SPACEY

VALENTYN SIANSKYI

GitHub

30 OCTOBER 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Visualizations
- Dashboard
- Discussion
- Conclusion
- Appendix

EXECUTIVE SUMMARY

This project focuses on SpaceY, a rival company to SpaceY, which utilizes data from the SpaceY Falcon 9 rocket to assess first stage landing **success rates and launch costs**. The key methodologies and results are outlined as follows:

- Summary of Methodologies:
  - **Data Collection**
  - **Data Wrangling**
  - **Exploratory Data Analysis (EDA)**
    - Includes data visualization and SQL
  - **Interactive Map Development**
    - Using Folium
  - **Dashboard Construction**
    - With Plotly Dash
  - **Predictive Analysis**
    - Classification models to predict outcomes
- Summary of Results
  - **Exploratory Data Analysis Results**
  - **Interactive Analytics Demonstration**
    - Presented through screenshot
  - **Predictive Analysis Outcomes**

INTRODUCTION

This project explores the prediction of **Falcon 9's first-stage landing success**, a crucial factor in **SpaceY**'s ability to offer rocket launches at a significantly **lower cost** compared to competitors.

While *SpaceY* charges **$62 *million*** per launch, other providers offer similar services at upwards of **$165 *million***, primarily because *SpaceY* can reuse the rocket's first stage.

Accurately predicting whether the first stage will land successfully not only influences the cost of a launch but also provides valuable insights for potential competitors aiming to bid against *SpaceY* for rocket launch contracts.

- The primary focus of this project is to address key questions, such as:
  - **What factors influence the success of a rocket's landing?**
  - **How different rocket variables affect the landing success rate?**
  - **What conditions SpaceY needs to meet to ensure the highest landing success rate?**

  we can help other companies improve their competitive stance against *SpaceY*.

- Through a combination of data collection, data wrangling, and exploratory data analysis, we built interactive tools for visualizing and analyzing the data, such as a Plotly Dash dashboard and a Folium map.
- Machine learning techniques were then employed to predict the likelihood of a successful landing, offering data-driven insights into optimizing rocket performance and minimizing launch costs.

METHODOLOGY

Data Collection Sources:

GitHub API

- Utilized SpaceX REST API to gather launch data.
- Conducted **web scraping** from Wikipedia for additional information on Falcon 9 launches.

GitHub WebScraping

Data Wrangling:

- Transformed and prepared the data for machine learning analysis.
- Applied **One-Hot Encoding** to categorical fields and removed irrelevant columns.

GitHub Data Wrangling

Exploratory Data Analysis (EDA):

GitHub SQL

- Visualized data using scatter plots, bar graphs, and SQL queries to identify relationships and patterns in the variables.

GitHub Visualizations

Interactive Visual Analytics:

GitHub Folium

- Developed an interactive map using **Folium** and created a dashboard with **Plotly Dash** to explore and analyze launch data.

GitHub Dash

Predictive Analysis:

- Built, tuned, and evaluated classification models to predict the success of Falcon 9 first-stage landings.

GitHub Machine Learning

METHODOLOGY

Data Collection Sources:

Utilized SpaceX REST API to gather launch data.

The dataset used in this project was collected from the following source:

- We utilized SpaceY launch data retrieved via the SpaceY REST API.
- The API provided comprehensive details about each launch, including the rocket type, payload, launch specifications, landing information, and the outcome of the landing attempt.
- Our main goal was to analyze this data and predict whether SpaceY will attempt to land the rocket's first stage.
- The SpaceY REST API endpoints are accessed through URLs beginning with *api.spacexdata.com/v4/*.

Request to SpaceY REST API → API returns SpaceX data in .JSON format file → Normalize data into flat data file such as .csv

METHODOLOGY

Extract a Falcon 9 launch records HTML table from Wikipedia, parse the table, and convert it into a Pandas data frame.
To do this, we scrape the data from a snapshot of the "List of Falcon 9 and Falcon Heavy launches" Wikipage updated on 9th June 2021.
Using an HTTP GET request, the page is fetched, and a `Beautiful Soup` object is created to parse the table for further processing in Python.

Data Collection Sources:

Conducted **web scraping** from Wikipedia for additional information on Falcon 9 launches.

Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response:

```
response = requests.get(static_url)
html_response = response.text
print(html_response[:500])
```

Create a `BeautifulSoup` object from the HTML `response` :

```
soup = BeautifulSoup(html_response, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly:

