Winning Space Race with Data Science

Amaan Ahmed
November 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

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

Section 1

Methodology

Methodology

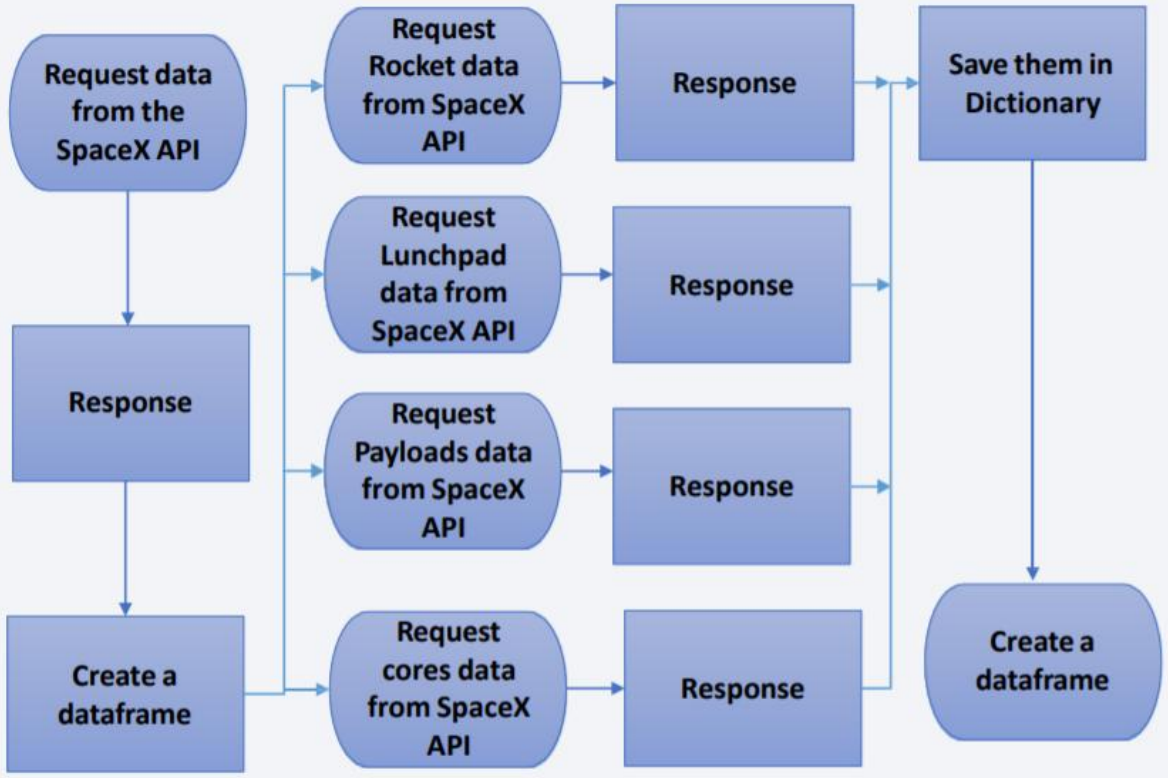
Executive Summary

- Data collection methodology:
  - SpaceX Launch data was collected from the SpaceX REST API using Python Web Scraping
- Perform data wrangling
  - Cleaned SpaceX data: removed errors, standardized types, coded landing with 0 or 1
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built classification models: Logistic Regression, SVM, Decision Tree, and KNN
  - Hyperparameters were tuned by GridSearchCV
  - Evaluated using confusion matrix, accuracy, precision, recall

Data Collection

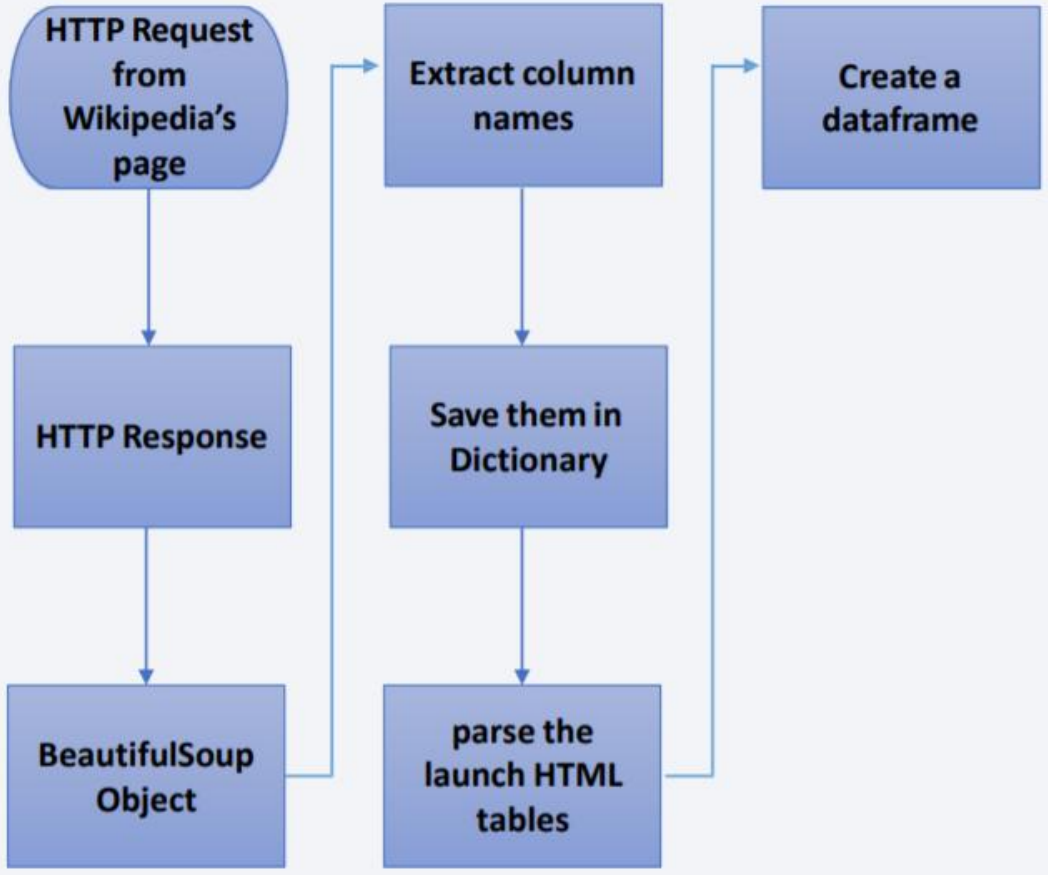
- Data Source: Got historical launch data from SpaceX API and a static JSON file
- Key Features: Pulled out rocket, payloads, launchpad, cores, flight number, and date
- API Details: Used the IDs in the data to get extra info like booster name, payload mass, orbit, launch site, and core landing outcomes
- Cleaning Data: Kept only Falcon 9 launches with one core and one payload and converted dates to a proper format
- Handling Missing Values: Filled missing PayloadMass with the average and kept None for unused landing pads

Data Collection – SpaceX API



https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/Data%20Collection/data_collection_api.ipynb

Data Collection - Scraping



https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/Data%20Collection/data_collection_webscraping.ipynb

Data Wrangling

- Data Cleaning: Removed unnecessary columns and handled missing values
- Data Standardization: Standardized dates, payload masses, and other data types
- Landing Outcome Encoding: Converted landings to binary (1 = success, 0 = failure/no attempt)
- Prepared for Analysis: Dataset ready for visualization, SQL queries, and predictive modeling
- https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/Data%20Collection/data_wrangling.ipynb

EDA with Data Visualization

- Launch Sequence vs Outcome: Scatter plots of FlightNumber vs PayloadMass and LaunchSite to see how success rates changed with more launches or different sites
- Orbit and Launch Factors: Plots of Orbit vs FlightNumber and PayloadMass to examine how orbital type affected landing success
- Success by Site and Payload: Scatter plots showing landing outcomes across LaunchSites and varying PayloadMass
- Trend Over Time: Bar plot of average yearly success rates to show improvement in landing reliability
- https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/EDA/eda_visualization.ipynb

EDA with SQL

- Filter & Explore Payloads: Queried launches within specific payload ranges and identified the heaviest payloads for each landing outcome
- Landing Success Counts: Counted successful landings by launch site, orbit type, and specific time periods (month or quarter)
- Compare Payloads by Outcome: Calculated average payload mass for different landing outcomes to see how success related to payload size
- Success Rates by Orbit: Determined the number and percentage of successful missions for each orbit type
- https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/EDA/eda_sql.ipynb

