IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Omar Jisr

September 29, 2021

Outline

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

Executive Summary

- The data for this study was collected using SpaceX's REST API and web scraping launch records from Wikipedia.
- The data was then wrangled to perform an exploratory data analysis (EDA) using Python and SQL to extract information and produce visualizations.
- Visualizations include an interactive map using Folium and an interactive dashboard using DASH.
- Machine learning techniques were used to analyze the data and predict the successful landing of future Falcon 9 launches.
- The best launch success rates are for launches from KSC LC-39A, launches to SSO, GEO, HEO and ES-L1 orbits, launches with 1,900 Kg to 3,700 Kg payloads, and launches using booster version FT.

Introduction

- 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.
- Therefore, 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.
- In this study we want to predict the success of recovery and reuse of the first stage of a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection using Space X API
  - Data collection using web scraping
- Perform data wrangling
- 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 Collection

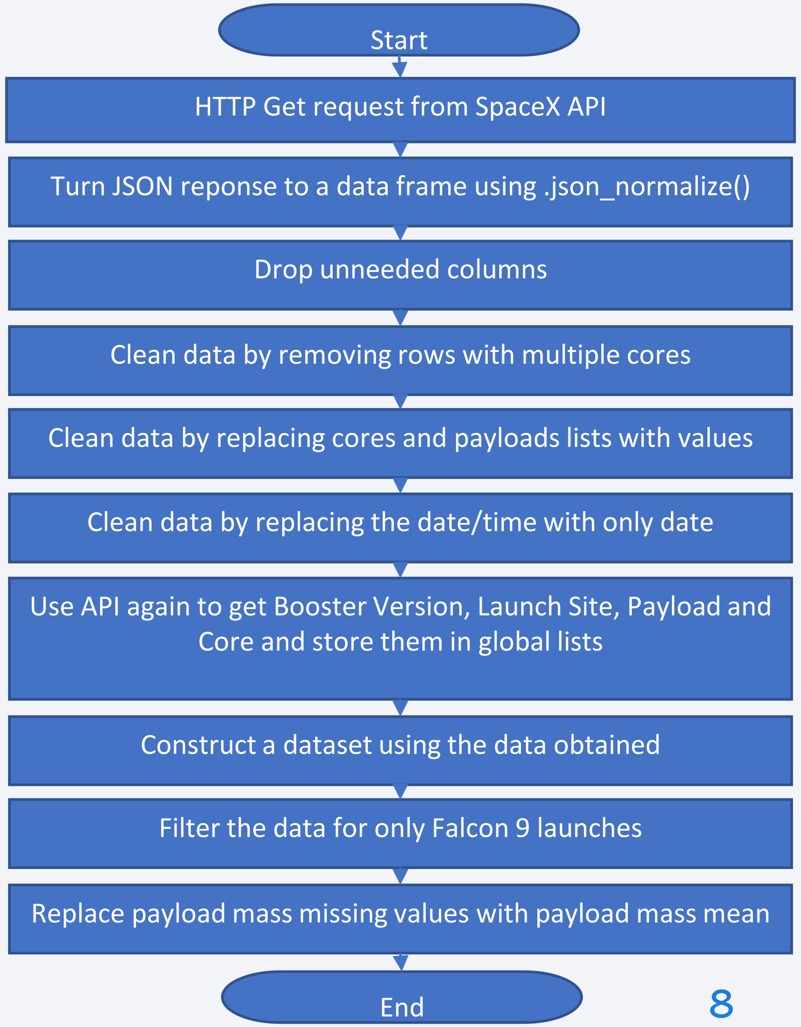
The data was collected using two methods: through the SpaceX REST API and by web scraping the SpaceX launch Wikipedia page.

The result of the API method is a data frame containing dates, payload masses, orbits, launch sites, outcomes, number of flights and other details that relate to Falcon 9 booster launches.

The result of the web scraping method is a data frame containing launch sites, payloads, orbits, customers, launch outcomes, booster versions, landing outcome, date and time of Falcon 9 booster launches.

Data Collection – SpaceX API

- Use a Get request using the REST API to get data in the form of a JSON response.
- Turn JSON response to a data frame using Pandas.
- Clean data by dropping unneeded columns, removing rows with multiple cores, replacing core and payload lists with single values, and replacing date/time with only date.
- Use API to get booster version, launch sites, payloads and cores.
- Combine both sets of data into a single dataset.
- Filter data for Falcon 9 launches.
- Fill in missing values with mean values for that column.
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/4beca6e9c63559981487f69fa69ad2cf8eb6d95a/SpaceX_Falcon_9/Week%201%20Lab%201%20Data%20Collection%20API.ipynb



Start

HTTP Get request from SpaceX API

Turn JSON reponse to a data frame using .json_normalize()

Drop unneeded columns

Clean data by removing rows with multiple cores

Clean data by replacing cores and payloads lists with values

Clean data by replacing the date/time with only date

Use API again to get Booster Version, Launch Site, Payload and Core and store them in global lists

Construct a dataset using the data obtained

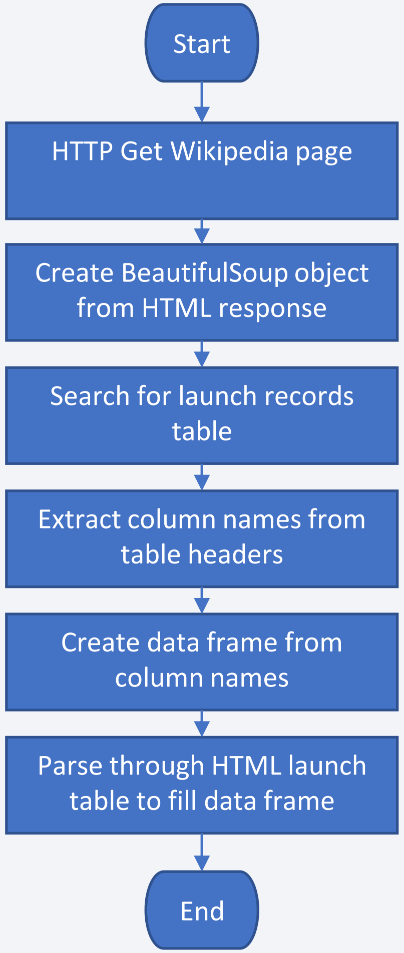
Filter the data for only Falcon 9 launches

Replace payload mass missing values with payload mass mean

End

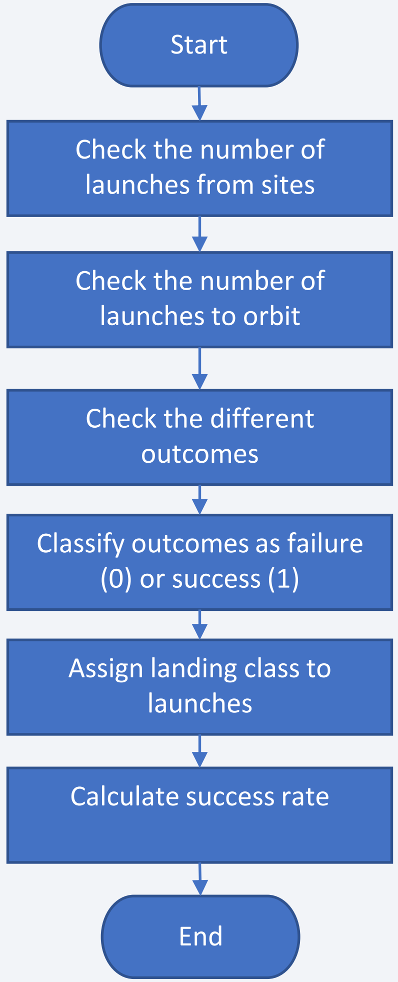
Data Collection - Scraping

- Use HTTP Get request to get an HTML from the Wikipedia page
- Create a BeautifulSoup object from the HTML response
- Search through the BeautifulSoup object for the launch records table
- Extract column names from table headers (<th> tag) and store in a list
- Create data frame from list of column names
- Parse through table to fill data frame
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/4beca6e9c635599814 87f69fa69ad2cf8eb6d95a/SpaceX_Falcon_9/Week%201%20Lab%202%20Data%20Collection%20with%20Web%20Scraping.ipynb


Start
HTTP Get Wikipedia page
Create BeautifulSoup object from HTML response
Search for launch records table
Extract column names from table headers
Create data frame from column names
Parse through HTML launch table to fill data frame
End

Data Wrangling

- To understand the data, we check the number of launches on each site, the number of launches to each orbit, and the number of different types of outcomes.
- There are 8 different types of outcomes, but we are only interested in whether the outcome is a success or a failure.
- Create a set of the bad outcomes.
- Assign a landing class of 0 if the outcome is in the bad outcomes set and a landing class of 1 otherwise.
- Add the landing class as a column at the end of the data set.
- We can now find the average of the landing class to determine the success rate.
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/4beca6e9c63559981487f69fa69ad2cf8eb6d95a/SpaceX_Falcon_9/Week%201%20Lab%203%20EDA.ipynb



Start
Check the number of launches from sites
Check the number of launches to orbit
Check the different outcomes
Classify outcomes as failure (0) or success (1)
Assign landing class to launches
Calculate success rate
End

EDA with Data Visualization

- A flight number vs payload mass scatter plot overlaid with the outcome was plotted to see the effect of the payload mass and the flight number on the outcome.
- A flight number vs launch site scatter plot overlaid with the outcome was plotted to see the effect of the launch site on the outcome.
- A payload mass vs launch site scatterplot overlaid with the outcome was plotted to see the correlation between the mass, the site and the outcome.
- A bar chart for the success rate of each orbit was plotted to see the relationship between the orbit and the outcome.
- A flight number vs orbit scatter plot overlaid with the outcome was plotted to see the relationship between the flight number and the orbit type.
- A payload mass vs orbit scatter plot overlaid with the outcome was plotted to see the relationship between the payload and orbit.
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/4beca6e9c63559981487f69fa69ad2cf8eb6d95a/SpaceX_Falcon_9/Week%202%20Lab%202%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Displayed the names of unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string ‘CCA’
- Displayed the total payload mass carried by the boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster version which carried the maximum payload
- Listed the failed landing outcomes in drone ships, their booster version, and launch site names in 2015
- Ranked the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/4beca6e9c63559981487f69fa69ad2cf8eb6d95a/SpaceX_Falcon_9/Week%202%20Lab%201%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

