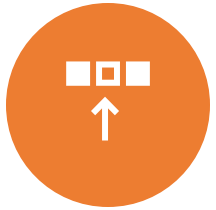
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shubham Jain
October 18, 2025

Outline



EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

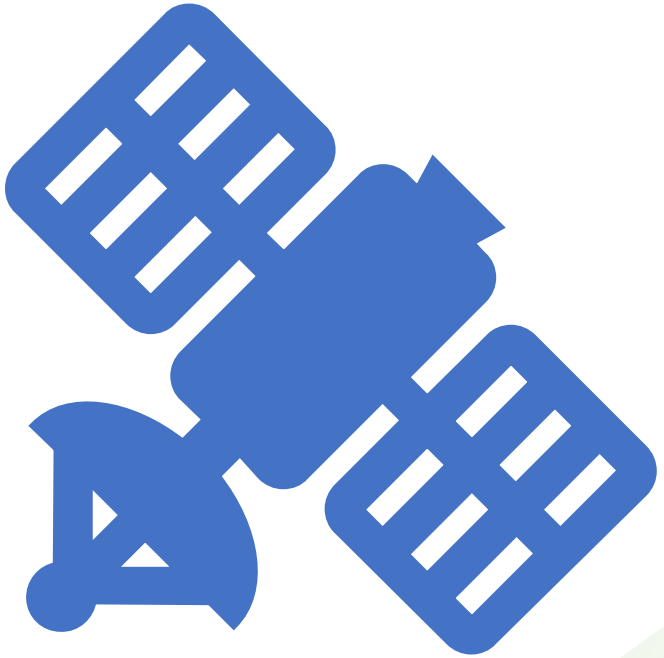
CONCLUSION

APPENDIX

Executive Summary

Summary of Methodologies:

To analyze SpaceX launch data, multiple data collection and analytical methodologies were applied:



Data Collection: SpaceX REST API was used to retrieve detailed information about past launches, including payload mass, launch site, orbit type, rocket version, and landing outcomes. Data was programmatically extracted, cleaned, and merged into structured datasets for further analysis.

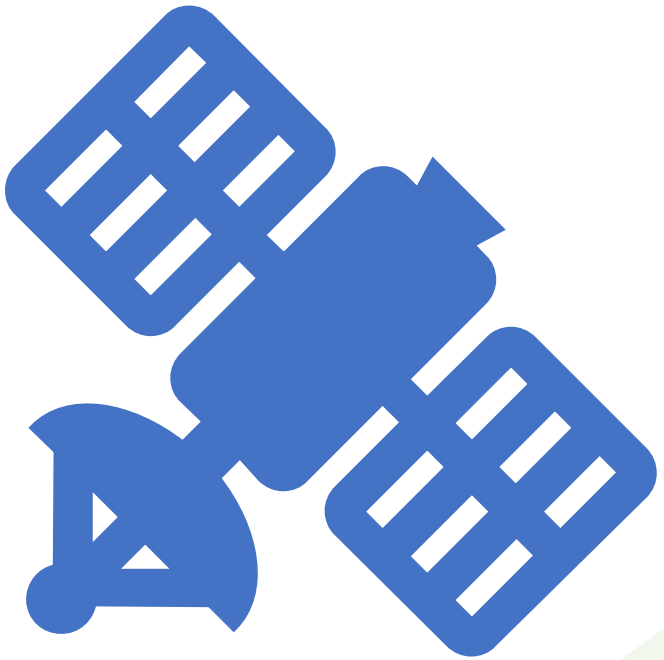
Data Wrangling & Exploration: Data inconsistencies were handled through normalization, null value imputation, and encoding of categorical variables. Exploratory Data Analysis (EDA) was conducted using Python libraries such as Pandas, Matplotlib, and Seaborn to identify patterns in launch success, orbit types, and payload distributions.

Visualization & Dashboarding: Interactive dashboards were developed using Plotly Dash and Folium to visualize launch success rates by site, payload mass trends, and geographic launch patterns. These visual tools provided actionable insights for stakeholders.

Continued....

- **Predictive Modeling:** Machine Learning models, including Logistic Regression, Decision Tree, Support Vector Machine, and K-Nearest Neighbors, were trained to predict the success of first-stage landings. Model performance was evaluated using accuracy, F1-score, and confusion matrices.

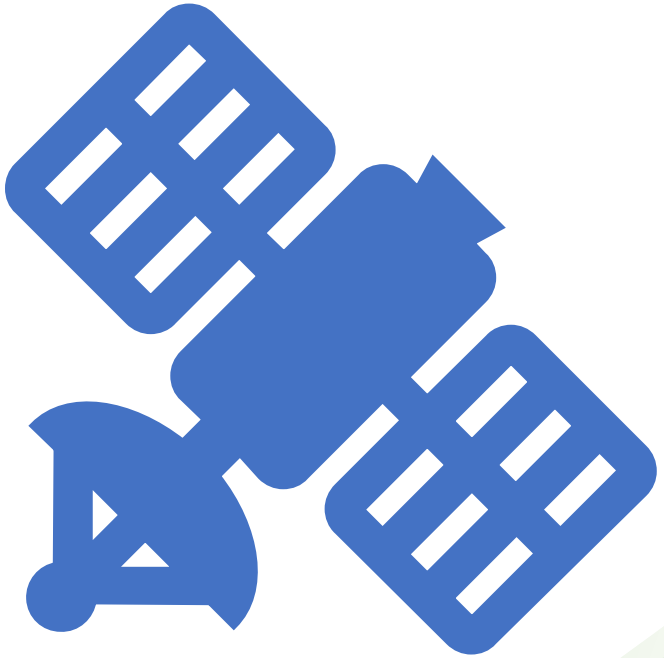
- **Model Selection & Optimization:** Hyperparameter tuning and cross-validation techniques were used to optimize model accuracy and generalization. The best-performing model was selected for deployment.



Continued....

Summary of Results

- **Key Findings**:
  - Launch success rates were strongly correlated with payload mass and orbit type.
  - Certain launch sites (notably *CCAFS SLC 40* and *KSC LC 39A*) showed consistently higher success probabilities.
  - Falcon 9 versions with reusable boosters demonstrated a steady improvement in reliability over time.
- **Machine Learning Results**:
  - Decision Tree Classifier achieved the highest prediction accuracy (~88.89%).
  - The predictive model successfully identified key factors influencing first-stage recovery, helping estimate potential launch cost efficiencies.
- **Business Impact**:
  - The analysis enables SpaceY to predict SpaceX's reusability outcomes and better estimate competitive bid prices.
  - Data-driven insights support strategic planning, cost forecasting, and market entry positioning in the commercial space sector.



Introduction

- SpaceX has revolutionized space travel with a series of groundbreaking achievements. It became the first private company to return a spacecraft from low-Earth orbit in December 2010. Its Falcon 9 rockets, priced at $62 million per launch—far below competitors' $165 million—owe much of their cost efficiency to the reusability of the first stage.

- For this project, as a data scientist at **SpaceY**, a rival company founded by visionary industrialist **Allon Musk**, my goal is to analyze SpaceX launch data to estimate launch prices and predict first-stage landings. Using **SpaceX's public REST API**, I've collected and processed data on past missions to build **interactive dashboards** and **machine learning models** that forecast reusability and inform competitive bidding strategies.

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - *Make GET request via Rest API*
- Perform data wrangling
  - *Replace missing values and create target labels for classification prediction*
- 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 Collection Process:
- Objective was to collect real-world dataset on SpaceX Falcon 9 launches
- Primary Source: SpaceX REST API
- GET request method
- Format: Data was returned in JSON
- Stored as data_falcon9.csv for data wrangling and visualization

Data Collection – SpaceX API

Helper function to extract information using identification number such as getBooterVersion()

SpaceX API Endpoint

Send GET Request using Python

Receive JSON Response

Normalize JSON using pd.json_normalize()

Clean & Select Relevant Column, dealing with missing values, applying helper function

Save as CSV File for Further Analysis using .to_csv()

Github link: https://github.com/sjain600/IBM-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

Created helper function such as date_time, booster_version to process web scraped HTML table

