IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bastián Astudillo Fica
June 3rd, 2023

Outline

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

Executive Summary

The commercial space age is here, which is why SPACEX . SPACEX is on a mission to make space travel possible for mankind.

Methodologies

- Data Collection API
- Data Wrangling
- EDA using SQL & Visualization
- Interactive Visual Analysis by creating Dashboard
- Machine Learning Predictive Analysis

Results

- Exploratory Data Analytics
- Interactive Analytics in screenshots
- Models to be used LOGISTIC, SVM, Decision Tree and KNN regression classifiers.
- The method that obtains the best results using training data

Introduction

The goal of SPACE Y is to create the technology necessary for safe space travel. This idea has always been in mankind and today it is being achieved.

Space Y advertises on its website launches of the Falcon 9 rocket at a cost of $62 million; SPACE Y can save millions without each launch of our Eagle rocket because we can reuse its first stage.

In addition, we can determine if our competitor's first stage will land and determine the cost of a launch using Data Science and Machine Learning models.

Section 1

Methodology

Methodology

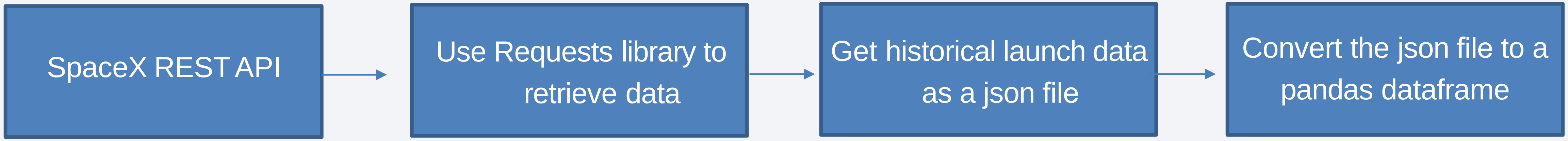
Executive Summary

- Data collection methodology:
    - The data was gathered through SpaceX API & Webscrapping from wikipedia
- Perform data wrangling
    - The collected data are retrieved in JSON format and HTML tables. After data extraction, the data were transformed into a pandas data frame for visualization and analysis..
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - We use machine learning algorithms such as regression, decision trees and K nearest neighbors to evaluate the success of the first phase of the launch.

Data Collection

The data was gathered from the SPACEX REST API and Web scraped from Wikipedia page

SPACEX REST API



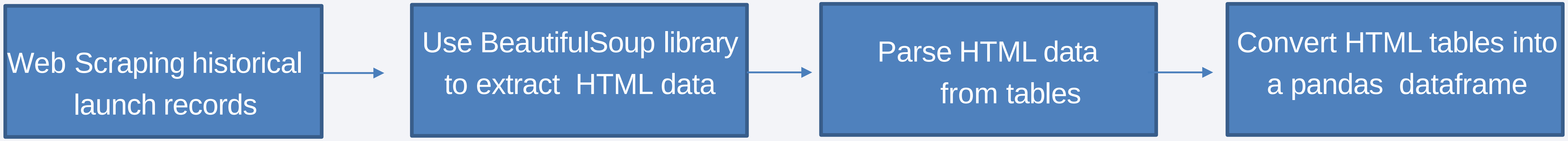
SpaceX REST API

Use Requests library to retrieve data

Get historical launch data as a json file

Convert the json file to a pandas dataframe

Web scraped



Web Scraping historical launch records

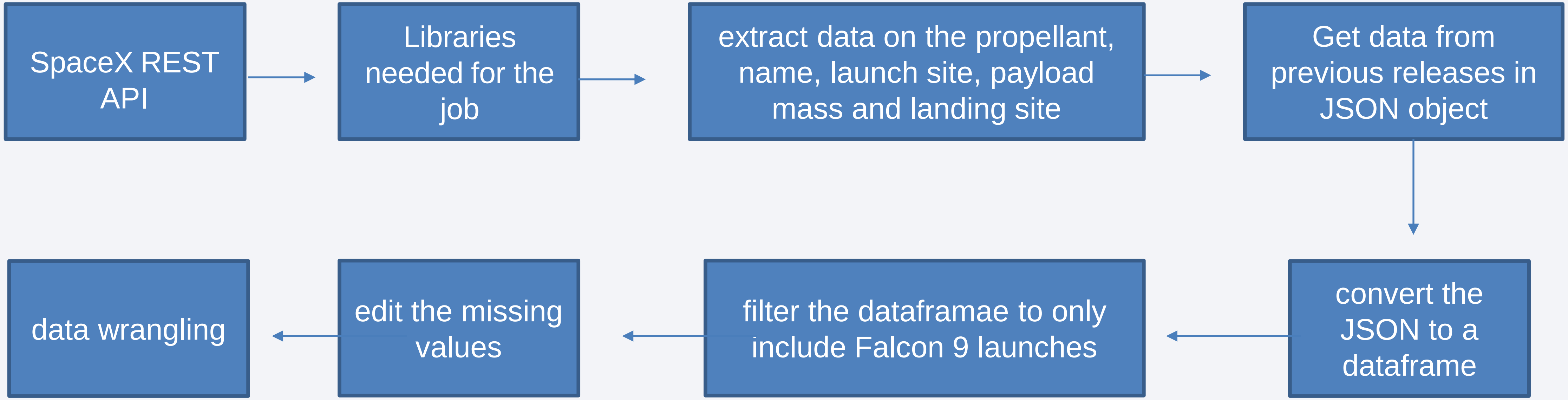
Use BeautifulSoup library to extract HTML data

Parse HTML data from tables

Convert HTML tables into a pandas dataframe

Data Collection – SpaceX API

Collect data via API and make sure it is in the correct format



SpaceX REST API

Libraries needed for the job

extract data on the propellant, name, launch site, payload mass and landing site

Get data from previous releases in JSON object

convert the JSON to a dataframe

filter the dataframae to only include Falcon 9 launches

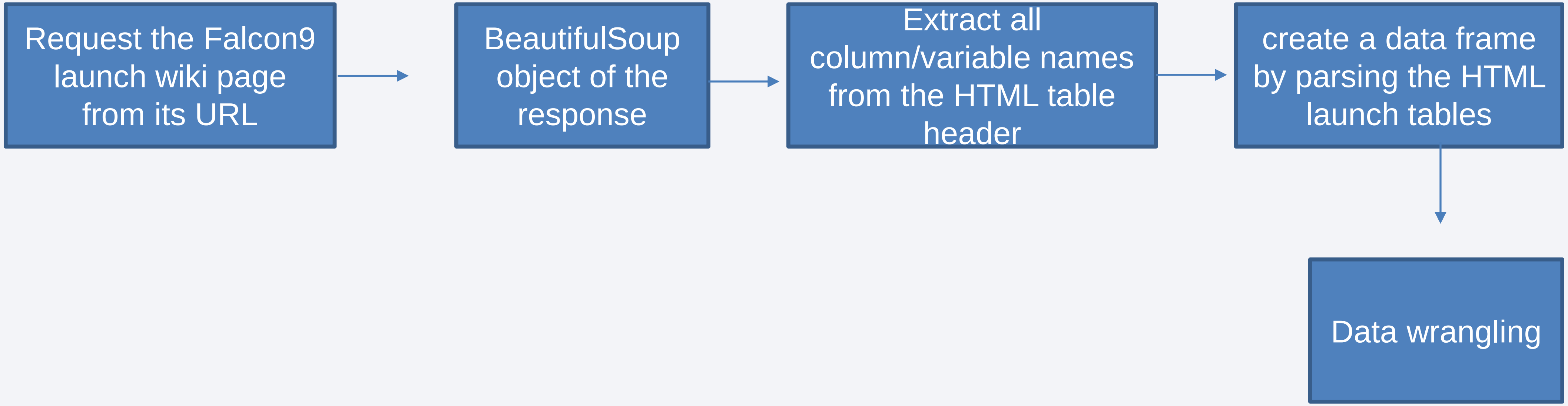
edit the missing values

data wrangling

[Data collection API notebook]

Data Collection - Scraping

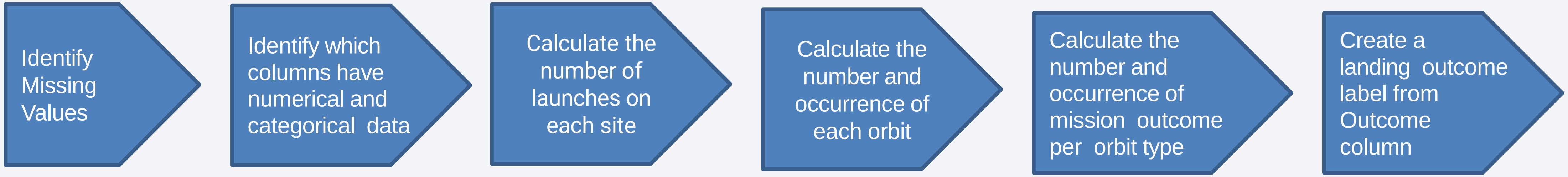
Perform web scrapping to collect Falcon 9 historical launch records from Wikipedia page


Request the Falcon9 launch wiki page from its URL
BeautifulSoup object of the response
Extract all column/variable names from the HTML table header
create a data frame by parsing the HTML launch tables
Data wrangling

[Data collection web scrapping notebook](#)

Data Wrangling

Perform data wrangling to explore, transform, and validate the dataset



Identify Missing Values

Identify which columns have numerical and categorical data

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Data wrangling notebook

EDA with Data Visualization

Summary of charts plotted:

- Scatterplot chart to visualize relationship between Flight Number and Payload
- Scatterplot chart to visualize relationship between Flight Number and Launch Site
- Scatterplot chart to visualize relationship between Payload and Launch Site
- Bar chart to visualize the success rate of each Orbit type
- Scatterplot chart to visualize the relationship between Flight Number and Orbit type
- Scatterplot chart to visualize the relationship between Payload and Orbit type
- Line chart to visualize the launch success yearly trend

Exploratory Data Analysis Data Visualization notebook

EDA with SQL

Summary of executed SQL queries:

- displays the name of each space mission launch site.
- displays five records where the launch site begins with the string "CCA".
- displays the total payload mass carried by NASA Launch Booster (CRS) Display
- List of booster version F9 v1.1 showing average payload mass
- List of dates that successfully landed on the ground pad
- List of names of boosters that successfully landed unmanned spacecraft with payload masses greater than 4000 and less than 6000
- List of total number of successful boosters in successful vs. unsuccessful mission results
- Comparison of names of names of booster versions with maximum payload masses
  Unsuccessful landing results on unmanned spacecraft, their booster versions and launch platforms in 2015
- List of names innumerical ranking of Landing_Outcomes by year 04/06/2010 and 20/03 /2017 descending.

