Investment & Intergovernmental Project Final Report

visualicity
VISUALIZING CITIES WITH SIMPLICITY

The University of British Columbia
Data Science for Social Good Fellowship 2017

The Fellows:

Rashedul (Rashed) Hoque

Tony Hui

Natasha Mattson

Sarah Neubauer

Table of Contents

0. SUMMARY

This report discusses the Investment and Intergovernmental project for the University of British Columbia's (UBC) 2017 Data Science for Social Good (DSSG) fellowship program. The purpose of this project was to create an economic profile of Surrey for the Economic Development Division at the City of Surrey, describing different topical features which have the potential to affect the economic health of Surrey as a whole and to find out what regions within Surrey (on a census tract level) are distinctive. The final product includes:

- A data matrix on a high granular geographic level (census tracts) and to visualize the data on an interactive browser called Visual Surrey;
- A classification of census tracts in Surrey using machine learning algorithms; and
- A Temporal Analysis to examine the determinants of business growth.

Each observation in the data matrix contains economic information about a particular census tract in Surrey. The data matrix contains business license, crime, demographic, job postings and commercial rental listings data compiled from various open data sources as well as private data provided to us by the City of Surrey. The data is then visualized using an interactive online mapping tool called Visual Surrey[1].

Using machine learning algorithms, we were able to classify census tracts based on high dimensional similarities. We were able to classify ninety-six census tracts into five main categories and two outlier categories. Once we had these classifications we could look at what variables were key for differentiating groups from one another. Finally, we could describe the clusters based on these key variables. In particular, we found that household income and education levels, along with commercial and home businesses were key for differentiating census tracts.

Temporal analysis was conducted to show how temporal analysis could be used. The results indicated that the growth of home based businesses and construction businesses are temporally related with the median gross land value per square feet of private properties. Only these two business types demonstrated such a relationship. Moving forward it is suggested to find a good proxy of the business growth other than the number of businesses and use more explanatory variables related to workers, transit and inhabitants.

[1] Visual Surrey is available here: bit.ly/visualsurrey

1. INTRODUCTION

The project began with the idea that the City of Surrey is a “City of Cities”. The municipal government of Surrey wanted to understand what made each of these individual cities, known as town centres, unique. This information would be useful for attracting investment, growing Surrey’s innovation economy, creating jobs and increasing Surrey’s competitive advantage as a work force. With consideration into what Surrey wanted, the data available and what would be feasible in a 14 week time frame, we decided to create Visual Surrey, an interactive economic profile of Surrey.

The City of Surrey is currently in the process of developing an interactive tool known as Site-Selector, which enables potential investors to view available commercial rentals, and see various statistics regarding this plot of land. Average age groups, types of households and incomes and more that are in the general area of this rental space. The problem with this tool is that in order to use it, one must already know and understand Surrey. Hence the primary goal for Visual Surrey was to create a searchable, browseable economic and social profile of Surrey that someone who knew nothing about Surrey could use.

We created an economic profile of Surrey by grouping together similar census tracts and then describing these groups based on their differences. Instead of describing the ninety-six census tracts as a whole, we could say whether an individual census tract was in a high income group, for example. These results are displayed on the Visual Surrey platform. The City of Surrey can visualize similar studies in the future. This ties in the project’s original goal of quantifying what makes each Town Centre distinct. By making this economic profile available, it documents knowledge that is implicitly known by the City of Surrey staff, making it easier for new staff to transition into their government role and understanding Surrey as a whole. In addition, Visual Surrey can provide new information and insights for existing City of Surrey staff.

2. METHODS

2.1 Summary and Visualization

Exploratory analysis was used to examine the characteristics and patterns of the compiled data matrix, using basic plottings like boxplots and histograms. After that, some of these summary statistics were used as new variables for in the data matrix, which could then be visualized using Visual Surrey. It was also important to conduct this exploratory analysis before incorporating variables into later analysis, since some data might have reporting errors. For example many business reported having 1 employee, which looked quite shocking in the exploratory analysis. The City of Surrey explained this is because data the data was typically collected at the time of issuing the license, hence many businesses would report having one employee. Further considerations had to be made when incorporating this data into the final data matrix, or to use with further analysis

2.2 Principal Component Analysis and Clustering

In order to classify and group together similar census tracts in the final data matrix *X*, Principal Component Analysis (PCA) and Clustering was used. Each row in *X* represents a census tract and each column represents a demographic or business variable. The data was first standardized across columns to get matrix *X'*. This was done because the variables measured in different units including percentage, dollar values, years, and business counts, and hence are not directly comparable to one another.

Clustering algorithms, like hierarchical clustering or k-means, groups together observations based on similarity, defined by some distance metric. However, running a clustering algorithm on *X'* would overemphasize the importance of correlated variables. *X'* contains variables that are correlated in two ways. The first is known,repeats in the measurement. For example, we have several measures of income, including the percentage of people under a given income bracket.The number of times a variable appears in our data is arbitrary. In addition, there might be variables that are correlated because of causality. For example, education and income might be correlated because higher education causes higher income, or people with higher income might receive higher education.

In order to deal with correlated variables, PCA was applied to *X'*. Then using *m* of the derived principal components *X'* can be represented as a new matrix *A*, where each row in *A* represents a census tract and each column represents one of the *m* selected principal components. In other words, each entry a_{ij} represents the factor loading of the *ith* census tract onto Principal Component *j*. A *N* by *M* data set returns *M* principal components where the first principal component explains maximum variation in data. The second principal component explains the second most variation orthogonal and is orthogonal to the first principal component. The third principal component explains maximum variation in the data, and is orthogonal to the first two, and so on. The purpose of being orthogonal means that none of the principal components will be correlated with one another. Because of this the principal components will be describing different aspects of the data. PCA represents the data as uncorrelated variables.

There is no rule for choosing *m*. However the purpose of PCA is to reduce the dimensionality of the data while preserving as much information as possible. Choose the number of principal components that explain most of the variation in the data and the least number. The elbow rule was used. The elbow rule says to plot the explained variance of the principal components in decreasing order, and then represent the data using principal components up to the “elbow” in the explained variance graph.

Bottom-up hierarchical clustering with complete-linkage was then run on the dimension reduced data set, *A*. Complete-link clustering defines the distance between two clusters as the maximum distance between points. Formally, this can be written as:

$$d(u,v)=\max(\text{dist}(u[i],v[j]))$$

For all points *i* in cluster *u* and all points *j* in cluster *v*, and where *dist* is euclidean distance. Complete-linkage was chosen because it tends to form more compact clusters. Based on

two-dimensional scatter plot visualizations (Appendix Figure 2), the data points seem to be close together in general, so we want to separate the points as much as possible.

Hierarchical clustering was used over k-means for several reasons. In this classification problem there are no “true” labels for the census tracts. Instead the clustering results are used to describe the data. Hierarchical clustering allows the user to examine the structure of the groupings and understand where it is reasonable to draw an initialization. K-mean requires that the number of clusters is known ex-ante. K-means is sensitive to initialization, meaning census tracts might be in different clusters for different runs of the algorithm. Having census tracts in different groups given the same variables doesn’t make sense for the fact that this exercise is for descriptive purposes. Because there are only 96 census tracts to classify, Finally, K-means is often used because it is faster than hierarchical clustering, however since there are only 96 data points to classify,speed is not an issue.

The clustering process is shown in a dendrogram, and then the maximum distance between clusters is decided based on what makes sense looking at the diagram. As discussed earlier, these clusterings are for descriptive purposes only, hence there are no “true” clusters so it is justifiable to choose distance based on what will best suit the description.

Finally, in order to give meaning to the clusters, boxplots for certain variables that greatly differed across clusters were examined. The differences between these variables would be driving the distance between groups, and hence would be key in defining the groups. Variables with high factor loadings onto the principal components and variables with statistically significantly different means across clusters based on ANOVA results were examined. Principal components are calculated to describe the maximum variance in the data. Hence, if census tracts differ greatly from one another across one variable, or several variables moving in the same direction, the principal components will capture this variation. These should be the same as the variables found by PCA, since variables with a different mean will probably be driving a lot of the variation in the data. Again, it is important to consider that we are using these statistical methods to describe the data, we are not looking for causation. These methods are for recommending variables to describe the clusters.

2.3 Temporal Analysis for Business Growth

Temporal analysis was used to examine how business growth changes across the census tracts with respect to particular variables over time. Business growth for different businesses might not be the same. Also, the variables might not affect all the businesses in a similar way. Temporal analysis will discover if such a pattern exists.

The number of businesses is used as a proxy for business growth, and is the response in our temporal analysis number of business break and enters, and gross assessment or gross land value per square feet are considered as the explanatory variables. Since the response is count, and we have a temporal setup, it is suggested to use the generalized mixed effects models or the generalized estimating equations as the methods to analyze the temporal data available in this project. as the temporal features for each census tract over the years 2011 to 2016.

If the City of Surrey is interested in identifying effects for each census tracts, then the generalized mixed effects models are more useful. Otherwise, for overall effect across the census tracts, the City can utilize the generalized estimating equations. Note that, to examine whether the business growth is affected by the explanatory variables, we need to use the interactions of the explanatory variables and the year in the model. The coefficient of the interactions will give a measure of the temporal relationship between the business growth and the explanatory variables.

3. Results

3.1 Data

The following types of data sets were used:

a. Geographic data,
b. 2011 National Household Survey (NHS),
c. Business data,
d. Commercial rental listing data,
e. Job postings,
f. Property detail listings,
g. Business break and enter data
h. New building permit data

The City of Surrey originally wanted the analysis to be conducted across town centres. However, analyzing the data at a higher granularity allows for more insight. The demographic data source, the 2011 National Household Survey (NHS), is available at the census tract[2] level only. Finer level data is released but contains fewer variables due to confidentiality issues. The remaining data is available as latitude-longitude points. Therefore all the variables of interest are described using census tracts as a unit of analysis. Individual geographic points are aggregated using census tract boundaries from Statistics Canada.

