Quad Squad Project Report

Inderjeet Singh(isb5), Naina Thapar(nta65), Nidhi Kantekar(nkk9), Rituraj Ojha(roa11)

Problem Definition:

Our project's goal is to examine Mobi Shaw's bike trends and suggest recommendations for maximizing profitability. We used Mobi Shaw Raw Data for the years 2020-2022. Further, we have also incorporated the Weather Data for Vancouver and merged them to get more insights. We then performed data cleaning, ETL, and feature engineering on the merged dataset. This dataset was used to create visualizations to depict metrics and trends related to the usage of the bikes.

Methodology:

Data Collection:

We used three types of datasets to create our final dataset.

1. Mobi Shaw system data Year 2020- 2022
2. Weather Data for Vancouver
3. Latitudes and Longitudes

Once we collected all this data, we merged it according to the date and locations and created our final dataset after performing feature engineering using Pyspark on the Amazon EMR cluster.

Architecture:

The fig. 1 shows the flow of our project.

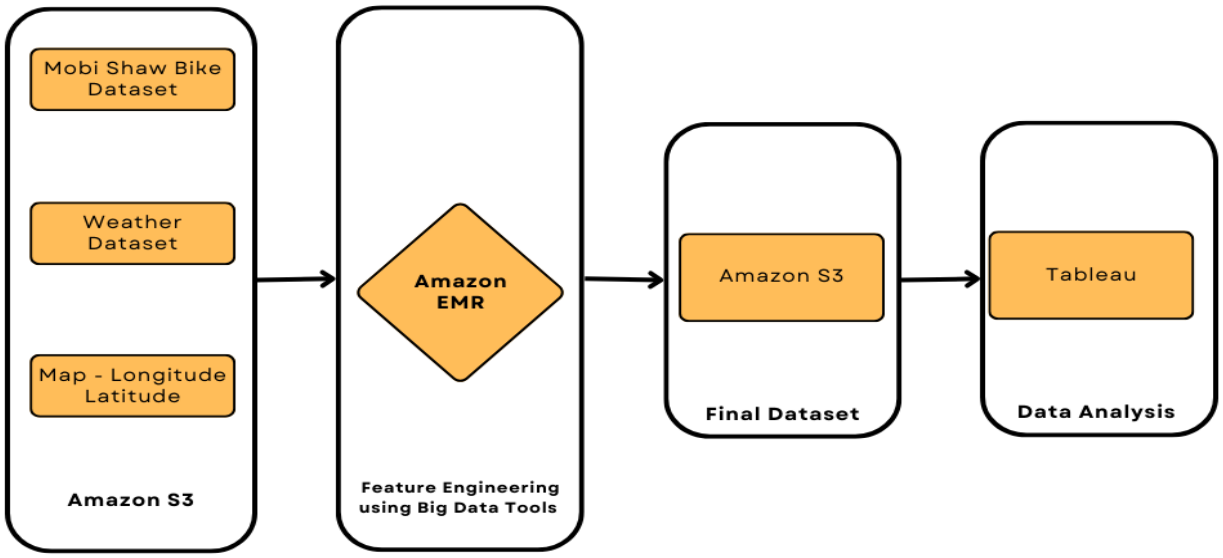


Fig: 1

ETL:
Three of the datasets we used were: Mobi Shaw Bike dataset, Weather Dataset, and Latitudes and Longitudes dataset (map). The Mobi Shaw Bike dataset provides us with the information about the start time, end time, duration of the ride, departure station, return station, type of membership, and more from January 2020 to October 2022. The Weather dataset provides us with the daily weather recorded in Vancouver. The Latitude and Longitude datasets provide the latitude and longitudes of the bike stands locations. These datasets and codes were stored in an Amazon S3 bucket. We did the ETL using Pyspark on the Amazon EMR cluster. The Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Spark, on AWS to process and analyze vast amounts of data. The Mobi Shaw Bike Dataset was taken and cleaned using several dataframe functions available in the Pyspark library. A few new columns such as 'Pass' and 'Seasons' were created based on the available columns. After the data cleaning step, the weather dataset was loaded, cleaned, and merged with the Mobi Shaw Bike Dataset. Also, the latitude and longitude dataset was merged to get the coordinates of the Departure Stations and the Return Stations. Furthermore, a few of the columns were dropped which contained a lot of null values. All these queries were run on several EC2 nodes of the Amazon EMR cluster. After all the above steps, the output was stored on the same bucket of Amazon S3. Furthermore, we created a machine learning code, named ml.py which will be used to predict the total count of bikes which will be used for the upcoming days based on few of the features. The input of the model will be the final dataset which we created using above feature engineering. This was one of our stretch goals.

