Applied Data Science Capstone Project

By

Ryan Squires

IBM Data Science Certification Program

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

This project integrates a variety of Data Science techniques and tools to better understand and predict outcomes for rocket launches by SpaceX.

Data was collected, refined, standardized, explored, and visualized.

By training four machine learning models a prediction can be made as to whether SpaceX will have a successful mission outcome, and if they will be able to reuse the first stage of Falcon 9, which enables them to save roughly 60 million dollars.

The most successful machine learning models produced 83% accuracy.

Introduction

Reusing the first stage of their Falcon 9 rocket is SpaceX's competitive advantage. When Space X can successfully reuse the first stage, it allows them a savings of roughly 60 million dollars.

Space Y in a new company entering the market. Space Y wants to learn more about the cost of rocket launches at Space X and be able to predict whether mission will be successful and if the first stage can be reused.

By training a machine learning model I will analyze the data and determine whether Space X can successfully reuse their first stage.

The outcome of this project is to determine how successful mission outcomes are affected by the variables; number of flights, launch site location, payload mass, number and type of orbits. The results of analysis will determine the optimal machine learning algorithm.

Methodology

- Data Collection through SpaceX Rest API and public Wiki page
- Data Wrangling by classifying launches data
- Explanatory data analysis using SQL and visualization
- Interactive visal analytics using Plotly dash and Folium maps
- Predictive analysis using classification models

Data Collection

Data is collected through the SpaceX Rest API using a GET command to parse through the json file.

Data is collected through a public source using the BeautifulSoap web scraping tool to pull from html tables.

Missing data is replaced with mean values and the dataframe was filtered to only include Falcon 9 launces.

Data Wrangling

Functions were used to

- calculate the number of launches at each site,
- occurrences of each orbit,
- mission outcomes per orbit type

A new outcome label is created for data sets

Explanatory Data Analysis Using SQL

After creating a database in Db2, the SpaceX dataset was loaded into a table. Python was used to perform explanatory data analysis through SQL queries and SQL magic commands.

SQL queries and commands yield

-success/failed mission outcomes

-landing outcome in drone ships

-payload mass

Explanatory Data Analysis with Visualization

Drew correlations between parameters by illustrating relationships. Identified the annual trend of successful SpaceX launches and illustrated the results.

-Clustering and relationships between payload, flight number, and launch site are illustrated in a scatterplot

-Success rates by each orbit type are illustrated in a bar graph.

-Success rates over extended time is illustrated in a line chart.

Interactive Visual Analytics and Dashboard

By building a Plotly Dashboard, a variety of data and information regarding SpaceX launches is aggregated and organized into a thorough visual representation. The dashboard includes drop-down components, pie charts, and scatter plots to illustrate the success/failed launches, distances between sites and proximities site. This dashboard serves as an effective means of communicating detailed technical information to the team at Space Y.

Interactive Visual Analytics and Dashboard

An interactive Follium map is used to highlight key attributes of site locations and the launches at each site.

-each site is highlighted with the circle object.

-site distance proximities are highlighted with the line object

-the success/failure of launches at each site is highlighted with red/green colors.

Predictive Analysis

Models are built by Standardizing data and splitting the dataset into training and test sets makes for machine learning.

Models are evaluated by testing for accuracy and by referencing the confusion matix

Models are evaluated by testing for the optimal hyperparameters.

Models are improved by refining the machine learning algorithm.

p

The best model is selected by meeting optimal hyperarameters.

Results

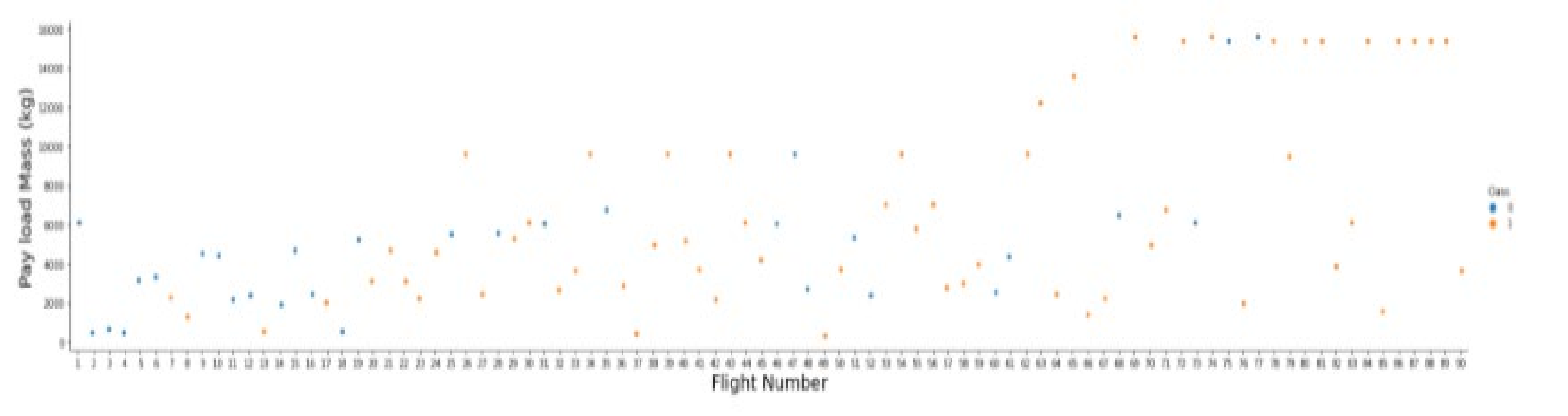
Exporatory Data Analysis with Visualization and SQL

Interactive Visual Analytics with Folium and Poltly

Prediction Analysis with Machine Learning

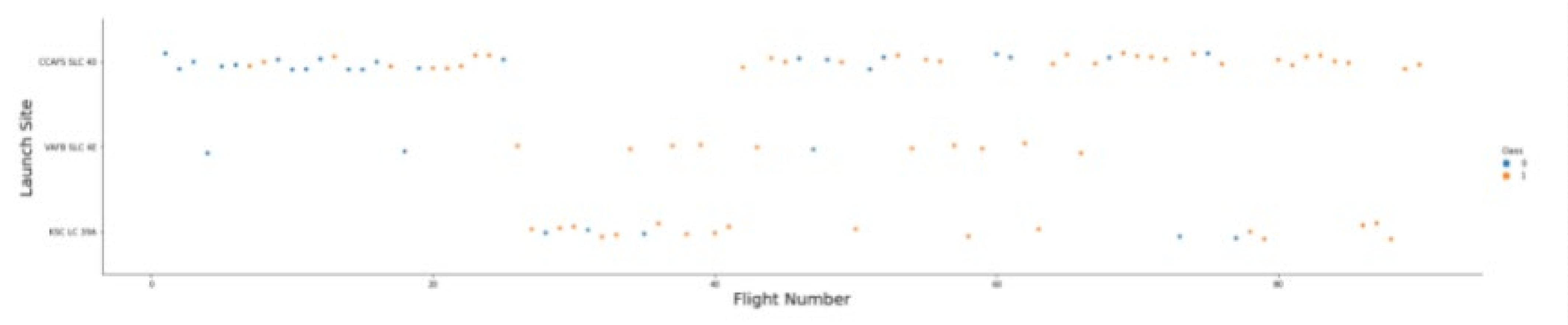
EDA with Visualization

Flight Number vs. Payload Mass



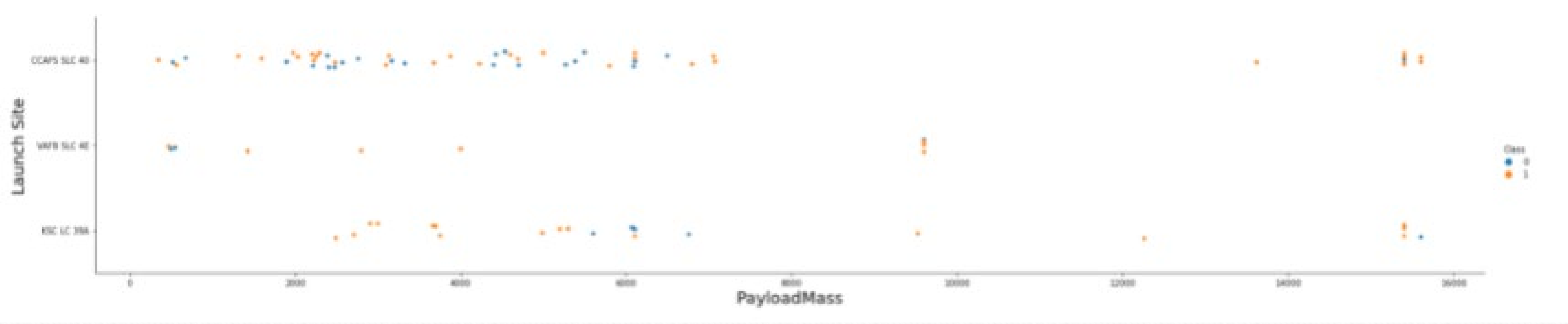
Missions with higher payloads are increasingly successful over time

