Tourism Resources Inventory Project & Data Science for Social Good

Roman, Raphael | Þórhallsson, Halldór | Wu, Hailey | Zhu, Gary

2017 Summer

Table of Contents

List of Figures

Executive Summary

This report is the result of a partnership between the UBC Data Science for Social Good (DSSG) Fellowship and the Tourism and Creative Sectors Branch (TCSB) of the B.C. Ministry of Tourism, Arts and Culture (MTAC). The DSSG Fellowship was looking for suitable projects, and the TCSB was in the early stages of the Tourism Resources Inventory Project (TRIP). Through a connection made by staff from the B.C. Centre for Data-Driven Innovation, a partnership was created that resulted in the work described in this report.

The TRIP has two main objectives: **(1)** to gather and make accessible data relevant to tourism for the purposes of planning, policy development, and decision making, and **(2)** to develop and assign a system of value attributes or ranking categories to the tourism assets in B.C. Through collaboration and stakeholders consultation, this project was able to successfully explore and test approaches for both of the goals in the TRIP. Work to develop a proof-of-concept and pipeline for the project will be valuable as the broader TRIP project progresses.

This project partnership was also successful in proving the concept of the DSSG Fellowship itself, and provided valuable insights into future opportunities in data science partnerships between governments and academic institutions. Fellows are able to get exposure to real-world data challenges, and governments are able to use a laboratory approach with minimal resources and staff time in order to test concepts and ideas.

Above all else, the present report is the final product of the diligent work performed by the fellows throughout the entire 14-weeks of the DSSG program. By incorporating an amalgam of both quantitative and qualitative analysis, the team tapped into the natural interdisciplinarity of its highly qualified members, spanning a wide array of research interests, from machine learning, scientific computing and business intelligence to environmental economics and public policy.

The structure of the report follows 6 main axes, which, combined together, provide a comprehensive and coherent narrative of the "short" time window that the fellows had the opportunity to fructify. Introducing the data science challenge proposed by MTAC, the first chapter defines the problem at stake and connects the dots between relevant stakeholders.

The background will then give a concise overview of the main statistical trends and valuation modelling research that the B.C. tourism industry has been witnessing since the end of the 20th century; while offering a snapshot of the pre-analysis literature review that the fellows undertook to inform their prototype of a tailored and multifaceted value ranking system for British Columbia. Having laid the groundwork, the third chapter will present an overview of the data gathering process, detailing the manifold data sources and distinct typologies that the team had to cope with, prior to listing the "data gaps" and future data collection priorities that the government ought to pursue. The results and overall data analysis process will be displayed in chapter *four*, laying out the diverse quantitative and statistical toolkit applied by the fellows in the elaboration process of their end product. Finally, the social good component of the DSSG program will be reflected upon in chapter *five*, prior to concluding and offering recommendations for decision-makers and land-use planners in chapter *six*.

1. Introduction & Problem Statement

In July 2017, fire sparkles and the smoke roars. A shocking scene on CBC shows people forced to abandon their home for safety. A painful year for local tourism and the economy as a whole.

When two forest resorts are on fire, which one do you save first?

- *Bruce Whyte*

Tourism Resources Inventory Project (T.R.I.P) is one of the four projects at DSSG that strives to create and leverage an inventory of the tourism facilities, infrastructures, and resources important to developing the tourism sector in British Columbia. The goal is to create tools that better inform policy development, investment attraction, marketing, and decision making, and increase the impact and effectiveness of tourism planning processes.

As the principal collaborator of this project, the Ministry of Tourism, Arts and Culture (MTAC) manages key lines of government services that help support the tourism sector and provide opportunities for economic growth in all areas of the province. This means seizing the opportunities and responding to the challenges of a globalized economy in order to support jobs creation as well as provincial long-term growth.

Recent alignment of provincial government tools (e.g. BC Data Catalogue[1], iMap BC[2] and BC Economic Atlas[3]), executive support, and opportunities for applying new data tools in tourism planning form the business case for this project. It supports the goals described in the provincial tourism strategy *Gaining the Edge[4],* by facilitating access to timely and accurate data about the facilities, infrastructures, and resources that form the foundation of the tourism industry.

Multiple factors make it difficult to present an accurate picture of the supply side of tourism in B.C. In addition to the dispersed small businesses that form the majority of the industry, the responsibility for permitting and regulation of the industry is shared among multiple government agencies. Reflecting on the intertwined and intrinsic relationships that connect the work of the fellows with tourism stakeholders in B.C., **Figure 1** below provides a simplified and informative stakeholders' map, depicting the multidimensional aspects that the DSSG project touched upon in the time span of summer 2017 (see **Appendix 2** for a detailed and comprehensive list of stakeholders).

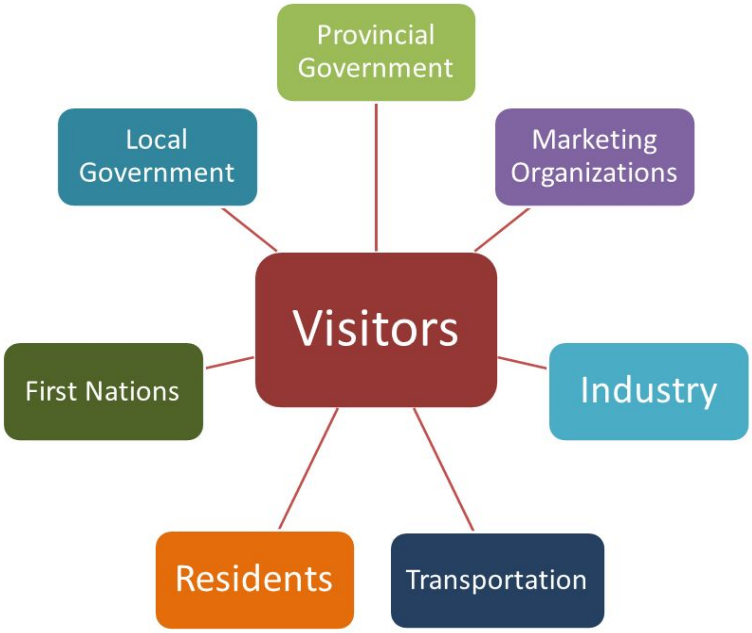


Figure 1: BC Tourism stakeholders' map

[1] BC Data Catalogue web link: https://catalogue.data.gov.bc.ca/dataset?download_audience=Public
[2] iMap BC web link: http://www2.gov.bc.ca/gov/content/data/geographic-data-services/web-based-mapping/imapbc
[3]BC Economic Atlas web link: http://www2.gov.bc.ca/gov/content/employment-business/economic-development/plan-and-measure/bc-economic-atlas
[4] Destination British Columbia. 2017. *Value of Tourism – Trends from 2005-2015*. See: http://www.destinationbc.ca/getattachment/Research/Industry-Performance/Value-of-Tourism/Value-of-Tourism-in-British-Columbia-(2015)/Value-of-Tourism_2015_FINAL.pdf.aspx

2. Background

Besides natural resources extraction, tourism is another major economic backbone of the province of British Columbia, supporting the socio-economic development of manifold rural communities scattered across a landmass of 940,000 square km. Due to its unusual topography, B.C. constitutes a unique setting for thriving natural assets, exceptional scenery and high quality of life, making it a prized destination for tourists seeking outdoor adventures or simply looking for nature connectedness.

2.1. B.C. Tourism Statistical Trends

To put it into numbers, in 2015, the tourism industry generated around $15.7 billion in revenues, corresponding to a 37.3% increase from 2005; while more than 127,000 people were employed in tourism-related businesses, which represents a 16% increase since 2005[5]. With its share of provincial GDP ratcheting up since the onset of the 21st century, tourism has witnessed an economic boon, which led to further investments in financial and political capital, with notably the creation of Destination BC in 2013, an industry-led Crown corporation that aims to market B.C. as a tourism destination to domestic, national and international travellers[6]. Since then, numerous strides have been reached, especially when it comes to attracting visitors from emerging and priority markets (e.g. China, India, Germany, California, Alberta…), with 2016 marking one of the most proliferous years on record, where 5.5 million international tourists visited the province of B.C. (which is more than the total population of the province) [7].

[5] Destination British Columbia. 2017. *Value of Tourism – Trends from 2005-2015*. See: http://www.destinationbc.ca/getattachment/Research/Industry-Performance/Value-of-Tourism/Value-of-Tourism-in-British-Columbia-(2015)/Value-of-Tourism_2015_FINAL.pdf.aspx

[6] Province of British Columbia, Minister of Jobs, Tourism and Skills Training and Minister Responsible for Training. 2015. *Gaining the Edge: 2015-2018, British Columbia's Tourism Strategy.* See: http://www2.gov.bc.ca/assets/gov/tourism-and-immigration/tourism-industry-resources/gainingtheedge_2015-2018.pdf

[7] Province of British Columbia, Minister of Jobs, Tourism and Skills Training and Minister Responsible for Training. 2017. *Gaining the Edge, A Progress Update, March 2017.* See:http://www2.gov.bc.ca/assets/gov/tourism-and-immigration/tourism-industry-resources/gainingtheedge_statusupdate_2017.pdf

2.2. Tourism Valuation Challenge: A Historical Perspective

Although the Tourism Resources Inventory Project (T.R.I.P.) is in its infancy, tourism valuation modelling techniques pertaining to B.C. have been developed since the midst of the 1990s, at a time where the main goal of such tourism-related projects was to devise and compile a comprehensive and simple GIS-based tourism inventory, investigating which features[8] are deemed of interest from both provincial and regional perspectives. With GIS becoming a widely accepted mapping technology, its functionalities have progressively allowed its users to reach greater levels of granularity, considerably facilitating planning strategies at the local or community scale for instance. Nonetheless, with much of the tourism developments happening at the site or asset level (1:1), meagre research record in that regard spearheaded the current Tourism Resources Inventory Project that the DSSG fellows have been working upon during summer 2017. Filling this gap will entail more than just simply locating every tourism asset[9] and determining the appropriate inventory valuation features, it will also allow policy-makers to make sound data-driven decisions at the tourism site or asset level.

2.3. Literature Review

Prior to focusing their attention on the case study of British Columbia, the team of DSSG fellows performed a "state-of-the-art" literature review on different value ranking systems that had been used across the world. Ranging from China[10], India[11] and Malaysia[12] to the Canadian province of Alberta[13], research pundits had been devising innovative and distinct value ranking methodologies, offering a quantitative measure of the potential and growth prospects of local and regional tourism economies. After accounting for the geographical, political, cultural and socio-economic context of their area of interest, the authors generally employed multi-criteria evaluation approaches while subjectively (and/or objectively) assigning relative weights to tourism features deemed relevant to the study. The **Figure 2** below offers a quick overview of the insights garnered during the review process.

[8] Here the notion of "features" relates to attributes or characteristics proper to the tourism assets at stake. Distance to the nearest airport, social media's attractiveness and job creation potential are, inter alia, a few examples of tourism-related features.

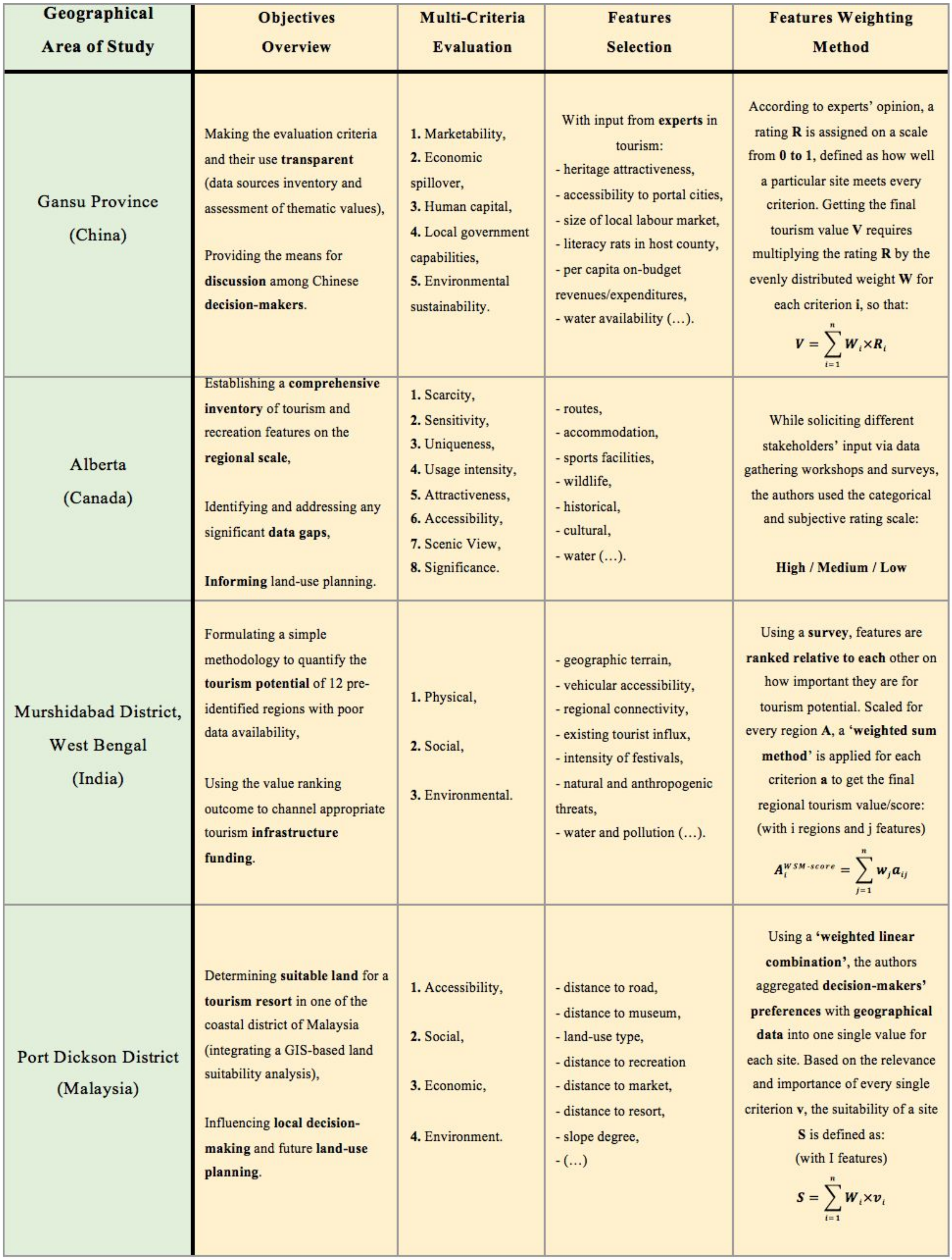
[9] A tourism "asset" is the report's "umbrella term" for all tourism sites, resources, facilities and infrastructures that are of interest to the main client of the DSSG team (e.g. parks, museums, hotels, lodges, airports, roads, ferries,...).

[10] Nick Novakowski, Rémy Tremblay and Edward Leman. 2008. Ranking Tourism Attractions According to their Suitability for Public Investment in Gansu Province, China. *Téoros*, 27-1, 59-66. See: http://teoros.revues.org/1597

[11] Al Mamun, Abdulla and Soumen Mitra. 2012. A Methodology for Assessing Tourism Potential: Case Study Murshidabad District, West Bengal, India. *International Journal of Scientific and Research Publications*, Volume 2, Issue 9.

[12] Liaghat Mahsa, Himan Shahabi, Bashir R. Deilami, Farshid S. Ardabili, Seyed N. Seyedi, and Hadi Badri. 2013. A multi-criteria evaluation using the analytic hierarchy process technique to analyze coastal tourism sites. *APCBEE Procedia* 5: 479-85.

[13] O2 Planning + Design Inc. 2010. Alberta Recreation and Tourism Features Inventory, Procedures and Standards Manual V1.02.



