IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrés Felipe Castañeda Vargas
25/07/2022

Outline

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

Executive Summary

Summary of methodologies

- Data Collection API Lab
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

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

Introduction

- **Project background and context**

- Space X stands out from the competition due to the low cost of launching the Falcon 9 rockets. Space X claims these launches cost $62 million, while other providers cost $165 million each. one. This saving is since the company can reuse the first stage of the launch. Consequently, if it is possible to predict whether the first stage will land successfully, the cost of a launch can be determined; which comes in handy if an alternative company wants to bid against Space X for a rocket launch. This project aims to create a machine learning process for this purpose.

- **Problems you want to find answers**

- Determine the factors that influence whether the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Through Space X rest API and web scraping from Wikipedia

Perform data wrangling

- Using one-hot encoding to categorical features

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- How to build, tune, evaluate classification models

Data Collection

- On the one hand, the data was collected through a get request to the SpaceX API. Next, the content of the response is decoded into a pandas data frame using the .json_normalize() function. Once the data was obtained, they were cleaned, and the missing values were filled.
- On the other hand, Wikipedia web scraping was performed for the Falcon 9 release logs with the BeautifulSoup library. With this, the launch records were extracted from an HTML table and converted into a screen data frame for analysis.

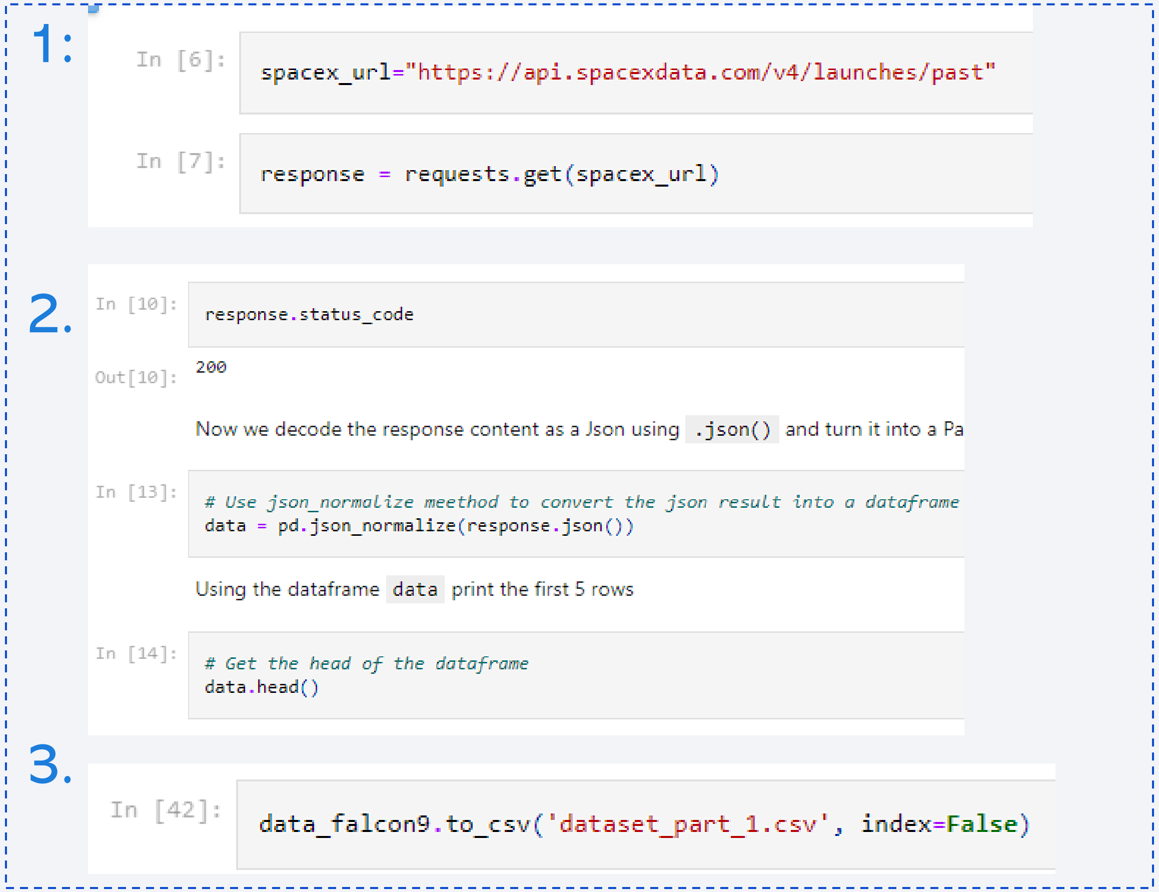
Data Collection – SpaceX API

1. Getting response from API.

2. Converting to a Pandas data frame.

3. Export to CSV file

- https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API%20Lab.ipynb



1:

```
In [6]:  spacex_url="https://api.spacexdata.com/v4/launches/past"

In [7]:  response = requests.get(spacex_url)
```

2.

```
In [10]:  response.status_code

Out[10]:  200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pa

```
In [13]:  # Use json_normalize meethod to convert the json result into a dataframe
          data = pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

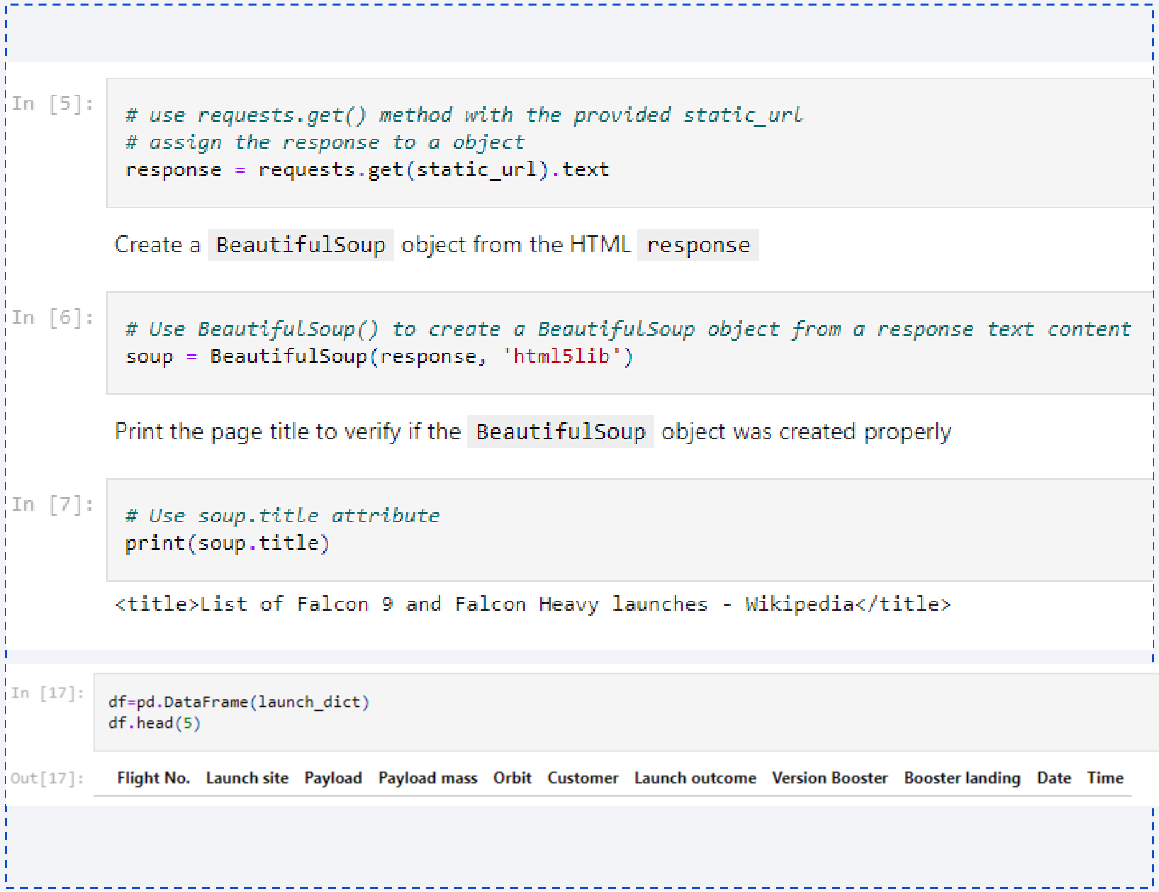
```
In [14]:  # Get the head of the dataframe
          data.head()
```

3.

```
In [42]:  data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Scraping

- Getting Response from HTMS and converting in Pandas data frame using the library BeautifulSoup
- https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



In [5]:

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `response`

In [6]:

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response, 'html5lib')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

In [7]:

```
# Use soup.title attribute
print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

In [17]:

```
df=pd.DataFrame(launch_dict)
df.head(5)
```

Out[17]:

Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time

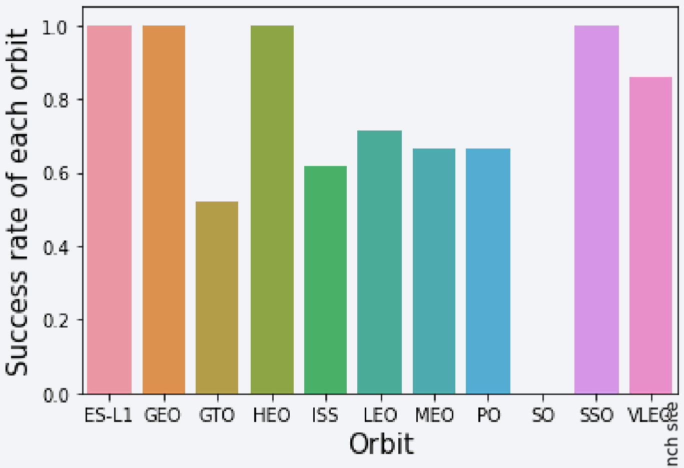
Data Wrangling

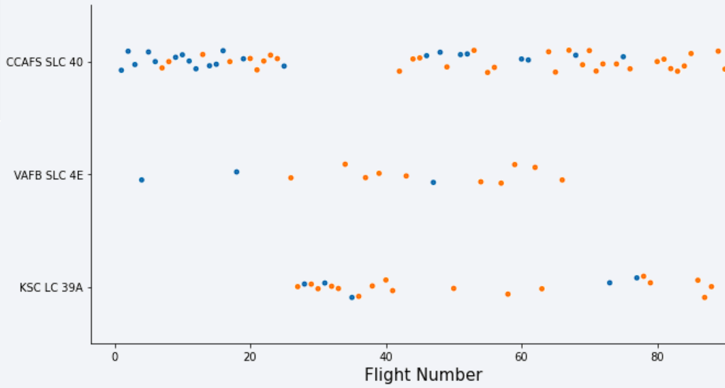
- Training labels were determined by exploratory data analysis.
- The number of launches at each site, and the number and occurrence of each orbit were determined.
- The results were exported to a CSV file by creating a landing results label from the results column.
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

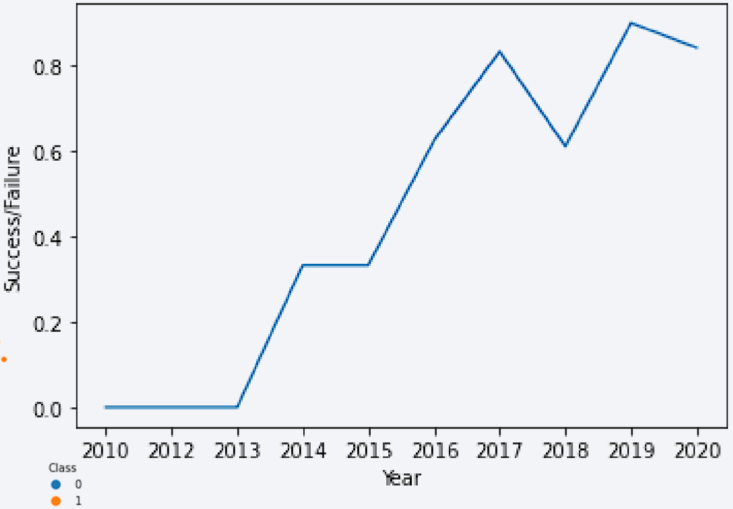
- https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/EDA.ipynb

