
DTU

DTU

Group 9

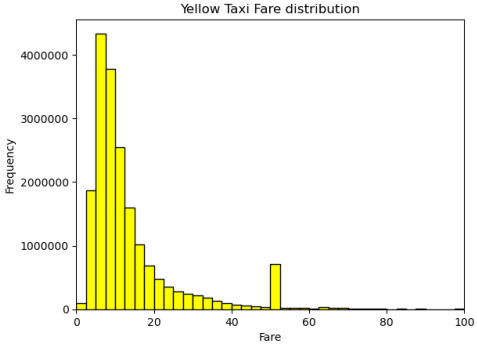
Data visualization and analysis

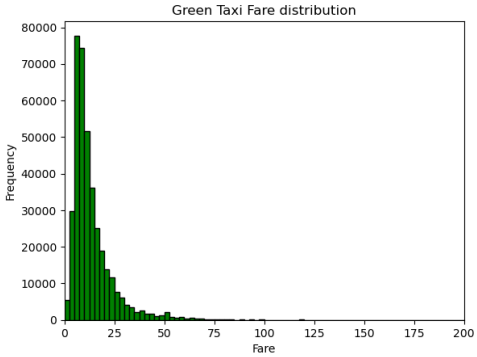
Project 1 - Taxis

Understanding the data

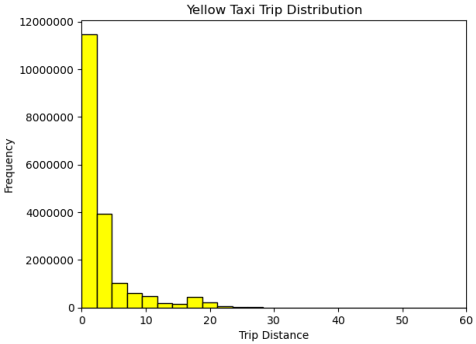
- The data shows records of each taxi trip for both yellow, and green taxi's in New York.
- The data has been pulled in the date interval 01/01/2022 - 30/06/2022
- The data shows some errors, as there are fare's there are negative, and unreasonably high ($>$ 1000\$).
- The same can be seen for trip distance.
- Therefore it is decided that data is only pulled for trips $0 \leq t \leq 500\$$ and $t \leq 500$ miles.

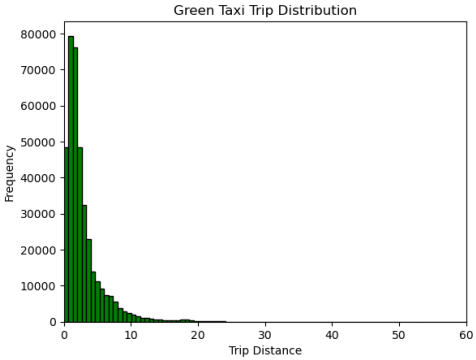
Project 1 - Distribution of fare



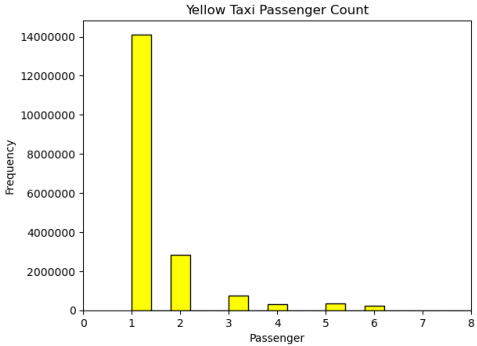


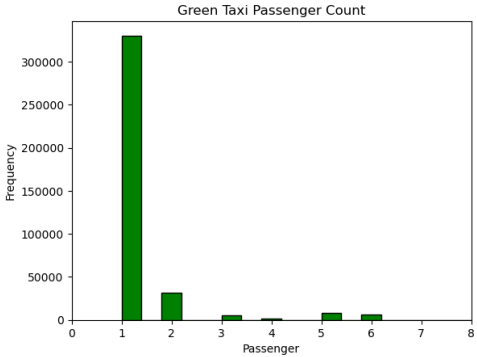
Project 1 - Distribution of Trip distance





Project 1 - Distribution of Passengers pr. Trip



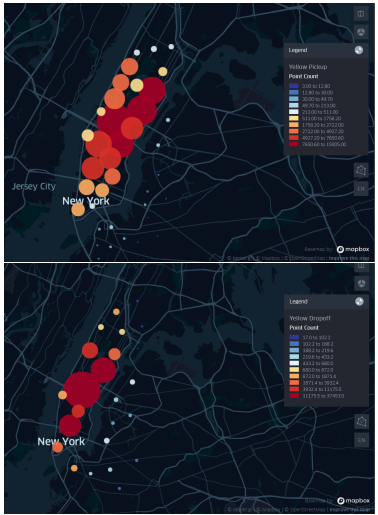


Project 1 - Summary of Exploratory Data Analysis

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.

