Ranking Neighbourhoods in the City of Vancouver

Parshan Pourbakht, 301429585
Tudor Stoiu, 301425880

Problem

Introduction

Within the city of Vancouver there are a total of 22 different neighborhoods. Each neighbourhood has their pros and cons, our goal is to analyze the neighbourhoods and rank them. We base our ranking off a scoring system that considers 3 factors: average rent prices, crime count, and amenity accessibility. The scoring system will be out of 30, where each of the 3 factors will contribute a maximum of 10 points to the total score. Given a score we can then rank the neighbourhoods and determine the good and not so good Vancouver neighbourhoods to live in.

Data

Retrieval Finding data

We first began by finding relevant datasets for each of the 3 factors we will base our scoring system off of. Our datasets/data sources are listed below:

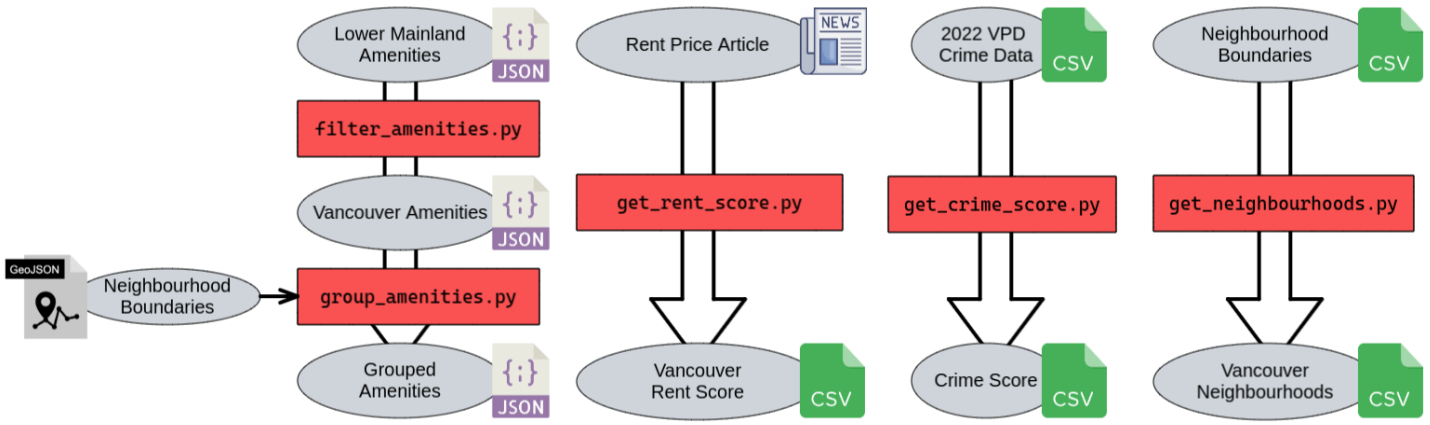
- OSM Lower Mainland Amenities (provided for Topic 2 projects)
  - `0-raw-data/amenities-lower-mainland.json`: A `JSON` file containing all amenities in the Lower Mainland
  - Used to create amenity score
- City of Vancouver Local area boundary
  - `0-raw-data/neighbourhood-boundaries.geojson:` A `GeoJSON` file containing geometric polygons representing the exact area belonging to each Vancouver neighbourhood
  - `0-raw-data/neighbourhood-boundaries.csv:` A `CSV` version of the above file that also contains the central coordinate of each Vancouver neighbourhoods
- VPD Crime dataset
  - `0-raw-data/crimedata_csv_AllNeighbourhoods_2022.csv`: A `CSV` file containing criminal activity data for each Vancouver neighbourhood
  - Used to create crime score
- Average Vancouver rent price by neighbourhood
  - A website containing the average rental price for each Vancouver neighbourhood

    - Used to create rent score
    - Provided as a website (With which we will web scrape from)
- [List of neighbourhoods in Vancouver]
    - Used for the retrieval of the official Vancouver neighbourhoods
    - Provided as a website (With which we will web scrape from)

Cleaning/Filtering Prepare/acquire relevant parts of data

The first step in our pipeline is to clean the provided `0-raw-data/amenities-lower-mainland.json`, which contains amenities all over the Lower Mainland extending as far as Chilliwack. For the purpose of our analysis, we only want the amenities in the 22 neighbourhoods of Vancouver and we want them classified by the neighbourhood they are in.

Data Pipeline Diagram



Furthermore we only want the major amenities. We decided on doing so as we wanted to focus on only the essential amenities for scoring purposes, excluding the non essential ones.

