Exploratory Data Analysis of Montreal BIXI Bike Sharing Data and Historical Weather Data - Final Report

Pushkar Sinha

Table of Contents

1 Overview

This report starts with the scope of this project being exploratory data analysis of Montreal BIXI data merged with historical weather data giving insights over behaviour and different patterns of the bike users, followed by the audience of this work who are likely to get benefited with these insights or offer help to improve the bike sharing system. This is followed by the questions which are likely to be asked and cover the scope of the project. Then the data, i.e. its sources, dimensions, type of these dimensions and cleaning of data which will contain the whole process of data preparation before plotting it over the charts.

Visualization design will follow next to explain the steps taken to answer the questions asked with the proof of data over an interactive visualization environment. The design explain the choices of plots and the axes to represent the data followed by the interactions which show the changes in the plots based on the subset of data selected. Design is followed by future work which includes further visualizations to support the design and the scope of this project but have not been implemented in this work and conclusions to reassure that this project has been able to answer the questions under the scope. Finally the references section mentions the those sources without whose help this work wouldn't have been possible and appendices section gives the steps to run the main project as well as the structure of the notebook with sample visualizations.

2 Introduction (Description of Criterion)

Being a 4 year long data available for Montreal Bike sharing system, it holds a huge scope to infer the salient patterns and behaviour of the bike users. As North America's first largest Bike sharing system it provides an essential mode of transportation to the residents of Montreal city.

2.1 General Scope of this project

To explore the different patterns and behaviour of the bike users through exploratory data analysis of the BIXI bike sharing data coupled with weather data. We would get the patterns differentiating members from non-members, good from bad weather as well as obtaining information of different stations with respect to the previous.

2.2 Audience

- **Tourists** as well as those **moving to the famous city of Montreal** are very likely to use the BIXI bike sharing system and would find the insights from this project very helpful.
- The **policy makers** would be able to augment such a prominent transportation system with the help of different patterns and behaviour pruned

out of the data and represented through visualization.

- Further **private business owners** can bring in mutual benefit plans to generate revenue and job opportunities adding to the economy.
- As the design being big, there would be more questions and further development of this project possible any enthusiastic **individual in visualization domain** to not pick it all up from the scratch.

3 Problem/Domain Questions (Description of Criterion)

1. What are the rush hours and the most frequent stations in those hours. How many average bikes are available in those hours? **(tourists, policy makers and private business owners)**
2. How members and non-members differ in counts and average duration in these rush hours? **(private business owners)**
3. How does the members differ from non-members based on the popularity of stations over duration of bike taken and frequency to visit those stations? **(policy makers and private business owners)**
4. How does the above mentioned visualizations differ in a non-rush hour or a different rush hour? **(tourists, policy makers and private business owners)**
5. How frequencies change due to a bad weather compared to a good one? What are the hours in which the bike travels continue even in a bad weather and whether those are member or non-members? **(policy makers and private business owners)**
6. If there are people travelling at high wind speed are they more likely to be members or non-members and what kind of hour they most likely travel? **(policy makers and private business owners)**
7. What are those stations being used even in a bad weather and what are those routes being taken from start station to the end ones? **(policy makers and private business owners)**

4 Data

4.1 Visualization 1

Data in this visualization is only the BIXI Montreal data and doesn't include the weather data.

- **Source :** kaggle.com

- **Description :** BIXI Montreal (Public bicycle sharing system)- data on North-America's first large scale bike sharing system.
(url : https://www.kaggle.com/aubertsigouin/biximtl/activity), 12 million rows with and 7 columns for main data while 546 rows and 4 columns for station data.

- **Dimensions :-** Main Data- `start_date` : Nominal (as kept in hours), `start_station_code` : Nominal, `end_station_code` : Nominal , `end_date` : Nominal (as kept in hours), `is_member` : Nominal, `duration_sec` : Quantitative, as well as the counts will be Quantitative. Station Data- code : Nominal, name : Nominal.

