IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tan Zhen Xuan
06/07/2022

Outline

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

Executive Summary

- Summary of methodologies used to analyze the data
  - Data collection using web scraping and the SpaceX API;
  - Exploratory Data Analysis (EDA), using data wrangling, data visualization, SQL and interactive visual analytics;
  - Making predictions using machine learning algorithms.
- Summary of all results
  - EDA enabled us to identify prominent features that can better predict success of future launches;
  - The machine learning algorithms enabled us to select the best model to predict those prominent features.

Introduction

- The objective of this project is to determine the viability of Space Y to compete with Space X.
- Problems to be answered in the project:
  - The cost of the launch based on different factors;
  - Chances of the first stage landing successfully;
  - Best places to conduct launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was collected from 2 main sources, namely;
    - Using the SpaceX API
    - Web Scraping from Wikipedia
- Perform data wrangling
  - Data collected is analyzed and reinforced by creating a new label called “Landing Outcome”.
  - This would be used for training the supervised machine learning models down the line.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected up till now is preprocessed and divided into training and test data sets to be used by the 4 different classification models.
    - The accuracy of each model will be determined by different combinations of parameters.

Data Collection

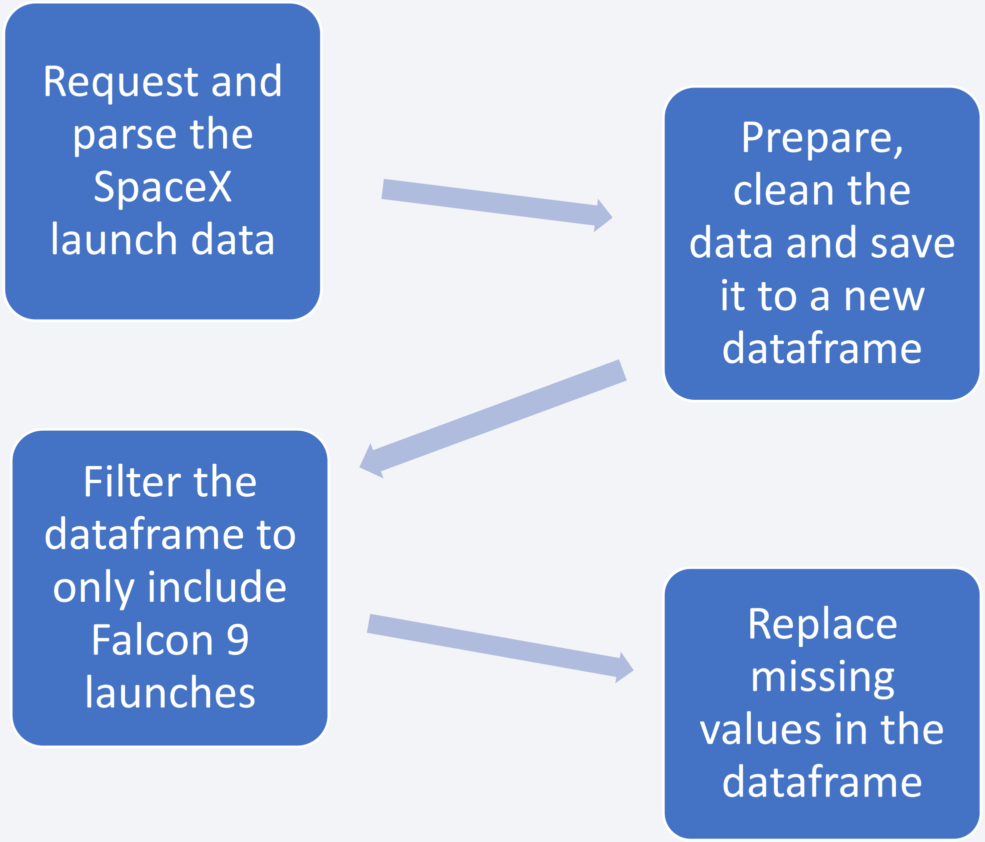
- The datasets are collected using the following ways, namely;
  - SpaceX API
    - (https://api.spacexdata.com/v4/rockets/)
  - From Wikipedia using Web Scraping
    - The Wikipedia page used is (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- The SpaceX REST API is an open-source REST API for obtaining various data about SpaceX.

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

- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/Data%20Collection%20API.ipynb



Request and parse the SpaceX launch data

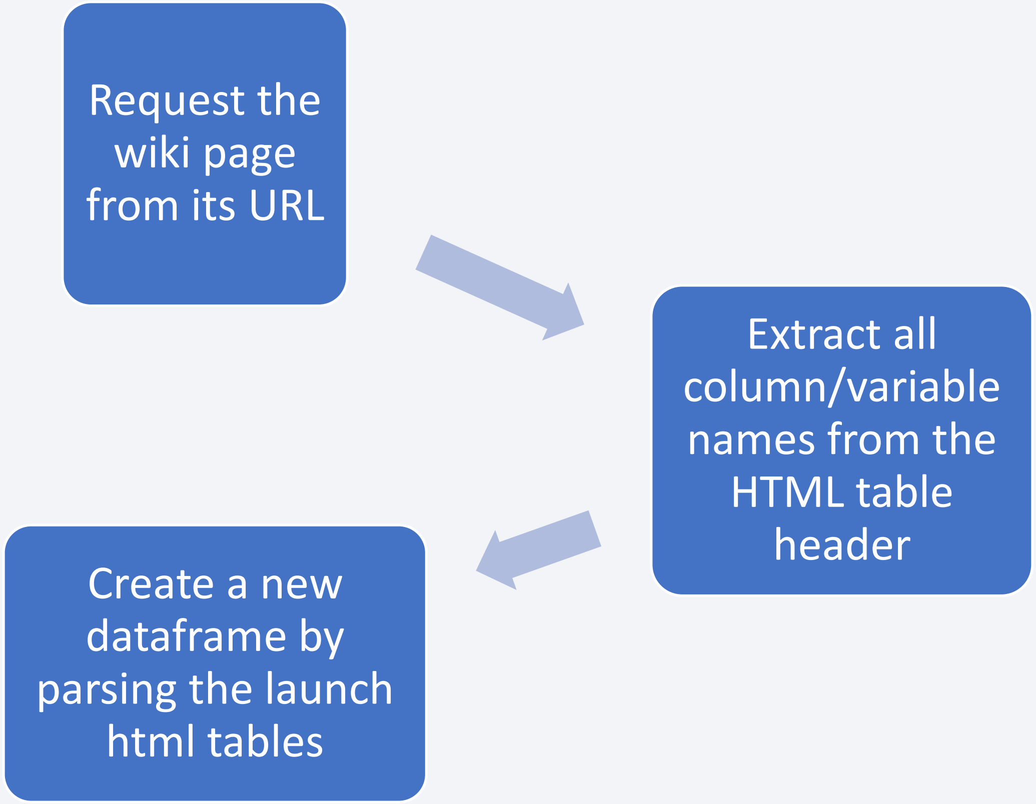
Prepare, clean the data and save it to a new dataframe

Filter the dataframe to only include Falcon 9 launches

Replace missing values in the dataframe

Data Collection - Scraping

- Data from SpaceX launches are also obtained from Wikipedia.
- Data is scrapped using BeautifulSoup and cleaned, then saved to a new dataframe.
- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Request the wiki page from its URL

Extract all column/variable names from the HTML table header

Create a new dataframe by parsing the launch html tables

Data Wrangling

- Firstly, some EDA is done on the dataset.
- Then, some calculations are done to determine the number of launches on the different sites and the number of occurrence of each orbit.
- Lastly, a landing outcome label is created using data from the Outcome column.

- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- To explore and visualize the collected data and the relationship between the different features, scatterplots, bar charts and line charts are used.
- The following relationships are plotted in the notebook;
    - Flight Number against Launch Site, and against Orbit Type
    - Payload Mass against Launch Site, and against Orbit Type
    - Orbit Type against Success Rate
    - Year against Success Rate

- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- The following SQL queries are performed:
  - Names of the unique launch sites used in the space mission
  - Display 5 records where launch sites begins with “CCA”
  - Display the total payload mass carried by boosters launched by NASA
  - Display the average payload mass carried by booster version “F9 v1.1”
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions that have carried the max payload mass
  - List the failed landing outcomes in drone ship, their booster versions and launch site names in the year 2015
  - Rank the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order
- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map objects such as markers, marker clusters, circles and lines are used together with Folium Maps
  - Markers indicate points of interest on the map
  - Marker clusters indicate group of events in each coordinate
  - Circles indicate highlighted areas around specific coordinates
  - Lines indicate distances between 2 point of interests or coordinates

- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

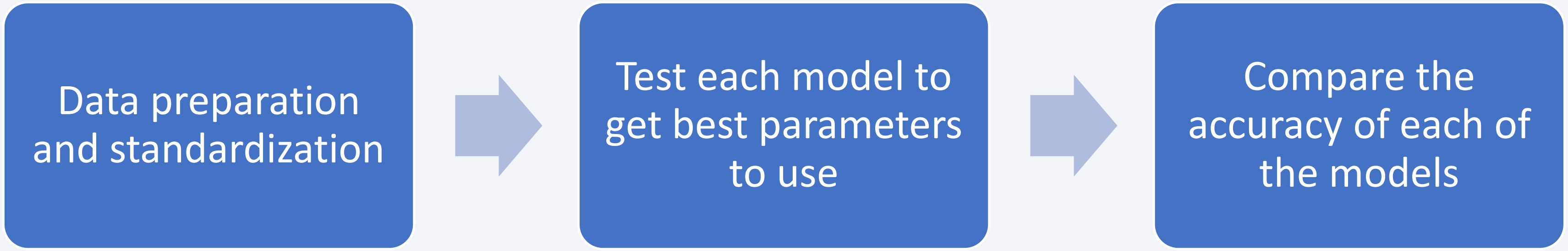
Build a Dashboard with Plotly Dash

- The graphs and plots that were used to visualize the data are as follows;
  - Pie chart of the Percentage of launches by the different sites
  - Scatterplot of how payload mass and booster version affects success rate
- The plots enables us to analyze the relation between the different features, which allows for quick identification of good launch sites and payload mass to use for successful launches.

- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- 4 classification models were compared, namely;
  - Logistic regression
  - Support Vector machine (SVM)
  - Decision tree
  - K Nearest Neighbour (KNN)



Data preparation and standardization

Test each model to get best parameters to use

Compare the accuracy of each of the models

- Source code: https://github.com/Lonpeace/IBM_data_science_applied_capstone/blob/master/Machine%20Learning%20Prediction.ipynb

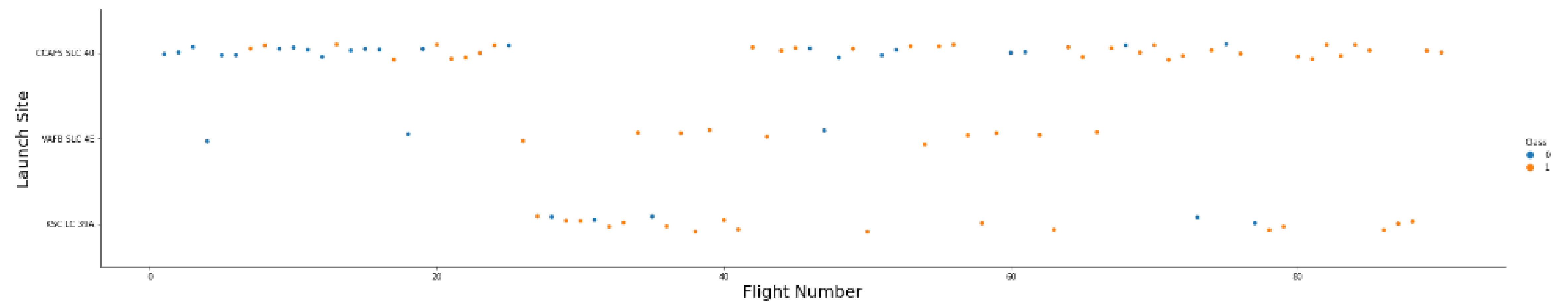
Results

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

Section 2

Insights drawn from EDA

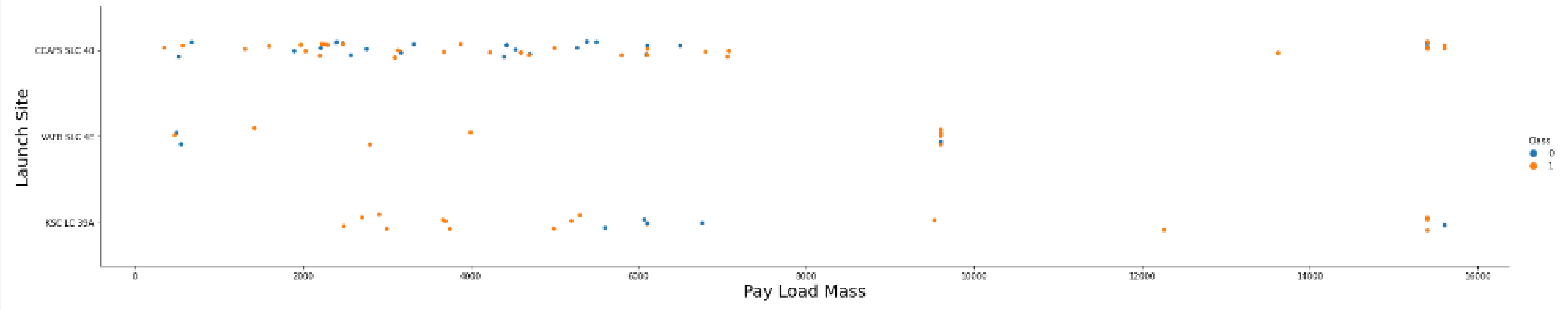
Flight Number vs. Launch Site



From the plot, we can observe that the CCAFS SLC 40 launch site is the more popular choice among the 3 launch sites, as a lot more launches have been made from that launch site as compared to the other 2.

Also, we can derive that the more frequently that a launch site is used, the success rate of that particular launch site will increase.

Payload vs. Launch Site

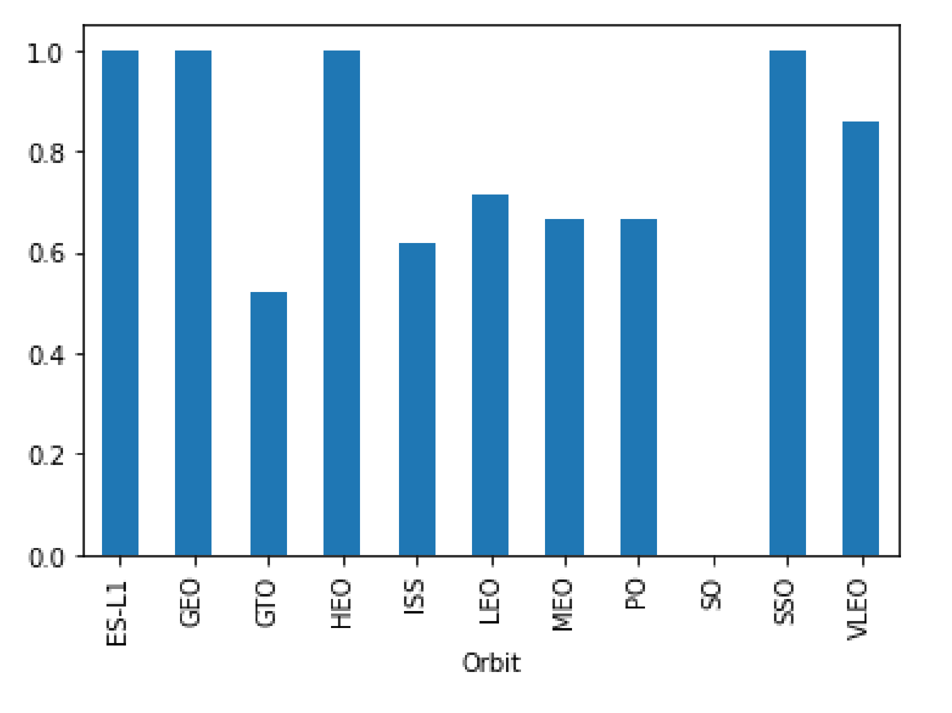


From the chart, we can observe that payloads over the weight of 10,000 kg are only launched from the CCAFS SLC 40 and KSC LC 39A launch sites.

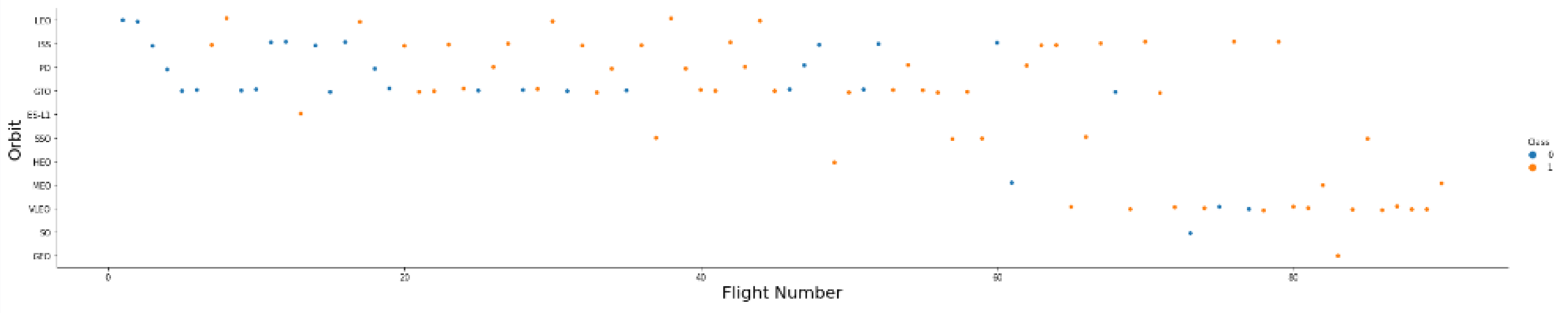
Success Rate vs. Orbit Type

Based on the bar chart, we can observe that the orbit types that have the highest success rates are as follows;

- ES-L1
- GEO
- HEO
- SSO
- VLEO



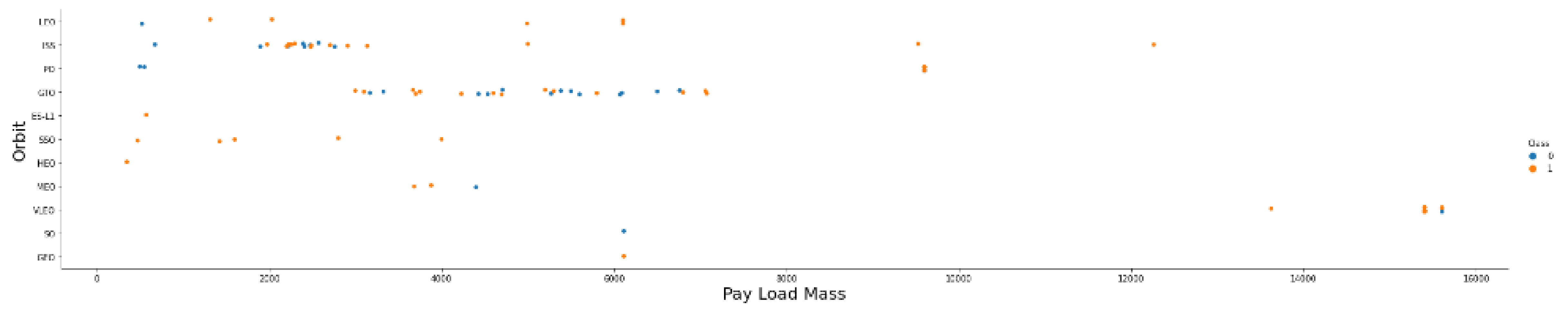
Flight Number vs. Orbit Type



From the chart, we can observe that the success rates for flights increases over time. This is prevalent in the orbit type of LEO, whereas the same cannot be said for the orbit type GTO.

