Exploring the Cosmos: Data Science Initiatives for SpaceX

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Gathering data through web scraping and the SpaceX API.
- Conducting Exploratory Data Analysis (EDA), encompassing data cleaning, visualization, and interactive visual analytics.
- Employing Machine Learning for prediction.
- Summarizing all findings.
- Valuable data were sourced from publicly available repositories.
- EDA facilitated the identification of optimal predictors for launch success.
- Machine Learning prediction revealed the most effective model for determining key factors influencing launch success, leveraging all available data.

Introduction:

- The aim is to assess the potential competitiveness of the emerging company Space Y in comparison to Space X.
- Desired outcomes include:
- Determining the optimal method for estimating the overall cost of launches, achieved through predicting successful landings of the first stage of rockets.
- Identifying the most favorable location for conducting launches.

Methodology:

- Data collection methodology:
- Data from Space X was obtained from 2 sources:

- Space X API (https://api.spacexdata.com/v4/rockets/)
- WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

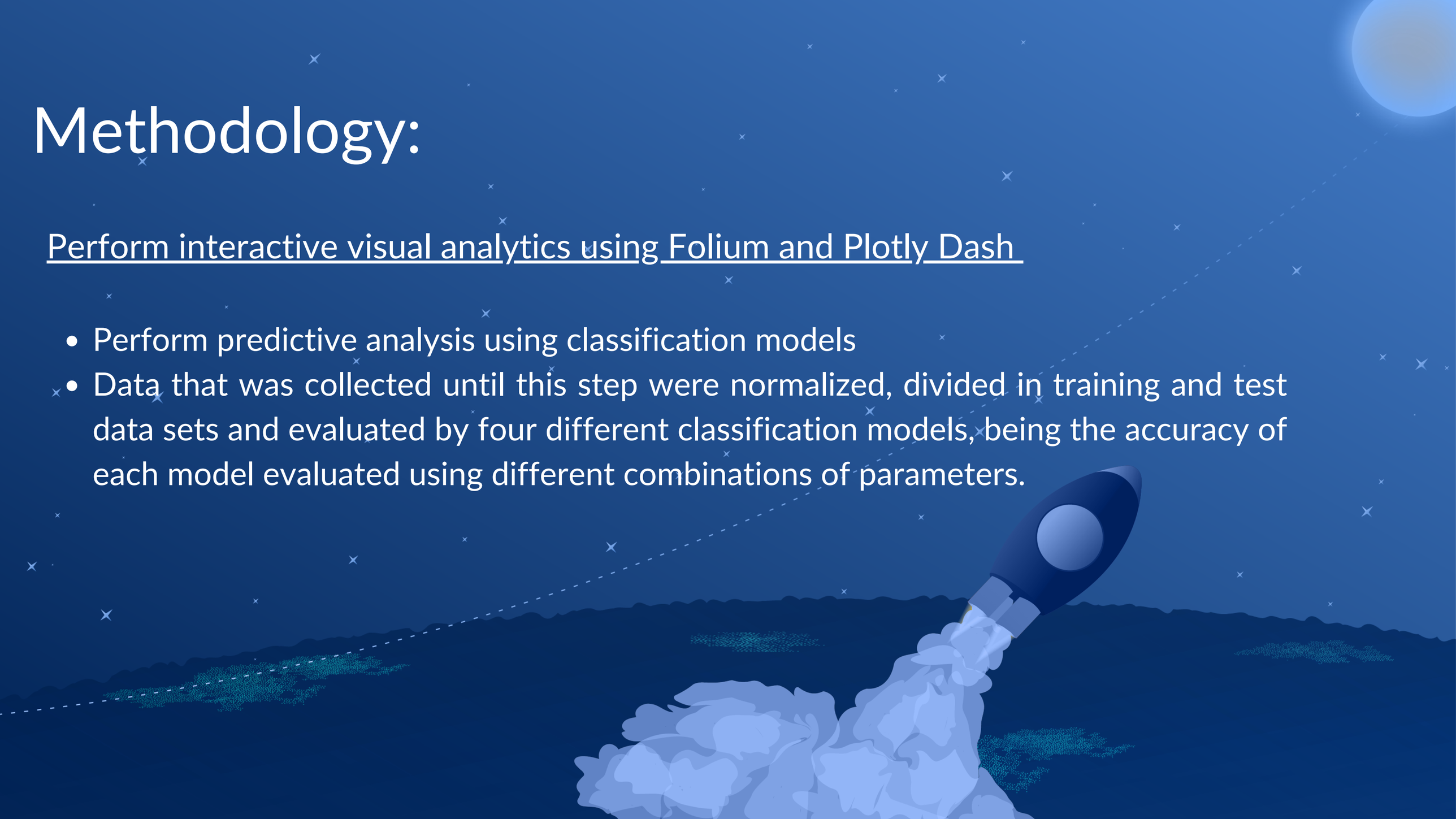
Perform data wrangling

- Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features •
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology:

Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models
- Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

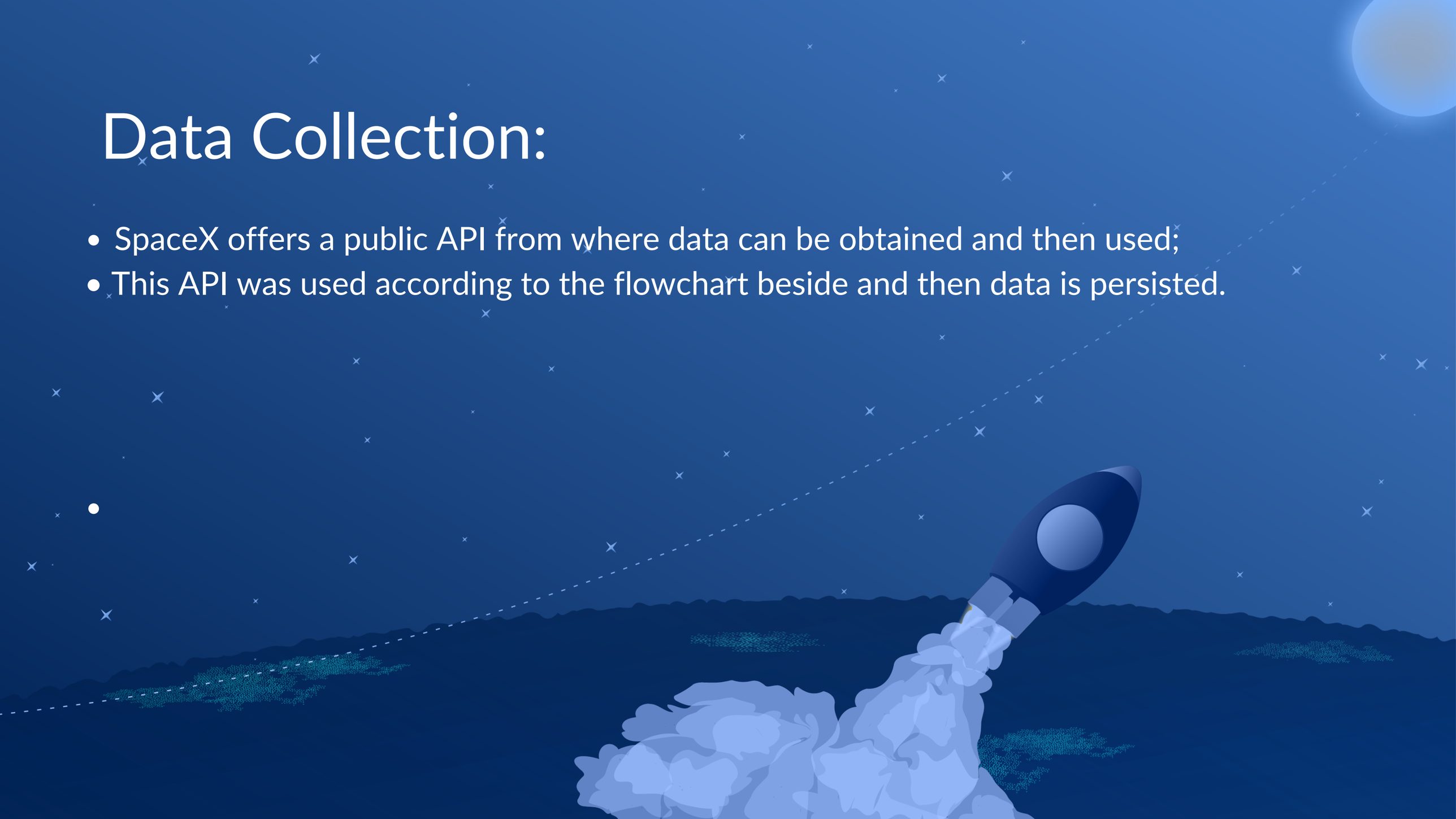


Data Collection:

- Datasets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) andfromWikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

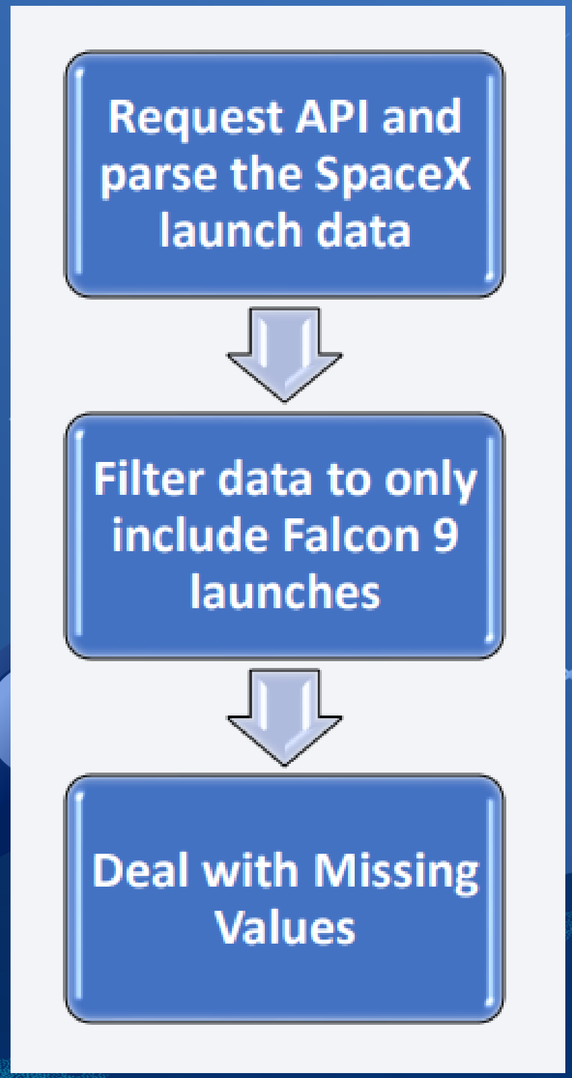
Data Collection:

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
-



Data Collection:

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.

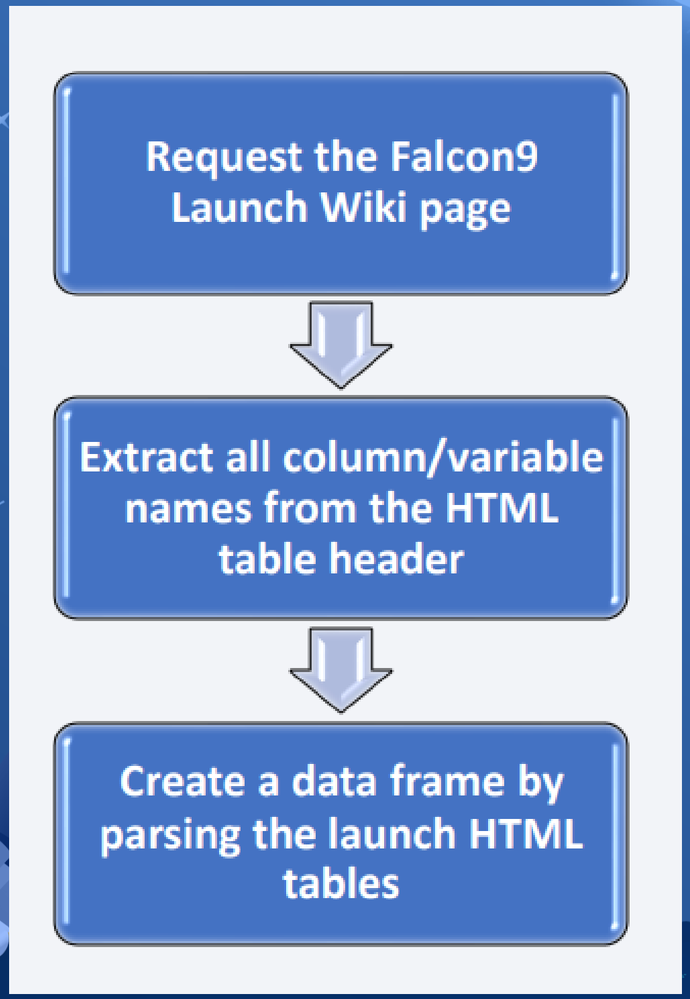


Source code

- https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/Data%20Collection%20API.ipynb

Data Collection - Scraping:

Data from SpaceX launches obtained from Wikipedia;.

