New York University

Center of Urban Science + Progress

Capstone Project

Smart Monitor for Accelerating Regional Transformation (SMART)

Author:
Jianqi Tang (jt2900)
Ram Sowmya Narayanan (rsn293)
Zehui Xiang (zx742)
Yanyan Xu (yx2193)
Zheyuan Zhang (zz2498)

Sponsor:
US Ignite
Mentor:
Dr. Martina Balestra

July 21, 2020

Abstract

Many cities have identified gaps between big data, Internet-Of-Things (IOT) and the potential of using it for economic development decisions. Unlike multinational corporations, small businesses lack the capacity, resources, and budget to conduct research and compare the pros and cons of locations. The purpose of this capstone is to develop a Decision Support System (a recommendation system), including a smart framework and an online interactive tool, for small businesses, using the City of Portland, Oregon to test our product. The system will highlight, and rank locations based on their economic and demographic information to help small business owners make better decisions using a scoring mechanism similar to that used by multinational corporations. We expect this platform to empower small businesses in their decision-making process and city officials to get an idea of how well their city is supporting small businesses.

Contents

1 INTRODUCTION

Small businesses have difficulty in choosing suitable locations and evaluating their economic development plan due to lack of resources and budgets. In this case, an economic development plan is needed to bridge the information gap between small businesses and the market. US Ignite and partners have been working on a new economic development research framework, Smart Monitor for Accelerating Regional Transformation (SMART). This has provided a great opportunity to enable local policies to support smart development decisions and the City of Portland has been identified as a start point.

Working with US Ignite, our team has created a decision support platform which provides an integrated perspective through a smart visualization tool and a recommendation system. Based on multiple open datasets including sensor data, federal data, transit data and user review data from Portland, Oregon, we used different models and techniques (e.g. random forest, sentiment analysis and clustering techniques) to analyze current small business conditions and predict the suitable areas based on user NAICS code selection. In addition to this, an interactive map containing all the features will be provided to users to get and filter results in real-time. In this case, the recommendation system will help small business owners to understand areas of strong performance and areas for improvement, consider and evaluate different metrics in the suitable neighborhoods, by providing them multiple choices to achieve smarter decision making progress.

2 PROBLEM STATEMENT

The research question is, can we provide information that the small business owners need in order to help them make better site selection decisions. This project aims at helping small business owners decide where (at the granularity of GeoID) their busi-

ness locations could be, based on data introduced in this article. In order to achieve this purpose, we have developed a decision support model (a recommendation system) that aggregates information and highlights a range of potential location choices when small business owners select their business type, using the output of the scoring method as the metric. Ideally, this model will help small business owners to think through these parameters and present aggregated data in an easily understandable way through an interactive map and related graphs.

Hypothesis: If small businesses had access to a free, accessible open source tool that uses data we provide, they could make better decisions on choosing the right location for their business.

3 LITERATURE REVIEW

Statistically, 80% of all data of a business is location data or containing location components. A better understanding of location can provide insights and improve decision making from operations to marketing and supply chain (Bowes 2015). Companies which need a lot of logistics may place themselves near places with strong transportation capacities, including truck networks and deep-water ports. Small companies which produce knowledge based products may have more freedom to choose a location based on their own need. Knowledge-based, fast growing companies which need to be surrounded by other companies in their industry should place themselves in technology centers (Gordan 2017). For all businesses, the most important factors in influencing business site selection decisions are availability of skilled labor and transportation facilities (Karakaya and Canel 1998). Stofan introduces the method to analyze a business's location by exploring foot traffic with actionable insights from standard street cameras. Measuring foot traffic gives the business an estimate of how many customers a potential location can bring, so as to choose an ideal spot (Sto-

fan 2018). Thus, in our research on businesses of Portland, traffic information and industry types are considered as metrics in the recommendation system.

4 DATA

4.1 HUD (Department of Housing and Urban Development) USER API

This federal dataset provides basic identification information like year, Zip, GEO ID and so on, essentially the boundary of Portland. This is joined with the CBP data to add the visual component to the information.

More Information on how to use the API at https://data.hud.gov/section3.html

4.2 CBP(County Business Patterns) DATA

This contains data for establishments from each year and presented by geographic area, 2-through 6-digits NAICS industry, legal form of organization and employment size class. Information is available on the number of establishments, employment, first quarter payroll, and annual payroll.

API can be called using https://api.census.gov/data/

4.3 CARTO DATA OBSERVATORY

With the help of our Sponsor, US Ignite, we have access to Carto's datasets for data, spatial analyses and other premium features. We also used it to get the ratio of ages of the population.

More info on how to use the API at https://carto.com/developers/cartoframes/

4.4 ArcGIS OpenData

From the ArcGIS Open Data portal, we use Portland sensors location data to provide the boundary box feature for CityIQ credentials.

Data source: https://gis-pdx.opendata.arcgis.com/datasets/

4.5 CityIQ

CityIQ is a repository of sensor data from the various cities. We use this to get pedestrian count and vehicle count in a month time period at the granularity of Geo-ID level in order to capture how busy the traffic is. It also plays a role in determining the score to rank the Geo-IDs.

More info on how to use the API at https://github.com/CityIQ

4.6 American Community Survey (ACS)

ACS is used for granular demographic information such as household income, race and ethnicity, age by sex, and other such factors. This information on the Geo-ID level provides us more insight into the characteristics of the location.

Data source at https://www.census.gov/programs-surveys/acs/data.html

4.7 Public Review System

To get an idea about how various businesses are doing in Portland, we used the public review portals Yelp and Google Review. The API provides us with a list of features used as the input for the NLP model to get the sentiment score and review rating for each business type in all the Geo-IDs which is merged with the ACS data to arrive at the final dataset.