- Cleaning/Preparation

  1. Main data were **unioned** for 4 years as they were available for different years.
  2. Hour were retrieved from the `start_date` and `end_date` to be used as nominal values as they were in timestamp. Station Data was joined over `station_code` to get the names of the stations.
  3. Groupby on the main data over `start_hour`, `start_station_code` and `is_member` taking aggregates as count and mean over `duration_sec` to get `duration_mean`. This helped to get the average counts on a particular hour of the day and for a specific station as well as if those counts were by the members or not. Similarly taking average duration under all these specific values of `start_date`, `start_station_code` and `is_member`. Groupby again on the main data but this time over `end_hour`, `end_station_code` and `s_member` to get the counts. Then these two tables are joined over `start_hour` and `end_hour` and the the field vacancy is created by subtracting end count from the start count which basically means no. of bikes coming to the station minus no. of bikes going away.
  4. Percentile bar which represents the popularity of stations is created by using quantile method from pandas over the average counts obtained from the previous process. This cuts the stations in a specific hour based on the percentile of their counts.
  5. Few more further cleaning has been done like changing seconds to minutes and `is_member` values to 'Y'and 'N'etc.

4.2 Visualization 2

Data in this visualization is both BIXI Montreal data and the historical weather data.

- **Source :** kaggle.com
(url : https://www.kaggle.com/selfishgene/historical-hourly-weather-data)

- **Description :** 6 different tables with 45.3k rows and 37 tables for different weathers i.e. `weather_description`, humidity, pressure, temperature, `wind_direction`, `wind_speed` out of which only three tables i.e. `wind_speed`, `temperature` and `weather_description` has been used.
- Cleaning/Preparation
  1. Different weather tables are joined based on timestamp to get the three columns `wind_speed`, temperature and `weather_description` in one table and only for the city Montreal. Further this table is joined with the BIXI Montreal table over the hour column to make up a one big main table consisting data for 4 longs years.
  2. As this data is also visualized for different hours of the day aggregated over all the days in the data, `temperature` and `wind_speed` had to be made into **buckets** using cut in pandas.
  3. Then a groupby was done over columns 'hour', `'weather_desc'`, `'is_member'`, `'temp_bin'`, `'wind_bin'` and data was aggregated as mean over `duration_sec` (i.e. `duration_mean`) and counts. This means that based on particular hour of the day, weather description, membership, the bin of temperature and bin of the `wind_speed` it was calculated that how many people were using the bikes on an average and how much average time they took in making those trips.
  4. As there were many different types of `weather_description` those which had *rain , snow, thunderstorm and clouds* in their description were converted to only those words as their description respectively.
  5. Further the data was normalized i.e. `wind_speed` was changed from m/sec to km/hr, `temperature` was changed from kelvin to fahrenheit, `duration_mean` was changed from seconds to minutes etc.

4.3 Visualization 3

Same data which was used for visualization 2 but with groupby on different columns i.e. `'start_station_code'`, `'weather_desc'`, `'is_member'`, `'end_station_code'` . This will give all the routes and the stations under a specific weather such that we can find out all those stations and routes which are used in bad weather and what is their frequency and whether whose use it are members or not.

5 Design

5.1 Visualization 1

5.1.1 Components

1. A **scatter plot at the top** to represent different stations along different hours of the day and their counts. Design choices :

- **X-axis :-** hour : Nominal. Hour is chosen nominal as different hours along the day was to be considered as different groups to see how change in hour changes the scenario.
- **Y-axis :-** counts : Quantitative. Counts being in the y-axis juxtaposed to hours over the plot of different stations give the position to those stations based on how frequently they were used as well as picks out that hour which has high/low frequent stations.
- **Color :-** Color doesn't depend on any column values but if an hour is selected then rest of the plotted station fades to grey as there is an interaction involved in that choice.
- **Tooltip :-** Tooltip contains the name of the station hovered on.

2. Another scatter plot just below the previous scatter plot to represent the selected stations based on the hour plotted against vacancy and `duration_mean` to show the stations with average number of bikes present in those stations in the hour selected and the average duration taken by the bikers in this scenario. Design choices :

    - **X-axis :-** vacancy : Quantitative. These values show the average number of bikes likely to be present in the chosen hour and it is fractional and also with negative values. This scale can easily represent those stations which has average excess/deficit number of bikes in the chosen hour.
    - **Y-axis :-** `duration_mean` : Quantitative. These are the values which represent the average time taken by the rider while making the journey. This scale can well different between those who took less time and those who took more along their journey.
    - **Color :-** `is_member` : Nominal. There are only two i.e. members or non-members and can easily support the difference on the plot.
    - **Tooltip :-** `station_name`, counts (As average number of bikes taken).

