IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Salina K Hall
October 29th, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis
    - Exploratory Data Analysis Using SQL
    - Exploratory Data Analysis with Data Visualization using Pandas and Matplotlib
  - Interactive Visual Analytics and Dashboards
    - Interactive Visual Analytics with Folium
    - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context
  - SpaceX, a leader in the space industry, aims to make space travel more affordable and accessible. Its major accomplishments include:
  - Delivering spacecraft to the **International Space Station (ISS)**
  - Launching **Starlink**, a global satellite internet constellation
  - Conducting **manned space missions**
  - A key reason for SpaceX's success is its **cost-efficient Falcon 9 launches**, priced at approximately **$62 million** each. Other providers charge over **$165 million** per launch, largely because they cannot reuse the rocket's **first stage**.
  - This project focuses on predicting whether the first stage will land successfully, which directly affects **launch costs**. Using **publicly available data** and **machine learning models**, we aim to create a pipeline to predict first-stage landing outcomes, providing insights into launch economics and potential competitive strategies for other companies in the commercial launch market.
- Problems you want to find answers
  - **Key Question:** What factors determine first-stage landing success?
  - **Critical Variables:** Payload mass, launch site, number of prior flights, orbital parameters
  - **Landing Trends:** Success rates over time
  - **Predictive Modeling:** Develop the best machine learning model to classify landing outcomes

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API was used for SpaceX launch data
  - Web scraping using BeautifulSoup was used to collect Falcon 9 historical launch records from a Wikipedia page
- Perform data wrangling
  - The collected data was cleaned and wrangled by filtering records, handling missing values, and applying one-hot encoding to prepare it for analysis and modeling. The dataset was further enriched by creating a landing outcome label based on the outcome data, after summarizing and analyzing key features.

Methodology

Executive Summary

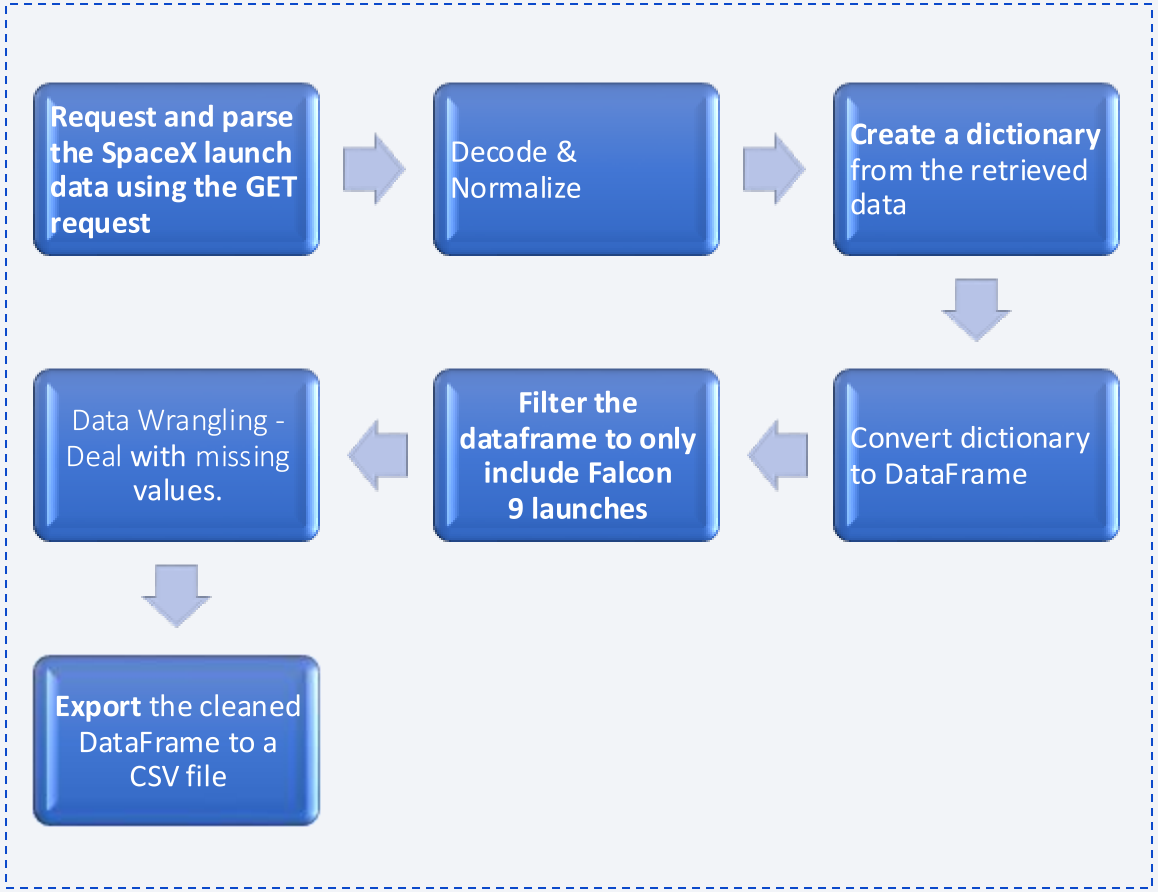
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Collected and prepared data was normalized and split into training and test sets. Multiple classification models were built to predict landing outcomes, and their performance was evaluated across different parameter combinations to identify the best model and optimal settings.

Data Collection

- Describe how data sets were collected.
    - SpaceX launch data was gathered from the SpaceX REST API
        - https://api.spacexdata.com/v4/.
    - Web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches using BeautifulSoup
        - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Data for this project was obtained from the **public SpaceX API**, which provides detailed information on rocket launches. The process involved the following steps:
    - **Request launch data** from the SpaceX API.
    - **Decode API response** using .json() and convert to a **DataFrame** with .json_normalize().
    - **Request detailed launch information** using custom functions.
    - **Create a dictionary** from the retrieved data.
    - **Convert dictionary to DataFrame**.
    - **Filter** to include only **Falcon 9 launches**.
    - **Handle missing values** in payload mass by replacing them with the **mean**.
    - **Export** the cleaned DataFrame to a CSV file.
- **Key Points:**
    - SpaceX's public API provides structured and accessible data.
    - Data is persisted locally for analysis and modeling.
    - Source code:
        - GitHub Notebook



Data Collection - Scraping

In addition to the API, SpaceX launch data was collected from **Wikipedia** using web scraping. The process involved the following steps:

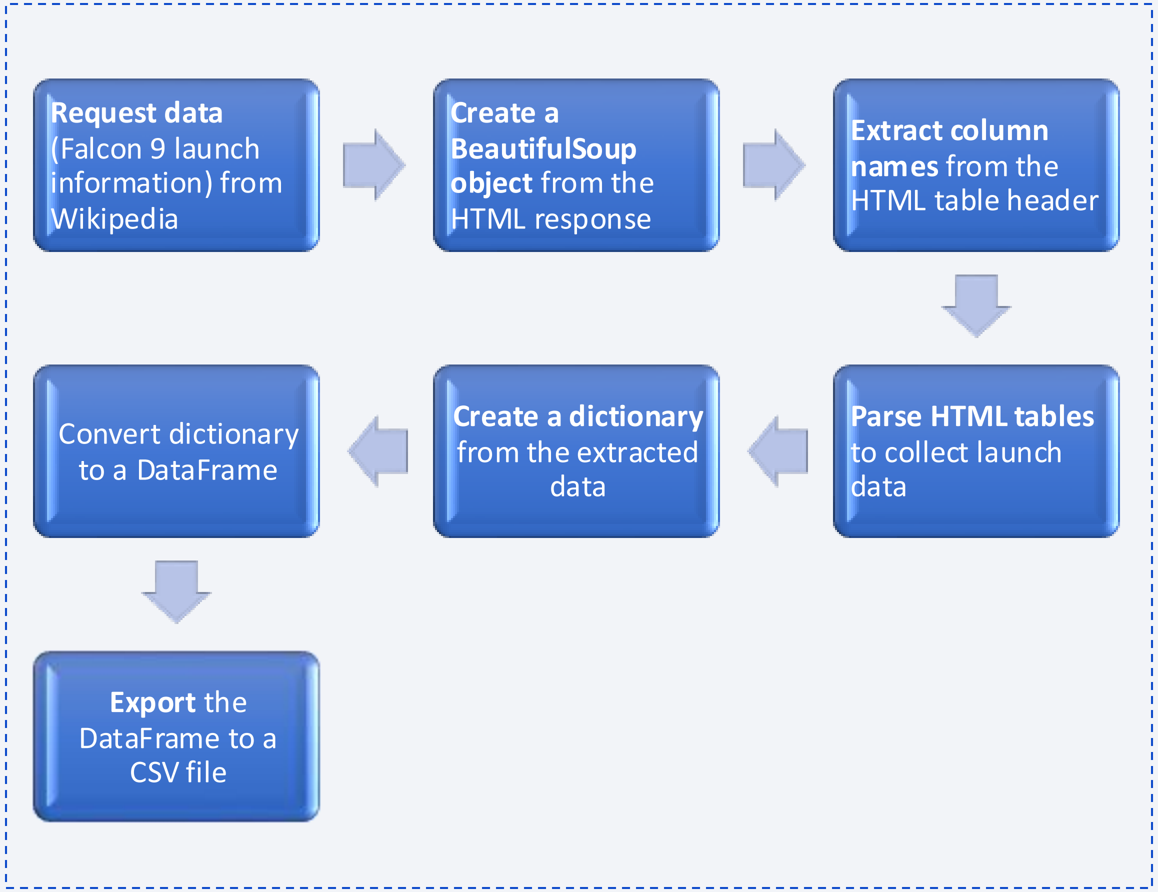
- **Request data** (Falcon 9 launch information) from Wikipedia.
- **Create a BeautifulSoup object** from the HTML response.
- **Extract column names** from the HTML table header.
- **Parse HTML tables** to collect launch data.
- **Create a dictionary** from the extracted data.
- **Convert dictionary to a DataFrame**.
- **Export** the DataFrame to a CSV file.

Key Points:

- Wikipedia provides publicly accessible data on SpaceX launches.
- Data is persisted locally for further analysis and modeling.

Source code:

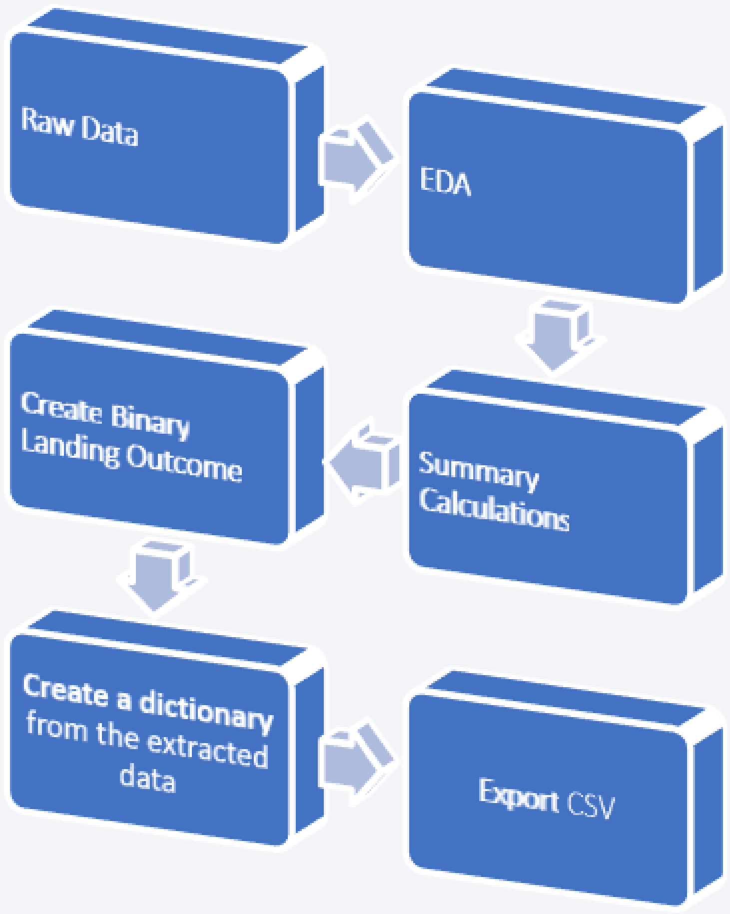
- [GitHub Notebook]



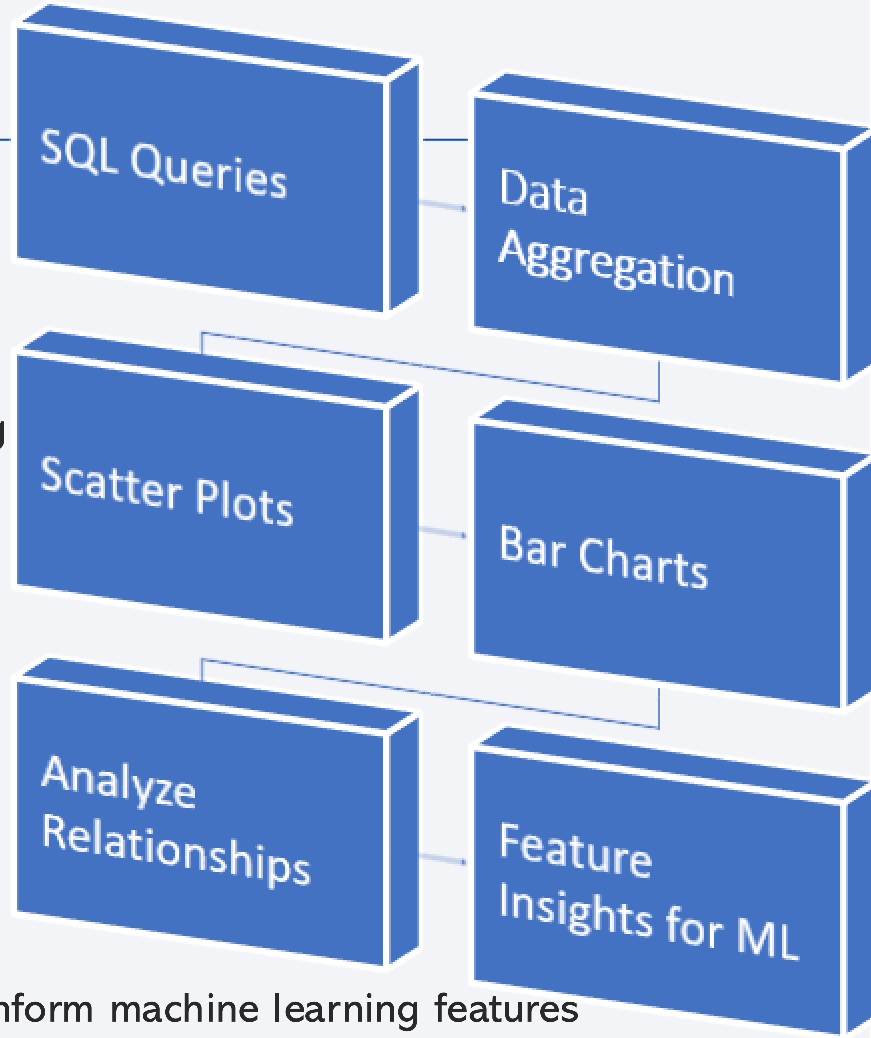
Data Wrangling

The collected datasets were cleaned, summarized, and prepared for modeling through the following steps:

