IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yuxuan Dong
25.10.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting data via API and Web Scraping
  - Visualizing data for Exploratory Data Analysis
  - Creating interactive map using Folium
  - Creating Dashboard using Plotly and Dash
  - Deploying predictive models for analysis
- Summary of all results
  - Interactive maps
  - Dashboard
  - Predictions

Introduction

- Project background and context
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - If I was given some relevant data of one launch, can I predict the possibility of successful landing?
  - What features are important for the training of the predictor?
  - Finally, what conditions are essential for a successful landing for Falcon 9?

Section 1

Methodology

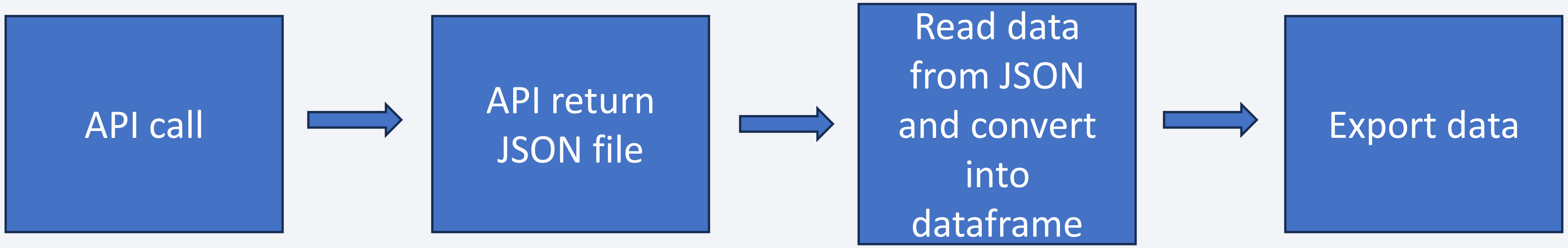
Methodology

Executive Summary

- Data collection methodology:
    - From SpaceX API
    - Scraping from Wikipedia
- Perform data wrangling
    - Clean the data by dropping unnecessary columns
    - Perform one-hot encoding for classifier training
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Binary Classification Model

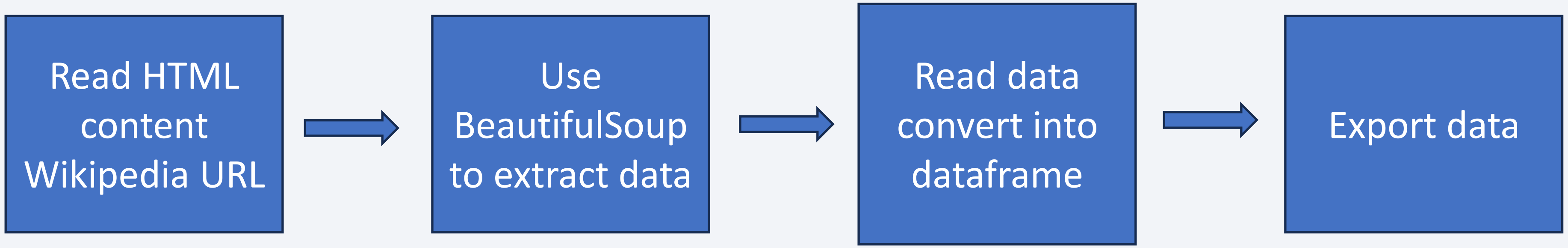
Data Collection

- From SpaceX Rest API



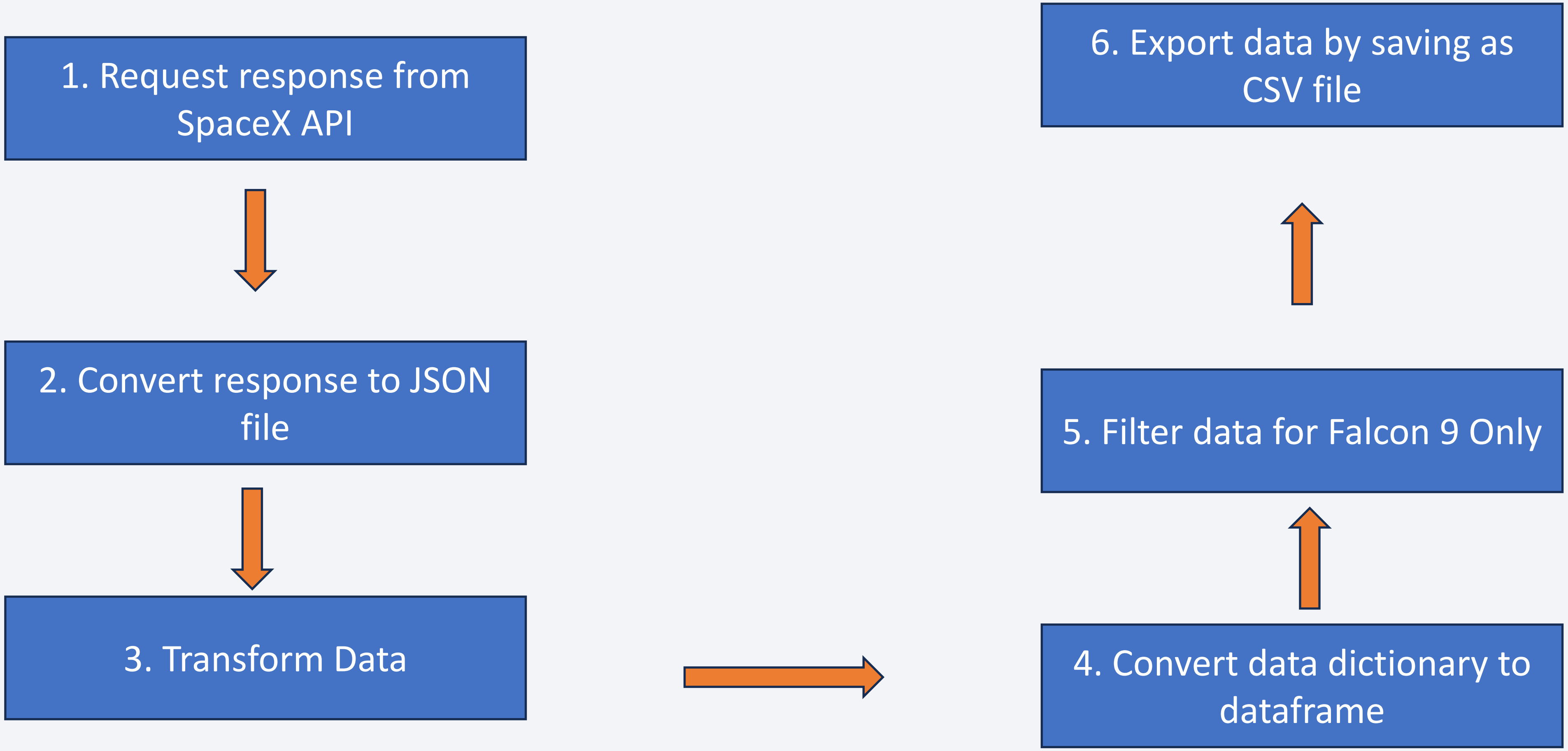
API call → API return JSON file → Read data from JSON and convert into dataframe → Export data

- Web scrap from Wikipedia



Read HTML content Wikipedia URL → Use BeautifulSoup to extract data → Read data convert into dataframe → Export data

