IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Davide De Carlo
08 June 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix/Additional insights

Executive Summary

- Summary of methodologies
  - Data collected through SpaceX API and web-scraping
  - Data prepared and transformed into a useable subset
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
  - Conclusion
- Summary of all results
  - EDA identified the best features to predict successful landings;
  - Machine Learning model predicted if the Falcon 9 first stage will land successfully.

Introduction

Project background and context

In this project, the goal is to predict if the Falcon 9 first stage will land successfully. The context revolves around predicting the success of Falcon 9 first stage landings, which is crucial in determining the cost of a launch. SpaceX, offering launches at a significantly lower cost compared to other providers, achieves these savings through the reuse of the first stage. Consequently, understanding the determinants of successful landings can be valuable for companies competing with SpaceX in bidding for rocket launches.

Problems you want to find answers

- What factors influence the successful landing of a rocket?
- How do different relationships with specific rocket variables impact the success rate of a landing?
- What conditions must SpaceX fulfill to optimize rocket success rates?
- Can we develop a model capable of predicting landing success to accurately calculate launch costs?
- Which launch site(s) offer the best performance in terms of successful landings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web-scraping of Falcon 9 Wikipedia page
- Perform data wrangling
  - incorporated a landing outcome label derived from outcome data, followed by the summarization and analysis of features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data were normalized, divided into training and test data sets, and evaluated by four different classification models, the accuracy of each model was assessed.

Data Collection

We collected data sets from SpaceX API and, via web scraping techniques from the Falcon9 Wikipedia page.

To gather launch-related information, we utilized the API once again by querying the launch IDs. Specifically, we focused on extracting data from the following columns: rocket, payloads, launchpad, and cores.

- From the rocket data, we obtained details such as the name of the booster.
- By analyzing the payload data, we obtained insights into the payload's mass and the intended orbit.
- The launchpad data provided us with valuable information, including the name of the launch site, as well as its longitude and latitude coordinates.

Through examining the cores data, we gained knowledge about the landing outcome, the type of landing, the number of flights associated with a specific core, the usage of grid fins, reusability status of the core, and other relevant details.

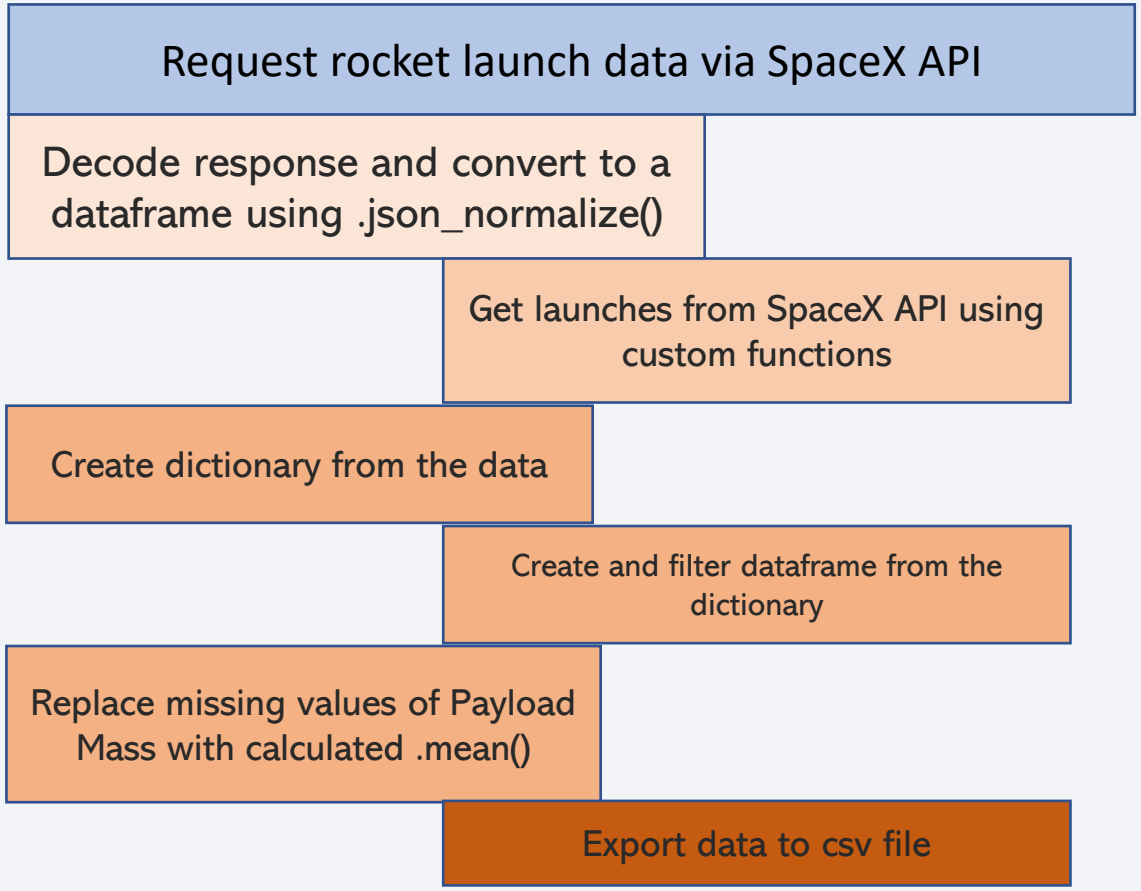
Data Collection – SpaceX API

Data is collected from SpaceX REST API which are about launch and landing specifications, rocket usage.

Data was further enriched via SpaceX API calls with additional API data

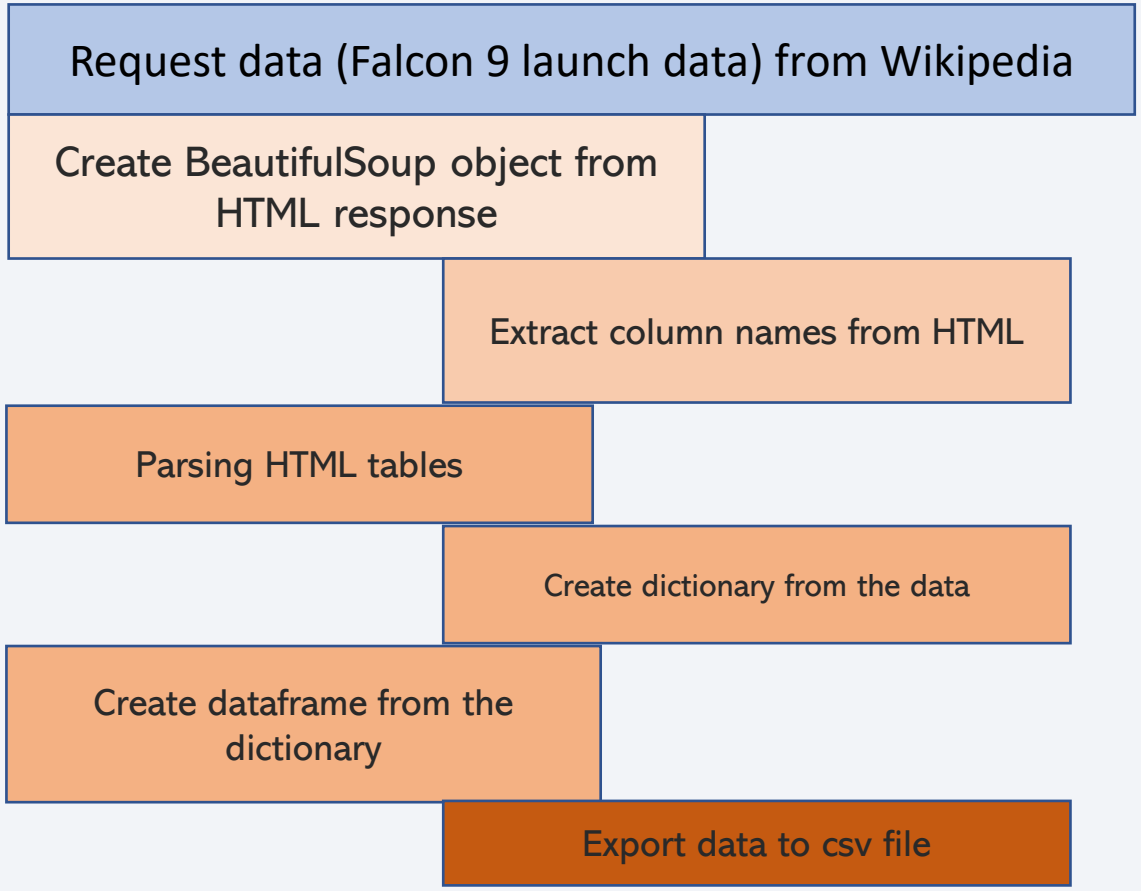
- Rocket
- Launchpad
- Payload Mass
- Orbit

Collecting the data GitHub URL



Data Collection - Scraping

Web scraping performed from the Wikipedia page to collect Falcon 9 historical launch records.



