IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eleanor Janine Juan

06 August 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data using SQL
  - Exploratory Data using Pandas and Matplotlib
  - Interactive Visual Analytics with Folium Lab
  - Interactive Visual Analytics and Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results

Introduction

- We are in the **commercial space age**, with companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX **making space travel more affordable**.
- SpaceX stands out by **reusing rocket parts**, advertising Falcon 9 launches at $62 million compared to $165 million from other providers.
- A significant portion of these **cost savings** comes from **reusing the first stage of the rocket**.

Introduction

SPACE Y

A new rocket company, Space Y, aims to compete with SpaceX.
To achieve this, Space Y needs to address several key objectives:

1. **Determine the Price of Each Launch.**

   Understanding the cost structure is essential for competitive pricing.

2. **Predict Falcon 9's First Stage Landing Success.**
   Can machine learning be used to predict if the Falcon 9's first stage will land and be reused? Accurate predictions of first stage landing success are crucial for determining overall launch costs.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Collecting data using the **SpaceX REST API** and **Web Scraping** from Wikipedia
- Perform data wrangling
    - Filtering, handling missing values, and structuring data for 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
    - Building, tuning and evaluating each model to achieve the best results

Data Collection

Rocket Launch Data is collected by

Requesting to the
SpaceX REST API

Web Scraping
additional Falcon 9 launch records from
Wikipedia with **BeautifulSoup**

All obtained data were stored in DataFrames using Pandas

Data Collection – SpaceX API

1. Request to SpaceX Launch Data to **SpaceX API**
2. Decode the response content as a **JSON**
3. Turn the JSON into **Pandas DataFrame**
4. Request to the API again to get more information about the launches using the previously obtained IDs
5. **Construct the dataset** using the data obtained by combining columns into a dictionary
6. Convert the dictionary to **Pandas DataFrame**

Data Collection using SpaceX API

[GitHub Link]()

Data Collection - Scraping

Request the **Falcon9 Launch Wiki** page from its URL

Create a **BeautifulSoup** object from the HTML response

Extract all **variable names** from the HTML table header

Create an **empty dictionary** with keys from the extracted variable names

Fill the dictionary with data parsed from HTML tables

Convert the dictionary to **Pandas DataFrame**

Data Collection with Web Scraping

[GitHub Link]

Data Wrangling

- Basic data processing was done on the previous two notebooks which involved
  - Filtering data, extracting only relevant data
  - Dealing with Missing Values
- In this notebook, we mainly simplify and standardize data for analysis and model training. We do this by performing **Exploratory Data Analysis and Determine Training Labels**
  - Convert several landing outcomes into just two outcome labels
    - 1: Successful landing (any location)
    - 0: Unsuccessful landing (any location)

- **Calculate the number of launches on each site**
- **Calculate the number and occurrence of each orbit**
- **Calculate the number and occurence of mission outcome of the orbits**
- **Create a landing outcome label from Outcome column**

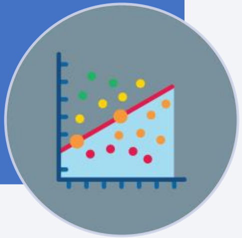
Data Wrangling
[GitHub Link]

EDA with Data Visualization

Different charts were used to make the data easier to understand.

- To visualize the relationship b/n:
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - FlightNumber and Orbit type
  - Payload Mass and Orbit type

Scatter Plot



- To visualize the relationship b/n Success Rate of Each Orbit Type

Bar Chart

- To visualize the Launch Success Yearly Trend

Line Plot

Exploratory Data Analysis with Visualizations

[GitHub Link]()

EDA with SQL

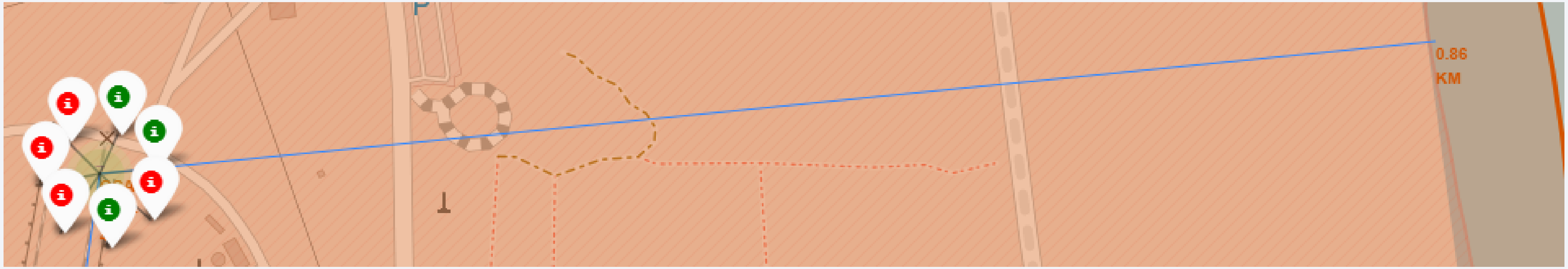
Summary of the SQL queries that were performed

- 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 the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Exploratory Data Analysis with SQL

