IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ong Hui Ling
20/4/2024

Outline

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

Executive Summary

This data science project aimed to **analyze space launch data** to derive insights into **success factors and strategic considerations** within the aerospace industry. Leveraging techniques such as **data collection through APIs and web scraping**, as well as **data manipulation and analysis using Python libraries like Pandas and Scikit-Learn**, the project examined key questions surrounding launch success rates, payload and launch site impacts, and temporal trends of success rates.

Key Findings:

- **Positive correlation** between **flight number and first-stage landing success** observed in Low Earth Orbit (**LEO**), while **no discernible relationship** found in Geostationary Transfer Orbit (**GTO**).
- **Increasing success rates noted over time**, with fluctuations in certain years.
- **KSC LC-39A** identified as the launch site with the **highest proportion of successful launches**.
- Launch sites strategically positioned **near infrastructure** for logistical **connectivity,** while maintaining **safe distances from urban centers**.

By providing actionable insights into launch success factors and strategic site considerations, this project contributes to informed decision-making within the aerospace industry, facilitating the advancement of space exploration endeavors.

Introduction

Project background and context

In an era where space exploration is becoming increasingly accessible, SpaceX stands out as a trailblazer. SpaceX's cost-effective approach, particularly through reusing rocket components like the Falcon 9 first stage, has revolutionized the industry. Our project focuses on **analyzing SpaceX's first-stage launch details to predict its launch success rate accurately**.

Problems to find answers

1. What are the probability of successful first-stage landings for SpaceX Falcon 9 rocket launches?
2. Which factors influence the success rates of SpaceX launches?
3. How strategically are launch sites positioned, and what logistical considerations influence their locations?

Section 1

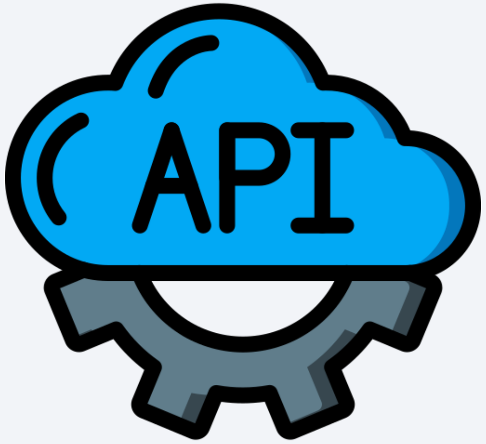
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scrapping using GET request and BeautifulSoup
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

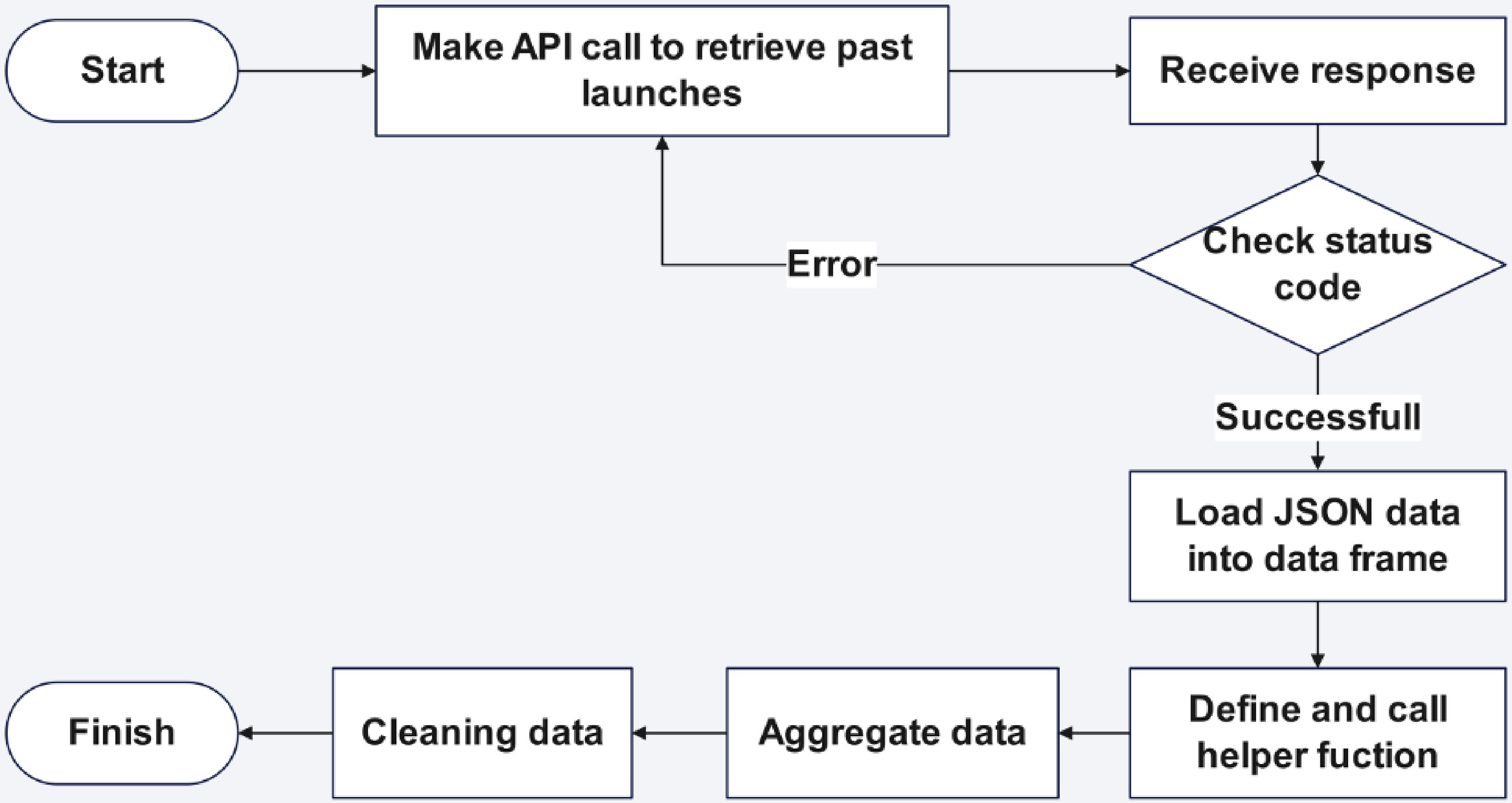


Utilize the SpaceX API to access current and official launch records, including booster versions and payload specifics, ensuring the dataset remains current and precise.

Employ web scraping techniques to efficiently extract structured data from unstructured HTML, particularly from Wikipedia.

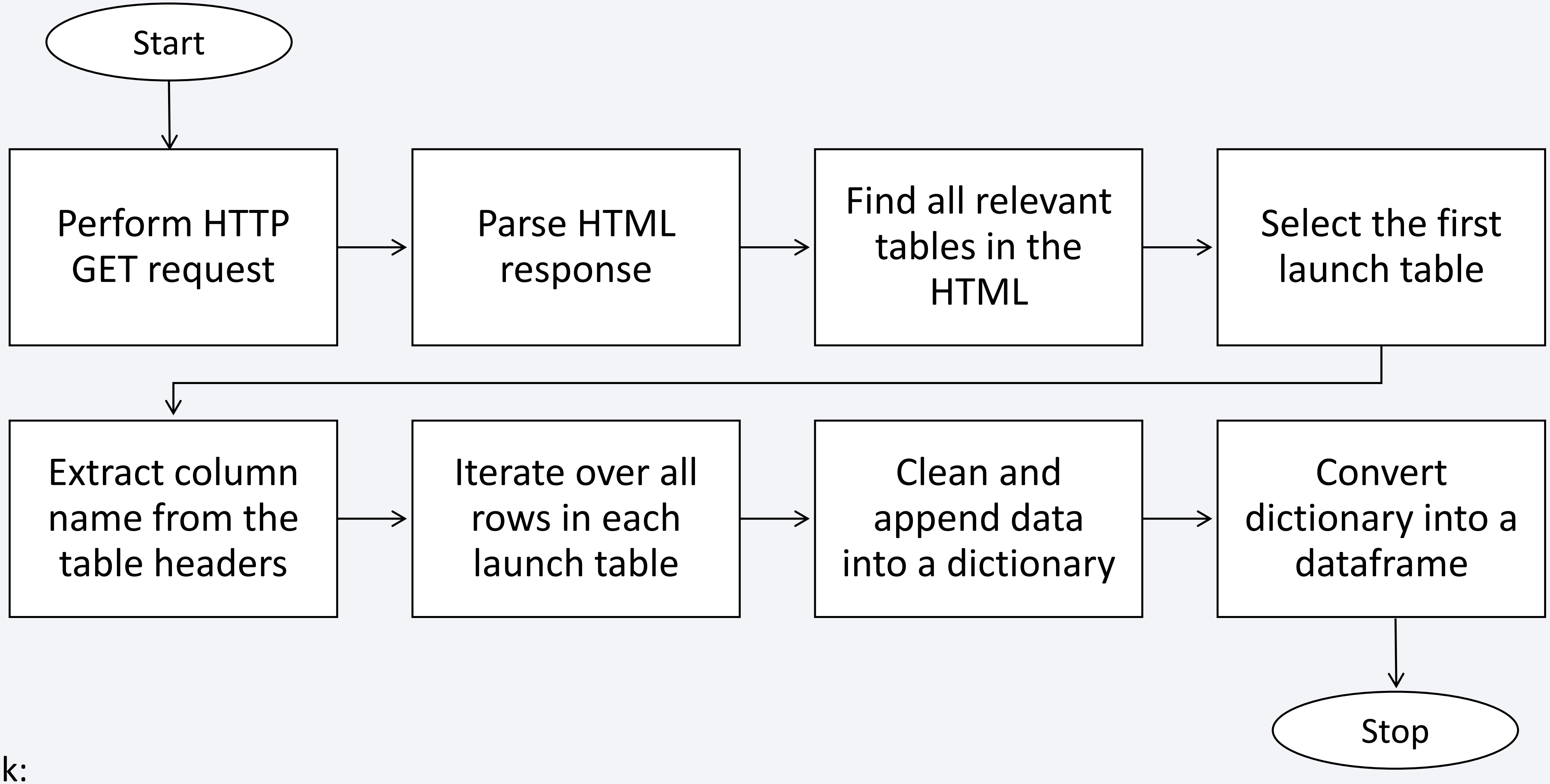
Data Collection – SpaceX API

1. Utilized the SpaceX API to dynamically enhance launch data with detailed information such as booster versions and payload details.
2. Employed Pandas to refine the dataset structure.



