IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

CHEE Shang Yao
31st of August 2022

Outline

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

Executive Summary

- Methodologies for data analysis:
  - Data collection with Web scraping and SpaceX API,
  - Data Wrangling,
  - Exploratory Data Analyses with SQL, Data Visualisation
  - Interactive Visual Analysis with Folium, Dashboarding with Poltly, Dash,
  - Machine Learning for classification
- Summary of all results
  - Results for Exploratory Data Analyses
  - Results for Data Visualisations
  - Results for Machine Learning classification

Introduction

- Project background and context
  - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. If we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.
  - This capstone project seeks to determine, with the use of Data Science and the available resources, if first stages of rocket launches would successfully land.
- Problems you want to find answers
  - Using available data to assess if rockets will successfully land.
  - Explore variables that may affect successful landing rates.
  - Explore classification models best suited for the project.

Section 1

Methodology

Methodology

Executive Summary

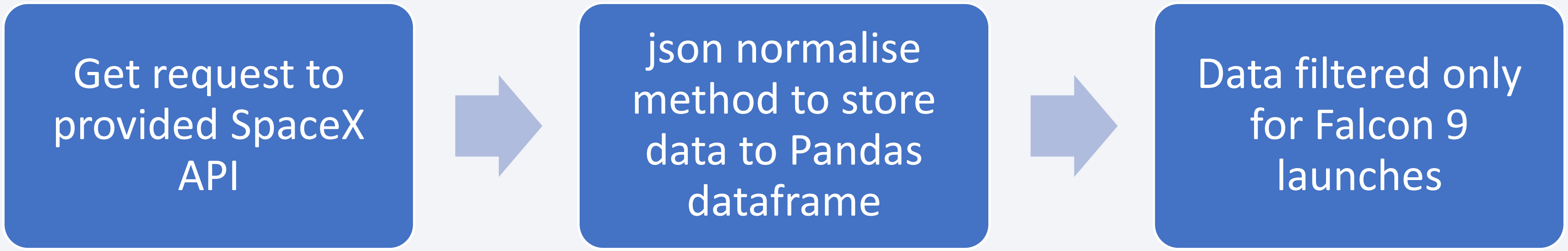
- Data collection methodology:
  - Data was collected with SpaceX API and Web Scrapping SpaceX's Wikipedia page.
- Perform data wrangling
  - Data were filtered, null-values were addressed. One-hot encoding applied to categorical values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data were split for testing and training. Classification algorithms were then evaluated.

Data Collection

- Data collection processes are summarised as follows:
    - Get request to SpaceX API, data parsed and transferred to Pandas dataframe.
    - Data filtered and missing valued of Falcon 9 launches of Payload Mass replaced with mean.
    - Web Scrapping of SpaceX's Wikipedia page regarding launch data using BeautifulSoup then transferred to Pandas dataframe.

Data Collection – SpaceX API

flowchart of SpaceX API calls



Get request to provided SpaceX API

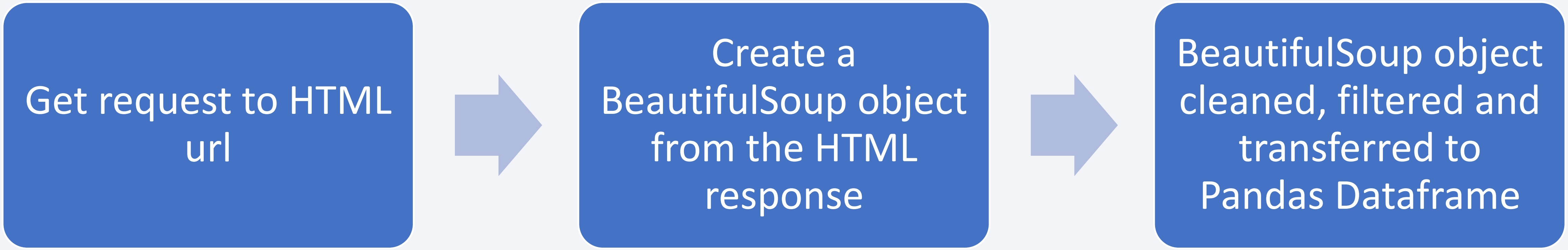
json normalise method to store data to Pandas dataframe

Data filtered only for Falcon 9 launches

GitHub URL https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%201%20-%20Data%20Collection%20API.ipynb

Data Collection - Scraping

flowchart of web scraping



Get request to HTML url

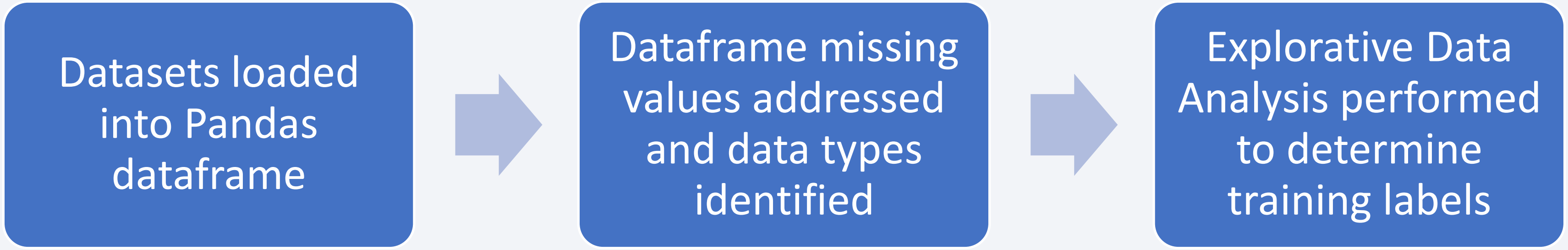
Create a BeautifulSoup object from the HTML response

BeautifulSoup object cleaned, filtered and transferred to Pandas Dataframe

GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%201%20-%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

flowchart of data wrangling

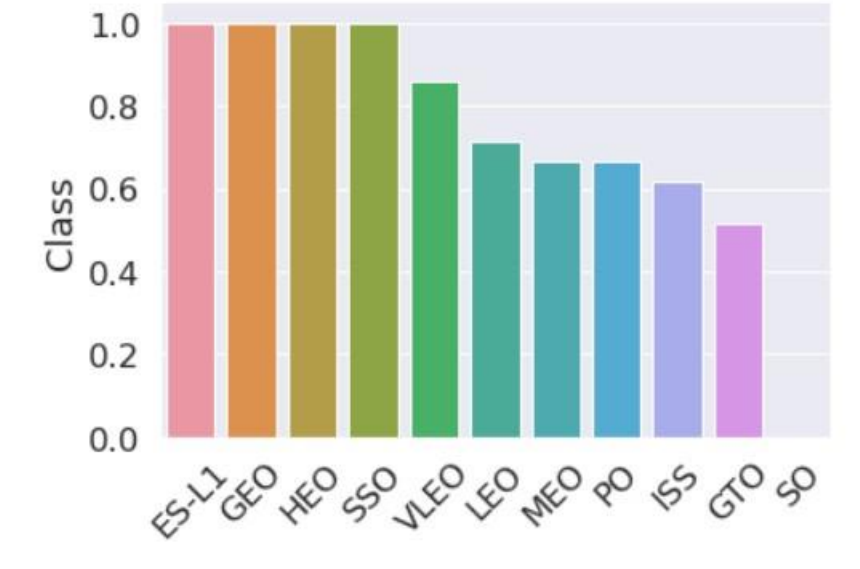

Datasets loaded into Pandas dataframe
Dataframe missing values addressed and data types identified
Explorative Data Analysis performed to determine training labels

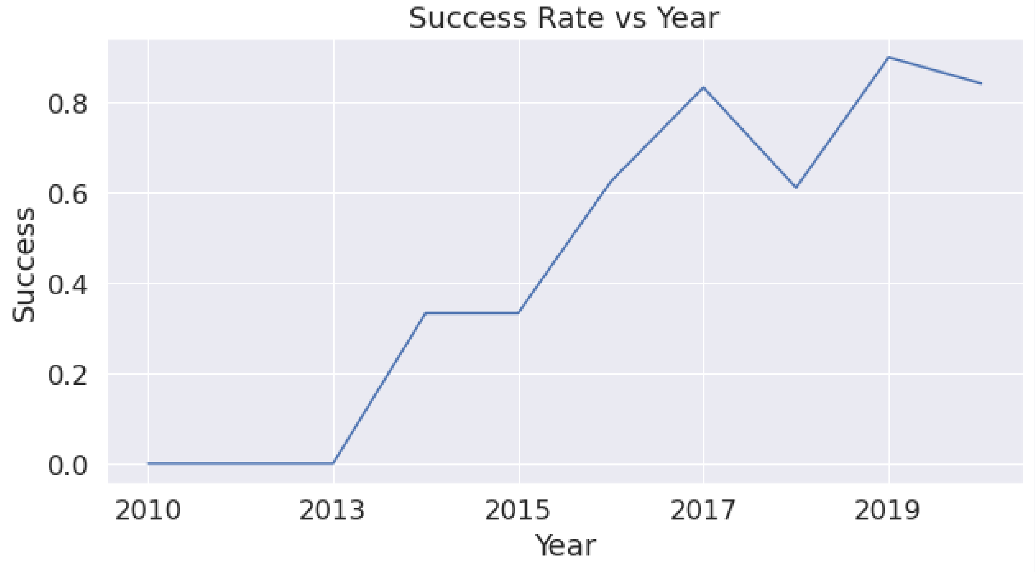
- GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%201%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

Using EDA allowed for visualisation of relationship between flight number, launch sites, payload, types of orbits in a form of a bar chart.

Success rate changes over the years were plotted as a line chart.





GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%202%20-%20EDA%20with%20Visualization.ipynb

EDA with SQL

- Datasets were loaded to DB2 and connection with Jupyter notebook was made.
- Using SQL, queries as followed were made:
  - Names of unique sites
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Listed date when the first successful landing outcome in ground pad
  - Listed names of the boosters successful in drone ship with payload mass > 4000 but < 6000 kgs
  - Listed total number of successful and failure mission outcomes
  - Listed names of booster versions which have carried the maximum payload mass
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%202%20-%20EDA%20and%20SQL.ipynb

Build an Interactive Map with Folium

- Markers used to incidicate launch sites
- Circles used to highlight areas around coordinates
- Colour labeled marker clusters used to indicate groups of events at specific coordinates
- Lines used to indicate and calculate distances between two coordinates of interest. Used to answer questions of proximity between two points.

