DTU Data Analysis and Visualization [Summer 2023]

Magnus Ahasverusen, (s190600)

Zakir H. Shahoo (s194054)

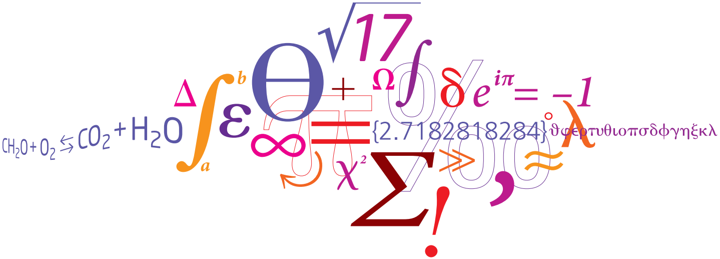
Chengjie Li (Jeff) (s231387)

Nicole Giordano (s231378)

Hannah Schweitzer (s231384)

Group 10

Technical University of Denmark (DTU)



DTU Environment
Department of Environmental Engineering

Outline

- Project 1: Analysis and Forecasting of NYC Taxi Rides
  - Task 1
  - Task 2
  - Task 3
  - Task 4
  - Task 5
- Project 2: NASA Data Acquisition, Visualization, and Analysis
  - Task 1
  - Task 2
  - Task 3
  - Task 3
  - Task 3
  - Task 4
  - Task 5

Task 1: Understanding the Data

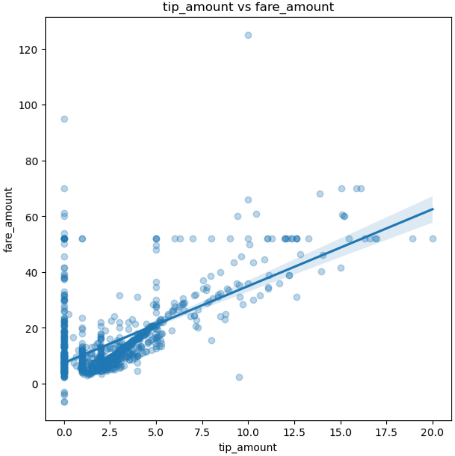
Important Data Given

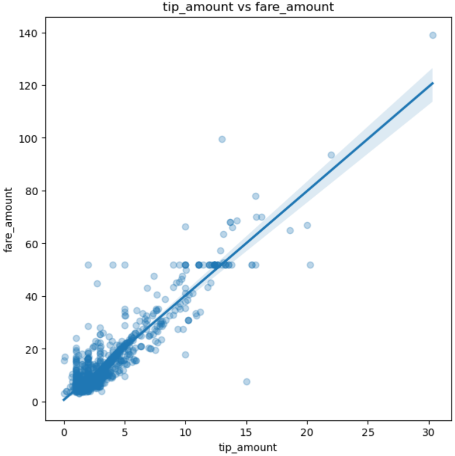
- Pickup/Dropoff Date and Time
- Pickup/Dropoff Location ID
- Passenger Count
- Trip Distance
- Fare Amount
- Tip Amount