Github link: https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/1.1-data-collection-api.ipynb

Data Collection - Scraping



Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/1.2-webscraping.ipynb

Data Wrangling

Import Data

Data Cleaning and Preprocessing

- Handle missing data, identify column data types etc.

Exploratory Data Analysis (EDA)

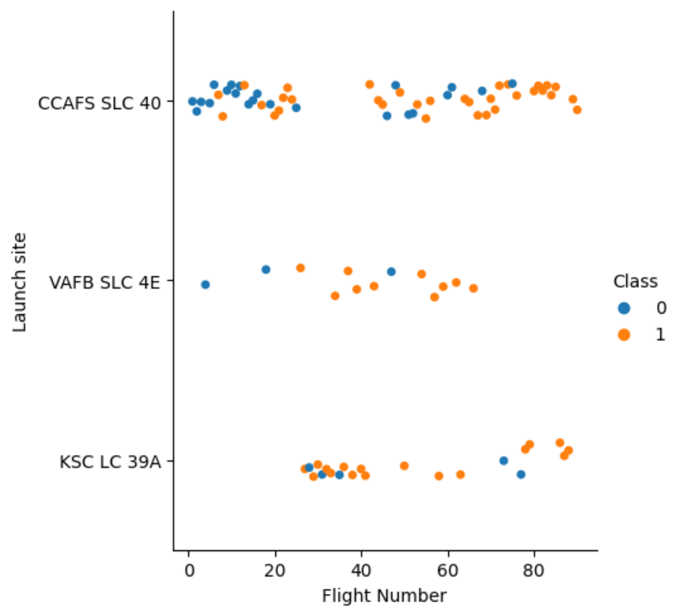
- Data visualization and SQL query

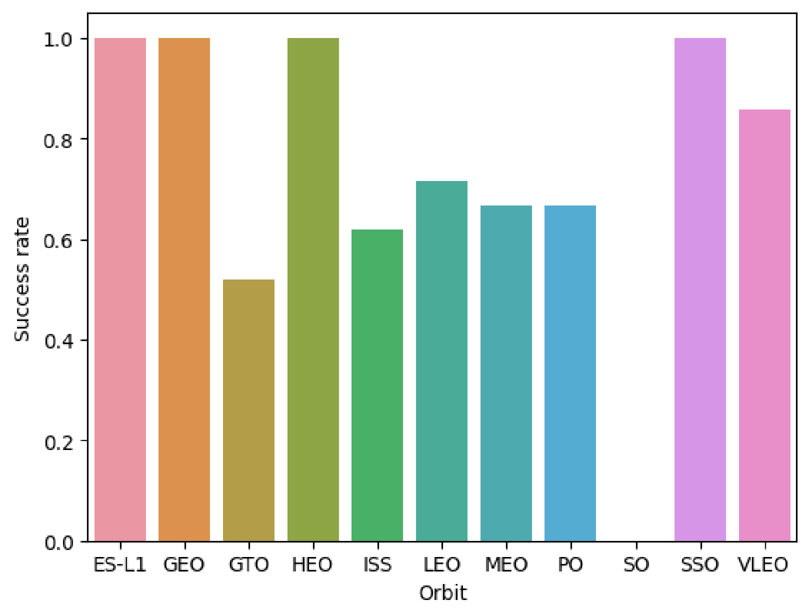
Feature Engineering

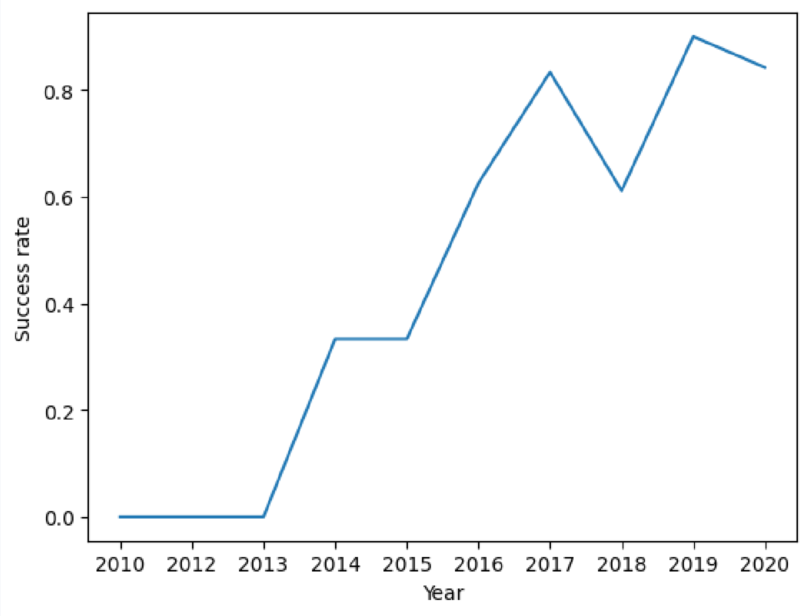
- Handle categorical data by creating dummies data

Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/1.3-Data%20wrangling.ipynb

EDA with Data Visualization







Scatter plot: Visualize the relationship between Flight Number and Launch Site on success rate

Bar plot: Visualize the relationship between success rate of each orbit type

Line plot: Visualize the launch success yearly trend

Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/2.2-eda-dataviz.ipynb

EDA with SQL

1. Identified **unique launch sites** in the dataset.
2. Filtered records to display **5** records where **launch sites begin with 'CCA'.**
3. Calculated the **total payload mass** carried by boosters launched by **NASA (CRS).**
4. Computed the **average payload mass** carried by booster version **F9 v1.1**.
5. Determined the **date of the first successful landing outcome on a ground pad**.
6. Listed **boosters** with **successful drone ship landings** and payload masses between **4000 and 6000 kg**.
7. Conducted frequency analyses of **successful and failure** mission outcomes.
8. Utilized a subquery to identify **booster versions** carrying the **maximum payload mass**.
9. Examined records to display month names, failure landing outcomes in drone ships, booster versions, and launch sites for the year **2015**.
10. Ranked the count of **landing outcomes** (e.g., failure on drone ship or success on ground pad) within the date range from **2010-06-04 to 2017-03-20.**

Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/2.1-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

1. Established **NASA Johnson Space Center** as the initial reference point.
2. Implemented **marker clustering** to condense site markers and enhance map clarity.
3. Employed **color-coded indicators** to swiftly differentiate launch outcomes.
4. Incorporated **proximity indicators** to highlight distances between launch sites and coastlines.
5. Illustrated **geographical context lines** connecting launch sites with nearby strategic locations.

Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/3.1_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

1. **Dropdown for launch site selection:** To offer users the flexibility to choose a specific site or view data for all sites.

2. **Pie chart for total successful launches:** To visually represent the distribution of total successful launches, providing quick insights for all sites or a selected site.