As a side note, SpaceX seems to be taking an interest in the orbit type VLEO, as its frequency has increased as of late.

Payload vs. Orbit Type



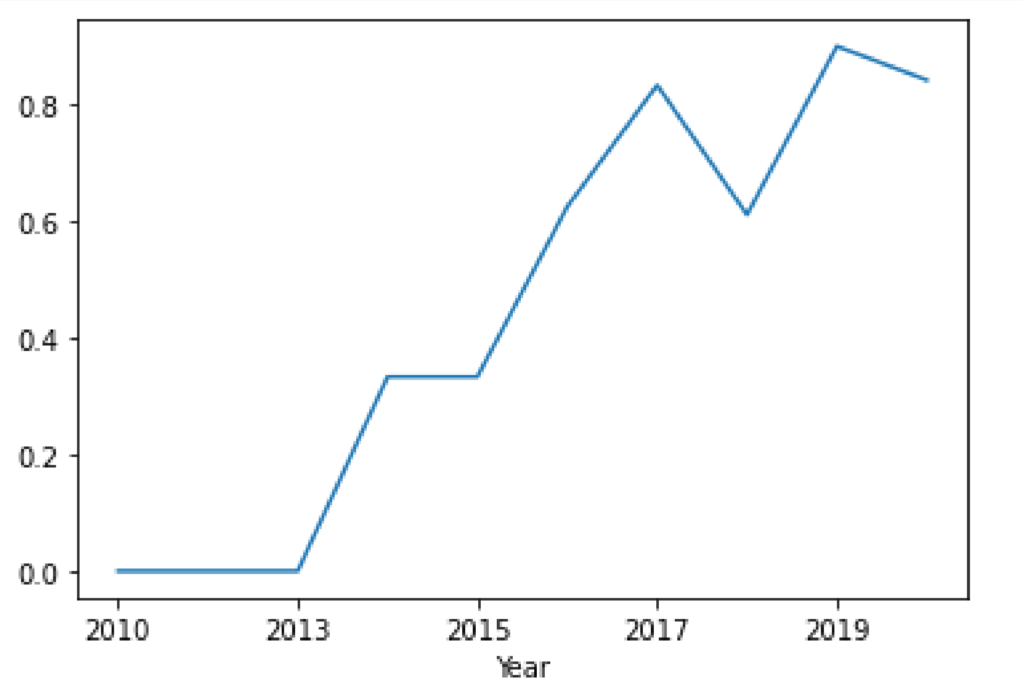
For the orbit type GTO, we cannot distinguish the whether payload mass influence the success rate of the launch.

For the orbit types Polar, LEO and ISS, it is observed that a heavier payload increases the success rate of the launch.

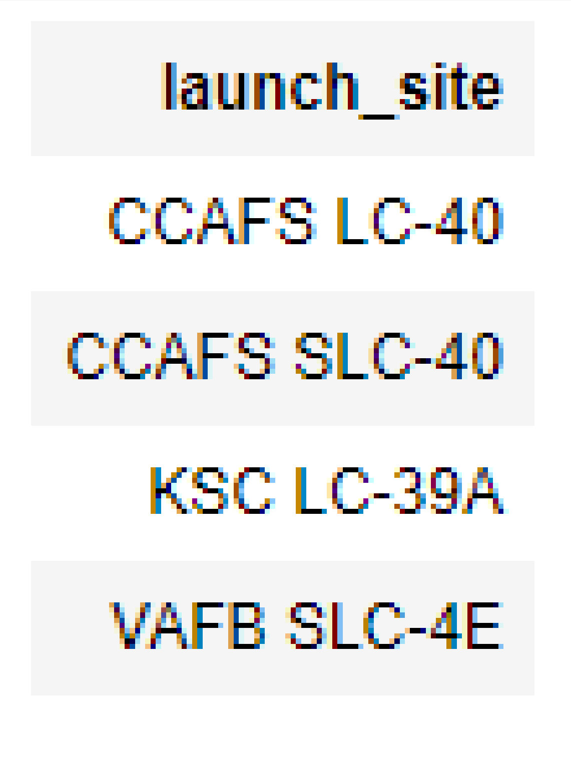
Launch Success Yearly Trend

From the chart, we can observe that the success rate of launches have been steadily improving over the years.

In 2018, it is observed that there was a drop of close to 20% in the success rates of launches.



All Launch Site Names



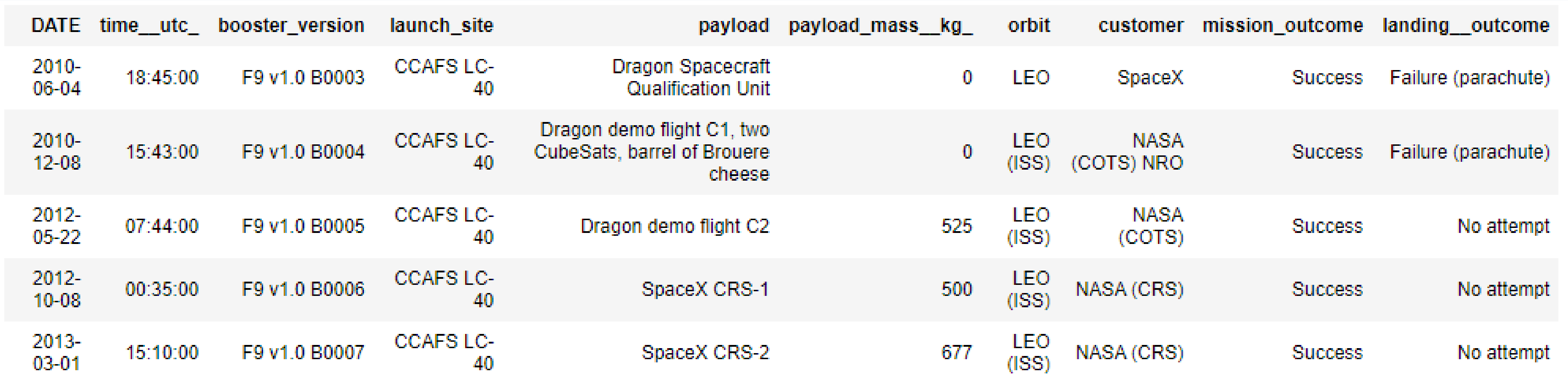
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Based on the query results, there are a total of 4 unique launch sites names, namely;

1. CCAFS LC-40
2. CCAFS SLC-40
3. KSC LC-39A
4. VAFB SLC-4E

The names are obtained by selecting unique occurrences of each value in the ‘launch_site’ column.

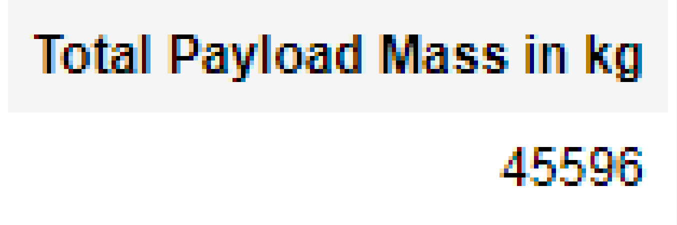
Launch Site Names Begin with 'CCA'



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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Based on the query results, we can observe 5 records that has a launch site name that begins with ‘CCA’.

Total Payload Mass

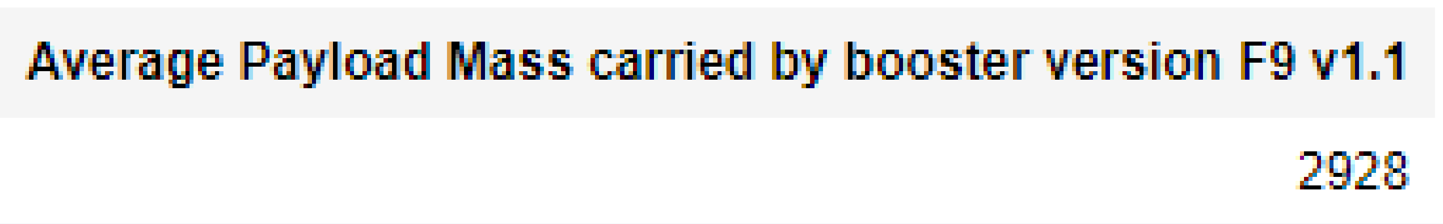


Total Payload Mass in kg
45596

Based on the query result, the total payload mass that is carried by boosters launched by NASA (CRS) is 45,596kg.

