IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Murilo Ferreira Pires
09/12/2024

Outline

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

Executive Summary

Summary of Methodologies:

This project aimed to predict the successful landing of the Falcon 9 first stage. The methodologies employed included:

- **Data Collection**
- **Data Wrangling**
- **Interactive Analysis**
- **Machine Learning Models**
- **Hyperparameter Tuning**

Summary of All Results:

- Data Insights
- Model Performance
- Business Impact

Introduction

Project Background and Context:

- SpaceX's Falcon 9 rocket has revolutionized space travel by reducing launch costs through the reuse of its first stage. Each Falcon 9 launch costs around 62 million dollars, significantly less than competitors whose costs exceed 165 million dollars per launch. The key to this cost-saving is the successful landing and reuse of the rocket's first stage. This project focuses on predicting the success of these landings to provide insights for potential competitors looking to enter the market.

Problems You Want to Find Answers:

- Can we accurately predict if the Falcon 9 first stage will land successfully?
- What factors significantly influence the landing outcome?
- How can machine learning models be used to forecast landing success and aid in cost-effective decision-making for future rocket launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used RESTful APIs and web scraping to gather historical data on Falcon 9 launches.
- Performed data wrangling
  - Cleaned and transformed the collected data into a structured dataframe.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Developed and tested various machine learning models, including SVM, Classification Trees, and Logistic Regression, to predict landing outcomes.

Data Collection

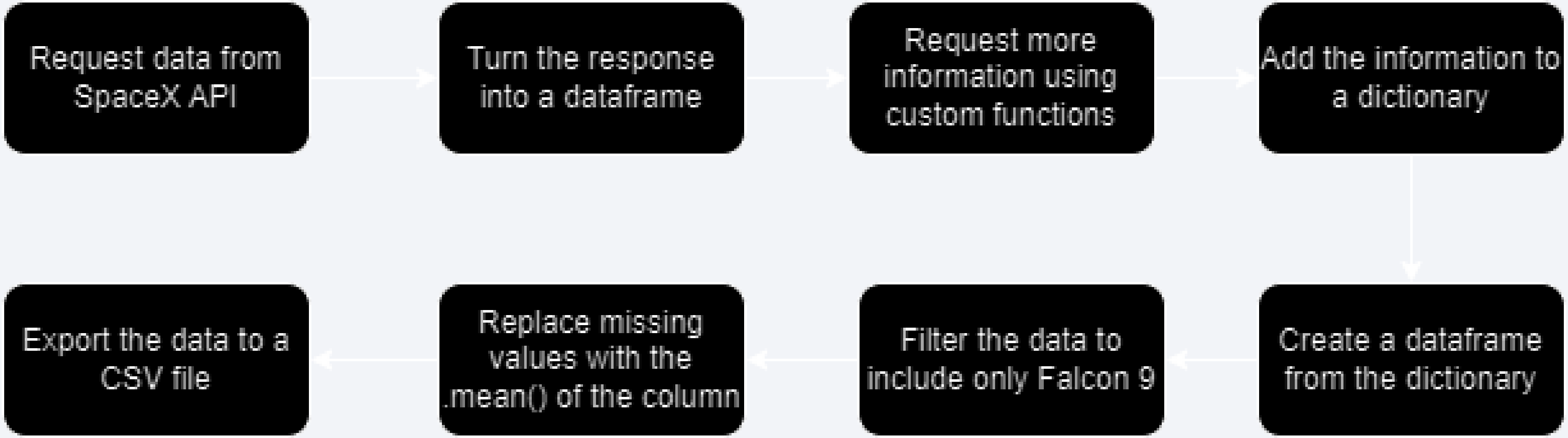
Description of Data Collection:

Data for this project was collected using a combination of RESTful API calls and web scraping techniques to gather historical launch records of Falcon 9 first-stage landings. The primary data sources included:

- **SpaceX Launch Data API:** Utilized a RESTful API to retrieve detailed information about each Falcon 9 launch, including date, launch site, payload, and landing outcome.
- **Web Scraping:** Supplemented the API data by scraping additional details from SpaceX's official website, such as launch site proximities and rocket specifications.

Data Collection – SpaceX API

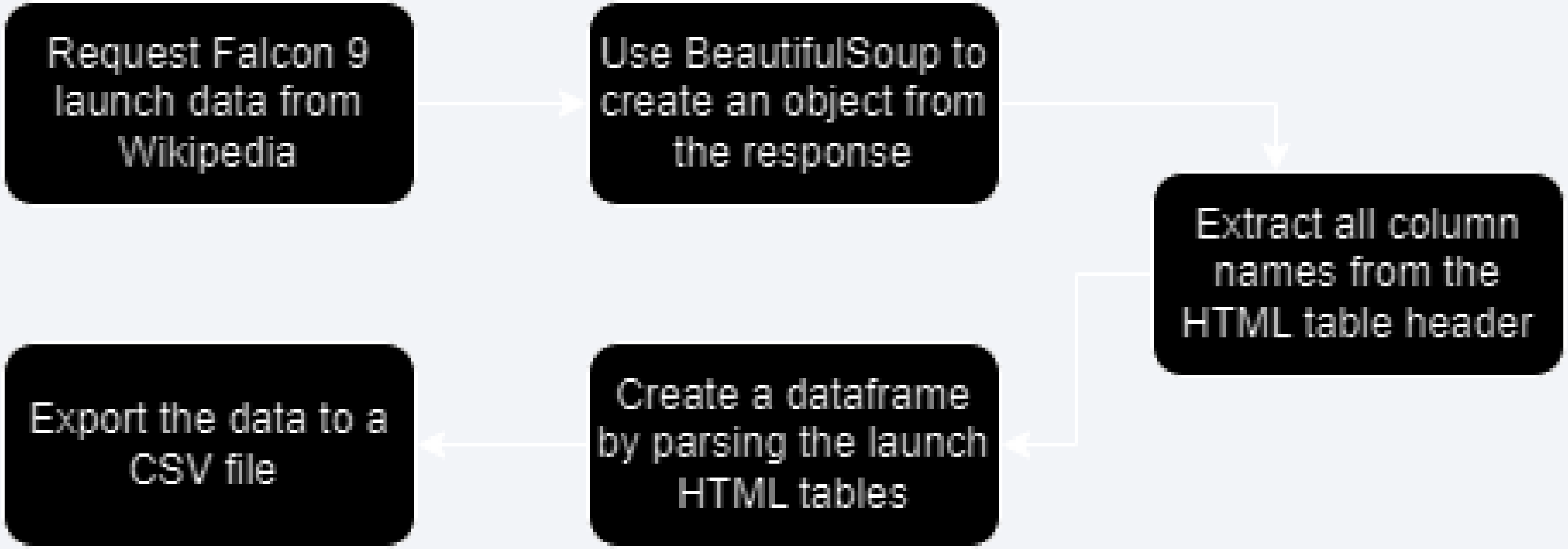
- Flowchart for the process of Data Collection using the SpaceX API



[GitHub URL for the notebook file]

