Exploratory Data Analysis on NYC Metropolitan Area

Anusha Muddapati, Harshith Sesham, Deekshit Vedula, Tejeshwine Viswanathan, Loren Young

Introduction

- This project analyzes the correlation and influences that weather has on public bicycling in the New York City (NYC) metropolitan area. Using weather data taken from the National Oceanic and Atmospheric Administration (NOAA) for the month of January, 2022, these influences were examined on bicycle riders under the Citi BikeShare community.
- Through these BikeShare systems, a user can rent a bike from a particular station and return it at another station. Currently, there are about over 500 bike-sharing programs around the world. There exists great interest in these systems due to their important role in traffic, environmental and health issues.
- Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems, which allows for extensive analysis.
- The following analysis allows for the understanding and interpretation of how weather impacts commuting via bicycle for a large American city in contemporary times, amidst the metropolitan hub of daily commuters.

Data

- Citi BikeShare is a rental company which allows users to rent bicycles. A rider unlocks a bike and pays per time or distance that they use the bike. After a ride is ended at a designated BikeShare location, thus ending the ride, the data is recorded in the BikeShare database.
- Parameters recorded include the type of bike used, the start and end times, and the start and end geographical coordinates and the start and end stations.
- The weather data includes the type of cloud cover, precipitation, and temperature for different times of day during the month of January, 2022.
- Both data-sets thus align in their respective time-frames, with the BikeShare data-set recording over one million events for the month.

Attributes of NYC Bike Data-Set

Attributes of NYC Bike Data
ride_id
rideable_type
started_at
ended_at
start_station_name
start_station_id

Attributes of NYC Bike Data
end_station_name
end_station_id
start_lat
start_lng
end_lat
end_lng
member_casual

Attributes of NYC Weather Data-set

Attributes of NYC Weather Data
name
datetime
tempmax
tempmin
temp
feelslikemax
feelslikemin
feelslike
dew
humidity
precip
precipprob
precipcover
preciptype
snow
snowdepth
windgust
windspeed
winddir
sealevelpressure
cloudcover
visibility
solarradiation
solarenergy
uvindex
severerisk
sunrise
sunset
moonphase
conditions
description
icon
stations

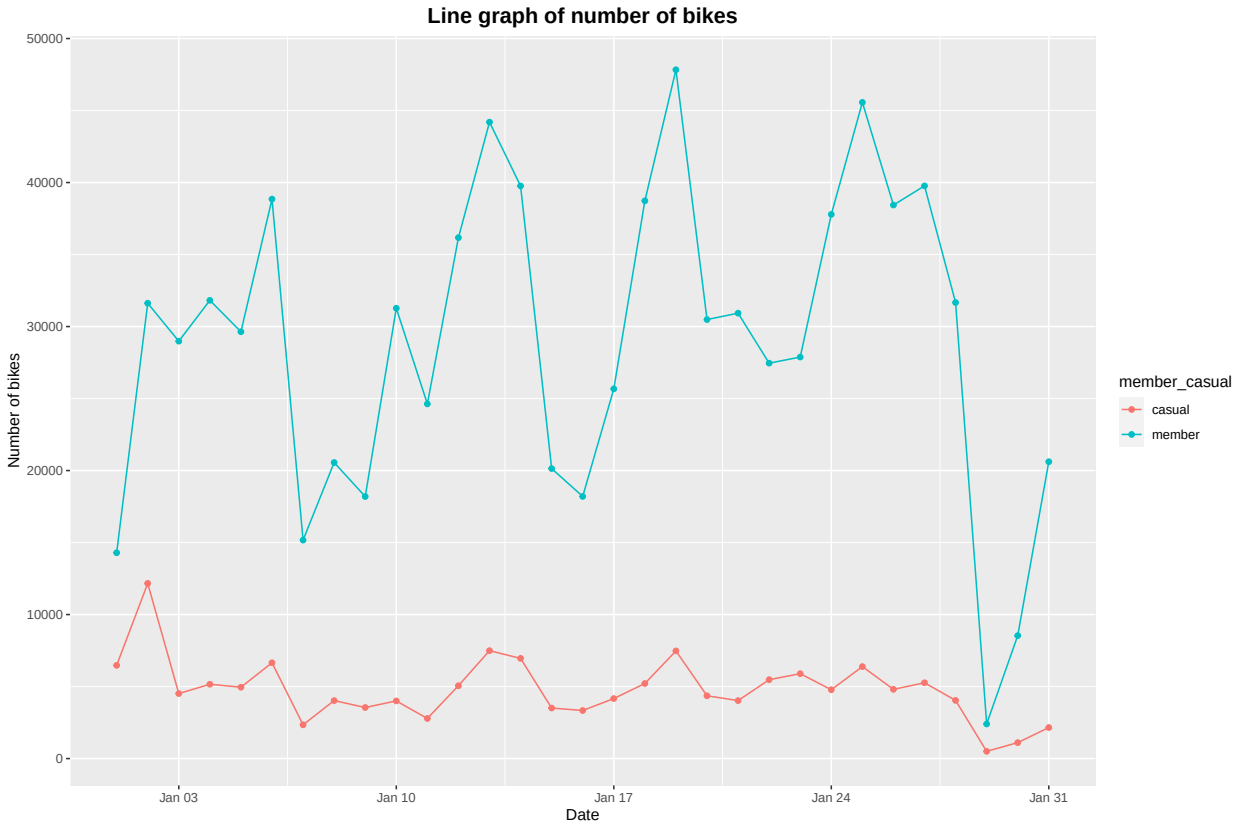
Analysis

- The premise of the analysis is to understand the correlation between several weather variables and BikeShare riders for the NYC area.

