Winning Space Race with Data Science

Andrii Zapukhlyi
26.05.2024

Outline

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

Executive Summary

Summary of Methodologies:

1. Data Collection and Wrangling:
   - The data was collected from a variety of sources, including the SpaceX API and web scraping methods.
   - Data cleaning involved handling missing values, correcting inconsistencies, and ensuring the dataset was suitable for analysis.
2. Exploratory Data Analysis (EDA):
   - Conducted EDA using SQL and Python to understand the distribution and relationships between variables.
   - Key insights were visualized using Python graphing libraries like Matplotlib and Seaborn.
3. Interactive Dashboard:
   - Developed an interactive dashboard using Plotly Dash to visualize success ratio for each launch site.
   - Created an interactive map using python library Folium to visualize the launch sites.
4. Predictive Analysis:
   - Implemented four machine learning classification algorithms: Logistic Regression, Decision Trees, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).
   - Each model was trained on a dataset split into training and testing sets to ensure robust evaluation.
   - Hyperparameter tuning was performed to optimize model performance.

Executive Summary

Summary of Results:

1. Data Insights:
   - The success rate of Falcon 9 first stage landings has improved over time, correlating with technological advancements and experience.
   - Certain launch sites and payload masses were more likely to result in successful landings.
2. Model Performance:
   - The SVM model merged as the most accurate model, with a balanced performance across all evaluation metrics.
   - SVM demonstrated superior precision and recall, making it the best choice for predicting the landing success of Falcon 9’s first stage.
3. Dashboard Utility:
   - The interactive dashboard enables stakeholders to visualize launch data in real-time.
   - It serves as a valuable tool for decision-makers in the aerospace industry, providing data-driven insights for strategic planning and competitive analysis.

Introduction

- **Project Background and Context**

  SpaceX has revolutionized the space industry by significantly lowering launch costs through the reuse of the Falcon 9 first stage. While SpaceX offers launches for about $62 million, other providers charge around $165 million, mainly due to the inability to reuse rocket stages. This project aims to leverage data science to predict the landing success of Falcon 9's first stage, providing insights for companies looking to compete with SpaceX.

- **Problems to Address**

  1. Can we predict the success of Falcon 9's first stage landing?

  2. What factors most influence landing success?

  3. How can these predictions inform competitive strategies?

Section 1

Methodology

Methodology

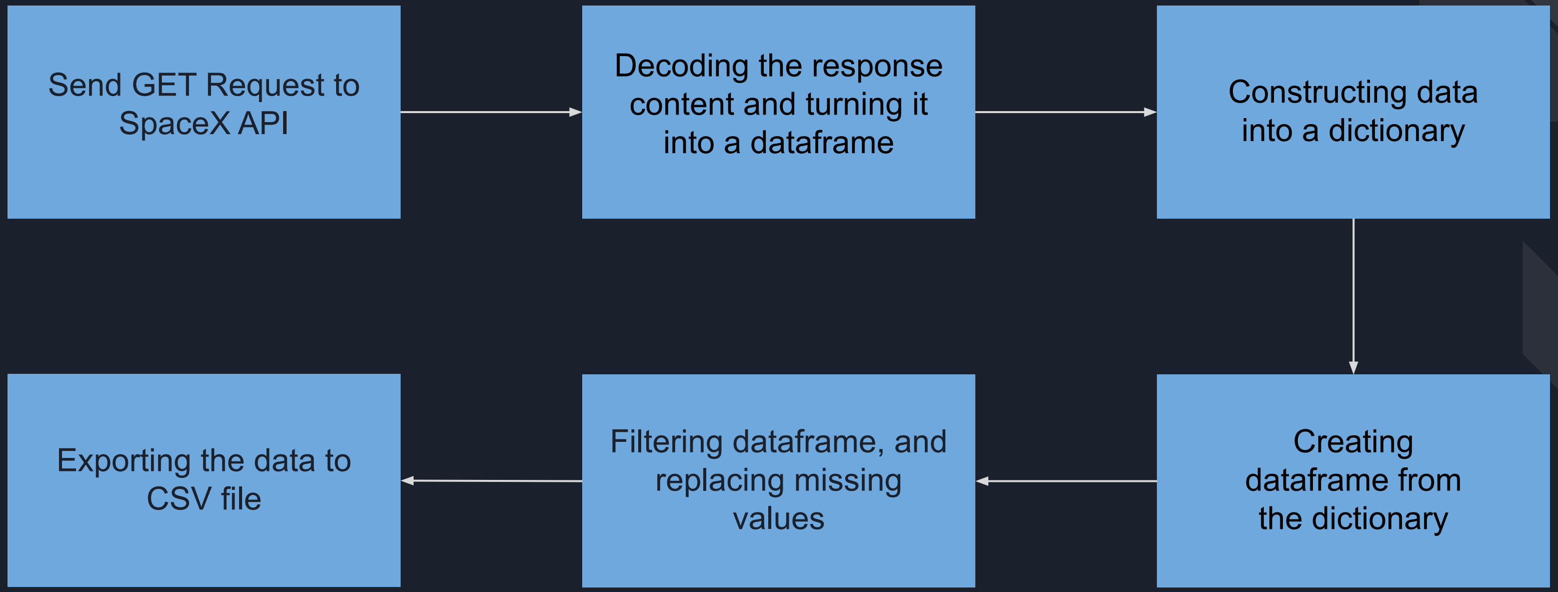
Executive Summary

- Data collection methodology:
  - The data was collected from SpaceX API and scraped from www.wikipedia.com
- Perform data wrangling
  - Replaced a column with a string type with a Boolean type. We are not interested in the reasons for the failure, so the result is the same for each failure - 0.
  - Replaced missing values for ‘Payload Mass’ column with an average payload
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Implemented four machine learning classification algorithms: Logistic Regression, Decision Trees, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)

Data Collection

- The data for this project was collected from two primary sources: the SpaceX API and web scraping from Wikipedia.
- The process involved accessing detailed information about past and upcoming rocket launches from the SpaceX API, extracting data such as flight number, booster version, launch site, payload mass, orbit, and mission outcome, and storing it in a structured format for further processing.
- Additionally, HTML content was retrieved from Wikipedia pages related to SpaceX launches, and data points like launch dates, payload names, launch sites, and launch outcomes were extracted and cleaned to handle inconsistencies and ensure uniform formatting.
- Finally, the data from both sources were merged for further analysis.

Data Collection – SpaceX API



Send GET Request to SpaceX API

Decoding the response content and turning it into a dataframe

Constructing data into a dictionary

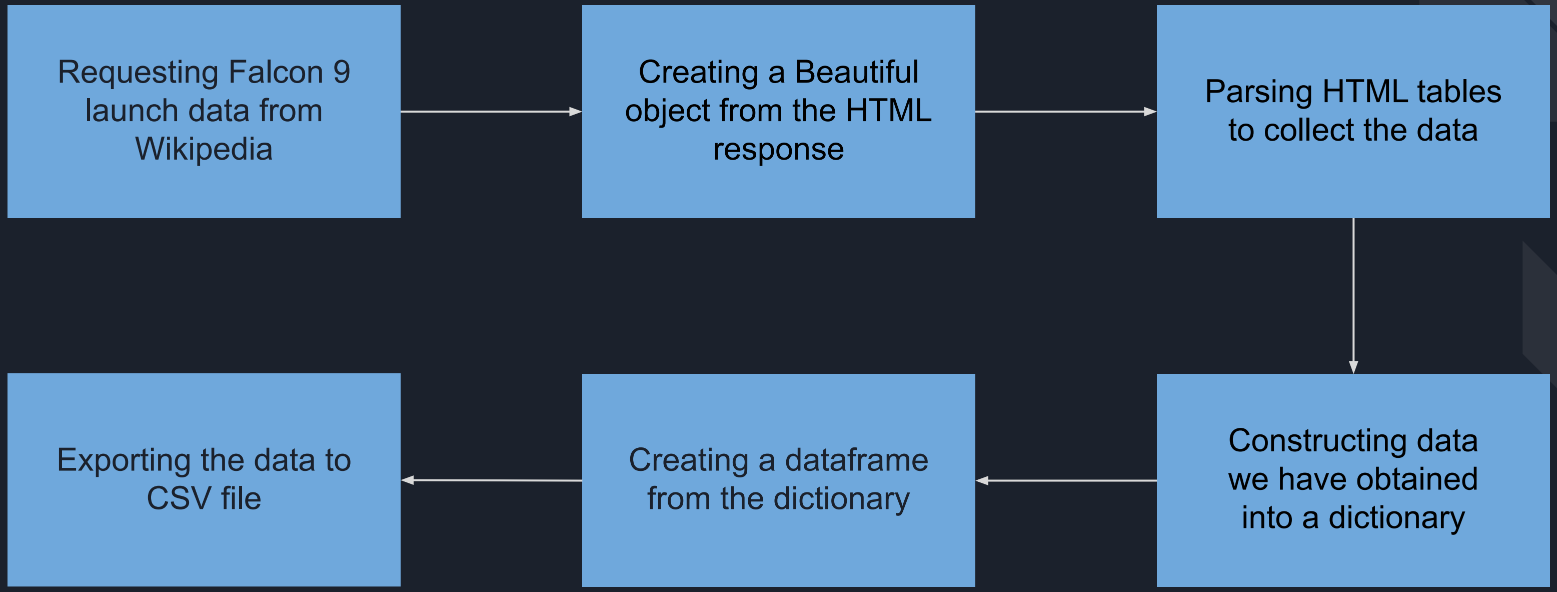
Creating dataframe from the dictionary

Filtering dataframe, and replacing missing values

Exporting the data to CSV file

GitHub URL on file: [Link]()

Data Collection – Scraping



Requesting Falcon 9 launch data from Wikipedia

Creating a Beautiful object from the HTML response

Parsing HTML tables to collect the data

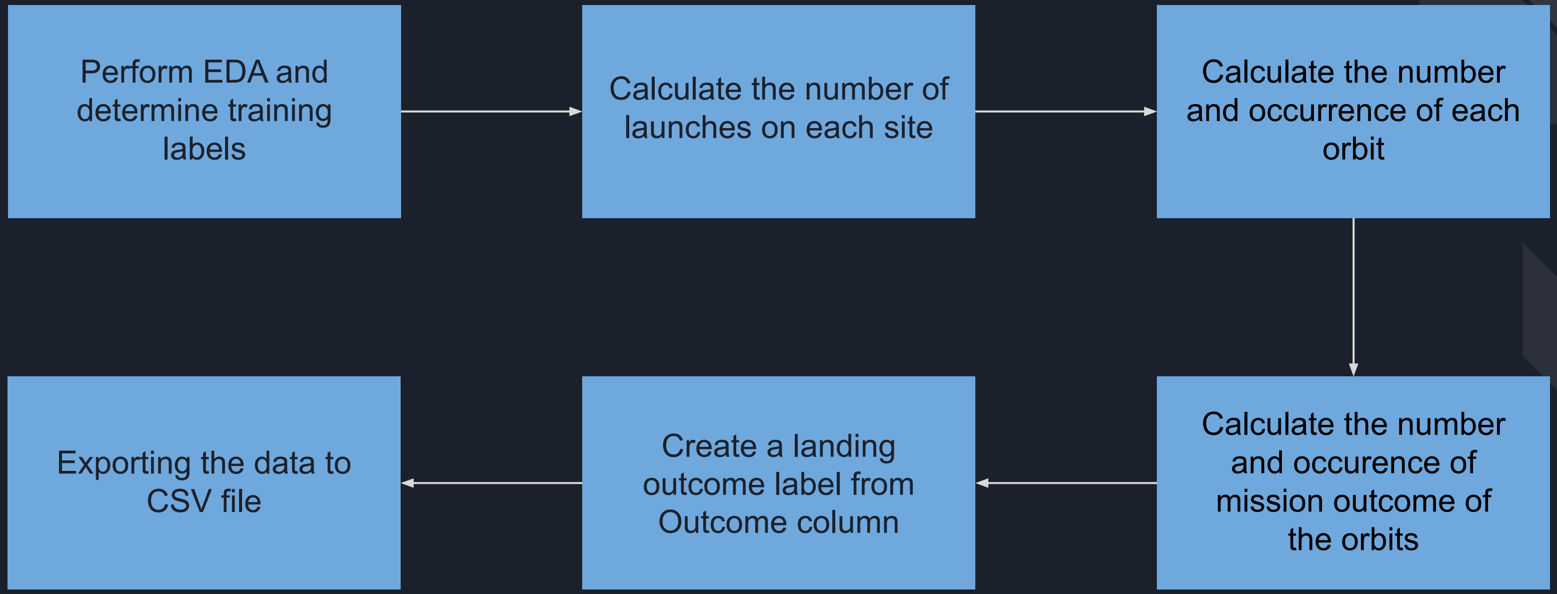
Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Exporting the data to CSV file

GitHub URL on file: [Link]()

Data Wrangling



Perform EDA and determine training labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome of the orbits

Create a landing outcome label from Outcome column

Exporting the data to CSV file

GitHub URL on file: Link

EDA with Data Visualization

The charts were selected to visually explore different aspects of the SpaceX launch data and derive insights. To summarise:

1. Flight Number vs. Payload Mass: To observe trends over time, distinguishing between successful and unsuccessful flights.
2. Flight Number vs. Launch Site: To compare launch distributions across sites, highlighting success outcomes.
3. Payload vs. Launch Site: To analyze payload distribution and its impact on launch outcomes across sites.
4. Success Rate of Each Orbit: To compare success rates across different orbit types.
5. Flight Number vs. Orbit: To examine orbit success rates over time.
6. Payload Mass vs. Orbit: To explore the relationship between payload mass and orbit success.
7. Mean Success Rate by Year: To visualize trends in launch success rates over time.