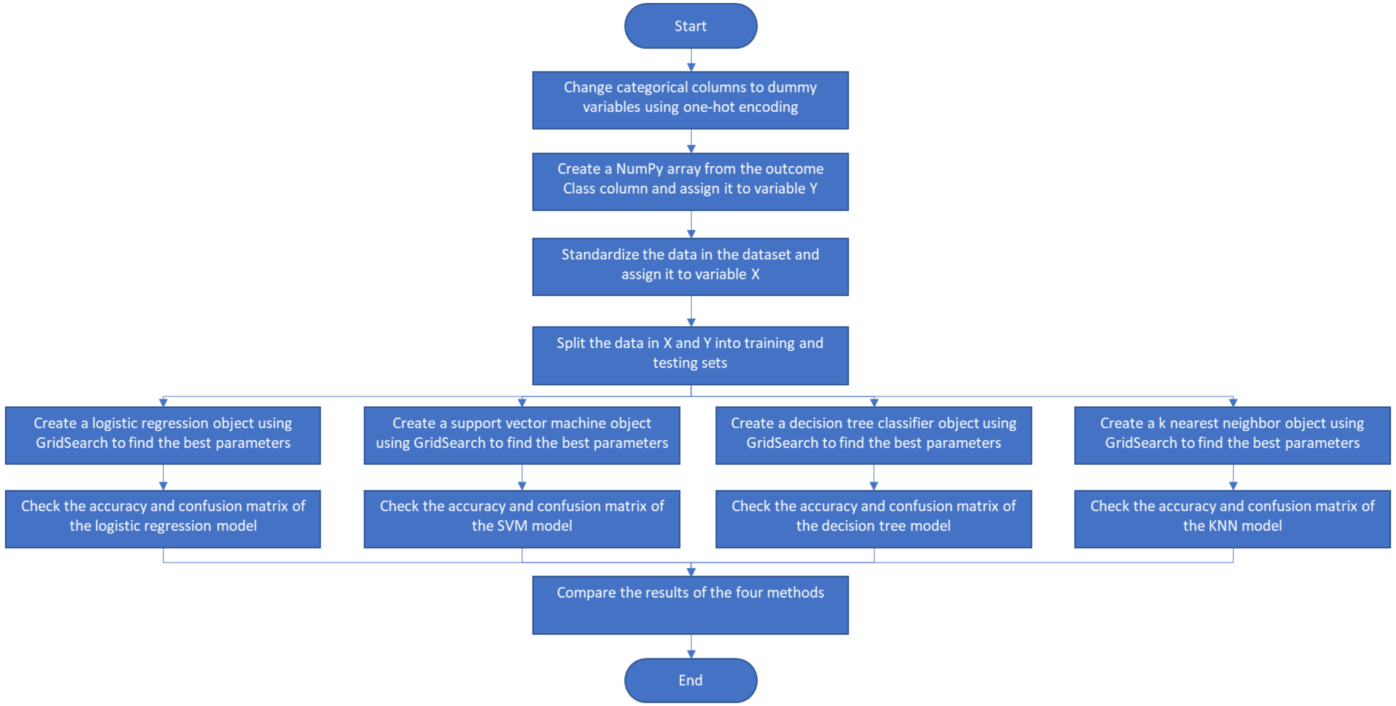
- A circle was added to show the location of the NASA Johnson Space Center in Houston, Texas.
- Markers were added to the USA map to show the locations of the launch sites.
- Cluster markers were added at the launch sites to indicate the successful and failed launches.
- A line was added to show the distance between KSC LC-39A and a nearby railway.
- A line was added to show the distance between CCAFS SLC-40 and the nearby coastline.
- These indicators illustrate the proximity of the launch centers to their surroundings.
- https://nbviewer.jupyter.org/github/omarjisr/IBM-Data-Science-Capstone-Project/blob/51d3b3906be001bb481c0560d6f0c5c42164b4a5/SpaceX_Falcon_9/Week%203%20Lab%201%20Interactive%20Visual%20Analytics%20Folium.ipynb

Build a Dashboard with Plotly Dash

- We added a dropdown list containing the names of all launch sites. The list determines what information the plots show.
- A pie chart was added to show the success/failure rates for the selected launch site, or if all sites were selected shows the successes of all sites.
- A payload mass vs outcome class scatter plot overlaid with booster version was added to show the correlation between the payload mass and the booster version and how they affect the outcome.
- A range slider was added to allow for zooming into a specific range of payload mass to allow for better visibility of the points on the chart.
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/51d3b3906be001bb481c0560d6f0c5c42164b4a5/SpaceX_Falcon_9/spacex_dash_app.py

Predictive Analysis (Classification)

- Split the one-hot encoded data set into training and test sets
- Use the training set to train four models: logistic regression, support vector machine, decision tree classifier, and k nearest neighbor
- Test the models using the testing set
- Compare the accuracies and confusion matrices of all four models to determine the best one.
- https://github.com/omarjisr/IBM-Data-Science-Capstone-Project/blob/51d3b3906be001bb481c0560d6f0c5c42164b4a5/SpaceX_Falcon_9/Week%204%20Lab%201%20Machine%20Learning%20Prediction.ipynb



Start

Change categorical columns to dummy variables using one-hot encoding

Create a NumPy array from the outcome Class column and assign it to variable Y

Standardize the data in the dataset and assign it to variable X

Split the data in X and Y into training and testing sets

Create a logistic regression object using GridSearch to find the best parameters

Create a support vector machine object using GridSearch to find the best parameters

Create a decision tree classifier object using GridSearch to find the best parameters

Create a k nearest neighbor object using GridSearch to find the best parameters

Check the accuracy and confusion matrix of the logistic regression model

Check the accuracy and confusion matrix of the SVM model

Check the accuracy and confusion matrix of the decision tree model

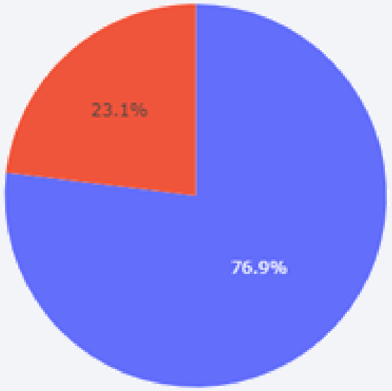
Check the accuracy and confusion matrix of the KNN model

Compare the results of the four methods

End

Results

- Exploratory data analysis shows:
  - Earlier flights had lower success rates for all launch sites
  - Launches from CCAFS SLC 40 have a lower success rate than other launch sites
  - Launches from KSC LC 39A tend to fail with payloads around 6000 kg
  - Launches to ES-L1, GEO, HEO, SSO and VLEO have higher success rates than other orbits
  - With a few exceptions, the success rate for each orbit is not related to the flight number
  - Heavy payloads fail for GTO orbit, but the reverse is true for ISS, LEO and PO orbits
  - The success rates tend to increase since 2013
  - KSC LC-39A has the highest successful launch rate at 76.9% and its failures are in payloads that are more than 5500 kg
  - The total payload mass launched by NASA (CRS) is 45,596 kg

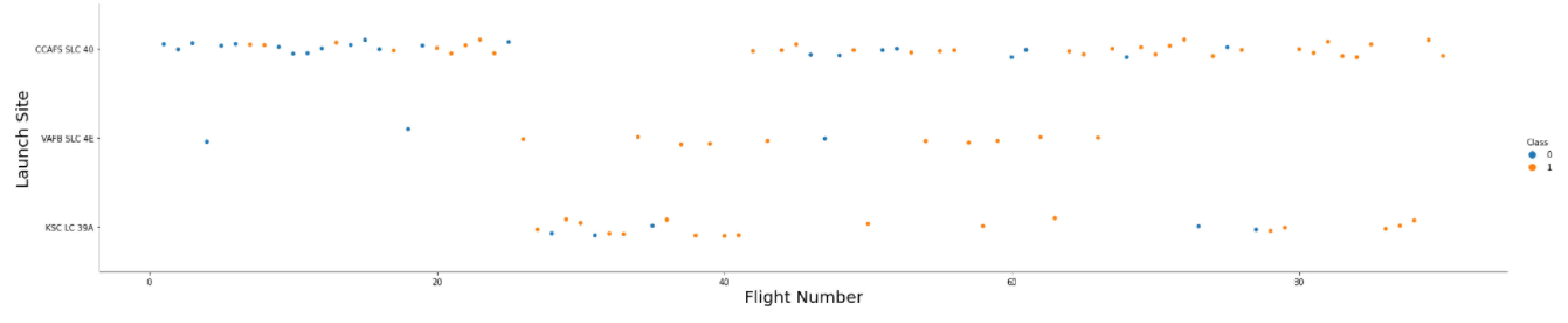


- Predictive analysis shows that due to the small size of the dataset, all four models (logistic regression, support vector machine, decision tree classifier, and k nearest neighbor) have the same accuracy score of 83.3% with the problem being false positives.

Section 2

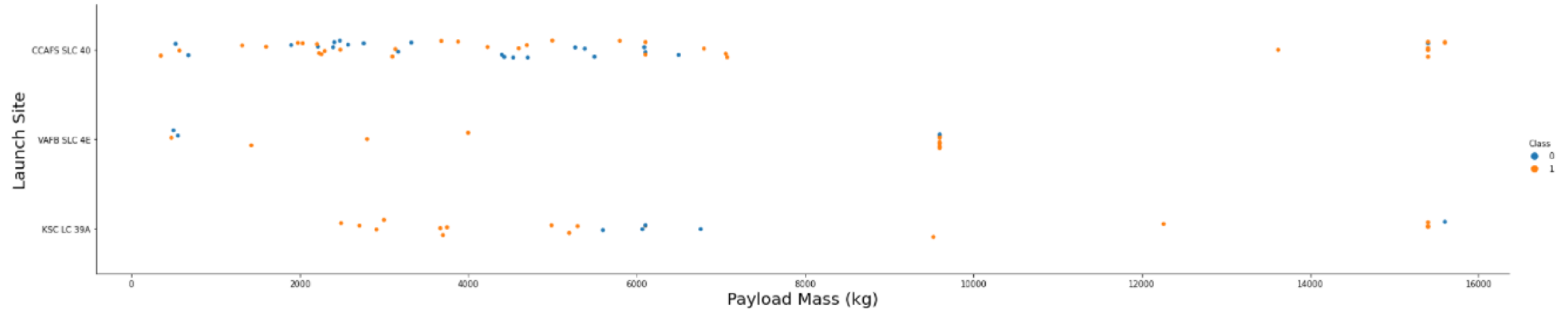
Insights drawn from EDA

Flight Number vs. Launch Site



- Failures are shown in blue, successes in orange.
- Earlier flights at each launch site tend to have more failures than later flights.
- The rate of success for KSC LC 39A is the highest while for CCAFS SLC 40 it is the lowest.

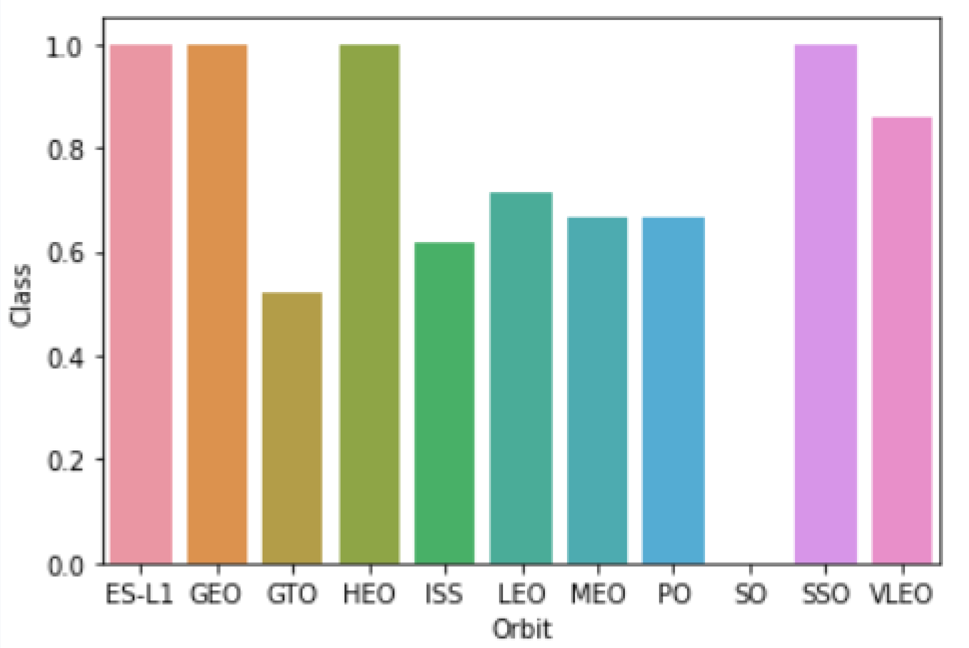
Payload vs. Launch Site



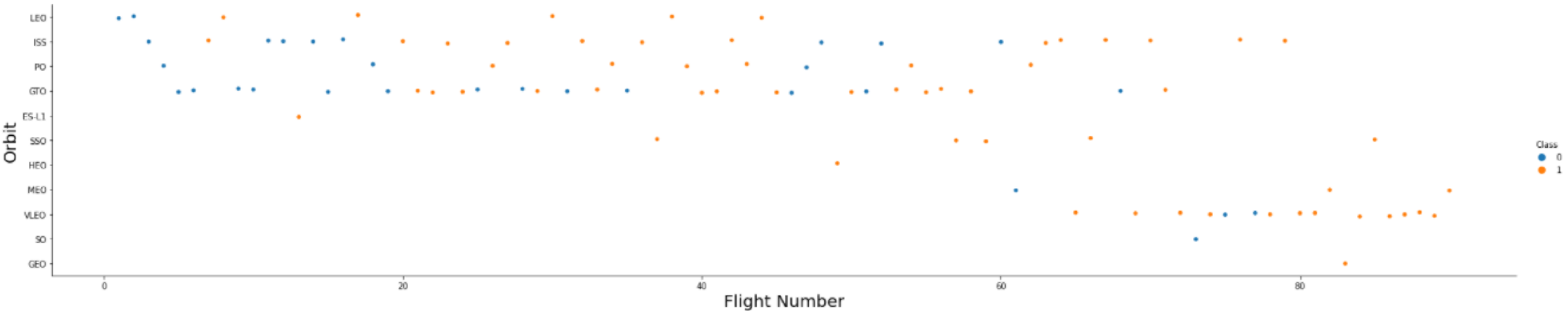
- Lower payloads are problematic at VAFB SLC 4E.
- Launches at KSC LC 39A only show problems around the 5500 kg to 7000 kg range.
- Outcome of launches from CCAFS SLC 40 mostly is not affected by payload mass.

