IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

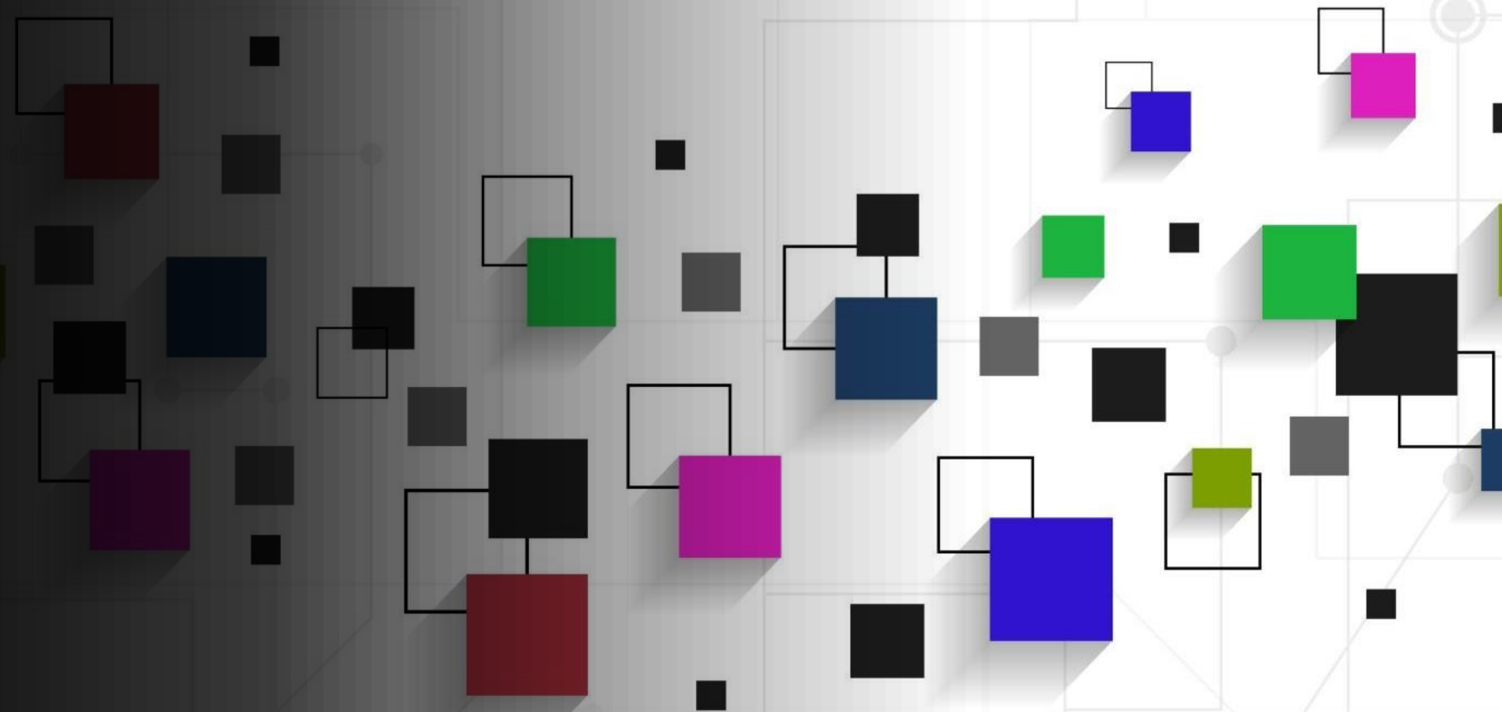
Floribert Ikuzwe
April 13, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

This analysis is aimed at predicting whether the Falcon 9 first stage will land successfully or not, and help us determining the cost of a launch. The reusability of the Falcon 9 launch's first launch makes it cost effective. It costs 62 million dollars, whereas other competitors cost more than 165 million dollars each. In this analysis, we used different data science methods like machine learning techniques to build a predictive model. The datasets used contain information on different features like flight number, launch date, payload type, booster category, launch outcomes, etc. Data cleaning, exploratory data analysis, feature engineering, model selection, and evaluation comprise the methodology used. The results indicate that the predictive model reached an accuracy score of 86% in predicting the successful landing the first stage for Falcon 9.



Introduction

Exploring the space has always been the humanity's never ending and ambitious to-do list. Recently, different companies started to heavily investing in the industry through reusable rockets, making it more affordable for the wealthy population.

One of those companies, SpaceX has revolutionized the industry by reducing the cost dramatically due to its ability to reuse the first stage of the Falcon 9 rockets. This is responsible for initializing the launch and separation from the second stage. After this separation, it returns back to Earth. This landing can succeed or fail but SpaceX has increased the odds of succeeding which means that the first stage can be reused hence the its low cost compared to others.

The objective of this study is to predict the landing outcome of the Falcon 9 first stage, which will enable us to determine the cost of a launch. Through this, valuable insights can be provided to companies that are willing to compete with SpaceX. This will be achieved through building a predictive model using machine learning techniques that can use various features like flight number, launch date, payload type, launch location, booster category, etc to accurately predict whether Falcon 9's first stage will land successfully or not.

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data were corrected through different approaches such as SpaceX API and web scraping.
- Perform data wrangling
  - After getting data from the internet sources, they were processed to remove unnecessary features or observations, dealing with missing values, etc.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models such as logistic regression, decision trees, SVM, and KNN were build, after training the data using train_test_split, tuned using GridSearchCV, and then evaluated through the accuracy score and confusion matrix

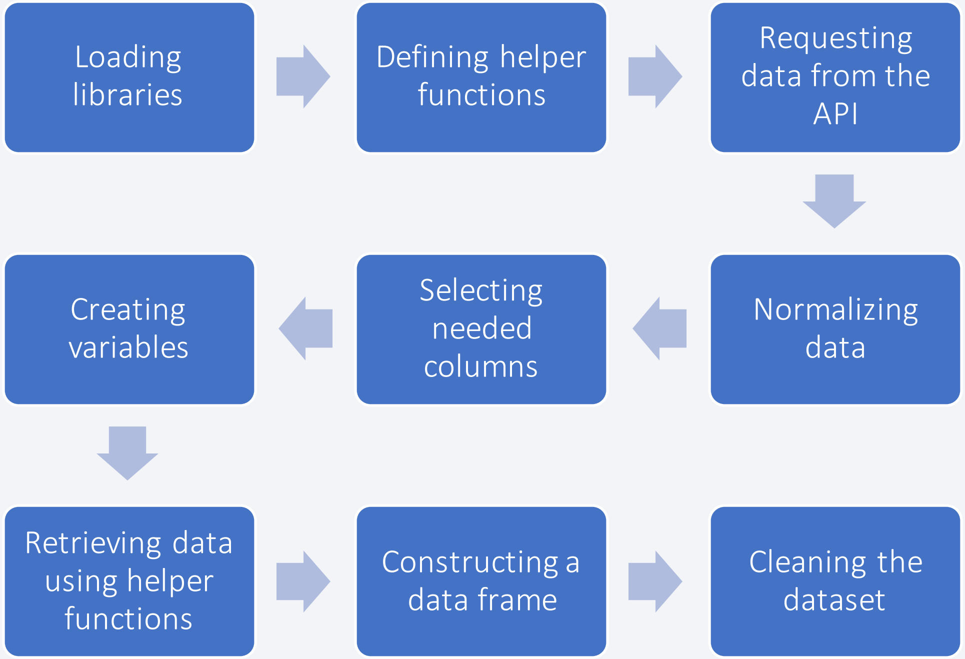
Data Collection

There are 2 datasets used in this study. The first one was gotten from SpaceX's API while the second one was retrieved from wikipedia through web scraping. The details of how both processes were performed are provided in the next slides.

Data Collection – SpaceX API

- After loading required libraries, helper functions were defined to help extract information using identification numbers in the launch data, the retrieved data are normalized using the json_normalize method (to transform them into a dataframe), needed columns were selected, and unneeded rows removed, needed columns/variables are created, needed data are retrieved using the helper functions and a dictionary is used to combine the variables with data into a data frame, only falcon 9 rows are selected and then missing values are dealt with by replacing them with the mean

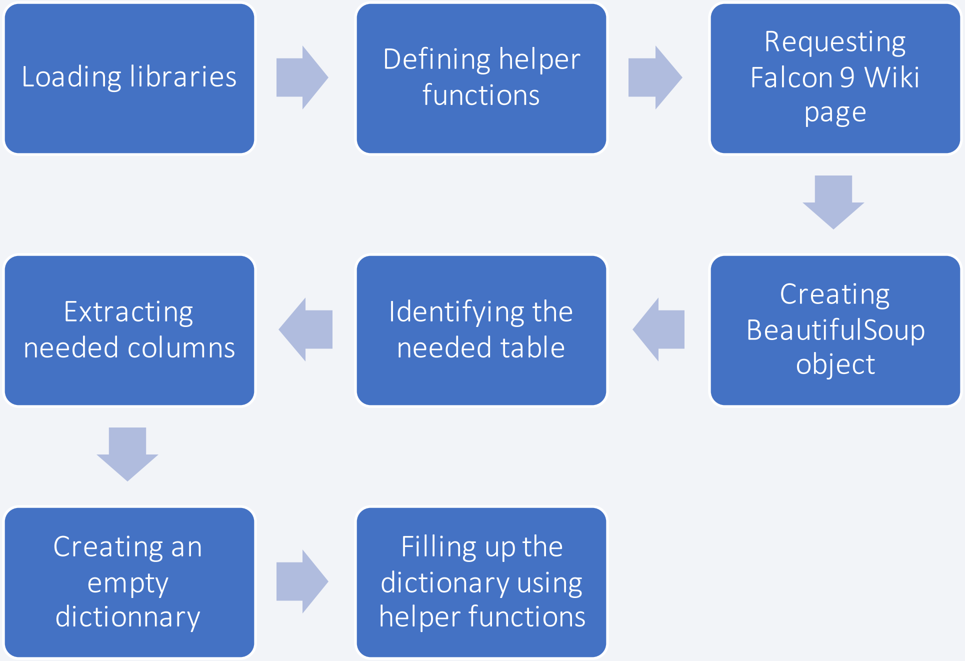
- GitHub URL of the completed SpaceX API calls notebook (https://github.com/Floribert-lang/ibm_cap/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb)



Loading libraries
Defining helper functions
Requesting data from the API
Normalizing data
Selecting needed columns
Creating variables
Retrieving data using helper functions
Constructing a data frame
Cleaning the dataset