- **Exploratory Data Analysis (EDA):**
  - Investigated dataset structure and identified relevant features.
  - Determined key data labels.
- **Summary Calculations:**
  - Count of **launches per site**
  - Count and occurrence of **orbit types**
  - Count and occurrence of **mission outcomes per orbit type**
- **Landing Outcome Labeling (Dependent Variable):**
  - Created a **binary landing outcome column** for modeling:
    - **True Ocean:** Successful landing in ocean
    - **False Ocean:** Unsuccessful landing in ocean
    - **True RTLS:** Successful landing on ground pad
    - **False RTLS:** Unsuccessful landing on ground pad
    - **True ASDS:** Successful landing on drone ship
    - **False ASDS:** Unsuccessful landing on drone ship
  - Converted outcomes into **1 = successful**, **0 = unsuccessful**
- **Data Export:**
  - Exported the processed dataset to **CSV** for modeling.
- **Source Code:**
  - GitHub Notebook



Exploratory Data Analysis with SQL and Data Visualization



SQL Queries Performed:
- Unique launch site names and top sites starting with 'CCA'
- Total and average **payload mass** for specific boosters
- First successful **ground pad landing** date
- Boosters with successful **drone ship landings** and payload between 4000–6000 kg
- Total number of **successful vs. failed missions**
- Booster versions carrying **maximum payload**
- **Failed drone ship landings** in 2015, with booster versions and launch sites
- **Ranked landing outcomes** between 2010–06–04 and 2017–03–20

Visualizations Plotted:
- **Flight Number vs. Payload** – scatter plot to observe trends over time
- **Flight Number vs. Launch Site** – scatter plot for launch site distribution
- **Payload Mass vs. Launch Site** – scatter and bar charts to compare categories
- **Payload Mass vs. Orbit Type** – to explore payload-orbit relationships

Analysis Purpose:
- Scatter plots help identify **relationships between numerical variables**, which may inform machine learning features
- Bar charts provide **comparisons among categorical variables** and highlight trends in launch outcomes
- **Line Chart** to visualize the launch success yearly trend.

Source Code:
GitHub Notebook - EDA with Visualization

EDA with SQL

To explore and better understand the SpaceX launch data, both **SQL queries** and **visualizations** were used.

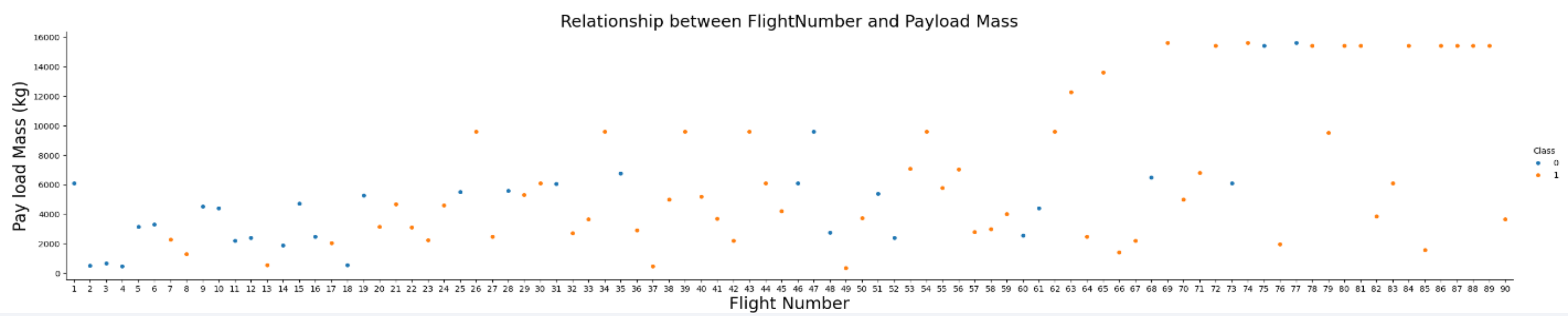
SQL Queries Performed:

- Retrieved names of **unique launch sites** and **top 5 sites** beginning with “CCA”
- Calculated and displayed **total payload mass** carried by boosters launched by NASA (CRS)
- Calculated and displayed the **average payload mass** carried by booster version F9 v1.1.
- Identified the **first successful ground pad landing date**
- Listed **boosters with successful drone ship landings** and **payloads between 4,000–6,000 kg**
- Listed **booster versions** which have carried the max **payload mass**
- Counted **successful vs. failed mission outcomes** and **ranked landing outcomes** (2010–2017)
- Retrieved **failed drone ship landings** in 2015, including **booster versions** and **launch sites**
- 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

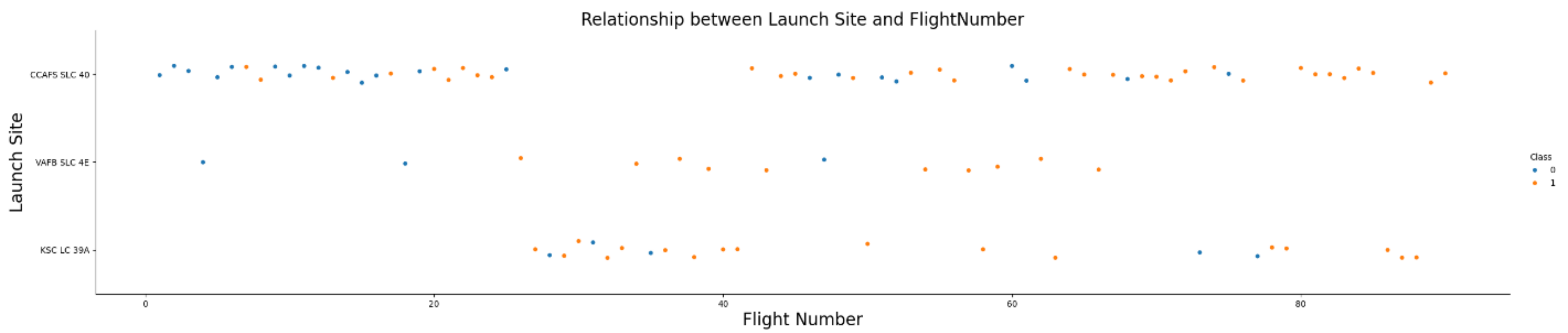
Source Code:
GitHub Notebook – EDA with SQL

EDA with Data Visualization

- **Scatter Plots** to explore relationships between features:
  - Payload Mass × Flight Number

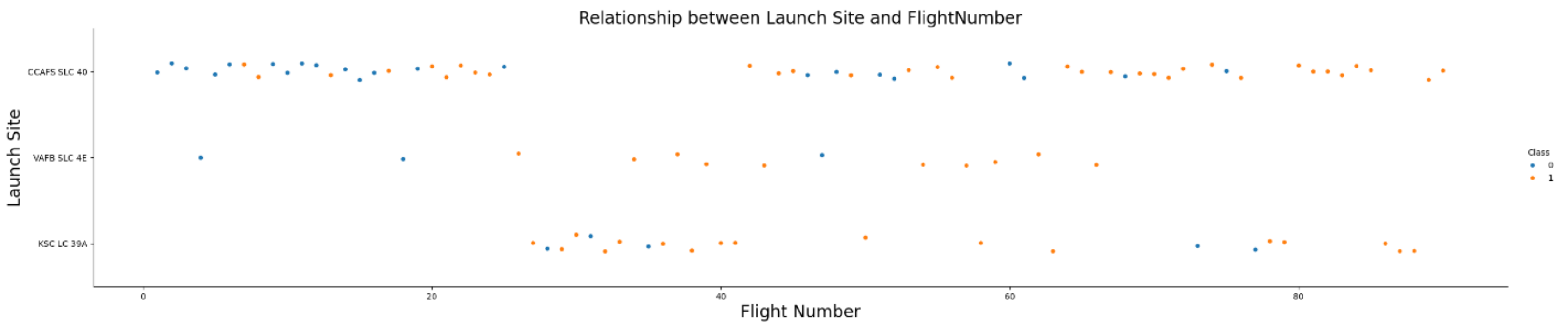


  - Launch Site × Flight Number

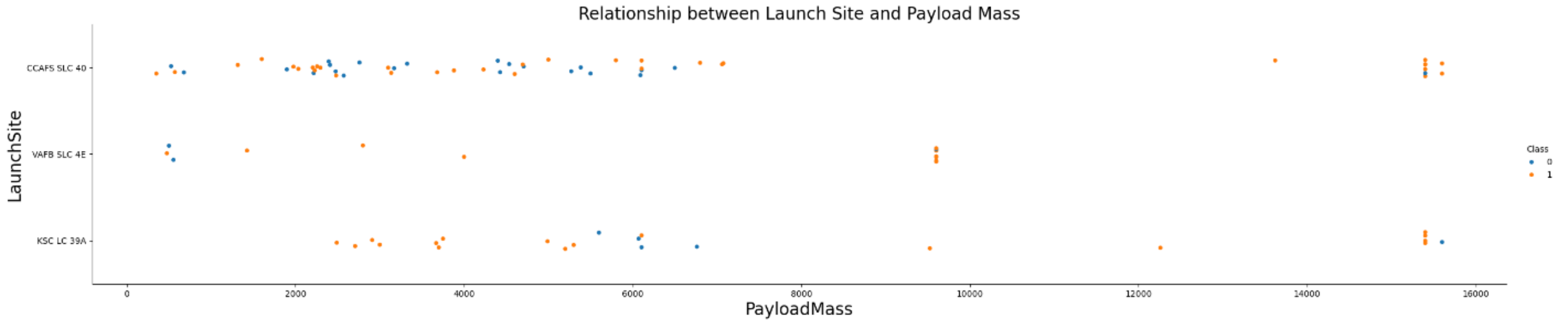


EDA with Data Visualization

- **Scatter Plots** to explore relationships between features:
  - Launch Site × Flight Number

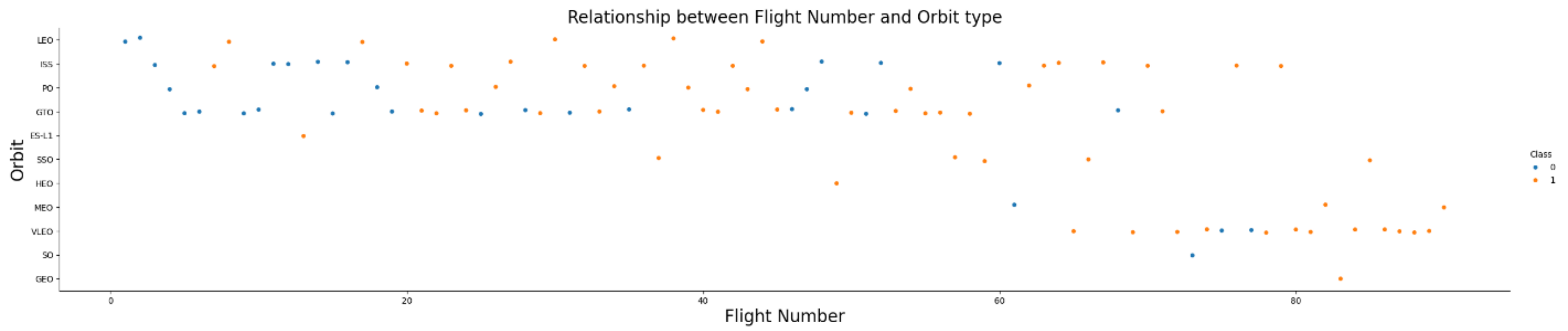


  - Launch Site × Payload Mass

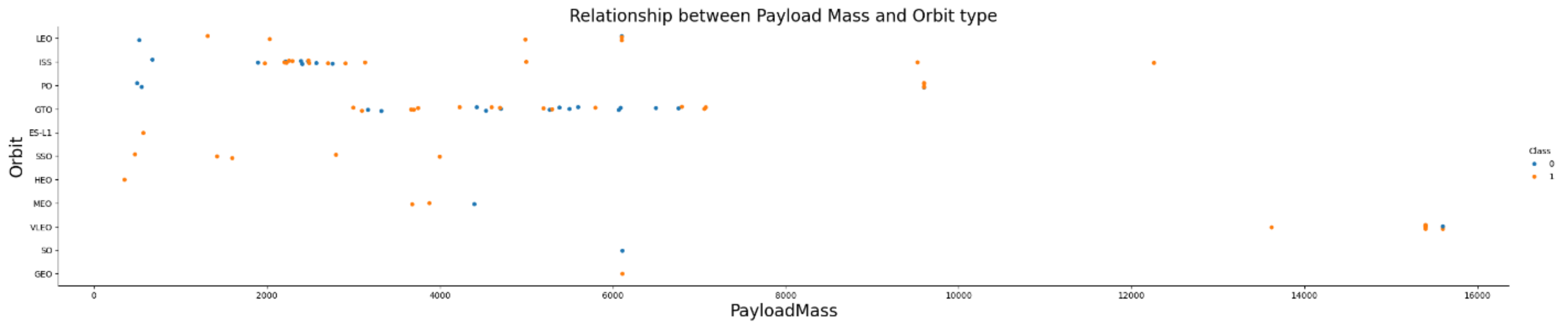


EDA with Data Visualization

- **Scatter Plots** to explore relationships between features:
  - Orbit × Flight Number



