IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bryan Headrick
10/6/2024

Outline

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

Executive Summary

Summary of methodologies

Data was collected from the SpaceX REST API and through web scraping. Data was processed using pandas to standardize and group the data in meaningful ways and perform statistical analysis. Exploratory Data Analysis was performed to asses predictors of successful launches/landings. Visual mapping was performed to analyze additional factors that promoted a successful launch. An interactive dashboard was created to analyze how success corresponded to site, payload size, and booster version. And finally, machine learning models were created to predict successful launches based on all of the data fields collected.

Summary of all results

Payloads from 2000 to 4000 appear to have the highest success rate. FT appears to be the most successful Booster Version category. GTO and ISS were the two most successful orbit types. Logistic Regression and K Nearest Neighbor are tied at 94% for the most accurate predictive model in predicting successful landing. The most successful launch site was the furthest inland

Introduction

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

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX REST API and through webscraping
- Perform data wrangling
  - Data was processed using pandas to standardize and group the data in meaningful ways and perform statistical analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV, various alrorithms were tested to determine the best for predicting successful launches

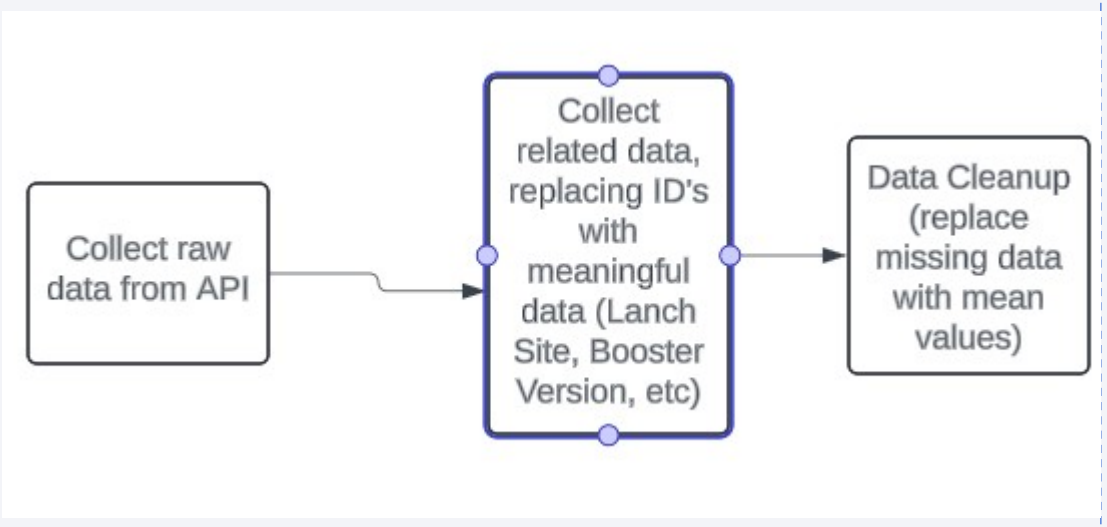
Data Collection

Describe how data sets were collected.

Present data collection process using key phrases and flowcharts

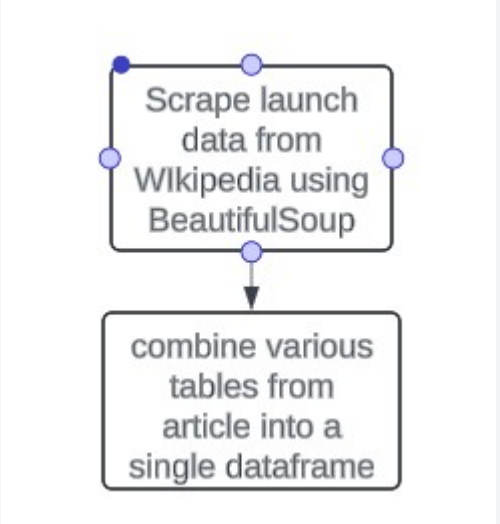
Data Collection – SpaceX API

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Identify missing values in each column
- Identify the data types of each column
- Analyze the data getting the counts of:
  - Orbit Types
  - Launch Sites
  - Landing Outcomes
- Add a class field to the dataframe to identify the overall landing success/failure
- Analyze the overall landing success rate

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter charts were used to illustrate the relationship between flight number, payload mass, and landing success. The data shows that landings became increasingly more likely to be successful, and larger payloads were used as there were more launches.

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Displayed overview of dataset
- Listed launch site names and booster names
- Displayed high level statistics of dataset, such as overall successes, failures, average payload mass, count of each outcome

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Added circles and markers to identify launch sites.
- Added marker clusters to indicate successful/unsuccessful launches
- Added other markers to indicate distance from various landmarks, such as:
  - Coastlines
  - Cities

The rationale for this is to illustrate some of the factors that go into finding an optimal location for a launch site.

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

First, there is a pie chart to show success/failure rate for each launch site as well as a pie chart to show the success rate for each site as a percentage of all successes.

Next is a scatter chart to show the relationship between booster version, payload mass, and launch success.

The charts are intended to answer 5 questions:

1. Which site has the largest successful launches?
2. Which site has the highest launch success rate?
3. Which payload range(s) has the highest launch success rate?
4. Which payload range(s) has the lowest launch success rate?
5. Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data was gathered from csv files. The dependent feature (class) was assigned to a separate variable. The data was then standardized to ensure all features were weighted the same. Data was, then split into train and test sets. Various predictive models were created, including logistic regression, support vector machine, decision tree classifier, and k nearest neighbor.

https://github.com/catmanstudios/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