GitHub URL on file: Link

EDA with SQL

- There are 4 distinct landing sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
- The total payload mass carried by boosters launched by NASA (CRS): 45596 kg
- An average payload mass carried by booster version F9 v1.1: 2534.6667 kg
- The date when the first successful landing outcome in ground pad was achieved: 2015-12-22
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- There are 100 successful mission outcomes and 1 failure
- There are 12 booster versions which have carried the maximum payload mass 15600 kg.
- There are 2 records which have failure (drone ship) landing outcome in 2015
- In the period from 2010-06-04 to 2017-03-20, there were 9 successful landings on the ground and 5 unsuccessful landings on the drone ship

GitHub URL on file: Link

Build an Interactive Map with Folium

In the data visualization process using Folium, various map objects such as markers, circles, and lines were created and added to the map to provide a clear and informative visualization of SpaceX launch data. Here is a summary of each object:

- Markers for launch sites:

  - Added markers for each SpaceX launch site to clearly indicate the geographical locations of all launch sites on the map, allowing for easy identification and comparison.

- Marked the success/failed launches for each site:

  - Added markers for each launch at the respective launch sites, color-coded by success (e.g., green for success, red for failure) to provide a visual representation of the outcome of launches at each site, making it easy to see patterns of success and failure geographically.

- Calculated the distances between a launch site to its proximities:

  - Added circles around each launch site to indicate proximity ranges to visually represent areas around each launch site, helping to understand the geographical spread and potential impact areas.

  - Added lines between the launch sites and the nearest highway, railway, city have been added to indicate distances to provide a visual assessment of the distances from the launch sites to significant objects, which helps in spatial analysis.

GitHub URL on file: Link

Build a Dashboard with Plotly Dash

- Added dropdown menu for "All sites" and individual launch sites to provide flexibility in the data being viewed, allowing users to focus on overall trends or drill down into specific launch sites for detailed analysis.

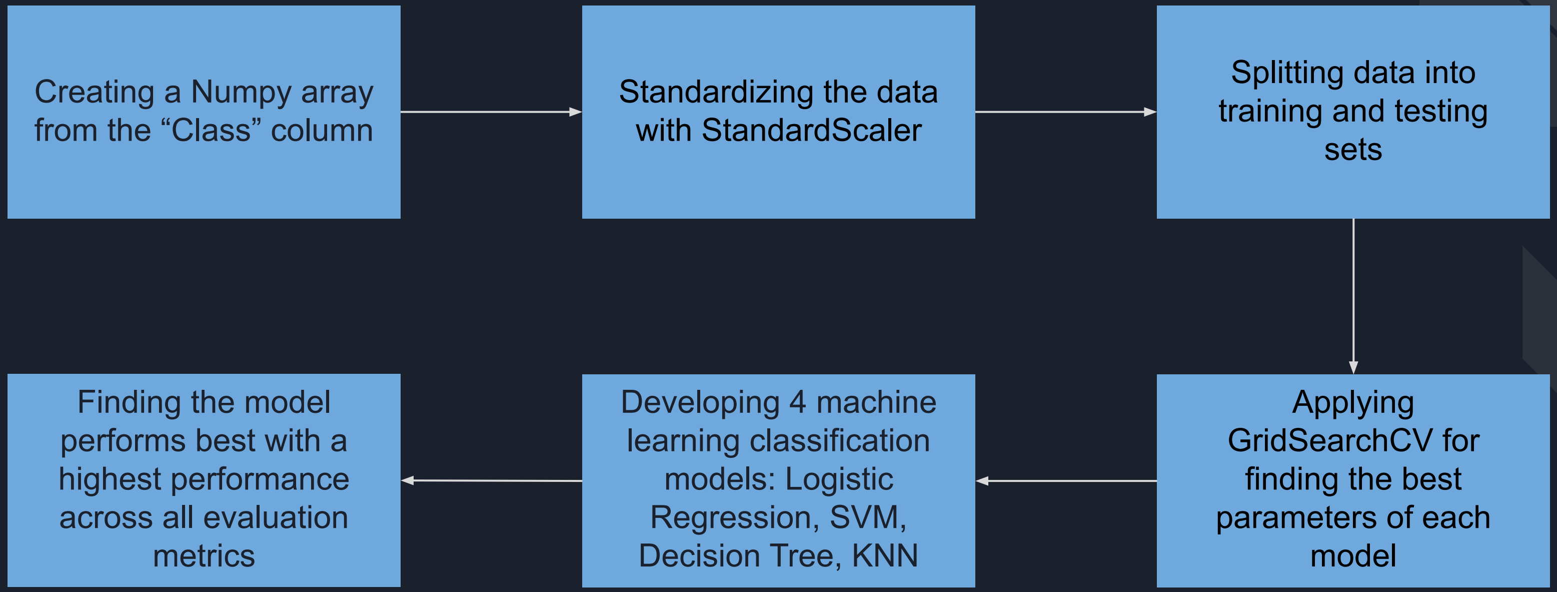
- Added a pie chart for the selected launch site (there is an option for all sites) to give a clear visual representation of the success ratio, helping users quickly grasp the performance of SpaceX launches either in aggregate or at specific sites.

- Payload Mass Range Slider: Added a range slider for payload mass from 0 to 10,000 kg to enable detailed analysis of how payload mass affects launch success, allowing users to investigate specific payload ranges and their corresponding outcomes.

- Added a scatter plot for Payload Mass vs Success that updates dynamically based on the selected launch site and payload mass range, with different colors for different booster versions. This helps in visually exploring the relationship between payload mass and launch success, and understanding how different factors influence the success of launches.

GitHub URL on file: [Link]()

Predictive Analysis (Classification)



Creating a Numpy array from the “Class” column

Standardizing the data with StandardScaler

Splitting data into training and testing sets

Applying GridSearchCV for finding the best parameters of each model

Developing 4 machine learning classification models: Logistic Regression, SVM, Decision Tree, KNN

Finding the model performs best with a highest performance across all evaluation metrics

GitHub URL on file: Link

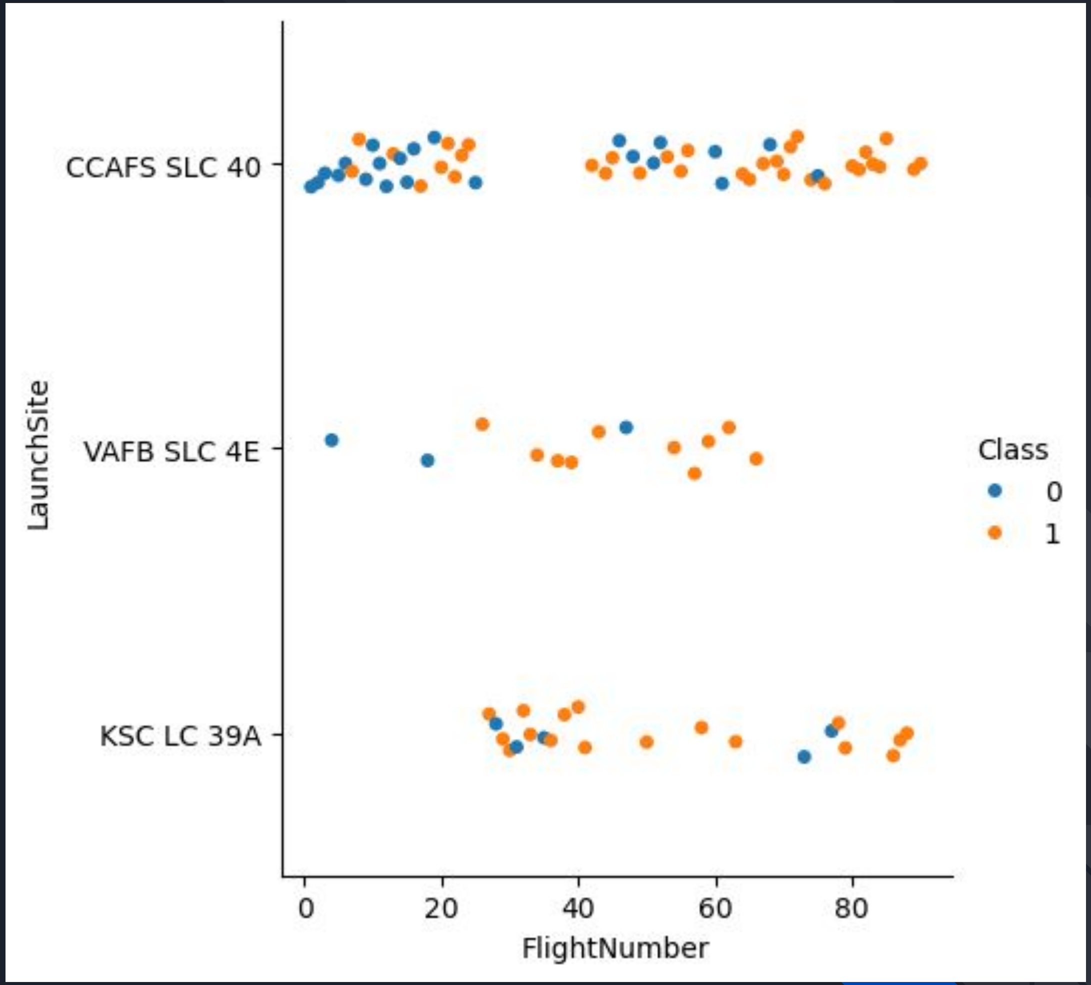
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Explanation:

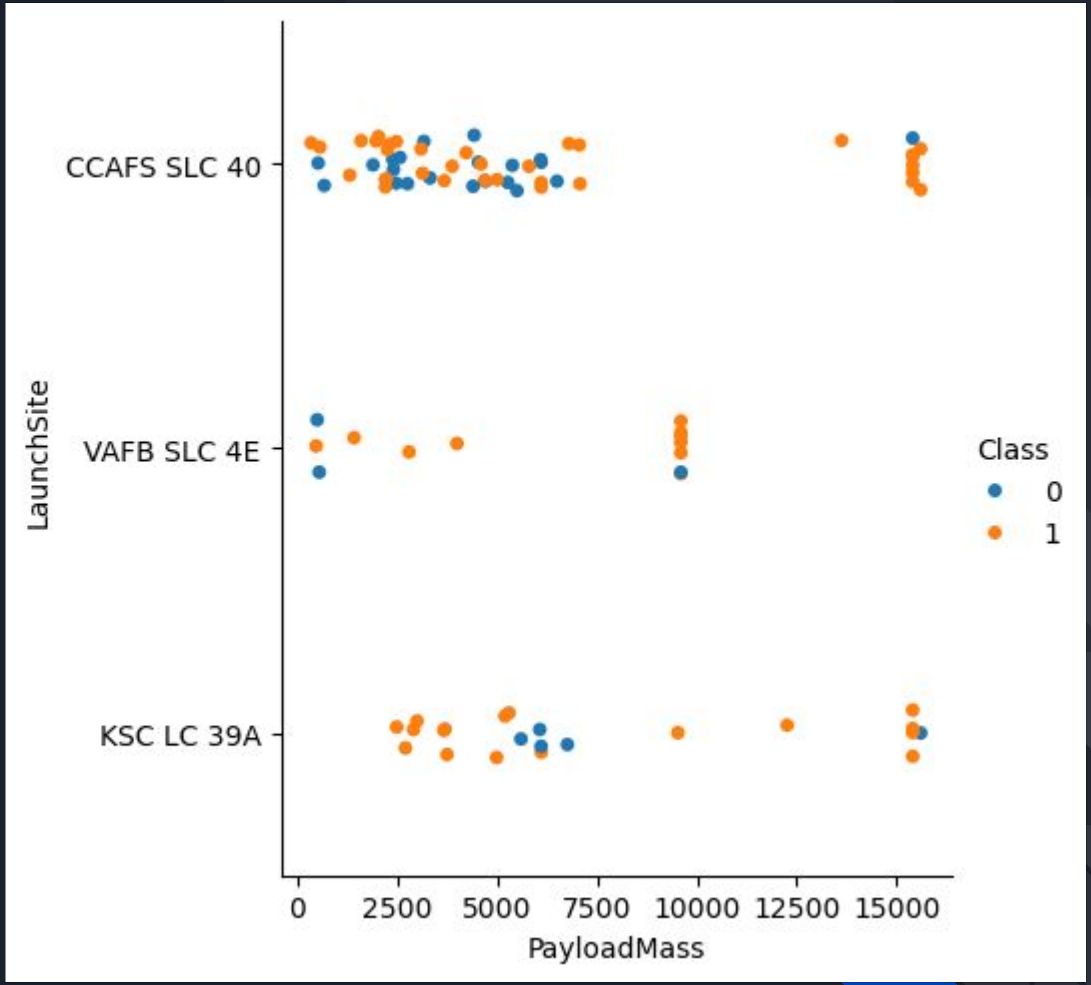
- The first launches were unsuccessful, while the latest ones are more successful
- The CCAFS SLC 40 launch site has the most launches
- However, VAFB SLC 4E and KSC LC39A have higher success rates



Payload vs. Launch Site

Explanation:

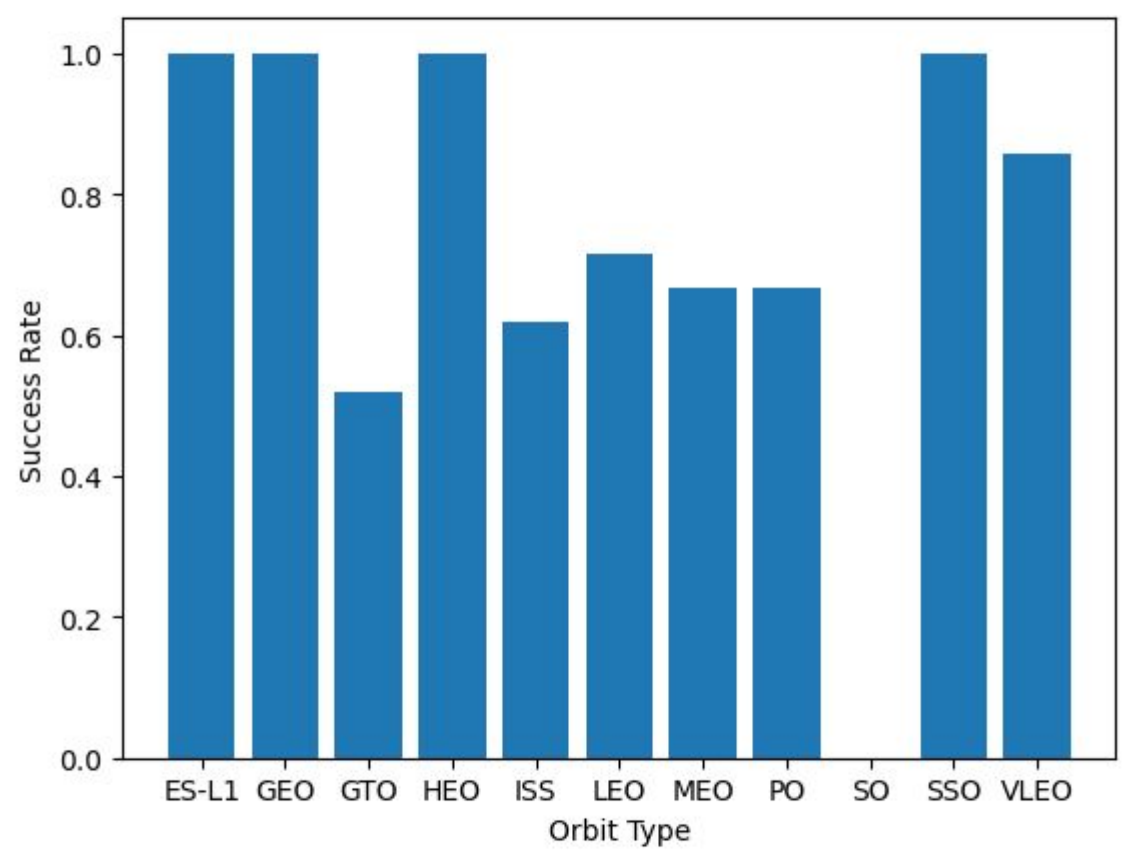
- For each launch site the higher payload mass gives higher success rate
- Most launches with payload mass more than 7500 kg were successful
- KSC LC 39A has also high success rate for payload mass less than 5000 kg



Success Rate vs. Orbit Type

Explanation:

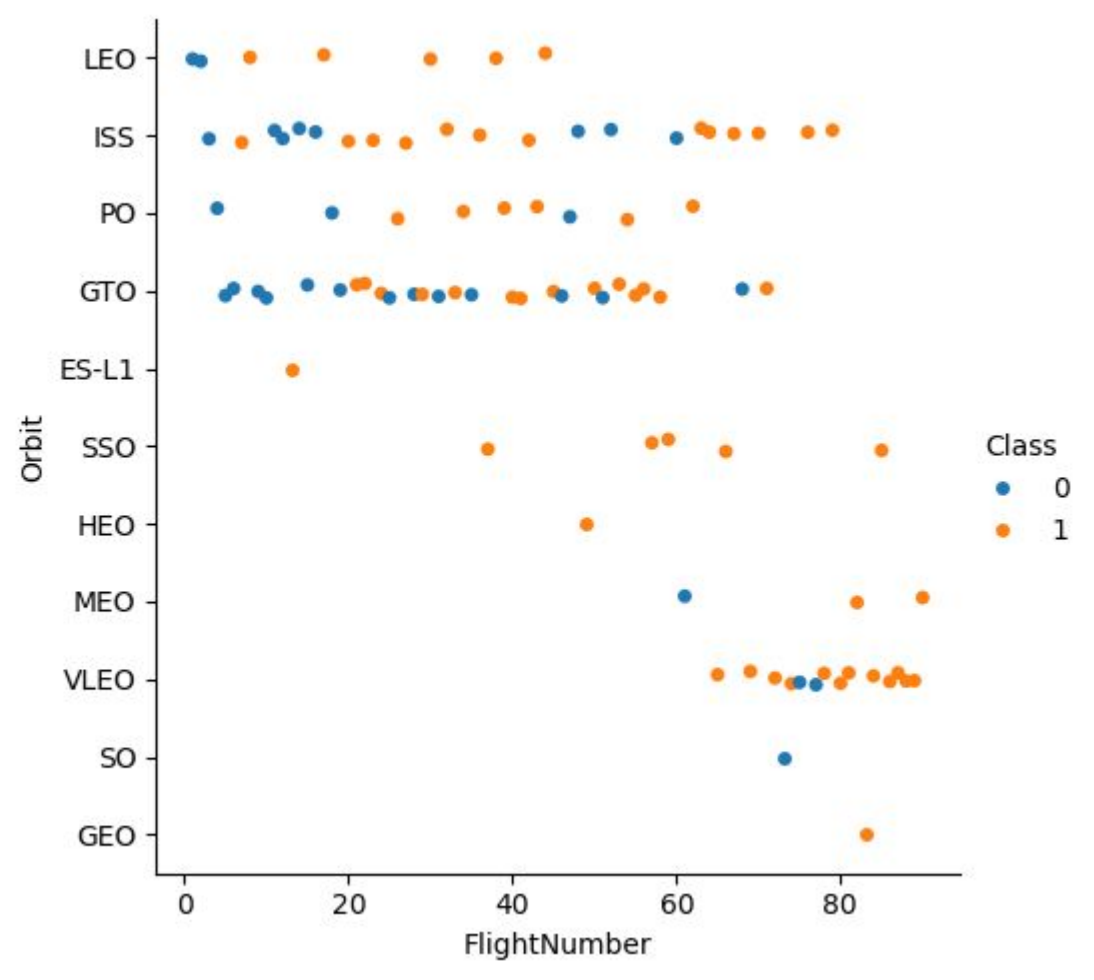
- There are 4 orbits with maximum success rate: ES-L1, GEO, HEO, SSO
- SO orbit has 0% success rate
- GTO, ISS, LEO, MEO, PO and VLEO orbits have success rate between 50% and 85%



Flight Number vs. Orbit Type

Explanation:

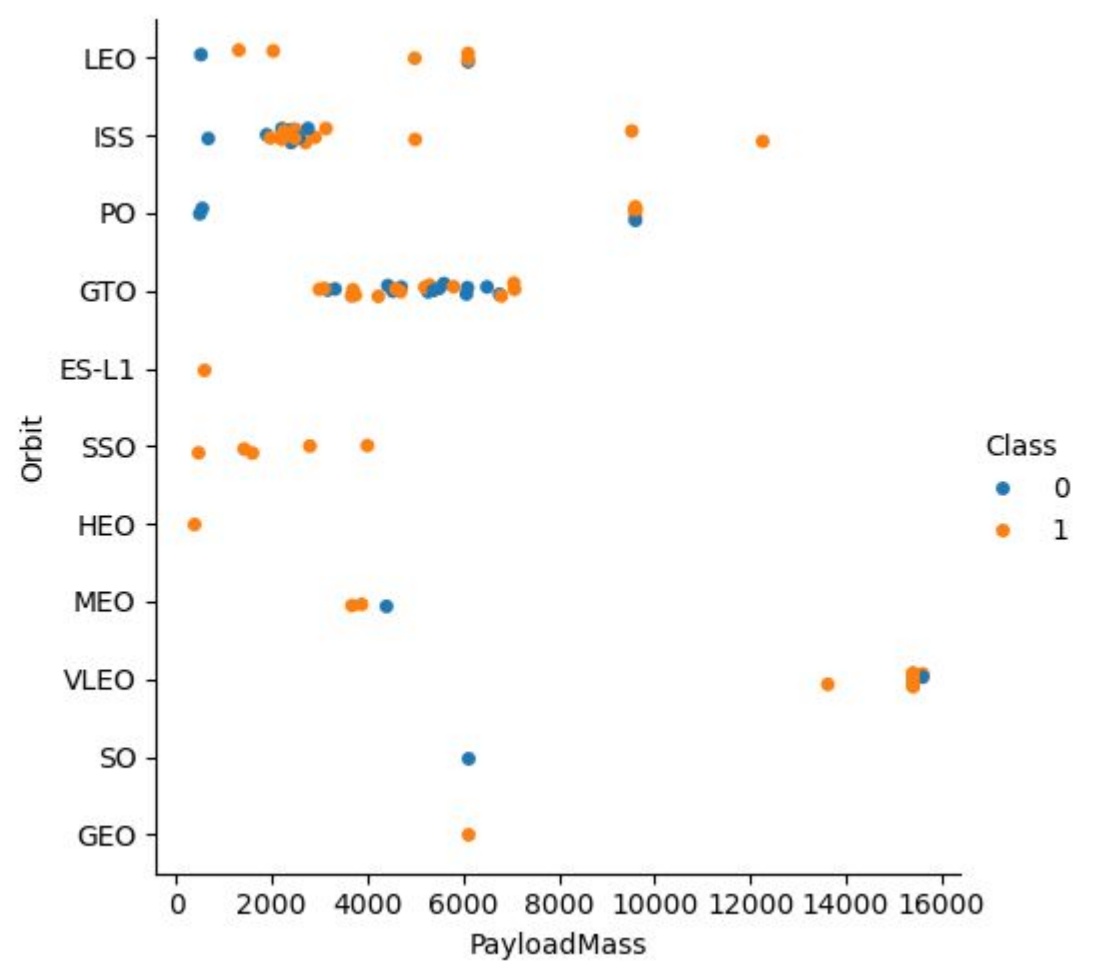
- For LEO orbit the success appears related to the number of flights
- However, for GTO orbit there isn't no relationship between flight number and success
- We can again see 100% success rate for 4 orbits