The 2011 NHS is publicly available on the Statistic Canada website. It contains demographic information, including information about a given population's income, languages spoken, citizenship status, education, ethnic origins, age, commuting duration, and residence. The data is divided by Topic and Characteristic, where one Topic (e.g. Income) will have several Characteristics pertaining to it. For example for Income, different characteristics will include different income ranges (such as the variable "number of people who make under 5,000 dollars after tax"). In the original CSV file, there are approximately 29 Topics, each with various Characteristics pertaining to the topic. The final data matrix contains 16 of the 29 Topics,

[2] Census tracts are areas with a population between 2,500 and 8,000 persons. Located in census metropolitan areas and in census agglomerations that have a core population of 50,000 or more, like Vancouver (Statistics Canada).

assuming these would be these most useful. In the data cleaning process, the data was standardized in order to account for different types of units and differences of population sizes. Finally the data frame needed to be changed from long to wide form.

The business data was comprised of two data sets containing all commercial and home business licenses issued by the City of Surrey, and the City of Surrey's classification of those business licenses with the six-digit North American Industry Classification System (NAICS) code [3]. Business information regarding home businesses was provided by the the City of Surrey's Economic Development Division for this project, and is not publically available due to privacy concerns.

The business license data was not collected for the purpose of analysis, and as a result, the following issues were encountered. Some business entities might have multiple business licenses. For example, a gas station might have a license for both the ATM and the gas station. Because the NAICS classifications were done by hand after the time of collection of business licenses a business might not have a corresponding entry in the NAICS file. In addition, some businesses might have several NAICS codes. Variables of interest include: types of businesses (Home or Commercial, and the NAICS classification), years of operation, and number of employees.

The Economic Development Division at the City of Surrey also supplied commercial rental data and the new building permit data for commercial and industrial properties. The commercial rental data is all the commercial suites posted on Spacelist[4] from June 2017. We have the value and the number of new building permits issued from 2013 to 2017 by the City of Surrey. We hypothesize these variables will have relationship with business growth, and can distinguish the town centres and census tracts from one another in terms of business opportunity.

Job posting data was obtained from WorkBC upon request. This data contained job postings from the WorkBC website from June 2017. Since the job posting data contained latitude and longitude, it was possible to find the jobs located in Surrey only.

Property detail listings for 2011 to 2012 were provided to by the City of Surrey. Data for the years 2013 to 2016 can be obtained through the City of Surrey's Open Data catalogue. The gross assessment value per square feet for the businesses and other properties is the variable of interest from the property listing data. Temporal analysis can be used to examine how the change in the gross assessment values affect businesses over time.

Number of business break and enters are taken from the a crime data set originally obtained through the City of Surrey's Open Data catalogue, but is no longer available online at this time. This data set covers the years 2011 to 2016 and gives an the address of where each incident was reported.

---

[3] The NAICS is an industry classification system developed by the statistical agencies of Canada, Mexico and the United States designed in order to facilitate the industrial structure of the three countries (Statistics Canada). The six-digit codes can be generalized up to 2-digits. However, it is often criticized for failing to capture information about new industries. That being said, we were unable to find a better classification of businesses. This system is currently used by the City of Surrey and has been used for other projects studying businesses (ex. Currid and Connoly, 2008).

[4] A commercial real estate website, that allows building owners to post post suites available to rent, and allows businesses to search for available rental spaces.

3.2 Exploratory Analyses

Figure 3.1 shows the number of active home and commercial businesses in 2011 and 2016. Figure 3.1 shows that most commercial and home businesses are in the Newton (NTC) in 2016. The number of commercial businesses is relatively high compared to the home businesses in South Surrey (SSTC), Newton (NTC), Guildford (GTC), Surrey City Centre (SCC), and Cloverdale (CTC). On the other hand, Whalley (WTC) and Fleetwood (FTC) have more home businesses than the commercial businesses. A similar pattern exists in 2011 except for Whalley, which had more commercial businesses than home businesses.

One important thing to note from Figure 3.1 is that home businesses are growing fast across almost all town centres over the years. Another important observation is that the number of total businesses almost doubled in 2016 compared to 2011. A similar plot can be shown for the types of businesses according to NAICS classification (Figure 1 Appendix 1).

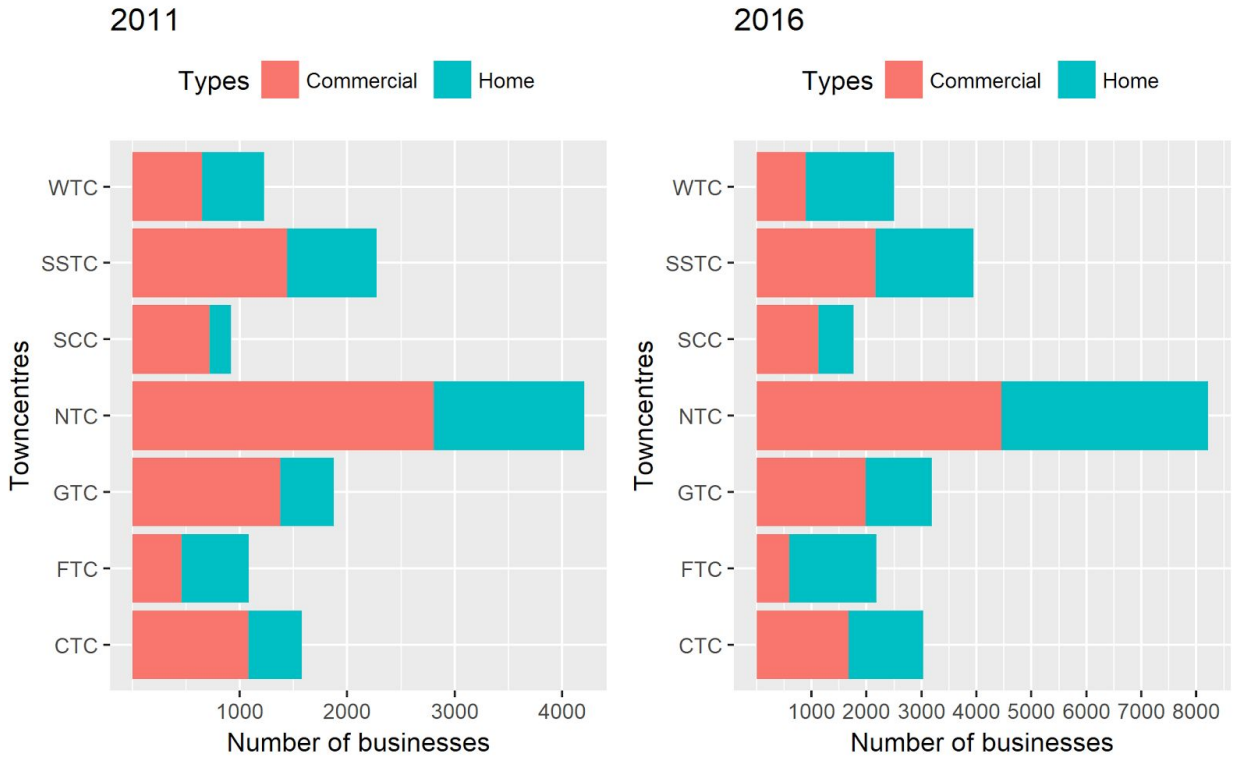


Figure 3.1: Number of active businesses in 2011 and 2016

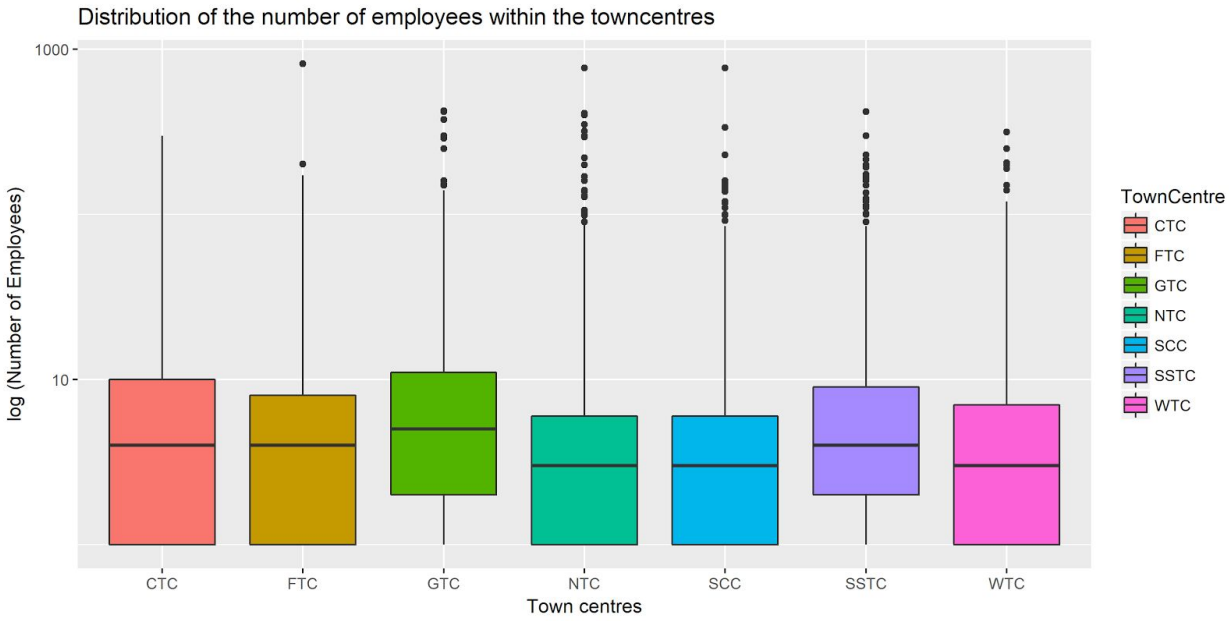


Figure 3.2: Distribution of log number of employees for active commercial businesses

The distribution of the number of employees for active commercial businesses is also interesting (Figure 3.2). The number of employees is log scaled as some businesses have substantially more employees than most of the businesses. Log scaling makes it possible to plot those very large numbers. From Figure 3.2, a heavy concentration of the log number of employees is observed at left side of the boxplots, precisely at zero. This implies many commercial businesses have just one employee in each town center. It turns out this is because the number of employees is only collected once, usually at the time when the business applied for license. In addition, the accuracy is not enforced. For calculating summary statistics about the number of employees, businesses with one employee were dropped.

