
# Winning Space Race with Data Science

Cody Pewarchuk
12/26/2021

# Outline

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

# Executive Summary

- Question: Can we determine optimal conditions for recapturing rockets for aerospace start up SpaceY?
- We assume SpaceY would have similar stats to current company SpaceX.
- Data acquired from
  - Webscraping Wikipedia SpaceX launch data tables
  - SpaceX API
- Exploratory data analysis and transformation
- Used KNN to create model and to predict whether rockets would land successfully
- Trend was for better results from:
  - CCAFS SLC-40 launch site
  - 2000kg to 5000kg sized payloads
  - Results improved as more launches were performed
  - Orbits of ES-L1, GEO, HEO, SSO, and VLEO were most successful.

# Introduction

- Classification analysis aimed at determining whether rockets would land successfully or not.
- Are there factors that can predict a positive vs a negative outcome?
- What is the best launch site to use?
  - CCAFS LC-40
  - KSC LC-39A
  - VAFB SLC 4E
  - CCAFS SCL-40

Section 1

# Methodology

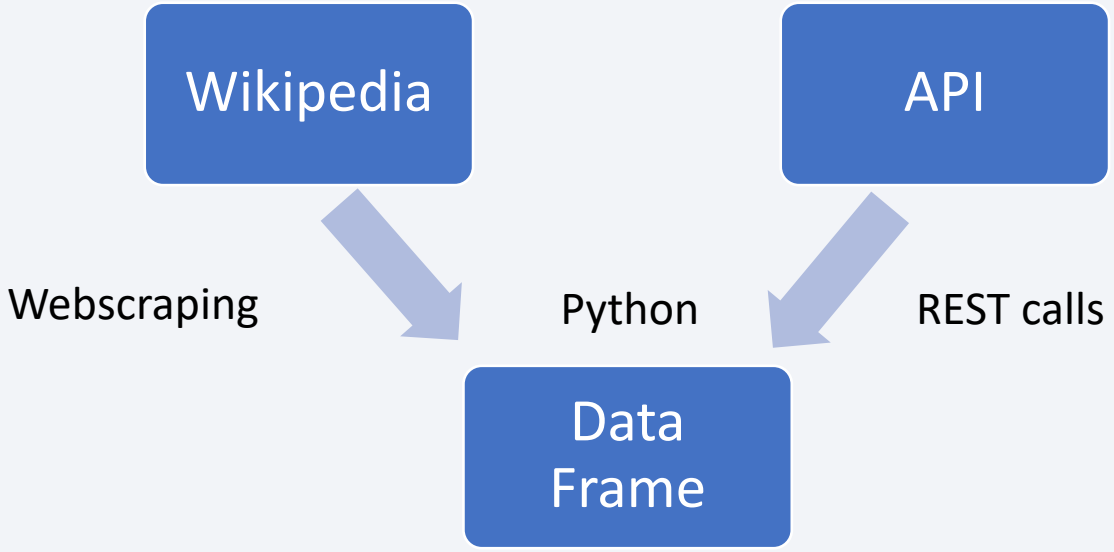
# Methodology

## Executive Summary

- Data collection methodology:
  - Webscraping Wikipedia
  - SpaceX API
- Perform data wrangling
  - Cleaned data with pandas, one hot encoding, etc.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used visual and statistical analysis to determine best models

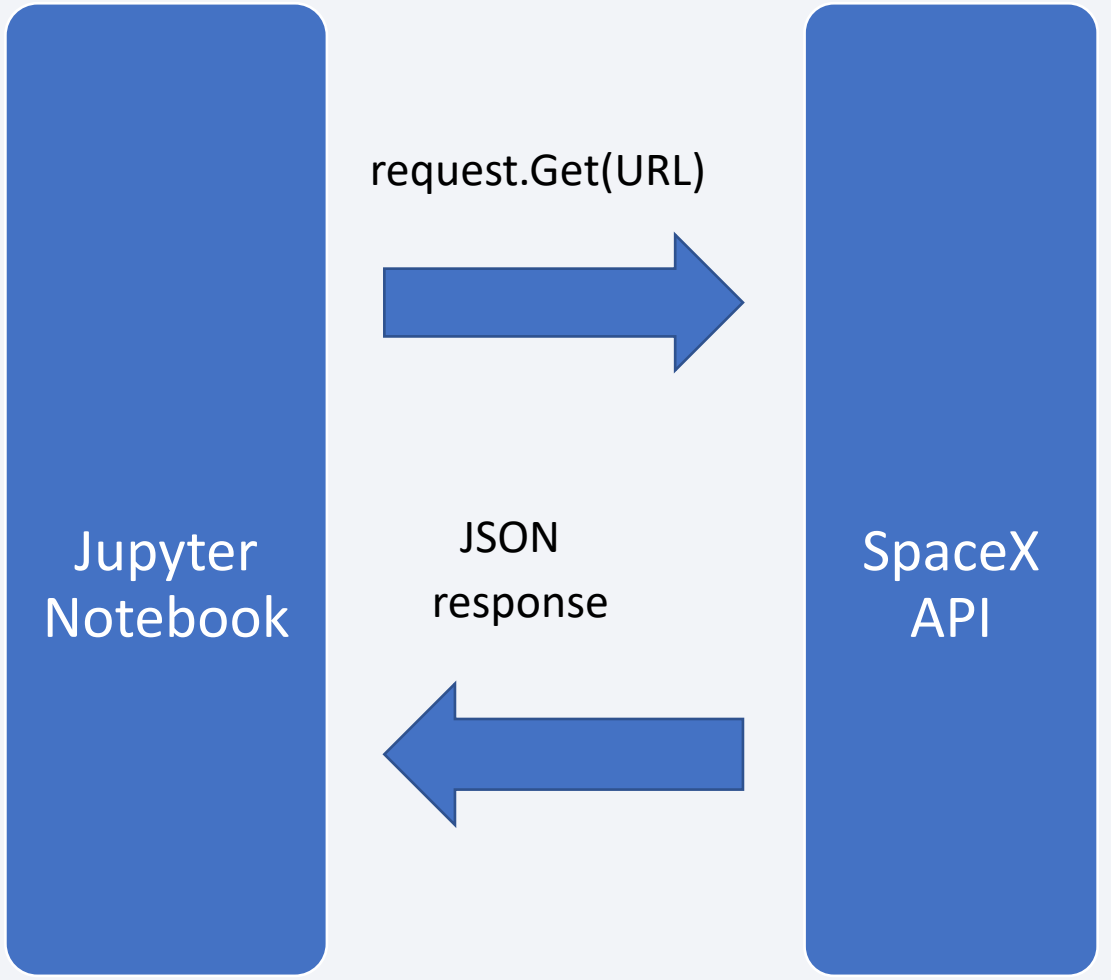
# Data Collection

- Data was collected through:
  - Webscraping SpaceX tables from Wikipedia
  - API calls.



