IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Carlos Alejandro Hernandez Gomez
Jan-13-2024

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data
  - Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

# Introduction

- Project background and context
- Problems you want to find answers

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Performed data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

# Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

**Data Columns are obtained by using SpaceX REST API:**

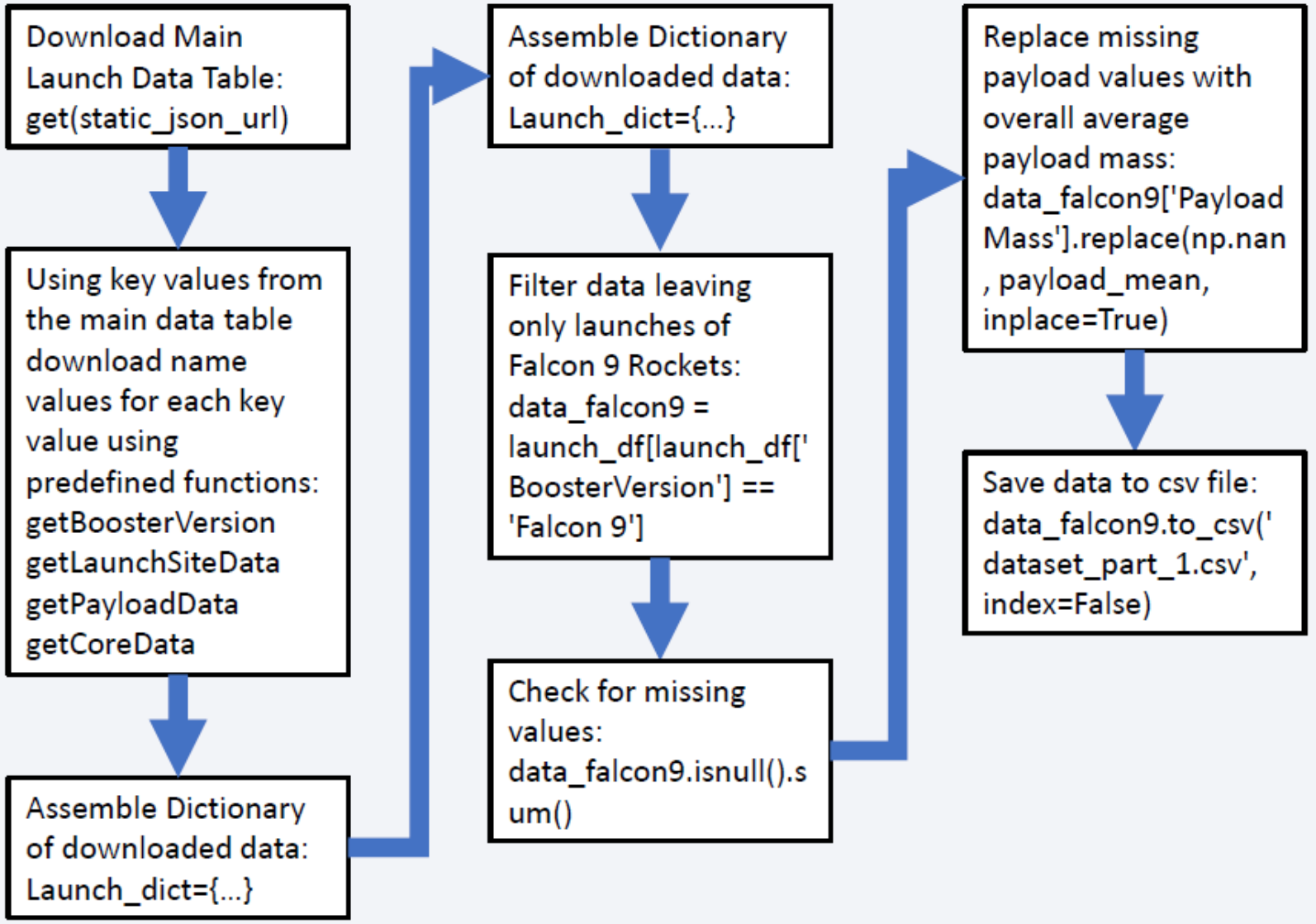
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

**Data Columns are obtained by using Wikipedia Web Scraping:**

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

# Data Collection – SpaceX API

[GitHub: Data Collection API]



# Data Collection - Scraping

Import libraries including Beautifulsoup4
**from bs4 import BeautifulSoup**

HTTP GET request is made to the relevant Wikipedia url
**page=requests.get(static_url)**
See Appendix 1.2 for more info

A BeautifulSoup object is created from the HTML response.
**soup = BeautifulSoup(page.text, 'html.parser')**

Find all tables in the HTML response.
**html_tables=soup.find_all('table')**
Iterate through the table header elements to extract column names.

Create a dictionary with the column names as keys
**launch_dict= dict.fromkeys(column_names)**
**launch_dict['Flight No.'] = []**
**launch_dict['Launch site'] = []**
**launch_dict['Payload'] = [] ...**

Fill the dictionary with launch records
Create a pandas dataframe from the dictionary

- [GitHub link: Data Collection - Scraping]

# Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed

to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means

the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

GitHub link: Data Wrangling

# EDA with Data Visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series).

GitHub link: EDA with Data Visualization

# EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order



GitHub link: EDA with SQL

# Build an Interactive Map with Folium

**Markers of all Launch Sites:**

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

**Coloured Markers of the launch outcomes for each Launch Site:**

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

**Distances between a Launch Site to its proximities:**

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.



GitHub link: Interactive Visual Analytics with Folium

# Build a Dashboard with Plotly Dash

**Launch Sites Dropdown List:**

- Added a dropdown list to enable Launch Site selection.

**Pie Chart showing Success Launches (All Sites/Certain Site):**

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

**Slider of Payload Mass Range:**

- Added a slider to select Payload range.

**Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:**

- Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub link: SpaceX Dash App

# Predictive Analysis (Classification)

**Model building**

- Import libraries, load data
- Standardize and transform data (preprocessing.StandardScaler())

**Evaluating and improving the model**

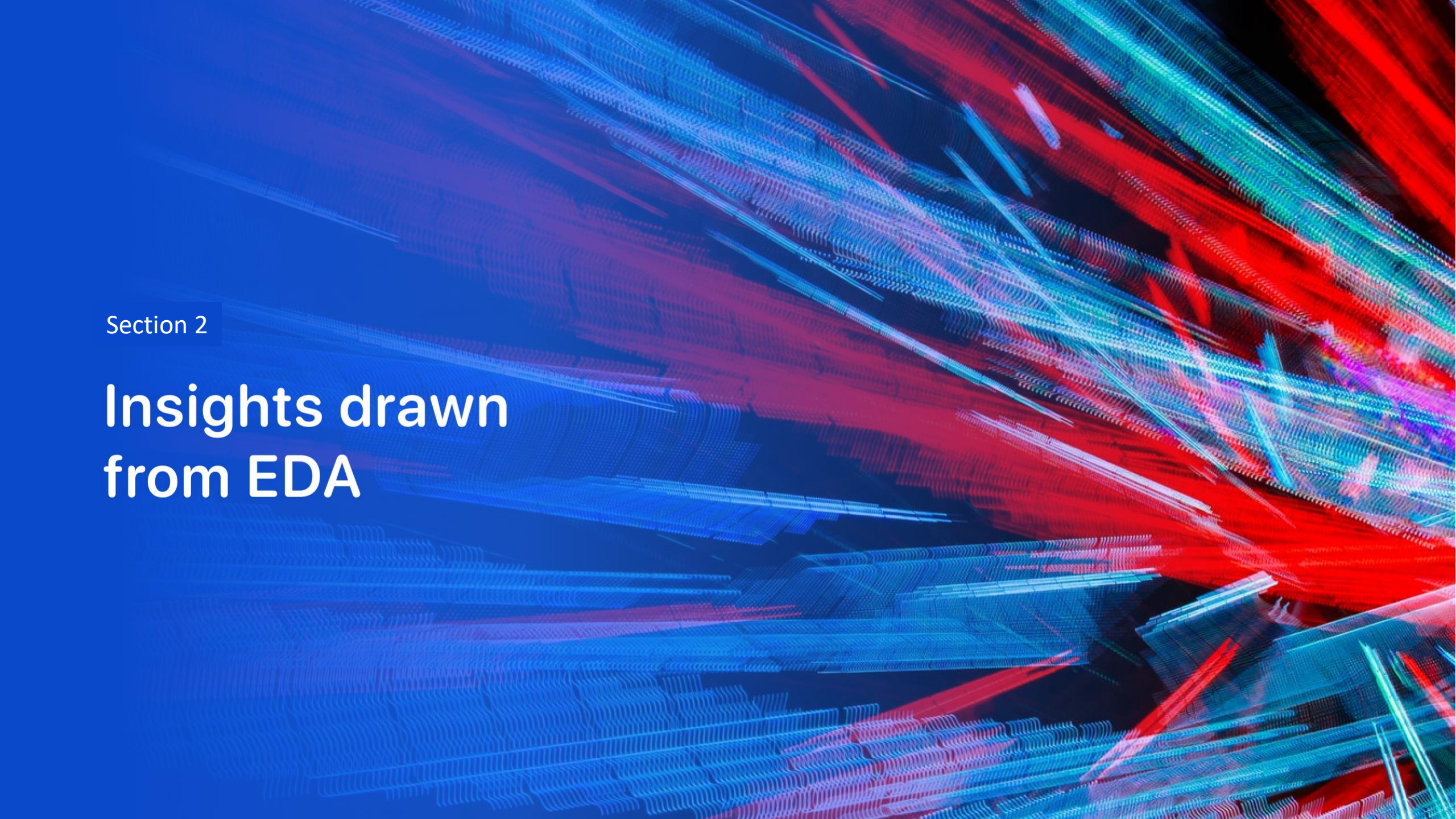
- Split data into Train, Test splits using train_test_split()
- For each algorithm tested: Logistic regression, Support Vector Machines, Decision tree, K nearest neighbors,
    - Create a GridSearchCV object. Use Grid Search to Fit this object to find the best model parameters from the dictionary parameters