```
# Use soup.title attribute
print(soup.title.string)
```

```
List of Falcon 9 and Falcon Heavy launches - Wikipedia
```

METHODOLOGY

Data Wrangling:

Transformed and prepared the data for machine learning analysis.

Applied One-Hot Encoding to categorical fields and removed irrelevant columns.

In the dataset, various cases show when a Falcon 9 booster either landed successfully or failed.
For example, a "True Ocean" outcome means the booster landed successfully in a designated ocean region,
while "False Ocean" indicates a failed landing attempt in the ocean.
Similarly, "True RTLS" refers to a successful landing on a ground pad, and "False RTLS" indicates a failure on the ground pad.
For drone ships, "True ASDS" signifies a successful landing on a drone ship, while "False ASDS" marks an unsuccessful attempt.
These outcomes are then converted into training labels where 1 represents a successful landing, and 0 indicates failure.

We applied **one-hot encoding** to categorical columns like `Orbit`, `Launch Site`, `Landing Pad`, and `Serial`, resulting in a transformed dataset with 80 columns. Finally, we cast all numeric columns to **float64**.

Exploratory EDA data analysis on a dataset

- Calculation the number of launches at each site
- Calculation the number and occurrence of each orbit
- Calculation the number and occurrence of mission outcome per orbit type
- Creation a landing outcome label from Outcome column
- Determine the success or failure rate
- Export dataset as .CSV

METHODOLOGY

Exploratory Data Analysis (EDA):

Visualized data using scatter plots, bar graphs, and SQL queries to identify relationships and patterns in the variables.

For the Exploratory Data Analysis (EDA), I performed several key steps:

- Data Visualization:
  - I created **scatter plots** to investigate the correlation between payload mass and the success of launches. This helped identify trends, such as how payload mass might affect the likelihood of landing success.
  - **Bar graphs** were used to compare the success rates across different launch sites and boosters, providing insight into which locations or types of boosters had higher success rates.
- SQL Queries:
  - I ran **SQL queries** to extract specific data subsets, such as successful vs. failed landings, as well as comparing launches based on their landing locations (ocean, ground pad, drone ship).
  - By filtering and grouping the data, I could better understand patterns in outcomes and booster performances.

Through these steps, I explored relationships between variables like **payload mass**, **launch site**, and **landing success**, using visualizations and queries to uncover underlying patterns.

METHODOLOGY

Exploratory Data Analysis (EDA):

Visualized data using scatter plots, bar graphs, and SQL queries to identify relationships and patterns in the variables.

Summary of tasks performed using SQL queries:

1. Displayed unique launch sites.
2. Displayed 5 records where launch sites start with "CCA."
3. Calculated the total payload mass carried by NASA (CRS) – 45,596 kg.