Technologies

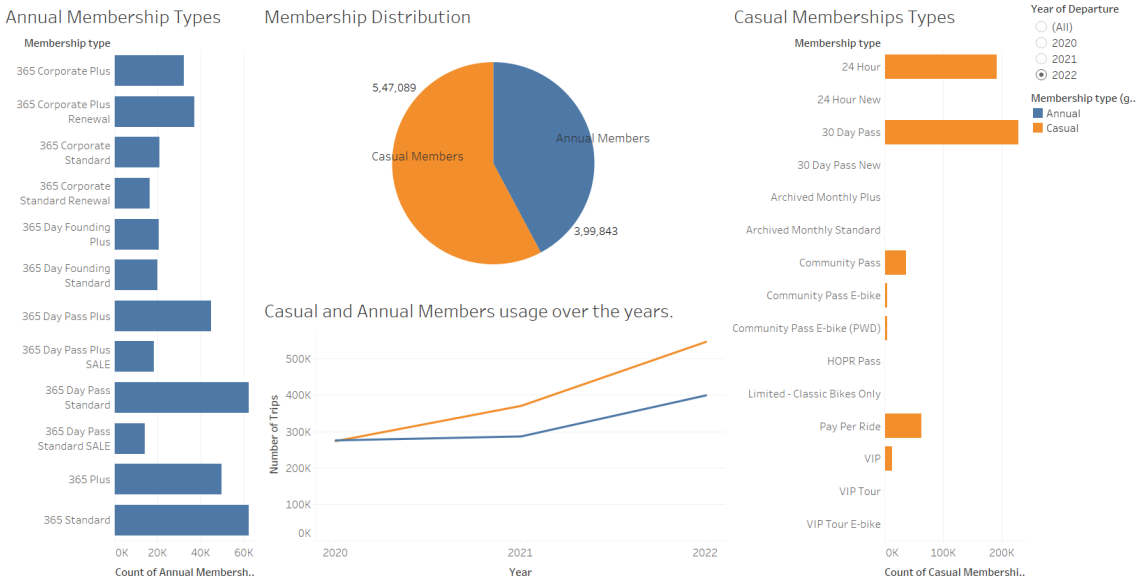
1. Languages: Python, SQL
2. Frameworks and Libraries: PySpark, Apache Spark, Spark ML, Spark DataFrames, Pandas
3. Software Tools: Amazon S3, Amazon EMR, Amazon EC2, Git, Tableau, Jupyter Notebook

Pyspark was very useful in performing the exploratory data analysis and building Machine Learning and ETL pipelines for Tableau visualizations. It helped us in creating scalable analysis and pipelines for the huge bike dataset for past years. Other benefits included In-memory computations, Fault Tolerance, and its dynamic nature. We further used AWS S3 bucket to store this data and the Pyspark code for its budget friendly high scalability and durability. The ETL code was run on the EMR cluster which provided elasticity, reliability and security for the data processing.