Success Rate vs. Orbit Type

- Launches to ES-L1, GEO, HEO and SSO have a perfect success track record.
- Launches to VLEO have a high success rate.
- Launches to SO have no successes.
- Launches to other orbits have a success rate ranging around 50% to 70%.

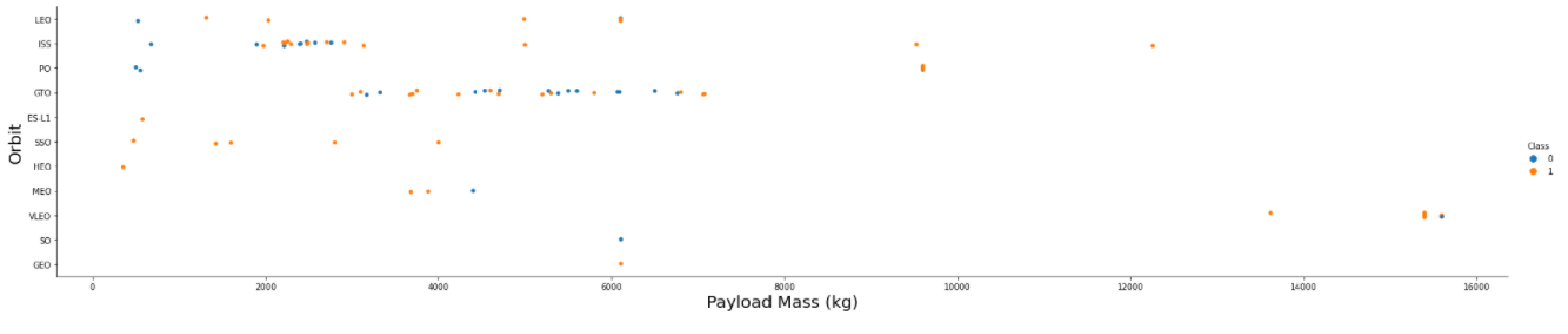


Flight Number vs. Orbit Type



- There seems to be a correlation between the flight number and the success of the flights for orbits LEO, PO, and VLEO.
- There doesn't seem to be a correlation between the flight number and the success of the launches for orbits ISS and GTO.

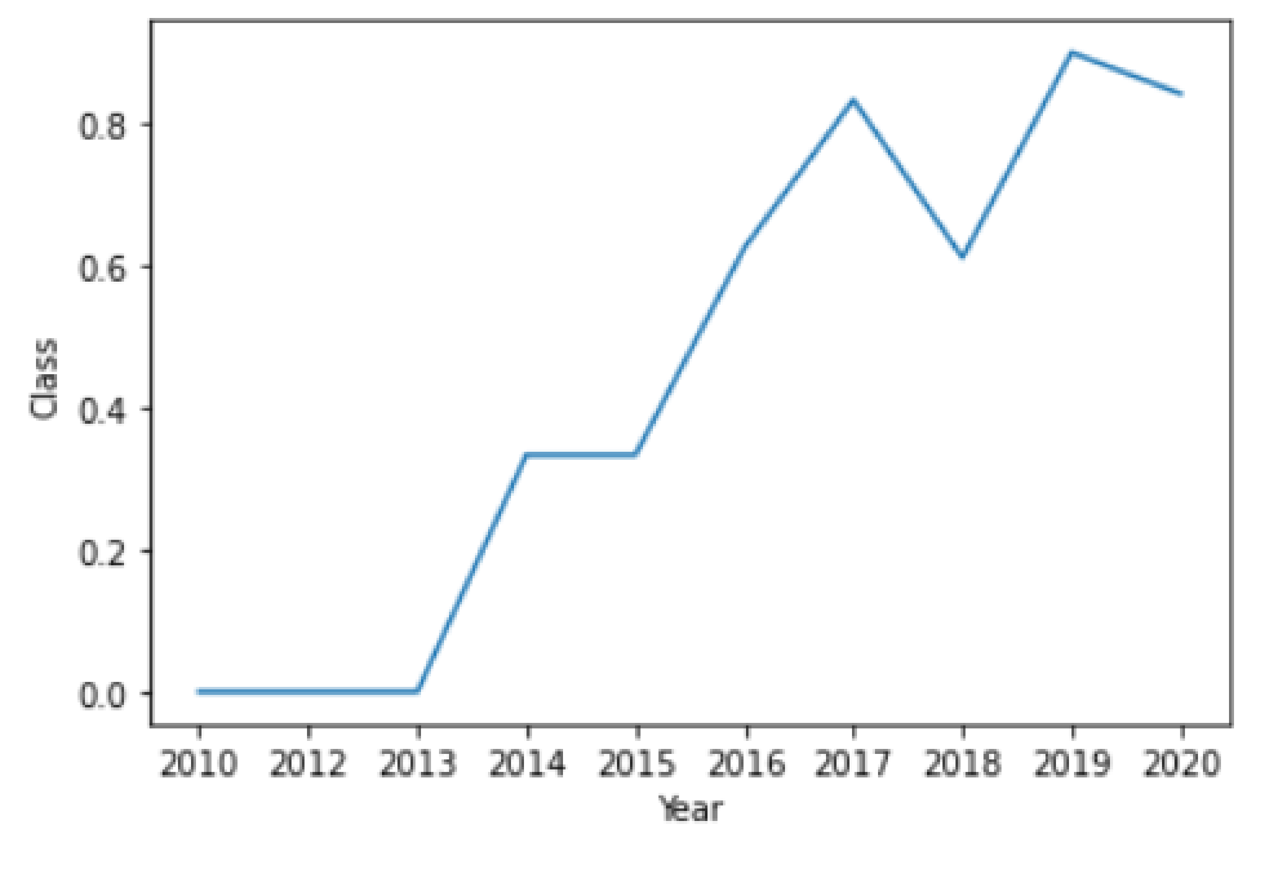
Payload vs. Orbit Type



- Heavy payloads seem to have a negative impact on GTO orbits.
- Heavy payloads seem to have a positive impact on LEO, ISS and PO orbits.

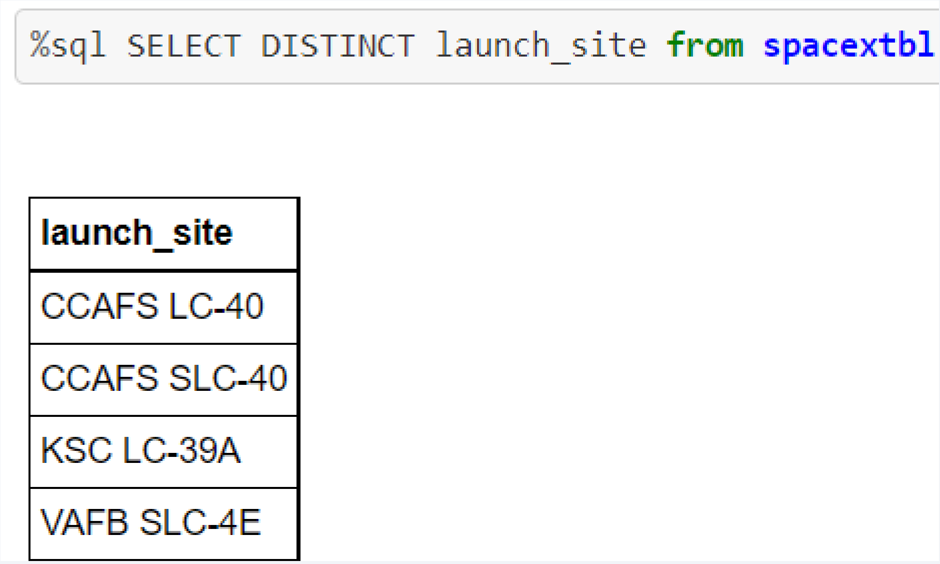
Launch Success Yearly Trend

- The Year vs Outcome linear trend shows that the success rate kept on increasing from 2013 to 2020 with an exception in 2018 and a slight dip in 2020.



All Launch Site Names

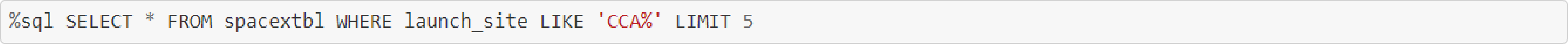
- Using a SQL query, we can find the names of the unique launch sites.
- There are four distinct launch sites:
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
    - VAFB SLC-4E



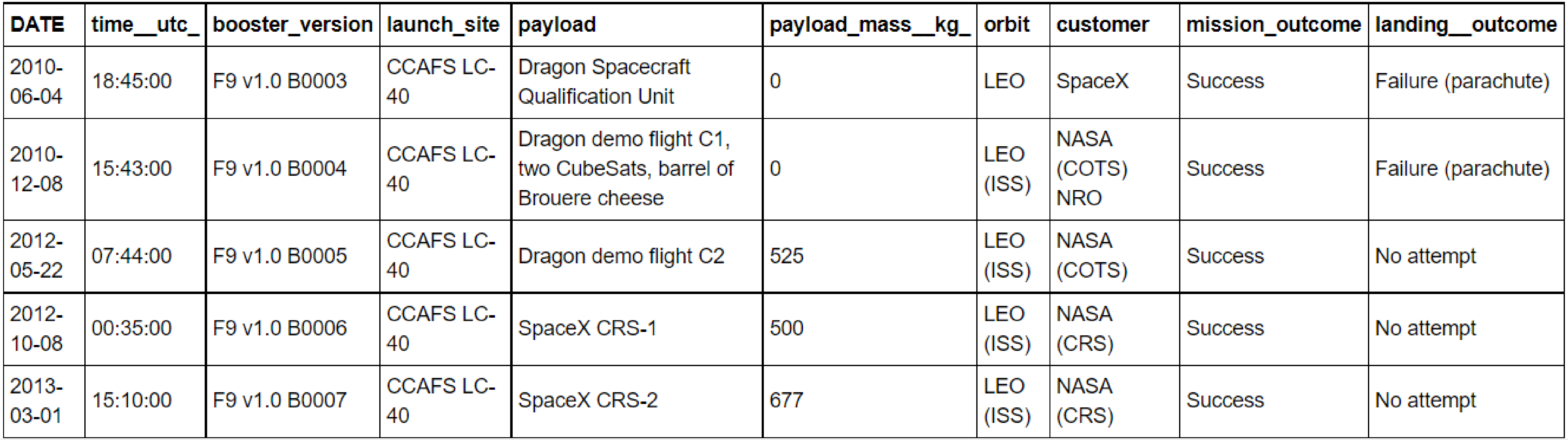
```
%sql SELECT DISTINCT launch_site from spacextbl
```

