IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Taras Trofymchuk

Aug 30, 2023

Outline

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

Executive Summary

Summary of Methodologies

- Data collection using web scraping and SpaceX API
- Exploratory Data Analysis (EDA) including data wrangling, data visualization, and interactive visual analytics
- Predictive Analysis (Classification)

Summary of Results

- EDA results
- Interactive analytics (screenshots)
- Predictive analysis (Classification)

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website for a cost of 62 million dollars; while other providers offer rocket launches for up to 165 million dollars each. Much of the savings is because Space X can reuse the first stage.
  Space Y, a new competitor on the market, wants to bid against Space X. In order to do it successfully we need to predict weather the first stage will land, which determines the cost of the launch. This objective of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What influences if the rocket will land successfully?
  - The relationship between various features that determine the success rate of a landing.
  - What conditions need to be in place to ensure successful landing

Section 1

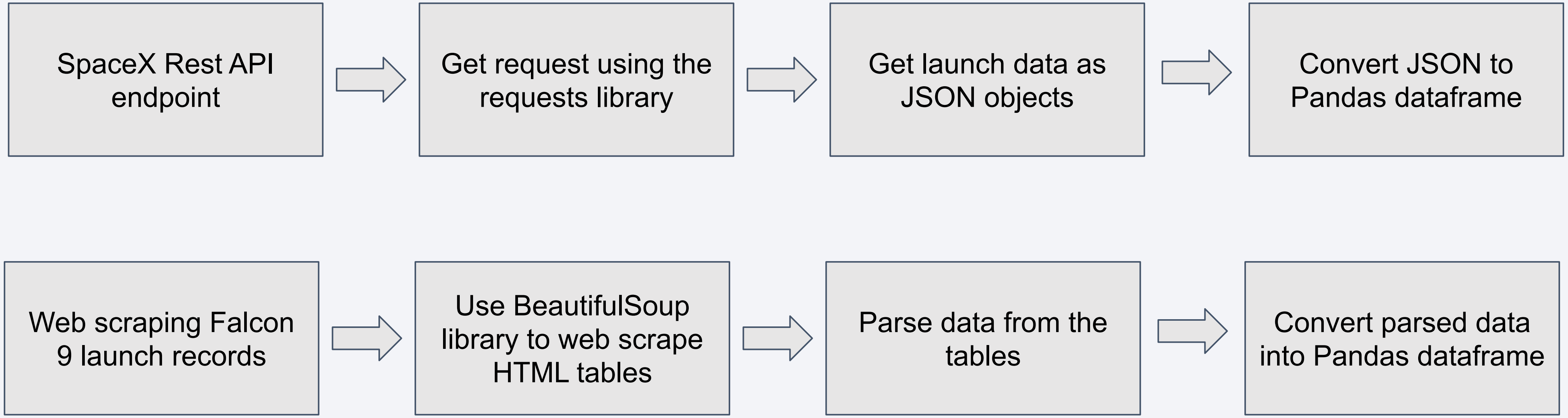
Methodology

Methodology

- Data collection methodology:
  - SpaceX Rest API
  - Web scraping from open sources (Wikipedia)
- Perform data wrangling
  - The data collected as JSON objects and HTML tables, cleaned and converted into Pandas dataframe for further analysis and visualization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - SVM, Classification Trees and Logistic Regression

Data Collection

Data was collected using SpaceX REST API and web scraped from wikipedia

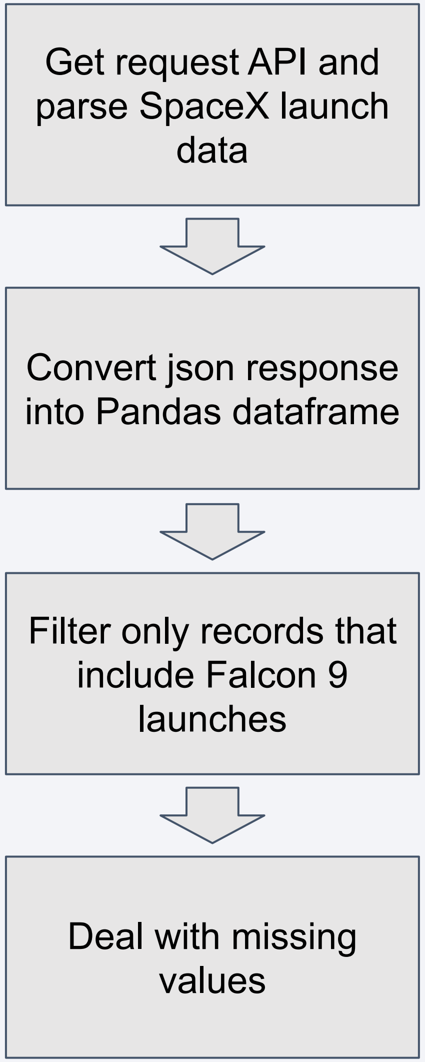


SpaceX Rest API endpoint → Get request using the requests library → Get launch data as JSON objects → Convert JSON to Pandas dataframe

Web scraping Falcon 9 launch records → Use BeautifulSoup library to web scrape HTML tables → Parse data from the tables → Convert parsed data into Pandas dataframe

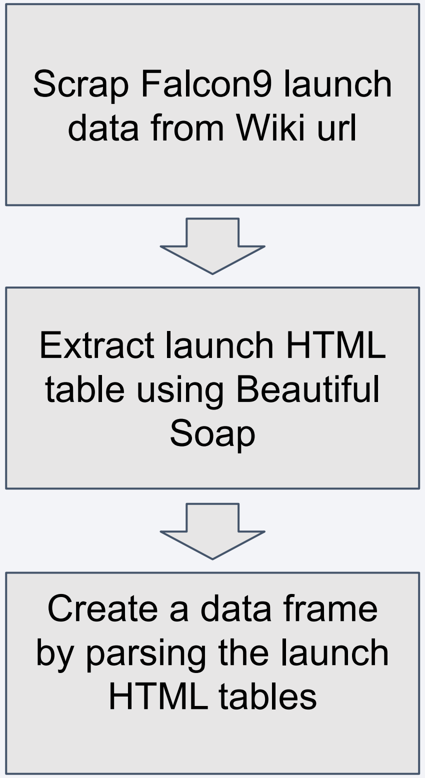
Data Collection – SpaceX API

- SpaceX offers a public REST API from where data was obtained.
- Follow the flow chart for the step by step description
- Data Collection Notebook: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/data-collection-api.ipynb


Get request API and parse SpaceX launch data
Convert json response into Pandas dataframe
Filter only records that include Falcon 9 launches
Deal with missing values

Data Collection - Scraping

- Alternatively data for SpaceX launches can also be obtained from Wikipedia using web scraping
- Follow the flow chart for the step by step description
- Web Scraping Notebook: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/data-webscraping.ipynb



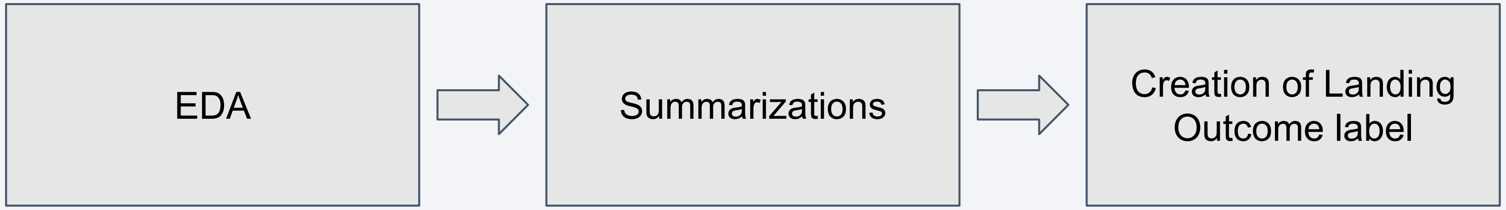
Scrap Falcon9 launch data from Wiki url

Extract launch HTML table using Beautiful Soap

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initial Exploratory Data Analysis (EDA) was performed on the data set.
- Summarized launches per site, occurrences of each orbit and mission outcomes grouped by orbit type.
- The Landing Outcome label was created from Outcome column
- Data Wrangling Notebook: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/data-wrangling.ipynb



EDA → Summarizations → Creation of Landing Outcome label

EDA with Data Visualization

- To explore data, scatter plots and bar plots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number
  - Launch Site X Flight Number
  - Launch Site X Payload Mass
  - Orbit and Flight Number
  - Payload and Orbit
- Data Visualization Notebook: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
    - Names of the unique launch sites in the space mission
    - Top 5 launch sites whose name begin with the string 'CCA’
    - Total payload mass carried by boosters launched by NASA(CRS)
    - Average payload mass carried by booster version F9 v1.1
    - Date when the first successful landing outcome in ground pad was achieved
    - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
    - Total number of successful and failure mission outcomes
    - Names of the booster versions which have carried the maximum payload mass
    - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
    - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20
- SQL Notebook: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/eda-sql.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates

- Folium Notebook:https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/launch-site-location-analysis.ipynb

Build a Dashboard with Plotly Dash

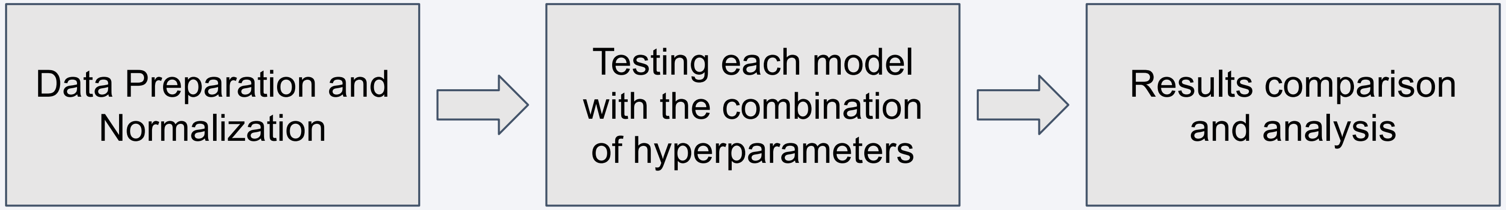
- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range

- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- Dashboard Script: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Machine Learning Notebook: https://github.com/taras-trofymchuk/AppliedDataScienceCapstone/blob/main/machine-learning-prediction.ipynb



Data Preparation and Normalization

Testing each model with the combination of hyperparameters

Results comparison and analysis

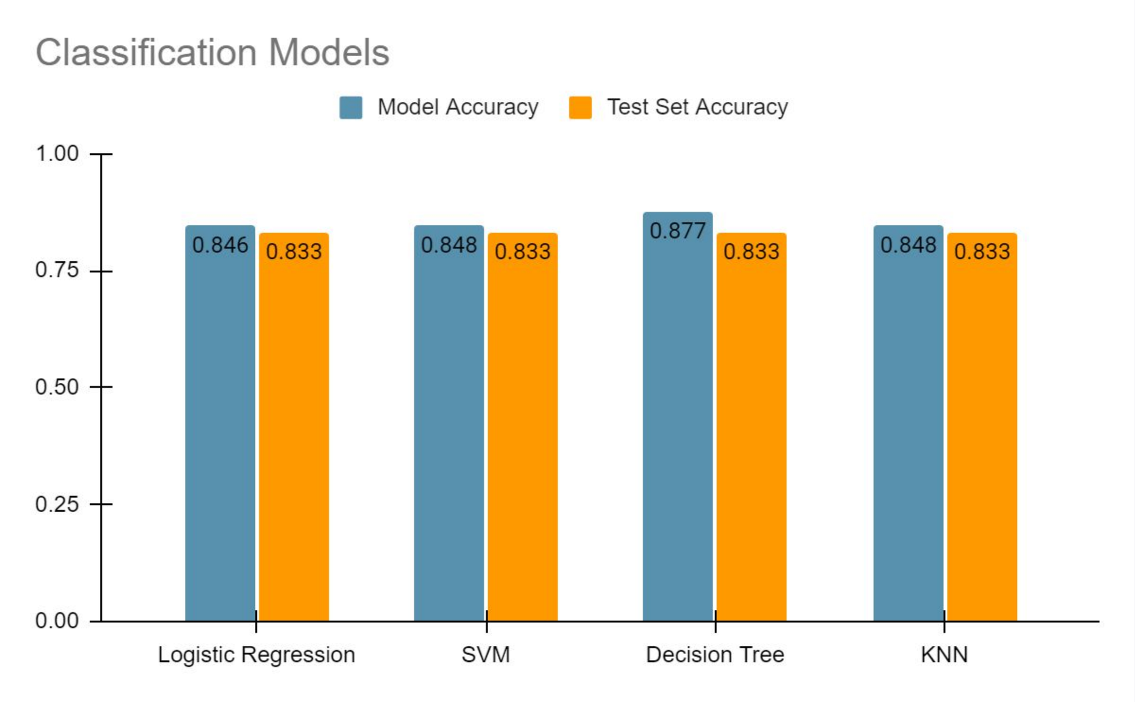
Results

Exploratory data analysis results:

- SpaceX uses 4 different launch sites
- The first launches were done to SpaceX itself and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first successful landing outcome happened in 2015 fiver year after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- Almost 100% of mission outcomes were successful
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- The success rate of landing outcomes improves with time

Results

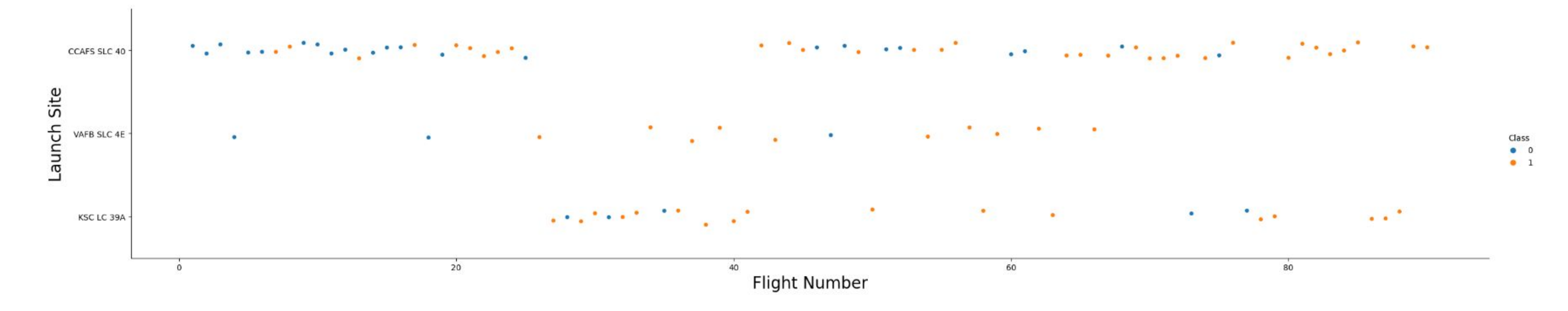
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, with accuracy over 87% and accuracy for test data over 83%.
- However the other classification models showed comparable results



Section 2

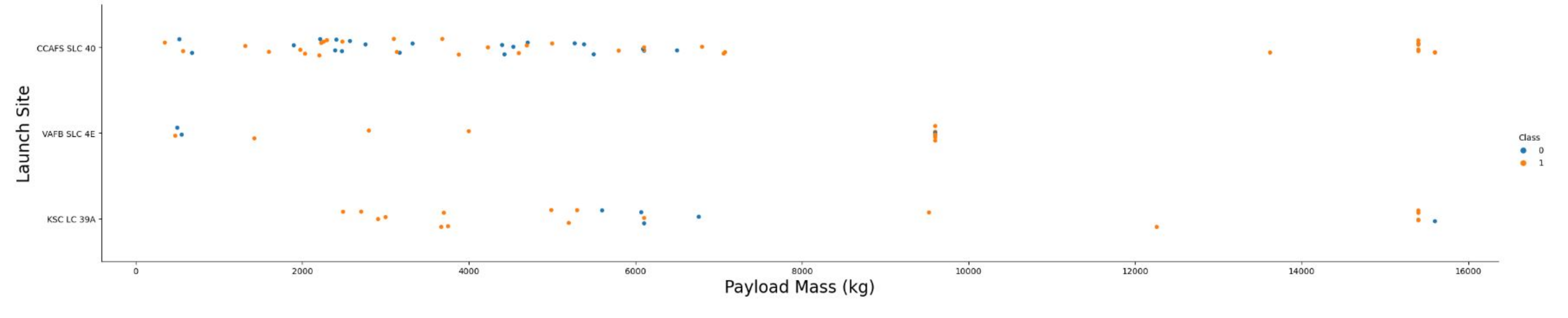
Insights drawn from EDA

Flight Number vs. Launch Site



- With the flight number (read the time) the success rate has increased for each launch site
- The majority of launches were done from CCAFS SLC40, the other two launch sites represent around one third of the launches

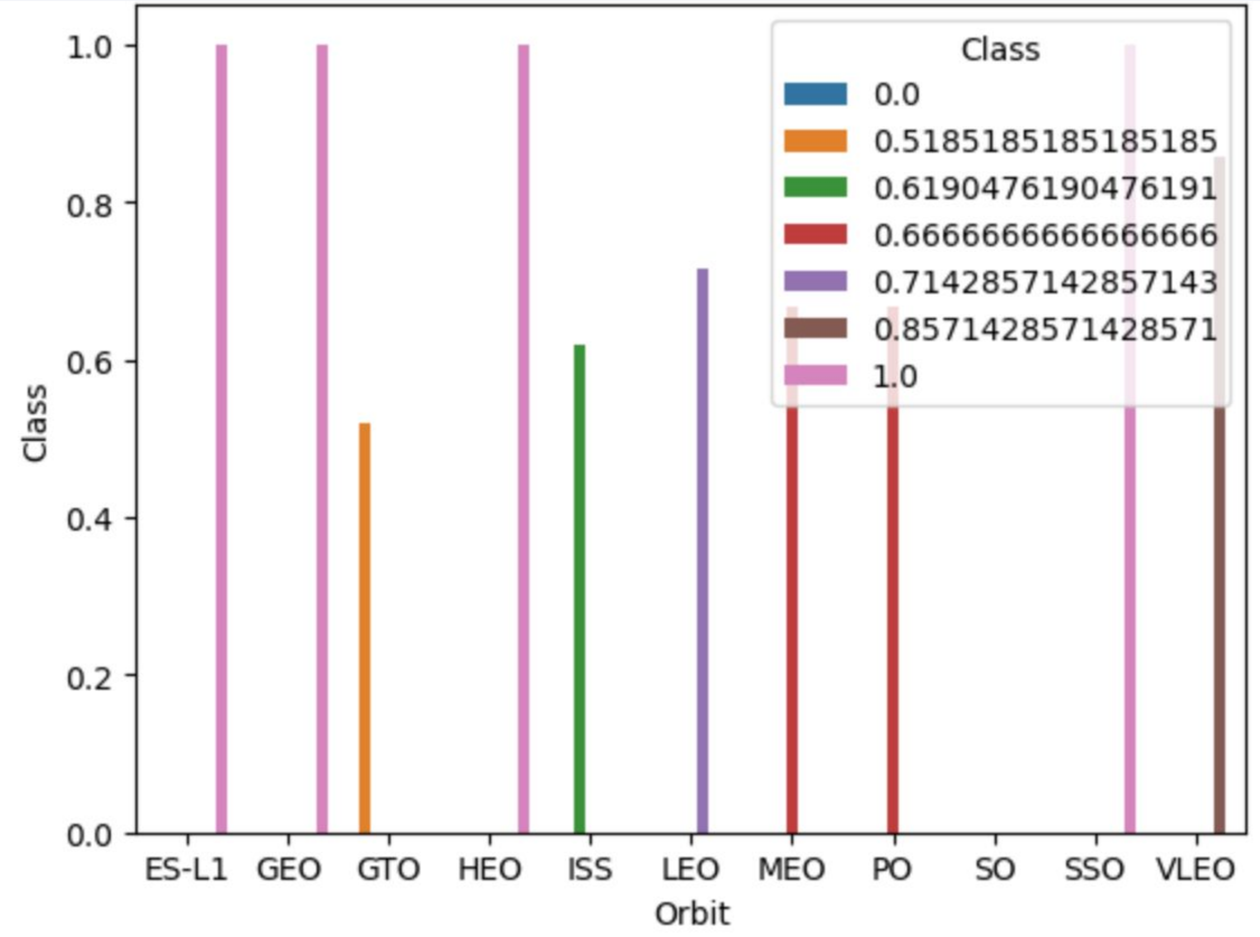
Payload vs. Launch Site



- With the flight number (read the time) the success rate has increased for each launch site
- The majority of launches were done from CCAFS SLC40, the other two launch sites together represent around one third of the launches

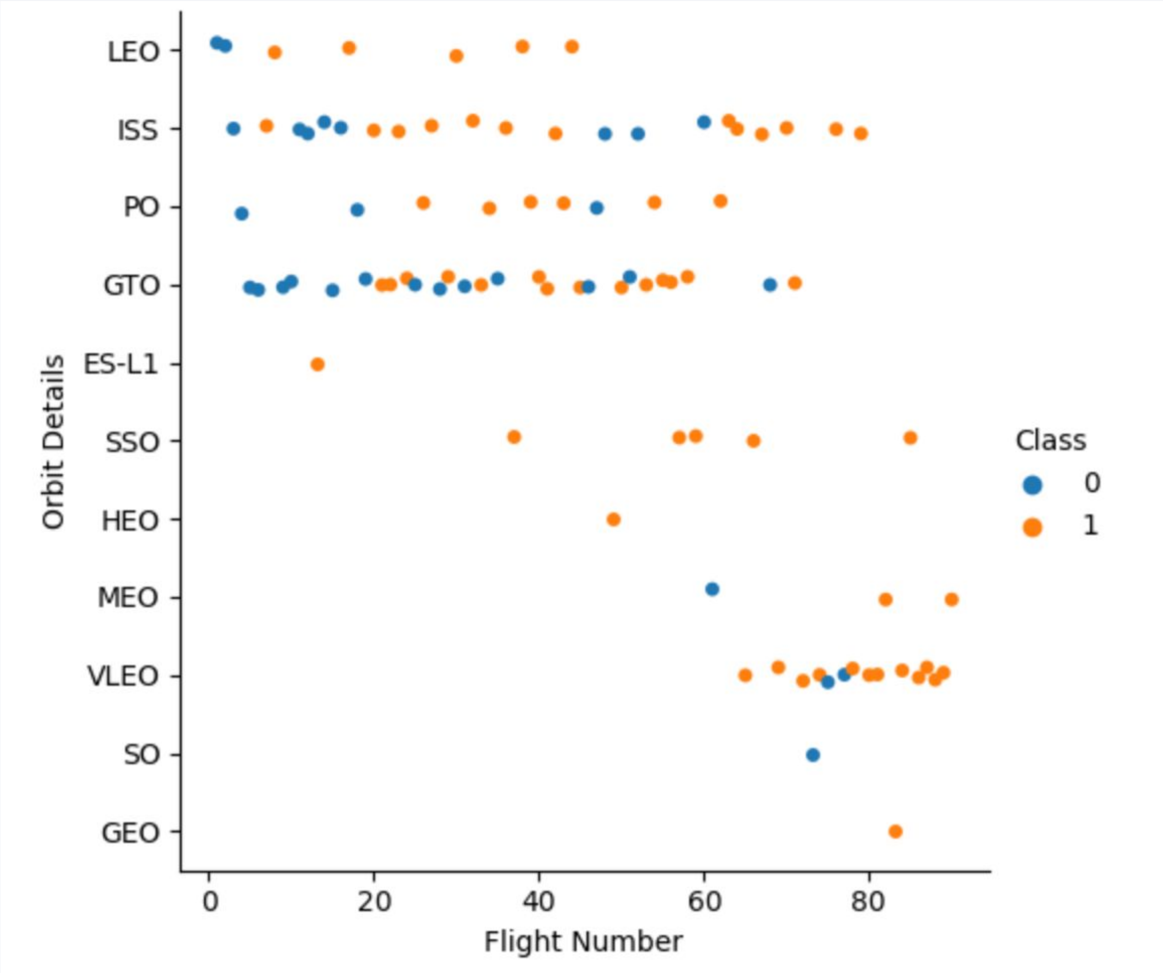
Success Rate vs. Orbit Type

- There are 4 Orbit types that have 100% success rate. However, this may not be representative of anything as the number of launches is not clear
- This chart should be interpreted in connection with the number of launches per orbit type