4. Calculated the average payload mass for booster version F9 v1.1 – 2534.67 kg.
5. Found the date of the first successful ground pad landing – December 22, 2015.
6. Listed boosters with successful drone ship landings and payloads between 4000 and 6000 kg.
7. Counted the total number of missions – 101.
8. Displayed boosters that carried the maximum payload mass – 15,600 kg.
9. Displayed records of failed drone ship landings for 2015.
10. Ranked landing outcomes by count between 2010 and 2017.

METHODOLOGY

Exploratory Data Analysis (EDA):

Visualized data using scatter plots, bar graphs, and SQL queries to identify relationships and patterns in the variables.

Interactive Map with Folium

We created an **interactive map** using **Folium** to visualize the SpaceY launch data.
The map marks each **launch site** with **circle markers** representing their coordinates, and labels are added for the launch site names.
Using **Marker Cluster**, launch outcomes (successes and failures) are displayed, with **green markers** for successes (class 1) and **red markers** for failures (class 0).
Additionally, using **Haversine's formula**, we calculated the distance from launch sites to nearby landmarks and drew lines on the map to analyze geographical patterns.

Key trends identified:

- **Proximity to railways**: No
- **Proximity to highways**: No
- **Proximity to coastline**: Yes
- **Distance from cities**: Yes

This analysis allowed us to extract critical insights from the data, setting the foundation for further prediction models on launch success.

METHODOLOGY

Interactive Visual Analytics:

Developed an interactive map using **Folium** and created a dashboard with **Plotly Dash** to explore and analyze launch data.

Interactive Dashboard with Flask and Dash

We created an **interactive dashboard** using **Flask** and **Dash**, hosted on **Python Anywhere** for 24/7 access to SpaceY launch data.

Key Features

- Total Launches Pie Chart
  - Displays total launches by site, with slice sizes proportional to the number of launches, illustrating site activity.

1. Scatter Graph
   - Shows the relationship between **launch outcomes** and **payload mass** (kg) for different booster versions, highlighting trends and patterns.

Summary

This dashboard allows users to explore and visualize SpaceY launch data in real time, offering insights into launch trends and relationships.

METHODOLOGY

Predictive Analysis:

Built, tuned, and evaluated classification models to predict the success of Falcon 9 first-stage landings.

Predictive Analysis (Classification)

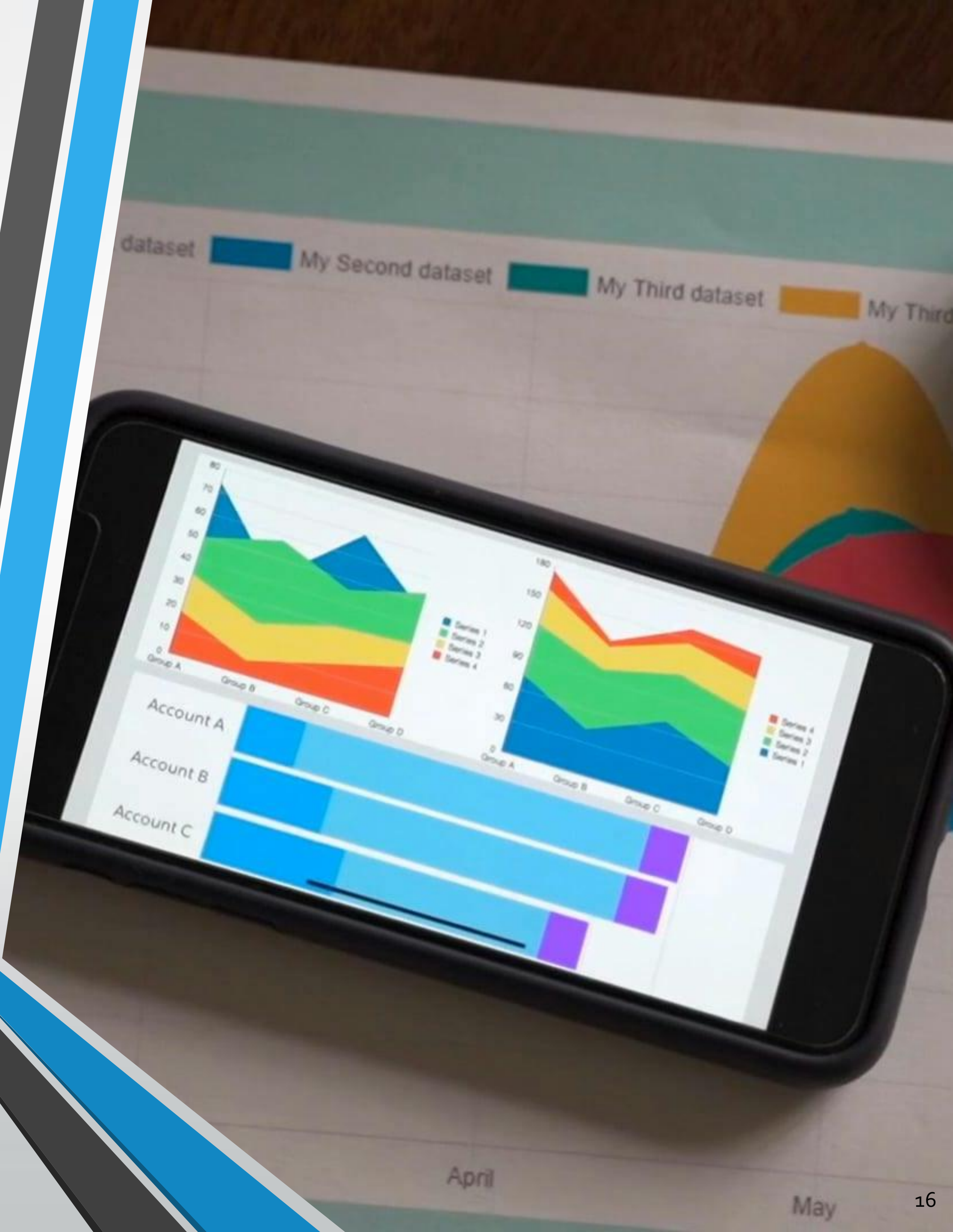
- **Building Model:**
  - Loaded the SpaceX dataset into NumPy and Pandas.
  - Transformed categorical variables using One-Hot Encoding.
  - Split data into training (80%) and test (20%) sets.
  - Selected classification algorithms (e.g., Logistic Regression, Random Forest) and optimized hyperparameters with GridSearchCV.

- **Evaluating Model:**
  - Assessed accuracy for each model and extracted tuned hyperparameters.