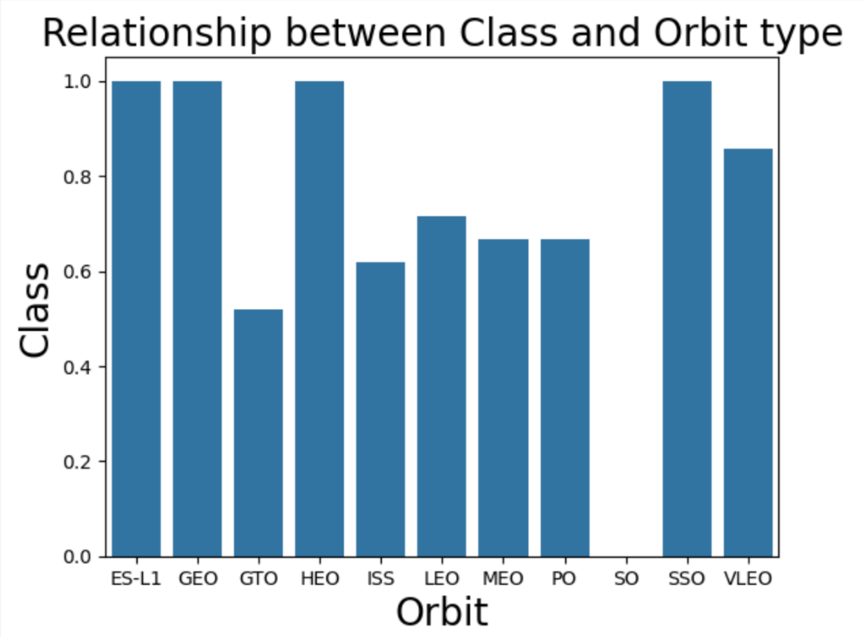
  - Payload × Orbit



EDA with Data Visualization

Visualizations Created:

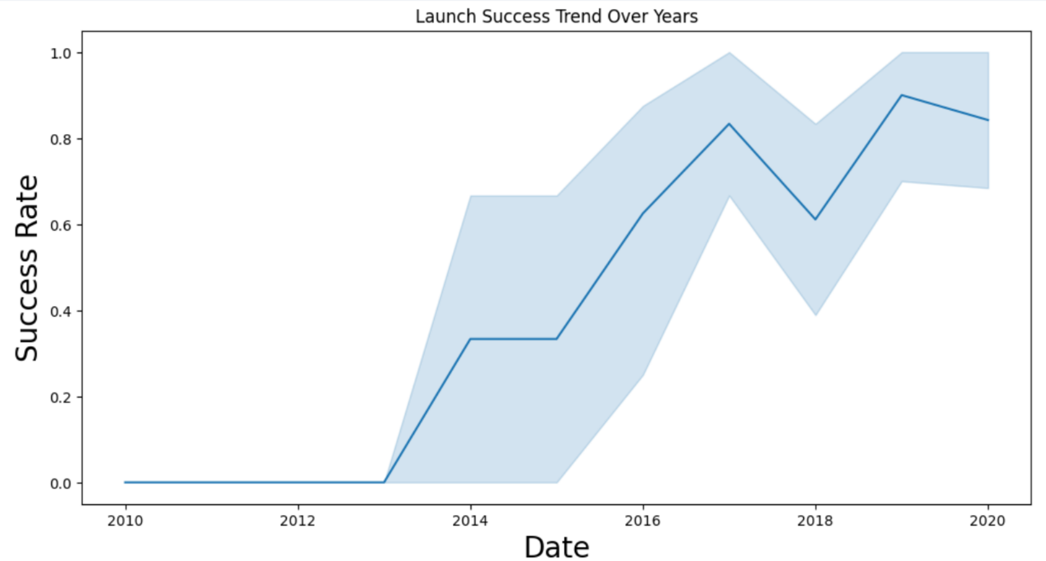
- **Bar Chart** to compare categorical features and highlight mission trends :



EDA with Data Visualization

Visualizations Created:

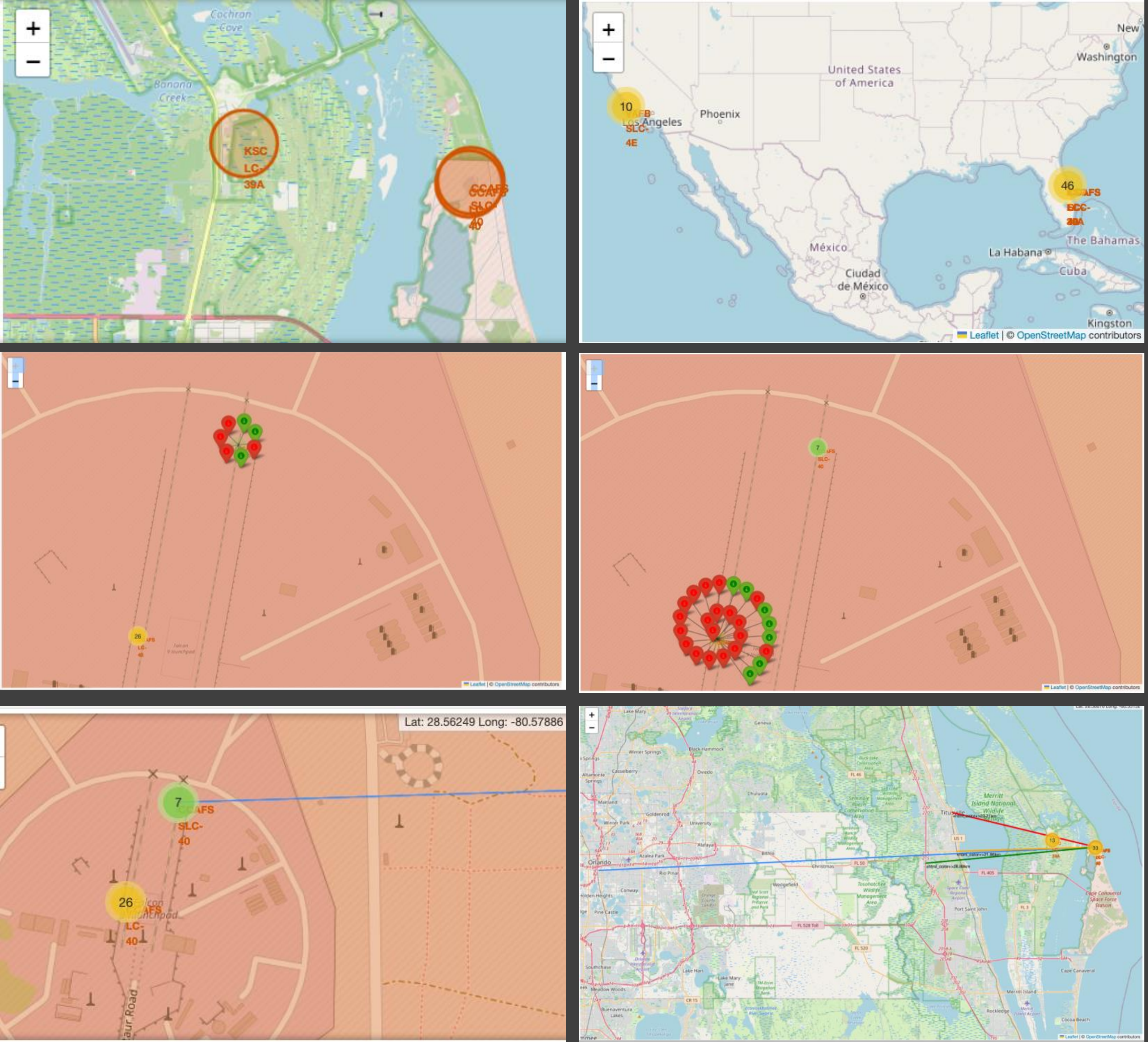
- **Line Chart** to visualize the launch success yearly trend:



Purpose:
To identify key patterns and relationships that could influence first-stage landing success and support machine learning model development.

Observation:
The success rate since 2013 kept increasing till 2020.

Source Code:
[GitHub Notebook - EDA with Visualization]



Build an Interactive Map with Folium

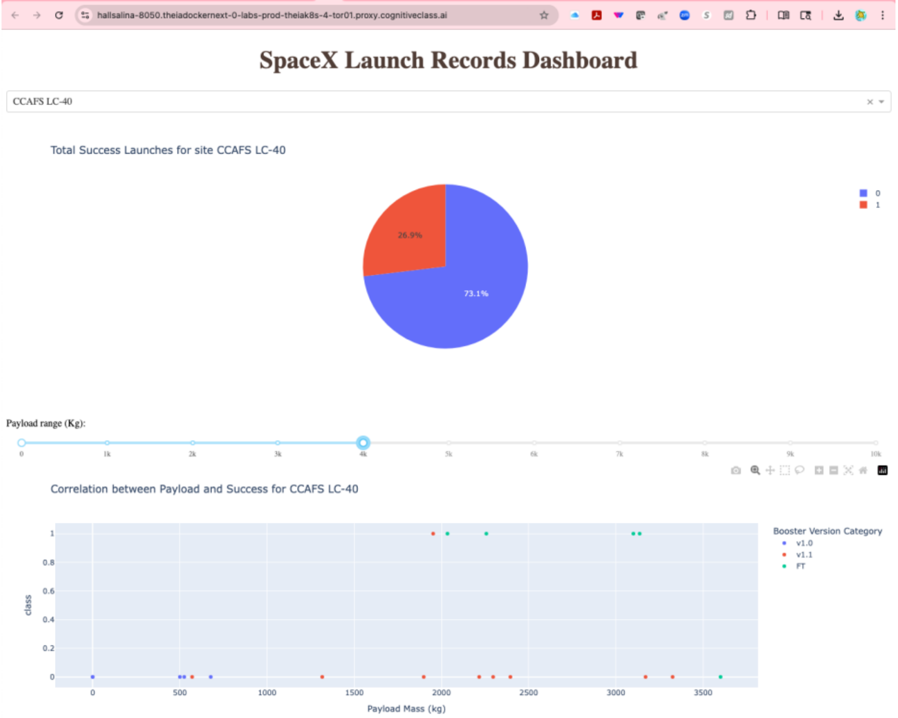
- In order to find some geographical patterns about launch sites, more interactive visual analytics using Folium was performed.
  - All launch sites on a map were marked as:
    - **Blue** circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
    - **Red** circles at all launch sites coordinates with a popup label showing its name using its latitude and longitude coordinates
    - showing its name using its name using its latitude and longitude coordinates
  - Sites with successful or failed launches were marked on the map as:
    - **Green** marker denoted successful launches
    - **Red** marker denoted failed launches.
    - From these color-labeled markers in marker clusters, it was easy to identify which launch sites have relatively high success rates.

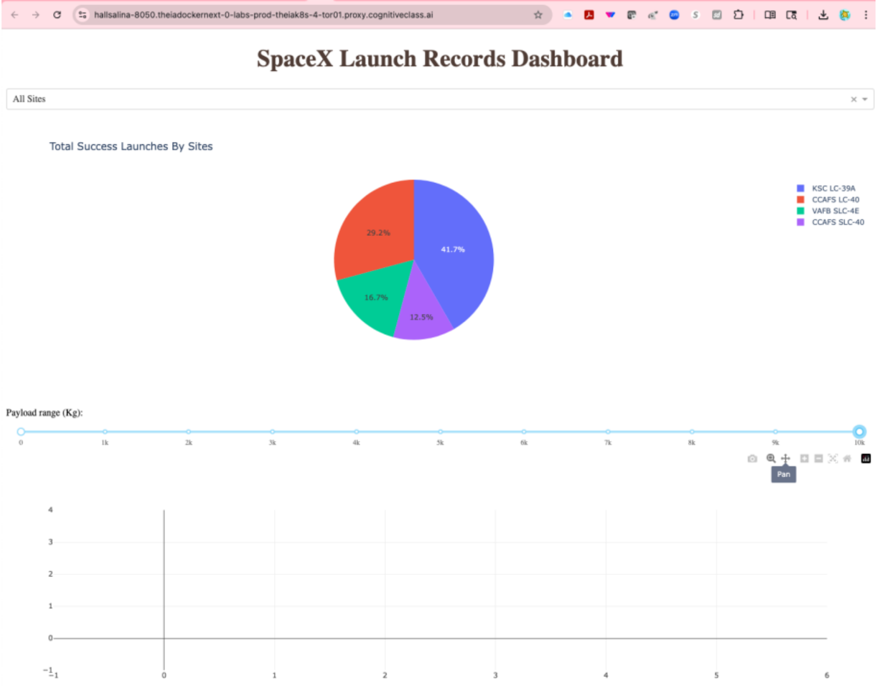
  - The distances between a launch site to its proximities were then calculated to explore and analyze the proximities of launch sites by plotting distance lines to the following:
    - City Distance 23.21
    - Railway Distance 21.96
    - Highway Distance 26.89
    - Coastline Distance 0.87

- [GitHub Notebook – Interactive Visual Analytics with Folium]

Build a Dashboard with Plotly Dash

- A Plotly Dash application was built for users to perform interactive visual analytics on SpaceX launch data in real-time.
- This dashboard application contained input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- These are the tasks performed:
    - Launch site drop-down list to allow users to select 1 or all launch sites.
    - Pie chart showing successful launches based on selected site in the launch site drop-down list.
    - A Range slider to select the Payload
    - A scatter plot with the x axis to be the payload and the y axis to be the launch outcome to visually observe how payload may be correlated with mission outcomes for selected site(s).
- [GitHub Notebook – Build a Dashboard with Ploty Dash]

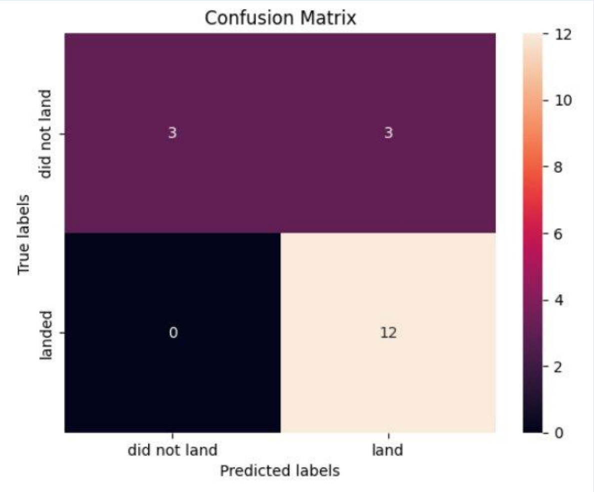




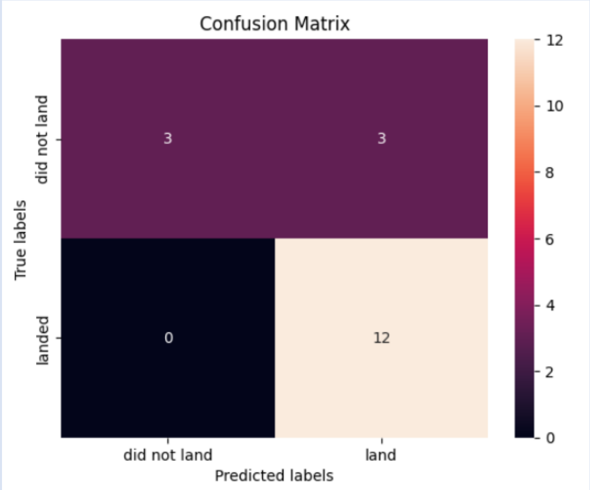
Predictive Analysis (Classification)

