CMPT 732 G100 – Big Data Lab 1 Final Project:

Smart City “YVR”

Problem Statement

For any city to ensure a sustainability and quality of life needs, it needs to answer some critical questions. We have taken up three major issues:

- Housing Affordability
- Energy Consumption
- Public Transport Connectivity

The aim of this project is to use a data driven approach to address them. The goal is to identify areas of concerns, provide with data backed trends and concerns that might arise in future and give insights helping us understand. We aim to use the Big Data Technologies to process the Raw data into useful insights.

a little more about the problem

The biggest challenge towards a data driven smart city model, is finding data. The reason we choose Vancouver, is because the abundant open source data available online. Also, over the last few months we have become familiar with this city.

In this project, we will be taking up large volume of raw data and Transform it to make it more sensible and understandable; The major work is done on ETL and Analytics part to develop various key metrics to understand all 3 Problems. We have used data visualization to see the patterns and trends. This has potential to guide stakeholders towards more efficient resource allocation and policy formulation.

Methodology

To solve this problem, we decided to tackle them individually, due to the large volume of data present. In general, we have built 3 Separate Data Pipelines for each of the Problem Area: Housing, Energy, Transportation.

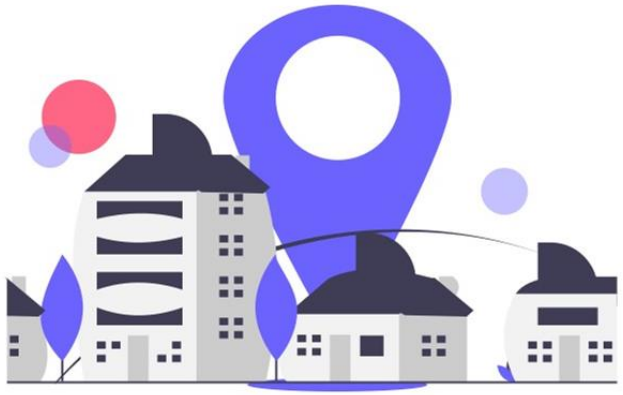
To get better understanding and for identifying trends we had to combine multiple data sets, perform Data Transformation, cleaning the data. We created analytics metrics to gain understanding of data trends and patterns and visualized the results. We have used Machine Learning for Predictions and have done a comparisons of different models and results to understand which is best suited for our dataset

Task and Technologies used:

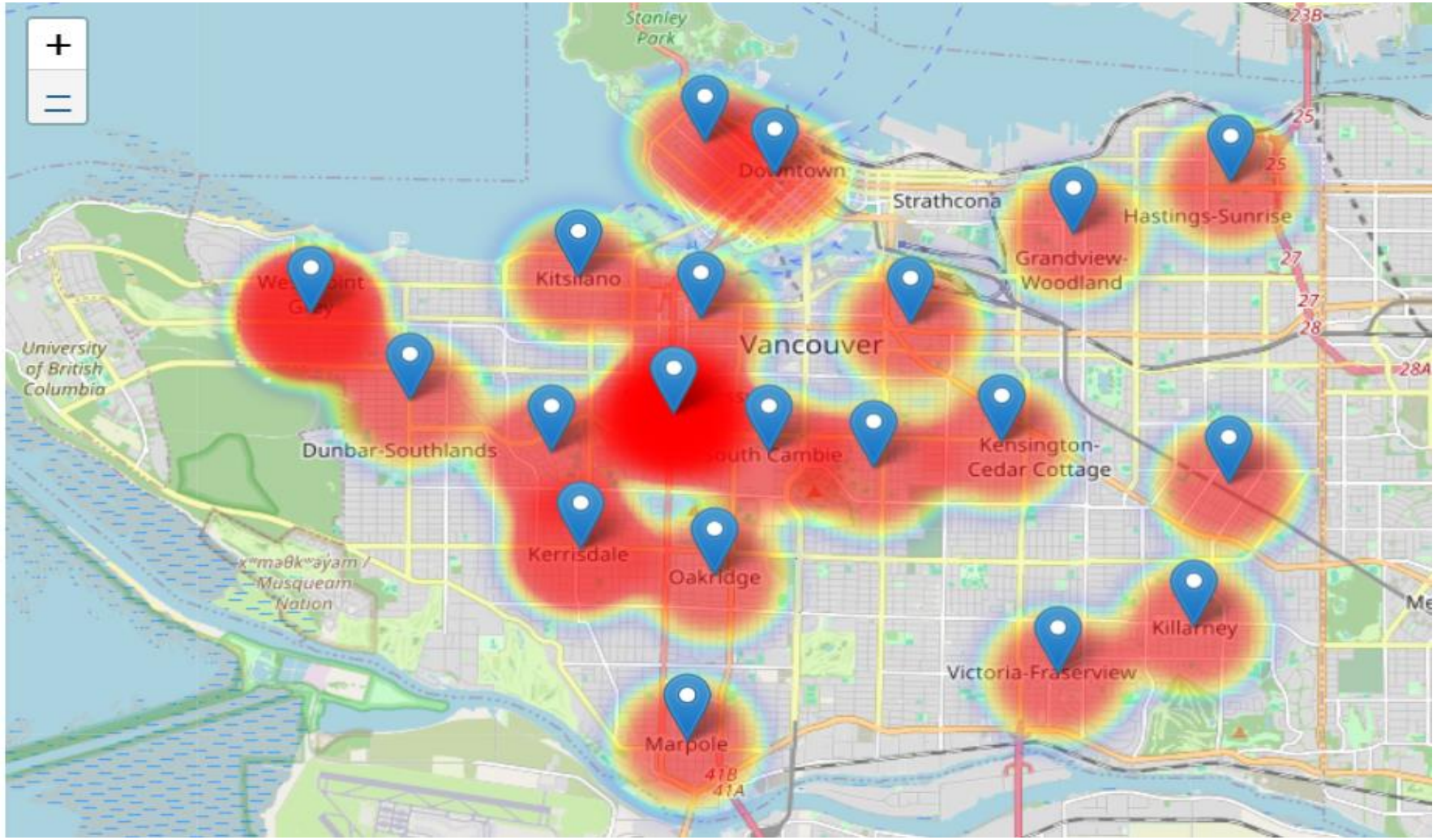
- ETL, Analytics, Machine Learning: Python, Spark, Pandas
- Data Visualization: Matplotlib, Folium, NumPy

Housing Affordability

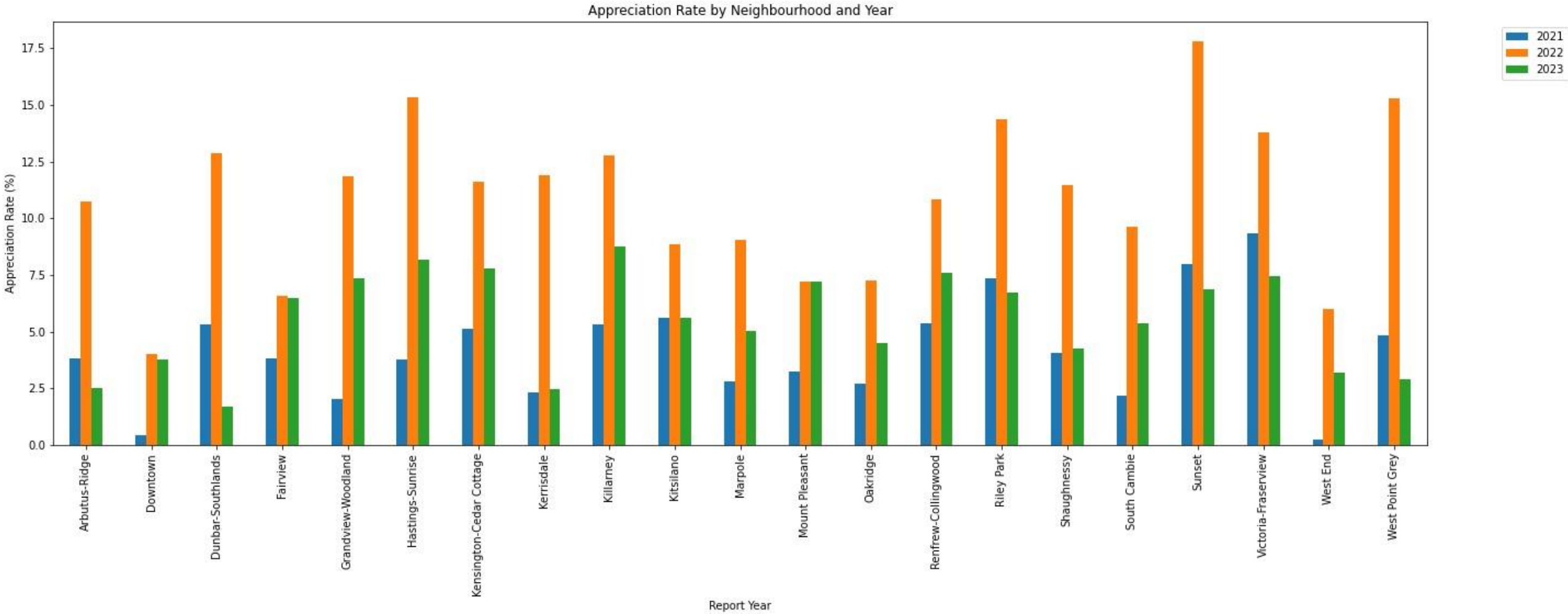
To understand the situation regarding housing affordability. We have done data-driven analysis of Vancouver's housing landscape involved an Extract, Transform, Load (ETL) process, cleansing data, and enriching it for insights. Key metrics such as property value appreciation rates, tax burdens based on zoning, and the impact of property age on value were derived and visualized through diverse graphs. Three Machine learning models were used to predicted property values and were evaluated for accuracy where we did algorithm comparison. Our findings highlighted varied property trends, tax implications, and affordability across neighborhoods.