  - Plotted confusion matrices to visualize performance.
- **Improving Model:**
  - Conducted feature engineering to select relevant features.
  - Tuned algorithms to enhance performance.
- **Finding the Best Performing Model:**
  - Identified the model with the highest accuracy score (e.g., Random Forest at 95%) as the best performer.
  - Included a summary of algorithm scores in the notebook for comparison.

RESULTS

- ❖ Visualizations and SQL queries for key insights.
- ❖ Folium map to showcase launch sites.
- ❖ Dynamic interactive dashboard.
- ❖ Classification to identify the best model.

Exploratory Data Analysis (EDA) with Visualizations



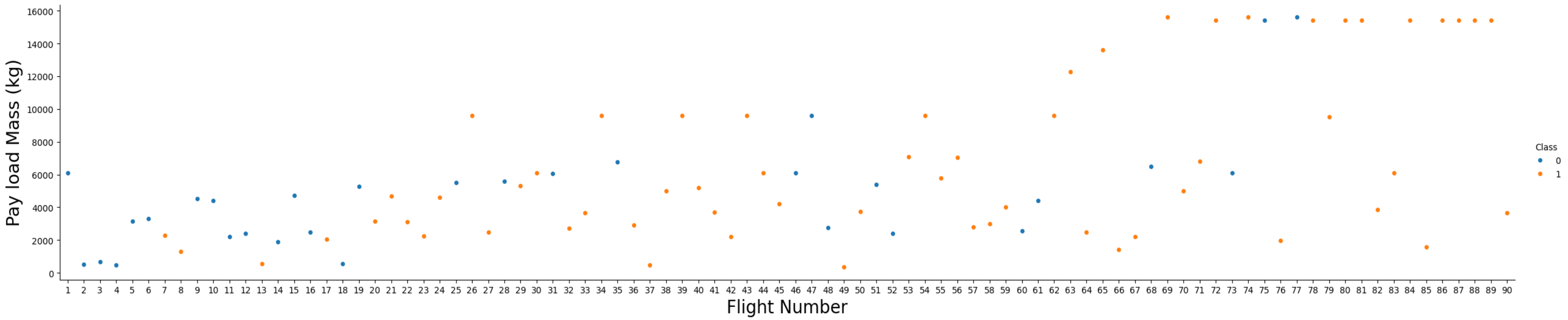
Question:

How do the Flight Number (indicating the number of launch attempts) and Payload Mass affect the outcome of a rocket's first stage landing?

Results:

Launch Success: The graph displays two categories of outcomes — failed landings (blue dots, Class = 0) and successful landings (orange dots, Class = 1). As the flight number increases, there is a clear trend towards more successful landings of the first stage.

Impact of Payload Mass: Heavier payloads are mostly found in the successful landing category, but for some payloads over 10,000 kg, there are still failed landings. This suggests that while higher payload mass may reduce the likelihood of a successful landing, it is not the only determining factor.



Findings:

As the number of launches increases, the probability of a successful landing improves. However, heavier payloads seem to introduce more risk, as evidenced by some failed landings with larger masses, although this is not a consistent rule.

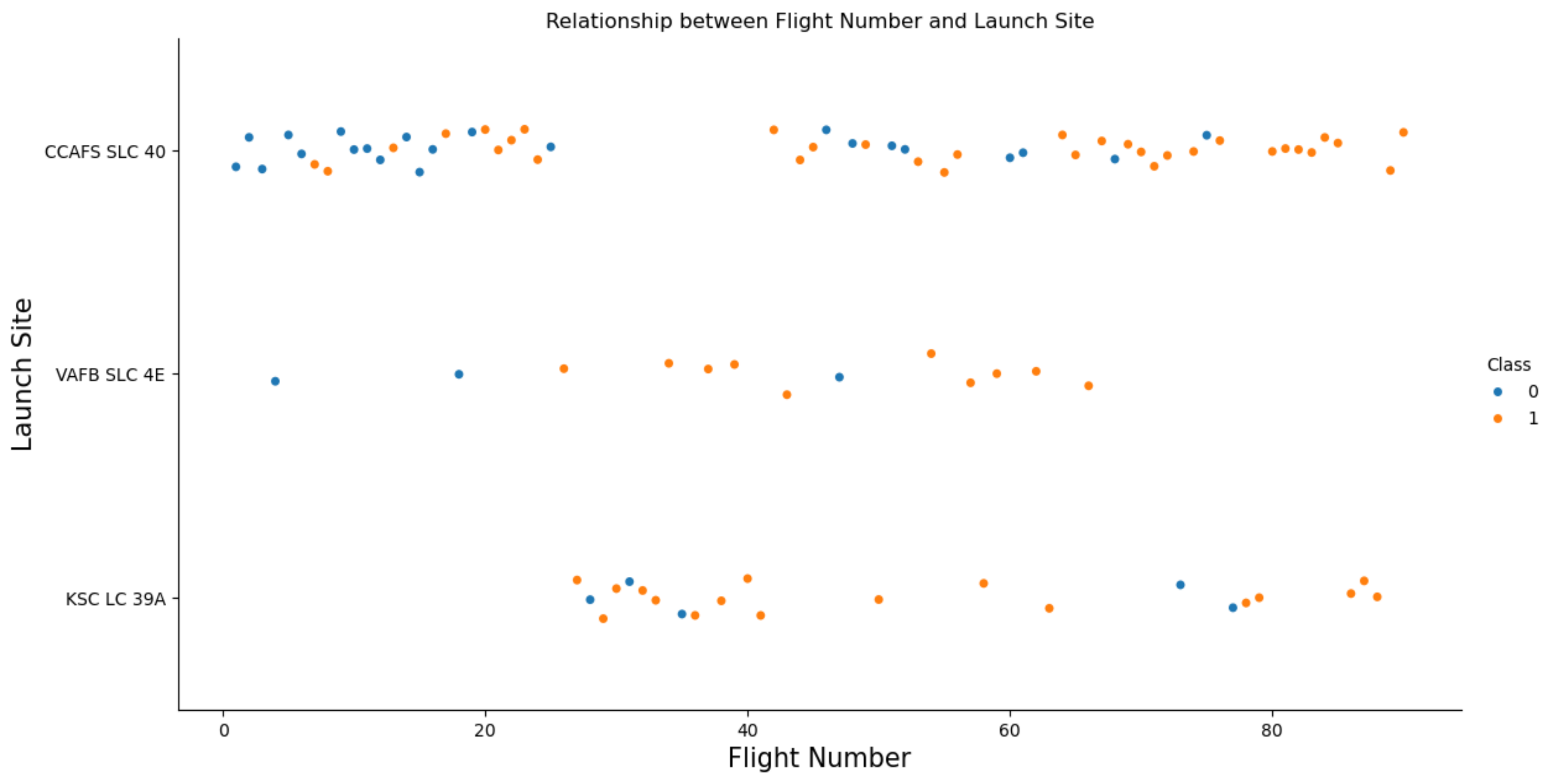
Question:

What is the relationship between the Flight Number and the Launch Site, and how does it impact the success of the rocket landings?

Explanation:

To explore this, I used Seaborn's `catplot` to create a scatter plot with `Flight Number` on the x-axis and `Launch Site` on the y-axis, while the `Class` (successful or failed landing) is represented by color.

- **Class 1 (orange)** indicates a successful landing, while **Class 0 (blue)** indicates a failure.

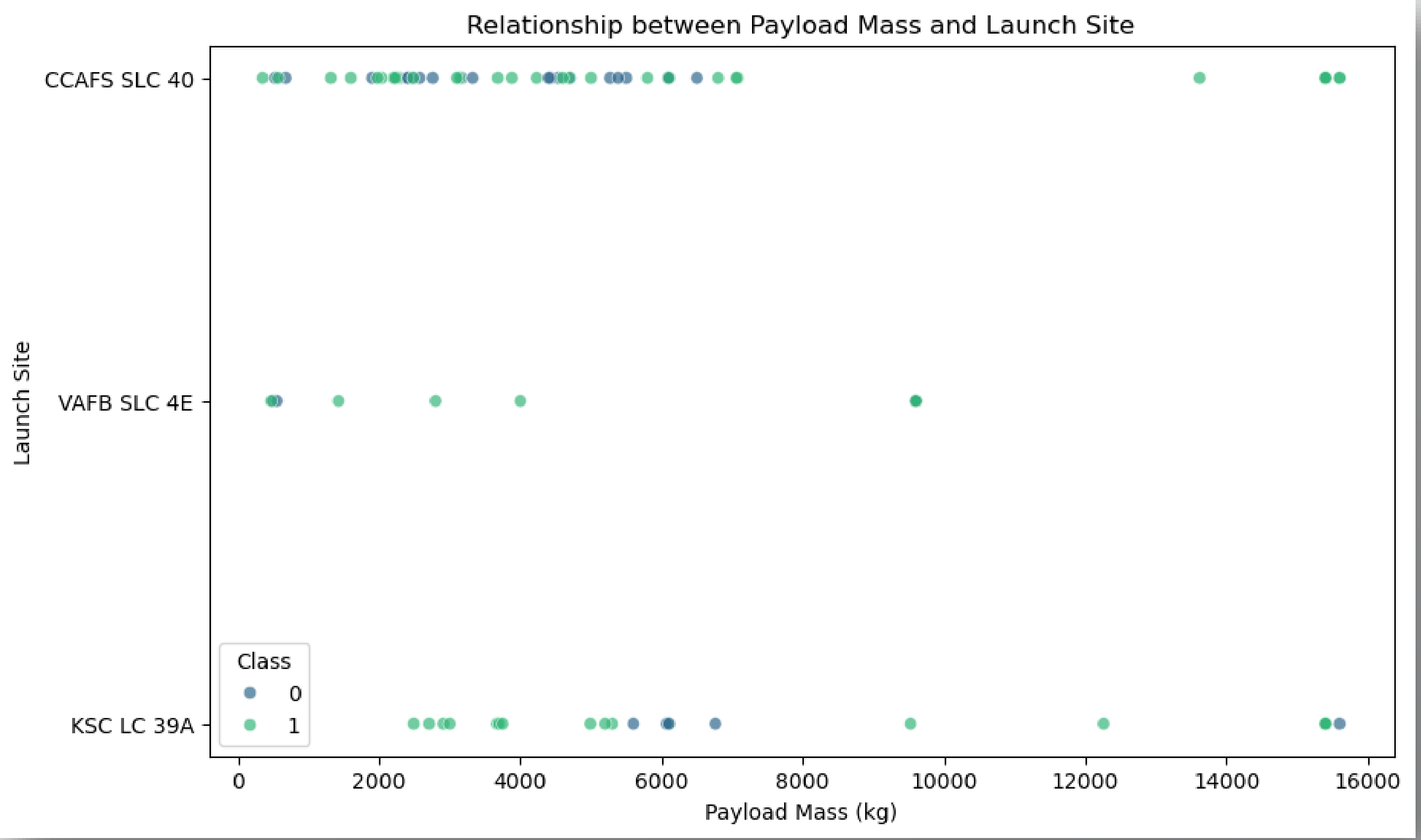


Findings:

The plot shows that most launches from **CCAFS SLC 40** and **KSC LC 39A** have a mix of successes and failures, though the frequency of successes increases with higher flight numbers.
For **VAFB SLC 4E**, fewer launches occurred, but it shows some successful landings as well.
The overall trend suggests that as the number of flights increases, there is a higher chance of success, but this varies depending on the launch site.

Question:

What is the relationship between Payload Mass and Launch Site, and how does it affect the success of the rocket landing?



Explanation:

To visualize this, I used a scatter plot with Payload Mass on the x-axis and Launch Site on the y-axis. The hue is set to Class to indicate the success (Class 1) or failure (Class 0) of the landing.

Findings:

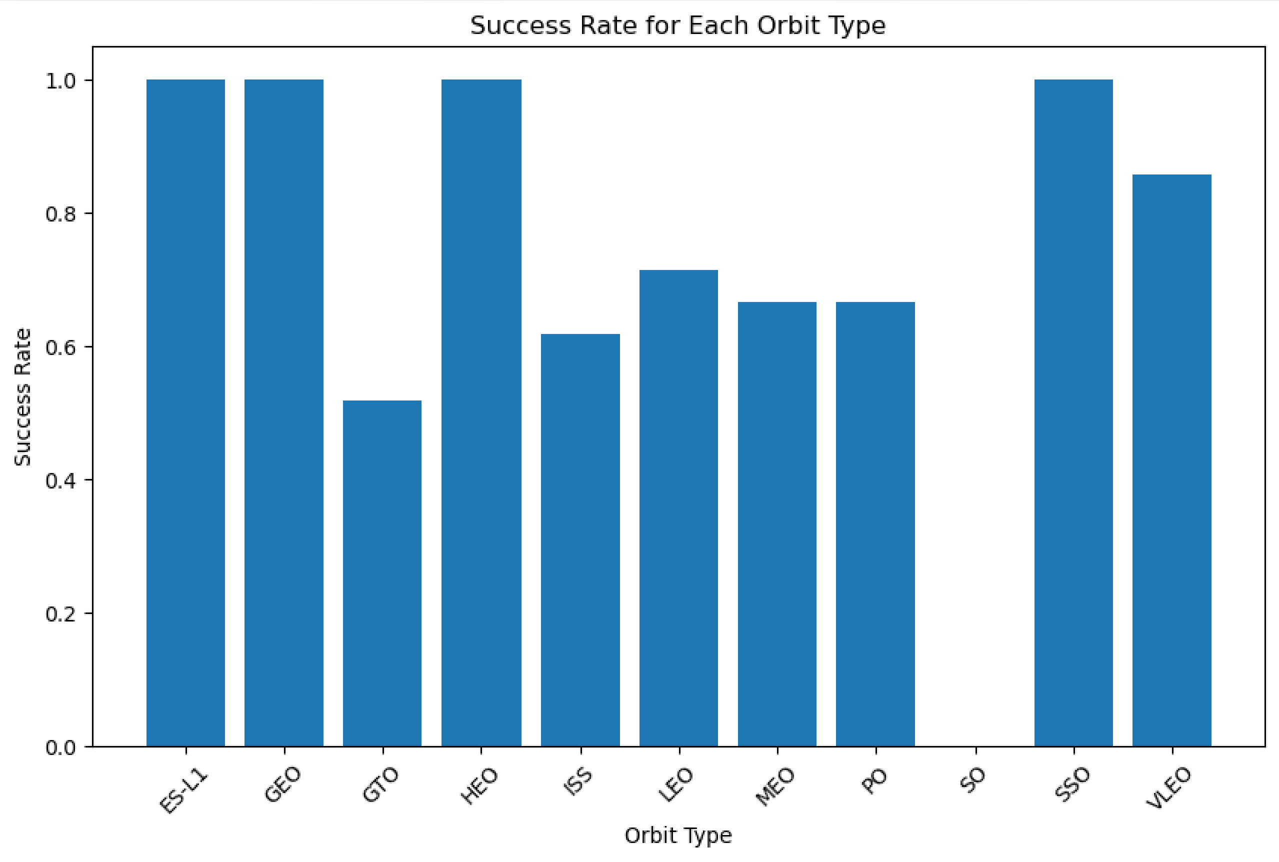
