

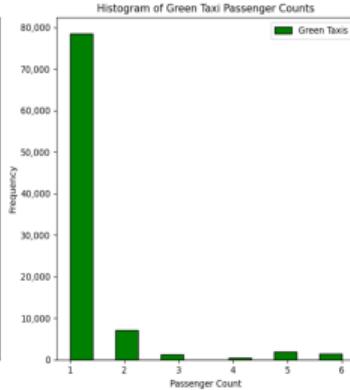
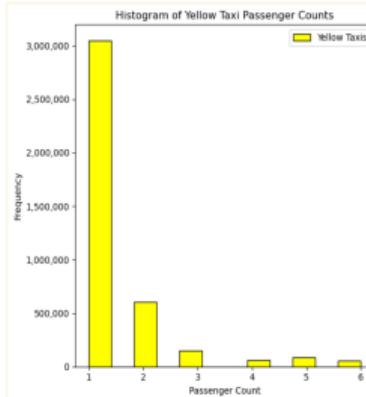
**DTU**



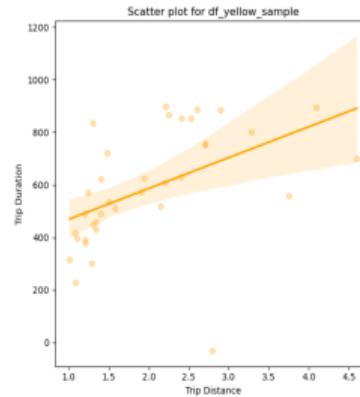
DTU Course no. 62444, Project, Group 11.

# Data Visualization and Analysis

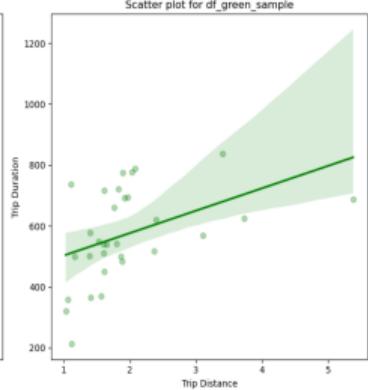
# Exploratory Data Analysis



(a) Histogram



(b) Scatterplot



- Histograms - For yellow taxis, the majority of trips have a passenger count of 1, with over 3 million occur For green taxis, single-passenger trips also dominate the distribution where 78,414 trips occur
- Scatterplots - For the green taxi scatter plot, the correlation coefficient is 0.57. In the yellow taxi scatter plot, the correlation coefficient is at 0.53.

# Spatial analysis - Green Taxis



(a) Heat-map



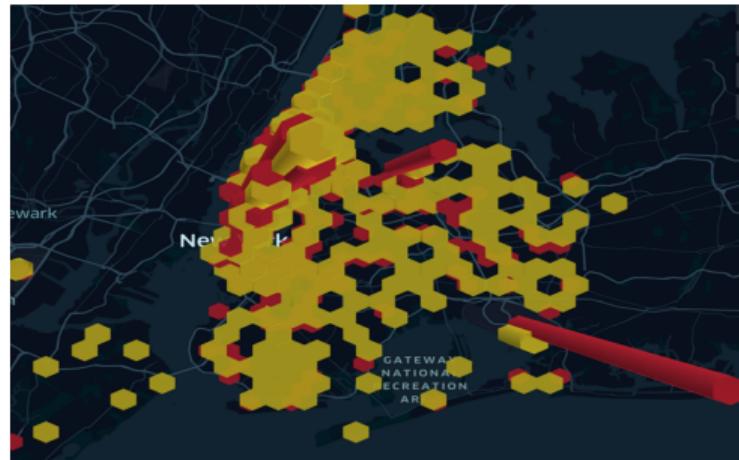
(b) Hexbin map

- Central Park hotspot
- Pickup (green) vs Dropoff (red)

# Spatial analysis - Yellow Taxis



(a) Heat-map



(b) Hexbin map

- Central Park hotspot
- Pickup (yellow) vs Dropoff (red)

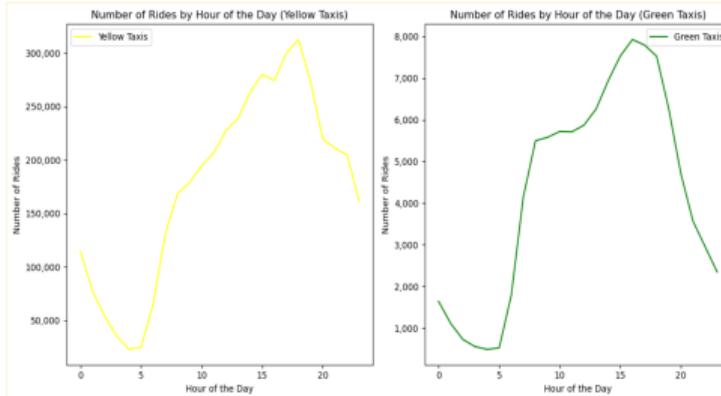
# Spatial analysis - Yellow and Green Taxis



