IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Joseph Line>
<3/13/24>

Outline

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

Executive Summary

Summary of methodologies

In an effort to identify the impact of multiple variables on the success of a reusable rocket landing, the following research methodologies were used:

- **Collecting** data using SpaceX Data Collection API and webscraping
- **Structuring** data to allow for better exploratory data analysis
- **Exploring/Analyzing** data using SQL to create graphs and identify possible trends
- **Building** machine learning models based on the data set to predict landing outcomes using logistic regression, support vector machine (SVM), decision trees, and K-nearest neighbor (KNN)

Summary of all results

- Successful landing has improved over time
- Payload has a significant impact on success rate
- Predictive analysis models built based on data can provide an **83%** accurate prediction on landing success

Introduction

Background

In an effort to make space travel affordable for everyone, SpaceX has developed a reusable first stage rocket allowing for relatively inexpensive launches ($62 million per launch, while other providers, which are not able to reuse the first stage, cost upwards of $165 million each). However, the first stage rocket does not have a 100% successful landing rate. By determining if the first stage will land, we can determine the effective price of the launch. To do this, we can use public data, web-scrapping, and machine learning models to predict how likely it is that the first stage rocket will land and what determining factors, if any, contribute most significantly to a successful or unsuccessful outcome.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using SpaceX REST API and web-scraping
- Perform data wrangling
  - Filtered data including normalization of missing values and preparing necessary transformation of data via one hot encoding for data modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models built and tested with multiple train/test splits to ensure optimal model parameters are met
  - Once train/test split of data is complete, use of multiple models (to include using logistic regression, support vector machine (SVM), decision trees, and K-nearest neighbor (KNN)) was employed to verify model accuracy

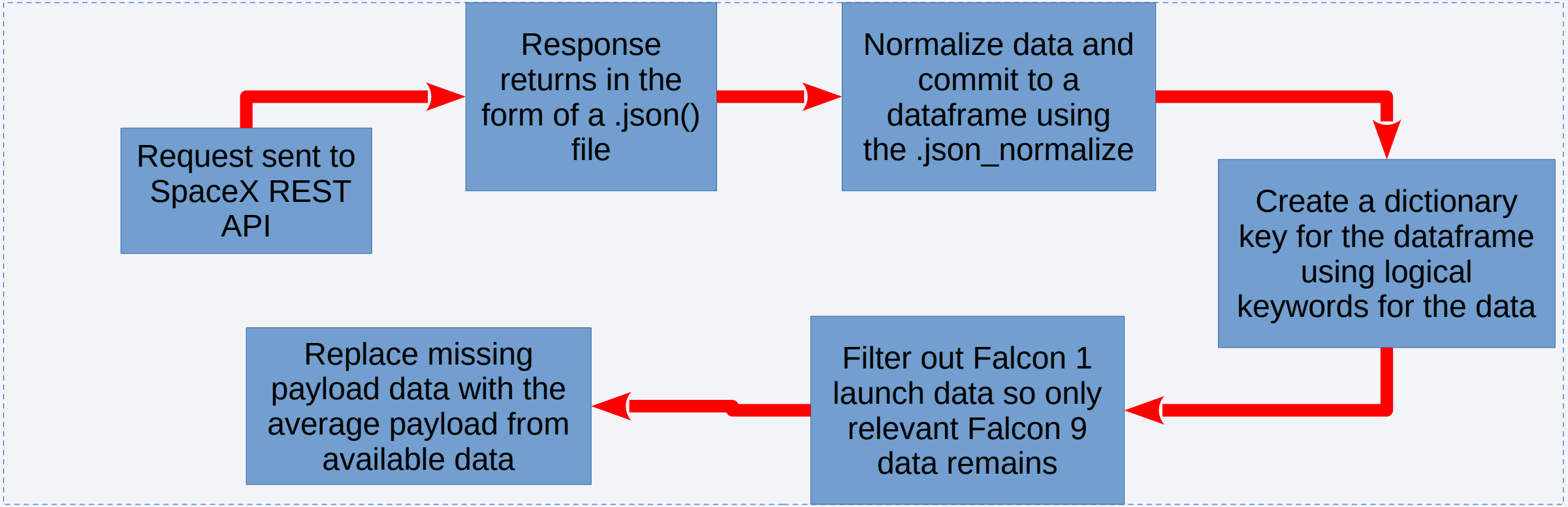
Data Collection

Steps taken to collect data include:

- Requesting data from SpaceX REST API
- Decoding data in python using .json_normalize() to set information to a flat table
- Transforming data to remove any N/A information and replacing ID# in our data set to the corresponding true data
  - Removal of Falcon 1 data as we are only concerned with Falcon 9 data
  - Replacing missing Payload data with the mean value for the available payload data

Data Collection – SpaceX API

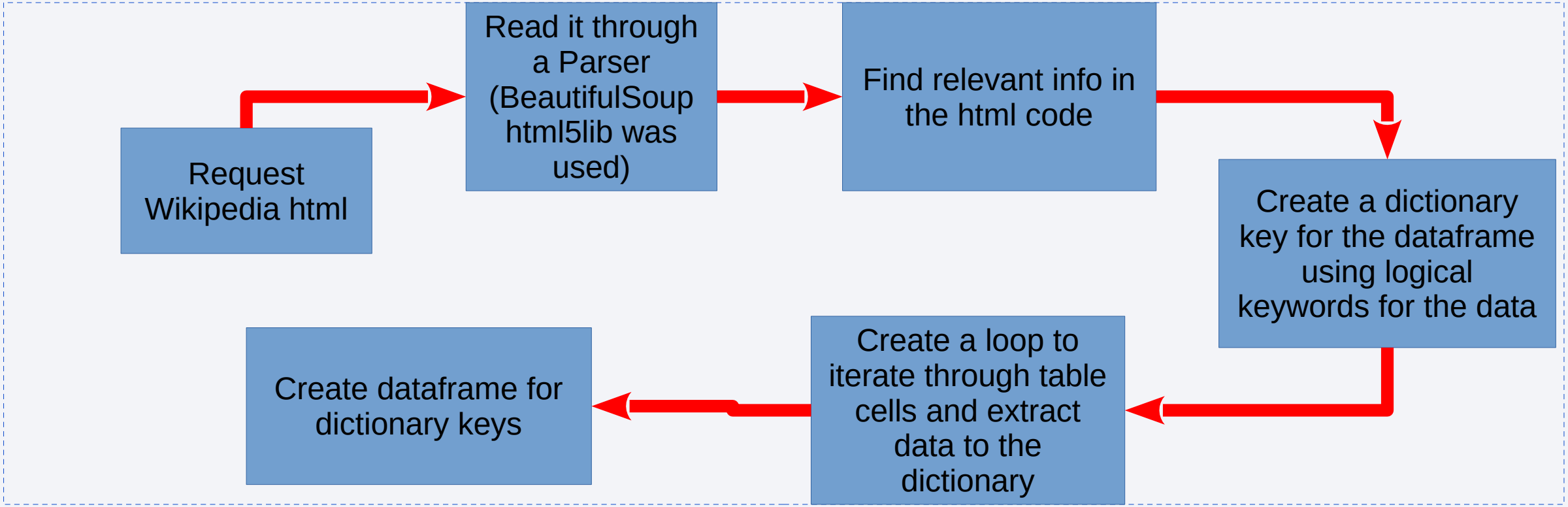
Data Collection description flow chart and link to Data Collection below



Github URL:

https://github.com/jline0708/DS_Capstone/blob/main/01_Data_Collection_API_Lab%20.ipynb

Data Collection cont. – Scraping



Github URL:

https://github.com/jline0708/DS_Capstone/blob/main/02_Webscraping_Lab.ipynb

Data Wrangling

Data Wrangling was processed through:

- OneHotEncoding to translate multiple ‘string’ values (True ASDS, False ASDS, True Ocean, False Ocean, etc..) into the corresponding digits represented by a success (1) or failure (0) and creating a new column with those values for streamlined data management.
- Missing data was managed using the average value collected from the remaining data as to not lose valuable entries

