IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elizabeth Mano
06/19/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Analysis Using SQL
  - Exploratory Analysis Using Panda and Matplotlib
  - Interactive Visual Analytics with Folium
  - Building an Interactive Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Visual Analytics Demo
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX has emerged as a leader in the commercial space sector by substantially lowering the cost of orbital launches. Traditional launch providers typically charge upwards of $165 million per mission, whereas SpaceX offers Falcon 9 launches at approximately $62 million. This significant cost advantage is largely attributable to the successful recovery and reuse of the Falcon 9 first stage booster. Accurately forecasting the outcome of first stage recovery is essential for precise cost estimation and strategic mission planning. By utilizing publicly available launch data in conjunction with advanced machine learning methodologies, this project seeks to develop a robust predictive model to determine the probability of successful first stage landing. Such predictive capability will enable more informed financial and operational decisions, thereby contributing to greater efficiency and sustainability in launch operations.

- Problems you want to find answers
  - How do variables such as payload mass, number of flights, orbits, and launch sites affect the success rate of the first stage landings?
  - Does the rate of successful first stage landings increase over time?
  - What is the best procedures that can be used for binary classification in this case study?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Using SpaceX API
  - Webscraping from Wikipedia
- Perform data wrangling
  - Filtering Data
  - Dealing with missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluation of classification models to ensure best results.

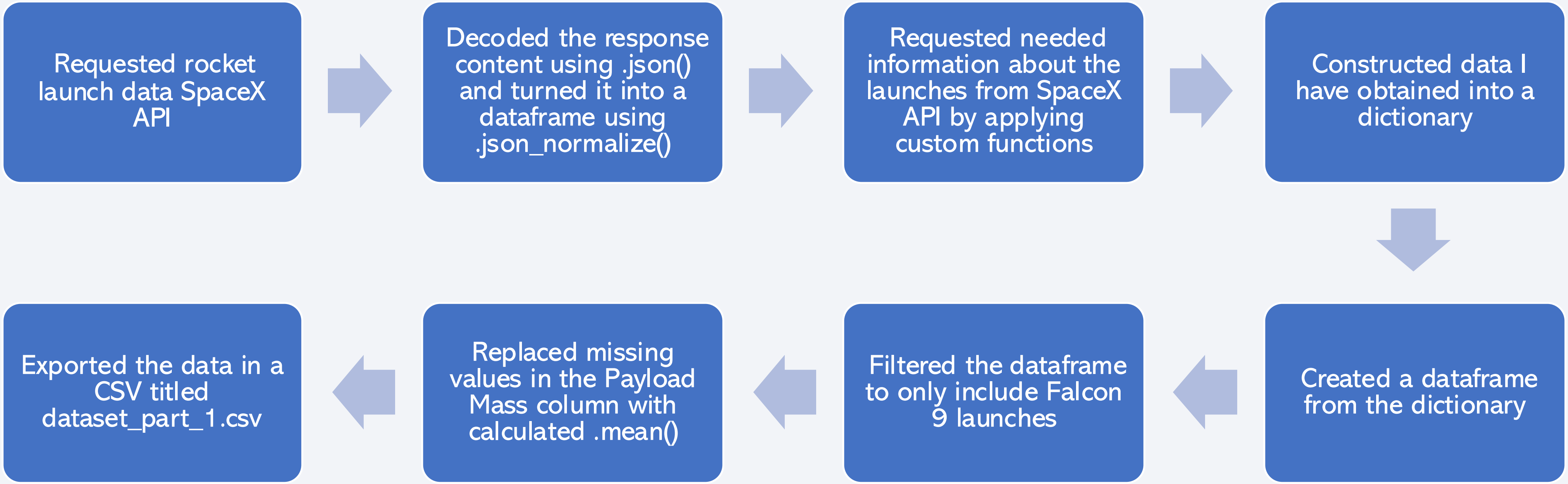
Data Collection

The data collection process involved a combination of API requests from SpaceX API and Webscraping data from SpaceX's Wikipedia page.

I used both data collection methods to gather complete information about the launches, which helped me get a better understanding and perform a more detailed analysis.

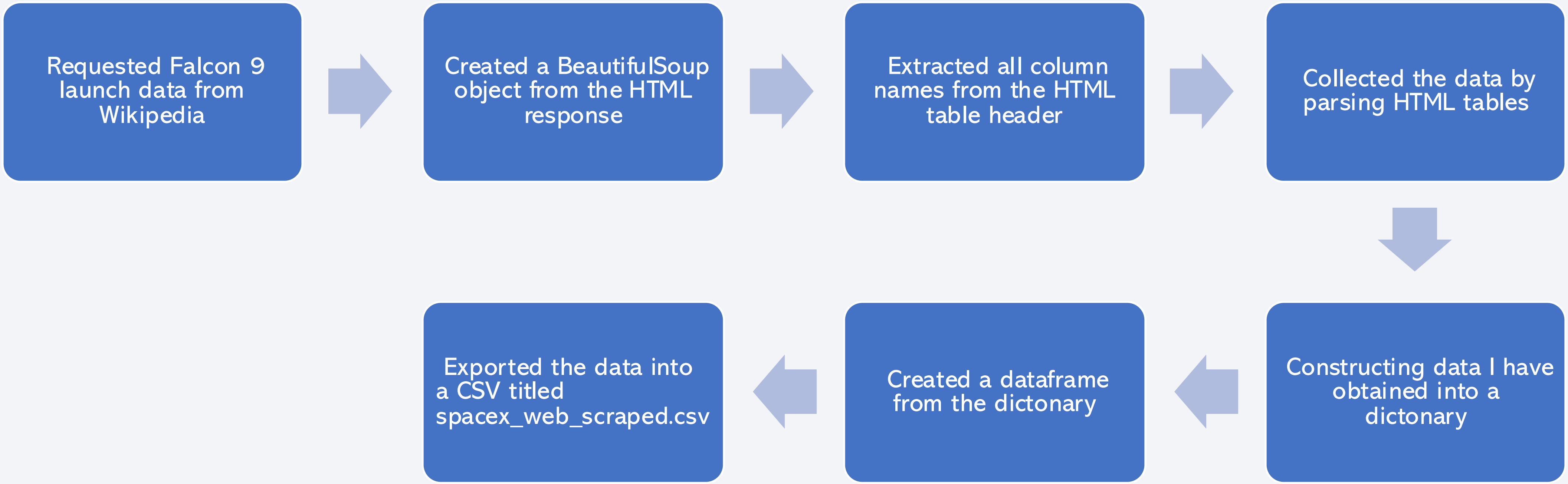
- Data Columns obtained from SpaceX API
  - Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude
- Data Collection from Webscraping
  - Date and Time (UTC) Flight No., Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API



