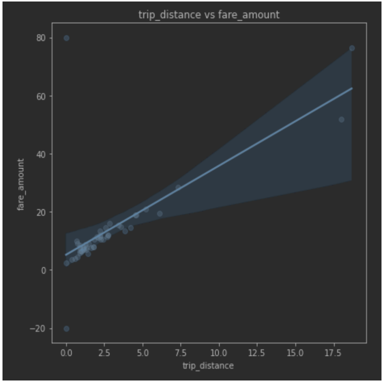
DTU

DTU LaTeX Support Group - latex.dtu.dk DTU

Beamer template

Project 1: Analysis and Forecasting of NYC Taxi Rides

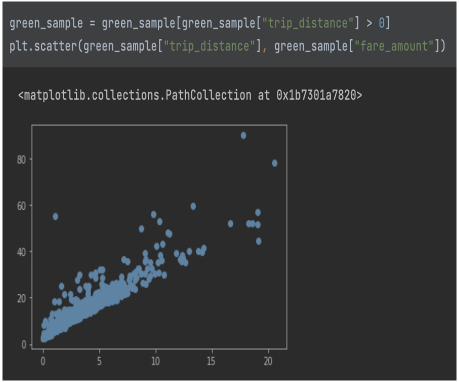
Understanding the Data



Insight in the relation between distance and fare amount
The Scatter Plot shows a linear relationship between the trip distance and the fare amount

Exploratory Data Analysis

Patterns and relationships in the data



```
green_sample = green_sample[green_sample["trip_distance"] > 0]
plt.scatter(green_sample["trip_distance"], green_sample["fare_amount"])
```

```
<matplotlib.collections.PathCollection at 0x1b7301a7820>
```

The plot shows a linear relationship between the trip distance and the fare amount

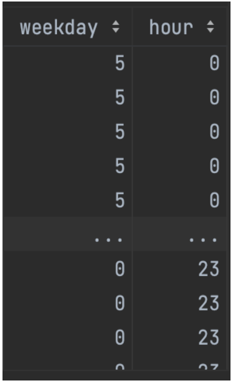
Several trips have a trip distance of zero:

those were filtered out

Outliers could be due to special fees

Spatial Analysis (Kepler)

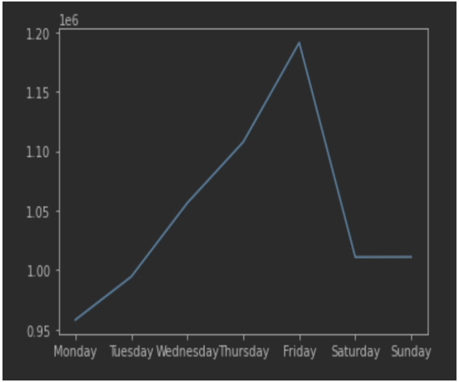
Temporal Analysis



weekday	hour
5	0
5	0
5	0
5	0
5	0
...	...
0	23
0	23
0	23
0	23

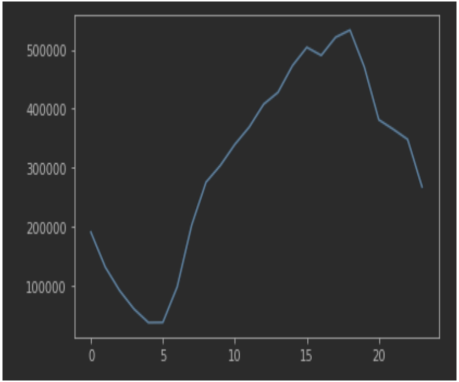
- Temporal patterns
- Added culumn (timeframe)

Number of taxi rides for each weekday



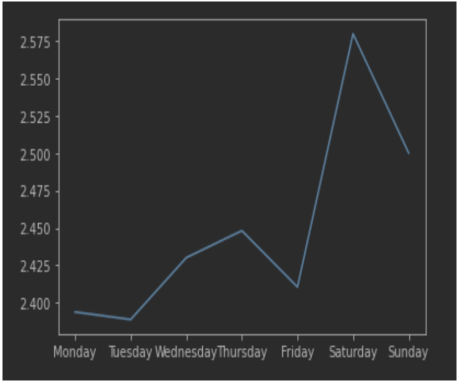
- Saturday and Sunday similar demand
- increase of demand during the week and tops Friday.