1. `1-cleaning/filter_amenities.py`
   First, we filtered our data down to the city of Vancouver, keeping only the major amenities listed on the right. Output of this step is found in `0-working-data/amenities-vancouver.json`
2. `1-cleaning/group-amenities.py`
   Next, we group amenities into their respective neighbourhood. The [GeoPandas] library allows us to store the Vancouver local area boundary GeoJSON dataset in a GeoDataFrame and perform operations like checking whether a point is contained `.within()` a polygon, this is exactly how

Major Amenities

1. Schools
2. Restaurants
3. Pharmacies
4. Banks
5. Community Centres
6. Fuel
7. Dentists
8. Doctors
9. Hospitals

`1-cleaning/group-amenities.py` finds the neighbourhood of each amenity. Output of this step is found in `0-working-data/grouped-amenities.geojson`
(visualized in the figure “Amenities by Vancouver Neighbourhood” at the top of page 5)

The next part in our pipeline is for us to retrieve the average monthly rent amounts for each neighbourhood. The `1-cleaning/get_rent_score.py` file does this by scraping an article on Vancouver renting prices using the Beautiful Soup library. Regarding the retrieval of Vancouver crime data, the `1-cleaning/get_crime_score.py` file is responsible for reading the City of Vancouver Crime dataset. With this dataset we filter the dataset to only contain the type of crime, neighbourhoods names, and the x and y coordinates given in UTM Zone 10 coordinates. The final step is to get the neighbourhoods and their central coordinate which will be used in the Haversine distance calculation in our analysis. This is done by `1-cleaning/get_neighbourhoods.py` and is stored in `0-working-data/vancouver_neighbourhoods.csv`

Analysis Using data

Scoring System

Our goal is to rank the neighbourhoods by a **3 factor scoring metric**. These factors are rent prices, crime activity, and amenity rarity and accessibility which we will derive from datasets in the form of 3 scores on a scale from 0 through 10 for each neighbourhood. These scores then combine to give an overall score to each neighbourhood for us to rank by. The `1-cleaning/get_rent_score.py, 1-cleaning/get_crime_score.py, 2-analysis/get_amenities_score.py` files are responsible for producing these 3 scores: rent score, crime score, and the amenity score.

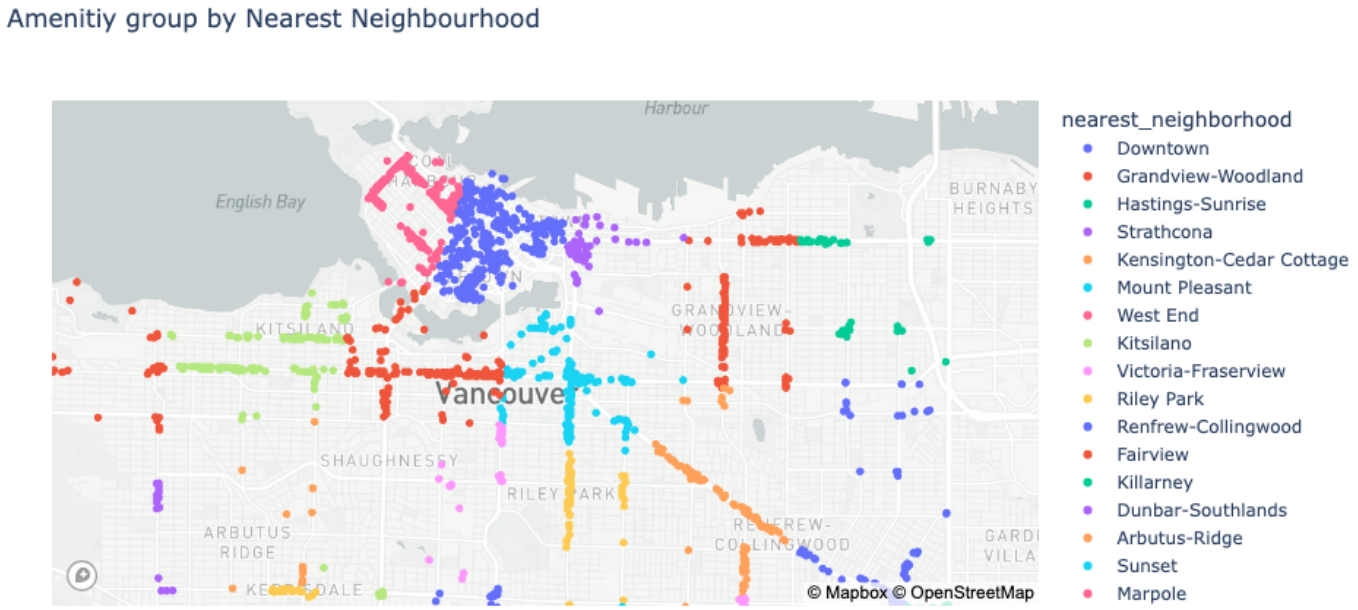
The rent, crime, and amenity scores for each neighbourhood are created by normalizing the rent prices, crime counts, and amenity rarity, respectively, to a scoring scale from 0 through 10 relative to the other neighbourhoods. Firstly, the rent score goes from 0 for the most expensive average rent price to 10 for the least expensive average rent price. Secondly, the crime score goes from 0 for the most criminal activity to 10 for the least criminal activity. Lastly, the amenity score goes from 0 for the least amenity rich neighbourhood to a score of 10 for the most amenity rich neighbourhood. Amenity richness is based on the accessibility and rarity of the major amenities (listed above) in a particular neighbourhood. Rarer amenities contribute more to the score of a neighbourhood than less rarer amenities. For all 3 criterias, the higher the score the better the neighbourhood. These scores are then passed into `2-analysis/get_overall_score.py` to calculate the overall score for each neighbourhood by summing the rent, crime, and amenity scores. The overall scores allow us to rank the neighbourhood from best (highest score) to worst (lowest score).