Web Scraping [GitHub]

Data Wrangling

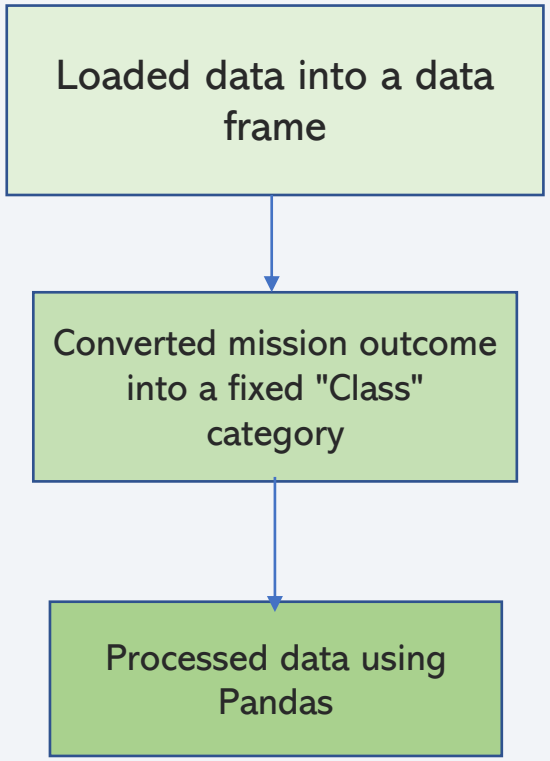
1. Loaded previously acquired data into a data frame.
2. Converted mission outcome text values into a fixed "Class" category (0 for fail, 1 for success) and added a new column.
3. Processed data using Pandas:

Calculated the number of launches for each site.

Determined the number and occurrence of different orbits.

Replaced missing data in critical columns with the average of existing data.

The processed data is now prepared for further analysis.



Data Wrangling [github]

EDA with Data Visualization

Scatter Charts

Flight Number vs. Launch Site: Identify successful launch sites and observe trends over time.

Payload vs. Launch Site: Determine launch site requirements based on payload weight.

Flight Number vs. Orbit Type: Analyze mission outcome by orbit type over time.

Payload vs. Orbit Type: Investigate if payload weight influences optimal orbit.

Bar Chart:

Orbit vs. Success Rate: Determine which orbits have higher success rates.

Line Chart:

Launch Success Trend: Track the yearly trend of average success rates.

EDA with Data Visualization [github]

EDA with SQL

List:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Display:

- Names of unique launch sites
- 5 records where launch site begins with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

EDA with SQL [github]

Build an Interactive Map with Folium

Markers Indicating Launch Sites:

- Placed a blue circle at the coordinate of NASA Johnson Space Center, displaying its name as a popup label using latitude and longitude coordinates.
- Added red circles at the coordinates of all launch sites, showing their names as popup labels using latitude and longitude coordinates.

Map with Folium:

Created an interactive map using Folium library.

Colored Markers of Launch Outcomes:

Represented successful launches with green markers and unsuccessful launches with red markers at each launch site.

This color coding helps visualize the launch sites with high success rates.

Distances Between a Launch Site and Proximities:

Incorporated colored lines to demonstrate the distance between the launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city. These lines provide visual context regarding the location of the launch site and its surroundings.

Build an Interactive Map with Folium [github]

Build a Dashboard with Plotly Dash

- **Launch Site Dropdown**: Enables users to select either all launch sites or a specific launch site for analysis.
- **Payload Mass Range Slider**: Allows users to define the range of payload masses to consider in the analysis.
- **Pie Chart: Success Rates**. Provides a visual representation of successful and unsuccessful launches as a percentage of the total, based on the selected launch site or all sites.
- **Scatter Chart: Payload Mass vs. Success Rate by Booster Version**

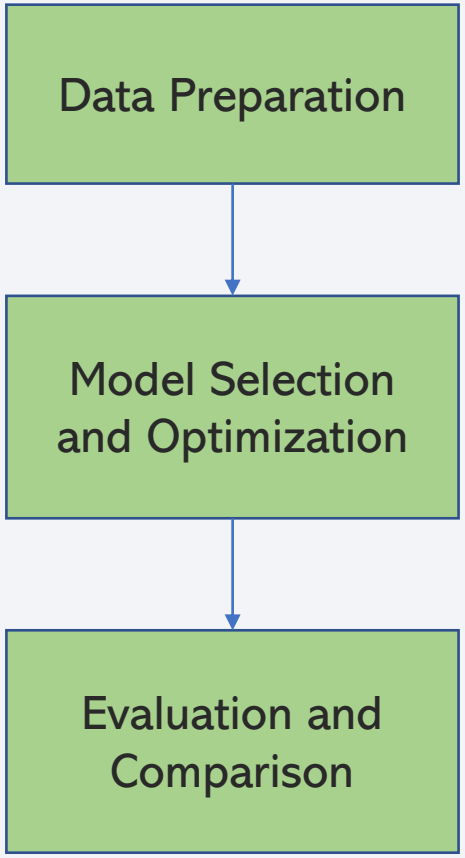
Presents a scatter chart that demonstrates the correlation between payload mass and launch success, categorized by booster version.

Build a Dashboard with Plotly Dash

The dashboard facilitates efficient analysis of the relationship between payload characteristics and launch sites, enabling the identification of optimal launch sites based on payload mass.

Build a Dashboard with Plotly Dash [github]

Predictive Analysis (Classification)



- Convert the "Class" column into a NumPy array.
- Standardize the data using StandardScaler by fitting and transforming the data.
- Split the data into training and testing sets using train_test_split.

- Utilize GridSearchCV for parameter optimization, with a cross-validation value (cv) of 10.
- Apply GridSearchCV on various algorithms: Logistic Regression (LogisticRegression()), Support Vector Machine (SVC()), Decision Tree (DecisionTreeClassifier()), and K-Nearest Neighbors (KNeighborsClassifier()).

- Calculate the accuracy of all models on the test data using the .score() function.
- Assess the confusion matrix for each model.
- Determine the best model based on evaluation metrics: Jaccard Score, F1 Score, and Accuracy.

Predictive Analysis (Classification) [github]

Results

Exploratory Data Analysis:

- Launch success rates have shown improvement over time.
- KSC LC-39A stands out with the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO, and SSO have achieved a perfect success rate of 100%.

Visual Analytics:

Most launch sites are situated near the equator and in close proximity to coastal areas.

The launch sites are strategically located far enough from potentially vulnerable targets such as cities, highways, and railways.

The sites remain accessible for logistical support, facilitating the transportation of personnel and materials.

Results

Predictive Analytics:

Based on the dataset, the Decision Tree model has demonstrated superior predictive capabilities.