Data Collection - Scraping

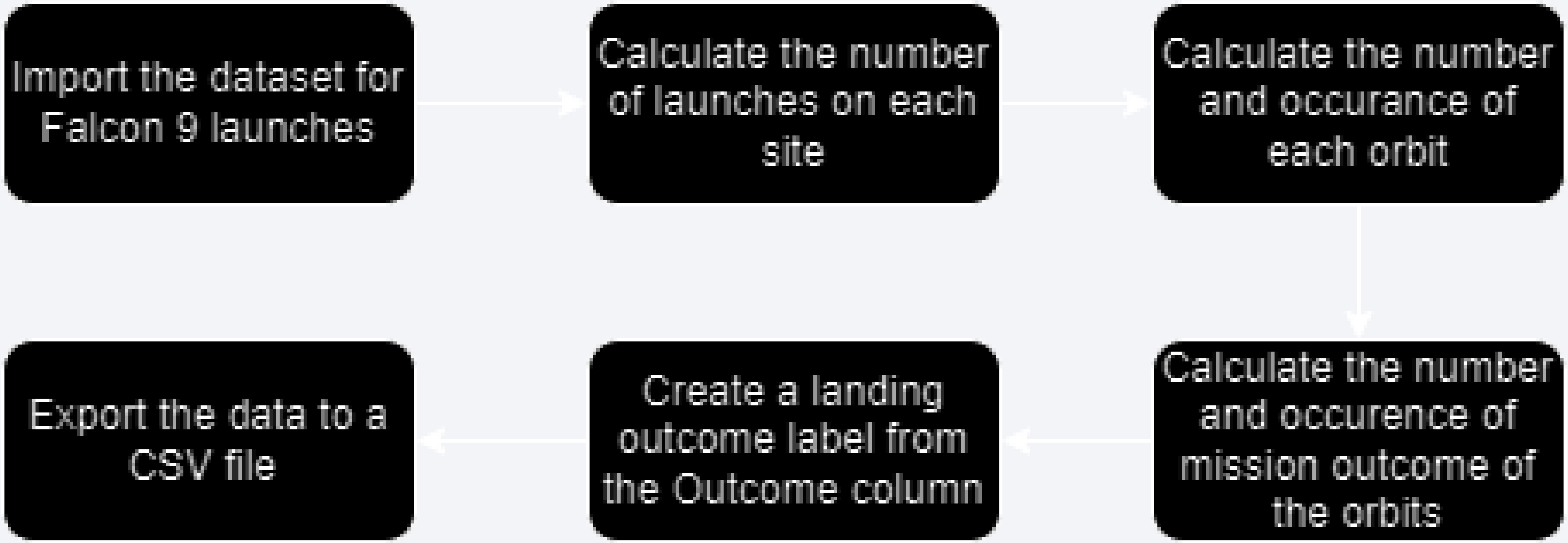
- Flowchart for the process of Data Collection using Web Scraping



GitHub URL for the notebook file

Data Wrangling

- The data was processed through Exploratory Data Analysis (EDA) to identify patterns and convert landing outcomes into training labels. Various landing scenarios, such as True Ocean, True RTLS, and True ASDS, were standardized into binary labels for model training.



[GitHub URL for the notebook file]()

EDA with Data Visualization

Summary of Charts Plotted:

1. **Scatter Plot (Flight Number vs. Launch Site):** Used to explore the distribution of flight numbers across different launch sites, helping to identify any patterns in launch frequency and location.
2. **Scatter Plot (Payload Mass vs. Launch Site):** Visualized the relationship between payload mass and launch site to assess how payload capacity varied across different locations.
3. **Bar Plot (Success Rate by Orbit Type):** Illustrated the success rate of each orbit type to identify which orbits had higher success rates, providing insights into mission planning and risk assessment.
4. **Scatter Plot (Flight Number vs. Orbit Type):** Analyzed how the flight number correlated with different orbit types, revealing trends in the types of missions conducted over time.
5. **Scatter Plot (Payload Mass vs. Orbit Type):** Investigated the impact of payload mass on orbit type to determine if certain orbits were more suited for heavier payloads.
6. **Line Chart (Yearly Launch Success Trend):** Displayed the success trend of launches over time, providing a historical perspective on SpaceX's landing success and technological improvements.

GitHub URL for the notebook file

EDA with SQL

In this lab, there were several SQL queries performed for Exploratory Data Analysis, such as:

- Displaying the names of unique entries;
- Displaying the total (sum) and average values of some attributes;
- Listing successful landings filtering by date and payload mass;
- Listing the number of successful and failure mission outcomes;
- And more, including advanced queries using sub-queries and multiple filtering conditions.

GitHub URL for the notebook file

Build an Interactive Map with Folium

Map Objects Created and Added:

- **Markers for Launch Sites:** Placed markers at each launch site to provide a visual reference for their geographical locations. This helped in identifying spatial patterns and proximity relationships between the sites and their surroundings.
- **Circle Markers for Success/Failed Launches:** Added circle markers with color coding to indicate the success or failure of launches at each site. This visualization made it easy to identify which sites had higher success rates, offering insights into location-based performance.
- **Lines for Distance Calculations:** Drew lines to measure the distances between each launch site and its proximities (e.g., nearby cities or significant landmarks). This was crucial for understanding how geographical factors like proximity to populated areas or water bodies might influence launch outcomes.

Explanation for Adding These Objects:

- **Markers** were used to pinpoint the exact locations of launch sites, allowing for a straightforward geographical analysis.
- **Circle Markers** provided a visual representation of launch success rates, helping to uncover patterns that may be influenced by location.
- **Lines** helped in assessing the impact of proximity factors on launch success, offering a deeper understanding of how location dynamics affect launch outcomes.

[GitHub URL for the notebook file]()

Build a Dashboard with Plotly Dash

Summary of Plots/Graphs and Interactions:

- **Success Pie Chart:** Displays the success rate for each launch site, updated via a dropdown menu. This allows quick comparison of site performance.
- **Payload vs. Success Scatter Plot:** Shows the relationship between payload mass and launch success, with a range slider for filtering. This helps explore how payload size impacts success rates.

Why These Plots and Interactions:

- **Pie Chart:** Offers an intuitive overview of success rates by site, enabling easy analysis of location-based performance.
- **Scatter Plot:** Provides insights into the effect of payload mass on launch outcomes, allowing users to identify patterns and trends through interactive filtering.

[GitHub URL for the dashboard file]()

Predictive Analysis (Classification)

Summary of Model Development Process:

- **Data Preparation:** Performed exploratory data analysis to identify key features and created a target column for the classification task, indicating the success or failure of the Falcon 9 first-stage landing.
- **Data Standardization:** Standardized the features to ensure uniformity in the scale, improving the performance of machine learning models.
- **Data Splitting:** Split the dataset into training and testing sets to evaluate model performance objectively.
- **Model Building:** Built multiple classification models, including Support Vector Machine (SVM), Classification Trees, and Logistic Regression.
- **Hyperparameter Tuning:** Applied cross-validation and hyperparameter tuning to optimize each model for the best performance.
- **Model Evaluation:** Evaluated the models using the test data and identified the best-performing model based on accuracy scores.

GitHub URL for the notebook file

Results

Exploratory Data Analysis Results:

- Lower payload mass launches showed a higher chance of successful landings.
- The success rates of launches have been increasing steadily over the years.

Interactive Analytics (Screenshots on Slides 39-41):

- The launch sites are strategically located near the Equator, close to highways, railroads, and coastlines to optimize launch conditions.
- The launch site with the highest success rate identified was KSC LC-39A.