**Finding the best performing classification model**

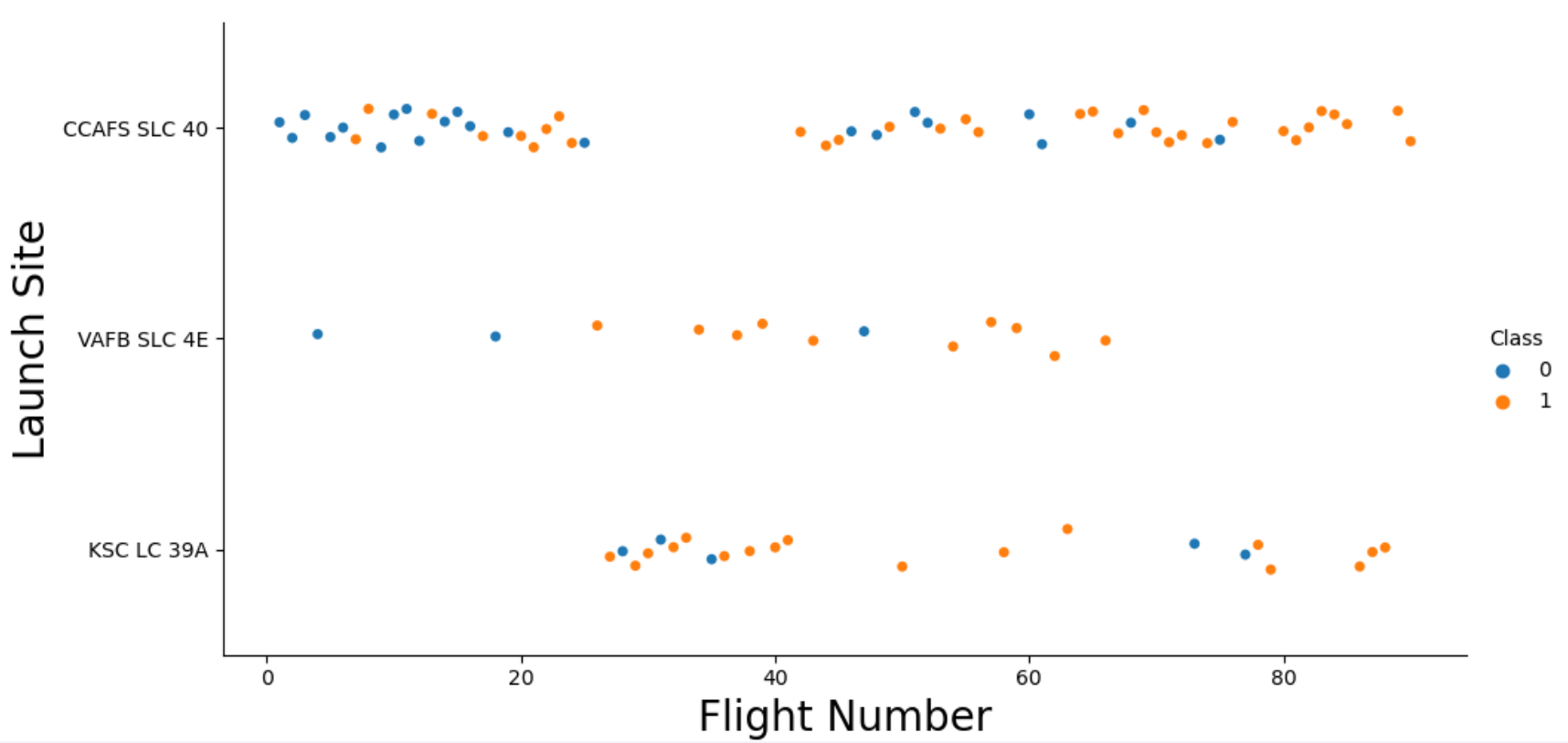
- For each algorithm tested:
    - From the output of the **GridSearchCV** object for the algorithm, We display the best parameters using the data attribute **best_params_** and the accuracy on the validation data using the data attribute **best_score_**.
    - The accuracy on the test data is calculated using the method **score**:
    - A **confusion matrix** is created and used to assess the performance of the classification model. It summarizes the results of classification by showing the counts of true positive, true negative, false positive, and false negative predictions.
    - Summarize how you built, evaluated, improved, and found the best performing classification model
    - Displaying the Scores of all the algorithms reveals the best model



GitHub link: Machine Learning Prediction

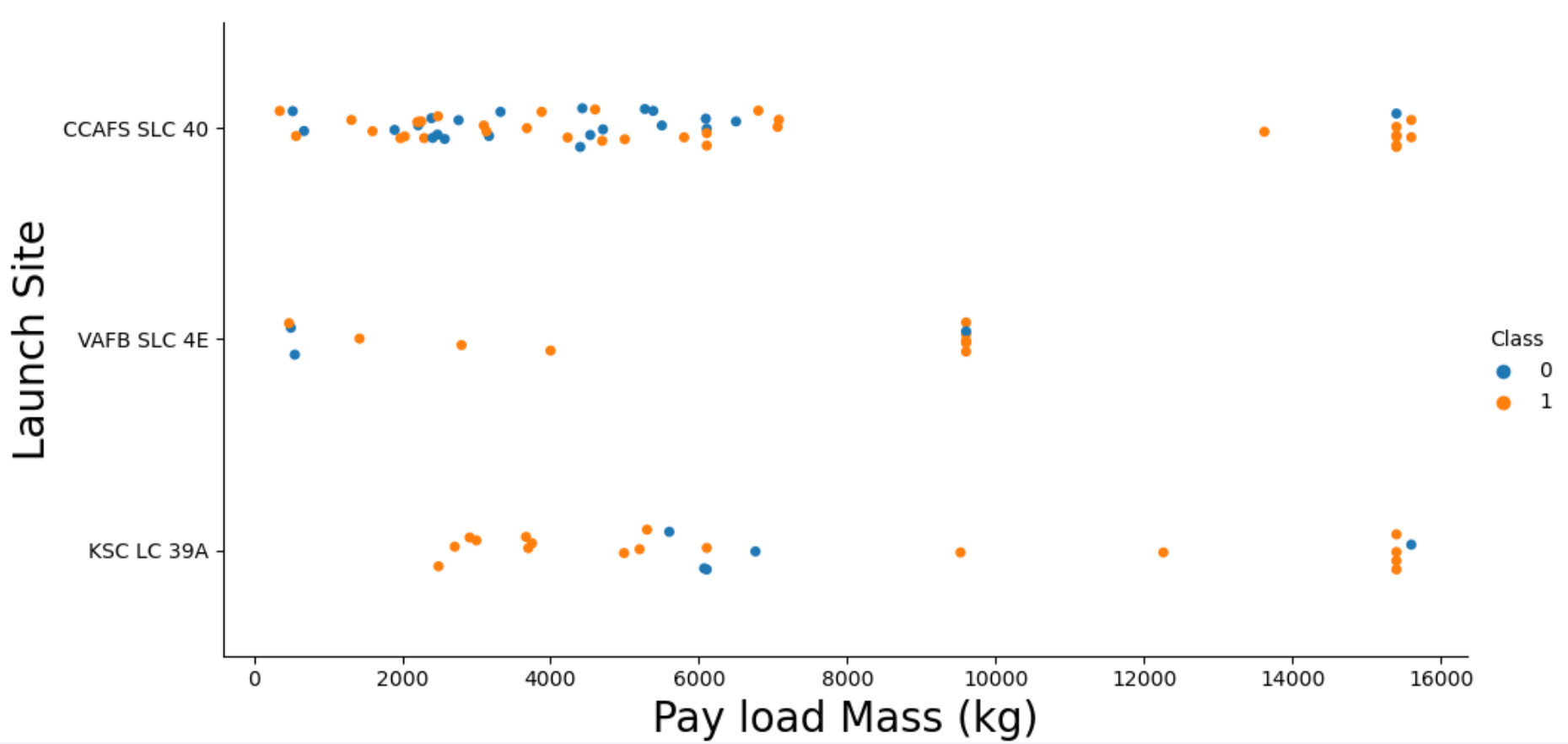

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site



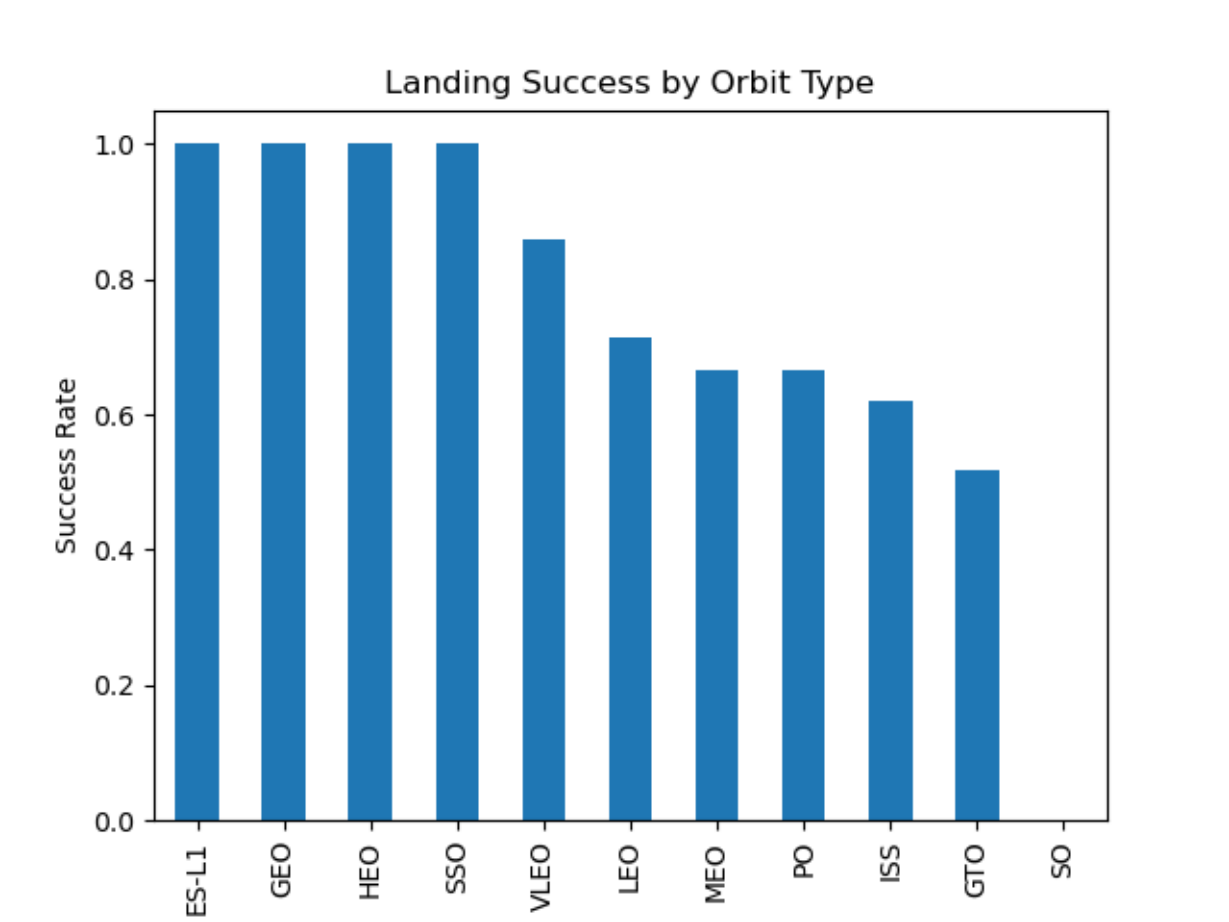

- Landing success is not strongly dependent upon launch site.
- Landings are more successful as more launches are conducted.