Project 1: Analysis and Forecasting of NYC Taxi Rides

Task 2: Exploratory Data Analysis I

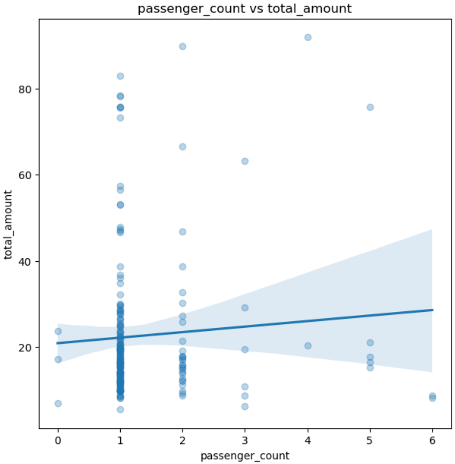
Exploring Tipping





Task 2: Exploratory Data Analysis II

Passenger Amount



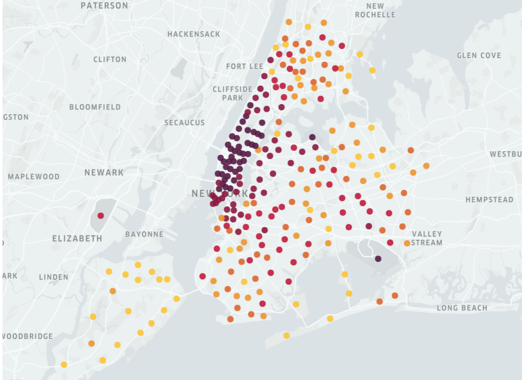
Task 2: Exploratory Data Analysis III

Other Curiosities

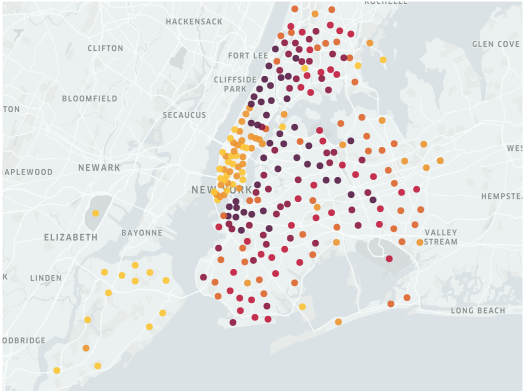
- Average Tip Amount (Yellow vs Green) = \$7.23 vs \$2.00
- Amount of Rides (Yellow vs Green) = 39,656,098 vs 840,402
- Average Distance (Yellow vs Green) = 5.96 miles vs 84.45 miles

Task 3: Spatial Analysis

Identifying PU/DO Hotspots using Kepler Maps



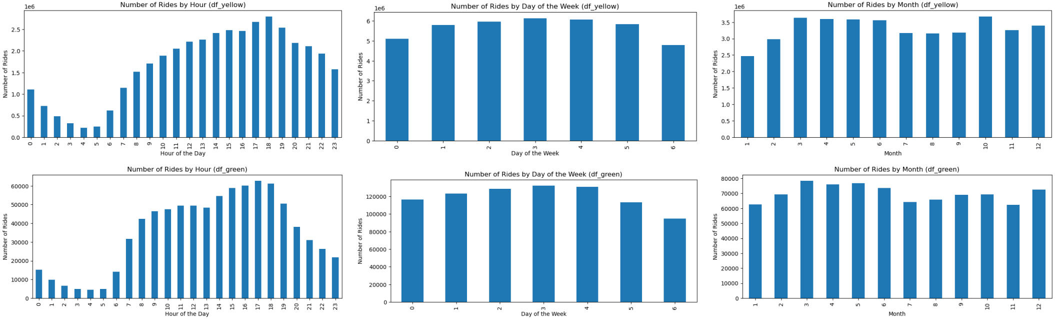
(a) Yellow Taxis



(b) Green Taxis

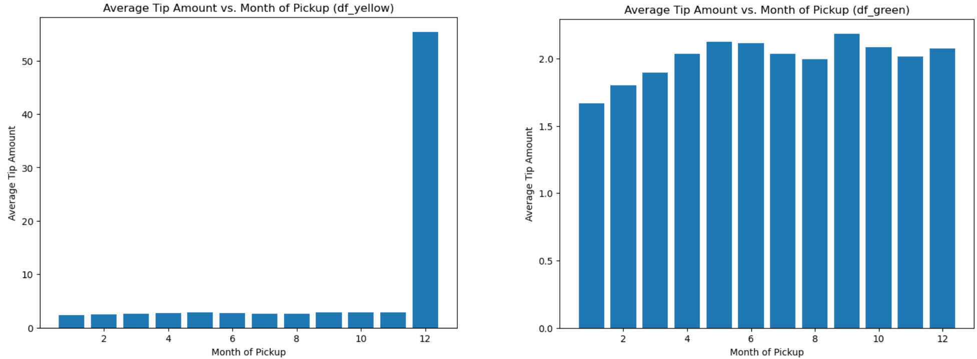
Task 4: Temporal Analysis I

Yellow vs Green Temporal



Task 4: Temporal Analysis II

Yellow vs Green (Month vs Tip Amount)