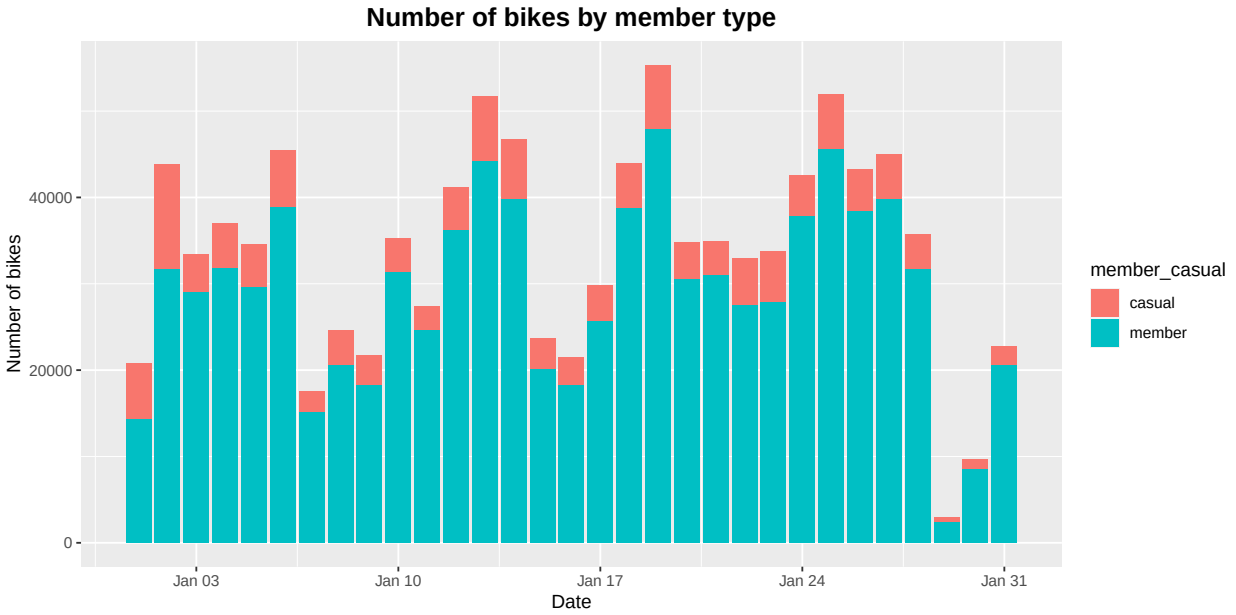
- Several weather related parameters were plotted against the types of bikes and number of riders for each day of the month.
- Regression analysis was employed to determine the correlations between parameters in the two data-sets to observe trends in the data.

Results

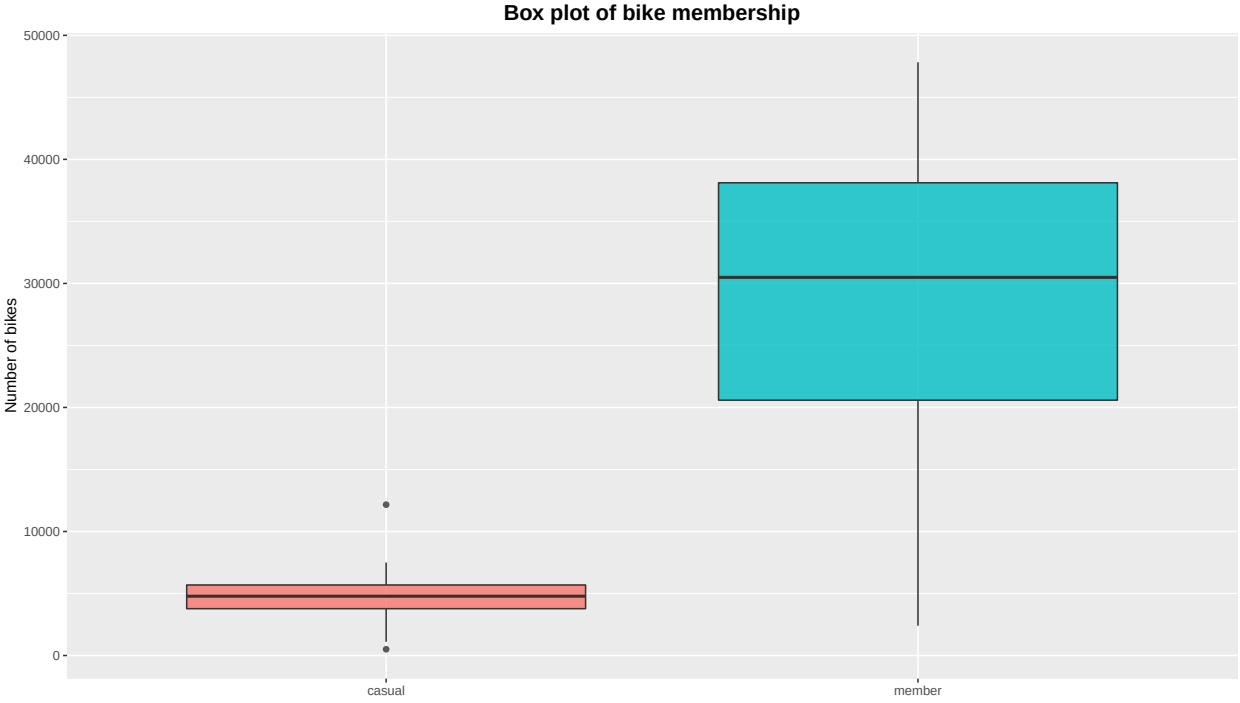
This is an accumulation of our data analysis done with the attributes of both the aforementioned data-sets.