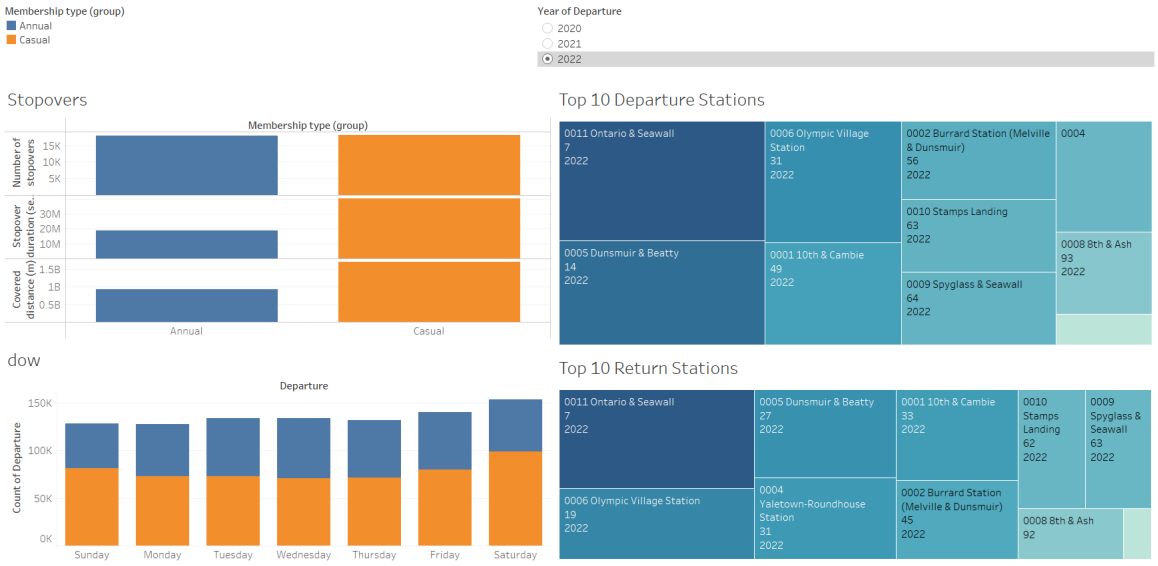
Visualization
Our visualizations were done using Tableau which assisted us in answering the business queries in our goals. We used Tableau because it handles copious amounts of data without impacting the performance of the dashboards and allows performing complex tasks.Tableau dashboards are mobile friendly and can be viewed and operated on several devices such as laptop, mobile, or even tablets. Tableau automatically detects the device being used and makes adjustments accordingly.

You can view the complete interactive dashboards here- Tableau Dashboards

- To understand how casual riders and annual members use Mobi Shaw bikes differently.



From our analysis of the datasets from the last 3 years, we observed that there were more casual members than annual members. The most popular Casual membership is the 30-day pass and the most popular Annual membership type is the 365 Plus. The number of casual and annual members, both had seen an increase in the number of members over the years, showing popularity among bike share. We even observed that several new stations were installed this year and the previous year as well showing a clear growth in popularity.



The stopovers made by the casual members were more than those made by the annual members.

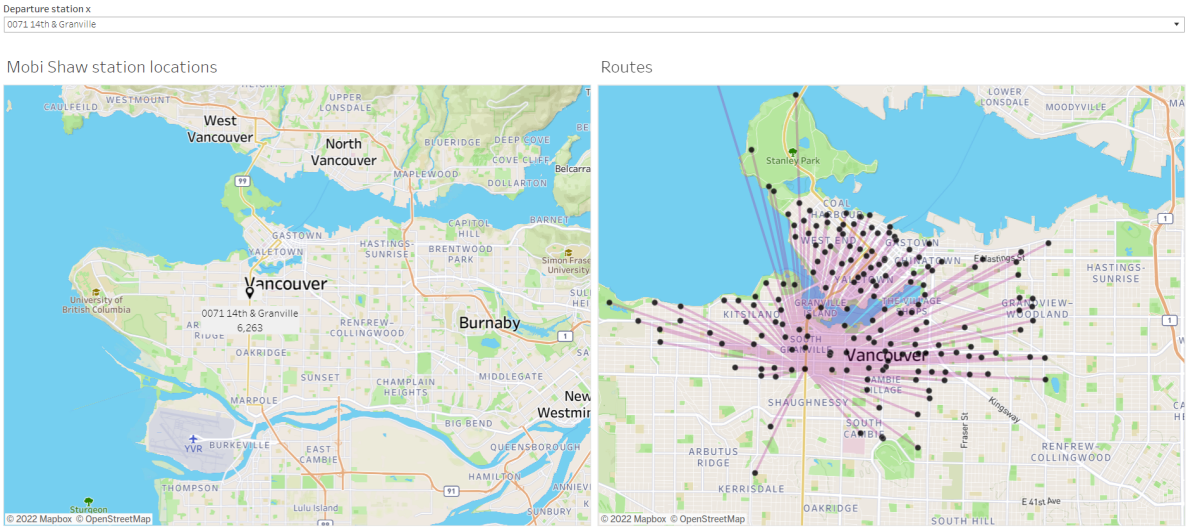
- To maximize the number of annual members.