Key Findings from Exploratory Data Analysis:

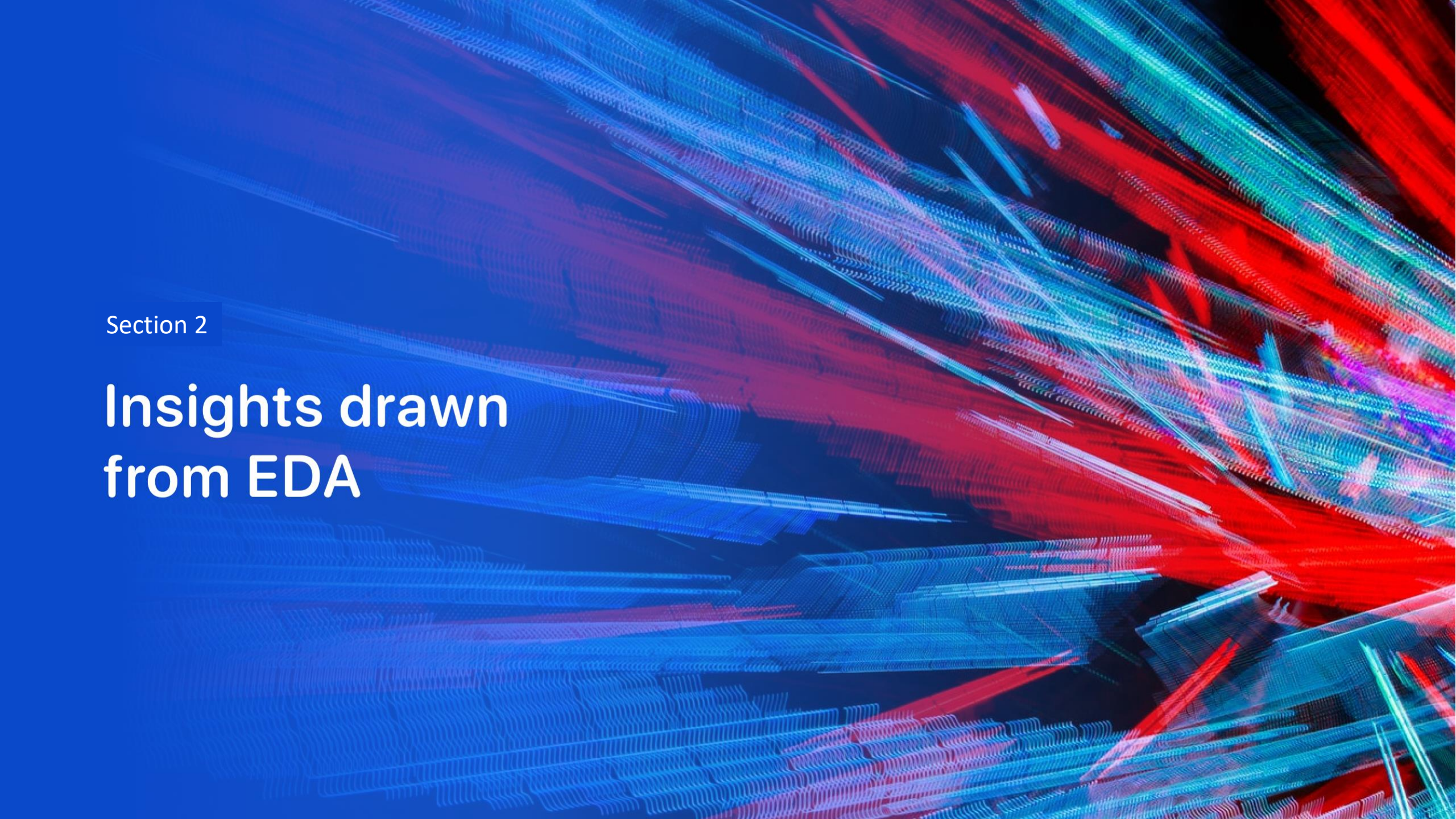
SpaceX utilizes four launch sites.

The average payload for the F9 v1.1 booster is 2,928 kg.

Falcon 9 booster versions exhibited successful landings on drone ships for payloads above the average.

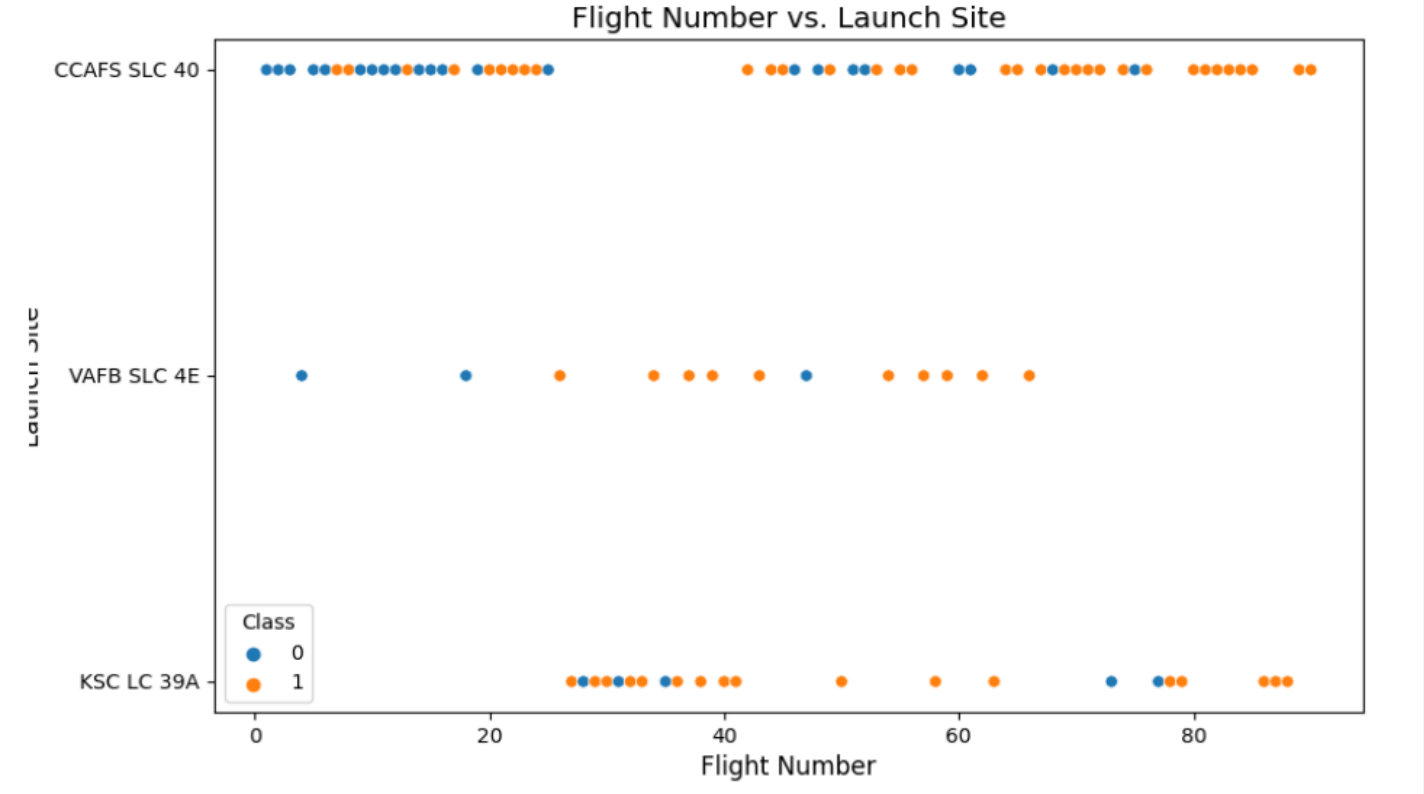
Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failures when attempting landings on drone ships in 2015.

The number of successful landing outcomes has increased progressively each year.


