Winning Space Race with Data Science

Binaryxx Sune

01/27/2023

Outline

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

Executive Summary

- Payload Mass, Orbit Type, and Booster Versions were key elements in determining on whether the first stage rocket would land successfully
    - Another aspect to consider is that the more recent (such as future launches) will tend to have a higher success rate compared to launches with similar conditions during the first few launches by SpaceX
- The 3 models selected for predicting landing success rate, LogReg, SVM, and KNN, are able to determine whether launches would result in successful landing but with a slight problem with False Positives (predicting a successful landing when it will not). Regardless, all three models sport a 90% plus accuracy for out of sample data

Introduction

- SpaceX is able to conduct rocket launches at competitively low prices due to the fact that they are sometimes able to reuse the first stage rocket. The first stage rocket is the main rocket which sends the fairing and second stage rocket into orbit. The fairing contains the payload, and the second stage rocket helps guide the fairing once in orbit. The first stage of the rocket is significantly larger than the other two components, which results in bearing most of the costs due to the resources it requires.
- By determining the probability of a successful salvage of the first stage rocket, the costs can be calculated for a given launch. A successful salvage results in cost savings since the rocket can be reused for future launches rather than making a new first stage rocket.

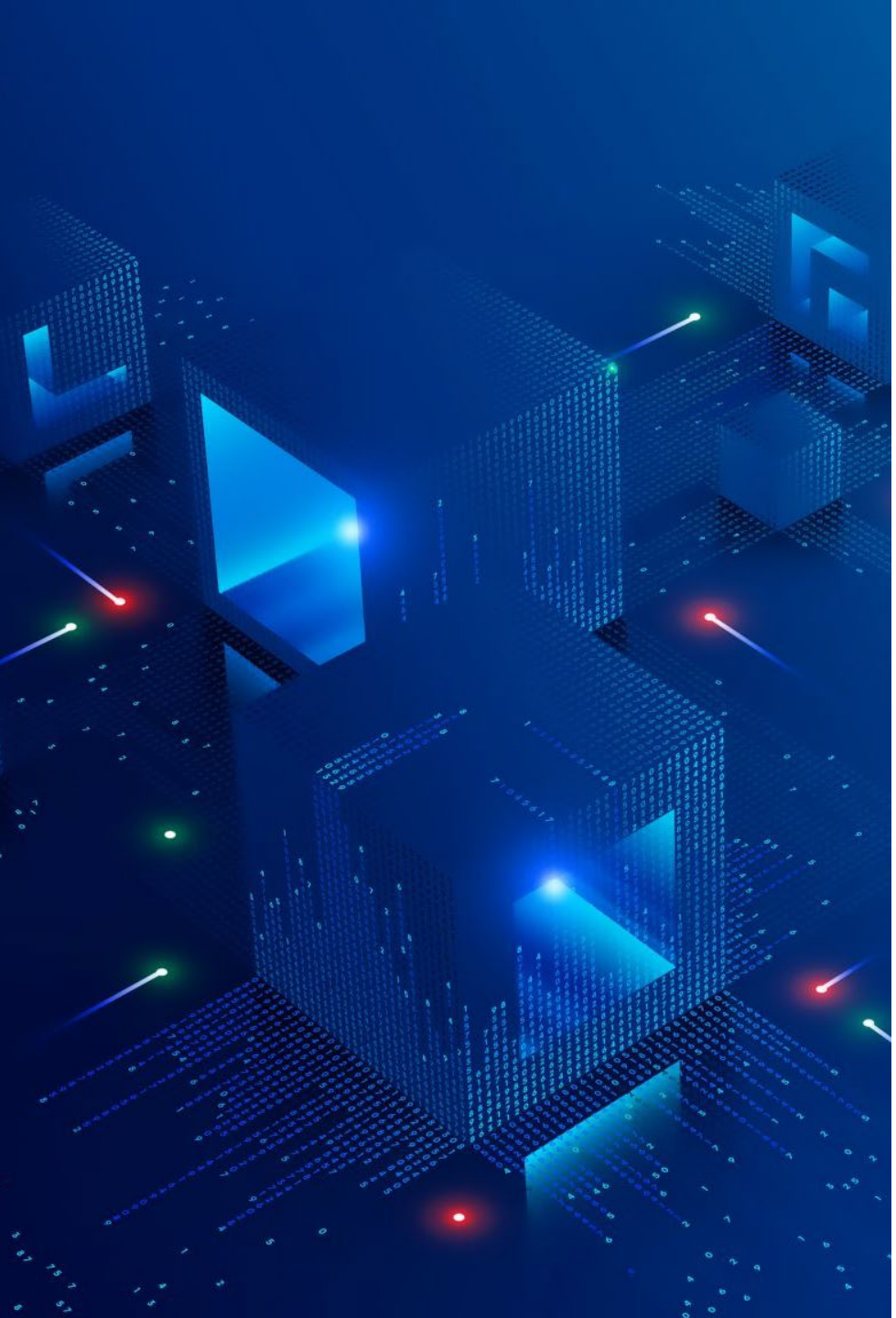
Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was gathered primarily from the SpaceX REST API, which provides data about launches, such as landing outcome.
  - Data was also gathered by web scraping wiki pages relating to the launches
- Perform data wrangling
  - Some data information from the API were IDs rather than the actual values, resulting in the use of the API again to relate those IDs to their respective values.
  - Only the relevant data needs to be kept, so proper filtering of the data set needs to be done so that the appropriate rocket boosters are selected (Falcon 9)
  - Null values for the rocket boosters are set to its mean value, while landing pad values were left null since that column was irrelevant
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built models based on optimal hyperparameters with the help of GridSearchCV to select the parameters which provided the highest in sample accuracy.
  - Evaluation of these models are done through a confusion matrix and it's out of sample prediction accuracy



Data Collection

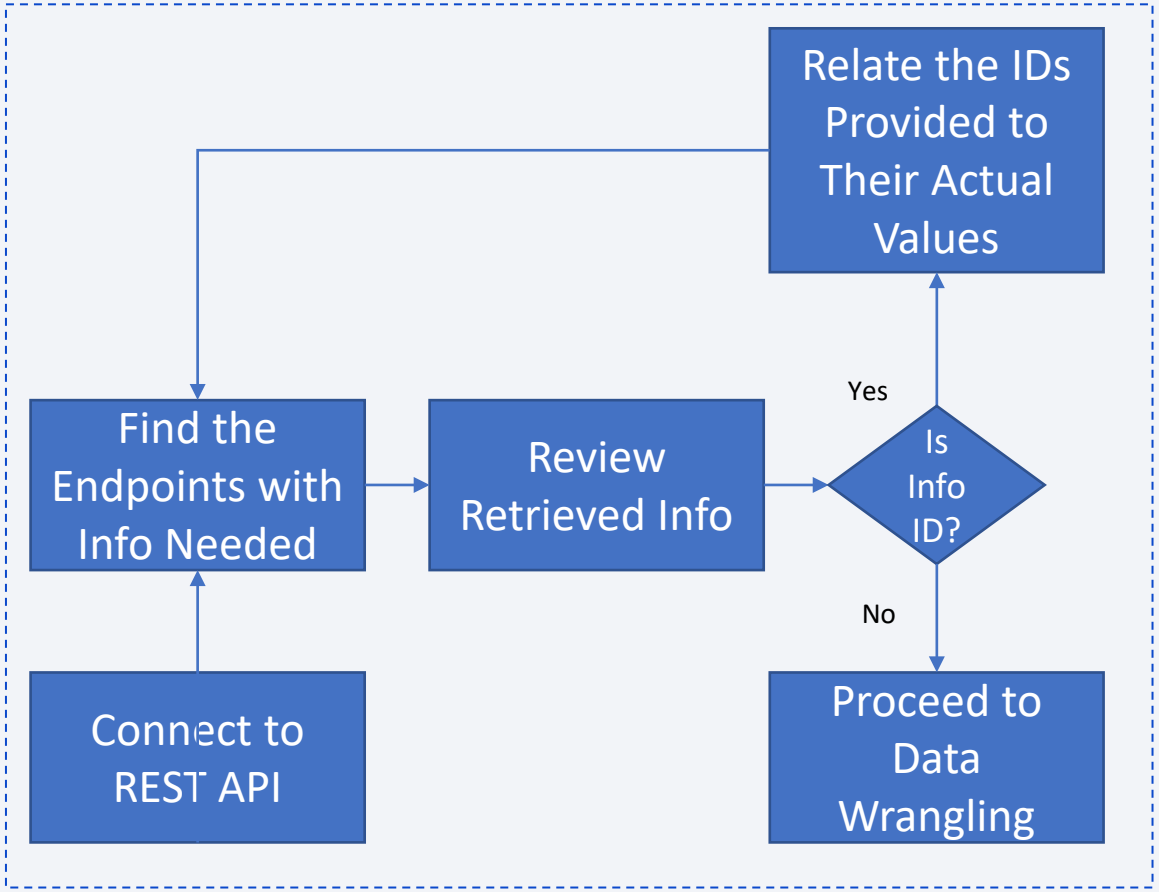
- Data was gathered primarily from the SpaceX REST API, which provides data about launches, such as landing outcome.
- Data was also gathered by web scraping wiki pages relating to the launches with the help of BeautifulSoup library; information regarding the launches similar to the info provided by the API was collected through this method

Data Collection – SpaceX API

Using the SpaceX API, relevant data was collected. There were instances where the values retrieved for certain columns were ID references, not inherently containing information. This would then require a search of the appropriate endpoints to relate the IDs to the corresponding values.