A histogram of the gross assessment value per square feet for non-government businesses in 2011 and 2016 is shown in Figure 3.3. Government owned properties are excluded since their value is not assessed since they do not pay taxes on the properties. There are two results from the histogram. First, it shows an increasing number of properties with a high gross assessment value. Second, from the area under the histogram, it is clear that Newton (NTC) has most properties followed by South Surrey (SSTC), Cloverdale (CTC), and Whalley (WTC). This is consistent with the previous exploratory analyses.

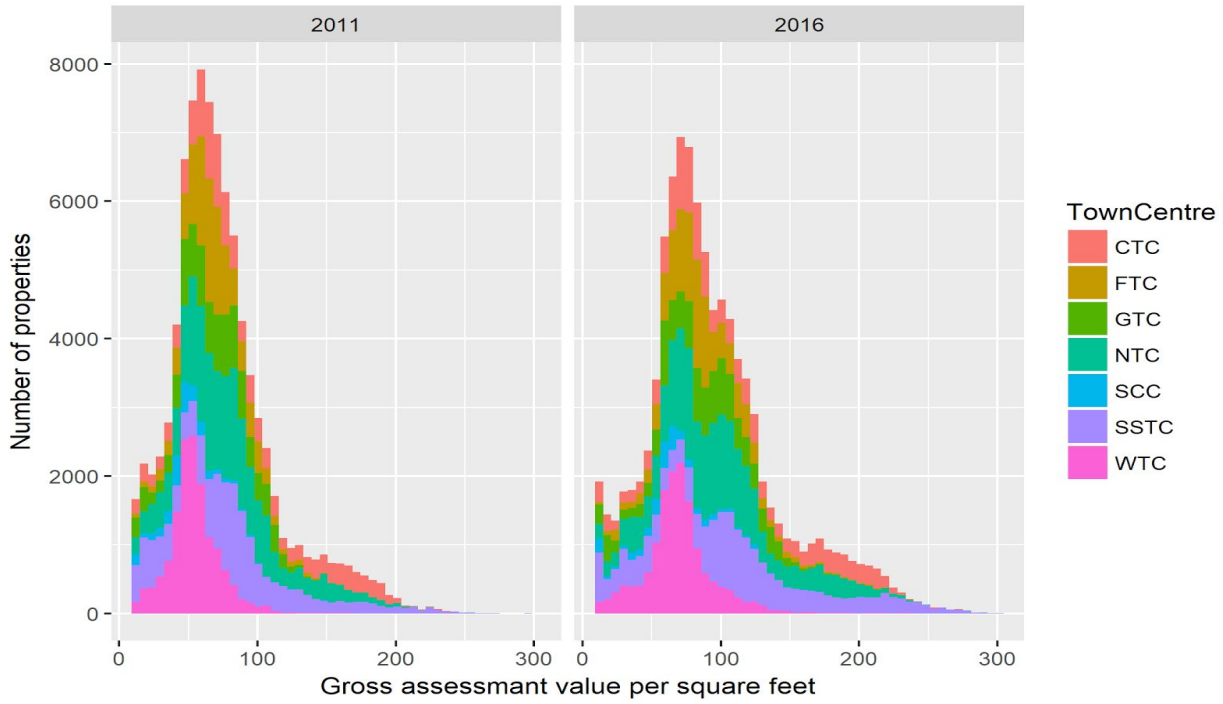


Figure 3.3: Distribution of gross assessment value per square feet for non-government businesses.

Finally, exploratory analysis was conducted for the crime data. Figure 3.4 show a time series plot of the number of break and enters across Town Centres, from January 2011 to October 2016. It is clear from Figure 3.3 that Newton has most business break and enters among all town centres followed by Surrey City Centre and South Surrey. It makes sense though as Newton and South Surrey have most and second most businesses respectively. Cloverdale is the least prone to crime among the city centres in terms of raw numbers.

The number of crimes is then normalized by the number of commercial businesses in Table 3.1. From Table 3.1, it is clear that Whalley has the highest number of business break and enters per business entity in all six years followed by Surrey City Centre, Newton and Guildford. Like as previous, Cloverdale is the least prone to crime according to the number of business break and enters per business entity. Over the years from 2011 to 2016, the number of business break and enters per business entity decreases. However, we see some interesting pattern as the number of business break and enters per business entity increases at 2014 compared to 2013 and then decreases again for almost all the town centres. This is quite interesting though we do not know the real reason that accelerated the break and enters in 2014.

Figure 3.4: Business break and enters across the town centres

	Number of business break and enters per business entity					
Town centers	2011	2012	2013	2014	2015	2016
CTC	0.06	0.09	0.10	0.14	0.09	0.08
FTC	0.98	0.84	0.99	1.20	0.83	0.72
GTC	0.49	0.46	0.50	0.61	0.46	0.38
NTC	0.93	0.86	0.82	0.99	0.72	0.60
SCC	1.91	1.43	1.39	1.66	1.23	1.09
SSTC	0.57	0.58	0.50	0.59	0.41	0.38
WTC	2.30	2.10	2.04	2.18	1.84	1.40

Table 3.1 Number of business break and enters per business entity from 2011 to 2016

This section does not discuss all the exploratory analysis conducted for this project. More results can be found in the Appendix.

3.3 Visual Surrey

Visual Surrey was created in an effort to create an interactive and browseable visual representation of the economic profile of Surrey. Visual Surrey can represent various features from the datasets, including demographic, crime, business and land value information. Figure 3.5 gives a snapshot of what Visual Surrey looks like.

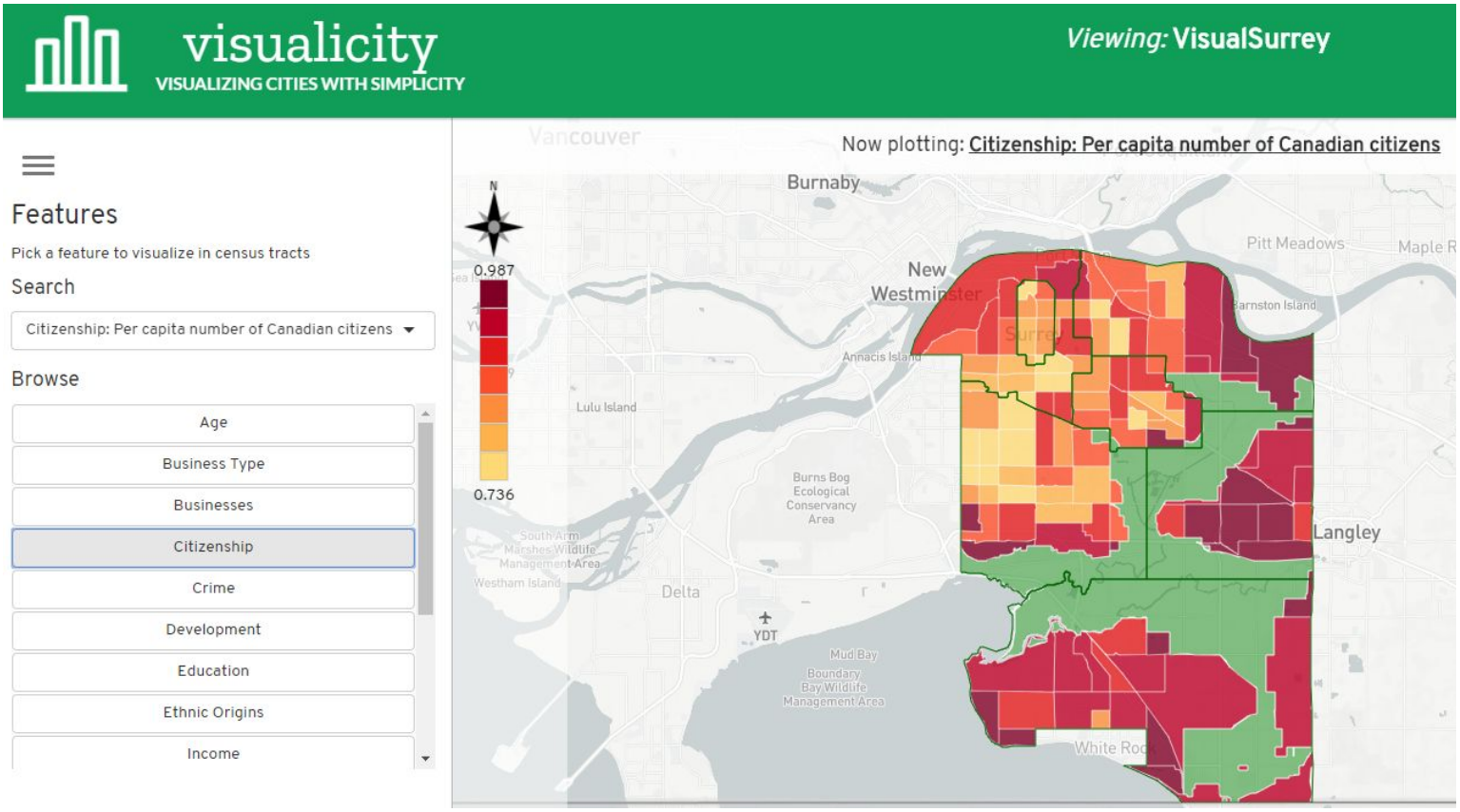


Figure 3.5: Visual Surrey showing Citizenship: Per capita number of Canadian Citizens

Visual Surrey consists of a front-end and a back-end. The front-end user interface was built using Bootstrap to make it mobile friendly. Geographical shapes were constructed using D3.js. Map data is first read via a RESTful API service on the backend (built in python/Flask), which returns data in JSON format to the frontend. D3.js was used to parse the resulting data and color in the shapes; interactivity and styling was also accomplished using D3.js. Leaflet.js was used to display the map on the background for context.

The features available are connected to a spreadsheet with all of our datasets discussed in Section 3.1, with all cleaned datasets merged into a single sheet. The sheet (powered by

Google Spreadsheets) can be updated and will be automatically added to Visual Surrey, available to physically see where the feature is more or less dense across the geographical area of Surrey by individual census tract, as well as within the larger grouping of Town Centres which are the divided census tracts separated by the dark green lines. The green areas within the map are Marshlands, which are empty and protected and are not available for residential or commercial use. The map is interactive, and you can click on specific census tracts to find the specific number for that area if so desired.