[Data Collection using SpaceX API Github Link]()

Data Collection - Scraping



[Data Collection Using Webscraping Github Link]

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

I will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

1. Performed exploratory Data Analysis and determine Training Labels
2. Calculated the number of launches of each site
3. Calculated the number and occurrence of each orbit
4. Calculated the number and occurrence of mission outcomes per orbit type
5. Created a landing outcome label from Outcome column
6. Exported the data to a CSV titled dataset_part_2.csv

[Data Wrangling Github Link]

EDA with Data Visualization

- Charts that were plotted
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type and Success Rate Yearly Trend
- Scatter plots are a type of data visualization that displays the relationship between two numerical variables. It can be used in machine learning models.
- Bar charts are a type of data visualization that shows comparisons among discrete categories.
- Line charts are a type of data visualization that show trends in data over time.

EDA with Matplotlib and Pandas Github Link

EDA with SQL

- Displayed the names of unique launch sites in the space missions.
- Displayed 5 records where launch sites begin with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed the average payload mass carried by booster version F9 v1.1.
- Listed the date when the first successful landing outcome in the ground pad was achieved.
- Listed the names of the boosters which have successful drone ships and payload masses greater than 4000 kg but less than 6000 kg.
- Listed the total number of successful and failure mission outcomes.
- Listed the names of the booster versions which have carried the maximum payload mass.
- Listed the failed landing outcomes in drone ships, their booster versions, and the launch site names for the months the year 2015.
- Ranked the count of landing outcomes (Failure (drone ship), or Success (ground pad)) between the dates of 06.04.2010 to 03.20.2017 in descending order.

[EDA with SQL Github Link]

Build an Interactive Map with Folium

- Markers of all Launch Sites (Space X and NASA)
  - Added Markers with Circle, Popup Label, and Text Label of NASA Johnson Space Center using it's latitude and longitude coordinates as a starting location.
  - Added Markers with Circle, Popup Label, and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and their proximity to the Equator and coastlines.
- Colored Markers of the launch outcomes for each Launch Site
  - Added colored Markers of successful (green) and failed (red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distance between a Launch Site to Railways, Highways, Coastline, and Closest City.
  - Added colored Lines to show distances between Launch Site KSC LC-39A (as an example) and its proximity to Railways, Highways, Coastlines, and Closest City.

Interactive Map with Folium Github Link

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown
  - Added a dropdown list to allow Launch Site Selections
- Pie Chart showing Successful Launches (All Sites/Or Certain Sites.
  - Added a pie chart showing the total successful launches count for all sites and the Success vs. Failed counts for each site.
- Slider of Payload Mass Range
  - Added a slider to select Payload Mass Range
- Scatter Chart of Payload Mass vs. Success Rate for Booster Versions
  - Added a scatter chart to show the correlation between Payload Mass and Launch Success Rate.

Dashboard with Plotly Dash Github Link

Predictive Analysis (Classification)

Created a NumPy array from the column 'Class' in the data

Standardized the data with StandardScaler, then fitted and transformed the data

Split the data into training and testing sets with train_test_split function

Created a GridSearchCV object with cv=10 to find the best parameters

Applied GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculated the accuracy of the test data using method.score() for all models

Examined the confusion matrix for all models

Found the method performs best by examining the Jaccard_score and F1_score metrics.

