SPACE Y: The Next Frontier in the Space Race

Bhuvanavarsh Shivaprasad Hiremath
29-09-2024

Table of Contents

- Executive Summary

- Introduction

- Methodology

- Results

- Conclusion

- Appendix

Executive Summary

- The following methodologies were used to analyze data:

  - Data Collection using web scraping and SpaceX API;

  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics

  - Machine Learning Prediction.

- Summary of all results

  - It was possible to collected valuable data from public sources;

  - EDA allowed to identify which features are the best to predict success of launchings;

  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The aim is to assess the potential of the new company Space Y to rival Space X.

- QUESTIONS:
- 1. The most effective method for calculating the overall expenses associated with launches involves forecasting the successful recovery of the first stage of rockets.
- 2. Identifying the optimal location for conducting launches is essential.

Methodology

- Data collection methodology:

    We have obtained Data from Space X was obtained from 2 sources:

    - Space X API (https://api.spacexdata.com/v4/rockets/)

    - Also, WebScraping
      (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_la
      unches)

- Data wrangling:

    - We collected data was enriched by creating a landing outcome label based on outcome
      data
      after summarizing and analyzing features.

- 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. It was then divided into 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

- Data sets were taken from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping techniques.
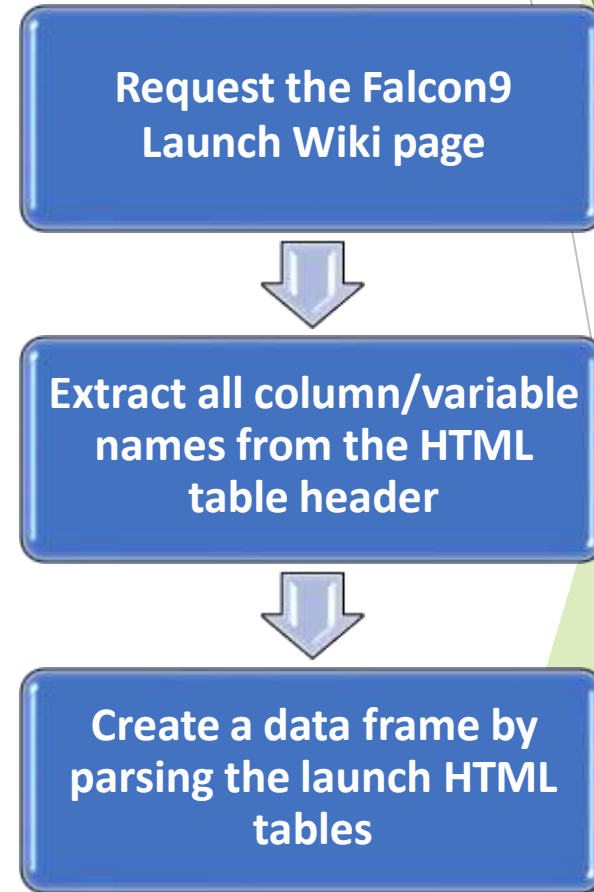
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used.

- This API was used according to the flowchart on the right.

- SOURCE CODE:
- https://github.com/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

```
┌─────────────────────┐
│  Request API and    │
│  parse the SpaceX   │
│    launch data      │
└─────────────────────┘
          ↓
┌─────────────────────┐
│ Filter data to only │
│  include Falcon 9   │
│      launches       │
└─────────────────────┘
          ↓
┌─────────────────────┐
│  Deal with Missing  │
│       Values        │
└─────────────────────┘
```
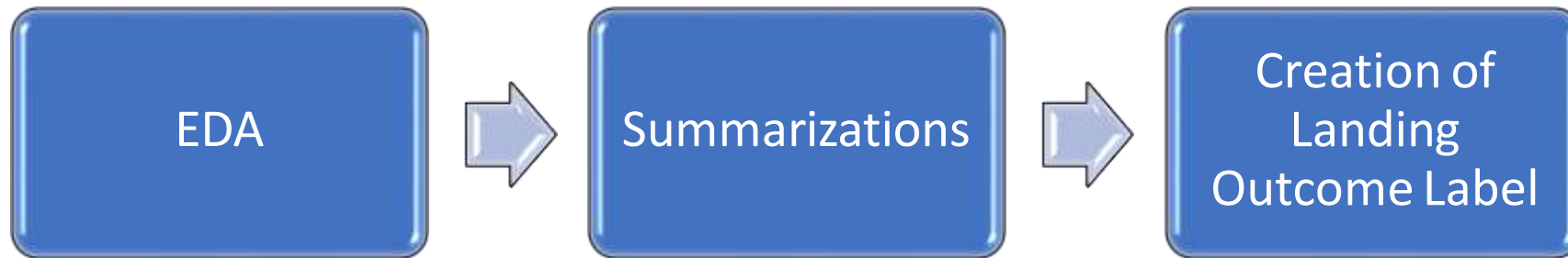
Data Collection - Scraping