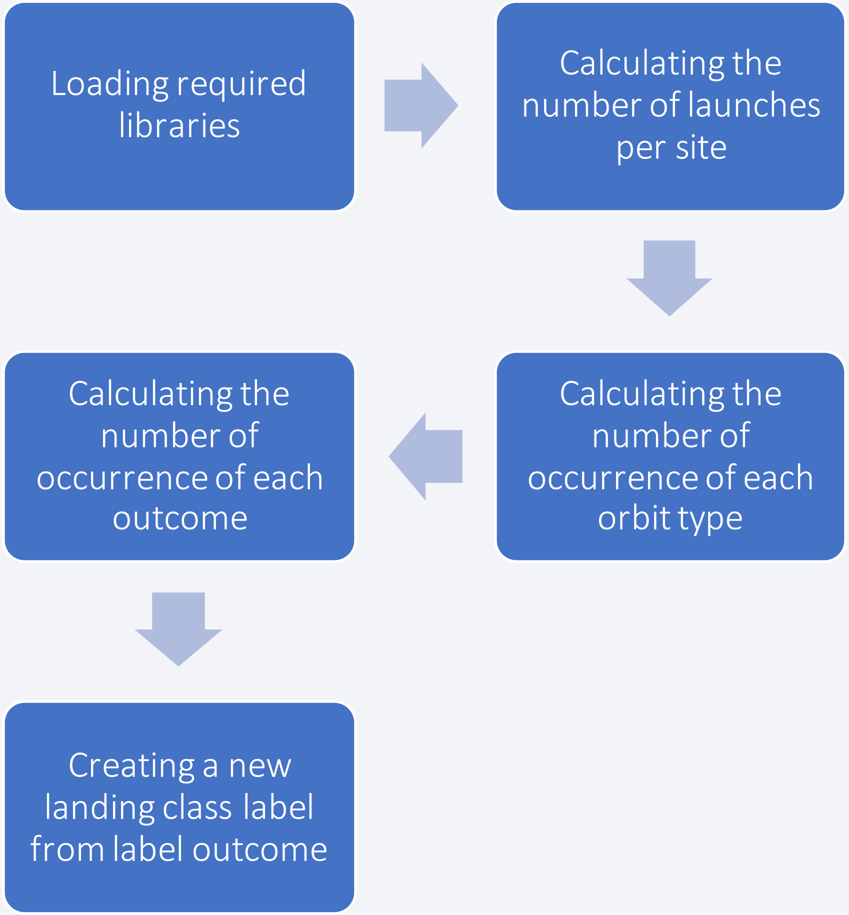
Data Collection - Scraping

- After loading required libraries, helper functions were defined to process web scraped HTML table, the Falcon9 Launch Wiki page is requested from its URL, BeautifulSoup object from the HTML response is created, the needed table was identified and selected from all the tables of the page, column names are extracted, empty dictionary with keys from the extracted column names were created and transformed into a dataframe, and the dictionary is filled up with data extracted using the helper functions.
- GitHub URL of the completed web scraping notebook (https://github.com/Floribert-lang/ibm_cap/blob/main/jupyter-labs-webscraping.ipynb)


Loading libraries
Defining helper functions
Requesting Falcon 9 Wiki page
Creating BeautifulSoup object
Identifying the needed table
Extracting needed columns
Creating an empty dictionary
Filling up the dictionary using helper functions

Data Wrangling

- While some of the data wrangling steps were described in the previous data collections slides, other tasks were performed in this lab too. After loading required libraries, the number of launches per each site was calculated together with the number of occurrence of each orbit as well as the number of occurrence of each outcome, then a new label "landing_class" were created from the outcome column to show 0 if the landing failed and 1 if it succeeded.
- GitHub URL of the completed data wrangling notebook https://github.com/Floribert-lang/ibm_cap/blob/main/jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb).



Loading required libraries

Calculating the number of launches per site

Calculating the number of occurrence of each orbit type

Calculating the number of occurrence of each outcome

Creating a new landing class label from label outcome

EDA with Data Visualization

- To explore the data, difference charts were used. Those charts are described below;
  - FlightNumber vs. PayloadMass: this shows how the both flight number and payload mass affect the landing outcome individually or in combination.
  - FlightNumber vs. LaunchSite: this shows how the both flight number and launch site affect the landing outcome individually or in combination.
  - Payload vs LaunchSite; this shows how the both payload mass and launch site affect the landing outcome individually or in combination.
  - Orbit vs Success rate: this shows which orbits with high/low success rate.
  - FlightNumber vs Orbit: this shows if there is any relationship between FlightNumber and Orbit type
  - Launch success yearly trend: this shows how SpaceX has been doing on these launches across years.
- GitHub URL of the completed EDA with data visualization notebook(https://github.com/Floribert-lang/ibm_cap/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb)

EDA with SQL

- Different queries were performed to understand the data and get insights. Those queries include;
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Displaying the date when the first successful landing outcome in ground pad was achieved.
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Displaying the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Ranking the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL of the completed EDA with SQL notebook (https://github.com/Floribert-lang/ibm_cap/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb)

Build an Interactive Map with Folium

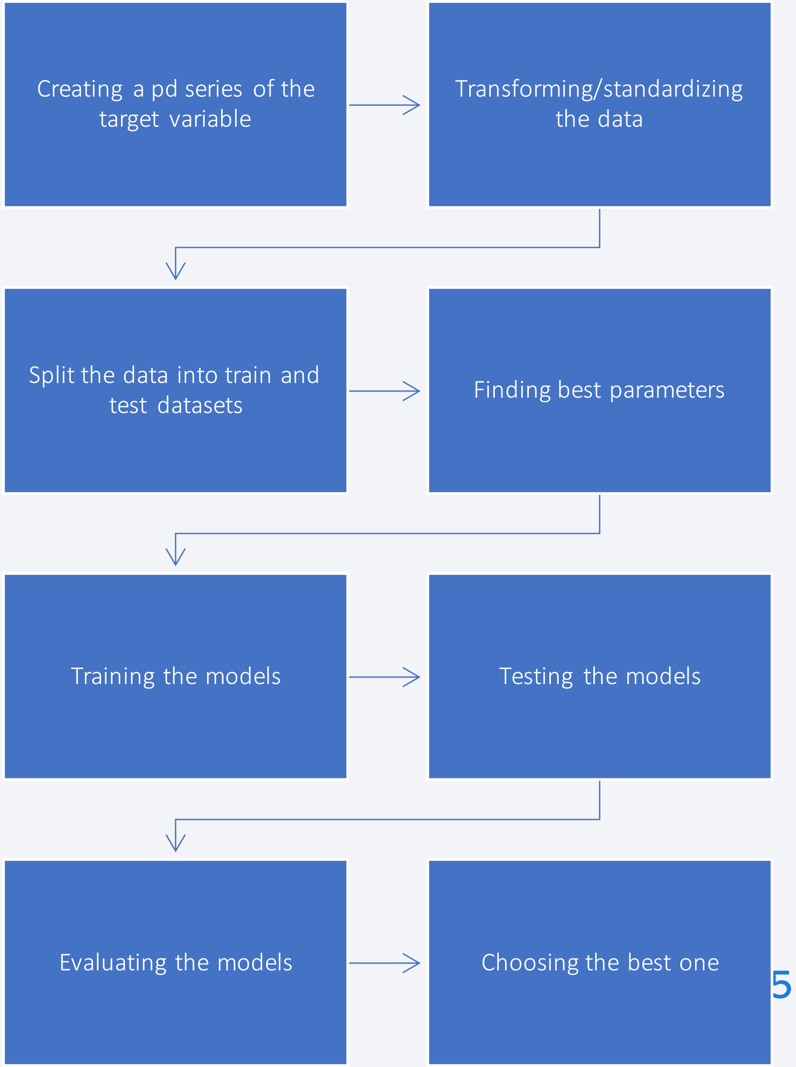
- Different map objects such as markers, circles, lines, etc. were created and added to a folium map;
  - After adding a circle, a popup, and a marker to the Nasa Johnson Center, the same is done to all launch sites for clear visibility.
  - A cluster of launches at each site were created and launches are distinguished using the colors (i.e. green for successes and red for failures)
  - The distance between the launch sites and proximities were calculated and added to the map too.
- GitHub URL of the completed interactive map with Folium map(https://github.com/Floribert-lang/ibm_cap/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb)

Build a Dashboard with Plotly Dash

- Using plotly Dash, plots/graphs and interactions were added to a created dashboard;
  - A dropdown list to enable Launch Site selection (i.e. you can select sites you want to visualize only).
  - A callback function for updating the pie chart (i.e. anytime the site is selected from the dropdown field, the pie chart get updated)
  - A slider to select payload range
  - A callback function for \`site-dropdown\` and \`payload-slider\` as inputs, \`success-payload-scatter-chart\` as output for updating the scatter chart when a site is changed from the drop down or any change happens to the payload slider.
- GitHub URL of the completed Plotly Dash lab(https://github.com/Floribert-lang/ibm_cap/blob/main/plotly_Dash_lab.py)

Predictive Analysis (Classification)

- The models were built, evaluated, improved, and found the best performing classification one through the following steps
  - After reading the data, the target column (class), was transformed into a numpy array then saved as a pd series
  - The data were transformed using standardScaler then split into train and test dataset using train_test_split, then the models were trained and hyperparameters were selected using the function GridSearchCV.
  - The used models are logistic regression, SVM, decision tree, and KNN. After training and testing each model, they were evaluated using the accuracy score and the confusion matrix.
- GitHub URL of the completed predictive analysis lab(https://github.com/Floribert-lang/ibm_cap/blob/main/Machine%20Learning%20Prediction.ipynb)


Creating a pd series of the target variable
Transforming/standardizing the data
Split the data into train and test datasets
Finding best parameters
Training the models
Testing the models
Evaluating the models
Choosing the best one