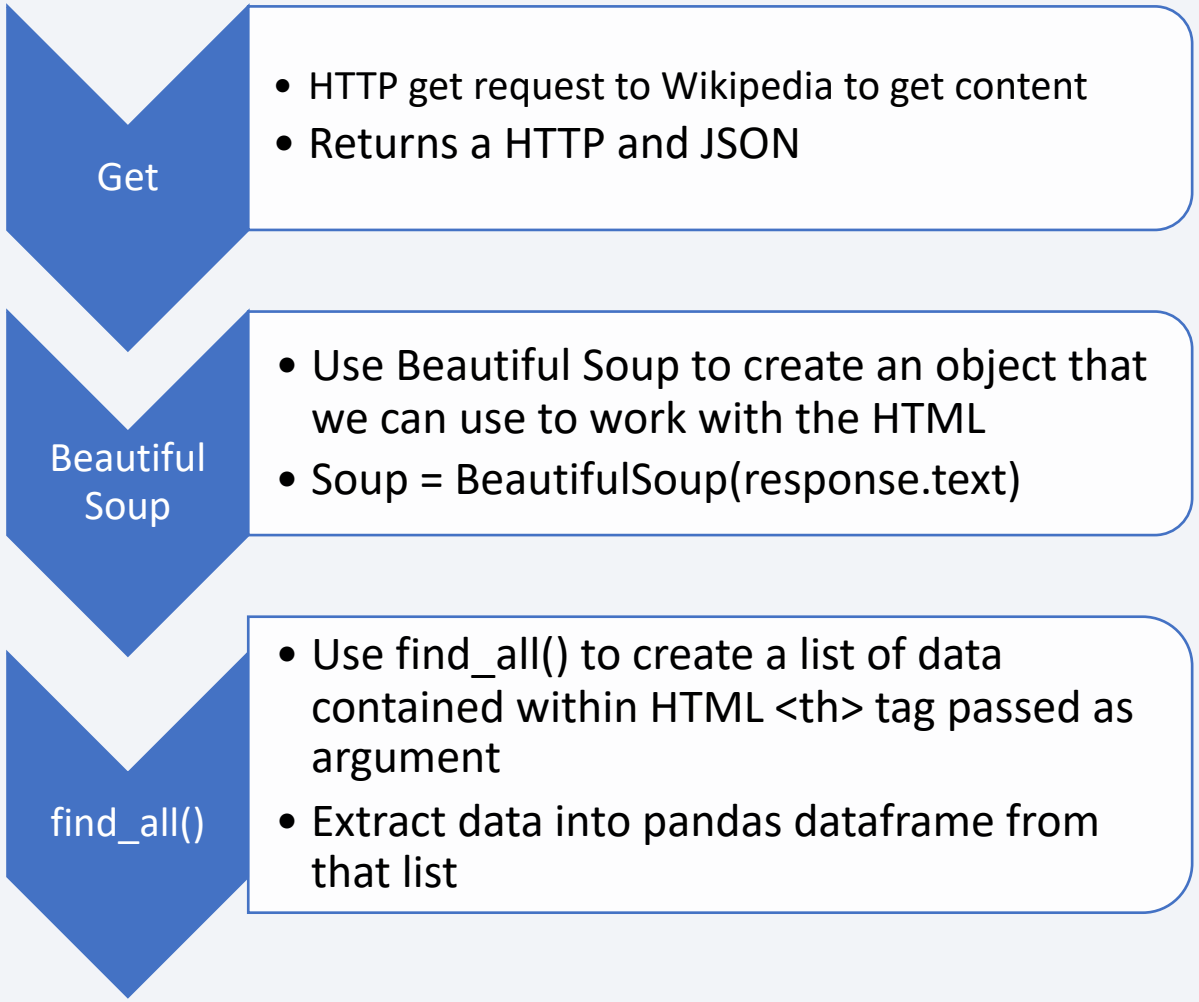
# Data Collection – SpaceX API

- SpaceX REST calls flow chart
- GitHub URL of the completed SpaceX API calls notebook:
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20Collection.ipynb



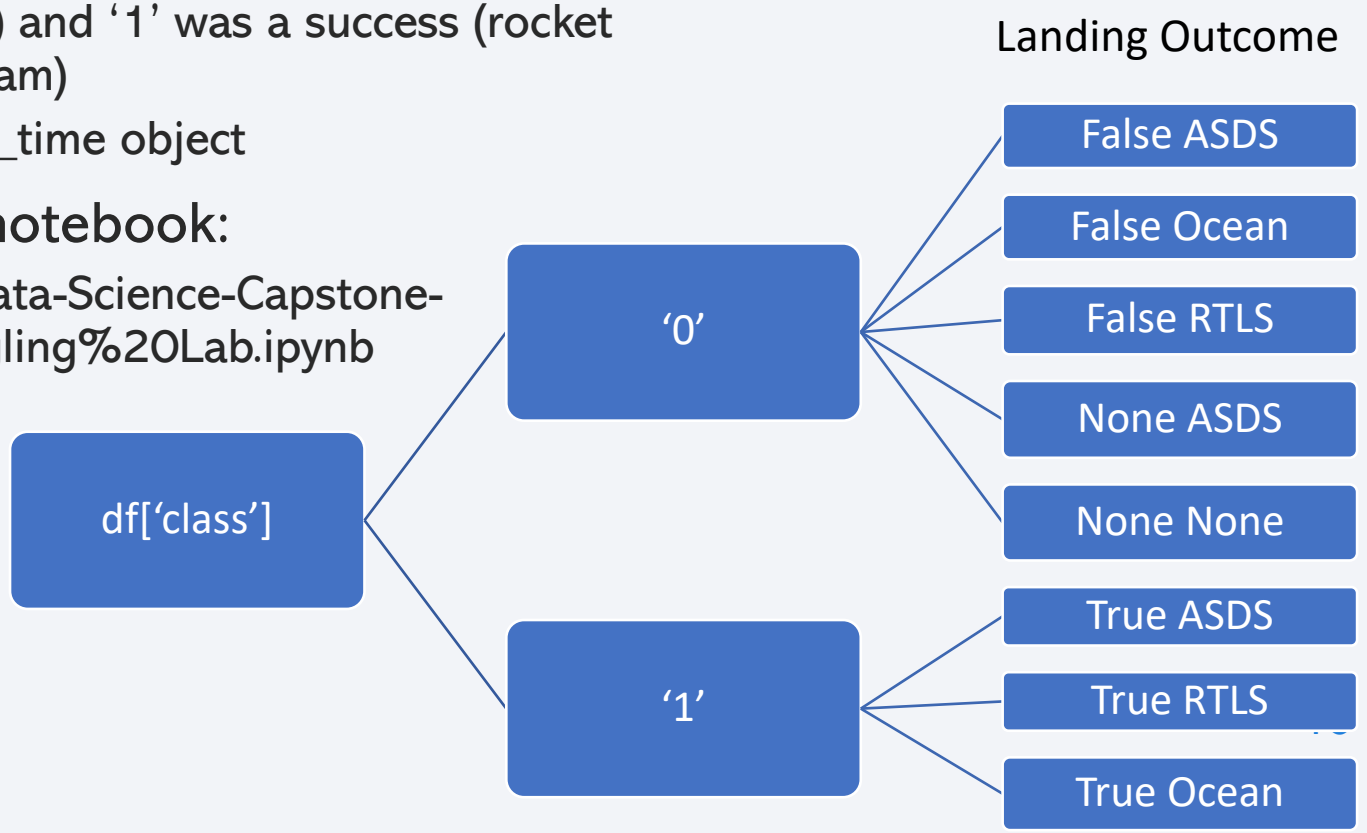
# Data Collection - Scraping

- Webscraping process
- GitHub URL of the completed web scraping notebook:
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/Data%20Collection%20With%20Webscraping.ipynb



# Data Wrangling

- Collected data was processed by:
  - Eliminating Null values in certain columns
  - One hot encoding certain launch outcomes to create column 'class' where '0' was a failure (rocket lost) and '1' was a success (rocket reusable/landed safely). (See diagram)
  - Also changed date value to a date_time object
- GitHub URL for data wrangling notebook:
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/Data%20Wrangling%20Lab.ipynb



# EDA with Data Visualization

- Charts plotted
  - Scatter plots in order to see relationship b/t the variables plotted and the success rate
    - ‘Flight number’ vs. ‘Payload Mass’, color success/failure
    - ‘Flight number’ vs ’Launch Site’, color success/failure
    - ‘Launch Site’ vs ‘Payload Mass’, color success/failure
    - ‘Flight number’ vs ‘Orbit’ type, color success/failure (More flights meant lower orbits)
  - Bar graph
    - Success rate depending on orbit to see if orbit type has an impact on success
  - Line plot
    - Success rate by year showed that as time went on, they had more success
- GitHub URL of completed EDA with data visualization notebook:
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/Exploratory%20data%20analysis%20with%20visualization.ipynb

# EDA with SQL

- SQL Queries performed (summary)
    1. Names of launch sites
    2. Total payload mass (kg) launched by NASA
    3. Average Payload Mass launched by booster F9 V1.1
    4. Date of first successful landing
    5. Names of boosters that were success in landing on drone ship with Payload Mass b/t 4000-6000kg
    6. Total number of successful landings
    7. Names of boosters that carried max payload mass
    8. Failed landings in drone ship for 2015
    9. Rank the landing outcomes by between June 4/2010 and March 20/2017
- GitHub URL of your completed EDA with SQL notebook:
    - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/Exploratory%20Data%20Analysis%20with%20SQL.ipynb

# Build an Interactive Map with Folium

- Map objects
  - Markers
    - Launch sites to show from where rockets were launched
    - Marker Cluster of successful launch types to see where the most successful areas were.
  - Circles
    - To more easily visualize launch sites
  - Lines
    - From launch sites to oceans and launch sites to highways and railroads to visualize the proximity of these structures
- Explain why you added those objects
- GitHub URL of your completed interactive map with Folium map:
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/Launch%20Site%20Location.ipynb