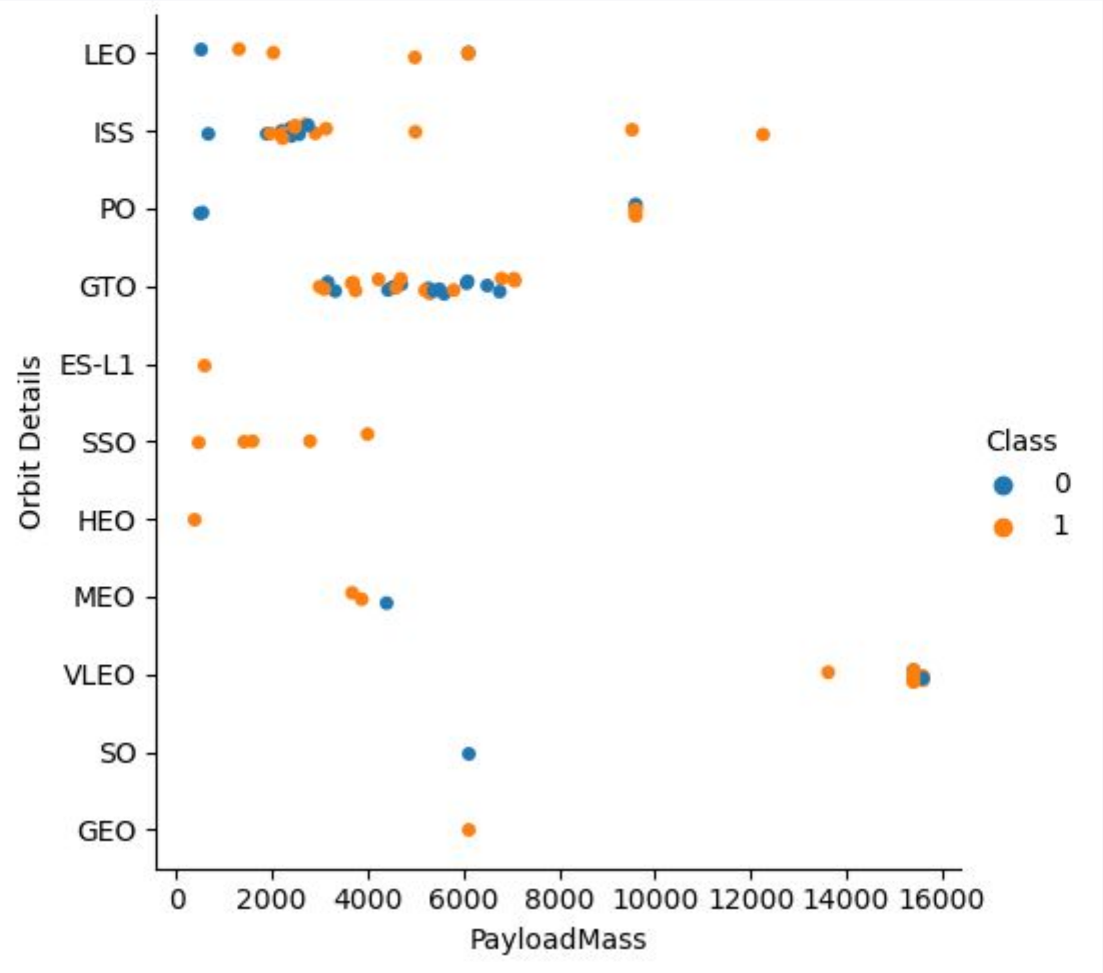
Flight Number vs. Orbit Type

- As expect the success rate improves with the time (flight number)
- It looks the tipping point was around 40th launch, when both success rate and variety of orbit types significantly improved
- GTO and ISS orbit types have the highest number of launches, although the highest number of failures
- The orbit type with 100% success rate have less than 6 launches each



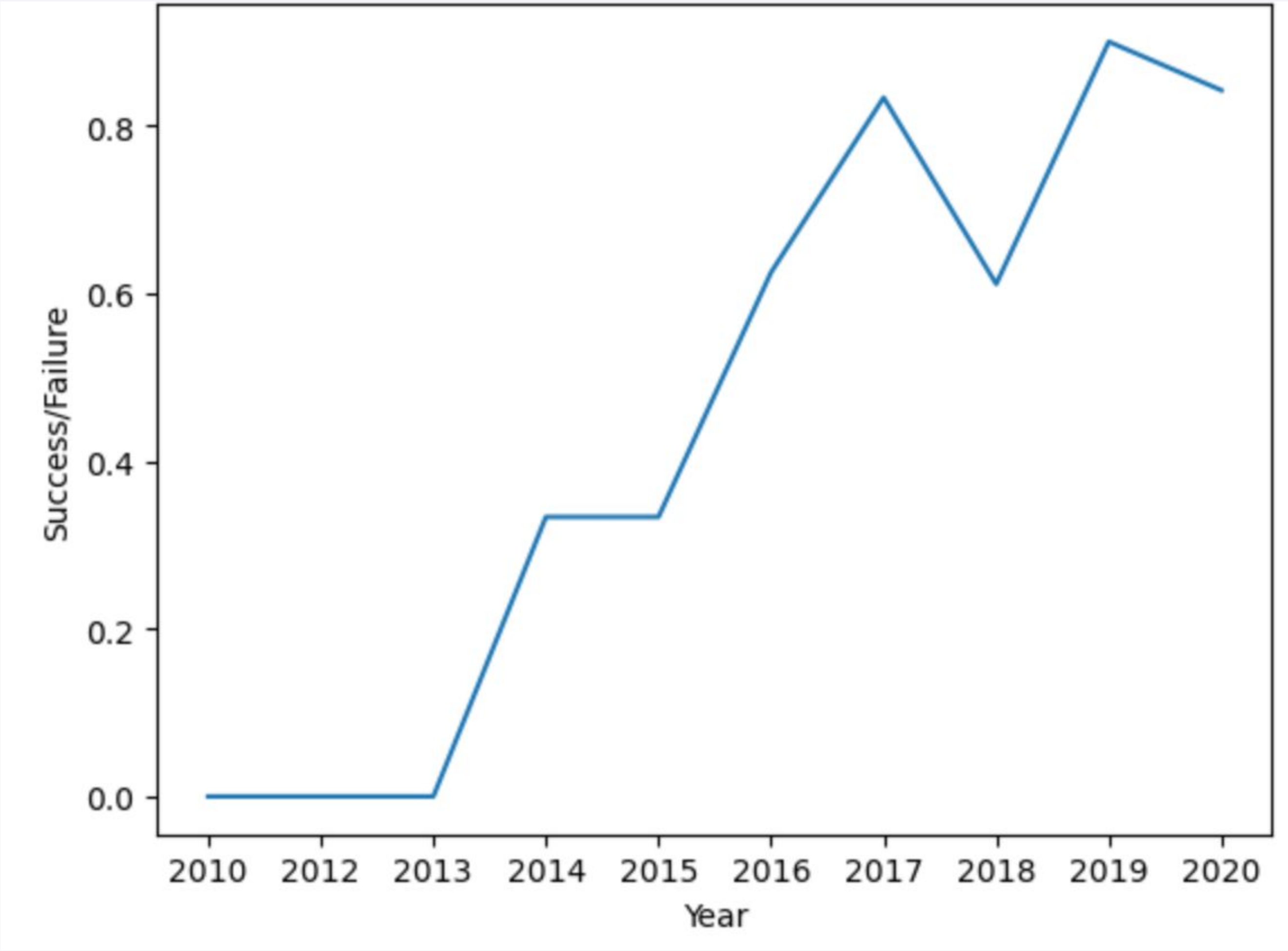
Payload vs. Orbit Type

- 7600 kg seems to represent some kind of the threshold, with barely over 10 launches exceeding this limit
-



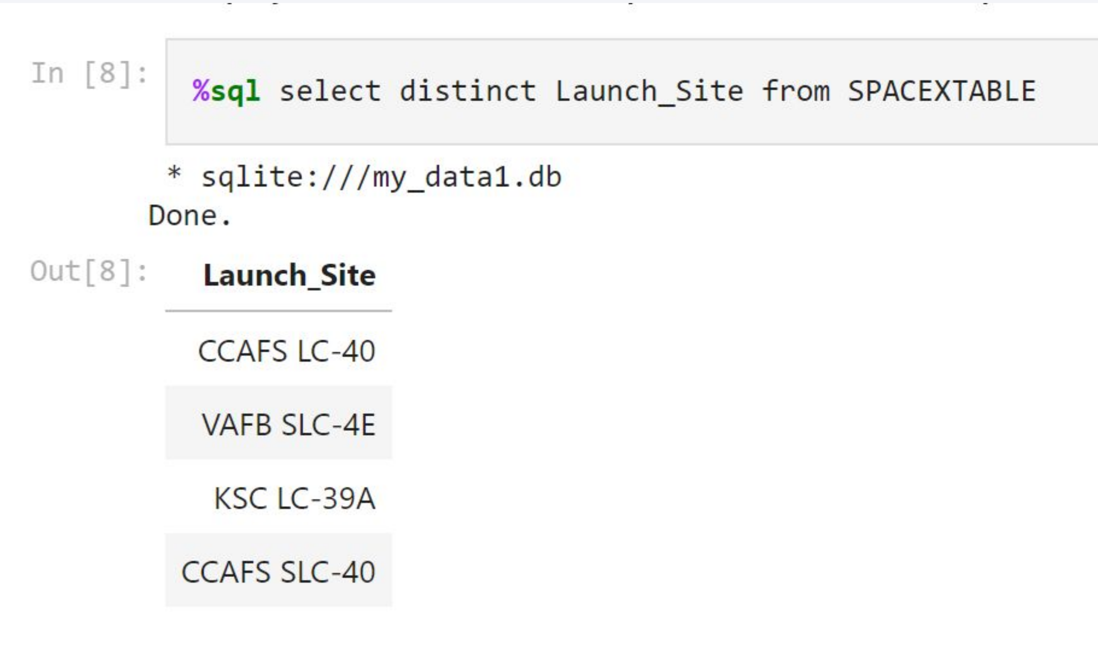
Launch Success Yearly Trend

- The success rate keeps improving since 2013



All Launch Site Names

- 4 unique launch sites were used to launch space mission
- Used distinct statement to identify unique values in the launch site column