GitHub Link

Build an Interactive Map with Folium

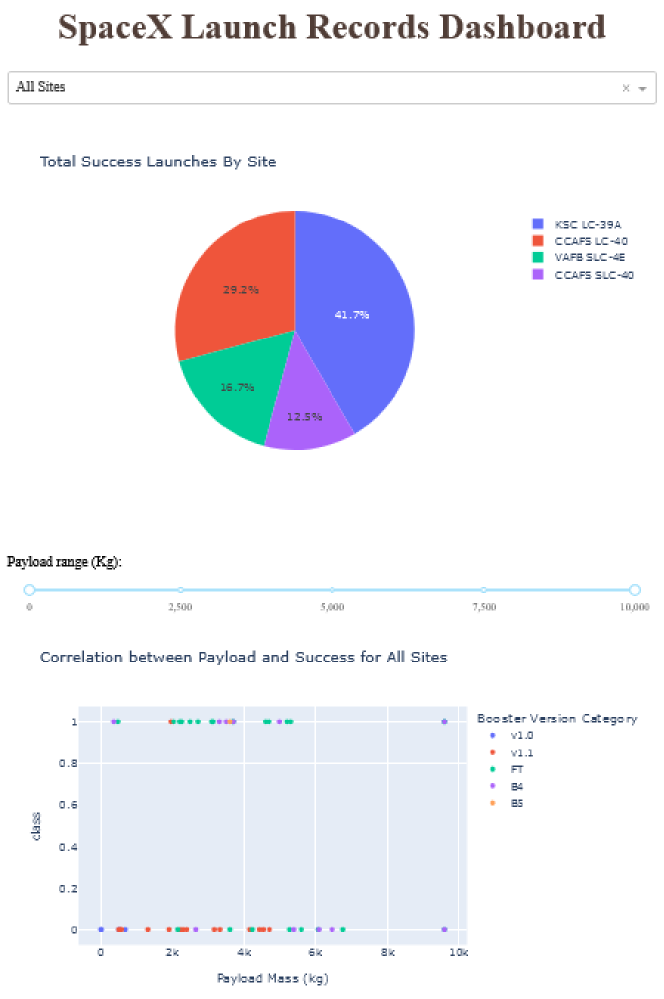


- Circles and Markers were used to mark launch sites on the map
- Markers were used to mark success/failed launches for each launch site
- Markers and Polylines were used to mark closest coastline, railway, highway, and city to the launch site

Interactive Visual Analytics with Folium
[GitHub Link]

Build a Dashboard with Plotly Dash

- Added a **dropdown list** for Launch Site selection
- Added a **pie chart** to show the total successful launches count for all and specific sites
- Added a **slider** to select payload range
- Added a **scatter chart** to show the correlation between payload and launch success count for all and specific sites and selected range



Interactive Visual Analytics with Folium
[GitHub Link]

Predictive Analysis (Classification)

- **Scikit-learn Library** was used to implement several Machine Learning Models
  - Logistic Regression
  - Support Vector Machine
  - Classification Trees
  - K-Nearest Neighbors
- The method that perform best was also determined

For each model:

Create a column for the class

Standardized the data

Split data into training and test data

Find the best Hyperparameter using GridSearchCV and train data

Calculate the accuracy on the test data

Machine Learning Prediction
GitHub Link

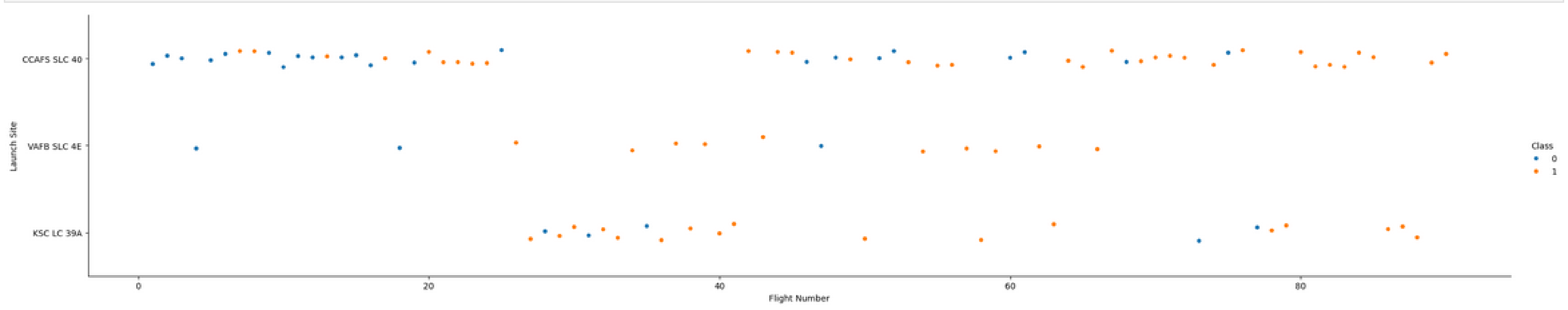
Results

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

Section 2

Insights drawn from EDA

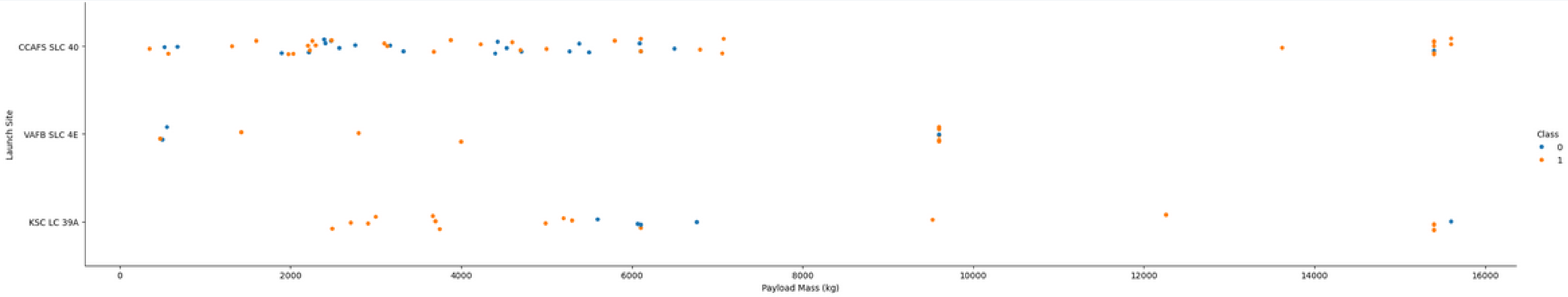
Flight Number vs. Launch Site



Findings:

- Launch Site CCAFS SLC 40 has the most launches out of all the sites (55/90) and successful landings (33/60).
- Launch Site KSC LC 39A has the highest success rate (77.2%), followed by VAFB SLC 4E (76.9%).
- The earliest flights failed, but the success rate improved over time.