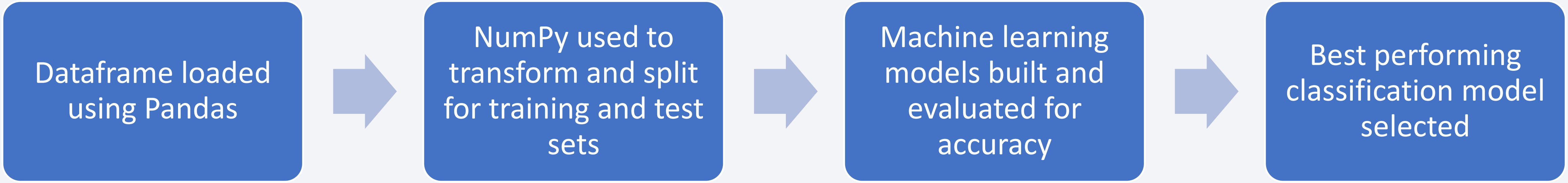
GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%203%20-%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Interactive dashboard built with Poltly Dash lab.
- Dropdown menus and range sliders created.
- Charts plotted with Plotly
  - Pie charts for number of launches from specific launch sites.
  - Scatter graphs indicating relationship between payload masses and outcome for specific booster versions

GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%203%20-%20SpaceX%20Dash%20App.py

Predictive Analysis (Classification)



Dataframe loaded using Pandas

NumPy used to transform and split for training and test sets

Machine learning models built and evaluated for accuracy

Best performing classification model selected

GitHub URL - https://github.com/sycheegithub/Applied-Data-Science-Capstone/blob/a67a757de98a23beaed3fe75486138d78a873021/Week%204%20-%20Machine%20Learning%20Prediction.ipynb

Results

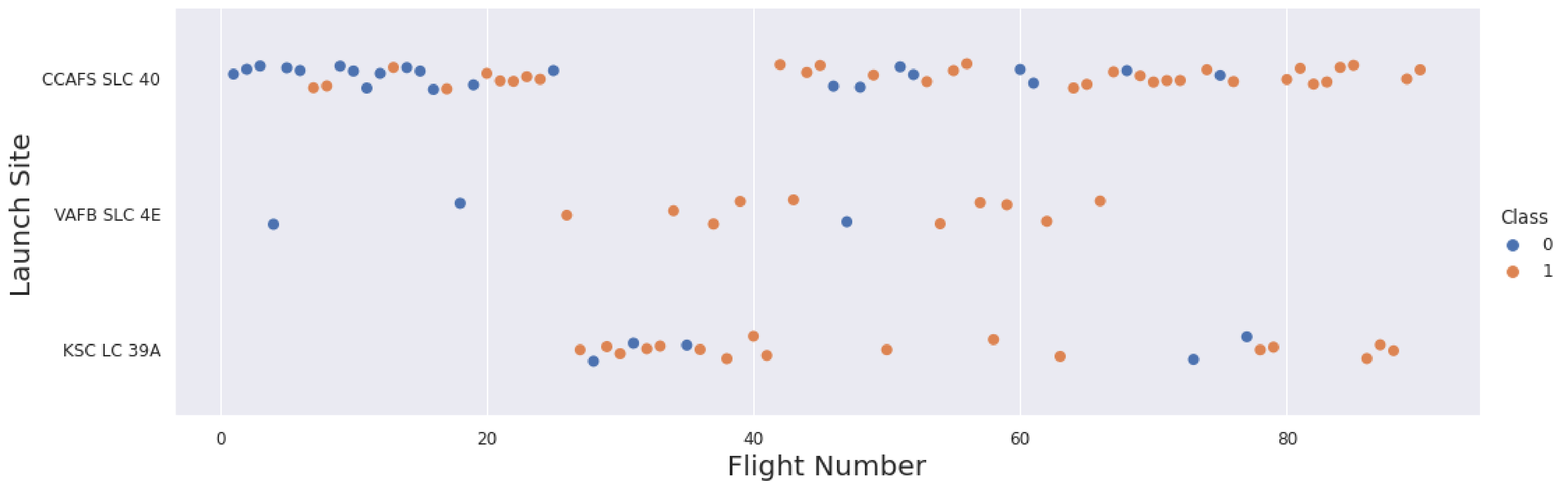
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

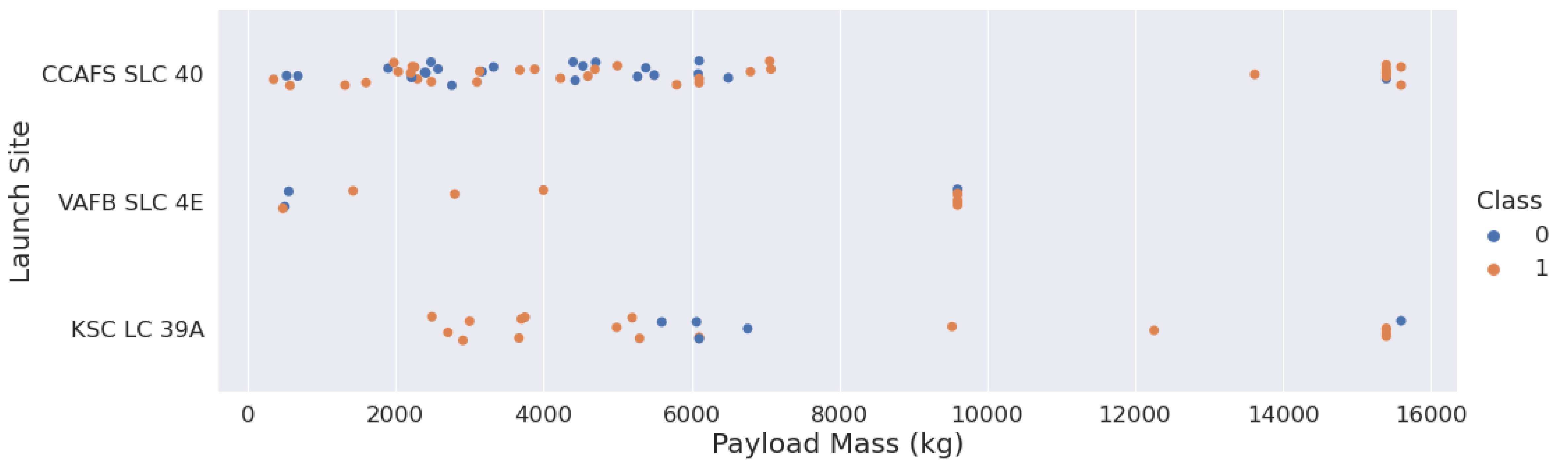
Flight Number vs. Launch Site

- Most number of flights originated from CCAFS SLC 40 as compared to KSC LC 39A and VAFB SLC 4E



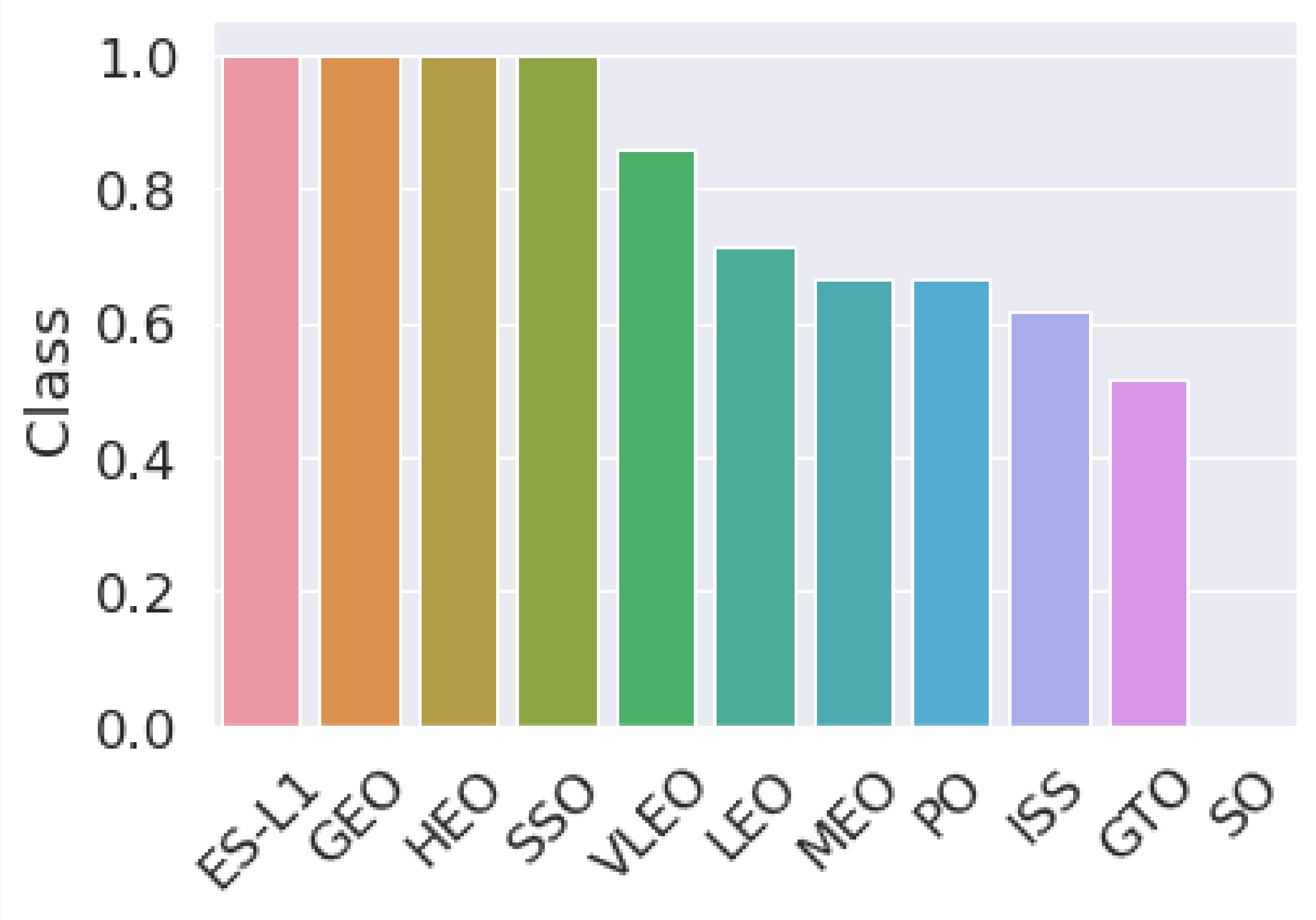
Payload vs. Launch Site

- Higher payloads have resulted in higher rates of successful outcomes.
- VAFB SLC 4E does not seem to be able to support higher payloads masses (>12000 kg).