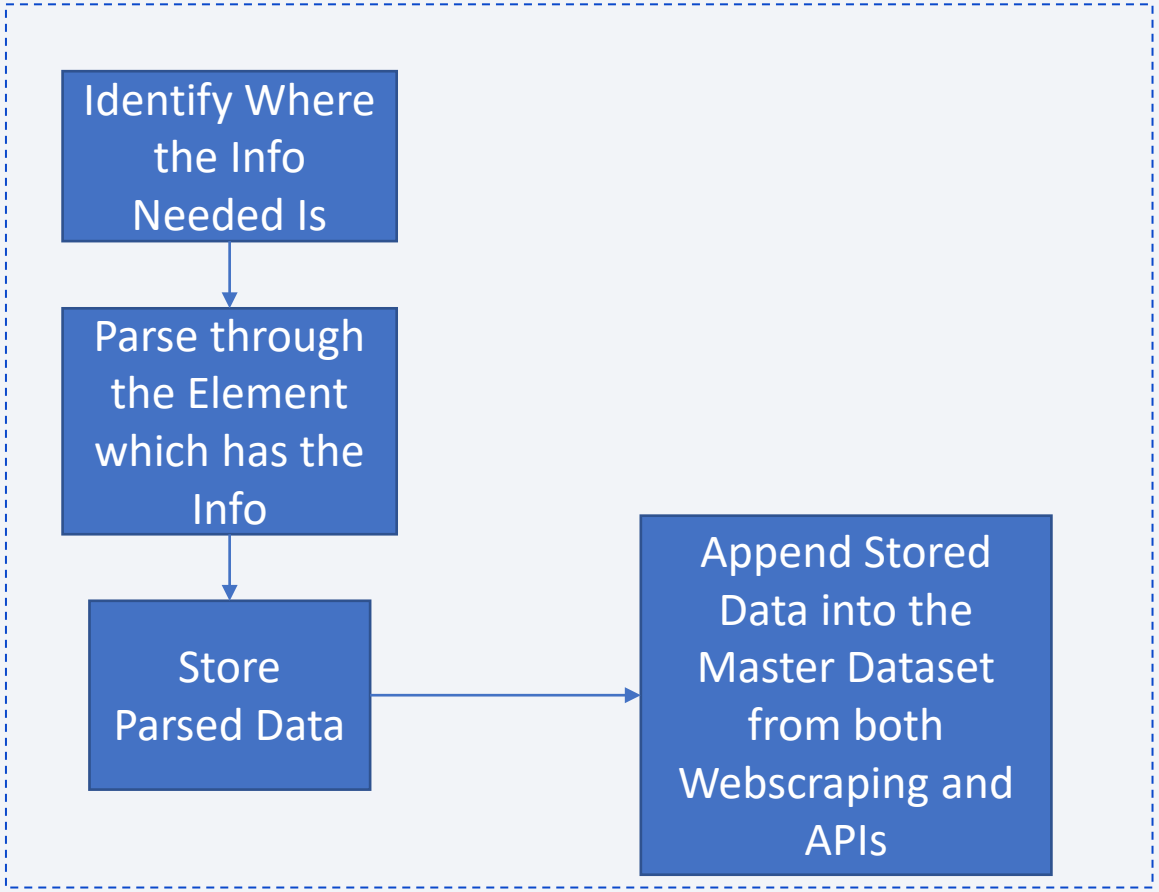
Below is a link to the notebook to conduct this process:

GitHub URL SpaceX API



Data Collection - Scraping

With the use of the Beautiful Soup library, information available in wikis relating to SpaceX launches were collected; the information collected is similar to what was provided in the API (successful retrieval of the first stage rocket, launch sites, etc.). The code parsed through a table which contained the useful information.



Below is a link to the notebook to conduct this process:

GitHub URL SpaceX Webscraping

Data Wrangling

- Some data information from the API were IDs rather than the actual values, resulting in the use of the API again to relate those IDs to their respective values.
- Only the relevant data needs to be kept, so proper filtering of the data set needs to be done so that the appropriate rocket boosters are selected (Falcon 9 Rockets Only)
- Null values for the rocket boosters are set to its mean value, while landing pad values were left null since that column was irrelevant

Below is a link to the notebook to conduct this process:

GitHub URL Data Wrangling

EDA with Data Visualization

11

- A scatterplot was used to visualize any correlation between the following feature pairs and the first stage rocket landing success rate:
    - Flight Number and Launch Site Location
    - Payload Mass and Launch Site Location
    - Flight Number and Orbit Type
    - Payload and Orbit Type
        - Flight Number effectively represents when the launch was made; higher flight numbers are the most recent ones while lower flight numbers represent the earliest launches
- A category plot was used to visualize any correlation between the following feature pairs and the first stage rocket landing success rate:
    - Flight Number and Payload Mass
    - Flight Number and Launch Site
- A bar graph was used to compare the success rate of the first stage rocket landing of each Orbit Type
- A line graph was used to identify trends of success rate over time

Below is a link to the notebook to conduct this process:

GitHub URL Data Visualization

EDA with SQL

- Identified Launch Sites
- Identified the statistics about the Payload Mass, such as max value
- Identified successes in relation to time
- Identified number of successful landings for each type of successful landing

Below is a link to the notebook to conduct this process:

GitHub URL SQL EDA

Build an Interactive Map with Folium

- Create a circle which captured the 4 primary areas of launches to indicate where launch sites can be
- Marked the launch site and its landing outcome to see the relation between successes and location
- Identified major geographical elements such as cities, major roads, rails, and coasts to identify minimum distance pattern of launch areas

Below is a link to the notebook to conduct this process:

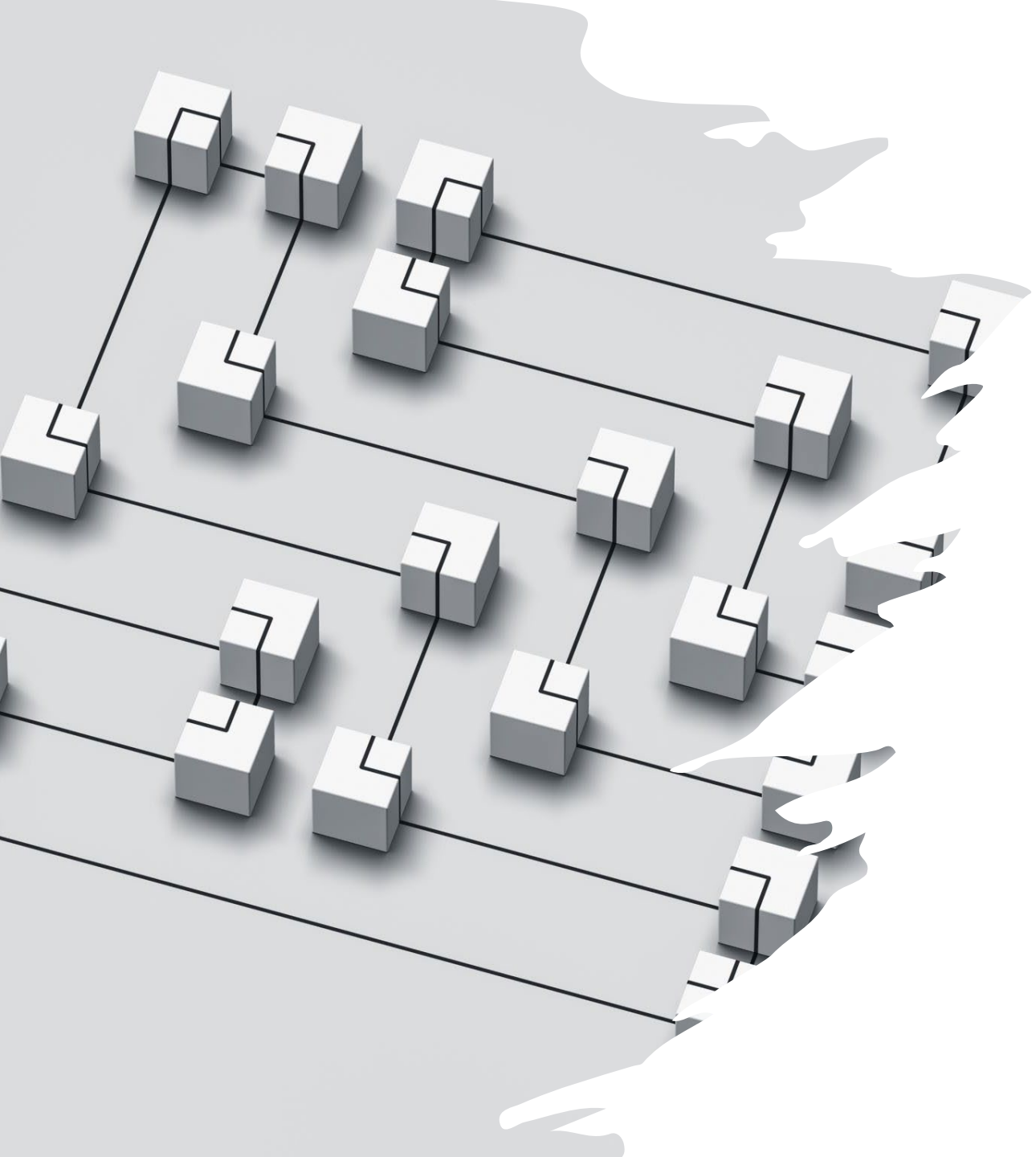
GitHub URL Map

Build a Dashboard with Plotly Dash

- The dashboard had a dropdown for different launch sites (or all launch sites) to examine the different performances of each one in regards to successful first stage rocket landings
- For each selected configuration of view, a pie chart displayed the percent of successful landings to total landings (or the percent that each site contributed to successful landings in the case of all launch sites view)
- Below the pie chart was a scatterplot which displayed the relationship between payload mass and booster version in regards to successful first stage rocket landings
    - The scatterplot can be configured with a slider to determine the range of payload mass values