Build an Interactive Map with Folium

- Base Map: Created an interactive Folium map centered on a starting launch site to serve as the base for all markers and overlays
- Launch Site Markers: Added markers for all four SpaceX launch sites with pop-ups showing their names
- Success/Failure Visualization: Used colored circle markers (green for success, red for failure) to show landing outcomes at each site
- Proximity Analysis: Added lines and markers to nearby roads, railways, and coastlines to explore geographic factors affecting site selection
- https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/Data%20Visualization/folium_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

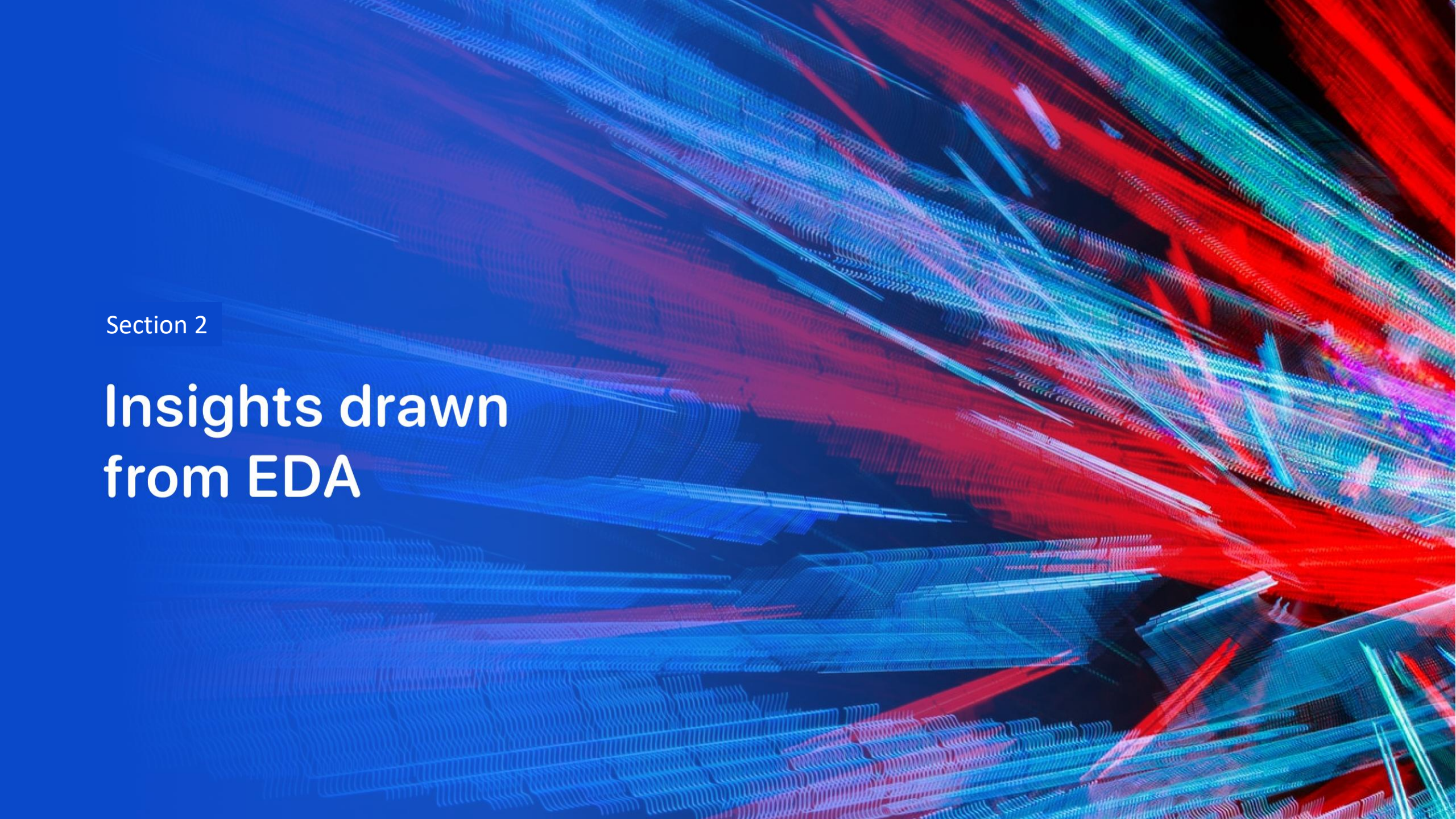
- Launch Site Dropdown: Lets users select a specific site or all sites to filter data for the charts
- Success Pie Chart: Shows the proportion of successful vs. failed launches for the selected site(s)
- Payload Range Slider: Allows filtering the scatter plot by minimum and maximum payload mass
- Payload vs Outcome Scatter Plot: Displays success/failure versus payload, colored by booster type, showing correlations and patterns
- https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/Data%20Visualization/dash.py

Predictive Analysis (Classification)

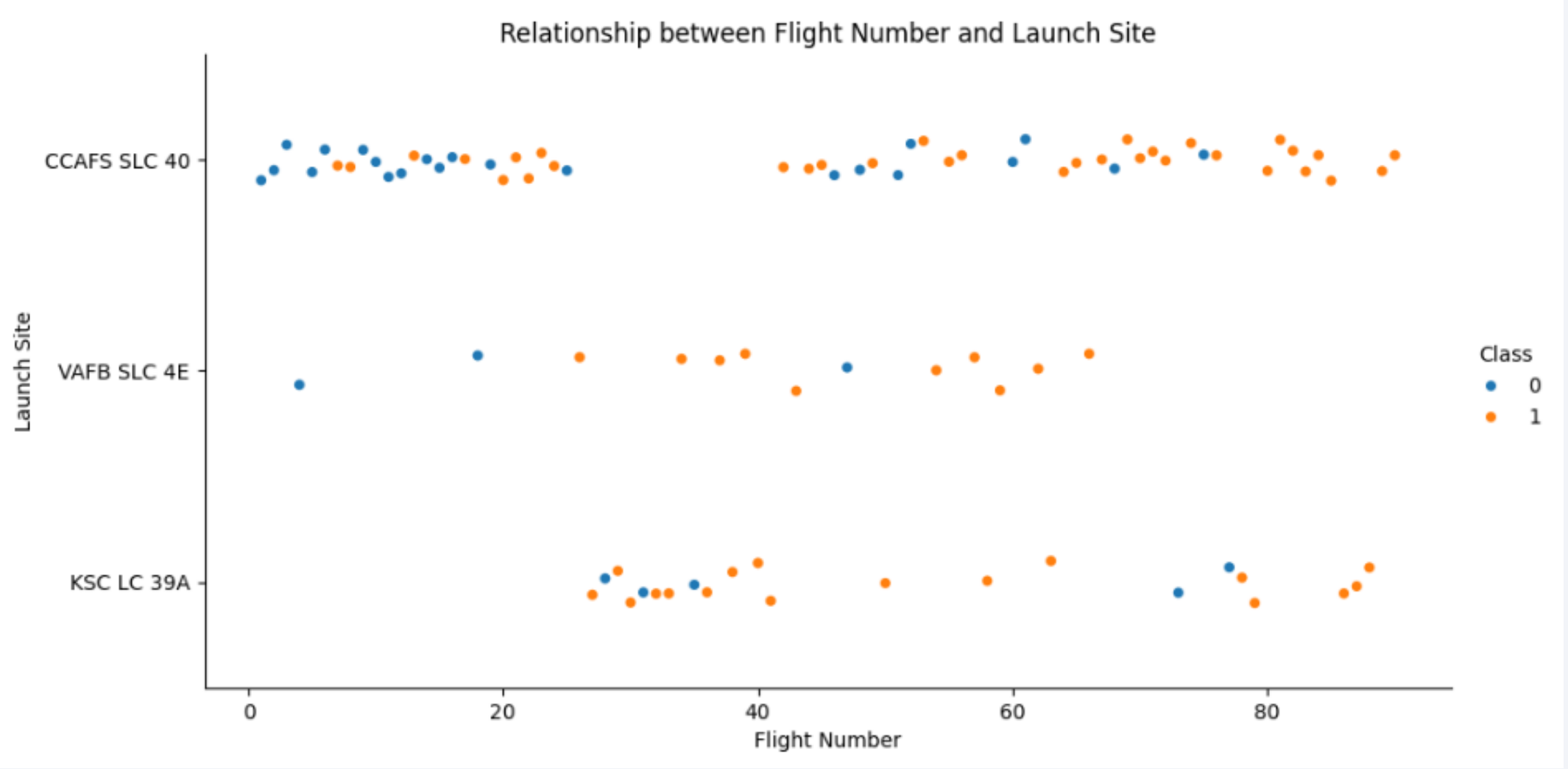
- Data Preparation: Features were one-hot encoded and standardized; dataset split into training and testing sets
- Model Training & Tuning: Four classifiers (Logistic Regression, SVM, Decision Tree, KNN) were tuned using GridSearchCV with 10-fold cross-validation
- Evaluation: Final models were tested on unseen data; accuracy scores and confusion matrices measured performance
- Best Model Selection: Decision Tree achieved the highest test accuracy and was chosen as the top-performing model
- https://github.com/amaanh2/Predicting-SpaceX-Falcon9-Landing-Outcomes/blob/main/ML%20Predictions/ml_predictions.ipynb

