IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Victor Diirr
November 6th, 2024

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Visual Analytics with Folium
  - Predictive Analysis with Machine Learning
- **Summary of all results**
  - Exploratory Data Analysis
  - Interactive Analytics in Screenshots
  - Predictive Analytics Results

Introduction

- **Project background and context**

SpaceX advertises Falcon 9 rocket launches on its website at a cost of $62 million. In comparison, other providers charge upwards of $165 million per launch. Much of this savings is due to SpaceX's ability to reuse the first stage of the rocket. Therefore, by being able to predict whether the first stage will land successfully, we can determine the true cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Problems you want to find answers**

  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be fulfilled to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  - Data was gathered using the SpaceX API, supplemented by web scraping for additional details.

- Perform data wrangling

  - The data processing involved cleaning and transforming the raw data, removing inconsistencies, and formatting the information appropriately for analysis.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

- Building Classification Models: Constructing models using various algorithms suitable for classification tasks.

- Tuning Models: Optimizing model performance by adjusting hyperparameters to achieve the best results.

- Evaluating Models: Assessing model effectiveness through metrics such as accuracy, precision, recall, and F1-score to ensure robust predictions for future landing outcomes.

Data Collection

- **Request to the SpaceX API**

  - Use a GET request to gather data from the SpaceX API.

  - To ensure consistent JSON results for this project, was utilized a static response object.

  - Decode the response content as JSON using .json() and convert it into a Pandas DataFrame using .json_normalize().

- **Data Cleaning**

  - Inspect the obtained DataFrame for missing values.

  - Replace missing values as needed to maintain data integrity.

- **Web Scraping Falcon 9 Launch Records with BeautifulSoup**

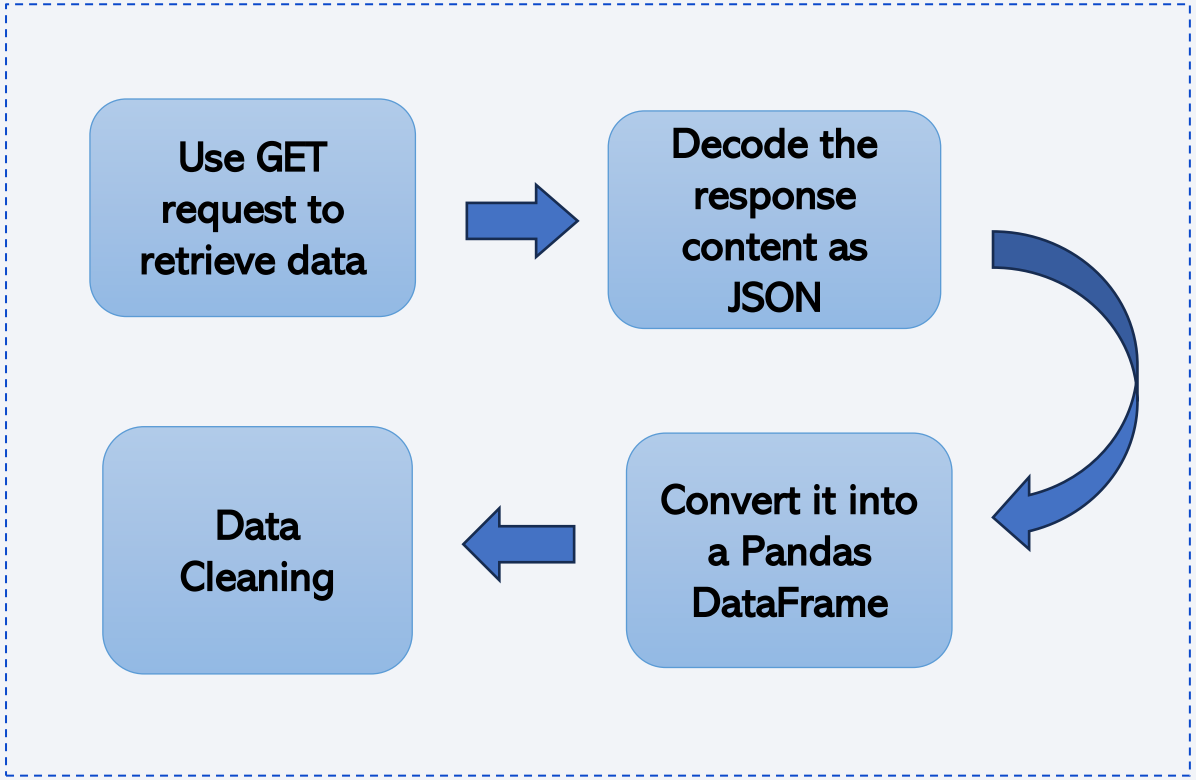
  - Extract the HTML table of Falcon 9 launch records from Wikipedia.

  - Parse the table and convert it into a Pandas DataFrame.

Data Collection – SpaceX API

- Use GET request to retrieve data
- Use Json on a static response object and convert into pandas DataFrame
- Check the data for missing values, replace than to maintain integrity
- GitHub link:

[GitHub - Data Collection]

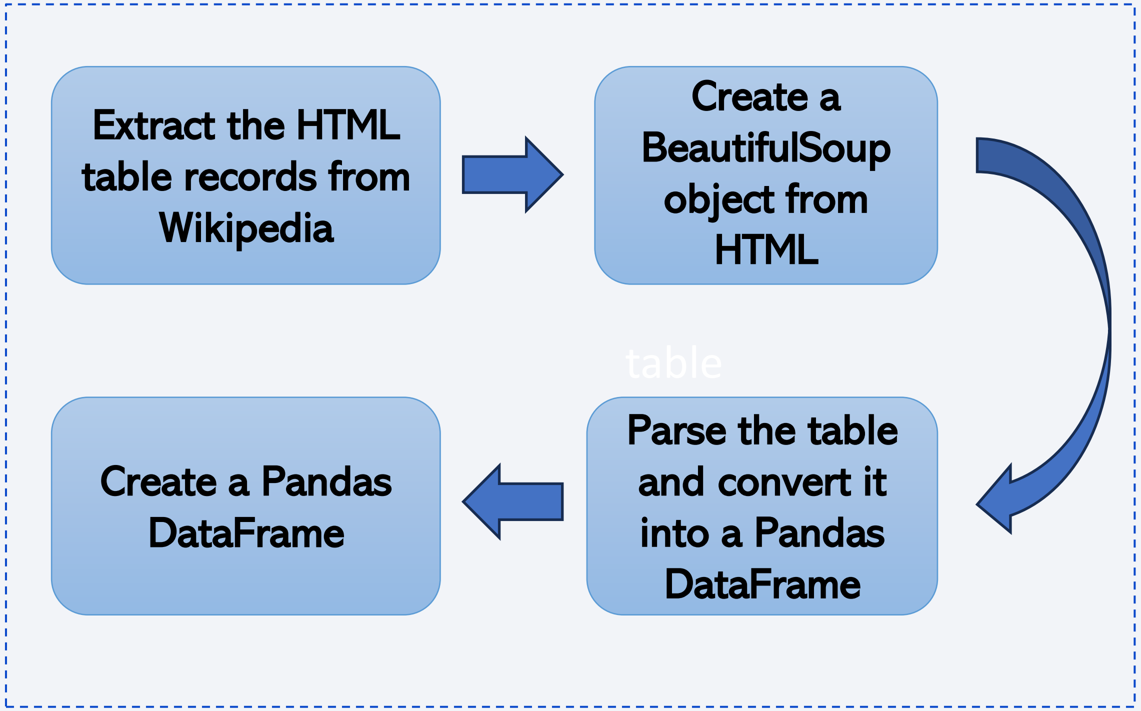


Data Collection - Scraping

- Extract HTML table records of Falcon 9 Launch,
- Parse the table and convert into a Data Frame