Send GET request to list of Falcon 9 and Falcon Heavy Launches Wikipage

Created a BeautifulSoup() object to parse response text

Apply find_all() function to extract columns using th element

Create a dataframe by parsing the launch HTML tables

Save csv files as spacex_web_scraped.csv

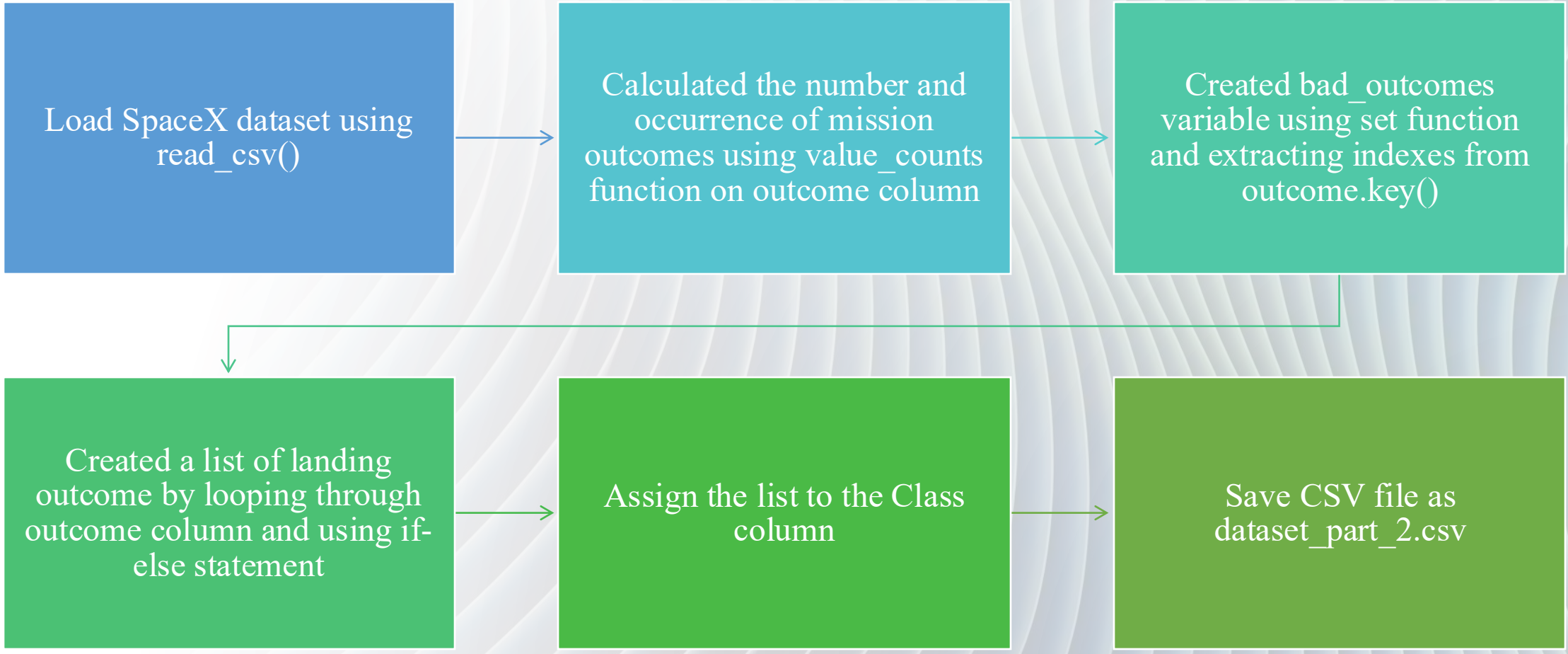
GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

The main purpose of data wrangling is to convert landing outcomes such as:

- True ASDS,
- True RTLS,
- False ASDS,
- True Ocean,
- False Ocean,
- None ASDS, and
- False RTLS into training labels with 1 means the booster successfully landed 0 means it was unsuccessful

Flowchart for Data Wrangling



GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Scatterplot

Using catplot to plot the relationship between following two variables where hue = class:

- Flight number and Payload Mass, as the flight number increases, the first stage is more likely to land successful and it seems the more massive the payload, the less likely the first stage will return.

- Flight number and Launch Sites:
  - ❑ CCAFS SLC 40 - Most flights are from this site. There is mix of success and failure launch, but success rate seems to be increasing for higher flight numbers
  - ❑ VAFB SLC 4E - This site has least number of flights but mostly successful at later flights
  - ❑ KSC LC 39A - This site has seen from mostly successful launch, there is few failures on this site around 5, indicating reliability.

Continued...

- **Payload Mass and Launch Site:**
  - ❑ VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
  - ❑ CCAFS SLC 40 site, there has been full success rate for payload greater than 12000.
  - ❑ VAFB-SLC might have used for polar or special missions. For heavier loads, nearly all success have occurred.

- **Flight Number and Orbit Type:** You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit. There has been no failure outcome for SSO. VLEO appears to be successful after 78 flights. ISS seems to be successful after 63 flights

- **Payload Mass and Orbit Type:** With heavy payloads the successful landing or positive landing rate are more for Polar, LEO, and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

EDA with Bar Plot

Visualize the relationship between success rate and orbit type:

From the bar plot, we can see that orbit types such as ES-L1, GEO, HEO, and SSO have the highest success rate.

EDA with Line Plot

You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

- GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

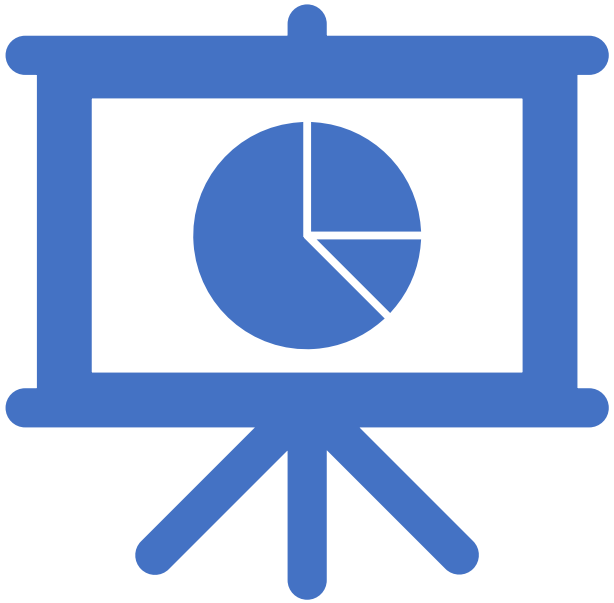
Build an Interactive Map with Folium

- Created and added circles and marker for each launch site on the map to locate them easily
- Created and added marker cluster to store marker colors based on the class value to find which launch sites have highest success rate
- Added MousePosition on the map to get coordinate over a point on the map and easily find any points of interests such as railways, highways, etc

GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

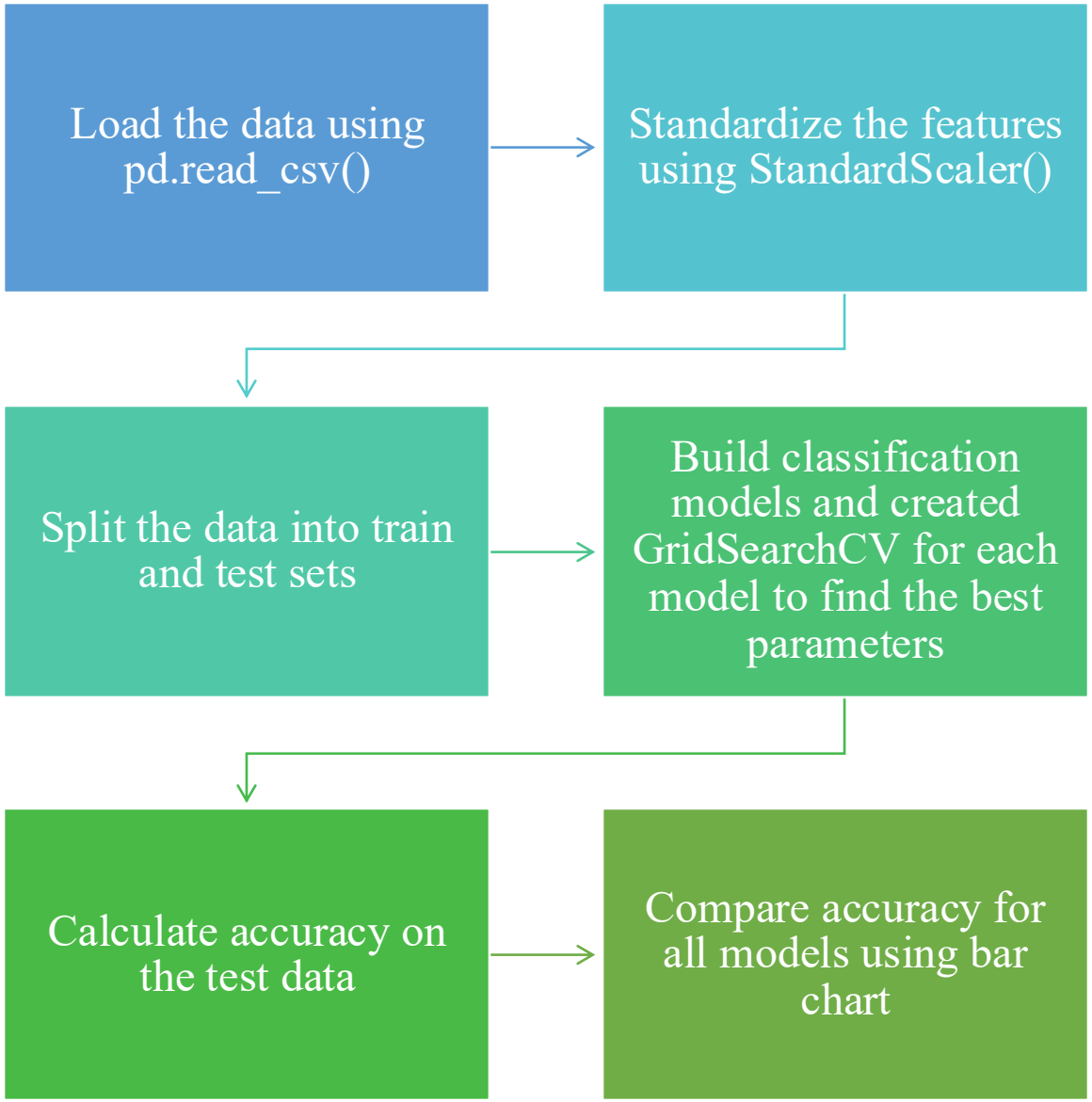
Here is the summary of the following plots and interactions added to my dashboard:



- Pie chart for Success Launches for all sites – to compute which site has the highest success rate
- Pie Chart for Highest Launch Success Ratio – to compute which site has the highest success ratio
- Scatterplot: Payload vs Launch Outcome – to plot the relationship between payload and successful outcome using booster version as the third variable

GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/dash_interactivity_cap.py

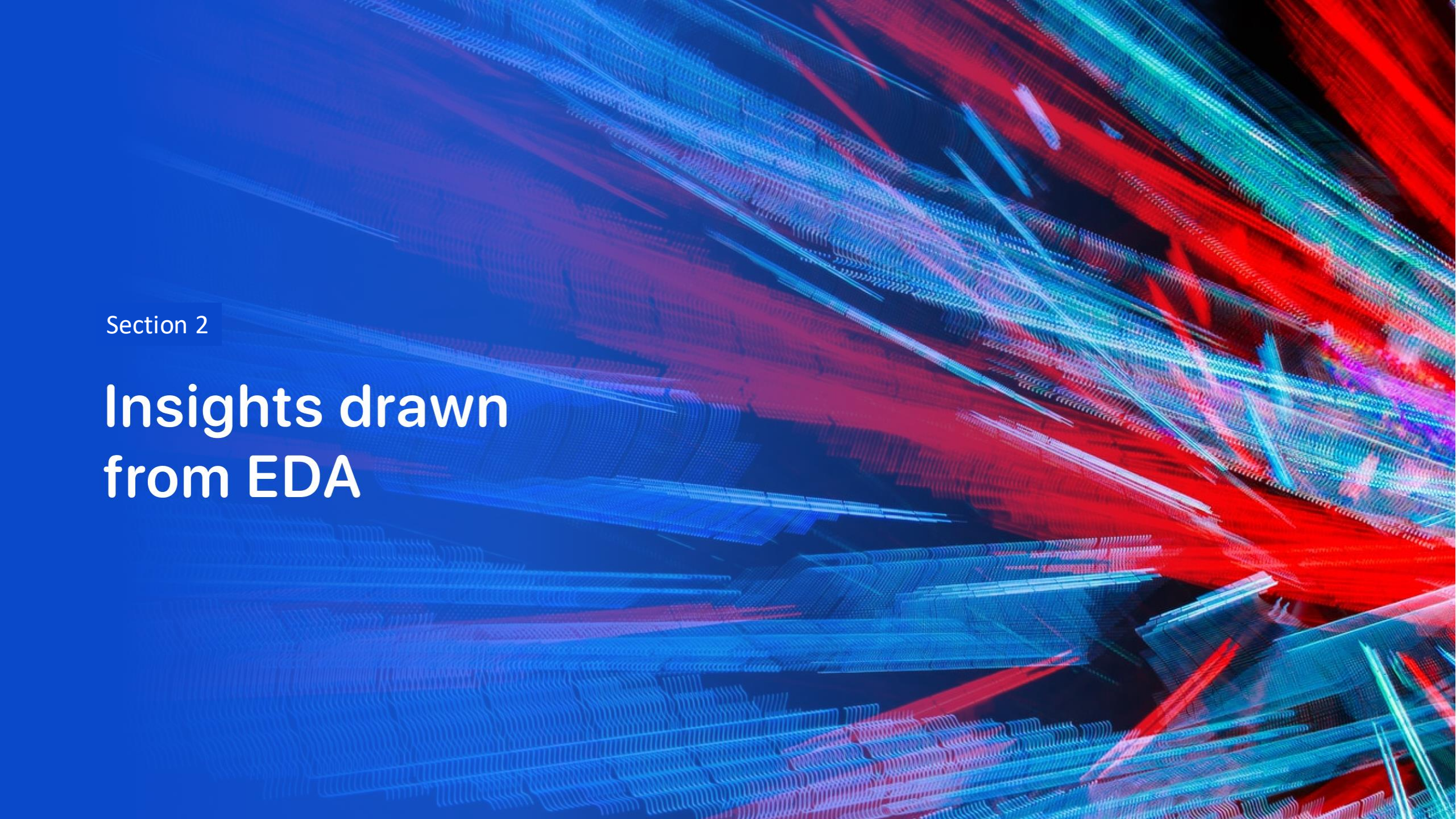
Flowchart for Predictive Analysis

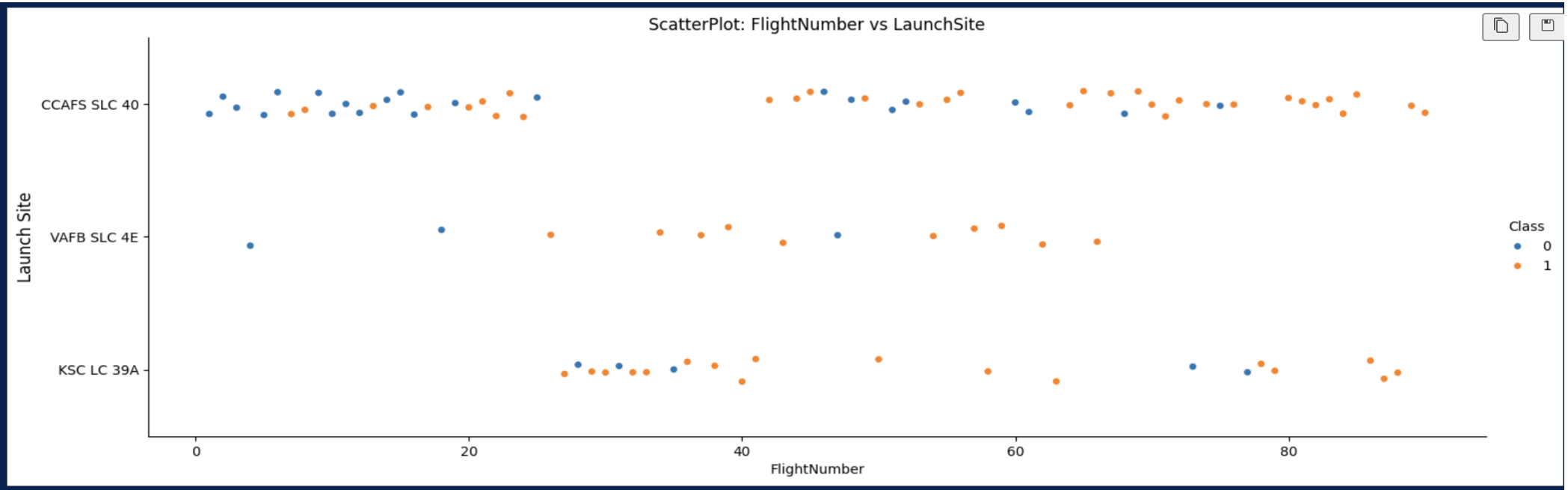


GitHub: https://github.com/sjain600/IBM-Capstone-Project/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

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


Section 2
Insights drawn from EDA

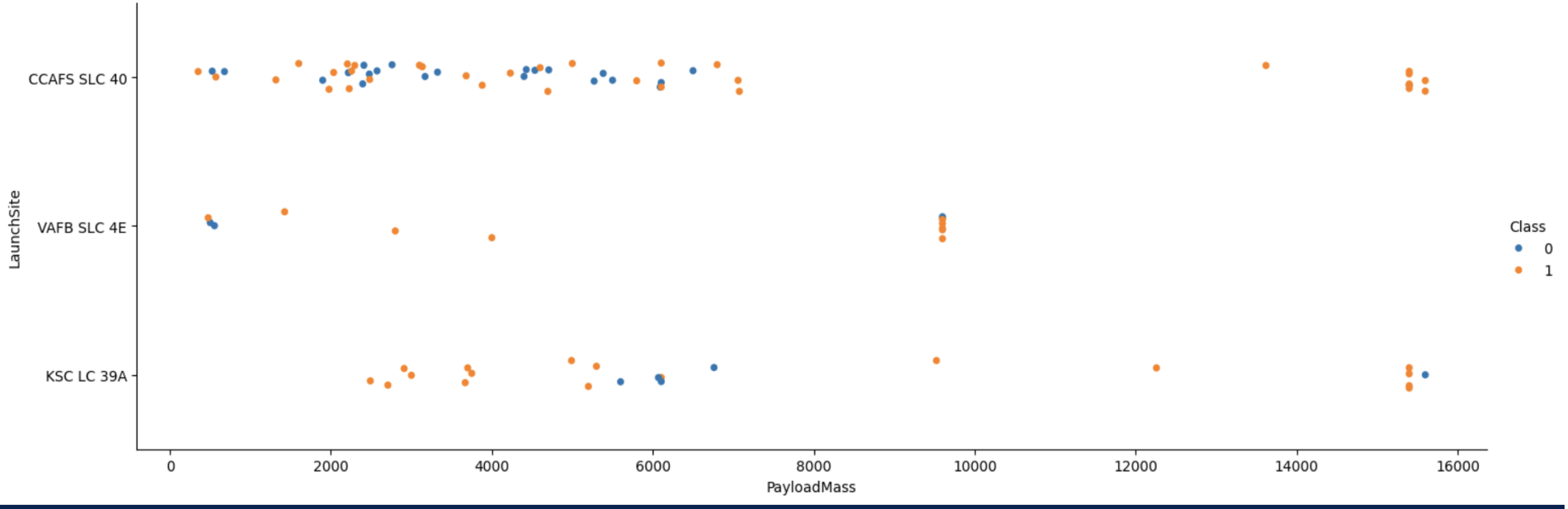


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots

From the scatterplot above, we can inference from the following launch sites below:

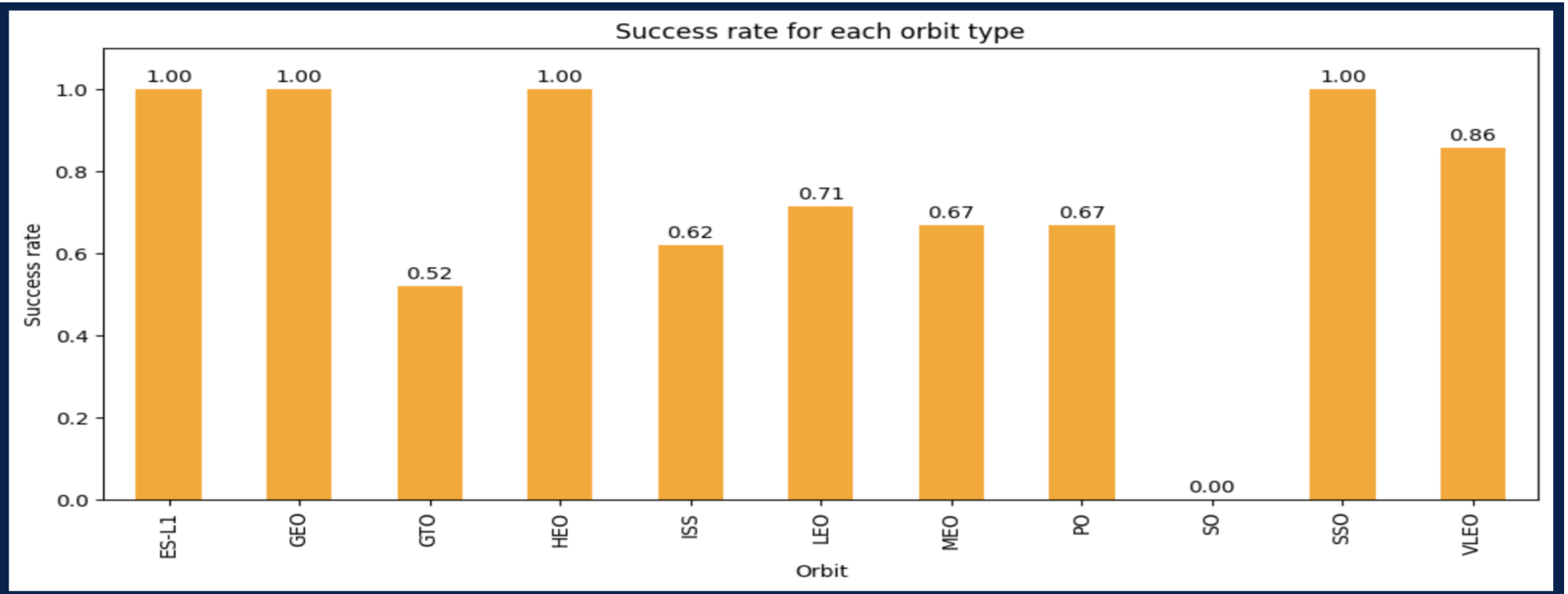
- CCAFS SLC 40 - Most flights are from this site. There is mix of success and failure launch but success rate seems to be increasing for higher flight numbers.
- VAFB SLC 4E - This site has least number of flights but mostly successful at later flights
- KSC LC 39A - This site has seen from mostly successful launch, there is few failures on this site around 5, indicating reliability

Flight Number vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000). For the CCAFS SLC 40 site, there has been full success rate for payload greater than 12000. VAFB-SLC might have used for polar or special missions. for heavier loads, nearly all success have occurred.

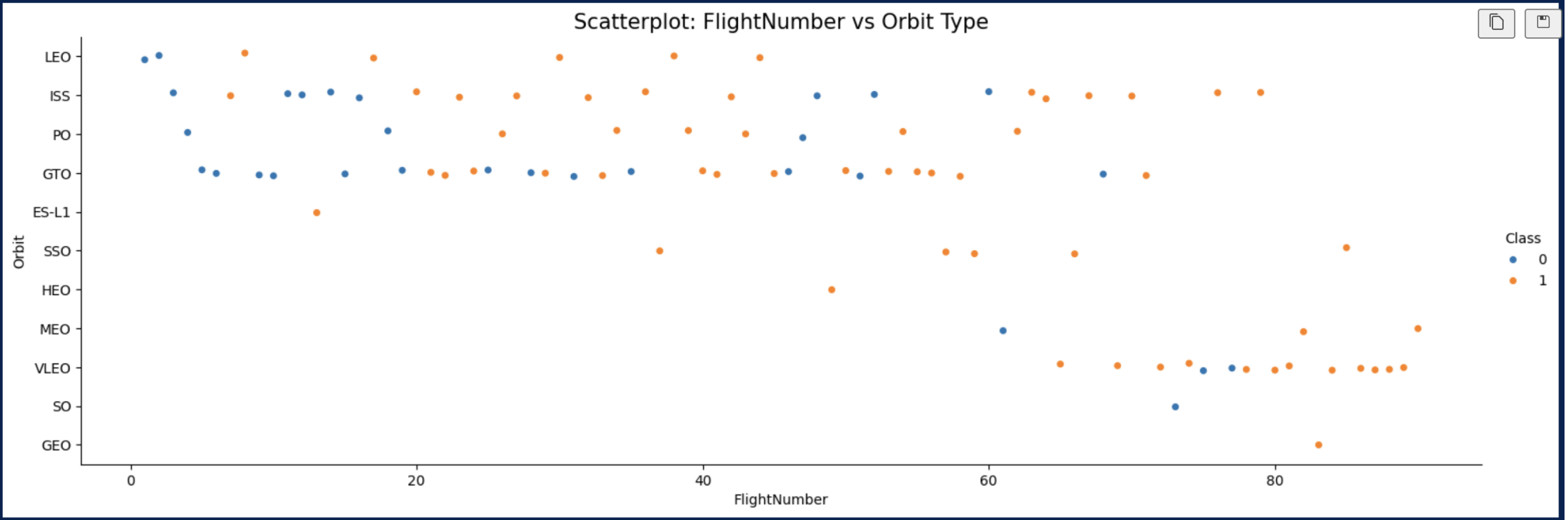
Payload vs. Launch Site



Analyze the ploted bar chart try to find which orbits have high sucess rate.

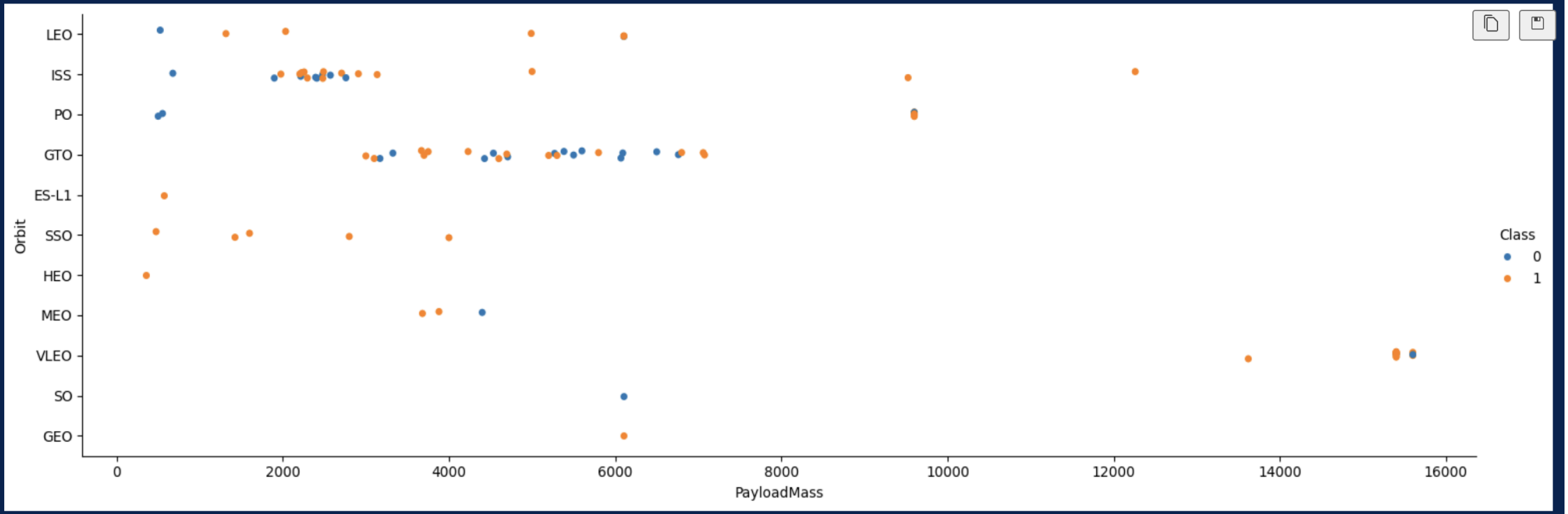
From the bar chart above, we can see that orbit types such as ES-L1, GEO, HEO, and SSO have the highest success rate.