Exploratory Data Analysis SQL notebook

Build an Interactive Map with Folium

To visualize the location of the launch pads and observe features in the area to see if there is a pattern in the nature of the launch environment, the latitude and longitude coordinates of each launch site were used to add circular markers and labels to the interactive map. The markers

have been colored with execution result labels (green for success and red for failure).

Haversin's formula was used to calculate the distance from some launch pads to landmarks such as coast, road, railroad, and city. These distances were drawn with lines on the Map.

[Interactive map with Folium notebook]

Build a Dashboard with Plotly Dash

The dashboard application includes input components such as drop-down lists and a range slider to interact with the pie chart and scatter plot.

- A launch Site dropdown component.There are four different launch sites and a dropdown menu let us select the different launch sites.
- A callback function to render success-soporte-chart based on the selected dropdown option. The inconcreto percepción of this callback function is to get the selected launch site from site-dropdown and render a soporte chart visualizing launch success count.
- A range Slider to select Payload. The slider is to be able to easily select the different payload range and see if we can identity some óptico patterns
- A callback function to render the success-payload-scatter-chart scatter plot. To visually observe how payload may be correlated with mission outcomes for selected site(s).

Interactive dashboard with Plotly

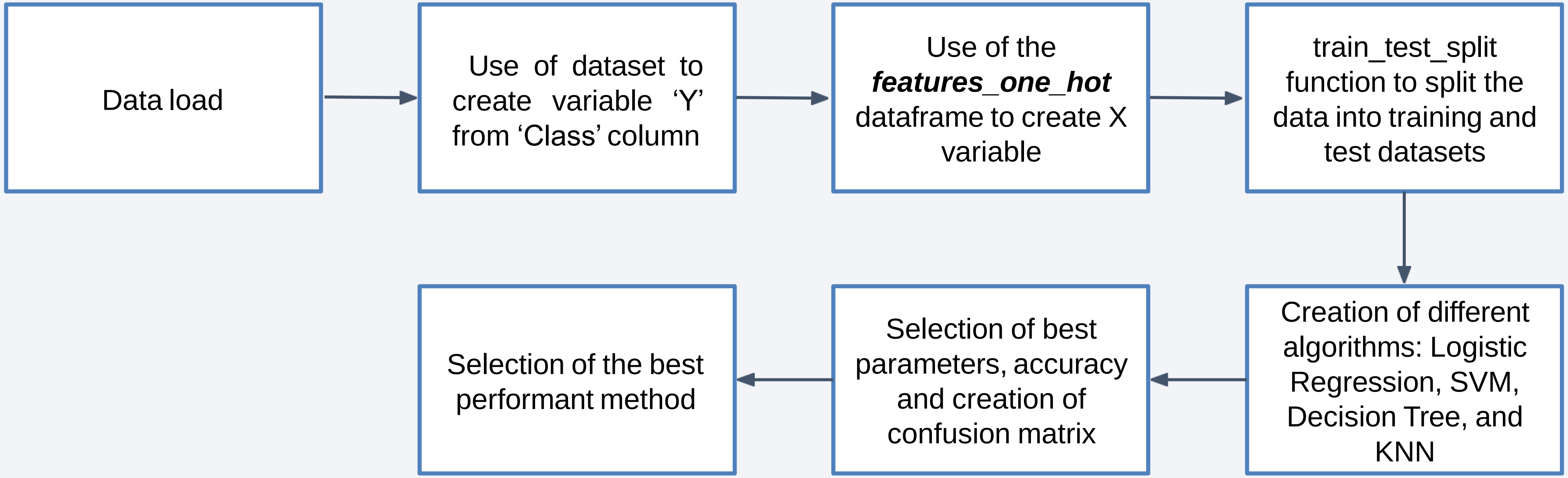
Predictive Analysis (Classification)

The following steps were taken to create the model:

- creating a NumPy matrix from the Class column in the data frame
- standardize the data
- use the train_test_split function to split the x and y data into training and test datasets, the best hyperparameters for the logistic regression, SVM, decision tree and ANN algorithms
- Find the method with the best performance (exactly)

Machine Learning prediction notebook

Predictive Analysis (Classification)



Data load

Use of dataset to create variable ‘Y’ from ‘Class’ column

Use of the ***features_one_hot*** dataframe to create X variable

train_test_split function to split the data into training and test datasets

Creation of different algorithms: Logistic Regression, SVM, Decision Tree, and KNN

