IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Johnny Diaz
Predictive Aerospace Analytics
Engineer- Space Y
August 18, 2024

Outline

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

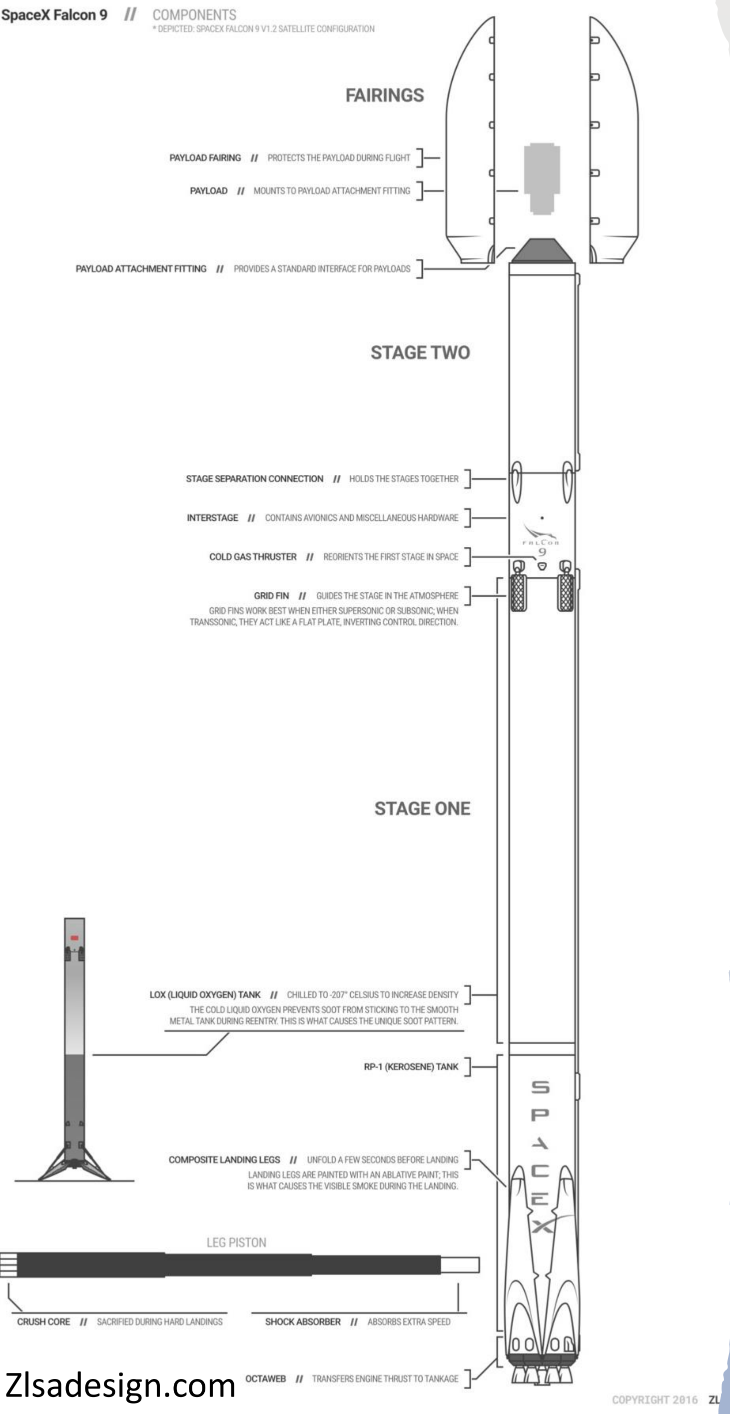
Executive Summary

This report analyzes the commercial space industry's growing competitiveness, focusing on SpaceX's Falcon 9 rocket. The primary objective is to identify factors that influence the successful landing of Falcon 9's first stage and to leverage this information for competitive bidding against SpaceX.

Data was collected from SpaceX's REST API, including web scraping techniques on other publicly available dataframes to gather historical launch data. The data was processed through various steps, including data wrangling and exploratory data analysis (EDA) using SQL and visualization tools like Folium and Plotly Dash. Predictive analysis was conducted using machine learning classification models, where a GridSearch revealed the Decision Tree Classifier emerged to be the most effective model for predicting the success of the first stage landing, thus providing Space Y a competitive advantage in the race to space.

Key findings from the analysis reveal that:

- The success rate of Falcon 9 launches increases with the number of flights.
- Certain launch sites, orbit types, and payload mass figures significantly affect the success of first-stage landings.
- The Decision Tree Classifier achieved the highest accuracy in predicting landing outcomes, providing a reliable model for forecasting launch costs.
- This analysis provides critical insights for a new rocket company, Space Y, enabling it to estimate launch costs more accurately and compete effectively with SpaceX by predicting the likelihood of first-stage recovery.



Zlsadesign.com

Introduction

The commercial space industry is becoming increasingly competitive as continuing resource-saving innovations are on the rise.

- Rocket Launches cost up to $165 M per launch
- Space X Falcon 9 charges $62 M
    - Cost Saving: Space X reuses 1st Stage of Falcon 9

OBJECTIVE:

- Determine factors for successful landing of Falcon 9 – Stage one
- Use cost information to bid against Space X for comparable rocket launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was sourced from historical information of Falcon 9 rocket launches.

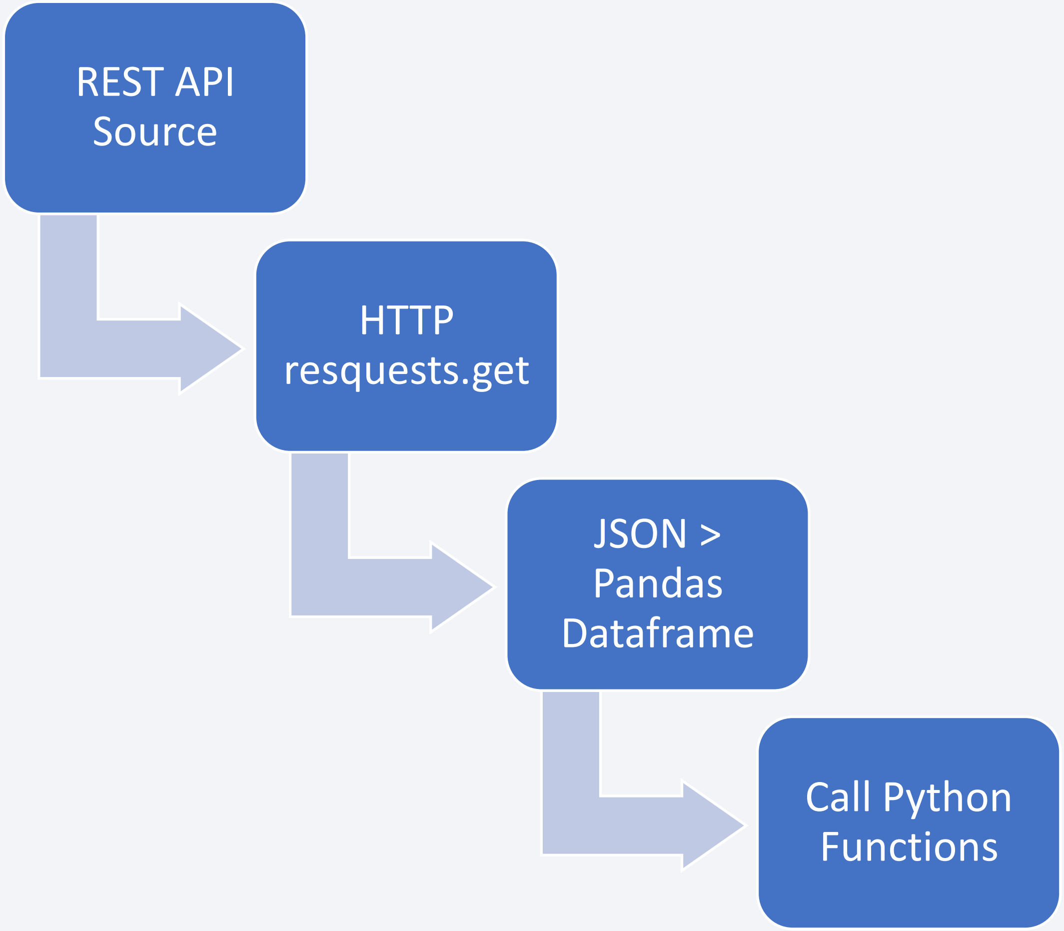
Space X Rest API: *https://github.com/r-spacex/SpaceX-API*

<u>Overview of Methods</u>

- HTTP requests.get calls
- Functions & parsing of JSON
- Webscraping via BeautifulSoup
- Parsing HTML Tables