# Build a Dashboard with Plotly Dash

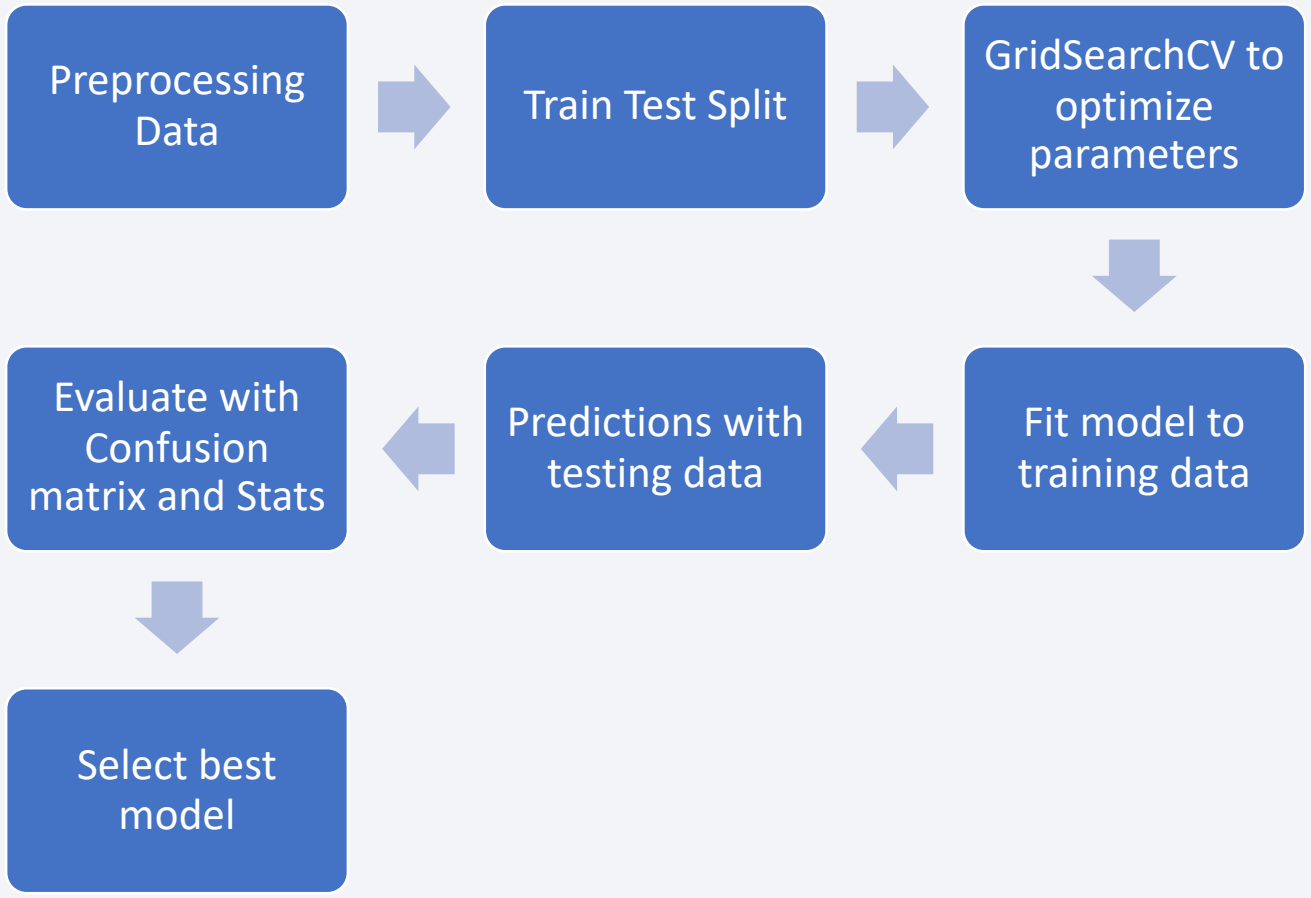
- Plotly Dash used to create:
  - Dropdown menu to select launch site location
  - Pie chart to compare success rates with launch site location
  - Slider to select payload mass range
  - Scatter plot to compare payload mass with success outcome depending on the launch site
- Add the GitHub URL of completed Plotly Dash lab code (screenshots to follow):
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

# Predictive Analysis (Classification)

- Preprocessing data
  - Transferred 'Class' data to numpy array
  - Standardized data so the models would be more accurate with preprocessing.StandardScaler()
  - Split data into testing and training sets using train_test_split with 20% of data reserved for testing
- Finding the right model
  - Used set of parameters and gridsearch to determine best parameters for
    - Logistic Regression Model
    - Support Vector Machine
    - Decision Tree
    - K Nearest Neighbr
- Evaluated each model by fitting with training data and testing with testing data. Then found the accuracy score for each as well as graphed the results using a confusion matrix to find where models succeeded and where they were deficient.
- Add the GitHub URL of completed predictive analysis lab:
  - https://github.com/cjpewarchuk/Data-Science-Capstone-Project/blob/main/Machine%20Learning%20and%20Prediction.ipynb

