IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Spencer Smith
5/3/2023

Outline

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

Executive Summary

- We collected data from the SpaceX API and the SpaceX Wikipedia page to develop a predictive model for successful landings of SpaceX rockets. Our exploratory analysis used SQL, visualization, folium maps, and dashboards. To prepare the data, we transformed categorical variables to binary, standardized the data, and used GridSearchCV to optimize several different machine learning models. We evaluated our models based on accuracy scores.
- Four machine learning models were employed in our study, including Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. While the first three models achieved an accuracy rate of approximately 83.33%, the Decision Tree Classifier produced a lower accuracy rate of 66%. Notably, all models over-predicted successful landings, indicating the need for more data to improve model accuracy. Our study highlights the importance of continuous improvement and iteration in developing robust predictive models.

Introduction

BACKGROUND

- SpaceX founder Elon Musk recognized the potential cost savings of developing reusable rockets and set a goal to make it a reality.
- The first successful landing of a Falcon 9 rocket booster occurred in 2015, marking a significant achievement in reusable rocket technology.
- Reusable rockets have the potential to significantly reduce the cost of space exploration and commercial satellite launches, making it more accessible to a wider range of organizations and researchers.

PROBLEM

- Space Y wants us to figure out how to predict Stage 1 recoveries

Section 1

Methodology

Methodology

Executive Summary

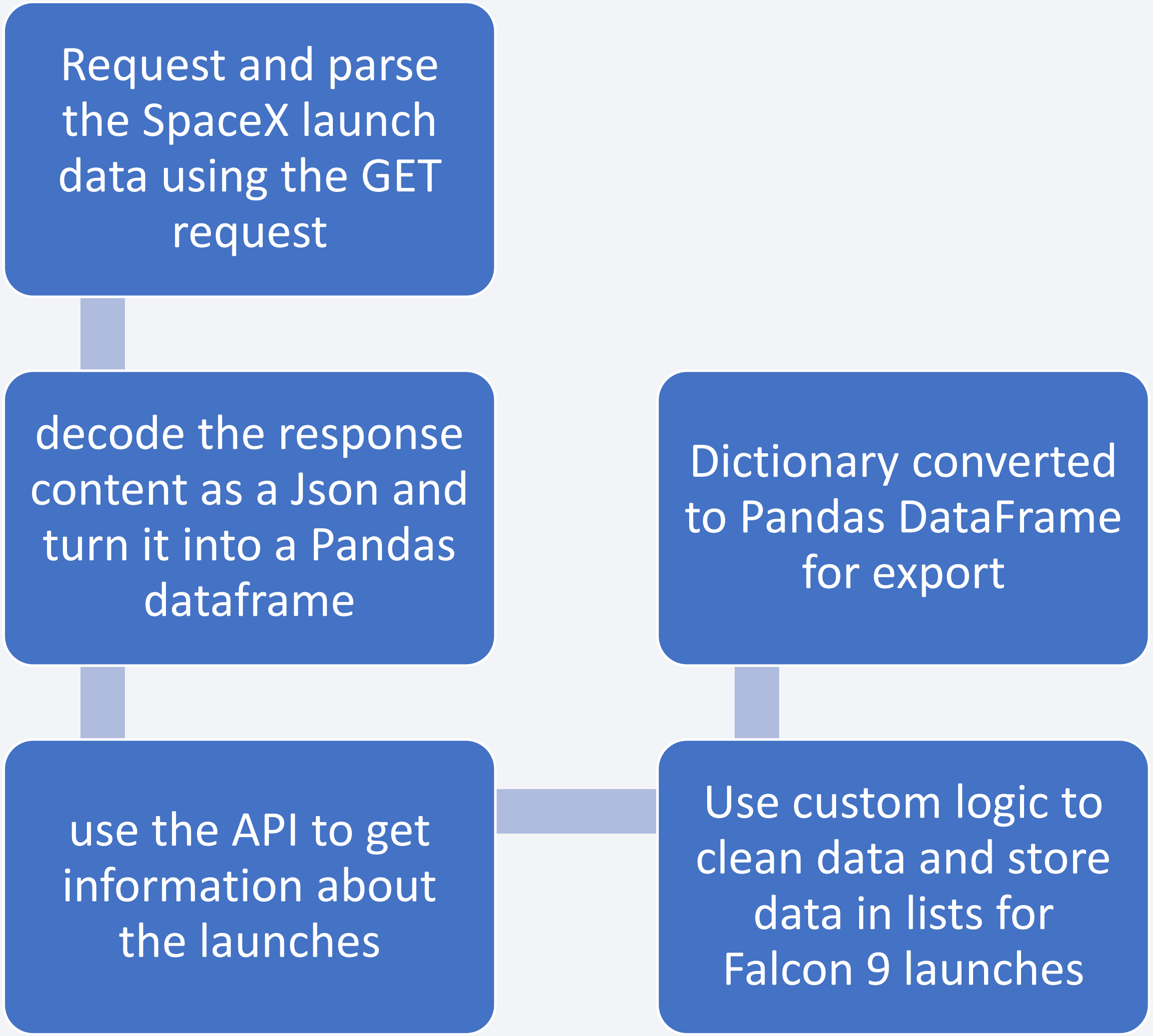
- Data collection methodology:
    - Data was obtained through SpaceX public API and the SpaceX Wikipedia page
- Perform data wrangling
    - Replace null values
    - determine number of launches and occurrence of landing outcome (successful vs. not)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Use Scikit-Learn to standardize data, train machine learning models, fine tune machine learning models, assess accuracy of models

Data Collection

- The data collection process for this project utilized a combination of methods, including API requests from SpaceX's public API and web scraping data from a table on SpaceX's Wikipedia page.
- The SpaceX API provided several columns of relevant data, including FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- For the Wikipedia web scraping process, we collected Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time columns.
- To better illustrate the data collection process, the following slide will display flowcharts for the API and web scraping methods. This comprehensive data collection approach allowed us to gather a wide range of data points and facilitate the development of our predictive model.

Data Collection – SpaceX API

- GitHub URL:
  - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_Collecting_Data.ipynb



Request and parse the SpaceX launch data using the GET request

decode the response content as a Json and turn it into a Pandas dataframe

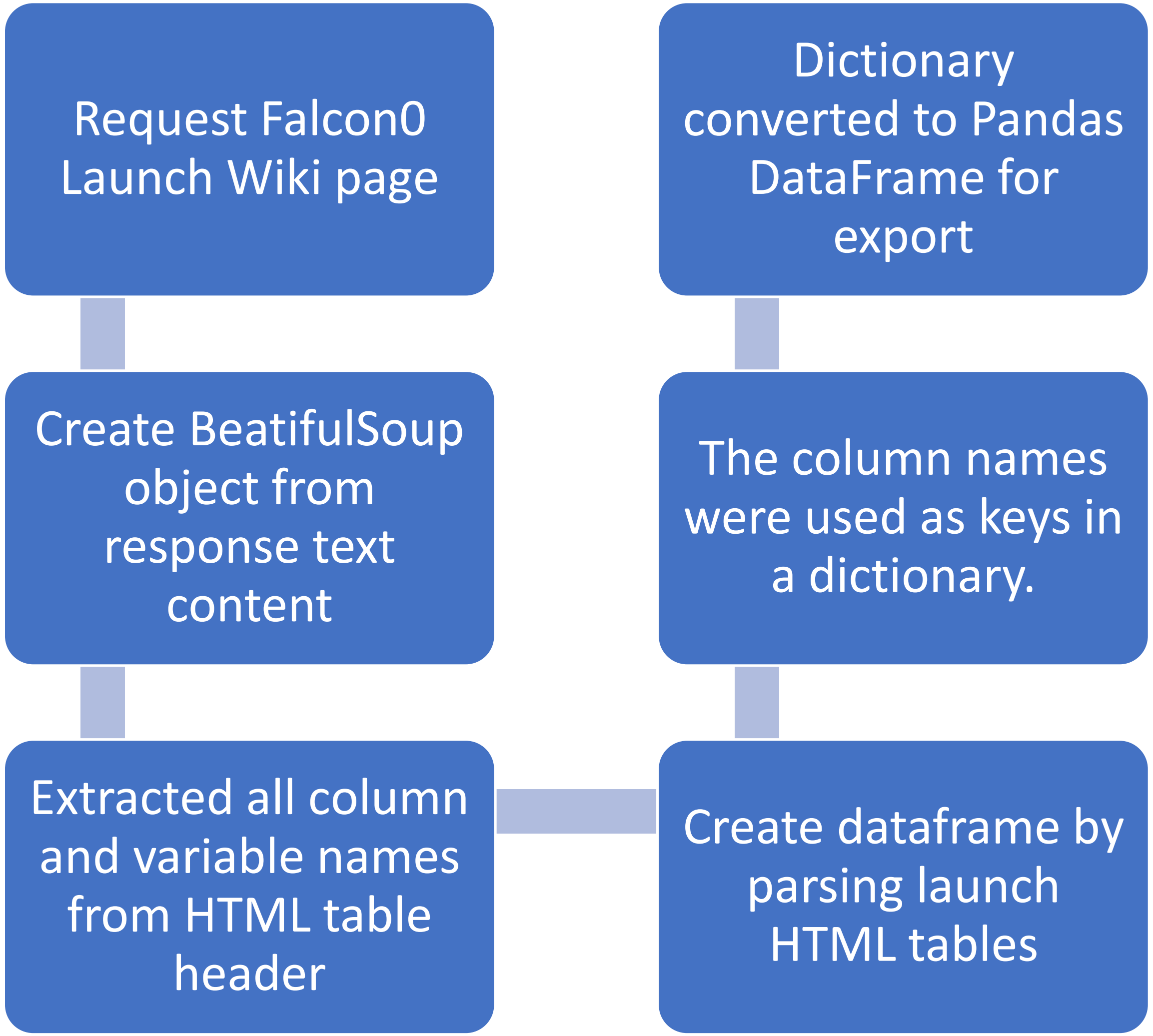
use the API to get information about the launches

Use custom logic to clean data and store data in lists for Falcon 9 launches

Dictionary converted to Pandas DataFrame for export

Data Collection - Scraping

- GitHub URL:
  - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_Web_Scraping.ipynb



Request Falcon0 Launch Wiki page

Create BeatifulSoup object from response text content

Extracted all column and variable names from HTML table header

Create dataframe by parsing launch HTML tables

The column names were used as keys in a dictionary.

Dictionary converted to Pandas DataFrame for export

Data Wrangling

- Process

1. Determine Successful outcomes
    1. Successful: True ASDS, True RTLS, True Ocean
    2. Unsuccessful: None None, False ASDS, False Ocean, False RTLS

2. Create successful outcome variable
    1. Successful = 1; Unsuccessful = 0

- GitHub URL:
    - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_Data_Wrangling.ipynb

EDA with Data Visualization

- Scatter Plots
  - Used to visualize relationships/correlations between 2 continuous variables: Flight Number vs. Launch Site, Payload vs. Launch Site, Orbit Type vs. Flight Number, and Payload vs. Orbit Type
- Bar Chart
  - Used to compare numeric value to categorical variable: Success Rate vs. Orbit Type
- Line Charts
  - Used to compare numerical values' change over time: Success Rate vs. Year
- GitHut URL:
  - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_Explore_Prepare_Data.ipynb

EDA with SQL

SQL QUERIES USED:

- The names of unique launch sites in the space mission were displayed.
- Five records were displayed where launch sites begin with the string 'CCA'.
- The total payload mass carried by boosters launched by NASA (CRS) was displayed.
- The average payload mass carried by booster version F9 v1.1 was displayed.
- The date of the first successful landing outcome on a ground pad was listed.
- The names of boosters with success on a drone ship and a payload mass between 4000 and 6000 kg were listed.
- The total number of successful and failed mission outcomes was listed.
- The names of booster versions that have carried the maximum payload mass were listed.
- Failed landing outcomes on drone ships, their booster versions, and launch site names for 2015 were listed.
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 was ranked in descending order.

