IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vedika Jain
12th July, 2023

Outline

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

Executive Summary

This project seeks to explore the probability of SpaceX reusing the first stage of its Falcon 9 rocket. The success of a rocket mission depends on multiple factors. Among the ones with which we were provided, like the landing pad, the number of grid fins, and the booster version, this exercise zeroes down on which ones were the most significant.

The first stage landing data was collected through a SpaceX API, and Falcon 9 launch data was scraped from Wikipedia, since information about the launches wasn’t provided by SpaceX. The data was then wrangled, and analyzed using SQL. This helped us find which variables affect success. Better understanding of their influence was obtained by visualising the launch sites on a map and plotting graphs between multiple variables. A dashboard was created to present the findings, and a classification model was trained to predict the success of future launches.

It was found that the proximity of the launch site to the equator increased the chances of success for a launch, and launches with heavier payloads were more likely to fail.

Introduction

SpaceX's Falcon 9 rocket is a game changer in terms of capital and resource optimisation in the private space exploration market. It reuses the first stage, which more than halves mission expenses, when compared to its competitors. Multi-billionaire Allon Mask wants to challenge Elon Musk's hegemony with his start-up, SpaceY. For a new entrant to the highly competitive sector, it would be imperative to gain knowledge of the savings compounded over years, and any overheads or caveats that such a practice might result in.

We stepped into the shoes of a data analyst at SpaceY to find out the efficiency of the retrieval of the first stage, and the number of times of reuse before it starts affecting operations, and the change in requirements in powering the rocket as it ages. We need to know if working with such a reusable rocket comes with specifications about mission conditions. This would provide us with insight about its use cases, and help us develop a better rocket. Our rocket would be more robust, more efficient, and more durable.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect landing data from SpaceX API
  - Web scrape launch data from Wikipedia page
- Perform data wrangling
  - Convert mission outcomes to a binary class
  - Replace missing payload mass values with mean
- 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

The first stage landing data was collected through a SpaceX API, and Falcon 9 launch data was scraped from Wikipedia, since information about the launches wasn’t provided by SpaceX.

Further details follow in the coming slides.

Data Collection – SpaceX API

A GET request was sent, the response JSON was normalised, and converted to a Pandas data frame. The actual values sourced from other API tables were plugged in the place of corresponding serial numbers in the main table. The reconstructed data frame was filtered to only contain Falcon 9 records.

The GitHub link for the notebook is provided in the Appendix.

Data Collection - Scraping

A GET request was sent, the response JSON was converted into a soup, the table tags were filtered and the required table obtained. The HTML table was carefully parsed through to avoid external references and extract the relevant data.

The GitHub link for the notebook is provided in the Appendix.

Data Wrangling

Missing data in the payloads column was substituted with the mean. The missing values in the landing pad column were left as they were, as they signified failures and landings which were not attempted. For successes and failures on different pads, a binary class was introduced.

The GitHub link for the notebook is provided in the Appendix.

EDA with Data Visualization

A variety of charts were plotted between different data variables such as payload, orbit type, flight number, to name a few, to better understand how different factors influenced each other and the success of a launch.

The GitHub URL is provided in the Appendix.

EDA with SQL

- Launch sites were identified.
- The type of launches at different sites were analysed.
- Payload for different customers and boosters were tallied.
- Different conditions for success were found.
- Success and failure counts for different periods and payloads were obtained.

The GitHub URL is provided in the Appendix.

Build an Interactive Map with Folium

The launch sites were visualised on a world map to see if their geography had any influence on mission success. Marker clusters were added to the sites to see their launch volume and success tally. Nearby landmarks were identified.

The GitHub URL is provided in the Appendix.

Build a Dashboard with Plotly Dash

A pie chart representing the success percentage for each site, and each site’s share in the total success count was created. A scatter chart was plotted for success of different payload missions using various boosters, for all sites cumulatively, and each site individually. This helped us know how the launch site, payload, and booster affected success.

The GitHub URL is provided in the Appendix.

Predictive Analysis (Classification)

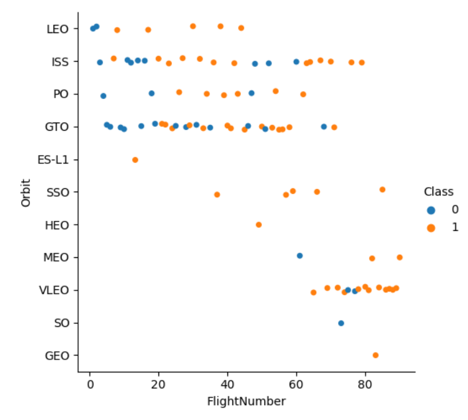
To find the best classification model, we used different classification techniques and derived optimum models based on them using grid search. The models were trained on the same training set, and their performance evaluated on the same testing set. The training and testing sets were obtained by an 80:20 split of the data. Confusion matrices were charted to find the models' shortcomings. The model with the maximum score was adopted.

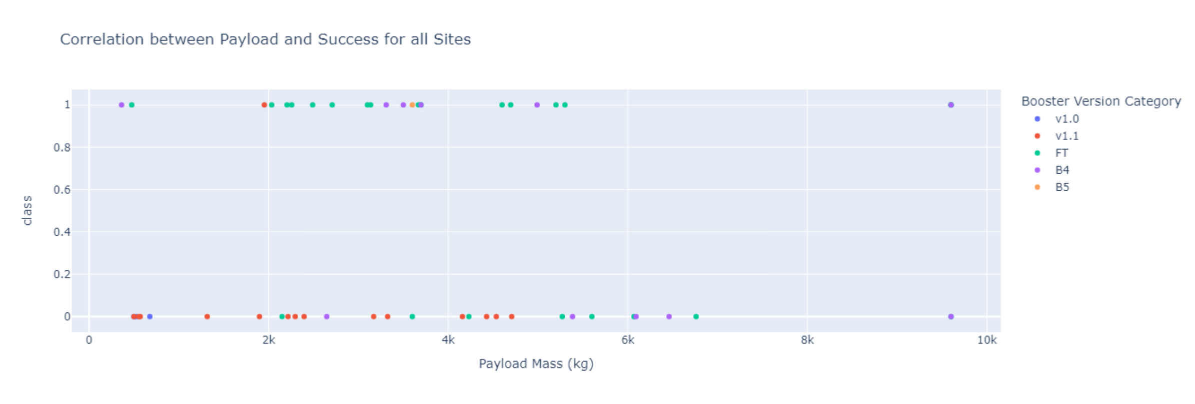
The GitHub URL is provided in the Appendix.

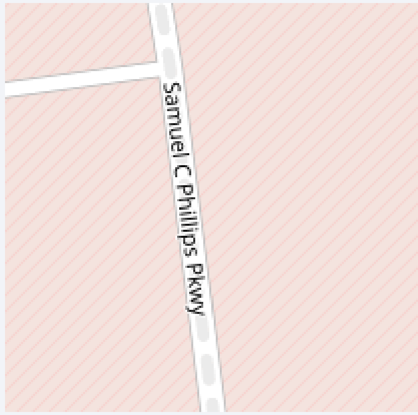
Results

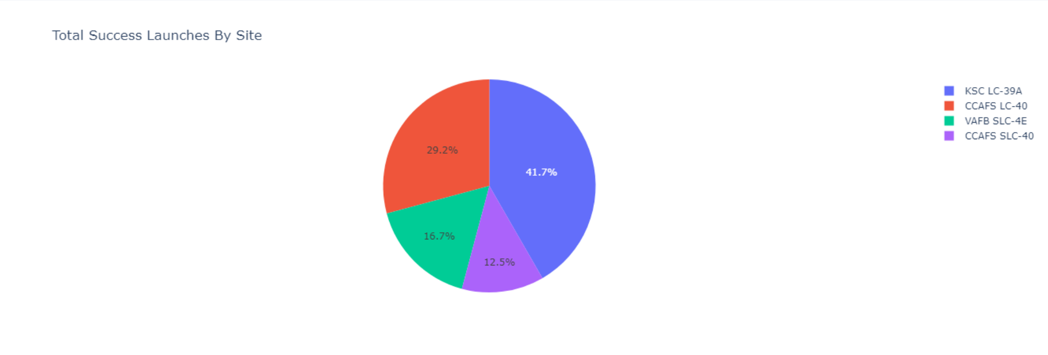
- The biggest factors influencing mission success were launch site, payload mass, booster, and orbit type.
- The decision tree technique yields us the best classification model for this problem.