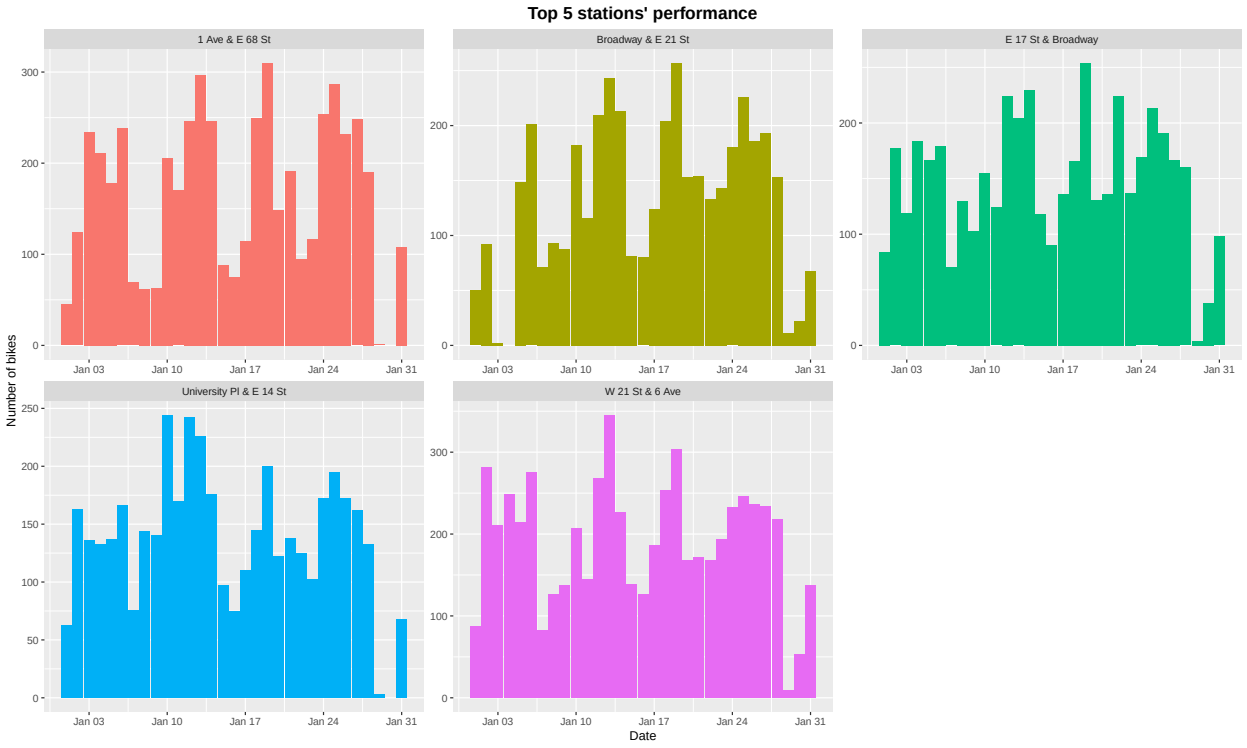
- The above graph shows the number of bikes being rented on each day of the month by membership.
- The number of people renting bikes with a membership out-weigh the people renting them casually.
- This helps in setting up the initial analysis of the data. Subsequent analysis can take place after understand the data structure and division.



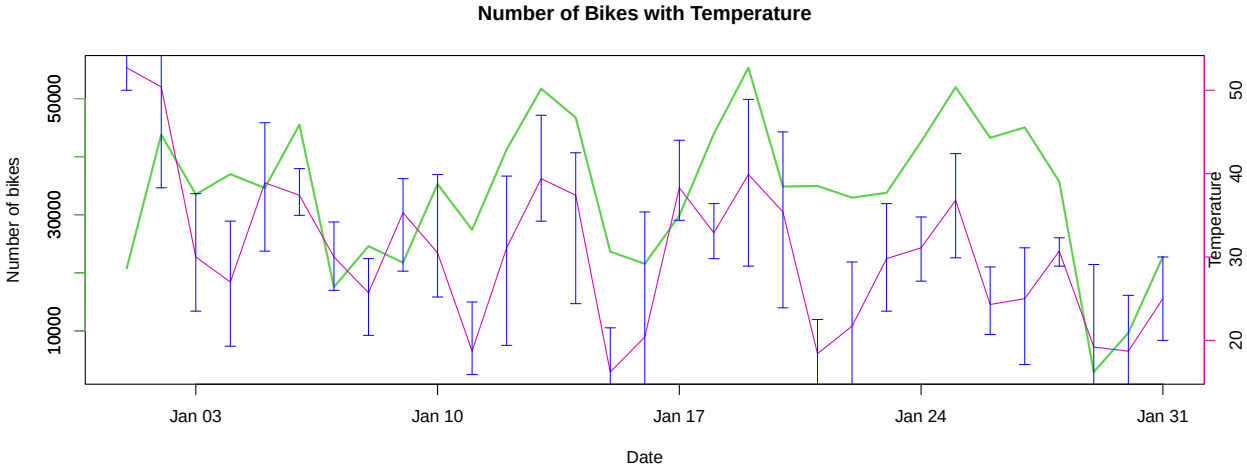
- This can also be visualized in the form of a bar plot above.



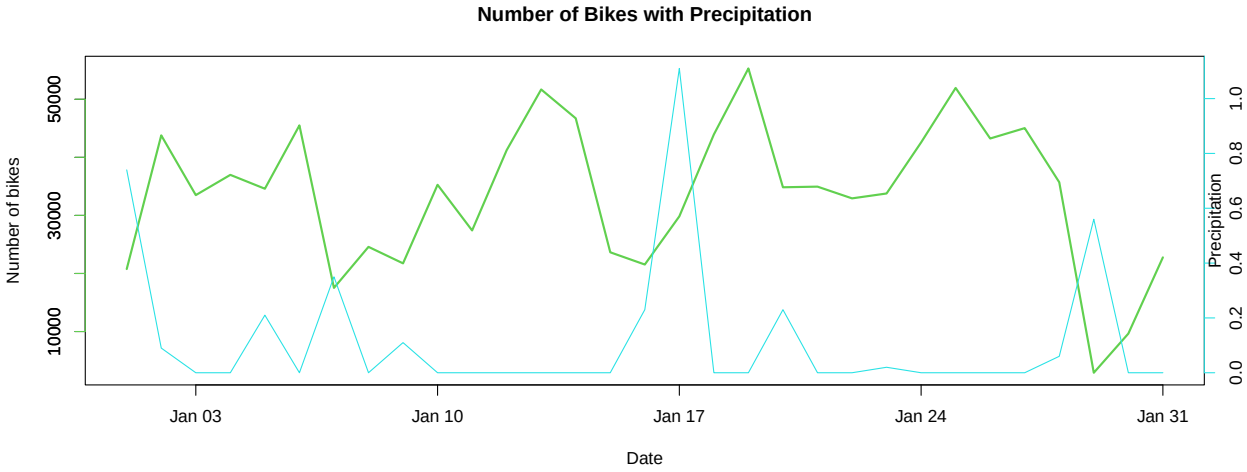
- The box plot below shows the spread of the number of bikes rented based on membership. On an average only 5,000 bikes were rented by casual customers while 30,000 bikes were rented by members.
- This information is helpful in stocking the bikes so that preference can be given to members until the bikes rented by members crosses 30,000.