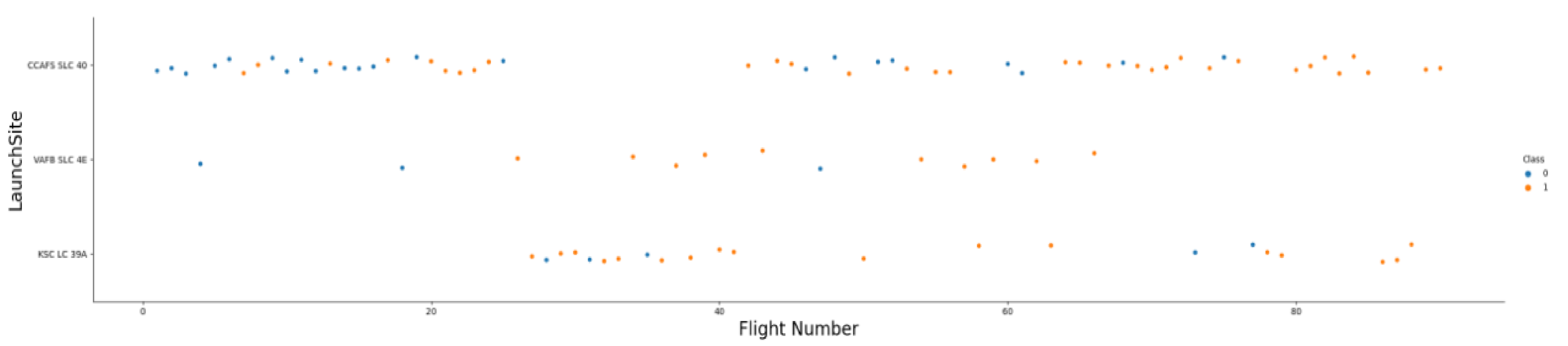
Results

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

Section 2

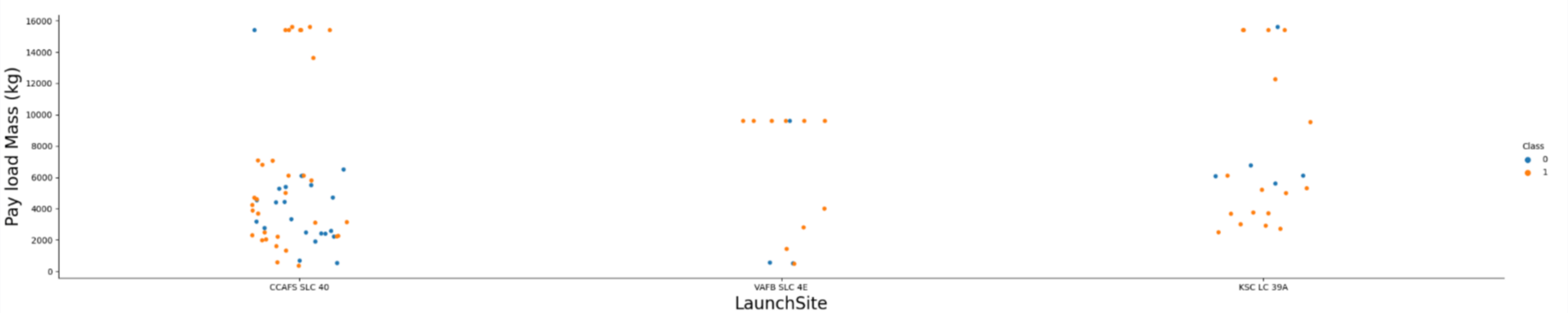
Insights drawn from EDA

Flight Number vs. Launch Site



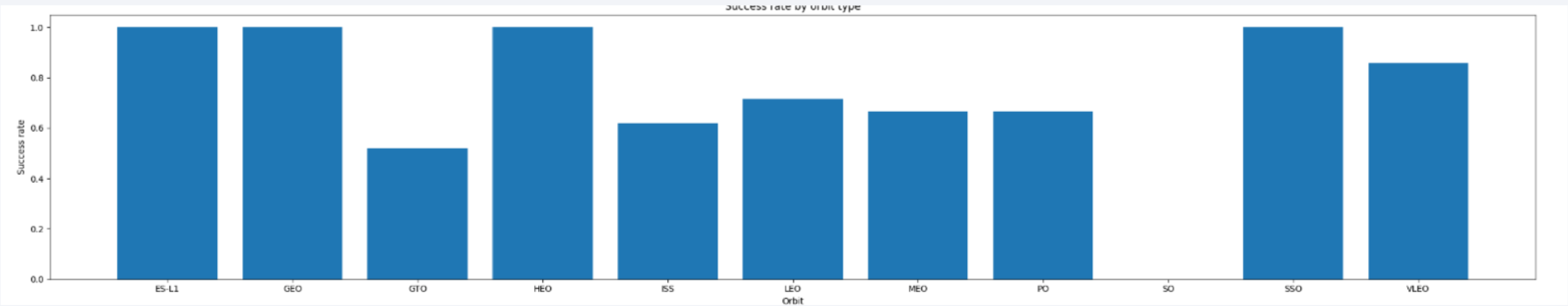
- According to this plot, the number of successes have been increasing with the number of launches across all launch sites.
- CCAFS SLC 40 has more flights than other sites as well as more failures rate in the beginning more than other sites when they were starting. Other sites might have learned from its early mistakes to improve their operations.
- KSC LC 39A is the second one with most launches while VAFB SLC 4E site is the one with few launches.

Payload vs. Launch Site



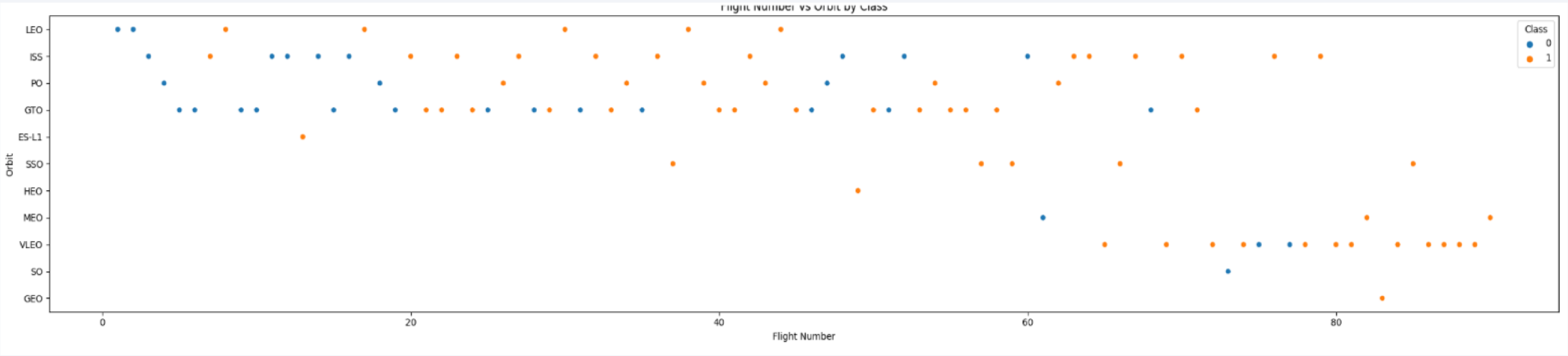
- In general, most of the launches were below 8,000kg across sites except for VAFB SLC 4E where most of its launches has 10,000 kgs.
- Also, VAFB SLC 4E site never had launch above 10,000 kgs.

Success Rate vs. Orbit Type



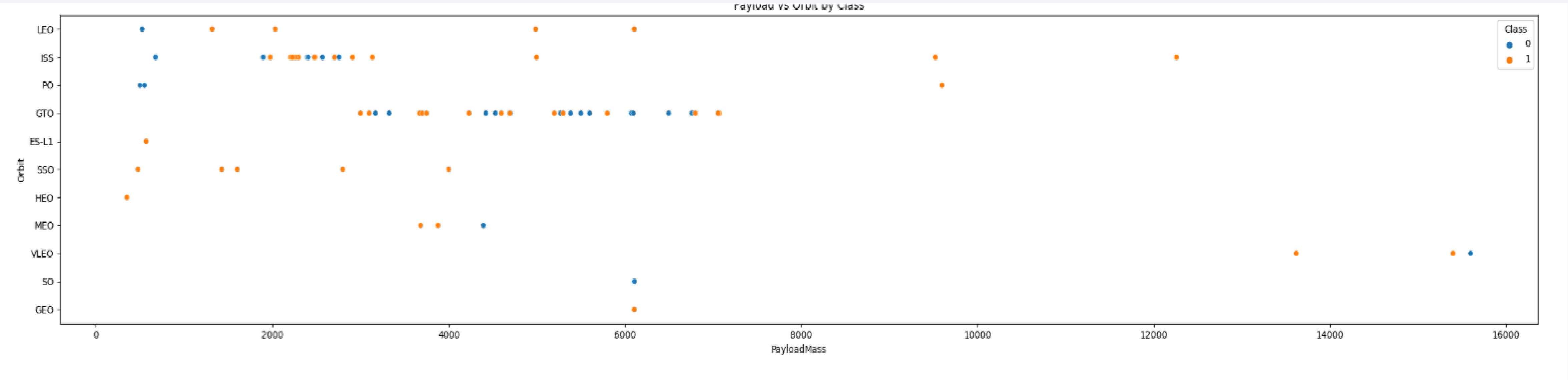
- ES-L1, GEO, HEO, and SSO have the highest success rate.
- Apart from SO with 0% success rate, GTO comes next on the list of losers.

Flight Number vs. Orbit Type



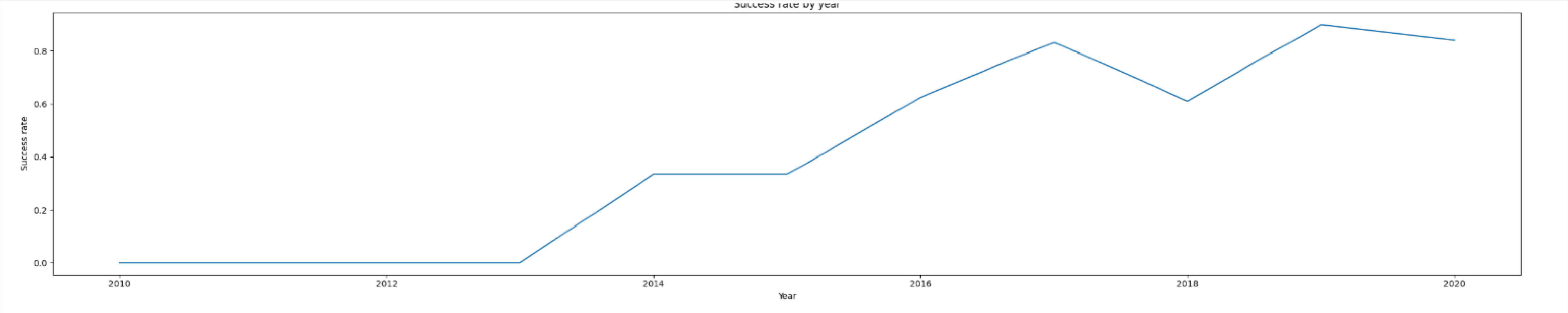
- LEO is the first orbit to be visited.
- Some orbits like LEO, and VLEO seems to have a relationship between success and flight number while in others like GTO and ISS seems to not have such relationship.

Payload vs. Orbit Type



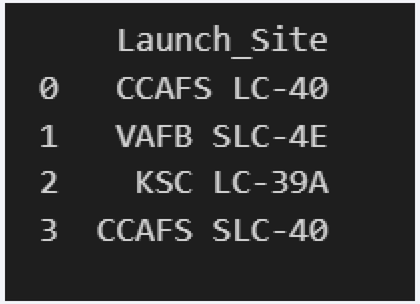
- There seems to be a high success rate in heavy payload (i.e. >8k kgs) and VLEO is the one that received more heavy payloads than the others.
- There seems to be a relationship between success rate and increase in payload except in GTO and ISS orbits.

Launch Success Yearly Trend



- The success rate increases (at variable velocity though) for as time goes except near 2018 and 2020. The early years had a constant low success rate.