- GitHub Link:
  - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_EDA_SQL.ipynb

Build an Interactive Map with Folium

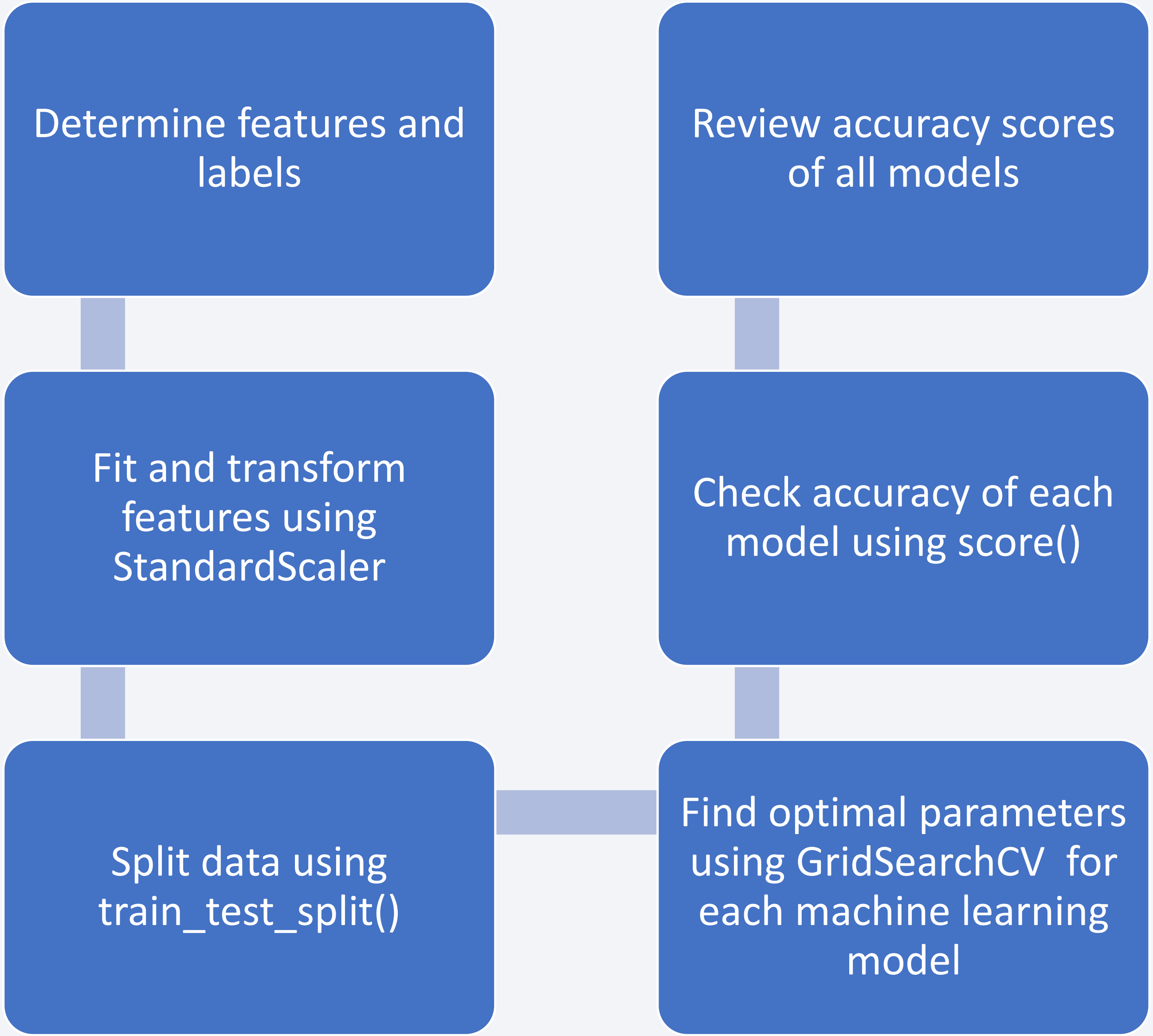
- Mapped all launch sites by initializing a Folium Map object and adding a folium.Circle and folium.Marker for each site on the launch map.
- Visualize the success/failure of launches for each site by clustering launches together based on their coordinates. Assign a marker color of green for successful launches (class = 1) and red for failed launches (class = 0) before clustering. Add a folium.Marker to the MarkerCluster() object for each launch, creating an icon as a text label with the icon_color set to the previously determined marker_colour.
- Calculate the distances between launch sites and their proximities by using the Lat and Long values. After marking a point with the Lat and Long values, create a folium.Marker object to display the distance. To show the distance line between two points, draw a folium.PolyLine and add it to the map.
- GitHub link:
  - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_Folium_Locations.ipynb

Build a Dashboard with Plotly Dash

- To visualize the success rate of each launch site, create a pie chart using px.pie() to show the total number of successful launches per site. This will help to identify the sites that have the highest success rates. Additionally, use dcc.Dropdown() to allow users to filter the chart and see the success/failure ratio for an individual site.
- To explore the relationship between the outcome of launches (success or failure) and the payload mass (kg), create a scatter plot using px.scatter(). This will allow users to see if there is any correlation between these two variables. To provide more insights, use RangeSlider() to filter the scatter plot by ranges of payload masses and by booster version.Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- GitHub link:
  - https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

GitHub link: https://github.com/spencersmith1/IBM_Course_Capstone/blob/main/Lab_Machine_Learning.ipynb



Determine features and labels

Fit and transform features using StandardScaler

Split data using train_test_split()

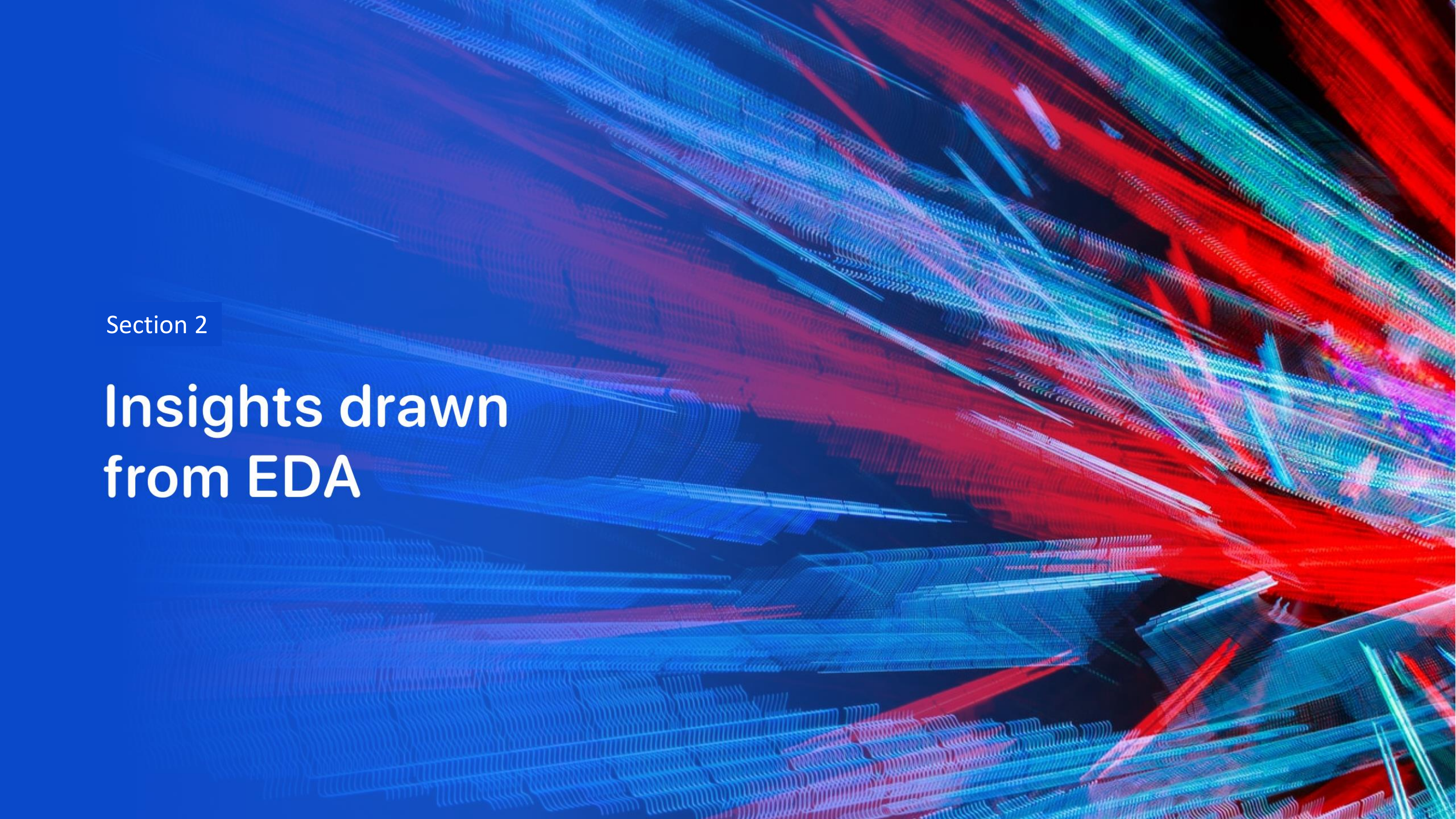
Find optimal parameters using GridSearchCV for each machine learning model

Check accuracy of each model using score()

Review accuracy scores of all models

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



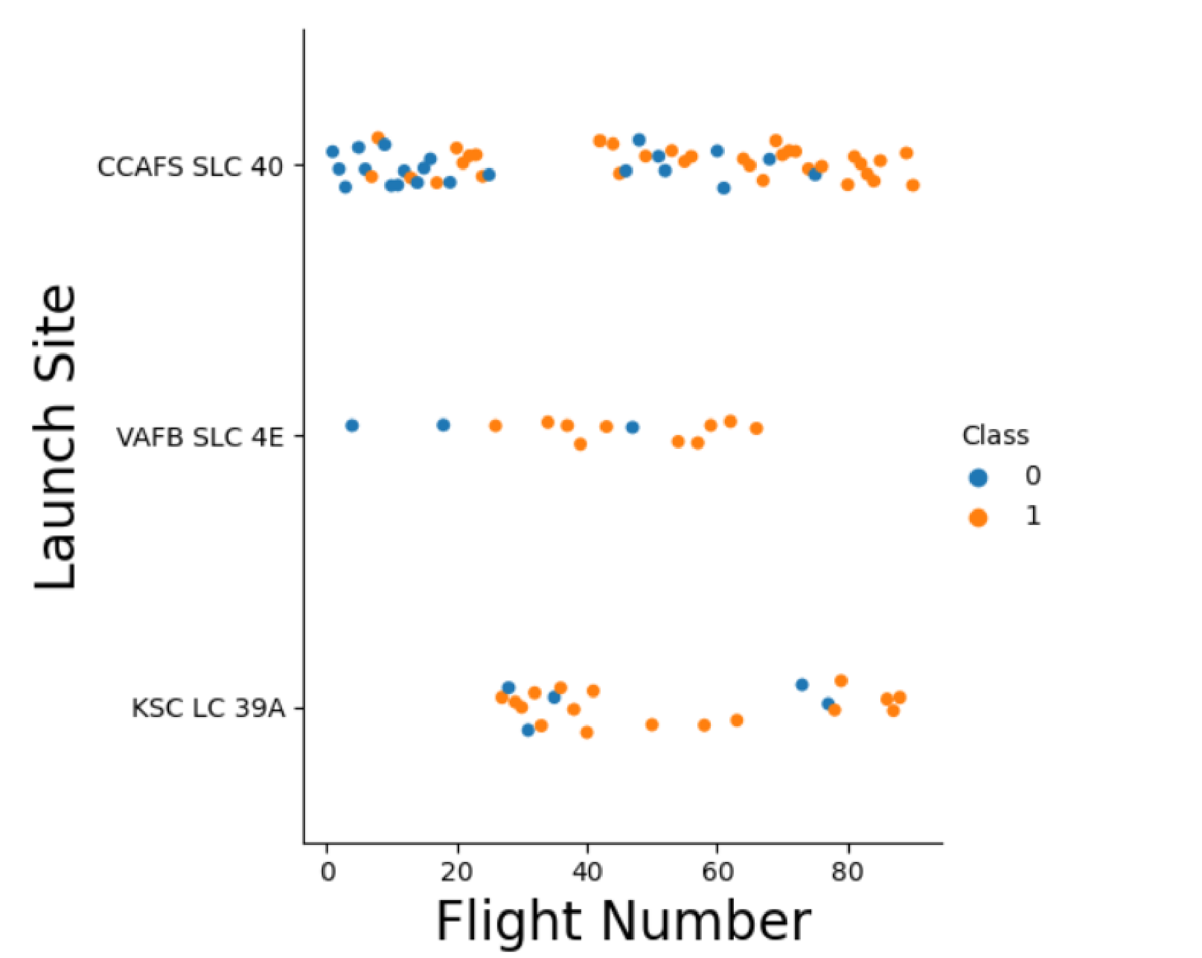
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Orange = successful launch
Blue = unsuccessful launch.