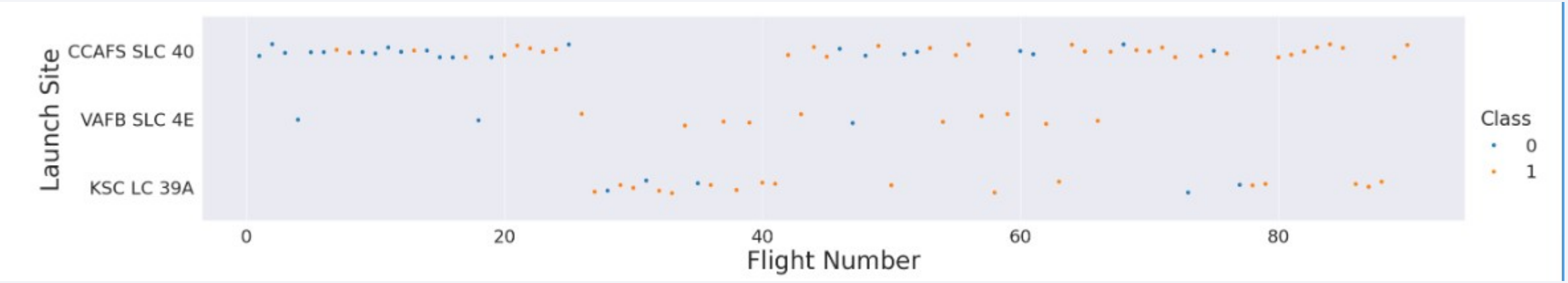
Section 2

Insights drawn from EDA

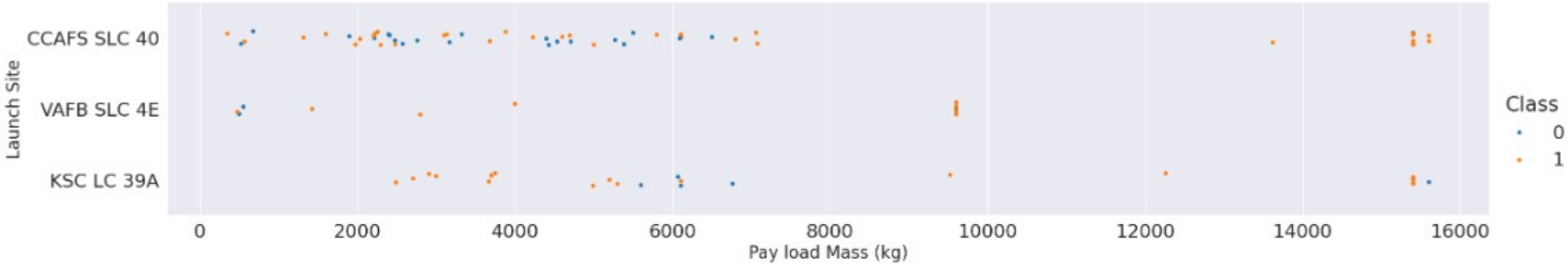
Flight Number vs. Launch Site

Flight Number vs. Launch Site

- The majority of the 1st 3rd of launches were from CCAFS
- The 2nd 3rd of launches had a comparable number of launches between all 3 sites with the majority of the initial launches in KSC LC 39A.
- The majority of the final 3rd of launches were from CCAFS

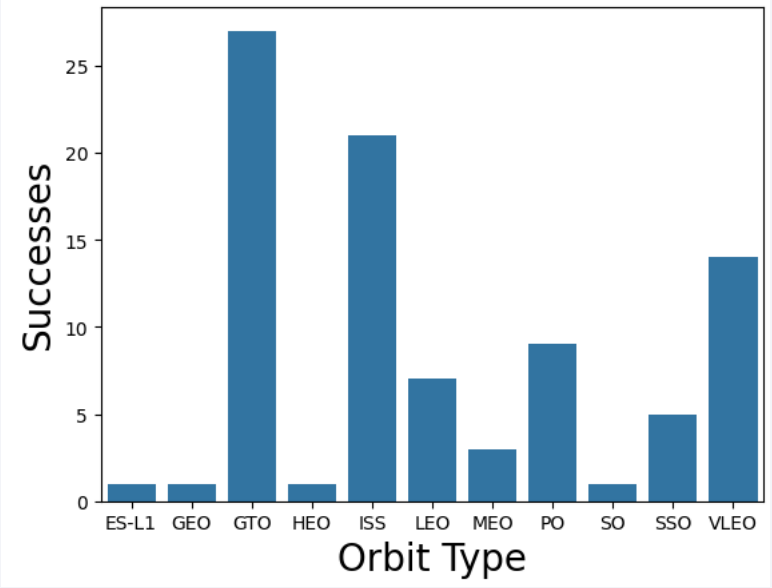


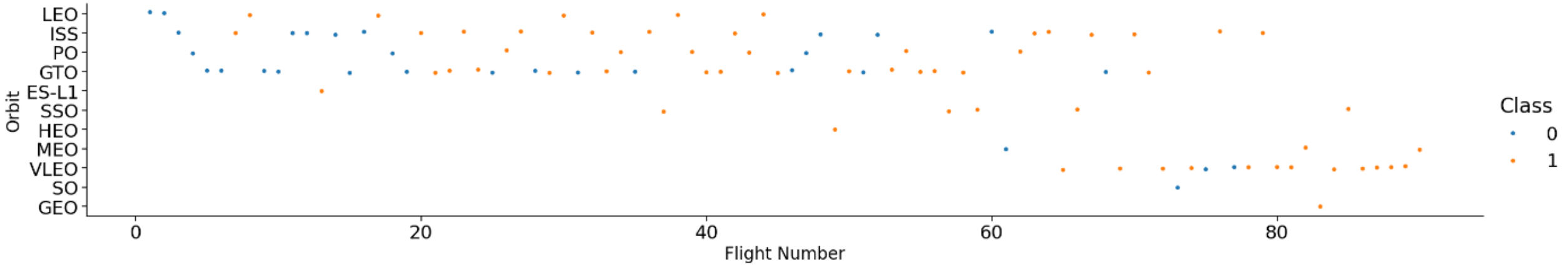
Payload vs. Launch Site



Now if you observe Payload Mass 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).