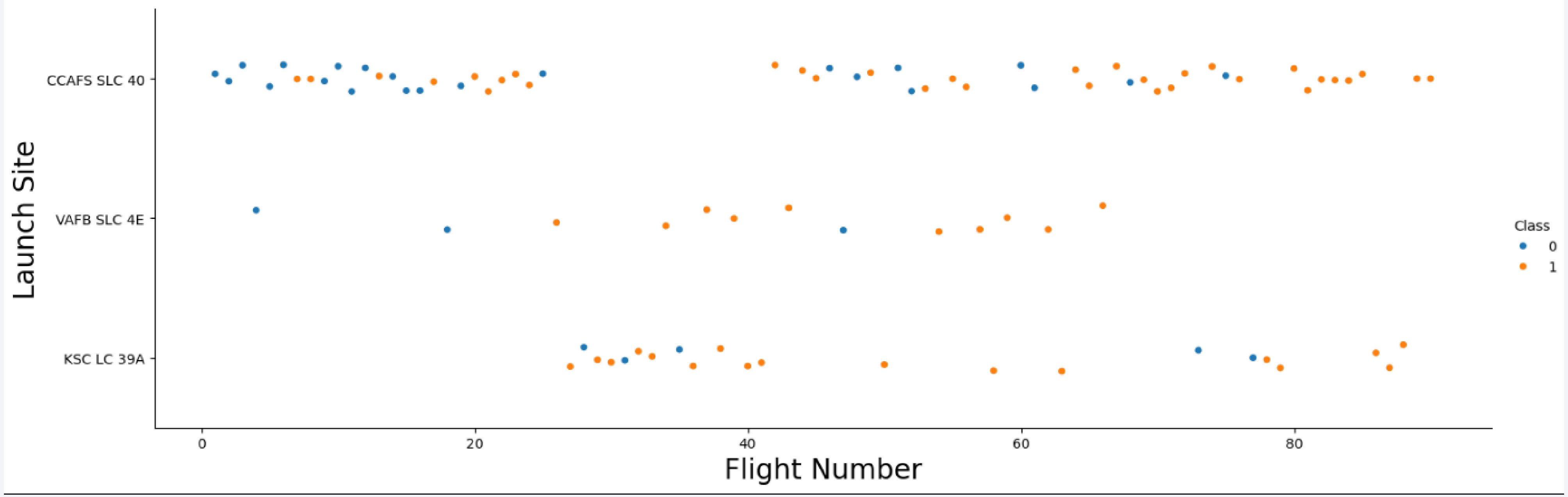
Predictive Analysis Results:

- Hyperparameter tuning significantly improved the accuracy of the models.
- Among the tested models, the Decision Tree model achieved the highest accuracy for predicting Falcon 9 first-stage landing success.

Section 2

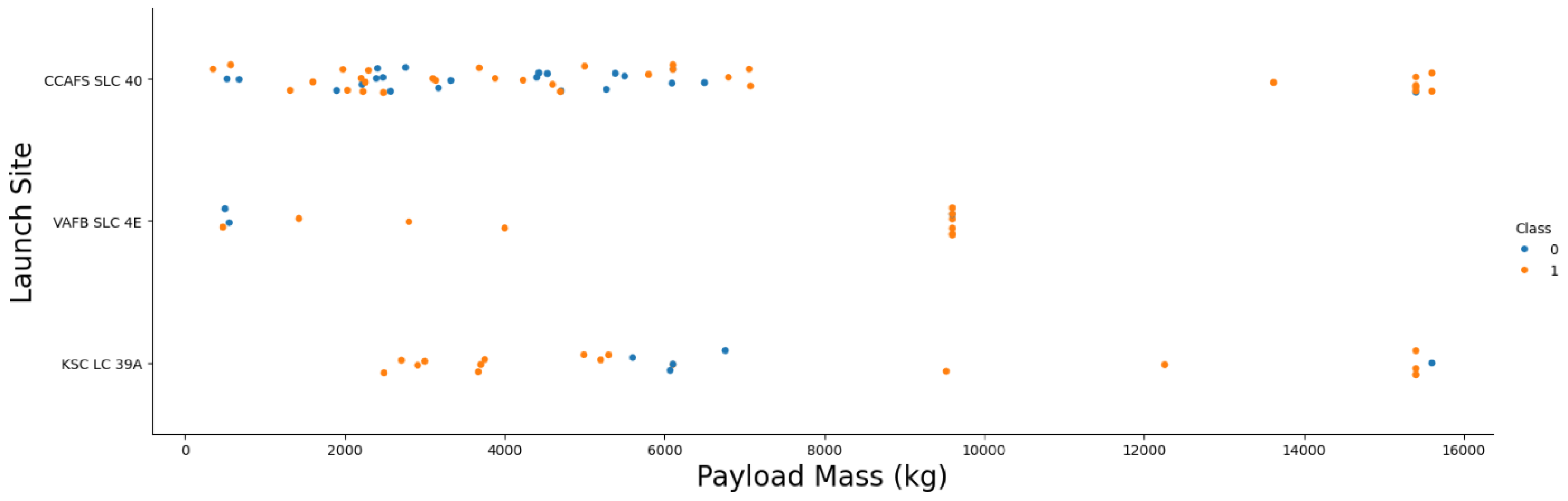
Insights drawn from EDA

Flight Number vs. Launch Site



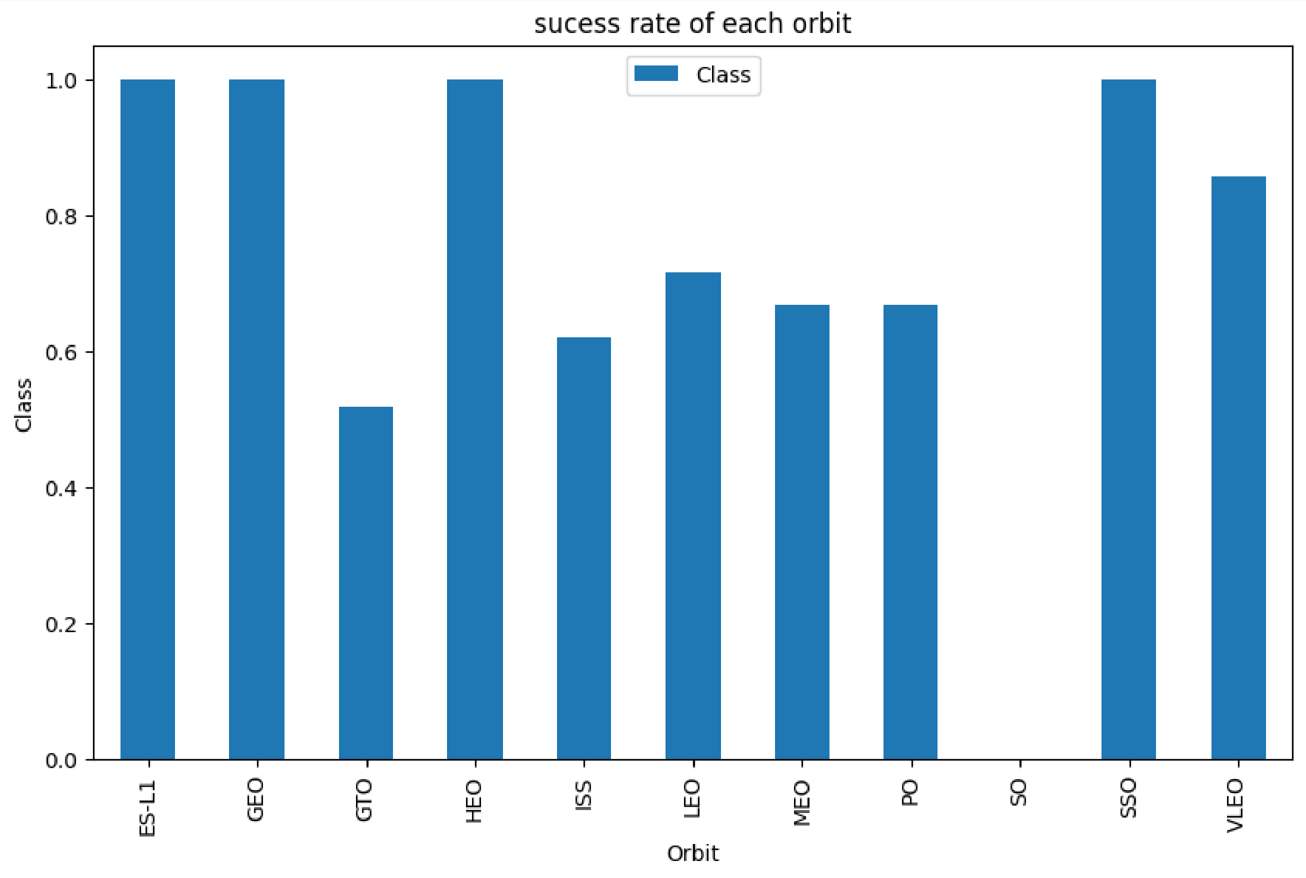
Used to explore the distribution of flight numbers across different launch sites, helping to identify patterns in launch frequency and location.

