IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Truman Tran
3/24/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Acquiring data via APIs & web scraping techniques
  - Cleaning and transforming data i.e. data wrangling
  - Analyzing data using SQL queries & visualizations
  - Making predictions using machine learning models
- Summary of all results
  - Valuable data can be gathered from publicly available sources.
  - Exploratory Data Analysis (EDA) can help determine the most relevant features for predicting launch success.
  - Machine learning can identify the most effective model for predicting key factors for success, leveraging collected data to optimize outcomes.

Introduction

- Project background and context
  - Predicting whether the first stage will land helps estimate launch costs. This project aims to build a machine learning model to predict successful landings,
- Problems you want to find answers
  - How different features interact to influence the likelihood of a successful landing.
  - The necessary operational conditions required to achieve a reliable landing program.

Section 1

Methodology

Methodology

Executive Summary

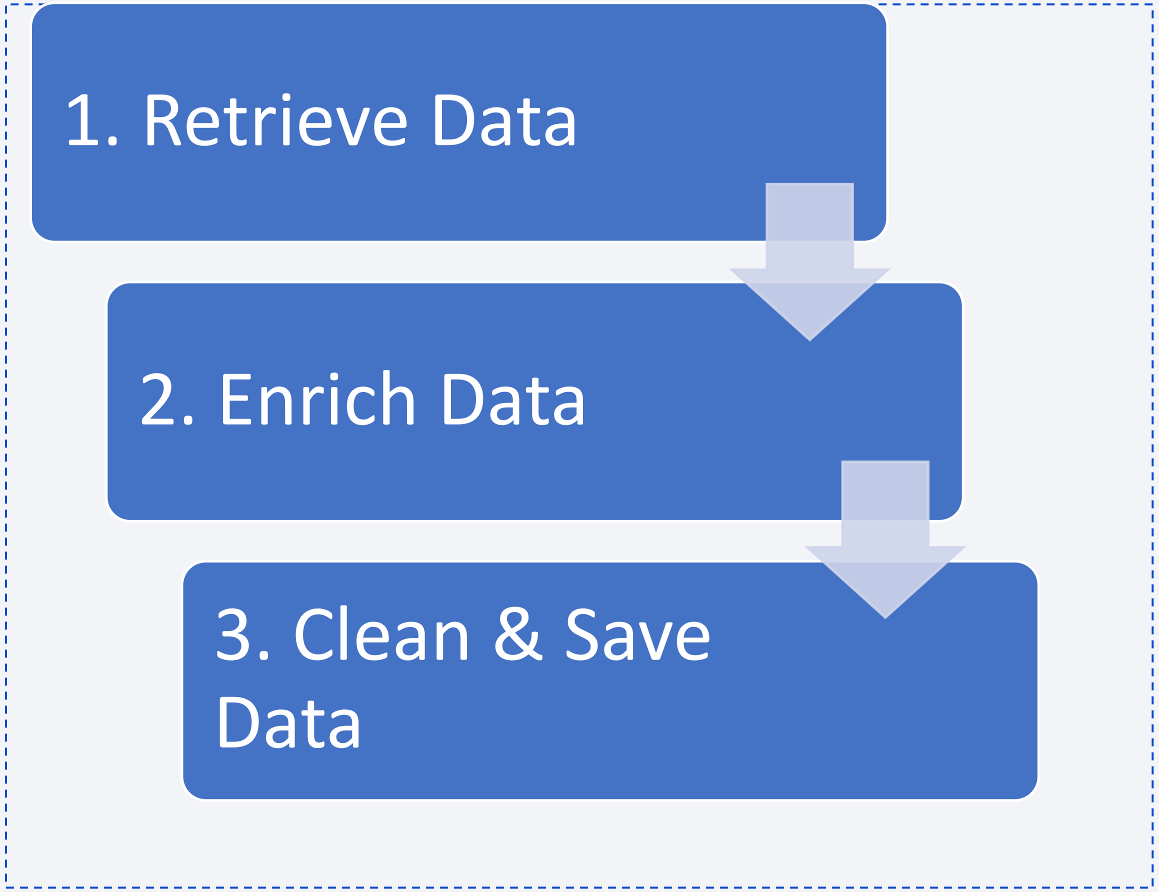
- Data collection methodology:
  - Using SpaceX API & perform web scraping on Wikipedia
- Perform data wrangling
  - Examine and transform raw data into a structured format suitable for analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data was normalized and split into training and testing sets. Four classification models were applied, and their accuracy was assessed using various parameter combinations.

Data Collection

- How data sets were collected.
  - Data was gathered through a GET request to the SpaceX API.
  - The response was parsed as JSON using the .json() function and transformed into a pandas dataframe with .json_normalize().
  - The dataset was then cleaned, and missing values were identified and filled where needed.
  - Additional data was obtained via web scraping from Wikipedia using BeautifulSoup, focusing on Falcon 9 launch records.
  - HTML tables containing launch data were extracted, parsed, and converted into pandas dataframes for further analysis.

