Data Science Institute, University of British Columbia

High-dimensional analysis of census tracts within the City of Surrey

Natasha Mattson, Sarah Neubauer, Rashedul Hoque, and Tony Hui

CASCADIA URBAN ANALYTICS COOPERATIVE

Microsoft

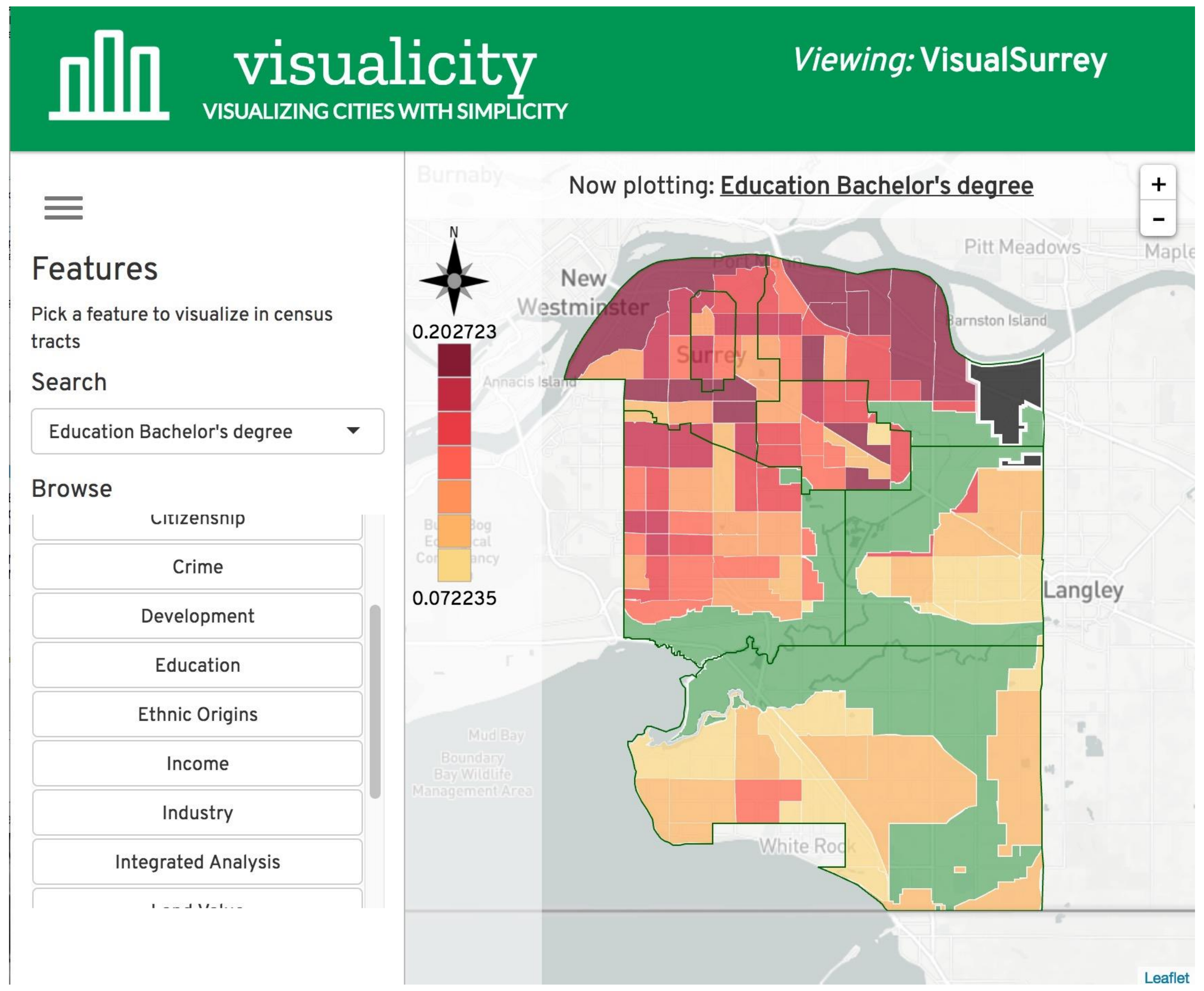
Introduction

PROJECT: Economic Development project with the City of Surrey for the University of British Columbia's (UBC) 2017 Data Science for Social Good (DSSG) fellowship program
PURPOSE: To create an economic profile of Surrey, describing different 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 with these different features.
DATASETS:

- Geographic data
- 2011 National Household Survey (NHS)
- Business licenses
- Commercial rental listings
- Job postings
- Property assessment data
- Business break and enters
- New building permit data

VisualSurrey allows for visualization of important metrics

- A data visualization platform written in Javascript/Python
- Hosted on Microsoft Azure
- Accessible at *bit.ly/visualsurrey*



visualicity
VISUALIZING CITIES WITH SIMPLICITY

Viewing: VisualSurrey

Features

Pick a feature to visualize in census tracts

Search

Education Bachelor's degree

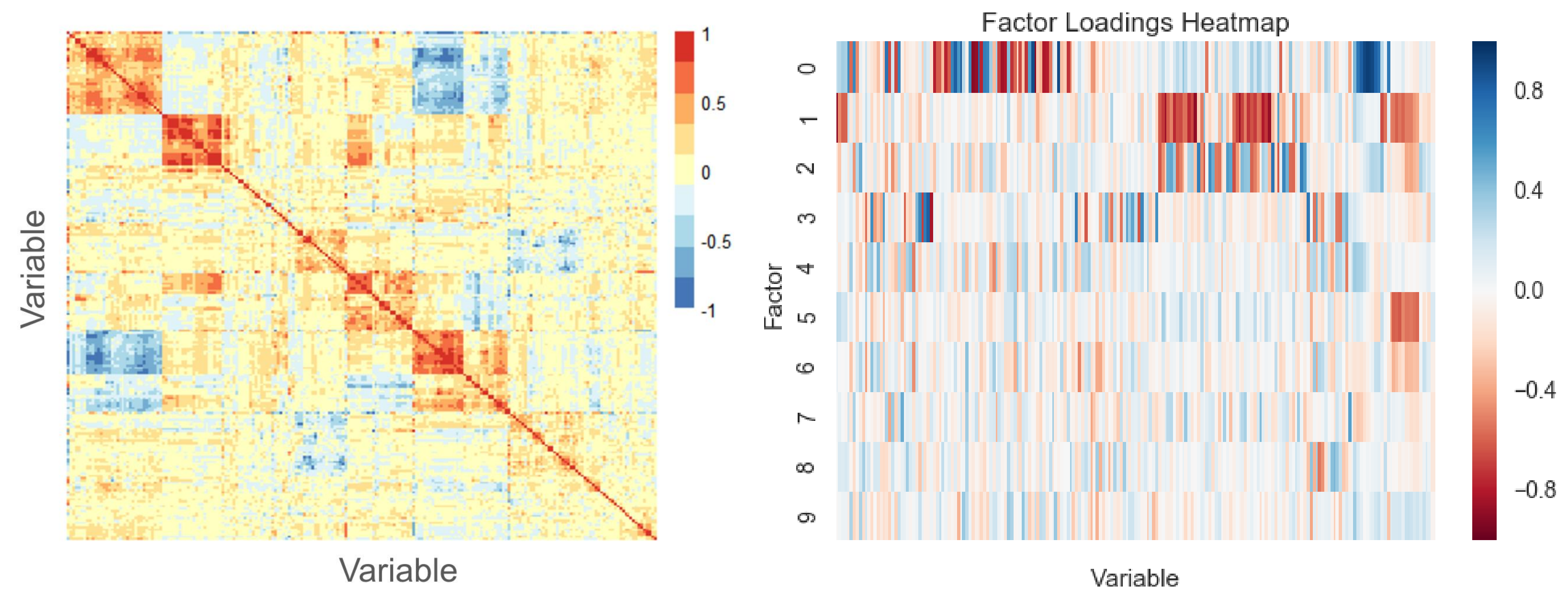
Browse

Citizenship
Crime
Development
Education
Ethnic Origins
Income
Industry
Integrated Analysis
Land Value