Results

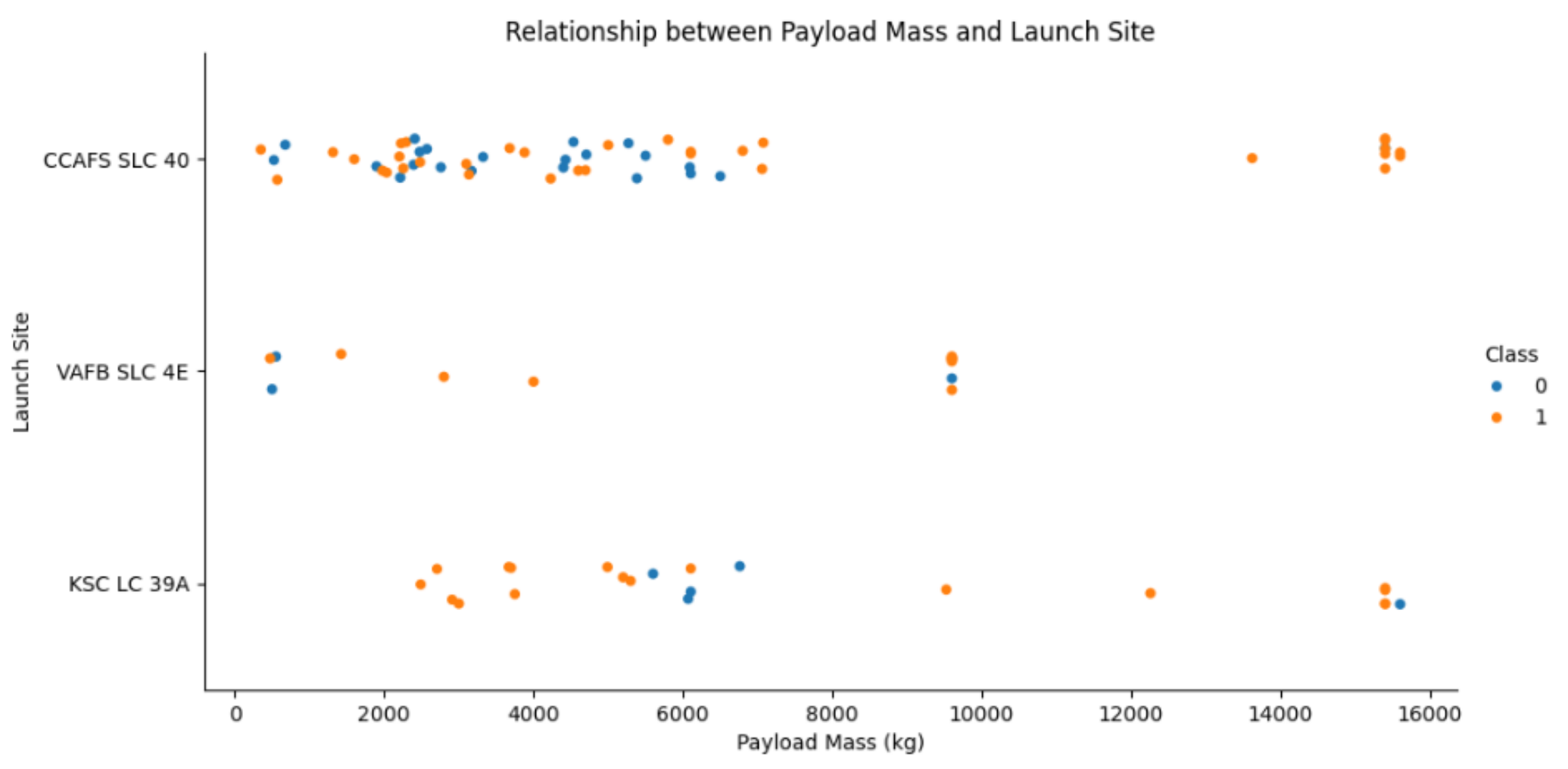
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