Samples







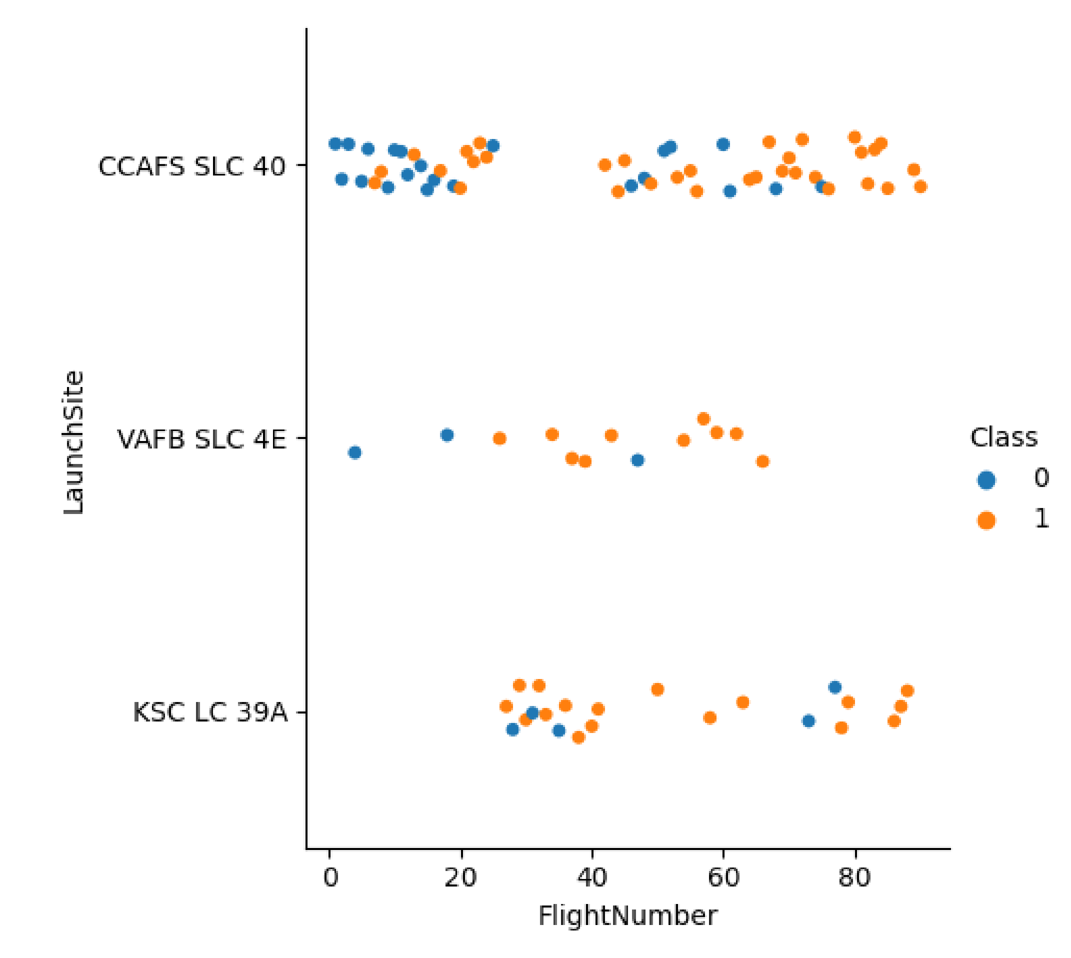


Section 2

Insights drawn from EDA

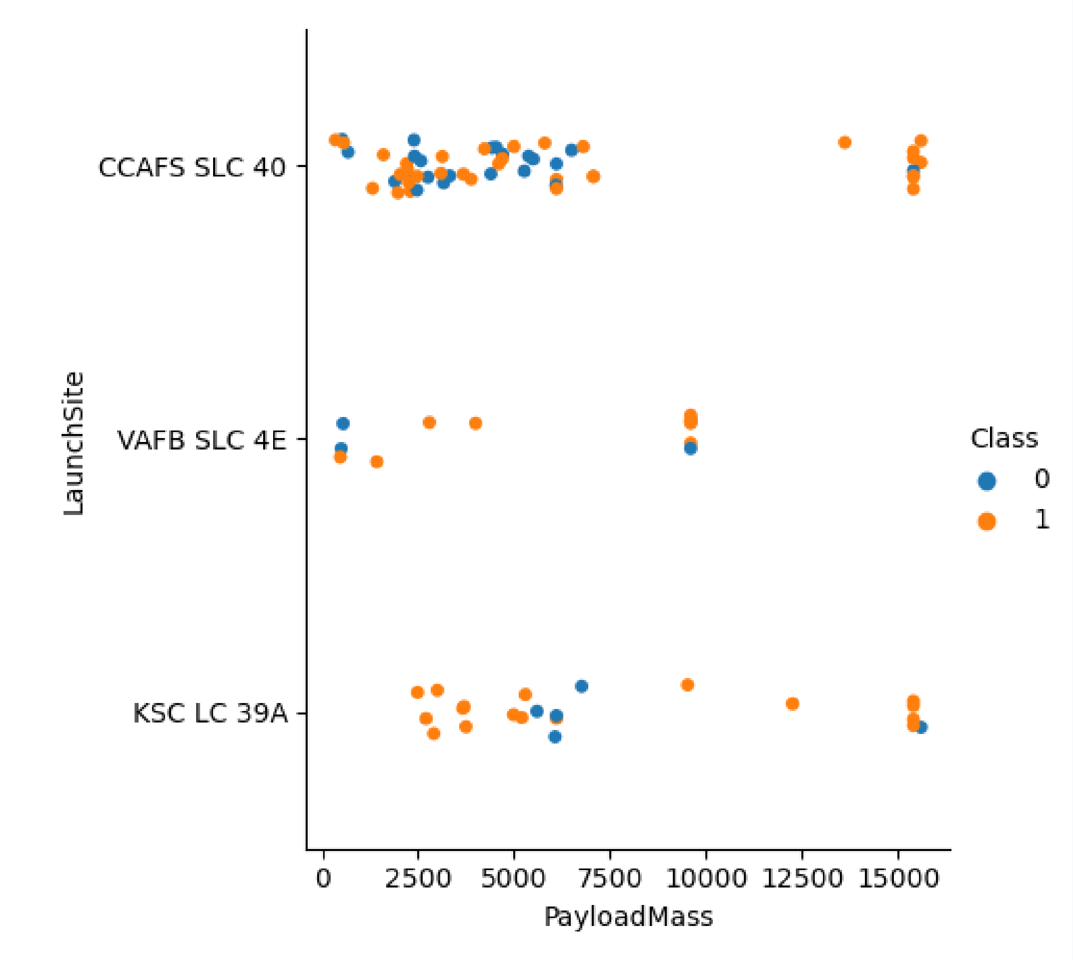
Flight Number vs. Launch Site

It can be seen that CCAFS SLC 40 witnessed successful launches for later flights, but early flights remained largely unsuccessful. VAFB SLC 4E launches have been mostly successful, and this site wasn't used for later attempts. KSC LC 39A has mixed results, and wasn't used for early flight launches.



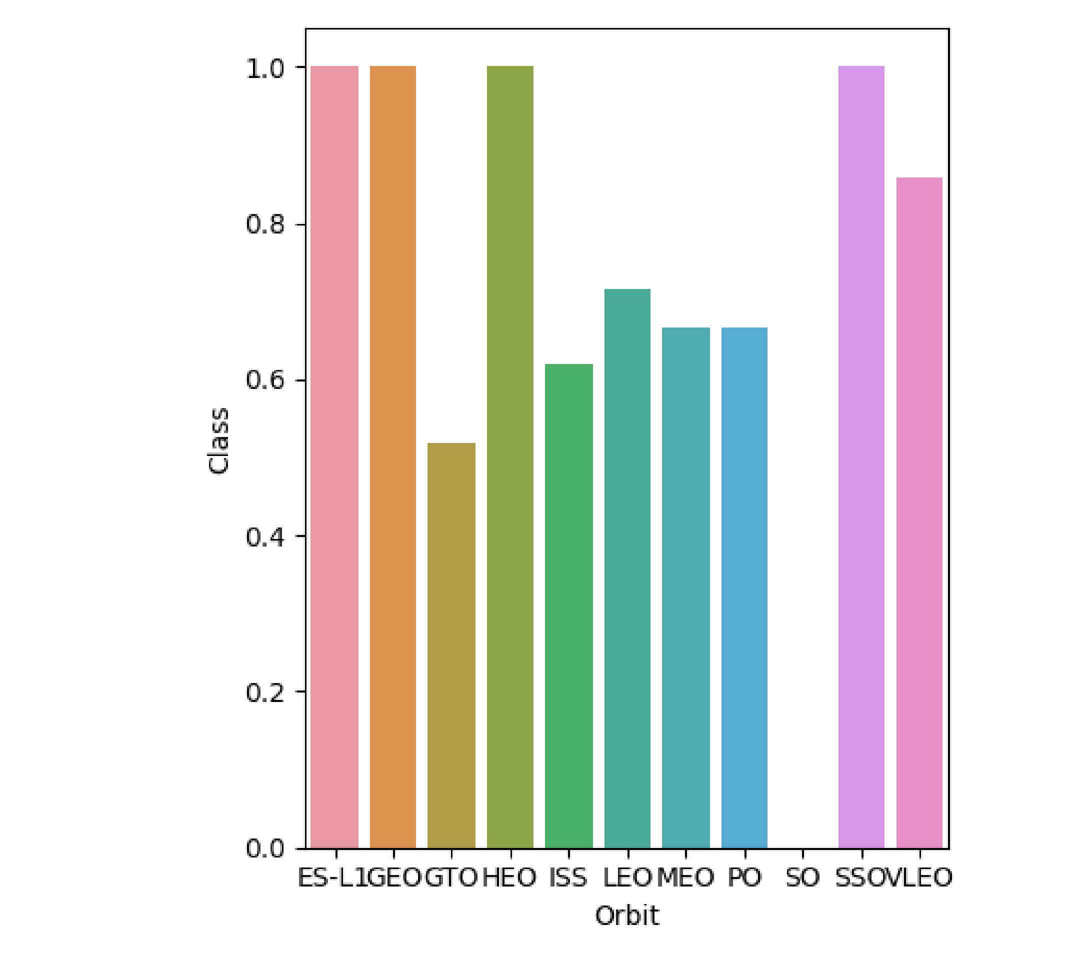
Payload vs. Launch Site

CCAFS SLC 40 had mixed results for low payloads but had good success rates at high payload. VAFB SLC 4E wasn’t used for high payloads, and KSC LC 39A had a good success rate for payloads below 5000 kg.