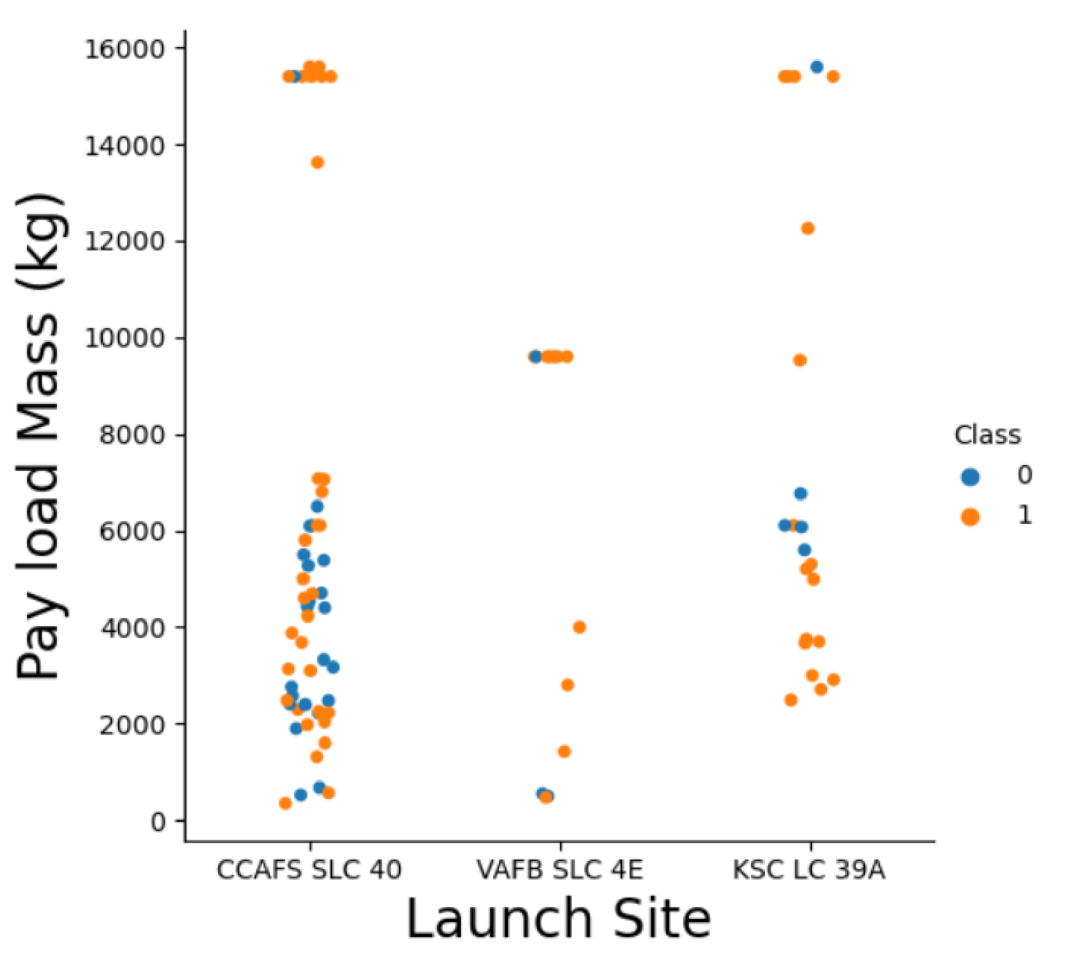
According to the graph, there has been a steady rise in the success rate of flights over time, as shown by the Flight Number. It is probable that there was a major breakthrough around the 20th flight, which resulted in a significant increase in the success rate. The launch site at CCAFS seems to be the primary location for launches, as it has the highest volume.



Payload vs. Launch Site

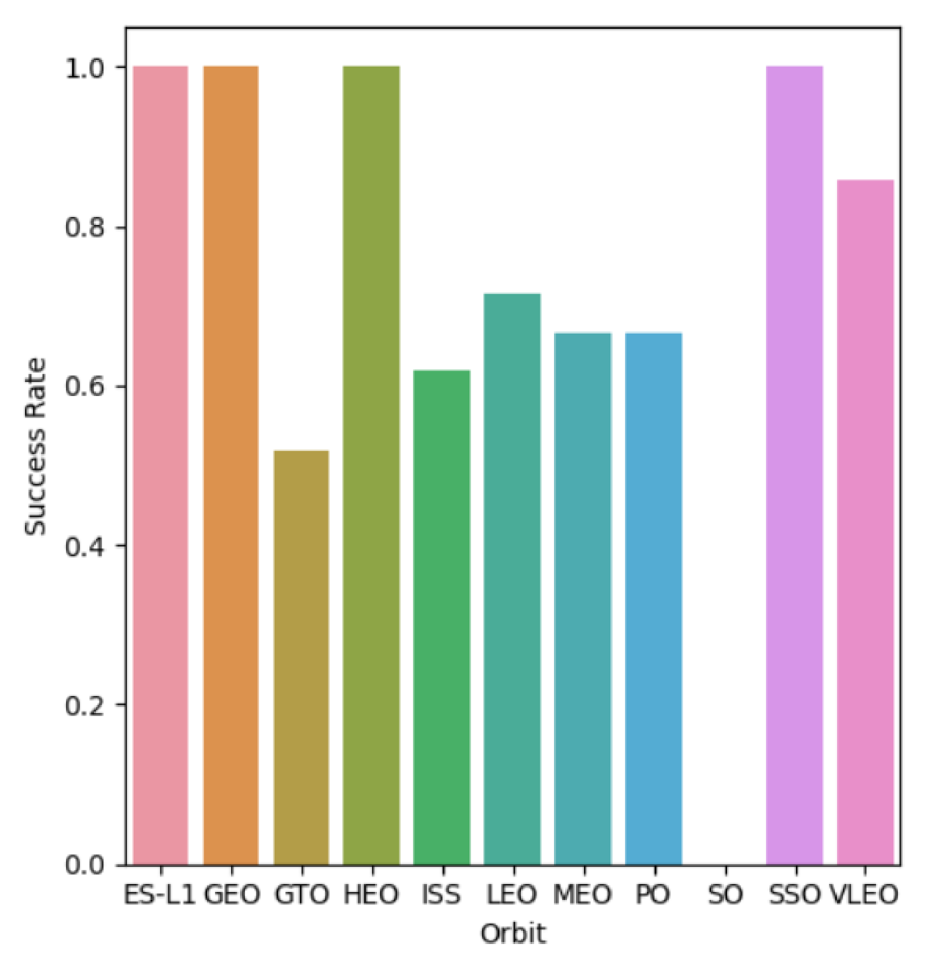
Orange = successful launch
Blue = unsuccessful launch

Different Launch sites use different payloads.



Success Rate vs. Orbit Type

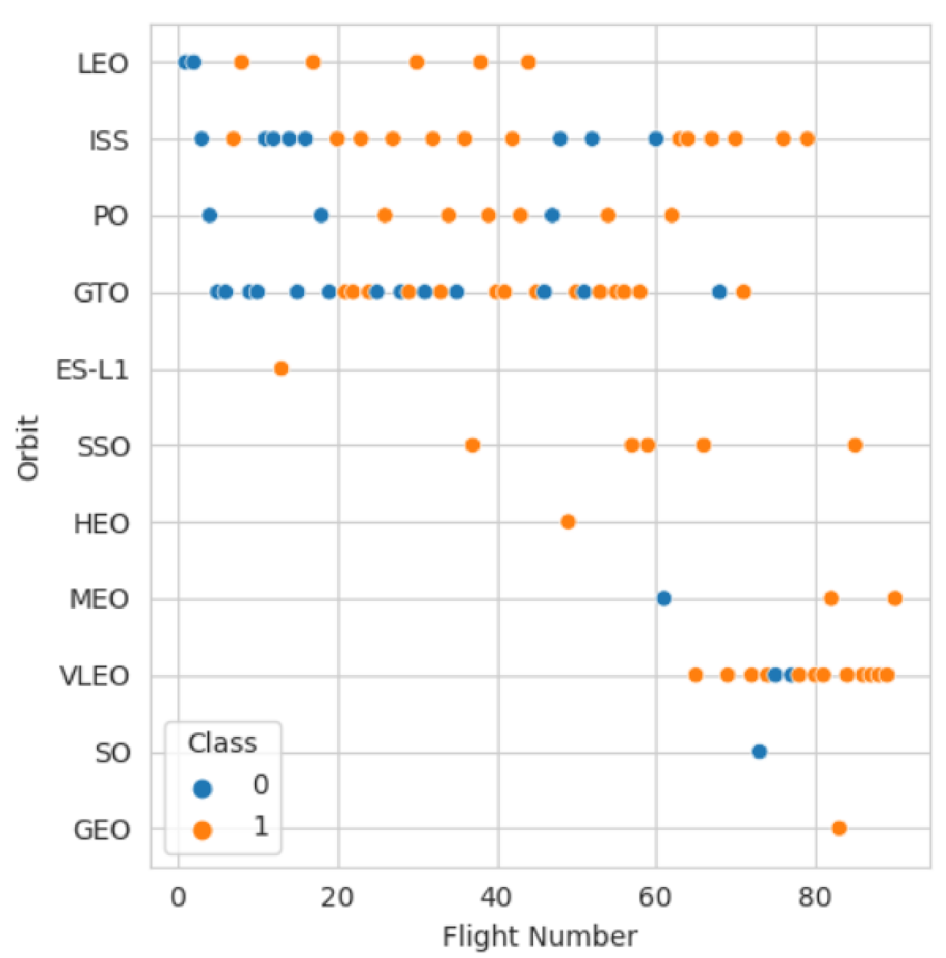
Different orbit types have different success rates



Flight Number vs. Orbit Type

Orange = successful launch
Blue = unsuccessful launch.

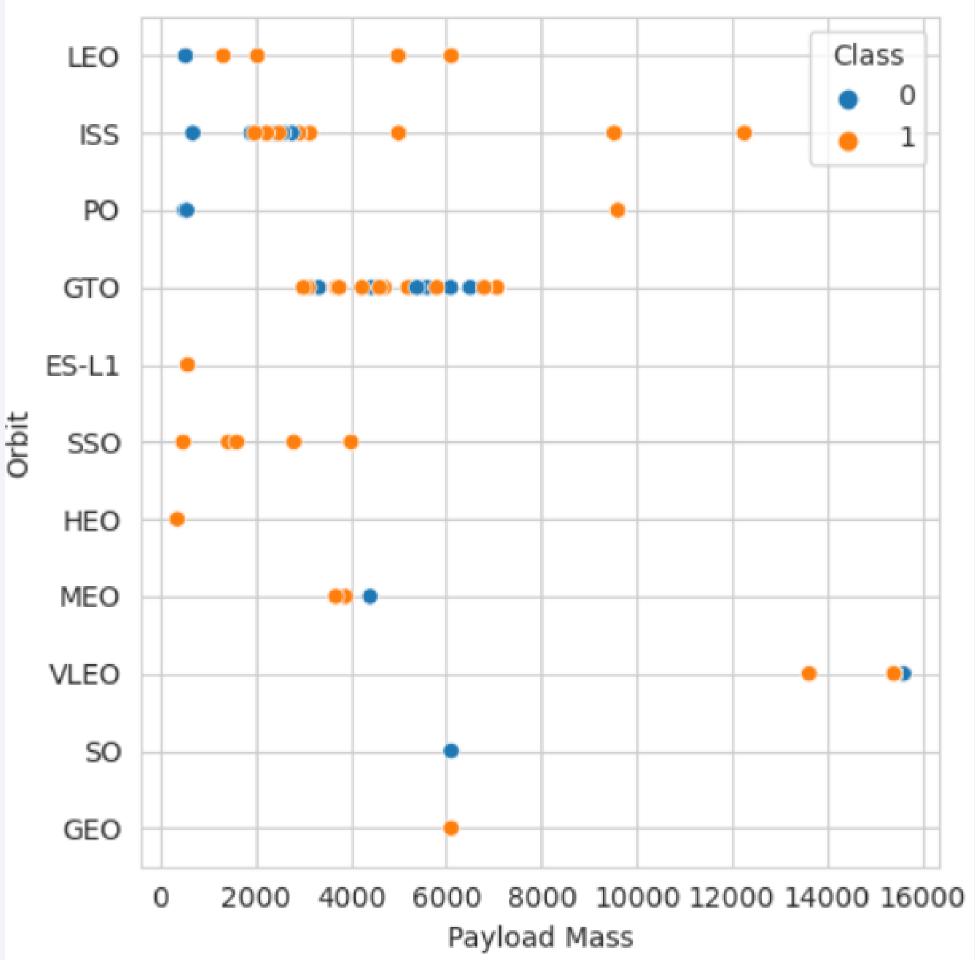
As the Flight Number increased, there was a change in the preferred Launch Orbit. This change in preference appears to be correlated with the Launch Outcome.