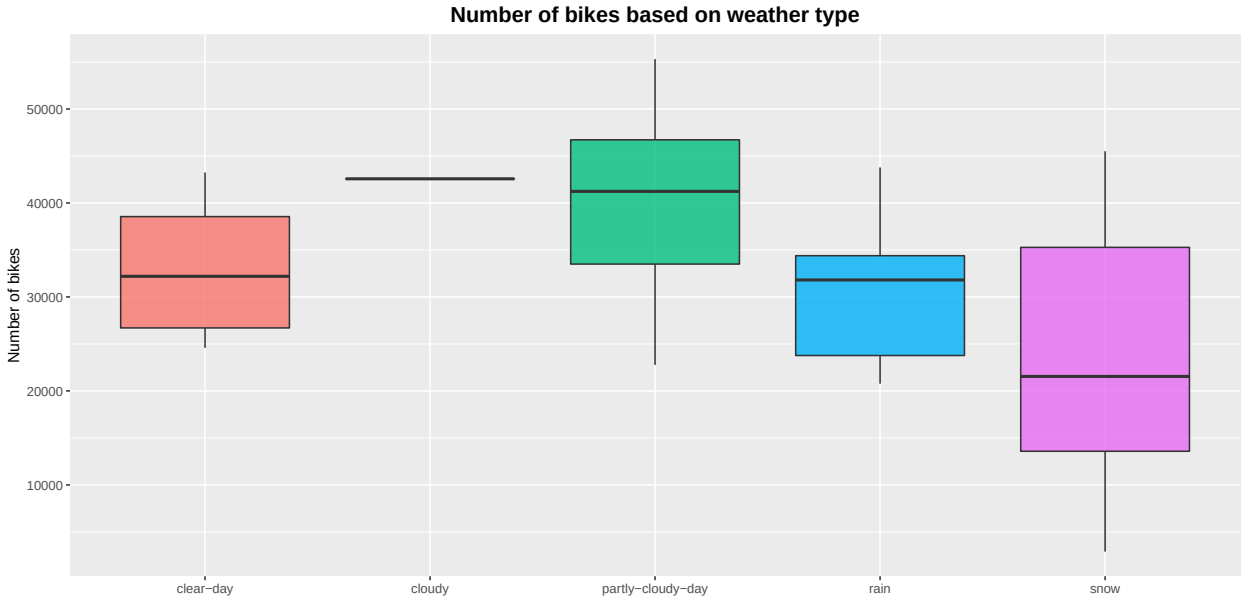
- There are total of more than 1500+ bike stations run by Citi Bike in NYC. The top 5 best performing stations can be seen above, all of which rent 100 bikes on an average everyday and can expect more during the weekdays.



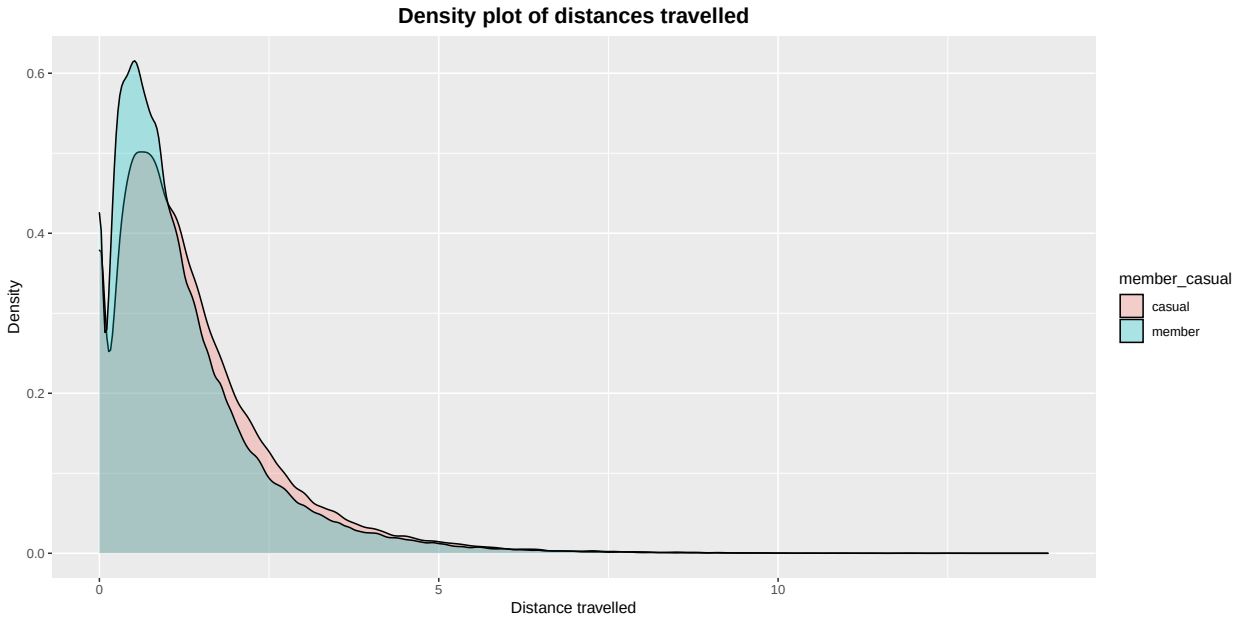
- The correlation between the number of riders per day and temperature is not extremely strong.
- But a general trend can be spotted where during high and low temperatures the number of riders seems to have reduced most of the time which can be expected.
- The above graph shows the temperature fluctuation (with max and min) and the number of bikes rented on a particular day.