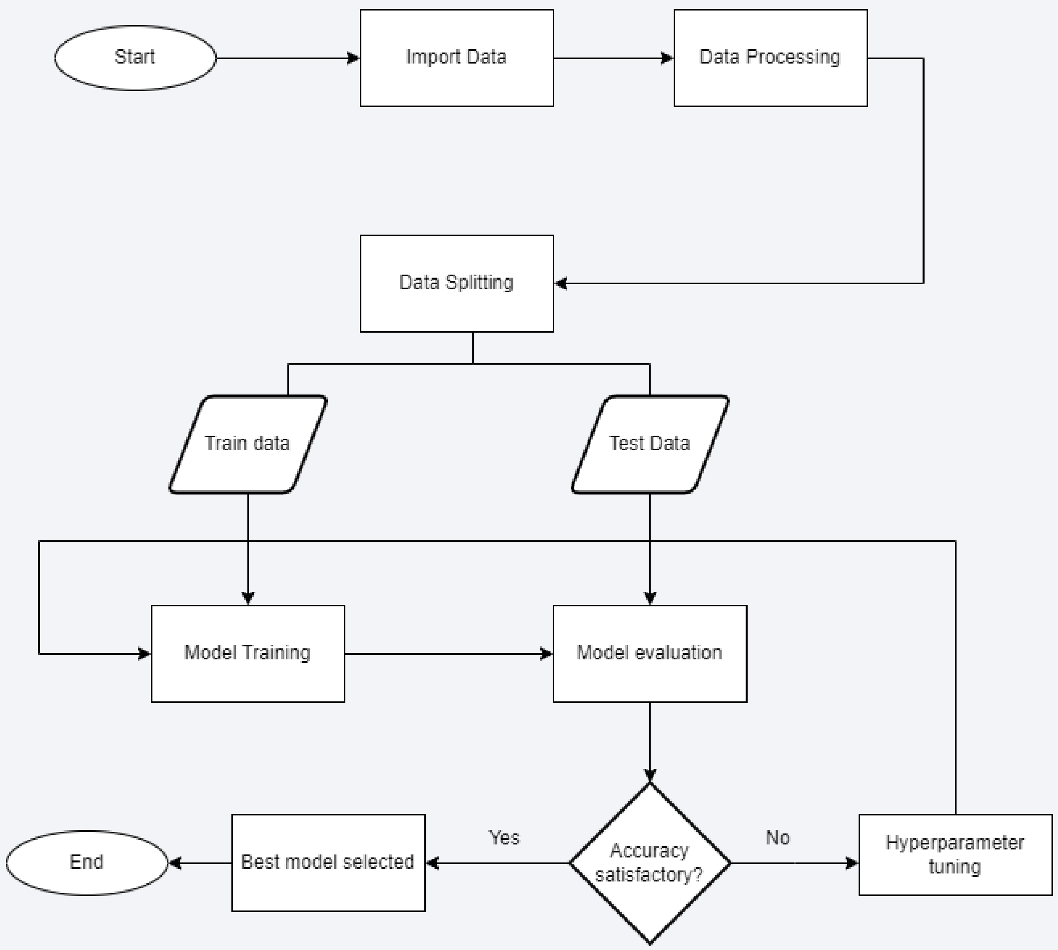
3. **Payload range slider:** To enable users to filter launches based on payload mass, and focus on specific mass ranges of interest.

4. **Scatter chart for payload vs. Launch success:** To visualize the relationship between payload mass and launch success, with color-coded differentiation by booster version category for enhanced insight.

Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/3.2-spacex_dash_app.py

Predictive Analysis (Classification)

1. **Model Creation**
   1. Logistic Regression
   2. SVM
   3. Decision Tree
   4. KNN
2. **Hyperparameter tuning**
   - Utilized GridSearchCV for model settings optimization.
3. **Model Evaluation**
   - Assessed model using test data, generating confusion matrices and accuracy scores.
4. **Best Model**
   - KNN and SVM (high accuracy score and high F1-score)



Github link:
https://github.com/HuiLing0511/AppliedDataScienceCapstone/blob/7202ee91a7a0adaa6517f5aa09d690e65b050a08/4.1-Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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

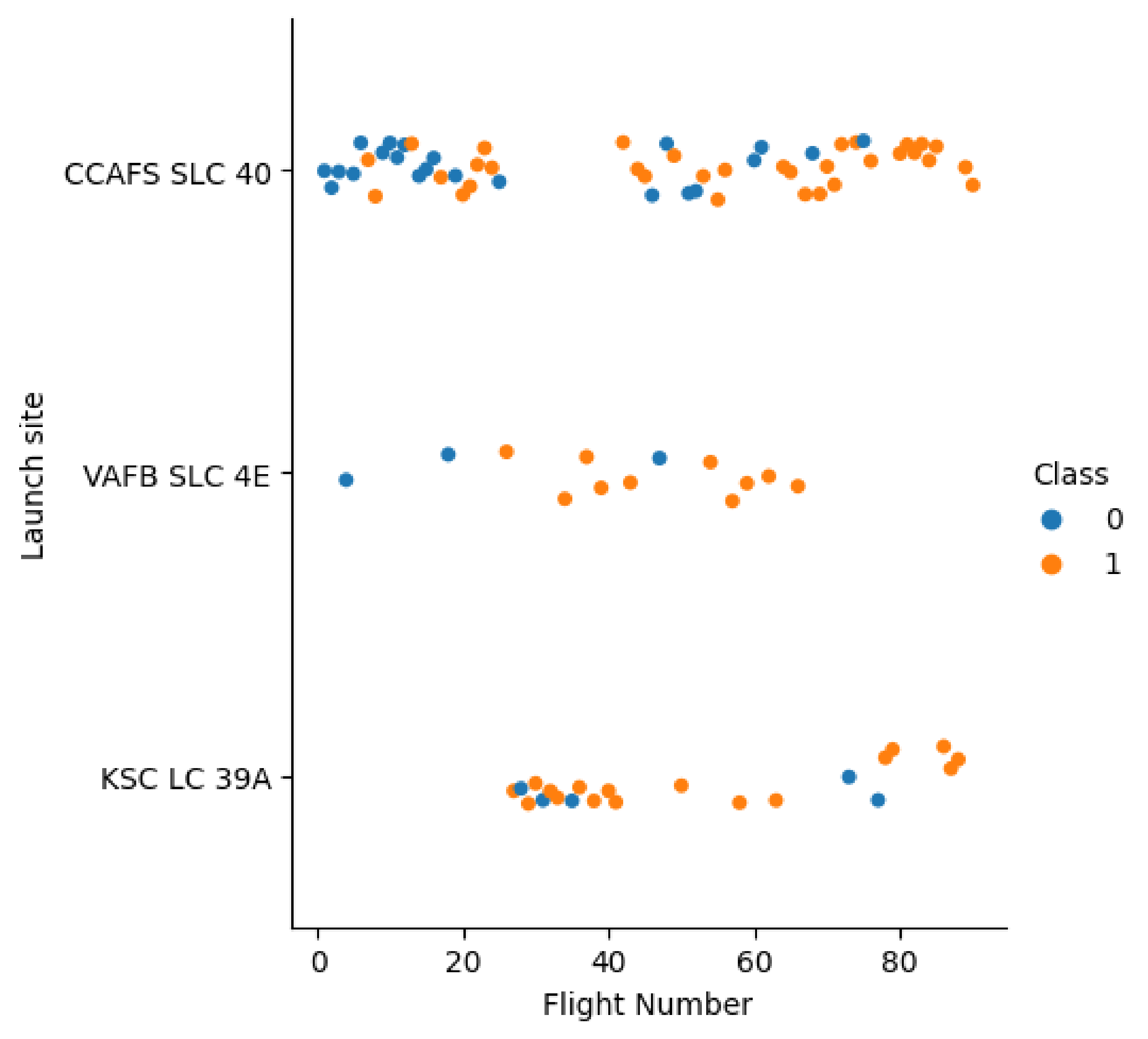
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Findings:

As the **flight number increases, the first stage is more likely to land successfully** across all three launch sites.