Success Rate vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Flight Number vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

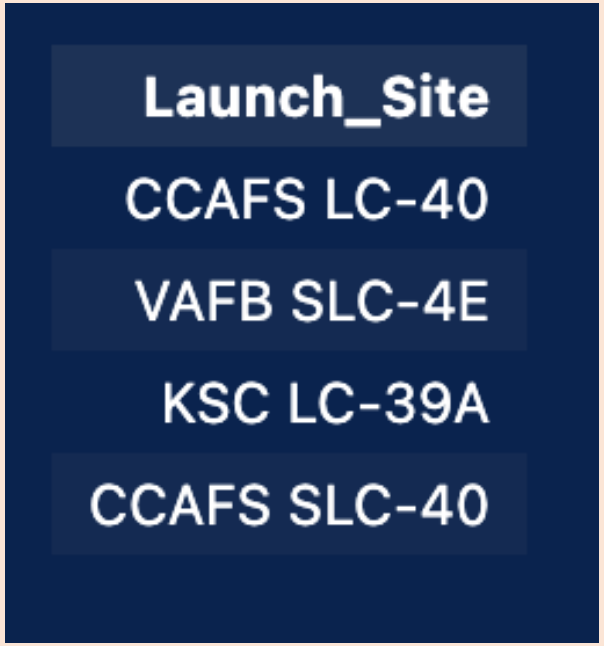
Payload vs. Orbit Type

Launch Success Yearly Trend

You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

We can observe from query result that they are four launch sites.



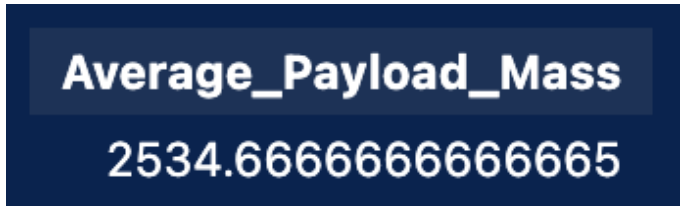
Launch Site Names Begin with 'CCA'

We can view from top 5 results that they are 5 successful mission outcomes for the orbit LEO.

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

Total Payload Mass by NASA(CRS) is 111,268 kgs



Average Payload Mass by F9 v1.1

The average payload mass by booster version F9 v1.1 is 2534.67 kgs.

First Successful Ground Landing Date

First Successful Ground Landing Date was on December 12, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

There are four boosters which had successful drone ship landing with Payload between 4000 and 6000 kgs.

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

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful mission outcomes is 100 and failed is only 1.

Boosters Carried Maximum Payload

Here is the list of boosters carried maximum payload of 15,600 kgs.

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

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

2015 Launch Records

We can see that there are two failed drone ship outcome on the launch site CCAFS LC-40.

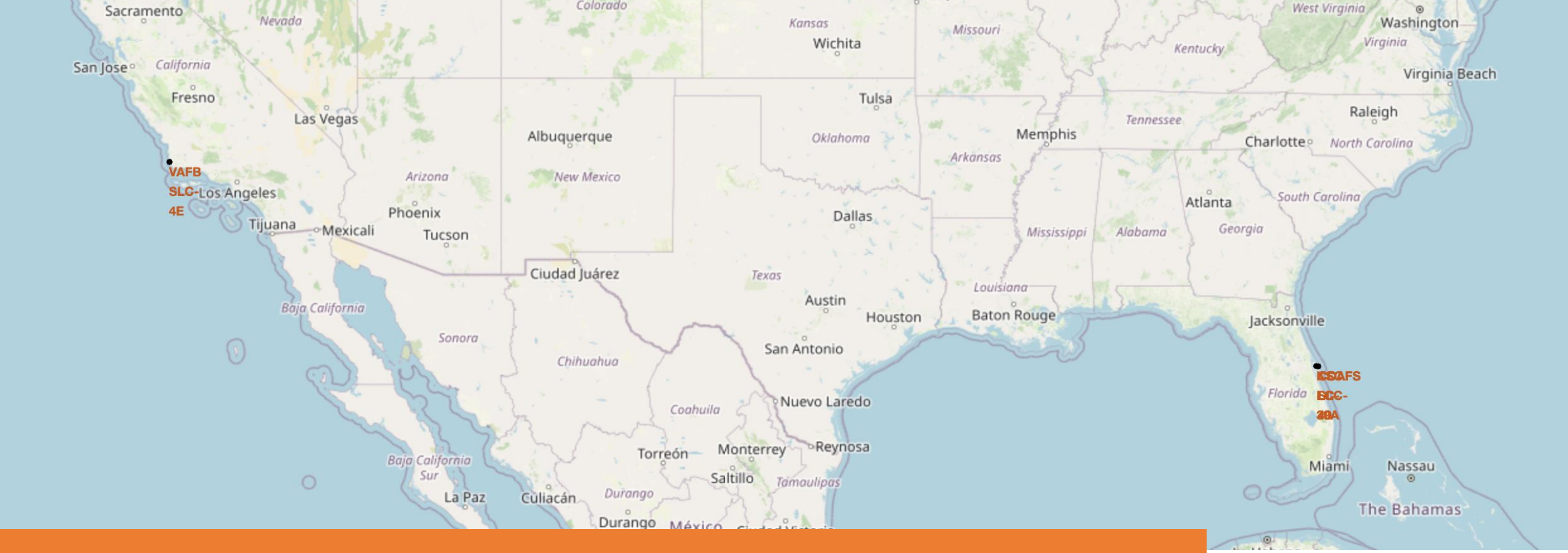
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

We can observe from the query result that majority of landing outcomes came from launch site CCAFS LC-40, where most counts (10) are no attempts.

Date	Launch_Site	Landing_Outcome	Count
2012-05-22	CCAFS LC-40	No attempt	10
2016-04-08	CCAFS LC-40	Success (drone ship)	5
2015-01-10	CCAFS LC-40	Failure (drone ship)	5
2015-12-22	CCAFS LC-40	Success (ground pad)	3
2014-04-18	CCAFS LC-40	Controlled (ocean)	3
2013-09-29	VAFB SLC-4E	Uncontrolled (ocean)	2
2010-06-04	CCAFS LC-40	Failure (parachute)	2
2015-06-28	CCAFS LC-40	Precluded (drone ship)	1

Section 3

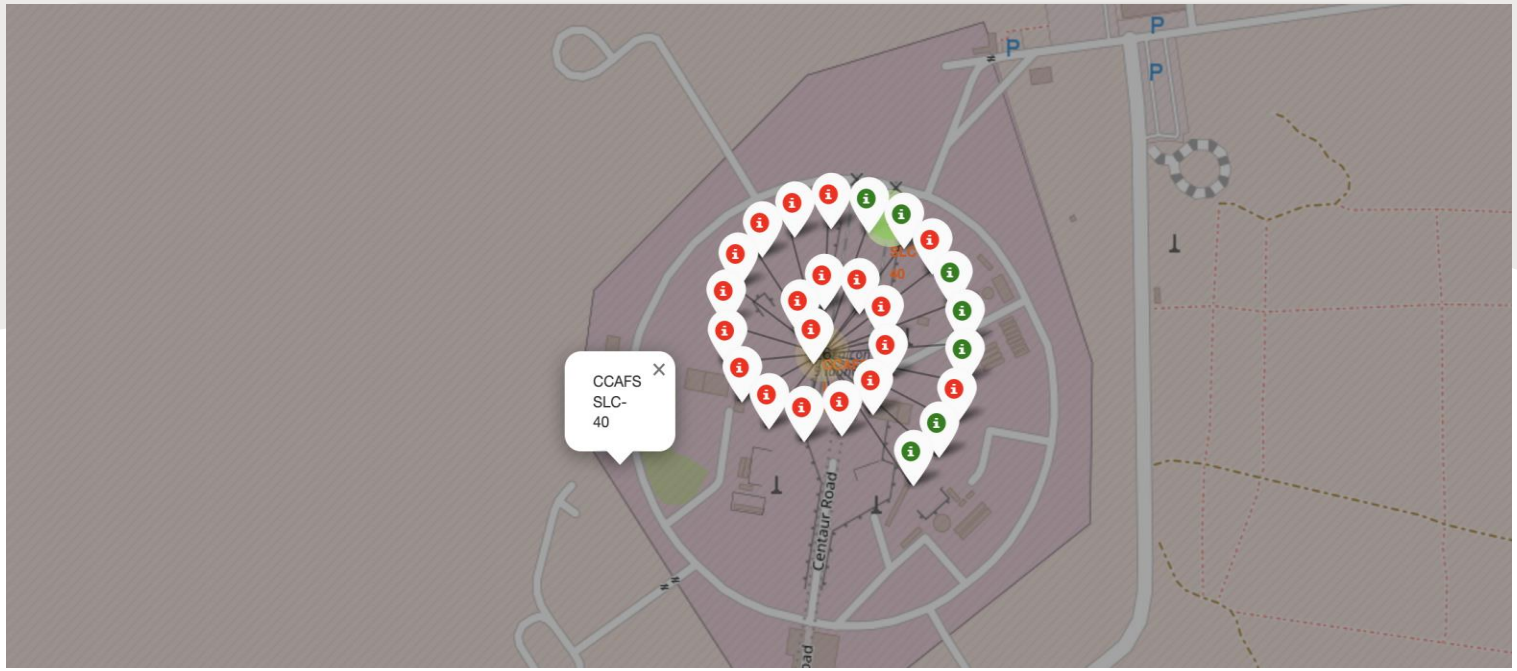
Launch Sites Proximities Analysis



Map of three launch sites

We can see from the map that there are three launch sites are in close proximity to the coast. However, there are not near to the equator line.