Payload vs. Launch Site



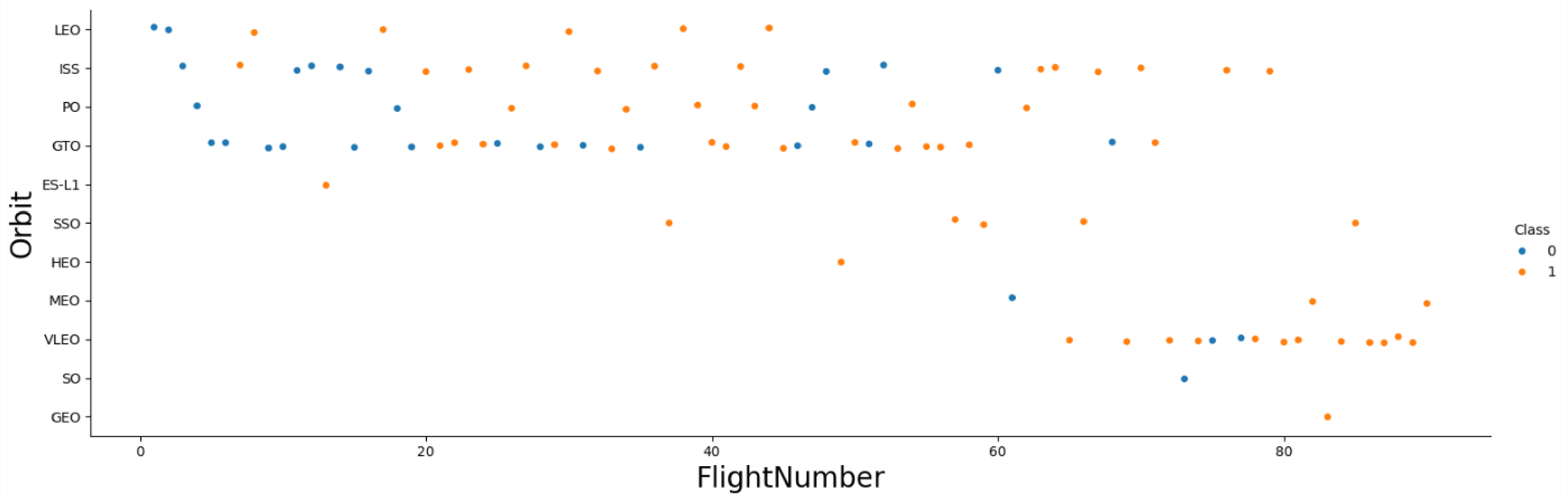
Visualized the relationship between payload mass and launch site to assess how payload capacity varied across different locations. It's noticeable that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type



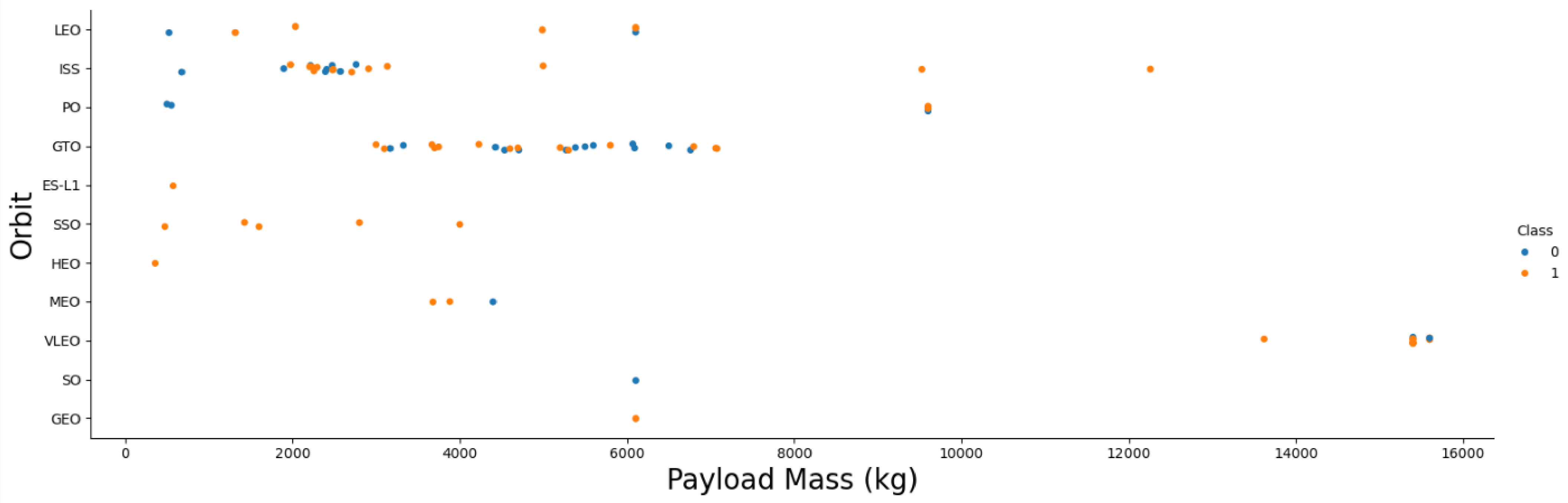
By analyzing this bar chart it's possible to identify which orbits have the highest success rates.