Now plotting: Education Bachelor's degree

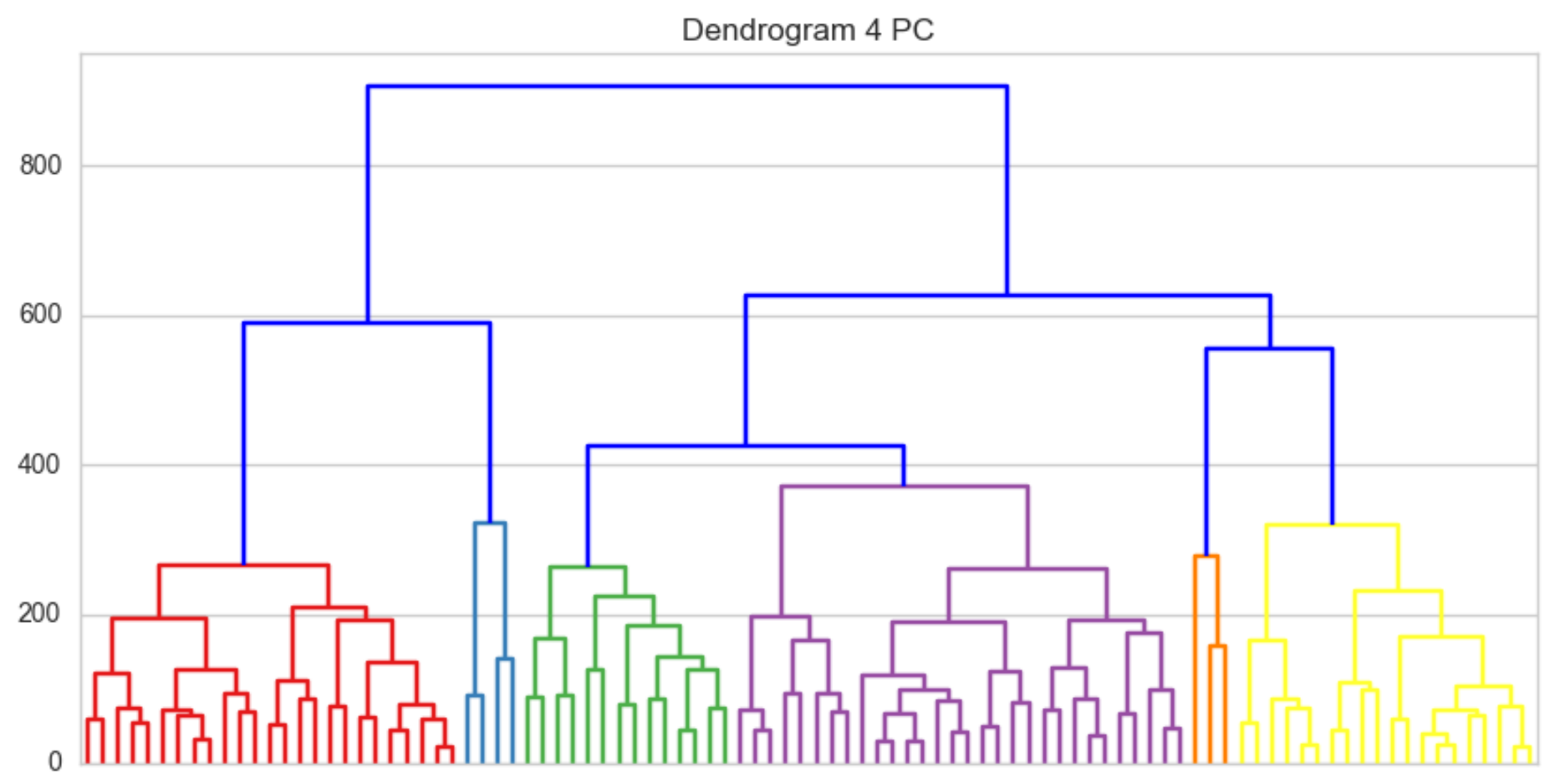
Automatic removal of redundant variables

- Many variables containing redundant information (*left*)
- Simultaneously reduce dimensionality and remove redundancy with Principle Component Analysis (PCA)
- Decided to use 4 PCs as PC5's loadings appeared to contain redundant information (*right*)

