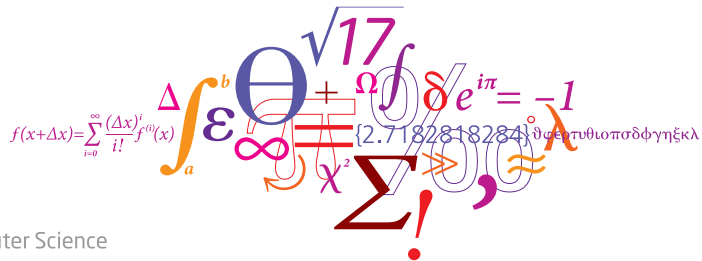
# 62444 - Data Visualization and Analysis

Group 21 - Project 1 & Project 2

Technical University of Denmark (DTU)




DTU Compute
Department of Applied Mathematics and Computer Science

# Outline

- Project 1 - Analysis and Forecasting of NYC Taxi Rides
  - Task 1 - Understanding the data
  - Task 2 - Exploratory Data Analysis
  - Task 3 - Spatial Analysis
  - Task 4 - Temporal Analysis
  - Task 5 - Time-Series Forecasting
- Project 2 - NASA Data Acquisation, Visualization, and Analysis
  - Task 2 - Data Analysis
  - Task 3 - Data Visualization Part A
  - Task 4 - Data Visualization Part B

# Project 1 - Analysis and Forecasting of NYC Taxi Rides

# Task 1 - Understanding the data I

- Dataframe of Taxis
- Months [01, 02, 03]
- Year [2023]
- Yellow and green taxis

5 rows × 19 columns

| | VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 2022-01-01 00:35:40 | 2022-01-01 00:53:29 | 2.0 | 3.80 | 1.0 | N |
| 1 | 1 | 2022-01-01 00:33:43 | 2022-01-01 00:42:07 | 1.0 | 2.10 | 1.0 | N |
| 2 | 2 | 2022-01-01 00:53:21 | 2022-01-01 01:02:19 | 1.0 | 0.97 | 1.0 | N |
| 3 | 2 | 2022-01-01 00:25:21 | 2022-01-01 00:35:23 | 1.0 | 1.09 | 1.0 | N |
| 4 | 2 | 2022-01-01 00:36:48 | 2022-01-01 01:14:20 | 1.0 | 4.30 | 1.0 | N |

Figure: Dataframe for yellow taxies

5 rows × 20 columns

| | VendorID | lpep_pickup_datetime | lpep_dropoff_datetime | store_and_fwd_flag | RatecodeID | PULocationID | DOLocationID |
|---|---|---|---|---|---|---|---|
| 0 | 2 | 2022-01-01 00:14:21 | 2022-01-01 00:15:33 | N | 1.0 | 42 | 42 |
| 1 | 1 | 2022-01-01 00:20:55 | 2022-01-01 00:29:38 | N | 1.0 | 116 | 41 |
| 2 | 1 | 2022-01-01 00:57:02 | 2022-01-01 01:13:14 | N | 1.0 | 41 | 140 |
| 3 | 2 | 2022-01-01 00:07:42 | 2022-01-01 00:15:57 | N | 1.0 | 181 | 181 |
| 4 | 2 | 2022-01-01 00:07:50 | 2022-01-01 00:28:52 | N | 1.0 | 33 | 170 |

Figure: Dataframe for green taxies

# Task 2 - Exploratory Data Analysis

- Distibution Analysis
  - Histograms
- Relationship Analysis
  - Scatter Plots (Sample size)