- GitHub link:

GitHub - Data Scraping

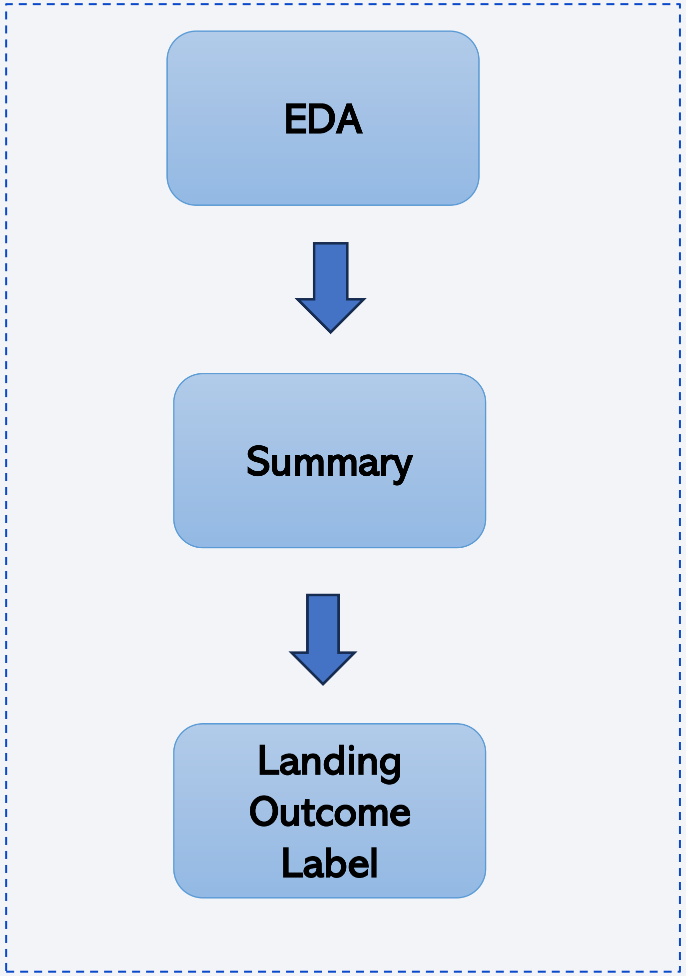


Data Wrangling

- **Exploratory Data Analysis (EDA)**
- **Summary Launch and Orbit Analysis**
- **Label Creation and Export**

- **GitHub link:**

[GitHub - Data Wrangling]()



EDA with Data Visualization

- **The data was explored by visualizing the relationship between:**

  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - success rate of each orbit type
  - FlightNumber and Orbit type
  - Payload Mass and Orbit type
  - Visualize the launch success yearly trend

- **GitHub link:**

[GitHub - Data Visualization]

EDA with SQL

- **The following SQL queries 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.

- **GitHub link:**

GitHub - SQL

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

- **GitHub link:**

GitHub - Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Add a Launch Site Drop-down Input Component
- Add a callback function to render success-pie-chart based on selected site dropdown
- Add a Range Slider to Select Payload
- Add a callback function to render the success-payload-scatter-chart scatter plot

- GitHub link:

GitHub - Dashboard with Ploty Dash

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
- Create a column for the class
- Standardize the data
- Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data

- GitHub link:

GitHub - Predictive Analysys

Results

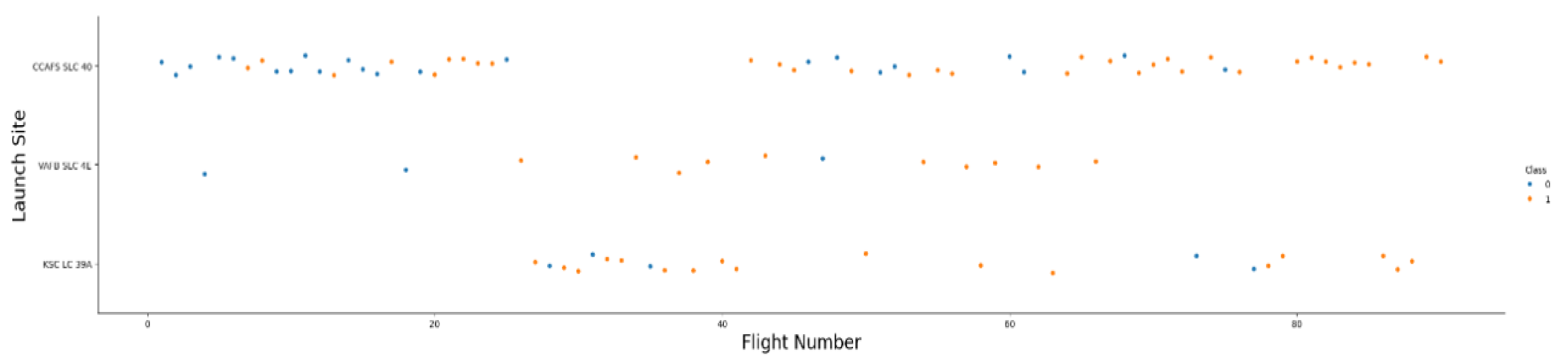
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

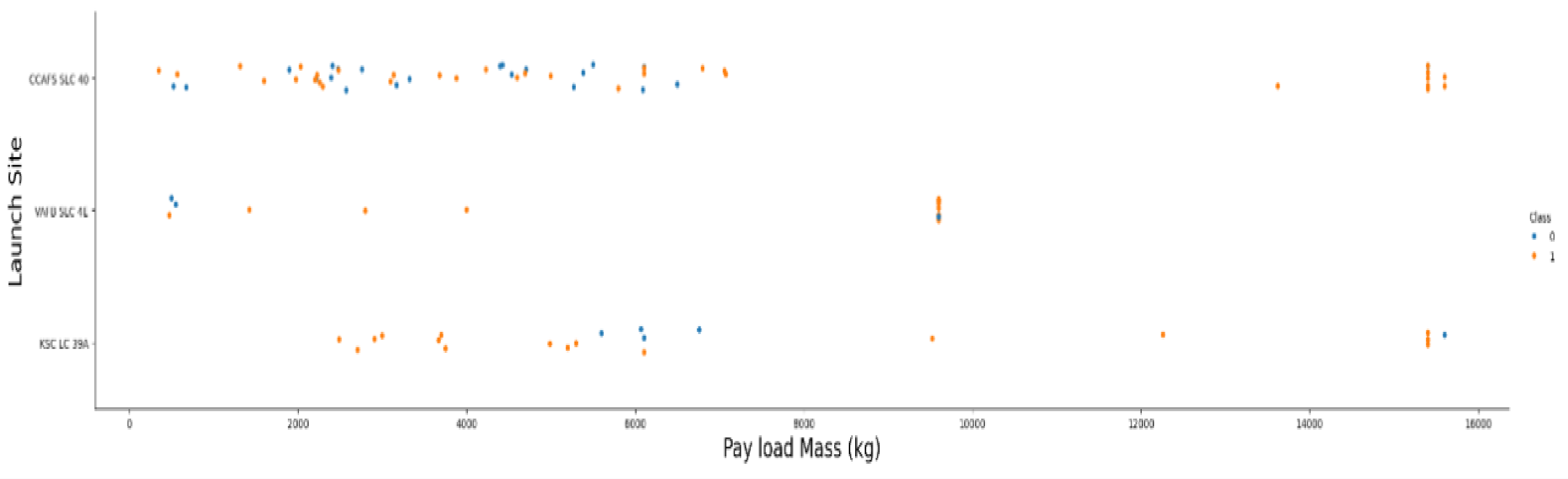
Flight Number vs. Launch Site

The larger the flight amount at a launch site, the greater the success rate at a launch site.