Data Collection – SpaceX API

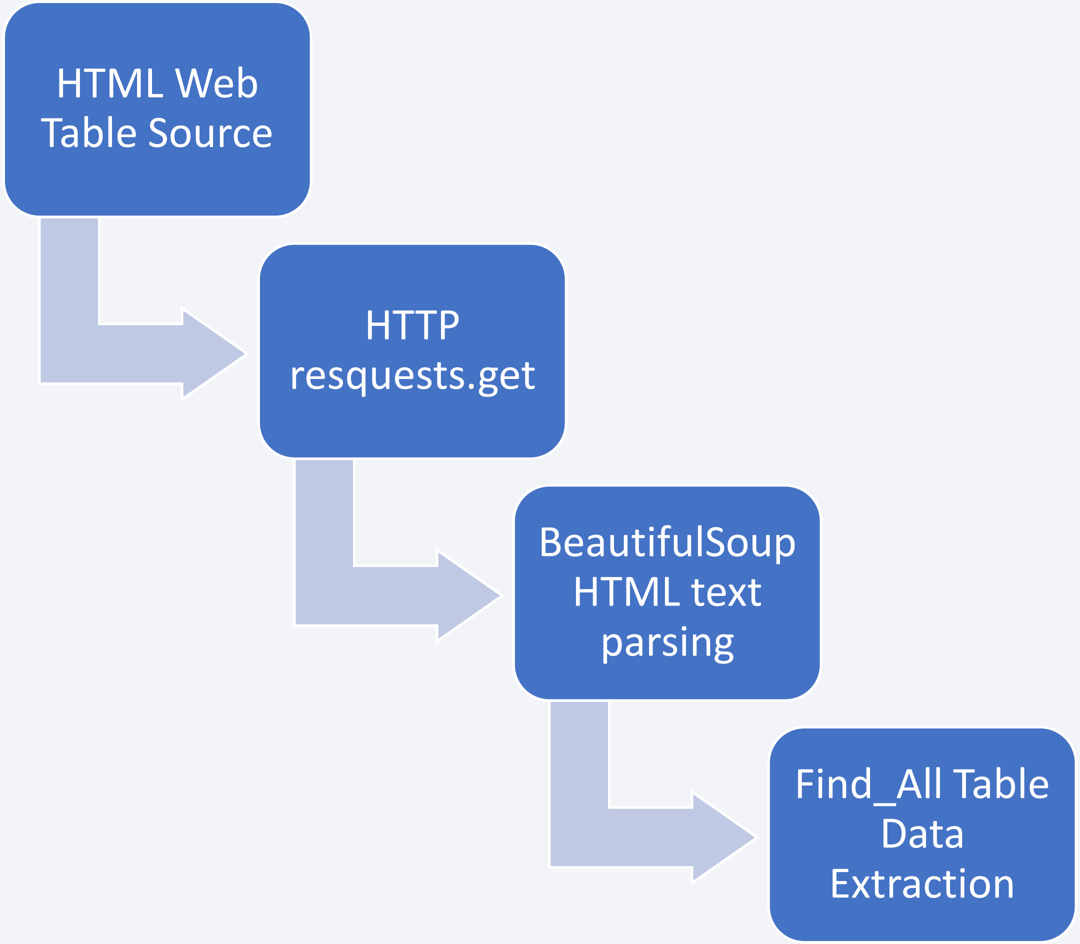
1. Request & Parse Space X data using GET requests
2. Decode HTTP response content as JSON
3. Convert JSON to data frame
4. Call functions to return Launch Site, Payload, and Core data from API



[GitHub URL]

Data Collection - Scraping

1. HTTP GET requests from HTML table
2. Response in BeautifulSoup HTML parser
3. Extract column names using find_all('table')
4. Create dataframes via parsing several HTML tables and running python functions

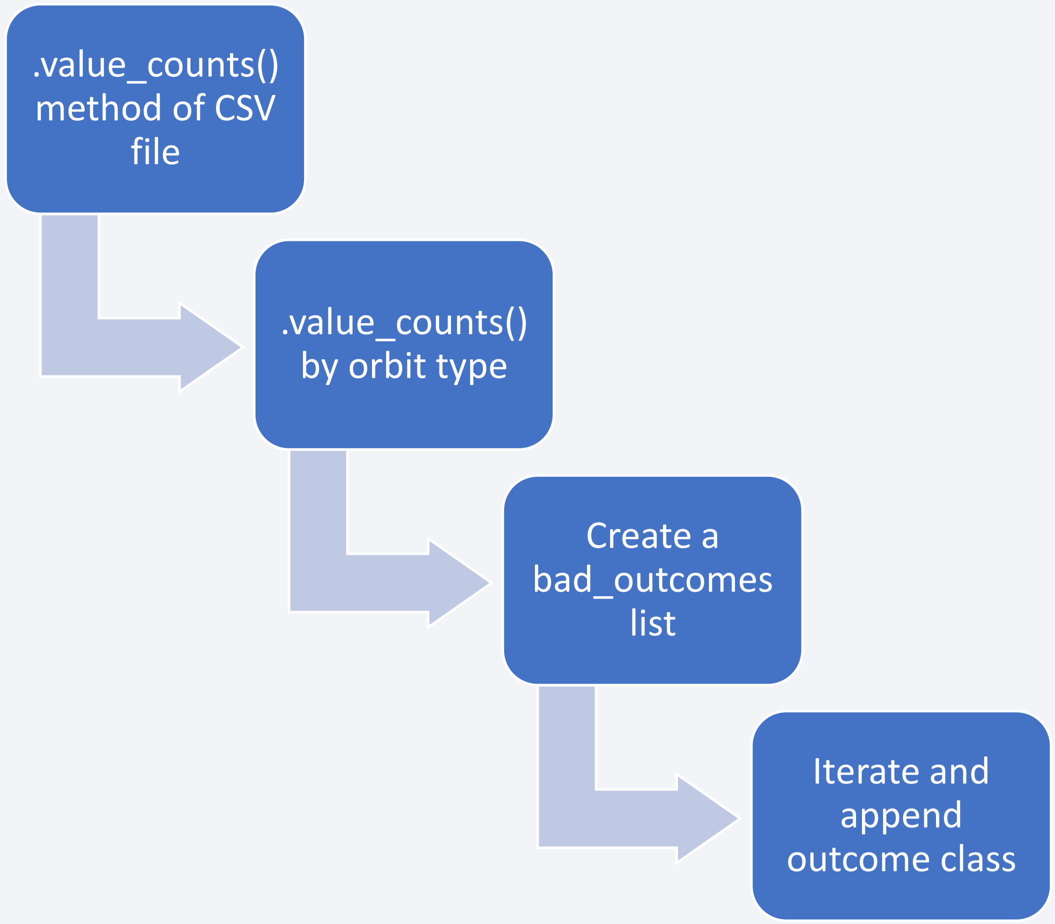


GitHub URL

Data Wrangling

1. Calculate number of launches by Site
2. Calculate number of launches by Orbit type
3. Calculate mission outcome counts
4. Filter landing outcomes into a class:

    0 = failure, 1 = success



GitHub URL

EDA with Data Visualization

Scatter Plot Relationships with Outcome Success Hue

- Flight Number & Payload Mass Kg
- Flight Number & Launch Site
- Payload Mass Kg & Launch Site
- Flight Number & Orbit Type
- Payload Mass Kg & Orbit Type

Bar Chart

- Orbit Type By Average Success %

Time Based Yearly Trend

- Year & Average Success %

GitHub URL

EDA with SQL

- Finding the 4 Distinct Launch Site Names
- Query of Total Payload Mass (Kg) by specific client name
- Average Payload Mass (Kg) by specific Falcon 9 Booster Version
- Date of 1st Successful Ground Pad Landing
- Booster names involved in Successful Drone ship Landings with Payload Mass between 4,000 & 6,000 Kg
- Distinct Booster Versions involved in the Max Payload Mass (15,600 Kg)

[GitHub URL]

Build an Interactive Map with Folium

- Marker, Circle, and Popup child objects were added to the main folium.Map() object to facilitate geographical identification of all 4 Launch Sites via coordinates.
- A MarkerClusters object was added to each site to visually differentiate successful from unsuccessful launch incidents
- A MousePosition object was implemented as a tool to collect coordinate information
- Lastly, individual Polyline objects with distance markers from Launch Site were added to document distances from nearby geographic attributes