Geographical Area of Study	Objectives Overview	Multi-Criteria Evaluation	Features Selection	Features Weighting Method
Gansu Province (China)	Making the evaluation criteria and their use **transparent** (data sources inventory and assessment of thematic values), Providing the means for **discussion** among Chinese **decision-makers**.	**1.** Marketability, **2.** Economic spillover, **3.** Human capital, **4.** Local government capabilities, **5.** Environmental sustainability.	With input from **experts** in tourism: - heritage attractiveness, - accessibility to portal cities, - size of local labour market, - literacy rats in host county, - per capita on-budget revenues/expenditures, - water availability (…).	According to experts' opinion, a rating **R** is assigned on a scale from **0 to 1**, defined as how well a particular site meets every criterion. Getting the final tourism value **V** requires multiplying the rating **R** by the evenly distributed weight **W** for each criterion **i**, so that: $V = \sum_{i=1}^{n} W_i \times R_i$
Alberta (Canada)	Establishing a **comprehensive inventory** of tourism and recreation features on the **regional scale**, Identifying and addressing any significant **data gaps**, **Informing** land-use planning.	**1.** Scarcity, **2.** Sensitivity, **3.** Uniqueness, **4.** Usage intensity, **5.** Attractiveness, **6.** Accessibility, **7.** Scenic View, **8.** Significance.	- routes, - accommodation, - sports facilities, - wildlife, - historical, - cultural, - water (…).	While soliciting different stakeholders' input via data gathering workshops and surveys, the authors used the categorical and subjective rating scale: **High / Medium / Low**
Murshidabad District, West Bengal (India)	Formulating a simple methodology to quantify the **tourism potential** of 12 pre-identified regions with poor data availability, Using the value ranking outcome to channel appropriate tourism **infrastructure funding**.	**1.** Physical, **2.** Social, **3.** Environmental.	- geographic terrain, - vehicular accessibility, - regional connectivity, - existing tourist influx, - intensity of festivals, - natural and anthropogenic threats, - water and pollution (…).	Using a **survey**, features are **ranked relative to each** other on how important they are for tourism potential. Scaled for every region **A**, a **'weighted sum method'** is applied for each criterion **a** to get the final regional tourism value/score: (with i regions and j features) $A_i^{WSM\text{-}score} = \sum_{j=1}^{n} w_j a_{ij}$
Port Dickson District (Malaysia)	Determining **suitable land** for a **tourism resort** in one of the coastal district of Malaysia (integrating a GIS-based land suitability analysis), Influencing **local decision-making** and future **land-use planning**.	**1.** Accessibility, **2.** Social, **3.** Economic, **4.** Environment.	- distance to road, - distance to museum, - land-use type, - distance to recreation - distance to market, - distance to resort, - slope degree, - (…)	Using a **'weighted linear combination'**, the authors aggregated **decision-makers' preferences** with **geographical data** into one single value for each site. Based on the relevance and importance of every single criterion **v**, the suitability of a site **S** is defined as: (with I features) $S = \sum_{i=1}^{n} W_i \times v_i$

Figure 2: Synopsis of reviewed literature with associated value ranking methodologies

2.4. B.C. Value Ranking System

Following a careful and organized inspection of the insights garnered from the aforementioned literature review, the team of DSSG fellows ended up devising their own value ranking system tailored to tourism in the province of British Columbia. With the ultimate goal of computing a multifaceted value for each B.C.'s tourism asset, this custom-built ranking system's initial purpose was to serve as a guiding analytical framework for strategic modelling purposes. Accounting for geographical and contextual idiosyncrasies, the B.C. ranking system is supported by *five* main themes or categories, namely: **Accessibility, Physical Capital, Human Capital, Natural and Cultural Resources as well as Local Government Capabilities.**

Within each category, the team brainstormed an optimal and diverse list of tourism-related data that would be valuable to garner, as is depicted by the analytic hierarchical process shown in **Figure 3** below. Such data sets could be further used for features computations and thus value ranking modelling, as we will cover in chapter 4, while chapter 3 will provide additional details on the actual data gathering process.

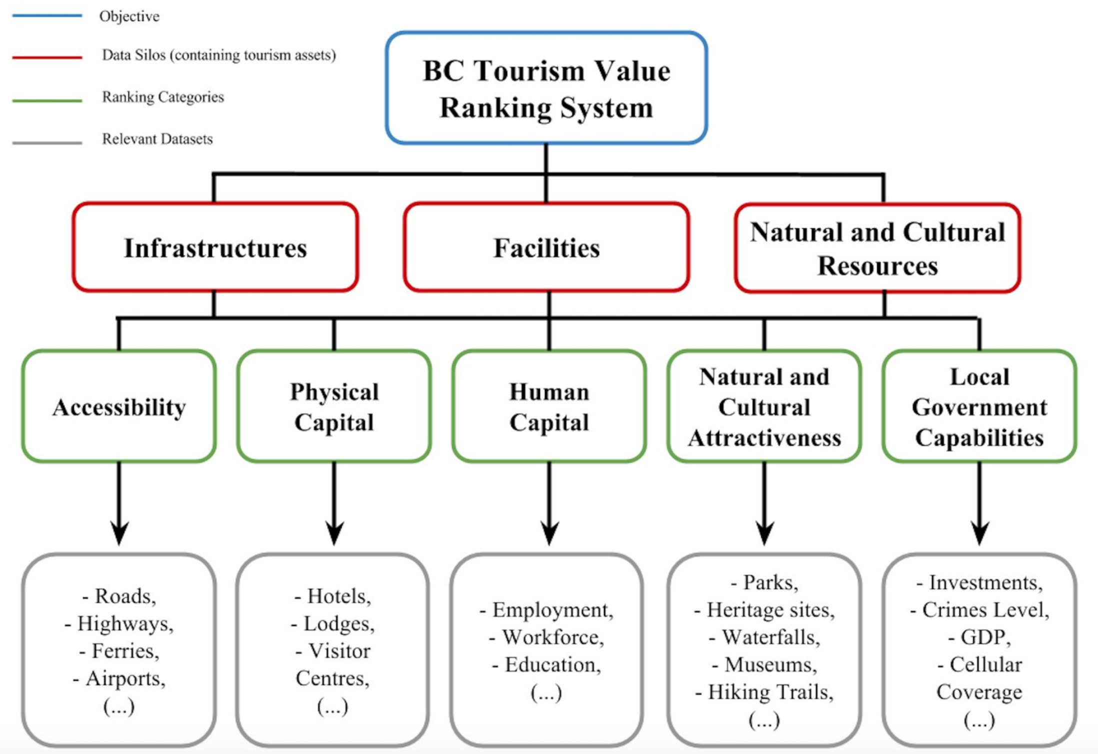

Objective
Data Silos (containing tourism assets)
Ranking Categories
Relevant Datasets
BC Tourism Value Ranking System
Infrastructures
Facilities
Natural and Cultural Resources
Accessibility
Physical Capital
Human Capital
Natural and Cultural Attractiveness
Local Government Capabilities
- Roads, - Highways, - Ferries, - Airports, (...)
- Hotels, - Lodges, - Visitor Centres, (...)
- Employment, - Workforce, - Education, (...)
- Parks, - Heritage sites, - Waterfalls, - Museums, - Hiking Trails, (...)
- Investments, - Crimes Level, - GDP, - Cellular Coverage (...)

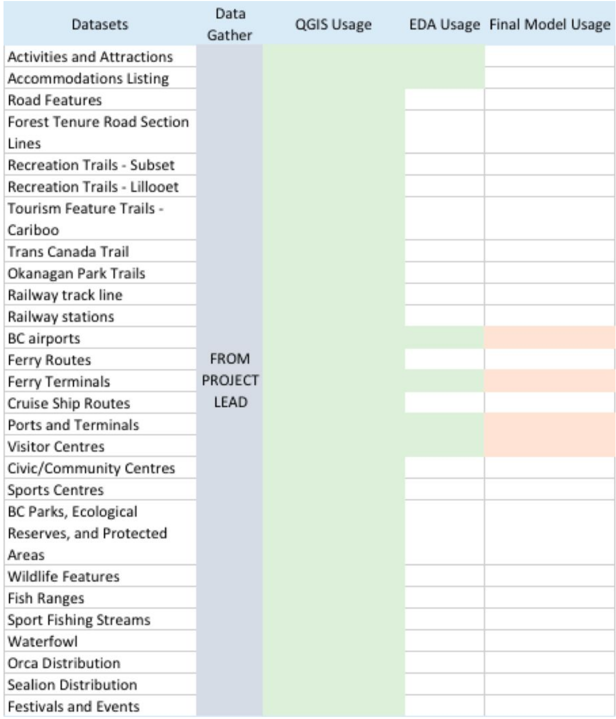
Figure 3: Analytic hierarchical process for categorizing B.C.'s ranking system

3. Data Gathering Process

3.1. Open Source Data Sets from BC Provincial Government

At the onset of the fellowship, the team received 27 initial government and public data sets from the project lead[14] (see **Figure 4**). Those data sets were divided into three major categories or silos: *Facilities*, *Infrastructures* and *Resources*. Specifically, *Facilities* included activities, attractions, and accommodations, while *Infrastructures* contained, among other things, data on roads, trails, airports and ferry terminals. Last but not least, *Resources* consisted of parks, wildlife, sport fishing streams et cetera.

The data was downloadable in various formats, such as CSV[15] and shapefile[16]. While those data sets were either public or restricted[17], it is important to note that the fellows did not utilize all of the initial data provided to them. Some of the main reasons involved a lack of information on how to incorporate high-dimensional vector data (e.g. lines and polygons) as well as an inability to consider meaningful features from limited and imperfect data sets (see **Appendix 3** for more details in that regard).



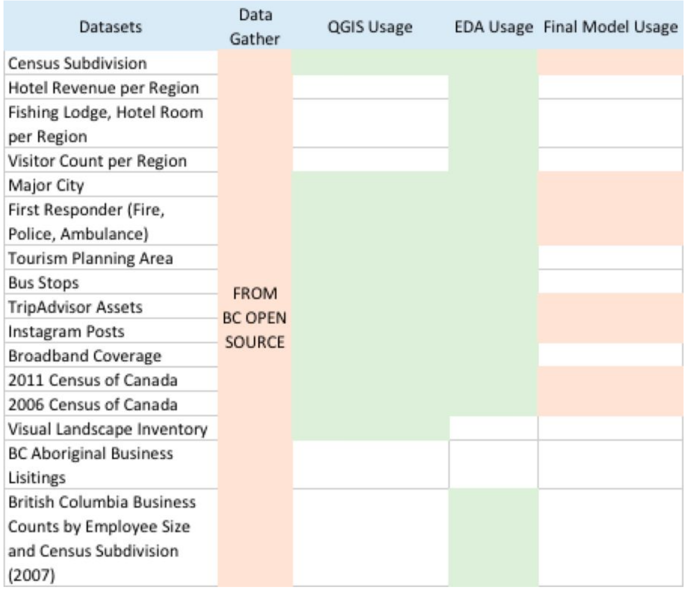
Datasets	Data Gather	QGIS Usage	EDA Usage	Final Model Usage
Activities and Attractions	FROM PROJECT LEAD	🟩	🟩	
Accommodations Listing		🟩	🟩	
Road Features		🟩		
Forest Tenure Road Section Lines		🟩		
Recreation Trails - Subset		🟩		
Recreation Trails - Lillooet		🟩		
Tourism Feature Trails - Cariboo		🟩		
Trans Canada Trail		🟩		
Okanagan Park Trails		🟩		
Railway track line		🟩		
Railway stations		🟩		
BC airports		🟩	🟩	🟧
Ferry Routes		🟩		
Ferry Terminals		🟩	🟩	🟧
Cruise Ship Routes		🟩		
Ports and Terminals		🟩	🟩	🟧
Visitor Centres		🟩	🟩	🟧
Civic/Community Centres		🟩		
Sports Centres		🟩		
BC Parks, Ecological Reserves, and Protected Areas		🟩		
Wildlife Features		🟩		
Fish Ranges		🟩		
Sport Fishing Streams		🟩		
Waterfowl		🟩		
Orca Distribution		🟩		
Sealion Distribution		🟩		
Festivals and Events		🟩		

Figure 4: Data initially garnered by the project lead

Besides the aforementioned list of data provided by the project lead, the team extracted additional data sets from both the BC Data Catalogue and the 2006 Census of Canada (see **Figure 5**); containing, inter alia, city populations, employment rates, fire stations, protected areas and Aboriginal businesses. With support from various stakeholders, the team was able to access both public and restricted data sets, allowing them to use a substantial fraction of the available data during the Exploratory Data Analysis (EDA) process.

[14] The project lead's name is Ben Clark, a senior policy analyst in the Tourism policy branch of the Minister of Jobs, Trade and Technology. Although physically located in Victoria, Ben held weekly digital meetings (2 hours) with the team of DSSG fellows.
[15] CSV stands for "Comma Separated Values", which can be opened with spreadsheet softwares such as Excel for instance.
[16] Shapefiles are a type of data format that consists of geospatial vectors that can be used with the GIS software.
[17] Publicly restricted data sets refer to data sets that require usage permission from data custodians prior to public download.

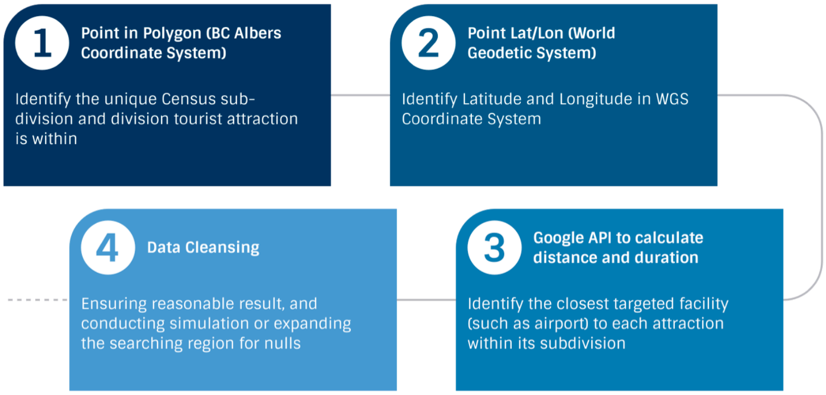
Nonetheless, due to unsuitable granularity levels and time constraints, not all the data used in the EDA process have been incorporated in the final valuation modelling approach (see **Appendix 4** for further details in that regard).



Datasets	Data Gather	QGIS Usage	EDA Usage	Final Model Usage
Census Subdivision	🟧 FROM BC OPEN SOURCE	🟩	🟩	🟧
Hotel Revenue per Region	🟧		🟩	
Fishing Lodge, Hotel Room per Region	🟧		🟩	
Visitor Count per Region	🟧		🟩	
Major City	🟧	🟩	🟩	🟧
First Responder (Fire, Police, Ambulance)	🟧	🟩	🟩	🟧
Tourism Planning Area	🟧	🟩	🟩	
Bus Stops	🟧	🟩	🟩	
TripAdvisor Assets	🟧	🟩	🟩	🟧
Instagram Posts	🟧	🟩	🟩	🟧
Broadband Coverage	🟧	🟩	🟩	
2011 Census of Canada	🟧	🟩	🟩	🟧
2006 Census of Canada	🟧	🟩	🟩	🟧
Visual Landscape Inventory	🟧	🟩		
BC Aboriginal Business Lisitings	🟧			
British Columbia Business Counts by Employee Size and Census Subdivision (2007)	🟧		🟩	

Figure 5: Data collection from open sources

At this stage, a majority of the data garnered by the fellows was containing “static” variables, such as the location of airports, ferry terminals (...). Using external resources and services such as Google API[18] (see **Figure 6**) and OpenStreetMap[19], the team was able to explore further some of the meaning hidden behind the static data, computing features such as distance and duration between two different tourism assets (e.g. between a hotel and the nearest airport).



1 Point in Polygon (BC Albers Coordinate System)

Identify the unique Census sub-division and division tourist attraction is within

2 Point Lat/Lon (World Geodetic System)

Identify Latitude and Longitude in WGS Coordinate System

3 Google API to calculate distance and duration

Identify the closest targeted facility (such as airport) to each attraction within its subdivision

4 Data Cleansing

Ensuring reasonable result, and conducting simulation or expanding the searching region for nulls

Figure 6: Google API usage procedure flow

[18] API stands for Application Program Interface, see: https://en.wikipedia.org/wiki/Application_programming_interface

[19] OpenStreetMap: https://en.wikipedia.org/wiki/OpenStreetMap

3.2. Social Media Data from Instagram

3.2.1. Defining the Relative Value of Tourism

One of the predominant concerns during the early stages of the team's data gathering process was the lack of a "qualified response variable"[20]. The failure to find a reasonable response (or dependent) variable from the aforementioned open-data sources (i.e. government and public) made it complex to predict and approximate a final tourism score or value that the project lead anticipated at the initial stage. Indeed, while the task was described as an exploratory trial to see the possibility of evaluating each tourism asset from all existing data sources, limitations in the availability of relevant "proxies" such as the total number of visitors or annual revenues per asset prevented the team from running meaningful analysis. In addition, due to the intangible, tacit and multifaceted nature of the notion of "value" the team had been reflecting upon, finding a coherent proxy variable became a challenge that required further communication with the project lead. The resulting thought process incentivized the team to seek out alternative data sources, naturally leaning towards social media such as Instagram, Twitter, Airbnb and TripAdvisor, to mention a few. Investigating data from Instagram could provide the fellows with valuable and geolocalized point data (or Instagram counts), offering an optimal granularity level as well as a promising response variable to value tourism assets across B.C.

3.2.2. Bias Concerns

With Instagram closing its API for research purposes in 2017, the team ended up extracting one-month (June to July) worth of Instagram statistics, pictures and captions that were associated with the hashtag "ExploreBC". The resulted data set was highly biased towards not only natural resources such as provincial and national parks, but also towards Instagram posts related to summer in general, ruling out any seasonality analysis (see section 4.3. for spatial and natural language processing analysis that the fellows performed on Instagram data) .

[20] In other words, based on the data garnered from government and additional public sources, the team was unable to identify an appropriate (and statistically robust) proxy variable for the relative value of tourism regarding each asset.

3.3. Social Media Data from TripAdvisor

3.3.1. Lack of Integrated Data Set on Natural and Cultural Resources

Acquiring a comprehensive and exhaustive enough list of tourism assets for British Columbia was critical to the project. Nonetheless, following stakeholders' consultation, and accounting for the fact that former B.C. tourism studies have primarily focused their attention on *Facilities* and *Infrastructures*, the team recognized the high value-added from *Natural and Cultural Resources*. Narrowing the scope of the project (as detailed in section 3.6.), the fellows concentrated their efforts on garnering a detailed list of B.C.'s natural and cultural assets, without ruling out future integration of facilities and infrastructures.

Data sets on *Natural and Cultural Resources* were scattered in distinct open government data sources[21], but due to the heterogeneity of the data, integrating all sources would have been a resource intensive task. TripAdvisor, on the other hand, contained a homogeneous structured data set on *Natural and Cultural Resources* across B.C., albeit using a different typology (see section 3.4.). Data from TripAdvisor included granular variables, such as, inter alia, bubble rating, review count and location of each natural and cultural asset, which allowed the team to compute a substantial amount of features, as the report shows in chapter 4.