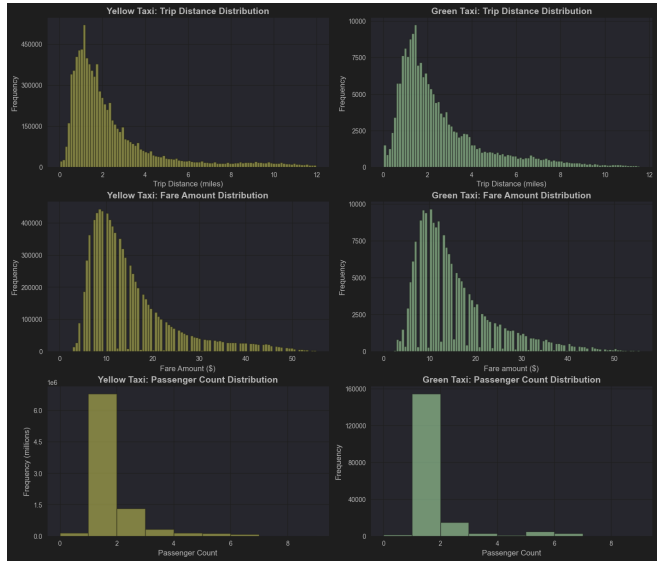



Figure: Histograms

# Task 2 - Exploratory Data Analysis

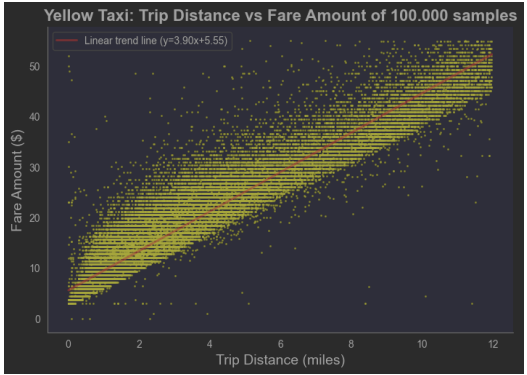



Figure: Trip Distance (miles)


Figure: Trip Distance (miles)

# Task 3 - Spatial Analysis

- Green Taxi's on the left
- Yellow Taxi's on the right
- As the color gets lighter, the number of trips increases.
- Interaktiv map