Selection of best parameters, accuracy and creation of confusion matrix

Selection of the best performant method

Machine Learning prediction notebook

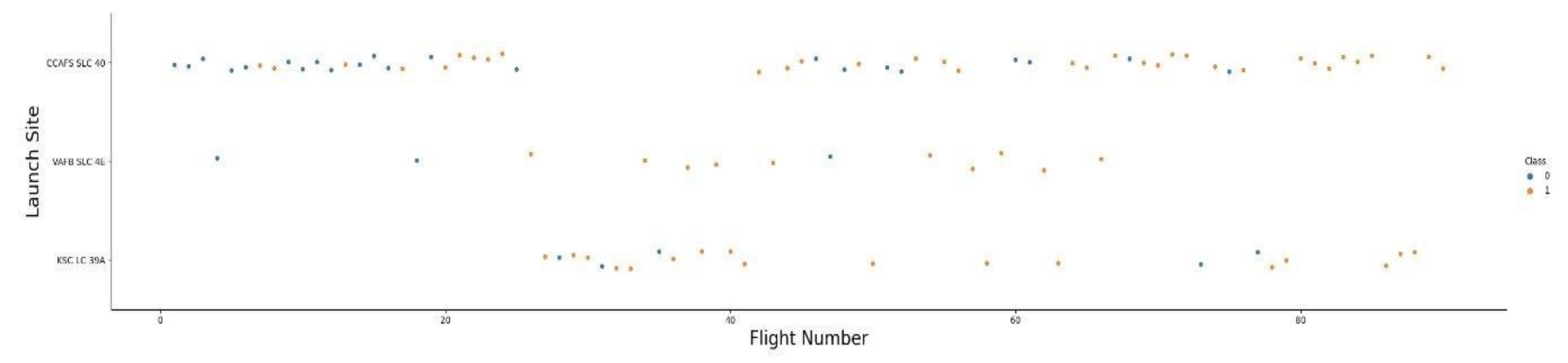
Results

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

Section 2

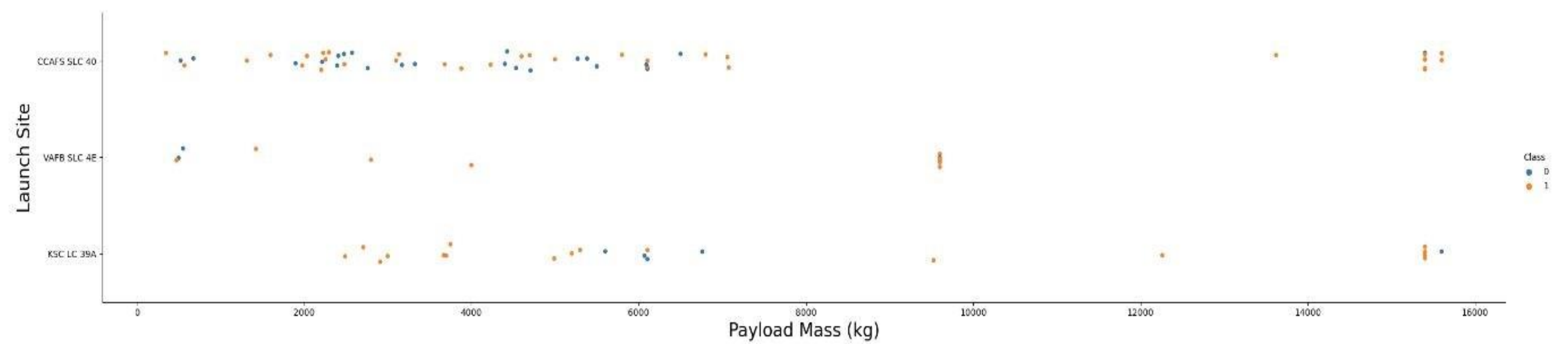
Insights drawn from EDA

Flight Number vs. Launch Site



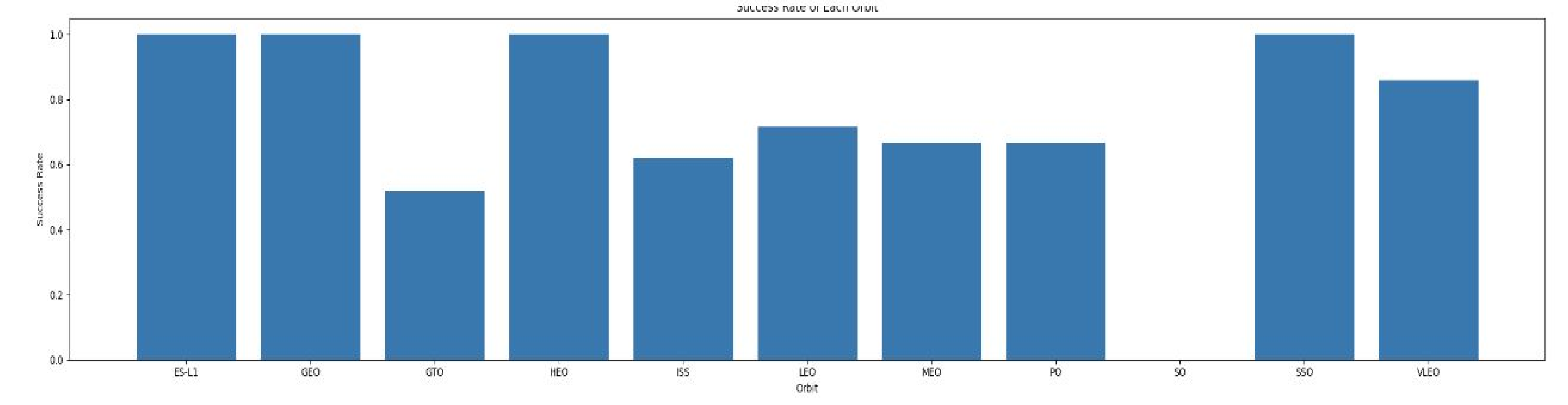
- The success rate seems to increase as the number of flights increases.
- Most of the flights were launched from his CCAFS SLC 40

Payload vs. Launch Site



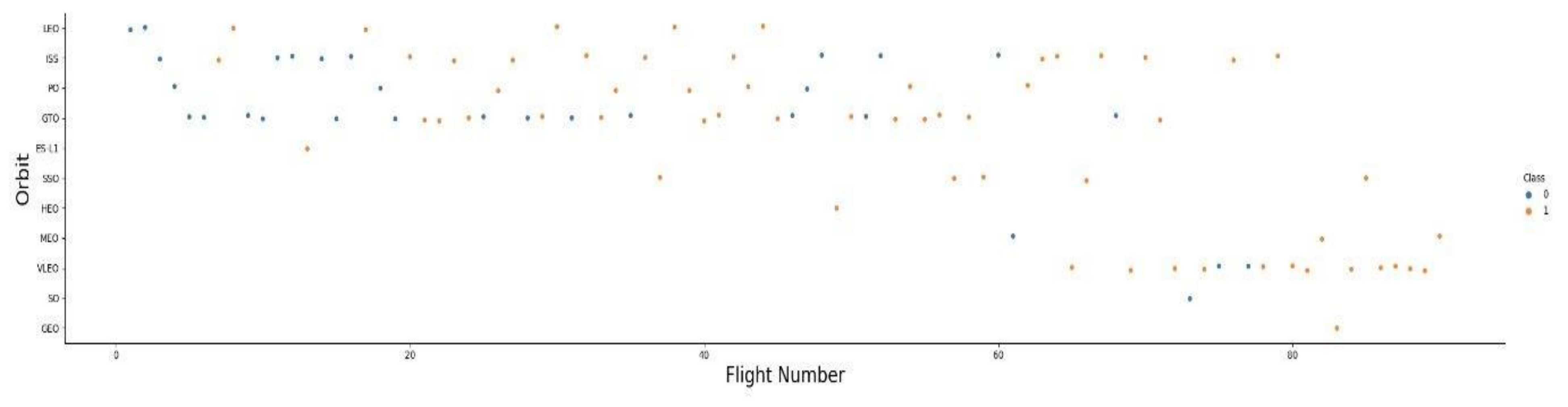
- Few launches with very high payload mass
- There is no clear pattern indicating whether launch site success depends on payload

Success Rate vs. Orbit Type



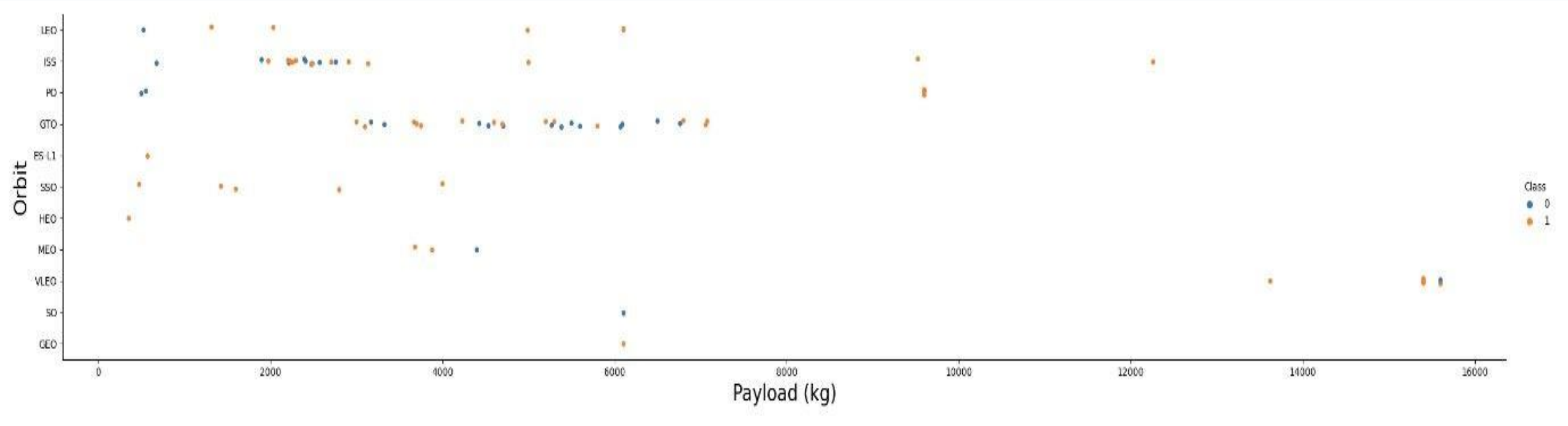
- GEO, HEO, SSO, ES-L1 have the highest success rate
- SO Orbit has no success
- The success rate on GTO orbits is considerably lower than on other orbits.

Flight Number vs. Orbit Type



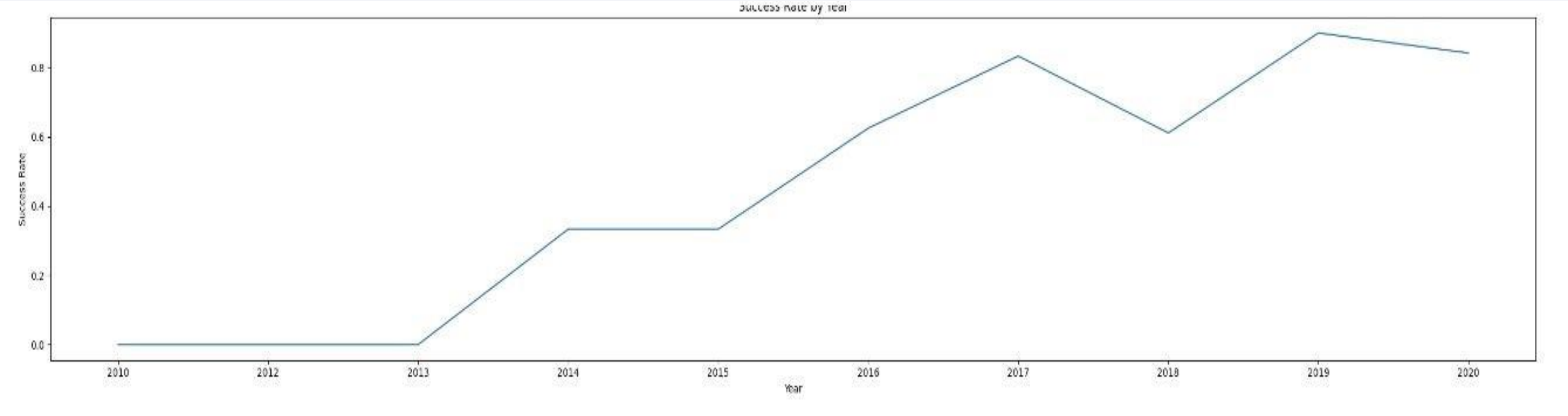
- In the LEO orbit the success appears related to the number of flights
- Flight number seems to have no impact on the success of the GTO orbit

Payload vs. Orbit Type



- Heavy payloads have a negative influence on GTO orbits and positive on Polar LEO (ISS) orbits