Github URL:
https://github.com/jline0708/DS_Capstone/blob/main/03_Data_Wrangling_Lab.ipynb

EDA with Data Visualization

Exploratory Data Analysis was performed on multiple variables to include:

Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots arguments include:

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

Plot types include:

Scatter plots, line charts, and bar plots

Many different plot types and arguments were considered to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

https://github.com/jline0708/DS_Capstone/blob/main/05_EDA_DataViz_Lab.ipynb

EDA with SQL

Exploratory Data Analysis was continued using SQL queries to include:

- Names of unique launch sites
- Flights where the launch site began with “CCA”
- Total payload in kg from all flights attributed to “NASA (CRS)”
- The average payload in kg carried by booster version “F9 v1.1”
- Total successes and failures for landing the stage 1 rocket
- The count of landing outcomes (detailed ex: True Ocean, False RLTS, etc) between dates June 4, 2010 and March 20, 2017
- The names of booster versions which have carried the maximum payload
- Names of boosters which have successes in drone ship and carried greater than 4000 kg and 6000 kg

Github URL: https://github.com/jline0708/DS_Capstone/blob/main/04_EDA_SQL_Lab.ipynb

Interactive Folium Map

Using Folium, an interactive map was created to visualize certain attributes of the data to include:

Launch sites:

Markers were added to show launch site coordinates with popup labels to show the site name

Outcomes:

Green and red markers were utilized at each launch site to show successes and failures respectively as a way to visualize which launch sites had higher success rates

Distances:

Colored lines were used to show how far a launch site was from major landmarks including coastline, highways, and cities

https://github.com/jline0708/DS_Capstone/blob/main/06_Site_Locations_w_Folium_Lab.ipynb

Interactive Dashboard using Plotly Dash

Through the use of the Plotly Dash application, an interactive visual dashboard was created to allow for quick querying of specific data represented via graphs and plots.

The dashboard included:

- Dropdown for selecting launch site(s) which would update the rest of the dash in real time based on the launch site(s) selected
- A Pie chart showing the ratio of successes to failures for selected launch site(s)
- A slider to narrow or widen the range of payload mass that was represented in the graphs
- A scatter chart of Payload vs Success Rate by Booster Version to see if a correlation exists

https://github.com/jline0708/DS_Capstone/blob/main/07_Dashboard_App_w_Plotly_Lab.py

Predictive Analysis (Classification)

After the use of exploratory data analysis to confirm that a correlation exists between success/failure and the other variables, Predictive Analytics was employed to build a model that would accurately predict if a mission would succeed. To do this:

- A NumPy array was created from the class (success/failure) column
- Data was standardized with the StandardScaler.Fit function
- The data was then split into a Train and Test subsection with a test size of 20%
- Multiple Predictive Analysis methods were tested to see which would provide the best score including
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors

https://github.com/jline0708/DS_Capstone/blob/main/08_Machine_Learning_Prediction_Lab.ipynb

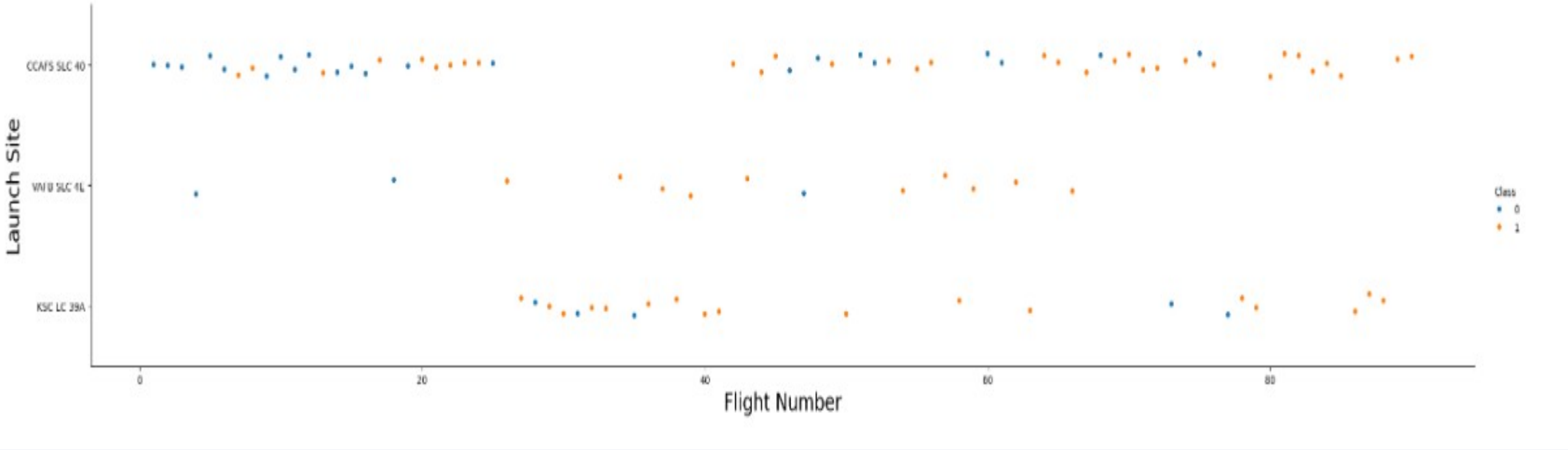
Results

- Successful landing has improved over time
  - As the flight number increased, a successful landing was more likely
- Payload has a significant impact on success rate
  - Rockets with payload of over 6000kg had a significantly higher success rate than those under 6000kg
- Predictive analysis models built based on data can provide an **83%** accurate prediction on landing success
  - With the data provided, our model can determine the success or failure rate of a future rocket based on its payload, launch site, and orbit type and thus give a better estimated cost

Section 2

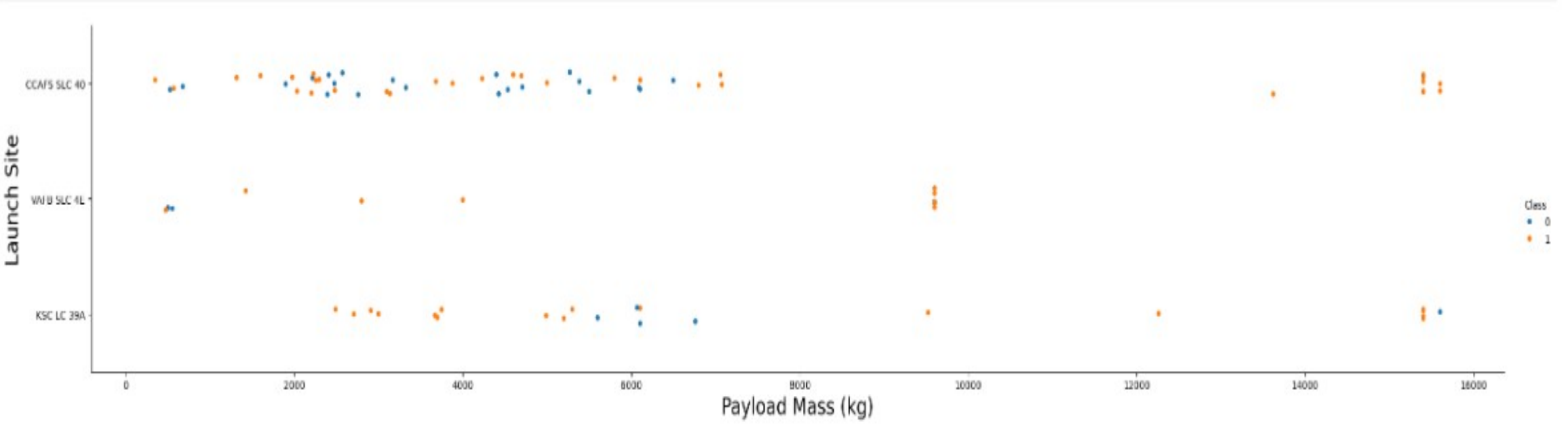
Insights drawn from EDA

Flight Number vs. Launch Site



- Of the first 25 missions, 23 were at CCAFS SCL 40 and over 60% were failures.
- Once multiple successes were strung together, the other 2 sites were used with more frequency
- Of the final 30 missions, failure rate had been decreased to 20%