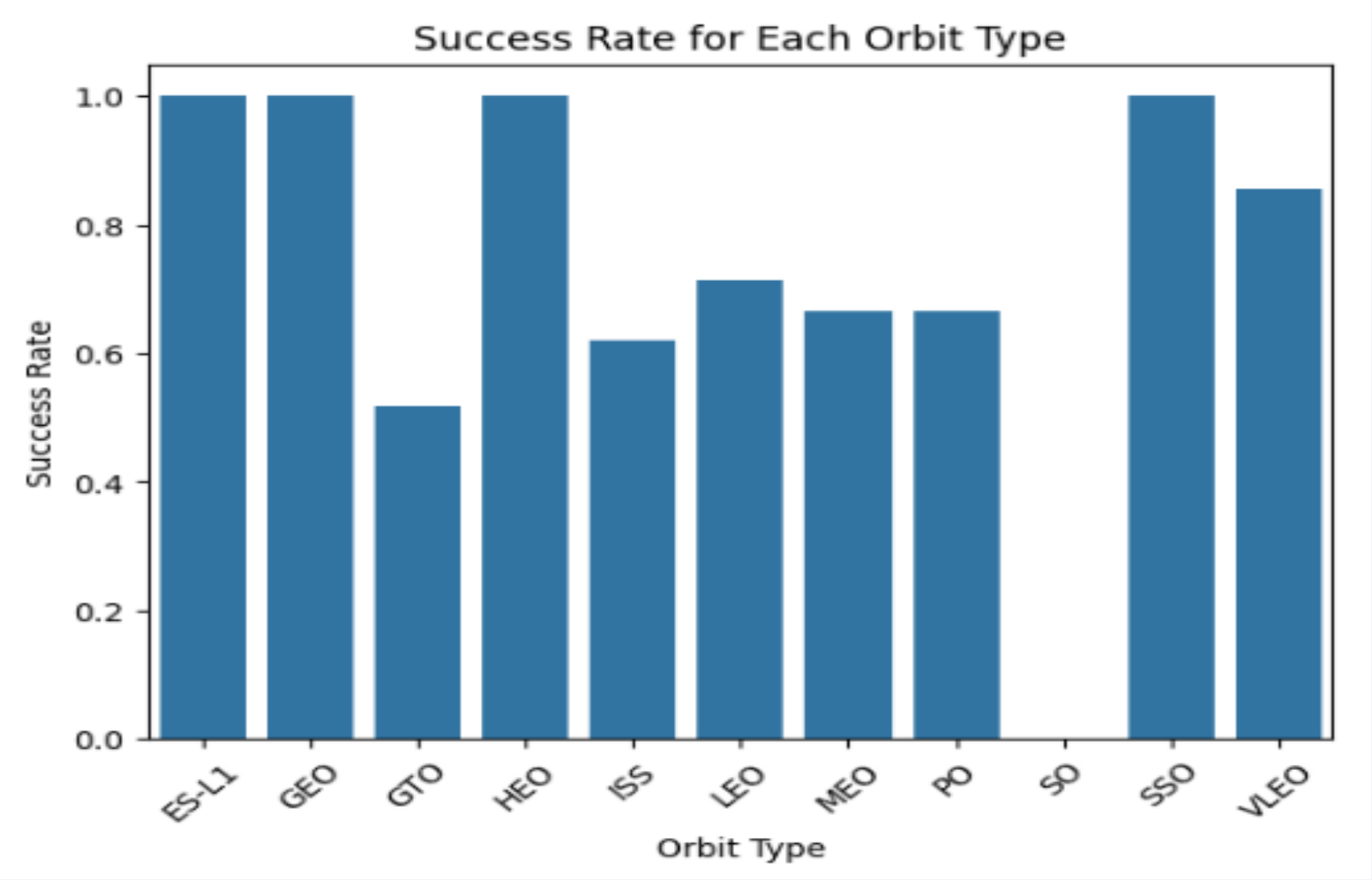
Flight Number vs. Launch Site



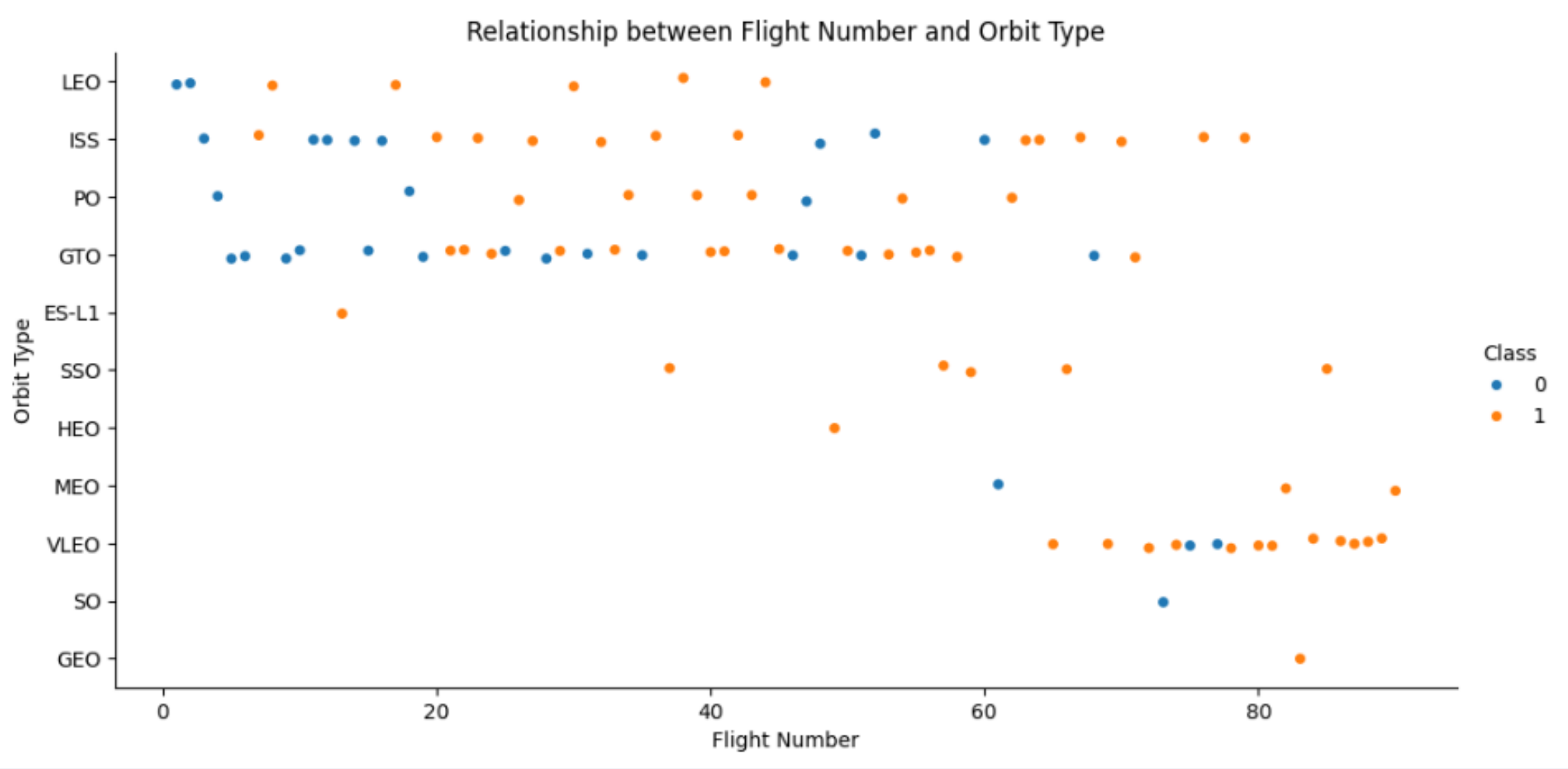
Payload vs. Launch Site



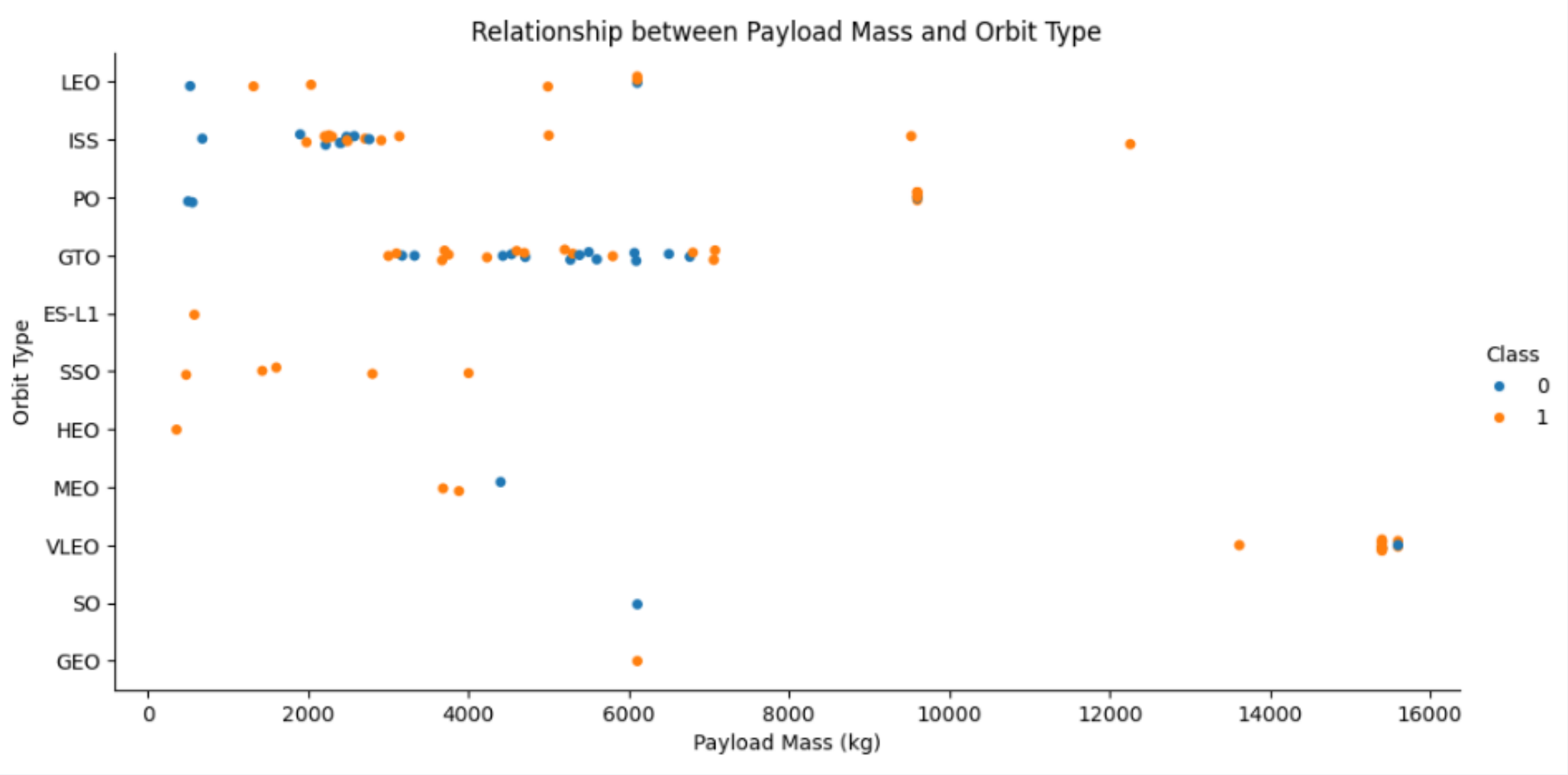
Success Rate vs. Orbit Type



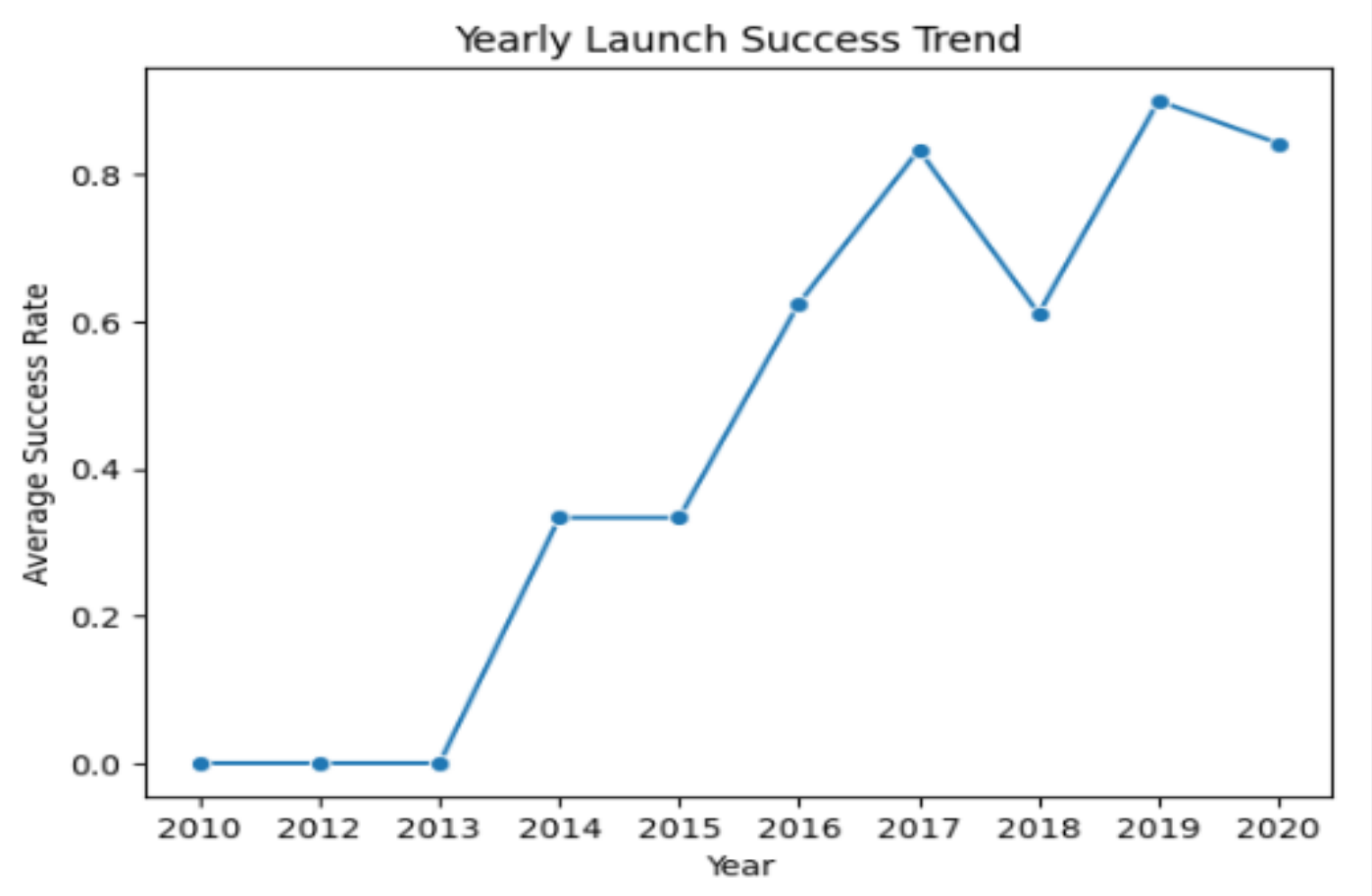
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
%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'

```
%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

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER == "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version == "F9 v1.1"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE Mission_Outcome == "Success"
```

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

MIN(DATE)