Payload vs. Orbit Type

Explanation:

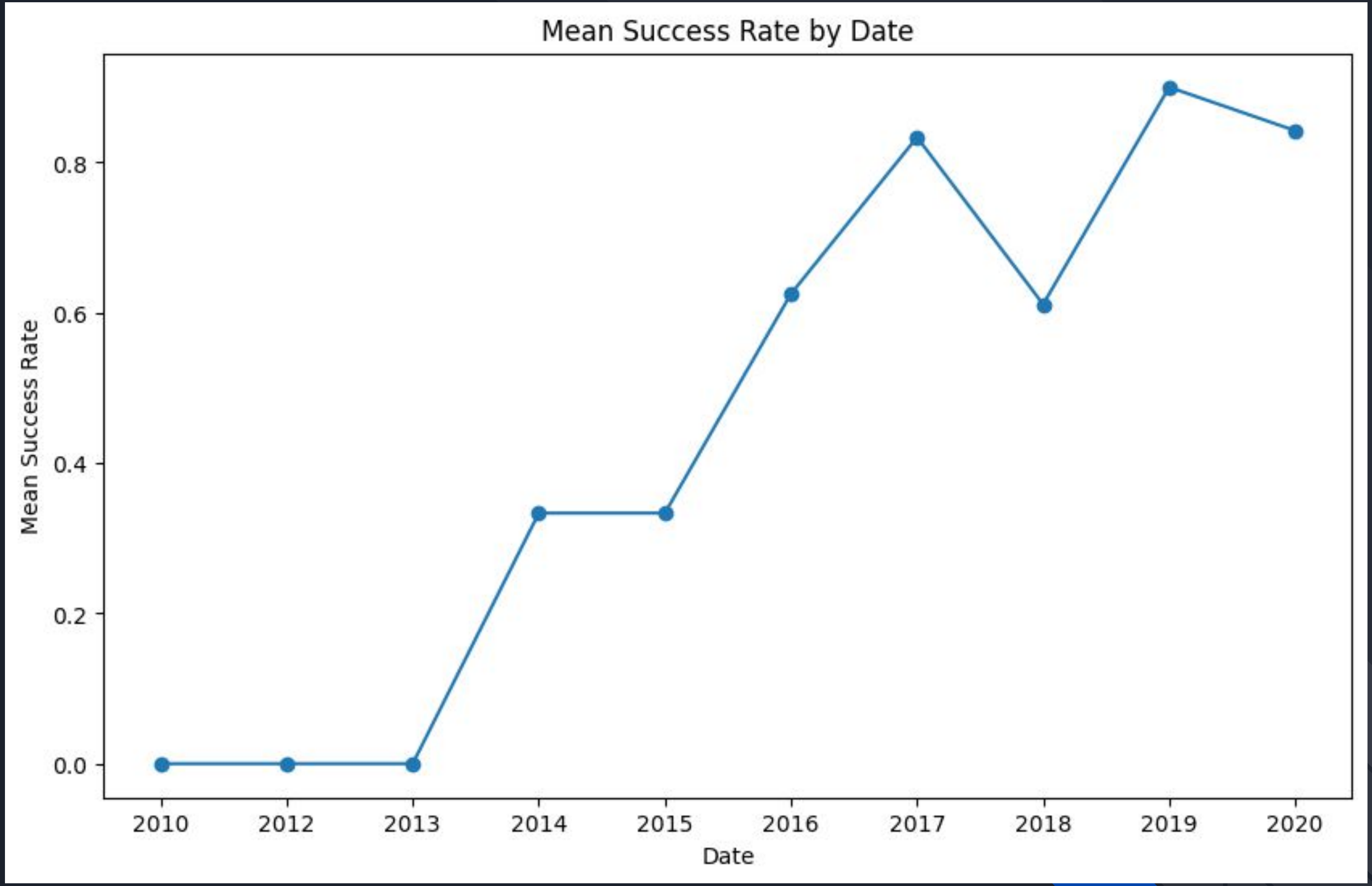
- Heavy payloads have a negative influence on GTO orbit and positive on ISS



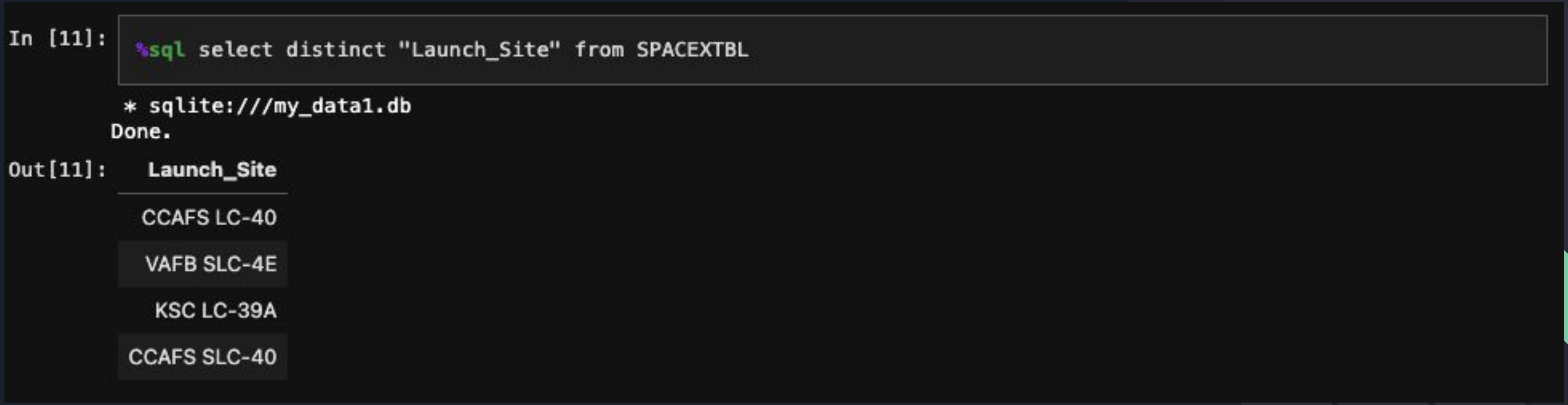
Launch Success Yearly Trend

Explanation:

- The success rate 2013 kept increasing till 2020
- There is decline in 2018
- From 2010 to 2013 there is 0% success rate



All Launch Site Names

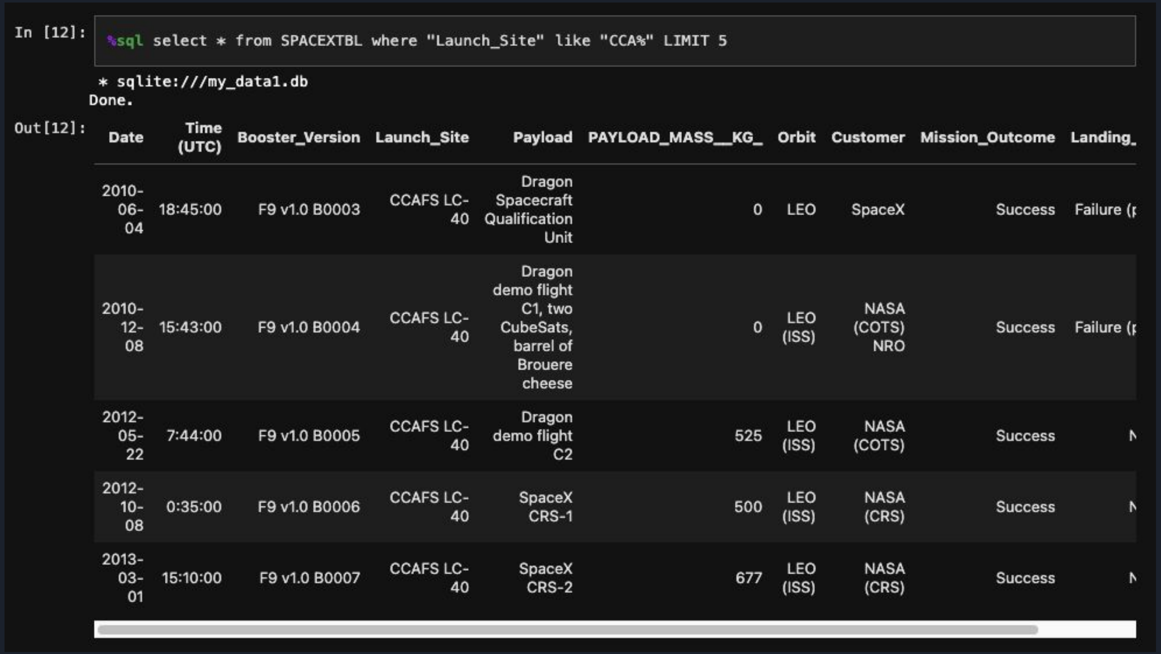


```
In [11]:  %sql select distinct "Launch_Site" from SPACEXTBL

           * sqlite:///my_data1.db
          Done.

Out[11]:      Launch_Site
             CCAFS LC-40
             VAFB SLC-4E
             KSC LC-39A
            CCAFS SLC-40
```

Explanation:

- Displaying the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'



```
In [12]: %sql select * from SPACEXTBL where "Launch_Site" like "CCA%" LIMIT 5

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

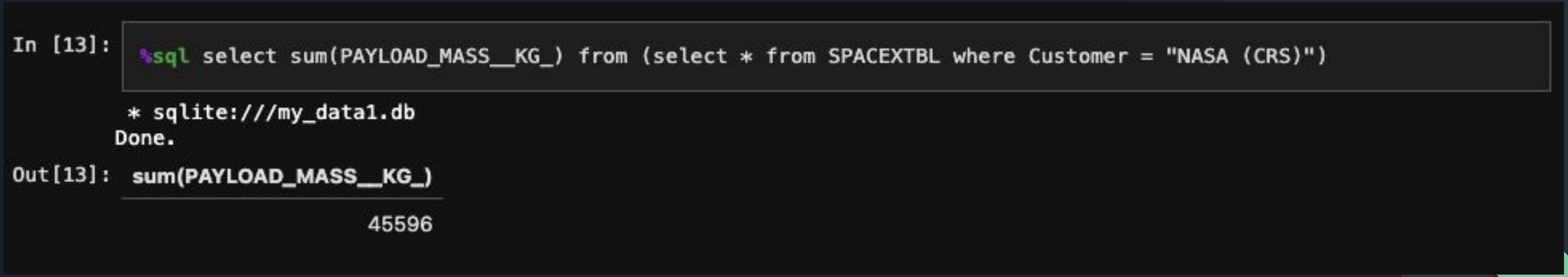
Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Explanation:

- Displaying first 5 records where launch sites begin with the string 'CCA'

Total Payload Mass