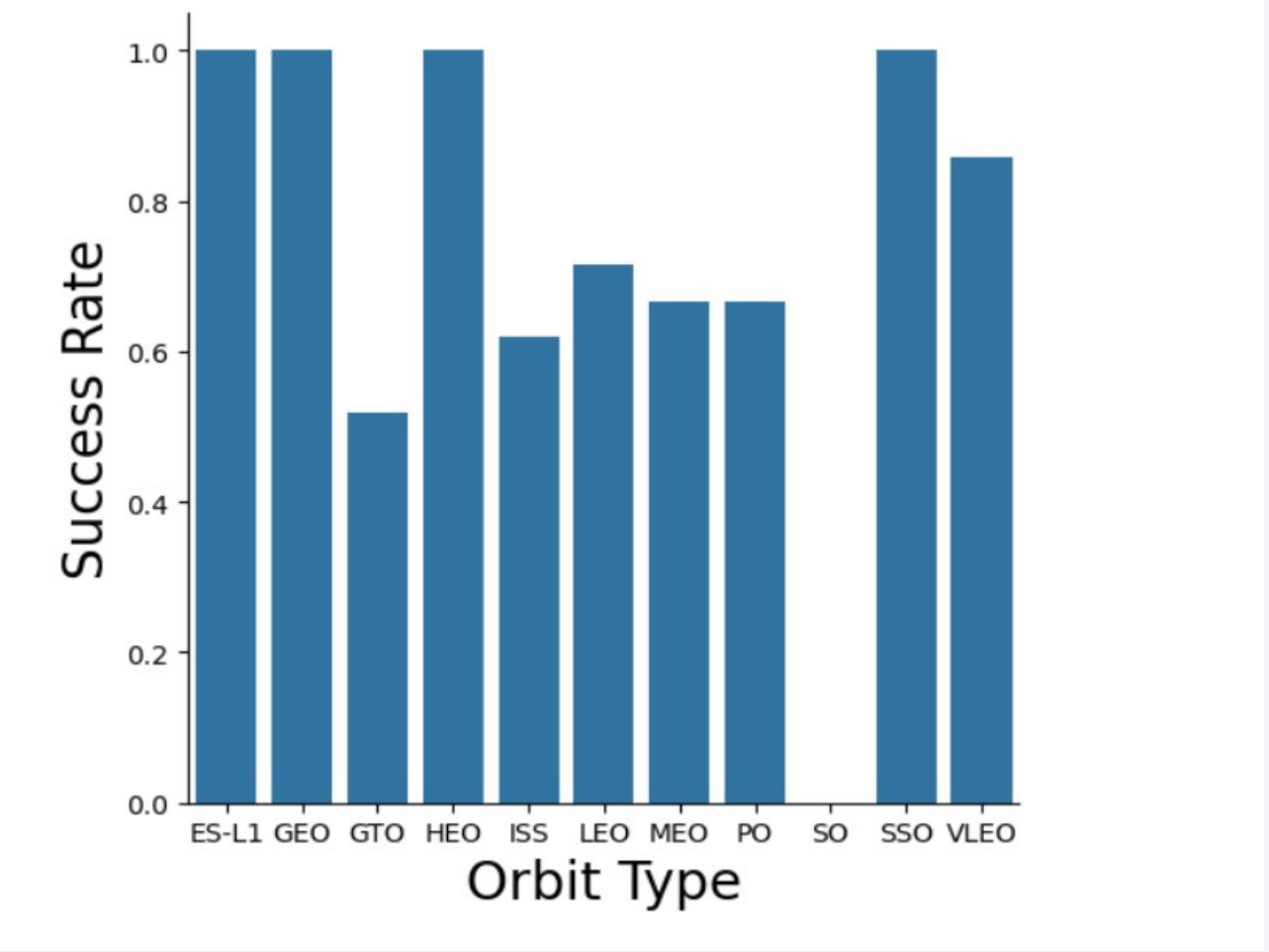
Payload vs. Launch Site



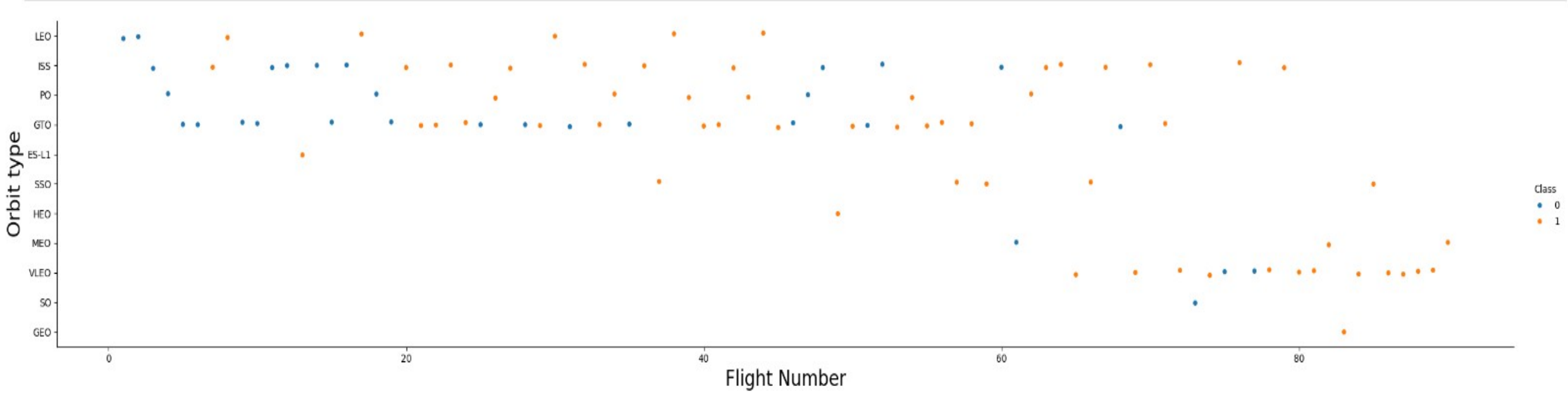
- Heavy Payloads (over 7,000kg) had near 100% success rate
- VAFB never had a launch with payload greater than 10,000kg
- The lightest (<1,000kg) and heaviest(>15,000kg) launches were from CCAFS SLC 40

Success Rate vs. Orbit Type

- 4 Orbit types (ES-L1, GEO, HEO, and SSO) had a 100% success rate
- SO had a 0% success rate
- All other orbits had a >50% success rate

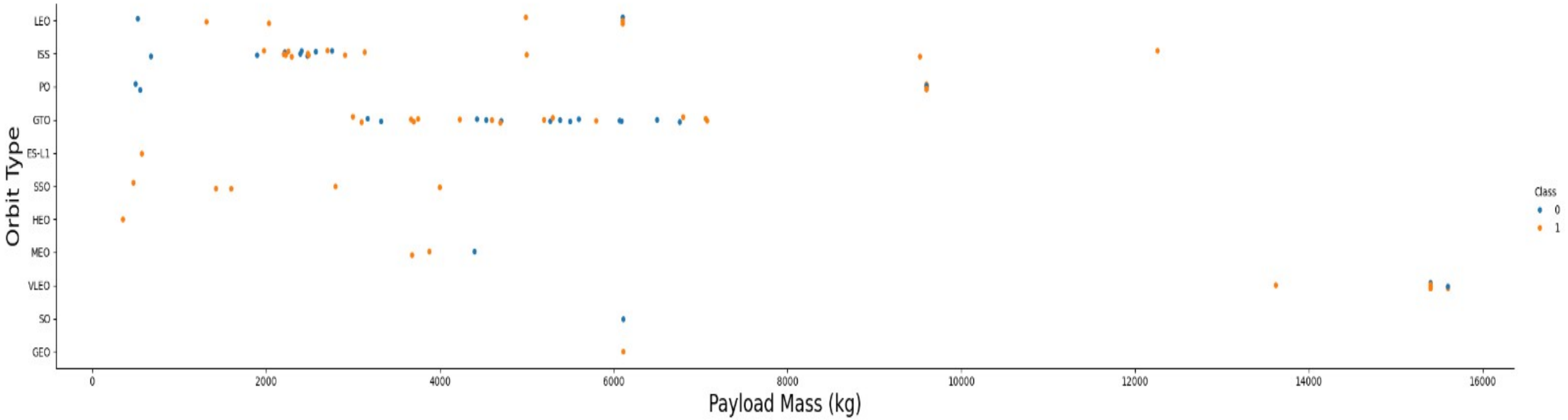


Flight Number vs. Orbit Type



- Most launches fell under LEO, ISS, GTO, or VLEO orbit types
- SO only had one launch and it failed meaning there is not enough data to infer all future launches will also fail
- Of the 4 100% success orbit types, only SSO had multiple launches

