IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Caio Azevedo
20/03/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- The analysis was conducted using the following methodologies:
  - Data collection through web scraping and the SpaceX API.
  - Exploratory Data Analysis (EDA), encompassing data cleaning, visualization, and interactive analytics.
  - Machine Learning for predictive modeling.
- Summary of Findings:
  - Valuable data was successfully gathered from public sources.
  - EDA helped identify the most relevant features for predicting launch success.
  - Machine learning models determined the key characteristics that influence successful launches, leveraging all collected data to optimize predictions.

Introduction

- Context:
  - We aim to predict whether the Falcon 9 first stage will land successfully. SpaceX offers Falcon 9 rocket launches at a cost of $62 million, whereas other providers charge upwards of $165 million per launch. A significant portion of SpaceX's cost advantage comes from its ability to reuse the first stage. Therefore, by determining whether the first stage will land successfully, we can estimate the overall launch cost. This information could be valuable for competing companies looking to bid against SpaceX for rocket launches.
- Desired Outcomes:
  - The most effective method for estimating total launch costs by predicting the successful landings of a rocket's first stage.
  - Identifying the optimal location for rocket launches.

Section 1

Methodology

Methodology

Executive Summary

- 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
  - The collected data was enhanced by generating a landing outcome label derived from outcome data after summarizing and analyzing key features.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The collected data was normalized, split into training and test datasets, and assessed using four different classification models. The accuracy of each model was evaluated based on various parameter combinations.

Data Collection

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

Data Collection – SpaceX API

- Where using a public API available by SpaceX where data can be obtained and used for this capstone.

Request API and parse the SpaceX launch data → Filter data to Falcon 9 launches → Treat the missing values

- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/commit/7d4d5222a23088555fd4590cfc9a4c79456e5fdf

Data Collection - Scraping

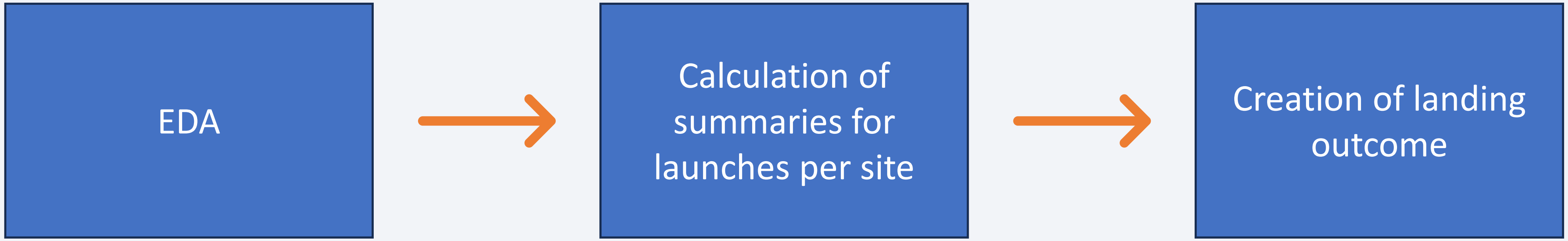
- Data are downloaded from Wikipedia to start our data frame

Request the Falcon9 launch Wikipedia page → Extract all data from the HTML table header → Create a data frame by parsing the launch HTML tables

- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/commit/7d4d5222a23088555fd4590cfc9a4c79456e5fdf

Data Wrangling

- Exploratory Data Analysis (EDA) was first conducted on the dataset, followed by the calculation of summaries for launches per site, the frequency of each orbit type, and mission outcomes by orbit type. Finally, the landing outcome label was created based on the Outcome column.



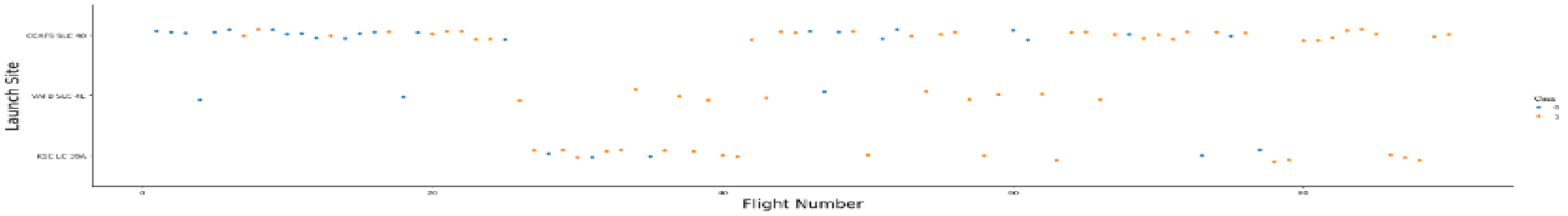
- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/commit/cf5773b69c6939b797ca75dcd2b66cbb6b6446c3

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between the Flight Numbers and Launch Sites.

```
In [8]:

# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch sit
sns.catplot(x='FlightNumber', y='LaunchSite', hue='Class', data=df, aspect=5)
plt.xlabel('Flight Number', fontsize=20)
plt.ylabel('Launch Site', fontsize=20)
plt.show()
```



- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/blob/main/edadataviz.ipynb

EDA with SQL

- The following SQL queries were executed:
  - Retrieved the names of unique launch sites involved in space missions.
  - Identified the top five launch sites with names starting with 'CCA'.
  - Calculated the total payload mass carried by boosters launched under NASA (CRS) missions.
  - Determined the average payload mass carried by the F9 v1.1 booster version.
  - Found the date of the first successful landing on a ground pad.
  - Listed the names of boosters that successfully landed on a drone ship and carried payloads between 4,000 and 6,000 kg.
  - Counted the total number of successful and failed mission outcomes.
  - Identified booster versions that carried the maximum payload mass.
  - Retrieved failed landing outcomes on drone ships, including booster versions and launch site names, for the year 2015.
  - Ranked landing outcomes, such as "Failure (drone ship)" and "Success (ground pad)," for launches between June 4, 2010, and March 20, 2017.
  - GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium Maps were utilized with various visual elements, including markers, circles, lines, and marker clusters.
- Markers represented locations such as launch sites, while circles highlighted specific areas around designated coordinates, such as the NASA Johnson Space Center. Marker clusters grouped multiple events at the same location, such as launches from a particular site. Lines were used to depict distances between two coordinates.
- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Various graphs and plots were used to visualize the data, including the percentage of launches by site and the payload range.
- This combination enabled a quick analysis of the relationship between payloads and launch sites, helping to identify the optimal launch locations based on payload capacity.
- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/blob/main/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Were used the following classification models: Logistic regression, support vector machine, decision tree and k nearest neighbors.