The plot shows that the **CCAFS SLC 40** site handled a wider range of payload masses, but both successful and failed landings occurred across this range. **KSC LC 39A** had more successful landings at lower payload masses, while **VAFB SLC 4E** handled fewer launches, showing limited but mostly successful outcomes. There doesn't seem to be a clear pattern of payload mass influencing success consistently across all sites, though larger payloads tend to have mixed results.

Question:

What is the success rate for each orbit type, and how does it vary across different orbits?

Explanation:

This bar chart shows the success rate for different orbit types, with each orbit type represented on the x-axis and success rate on the y-axis.



Findings:

- **Highest Success Rates:** ES-L1, GEO, and HEO orbits have a perfect success rate of 1.0.
- **Moderate Success Rates:** LEO, MEO, PO, and VLEO orbits have success rates between 0.7 and 0.9.
- **Lower Success Rate:** GTO has a lower success rate of around 0.5, suggesting challenges with launches.
- **High Success:** SSO also shows a high success rate near 1.0.

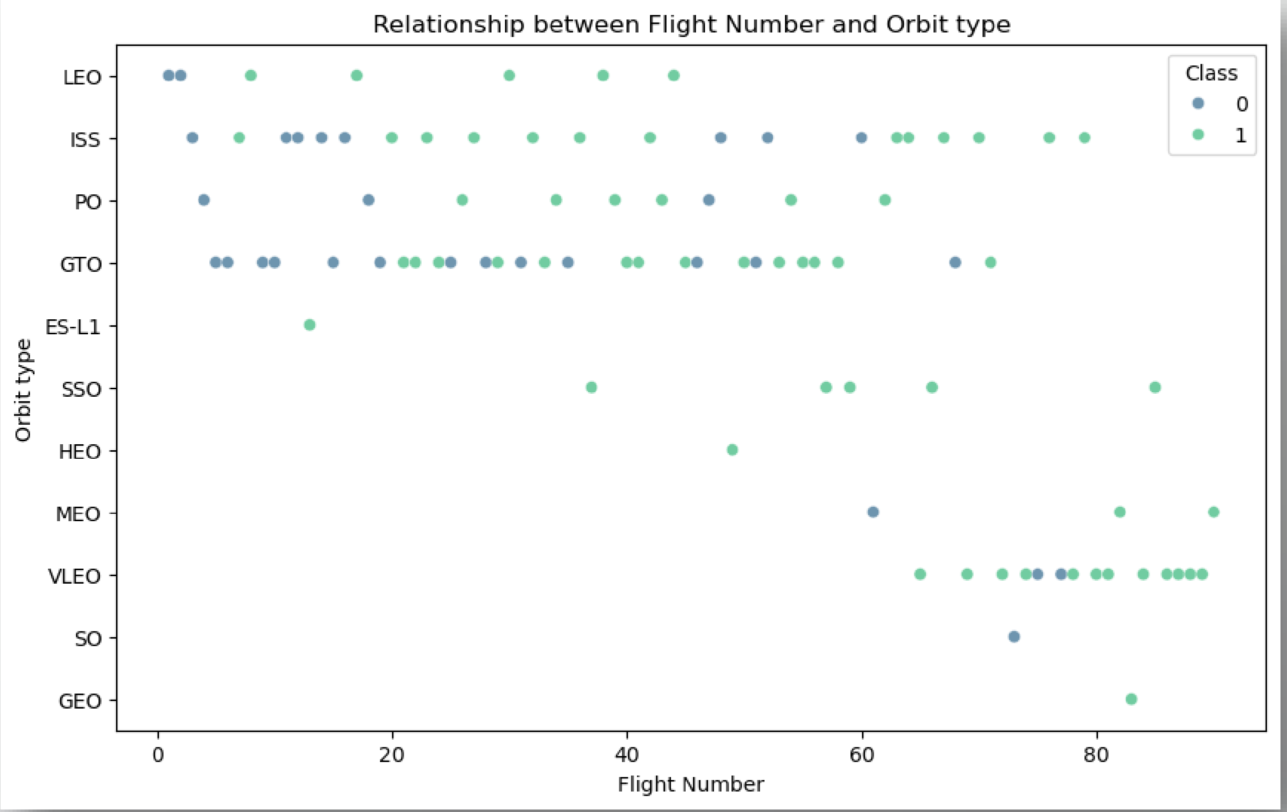
In summary, while most orbits are successful, GTO faces the most challenges, likely due to its higher energy requirements.

Question:

What is the relationship between flight number and orbit type, and how does it impact the success of missions?

Explanation:

This scatter plot shows flight numbers on the x-axis and different orbit types on the y-axis. Each point represents a mission, with the color indicating its outcome: Class 0 for failure and Class 1 for success.



Findings:

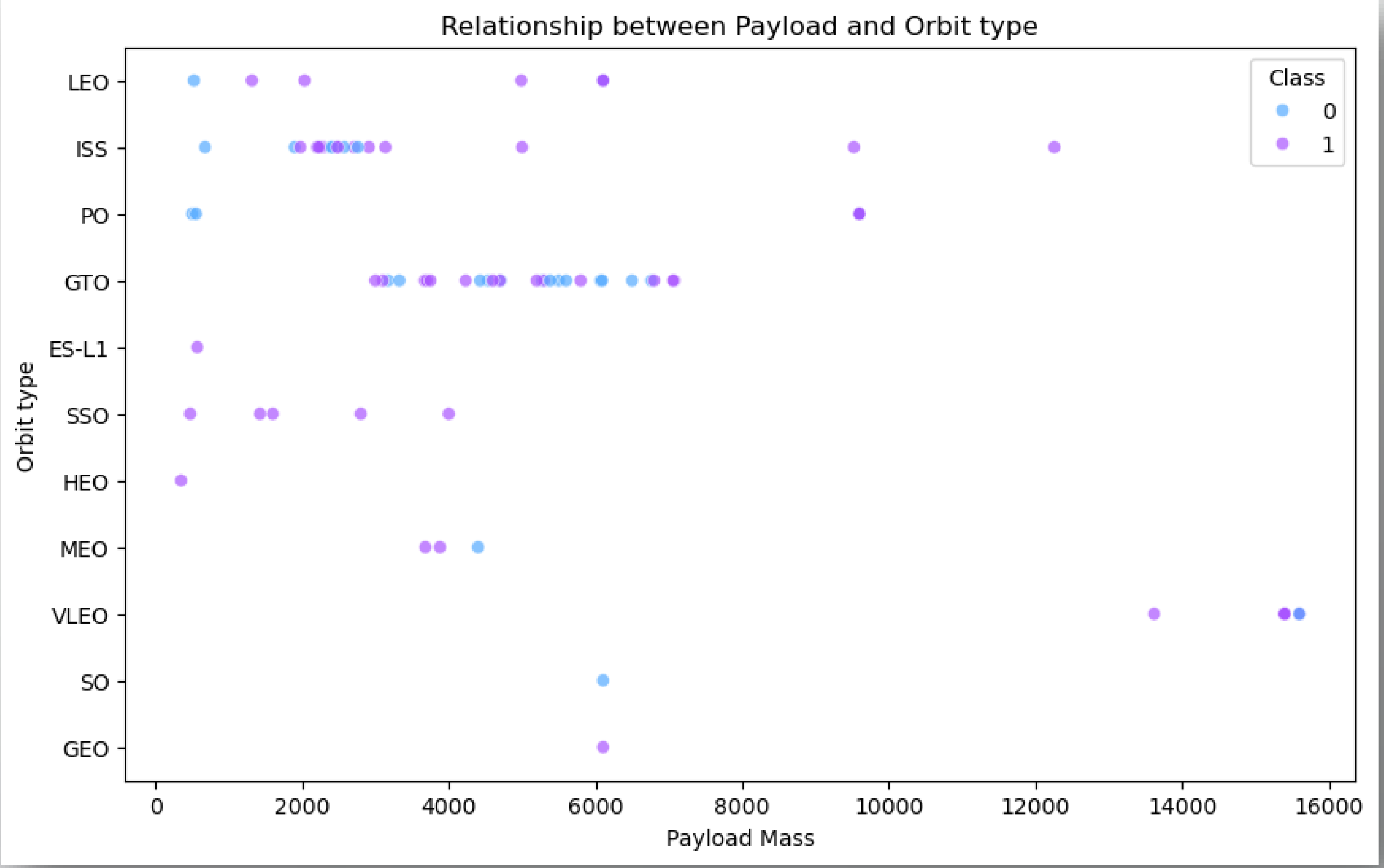
- **LEO, ISS, PO:** Mixed success and failure rates across flight numbers, indicating no strong link to experience.
- **GTO:** Higher failure rate, especially at lower flight numbers, suggesting early challenges.
- **SSO, VLEO:** Mostly successful at higher flight numbers, showing improved reliability.
- **GEO, ES-L1:** Limited data but high success rates, suggesting stable performance.
- **Summary:** Success rates generally improve with flight experience, highlighting the role of advancements over time.

Question:

What is the relationship between orbit type, payload mass, and class?

Explanation:

The graph shows the relationship between payload mass (on the X-axis) and orbit type (on the Y-axis). Different colors indicate classes: 0 and 1.



Findings:

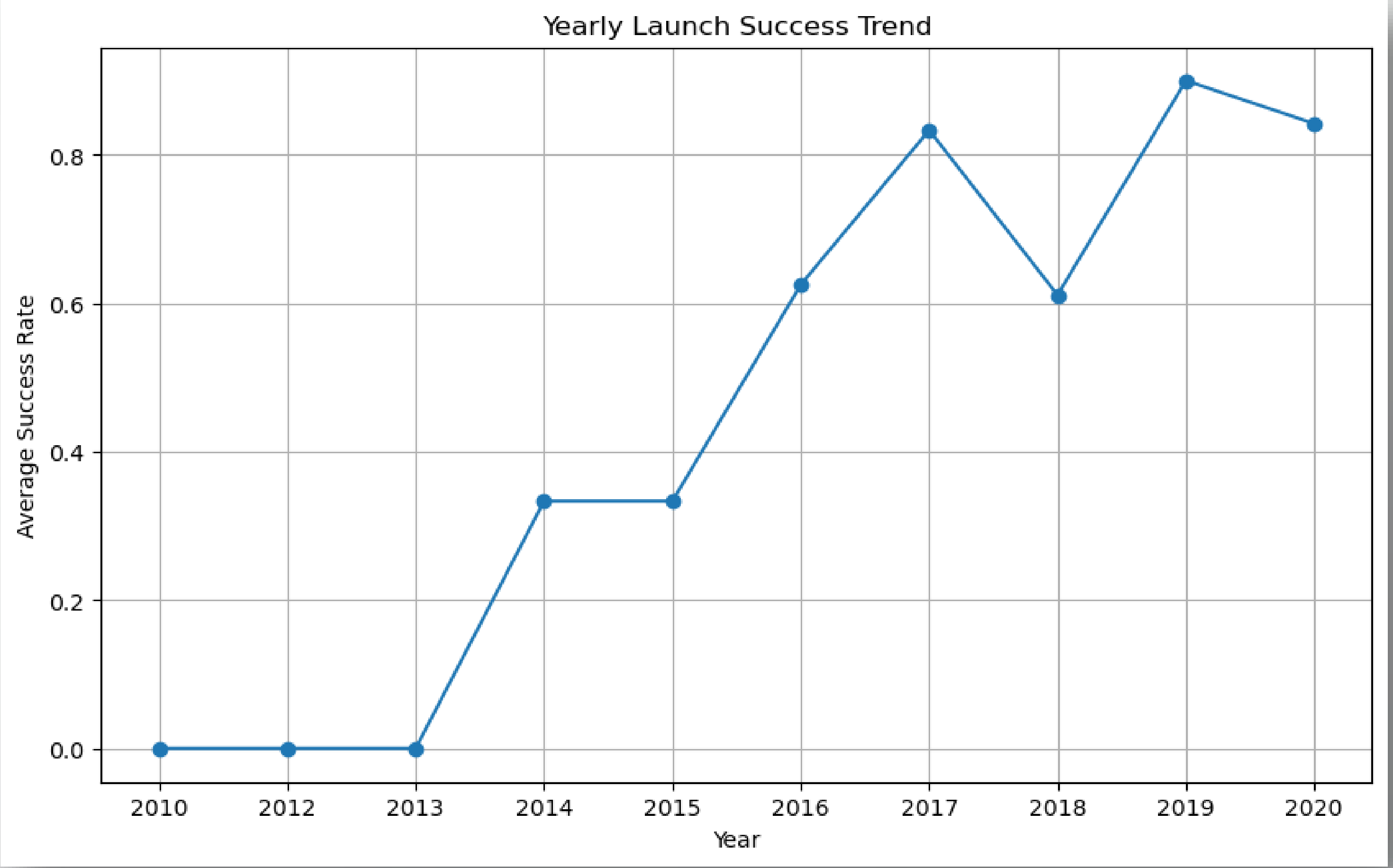
- The graph suggests that certain orbit types (e.g., GTO and LEO) have a higher concentration of launches with varying masses. The class appears to correlate with specific orbit types.

Question:

What is the trend in launch success rates over the years?

Explanation:

The graph displays the yearly average success rate of launches from 2010 to 2020.



Findings:

- The success rate shows an upward trend, with notable increases after 2015, reaching a peak in 2019. This suggests improvements in launch reliability over time.

Exploratory Data Analysis (EDA) with SQL