Payload vs. Launch Site



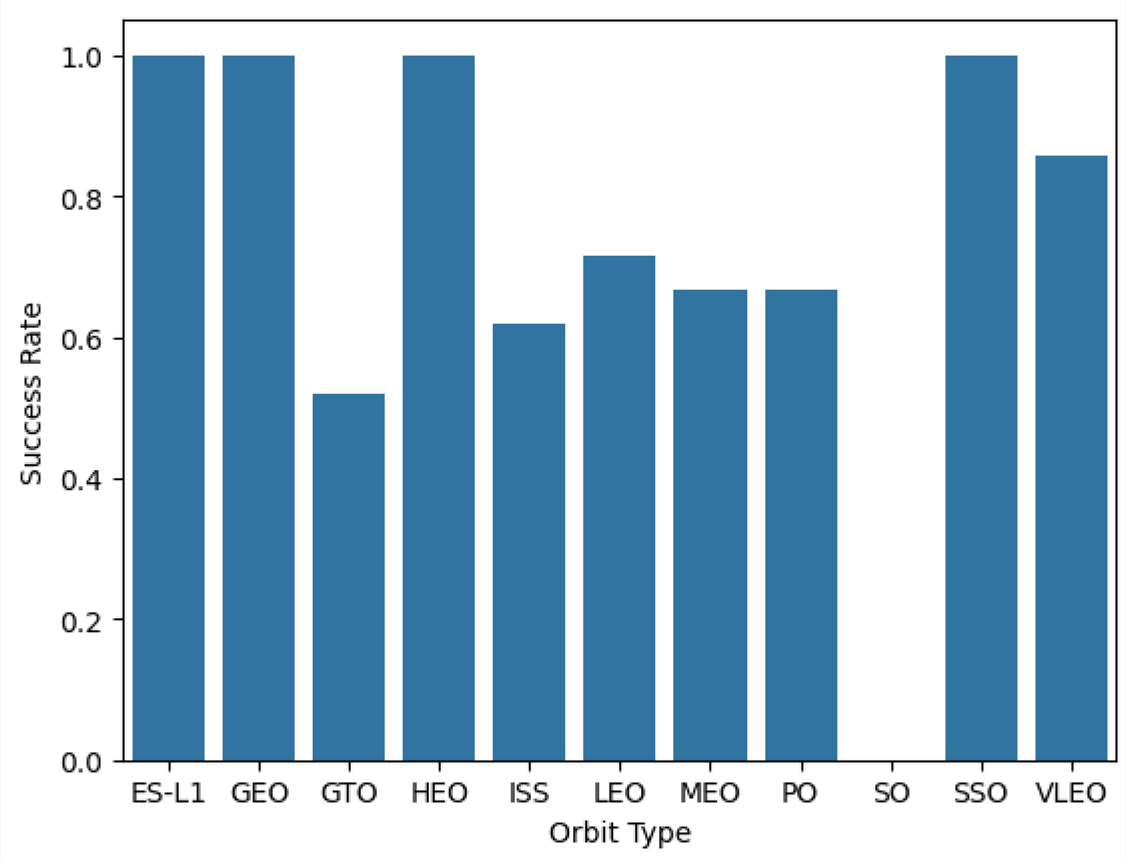
Findings:

- Most of the launches with payload mass over 8000 kg were successful.
- All the launches from KSC LC 39A with payload mass less than 5500 kg were successful.
- Most of the launches were in CCAFS SLC 40 and in under a payload mass of 8000kg.
- VAFB SLC 4E has no rockets launched for heavy payload mass (greater than 10000).

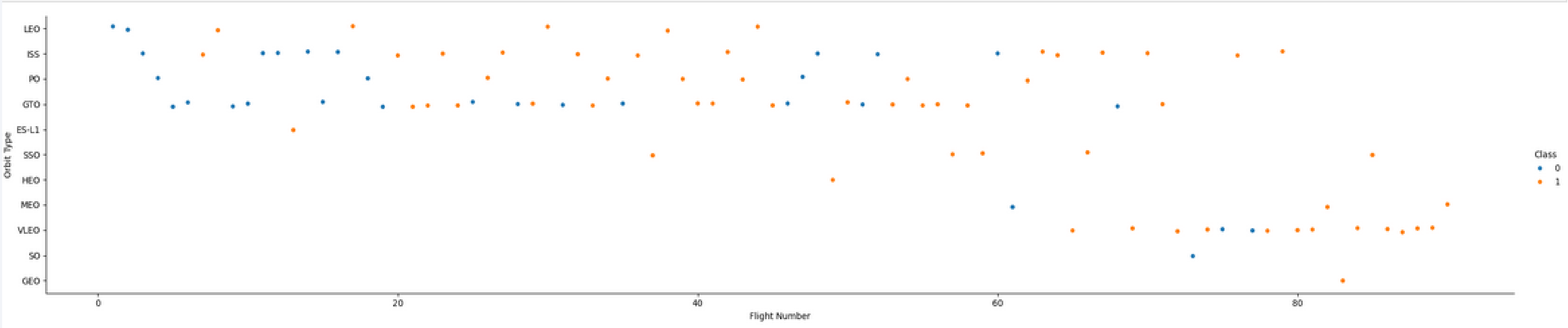
Success Rate vs. Orbit Type

Findings:

- The orbit type with 100% success rate are:
    - ES-L1, GEO, ISS, SSO
- The orbit type with 0% success rate is:
    - SO



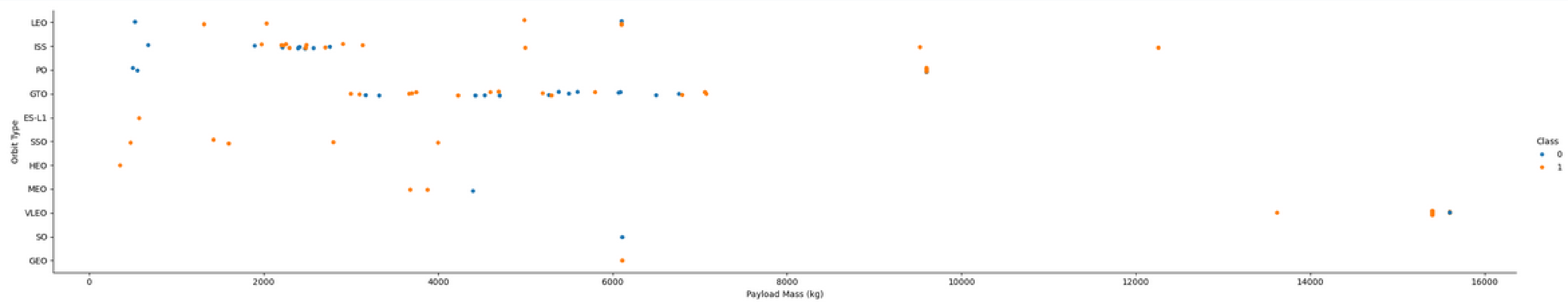
Flight Number vs. Orbit Type



Findings:

- In the LEO orbit, success seems to be related to the number of flights
- In the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