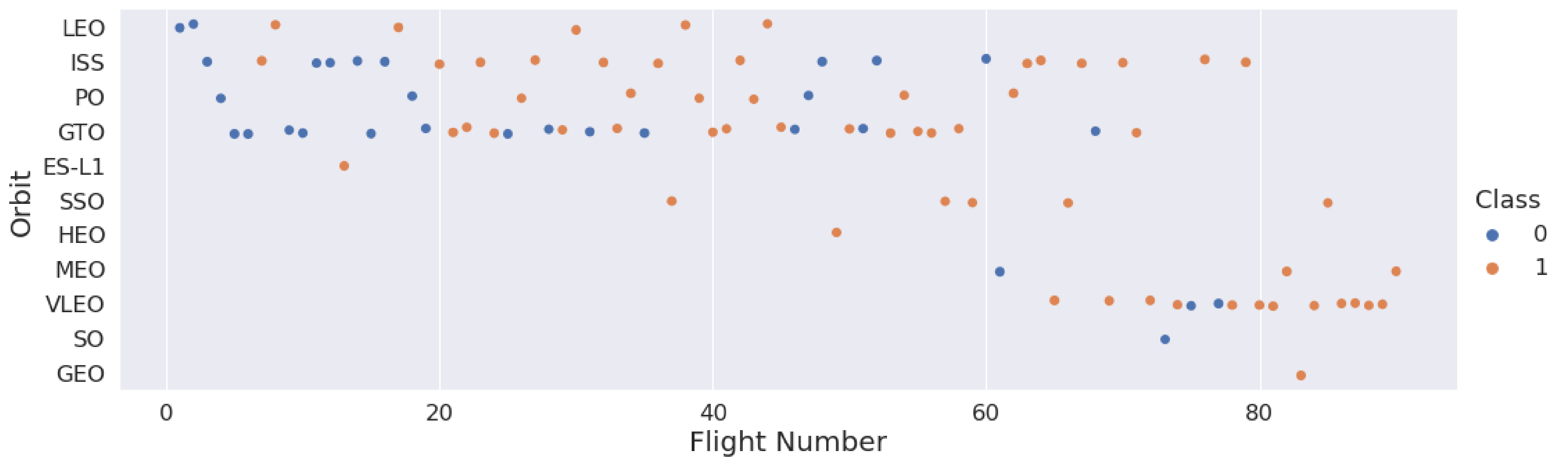
Success Rate vs. Orbit Type

- Orbit tyes of ES-L1, GEO, HEO and SSO had certainty in success rates. This is followed by VLEO, LEO, MEO, PO, ISS and GTO.



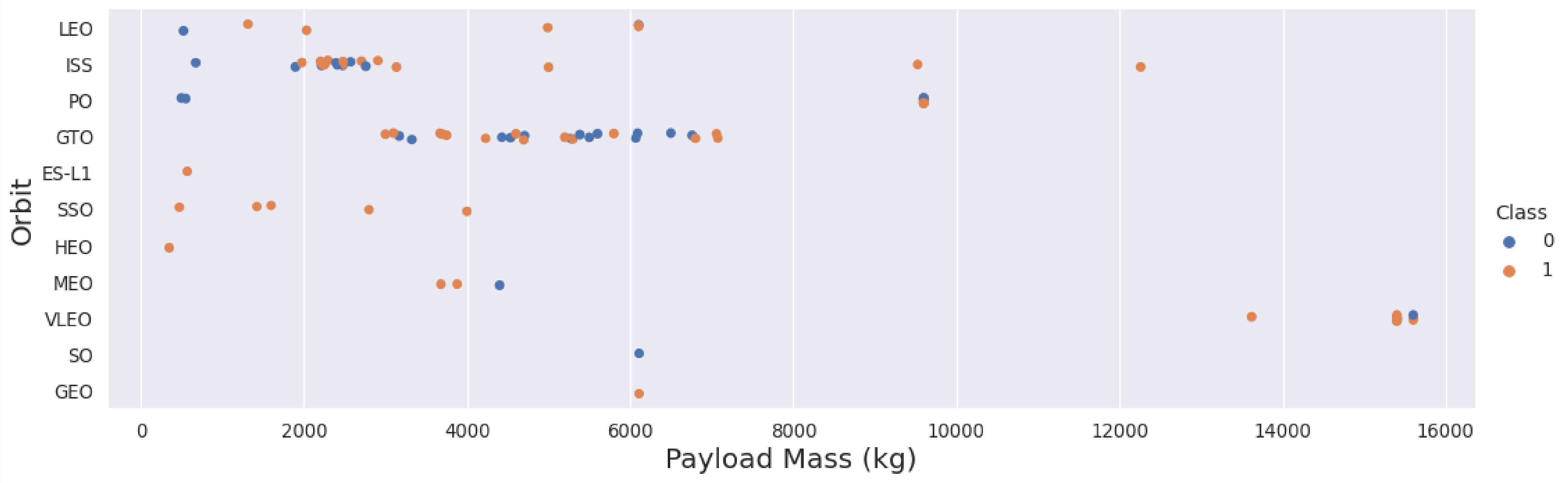
Flight Number vs. Orbit Type

- We see that in the LEO orbit the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.



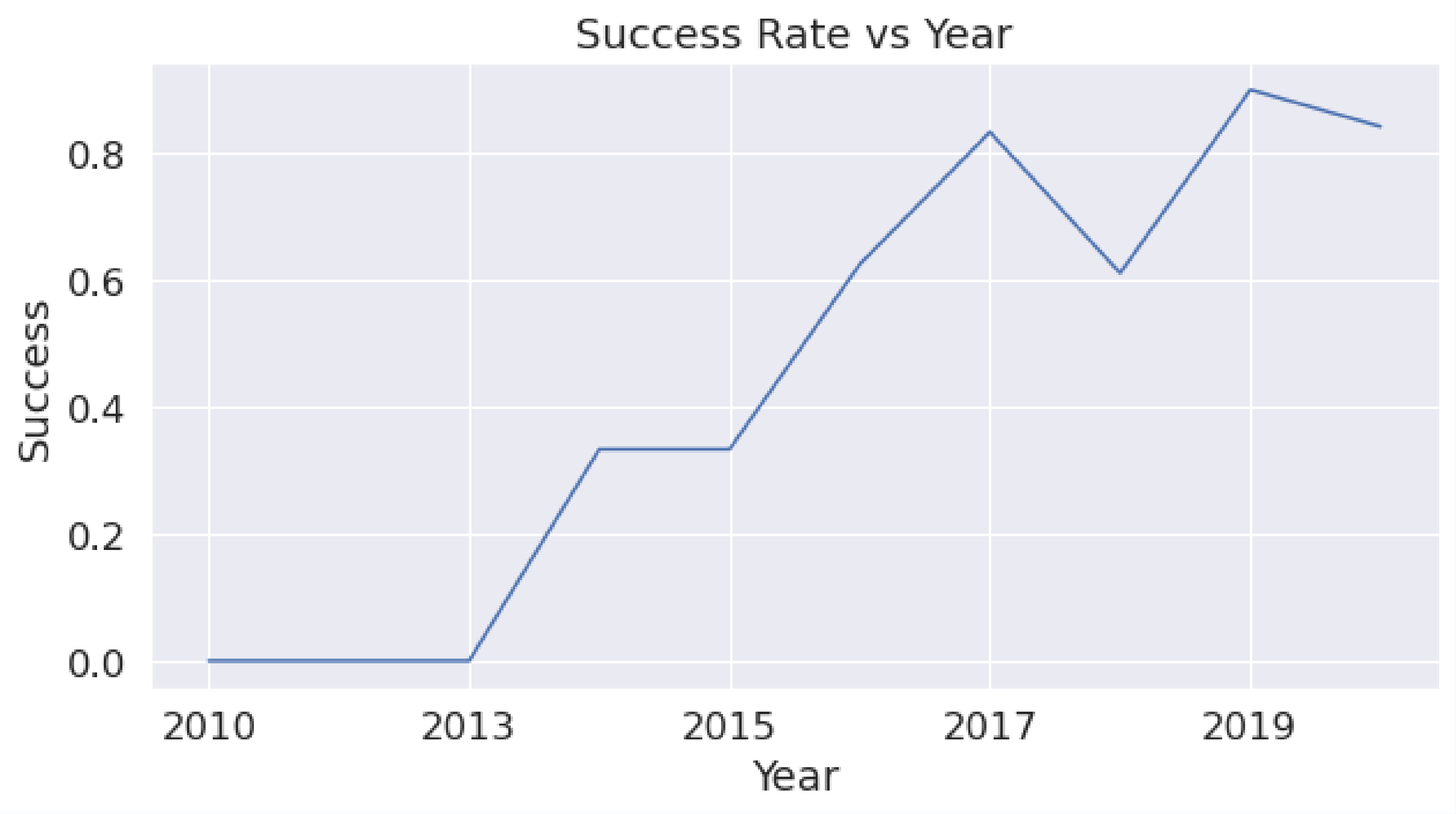
Payload vs. Orbit Type

- With heavy payloads the successful landing are more for Polar,LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here.