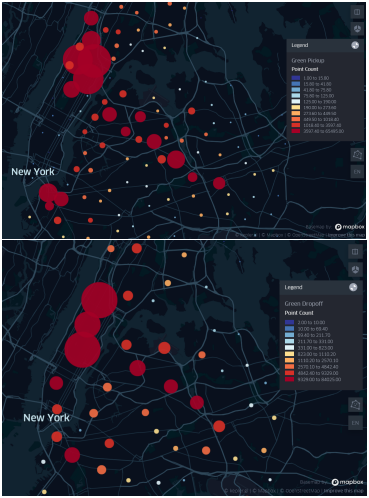
Project 1 - Spatiel Analysis



Yellow Pickup & Dropoff

- The data shows that the yellow taxis are primarily based in the center of NYC.
- The data shows that many of the trips are also inside the city center, and not many of the trips goes to the outskirts of the city.

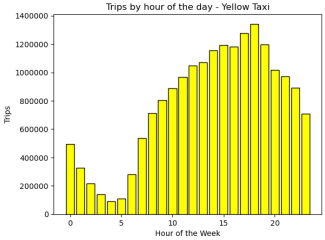
Project 1 - Spatiel Analysis

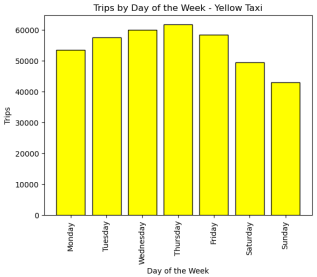


Green Pickup & Dropoff

- 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

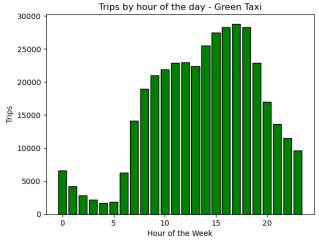




Time of trips for yellow taxi's

- 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.

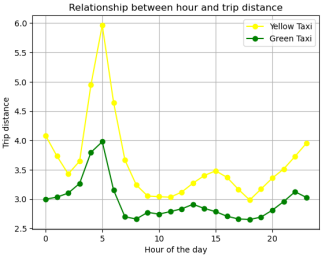
Project 1 - Temporal analysis

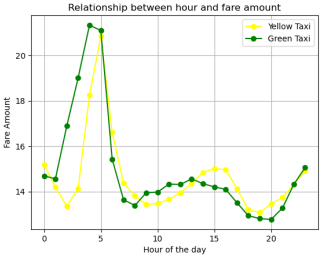


Time of trips for green taxi's

- 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

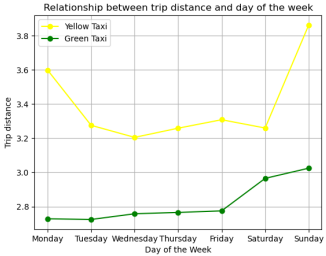


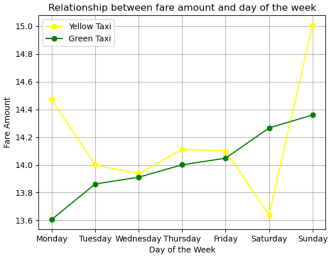


Fare and distance

- 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

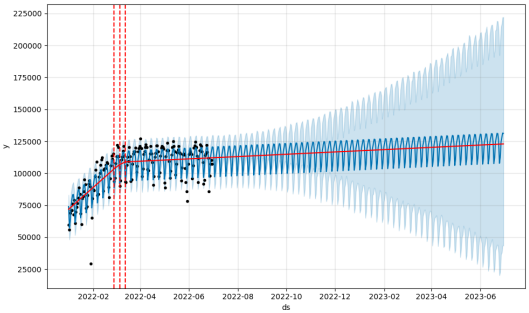




Fare and distance

- 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.