Data source: 1. https://www.yelp.com/developers/documentation/v3

2. https://developers.google.com/places/web-service/

- **Data Limitation**

Only a small part of the city is covered with CityIQ sensors, and sensors are not evenly distributed in the covered area, which leads to potential bias in the estimation of traffic data. Though there is five years data in the carto data observatory but the most recent data is in 2017. The thing could change violently since then. It will be much better if the data could be real-time.

5 METHODS

5.1 Data Preparation and Engineering

5.1.1 CityIQ and Cartoframe

We queried data from CityIQ and cartoframes observatory, cleaned the data and aggregated them on an hourly basis. For CityIQ, we wrote scripts to extract the location to set up the boundary box of the smart sensors. This way, we fetched event data produced by smart sensors inside the boundary. Finally, a pipeline that connects to the API and collects formatted data was built.

5.1.2 Autocensus API

We used the Socrata library to get the Census data about Portland for the years 2010 to 2017. Then, we fetched the maximum NAICS code for every zip code in Portland for those years using the CBP data and merged them on Zip Code and year. Once this was done, we had the final dataset on which machine learning models are implemented. Finally, we use this dataset to make a visualization to check if all the zipcodes are present with the appropriate values.

5.2 Exploratory Data Analysis (EDA)

5.2.1 Correlation Analysis

We apply correlation analysis on economic factors like the GINI Index, employment ratio with other metrics like population density, public infrastructure level, criminal risk, education level etc. to get correlations of the metrics in our system. As the Sponsor emphasized, we created a correlation matrix that compares factors of employment such as ratio of

the employed population, the ratio of skilled employees, ratio of population out of the workforce, etc. against economic factors such as the GINI Index, spending potential, etc. This way, we derived insights on how well neighborhood functions and what are the types of positions that a neighborhood supports.

5.2.2 Clustering

In this process, we cluster all the business types according to the information like business types, employee size etc,. We perform this to understand the data and identify any potential clusters. We identify the average characteristics of all the items in that cluster when they have been split into discrete clusters. After obtaining the standard characteristics for a given business type, we highlight potential spots on the map that satisfied those conditions. We have identified Gaussian Mixture as the best method that highlights underlying clusters.

5.3 Recommendation System

5.3.1 Supervised Learning

We used decision trees, random forest, support vector machines to train models to predict the Maximum NAICS code of establishments for each neighborhood of Portland. We checked the feature importance of each variable in our dataset which is helpful for us to build the dashboard of the recommendation system. The target is the maximum NAICS code of all industries in the corresponding US Census tract and year. We also used Logistic Regression, Ada Boost, XGBoost, and Naive Bayes to train models to predict NAICS code of each neighborhood. Of all 7 methods, random forest had the highest accuracy in prediction. Therefore, we decided to use

the labels predicted by random forest on our final computational model.

5.3.2 Sentiment Analysis

We use the public user review system (Yelp and Google review) to obtain categories, ratings and review texts of small businesses in each GeoID in the City of Portland. Based on the review datasets, we analyzed the distributions, ratings scores and words frequency for different types of GeoID. Then, we fine-tuned the BERT model and applied LDA modeling for sentiment scores, attitude aspects analysis and word frequency analysis among different categories. Finally, we merged the sentiment scores and ratings with NAICS codes.

5.3.3 Implementation

To solve our problem statement, we came up with an approach that involved aggregating output from multiple sources and developing a scoring mechanism to convey the result. First, the data source is created. Second, multiple clustering techniques are applied to see if the data has any inherent clusters and if so, how many are there. From this, a Gaussian Mixture with 4 centroids was identified to describe the data most aptly. Third, multiple supervised learning techniques are applied on the dataset to see which technique provided us with highest accuracy against the target column of Maximum NAICS code for the Geo-ID. We arrived at Random Forest as the model to use. Then, the data from 2012 to 2016 was taken as the train dataset and the 2017 dataset was taken as the test dataset. A new column is created on the 2017 test dataset that contains the predicted NAICS code from the supervised model. Finally, a scoring model is developed with all the available data and the locations are ranked with

the highest scores as the most preferable.

If the actual and predicted NAICS code matches a score of 10 is given, if not, a 5 is given. The pedestrian and vehicle counts are scaled to 1 to 5 using a Min-Max Scaler to arrive at a score. A major hurdle we faced here was that the sensors for the pedestrian and vehicles were not uniformly spread across the Portland Metro Area, so the counts had to be assumed for non-present Geo-IDs as suggested by our Sponsor. Then the sentiment and ratings from the NLP model are merged to the dataframe to get a score for each Geo-ID. The sentiment score is scaled from -2.5 to 2.5 and rating from 1 to 5. The NaNs are filled with zero and the median respectively. Finally, we chose the Establishment Ratio ("estabratio") as an explicit feature to include in our scoring process since the density of businesses gives us a good idea about the Geo-ID. Once all the columns have been appropriately scaled, we computed the final score for each Geo-ID by performing a linear summation of all the scores and we rank the Geo-IDs which are split by the NAICS code based on the score.

$$final\_score = NAICS\_score + pedestrian\_score\_scaled + vehicle\_score\_scaled + sentiment\_score + ratings + estabratio$$

5.3 Data Visualization

We used Carto Dashboard for the data visualization and implemented more advanced features that were provided through the relationship our Sponsor had maintained with Carto. We uploaded the final Dataset into the dashboard and created a map based on the GeoID, Final Score, NAICS code, etc., which was then coded with different colors. We proceeded to add widgets to the map which

will allow the users to choose the business types and also select ranges of features from histograms. A tool-tip was integrated to show detailed information such as Population Score, Income Score, etc. Since all information is summarized and displayed through the interactive map, business owners can interact with different metrics and analyze locations in a more intuitive manner. During the process, we also tried other visualization tools like "MapBox", "Tableau", but found Carto has better interactivity and flexibility features.

