IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

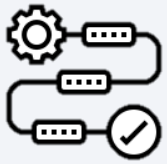
Alessandro Boldrin
1 August 2025

GitHub Repository:
https://github.com/Ragazzoatomico/testrepo

# Outline

Executive Summary

Introduction



Methodology

Results

Conclusion

Appendix

# Executive Summary

## Methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

## Results

- Exploratory Data Analysis Results
- Interactive Analytics in screenshots
- Predictive Analytics Results from Machine Learning Lab

# Introduction

SpaceX is a revolutionary company who has disrupted the space industry by offering a rocket launch, specifically Falcon 9, as low as 62 million dollars; while other providers cost upward of 165 million dollar each.

Most of this saving is due to SpaceX astounding idea to reuse the first stage of the launch by re-land the rocket to be used on the next mission. Repeating this process will make the price down even further.

As a data scientist of a startup rivaling SpaceX, the goal of this project is to create the machine learning pipeline to predict the landing outcome of the first stage in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch.

The problems included:

- Identifying all factors that influence the landing outcome
- The relationship between each variables and how it is affecting the outcome
- The best condition needed to increase the probability of successful landing

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX Open Rest API
  - Web Scraping from Wikipedia page “List of Falcon 9 and Falcon Heavy Launches”
- Perform data wrangling:
  - Transforming categorical data using One Hot Encoding for Machine Learning algorithms and removing any empty or unnecessary information from the dataset
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - Logistic Regression, Support Vector Machines, Decision Tree and K-Nearest Neighbors models have been developed to determine the most effective classification method

# Data Collection

## Definition:

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.

## What we did?

We collected data through REST API and web scraping from Wikipedia.
Using a GET request, we decoded the JSON response and transformed it into a pandas DataFrame with `json_normalize()`.
After cleaning and filling missing values, we used BeautifulSoup to scrape launch records, converting them into a pandas DataFrame for analysis.

# Data Collection – SpaceX API

**Request Launch Data**
Sent a GET request to the SpaceX API endpoint for launch data

**Decode Response**
Convert the JSON response from the API into a Python dictionary using `.json()`

**Normalize Data**
Use `.json_normalize()` to create a Pandas dataframe from the dictionary

**Request Additional Data**
Call SpaceX API endpoints for specific launch information using custom functions.

**Construct Dictionary**
Convert the dictionary containing all launch details into a dataframe.

**Create Dataframe**
Convert the dictionary containing all launch details into a dataframe.

**Filter Dataframe**
Filter the dataframe to include only Falcon 9 launches.

**Clean Data**
Impute missing values in the "Payload Mass" column using the mean value.

**Export Data**
Save the cleaned dataframe as a CSV file for further analysis.

**GitHub Link:**
Data Collection API

# Data Collection - Scraping

**Request Data**
Send a GET request to the Wikipedia page containing Falcon 9 launch data.

**Parse Response**
Use Beautiful Soup to create an object representing the HTML structure of the webpage.

**Extract Headers**
Identify and extract all column names (headers) from the HTML table containing launch data.

**Parse Table Data**
Iterate through the HTML table rows and columns, collecting launch information.

**Construct Dictionary**
Organize the collected launch data into a Python dictionary with headers as keys.

**Create Dataframe**
Convert the dictionary containing all launch details into a dataframe

**Export Data**
Save the cleaned dataframe as a CSV file for further analysis.

**GitHub Link:**
Data Collection Scarping

# Data Wrangling

## Definition:

Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA).

## What we did?

1. Calculate the number of launches on each site
2. Determine mission outcomes per orbit types
3. Create landing outcomes labels
4. Export results to CSV

# EDA with Data Visualization

**Scatter Plot: Flight Number VS Launch Site**
Identify any patterns between the number of flights (Flight Number) and the location from where they took off (Launch Site)

**Scatter Plot: Payload VS Launch Site**
Explore any potential relationships between the weight of the cargo carried (Payload) and the launch location (Launch Site)

**Line Plot: Launch Success Yearly Trend**
Identify trends in launch success rates over time (yearly). It reveals whether the success rate is improving, declining, or staying consistent across years.

**GitHub Link:**
EDA Data Visualization

**Scatter Plot: Flight Number VS Orbit Type**
Uncover any patterns between the launch order (Flight Number) and the type of orbit the mission aimed to achieve (Orbit Type)

**Scatter Plot: Payload VS Orbit Type**
Explores any potential relationships between the cargo weight (Payload) and the type of orbit targeted by the launch (Orbit Type)

**Bar Chart: Success Rate by Orbit Type**
Visualize the success rate of missions for each specific orbit type

# EDA with SQL

## Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes, such as Failure (drone ship) or Success (ground pad), between the date 2010-06-04 and 2017-03-20 in descending order

**GitHub Link:**
EDA SQL

# Build an Interactive Map with Folium

## Launch Sites

Markers with circles highlight all launch site locations.
Text labels provide additional information like "NASA Johnson Space Center." These markers showcase the geographical distribution of launch sites and their proximity to the equator and coastlines.

## Launch Outcomes

Green and red markers represent successful and failed launches, respectively. Marker clusters group these markers, allowing viewers to identify launch sites with high success rates immediately.

## Distances

Lines connect a specific launch site to nearby features like railways, highways, coastlines, and the closest city.
These lines illustrate the proximity of launch sites to essential infrastructure and population centers.

**GitHub Link:**
[Interactive Map with Folium]

# Build a Dashboard with Plotly Dash

## Dropdown List

This dropdown list allows users to filter data by a specific launch site. By selecting a site, the dashboard focuses on that location's launch information.

## Pie Chart

This pie chart displays the total number of successful launches across all sites. When a specific site is chosen from the dropdown, the pie chart dynamically updates to show the success and failure rates for that specific location, providing a focused view.

## Slider

This slider enables users to select a specific range of payload mass. This allows for focused analysis on how launch success rates correlate with the weight of the cargo carried.

## Scatter Plot

This scatter plot visualizes the relationship between payload mass (weight) and launch success rate for different booster versions. By filtering the data with the slider and potentially the dropdown, users can explore these correlations for specific launch sites or booster types.

**GitHub Link:**
Dashboard with Plotly Dash

# Predictive Analysis (Classification)

**Data Preparation**

- Select relevant features
- Standardize data for improved model performance (`StandardScaler`)

**Train and Test Split**

Divide the data into training and testing sets (`train_test_split`) for model training and evaluation.

**Model Development**

Apply `GridSearchCV` with cross-validation to various models (Logistic Regression, SVM, Decision Tree, KNN) in order to identify the optimal hyperparameter settings for each model.