Housing Affordability Index over different Neighbourhood



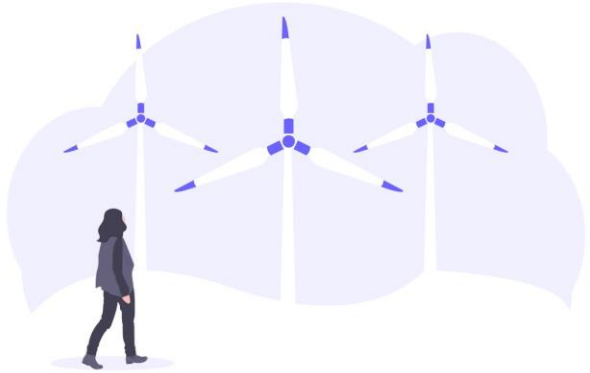
Property Appreciation Rate over the Years



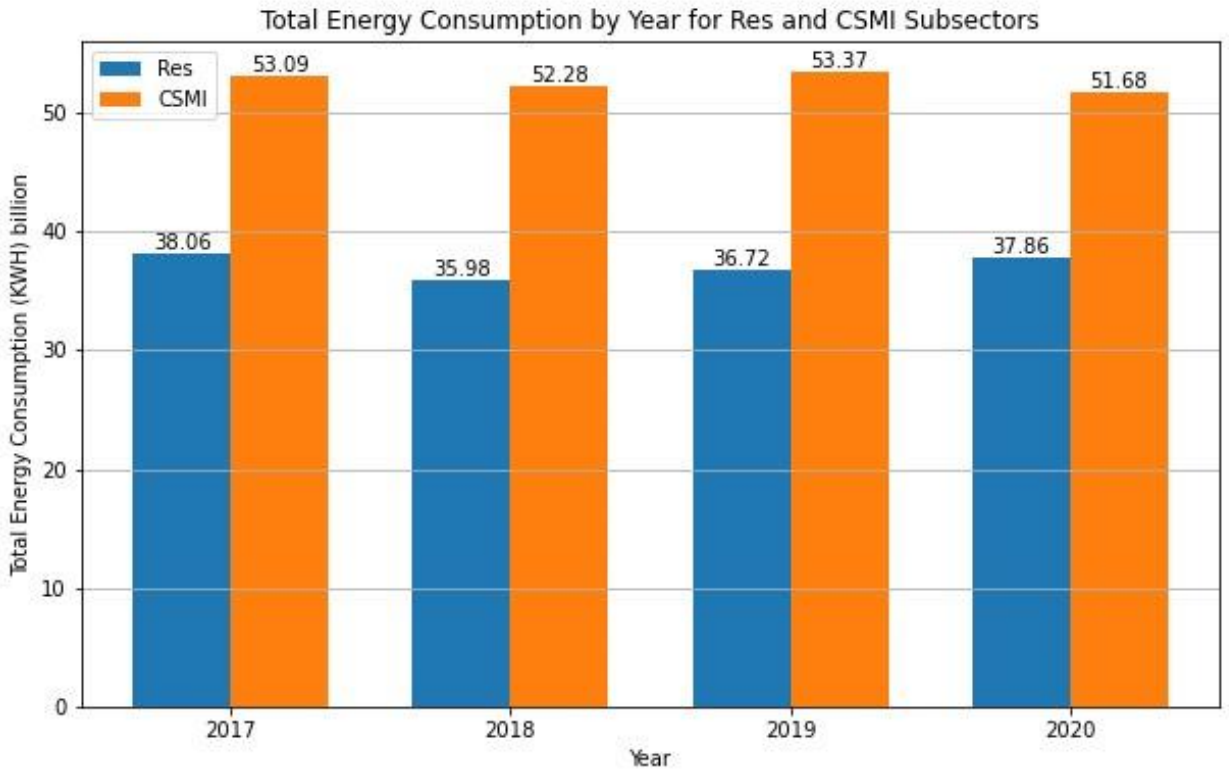
Energy Consumption

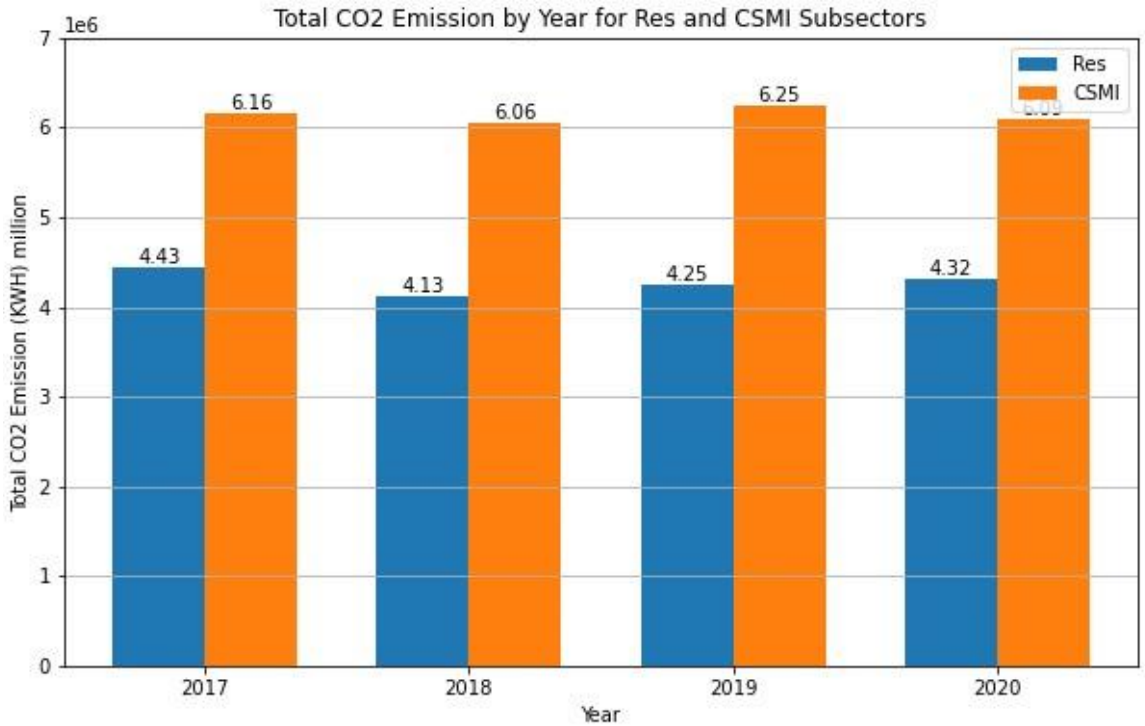
We've utilized PySpark functionalities to transform, clean, and filter the dataset, focusing on specific conditions and refining the data for analysis. We used Spark SQL to extract relevant information based on various quality checks and restrictions. We performed analytics tasks, including computing CO2 emissions by organization types, assessing total energy consumption across different subsectors, and examining CO2 emissions for specific regions.