Amenity Neighbourhood Classification

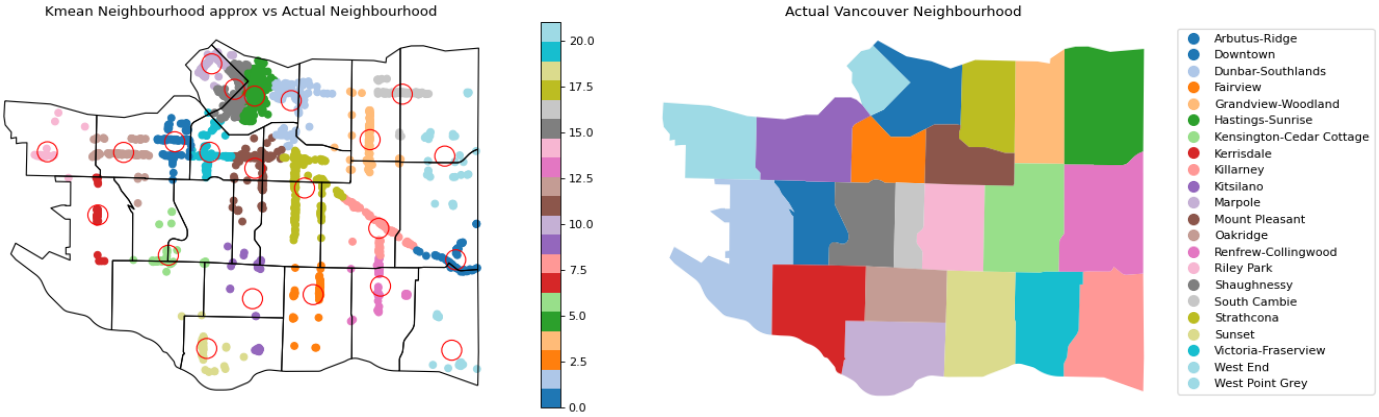
Given all major amenities in Vancouver we had to come up with a way of classifying what neighbourhoods each of them fell under. Regarding the classification of the amenities we used 3 methods:

1. *Nearest Neighbourhood*
2. *K-means Clustering*
3. *Point in Polygon Query*

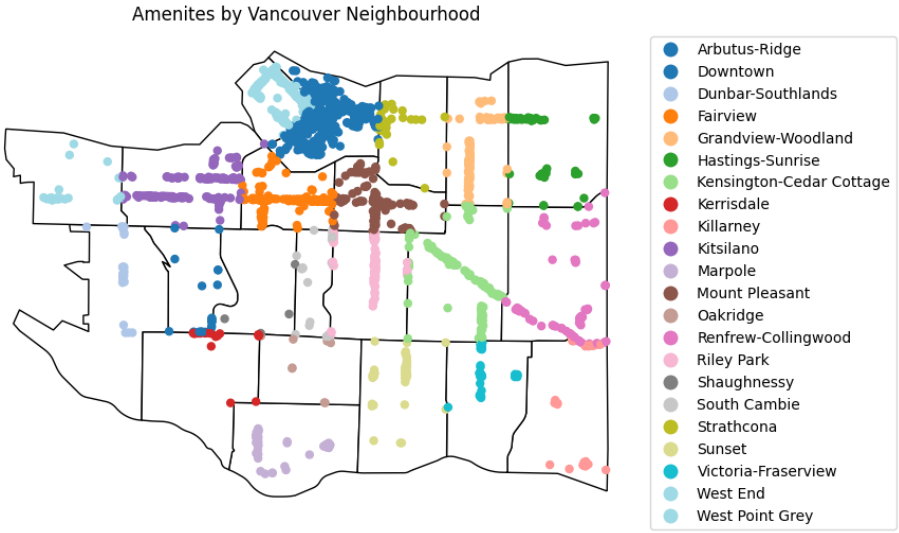
The **Nearest Neighbourhood** method classifies an amenity by the nearest neighbourhood. The nearest neighbourhood to a particular amenity is found by calculating the Haversine distance from the amenity's coordinate to each neighbourhood's central coordinate and selecting the neighbourhood with the minimum Haversine distance.