**Model Evaluation**

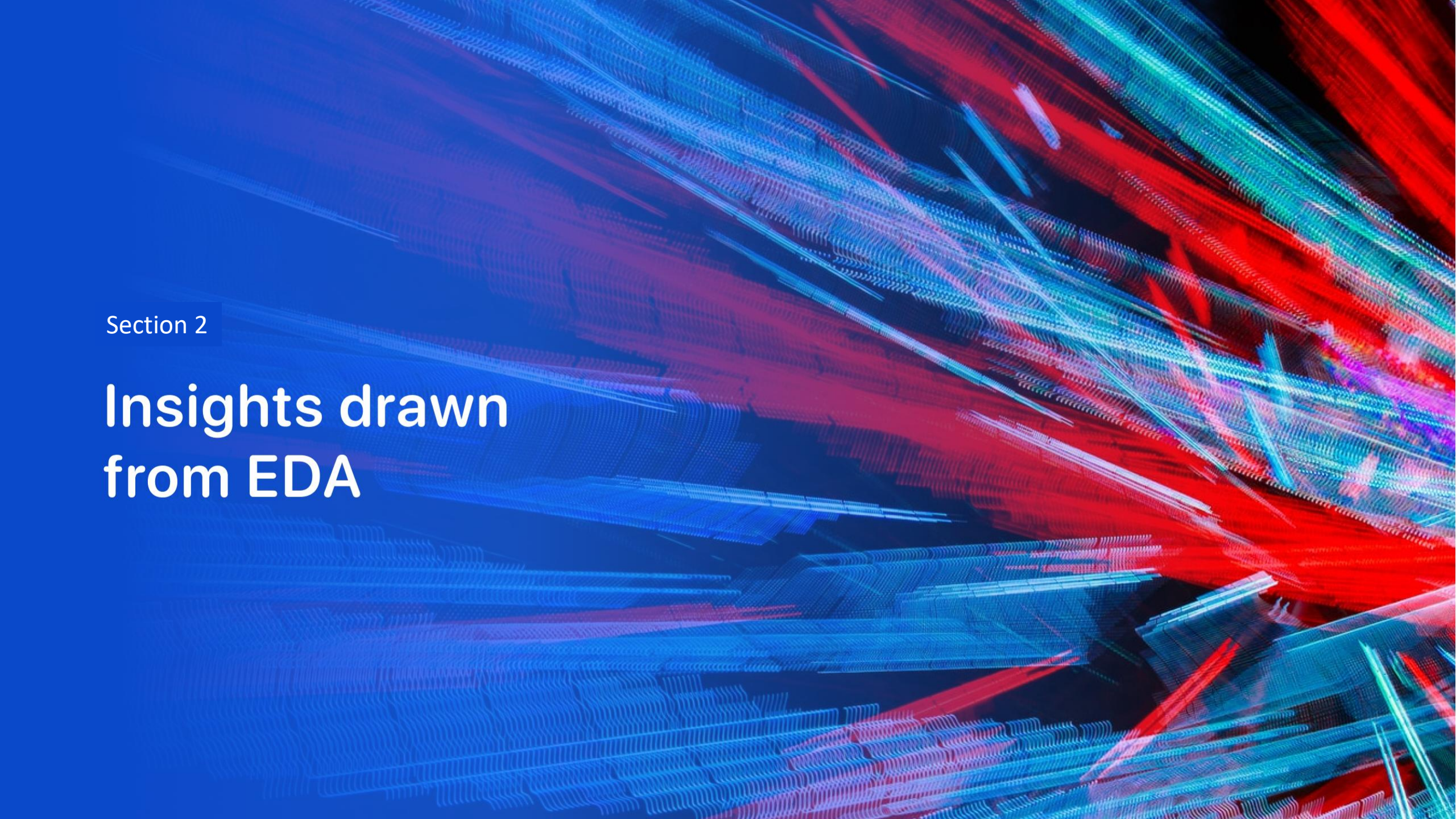
- Evaluate each model using accuracy (`.score()`) and confusion matrices to assess performance on unseen data
- Select the model with the highest overall performance based on the combined evaluation metrics

**GitHub Link:**
Predictive Analysis

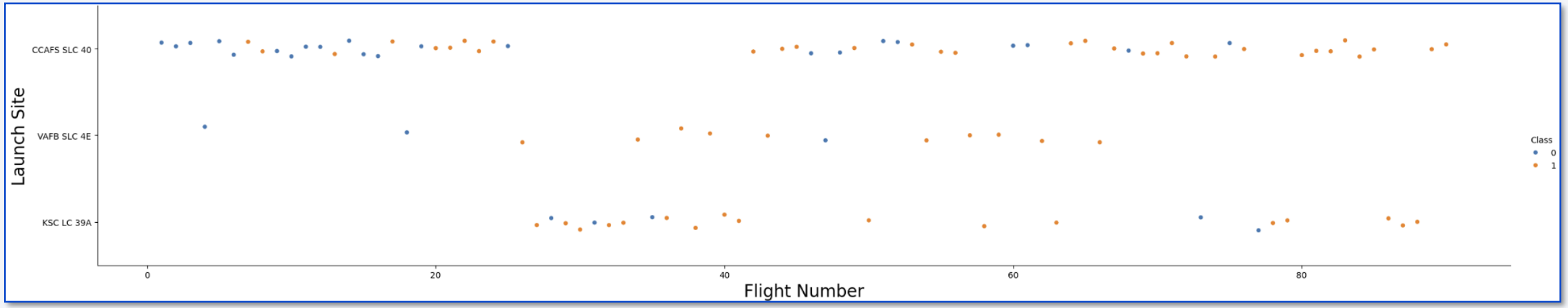
# Results

The results will be categorized to 3 main results which are:

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


Section 2
Insights drawn from EDA

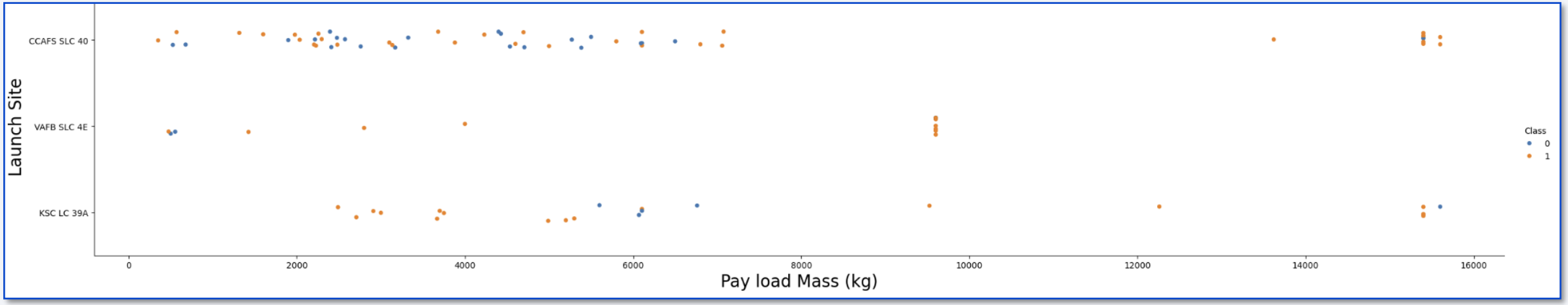
# Flight Number vs. Launch Site




## Explanation:

This scatter plot shows that CCAFS SLC 40 launches the most, while VAFB SLC 4E and KSC LC 39A have higher success (i.e., Class = 1).