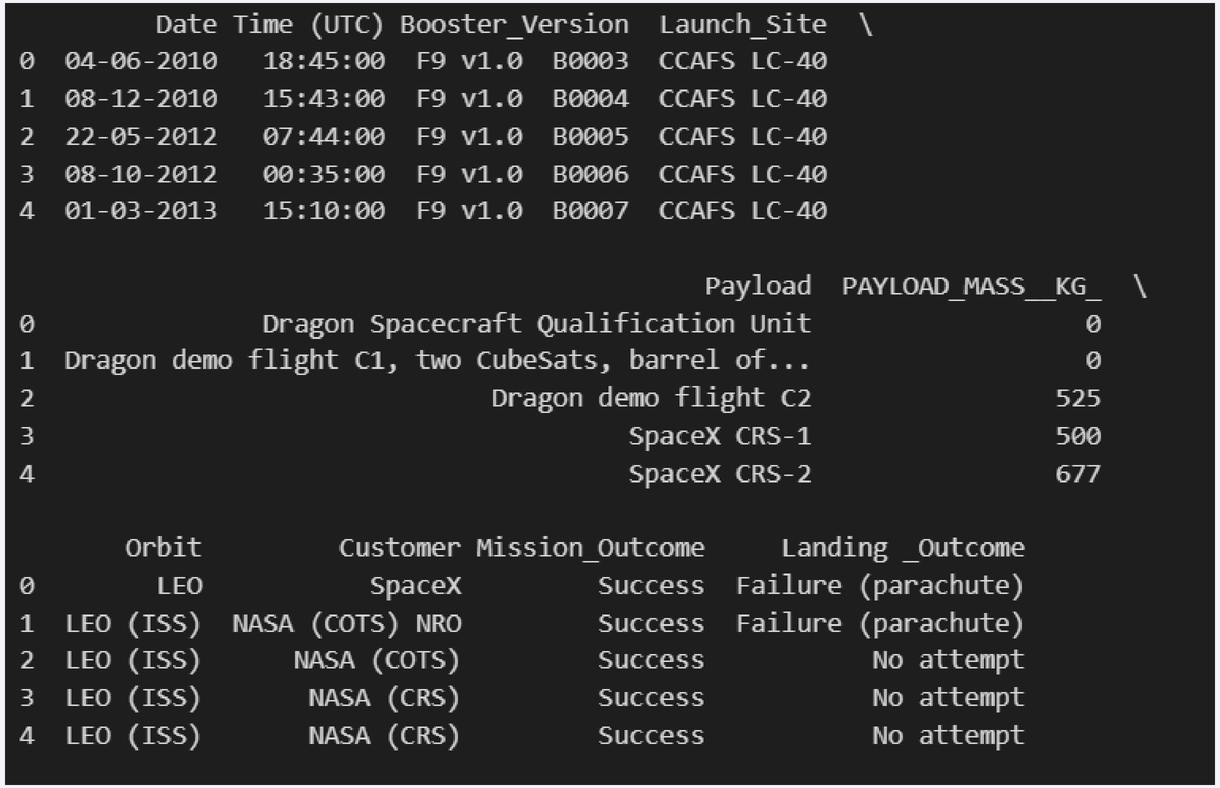
All Launch Site Names



```
     Launch_Site
0    CCAFS LC-40
1    VAFB SLC-4E
2     KSC LC-39A
3   CCAFS SLC-40
```

- There are 4 unique launch sites

Launch Site Names Begin with 'CCA'



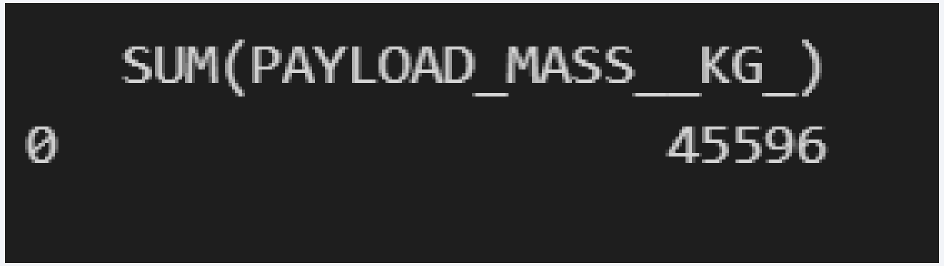
```
         Date Time (UTC) Booster_Version  Launch_Site  \
0  04-06-2010   18:45:00   F9 v1.0  B0003  CCAFS LC-40
1  08-12-2010   15:43:00   F9 v1.0  B0004  CCAFS LC-40
2  22-05-2012   07:44:00   F9 v1.0  B0005  CCAFS LC-40
3  08-10-2012   00:35:00   F9 v1.0  B0006  CCAFS LC-40
4  01-03-2013   15:10:00   F9 v1.0  B0007  CCAFS LC-40

                                             Payload  PAYLOAD_MASS__KG_  \
0                Dragon Spacecraft Qualification Unit                  0
1  Dragon demo flight C1, two CubeSats, barrel of...                  0
2                              Dragon demo flight C2                525
3                                       SpaceX CRS-1                500
4                                       SpaceX CRS-2                677

       Orbit         Customer Mission_Outcome     Landing _Outcome
0        LEO           SpaceX         Success  Failure (parachute)
1  LEO (ISS)  NASA (COTS) NRO         Success  Failure (parachute)
2  LEO (ISS)      NASA (COTS)         Success           No attempt
3  LEO (ISS)       NASA (CRS)         Success           No attempt
4  LEO (ISS)       NASA (CRS)         Success           No attempt
```

- 5 records where launch sites begin with `CCA`

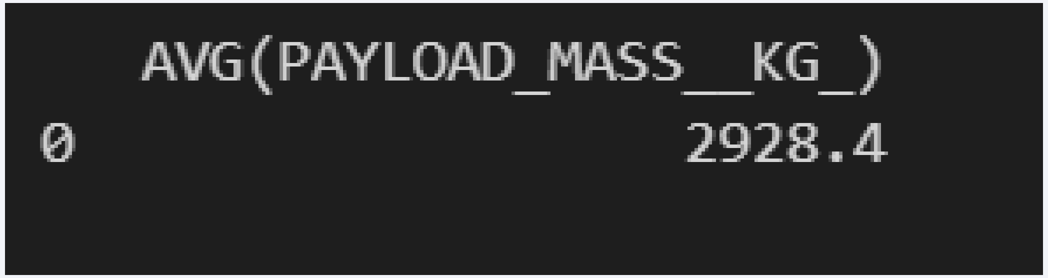
Total Payload Mass



```
   SUM(PAYLOAD_MASS__KG_)
0                   45596
```

- The total payload carried by boosters from NASA

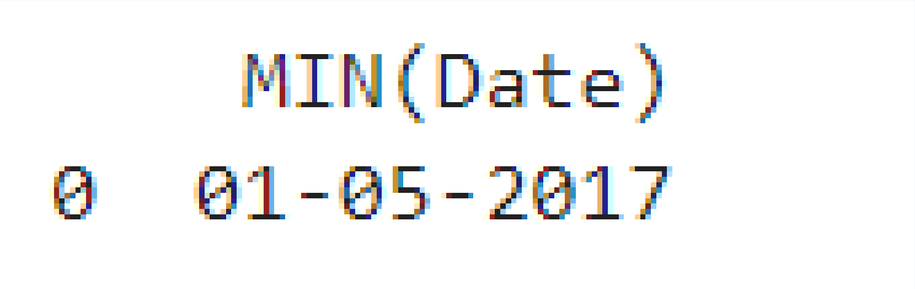
Average Payload Mass by F9 v1.1



```
   AVG(PAYLOAD_MASS__KG_)
0                  2928.4
```

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

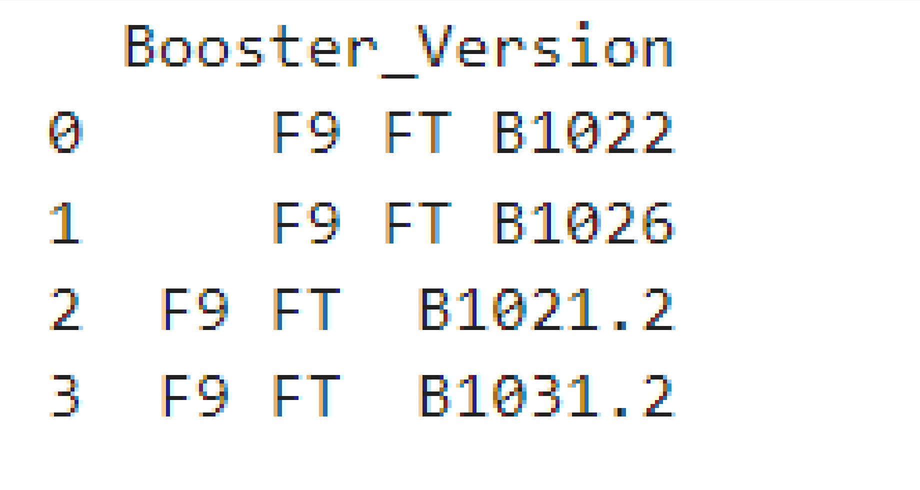
First Successful Ground Landing Date



```
     MIN(Date)
0  01-05-2017
```

- The dates of the first successful landing outcome on ground pad

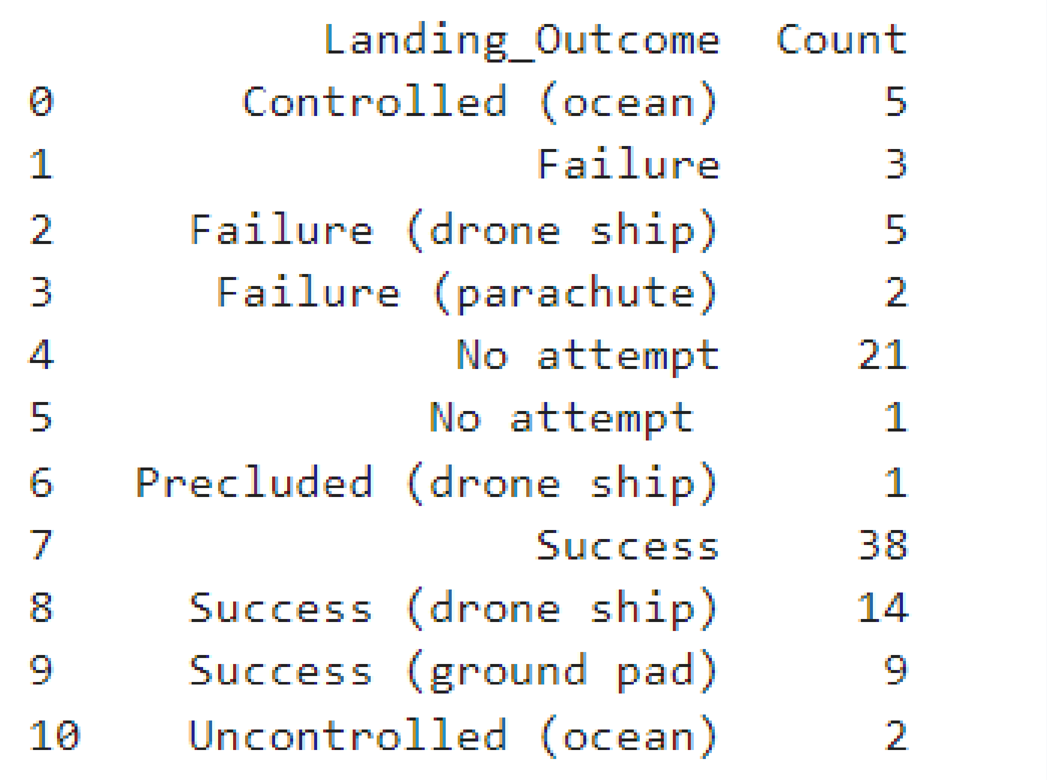
Successful Drone Ship Landing with Payload between 4000 and 6000



```
  Booster_Version
0     F9 FT B1022
1     F9 FT B1026
2  F9 FT  B1021.2
3  F9 FT  B1031.2
```