Below is a link to the python file to conduct this process:

GitHub URL Dashboard



Predictive Analysis (Classification)

- Built models based on optimal hyperparameters with the help of GridSearchCV to select the parameters which provided the highest in sample accuracy.
- The following models were used and compared with each other:
    - Logistic Regression
    - K Nearest Neighbors
    - Decision Trees
    - Support Vector Machines
- Evaluation of these models are done through a confusion matrix and it's out of sample prediction accuracy
    - Every model except for the Decision Trees Model were very successful (equally high success) in making an accurate prediction

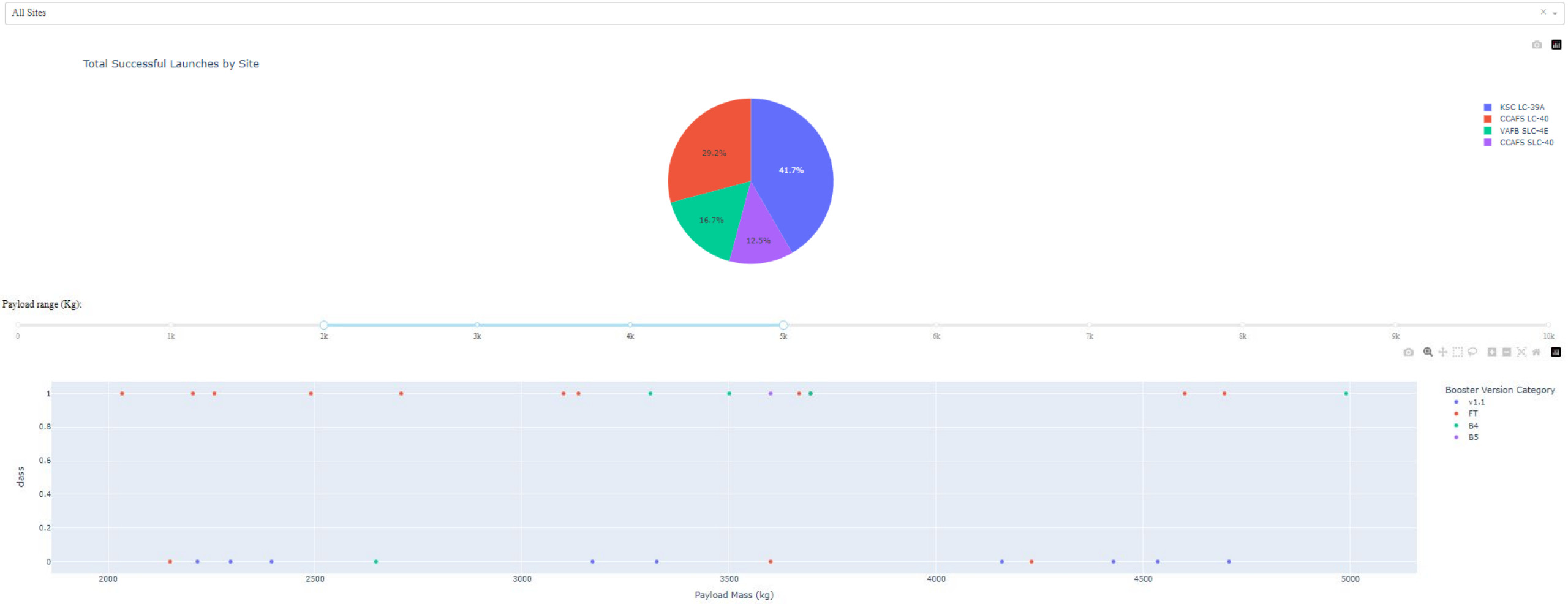
Below is a link to the python file to conduct this process:

GitHub URL Dashboard

Results

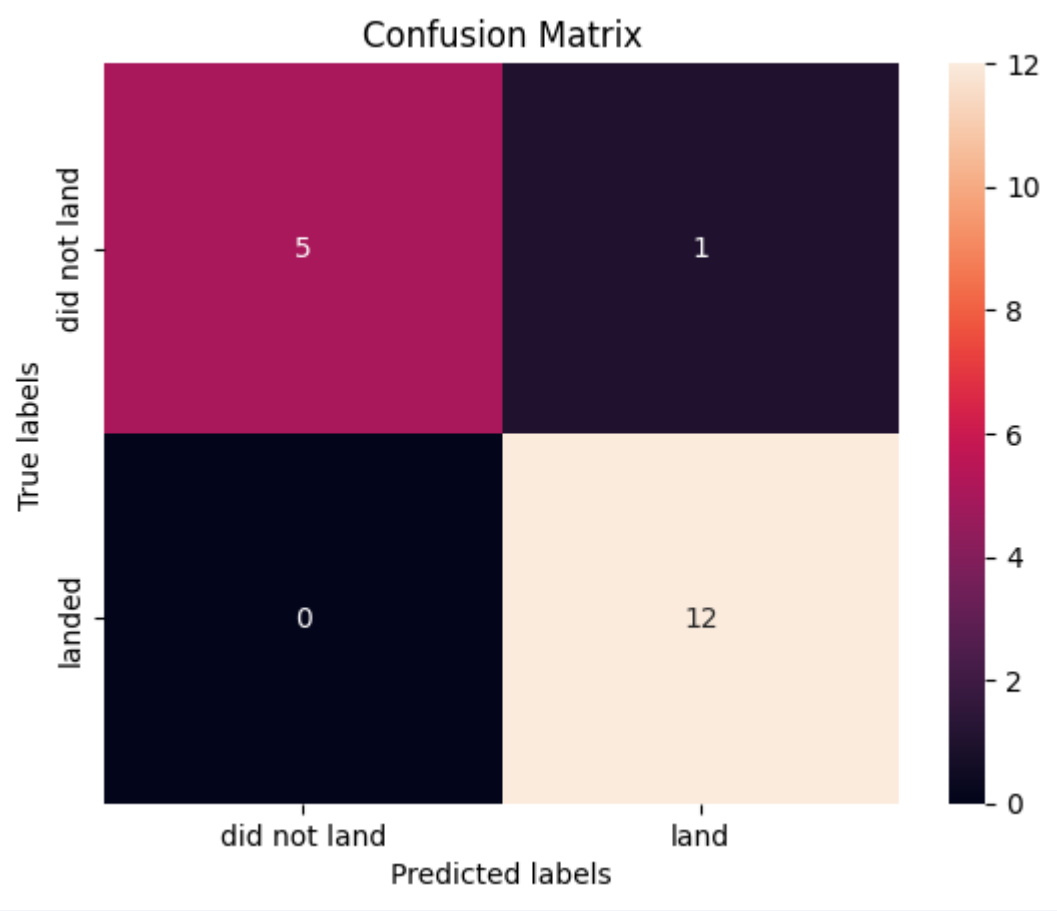
- Through the exploratory data analysis, it was observed that:
  - As the flight number increased, the first stage was more likely to land successfully; this would mean that more recent and future launches are more likely to land successfully
    - This was confirmed and supported by analyzing the success rate in relation to time
  - As payload mass increased, the less likely the first stage will return
    - The opposite holds true for launches for Polar, LEO, and ISS orbits
  - Certain launch sites, such as VAFB SLC 4E has a higher success rate than the rest
    - Upon further analysis, it was noted that this launch site never launched rockets with payload mass greater than 10000