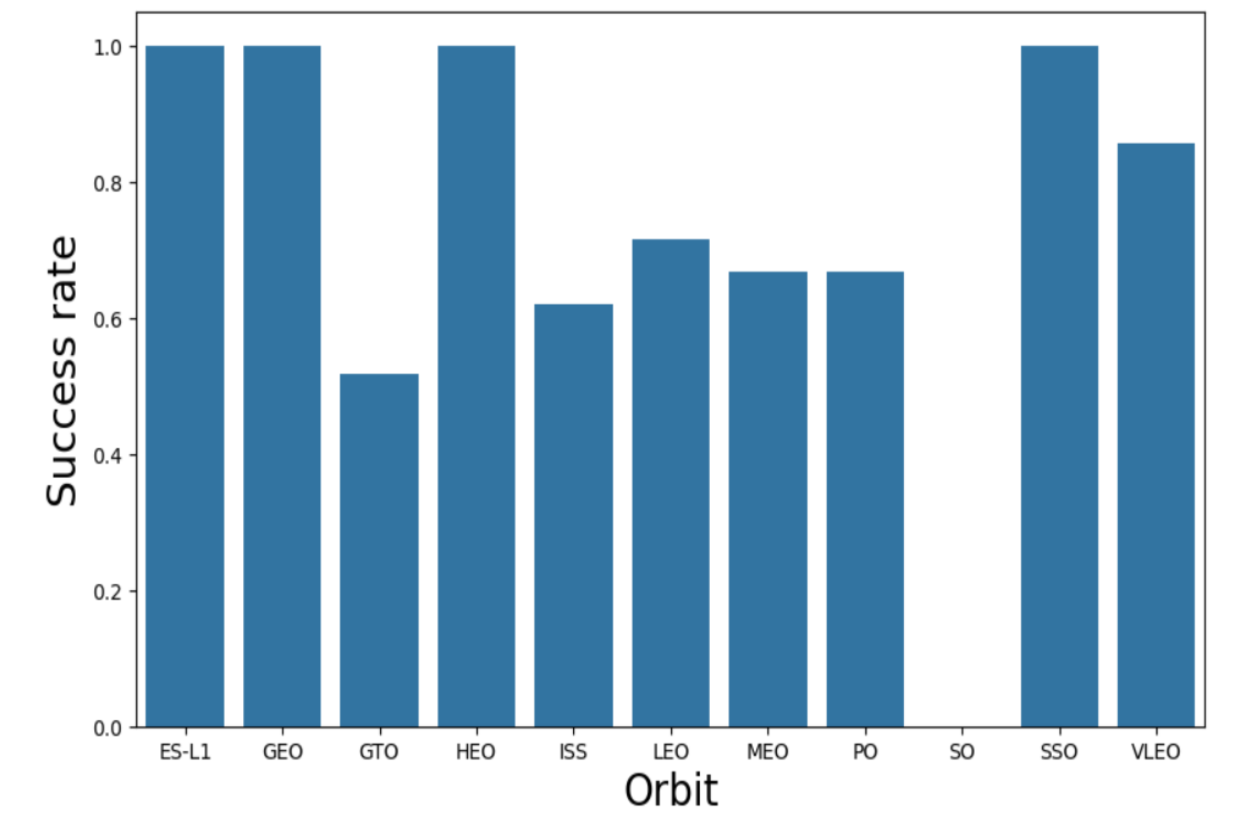
Payload vs. Launch Site

The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



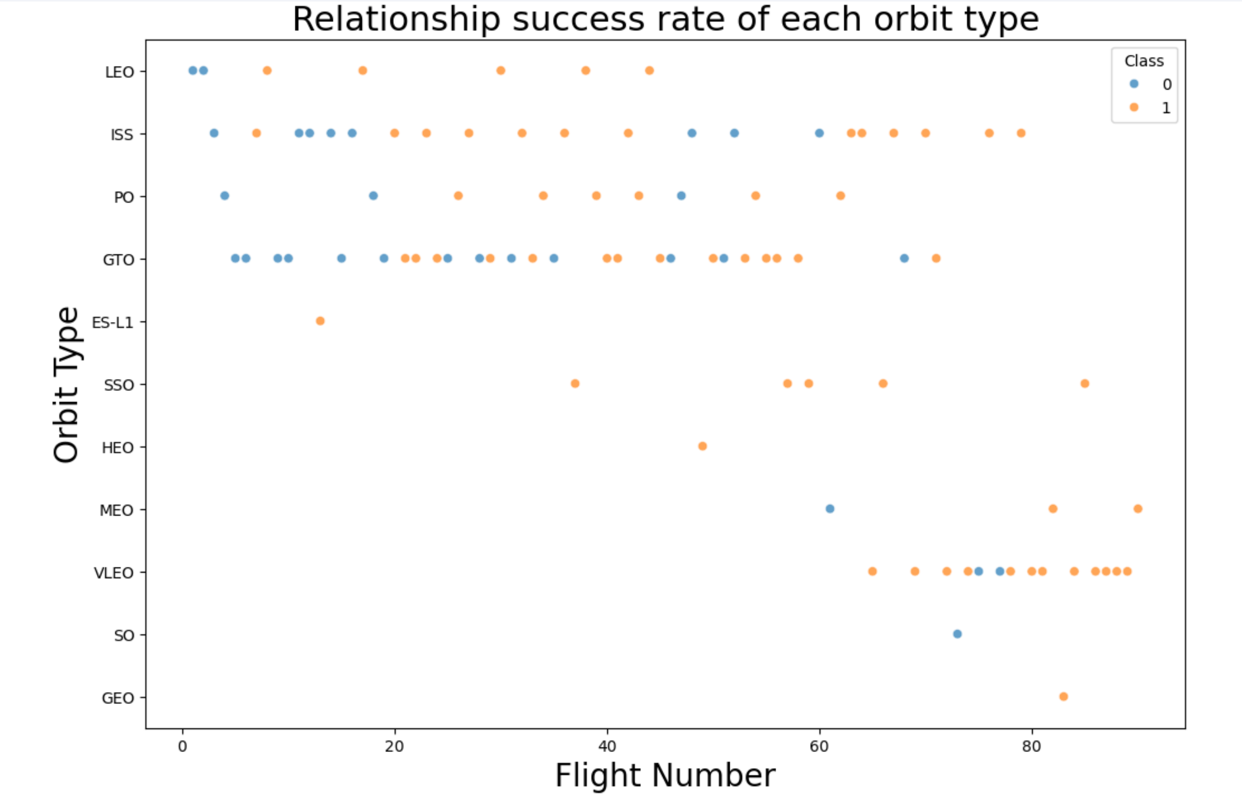
Success Rate vs. Orbit Type

From the plot, we can see that ES-L1, GEO, HEO and SSO had the most success rate



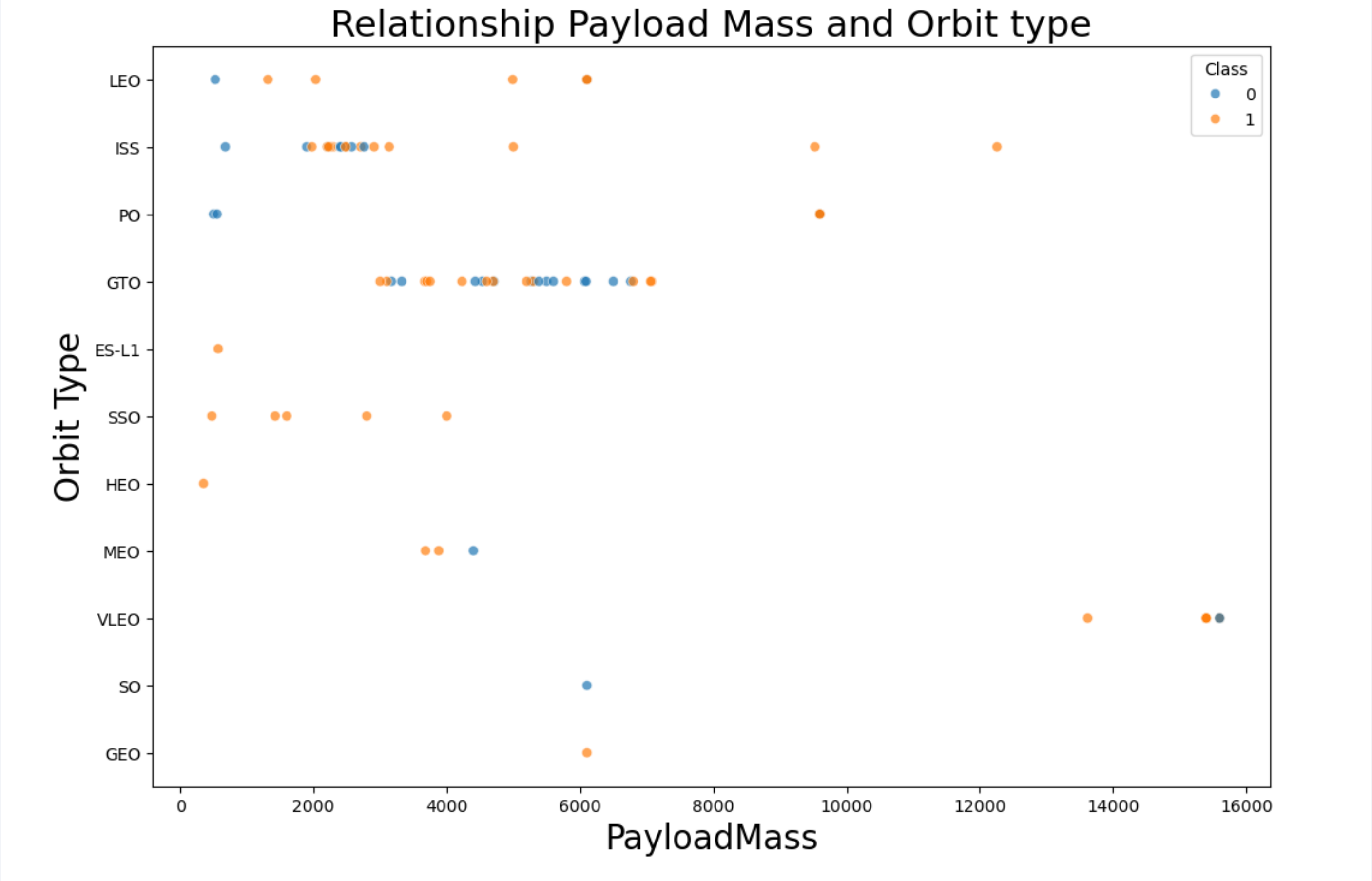
Flight Number vs. Orbit Type

- We can see if there is any relationship between Flight Number and Orbit type.
- GTO appears to be no relationship between Flight Number and Orbit type



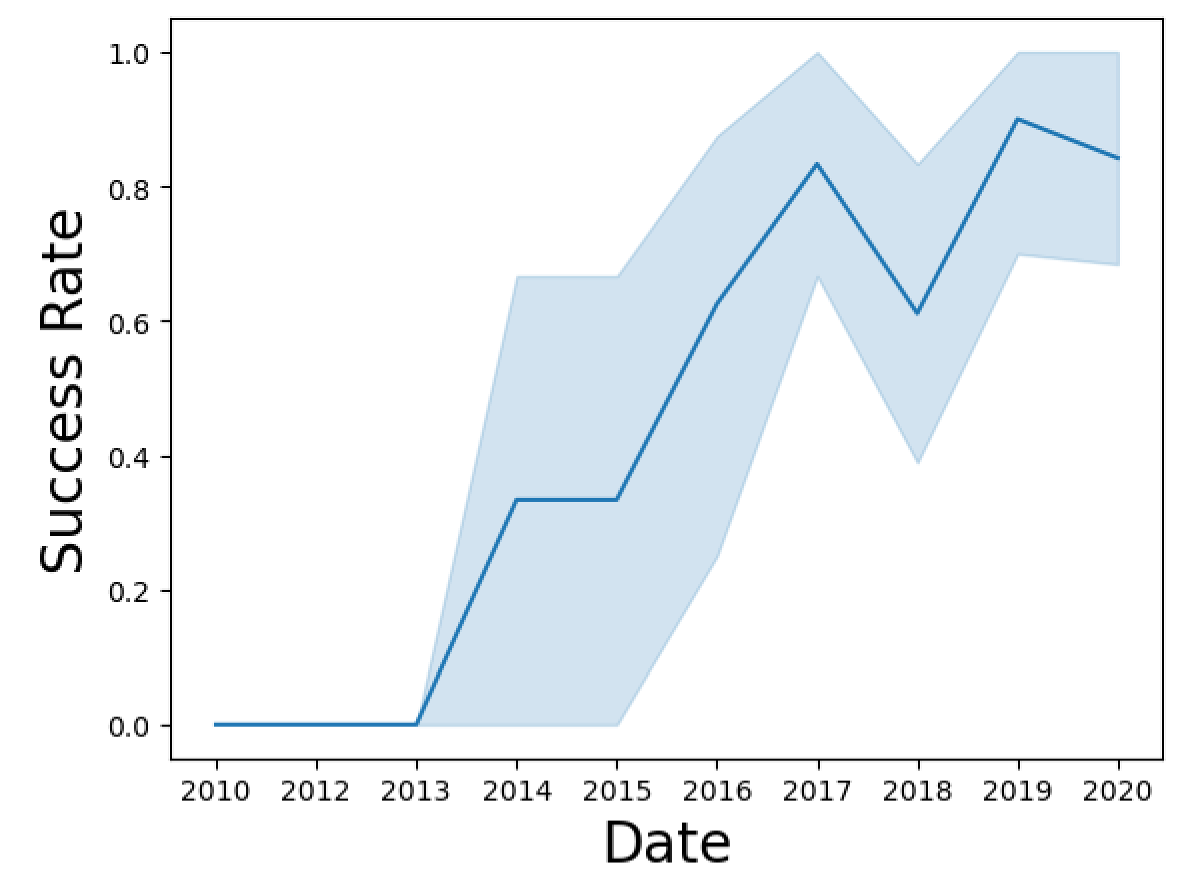
Payload vs. Orbit Type

- 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

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



All Launch Site Names

Key word DISTINCT was used to show only unique launch sites from the SpaceX data.

```
%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'

Display 5 records where launch sites 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

Total Payload Mass

Calculate the total payload carried by boosters from NASA

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

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

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

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

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

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

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 kg

