Vancouver Parking Tickets - Data Analysis

Manvir Heer, Kevin Chung

Submitted as part of the term project for CMPT 353: Computational Data Science Summer 2024

Overview

In July 2024, TransLink, Metro Vancouver's primary transit provider, reported a significant funding shortfall that could lead to drastic service cuts by late 2025. Without a new funding strategy, the region could see bus services halved and SkyTrain and SeaBus frequencies reduced by a third. These cuts may force more residents to rely on personal vehicles for their commutes, adding pressure to Vancouver’s already strained parking system, particularly in the downtown core.

The likely increase in car usage poses a significant challenge to Vancouver's already overburdened parking system, especially in the downtown area. More vehicles on the road raise concerns about the sufficiency of current parking policies, enforcement, and public awareness of parking rules. Pinpointing where parking infractions are most common and understanding the causes behind these issues are essential for enhancing the city's parking management. A focused analysis could identify areas with the highest rates of parking infractions, helping the City of Vancouver target improvements. This data-driven approach would allow the city to enhance signage, introduce new regulations where necessary, and improve compliance, ultimately reducing the strain on the parking infrastructure.

Data Collection

For this project, we sourced multiple datasets from the Vancouver Open Data Portal, focusing on parking tickets issued across Vancouver.

Primary Dataset: Parking Tickets

The primary dataset includes detailed records of over 3 million parking tickets issued in Vancouver from 2017 to 2024. Key details include street names, block numbers, ticket issuance year, ticket status, violated bylaw, and the corresponding section.

Geocoding with Google Maps API

To enable spatial analysis, we used the Google Maps API to geocode addresses in the parking ticket data. By converting street names and block numbers into latitude and longitude coordinates, we were able to map tickets and perform a detailed spatial analysis. This process involved sending API requests for each unique address, returning geographic coordinates that were then appended to the dataset. We implemented rate limiting and error handling to manage the volume of requests, ensuring efficiency and accuracy.

Supplementary Dataset: Neighborhood Boundaries

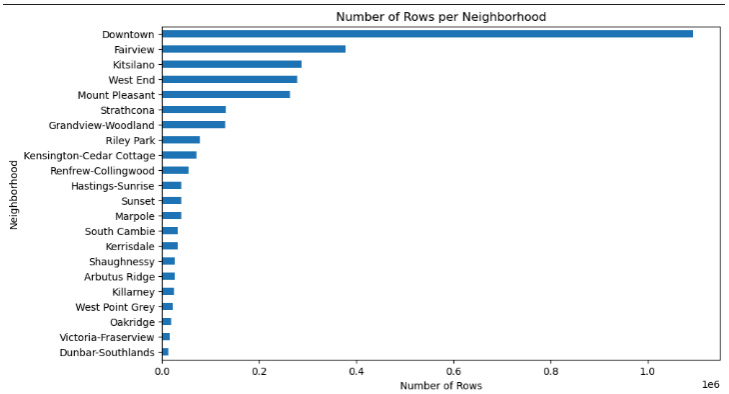
We also used a supplementary dataset showcasing neighborhood boundaries within Vancouver. This allowed us to facilitate targeted analysis within defined geographic sections. We performed a spatial join between the geocoded parking ticket data and neighborhood boundaries using a GeoDataFrame, ensuring both datasets were aligned in their coordinate reference systems (CRS). The join used a 'within' predicate to accurately assign each ticket to its corresponding neighborhood. Afterward, we cleaned the dataset by removing unnecessary columns to streamline the data for analysis.

Data Preparation/Cleaning

We focused our analysis on parking infractions related to **Bylaw 2952** and **Bylaw 2849** to concentrate on the most common violations. This filtering reduced noise and ensured our analysis remained relevant. To enhance the dataset, we performed a spatial join between the geocoded ticket data and neighborhood boundaries, allowing us to associate each parking ticket with a specific neighborhood. Our analysis concentrated on **Downtown** and **West End** due to their high infraction density. After the spatial join, unnecessary columns, such as geometry and index columns, were removed to streamline the dataset. We also added a “Quarter” column to categorize infractions by the quarter of the year, enabling us to examine seasonal trends in parking violations. To facilitate temporal analysis, we organized the data into separate CSV files for each year, specifically for the Downtown and West End neighborhoods.