- **Model Development and Evaluation**
  - **Data Preparation:**
    - Created target labels from the *Class* column and converted to NumPy arrays.
    - Standardized the feature data using **StandardScaler**.
    - Split the dataset into **training (80%)** and **test (20%)** sets using train_test_split.
  - **Model Building & Optimization:**
    - Built and tuned four classification algorithms using **GridSearchCV (cv=10)**:
      - Logistic Regression
      - Support Vector Machine (SVM)
      - Decision Tree Classifier
      - K-Nearest Neighbors (KNN)
    - Performed hyperparameter tuning to identify optimal model settings.
  - **Model Evaluation:**
    - Evaluated models using **Accuracy**, **Jaccard Score**, and **F1 Score**.
    - Visualized performance through **confusion matrices**.
    - Determined the **best-performing classification model** for predicting first-stage landing success.
- GitHub Notebook – Predictive Analysis

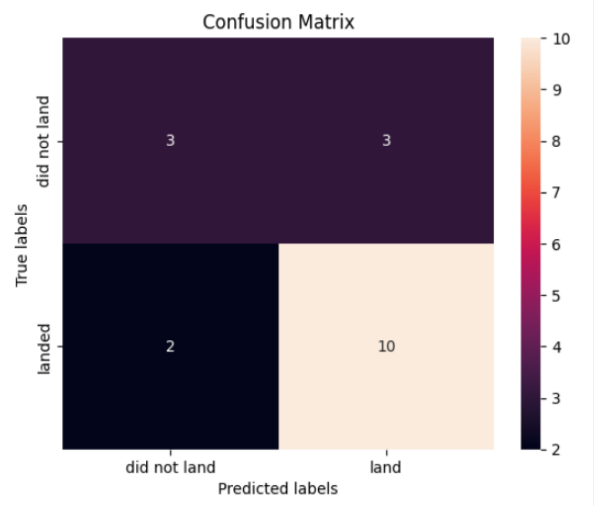
Logistic Regression



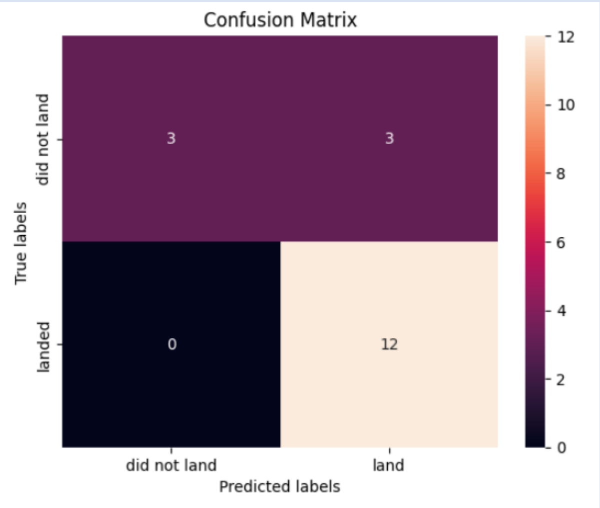
Support Vector Machine (SVM)



Decision Tree Classifier

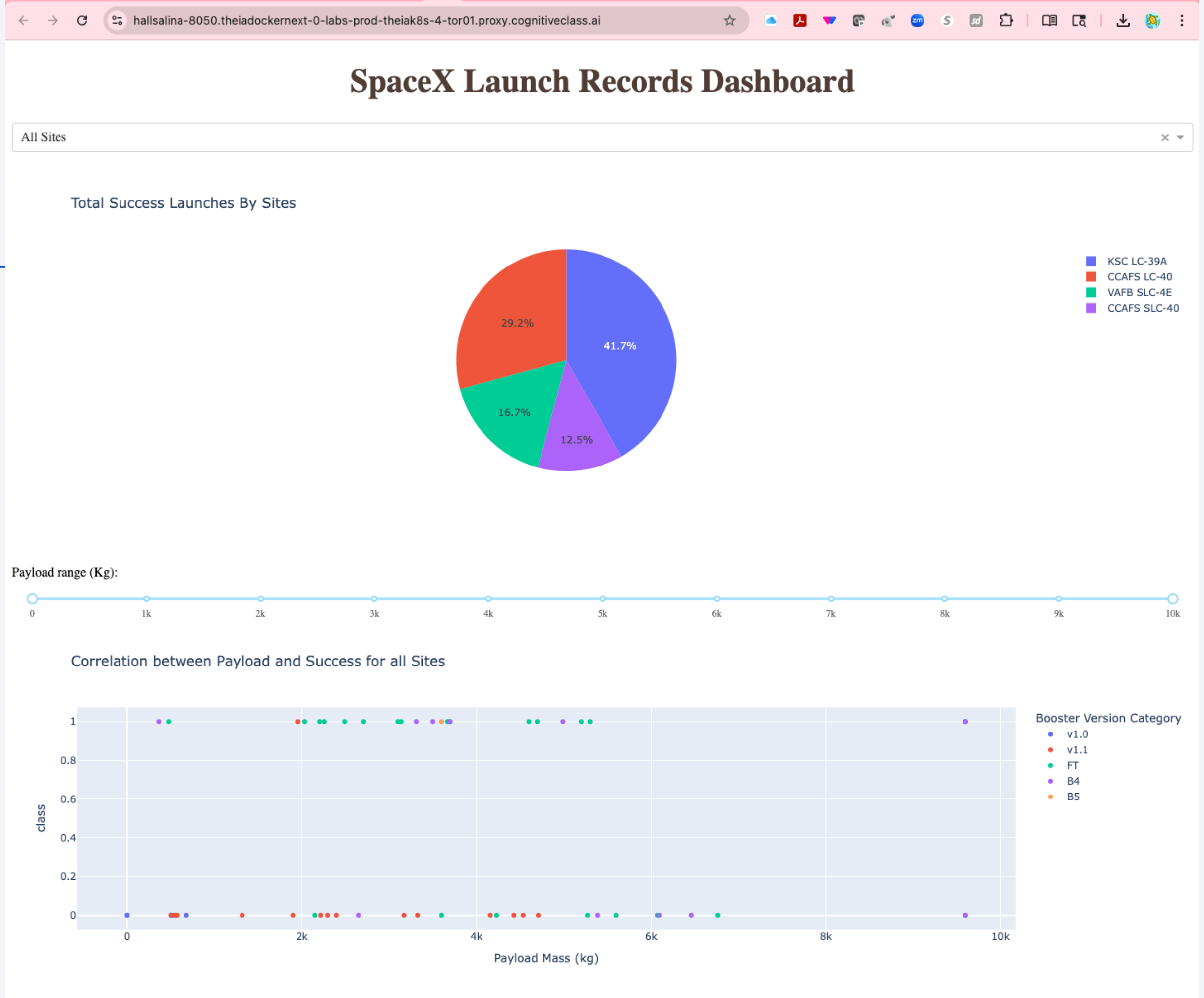


K-Nearest Neighbors (KNN)

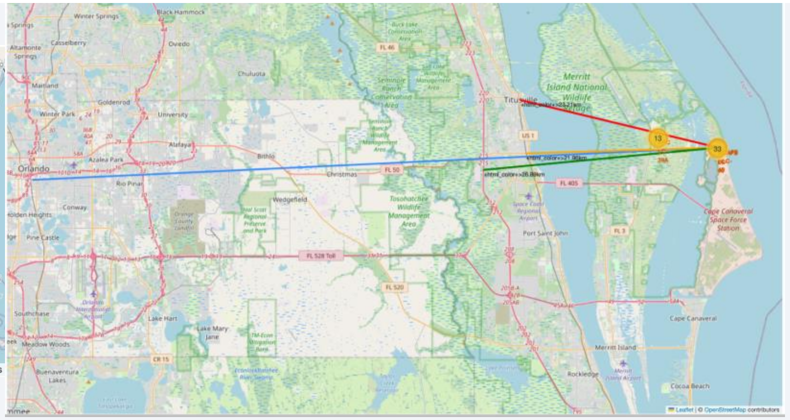


Results

- **Exploratory data analysis Findings:**
  - SpaceX uses **four launch sites**, mostly near the coast for safety and logistics.
  - **Launch success rates** have steadily increased since **2013**, reaching near 100% by 2020.
  - **KSC LC-39A** has the highest success rate among all launch sites.
  - **Payload mass** is **inversely** correlated with success — heavier payloads reduce landing success probability.
  - **Orbit type** influences outcomes: **ES-L1, GEO, HEO,** and **SSO** show the highest success rates, while **SO** has the lowest.
  - The **first successful landing** occurred in **2015**, five years after the first launch.



- **Interactive analytics demo and Insights:**
  - Most launches are from **East Coast sites**, strategically located near the ocean.
  - **Falcon 9 v1.1 boosters** averaged a payload of **~2,928 kg**, with several successful drone-ship landings above th average.
  - Two booster versions (**F9 v1.1 B1012** and **B1015**) failed during 2015 drone-ship landings.



Results

- **Predictive analysis results**
  - Findings: Logistic Regression, SVM, and KNN achieved equal test accuracy (83.3%), while Decision Tree performed lowest.
  - Best Model: SVM achieved the highest overall performance, with the strongest balance between training and test accuracy..
  - Feature relationships (launch site, orbit, payload mass) proved important predictors of first-stage landing outcomes.
  - Accuracy Results for Logistic Regression, SVM, and KNN Model for their Test Score and Train Score:

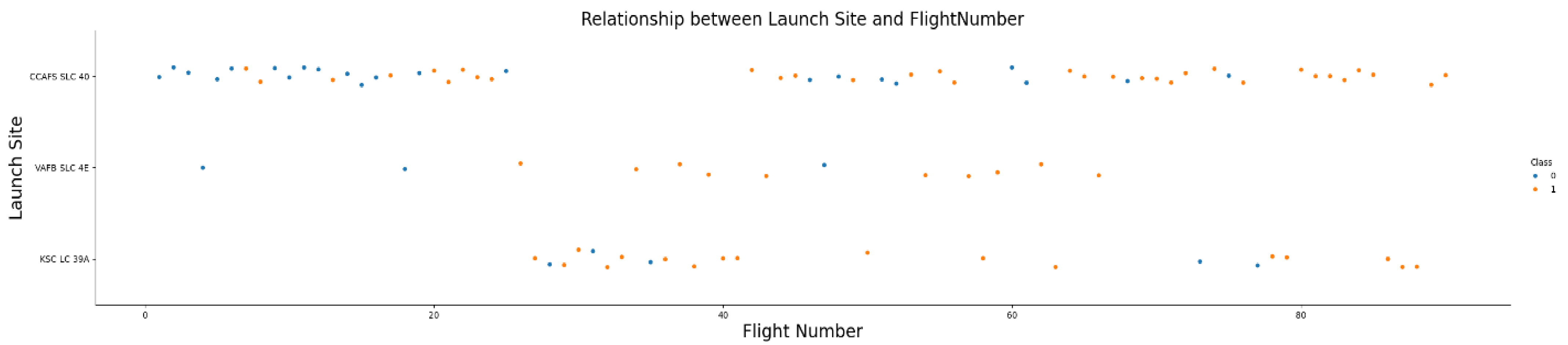
	Model	Test Score	Train Score
0	Logistic Regression	0.833333	0.875000
1	SVM	0.833333	0.888889
2	Decision Tree	0.722222	0.861111
3	KNN	0.833333	0.861111

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

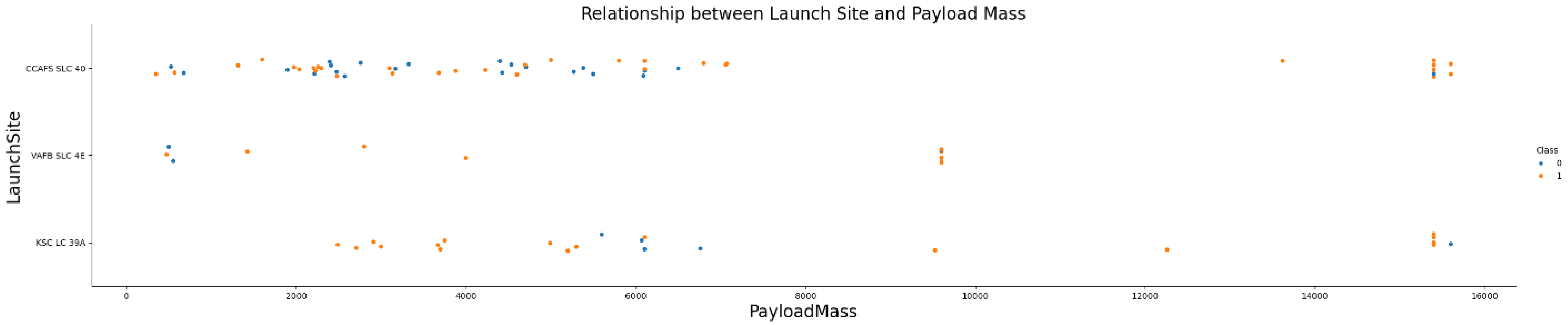
- Scatter plot of Flight Number vs. Launch Site
  - Blue = Fail
  - Orange = Success
- Launch sites with more frequent flights show higher success rates.
- Earlier launches had more failures, while recent launches show greater success.
- About half of all launches occurred at CCAFS SLC-40.
- VAFB SLC-4E and KSC LC-39A demonstrate the highest landing success rates.
- Overall, newer launches achieve higher success, reflecting improvements in technology and operations.



Payload vs. Launch Site

Scatter plot of Payload Mass vs. Launch Site

- Generally, heavier payloads correspond to higher success rates.
- Most launches over 7,000 kg were successful.
- KSC LC-39A achieved 100% success for payloads under 5,500 kg.
- VAFB SLC-4E has not launched payloads above ~10,000 kg but the majority of the launches with payloads above 1000 kg have been successful.
- At CCAFS SLC-40, most launches carry payloads under 7,000 kg, where the success rate is roughly 50/50. However, for launches above 7,000 kg, the majority have been successful.