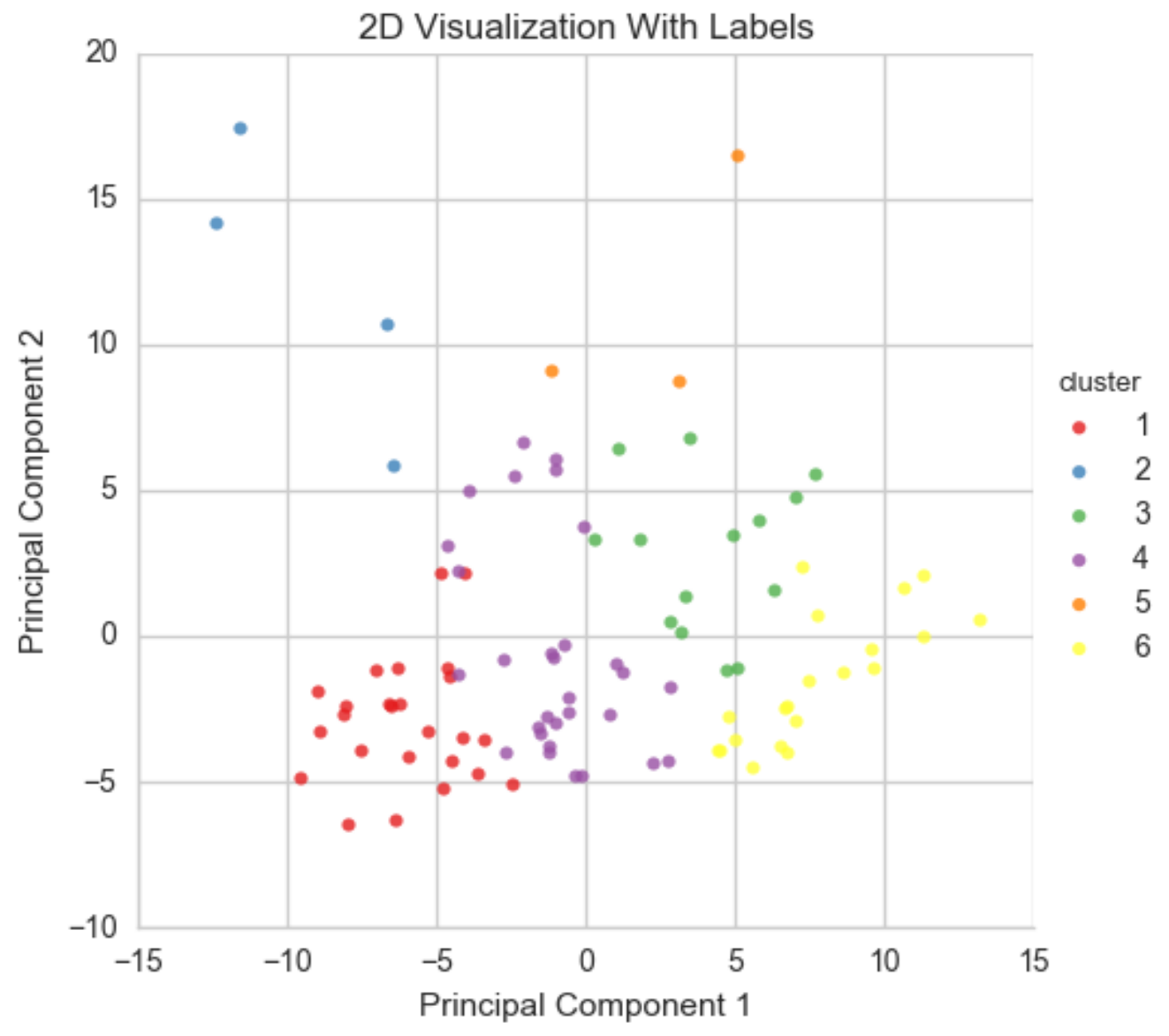


Hierarchal clustering reveals distinct clusters

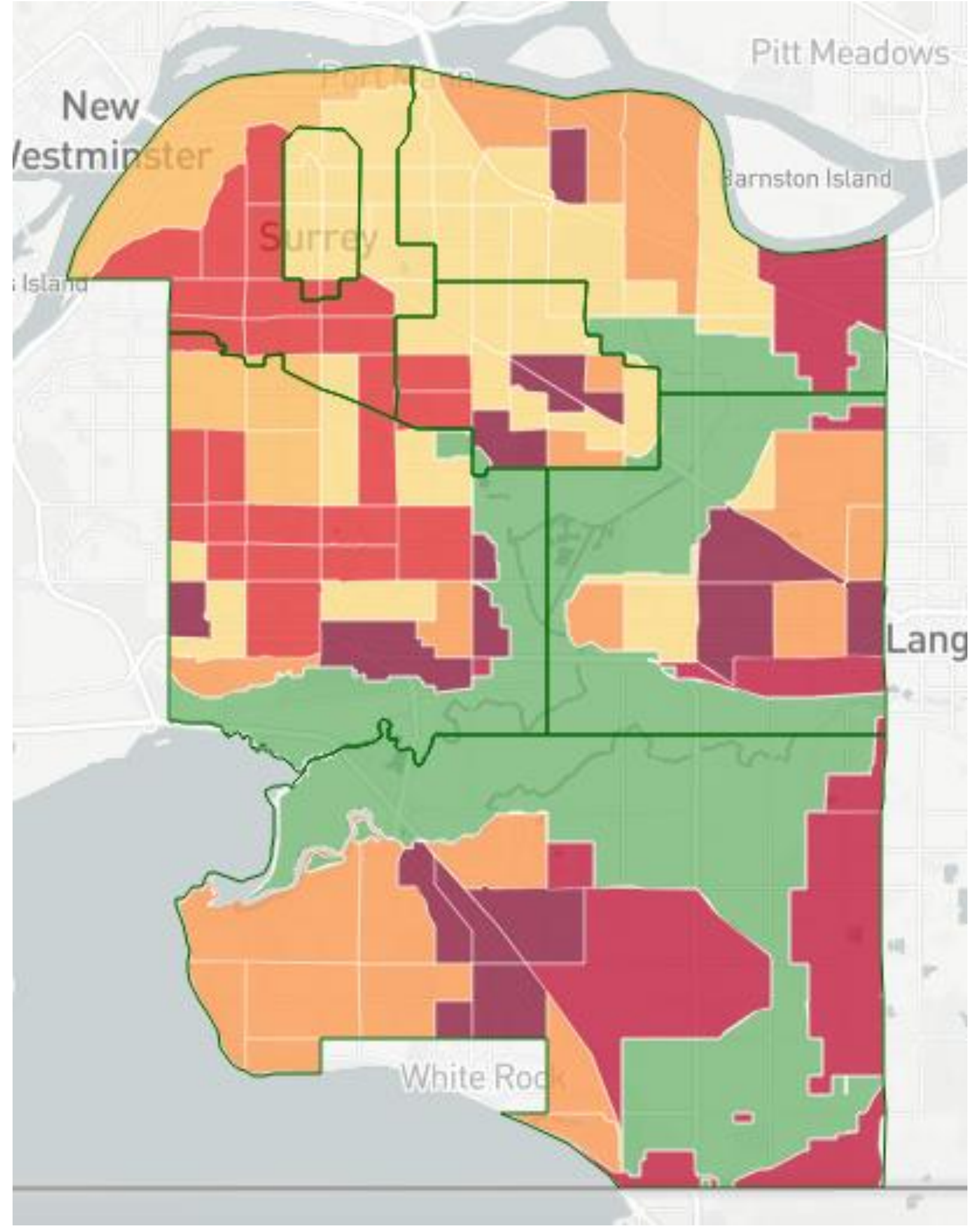
- Upon inspection of the resulting dendrogram, we decided to go with 5 clusters (as colored).