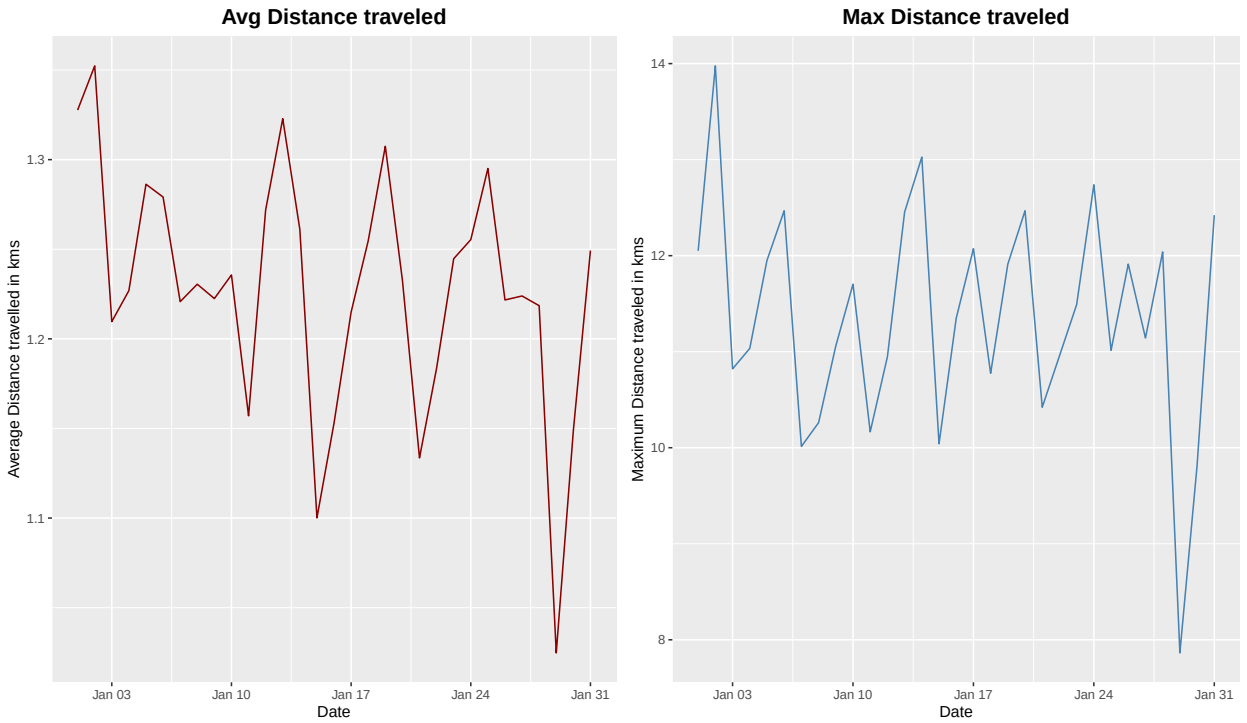
- The correlation between the number of bikes rented and precipitation is mostly sporadic.
- A very weak correlation can be observed when the precipitation increases the number of rides reduces in most of the cases.



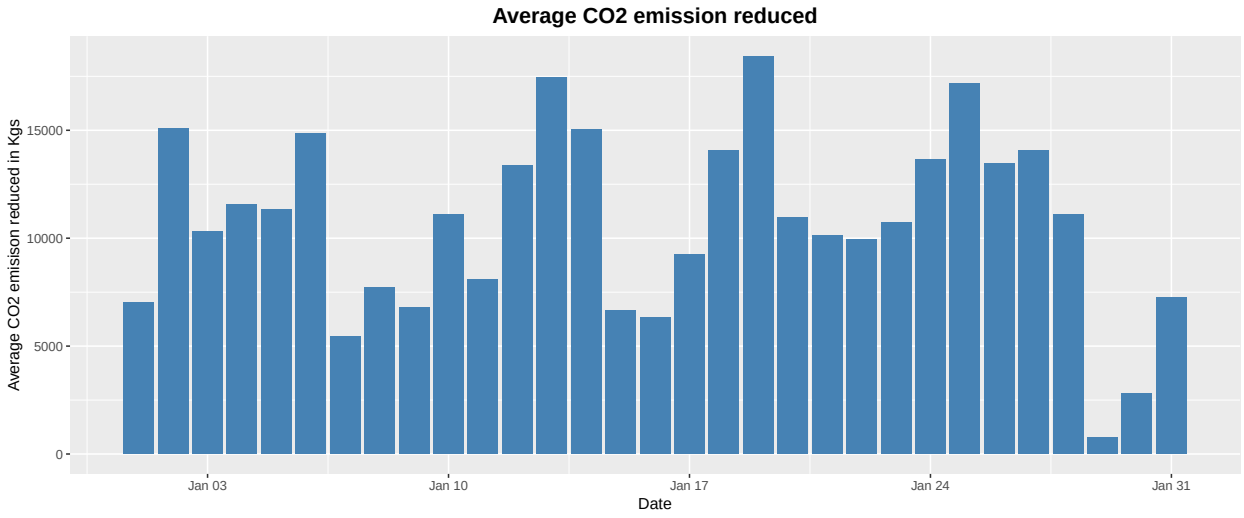
- The above graph shows the box plot of the number of bikes rented and the type of weather on a particular day.
- As can be seen here not a lot of bikes seem to have been rented when the weather is cloudy but only 1 day had a cloudy forecast in NYC in Jan 2022.
- So the only a few bikes would fall in that category in this graph.



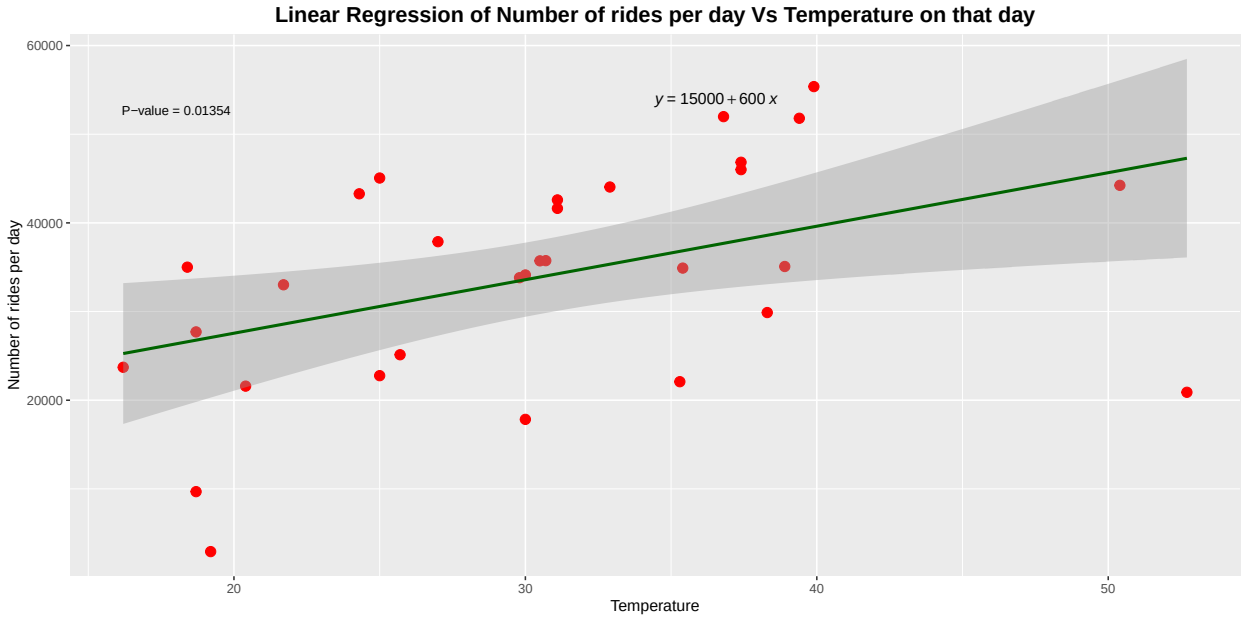
- The density plot can also be seen for both casual and member riders and both of these plots seem to follow a very similar trend. The distances are calculated using the Haversian formula which assumes that the Earth is a perfect sphere. This would not effect the distances calculated at all because the deviation of curvature of NYC would be negligible from a perfect sphere for such a small area.