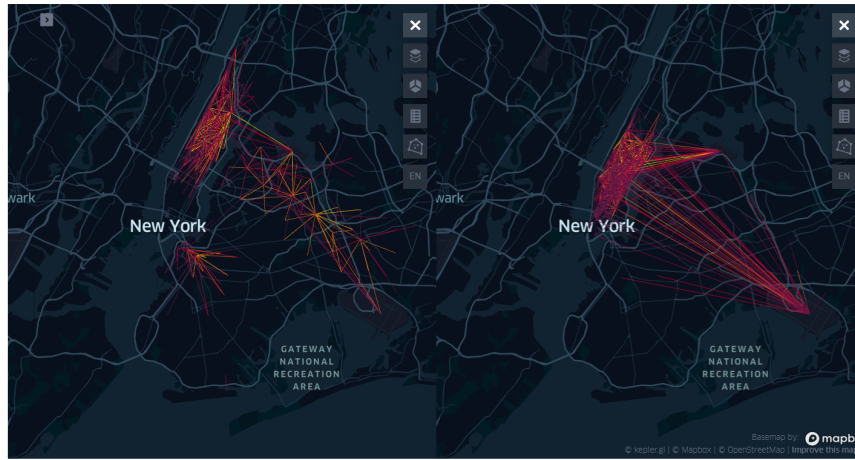



Figure: Visualization of green and yellow taxi trips

# Task 4 - Temporal Analysis

- Yellow and Green Taxis
- Hourly, weekdays, months
- Trip count
- Bar graph

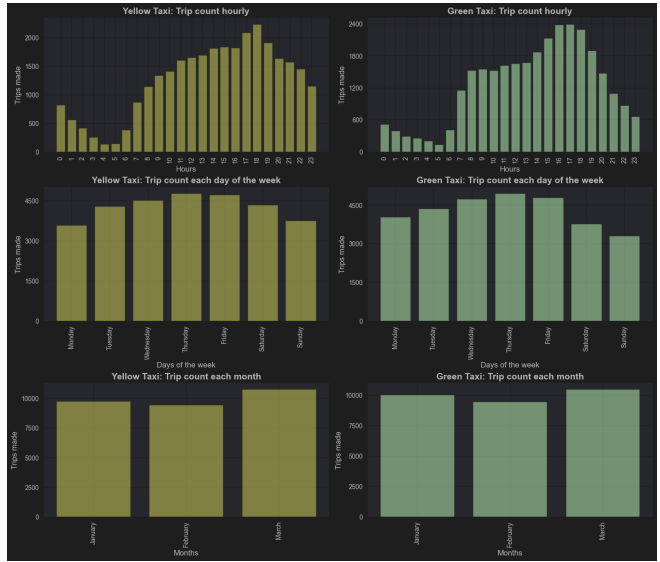
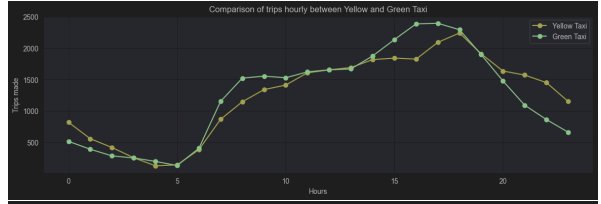
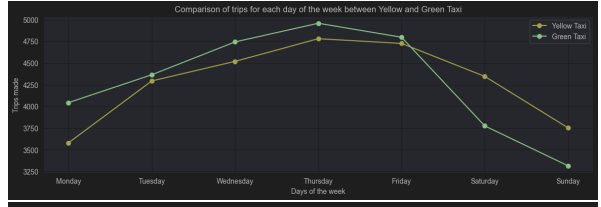


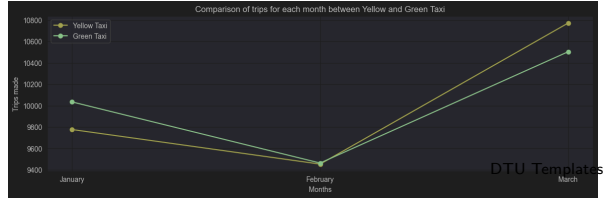

Figure: Histogram Plots

# Task 4 - Temporal Analysis

- Yellow and Green Taxis
- Hourly, weekdays, months
- Trip count
- Line graph



# Task 4 - Temporal Analysis

- Yellow and Green Taxis
- Hourly, weekdays, months
- Average distance traveled
- Bar graph

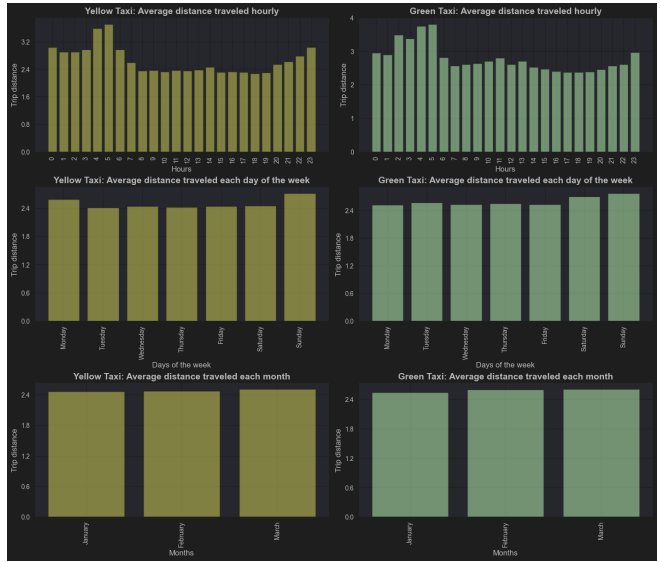
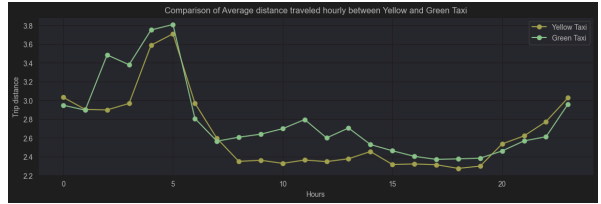
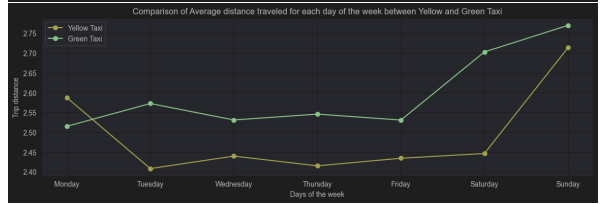