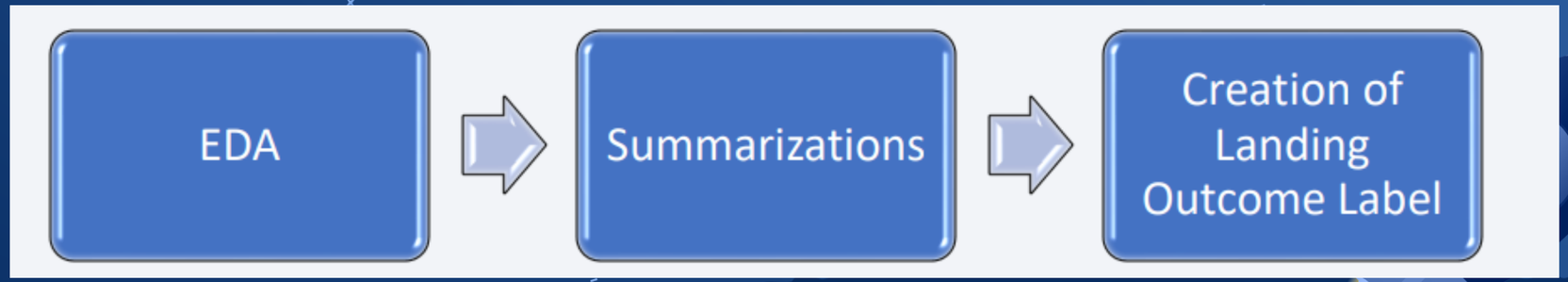


Source code
https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling:

Initially some Exploratory Data Analysis (EDA) was performed on the dataset.

- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



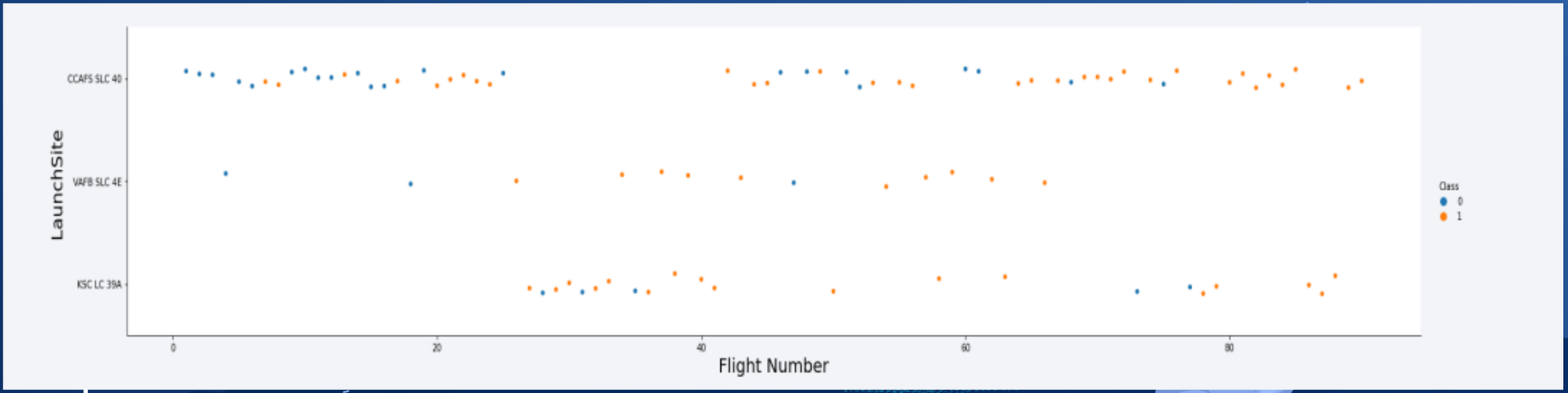
Source code

- https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/Data%20Wrangling.ipynb