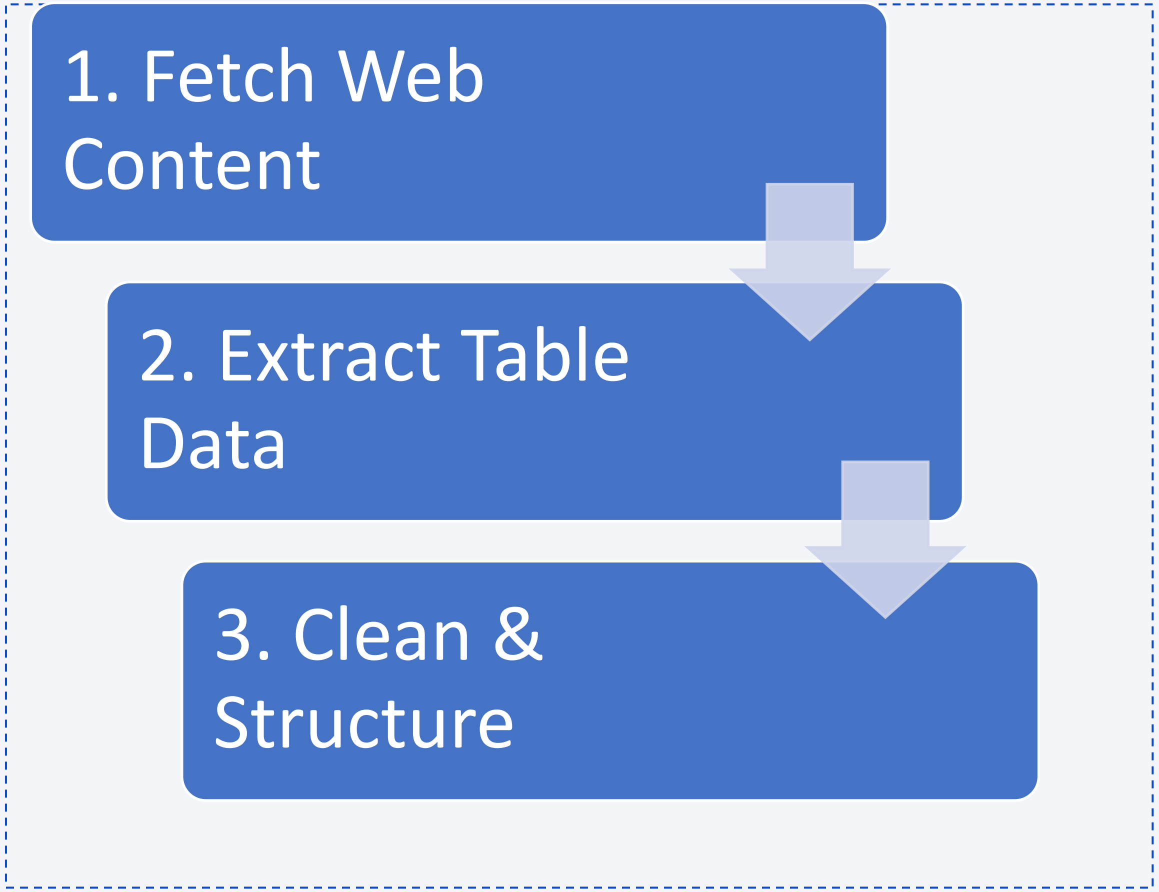
Data Collection – SpaceX API

- 1. Use SpaceX API to fetch launch data. Normalize the returned JSON into a flat table.
- 2. Use helper functions to extract rocket, launchpad, payload, and core details via additional API calls.
- 3. Filter for Falcon 9 launches, handle missing values, and export the final dataset to CSV.
- https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/jupyter-labs-spacex-data-collection-api(1).ipynb



Data Collection - Scraping

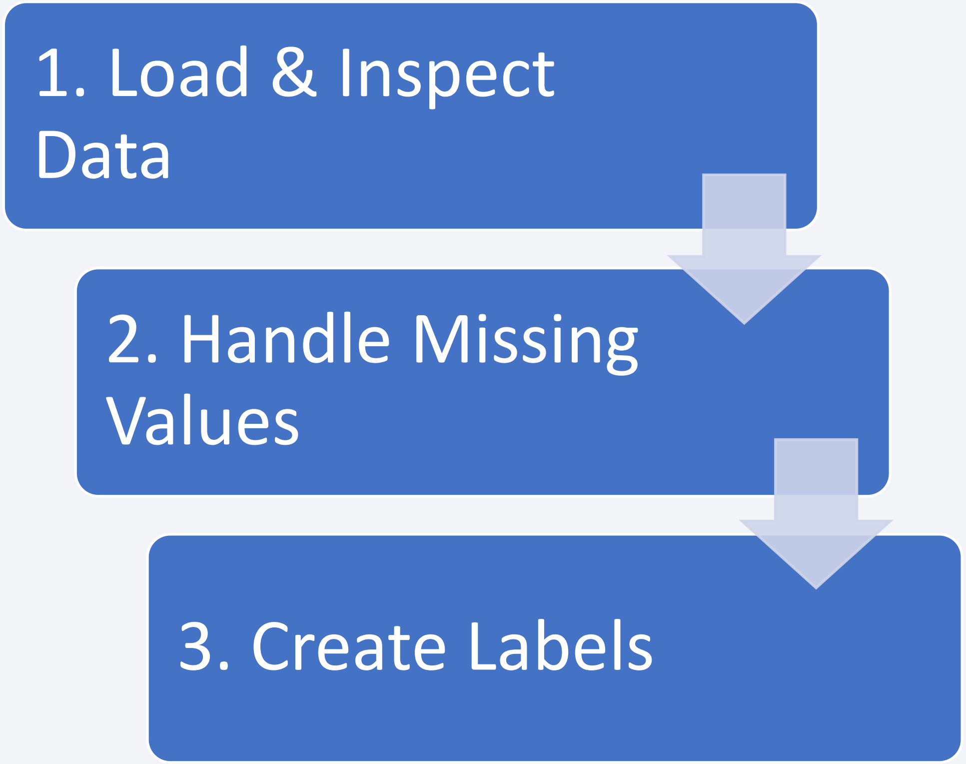
- 1. Send a request to Wikipedia page & parse it using BeautifulSoup.
- 2. Locate launch tables, loop through rows, and extract key fields (e.g., date, payload, booster, outcome).
- 3. Process raw strings, apply helper functions, and store the data into a structured dictionary or DataFrame.
- https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/jupyter-labs-webscraping(1).ipynb



Data Wrangling

1. **Load & Inspect Data**: Import the SpaceX dataset & explore structure and content.
2. **Handle Missing Values**: Identify & calculate the percentage of missing values per column.
3. **Create Labels**: Transform the Outcome column into binary landing success labels (1 for success, 0 for failure).

