Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- **Goal:** The aim of this project was to analyze SpaceX launch data, process it to extract insights, and build predictive models to predict the success of future launches.
- **Key Findings:**
  - Web scraping and API methods were used to collect launch data.
  - Exploratory analysis identified key success factors in payload mass, launch sites, and orbital types.
  - Classification models were built and evaluated, with the best model achieving an accuracy of XX%.

Introduction

- **Problem Statement:**
  - SpaceX has successfully reduced launch costs through reusable rocket technology. The goal is to predict launch outcomes (success/failure) based on historical data.
- **Data Sources:**
  - SpaceX API for launch details.
  - Web scraping for complementary data.
- **Objective:**
  - To process, visualize, and analyze the data, leading to a predictive model for future launches.

Section 1

Methodology

Methodology

Executive Summary

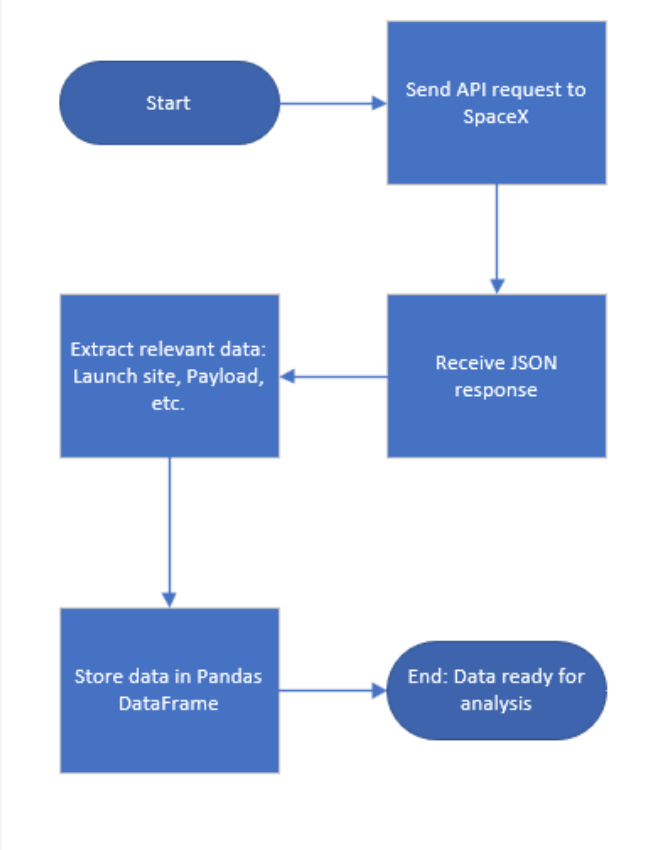
- Data collection methodology:
  - Data was collected through **API collection** and **Web Scraping**
- Perform data wrangling
  - Data **formatting, missing values, duplicates, outliers** were identified and handled.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built classification models (i.e. **Logistic Regression**) to predict launch outcomes
  - Used **GridSearchCV** for hyperparameter tuning to find the best model.
  - Evaluated the models using **accuracy, confusion matrix**, and other metrics

Data Collection

- Data for the SpaceX project was collected through two main methods:
  - **API Data Collection:** Using the SpaceX public API, launch data such as launch sites, outcomes, and payload information were gathered programmatically.
  - **Web Scraping:** For additional details not available via the API, such as launch articles or real-time updates, a web scraping approach was implemented using Python libraries like BeautifulSoup and requests.
- Key Flow
  - **Step 1:** Identify required data points (launch site, payload mass, outcome).
  - **Step 2:** Use the SpaceX API to retrieve data.
  - **Step 3:** Complement missing or unavailable data through web scraping techniques.

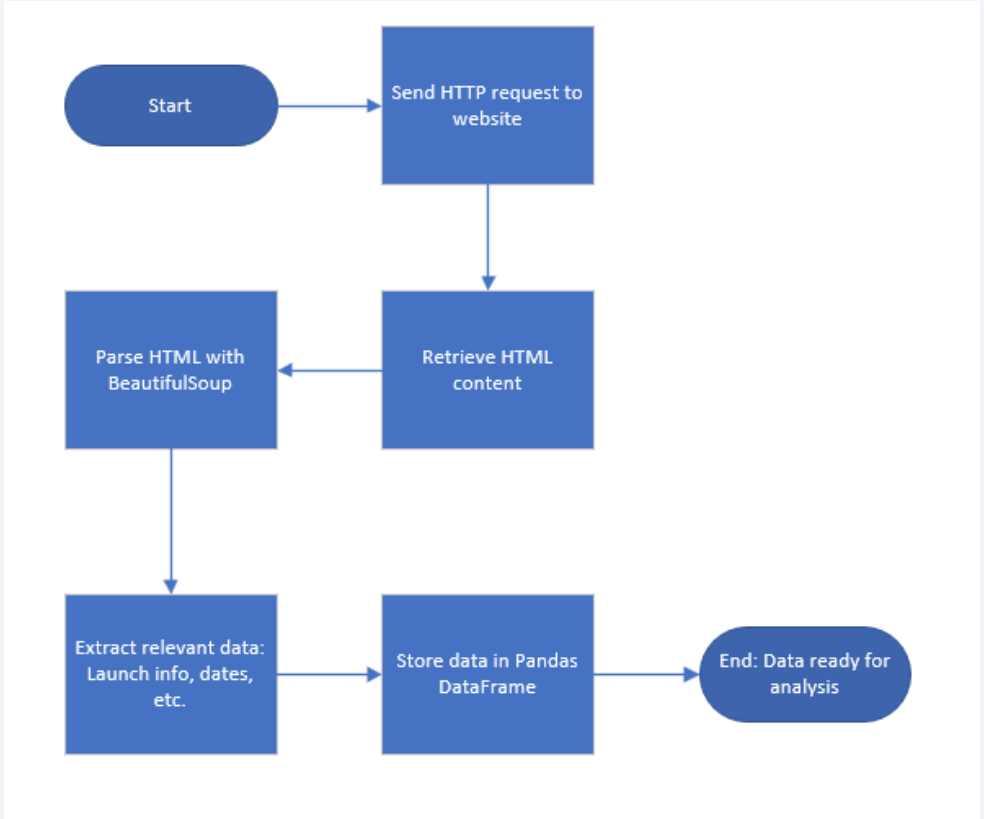
Data Collection – SpaceX API

- The SpaceX API was used to retrieve data about launches, including:
  - Launch Sites
  - Launch Outcomes (Success/Failure)
  - Payload Mass and Type
  - Orbit Type
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/1%20jupyter-labs-spacex-data-collection-api-v2.ipynb