The result is obtained by summing up all the payload mass of rows that have NASA (CRS) as the customer.

Average Payload Mass by F9 v1.1

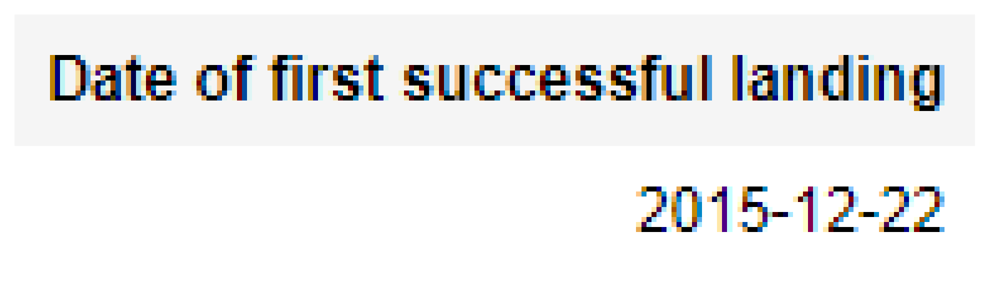


Average Payload Mass carried by booster version F9 v1.1
2928

Based on the query result, the average payload mass carried by the booster version “F9 v1.1” is 2,928kg.

The result is obtained by filtering the table for rows that have booster version = ‘F9 v1.1’ and taking the average of the payload mass.

First Successful Ground Landing Date

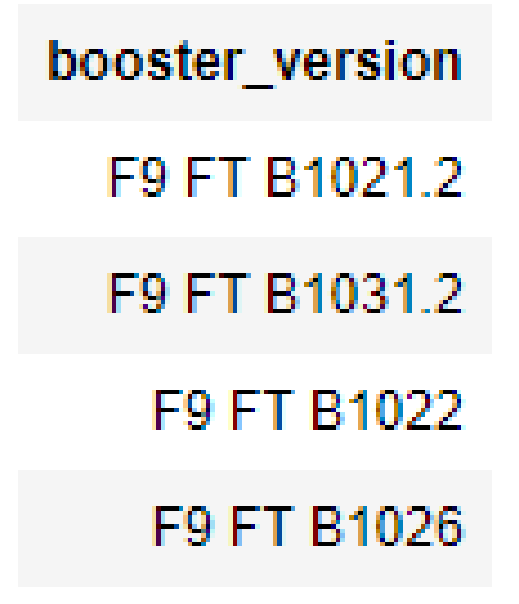


Date of first successful landing
2015-12-22

Based on the query result, the first successful landing occurred on the date 22/12/2015.

This result is obtained by filtering the table for rows with a successful landing outcome and finding the earliest date from those rows.

Successful Drone Ship Landing with Payload between 4000 and 6000



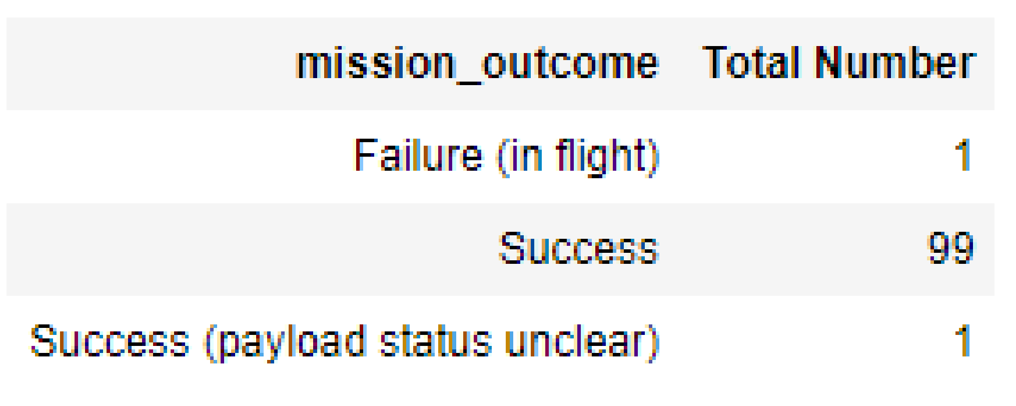
booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Based on the query results, the booster versions that have a successful drone ship landing with a payload between 4000kg and 6000kg are as follows;

1. F9 FT B1021.2
2. F9 FT B1031.2
3. F9 FT B1022
4. F9 FT B1026

This result is obtained by filtering the table for rows that have a successful landing outcome on a drone ship and a payload mass between 4000kg and 6000kg. We then select each unique occurrences of the booster version.

Total Number of Successful and Failure Mission Outcomes

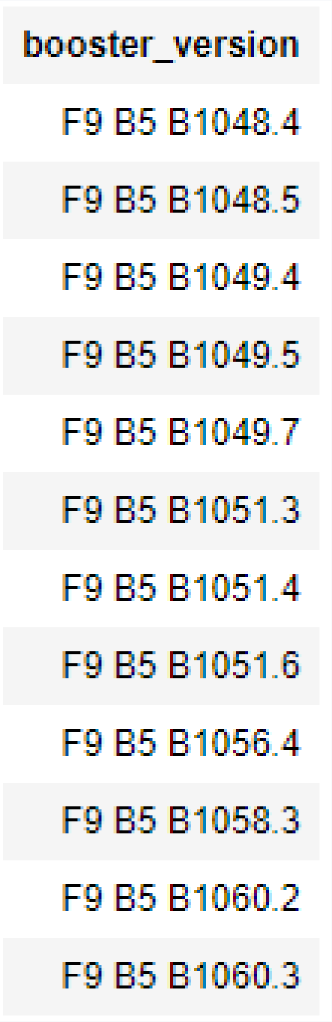


mission_outcome	Total Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Based on the query results, we can observe that there are 100 successful missions while having only 1 failure.

This result is obtained by grouping the rows with the mission outcome column and counting each occurrence.

Boosters Carried Maximum Payload

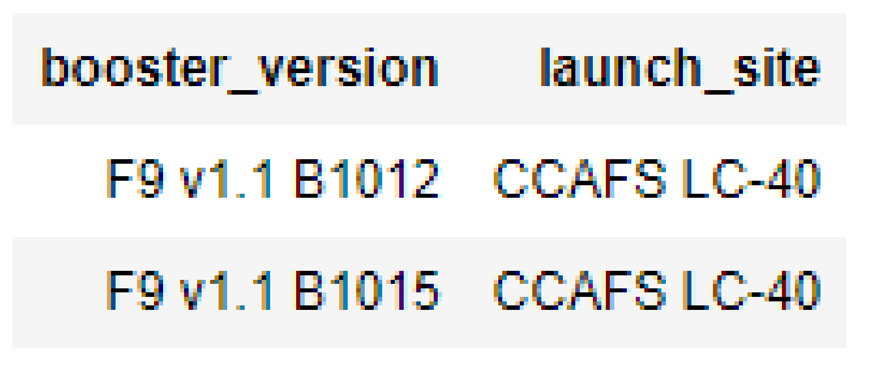


booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Based on the query result, we can find all the booster versions that have carried the maximum payload mass.

We can obtain the value of the maximum payload mass by using a subquery.

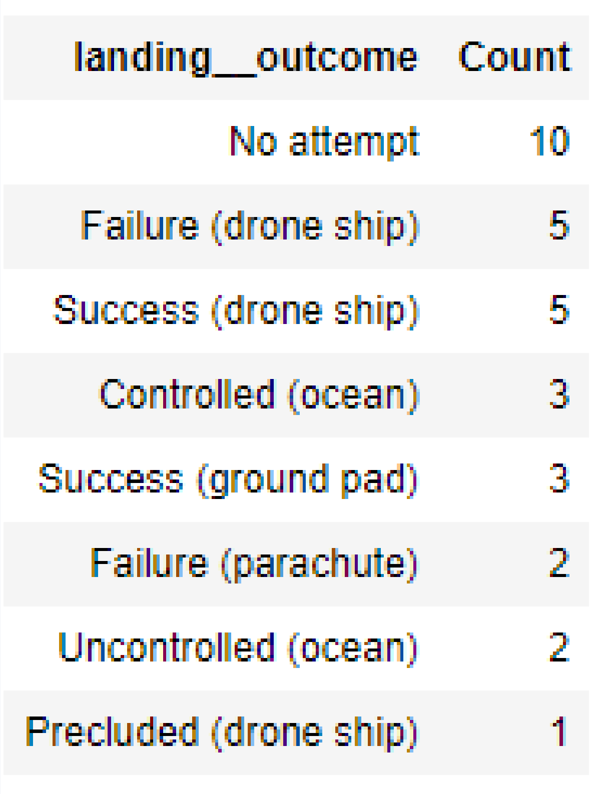
2015 Launch Records



booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Based on the query result, the booster versions and launch sites that have a failed landing outcome on the drone ship in the year 2015 can be observed above.

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



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

Between 04/06/2010 and 20/03/2017, it can be observed that the number of times that no attempt was made to land the booster was the highest.

This is followed by an equal number of successful landings and unsuccessful landings on drone ships.