Flight Number vs. Orbit Type



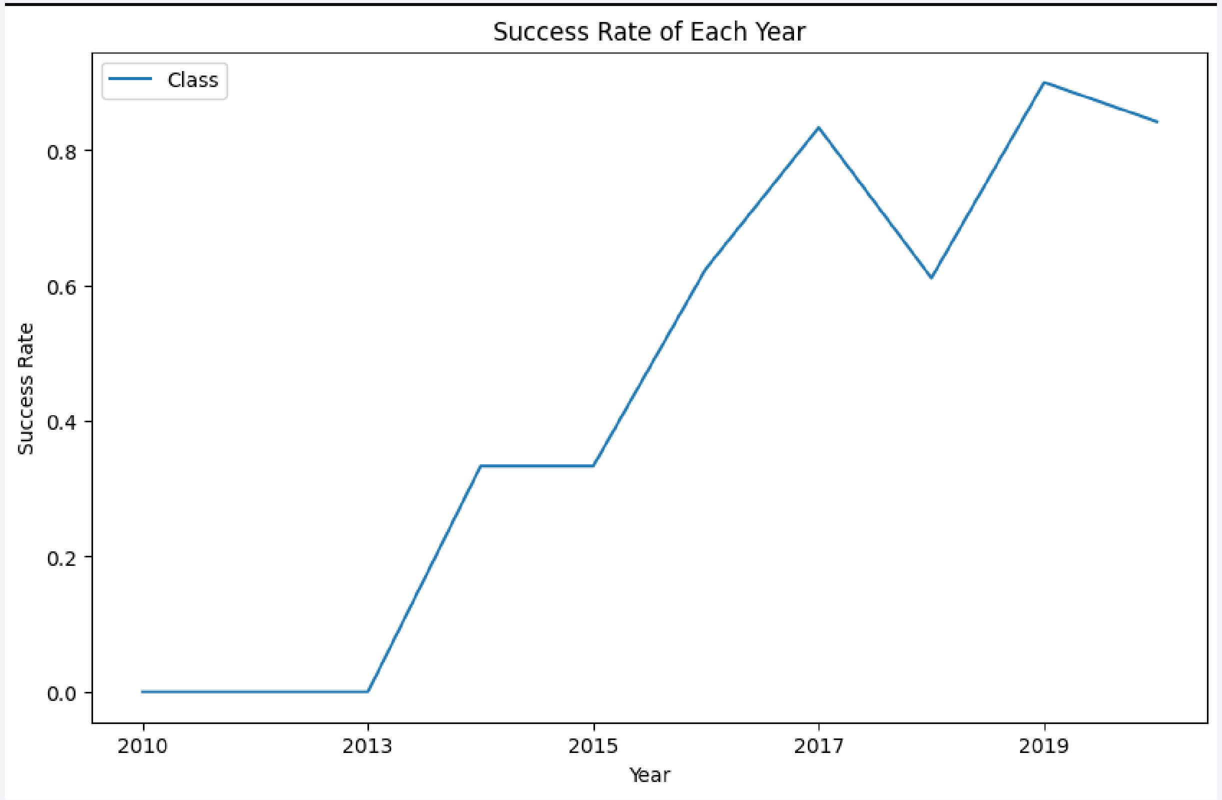
Analyzed how the flight number correlated with different orbit types, revealing trends in the types of missions conducted over time. You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



With this chart, we can investigate the impact of payload mass on orbit type to determine if certain orbits were more suited for heavier payloads. 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



This chart displays the success trend of launches over time, providing a historical perspective on SpaceX's landing success and technological improvements. You can observe that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

```
In [13]: %sql SELECT DISTINCT Launch_Site from SPACEXTBL

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

Out[13]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

This SQL Query displayed the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

In [15]:
```
%sql SELECT * from SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5
```

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

Out[15]:

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

This SQL Query displays 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

```
In [16]: %sql SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Customer LIKE "NASA (CRS)";

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

Out[16]:

SUM(PAYLOAD_MASS__KG_)
45596

This SQL Query displays the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
In [26]: %sql SELECT avg(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Booster_Version LIKE "F9 v1.1 %"

 * sqlite:///my_data1.db
Done.

Out[26]:
```

avg(PAYLOAD_MASS__KG_)
2337.8

This SQL Query displays average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
In [23]: %sql SELECT min(Date) as "First Success Ground Pad" from SPACEXTBL WHERE Landing_Outcome LIKE "Success (ground pad)";

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

Out[23]:

First Success Ground Pad
2015-12-22

This SQL Query lists the date when the first successful landing outcome in ground pad was acheived.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [25]: %sql SELECT Booster_Version from SPACEXTBL WHERE Landing_Outcome LIKE "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN
```

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

Out[25]:

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

This SQL Query lists the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
In [27]: %sql SELECT Mission_Outcome, Count(*) AS Total FROM SPACEXTBL GROUP BY Mission_Outcome;

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

Out[27]:

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

This SQL Query lists the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

In [32]:
```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) from
```

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

Out[32]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

This SQL Query lists the names of the booster_versions which have carried the maximum payload mass using a subquery

2015 Launch Records

```
In [36]: %sql SELECT strftime('%m', Date) AS Month, Mission_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE strftime('%Y'_
```

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

Out[36]:

Month	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

This SQL Query lists 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 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.

In [39]:
```
%sql SELECT Landing_Outcome, Count(*) AS TOTAL from SPACEXTBL WHERE Date BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY Lanc
```

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

Out[39]:

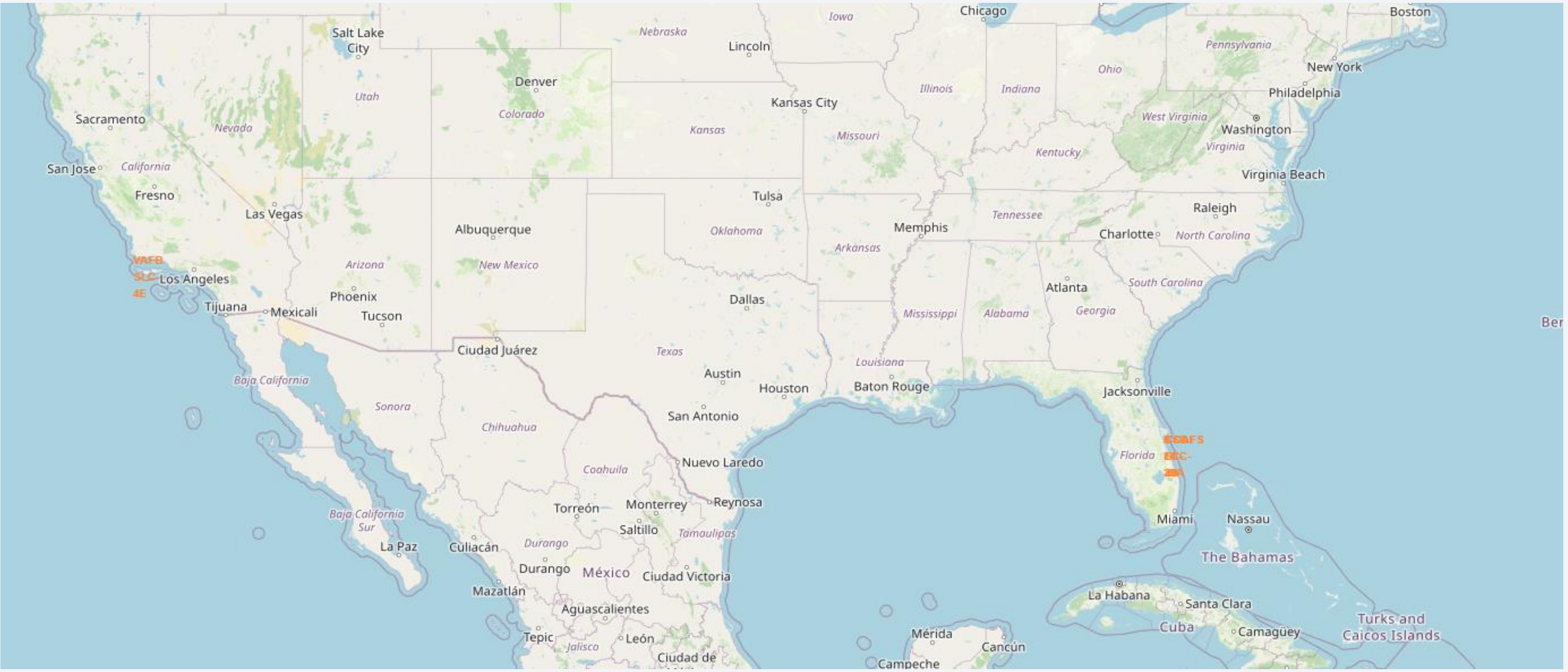
Landing_Outcome	TOTAL
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

This SQL Query ranks 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.