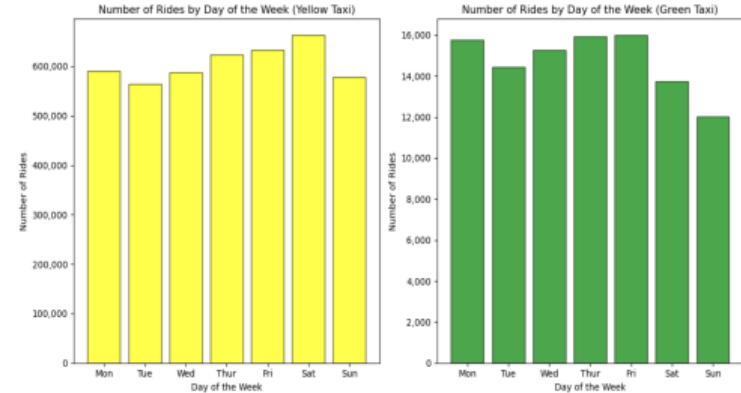
(a) Heat-map

- Differences in hotspots
- Green vs yellow

# Temporal analysis



(a) Line-plot



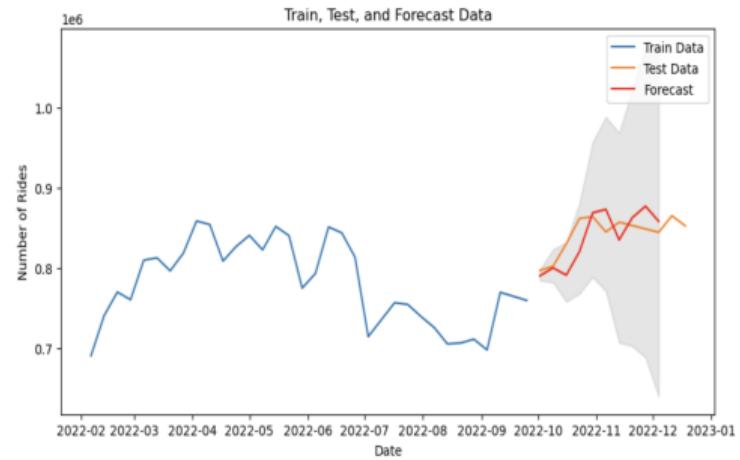
(b) Histogram

- The line plot compares the number of trips by hour of the day
- The number of rides by day of the week - For the yellow taxis the highest number of rides occurs on Saturday. For green taxis the highest number of rides occurs on Thursday