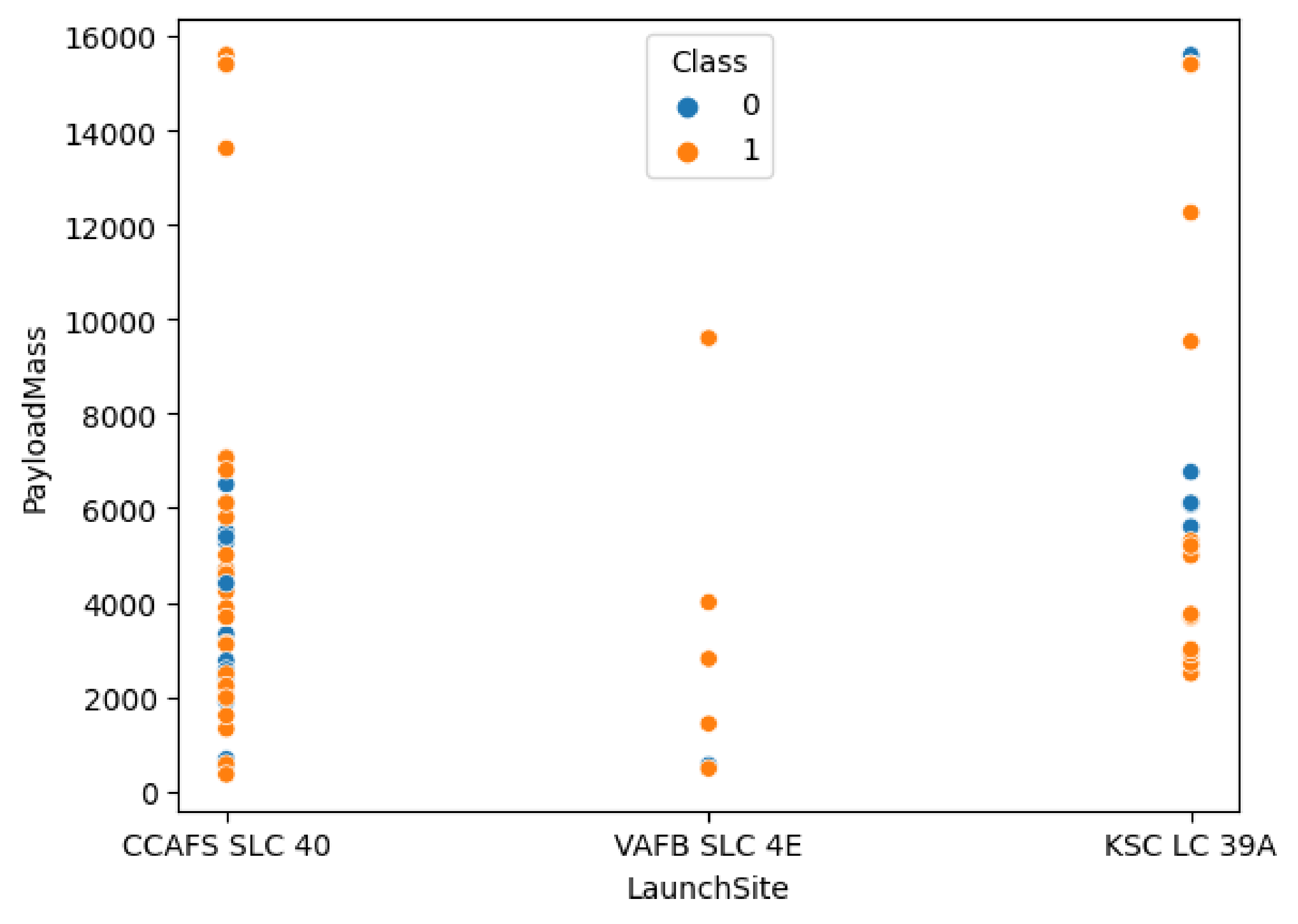
Payload vs. Launch Site

Key Findings:

As the **payload mass increases, the first stage is more likely to land successfully**.

Detailed Findings:

1. CCAFS SLC 40 launch site shows 100% success rate with payload mass >12000 kg.
2. VAFB SLC 4E launch site does not have records with payload mass above 10000 kg.
3. KSC LC 39A shows weak relation between payload mass and success rate.



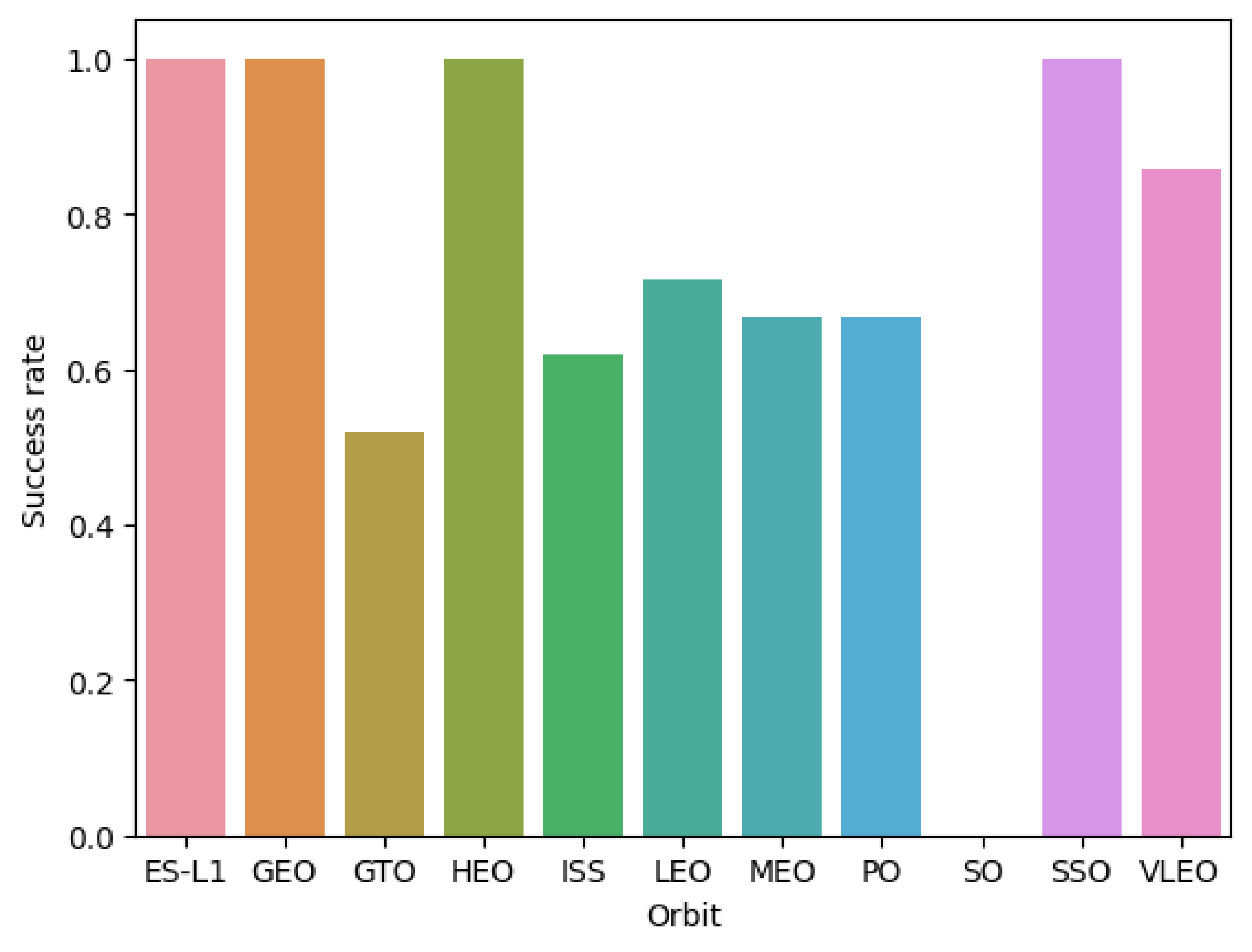
Success Rate vs. Orbit Type

Findings:

1. **ES-L1, GEO, HEO and SSO** have the **highest success rate** at **1**.

2. **VLEO** also show good success rate at about **0.9**.

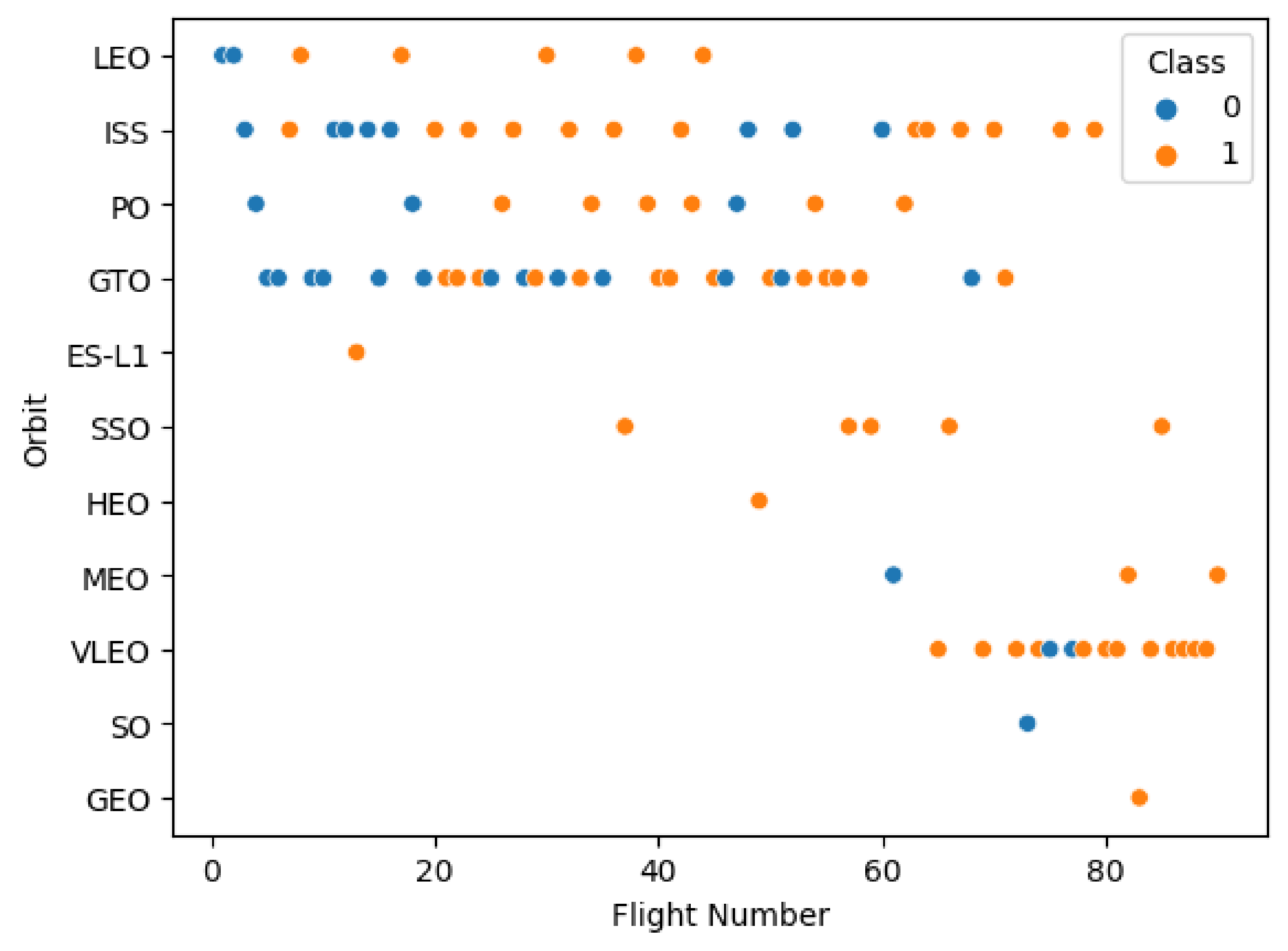
3. **GTO** have the **lowest** success rate at about **0.55**.