Exploratory Data Analysis:

We decided to check whether certain locations on average had more parking tickets given based on the neighborhood.

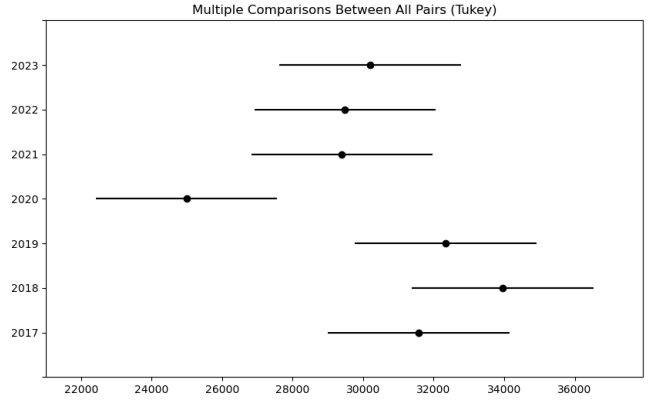


From the above graph we see that a majority of the parking tickets are in downtown Vancouver. As such we decided to limit our clustering to mainly the first 2 neighborhoods and see if there are any significant areas within those neighborhoods as they

Findings

Question 1: Are there any differences in getting parking tickets at different years

We first checked to see if our grouped data was all normally distributed in order to check if we could use anova testing. Luckily, we found that all data had a p-value > 0.05 using the stats.normaltest to check individual data. We used a one way anova test to see if there were any significant differences in the parking ticket count and we got a p-value of 6.929164932413644e-05 < 0.05 meaning that there were statistical differences between the count of parking tickets in each year. However, to get a better understanding of how they differed we did a Post HOC analysis to get a better understanding of the differences. Although the Anova testing showed us there is a significant difference the Post HOC analysis illustrates differently as only the year 2020 bears a statistical significance in difference compared to all years whereas all the other comparisons seemingly does not bear a statistical difference in means that cause us to reject the null hypothesis. This makes sense as during 2020 was the beginning of the isolation and we see that in later years it is slowly returning back to normal. (Use the graph below for reference)



Multiple Comparison of Means - Tukey HSD, FWER=0.05

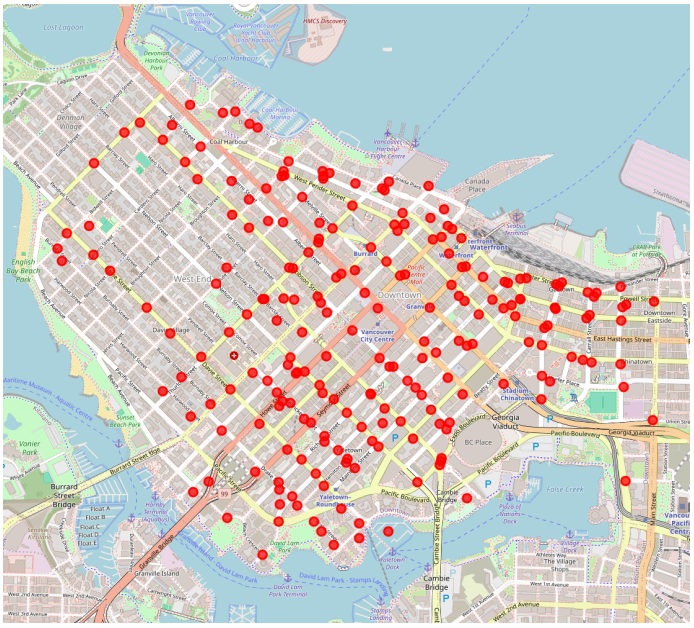
group1	group2	meandiff	p-adj	lower	upper	reject
2017	2018	2377.0	0.8015	-2772.7751	7526.7751	False
2017	2019	770.75	0.9993	-4379.0251	5920.5251	False
2017	2020	-6586.8333	0.0041	-11736.6084	-1437.0582	True
2017	2021	-2178.25	0.8587	-7328.0251	2971.5251	False
2017	2022	-2091.0	0.8805	-7240.7751	3058.7751	False
2017	2023	-1375.8333	0.9834	-6525.6084	3773.9418	False
2018	2019	-1606.25	0.9639	-6756.0251	3543.5251	False
2018	2020	-8963.8333	0.0	-14113.6084	-3814.0582	True
2018	2021	-4555.25	0.1179	-9705.0251	594.5251	False
2018	2022	-4468.0	0.1323	-9617.7751	681.7751	False
...						
2021	2022	87.25	1.0	-5062.5251	5237.0251	False
2021	2023	802.4167	0.9991	-4347.3584	5952.1918	False
2022	2023	715.1667	0.9995	-4434.6084	5864.9418	False