The **K-means clustering** method classifies an amenity by clustering all amenities into K groups of amenities based on their location. Since we know that Vancouver has 22 neighbourhoods we use k = 22 with the goal of roughly approximating each of the neighbourhoods with each cluster. The purpose of this is to see if the natural grouping of amenities relate to the actual neighbourhood segments.

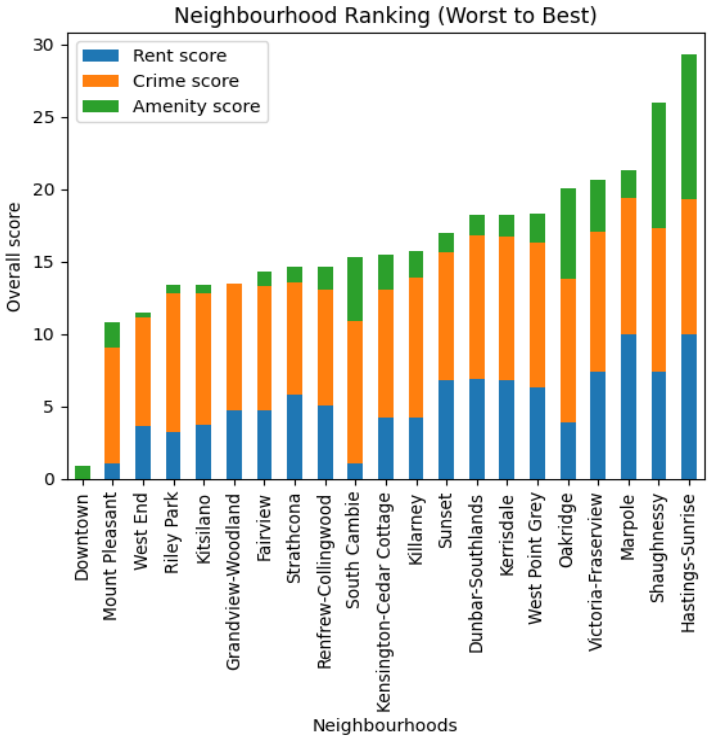


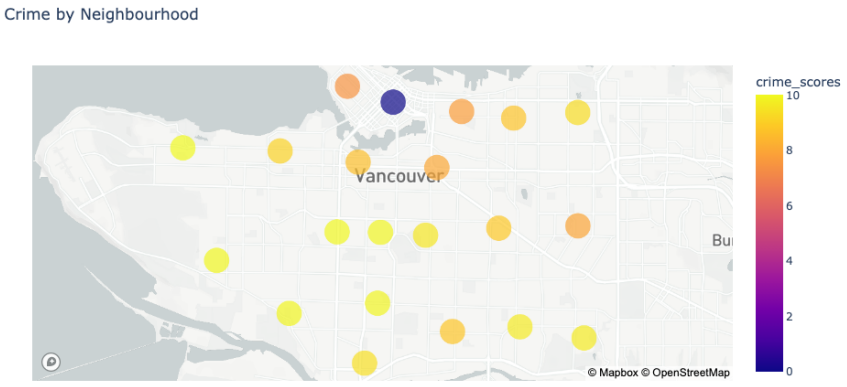
The **Point in Polygon Query** method resulted in the most accurate classification as it uses the Vancouver Neighbourhood Boundary GeoJSON dataset and checks if the location of an amenity (a **point**) is within a neighbourhood's boundary (defined as a **polygon**). This is done in `1-cleaning/group-amenities.py` in the Cleaning step of the pipeline.