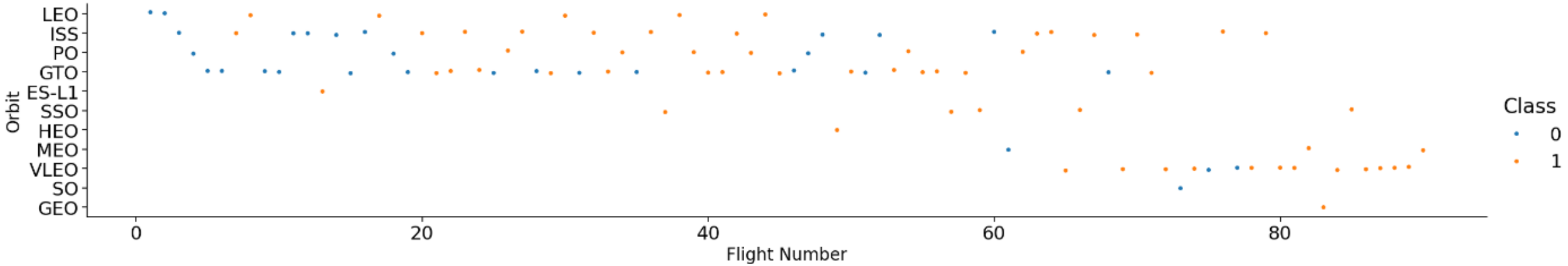
Success Rate vs. Orbit Type





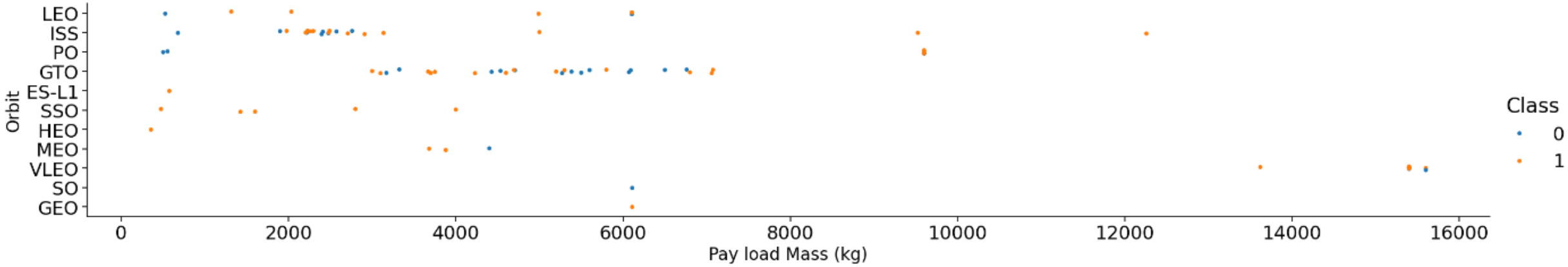
You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

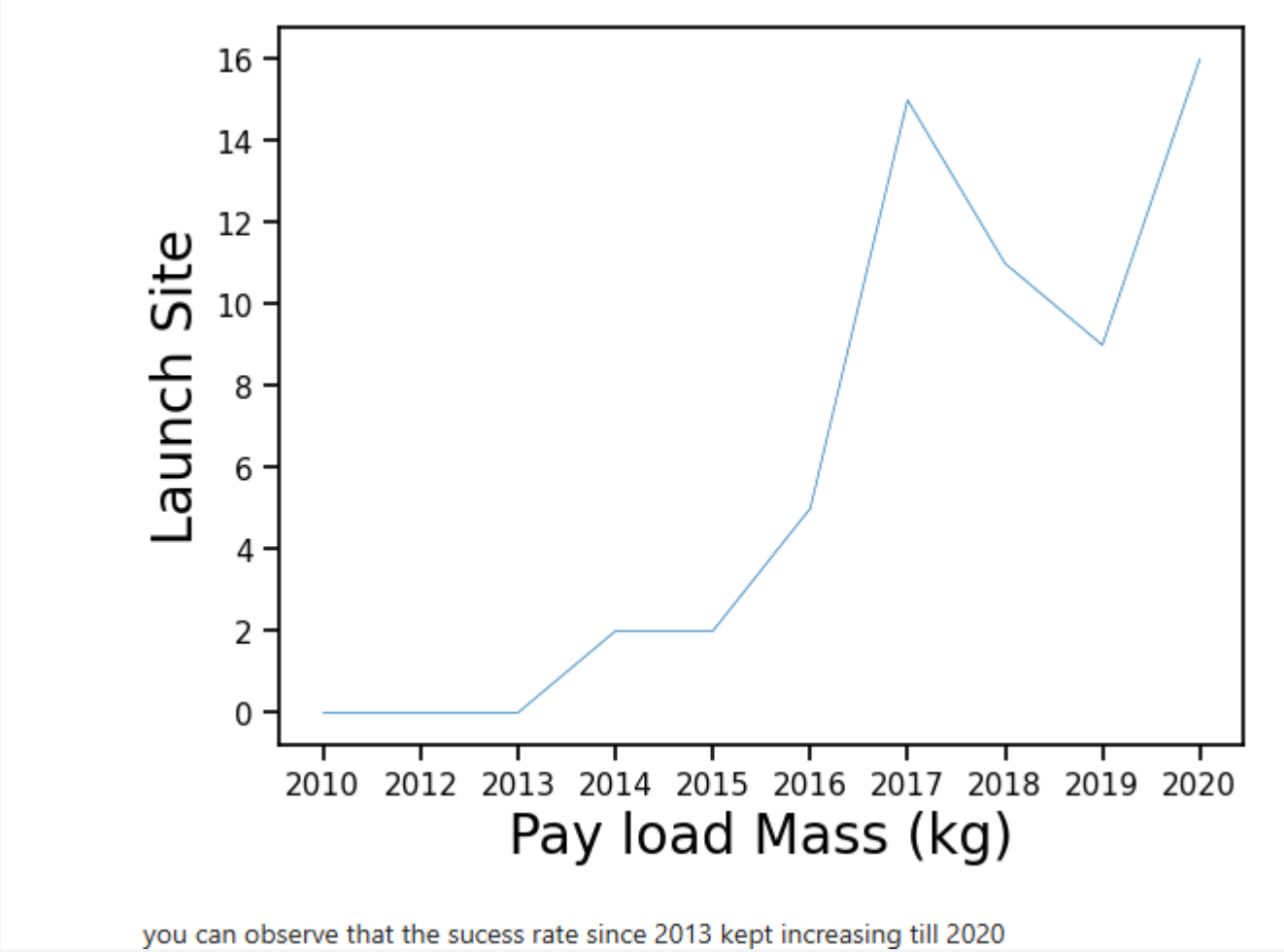
Payload vs. Orbit Type



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

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Query:

```
select distinct Launch_Site from SPACEXTBL
```

Explanation:

Launch_Site is the column name where the launch site names are stored, the distinct directive ensures no duplicates are listed

Launch Site Names Begin with 'CCA'

Query:

```
select * from SPACEXTBL WHERE Launch_Site like 'CCA%' limit 5
```

Explanation:

The ‘like’ directive looks for partial text matches, and the % is the wildcard character in the expression.