3.3.2. Point versus Polygon Spatial Data and Typology Distinction

TripAdvisor structures all tourism assets as point data on Google map, which lost the spatial information such as the area and the perimeter. For a specific type of assets such as parks, a reasonable assessment of travelling distance with a car should mark the destination location at the park entrance. Nonetheless, with only the information of latitude and longitude of the spatial point located at the center of a park, Google Matrix API failed to give a precise estimation of accessibility features (e.g. distance to nearest park).

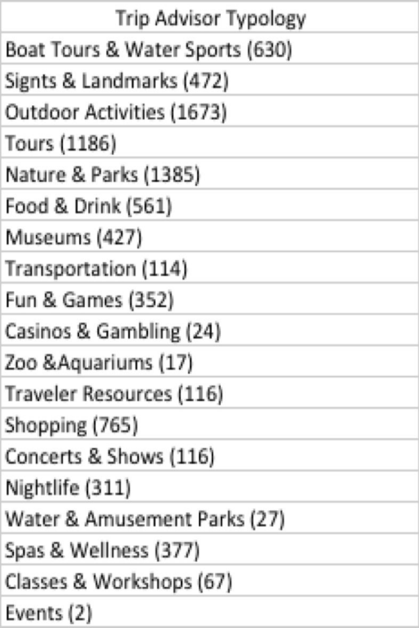
Furthermore, TripAdvisor categorizes the assets in a different typology than the one used by open government data sources (see section 3.4.) . There was no unique shared identifier between the two data sets. The only method to identify common assets between the two data sources was to perform a manual check. Thus, within the time constraint, the team decided to postpone the labelling and filtering processes of TripAdvisor assets and see whether or not it belongs to the government data set.

[21] For instance, "DataBC" was frequently used by the fellows, providing access to manifold government's data holdings, such as the B.C. Data Catalogue or iMap B.C., to name a few. http://www2.gov.bc.ca/gov/content/data/about-data-management/databc

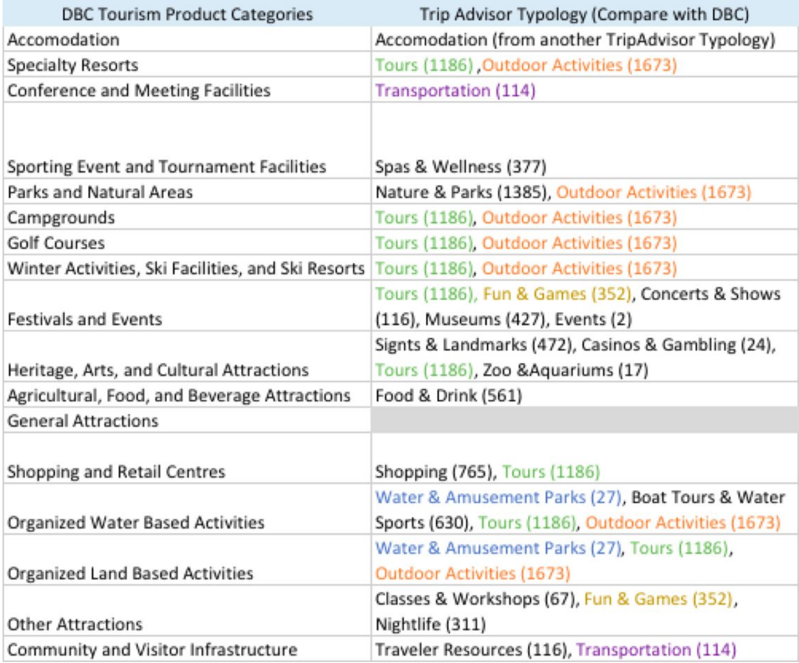
The team also noted some discrepancies regarding the geolocation of particular tourism assets (e.g. Vargas Island). Although included into the current analysis, data from TripAdvisor will need to go through a quality assurance process for location accuracy.

3.4. Comparing Typologies

Figure 7 (top table) refers to the data typology of TripAdvisor. The number within brackets is the count of relevant B.C.'s tourism assets present in TripAdvisor. Each "category" contains distinct subcategories, while it is important to note that one asset can appear several times under different TripAdvisor categories. For instance, "boat trip" can be spotted both in *Boat Tours & Water Sports (630)* as well as in *Tours (1186).* **Figure 7 (right table)** compares the data typology of TripAdvisor and "Destination BC (DBC) Tourism Product Categories"[22]. As outlined below, each TripAdvisor category successfully fits into one particular category of the DBC typology, although there are no clear one-to-one relationships between the two typologies when it comes to *Natural and Cultural Resources*.



Trip Advisor Typology
Boat Tours & Water Sports (630)
Signts & Landmarks (472)
Outdoor Activities (1673)
Tours (1186)
Nature & Parks (1385)
Food & Drink (561)
Museums (427)
Transportation (114)
Fun & Games (352)
Casinos & Gambling (24)
Zoo &Aquariums (17)
Traveler Resources (116)
Shopping (765)
Concerts & Shows (116)
Nightlife (311)
Water & Amusement Parks (27)
Spas & Wellness (377)
Classes & Workshops (67)
Events (2)



DBC Tourism Product Categories	Trip Advisor Typology (Compare with DBC)
Accomodation	Accomodation (from another TripAdvisor Typology)
Specialty Resorts	Tours (1186) ,Outdoor Activities (1673)
Conference and Meeting Facilities	Transportation (114)
Sporting Event and Tournament Facilities	Spas & Wellness (377)
Parks and Natural Areas	Nature & Parks (1385), Outdoor Activities (1673)
Campgrounds	Tours (1186), Outdoor Activities (1673)
Golf Courses	Tours (1186), Outdoor Activities (1673)
Winter Activities, Ski Facilities, and Ski Resorts	Tours (1186), Outdoor Activities (1673)
Festivals and Events	Tours (1186), Fun & Games (352), Concerts & Shows (116), Museums (427), Events (2)
Heritage, Arts, and Cultural Attractions	Signts & Landmarks (472), Casinos & Gambling (24), Tours (1186), Zoo &Aquariums (17)
Agricultural, Food, and Beverage Attractions	Food & Drink (561)
General Attractions	
Shopping and Retail Centres	Shopping (765), Tours (1186)
Organized Water Based Activities	Water & Amusement Parks (27), Boat Tours & Water Sports (630), Tours (1186), Outdoor Activities (1673)
Organized Land Based Activities	Water & Amusement Parks (27), Tours (1186), Outdoor Activities (1673)
Other Attractions	Classes & Workshops (67), Fun & Games (352), Nightlife (311)
Community and Visitor Infrastructure	Traveler Resources (116), Transportation (114)

Figure 7: (top) TripAdvisor typology, (right) Typology distinction between TripAdvisor and DBC

-

Note: (1) Accommodation (hotels, lodges…) is part of the *Facilities* typology,

(2) The *General Attractions* category is greyed out due to unclear definition.

[22] "DBC Tourism Product Categories" is used as the provincial tourism data typology for policy, marketing and economic analysis.

Another typology distinction/comparison, this time focusing upon TripAdvisor and the "North American Industry Classification System" (NAICS) is available in **Appendix 5**. In a nutshell, the two typologies do not correlate well (due to just a few matching categories), while there are no clear one-to-one relationships between the two typologies.

3.5. Data Gaps

Identifying "data gaps" required investigating missing data sets that the team deemed valuable and insightful to strengthen and upheld the initial value ranking system showed in section 2.4. Specifically, within each of the five ranking categories, the fellows brainstormed and listed all relevant and realistic data that could add value to the project and enhance the predictive power of their statistical modelling. After having garnered all available data sets from both open and alternative sources, the team did the sum and identified which data sets were missing, sharing a non-exhaustive, albeit meaningful list in **Figure 8**, providing a guidance for tourism authorities regarding future data gathering processes. Three particular data gaps are justified below:

Accommodation Bed Count

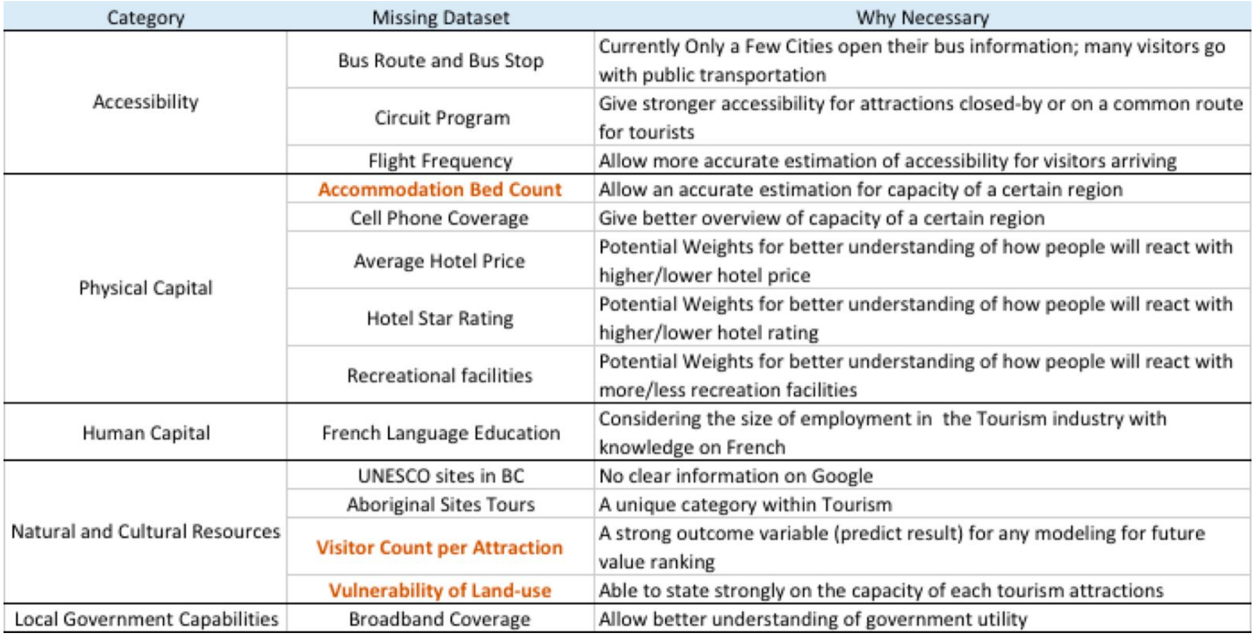
Accounting for distinctive accommodations (e.g. hotels, lodges as well as Bed and Breakfasts) ought to be a predominant aspect of the "Physical Capital" ranking category (see section 2.4.). Nonetheless, available data on accommodations only focuses on their location, count and thus geographical accessibility. Reporting and collecting data on accommodation bed count would be valuable in providing a better estimation of the physical or carrying capacity of each tourism facility.

Visitor Count per Asset

The number of visitors per tourism asset has been considered as one of the best estimation of the relative value of tourism in B.C. Using it as a statistical proxy (or response variable) would answer a few critical questions such as: "How can tourism authorities improve visitors' experiences?", "What features weigh the most in attracting further tourists?", "What are the local economic and environmental impacts of an increase in the number of tourists/visitors?".

Vulnerability of Land-use

Vulnerability of land-use, or land carrying capacity, can provide key insights on how many visitors a particular site/asset can hold and sustain during a certain period of time. Gathering such data will require collaboration with specialized research groups in order to conduct accurate estimations.



Category	Missing Dataset	Why Necessary
Accessibility	Bus Route and Bus Stop	Currently Only a Few Cities open their bus information; many visitors go with public transportation
	Circuit Program	Give stronger accessibility for attractions closed-by or on a common route for tourists
	Flight Frequency	Allow more accurate estimation of accessibility for visitors arriving
Physical Capital	**Accommodation Bed Count**	Allow an accurate estimation for capacity of a certain region
	Cell Phone Coverage	Give better overview of capacity of a certain region
	Average Hotel Price	Potential Weights for better understanding of how people will react with higher/lower hotel price
	Hotel Star Rating	Potential Weights for better understanding of how people will react with higher/lower hotel rating
	Recreational facilities	Potential Weights for better understanding of how people will react with more/less recreation facilities
Human Capital	French Language Education	Considering the size of employment in the Tourism industry with knowledge on French
Natural and Cultural Resources	UNESCO sites in BC	No clear information on Google
	Aboriginal Sites Tours	A unique category within Tourism
	Visitor Count per Attraction	A strong outcome variable (predict result) for any modeling for future value ranking
	Vulnerability of Land-use	Able to state strongly on the capacity of each tourism attractions
Local Government Capabilities	Broadband Coverage	Allow better understanding of government utility

Figure 8: Missing data sets

3.6. Narrowed Focus

In the midst of the DSSG project, due to time constraints and lack of appropriate data vis-à-vis particular categories of the B.C. value ranking system (see data gaps in section 3.5.), the team switched gears and decided to narrow their approach by reshuffling their value ranking system while retaining only natural and cultural assets[23], leading to the selection of *three* main overarching themes of interest or ranking categories, that is: **Accessibility, Significance** and **Capacity**. Although the previous ranking system had not been ruled out (and will still be functional for future data-gathering recommendations), the newly devised one will befit the natural limitations of the project (time and resources wise) and allow the team to test the efficacy and potential of their valuation model (see end product in chapter 4). Additional details pertaining to the team's multifaceted value ranking system are available in **Figure 9** below, notably regarding instances of computable tourism features that can be incorporated within each new ranking category/theme.

[23] While the initial list of tourism assets is broken down into three main data silos, namely: facilities, infrastructures as well as natural and cultural resources, the latter appeared to bring more value-added to the main client of the project (i.e. MTAC). Indeed, previous works have been specifically devoted to tourism infrastructures and facilities, while research gaps pertaining to natural and cultural resources were still predominant prior to the start of the DSSG fellowship.

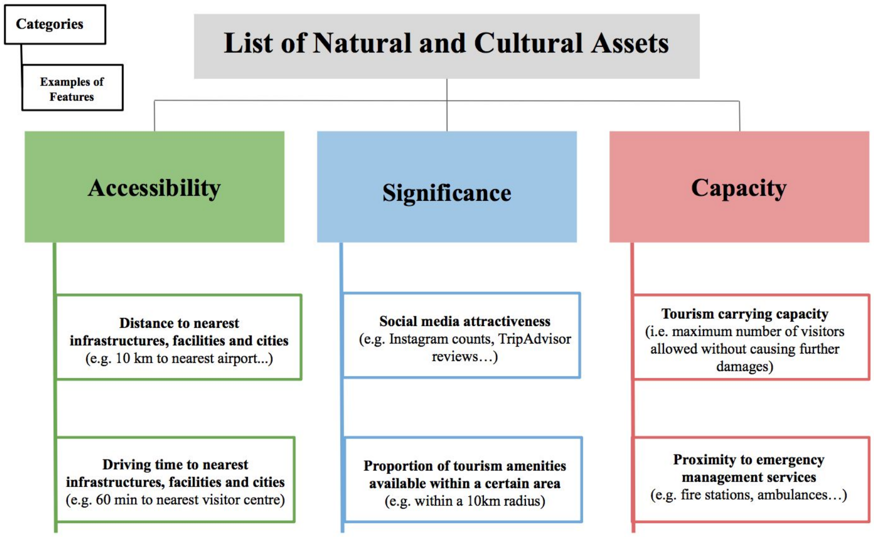

Categories
Examples of Features
List of Natural and Cultural Assets
Accessibility
Significance
Capacity
Distance to nearest infrastructures, facilities and cities (e.g. 10 km to nearest airport...)
Driving time to nearest infrastructures, facilities and cities (e.g. 60 min to nearest visitor centre)
Social media attractiveness (e.g. Instagram counts, TripAdvisor reviews…)
Proportion of tourism amenities available within a certain area (e.g. within a 10km radius)
Tourism carrying capacity (i.e. maximum number of visitors allowed without causing further damages)
Proximity to emergency management services (e.g. fire stations, ambulances…)

Figure 9: New narrowly devised B.C. value ranking system

4. Data Analysis Process & Results

4.1. Challenges

A predominant objective of the DSSG project was to build, devise and arrange a shareable data set ready for analysis, and primarily made from various public and alternative data sources that contain semi-structured data[24]. Due to the heterogeneous nature of many of those data sets, integration proved to be a difficult process (see chapter 3). The heterogeneous nature of the data was largely due to the distinct spatial granularity levels (e.g. regions, subdivisions and tourism assets). For instance, socio-economic data was mainly available at the census subdivision level, displaying a single value per area (e.g. unemployment rate for "Greater Victoria"). On the other hand, many tourism assets were available at the highest granularity level (1:1); as an example, the location of a museum or a summer festival would be attached to a single geocoordinate[25]. Following stakeholders' consultation, the team promptly learnt that it was essential to value and rank each tourism asset per se, prior to scaling it up to an aggregate value at the census subdivision level. This posed two immediate challenges:

(1) How could the team transform data on a lower granular level such as a census subdivision down to a higher level such as a polygon, line or a point coordinate?

[24] Semi-structured data can be defined as data that has not been organized into a specialized repository, such as "relational" databases and other forms of data tables, but that nevertheless contains information associated with it, such as metadata tagging, that allows contained elements to be addressed accordingly.

[25] Geographic coordinate such as latitude and longitude.

(2) How could the team deal with the heterogeneous spatial nature of the tourism assets? It would require significant resources to compute features for all the various spatial forms (line/point/polygon) of a tourism asset.

