NYC Citi Bike Data Analysis

Prathiba

30/08/2021

1. Documentation

Domain knowledge (1.1)

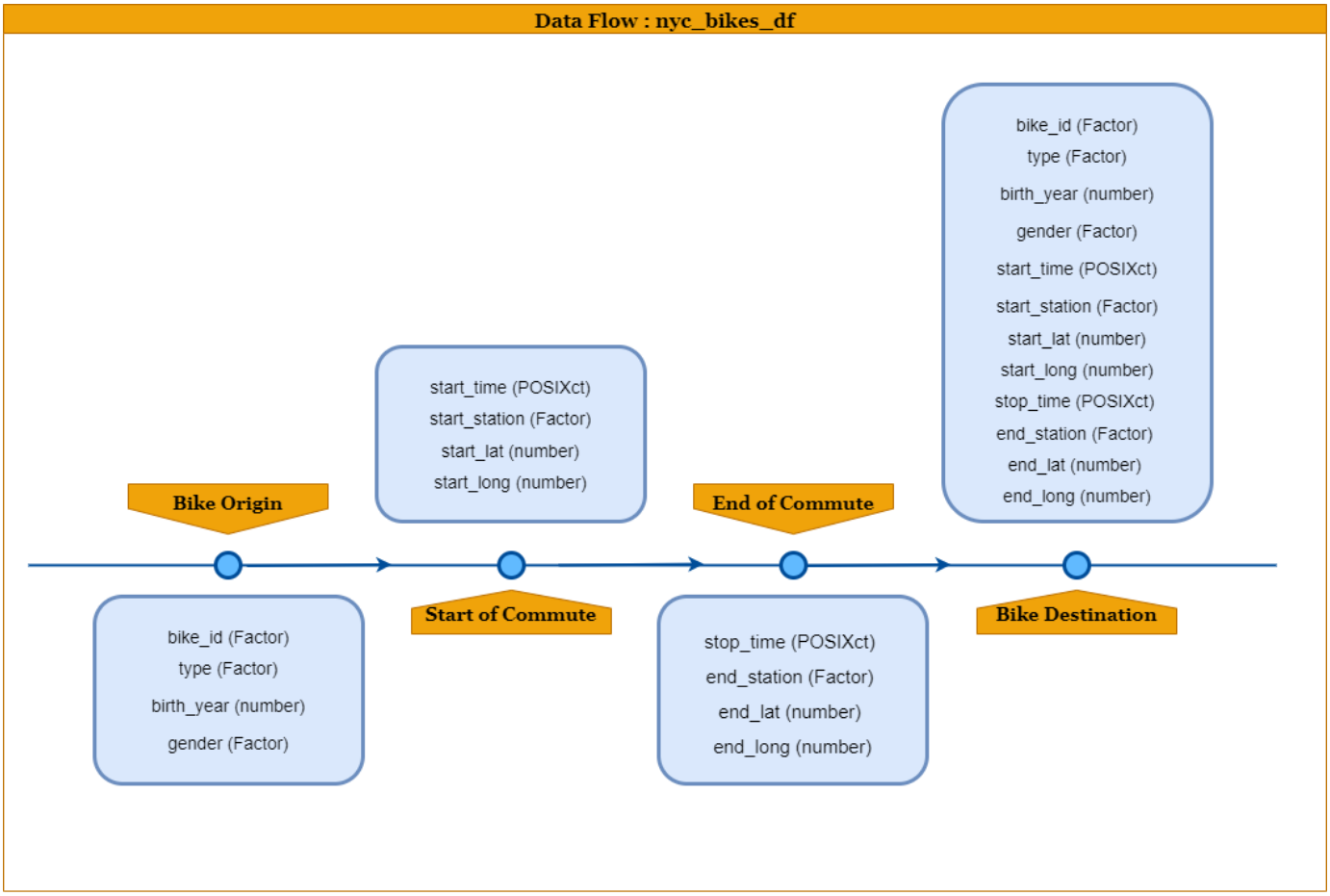
Citi Bike is New York City's bike share system, and the largest in the nation. Citi Bike launched in May 2013 and has become an essential part of the transportation network. It's fun, efficient and affordable. The bikes can be unlocked from one station and returned to any other station in the system, making them ideal for one-way trips.Citi Bike is available for use 24 hours/day, 7 days/week, 365 days/year. You can read more about the company the data comes from on the NYC Citi Bike website (https://www.citibikenyc.com/about)