Section 3

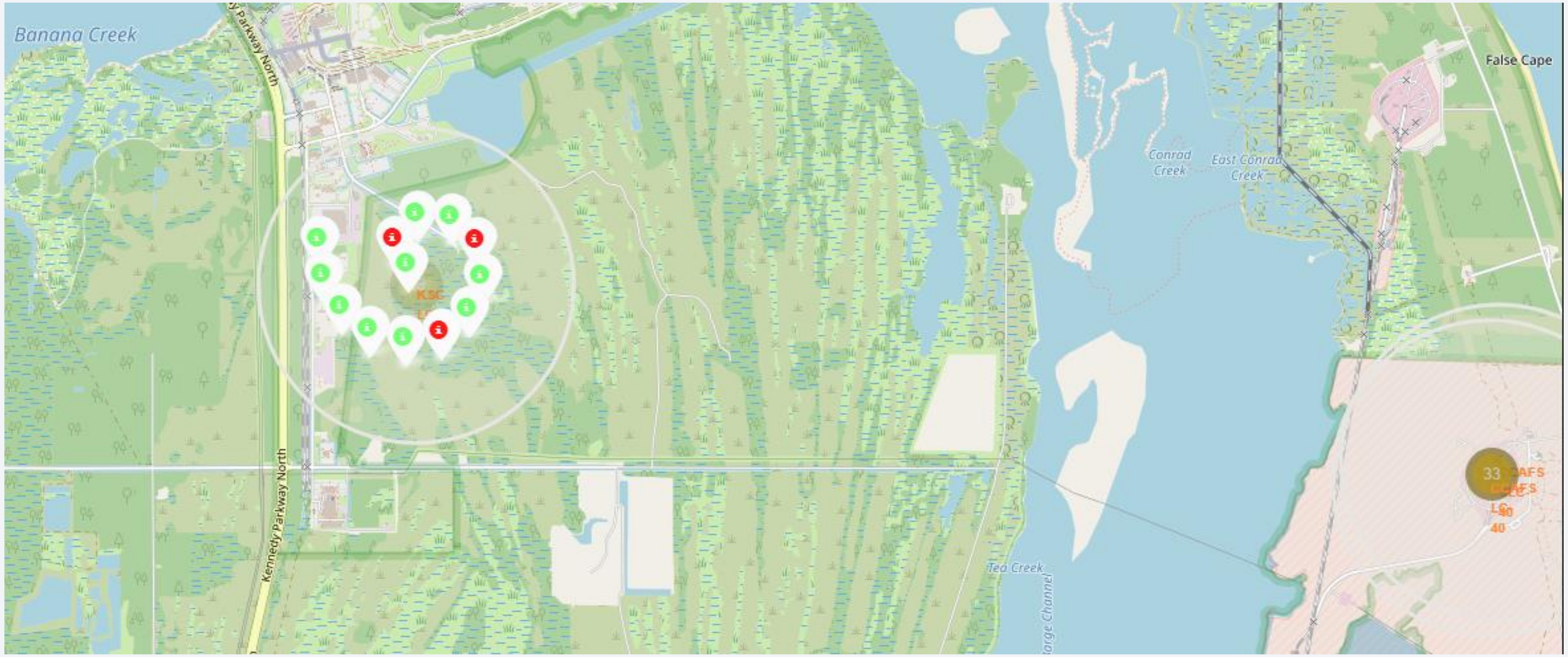
Launch Sites Proximities Analysis

Launch Site Locations



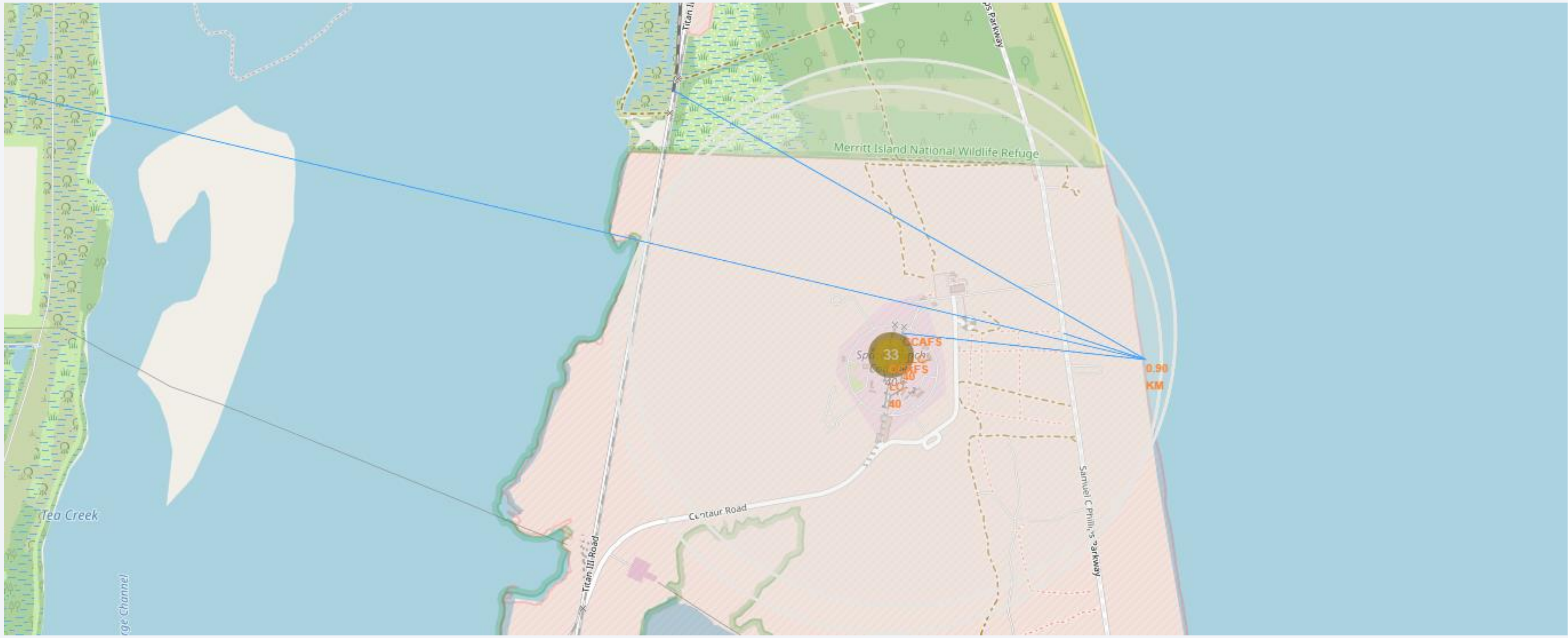
These are the launch site locations marked on the map using Folium. We can observe that the launch sites are close to the Equator line and to the coast.

Launch Outcomes on the Map



Here are the launch outcomes marked on the map in their corresponding launch locations. Green indicates success and red means failure. If we zoom out, the markers will cluster together, as shown in the other launch locations on the far right.