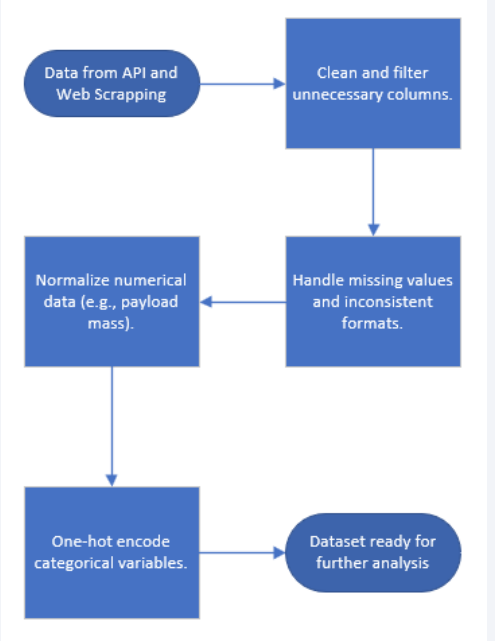
Data Collection - Scraping

- Web Scraping Process:
  - Identify the target website to scrape.
  - Use BeautifulSoup and requests to extract relevant HTML elements.
  - Parse and store the data into a structured format.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/2%20jupyter-labs-webscraping.ipynb



Data Wrangling

- The SpaceX launch data underwent a series of data wrangling steps to ensure it was clean, structured, and ready for analysis:
  - **Handling Missing Values**: Missing values for payload mass, launch sites, and outcomes were addressed. We either imputed appropriate values or excluded records that were incomplete.
  - **Data Cleaning**: Non-essential columns were removed, and we handled any discrepancies between API data and web-scraped data (e.g., different formats or units).
  - **Data Normalization**: Standardized payload mass and other numerical fields to a uniform format for easy analysis and visualization.
  - **Categorical Data Encoding**: Categorical variables such as launch site and orbit type were one-hot encoded for use in predictive modeling.
  - **Duplicate Removal**: Checked and removed any duplicated records to ensure the integrity of the dataset.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/3%20labs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with Data Visualization

- **Flight Number vs. Launch Site (Scatter Plot):** This chart was used to visualize the distribution of launches across different launch sites. It helps identify the frequency and success patterns for each site.
- **Payload vs. Launch Outcome (Scatter Plot):** This chart shows how payload mass affects the success of the launch at different launch sites. It was important to explore any correlations between payload and launch outcome.
- **Success Rate by Orbit Type (Bar Chart):** The bar chart displays the success rates for different orbit types, providing insight into the reliability of each orbit type in terms of launch success.
- **Yearly Success Rate (Line Chart):** This line chart visualizes the trend in launch success rates over time, helping identify improvements or setbacks in SpaceX's performance over the years.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/5%20jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- **Unique Launch Site Names:** Queried to retrieve the names of all launch sites.
- **Total Payload by NASA:** Calculated the total payload mass carried by SpaceX for NASA.
- **Average Payload Mass by Booster Version F9 v1.1:** Calculated the average payload mass for the F9 v1.1 booster version.
- **First Successful Landing on Ground Pad:** Queried for the date of the first successful ground pad landing.
- **Total Successful and Failed Mission Outcomes**: Calculated the total number of successful and failed mission outcomes across all launch sites.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/4%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

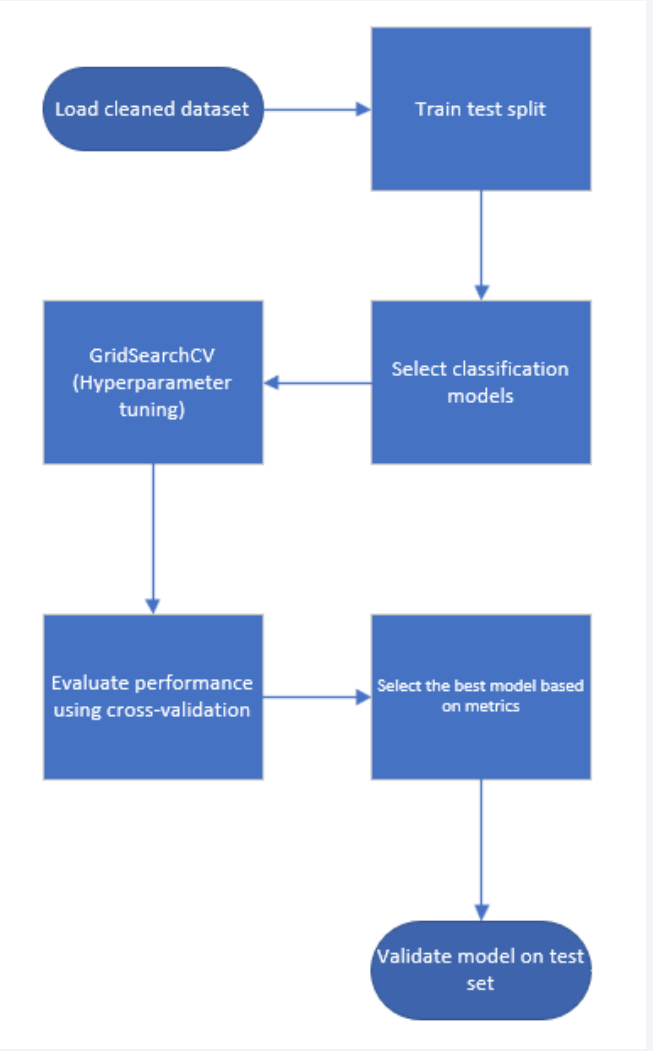
- **Launch Site Markers**: Added markers for each SpaceX launch site to visualize their geographical locations.
- **Circle Markers:** Added circle markers around each launch site to show proximity areas and launch influence zones.
- **Lines to Proximities:** Added lines from launch sites to key infrastructure such as highways, railways, and coastlines to visualize distances.
  - The **markers** were used to pinpoint each launch site's location, helping to understand spatial distribution.
  - **Circles** were added to represent areas of influence or safety zones around each launch site.
  - The **lines** were drawn to analyze the proximity of launch sites to critical infrastructure, which could be factors affecting logistics and safety.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/6%20lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown for Launch Site Selection**: A dropdown menu was added to allow users to filter data by specific SpaceX launch sites or view data for all launch sites.
- **Pie Chart** visualizes the total number of successful launches across different sites
- **Range Slider for Payload Selection**: A slider allows users to filter the data based on a range of payload mass values (0 to 10,000 kg).
- **Scatter Plot:**A scatter plot visualizes the relationship between payload mass and launch success.
    - Dropdown for Launch Site: This allows users to focus on specific launch sites, providing flexibility to analyze data site by site or overall.
    - Pie Chart: The pie chart gives a clear, immediate overview of launch outcomes (success vs. failure) either across all sites or for individual sites.
    - Range Slider: The slider helps users explore how payload mass affects launch outcomes by filtering the data to a specific payload range.
    - Scatter Plot: The scatter plot highlights the correlation between payload mass and launch success. The color-coding by booster version provides further insight into the booster's role in launch outcomes.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/spacex_dash_app.py