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

Launch Site Names Begin with 'CCA'



```
%sql SELECT * FROM spacextbl WHERE launch_site LIKE 'CCA%' LIMIT 5
```

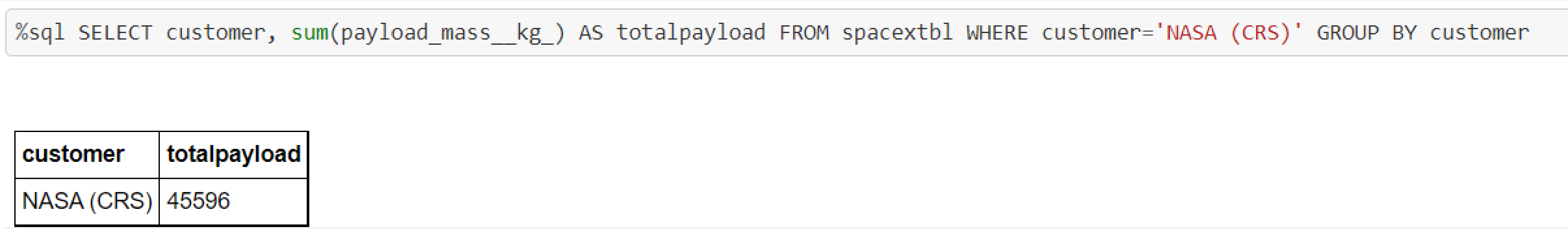


DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- The query result above shows 5 launches where the launch site name begins with ‘CCA’.

Total Payload Mass

- Using a simple SQL query, we can calculate the total payload carried by boosters from NASA.
- The total payload is 45,596 Kg.

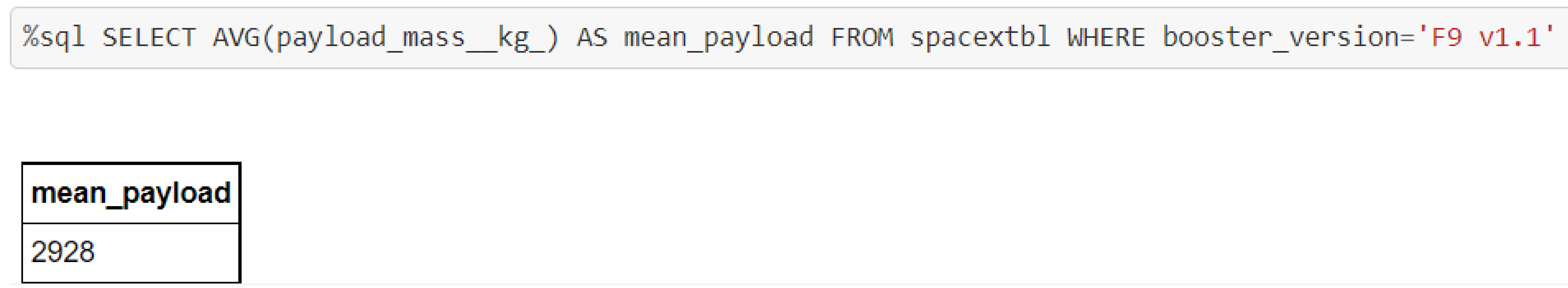


```
%sql SELECT customer, sum(payload_mass__kg_) AS totalpayload FROM spacextbl WHERE customer='NASA (CRS)' GROUP BY customer
```

customer	totalpayload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated using the below SQL query.
- The mean payload is 2,928 Kg.

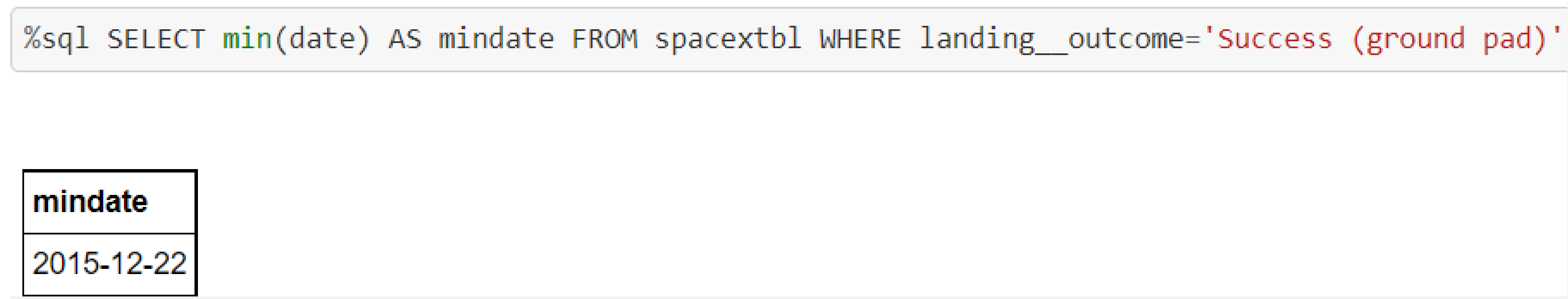


```
%sql SELECT AVG(payload_mass__kg_) AS mean_payload FROM spacextbl WHERE booster_version='F9 v1.1'
```

mean_payload
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 22 December 2015.

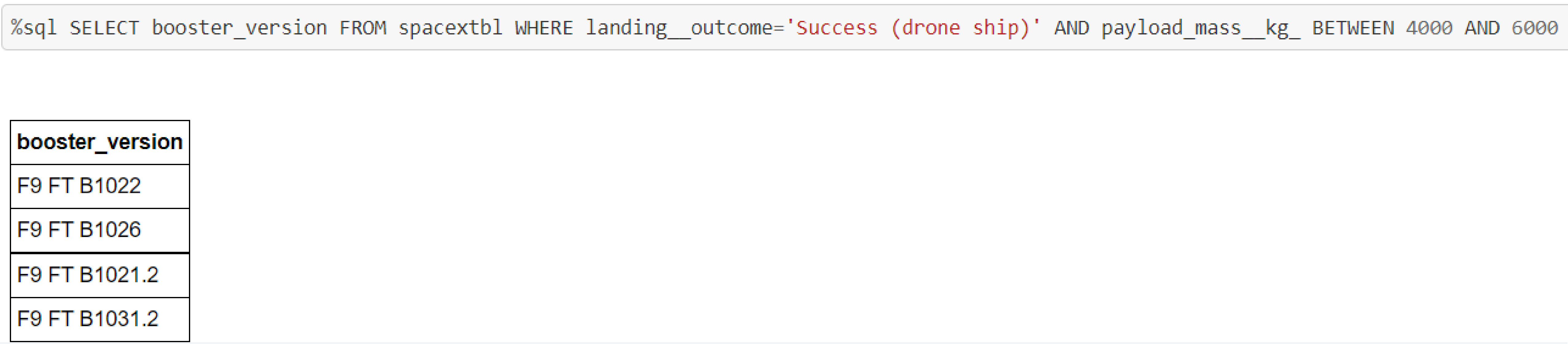


```
%sql SELECT min(date) AS mindate FROM spacextbl WHERE landing__outcome='Success (ground pad)'
```

mindate
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are shown in the SQL query result below.

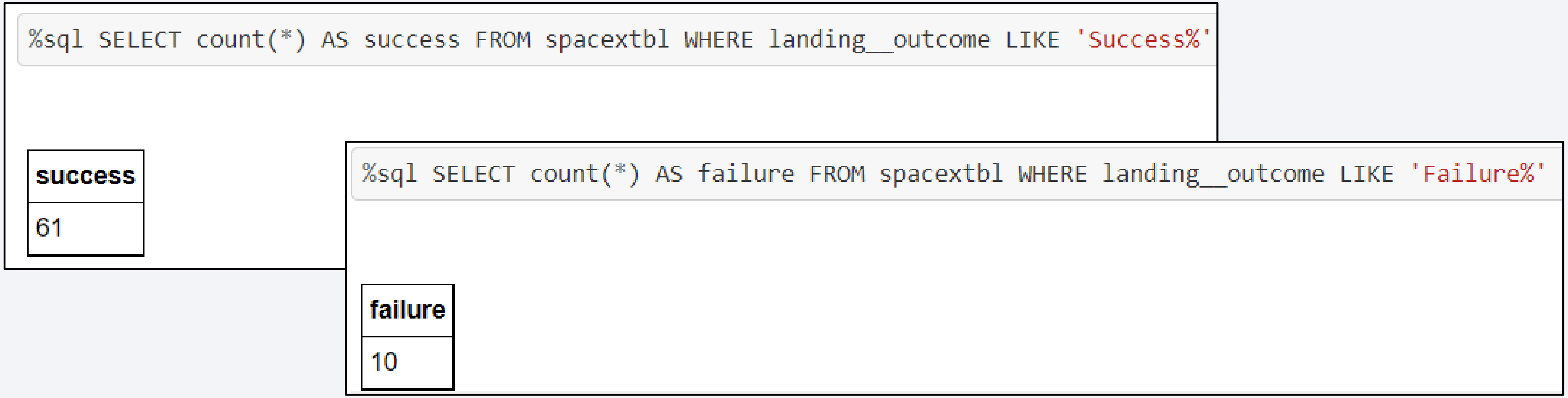


```
%sql SELECT booster_version FROM spacextbl WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes is 61.
- The total number of failure mission outcomes is 10.



```
%sql SELECT count(*) AS success FROM spacextbl WHERE landing__outcome LIKE 'Success%'
```

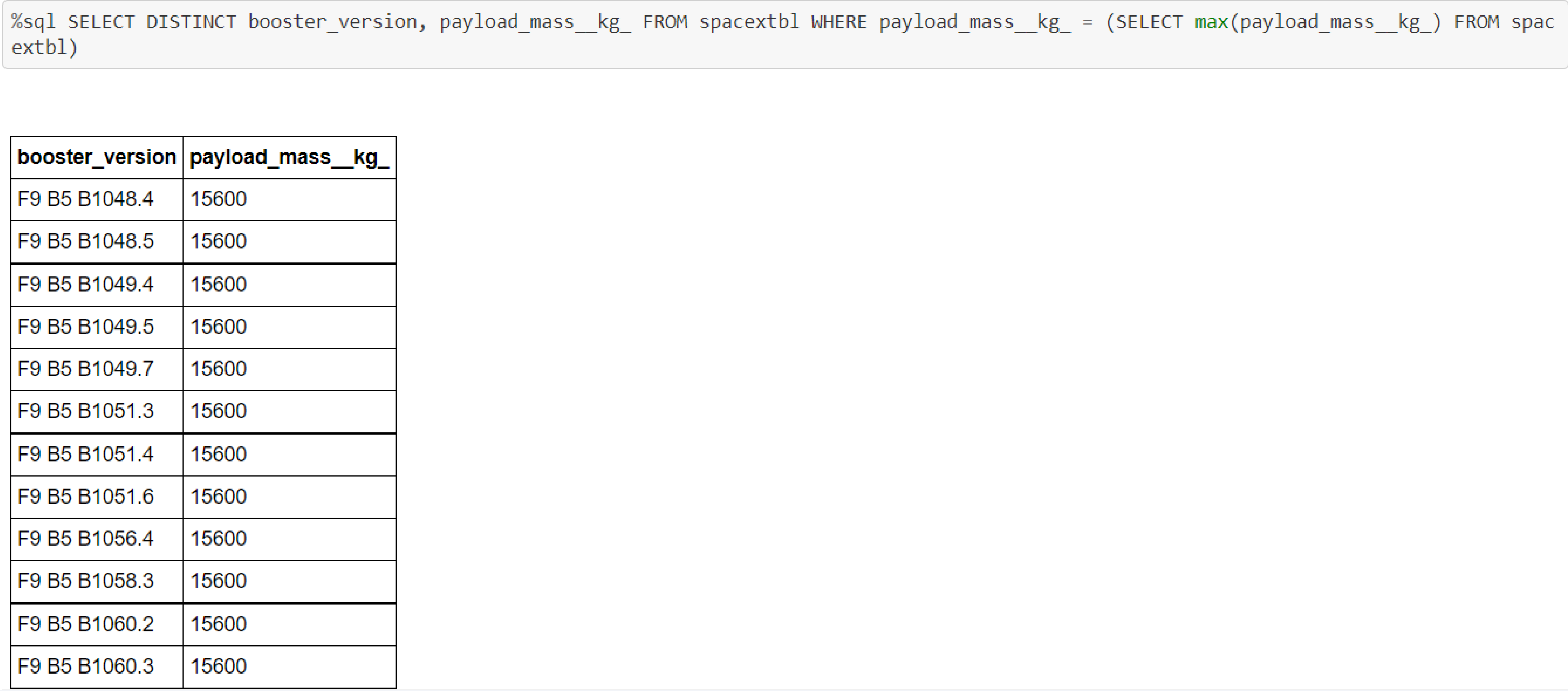
success
61