Flight Number vs. Launch site



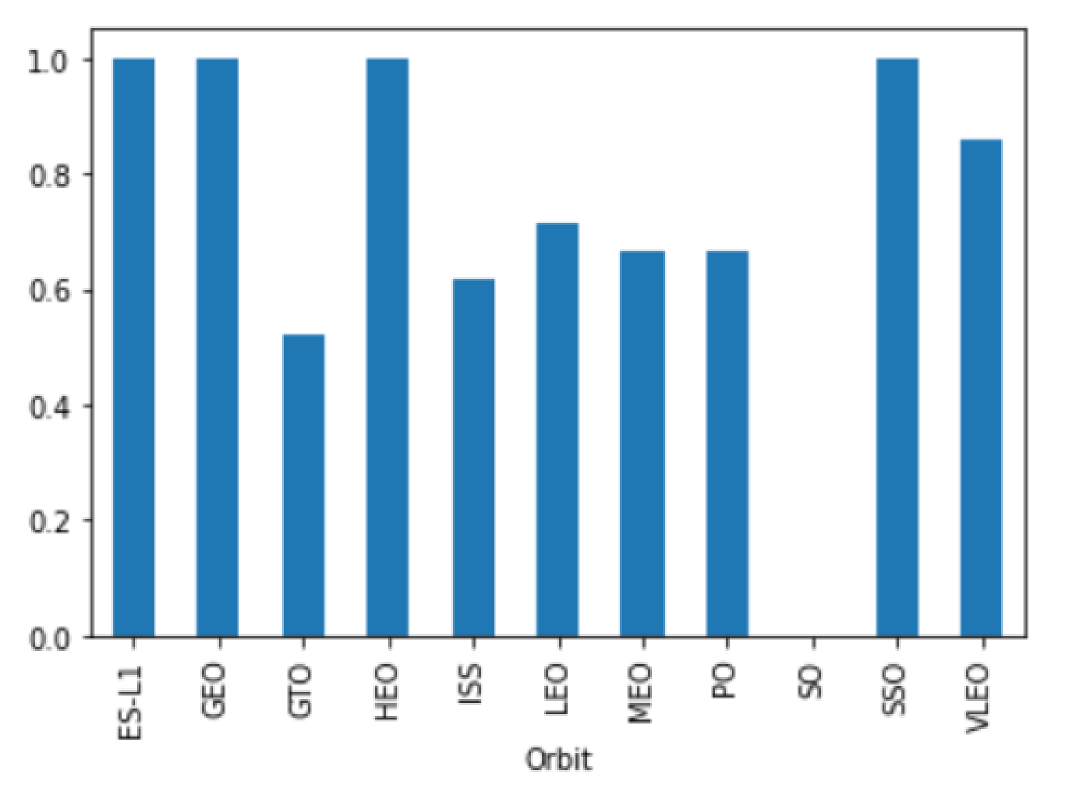
CCAFS launches more missions than other launch sites

Payload vs. Launch Site



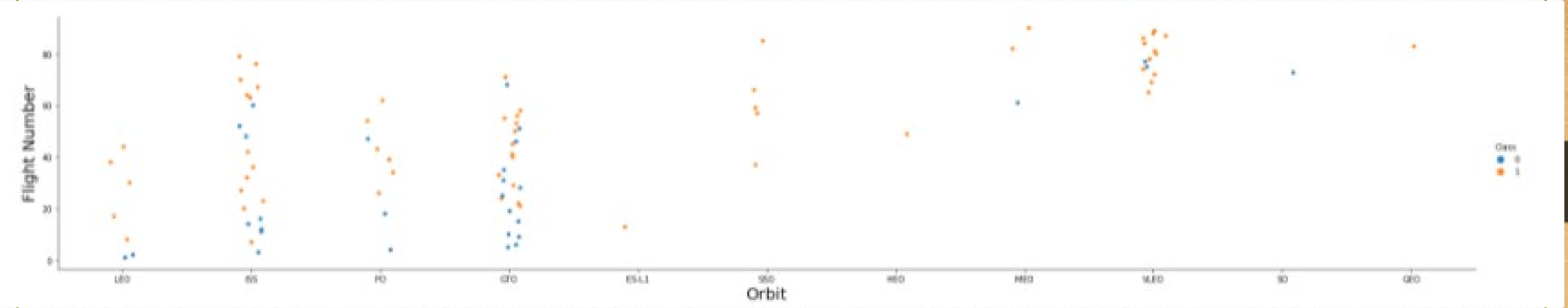
There are more frequently sites launching lower payloads

Success Rate vs. Orbit Type



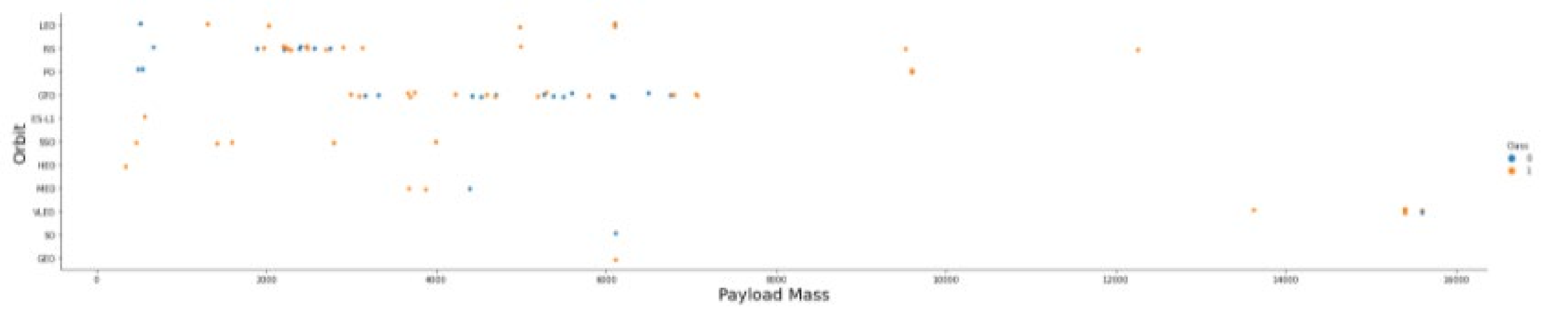
There are four launch sites with 100% success rate SSO, HEO, GEO, ES-L1