Payload vs. Orbit Type

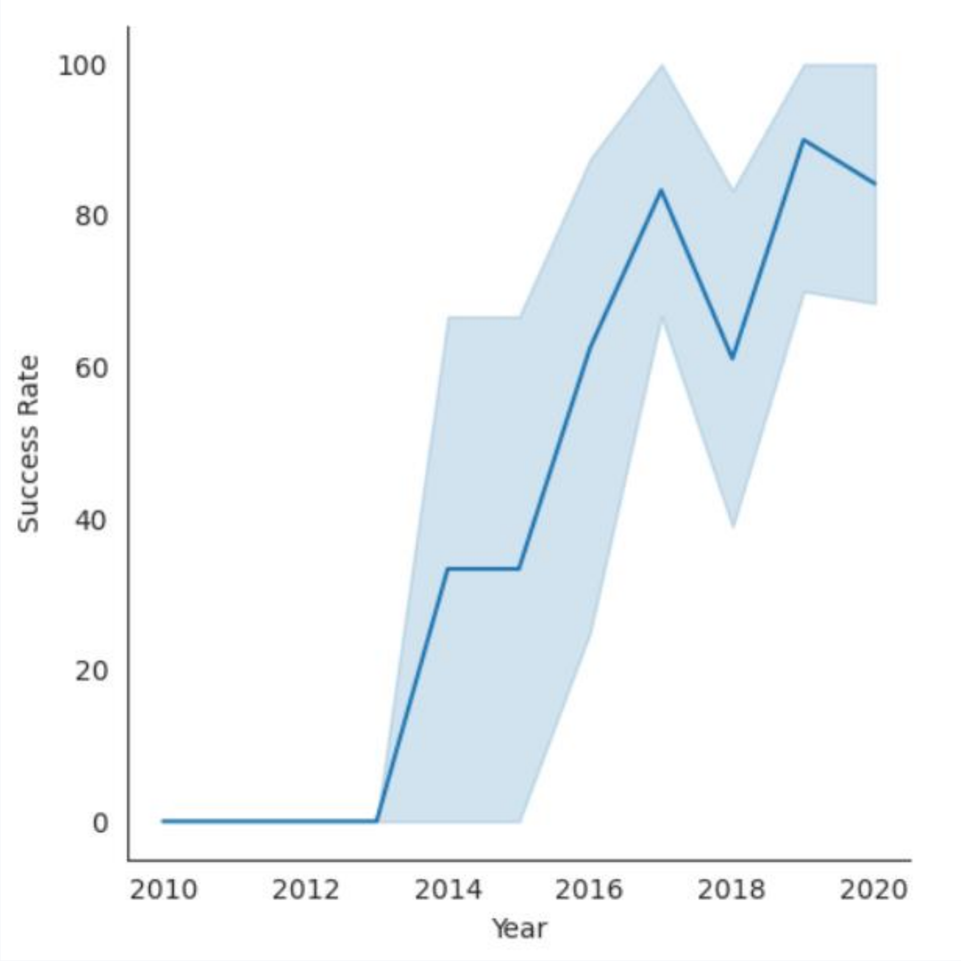
Orange = successful launch
Blue = unsuccessful launch.

The mass of the Payload appears to have a correlation with the Orbit type. Orbits such as LEO and SSO tend to have Payloads with lower mass, while VLEO - which is one of the most successful Orbits - tends to have Payloads with higher mass values.



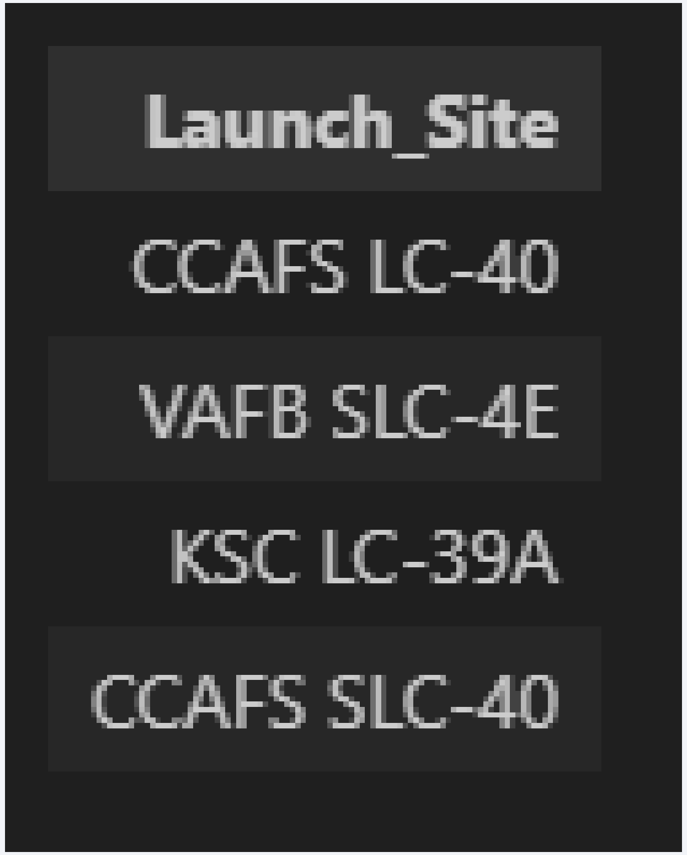
Launch Success Yearly Trend

- As years have gone on, success rates have increased in a linear fashion with a slight decrease in 2018



All Launch Site Names

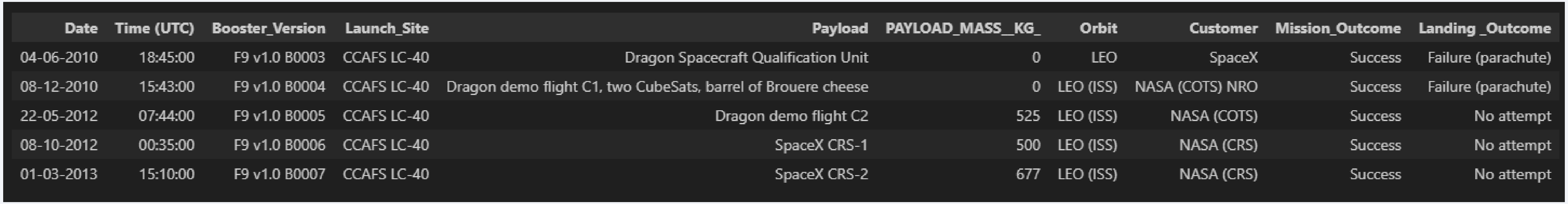
- The following are the launch names queried from database



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

Launch Site Names Begin with 'CCA'

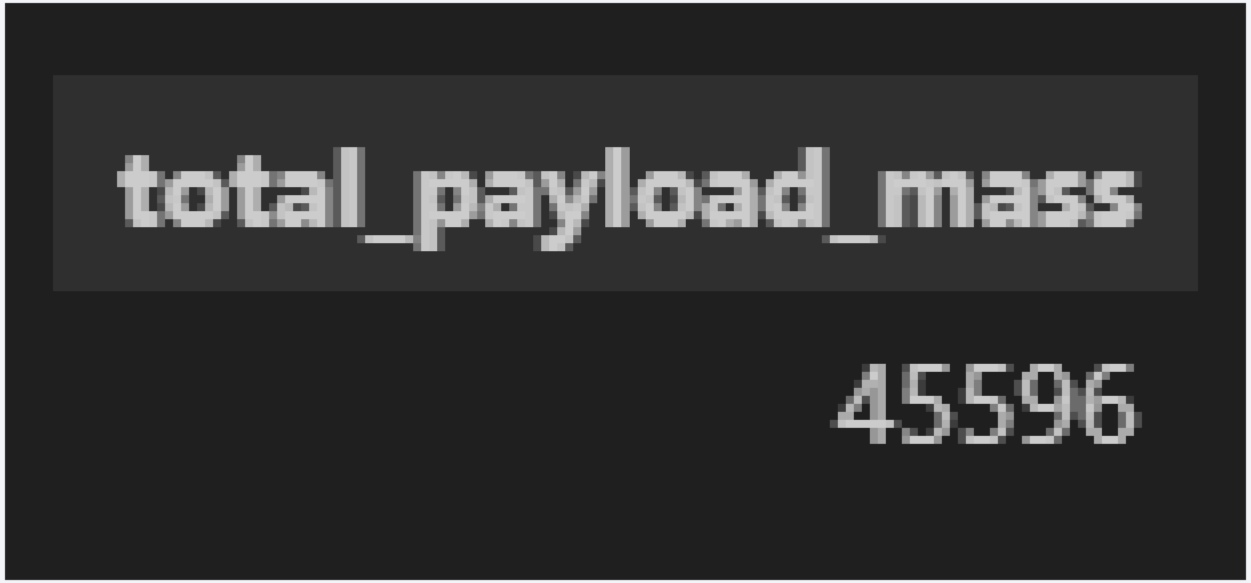
- Below are 5 records with launch site names that begin with CCA



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

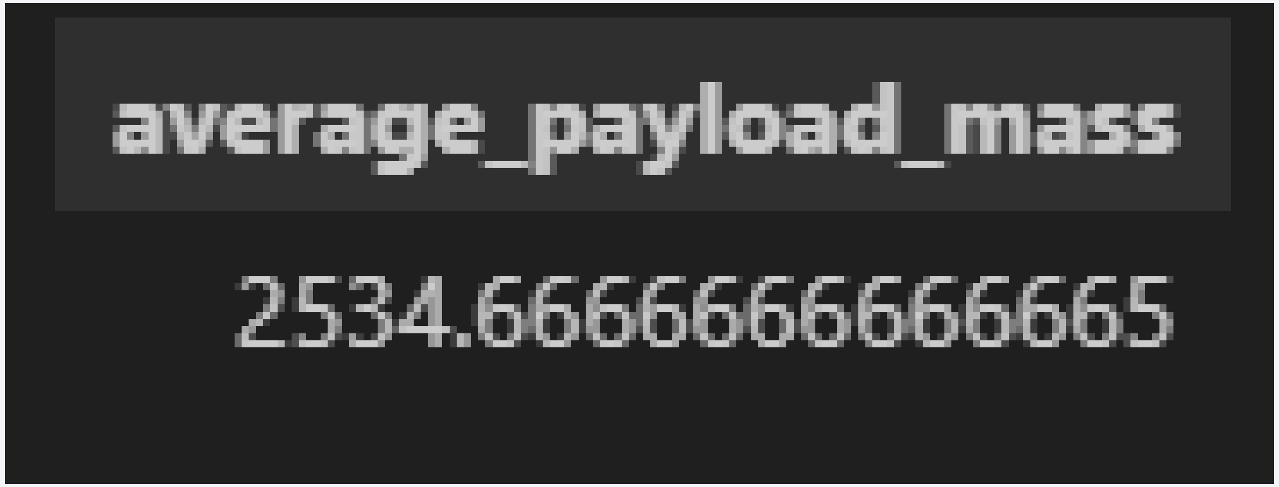
- The total payload mass in kilograms represents the total payload mass carried by bossters launched by NASA (CRS)