# Payload vs. Launch Site



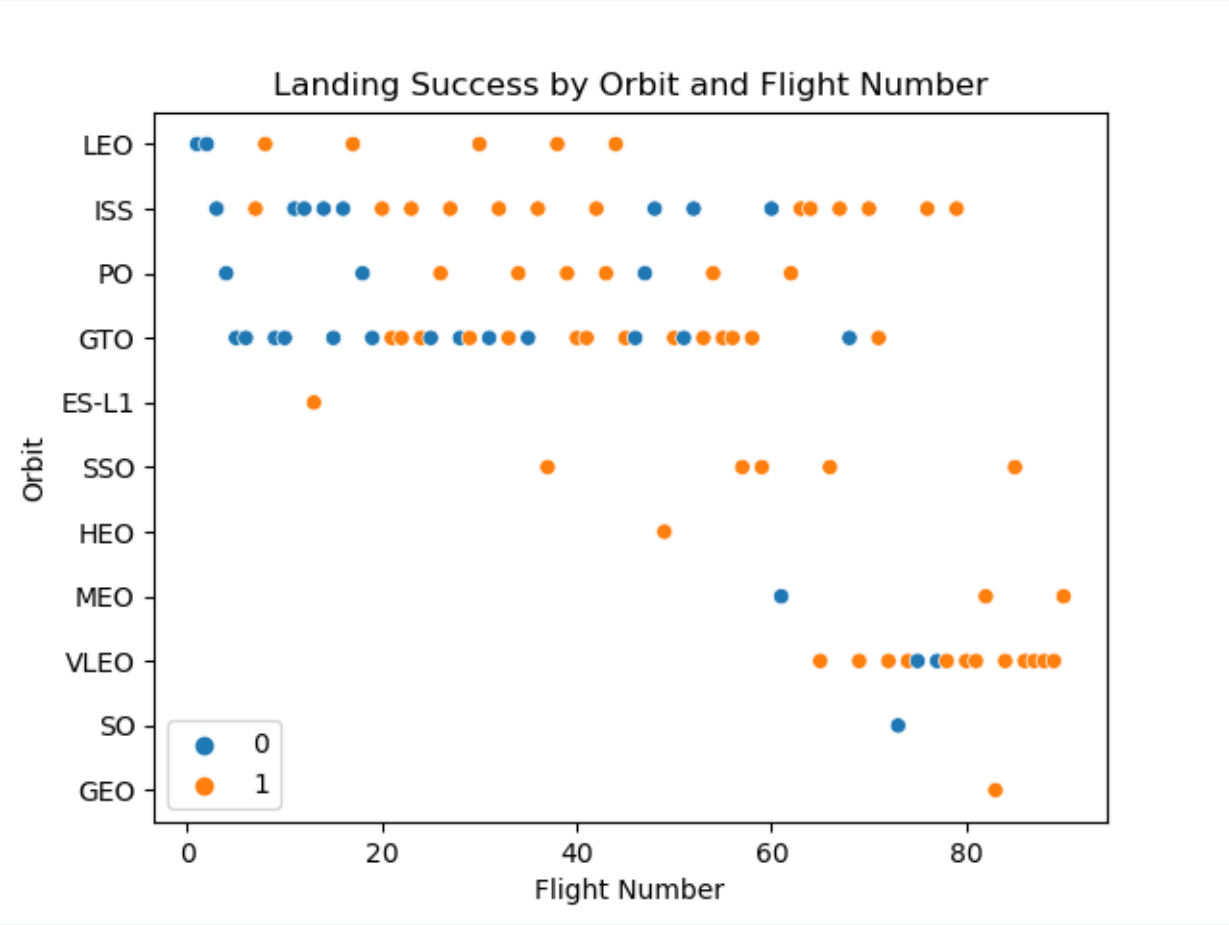

- Landing success at KSC an VAFB are not influenced by payload mass.
- Landing success at CCAFS improves with increasing payload mass.

# Success Rate vs. Orbit Type



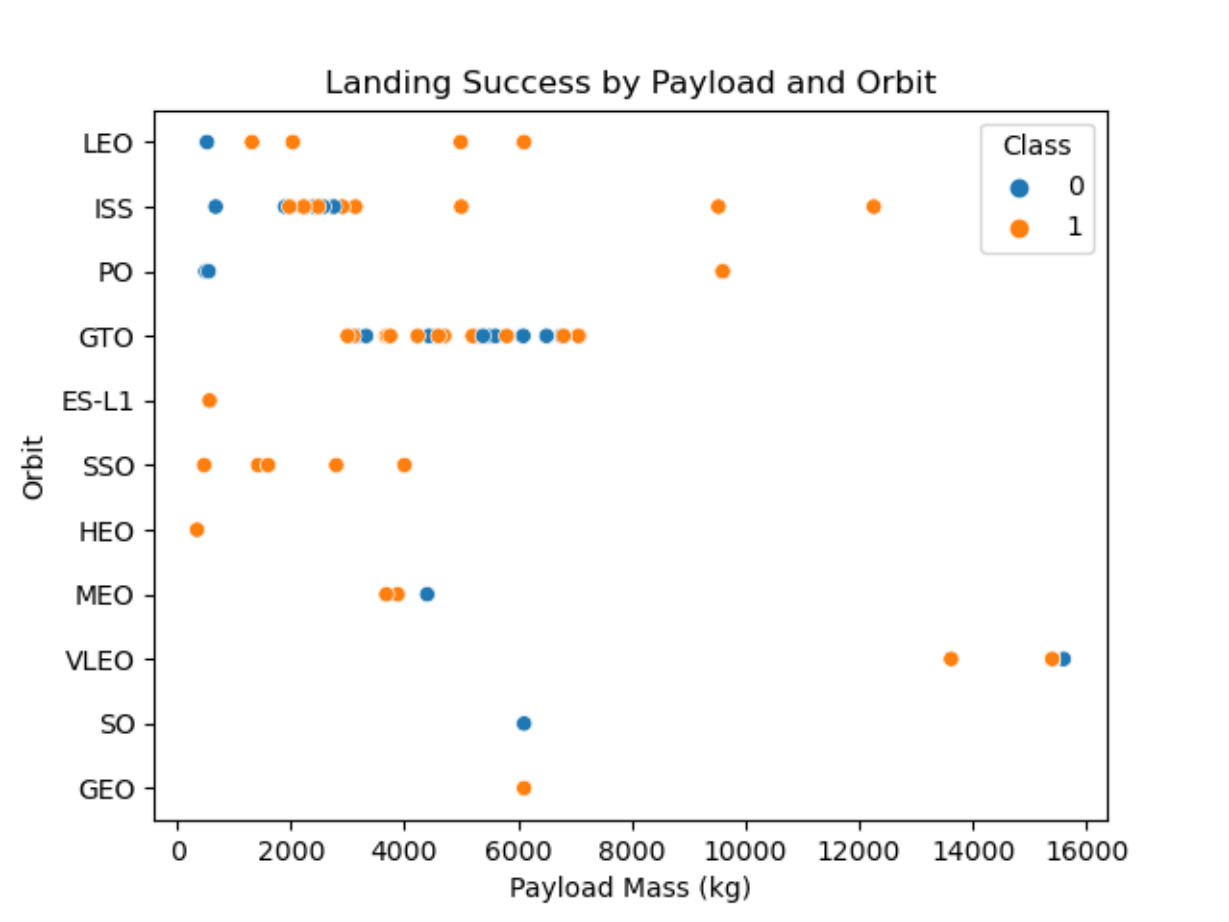

- Orbits ES-L1, GEO, HEO, and SSO have high landing success rates.
- SO (Solar Orbit) has not had any successful landings.

# Flight Number vs. Orbit Type



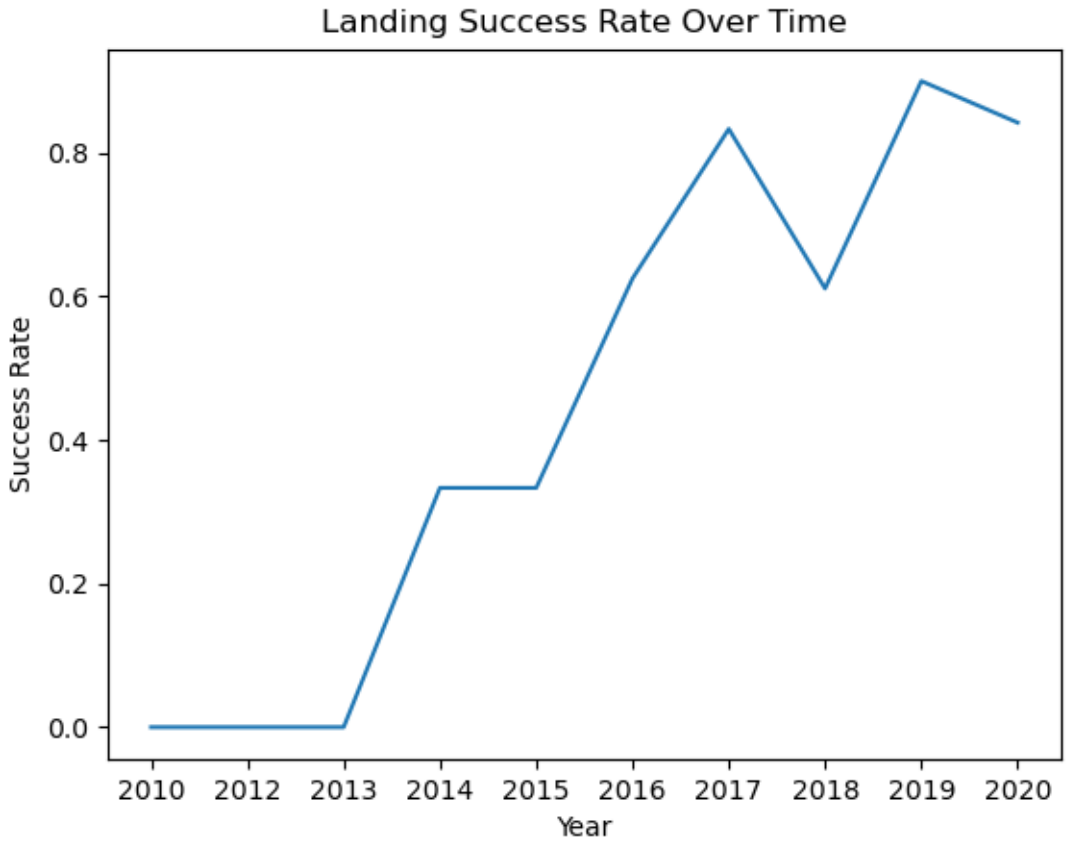

- Success by orbit I may be misleading due to small sample sizes for GEO,SO,MEO,HEO, and ES-L1.
- Successful landings are dependent upon launch number more than orbit type.