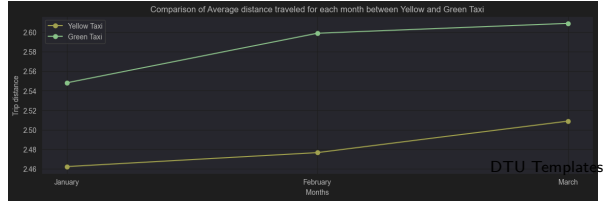

Figure: Histogram Plots

# Task 4 - Temporal Analysis

- Yellow and Green Taxis
- Hourly, weekdays, months
- Average distance traveled
- Line graph



# Task 4 - Temporal Analysis

- Yellow and Green Taxis
- Hourly, weekdays, months
- Fares
- Bar graph

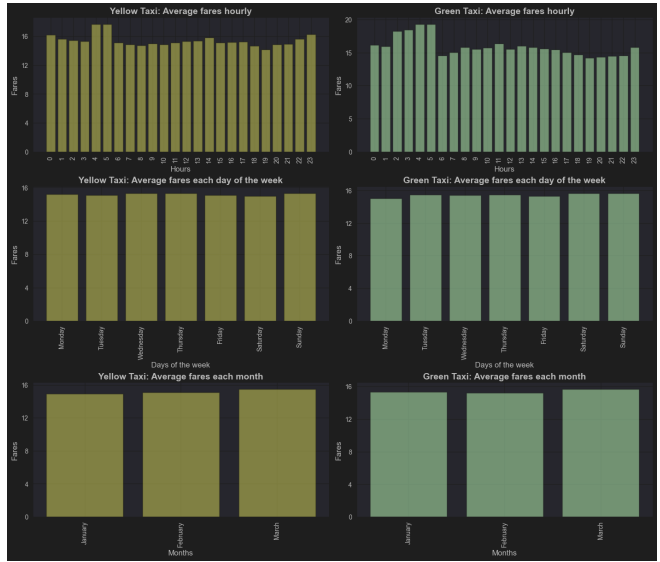
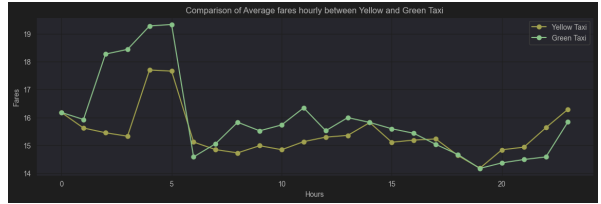
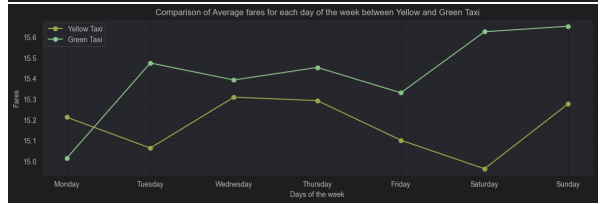