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE 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

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE Payload_Mass__kg_ = (
    SELECT MAX(Payload_Mass__kg_)
    FROM SPACEXTABLE
);
```

```
 * 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

```
%%sql SELECT
    substr(Date, 6, 2) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE
    Landing_Outcome LIKE '%Failure (drone ship)%'
    AND substr(Date, 1, 4) = '2015';
```

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

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

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

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

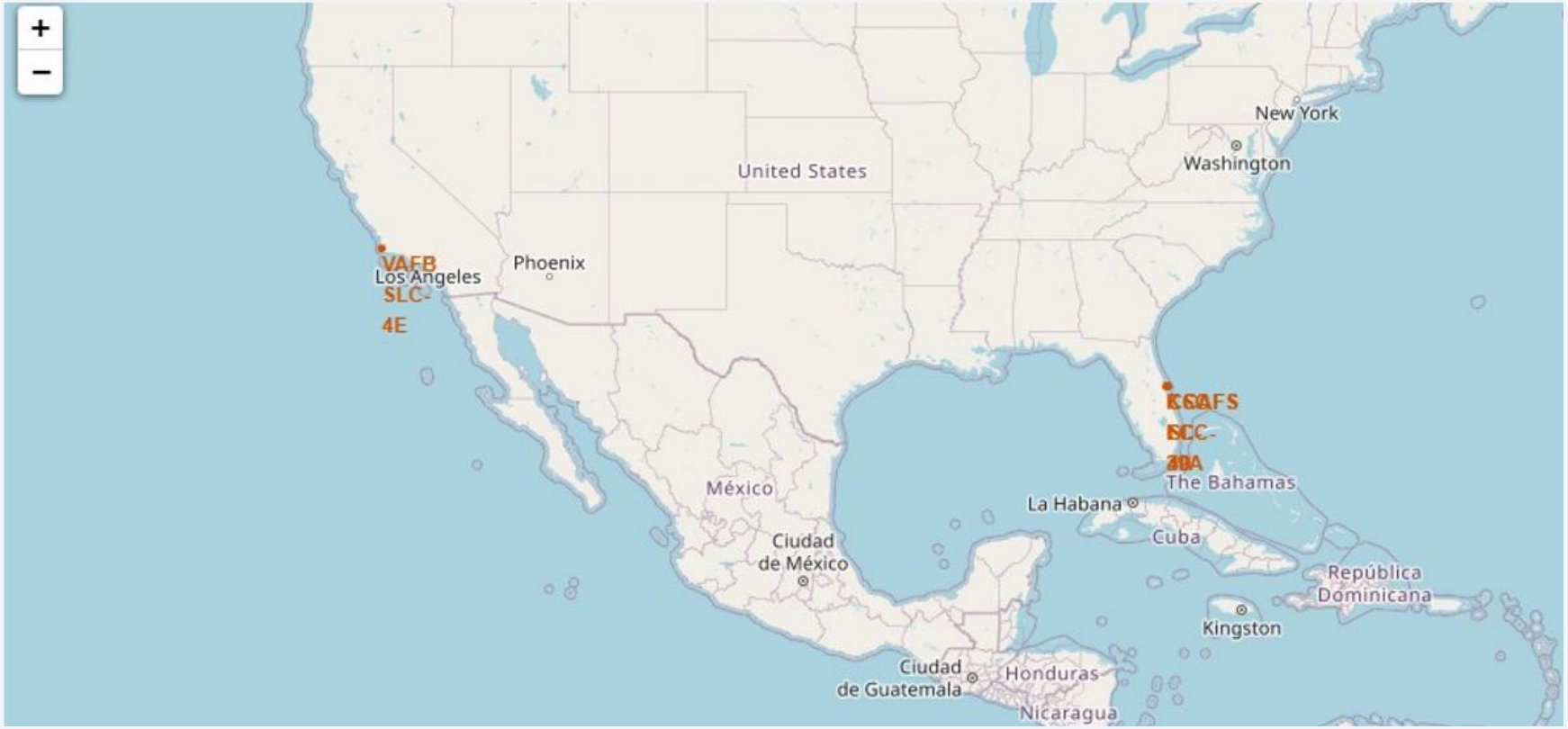
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