total_payload_mass
45596

Average Payload Mass by F9 v1.1

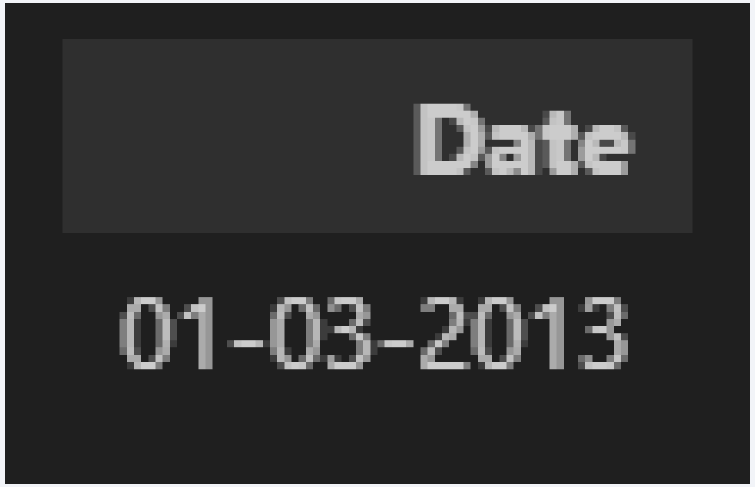
- The average payload mass in kilograms carried by booster version F9 v1.1



average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

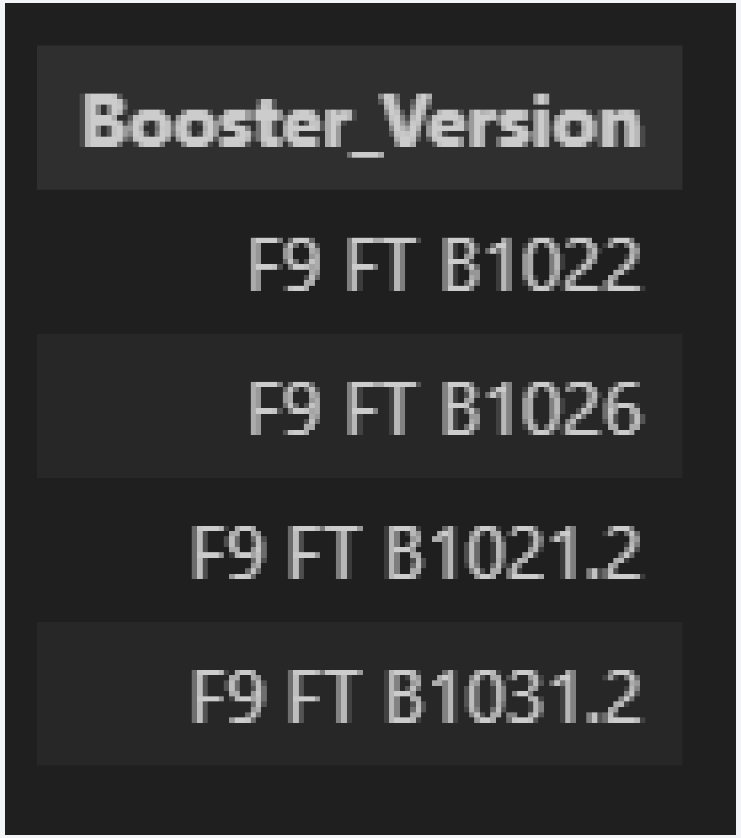
- Below is the date of the first successful landing outcome on ground pad



Date
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

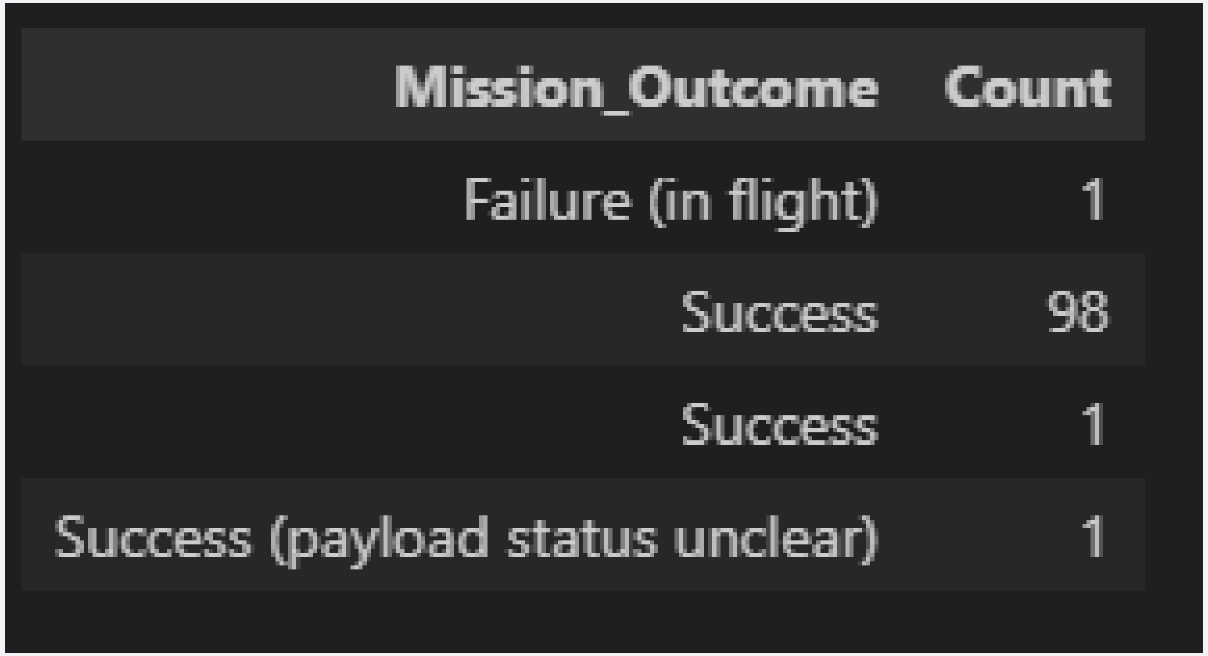
- These are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



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

Total Number of Successful and Failure Mission Outcomes

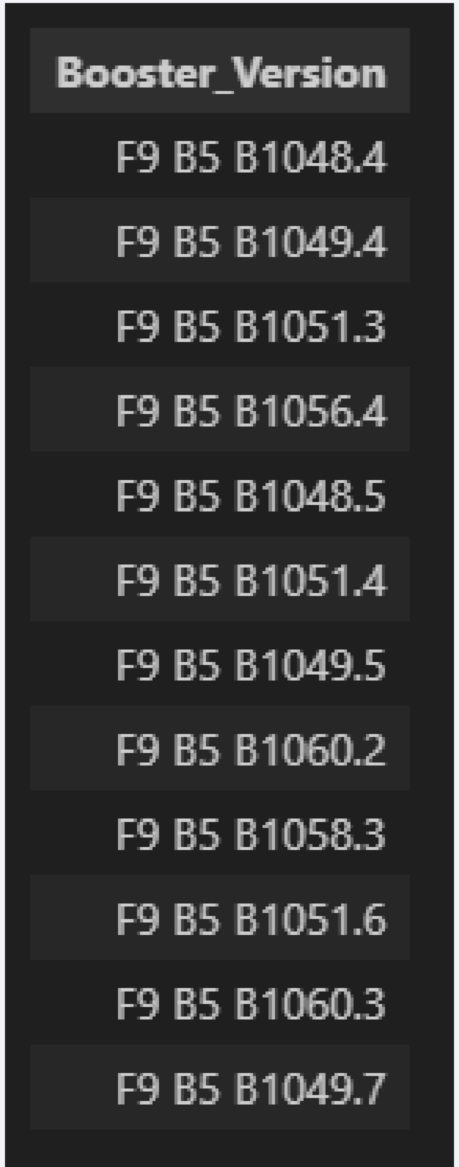
- Here are the total number of successful and failure mission outcomes



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

Boosters Carried Maximum Payload

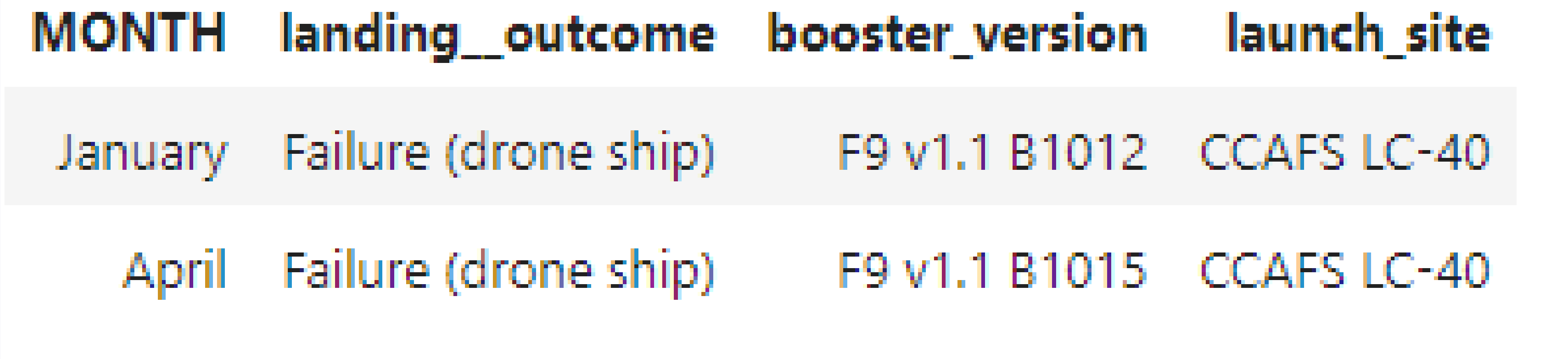
- These are the names of the booster which have carried the maximum payload mass



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

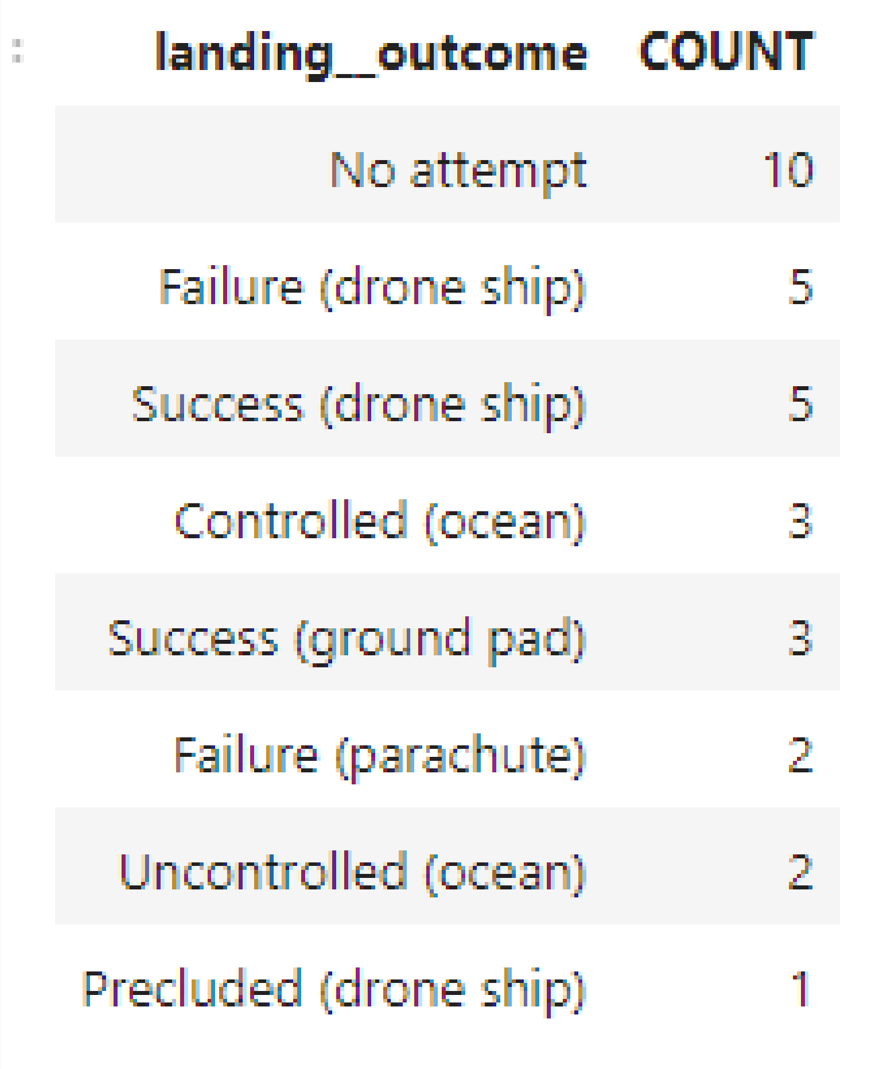
- The following are the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