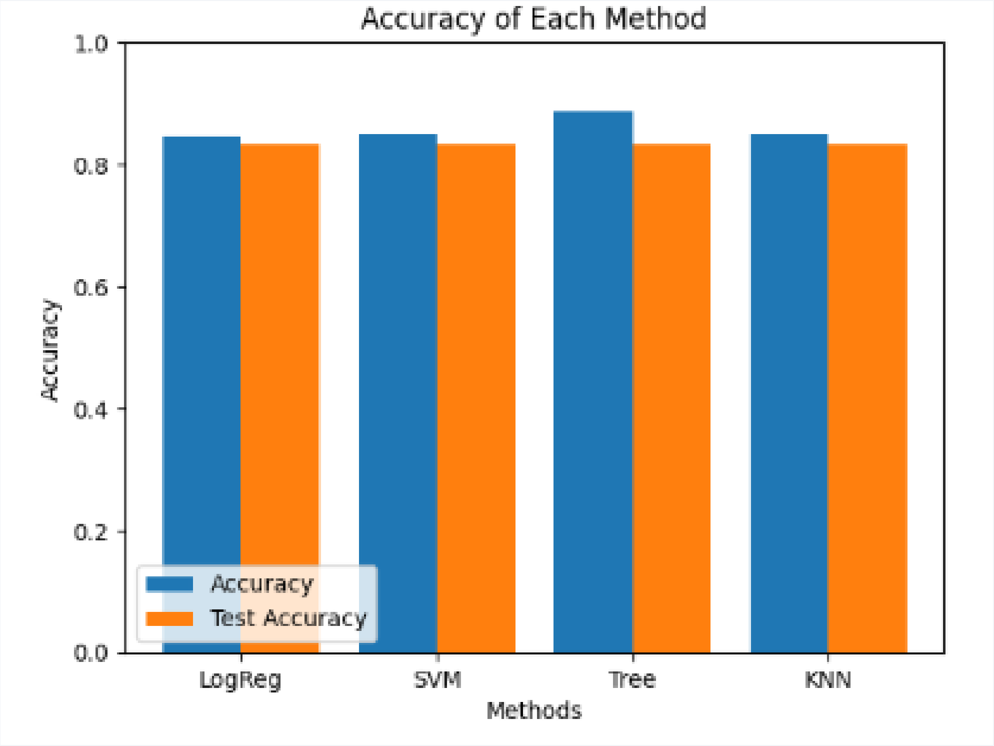
Data preparation → Test of each model → Comparison of results

- GitHub URL: https://github.com/O-Caio-Cesar/IBM_DS_CAPSTONE/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Here are the results from the exploratory data analysis:
  - SpaceX operates four distinct launch sites.
  - The initial launches were carried out for SpaceX and NASA.
  - The average payload of the F9 v1.1 booster is 2,928 kg.
  - The first successful landing occurred in 2015, five years after the first launch.
  - Several Falcon 9 booster versions successfully landed on drone ships with payloads above the average.
  - Nearly all mission outcomes were successful.
  - Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015.
  - The success rate of landing outcomes improved over time.

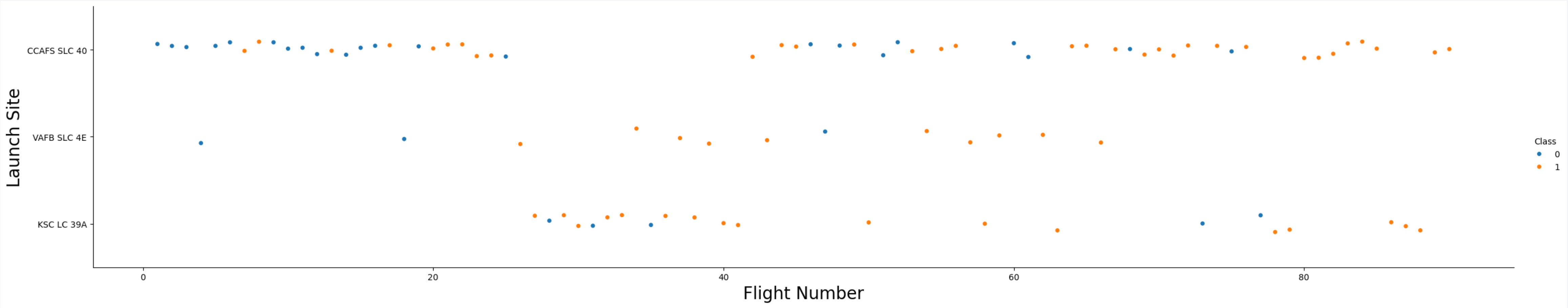
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%..



Section 2

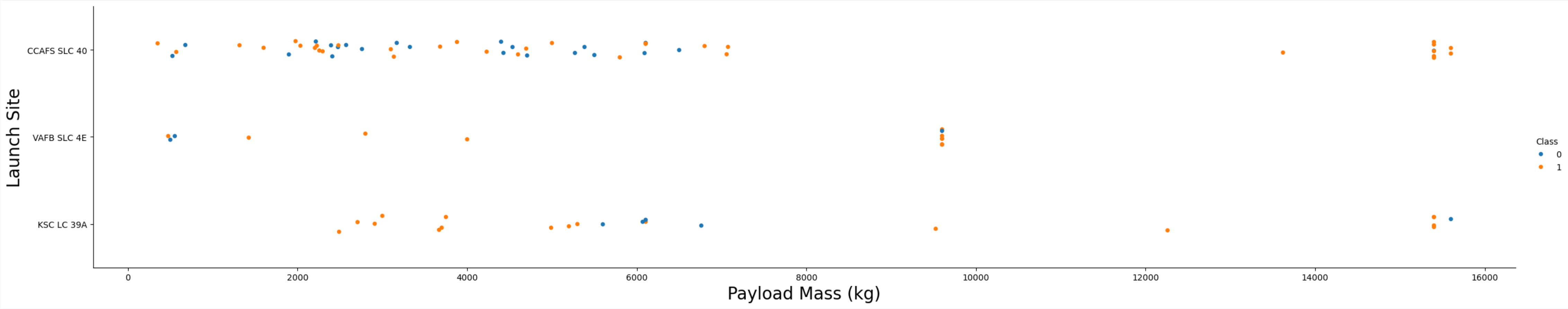
Insights drawn from EDA

Flight Number vs. Launch Site



- The plot above shows that the top launch site currently is CCAF5 SLC 40, where the majority of recent launches have been successful, followed by VAFB SLC 4E in second place and KSC LC 39A in third. Additionally, it's evident that the overall success rate has improved over time.