Business requirements (1.2)

The bike sharing system is becoming more popular in the process of greener initiative. This report helps to

- Understand the pattern of bike hires over a period of time to identify when it is widely used.
- Analyse the pattern of bike hires with respect to the rider demographics such as age, gender or the type of the user.
- Identify the geographical spread of the starting points of bike hires

Business processes and data flow (1.3)



This diagram was created using draw.io at https://app.diagrams.net (https://app.diagrams.net)

Data visualisation as a tool for decision-making (1.4)

One of the main challenges in the bike sharing system is to effectively plan the resource usage. This report helps to visualise the traffic level at each station on a daily basis. Based on the traffic level, the demands can be predicted and the number of available bikes at that station can be intepreted or forecasted.

Data types (1.5)

Attributes	Description	Data Types
bike_id	Unique ID for the bike	Factor
start_time	Starting time from the origin	POSIXct
stop_time	Ending Time at the destination	POSIXct
start_station	Origin Station	Factor
start_lat	Origin Station's Latitude	Number
start_long	Origin Station's Longitude	Number
end_station	Destination Station	Factor
end_lat	Destination Station's Latitude	Number
end_long	Destination Station's Longitude	Number
type	Type of user	Factor
birth_year	Birth Year of the user	Number
gender	Gender of the user	Factor

Data quality and data bias (1.6)

The dataset description says, **" A sample from NYC Citi Bike usage of 10 bikes throughout 2018"** which may result in *selection bias*. However we assumed that the sample seeks to accurately reflect the characteristics of the larger group and hence doesn't affect the quality of the data.

2. Data cleaning

Preparing data for visualisation (1.7)

- The time attributes like month, week, day & year are derived from the start time.
- The gender with type "Unknown" are considered as NA as a part of cleaning process.

3. Data visualisation

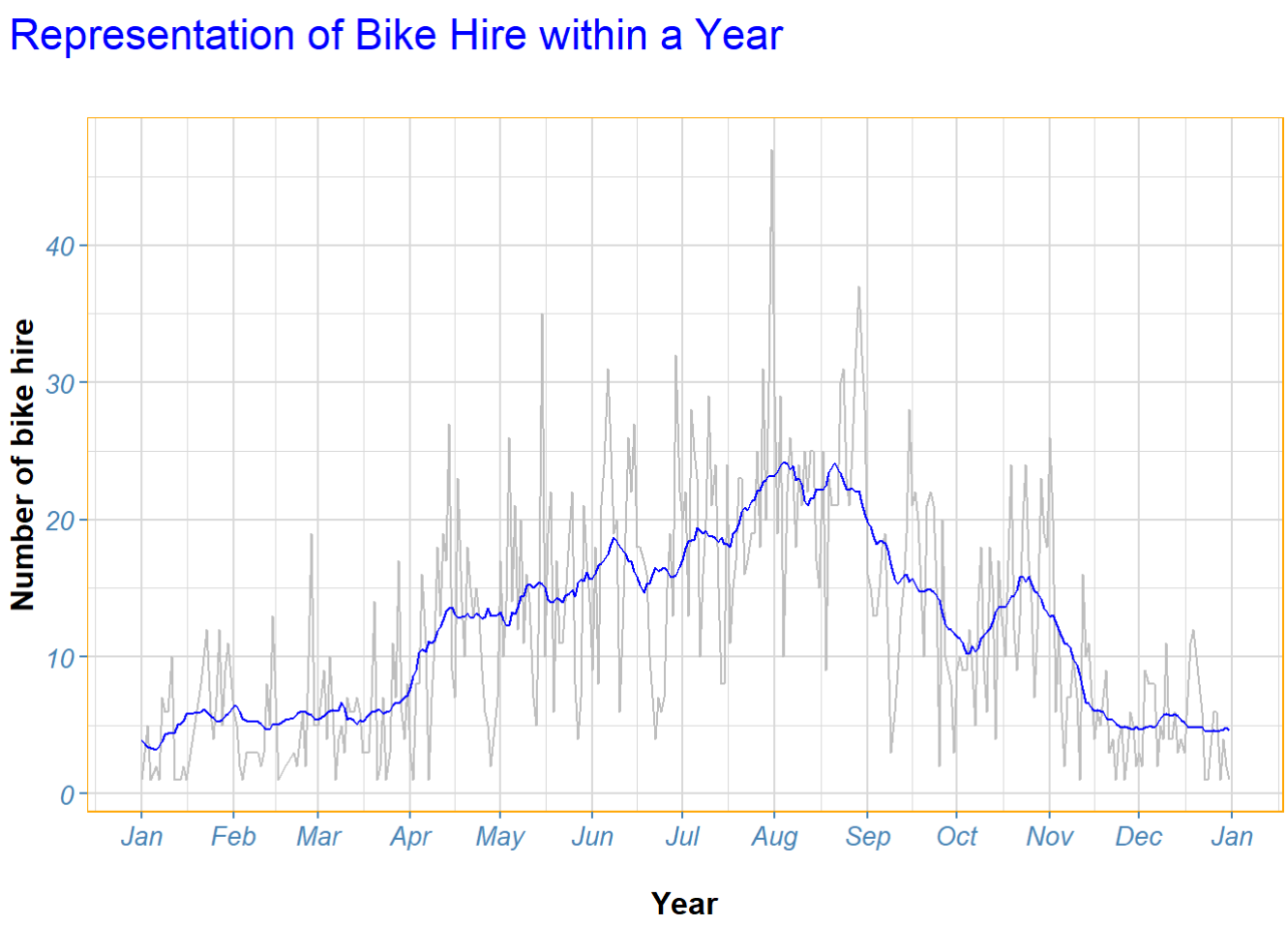
Process and design (2.1, 2.7, 2.8, 2.9)