Flight Number vs. Orbit Type



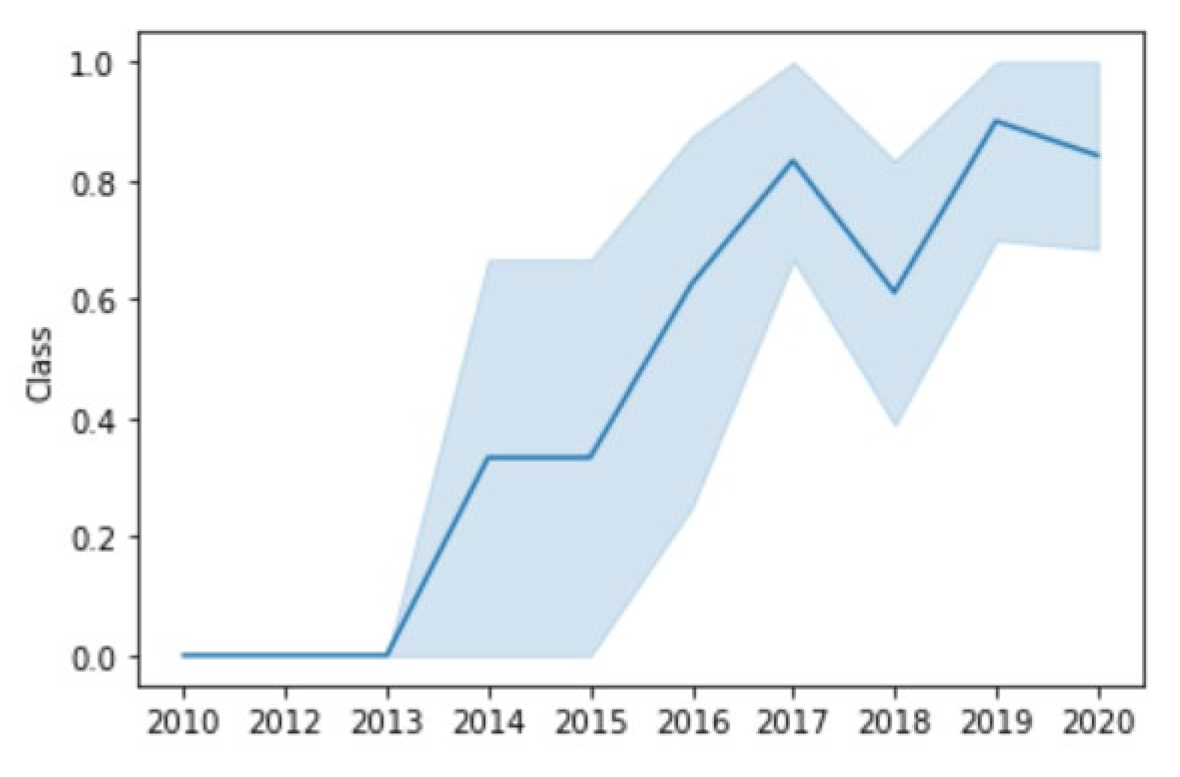
More recent launches have different orbit types than earlier launches

Payload vs. Orbit Type



The SO and GEO orbit types share virtually identical payload capacities

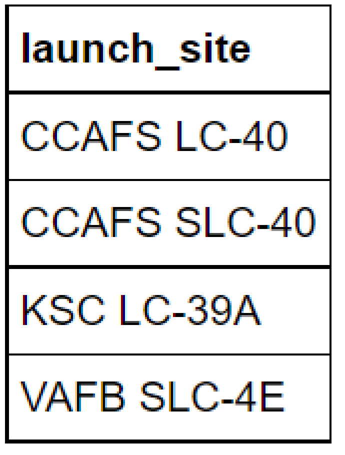
Launch Success Annual Trend



While there are still failures, mission outcomes are becoming increasingly more successful

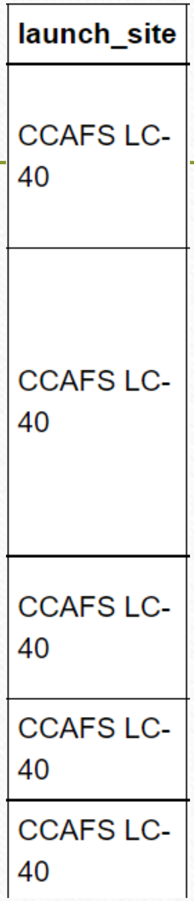
EDA with SQL

Unique Launch Site Names



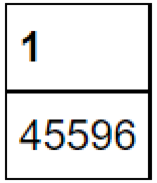
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

5 launch sites beginning with ‘CCA’



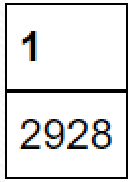
launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass Carried by NASA Boosters



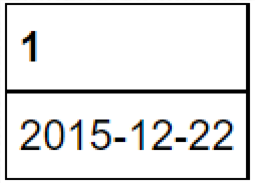
1
45596

Average Payload Mass Carried by Booster Version F9 v1.1



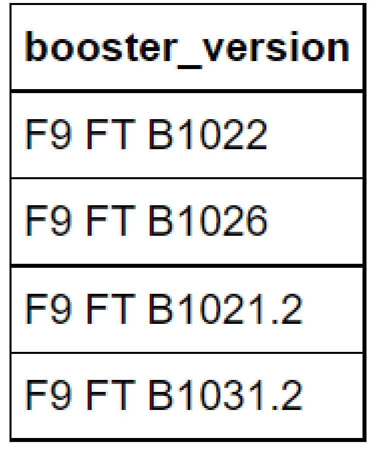
1
2928

Date of The First Successful Landing Outcome



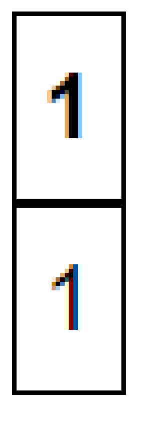
1
2015-12-22

Boosters With Successful Drone Ship Landing And Payload between 4000 and 6000



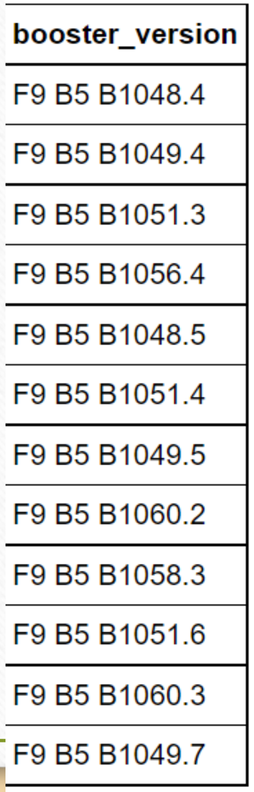
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number Mission Outcomes



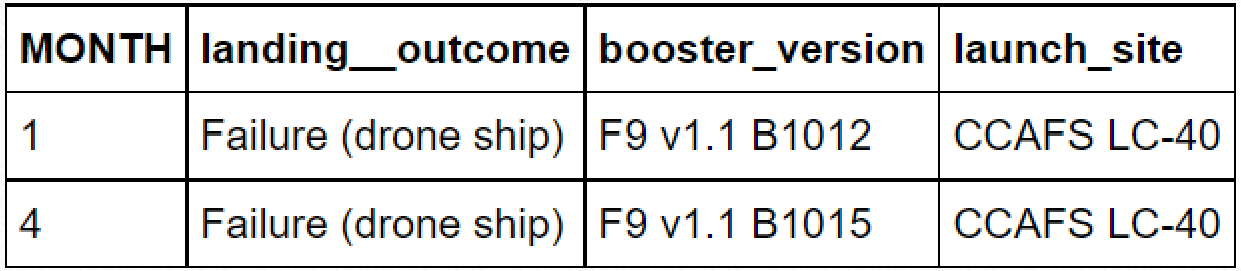
1
1

Boosters Which Have Carried Maximum Payload Mass



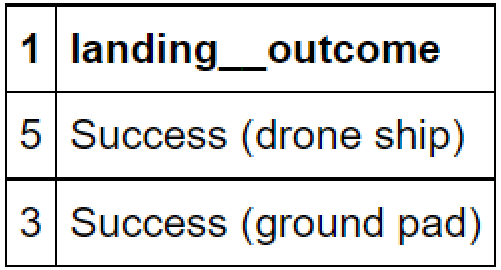
booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Failed Landing Outcomes for year 2015