Predictive Analysis (Classification) Github Link

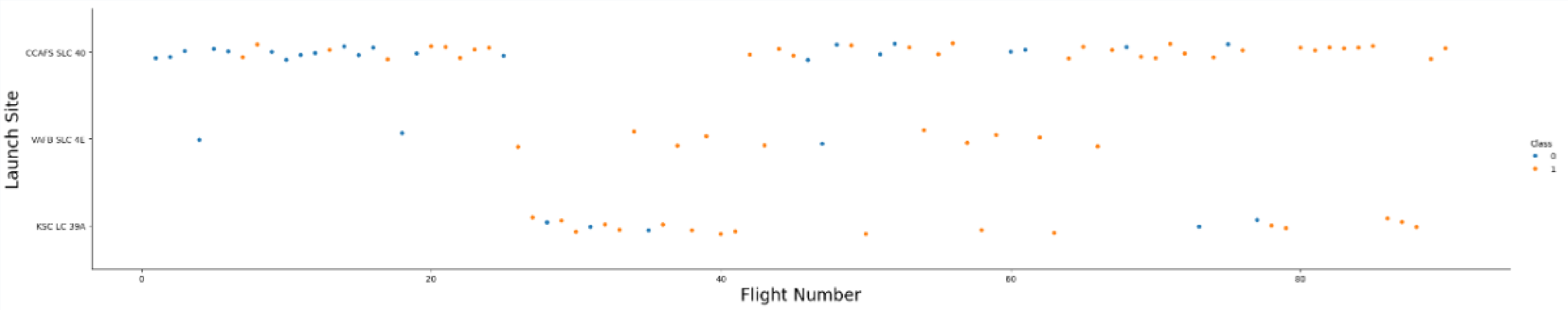
Results

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

Section 2

Insights drawn from EDA

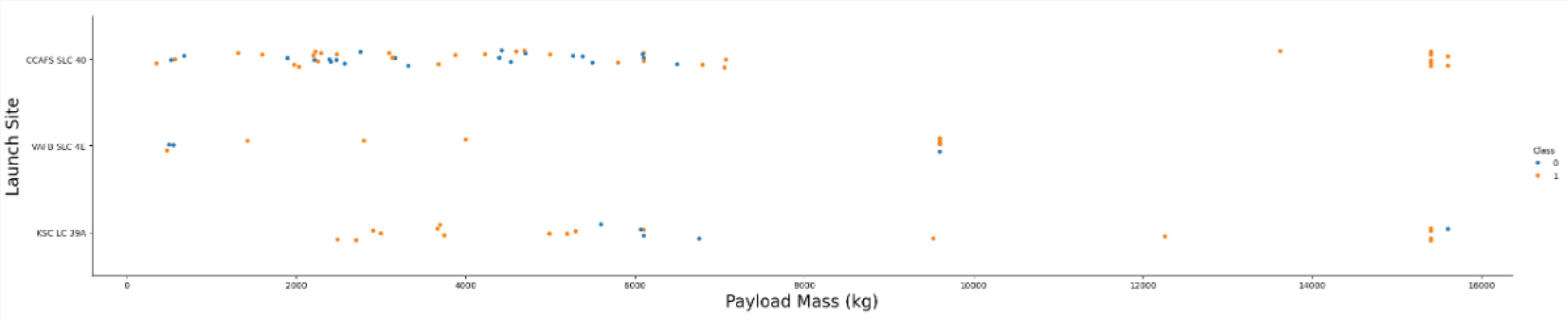
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch will have a higher rate of success.

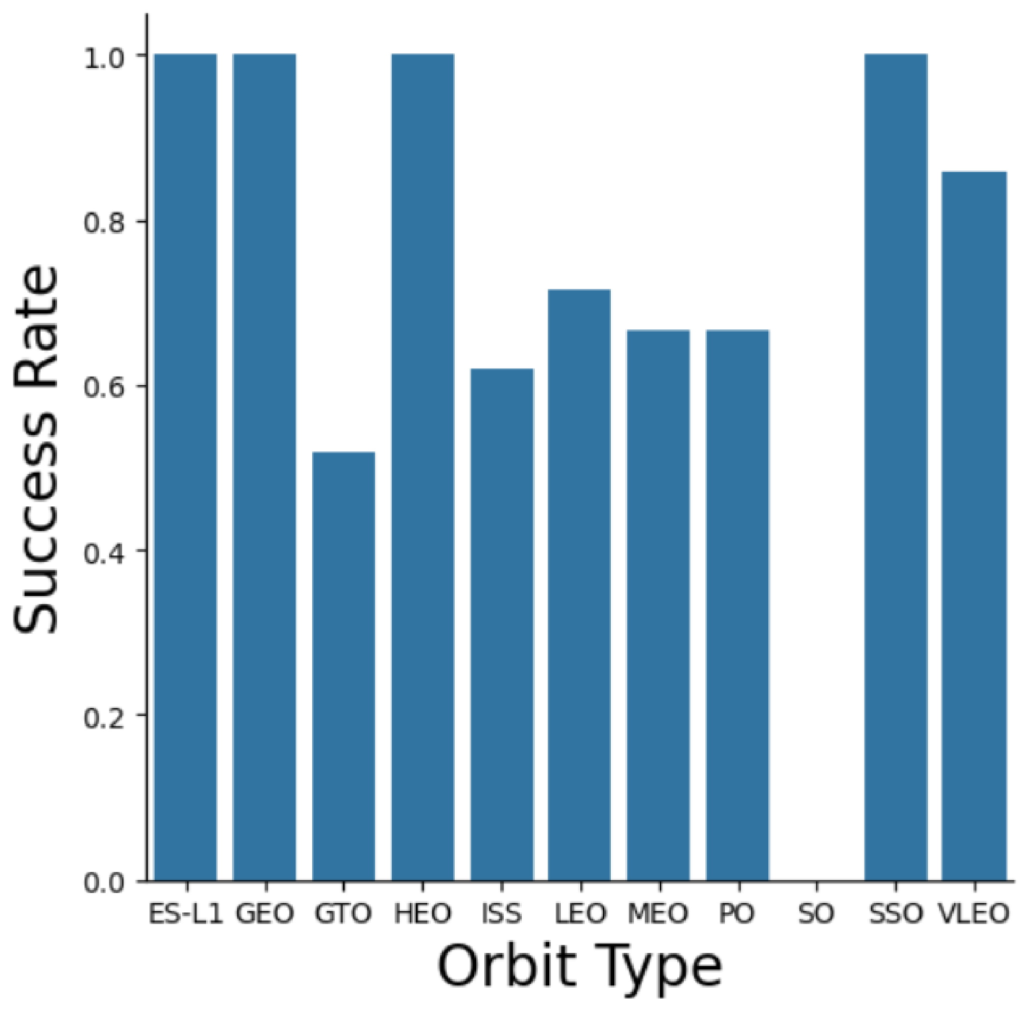
Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg

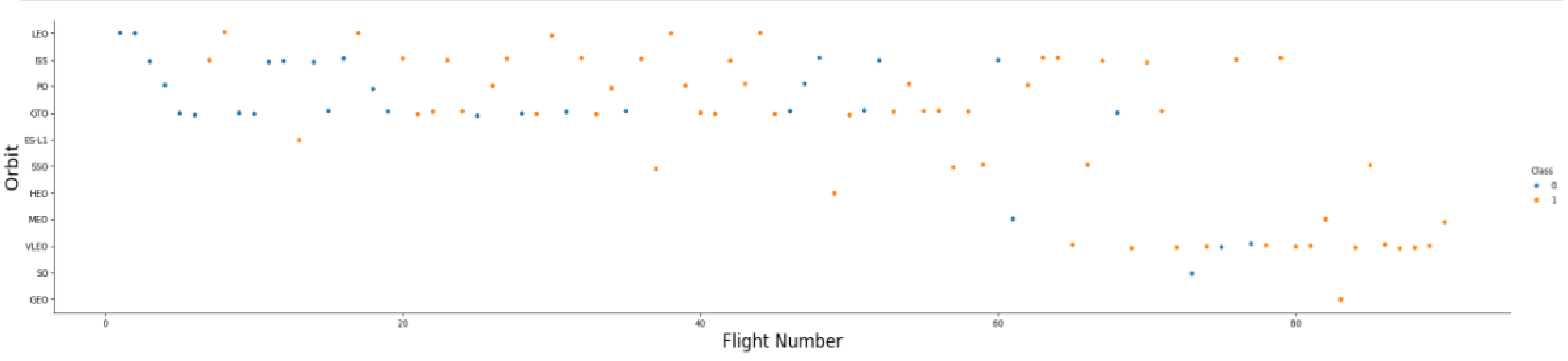
Success Rate vs. Orbit Type



Explanation:

- SSO, HEO, GEO, ES-L1 have the highest success rate by orbit at 100%
- VLEO has the second highest success rate at 85%.
- GTO, LEO, ISS, PO all are lower success rates betwee 50% and 70%.
- SO has 0% success rate.

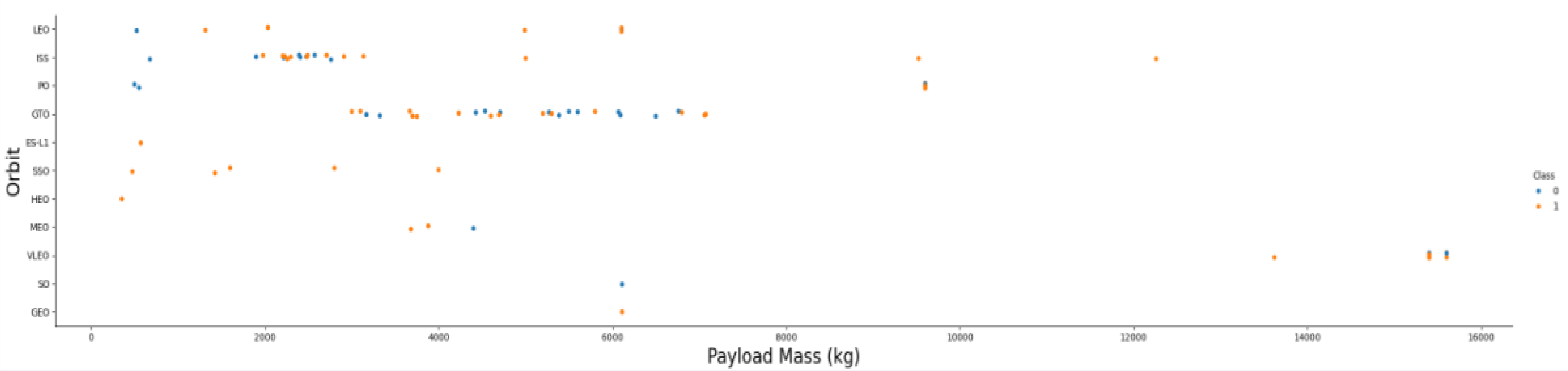
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit the success rate appears to be related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