# Payload vs. Launch Site




## Explanation:

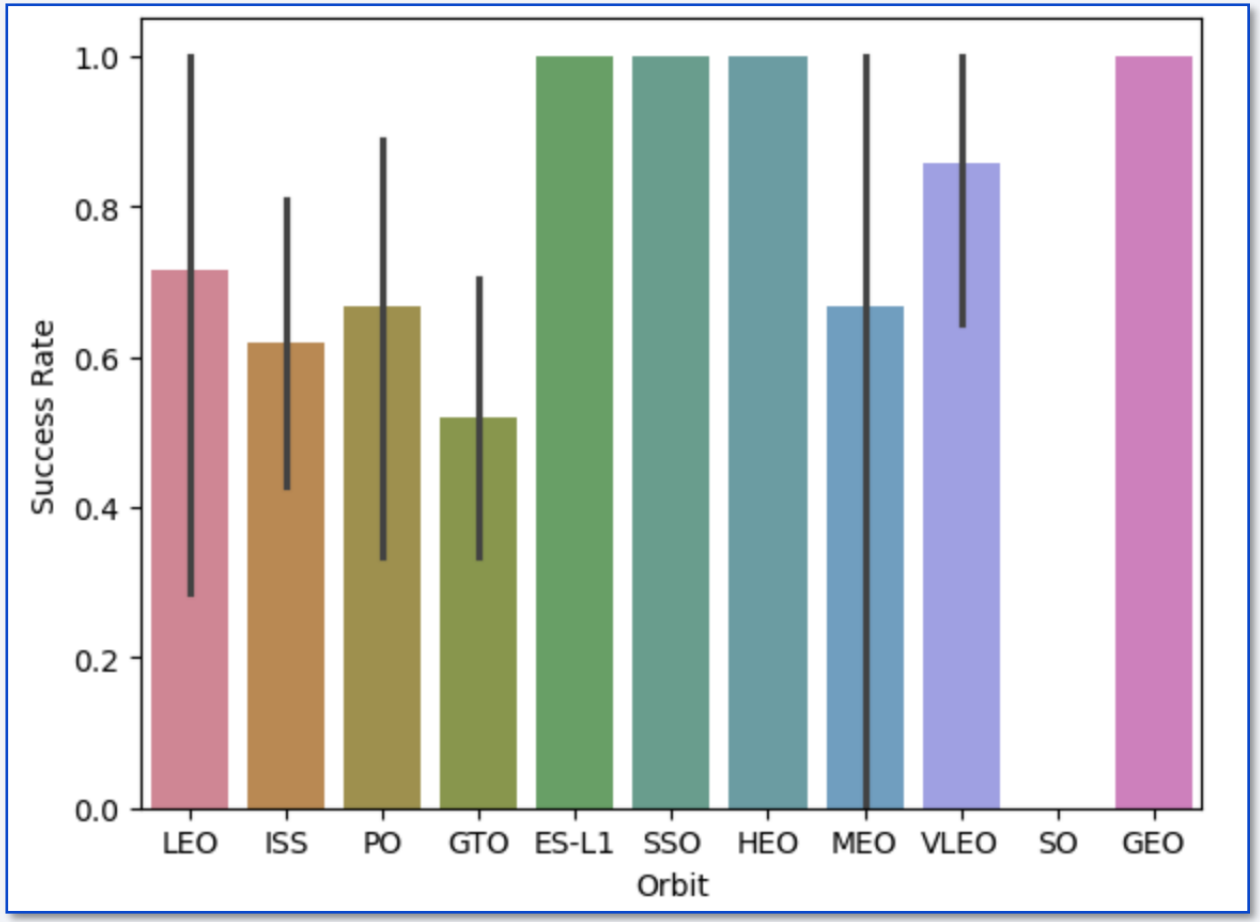
This scatter plot shows that heavier payloads masses see higher successes across sites (i.e., Class = 1).
A particular case, indeed, is the one of KSC LC 39A, because it excels also for lighter payloads masses.

# Success Rate vs. Orbit Type

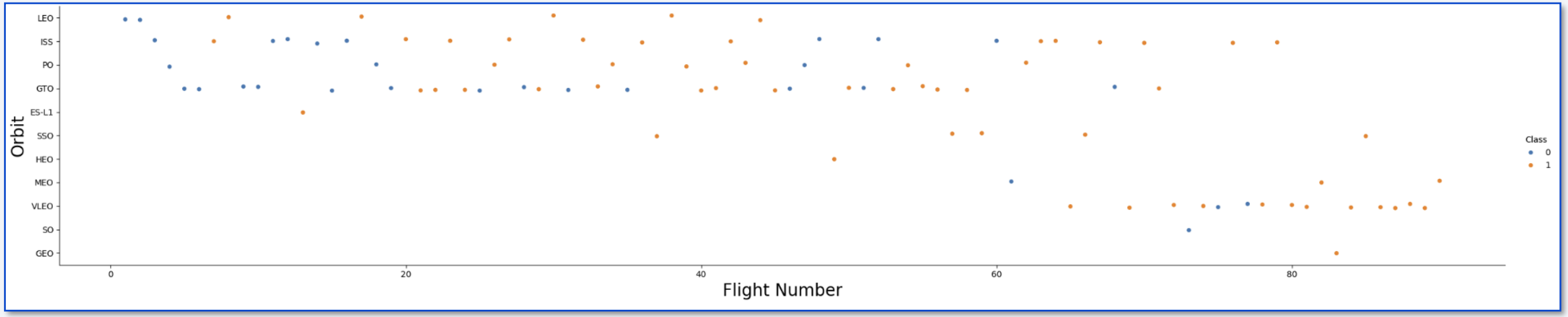
## Explanation:

This bar char shows the possibility of the orbits to influences the landing outcomes, for example:

- some orbits has 100% success rate such as SSO, HEO, GEO and ES-L1
- while SO orbit produced 0% rate of success



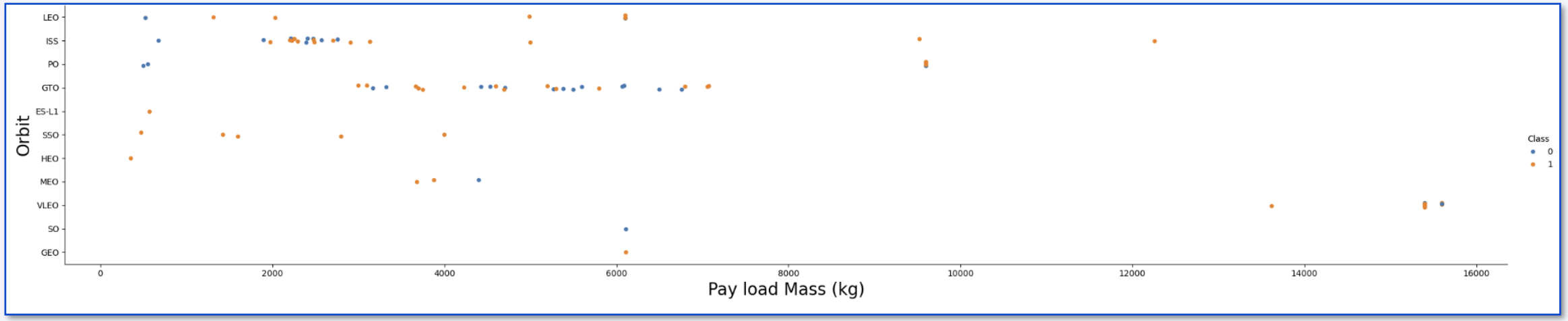
# Flight Number vs. Orbit Type




## Explanation:

This scatter plot shows that, generally, the larger are the flight numbers on each orbits, the greater the success rate (i.e., Class = 1), except for GTO orbit which it seems to not have any relationship between both attributes.