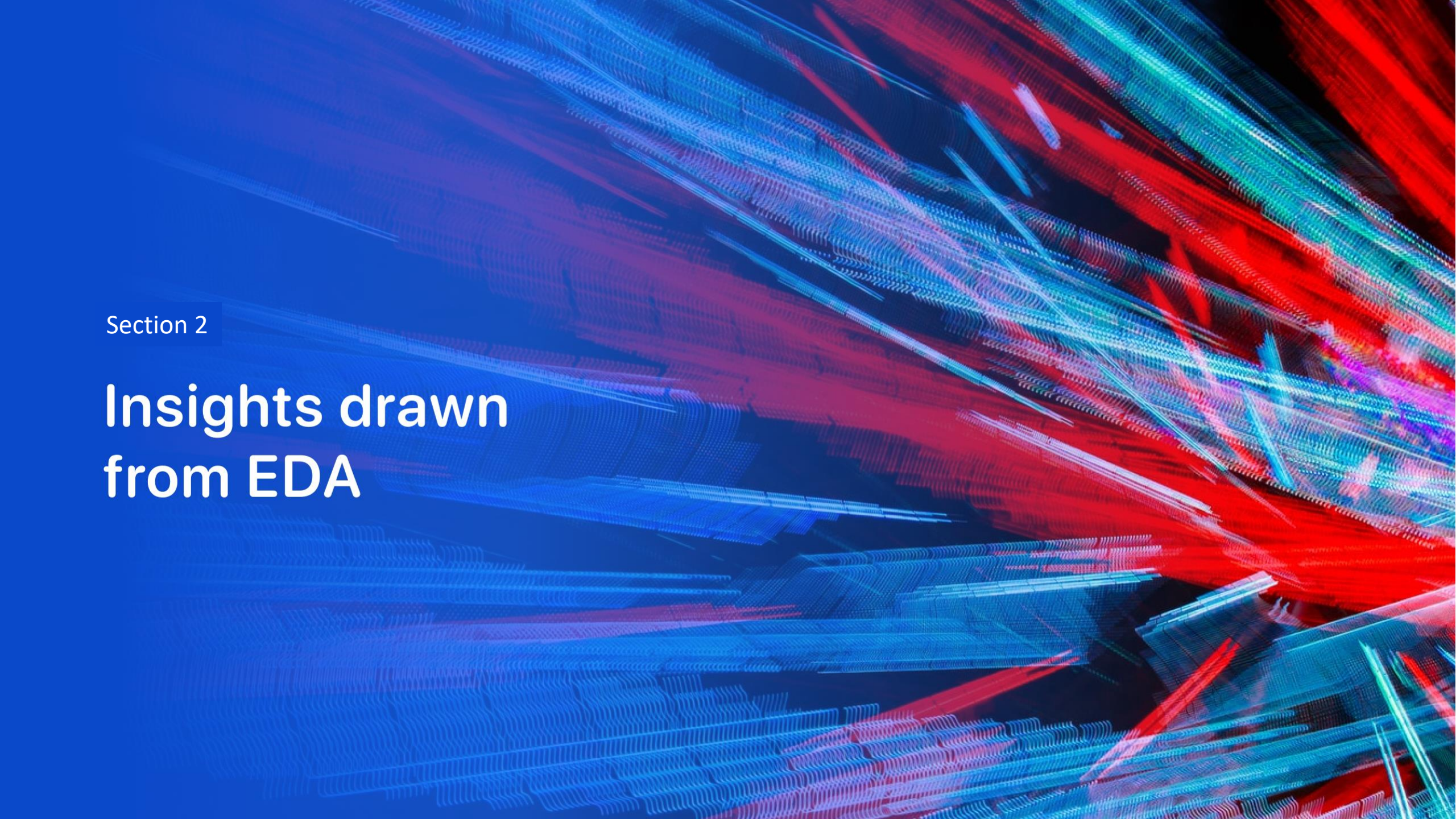
Predictive Analysis (Classification)

- **Model Selection:** Multiple classification models were considered, including Logistic Regression, Decision Trees, SVM and KNN. The goal was to predict the success of SpaceX launches based on features like payload mass, launch site, and booster version.
- **Hyperparameter Tuning:** A GridSearchCV was used to tune model hyperparameters. For Logistic Regression, the best regularization parameter (C) was found by testing values [0.01, 0.1, 1].
- **Evaluation:** Models were evaluated using cross-validation, with accuracy and other metrics like precision, recall, and F1-score. Confusion matrices were used to visualize classification performance.
- Best Model: All returns the same score suggesting different approach (i.e. more data) is needed.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM/blob/main/7_SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