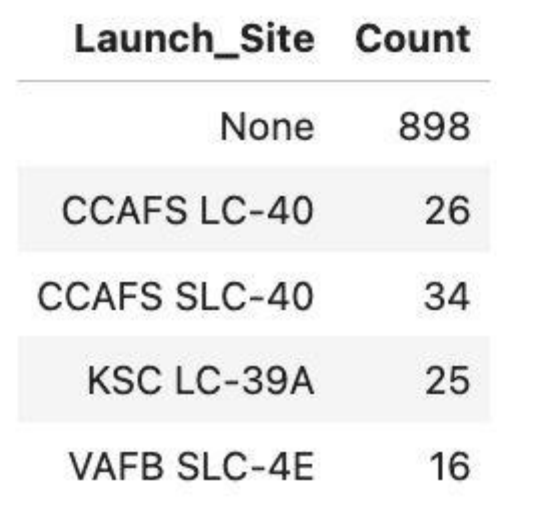
Launch Success Yearly Trend



- Success rate has been increasing since 2013 till 2020

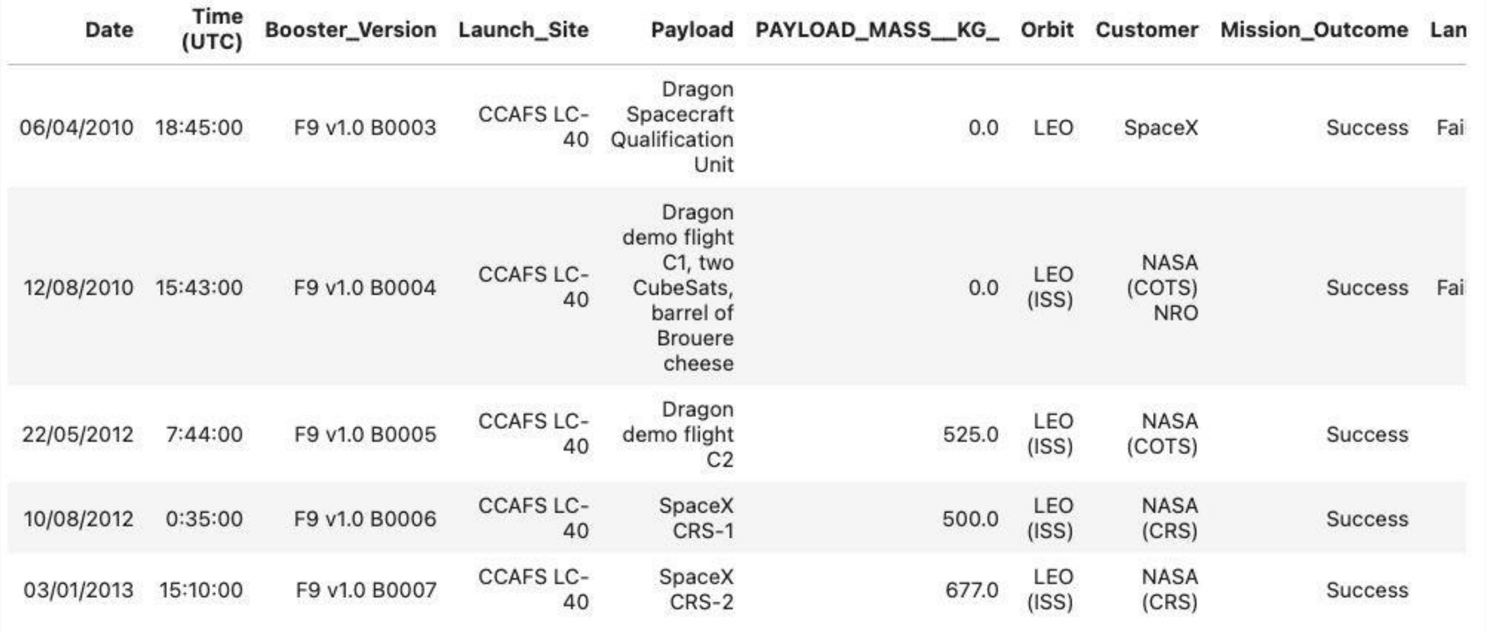
All Launch Site Names

- **select** Launch_Site, **count**(*****) **as** 'Count' **from** SPACEXTBL **group by** Launch_Site
- Select and count all the Launch Site registers, then group by Launch_Site to count how many launches each site has.



Launch_Site	Count
None	898
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

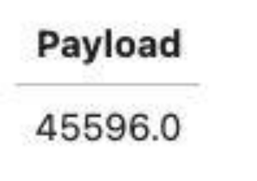
- **select * from** SPACEX **where** Launch_Site **like** 'CCA%' **limit** 5

- The like keyword enables entering string patterns, and the limit keyword enables displaying only the first 5 rows

Total Payload Mass

- **SELECT SUM**(PAYLOAD_MASS.KG_) **as** Payload **FROM** SPACEX
  WHERE Customer **=** 'NASA (CRS)'

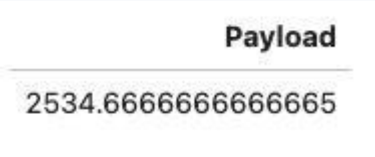
- sum is used to aggregate the total by customer group having the string required



Payload
45596.0

Average Payload Mass by F9 v1.1

- **SELECT AVF**(PAYLOAD_MASSKG_) **AS** Payload **FROM**
  SPACEX
  WHERE Booster_Version **LIKE** 'F9 v1.1%'



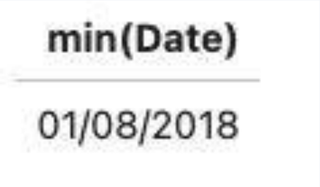
Payload
2534.6666666666665

- avg is the function for aggregating the average value in groups

First Successful Ground Landing Date

- **SELECT MIN**(Date) **FROM** SPACEX
  WHERE Landing_Outcome **=** 'Success (ground pad)'

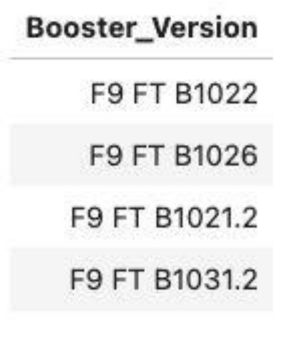
- min function is used to find the minimum value



min(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- **SELECT DISTINCT** Booster_Version **FROM** SPACEX **WHERE** Landing_Outcome **=** 'Success (drone ship)’ **AND** PAYLOAD_MASS_KG_ **BETWEEN** 4000 **AND *** 6000



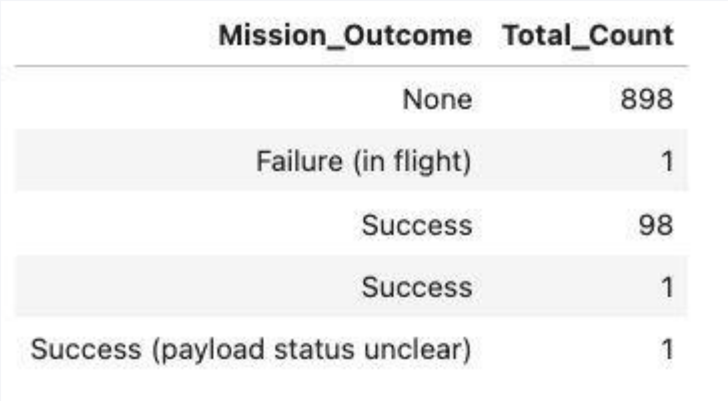
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Two conditions are used in a single query. Four entries are returned as result

Total Number of Successful and Failure Mission Outcomes

-
  ```
  SELECT Mission_Outcome, COUNT(*) AS Total_Count
  FROM SPACEX
  GROUP BY Mission_Outcome;
  ```

- 100 successes have been recorded in the data, and one 1 failure

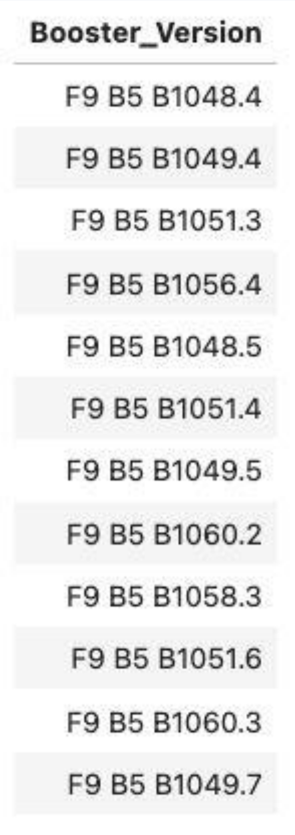


Mission_Outcome	Total_Count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- **SELECT DISTINCT** Booster_Version **FROM** SPACEX **WHERE** PAYLOAD_MASS_KG_ **=** (**SELECT MAX**(PAYLOAD_MASS__KG_) **FROM** SPACEXTBL)

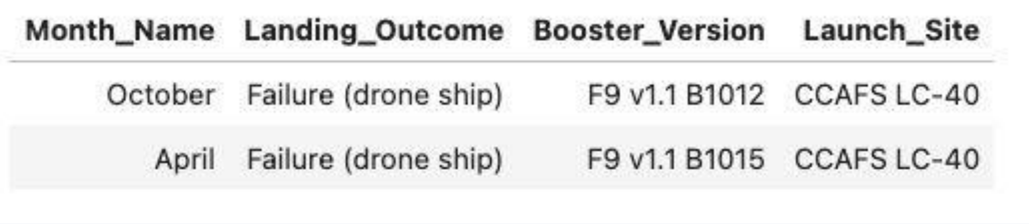
- A subquery was used to obtain the desired result. A total of 12 booster versions have carried maximum payload



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

```
SELECT
        CASE SUBSTR(Date, 4, 2)
                WHEN '01' THEN 'January'
                WHEN '02' THEN 'February'
                WHEN '03' THEN 'March'
                WHEN '04' THEN 'April'
                WHEN '05' THEN 'May'
                WHEN '06' THEN 'June'
                WHEN '07' THEN 'July'
                WHEN '08' THEN 'August'
                WHEN '09' THEN 'September'
                WHEN '10' THEN 'October'
                WHEN '11' THEN 'November'
                WHEN '12' THEN 'December’
        END AS Month_Name, Landing_Outcome,
        Booster_Version,
        Launch_Site FROM SPACEX
WHERE SUBSTR(Date, 7, 4) = '2015’ AND Landing_Outcome =
'Failure (drone ship)';
```



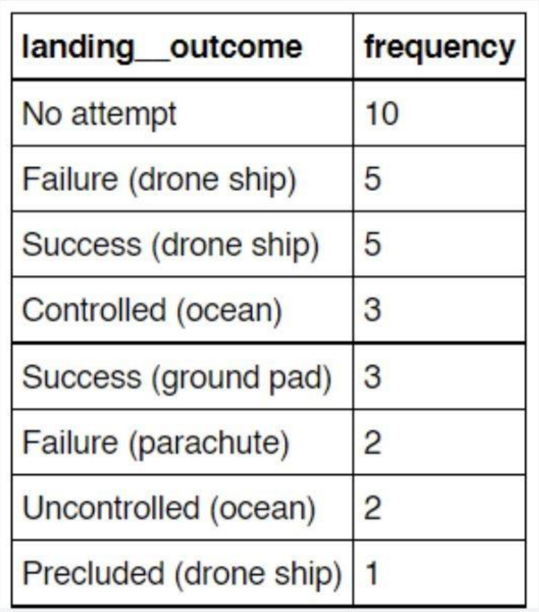
Month_Name	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The CASE function was used to change the number of the month in the Date column for the name of the month