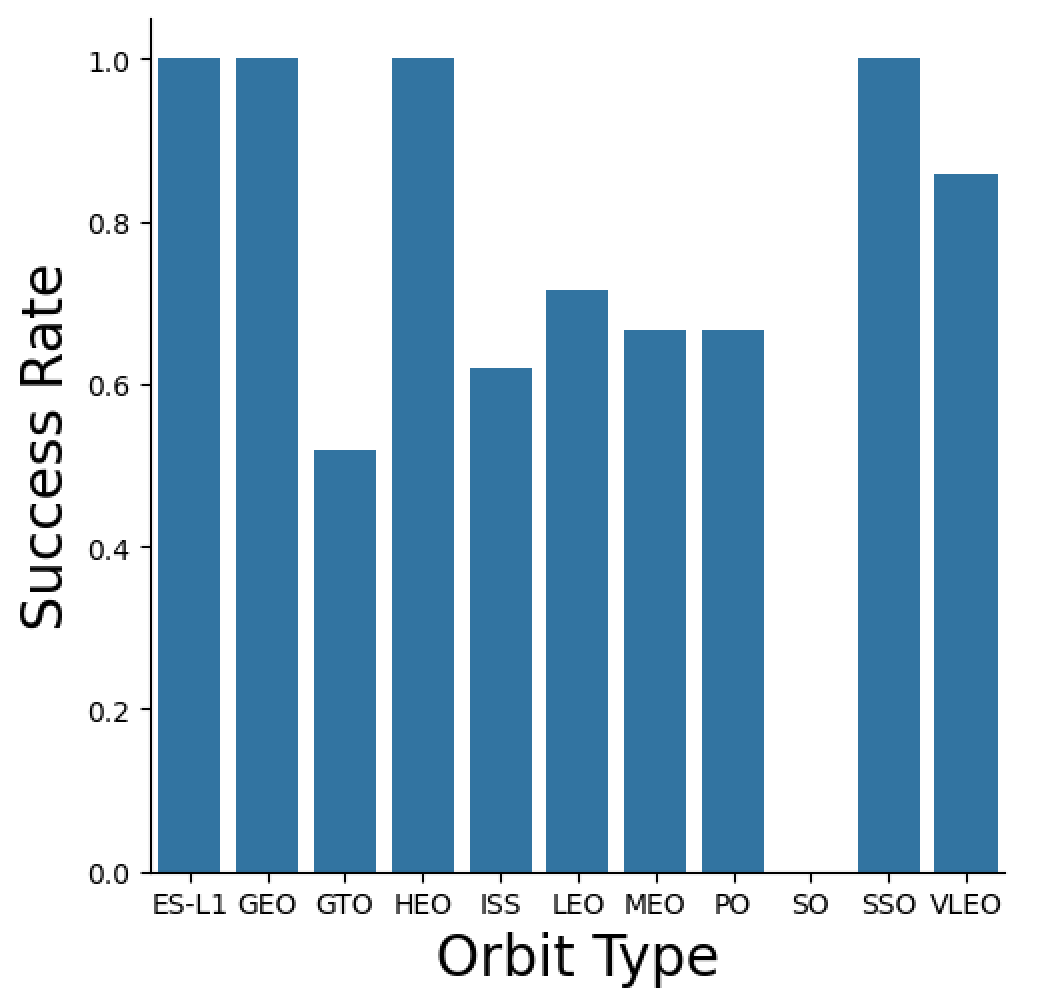
Payload vs. Launch Site



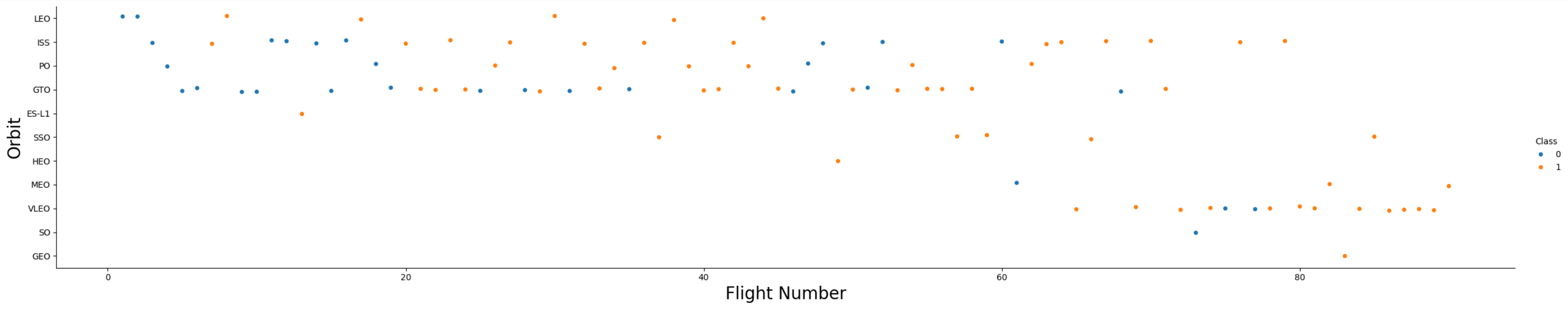
- Payloads weighing more than 9,000 kg (approximately the weight of a school bus) show an excellent success rate. Payloads exceeding 12,000 kg appear to be feasible only at the CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
- ES-L1;
- GEO;
- HEO.