- https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Catplot: PayloadMass vs FlightNumber – Shows how payload varies by flight and its relation to mission success.
- Catplot: LaunchSite vs FlightNumber – Examines if launch site affects mission outcome.
- Catplot: Orbit vs PayloadMass – Analyzes how orbit type impacts payload and success.
- Bar Plot – Compares average success rates across different orbit types.
- https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/edadataviz%20(1).ipynb

EDA with SQL

- 1. List all unique launch site names.
- 2. Show 5 launches where the site name starts with "CCA".
- 3. Calculate the total payload mass carried by NASA (CRS) missions.
- 4. Find the average payload mass for booster version F9 v1.1.
- 5. Identify the earliest date a successful landing occurred on a ground pad.
- 6. List booster versions with successful drone ship landings and payloads between 4000 and 6000 kg.
- 7. Count how many missions were successful or failed.
- 8. Find the booster versions that carried the heaviest payload.
- 9. Retrieve 2015 records of failed drone ship landings, showing month, booster version, and launch site.
- 10. Rank landing outcomes by how often they occurred between June 4, 2010, and March 20, 2017.

https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- We plotted all launch sites and used Folium map elements like markers, circles, and lines to visualize whether each launch was successful or failed.
- We encoded launch outcomes into a `class` column, where 0 represents a failure and 1 represents a success.
- By using color-coded marker clusters, we were able to identify which launch sites demonstrated higher success rates.
- We also measured distances from each launch site to nearby features such as railways, highways, coastlines, and cities to explore spatial relationships. This helped us answer questions like:
    - Are launch sites typically located near transport infrastructure or coastlines?
    - Do launch sites tend to maintain a certain distance from urban areas?

https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/lab_jupyter_launch_site_location.ipynb

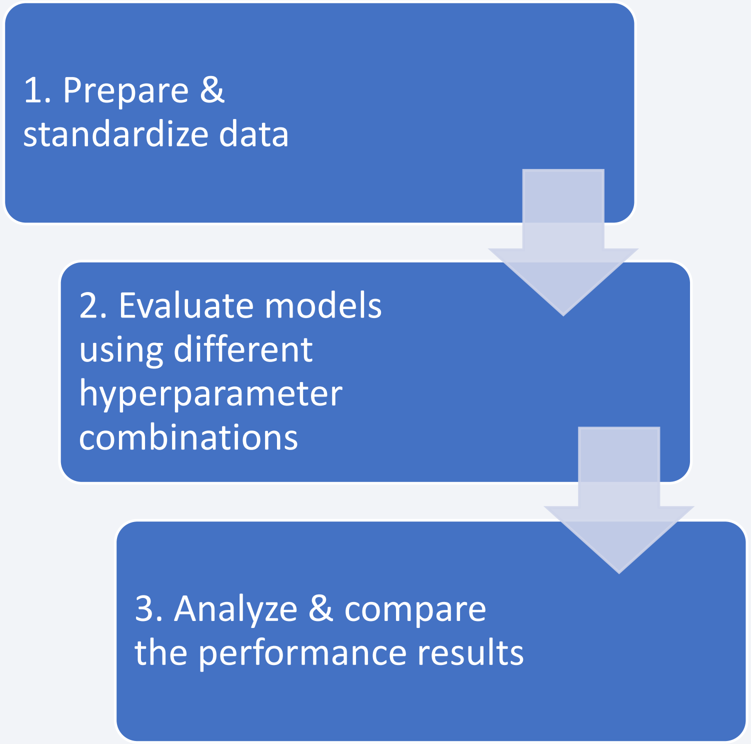
Build a Dashboard with Plotly Dash

- https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/spacex-dash-app%20(1).py
- We created an interactive dashboard using Plotly Dash.
  It includes pie charts displaying the total number of launches by specific sites, and a scatter plot illustrating the relationship between launch outcomes and payload mass (in kg) across different booster versions.

Predictive Analysis (Classification)

- We loaded the dataset with NumPy and Pandas, performed data transformation, and split it into training and testing sets.
- We compared several machine learning models (compared: logistic regression, support vector machine, decision tree and k nearest neighbors) and optimized their hyperparameters using GridSearchCV.
- Model performance was evaluated using accuracy as the primary metric, and we enhanced results through feature engineering and algorithm tuning.

- https://github.com/trumantnt/IBM-Data-Science-Capstone-SpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



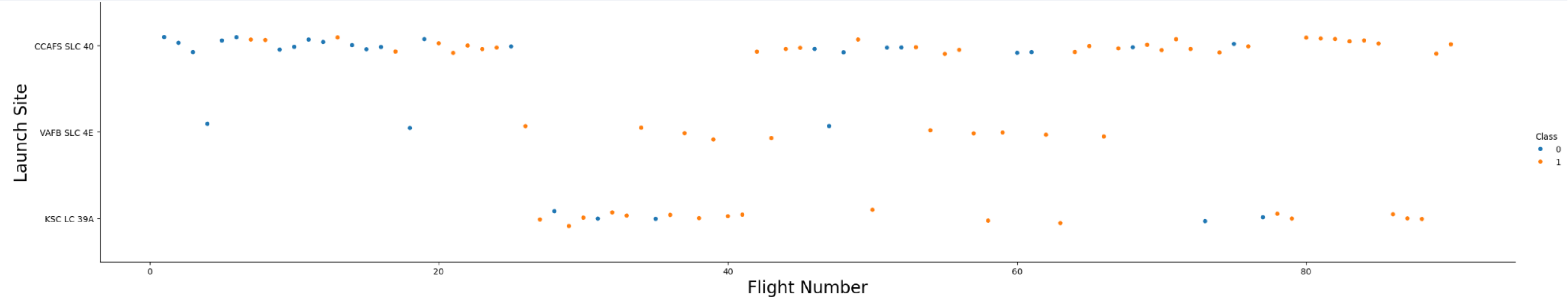
Results

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

Section 2

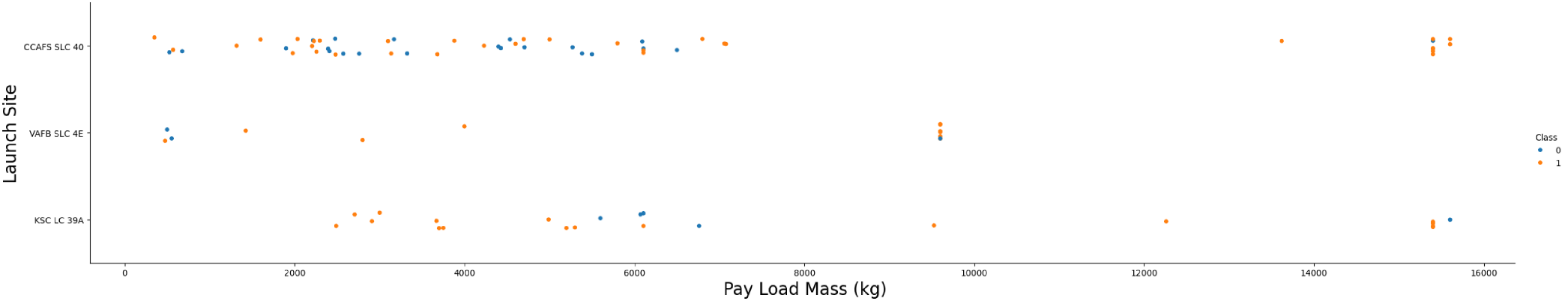
Insights drawn from EDA

Flight Number vs. Launch Site



- The plot indicates that the most reliable current launch site is CCAF5 SLC-40, with the highest number of recent successful launches.
- VAFB SLC-4E ranks second, followed by KSC LC-39A in third place.
- Overall, the success rate of launches has shown consistent improvement over time.