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

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

- The following is the count 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

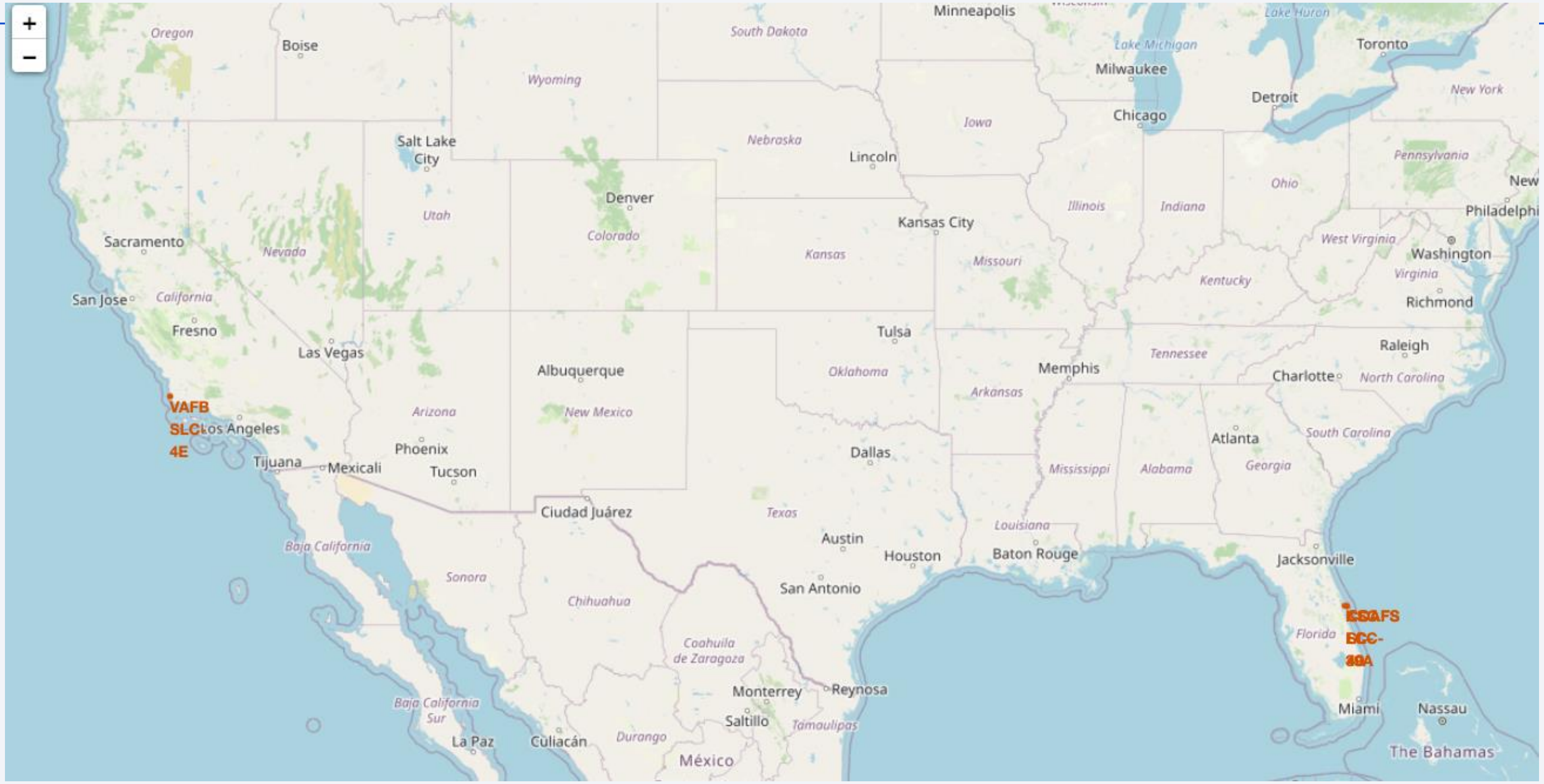


landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

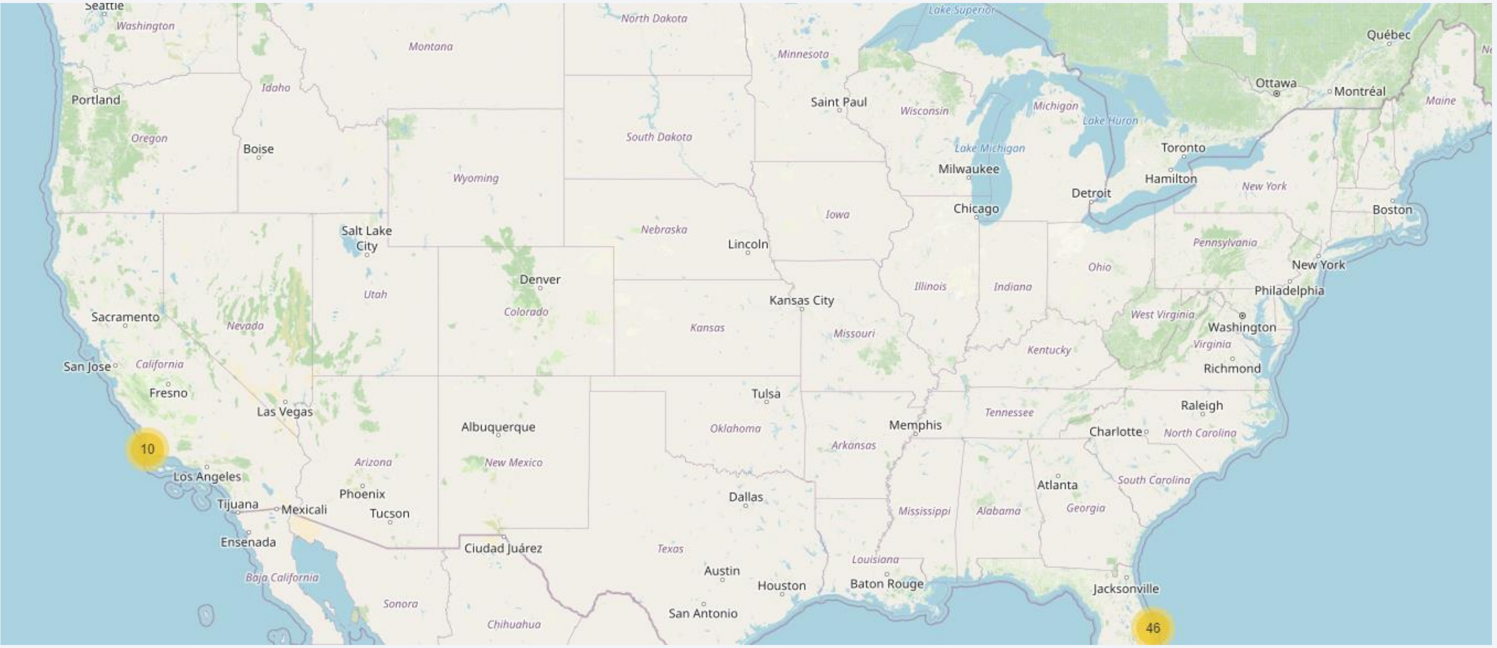
Locations of Launch Sites

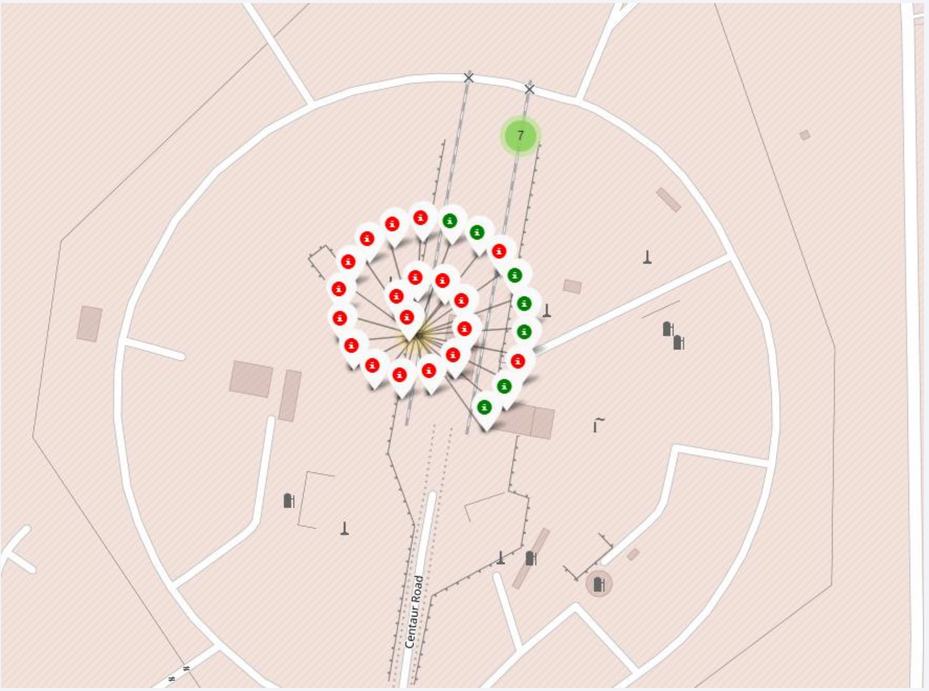


- SpaceX launch sites are exclusively located on the coastlines of the United States of America, more specifically in Florida and California.

Successful and Failed Launches for Each Site

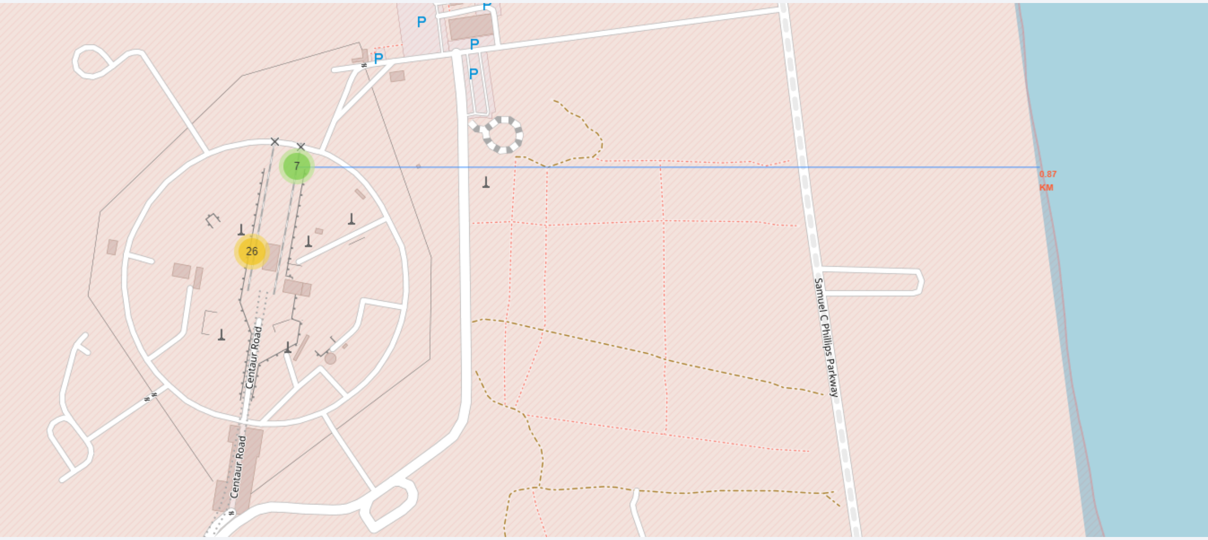
- The launches have been categorized into clusters and labeled with green icons for successful launches and red icons for failed launches.