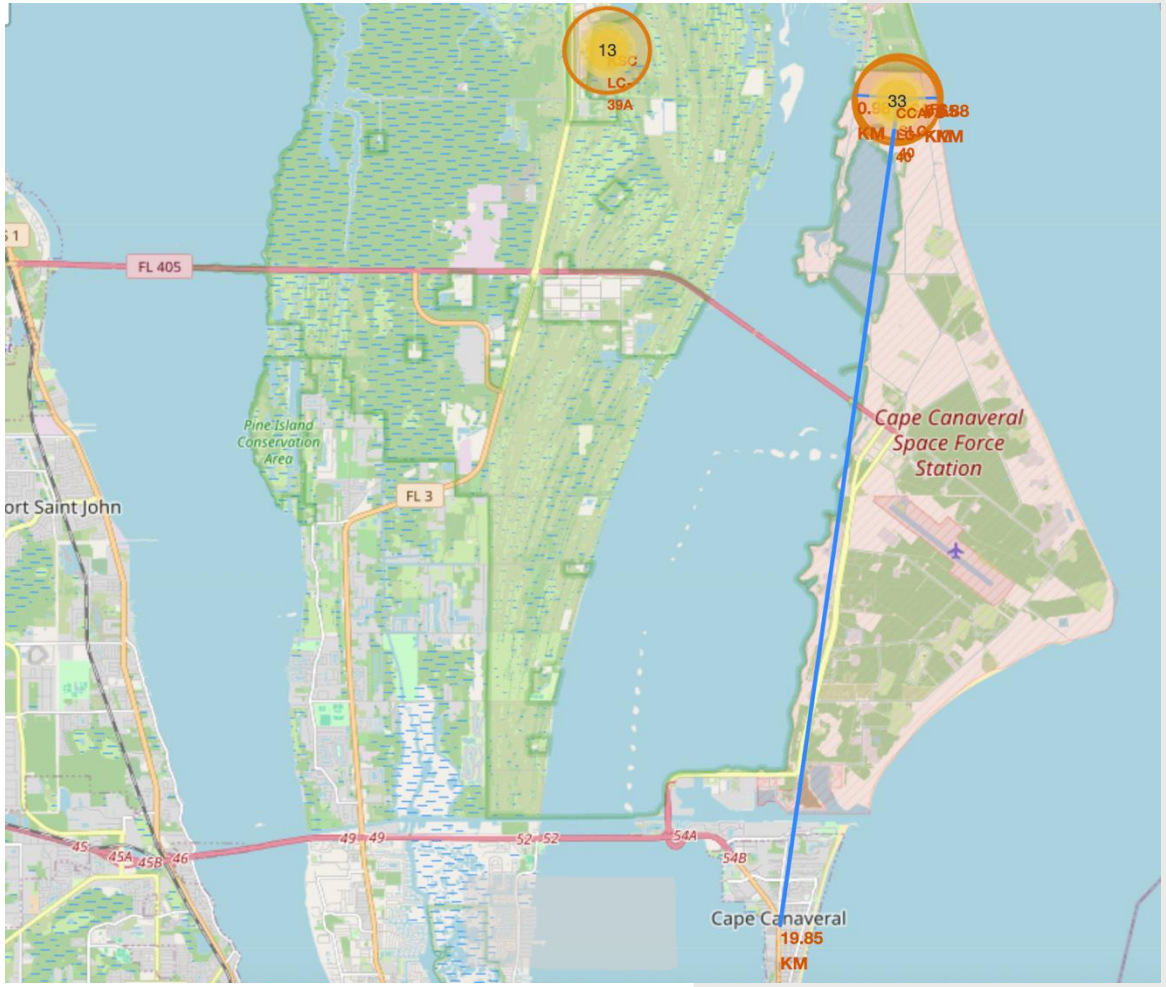
Launch outcomes for CCAFS SLC-40

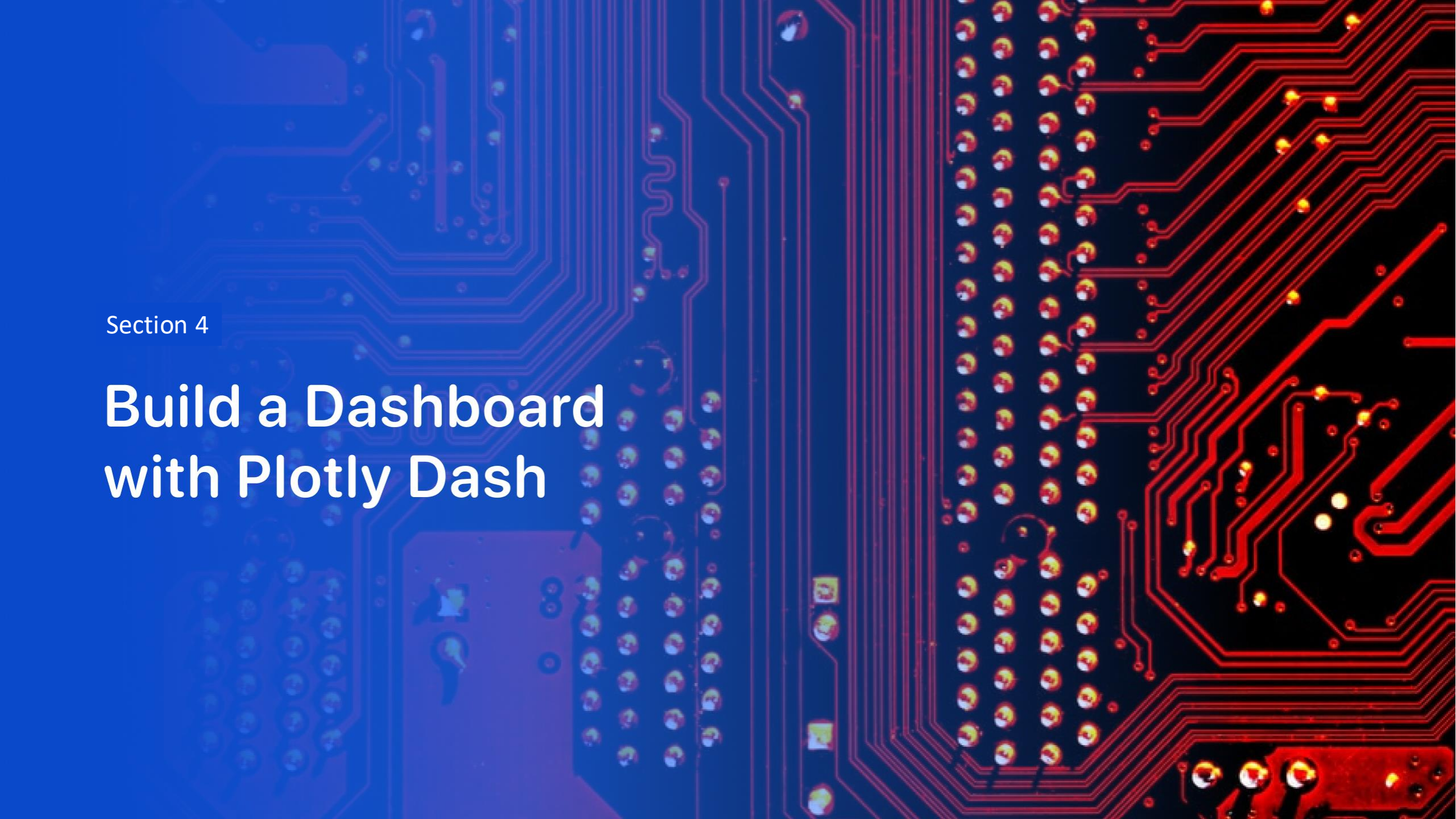


We can see from map that the success rate for CCAFS SLC-40 is low since there are higher number of failed outcomes.