The ‘limit 5’ directive means we’re only displaying the first 5 results.

Total Payload Mass

Query:

```
select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

Explanation:

The where clause isolates records matching that condition, and the sum directive sums values in that column across all filtered records.

Average Payload Mass by F9 v1.1

Query:

```
 select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where
Booster_Version like '%F9 v1.1%'
```

Explanation:

The where clause filters records matching the condition where the Booster_Version partially matches the ‘F9 v1.1’ string, and the avg directive averages values in the specified column across all filtered records.

First Successful Ground Landing Date

Query:

```
 select min(Date) from SPACEXTBL  where
Landing_Outcome='Success (ground pad)'
```

Explanation:

The where clause filters records matching the specfied value for Landing_Outcome, and the min directive selects the lowest value in the date column within the filtered results.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
select Booster_Version from SPACEXTBL where  Landing_Outcome
='Success (drone ship)' and PAYLOAD_MASS__KG_ >4000 and
PAYLOAD_MASS__KG_ <6000
```

Explanation:

This lists the Booster_Version values with landing_outcomes matching values for successful drone ship and payload mass is greater than 4000kg and less than 6000kg

Total Number of Successful and Failure Mission Outcomes

Query:

```
select sum(success) as success, sum(failure) as failure from ( select case
WHEN Mission_Outcome ='Success' then 1 else 0 end success, case
WHEN Mission_Outcome ='Success' then 0 else 1 end failure from
SPACEXTBL) t
```

Explanation:

This uses a subquery with case statements to independently count successes and failures (based on whether the mission_outcome field is ‘Success’)

Boosters Carried Maximum Payload

Query:

```
select distinct Booster_Version FROM spacextbl where
PAYLOAD_MASS__KG_ =(select max(PAYLOAD_MASS__KG_) from
spacextbl)
```

Explanation:

This lists all booster_version values having a record where the payload mass is at the maximum value for the entire dataset, and only lists each booster_version once.

2015 Launch Records

Query:

```
select substr(Date, 6,2) as month, Booster_Version,
Landing_Outcome, Launch_site from spacextbl where
substr(Date,0,5)='2015' and Landing_Outcome like 'Failure%'
```

Explanation:

This extracts the month portion of the date, booster version, landing outcome (failure type), and launch site for launches that failed to land in 2015

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

Query:

```
select count(Date) as "outcome count", Landing_Outcome from
spacextbl where Date between  '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by count(Date) desc
```

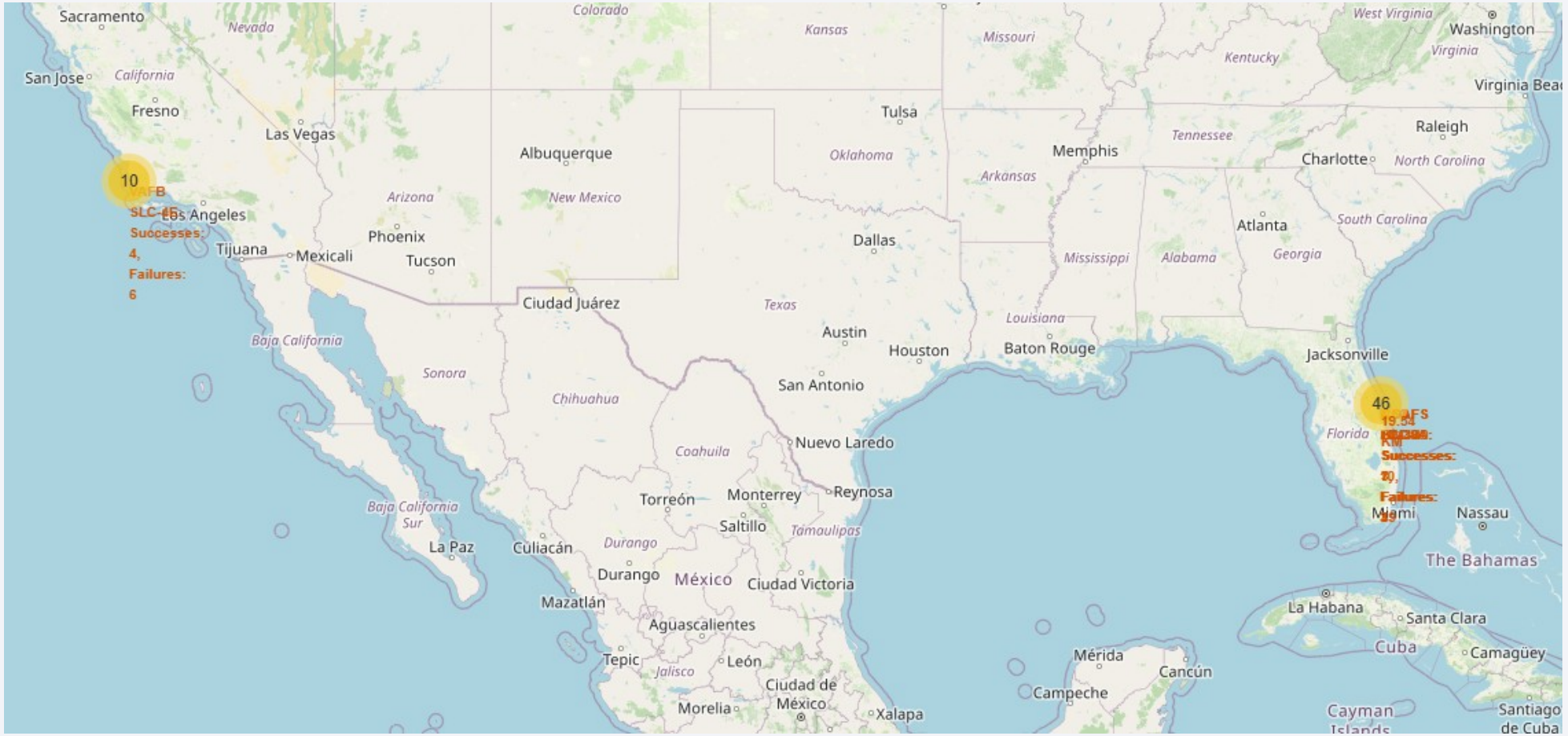
Explanation:

This returns the count of records in the above date range grouped by landing outcome and sorted with the highest counts first.