Number of passengers riding the taxis for each hour



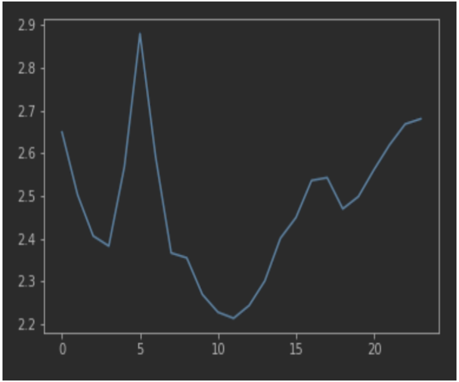
- Increase during day (5-18)
- Fewest during night

When is the tip amount the highest?



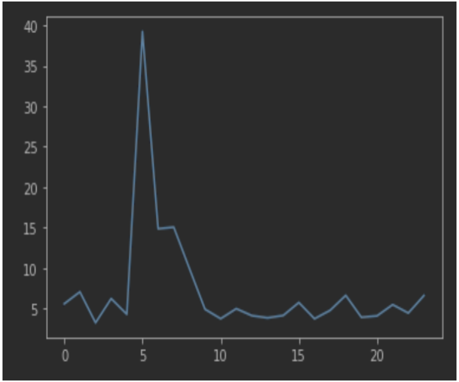
- Weekends

The plot shows the average tip amount on each hour



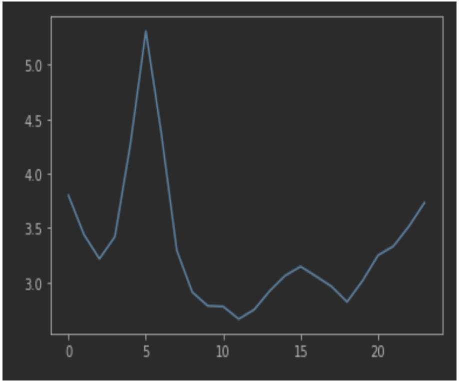
- Worst during day, best during night
- Tops at 5

Average distance on each hour.



- significant larger at 5
- warrants further investigation

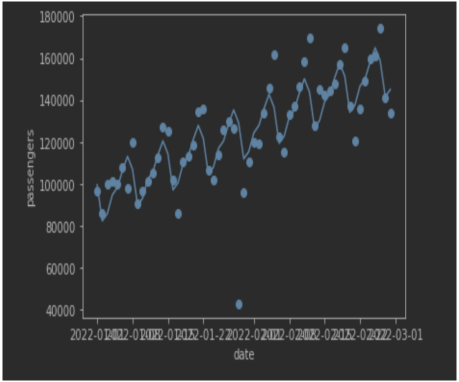
average distance on each hour, without the outliers



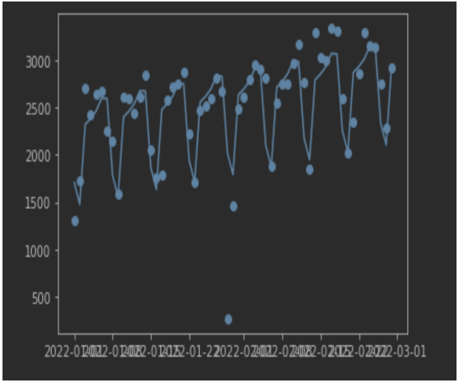
< 100

- largest in the morning. (correlated to the large tips?)

Time-Series Forecasting



Yellow taxis
Predictions are made with data from 15Th of February
Model find an increase in passenger amount



Green taxis
The algorithm finds the temporal patterns in an good way however there is room for improvement.