Flight Number vs. Orbit Type

Findings:

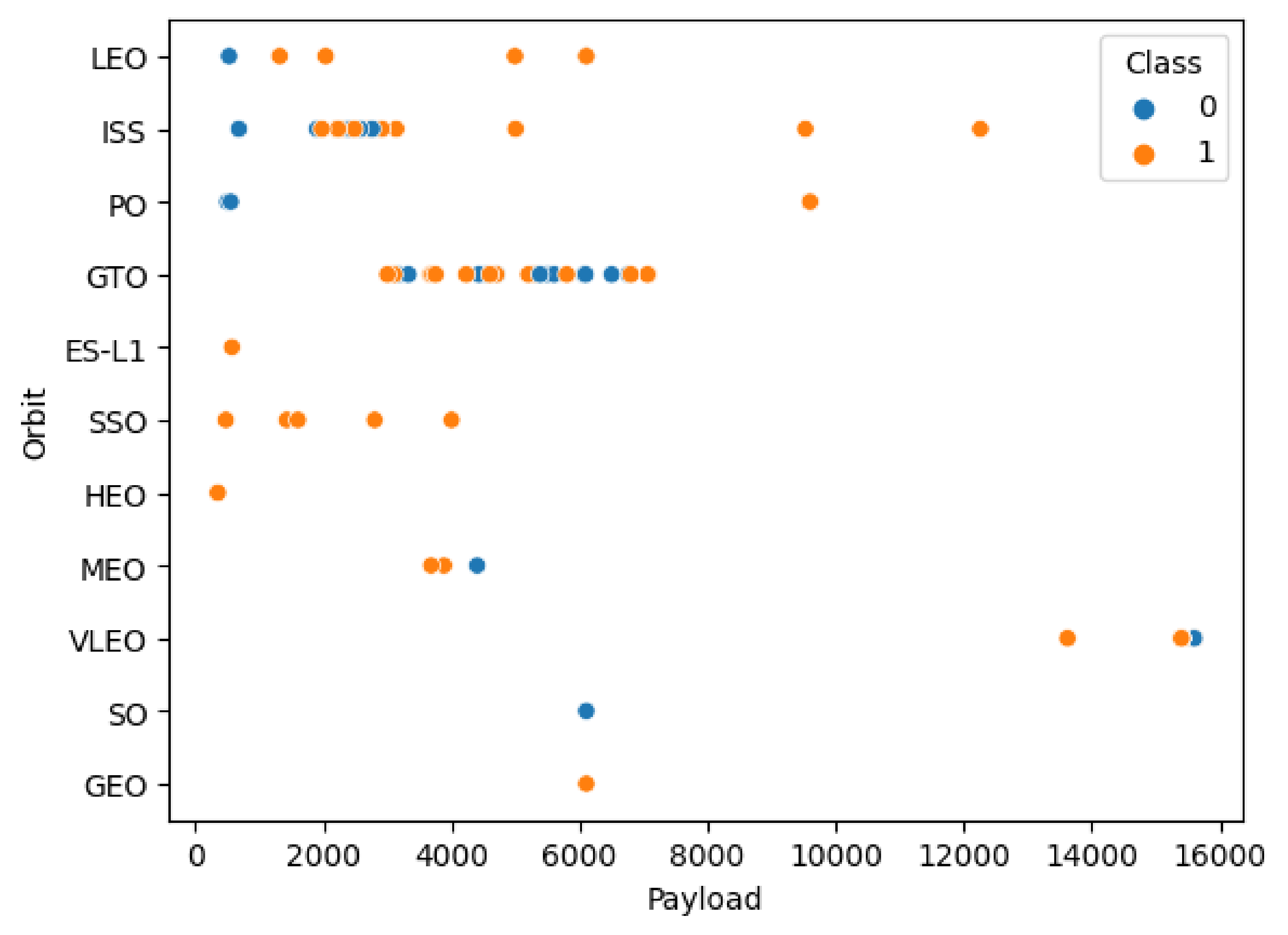
1. **LEO orbit:** Success rates show a **positive correlation** with the number of flights.
2. **GTO orbit**: **No apparent relationship** between flight number and success rates.



Payload vs. Orbit Type

Findings:

1. As **payload mass increases**, **Polar, LEO,** and **ISS orbits** show a **higher success rate**.
2. However, in **GTO**, payload mass has **no effect** on success rate, as both successful and unsuccessful landings occur.



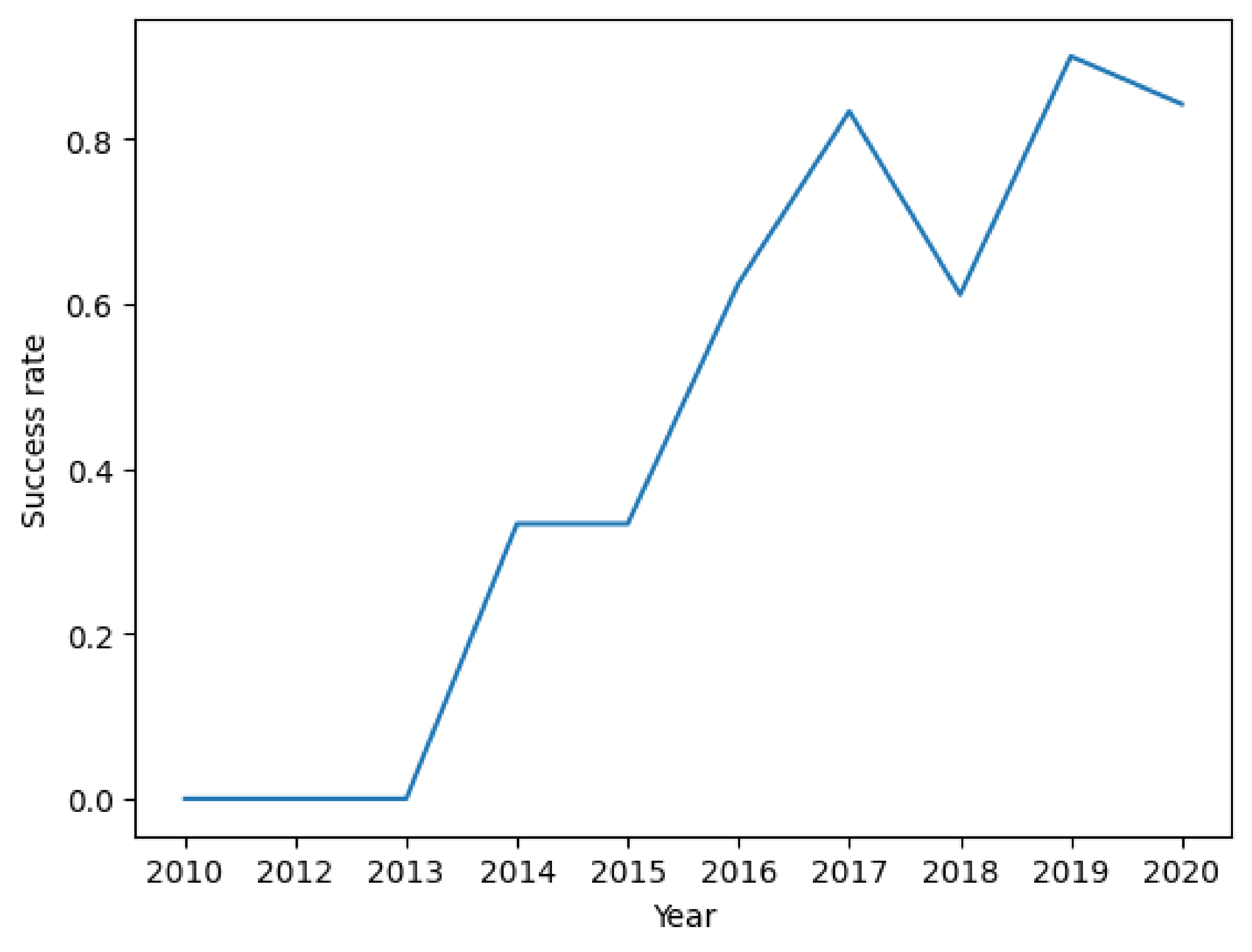
Launch Success Yearly Trend

Findings:

1. The success rate has shown a **consistent increase** since 2013, with **constant** observed in **2014**.

2. Notably, **from 2015 to 2017**, there was a **significant uptrend** in success rates.

3. In **2018**, there was a **slight decrease** in success rate, followed by a **subsequent increase** in **2019**.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%sql select distinct Launch_Site from SPACEXTBL
```

[9]

* sqlite:///my_data1.db
Done.

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

All launch sites names:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCCAFS SLC-40

Launch Site Names Begin with 'CCA'

Launch sites begin with `CCA`: CCAFS LC-40

Task 2

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