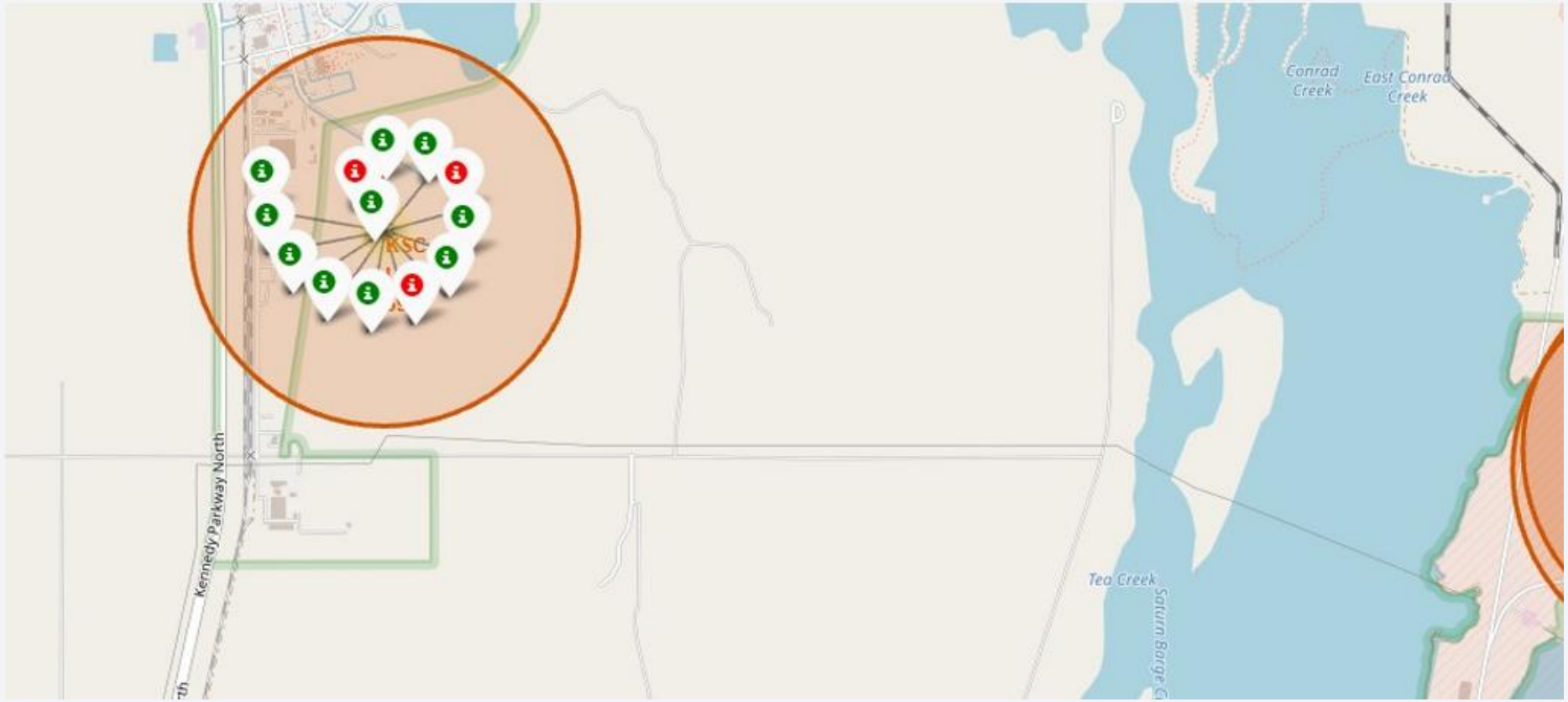
Launch Sites Marking

The markers on this maps show the launch site locations on the map.



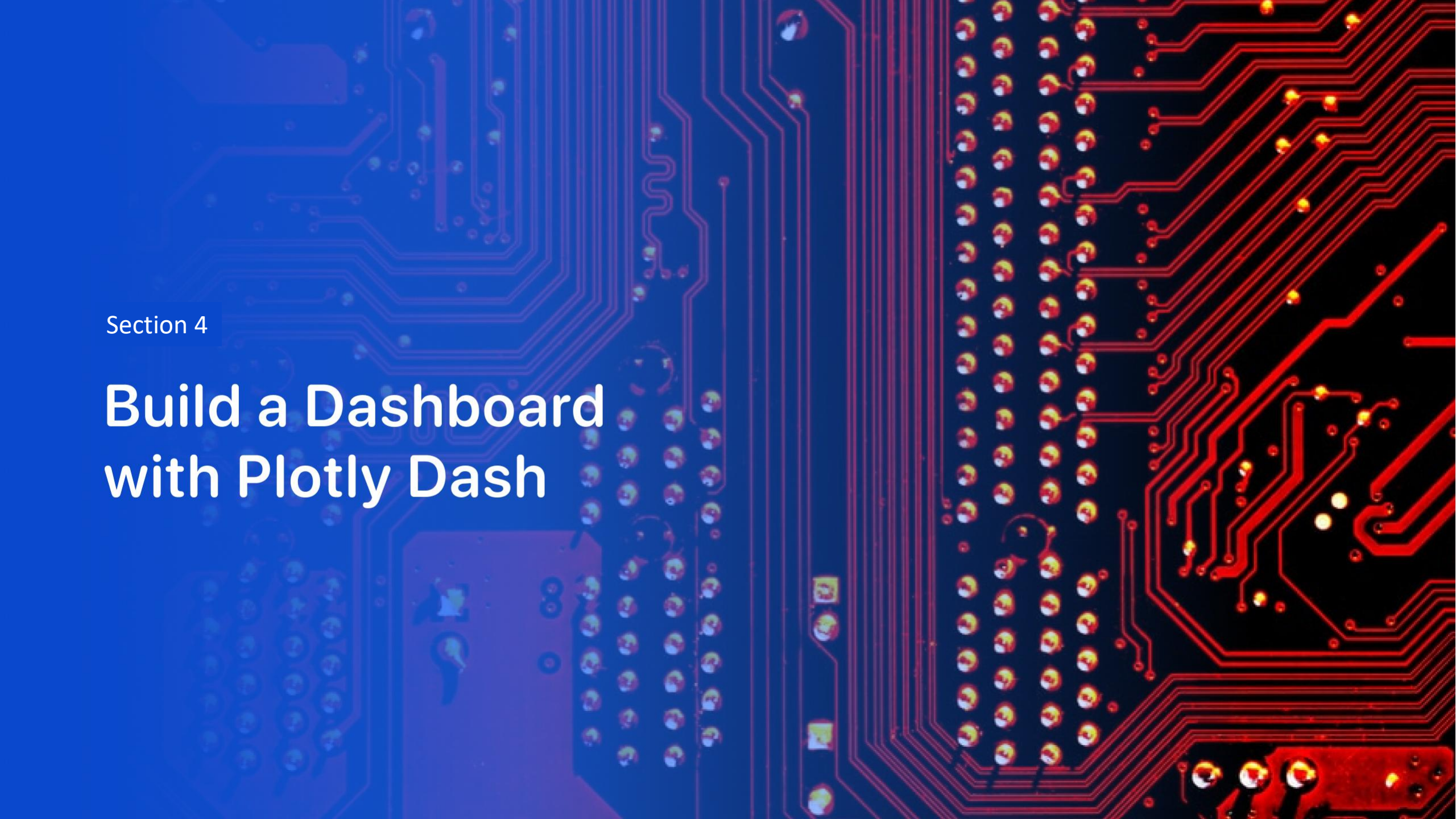
Marking Successful/Failed Launches

A green marker represents a successful landing outcome, while a red one represents failure.