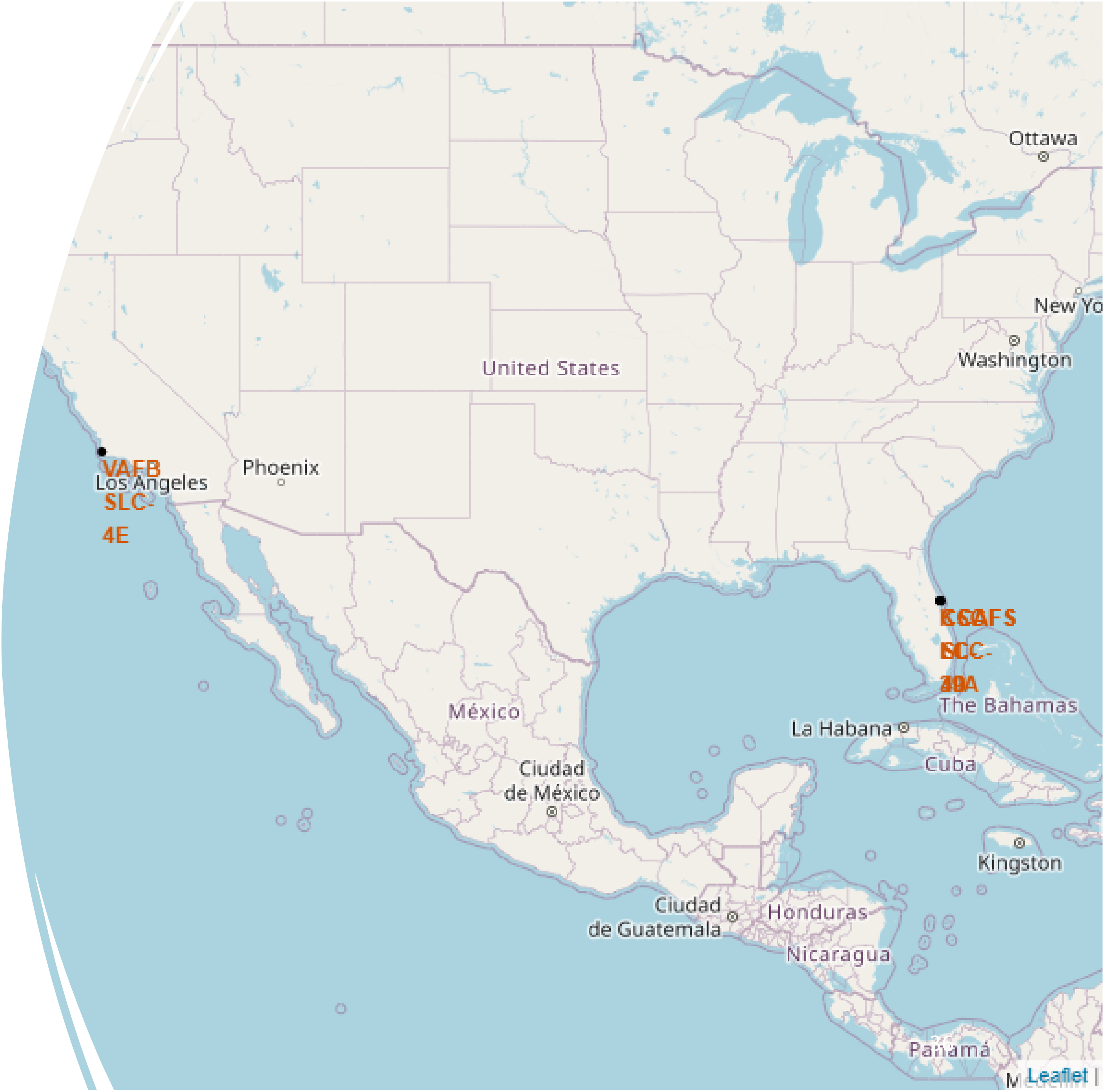
Section 3

Launch Sites Proximities Analysis

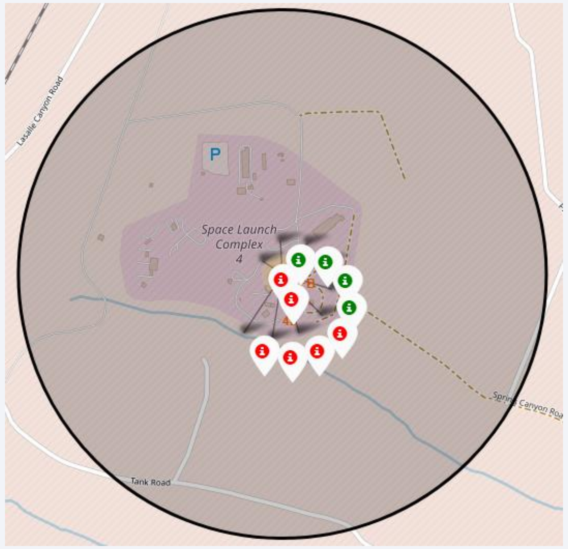
All launch site's location

The screenshot on the left shows all the launch sites used by SpaceX for their rocket launches.

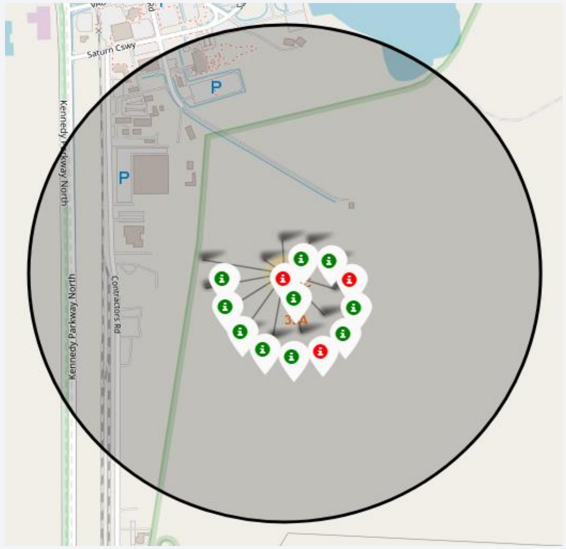
All the launch sites are located near the coastlines.



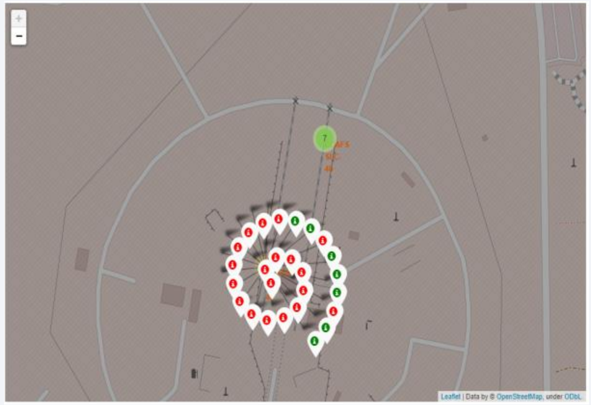
Launch sites with color coded launch outcomes



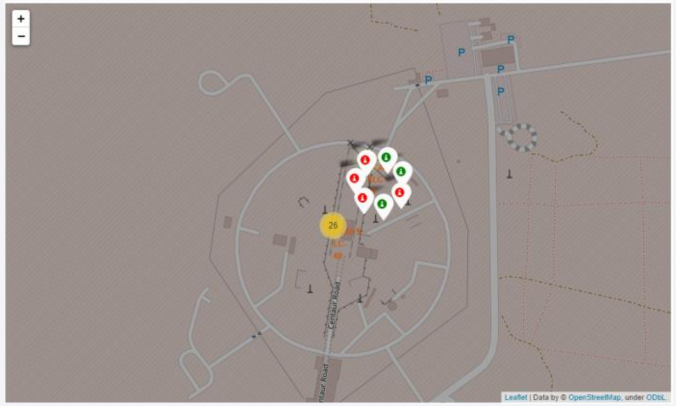
VAFB SLC-4E



KSC LC-39A



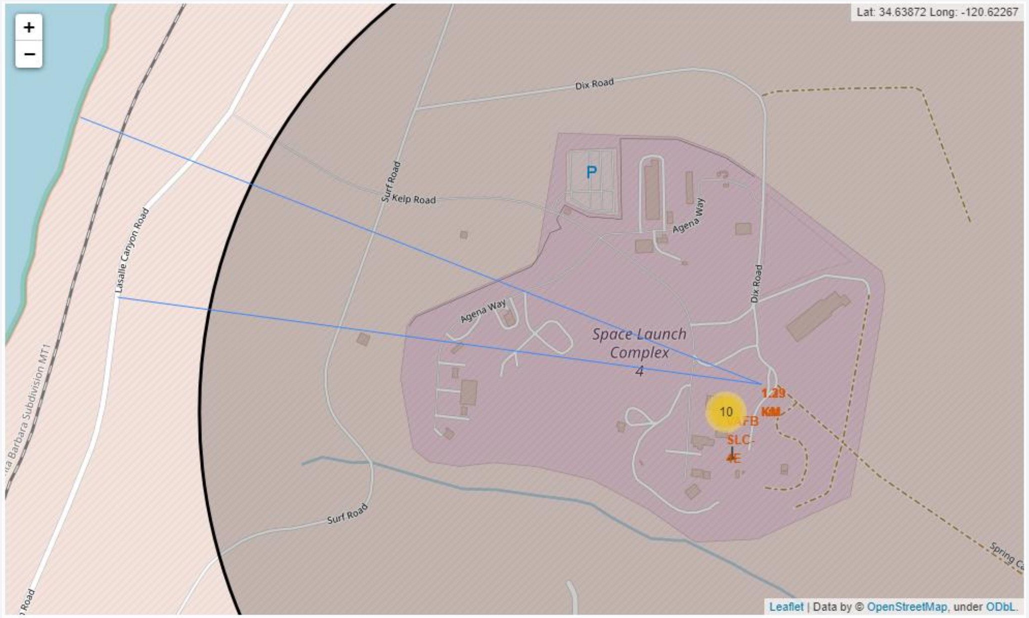
CCAFS LC-40



CCAFS SLC-40

The green markers represent successful launches while the red markers represent unsuccessful launches.