```
[13] %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

Python

```
* 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

Total payload carried by boosters from NASA: 45596 kg

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[14]
```
%sql select sum (PAYLOAD_MASS__KG_) from SPACEXTBL where Customer == 'NASA (CRS)'
```

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

sum (PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1: 2928.4 kg

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[16] %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version == 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Date of the first successful landing outcome on ground pad: 2015-12-22

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[19]

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome == 'Success (ground pad)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

Task 6

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

[21]
```
%sql select Booster_Version from SPACEXTBL 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

1. Total number of successful mission outcomes: 98
2. Total number of failure mission outcomes: 0

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT SUM(CASE WHEN Mission_Outcome = 'Success' THEN 1 ELSE 0 END) AS Successful_Outcomes, SUM(CASE WHEN Mission_Outcome = 'Failure' THEN 1 ELSE 0 END) AS Failed_Outcomes F
```

[27] Python

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

Successful_Outcomes	Failed_Outcomes
98	0

Boosters Carried Maximum Payload

Booster that have carried the maximum payload mass

1. F9 B5 B1048.4
2. F9 B5 B1049.4
3. F9 B5 B1051.3
4. F9 B5 B1056.4
5. F9 B5 B1048.5
6. F9 B5 B1051.4
7. F9 B5 B1049.5
8. F9 B5 B1060.2
9. F9 B5 B1058.3
10. F9 B5 B1051.6
11. F9 B5 B1060.3
12. F9 B5 B1049.7

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[29]

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * 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

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%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
```

[30] Python

```
 * 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

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

Task 10

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.

[31]

```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTBL 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

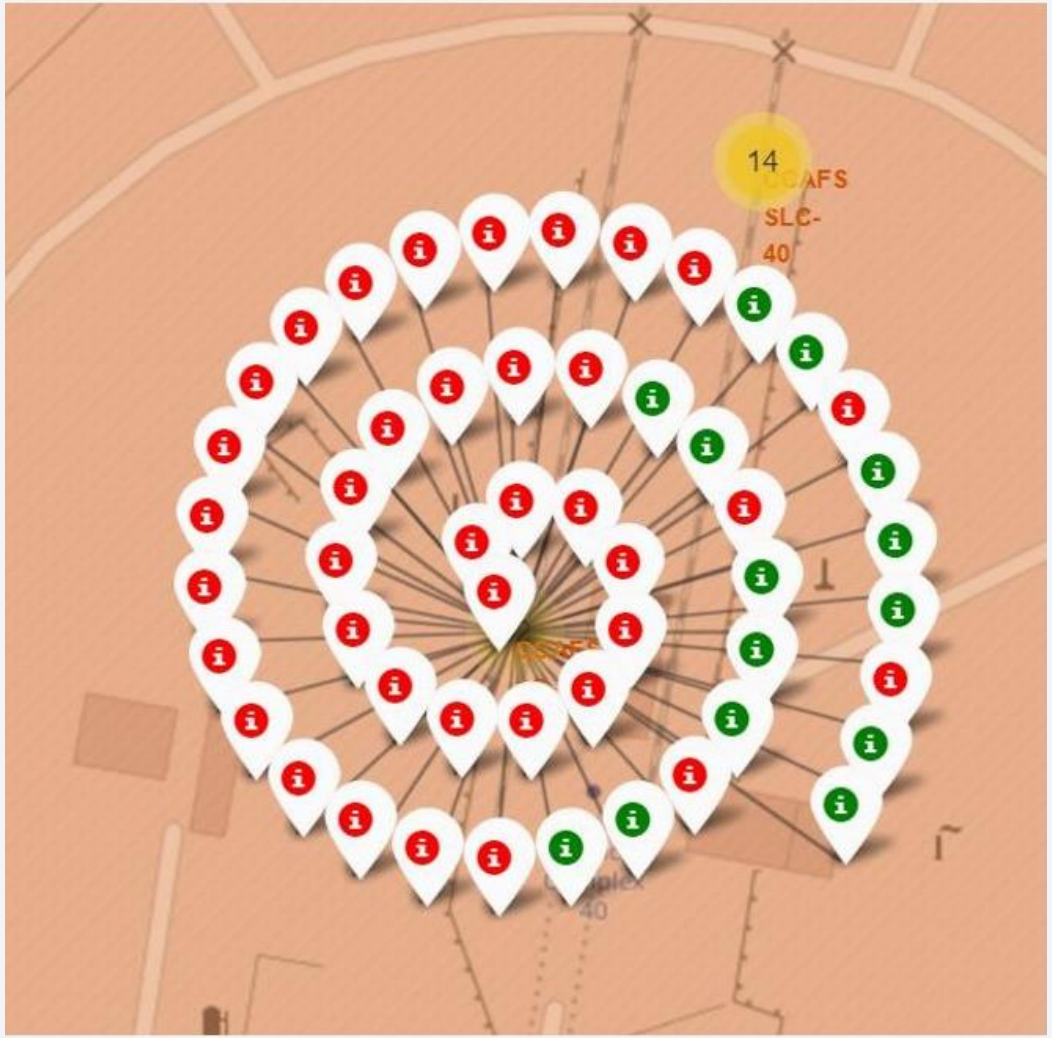
Section 3

Launch Sites Proximities Analysis

Regional Overview of Launch Sites

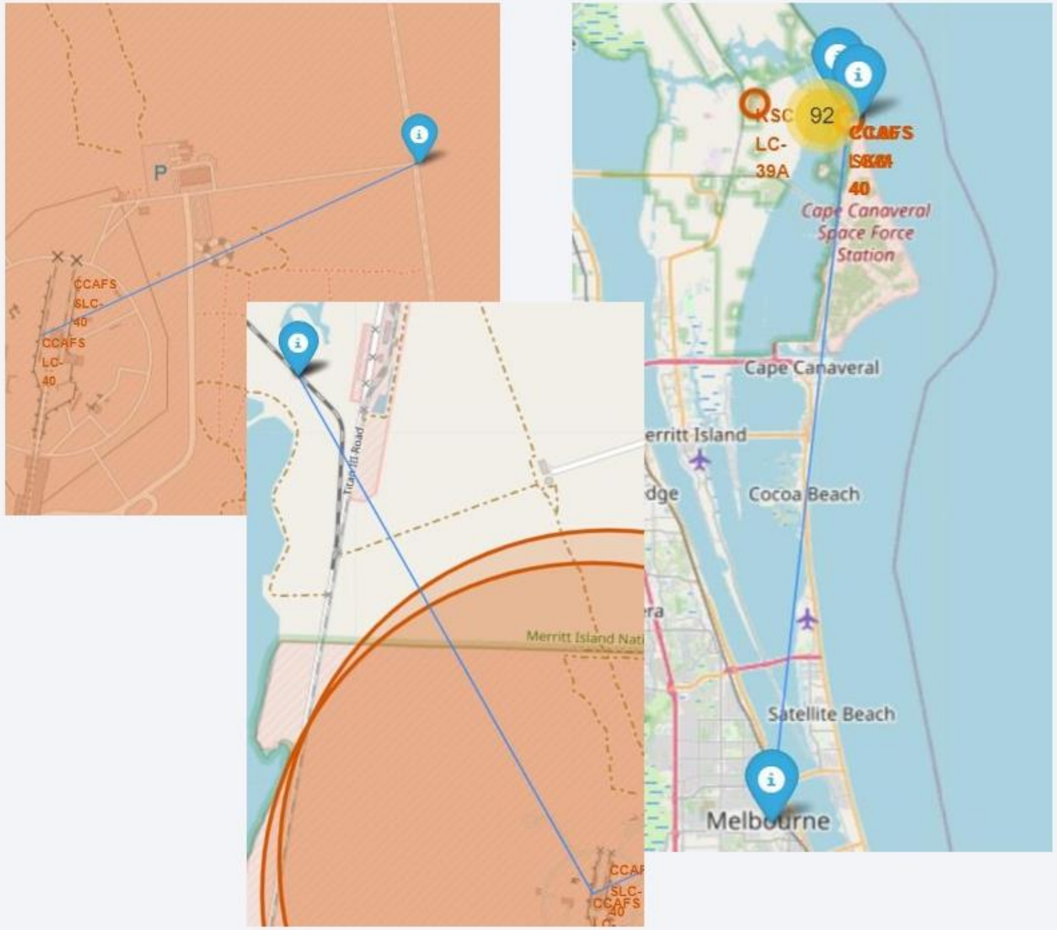
The location markers on the global map depict the **distribution of launch sites worldwide**, emphasizing the geopolitical significance of space launches.

Launch Outcomes at a Centralized Launch Site



The color-coded launch outcomes on the map **provide insight into the historical success or failure rates at various launch sites**, offering valuable information on the reliability and performance of different locations for space launches.