# Forecasting of NYC Taxi Rides - Yellow Taxis



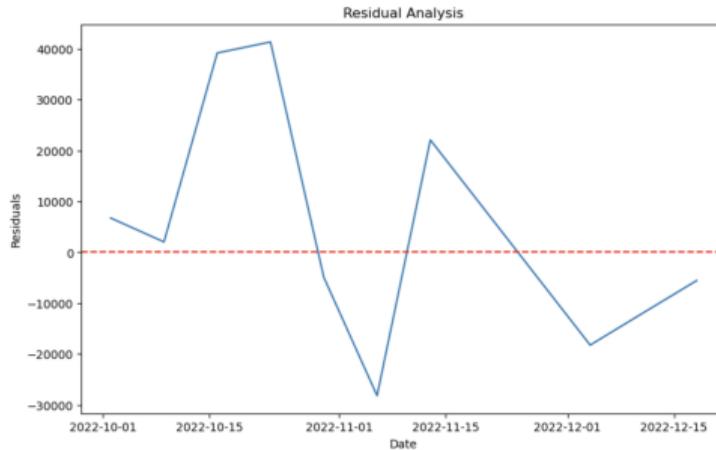
(a) Test Train data



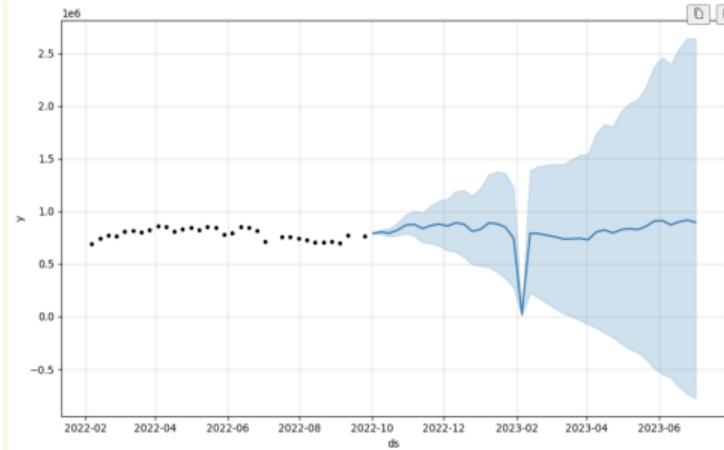
(b) Forecast Model Testing

- Outliers
- Forecast Test Model – Forecast vs Test Data

# Forecasting of NYC Taxi Rides - Yellow Taxis



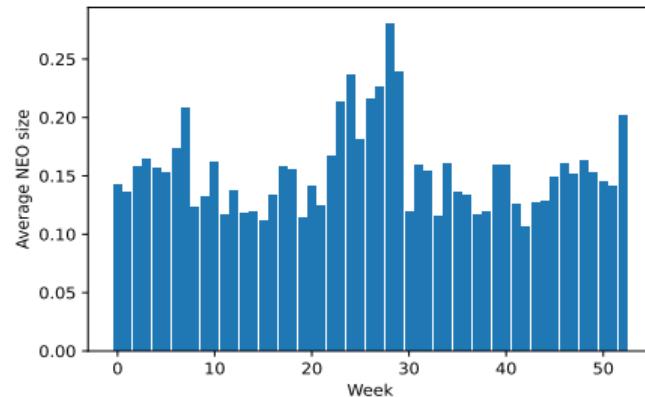
(a) Residual plot (Forecast Test Model)



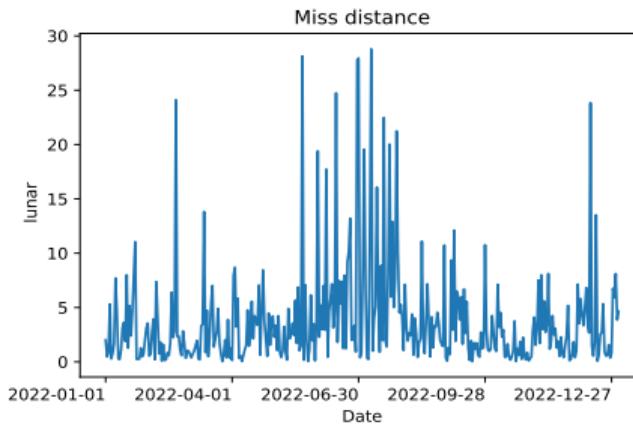
(b) Forecasting the Future

- Mean Absolute Error (MAE): 18.017,62
- Predicted values and Actual values
- Number of Taxi Rides in the future – period with unrealistic decrease

# NASA Data Acquisition, Visualization, and Analysis



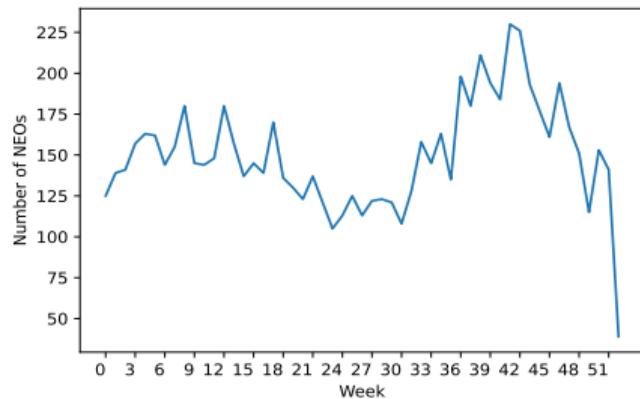
(a) Avg NEO size



(b) min miss distance

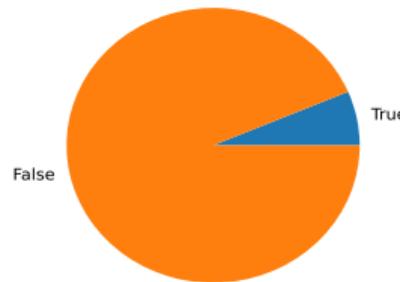
- Larger and further away from Earth in the Summer