First, I investigated the bike hire pattern with simple line chart starting with year and later aggregated with month to identify when it is widely used. Then I analysed the bike hire pattern with respect to the rider demographics such as age, gender or the type of the user to see if there were any relation. I have also represented the geographical spread using spatial plots with cluster option. I ensured that the key visualizations are accurate and depicts the clear understanding of the requirement and not misleading. I also wanted to convey that the plots were aesthetically pleasing. All visualizations were made in RStudio using the ggplot and leaflet package.

Visualisations (2.2, 2.3, 2.4, 2.5, 2.6)

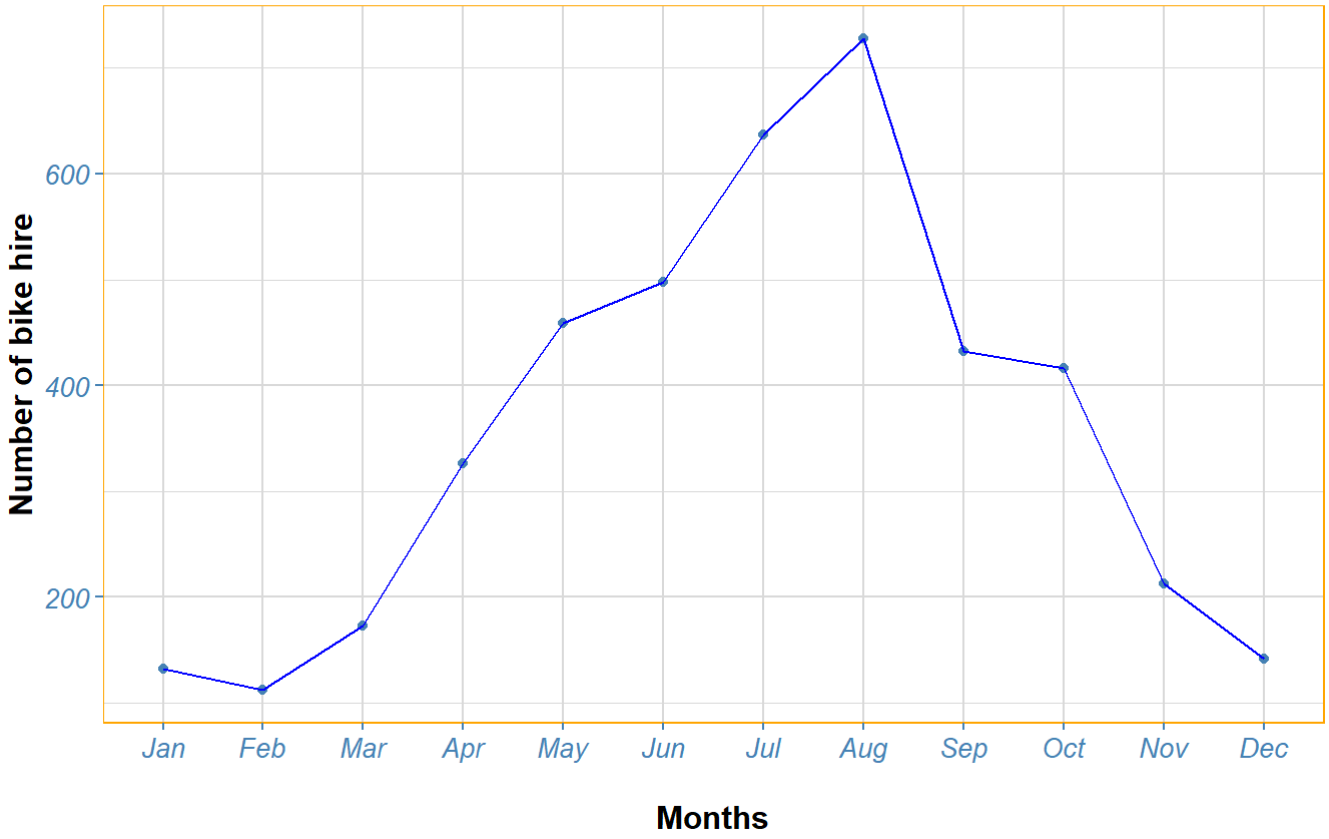
Understand the pattern of bike hires over a period of time

Let's first analyse the bike hire pattern over year.



The above plot shows the representation of number of bike hire on a daily basis within a year where the blue line shows the moving average of the number of bike hire. There seems to be more number of bike hire between July and October. Let's try to aggregate this to month's data and visualize further.