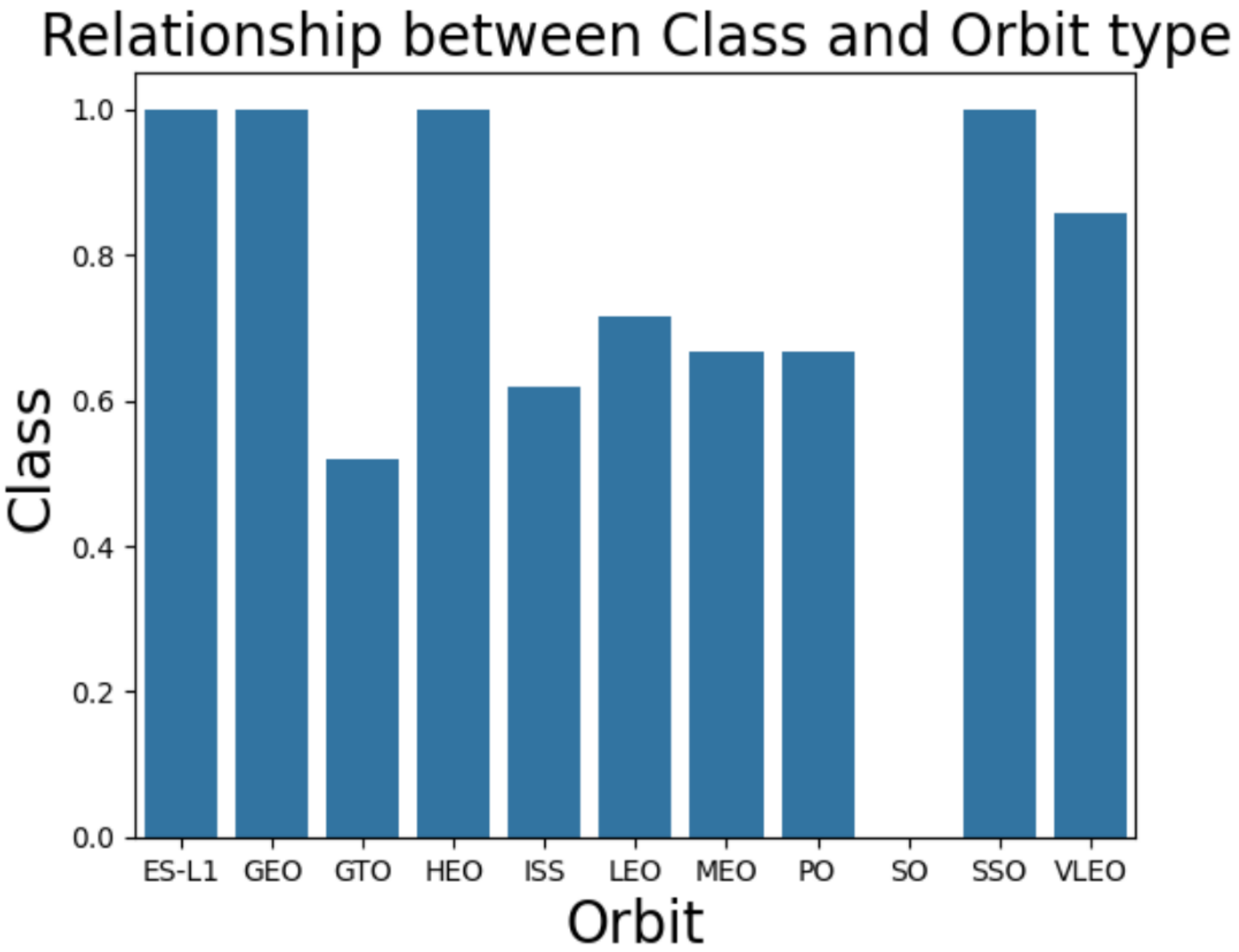


Success Rate vs. Orbit Type

Bar chart for the success rate (Class) of each orbit type findings:

- From the plot we can see that orbits ES L1, GEO, HEO and SSO Were the most successful at 100%

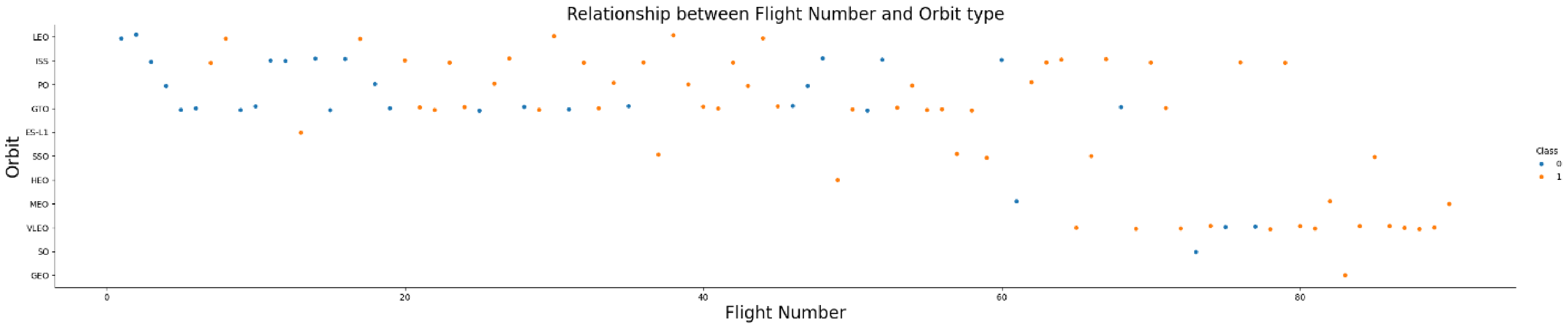
Success Rate	Orbits
100%	ES L1, GEO, HEO and SSO
50% - 90%	GTO, ISS,LEO, MEO, PO, VLEO
0%	SO



Flight Number vs. Orbit Type

Scatter point of Flight number vs. Orbit type findings:

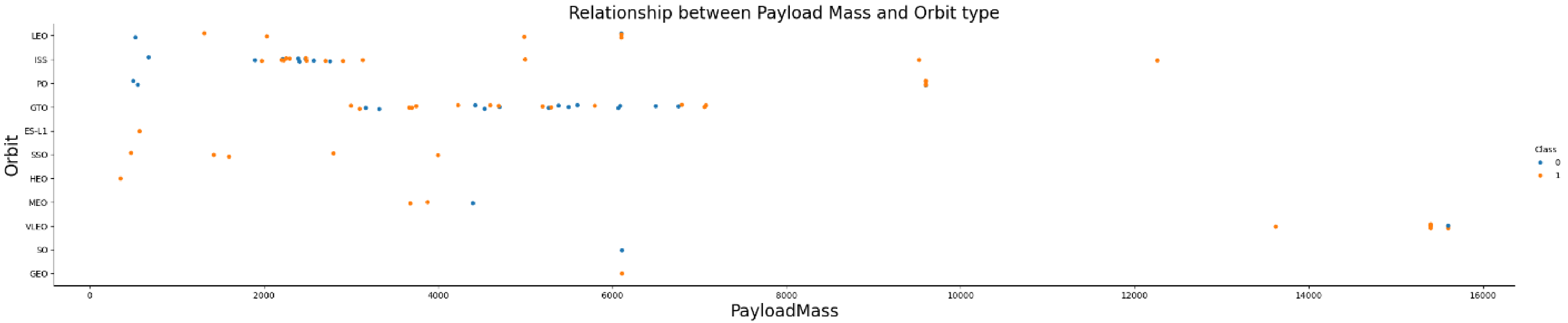
- Success generally increases with the number of flights.
- In the LEO orbit, success seems to be related to the number of flights.
- In GTO orbit, there is no apparent relationship between flight count and success.



Payload vs. Orbit Type

Scatter point of payload vs. orbit type findings:

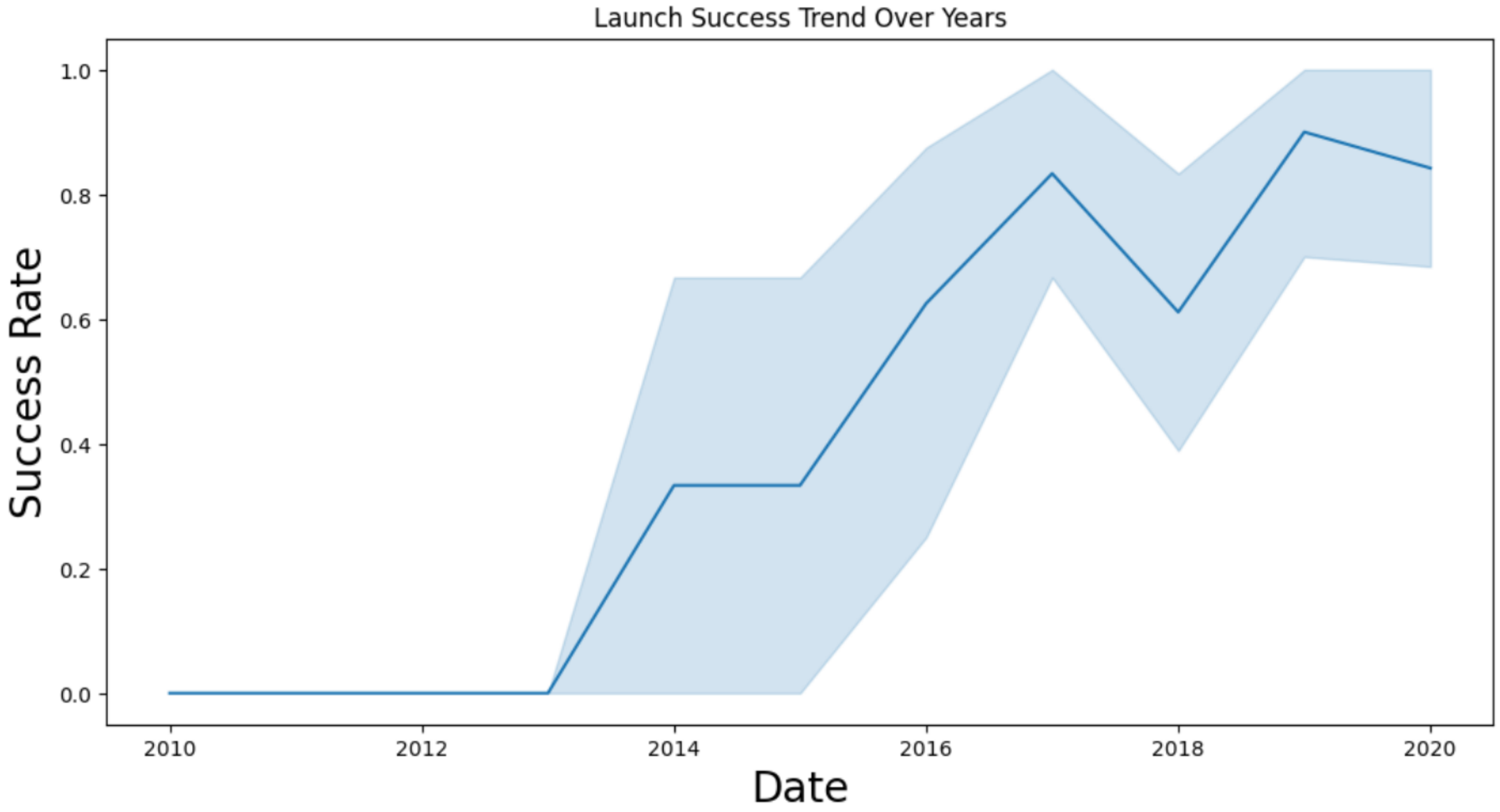
- Heavier payloads show higher success rates in LEO, ISS, and PO orbits.
- GTO orbits show mixed success with heavy payloads.
- Overall, low-Earth and station orbits perform best for heavy launches.



Launch Success Yearly Trend

Line chart of yearly average success rate findings:

- Overall success rate has steadily improved since 2013.
- Notable increases from 2013–2017 and 2018–2019.
- Slight declines observed between 2017–2018 and 2019–2020.



All Launch Site Names

- This query returns a **unique list of all launch sites** from the dataset by removing duplicates.
  It helps identify **how many and which distinct SpaceX launch locations** are included in the data; providing insight into the **geographic distribution of launches**.
  - `%sql select distinct Launch_Site from SPACEXTABLE;`

Task 1

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

In [10]:
```
%sql select distinct Launch_Site from SPACEXTABLE;
```

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

- This query retrieves the first five records from the dataset where the launch site name begins with "CCA" — filtering for launches that occurred at Cape Canaveral (CCAFS) locations.
  - `%sql select * from SPACEXTABLE WHERE Launch_Site like 'CCA%' limit 5;`
- It helps identify the subset of missions launched from CCAFS SLC-40 and CCAFS SLC-41, which are key SpaceX launch sites.

Task 2

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

In [11]:
```
%sql select * from SPACEXTABLE WHERE Launch_Site like 'CCA%' limit 5;
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- The total payload carried by boosters from NASA was calculated using the query:
  - `%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';`
- **Total Payload Mass:**
  - 45,596 kg carried by boosters launched for **NASA (CRS)**

Task 3

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

```
In [26]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';

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

Out[26]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated using the query:
  - `%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1';`
- **Average Payload Mass:**
  - 2,928 kg for **booster version F9 v1.1.**
- In summary, NASA (CRS) launches carried a total of 45,596 kg, and F9 v1.1 boosters had an average payload of 2,928 kg, showing consistent payload capacity across missions.

Task 4

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

In [18]:

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1';
```

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

Out[18]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Using the below query, the dates of the first successful landing outcome on ground pad was retrieved.
  - `%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';`
- The first successful landing on a ground pad occurred on 22nd December 2015.

```
In [23]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.

Out[23]:
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query uses the WHERE clause and AND for more than one criteria to filter for the list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - `%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;`
- Four boosters meet these criteria, as shown in the screenshot.

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

```
[28]: %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.
```

[28]:

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

Total Number of Successful and Failure Mission Outcomes

- This query counts the total number of missions for each Mission_Outcome, grouping them by outcome and ordering the results from most to least frequent. Wildcards like % can be used with LIKE in the WHERE clause to filter for outcomes containing specific text, such as success or failure.
    - `%sql select Mission_Outcome, count(Mission_Outcome) as 'Count' from SPACEXTABLE GROUP BY Mission_Outcome ORDER BY count DESC;`
- The total number of successful outcomes was 101 and failure mission outcomes was 1.

```
[32]: %sql select Mission_Outcome, count(Mission_Outcome) as 'Count' from SPACEXTABLE GROUP BY Mission_Outcome ORDER BY count DESC;

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

[32]:

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

Boosters Carried Maximum Payload

- Using a subquery in the WHERE clause and MAX() function, this query listed the booster versions that have carried the maximum payload.
  - %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
- 12 boosters were identified that have carried the maximum payload. These are the names of these booster:
  - **Booster_Versions**

• F9 B5 B1048.4	• F9 B5 B1049.5
• F9 B5 B1049.4	• F9 B5 B1060.2
• F9 B5 B1051.3	• F9 B5 B1058.3
• F9 B5 B1056.4	• F9 B5 B1051.6
• F9 B5 B1048.5	• F9 B5 B1060.3
• F9 B5 B1051.4	• F9 B5 B1049.7

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
[33]: %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

[33]:

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

- **Using substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year, the following query lists the** failed landing_outcomes in drone ship, their booster versions, the month name, and launch site names for in year 2015.
  - %sql select substr(Date,6,2) as 'Month', Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome ='Failure (drone ship)' and substr(Date,0,5)='2015';
- Two failed drone ship landings in 2015 were identified.
- Both failures occurred at CCAFS LC-40, involving different booster versions in different early-year months.