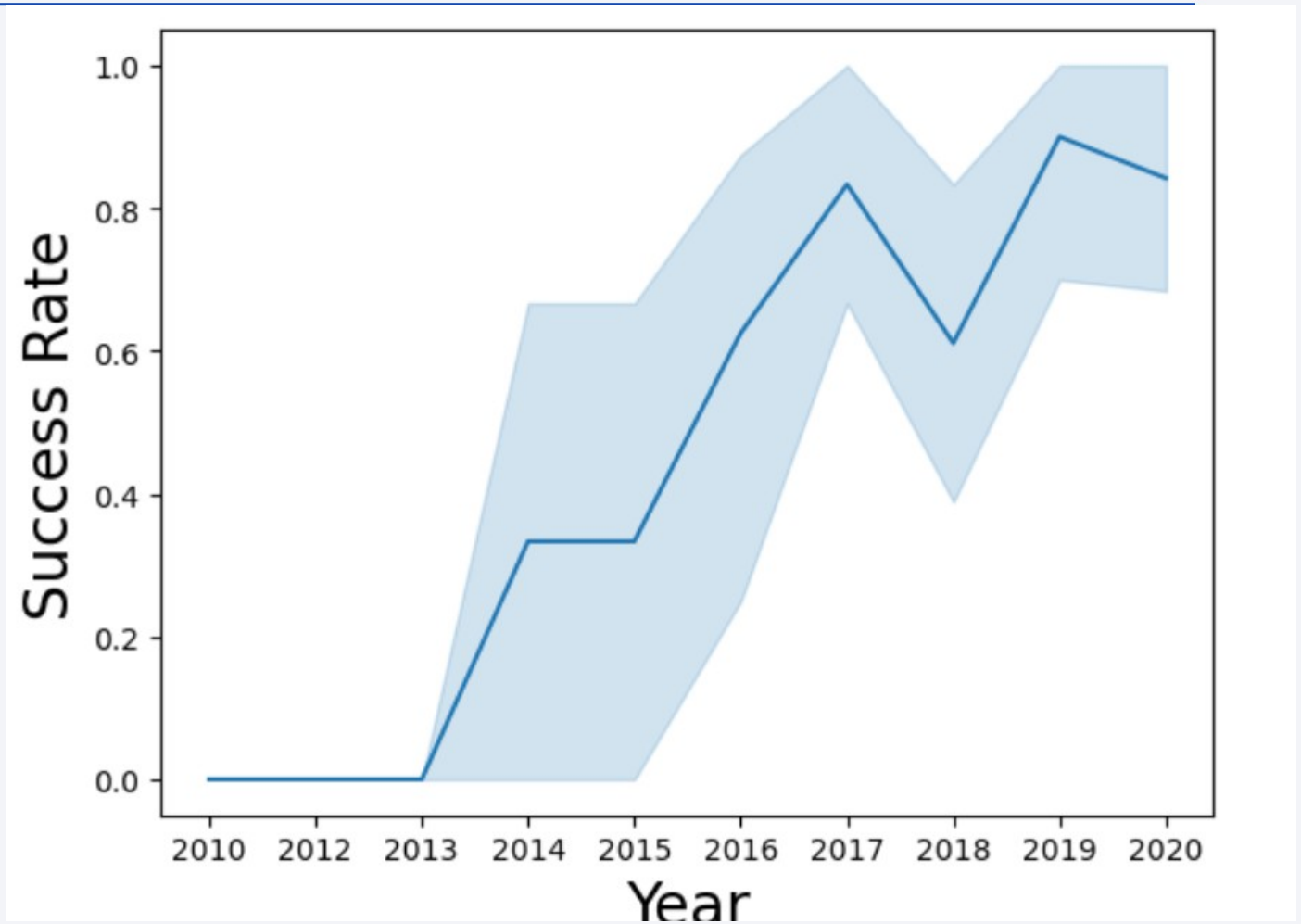
Payload vs. Orbit Type



- Payloads larger than 13,00kg were all VLEO
- Most launches less than 3,500kg were ISS
- Launches between 3,500kg and 8,000kg were GTO

Launch Success Yearly Trend

- Successes increased as time progressed
- The best launch year was 2019 with ~90% success rate
- The graph deviates from the trendline the most between 2013 and 2016 implying the results were much more volatile



All Launch Site Names

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

The four* launch sites used were:

Cape Canaveral Launch Complex 40

Vandenberg Space Launch Complex 4

Kennedy Space Center Launch Complex 39A

Cape Canaveral Space Launch Complex 40

*CCAFS LC-40 was later renamed CCAFS SLC-40 and results from the two should be grouped together

Launch Site Names Begin with 'CCA'

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

The total payload mass carried by boosters launched by NASA (CRS) was 45,596kg

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1 was 2,928.4kg

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The first successful groundpad landing was on December 22, 2015

First Successful GroundPad Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The 4 boosters which have been successful with drone ships and a payload mass between 4,000kg and 6,000 kg are listed below

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

Nearly all mission outcomes (as opposed to landing outcomes) were a success as seen below

Mission_Outcome	TOTAL
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Booster versions that carried the Maximum payload listed

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

2015 Launch Records

There were 2 failed drone ship landings in 2015 (January and April launches) with booster version and launch site info below