# Payload vs. Orbit Type




- There is only weak correlation between payload mass and landing success.
- SSO and LEO appear to be the most successful orbits with any payload mass.

# Launch Success Yearly Trend




- Landing success was not achieved before 2012.
- Success generally increases with time.
- Success dip in 2018 is suspicious and may not reflect larger trend.
- Success rate may be reaching a stable level around 0.8

# All Launch Site Names

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

- Four distinct launch sites are present in the spacextable dataset.

# Launch Site Names Begin with 'CCA'

| Launch Site |
| --- |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

- The first 5 records in the dataset where launch sites begin with `CCA` are form the same site, CCAFS LC-40

# Total Payload Mass

| Total NASA Payload (kg) |
| --- |
| 45596 |

- The total payload for all launches with NASA (CRS) as the customer is 45596 kg.

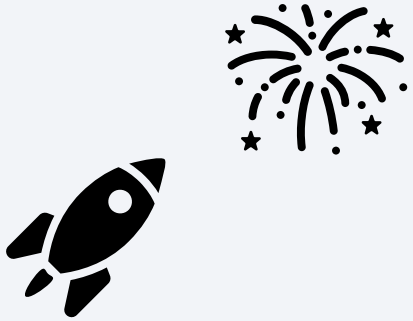
# Average Payload Mass by F9 v1.1

| Falcon9 v1.1 Average Payload (kg) |
| --- |
| 2534.6666666666665 |

- The average payload for a Falcon 9 v1.1 booster is approximately 2535 kg.

# First Successful Ground Landing Date

| Frist successful landing date |
| --- |
| 2015-12-22 |



- By sorting the data by landing date, and filtering for landing success, the date of the first successful landing is found by retrieving only the first value in the returned query results.

# Successful Drone Ship Landing with Payload between 4000 and 6000

| Boosters on Boats |
|---|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

- Four different boosters successfully landed on drone ships after launching with a payload between 4000 and 6000 kg.

# Total Number of Successful and Failure Mission Outcomes

| Count | Type |
|---|---|
| 100 | Missions Successful |
| 1 | Missions Failed |

- Only one of the Falcon 9 launches was considered by SpaceX to be unsuccessful.
- This suggests that most of the failed landings were expected or acceptable to SpaceX.

# Boosters Carried Maximum Payload

| Booster |
| --- |
| 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 |

- Only the Block 5 version of the Falcon 9 booster has been launched with the maximum payload.

# 2015 Launch Records

| Month | Outcome | Booster | Launch Site |
|---|---|---|---|
| 01 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 03 | Failure (drone ship) | F9 FT B1020 | CCAFS LC-40 |
| 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |
| 06 | Failure (drone ship) | F9 FT B1024 | CCAFS LC-40 |

- The only landing failures in 2015 occurred with drone ships as the landing site.
- No landing failures occurred after June 2015.

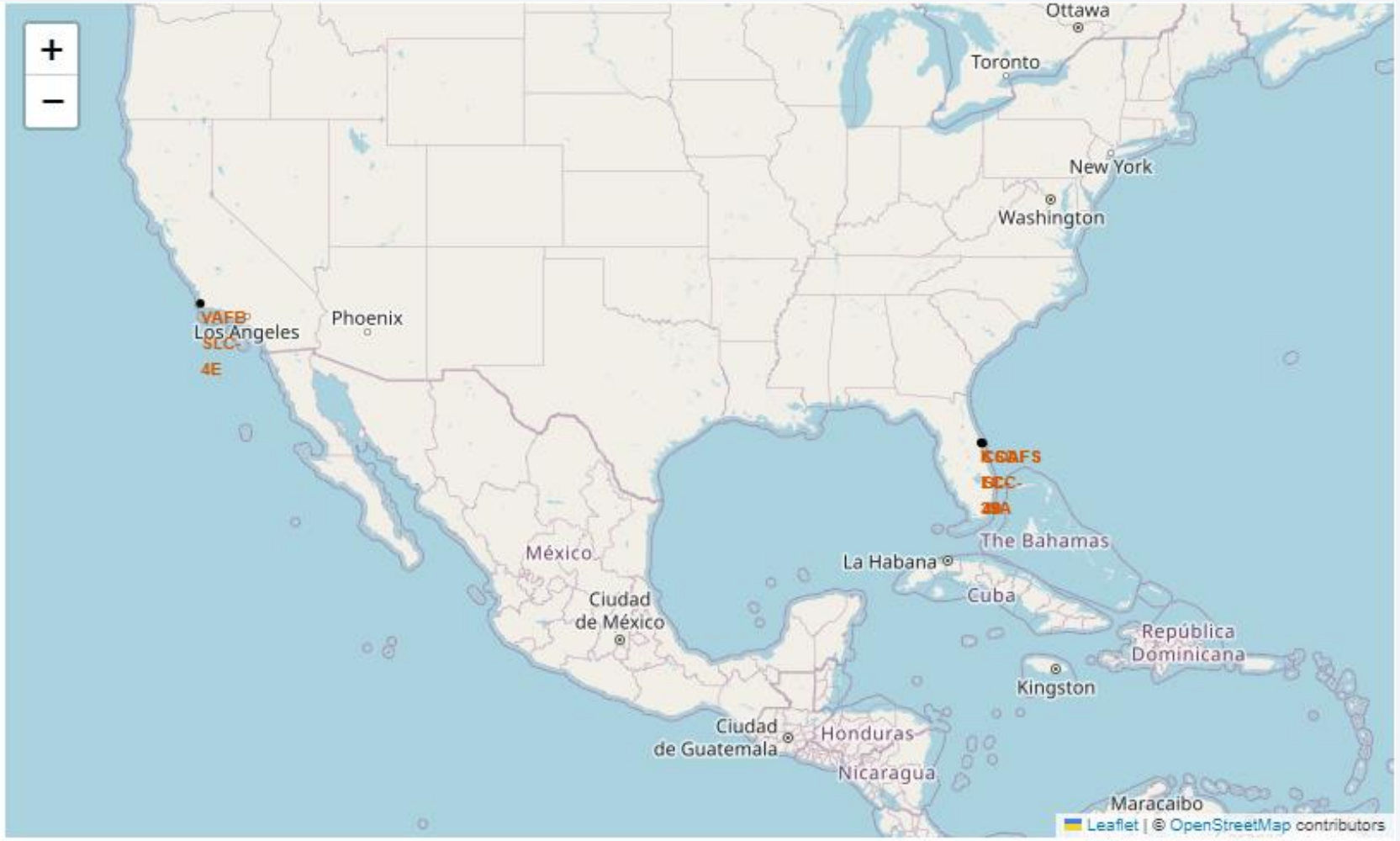
# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

- The largest group of landing outcomes indicates that landing was not attempted.
- Of the landings attempted, approximately half were successful.