GitHub URL

Build a Dashboard with Plotly Dash

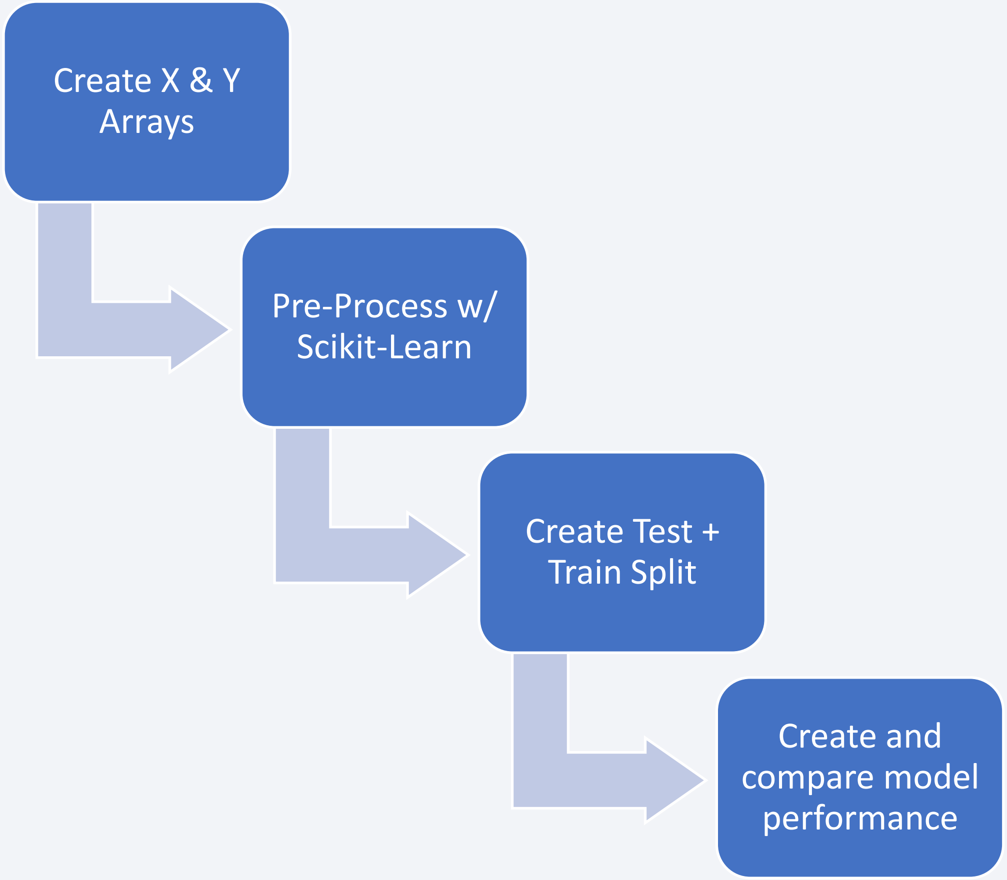
- Built Plotly Dashboard application for users to perform interactive visual analytics on SpaceX launch data in real-time
- Launch Site Drop-down Input menu to compare combined or individual site launch success rates via Pie Chart
- Interactive Payload Mass Range Slider to react to correlation graph
- Scatter plot between Payload Mass and Success Rate grouped by Booster Version for enhanced insight and interaction

GitHub URL

Predictive Analysis (Classification)

Determine Best Performing Predictive Method (Will First Stage Land?)

1. Convert Y Dependent Variable into Array
2. Standardize & fit (X) Independent Variable to normalize scale
3. Create a 20% Test vs. Train split to fit several models
4. Perform a series of GridSearchCV > Model fitting > hyperparameters > Score Tests > Confusion Matrices to compare predictive performance



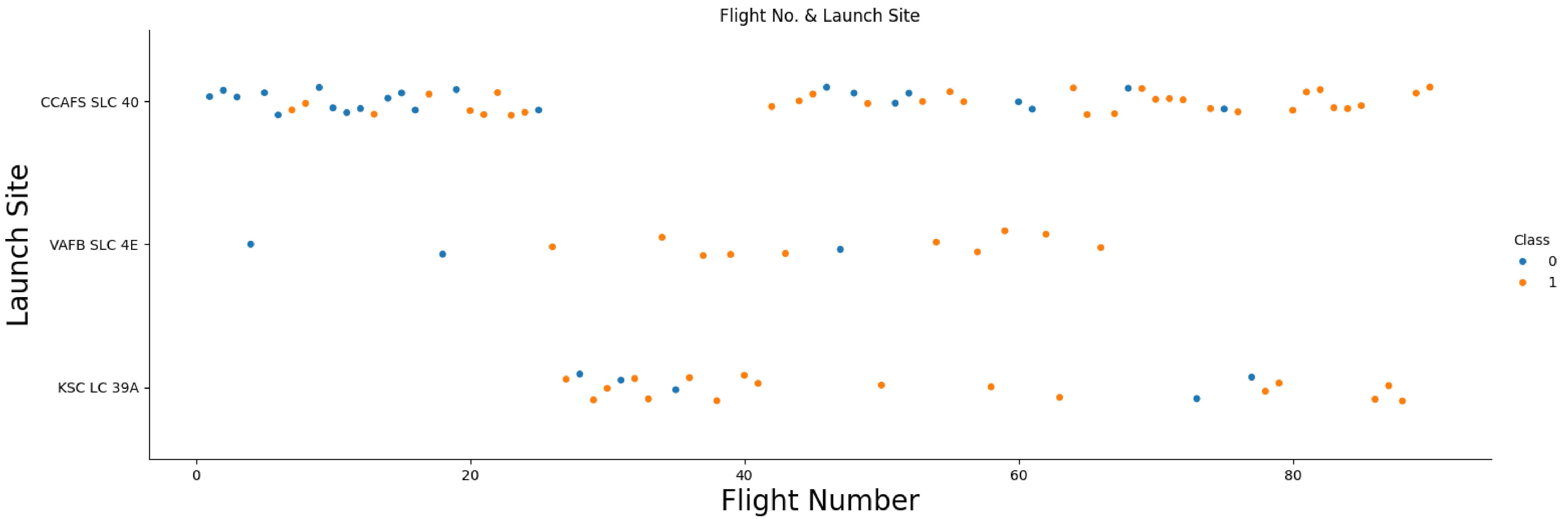
GitHub URL

Results

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

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

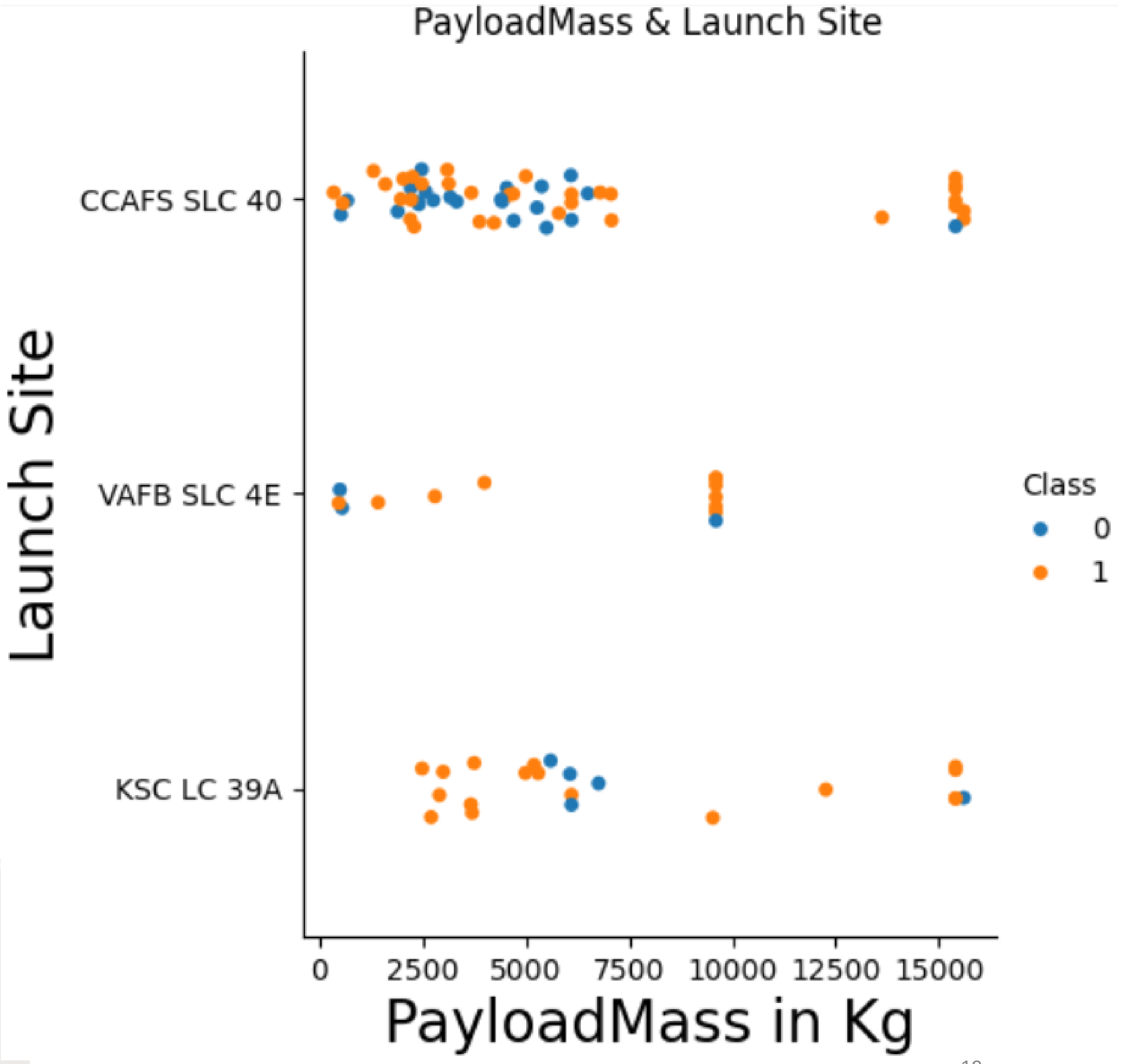
The 3 Launch Sites experience an increase in successful (orange) launches as flight number increased. Whereas with lower flight numbers, the results appear more mixed.