Payload vs. Launch Site



- Payloads exceeding 9,000 kg tend to have a very high success rate.
- Only CCAFS SLC-40 and KSC LC-39A appear capable of handling payloads over 12,000 kg.

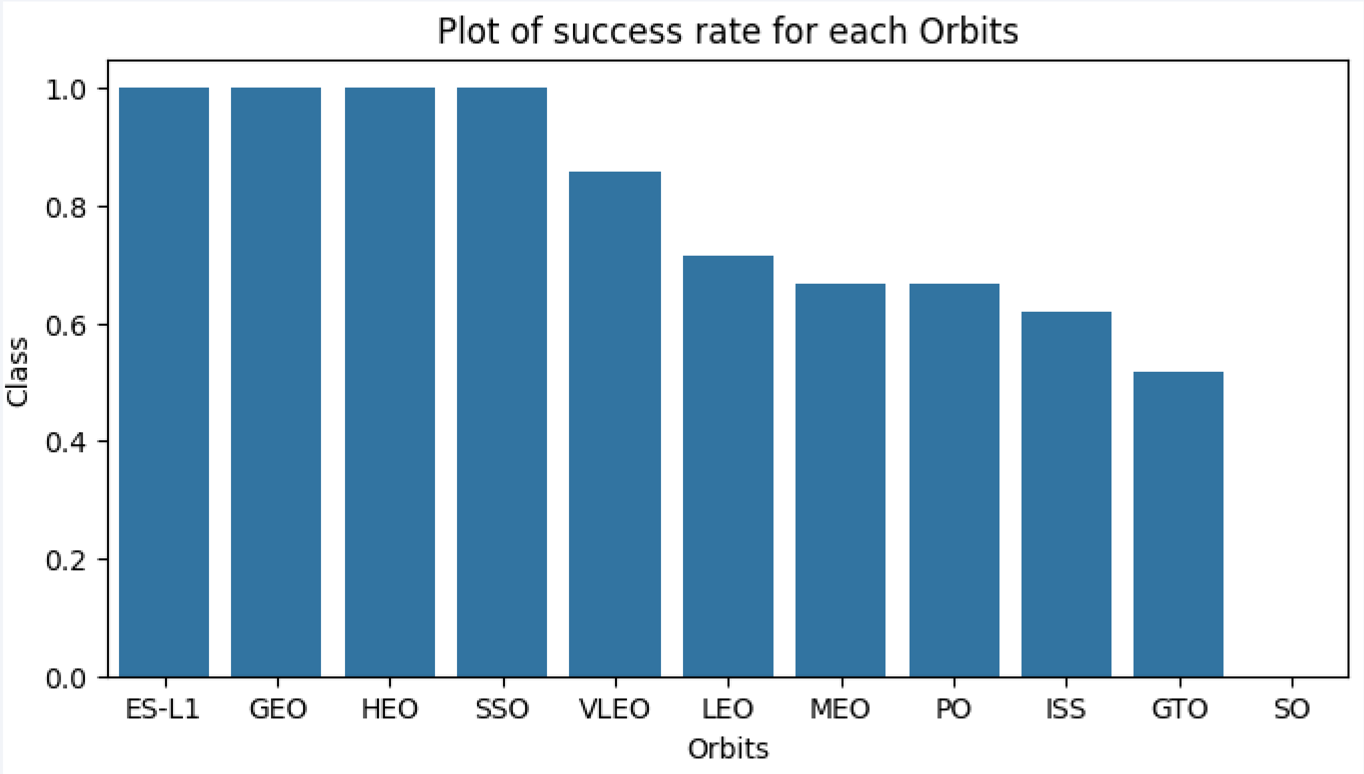
Success Rate vs. Orbit Type

The highest success rates are associated with the following orbits:

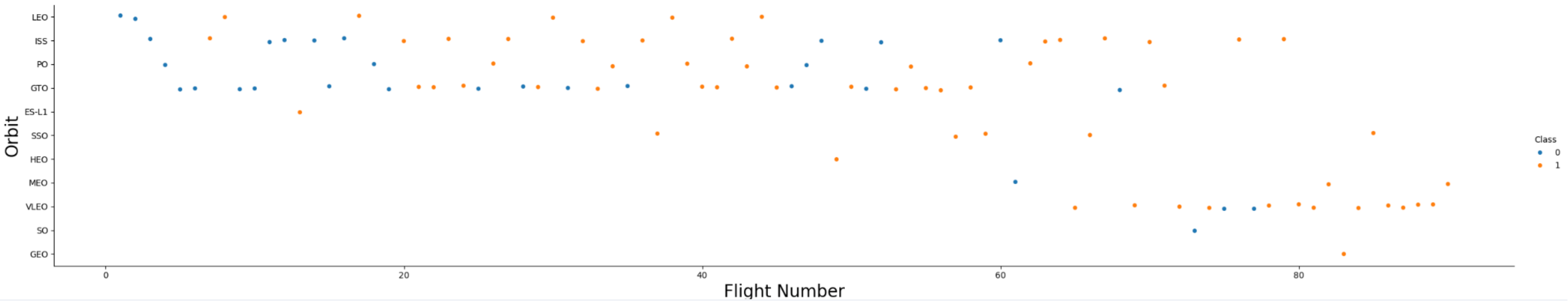
- ES-L1
- GEO
- HEO
- SSO

These are followed by:

- VLEO
- LFO

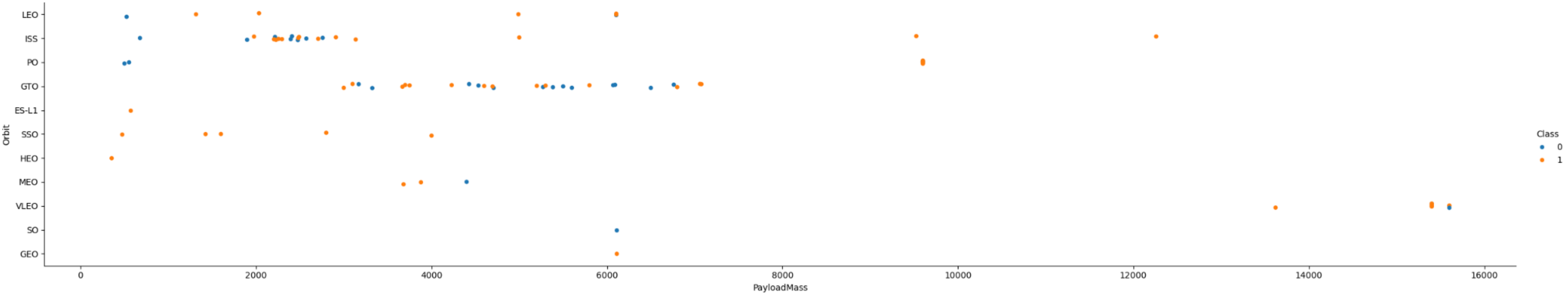


Flight Number vs. Orbit Type



- Success rates have improved over time across all orbit types.
- The VLEO orbit, in particular, shows potential as an emerging business opportunity due to its recent rise in launch frequency.

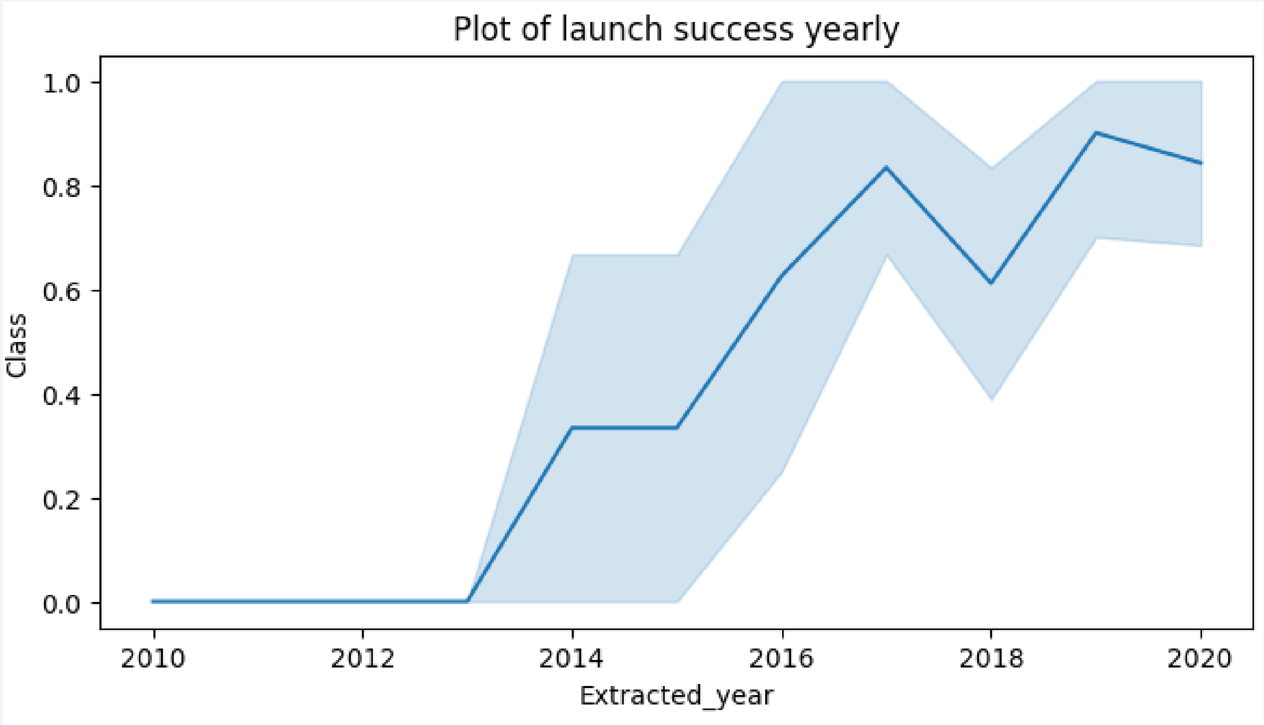
Payload vs. Orbit Type



- There appears to be no clear correlation between payload size and success rate for the GTO orbit.
- The ISS orbit supports the broadest range of payloads and maintains a strong success rate.
- Launches to the SO and GEO orbits remain relatively limited in number.

Launch Success Yearly Trend

- The success rate began rising in 2013 and continued to improve through 2020.
- The initial three years appear to have been a phase of technological development and refinement for SpaceX.



All Launch Site Names

SELECT DISTINCT: This ensures that only unique (non-duplicate) values are returned.

"Launch_Site": The column from which unique values are being selected.

FROM SPACEXTABLE: Specifies the table containing the data.

Task 1

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

[10]:
```
%%sql
SELECT DISTINCT "Launch_Site"
FROM SPACEXTABLE;
```

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

[10]:

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

Launch Site Names Begin with 'CCA'

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

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

[11]:

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

SELECT *: Selects all columns from the matching rows.

WHERE "Launch_Site" LIKE 'CCA%': Filters the results to only rows where the "Launch_Site" begins with 'CCA'.

- The % is a wildcard character that means "any sequence of characters".
- 'CCA%' matches strings like 'CCAFS SLC 40', 'CCAFS LC-41', etc.

LIMIT 5: Returns only the first 5 matching records.

Total Payload Mass

```
%%sql
SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

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

Total_Payload_Mass
45596

- 45596 is the single number showing the total mass of all payloads launched for NASA's Commercial Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG("Payload_Mass__kg_") AS Average_Payload_Mass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

- We found that the booster version F9 v1.1 carried an average payload mass of approximately 2,928.4 kg.

First Successful Ground Landing Date

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

Hint:Use min function

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

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

First_Successful_Landing_Date
2015-12-22

- We observed that the first successful ground landing occurred on December 21, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have 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;
```

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

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

- **WHERE "Landing_Outcome" = 'Success (drone ship)':** Filters the rows to only include those where the booster successfully landed on a drone ship.
- **AND "Payload_Mass__kg_" BETWEEN 4000 AND 6000:** Further filters the results to include only rows where the payload mass is between 4,000 and 6,000 kilograms (inclusive).