Dashboard Screenshots

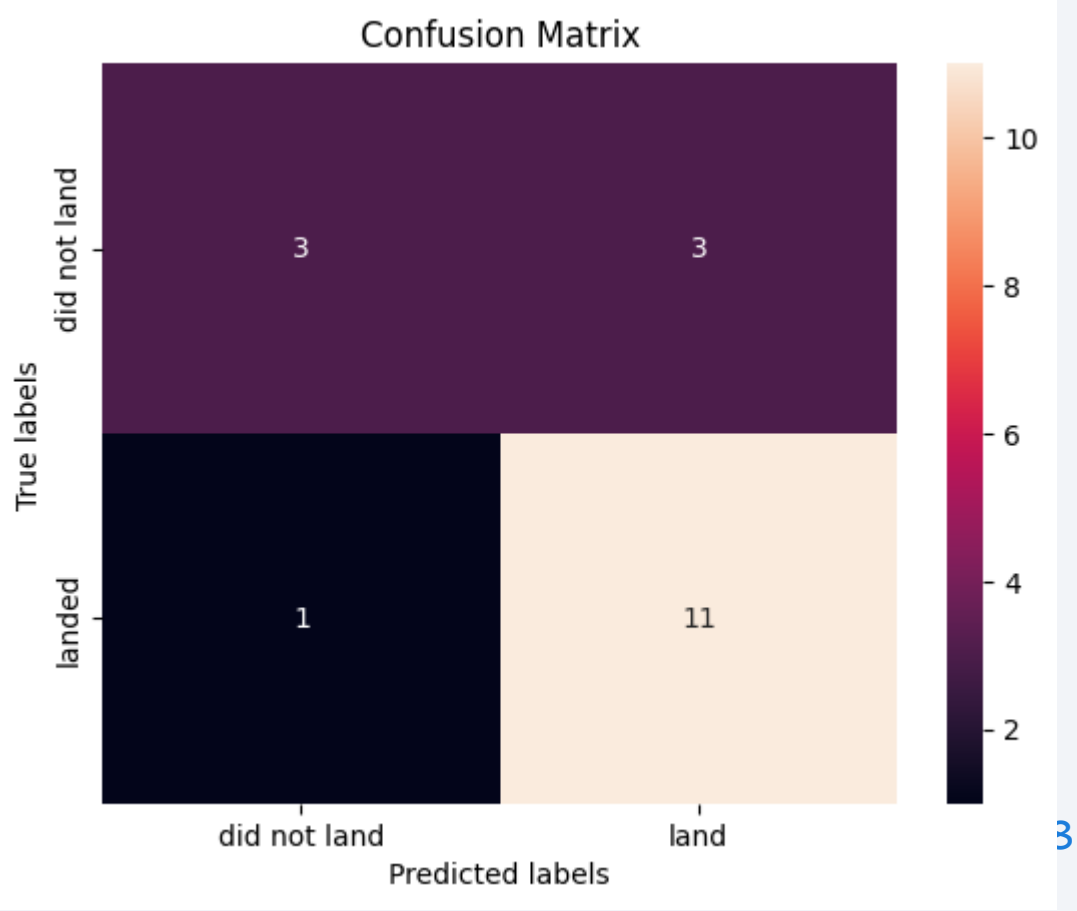


Results

SVM, KNN, LogReg



Decision Tree

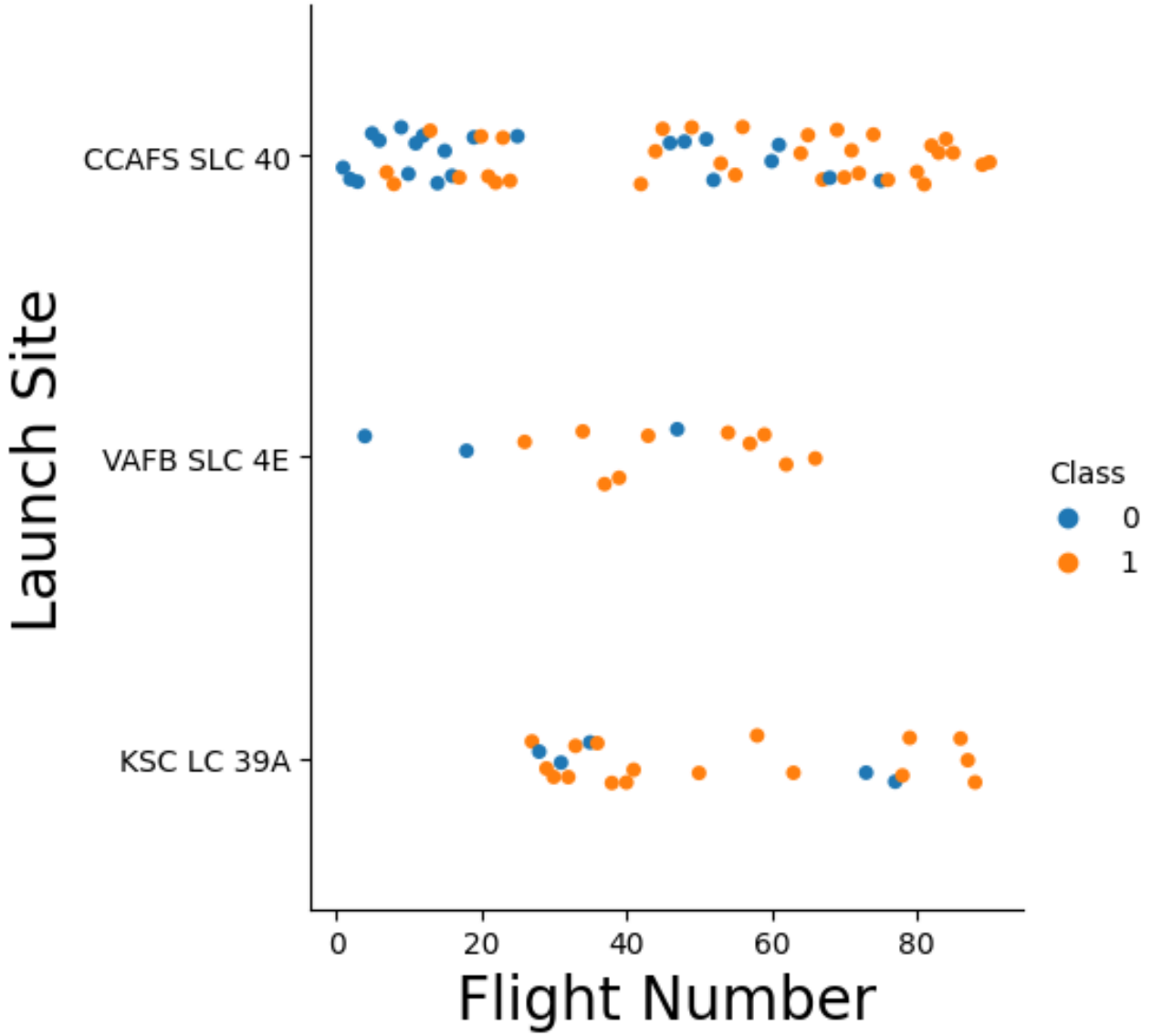


Section 2

Insights drawn from EDA

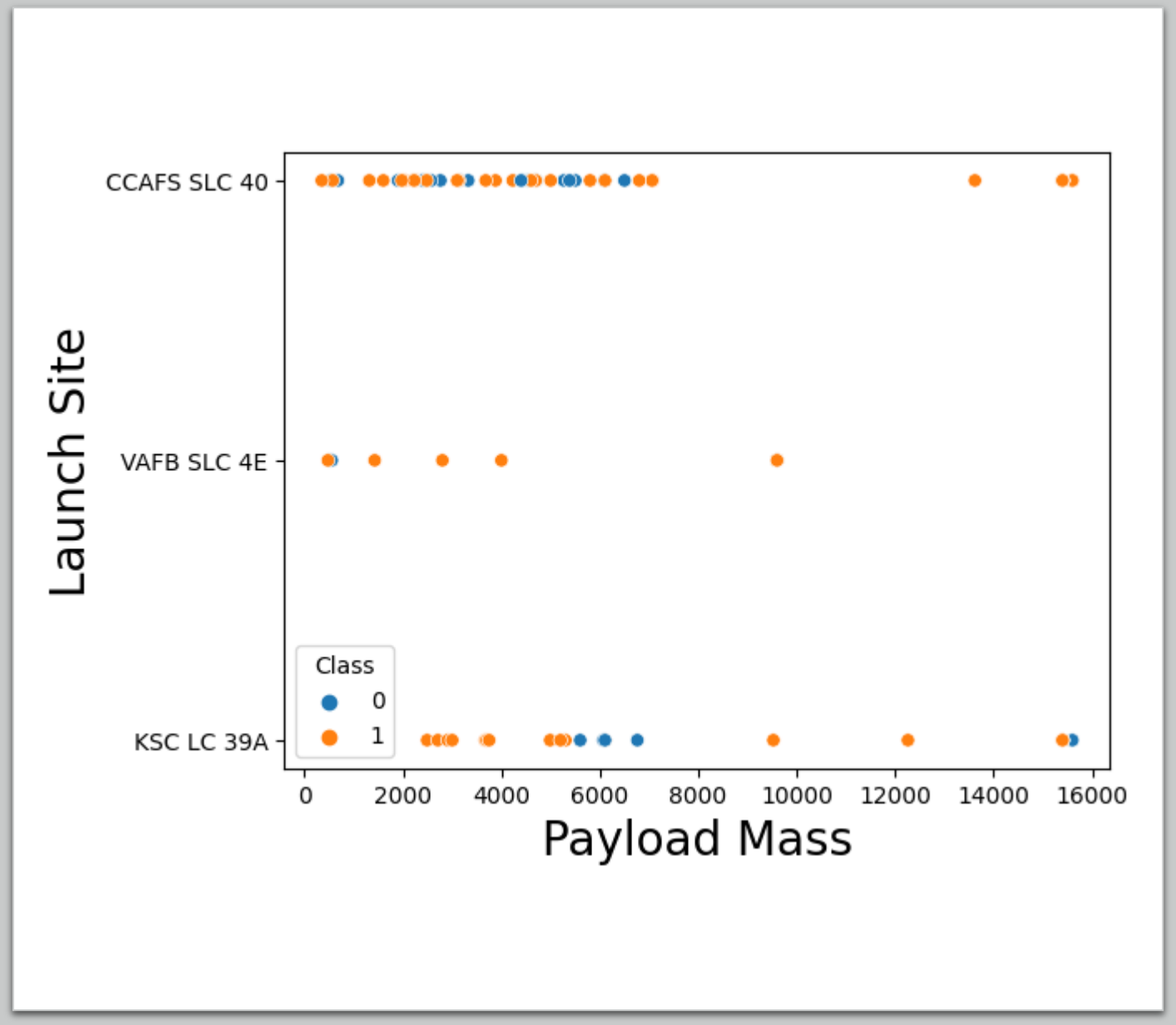
Flight Number vs. Launch Site

- The scatterplot demonstrates that SLC 40 had more successful landings through its early iterations; other launch sites did not display any noticeable pattern