Representation of Bike Hire over months

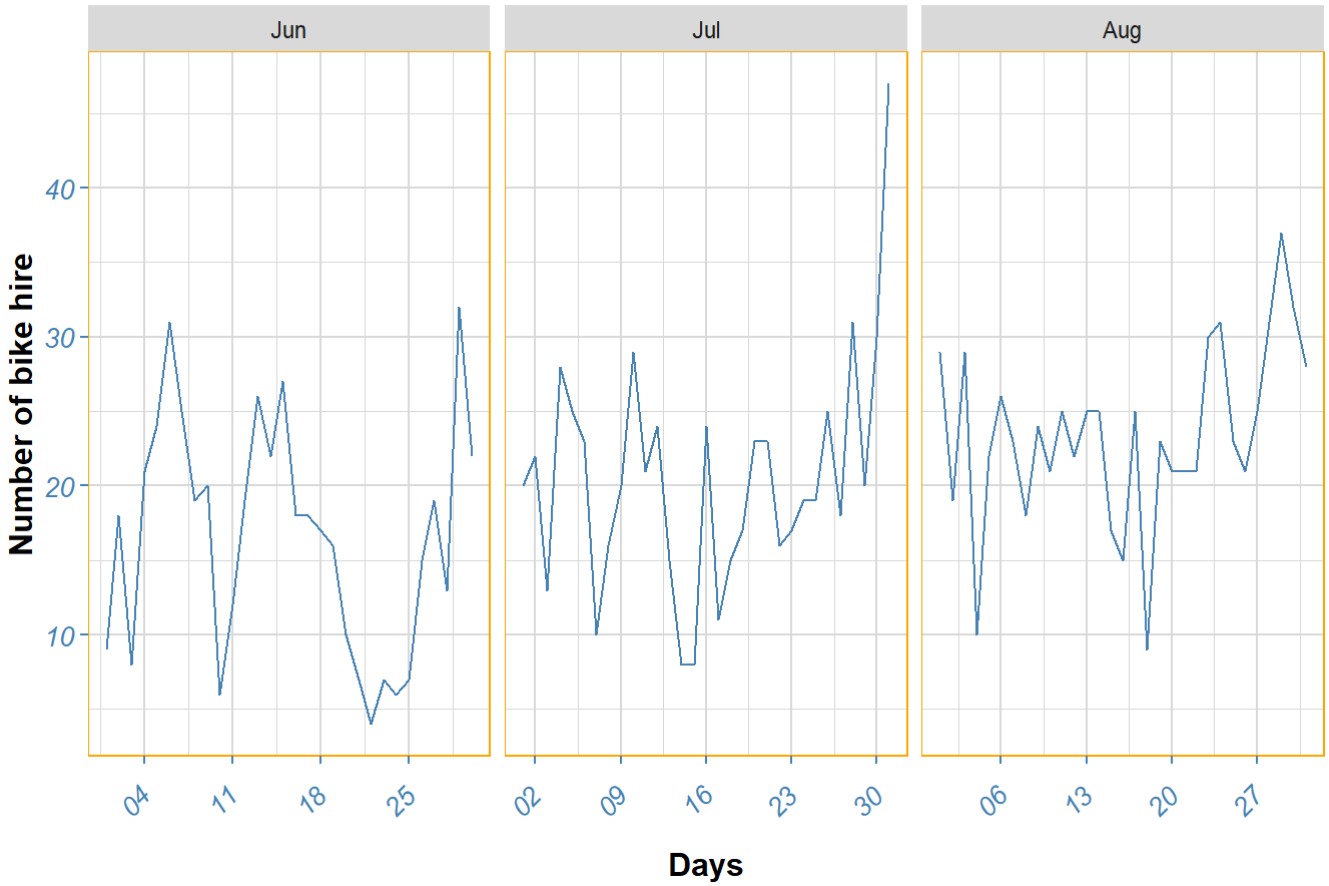


The line chart above shows the pattern of bike hire over a period of 12 month of the year 2018. As month progresses from left to right, points connect the number of bike hire . We can read from the general slope of the line and its vertical positions that the count is improved from March to August , then dropped gradually to December.

This shows that the bike is hired predominantly from Spring to Autumn, with highest being the Summer time.

Let’s further explore the usage of bike over the weeks. Since the number of bike hire is predominantly high between June and August, let’s visualise the data for those month,