Data Collection – SpaceX API



1. Request response from SpaceX API

2. Convert response to JSON file

3. Transform Data

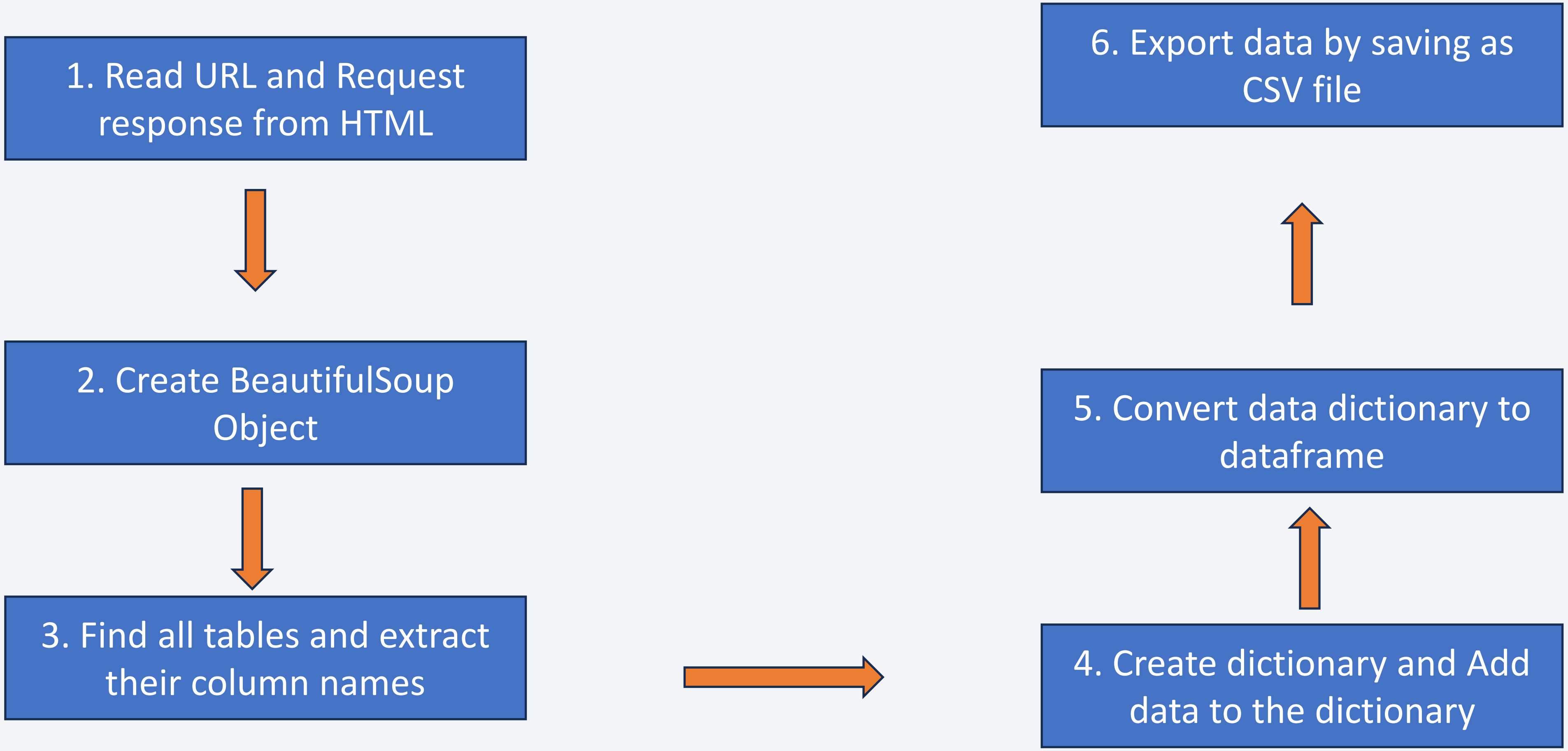
4. Convert data dictionary to dataframe

5. Filter data for Falcon 9 Only

6. Export data by saving as CSV file

Click here for the completed notebook

Data Collection - Scraping



1. Read URL and Request response from HTML

2. Create BeautifulSoup Object

3. Find all tables and extract their column names

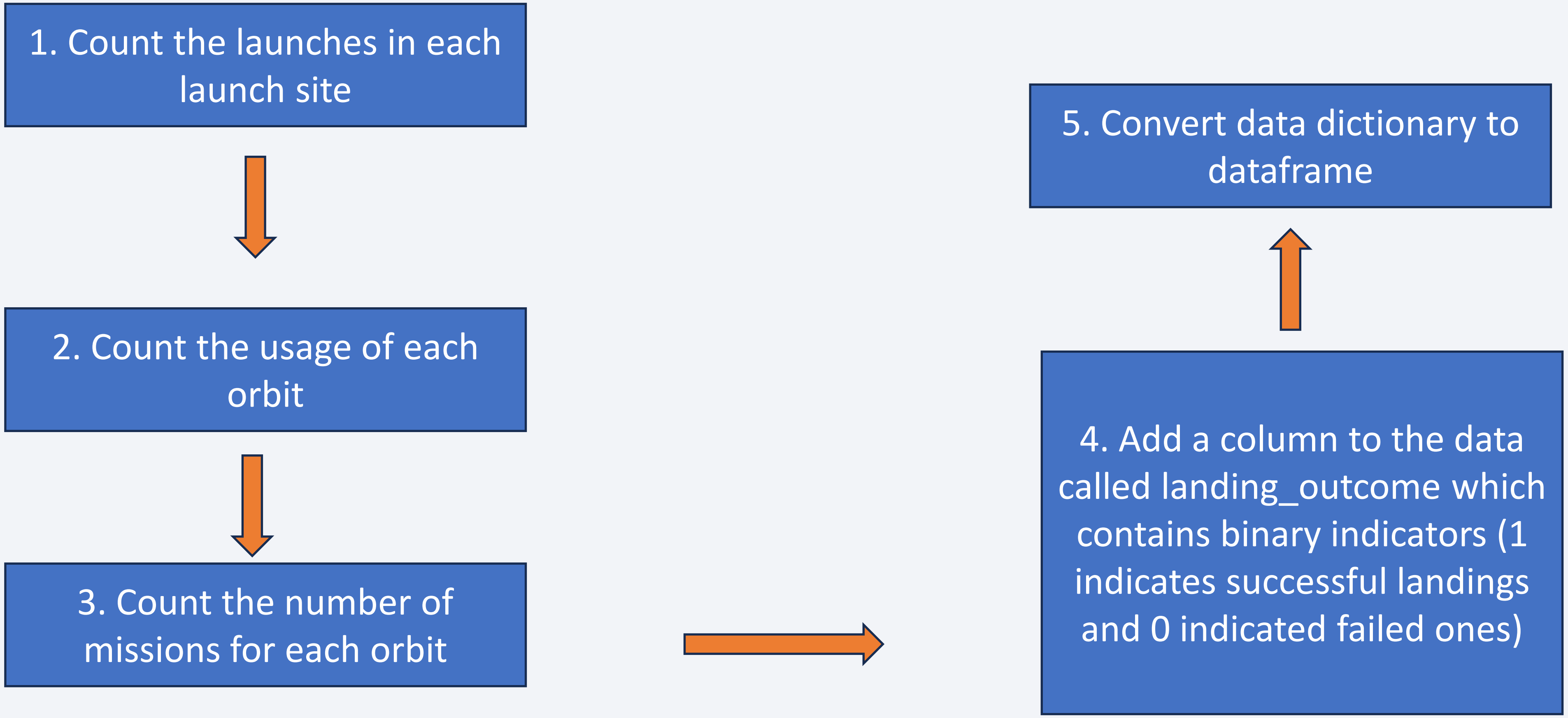
4. Create dictionary and Add data to the dictionary

5. Convert data dictionary to dataframe

6. Export data by saving as CSV file

Click here for the completed notebook

Data Wrangling



1. Count the launches in each launch site

2. Count the usage of each orbit

3. Count the number of missions for each orbit

4. Add a column to the data called landing_outcome which contains binary indicators (1 indicates successful landings and 0 indicated failed ones)

5. Convert data dictionary to dataframe

Click here for the completed notebook

EDA with Data Visualization

Chart	Plot	Why
Scatter Plot	• Flight No. vs Payload Mass • Flight No. vs Launch Site • Payload Mass vs Launch Site • Orbit vs Flight No. • Payload Mass vs Orbit • Orbit vs Payload Mass	Scatter plot illustrates the distribution of the data and shows the correlation between variables.
Bar Chart	Success rate vs Orbit	Bar chart shows the data distribution of categorial variables.
Line Plot	Success rate vs Year	Line plot shows the trend of data and in this case, it gives intuitive view of how the success rate been improved for years.

Click here for the completed notebook

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display 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 greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
- 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.

Click here for the completed notebook

Build an Interactive Map with Folium

- Red Circles label the locations of objects we interested, such as the NASA Space Centre and different launch sites.
- Green markers indicate the successful landing and red ones indicate failed landings.
- Some other markers are to show the position relation between the launch sites and some important locations.
- These circles and markers help user to have an overview of the launch sites and the landing outcomes at each launch sites for further analysis about the impact of locations of launch sites of the success rate of landing.

Click here for the completed notebook

Build a Dashboard with Plotly Dash

- The dashboard includes dropdowns, pie charts and scatter plots.
- The user can choose different launch sites from the dropdown to obtain the information and plots at those launch sites respectively.
- The pie charts show the success rate of landing at different launch sites.
- The scatter plots shows the relation between the payload weight and the successful landing rate.

Predictive Analysis (Classification)

- Data preparation
  - Load Dataset
  - Clean and preprocess the data
  - Split data into training and validation sets
- Model development
  - Train models by choosing different machine learning algorithms (GridSearchCV)
- Model Evaluation
  - Compute the accuracy scores of each model
  - Compute the confusion matrices for each model
- Model Comparison
  - Plot graphs to compare the accuracy score of each model to find the best model for production