- Source code:
https://github.com/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb
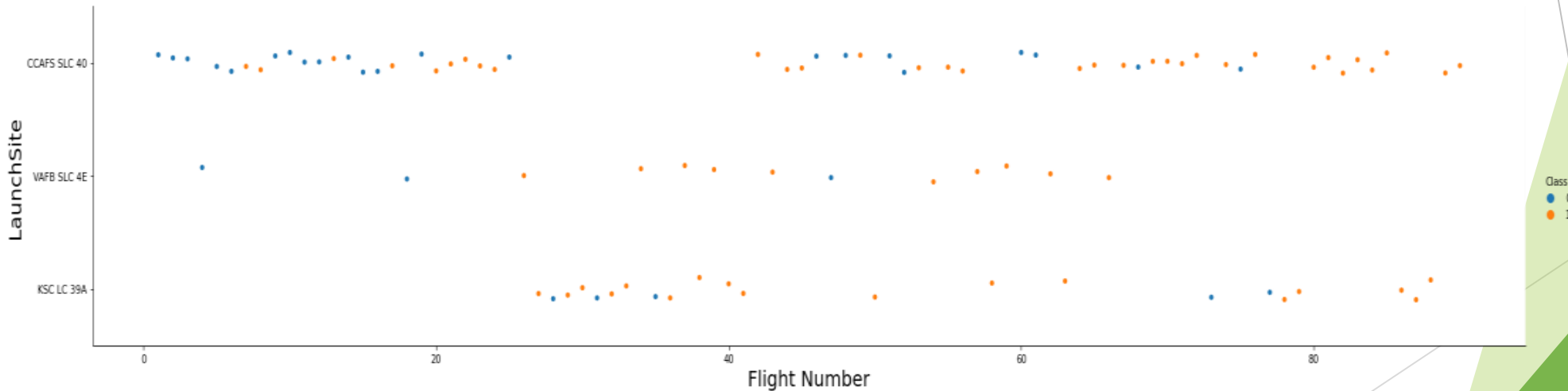
Request the Falcon9 Launch Wiki page

⬇

Extract all column/variable names from the HTML table header

⬇

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially, an Exploratory Data Analysis (EDA) was conducted on the dataset. Subsequently, summaries were generated for each site, along with the frequency of each orbit and the occurrences of mission outcomes categorized by orbit type. Ultimately, the landing outcome label was derived from the Outcome column..

EDA	→	Summarizations	→	Creation of Landing Outcome Label

- Source code: https://github.com/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/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/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

12

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.
- Source code: https://github.com/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb
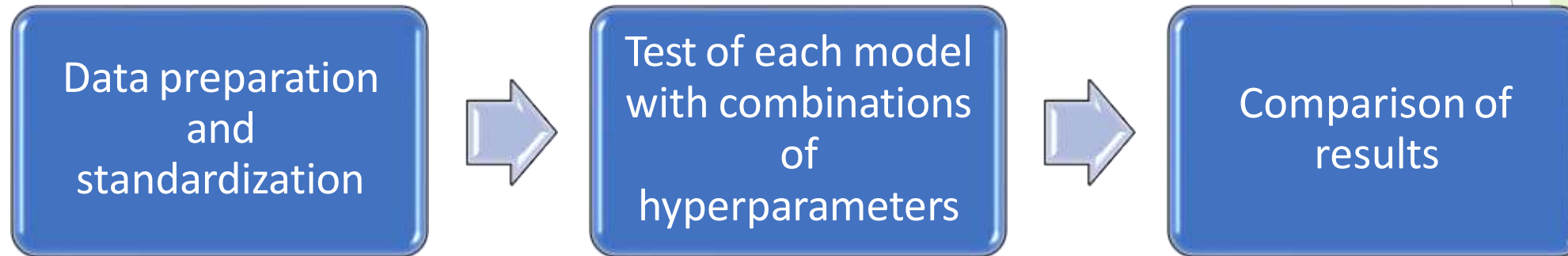
Build an Interactive Map with Folium

- ## 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/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/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/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

14

Predictive Analysis (Classification)

- Four classification models : logistic regression, support vector machine, decision tree, K-nearest neighbors.

Data preparation and standardization ➡ Test of each model with combinations of hyperparameters ➡ Comparison of results

- Source code: https://github.com/bhuvanhiremath471/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis findings indicate the following:
- SpaceX operates four distinct launch sites.
- Initial launches were conducted for SpaceX and NASA.
- The average payload capacity of the Falcon 9 v1.1 booster is 2,928 kg.
- The first successful landing occurred in 2015, five years after the inaugural launch.
- Numerous versions of the Falcon 9 booster achieved successful landings on drone ships, often carrying payloads exceeding the average.
- Nearly all mission outcomes have been successful.
- In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced landing failures on drone ships.
- The success rate of landing outcomes has improved over the years.