```
%sql SELECT count(*) AS failure FROM spacextbl WHERE landing__outcome LIKE 'Failure%'
```

failure
10

Boosters Carried Maximum Payload

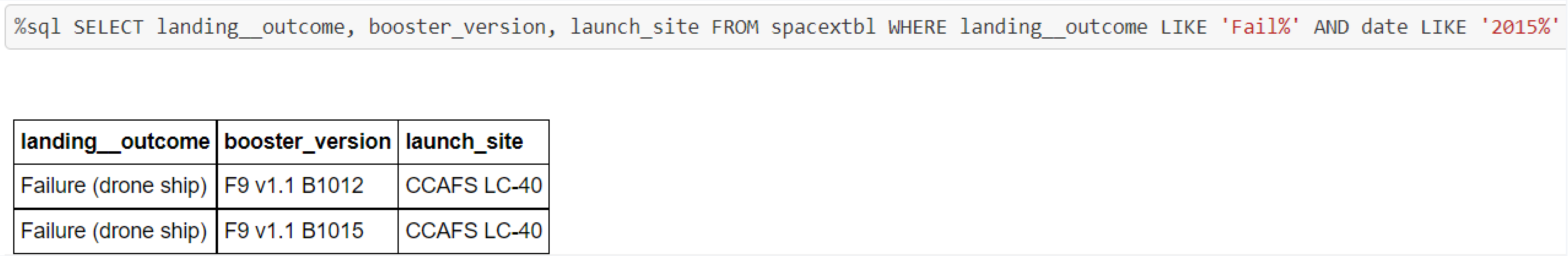
- The names of the boosters which have carried the maximum payload mass are listed below.



```
%sql SELECT DISTINCT booster_version, payload_mass__kg_ FROM spacextbl WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM spac
extbl)
```

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

2015 Launch Records



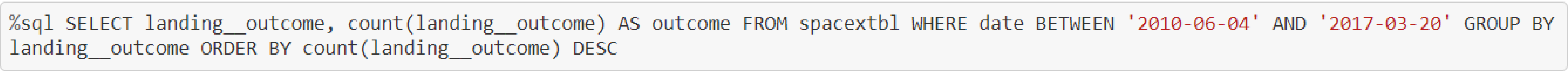
```
%sql SELECT landing__outcome, booster_version, launch_site FROM spacextbl WHERE landing__outcome LIKE 'Fail%' AND date LIKE '2015%'
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

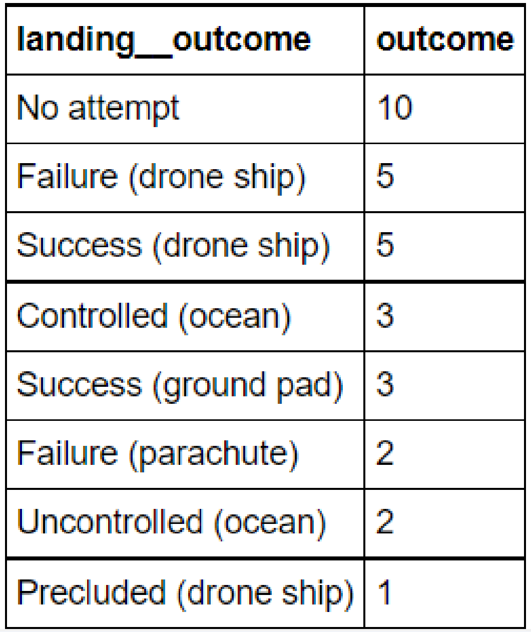
- The failed landing outcomes on drone ship, their booster versions, and launch site names for the year 2015 are listed above.

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

- A ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order is presented below.



```
%sql SELECT landing__outcome, count(landing__outcome) AS outcome FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY
landing__outcome ORDER BY count(landing__outcome) DESC
```

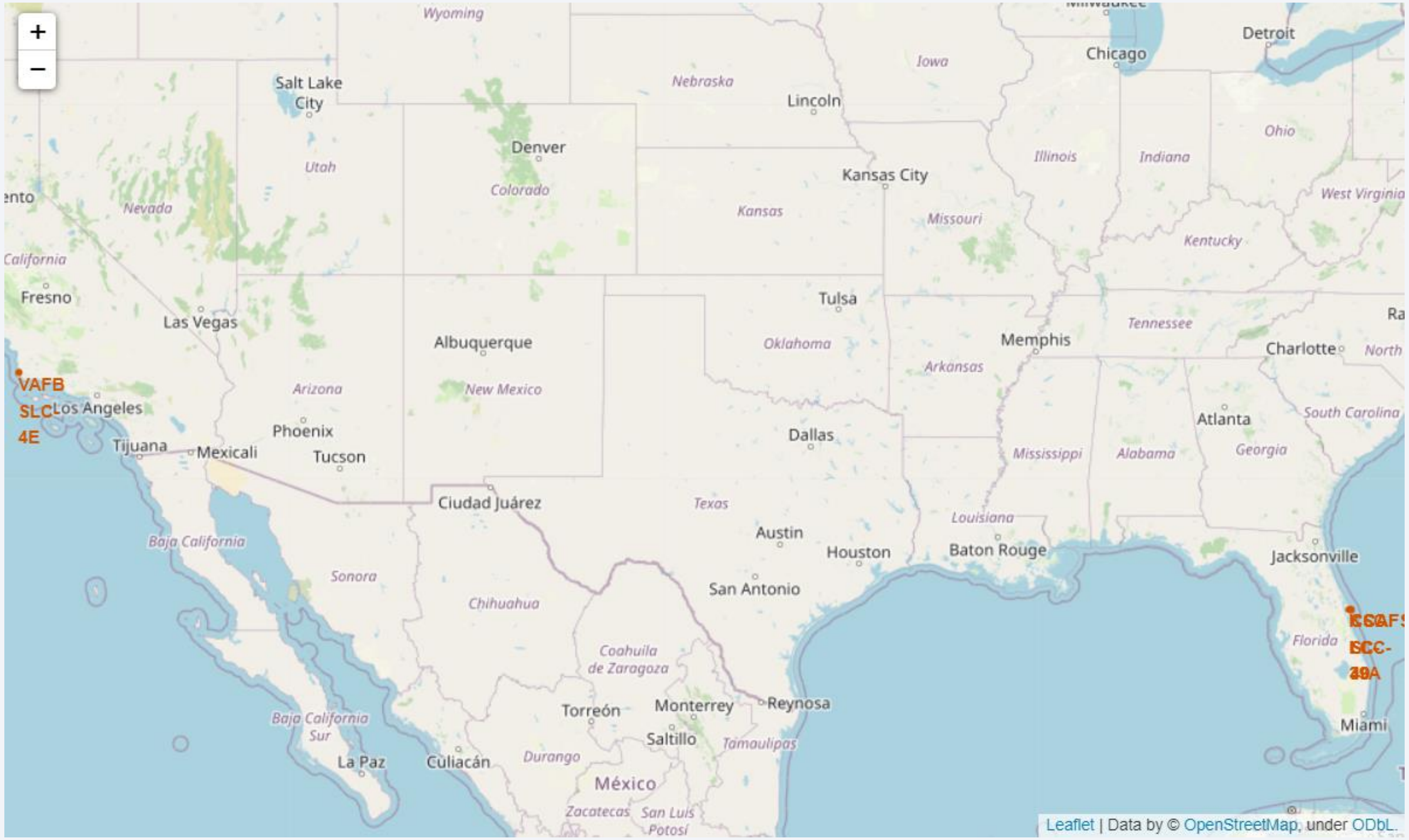


landing__outcome	outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

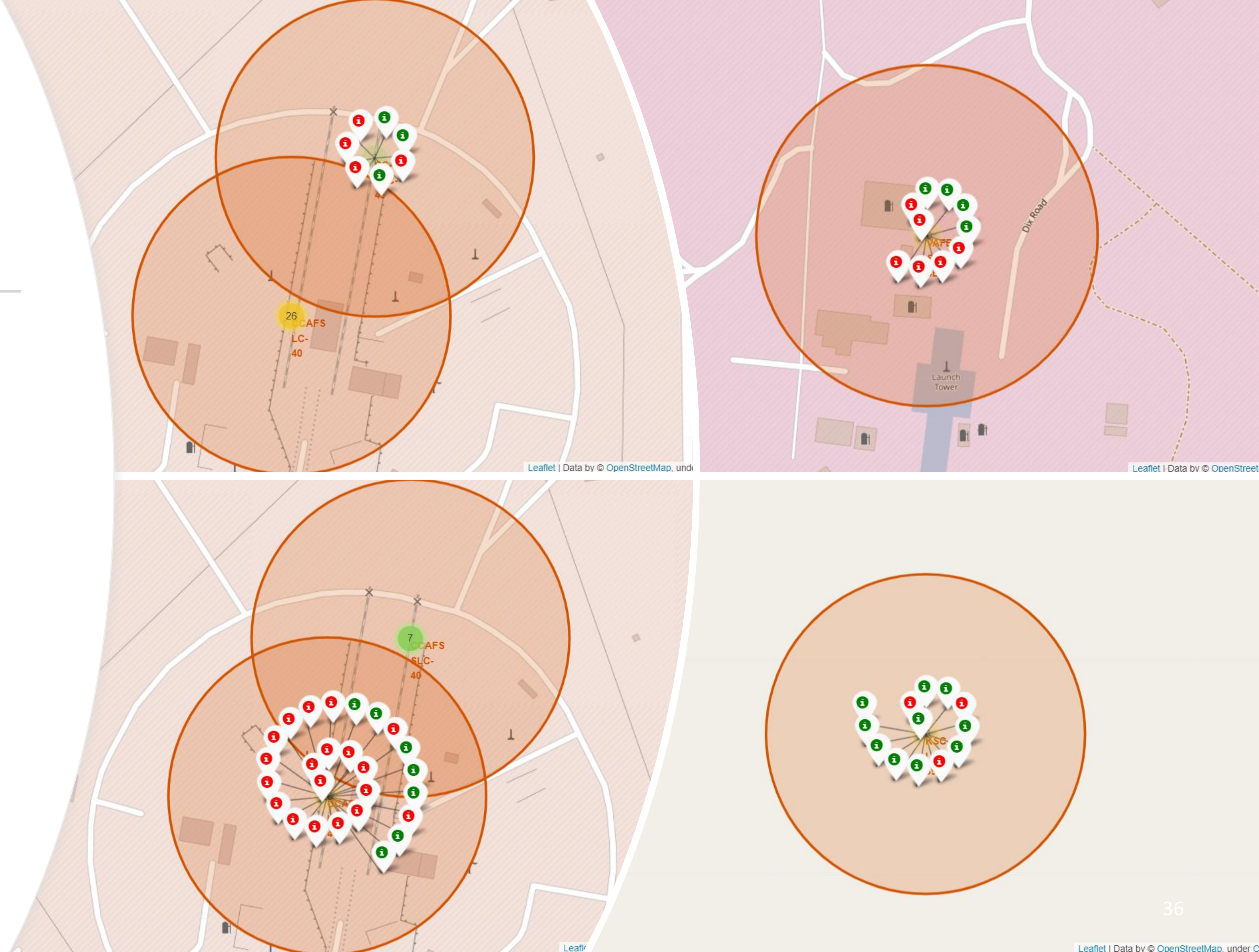
Launch Sites' Locations



- This folium map shows all launch sites' location markers on a global map.
- The launch sites shown are VAFB SLC-4E in California and CCAFS SLC-40, CCAFS LC-40, and KSC LC-39A in Florida.
- The map shows that the launch sites are all close to the ocean and as close as possible to the equator.