The legend indicates that the darker, red areas have a higher number of whatever feature Visual Surrey is showing. In Figure 3.5, the highest percentage of Canadian Citizens within a census tract in Surrey is 0.987, meaning that there is a census tract with 98.7% of people are Canadian Citizens. The higher the percentage, the darker the census tract will be. The lowest percentage of Canadian Citizens within a census tract is shown by the lighter, more orange or yellow colours. For Figure 3.5, the lowest percentage a census tract has is 0.736, meaning that there is a census tract with 73.6% of people within the tract are Canadian Citizens. This feature is on a per capita basis, so each census tract has been adjusted for population. As seen in Figure 3.5, the more southern and eastern parts of Surrey have more Canadian Citizens per capita generally speaking, with the northern and western parts of Surrey have less Canadian Citizens per capita.

Visual Surrey is searchable and browsable by topic, making it easy to look up many different features and get a better understanding of what the demographic and economic profile of Surrey is. This allows someone who may not know Surrey overly well to see Surrey on a broad scale. The map has additional controls, such as toggling whether or not you want Town Centres to be visible, changing the opacity of census tract colours, and the ability to search for a specific Census Tract ID.

Visual Surrey also has a Heat Map option for certain features, such as leasable properties (from Spacelist), break and enters in 2016 and the number of businesses of by type. The heat map is useful in the sense where you can visualize its available features and also visualize a different feature on the census tracts, allowing for comparisons. This is shown in Figure 3.6.

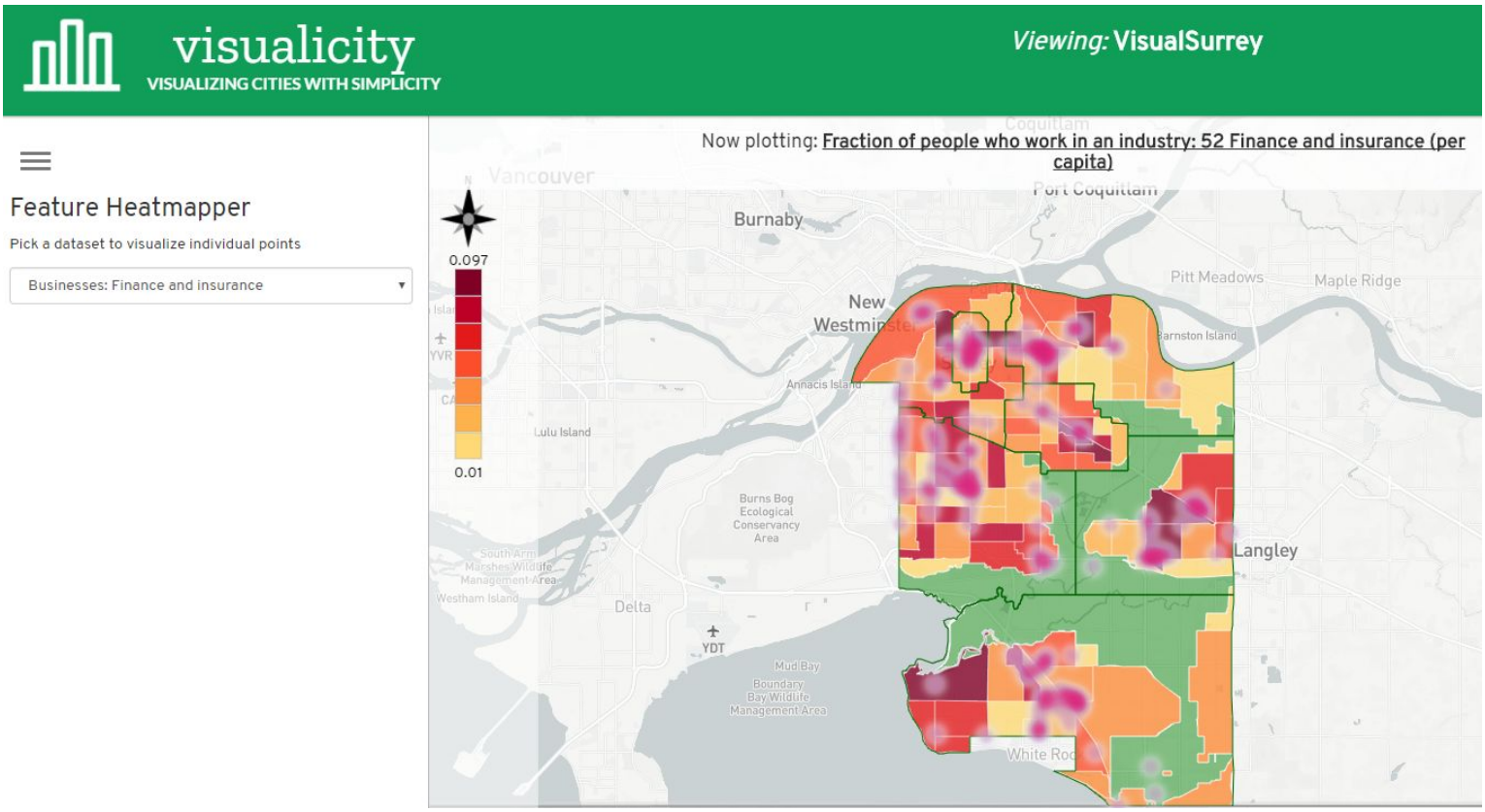


Figure 3.6: Visual Surrey plotting Fraction of People working in Finance & Insurance and Heatmapper plotting Number of Finance and Insurance businesses.

By being able to plot two types of features at once, we can see some valuable information that can only be found by observing data visualization. For example, the City of Surrey wanted to know if generally people lived where they worked. By just the census tract plotting alone (which is showing where people who work in Finance and Insurance live) we see there are darker regions in the south-western part of Surrey (within the town centre of South Surrey, as well as in the east (within the town centre of Cloverdale). With the overlap of the Heatmapper (showing where the physical finance and insurance businesses are) we see that a high percentage of people who work in finance and insurance live in Cloverdale, where there is also a large number of this type of businesses. On the other hand, in South Surrey where there is a high percentage of people who work in finance and insurance, there is a low number of businesses in this industry. Therefore, within this specific area, people do not generally live where they work in terms of census tracts. However, the Heatmapper has a downfall regarding its inability to show physical numbers on the map, so while we may see that there is a dense area of finance and insurance businesses in north-west side of Surrey, we will not know the exact number, and what one might consider a high number of businesses is relative.

This same question, do people live where the work, can be answered with Visual Surrey without using the Heatmapper as well, especially if you want to take into account different population densities.

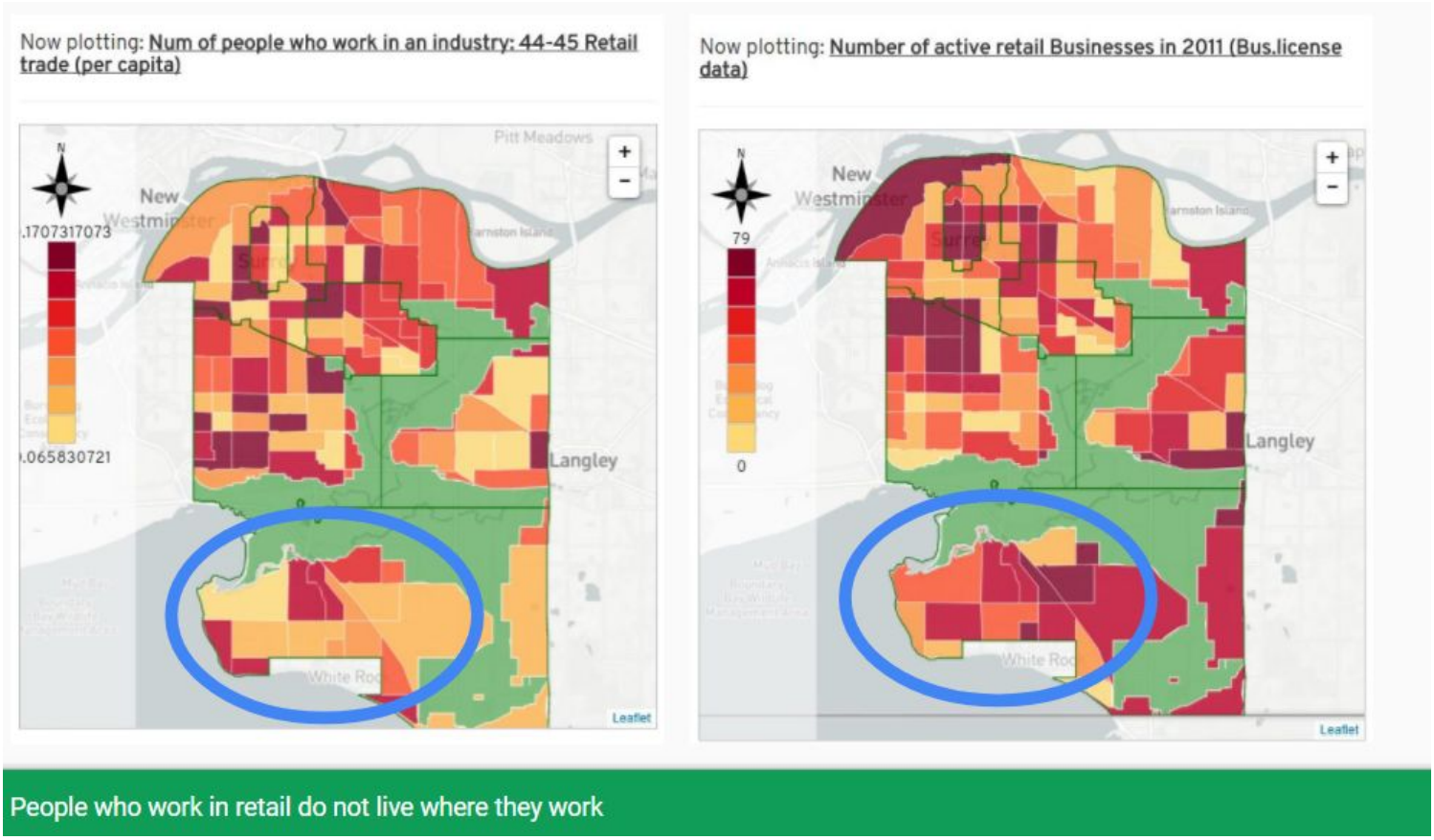


Figure 3.7: Visual Surrey plotting the number of people who work the retail industry per capita on the left hand side, and the number of active retail businesses is plotted on the right. Both maps have the most western side of the town centre South Surrey circled.