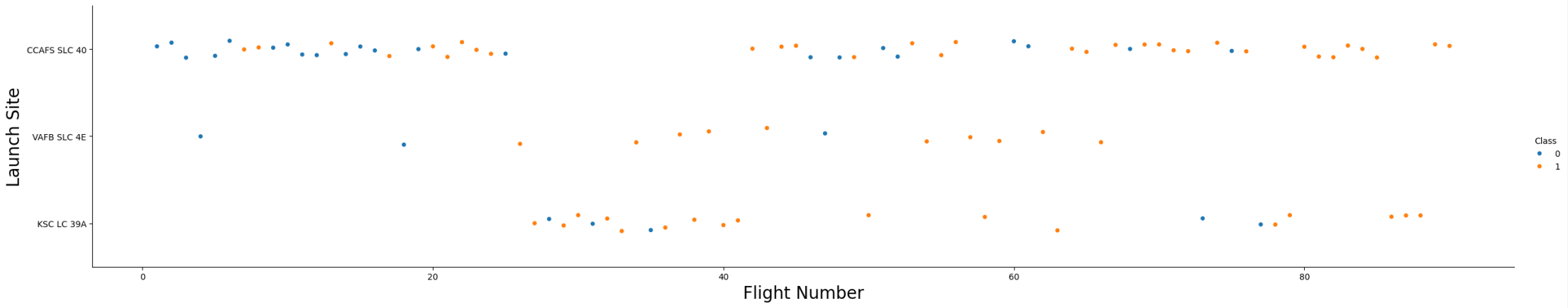


Results

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


Section 2
Insights drawn from EDA

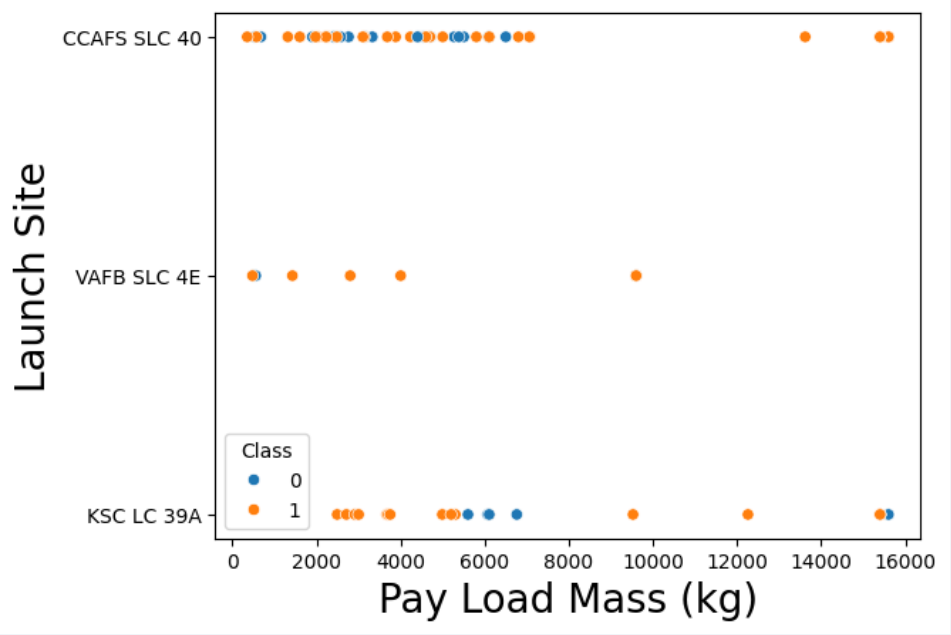
Flight Number vs. Launch Site



- Success rate improves with increasing **Flight Number**, indicating that as SpaceX gained more experience, the likelihood of launch success increased.
- Certain launch sites, like **CCAFS SLC 40**, have more launches and a higher success rate, suggesting these sites might be preferred for critical or frequent launches.

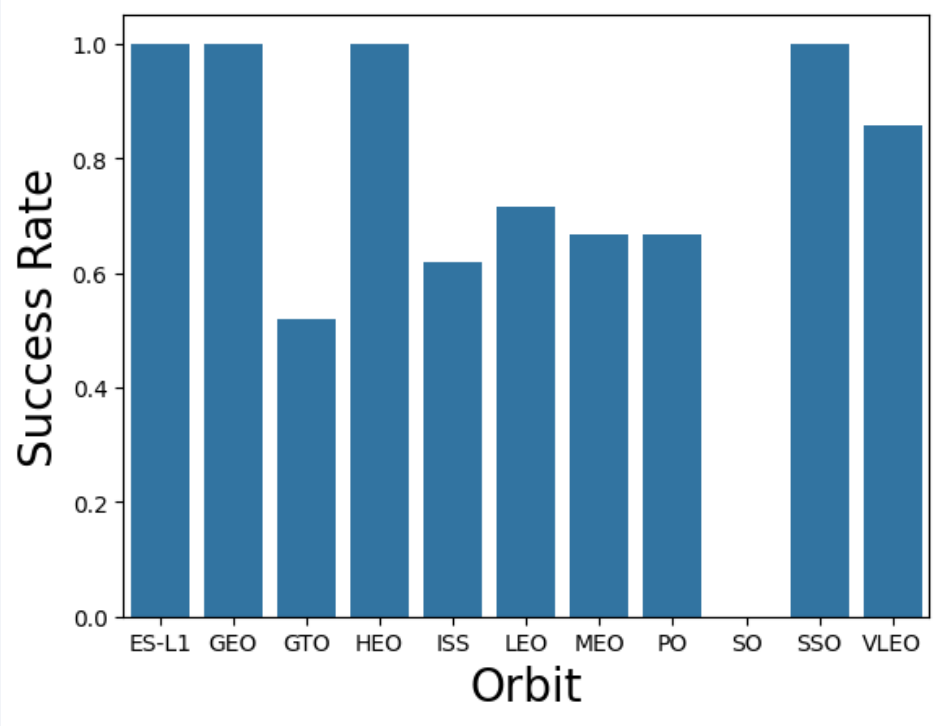
Payload vs. Launch Site

- The site **CCAFS SLC 40** has a diverse range of payload masses, but it also has more failed launches compared to the other sites.
- **VAFB SLC 4E** generally has successful launches with relatively smaller payloads.
- **KSC LC 39A** handles heavier payloads, with more successful launches at higher payloads, suggesting this site may be used for more critical or larger payload missions.