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Site Locations



All sites are near the coast, but roughly 1km or more inland

All sites are reasonably far away from major cities

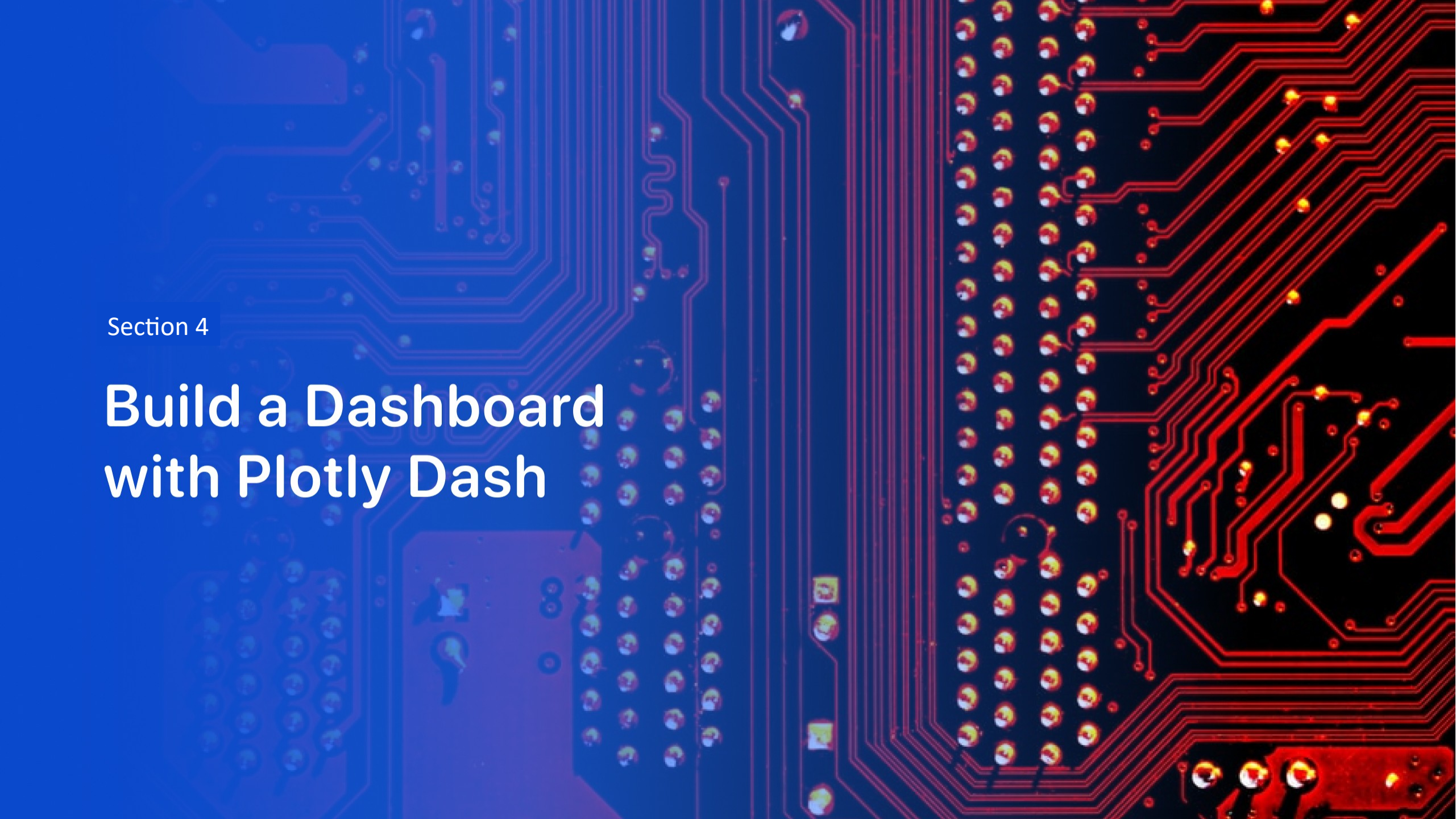
Success and Failure Markers

The markers make it easier to quickly identify sites with higher success rates

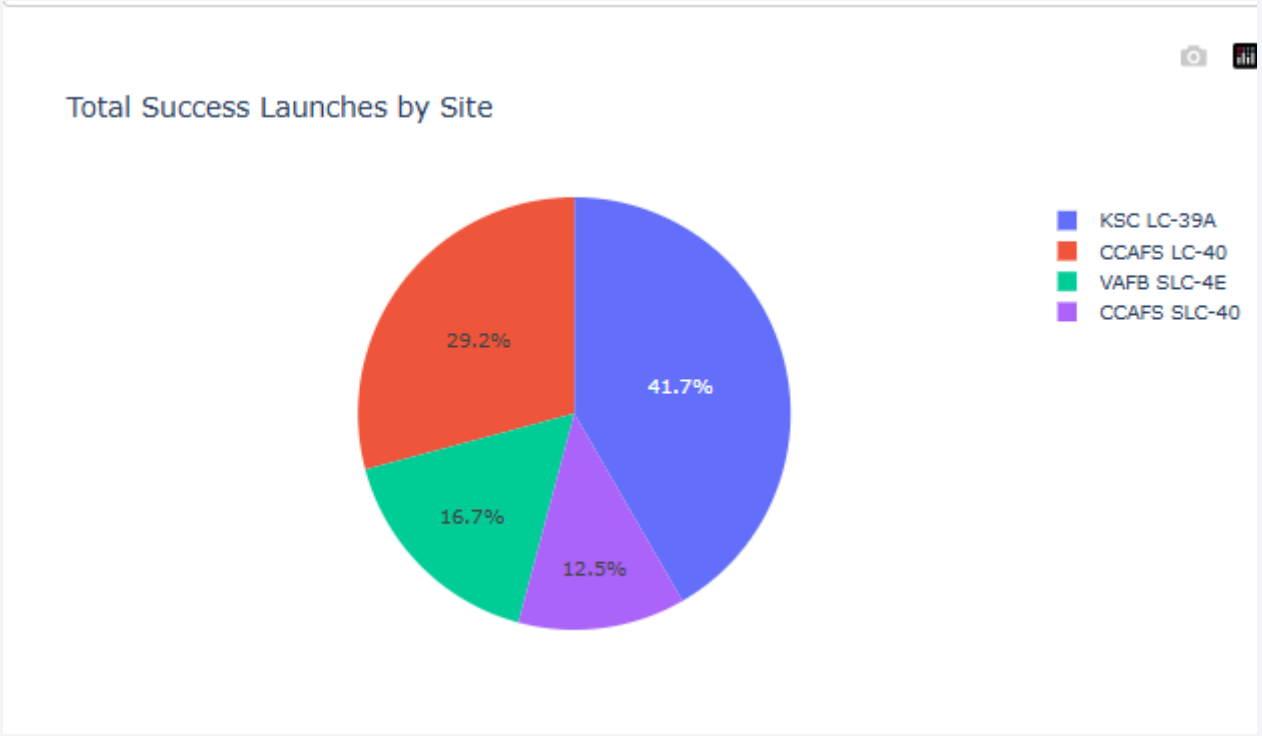
KSC LC-39A had a much higher success rate than other launch sites. It was also much further inland than the other launch sites, although this is not enough data to suggest a causal link