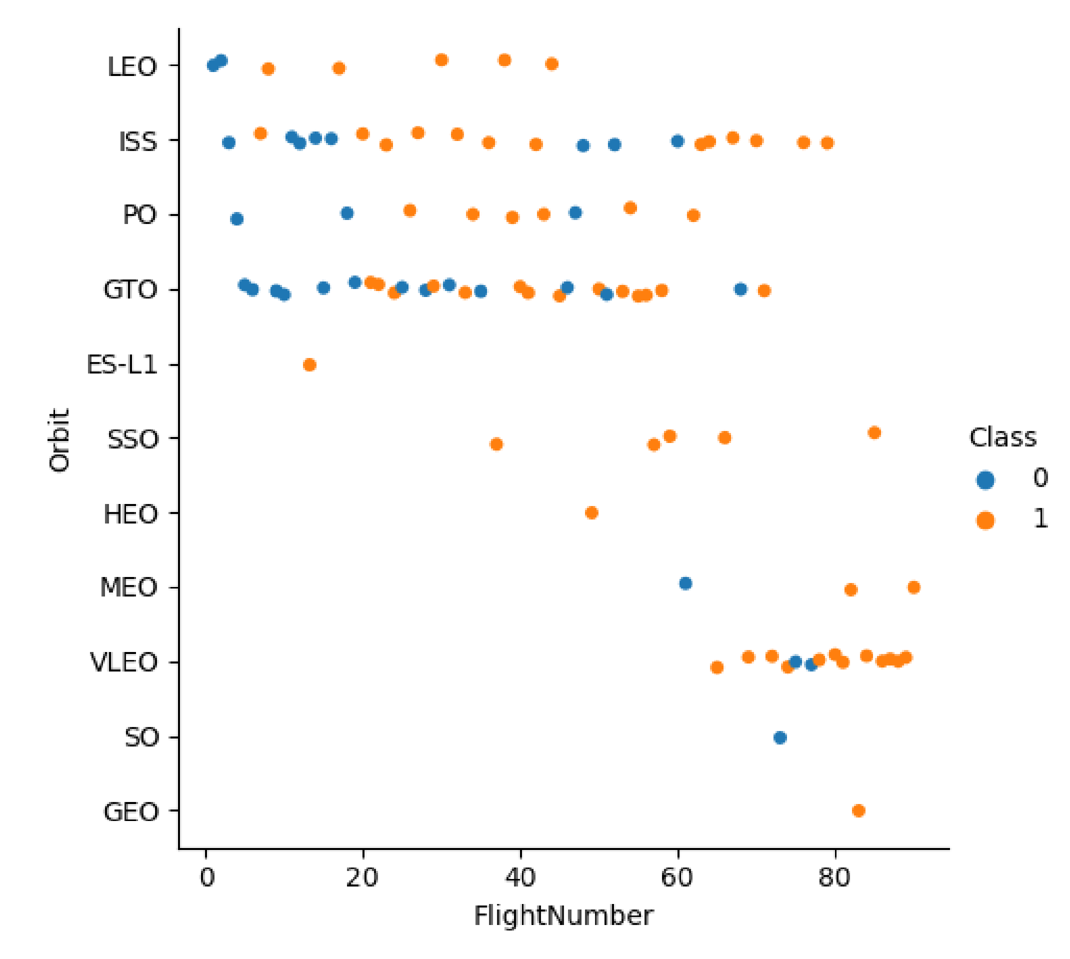
Success Rate vs. Orbit Type

There has been no successful launch in the SO orbit. Success rates are impressive for ES-L1, GEO, HEO, and SSO orbits. The GTO orbit has had equal number of successes and failures.



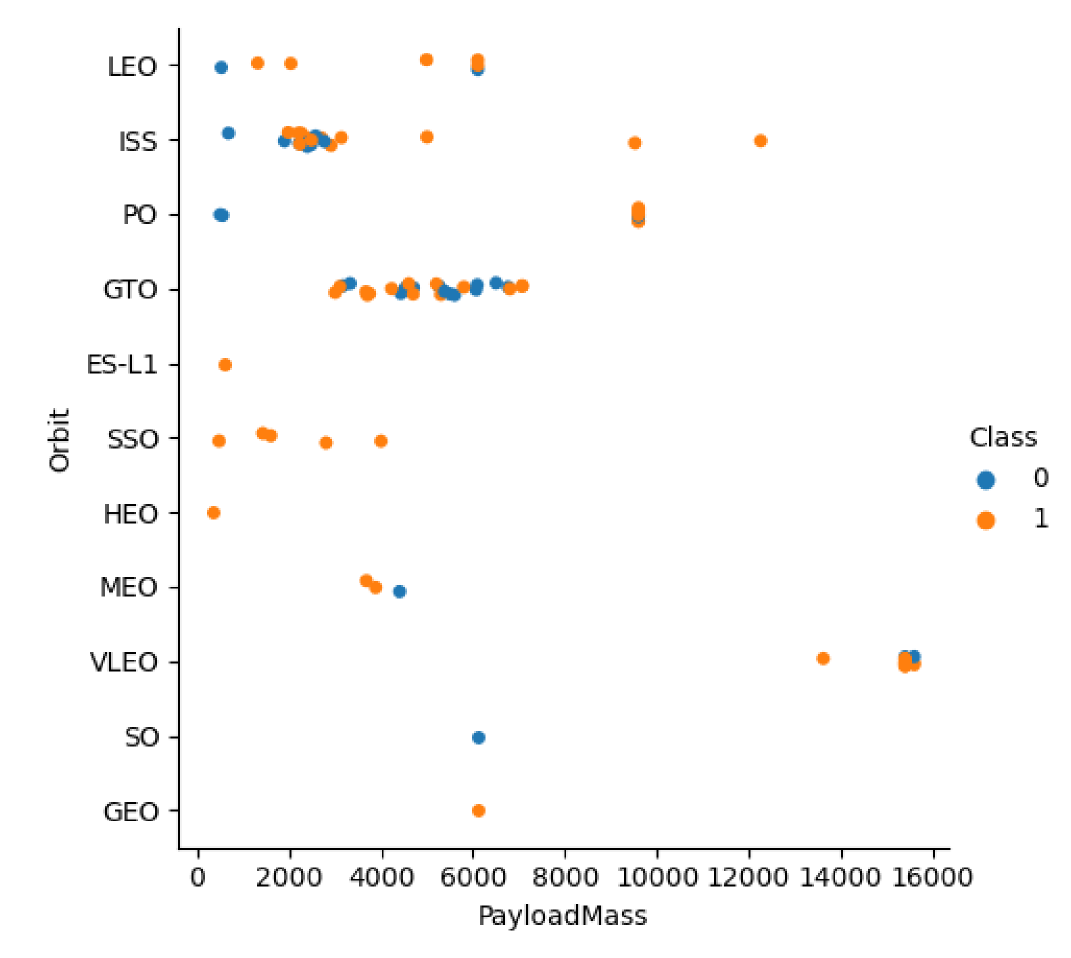
Flight Number vs. Orbit Type

Earlier flights in the GTO orbit are unsuccessful, and the results remain fairly mixed afterwards too. Later launches in the ISS orbit have been successful. VLEO has only witnessed later stage flights and they have been successful. Most of the flights have been to the LEO, ISS, PO, and GTO orbits.



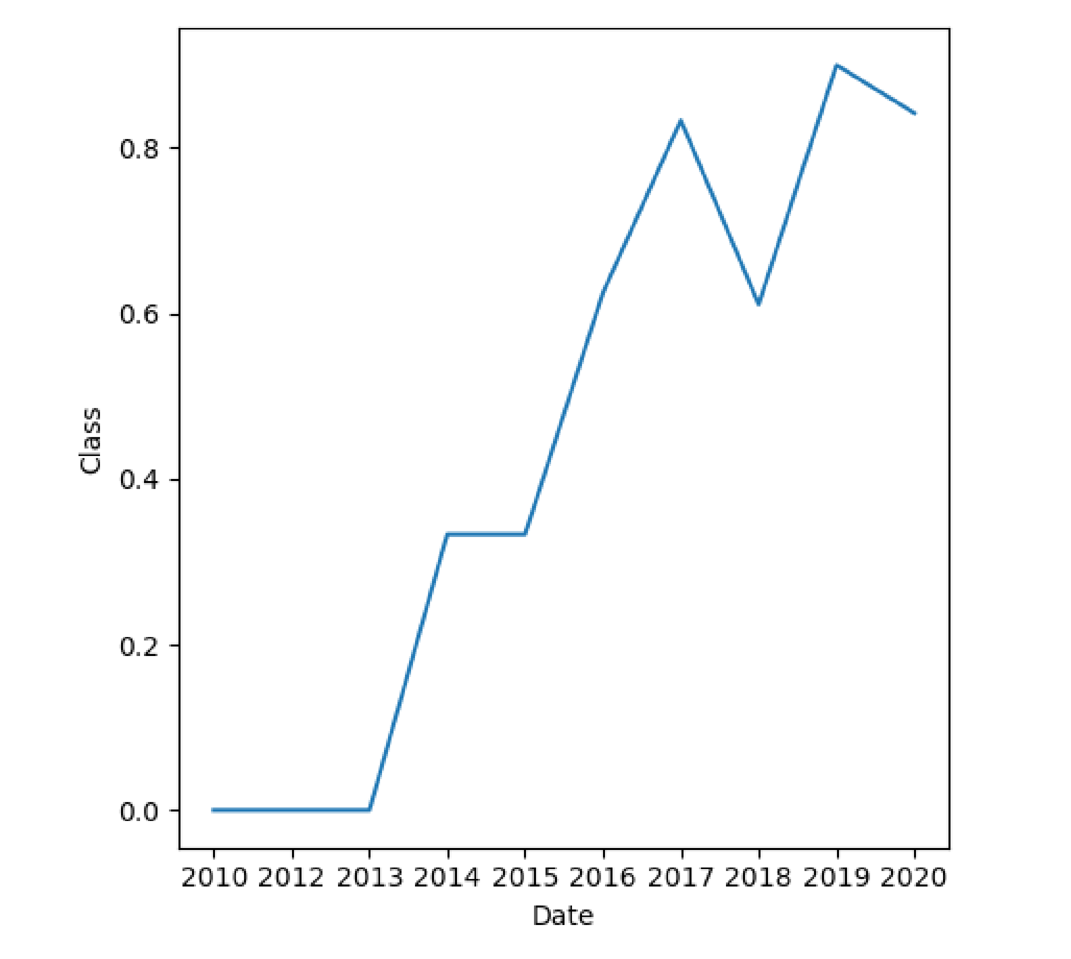
Payload vs. Orbit Type

VLEO has only had high payload launches. SSO had low payload missions, all of which have been successful. GTO has had mixed results. ISS had variety of payloads.

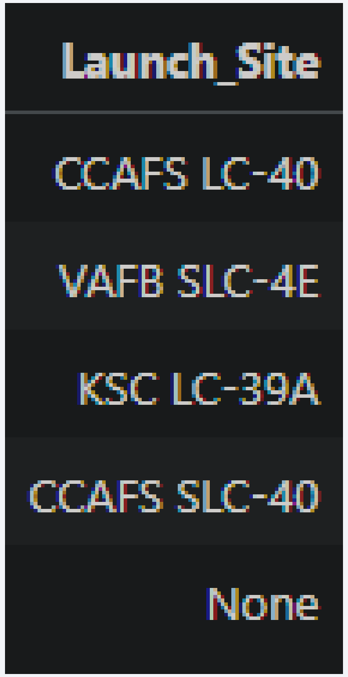


Launch Success Yearly Trend

There was a sharp increase in the success rate after 2013. The success rate stayed at 35% from 2014 to 2015. 2018 saw some failures, but the next year saw the highest success rate of all time.