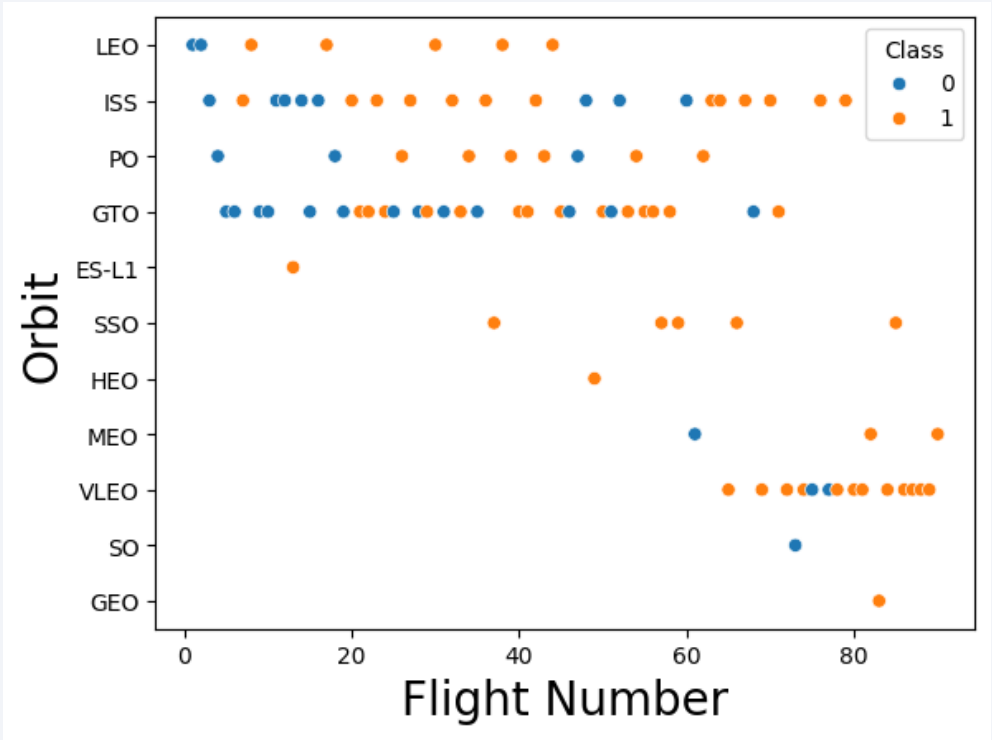
Success Rate vs. Orbit Type

- The chart highlights that **GTO (Geostationary Transfer Orbit)** presents the most challenges, with a significant number of failures.
- Orbits like **GEO (Geostationary Orbit)** and **SSO (Sun-synchronous Orbit)** are highly reliable, with all launches to these orbits being successful.
- This analysis helps identify the relative difficulty of achieving successful launches based on the target orbit type, potentially informing future mission planning.



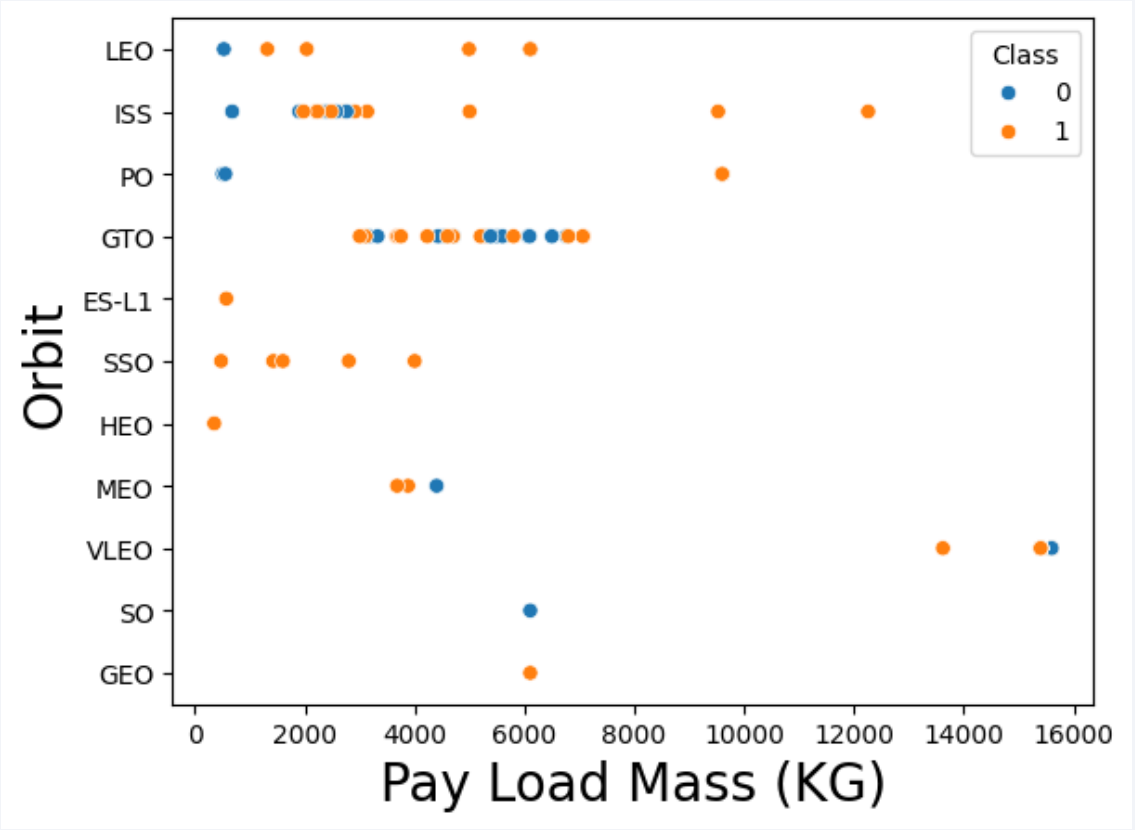
Flight Number vs. Orbit Type

- **Low Earth Orbit (LEO) and International Space Station (ISS)** are the most frequently targeted orbits, and despite their high use, they still show a few launch failures, especially in earlier flights.
- The **GTO** orbit stands out as particularly risky, with a larger proportion of failures relative to other orbits.
- Over time (as flight numbers increase), the success rate improves, suggesting that SpaceX has refined its technology and processes over the course of their flight history.