To address these challenges, as explained in section 3.6., the fellows decided to narrow the scope of their project, knowing that their model would ultimately be a "proof-of-concept" that could lead to further data-driven applications in provincial and international tourism policy-making. They ended up using "naive" approximations that were fast to compute and address the lower to higher granular level challenges. For instance, if a tourism asset (point) was located within a census subdivision, then particular subdivision attributes were automatically assigned to the point of interest. As for the tourism assets, the team decided to solely work with point data (the highest granular level on a homogeneous spatial form). Accordingly, they only performed quantitative analysis with *Natural and Cultural Resources,* as outlined in chapter 3, while assets that were not initially considered as point data but rather as polygons (e.g. parks, reserves), were ultimately represented as points while using TripAdvisor. Such approximations ought not to have a noticeable effect on the proof-of-concept model, however, if the data set is used as a decision-making tool, these approximations will have to be revised.

4.2. Tourism Assets Analysis

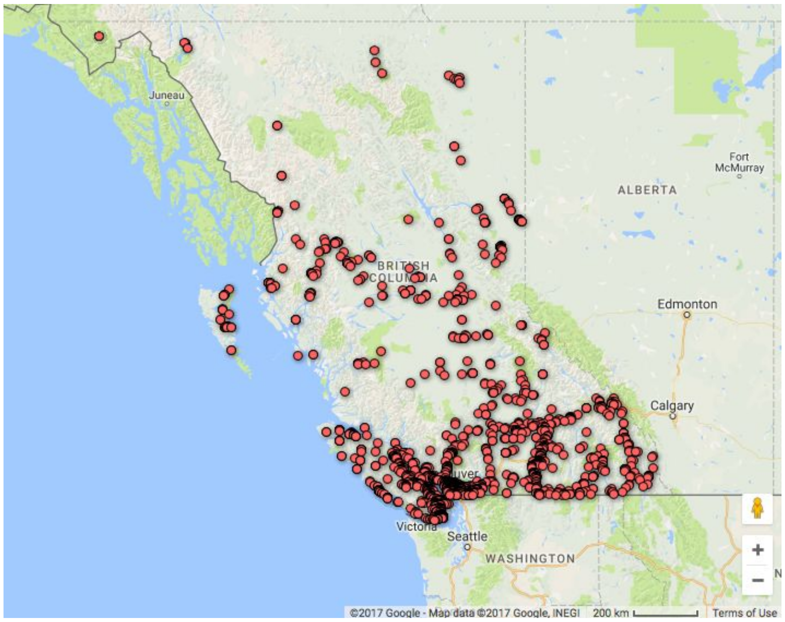


Figure 10: Map of natural and cultural tourism assets in B.C. using Google Fusion Tables[26]

-

Note: each natural and cultural tourism asset has been extracted from TripAdvisor.

[26] Gonzalez, Hector, et al. 2010. Google fusion tables: data management, integration and collaboration in the cloud. *Proceedings of the 1st ACM symposium on Cloud computing*.

On **Figure 10**, the spatial distribution of natural and cultural assets extracted from TripAdvisor is displayed. A key thing to note from it is that, using visual inspection, it seems that most of the natural and cultural resources are clustered in the Southern half part of B.C. They are especially close to cities and densely populated areas, such as, inter alia, Metro Vancouver, Greater Victoria and Kelowna. It might be evident for cultural resources such as museums, zoos and aquariums, but initially startling when it comes to natural resources. One possible reason might be related to the highly biased TripAdvisor data towards geographically accessible assets. However, the team was also confident that the majority of resources that qualified as tourism assets have had some sort of infrastructure built around them and therefore are bound to be in locations that are accessible and tend to be close to or in populated areas.

4.3. Social media analysis

4.3.1 Spatial Analysis of Instagram Posts

The team was able to extract roughly 41,000 geocodable Instagram posts in British Columbia, which were associated with one full summer month (namely, June to July 2017). As can be seen on **Figure 11**, the spatial distribution of Instagram posts closely follows the one related to tourism assets in **Figure 10**. The frequency of Instagram posts is heavily skewed towards southern B.C., or where the majority of the population is. It can be seen more clearly on **Figure 12**, that zooms-in on the Okanagan Valley, where the posts are heavily clustered in Kelowna and Kamloops, as well as around the Okanagan lake. This is a trend the team repeatedly observed when looking at Instagram posts, which are heavily clustered around natural assets like parks, lakes, rivers et cetera. This is due to a certain selection bias when extracting data from Instagram. Indeed, the fellows only extracted posts that were associated with the particular and popular hashtag **#explorebc**[27], inducing biases towards places with breathtaking scenery and pristine landscapes, without reflecting the population of Instagram posts in British Columbia. Thus, instead of using them as a proxy for visitor count, they might be a good proxy for the natural attractiveness of a tourism asset. Such a feature will be highly relevant to the *Significance* ranking category.

[27] See the main web page of Instagram posts that present the hashtag #exploreBC: https://www.instagram.com/explore/tags/explorebc/?hl=en

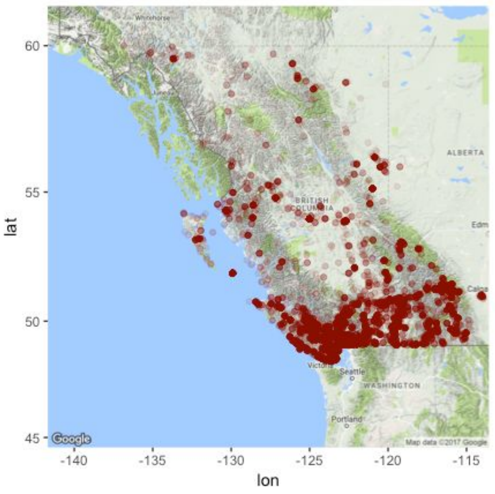


Figure 11: Spatial distribution of Instagram posts in British Columbia

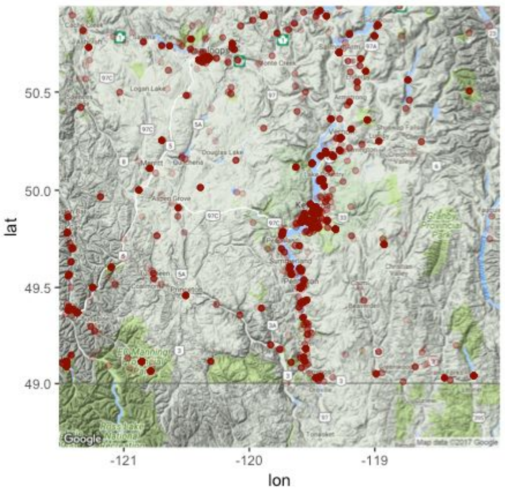


Figure 12: Spatial distribution of Instagram posts in the Okanagan Valley (zoom-in)

4.3.2. Natural Language Processing

As part of the Instagram's data extraction process, the fellows expressed interest in gathering "text data", representative of Instagram users' comments or captions that are attached to a particular photo taken in British Columbia. Using "text data" allowed the team to perform some valuable Natural Language Processing (NLP)[28] analysis, probing into the quality of tourism experiences offered by each particular site or asset in B.C.

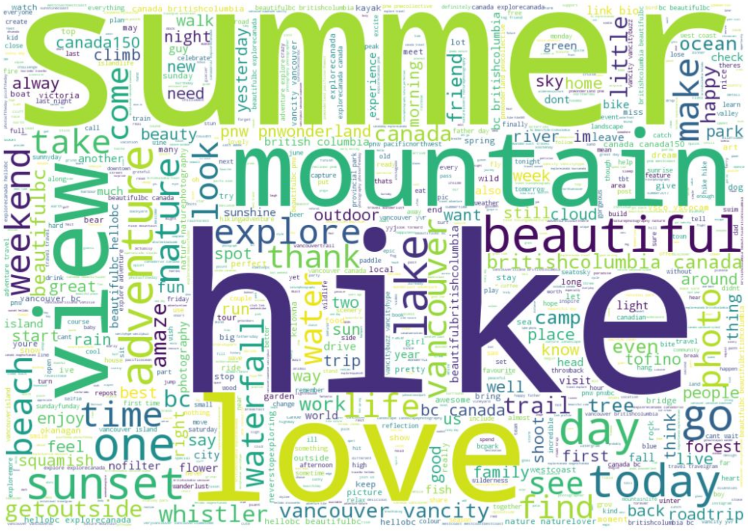


Figure 13: Word Cloud of most commonly referred keywords on Instagram

[28] Natural Language Processing (NLP) is a text mining methodology that aims to understand and process extensive natural language corpora, in a similar way that humans do. Combining the potential of computer science, artificial intelligence and computational linguistics, NLP is an ever evolving realm that allows its users to perform valuable analysis such as, among other things, text classification, machine translation, topic modelling and sentiment analysis.

As a result of further data cleansing[29] and polishing, the fellows investigated the most commonly referred keywords by Instagram users who mentioned the hashtag #explorebc in their captions. From this extensive text corpus, the team detected a word-frequency pattern that can be visualized via **Figure 13** above. Intuitively, it is not startling to observe such words as “hike”, “summer”, “mountains”, “view” and “sunset” making it to the top of our word-frequency list.

Although visually compelling, a word cloud is not sufficient in determining any geographical trends in the text data that could help assessing idiosyncratic topics. For instance, are visitors in the Okanagan region talking and posting about similar or distinct activities when compared with other visitors touring on Vancouver Island? Is it possible to identify any particular traits and customs for each tourism region[30] of B.C.? Employing document clustering techniques at the subdivision level, the team detected *five* distinct “topical clusters” as illustrated by **Figure 14** below. Interestingly, those clusters can be dissociated from each other when comparing their “topical keywords”. For instance, while in the Okanagan region the main topics of discussion revolved around wineries; fishing and camping were the main activities of interest for visitors touring the Northern part of British Columbia.

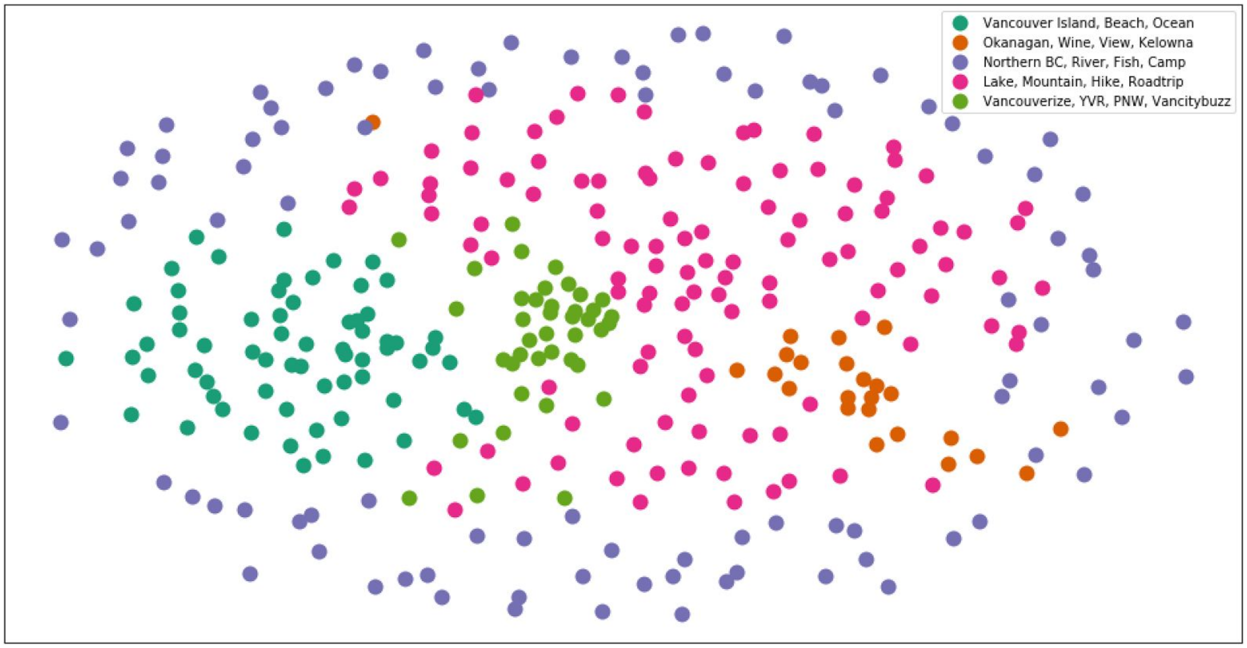


Figure 14: Document clustering performed on Instagram captions (point = subdivision)
*Legend (*top right corner*)*: synopsis of most commonly referred keywords within each cluster (non-exhaustive list)

[29] In the context of “text data”, data cleansing means the application of several filtering processes prior to the obtention of a final corpus on which the NLP analysis will be performed. Specifically, filtering methods include, among other things, the removal of “stopwords” (i.e. english words that do not bear any particular significance during search queries, e.g. I, we, what, who…), the “lemmatization” of the tokens used for analysis (i.e.returning the dictionary form of a word) and the filtering out of digits and words with less than 2 characters. Although not perfect, such filtering methods are necessary for a coherent text analysis.

[30] Here is a link to the geographical boundaries of each tourism region in British Columbia: http://www.destinationbc.ca/Programs/Regions-Communities-and-Sectors/Regional-Tourism-Programs/Regional-Partners.aspx

4.3.3 Tripadvisor Rating and its Implication on Significance

Besides what was already mentioned in section 3.3., there are a few other reasons the team ended up analyzing TripAdvisor data instead of the tourism data sets available in the BC Data Catalogue and other open data sources. First and foremost, the team was highly skeptical about the validity, comprehensiveness, and currency of data from BC Data Catalogue after observing certain inconsistencies from narratives' speech. Then, the team detected useful features, such as bubble rating and reviews[31], allowing the fellows to infer the attractiveness or significance potential associated with each tourism asset. Specifically, TripAdvisor bubble rating let visitors rate their personal experiences on a 1 to 5 scale, where 1 stands for "terrible", 3 is considered as "average" and 5 is "excellent", along with a comment that could potentially be used for Natural language Processing analysis. **Figure 15** below is a distribution of averaged ratings for all tourism assets in British Columbia .

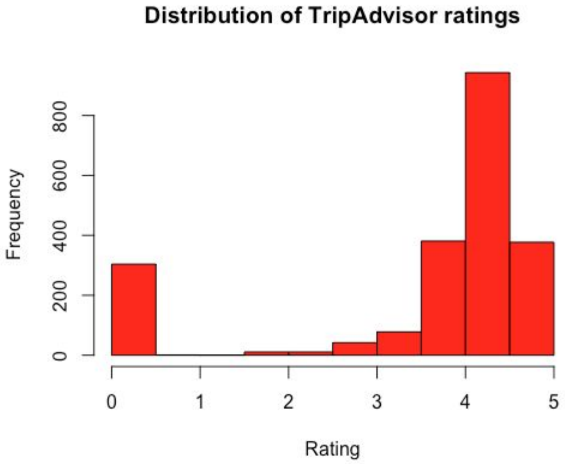


Figure 15: Distribution of TripAdvisor ratings

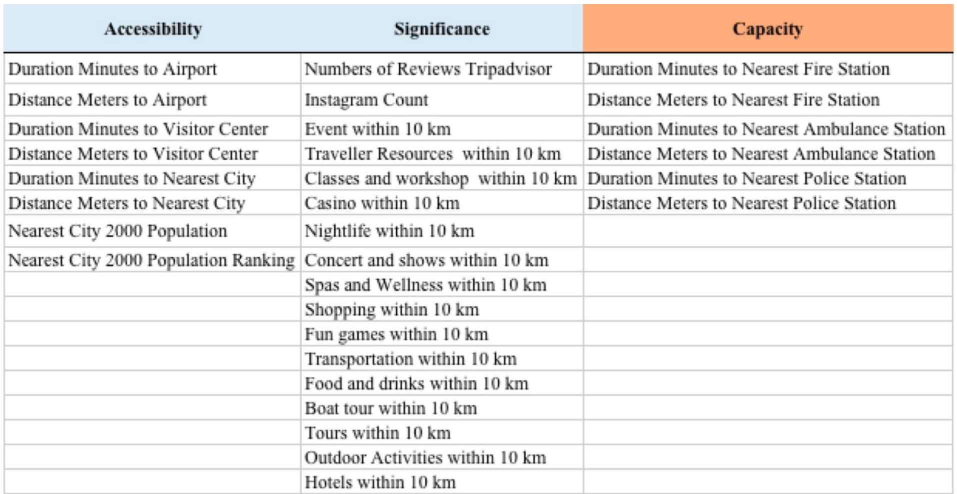
When people rate their experiences as "terrible", "poor" or "average", negative comments are easy to find, with people often complaining about high entrance fees, distance to travel, and unsatisfying camping conditions. It could be valuable to use TripAdvisor ratings as a feature assessing *Significance*, by looking at the percentage of each rating and see how much it varies across tourism assets. However, the downside of using TripAdvisor is its "incompleteness", namely, it only contains a small fraction of all provincial parks, beaches, and lakes, while excluding infrastructures like roads and railway stations (especially when looking at future data applications).

[31] Here is a link to the main TripAdvisor web page that the fellows extracted natural and cultural resources data from: https://www.tripadvisor.ca/Attractions-g154922-Activities-British_Columbia.html

4.4. Features Engineering

At this point, after having structured their final data set (although of varying spatial granularity), the team now had to come up with ways to compute meaningful features or attributes for the tourism assets they extracted earlier, and which could serve as an input into algorithms that output a score for each ranking category. A lot of these features were computed using spatial methods, such as counting the number of data points within a radius around an asset, checking whether an asset is present inside a polygon et cetera. This allowed for linking data sources to the tourism assets per se. For instance, concerning the Instagram data, the number of posts within a 10km radius of each tourism asset was assessed. This could prove to be another valuable input for measuring *Significance,* after what was previously mentioned in section 4.3. When it comes to the computation of features relevant to *Accessibility,* the team used the Google API for assessing travel duration from one point to another. As an example, for every tourism asset the travel duration to the nearest city was calculated using the API.