# Predictive Analysis (Classification) cont'd

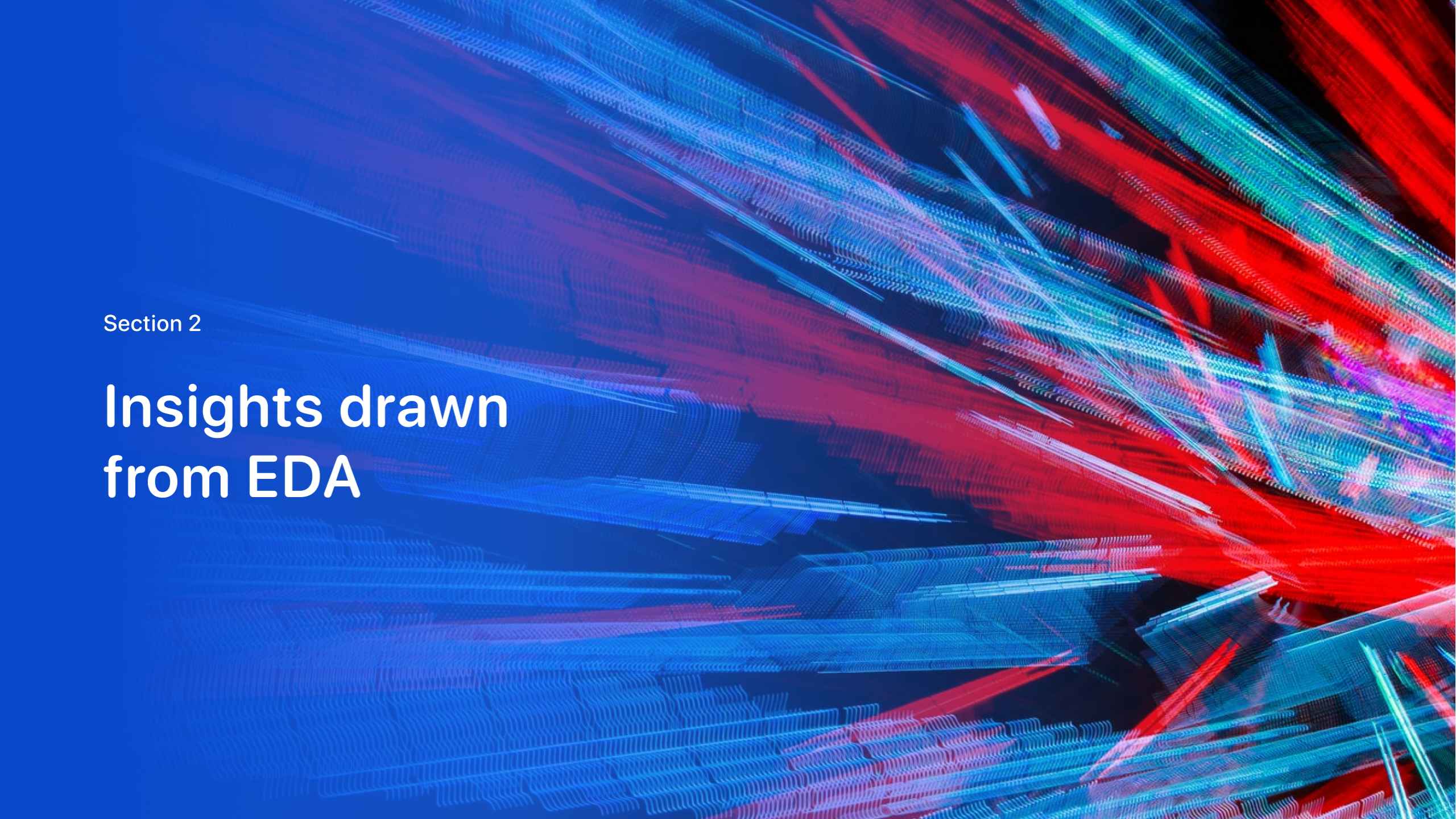
Flow chart of model development



# Results

- Exploratory data analysis results

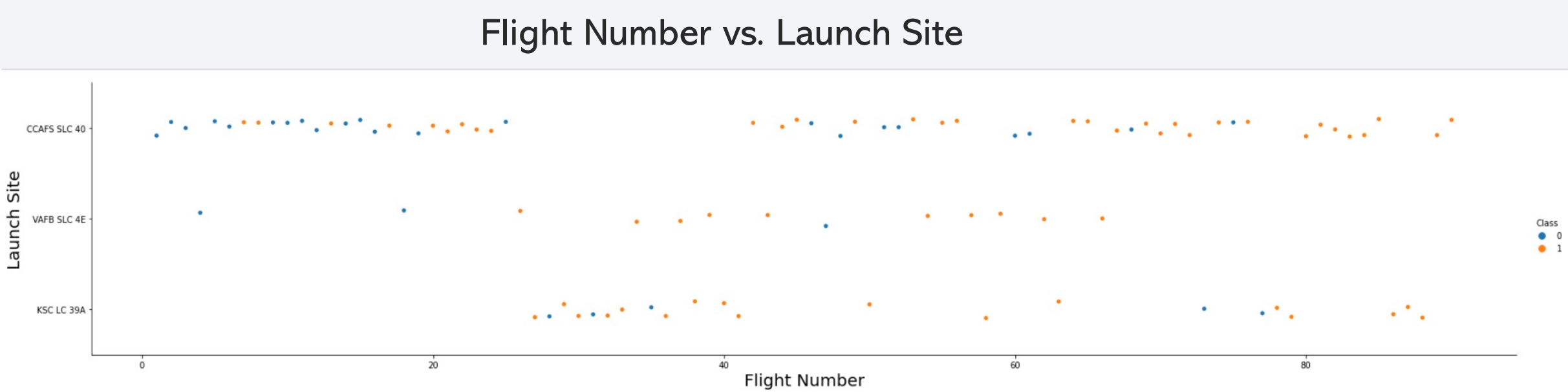
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

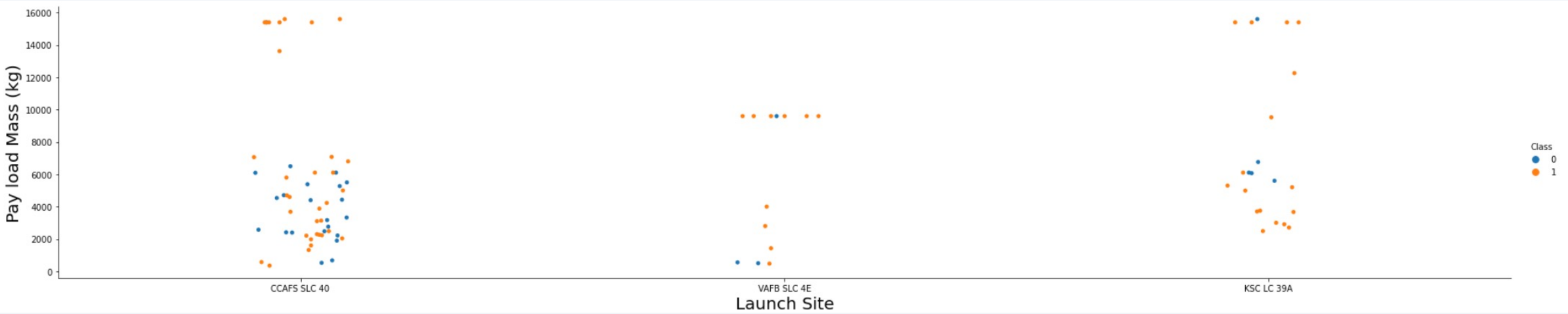
# Insights drawn from EDA

# Flight Number vs. Launch Site




CCAFS SLC 40 is the most frequently used site, though between missions 25 and 42, KSC LC 39A was used most frequently.  VAFB SLC 4E is less frequently used.  The success at all launch sites went up as time went on.

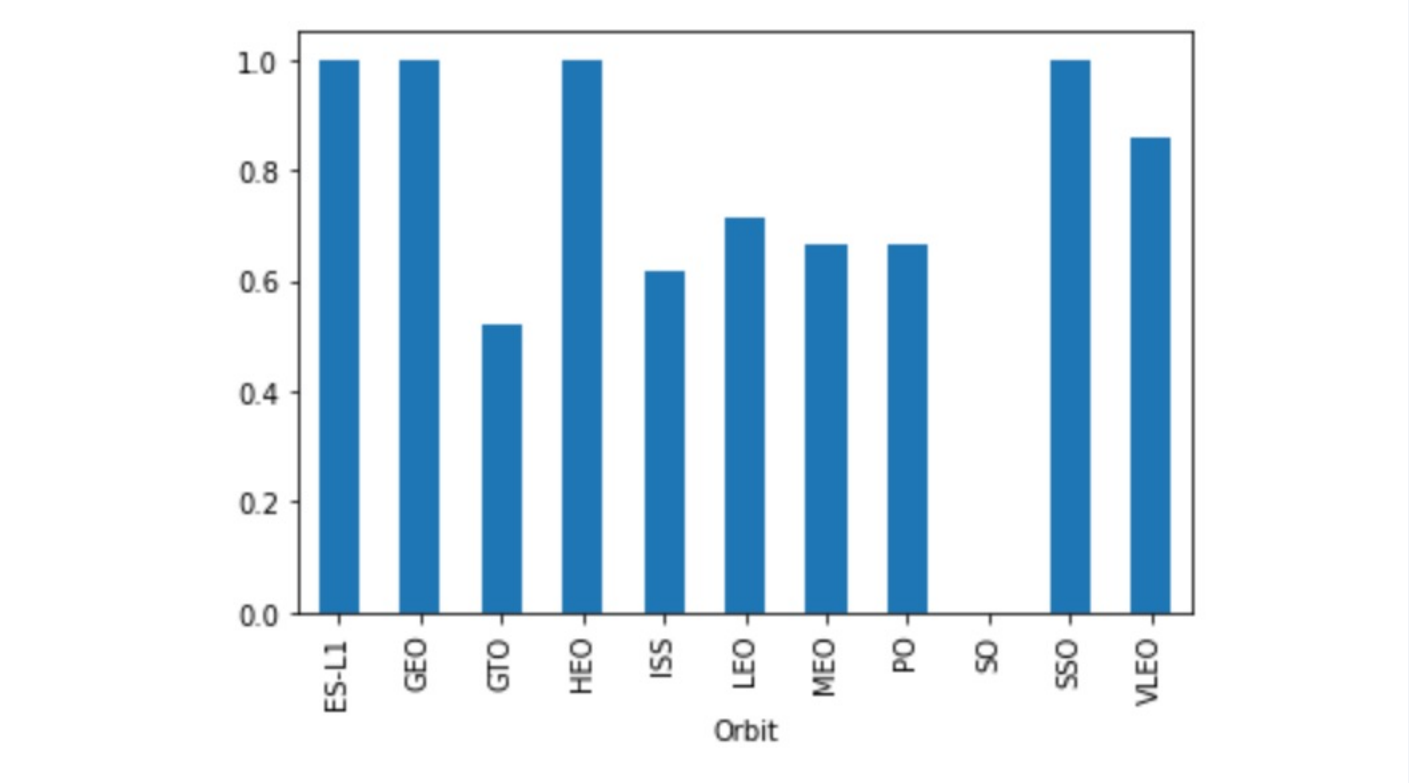
# Payload vs. Launch Site




CCAFS SLC 40 and KSC LC 39A had a range of small to large payload masses. VAFB SLC 4E payloads maxed out at 10 000 kg, but had few other values. There appears to be few payloads between 8,000 kg and 14,000 kg at either of the other two launch sites.

# Success Rate vs. Orbit Type

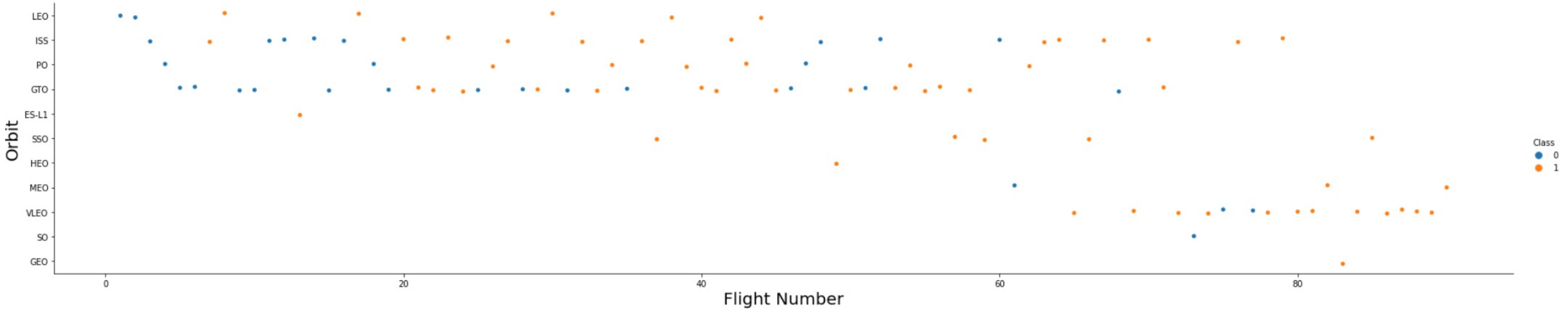
- Success Rate vs Orbit type




SO orbit was obviously a case where the rockets had to be sacrificed. Best orbit types for success were ES-L1, GEO, HEO, SSO, and VLEO.