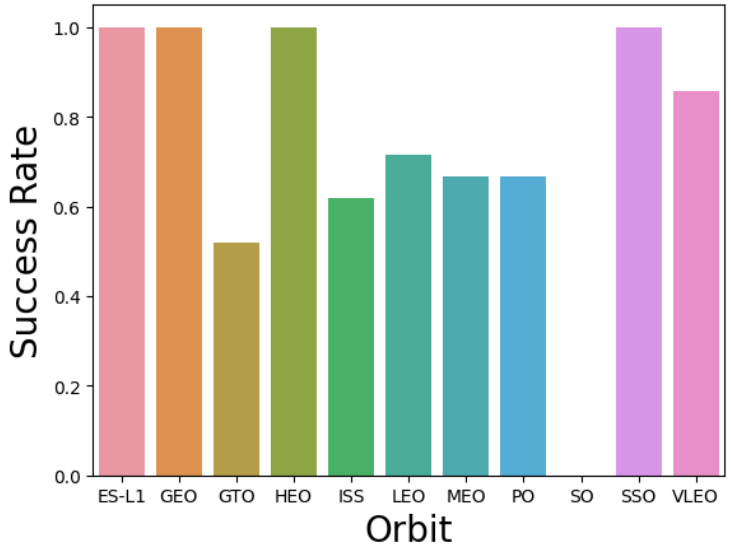
Payload vs. Launch Site

- A general pattern of most launches having mass lower than 8000kg



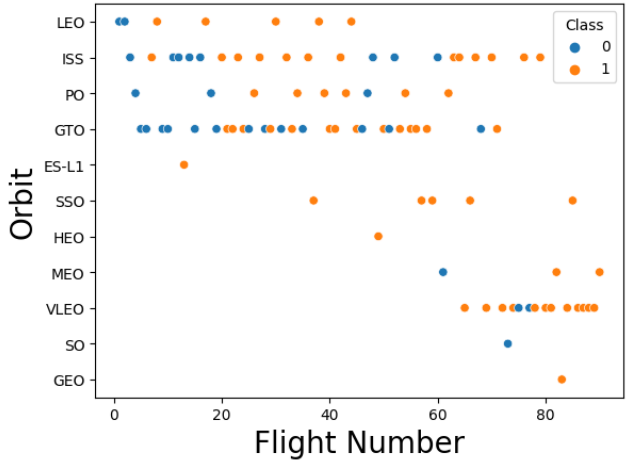
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations



Flight Number vs. Orbit Type

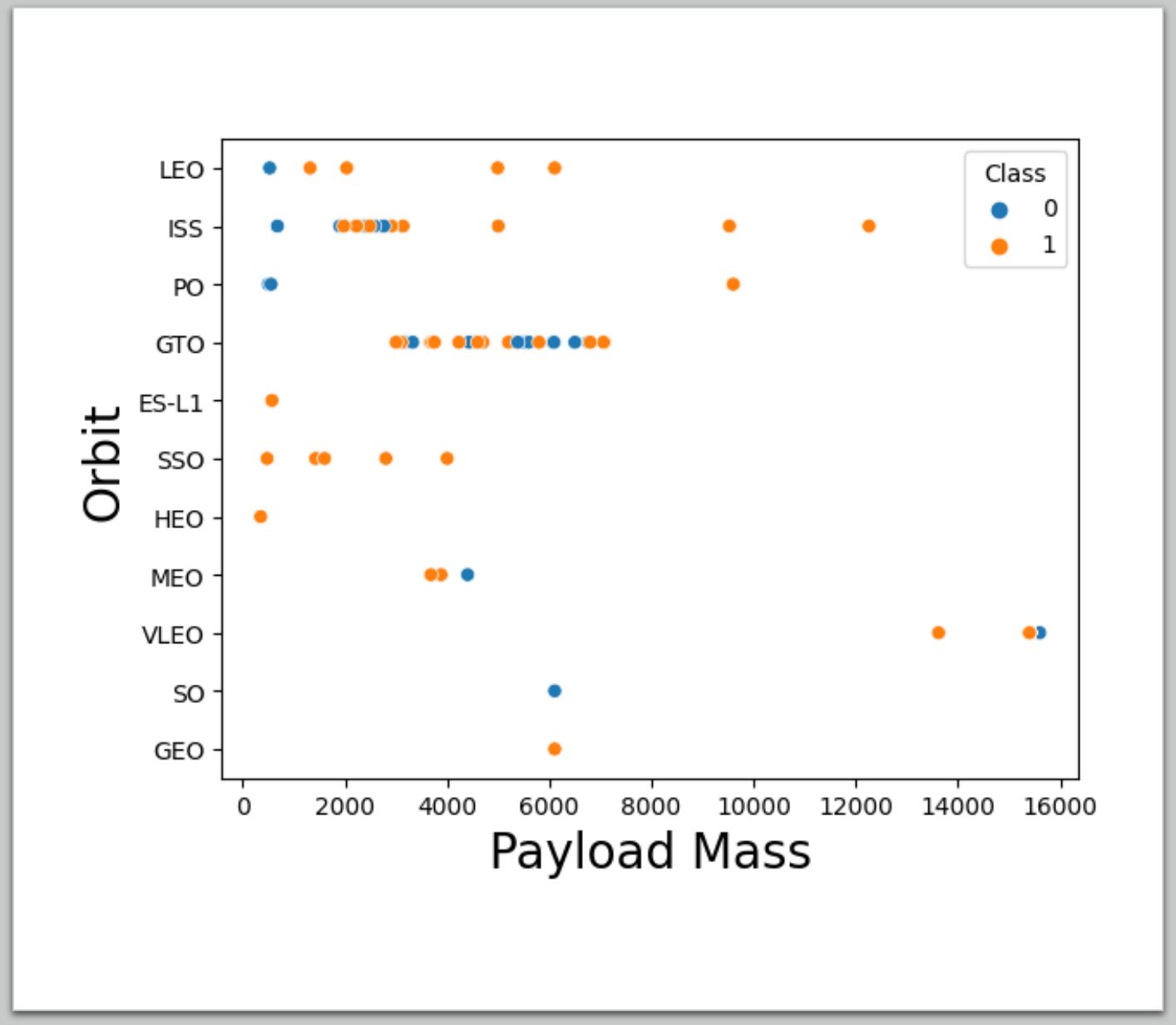
- Certain orbits have higher number of successful lands such as VLEO and ISS, while others have more failures such as GTO and ISS
- Certain launches to specific orbits were only available in later launches, possibly contributing to its higher success rate



Payload vs. Orbit Type

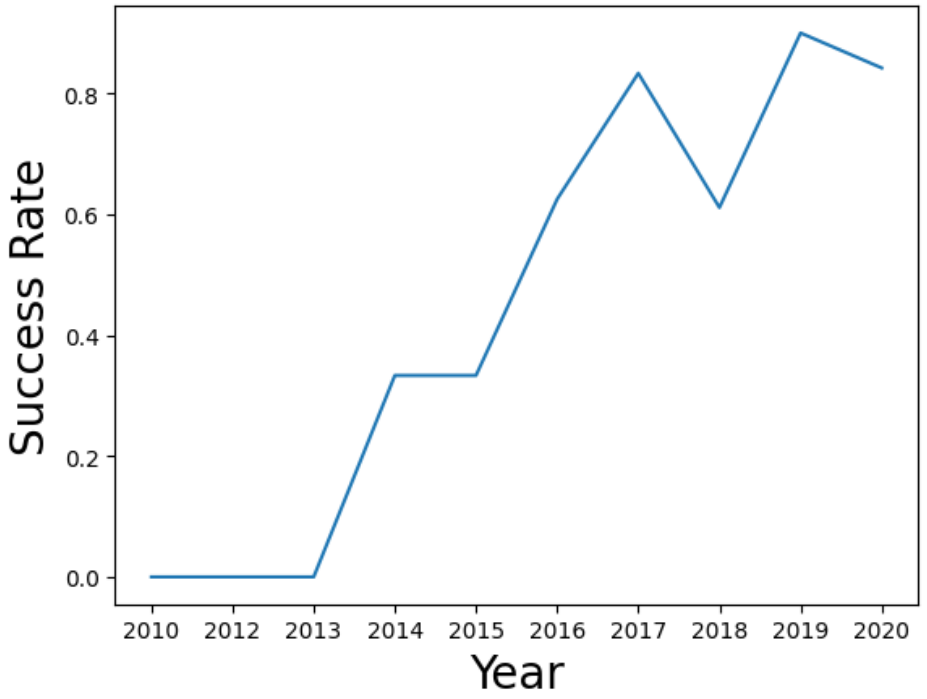
Higher payload mass related to higher success rate except for the following orbits:

- GTO
- VLEO



Launch Success Yearly Trend

- It can be observed that as time went on, the success rate of landings tended to increase, showing signs of plateauing at around 80%



All Launch Site Names

- There are 4 Launch Sites from the acquired data sets from which the rockets were launched

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

Launch Site Names Begin with 'CCA'

- The following are 5 records from Launch Sites beginning with “CCA”

Launch_Ste
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Launch Site Names Begin with 'CCA'

- The following are 5 records from Launch Sites beginning 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 following query displays the total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The following output is the average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

Date

22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following table is the list of 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