To provide a clear visual representation, we have created several graphs illustrating these metrics. Moreover, we applied Linear Regression to predict future CO2 emissions for the next four years based on historical patterns. This comprehensive analysis offers insights into energy usage trends, CO2 emissions, and predictive models, enabling informed decision-making.



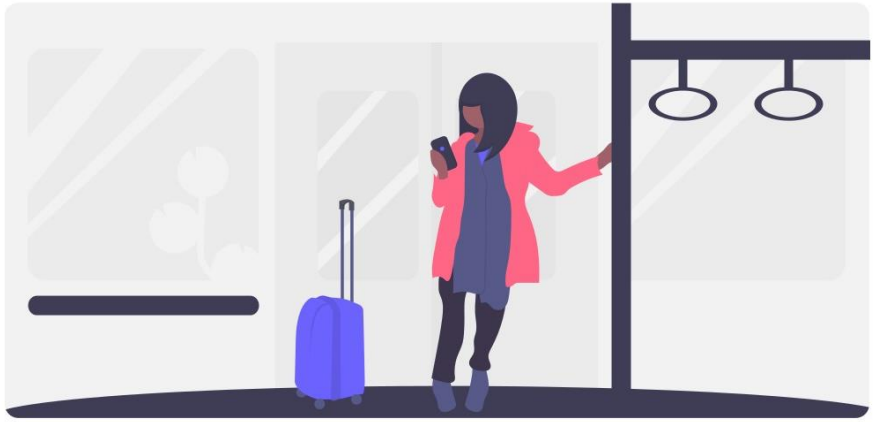
Comparison of Energy Consumption and CO2 Emission over the years