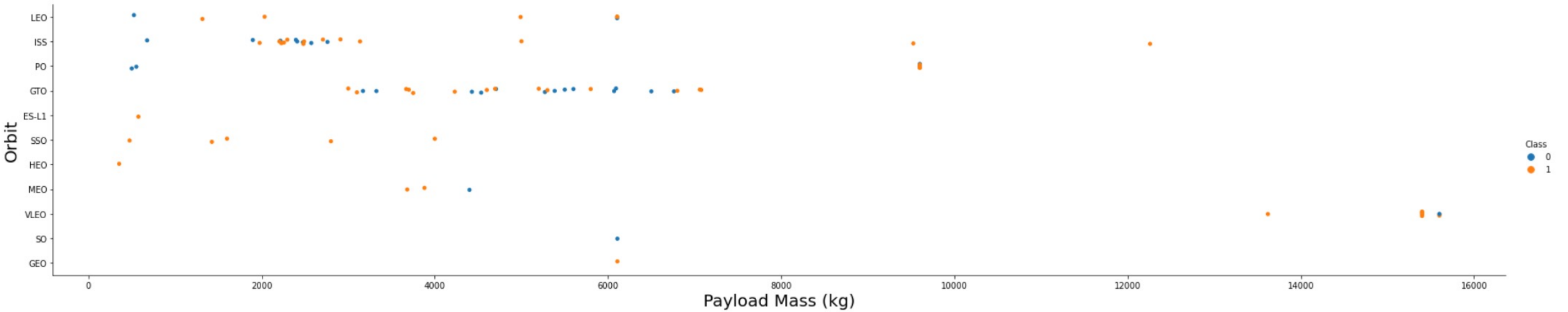
# Flight Number vs. Orbit Type

- Flight number vs. Orbit type



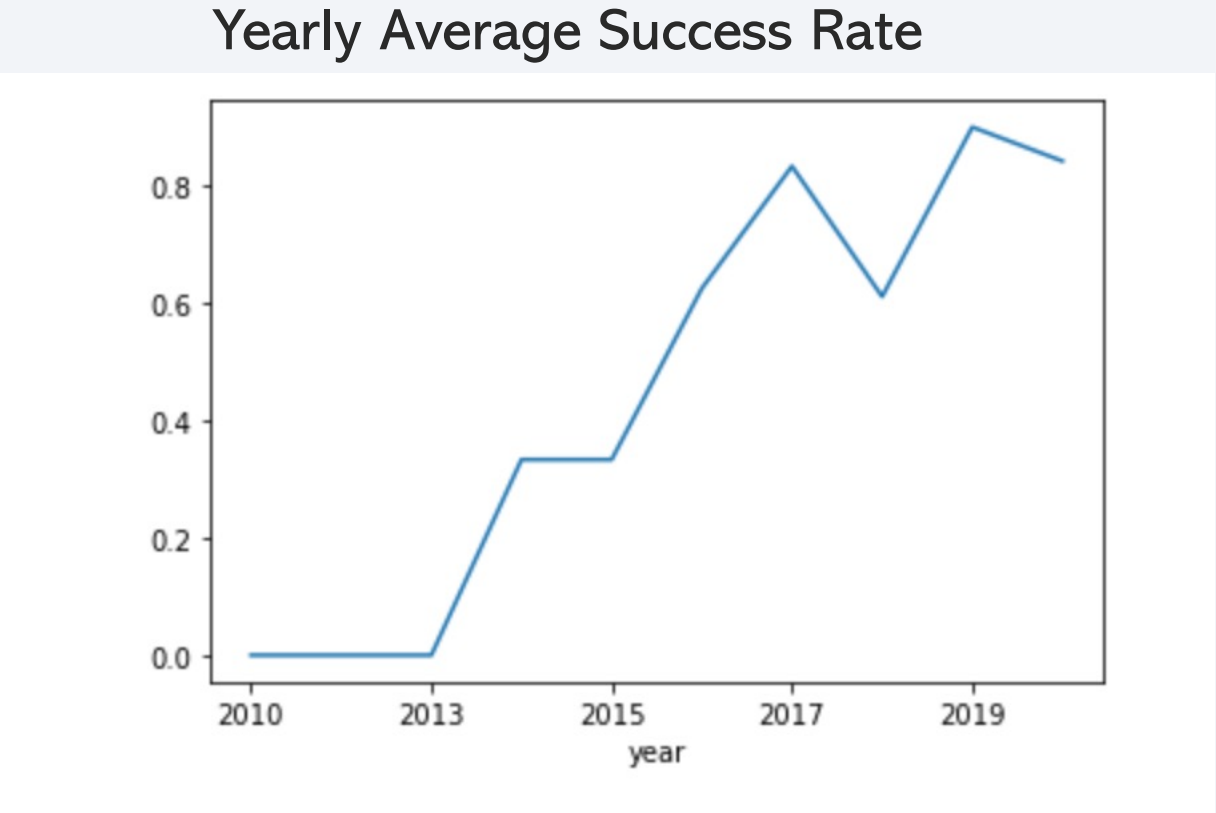

LEO, ISS, GTO were the most common orbit types at the beginning. Later in the flights, we see that VLEO becomes much more common, and in fact the most common orbit type after flight 60. The color shows that all orbit types success rates improved as more flights occurred.

# Payload vs. Orbit Type




With heavy payloads, the more successful launches were VLEO, PO, and ISS.

# Launch Success Yearly Trend




It is very clear from this graph that the average success rate rose significantly over time since 2013.

# All Launch Site Names

SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL

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

This SQL query provided the name of each unique launch site for SpaceX in the dataframe.

# Launch Site Names Begin with 'CCA'

SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

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

This SQL query returned the launch sites with CCA in their name with a limit of 5.

# Total Payload Mass

SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'

| total_payload_mass |
| --- |
| 45596 |

This SQL query returned the total payload mass in kg for when the booster was launched by NASA.

# Average Payload Mass by F9 v1.1

SELECT AVG(PAYLOAD_MASS__KG_) AS MEAN_PAYLOAD_MASS FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

| mean_payload_mass |
| --- |
| 2928 |

This SQL query returned the average payload mass carried by F9 1.1 booster rocket.

# First Successful Ground Landing Date

SELECT MIN(DATE) AS FIRST_SUCCESSFUL_LANDING FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'

| first_successful_landing |
| --- |
| 2015-12-22 |

This SQL query returned the date of the first successful landing on a ground pad.

# Successful Drone Ship Landing with Payload between 4000 and 6000

SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000

| booster_version |
| --- |
| F9 v1.1 |
| F9 v1.1 B1011 |
| F9 v1.1 B1014 |
| F9 v1.1 B1016 |
| F9 FT B1020 |
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1030 |
| F9 FT B1021.2 |
| F9 FT B1032.1 |
| F9 B4 B1040.1 |
| F9 FT B1031.2 |
| F9 B4 B1043.1 |
| F9 FT B1032.2 |
| F9 B4 B1040.2 |
| F9 B5 B1046.2 |
| F9 B5 B1047.2 |
| F9 B5 B1046.3 |
| F9 B5B1054 |
| F9 B5 B1048.3 |
| F9 B5 B1051.2 |
| F9 B5B1060.1 |
| F9 B5 B1058.2 |
| F9 B5B1062.1 |

This SQL query returned a list of booster versions that carried a payload mass in kg between 4000 kg and 6000 kg

# Total Number of Successful and Failure Mission Outcomes

SELECT COUNT(*) AS NUMBER_OF_SUCCESSES FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%'

| numberofsuccesses |
| --- |
| 100 |

This SQL query returned the number of successful mission outcomes.

# Boosters Carried Maximum Payload

SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

| 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 |

This SQL query returned a list of boosters that carried the maximum payload.

# 2015 Launch Records

SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = ‘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 |

This SQL query shows landing outcomes, booster versions, and launch sites that resulted in drone ship failures in 2015.