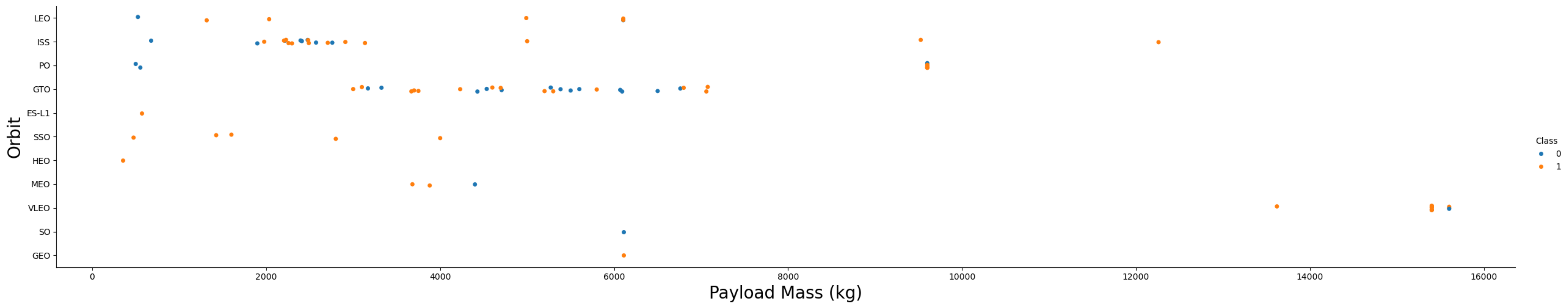


Flight Number vs. Orbit Type



- It seems that the success rate has improved over time for all orbits. The VLEO orbit appears to be a new business opportunity, as its frequency has recently increased.

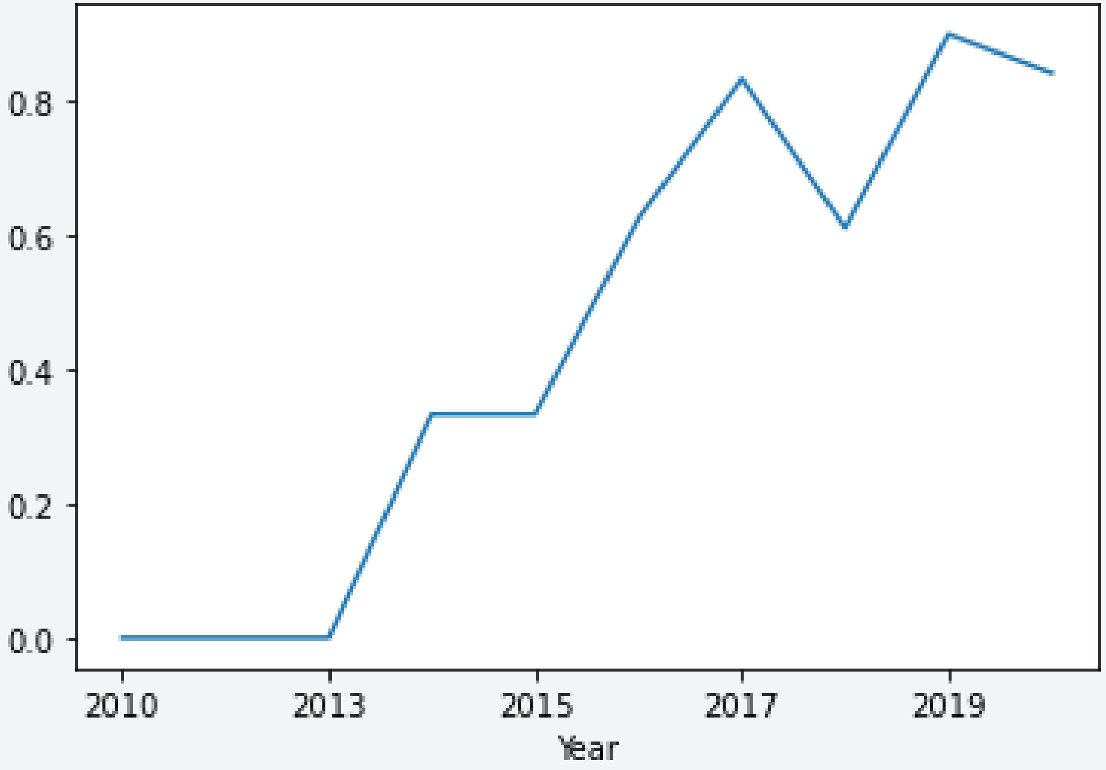
Payload vs. Orbit Type



- It appears that there is no correlation between payload and success rate for the GTO orbit. The ISS orbit has the broadest payload range and a strong success rate. Additionally, there have been few launches to the SO and GEO orbits.

Launch Success Yearly Trend

- The success rate began to rise in 2013 and continued to improve until 2020. It appears that the first three years were a period of adjustments and technological advancements.



All Launch Site Names

- There is four launch sites registered in this data:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin 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	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- First successful landing outcome on ground pad