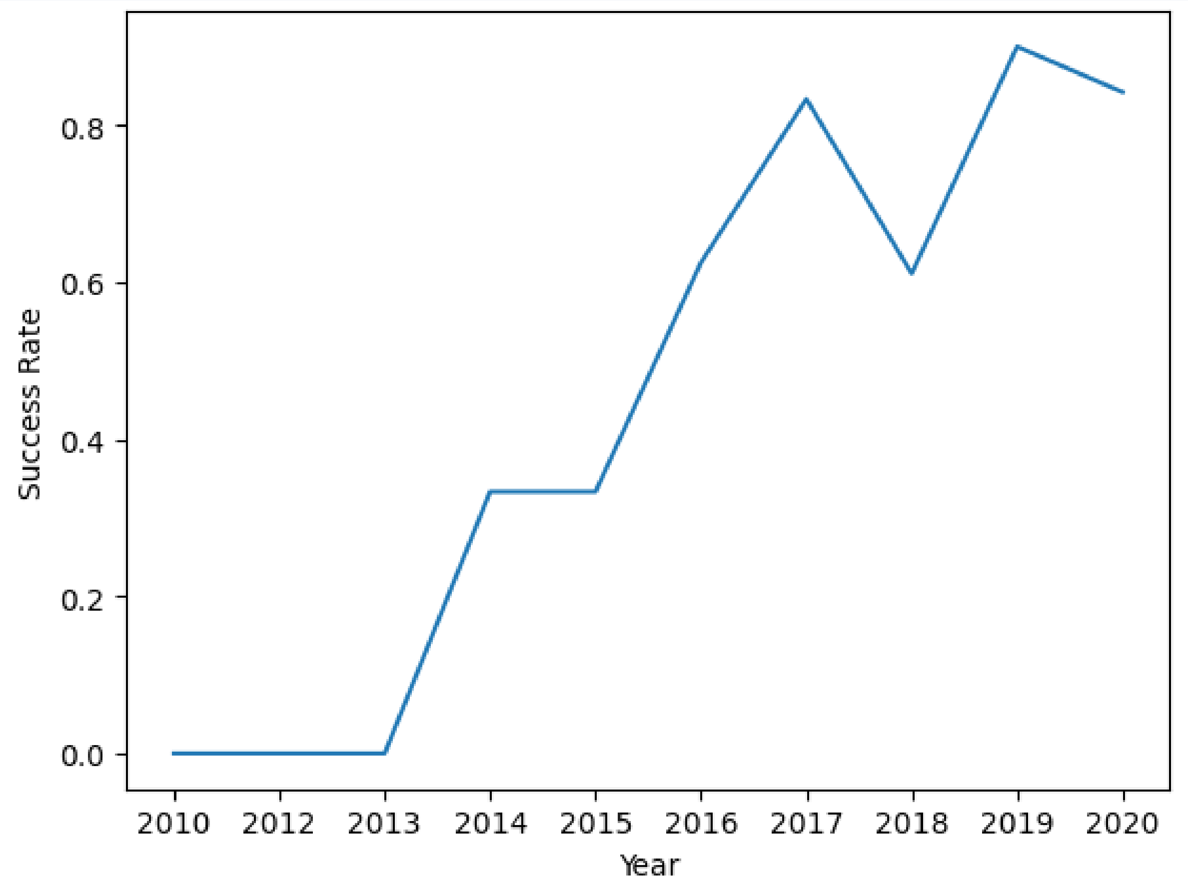
Findings:

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.
- The highest payloads are for flights to VLEO.

Launch Success Yearly Trend

Findings:

- Success rate since 2013 kept increasing till 2020



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

- Query and output for determining all the unique rocket launch sites

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

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

- Query and output for getting the first five records from launch sites beginning with ‘CCA’

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

- Query and output for determining the total payload mass carried by boosters launched by NASA (CRS)

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

- Query and output for determining the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql SELECT Date FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' ORDER BY Date LIMIT 1
```

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

Date
2015-12-22

- Query and output for determining the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

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

- Query and output for listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Query and output for calculating the total number of successful and failure mission outcomes

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

- Query and output for listing the names of the booster which have carried the maximum payload mass

2015 Launch Records

- Query and output for listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- SQLite does not have a built-in MONTHNAME function unlike other SQL databases so we make use of substr and CASE

```
%%sql SELECT
    CASE
        WHEN substr(Date, 6, 2) = '01' THEN 'January'
        WHEN substr(Date, 6, 2) = '02' THEN 'February'
        WHEN substr(Date, 6, 2) = '03' THEN 'March'
        WHEN substr(Date, 6, 2) = '04' THEN 'April'
        WHEN substr(Date, 6, 2) = '05' THEN 'May'
        WHEN substr(Date, 6, 2) = '06' THEN 'June'
        WHEN substr(Date, 6, 2) = '07' THEN 'July'
        WHEN substr(Date, 6, 2) = '08' THEN 'August'
        WHEN substr(Date, 6, 2) = '09' THEN 'September'
        WHEN substr(Date, 6, 2) = '10' THEN 'October'
        WHEN substr(Date, 6, 2) = '11' THEN 'November'
        WHEN substr(Date, 6, 2) = '12' THEN 'December'
    END AS Month,
    Landing_Outcome, Booster_Version, Launch_Site
    FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Failure (drone ship)'
    AND substr(Date,0,5) = '2015'
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Query and output for ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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