Payload vs. Launch Site

- The VAFB-SLC launch site does not have any rockets launched for heavy payload mass(greater than 10000).

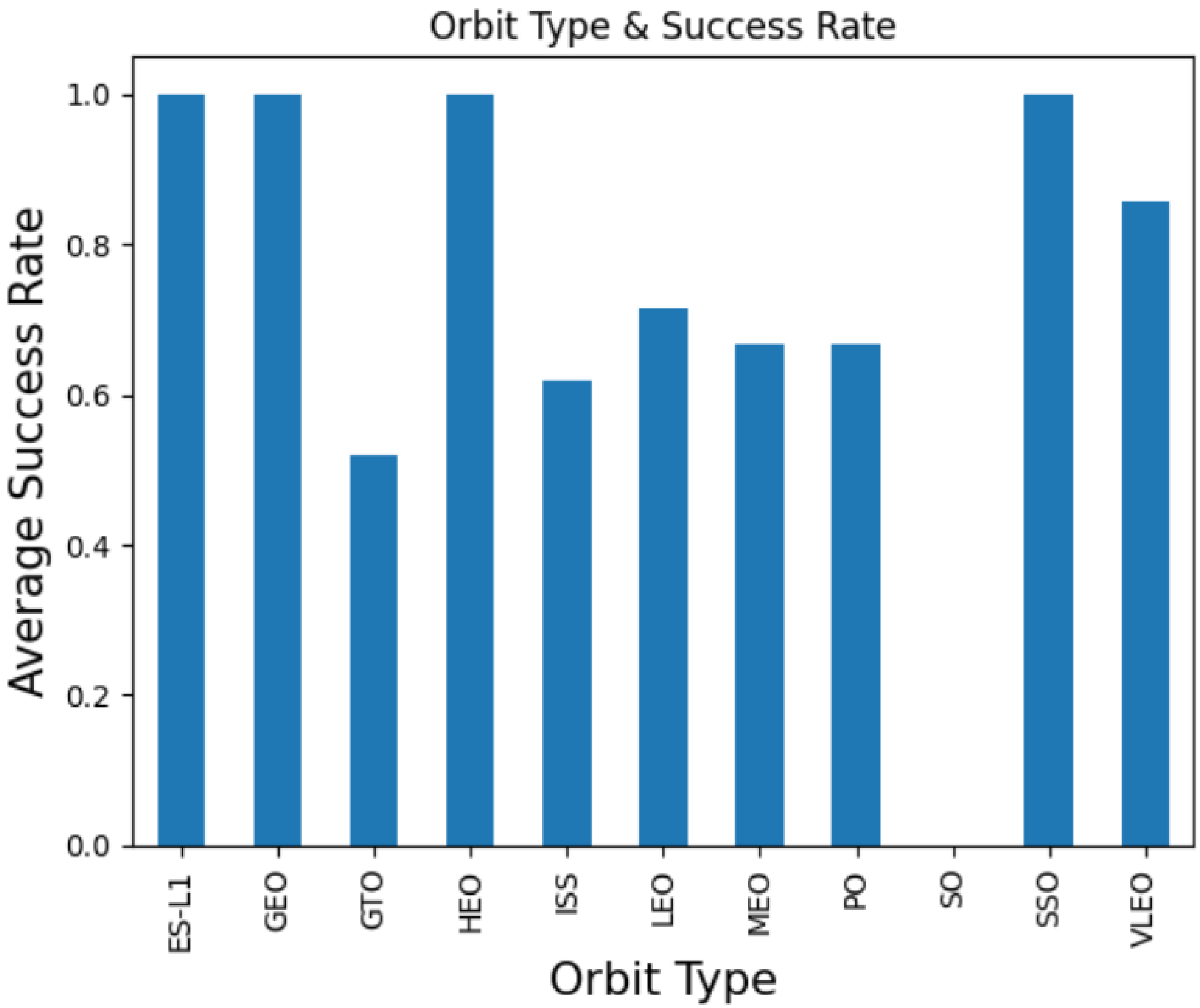
- CCAFS SLC 40 has mixed launch results for payload masses 7,500 kg and below. Yet, They have a high success rate for masses above 12,500 kg

- KSC LC 39A has a high success rate when payload masses are approximately 5,000 and below



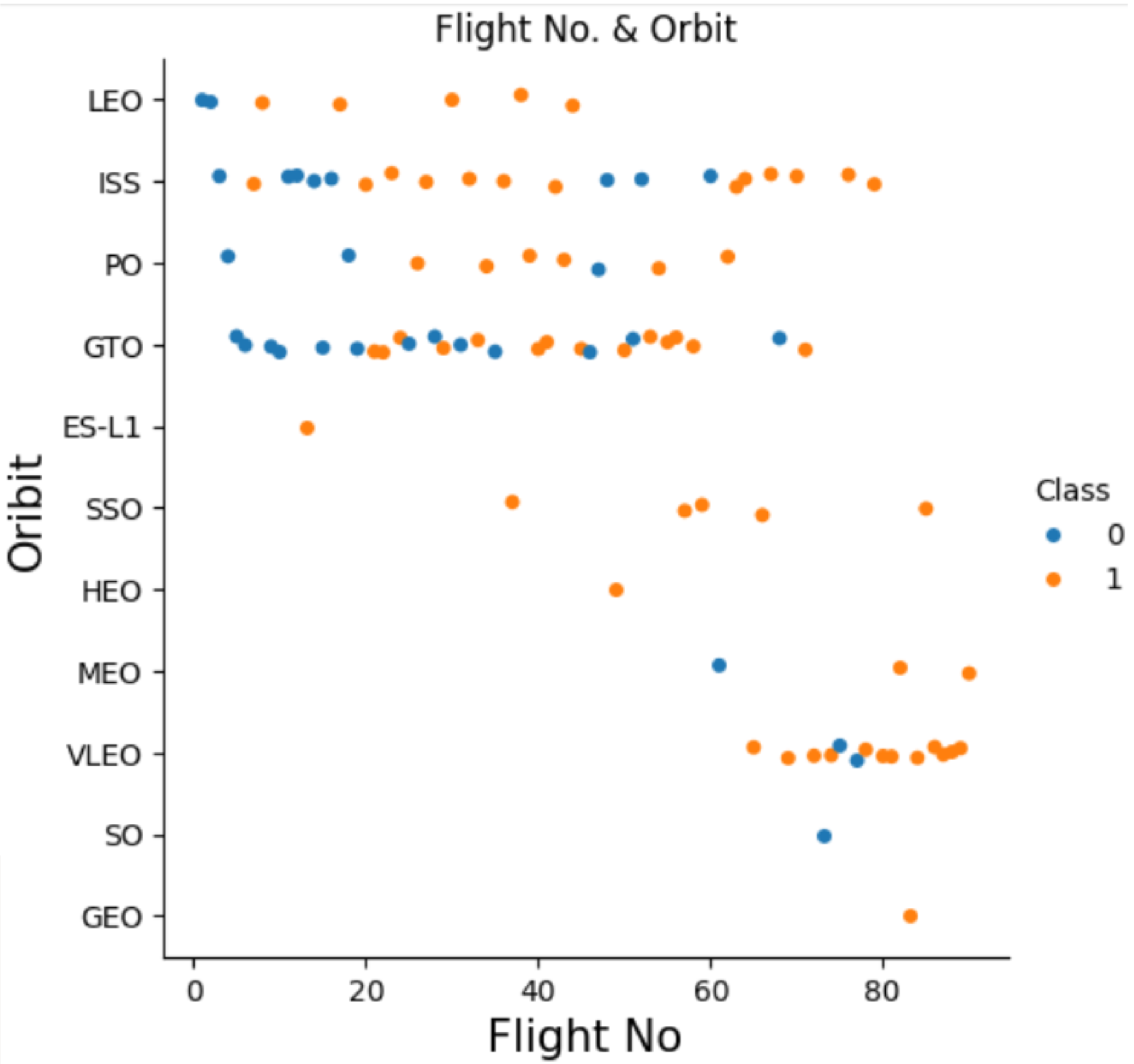
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO have the highest success rates.
- Orbit SO does not show any successful launches by average
- The remaining Orbit types appear to be approximately between 50%-70% successful