We recommend Mobi Shaw market its seasonal and annual memberships over the summer and fall seasons in order to increase the number of annual members. Students and the working class could be the target groups. Mobi Shaw already has corporate plans that are pretty popular among their annual memberships.

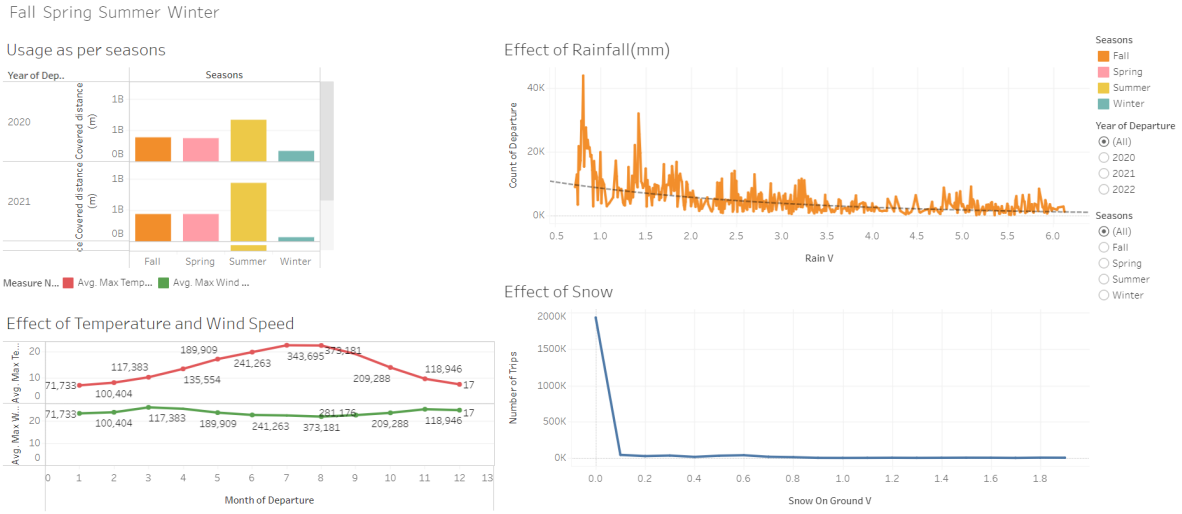
We observed that- Ontario and Seawall, Olympic Village, Yaletown Roundhouse Station, and Burrard Station are quite famous as departure and return stations as well. These stops could use more bikes and frequent repositioning of bikes is required here as well.

- To understand and plan the repositioning of bikes according to the demand and supply in order to increase the average utilization of bikes.

There were a lot of popular paths and bike routes, but most of the famous bike routes were in and around Vancouver downtown. And we even observed that Vancouver's downtown stations were the popular departing and returning stations. As we moved a little further away, the count of returns increased, but the number of departures was not that high, indicating that repositioning was required.