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

- **SELECT** Landing_Outcome, **COUNT**(*****) **AS** Success_Count
  FROM SPACEX
  WHERE Date **BETWEEN** '2010-06-04' **and** '2017-03-20’
  AND Landing_Outcome **LIKE** '%Success%’
  GROUP BY Landing_Outcome
  ORDER BY Success_Count **DESC**;

- No attempt has the highest rank while Precluded (drone ship) has the lowest rank



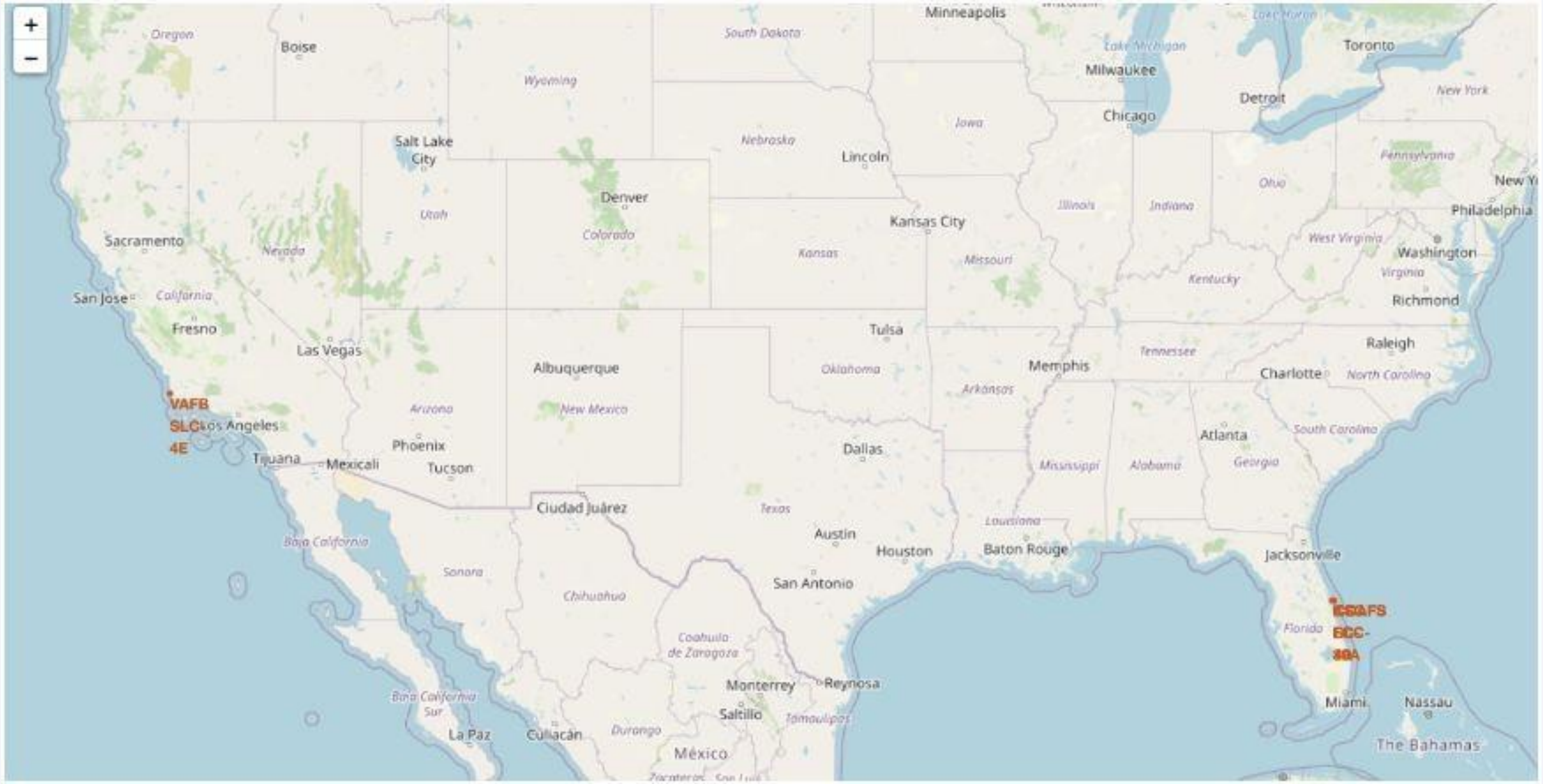
landing__outcome	frequency
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

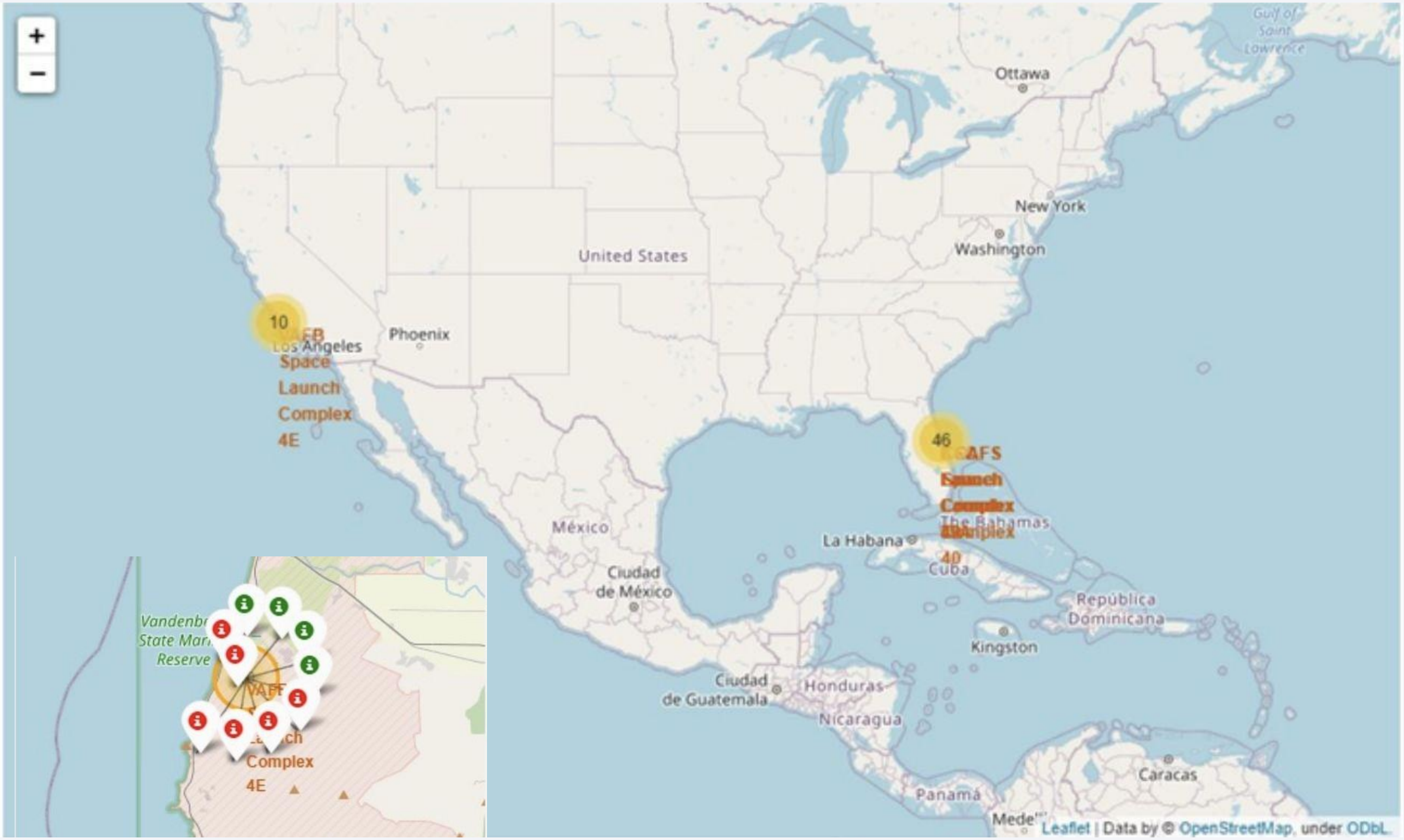
Map showing all launch sites

The launch sites are located in the United States, specifically in the states of Florida and California.