Proximity Analysis of a launch site

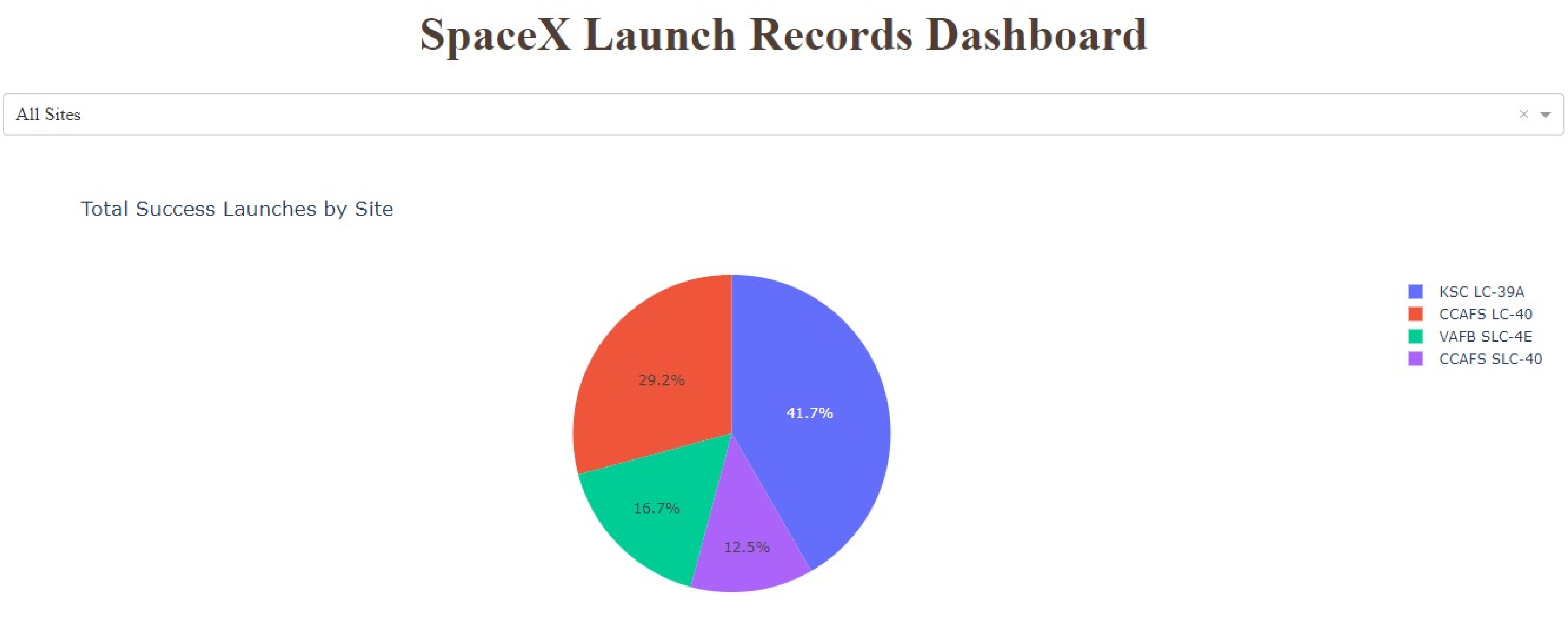


1. Launch sites are strategically located within 2 km of railways and less than 1 km from highways, ensuring vital **connectivity** for logistical operations.
2. Additionally, situated over 51 km from urban centers, these sites maintain a **safe distance from populated areas**.
3. Furthermore, their proximity to coastlines facilitates **safer launch trajectories and effective debris management**.

Section 4

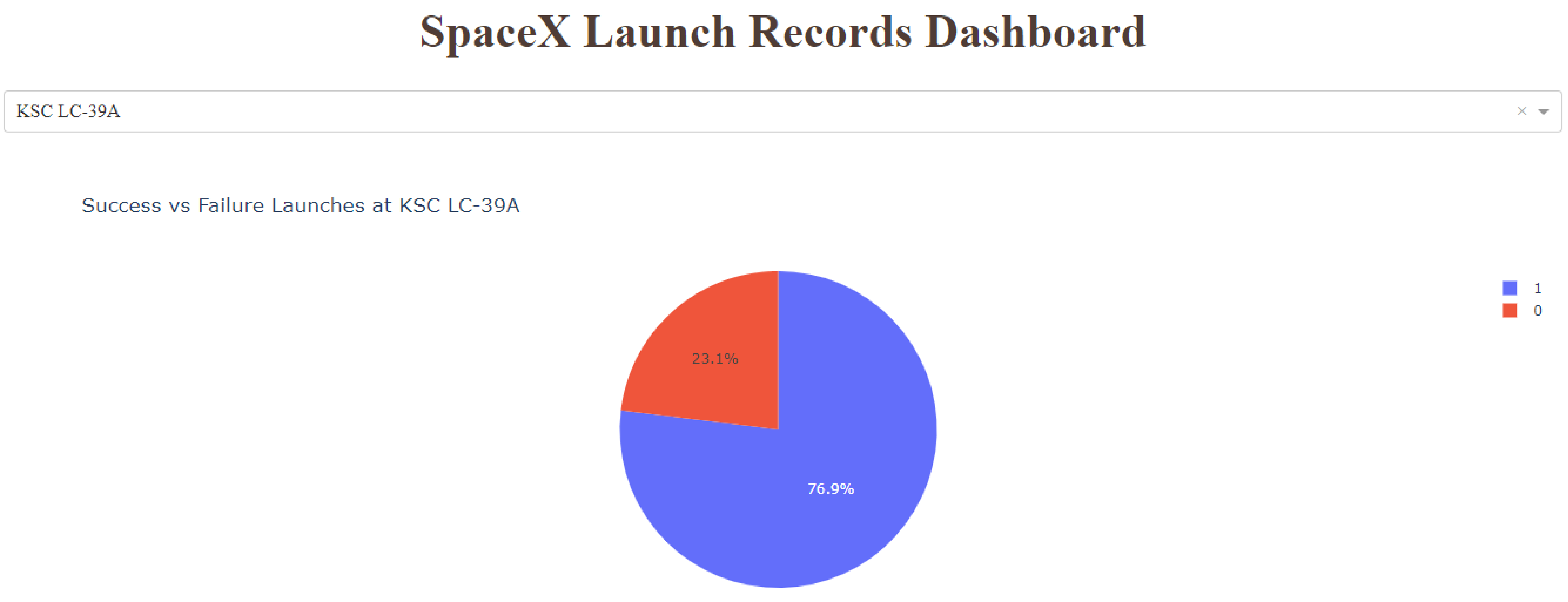
Build a Dashboard with Plotly Dash

Total Success Launches by site



1. **KSC LC-39A** contributes the **highest proportion to total successful launches**, accounting for **41.7%** of successes, followed by CCAFS LC-40 and VAFB SLC-4E.
2. Conversely, **CCAFS SLC-40 contributes the lowest proportion** to successful launches.

Highest launch success ratio (KSC LC-39A)



Highest launch site: KSC LC-39A

- 76.9% success, 23.1% failure.

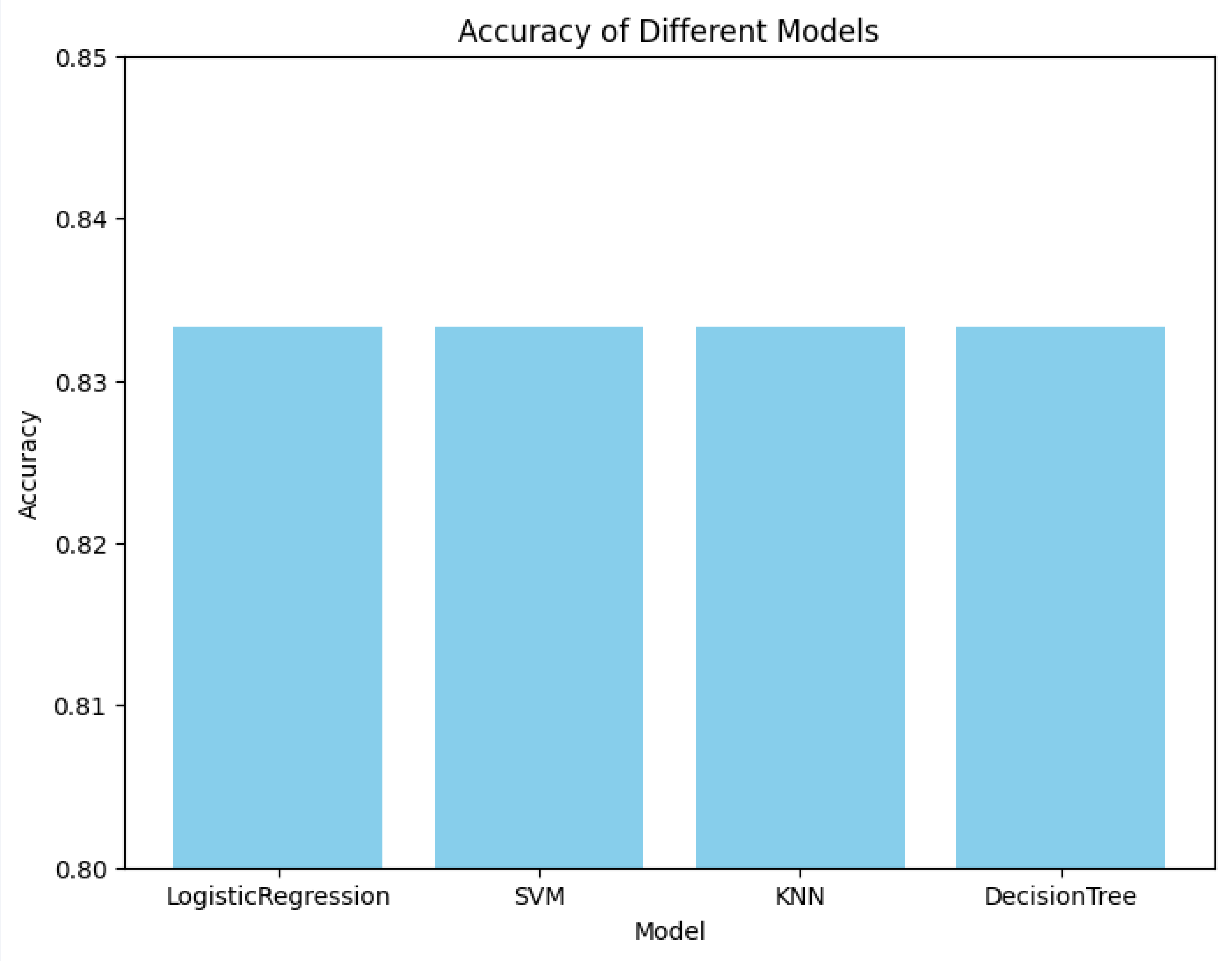
Payload vs. Launch Outcome scatter plot for all sites