Representation of Bike Hire over Days

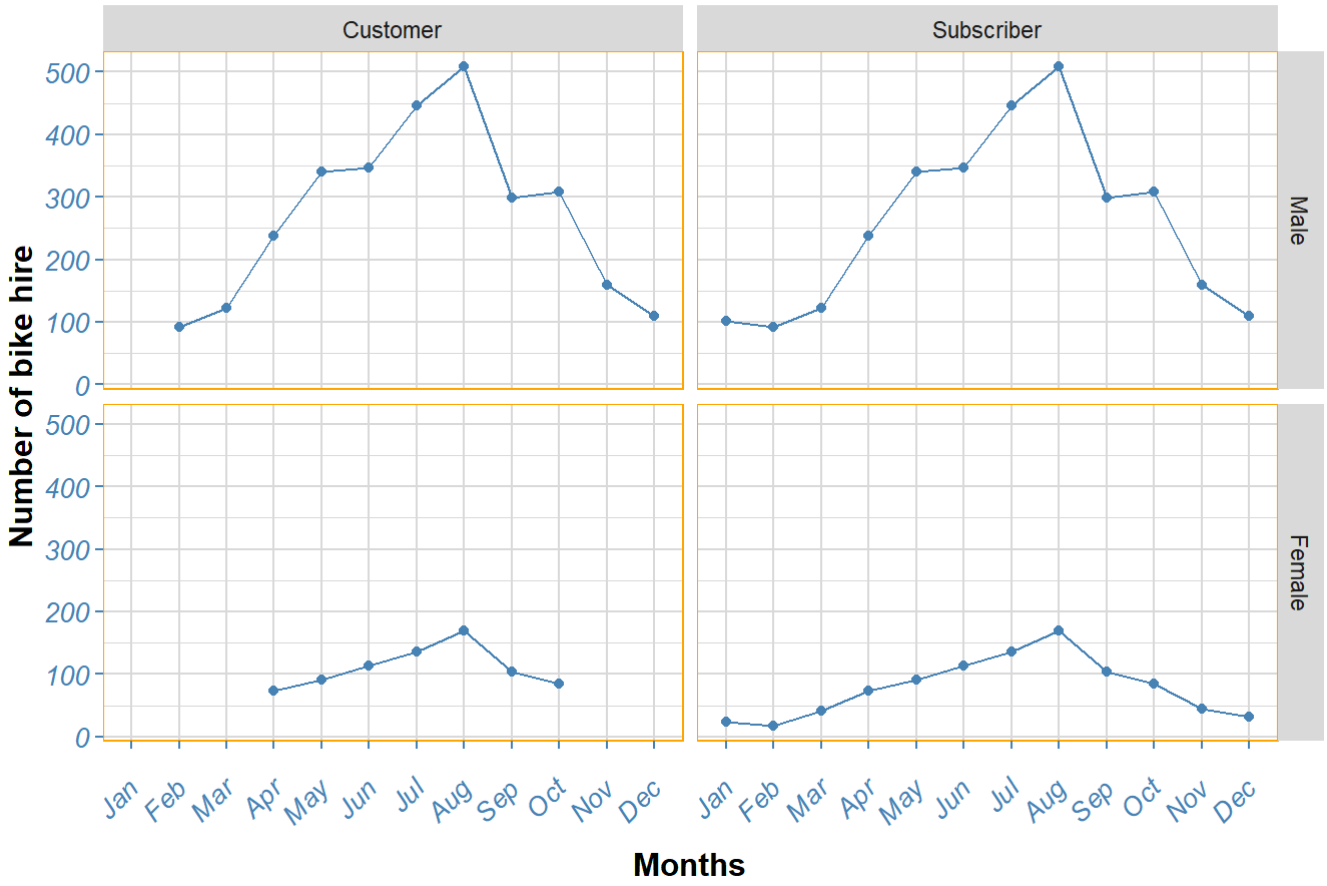


There is an obvious cyclical pattern, with weekdays (particulary the beginning of weeks) having about much traffic than weekends. Further drilling down to hours might give a view on the busiest hours of the day which might be helpful in further optimization.

Analyse the pattern of bike hires with respect to the rider demographics

Let’s check the hire patterns between the bike riders based on gender and the type of bike rider.

Representation of Bike Hire patterns based on Gender and Bike rider



This line plot is a overall representation of the bike hire patterns based on gender and type of bike rider over a period of 12 months for the year 2018.