All Launch Site Names

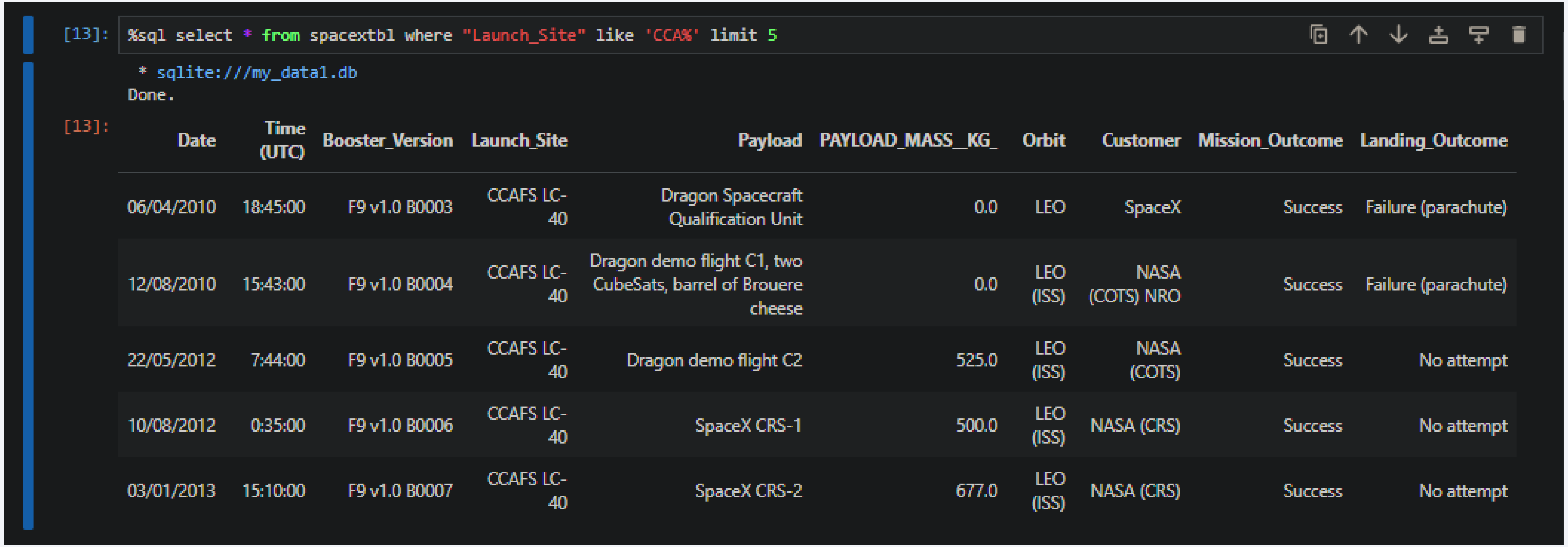


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

```
%sql select distinct "Launch_Site" from spacextbl
```

CCAFS LC-40 is code for the Cape Canaveral Space Launch Complex 40. Vandenberg Air Force Base Space Launch Complex 4E has the code VAFB SLC-4E. KSC LC-39A stands for Kennedy Space Center Launch Complex 39A.

Launch Site Names Begin with 'CCA'



```
[13]: %sql select * from spacextbl where "Launch_Site" like 'CCA%' limit 5
```

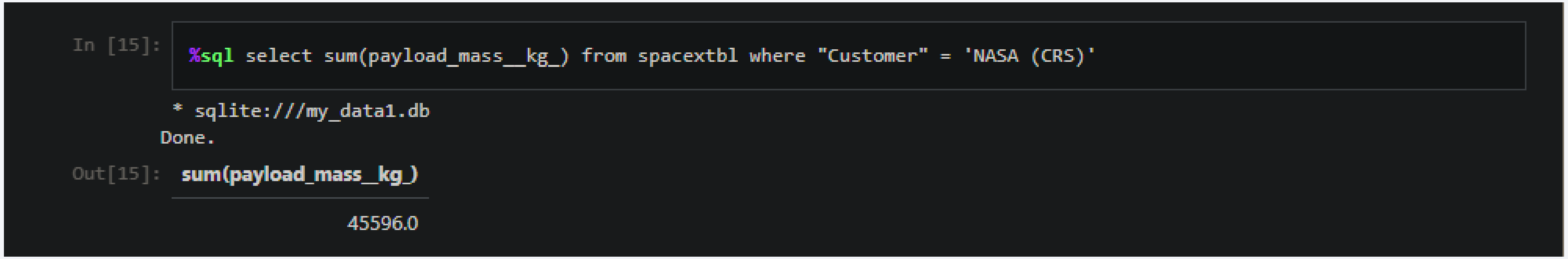
```
 * sqlite:///my_data1.db
Done.
```

[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Maximum launches have been from the CCAFS LC-40 site, five of them being presented here.

Total Payload Mass



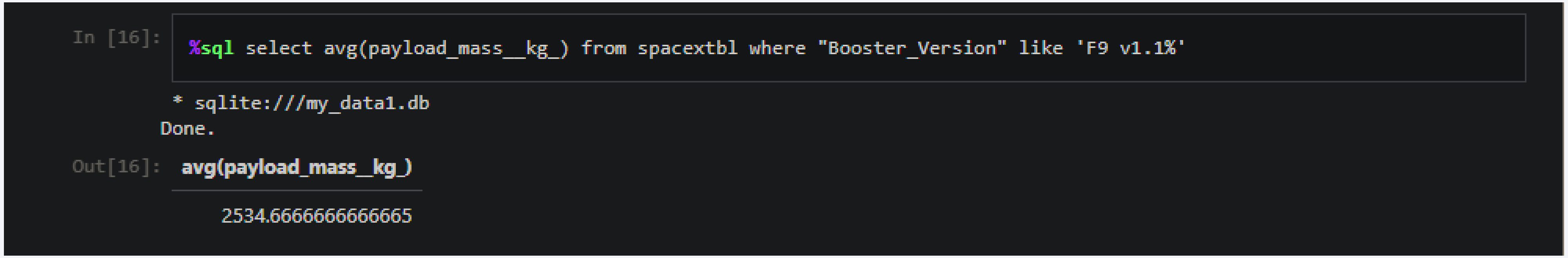
```
In [15]:  %sql select sum(payload_mass__kg_) from spacextbl where "Customer" = 'NASA (CRS)'

           * sqlite:///my_data1.db
          Done.

Out[15]:  sum(payload_mass__kg_)
                         45596.0
```

NASA has been a major customer of SpaceX, with a sizeable payload entrusted to be launched by the private entity.

Average Payload Mass by F9 v1.1



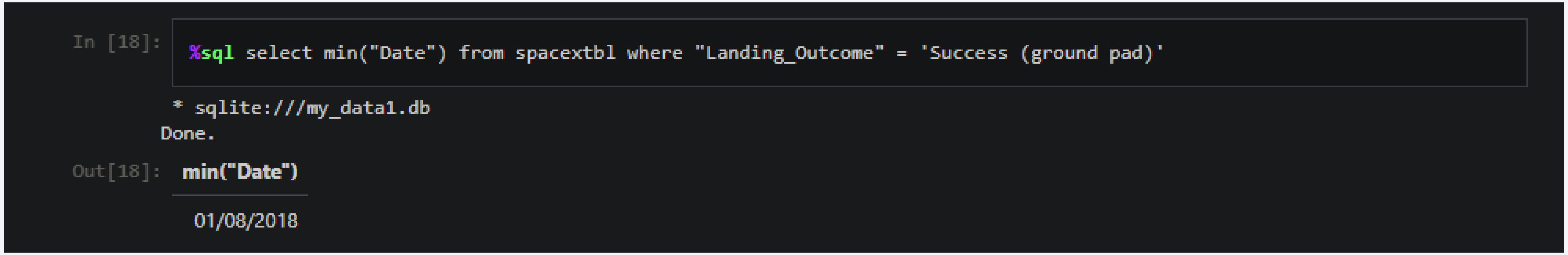
```
In [16]:  %sql select avg(payload_mass__kg_) from spacextbl where "Booster_Version" like 'F9 v1.1%'

 * sqlite:///my_data1.db
Done.

Out[16]:  avg(payload_mass__kg_)
                   2534.6666666666665
```

The F9 v1.1 booster is a lightweight carrier, carrying no more than 5000 kg of payload on average.

First Successful Ground Landing Date



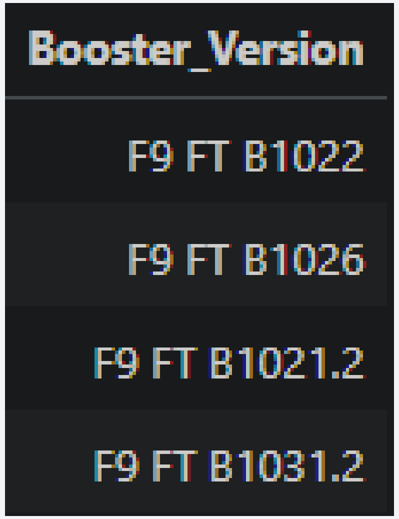
```
In [18]: %sql select min("Date") from spacextbl where "Landing_Outcome" = 'Success (ground pad)'

          * sqlite:///my_data1.db
         Done.

Out[18]: min("Date")
          01/08/2018
```

After all the years of trial, the first successful ground pad landing came in 2018.