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

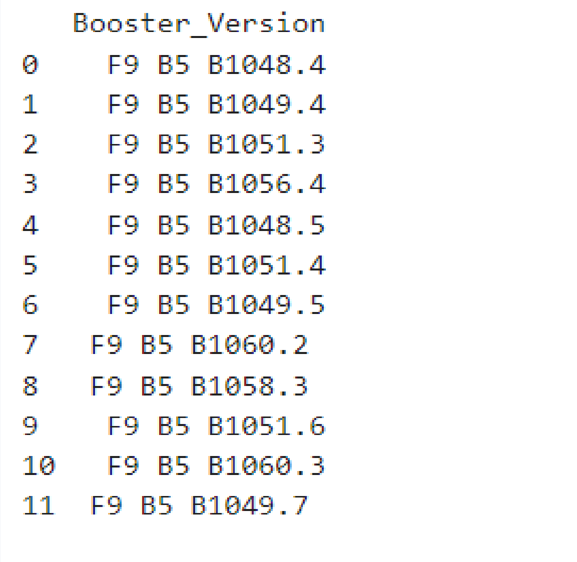
Total Number of Successful and Failure Mission Outcomes



```
           Landing_Outcome  Count
0       Controlled (ocean)      5
1                  Failure      3
2     Failure (drone ship)      5
3      Failure (parachute)      2
4               No attempt     21
5              No attempt       1
6   Precluded (drone ship)      1
7                  Success     38
8     Success (drone ship)     14
9     Success (ground pad)      9
10    Uncontrolled (ocean)      2
```

- The total number of successful and failure mission outcomes

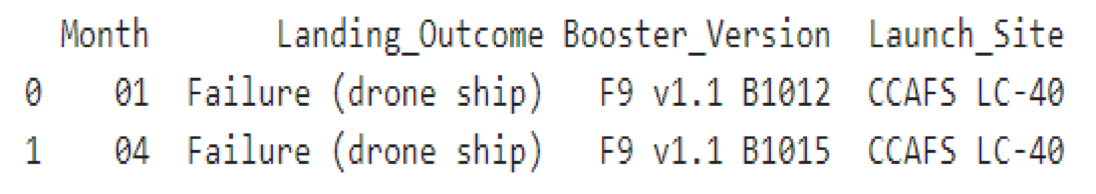
Boosters Carried Maximum Payload



```
   Booster_Version
0     F9 B5 B1048.4
1     F9 B5 B1049.4
2     F9 B5 B1051.3
3     F9 B5 B1056.4
4     F9 B5 B1048.5
5     F9 B5 B1051.4
6     F9 B5 B1049.5
7    F9 B5 B1060.2
8    F9 B5 B1058.3
9     F9 B5 B1051.6
10    F9 B5 B1060.3
11   F9 B5 B1049.7
```

- Names of the booster which have carried the maximum payload mass

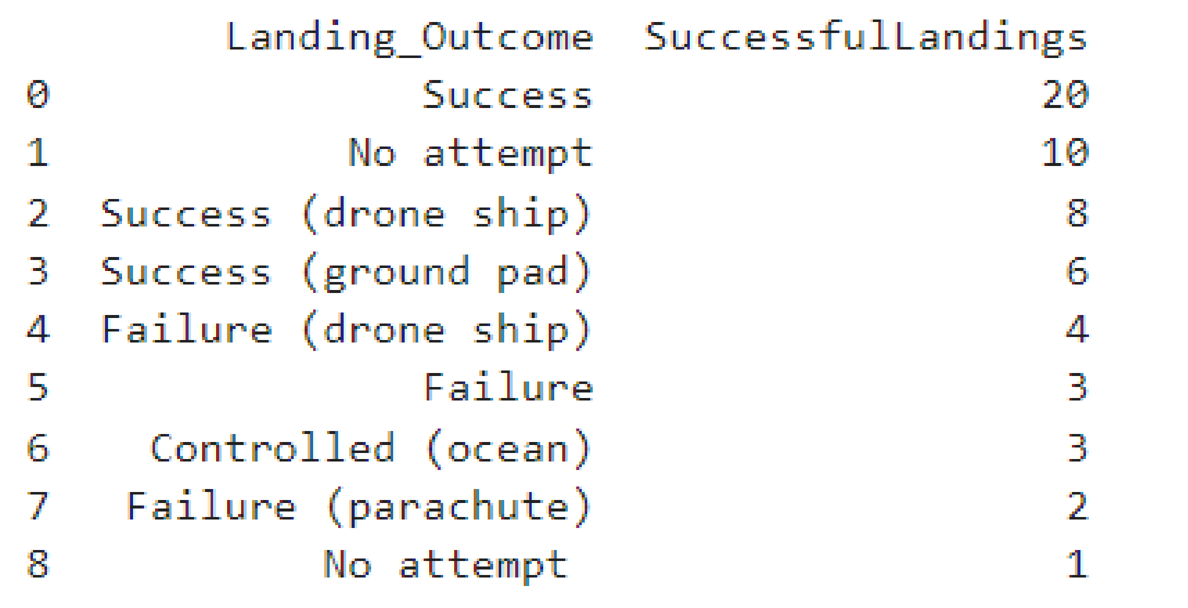
2015 Launch Records



```
  Month       Landing_Outcome Booster_Version  Launch_Site
0    01  Failure (drone ship)   F9 v1.1 B1012  CCAFS LC-40
1    04  Failure (drone ship)   F9 v1.1 B1015  CCAFS LC-40
```

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

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



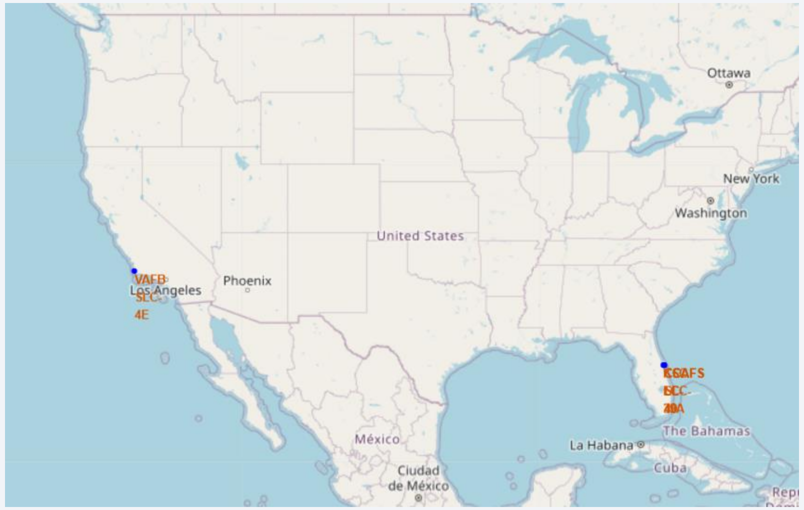
```
         Landing_Outcome  SuccessfulLandings
0                Success                  20
1             No attempt                  10
2   Success (drone ship)                   8
3  Success (ground pad)                    6
4   Failure (drone ship)                   4
5                Failure                   3
6     Controlled (ocean)                   3
7    Failure (parachute)                   2
8            No attempt                    1
```