In figure 3.7, we have circled a specific area to demonstrate how you can answer the question, do people who work in retail live where they work using Visual Surrey without the Heatmapper. The left image shows the number of people who work in retail, and from within the circle we can see that this area demonstrates predominantly light colours, indicating that a high percentage of people in this area do not work in retail. On the right we see the number of active retail businesses, and within the circled area this is predominantly darker colours, indicating a high number of retail businesses are in this area. Therefore, for this circled area we can draw the conclusion that within the retail industry, people generally do not live where they work.

Here we have shown just a few examples of how Visual Surrey can be used for simple visualization and to answering general questions about the economic profile of Surrey. The inspiration for Visual Surrey is that no matter the question, the data and the visualization can

provide some kind of insight. By being searchable and browseable, the City of Surrey as well as public users can understand Surrey better on a demographic, economic and social level, allowing for more influential and informed decision making.

3.4 PCA and Clustering Results

Based on the results of PCA, four principal components were chosen to represent the data matrix before running hierarchical clustering. The variance described clearly starts to decrease after the fifth principal component (Figure 3.8), so the “elbow rule” discussed earlier would suggest using five principal components. However, only 4 principal components were used for the following reasons. First, when looking at the highest factor loadings for each principal component, Principal Component 5 appears to contain information already captured by the previous principal components (Table 3.2). Secondly, it is difficult to justify using Principal Component 5 over Principal Component 6. Principal Component 5 captures elements already captured by the previous principal components(Table 3.2), while Principal Component 6 appears to capture new information, the number of building permits issued (Table 3.3). Variables have a lower factor loading for principal component 5 compared to principal component 6. In addition, Principal Component 5 explains 7.8 percent of the variance in the data, while Principal Component 6 explains 7.1 percent. One could easily change the variance by including or excluding different variables. Since this exercise is for descriptive purposes it makes sense to use only principal components 1, 2, 3, and 4.

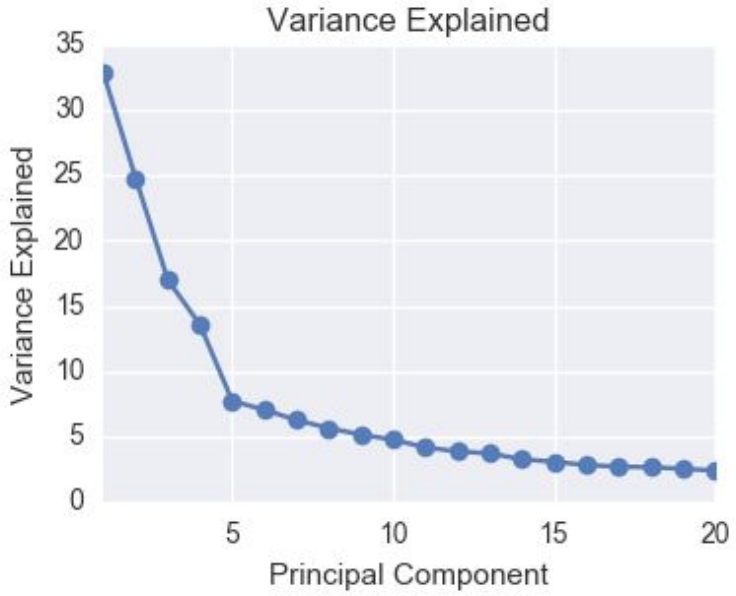


Figure 3.8 Variance explained by the first twenty principal components

Factor Loading	Variable
-0.494543	Median monthly shelter costs for owned dwellings in dollars
-0.476738	Income[5] of individuals $40,000 to $49,999
-0.474432	Occupied private dwellings with 4 or more bedrooms
0.440467	Ethnic origin - North American Aboriginal
-0.421344	Apprenticeship or trades certificate

Variance Explained: 7.8%

Table 3.2 Highest factor loadings of Principal Component 5

Factor Loading	Variable
-0.647260	Commercial building permits 2016
-0.643797	Commercial building permits 2014
-0.627396	Commercial building permits 2015
-0.564092	Industrial building permits 2017
-0.549579	Industrial permits 2016

Variance Explained: 7.1%

Table 3.3 Highest factor loadings of Principal Component 6

For each principal component used, the greatest magnitude factor loadings were plotted to show which variables contributed the most to each principal component, and hence what the clustering would be based on. The four principal components appear to have variables loaded onto it related to either business characteristics or demographic characteristics, but not both (Appendix 1 Tables 1-4).

Variables describing income, education, and occupation were loaded onto the first principal component. Income and education are correlated with each other based on economic literature, and we were worried about clustering the data on these variables. In addition, "redundant" variables, for example, the multiple measures of income, are loaded onto Principal Component 1. People working in occupations that often do not require a postsecondary degree

[5] All income is In 2010 dollars after tax

are correlated with people not having a postsecondary degree. People with an education degree is negatively correlated with people working in occupation 9. These results may make sense since someone with an education degree would not work in manufacturing unless they changed their career path. However, this relationship would not be obvious to us if we were to remove these variables by hand.

Principal Component 2 had high factor loadings for variables related to the number of businesses within a census tract (Appendix Table 2). This included variables that described the number of businesses of a certain type and in a particular year. Interestingly, it also had the percentage of of commercial and home businesses. Before performing this analysis, we could not confidently say would have not expected the percentage of commercial businesses to be correlated with number of businesses, for example, again exhibiting the importance of using PCA for variable selection.

Principal Component 3 has a high loading of variables related the percentage of certain kinds of businesses. Finally principal component 4 had high factor loadings for variables related to shelter costs and dwelling[6] characteristics.

Then hierarchical clustering was applied to the dimension reduced data set, where each row represents a census tract, and each column represents one of the four principal components. Based on the dendrogram produced, 17 was chosen as the cutoff distance between clusters. This yielded five main groups and two groups of outliers, where different colours indicate different groups (Figure 3.9).

[6] "A separate set of living quarters designed for or converted for human habitation in which a person or group of persons reside or could reside. In addition, a private dwelling must have a source of heat or power and must be an enclosed space that provides shelter from the elements, as evidenced by complete and enclosed walls and roof, and by doors and windows that provide protection from wind, rain and snow." (Statistics Canada)

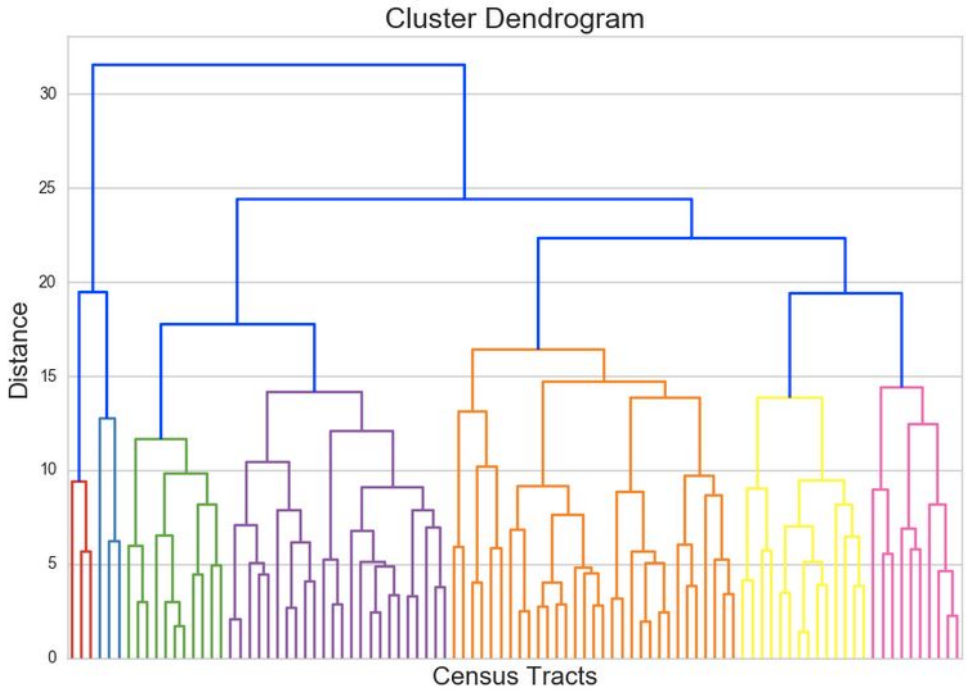


Figure 3.9 Clustering of census tracts on principal components

The clusters were then examined using different visualization methods to check if the clustering worked well. The census tracts labeled with their label were plotted in two dimensional space where using each principal component on the different axis (Appendix Figure 2). The labels seem to be well separated. Groups that are not well separated using a particular two principal components will be well separated using other principal components. The clusters were also visualized using Visual Surrey (Appendix Figure 3). Clusters of a similar type seem to group together geographically, which is particularly interesting considering that no information about the proximity of census tracts to one another was in the data set. There appears to be a residential-business divide.

3.5 Cluster Analysis

Boxplots for variables with high factor loadings onto the principal components and variables with statistically significantly different means across clusters based on ANOVA results were examined. These two methods returned the same variables (Appendix, Figure 3 and 4).

Looking at the two dimensional visualization of these clusters along Principal Components 1 and 2, (Appendix Figure 2) Cluster 1 and 2 are outliers with a small number of census tracts assigned to these clusters. When looking at the boxplots, clusters 1 and 2 were clearly separated by a number of variables including number of commercial businesses, manufacturing and wholesale businesses. In addition, Cluster 1 had a high businesses per square kilometer and high number of retail businesses. This makes sense as it appears to be in a business centre. Cluster 7 has a high number of businesses per square kilometer, and a high percentage of retail businesses. Again, cluster 7 appears to be a business cluster. Clusters 2, 3,

and 4 appear to have relatively high income and education while 1, 5 and 7 are relatively lower. Clusters 3, 4, and 6 have mostly mostly home businesses. Aside from these findings, it is difficult to construct a story about Clusters 2, 3, 4 and 5 because they do not clearly differ across many variables.