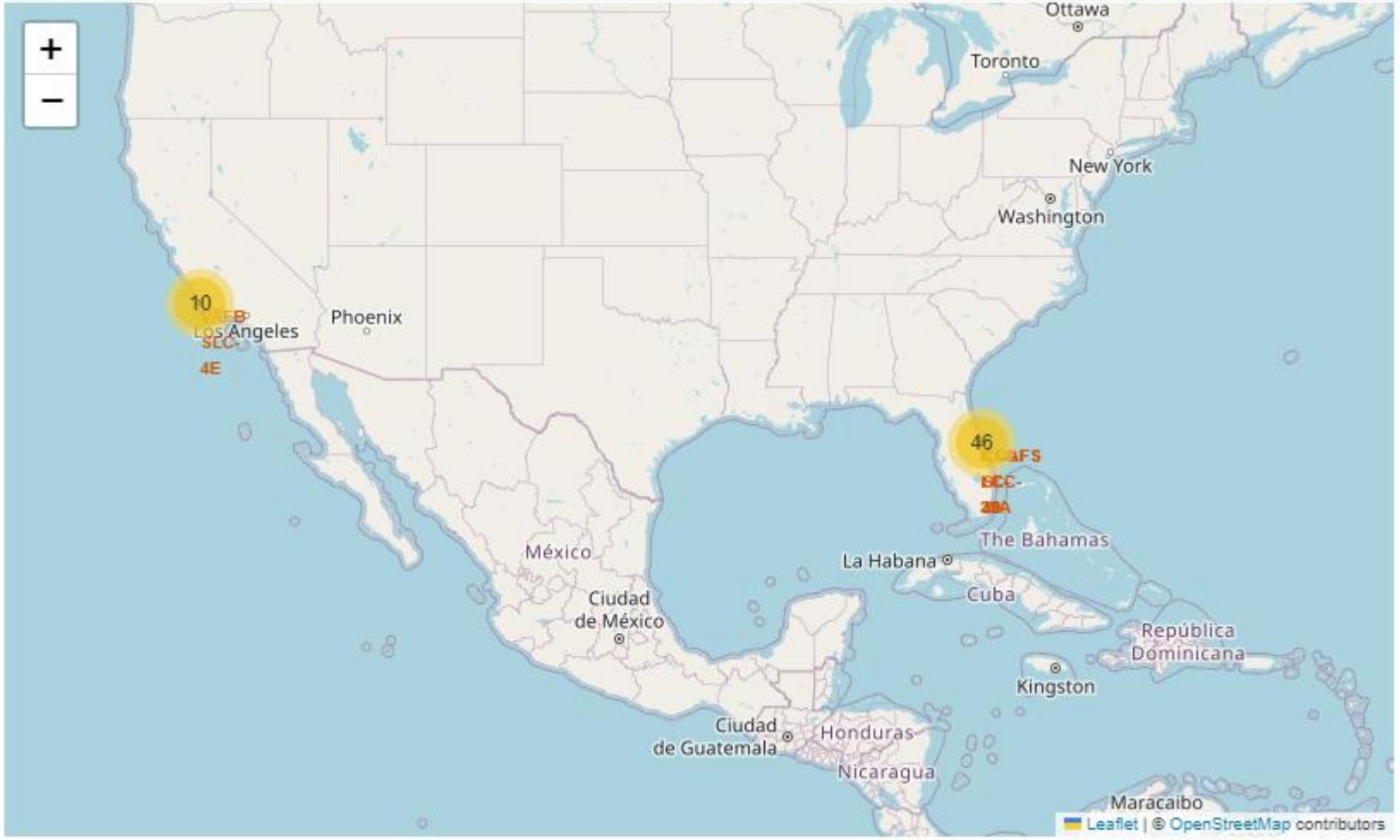

Section 3
Launch Sites
Proximities Analysis

# Falcon 9 Launch Sites



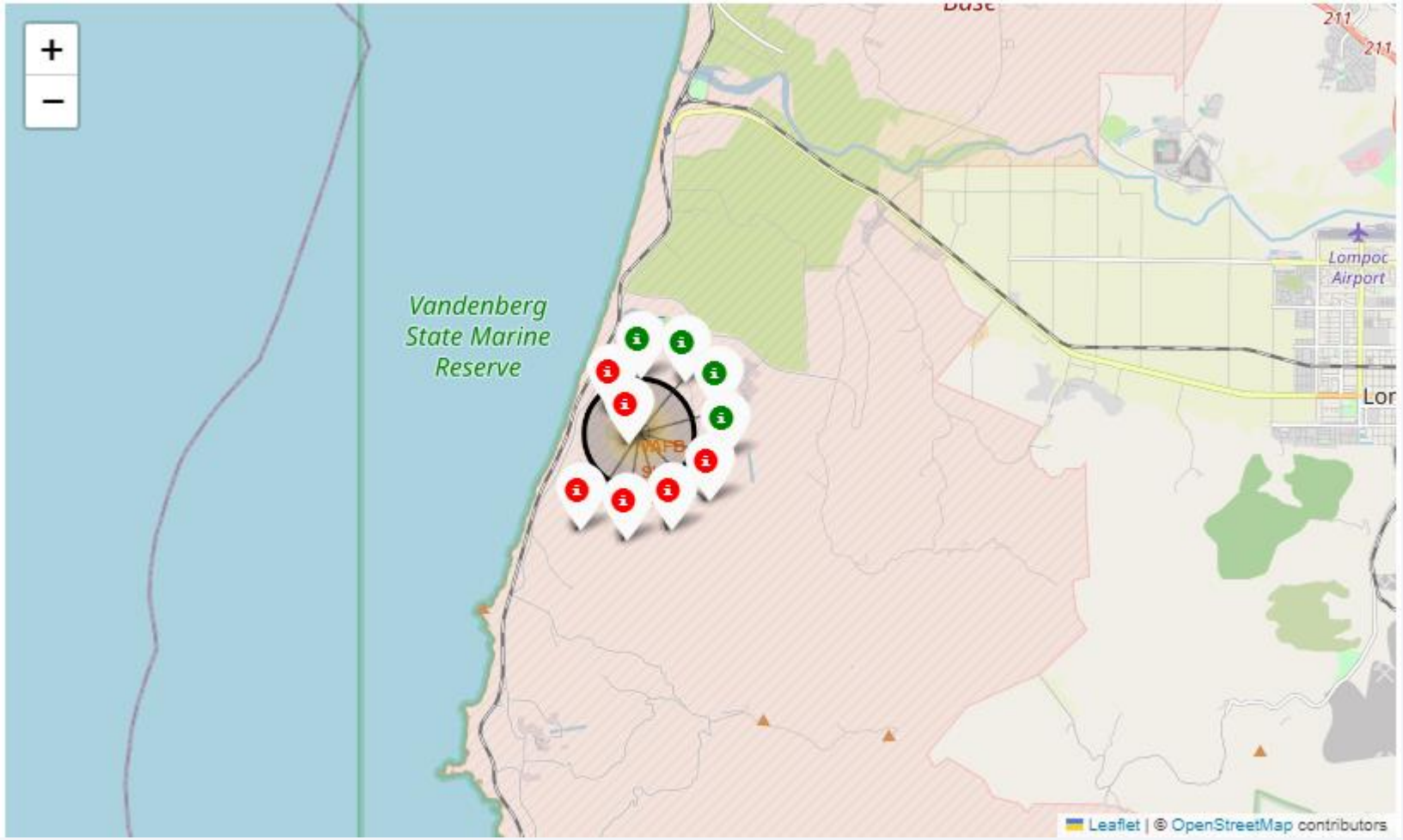

- All launch sites are in the Southern United States near the ocean coast.

# Launches at Each Launch Site



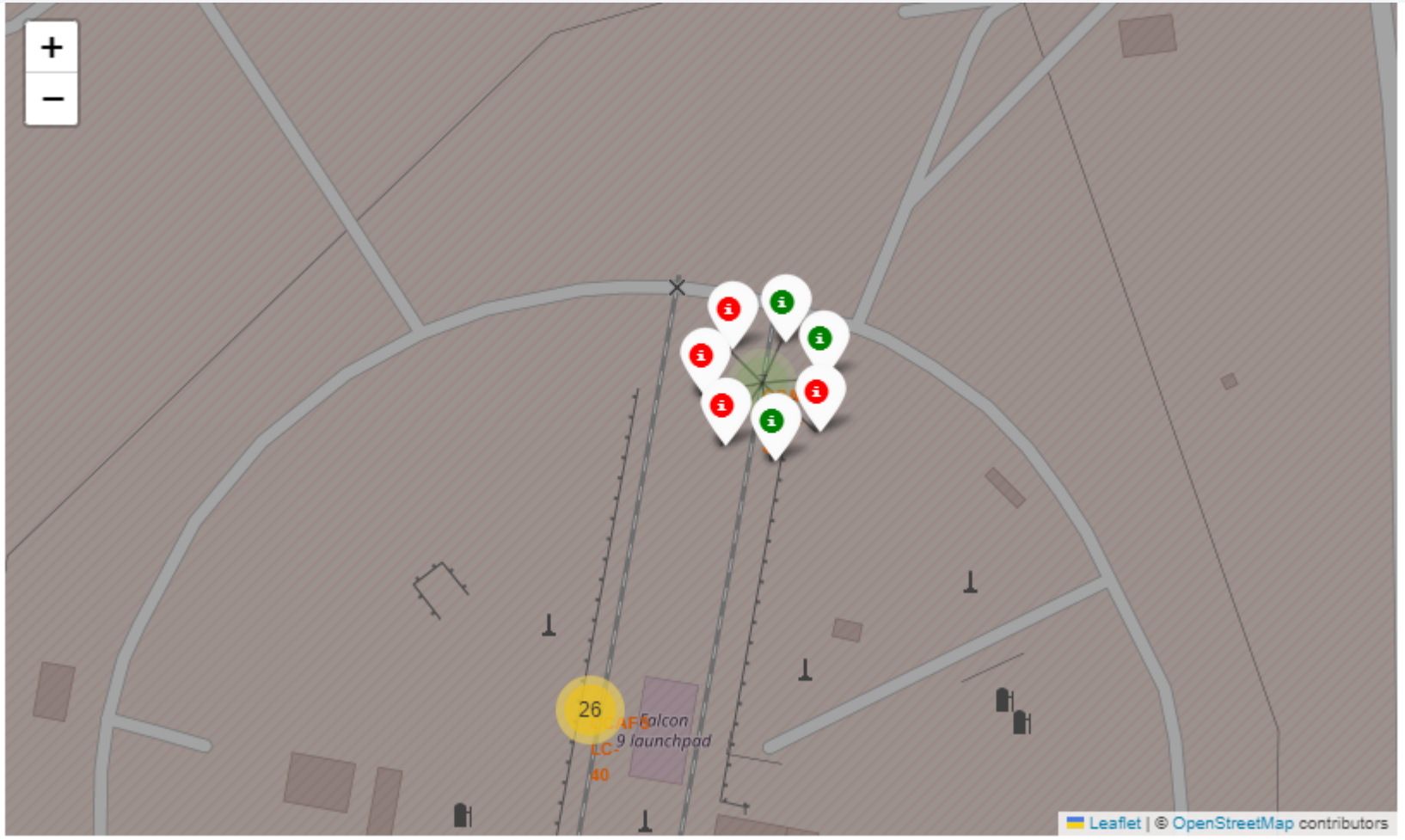

- The Kennedy Space Center launch pads are use much more often than the pad a Vandenburg AFB.