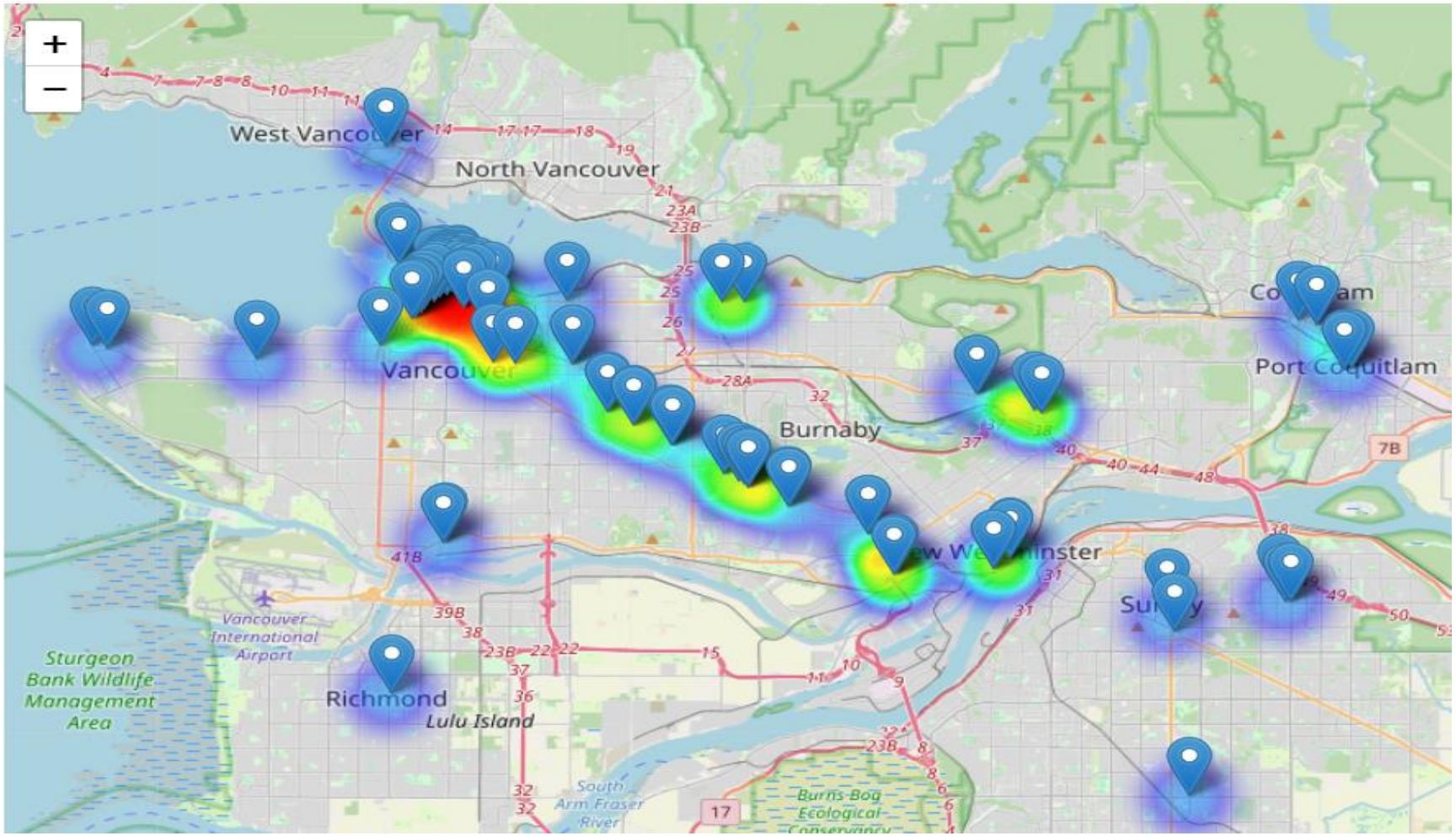


Public Transport Connectivity

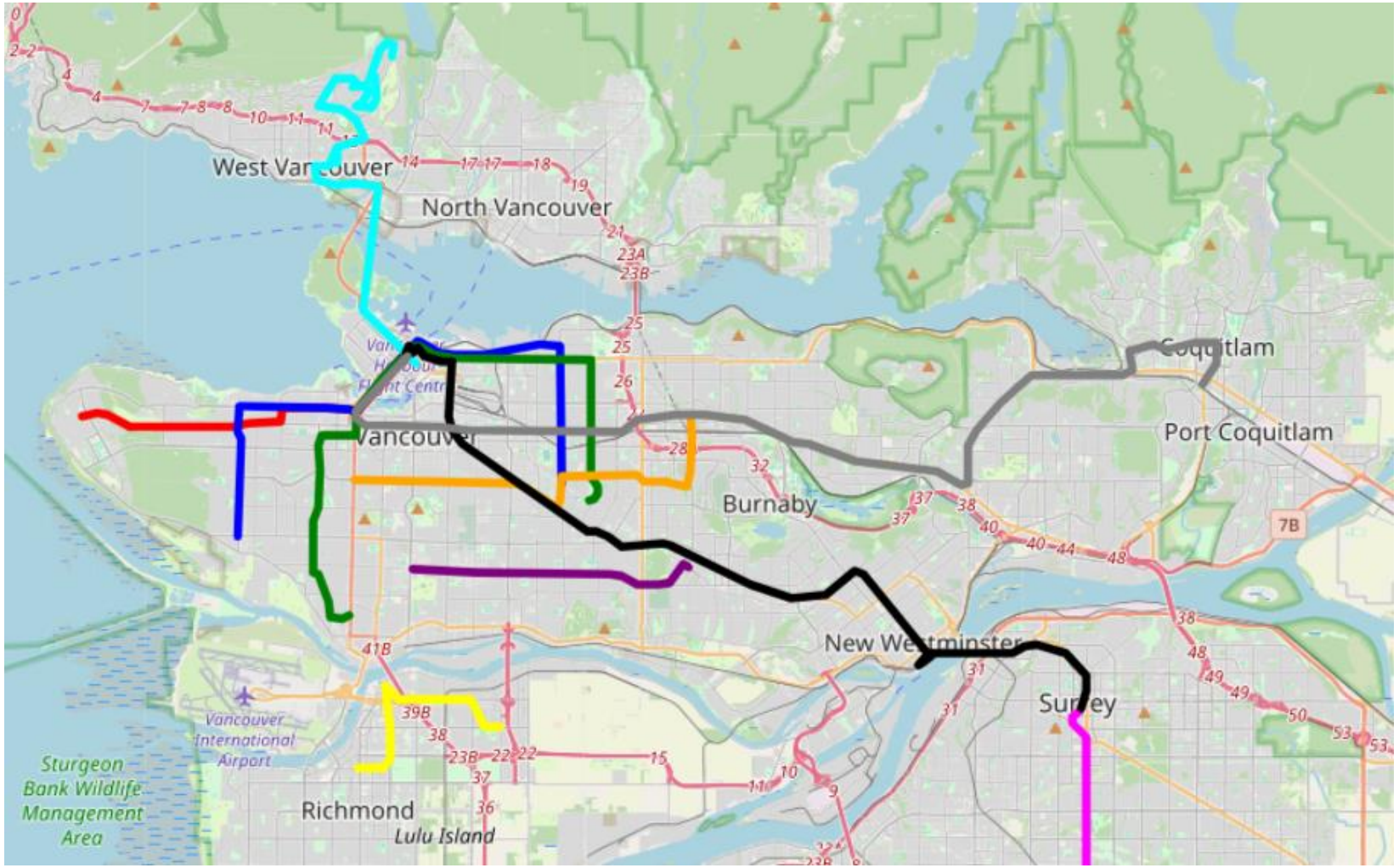
We combined multiple transport-related datasets using PySpark. Initially, we loaded and merged various tables like stops, stop times, trips, routes, calendar, and directions. This comprehensive data transformation enabled us to generate a combined dataset for analysis. The key tasks included assessing the busiest stops in Vancouver, plotting their frequency distribution on a heatmap using Folium, and displaying the number of active buses per hour throughout the week. Additionally, we visualized the longest bus routes connecting different locations by drawing these routes on an interactive map with distinct colors. These steps facilitated an insightful exploration of bus transit patterns, stop frequencies, and route lengths within the transportation network.