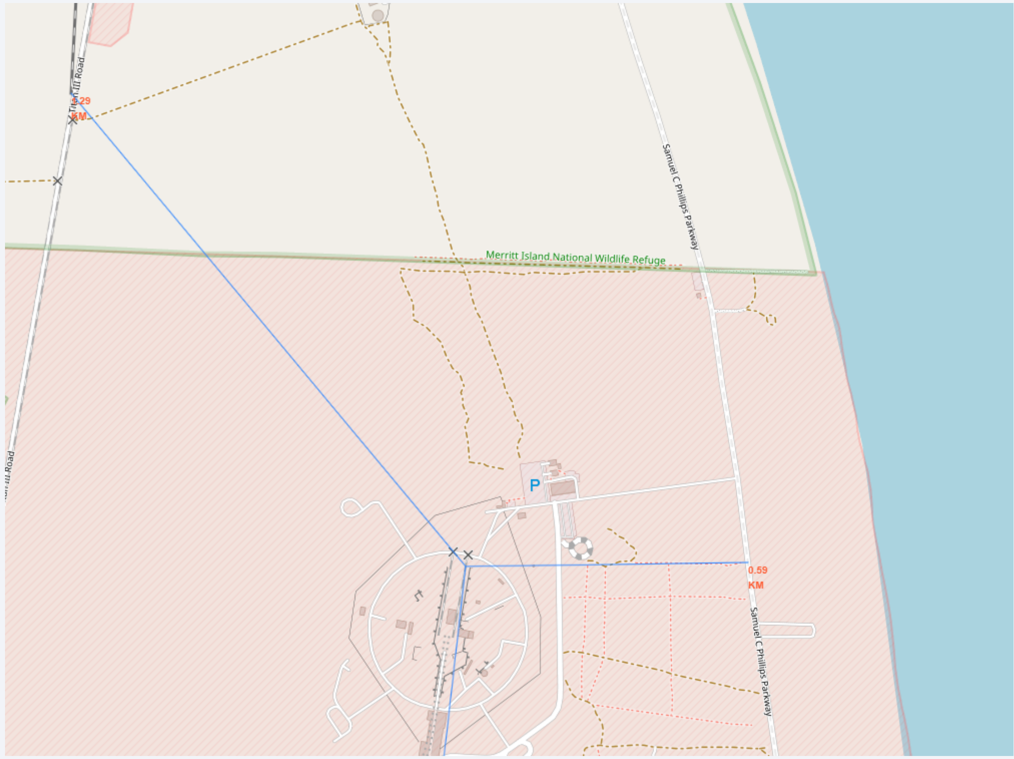


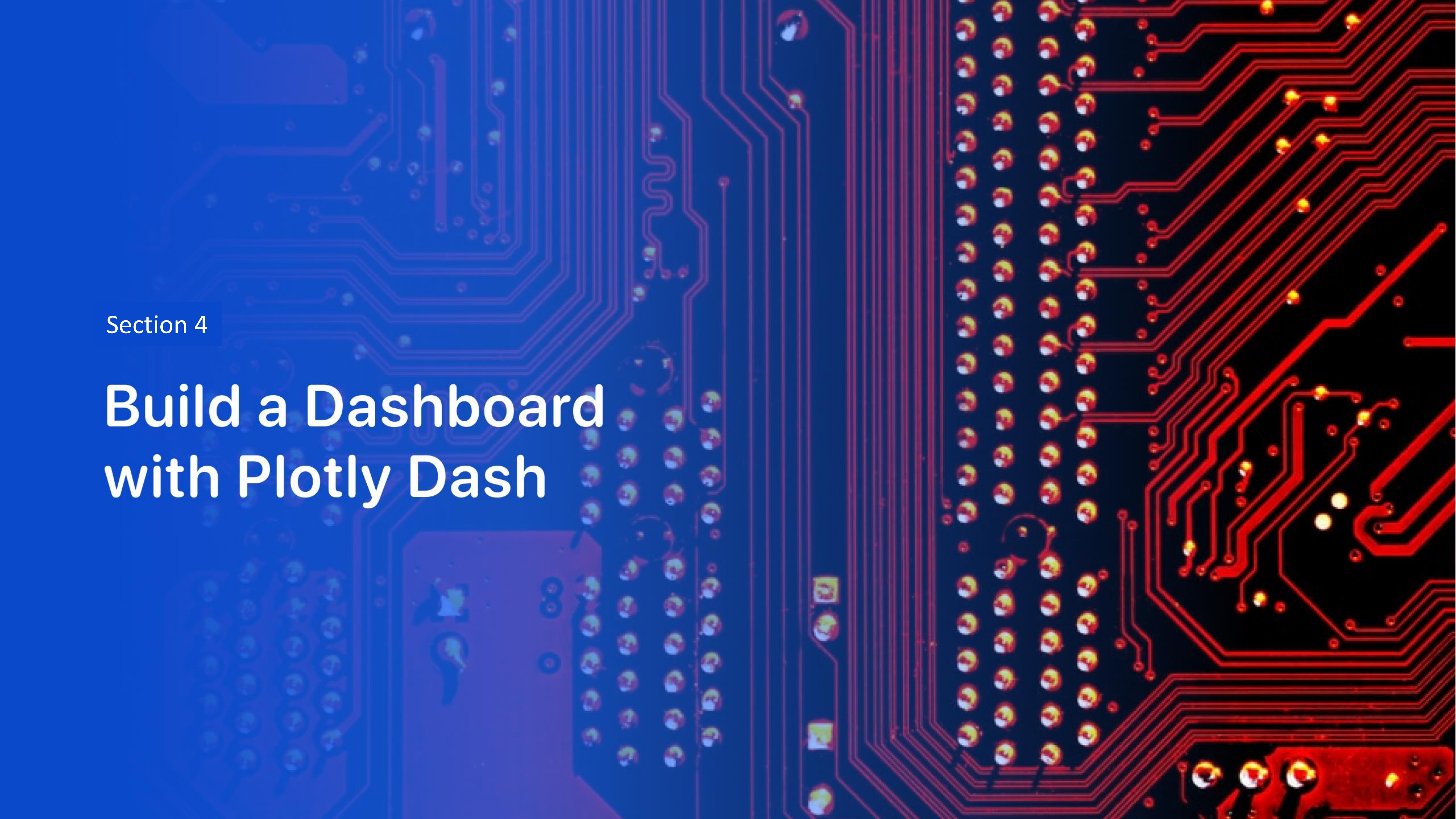


Proximity of Launch Sites to Notable Locations or Landmarks.

- Examining the CCAFS SLC-40 launch site as an example, enables us to gain deeper insight into the strategic placement and positioning of launch sites.



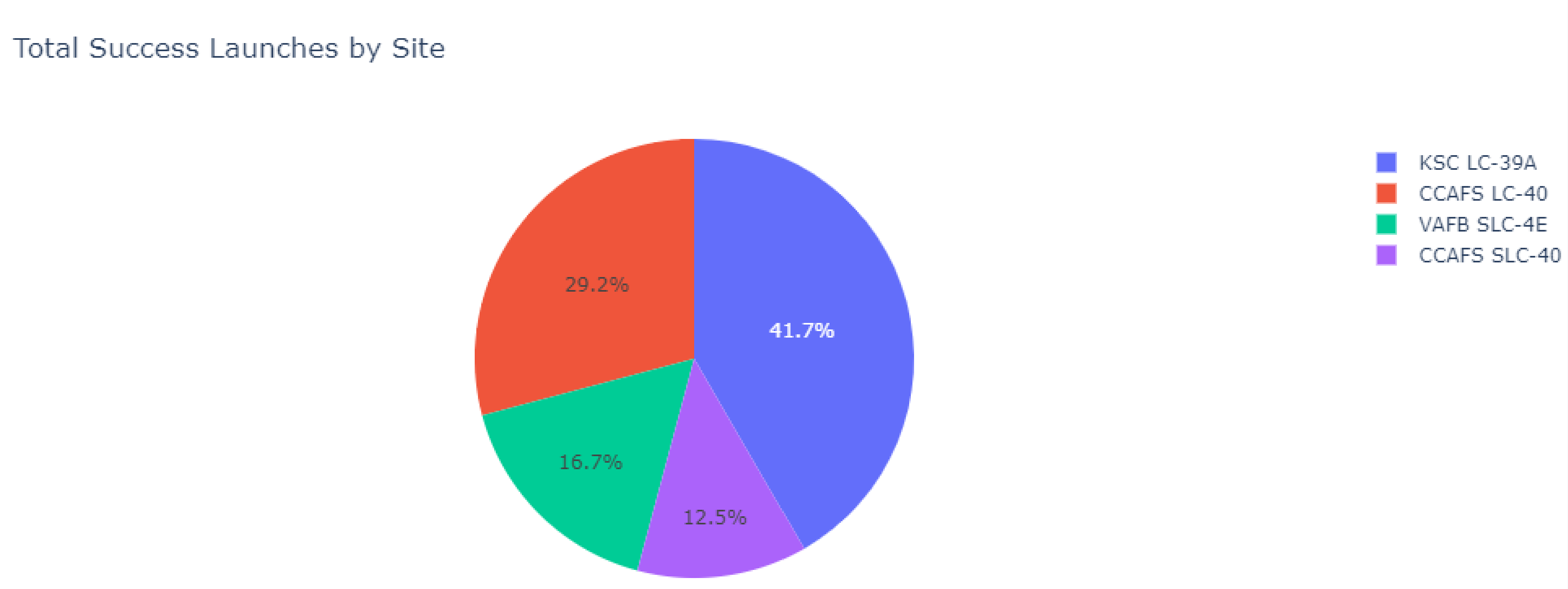

Titan III Road
1.29 KM
Merritt Island National Wildlife Refuge
Samuel C Phillips Parkway
P
0.59 KM



Section 4

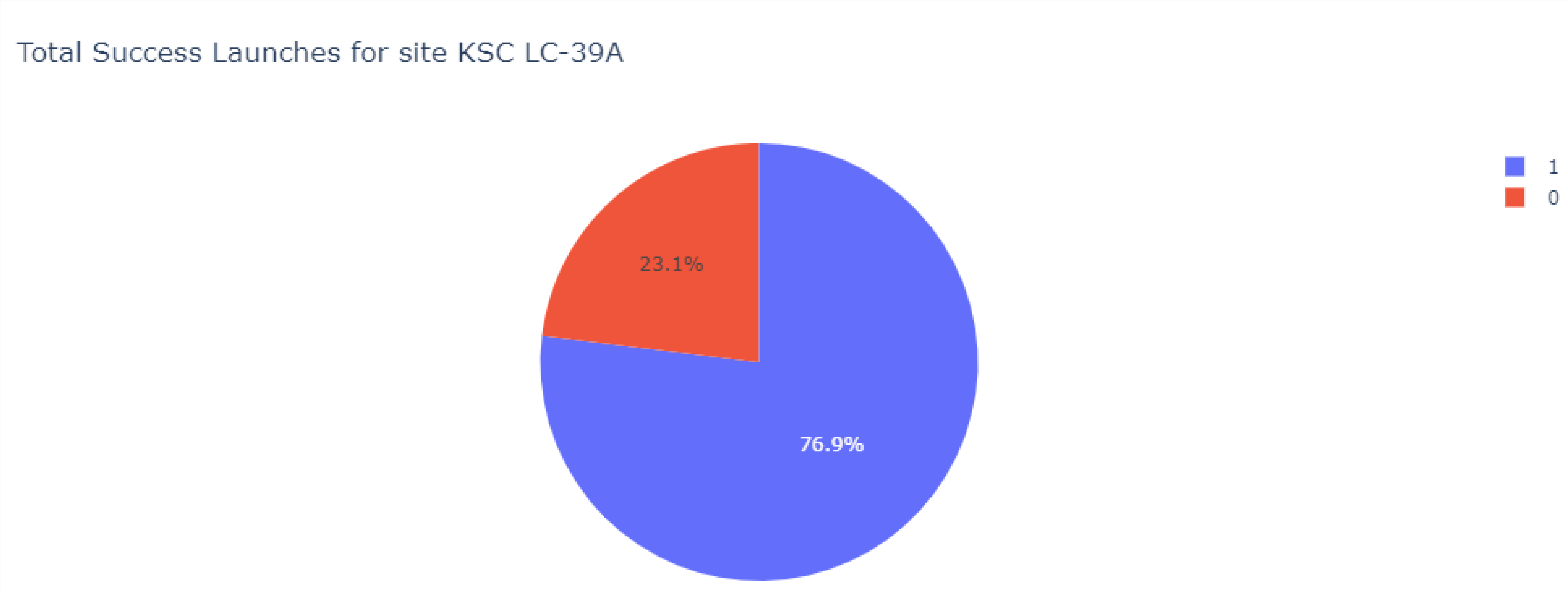
Build a Dashboard with Plotly Dash

Successful Launch Rates By Launch Sites



The distribution of successful landings across all launch sites indicates that CCAFS and KSC have an equal number of successful landings, as CCAFS LC-40 is the old name of CCAFS SLC-40. However, the majority of the successful landings occurred before the name change. On the other hand, VAFB has the lowest percentage of successful landings, which could be attributed to the smaller sample size and the increased difficulty of launching rockets on the West Coast.