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

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

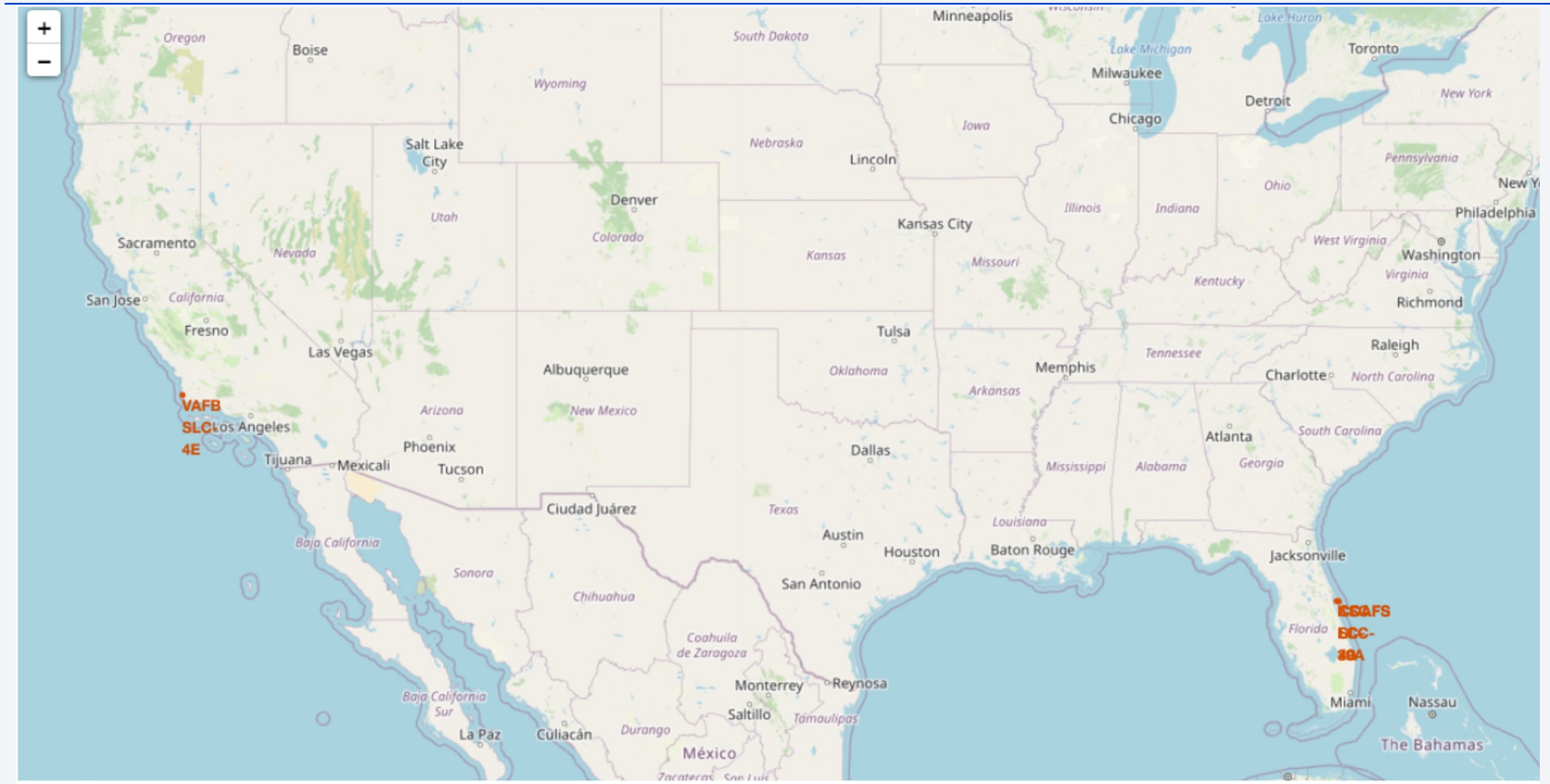
| landing__outcome | number_of_each_outcome |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

This SQL query shows the landing outcomes between June 4 2010 and March 3 2017 ranked in descending order of number of each outcome.