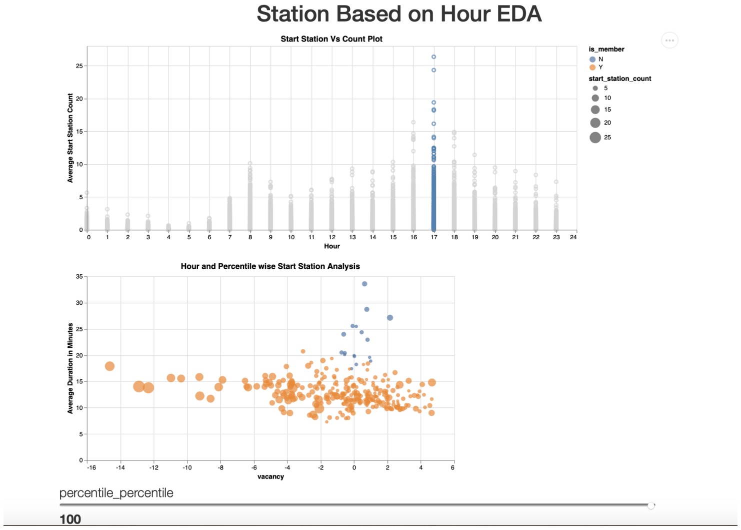
3. A range bar from 10-100 with steps of 10 representing the percentile of the counts of the stations plotted.

5.1.2 Interactions

- The first plot is linked with the second plot just below it, in which when an hour is selected stations only in those hours are plotted on the second plot. With this selection the rest of the plotted points in the first plot greys out.

- The range bar is connected with the second plot which when dragged horizontally changes the plotted station in the second plot based on the percentile selected in the range bar.

5.1.3 Screenshots of plots and how they relate to the questions asked (Visualization 1)



percentile_percentile

100

Figure 1: plot for most frequent hour and most popular stations

- **(Figure 1)**Here on the top graph, the most frequent hour is at 5pm which connects to the plot below showing that journeys are mostly made by the members rather than the non-members. **(question 1)**
- **(Figure 1)** Plots on the second graph can be hovered upon to see those stations and the x-axis would represent the average number of bikes available in that hour in those stations. Here they are very high with 3 stations on the left side having the average vacancy value of -15 which means that 15 deficit bikes are there on an average in those three stations at 5pm and one is rather unlikely to get a bike at that time. **(question 1 and 2)**
- **(Figure 1)** Based on the average duration we can see that the members are very steady compared to the non-members even with a much higher count. This shows that members travel those routes repeatedly and are stabilized.**(question 1)**
- **(Figure 1)** Based on the percentile bar we can see that these plots are for the most popular stations. **(question 3)**
- **(Figure 2)** Here the change that has been made from the previous figure is the percentile bar which has been dragged down to 50 percent which means the stations represented and much less popular as we can see in the the graph just above it having non-members much more than the members. **(question 3)**

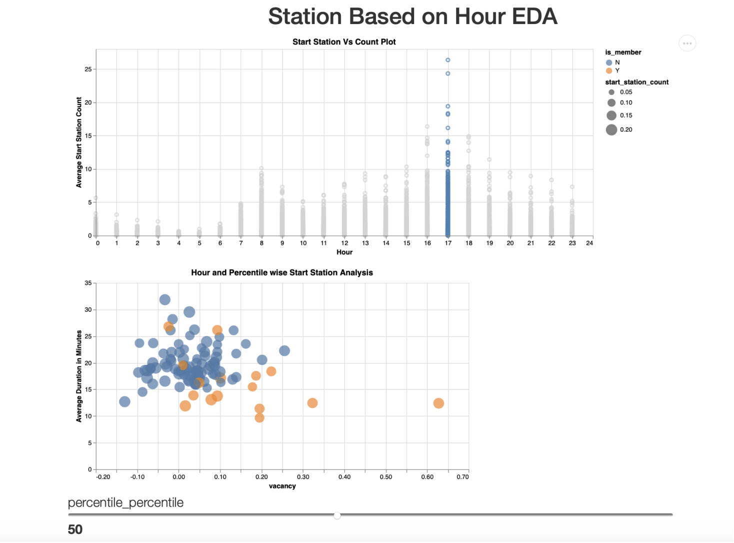


Figure 2: plot for most frequent hour and stations with 50 percent popularity

- **(Figure 2)** Even after the members being much less than the non-members, the time taken by members are still less showing that they are taking routes with which they have been customised to. **(question 3)**
- **(Figure 2)** Another peculiar observation here is the severe drop in counts if we see the axis representation as well as drop in vacancy. This can be considered quite related and usual due to the drop in popularity of the stations. **(question 3)**