- The query displays the total number of successful and failure mission outcomes

COUNT(DISTINCT("Landing_Outcome"))

11

Boosters Carried Maximum Payload

- The following are boosters 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 table lists launches that had failed landing outcomes in drone ship, their booster versions, and their launch site names for the 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

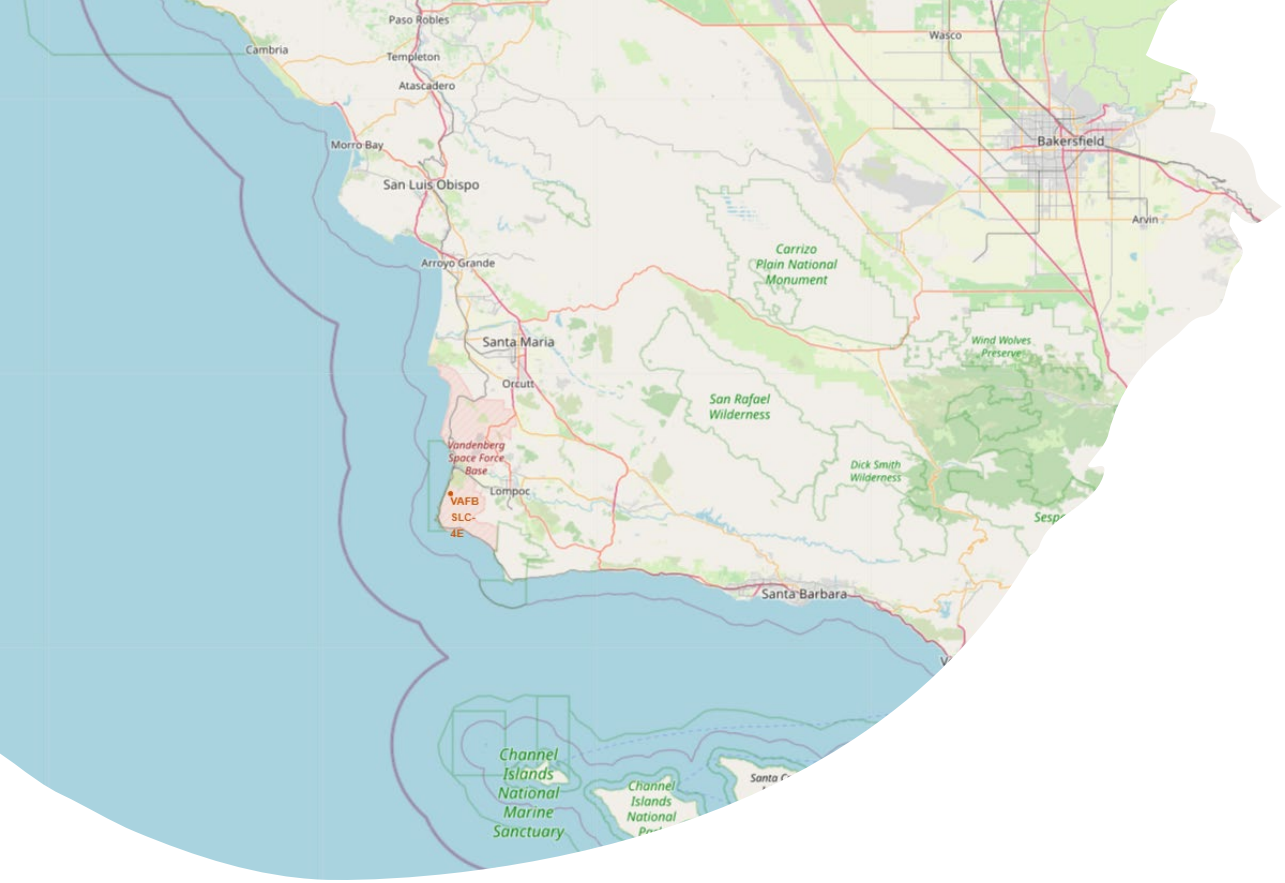
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

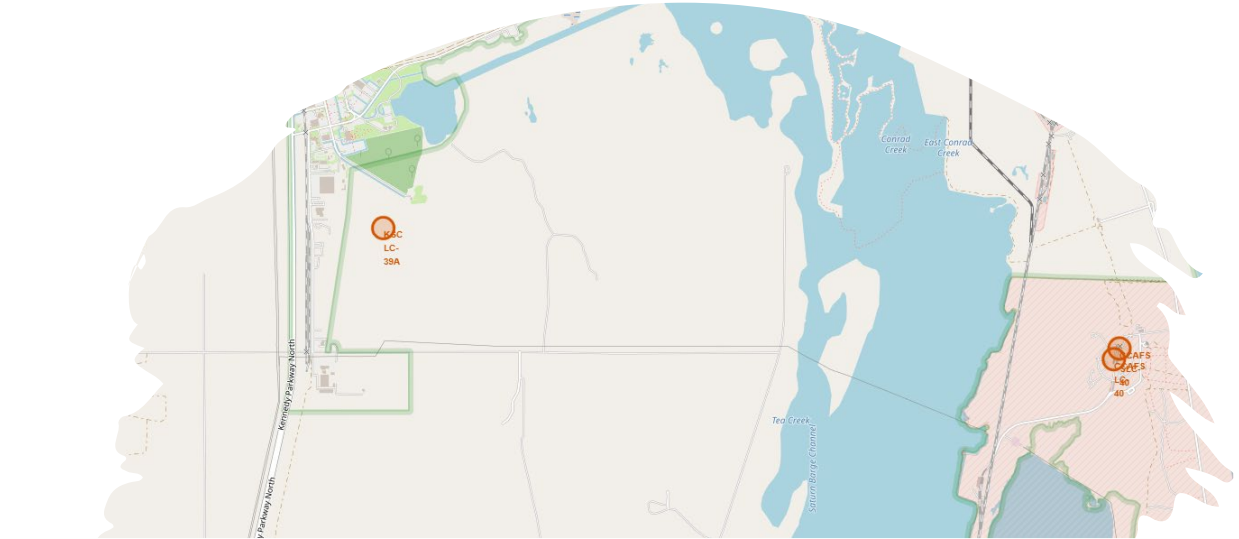
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Present your query result with a short explanation here

Landing _Outcome	COUNT("Landing _Outcome")
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis



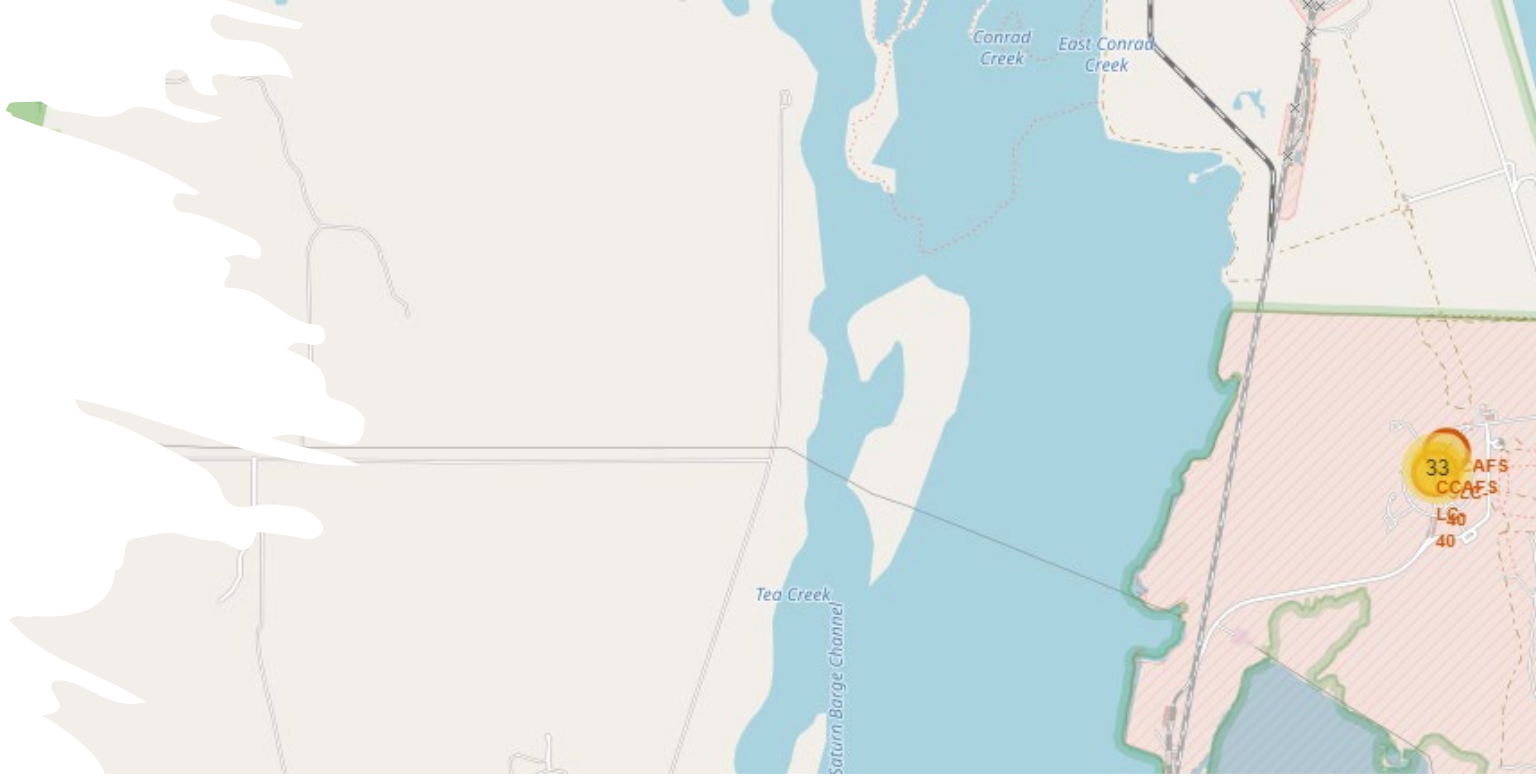


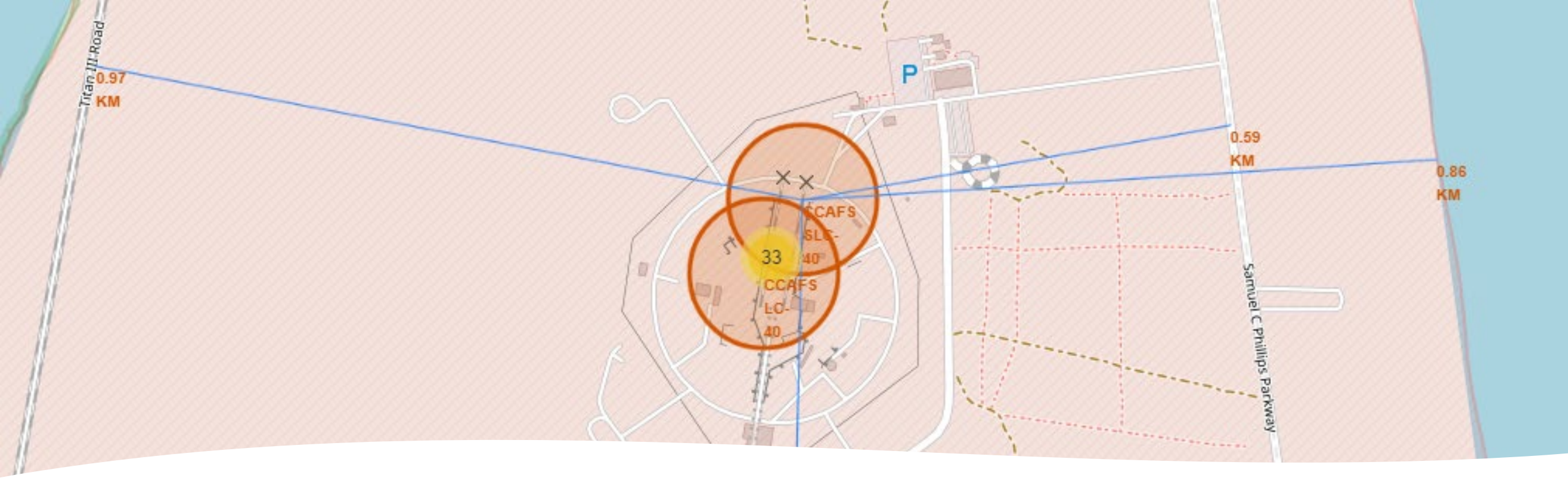
Launch Sites

The following screenshots have 4 circles which indicate the launch sites; 1 located on the west coast of the USA (California) and 3 located on the east coast of the USA (Florida)

Launch Outcome based on Launch Site

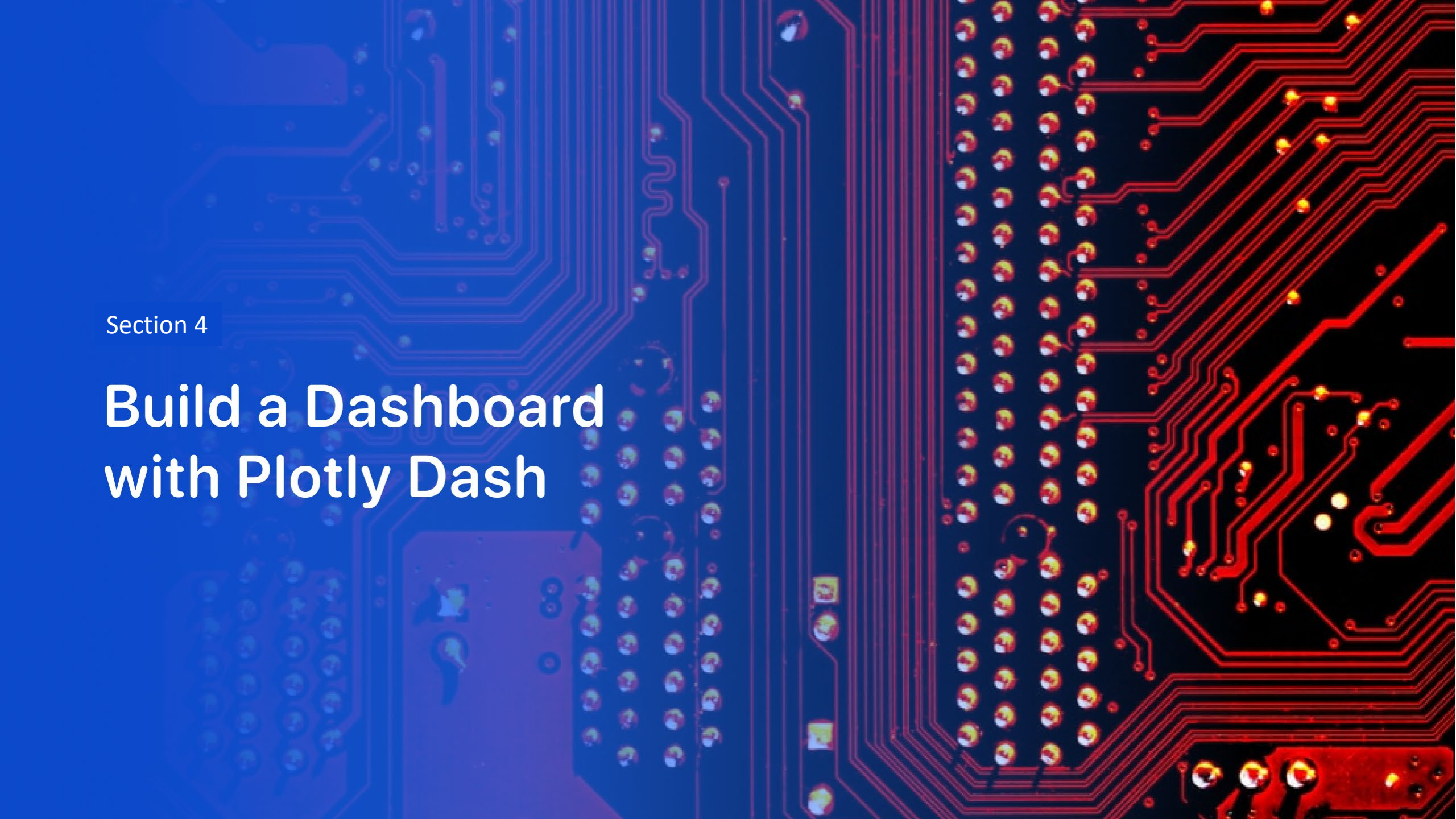
- The following screenshots shows a cluster which indicate the number of markers given a general area
- The 2nd screenshot shows markers colored in red or green to indicate whether the landing of the first stage rocket was successful or not





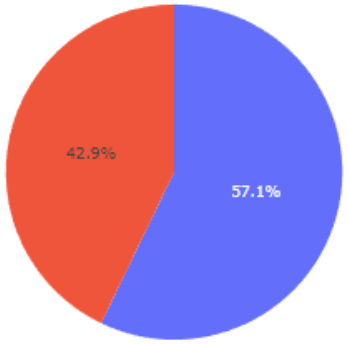
Major Geographical Elements' Proximity to Launch Sites

- The following map has blue lines pointing towards the nearest major coastline, roadway, rail, and city.
  - At the end of each line, there is the distance from the SLC-40 Launch Area


Section 4
Build a Dashboard with Plotly Dash

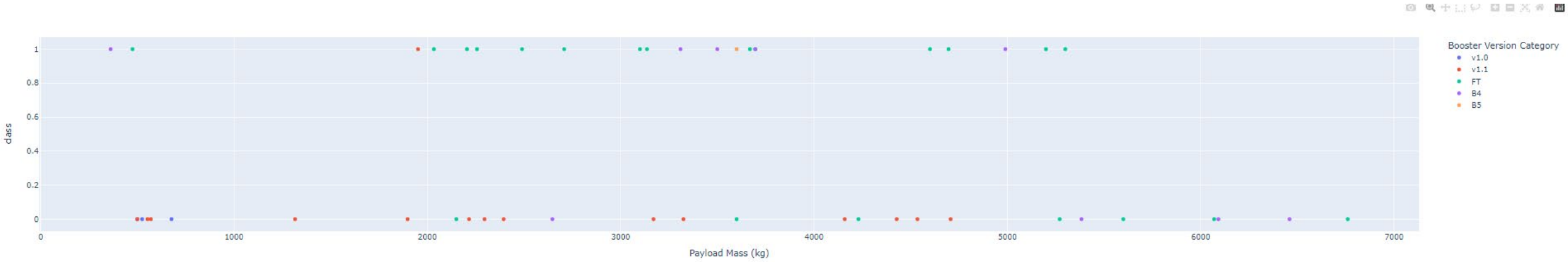
Total Successful Launches by Site

- Out of all the successful landings, LC-39A had the most successes while SLC-40 had the least
  - However, this only counts which launch site had the most successful launches out of the total successful launches; not necessarily which one is most sucessful