```
%%sql SELECT 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

Calculate the total number of successful and failure mission outcomes

```
%%sql SELECT "Mission_Outcome", COUNT(*) AS Total
FROM SPACEXTABLE
GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

```
%%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

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

Month_Name	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

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

```
%%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.
```

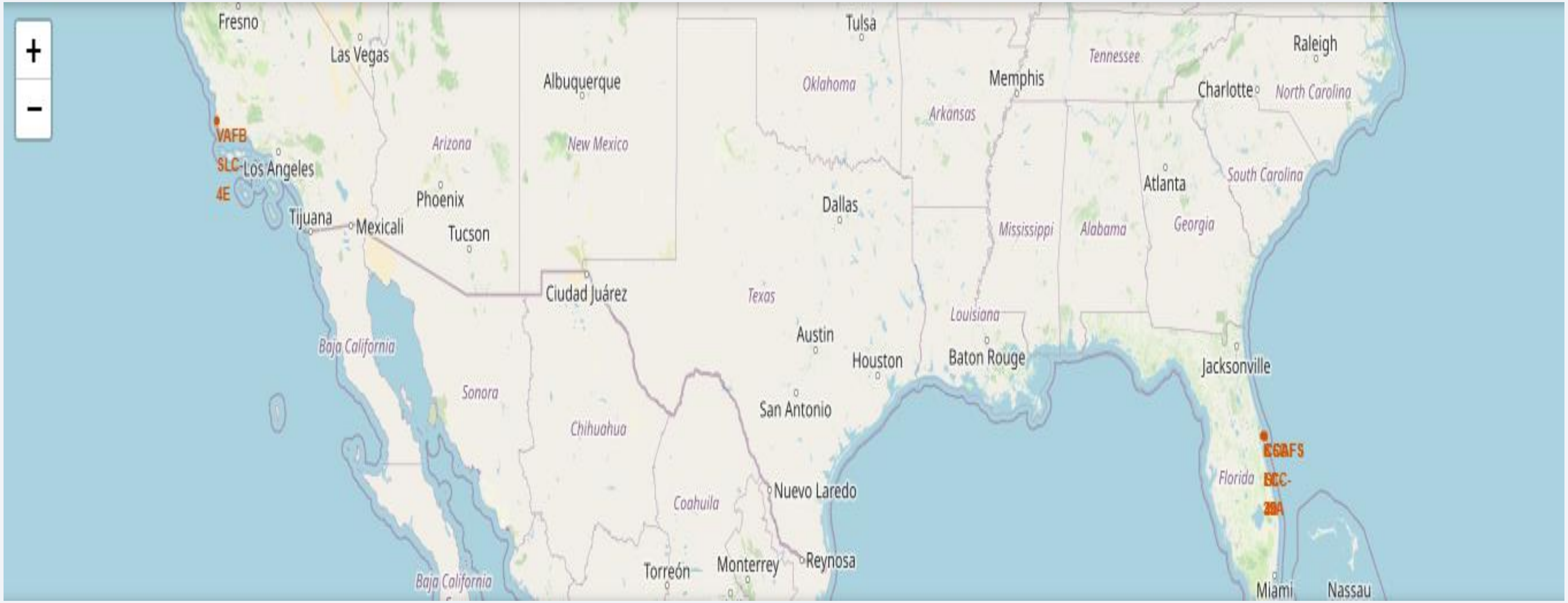
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