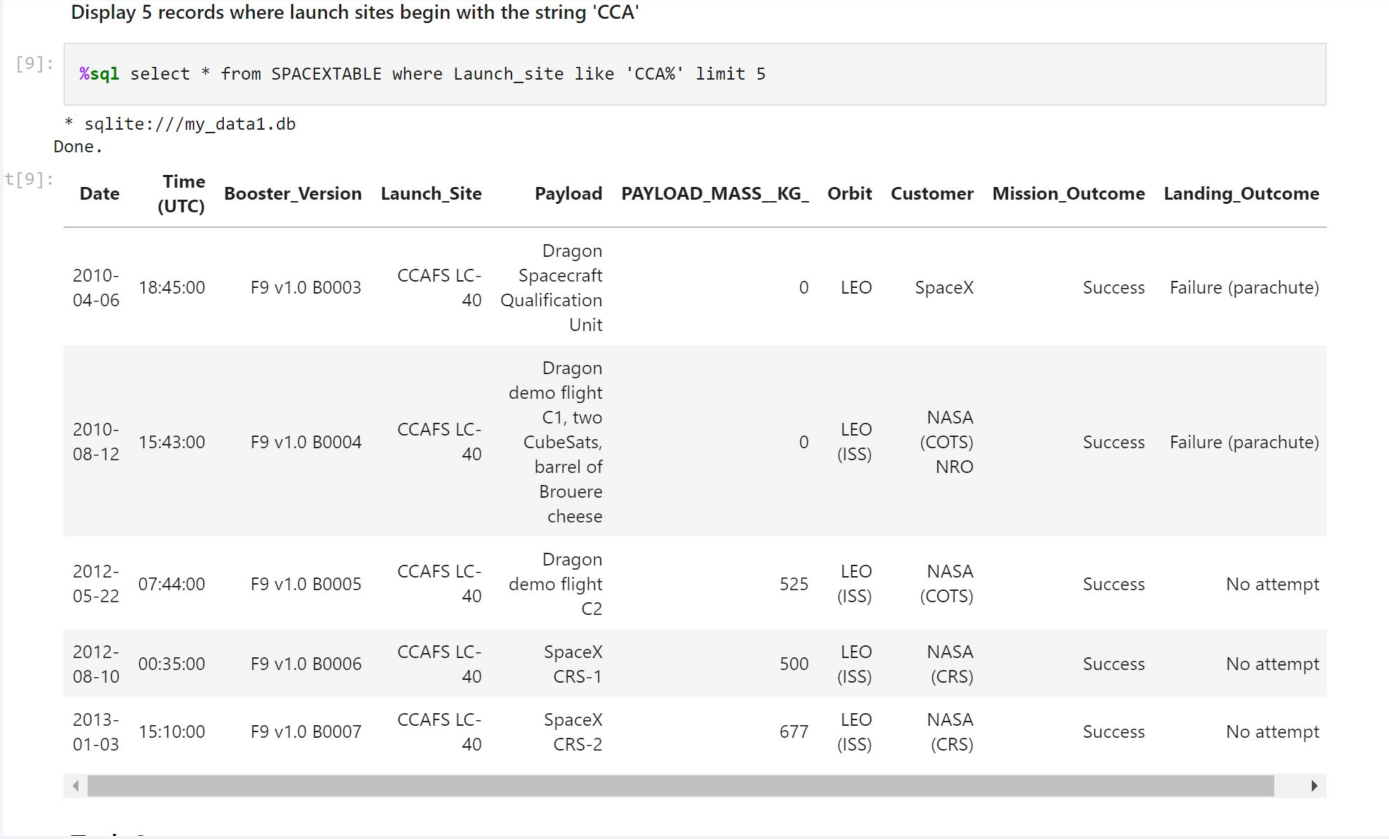
```
In [8]:  %sql select distinct Launch_Site from SPACEXTABLE

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

- Used limit clause to show only 5 records
- Used where and like clause in the query to find launch site names that begin with CCA



Display 5 records where launch sites begin with the string 'CCA'

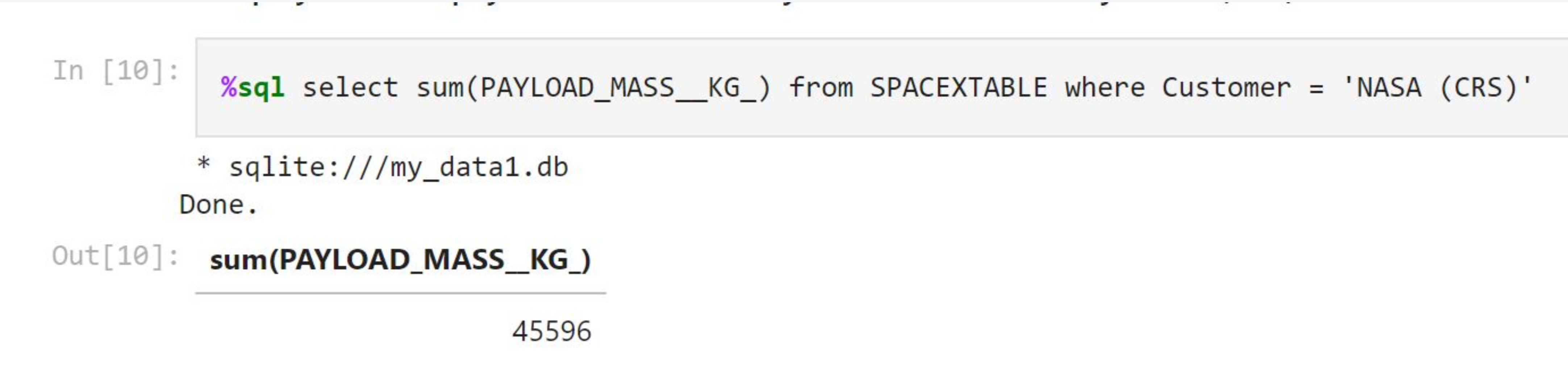
```
[9]: %sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5
```

```
 * 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

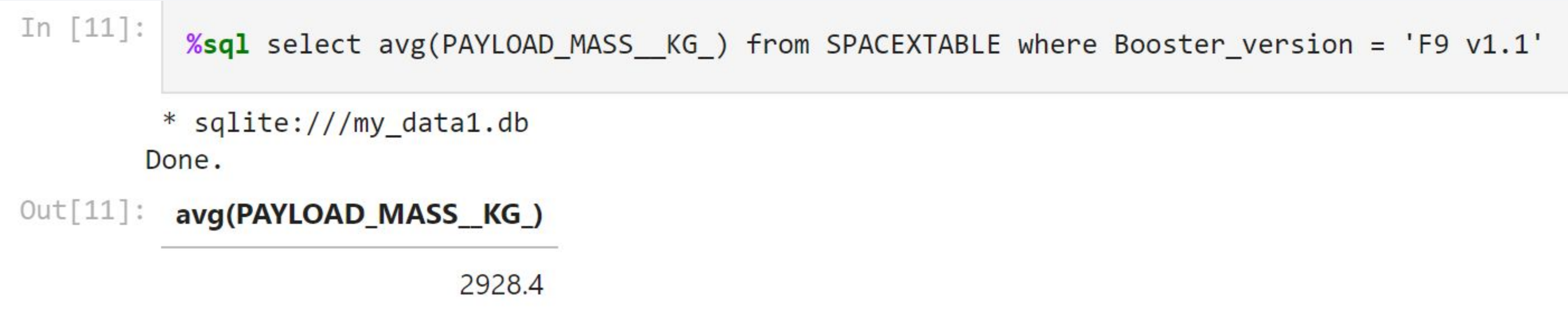
- The total payload mass carried by boosters launched by NASA (CRS) using sum function and where clause



```
In [10]:  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'

           * sqlite:///my_data1.db
          Done.

Out[10]:  sum(PAYLOAD_MASS__KG_)
                             45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by F9 v1.1. booster using avg function and where clause



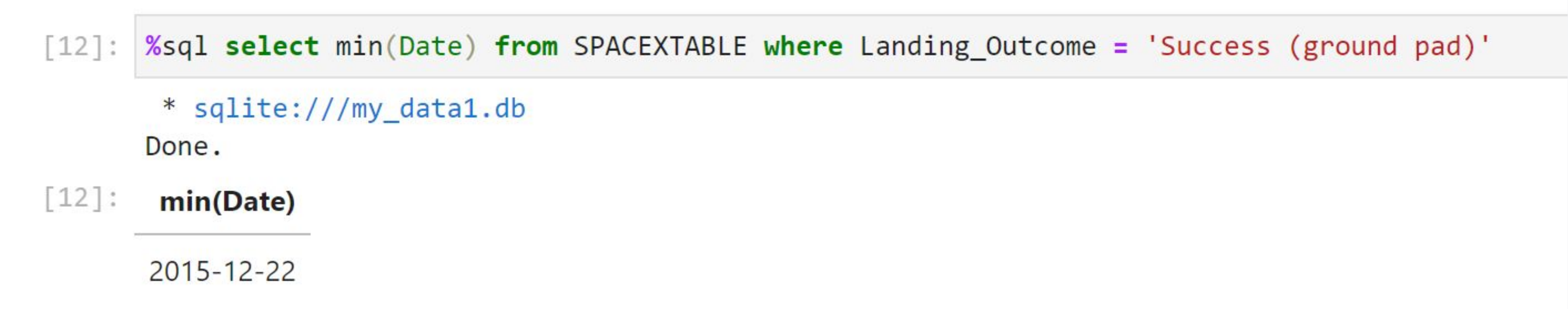
```
In [11]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_version = 'F9 v1.1'

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

Out[11]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

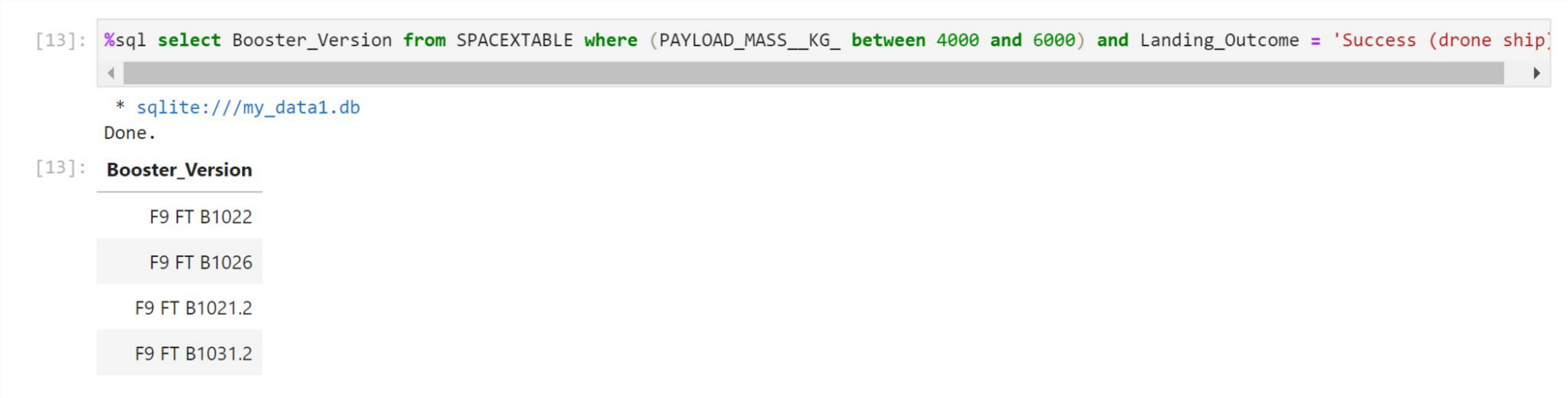
- First successful ground landing date using min function



```
[12]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'

 * sqlite:///my_data1.db
Done.

[12]:  min(Date)

       2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 using between clause