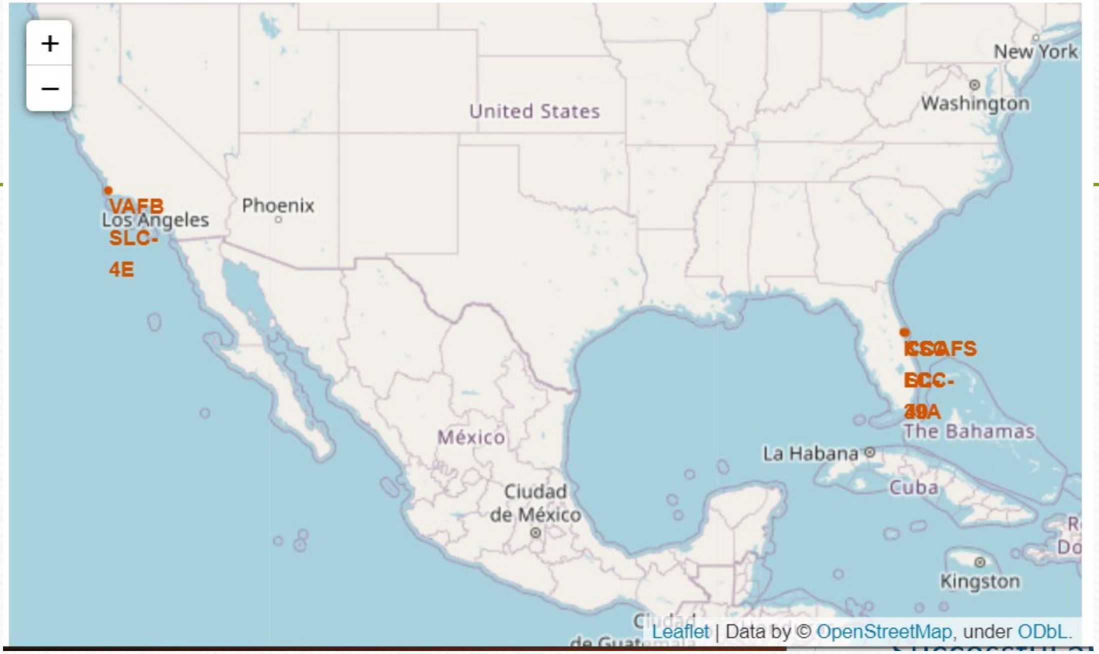
MONTH	landing__outcome	booster_version	launch_site
1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Ranked Count of Landing Outcomes Between 2010-06-04 and 2017-03-20



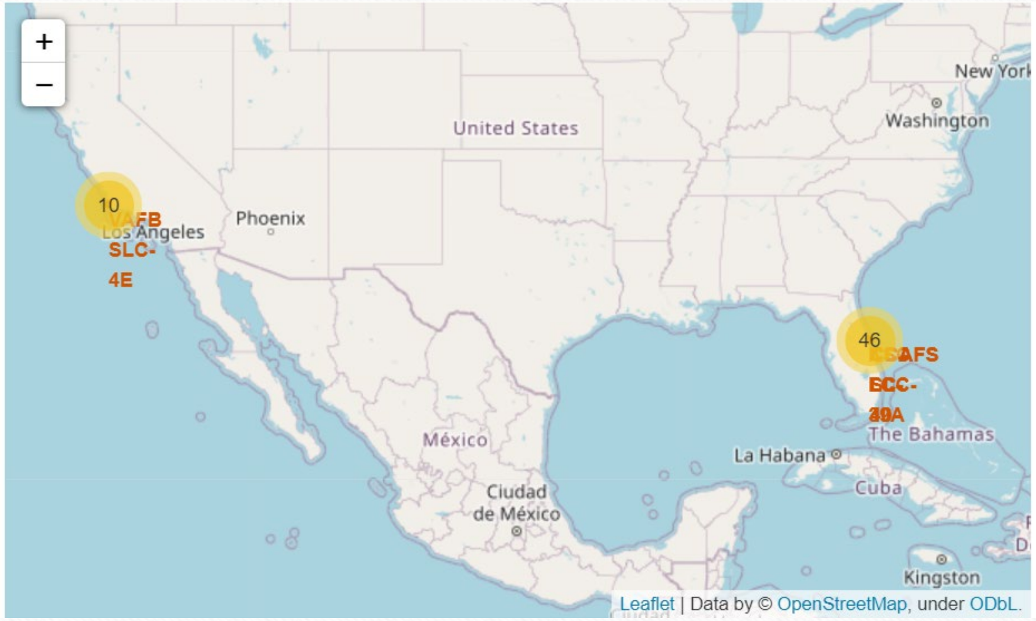
1	landing__outcome
5	Success (drone ship)
3	Success (ground pad)

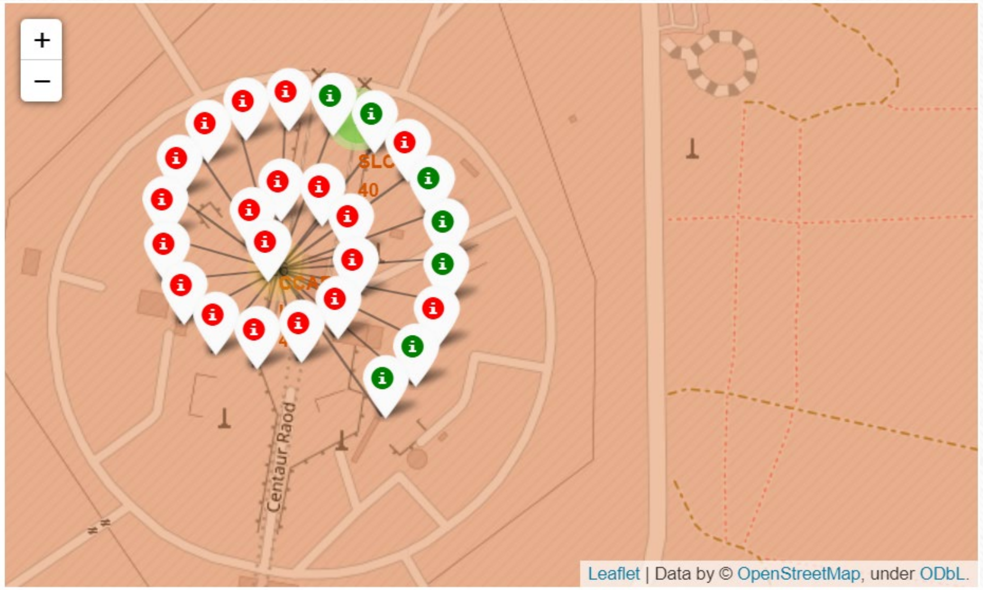
Interactive Folium Map



Marked site launches are seen in California and Florida, near the equator and near coastlines