```
In [13]: %sql select sum(PAYLOAD_MASS__KG_) from (select * from SPACEXTBL where Customer = "NASA (CRS)")

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

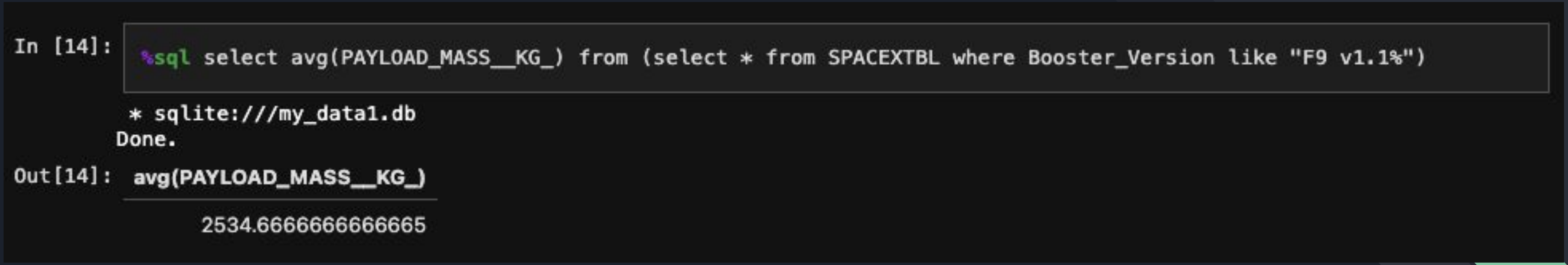
Out[13]:

sum(PAYLOAD_MASS__KG_)
45596

Explanation:

- Displaying the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

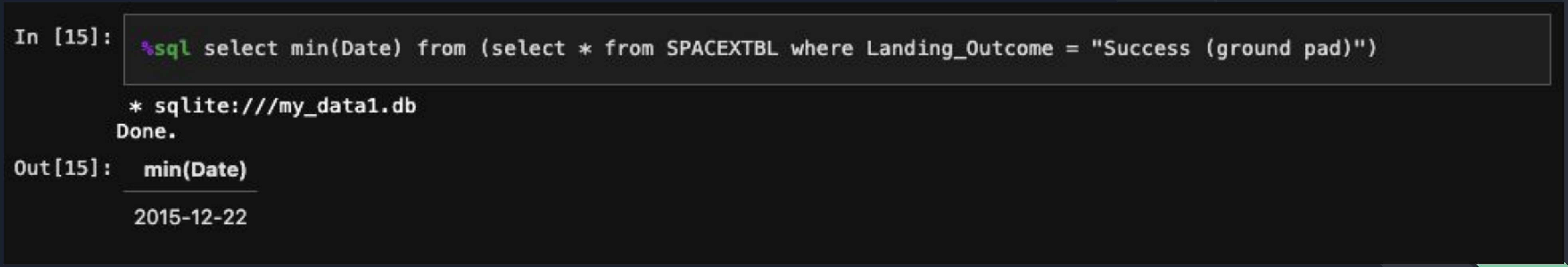


```
In [14]:  %sql select avg(PAYLOAD_MASS__KG_) from (select * from SPACEXTBL where Booster_Version like "F9 v1.1%")

           * sqlite:///my_data1.db
          Done.

Out[14]:  avg(PAYLOAD_MASS__KG_)
                2534.6666666666665
```

Explanation:

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

First Successful Ground Landing Date

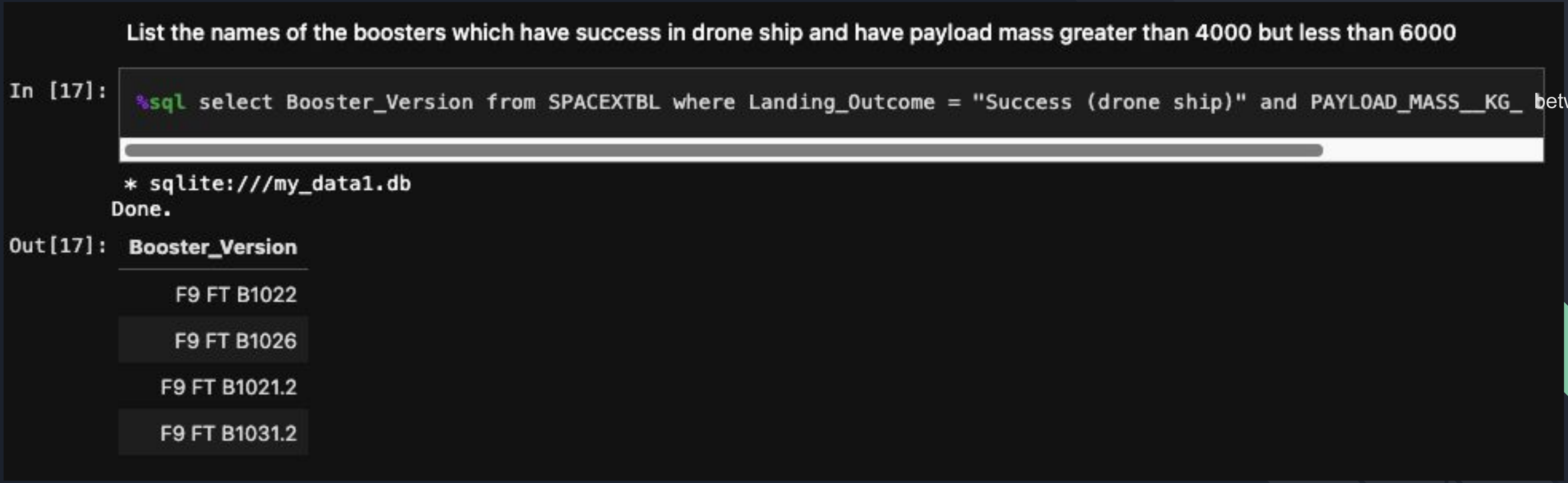


```
In [15]:  %sql select min(Date) from (select * from SPACEXTBL where Landing_Outcome = "Success (ground pad)")

           * sqlite:///my_data1.db
          Done.

Out[15]:  min(Date)

          2015-12-22
```

Explanation:

- The date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000



List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [17]: %sql select Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

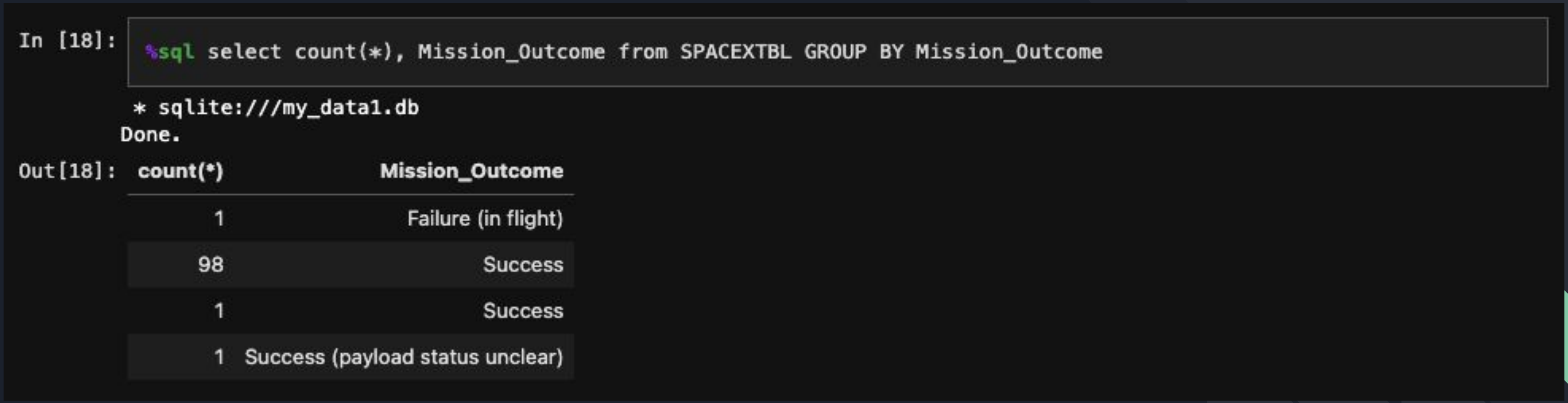
Out[17]:

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

Explanation:

- Displaying the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes



```
In [18]:  %sql select count(*), Mission_Outcome from SPACEXTBL GROUP BY Mission_Outcome

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

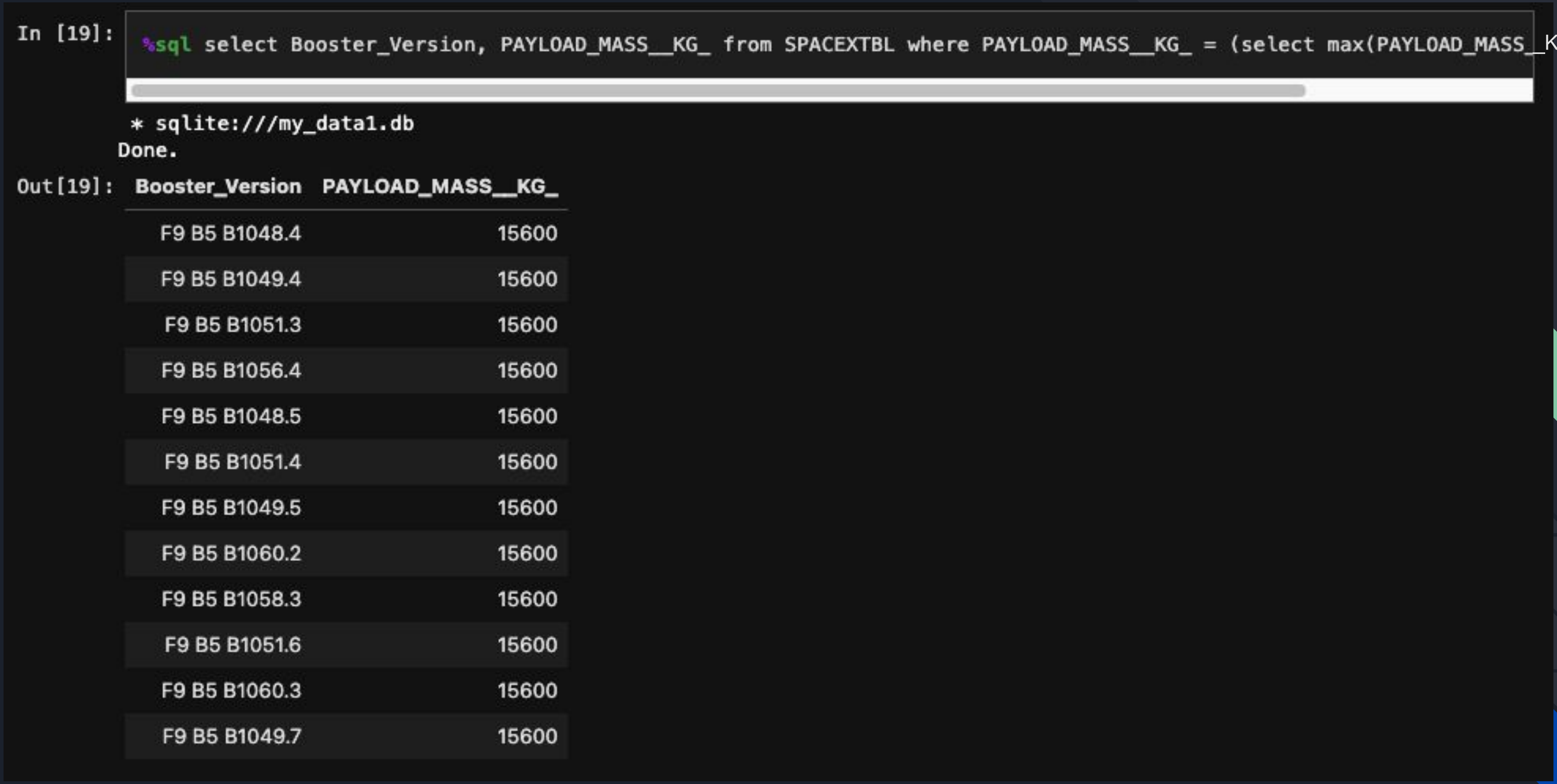
Out[18]:

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

Explanation:

- Displaying the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload



```
In [19]: %sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

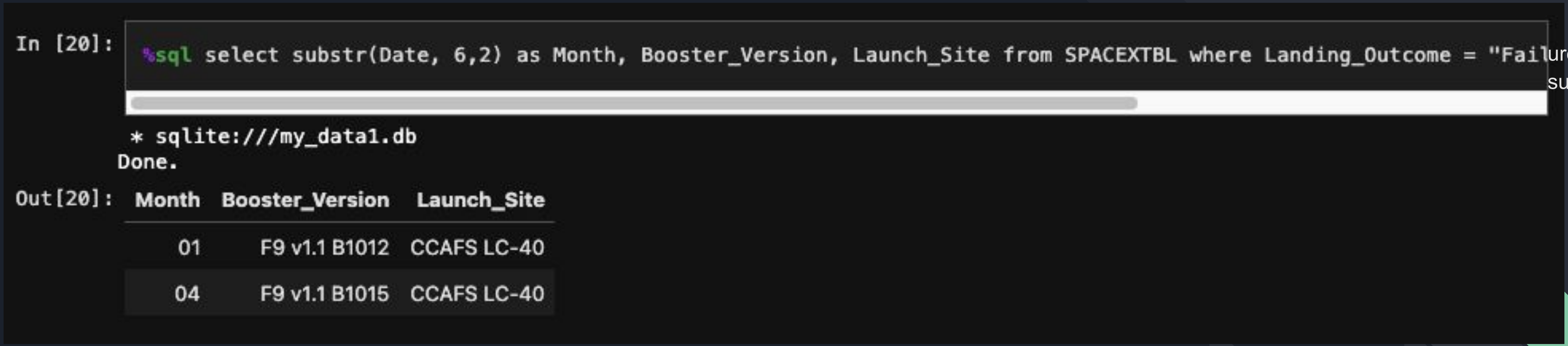
Out[19]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Explanation:

- Displaying the names of the boosters which have carried the maximum payload mass

2015 Launch Records



```
In [20]: %sql select substr(Date, 6,2) as Month, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome = "Failure (drone ship)" and
         substr(Date,0,5)='2015'

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

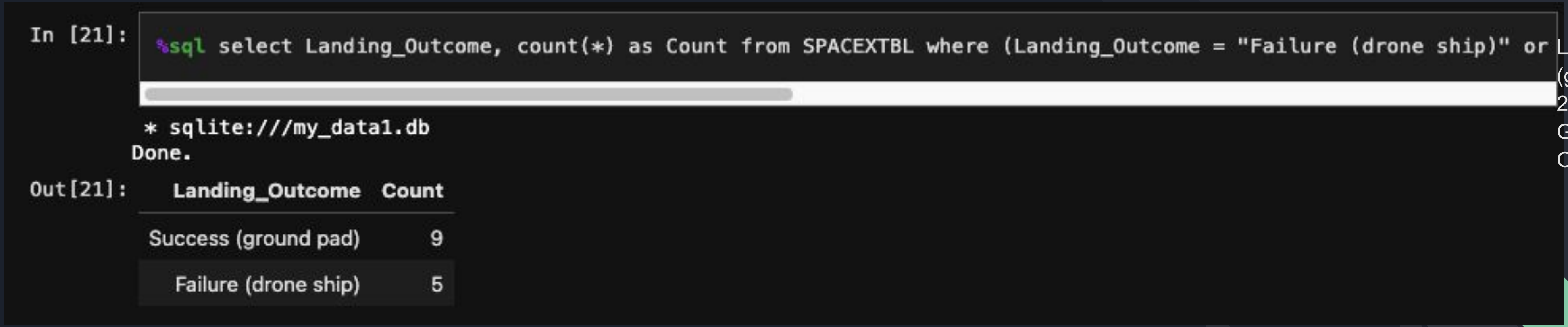
Out[20]:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Explanation:

- Displaying the names of the boosters which have been launched in 2015

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



```
In [21]:  %sql select Landing_Outcome, count(*) as Count from SPACEXTBL where (Landing_Outcome = "Failure (drone ship)" or