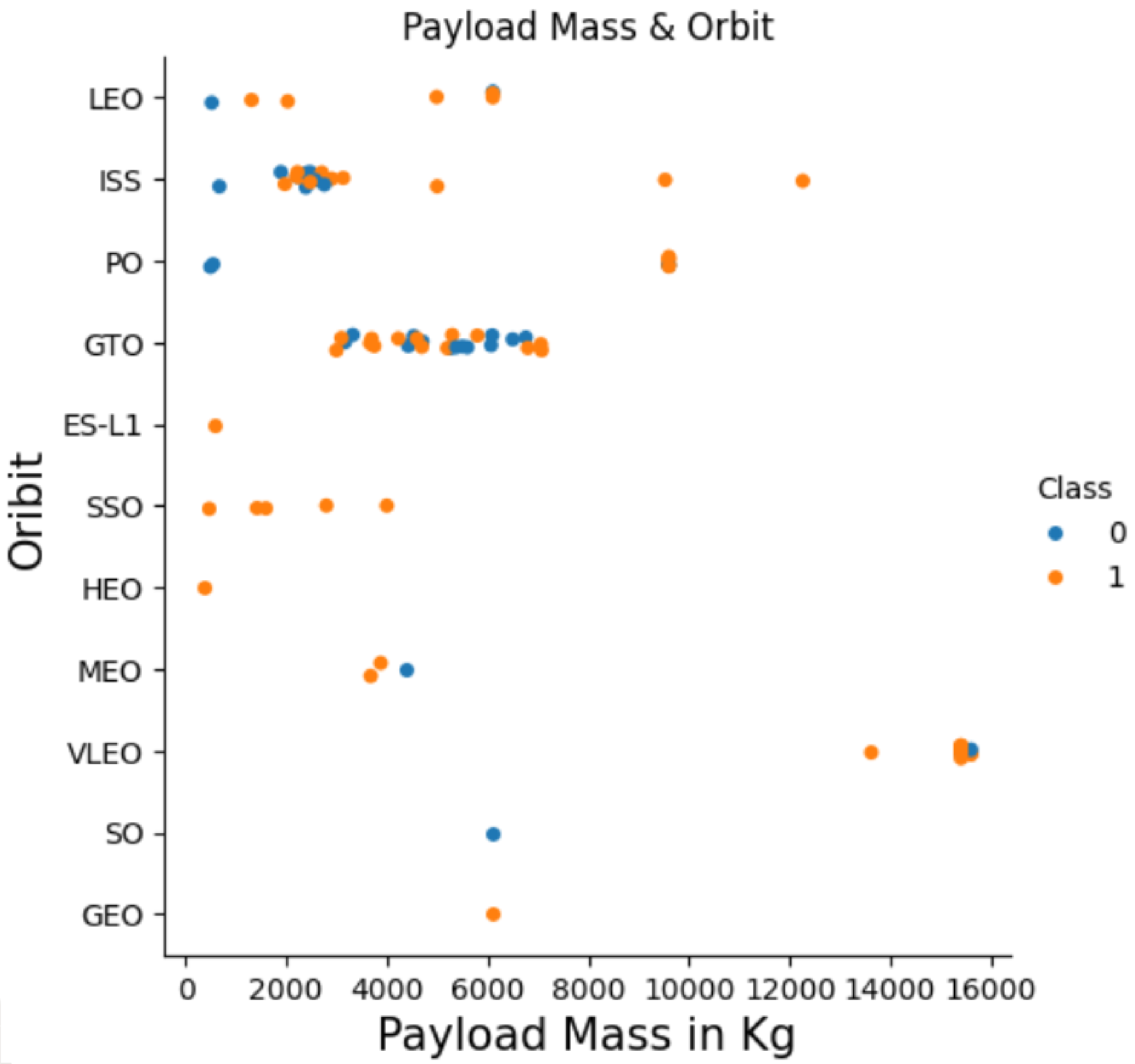
Flight Number vs. Orbit Type

- The results of the scatter plot compliment the bar chart results because the information is identical
- The scatter plot visualization illustrates the individual records as Flight Numbers increase
- GTO is illustrated with the most failure landing results



Payload vs. Orbit Type

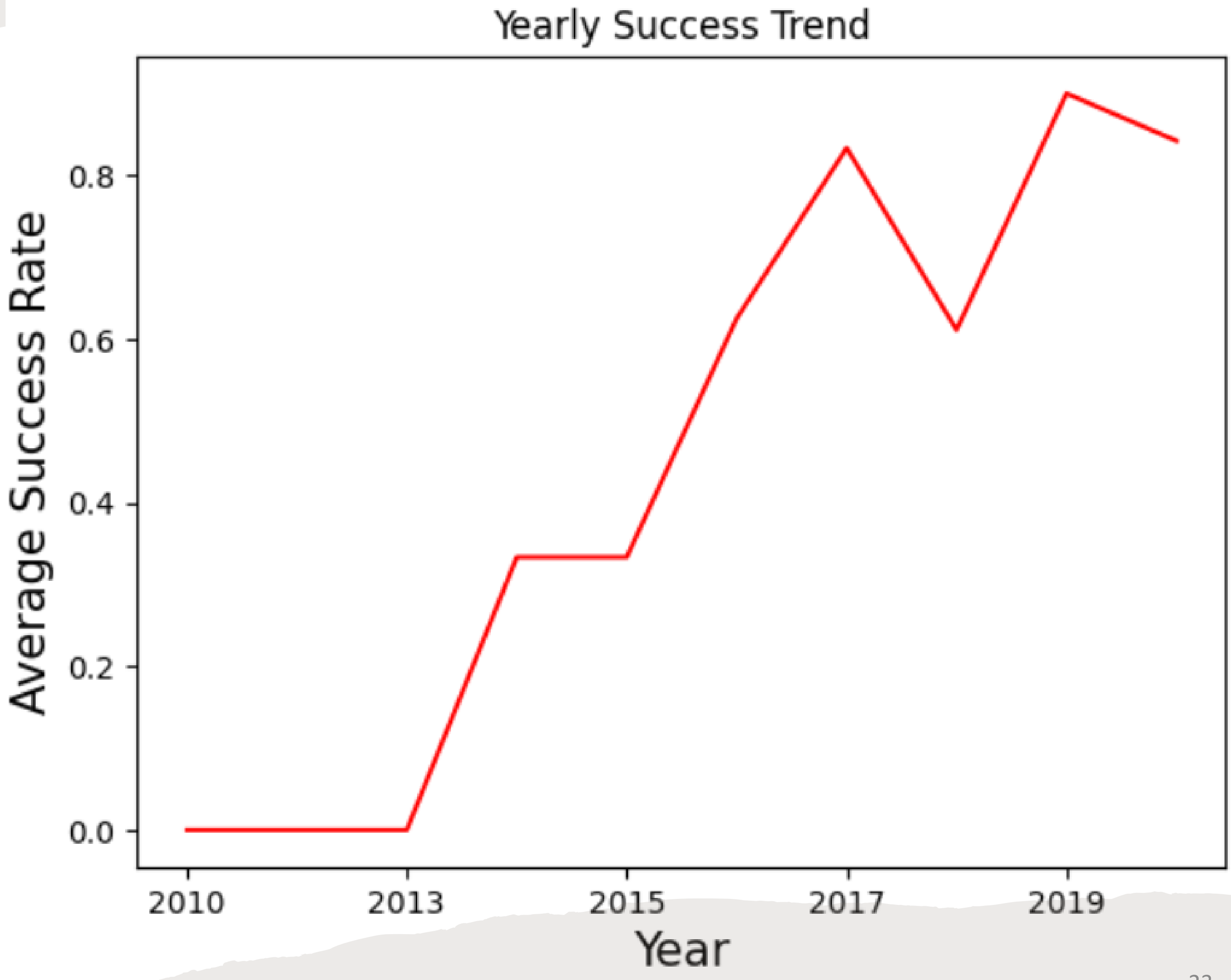
- Successful landings of heavy payloads are more frequent for Polar, VLEO, LEO and ISS Orbits
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present. This is consistent with the other graphs of GTO
- SSO shows high frequency of success with payloads under 6,000 kg



Launch Success Yearly Trend

- The average success rate since 2013 has been on an increasing trend

- There is a slight pull back in the short term after 2017 and before 2019, yet overall, the longer-term trend is positive