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

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

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

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

- **SELECT "Mission_Outcome", COUNT(*) AS Outcome_Count**: Retrieves each unique value in the "Mission_Outcome" column.COUNT(*) counts how many rows have that specific outcome.The result is labeled as Outcome_Count.
- **GROUP BY "Mission_Outcome"**: Groups the data by each distinct mission outcome so that the count is calculated for each group.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass

```
%%sql
SELECT "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

We identified the booster that carried the heaviest payload by using the MAX() function within a subquery in the WHERE clause.

2015 Launch Records

```
%%sql
SELECT substr(Date, 6, 2) AS Month,
       "Landing_Outcome",
       "Booster_Version",
       "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Failure (drone ship)'
  AND substr(Date, 0, 5) = '2015';
```

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

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

- We applied a combination of the WHERE clause with LIKE, AND, and BETWEEN conditions to filter records from 2015 that involved failed drone ship landings, along with their corresponding booster versions and launch site names.

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

Rank the count of landing outcomes (such as Failure (dro
in descending order.

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

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

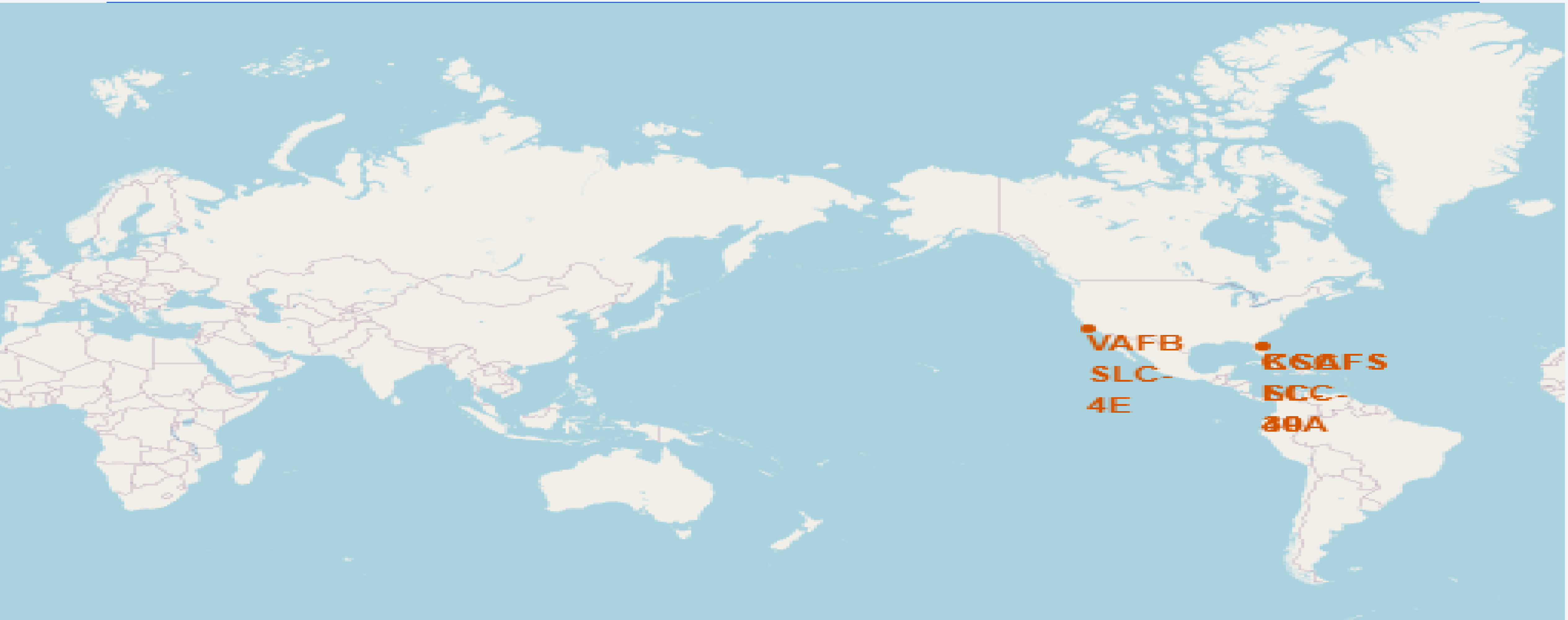
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- We retrieved landing outcomes and their counts by filtering the data using the WHERE clause for dates between June 4, 2010, and March 20, 2010.
- Then, we grouped the results by landing outcome using the GROUP BY clause and sorted them in descending order with the ORDER BY clause.

Section 3

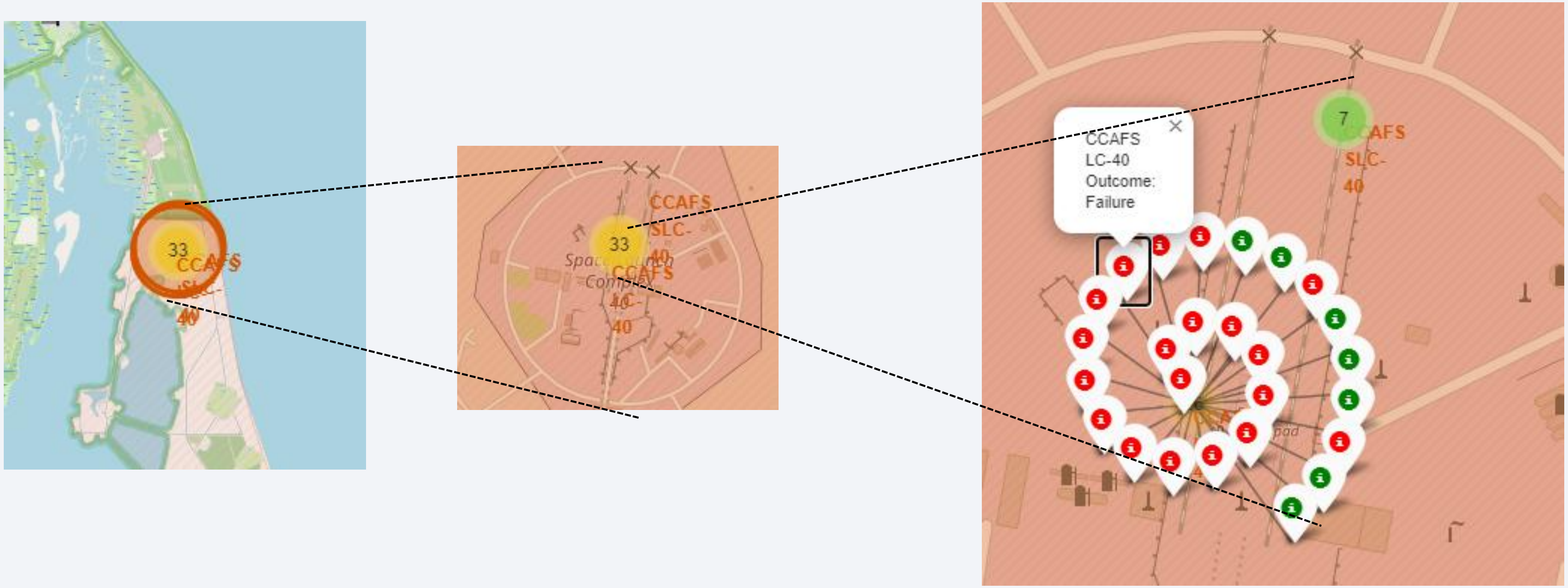
Launch Sites Proximities Analysis

All SpaceX Launch Sites



All SpaceX launch sites are in the US

Example Launch Outcome Interface



- Red (i) markers indicate failed launches
- Green (i) markers indicate successful launches

Example Distance Indicator

- Lines can be drawn to measure distance from the launch sites to different landmarks like coastline, city, railway station, etc.