Perhaps the PCA and clustering can produce more meaningful results by making further improvements. Different methods could be used to standardizing the variables before applying PCA. It would also be useful if more variables were available, with larger variation. This would hopefully create more distinctive clusters.

3.6 Temporal Analysis

As discussed earlier, temporal analysis was performed using the data with a time component, in order to examine the factors that contributed for the business growth over time. To do so, the response is the number of businesses, which can be thought of a proxy for business growth. The explanatory variables are year, gross assessment or land value per square feet, and number of business break and enters. Year is considered a factor in our temporal analysis. Also, the interaction between the year and the other explanatory variables is considered. These interactions allow the interpretation of whether the available explanatory variables have any effect on the increasing number of businesses over the year. Temporal analysis for each type of business was conducted separately.

A generalized mixed effects Poisson model was used to fit the count variables, number of businesses of any type in this temporal setup. As an example, the number of commercial businesses is considered the response. From the fitted result, the main effects of all years, with respect to 2011, is a statistically significant contributor to the increase of commercial business over the year at the 5% and 10% significance level. However, the interaction effects of the years and median gross land value per square feet and the main effect of the number of business break and enters are not contributing significantly to the increase of commercial business. The interaction plot is shown in Appendix 1 Figure 6.

The lines for different years are quite parallel. This implies that on an average, the change in the businesses for a specific median gross land value per square feet is the same for all the years for a specific census tract. Hence, the effect of years does not alter the effect of the median gross land value per square feet on the number of commercial businesses over time. However, using a mixed effects model allows for a census tract specific interpretation since there are different intercepts for the census tracts. For overall interpretation across census tracts, the generalized estimating equation (GEE) method was used. The results obtained using GEE method for commercial businesses are shown in Table 3.4. From the estimates using GEE method displayed the number of commercial businesses significantly changes over the year. For example, compared to 2011, there are exp(0.147) = 1.16 times the number of commercial businesses in 2012 when all other variables are held fixed. Similar interpretations hold for the other estimates.

Variables/Effects	**Estimates**	**Std. Error**	**P-value**

Intercept	4.478	0.249	<0.01
Median gross land value per square feet	-0.002	0.003	0.64
Year 2012	0.147	0.015	<0.01
Year 2013	0.237	0.029	<0.01
Year 2014	0.327	0.033	<0.01
Year 2015	0.390	0.037	<0.01
Year 2016	0.470	0.058	<0.01
Number of business break and enters	0.002	0.001	0.20

Table 3.4: Parameter estimates for fitting the number of commercial businesses using GEE method

One can repeat the same temporal analysis for home businesses and the types of businesses according to NAICS separately. For example, for construction businesses, the interaction plot is given in Appendix Figure 7. Figure 7 exhibits the presence of interaction of year and the median gross land value per square feet. From the GEE results, significant (at 5% level of significance) main effect and interactions of Year and the median gross land value per square feet. So, for construction businesses, the change in the median gross land value per square feet did have an effect on the growth over time.

So, for certain business types we have observed a temporal relationship between the business growth and the median gross land value per square feet, and for others we have not. However, we do not want to make a general comment from the results as we only have two explanatory variables. Rather, we have demonstrate a process of finding temporal relationship for future study that can help to understand the uniqueness of the City of Surrey across their geographical space.

4. Discussion

4.1 What is "Social Good" anyway?

Given that Visual Surrey was borne out of the Data Science for Social Good fellowship program, it is reasonable to assume that there would be a certain level of "social goodness" to this project. But what does social good really mean? In terms of DSSG, a program which has been launched across various universities in the United States, and now Europe and Canada, social good for us means creating social impact, dealing with a real-world problem and aiming to bring some kind of benefit to society.

It is a noble mission, one that we are happy to make an effort towards, but how does Visual Surrey bring social good to the people of Surrey? From an objective perspective, Visual Surrey alone does not really have a social good aspect. Its simply a tool that can help the City of Surrey use their data and help understand the city that they work for a little bit better. So the social good element of our project does not stem from the tool itself, but instead from the way it is used.

A strategic priority for the City of Surrey is to attract investment. Visual Surrey can help the City of Surrey better understand where investment money should go,within the city, based on already existing businesses and the demographics of a certain area. At base value, that is what Visual Surrey can offer: information. If the City of Surrey uses this tool to bring investment and opportunity to areas of lower income, lower education, higher number of break and enters per business and therefore areas of higher vulnerability (which, according to Visual Surrey seems to be the general north western area of Surrey) then the use of Visual Surrey would be a tool for bringing social good to the community. By bringing employment opportunity, education and prosperity to a more vulnerable region, Visual Surrey and thus the City of Surrey would be benefiting society.

However, while Visual Surrey has potential to be a tool to bring social good and benefit, it could also be a misused tool that creates social cost instead. Take into consideration their strategic priority to attract investment; if this money only goes to areas that are already prospering and are resident to higher income individuals, this would further diversify the rich and poor within Surrey. If lower income areas do not get the same kind of support, they can be privy to increased crime, drug use and other negative factors that affect economic development and growth. Therefore Visual Surrey must be used as a whole, examining all possibilities and not only those that generate the highest profit.

This product, Visual Surrey is a neutral tool that brings neither social good or bad to Surrey or any community for that matter. It has the potential to bring social good depending on it's use. Therefore it is up to the City of Surrey to ensure that Visual Surrey is used with the foresight of creating social benefit, and not increasing social cost.

4.2 Limitations

The data matrix was standardized before running PCA in order to cope with having variables in different units. However, many variables are in percentage units. For example, percentage of businesses classified as manufacturing and percentage of people whose primary form of transportation is bicycling are both in percentage units. PCA gives weight to variables with higher variance. In this case in particular, we would want to give higher weight to those variables that have a larger variance.

PCA can be used for variable selection when using a data set containing correlated variables. However, certain variables in our data, for example mode income and median income. This could be an issue since PCA is based on explained variance. The fact that we have removed some of the variance information by standardizing the data, will increase the importance of this variance derived from "repeated" variables.

There are some limitations in the temporal analysis. The number of businesses might is not the best proxy for business growth. Better proxies would be revenue of the businesses or

the number of employees. Second, there are zero counts of some business types across the census tracts. To solve this issue, one can use the zero inflated models. We omit this from the analysis because of the unavailability of the software packages and it hinders interpretability.

5. Recommendations

Talking with the Economic Development team we had a better understanding of the limitations they faced. Our primary recommendation stems from the way in which the business license data is collected. A structured questionnaire can be prepared to collect the business license data at each year. It would be useful if variables like the number of employees, revenue were can be collected each year. Those can serve as a good indicator of business growth over the time. This would allow for temporal analysis proxy variables.s