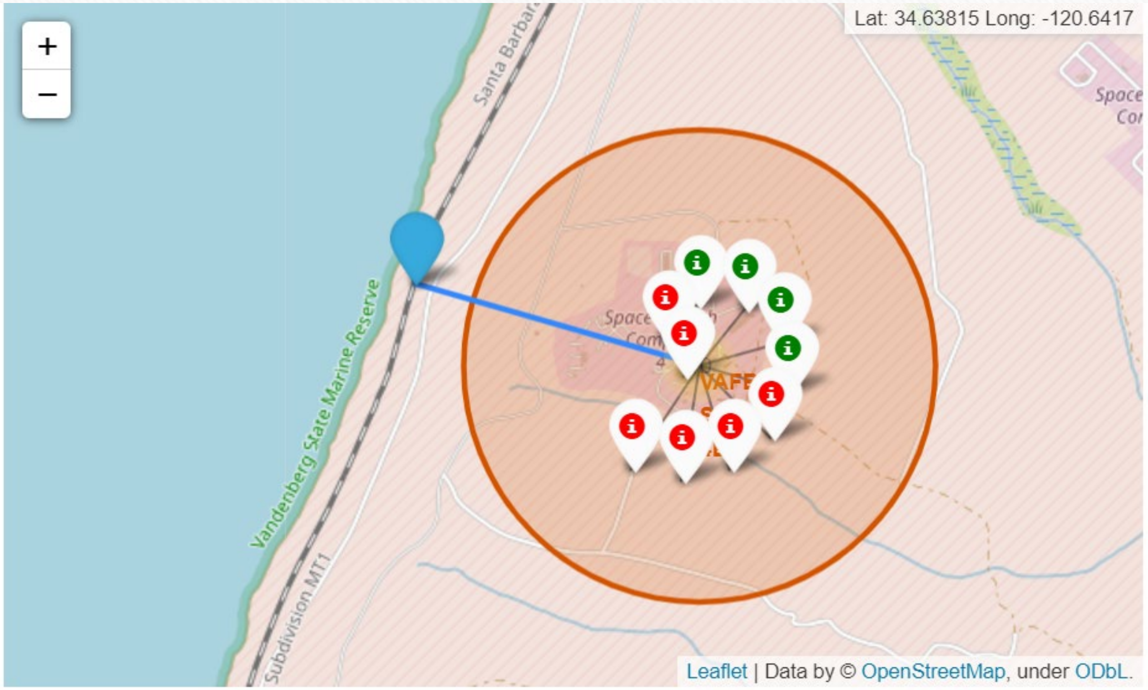
Success/Failed Launches





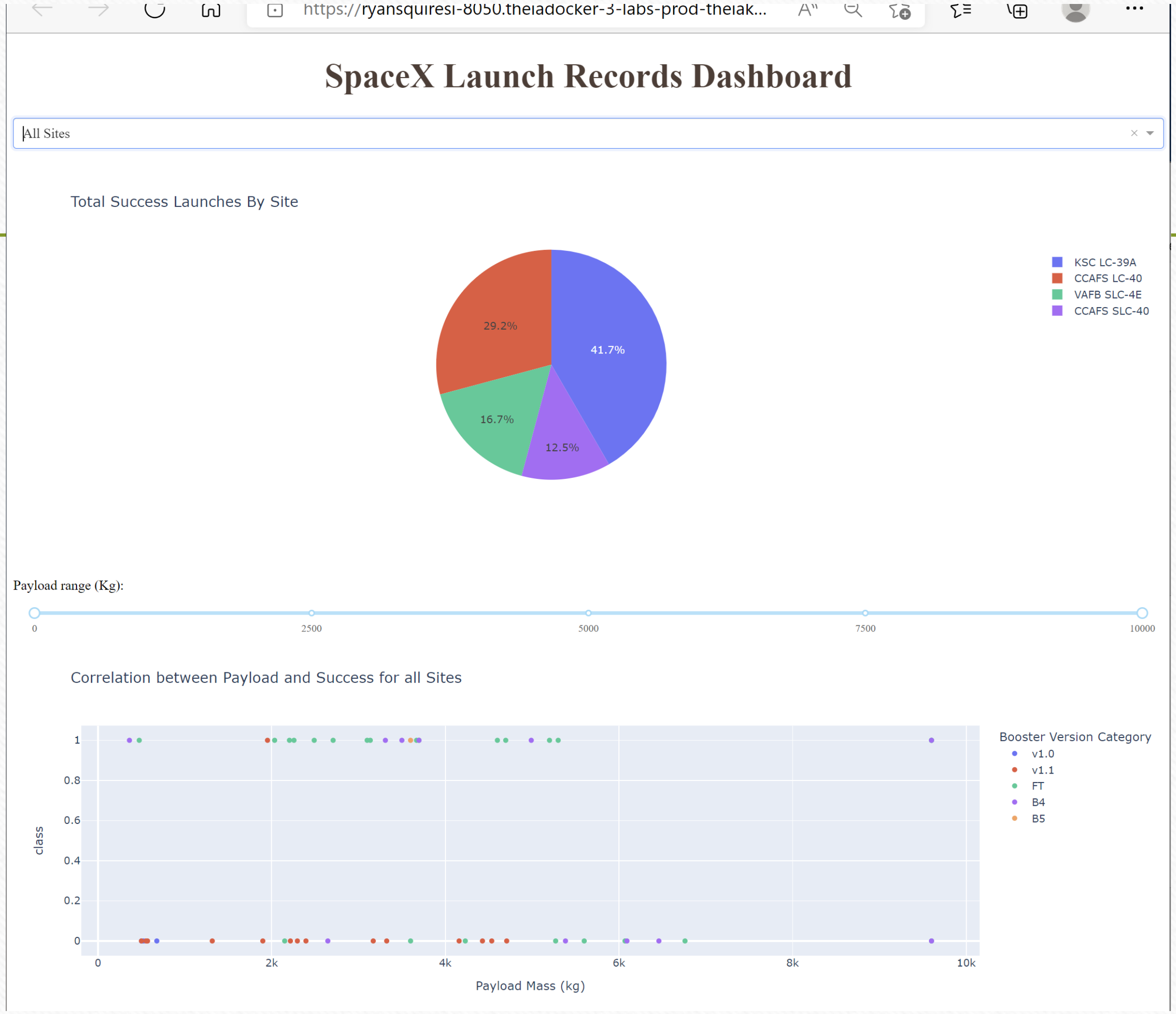
Each launch is numbered and assigned a color, green for successful and red for failed

Site Distance to Proximities


Lat: 34.63815 Long: -120.6417
+
−
Santa Barbara
Vandenberg State Marine Reserve
Subdivision MT1
Space
Com
VAFB
Leaflet | Data by © OpenStreetMap, under ODbL.

Launch site proximities to railways, highways, and coastlines are depicted with lines

Interactive Dashboard with Plotly



https://ryansquiresi-8050.theiadocker-3-labs-prod-theiak...

SpaceX Launch Records Dashboard

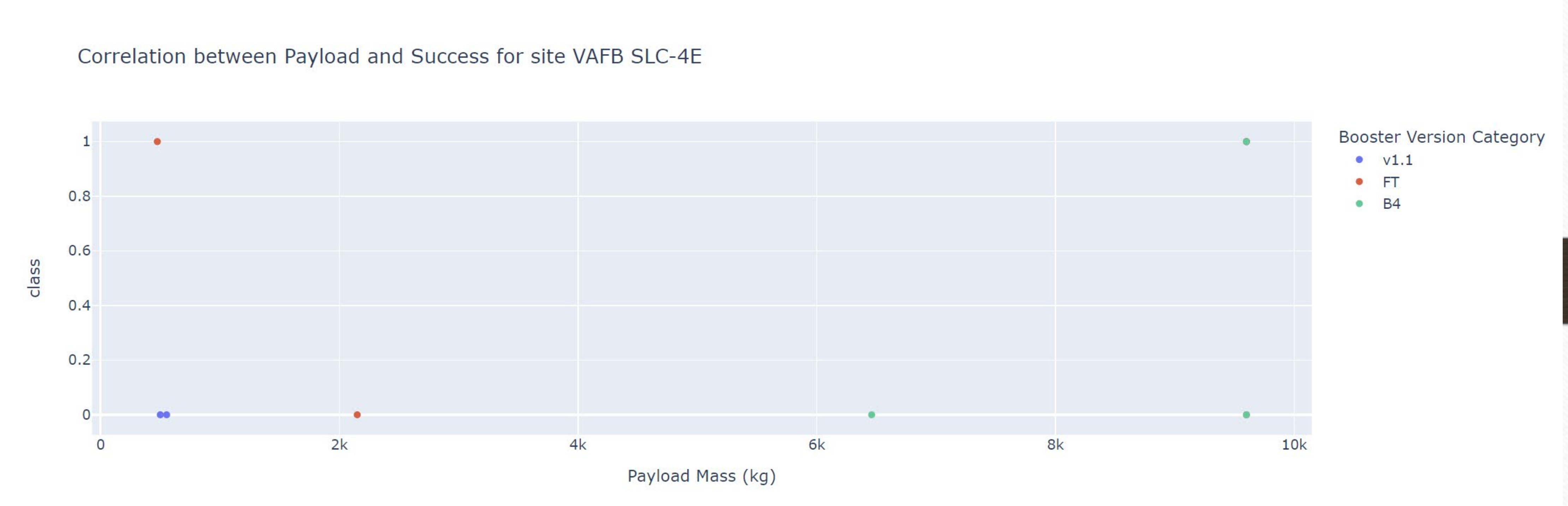
All Sites

Total Success Launches By Site

Payload range (Kg):