- **(Figure 3)** As one can see here that the percentile has all the way been dropped to 10 percent and we are left with only non-members as well as the counts and vacancies have dropped to minimum representing the least popular stations. **(question 3)**
- **(Figure 4)** Here the least frequent hour is select from the top graph and it is very intuitive to see that there are no stations which are as popular as the ones in the 90-100 percentile range. **(question 4)**
- **(Figure 5)** It is really interesting to see that at the least frequent hour (**4 am**) the plots are almost opposite to one in the highest frequent hour. At 50 percentile(range bar) popularity of stations members are more than the non-members which means if the hour is not good the likelihood of non-members making trips is almost nil. **(question 4)**
- **(Figure 6)** This plot is at 4am and for the least popular stations. We can see that the counts and vacancies are almost nil but still members take less time than the non-members comparatively. **(question 4)**

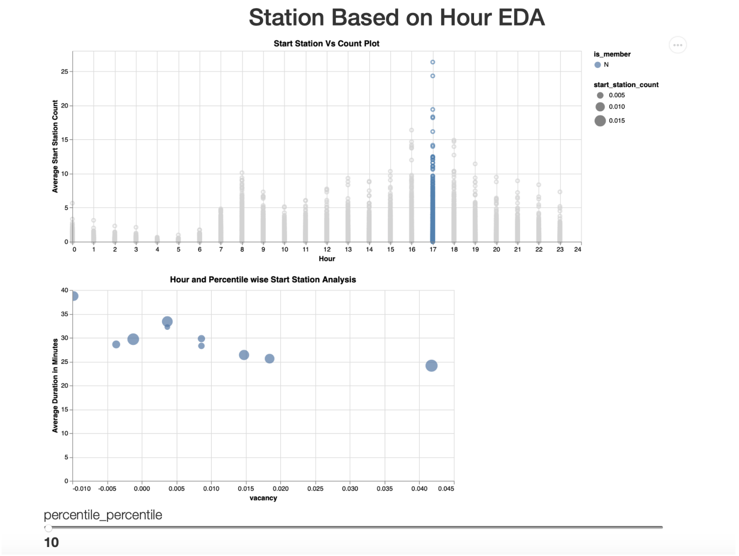


Figure 3: plot for most frequent hour and least popular stations

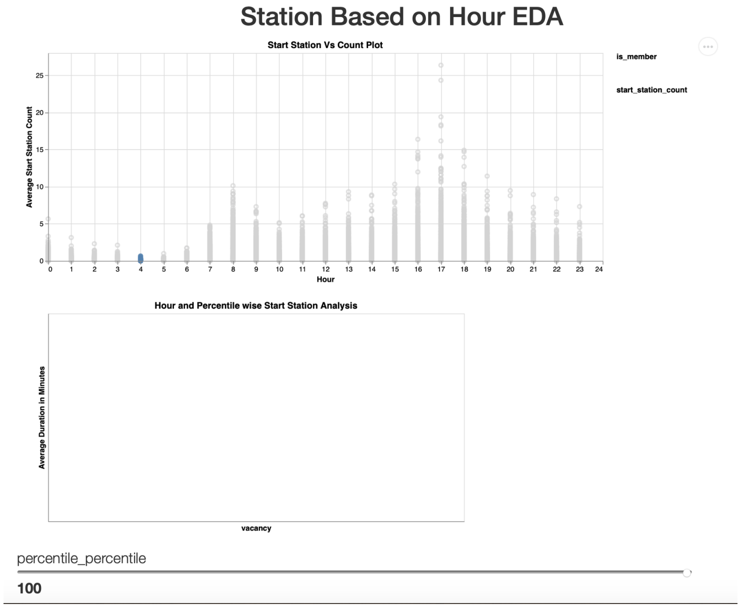


Figure 4: plot for least frequent hour and most popular stations

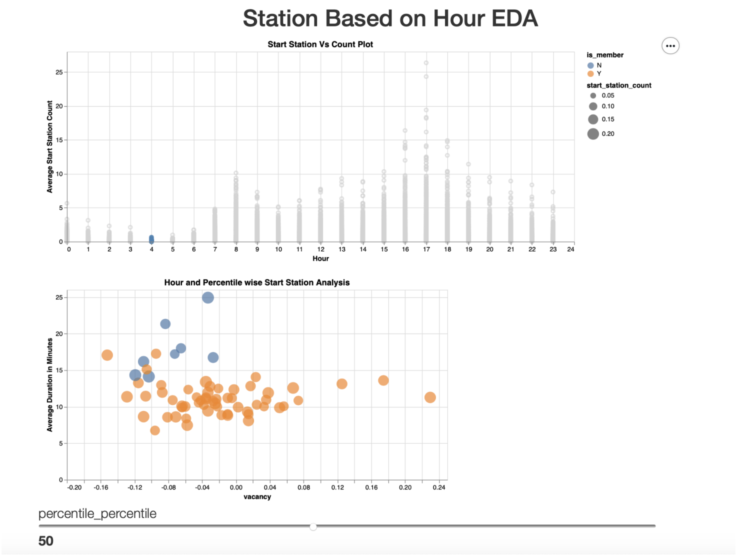


Figure 5: plot for least frequent hour and stations with 50 percent popularity

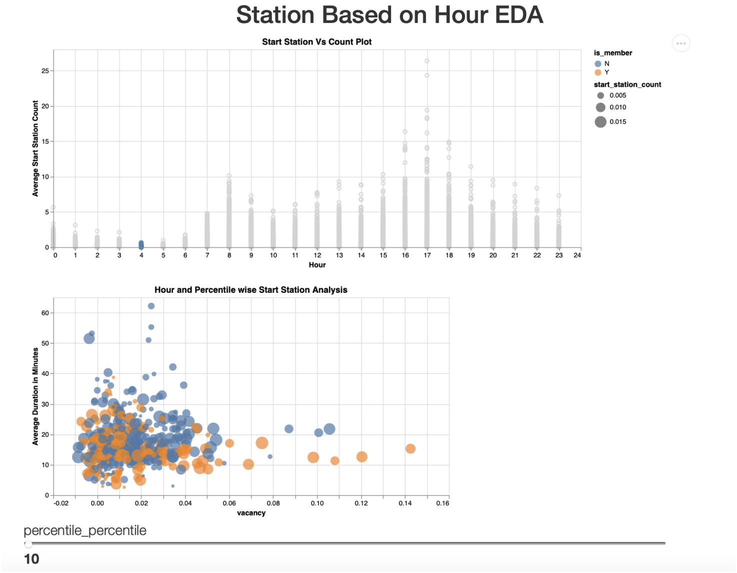


Figure 6: plot for least frequent hour and least popular stations

5.2 Visualization 2

5.2.1 Components

1. A **scatter plot on the left** which selects the plots which representing the different weather against the temperature bucket and the counts. If a

point is selected all the points with the same weather is selected :