Results

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

- 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%.
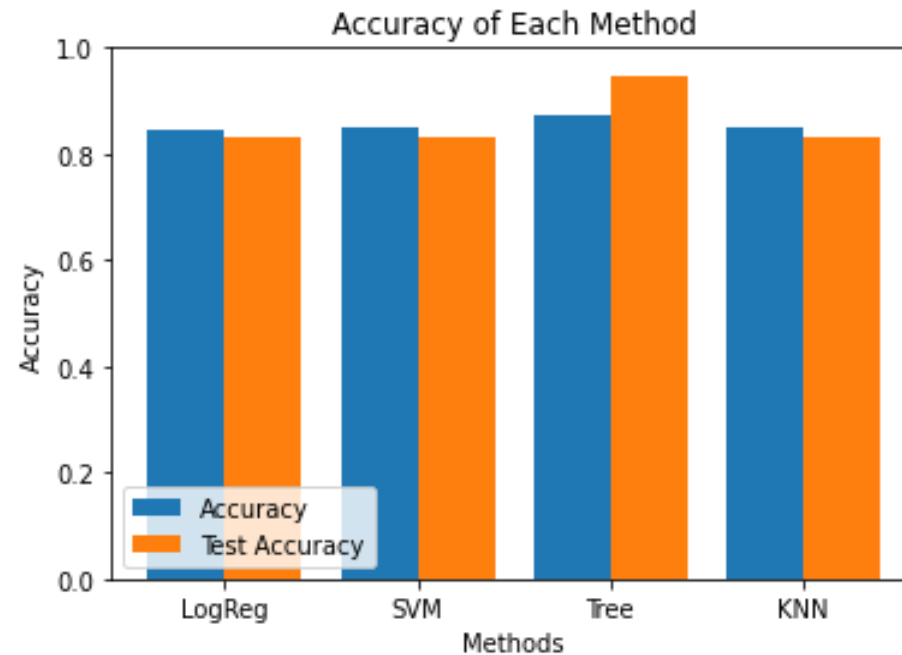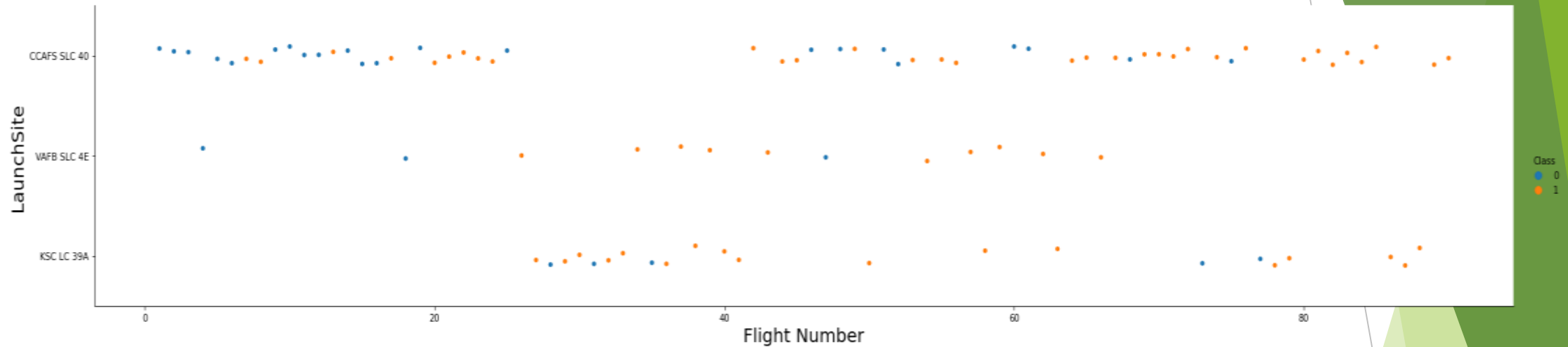
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- According to the plot above, it's possible to verify that the best launch site is CCAF5 SLC 40, where most of recent launches were successful.

- In second place VAFB SLC 4E and then comes KSC LC 39A.