```
[50]: %sql select substr(Date,6,2) 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.
```

[50]:

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

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

- This query ranks **landing outcomes** (e.g., *Success (ground pad)*, *Failure (drone ship)*) based on their **frequency** between the dates **2010-06-04** and **2017-03-20**, in descending order. By using **GROUP BY**, outcomes are grouped by type, and **ORDER BY** arranges them in **descending order** to highlight which outcomes occurred most often.
    - %sql select Landing_Outcome, Date, count(Landing_Outcome) as 'Count' from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' GROUP BY Landing_Outcome ORDER BY count DESC;
- Between **2010 and 2017**, most missions initially made **no landing attempts (10)**.
- As SpaceX advanced its technology, **drone ship and ground pad successes** became more frequent — with **5 successful drone ship** and **3 successful ground pad** landings recorded.
- This trend reflects **steady improvement in landing reliability** and **transition from test flights to successful recoveries**.

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.

```
[48]: %sql select Landing_Outcome, Date, count(Landing_Outcome) as 'Count' from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' GROUP BY Landing_Outcome ORDER BY count DESC;

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

[48]:

Landing_Outcome	Date	Count
No attempt	2012-05-22	10
Success (drone ship)	2016-04-08	5
Failure (drone ship)	2015-01-10	5
Success (ground pad)	2015-12-22	3
Controlled (ocean)	2014-04-18	3
Uncontrolled (ocean)	2013-09-29	2
Failure (parachute)	2010-06-04	2
Precluded (drone ship)	2015-06-28	1

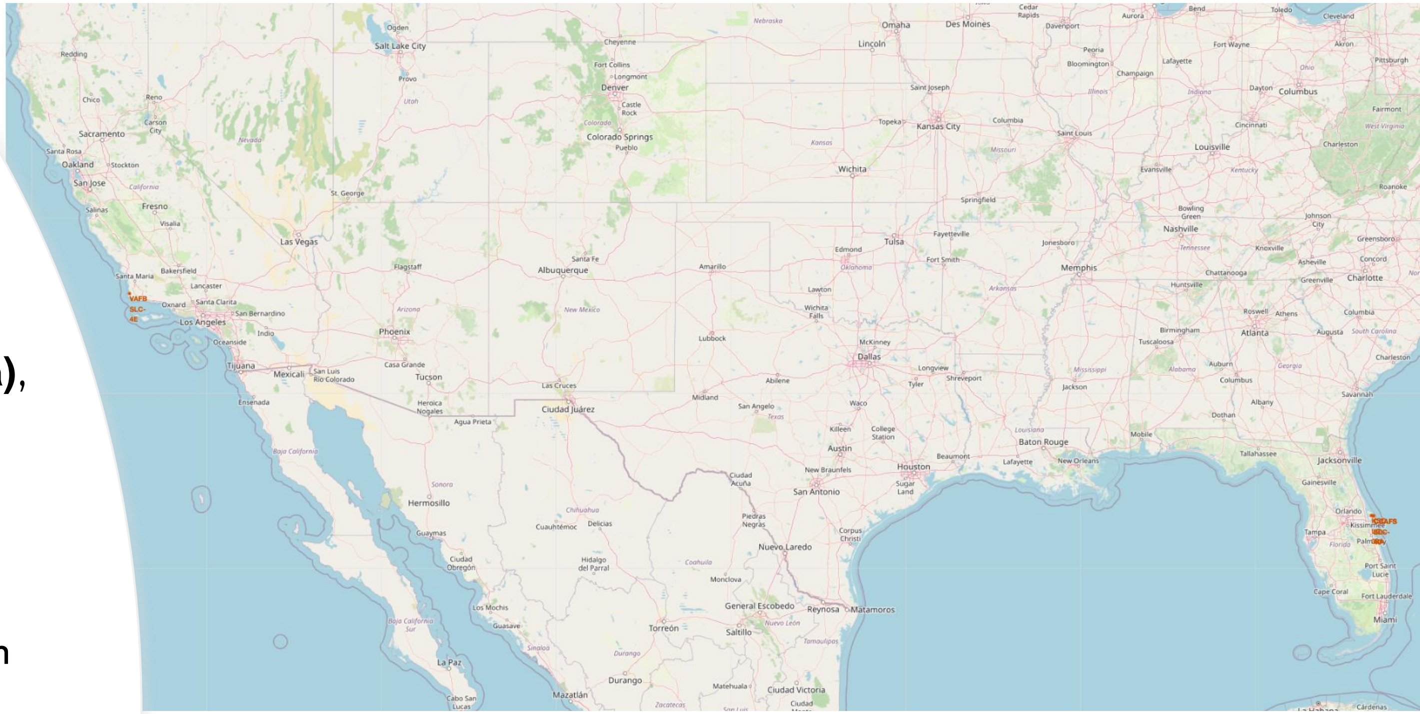
Section 3

Launch Sites Proximities Analysis

Global Launch Sites - Folium Map Overview

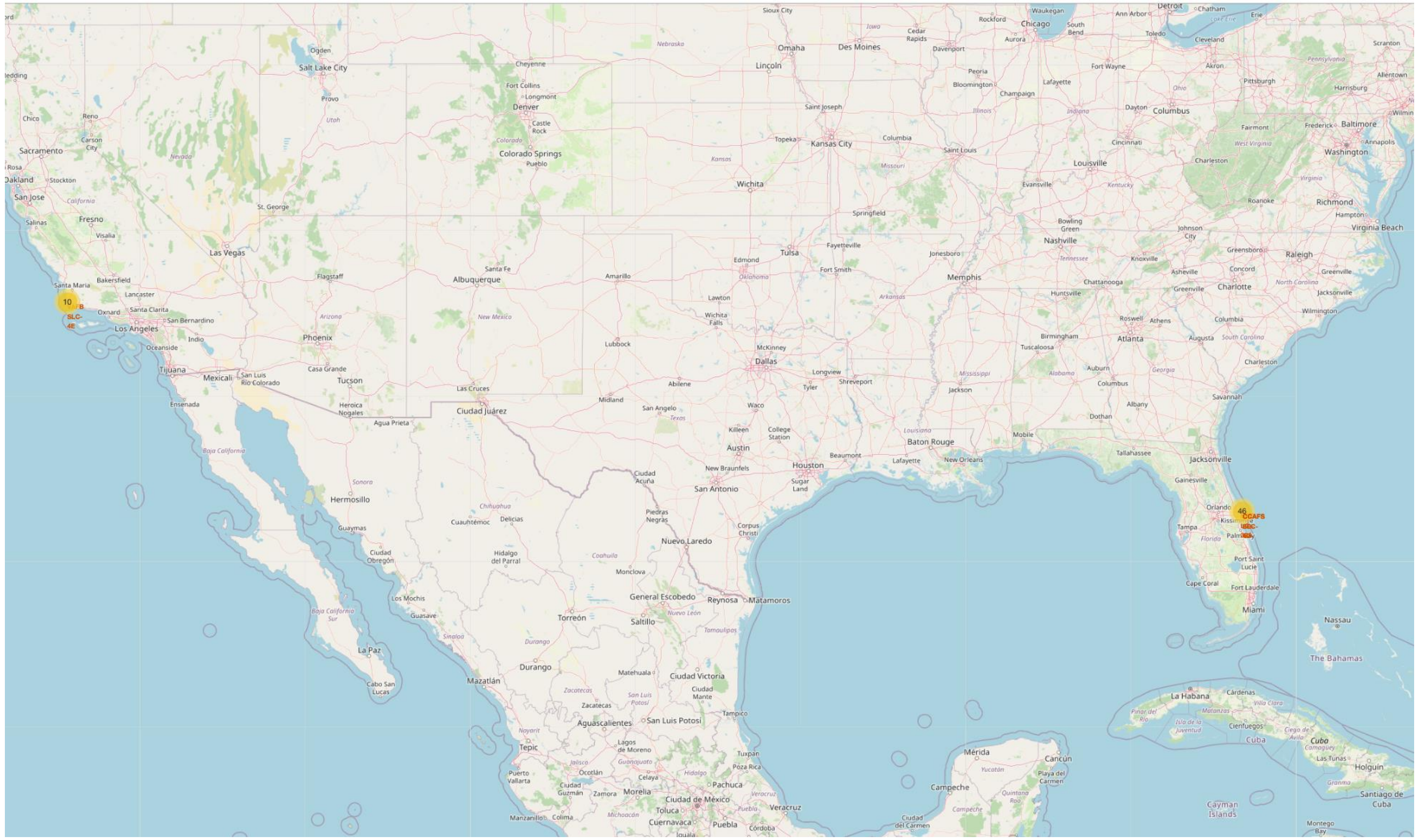
- The Folium map displays all **SpaceX launch sites** marked globally, primarily located across the **United States** — including **CCAFS SLC-40 (Florida)**, **KSC LC-39A (Florida)**, and **VAFB SLC-4E (California)**.

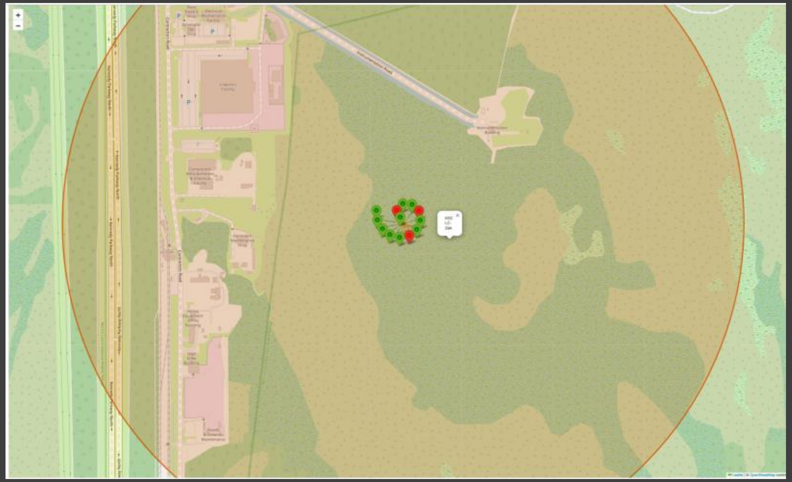
- **Key Observations**
- 🌍 **Proximity to the Equator:**
 Not all launch sites are directly on the Equator, but they are positioned in lower latitudes, particularly those in **Florida**, which are **closer to the Equator** and gain a **rotational boost from Earth's spin** — improving efficiency for **prograde orbits**.
- 🌊 **Proximity to the Coast:**
 Yes, all launch sites are located near coastlines. This allows rockets to launch safely over the ocean, **minimizing risk to populated areas** in case of failure.
- 🚀 **Strategic Advantage:**
 Launch sites near the Equator and coastlines help reduce fuel costs, **improve launch safety**, and **enable efficient orbital access** for various mission types.
- **Summary**
 The Folium map illustrates that while not all launch sites are on the Equator, they are **strategically located at low latitudes and close to the coast**, balancing **safety, cost efficiency, and orbital performance**.

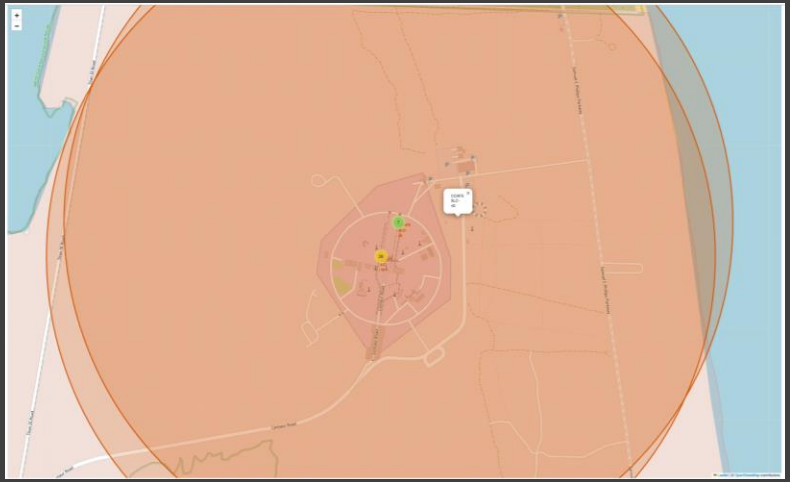


US Launch Sites

- The map displays markers indicating launch sites in the United States, primarily located on the East Coast (Florida) and the West Coast (California), all in the southern regions of the country.