# Launches at VAFB



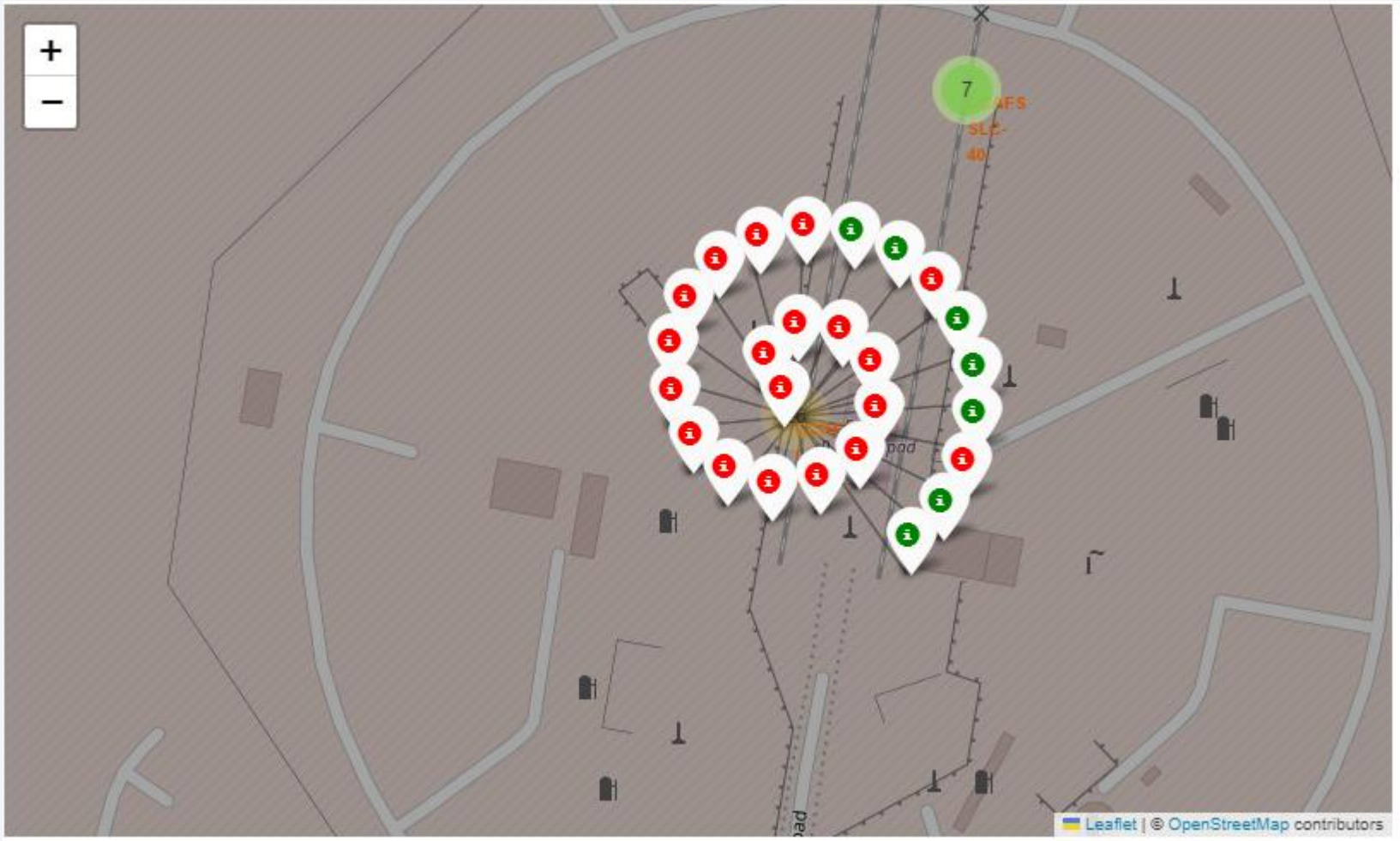

- VAFB launches have a roughly 50% success rate

# Launches at CCAFS SLC-40




- SLC-40 has an approximately 50% success rate

# Launches at CCAFS LC-40

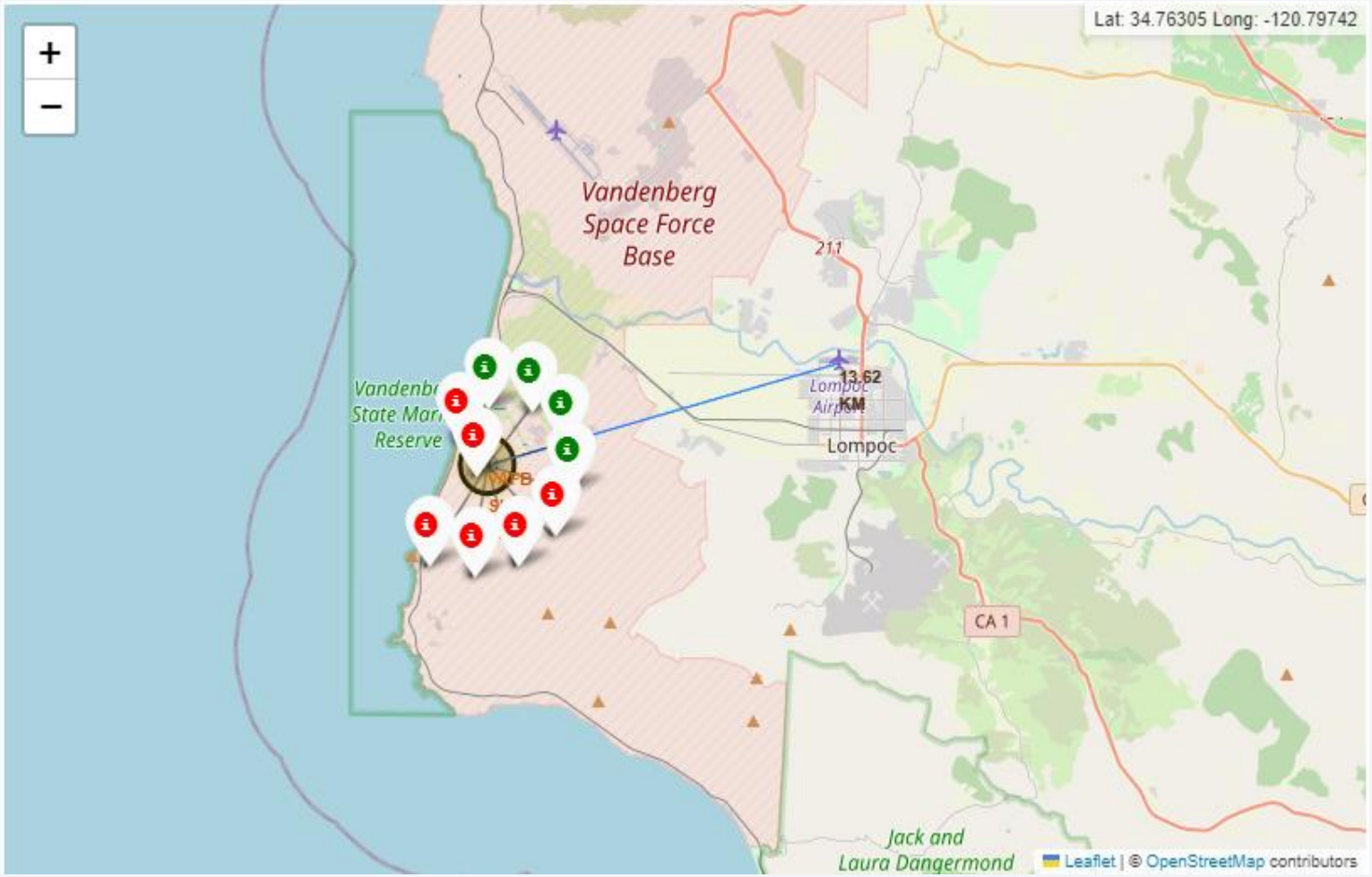



- LC-40 has launched mostly rockets with failed landings

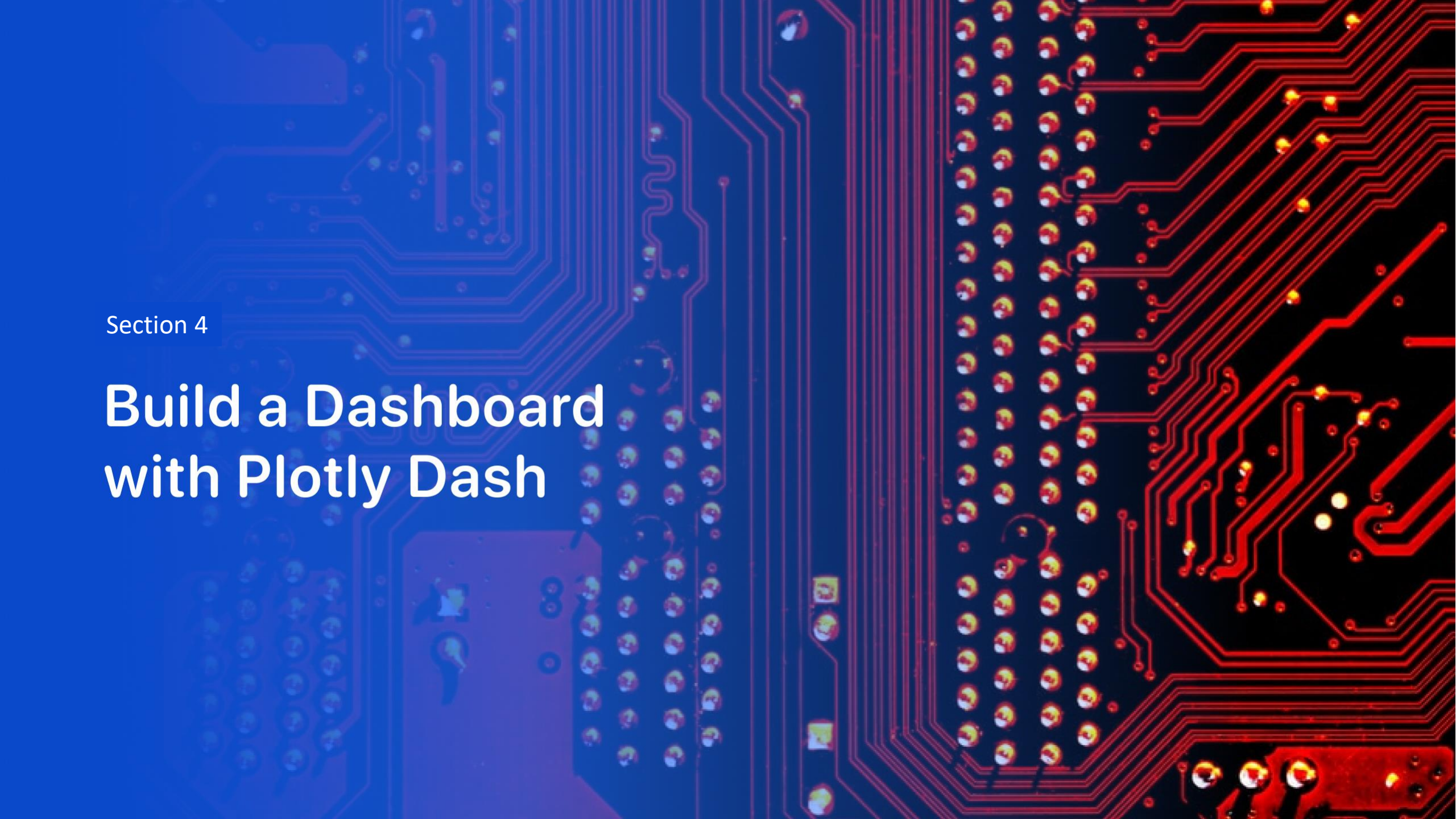
# Launches at CCAFS LC-39A


- LC-39A has launched mostly rockets with successful landings.