- **X-axis :-** `Temperature_bin` : Nominal. When a particular weather is chosen this can show the temperature buckets with the highest to the lowest counts. A high count in a particular weather show that the particular temperature range is most favourable by the bikers.
- **Y-axis :-** Count : Quantitative. Give the counts.
- **Color :-** `Weather_description` : Nominal. When selecting the weather all the plotted points with same weather description are highlighted with the same color. This makes the user feel that rest of the plots come under this particular weather.
- **Tooltip :-** count : Quantitative, `weather_desciption` : Nominal

2. A **range bar at the bottom** showing hours from 1 to 24 and filters the data based on hours on all of the plots.

3. A **scatter plot on the right** which represents the member and non-members based plotted points plotted against `wind_speed` bucket and average duration in minutes. This shows the plotted point under the selected hour and weather to show the members or non-members to make the journeys even at higher wind speeds.

   - **X-axis :-** `wind_bin` : Nominal. Shows the speed of the wind under a particular weather and hour.
   - **Y-axis :-** `duration_mean` : quantitative. Showing the average time of the trips by the bikers.
   - **Shape :-** `is_member` : Nominal.
   - **Tooltip :-** average duration, temperature bucket, weather description, hour

5.2.2 Interactions

- If a point is selected in the left scatter plot, the data filters on the right scatter plot and shows only the ones under the selected weather.

- When the hour slider slides, data is filtered on both the scatter plots but the weather doesn't change in the left plot thus we can find on a particular hour in a particular weather the changes in the right plot.