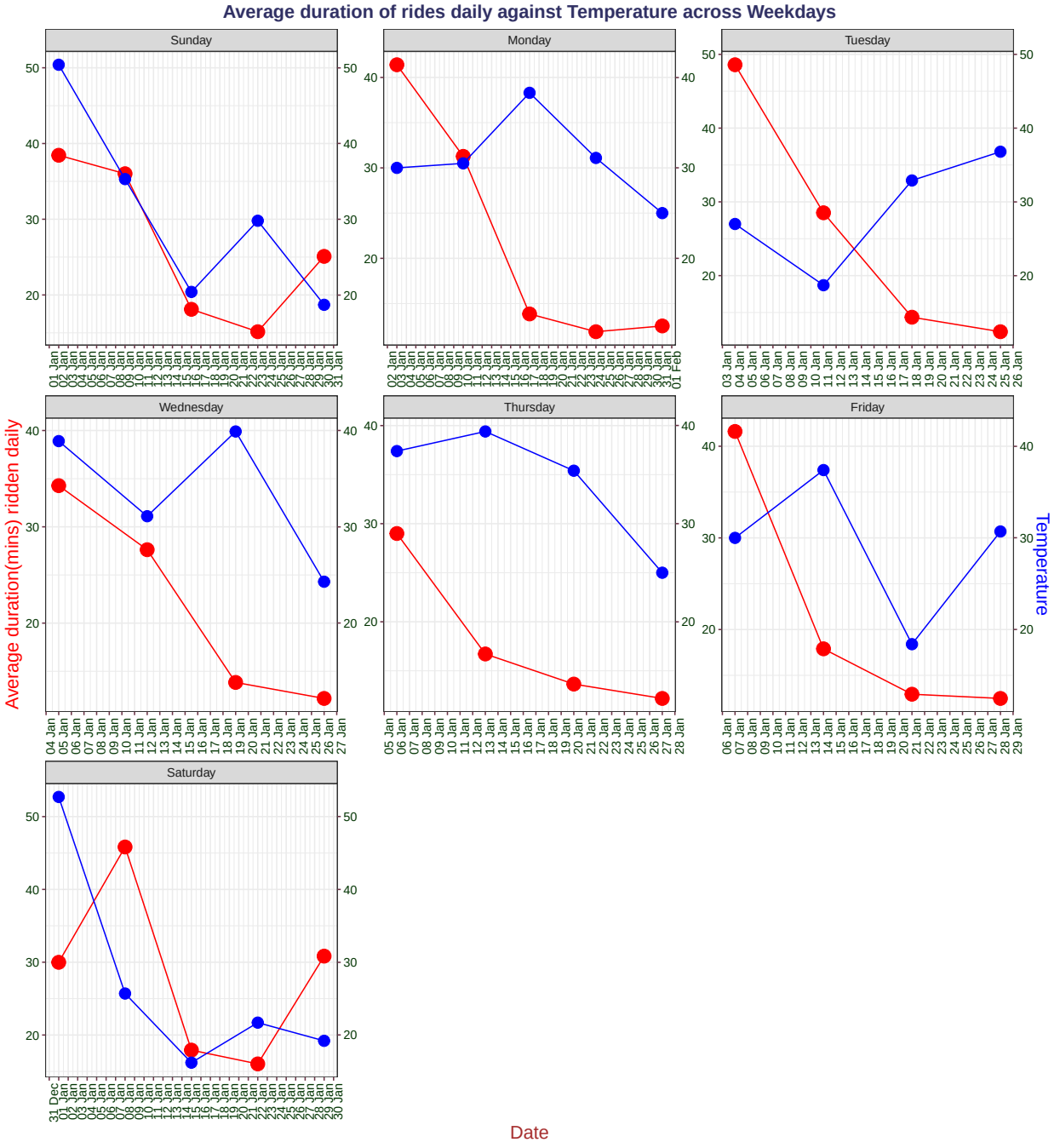
- The above plots show the average distance travelled by riders which is 1-1.5 Kms as seen in the density plots before and the maximum distance travelled by rider on each day.



- An estimated amount of 255 gms of CO2 is released for every Km travelled by an average car. The below graph shows the average amount of CO2 emission reduced in Kgs on each day by Citi Bike customers.