```
[13]: %sql select Booster_Version from SPACEXTABLE where (PAYLOAD_MASS__KG_ between 4000 and 6000) and Landing_Outcome = 'Success (drone ship
```

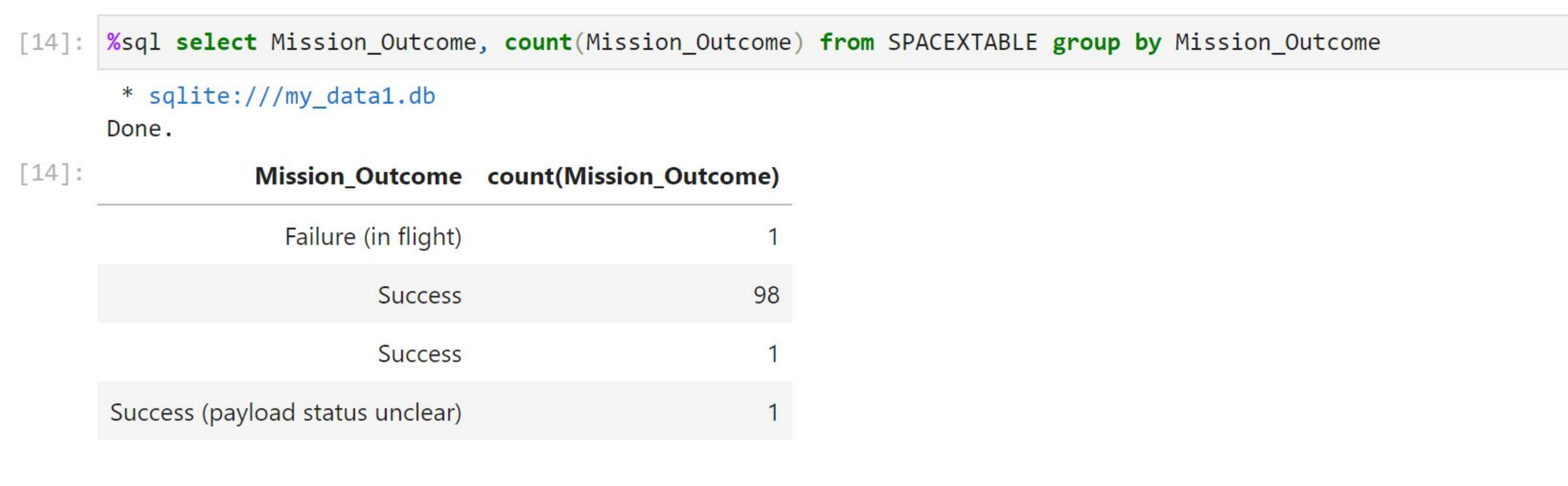
```
 * sqlite:///my_data1.db
Done.
```

[13]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes using count function and group by clause



```
[14]: %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

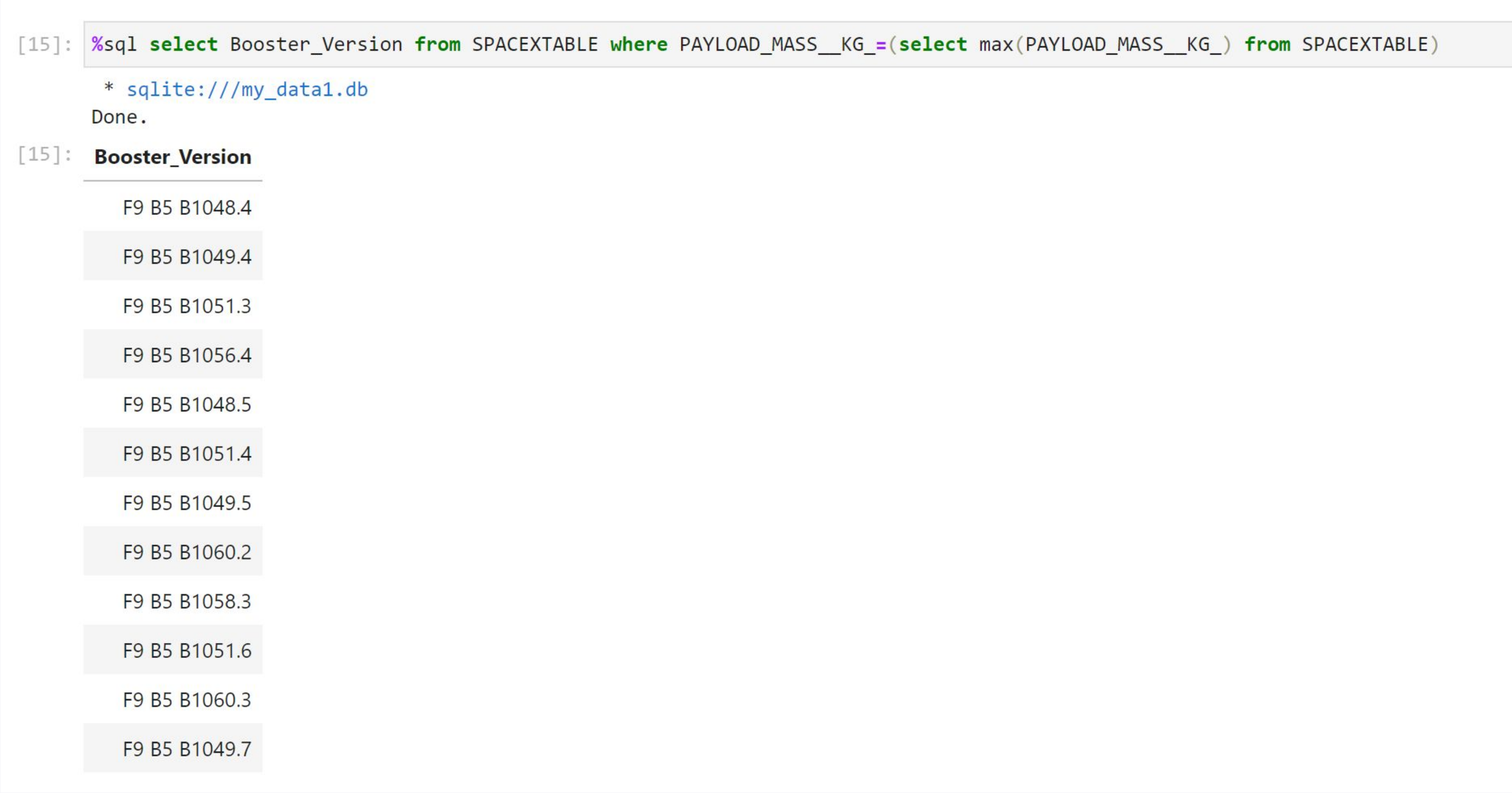
```
 * sqlite:///my_data1.db
Done.
```

[14]:

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listed the names of the booster which have carried the maximum payload mass using a subquery



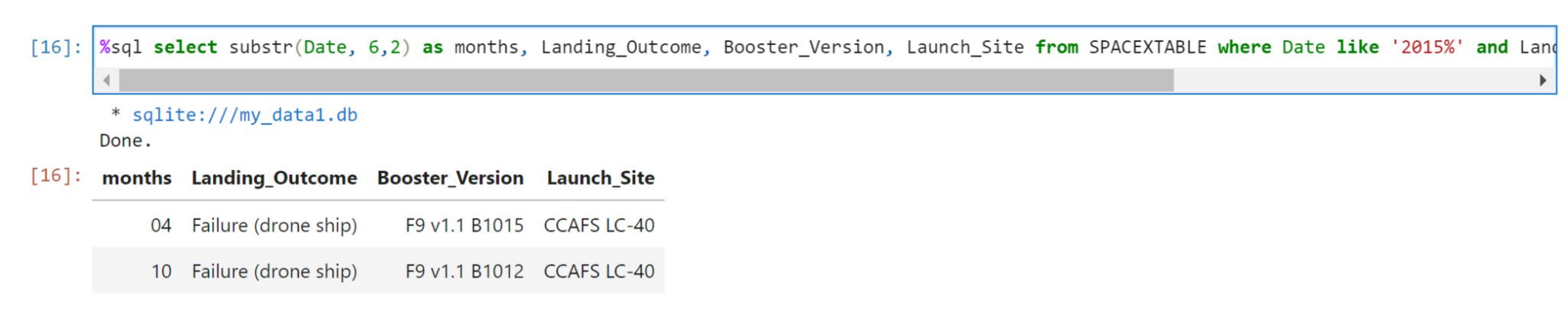
```
[15]: %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

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

[15]:

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

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



```
[16]: %sql select substr(Date, 6,2) as months, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Date like '2015%' and Land
```

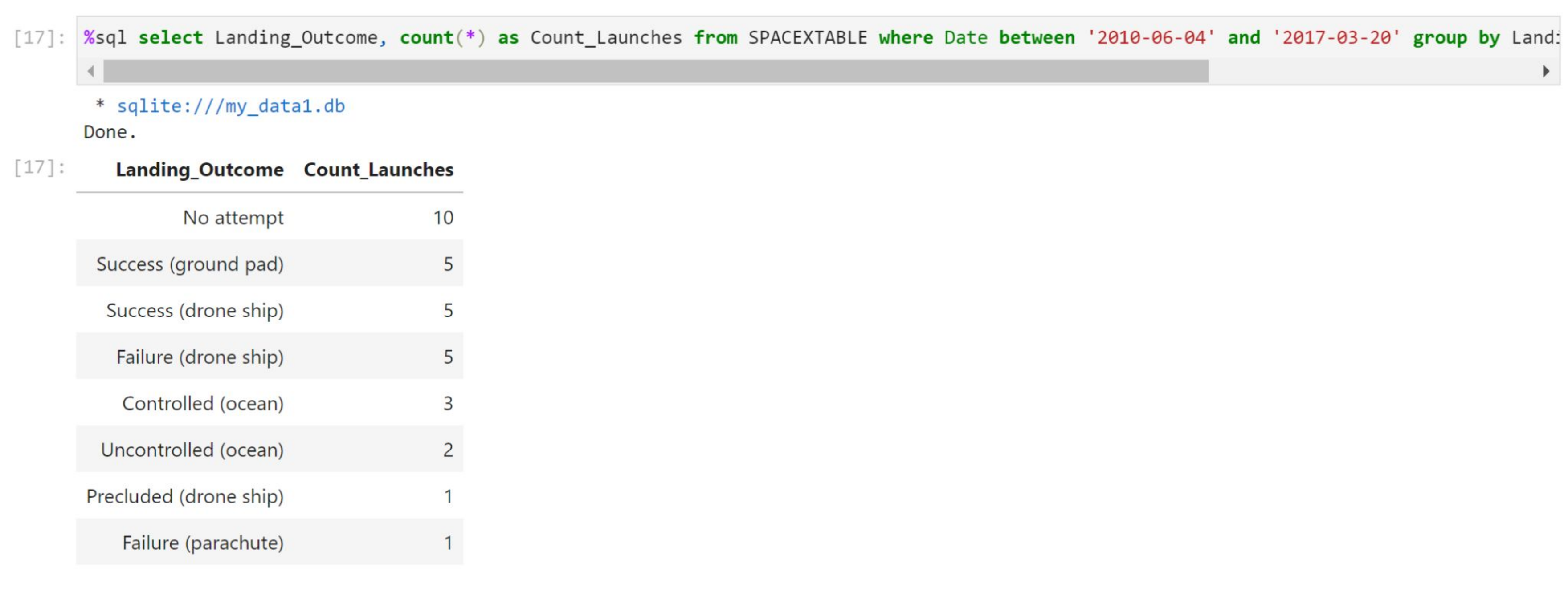
```
 * sqlite:///my_data1.db
Done.
```

[16]:

months	Landing_Outcome	Booster_Version	Launch_Site
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
10	Failure (drone ship)	F9 v1.1 B1012	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. The query uses count, where, between and group by