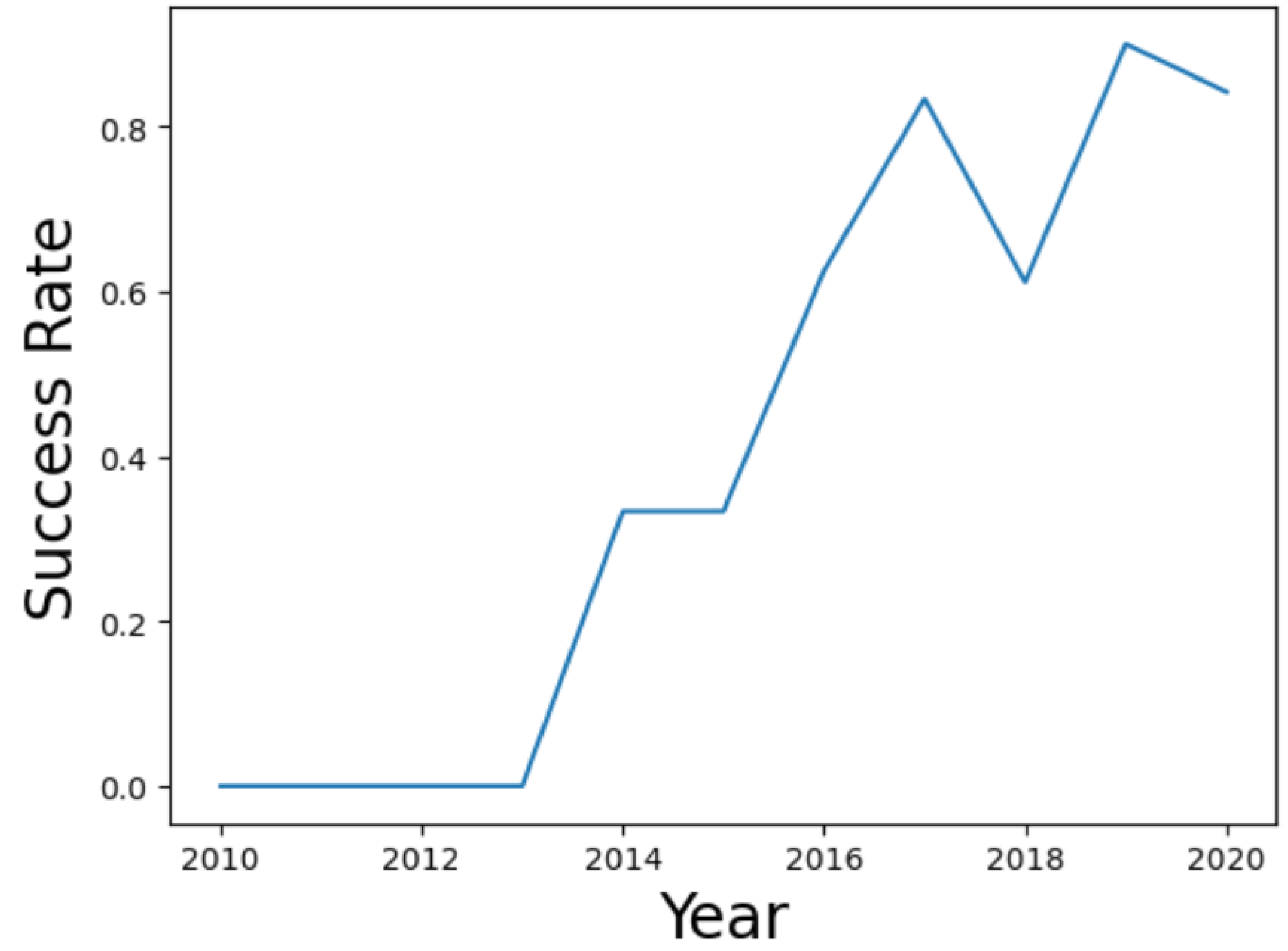
Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive influence on Polar LEO (ISS) orbit.

Launch Success Yearly Trend

Explanation:

- The success rate since 2013 has kept increasing until 2020.



All Launch Site Names

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Explanation:

- The unique launch site names are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

Explanation:

- The top 5 records that begin 'CCA'.

```
        Date Time (UTC) Booster_Version  Launch_Site  \
0  2010-06-04   18:45:00   F9 v1.0  B0003  CCAFS LC-40
1  2010-12-08   15:43:00   F9 v1.0  B0004  CCAFS LC-40
2  2012-05-22    7:44:00   F9 v1.0  B0005  CCAFS LC-40
3  2012-10-08    0:35:00   F9 v1.0  B0006  CCAFS LC-40
4  2013-03-01   15:10:00   F9 v1.0  B0007  CCAFS LC-40

                                              Payload  PAYLOAD_MASS__KG_  \
0              Dragon Spacecraft Qualification Unit                   0
1  Dragon demo flight C1, two CubeSats, barrel of...                   0
2                              Dragon demo flight C2                 525
3                                     SpaceX CRS-1                 500
4                                     SpaceX CRS-2                 677

       Orbit         Customer Mission_Outcome      Landing_Outcome
0        LEO           SpaceX         Success  Failure (parachute)
1  LEO (ISS)  NASA (COTS) NRO         Success  Failure (parachute)
2  LEO (ISS)      NASA (COTS)         Success           No attempt
3  LEO (ISS)       NASA (CRS)         Success           No attempt
4  LEO (ISS)       NASA (CRS)         Success           No attempt
```

Total Payload Mass

Explanation

- The total Payload Mass is 45596 kg.

	total_payload_mass
0	45596

Average Payload Mass by F9 v1.1

	average_payload_mass
0	2928.4

Explanation

- The average Payload Mass is 2928.4 kg.

First Successful Ground Landing Date

Explanation

- The first successful landing date was 12.22.2015.

	first_successful_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Explanation

- The F9 FT B1022, F9 FT B1026, F9 Ft B1021.2, and F9 Ft B1031.2 had successful drone ship landings with a payload between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Explanation

- The successful missions are 98, while the failure (in flight) is 1.

