Russel Lutke: 301169861
Sean Yiang: 301364025

CMPT 353 Final Project Report

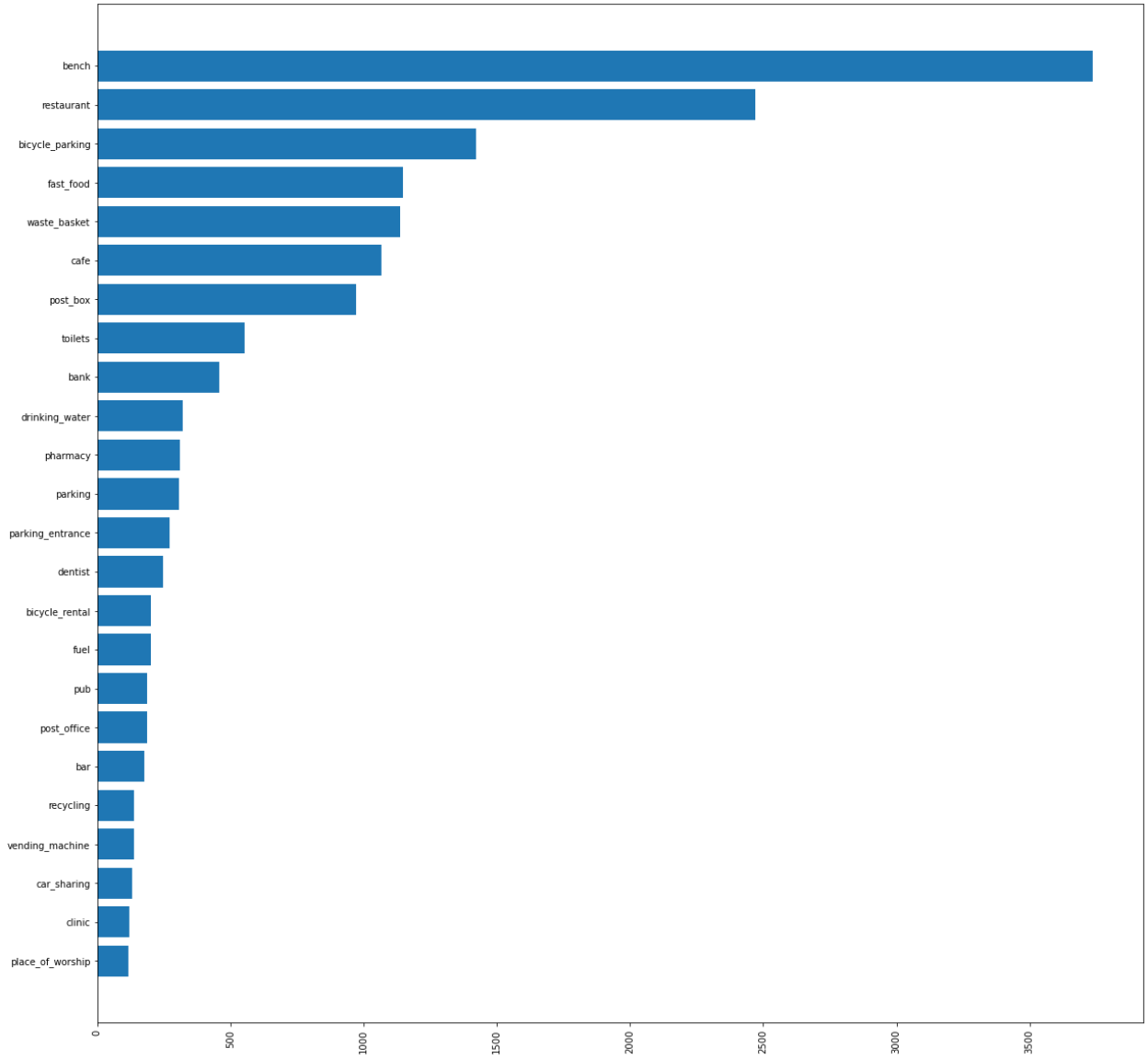
Introduction

For our project, we decided to use the provided OSM data in conjunction with Vancouver census data to try and predict features or properties of various neighbourhood districts. The census data consisted of various income and population statistics, conveniently broken down into each of the 22 official neighborhoods of Vancouver. After dividing the original OSM data amenities into larger categories (e.g. “financial”, “education”, “community”), we were able to generate scores for each of these categories for each of the 22 neighbourhoods by combining the coordinate data from the OSM dataset with the geolocation boundary data for each neighbourhood. This yielded a dataset of 22 points, each with the scores or ratings of each amenity category along with some feature of the census data for each neighbourhood. We then fitted a linear regression model to this dataset, attempting to predict the chosen property of the neighbourhood using some subset of the amenity scores as variables/features. We also constructed a function that takes an arbitrary coordinate point and a radius and computes a set of scores for each category within a circle of the given radius, with higher scores indicating a larger amount of more densely packed amenities of the given category.

Data Cleaning

When loading in the provided ‘amenities-vancouver.json.gz’ dataset into our jupyter notebook, we first dropped the tags and timestamp columns since we decided that our topic of interest is unrelated to the wikipedia tags nor timestamp. The dataset now contains the columns: ‘lat’, ‘lon’, ‘amenity’, and ‘name’. Upon looking at the data, we noticed that there are different types of amenity, and the range of numbers of different types of amenity is large. And there are many similar types of amenities, such as ‘parking’ and ‘parking entrance’.

Russel Lutke: 301169861
Sean Yiang: 301364025



The above plot showcases the distribution of types of amenity based on their numbers. Note that this plot only shows amenities that have more than 100 counts. Based on the graph, we also found out that most of the entries in the datasets are benches, restaurants, and bicycle_parking.

We then combined similar amenity types and categorized them as a single type of amenity into dataframes. For example, we categorized all the restaurants, bistros, cafes, etc. as food and drink. Each data frame is a single type of grouped amenity. We then concat all the data frames into a single data frame.

Russel Lutke: 301169861
Sean Yiang: 301364025

Neighbourhood Boundaries