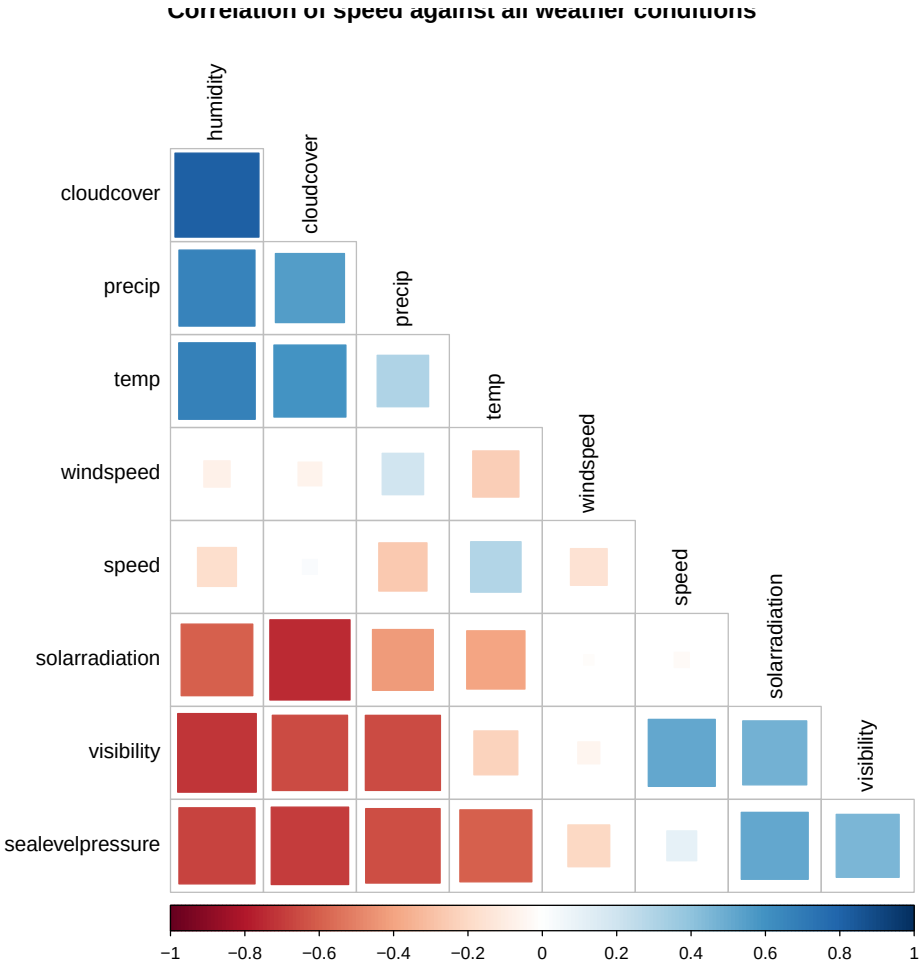


- A Linear regression model is used to understand the relationships between the biker density in various weather conditions.
- Among all the weather conditions, temperature is most highly associated with the number of rides per day.
- This is evident from the p-value obtained for this model, which is 0.013 and is lesser than the significance level.
- This indicates that there is a good amount of association between these two variables.
- Also, the regression line fits almost half of the data points which lie within the confidence intervals.

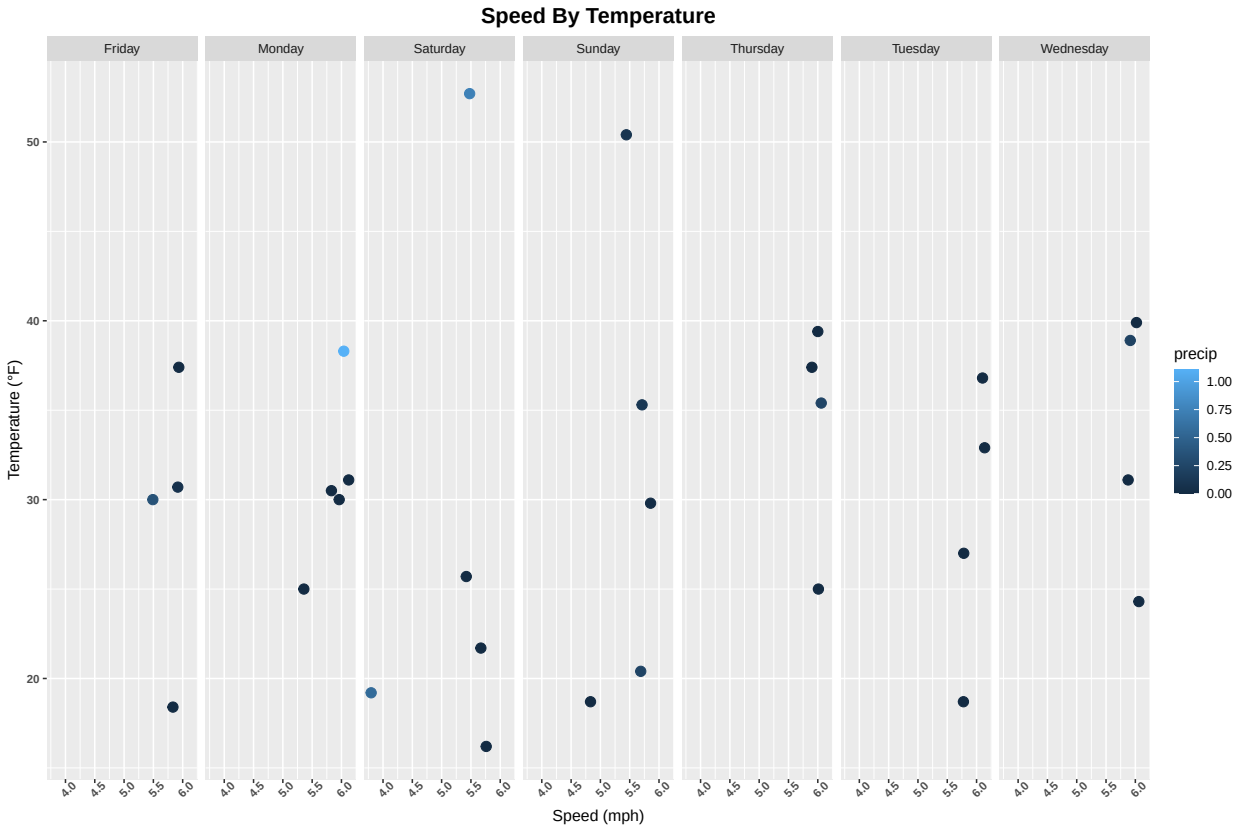


- This plot shows the average duration of bike rides on a daily basis across weekdays and their trends w.r.t temperature.
- The red line shows the avg duration in mins each day and the blue line is the temperature on that day.
- We observe a similar trend during the week as opposed to weekends, i.e., all the 5 days from Monday to Friday across the 4 weeks in January have a similar trend in the average trip duration.
- We notice that there is a decent amount of correlation between these two variables.