6. Appendix

Figure 1

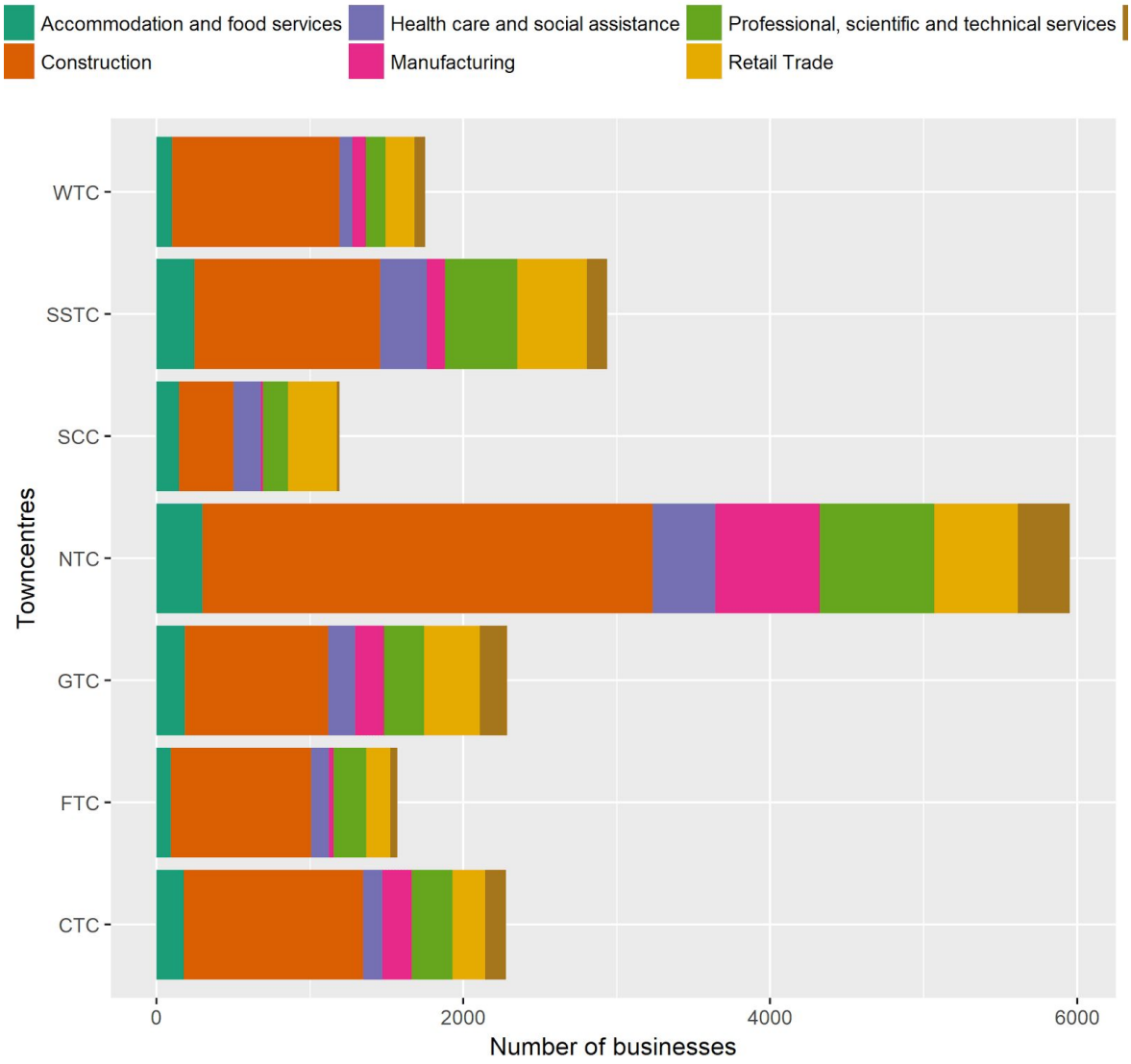
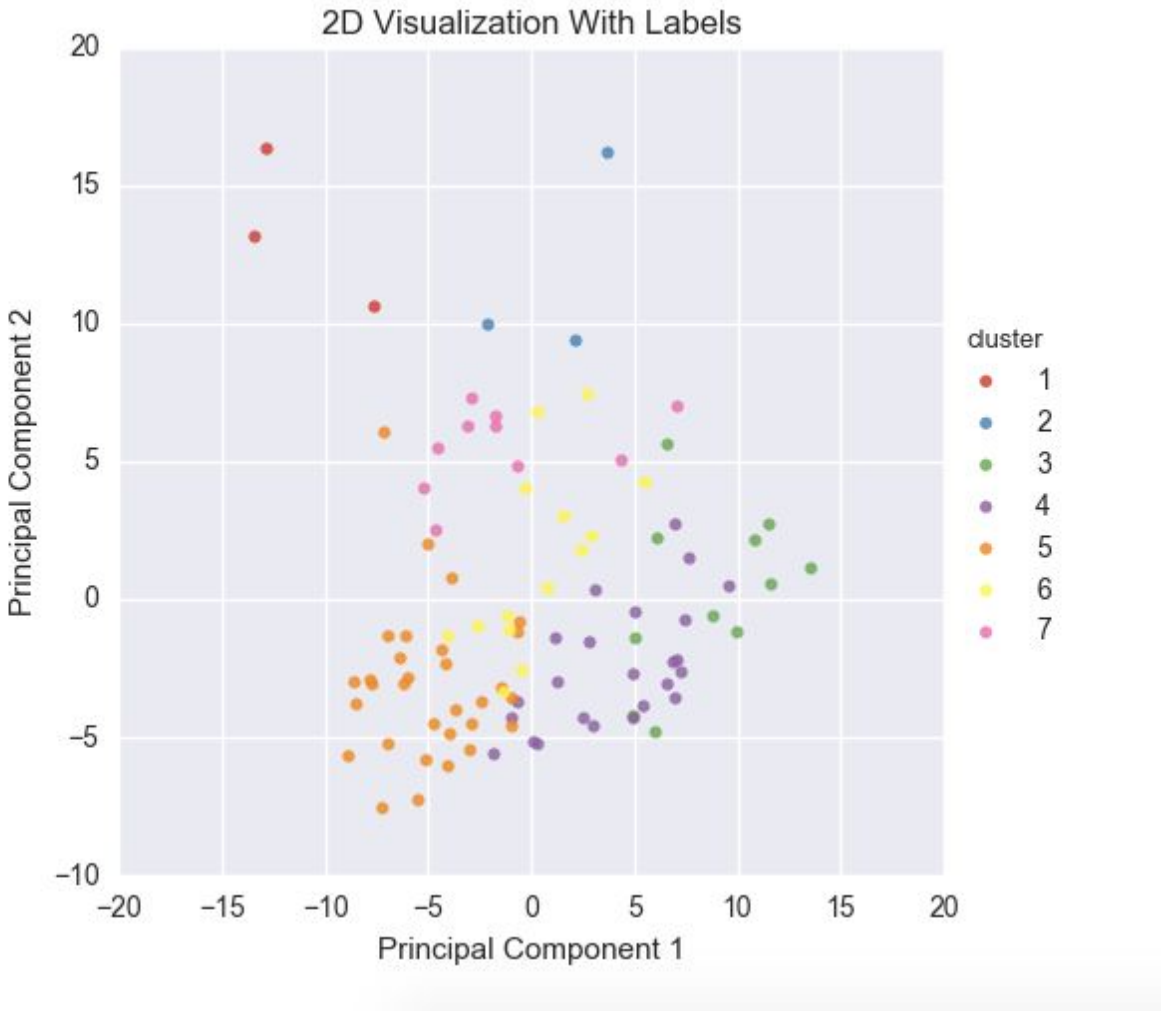


Figure 2


2D Visualization With Labels
Principal Component 2
Principal Component 1
20
15
10
5
0
−5
−10
−20
−15
−10
−5
0
5
10
15
20
cluster
1
2
3
4
5
6
7

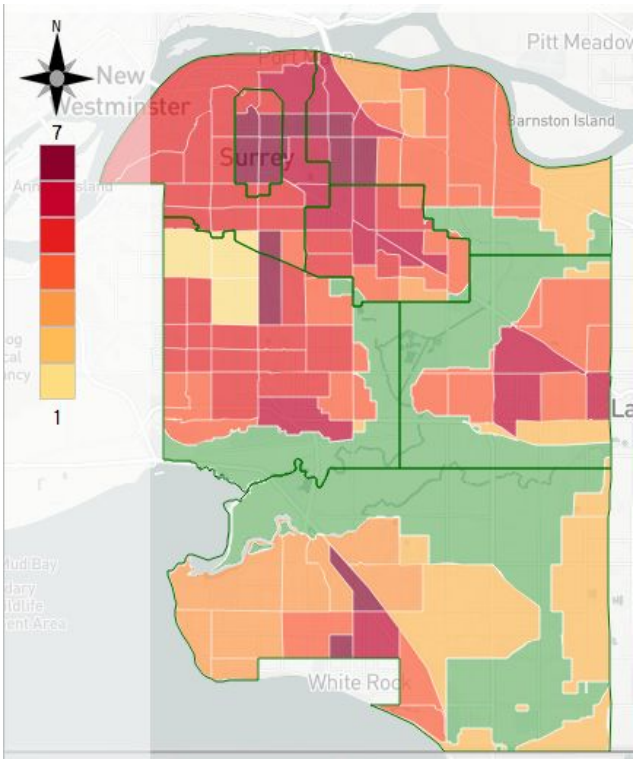


Figure 3 - Cluster Results

Figure 4

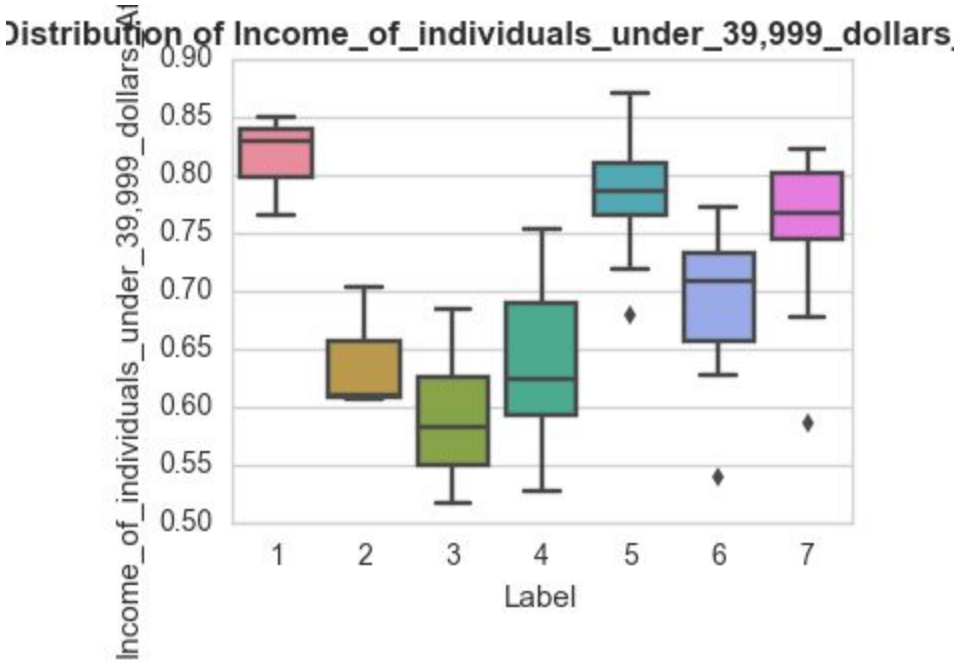


Figure 5

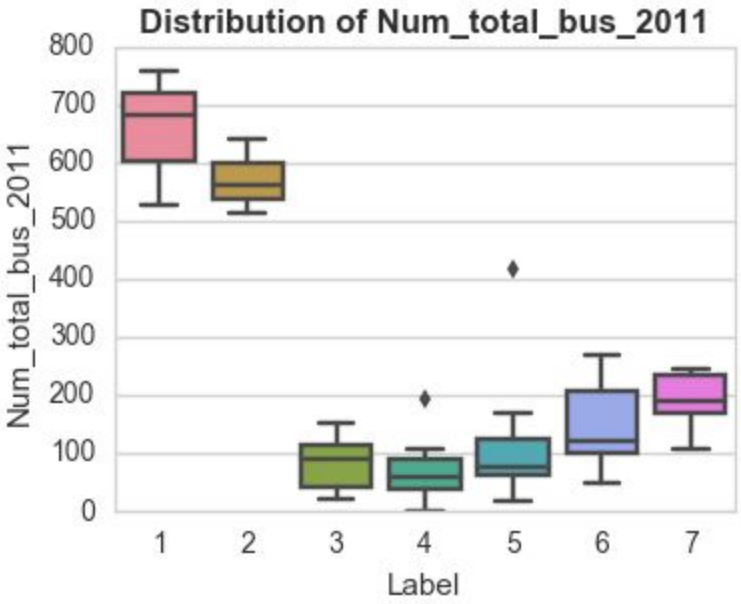


Figure 6

Interaction plot between the Year and the median gross land value per square feet while predicting the business growth for commercial businesses