```sql
SELECT DISTINCT Launch_Site
  FROM SPACEXTABLE;
```

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

- Names of the unique launch sites in the space mission

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

Total_Payload_Mass
45596

- Total payload mass carried by boosters launched by NASA (CRS)

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

- 5 records where launch sites begin with the string 'CCA'

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

```sql
SELECT
AVG(PAYLOAD_MASS__KG_) AS
AVG_Payload_Mass
  FROM SPACEXTABLE
WHERE Booster_Version LIKE
'F9 v1.1%'
```

AVG_Payload_Mass
2534.6666666666665

- Average payload mass carried by booster version F9 v1.1

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

MIN(Date)
2015-12-22

- Date when the first succesful landing outcome in ground pad was acheived.

```sql
SELECT Booster_Version
  FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success
(drone ship)'
     AND PAYLOAD_MASS__KG_ > 4000
     AND PAYLOAD_MASS__KG_ < 6000;
```

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

- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```sql
SELECT
count(Mission_Outcome) as
Total_Number_Missions
  FROM SPACEXTABLE
```

Total_Number_Missions
101

- Total number of successful and failure mission outcomes

```sql
SELECT Booster_Version, PAYLOAD_MASS__KG_
    FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                    FROM SPACEXTABLE);
```

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

- Names of the booster_versions which have carried the maximum payload mass.

```sql
SELECT
   strftime('%m', Date) AS Month,
   Landing_Outcome,
   Booster_Version,
   Launch_Site
     FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure (drone ship)'
     AND substr(Date, 0, 5) = '2015';
```

- The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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

```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;
```