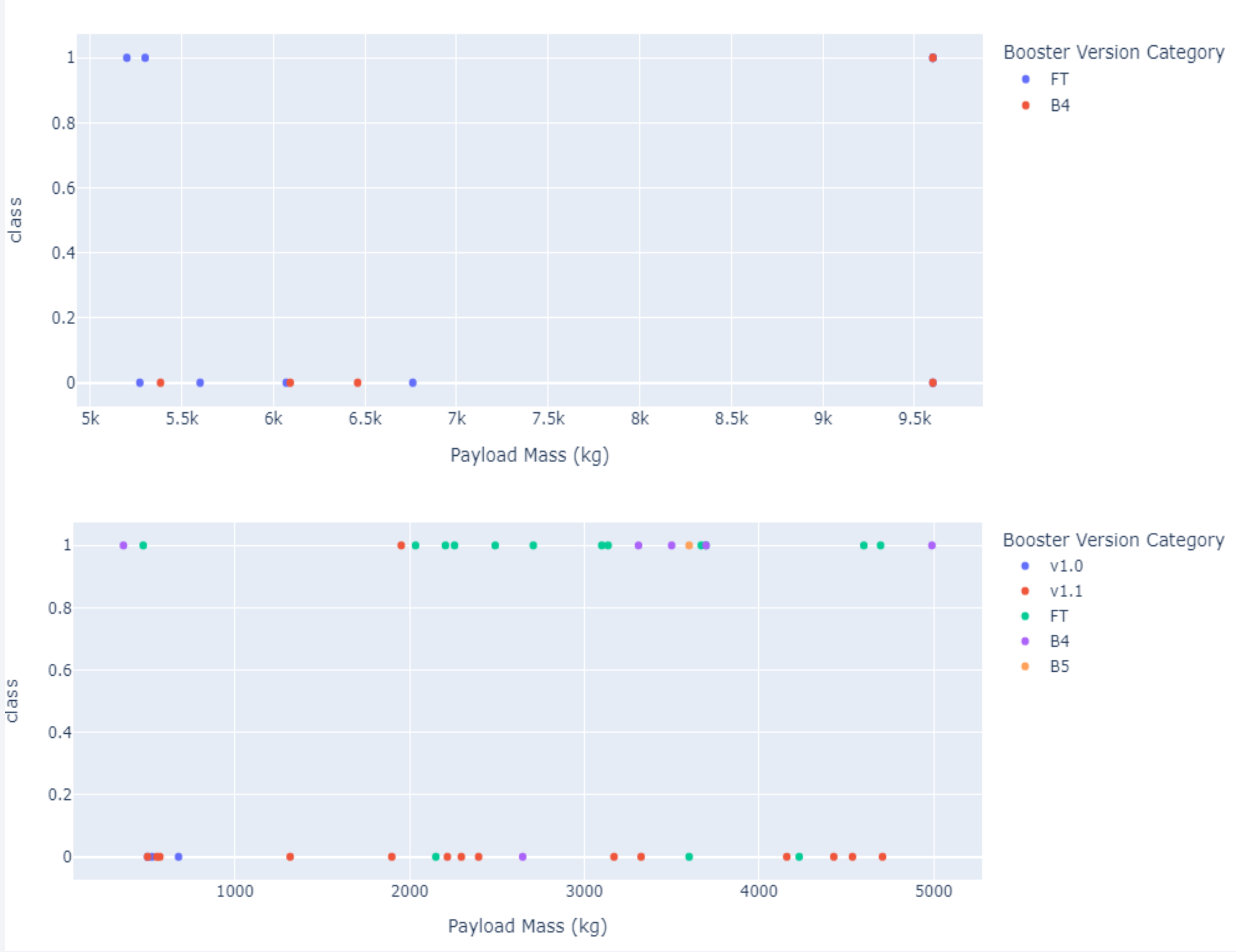
Launch Site with Highest Success Rate



Blue = Success; Red = Failure

- Out of all launch sites, KSC LC-39A has the highest success rate with ten successful landings and three failed landings.

Payload Mass Vs Success Vs Booster Version



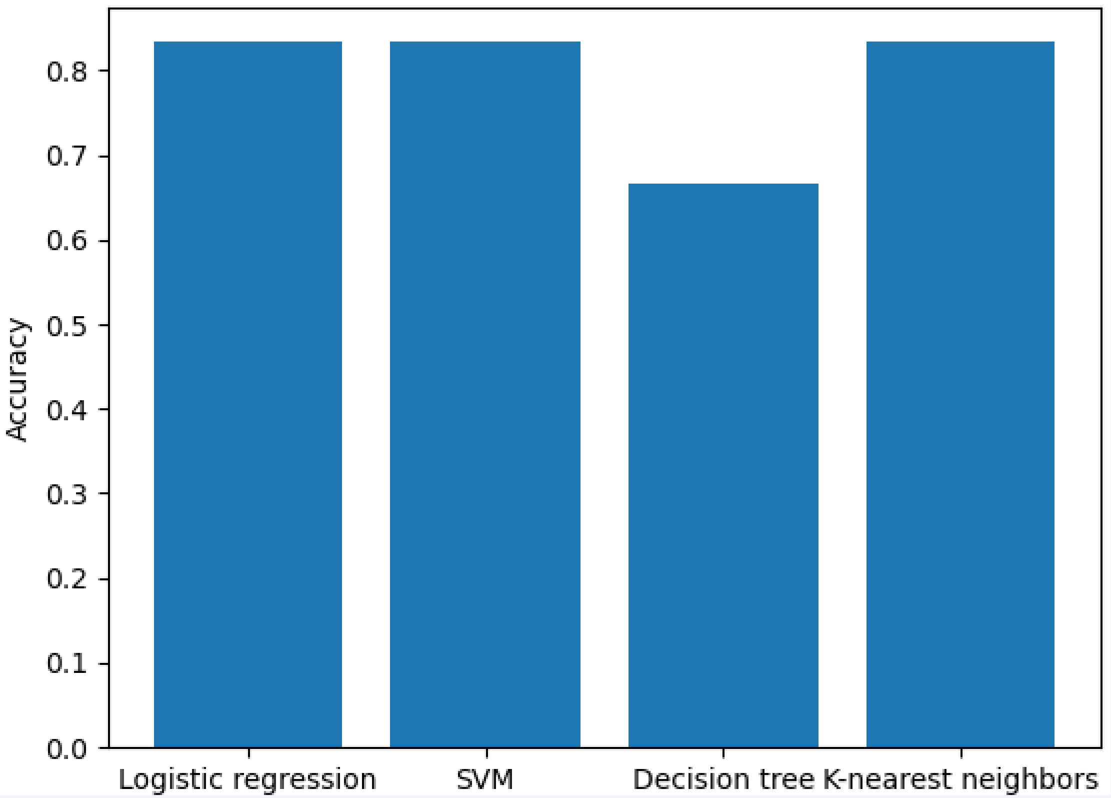
- Plotting launch outcome vs. payload for all sites reveals a gap around 4000 kg.
- Split data into two ranges:
  - 0 - 5000 kg (light payloads)
  - 5000 - 10000 kg (heavy payloads)
- Two plots indicate that success rate for massive payloads is lower than that for low payloads.
- Noteworthy observation: booster types v1.0 and B5 have not been launched with massive payloads.

Section 5

Predictive Analysis (Classification)

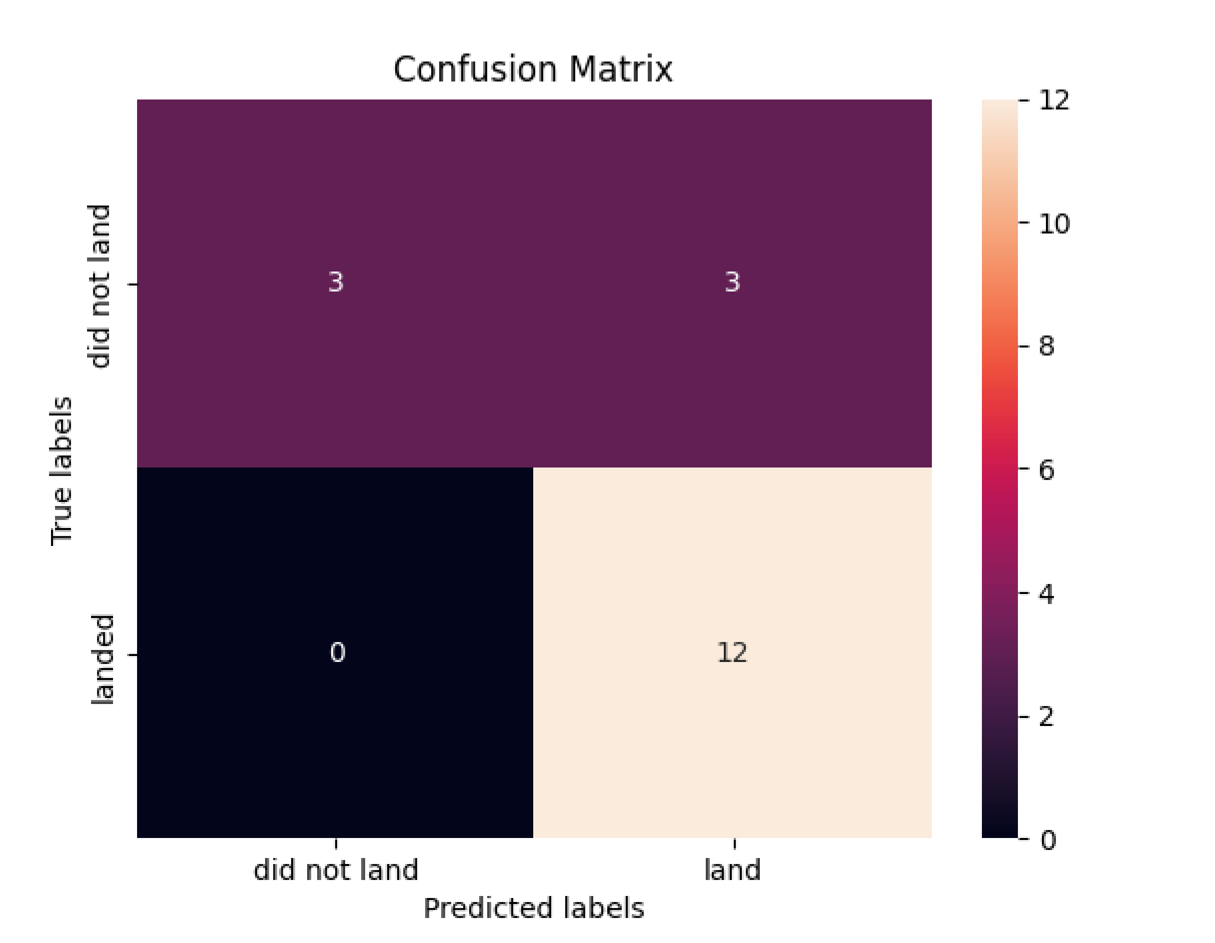
Classification Accuracy

- All models but Decision Tree classification model had an accuracy of 83%



Confusion Matrix

- The confusion matrix is identical across the 3 good models as they performed equally well on the test set.
- Out of the total successful landings, all 3 models predicted 12 correctly.
- In case of unsuccessful landings, the models predicted 3 correctly.
- The models falsely predicted 3 successful landings when the actual outcome was unsuccessful (false positives).
- The models have a tendency to over-predict successful landings.



Conclusions

- Objective: Develop a machine learning model for Space Y to compete with SpaceX by predicting successful Stage 1 landing and saving approximately $100 million USD.
- Data collection: Obtained data from a public SpaceX API and scraped information from SpaceX's Wikipedia page.
- Data labeling and storage: Collected data was labeled, stored in a DB2 SQL database, and visualized through a dashboard.
- Model accuracy: Achieved a satisfactory accuracy rate of 83%.
- Model application: SpaceY's CEO can use the model to predict the probability of a successful Stage 1 landing before launch and decide whether to proceed with the launch.
- Further improvements: Additional data should be collected to refine the model and improve its accuracy.

Thank you!