Click here for the completed notebook

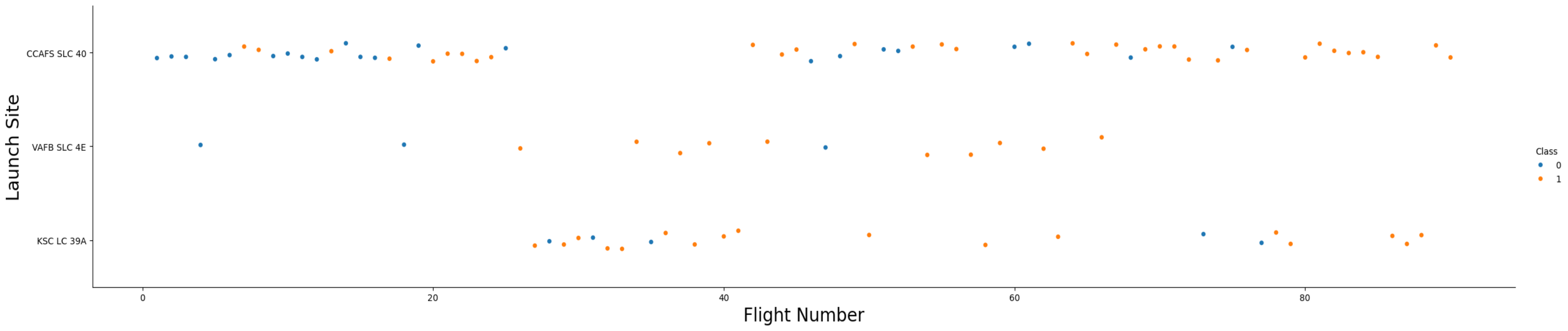
Results

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

Section 2

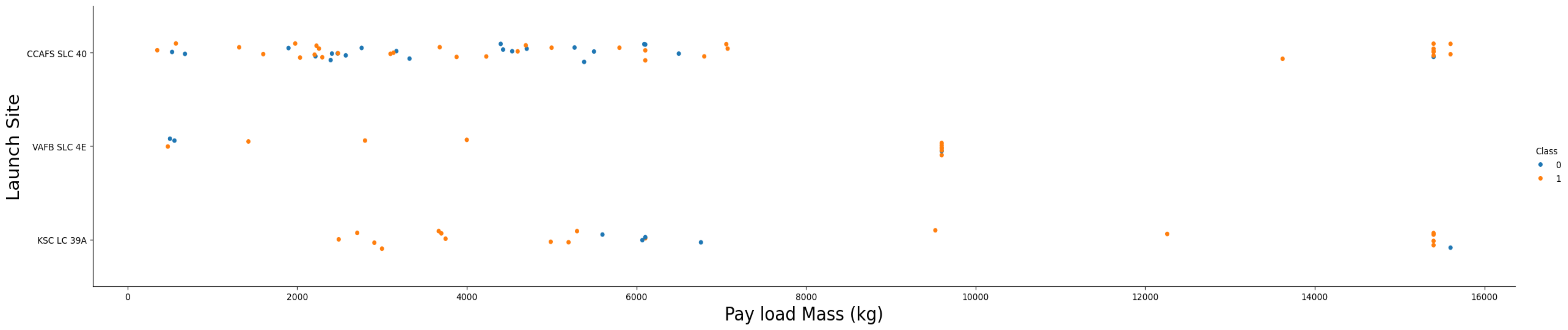
Insights drawn from EDA

Flight Number vs. Launch Site



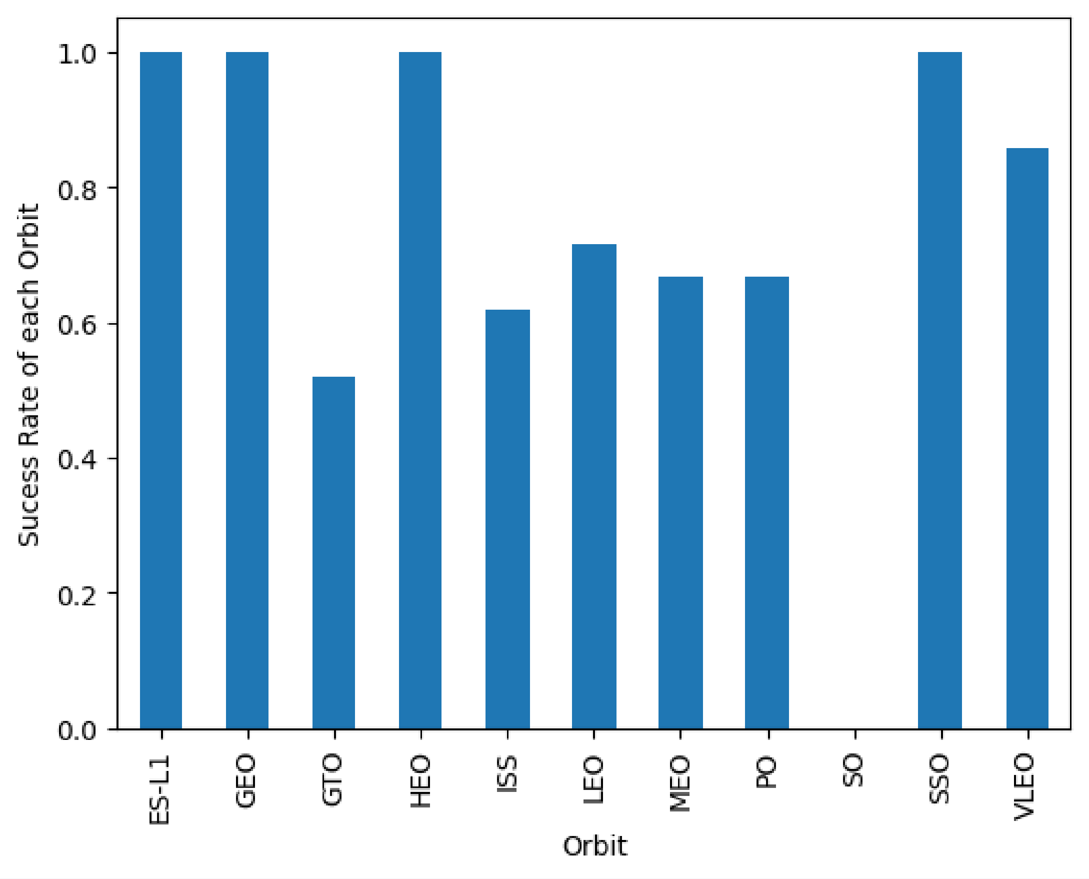
- For each site, the landing success rate increases as the number of launches increases. And also, the launch site, CCAFS SLC 40, is the first launch site been used.

Payload vs. Launch Site



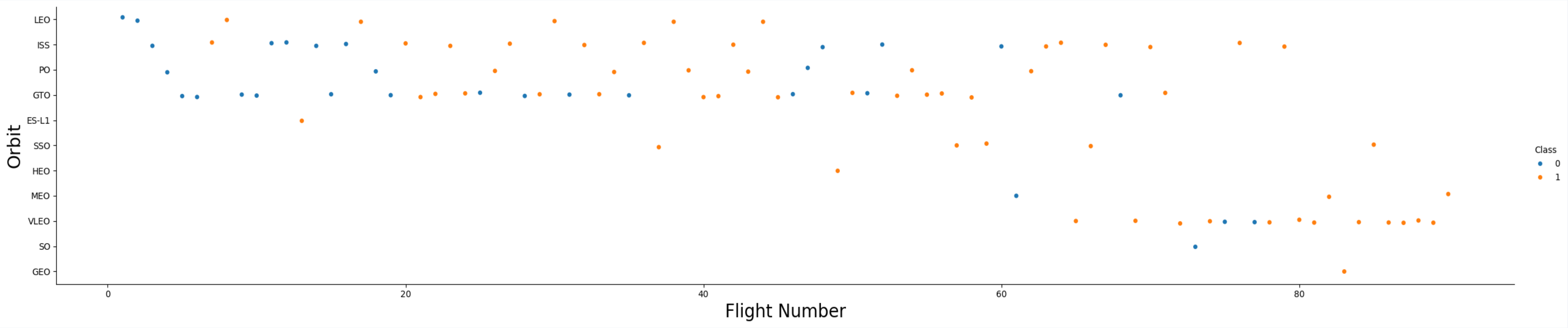
- The launches with heavier payload are more likely to land successfully, however, once the weight of the payload is beyond some threshold, it may cause a failed landing.

Success Rate vs. Orbit Type



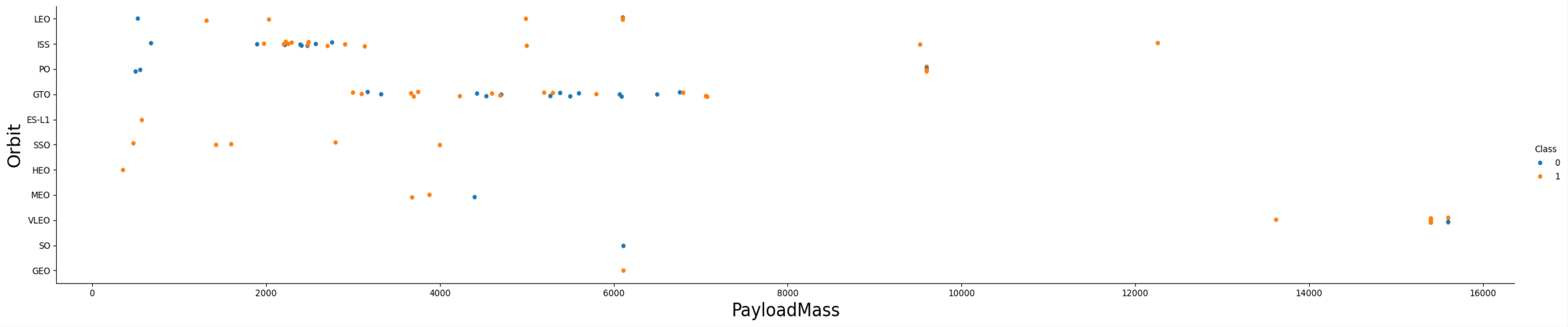
- The launches with orbit, ES-L1, GEO, HEO and SSO have the same rate of successful landing of 1, while the orbit GTO has the lowest success rate of only about 0.55.

Flight Number vs. Orbit Type



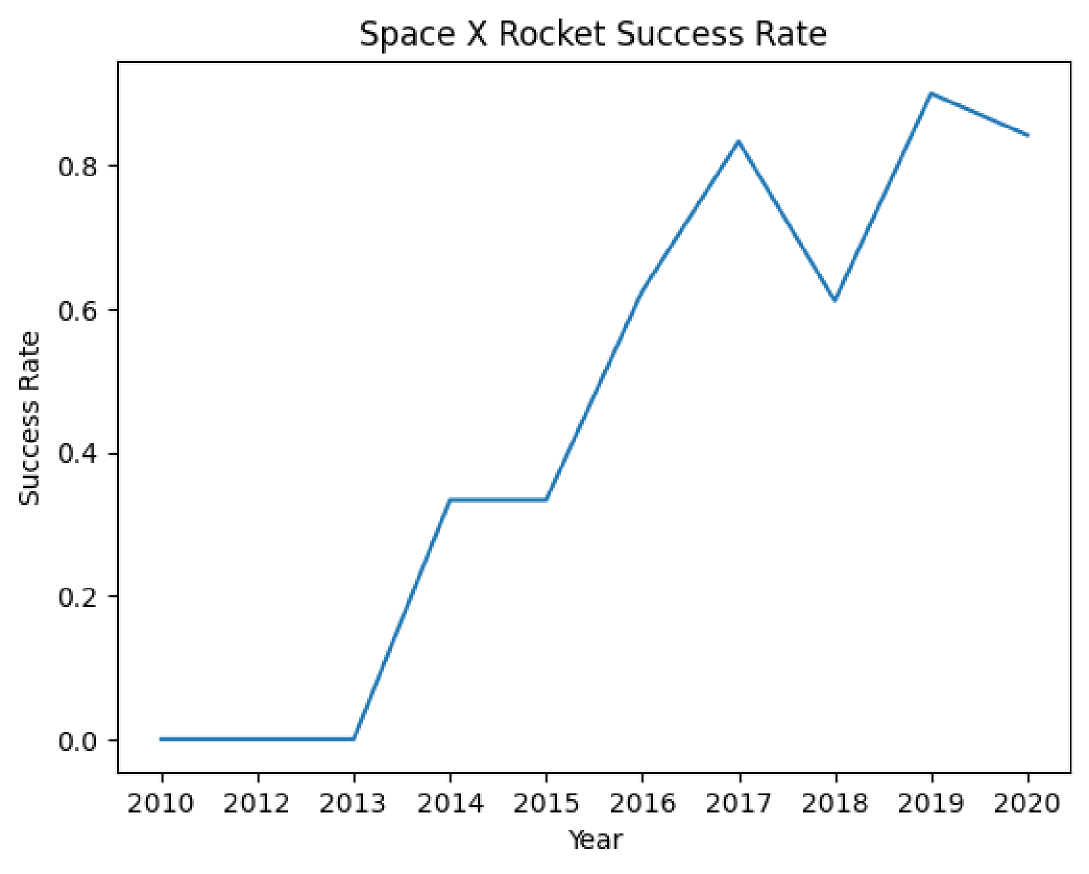
- For all the orbits, the successful landing rate increased.
- Some of the orbits are used from middle point of the whole journey, so their high success rate shows the importance of learning from the previous failures.