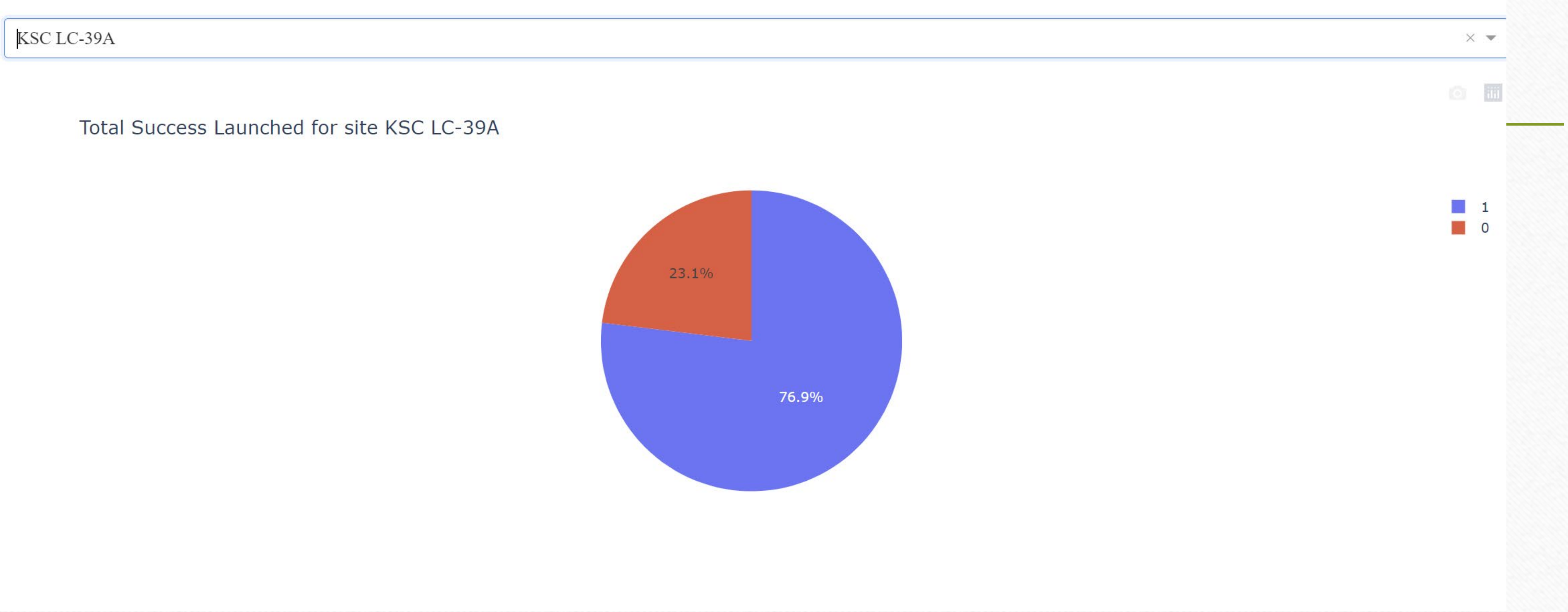
Correlation between Payload and Success for all Sites

Site With Largest Successful Launches



Site VAFB launched successful missions with the highest payload capacity

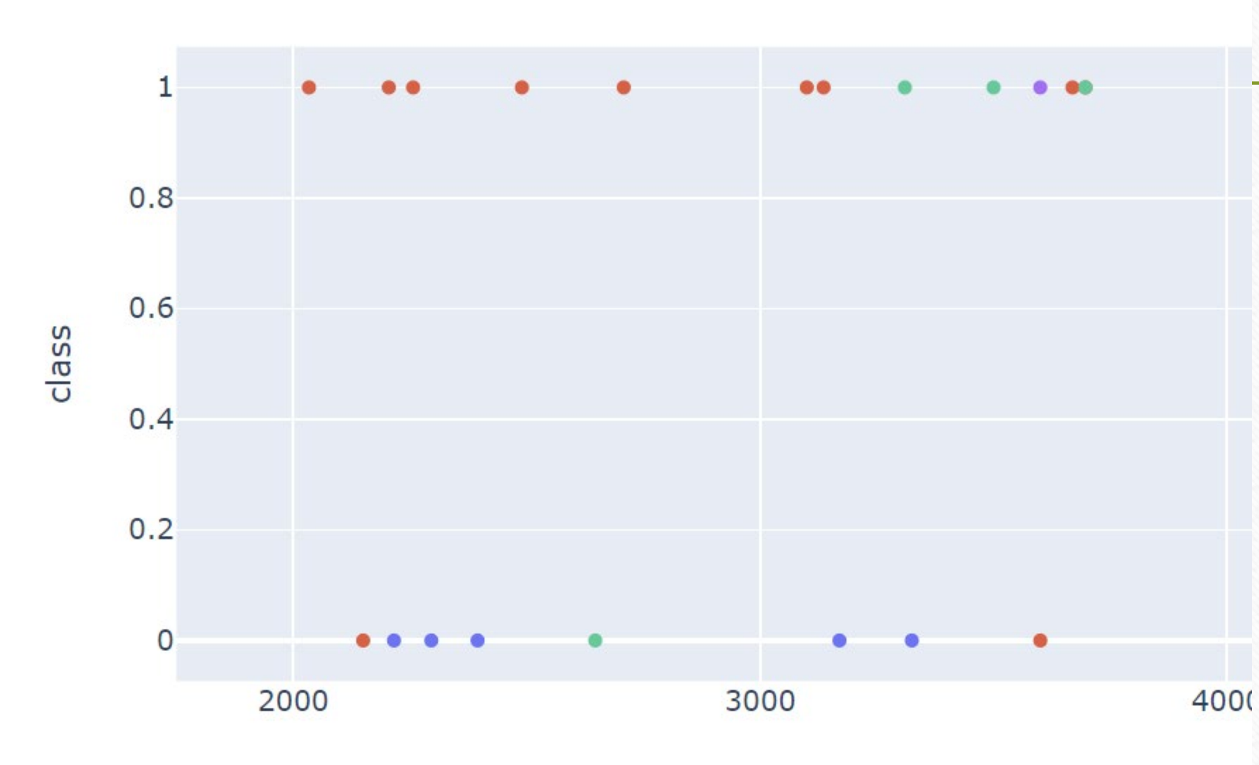
Site with Highest Launch Success Rate



KSC LC-39A

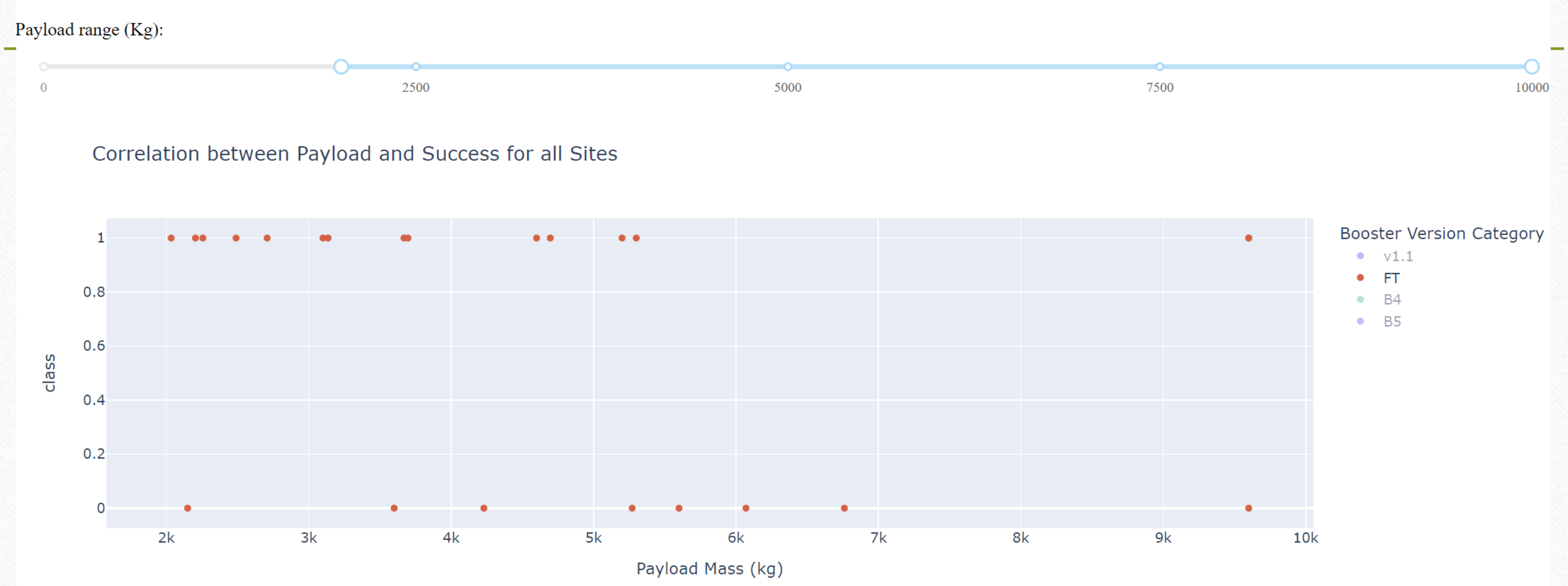
Site KSC LC-39A had the highest success rate