Successful Drone Ship Landing with Payload between 4000 and 6000

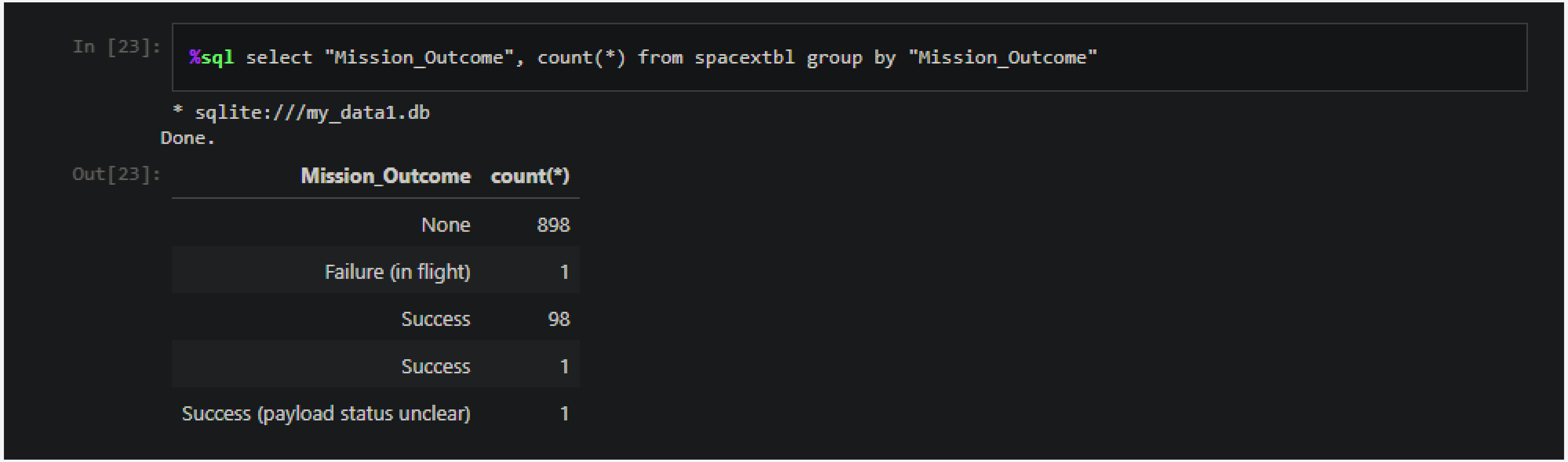


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

```
%sql select distinct "Booster_Version" from spacextbl where
"Landing_Outcome" = 'Success (drone ship)' and
payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

The F9 FT B series of boosters have shown a good track record in terms of drone ship landing.

Total Number of Successful and Failure Mission Outcomes



In [23]:

```
%sql select "Mission_Outcome", count(*) from spacextbl group by "Mission_Outcome"
```

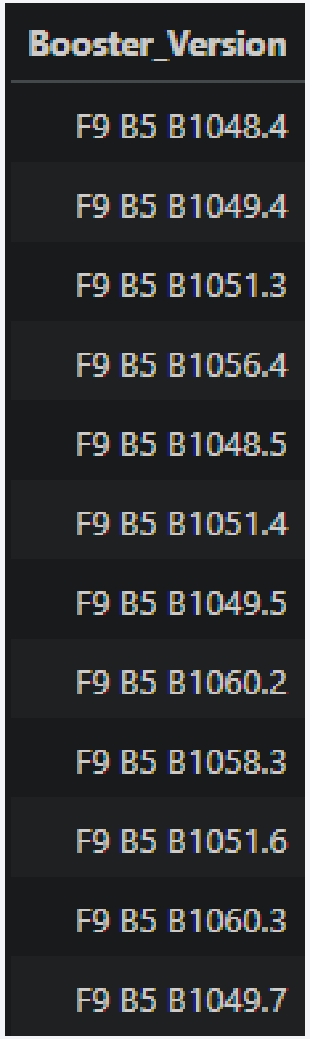
```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

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

The total number of successes is dwarfed by the number of non-successful launches, one of them also being an in-flight failure.

Boosters Carried Maximum Payload



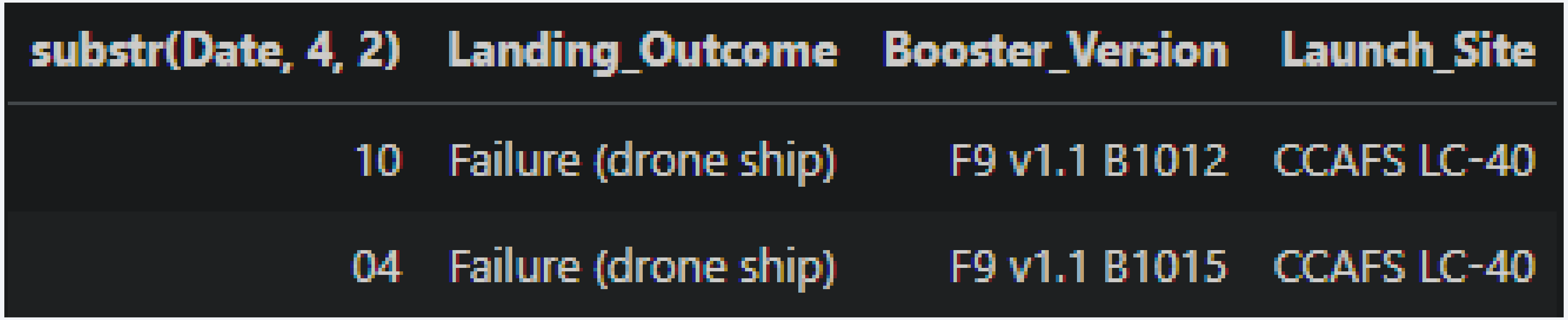
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

%sql select distinct "Booster_Version" from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)

The F9 B5 B series boosters are the beasts powering Falcon 9 heavyweight missions.

2015 Launch Records

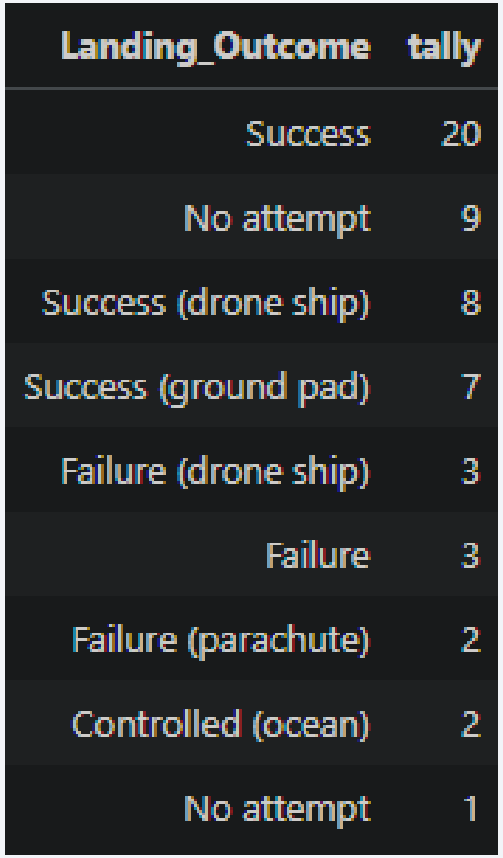
%sql select substr(Date, 4, 2), "Landing_Outcome", "Booster_Version", "Launch_Site" from spacextbl where "Landing_Outcome" = 'Failure (drone ship)' and substr(Date,7,4)='2015‘



substr(Date, 4, 2)	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The drone ship failures occurred with the F9 v1.1 B series boosters.

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



Landing_Outcome	tally
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

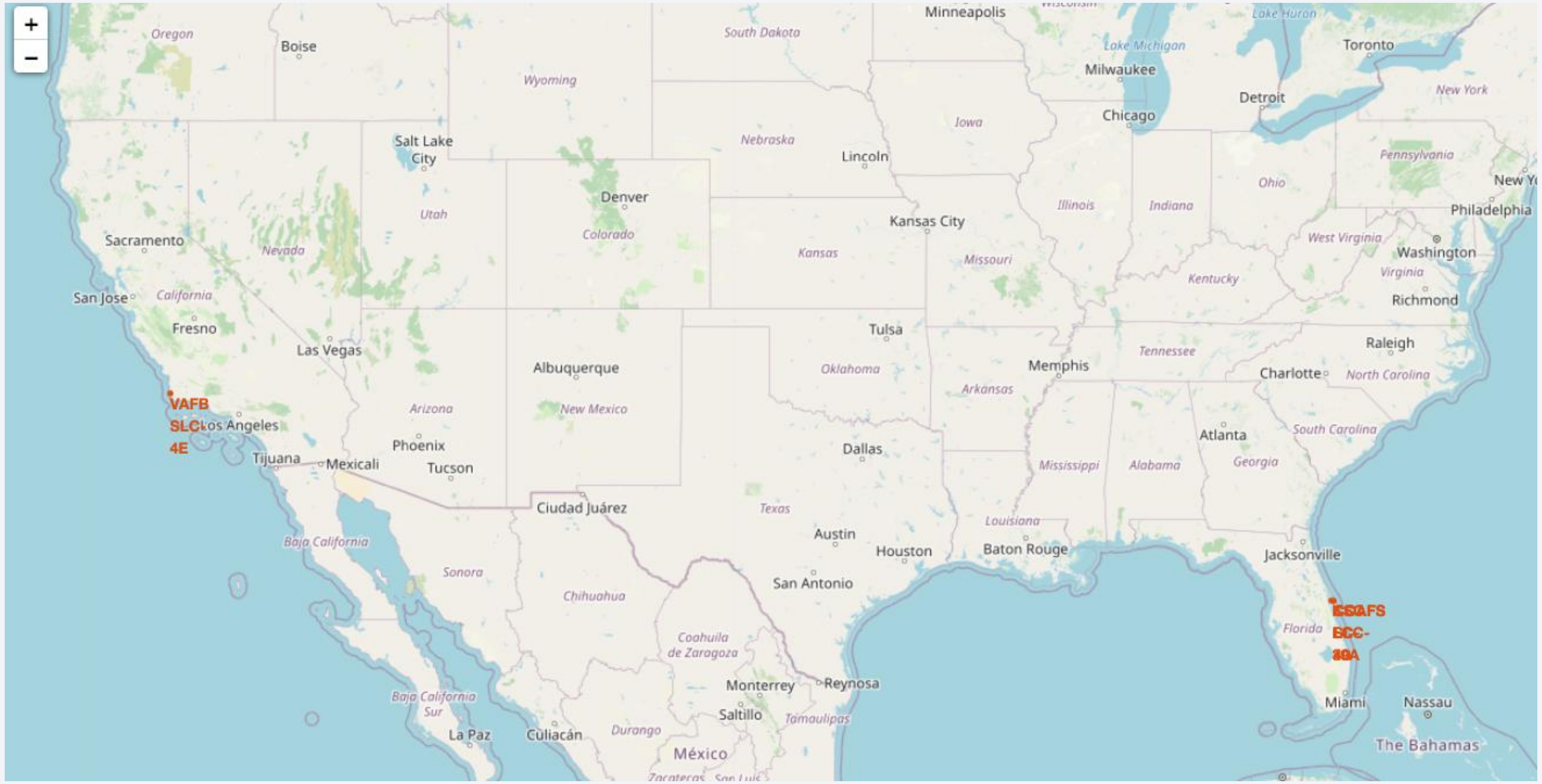
```
%sql select "Landing_Outcome", count("Landing_Outcome") as 'tally' from spacextbl where "Date" between '04/06/2010' and '20/03/2017' group by "Landing_Outcome" order by tally desc
```

Most of the missions have ended successfully. Apart from the failures, some launches were also abandoned midway.