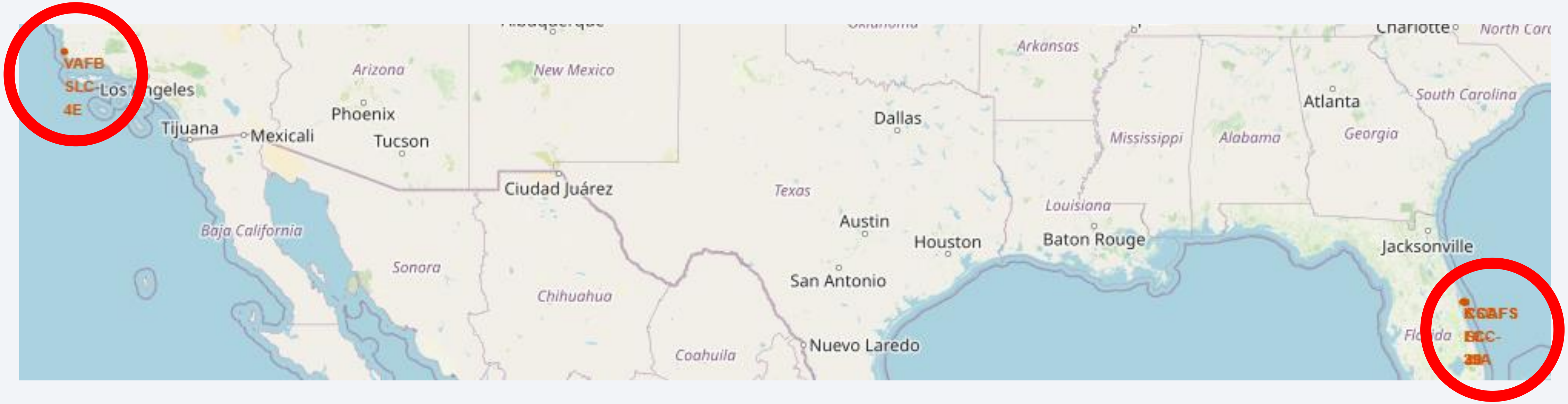
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Count(Landing_Outcome)
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

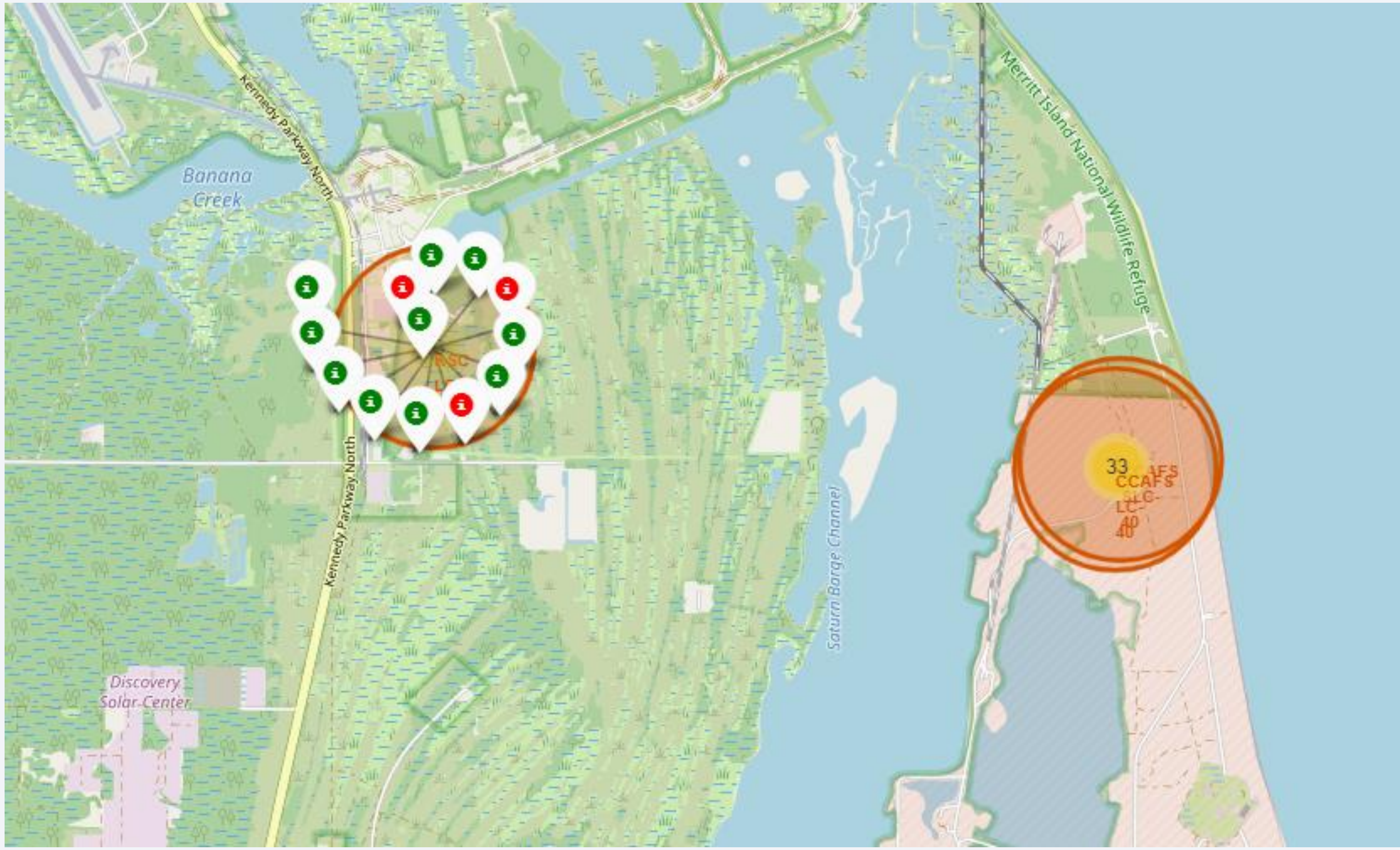
Launch Sites Proximities Analysis

Launch Site Locations



- Launch sites are **close to the equator** because the Earth's rotation is fastest there. The surface speed at the equator is approximately 1670 km/hour, providing an initial boost to the spacecraft, aiding it in reaching orbit due to inertia.
- Launch sites are **near the coast** to minimize risks. Launching rockets over the ocean reduces the danger of debris or explosions affecting populated areas.

Rocket Launches on Site KSC LC 39A



successful

failure

- Successful launches are indicated by a green marker, while a red marker denotes failed rocket launches.