Highest Launch Success Ratio

- CCAFS SLC-40 had the highest successful landing rate out of the 4 landing sites
  - Despite having the least number of successful launches, out of the launches that this launch site did have, around 43% resulted in a successful landing



Payload and Booster Version to Success Rate

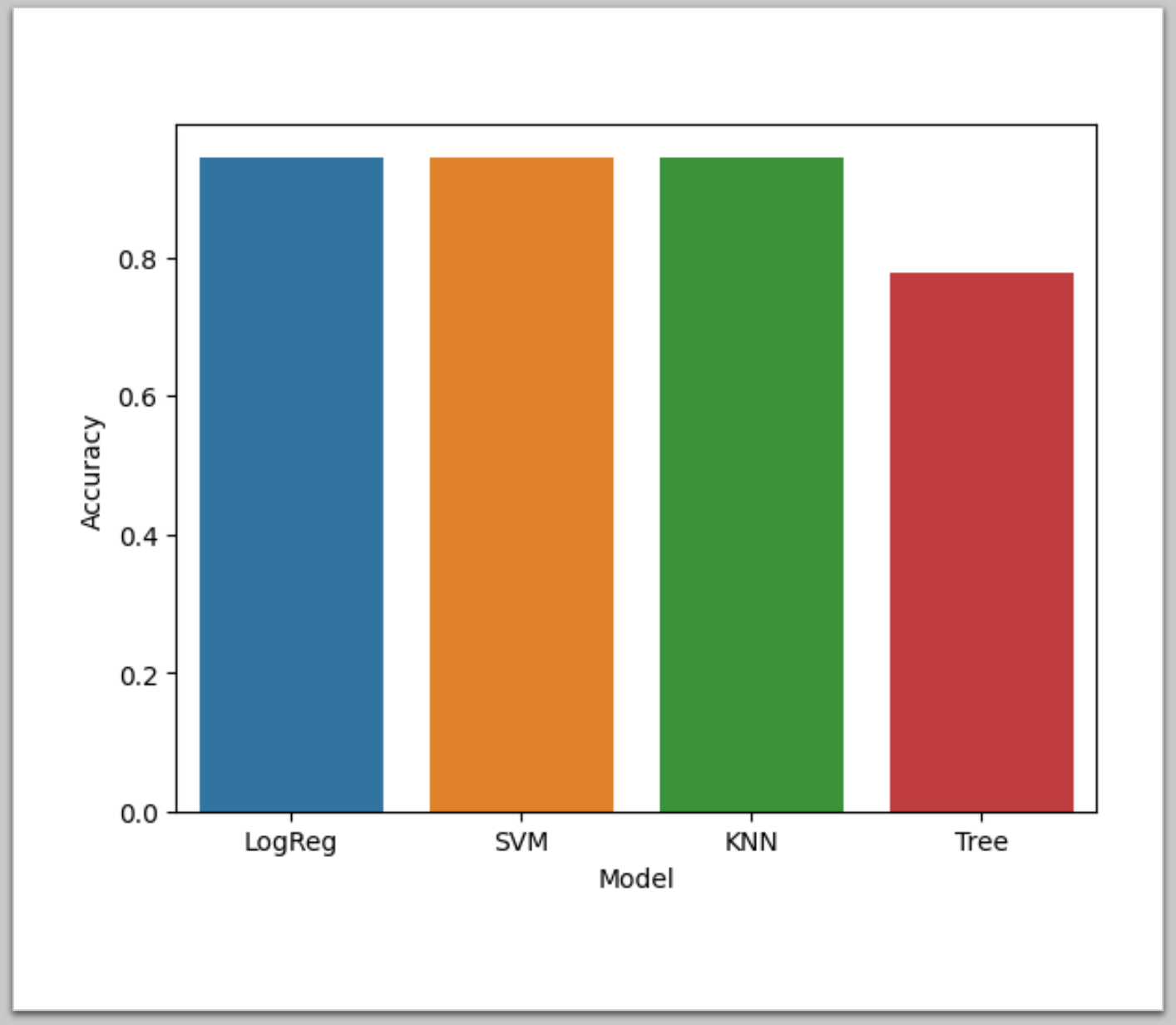
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.
- FT and B4 Boosters were the boosters which experienced the most success when the payload is between a range of 2000kg and 5500kg
- In general, those with a payload mass higher than 6000kg were unsuccessful
- For those with a payload mass of less than 2000kg, it’s success depended on the booster version; FT and B4 being more sucessful

Section 5

Predictive Analysis (Classification)

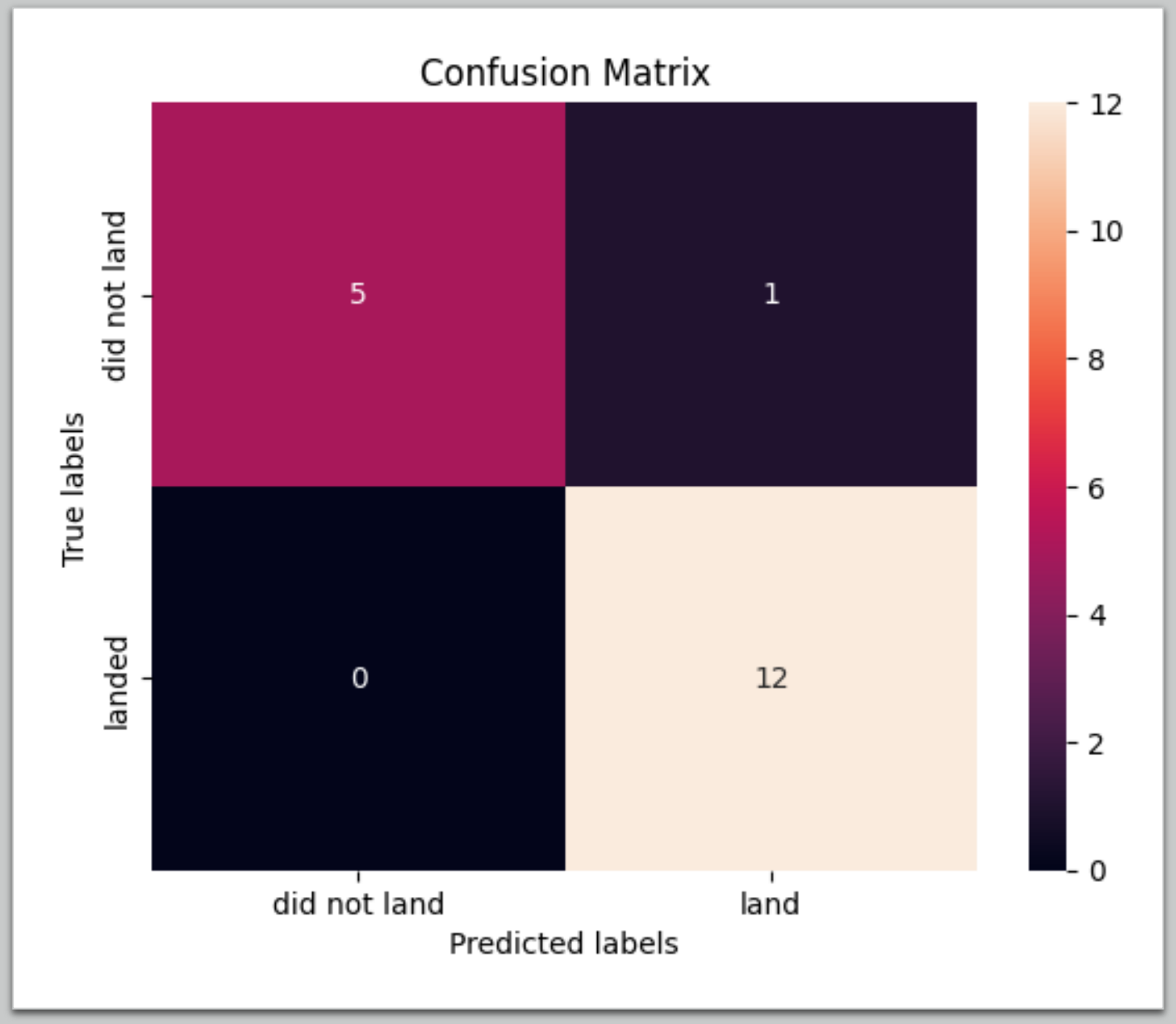
Classification Accuracy

- Logistic Regression, Support Vector Machines, and K Nearest Neighbors had the highest accuracy of 94.44%



Confusion Matrix

- The following Confusion Matrix is shared by the 3 models who had the highest accuracy
- The model correctly predicted 100% of the time when the first stage rocket would land successfully, but would have issues of predicting those that did not land successfully as successful, albeit a low percent



Conclusions

- From the analysis conducted on this scenario, the following conclusions should be particularly noted:
    - Payload Mass, Orbit Type, and Booster Versions were key elements in determining on whether the first stage rocket would land successfully
        - Another aspect to consider is that the more recent (such as future launches) will tend to have a higher success rate compared to launches with similar conditions during the first few launches by SpaceX
    - The 3 models selected, LogReg, SVM, and KNN, are able to determine whether launches would result in successful landing but with a slight problem with False Positives (predicting a successful landing when it will not). Regardless, all three models sport a 90% plus accuracy for out of sample data

Thank you!