- This plot shows the number of rides varying w.r.t the time of day, i.e., morning, evening, day and night. It is evident that members of the Citibike rental system are the ones who had the maximum rides compared to non-members throughout the month of January.
- These numbers are high during the week probably because the members consider the rental system as their common mode of commute.



- This is a correlation plot which shows that speed is positively correlated with temperature.
- This indicates that a higher speed is associated with a higher temperature value. On the contrary, it also indicates a lower speed associated with a lower precipitation which indicates positive correlation.
- The correlation will increase if the data is extrapolated for more than a month's worth of data (over the course of the whole year).
- More significant statistical conclusions can be made with regard to speed against the weather attributes for a data-set collected over a longer duration of time.



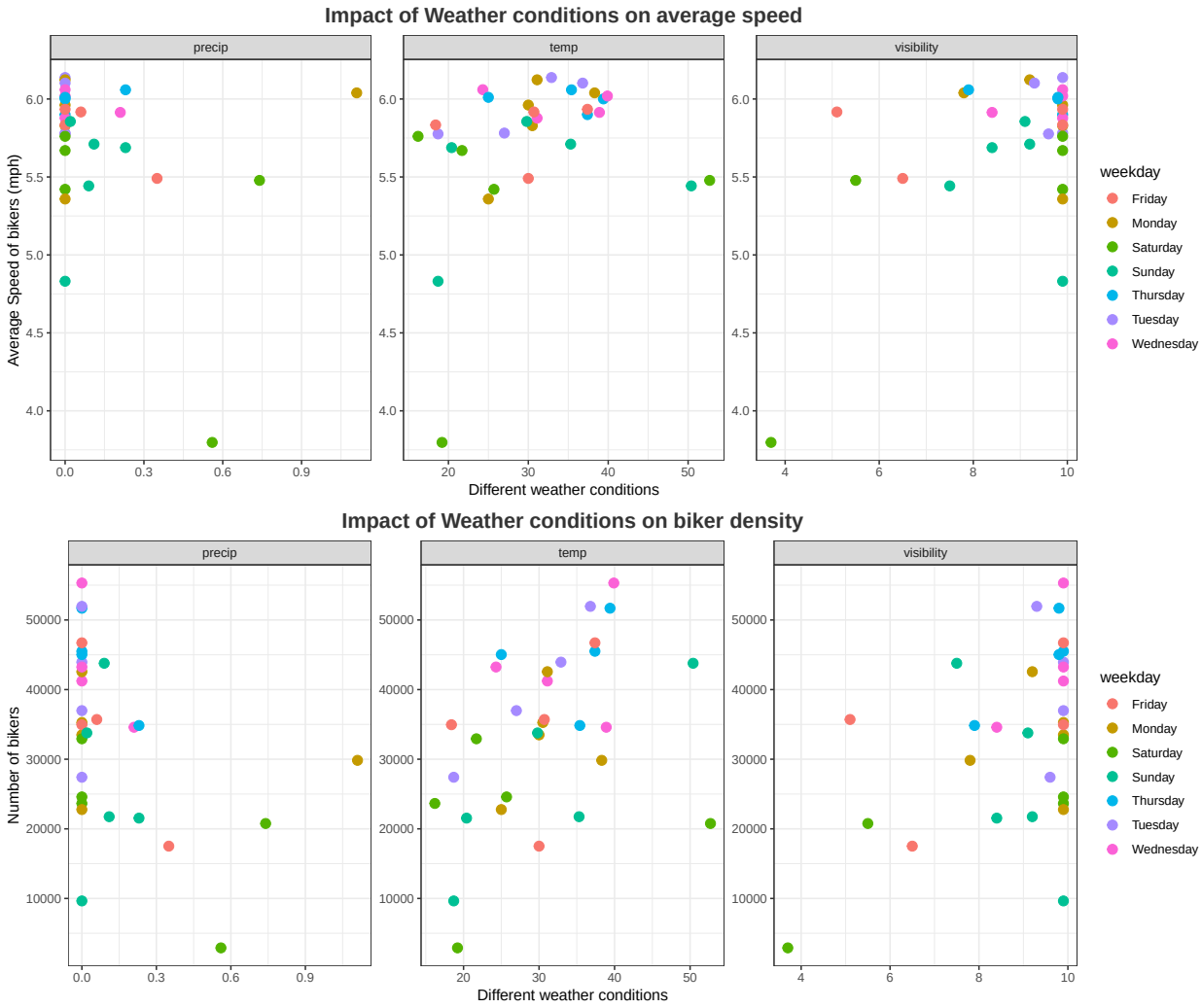
- This is a plot which displays the temperature vs average speed for the day, with precipitation as the colour coded values.
- This shows outliers during the weekend which indicates that despite a higher temperature value, riders tend to venture out during the weekends as opposed to weekdays.
- Higher precipitation levels seem indicative of higher speed which denotes that there is a positive correlation between the two, which can be seen in the previous correlation plot as well.

KILLER PLOT

- This shows our "killer plot" in a grid format. Our killer plot analyzes biker density and average speed of bikers per day against the top 3 weather parameters.

1. temperature
2. visibility
3. precipitation

- The speed had to be calculated using the start and end latitudes and longitudes.
- The haversian distance was calculated and the timestamps were extracted and formatted to give rise to the duration of the trip. Using this, average speed was calculated and plotted against the weather parameters with the highest correlation.



Conclusions

- Throughout the analysis, regardless of membership status for CitiBike Share, riders continued to ride in substantial amounts and frequency during average weather conditions. Weather and riders per day did seem to align with temperature in that as temperature rose or fell, the number of riders decreased.
- However, riders typically rode in uncorrelated fashion when compared to different forms of cloud cover as more riders tended to be active during partly cloudy weather rather than on clear days as expected, with the least number of riders occurring under harsher conditions like rain and snowfall.
- Bike traffic across the city also tended to align between different streets and areas of NYC and throughout the month of January. Categorized based on time of day, mornings saw increased activity across the week regardless of weather for both members and non-members on weekdays.
- Even for a large metropolitan area such as NYC, certain key weather conditions have observable effects on how many riders participate in the BikeShare community.