6 RESULTS

6.1 Establishment Size Spatial Analysis

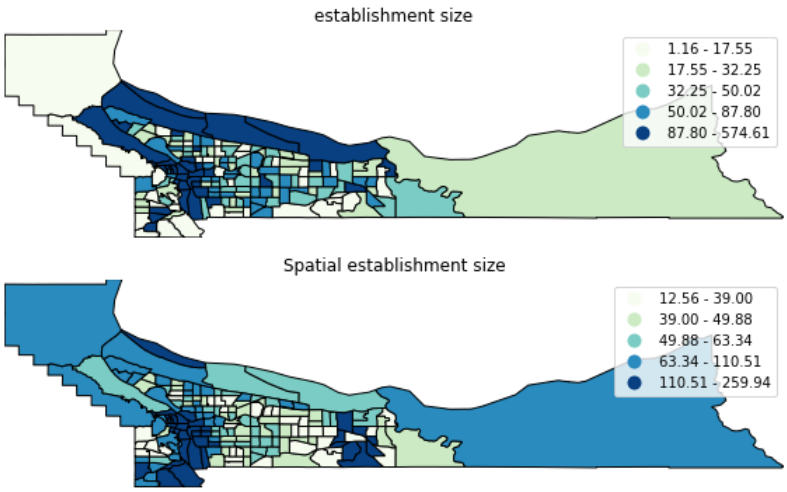

Figure 1: Spatial Distribution of establishment size

We used the final dataset to measure how various features influence the establishment ratio in each area as part of the EDA. We check the distribution of features normalized by population. We used the PySAL package to quantify the geographic information, getting the adjacent weights and lag spatial weights of features from the GeoDataframe. Finally, we map the established size to five classes and compare the difference between weighted establishment size and the original one.

6.2 User Reviews Spatial Analysis

We used user review datasets to cluster and aggregate the distribution of different types of small businesses, average rating, high rating (score ≥ 4.5) distribution, review counts, sentiment scores distribution (Fig 2) within each geo id in Portland (Append. I). Also, we calculate sentiment scores among different types in different GeoID (Fig 3). From these spatial analyses, we derive an understanding of current small business conditions and feedback. We have also used a version of these maps on the final product to support our recommendations.

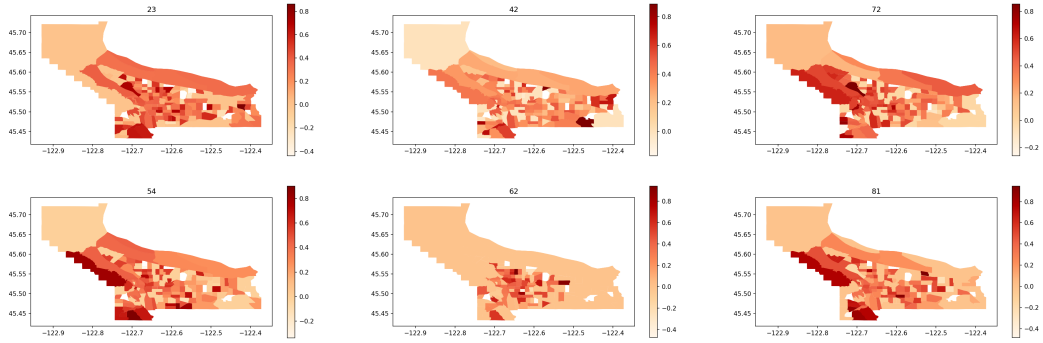

Figure 2: Distribution of different code of sentiment scores

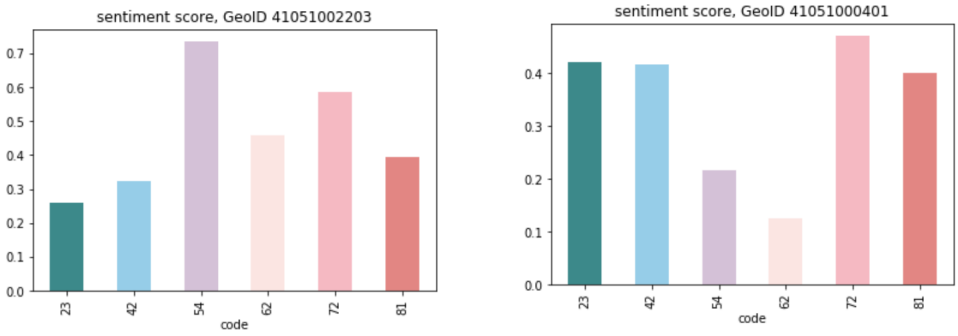

Figure 3: Sentiment scores of NAICS codes in different GeoID

6.3 Interactive map

From the interactive map, users can choose one or more business types from six categories (derived from the dataset) on the right side to see the final score on each GeoID within the selections. Users can also select the range of final score, population and income on the histograms to obtain more information on the specific range. When clicking the button of "auto style", the map will be visualized by luminosity. When the mouse hovers over any area on the map, it will show a tool-tip which has geo_id, final score, ranking, etc. Through this interaction, users will acquire the recommended areas, rank scores and demographic information which will help their decision on the suitable locations.