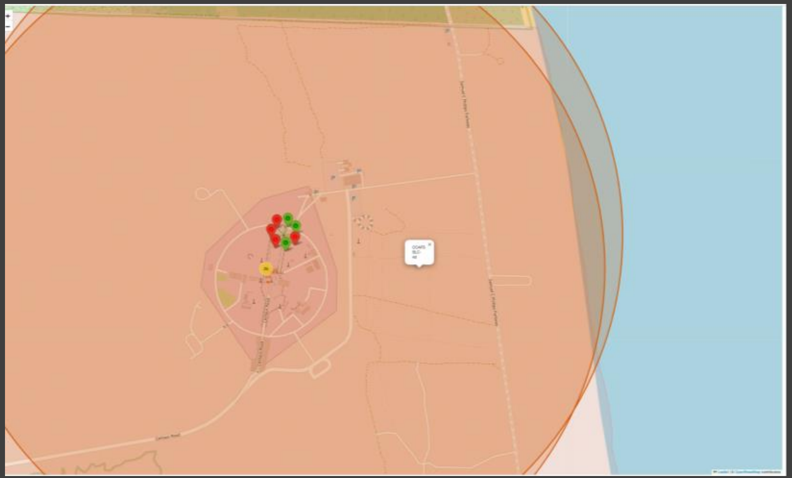


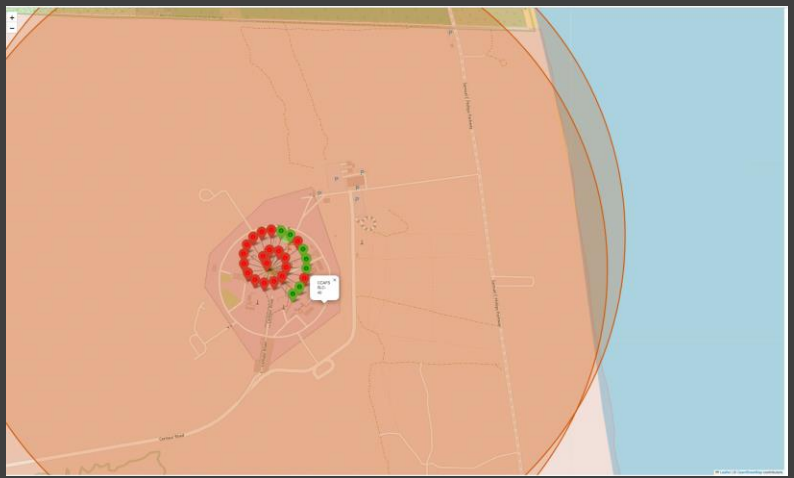


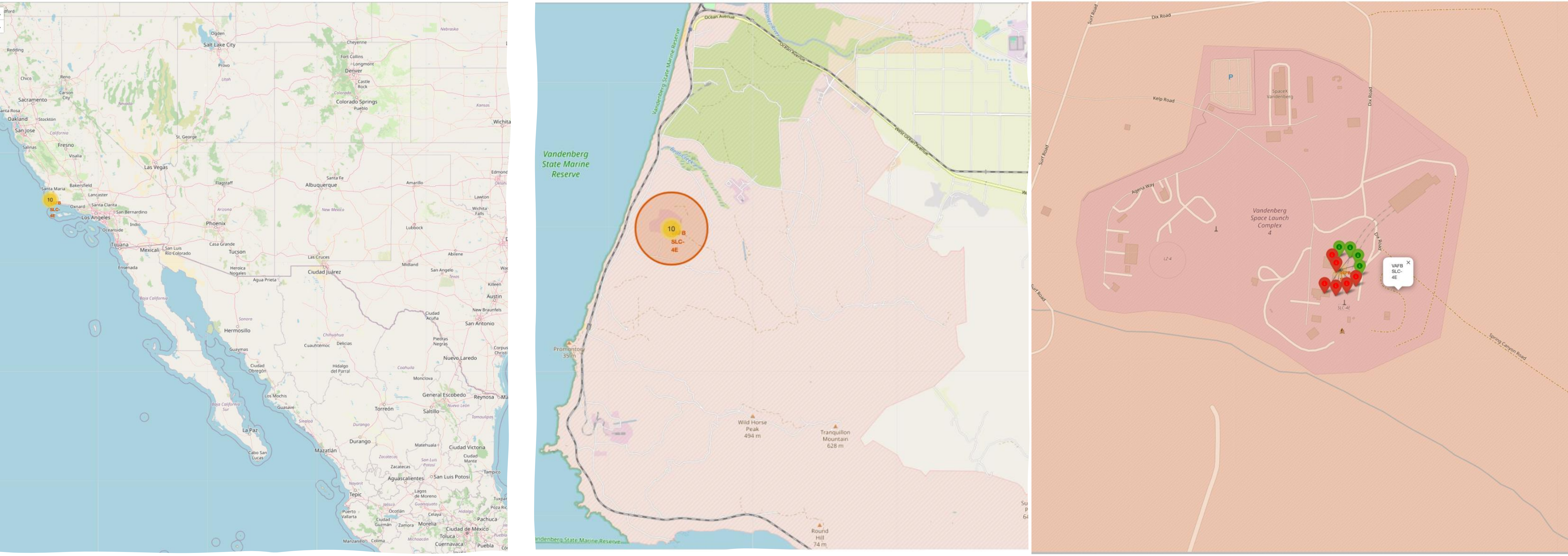


Florida Launch Sites and Success Rates

- Markers are color-coded to represent launch outcomes—success or failure—and are grouped in clusters for easier visualization.
- In Florida, Cape Canaveral Space Force Station (CCAFS) SLC-40 shows a higher proportion of failed launches compared to Kennedy Space Center (KSC) LC-39A, making it clear that KSC LC-39A has a relatively higher success rate.





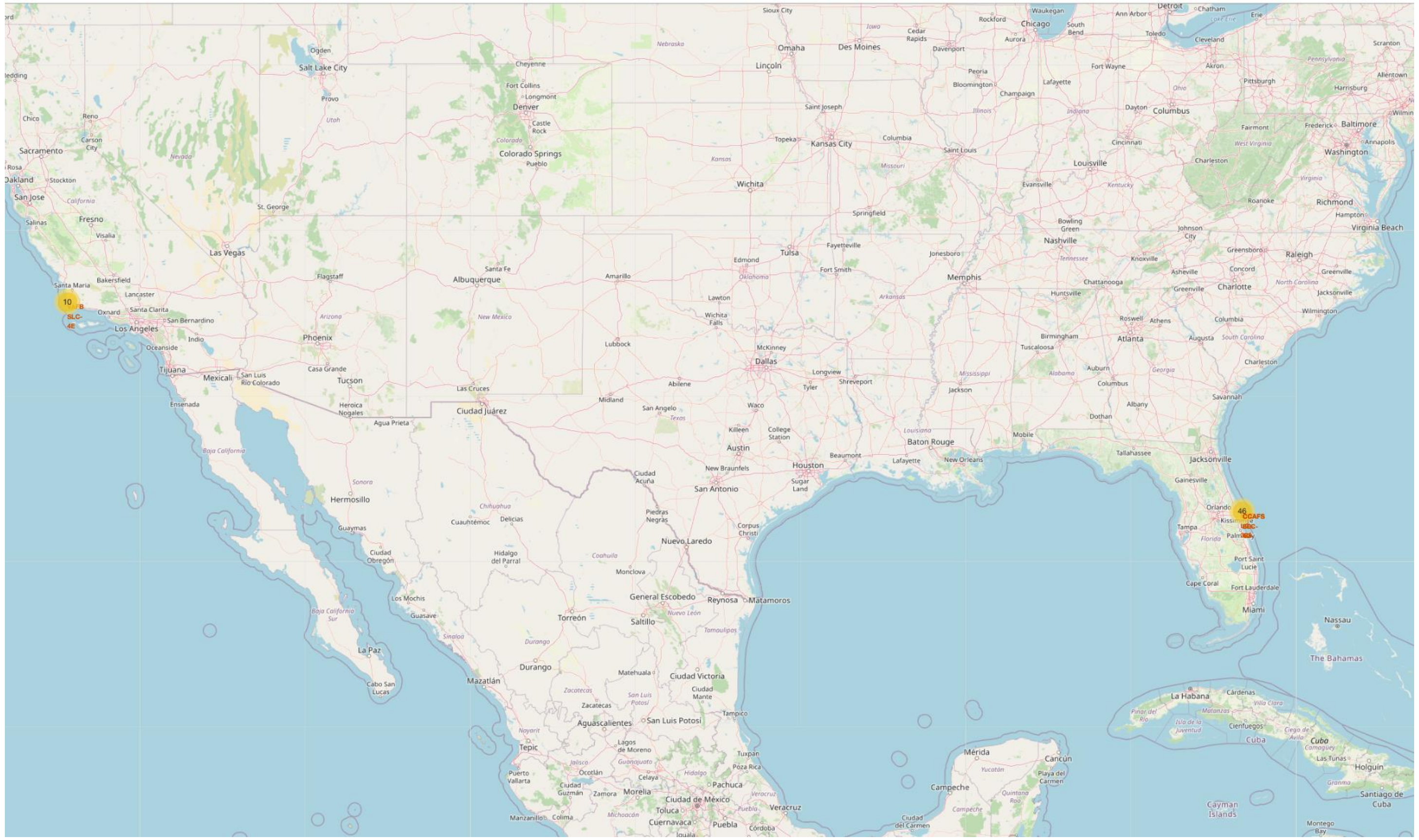


California Launch Sites and Success Rates

- In California, Vandenberg Air Force Base (VAFB) SLC-4E has experienced more failures, with 6 out of 10 launches unsuccessful.

US Launch Sites Summary

- US launch sites are concentrated in the southern regions of Florida and California.
- Success rates vary by site: KSC LC-39A in Florida demonstrates relatively high reliability, whereas CCAFS SLC-40 and VAFB SLC-4E show higher proportions of failed launches.
- Overall, the map highlights that even within the same state, launch site performance can differ significantly, emphasizing the importance of site-specific operational factors.



SpaceX Launch Site Proximity Analysis

```
print("City Distance", city_distance)
print("Railway Distance", railway_distance)
print("Highway Distance", highway_distance)
print("Coastline Distance", distance_coastline)
```

```
City Distance 23.21117556141272
Railway Distance 21.964008936972807
Highway Distance 26.890532380918604
Coastline Distance 0.8718314338759485
```

- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

The Map shows the selected launch site with markers for nearby city, railway,highway, and coastline, including measured distances from the pad.

- Key Findings
  - Near the Coast: Launch sites are intentionally located near coastlines to ensure safe launch trajectories over water.
  - Easy Access: Ground transportation access (highways, railways) is strategically integrated to support logistics and heavy equipment movement.
  - Efficient Placement: The proximity measurements help confirm that site placement is optimized for: safety; fuel and performance efficiency; access to orbital paths; and operational logistics.

Section 4

Build a Dashboard with Plotly Dash

Launch Success Distribution by Site

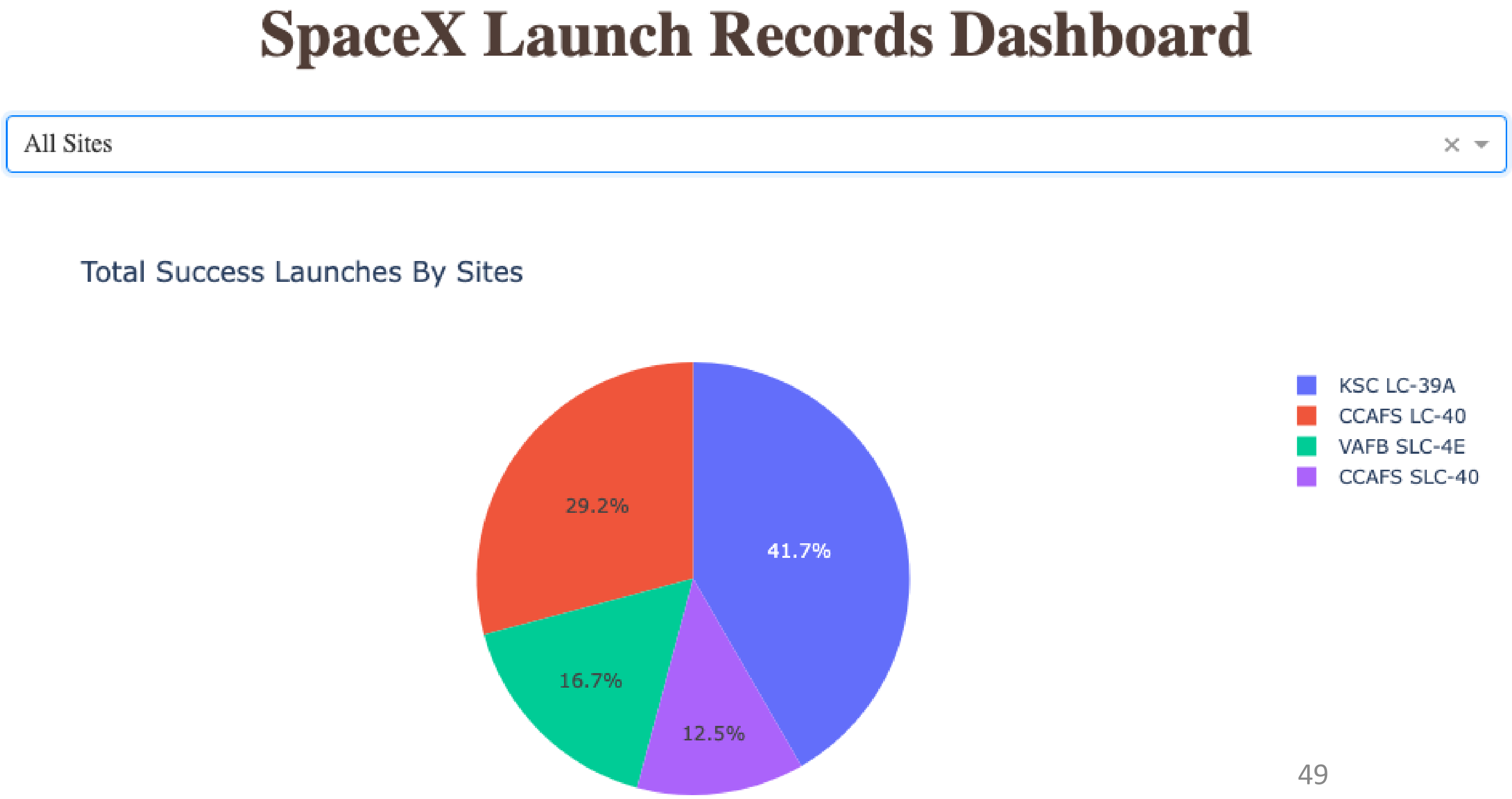
Screenshot Summary

- The pie chart displays the success counts for all SpaceX launch sites.
- Each slice represents a launch site's percentage contribution to total successful launches.