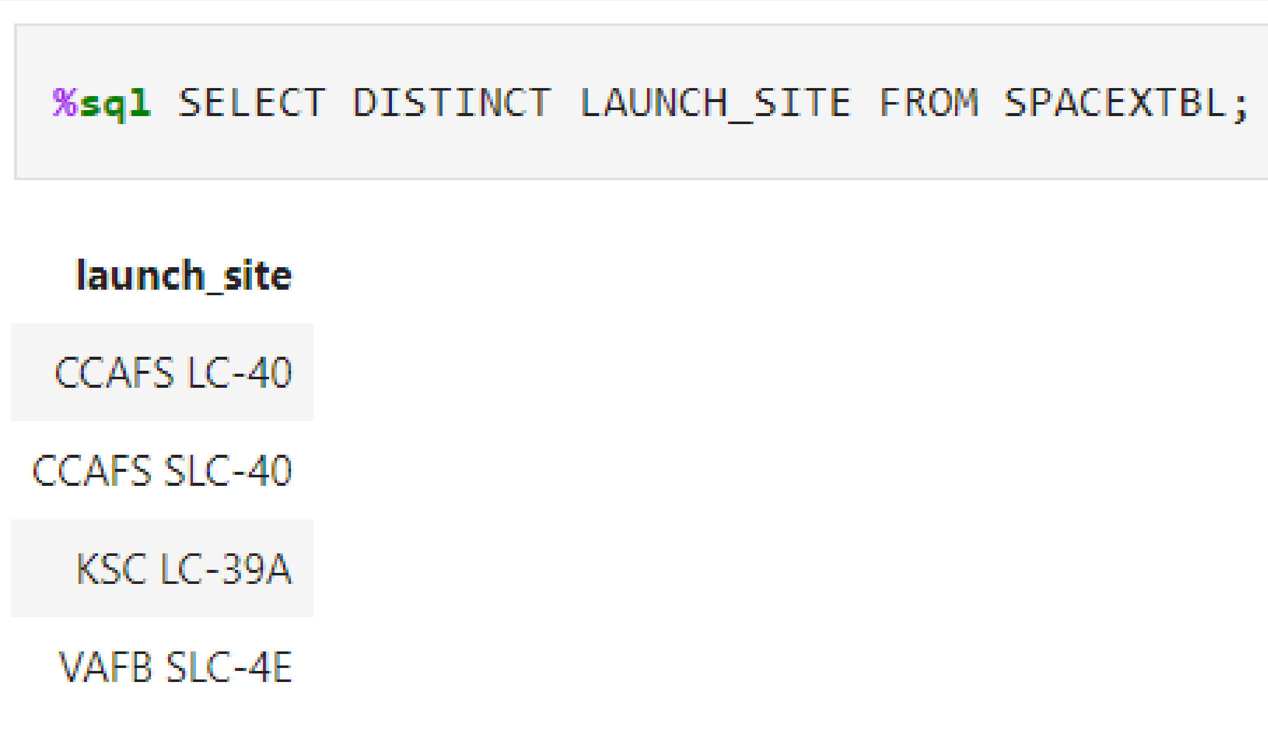
Launch Success Yearly Trend

- Sucess rate since 2013 have been improving over time.



All Launch Site Names

- SELECT DISTINCT sql command was used to get unique names of launch sites.

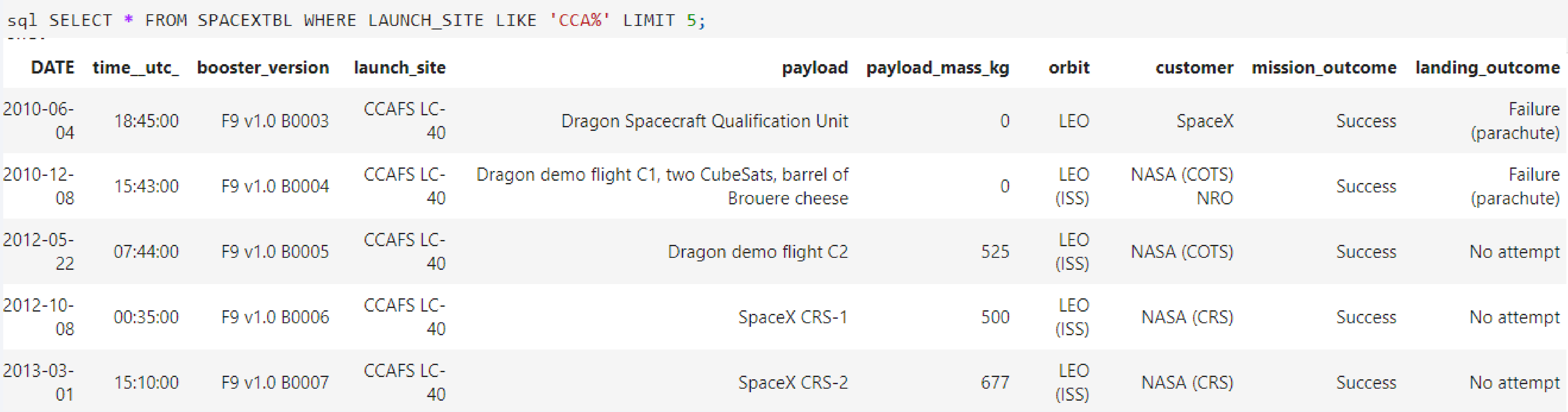


```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- Using SQL command LIKE to select launches beginning with CCA
- SQL command LIMIT was used to display five records

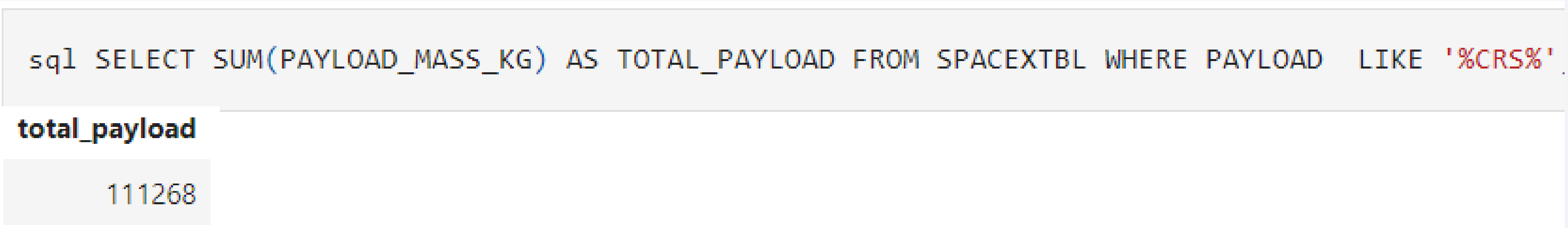


```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass_kg	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- SQL command SUM was used to calculate the total payload.
- SQL command LIKE %CRS% to select NASA (CRS) boosters.

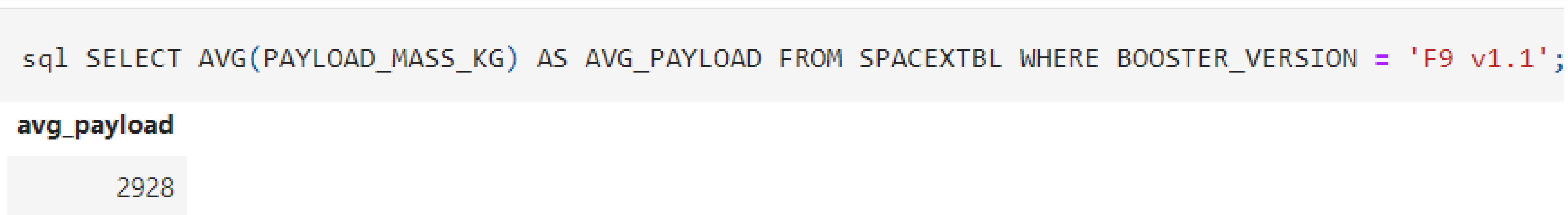


```
sql SELECT SUM(PAYLOAD_MASS_KG) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%'
```

total_payload
111268

Average Payload Mass by F9 v1.1

- Average payload mass was calculated using SQL command AVG.

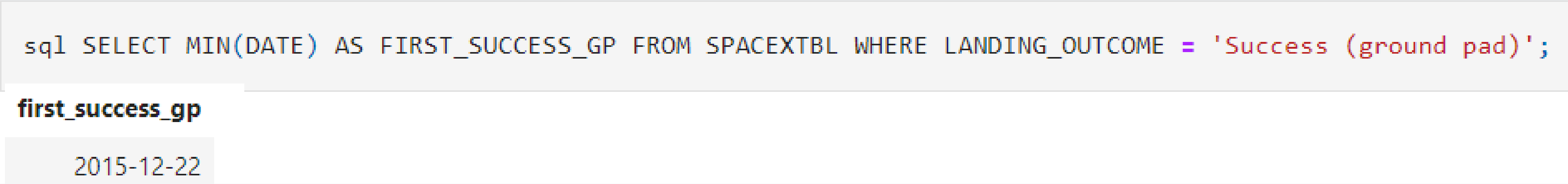


```
sql SELECT AVG(PAYLOAD_MASS_KG) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

avg_payload
2928

First Successful Ground Landing Date

- SQL command MIN(DATE) was used to identify the earliest date of an event.

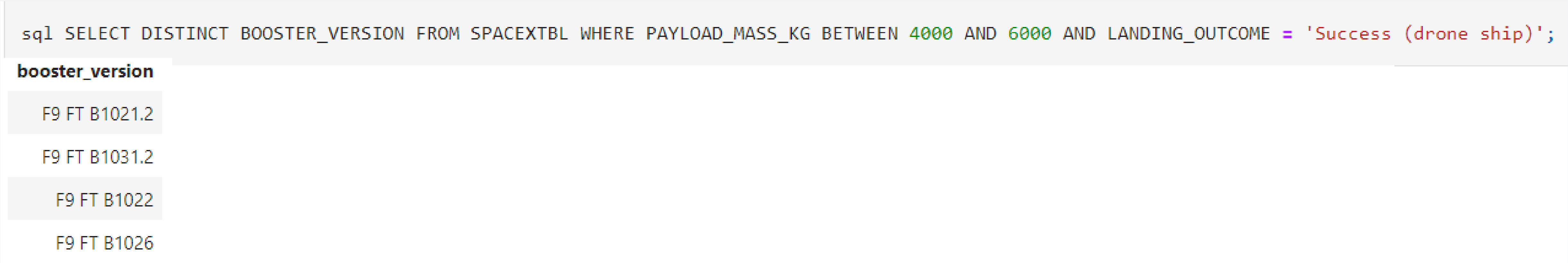