Payload vs. Launch Site
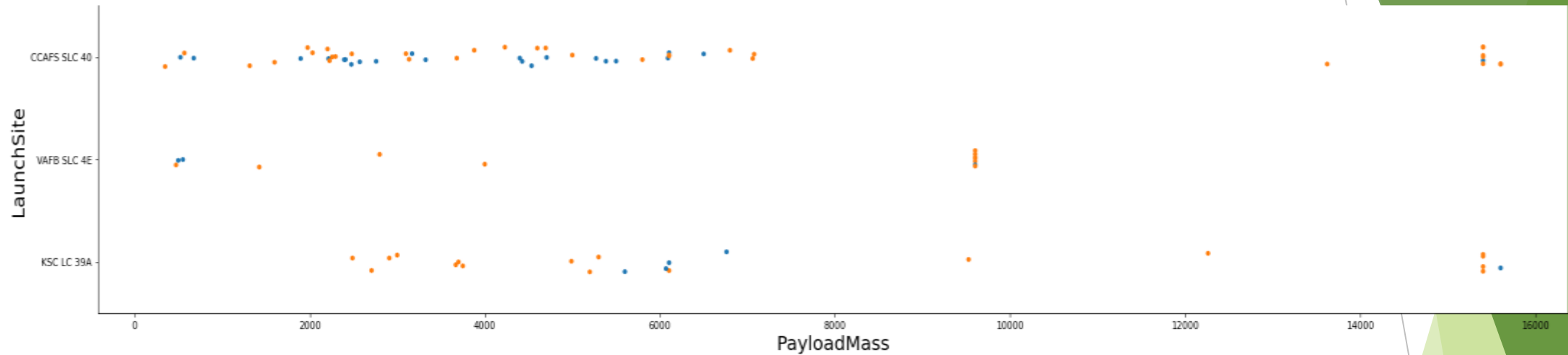
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;

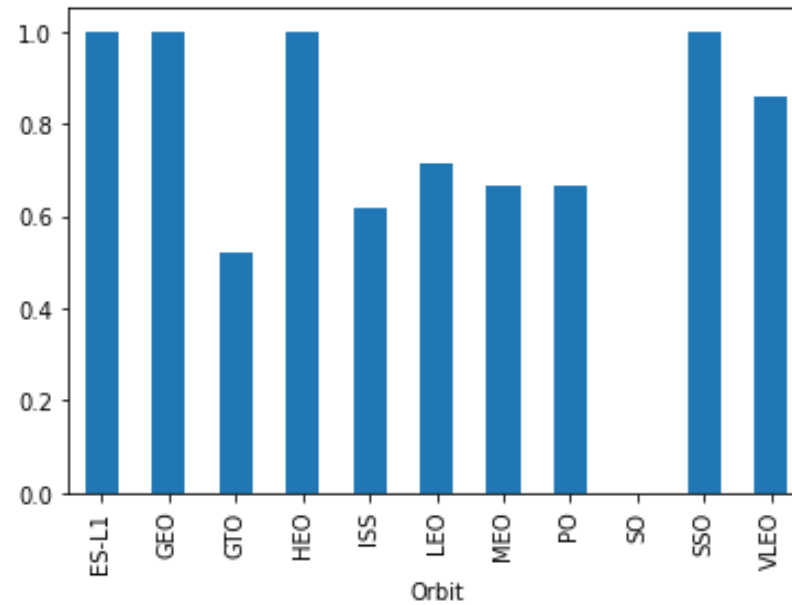- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type
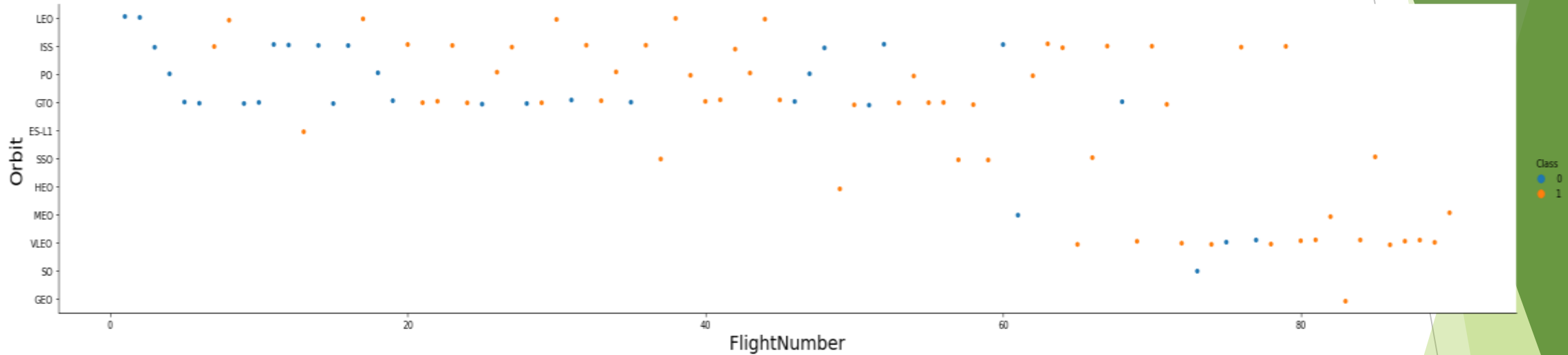
- The biggest success rates happens to orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO.
- Followed by:
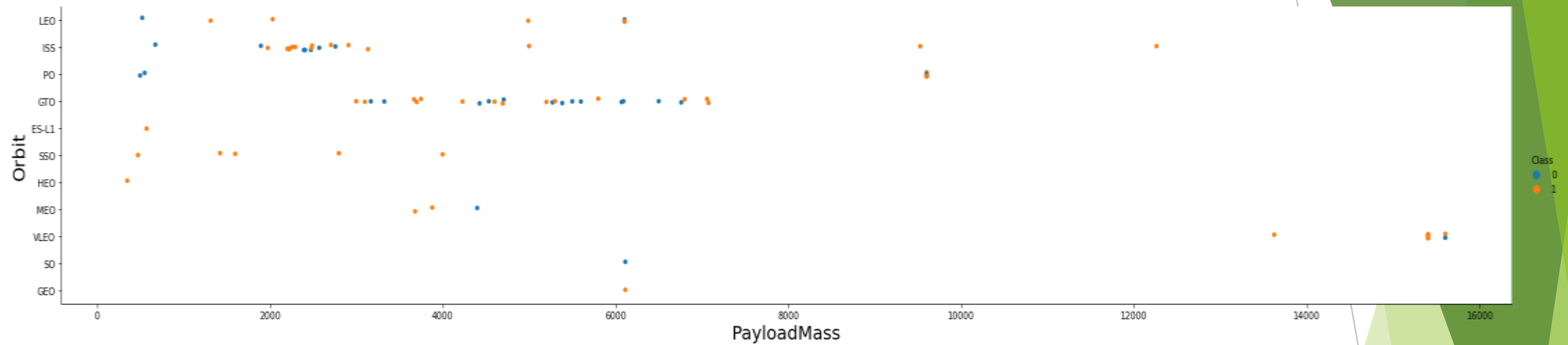 - VLEO (above 80%)
  - LFO (above 70%).