EDA with Data Visualization

In this section, line graphs, scatter graphs and bar graphs are used to visualize the results of the problem.







https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Loaded the SpaceX dataset into a db2 database within the jupyter notebook. EDA with SQL was applied to obtain information from the data. Among the queries applied, information was obtained such as:
  - The names of launch sites unique to the space mission.
  - The total mass of the payload carried by the boosters launched by NASA
  - The average payload mass carried by the booster version.
  - The total number of successful and failed mission results.
  - The results of the unsuccessful landing on the drone, its booster version, and the names of the launch sites.
- https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL.ipynb

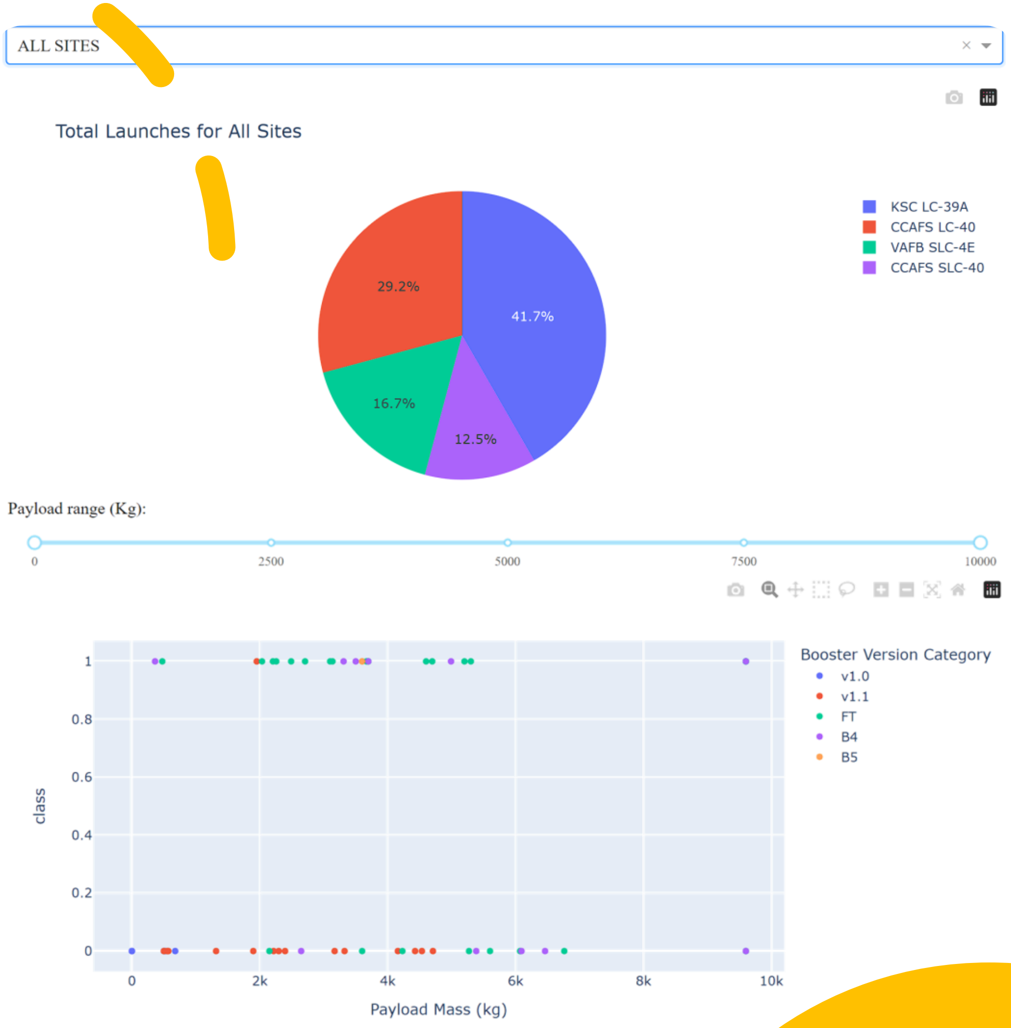
Build an Interactive Map with Folium

- All launch sites were marked with objects such as markers, circles, or lines to mark the success or failure of launches within the map. Assigning the class 0, if it fails, and 1, if it is successful.
- Groups of color-labeled markers were also assigned to identify sites with the highest success rate.
- The distance between the launch site and its surroundings was also calculated.

- https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Firstly, a dropdown input menu was added with which you can select the specific launch site you want to analyze.
- Next, added a pie chart to visualize launch success counts and a range slider to select the payload.
- Additionally, a scatter plot was added between the payload, on the X axis, and the launch result, on the Y axis, to visually identify their correlation.
- https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py



Predictive Analysis (Classification)

Using the Pandas and Numpy libraries, the data transformation was performed, and they were divided into training and test data.

With this, different machine learning models were built, taking into account accuracy as the model metric; improving them with some feature engineering techniques and algorithm tuning.

https://github.com/Soka0/Applied-Data-Science-Capstone/blob/master/Machine%20Learning%20Prediction.ipynb

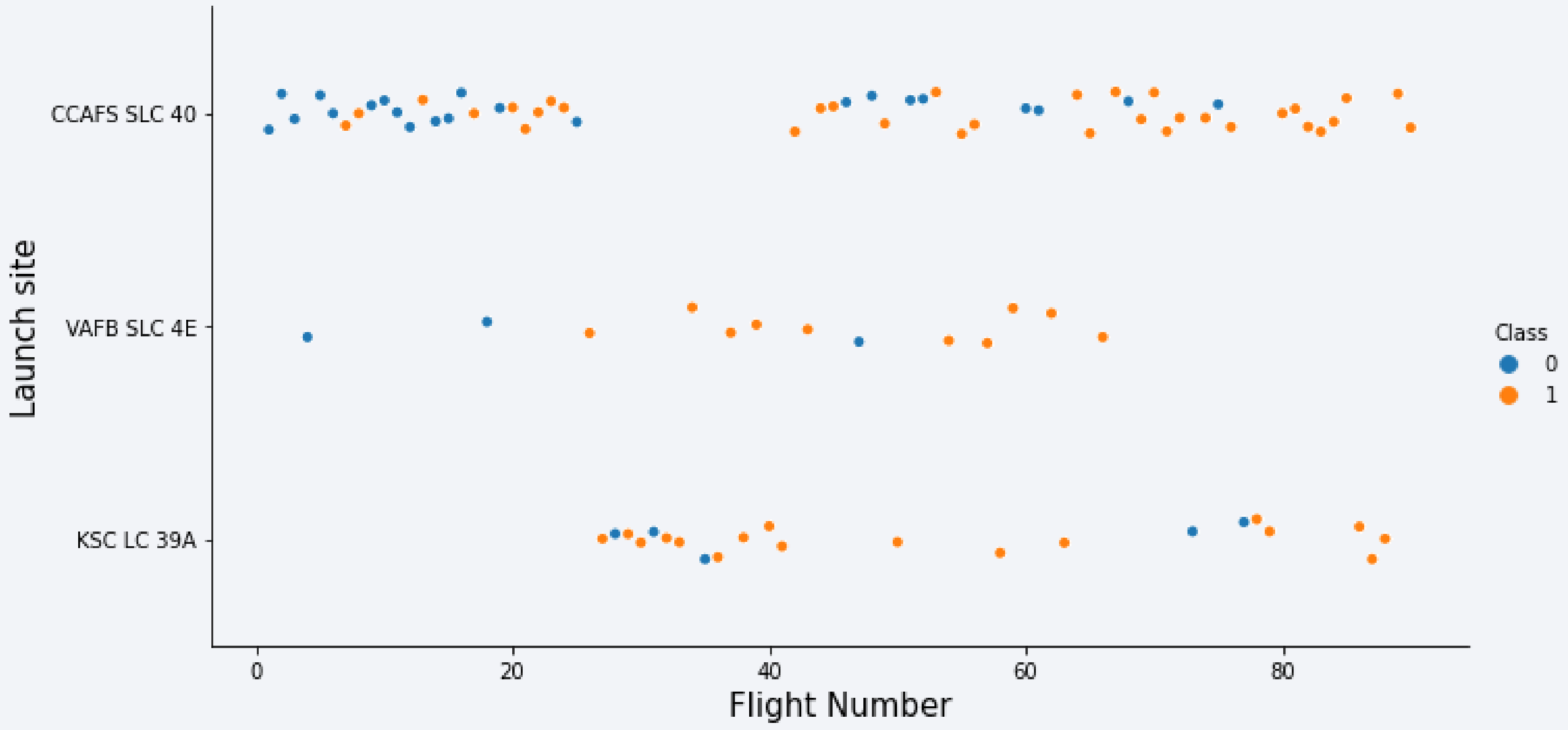
Results

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

Section 2

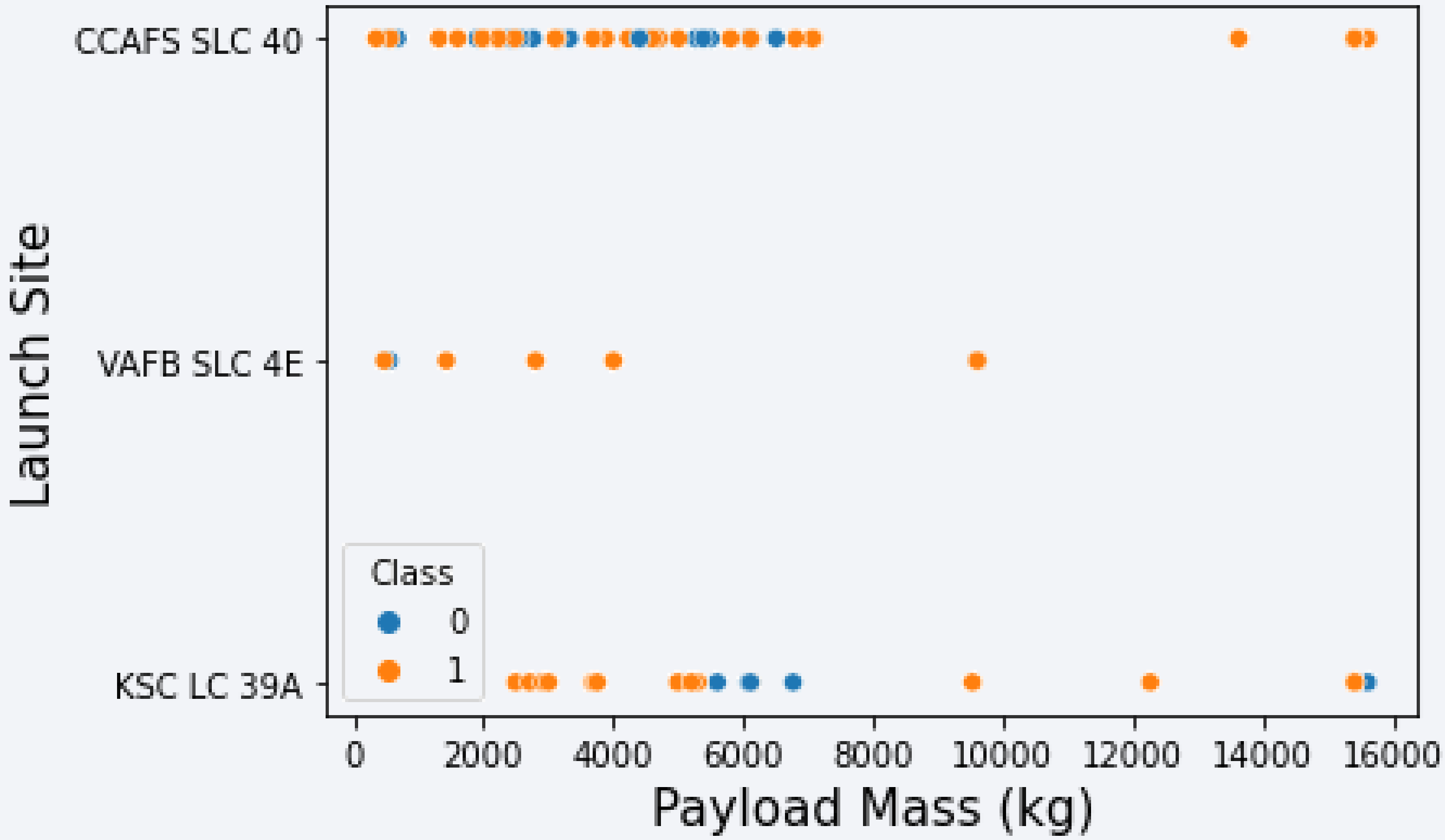
Insights drawn from EDA

Flight Number vs. Launch Site



Launches from the CCAFS LC-40 launch site are noticeably more successful.

