

DTU



Group 9

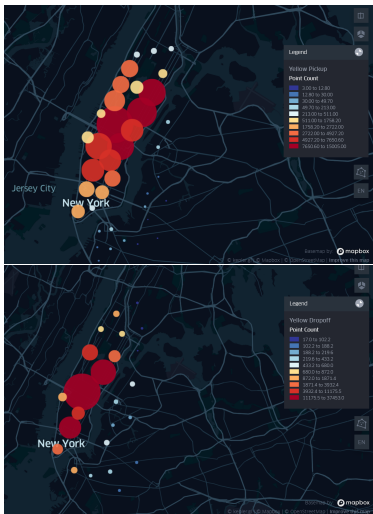
Data visualization and analysis

Project 1 - Summary of Exploratory Data Analysis

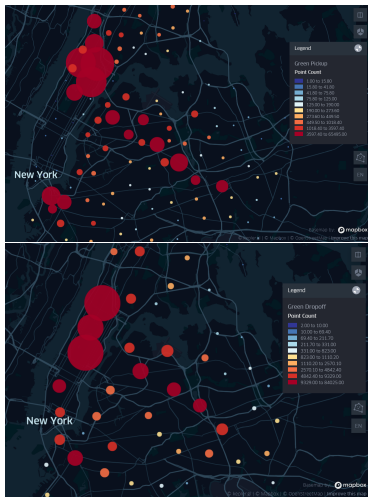
Summary

- Both the yellow and green taxi's roughly follows the same distribution for fare, trip and passenger counts.
- There is a strong correlation between fare amount and the trip distance which is expected.
- However there are no correlation between the number of passengers and the fare amount.
- The same can be said about the number of passengers and the trip distance.

Project 1 - Spatial Analysis



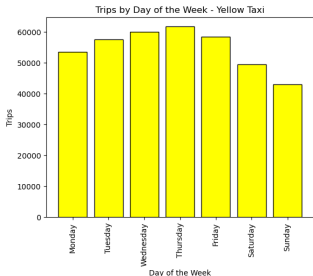
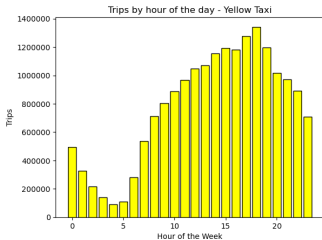
Project 1 - Spatial Analysis



Green Pickup & Dropoff

- The data shows that green-taxis are based both in the center of NYC, but also very much in the outskirts of the city.
- This can also be seen in the clustering of the drop-off zones which are primarily in the center of the city.

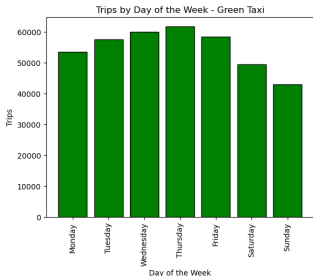
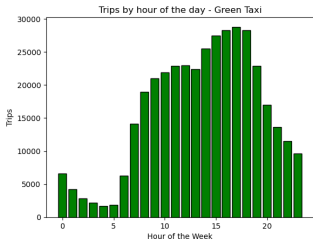
Project 1 - Temporal analysis



Time of trips for yellow taxi's

- The data shows that most of the trips are being conducted in the rush hours.
- The busiest hours of the day are between 17-19.
- The data shows that the busiest days of the week are in the mid-week.
- Surprisingly, the least busiest day of the weeks are in the weekends.

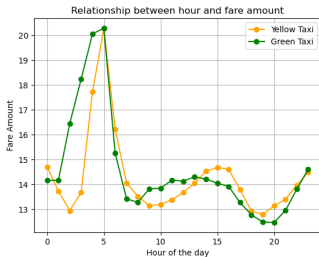
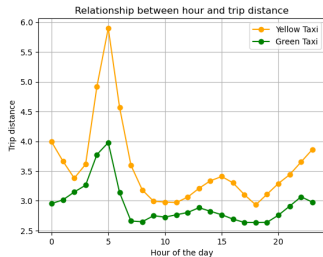
Project 1 - Temporal analysis



Time of trips for green taxi's

- The data for the green taxi's mostly follows the same trend as the yellow taxi's.
- The only difference is that the busiest hours for the green taxi's are between 15-19.

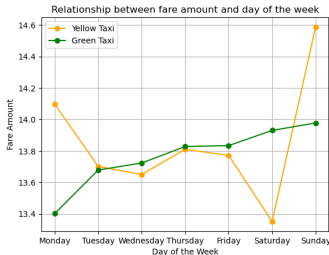
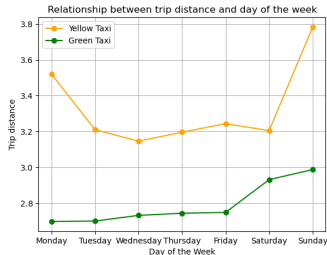
Project 1 - Temporal Analysis



Fare and distance

- Data shows that longer trips are being conducted at that time earlier morning and at night.
- Maybe this can be because of the limited public transport.

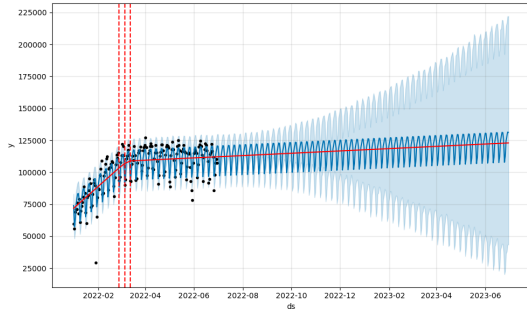
Project 1 - Temporal Analysis



Fare and distance

- Data shows that the fare amount for the green taxi's incline doing the week.
- The steep incline in the weekends suggest that most of the trips are longer.