# Payload vs. Orbit Type



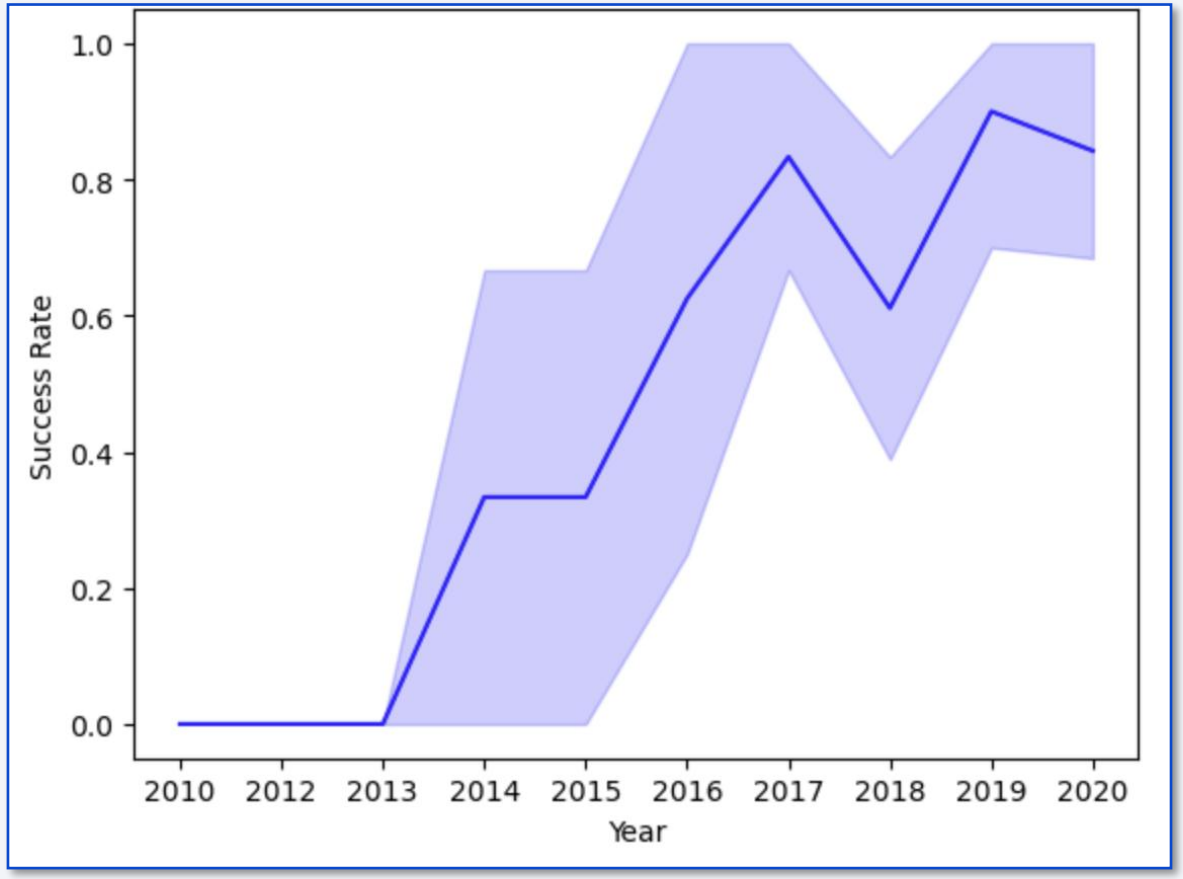

## Explanation:

This scatter plot shows that heavier payloads have a more successful impact (i.e., Class = 1) on LEO, ISS and PO orbits. On the contrary, lighter payloads have a more successful MEO and VLEO orbits.
Regard GTO orbit, indeed, it seems to not have any relation between the two attributes.

# Launch Success Yearly Trend

## Explanation:

According to the line plot, the success rate kept increasing since 2013 to 2020



# All Launch Site Names

## Explanation:

The key word DISTINCT shows only unique launch sites from the SpaceX data, which are CCSFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

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

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

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

# Launch Site Names Begin with 'CCA'

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

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

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

## Explanation:

The query displays the (first) 5 records of launches done from the sites whose name begin with ‘CCA’

# Total Payload Mass

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

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

| total_payload_kg |
|---|
| 45596 |

## Explanation:

The query displays the total payload mass carried by the NASA (CRS) which is 45596 kg.

# Average Payload Mass by F9 v1.1

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

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

| avg_payload_kg |
| --- |
| 2928.4 |

## Explanation:

The query displays the average payload carried by the Booster Version F9 v1.1 which is 2928.4 kg.

# First Successful Ground Landing Date

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

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

| first_success_ground_pad |
| --- |
| 2015-12-22 |

## Explanation:

The query displays the first successful ground landing date which was 22 December 2015.

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)'AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

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

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

## Explanation:

The query displays the Booster Versions which were successful in landing with payload between 4000 and 6000 kg.

# Total Number of Successful and Failure Mission Outcomes

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

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

| Mission_Outcome | count |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

## Explanation:

The query displays the groups and displays total number of mission outcomes be it Success or Failure.

# Boosters Carried Maximum Payload

## Explanation:

The query shows a list of Booster Versions carrying maximum payload.

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

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

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

# 2015 Launch Records

```
%sql SELECT SUBSTR("Date", 6, 2) AS Month, Date, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR("Date", 1, 4) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)';
 * sqlite:///my_data1.db
Done.
```

| Month | Date | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|---|
| 01 | 2015-01-10 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | 2015-04-14 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

## Explanation:

The query displays a list of Booster Versions and Launch Site for all launches in 2015 (January and April) where the Landing Outcome was failure on drone ship.

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

```
%sql SELECT "Landing_Outcome", outcome_count, RANK() OVER (ORDER BY outcome_count DESC) AS rank FROM (SELECT "Landing_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome");
```

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

| Landing_Outcome | outcome_count | rank |
|---|---|---|
| No attempt | 10 | 1 |
| Failure (drone ship) | 5 | 2 |
| Success (drone ship) | 5 | 2 |
| Controlled (ocean) | 3 | 4 |
| Success (ground pad) | 3 | 4 |
| Failure (parachute) | 2 | 6 |
| Uncontrolled (ocean) | 2 | 6 |
| Precluded (drone ship) | 1 | 8 |

## Explanation:

The query ranks landing outcome between specific dates and groups them by landing outcome with their counts.