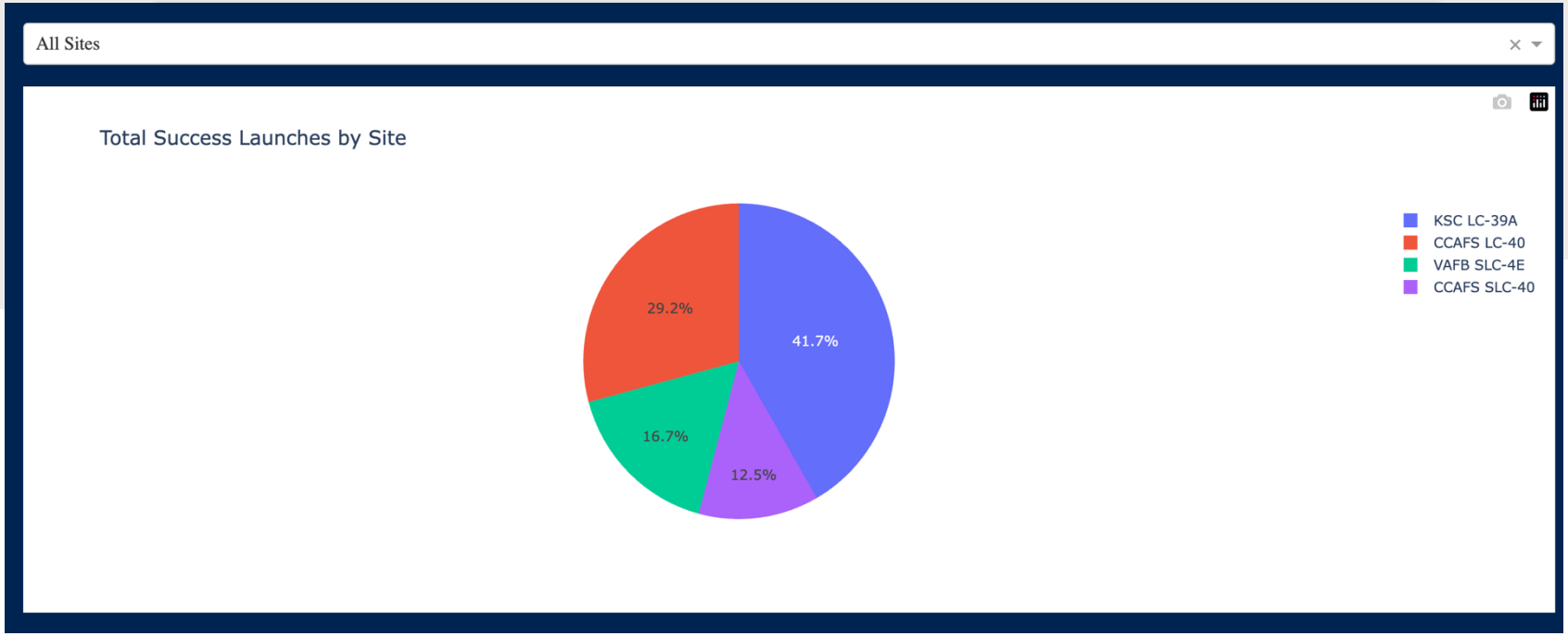
Proximities to launch sites

- Few insights gained from Folium Map is that:
- Nearest city to CCAFS is Cape Canaveral, which is 19.85 km away
- Nearest highway is Samuel C Phillips Parkway (0.65 km away)
- Coastline is 0.88 km away
- Railways is 0.98 km away



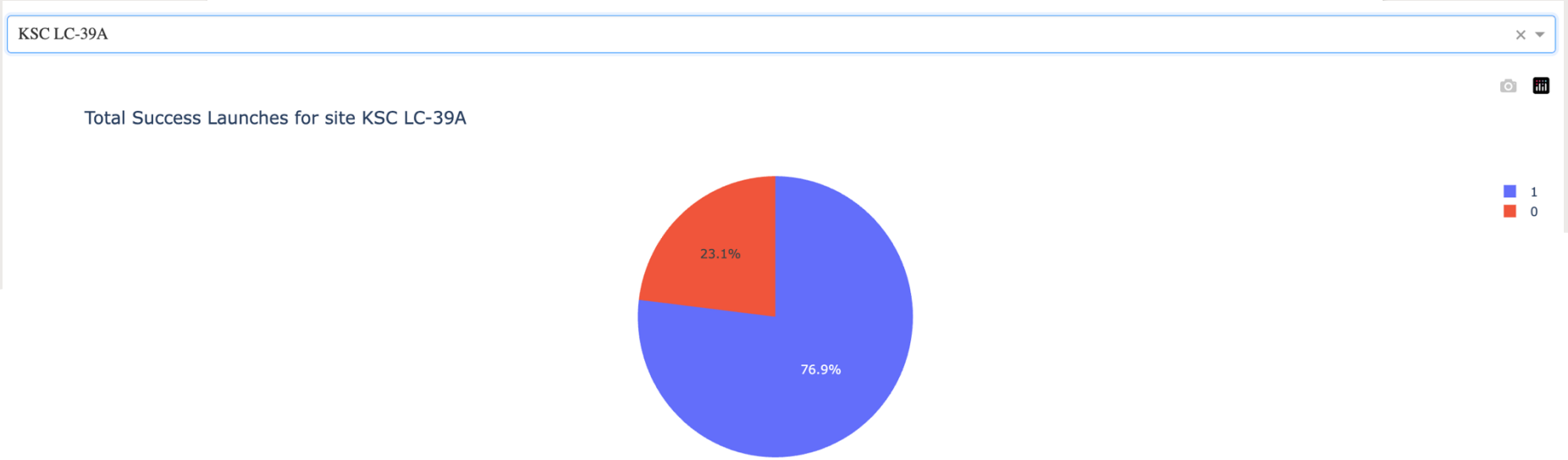

Section 4
Build a Dashboard with Plotly Dash

Pie chart for Success Launches for all sites



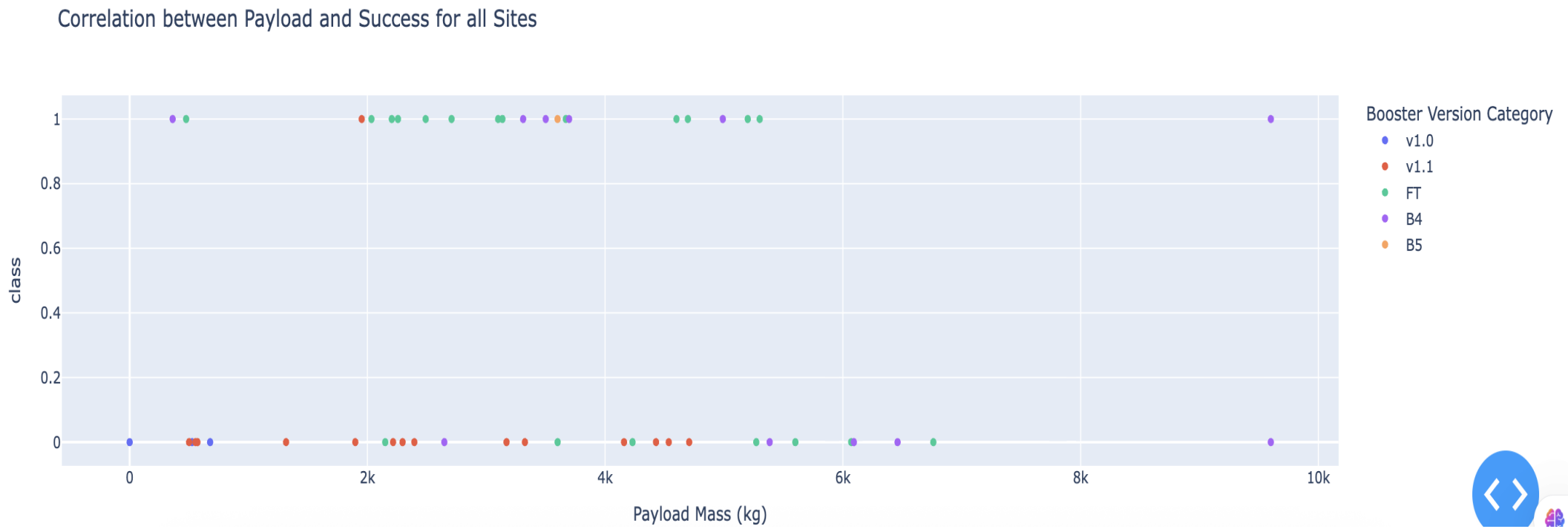
From the pie chart above, we can see that the success rate of KSC LC-39A is highest amongst other launch sites, which is 41.7%.

Pie Chart for Highest Launch Success Ratio



We can observe from the pie chart that total success launches for KSC LC-39A is highest percentage of 76.9%.