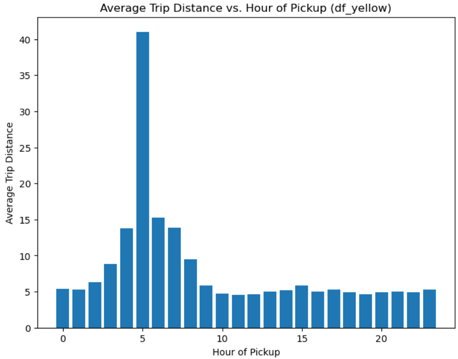


(a) Yellow

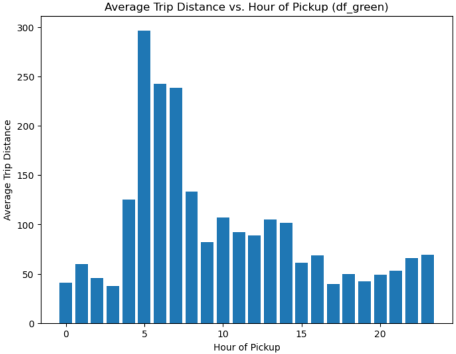
(b) Green

Task 4: Temporal Analysis III

Yellow vs Green (Distance vs Hour)



(a) Yellow

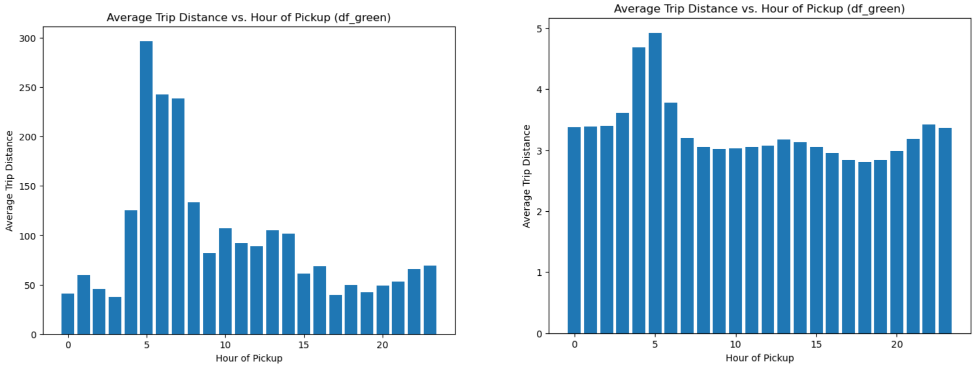


(b) Green

Project 1: Analysis and Forecasting of NYC Taxi Rides

Task 4: Temporal Analysis IV

Errors

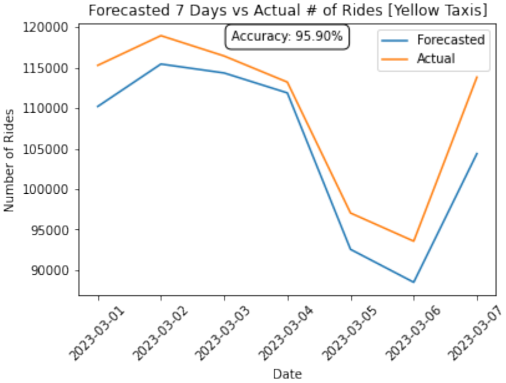


(a) Original

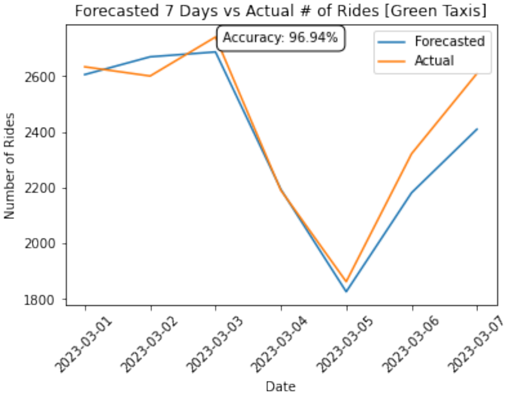
(b) Fixed

Task 5: Time Series Forecasting I

Forecasting 7 Days using Prophet trained on Jan/Feb 2023 data



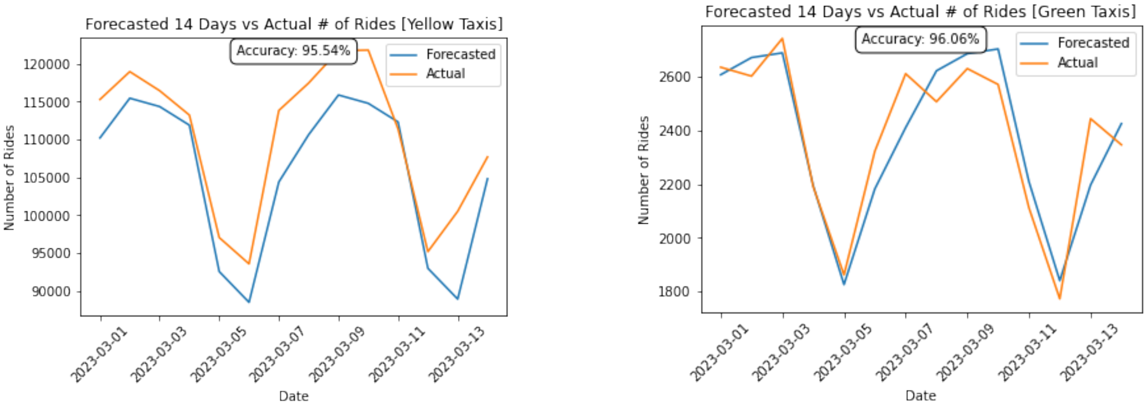
(a) Yellow Taxis



(b) Green Taxis

Task 5: Time Series Forecasting II

Forecasting 14 Days using Prophet trained on Jan/Feb 2023 data



(a) Yellow Taxis

(b) Green Taxis

Project 2

Project 2: NASA Data Acquisition, Visualization, and Analysis

Task 1 Understanding the NASA API and Data Collection

Data acquisition and Data analysis

- Fetched JSON data about Near Earth Objects (NEOs) using Python
- Cleaned up data by removing extra date in a "week" (8 days) of data, e.g. $neo\_df = neo\_df.drop\_duplicates()$
- Extracted and analyzed distinct data elements for each NEO via pd.json_normalize, e.g. $expanded\_neo\_df = pd.json\_normalize(neo, record\_path = neo\_entry\_date)$
- Converted extracted data into pd.DataFrame

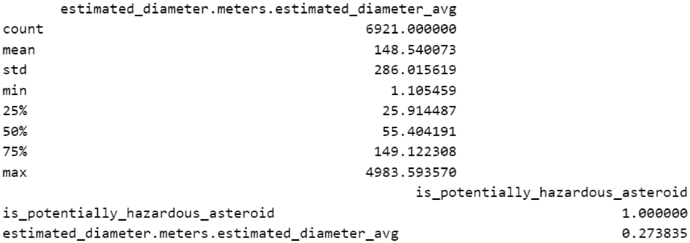
Task 2: Data Science and Analytics Works

Average size, hazards correlation, statistics