Figure 16 below displays all the "numerical features" that were computed and devised to test the model with. As indicated, many features ended up being highly correlated, such as the duration and distance to a certain target location, making it necessary to conduct features selection prior to any predictive modeling purposes. In addition, the *Capacity* ranking category only contained features associated with the distance and duration to the nearest fire and ambulance stations, missing on other key inputs such as tourism carrying capacity and vulnerability of land-use, to mention a few. As the team could not conduct any meaningful analysis on this particular ranking category, they offered recommendations for future data gathering processes in that regard. Consequently, the fellows focused their features engineering process and ranking modeling on both *Accessibility* and *Significance*.



Accessibility	Significance	Capacity
Duration Minutes to Airport	Numbers of Reviews Tripadvisor	Duration Minutes to Nearest Fire Station
Distance Meters to Airport	Instagram Count	Distance Meters to Nearest Fire Station
Duration Minutes to Visitor Center	Event within 10 km	Duration Minutes to Nearest Ambulance Station
Distance Meters to Visitor Center	Traveller Resources within 10 km	Distance Meters to Nearest Ambulance Station
Duration Minutes to Nearest City	Classes and workshop within 10 km	Duration Minutes to Nearest Police Station
Distance Meters to Nearest City	Casino within 10 km	Distance Meters to Nearest Police Station
Nearest City 2000 Population	Nightlife within 10 km	
Nearest City 2000 Population Ranking	Concert and shows within 10 km	
	Spas and Wellness within 10 km	
	Shopping within 10 km	
	Fun games within 10 km	
	Transportation within 10 km	
	Food and drinks within 10 km	
	Boat tour within 10 km	
	Tours within 10 km	
	Outdoor Activities within 10 km	
	Hotels within 10 km	

Figure 16: Detailed description of computed features for each ranking category

4.5. Assessing the Ranking Categories of a Tourism Asset Using Automatic Methods

In order to aggregate and make sense of the computed features that have been classified within each ranking category of interest (i.e. *Accessibility* and *Significance*), the fellows opted for a clustering methodology named "K-Means clustering"[32]. By partitioning features similarities within each ranking category, it is then possible to visualize and determine a categorical or numerical scale that will be ultimately employed for the value ranking system of tourism assets. A typical instance of ranking scale such as "High/Medium/Low" can help assessing the relative potential of each tourism asset regarding accessibility and social media attractiveness (or significance).

4.6.1. Accessibility Clustering

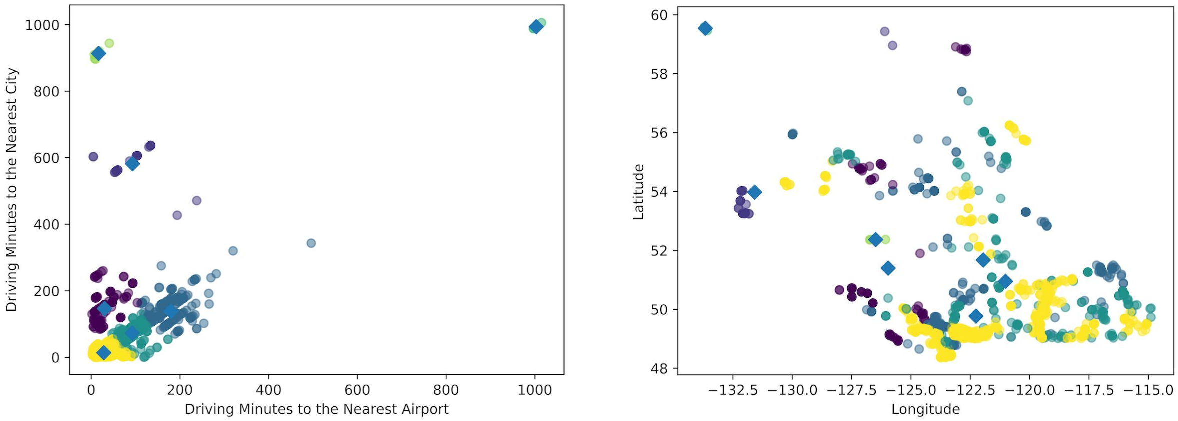


Figure 17: K-Means clustering on accessibility features. (Left top panel) Scatter plot of the features where points are color coded per cluster. (Right top panel) Points plotted on their geographic location (latitude against longitude).

With the assumption that assets with similar travelling time to the nearest city and airport have a comparable level of accessibility, the team performed K-Means clustering on driving duration features. It turned out that the output did not represent tourism assets spatially close to each other as similarly accessible. As a consequence, the team decided to incorporate a spatial component on top of the driving duration feature.

[32] For more information about K-Means clustering and its applications, see: https://en.wikipedia.org/wiki/K-means_clustering

With educated input from the project lead, the team investigated the influence of population on accessibility values and ended up using it as their fifth dimensional variable[33] (so that accessibility can be measured with regard to population, which means that, for instance, an asset located 10 minutes from a densely populated city centre can be associated with higher accessibility, while an asset located less than 10 minutes from a sparsely populated city would then obtain a lower accessibility value). Since population is a large number when compared with driving minutes and geographic coordinates, it had been normalized and a weight factor had been calibrated so that the result looks reasonable. **Figure 17** illustrates the final output of the team's K-Means clustering; specifically, Metro Vancouver (depicted as yellow) is defined as highly accessible to both cities and airports, on the other hand, Whistler and Squamish are considered as moderately accessible, while Pemberton[34] presents the least accessibility. When it comes to tourism sites and islands located alongside the coast, it is interesting to notice their high inaccessibility due to natural remoteness from both cities and airports (Bella Coola is an intriguing case, as it takes the advantage of its airport but requires a long trip to reach Williams Lake, its nearest major city). One drawback of a clustering method like K-Means is its relative sensitivity to outliers, which led the team to consider atypical algorithms such as CLARANS[35], providing a better and more robust representation of the clusters (see **Appendix 10**).

4.6.2. Significance Clustering

Focusing attention on *Significance*, the fellows incorporated the features displayed on the second column of **Figure 16**. Similarly to what had been performed in sub-section 4.6.1., they ran a K-Means clustering algorithm on this particular features' group, aiming for conspicuous and consistent clusters that can be sorted out and ordered accordingly. Employing Principal Components Analysis (PCA)[36], **Figure 18 (top panel)** helps visualizing the way *Significance* features are partitioned into *five* distinct clusters of interest (specifically, after running the "Elbow method"[37], the team found out that *five* clusters was the optimal number of clusters for modeling *Significance* features). Using the geocoordinates attached to each tourism asset, **Figure 18 (bottom panel)** depicts the geographical location and trend associated with each of the *five* clusters, providing first clues on scale matching.

[33] The other 4 dimensional variables are: driving minutes from both nearest cities and airports, as well as latitude and longitude.

[34] To locate Pemberton, see: http://www.hellobc.com/pemberton.aspx

[35] CLARANS stands for "Clustering Large Applications based upon RAndomized Search". For more technical information, see : http://www.cs.ecu.edu/dingq/CSCI6905/readings/CLARANS.pdf

[36] Principal Components Analysis (PCA) was used to create a 2-dimensional picture of the K-Means clustering method, with the aim of detecting and revealing the five optimal clusters that were initially defined. For more information on PCA, see: https://en.wikipedia.org/wiki/Principal_component_analysis and http://setosa.io/ev/principal-component-analysis/

[37] The Elbow method is a popular method in clustering analysis, especially when it comes to determine the optimal number of clusters present in a particular data set, validating the consistency of the clustering method at stake.

Using the categorical scale: “Very Low / Low / Medium / High / Very High”, where “High” means high significance, namely, large number of Instagram counts, TripAdvisor reviews, as well as tourism amenities and attractions located within a 10km radius of an asset of interest; the fellows were able to assign each cluster to each aforementioned scale value. Although this process required further robustness and consistency checks, it allowed the team to specifically identify which tourism assets have very low, medium or high significance, by simply looking at the cluster they belong to. As a consequence, assets with medium, high and very high significance are mainly concentrated around densely populated areas, such as Metro Vancouver, Greater Victoria and Kelowna, while assets with relatively low significance are majoritarily located in the backcountry. It is interesting to note that such observations do not vary if the team accounts for population while performing K-Means clustering.

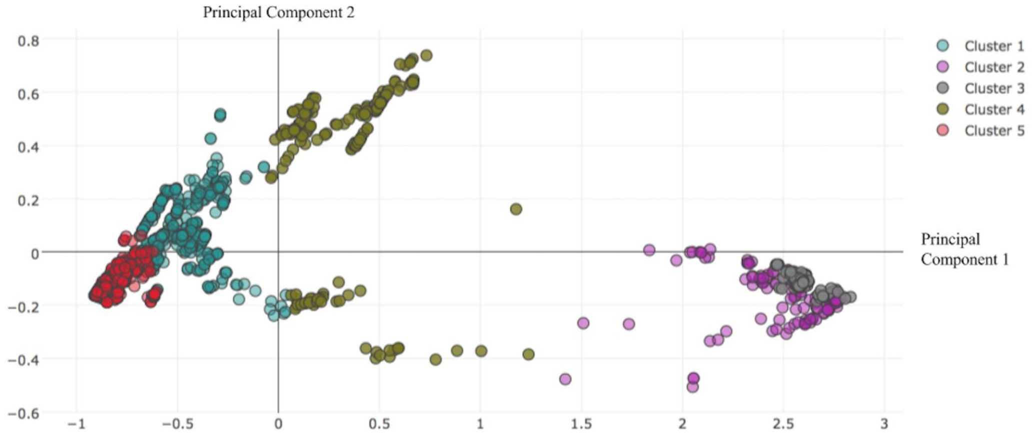
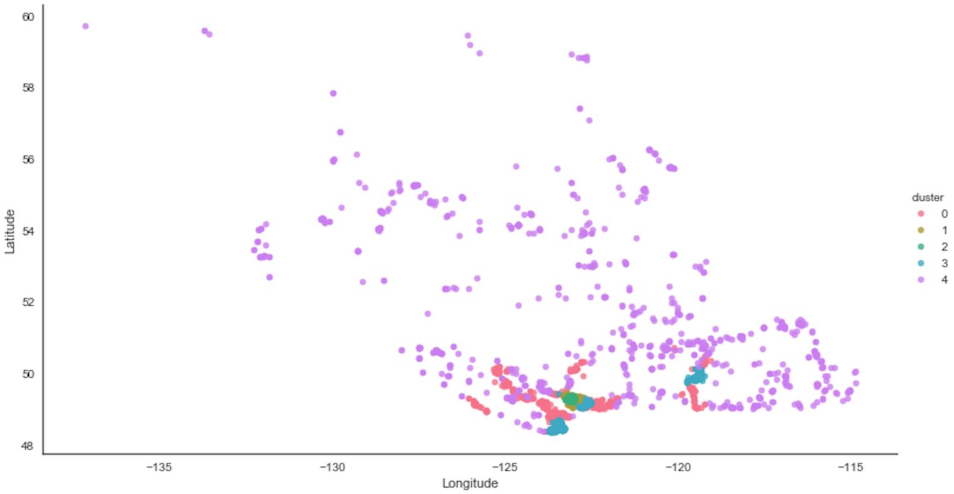


Figure 18: K-means clustering on significance features,
(top panel) Significance clustering using Principal Components Analysis (PCA),
(bottom panel) Significance clustering with latitude against longitude



4.6. Learning to Rank: A Machine Learning Approach

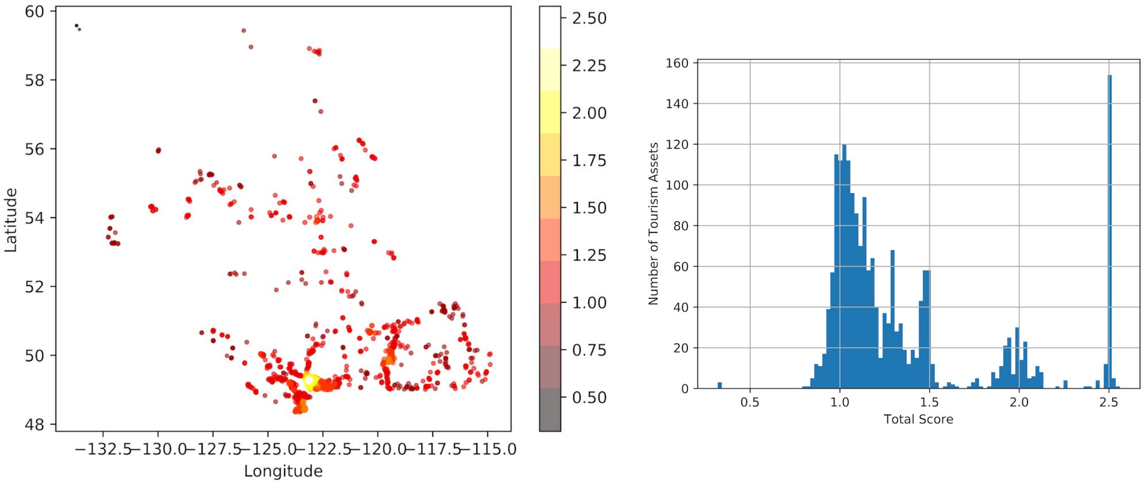
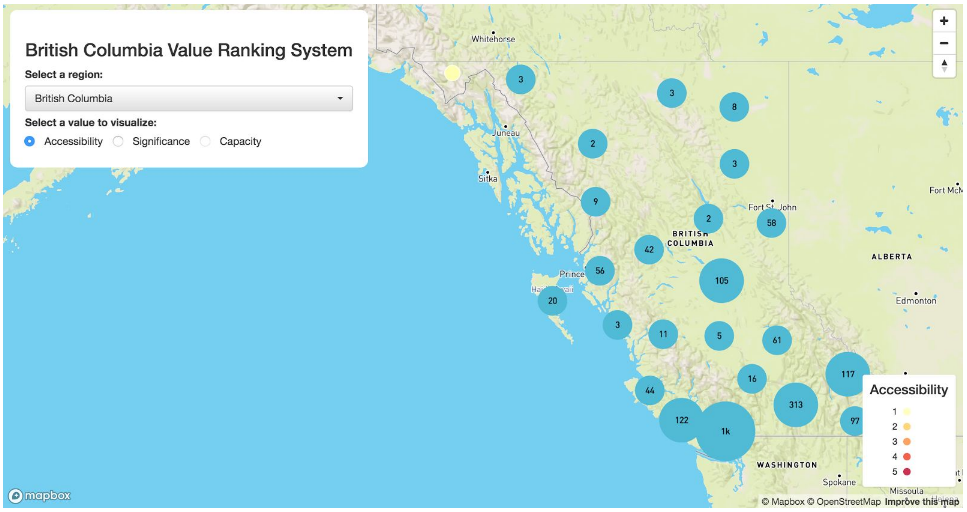


Figure 19: Distribution of equally weighted value score

Prior to using automatic means such as K-Means clustering (see section 4.5.), one of the first attempts to rank tourism assets was to compute equally weighted features in each ranking category as a prototype of relative tourism value. The results showed that Lower Mainland was extensively high, while most assets presented a value between 0.7 and 1.5 out of 3, as shown in **Figure 19**. To design a value ranking system, it is often assumed that realm experts would provide a set of rubric scores, with the derived results coinciding or not with their prior knowledge. In that regard, they would need to perform substantial computations for each tourism asset, including the ones they do not have prior information upon, while ensuring consistency between such values so that they are comparable. In this context, applying a machine learning approach to answer such questions would be recommended. Recent literature shows that it is often assumed that the value follows some kind of spatial pattern, while the degree of "smoothness" and correlation between each observation are controlled by some parameters. One can also specify a rule of how these labels of the known sites propagate to the entire data set. In other words, the program emulates the behaviour of a domain expert via a responsible and intelligent system that projects its expertise into routines. With modern computing and storage power, a system like this can even be designed to support real-time value computations based on massive data sets, ensuring currency of the system, which can then be used for various purposes including, among other things, emergency management and policy-making.

4.7. Web Mapping Tool

As a final product, the team created a web map tool, which displays the results of their data analysis on an interactive map visualization. Specifically, users can zoom-in on a specific tourism planning area and observe the distinctive tourism assets on the map. Each asset is clickable, providing the user with a full view of a set of relevant features for that asset. The user can also choose to visualize one of the ranking categories, be it accessibility or significance. The assets are then color coded using a gradient color scheme to depict the value of the ranking category for the corresponding asset. **Figures 20** and **21** illustrate the web map interface.