Project 2: NASA Data Acquisation, Visualization, and Analysis

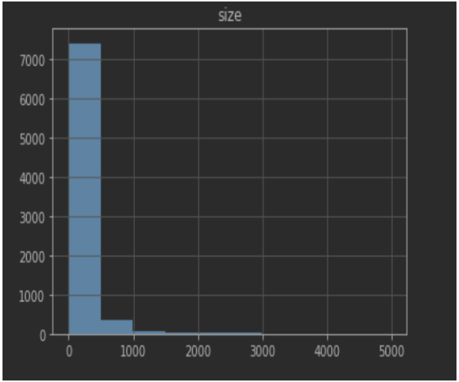
Understanding the Data

We flatten the JSON-data and create a pandas dataframe

We pick the following features: size, is hazardous, date, closest approach distance. Later on we include the velocity of the NEO

Data Analysis

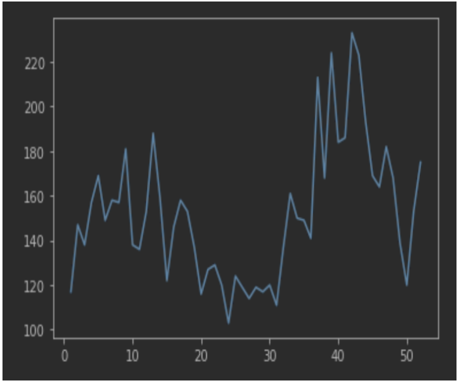
Neo’s observed with size



- exponential distribution

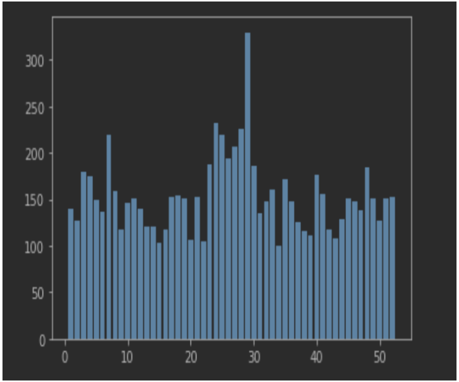
Data Visualization

NEO’s observed weekly

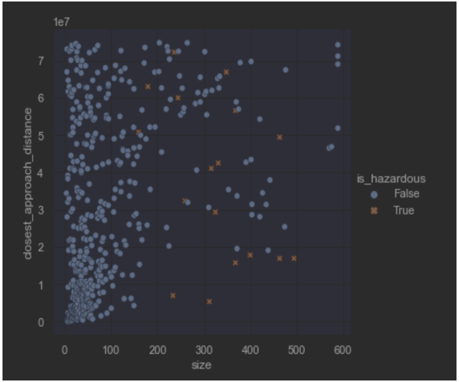


- Season

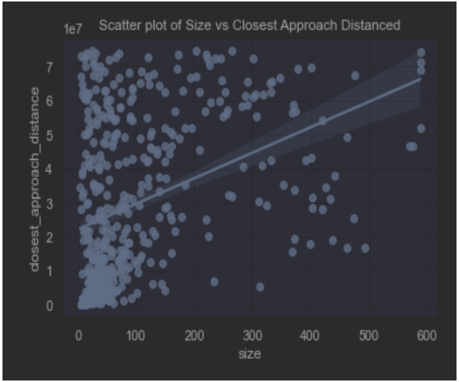
Average size observed weekly



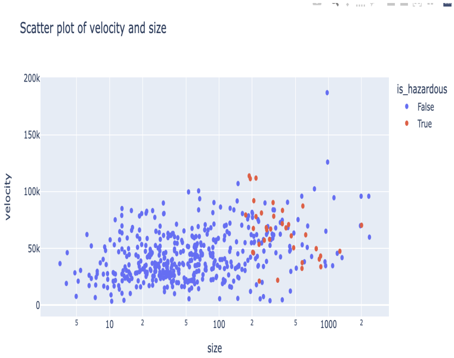
Distance and size
Hazardous or not



- below 150



Connection between size and closest approach distance?
Not clear at all



This plotly chart showing the hazardousness against size and velocity
Conclusion: Size is the predominant factor

Summary