In order to categorize each amenity by their geo-locations, we need to get the boundary for Vancouver districts. We used the 'local-area-boundary.geojson' dataset available on the City of Vancouver's website. We consider the 2019 dataset to be reliable given that the dataset is published by the city of Vancouver, on the specific topic of Vancouver neighbourhoods' geographic data. The dataset contains the boundaries of Vancouver's 22 official areas as a geojson file.

First, we load the 'local-area-boundary.geojson' into the Jupyter Notebook using the geopandas package. The data contains three columns: mapid, (the abbreviation of the names of the neighbourhoods), name of the neighbourhood, and the geometry (boundaries represented by POLYGON objects) of the neighborhood.

Our objective now is to detect if an amenity is within the boundary of a certain neighbourhood. So we need to change the previous 'return_near_by_structures()' function. The changed function 'return_near_by_structures_boundary()' takes a given coordinate, a name of the neighborhood, and the dataset, and returns all the amenities as a data frame that is close to the given point, within the specified boundary. However, it is not efficient to detect if each of the amenities is within the boundary this way. Thus, we undated the function to 'point_into_neighbourhood()'. The 'point_into_neighbourhood()' function takes only the original OSM dataset for amenities and the boundary dataset. The function returns the number of total amenities in each neighbourhood.

Analysis

Our topic of interest is whether we can use the scores for amenities in each neighbourhood to predict the values in the census data. There are many variables in the census data such as income per neighbourhood, marriage status per neighbourhood, and so on. We narrow down to income before tax, and number of single person households. We want to see if the scores for amenities have a correlation with the income per neighborhood, and if so, can we use the scores to predict the income in each neighborhood.

Russel Lutke: 301169861
Sean Yiang: 301364025

In the census dataset, each column represents a neighbourhood in Vancouver; each row represents a topic variable. For the income prediction, we first selected the row " Median total income in 2015 among recipients ($)" and dropped the unneeded columns. We then transposed the row and sorted by the neighbourhood name. Note that the previous neighbourhood datasets are also sorted by the neighbourhood name, this way we are able to just append the column into the target datasets without using join() statements.

In the next step, we applied the 'neighbourhood_score()' function to a copy of the original dataset, the function returns the score for each amenity in the neighbourhood in separate columns. The score is based on the number of amenities per type in the neighbourhood divided by the total number of amenities per type in the entire dataset times a scaling constant. We then append the medium income to the result dataframe of 'neighbourhood_score()'.