British Columbia Value Ranking System

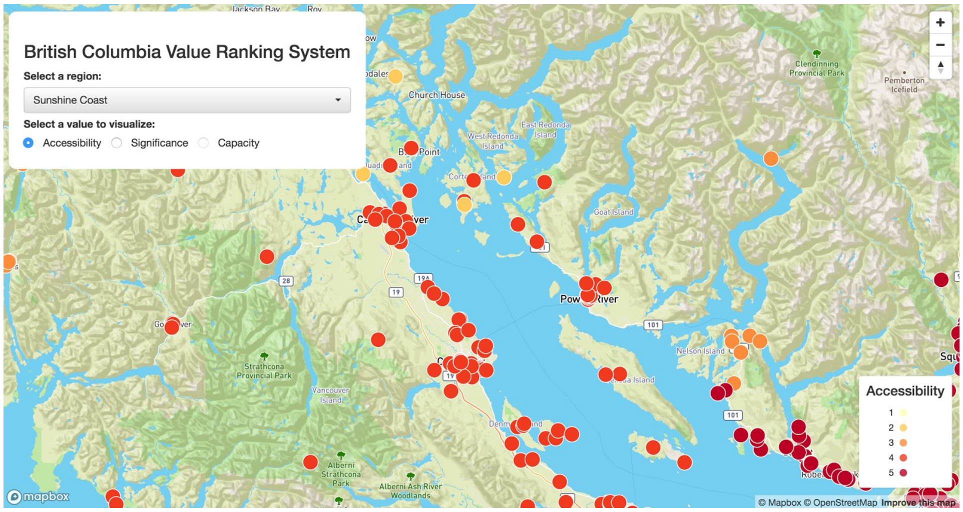
Select a region:

British Columbia

Select a value to visualize:

● Accessibility ○ Significance ○ Capacity

Figure 20: Public web map tool in its default view



British Columbia Value Ranking System

Select a region:

Sunshine Coast

Select a value to visualize:

● Accessibility ○ Significance ○ Capacity

Figure 21: Public web map tool, with zoom-in on the Sunshine coast, visualizing accessibility

5. T.R.I.P. for Social Good

A predominant aspect of the DSSG program lies into its social good component, which provides the fellows with a unique opportunity to reflect on the multifaceted impacts of their data-driven project and associated end product. Balancing out a broad array of stakeholders' needs while conveying a simple but impactful narrative is no easy task; nonetheless, the team believes firmly in the environmental, socio-economic, cultural, financial and reputational benefits that their work could potentially bring to the future of tourism in British Columbia.

5.1. Local Economic Development and Government Channelled Investments

With much of tourism developments happening at the site or asset level, the team's end product will help identifying unique amenities, resources and tourism experiences that a particular location has to offer. Such granularity level, associated with innovative inputs from social media analysis, can detect intriguing "outliers" that the government of British Columbia had not properly identified; for instance, a natural asset such as a waterfall could be located in an area with low tourism development[38] but with substantial social media coverage, indicating a potential opportunity for government channelled investments. As a consequence, local communities living in the vicinity of promising tourism assets could see an increase in the number of yearly visitors, fostering local economic development, via, among other things, job creation, access to new markets and clientele, as well as support to niche businesses.

On the other hand, if the focus is on the demand side, promoting the development of community infrastructures could lead to enhanced quality of visitors' experiences, generating a positive feedback loop that can then be scaled-up at the regional and provincial scales.

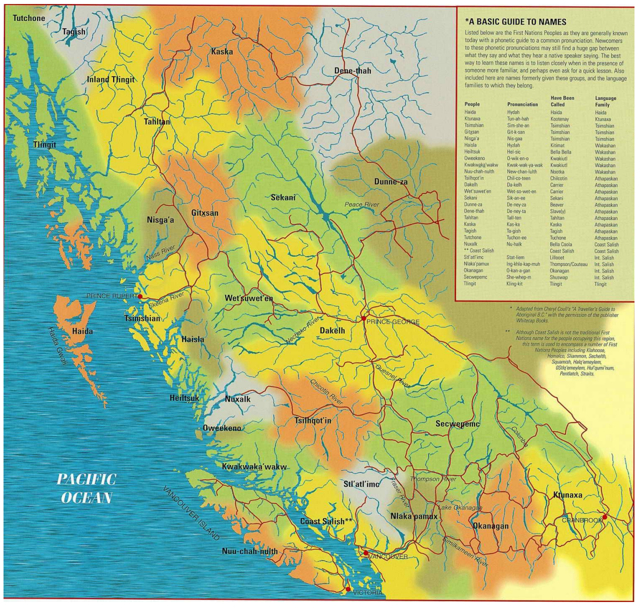
[38] Here low tourism development refers to an area with little infrastructures (roads, local airports, railways…) and facilities (hotels, lodges…). Such assets are both relevant when it comes to the overall quality of visitors' experiences.

5.2. Dampening the Risks and Side-Effects of Identified Tourism Potential

Notwithstanding the list of positive effects that the tourism industry can bring to (insular) local communities, it is important to not lose sight of the varied challenges and potentially negative side-effects that increased tourism activities can generate in both the short- and long-run. Although non-exhaustive, the list of potential risks displayed below can be considered as a starting point for decision-makers and land-use planners keen to integrate the needs of critical stakeholders whose participation would leverage invaluable inter-generational benefits.

5.2.1. First Nations' Native Land Claims

As one of the main cultural concerns of tourism expansion in British Columbia, the respect of First Nations' claimed territories needs to be accounted for while planning new developments across the province. Facilitating a symbiotic relationship between First Nations and tourists could lead to substantial socio-cultural and economic benefits, notably vis-à-vis the numerous **cultural assets** (e.g. heritage sites, Aboriginal museums, First Nations' small businesses…) that are scattered across B.C.[39] Negative spillover effects originating from site development, expansion and enhanced attendance could be mitigated and avoided if a common ground is reached between Aboriginal communities and tourism authorities. First Nations' cultural sites and reserves constitute an intrinsic part of British Columbia's identity; the eclectic distribution of Aboriginal languages and communities (as illustrated by **Figure 22**) ought to play in favour for greater socio-cultural integration vis-à-vis future tourism planning.



***A BASIC GUIDE TO NAMES**

Listed below are the First Nations Peoples as they are generally known today with a phonetic guide to a common pronunciation. Newcomers to these phonetic pronunciations may still find a huge gap between what they say and what they hear a native speaker saying. The best way to learn these names is to listen closely when in the presence of someone more familiar, and perhaps even ask for a quick lesson. Also included here are names formerly given these groups, and the language families to which they belong.

People	Pronunciation	Have Been Called	Language Family
Haida	Hydah	Haida	Haida
Ktunaxa	Tun-ah-hah	Kootenay	Ktunaxa
Tsimshian	Sim-she-an	Tsimshian	Tsimshian
Gitxsan	Git-k-san	Tsimshian	Tsimshian
Nisga'a	Nis-gaa	Tsimshian	Tsimshian
Haisla	Hyzlah	Kitimat	Wakashan
Heiltsuk	Hel-sic	Bella Bella	Wakashan
Oweekeno	O-wik-en-o	Kwakiutl	Wakashan
Kwakwaka'wakw	Kwak-wak-ya-wak	Kwakiutl	Wakashan
Nuu-chah-nulth	New-chan-luth	Nootka	Wakashan
Tsilhqot'in	Chil-co-teen	Chilcotin	Athapaskan
Dakelh	Da-kelh	Carrier	Athapaskan
Wet'suwet'en	Wet-so-wet-en	Carrier	Athapaskan
Sekani	Sik-an-ee	Sekani	Athapaskan
Dunne-za	De-ney-za	Beaver	Athapaskan
Dene-thah	De-ney-ta	Slave(y)	Athapaskan
Tahltan	Tall-ten	Tahltan	Athapaskan
Kaska	Kas-ka	Kaska	Athapaskan
Tagish	Ta-gish	Tagish	Athapaskan
Tutchone	Tuchon-ee	Tuchone	Athapaskan
Nuxalk	Nu-halk	Bella Coola	Coast Salish
** Coast Salish		Coast Salish	Coast Salish
Stl'atl'imc	Stat-liem	Lillooet	Int. Salish
Nlaka'pamux	Ing-khla-kap-muh	Thompson/Couteau	Int. Salish
Okanagan	O-kan-a-gan	Okanagan	Int. Salish
Secwepemc	She-whep-m	Shuswap	Int. Salish
Tlingit	Kling-kit	Tlingit	Tlingit

* *Adapted from Cheryl Coull's "A Traveller's Guide to Aboriginal B.C." with the permission of the publisher Whitecap Books.*

** *Although Coast Salish is not the traditional First Nations name for the people occupying this region, this term is used to encompass a number of First Nations Peoples including Klahoose, Homalco, Sliammon, Sechelth, Squamish, Halq'emeylem, OStlq'emeylem, Hul'qumi'num, Pentlatch, Straits.*

Figure 22: First Nations people (from British Columbia Ministry of Education).

[39] A detailed map of the geographical distribution of First Nations communities and cities, as well as indian reserves in British Columbia is available here: https://www.aadnc-aandc.gc.ca/DAM/DAM-INTER-BC/STAGING/texte-text/inacmp_1100100021016_eng.pdf

5.2.2. “False Positives”

The value ranking analysis performed by the team could also hide some unexpected pitfalls when it comes to evaluate the tourism potential of a particular site or asset. Indeed, although the prediction model has been tested for robustness, it is important to not rule out the possibility of encountering so-called “false positives”; namely, tourism assets that have been associated with high returns on investment, but which “actual” or pragmatic potential does not match the quantitative expectations of the model. Thus, as a precautionary measure, if the local context does not seem to align with the outcome of the statistical model, it would be recommended to ascertain the underlying drivers of the high tourism value attached to a particular site of interest. It would prevent any unwarranted damages and financial losses that could jeopardize the reputation of local, regional and even provincial tourism authorities.

5.2.3. Tourism Carrying Capacity and Ecotourism Opportunities

British Columbia’s global tourism reputation has been partly built on its pristine natural resources and breath-taking sceneries. While attracting ever more international visitors to its provincial parks, hiking trails, lively rivers, snowy mountains and more; it is nonetheless critical to effectively manage the resources at stake for long-term sustainability purposes. Assessing the “sensitivity” of a particular natural asset requires, among other things, to get a sense of its relative tourism carrying capacity; specifically, the maximum amount of visitors that the asset can sustain prior to further ecological degradation. If not carefully monitored, over-crowding can lead to substantial environmental damages and financial burdens during the overturn process. By gathering significant data on tourist attendance all-year long, it would provide land-planning authorities and decision-makers with invaluable ecotourism insights for sound and sustainable policy implementations in the near-future.

Figure 23: “Free Spirit Sphere” or Treehouse, Vancouver Island, British Columbia[40]

[40] See Treehouse map weblink here: http://treehousemap.com/listing/treehotel-canada-free-spirit-spheres/

5.2.4. Government Funding

Identifying promising tourism sites or assets for future investments is an important but not sufficient input prior to implementing local tourism action plans. Indeed, bureaucratic approval and funding processes are an essential and unpredictable part of the overall tourism resources inventory project. While focusing on the supply side of B.C.'s tourism industry; such funding decisions, due to budget restrictions or other external factors, do not always align with the ever increasing tourism demand stemming from domestic, national and international visitors. Accounting for such implementation uncertainties could be helpful in determining operational alternatives to manage the quality and quantity of tourism services that match tourists' expectations while visiting a particular site of interest.

5.3. Spillover Effects on Neighbouring Visitor Markets

While the potential of B.C.'s tourism industry has been recently tapped into by its provincial Destination Marketing Organization (DMO), namely, Destination British Columbia (see section 2.1.); the prospects of garnering further attention from B.C.'s neighbouring jurisdictions is an important reputational and economic aspect that this report's end product could initiate and pass on to the professionals responsible with furthering the DSSG project. Specifically speaking, with the province of Alberta, as well as the States of Washington and Oregon in mind, attracting governments officials' interests to B.C.'s destination development planning, and overall value ranking system for tourism assets could boost B.C.'s international tourism reputation, upholding its current strategy and enhancing future visitors' prospects. If successful, neighbouring jurisdictions and DMOs could emulate particular facets of the Tourism Resources Inventory Project, indicating potential value in extrapolating further the tourism know-how of British Columbia.

5.4. Data-Science Driven Decision-Making

Last but not least, Data Science for Social Good showed collaboration promises between academics and government officials. As a matter of fact, via productive, insightful and appreciative weekly meetings between the DSSG team members and their project lead, the fellowship became an experimental setting for interdisciplinary reflections and innovative ideas. Boasting the merits of data-driven decision-making, the team firmly believes that bridging such a long-lasted communication gap can lead to more effective and insightful policy implementation.

6. Conclusion & Recommendations

Methodologies arising from Data Science can be a new way of tackling tourism research challenges if enough attention has already been addressed to traditional GIS approaches. Data analysis contains a clear timeline: problem statement, data collection, data pre-processing, exploratory data analysis, modeling, model tuning, scoring, prediction and presentation. The following sections will list a few major recommendations that the fellows deem relevant to address.

6.1. Joint Communication on the Objective

Tourism stakeholders ought to continuously communicate between each other if they want to anticipate and cope with any challenges and pitfalls along the way. Enhancing interactions between stakeholders can leverage unvaluable economic, environmental, political and cultural benefits for the province of British Columbia. In addition, with novel data approaches and ever increasing information stemming from the realm of data science, the current project's analysis ought to incentivize stakeholders to reevaluate their "initial objective" (or problem statement) and better tailor it to the goals of data-driven decision making.

6.2. Data Collection - Reduce the Data Gap

Section 3.5 offers a table of suggested data collection for tourism and data professionals responsible with furthering the DSSG project. This section will reason on how to have a better data gathering process.

Statistical Analysis

To address the relative value of each tourism asset, it is critical to acknowledge the need for a proper "proxy" or response variable. Throughout this report, the team has underlined the suitability of measuring and gathering visitors count per tourism asset at a given time. Statistical analysis allows a one-variable comparison of the value among distinct tourism assets, and although the "overall" value is multifaceted, the one-variable result provides higher granularity.

Time Series

Current data sets lack the ability to conduct time series analysis, which is crucial for comprehending the dynamic changes of tourism-related data over a certain period of time.

Assessing the impact of a new tourism facility or targeted policy through time would provide unvaluable information to decision-makers and land-use planners.

Demand-Side Analysis/Forecast

Including the demand-side of tourism within the team's data analysis might prove worth the effort when it comes to better understanding future policy implications. This includes, among other things, customer satisfaction and customer reviews of a particular tourism asset. In that regard, the fellows conducted an initial Natural Language Processing (NLP) analysis on the Instagram posts, with the aim of probing into the diversity of visitors' experiences. The NLP results were highly positive due to the scenery and landscapes bias of the Instagram data. Future studies should include a more detailed sentiment analysis, which would cover more than one month of data, thereby increasing the overall sample of Instagram captions while accounting for seasonality. Other potential data sources include TripAdvisor reviews as well as specific customer surveys targeting a certain tourism asset. The demand-side analysis can also include business perspectives such as the requests for new businesses in a certain area.

Agent-Based Studies/Simulation

This requires a definition of a 'Trip'. Using such definition, it is then possible to simulate particular agents' behaviors, for instance, booking a hotel, airline, shopping, outdoor decisions et cetera. The analysis on visitors' "common routes" would allow a greater understanding of the potential associated with circuit programs. In addition, a cost-benefit analysis with some proxy variable for leisure against total cost can also help comprehend why and how people choose to travel.

Capacity Features

Capacity has been marked as a crucial aspect of each tourism asset's multifaceted value, but due to some natural measurement intricacies, this ranking category lacked features of critical importance. For instance, questions such as how to measure the carrying capacity of a park, the vulnerability of land-use, and the capability of physical capital remain to be answered. At this point, the team would recommend greater collaborations and partnerships with appropriate experts in order to collect insightful knowledge on the aforementioned queries.

6.3. Modeling: Unsupervised Learning

Amongst the wide range of available and ever evolving data science methodologies and algorithms, unsupervised learning turned out to be best suited for this particular DSSG project. Unsupervised learning refers to clustering or classification. Without proper knowledge of the proxy variable or weights associated with each feature of interest, machine learning can cluster groups of points based on their similarities, while the output can be compared with professional insights to observe the difference and draw meaningful conclusions.

6.4. Future Research Directions

Suitable Proxy Variable

Considering the current features collection as well as future analysis promises, it would be possible to perform statistical analysis on different variables to test whether certain variables perform better than others when it comes to fill the role of a proxy variable. As an example, focusing on *significance*, there could be a linear regression analysis being performed on all of its features, using Instagram counts as the dependent or response variable. A similar procedure can be conducted with regard to TripAdvisor reviews this time, before comparing which variable stands better as a proxy for the relative value of tourism.

Further Exploratory Data Analysis regarding Each Cluster

Additional research could also target each cluster and its associated categorical scale, in an attempt to explain why they present particularly low or high values. Such an analysis could further support decision making and local tourism developments, among other things.

Distinct Focus and Additional Value's Facets

The current data sets only allowed the team to conduct initial analysis on *significance* and *accessibility*. With additional data collection, they hope that vulnerability of land-use and tourism carrying capacity will progressively complement their original project, while being open to newly devised ranking categories as well. Finally, incorporating tourism facilities and infrastructures will certainly benefit the current prototype, and initiate new research on interlinkages between the distinct categories of tourism-related data.