EDA with Data Visualization:

To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:

Payload Mass X Flight Number, Launch Site X Flight Number,
Launch Site X Payload Mass, Orbit and Flight Number
, Payload and Orbit



Source code

- https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL:

• The following SQL queries were performed: •
Names of the unique launch sites in the space mission;
• Top 5 launch sites whose name begin with the string 'CCA';
• Total payload mass carried by boosters launched by NASA (CRS);
• Average payload mass carried by booster version F9 v1.1;
• Date when the first successful landing outcome in ground pad was achieved;
• Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg; • Total number of successful and failure mission outcomes;
• Names of the booster versions which have carried the maximum payload mass;
• Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
• Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/EDA.ipynb

Build an Interactive Map with Folium:

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates.

Source code

- https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash:

The following graphs and plots were used to visualize data

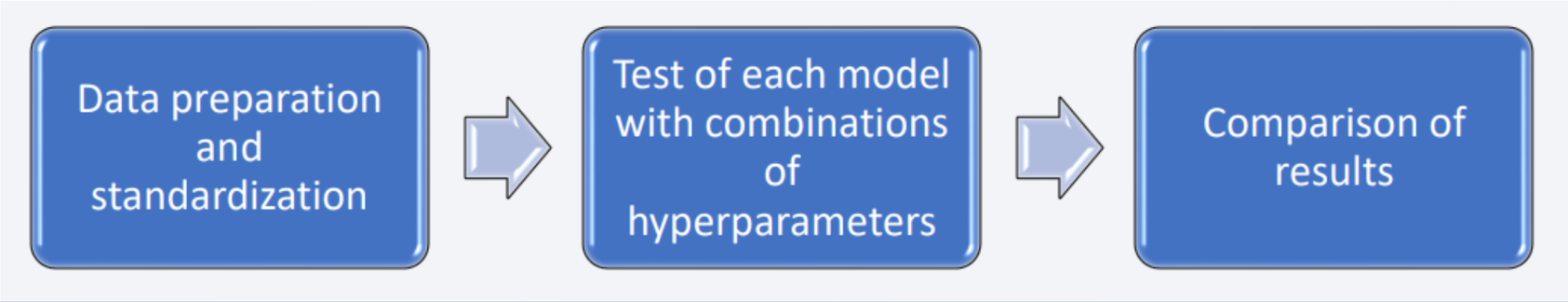
- Percentage of launches by site
- Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads. .

Source code

- https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/spacex_dash_app.py

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors. .



Source code

- https://github.com/preethakl/IBMdatascience/blob/881a0c927f55af670b801de4541aec4f2936dabc/SpaceX_Capstone_Project/Machine%20Learning%20Prediction.ipynb

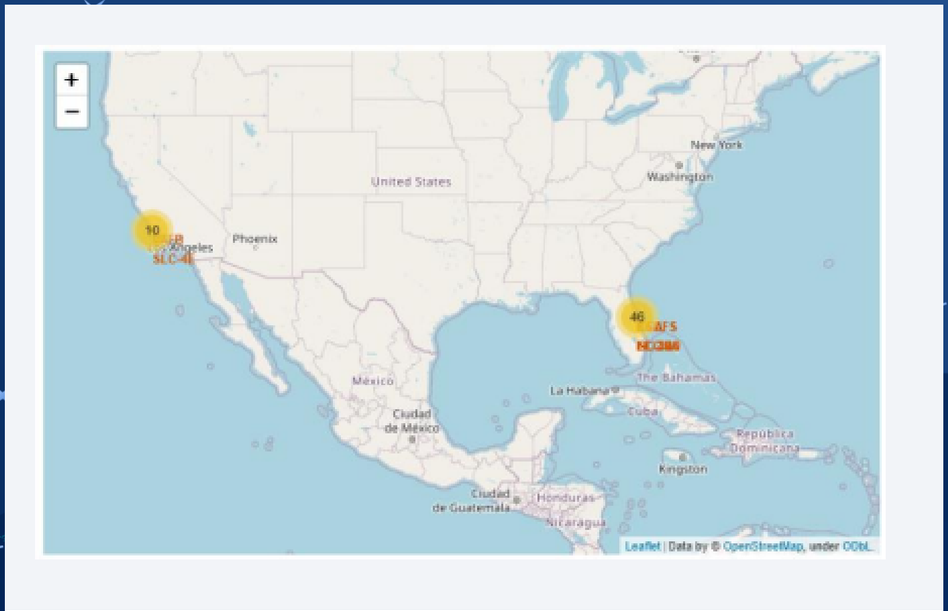
Results

- Exploratory data analysis results:
- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Results