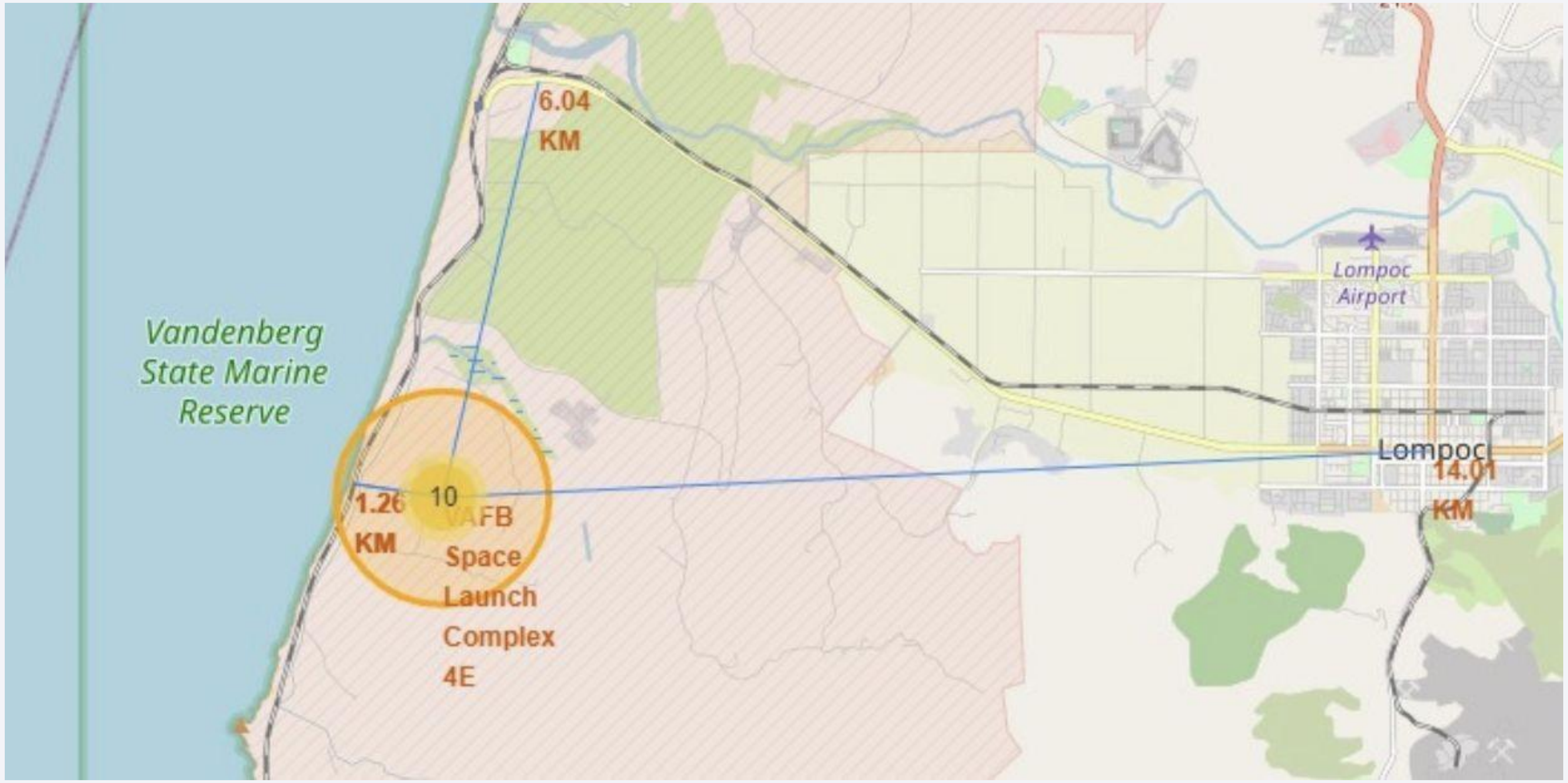
Map showing launch outcomes at each site

The first map shows clusters for every launch site, the second shows a green marker if a launch was successful, and a red marker if a launch was failed



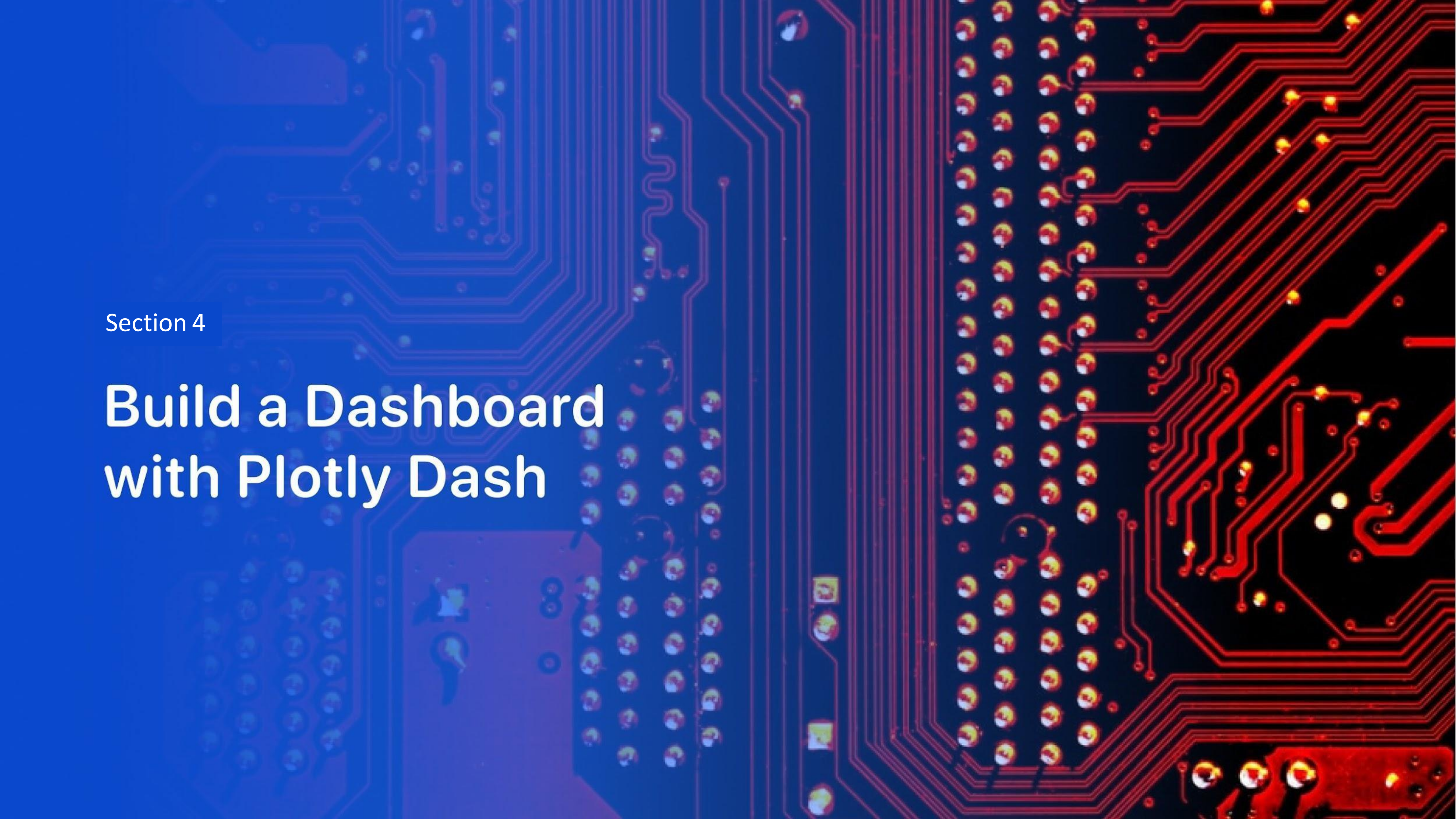
Launch site proximity to Landmarks

Launch sites are located near railroads, roads, highways and coasts. We understand the importance of not only ease of supply and accessibility, but also a safe distance from neighboring cities