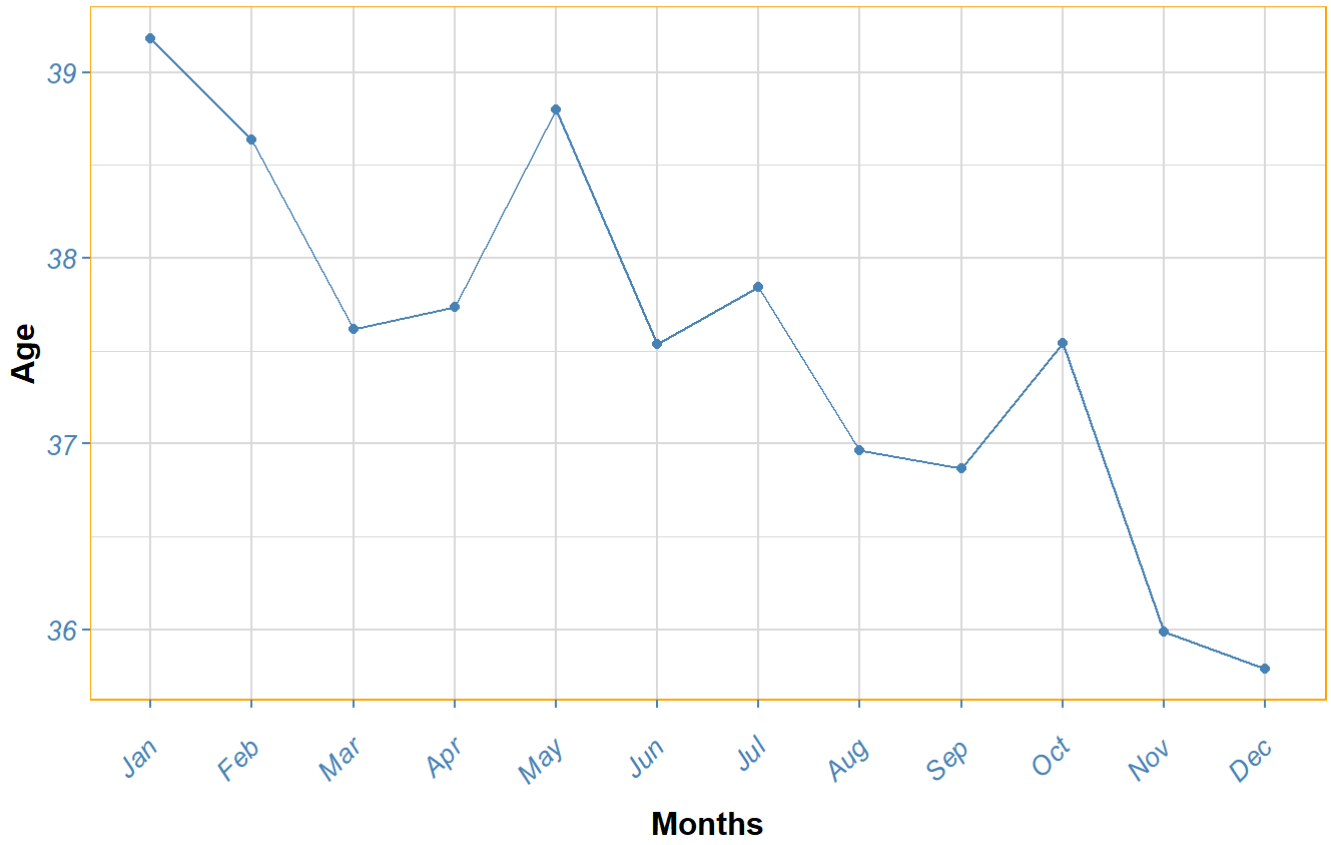
It gives a brief summary stating

- Male tend to hire more bikes than female.
- Subscriber tend to hire more bikes when compared to Customer.

Let’s explore which age group hire more bikes

Representation of Bike Hire patterns based on demographics

(Data taken for the year 2018)



The above line chart shows that on an average, riders between age 36 and 38 hire often.

To summarise, though the bike rider's demographics doesn't seems to have a direct impact on the hire pattern, they are dependent to each other.

Identify the geographical spread of the starting points of bike hires

Now we understood the maximum usage of the bike hire pattern and analysed with respect to the rider demographics, let's analyze the busiest station on a daily basis. It is calculated based on the total number of bike hire per day. This measure helps to identify the traffic level of a particular station.

Since there are more bike hires between June and August, Let's visualize those data using map based display.

The map shows the location of each station grouped with markers together (called as clusters) . Each cluster is identified by its color scale which really helps to visualize the number of bike hire (check-in) at that station. The orange circle indicates the busier station with green being, comparatively lowest traffic level. Based on that, the Jersey City seems to have the busiest traffic,especially Groove Street (to be precise). This helps to give an estimate of how to forecasts the future demand. ie. the minimum available number of bikes at that station.

Conclusion:

The bike usage pattern for the year 2018 is analysed over time. Based on the analysis, the bike hire is most frequently used from Spring to Autumn, especially in the months between June and August. It is also observed that the bike usage is more during weekdays than weekends. Considering the bike rider demographics, male seems to hire more frequently than the female. Also subscribers hire more frequently when compared to customers.Finally the busiest station is narrow down using the clustering analysis from the spatial plot by which the resource usage can be planned more effectively .

To conclude, these results provide a framework for the business planners to assess the current process and to develop more prediction analysis on the number of available bikes at any given station.

Other Insights:

- In our analysis, if we narrow down further to hours the bike was hired, it might help to identify the busiest timings of the day. It gives a better view to optimize the availability of bikes.
- Adding weather conditions to our analysis might give a different insights on the rider’s perspective.