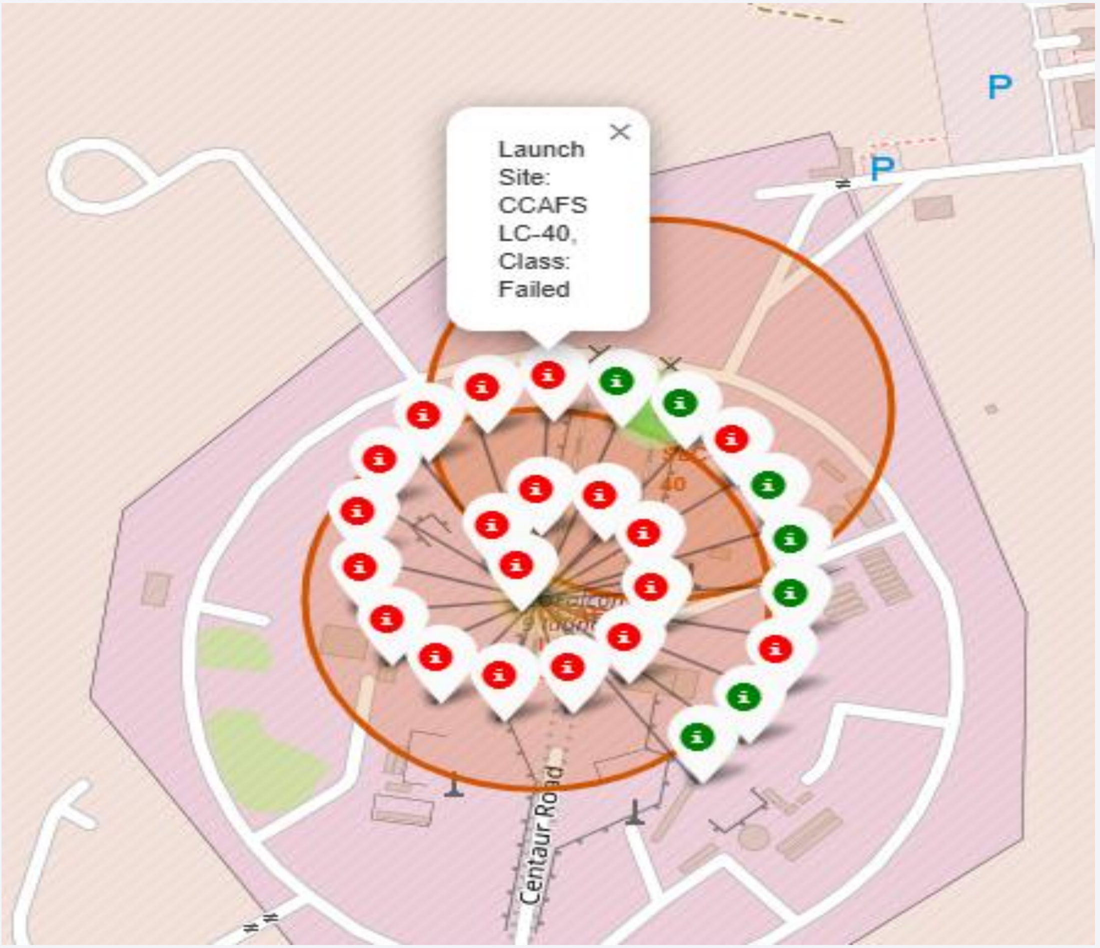
Launch Sites on a Map

SpaceX launch sites are located on the coast of USA



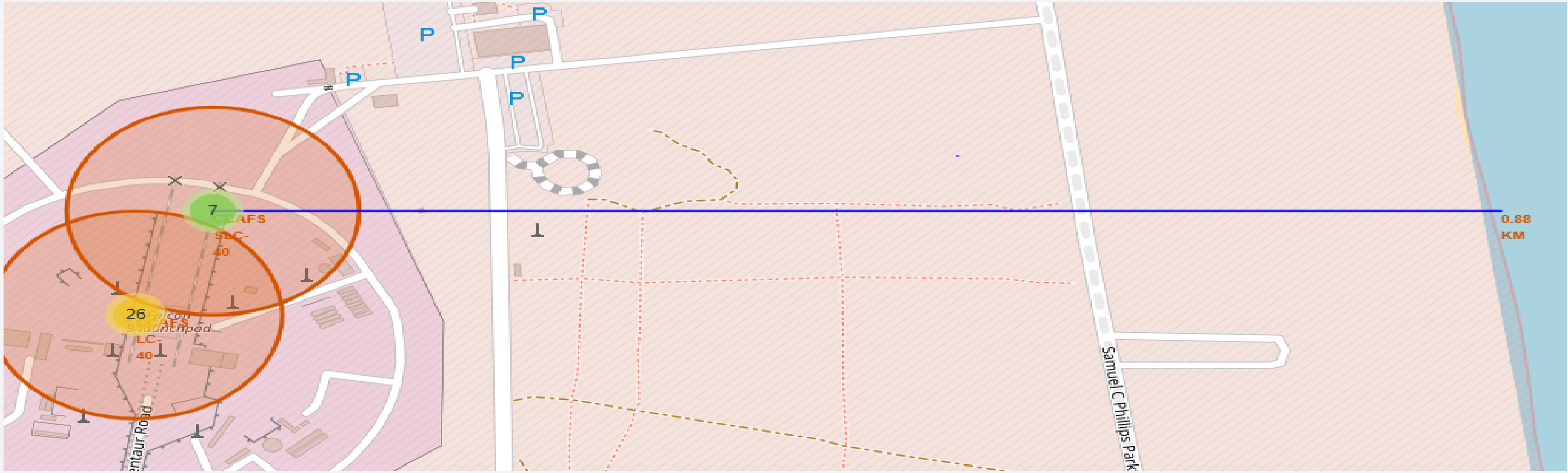
Success/Failed Launches for Each Site

- Adding the launch outcomes for each site, and see which sites have high success rates.
- Should be able to easily identify which launch sites have relatively high success rates



Calculate the Distances

- Distances between a launch site to proximities such as railway, highway, coastline, with distance calculated and displayed
- Distance calculated and displayed