Distance from Launch Site to Important Elements

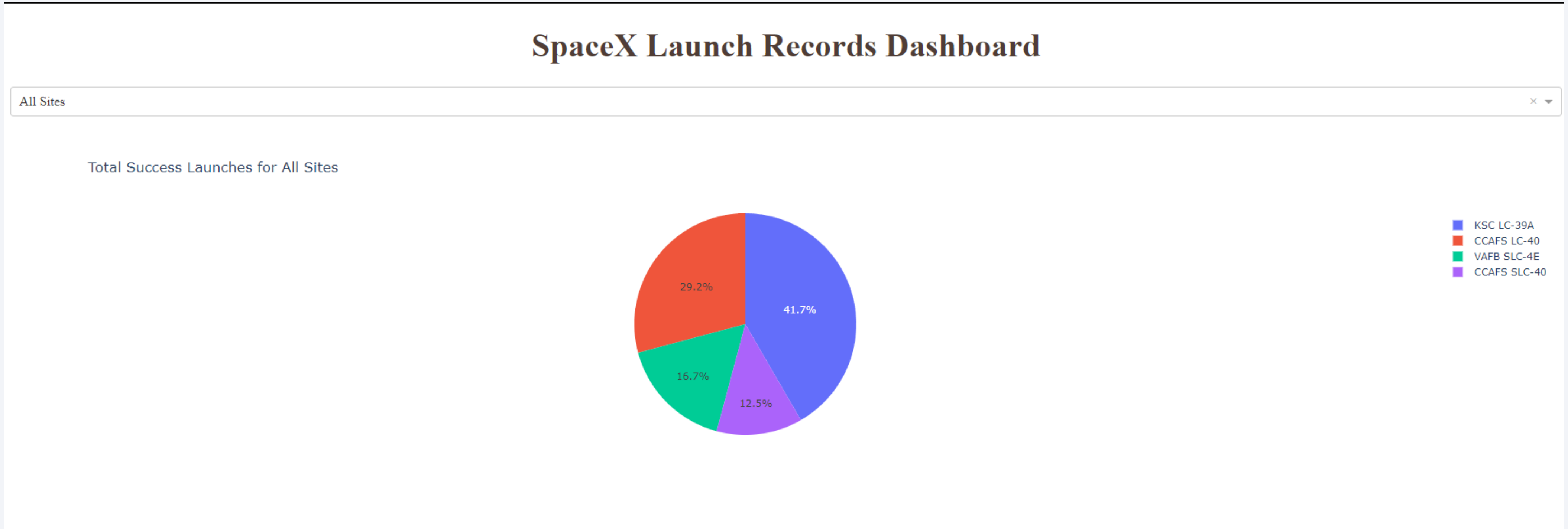


This is a screenshot of the Folium map indicating the distance of the launch site to important elements, such as railway, highway, coastline and the closest city.

Section 4

Build a Dashboard with Plotly Dash

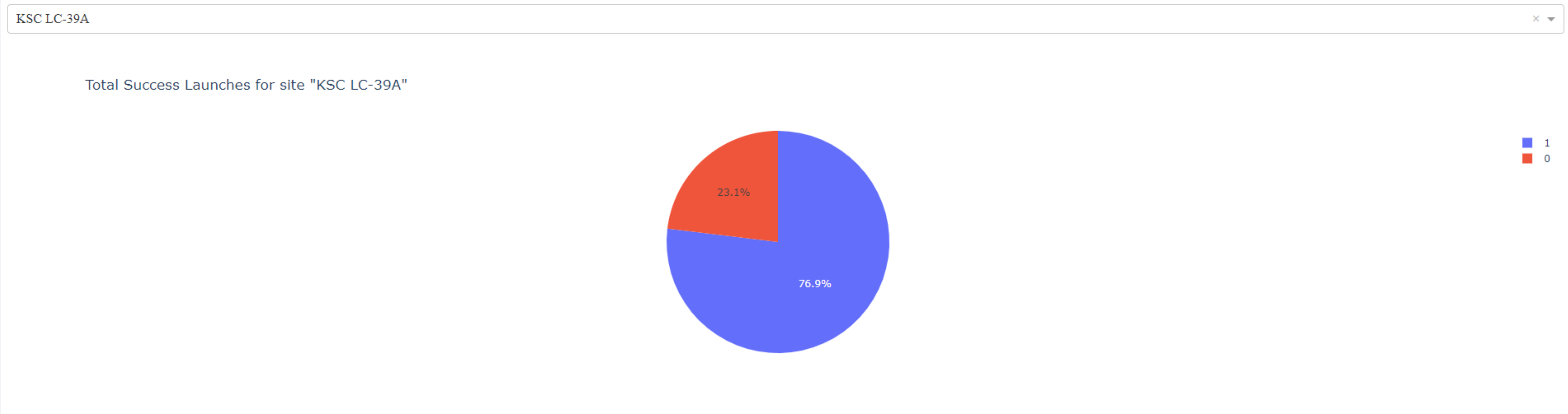
Pie Chart – Launch Success for All Sites



In this pie chart we can see the launch success rates for all sites. Each site is represented by a different color, indicated on the right side of the screen.

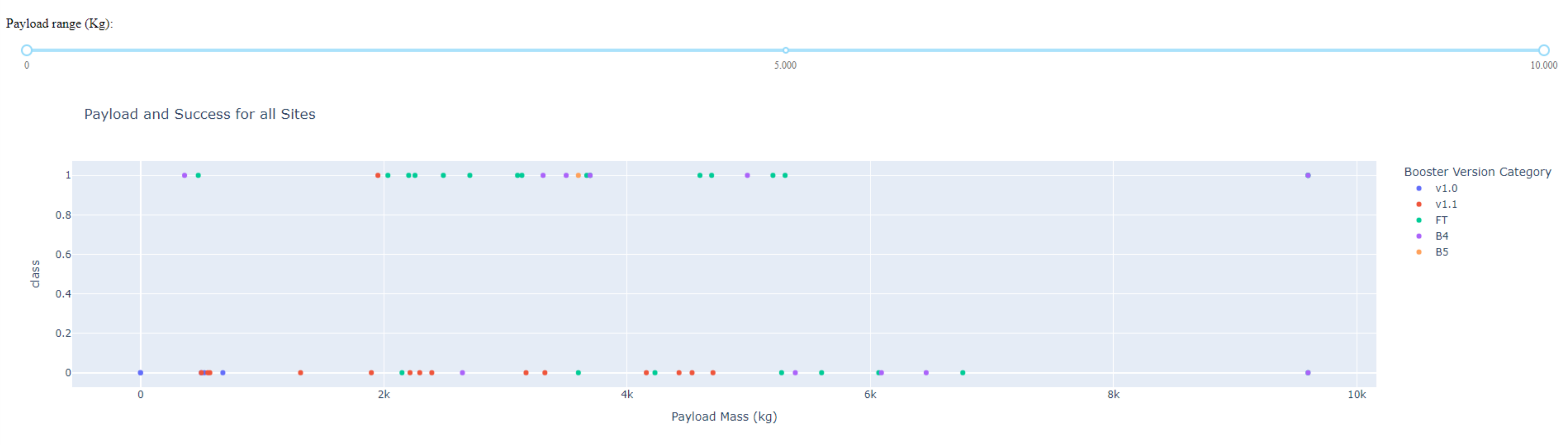
By analyzing this chart, we can determine which sites had the highest launch success rates.

Pie Chart of the highest Launch Success ratio Site



Here is the pie chart for the site with the highest launch success ratio, KSC LC-39A. We can see that the success rate is 76.9%, as indicated by the blue color with a label of "1".