```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

```
first_success_gp
      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL command BETWEEN was used to identify instances of payload mass that falls in that range.
- SQL command AND was used to further filter successful landings on a drone ship.

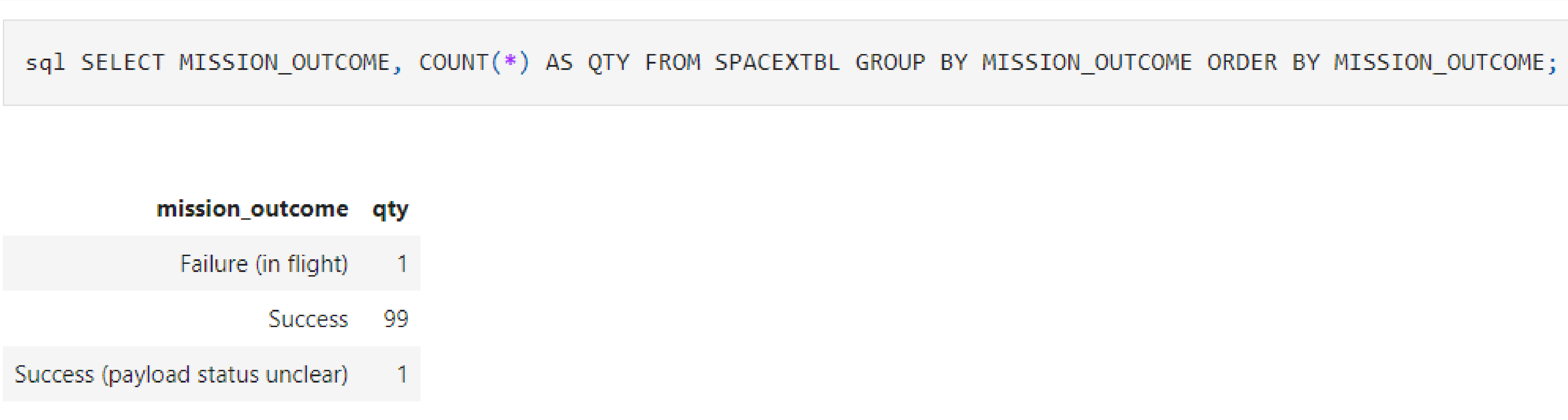


```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS_KG BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)';
```

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

Total Number of Successful and Failure Mission Outcomes

- SQL command AS was used for header of quantity.
- SQL command group by was used to categorize the counts.

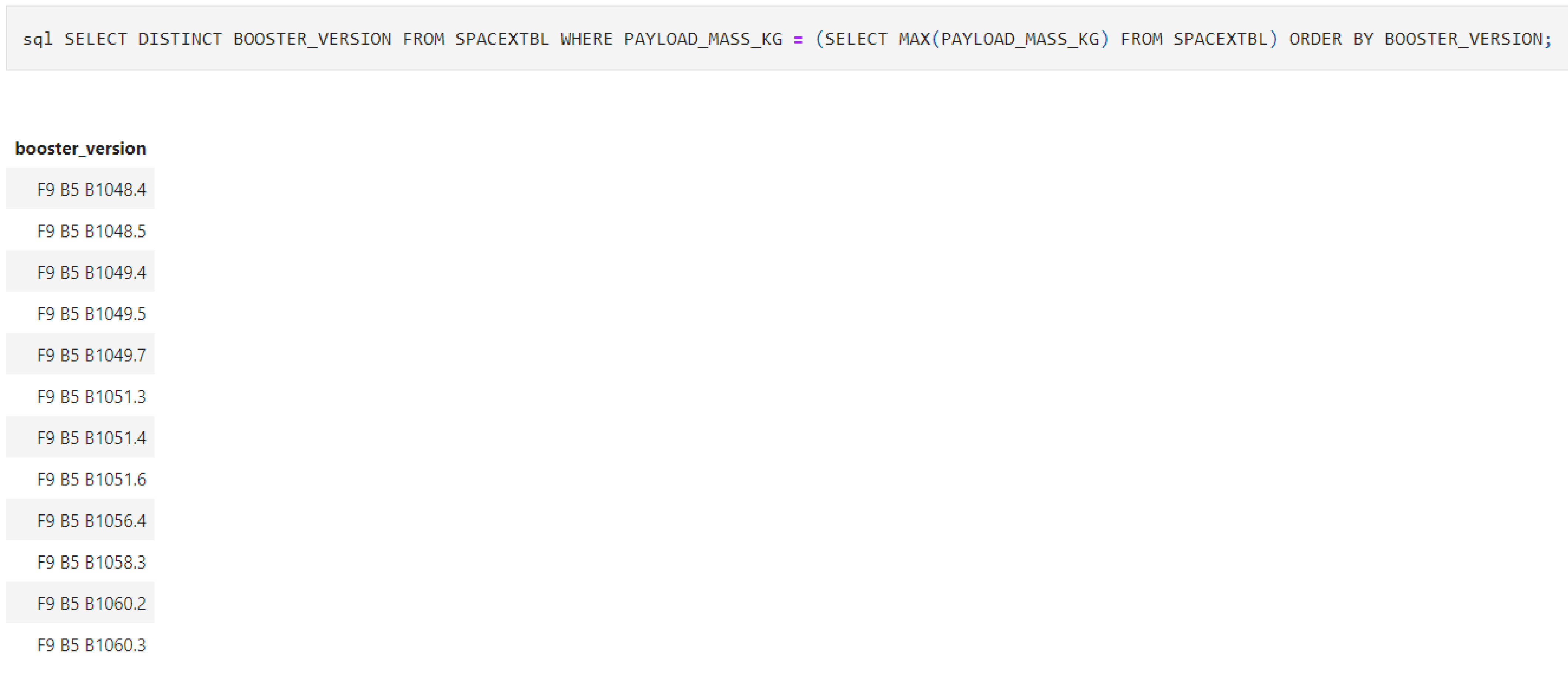


```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- A nested SQL query was performed where SQL command MAX was nested within a WHERE command. This was used to find boosters carrying maximum payloads.

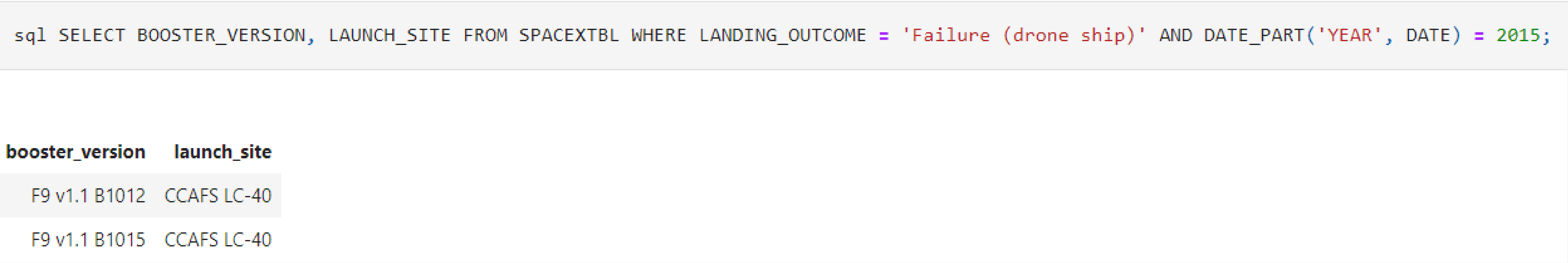


```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS_KG = (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

2015 Launch Records

- SQL Command AND was used to identify two specific criteria (landing outcome and date)

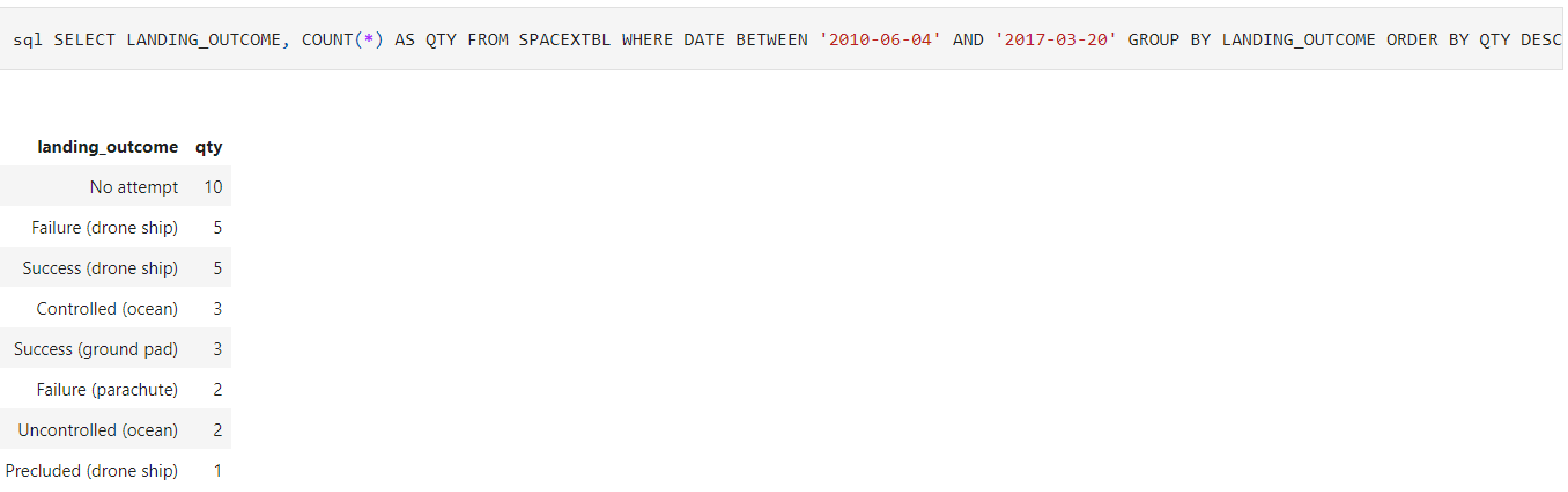


```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- SQL command ORDER BY DESC was used to rank landing outcomes from highest to lowest.



```
sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY QTY DESC
```