Payload vs. Orbit Type



- The launches with heavier payload are less likely to land successfully, also, once the weight of the payload is beyond some threshold, it may cause a higher failure landing rate.

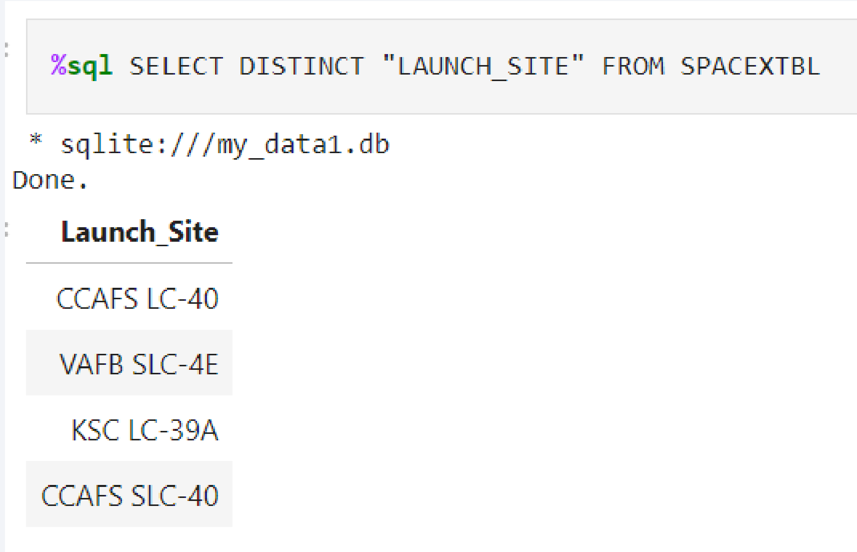
Launch Success Yearly Trend



- The success rate of landing increased with fluctuation from 2010 to 2020. The success rate increased by almost 0.9 within 10 years.

All Launch Site Names

- DISTINCT gives unique items in a column.



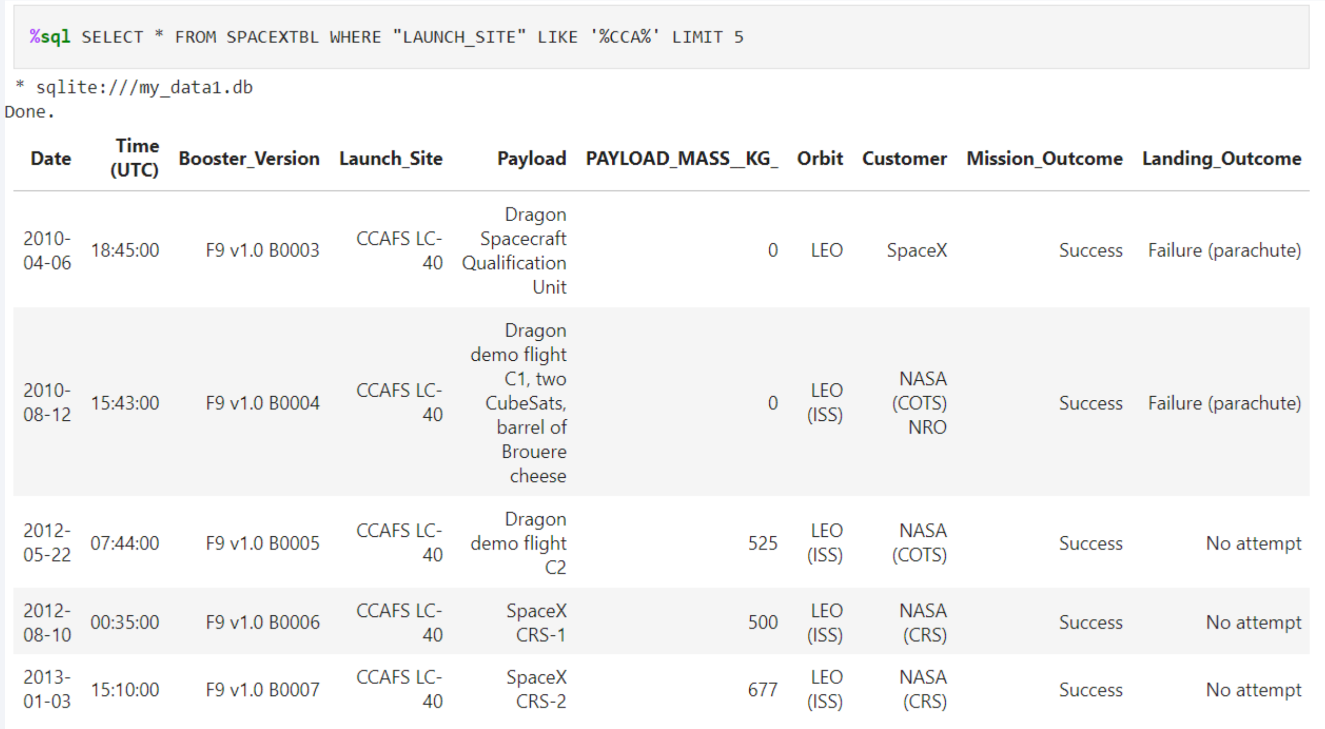
```
%sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Use LIKE to find the rows has CCA in its Launch Site column. Use LIMIT 5 to only print out the first 5 rows meet the condition.



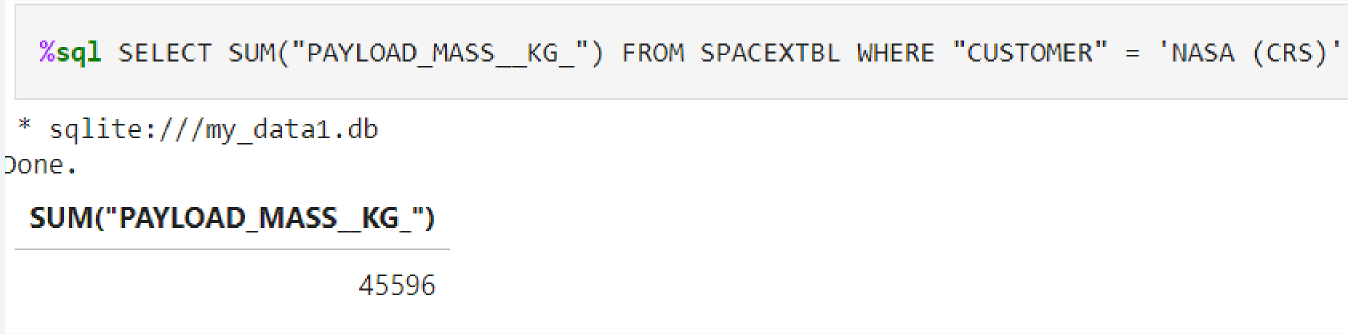
```
%sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Payload Mass

- Use WHERE to apply a condition of the CUSTOMER is NASA(CRS) and use SUM to calculate the total values of payload mass of rows meet the conditions.