Interactive Scatter Plot for Payload and Success



This is an interactive scatter plot with Payload Mass x Success Rate for all sites, with each entry marked by a color, as to indicate the booster version on the right side. It's possible to filter the Payload Mass range by using the slider above.

By analyzing this plot, we can see that how each booster version performs under different payloads, and which ones have the highest success rates.

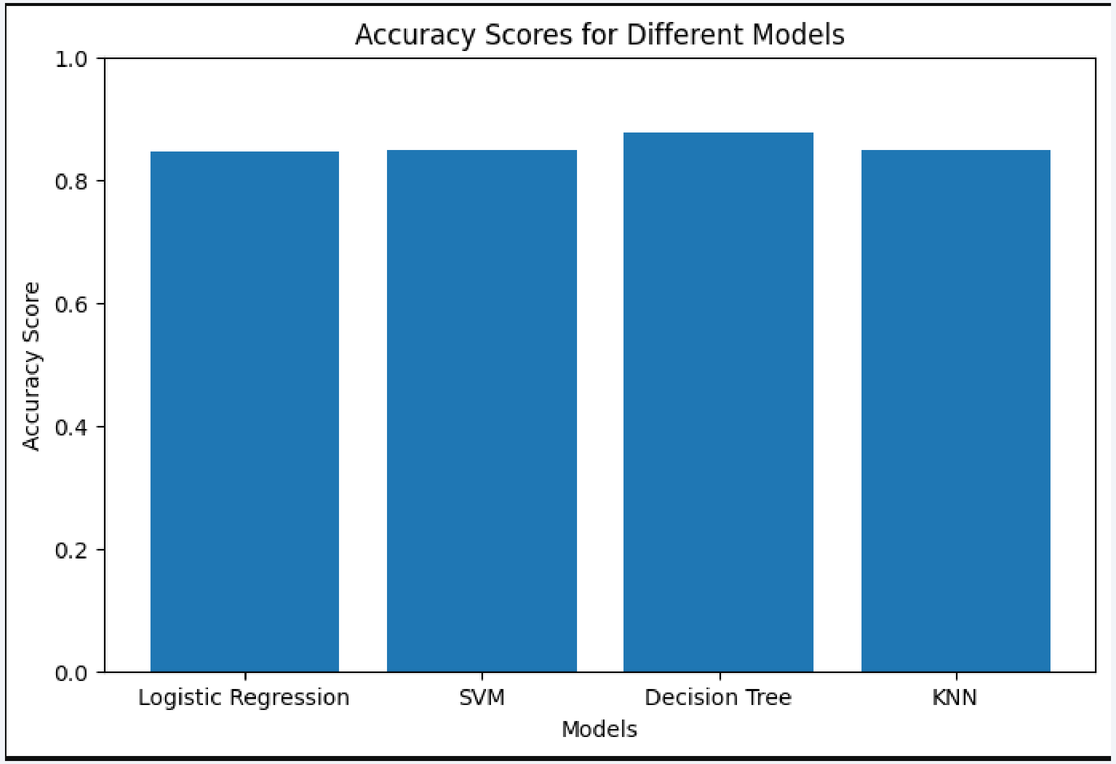
Section 5

Predictive Analysis (Classification)

Classification Accuracy

This is a bar chart showing each of the models and their final accuracy scores after the hyperparameter tuning.

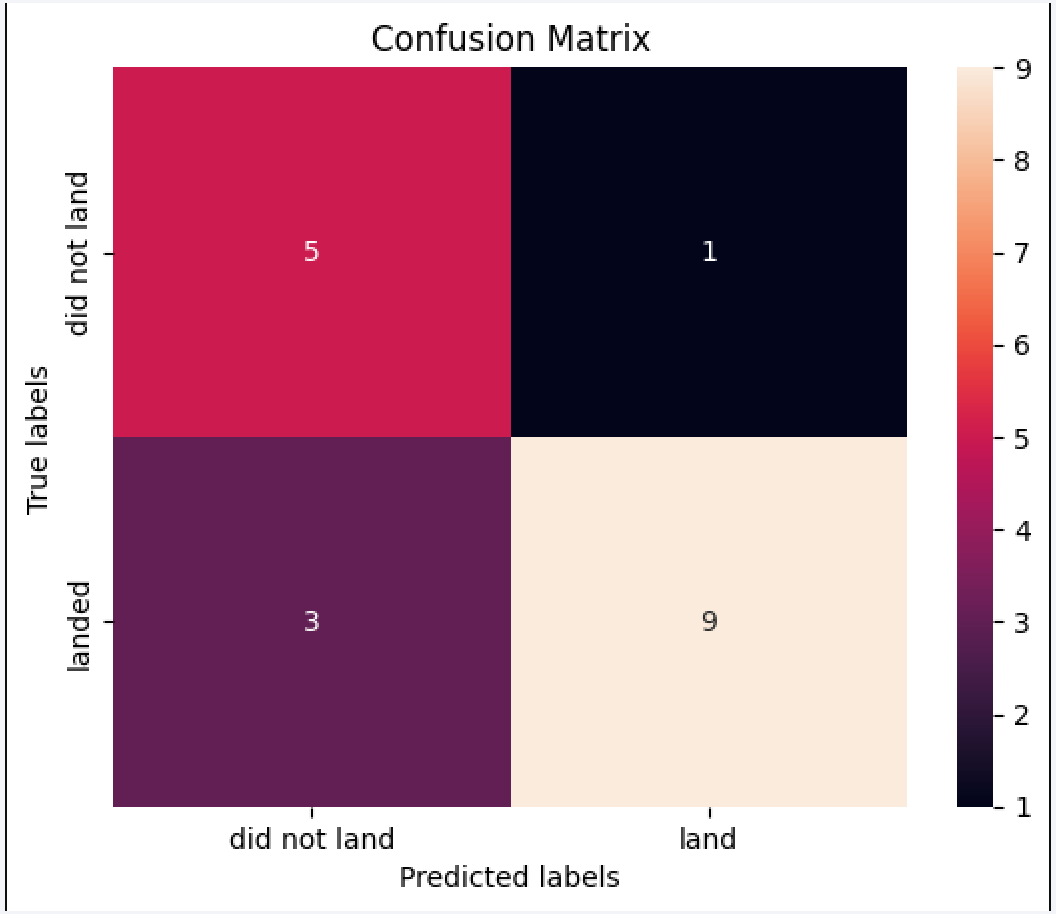
The Decision Tree model had the highest accuracy, reaching 0.876



Confusion Matrix

Here is a Confusion Matrix for the Decision Tree model, which had the highest accuracy.

This matrix compares the output the model predicted vs the true labels for the test set.



Conclusions

- Lower payload mass launches showed better chances of successful landings.
- The success rates of launches are increasing over the years.
- The launch sites are close to the Equator line, close to highways and railroads, and in proximity of the coast.
- The launch site with the highest success rate is KSC LC-39A.
- Based on the lab results, we can conclude that the models' accuracy had a significant improvement after hyperparameter tuning.
- The model with the highest accuracy for this dataset was the Decision Tree.

Appendix

All of the files for this project, including notebook files, code snippets and datasets are organized in this Github Repository.

For further information on the resources and technologies used, check:

- SpaceX API
- BeautifulSoup 4
- Matplotlib
- Seaborn
- Plotly
- Numpy
- Pandas
- Dash
- Folium
- Scikit-Learn

Thank you!