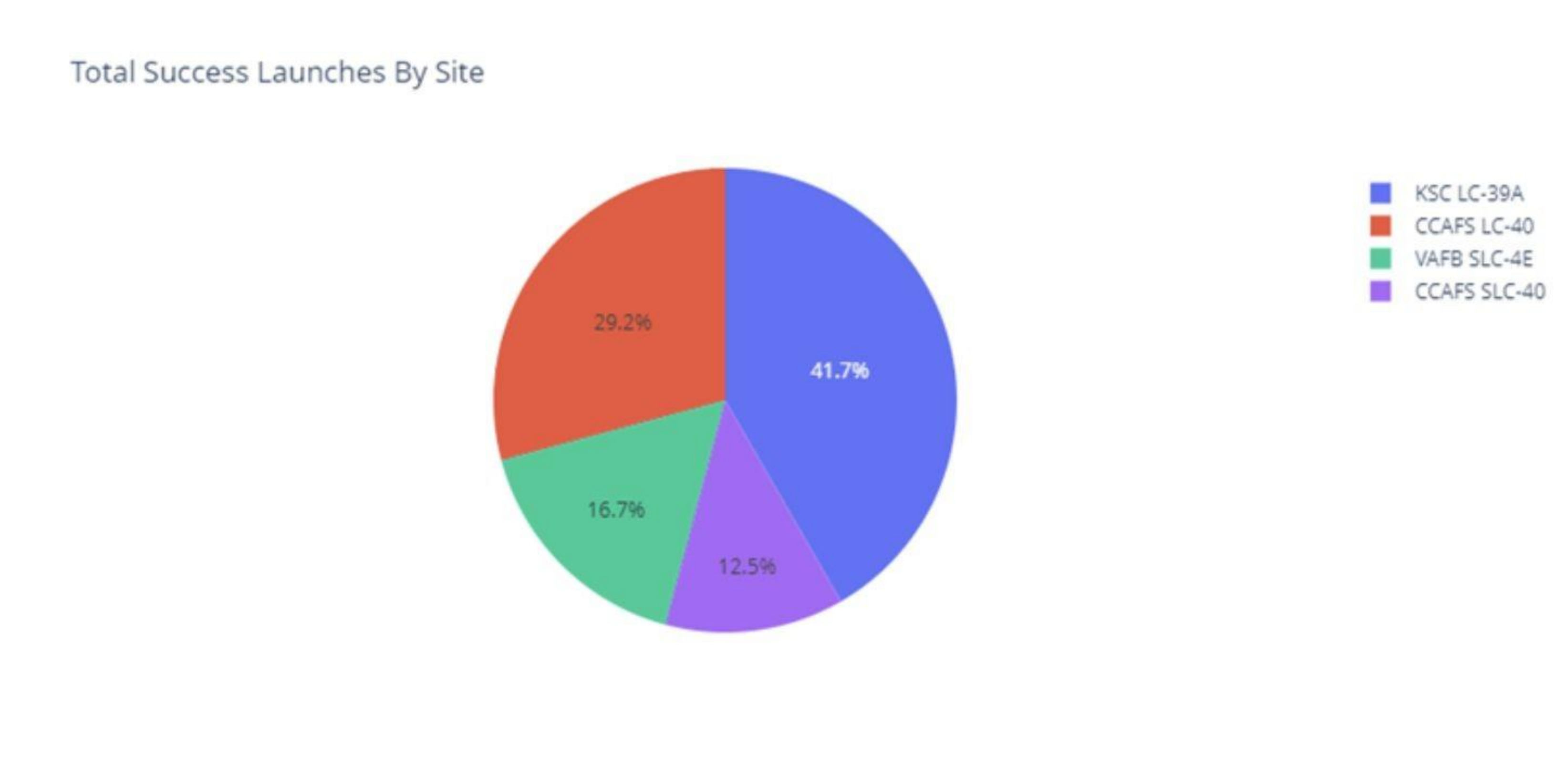


Section 4

Build a Dashboard with Plotly Dash

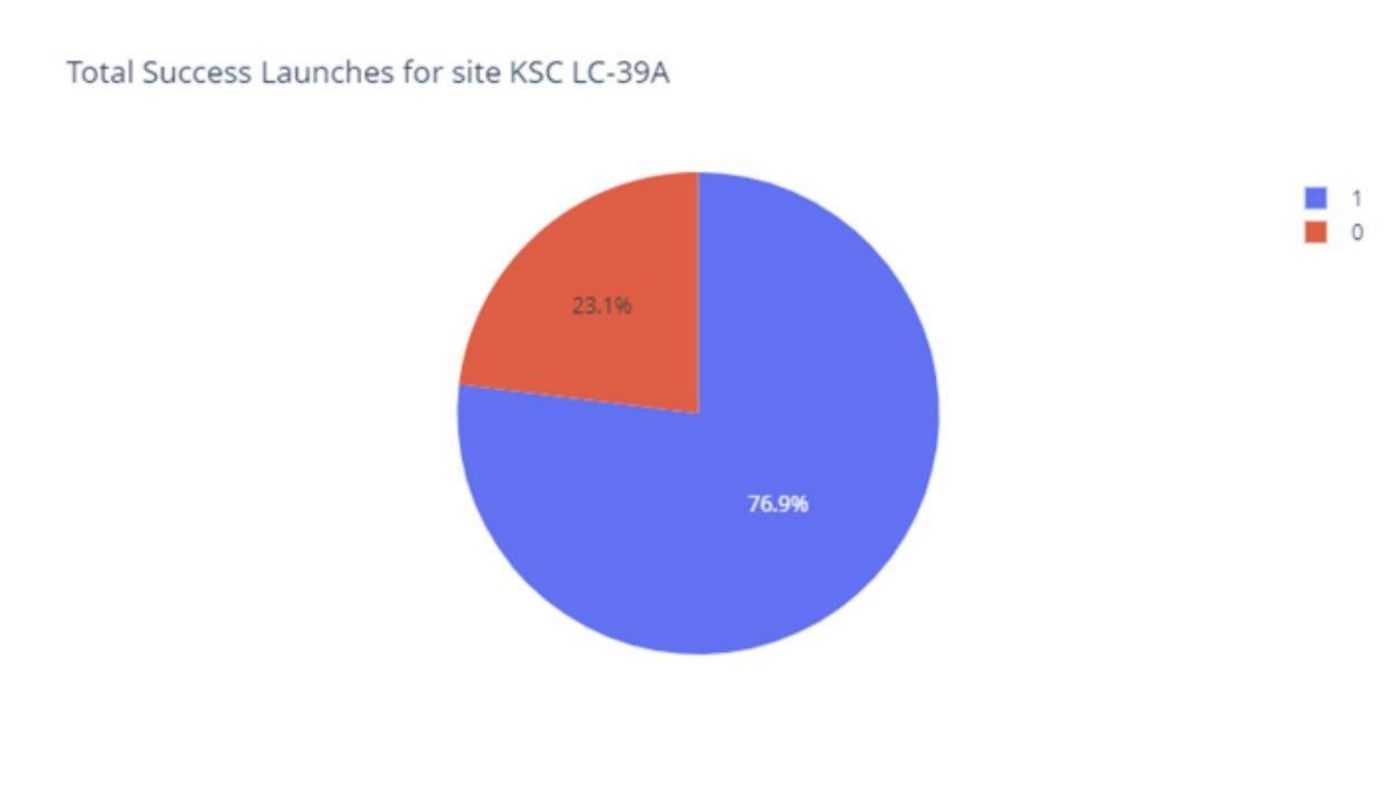


Total Success Launches by Site



KSC LC-39A is the most successfully launched site, followed by CCAFS LC-40.

KSC LC-31A Launches Success

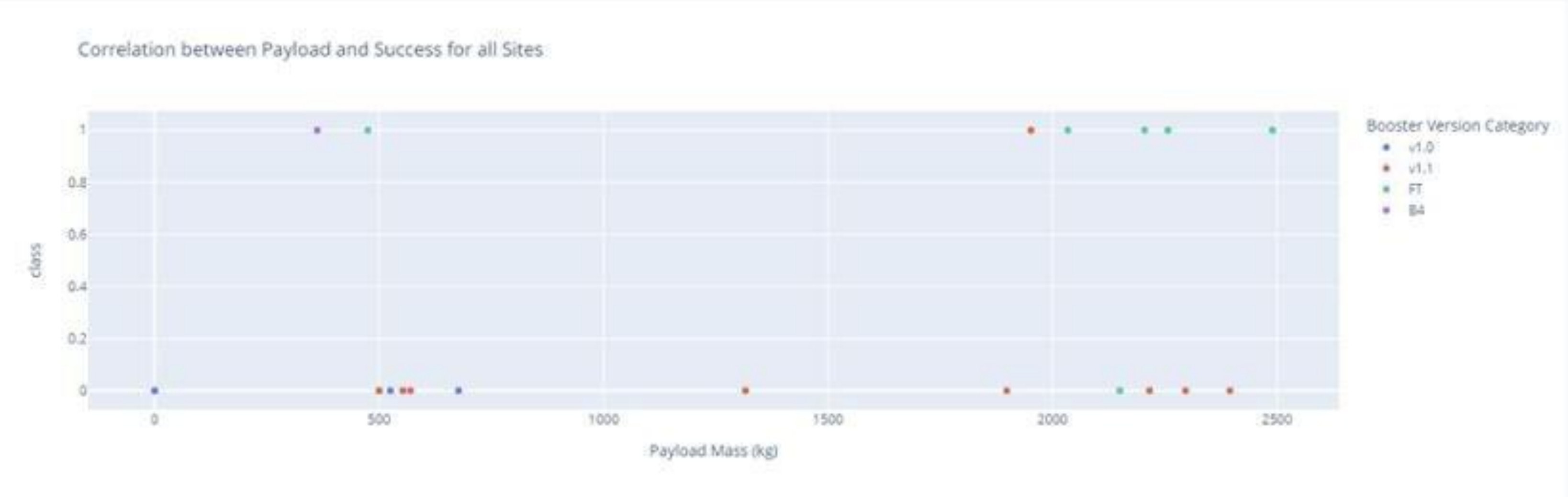


The KSC LC-39A launch site pie chart shows the launch sites with the highest success rates.