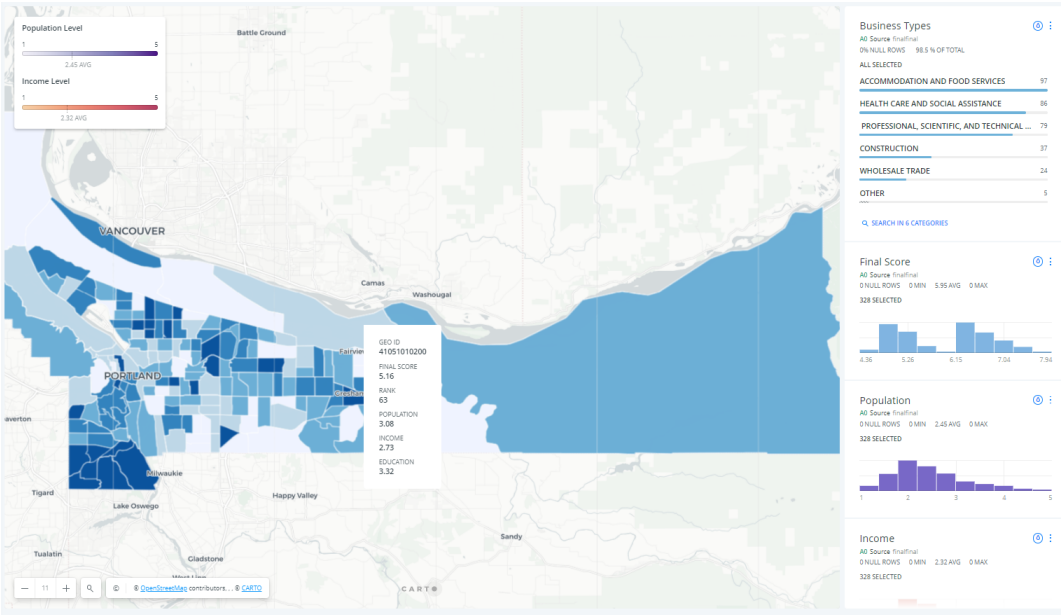

Figure 4: Interactive map of final score in each GeoID

7 CONCLUSION AND FUTURE WORK

Decision Support Systems for locations is a helpful tool in the hands of small-business owners since it gives an opportunity to access information on the level of multinational companies. Organizations have long resorted to scoring mechanisms to

quantitatively compare against locations in viable markets. In this project, we have aggregated data from multiple sources to come up with a score that ranks Geo-IDs based on said score, similar to that used by large organizations. Although we have tried to emulate the process done by larger organizations, we have also tried to tailor the approach to suit the needs of small-businesses. It has to be pointed out that these results are not final and error-free, with further research and effort this product can become a reliable go-to option.

For future work, it is promising that the research could further be carried on from different perspectives. We have been doing analyses mostly as a data and location analyst for the small business owners in this project. Using the results we have at this stage, this project could be much more comprehensive and practical from the perspective of community leaders. From this perspective, one could further analyze if the data could help community leaders get information they need to attract new small businesses, retain existing ones and track the economic impact of employment trends across targeted community neighborhoods.

8 TEAM ROLES

- **Jianqi Tang**: Spatial and Temporal Analyses; Data Visualization and Web UI
- **Ram Sowmya Narayanan**: Data Cleaning, Preprocessing and Enrichment; Data Science
- **Zehui Xiang**: Data Modelling; Supervised Learning and Data Science
- **Yanyan Xu**: NLP and Sentiment Analysis ;Urban Modelling and User Review Analysis
- **Zheyuan Zhang**: Data Engineering; Version Control and Codebase Maintenance

9 LINKS

- GitHub: https://github.com/JasonZhangzy1757/CUSP-Capstone
- Website: https://smartcapstone.wixsite.com/nyucusp

REFERENCES

Bowes, Pitney. Forbes, Forbes Magazine, Mar. 2015, www.forbes.com/forbesinsights/location_intelligence/index.html.

Gordan, Kayla M. "Business Site Selection, Location Analysis, and GIS." Arcada University of Applied Science, 2017.

Karakaya, Fahri, and Cem Canel. "Underlying Dimensions of Business Location Decisions.' Industrial Management & Data Systems, vol. 98, no. 7, 1998, pp. 321329., doi:10.1108/02635579810205395.

Stofan, Daniel. "Location Analytics: Exploring Foot Traffic at Your New Business Location with GoodVision." Medium, GoodVision, 22 Feb. 2020, medium.com/goodvision/goodvision-location-analytics-exploring-foot-traffic-at-your-new-business-location - 805b8919b0ea.

Acknowledgements

US Ignite is an effort originated by the Office of Science and Technology Policy (OSTP), led by the National Science Foundation (NSF), to promote the development and deployment of next-generation applications with the potential for significant societal impact.

Thanks to Our POC in US Ignite, Mr. Praveen Ashok. He helped us understand the landscape and technicalities of the project. Mr. Praveen Ashok got in touch with several community leaders from the City of Portland to gain an insight into what small businesses look for in choosing their locations. The City of Portland was chosen as the start point of this project since it is not as populous as other cities but still is comparable in features and operations to the larger cities. This decision was made with the immense help of Mr. Praveen Ashok.

We would also like to thank Dr. Martina Balestra, our mentor, for guiding us throughout the course of our project and giving invaluable insights to develop our project to this stage.