Proximities Map

The launch sites are reasonably close to coasts and railways to facilitate ease of transporting massive parts, but very distant from neighboring cities to limit potential noise pollution and collateral damage in the event of crashes.

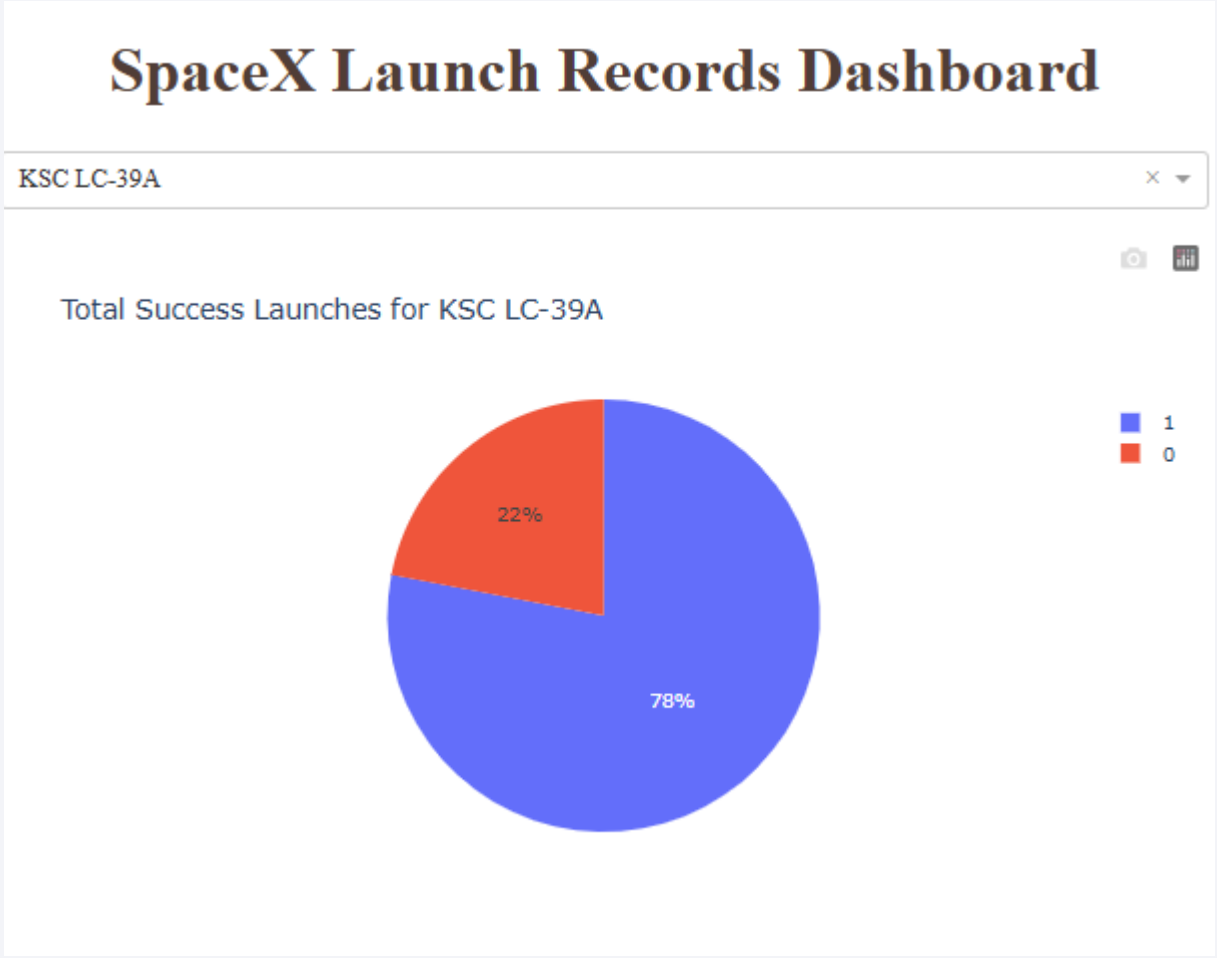

Section 4
Build a Dashboard with Plotly Dash

Launch Success Statistics



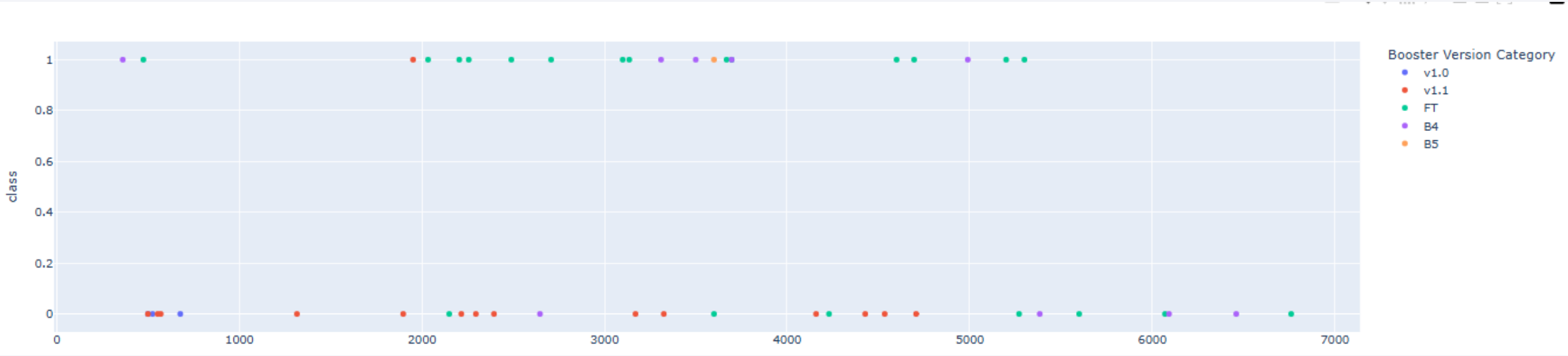