Launch site distance to landmarks



Distance from nearest coastline: 1.39km

Distance from nearest highway: 1.28km

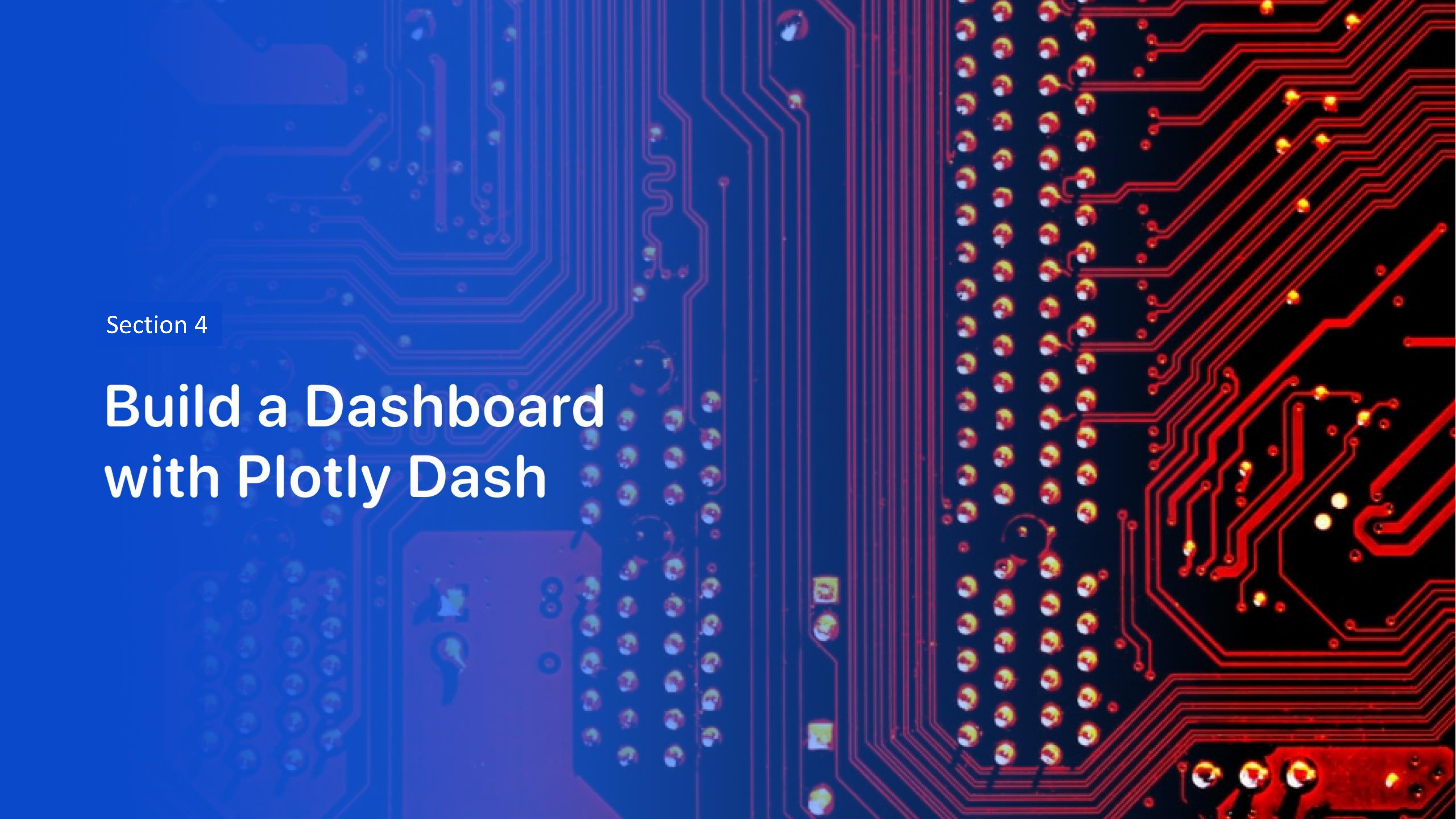
The launch site is located quite close to a coastline and a highway.

It could be located close to the coastline for easy retrieval of the booster after a successful landing.

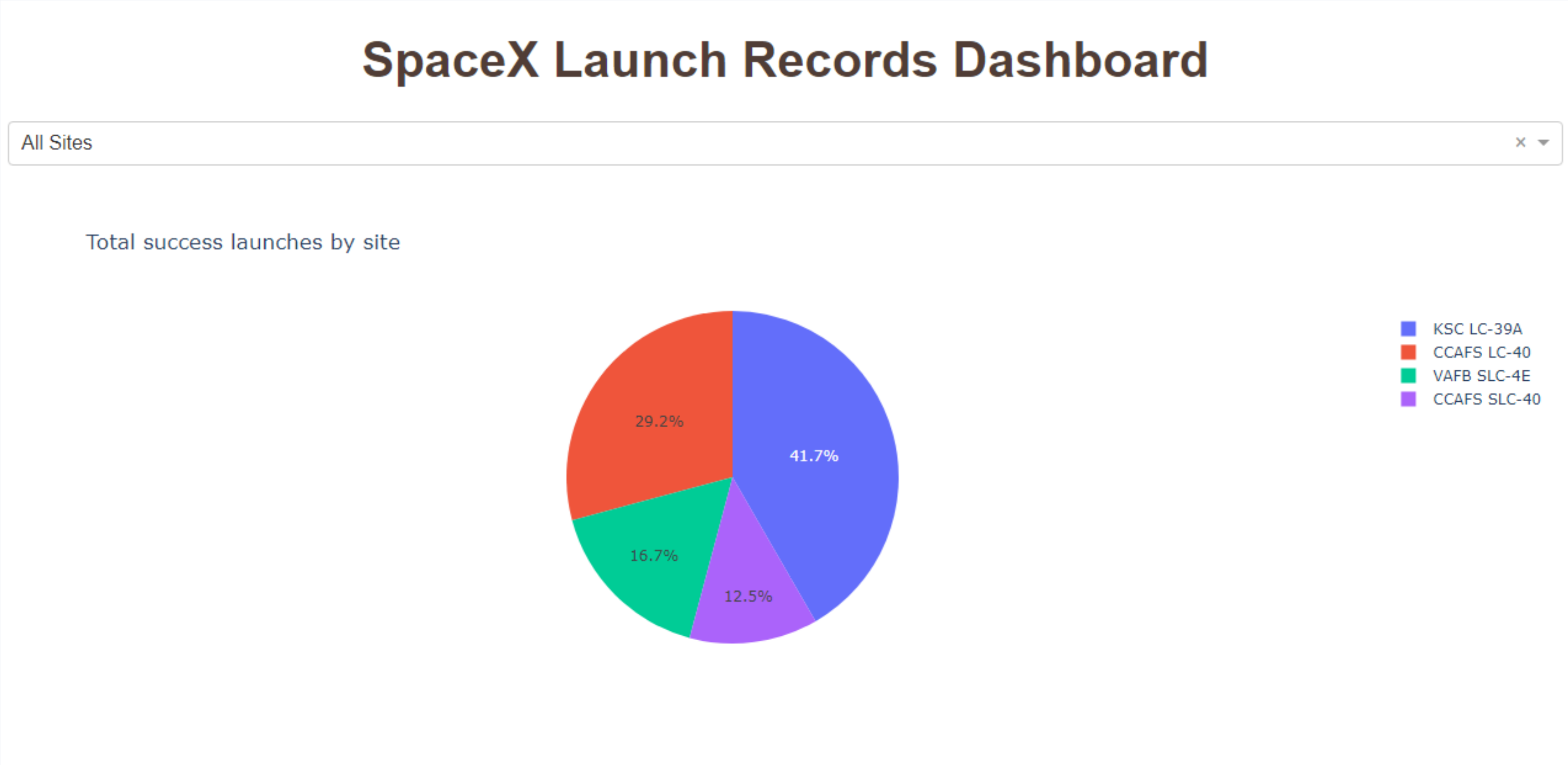
Being close to a highway also allows for more efficient logistical processes.