Flight Number vs. Orbit Type

- Apparently, success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

Payload vs. Orbit Type

- Apparently, there is no relation between payload and success rate to orbit GTO.

- ISS orbit has the widest range of payload and a good rate of success.

- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020.

- It seems that the first three years were a period of adjusts and improvement of technology.

All Launch Site Names

- According to data, these are the four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They are obtained by selecting unique occurrences of "launch_site" values from the dataset.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites that started with `CCA`

Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- Total payload calculated above, by summing all payloads whose codes contain 'CRS', which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- This view of data alerts us that "No attempt" must be taken in account.

Launch Sites Proximities Analysis

All launch sites
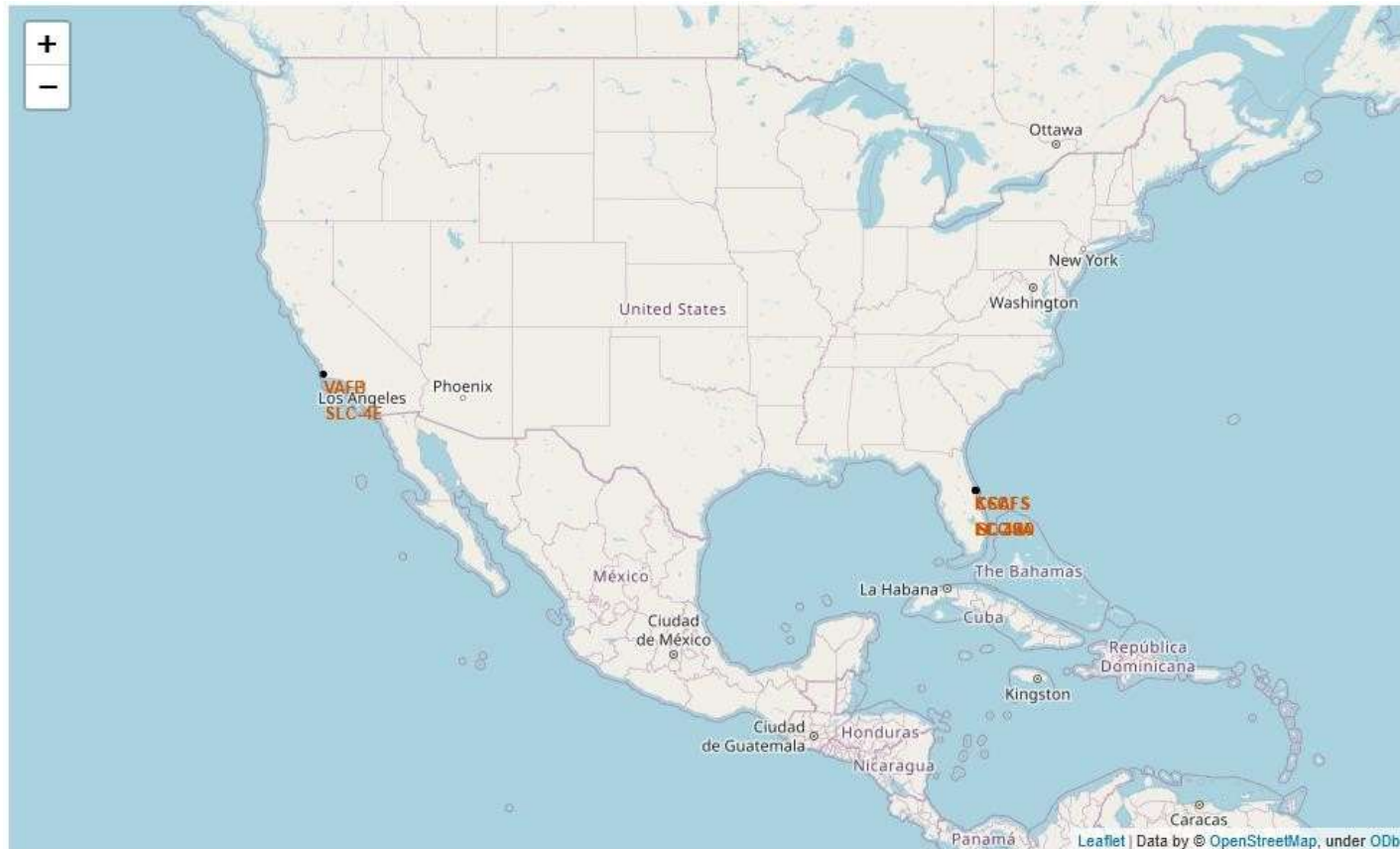
- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes

- Green markers = successful
- Red ones = failure.

Logistics and Safety
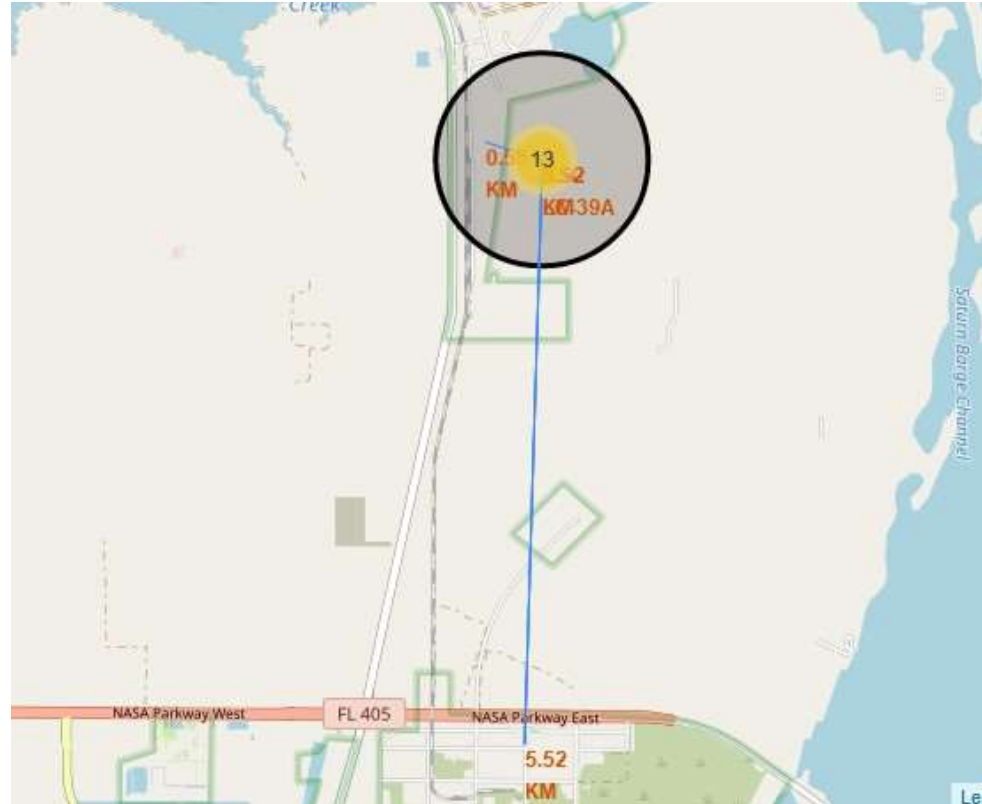
- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.
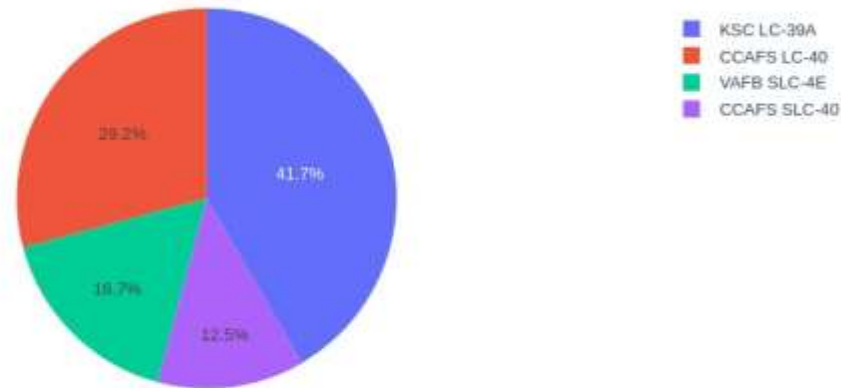
Section 5