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

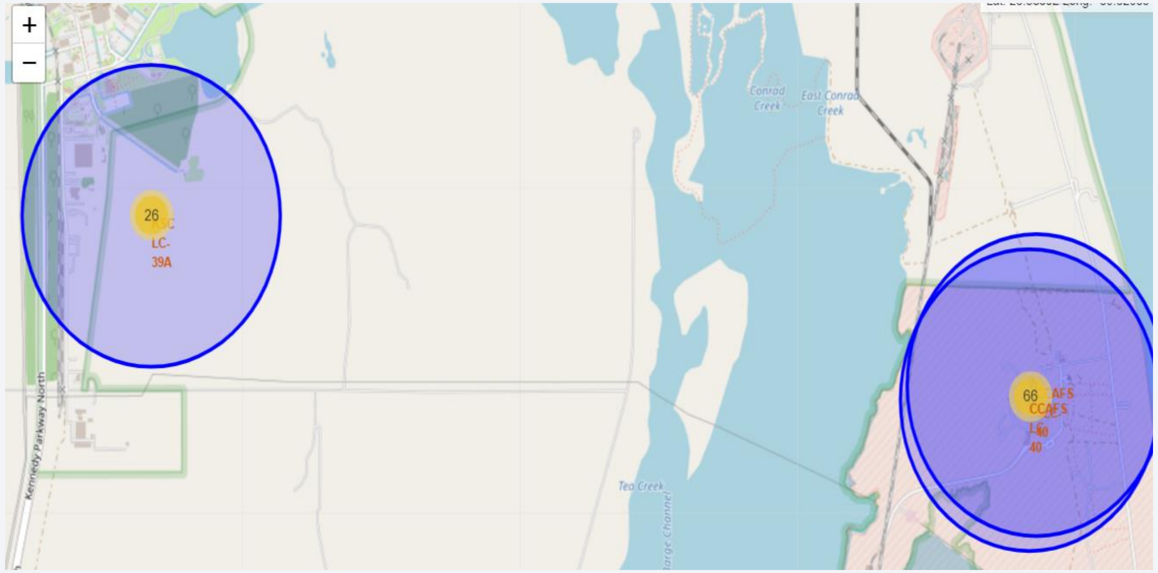
Section 3

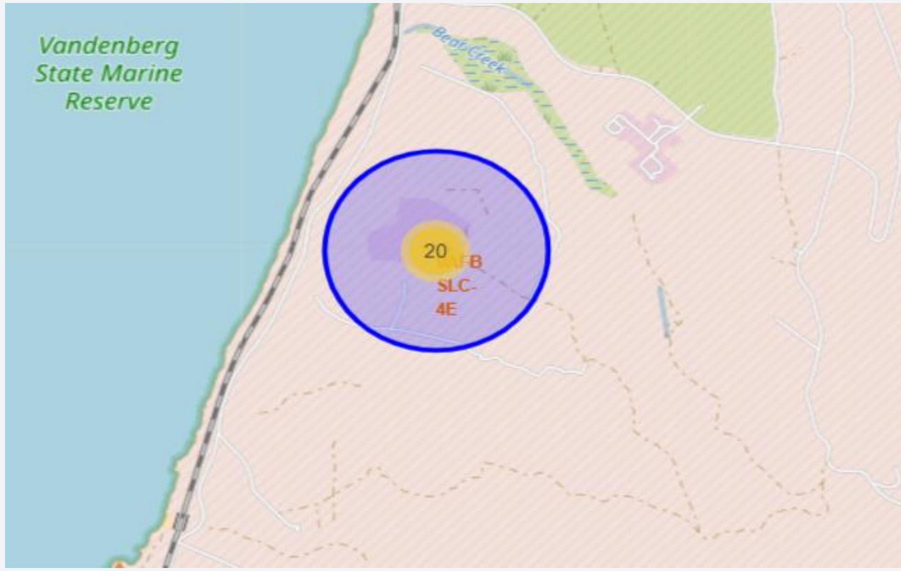
Launch Sites Proximities Analysis

Launch Sites Visualization

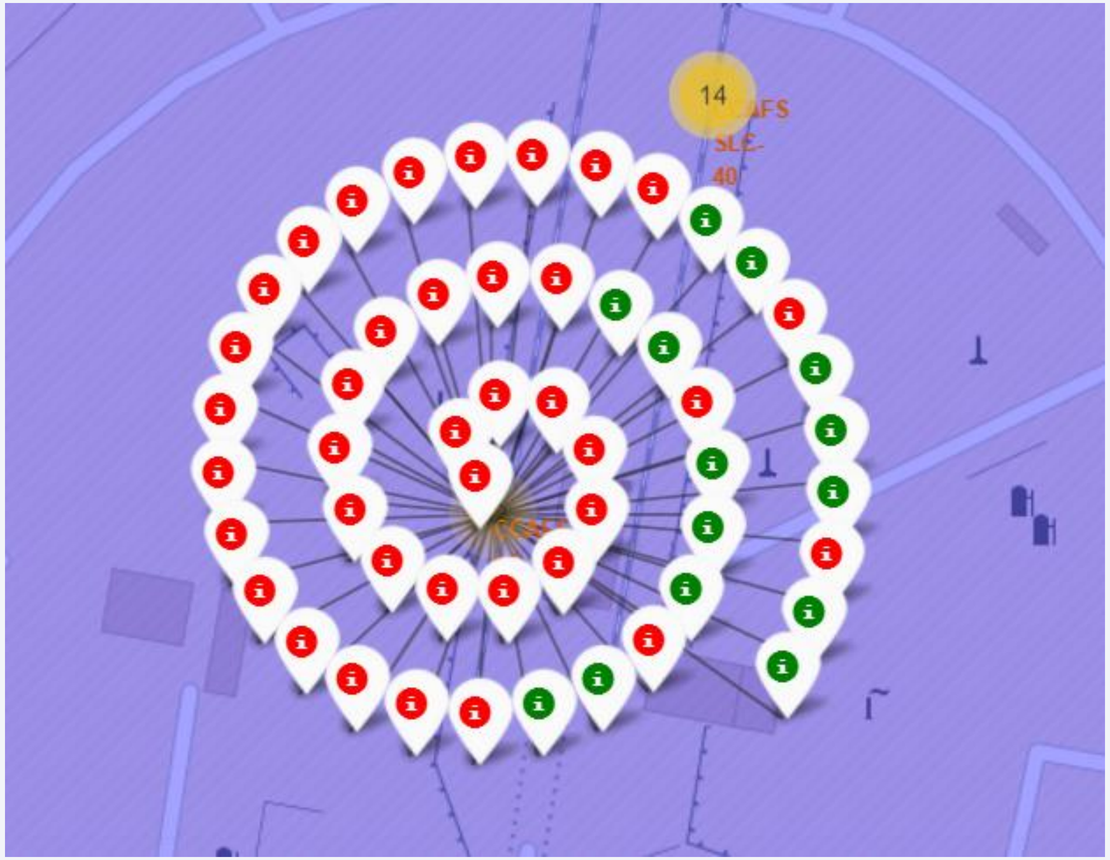


- This screenshot at the top tries to show all sites in one view but 3 of them are very close to each other, they are seen as one. IN the 2 below screenshots, I tried to zoom in so that sites can be seen separately.
- All launch sites are close to the coastline, though KSC LC-39A is not as close as the other 3.



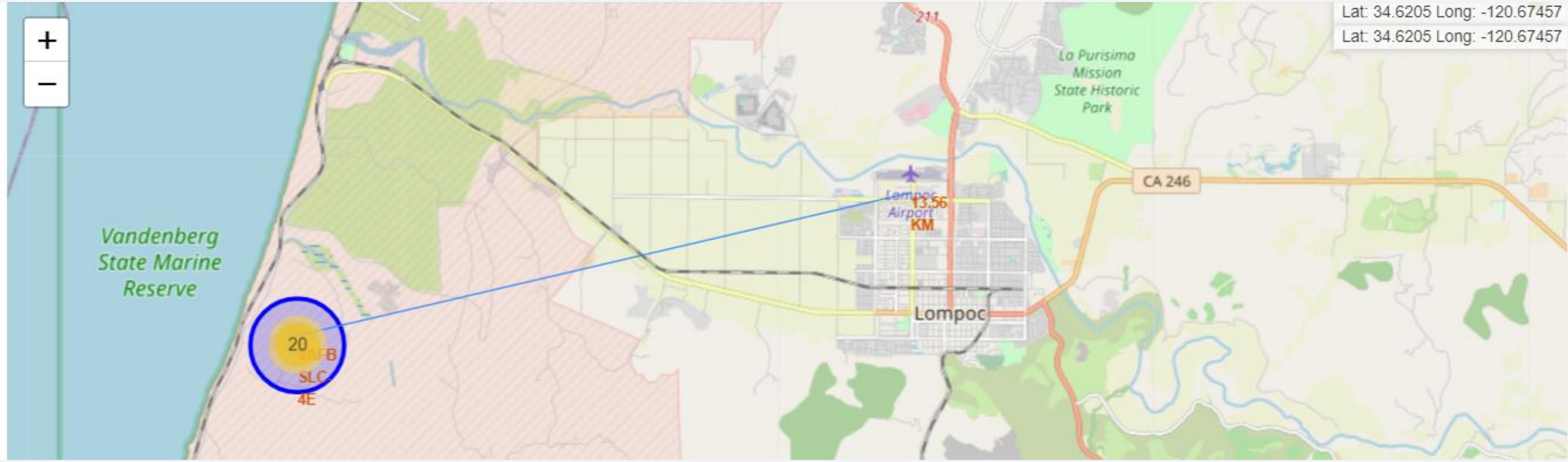


Cluster of launch outcomes in one of the sites



- The screenshot shows the numbers of launches performed at this specific site distinguished by the colors where green mean success while red is failure. Apparently, this site has more failures that successes.

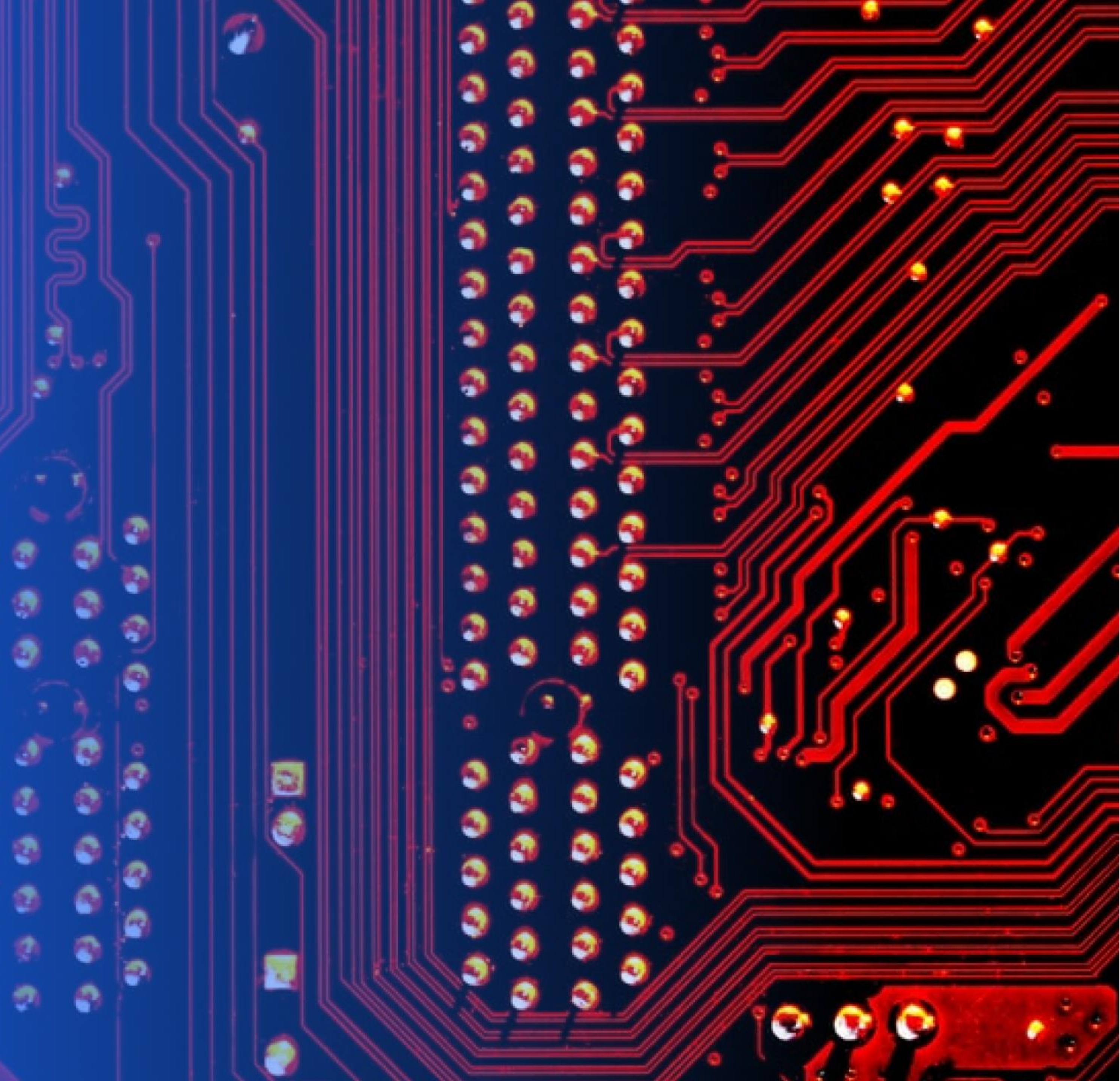
Distance between one site and its nearest Airport