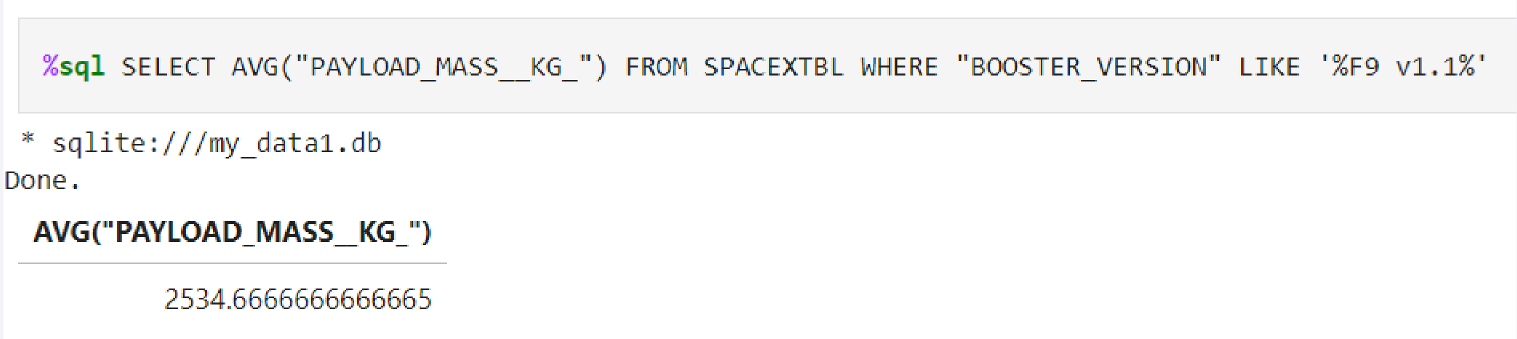
```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- Find rows with F9 v1.1 in their Booster version column and use AVG calculate the average payload mass of the rows meet the previous condition.



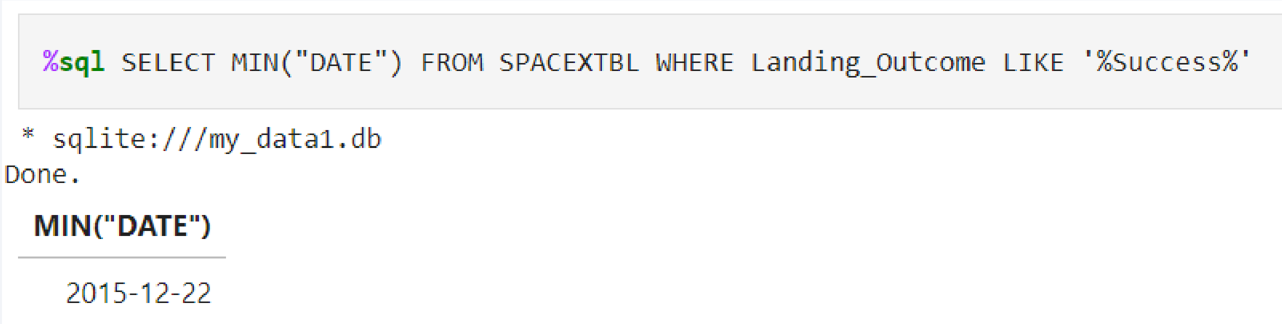
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- Find the rows with Success on their Landing_Outcome column and from those rows find the minimum Date by using MIN.



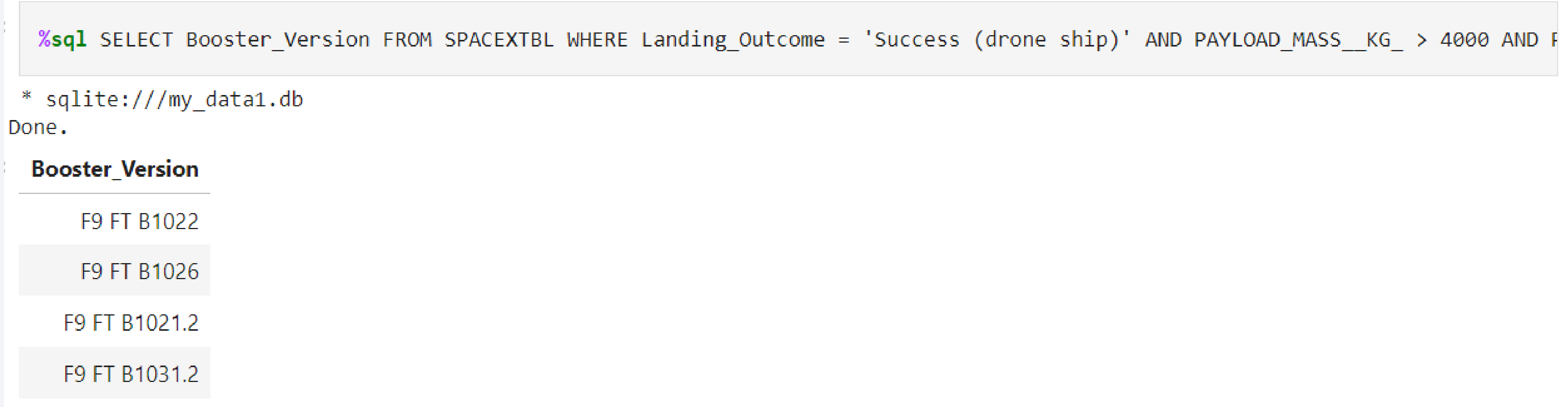
```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%'
```

```
 * sqlite:///my_data1.db
Done.
```

MIN("DATE")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Select the rows that meet the conditions of the landing outcome is Success (drone ship) and the payload mass is between 4000 and 6000 kg. Only print the booster version columns of the rows that meet the conditions.



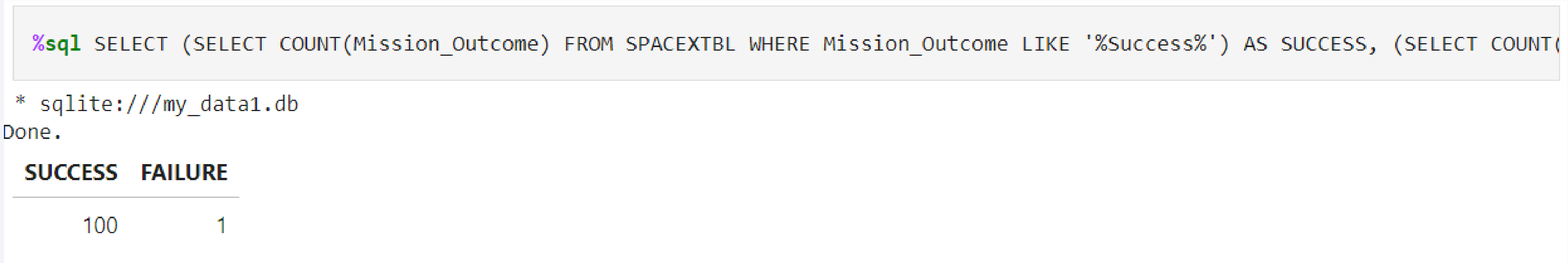
```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND P
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Use COUNT for count and SELECT FROM WHERE apply the filter for successful and failed landing

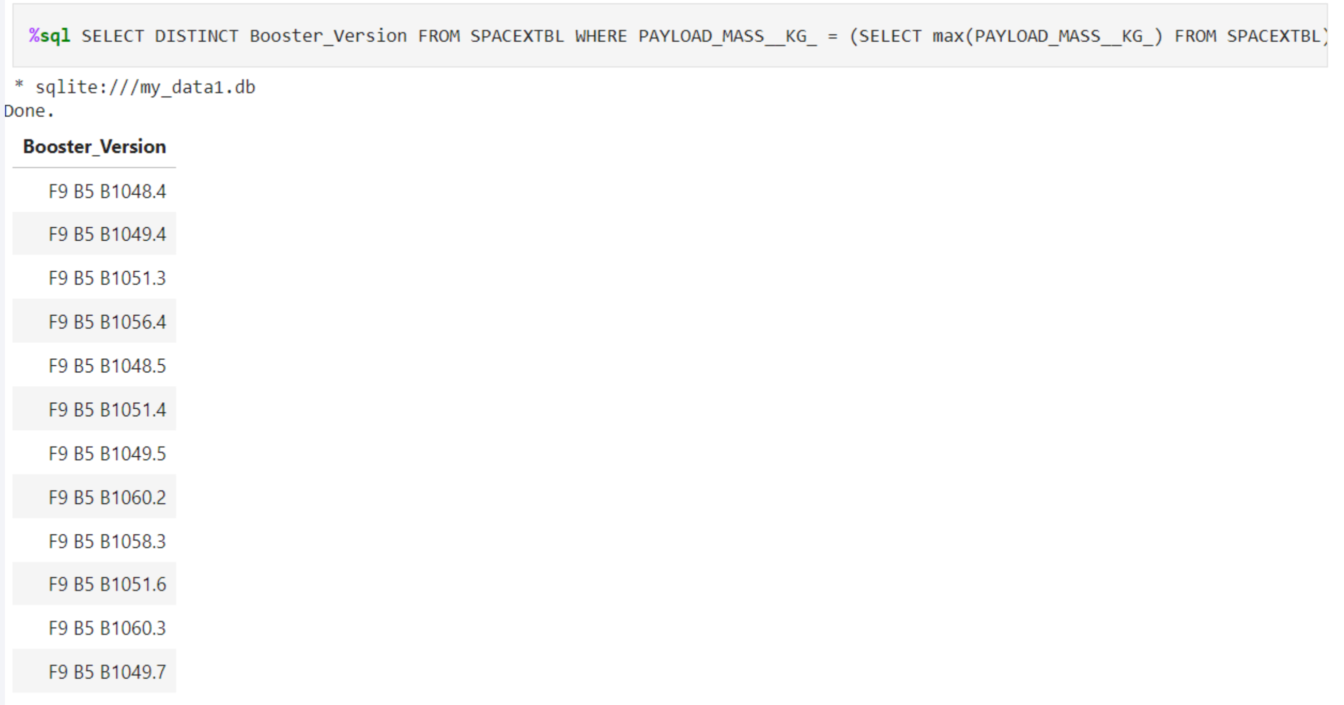


```
%sql SELECT (SELECT COUNT(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE '%Success%') AS SUCCESS, (SELECT COUNT(
```

```
 * sqlite:///my_data1.db
Done.
```

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload



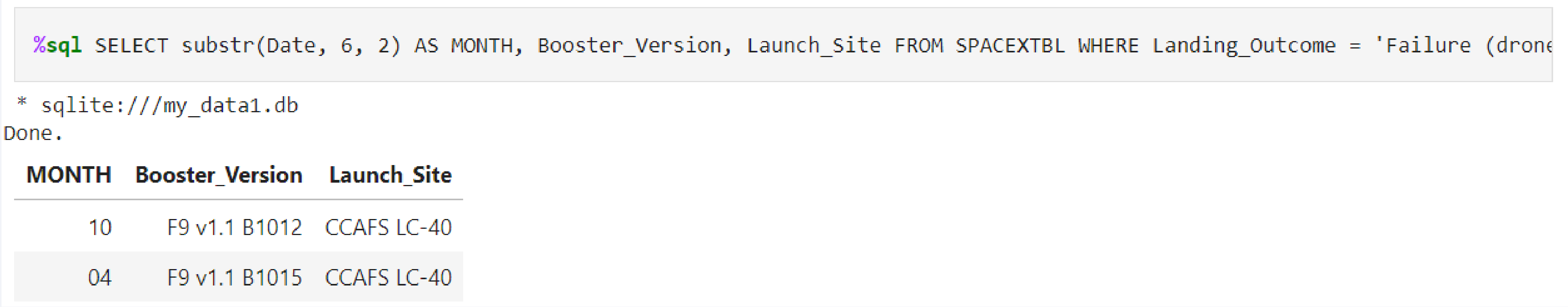
```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Firstly apply a MAX query to filter the data with highest weight of payload and them apply DISTINCT to find the unique booster version with heaviest payloads.

2015 Launch Records



```
%sql SELECT substr(Date, 6, 2) AS MONTH, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone
```