MONTH	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

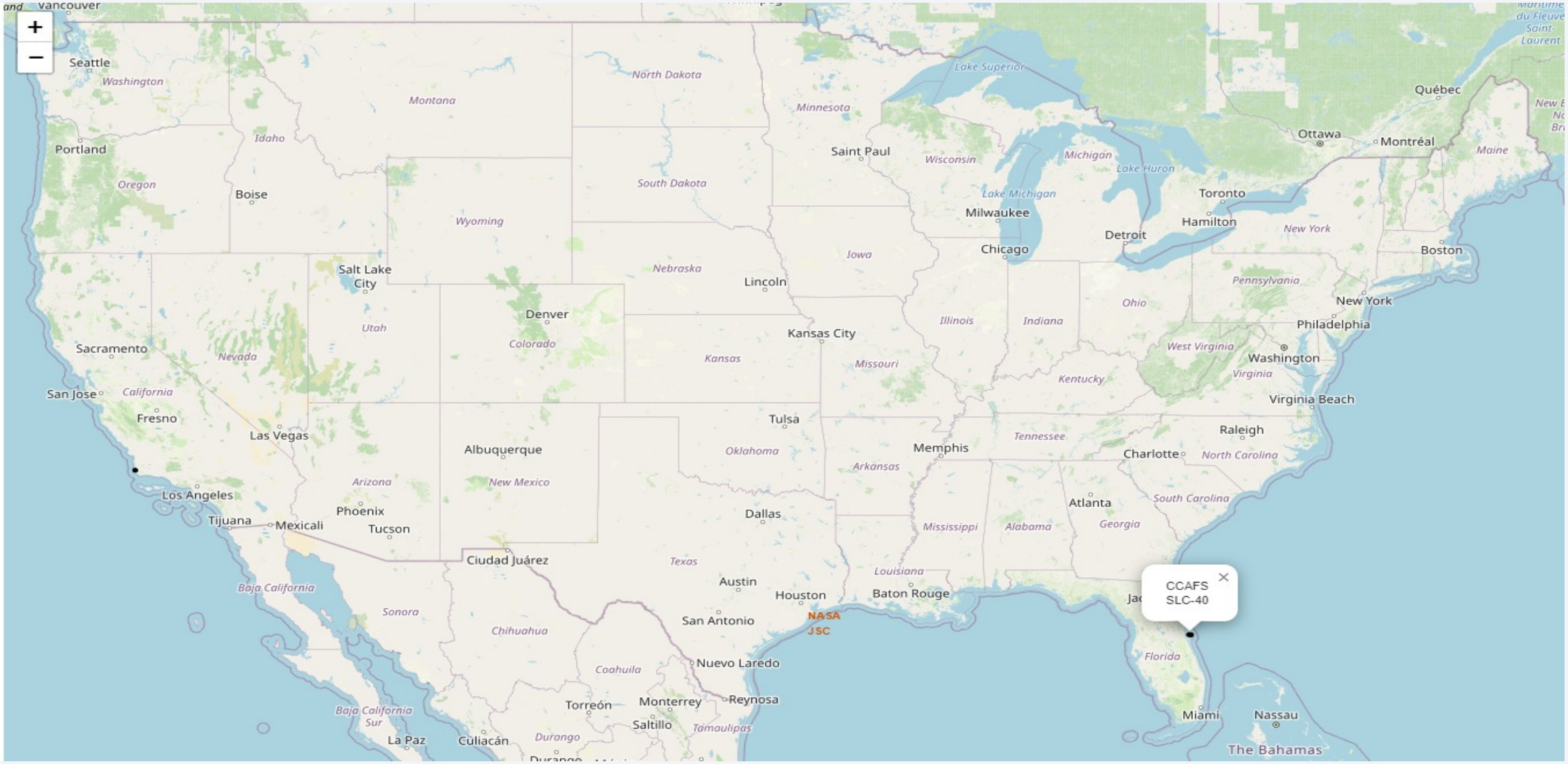
Landing outcomes between 2010-06-04 and 2017-03-20 listed below