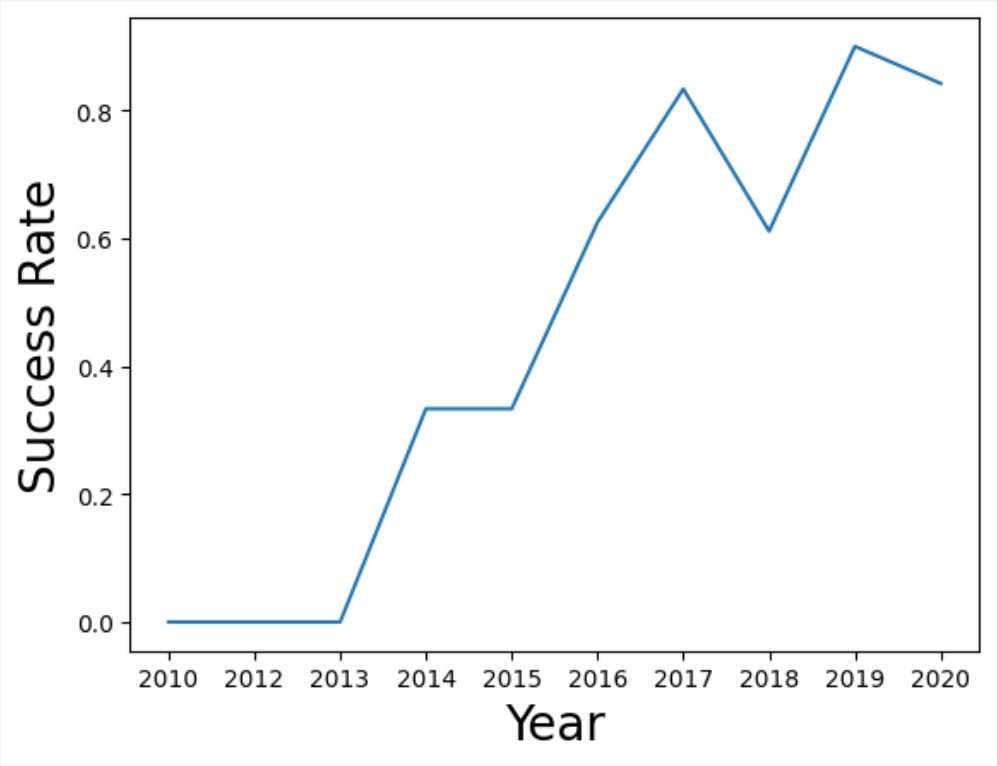
Payload vs. Orbit Type

- The chart shows that **LEO (Low Earth Orbit)** and **ISS (International Space Station)** have more launches with a mix of successes and failures, whereas orbits like GTO have a higher failure rate, especially at higher payloads.
- **Payload Mass** seems to be a factor influencing success for certain orbits, particularly GTO where higher payloads correspond to more failures.
- Orbits like **SSO and VLEO** are more reliable, with most launches being successful regardless of payload mass.



Launch Success Yearly Trend

- This chart demonstrates SpaceX's growing proficiency over the years. Starting from consistent failures in the early 2010s, the company managed to turn around its success rate, especially from 2014 onwards.
- By 2020, SpaceX had successfully achieved a stable success rate of around 80-90%, showcasing the maturation of its launch capabilities.



All Launch Site Names

SELECT DISTINCT Launch_Site

FROM SPACEXTABLE;

- Explanation: The SQL query selects distinct launch site names from the database. This ensures that only unique launch site names are displayed without duplicates.
- Query Result: The result shows four unique launch sites used in the space missions:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

```
SELECT *

FROM SPACEXTABLE

WHERE Launch_Site LIKE
"CCA%“

LIMIT 5;
```

- The query retrieves the first 5 records from the SpaceX dataset where the Launch Site field begins with 'CCA'.
- The specific launch site selected in these records is 'CCAFS LC-40'. This is a notable site used for many of SpaceX's launches.

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

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596

```
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)";
```

- **Query Result:** The total payload mass carried by boosters from NASA is **45,596 kg**.
- **Explanation**: The query sums up the payload mass for all launches where NASA was the customer. The total payload mass for all missions handled by NASA amounts to 45,596 kg.

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2337.8

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

- The query calculates the average payload mass carried by the Falcon 9 (F9) booster version 1.1.
- The result of the query shows that the average payload mass is 2337.8 kg for missions using this booster version.
- This information can help analyze the performance and capabilities of the F9 v1.1 in carrying payloads to orbit.

First Successful Ground Landing Date

MIN(Date)
2015-12-22

```
SELECT MIN(Date)

FROM SPACEXTABLE

WHERE Landing_Outcome = "Success (ground pad)";
```

- **Explanation**: The SQL query was executed to find the earliest date when the landing outcome was a success on a ground pad. The query specifically filters the data for records where the landing outcome is "Success (ground pad)" and retrieves the minimum date.
- **Result**: The first successful ground landing occurred on **December 22, 2015.**

Successful Drone Ship Landing with Payload between 4000 and 6000

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

```
SELECT Payload FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)"
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000;
```

- This query was used to filter successful drone ship landings where the payload mass was between 4000 and 6000 kilograms. The result provided the names of four payloads that met these conditions, indicating the specific missions that were within this payload range and successfully completed drone landings.

Total Number of Successful and Failure Mission Outcomes

```
SELECT 'Failure' AS mission_outcomes, COUNT(*)
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE "Failure%“
UNION
SELECT 'Success' AS mission_outcomes, COUNT(*)
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE "Success%"
```

mission_outcomes	COUNT(*)
Failure	1
Success	100

- **Explanation**: The query was used to count the total number of successful and failure mission outcomes from the SpaceX mission dataset. It involved using the COUNT(*) function and selecting outcomes that began with the word "Success" or "Failure" to classify the results.
- This breakdown demonstrates a high success rate for SpaceX missions in this dataset.

Boosters Carried Maximum Payload

```
SELECT Booster_Version

FROM SPACEXTABLE

WHERE PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTABLE);
```

- The query identified these booster versions based on the maximum payload mass carried, determined using a subquery to select the maximum payload value from the table.

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

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

SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE

WHERE substr(Date,0,5)='2015'

AND Landing_Outcome = "Failure (drone ship)"

- The records for 2015 show two failed landing outcomes on drone ships. Both launches were from "CCAFS LC-40" using booster versions F9 v1.1 B1012 and B1015 in **January and April** of 2015.