```
[17]: %sql select Landing_Outcome, count(*) as Count_Launches from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landi
```

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

[17]:

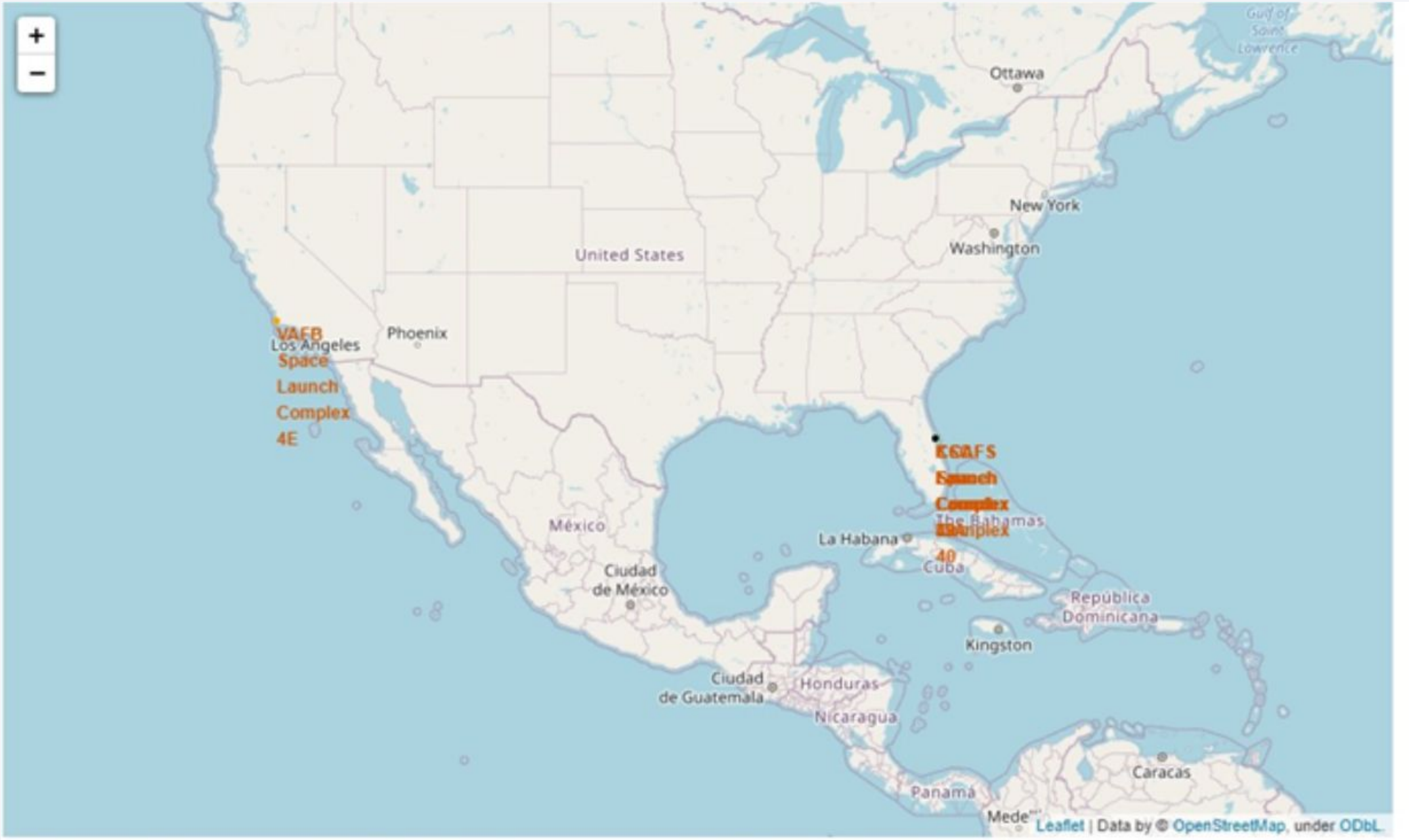
Landing_Outcome	Count_Launches
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

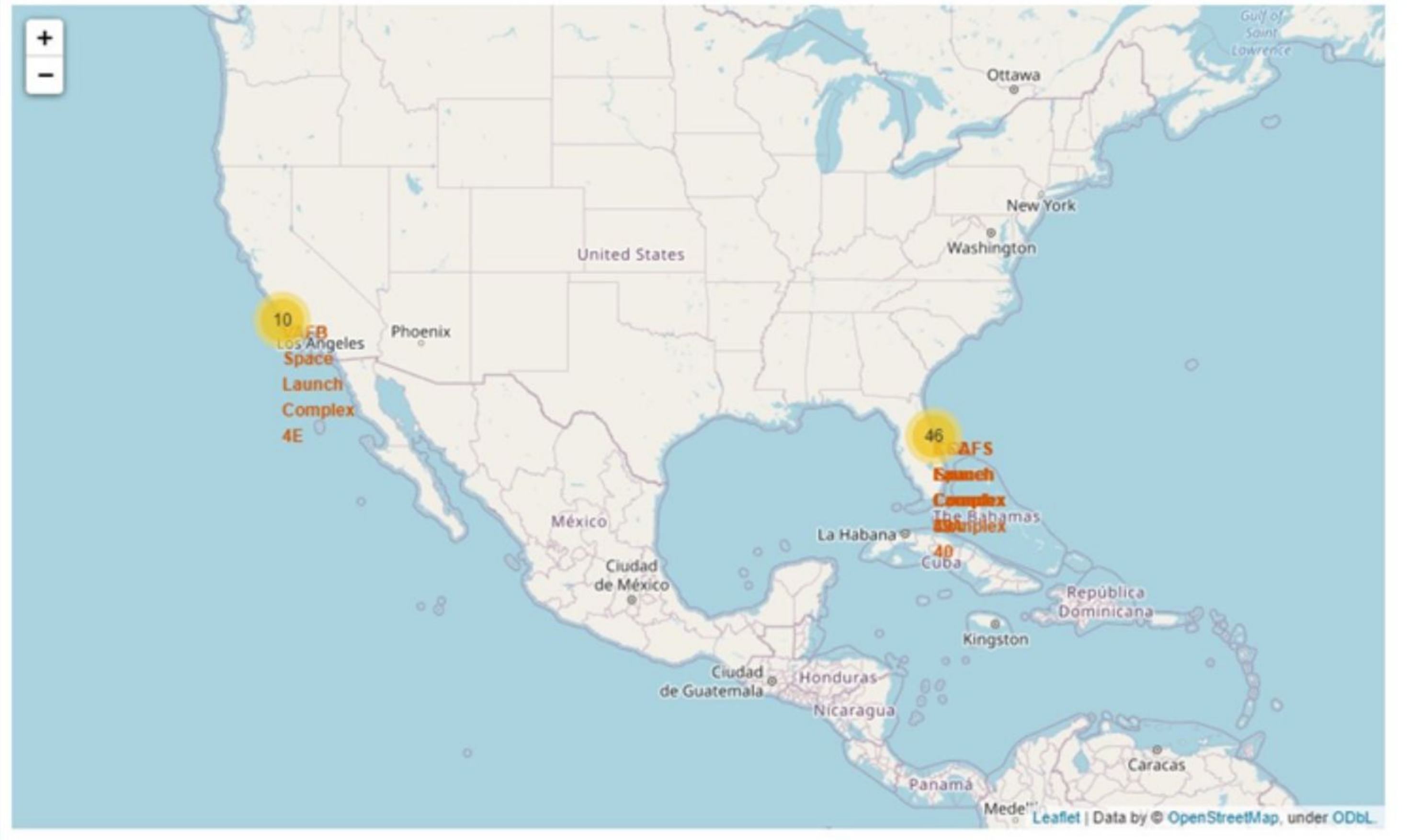
All Launch Sites

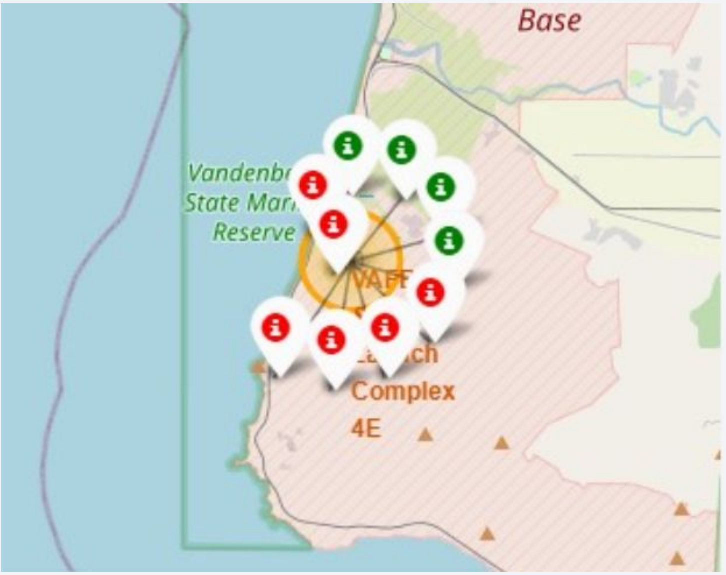
- All launch sites are in the US and in very close proximity to the coast



Success/Failed Launches For Each Site

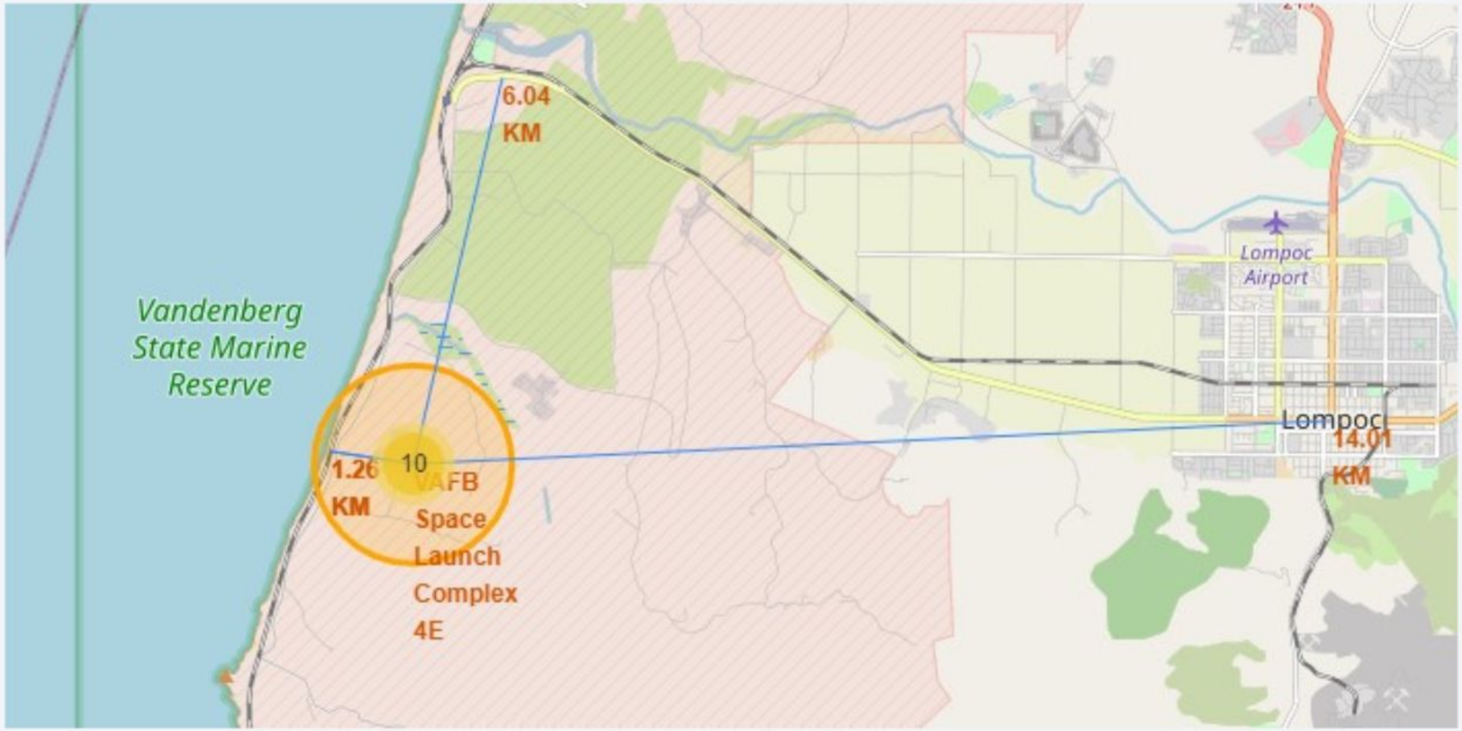
- The first map shows clusters for every launch site, the second map shows a breakdown of the launches for the particular site (red - failure, green - succes)





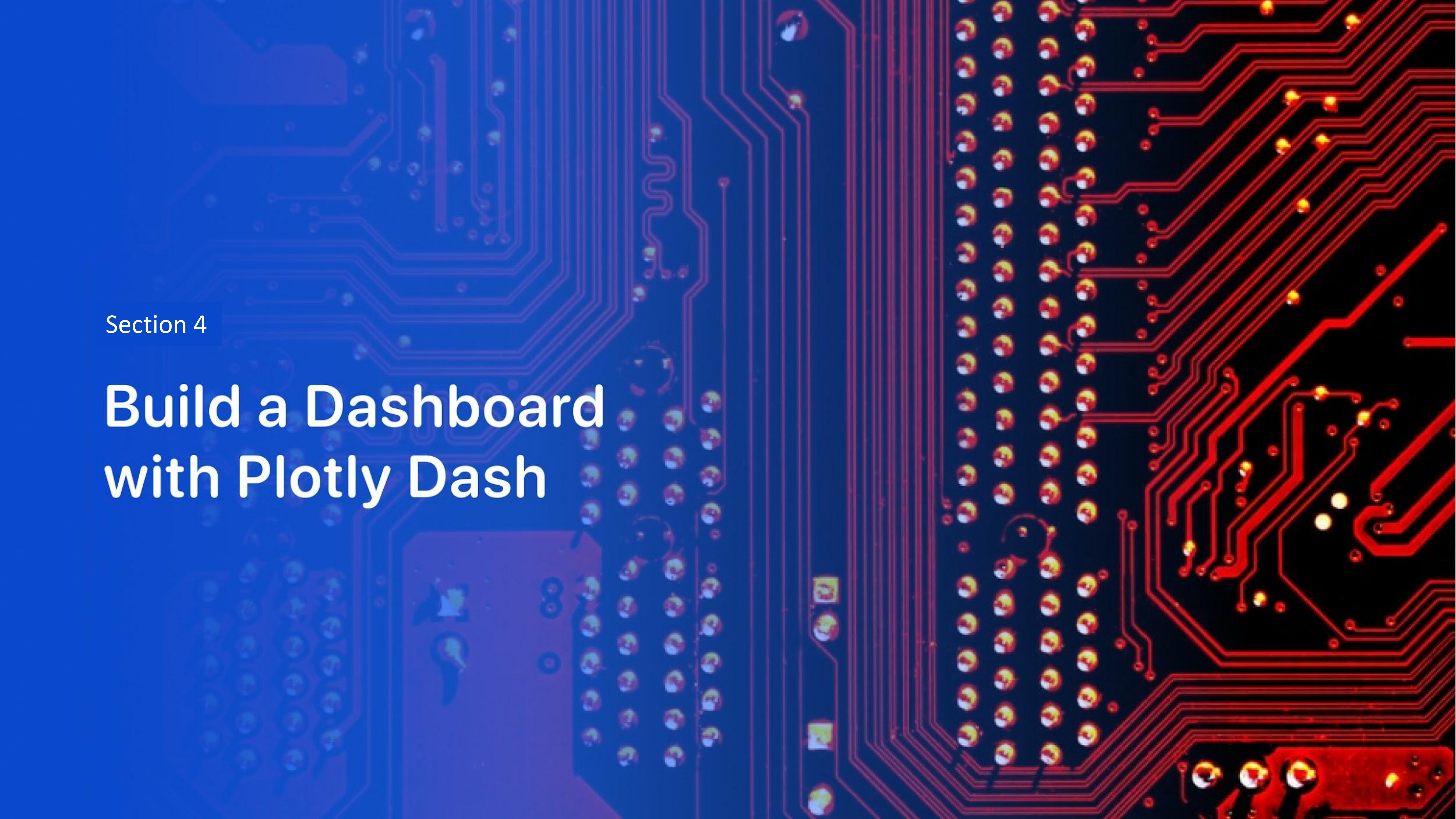
A Launch Site and Its Proximities

- Launch sites are located close to highways, railways likely to ease the logistics of delivering equipment and payload. They are also close to the coast and distant from major cities likely for safety reasons