All Launch Site Names

Specifying “distinct” names of Launch_Site yielded an efficient table of 4 results seen to the right

```
%sql select distinct Launch_Site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

Results of “CCAFS LC-50” following a modulo operator in the WHERE clause with a limit of 5 results

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select Launch_Site from SPACEXTABLE where Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select SUM(PAYLOAD_MASS__KG_) AS "NASA CRS Total Payload kg" from SPACEXTABLE where Customer is 'NASA (CRS)
```

```
 * sqlite:///my_data1.db
Done.
```

NASA CRS Total Payload kg
45596

Total Payload Mass by Client NASA (CRS)

This query highlights a Total Payload Mass of **45,596 kg** launched by Space X on behalf of NASA CRS

Display average payload mass carried by booster version F9 v1.1

```
elect AVG(PAYLOAD_MASS__KG_) AS "Avg Booster F9 v1.1 Payload KG" from SPACEXTABLE where Booster_Version is 'F9 v
```

```
 * sqlite:///my_data1.db
Done.
```

Avg Booster F9 v1.1 Payload KG
2928.4

Average Payload Mass by F9 v1.1

This query highlights an Average Payload Mass of **2,928.4 kg** launched where Booster Version F9 v 1.1 was involved

First Successful Ground Landing Date

Using the MIN aggregate for Date, a singular result is returned for the first Ground Pad Landing Outcome that was deemed Successful

```
%sql select MIN(Date) from SPACEXTABLE where Landing_Outcome is "Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)
2015-12-22

```
n from SPACEXTABLE where Landing_Outcome is "Success (drone ship)" AND PAYLOAD_MASS__KG_ Between 4000 and 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Successful Drone Ship Landing with Payload between 4000 and 6000

A query for Booster Version names with a robust WHERE CLAUSE that contains a range between number of kilograms & filters for successful drone ship landing

```
%sql select Mission_Outcome, COUNT(Mission_Outcome) As "Totals" from SPACEXTABLE GROUP BY Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

Total frequency of Mission Outcomes

Boosters Carried Maximum Payload

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

QUERY:

%sql select distinct Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (Select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE) Order By Booster_Version;

The Query returns the names of the distinct Booster Versions involved. Utilizing a sub query, we are able to return a table of those Booster Versions with the max Payload Mass of 15,600 Kg.

```
substr(Date, 6, 2) as Month, substr(Date, 1, 4) as Year, Landing_Outcome, Booster_Version, Launch_Site from SPAC
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