Payload Ranges with Highest Launch Success Rate



The payload range of 2000-3000 shared the same success rate as the payload range of 3000-4000

F9 Booster With Highest Launch Success Rate

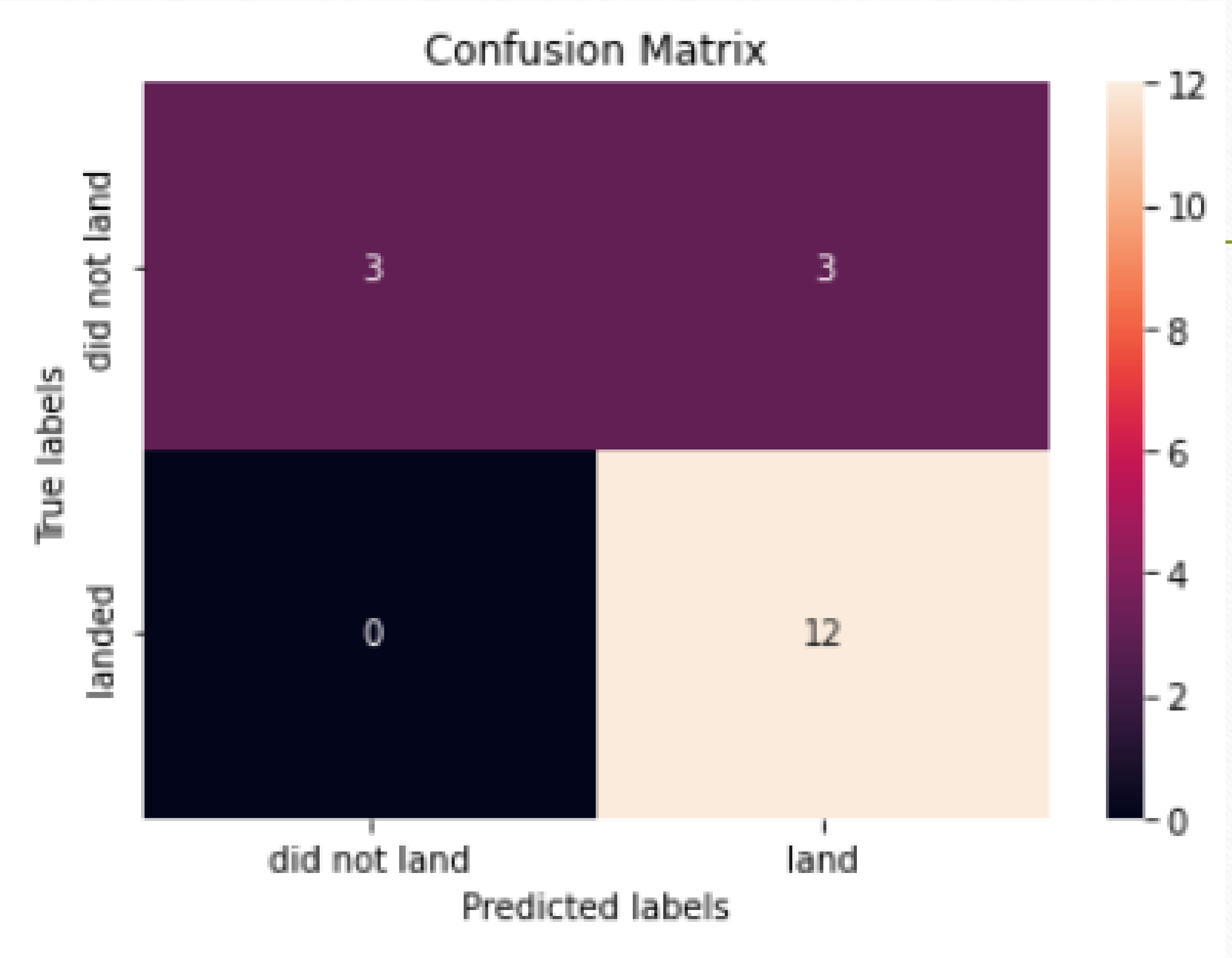


Payload range (Kg):

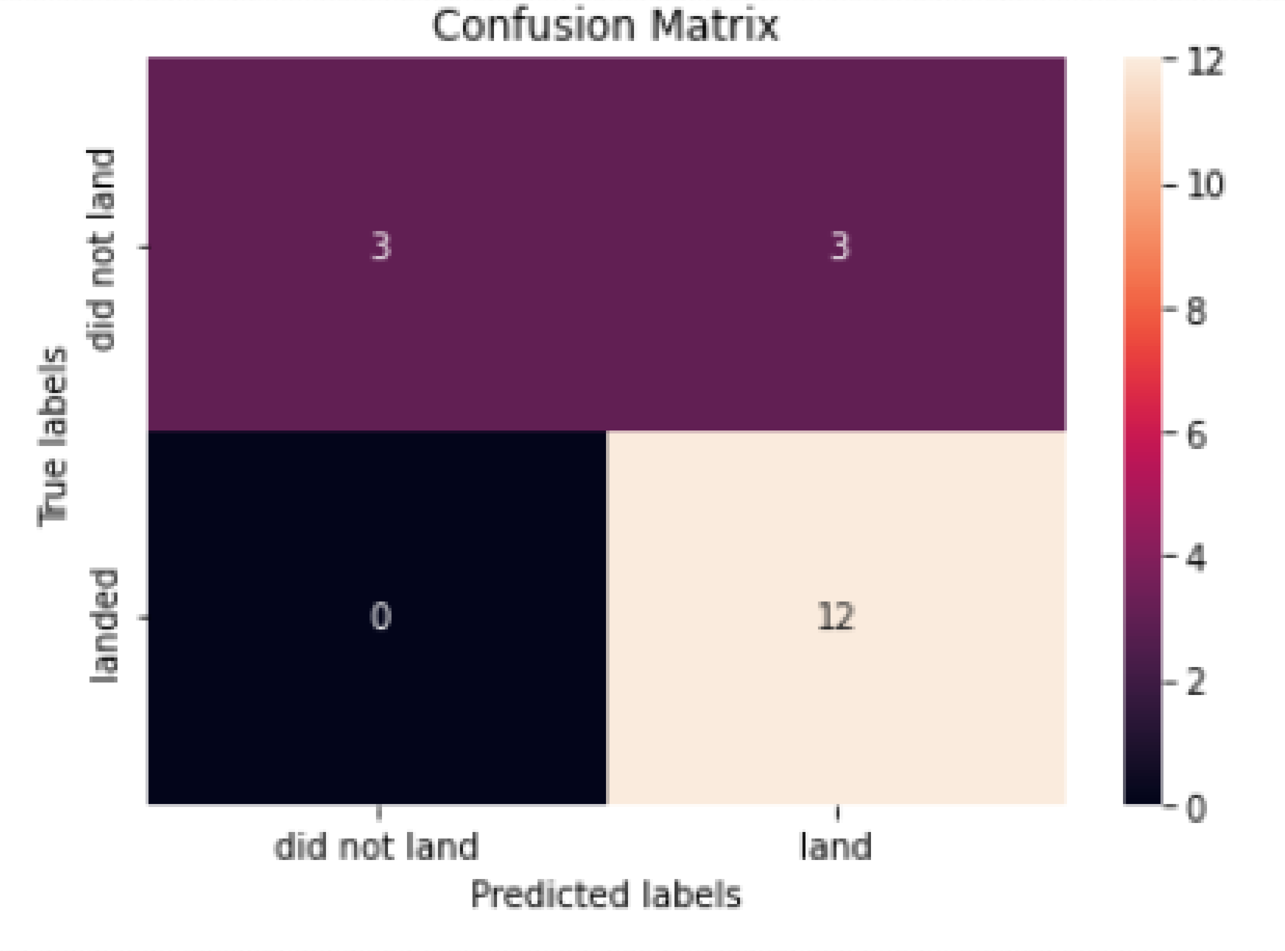
The FT F9 Booster had the highest launch success rate