Build a Dashboard
with Plotly Dash

Successful Launches by Site

SpaceX Launch Records Dashboard

All Sites   ✕ ▾

Total Success Launches By Site

- KSC LC-39A
- CCAFS LC-40
- VAFB SLC-4E
- CCAFS SLC-40

41.7%

28.2%

16.7%

12.5%

- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A

Total Launches for site KSC LC-39A
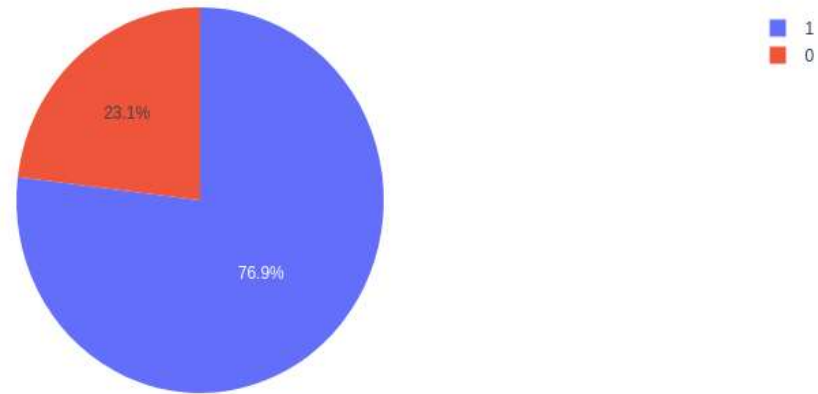
- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome
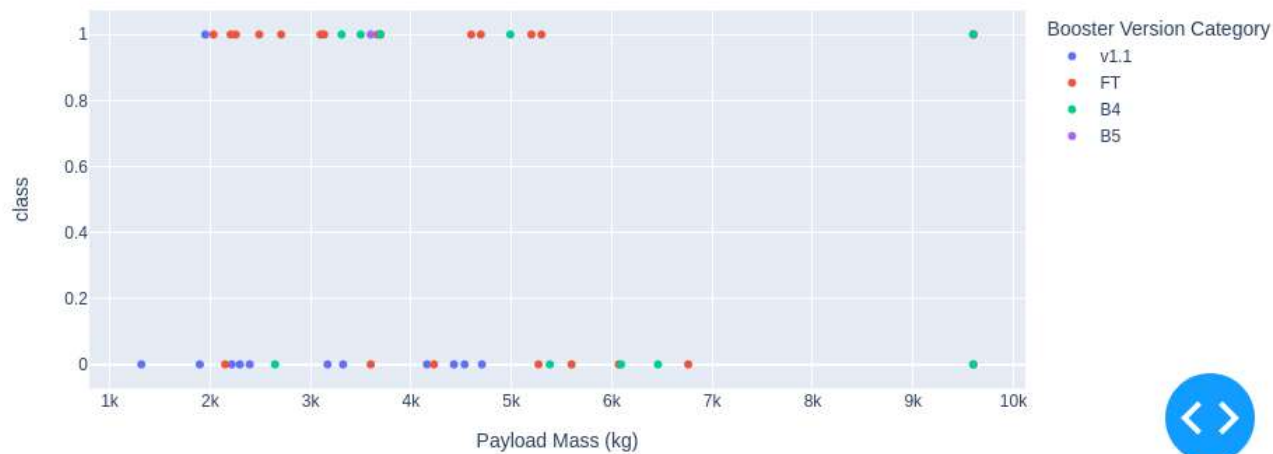
- Payloads under 6,000kg and FT boosters are the most successful combination.