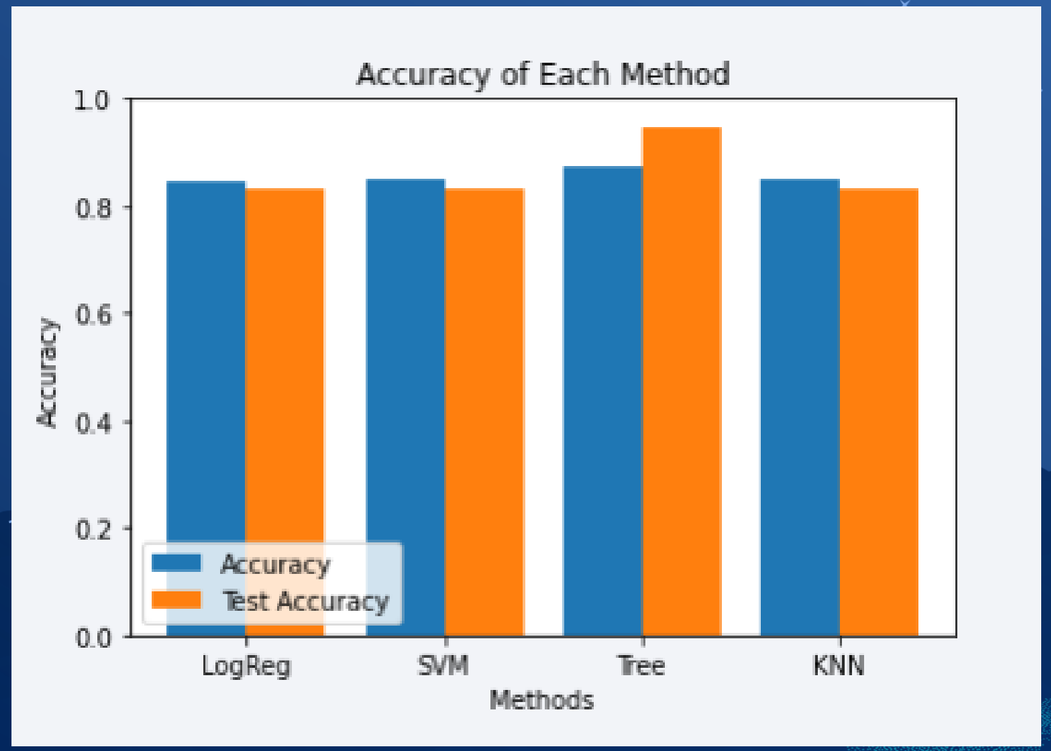
Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.

- Most launches happens at east cost launch sites.