Top 100 Busiest Bus Stops in Vancouver



Top 10 Longest Bus Routes in Vancouver

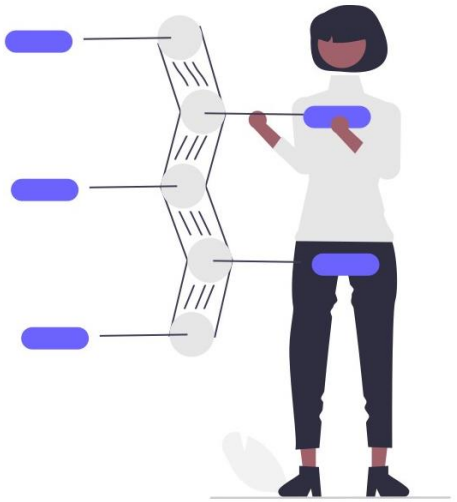


Challenges

- The initial problem we faced was finding data, although we were able to find data but most of the datasets were incomplete, so we had to find additional datasets that we could use and join to get complete information
- Once we were able to finalize the dataset we had to spend a significant amount of time in research to understand data, find anomalies, remove redundant data, and clean data to only have relevant portion as the data was huge in volume.
- Even we had done data cleaning we had to ensure to use big data principles for processing of data to it's size and time taken in computation
- Compatibility issues with AWS, a major part of project is Data visualization. We were using Matplotlib and Folium, but we found that AWS doesn't support them because of this we had to move our entire code to Jupyter Notebook
- Finally, we had to go through documentation and learn Matplotlib and Folium as it was our first time using them. Exploring through that was a good learning experience.

Results

- For Housing Affordability: Three biggest takeaways are the exponential increase in price of houses in 2022 has impacted the housing affordability. The burden of Property Tax on residential property has equal impact and it should be capped according to average income of demographics. Finally, the major increase in the land value with respect to improvement value of a house should be restrained.
- Analyzed the areas where the Co2 emission trends are high, energy consumption patterns, highlighting areas for the environmental sustainability and improved quality of life by moving more towards renewable energy than non-renewable sources.
- For Transportation: There is a lacking of connectivity between Vancouver with nearby areas like Richmond, Surrey. Burnaby has high concentration of busy hubs, while Surrey has fewer, indicating an opportunity to create more connections in Surrey for better accessibility across different locations



Thank You !!!