Section 4

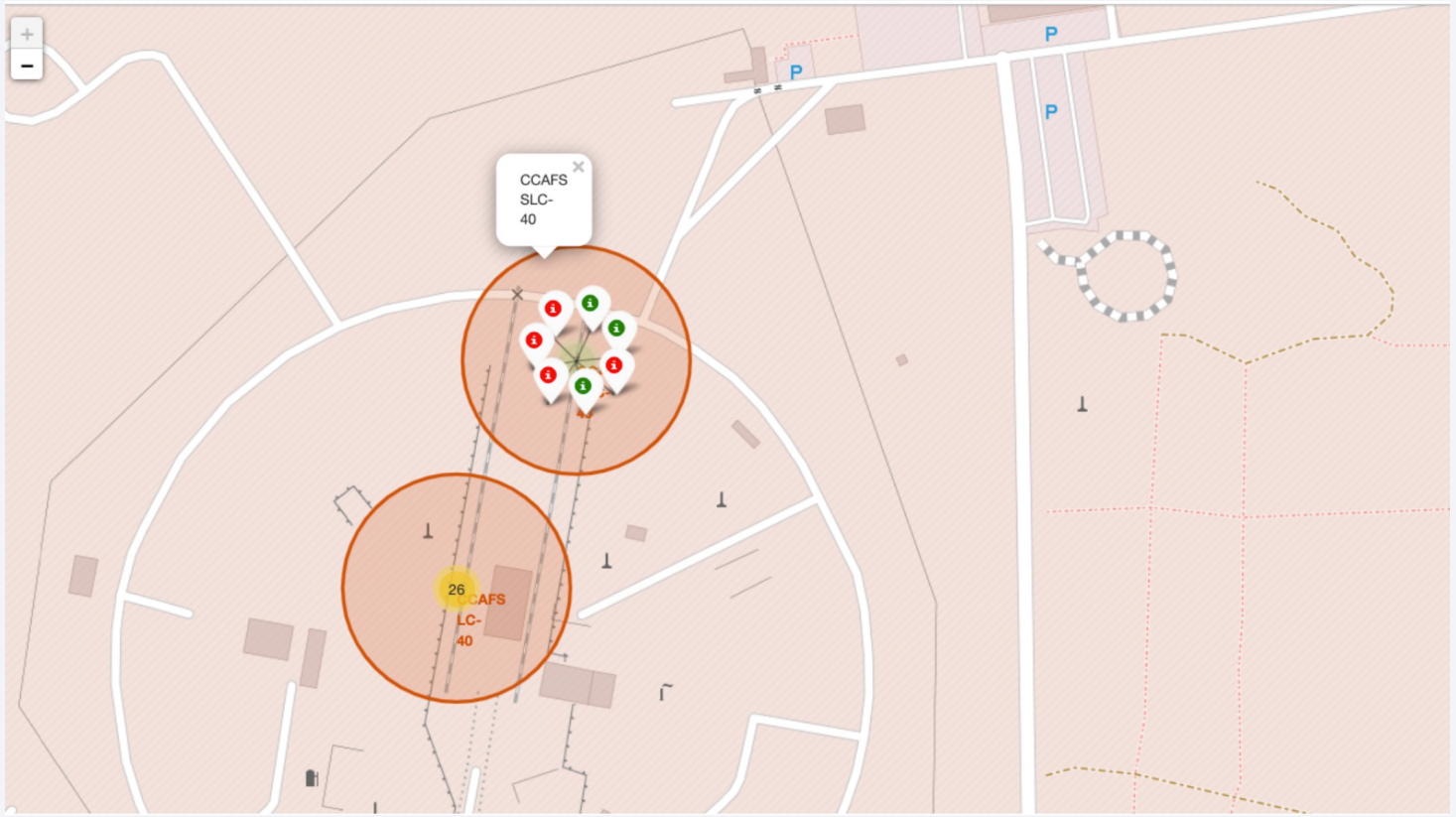
# Launch Sites Proximities Analysis

# Launch Site Locations



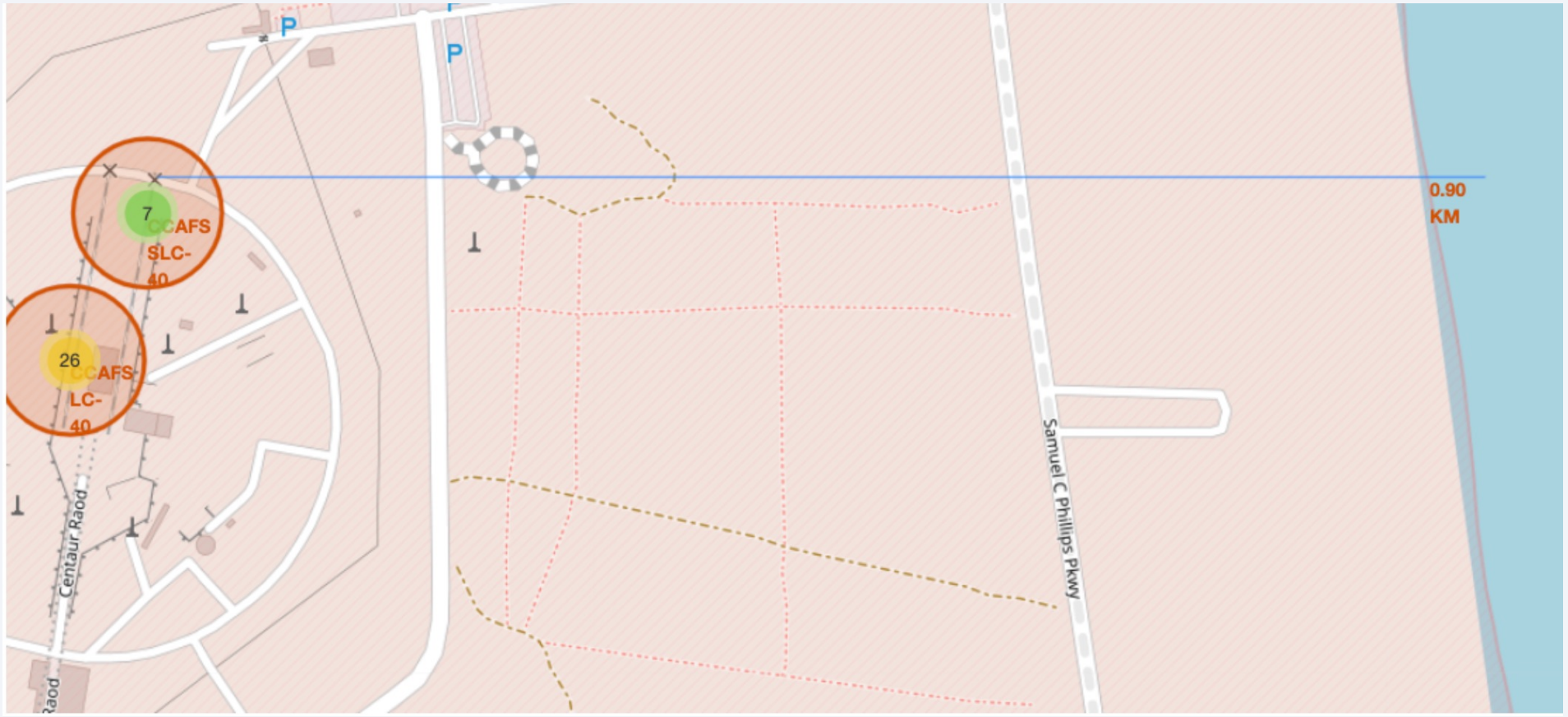

In red are the launch site locations.  Two are in California, and two are in Florida.

# Map Showing Launch Outcomes At CCAFS SLC-40

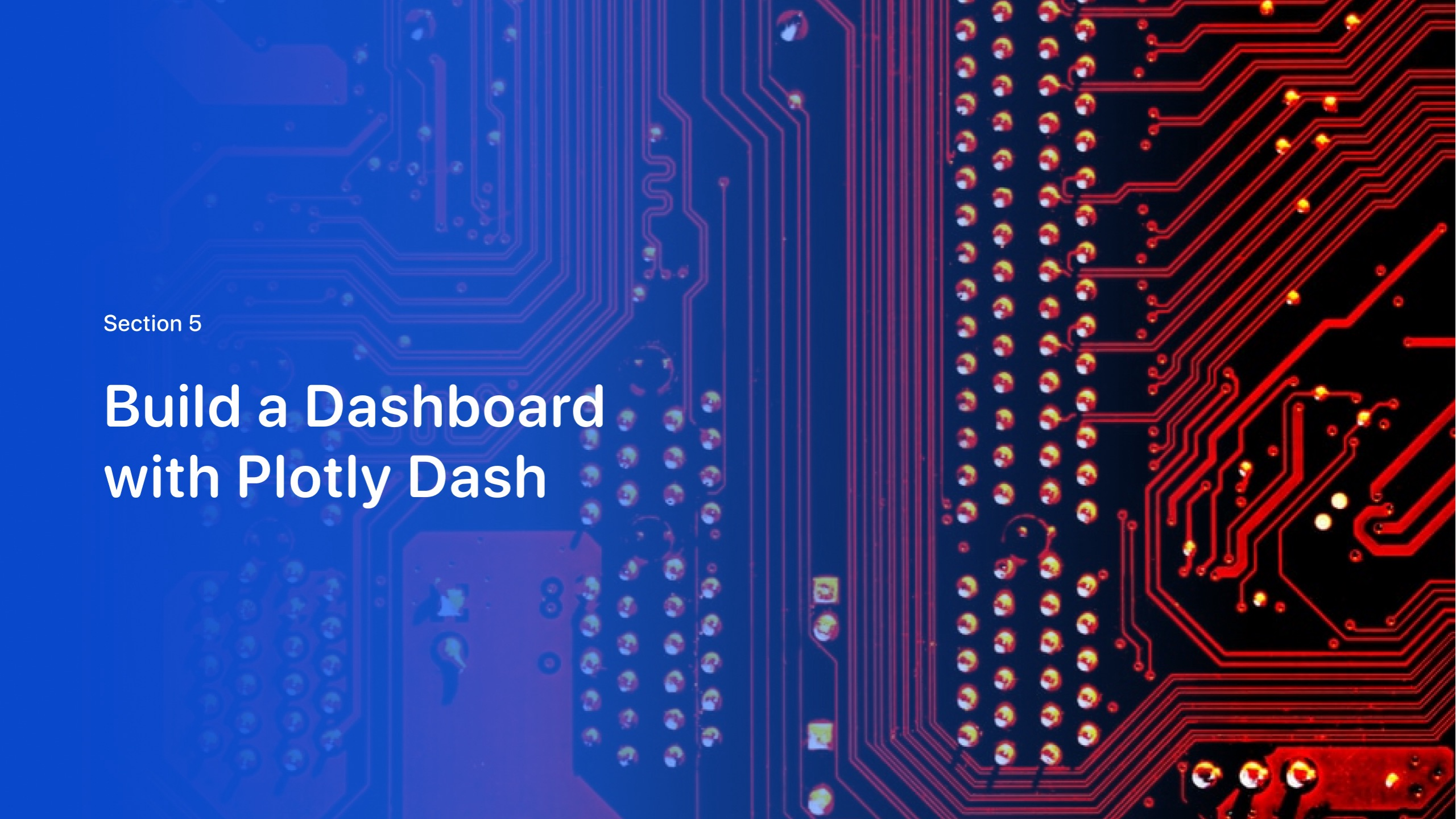



The top circle contains the marker cluster for CCAFS SLC-40. Green markers represent successes and red represent failures.

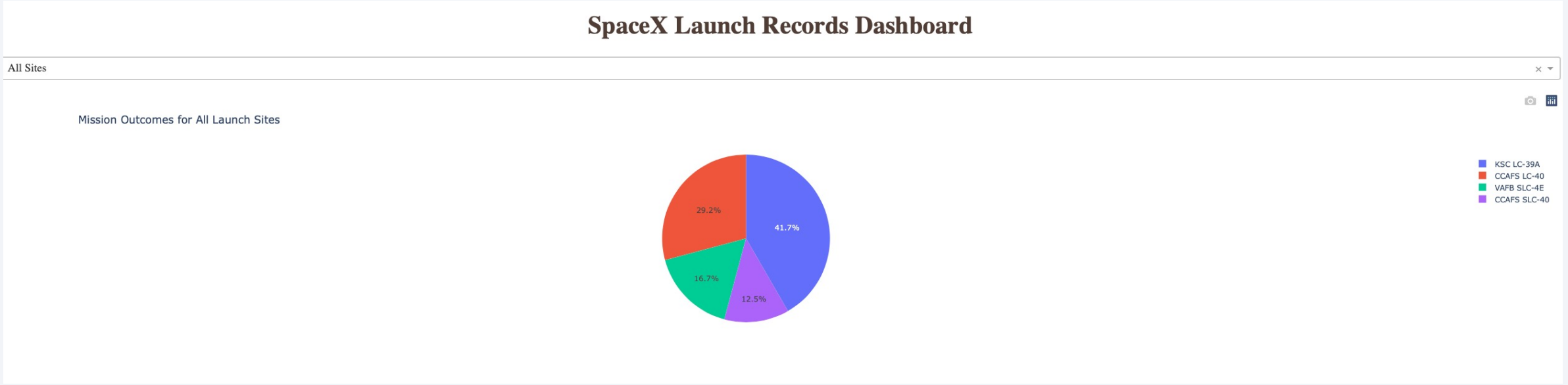
# Map showing distance to ocean from launch site




The Circles on the left show the launch sites, the blue line represents the distance from the ocean to the launch site CCAFS SLC-40 and the red number (0.90 km) is the calculated distance based on the longitudes and latitudes.