5.2.3 Screenshots of plots and how they relate to the questions asked (Visualization 2)

- **(Figure 7)** Above the weather selected is rain and the most frequent hour 5 pm and on the right side plot we can see mostly the members making the journey even at `wind_speed` as high as 48 km/hr. **(question 5 and 6)**

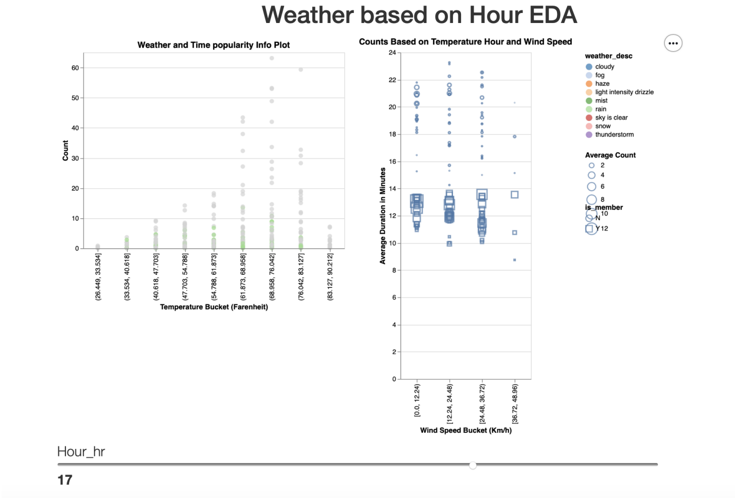


Figure 7: plot for a rainy weather and most frequent hour

- **(Figure 8)** The hour has been changed to the least frequent 4 am from the previous plot and we can see that the counts have dropped and still there are more members than non-members. This deviates from the previous visualization indicating that members travel even in bad weather as they going to their everyday work/school/college. **(question 5 and 6)**

- **(Figure 9)**Here the a pleasant weather is selected i.e. sky is clear and we can see the counts going really high. Further the most favourable temperature here is 68-76 degree fahrenheit. **(question 5 and 6)**

5.3 Visualization 3

5.3.1 Components

1. A **scatter plot on the left** which selects the plots representing the different stations against the weather and the counts. Here a interactive brush can be dragged and an area can be selected :

    - **X-axis :-** `Weather_description` : Nominal. When an area is selected these weather(s) will be considered for the plotted points.
    - **Y-axis :-** Count : Quantitative. Give the route counts.
    - **Color :-** `Weather_description` : Nominal. Selected points under the area will be highlighted and rest will be faded grey.
    - **Tooltip :-** `Start_station` : Nominal