+
−
Lat: 34.6205 Long: -120.67457
Lat: 34.6205 Long: -120.67457
211
La Purisima Mission State Historic Park
CA 246
Lompoc Airport
13.56 KM
Lompoc
Vandenberg State Marine Reserve
20
VAFB SLC-4E

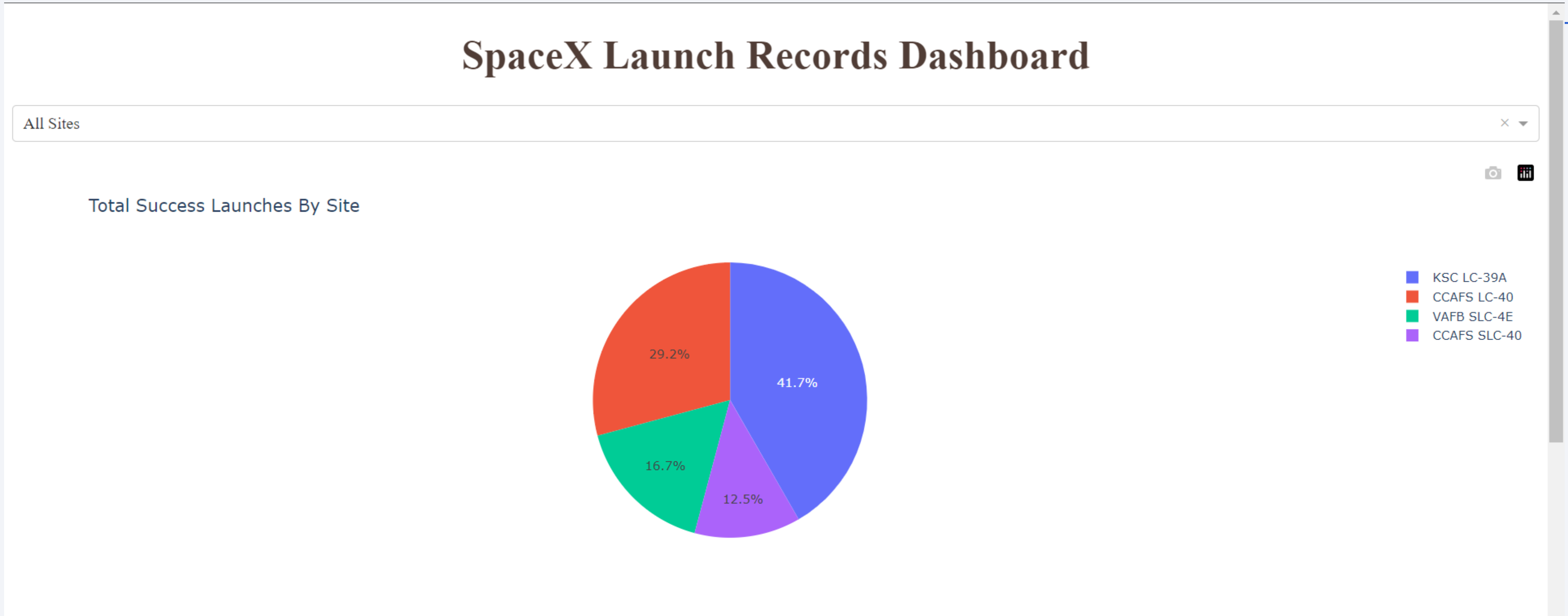
- This screenshot shows the distance (13.56 KM) between VAFB SLC-4E launch site and its nearest airport (Lompoc Airport)

Section 4

Build a Dashboard with Plotly Dash



Launch success count for all sites



SpaceX Launch Records Dashboard

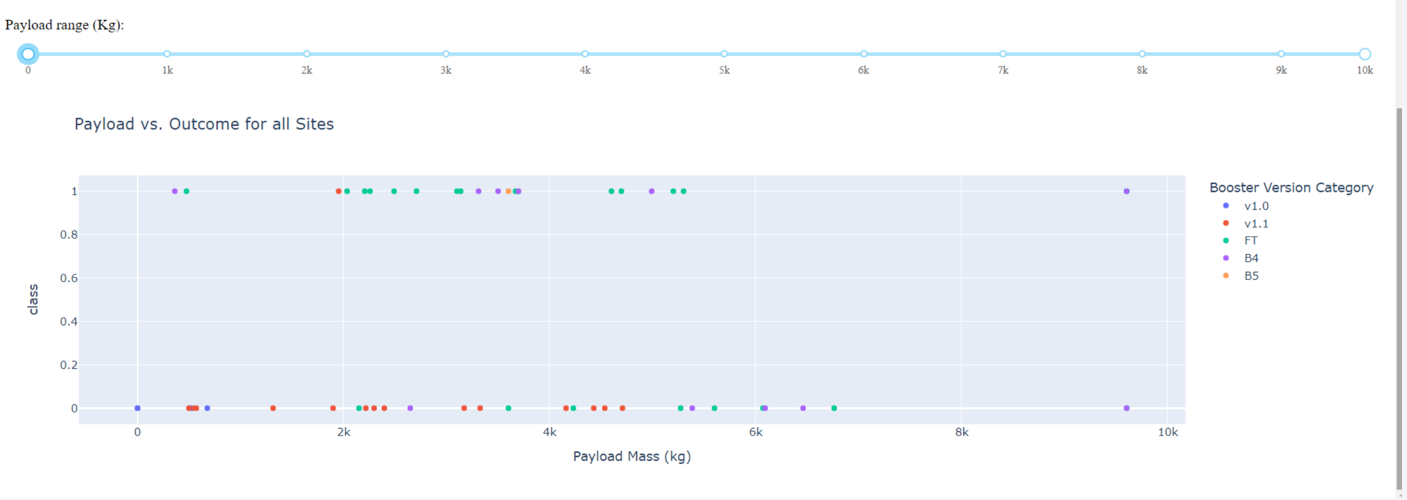
All Sites

Total Success Launches By Site

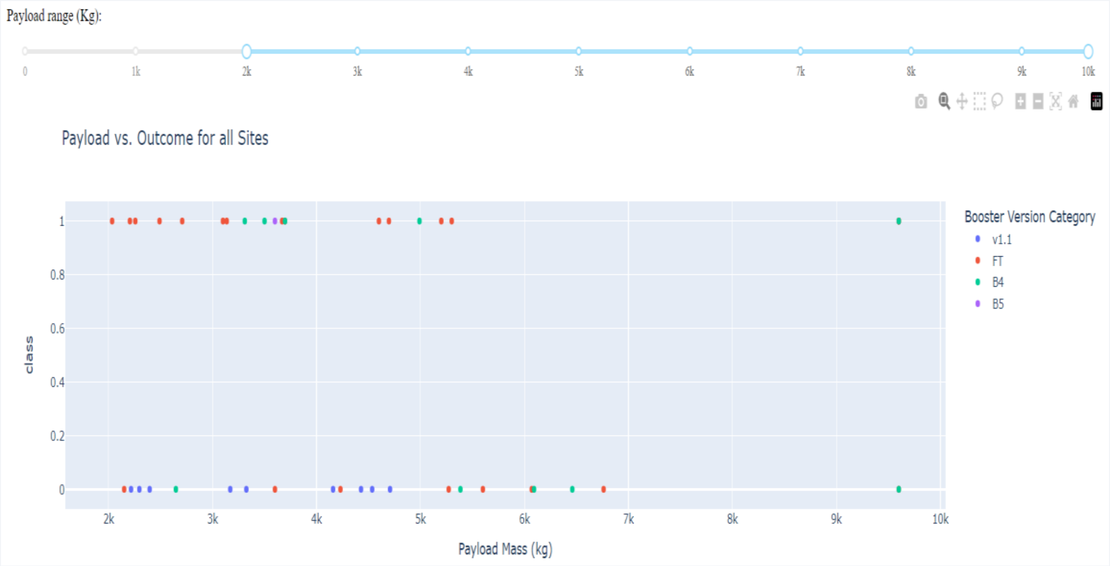
- Almost a half(41.7%) of all successes come from the KSC LC –39A launch site followed by CCAFS LC-40 with 29.9% while the other 2 shares the remaining third of the total successes.

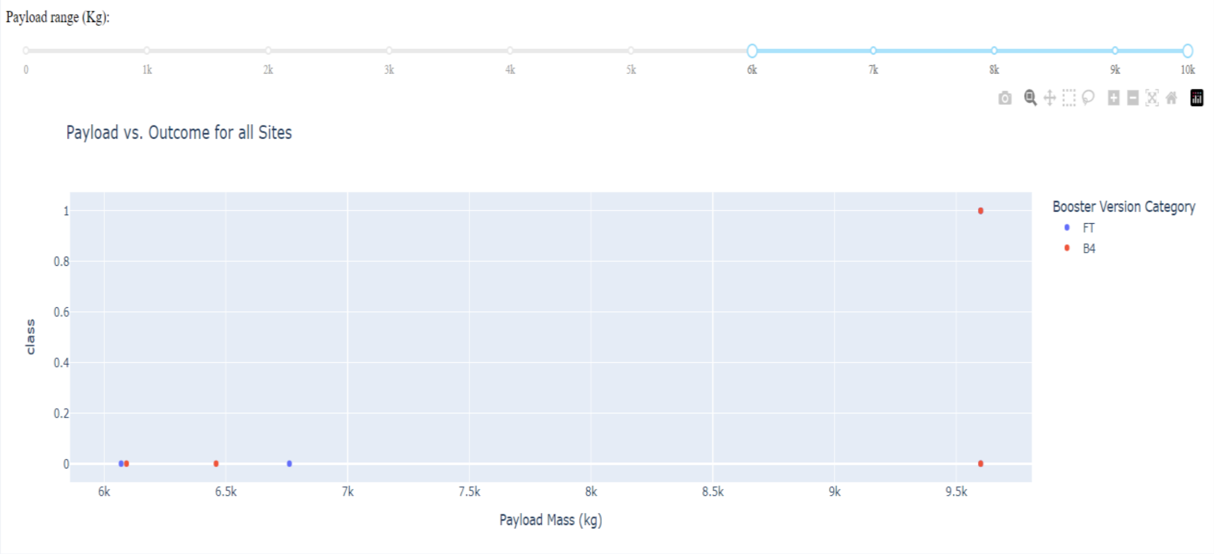
Site with highest launch success

Payload vs. Launch Outcome for all sites



- The first screenshot shows the outcome across all sites while the other 2 at the bottom shows the outcome ad varied payload masses.