	Mission_Outcome	outcome_count
0	Failure (in flight)	1
1	Success	98
2	Success	1
3	Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
Empty DataFrame
Columns: [Booster_Version, PAYLOAD_MASS__KG_]
Index: []
```

Explanation

- I couldn't get it to display properly.

2015 Launch Records

Explanation

- January and April of 2015 were both failure (drone ship).

	Month	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 06.04.2010 and 03.20.2017

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

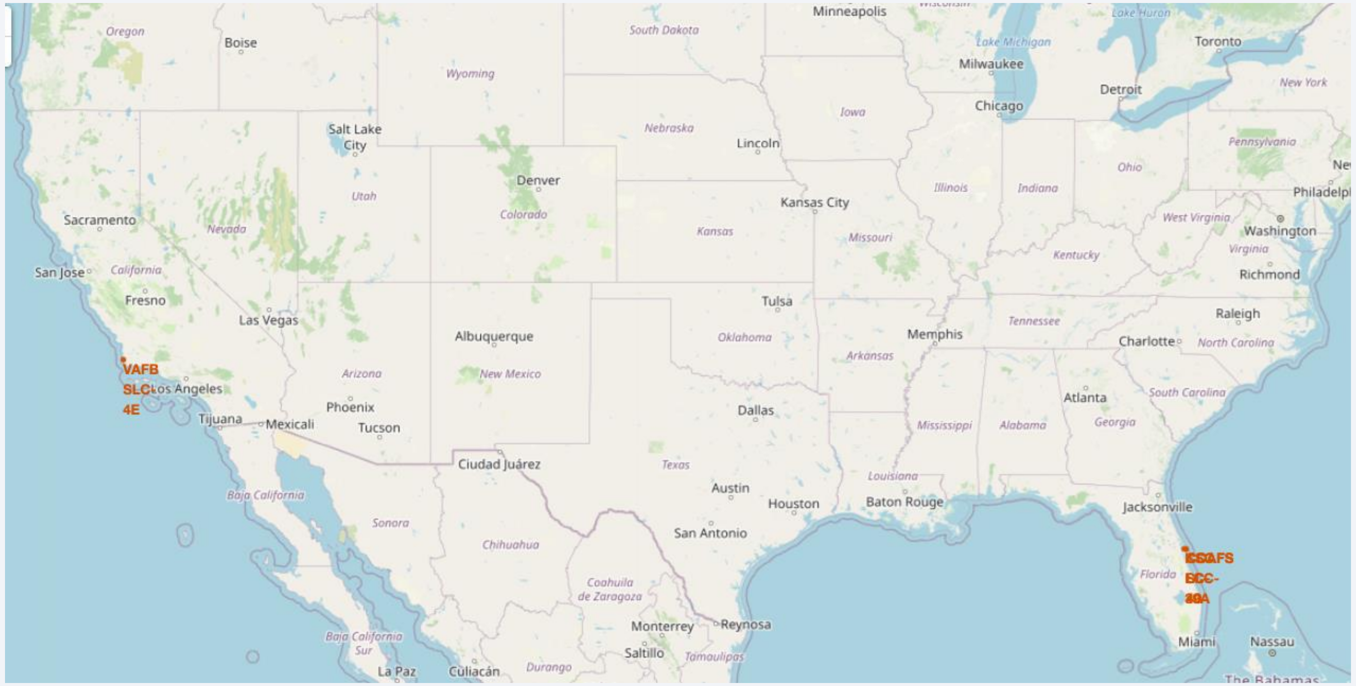
Explanation

- Between 06.04.2010 and 03.20.2017 there was
  - 5 success (drone ship)
  - 5 failure (drone ship)
  - 3 success (ground pad)
  - 3 controlled (ocean)
  - 2 uncontrolled (ocean)
  - 2 failure (parachute)
  - 1 precluded (drone ship)
  - 10 no attempts

Section 3

Launch Sites Proximities Analysis

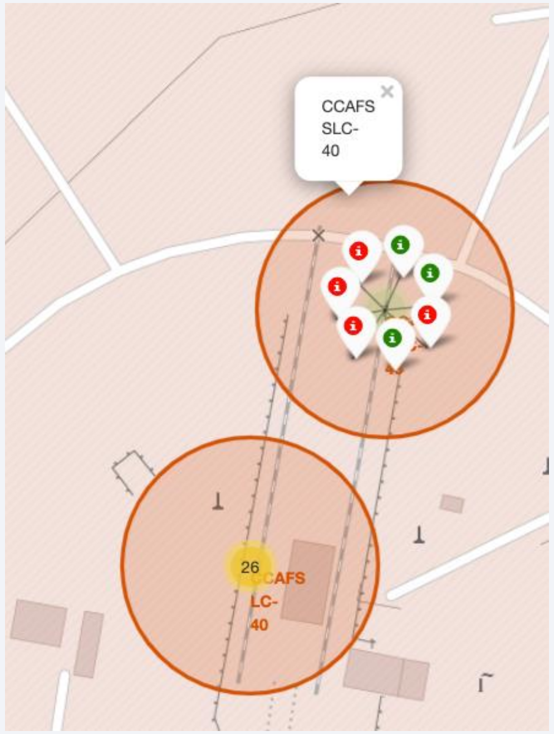
All launch sites' location markers on Global Map



Explanation:

- The launch sites are in proximity to the Equator line and nearby coastlines.

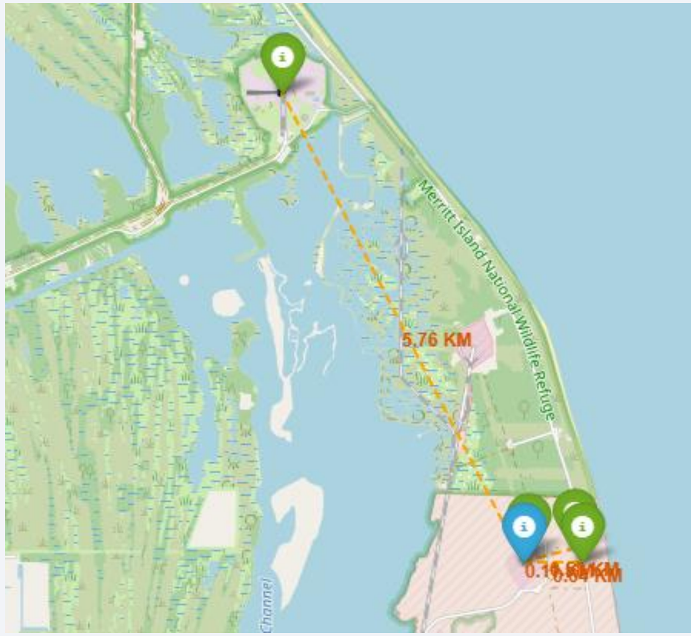