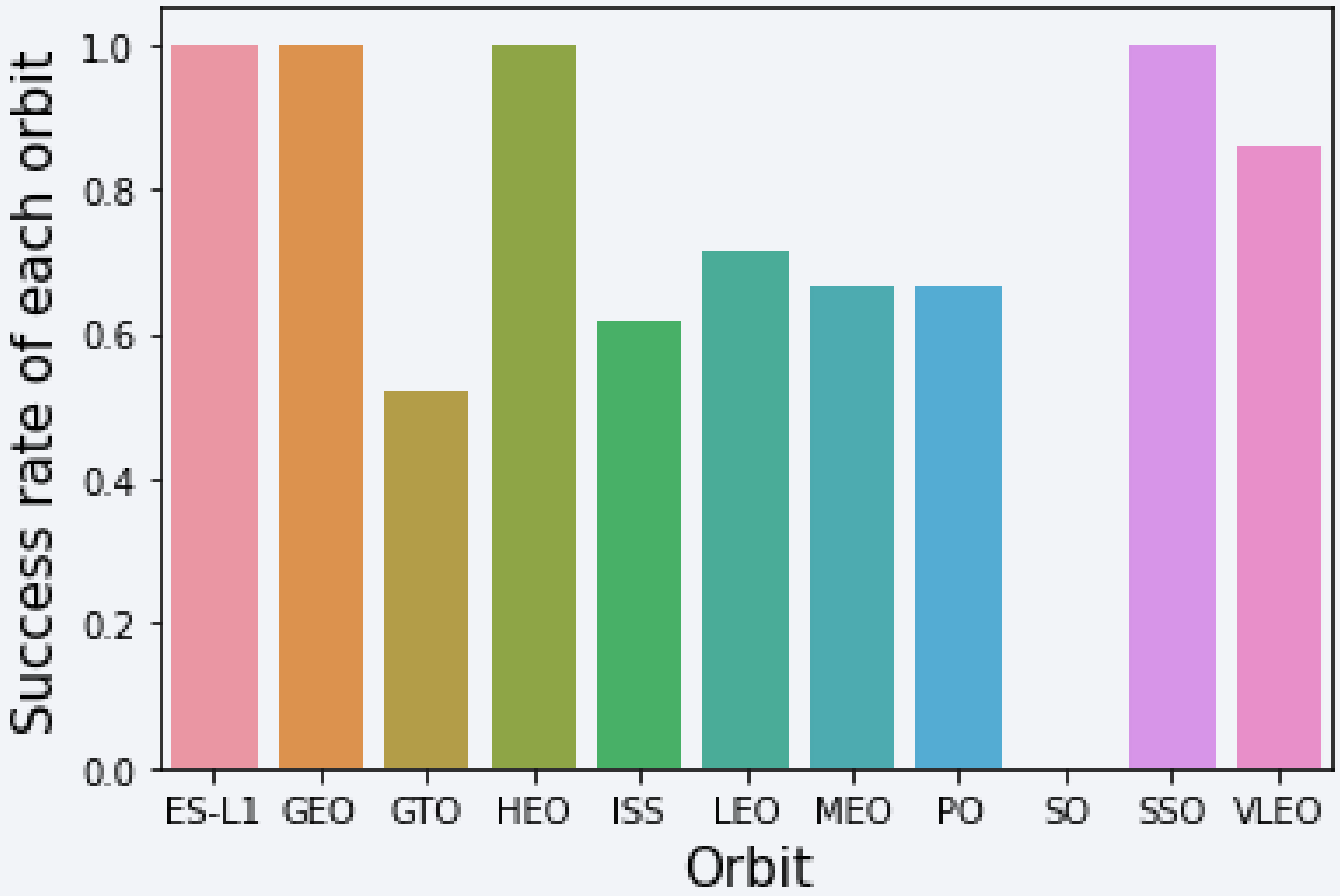
Payload vs. Launch Site



It is noted that most launches at the CCAFS SLC 40 launch site do not exceed 8000 kg, which may be a factor in its higher number of successful launches.

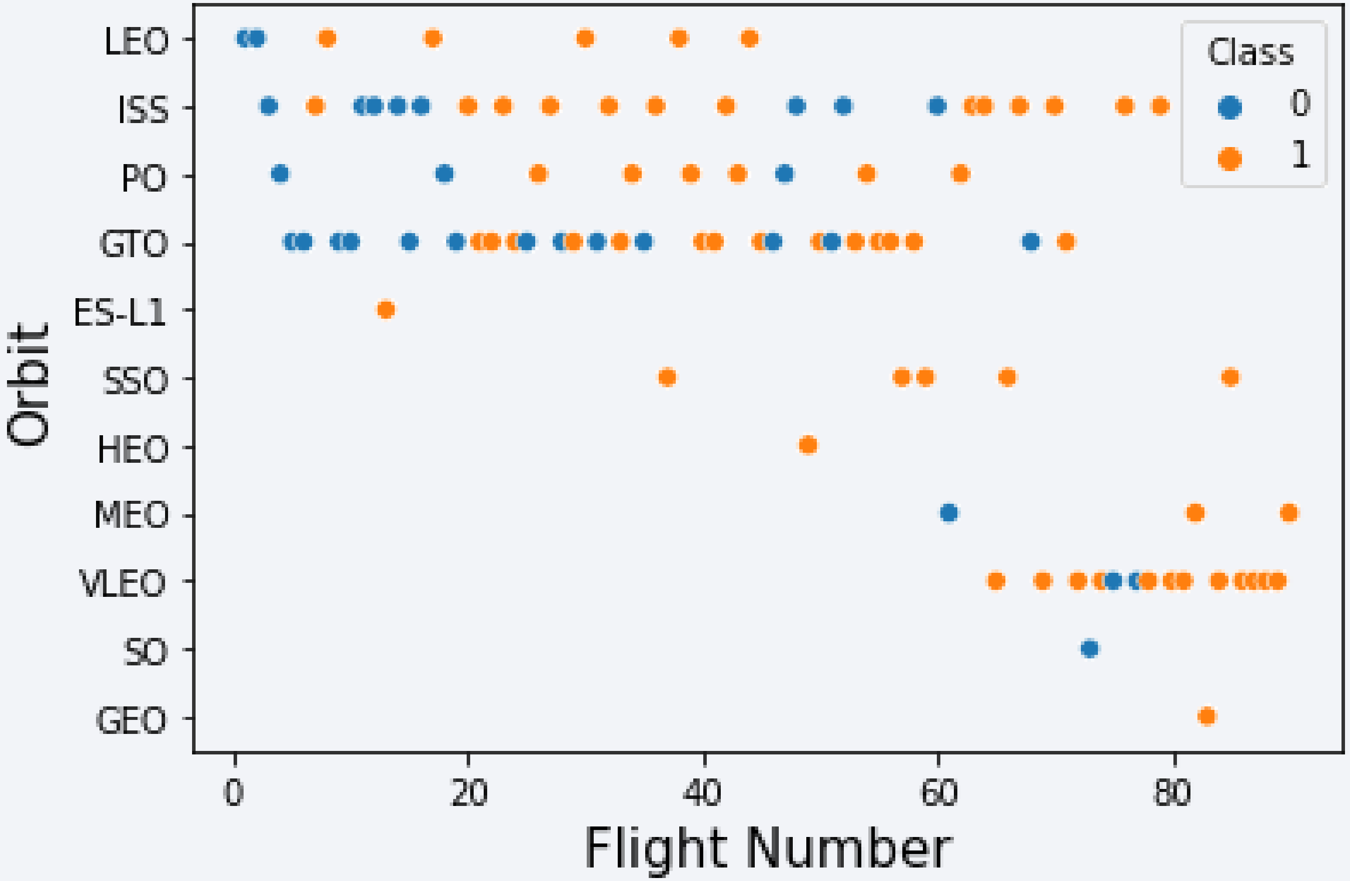
Success Rate vs. Orbit Type

Launches to orbits ES-L1, GEO, HEO and SSO have a success rate of 100%, while launches to orbits SO and GTO have the lowest success rate; being 0% for the case of the SO orbit.

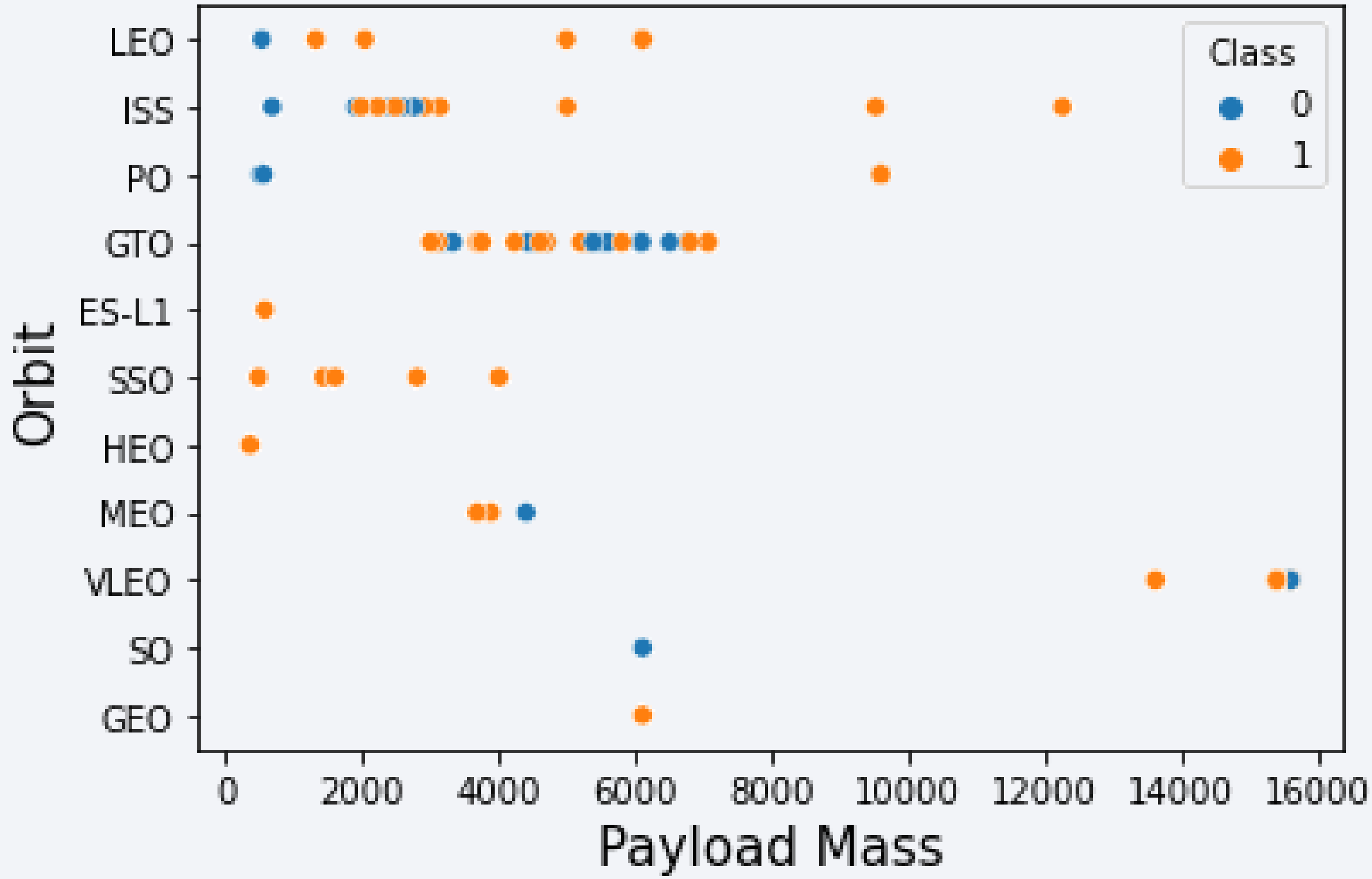


Flight Number vs. Orbit Type

It is well known that there is a relationship with the number of flights in the LEO orbit. However, in the GTO orbit, this same relationship does not seem to exist.