          * sqlite:///my_data1.db
         Done.
Out[21]:     Landing_Outcome   Count
           Success (ground pad)      9
             Failure (drone ship)    5
```

Landing_Outcome = "Success (ground pad)") and (Date > 2010-06-04 or Date < 2017-03-20) GROUP BY Landing_Outcome ORDER BY Count DESC

Explanation:

- Displaying the count of landing outcomes (Failure (drone ship) and Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

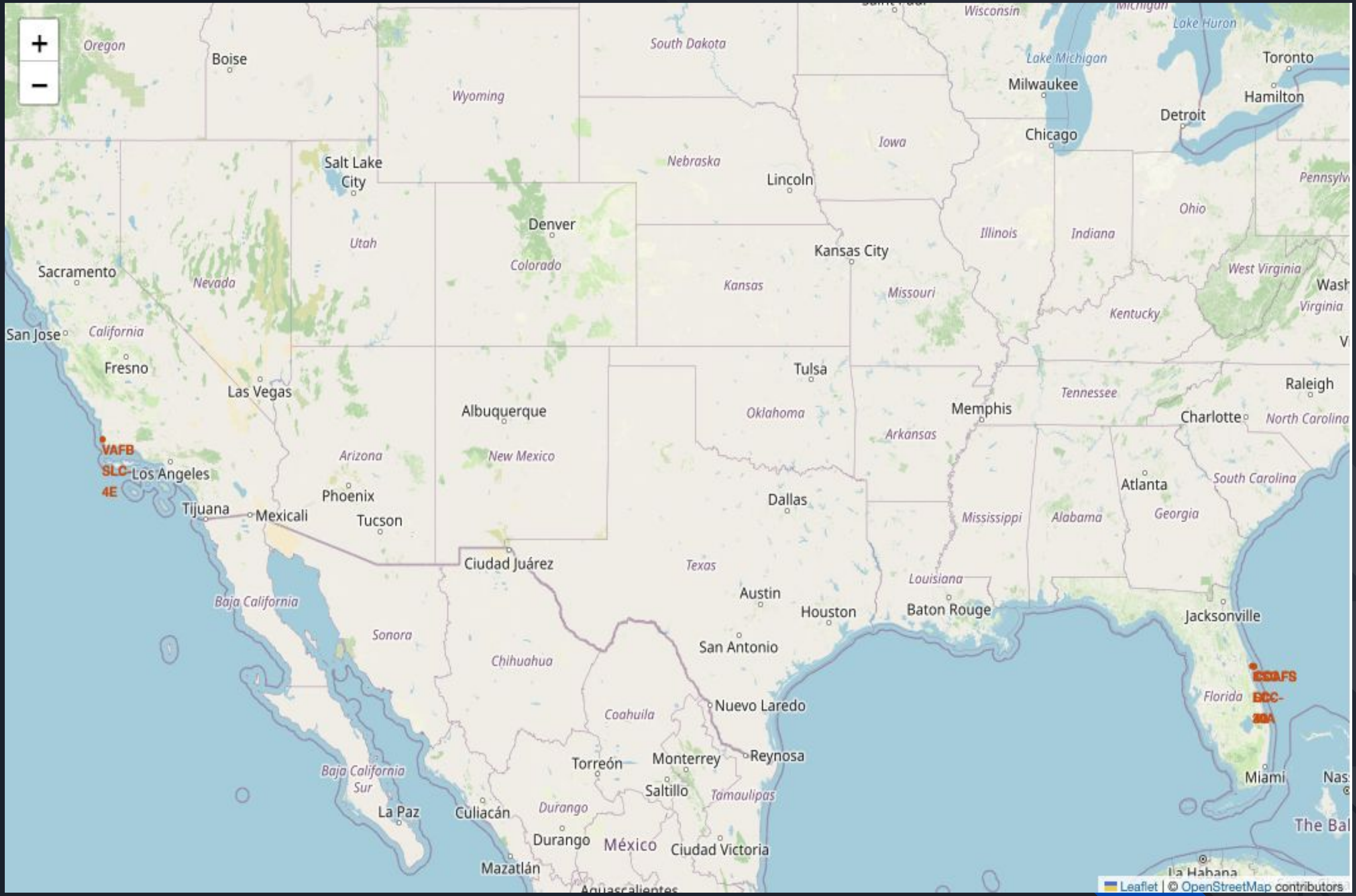
Section 3

Launch Sites Proximities Analysis

Map with marked launch sites

Explanation:

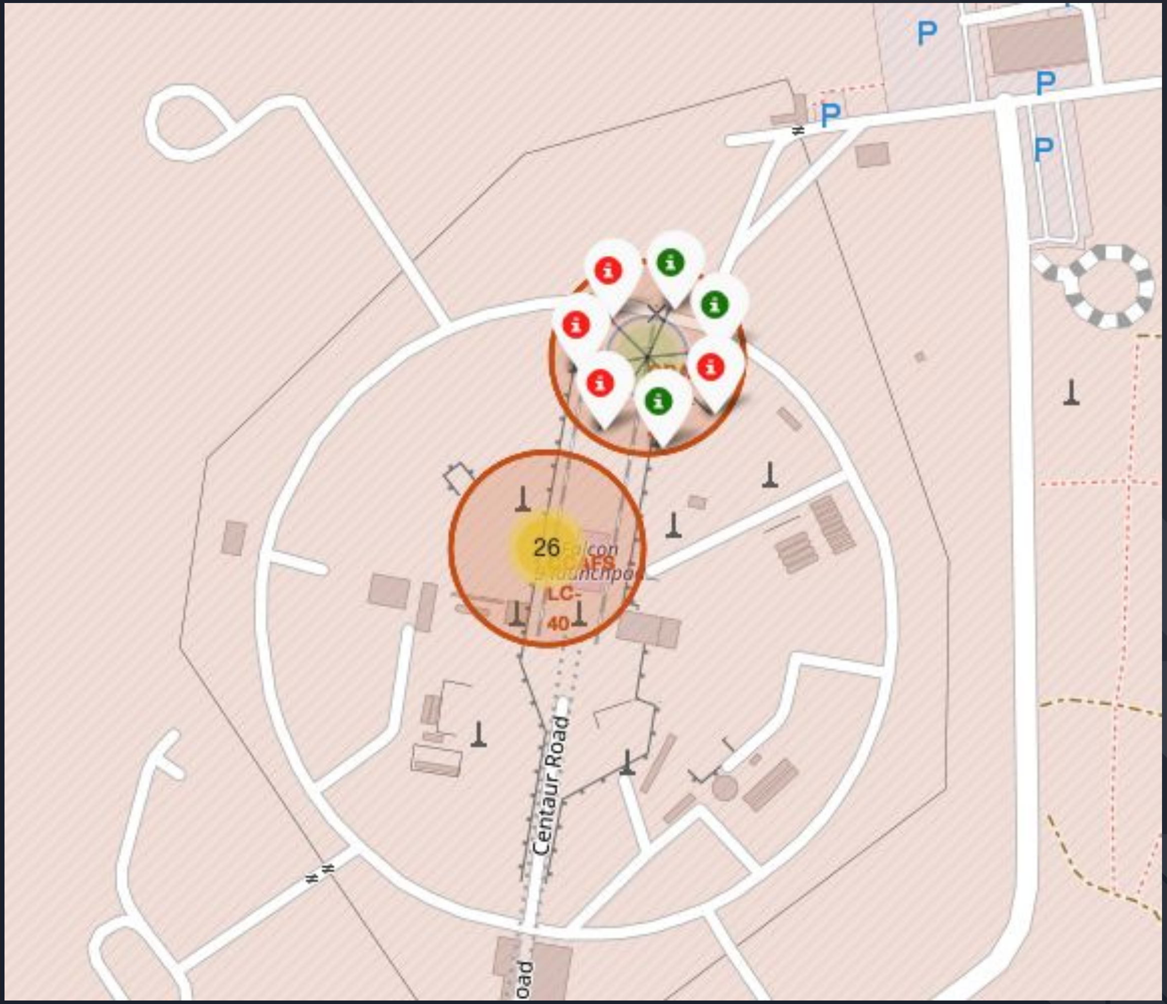
- All launch sites in proximity to the Equator line, because the Earth is moving faster at the equator than any other place on the surface. This speed helps the rockets keep up a good speed to stay in orbit
- All launch sites are very close to the coast, because it minimises the risk of exploding near people



Color-labeled launches on the map

Explanation:

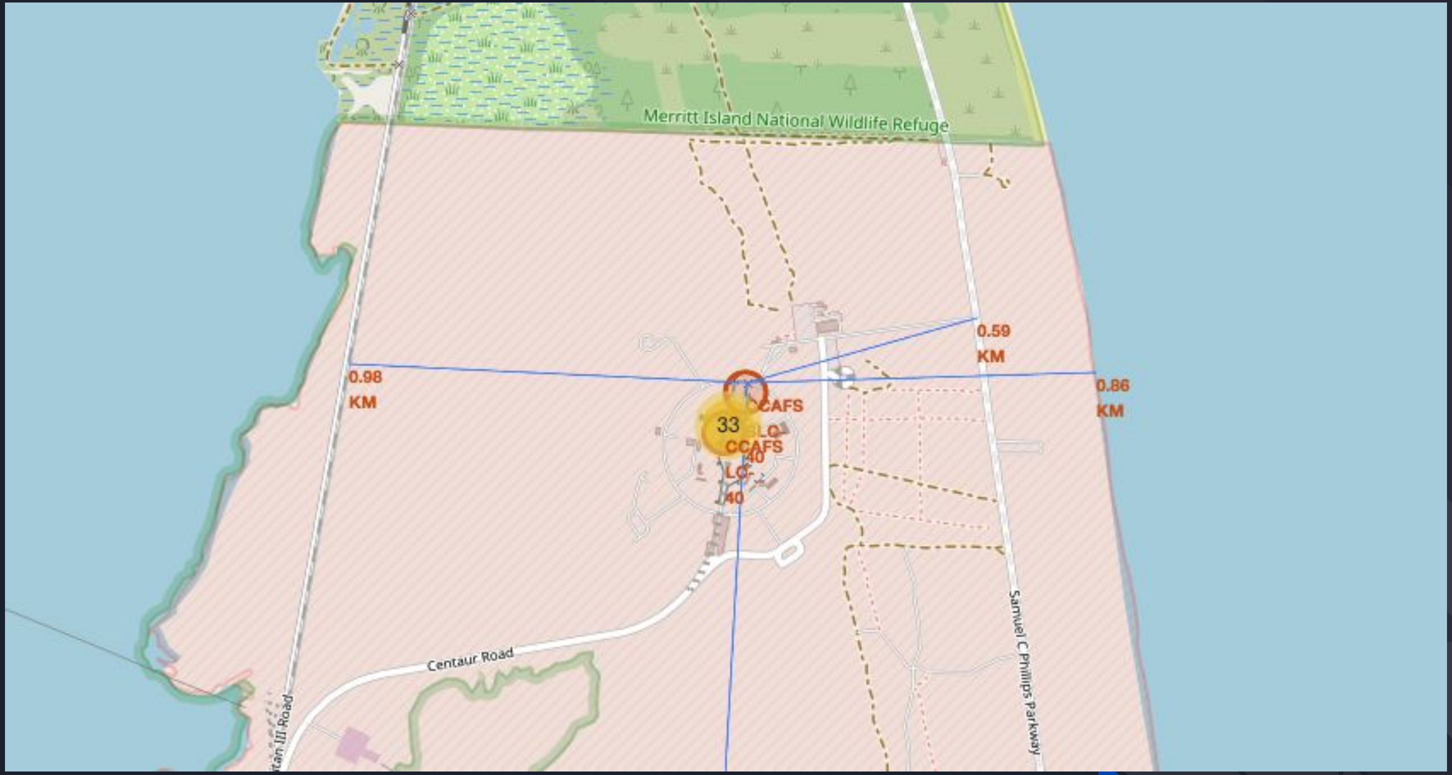
- All launch records are labeled into green or red (where green is successful launch and red - failure) markers, so we can easy identify which launch sites have high success rates
- The map also shows the number of all launches from each launch site



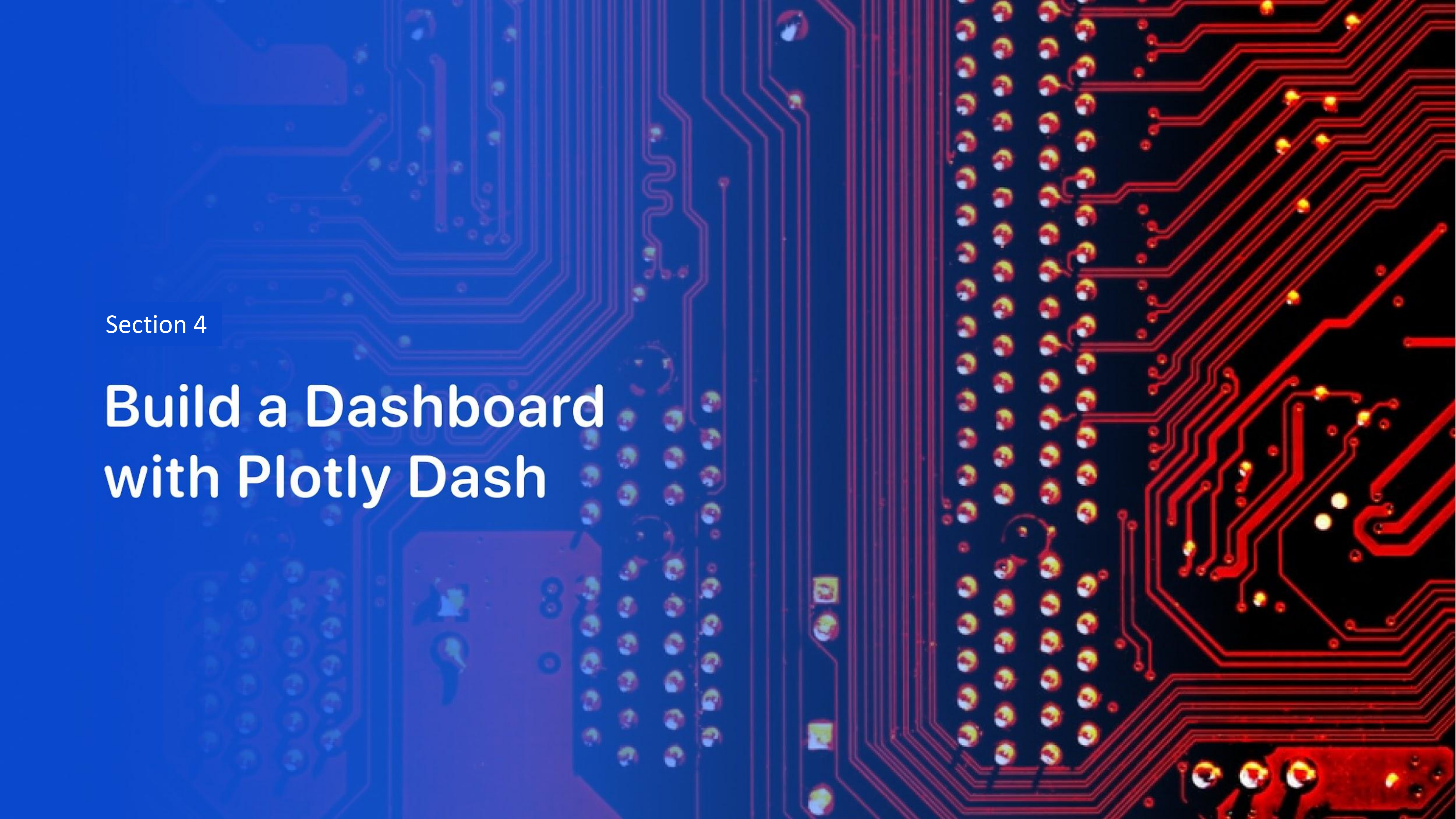
Distances from the launch site to its proximities

Explanation:

- For the launch site CCAFS SLC-40 we can see that:
  - distance to the nearest railway (0.98 km)
  - distance to the nearest highway (0.59 km)
  - distance to the nearest coastline (0.86 km)
  - distance to the nearest city (54.04 km)
- So this launch site is close to highway, railway and coastline but the distance to the nearest city is quite large



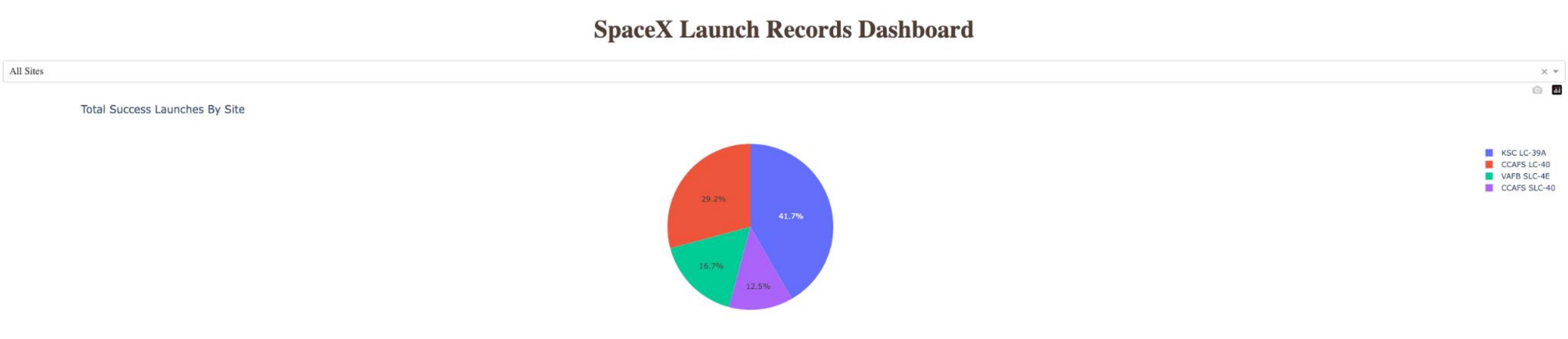
Melbourne
54.04 KM
Crane Creek



Section 4

Build a Dashboard with Plotly Dash

Launch success rate for all sites



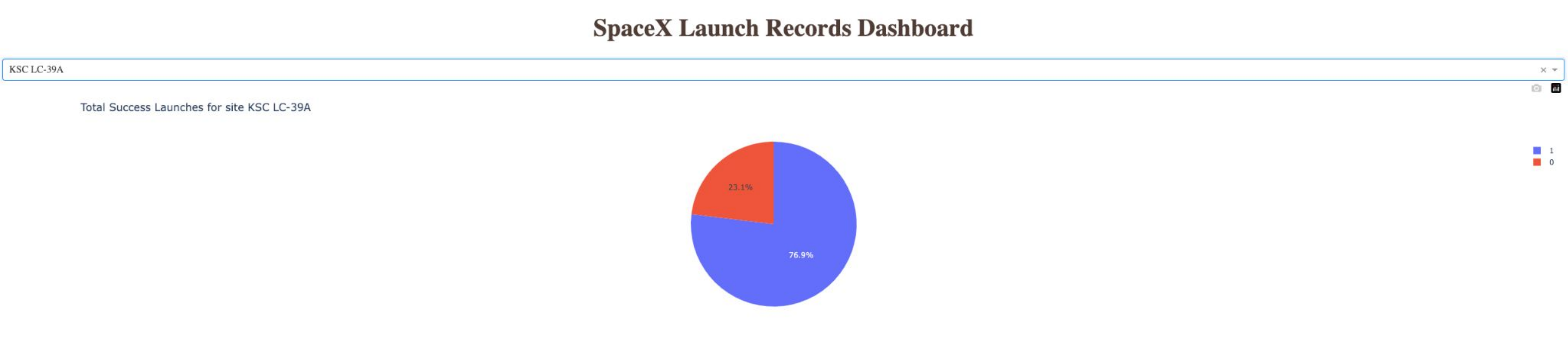
SpaceX Launch Records Dashboard

All Sites

Explanation:

- From the chart above we can identify which launch sites are more successful
- So, we can see, that KSC LC-39A launch site has the highest number of successful launches (41.7%)

KSC LC-39A - launch site with highest success rate



SpaceX Launch Records Dashboard

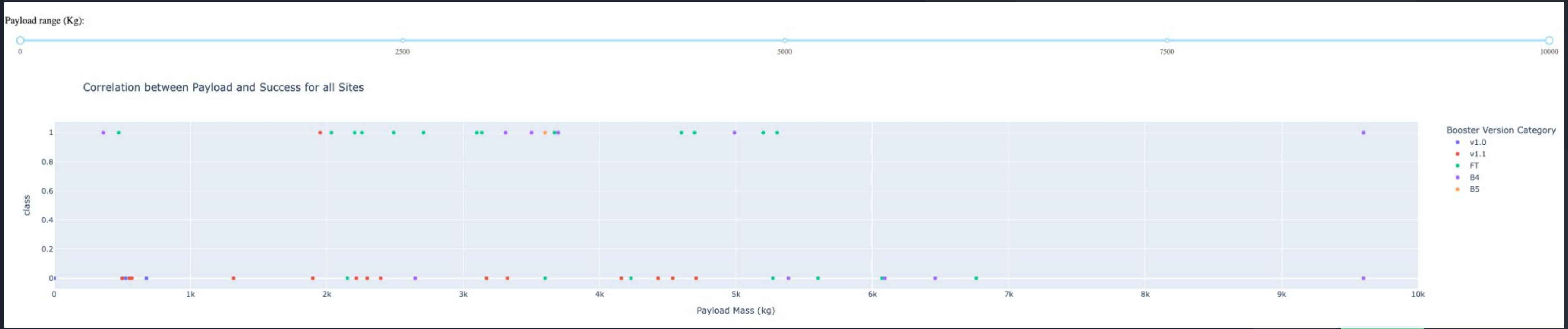
KSC LC-39A

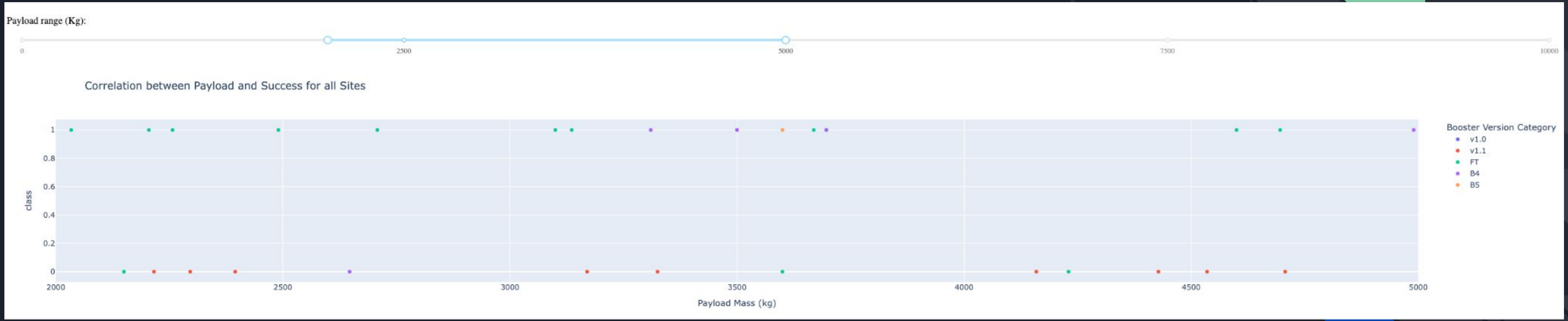
Total Success Launches for site KSC LC-39A

Explanation:

- The KSC LC-39A has a total of 13 launches, of which 10 were successful (76.9%) and only 3 failed

Payload mass vs Launch success for all sites





Explanation:

- From the scatter charts we can easily identify that payloads between 2000 and 5500 kg have the highest success rate
- We also see which versions of boosters are more successful

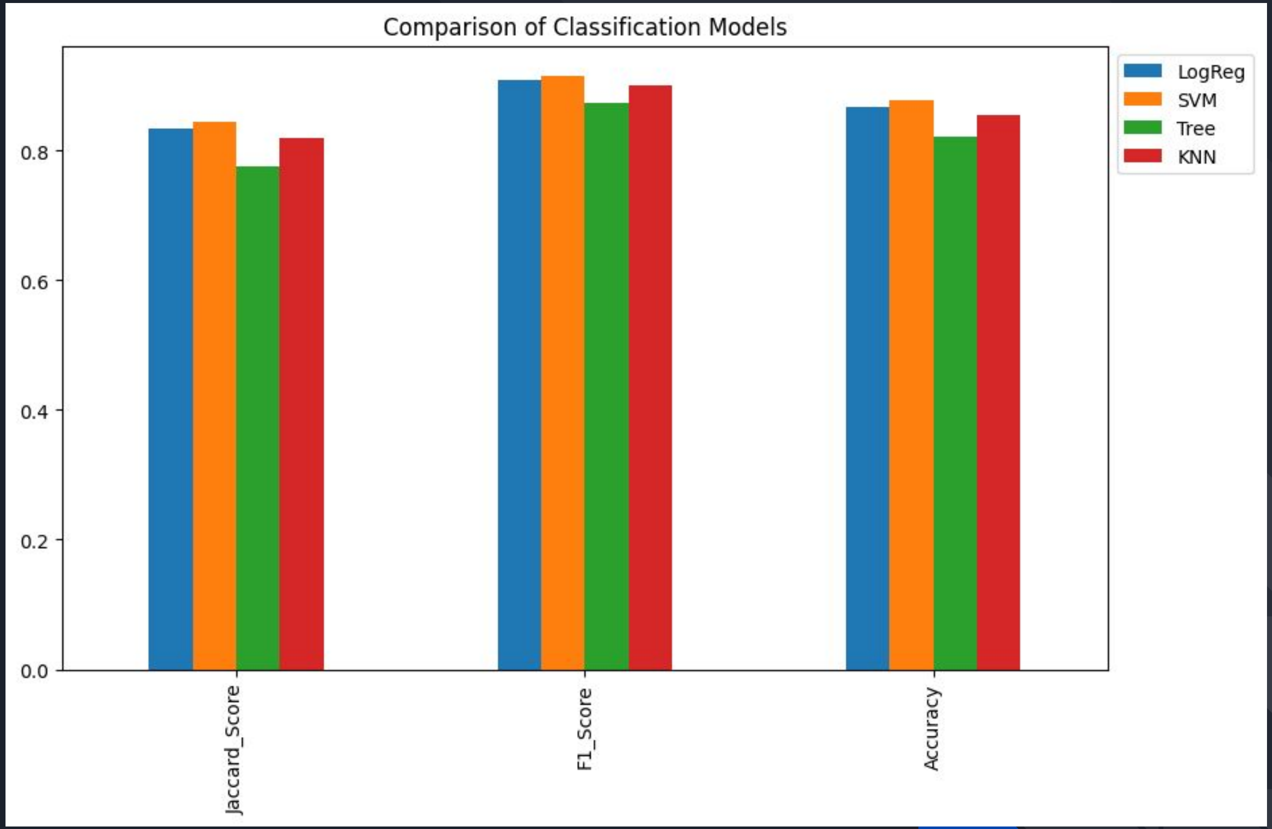
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

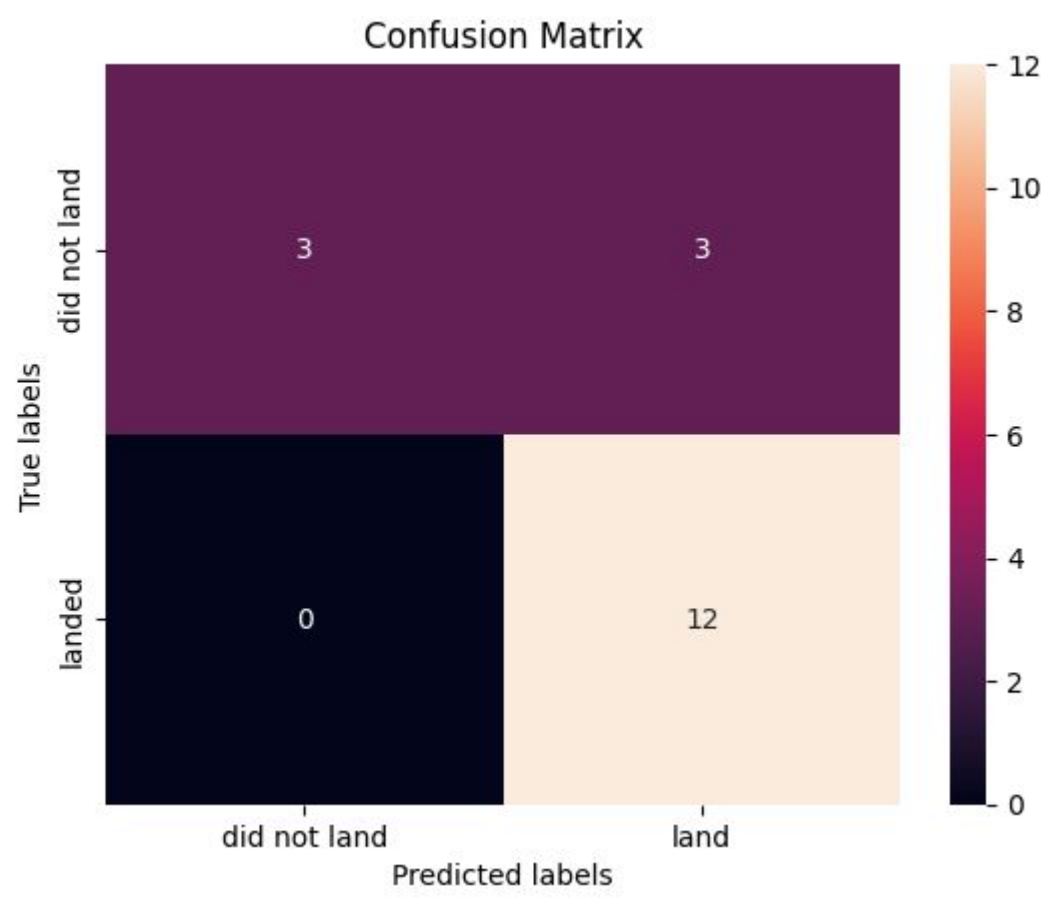
- On the bar chart we can see that SVM model (colored in orange) is the best model with highest Jaccard score, F1 score and Accuracy
- This scores is calculated for the whole dataset, as the accuracy of each classification model is the same on the test set (0.8333)



Confusion Matrix for SVM model

Explanation:

- Examining the confusion matrix, we see that Support Vector Machine (SVM) model can distinguish between the different classes. We see that the major problem is false positives (FP)



Conclusions

- Support Vector Machine (SVM) is the best machine learning classification algorithm for this dataset. Accuracy and scores of this model is well, so we can predict the success of Falcon 9's first stage landing.
- The most influential factors affecting the success of Falcon 9's first stage landing include:

  - Payload Mass (kg): Launches with a low payload mass show better results

  - Orbit Type: Orbits ES-L1, GEO, HEO and SSO have 100% success rate

  - Launch Site: KSC LC-39A has the highest success rate of the launches

Appendix

Special Thanks to:

IBM

Coursera

Thank you!