first_successful_landing
2015-12-22

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

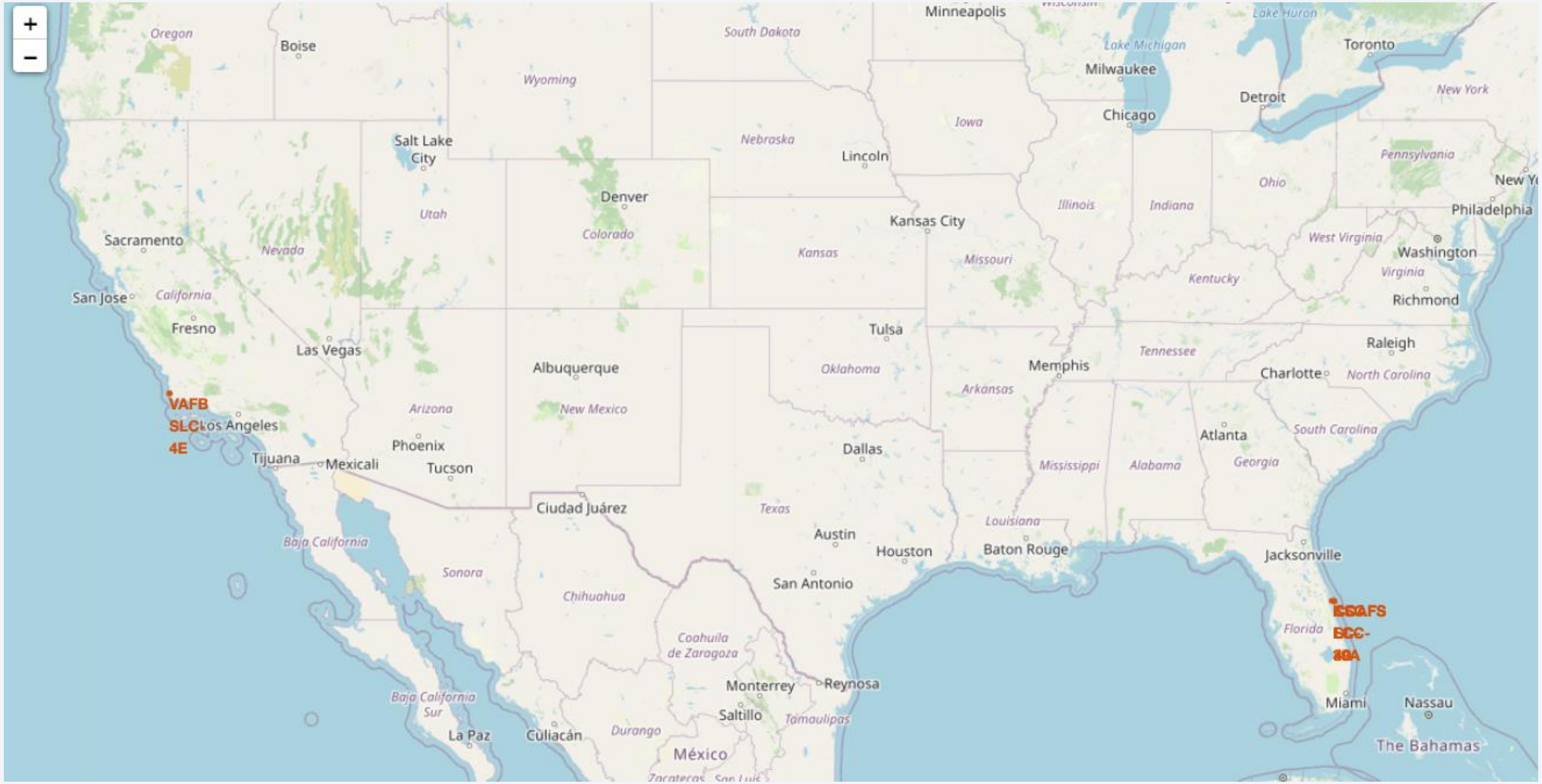
- Rank of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