Section 5

Predictive Analysis (Classification)

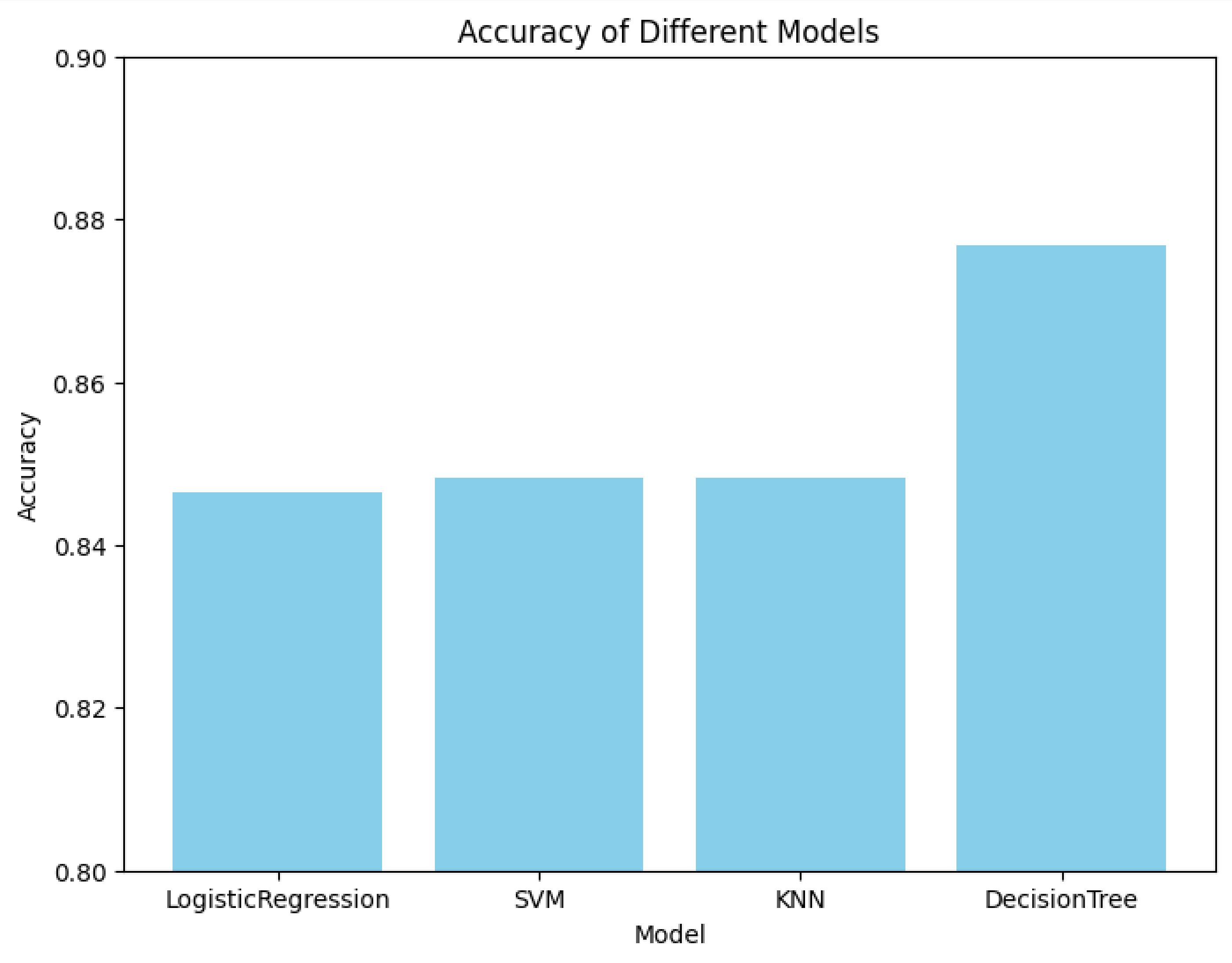
Classification Accuracy

All four models achieve the **same accuracy score** of **0.833** when evaluated using the score (X_test, Y_test) method.



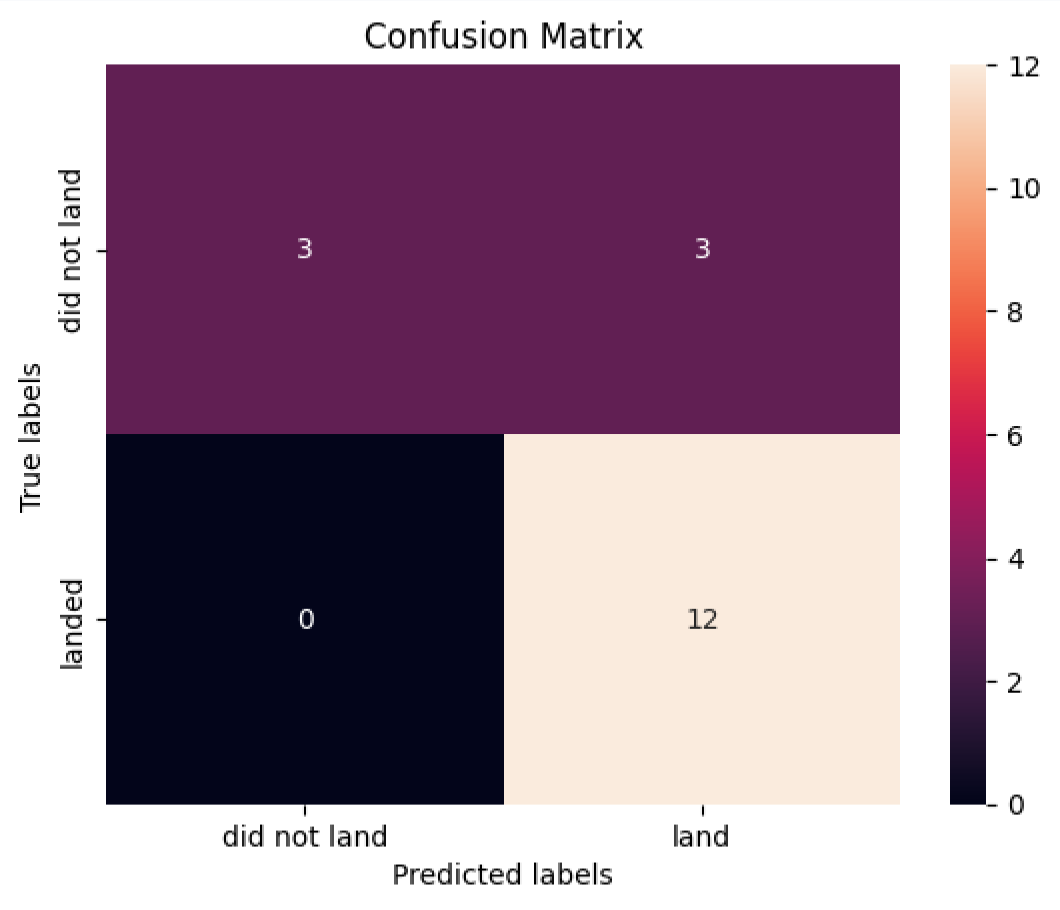
Classification Accuracy

If we utilize the model.best_score_ method to evaluate the models, the **Decision Tree** emerges as the best model, achieving an accuracy of around 0.88.



Confusion Matrix

- All four models have **same confusion matrix.**
- Explanation:
  - The model performs very good in predicting the TRUE label (land) with all TRUE label predicted correctly.
  - While in predicting the FALSE label (do not land), 3 out of 6 was predicted wrongly as landed.



Conclusions

This project successfully achieved the objective of analyzing space launch data to derive insights and answer key questions.

Key Findings:

1. **Positive correlation** observed between **flight number** and first-stage **landing success** in **LEO** orbit.
2. **No discernible relationship** between **flight number** and first-stage **landing success** in **GTO** orbit.
3. **Increasing success rates** observed **over time**, with slight fluctuations in certain years.
4. **KSC LC-39A** identified as the launch site with the **highest proportion of successful launches**.
5. Launch sites strategically positioned **near railways and highways** for logistical **connectivity**, while maintaining **safe distances from urban centers**.
6. **Decision tree** emerges as the **best model** in predicting the launch success rate.

Conclusion:

Through comprehensive analysis, we have gained valuable insights into the dynamics of space launches, contributing to a better understanding of success factors and strategic considerations in the aerospace industry.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!