(a) The average size of the NEOs for each day

```
            average_size
2022-01-01    164.069506
2022-01-02    113.283811
2022-01-03     28.178929
2022-01-04     80.179344
```

(b) Statistical analysis + correl



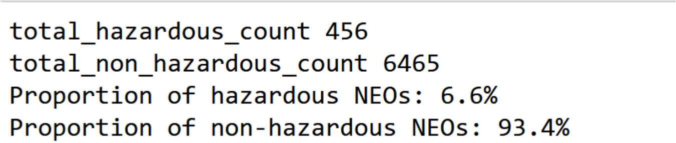
```
       estimated_diameter.meters.estimated_diameter_avg
count                                       6921.000000
mean                                         148.540073
std                                          286.015619
min                                            1.105459
25%                                           25.914487
50%                                           55.404191
75%                                          149.122308
max                                         4983.593570
                                                  is_potentially_hazardous_asteroid
is_potentially_hazardous_asteroid                                          1.000000
estimated_diameter.meters.estimated_diameter_avg                           0.273835
```

Project 2: NASA Data Acquisition, Visualization, and Analysis

Data Science and Analytics Works

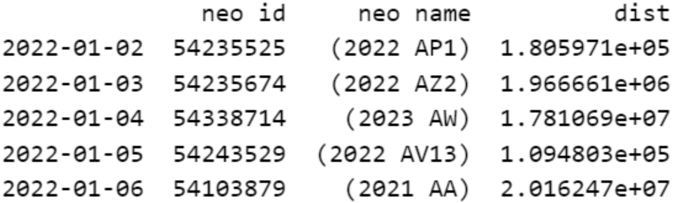
Closest approach size-potential hazard correlation.

(a) Proportion of NEOs that are potentially hazardous.



```
total_hazardous_count 456
total_non_hazardous_count 6465
Proportion of hazardous NEOs: 6.6%
Proportion of non-hazardous NEOs: 93.4%
```

(b) NEOs with the closest approach distance for each day.