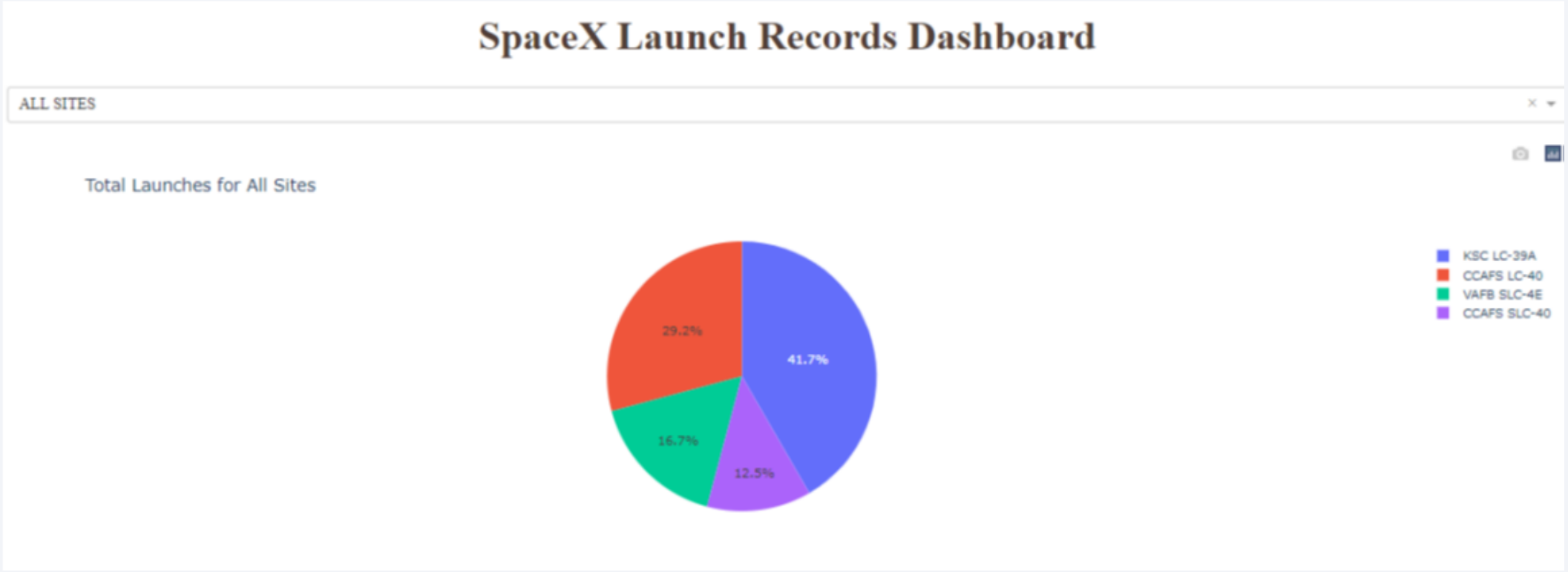
Launch Site Proximities

The blue line represents the distance between the lunch site and the closest coastline.


Section 4
Build a Dashboard with Plotly Dash

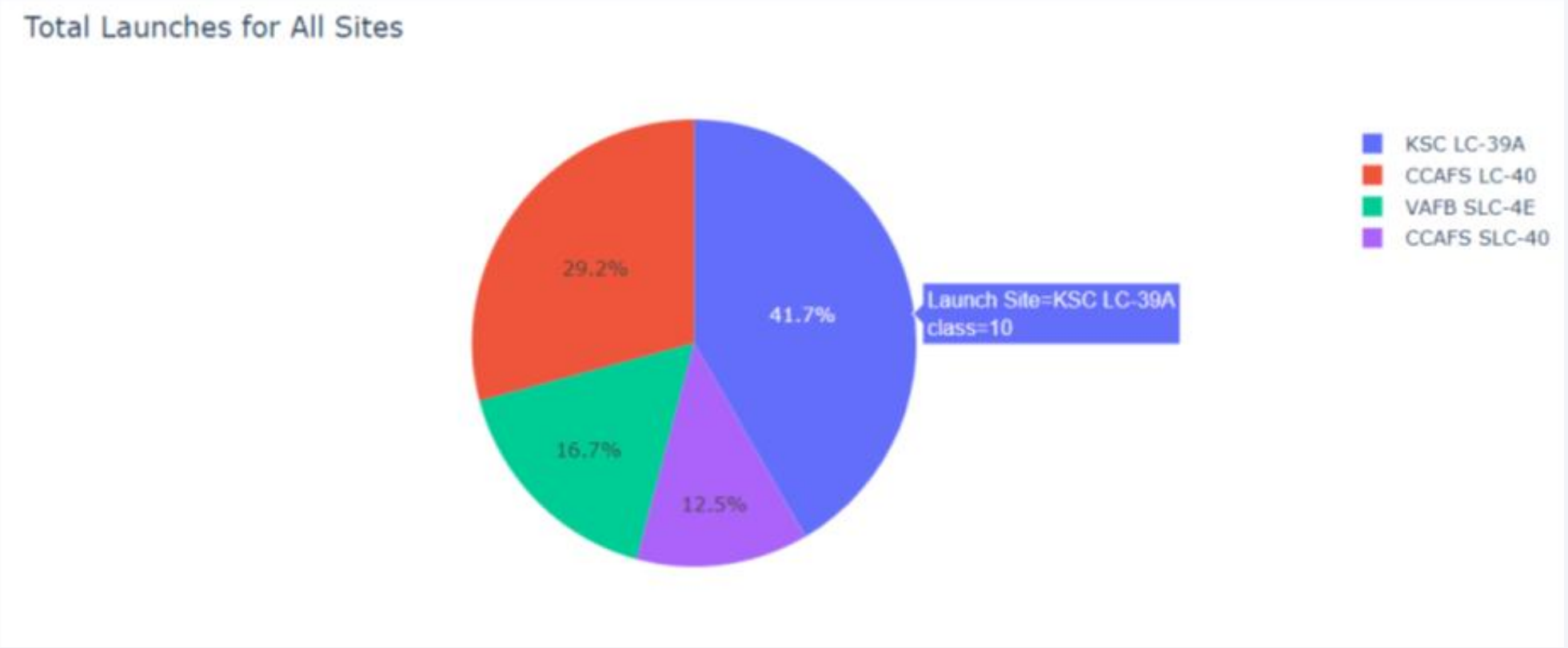
Launch Success Count for all Sites

Showing the screenshot of launch success count for all sites, in a Piechart.

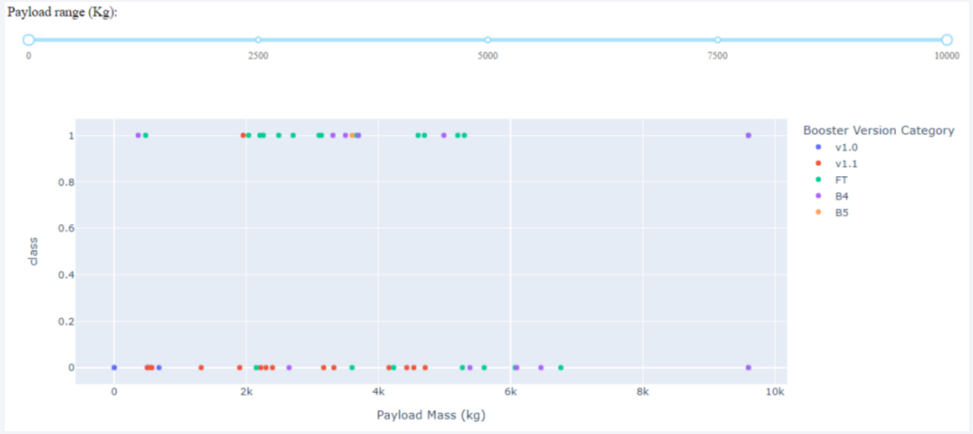


Highest Success Launch Site

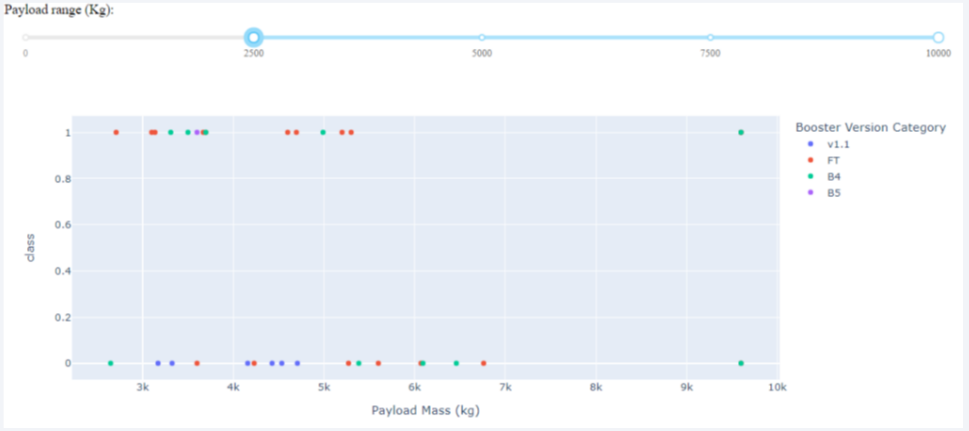
Showing the screenshot of the Piechart for the launch site with highest launch success ratio.