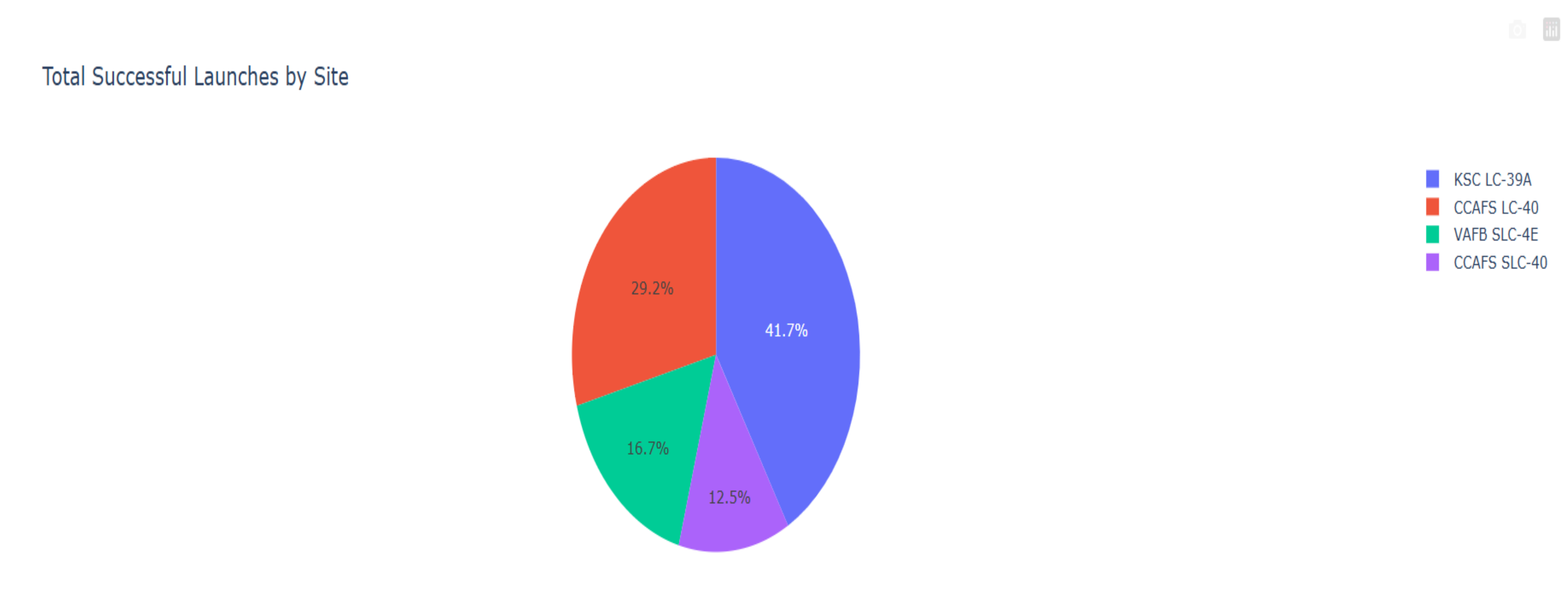


Section 4

Build a Dashboard with Plotly Dash

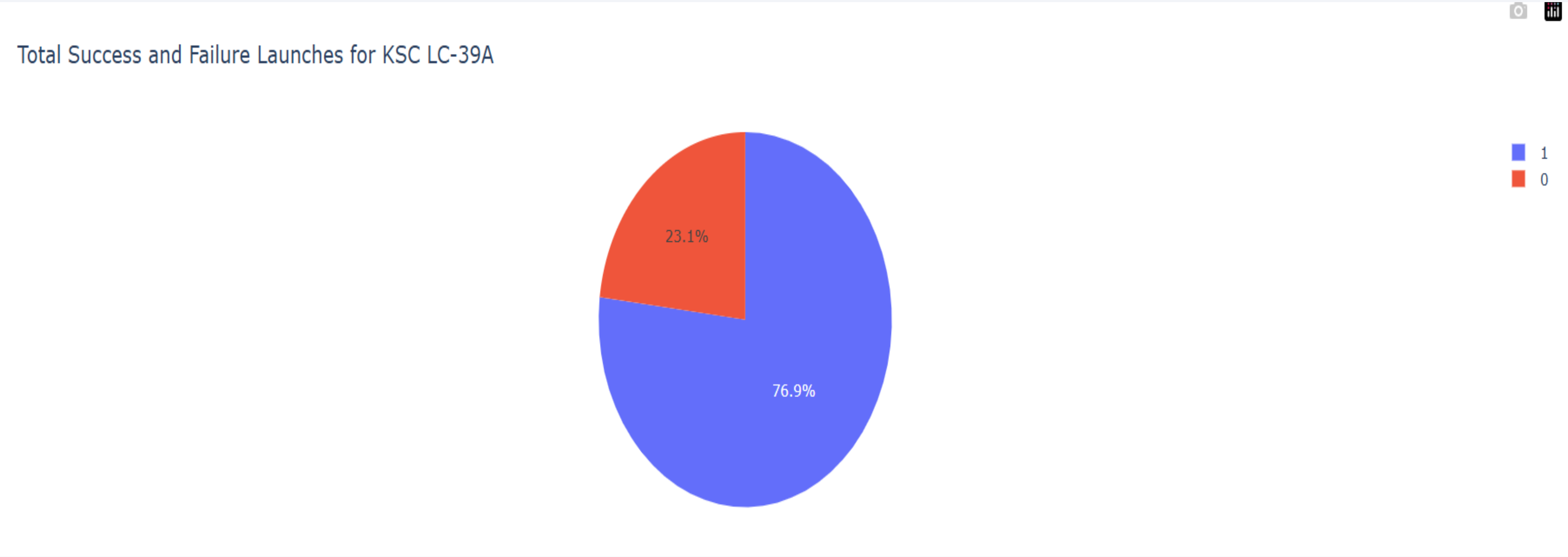
Total Successful Launches by Site

KSC LC-39A has the most successful launches with 41.7%



Launch Site with Hiest Launch Success Ratio

This pie chart shows the success and failure launches for KCS LC-39A



Payload vs Launch Outcome

- Payload vs. Launch Outcome scatter plot for all sites with
- Success rates for low weighted payloads (FIG 1) are higher than heavy weighted payloads (FIG 2).

FIG 1

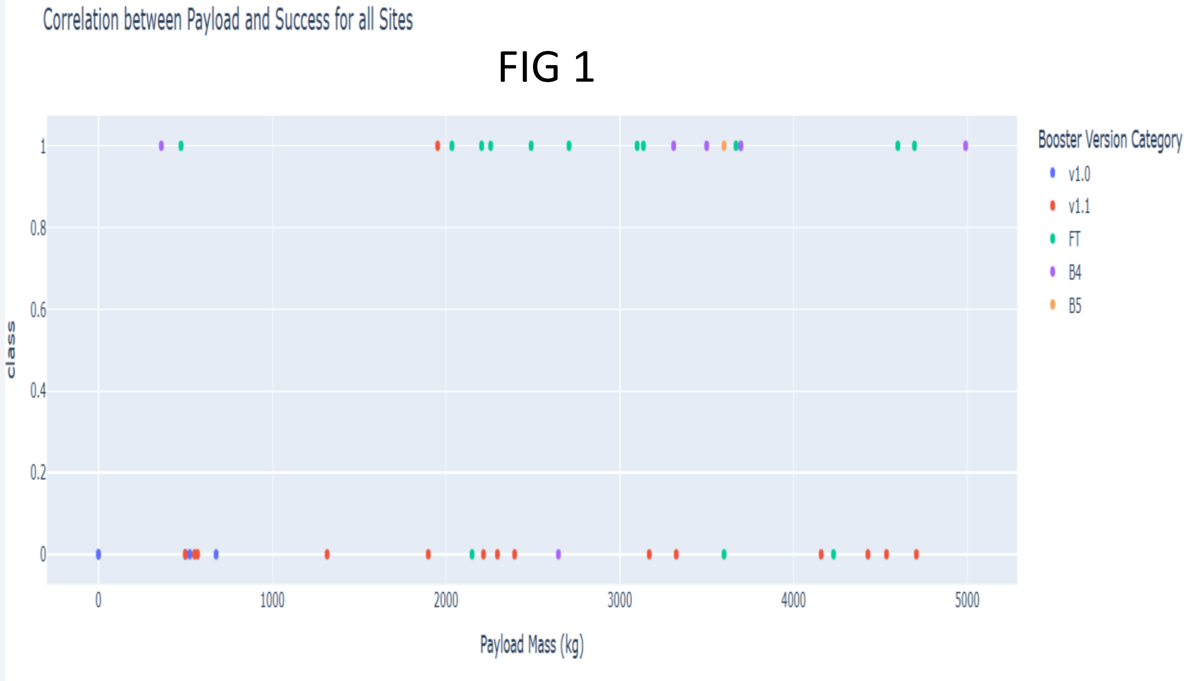
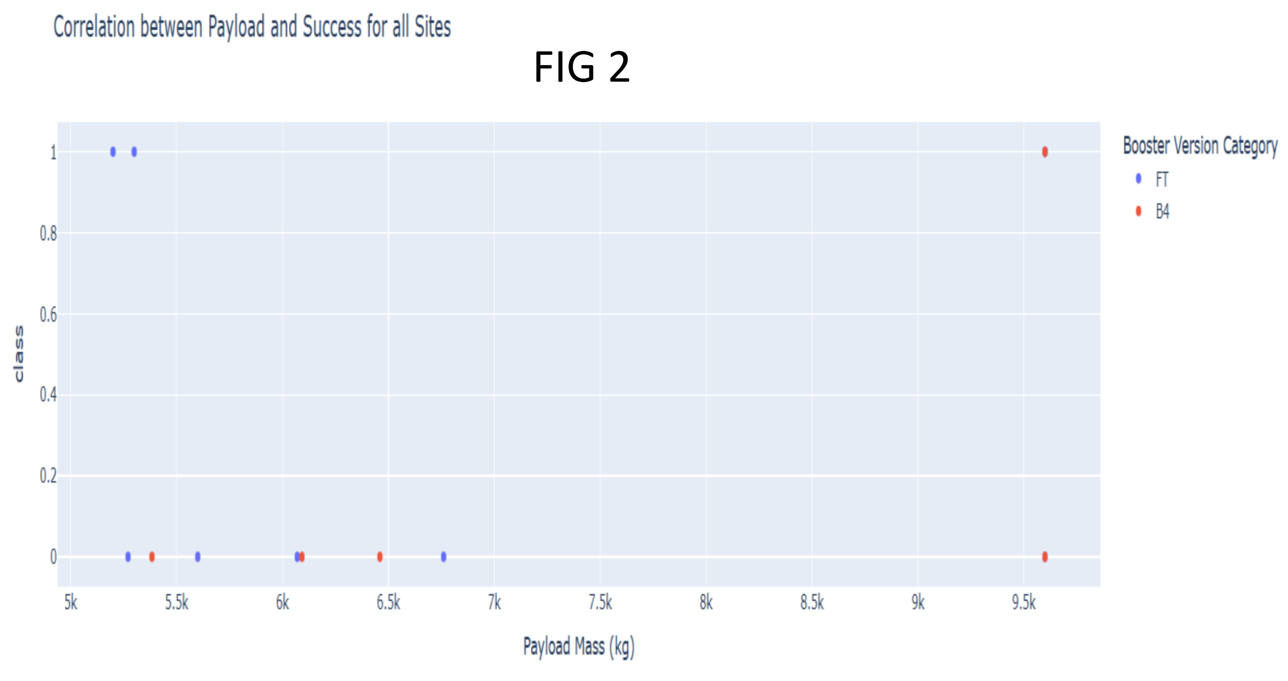