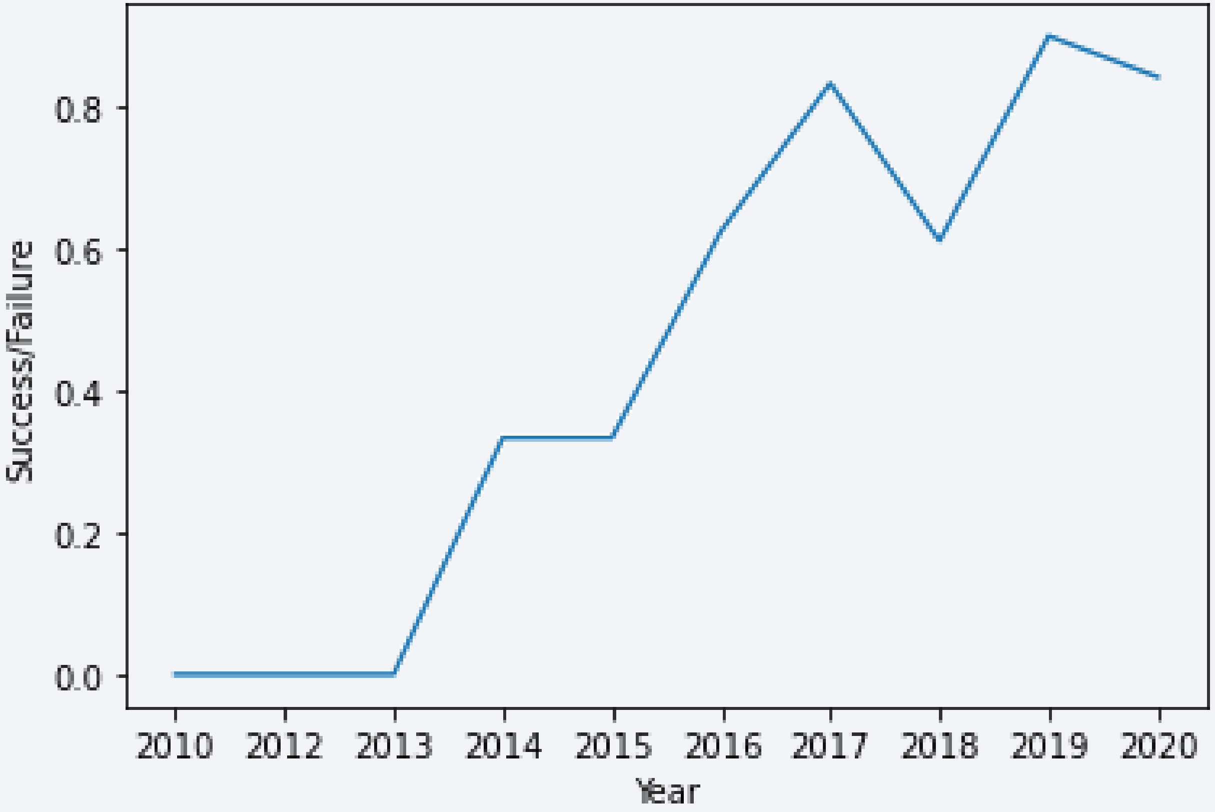
Payload vs. Orbit Type



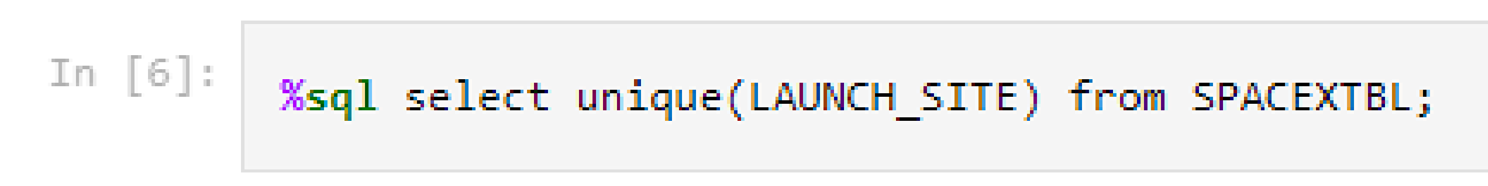
The highest successful landing rate is possessed by Polar, LEO and ISS orbits. For the GTO orbit part, it cannot be distinguished correctly because the successful and unsuccessful landing rate are found here.

Launch Success Yearly Trend

Since 2013, the rate of successful launches has increased markedly, with the highest success rate in 2019.

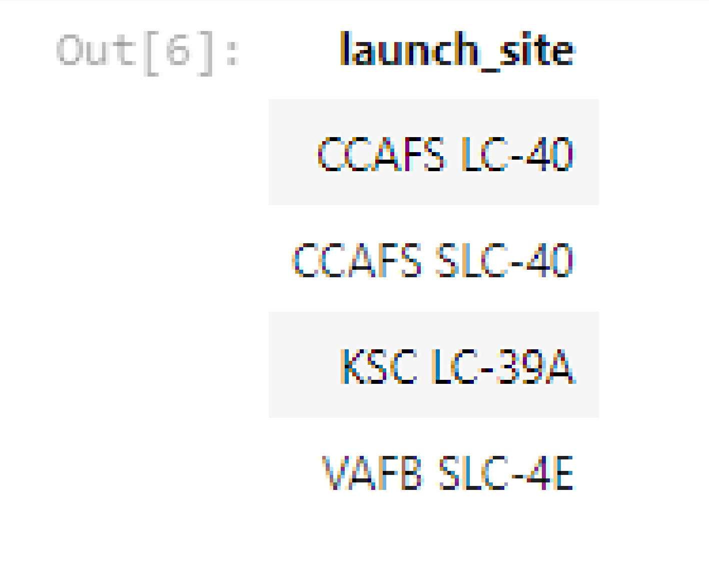


All Launch Site Names



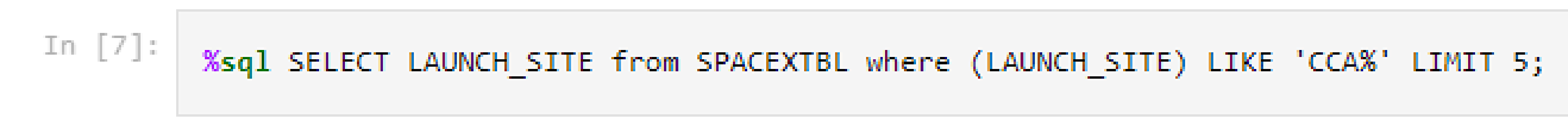
```
In [6]:  %sql select unique(LAUNCH_SITE) from SPACEXTBL;
```

Using the unique SQL query, the list of launch site names shown below was extracted.



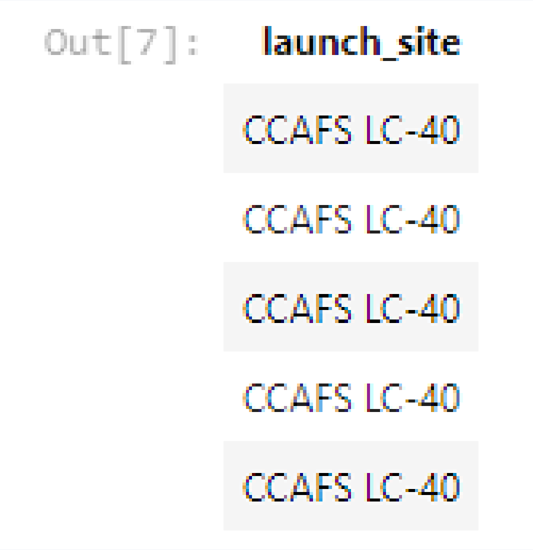
```
Out[6]:    launch_site
          CCAFS LC-40
         CCAFS SLC-40
           KSC LC-39A
          VAFB SLC-4E
```

Launch Site Names Begin with 'CCA'



```
In [7]:  %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

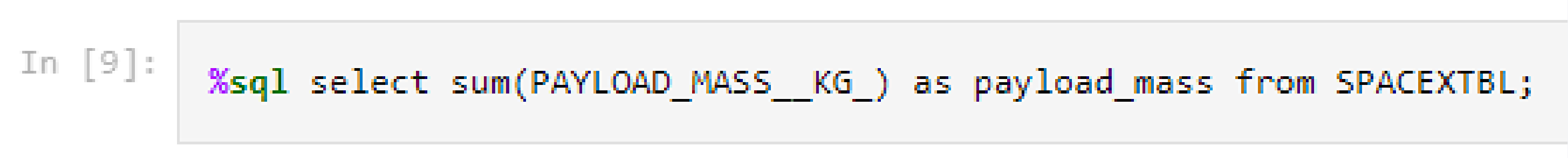
Using the SQL query, we got 5 records where the launch sites start with the string 'CCA'



Out[7]:

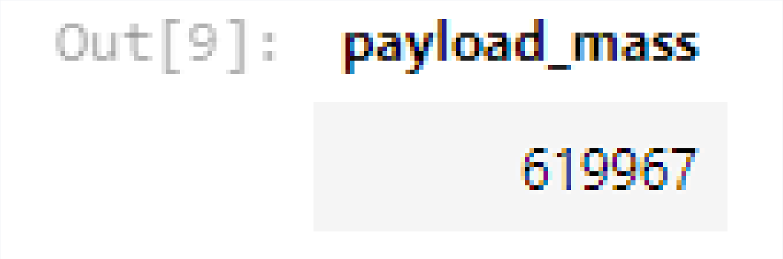
launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass



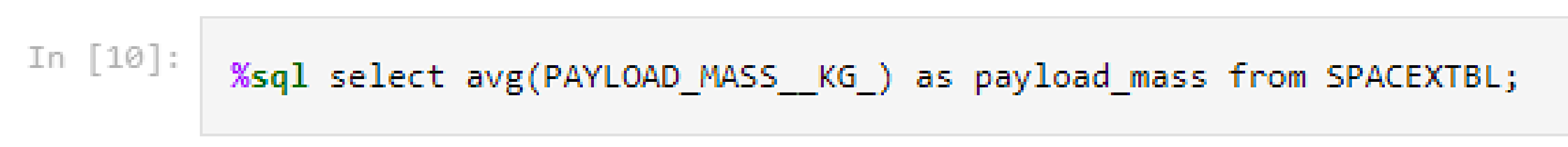
```
In [9]:  %sql select sum(PAYLOAD_MASS__KG_) as payload_mass from SPACEXTBL;
```

Using the SQL query, the total mass of the payload carried by the boosters launched by NASA (CRS) was obtained.



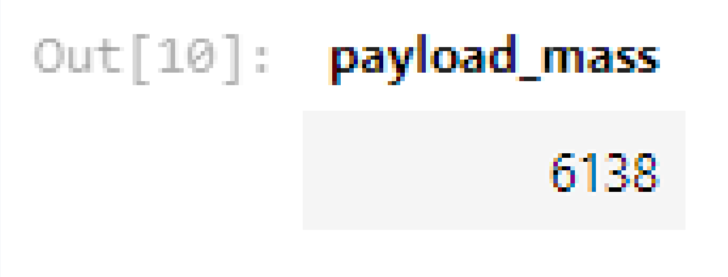
```
Out[9]:  payload_mass
               619967
```

Average Payload Mass by F9 v1.1