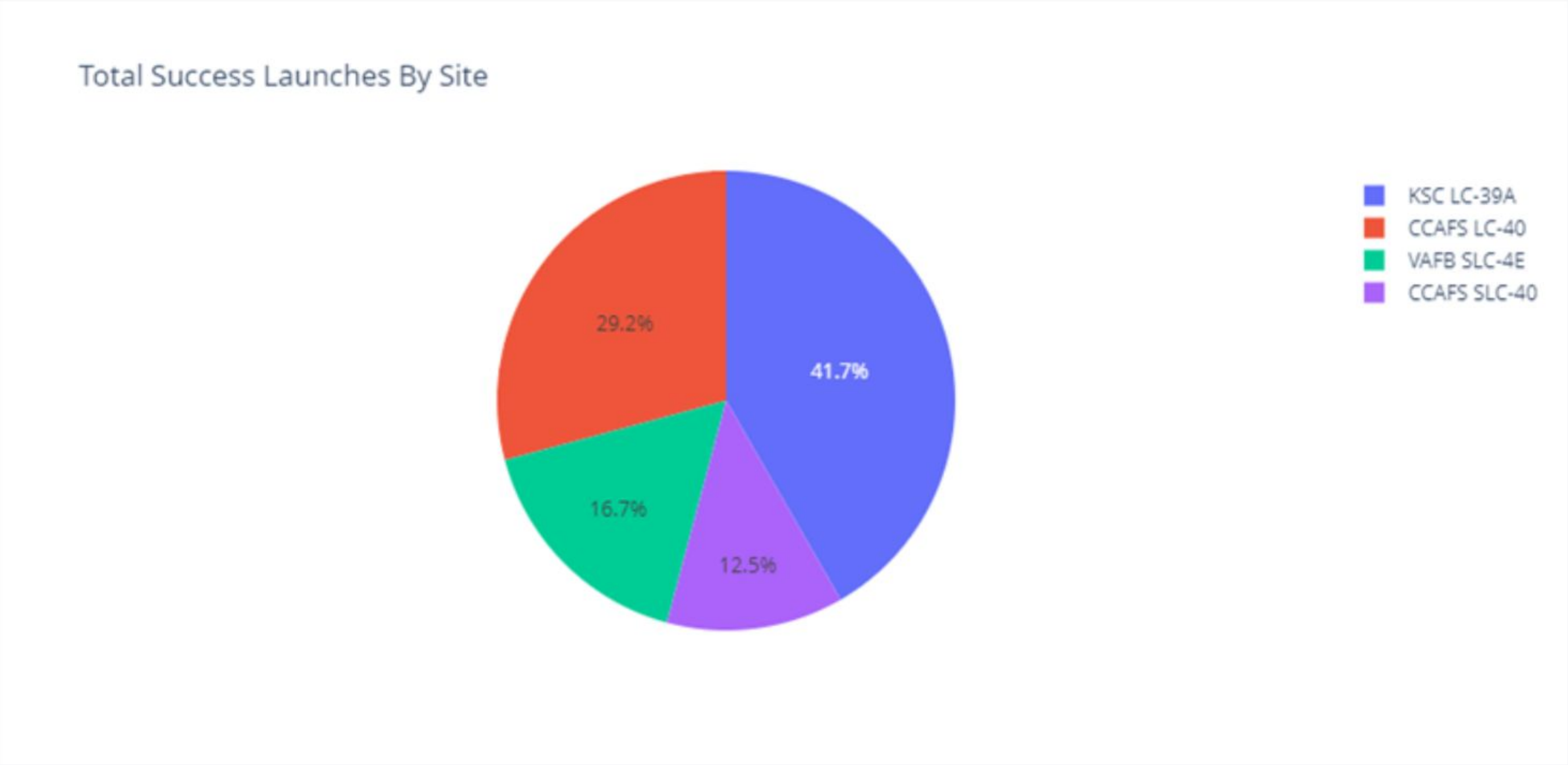
Section 4

Build a Dashboard with Plotly Dash



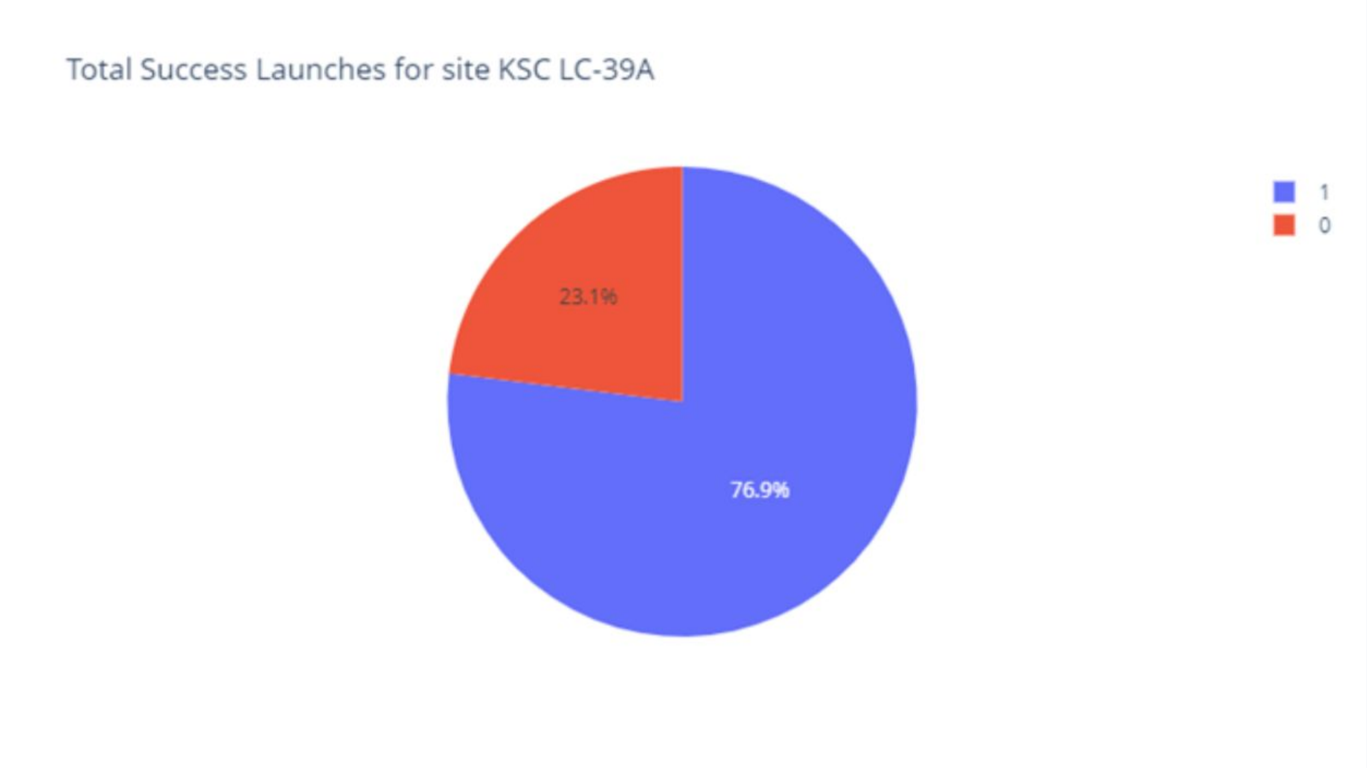
Total Success Launches by Site

- KSC LC-39A is the site with the highest number of successful launches, followed by CCAFS LC-40



KSC LC-39A

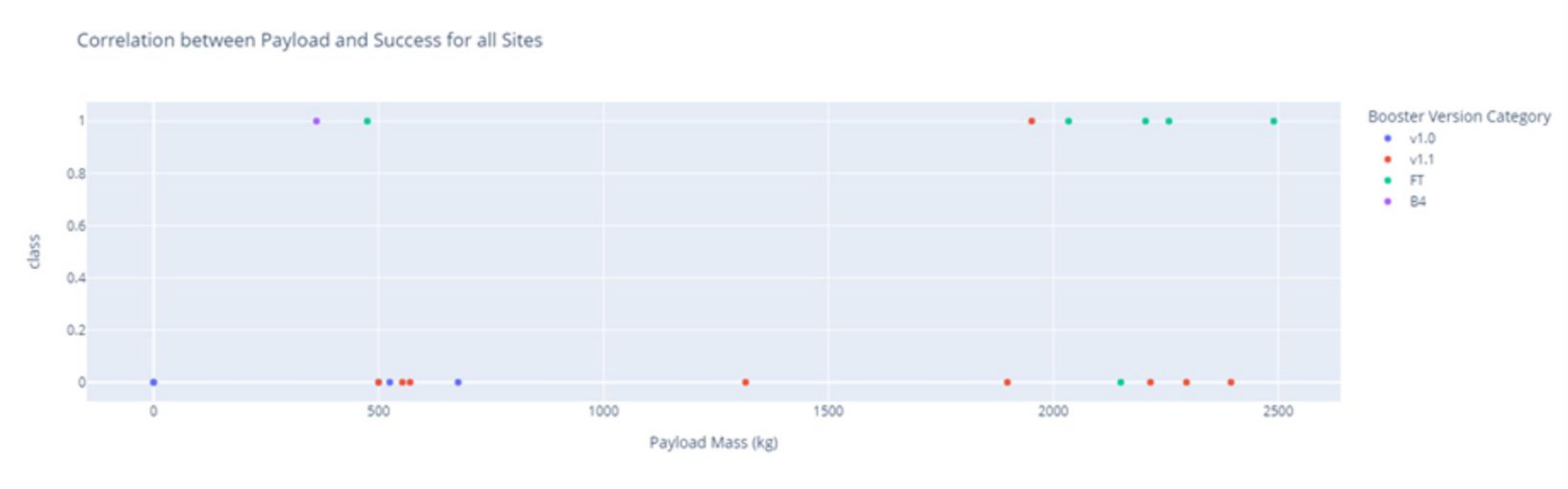
- The breakdown for KSC LC-39A shows that over ¾ of launches from the site were successful



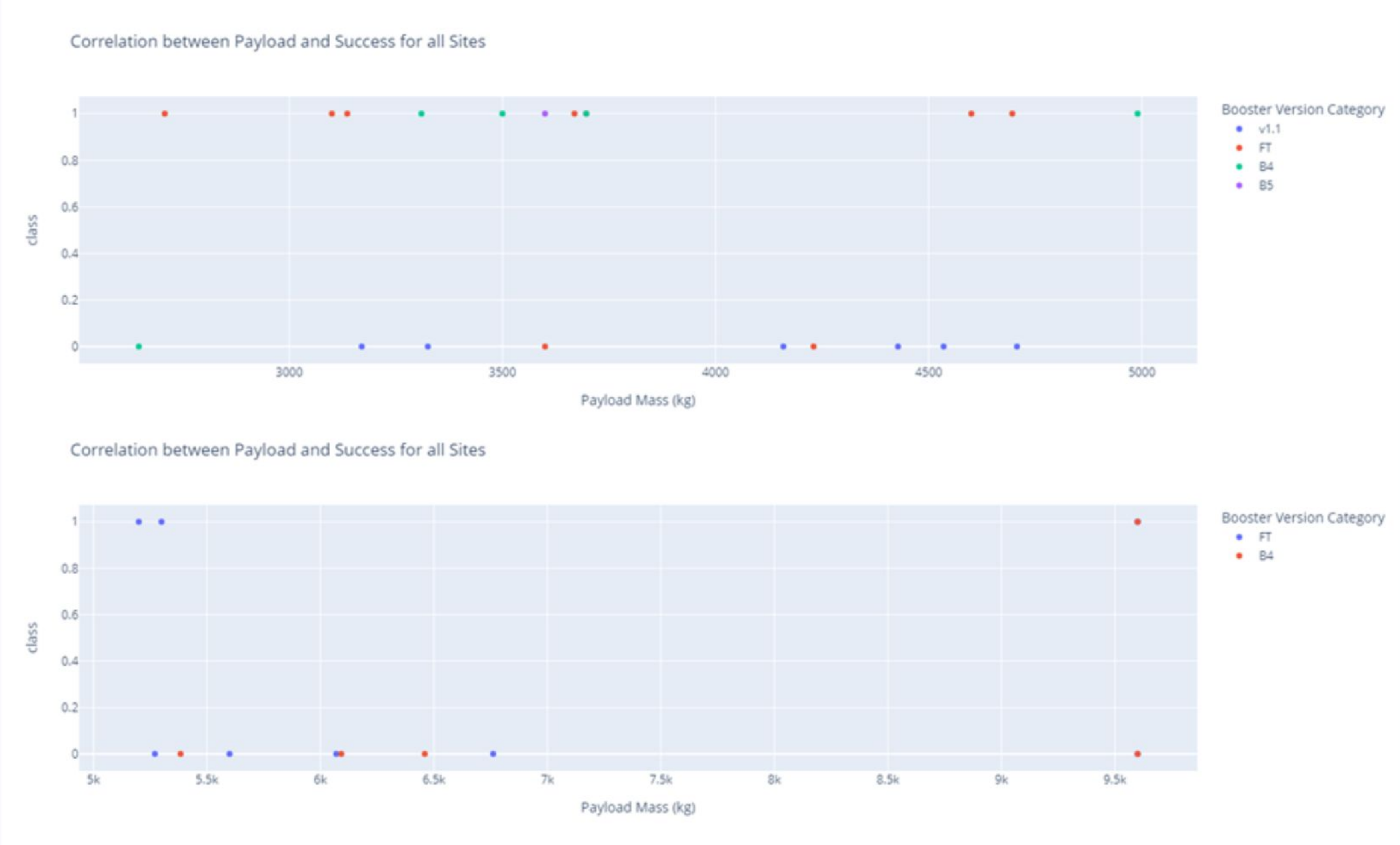
Payload vs. Launch Outcome

Scatter plot for all sites with 2500(kg), 5000(kg), and 10000(kg) payload ranges.

The 2500-5000(kg) range concentrate the majority of successful launches, the 0-2500(kg) range has most failed launches.



Payload vs. Launch Outcome



Correlation between Payload and Success for all Sites

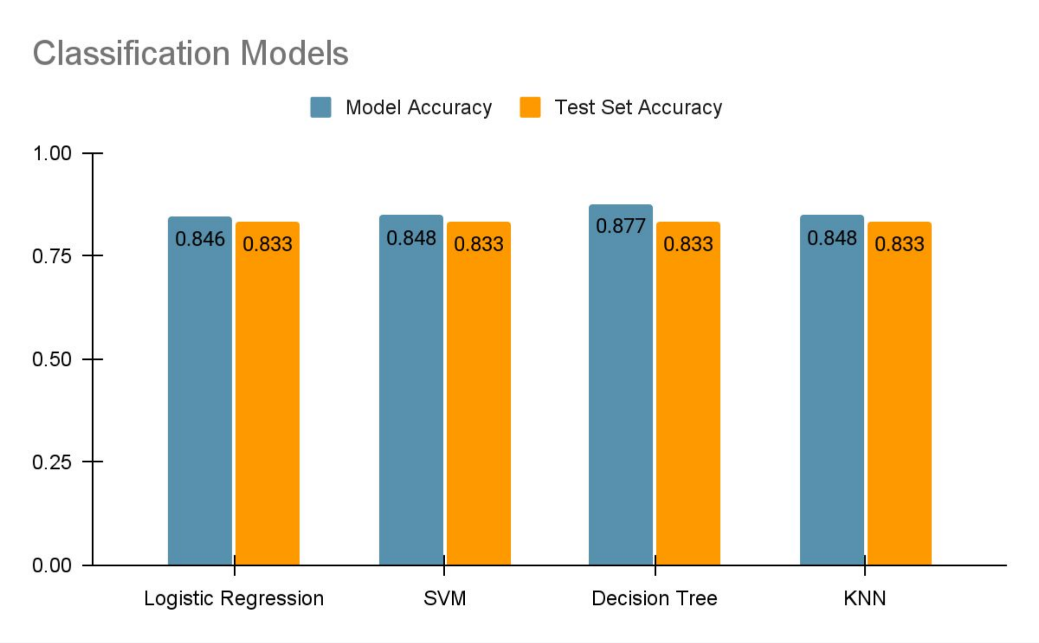
Correlation between Payload and Success for all Sites