Launch Outcomes

- The cluster markers at the launch sites allow us to quickly see the successes and failures at each launch site.
- It is easy to see in that KSC LC-39A has a larger number of successes than the remaining launch sites.
- CCAFS LC-40 has the highest failure rate.

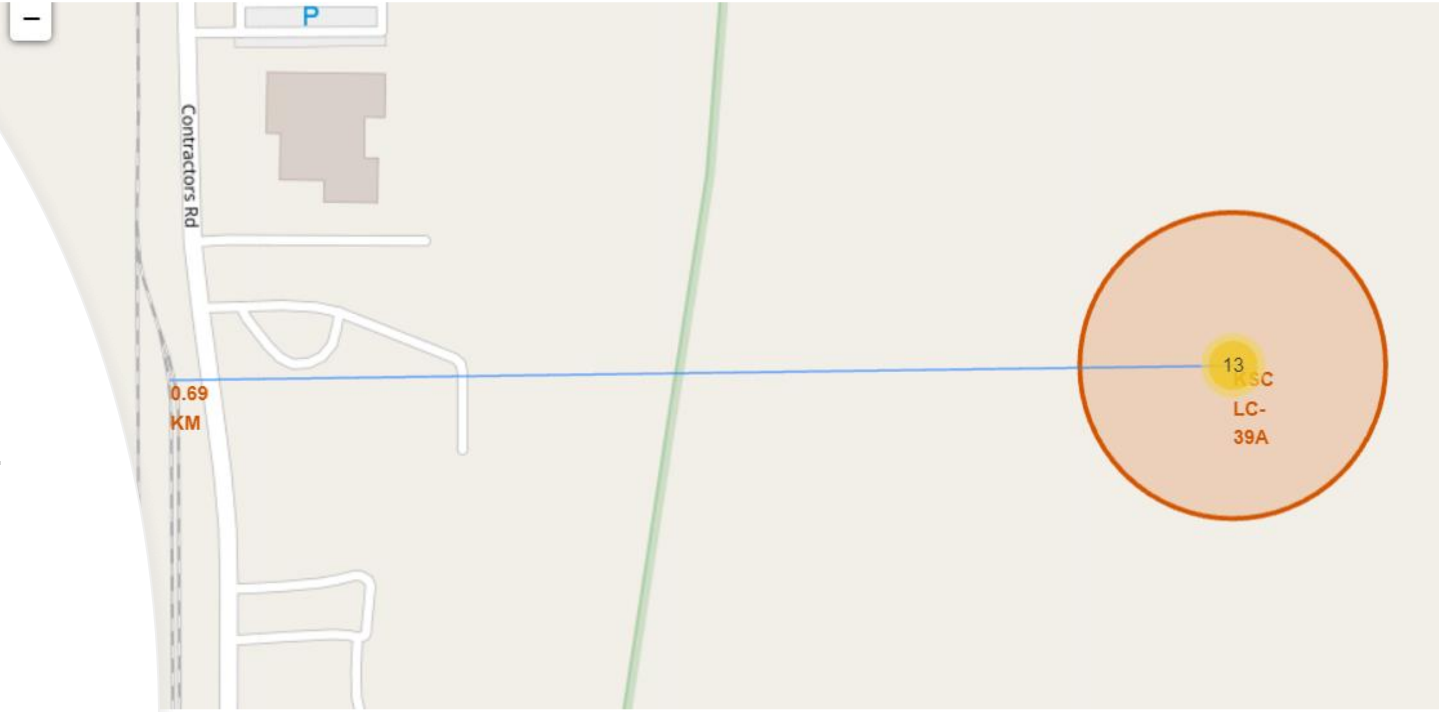

26
CCAFS
LC-
40
Leaflet | Data by © OpenStreetMap, und
VAFB
Dix Road
Launch
Tower
Leaflet | Data by © OpenStreet
7
CCAFS
SLC-
40
Leafl
KSC
Leaflet | Data by © OpenStreetMap, under C

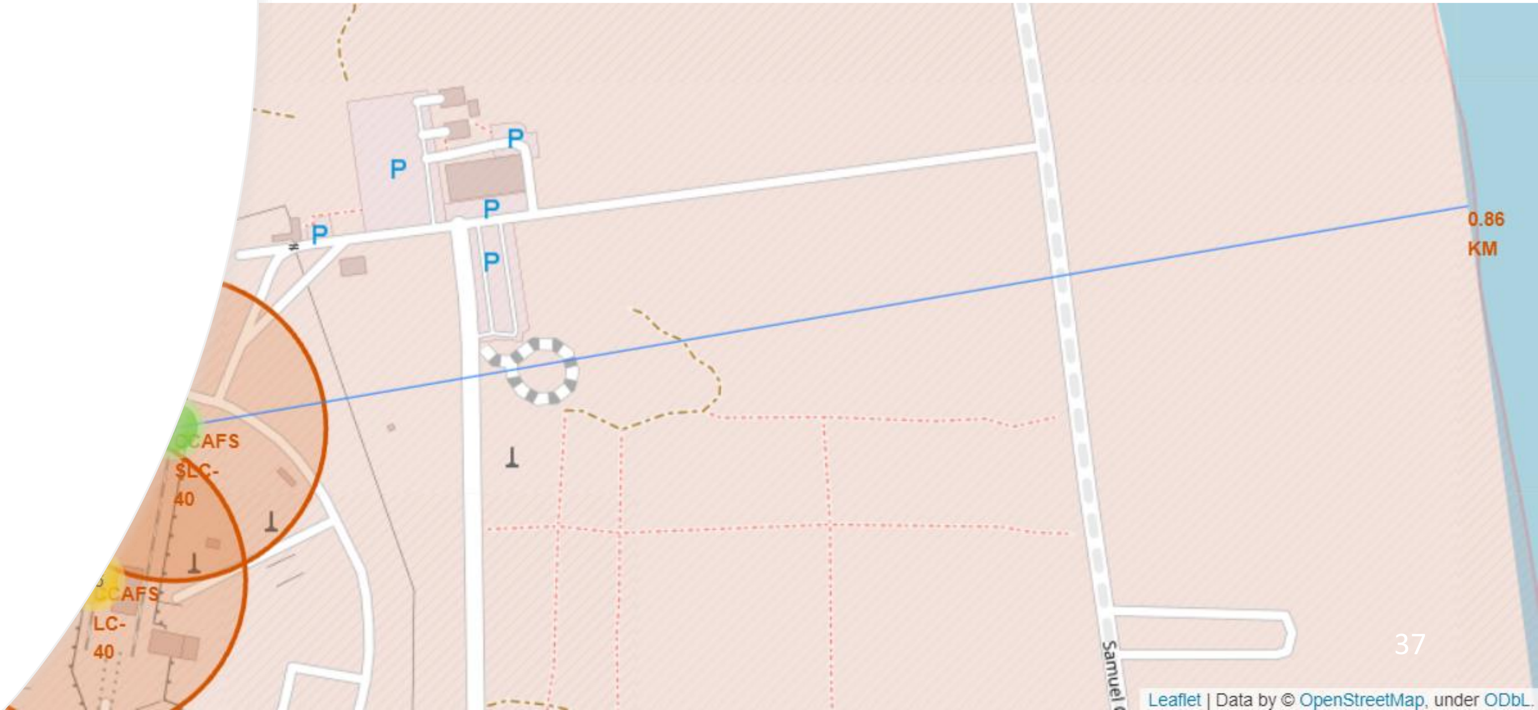
Launch Sites’ Proximity to Nearby Features

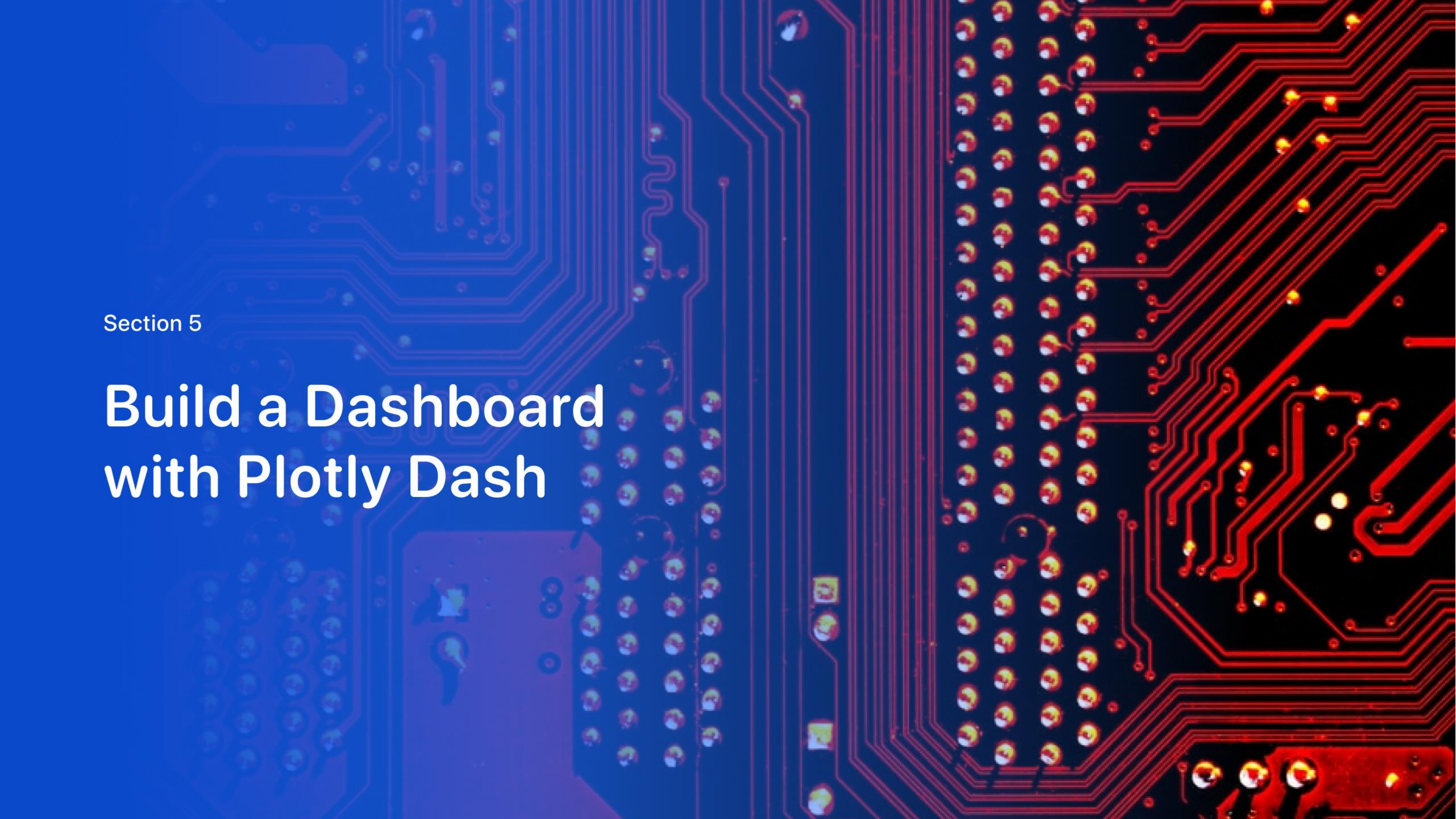
- These screenshots show some of the launch sites’ proximity to nearby features.