```
In [10]: %sql select avg(PAYLOAD_MASS__KG_) as payload_mass from SPACEXTBL;
```

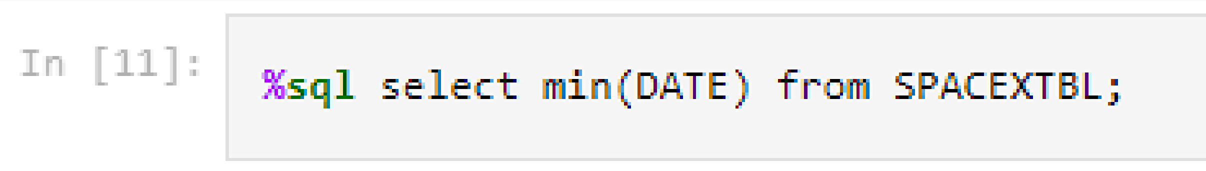
Through the SQL query, the average payload mass carried by the F9 v1.1 booster version was obtained.



Out[10]:

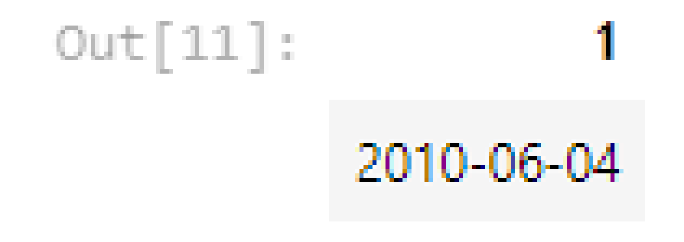
payload_mass
6138

First Successful Ground Landing Date



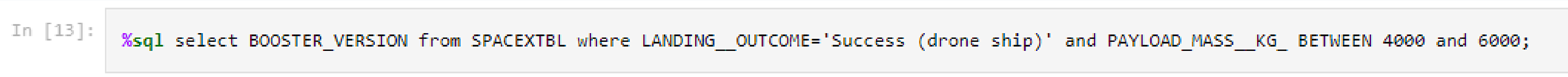
```
In [11]:  %sql select min(DATE) from SPACEXTBL;
```

Through the SQL query, the date of the first successful landing outcome on ground pad was obtained.



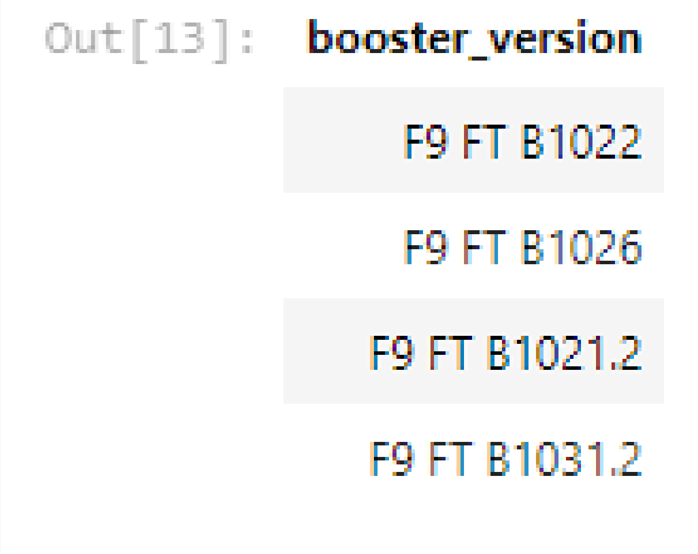
```
Out[11]:           1
          2010-06-04
```

Successful Drone Ship Landing with Payload between 4000 and 6000



```
In [13]: %sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

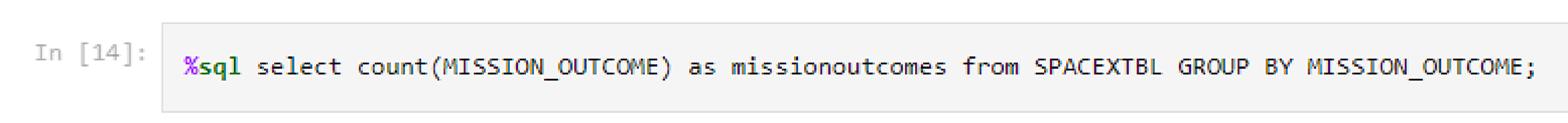
Through the SQL query, the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 was obtained.



Out[13]:

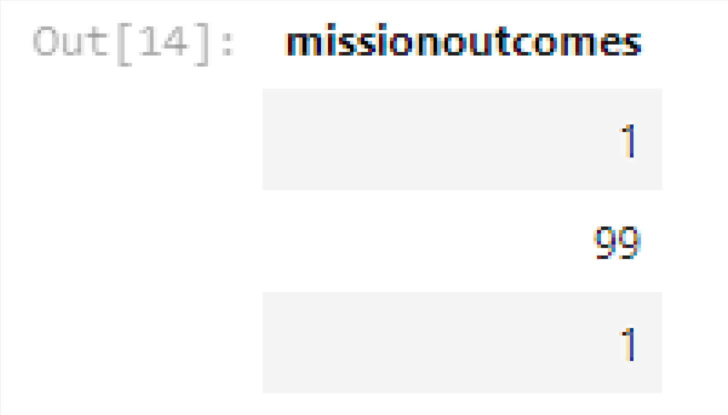
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes



```
In [14]:  %sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

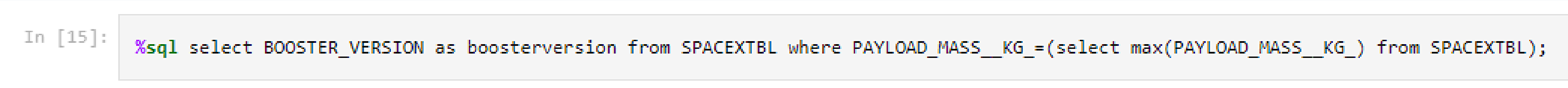
Through the SQL query, the total number of successful and failure mission outcomes was obtained.



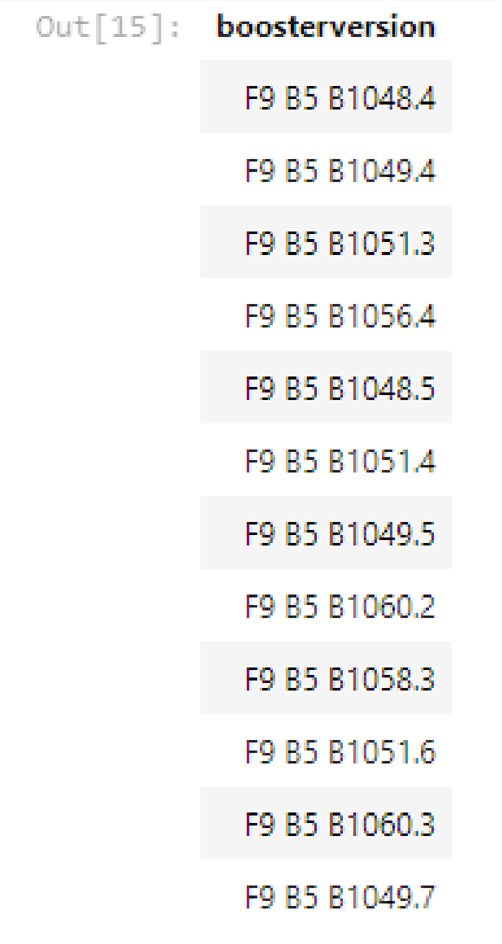
Out[14]:

missionoutcomes
1
99
1

Boosters Carried Maximum Payload



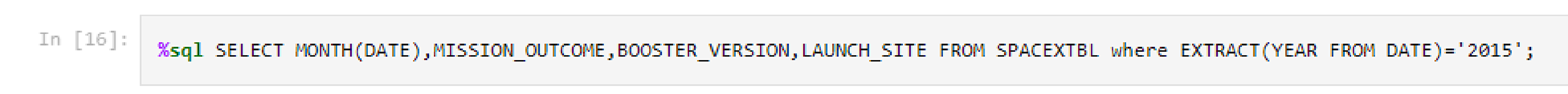
```
In [15]: %sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```



```
Out[15]: boosterversion
         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
```

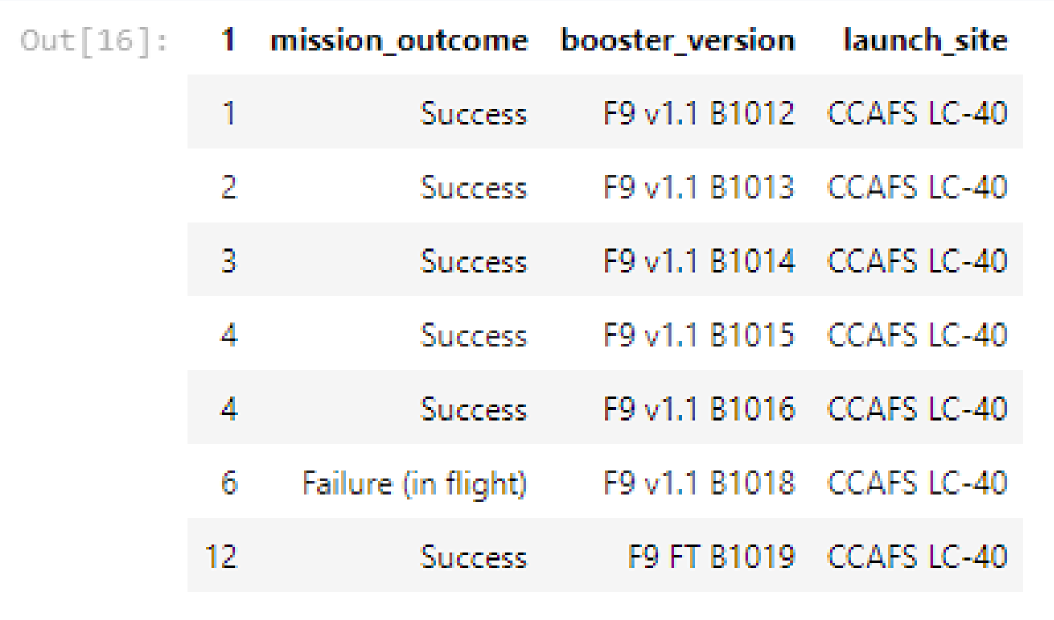
Through the SQL query, the names of the booster which have carried the maximum payload mass was obtained.

2015 Launch Records



```
In [16]:  %sql SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
```

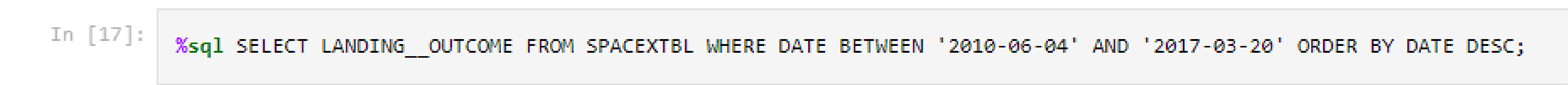
Through the SQL query, the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 was obtained.