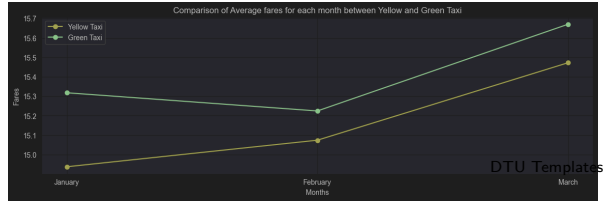

Figure: Histogram Plots

# Task 4 - Temporal Analysis

- Yellow and Green Taxis
- Hourly, weekdays, months
- Fares
- Line graph

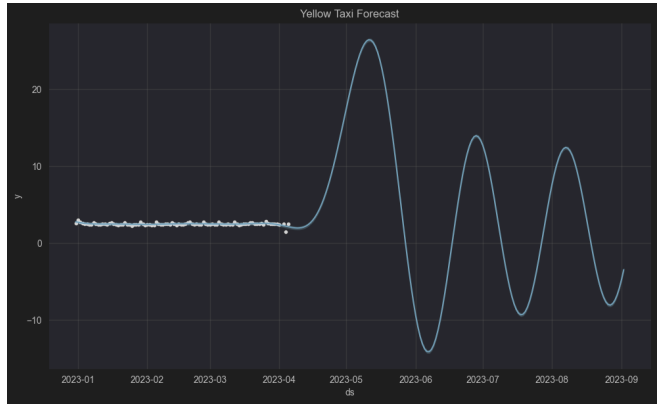


# Task 5 - Time-Series Forecasting

- 150 days data, 150 days future dataframe
- Fluctuation in the demand for taxi rides
- Unrealistic negative forecast values



# Task 5 - Time-Series Forecasting

- 150 days data, 150 days future dataframe
- Fluctuation in the demand for taxi rides
- Unrealistic negative forecast values

# Task 5 - Time-Series Forecasting

- Yearly seasonality showcased for yellow forecast components
- Upward trend - increasing trend over the forecast period

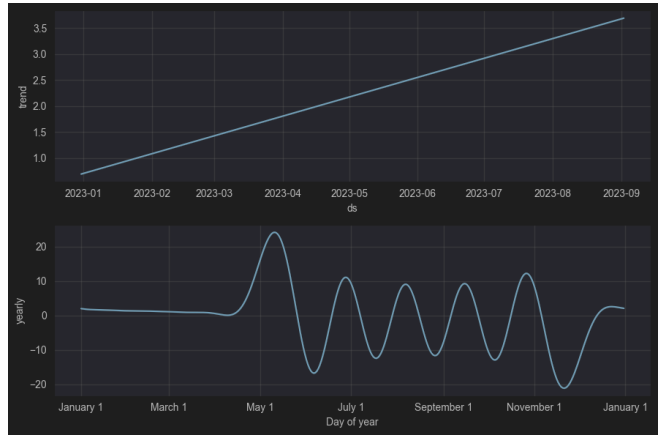



Figure: Yellow Forecast Components

# Task 5 - Time-Series Forecasting

- Yearly seasonality showcased for green forecast components
- Downward trend - decreasing trend over the forecast period