2015 Launch Records

Substr utilized to filter Date attribute and limit the year 2015 WHERE incidents of Failed drone ship landings were recorded

```
te,6,2) || substr(Date,9,2) between '20100604' and '20170320' GROUP BY "Landing_Outcome" ORDER BY "COUNT" DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

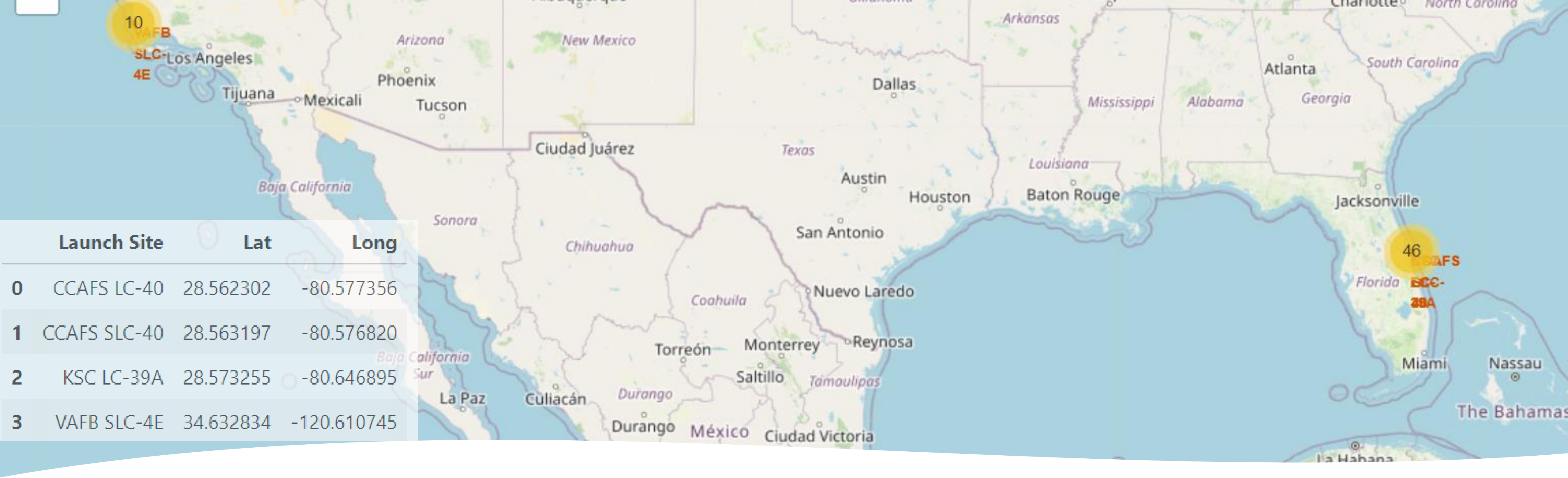
Landing_Outcome	COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

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

Substr utilized to filter Date attribute and group by Landing Outcome and Orby by a Count aggregate

Section 3

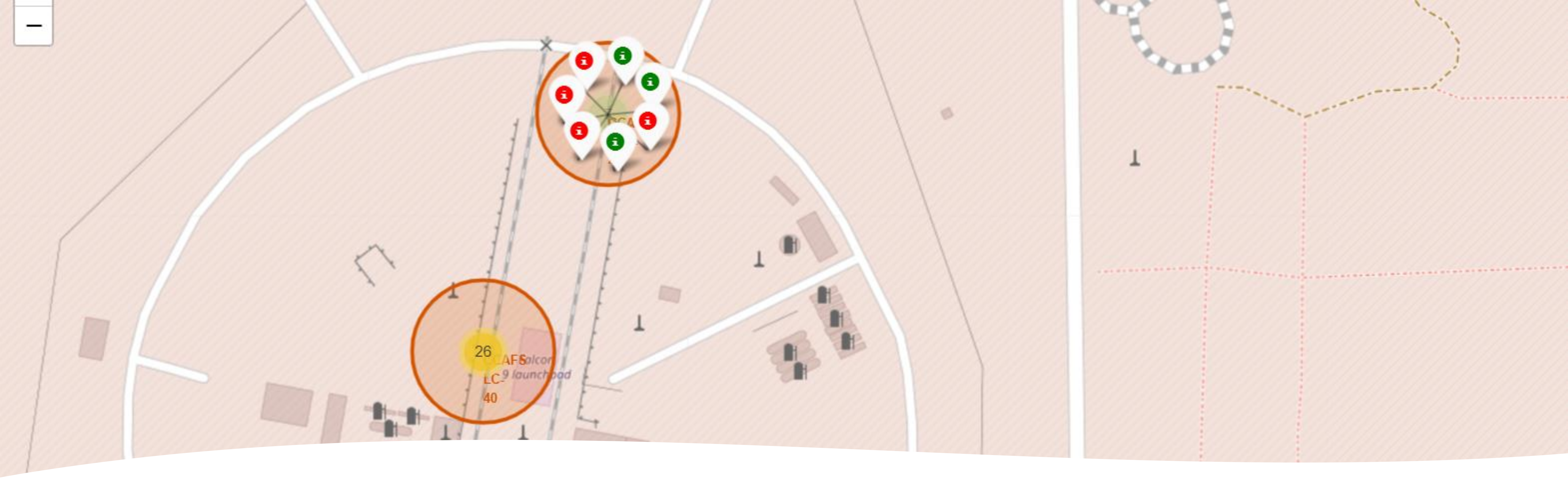
Launch Sites Proximities Analysis



	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

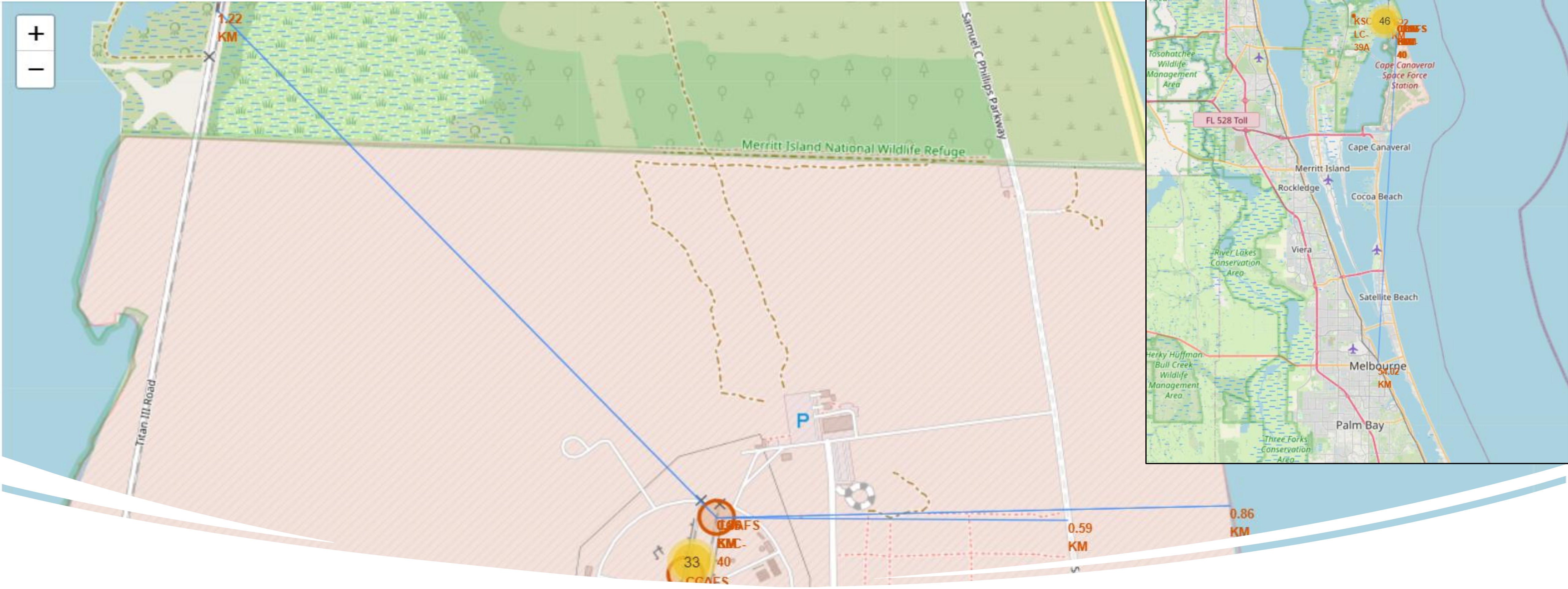
Site Marker & Circle Objects

- Generated global folium map which includes all 4 Launch Sites according to their coordinates. 1 in California & 3 in Florida
- Location circles, and text Markers were added for all 4 Launch Sites which are important in generating geographical data



Outcome Color-Coded Marker Cluster

- MarkerCluster child is implemented to color coded labels based on mission outcomes.
- Clicking on a circle object reveals the clusters and the appended colors based on the outcome class.



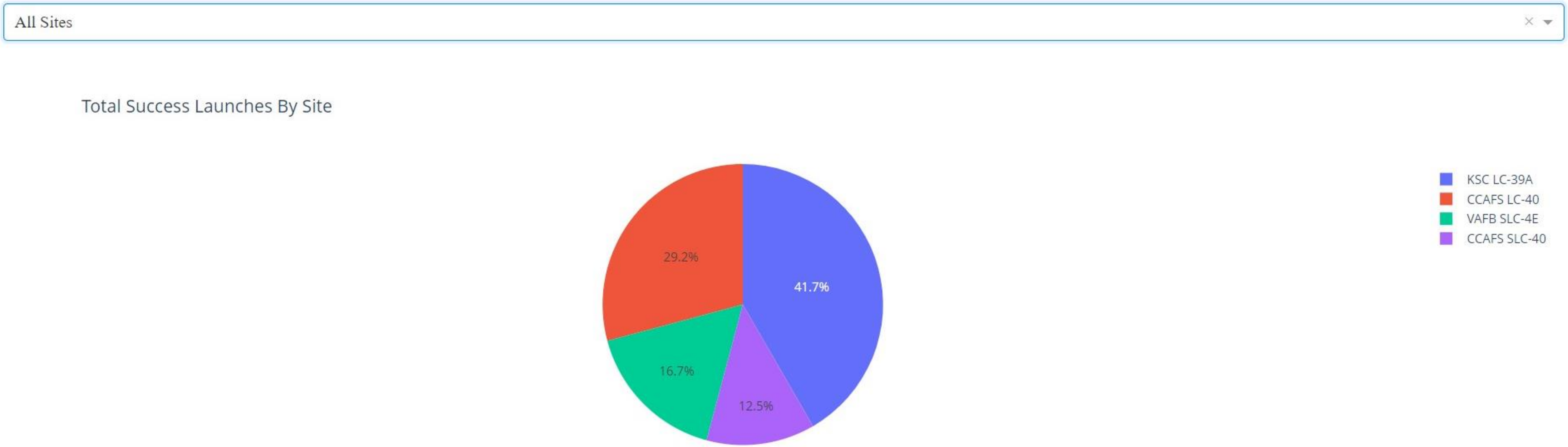
Nearest Geographical Proximities

- Polyline objects highlight distance markers (KM) of the nearest geographical proximities of interest
- Train tracks: 1.22KM, Coastline: 0.86M, Highway: 0.59M
- Melbourne city map: 52.02 KM

Section 4

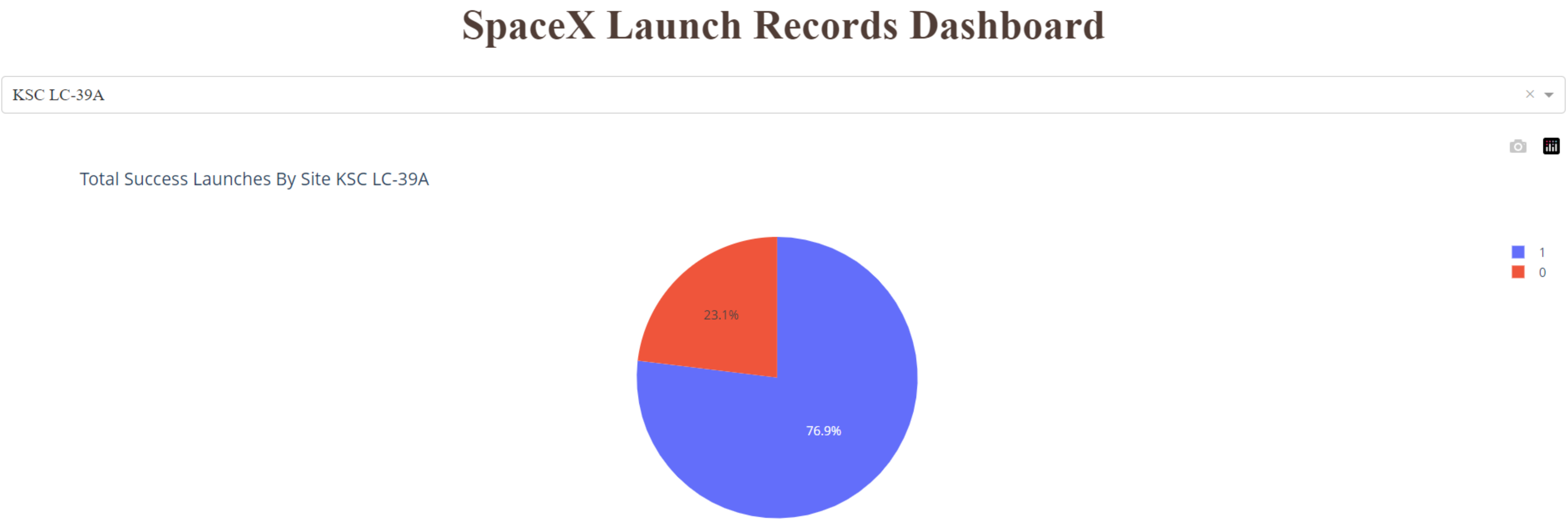
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



Drop-Down Pie Chart Dash

- Pie Distribution of All Launch Sites Based on Success Rate
- The Drop-Down Menu Enables End user to view success distribution of all sites or 1 site in isolation.
- The Legend on the right signals the colors of the individual sites



Highest Success Ratio

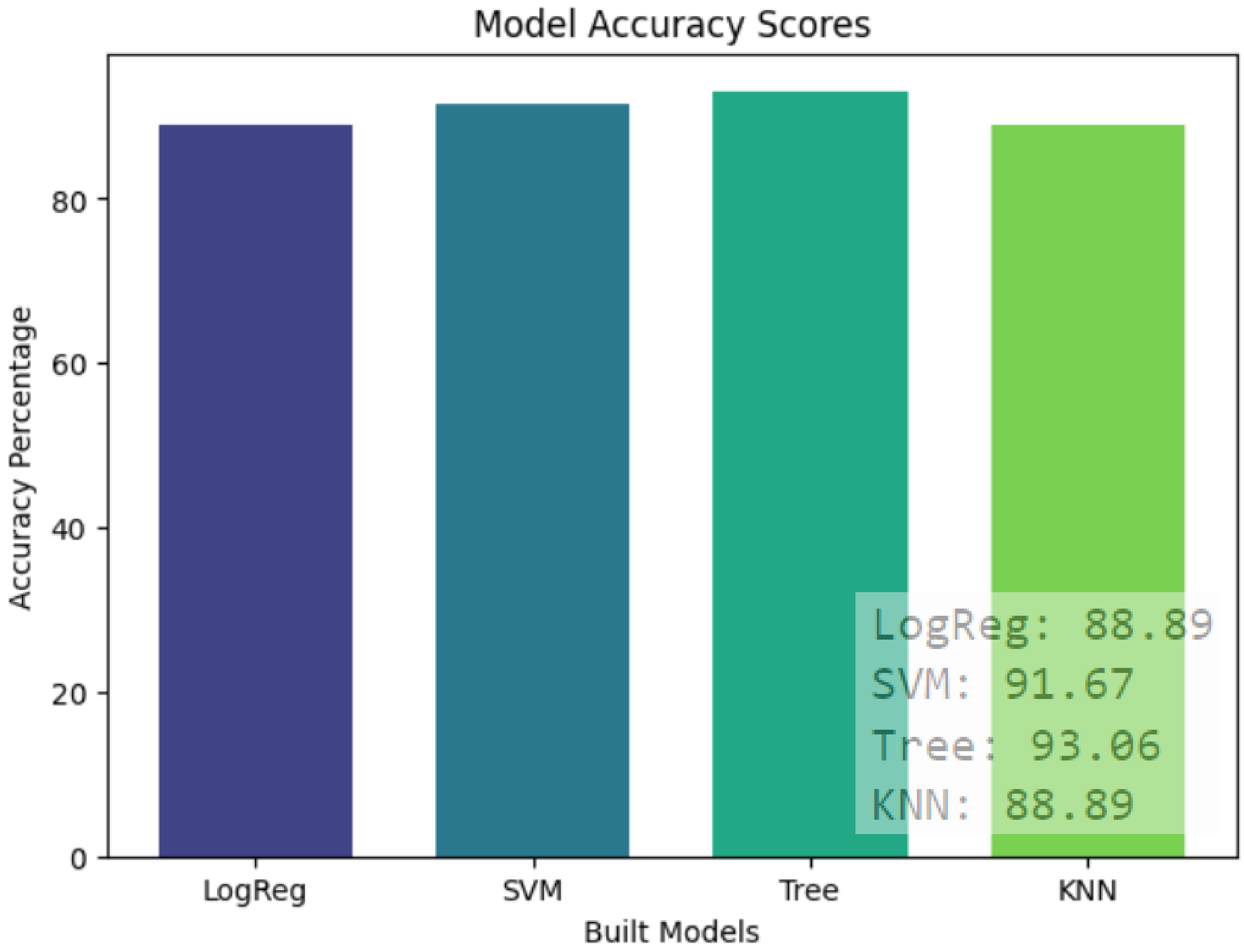
- The distribution of the "All Sites" drop down option reveals site "KSC LC-39A" as having the highest success ratio
- KSC LC-39A takes up 41.7% of total landing success outcome
- The Legend on the right signals the outcome class where blue symbolizes a 76.9% success rate