Section 5

Predictive Analysis (Classification)

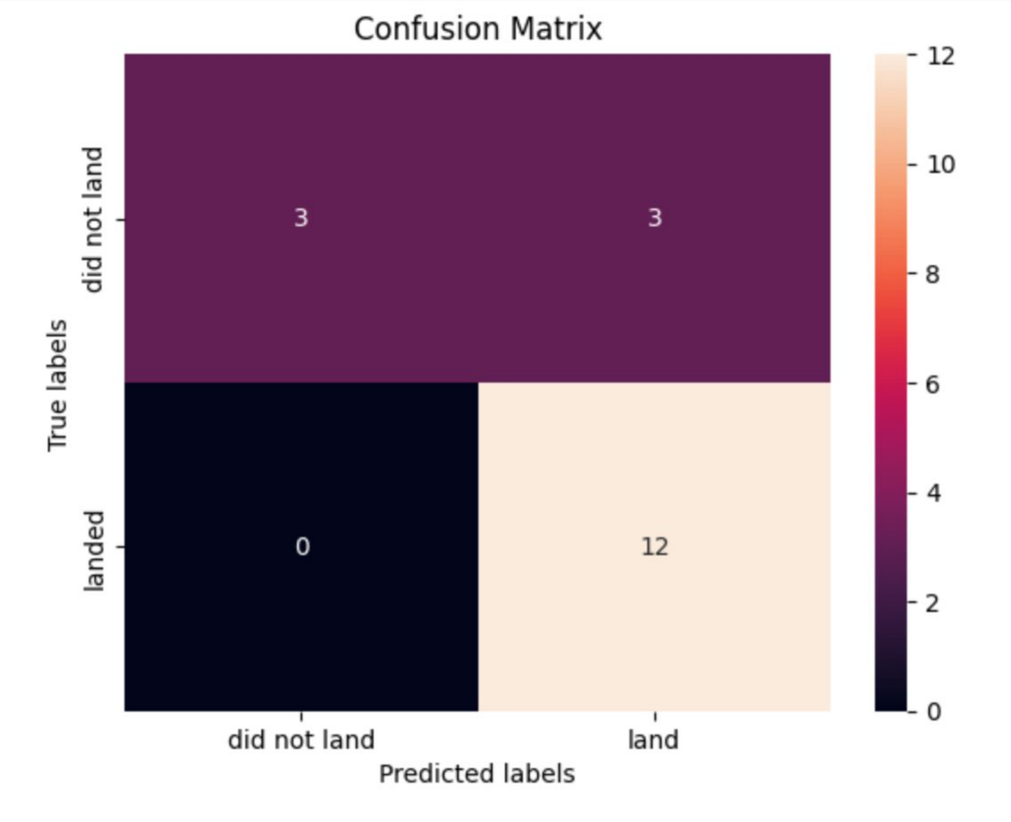
Classification Accuracy

- The Decision Tree model has slightly higher classification accuracy than the rest of the model. However, the advantage is negligible and test set accuracy is similar to all models



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.
- The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier
- It is worth to note that the other 3 classification models performed with the same level of accuracy on the test set.



Conclusions

- With the passage of time (flight number) the success rate continuously improved
- Launch success rate increased between 2013 and 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task. However, the other models gave comparable level accuracy.

Appendix

Capstone GitHub Repository:
https://github.com/taras-trofymchuk/AppliedDataScienceCapstone

Thank you!