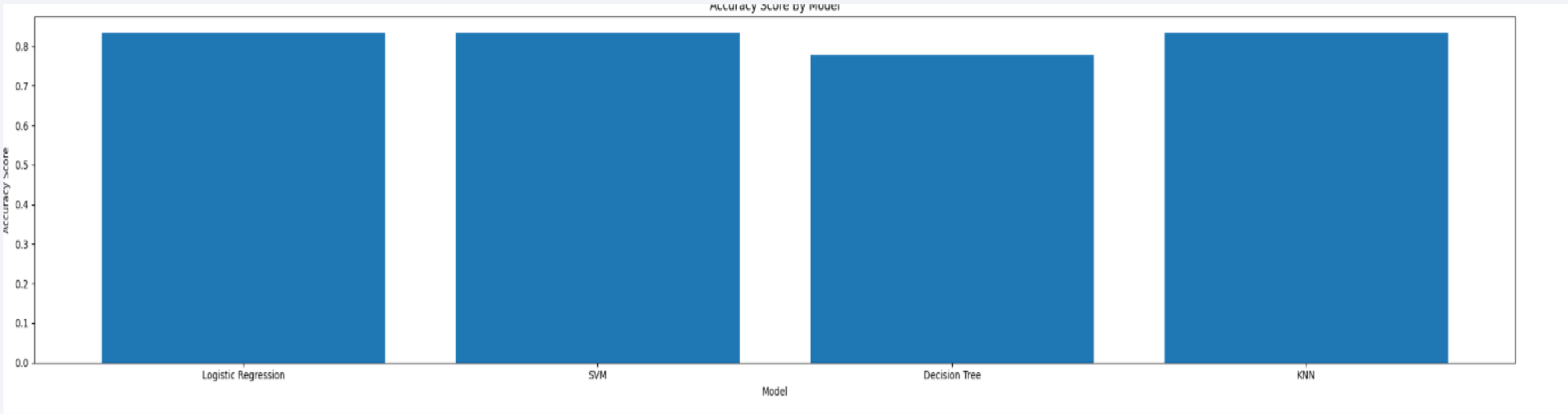




Section 5

Predictive Analysis (Classification)

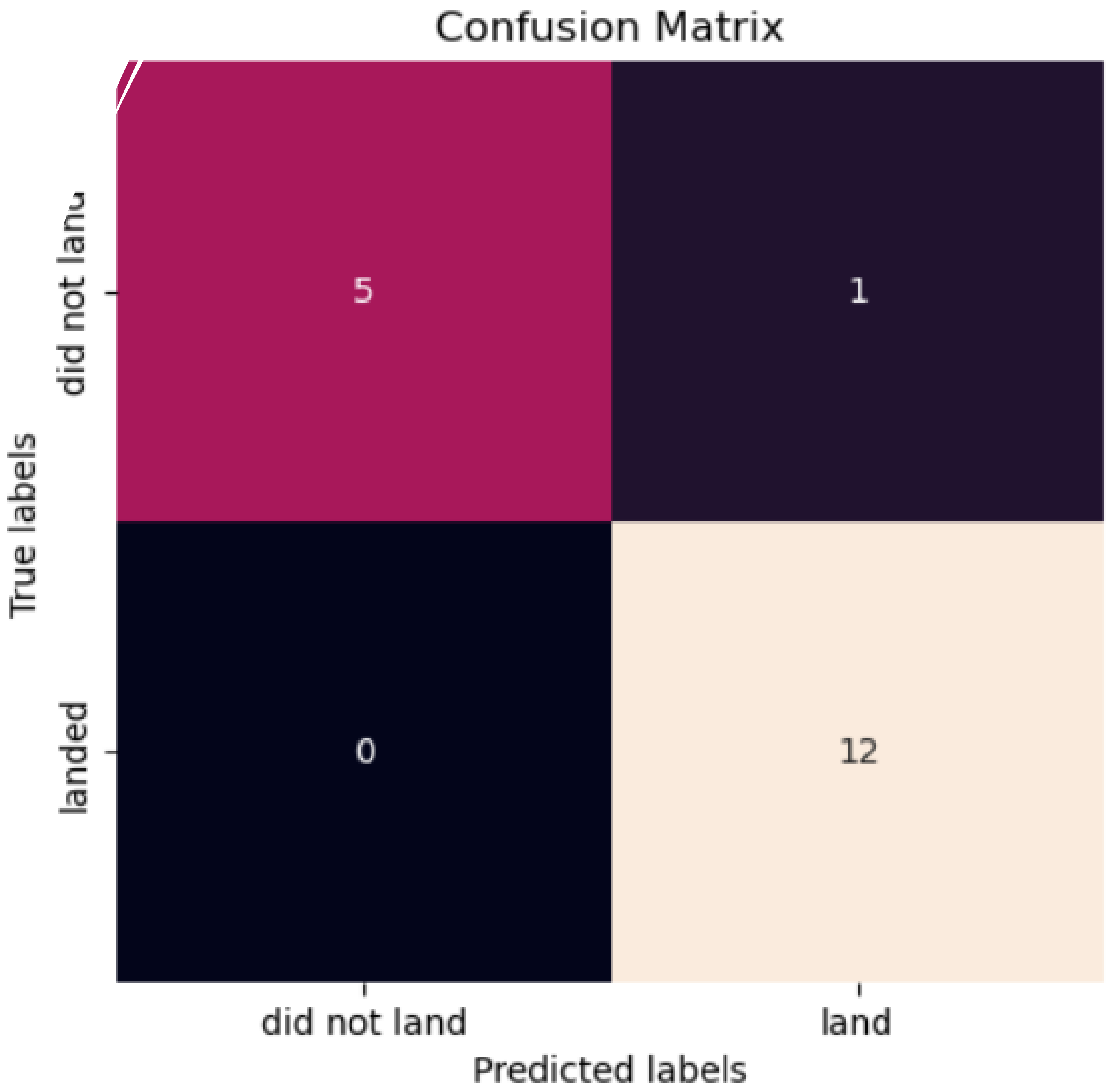
Classification Accuracy

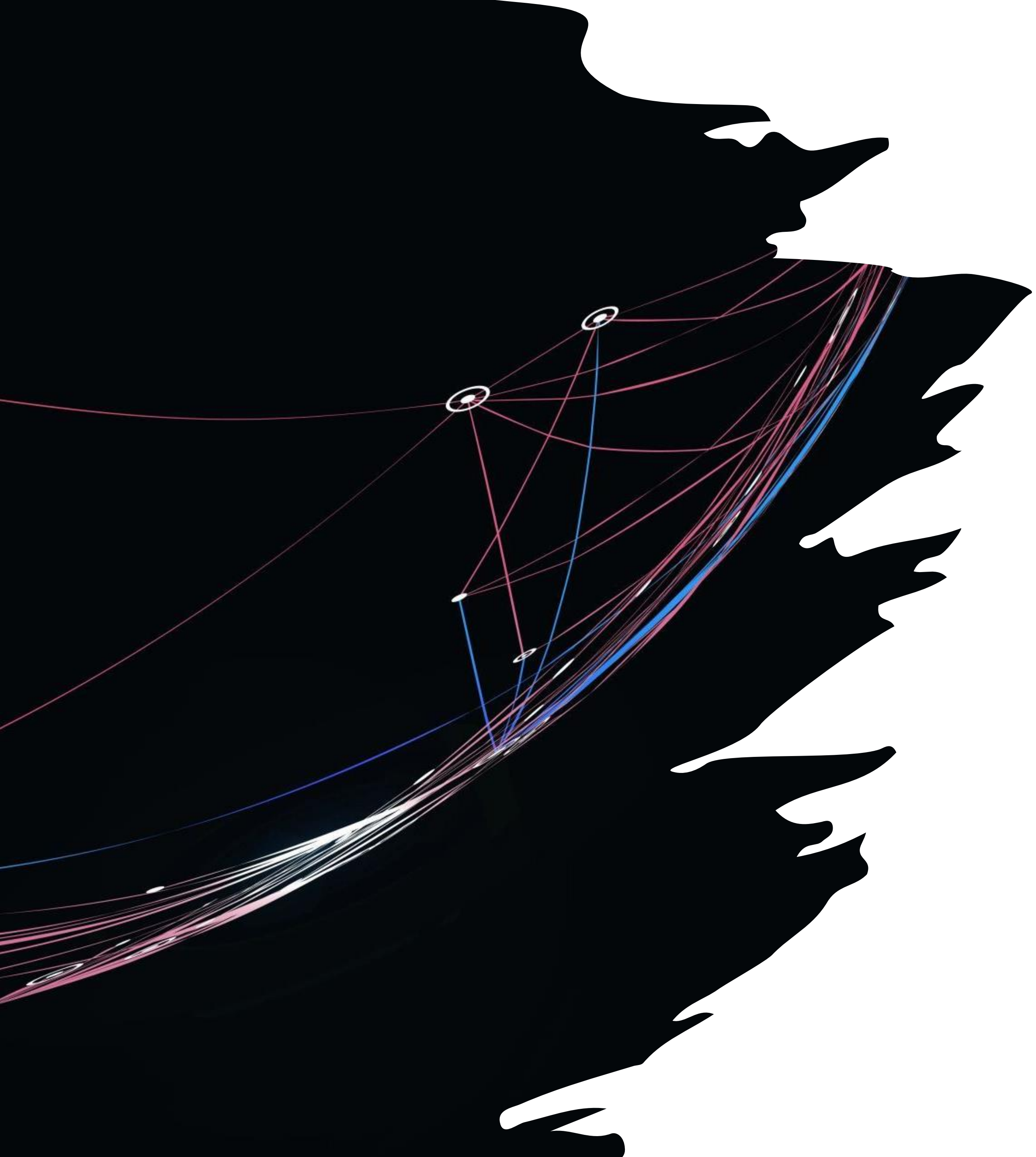


- Apparently, except the decision tree model with the score of 0.77, other 3 have the same score, 0.83 which means that all of the can provide the same predictions.

Confusion Matrix

- This is the confusion matrix of the LogisticRegression model.
- This model is able to distinguish different classes except one false positive.





Conclusions

In conclusion, in the race to the space, SpaceX with its Falcon9 rocket has revolutionized the industry by making is more affordable through the reusability of its first stage between the two. Using the data provided by SpaceX fetched through its API or scraped from the web (Wikipedia), we have developed a machine learning model to predict the success of the Falcon 9 first stage landing based on a variety of factors such as launch date, payload mass, flight number, launch location, etc. 3 of our 4 built models, achieved an accuracy of 83%, which demonstrates its effectiveness in predicting the landing outcome of the Falcon 9 first stage.

The accurate prediction of the landing outcome of different launches has many applications including determining the cost of a launch and make informed financial decisions like bidding in different competitions especially against SpaceX.

Finally, this study determines that in addition to rocket science, machine learning too can provide a big contribution in predicting the success or failure of different systems and demonstrates the importance of data analysis and modeling in making informed decisions in the space industry as well as other fields.

Thank you!