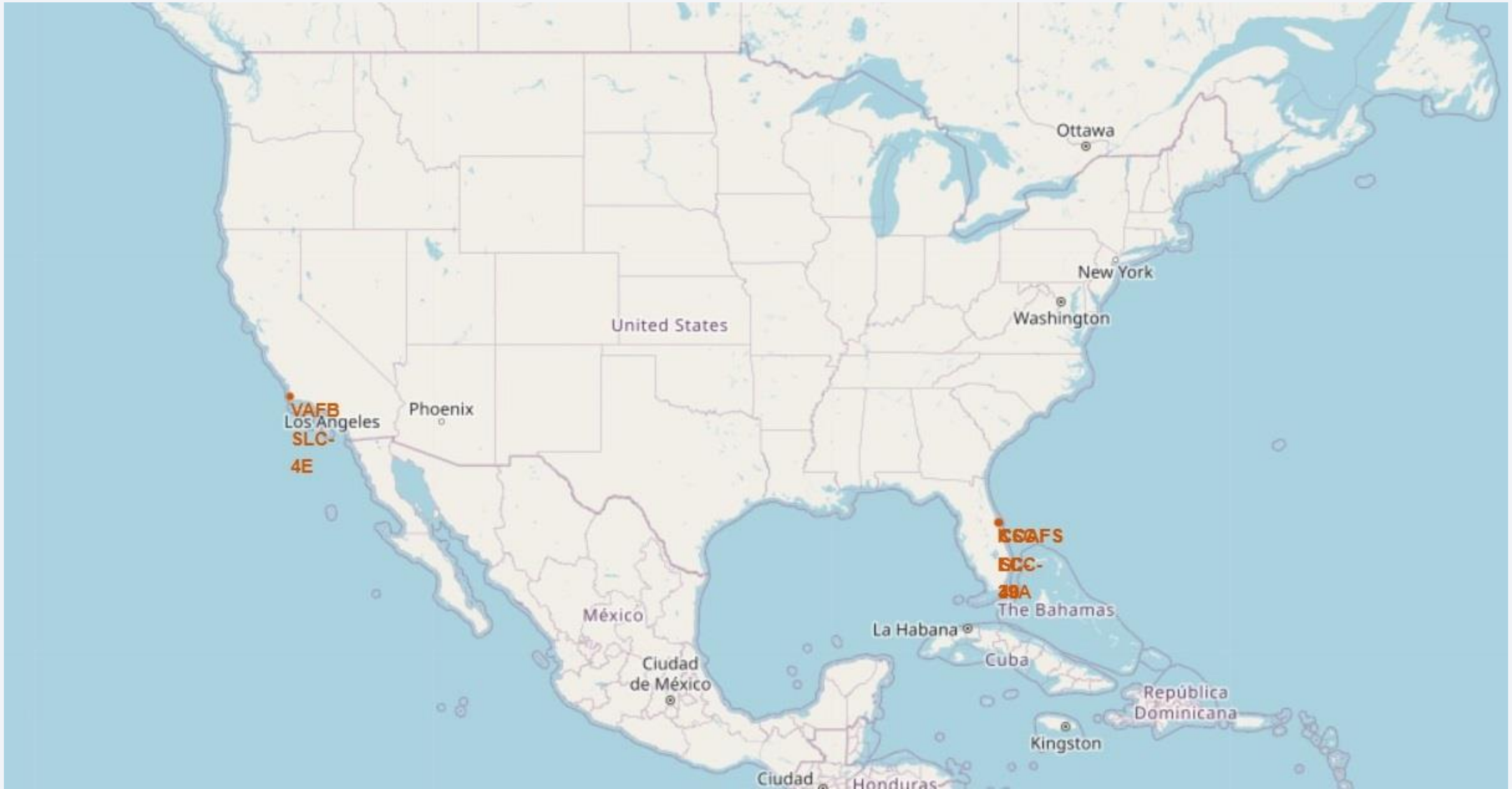
landing_outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