Colored label markers for success vs failure



Explanation:

- The successful launches are marked green and the failures are marked red.

Distance from the launch site its proximities.



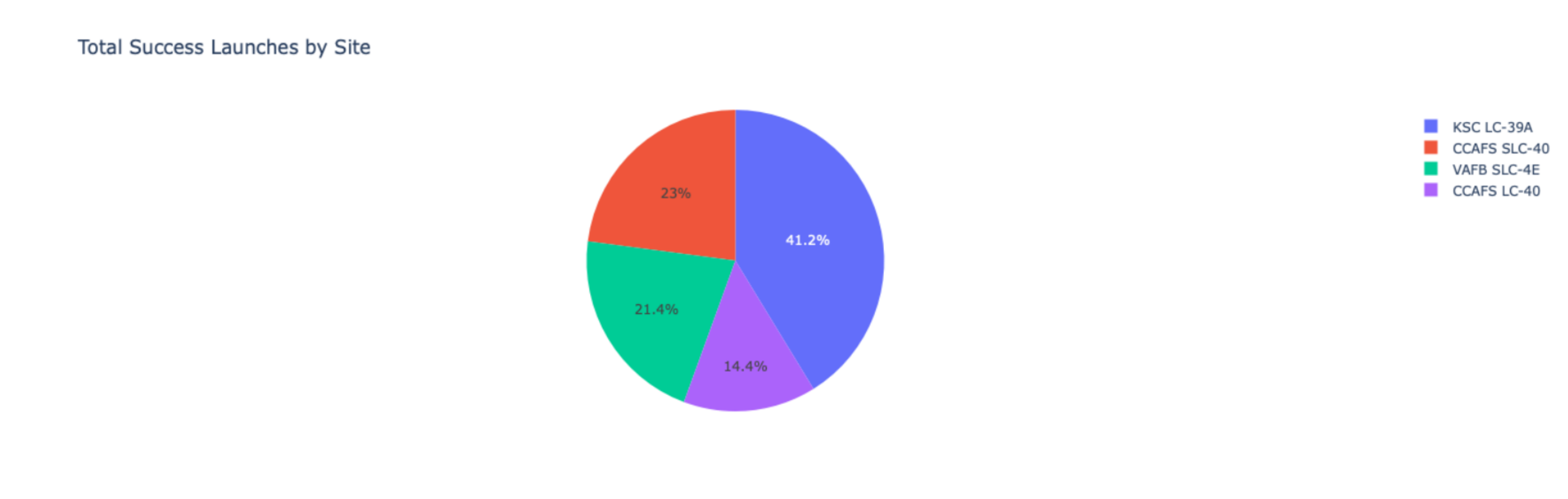
Explanation:

- The distance between Railways, Highways, Coastlines, and Nearest City to the launch site.

Section 4

Build a Dashboard with Plotly Dash

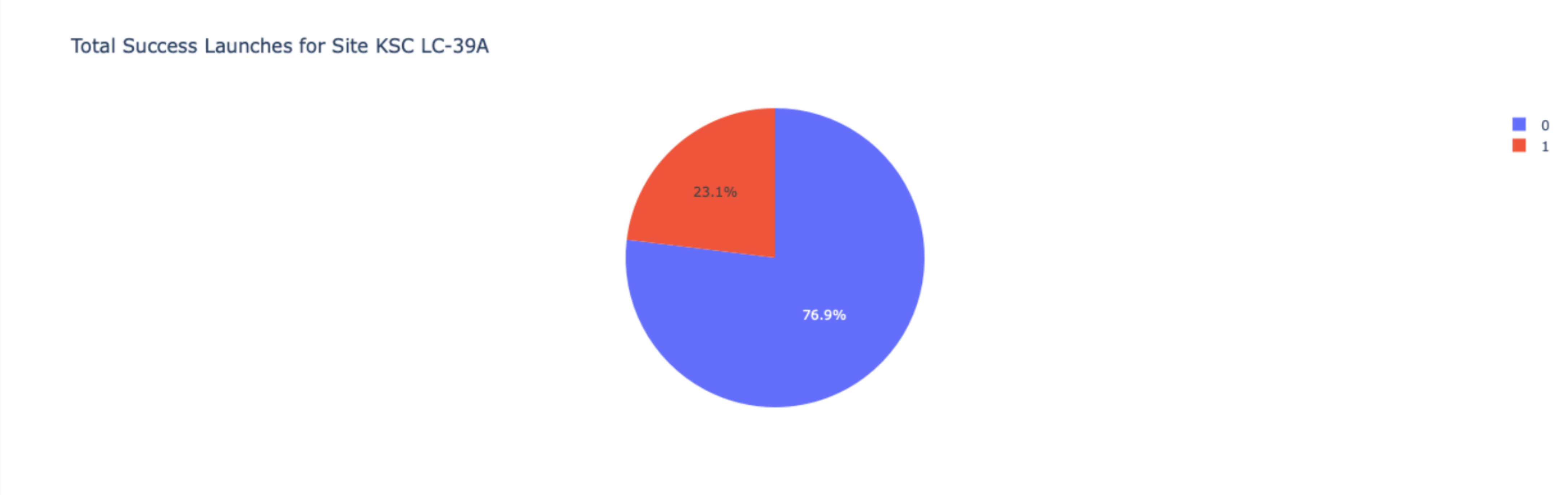
Launch success count for all sites



Explanation:

- Pie chart showing success rate for all sites.

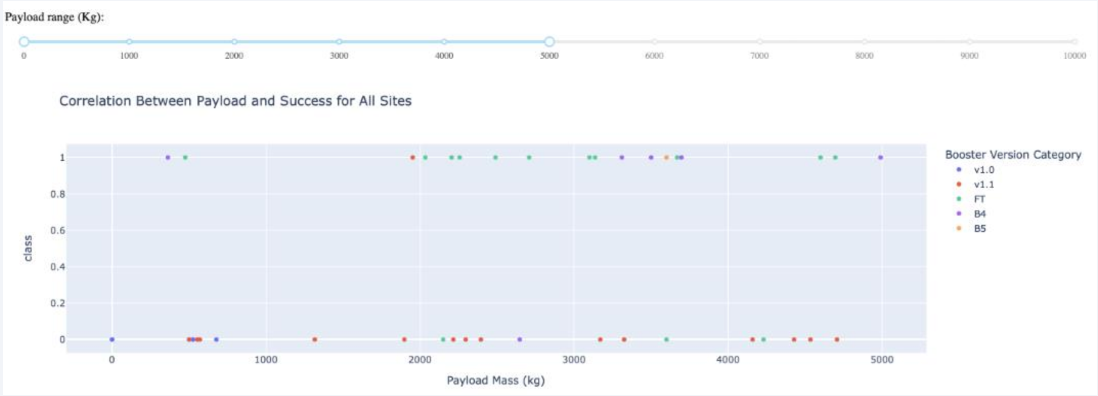
Successful Launches for Site KSC LC-39A



Explanation:

- Pie chart showing success rate for site KSC LC-39A.

Payload Vs. Lauch Outcome for all sites



Explanation:

- The charts show that payloads between 2000 kg and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- Based on the scores of the Test Set, I confirmed that Decision Tree was the best performing model with a 94% accuracy.
- The other models performed around 83% accuracy.

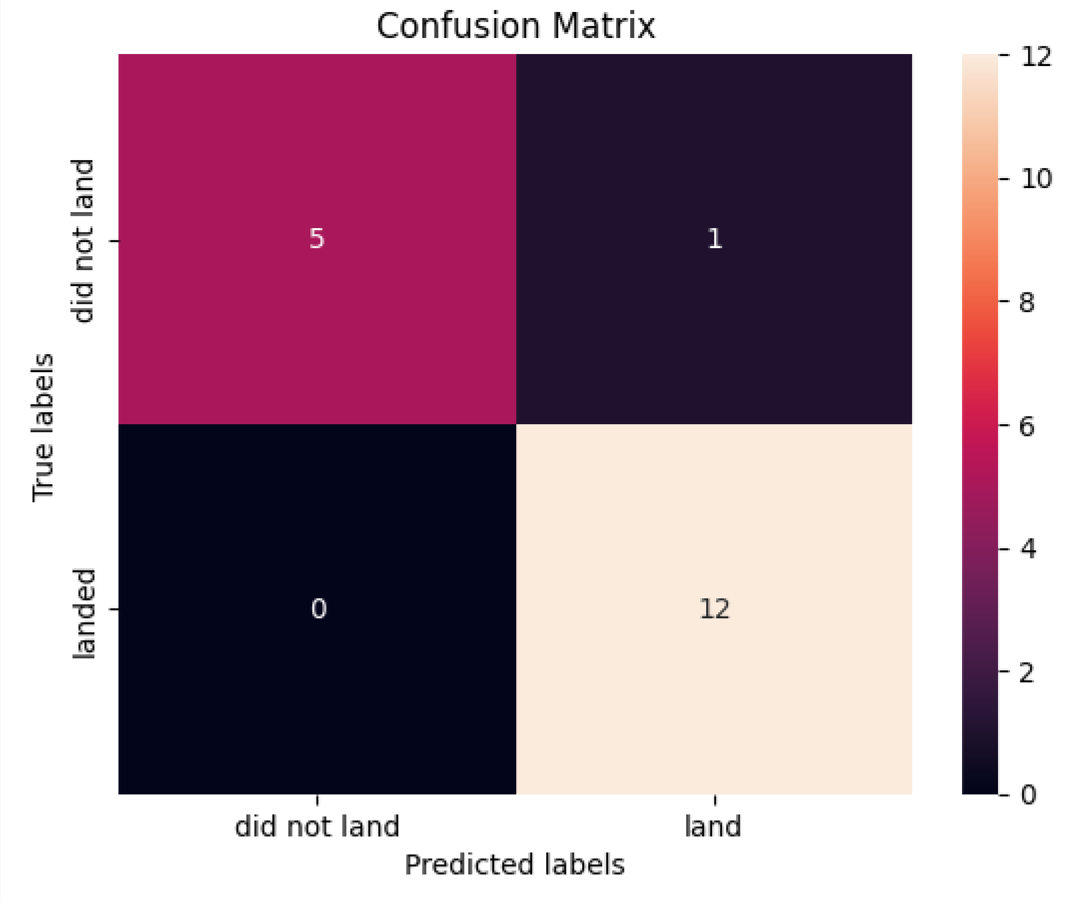
```
acc_logreg = logreg_cv.score(X_test, Y_test)
acc_svm = svm_cv.score(X_test, Y_test)
acc_tree = tree_cv.score(X_test, Y_test)
acc_knn = knn_cv.score(X_test, Y_test)

accuracies = {
    'Logistic Regression': acc_logreg,
    'SVM': acc_svm,
    'Decision Tree': acc_tree,
    'KNN': acc_knn
}

best_model = max(accuracies, key=accuracies.get)
best_accuracy = accuracies[best_model]

print(f"Best performing model: {best_model} with accuracy {best_accuracy:.4f}")
```

```
Best performing model: Decision Tree with accuracy 0.9444
```

Confusion Matrix

Explanation:

- The confusion matrix of the best performing model.



Conclusions

- Decision Tree Model has the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a higher payload mass.
- Most of the launch sites are in proximity to the Equator line and all sites are in very close proximity to a coastline.
- The success rate of launches has increased over the years.
- KSC LC-39A has the highest success rate of launches from all sites.
- Orbits ES-L1, GEO, HEO, and SSO have 100% success rates.

Appendix

Github Project Link: Data Capstone Project

Thank you!