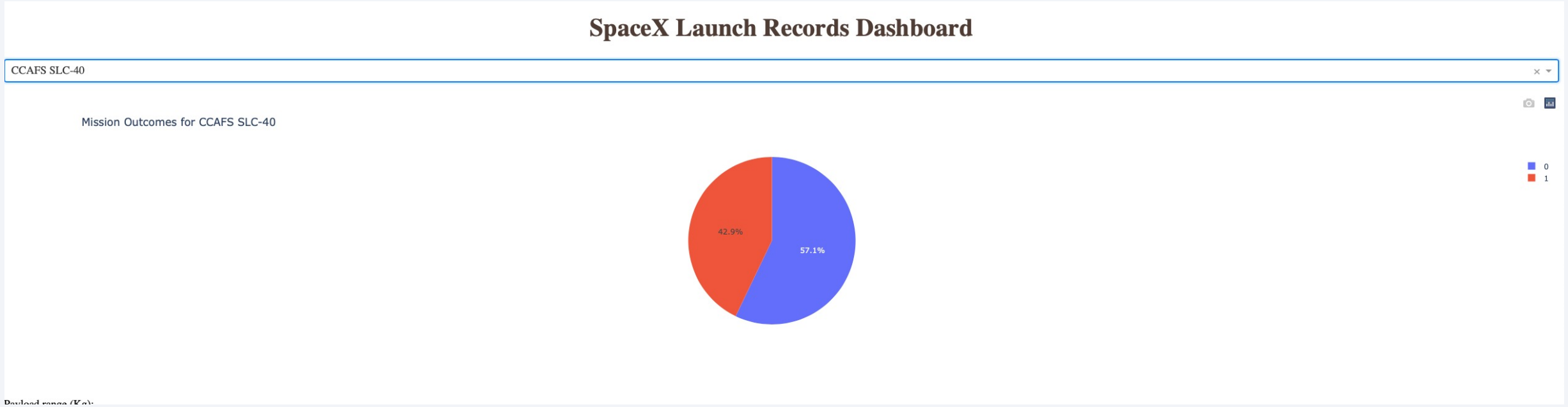

Section 5
Build a Dashboard with Plotly Dash

# SpaceX Dashboard PieChart (All sites)




SpaceX Dashboard Pie chart. The dropdown menu allows you to select launch site, which then changes the pie chart to reflect the success rate of that particular launch site. KSC LC-39A has the highest percentage of successful landings.

# Pie Chart For CCAFS SLC-40 Success Ratio




Above is the pie chart showing the success ratio for CCAFS SLC-40. This launch site has a success rate of 42.9%, which is the highest of the launch sites.

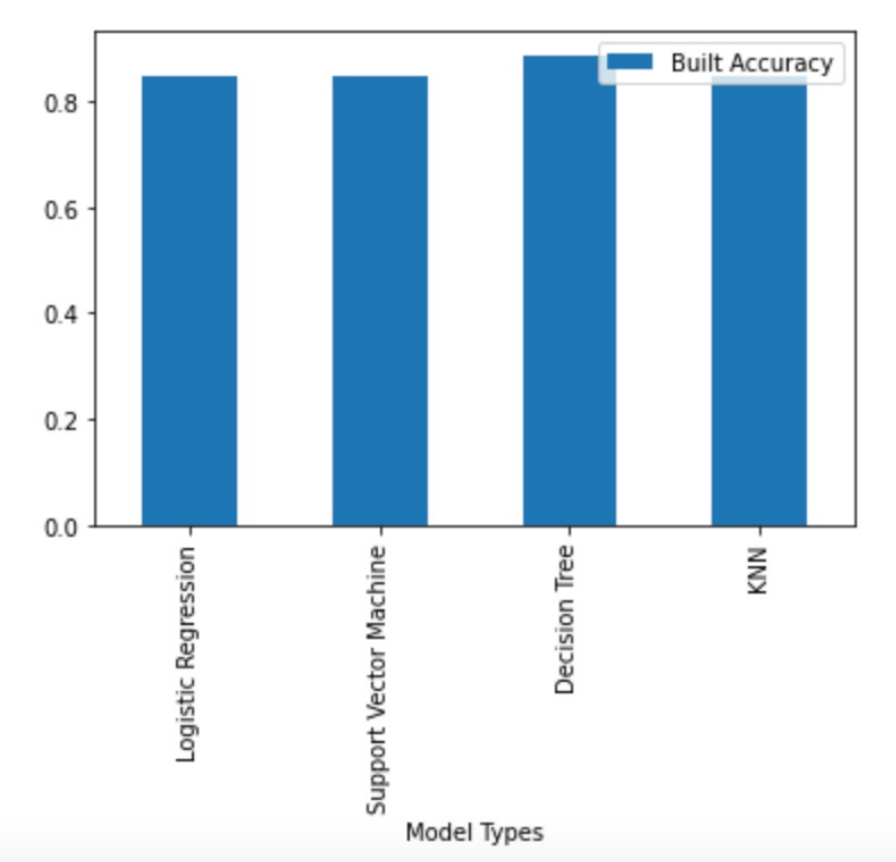
# Scatter Plot for Payload vs Launch Outcome


Scatter plot for Payload Mass vs. Launch Outcome for all sites. Color indicates the booster version category. As you can see, the values for launch outcome are 0 and 1, causing a discrete look to the scatter plot. The range slider at the top is set from about 2000 to 6000 kg, which has restricted the data set with respect to payload mass.

Section 6

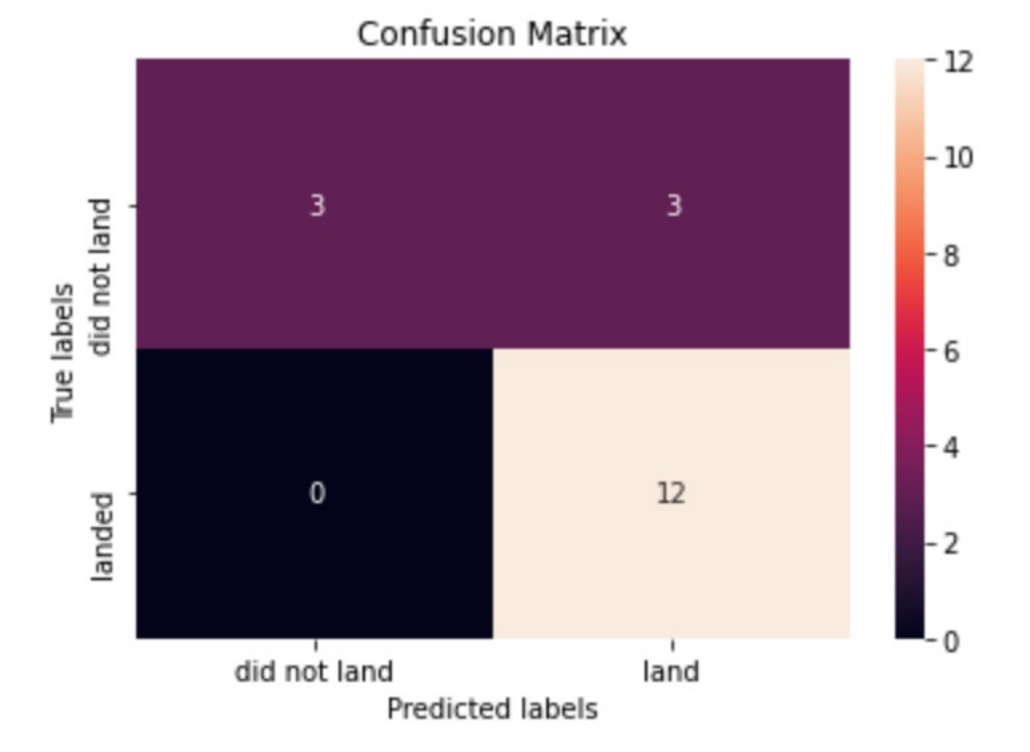
# Predictive Analysis (Classification)

# Classification Accuracy For Different Models




Above is a bar chart showing the build accuracy of the different model types that we used using the training data. The decision tree performed best on the training data, though it actually performed worst on the testing data that we used.

# Confusion Matrix




- This is the confusion matrix for the best performing model. The upper left cell contains the number of true negatives, the upper right contains the number of false positives, lower left is false negatives, and lower right is true positives.

# Conclusions

- The landing success increases significantly the more flights are done, so SpaceY can expect to have a bit of learning curve for saving the rockets for future launches unless they're able to poach some of SpaceX's people or learn from their mistakes.
- At first, while growing and needing to preserve money, SpaceY may want to consider launching with orbit types of ES-L1, GEO, HEO, SSO, and VLEO with moderate to heavy payloads. This should increase the chances of preserving the rockets.
- Though the decision tree model for determining launch success was the strongest for the training data, it had a lower fit for testing data. The models were fairly comparable for the testing data, so as more data is accumulated, the models should continue to be refined and improved in order to better predict future launch success.

# Appendix

- All relevant screen shots and charts have been included in above presentation. Normally, we would have extra charts and other items in the appendices.

# Thank you!