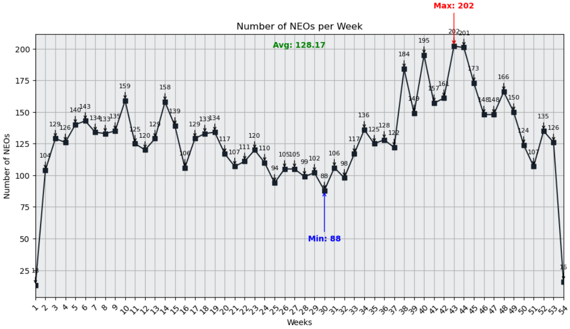


```
              neo id     neo name          dist
2022-01-02  54235525   (2022 AP1)  1.805971e+05
2022-01-03  54235674   (2022 AZ2)  1.966661e+06
2022-01-04  54338714    (2023 AW)  1.781069e+07
2022-01-05  54243529  (2022 AV13)  1.094803e+05
2022-01-06  54103879    (2021 AA)  2.016247e+07
```

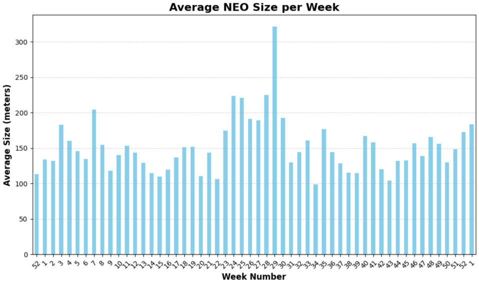
Project 2: NASA Data Acquisition, Visualization, and Analysis

Task 3: Data Visualization Part A (I)

line plot (count/week), bar plot (avg. size)



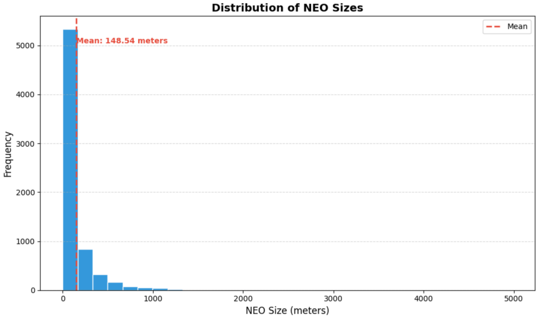
(a) line plot of the number of NEOs per week



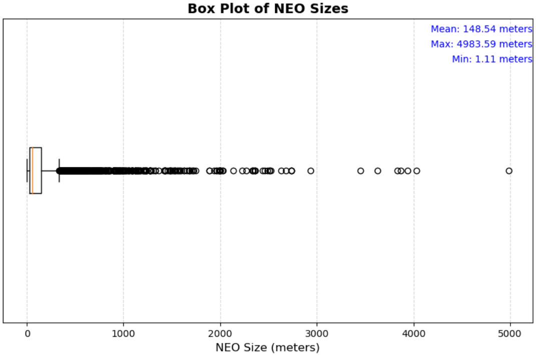
(b) bar plot of the average NEO size per week

Task 3: Data Visualization Part A (II)

Histogram (NEOs size),Seaborn box plot (NEO sizes)



(a) Histogram of the distribution of NEO sizes

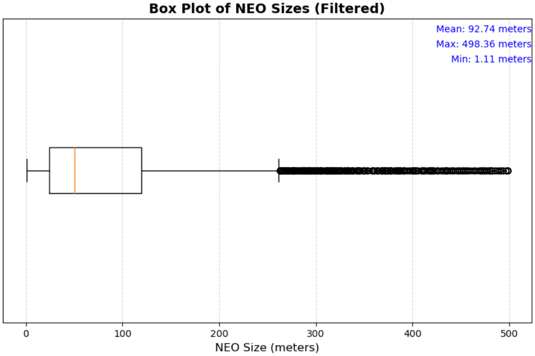


(b) box plot of the NEO sizes

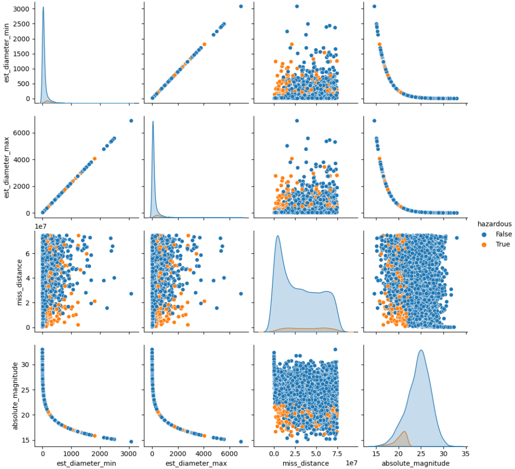
Project 2: NASA Data Acquisition, Visualization, and Analysis