Landing_Outcome	OCCURANCES
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

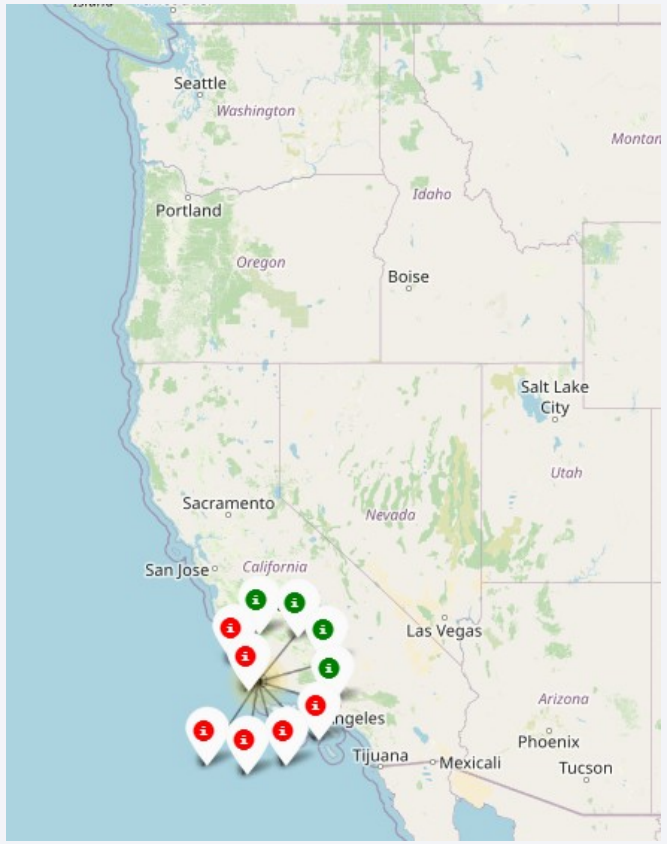
Section 3

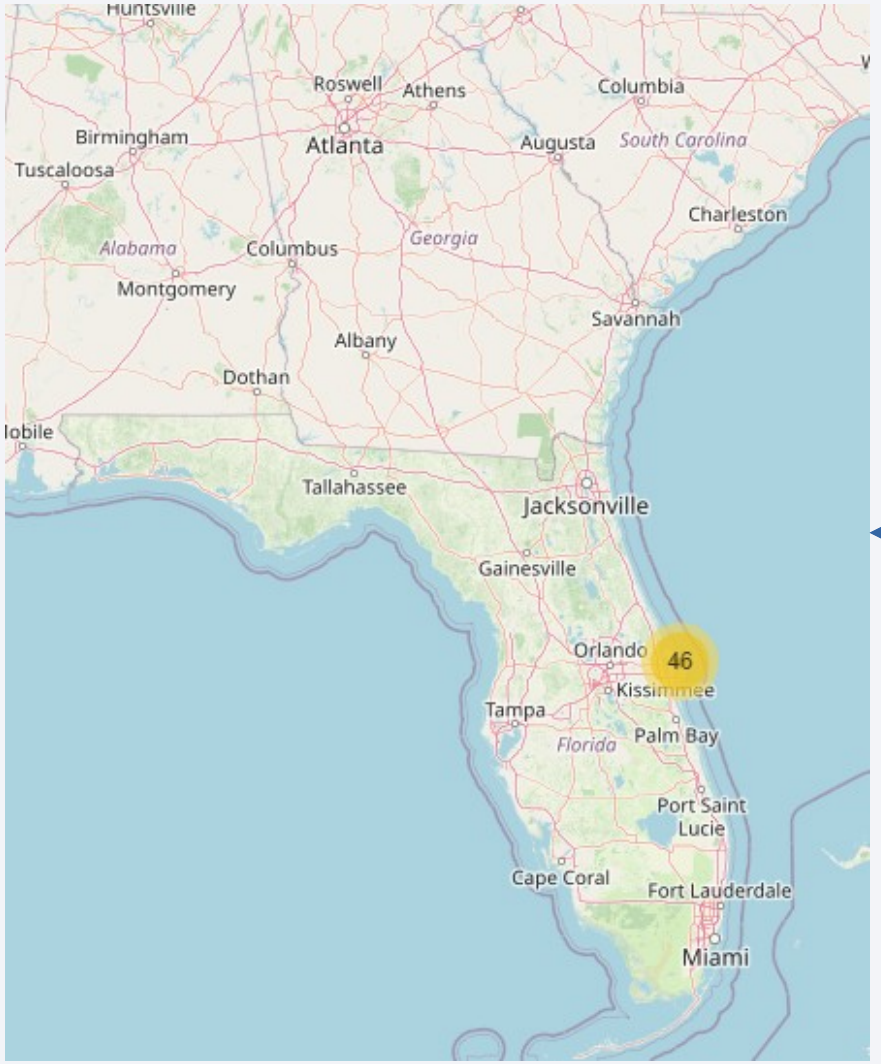
Launch Sites Proximities Analysis

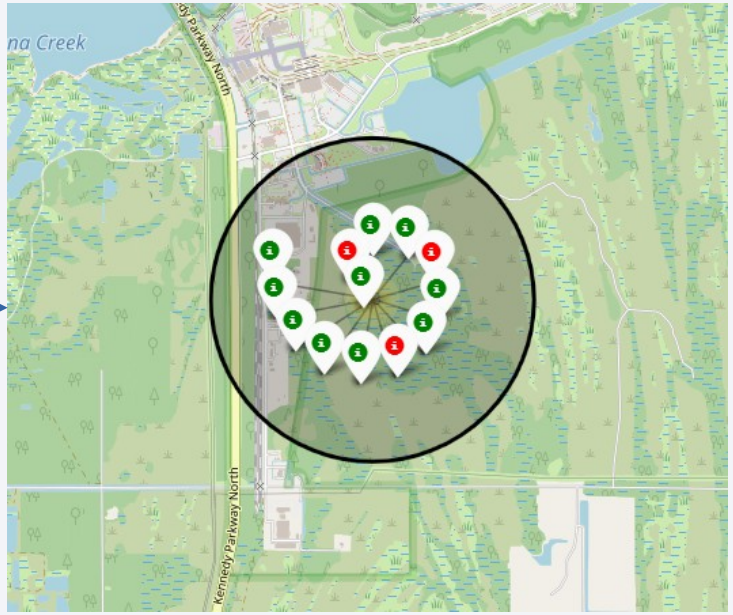
Launch Site Map

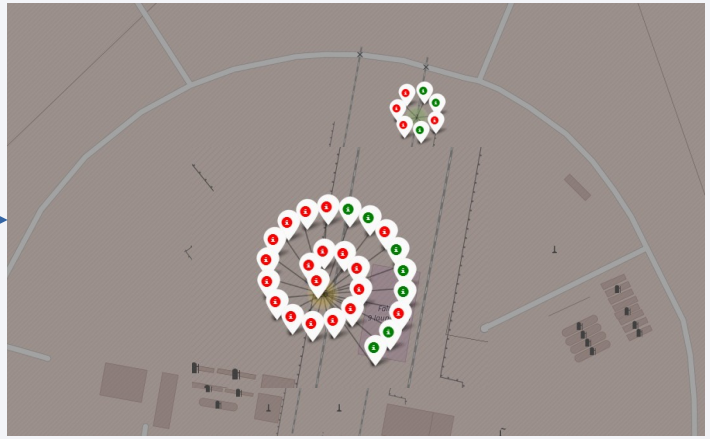