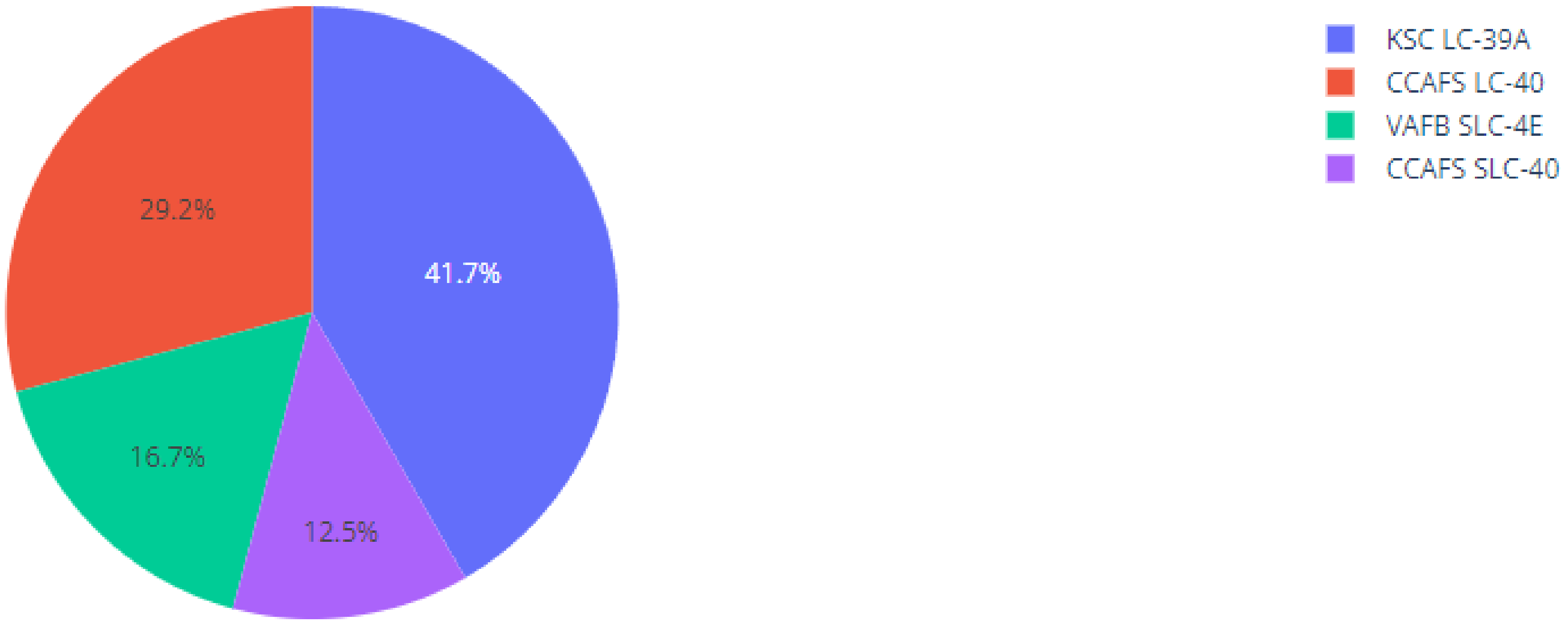
Section 4

Build a Dashboard with Plotly Dash

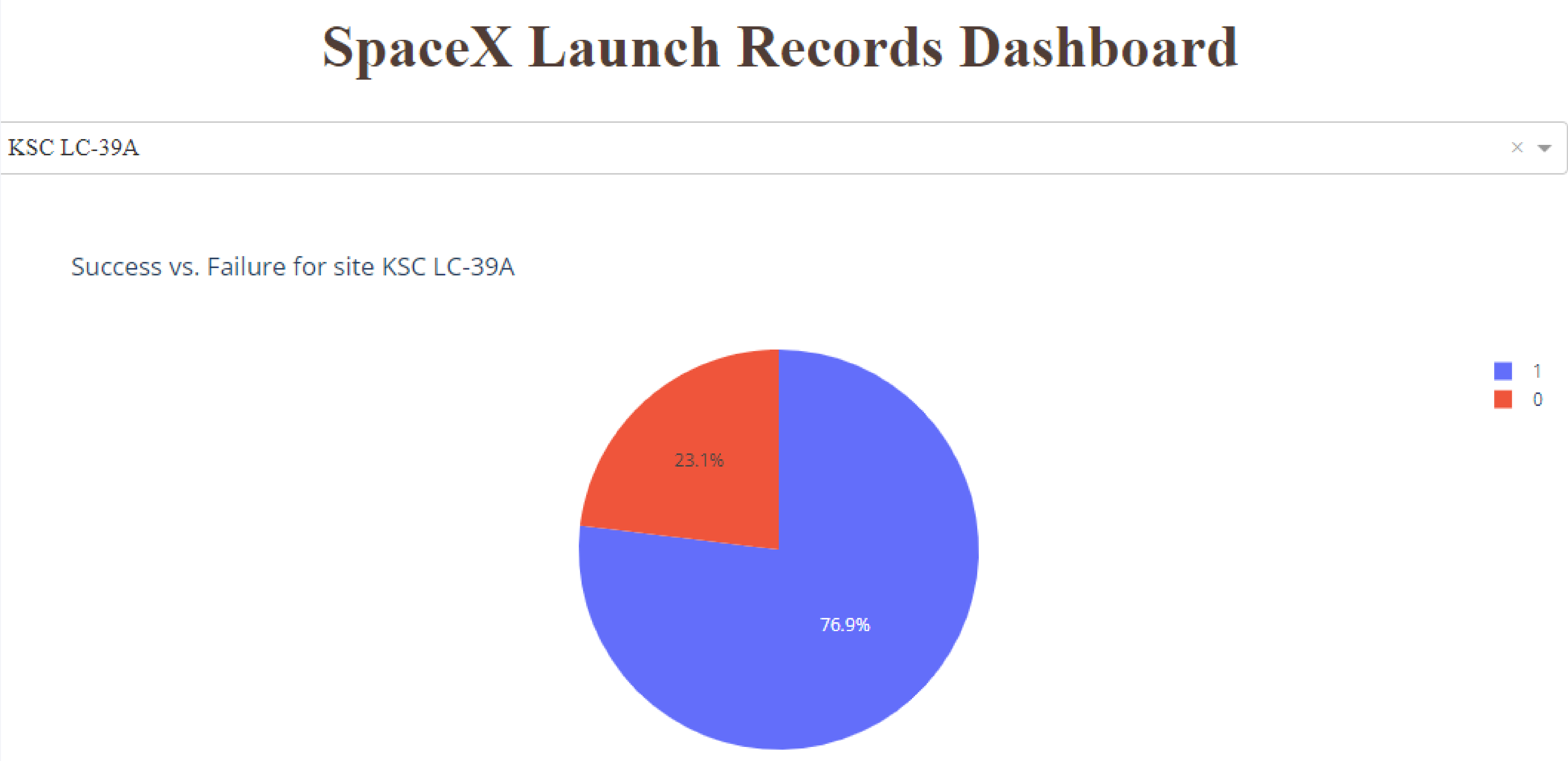
SpaceX Launch Records Dashboard

All Sites

Total Successful Launches by Site



- KSC LC-39A seems to have the most successful launches



- KSC LC-39A has a 76.9% success rate, about 3 success launches for every 4 tries.

- Success rates for low weight payloads is higher.
- V1.1has very low success rates, while FT has quite high success rates

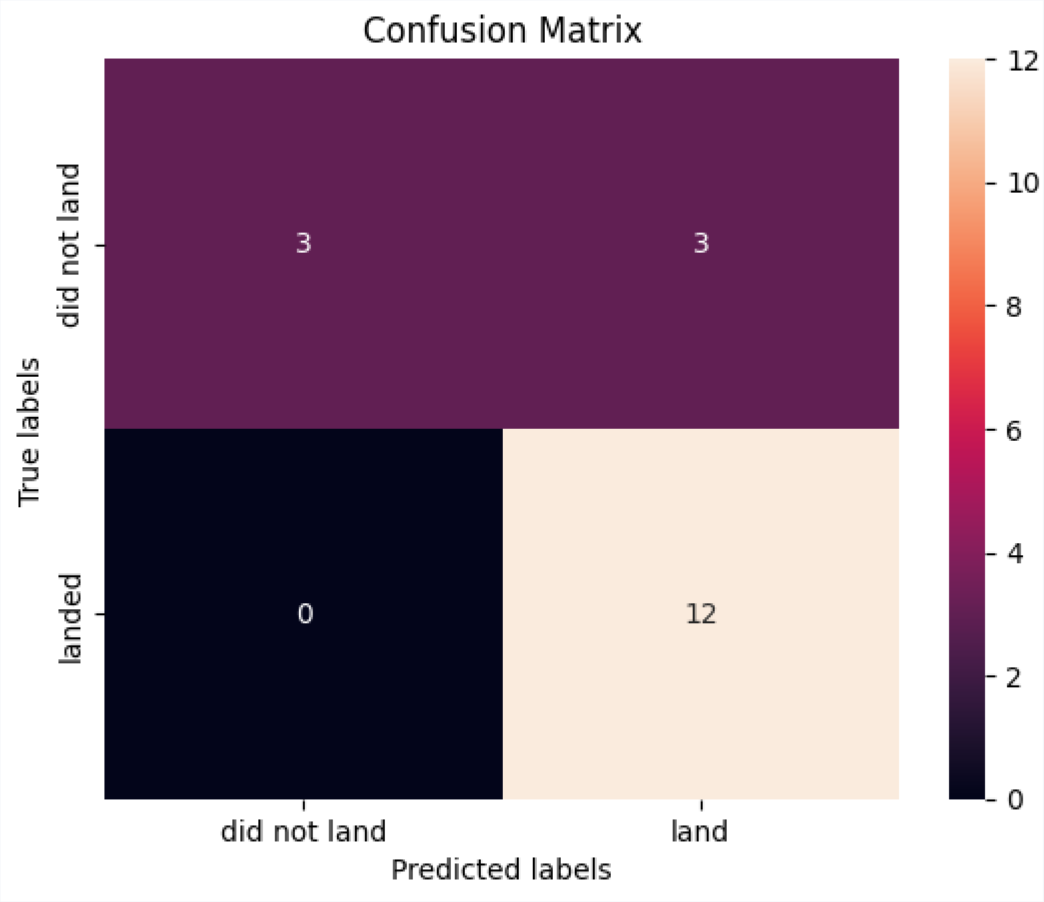
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

Confusion Matrix



- Either KNN or Support Vector machine has better accuracy due to ratio between True Positive vs False Positive and True Negative vs False Negative

Conclusions

- Sites with higher launch volumes tend to show higher success rates.
- KSC LC-39A stands out as the most successful and reliable launch site.
- Certain orbits—such as ES-L1, GEO, HEO, SSO, and VLEO—are associated with higher success rates.
- Launches with FT booster have highest success rates among boosters.
- The Decision Tree Classifier is the most effective machine learning algorithm for predicting successful launches and landings, which can help increase profitability.

Thank you!