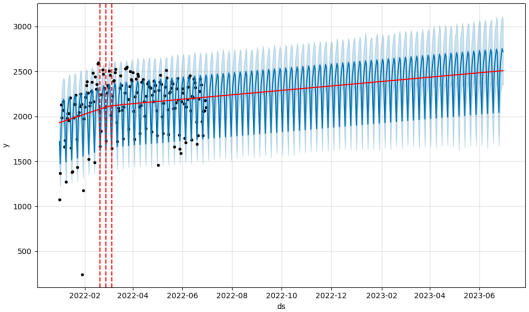
Project 1 - Time Series forecasting



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 ≈ 10.000 trips pr. day, which is a average error of 9,87%

Project 1 - Time Series forecasting



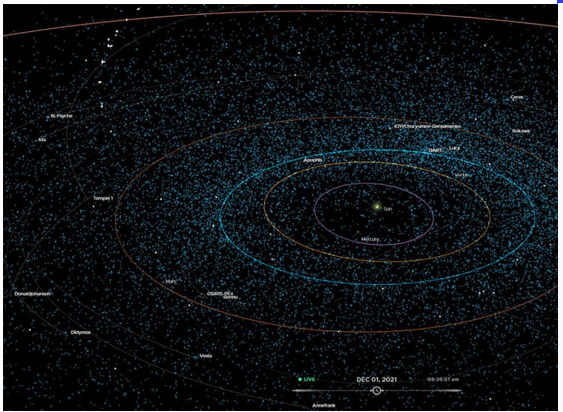
Time-Series Forecast (Green)

- 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 $\approx$ 89 trips pr. day, which is a average error of 4,13%

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 - NEO



Extracting Data

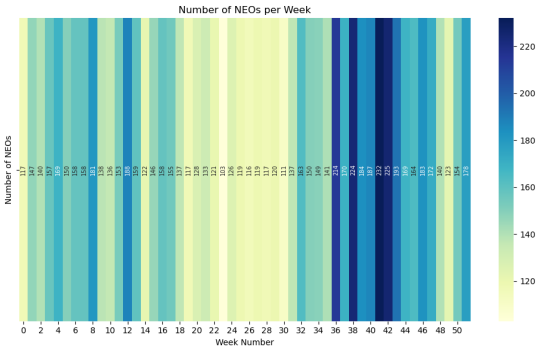
- The Data pulled from the NASA website shows near earth objects or "NEO"
- An API key is generated to get access to the data.
- The Data is then analysed and setup in a visual/graph to easier understand our findings.
- For the context of this project, we will work with one years worth of observations, starting first of January 2022.

Project 2 - Methodology

Statistical analysis

- The data received has been unpacked from JSON-format and extracted to a pandas dataframe.
- The data is herafter used to make statistical analysis to try and predict the behaviour of NEO's.
- The size, velocity and distance of the NEO's has been taken into account.
- Afterwards, summary and recommendations are made, and a relevant scientific paper has been analyzed for furture methods to classify NEO's.

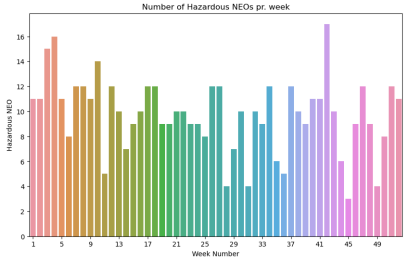
Project 2 - Number of NEO's

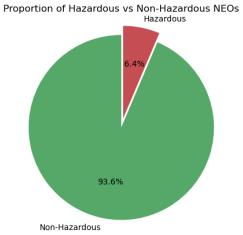


NEO's

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

Project 2 - Hazadeous NEO's





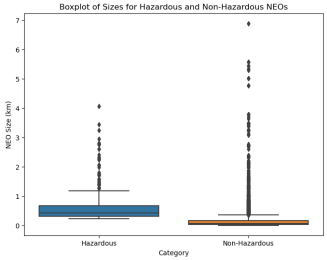
NEO's

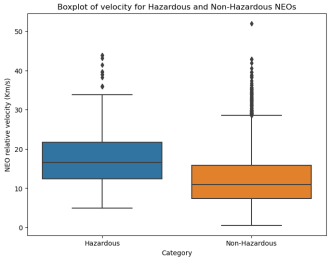
- The number of NEO's that are hazadeous doesn't seem to have a seasonal correlation.
- The number of NEO's that are hazadeous is also grealy lower then the onces that are classified as non-hazadeous.

Project 2 - NEO's attributes

Attributes

- The NEO's that are considered hazadeous are bigger, and faster then non-hazadeous.
- 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 hazardeous.
- The data shows that the NEO's are hazardeous are typically bigger in diameter.
- The data also shows that the NEO's that are hazardeous 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 considered hazardeous it is an accumalation 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 ressources will be prioritized to find both the speed and the size of the NEO.
- It can be recommended that ressources are prioritized to find methods to better predict and calculate the size of the NEO's.

Project 2 - Advanced classification

Support Vector Machines

- A paper written by C. Dechev, V. V. Orlov, I. Oprea, B. G. Williams in 2014 suggest that using SVM can greatly increase the classification of the NEOs
- This method trains a model to classify the NEOs into different known classes.
- Using predefined classes, the knowledge obtained is greatly increased.