Launch Site Successes/Failures

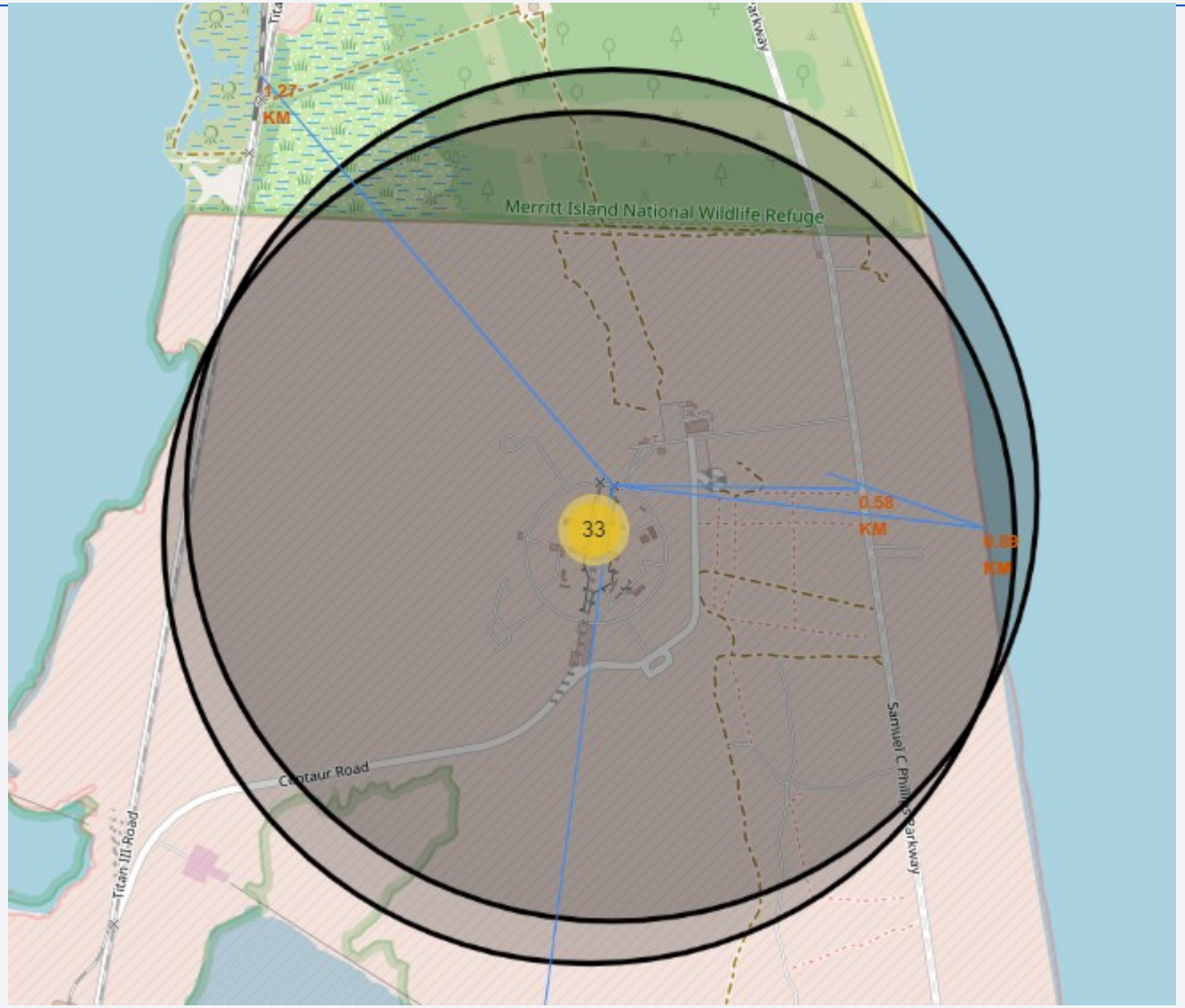




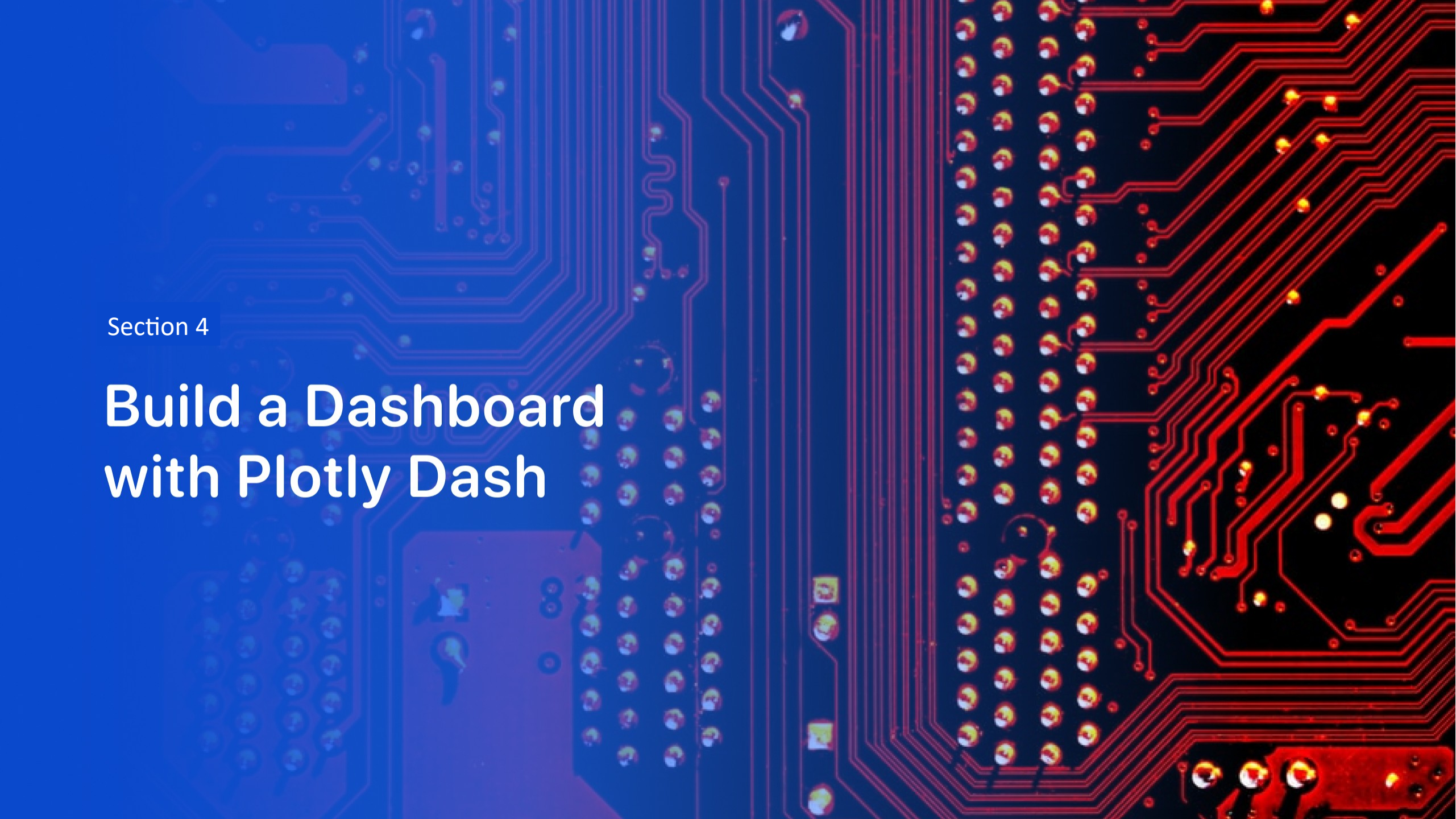




Points of interest Map

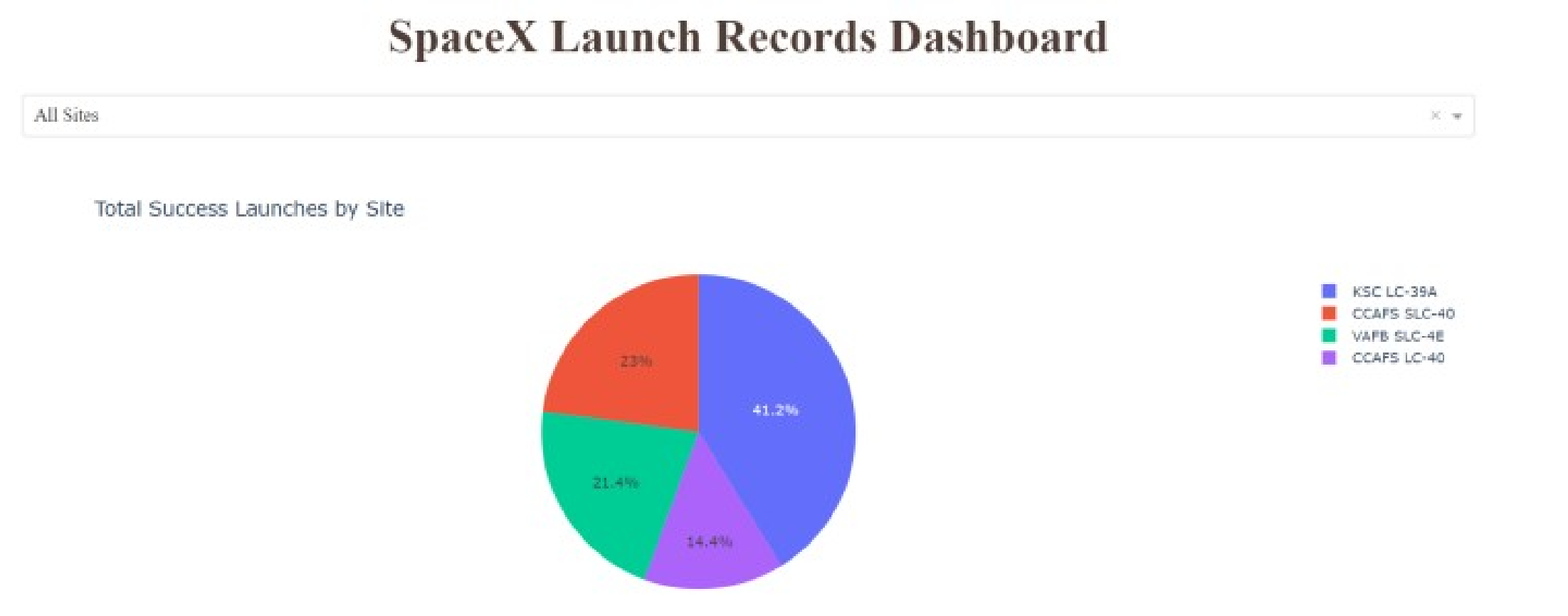


Launch site is 0.89km from the nearest coastline, 0.58km from nearest highway, 1.27km from the nearest railroad and 51.46km from the nearest airport