Our first attempt was to fit a multiple-linear regression model predicting the medium income per neighbourhood using all the scores as features. Our second attempt was to predict the medium income using only "community" and "education" amenity types.

We also built a similar multiple linear regression model using the same predictors to predict low_income per neighbourhood.

Our second topic of interest is whether the number of single person households is correlated with types of amenities in the neighbourhoods. For this, we selected 'entertainment', 'transportation', 'shopping', and 'postal' amenity types with the assumption "individuals that live by themselves consume more entertainment goods/services, and rely more on transportation and postal services. "

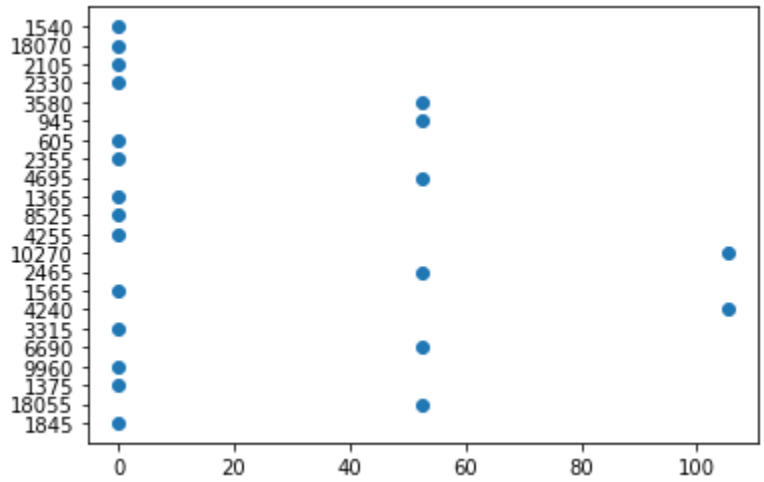
We built similar multiple linear regression models for the prediction.

Problems in the Analysis

The way that we conducted our analysis may neglect external factors. For example, we wanted to know if we can use 'transportations', 'entertainment', 'shopping', and 'postal' to predict the number of single person households. And the results from our linear regression models fail to predict the data accurately. However, despite the small sample size (only 22 entries to train_test_split()) contributes to the ineffectiveness of the model,

Russel Lutke: 301169861
Sean Yiang: 301364025

we did not consider the fact that some neighbourhoods have more entertainment related amenity due to its nature. Look at the following plot.

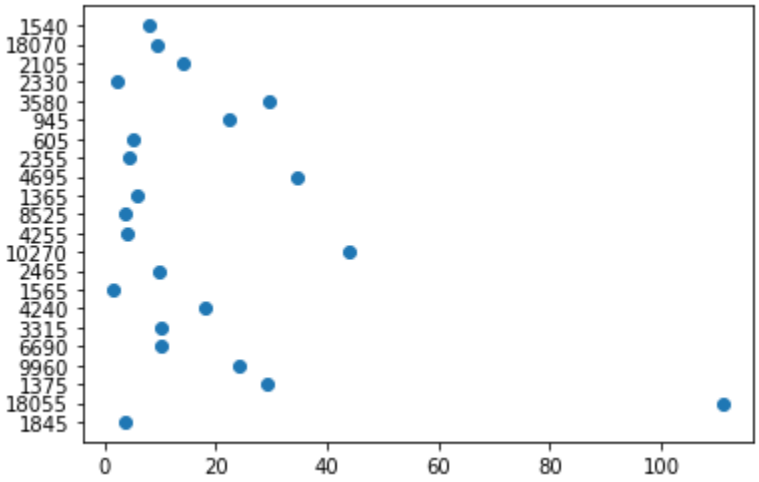


The x-axis indicates the score for shopping amenities, and the y-axis indicates the number of single person households in each neighborhood. If we compare the Sunset (second row with 18070 single person households) and Downtown (18055 single person households), it is easy to notice that the Downtown area has significantly more shopping amenities than that of Sunset. If we look at the geographic location of the two neighbourhoods:

Russel Lutke: 301169861
Sean Yiang: 301364025

It should be easy to explain the difference in the number of shopping amenities. As we consider it is commonly recognized that downtown areas of a city are usually the central business district, which have more malls, brand stores, or small stores. In addition to the information on wikipedia about Vancouver city neighbourhoods, Sunset is a residential area filled with single family homes, low rise apartments, and small stores.