Payload vs. Launch Outcome
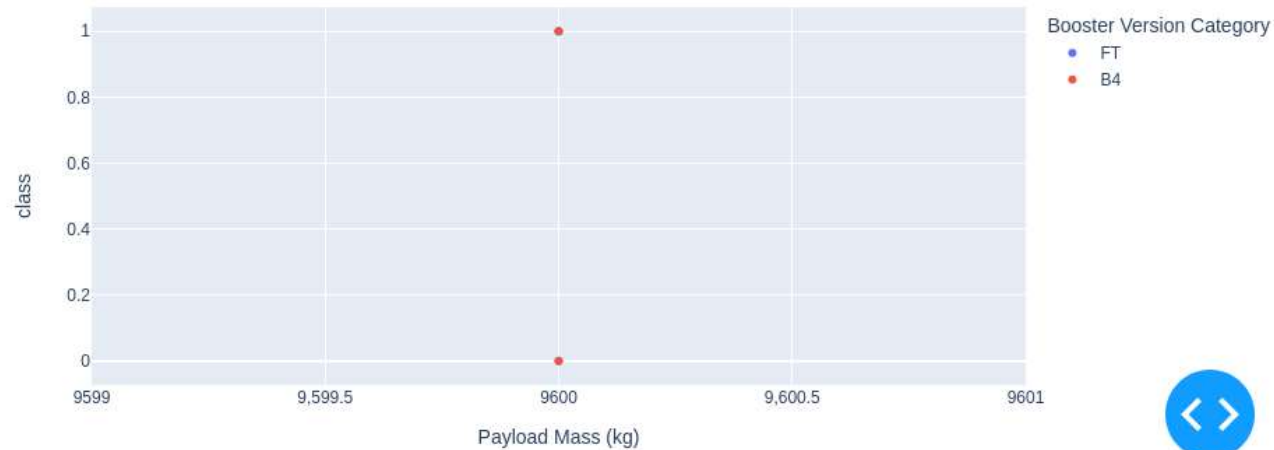
- There's not enough data to estimate risk of launches over 7,000kg
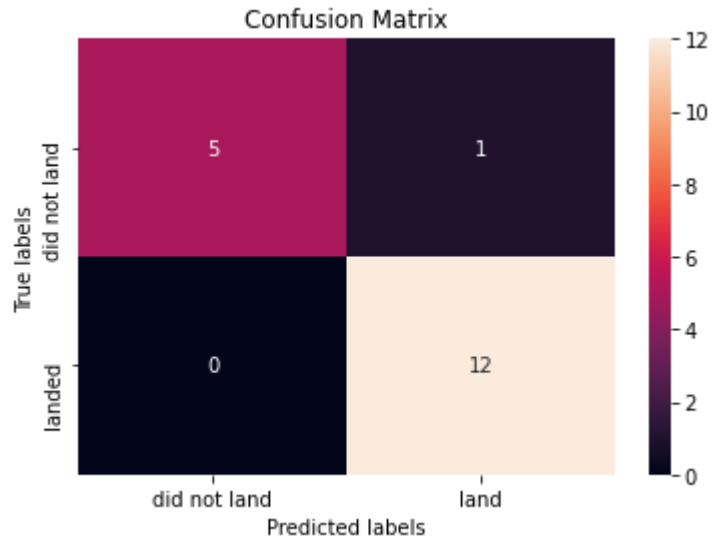
Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside.

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.
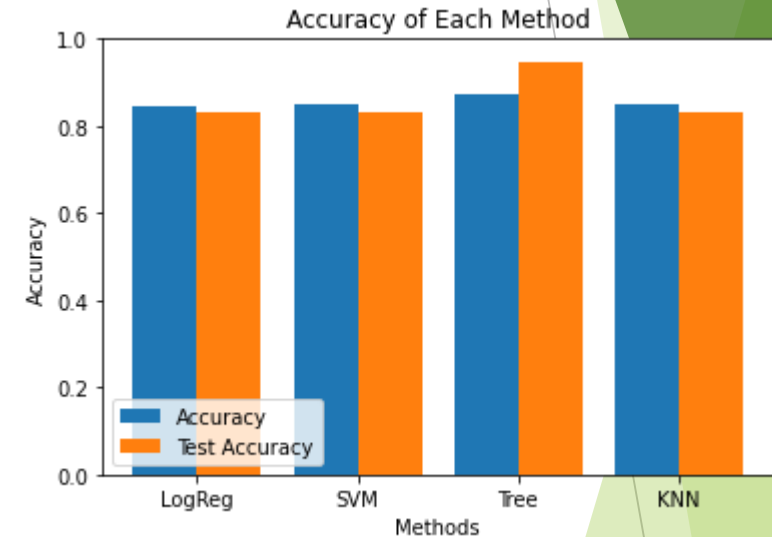
Confusion Matrix of Decision Tree Classifier

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Various data sources were examined, leading to refined conclusions throughout the process.
- The optimal launch site identified is KSC LC-39A.
- Launches exceeding 7,000 kg present a lower risk.
- While the majority of mission outcomes are successful, the rate of successful landings appears to improve over time, reflecting advancements in processes and rocket technology.
- A Decision Tree Classifier can be employed to forecast successful landings and enhance profitability.

Thank you!