Right: Validating clustering results in two-dimensional PC space. The individual points represents individual census tracts



Towards a new data-driven vocabulary



Right: Cluster results visualized in VisualSurrey. Each color represents one cluster. Clusters have a tendency to be grouped together in geographical space. This enables a new vocabulary to describe the different neighborhoods in the City

Right: High-level descriptions of the clusters

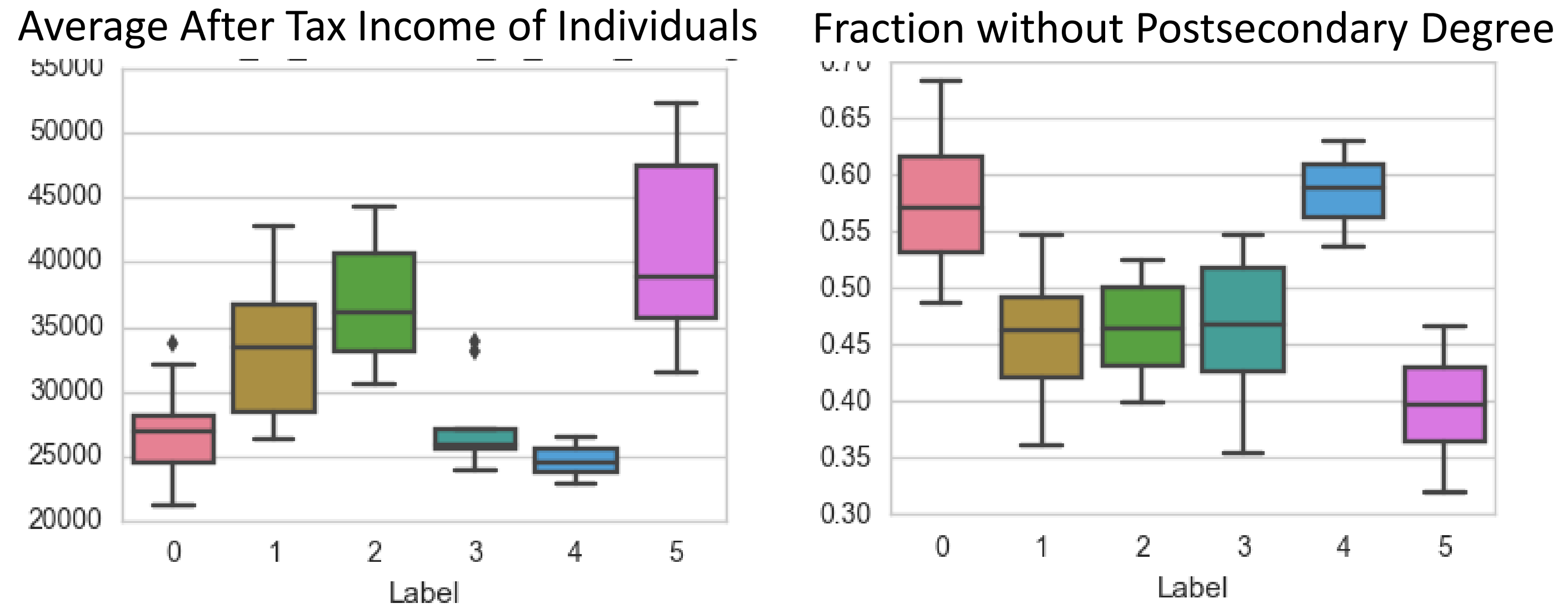
Below: Example boxplots of some key variables defining each cluster of census tracts

Residential - few businesses

- Cluster 0: Residential 1
  - low income
  - low education
- Cluster 1: Residential 2
  - high income
- Cluster 5: Residential 3
  - high income
  - high education

Business - many businesses

- Cluster 2: Business 1
  - high income
  - "professional service" businesses
- Cluster 3: Business 2
  - low income
  - retail, food, etc. businesses
- Cluster 4: Business 3
  - low income
  - manufacturing businesses



Conclusions and Future Directions

- Built a visualization tool to visualize arbitrary =
- Clustered census tracts in high dimensional space
- Interpreted the defining characteristics for each cluster

- More data to add to VisualSurrey
- More detailed descriptions of census tract clusters

UNIVERSITY of WASHINGTON

THE UNIVERSITY OF BRITISH COLUMBIA

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UBC Data Science Institute