- KSC LC-39A is 690m away from a nearby railway.

- CCAFS SLC-40 is 860m away from the nearest coastline.

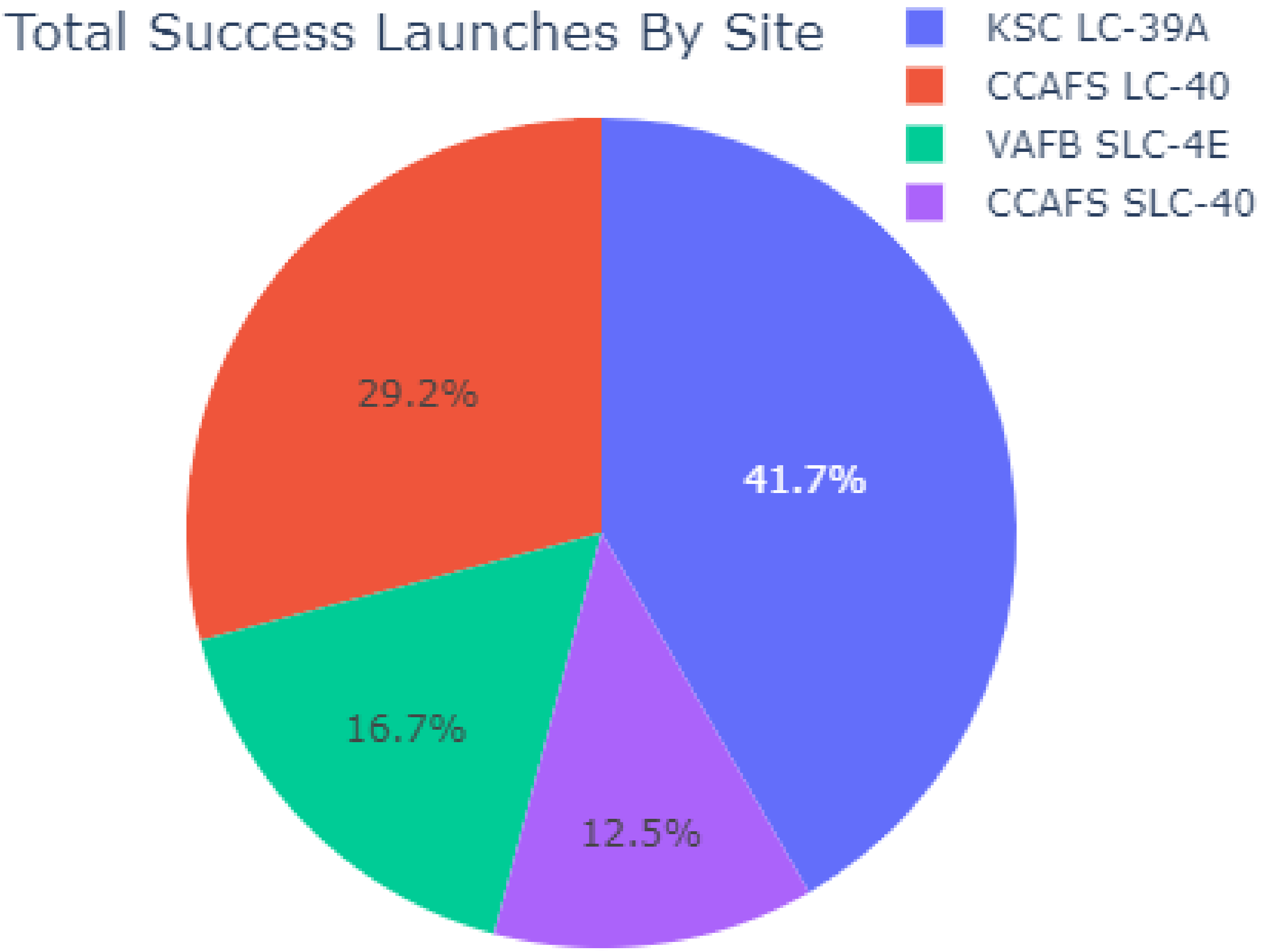






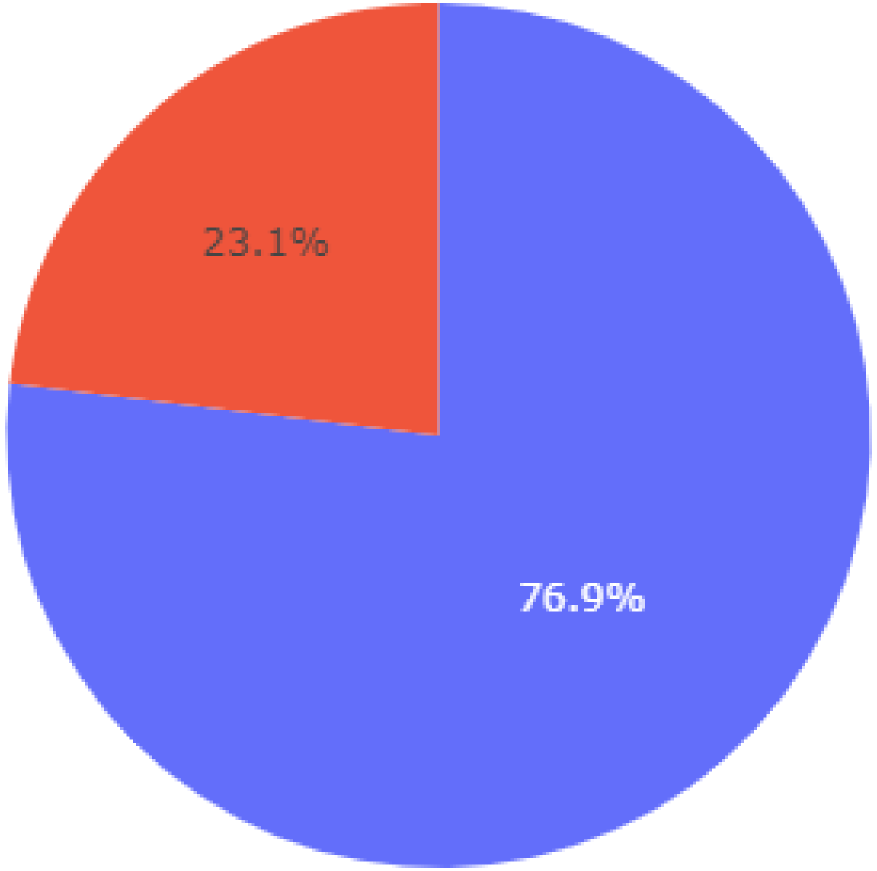
Section 5

Build a Dashboard with Plotly Dash



Launch Success Count for All Sites

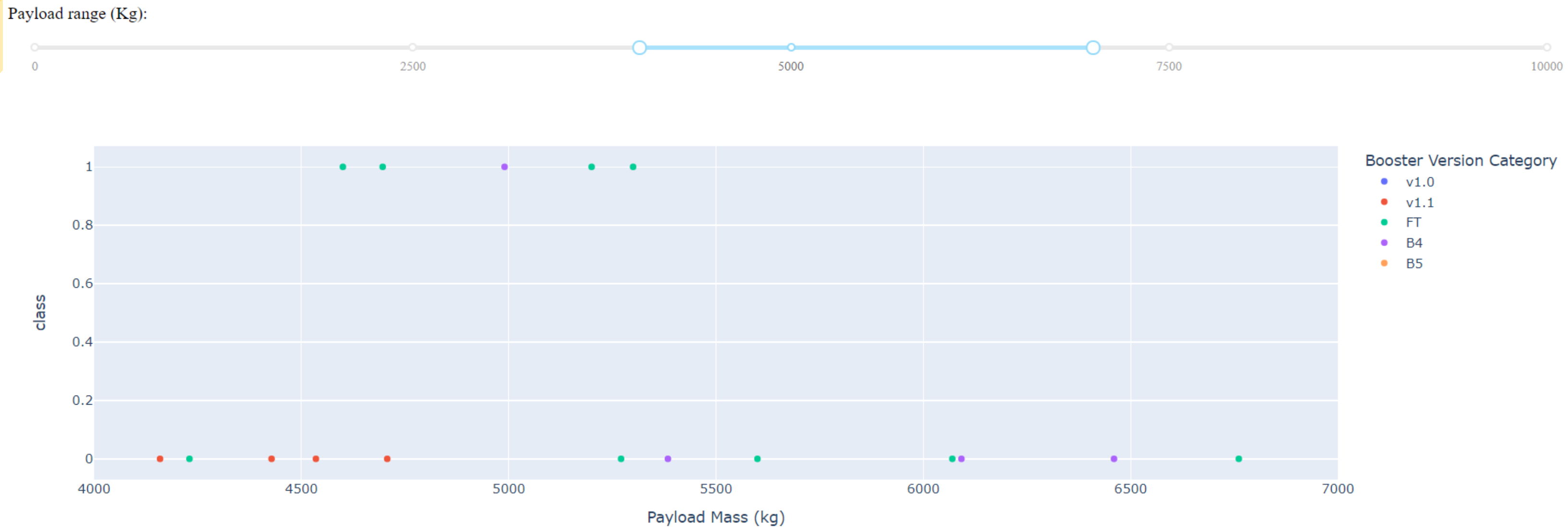
- KSC LC-39A has the highest success count among the four launch sites with 41.7% of all successful launches.
- CCAFS SLC-40 has the lowest success count among the four launch sites with only 12.5% of all successful launches.