# Intimidation Distance



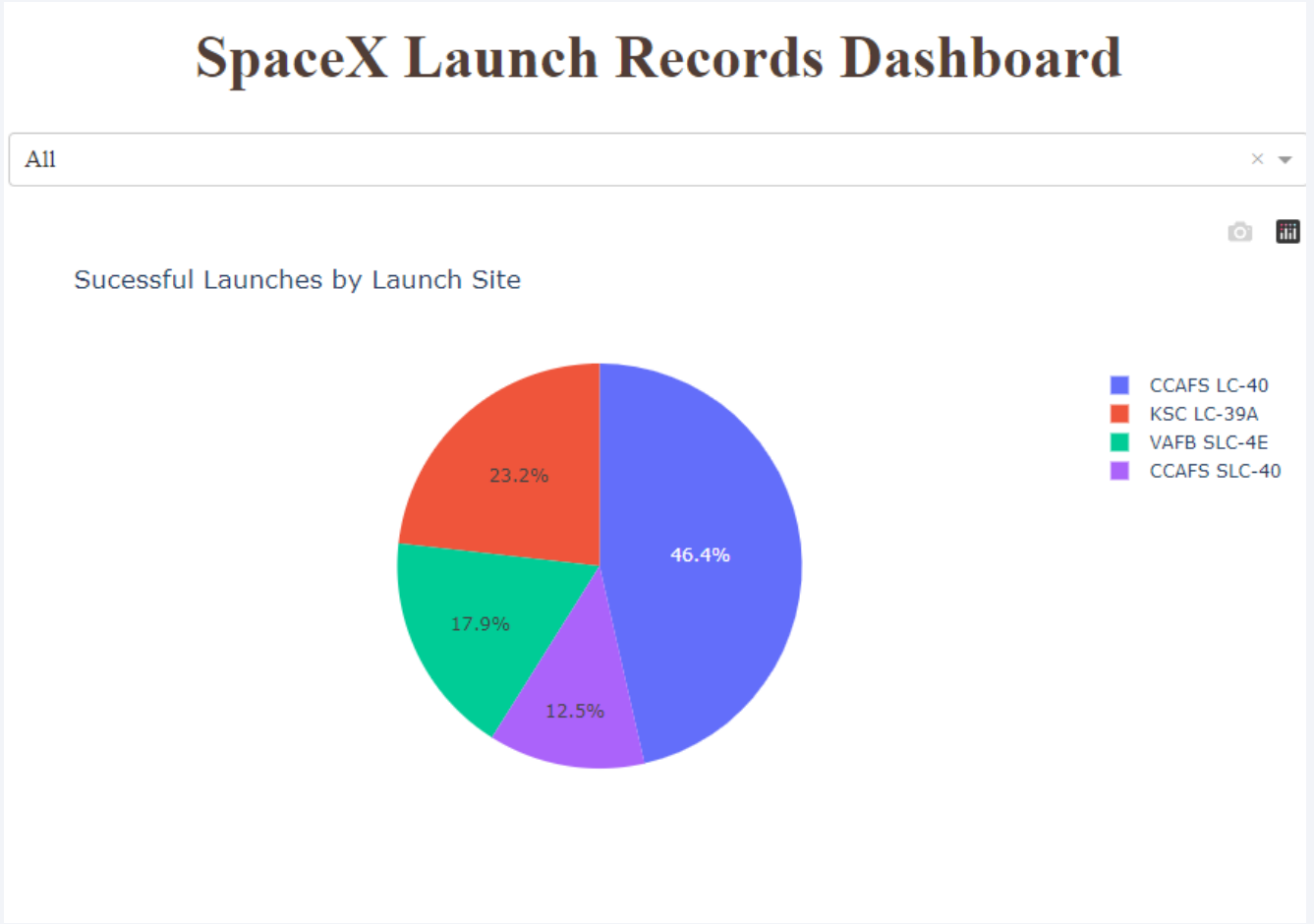

- The launch pad at Vandenbeg Space Force base is approximately 13.6 km from the nearest commercial airport in Lompoc.
- This small distance was chosen strategically to intimidate airline passengers with the large rocket trails.
- Passenger afraid to fly on airplanes are more likely to use spaceflight for their next vacation travels.


Section 4
Build a Dashboard with Plotly Dash

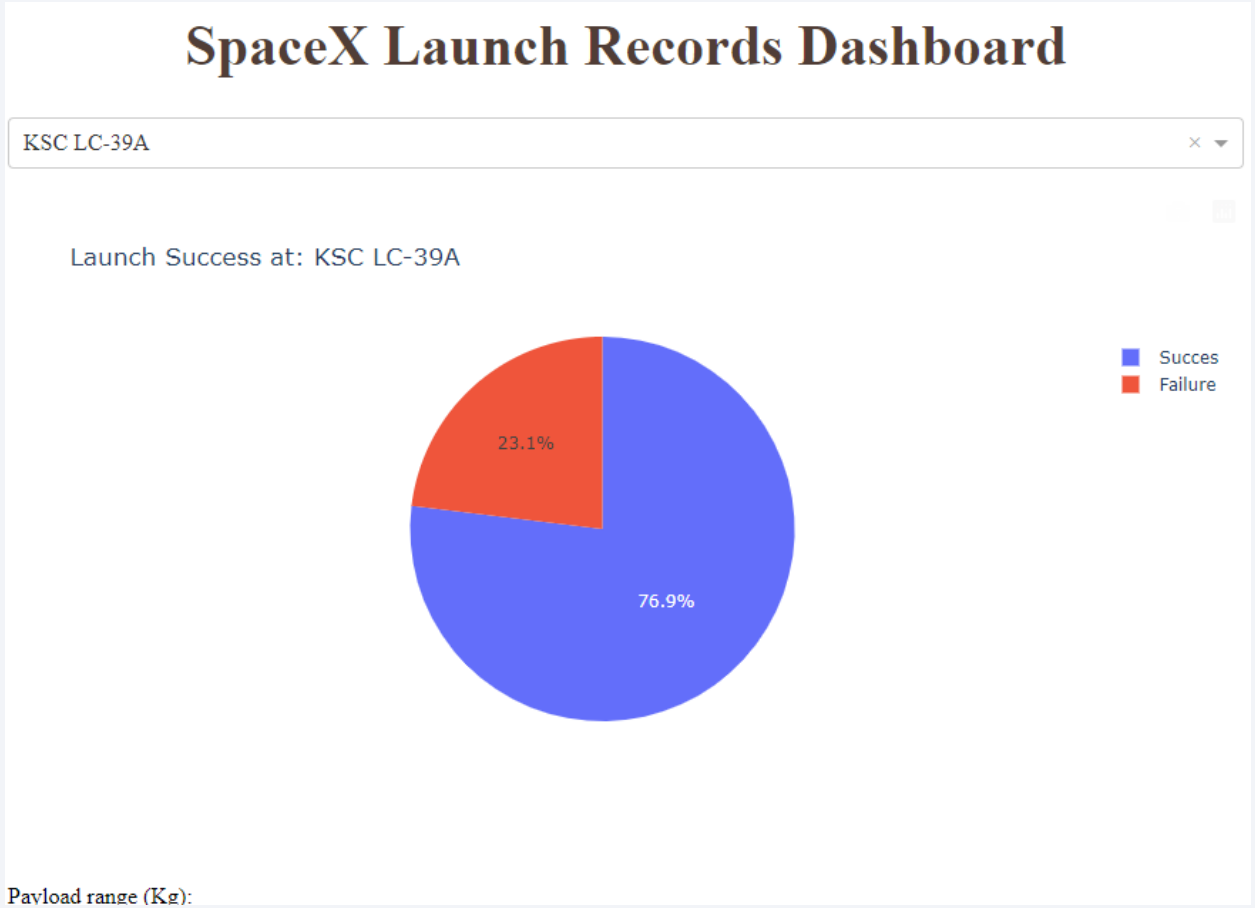
# Data Dashboard: Launch Success Pie Chart

- When “ALL” is selected from the launch site dropdown a pie chart is displayed to visualize the portion of successful launches contributed by each launch site.

- CCAFS-LC40 contributes the larges portion of successful launches. This may be due to the total number of launches and is not representative of the success rate at a given site.