KSC LC-39A had the greatest share of successes (41.7%) as well as the highest success rate, at 78%.

Launch Site With Highest Success Ratio



KSC LC-39A had 78% successes and 22% failures.

Payload vs Launch Outcome

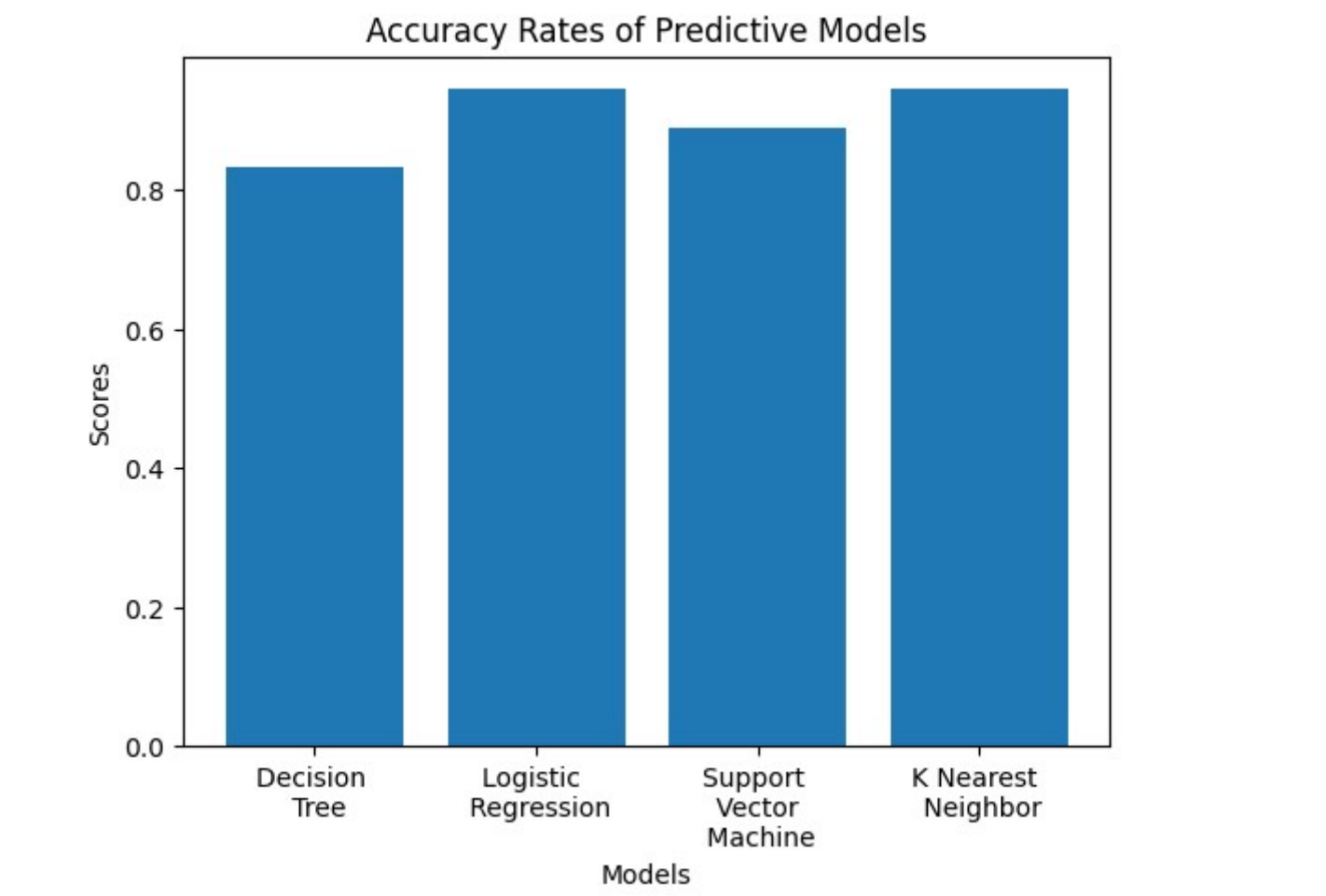


Payloads from 2000 to 4000 appear to have the highest success rate, and FT appears to be the most successful Booster Version category.