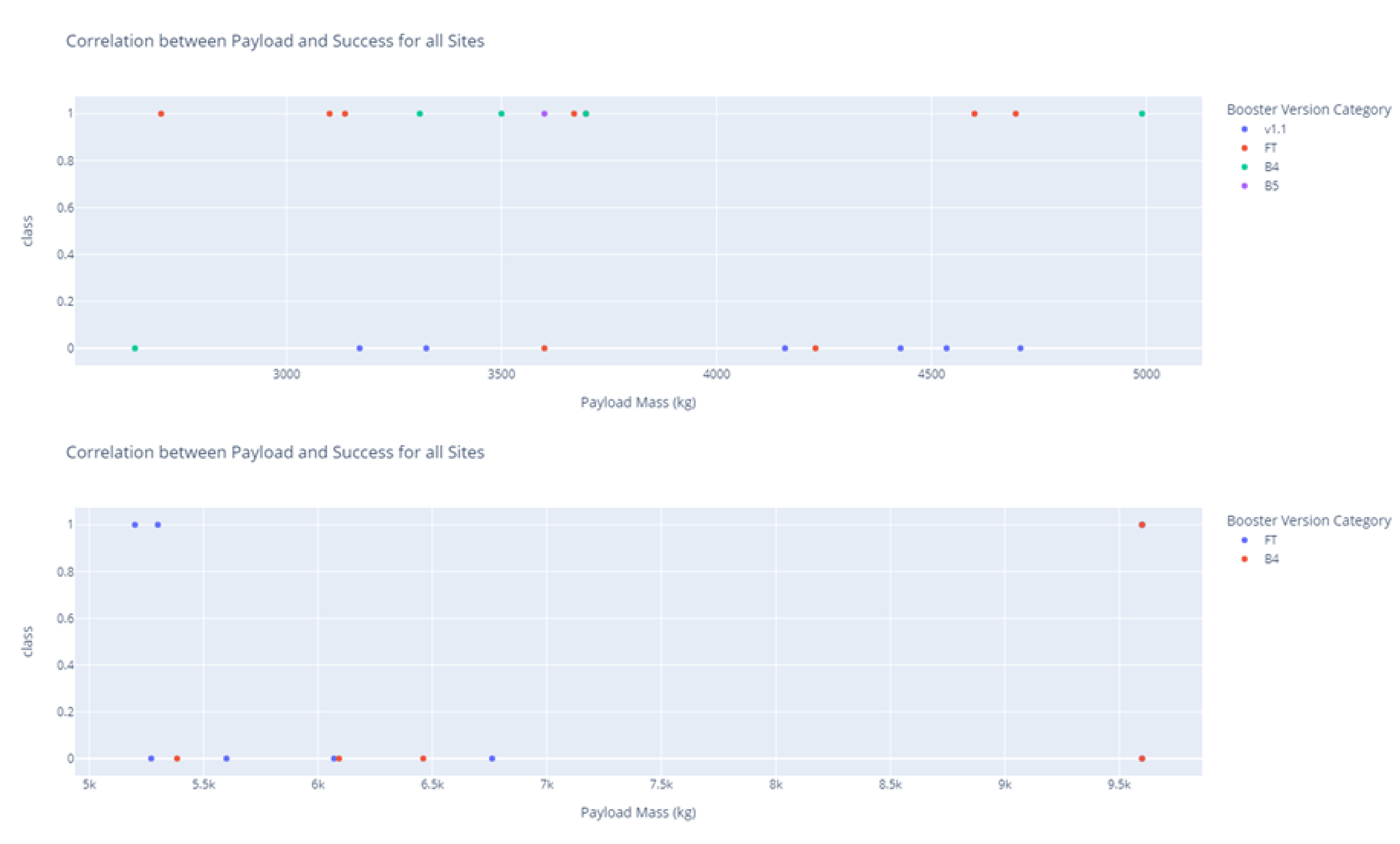
Payload vs Launch Outcome

Scatterplot for all locations with payload ranges of 2500 kg, 5000 kg and 10000 kg.

The 2500-5000kg range focuses on the most successful launches, and the 0-2500kg range focuses on the most unsuccessful launches, but all three are similar.



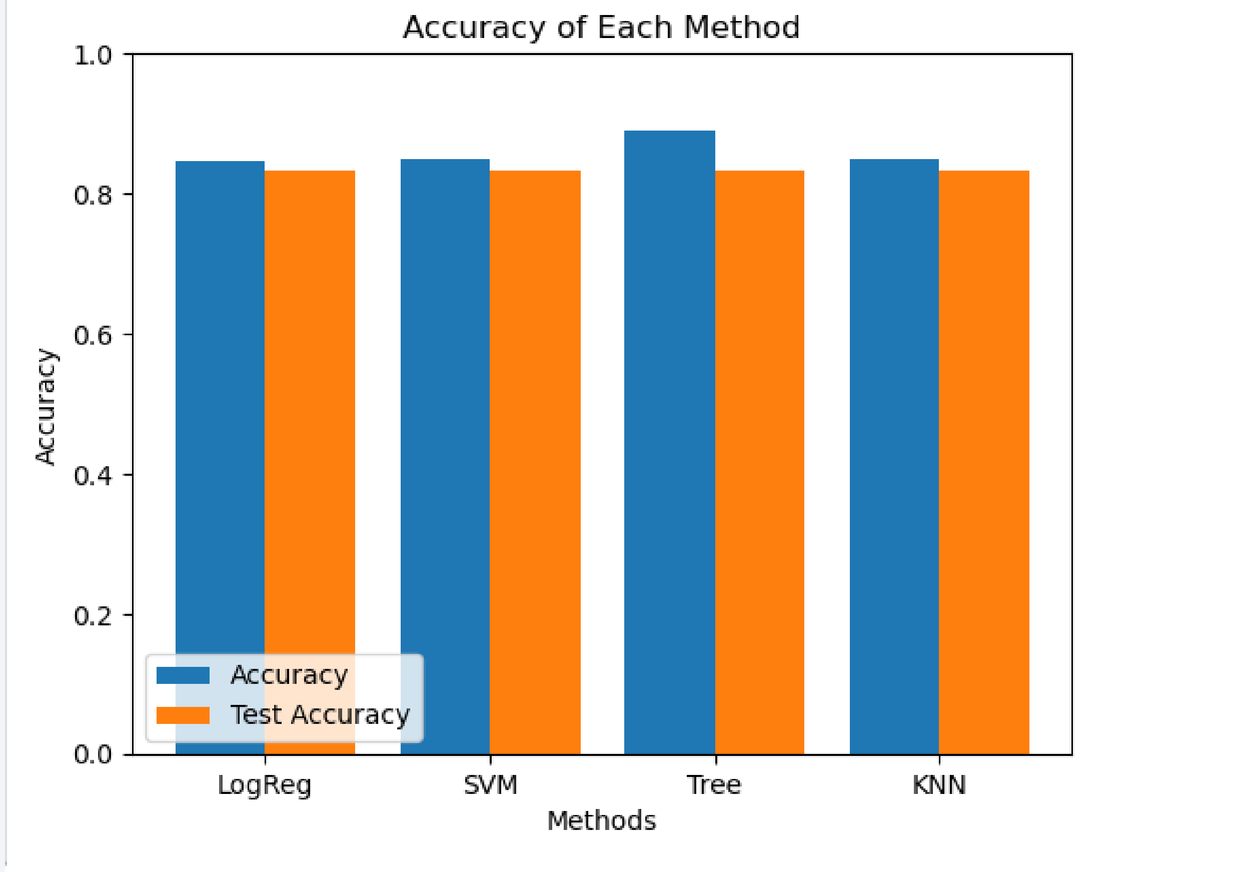
Payload vs Launch Outcome



Section 5

Predictive Analysis (Classification)

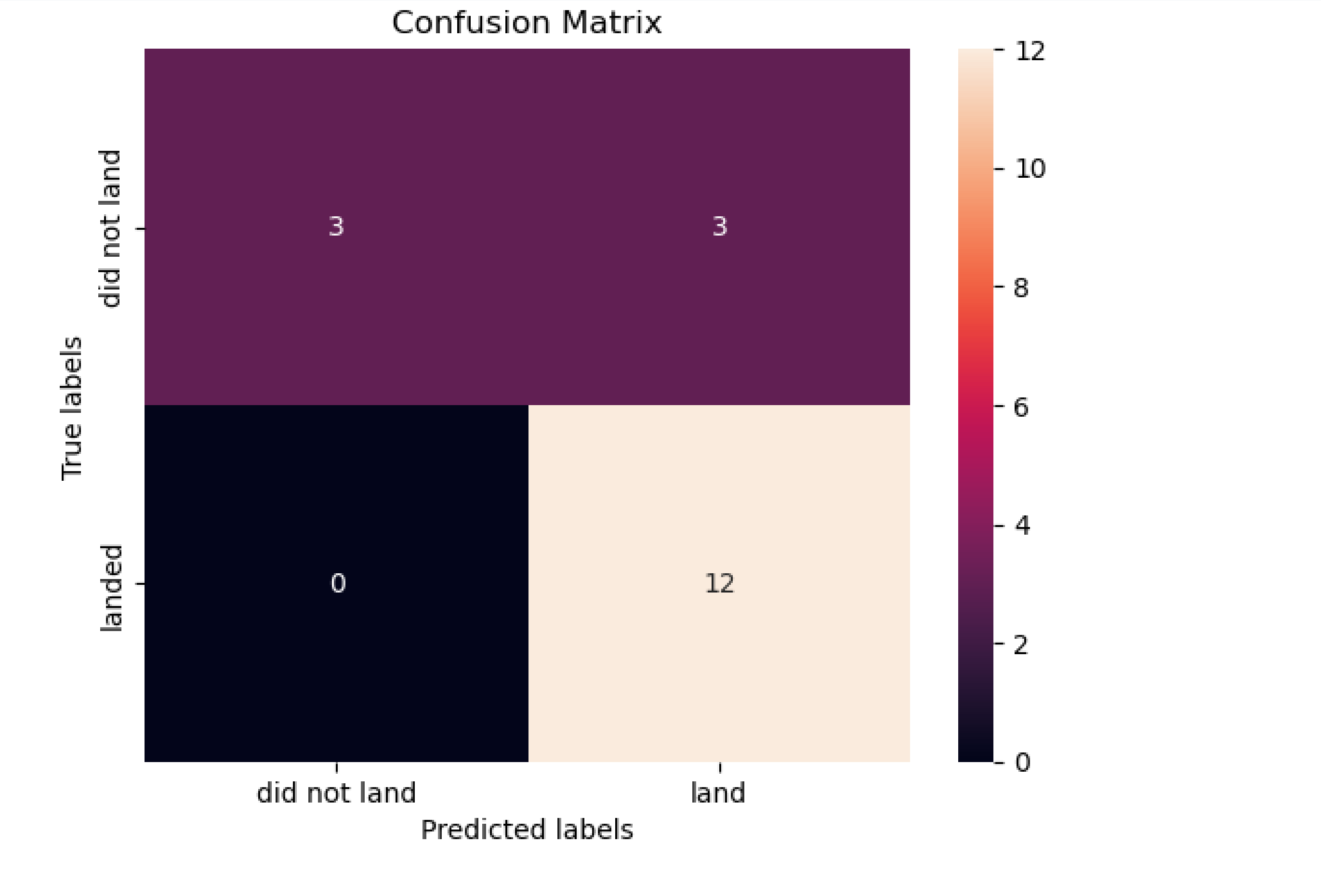
Classification Accuracy



The accuracy is the same for all the models (83,33%)

Confusion Matrix

The confusion matrix is the same for all the models



Conclusions

- All algorithms provide the same level of accuracy, so they all work practically the same.
- A machine learning model can be used to predict whether a competitor's first stage will end up at curacy 83.3.
- Smaller payloads have a higher landing success rate than larger and heavier payloads.
- The launch site with the highest success rate is the KSC LC-39A.
- spaceX's success rate increases over time.

Appendix

Click the following GitHub repository links for notebooks, datasets, and scripts.

↓

Applied Data Science Capstone

Thank you!