If we look at the distribution with foods and drinks type amenities:



It is easy to notice that Downtown has the highest score for food and drink amenities, while Sunset, as a residential area, has relatively low number of food and drink amenities.

Another potential problem is the way we categorize types of amenities. We grouped the types of amenities based on arbitrary decided similarities. For example, we grouped the ‘market_place’ and the ‘shop|clothes’ type amenity from the original ‘amenities-vancouver.json.gz’ dataset into ‘shopping’ type amenity. Although Downtown area may have more shop|clothes, it does not suggest that there are more market_place.

Summary

In general, our linear regression models did a fairly poor job of predicting anything about the neighbourhood census data, with coefficients of determination falling around zero or slightly less than zero. Several reasons may be hypothesized as to why this was the case. The first concerns the small size of the dataset. Given that the census data was divided into only 22 neighbourhoods, this placed a limitation on how many data points we could use. An insufficient amount of training data is therefore one potential cause of the observed results. A second explanation might be that the neighbourhood divisions

themselves were somewhat meaningless. While the City of Vancouver claims that the 22 provided subdivisions represent distinct cultural and economic centres, it is possible that this is not really the case and that the neighbourhood distinctions are more or less arbitrary. If the neighbourhood divisions fail to reflect real underlying differences in economy or culture then using them as a basis for the model is unlikely to be successful. A final explanation might be that, assuming the data points were sufficient and the neighbourhood divisions meaningful, there simply was no relationship between the category scores and any of the predicted properties (median income, proportion of people in low-income class, or density of single-person households). While one might expect that there is a correlation between, say, the density of educational or scientific amenities in a given area and the median income of that area's residents, this might not be the case. The various amenities could be spread fairly evenly across the city while each neighbourhood has differing values for income, poverty levels, and so on.

Russel Lutke: 301169861
Sean Yiang: 301364025

Project Experience Summary (Russel Lutke)

Project Description: Ran a linear regression analysis on OSM and City of Vancouver census data. Attempted to predict features of delineated neighbourhoods such as income based on the density of nearby amenities and venues. Also implemented a function that generated scores for each category of amenity for a given point and radius.

Skills Applied: A brief description of the various competencies employed during the course of the project work.

- Computer programming, primarily in Python. Used libraries related to data processing and analysis such as Pandas, numpy, and sklearn. An emphasis was placed on using more efficient library-specific methods for processing large amounts of data, as opposed to relying on generic control methods such as loops.
- General Research. While the OSM data was already provided, the neighbourhood boundary and census data had to be sought out. Research was done primarily through web searches using Google.
- Interpersonal Communication. While the project group included only two participants, communication was still key to successfully accomplishing the goal at hand. Work was done both in-person and via electronic media. Code was edited and shared primarily using Google Collaboratory.
- Project Management. This section encompasses skills such as prioritizing different aspects of the workload and deciding when certain avenues of the project must be abandoned or altered. It also includes time management and partitioning of the workload between group members.
- Idea Generation/Brainstorming. Due to the open-ended nature of the project, some time had to be spent generating ideas for what to implement/what problems to solve. As well as actually generating the ideas in the first place, this skill also encompasses efficiently determining which are feasible or interesting and discarding those which are neither.

Russel Lutke: 301169861
Sean Yiang: 301364025

Project Experience Summary (Sean Yang)

Project Description: Ran a multiple linear regression model on OSM and City of Vancouver census data to predict income and the number of single person households in each neighbourhood of Vancouver.

Skills Applied:

- Python: Used Geopandas and Sklearn libraries for data preprocessing and data analysis. Constructed efficient function for preprocessing large datasets, as opposed to loops.
- General Research: Explore additional relative datasets for analysis. Evaluate datasets based on reliability, relevance, and recency.
- Interpersonal Communication: Collaborate tasks and jobs both in-person and online. Finish coding tasks using Google Collaboratory.
- Time management: Schedule group meetings and in-person working sessions. Prioritize different objectives based on progress. Partition workloads.