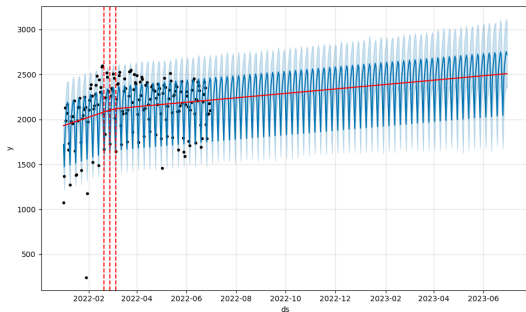
Project 1 - Time Series forecasting



Time-Series Forecast (Yellow)

- The forecast has been made for 365 days.
- The trend originally showed a decline.
- Using changepoints and making the trend less flexible the forecast to the left has been made.
- Forecast has been validated using actual data from July - September.
- It shows an average deviation of ≈ 8476 trips pr. day, which is a average error of 8,19%

Project 1 - Time Series forecasting



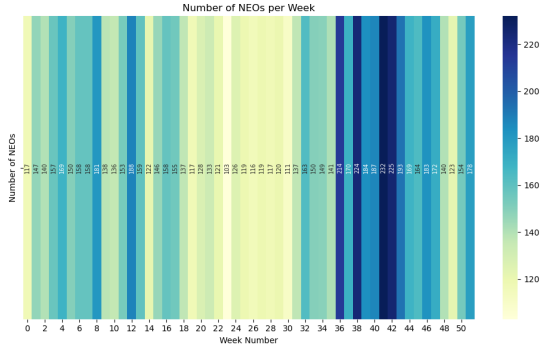
Time-Series Forecast (Green)

- The forecast has been made for 365 days.
- The trend originally showed a decline.
- Using changepoints and making the trend less flexible the forecast to the left has been made.
- Forecast has been validated using actual data from July - September.
- It shows an average deviation of ≈ 63 trips pr. day, which is a average error of 2,91%

Project 1 - Limitations and future research

- Too many data entries for Yellow Cabs for our PC's to handle (Hardware)
- There are many outliers in the data set, which needs to be evaluated
- Weather data could provide insights on if cabs are busier on days with bad weather.
- Improved seasonal data could provide better forecasting for high demand periods.
- Data on trip purpose could provide valuable insights to why and where people are taking cabs.

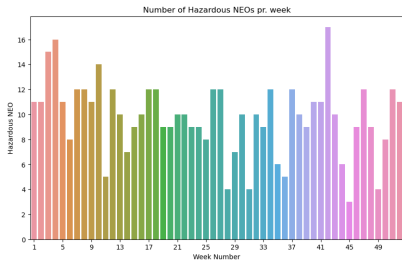
Project 2 - Number of NEO's



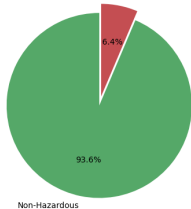
NEO's

- The number of NEO's seems to rise by the end of the year.
- However the number of hazardous NEO does not seem to have seasonal correlation.

Project 2 - Hazadeous NEO's



Proportion of Hazardous vs Non-Hazardous NEOs



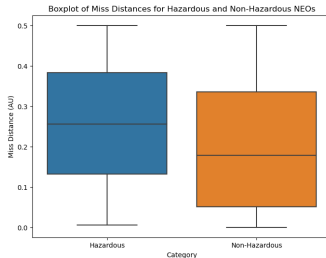
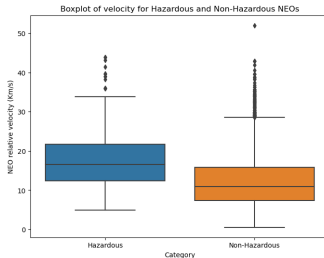
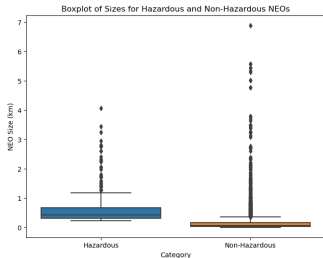
NEO's

- The number of NEO's that are hazadeous doesn't seem to have a seasonal correlation.
- The number of NEO's that are hazadeous is also grealy lower then the onces that are classified as non-hazadeous.

Project 2 - NEO's attributes

Attributes

- The NEO's that are considered hazardous are bigger, and faster than non-hazardous.
- However they have a greater miss distance.



Project 2 - Conclusion

Summary

- The data shows no real correlation between number of NEO's pr. week that are hazardous.
- There are far more non-hazardous NEO's than hazardous.
- The data shows that the NEO's are hazardous are typically bigger in diameter.
- The data also shows that the NEO's that are hazardous have a greater miss distance, and higher velocity.
- The data also shows that many of the NEO sizes fall as outliers.

Recommendation

- The data shows that for a NEO to be considered hazardous it is an accumulation of both speed, distance and size.
- Therefore it is recommended that when analyzing the NEO's one has to consider all of these factors, and if a NEO is found to be predicted for a near-miss that resources will be prioritized to find both the speed and the size of the NEO.
- It can be recommended that resources are prioritized to find methods to better predict and calculate the size of the NEO's.