Section 3

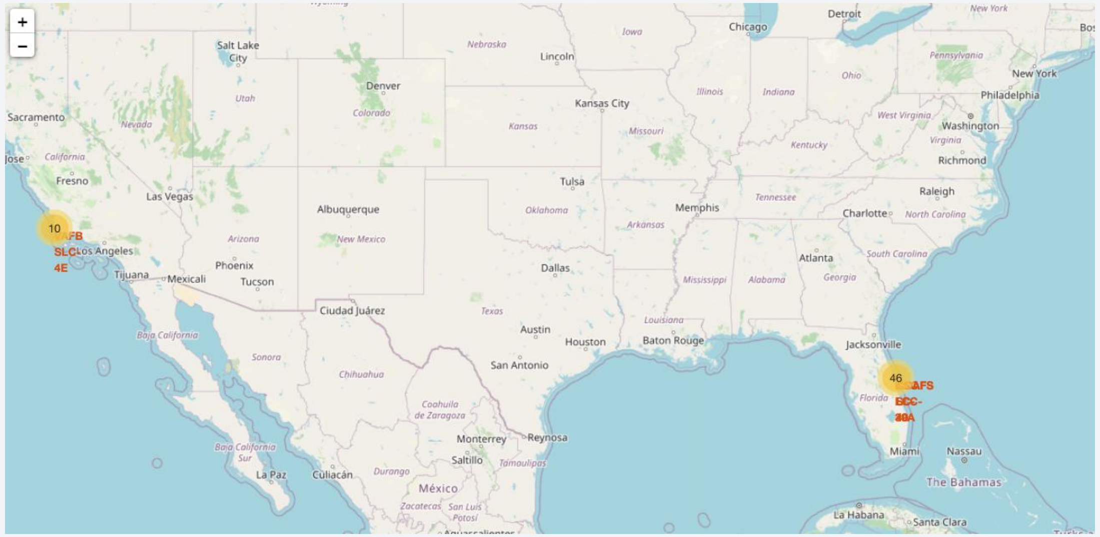
Launch Sites Proximities Analysis

Launch Sites

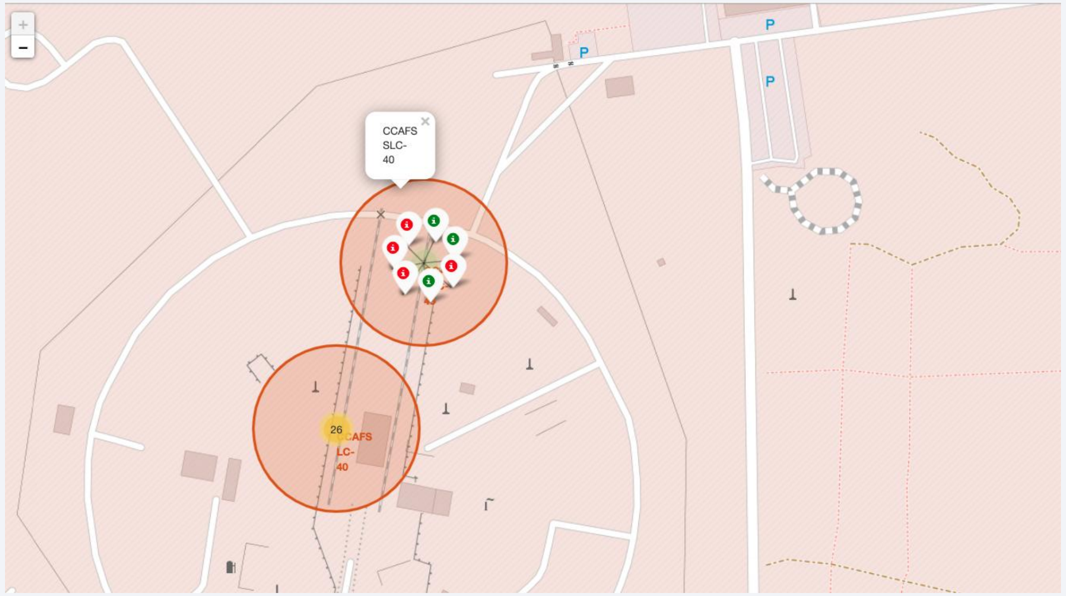


Three of the launch sites are in close proximity to each other. All are coastal sites.

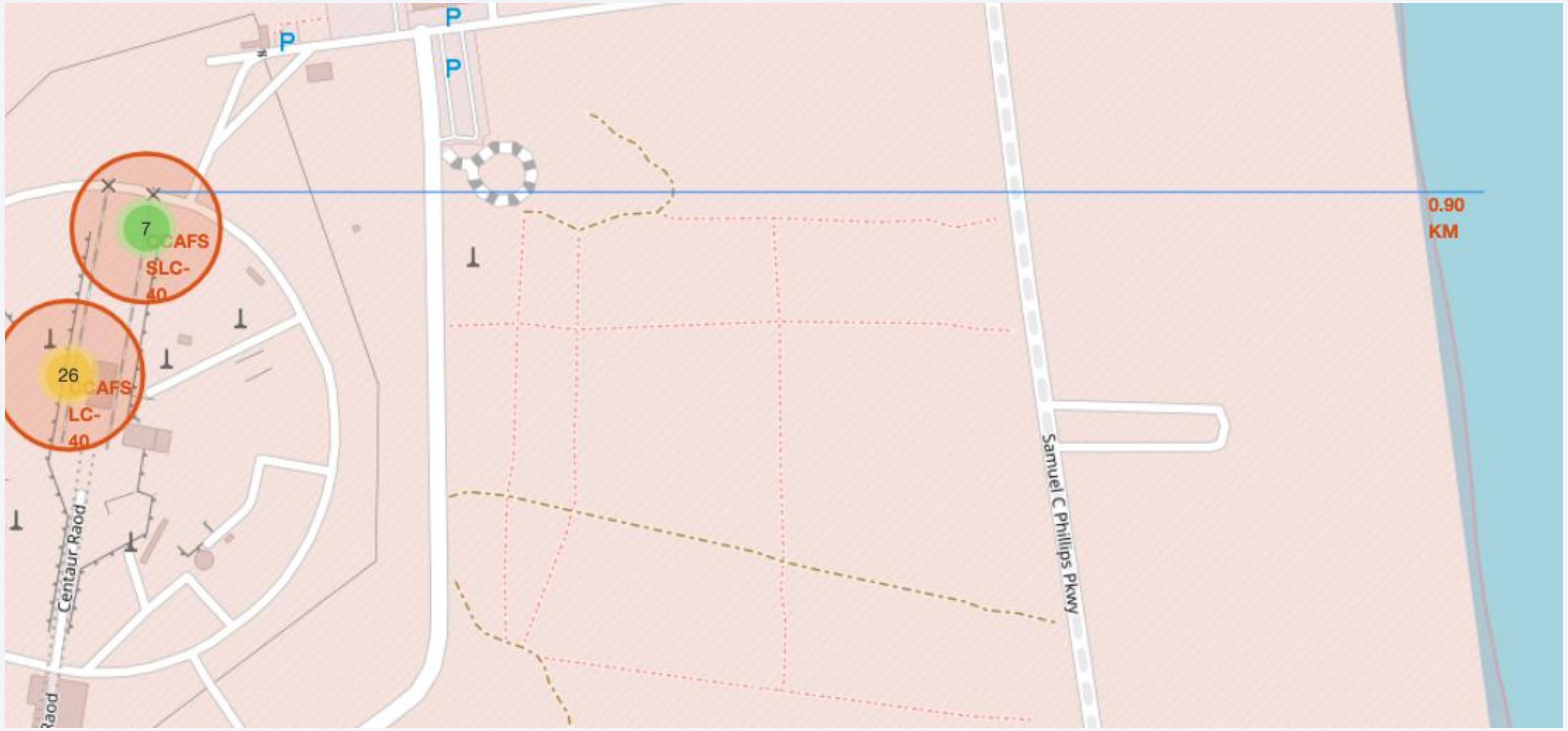
Launch Outcomes



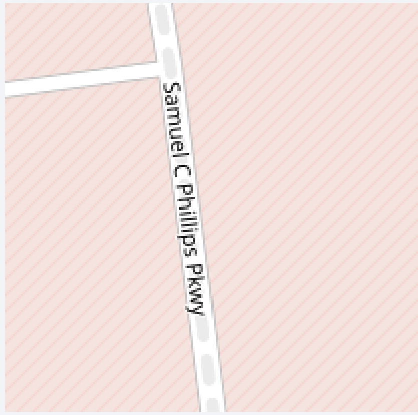
CCAFS SLC-40 has had mixed, if not largely unsuccessful, results. The number of launches are proportionally distributed among the west and east coast clusters.