Section 4

Build a Dashboard with Plotly Dash



Successful launches for all sites



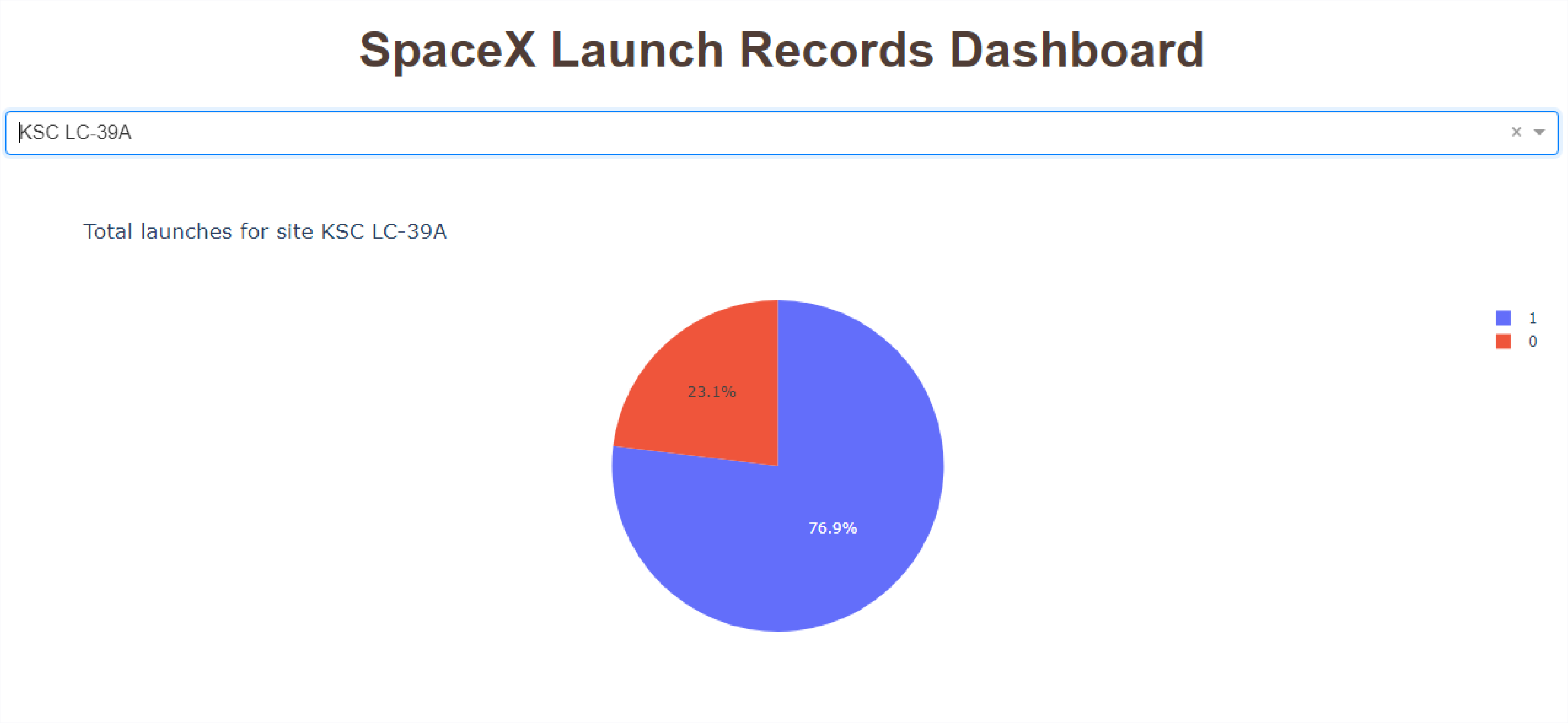
SpaceX Launch Records Dashboard

All Sites

We can observe that the launch site “KSC LC-39A” has the greatest number of successful launches among the launch sites.

It contributed 41.7% of the total successful launches conducted by SpaceX.

Launch site with the highest launch success ratio



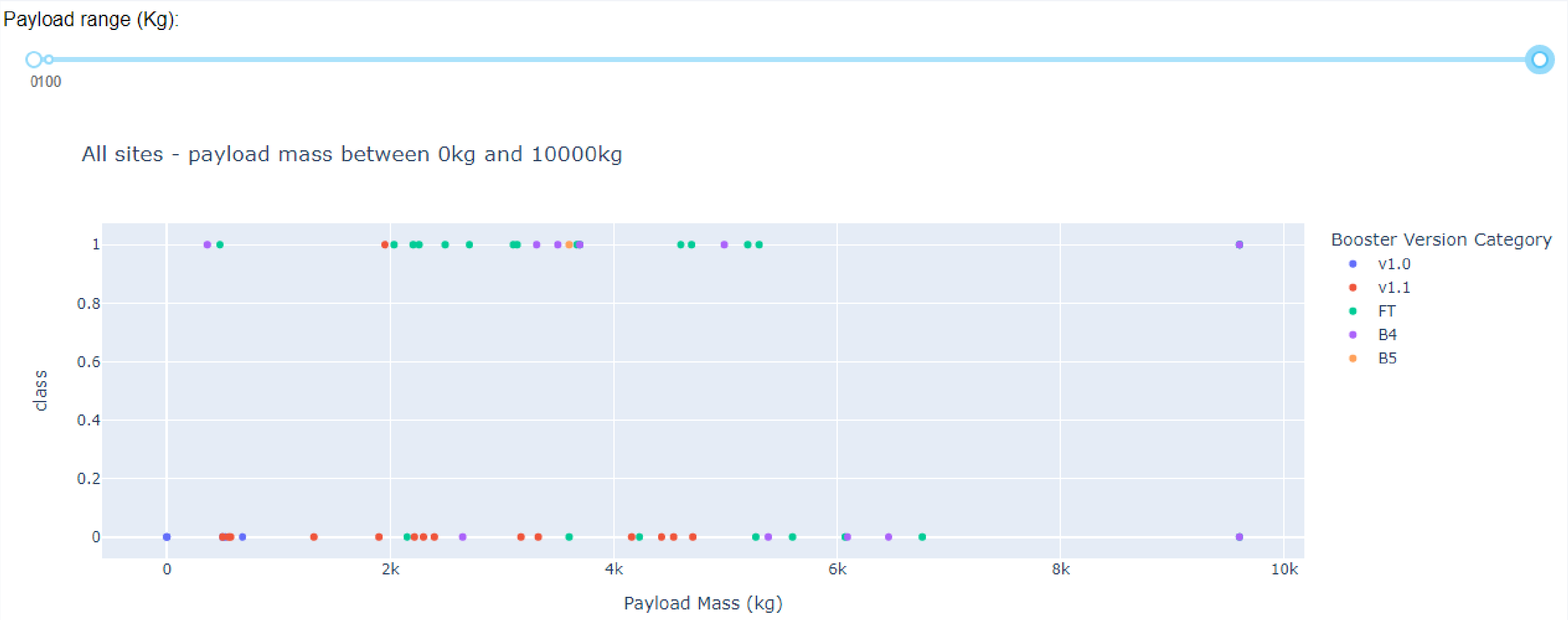
SpaceX Launch Records Dashboard

KSC LC-39A

Total launches for site KSC LC-39A

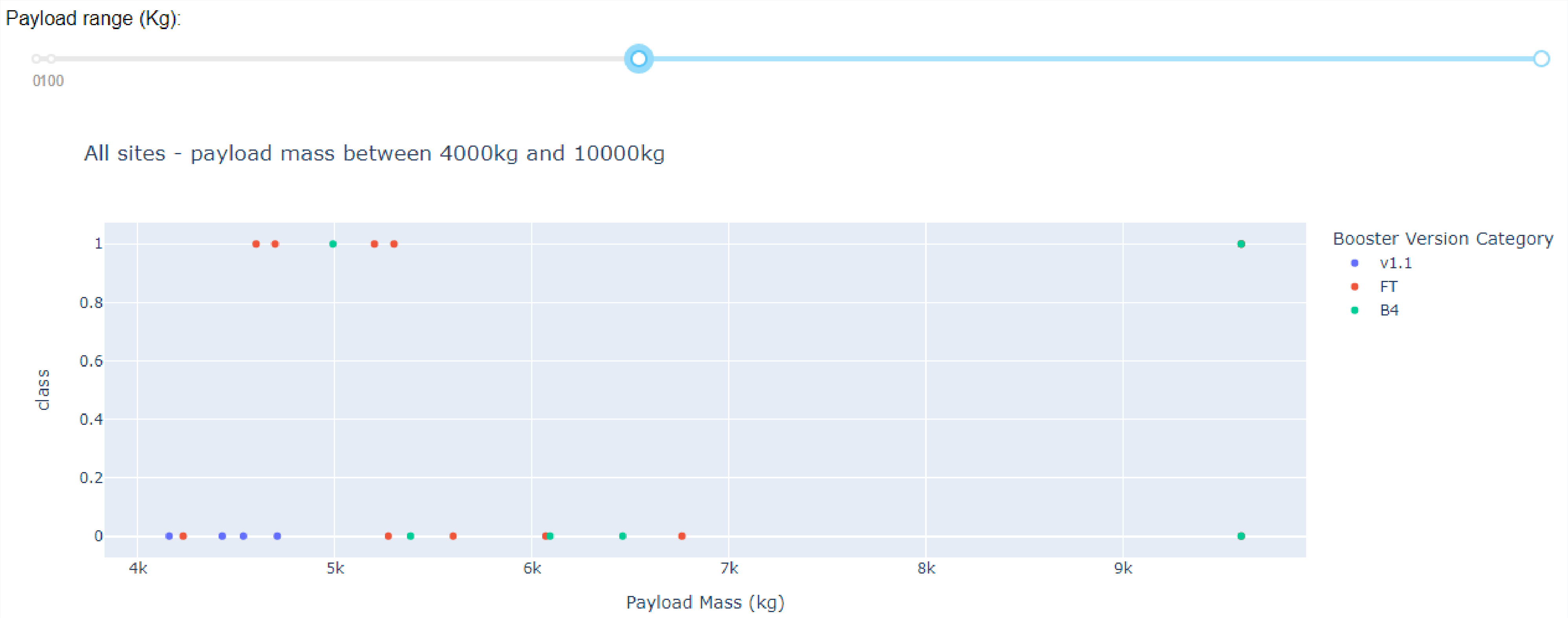
From the dashboard, it can be observed that 76.9% of the launches done in the site is successful.