Scatterplot & Payload Range Slider

- All Sites are selected from the drop-down, featuring 3 varying Payload Range (kg) slider positions
- Booster Version “FT” (green) shows the largest success rate for their payload range of 475 kg to 5,300 kg

Section 5

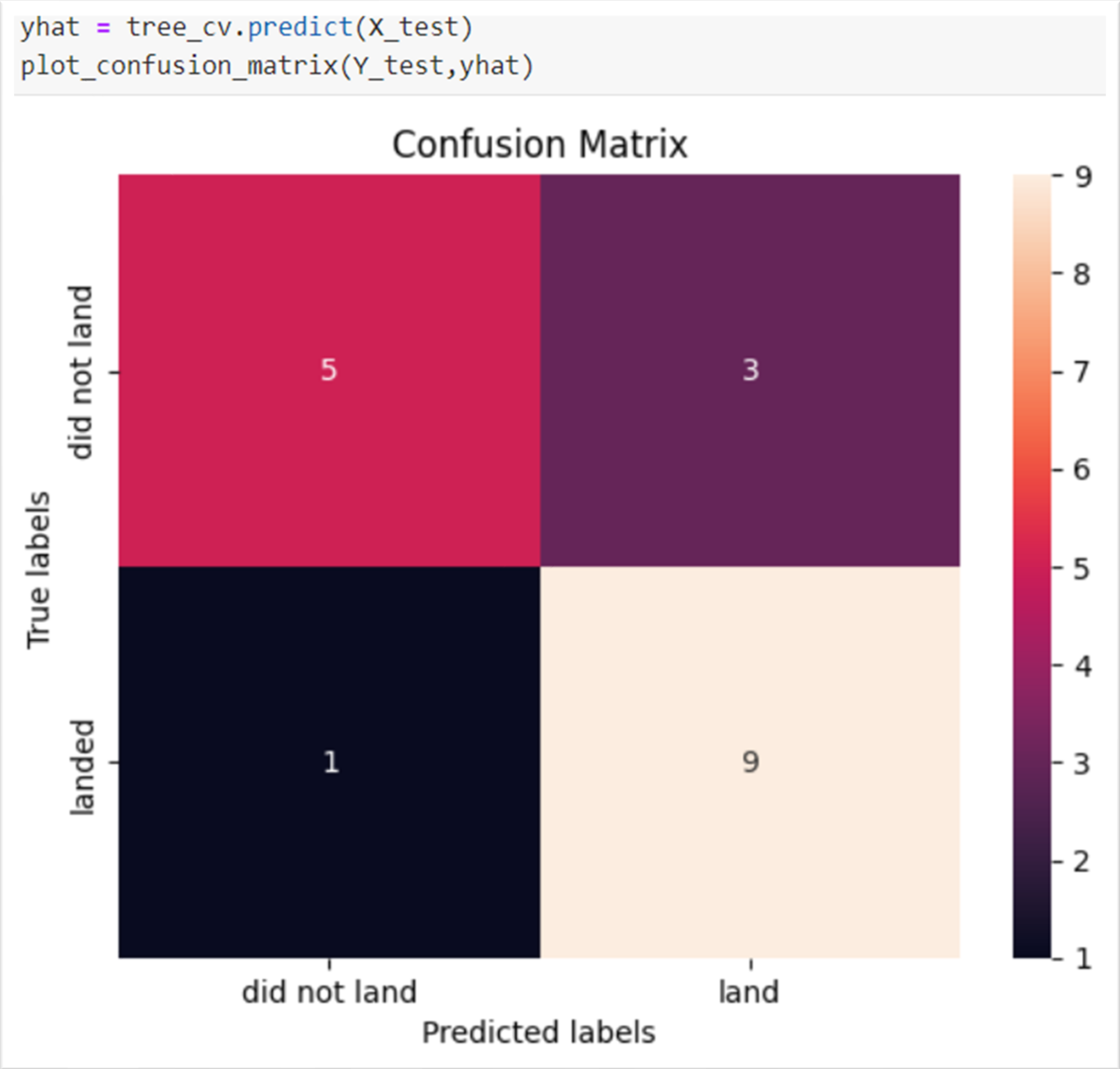
Predictive Analysis (Classification)



Classification Accuracy

This bar chart displays the distribution of Accuracy percentage organized by the different Machine Learning Models

- The **Decision Tree Classifier** has the highest Accuracy Score with Support Vector Machine coming in 2nd



Confusion Matrix

This **Confusion Matrix** features the best performing model: Decision Tree Classifier

Beginning with "Landed", the model predicted 9/10 incidents, consisting of 1 false negative

"Did not Land" features 5 true negatives, and 3 false positives

Conclusions

1. **Enhanced Predictive Accuracy for Cost Estimation**

The methods utilized in this research enable an examination process to discover the maximum potential forcost estimation accuracy. GridSearch analysis highlighted the Decision Tree Classifier machine learning model based on the data and objectives of the research. Thus, the results will ensure a competitive advantage in commercial space bids.

2. **Strategic Insights for Launch Planning**

Insights into the impact of launch sites and payload mass on success rates allow Space Y to optimize launch strategies, increasing mission success and customer satisfaction.

3. **Foundation for Future Research and Development**

This study sets the stage for further research and development, with potential to refine predictive models and enhance Space Y's technological edge in competing with Space X Falcon 9 rockets.

4. **Market Positioning and Competitive Edge**

By leveraging data science and machine learning, Space Y can position itself as a leader in the commercial space industry, offering cost-effective and reliable launch solutions.

Appendix

API: https://github.com/r-spacex/SpaceX-API/tree/master/docs#rspacex-api-docs

Falcon 9 Infographics: https://zlsadesign.com/tags/spacex-falcon9/page/2/

IBM Developer Skills Network

List of Falcon 9 and Falcon Heavy launches (Web Scraping)

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Thank you!