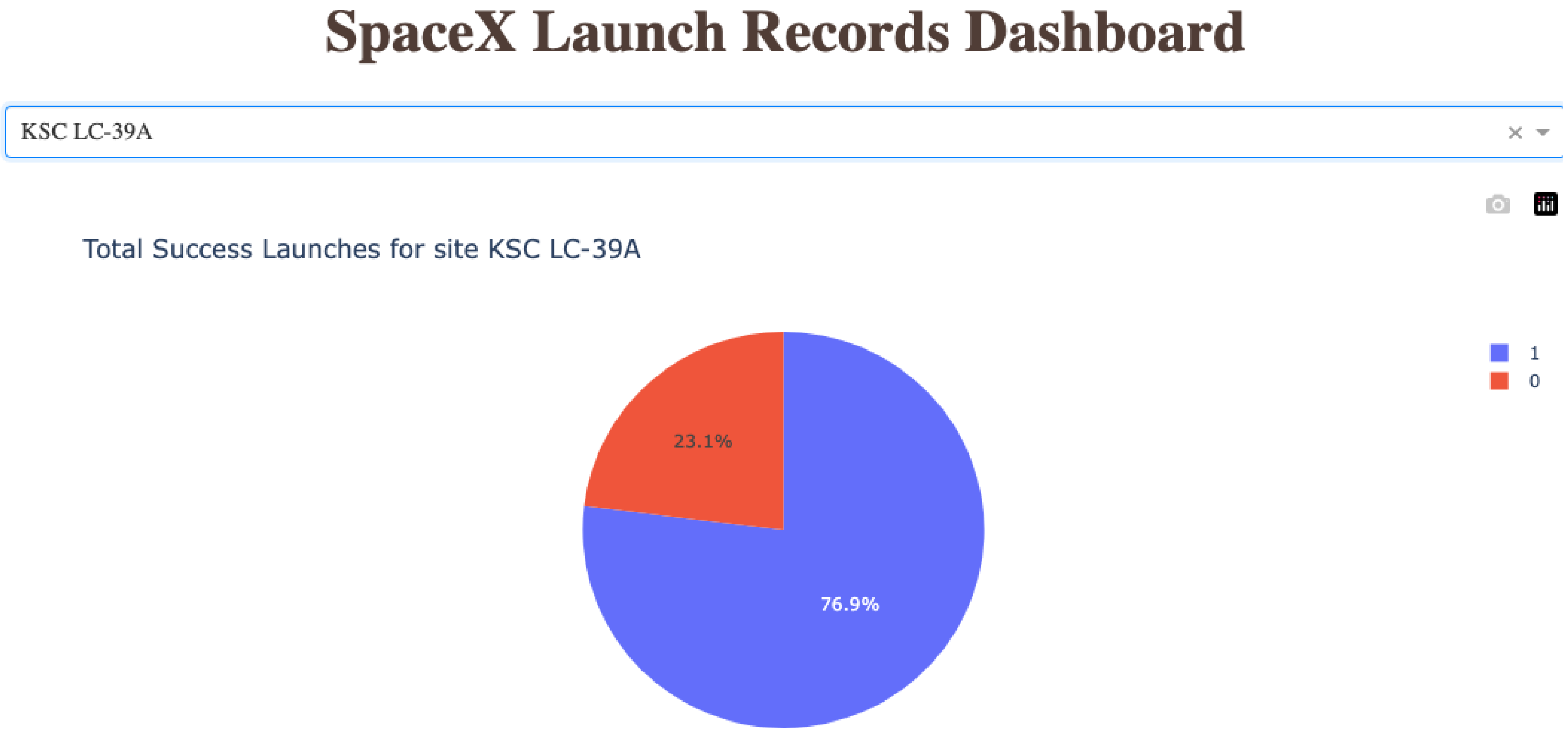
Key Findings

- KSC LC-39A has the highest share of successful launches at 41.7%, making it the top-performing site.
- CCAFS LC-40 contributes a substantial portion (29.2%) of total successes.
- VAFB SLC-4E accounts 16.7%.
- CCAFS SLC-40 represents the remaining 12.5% of successful launches.

Overall, the distribution shows that Florida launch sites (mainly KSC LC-39A and CCAFS LC-40) dominate SpaceX's successful missions.



Launch Site with Highest Launch Success Ratio



KSC LC-39A has the highest launch success ratio of all sites.

- Success Rate: 76.9%
  - 10 successful launches
- Failure Rate: 23.1%
  - 3 failed launches
- The pie chart clearly highlights the strong performance of KSC LC-39A, with the majority of launches shown in the success segment.
- This visual emphasizes KSC LC-39A reliability and its role as the top-performing SpaceX launch site.

Payload vs. Launch Outcome - Scatter plot for all sites, with different payload selected in the range slider

Key Elements & Findings

- The scatter plot shows at all sites, the payload mass on the x-axis and launch outcome (1 = success, 0 = failure) on the y-axis, filtered using the payload range slider.
- Payloads between ~2,000 kg and 5,000 kg show the highest concentration of successful launches (mostly outcome = 1).
- Lower-weight payloads generally have higher success rates across booster versions.
- Heavier payloads (above ~5,300 kg) show lower success rates and fewer cases overall, indicating reduced reliability at higher weights.
- When filtering by range, the plot shows that certain booster versions consistently achieve successful outcomes in the mid-weight range, reinforcing that success is more common with moderate payloads.

Booster Version vs Payload vs. Launch Outcome for all sites Findings

Booster Version Performance – Summary

- Booster Version FT
  - Most launches overall (22)
  - Highest number of successes: 14/22
  - Used at all launch sites, each showing at least some successful outcomes
- Booster Version v1.0
  - 4 launches – all failed (0% success)
- Booster Version v1.1
  - 15 launches – only 1 success
- Booster Version B4
  - 6 successes out of 11 launches
- Booster Version B5
  - 1 launch – failed

Overall Insight

- FT is the most reliable and widely used booster version, showing consistent success across every launch site.
- Early versions (v1.0, v1.1) had poor performance, while B4 shows moderate improvement.

Plotly Dash Application Findings

Key Findings

1. Which site has the largest number of successful launches?
   - KSC LC-39A
   - It accounts for 41.7% of all successful launches, the highest of any site.
2. Which site has the highest launch success rate?
   - KSC LC-39A
   - 76.9% success rate (10 successes out of 13 launches), making it the most reliable site.
3. Which payload range has the highest launch success rate?
   - ~2,000 kg to 5,000 kg
   - This payload band consistently shows the highest success rate across booster versions and launch sites.
4. Which payload range has the lowest launch success rate?
   - Heavier payloads above ~5,600 kg
   - As payload mass increases, success probability decreases, especially at certain sites.
5. Which F9 booster version has the highest launch success rate?
   - Booster Version FT
   - 14 successful launches out of 22
   - Highest success count and strong performance across all launch sites, unlike earlier versions.

Section 5

Predictive Analysis (Classification)

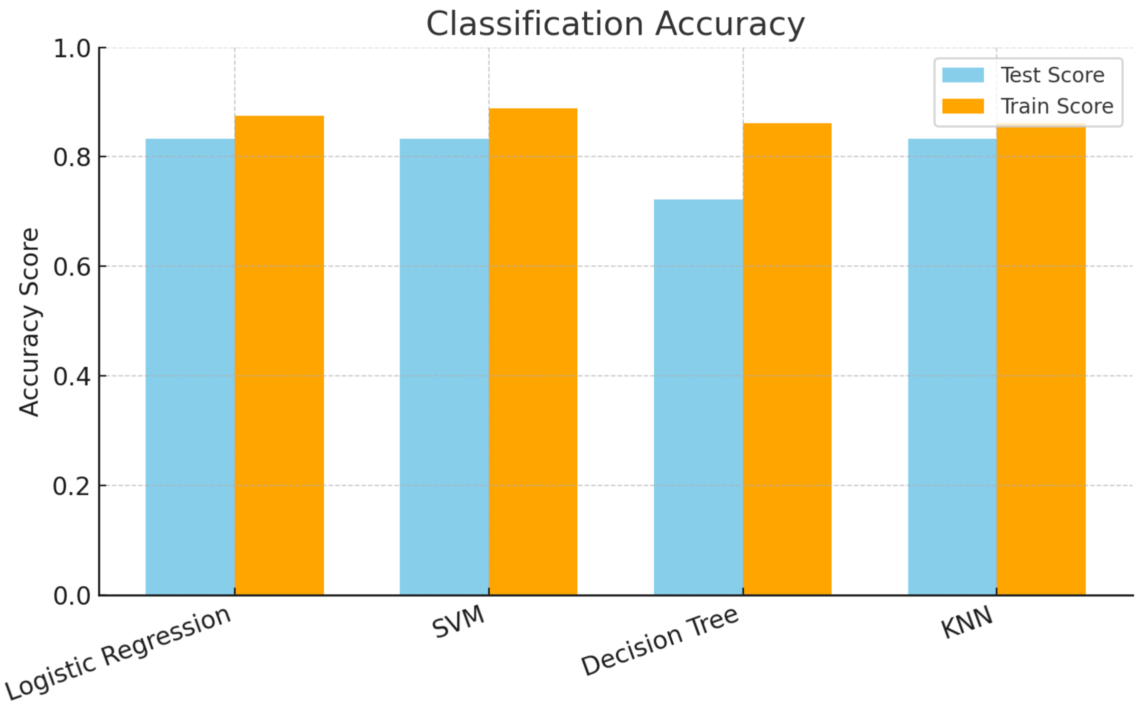
Classification Accuracy

- Below is a bar chart showing the classification accuracy for all models
- From the chart, it is observed that Logistic Regression, SVM, and KNN all achieved the highest test accuracy (0.833), while the Decision Tree scored lower (0.722).
- SVM is likely the best model overall, as it combines top test accuracy with the highest train score (0.889).

In [57]:

```python
# Create lists of the scores
scores = [logreg_score, svm_score, tree_score, knn_score]
scores_train = [logreg_score_train, svm_score_train, tree_score_train, knn_score_train]
model_names = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']

# Create a dictionary to hold the data
data = {
    'Model': model_names,
    'Test Score': scores,
    'Train Score': scores_train
}

# Create the pandas DataFrame
df = pd.DataFrame(data)

# Display the DataFrame
print(df)

# Find the index of the highest test score
best_model_index = df['Test Score'].idxmax()

# Get the row for the best model
best_model_row = df.loc[best_model_index]

# Display the best model
print("\nBest Model:")
print(best_model_row)


# Find and print the best model
#best_model_index = scores.index(max(scores))
#print(f"\nThe best score is from the '{model_names[best_model_index]}' model.")
```

```
                 Model  Test Score  Train Score
0  Logistic Regression    0.833333     0.875000
1                  SVM    0.833333     0.888889
2        Decision Tree    0.722222     0.861111
3                  KNN    0.833333     0.861111

Best Model:
Model          Logistic Regression
Test Score                0.833333
Train Score                  0.875
Name: 0, dtype: object
```



Confusion Matrix

Explanation:

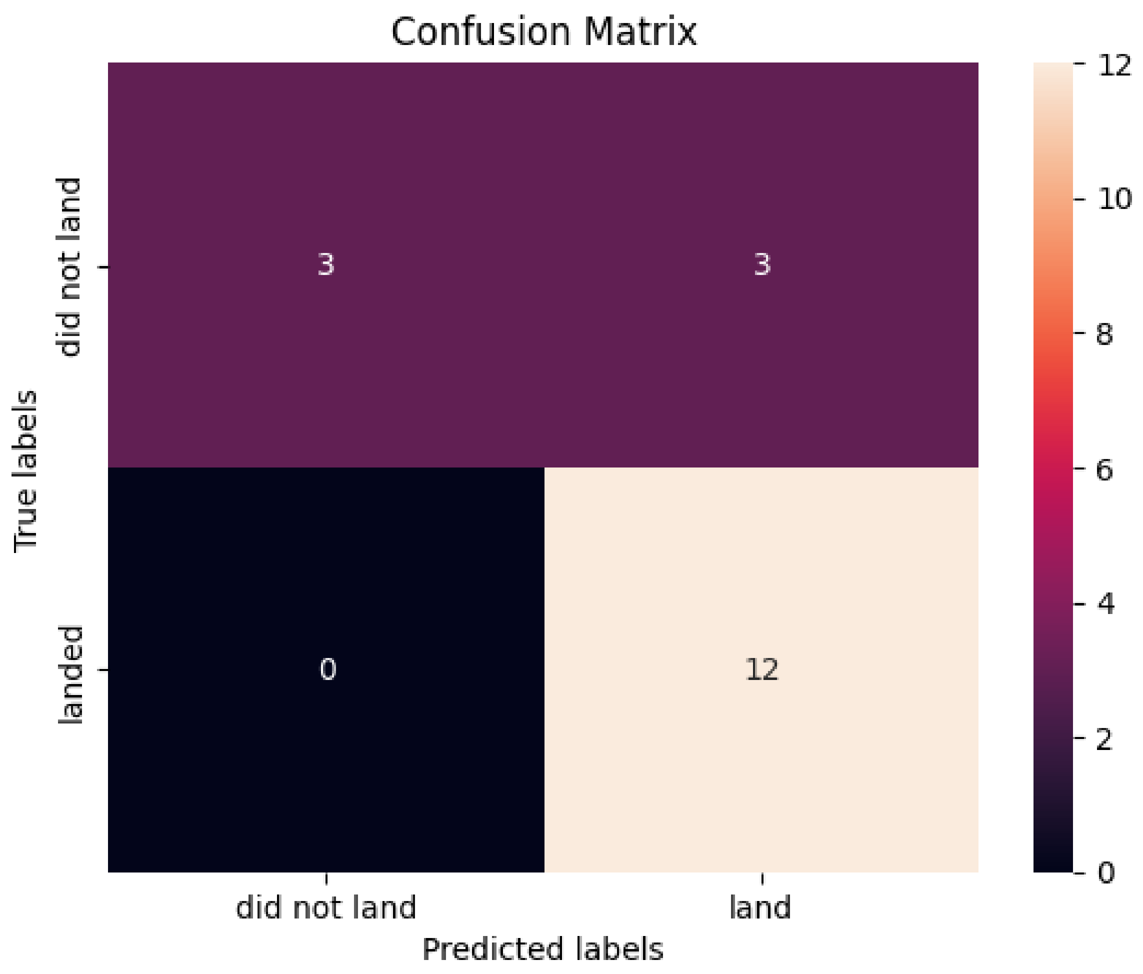
- The SVM model correctly classified most samples, with only 3 misclassifications.
- False positives (3 cases): the model predicted a successful landing when it was actually unsuccessful — this is a Type I error, which can be critical depending on the application.

Metrics:

- **Precision = TP / (TP + FP) =** 12 / (12 + 3) = 0.80 → 80% of predicted successes were correct
- **Recall = TP / (TP + FN) =** 12 / (12 + 0) = 1.00 → the model correctly identified all actual successes
- **F1 Score = 2 * (Precision * Recall) / (Precision + Recall) =** 2 × (0.8 × 1) / (0.8 + 1) = 0.89 → balances precision and recall
- **Accuracy = (TP + TN) / (TP + TN + FP + FN) =** (12 + 3) / (12 + 3 + 3 + 0) = 0.833 → 83.3% overall correct

Summary:

- The classifier can distinguish between successful and unsuccessful landings effectively.
- Main issue: false positives, which could overestimate successful landings.



Final Conclusion – SpaceX Launch Analysis

Key Insights:

- **Launch Sites:** KSC LC-39A is the most reliable (highest success rate and total successes); CCAFS SLC-40 and VAFB SLC-4E show higher failure rates. Sites are strategically located near coastlines and low latitudes for safety and efficiency.
- **Payload & Orbit:** Moderate payloads (~2–5k kg) have the highest success; heavier payloads reduce probability of successful landing. Best orbits for success: ES-L1, GEO, HEO, SSO.
- **Booster Performance & Trends:** FT booster is most reliable; early versions (v1.0, v1.1) had poor performance. Overall success has steadily increased since 2013, reflecting improved technology and operations.
- **Predictive Modeling:** SVM is the best model (Accuracy = 83.3%, F1 = 0.89), with launch site, payload mass, and orbit type as key predictors. False positives remain the main limitation.

Overall Conclusion:

- SpaceX landing success is strongly influenced by site, payload, orbit, and booster technology.
- Historical trends and predictive modeling confirm steady improvement in success rates and allow data-driven risk assessment and launch planning.
- Insights from this analysis provide a foundation for strategic decision-making in future launches.

Thank you!