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

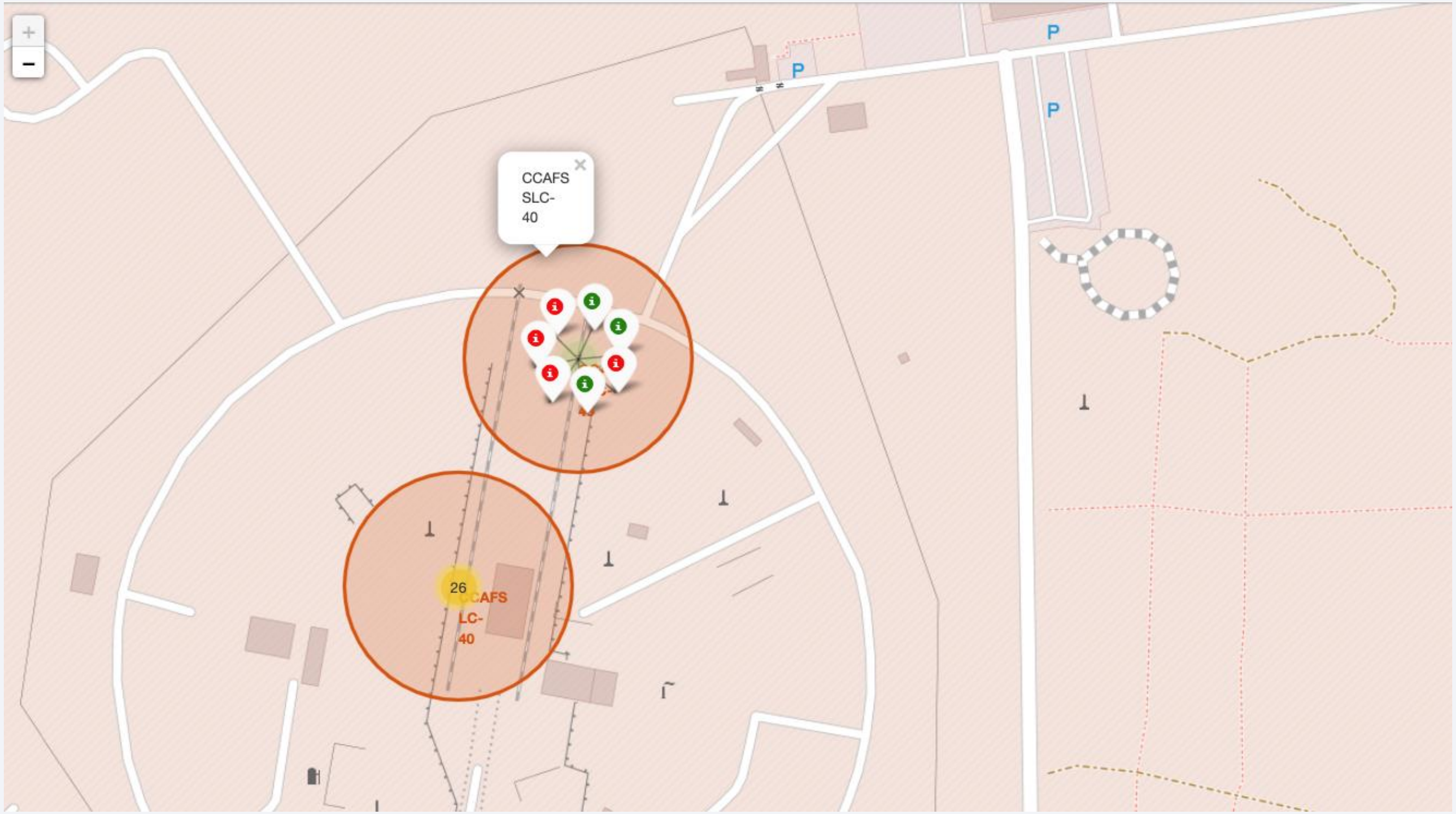
Section 3

Launch Sites Proximities Analysis

Launch Sites



Launch Outcomes

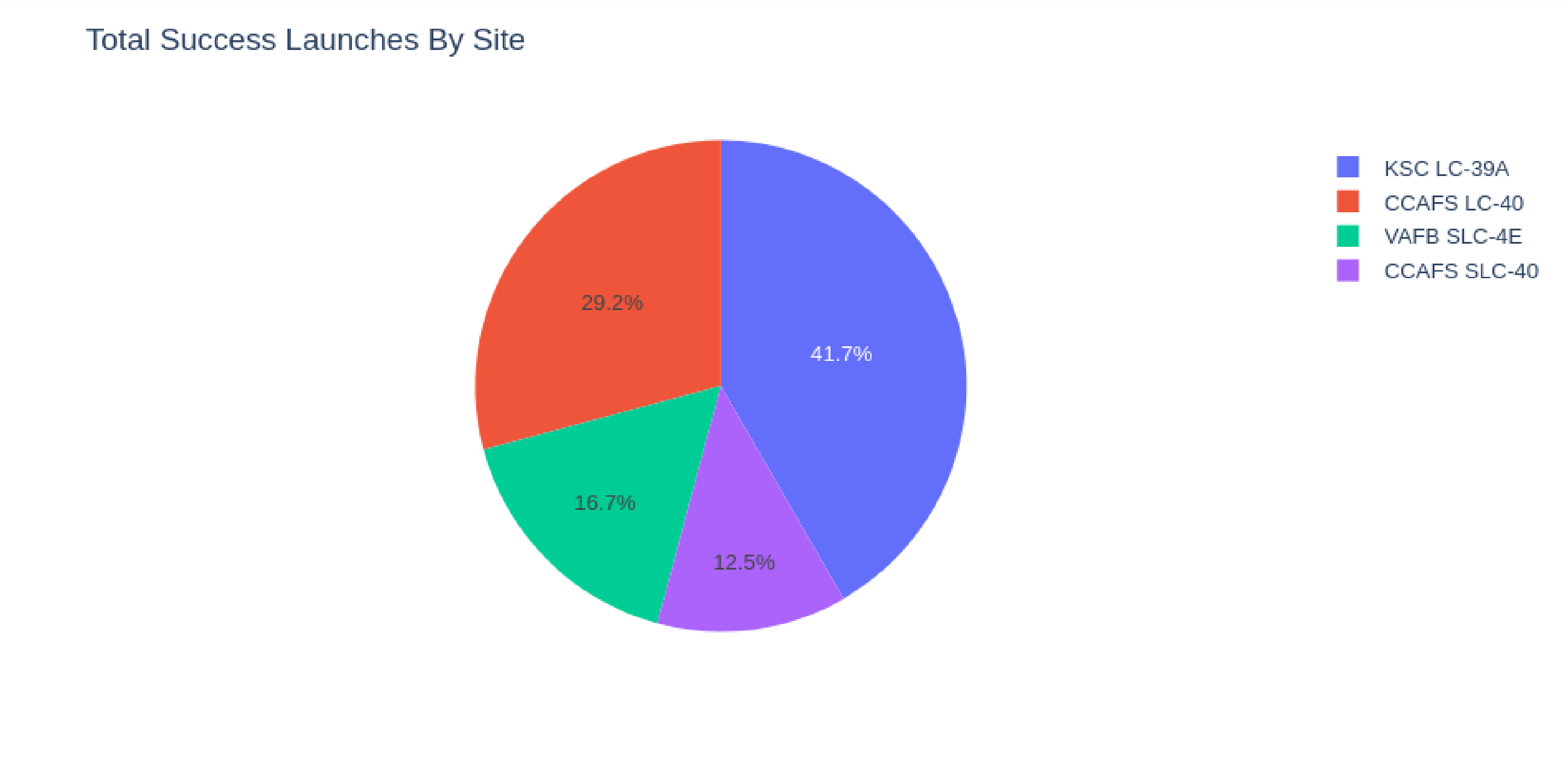


Distance Line

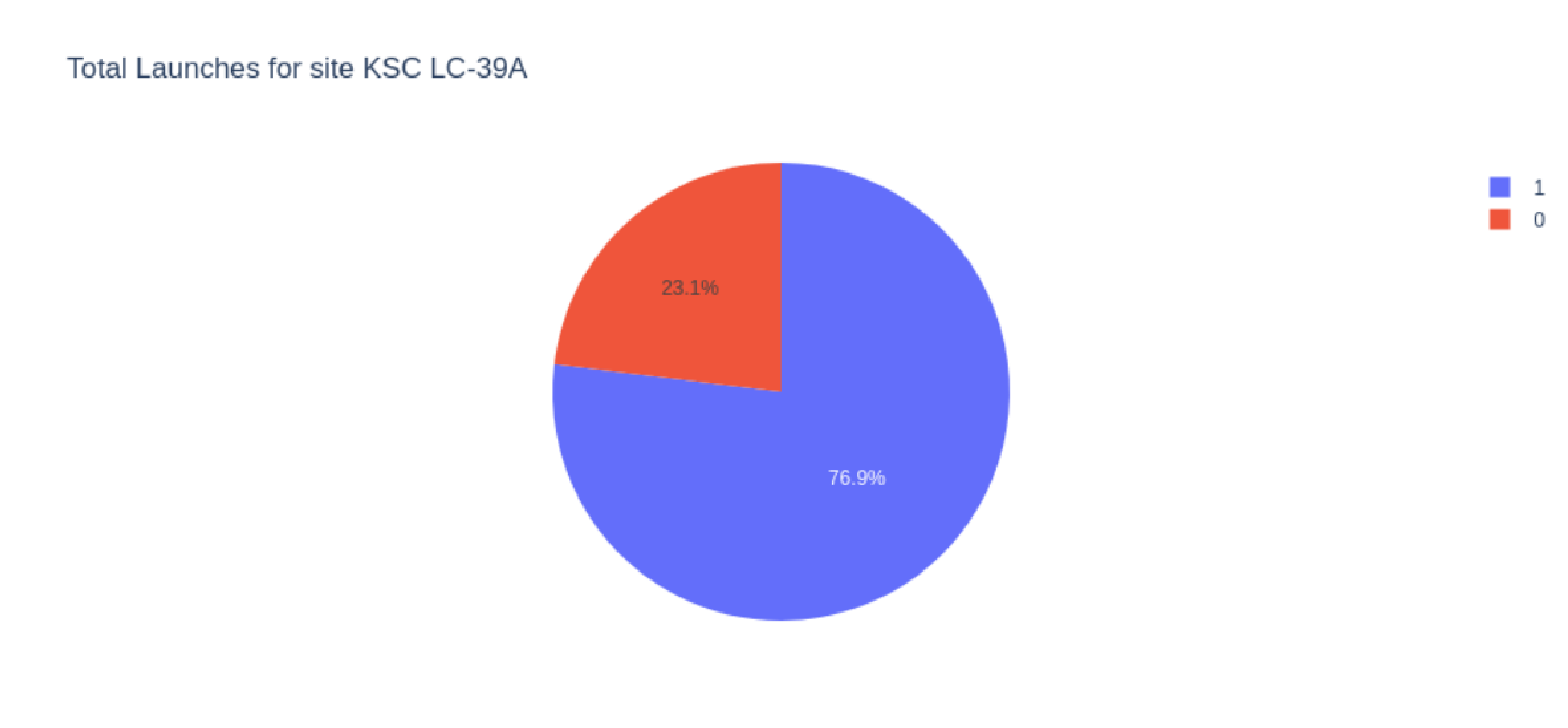
Section 4

Build a Dashboard with Plotly Dash

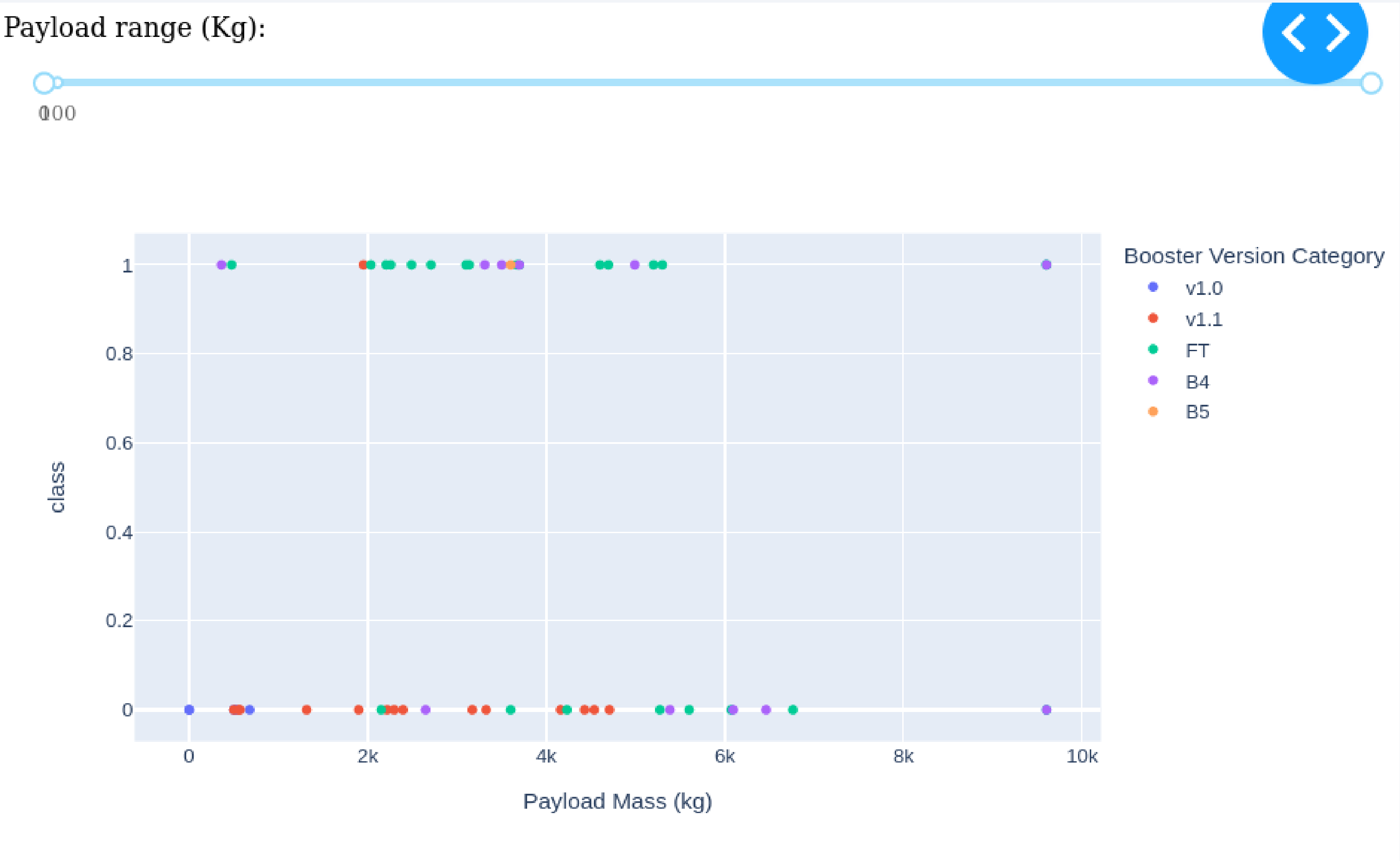
Launch Records



Launches for Site KSC LC-38A