Launch Site Proximities

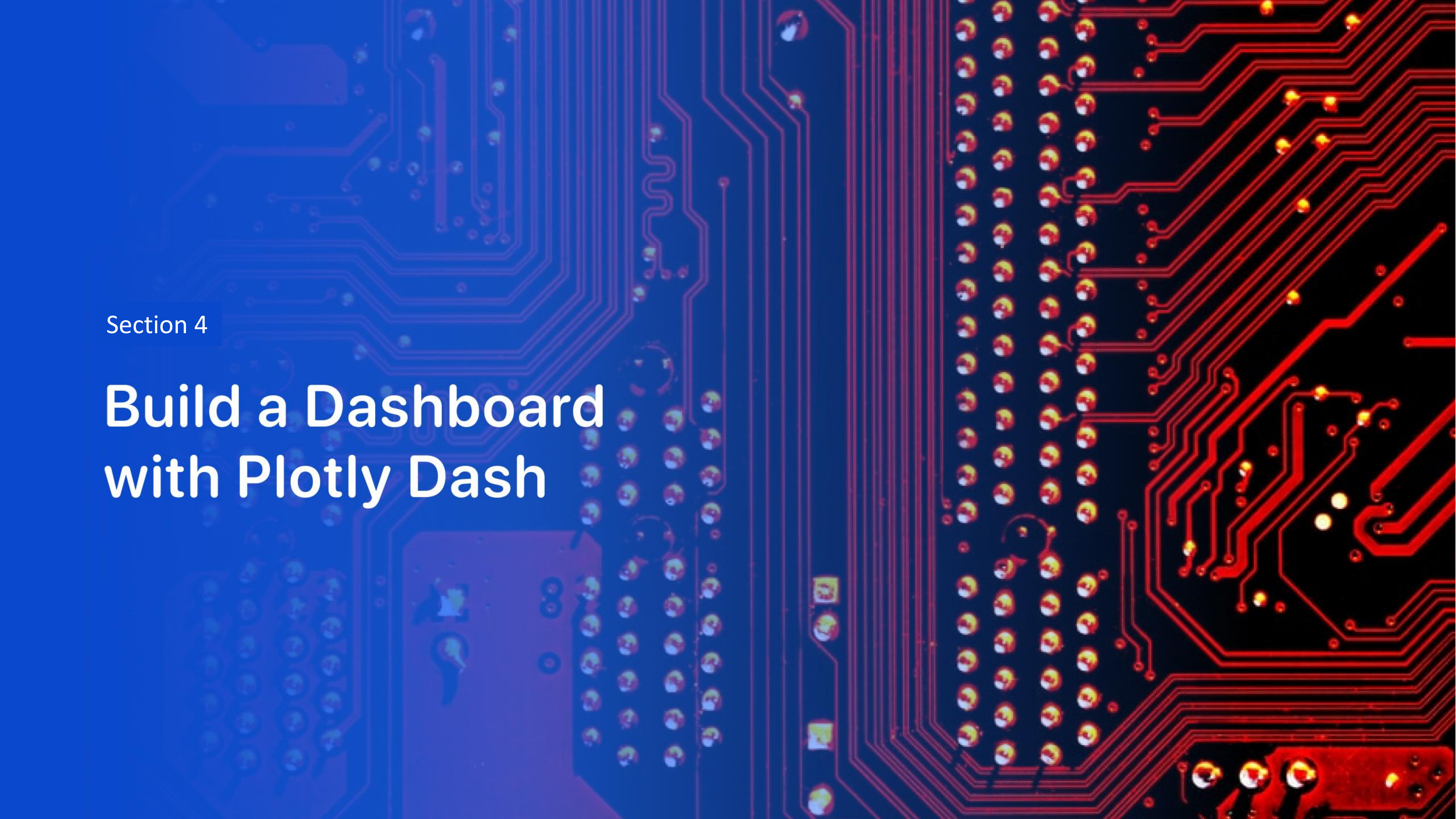


Launch sites are less than a kilometre away from the coast. They are well connected to neighbouring areas through roads and railways.

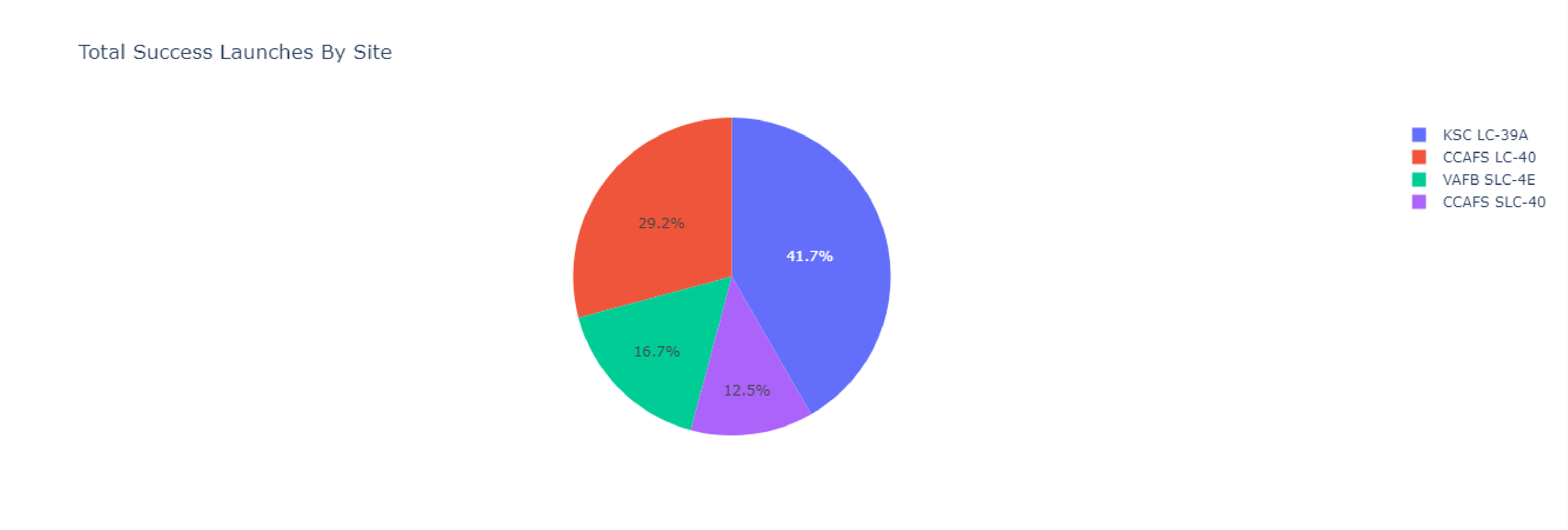


Section 4

Build a Dashboard with Plotly Dash

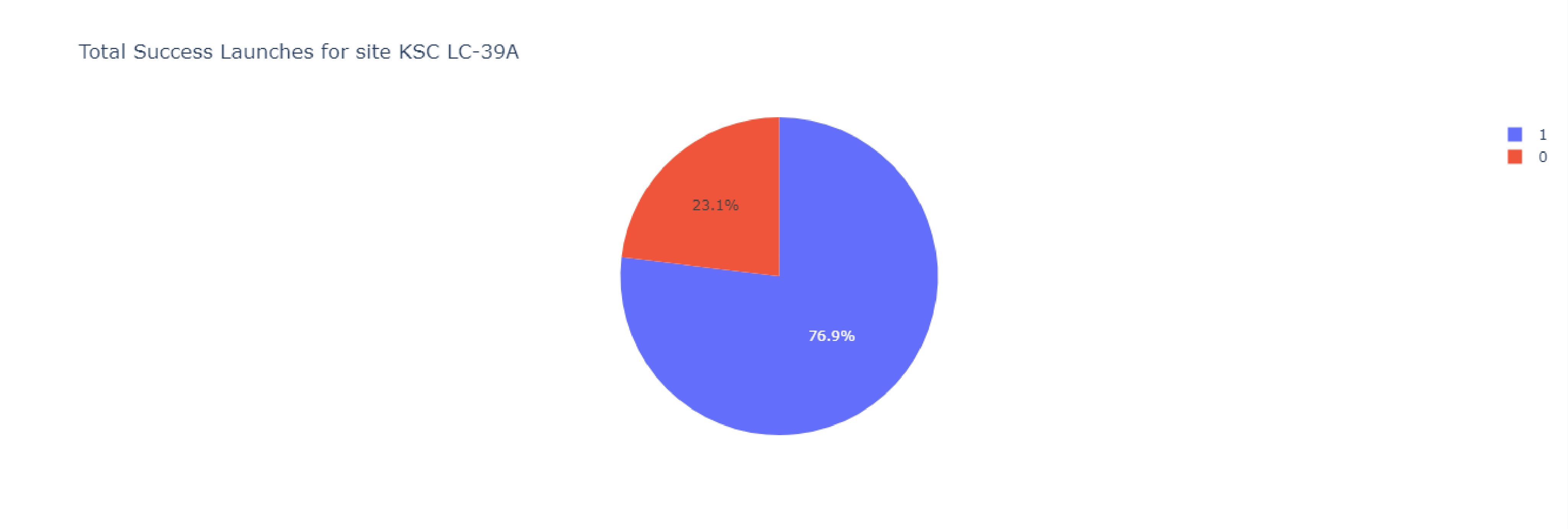


Launch Success Count for all Sites



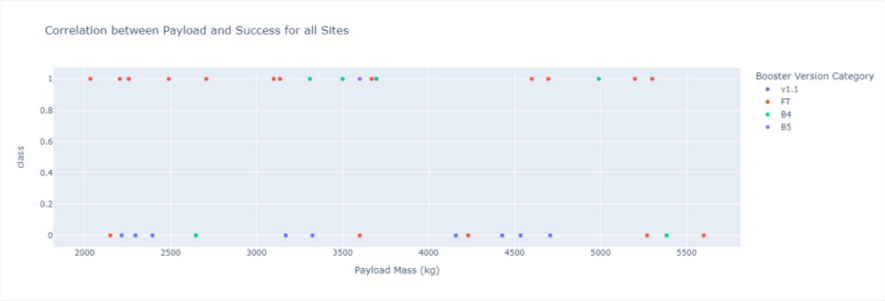
KSC LC-39A has the maximum contribution to successful launches, followed by the CCAFS LC-40, in spite of the latter's launch volume.