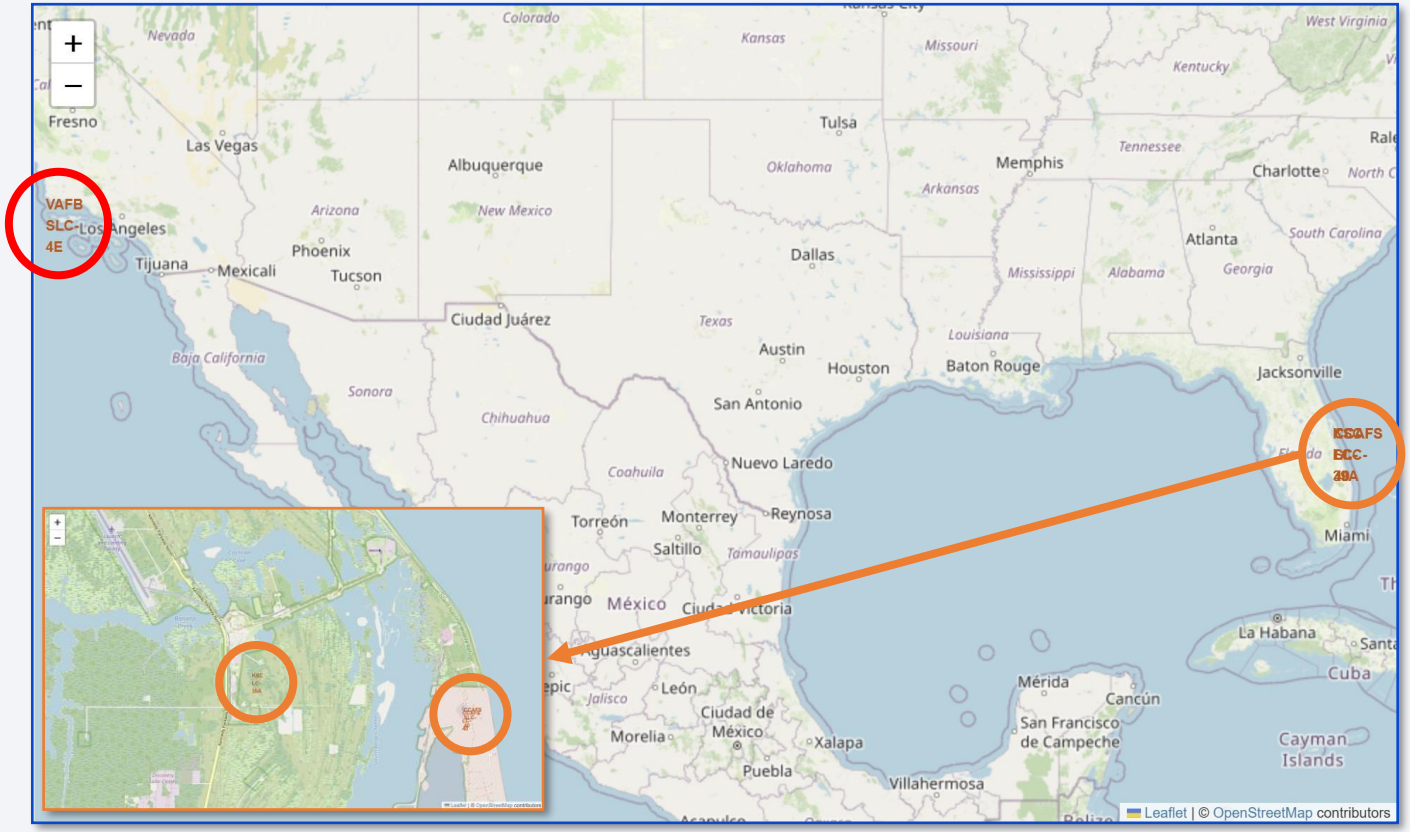
Section 3

# Launch Sites Proximities Analysis

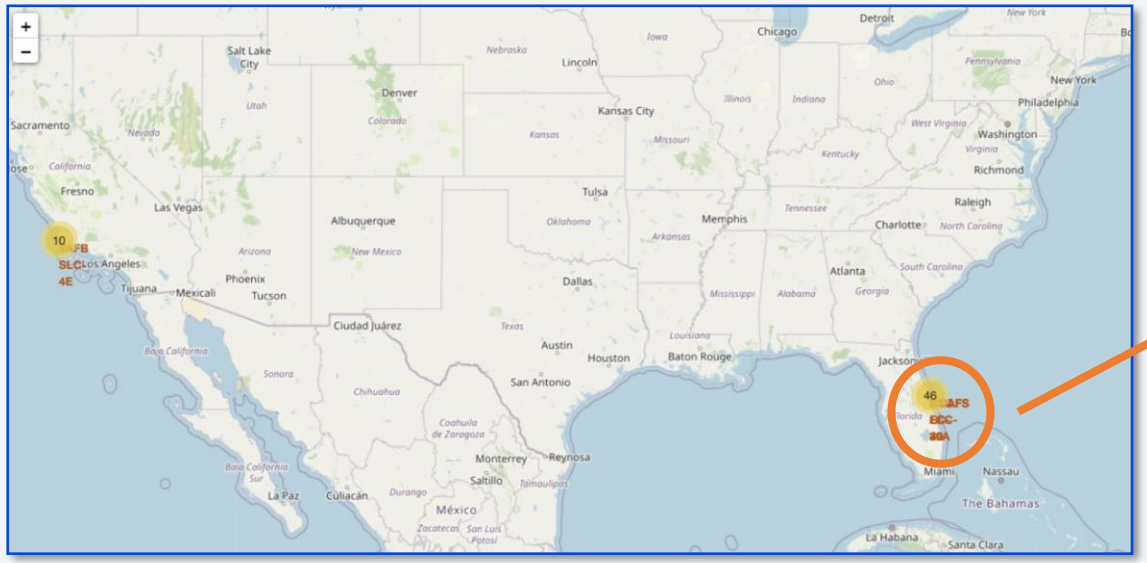
# Launch Sites' Locations On A Global Map

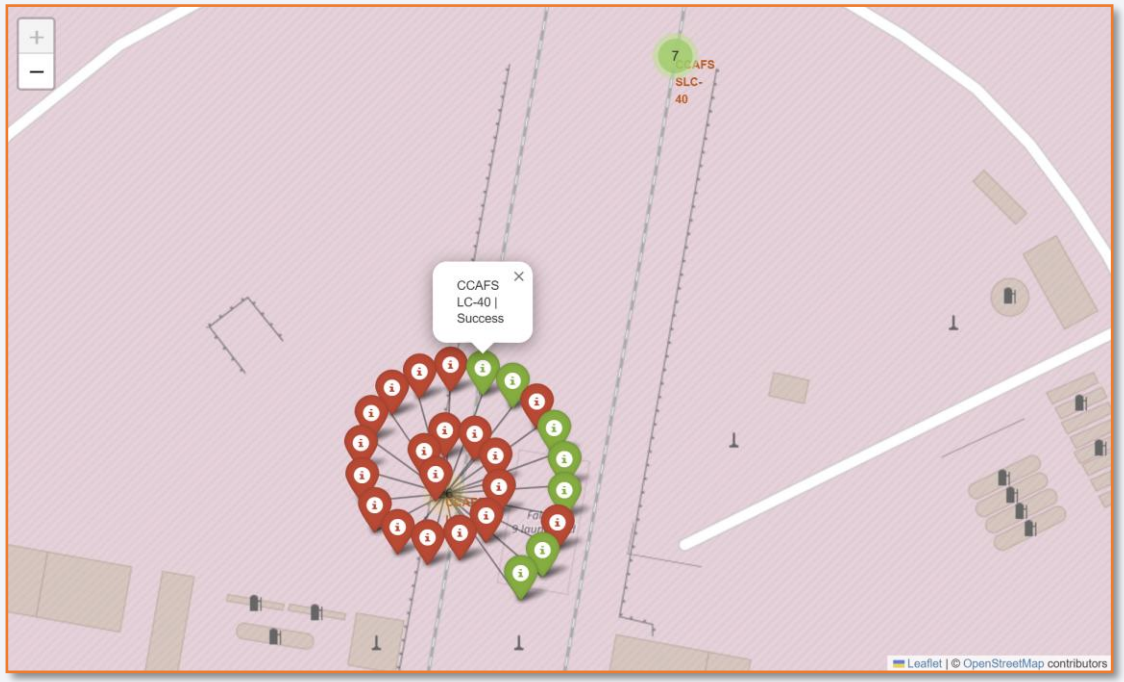
## Explanation:

All the SpaceX launch sites are located inside the United States, especially on the coasts of Los Angeles (VAFB SLC 4E) and of Florida (KSC LC 39A, CCAFS LC 40 and CCAFS SLC 40)