Out[16]:

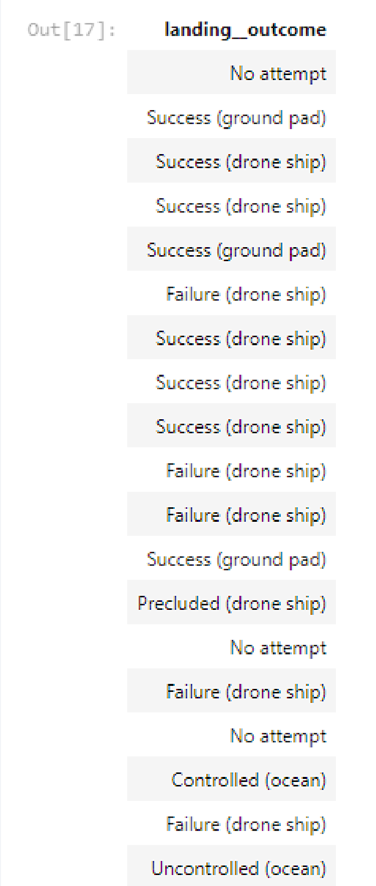
1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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



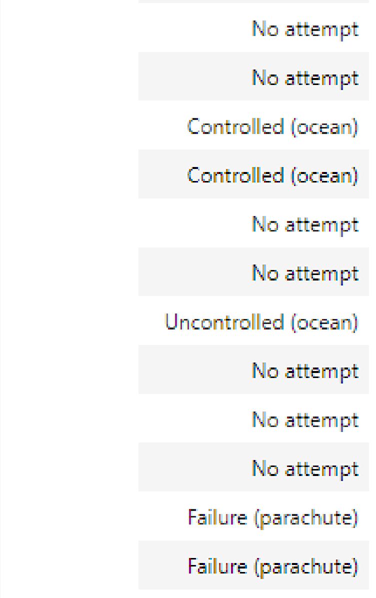
```
In [17]: %sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

Through the SQL query, 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 was obtained.



Out[17]:

landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)

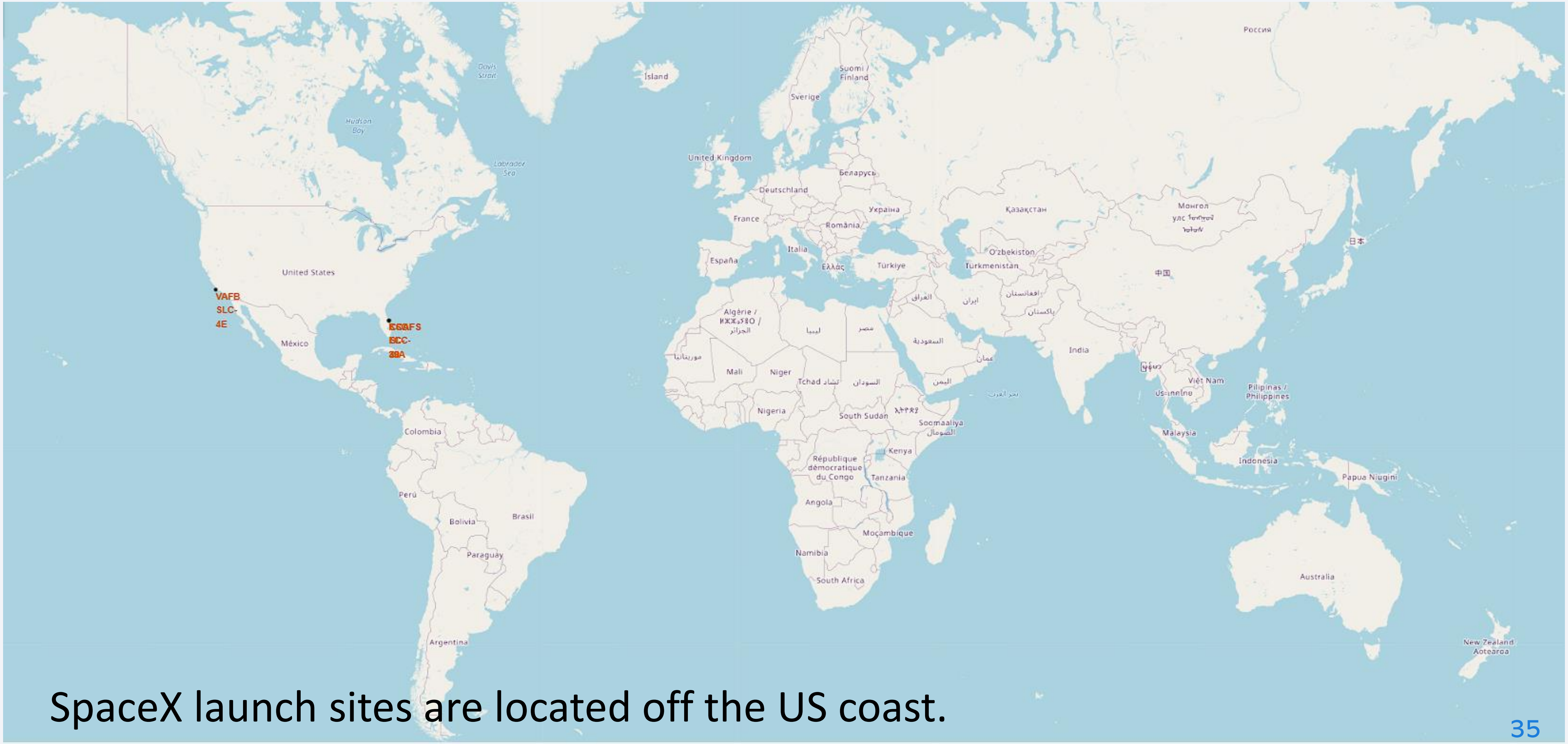


No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

Section 3

Launch Sites Proximities Analysis

Launch sites on a global map

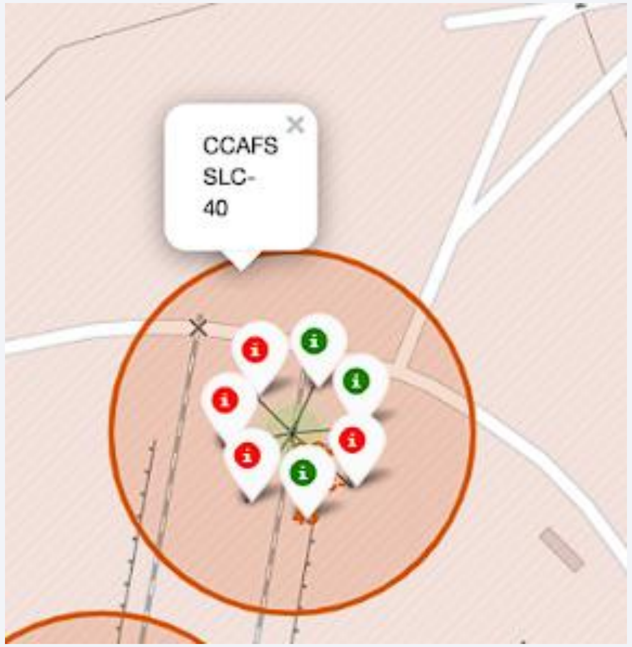


SpaceX launch sites are located off the US coast.

Color-labeled launch outcomes on the map

Two colored markers are added:

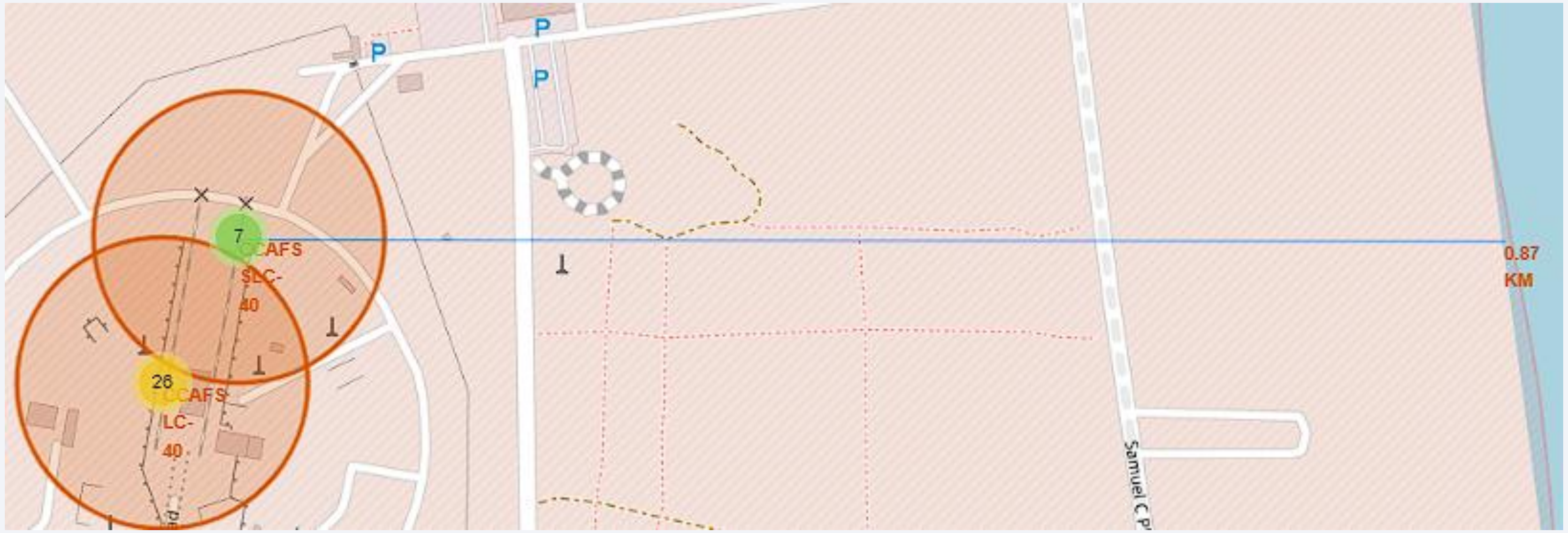
- Green for successful launch
- Red for unsuccessful launch