Section 2
Insights drawn from EDA

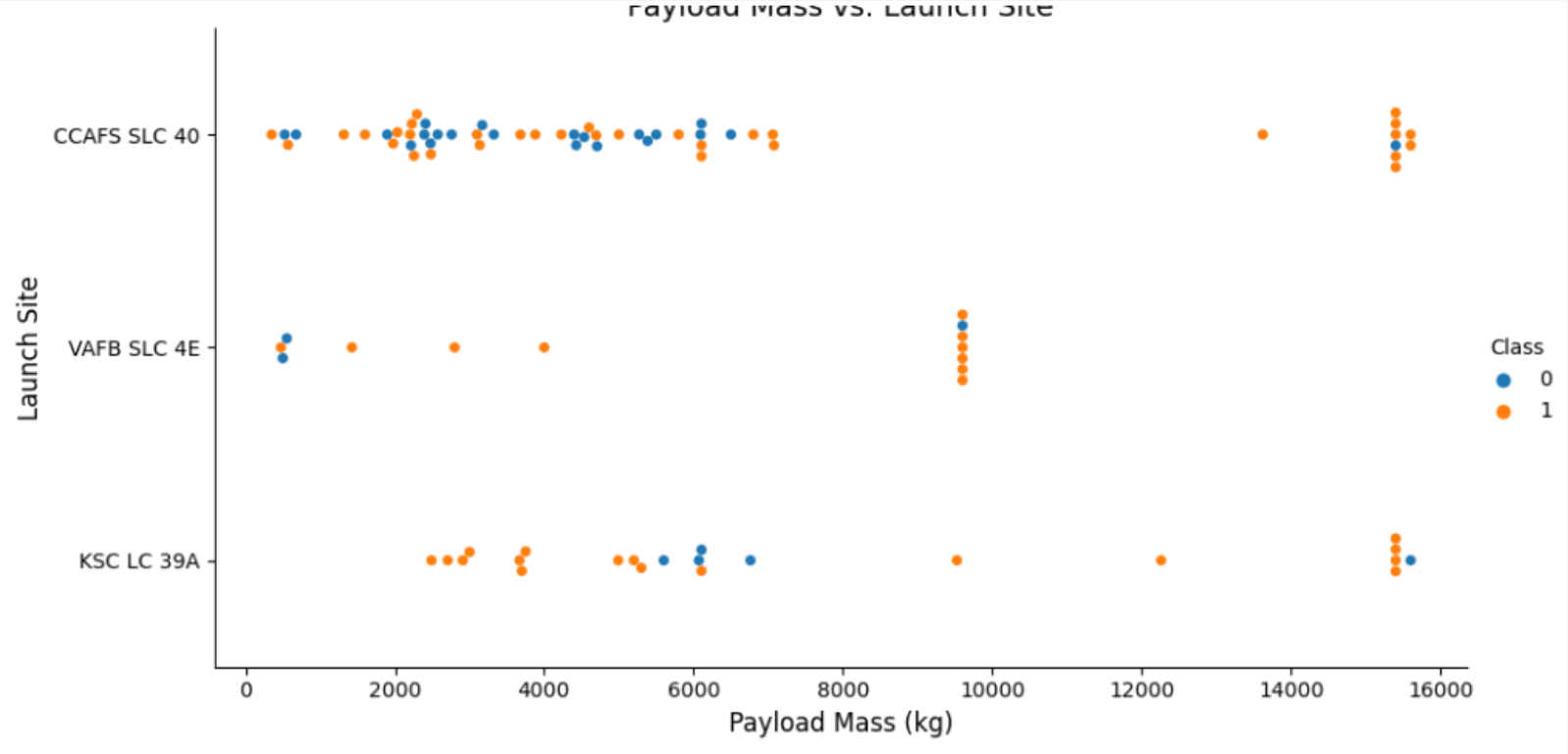
Flight Number vs. Launch Site

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- New launches have a higher success rate



Payload vs. Launch Site

- Higher payloads (over 12,000kg) have a better success rate at CCAFS SLC 40 and KSC LC 39A launch sites
- KSC LC 39A has a 100% success rate for launches < 5,500 kg
- VAFB SKC 4E has not launched anything >10,000 kg



Success Rate vs. Orbit Type

100% Success Rate:

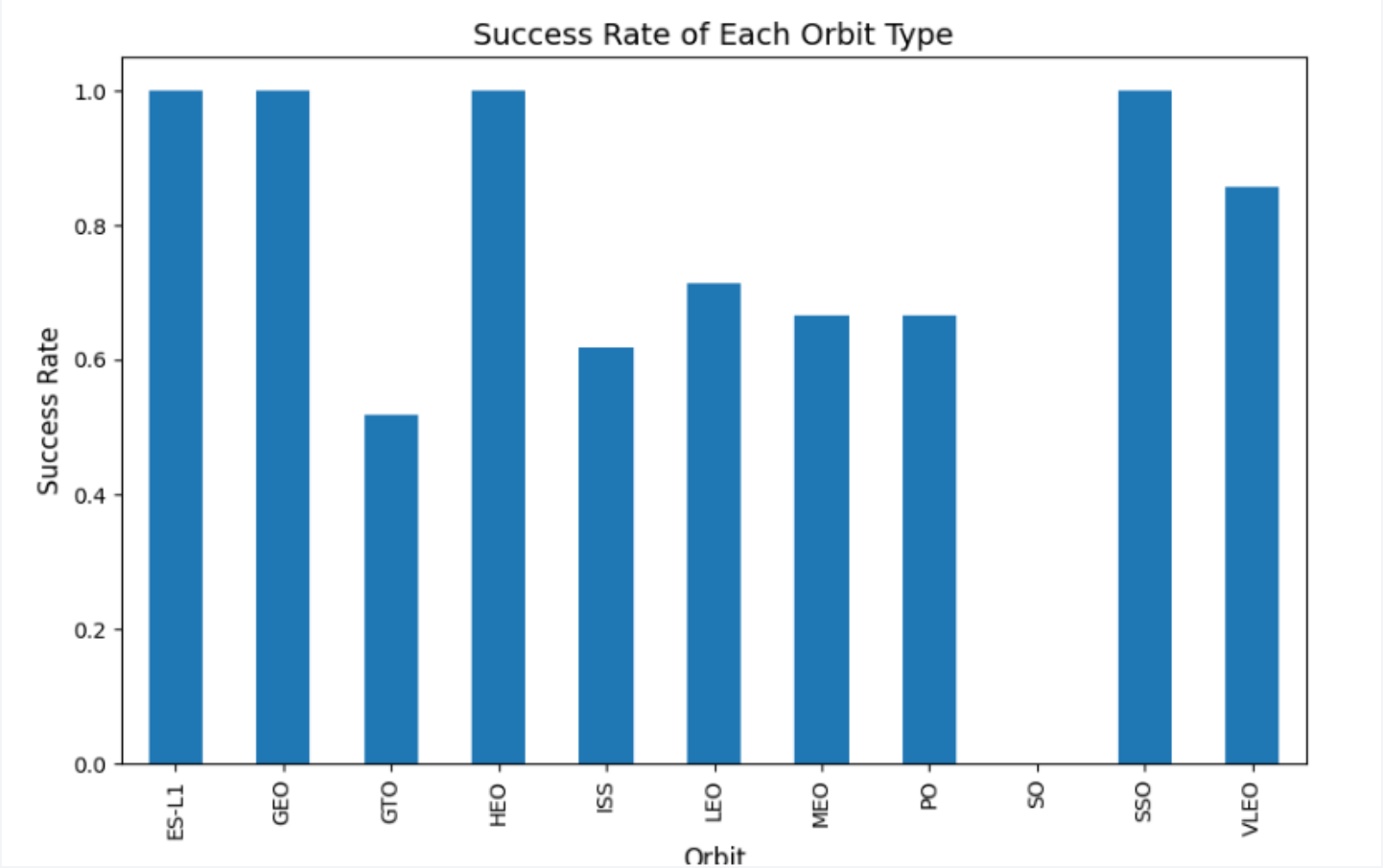
ES-L1, GEO, HEO and SSO

50%-80% Success Rate:

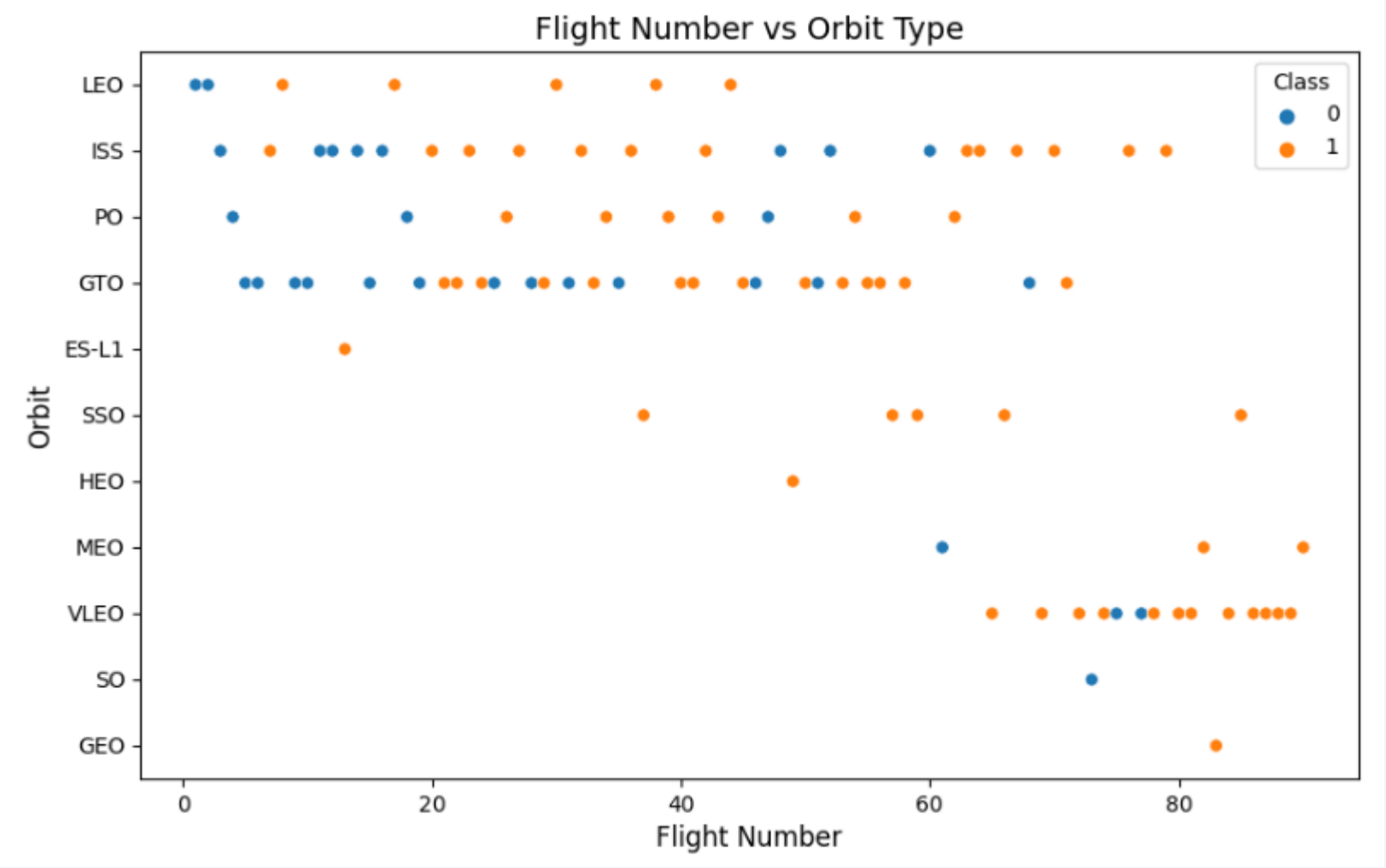
GTO, ISS, LEO, MEO, PO

0% Success Rate:

SO

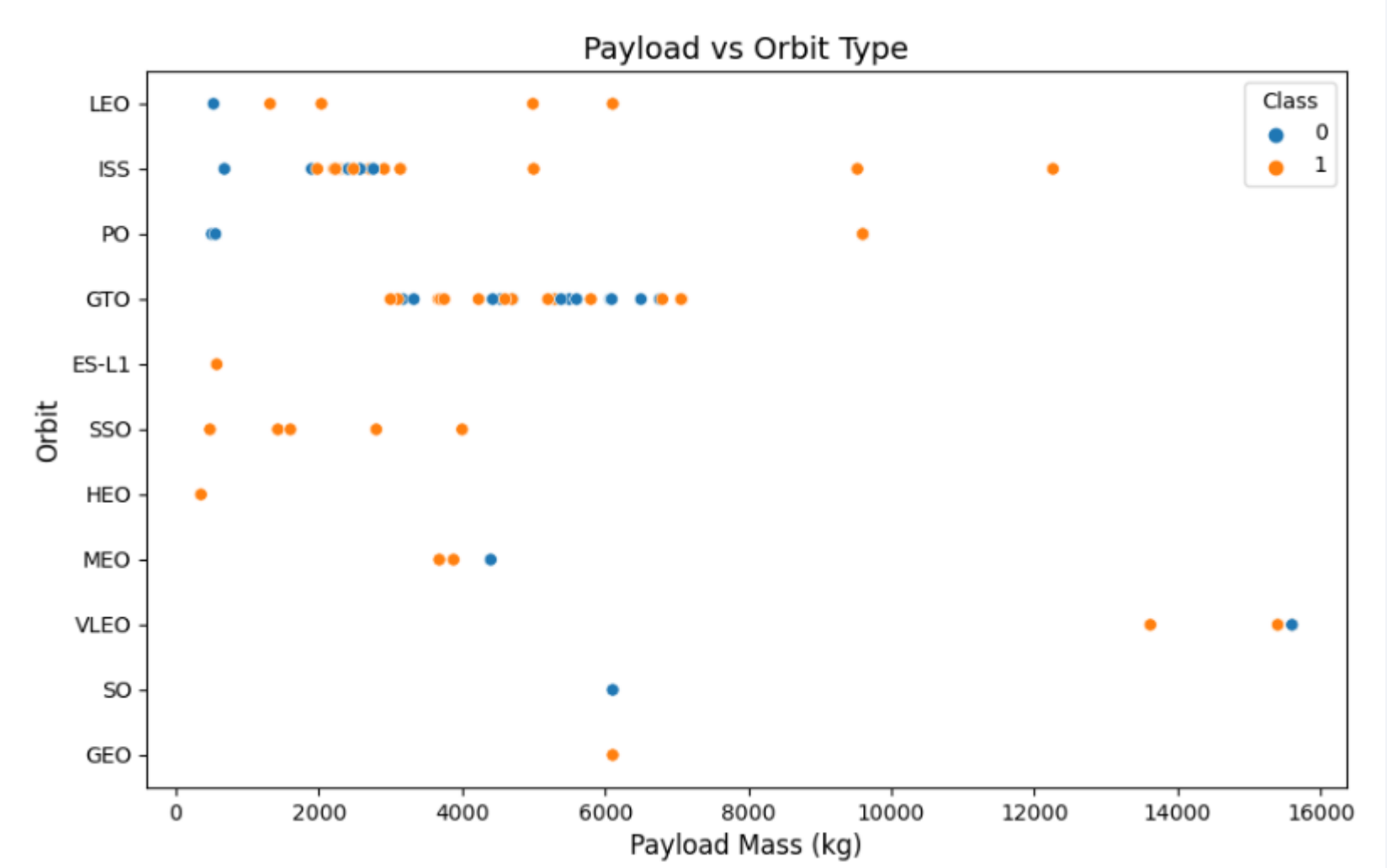


Flight Number vs. Orbit Type



The success rate typically increases with the number of flights for each orbit

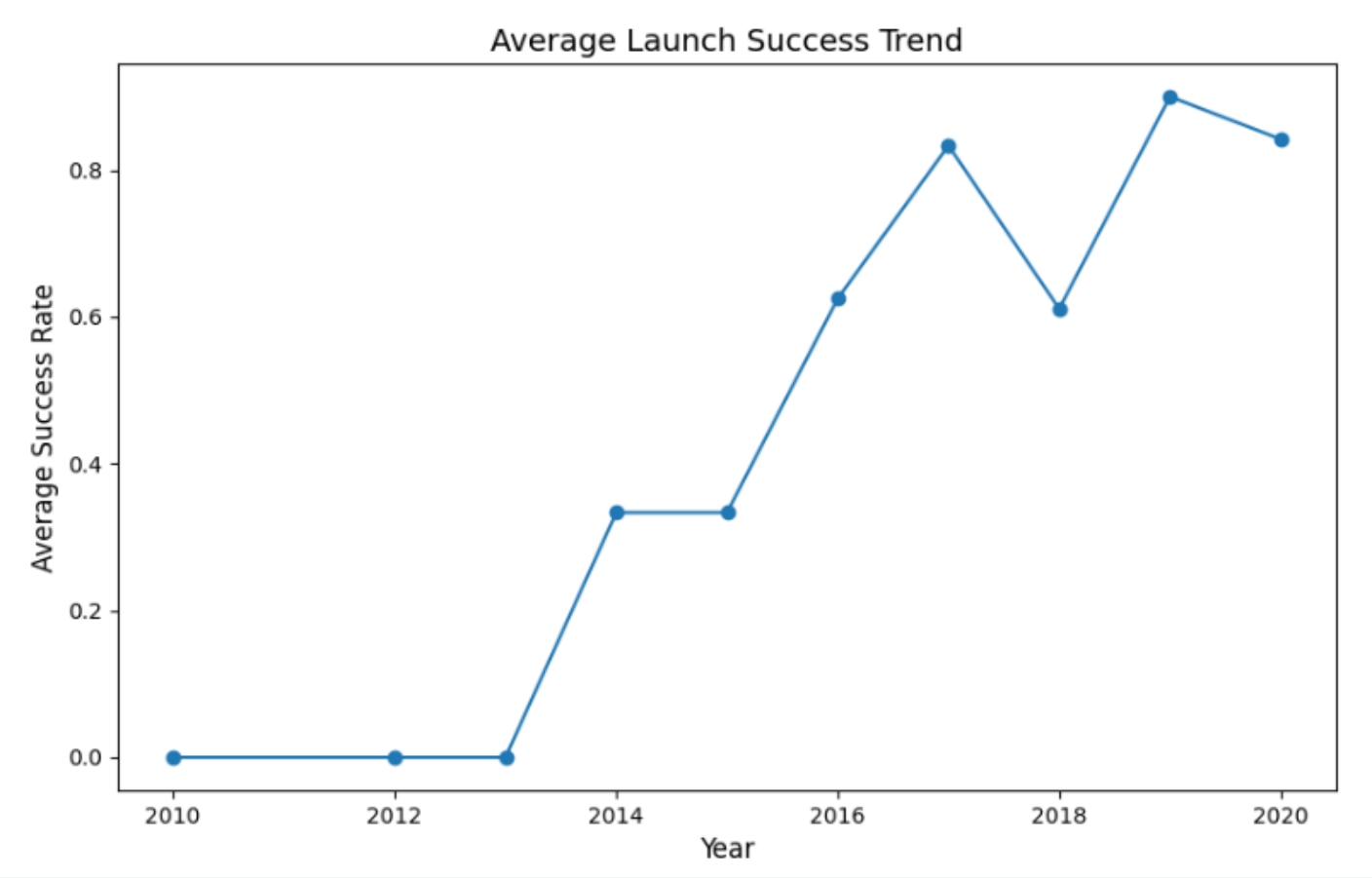
Payload vs. Orbit Type



Heavy payloads are better with LEO, ISS and PO orbits

Launch Success Yearly Trend

Overall, the success rate has improved since 2013



All Launch Site Names

There are 4 launch sites, which can be optained by selecting unique occurrencies of “launch_site” values from the dataset.

Out[5]:

	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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%';
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_(
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (pa
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (pa
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No

We can identify the first outcomes for all CCA launch sites.

Total Payload Mass

45,596 kg total carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

Total_Payload_Mass
45596.0

Average Payload Mass by F9 v1.1

2,928 kg (average) carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

01/08/2018

```
%%sql
SELECT MIN(Date) AS First_Successful_Landing_Date
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_Successful_Landing_Date
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster mass greater than 4,000 but less than 6,000: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%%sql

SELECT Mission_Outcome, COUNT(*) AS Count
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

Mission_Outcome	Count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- 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

```
%%sql

SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL
);
```

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

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

Records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

```
%%sql
SELECT SUBSTR(Date, 4, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE SUBSTR(Date, 7, 4) = '2015'
AND Landing_Outcome LIKE '%Failure (drone ship)%';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	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