# Successful/Failed Launches Markers on the Map



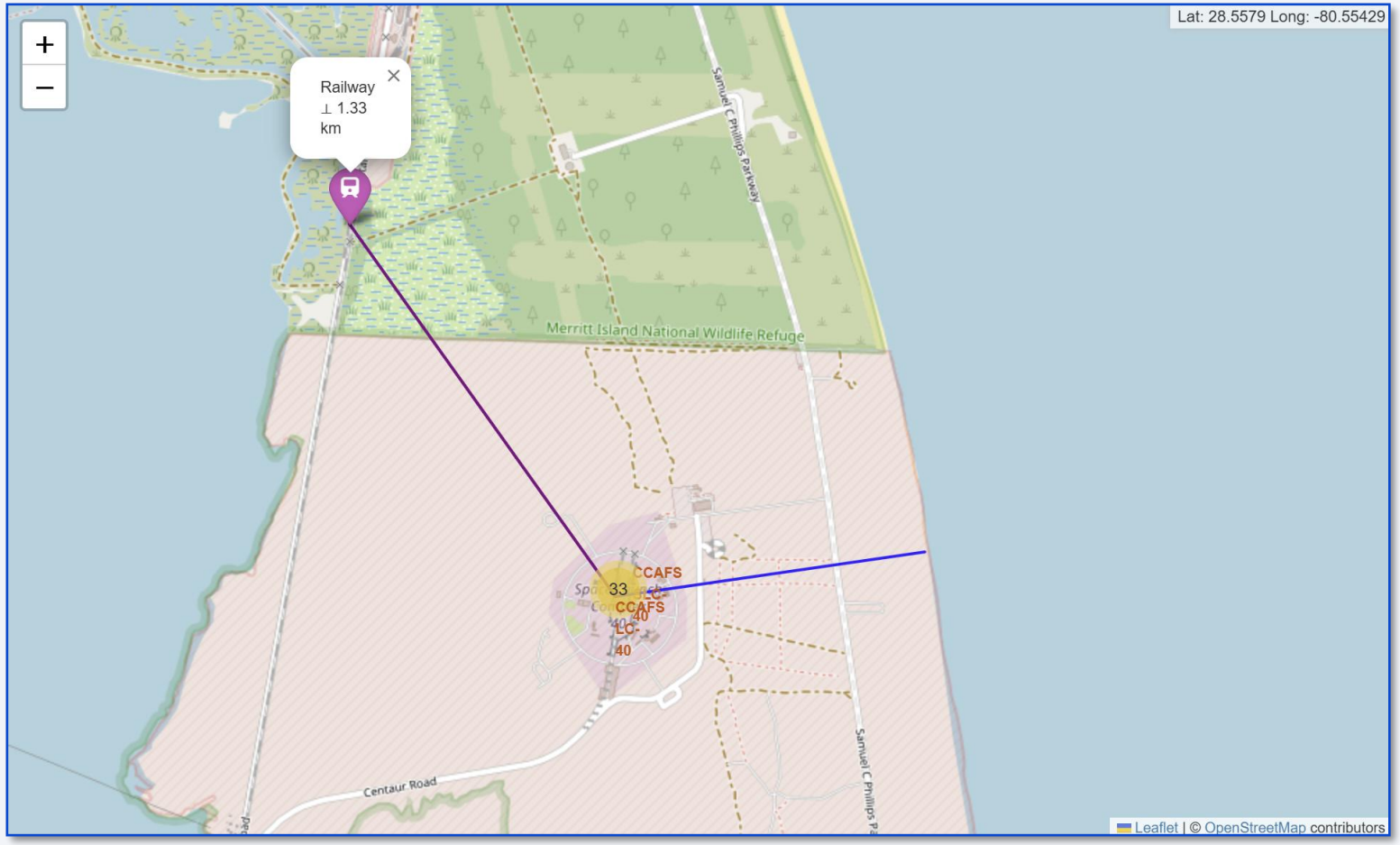

## Explanation:

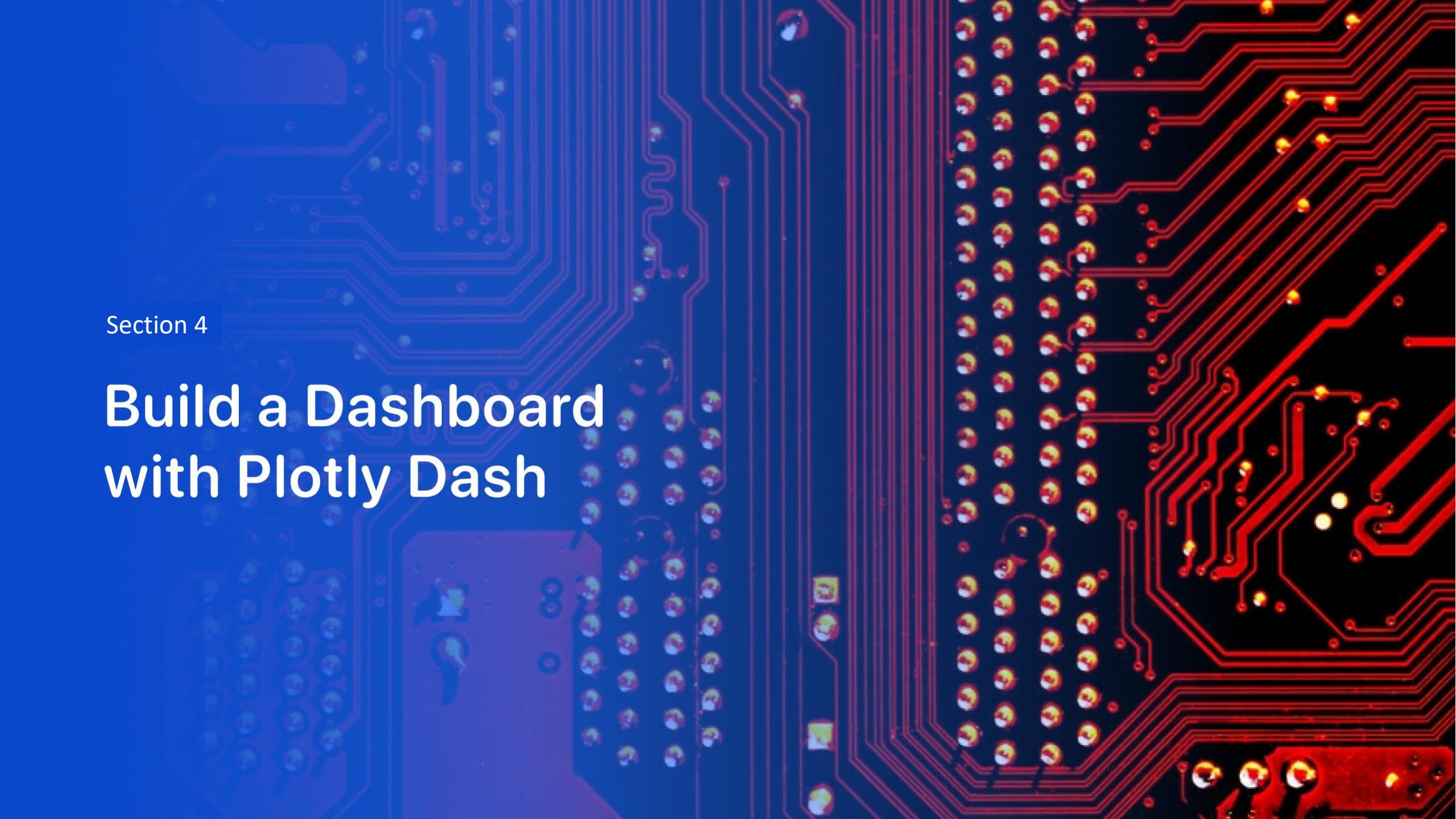
From the color-labeled markers in the various marker clusters, we can easily identify which launch site has relatively high success or failure rates (i.e., launch site's outcomes of CCAFS LC 40).

# Launch Sites Distance to Landmarks

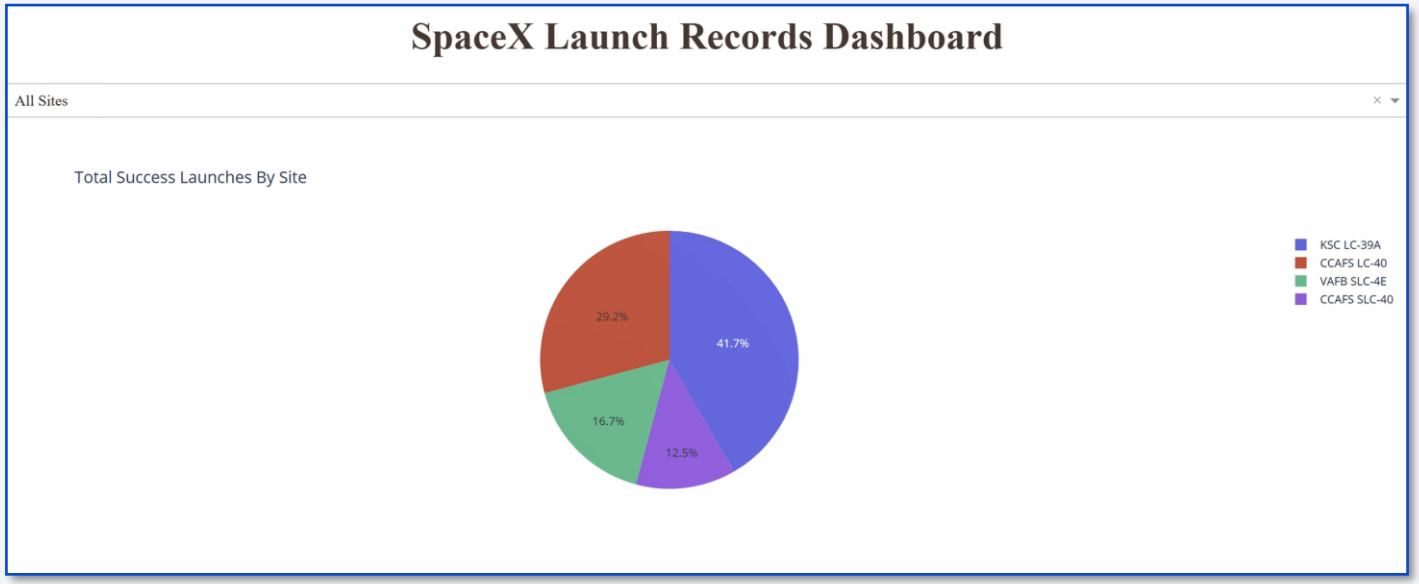
## Explanation:

Example of a distance between the launch site CCAFS LC 40 and two landmarks (i.e., a railway far 1.33 km from the launch site; and a the coast)




Section 4
Build a Dashboard with Plotly Dash

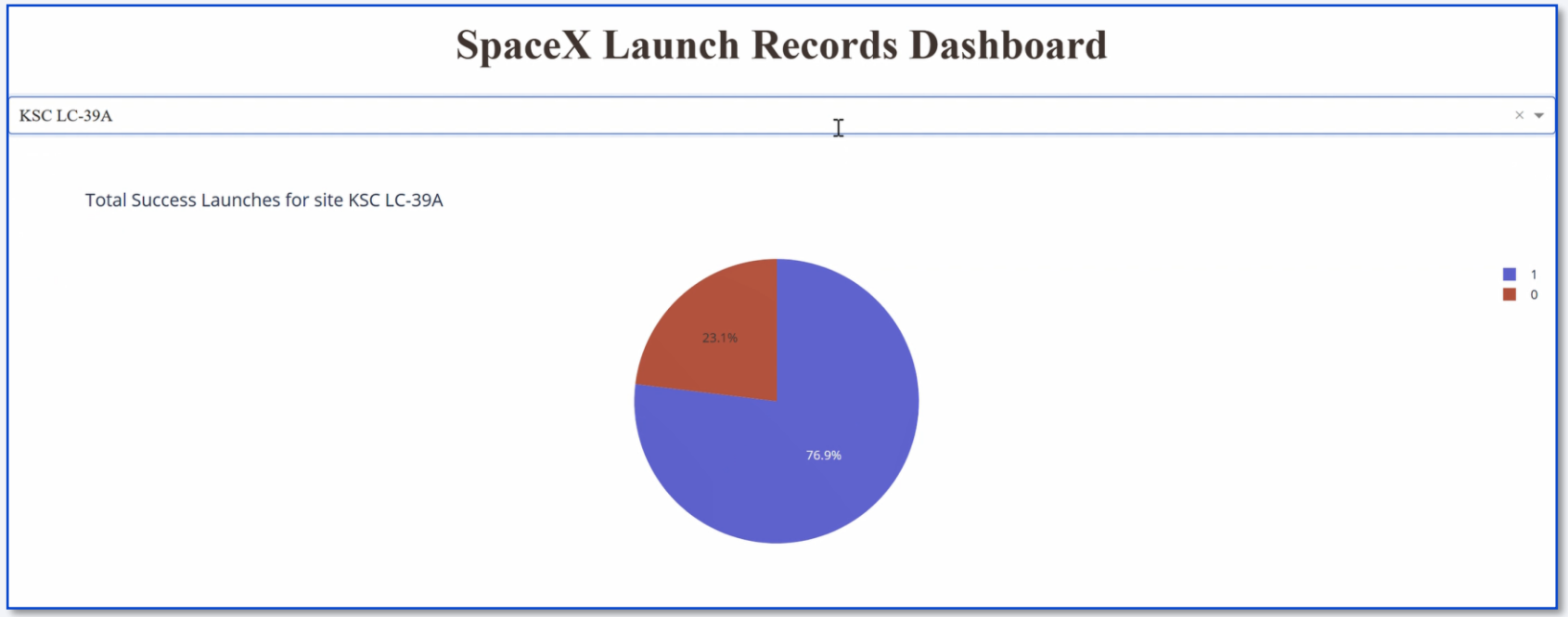
# Total Success Launches By Site




## Explanation:

According to the pie chart the KSC LC-39A contributes heavily for the greatest number of successful launches among other sites for about 41.7%, above CCAFS LC-40 which contribute around 29.2%. The rest number of successful launches goes from VAFB SLC-4E at 16.7% to the least with CCAFS SLC-40, who has the lowest 12.5% successful launches.

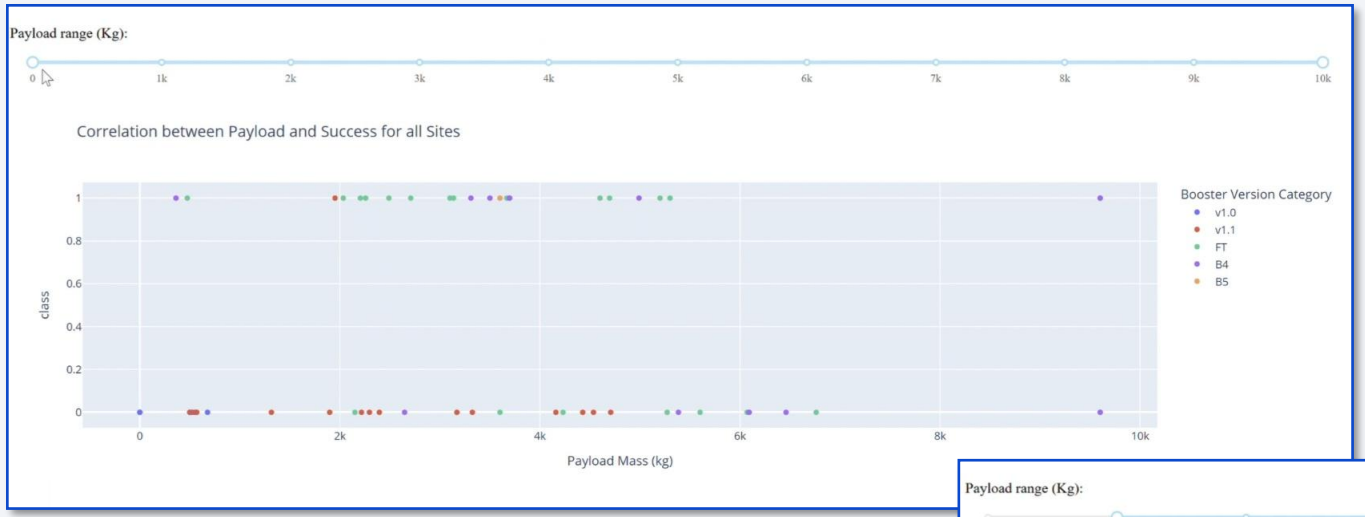
# Total Success Launches for KSC LC 39A




## Explanation:

According to the pie chart, the KSC LC-39A, which is the launch site with the highest amount of successful launches, has a 76.9% of successes against 23.1% of failures.

# Payload vs. Launch Outcome For Different Payloads




## Explanation:

The scatter plot on the left shows launch outcomes for payload mass between 0 to 10000 kg for various Booster Version Category