Payload vs. Launch Outcome

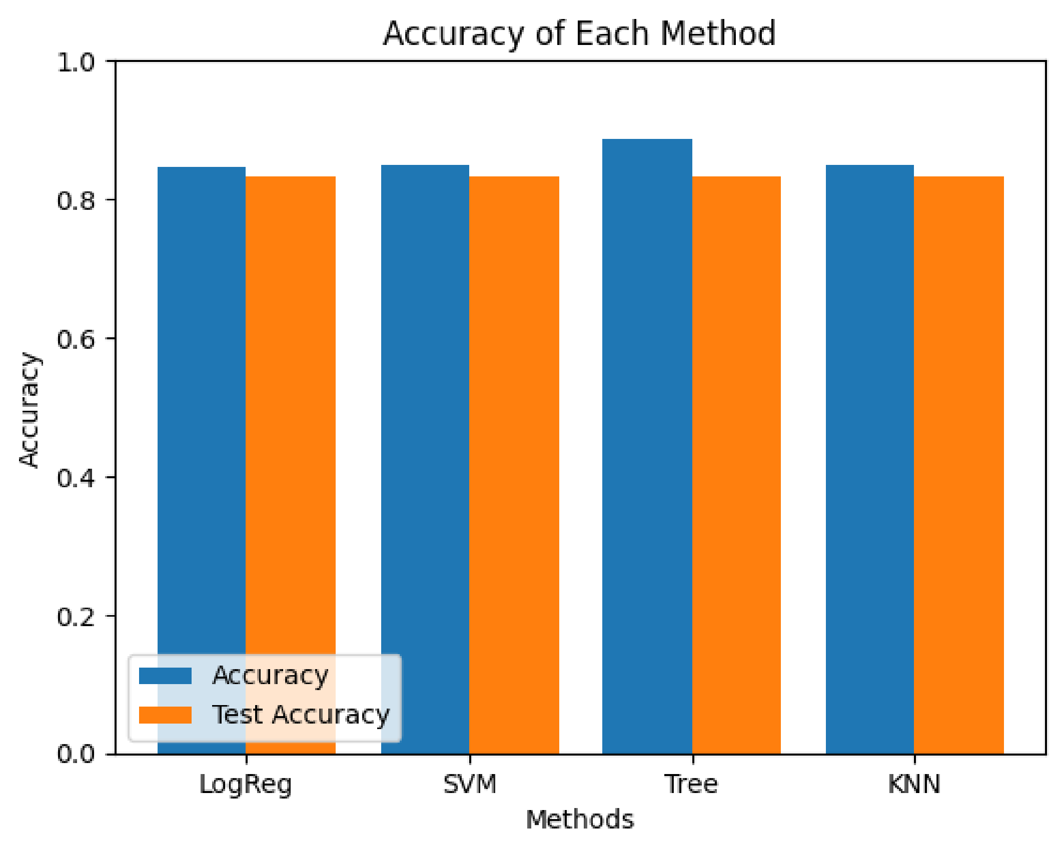


Section 5

Predictive Analysis (Classification)

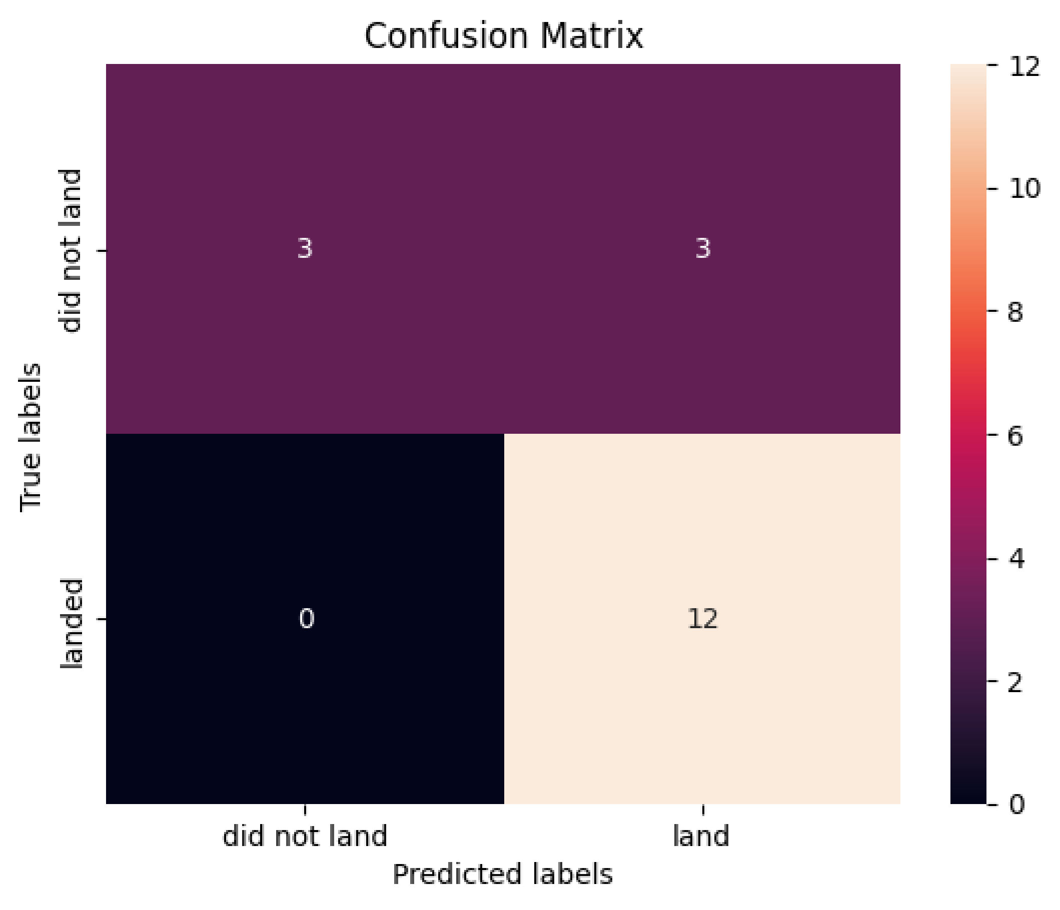
Classification Accuracy

- The highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87% but very near with the other models.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier



Conclusions

- This specialization-ending activity was an excellent and highly productive analysis of mission success factors. The success of a space mission depends on multiple elements, including the launch site, orbit, and notably, the experience gained from previous launches. The learning curve between launches often contributes to turning failures into successes. The experience of treating and getting results with the proposed data were great and very challenging.

Thank you!