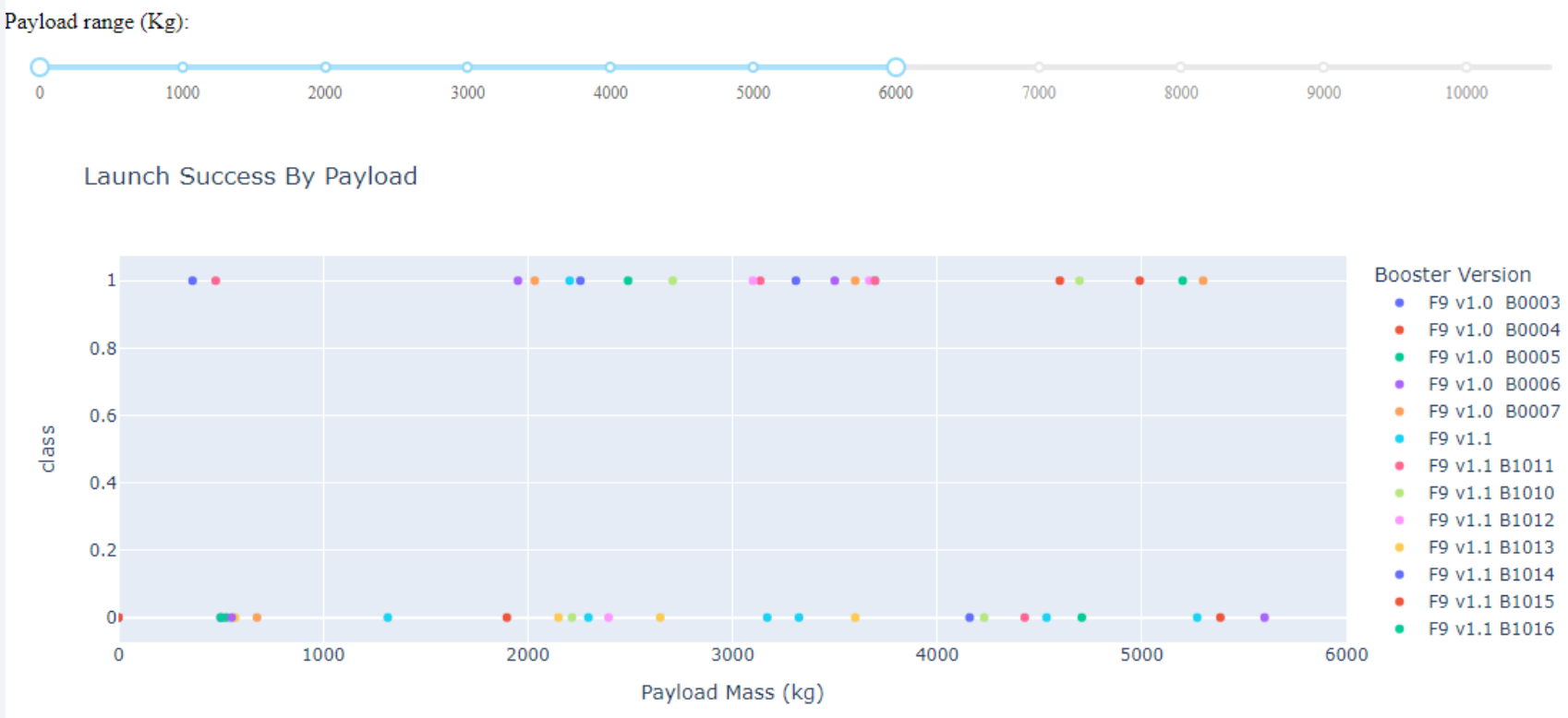


# Data Dashboard: Launch Site Success Pie Chart




- Selecting KSC –LC39A displays a pie chart illustrating launch success at the selected launch site.
- Launches from KSC LC-39A have a ~77% rate of success.
- This rate is approximately equal to the overall success rate for the most recent years in the dataset.

# Data Dashboard: Success By Payload




The payload slider limit the rang of payloads shown on the categorical scatter plot.

- Limiting the maximum payload displayed to 6000kg shows a fairly even distribution of success and failure for all displayed payloads.

# Data Dashboard: Success By Payload

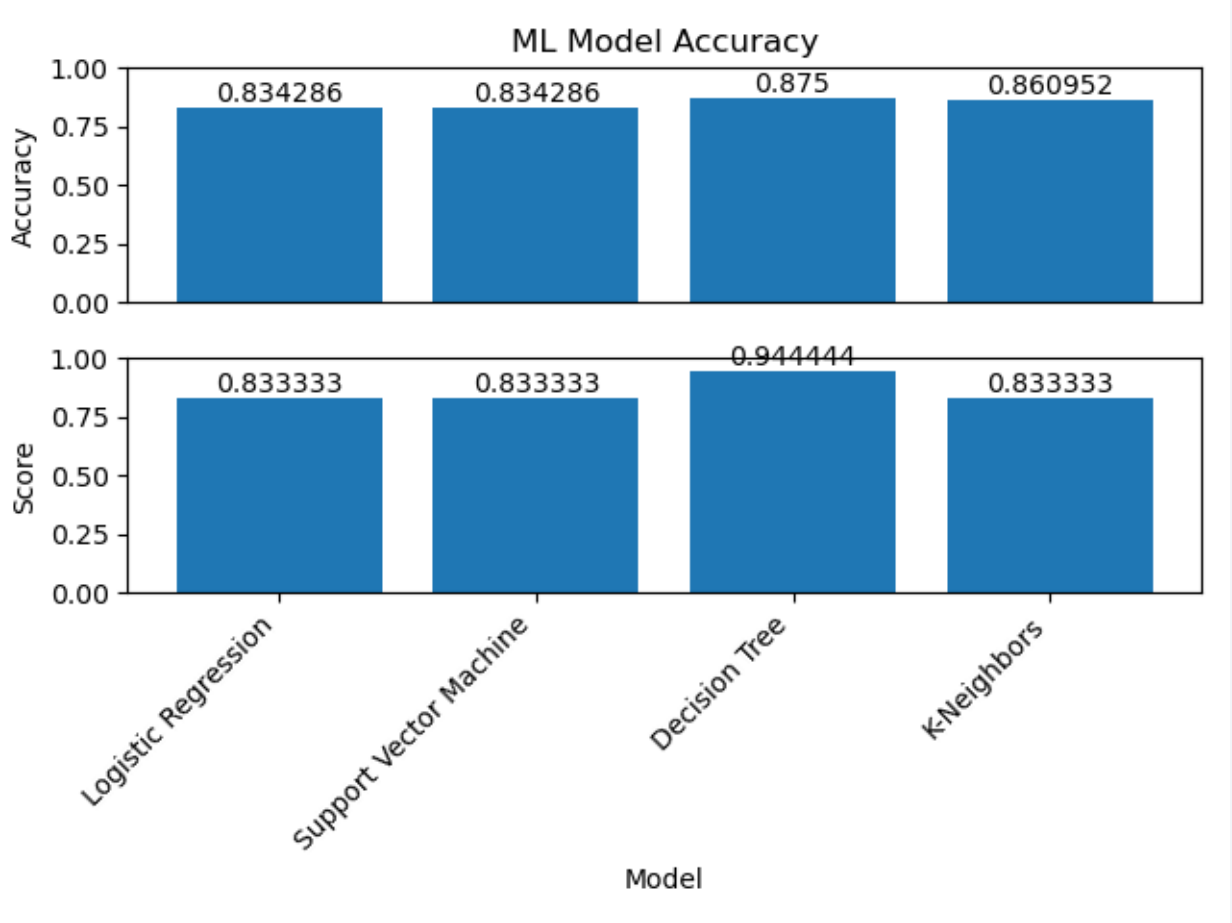

- With the payload slider set to a minimum of 6000kg a different patter is apparent.
- Launches with large payloads are almost entirely successful.

Section 5

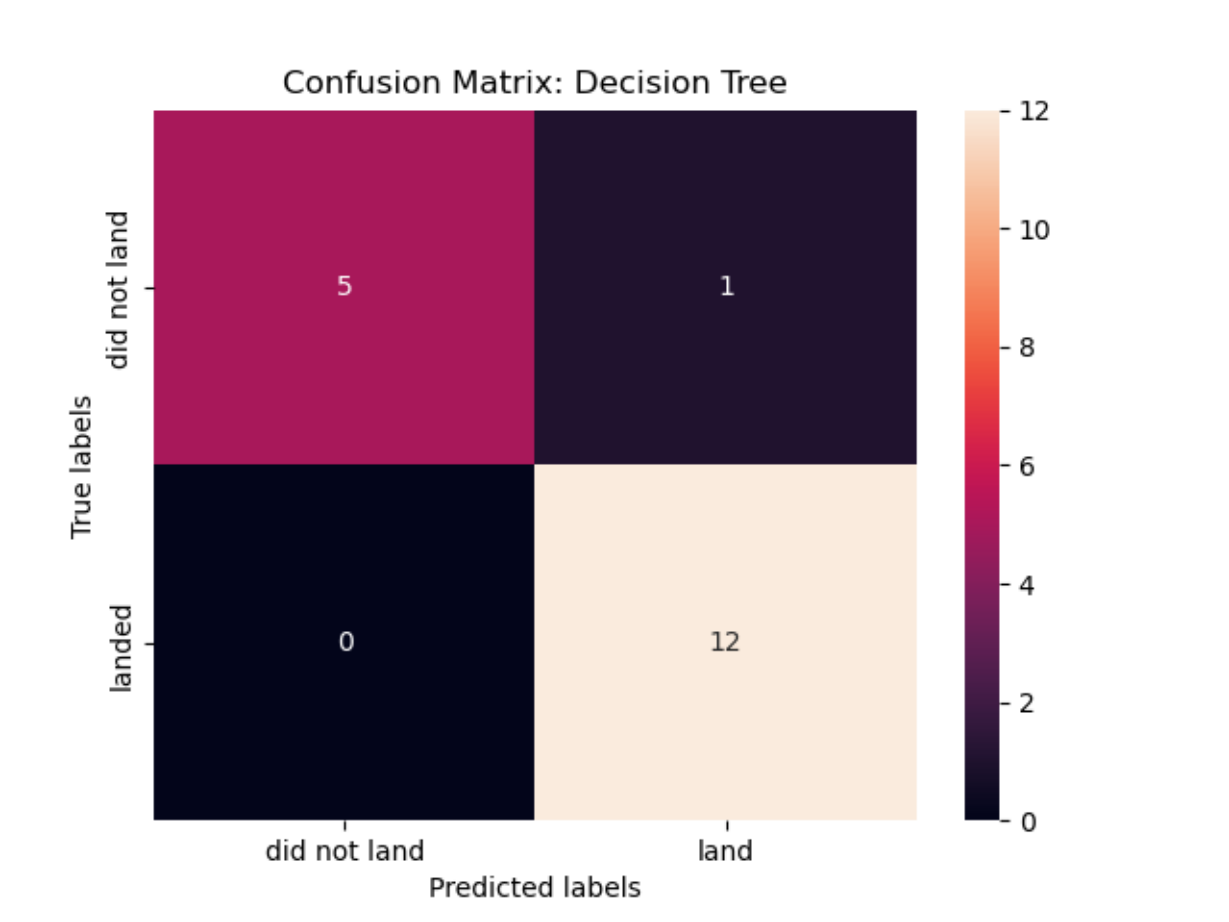
# Predictive Analysis (Classification)

# Classification Accuracy

- All Machine Learning classifier models perform at nearly the same level of accuracy.
- The Decision Tree classifier cross-validation found a slightly better fitting classified.
- When tested against the out of sample test data, the decision tree classifier performs significantly better then the other classifiers.



# Confusion Matrix




- When compared to the test data split, the decision tree classifier incorrectly classified only one launch outcome.
- One of the launches which did no land was predicted by the classifier to land.

# Conclusions

- Successful landings do not strongly depend on easily controllable factors (location, payload, booster type, etc.)
- Success rates can appear to be skewed based on the number of launches for a given launch site.
- The overall trend in landing success is that the success rate increases with more experience, number of launches, indicating continued improvement.

# Thank you!