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

- 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 distinct  Launch_Site
   FROM SPACEXTABLE
```

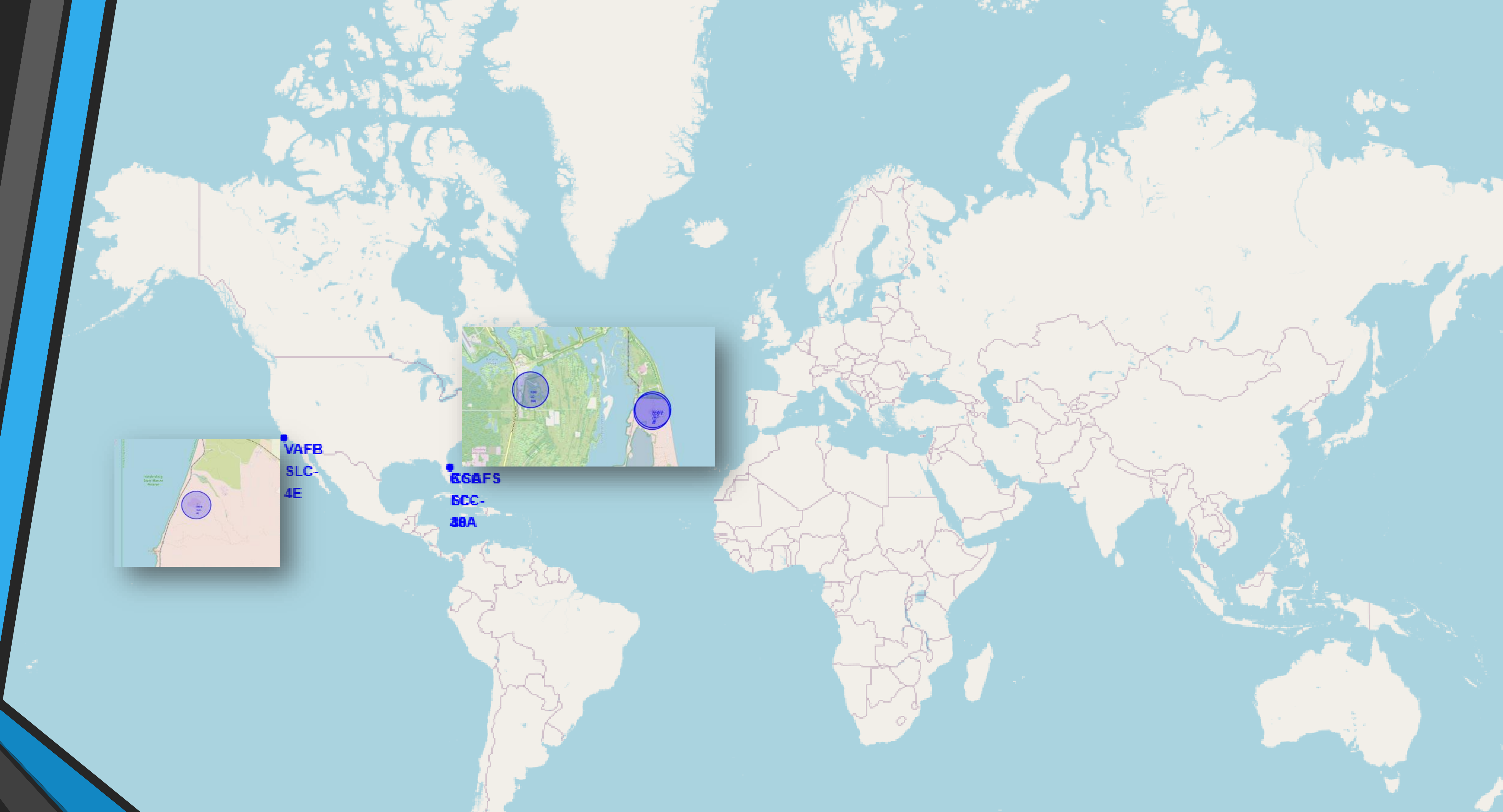
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- Unique Launch Site.

Interactive Map with Folium

The launch success rate may depend on many factors such as payload mass, orbit type, and so on.
It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

<The GitHub link of the Cognos dashboard goes here.>



As we can see that the SpaceX launch sites are in the United States of America coasts.
Florida and California

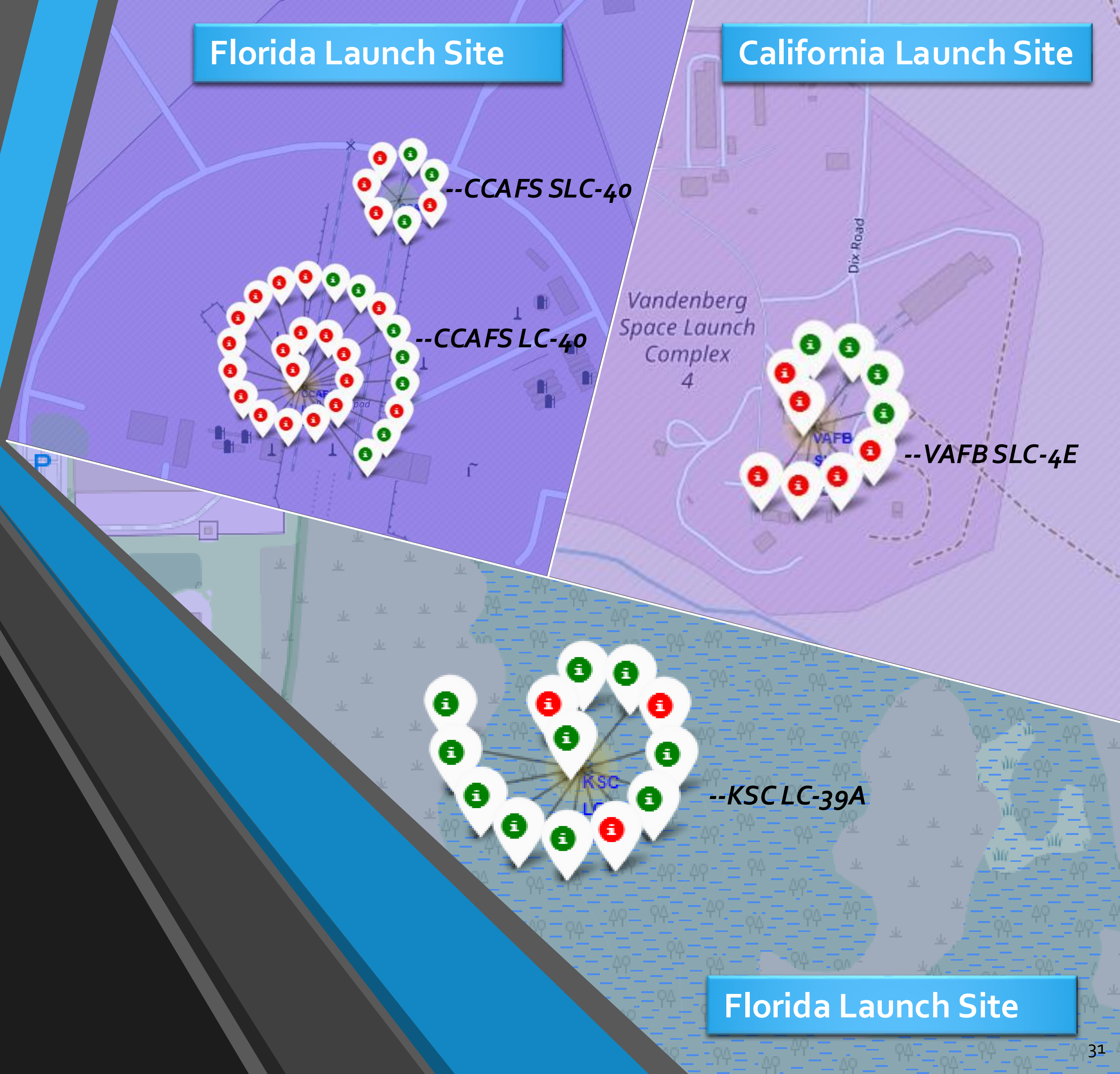
Markers

Key trends identified:

- **Proximity to railways**: No
- **Proximity to highways**: No
- **Proximity to coastline**: Yes
- **Distance from cities**: Yes
- Distance from airports: Yes

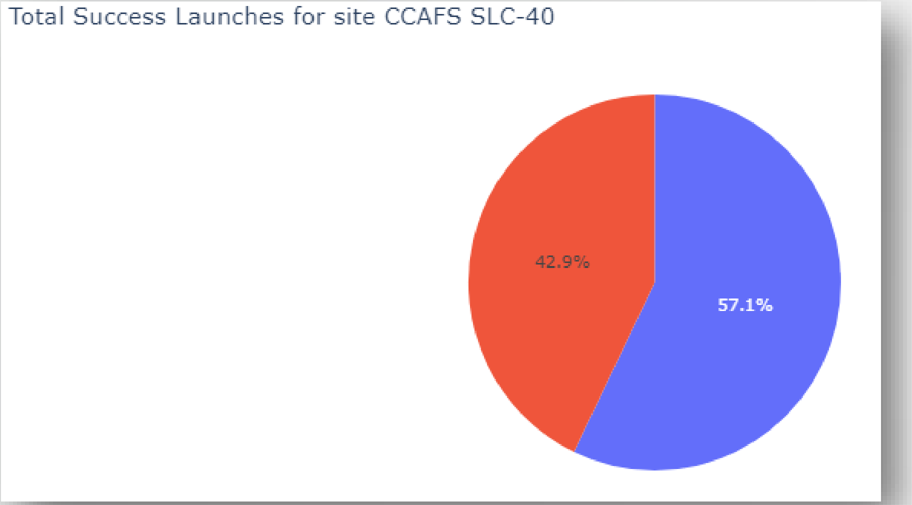
- Successful Launches

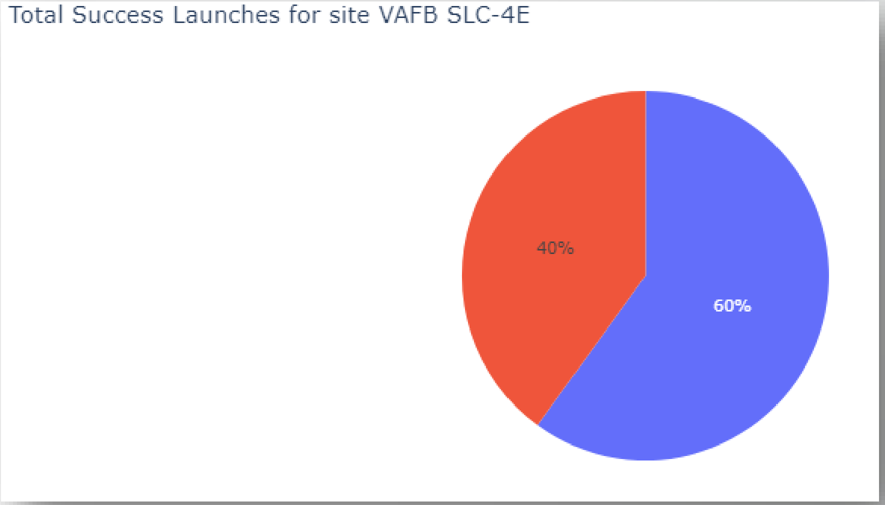
- Failures



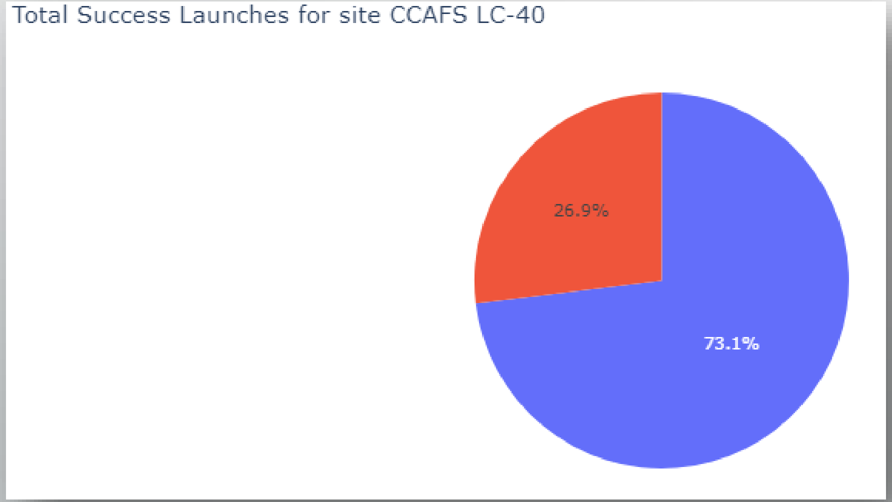
Dashboard with Plotly Dash

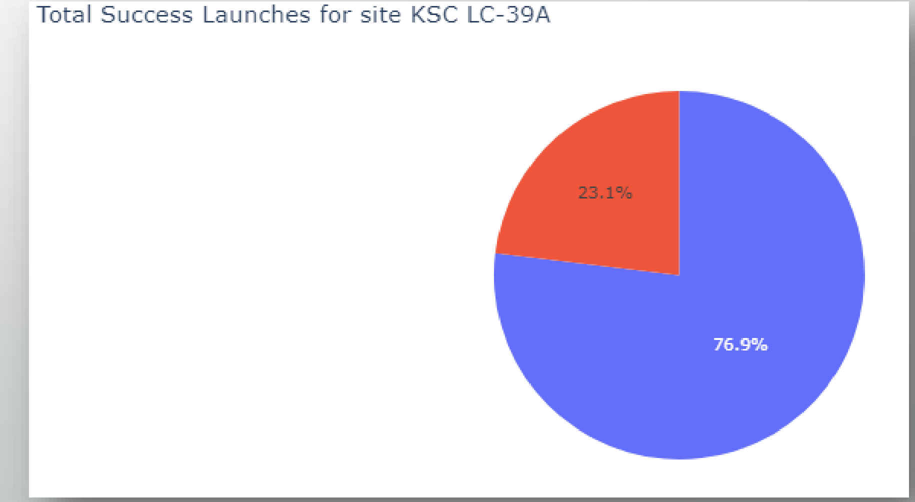
Pie chart displaying the success percentage of each launch site.





It can be observed that **KSC LC-39A** had the **most successful** launches among all the sites.

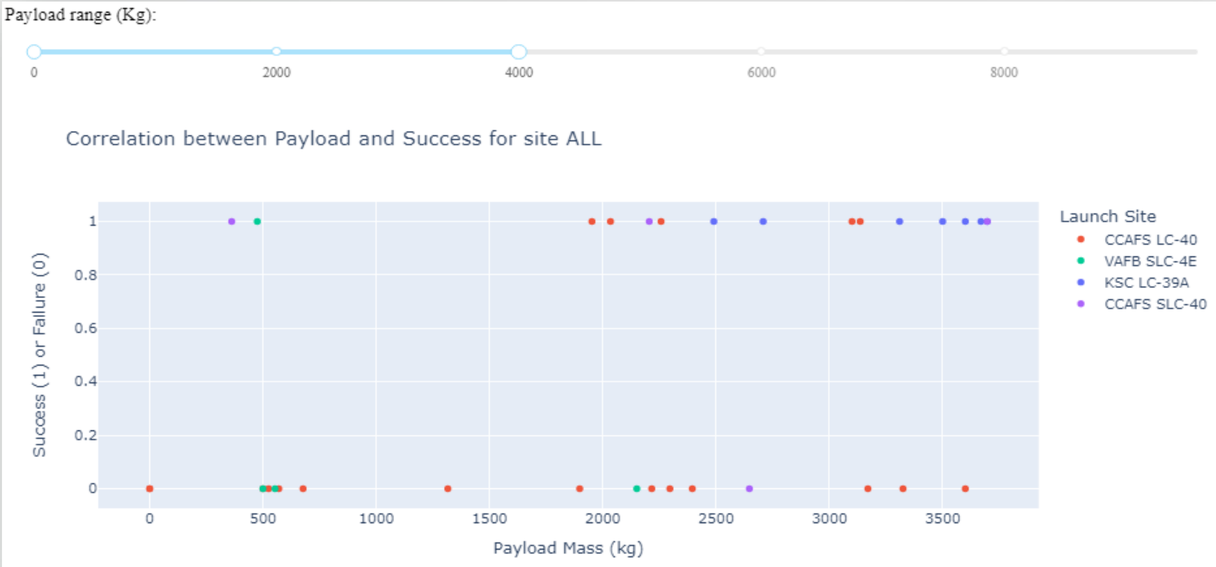




--Success

--Failure

Scatter plot of Payload *vs.* Launch Outcome for all sites, with a range slider to select different payload values.



It can be observed that **success** rates are **higher** for **lower-weight** payloads *compared* to heavier ones.

DISCUSSION

Predictive Analysis (Classification)

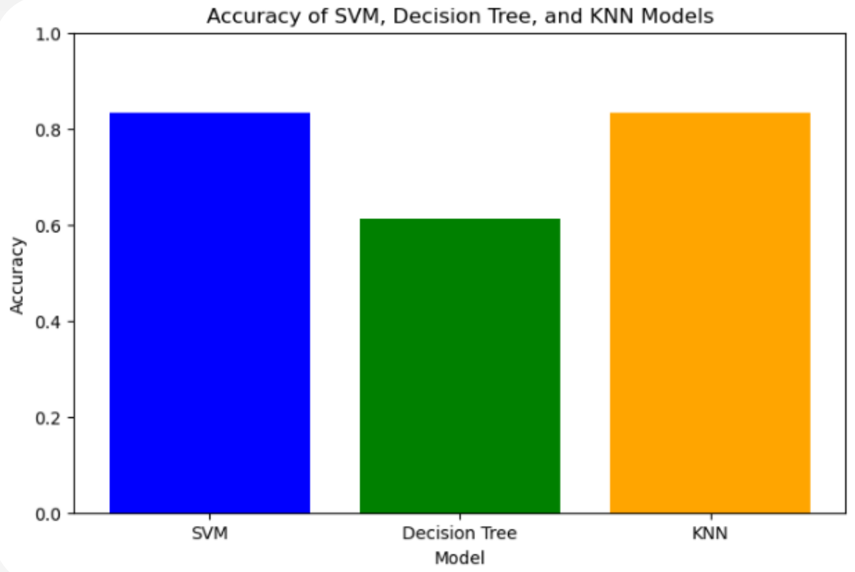
Classification Accuracy using Training Data.