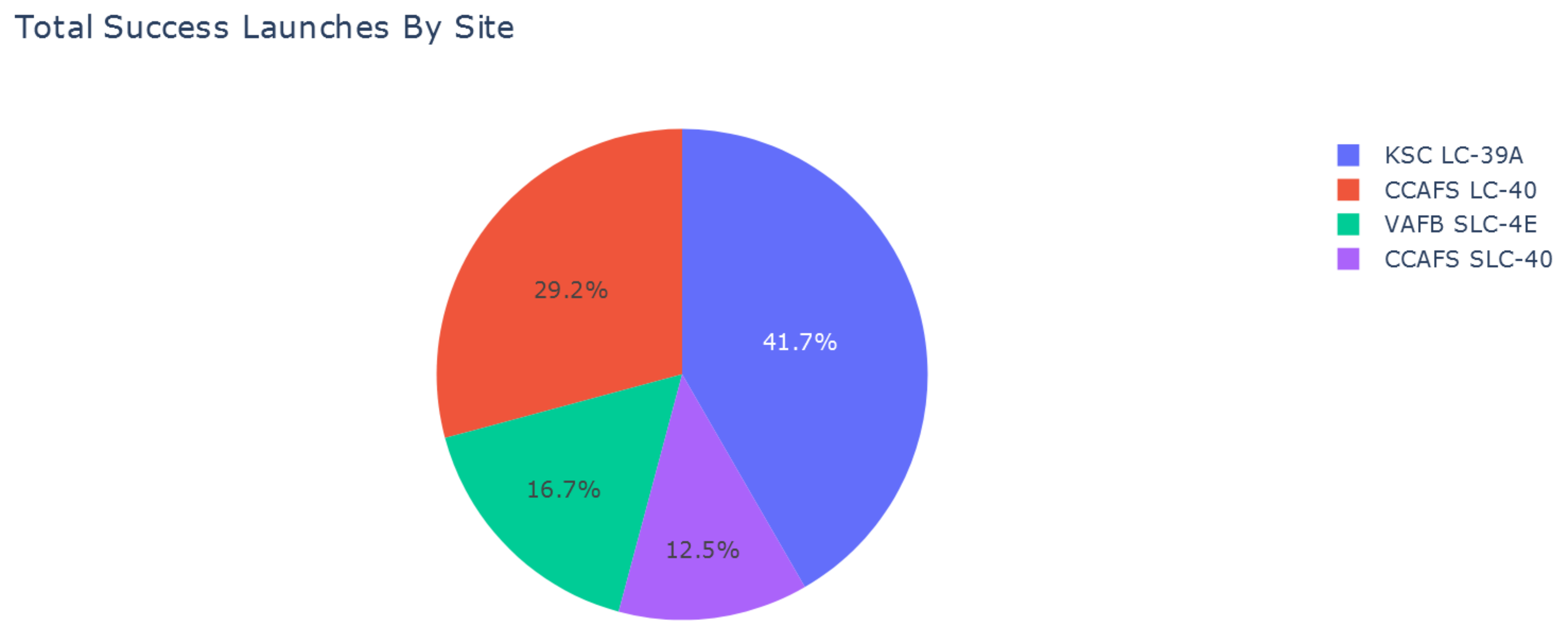
Proximity of Site CCAFS SLC 40 to Landmarks

- Launch sites are typically located **far from urban areas**, likely to reduce the risk of accidents near populated regions.
- Launch sites are often **near railways and highways**, which facilitates the transportation of rocket components.
- Launch sites are situated **near coastlines**, as demonstrated by numerous rocket landing tests conducted over bodies of water like the ocean.
- Included screenshots are the closest coastline, railway, highway and city to site CCAFS SLC 40.