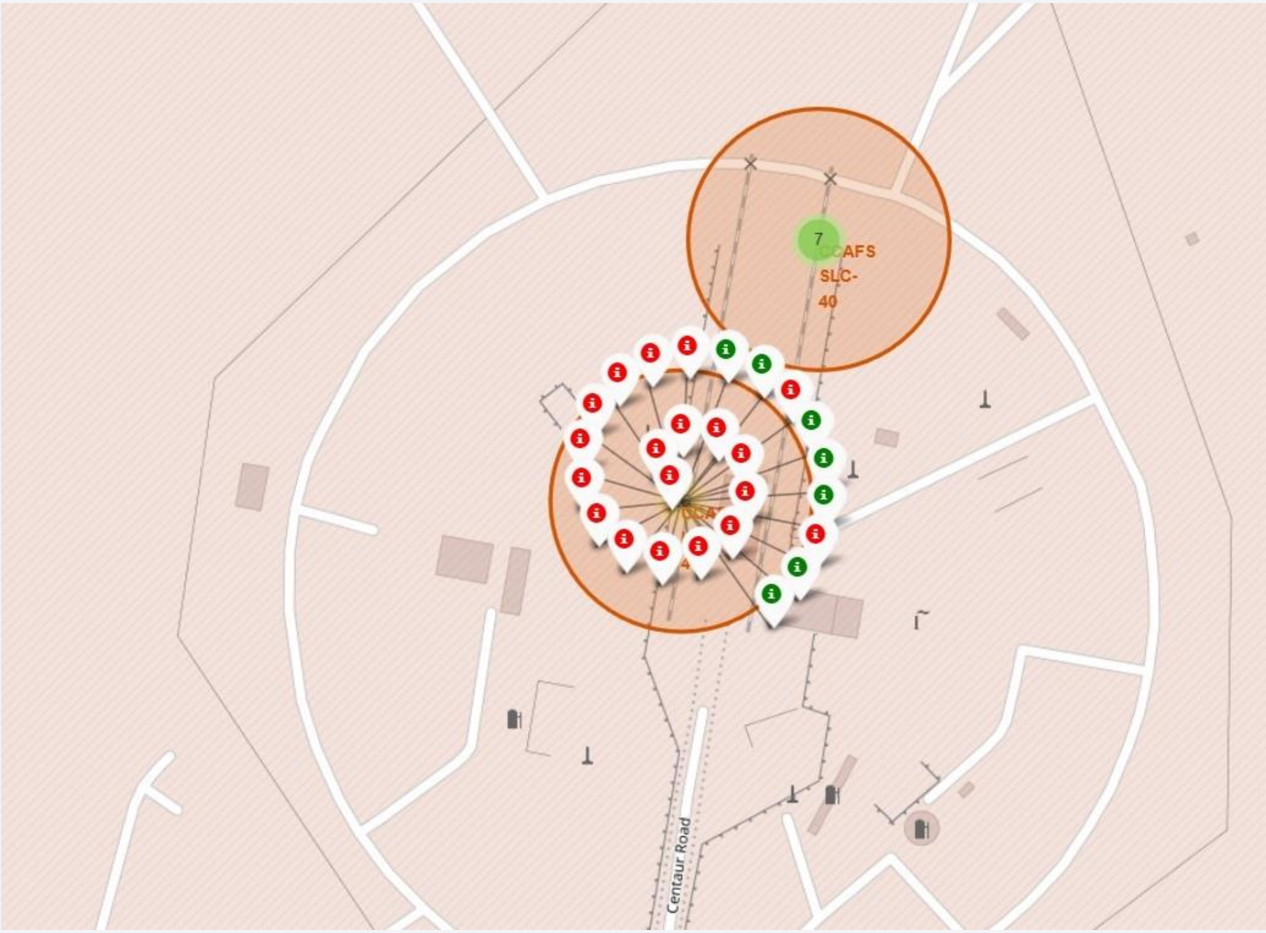
Launch Sites Proximities Analysis

Rocket Booster Launch Sites

- SpaceX launch sites marked

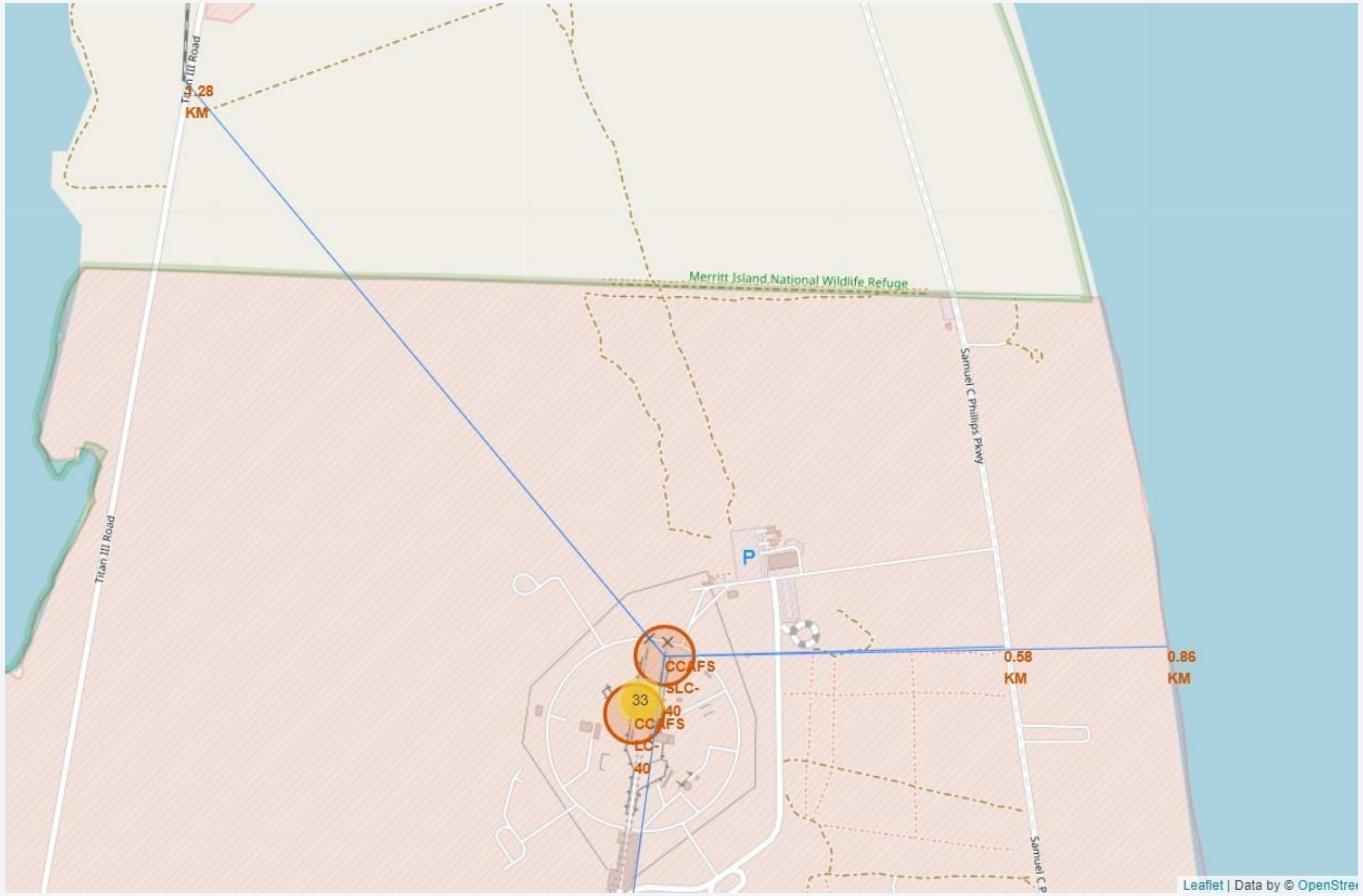


Launch outcomes of CCAFS LC-40



- Green marker – successful launches
- Red marker – failed launches

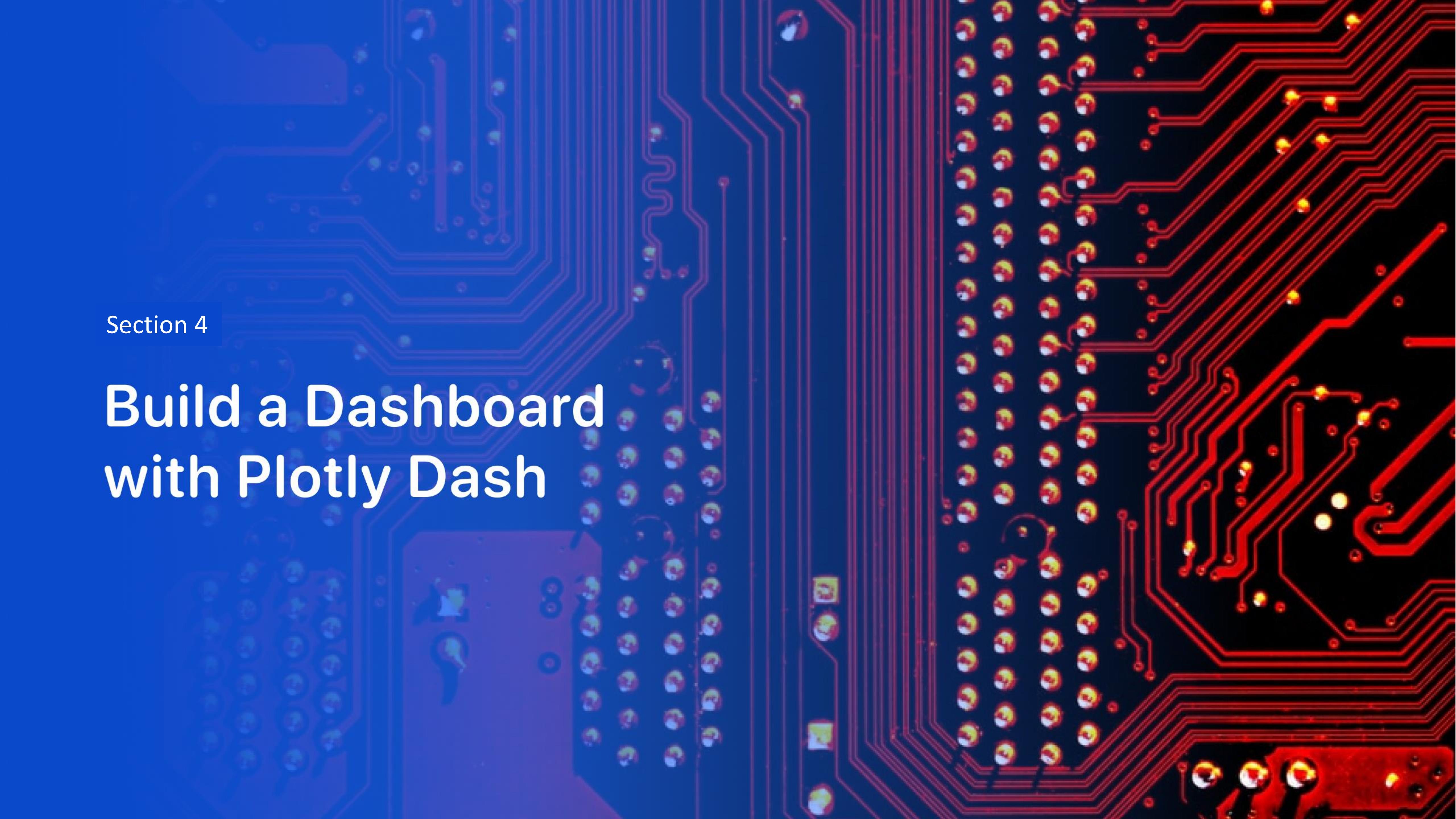
Logistics and distances



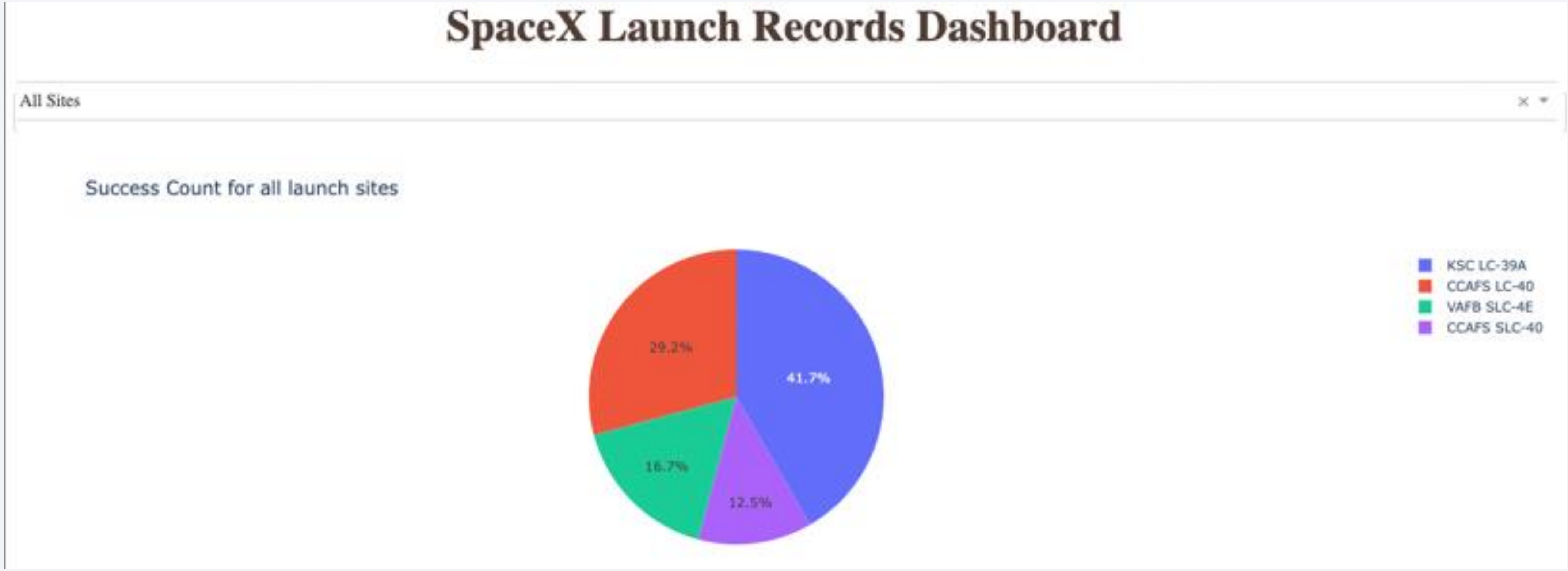
- We used CCAFS SLC-40 as a launchpad of reference.
- As calculated, the proximity are summarized below:
    - Highway – 0.58 km
    - Coastline – 0.86 km
    - Railway – 1.28km
- It appears to be an optimum site for launching as it is located close to the railway and highway yet providing a location for safe launching as it is next to the coastline.

Section 4

Build a Dashboard with Plotly Dash



Successful launches based on site



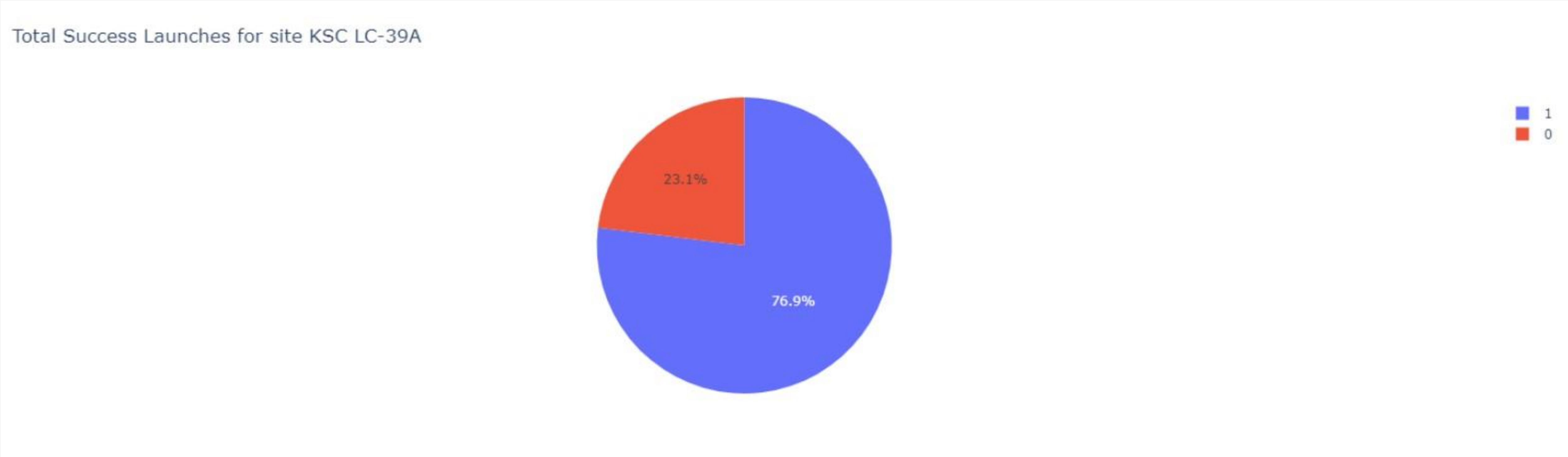
SpaceX Launch Records Dashboard

All Sites

Success Count for all launch sites

- KSC LC-39A had the most successful launches with 41.7%, followed by CCAFS LC-40 with 29.2%.