FIG 2

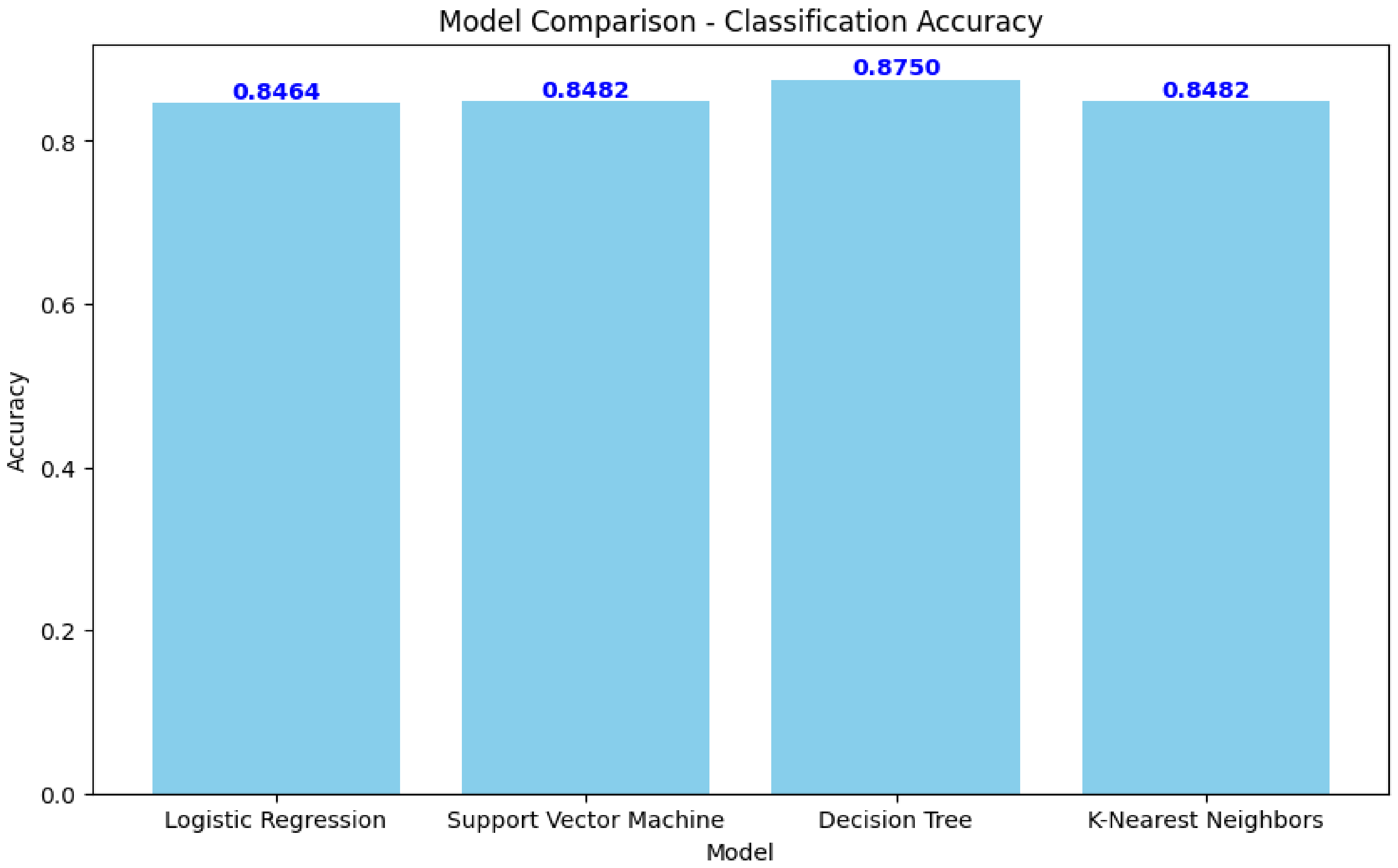


Section 5

Predictive Analysis (Classification)

Classification Accuracy

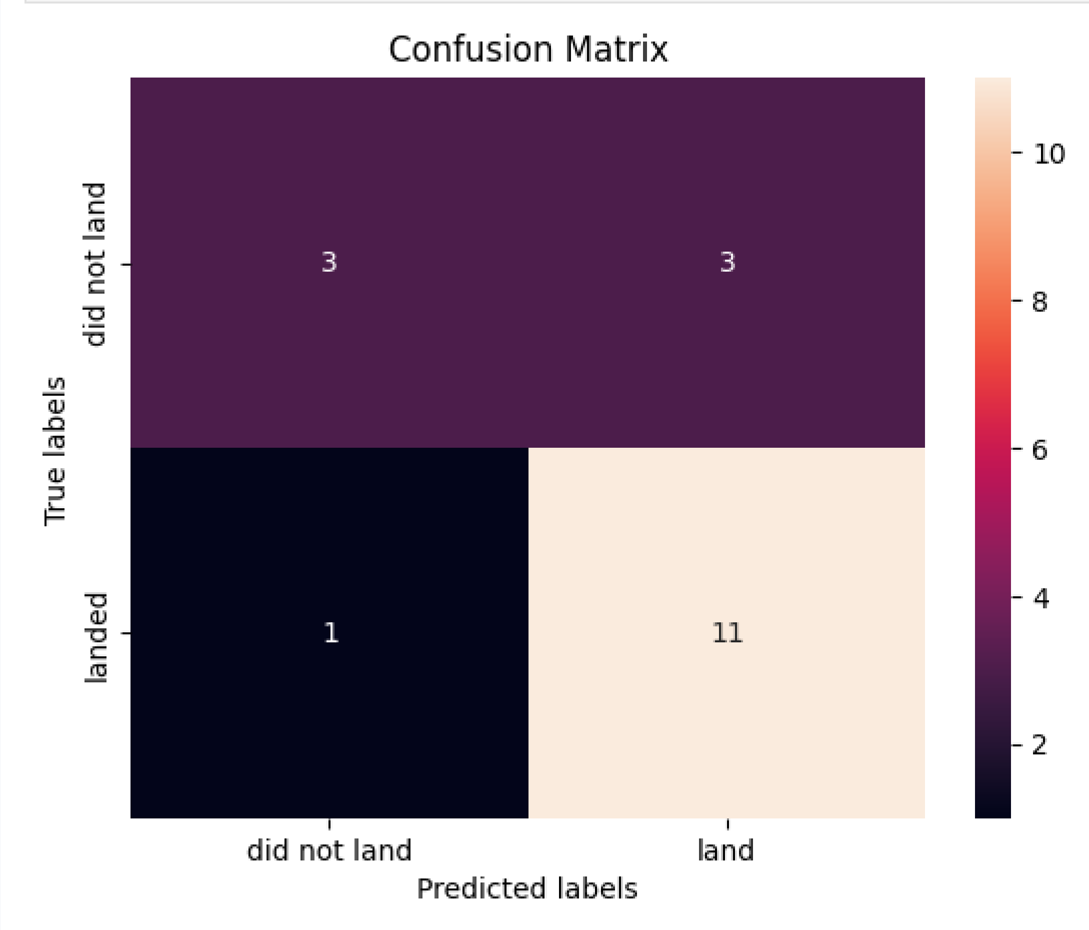
The decision tree classifier is the model with the highest accuracy



The model with the highest accuracy is: Decision Tree with an accuracy of 0.8750

Confusion Matrix

This confusion matrix indicates that the model has a high accuracy rate for the "landed" class and some confusion in the "did not land" class.



Conclusions

- The higher the number of flights at a launch site, the greater the success rate at that site.
- The launch success rate began increasing in 2013 and continued to rise until 2020.
- The orbits ES L1, GEO, HEO and SSO had the highest success rates.
- KSC LC-39A had the most successful launches among all sites.
- The Decision Tree classifier is the most suitable machine learning algorithm for this task.

Thank you!