Count of successful landing outcomes between the date 04-06-2010 and 20-03-2017

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Successful_Landings
FROM SPACEXTBL
WHERE Date BETWEEN '04-06-2010' AND '20-03-2017'
AND Landing_Outcome = 'Success'
GROUP BY Landing_Outcome
ORDER BY Successful_Landings DESC;
```

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

Landing_Outcome	Successful_Landings
Success	20

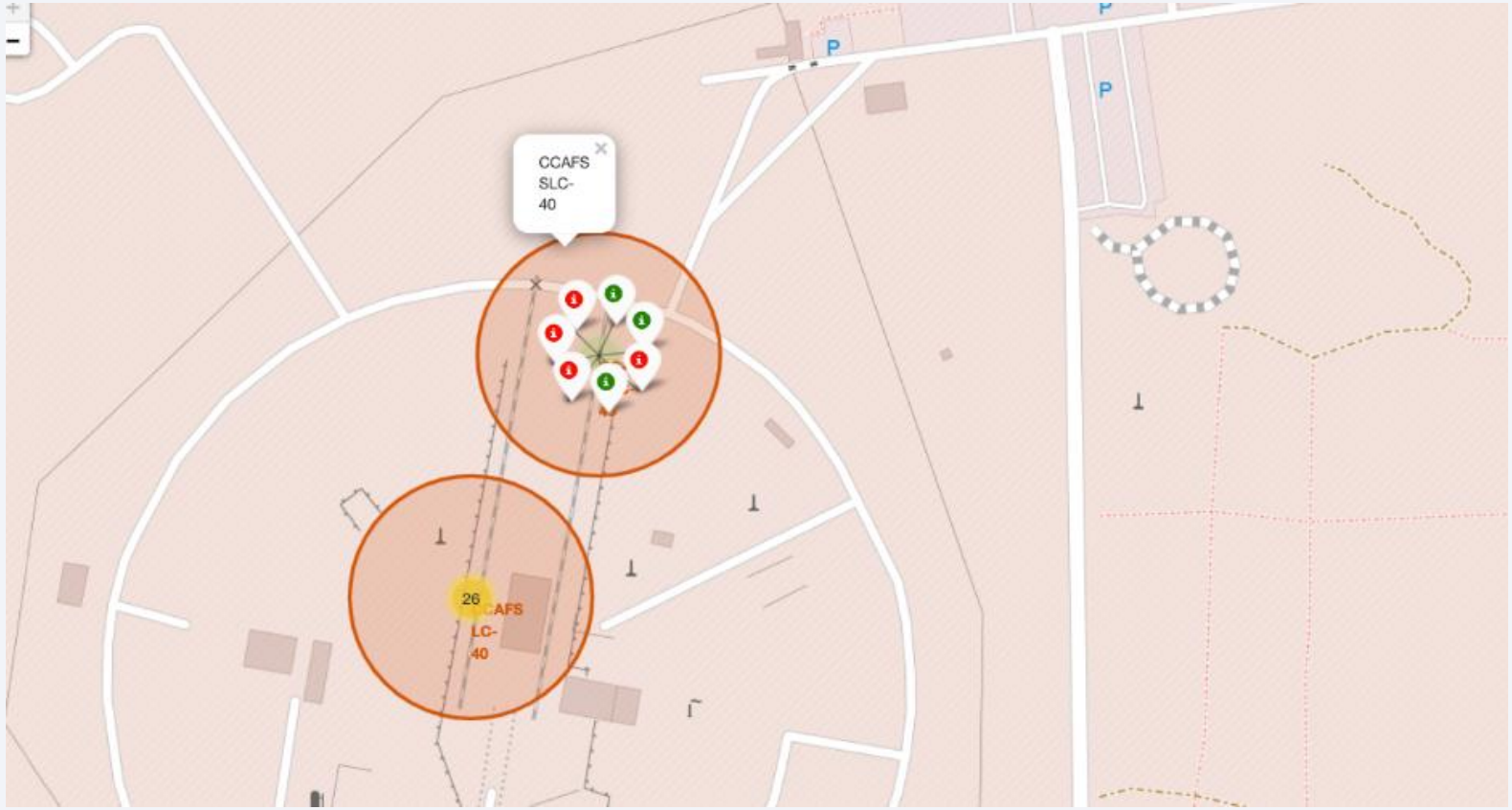
Section 3

Launch Sites Proximities Analysis

Launch Sites

Proximity to the equator benefits launches to equatorial orbit by leveraging Earth's rotation for a prograde trajectory. Launching from near the equator provides a natural advantage, reducing the need for additional fuel and boosters.

Launch Outcomes



Example of CCAFS SLC-40 which has a 3/7 success rate (42.9%)

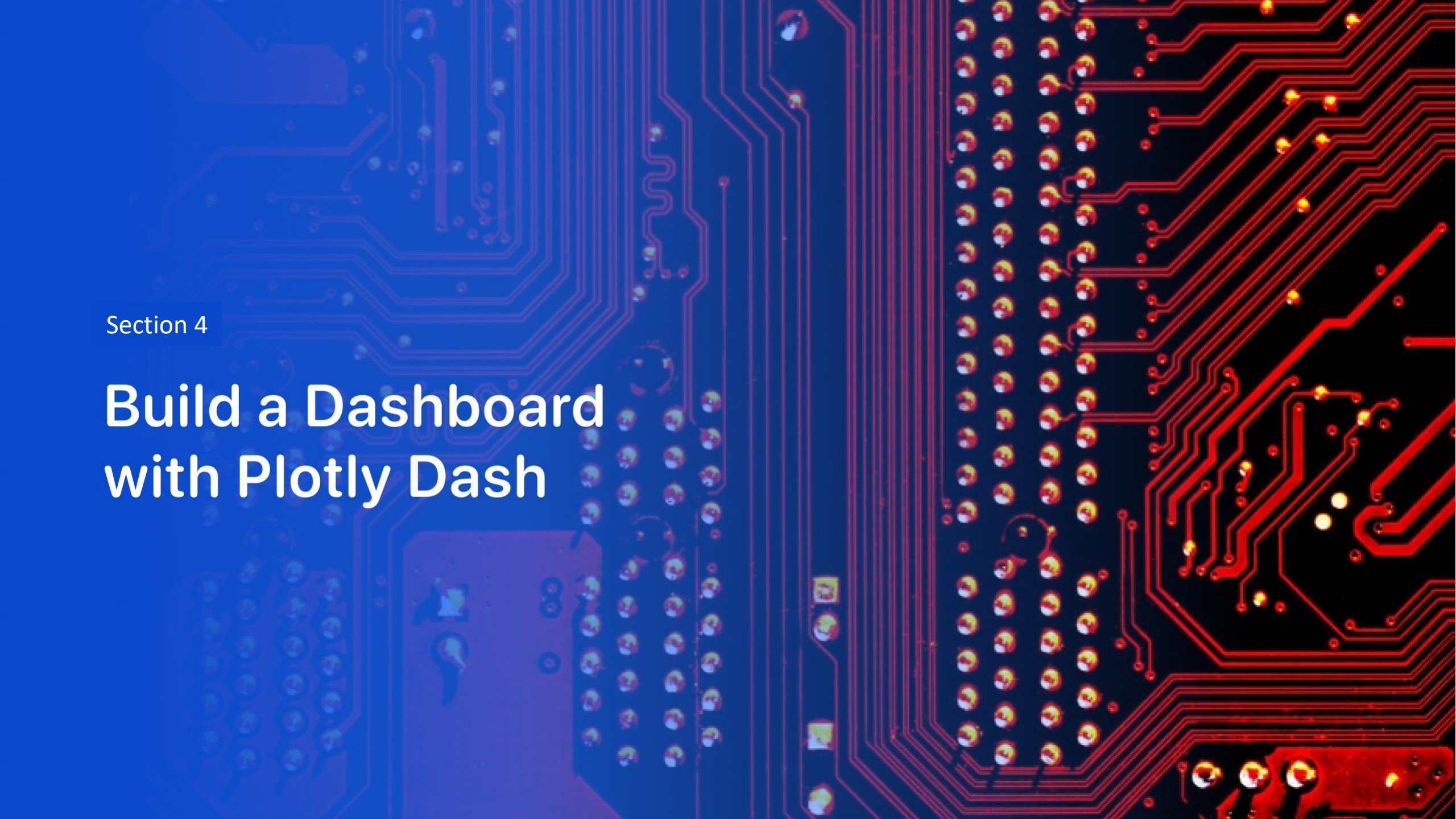
Green markers for successful launches

Red markers for unsuccessful launches

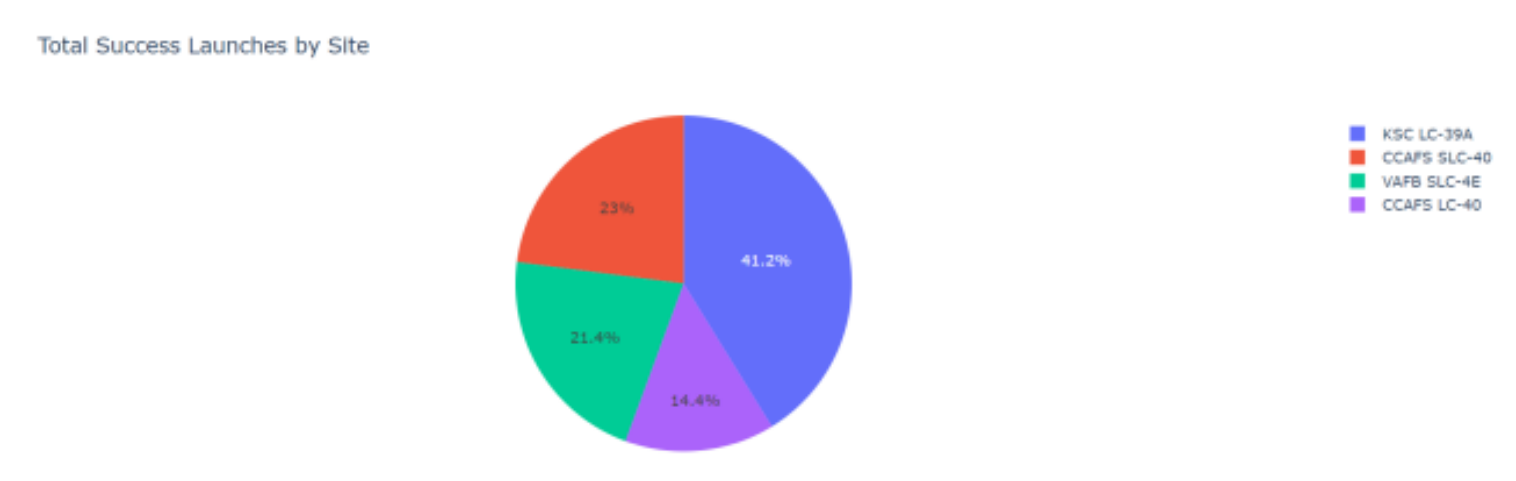
Distance to Proximities

CCAFS SLC-40

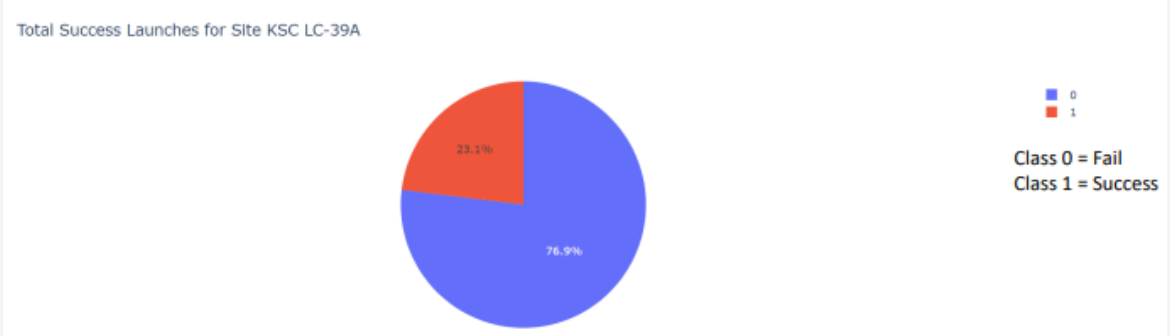
- City Distance 23.234752126023245
- Railway Distance 21.961465676043673
- Highway Distance 26.880385696814 92
- Coastline Distance 0.8627671182499878


Section 4
Build a Dashboard with Plotly Dash

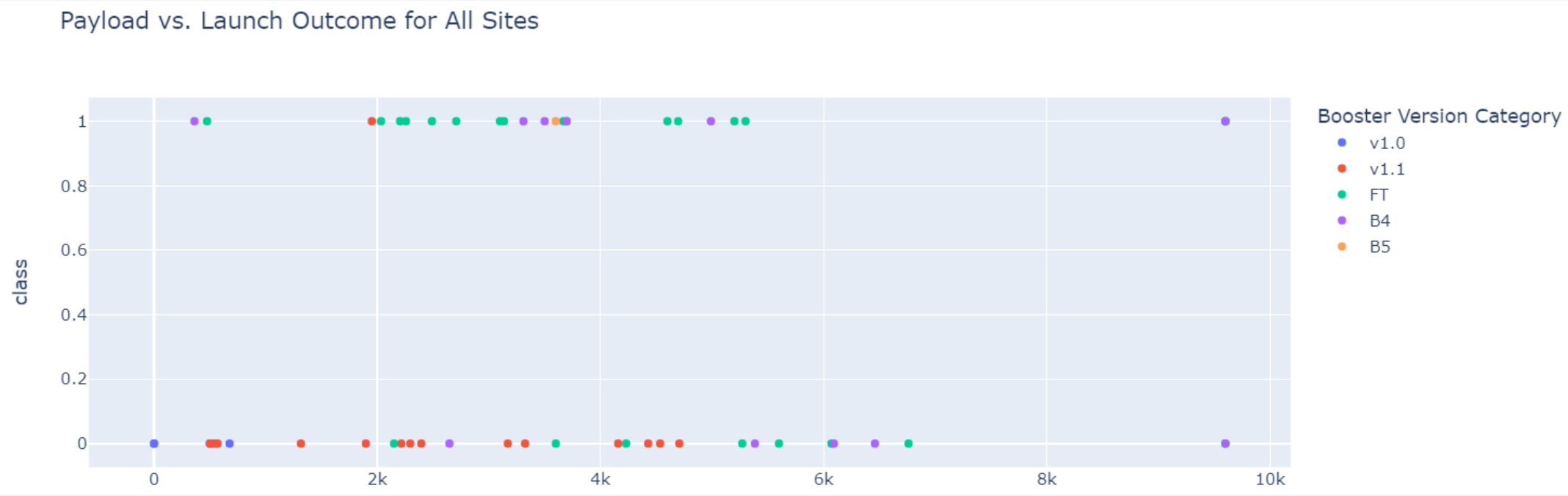
Launch Success by Site



KSC LC-39A is the most successful launch site (41.2%) with the highest success rate score.



Payload vs Launch Outcome



Payloads between 2,000 kg and 5,000 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
jaccard_scores = [
                  jaccard_score(Y, logreg_cv.predict(X), average='binary'),
                  jaccard_score(Y, svm_cv.predict(X), average='binary'),
                  jaccard_score(Y, tree_cv.predict(X), average='binary'),
                  jaccard_score(Y, knn_cv.predict(X), average='binary'),
                 ]

f1_scores = [
             f1_score(Y, logreg_cv.predict(X), average='binary'),
             f1_score(Y, svm_cv.predict(X), average='binary'),
             f1_score(Y, tree_cv.predict(X), average='binary'),
             f1_score(Y, knn_cv.predict(X), average='binary'),
            ]

accuracy = [logreg_cv.score(X, Y), svm_cv.score(X, Y), tree_cv.score(X, Y), knn_cv.score(X, Y)]

scores = pd.DataFrame(np.array([jaccard_scores, f1_scores, accuracy]),
                      index=['Jaccard_Score', 'F1_Score', 'Accuracy'],
                      columns=['LogReg', 'SVM', 'Tree', 'KNN'])
scores
```