Payload vs Launch Outcome – Full range



It can be observed that the booster version “FT” has the highest success rate among the different booster versions for payloads that are below 6000kg.

Payload vs Launch Outcome – 4k to 10k

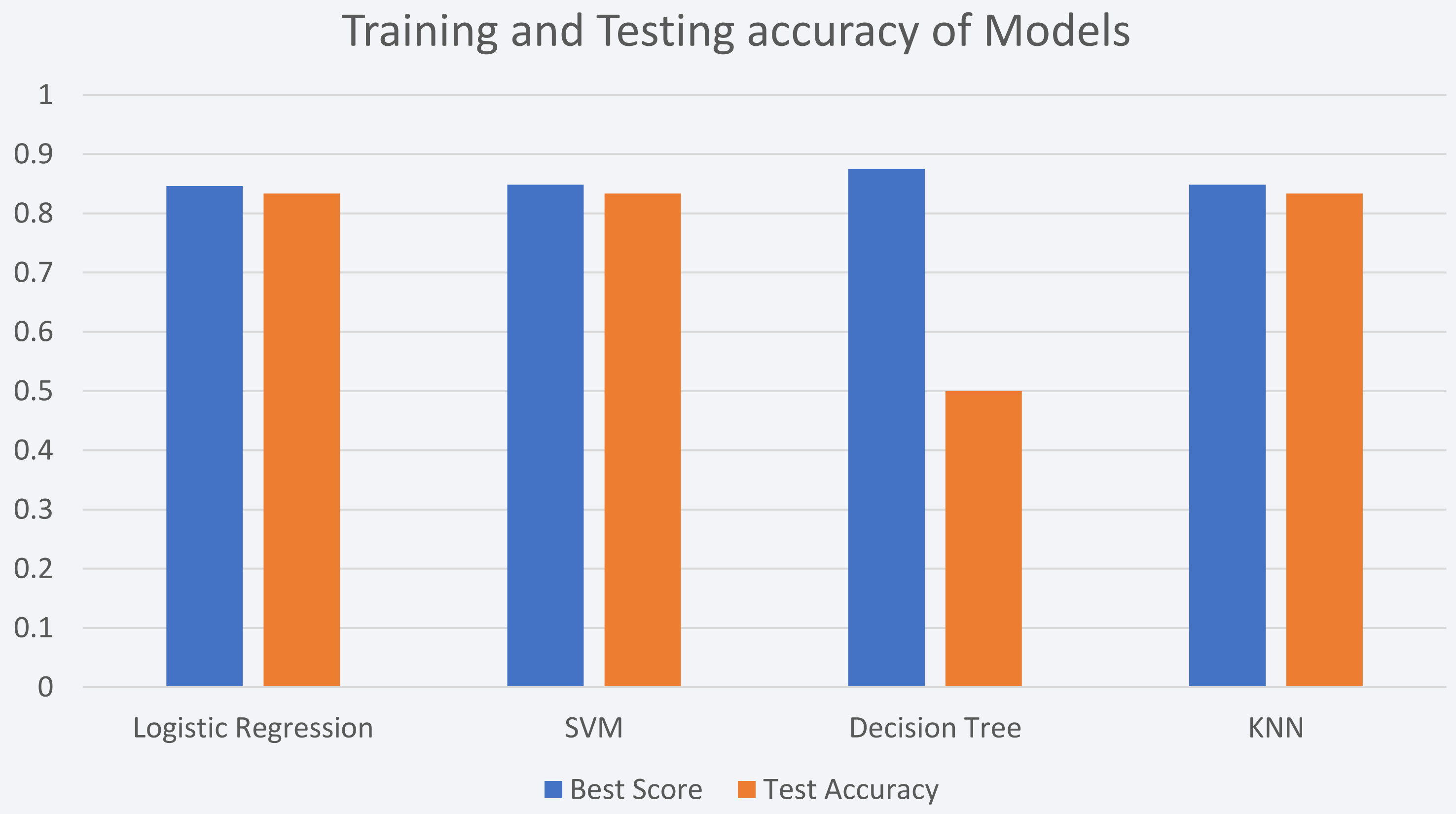


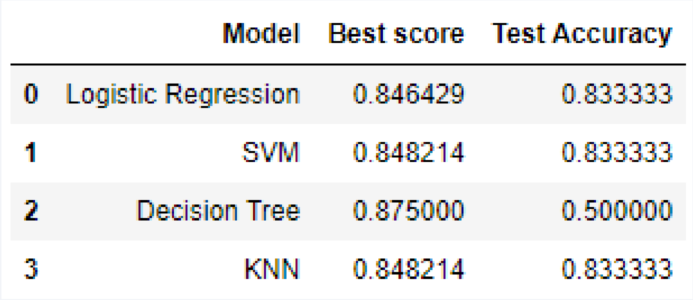
It can be observed that regardless of booster version, the success rate of launches decreases when the payload mass increases beyond 4000kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



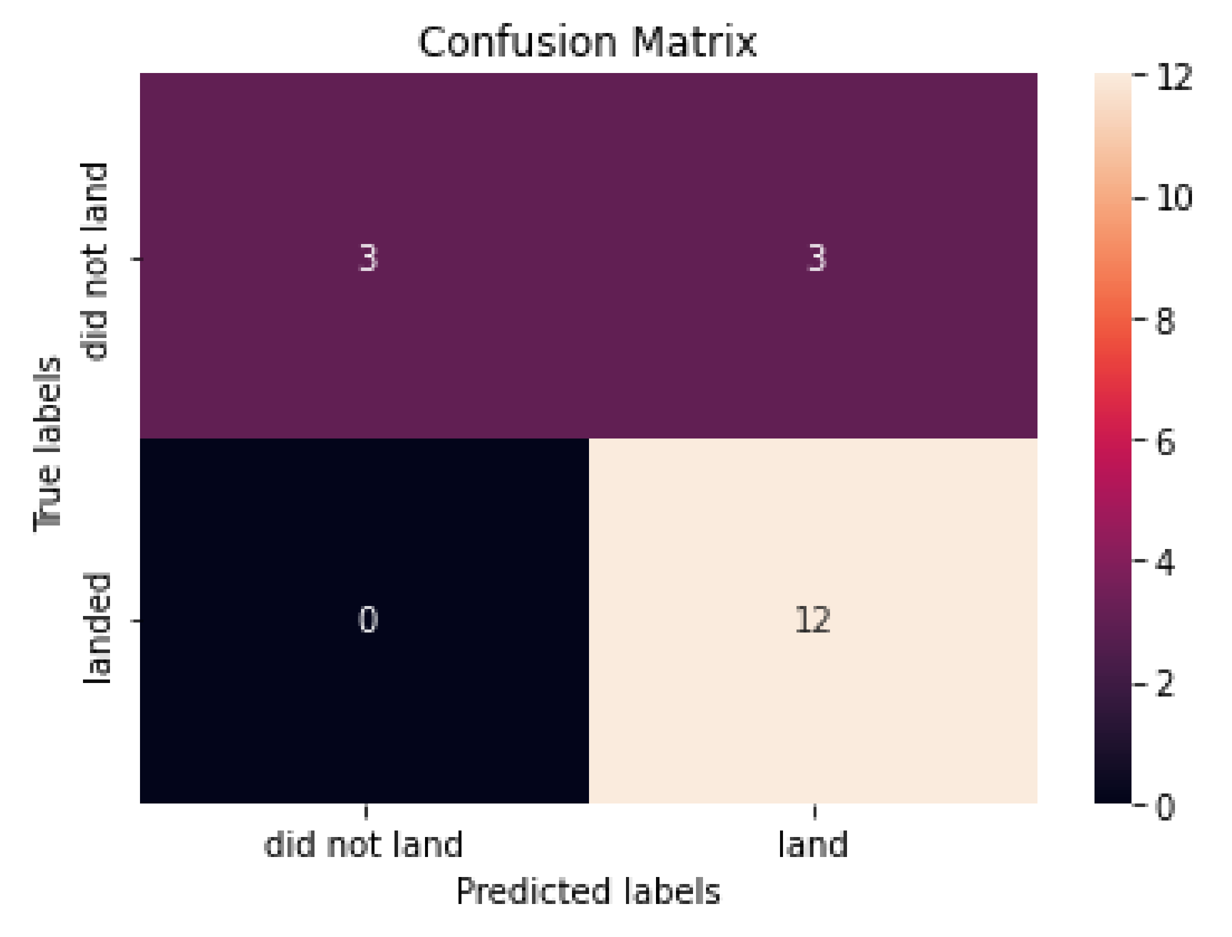


	Model	Best score	Test Accuracy
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.875000	0.500000
3	KNN	0.848214	0.833333

Among the different models used, the decision tree has the highest training accuracy of 0.875. However, its test accuracy is significantly lower than the rest of the models.

The other 3 models have the same testing accuracies.

Confusion Matrix



Out of the 4 models, Logistic Regression, SVM and KNN performed equally well with test accuracies of 0.833.

Their confusion matrix is shown on the right.

We can observe that these models can very accurately predict true positives, but the accuracy for true negatives is not very good.

Conclusions

To summarize;

- The launch site with the most successful launches is “KSC LC-39A”.
- The success rates of launches at a launch site increases over time, as more launches are made.
- The orbit types with the highest success rates are the ES-L1, GEO, HEO, SSO and VLEO.
- The Logistic Regression model, SVM model and K Nearest Neighbor model can be used to predict whether a future landing will be successful.

Thank you!