Section 4

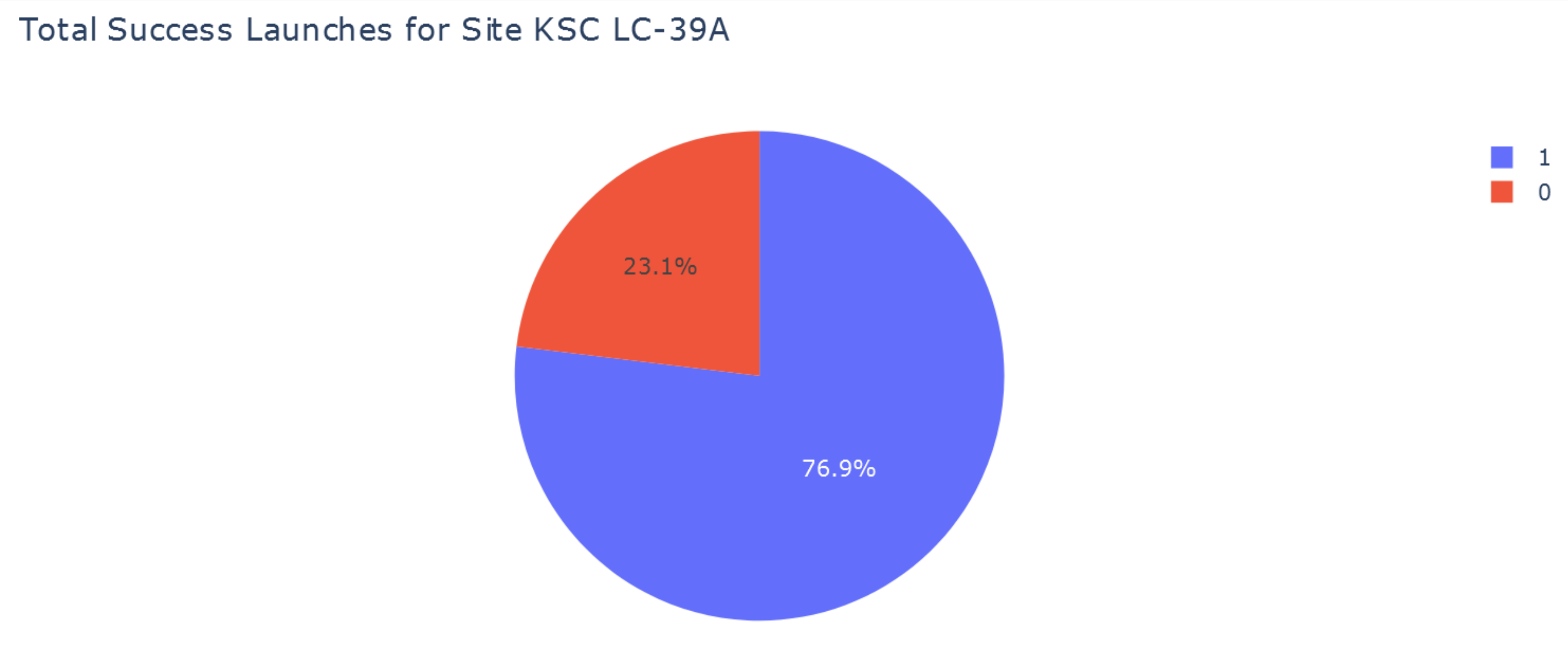
Build a Dashboard with Plotly Dash

Successful Launches by Site



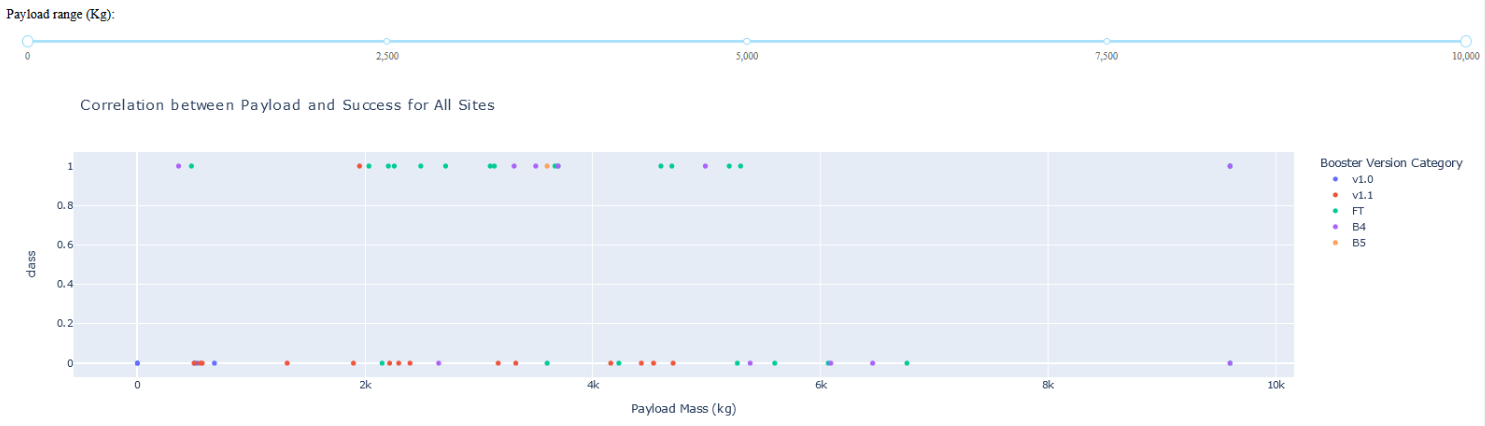
- We can see that KSC LC-39A has the most successful launches

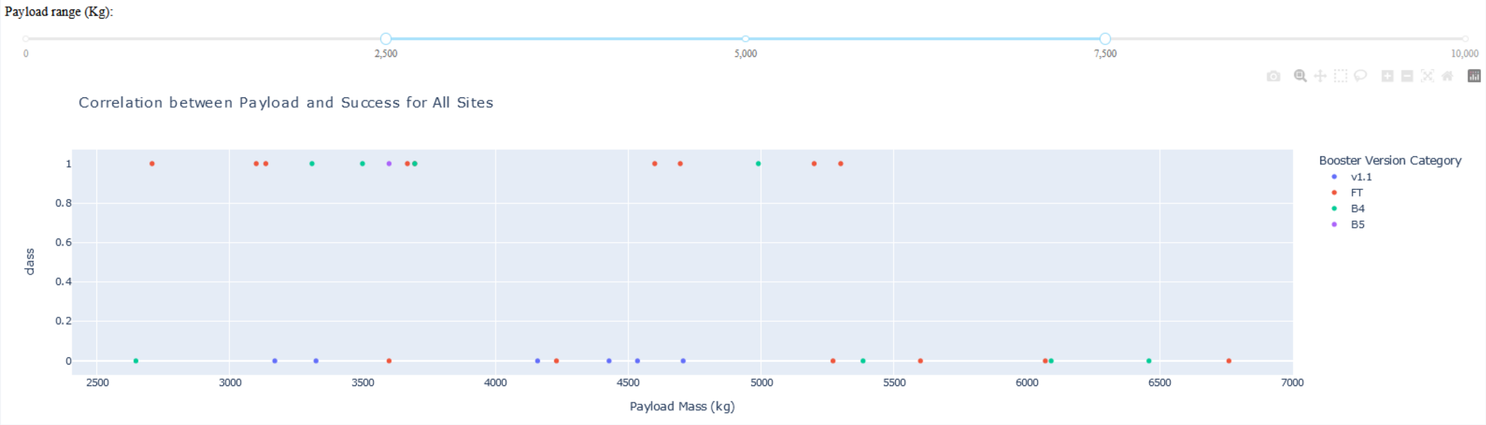
The Launch Site with the Highest Launch Success Ratio



- We can see that the success rate for Site KSC LC 39A is 76.9%, the highest among all sites with 10 successful landings out of 13 attempts