```
# Create a dictionary of models and their accuracy
accuracies = {
    "SVM": svm_test_accuracy,
    "Decision Tree": tree_test_accuracy,
    "KNN": knn_test_accuracy
}

# Find the best model in terms of accuracy
best_model = max(accuracies, key=accuracies.get)

print(f"The best model is {best_model} with accuracy: {accuracies[best_model]}")

The best model is SVM with accuracy: 0.8333333333333334
```



Test accuracy for **SVM**:	0.8333333333333334	- **Best Model**
Test accuracy for **KNN**:	0.8333333333333334	
Test accuracy for **Decision Tree**:	0.6111111111111112	

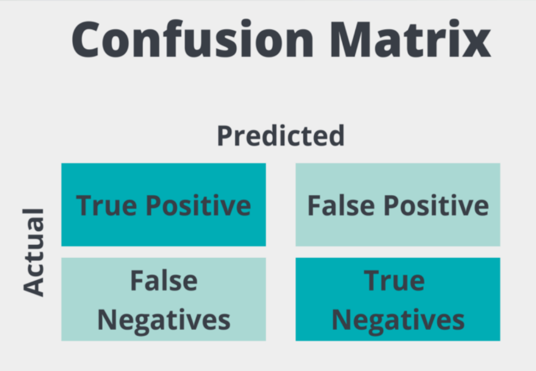
Both models — **SVM** and **KNN** — have the same test accuracy: **0.833**. While Decision Tree shows a lower accuracy of **0.611**.

Since SVM and KNN have the same accuracy, let's consider other aspects:

1. **Big data performance:**
   - SVM usually performs better on large data with a large number of features and is good at solving problems with non-linear boundaries.
   - KNN can be less efficient on large data sets because every time it needs to predict a class, it calculates the distance to each neighboring element, which affects the rate.
2. **Sensitivity to noise:**
   - SVM handles noise better by using optimal hyperplanes.
   - KNN can be sensitive to noise because the sample's nearest neighbors can be misclassified due to anomalous or atypical data.
3. **Setting parameters:**
   - SVM has many settings (such as kernels), which allows for more flexible tuning of the model.
   - KNN has only a few parameters, such as the number of neighbors (k), but choosing the optimal k can be critical.

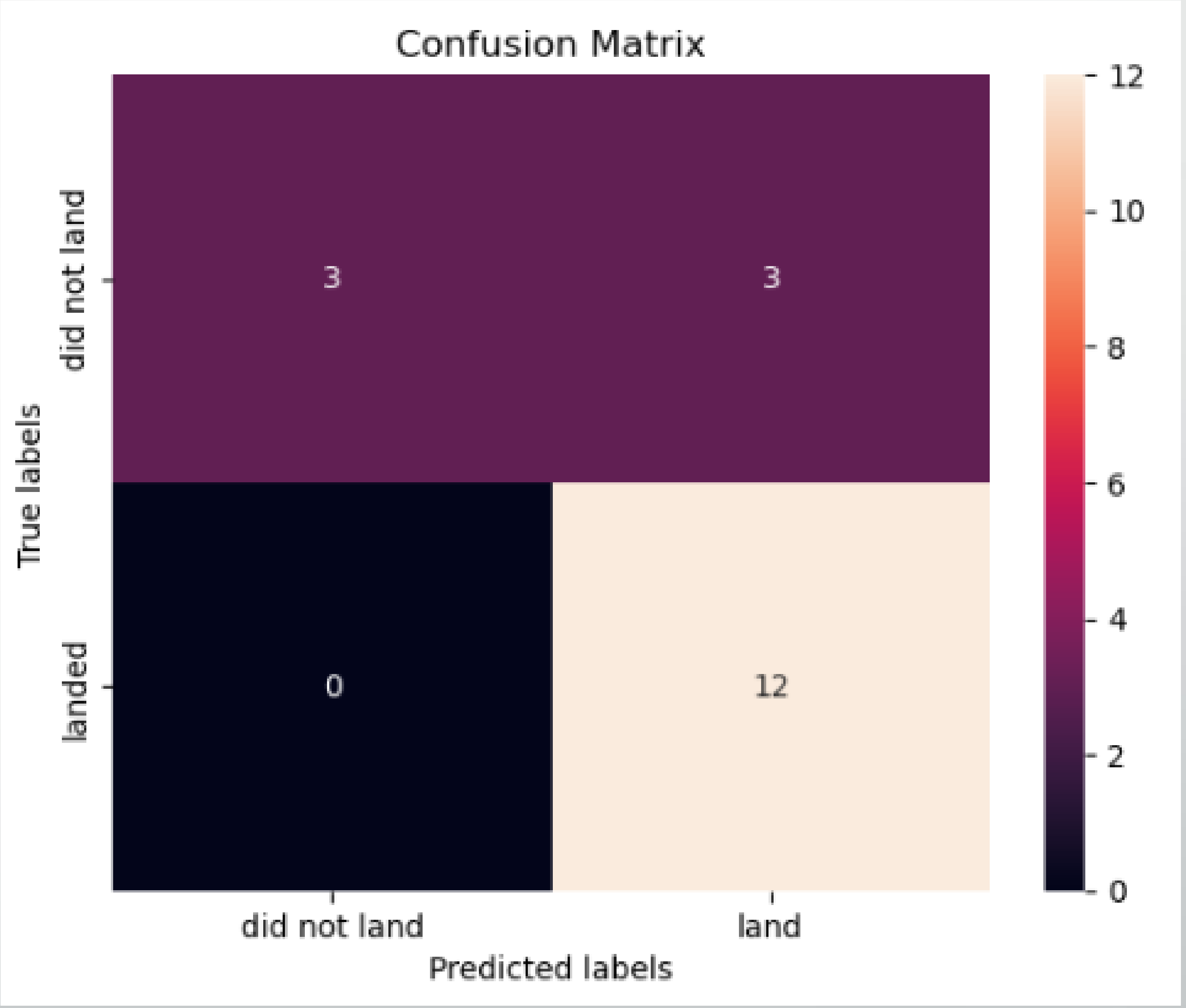
So both models are equal in terms of accuracy, but SVM may be a better option if you need to process more complex data or avoid noise issues.

Confusion Matrix for the SVM



- Accuracy - ≈0.833 (**83.3%**)
- Precision - =0.5 (**50%**)
- Recall - =1.0 (**100%**)
- F1-Score - ≈0.667 (**66.7%**)

The SVM model has good accuracy and recall.



CONCLUSION

The SVM Classifier Algorithm is the most effective for this dataset in Machine Learning.

Lower-weight payloads demonstrate better performance compared to heavier ones.

The success rates of SpaceY launches are directly proportional to the time invested in refining the launch processes.

It is evident that KSC LC-39A has recorded the highest number of successful launches among all sites.

APPENDIX

- Distance from the launch site **CCAFS LC-40** to the nearest airport