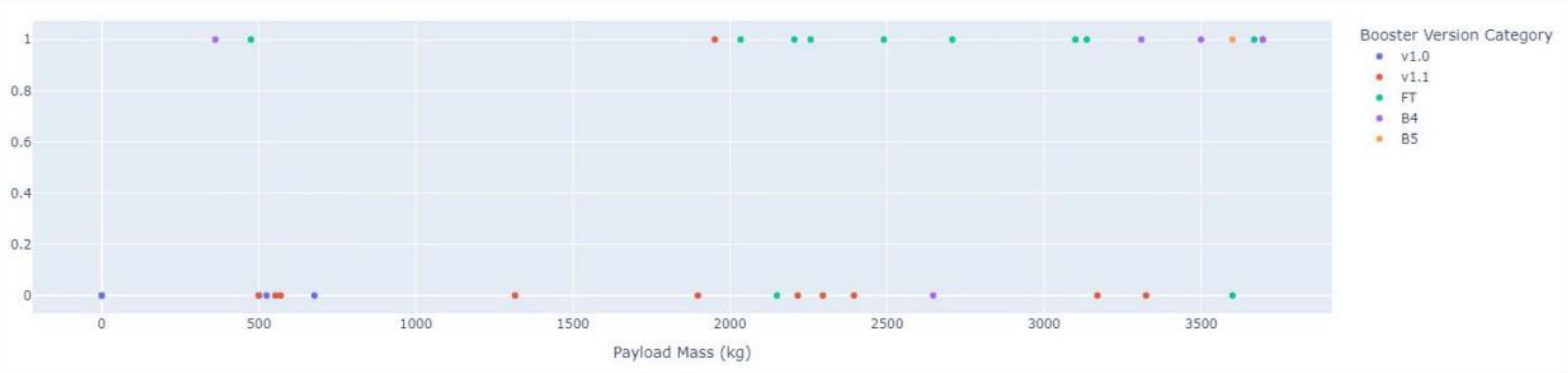
Launchsite KSC LC-39A by Launch Outcomes



- The KSC LC-39A produces successful launch outcomes of 76.9%

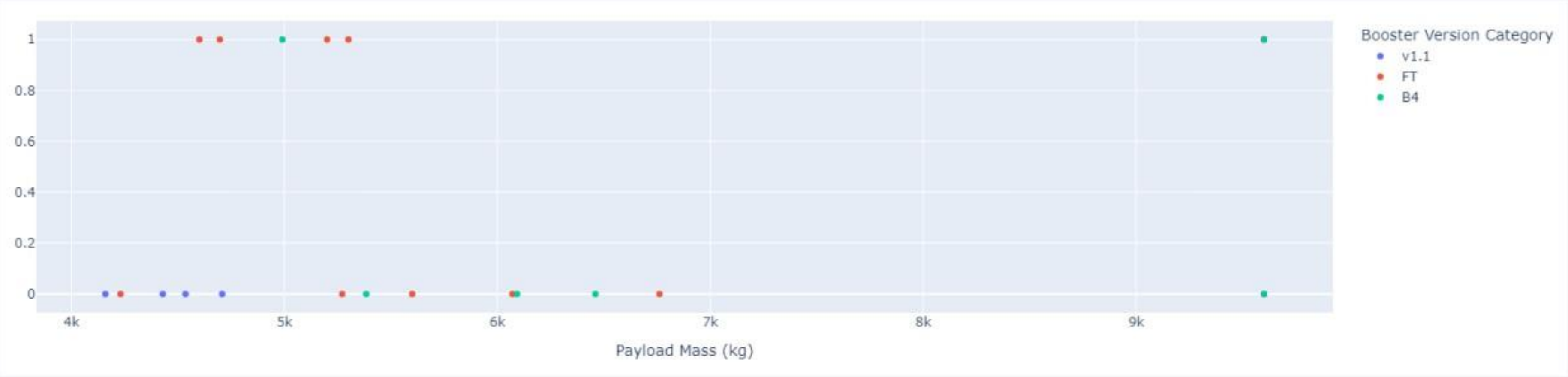
Payload vs Launch Outcomes Based on Payloads

Payload mass below 4000 kg



Payload vs Launch Outcomes Based on Payloads

Payload mass above 4000 kg



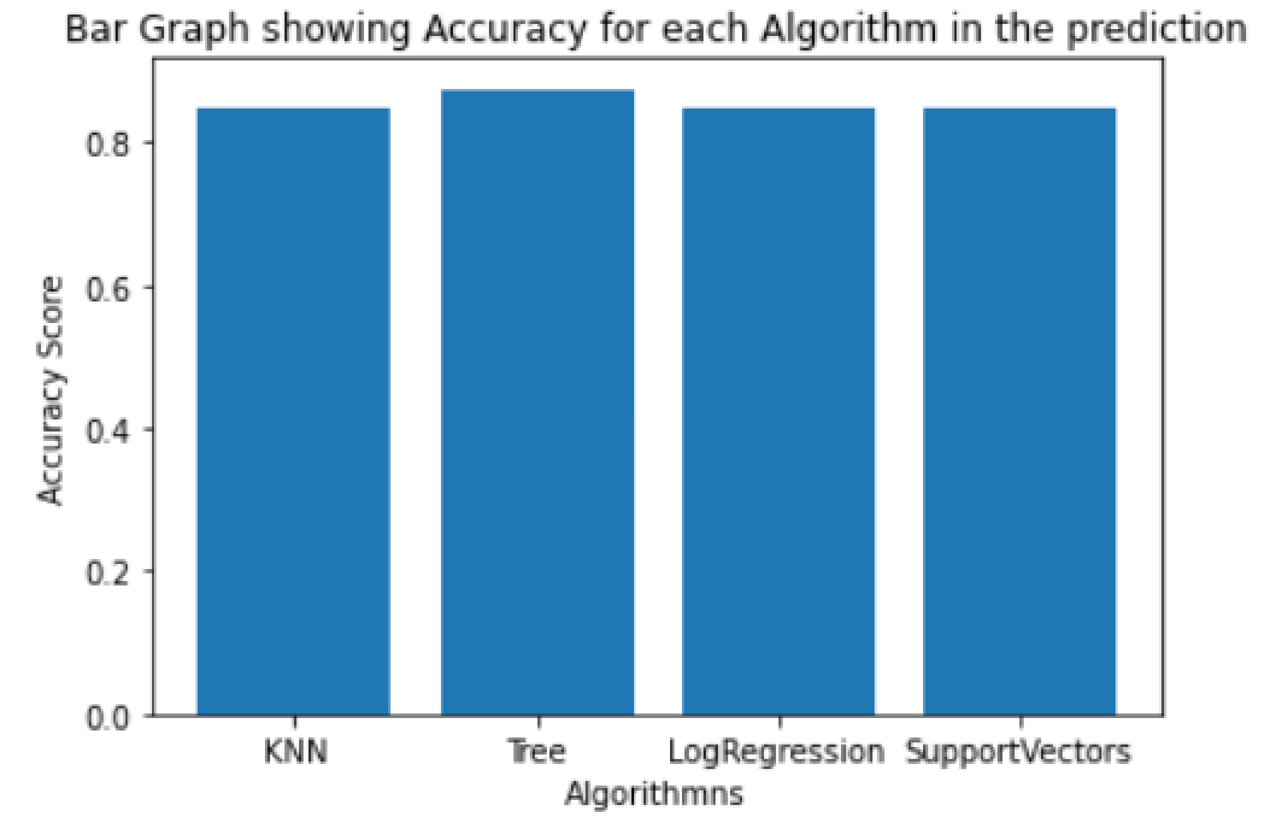
- We can visually appreciate that launches with payload masses below 4000 kg has a better launch outcomes.

Section 5

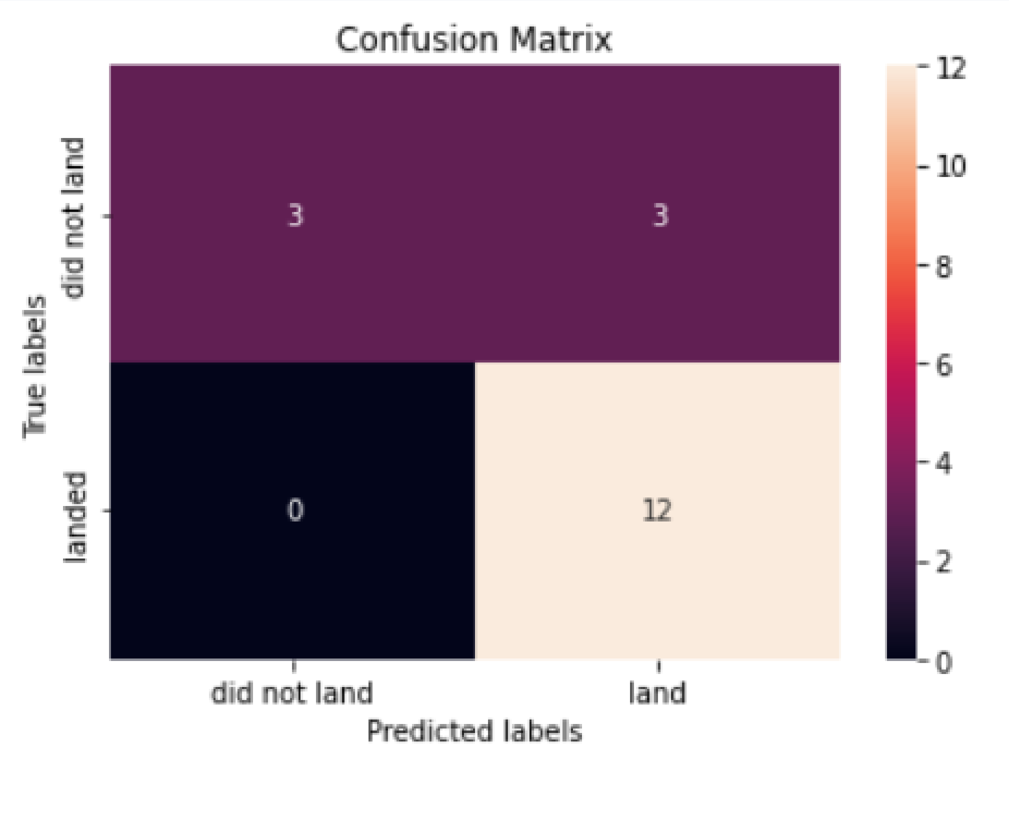
Predictive Analysis (Classification)

Classification Accuracy

- The bar chart visualized indicates that the decision tree classification has the highest accuracy whilst the others less, but, similarly accurate.



Confusion Matrix



- The confusion matrix of the decision tree classifier indicates the following:
  - 12 true-positives
  - 3 true-negatives
  - 3 false-positives
  - 0 false-negatives
- However, all the other confusion matrixes had identical accuracy.

Conclusions

- The project suggested that KSC LC-39A is the best site for launching rockets.
- Launches for orbits ES-L1, GEO, HEO, SSO, VLEO had the highest success rates.
- The Decision Tree Classifier is the most appropriate classifier for this project.
- There had been steady improvement in launch success rates over the years.
- Success in launch outcomes increases with increasing number of launches for any given launch site.

Thank you!