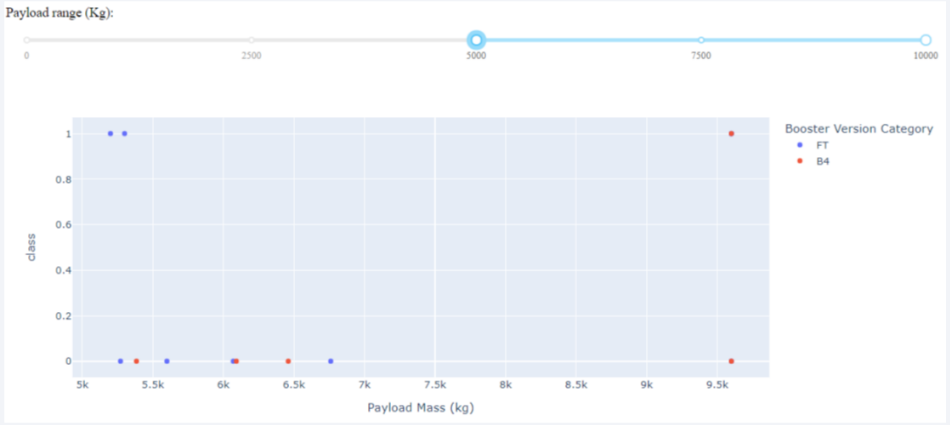
Payload vs Launch Outcome



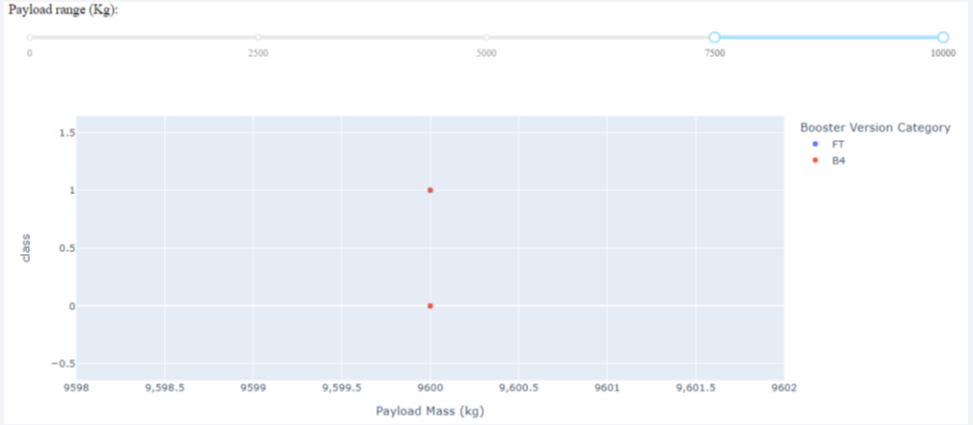
Payload Mass: 0 kg



Payload Mass: 2500 kg



Payload Mass: 5000 kg



Payload Mass: 7500 kg

Section 5

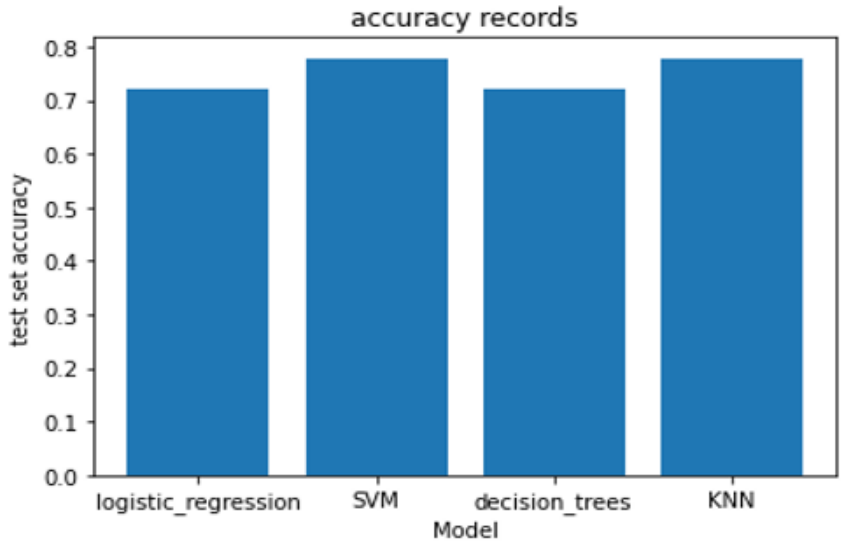
Predictive Analysis (Classification)

Classification Accuracy

SVM and KNN have the highest out of sample accuracy, thus they are the most accurate models.

```
In [44]: x = [lr.score(X_test,Y_test),svm.score(X_test,Y_test),tree.score(X_test,Y_test),KNN.score(X_test,Y_test)]
         y = ['logistic_regression','SVM','decision_trees','KNN']
         plt.bar(y, x)
         plt.title('accuracy records')
         plt.xlabel('Model')
         plt.ylabel('test set accuracy')
         plt.show
```

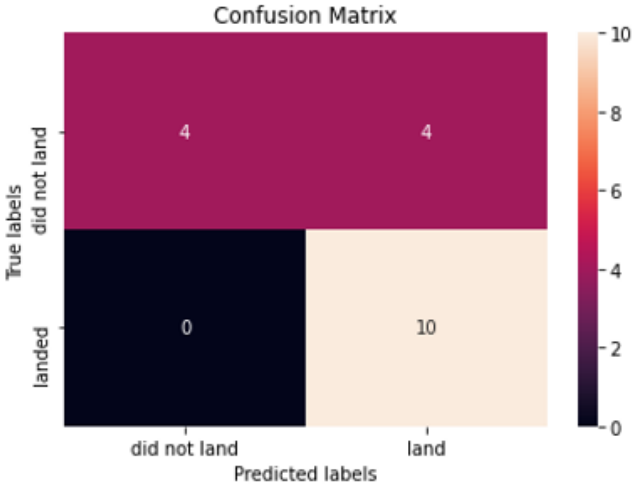
```
Out[44]: <function matplotlib.pyplot.show(*args, **kw)>
```



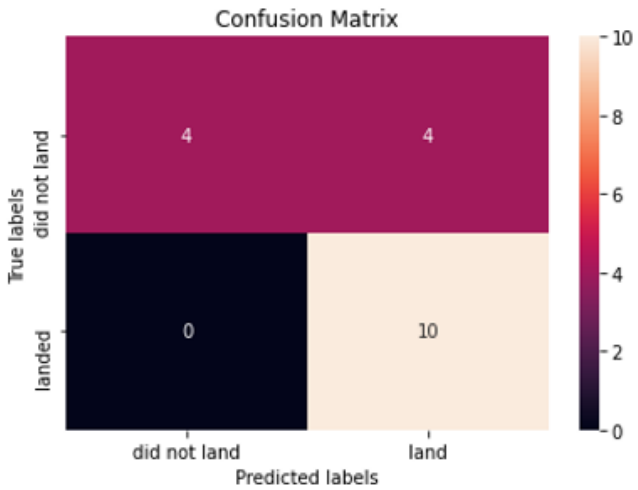
Confusion Matrix

These two graphs represent the confusion matrix for both the SVM and KNN models. These confusion matrices show the largest true positive and true negative values, as well as the least false positive and false negative values.

```
In [36]: yhat = knn_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```




```
In [22]: yhat=svm_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Improved Success Rate Over Time: SpaceX's average launch success rate has significantly improved over the years, with a major upward trend observed from 2014 to 2019
- Launch Site Correlation: The KSC LC-39A launch site had the highest total launch count, indicating it might be associated with a higher launch success ratio compared to other sites
- Orbit Type Impact on Success: Certain orbits, specifically ES-L1, GEO, and HEO, achieved a 100% success rate in the dataset, while GTO had a lower success rate of approximately 50%
- Predictive Model Performance: The machine learning models, specifically SVM and KNN, demonstrated the highest out-of-sample accuracy for predicting landing outcomes among the models tested

Thank you!