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

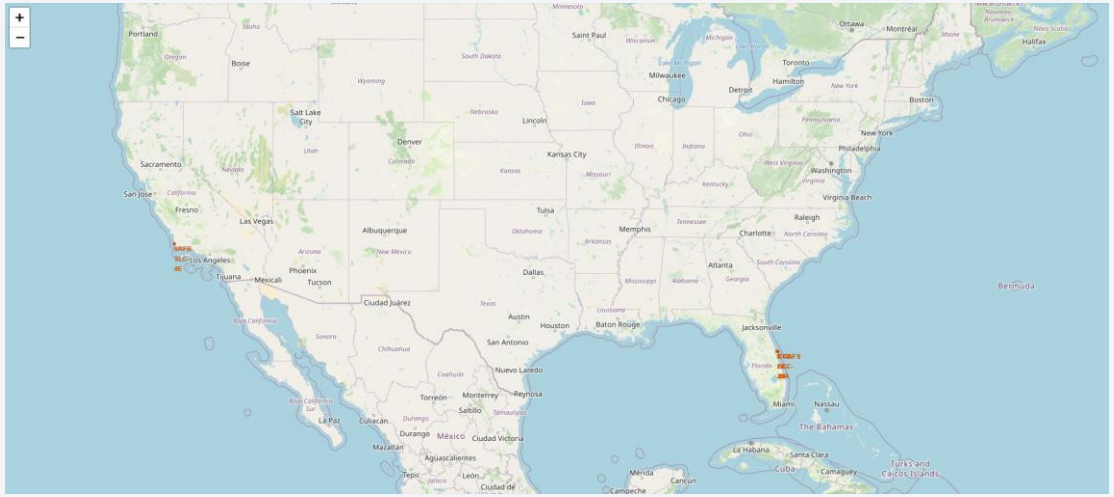
```
SELECT DENSE_RANK()OVER(ORDER BY COUNT(Landing_Outcome) DESC) AS rnk, Landing_Outcome, COUNT(Landing_Outcome) AS N

FROM SPACEXTABLE

WHERE Date BETWEEN "2010-06-04" AND "2017-03-20“

GROUP BY Landing_Outcome

ORDER BY COUNT(Landing_Outcome) DESC
```

- The SQL query ranks the different landing outcomes based on the count of occurrences between June 4th, 2010, and March 20th, 2017.
- This ranking helps to identify the frequency of successful and failed landings across different landing methods during this time period.

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


Section 3
Launch Sites
Proximities Analysis

<Folium Map Showing SpaceX Launch Sites>

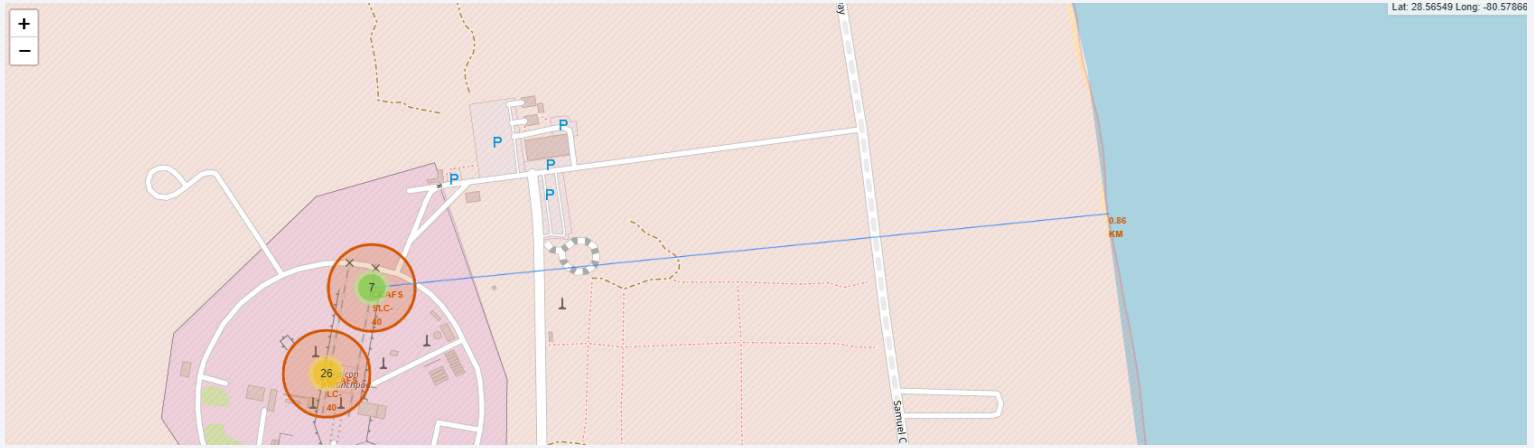


- The folium map displays the locations of SpaceX's launch sites across the United States.
- Location markers: The launch sites are represented by markers on the map, making it easy to identify their geographic positions.
- Geographic range: The map spans both the west and east coasts of the United States, emphasizing the strategic placement of the launch sites.