Appendix I: User Review Analysis

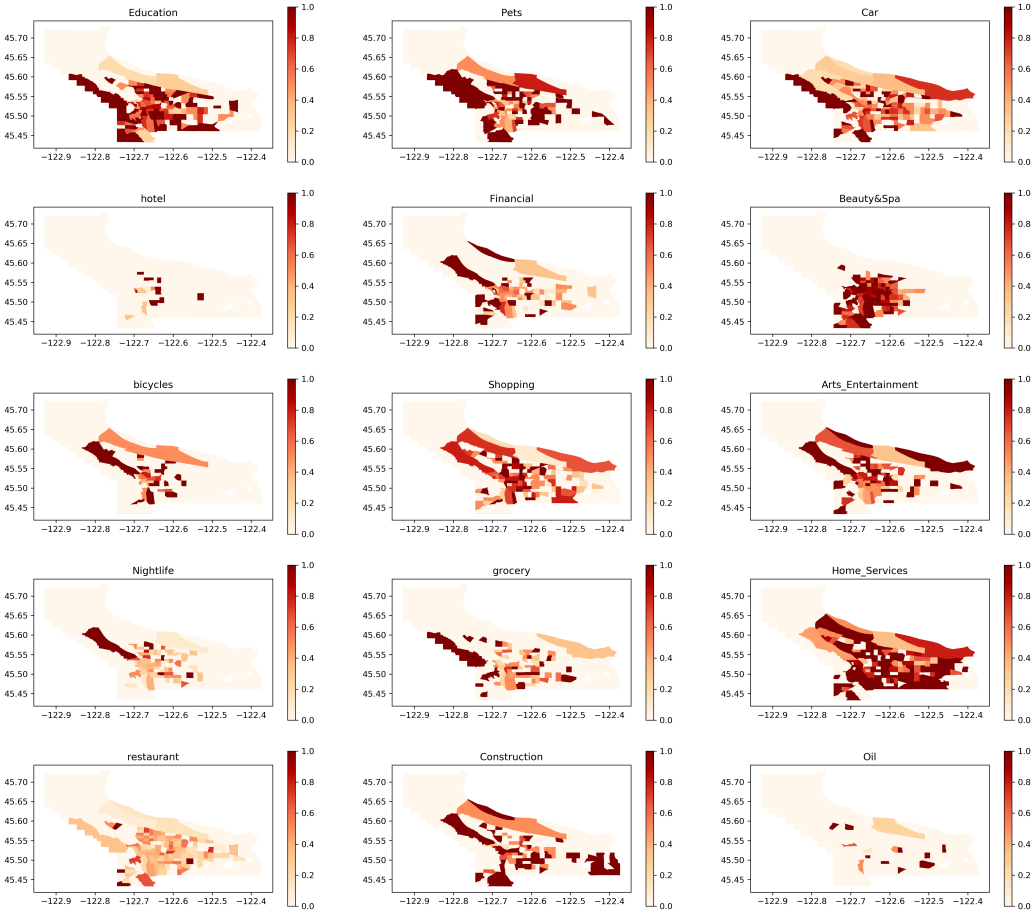

Figure 5: High Review Rating Ratio

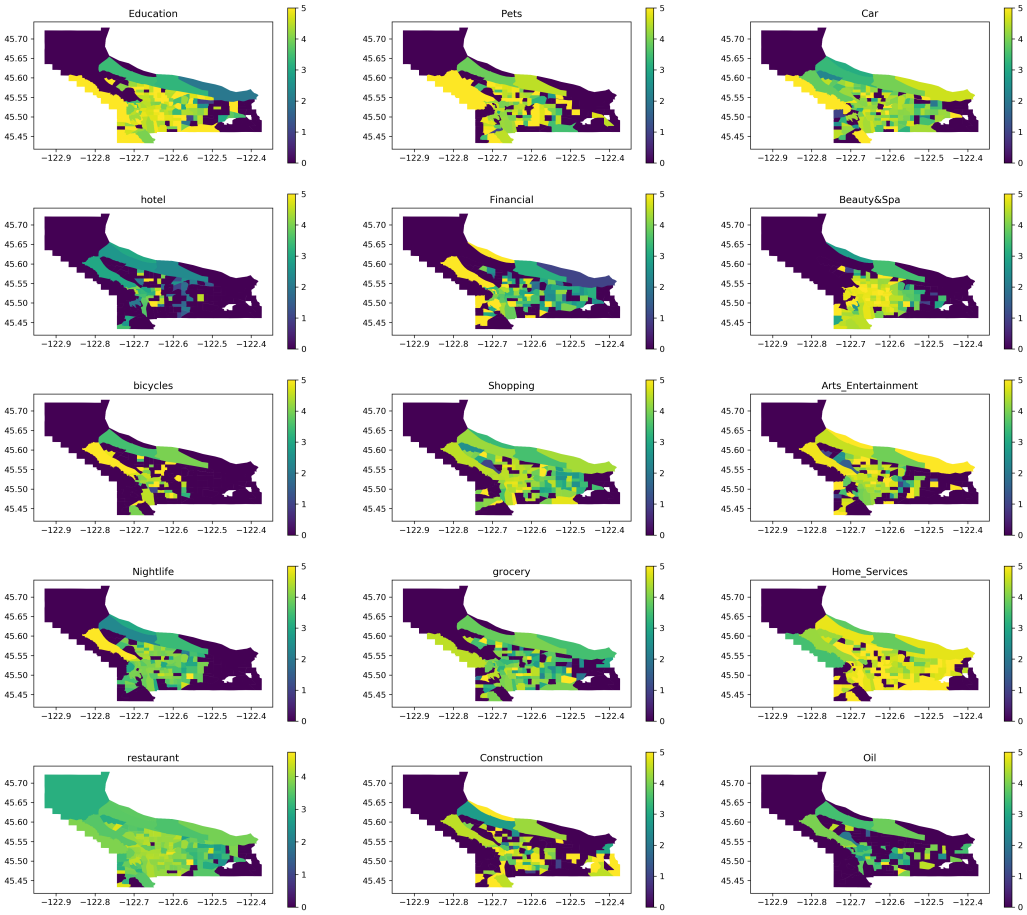

Figure 6: Average Rating of Different Types of Small Businesses

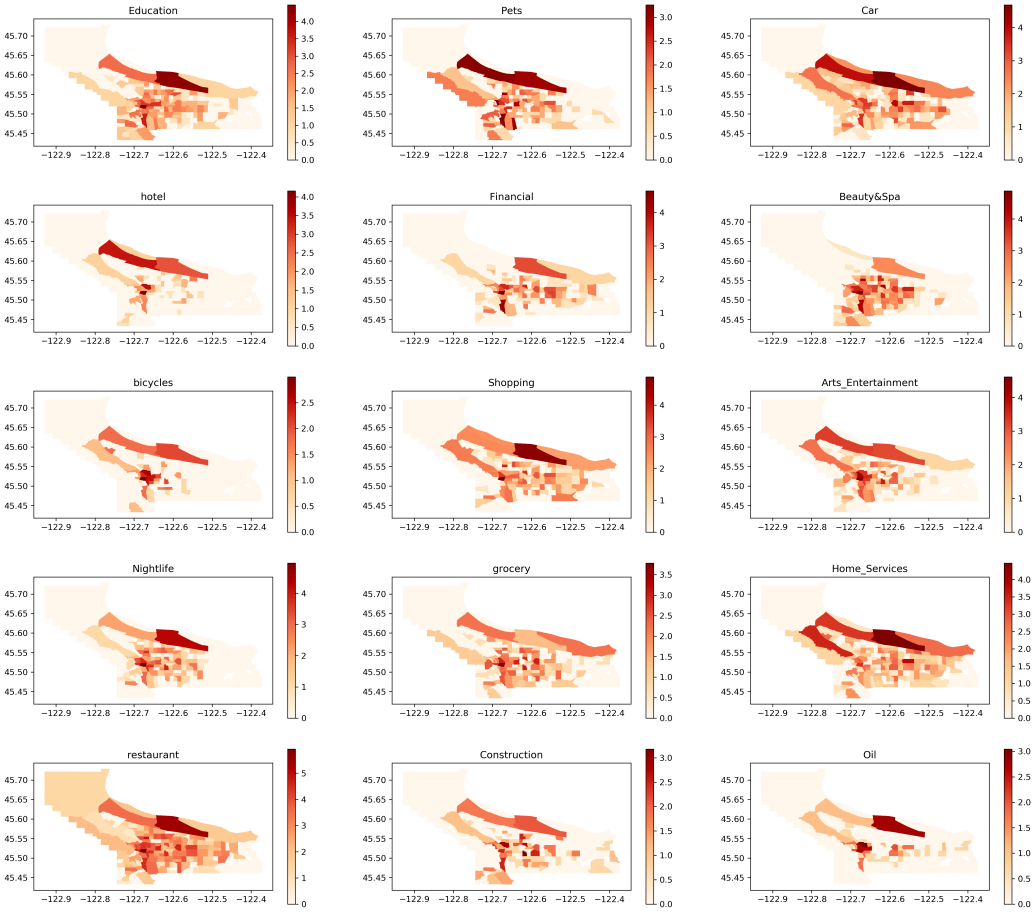

Figure 7: Distribution of Different Types of Small Businesses

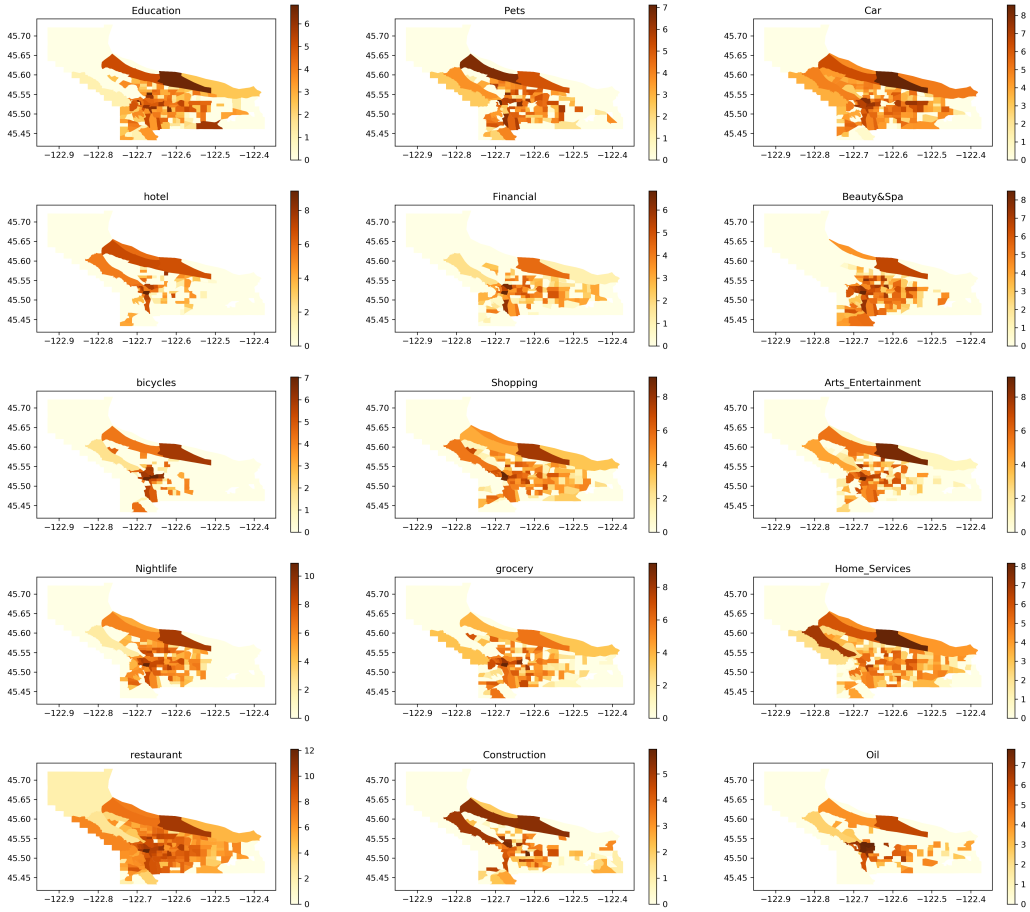

Figure 8: Review Counts Distribution of Different Types of Small Businesses

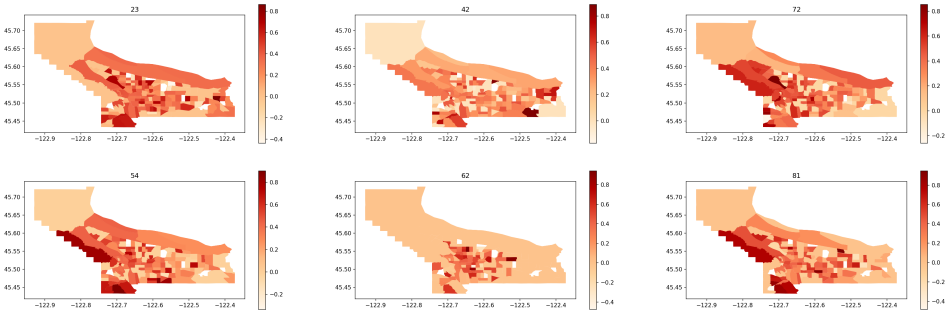

Figure 9: Distribution of Different Code of Sentiment Scores

Appendix II: Map with Prominent NAICS code

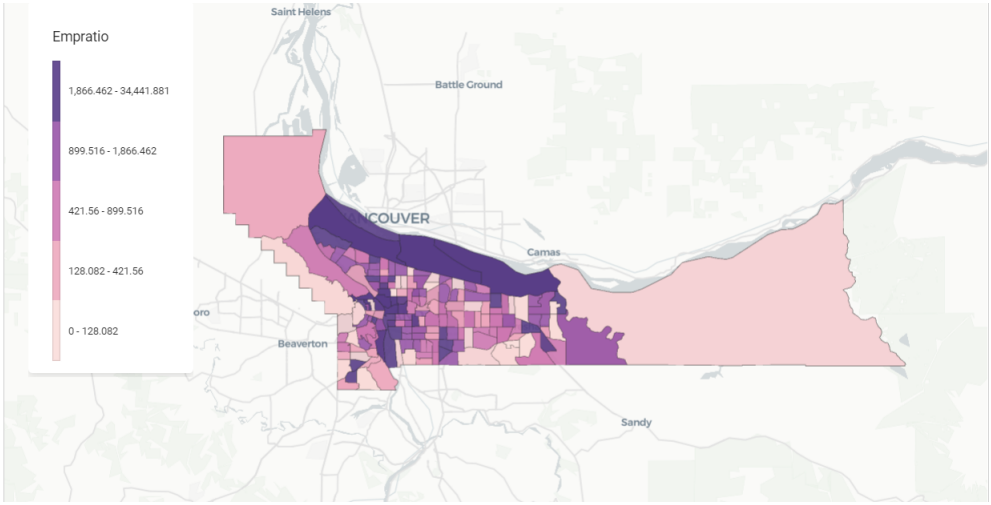