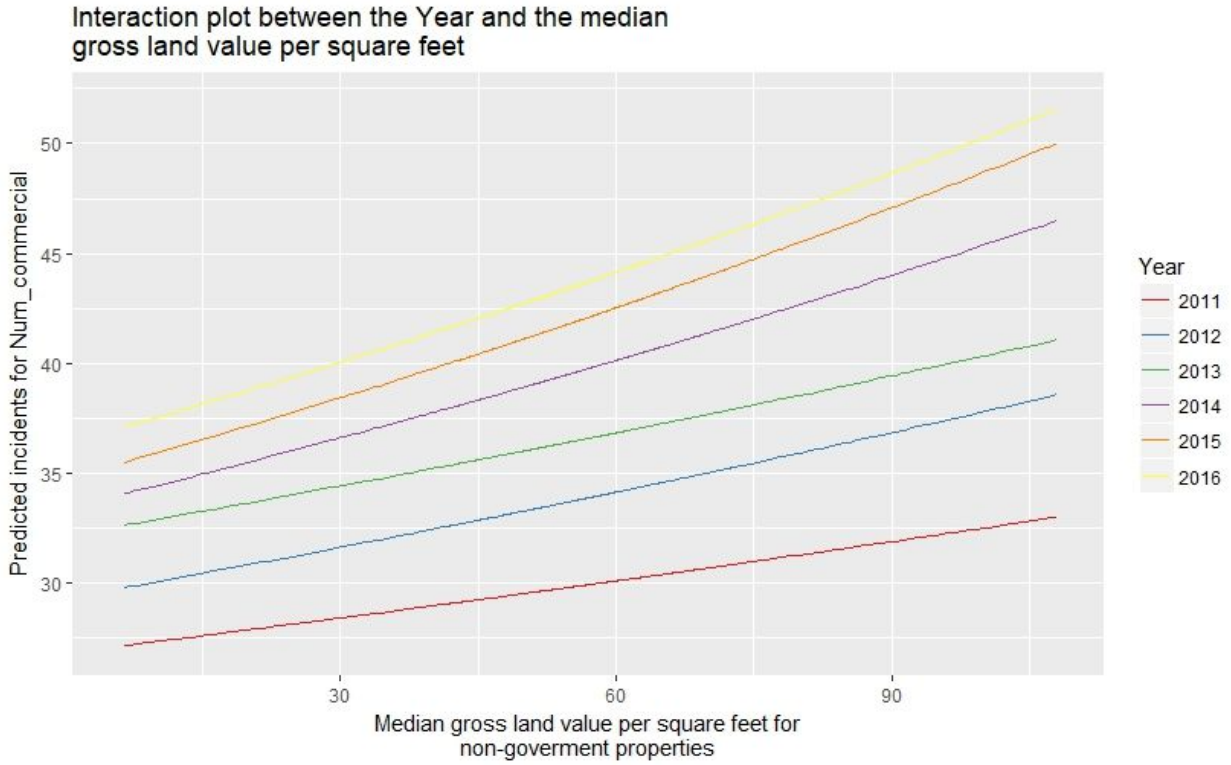


Figure 7

Interaction plot between the Year and the median gross land value per square feet while predicting the business growth for construction businesses

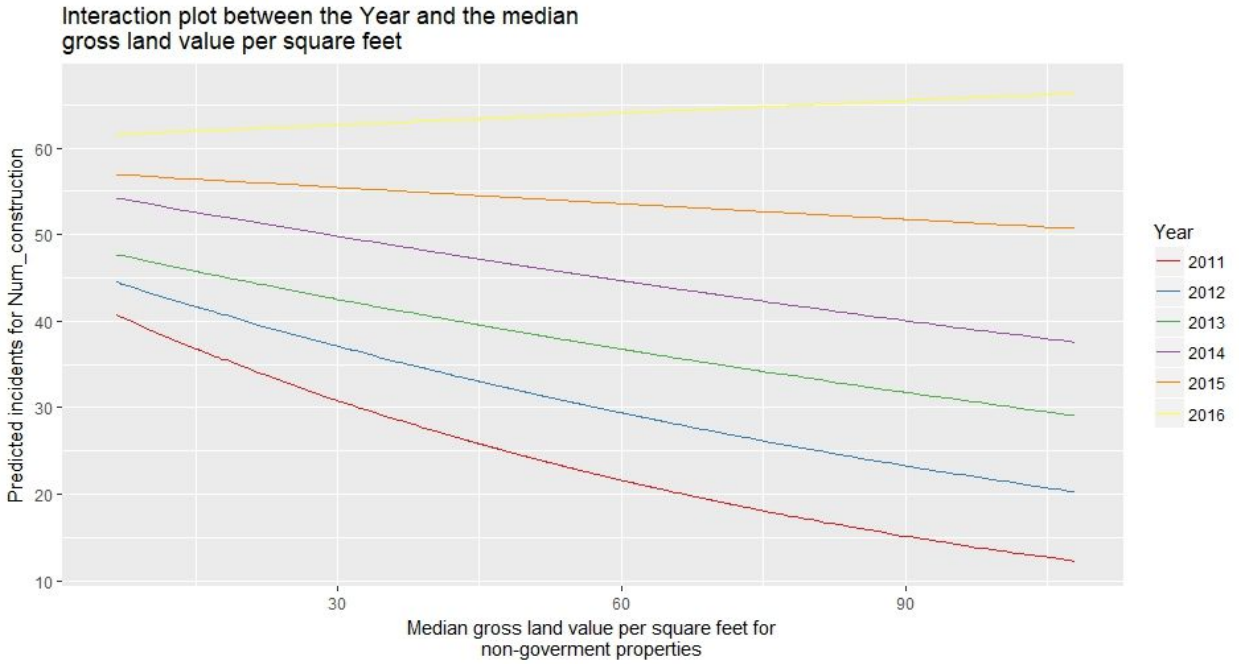


Table 1

Principal Component 1: Income and Education

Factor Loading	Variable
0.935692823	Income of individuals under $39,999
0.928170741	Income of individuals under $49,999
0.898543877	Income of individuals under $59,999
0.897718667	Average income of individuals
0.897513468	Income of individuals under $29,999
0.885983133	No postsecondary certificate, diploma or degree
0.885309347	Postsecondary certificate, diploma or degree
0.880847903	No certificate, diploma or degree
0.863708683	Median income of individuals
0.843590878	Occupation 9 : manufacturing and utilities

0.838764963	Income of individuals $60,000-79,999
0.821129199	Occupation 7 : Trades, transport and equipment operators and related occupations
0.81332313	Income of individuals $80,000- 99,999
0.810810412	Income of individuals under $79,999
0.807072047	Income of individuals under $19,999
0.796626702	Education: Business, management and public administration
0.782964215	Education: Education
0.780283137	Occupation 0: Management occupations
0.777406872	Income of individuals $10,000-14,999
0.768035624	Occupation 4: Education, law and social, community and government services

Table 2
Principal Component 2: Number of businesses, percentage of commercial and percentage of home businesses

Factor Loading	Variable
-0.869868	Number of commercial businesses in 2011
-0.869706	Number of commercial businesses in 2016
-0.855573	Number of total businesses in 2011
-0.828767	Number of professional service businesses in 2016
-0.821368	Number of total businesses in 2016
0.820565	Percentage of home businesses in 2016
-0.820565	Percentage of commercial businesses in 2016
-0.818697	Number of professional service businesses in 2011
-0.814067	Number of accommodation businesses in 2016
-0.769014	Number of retail businesses in 2011

-0.766145	Number of accommodation businesses in 2011
-0.734666	Number of retail businesses in 2016
-0.720208	Number of real estate businesses in 2016
-0.692613	Number of financial institutions 2016
-0.674443	Number of real estate businesses in 2011
-0.665472	Number of businesses per square kilometer
-0.661555	Number of wholesale businesses in 2016
-0.659497	Number of manufacturing businesses in 2011
-0.645145	Number of manufacturing businesses in 2016
-0.623128	Number of wholesale businesses in 2011

Table 3
Principal Component 3: Percentage of retail businesses

Factor Loading	Variable
0.714371505	Percentage retail businesses 2016
0.708419744	Number of construction businesses 2016
0.686762566	Percentage retail businesses 2011
0.669414192	Number of construction businesses 2011
0.657508344	Percentage of accommodation businesses 2016
0.627369329	Number of home businesses 2011
0.625305328	Percentage construction businesses 2011
0.624146069	Percentage of financial institutions 2016
0.621849466	Number of transportation businesses 2016
0.611295694	Number of transportation businesses 2011
0.608172987	Percentage of construction businesses 2016

0.596499885	Number of home businesses 2016
0.595216939	Percentage of financial institutions 2011
0.56806204	Percentage of wholesale businesses 2016
0.553561026	Percentage of transportation businesses 2016
0.548450708	Percentage of real estate businesses 2016
0.527864562	Percentage of accommodation businesses 2011
0.527513325	Percentage of retail businesses 2011
0.52656927	Percentage of wholesale businesses 2011
0.524630238	Number of wholesale businesses 2016

Table 4
Principal Component 4: Owner or tenant households, and other shelter cost info

Factor Loading	Variable
0.841071092	Number of tenant households
0.836984941	Number of owner households
0.744227254	Average number of_rooms per dwelling
0.717671948	Spending less than 30 percent of household total income on shelter costs
0.709899662	Spending 30 percent or more on household total income on shelter costs
0.700362991	Mode of transportation: car, truck or van
0.666023945	Industry 61: Educational services
0.665731927	Median value of dwellings
0.638455727	Mode of transportation: Public transit
0.620705078	Industry 56: Administrative_ and support, waste management and remediation services

Table 5

Principal Component 5: No clear interpretation

Factor Loading	Variable
0.494543118	Median monthly shelter costs for owned dwellings
0.476738317	Income of individuals $40,000-49,999
0.474431976	4 or more bedrooms
0.440467159	Ethnic origin: North American Aboriginal
0.42134417	Apprenticeship or trades certificate or diploma
0.396157482	Percent of tenant households spending 30 percent or more on household total income on shelter costs
0.390644512	Income of individuals under $9,999
0.37201238	Income of individuals under $14,999
0.371561946	Ethnic origin: Asian
0.371424974	Income of individuals $30,000-39,999

Table 6

Principal Component 6: Number of building permits issued

Factor Loading	Variable
0.647259949	Number of commercial building permits issued 2016
0.643797445	Number of commercial building permits issued 2014
0.627395889	Number of commercial building permits issued 2015
0.564092075	Number of industrial building permits issued 2017
0.54957912	Number of industrial building permits issued 2016
0.547438104	Number of industrial building permits issued 2013
0.545434924	Number of industrial building permits issued 2014
0.531745669	Number of commercial building permits issued 2017
0.522060972	Number of industrial building permits issued 2015

0.467610925	Number of commercial building permits issued 2013