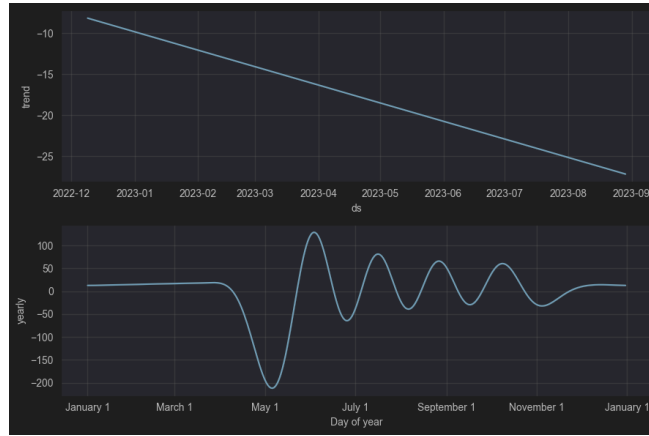



Figure: Green Forecast Components

# Task 5 - Time-Series Forecasting (Cross-Validation)

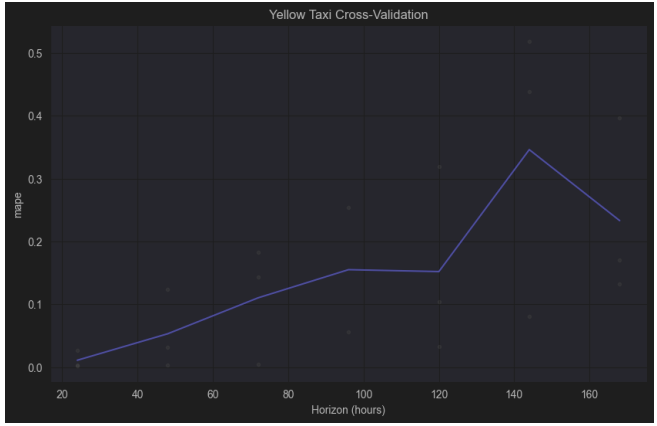



Figure: Yellow Cross-Validation

# Task 5 - Time-Series Forecasting (Cross-Validation)

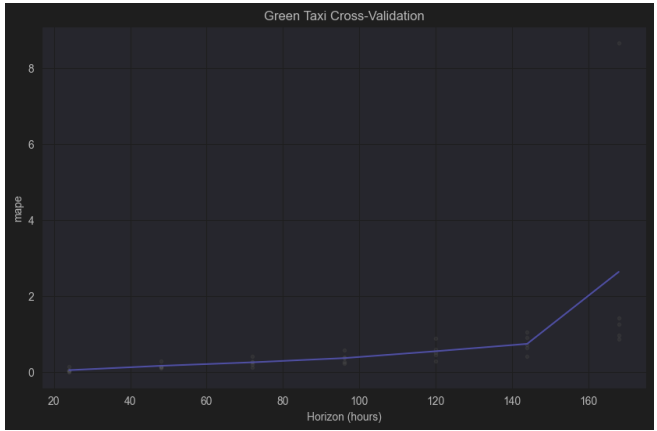



Figure: Yellow Cross-Validation

# Project 2 - NASA Data Acquisation, Visualization, and Analysis

# Task 2 - Data Analysis

6 rows × 5 columns

CSV

| | absolute_magnitude_h | estimated_diameter_min (m) | estimated_diameter_max (m) | relative_velocity (km/s) | miss_distance (lunar) |
|---|---|---|---|---|---|
| mean | 22.879868 | 164.607334 | 368.073189 | 14.430863 | 114.663061 |
| std | 2.971990 | 253.652908 | 567.185144 | 7.085677 | 51.810334 |
| min | 14.670000 | 1.214940 | 2.716689 | 0.536816 | 1.784416 |
| median | 22.705000 | 76.481453 | 171.017727 | 13.715671 | 118.694940 |
| max | 31.700000 | 3094.246986 | 6918.946600 | 52.005854 | 194.480567 |
| mode (rounded) | 21.000000 | 13.000000 | 36.000000 | 9.000000 | 185.000000 |

Figure: General statistical analysis

The 1500 objects with smallest maximum diameter:

1 row × 1 columns

| potentially_hazardous | count |
|---|---|
| False | 1500 |

Figure: The 1500 smallest objects

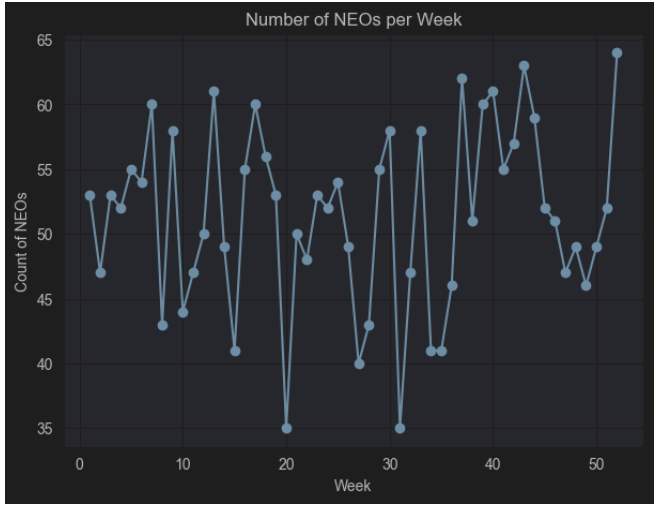
The 1000 objects with greatest maximum diameter:

2 rows × 1 columns

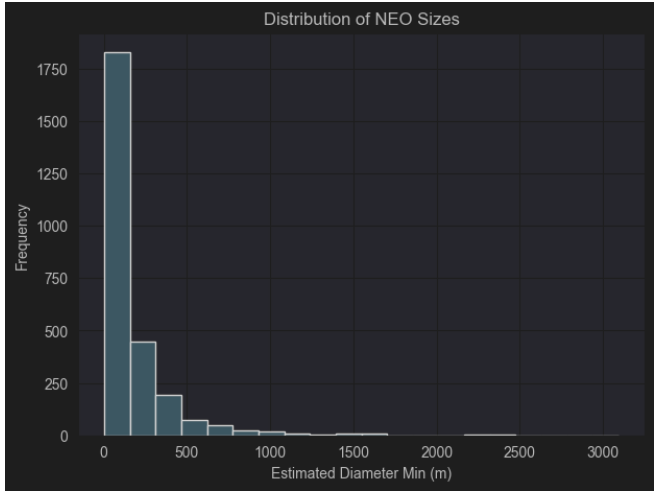
| potentially_hazardous | count |
|---|---|
| False | 684 |
| True | 316 |

Figure: The 1000 largest objects

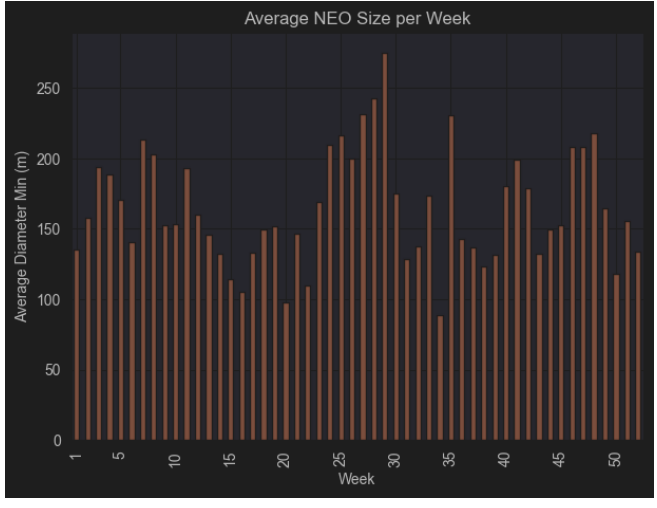
# Task 3 - Data Visualization Part A



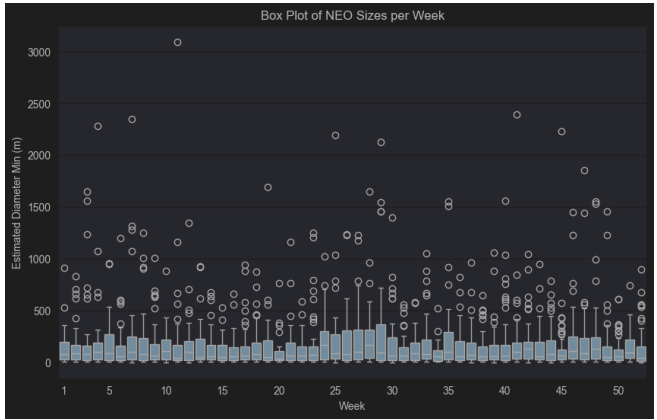
# Task 3 - Data Visualization Part A



# Task 3 - Data Visualization Part A

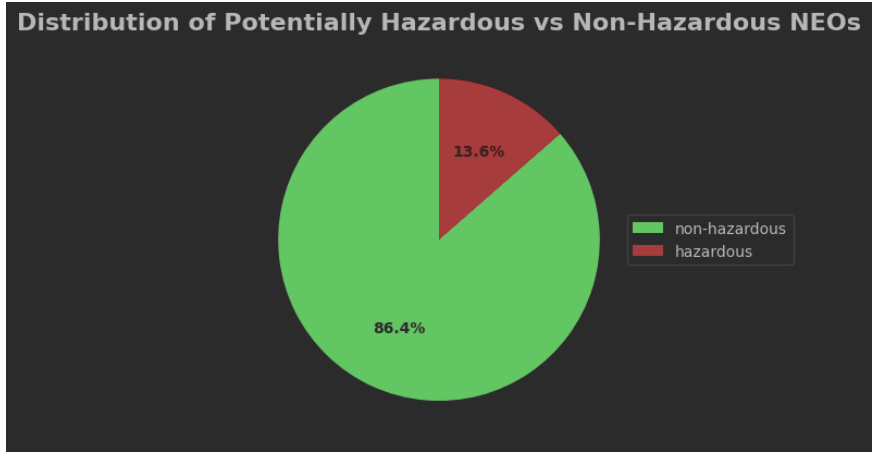


# Task 3 - Data Visualization Part A



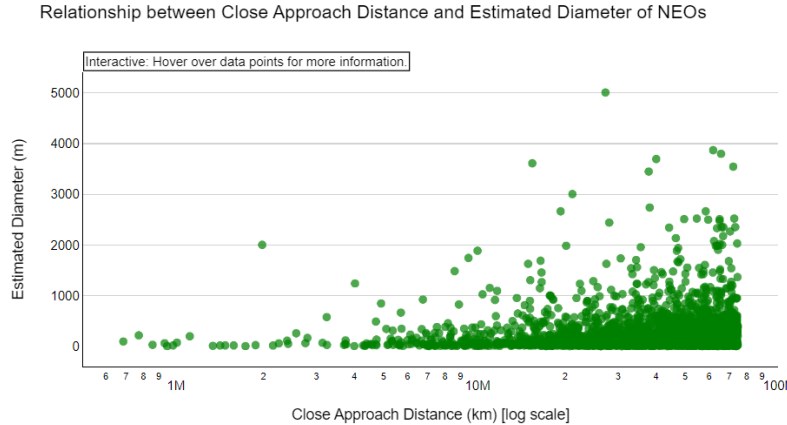