Task 3: Data Visualization Part A (III)

Pairwise Relationships and Hazardousness in NEO Data



(a) Histogram of the distribution of NEO sizes

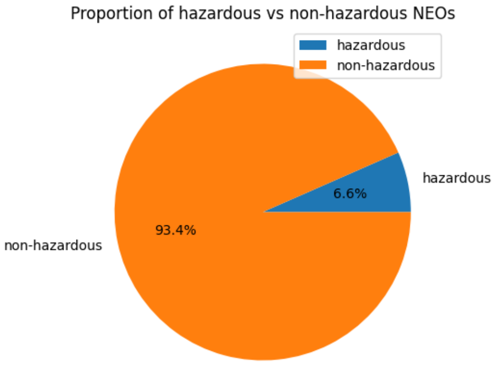


(b) Pair plot that visualizes the relationships between different variables

Task 4: Data Visualization Part B

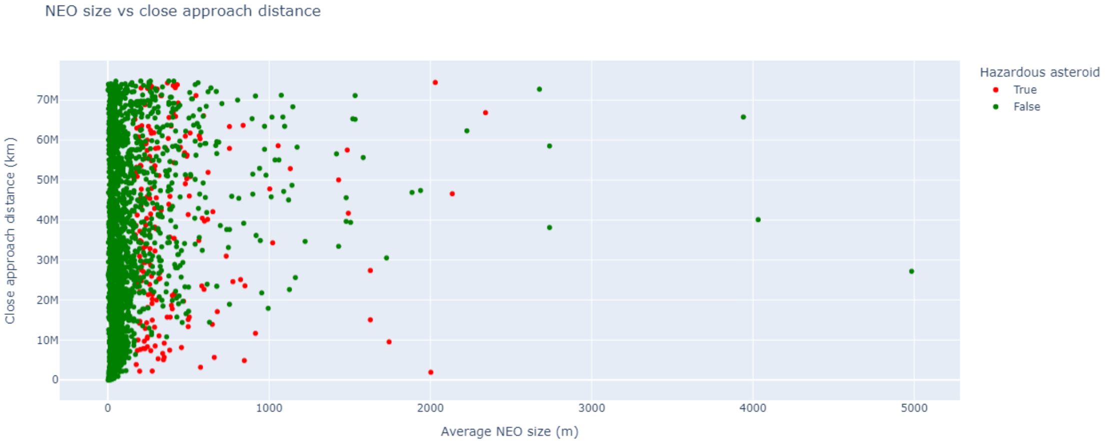
Pie chart: Hazardous vs. non-hazardous NEOs