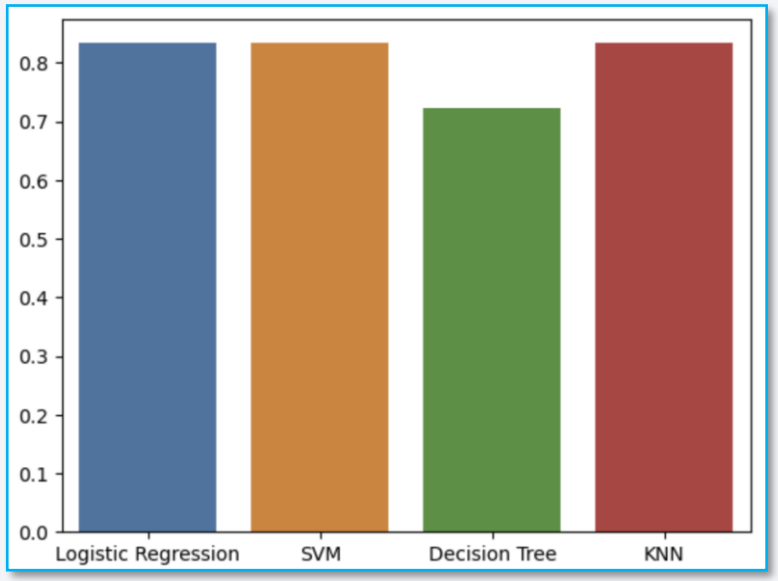

## Explanation:

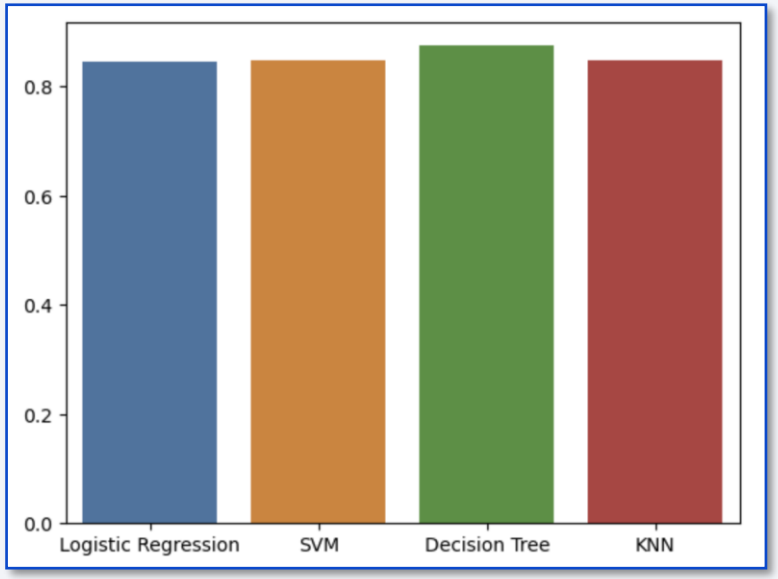
The scatter plot on the right shows launch outcomes for payload mass between 1000 to 8000 kg for various Booster Version Category

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy



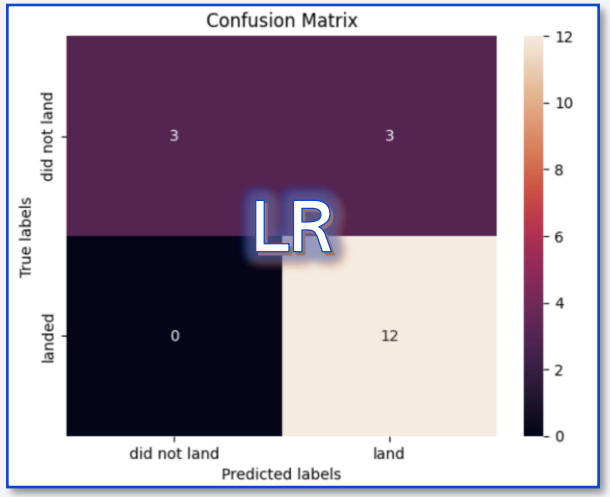

| Test Set Accuracy | | | |
|---|---|---|---|
| **Logistic Regression** | **SVM** | **Decision Tree** | **KNN** |
| 0.833333 | 0.833333 | 0.722222 | 0.833333 |

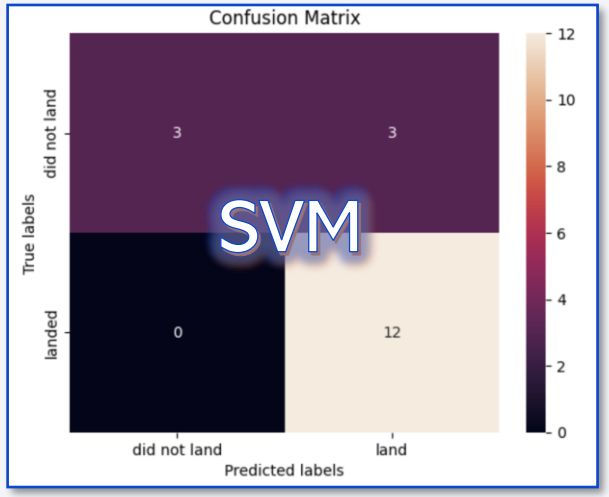
| Entire Data Set Accuracy | | | |
|---|---|---|---|
| **Logistic Regression** | **SVM** | **Decision Tree** | **KNN** |
| 0.846429 | 0.848214 | 0.875 | 0.848214 |

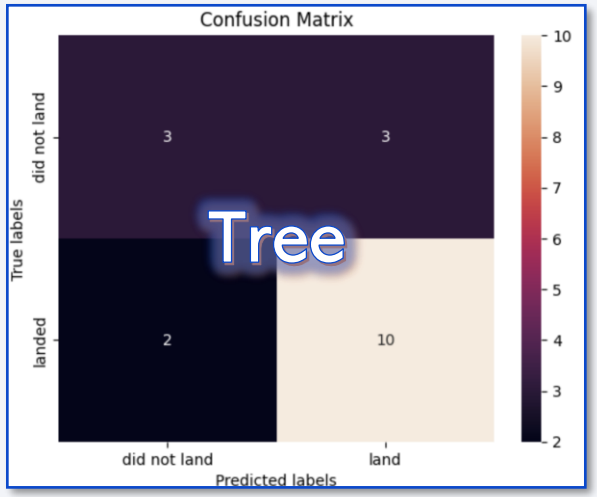
## Explanation:

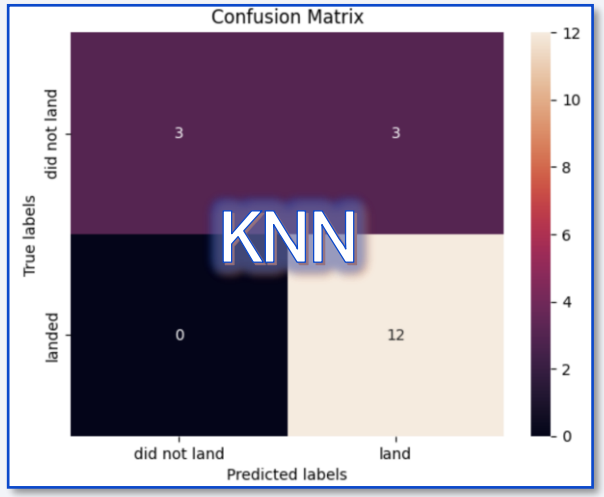
According to the bar charts above, the Decision Tree model has the lowest accuracy on the test dataset, while it has the highest on the entire data set. Regard the other models, they gave approximately the same results on both sets.

# Confusion Matrix



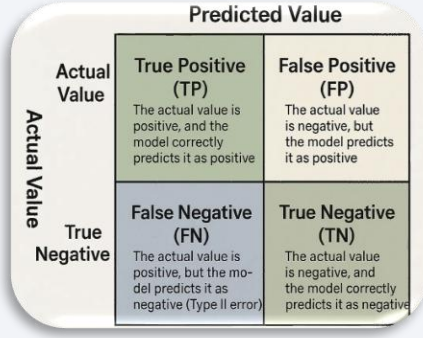




## Explanation:

- LR, SVM, KNN models are the bests as their confusion matrix show that they predicted all 12 successful landing correctly, with 0 error.
- However, the Decision Tree model only predicted 10 successful landing correctly, with 2 of them wrongly predicted as a failed/did not land.
- LR, SVM, KNN models have the same accuracy of 83.33% as displayed earlier, hence the same confusion matrix.



# Conclusions

- LR, SVM, KNN are top-performing models for forecasting outcomes in this data
- Lighter payloads have a higher performance compared to heavier ones.
- The likelihood of a SpaceX launch succeeding increases with the number of years of experience, suggesting a trend towards flawless launches over time
- The KSC LC 39A has the highest number of successful launches compared to other launch sites
- SSO, HEO, GEO and ES-L1 orbits types exhibit the highest rates of successful launches

# Appendix

All external links to Python and Jupiter Notebooks code snippets, SQL queries, plots, or data sets used in the project is provided in my GitHub repository.

Special thanks to all the Instructors, Coursera and IBM.

# Thank you!