Section 5

Predictive Analysis (Classification)

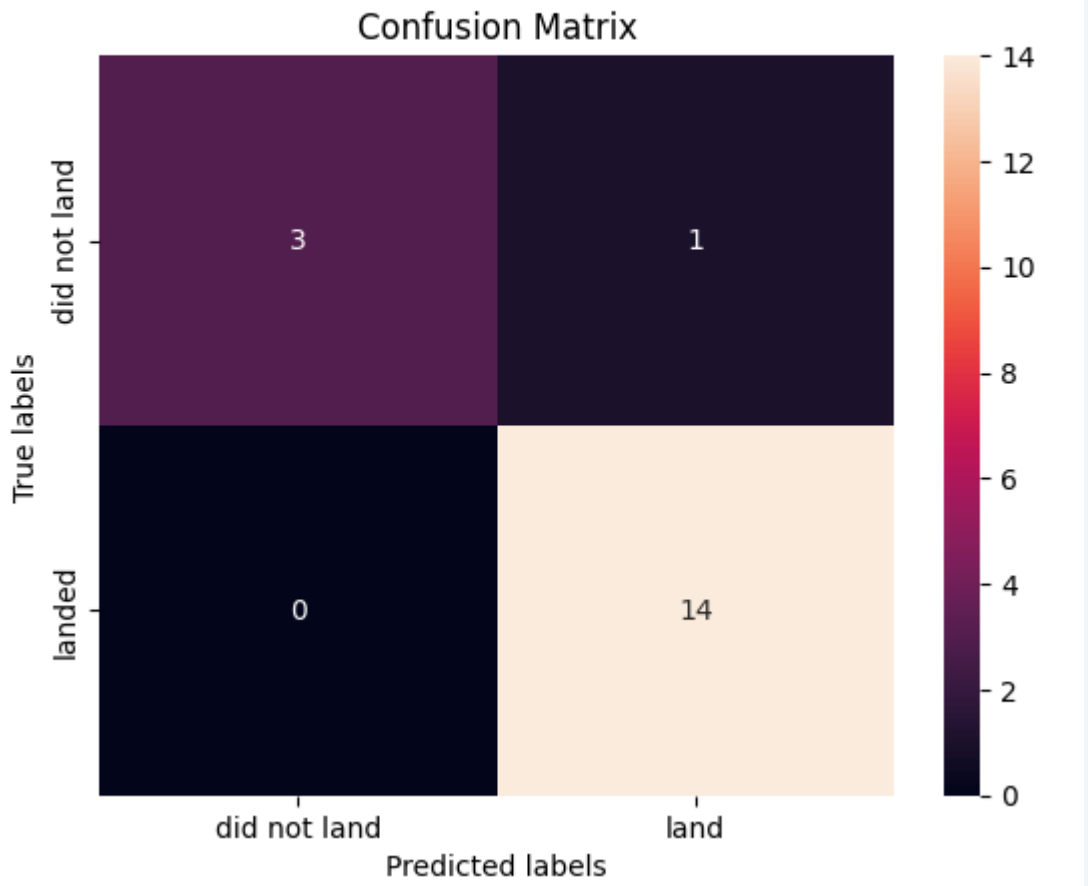
Classification Accuracy



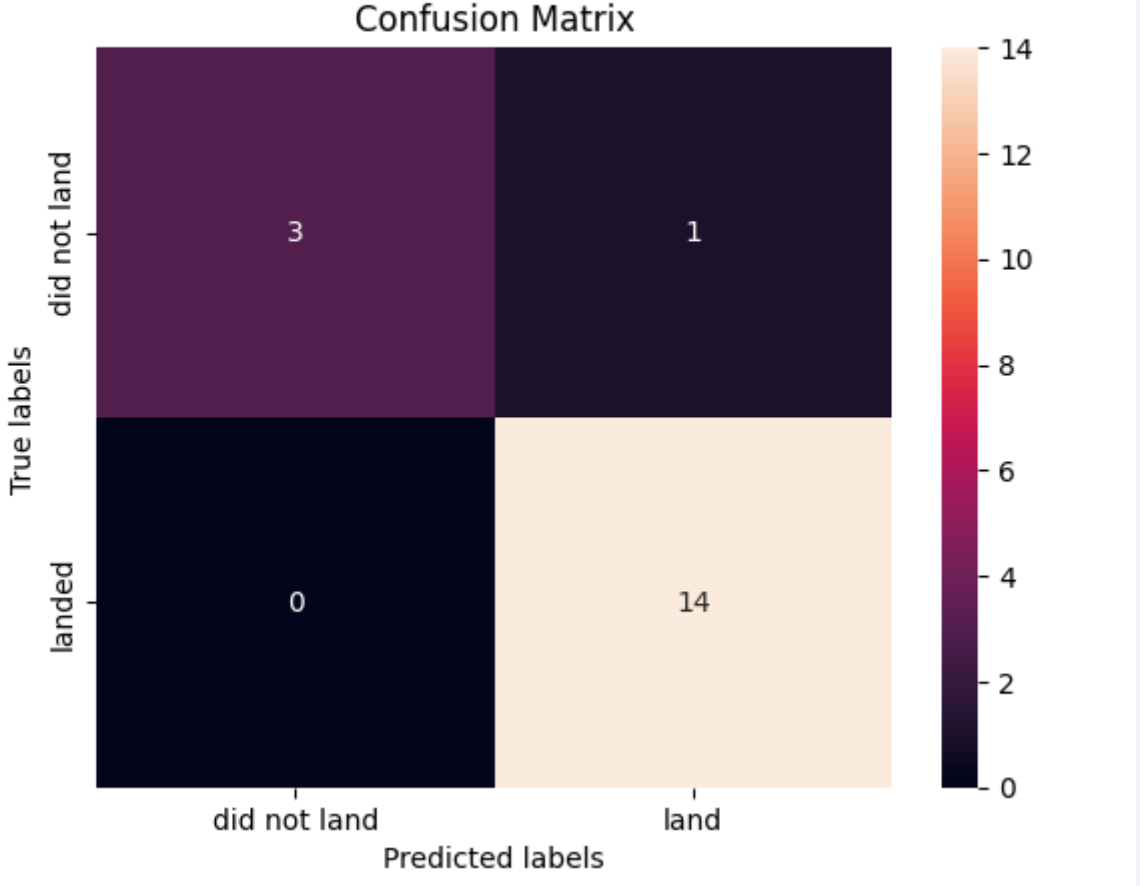
Logistic Regression and K Nearest Neighbor are tied at 94% for the most accurate predictive model

Confusion Matrix

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




```
[17]: yhat=logreg_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



Both KNN and Logistic Regression models only had a single false-positive prediction and no false-negative predictions.

Conclusions

Payloads from 2000 to 4000 appear to have the highest success rate

FT appears to be the most successful Booster Version category.

GTO and ISS were the two most successful orbit types

Logistic Regression and K Nearest Neighbor are tied at 94% for the most accurate predictive model in predicting successful landing

The most successful launch site was the furthest inland

Appendix

Thank you!