Launch site proximities to railway, highway or coastline

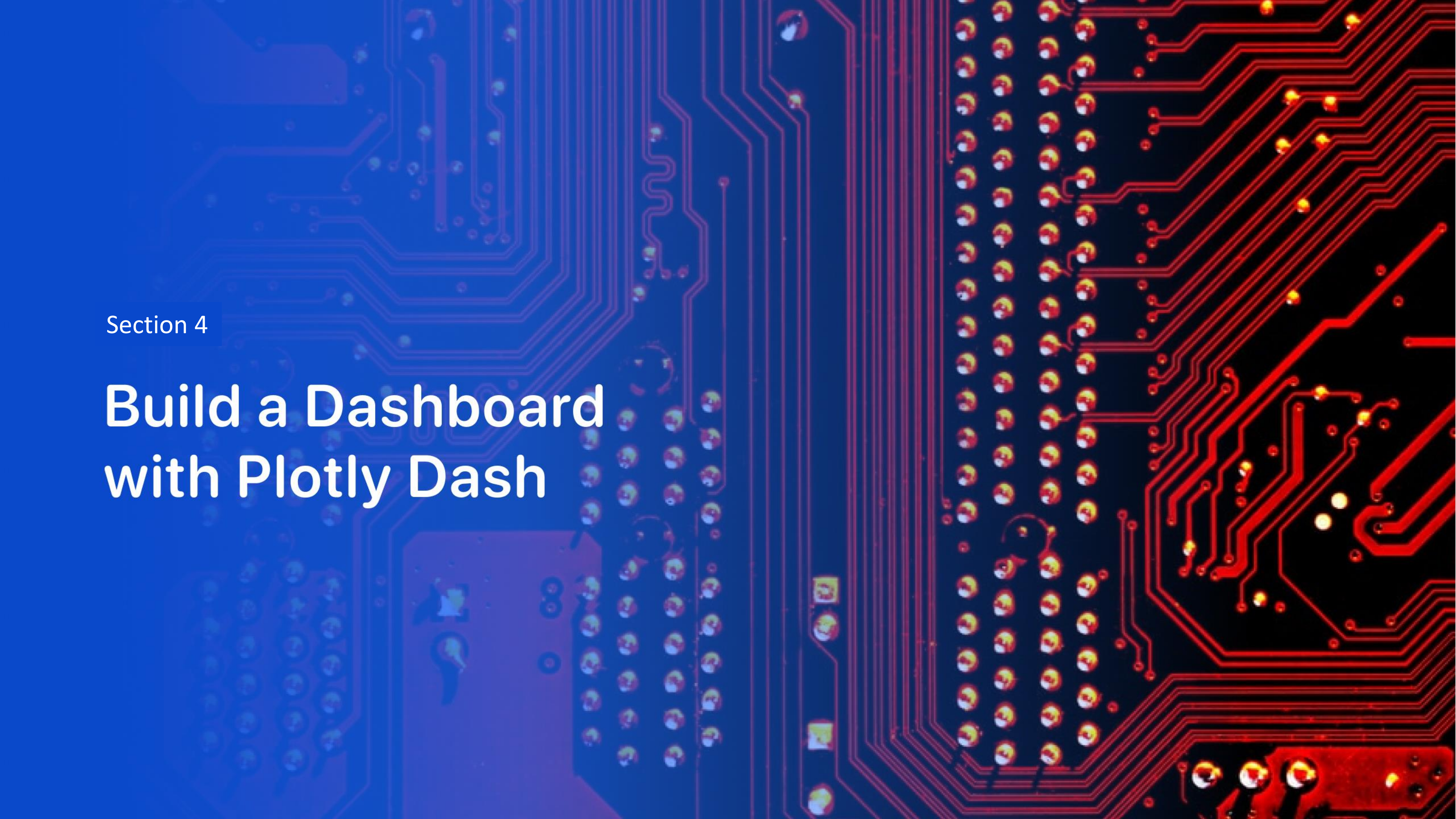
For the selected launch site:

- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? No



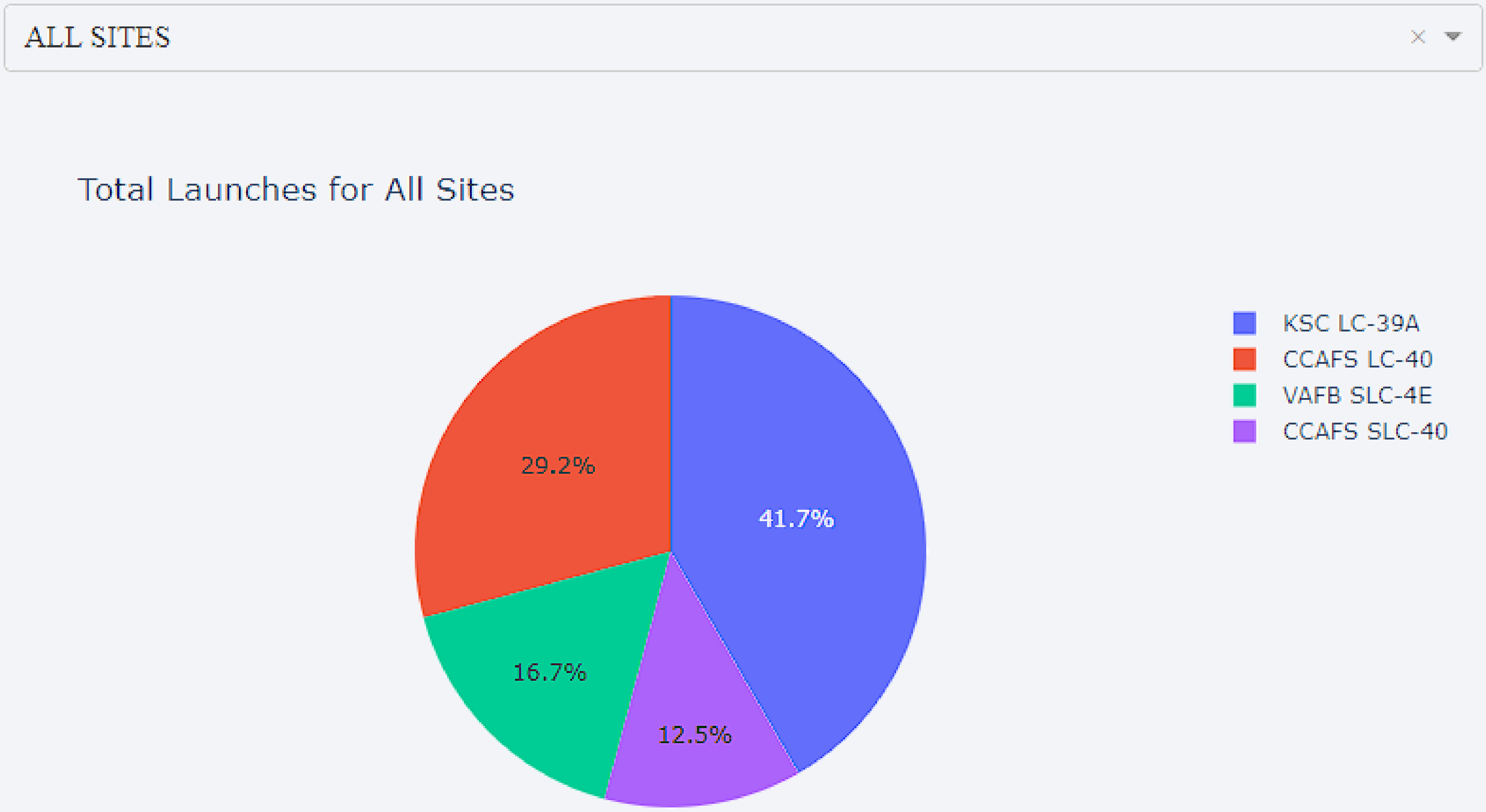
Section 4

Build a Dashboard with Plotly Dash



Percentage of total launches by each launch site

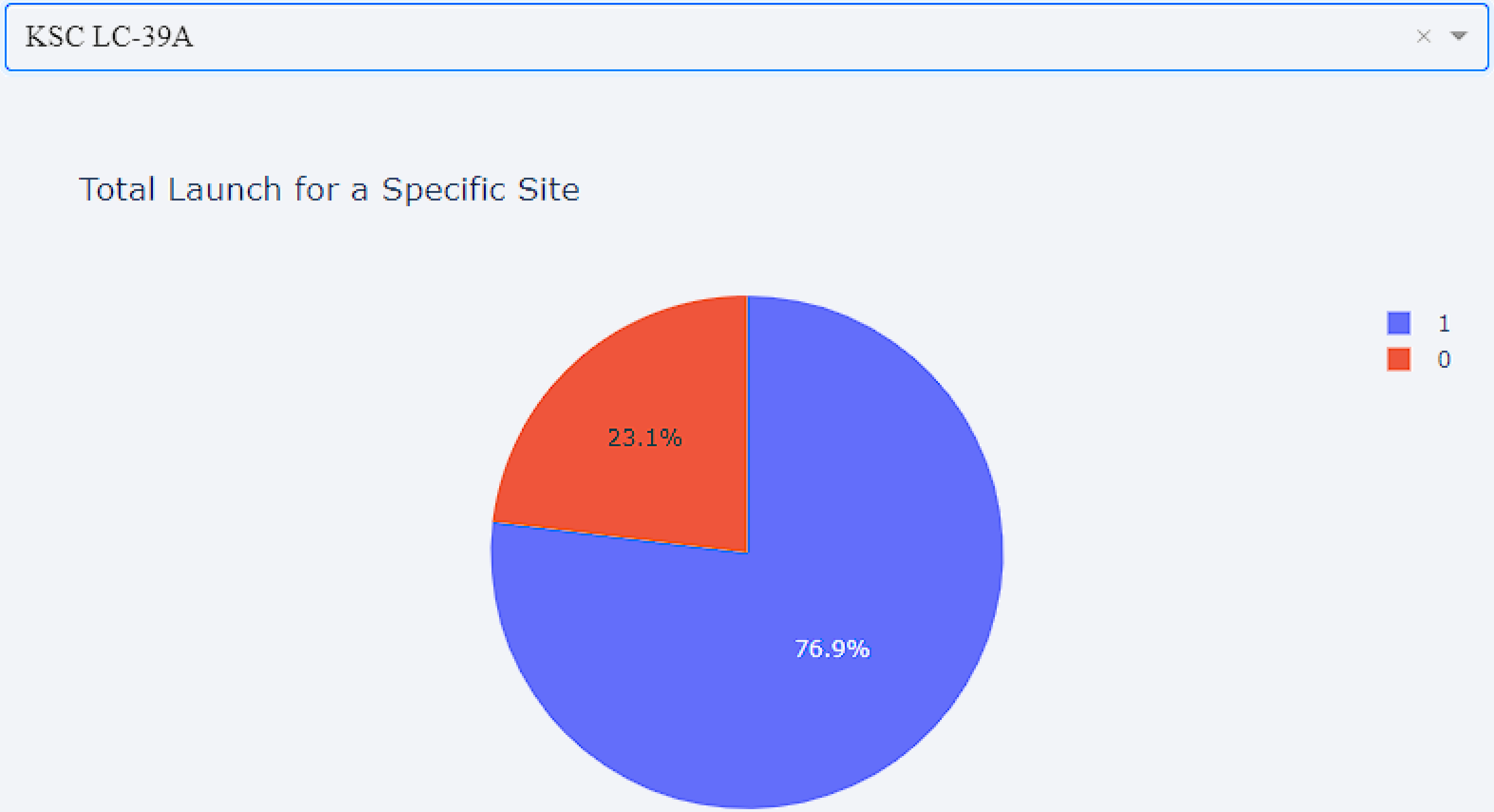
- The KSC LC 39A launch site occupies the largest launch share at 41.7%



ALL SITES × ▾

The launch site with highest launch success ratio

The launch site with the highest launch success rate corresponds to the KSC LC-39A launch site, with 76.9% successful launches.

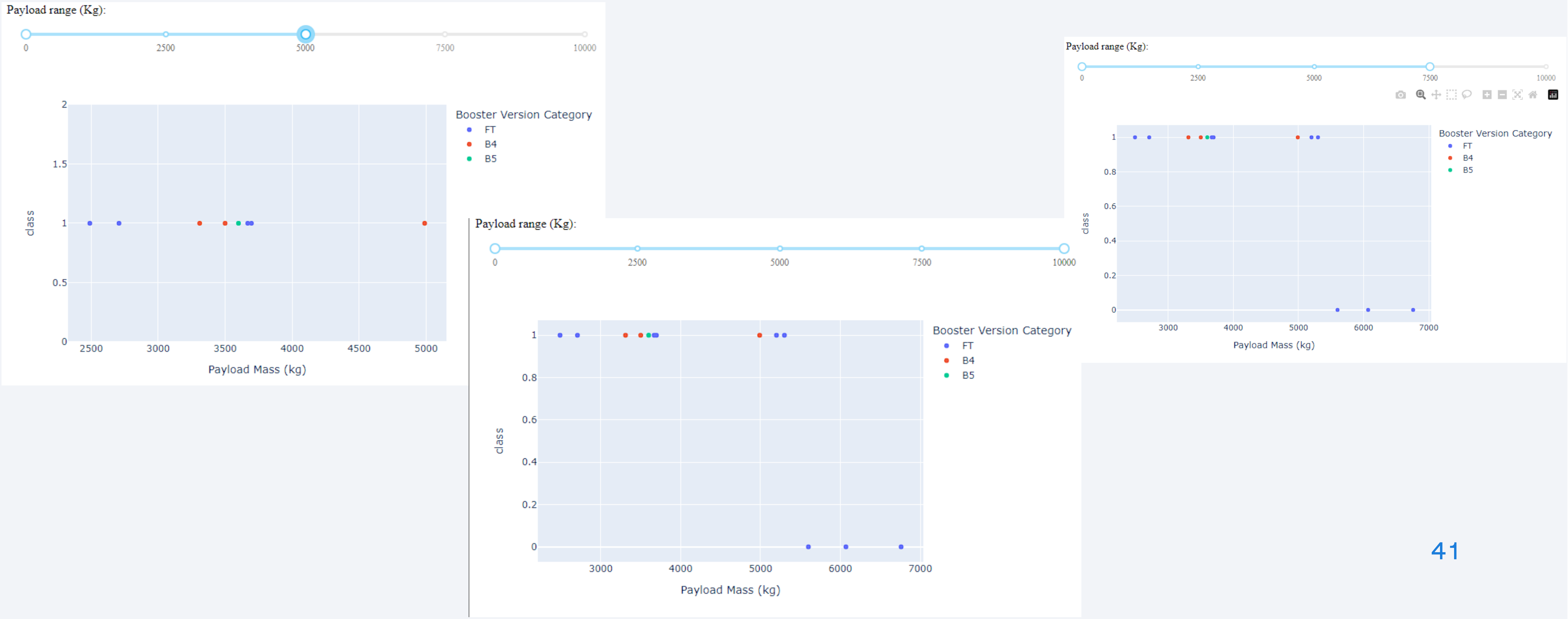


KSC LC-39A × ▾

Total Launch for a Specific Site

Payload vs. Launch Outcome with different payload

Payload vs. launch result with 5000 kg, 10000 kg and 7500 kg, respectively.



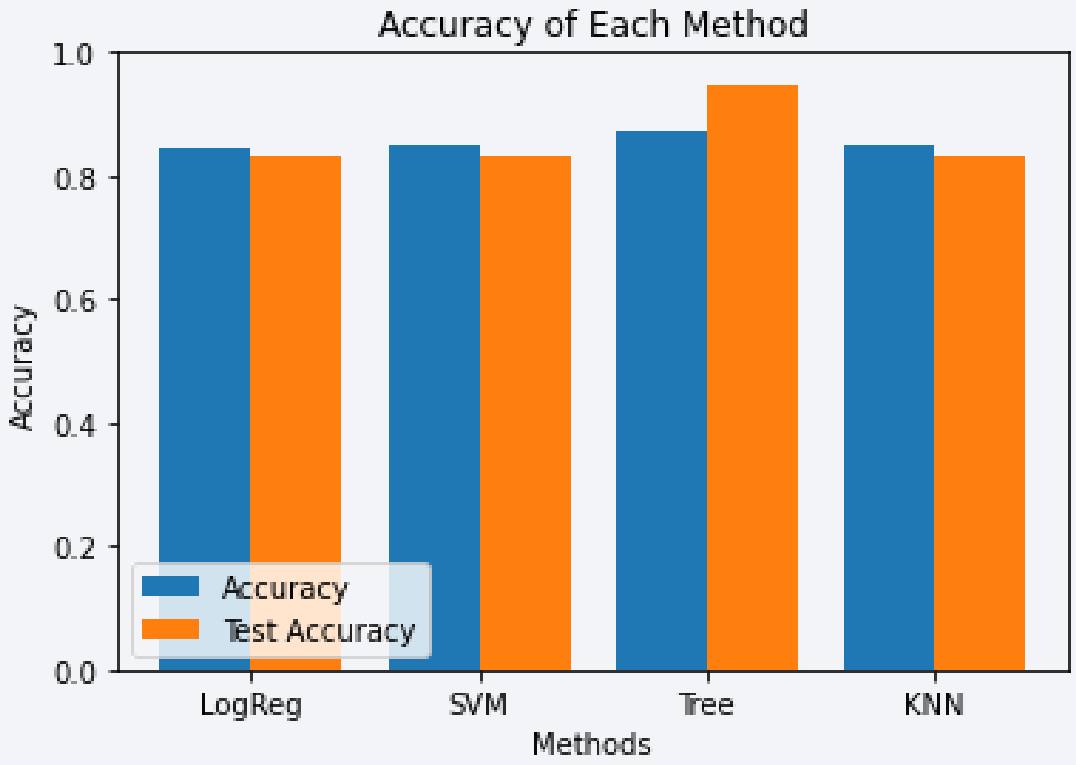
Section 5