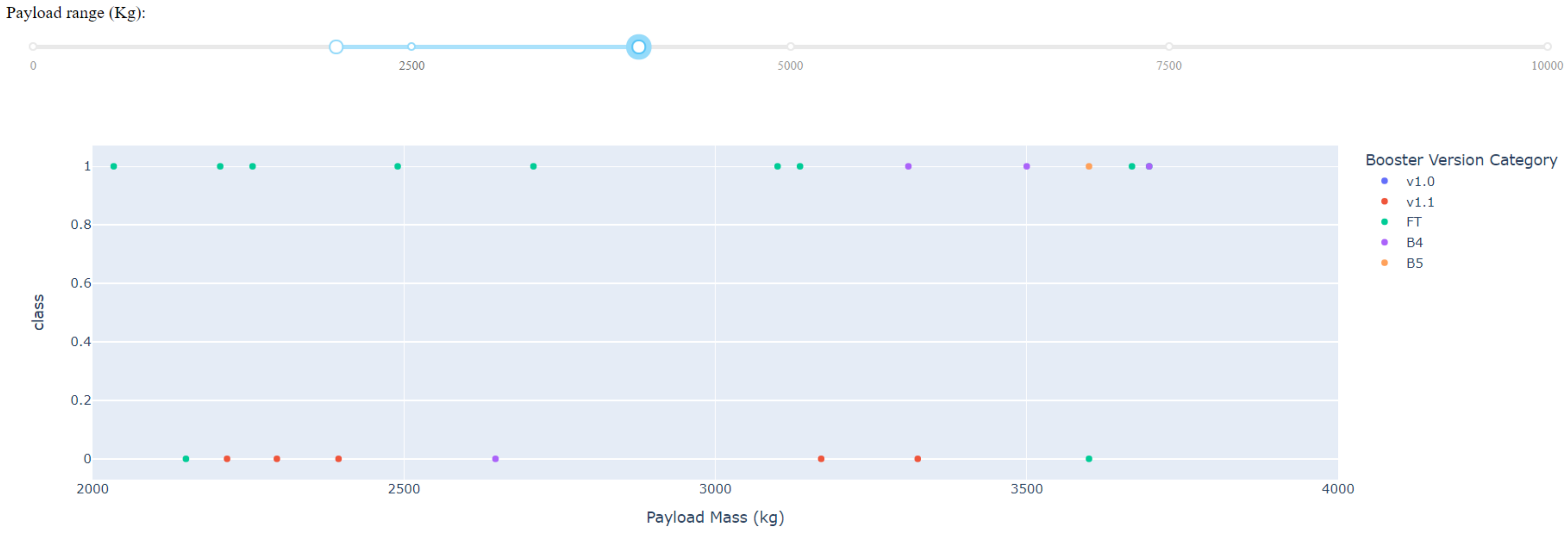


KSC LC-39A Success Rate

- KSC LC-39A has the highest success rate at 76.9%.

Relationship between Payload Mass and Launch Outcome





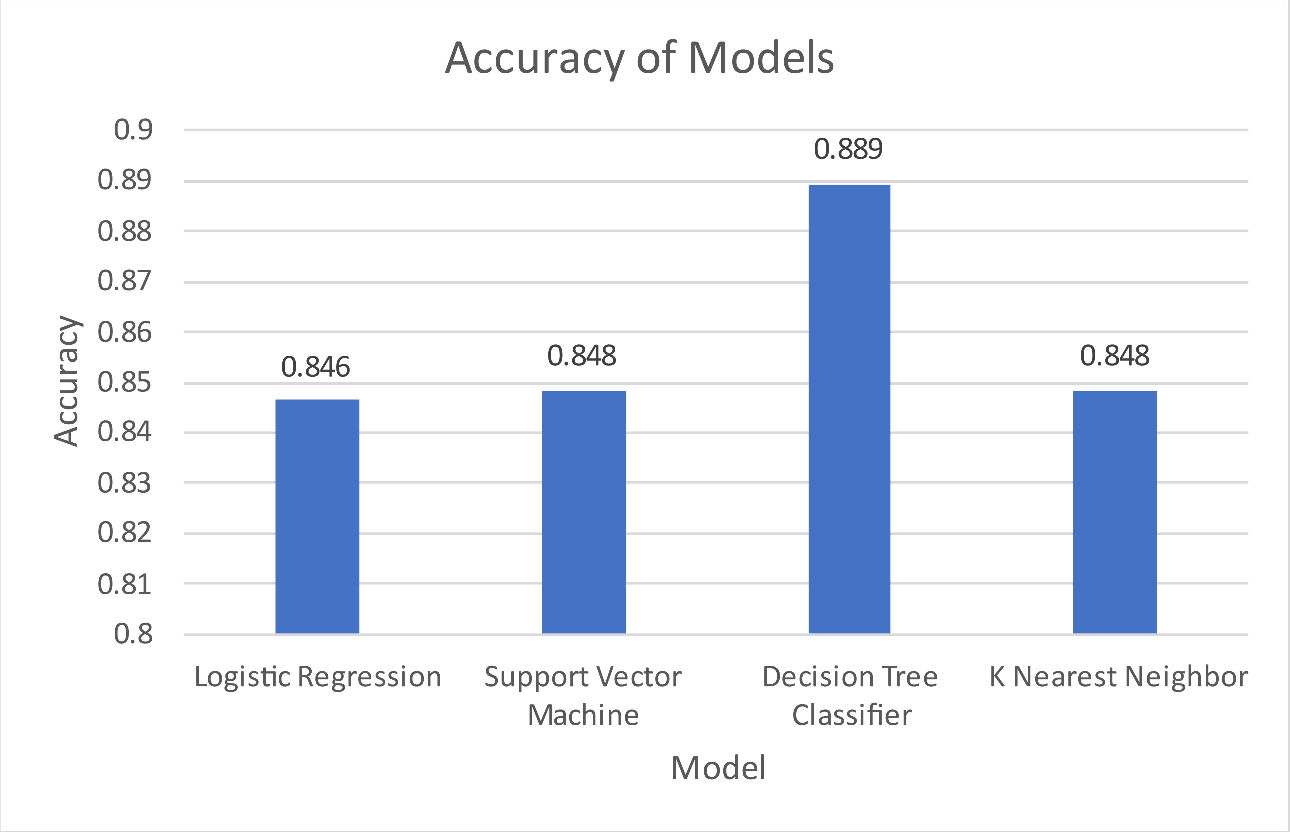
- The payload mass range with the highest success rate is between 1,900 Kg and 3,700 Kg.
- The payload mass range with the lowest success rate is between 4,100 Kg and 6,800 Kg.
- The FT booster version has the highest launch success rate.

Section 6

Predictive Analysis (Classification)

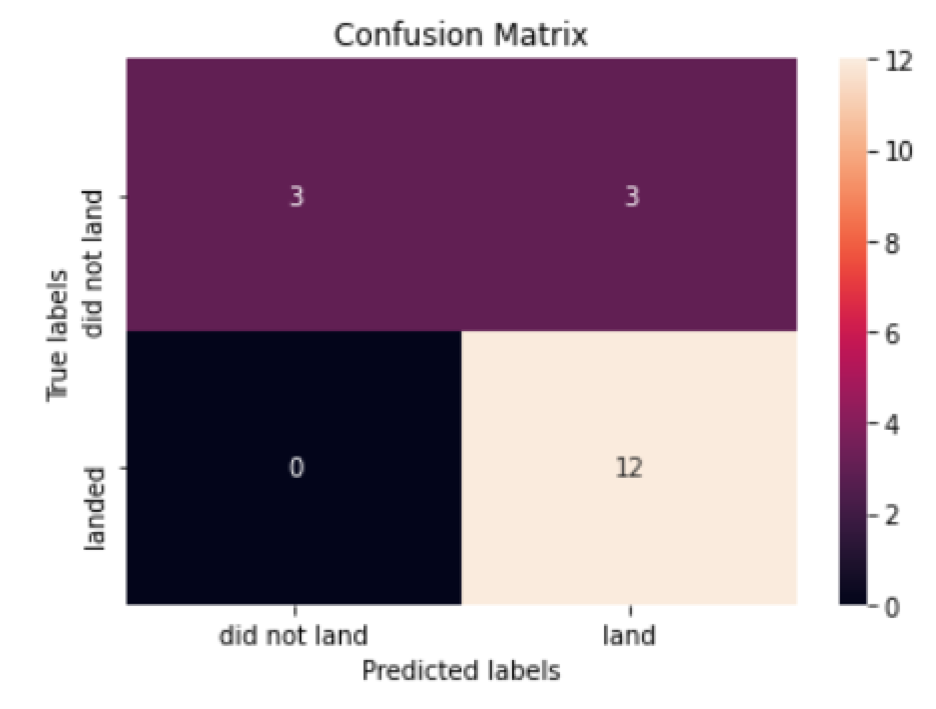
Classification Accuracy

- Four models are prepared to be used in classification of future launches.
- The Decision Tree Classifier model had the highest accuracy rate at 88.9%.
- The Logistic Regression model had the lowest accuracy rate at 84.6%.
- The Support Vector Machine and K Nearest Neighbor models are tied at 84.8%.



Confusion Matrix

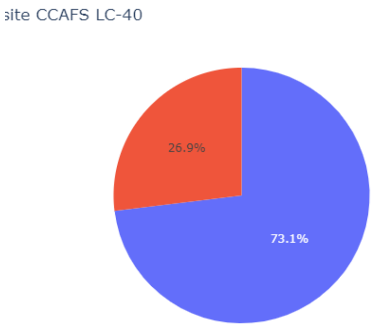
- The confusion matrix for a model plots the predicted labels generated by the model using a test sample against the true labels for the same sample.
- The confusion matrix for the Decision Tree Classifier model shows that the model generates some false positives (16.7% of the results) and no false negatives.

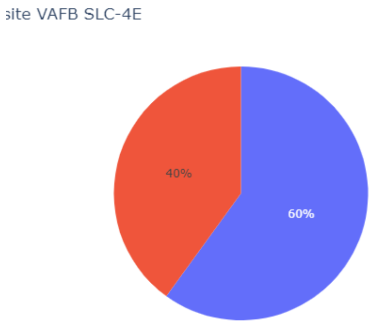


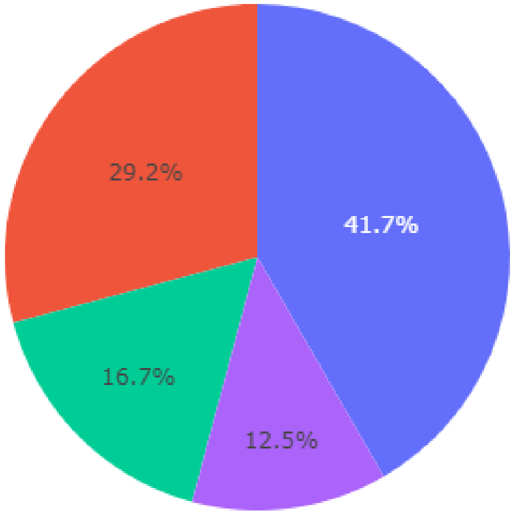
Conclusions

- The launch site with the most successes is KSC LC-39A.
- The orbits with the most successes are ES-L1, GEO, HEO and SSO.
- The payload mass range with the most successes is 1,900 Kg to 3,700 Kg.
- The booster version with the most successes is FT.
- The predictive model for future flights with the highest accuracy is the Decision Tree Classifier with the parameters:
  - Criterion: entropy
  - Max depth: 16
  - Max features: auto
  - Min samples leaf: 2
  - Min samples split: 10
  - Splitter: random

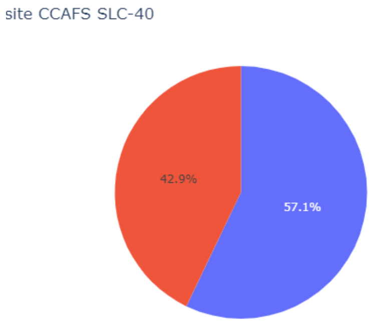
Appendix

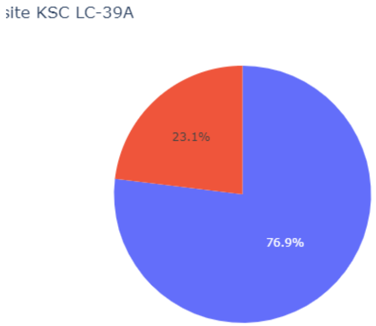

site CCAFS LC-40


site VAFB SLC-4E



- Pie charts for successes of all launch sites are shown on this slide.
- For other visualizations and code, please visit the links included in the Methodology section of this presentation.


site CCAFS SLC-40


site KSC LC-39A

Thank you!