```
 * sqlite:///my_data1.db
Done.
```

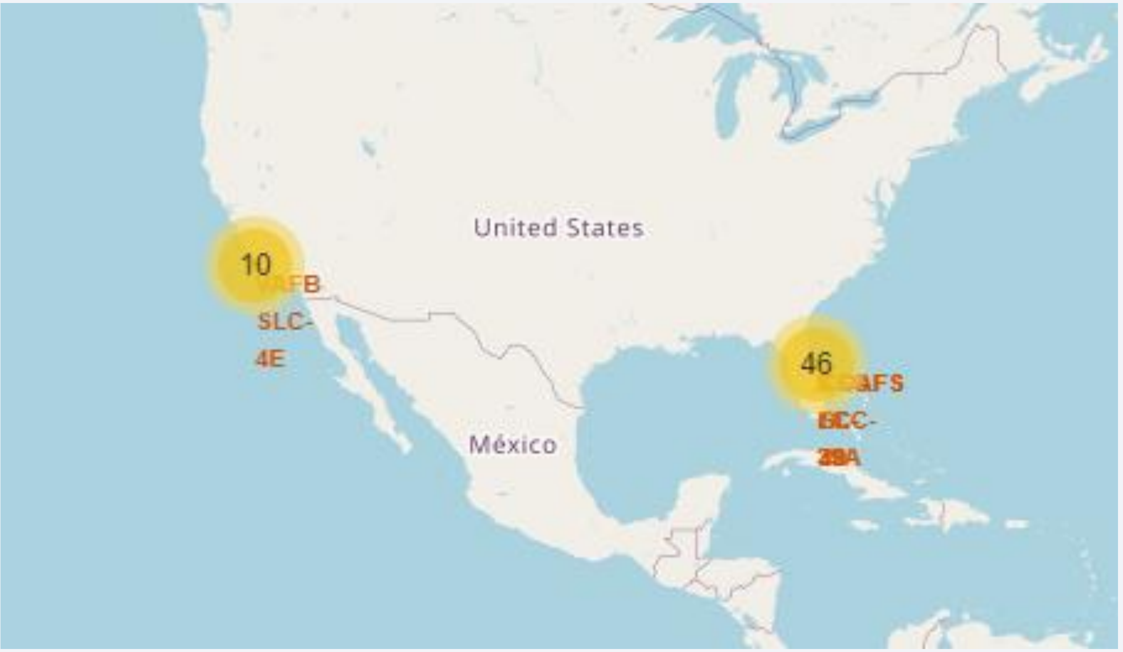
MONTH	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- There are two failed landing in drone ship, one was in April and one was in October of 2015

Section 3

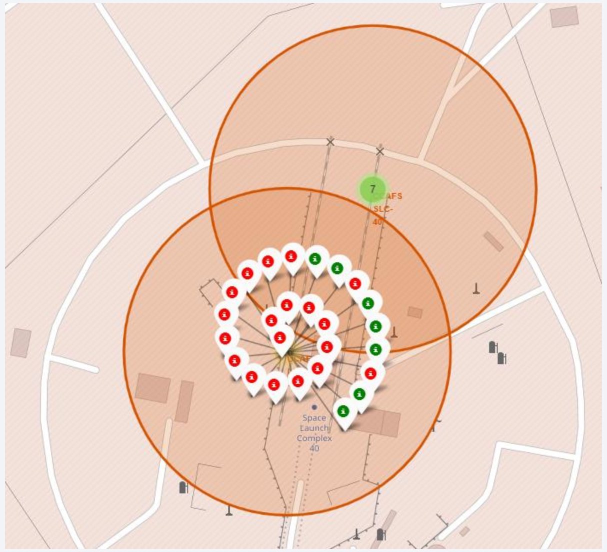
Launch Sites Proximities Analysis

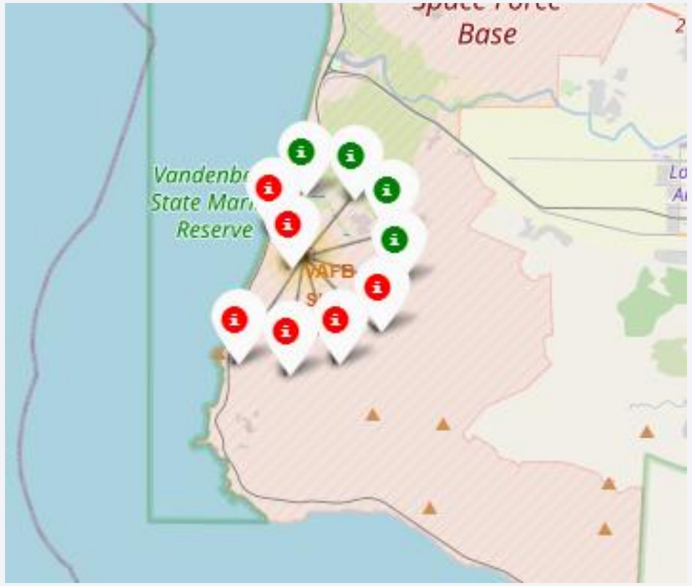
Overview of launch site locations

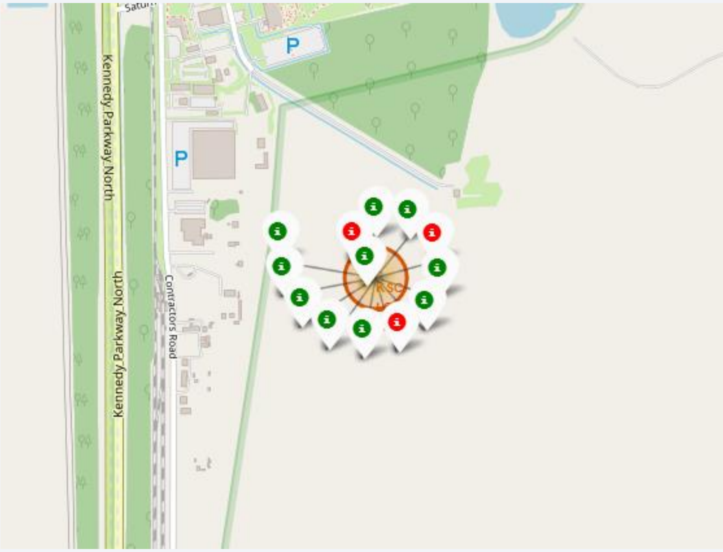


- The circles label the locations of launch sites.
- It can be seen that the launch sites are all located near the sea.

Launch Sites

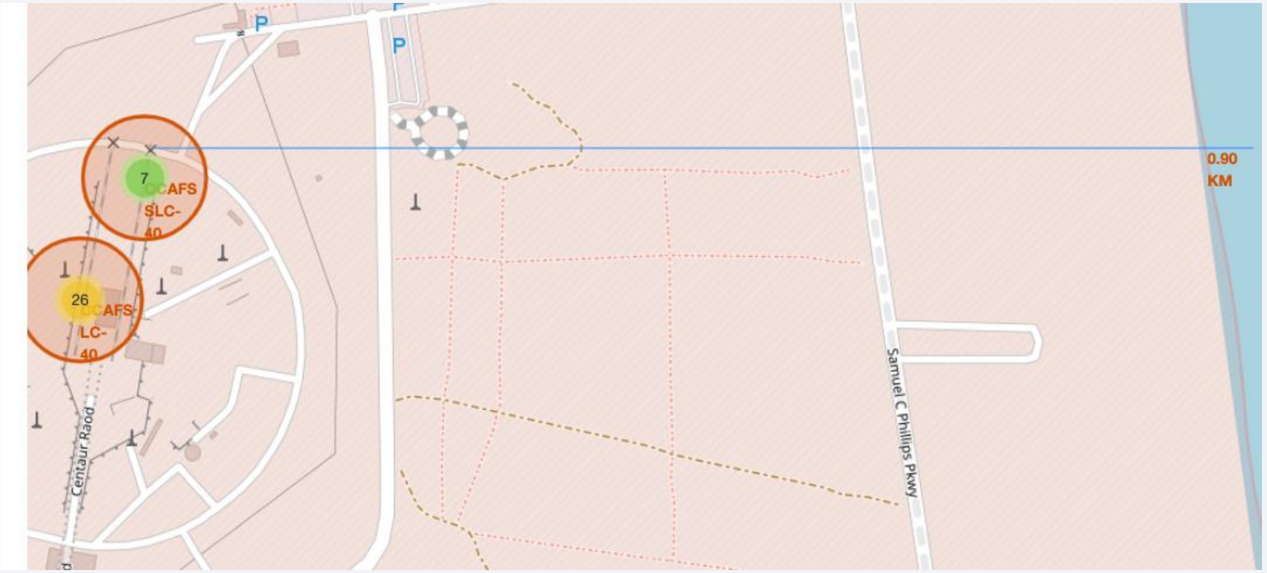




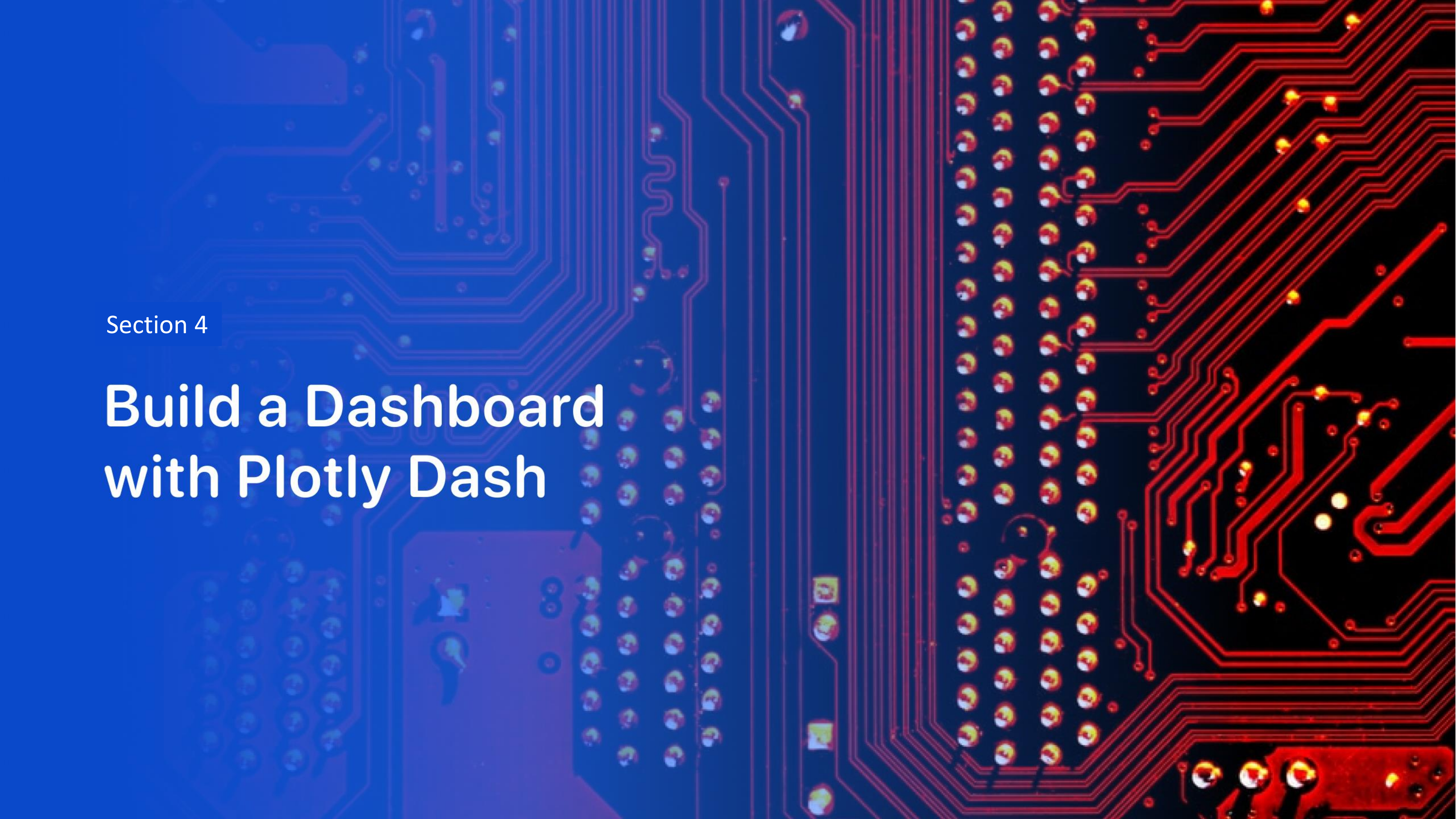


- The colored markers show each of the launches.
- Green indicates successful landing
- Red indicated failed landing

<Folium Map Screenshot 3>



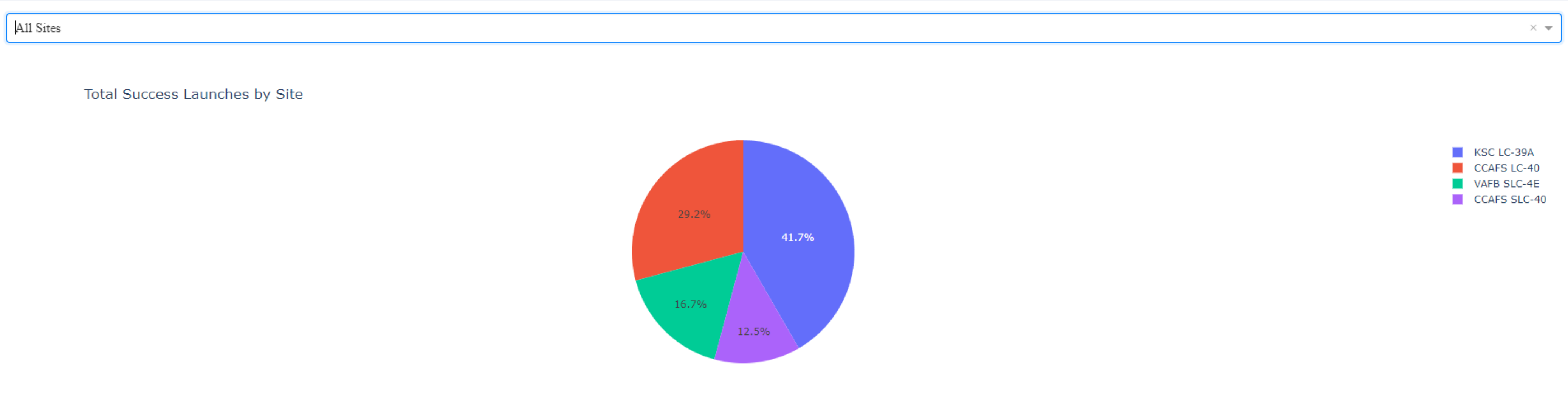
- This figure shows an example of labelling the launch site to some important locations.
- As discussed in the previous slides, the launch sites are all on the coast and are within a distance of 1km.