Payload vs. Launch Outcome Scatter Plot for All Sites



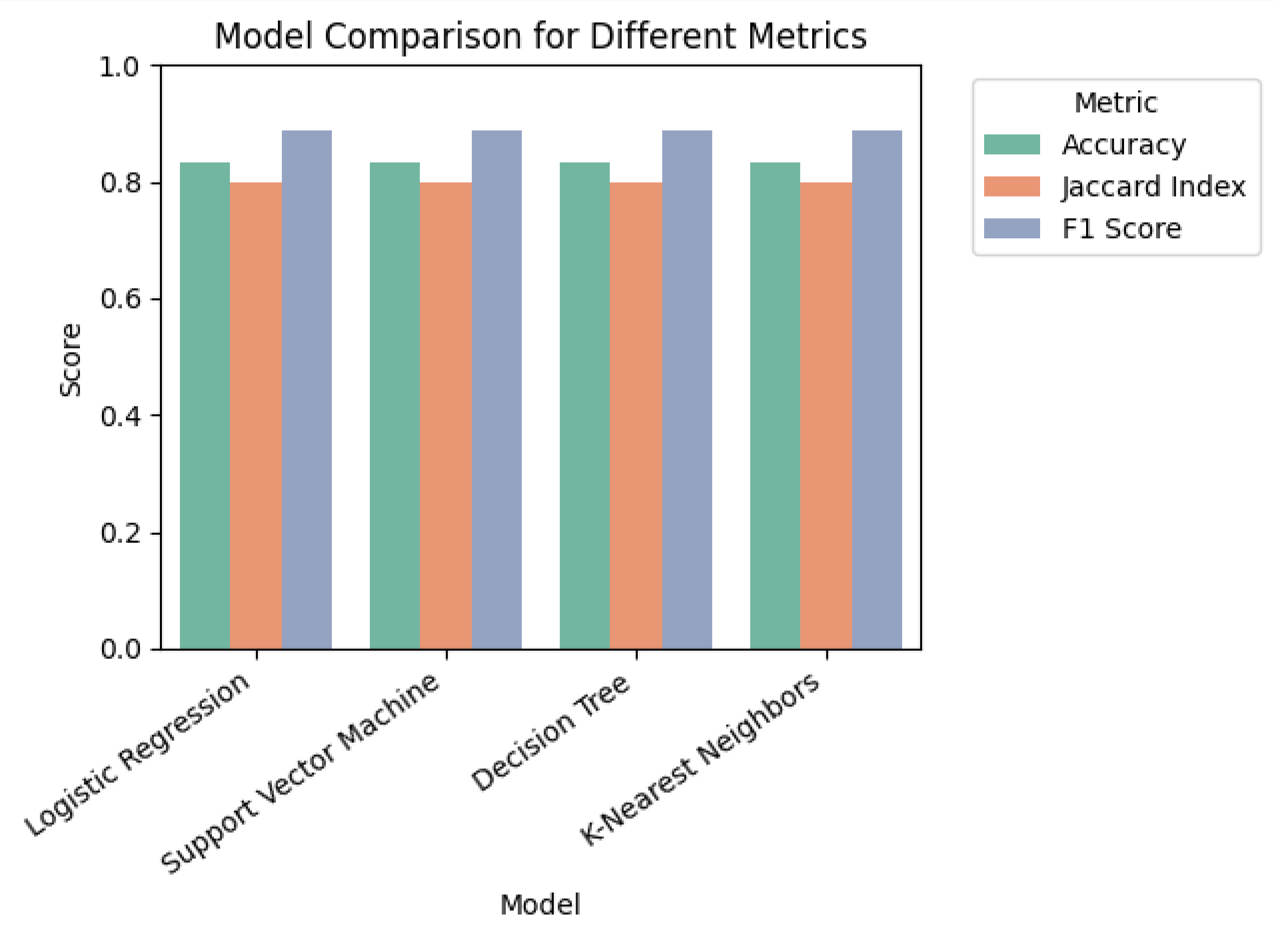


- The payload range of 2000 kg to 4000 kg shows the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

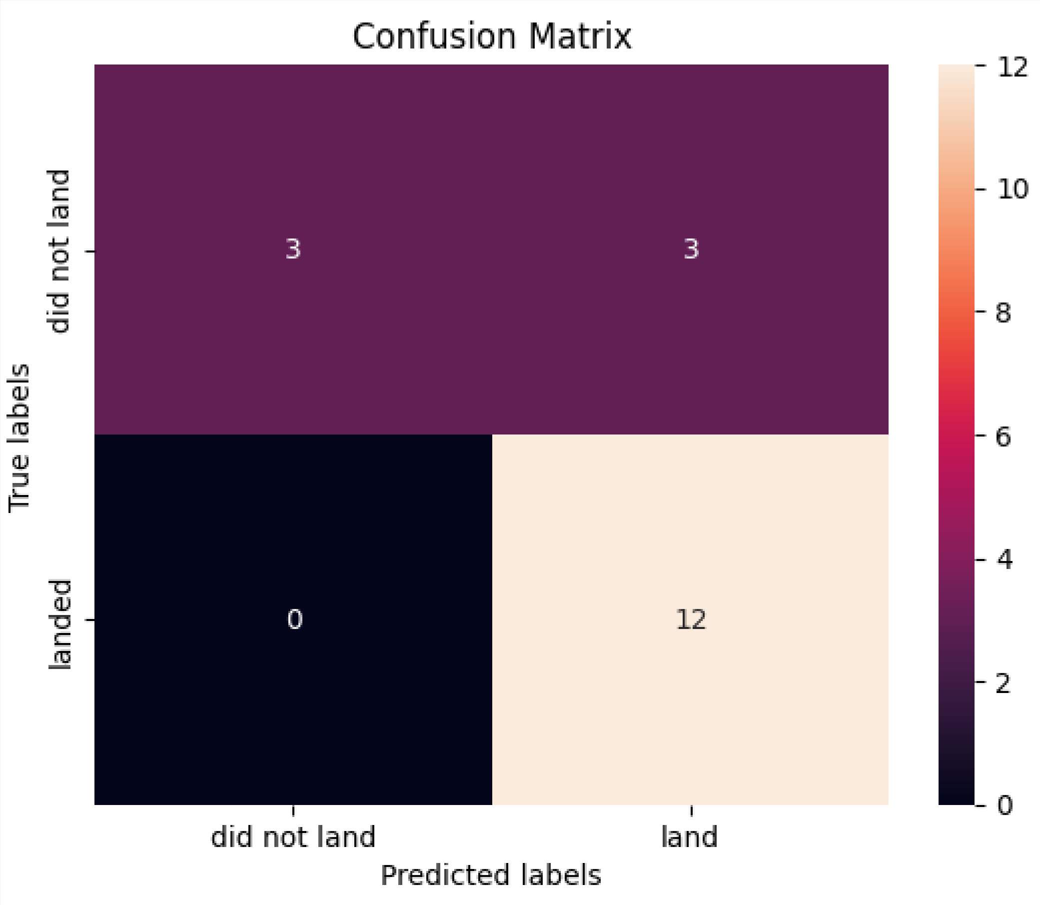


	Model	Accuracy	Jaccard Index	F1 Score
0	Logistic Regression	0.833333	0.8	0.888889
1	Support Vector Machine	0.833333	0.8	0.888889
2	Decision Tree	0.833333	0.8	0.888889
3	K-Nearest Neighbors	0.833333	0.8	0.888889

- All models performed equally, achieving identical Accuracy Scores, Jaccard Indices, and F1 Scores.

Confusion Matrix

- All models produced the identical confusion matrix.
  - All models achieve a 100% recall rate, correctly identifying all positive cases.
  - Precision is 80%, indicating that 80% of the positive predictions are correct.
  - At 50%, the model has an equal number of false positives and true negatives, suggesting the models may be over-predicting the positive class.



Conclusions

- The success rates of launches have improved over the years.
- Orbits ES-L1, GEO, HEO, and SSO each have a 100% success rate.
- Launch sites are located near the equator, coastlines, railways, highways, and are distant from populated areas.
- Site KSC LC 39A achieved the highest launch success rate among all the launch sites.
- None of the models outperformed the others, indicating that our data may be insufficient.

Thank you!