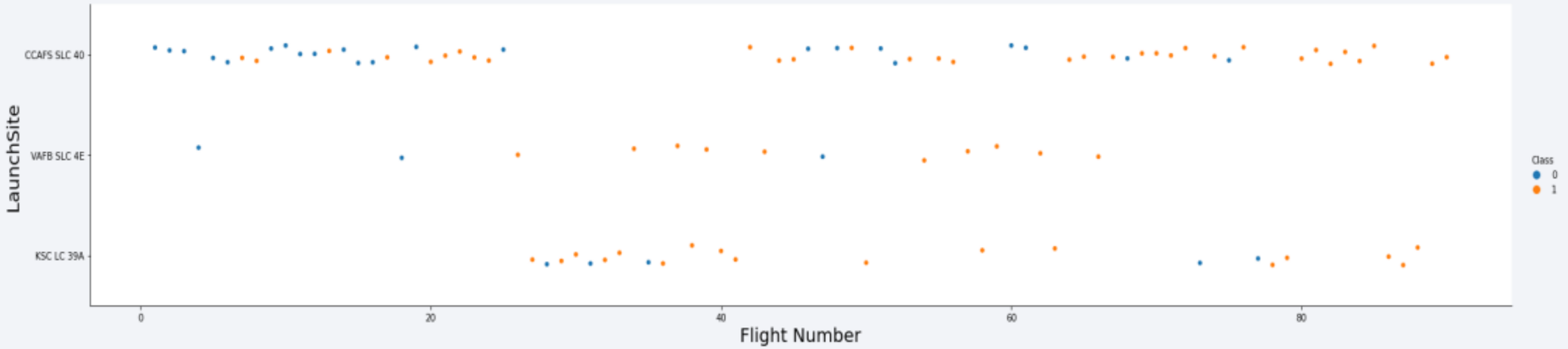
Results

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



Insights

Flight Number vs. Launch Site

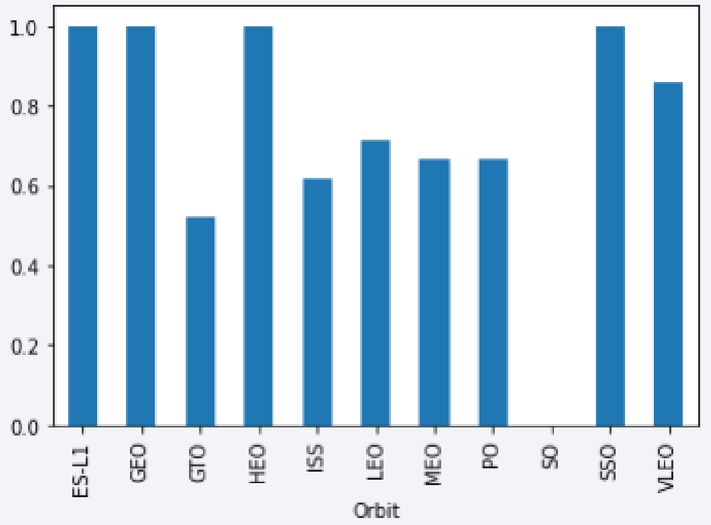


- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

Insights

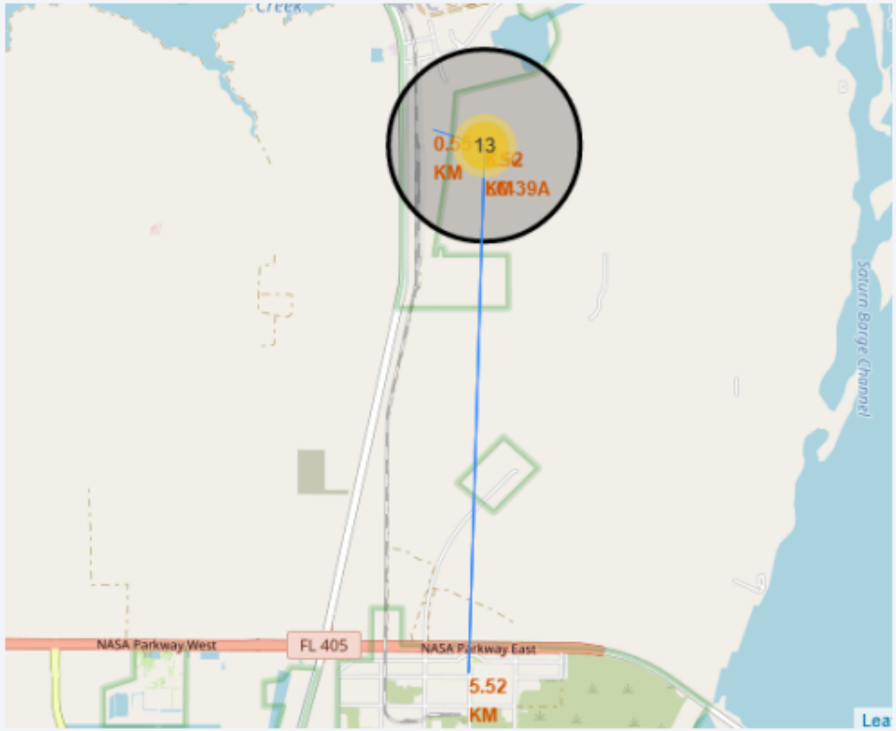
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).



Insights

Logistics and Safety



- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Conclusion

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits. C