7. Appendix

Appendix 1: Web Link to the Project’s GitLab Repository

https://gitlab.math.ubc.ca/halldorb/bc_tourism

Appendix 2: BC Tourism Stakeholder List

Provincial Government

- Ministries
- Crown Corporations

Local Governments

- Union of BC Municipalities
- Regional Districts
- Municipalities
- Resort Municipalities
- Islands Trust

First Nations

- Aboriginal Tourism Association of BC
- Union of BC Indian Chiefs
- BC Assembly of First Nations
- First Nations Summit
- Sector-specific councils
    - Forestry
    - Mining
    - Fisheries
- Individual First Nations Communities

Destination Marketing Organizations

- Destination Canada
- Destination British Columbia
- Regional & Community DMOs
    - Cariboo Chilcotin Coast Tourism Association
    - Kootenay Rockies Tourism Association
    - Thompson Okanagan Tourism Association
    - Vancouver Coast Mountains Tourism Association
    - Tourism Vancouver Island

- Visitor Centre Network (113 Visitor Centres in BC)

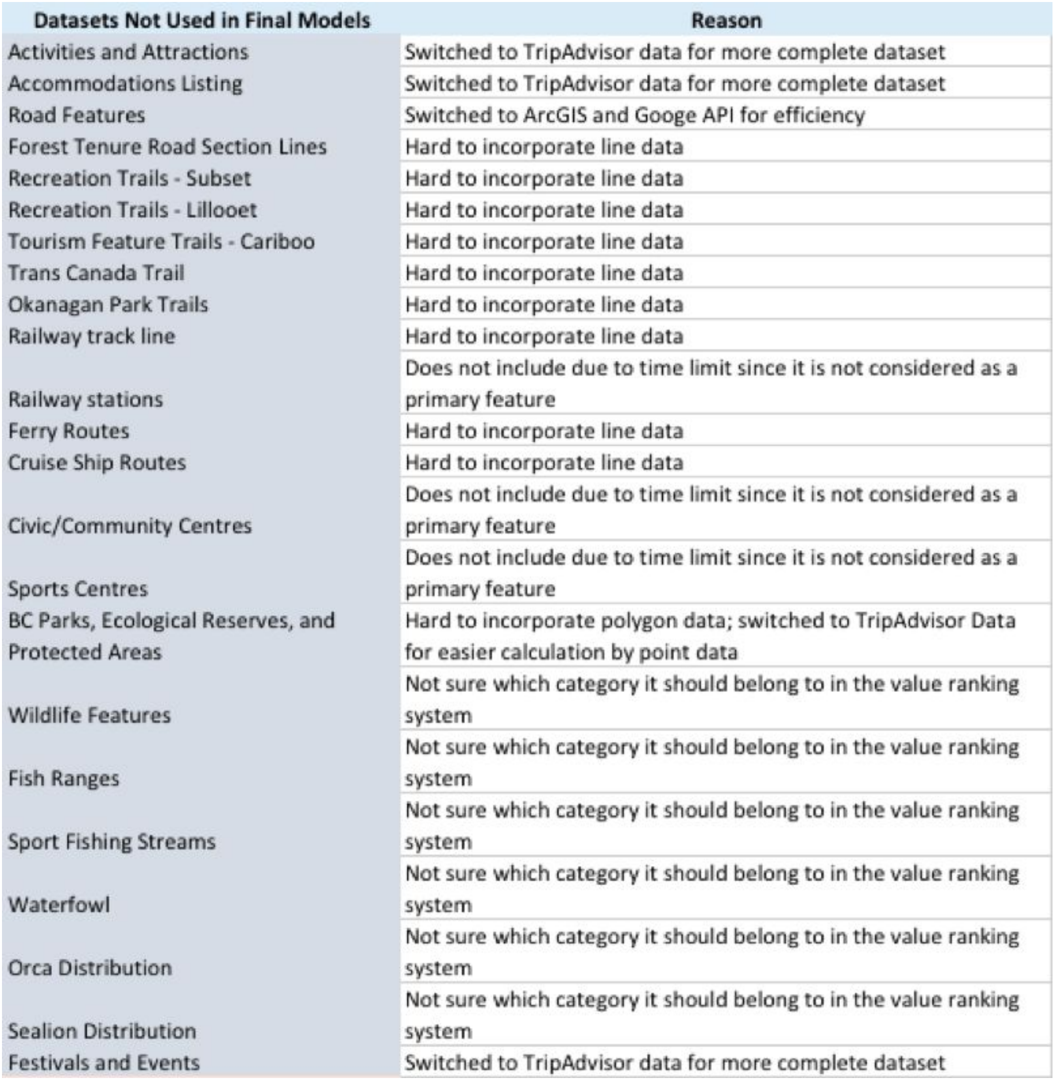
- Transportation Authorities
- Vancouver Airport Authority
- Victoria Airport Authority
- Regional Airports
- Greater Victoria Harbour Authority
- Vancouver Fraser Port Authority
- Nanaimo Port Authority

Industry Associations

- Tourism Industry Association of BC
- Tourism Industry Association of Canada
- Commercial Bear Viewing Association
- Guide Outfitters Association of BC
- Wilderness Tourism Association
- Canada West Ski Areas Association
- Heli-Cat Canada
- Association of Canadian Mountain Guides
- BC Commercial Snowmobile Operators Association
- Backcountry Lodges of BC
- Canadian Ski Guides Association
- Association of Canadian Mountain Guides
- BC Fishing Resorts and Outfitters Association
- BC Ocean Boating Tourism Association
- BC River Outfitters Association
- Boating BC
- Mountain Biking BC
- Sport Fishing Institute of BC
- Sea Kayak Guides Alliance of BC
- BC Hotel Association
- go2HR – Tourism workforce and labour association
- BC Restaurant and Foodservices Association
- BC Lodging and Campground Association
- North West & Canada Cruise Association
- Cruise Canada (RV rentals)
- RV Rental Association of Canada

Appendix 3: Reasons for not Using Certain Data Sets from the Project Lead's list

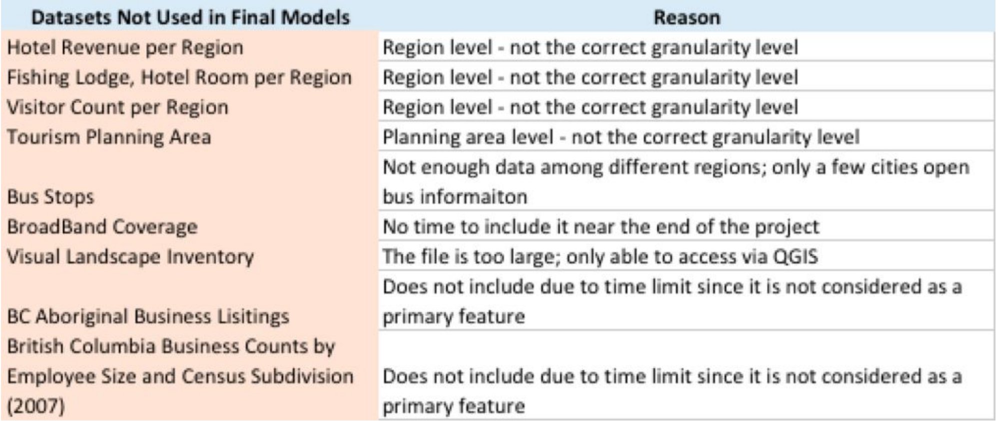
Note: In the final modeling process



Datasets Not Used in Final Models	Reason
Activities and Attractions	Switched to TripAdvisor data for more complete dataset
Accommodations Listing	Switched to TripAdvisor data for more complete dataset
Road Features	Switched to ArcGIS and Googe API for efficiency
Forest Tenure Road Section Lines	Hard to incorporate line data
Recreation Trails - Subset	Hard to incorporate line data
Recreation Trails - Lillooet	Hard to incorporate line data
Tourism Feature Trails - Cariboo	Hard to incorporate line data
Trans Canada Trail	Hard to incorporate line data
Okanagan Park Trails	Hard to incorporate line data
Railway track line	Hard to incorporate line data
Railway stations	Does not include due to time limit since it is not considered as a primary feature
Ferry Routes	Hard to incorporate line data
Cruise Ship Routes	Hard to incorporate line data
Civic/Community Centres	Does not include due to time limit since it is not considered as a primary feature
Sports Centres	Does not include due to time limit since it is not considered as a primary feature
BC Parks, Ecological Reserves, and Protected Areas	Hard to incorporate polygon data; switched to TripAdvisor Data for easier calculation by point data
Wildlife Features	Not sure which category it should belong to in the value ranking system
Fish Ranges	Not sure which category it should belong to in the value ranking system
Sport Fishing Streams	Not sure which category it should belong to in the value ranking system
Waterfowl	Not sure which category it should belong to in the value ranking system
Orca Distribution	Not sure which category it should belong to in the value ranking system
Sealion Distribution	Not sure which category it should belong to in the value ranking system
Festivals and Events	Switched to TripAdvisor data for more complete dataset

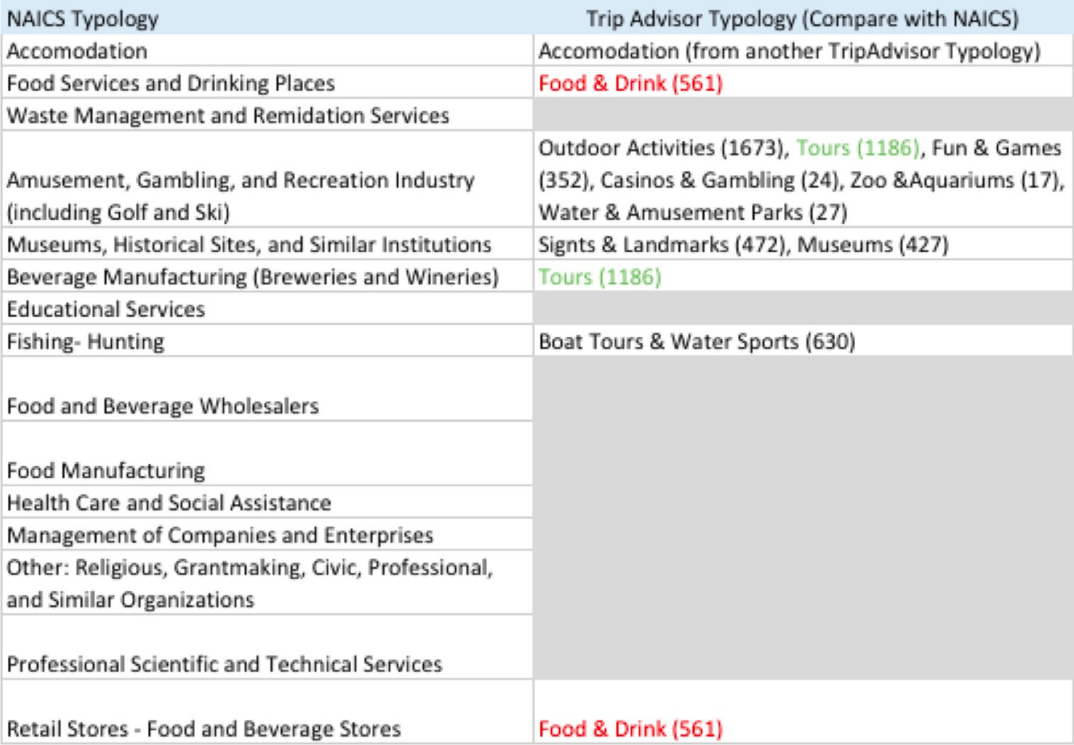
Appendix 4: Reasons for not Using Certain Data Sets from Open Data Sources

Note: In the final modeling process



Datasets Not Used in Final Models	Reason
Hotel Revenue per Region	Region level - not the correct granularity level
Fishing Lodge, Hotel Room per Region	Region level - not the correct granularity level
Visitor Count per Region	Region level - not the correct granularity level
Tourism Planning Area	Planning area level - not the correct granularity level
Bus Stops	Not enough data among different regions; only a few cities open bus informaiton
BroadBand Coverage	No time to include it near the end of the project
Visual Landscape Inventory	The file is too large; only able to access via QGIS
BC Aboriginal Business Lisitings	Does not include due to time limit since it is not considered as a primary feature
British Columbia Business Counts by Employee Size and Census Subdivision (2007)	Does not include due to time limit since it is not considered as a primary feature

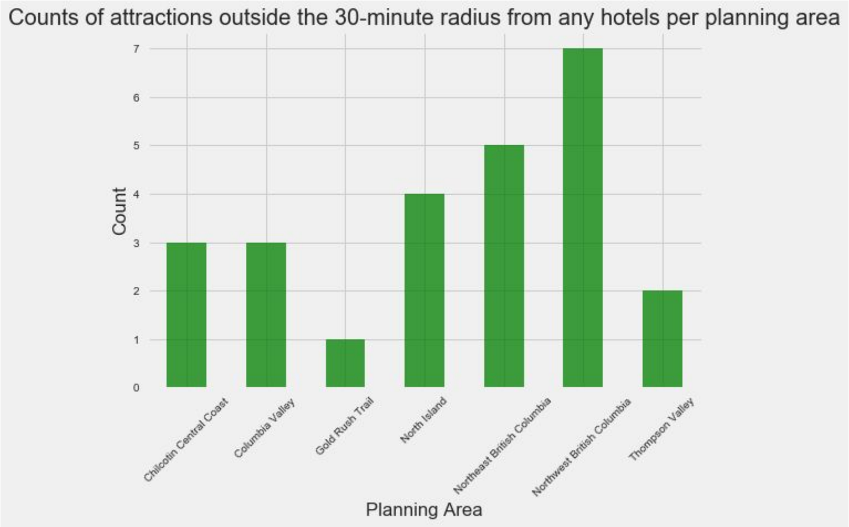
Appendix 5: Typology Comparison between TripAdvisor and NAICS



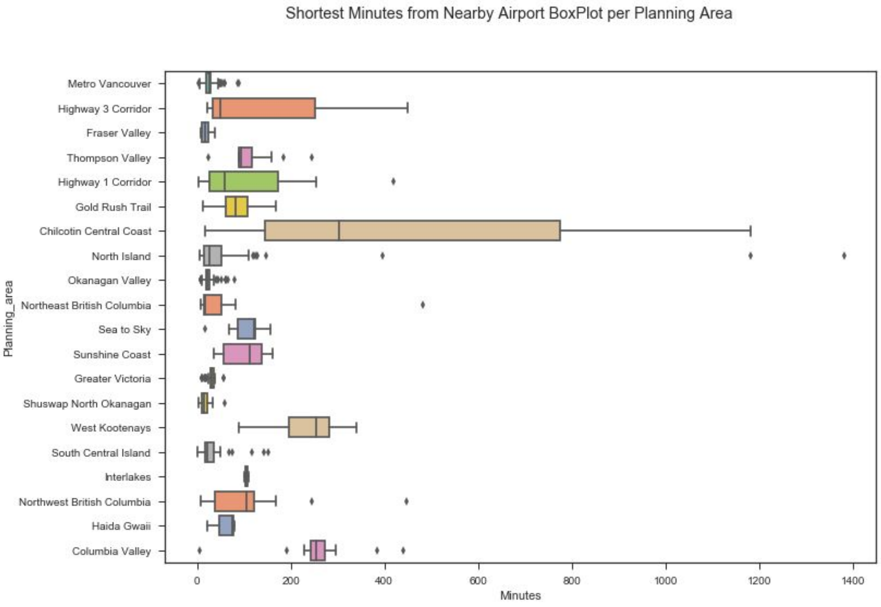
NAICS Typology	Trip Advisor Typology (Compare with NAICS)
Accomodation	Accomodation (from another TripAdvisor Typology)
Food Services and Drinking Places	Food & Drink (561)
Waste Management and Remidation Services	
Amusement, Gambling, and Recreation Industry (including Golf and Ski)	Outdoor Activities (1673), Tours (1186), Fun & Games (352), Casinos & Gambling (24), Zoo &Aquariums (17), Water & Amusement Parks (27)
Museums, Historical Sites, and Similar Institutions	Signts & Landmarks (472), Museums (427)
Beverage Manufacturing (Breweries and Wineries)	Tours (1186)
Educational Services	
Fishing- Hunting	Boat Tours & Water Sports (630)
Food and Beverage Wholesalers	
Food Manufacturing	
Health Care and Social Assistance	
Management of Companies and Enterprises	
Other: Religious, Grantmaking, Civic, Professional, and Similar Organizations	
Professional Scientific and Technical Services	
Retail Stores - Food and Beverage Stores	Food & Drink (561)