- Created a pie chart of the proportion of hazardous vs non-hazardous NEOs



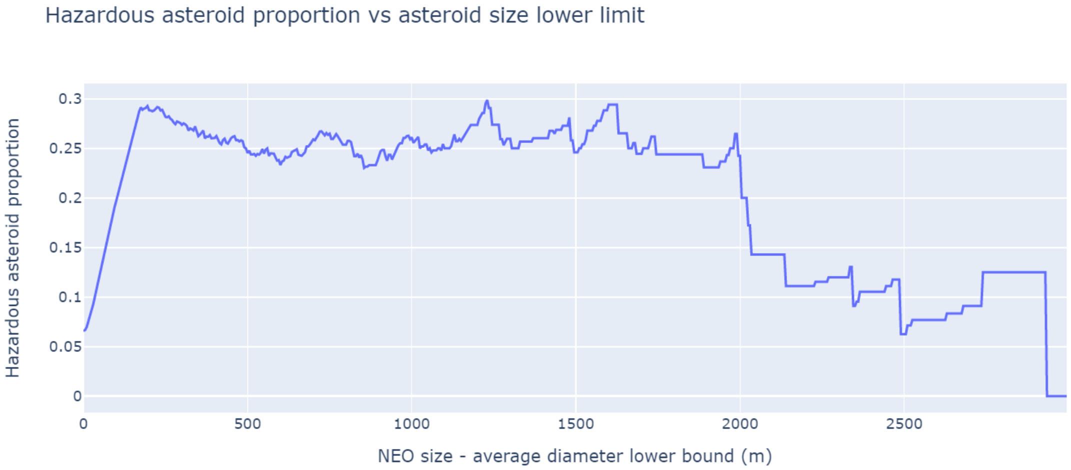
Task 4: Data Visualization Part B

Scatter plot with hover functionality for NEO data using Plotly



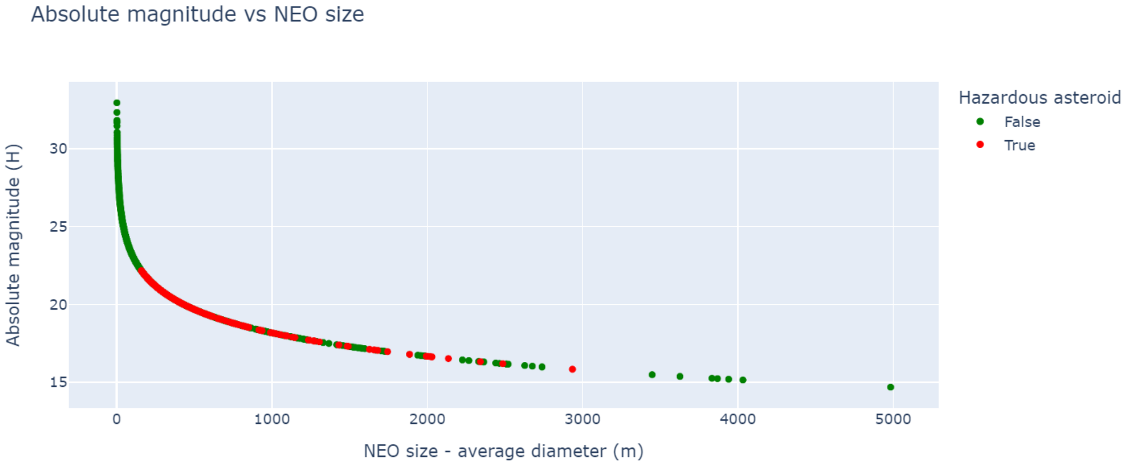
Task 4: Data Visualization Part B

Line chart: Hazardous asteroid proportion vs. NEO size's lower limit



Task 4: Data Visualization Part B

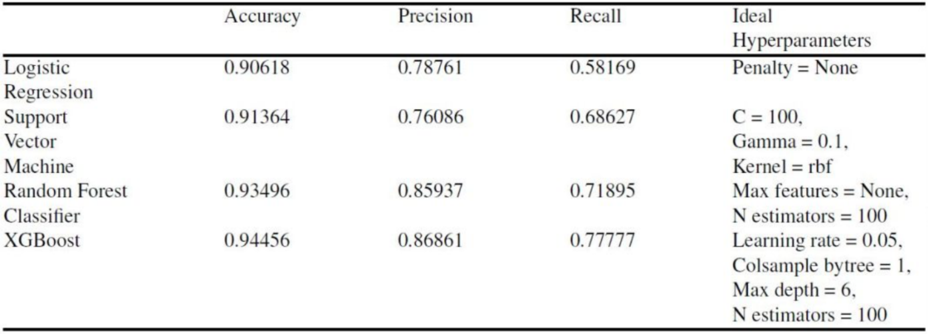
Interesting finding - Scatter plot: Absolute magnitude vs. NEO size



Task 5

Findings from NASA data visualizations to make predictions or recommendations

Classifying Hazardous and Non-Hazardous Asteroids Using Machine Learning



	Accuracy	Precision	Recall	Ideal Hyperparameters
Logistic Regression	0.90618	0.78761	0.58169	Penalty = None

Support Vector Machine	0.91364	0.76086	0.68627	C = 100,
				Gamma = 0.1,
				Kernel = rbf

Random Forest Classifier	0.93496	0.85937	0.71895	Max features = None, N estimators = 100
XGBoost	0.94456	0.86861	0.77777	Learning rate = 0.05, Colsample bytree = 1, Max depth = 6, N estimators = 100

Table 1* | *Results on the Test Set

Source for scientific paper.(NJS)