Section 4

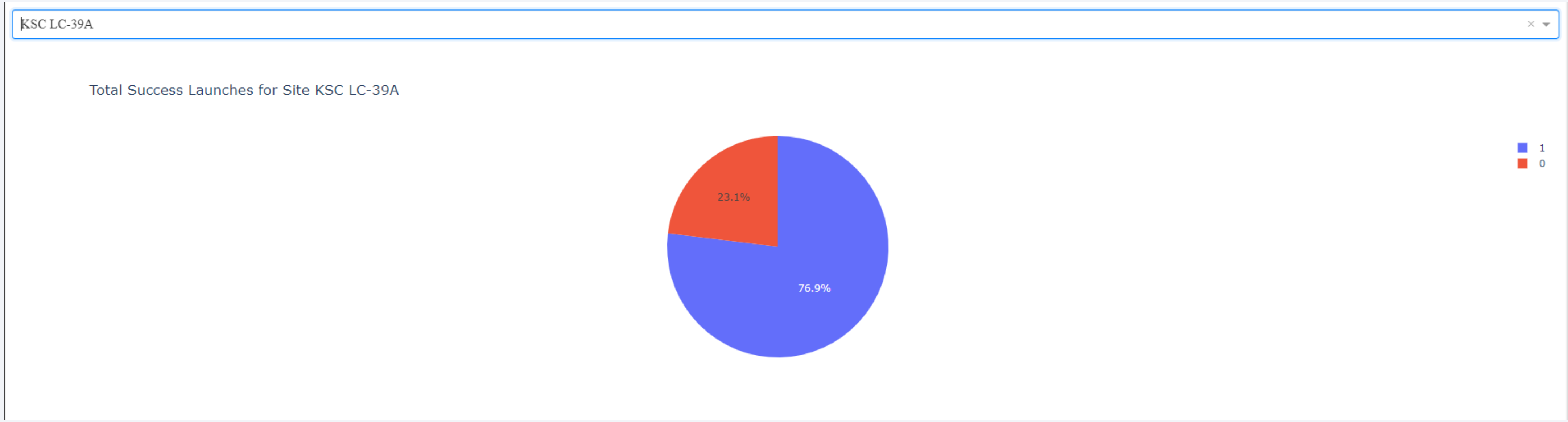
Build a Dashboard with Plotly Dash

Launch success count for all sites


All Sites

- It can be seen that the site KSC LC-39A has the highest overall success launch rate of 41.7%, while the site, CCAFS SLC-40 has the lowest overall success launch rate of 12.5%.

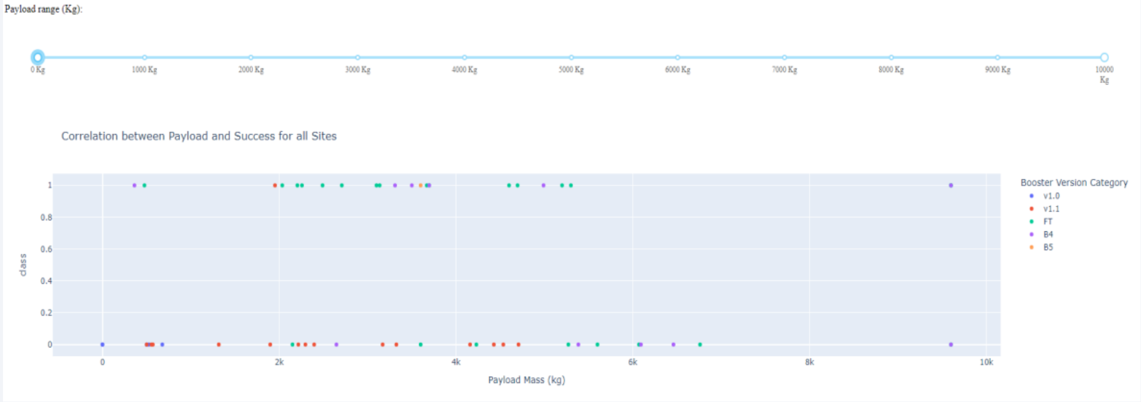
Launch success count for KSC LC-39A

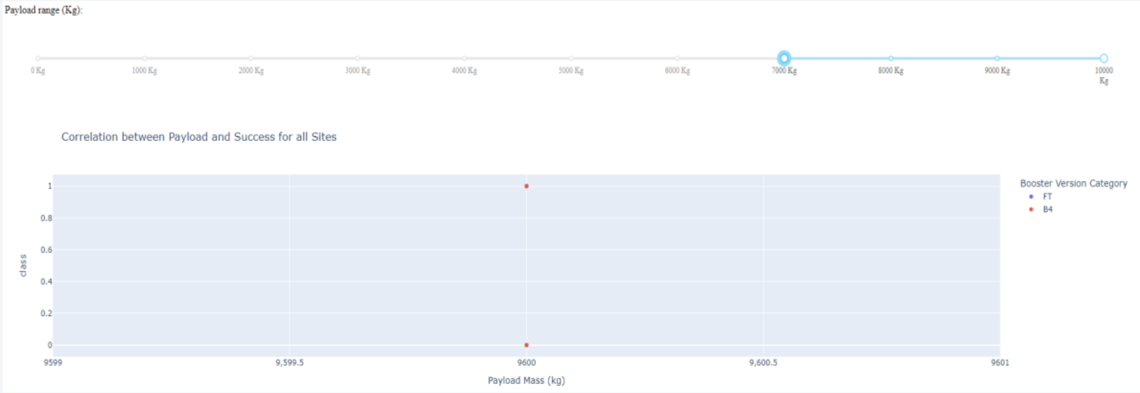


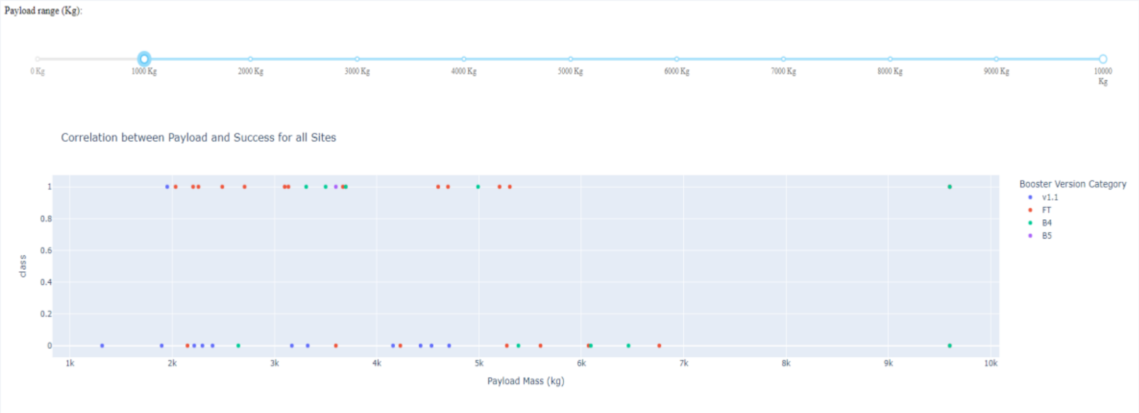
KSC LC-39A × ▾

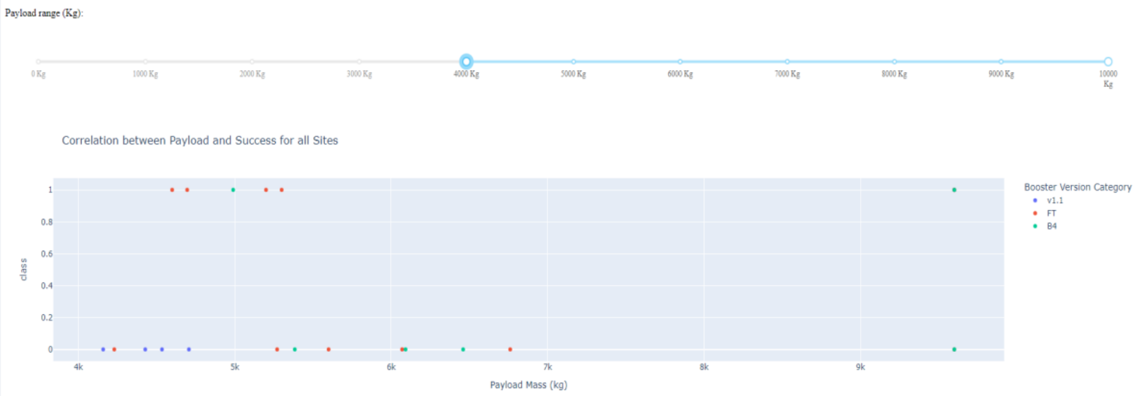
- Individually, the site KSC LC-39A has a successful launch rate of 76.9% and a failure rate of 23.1%.

Payload vs. Launch Outcome scatter plot for all sites







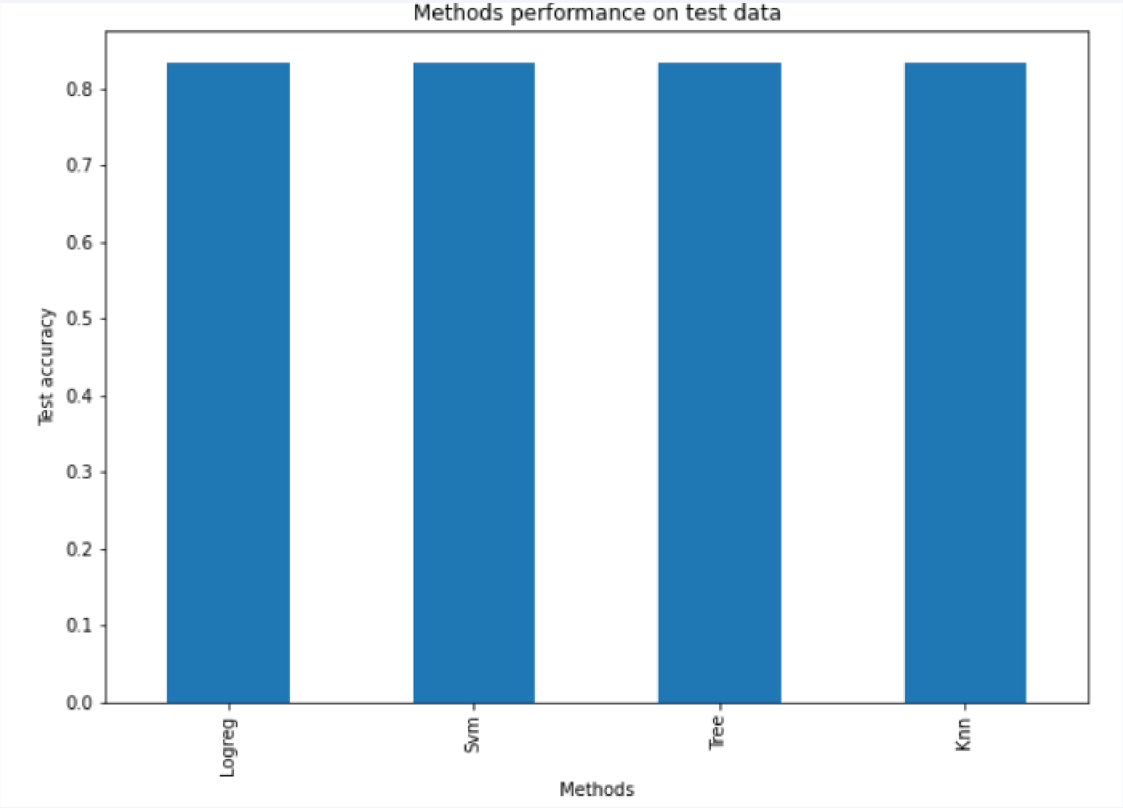


- Overall, launches with lower-weight payload have higher success rate of landing.

Section 5

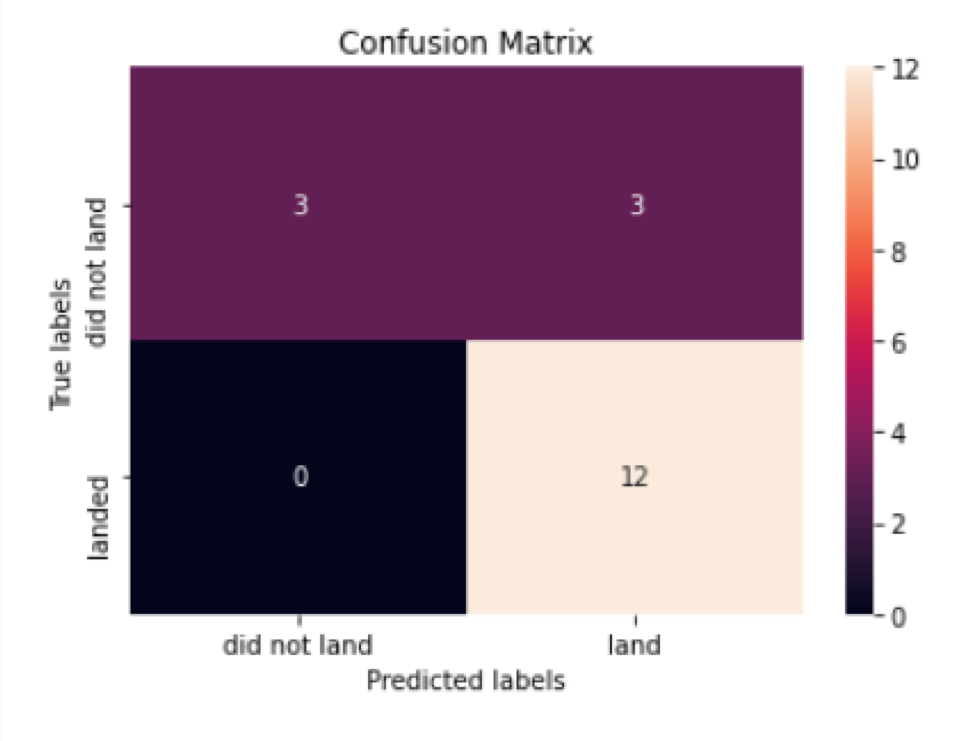
Predictive Analysis (Classification)

Classification Accuracy



- The four models have the same out of sample accuracy score of 0.8333.

Confusion Matrix



- The four models give the same confusion matrix

Conclusions

- The initial data visualization indicates that as the number of launch increases, the success rate of landing increases, which can be a result of data analysis.
- From the analysis, some important factors that impact the success rate of landing are orbit types, the location of launches and the weight of payload. Some other factor may include weather, time of landing and location of landing.
- According to the results, the launch site KSC LC-39A has the highest success rate of landing and the orbits, ES-L1, GEO, HEO and SSO, have the best landing records.
- To analyze the data, a logistic regression model was trained to predict the possibility of a launch to be successful or failed. The trained model has an accuracy of 83.3% which is acceptable.
- For a binary classification task in this case, the algorithms chosen in the lab, decision tree, logistic regression, KNN and SVM, have the same out of sample score, while decision tree has better performance on training set.

Thank you!