Appendix 6: Attractions within 30-minutes Radius from any Hotel

Note: only show the planning area with existed target

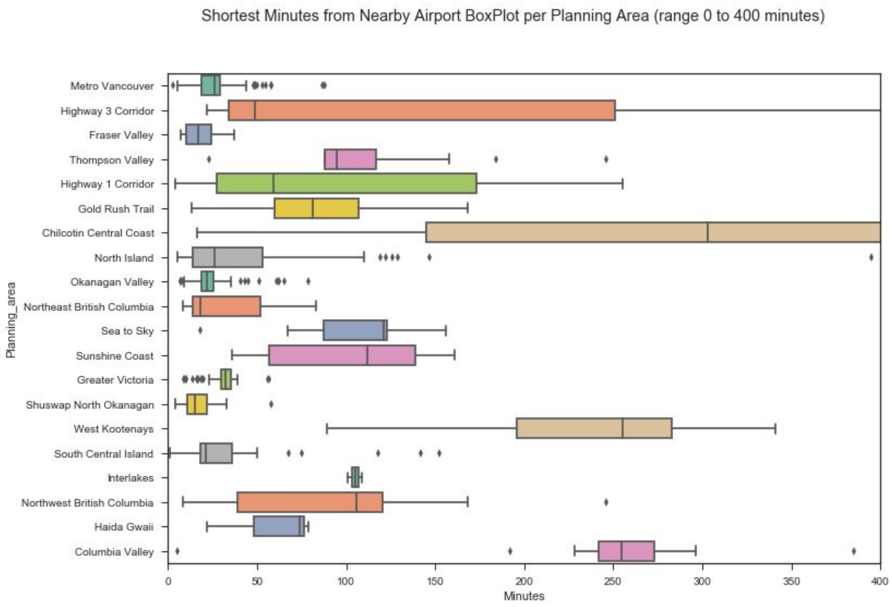


Appendix 7: Shortest Time (min) from Nearby Airport, Boxplot per Planning Area



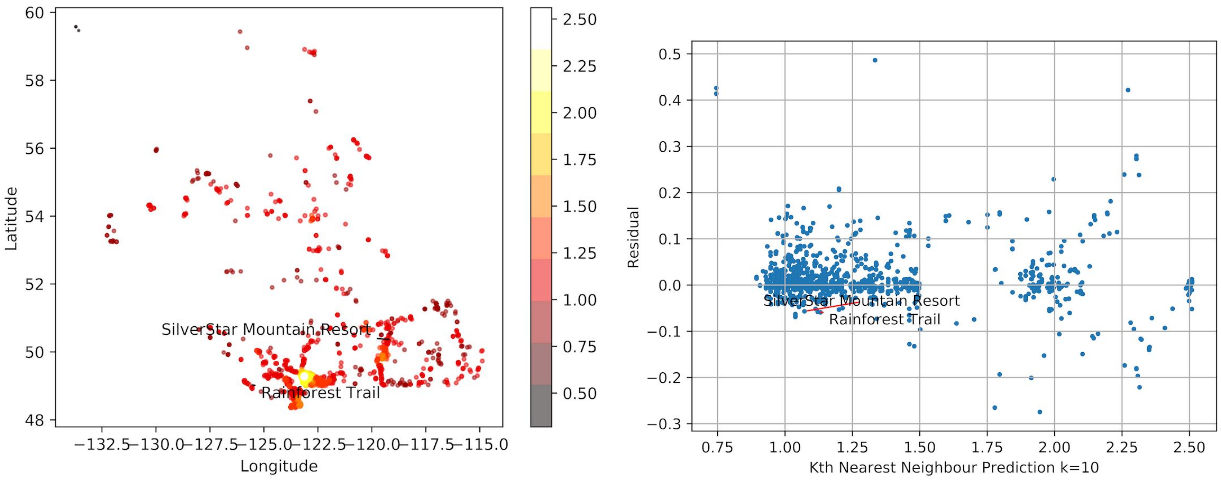
Appendix 8: Shortest Time (min) from Nearby Airport, Boxplot per Planning Area

Note: zoomed in



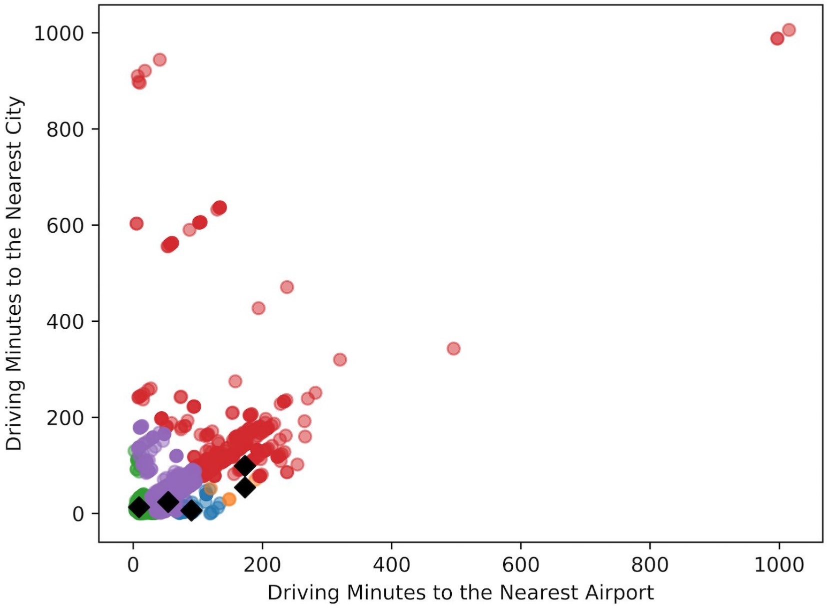
Appendix 9: Anomaly Detection

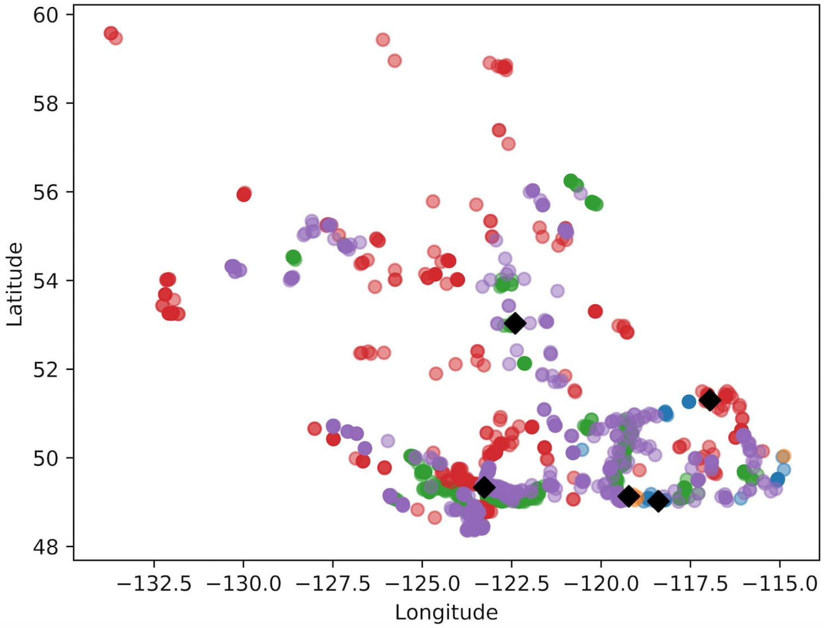
Note: More research can be done on why there are low value assets surrounded by high value ones



Appendix 10: CLARANS Algorithm on Accessibility Features

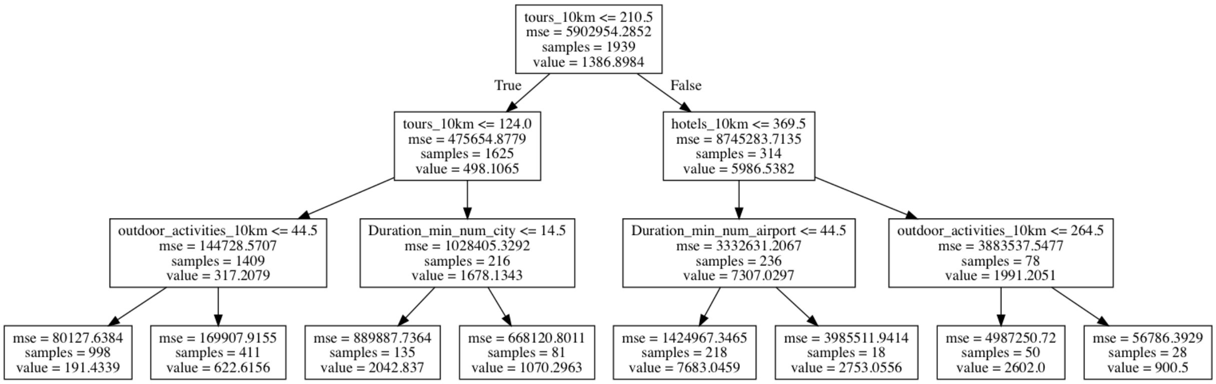
Note: (top panel) Scatter plot of the features where points are color coded per cluster, (bottom panel) Points plotted on their geographic location (latitude against longitude)





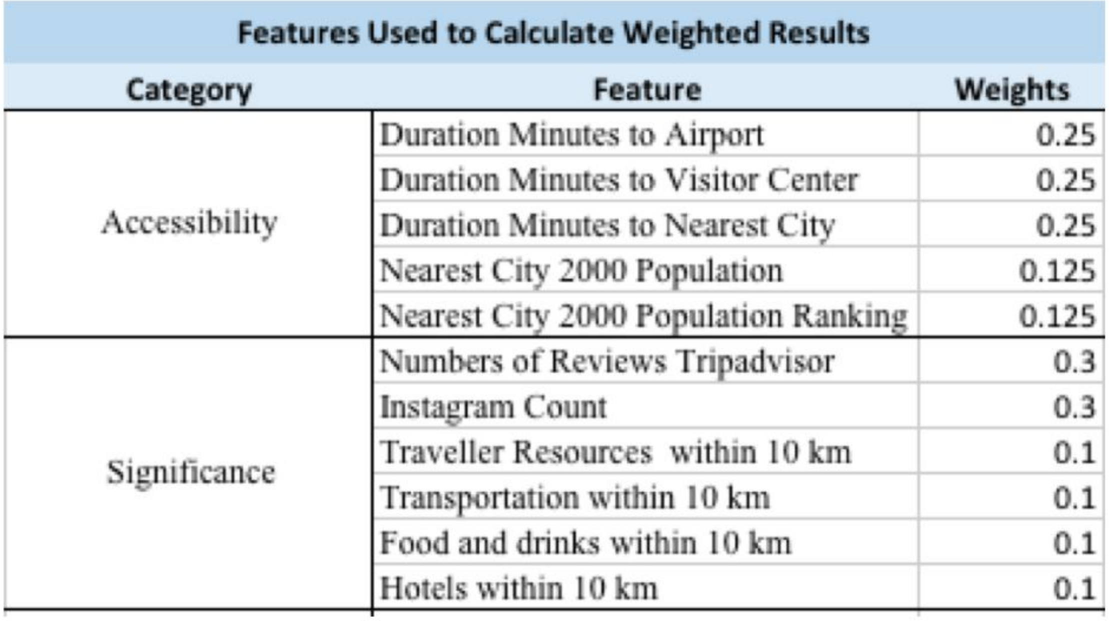
Appendix 11: An Instance of Decision Tree

Note: Decision trees can be used to emulate tourist's decision for planning a trip.



Appendix 12: Features Used to Compute Weighted Results

Note: Weights were given by the project lead



Features Used to Calculate Weighted Results

Category	Feature	Weights
Accessibility	Duration Minutes to Airport	0.25
	Duration Minutes to Visitor Center	0.25
	Duration Minutes to Nearest City	0.25
	Nearest City 2000 Population	0.125
	Nearest City 2000 Population Ranking	0.125
Significance	Numbers of Reviews Tripadvisor	0.3
	Instagram Count	0.3
	Traveller Resources within 10 km	0.1
	Transportation within 10 km	0.1
	Food and drinks within 10 km	0.1
	Hotels within 10 km	0.1

Appendix 13: Value Ranking System for British Columbia - Pipeline User Manual

Problem:

- To estimate the relative value of tourism assets in British Columbia:
    - Assets are separated into three data categories or silos: *Infrastructure*, *Facilities*, as well as *Natural and Cultural Resources.*
- To deliver a multifaceted value ranking system:
    - Brainstorming the facets of the final tourism value based on the goal and context of the value ranking system
    - A good choice had been to separate the value into three distinct ranking categories: *Significance, Accessibility* and *Capacity/Sensitivity.*

Data gathering:

- To acquire a comprehensive list of tourism assets from the internal government and external sources (the latter one could be private or public)
    - Each tourism asset includes spatial reference information (i.e. geo-coordinates, polygon coordinates, census subdivision, tourism region, ...). Assets without spatial reference information have limited utility but could potentially be used as proxy measures.
- To analyze and categorize data sets based on their relevance to the different ranking categories determined earlier (e.g. compiling airport location and bus route data for *Accessibility*)
    - Here is a non-exhaustive list of proper data elements serving as input for the value ranking system: socio-economic indicators (demographics and economic activity), social media (amount of activity, sentiment, reach, ratings), transit (distance and time to travel to nearby tourism assets), business (number of businesses in region by sector, economic output, employment), tourism performance indicators (airport arrivals, hotel occupancy, visitor count, net promoter score).

Data cleansing:

- To clean the data:
    - Ensuring all the missing values will be simulated accordingly (data pre-processing for machine learning)
    - Eliminating features that present undue missing values.

Feature engineering: (*see **Appendix 14** for a non-exhaustive set of computed features)

- To build and compute features relevant to the ranking categories that are being assessed (i.e. significance,accessibility and capacity/sensitivity):
    - Data flattening (per asset, census subdivision, regional district, tourism regions),
        - Collapsing data on different granularity levels to the scale of the tourism asset (1:1),
        - Producing a method or proxy measure to match data at different levels when data cannot be matched to the tourism asset scale. For instance, socio-economic data

is only available at the census-subdivision level, while another challenge lies in the transboundary character of parks (identified as polygons). Manual solutions are needed in some cases.

    - Computing the features via Google API and Geographic Information System (GIS),
        - Many geographical features can be calculated using GIS,
        - Google API can be used to compute the travel distance between two points in space (e.g. between a tourism asset such as a waterfall and the nearest airport),
        - Exploratory data analysis (EDA) can be applied to compute feature while accounting for major and unexpected outliers.
- To aggregate the selected features using automatic (and/or) manual methods, with the aim of obtaining a single score/value for each ranking category of interest:
    - Automatic: Using clustering methods to identify and attach a numerical and/or categorical scale vis-à-vis each ranking category,
        - Accounting for Metro Vancouver as a strong outlier, which can affect model accuracy,
            - Possible remedies: to exclude Metro Vancouver from the clustering analysis, or to reduce the impact of outliers by normalizing the data with regard to population or number of assets for instance.
        - Features selection from principle component analysis and recursive feature elimination,
        - Identifying the optimal number of clusters via inertia analysis,
        - Running K-Means clustering and/or Clustering Large Applications based on RAndomized Search (CLARANS),
        - Comparing clustering outcomes between K-Means and CLARANS, using statistical and exploratory data analysis,
        - Visually inspecting the clusters to identify and attach them to a particular categorical scale such as High/Medium/Low (or another scale) for each ranking category being included in the multifaceted value ranking.
    - Manual: Computing weighted sum with manual weights
        - Exploratory data analysis on the computed weighted sum, while checking and accounting for any major and unexpected outliers.
    - Semi-Automatic: Using artificial intelligence to learn assigning value to tourism assets
        - Manually input weights to calculate a total score for each asset.
        - Assign label to those assets according to total scores and user's expertise.
        - Split labeled data into training and testing, and train classifier/regressor.
        - Use the classifier/regressor to infer the unlabeled data.

Output:

- A layer including a score within every ranking category and attached to each tourism asset, in the form of shapefiles (or other flexible spatial data formats),
    - Potential input into GIS (and ArcGIS) visualization tool for further analysis or as a decision-making pipeline for future tourism development and land-use planning.

Appendix 14: Numerical Data Cleansing Documentation

Selected variables (or columns) of the final data spreadsheet:

Socio-economic Data: Total_population_CSD, Workforce_CSD, Employed_CSD, Unemployed_CSD, Occupations_art_culture_CSD, Occupations_natural_resources_CSD, Median_income_2005_CSD, Average_income_2005_CSD

Distance and Duration Data: Duration_min_num_airport, Distance_meter_airport, Duration_min_num_vc, Distance_meter_vc, Duration_min_num_city, Distance_meter_city, City_Pop_2000, City_Pop_Rank, Duration_min_num_ambulance, Distance_meter_ambulance, Duration_min_num_fire, Distance_meter_fire, Duration_min_num_police, Distance_meter_police

*(*Legend*: CSD* = Census Subdivision*, CD* = Census Divisions*, min* = minutes*, num* = number or numerical value*, vc* = visitor centre*)*

Methodology:

Data cleansing on Total_population_CSD (i.e. total population per subdivision)*:*
1. Identify the CSDs that do not have any population and socio-economic data (e.g. employment, workforce…) available for analysis, here is a way to simulate those "null" or missing values:
2. Get the total population on division-level (CD) from the latest census,
3. Sum the existed current population of all subdivisions within each division of the null data set,
4. Use total CD population - sum of CSD population to identify what is the remaining population to distribute among all null subdivisions; equally distribute the remaining population among all different null CSD in each CD,
5. If sum CSD population > total CD population, apply the mean of the current population to all different null CSD

Distance and Duration Data (how to cope with "null" or missing values):
1. Identify the Census Division (CD) attached to the row of interest (in final data spreadsheet),
2. Find the CSD within CD that already has the wanted duration value (e.g. airport duration),
3. Get the maximum duration value + 1 within CD for all null duration,
4. Update the distance as well with (max duration row's distance/maximum duration value) * row's new duration value.

Notes: *"maximum duration value + 1" is used* to emphasize that the null value simulation is even further than all the current known data points' duration.

Appendix 15: Research Team Introduction

Name: Raphaël Roman

Email: raphael.roman@alumni.ubc.ca

LinkedIn: https://www.linkedin.com/in/raphael-roman-2a5517122/

Education: Master of Public Policy, B.Sc. in Economics

Research Interests: Environmental and Ecological Economics, Ocean Stewardship, Sustainability Education and Data Science.

Name: Halldor Thorhallsson

Email: halldorb@alumni.ubc.ca

LinkedIn: https://www.linkedin.com/in/halldorbjarni/

Education: M.Sc. Student in Computer Science

Research Interests: Big data, Machine Learning, Efficient Big Data Pipelines, Data Science and Data Visualizations.

Name: Hailey Wu

Email: qianqianwu577@yahoo.com

LinkedIn: https://www.linkedin.com/in/qianqianhaileywu/

Education: Undergraduate in Business and Computer Science Combined

Research Interests: Machine Learning, Data Visualization, Big Data, Business Intelligence, Finance Quantitative Analysis.

Name: Gary Zhu

Email: garyzhubc@gmail.com

LinkedIn: https://www.linkedin.com/in/garyzhubc/

Education: Undergraduate in Combined Major in Economics and Statistics, Honours in Mathematics with a focus on Computer Science

Research Interests: Machine Learning, High Dimensional Inference, Scientific Computing, Economics & Finance.