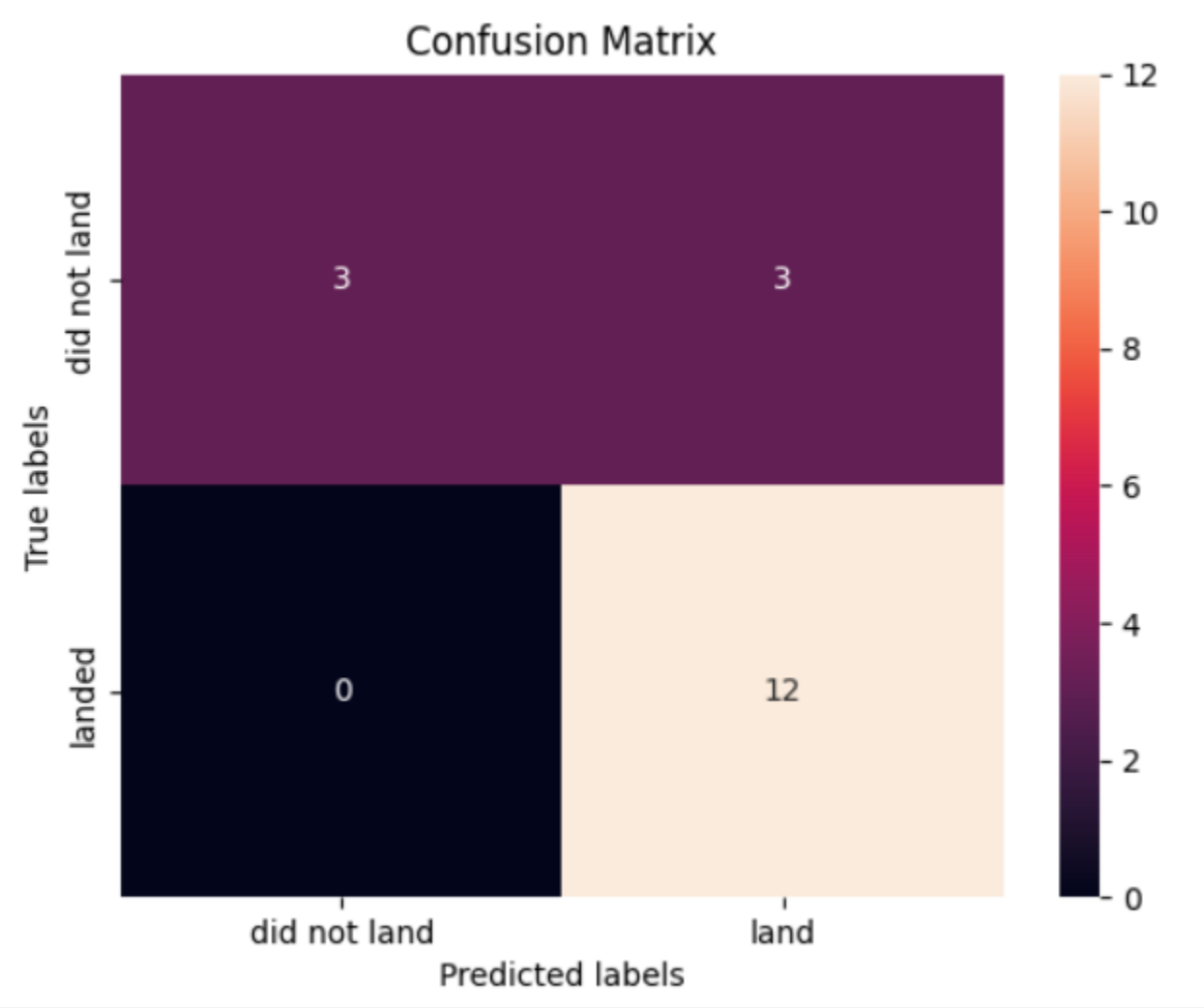
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.805556	0.819444
F1_Score	0.909091	0.916031	0.892308	0.900763
Accuracy	0.866667	0.877778	0.844444	0.855556

	Logistic Regression	SVM	Decision Tree	KNN
Jaccard Score	0.800000	0.800000	0.800000	0.800000
F1 Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

```
jaccard_scores = [
                  jaccard_score(Y, logreg_cv.predict(X), average='binary'),
                  jaccard_score(Y, svm_cv.predict(X), average='binary'),
                  jaccard_score(Y, tree_cv.predict(X), average='binary'),
                  jaccard_score(Y, knn_cv.predict(X), average='binary'),
                 ]

f1_scores = [
             f1_score(Y, logreg_cv.predict(X), average='binary'),
             f1_score(Y, svm_cv.predict(X), average='binary'),
             f1_score(Y, tree_cv.predict(X), average='binary'),
             f1_score(Y, knn_cv.predict(X), average='binary'),
            ]

accuracy = [logreg_cv.score(X, Y), svm_cv.score(X, Y), tree_cv.score(X, Y), knn_cv.score(X, Y)]

scores = pd.DataFrame(np.array([jaccard_scores, f1_scores, accuracy]),
                      index=['Jaccard_Score', 'F1_Score', 'Accuracy'],
                      columns=['LogReg', 'SVM', 'Tree', 'KNN'])
scores
```

- The accuracy is extremely close, this is likely due to the small dataset.

Confusion Matrix



All the confusion matrices were identical

There are 3 false positive, which is not a good news.

Conclusions

Model Performance: the models exhibited similar performance on the test set, with the decision tree model slightly outperforming the others. It can be employed to predict successful landings, ultimately contributing to increased profits.

Key Findings:

- Equator Proximity: Most launch sites are strategically located near the equator, leveraging the rotational speed of the Earth to save on fuel and boosters.
- Coastal Proximity: All launch sites are situated in close proximity to coastlines.
- Launch Success: The overall launch success rate has shown a positive trend over time.
- KSC LC-39A: This launch site demonstrated the highest success rate among all sites, particularly for launches with payloads weighing less than 5,500 kg.
- Orbit Success: Orbits including ES-L1, GEO, HEO, and SSO achieved a 100% success rate.
- Payload Mass: Across all launch sites, there is a positive correlation between higher payload mass (kg) and higher success rates.

Conclusions and additional insights

Additional Insights:

Lower weighted payloads tend to exhibit better success rates compared to heavier payloads.

The success rates of SpaceX launches increase proportionally with the passage of time, indicating a progressive improvement in launch performance.

The Decision Tree Classifier can be employed to predict successful landings, ultimately contributing to increased profits.

A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set

Thank you!