Figure 10: Employment Size With Prominent NAICS Code For Each Geo ID

Appendix III: Federal dataset EDA & Spatial analysis

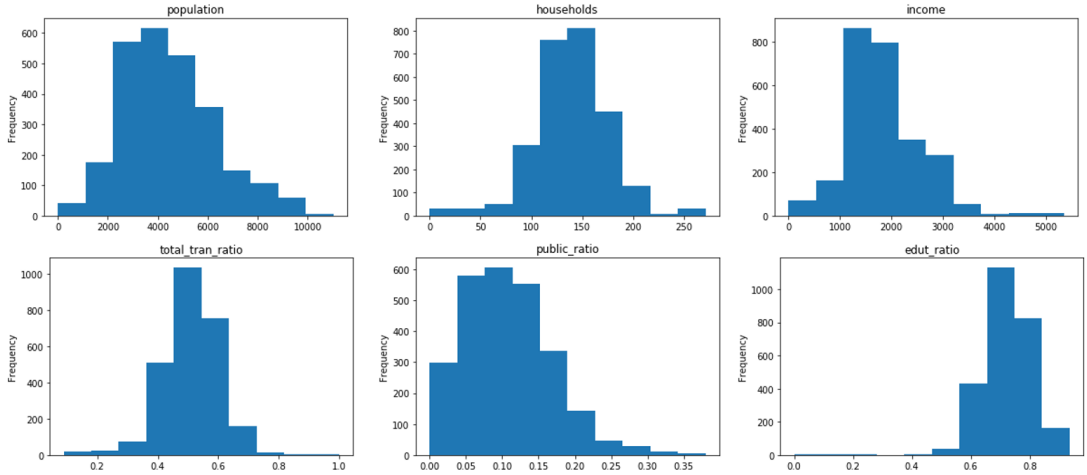

Figure 11: Frequency of Features

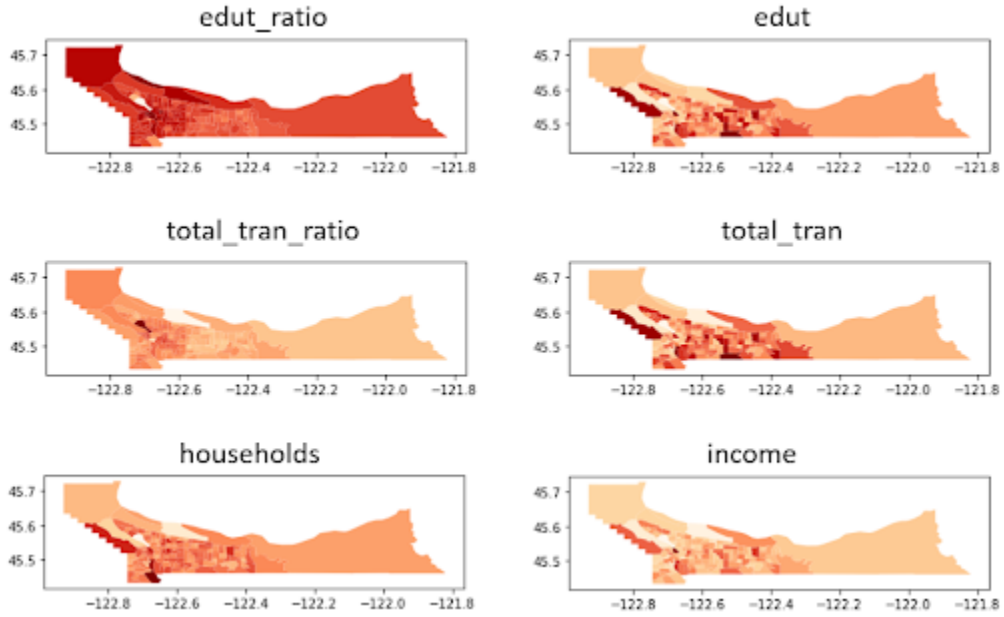

Figure 12: Distribution of Features

Appendix IV: Decision Trees Visualization

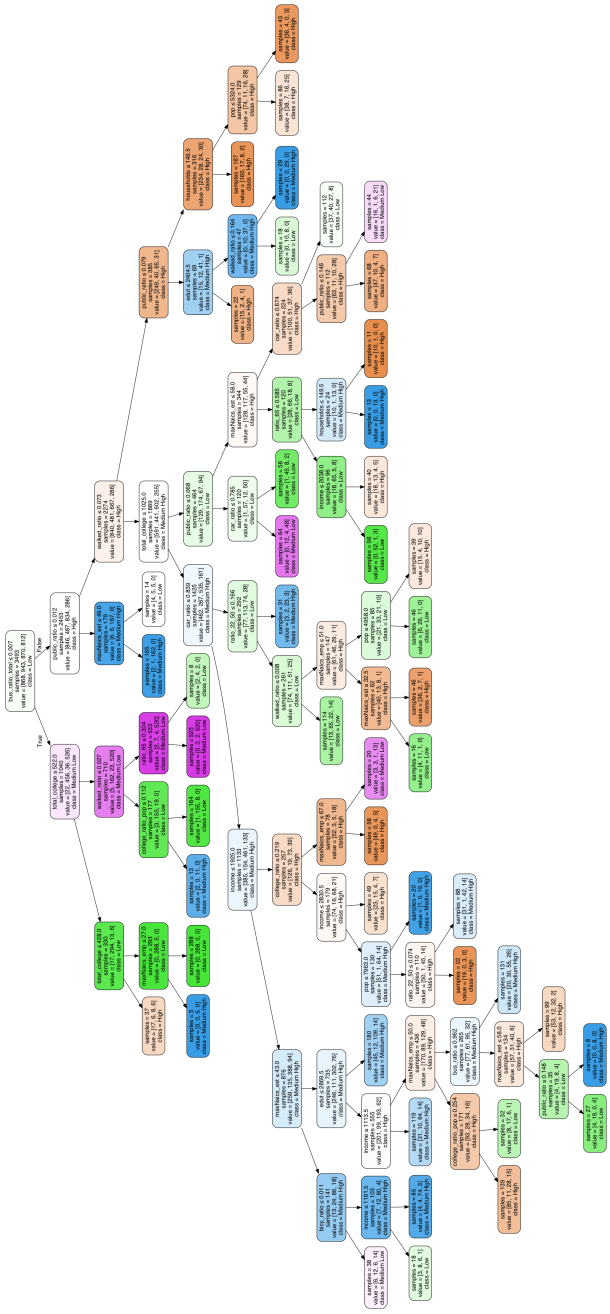

Figure 13: Decision Tree for Employment Size

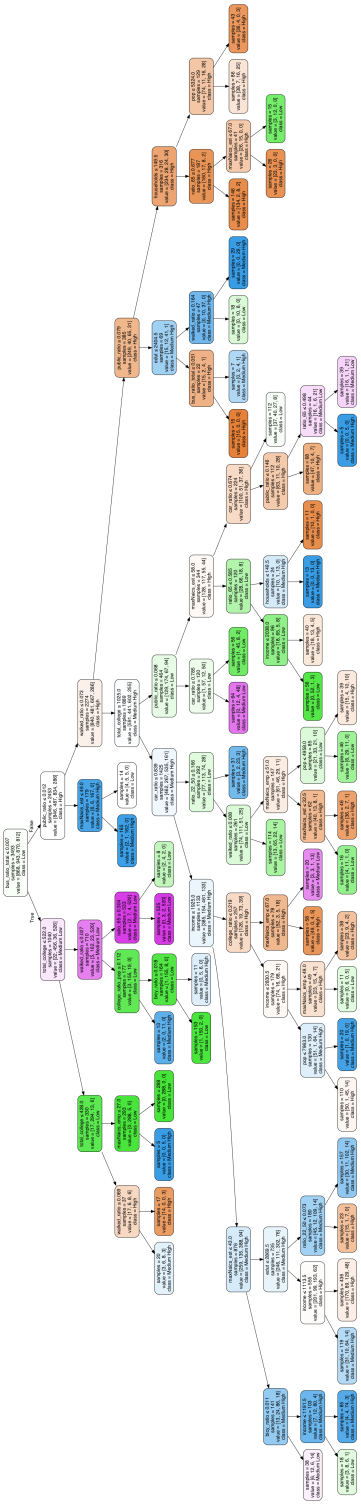

Figure 14: Decision Tree for Establishment Size

Appendix V: Geographically Weighted Regression

Using the same dataset of supervised learning model, we also did geographically weighted regression to calculate the local R squared values. In a GWR model, local R squared values represent the relationship between our dependent variables and independent variables, which fluctuates through the neighborhood of Portland.

We have local R squared values distribution graphs for each independent variable: employment size (Fig (a)) and establishment size (Fig (b)). Each neighborhood of Portland has its own R squared value which evaluates its regression relationship between dependent variables and independent variables.

(a) Local R squared for employment size

(b) Local R squared for establishment size

Appendix VI: Other Insights

It is important that we understand what the most prominent industry is in a respective geographical ID to make recommendations. In the supervised learning models, we find that the most important features of both establishment and employment sizes are the differences of employment size and establishment size from last year at the same census tract. Besides this, we also find that these variables matter: the ratio of bus usage by the total transportation, ratio of people who walked to work by total transportation usage, college population, and the ratio of

public transportation by total transportation.