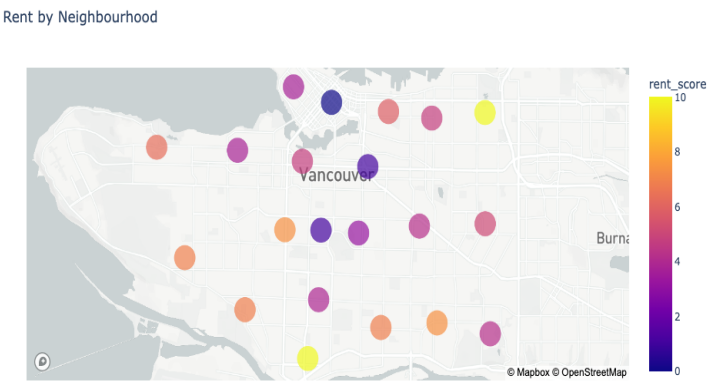
Results Findings and conclusions

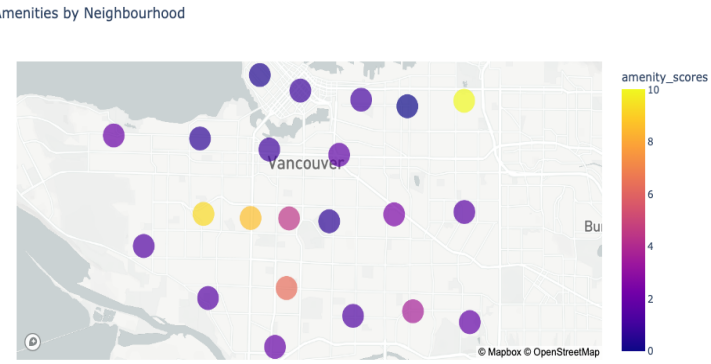
From the output of `2-analysis/get_overall_score.py` we were able to determine the overall score for each of the amenities. As shown in the diagram (right) the "best" neighbourhood to live in based on our 3 factor metric is the neighbourhood of Hastings-Sunshine, while the "worst" is Downtown. Note: Downtown was given a crime and rent score of 0 because it has the highest rent and crime activity relative to the other neighbourhoods. Grandview-Woodland was the least amenity rich neighbourhood with an amenity score of 0.





Regarding the crime scores for each neighbourhood, we noticed they all increase in criminal activity the closer they get to the Downtown neighbourhood (shown by the dark purple dot in the left diagram). Note the diagram to the left and the next 2 diagrams are interactive and can be accessed in the notebook: `3-results/visualizations.ipynb`.





For our rent scores (left) we noticed how rent scores/prices increase as we move more Downtown, while neighbourhoods found more on the outskirts of Vancouver (shown by the yellow colored, used to indicate good rent scores) are relatively cheaper. Additionally, regarding the amenities scores we noticed that a majority of the neighborhoods in Vancouver lack access to the major amenities we listed (shown by the majority of purple dots in the diagram to the right).

All the above visualizations and more can be found in `3-results/visualizations.ipynb` notebook. Note some of the plots are interactive and provide further details.

Lessons Learned

- To get the list of neighbourhoods, we initially scraped the wikipedia List of Vancouver neighbourhoods and then used geopy to reverse geocode the address of the neighbourhood to get a coordinate for the neighbourhood. This method was very slow and tedious. This was done in the `0-archive/neighbourhoods.py file.` After doing this we stumbled across the local area boundary dataset from the city of Vancouver. This dataset contained the official neighbourhoods of Vancouver and also contained geometric polygons representing the neighbourhoods.
- The above is an example of not spending enough time looking for data in the right format. Much of our time was wasted in the pursuit of manipulating our data into the format we needed for our analysis. When there exists data in the format we need, we just didn't find it at first.
- Our scoring system could be expanded to consider more than just the rent, crime, and amenities. We could possibly explore factors like transit, traffic, and weather.
- Never assume your data is complete and working. Originally with the OSM data we assumed the data consisted of all the amenities in Vancouver; however this turned out to not be the case. Only after plotting the amenity locations on the map of Vancouver were we able to truly see that our data was not what we expected and had to be cleaned and filtered.

Project Experience Summary

Parshan

Vancouver Neighbourhood Ranking **Oct. - Dec. 2022**

- Performed data analysis using Pandas to retrieve, clean/filter, and analyze datasets
- Immensely improved organization, by modularizing and executing a proper data pipeline system
- Visualized neighbourhoods based off their rent, crime, and amenities using MatPlot, Geopandas, and Plotly
- Performed web scraping using Beautiful soup to extract monthly Vancouver rent data

Tudor

Multi-Factor Vancouver Neighbourhood Ranking **Oct. - Dec. 2022**

- Evaluated and ranked the neighbourhoods of Vancouver based on cost of living, amenities accessibility, and safety data
- Leveraged the GeoPandas library to conduct analysis of the geospatial data in Vancouver
- Significantly improved developer productivity by creating Bash shell scripts to streamline the data analysis pipeline