Scatterplot: Payload vs Launch Outcome

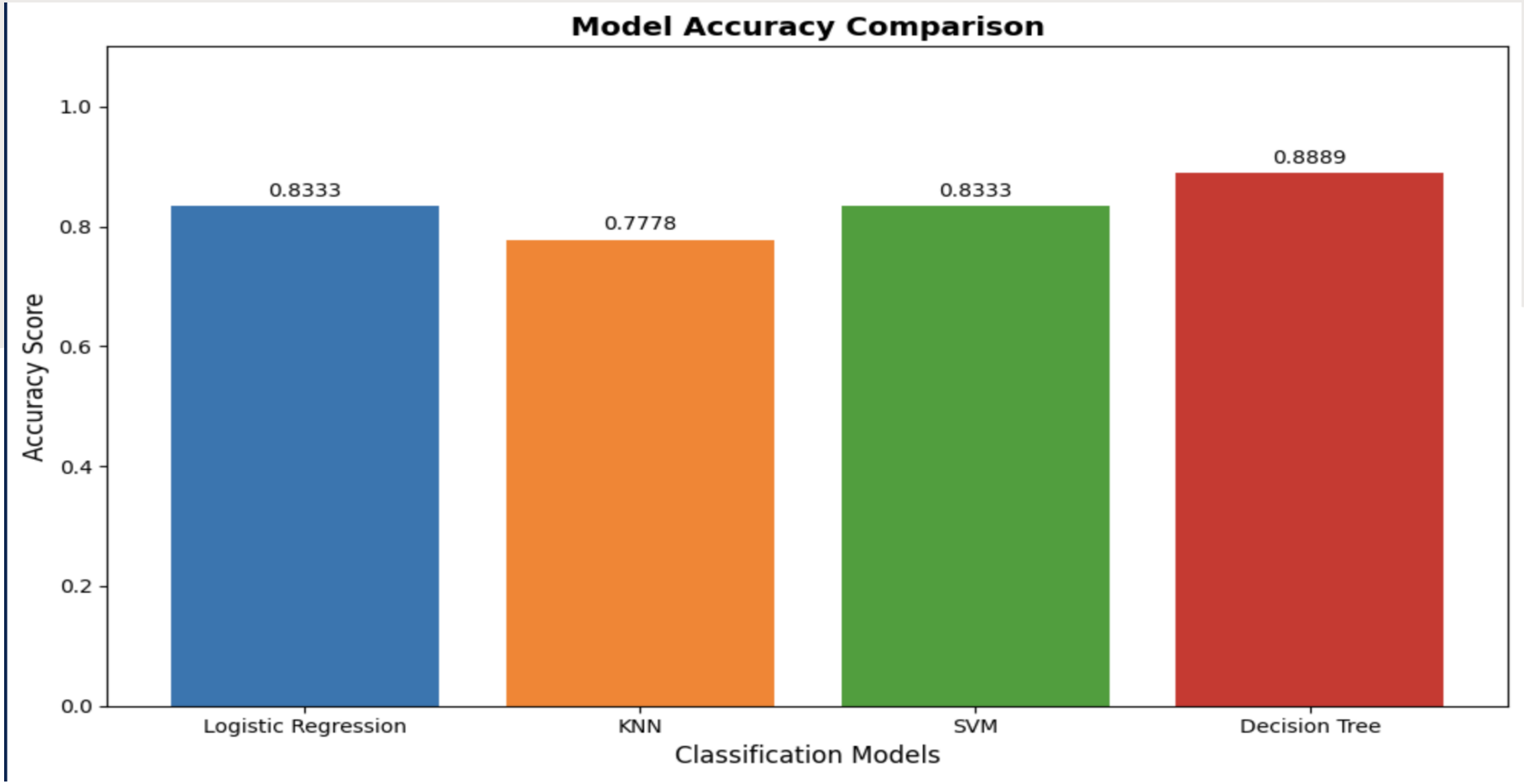


From the scatterplot above, it is hard to inference that which payload range or booster version has the highest success rate since the results seems to be mixed

Section 5

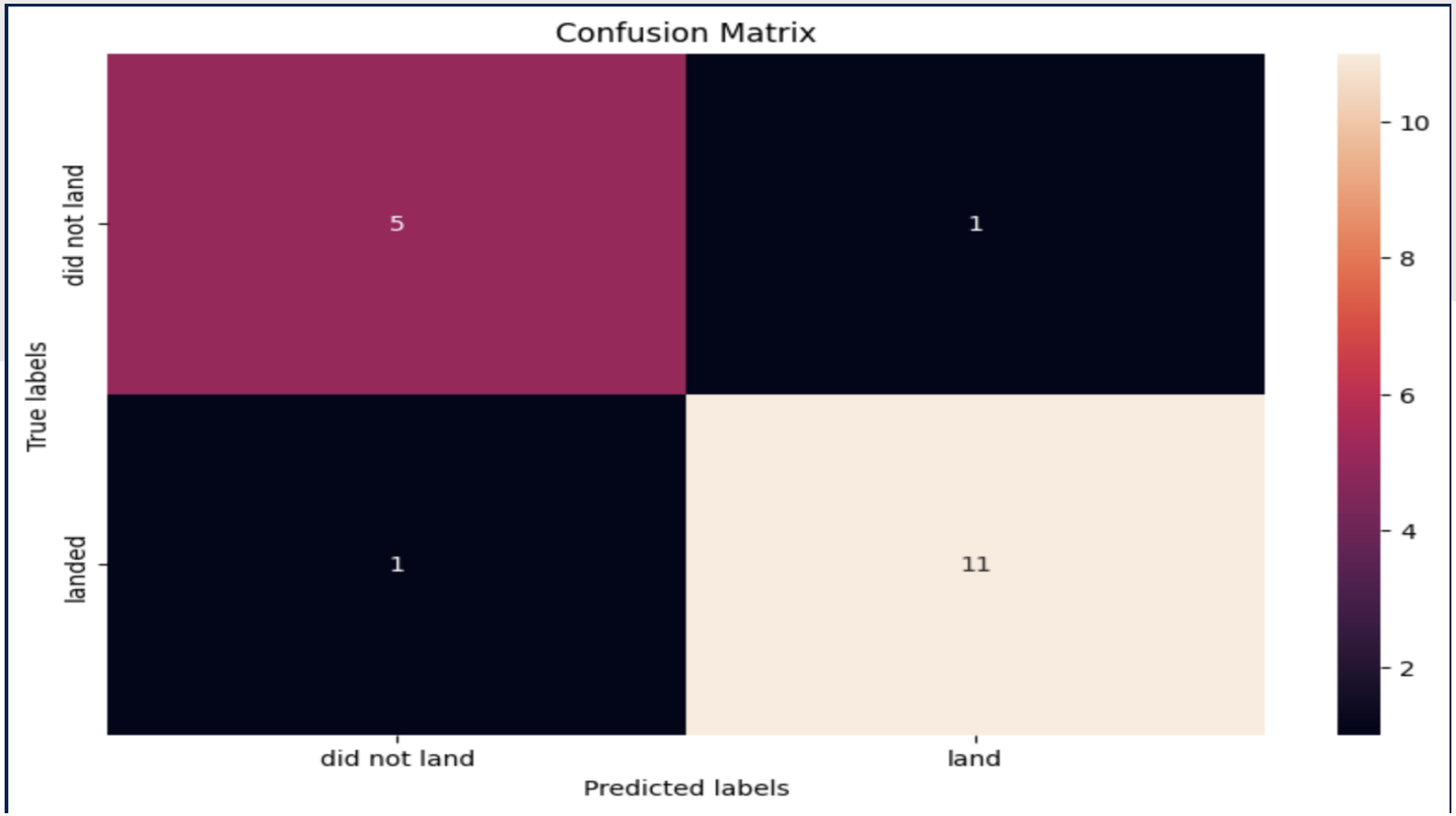
Predictive Analysis (Classification)

Classification Accuracy



We can inference from the bar chart above that Decision Tree Classifier has the highest accuracy rate of 88.89% on test data.

Confusion Matrix



Examining the confusion matrix shows that **true positive labels are 11 and false positive labels are 1**.

Conclusions

Through comprehensive data collection, exploratory analysis, visualization, and predictive modeling, this project provided actionable insights into SpaceX's launch performance and cost determinants.

These findings equip SpaceY with a competitive edge in forecasting launch costs and optimizing bidding strategies against SpaceX.

Appendix

Relevant Assets for the SpaceX Capstone Project

1. Python Code Snippets:

- *Data Collection:* Used the SpaceX REST API to fetch launch data using Python requests library.

```
import requests
spacex_url = "https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
data = response.json()
```

- *Data Wrangling:* Cleaned and transformed data using pandas.

```
import pandas as pd
df = pd.json_normalize(data)
df.to_csv('spacex_launch_data.csv', index=False)
```

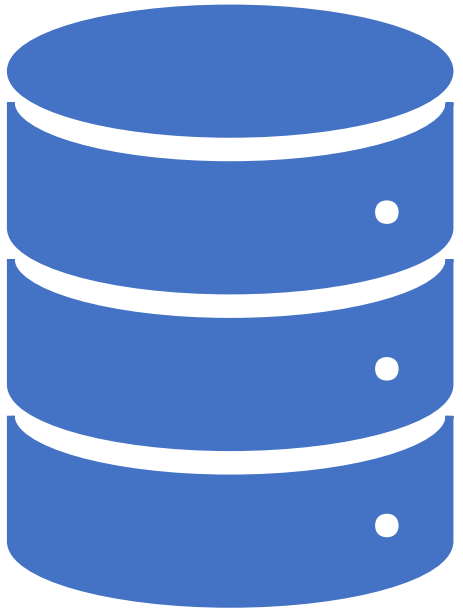
Feature Engineering: Extracted booster version categories, landing outcomes, and payload mass.

Continued...

2. SQL Queries:

Queried launch data stored in a database to extract specific launch outcomes and success rates.

```
SELECT LaunchSite, COUNT(*) AS Total_Launches,
    SUM(LandingOutcome='Success') AS Successful_Landings
FROM spacex_data
GROUP BY LaunchSite;
```



Continued...

3. Charts and Visualizations:

- *Launch Success by Site:* Bar charts and pie charts created using Plotly Express.
- *Payload Mass vs. Orbit Type:* Scatter plots to analyze correlation.
- *Success Trend Over Time:* Line graph showing improvement in landing success rate over years.

Continued...

4. Model Prediction Output:

Included tables showing top-performing models (KNN, Logistic Regression, SVM, Decision Tree).

Model accuracy comparisons from test data:

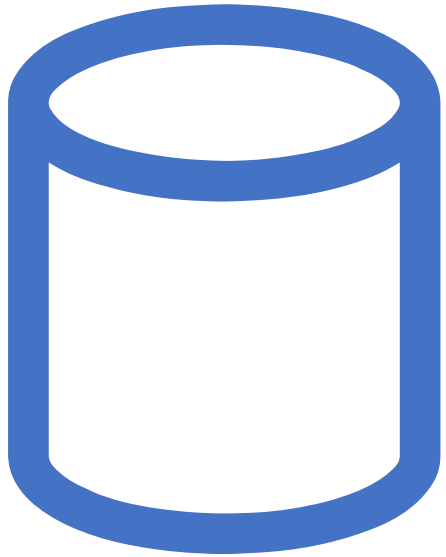
- Logistic Regression: 0.83
- KNN: 0.78
- SVM: 0.83
- Decision Tree: 0.89

Continued...

5. Datasets Used:

- *SpaceX API Data:* Retrieved directly from SpaceX REST API.
- *Enhanced Dataset:* Merged API data with additional CSV files containing orbit, payload, and booster information.

These assets together formed the foundation for data exploration, visualization, and predictive modeling in the SpaceX Capstone Project.



Thank you!