# NASA Data Acquisition, Visualization, and Analysis



(a) Number of NEO per week

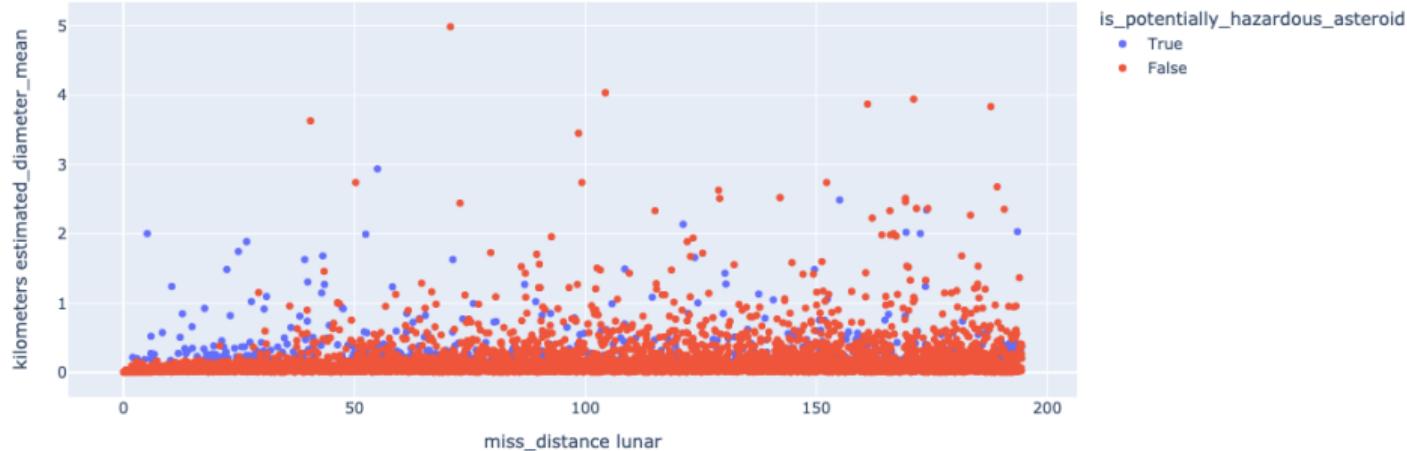
Distribution of hazardous vs non-hazardous NEOs



(b) Whether the asteroids are hazardous

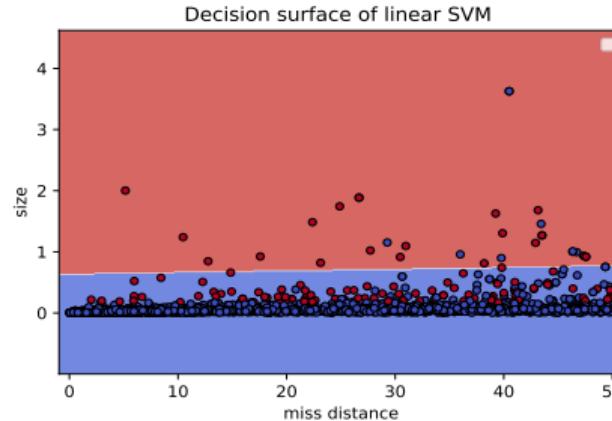
- Less NEO in the Summer
- Most are non-hazardous asteroids

# NASA Data Acquisition, Visualization, and Analysis

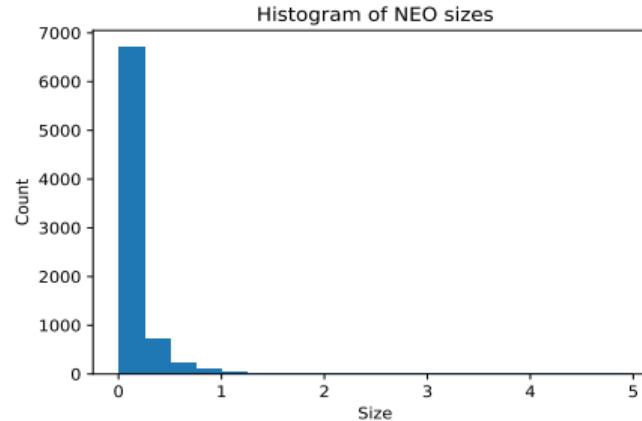


- Almost no correlation
- Near the earth, there are more hazardous objects.

# Support vector machine classifier



(a) Classifier



(b) Histogram size

- Hard to make clear boundaries for a classifier
- Most NEOs are small