Predictive Analysis Classification

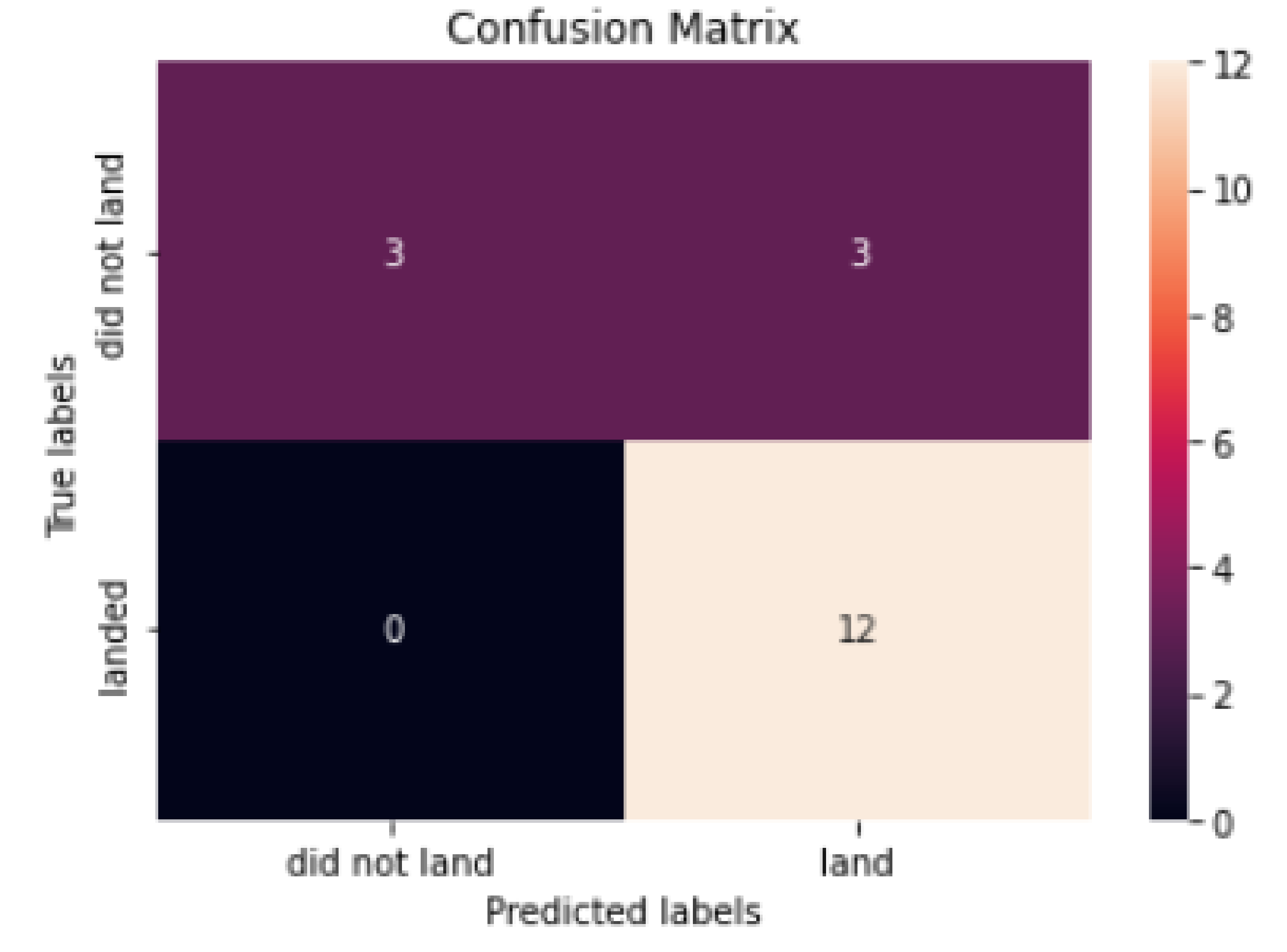
Logistics Regression Method



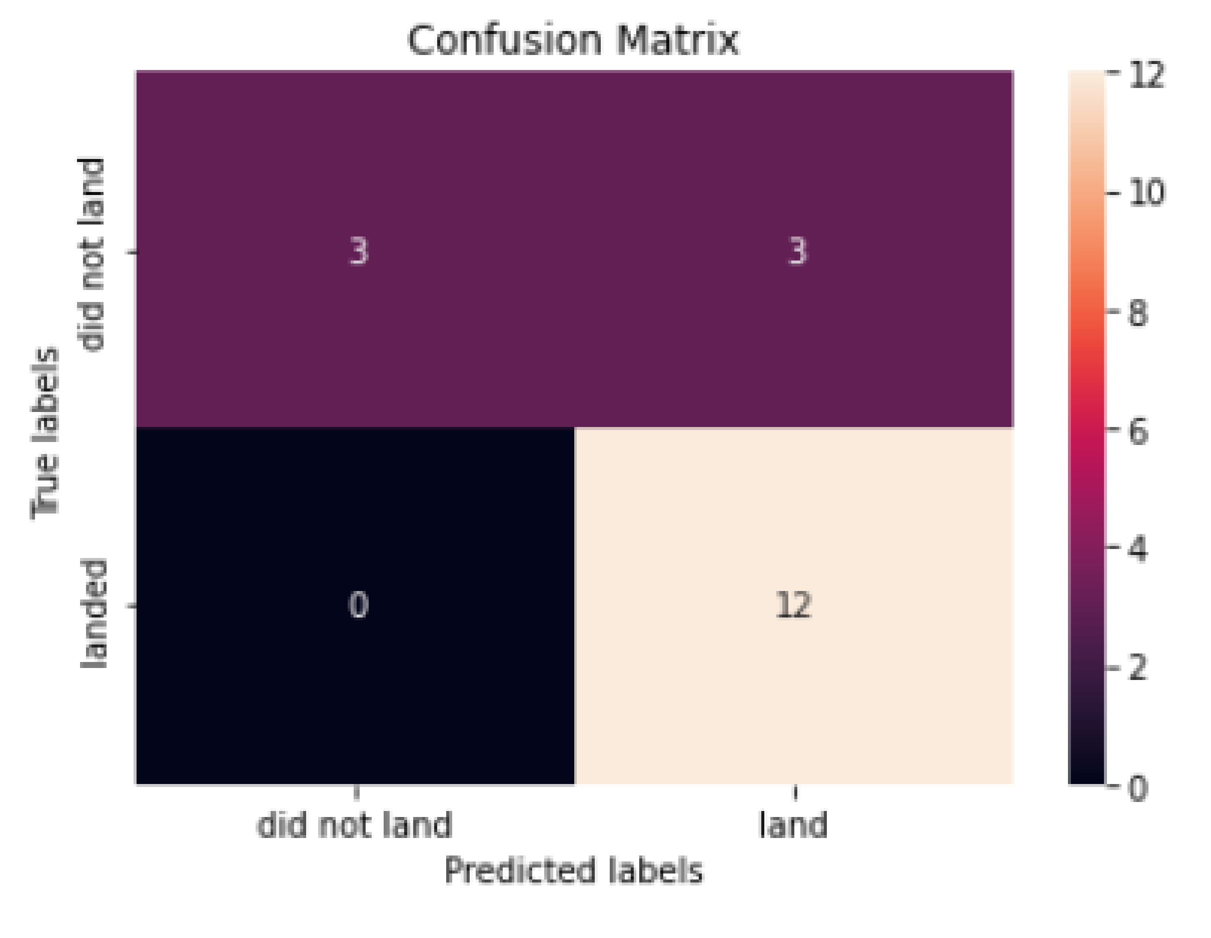
Support Vector Machine Method



Decision Tree Method



K Nearest Neighbors Method



Conclusion

Using data analysis and machine learning we were able to better understand SpaceX Launches and make predictions as to mission success and whether or not the first stage of the Falcon 9 could be reused.

- There is a 100% success rate for orbit types of SSO, HEO, GEO, ES-L1
- The KSC LC-39A launch site has the highest success rate
- In the early rocket launches, rocket launches were less successful when carrying higher payload capacities
- Over time, missions have become more successful have been able to carry higher payload capacities

Appendix

- Small sample size of 18
- Missing data was replaced with mean values

This project was completed through the IBM Cloud and Skills Network Labs