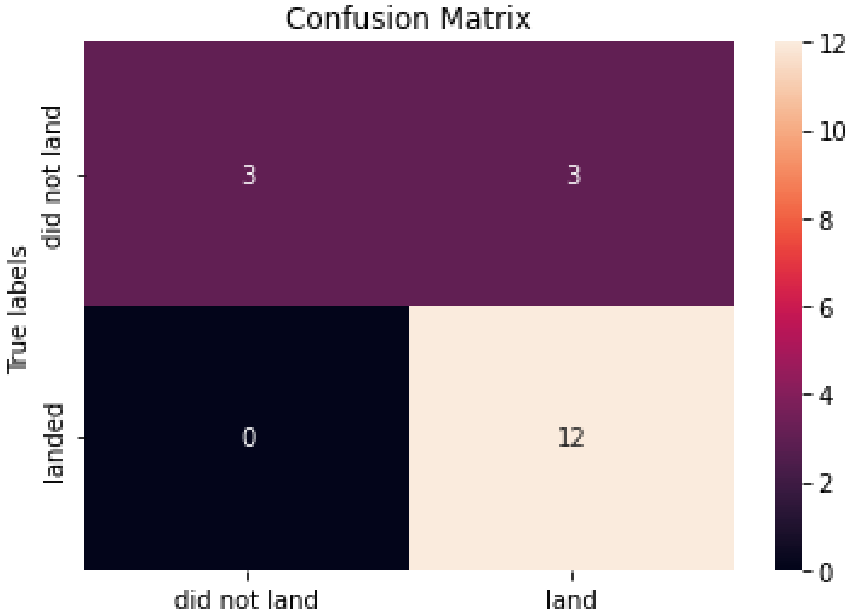
Predictive Analysis (Classification)

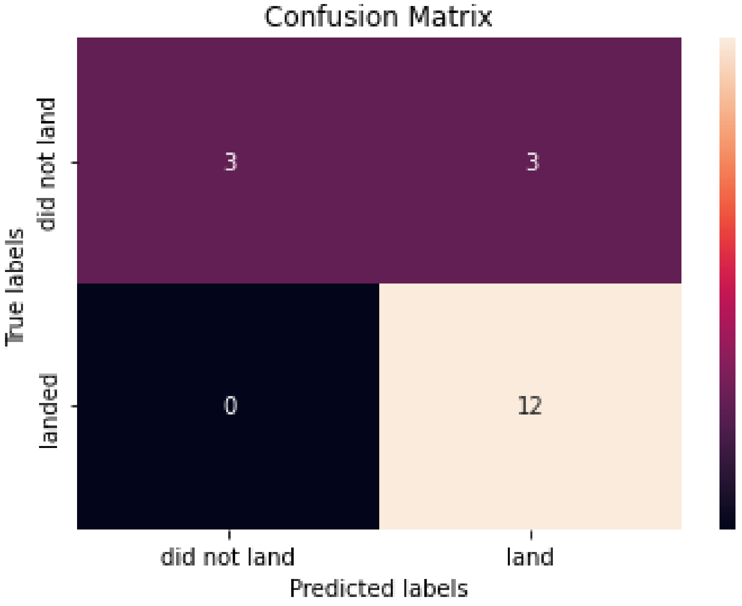
Classification Accuracy

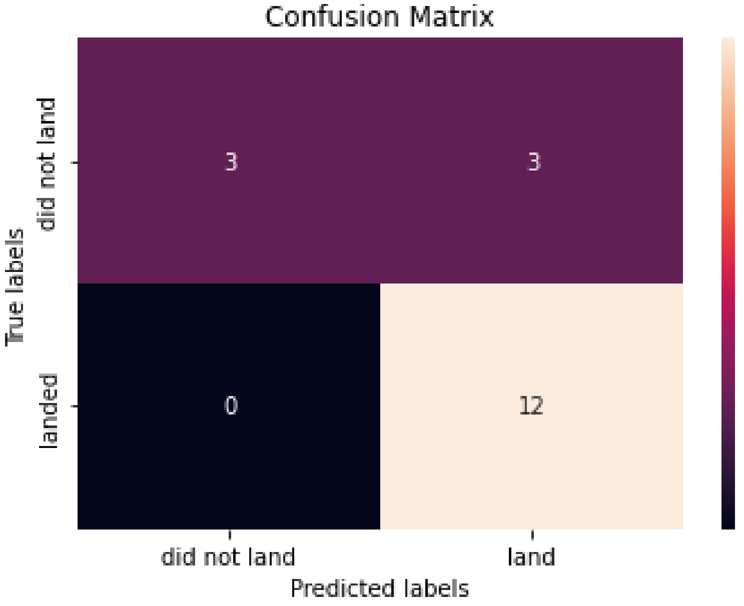
Accuracy of the constructed model for all the classification models used.

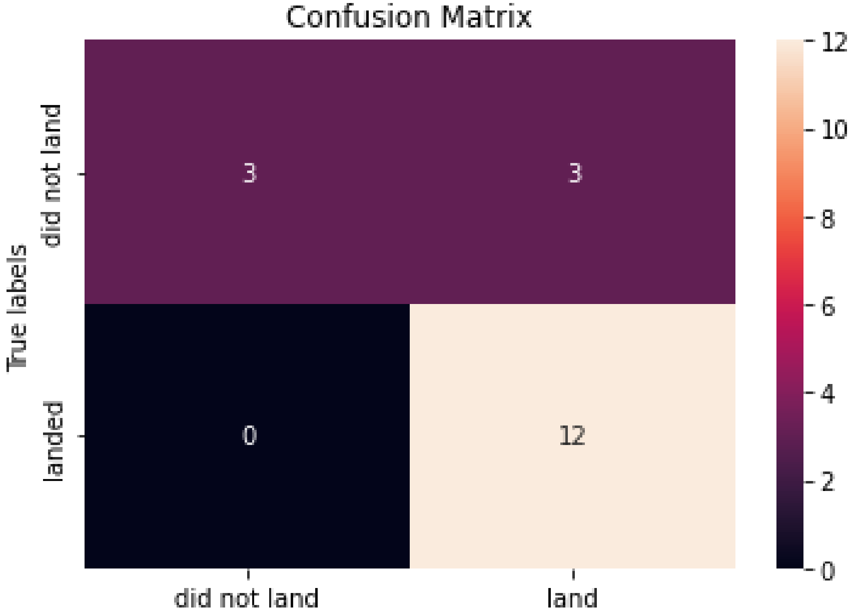
The highest accuracy is given by the decision tree model.











Confusion Matrix

Conclusions

- LogReg, SVM, Tree and KNN fit the model very well as they have a high accuracy level. However, the one that best fits is the decision tree.
- A notable relationship was found between low-weight payloads and launch success. Therefore, it is a determining factor for the object of study.
- Launch success was shown to have increased significantly over the years. Having the highest success rate in the ES-L1, GEO, HEO and SSO orbits.

Appendix

- https://github.com/Soka0/Applied-Data-Science-Capstone/tree/master

Thank you!