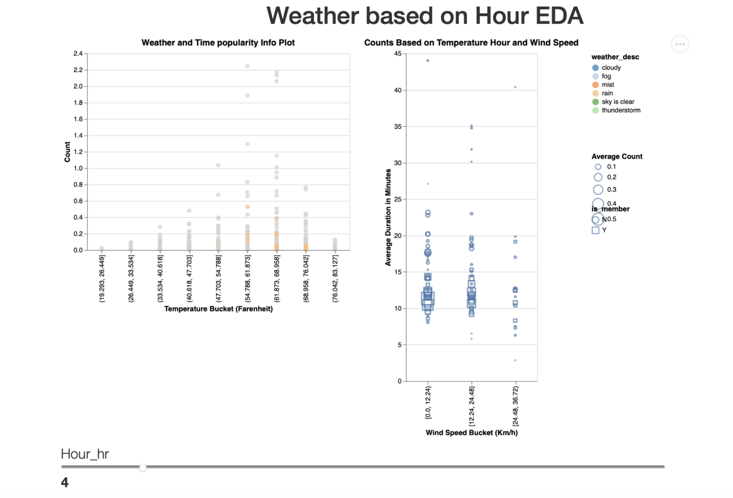


Figure 8: plot for a rainy weather and most frequent hour

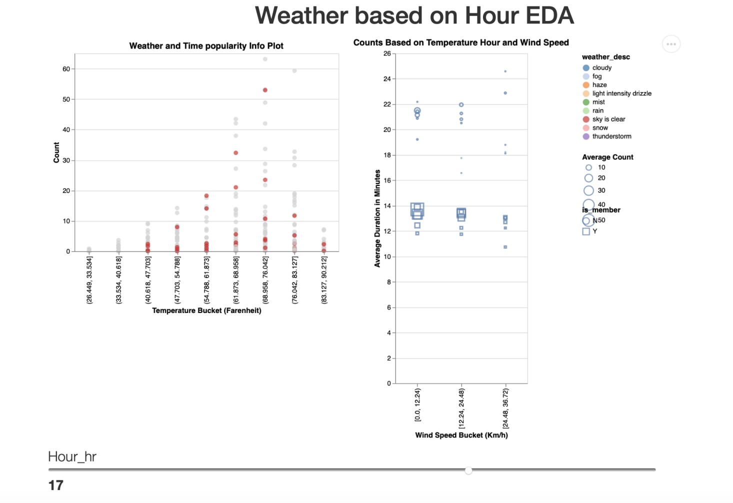


Figure 9: plot for a rainy weather and most frequent hour

2. A **horizontal bar graph** on the right which represents the routes i.e. `start_station`,`end_station` and the average time taken to cover the route . This shows the station pairs which are used in a selected weather.

   - **X-axis :-** `duration_mean` : quantitative. Showing the average time of the trips by the bikers.
   - **Y-axis :-** `end_station_name` : Nominal.

- **Tooltip :-** `start_station_name`, `end_station_name`.

5.3.2 Interactions

- After an area is selected using brush including the station plotted points for different weather(s) the data is filtered to show those start and end stations on the right bar plot with average duration taken for each pair of stations.

5.4 Screenshots of plots and how they relate to the questions asked (Visualization 3)

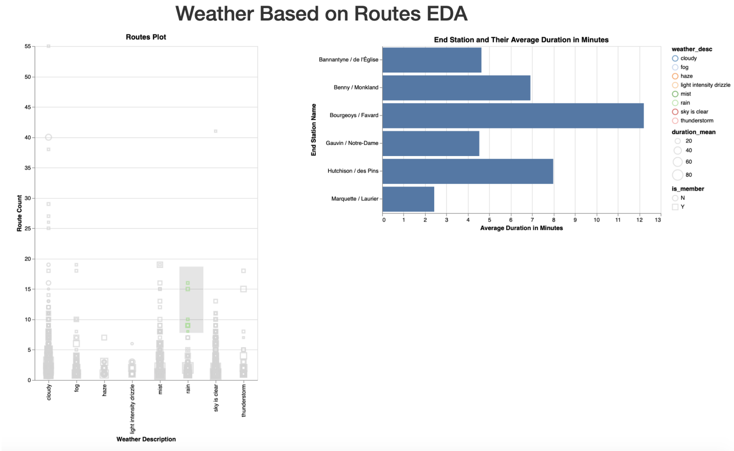


Figure 10: plot for a rainy weather and most frequent hour

- Here *rainy* weather plots are selected and we can see those routes as pair of stations when we hover on the right bar plot.

6 Further Work

- **Intend to do but didn't complete :** This exploratory data analysis doesn't differentiate between working and non-working days which can bring out many important insights. This can be done just by keeping an extra column which can be convert from the `start_date` and `end_date` to a working day or a non-working day. Similarly month-wise analysis will also give interesting results.
- **What didn't work :** to get the plots based on different stations and different weather for every hour was didn't work as number of records

increased to 500,000 and it was difficult to render on the browser, but a subset of this data has been visualized in the third visualization. This can be handled if asychronous AJAX can be made to bring only that subset of data which is to be shown in the plots instead of sending all the data beforehand.

- **Next version :** As I was able to get real estate property data for every city in canada with exact coordinates , this data could also have been merged with BIXI bike sharing data as it has the coordinates of the different stations. A density based clustering can be applied to cluster based on coordinates and highlight those stations which has more houses in their cluster. This could give insights that stations nearby more houses are more frequent or vise-versa.

7 Conclusion

After the exploratory data analysis of Montreal BIXI data with the historical weather data, we have come up with 3 different visualizations with numerous insights answering 7 different (or possibly more) questions pertaining to the biker's behaviour and their biking patterns.

8 Appendix

8.1 How to run the main project

1. Download and Unzip the `vis_project.zip` folder and navigate to that folder through **terminal**.

2. Execute these commands (Please install pip in your PC if it doesn't exist already):

    (a) pip install altair

    (b) pip install flask

    (c) `FLASK_APP=first.py flask run`

    (d) Navigate to http://127.0.0.1:5000/ (Please make sure that any other application is not running at http://127.0.0.1:5000/)