Section 4
Build a Dashboard with Plotly Dash

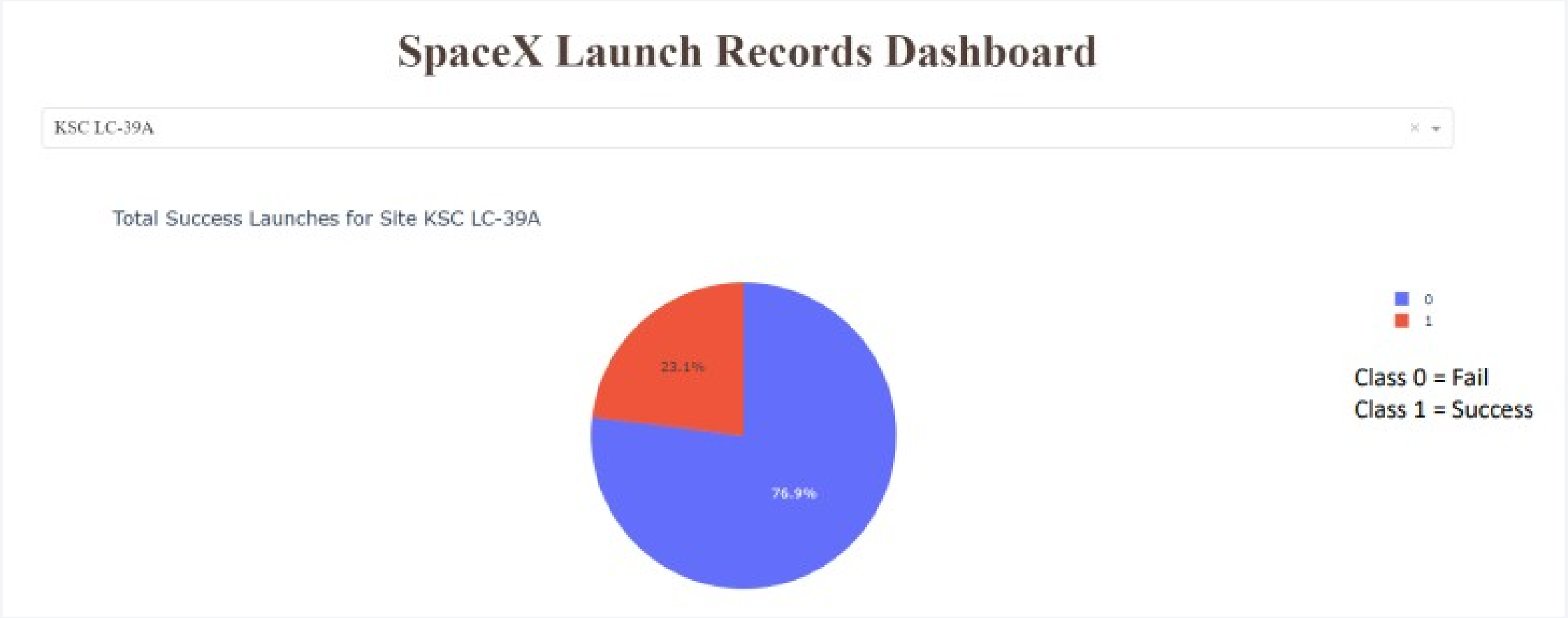
Total Successful Launches for all sites

Kennedy Launch Center had the most successful record of all launch sites



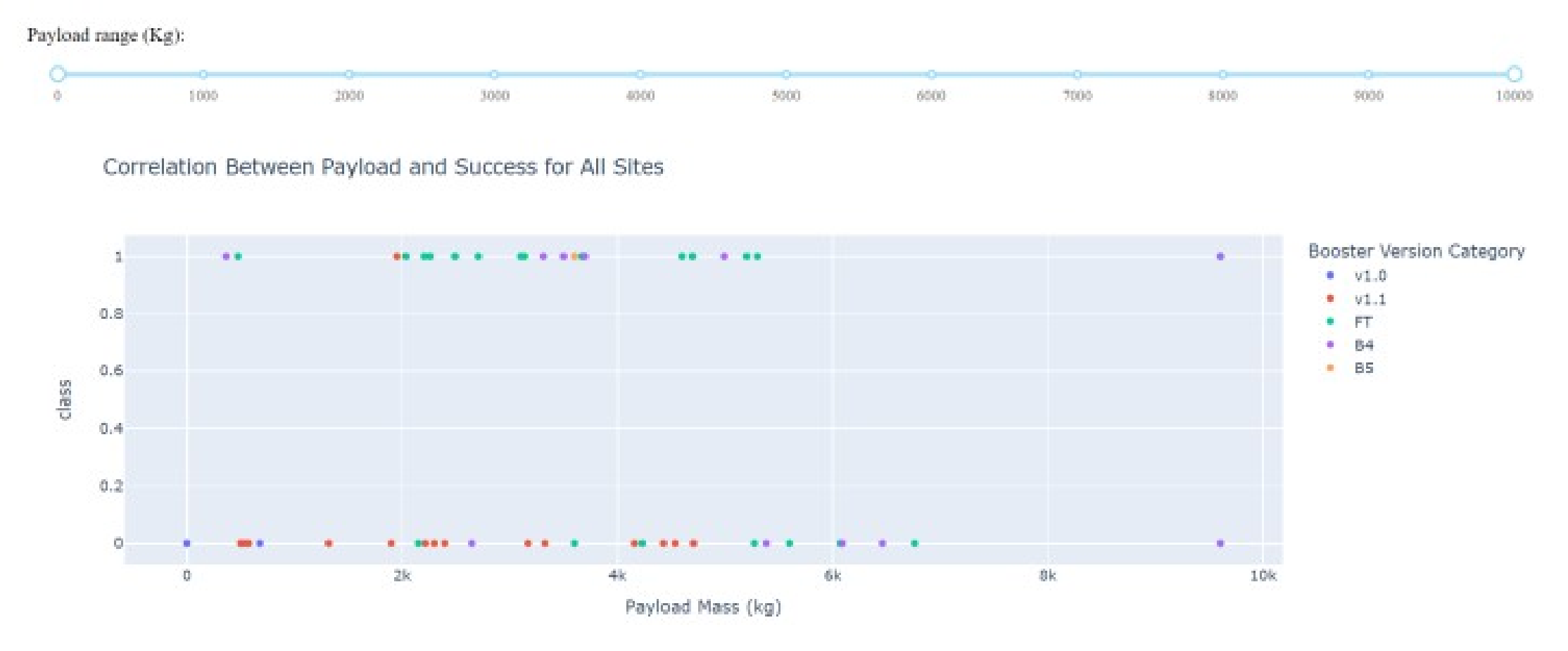
Launch success/failure for Kennedy Launch Center

Kennedy had a total of 13 launches with 10 successes and 3 failures



<Dashboard Screenshot 3>

Booster version FT was most successful while v1.0 and v1.1 were least successful

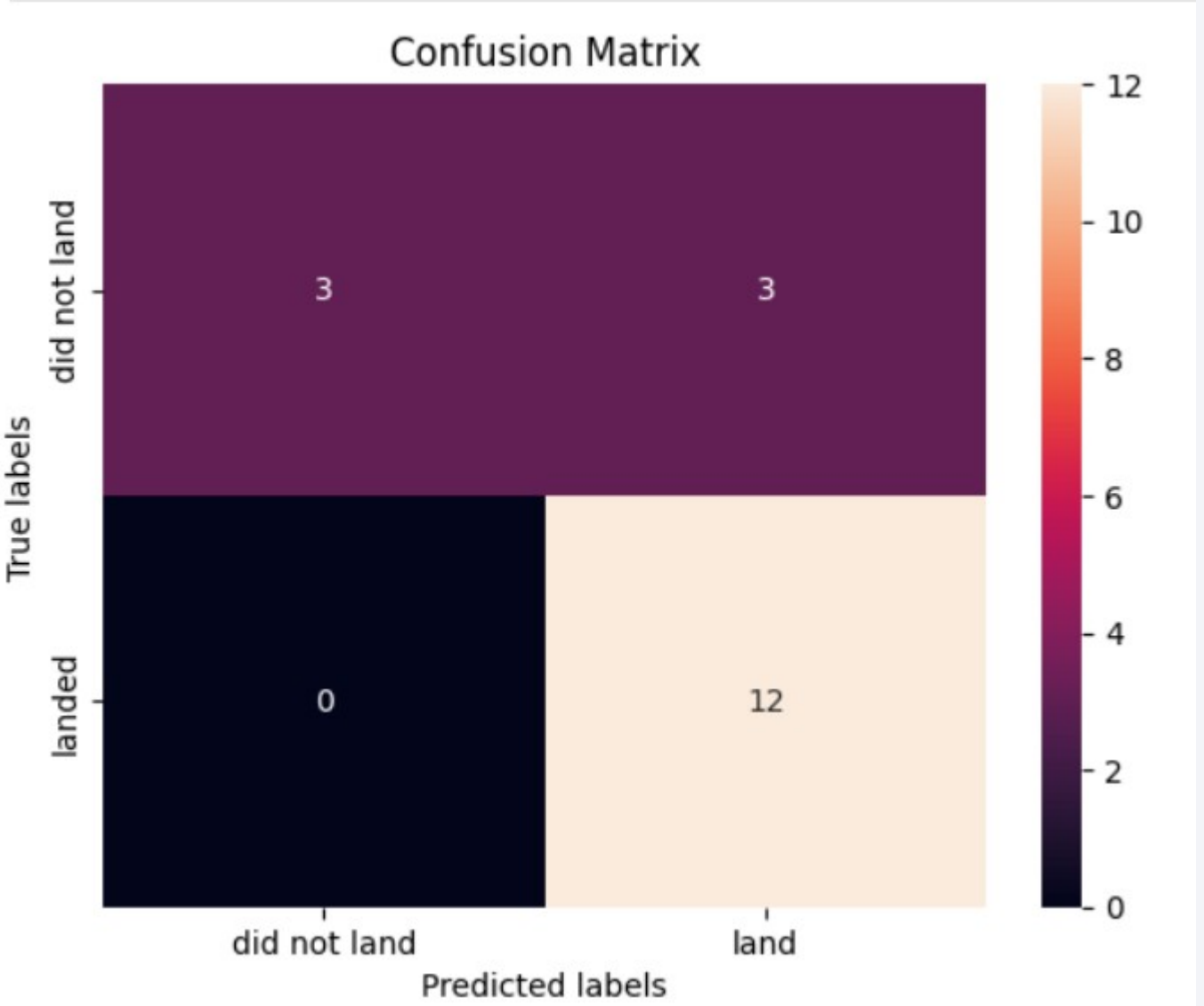


Section 5

Predictive Analysis (Classification)

Confusion Matrix

- The confusion Matrix (right) was generated by 3 of the 4 predictive models.
- Only the decision tree model performed slightly worse by incorrectly predicting 4 as landed when they did not land



Conclusions

- 100% success rate (and 0% success rate) can be misleading if taken at face value without considering the number of attempts
- Name changes during data keeping can skew results as they will not be correctly recorded as seen with CCAFS LC and CCAFS SLC being the same location, just a name change
- Most Launch sites are very near the equator and an ocean. This is likely due to a slingshot effect that the earth provides for launches and a safe failure zone in the event that a ship malfunctions

Thank you!