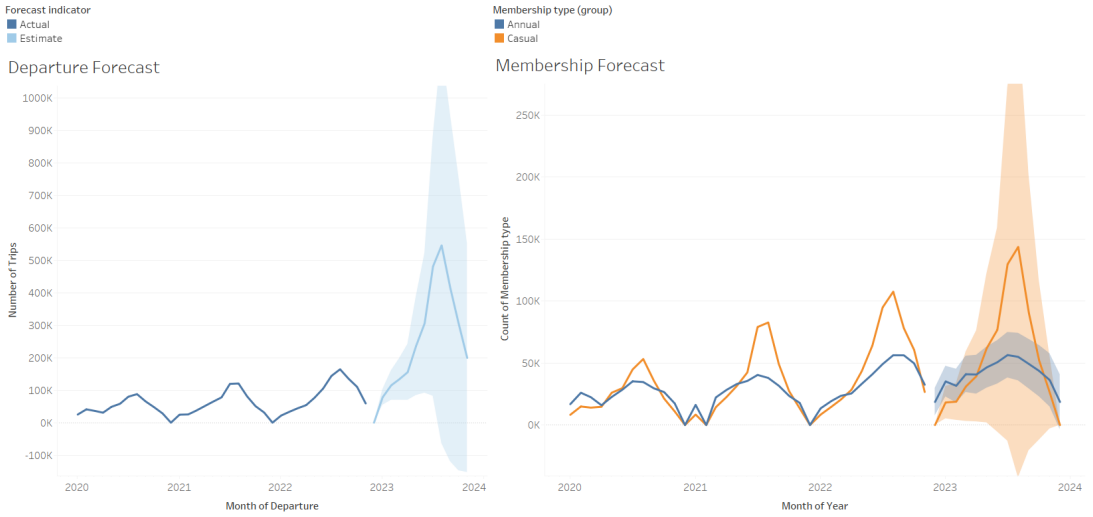


- To analyze how the temperature, wind speed, and seasons affect bike usage.



According to our analysis, people enjoy riding bikes during summer the most, followed by the fall, spring, and winter seasons. We can observe through the above visualizations that the trips decrease as the snow or rainfall increases. With wind speed, we can see that during winter months the departures are less as the weather is chilly whereas during warm months the departures are double in numbers.

We have also created a Tableau forecast model to predict/analyze the trends of number of trips over the next year and membership forecast of annual and casual members. We observe that there is a similar hike in the year 2023 following the month's trend we observed previously. We even observe that annual and casual members shall be increasing leading to more usage over the years. This prediction sits with the fact that Mobi shaw bikes over the years have gone increasing their number of stations as their popularity has increased.



Problems:

One of the problems we encountered was visualizing our repositioning of the bike's goal in Tableau. To represent stations on the map as a visualization, we wanted either the complete address, the pin code, or the latitude and longitude of the station. This was not present in Mobi Shaw's Raw Data.

- Using the pin code approach:
  We sourced a dataset online that had the pin codes for most of the stations but this approach was not successful as Tableau was only taking the first three characters for the location, which made the map generalized. Stations having the same first three characters were grouped by Tableau and the count of departures was summed up for these stations.
- Using the Latitude & Longitude approach:
  We made our dataset of Departure Station, Return Station, Latitude, and Longitude with the help of Google Maps & Mobi Shaw Raw Data. This helped us visualize the map in Tableau as we were getting the accurate station locations.

Another major problem we encountered was making the connection (pipeline) between S3 Bucket and Tableau. We were using the tableau public version and realized a lot of the connector options were unavailable. We then tried using the tableau desktop version, which provided us with a plethora of connectors like Hevo but didn't have the one for the S3 bucket. Another option we tried was using the Amazon Redshift connector provided by tableau, wherein we had to extract the final data into Amazon Redshift and further connect it to Tableau for pulling the data. While doing this we faced a lot of errors and hurdles and the configuration was the biggest one of them.

Results:
Main insights and finding conclusions:

- Casual members hold the highest proportion of the total rides.
- Every month, we have more casual members than annual members.
- The weekend has the highest data volume for casual users.
- The vast volume of bikes used was during Summer and Fall.
- The amount of snow affected the number of rides by a huge amount. Due to covid or not being available a lot of bike stations were not used during 2020 and 2021.

Project Summary:

- Getting the data: 2
- ETL: 4
- Problem: 3
- Algorithmic work: 2
- Bigness/parallelization: 3
- UI: 0
- Visualization: 4
- Technologies: 2

Tableau Dashboards:
The Tableau Dashboards are available on Tableau Public and can be accessed with the below link.
https://public.tableau.com/app/profile/inderjeet.singh.bhatti/viz/Mobi-ShawAnalysis/Story1

Dashboard 1: Behavior Trends of users as per membership types
Dashboard 2: Stopovers & Popular Stations
Dashboard 3: Mobi Shaw Station Locations & Routes
Dashboard 4: Weather affecting the usage of bikes
Forecast Dashboard: Forecasting growth of members as well as trips.

Video Presentation:
The Presentation for our Project can be accessed at below link:
https://drive.google.com/drive/folders/1fHlwhwYc3W-pyujbtNykRBT_IqMlReCU

Mobi Shaw Dataset:
https://drive.google.com/file/d/1ILN_LhfbIcYGZOE0LZfeV1wNxJB572Vh/view?usp=share_link