# Task 4 - Data Visualization Part B

- Overwhelming amount of non-hazadous



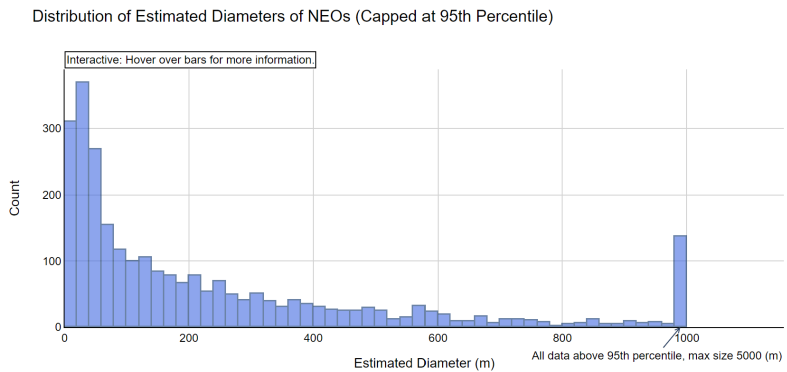
# Task 4 - Data Visualization Part B

- Large amount of NEO's are small and far away
- Needed to log scale the distance



# Task 4 - Data Visualization Part B

- Interactive histogram explore the distribution of NEO sizes.
- Histogram winsorized at 95th percentile



# Task 4 - Data Visualization Part B

- Interactive box plot explore the distribution of NEO sizes.
- Box plot logarithmic scale