Question 2: Are there any locations which had a higher rate of getting parking tickets

First, we processed the parking ticket data for each year from 2017 to 2024, focusing on the downtown area. After loading and cleaning the data specifically by removing entries without latitude and longitude coordinates we created yearly heatmaps centered on Vancouver. These heatmaps visually represented the frequency of parking infractions, with denser areas

indicating higher rates of tickets. The density of infractions was calculated for each year, allowing us to observe trends over time, which we then visualized using a bar chart showing the density of infractions per year.

Next, we applied DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to identify clusters of parking infractions. We set specific parameters for the clustering process, including the maximum distance between points (`eps`) and the minimum number of points required to form a cluster (`min_samples`). We aggregated the data by latitude and longitude, calculating the number of infractions at each location and the number of years each location appeared in the dataset. This allowed us to run DBSCAN and identify clusters of locations with persistently high rates of parking tickets across all years. We filtered for "persistent hotspots," or locations where the number of infractions exceeded 1,500 tickets and were consistent across the years.



Finally, we visualized these hotspots using Folium, marking locations on a map that consistently experienced higher rates of parking infractions. This combined approach of heatmap visualization and clustering provided a comprehensive view of where parking tickets were most frequently issued in Vancouver, offering valuable insights for potential improvements to parking policies and enforcement.

Results

For our results, we found that after running our code we created a lot of heatmaps but could not see any real patterns for our findings. However, we could see which general areas had the highest amount of parking tickets within the years and noticed that south of downtown generally had the highest counts of parking tickets. As for whether there were any statistical differences using anova testing we did get a p-value <0.05 meaning we reject the null hypothesis that all mean sizes are statistically not different. However, when using the Post HOC Analysis we found that the only ones that had a p-value that shows statistical difference were 2020 and all other years whereas the rest did not show enough of a difference to reject the null hypothesis that each year was significantly different to each other.

Limitations

Initially, we wanted to check if there was a difference in the prices of parking tickets at different locations and which times had the most parking tickets given out. However, we found ourselves limited by the lack of variables in our data set. Although we had over 3 million data points about parking tickets we were not given the price of each ticket which we could have used to answer more significant questions such as which locations were the most expensive to get a parking ticket. Although we were given the bylaw and section infraction of the parking tickets. Gathering the necessary information based on the law violation was difficult as the price can vary from lowest being $77 to a maximum of $10,000 dollars. As such we had to refine our problem quite a bit.

Additionally, due to the lack of variables we were not able to build any substantial model that could accurately predict the count of parking tickets based on location. For example if we were given a price variable for each parking ticket data we could have used it to predict the price of tickets based on the block and street given and used a machine learning model.

Project experience summaries

Kevin Chung

At the beginning of the project, it was a struggle to figure out an idea to make a project out of. At first, I wanted to find a data set that had a lot of points. I figured that it would be good enough to use for our project. However, this would not be very ideal as although the data set was very large the amount of useful variables were limited making and as such found it difficult to do data analysis. The main packages I used were pandas to clean the data into data I could use, matplotlib for some data visualisation, debugging with geopandas for converting addresses into longitude and latitude. Finally, I used the packages scipy.stats and statsmodels to do my anova testing and Post Hoc Analysis. Where I found the p-value of individual data sets and found whether there were significant differences in the parking data.

Manvir Heer

During the project, I was responsible for geocoding parking tickets using the Google Maps API, which allowed us to accurately map the locations of infractions. I then used GeoPandas to distinguish between different neighborhoods, enabling a more detailed spatial analysis. I utilized Folium and DBSCAN clustering to identify and visualize hotspots of parking infractions across the city. These efforts were crucial in providing a clear understanding of the areas most affected by parking violations, contributing to our team's overall analysis and recommendations. I also actively worked on the data cleaning and preparation where I did the distinction for downtown area and separate csvs for yearly data.