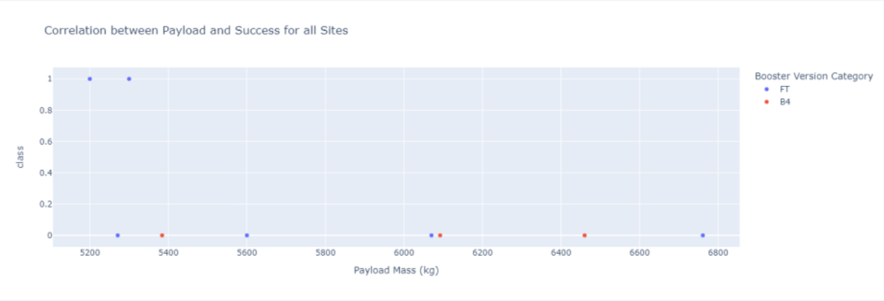
Highest Launch Success Ratio



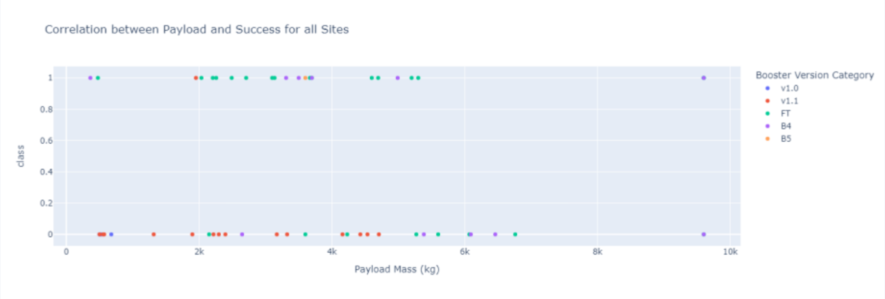
The site with the most successful launches also has an impressive launch success ratio, best of the four sites.

Payload vs. Launch Outcome

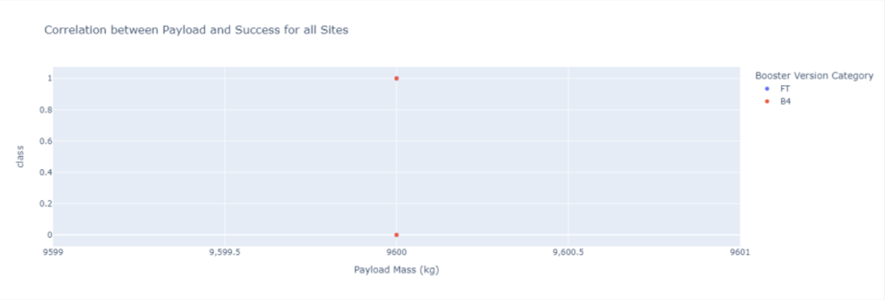


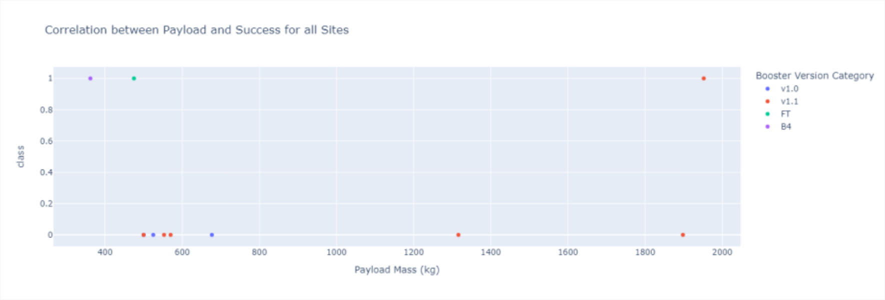


Most launches were in the 2k – 6k range, in which FT fared well, while v1.1 sank. B4 was exclusively reserved for the 9k – 10k high payload range, which has the least launches.



Most launches in the 5k – 7k midrange failed. They relied on FT and B4. Same fate was met by the 0 – 2k lightweight range, which majorly used v1.1 boosters.

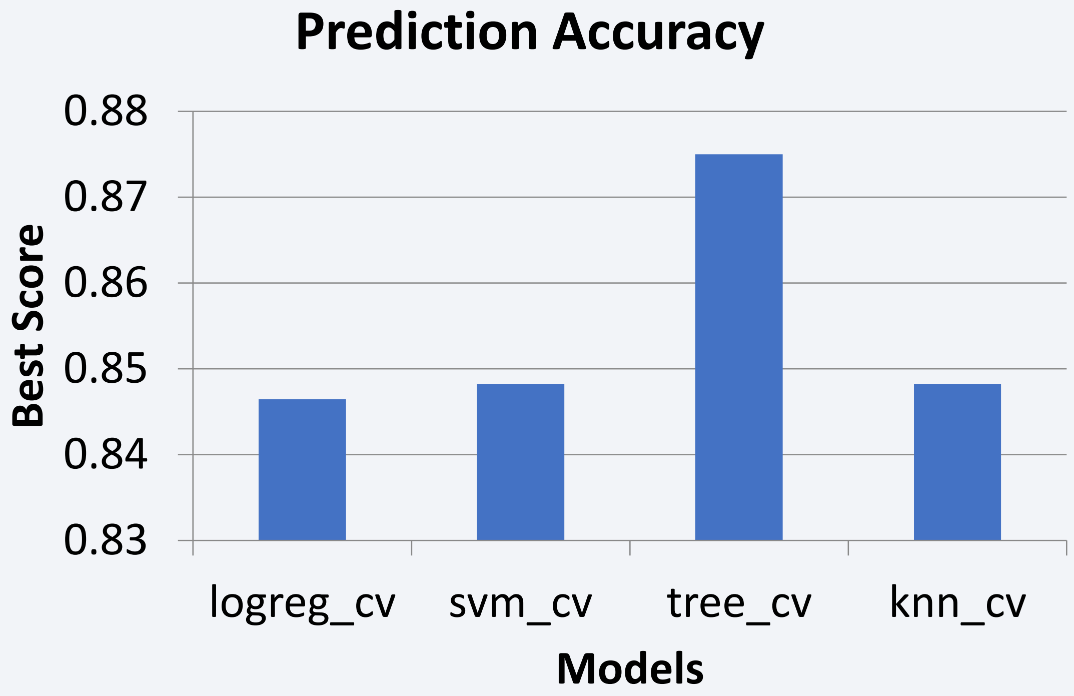




Section 5

Predictive Analysis (Classification)

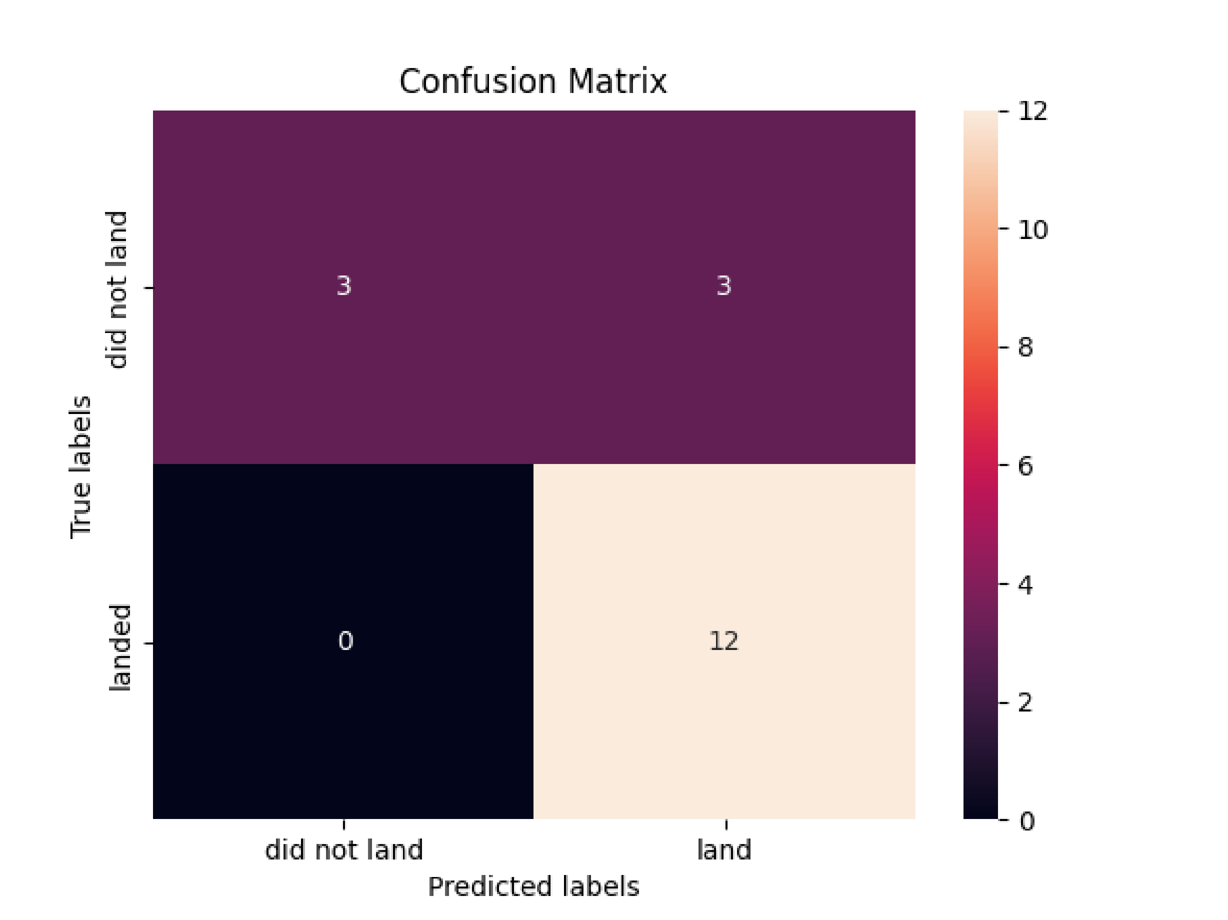
Classification Accuracy



The decision tree model has the highest accuracy that could be obtained by grid search optimisation.

Confusion Matrix

The R² score for all the models on the test set were same. There were no false negatives, but false positives cropped up as a problem, 20% of the positive predictions being mislabels.



Conclusions

- The year 2018 was of experimentation and learning. It had failures, but also the first successful ground pad landing.
- The booster v1.1 does not perform. FT is useful for small missions.
- Most missions have been light or medium loaded.
- Most experimental flights were from CCAFS SLC-40, and were to the LEO, ISS, PO, and GTO orbits.
- The site with the most launches, CCAFS LC-40, is well connected by roads and railways.
- The site with the most successful launches, KSC LC-39A, didn’t have much experimentation.

Appendix

- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-spacex-data-collection-api.ipynb
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-webscraping.ipynb
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/spacex_dash_app.py
- https://github.com/vedikajain2004/IBM-Data-Science-Professional-Certificate/blob/main/Applied%20Data%20Science%20Capstone/module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Thank you!