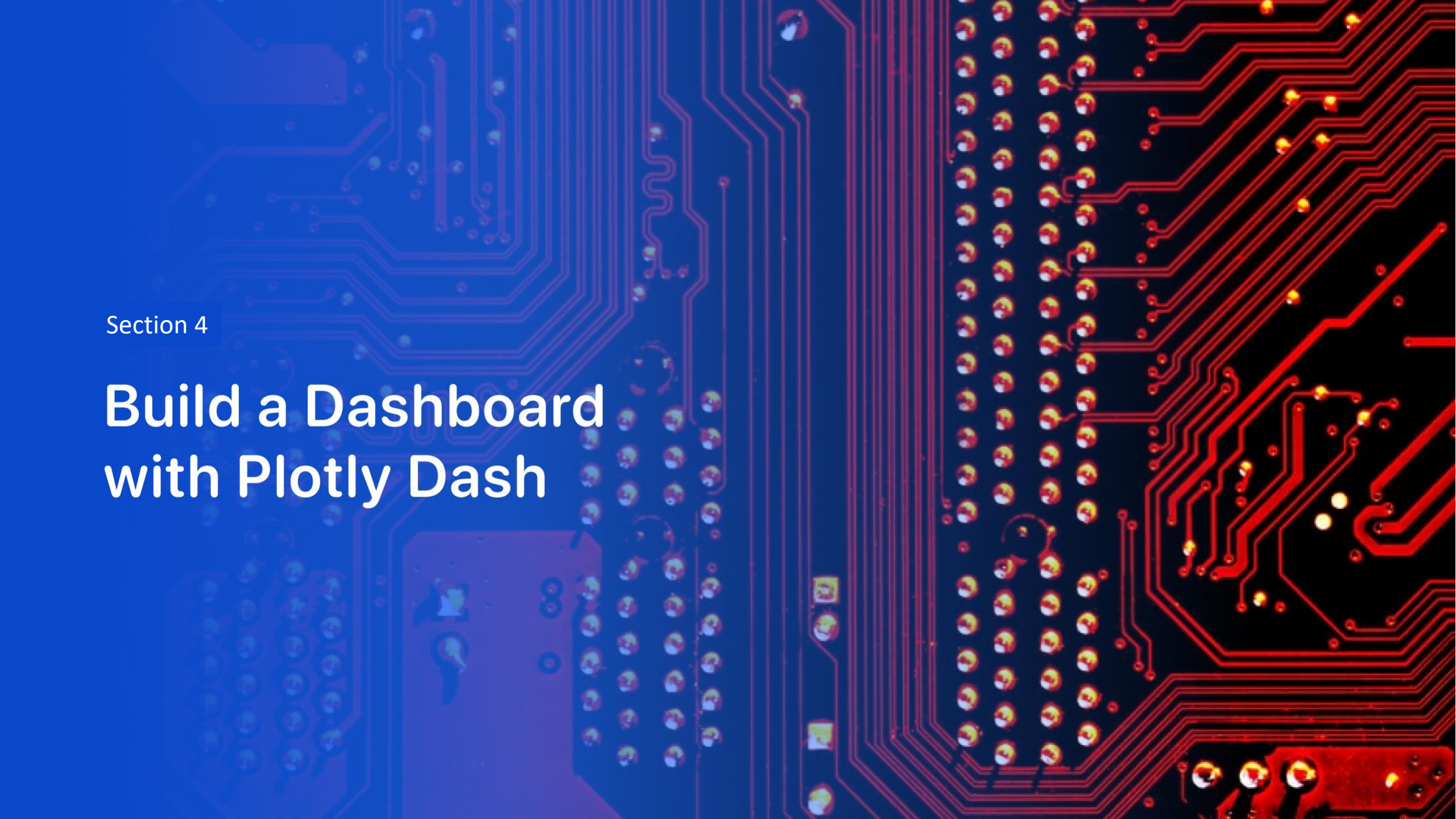
<Launch Outcomes Displayed on the Folium Map>

- The second folium map shows detailed color-coded outcomes of launches. This zoomed-in view specifically focuses on the CCAFS SLC-40 launch site.
- Key Elements:
  - Markers: The map uses colored markers to indicate the success or failure of launches:
    - Green markers represent successful landings.
    - Red markers represent failed landings or attempts.
  - Launch Clustering: The overlapping markers at specific launch pads indicate the concentrated number of launches from certain locations.

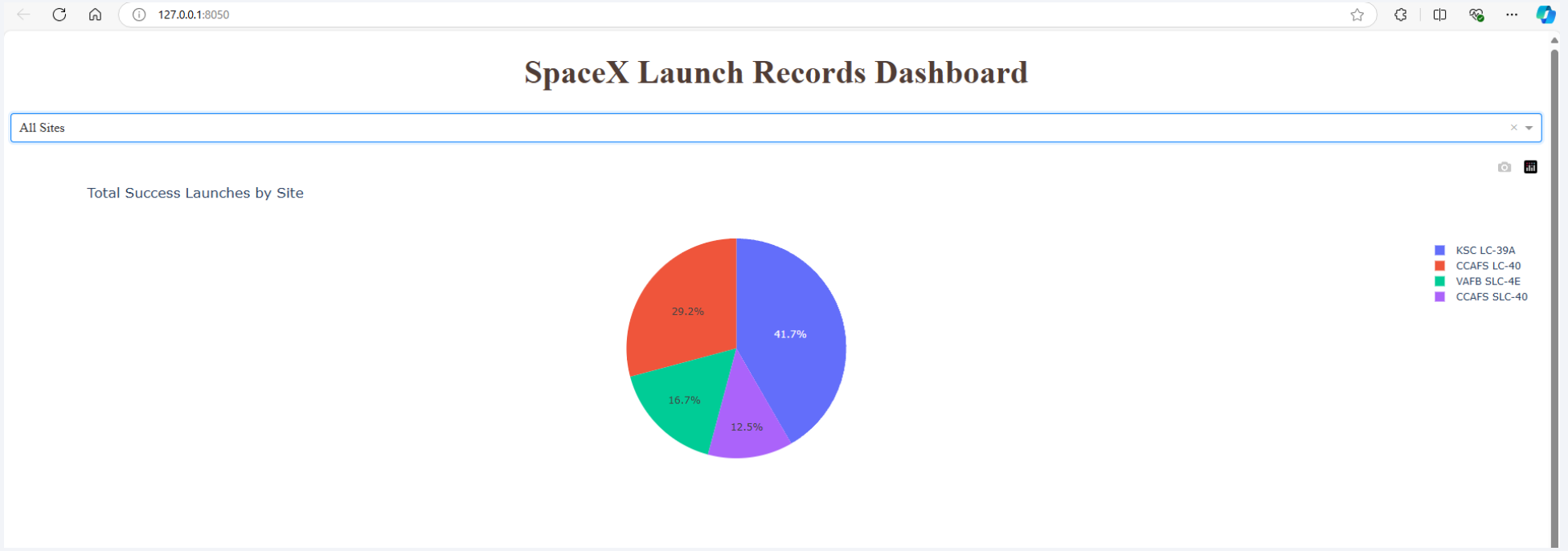
< Proximity Analysis of CCAFS SLC-40 to Coastline >



- **Launch Site:** The map shows the launch pad at CCAFS SLC-40 in Florida, an active launch site for several missions.
- **Proximities:**
  - The site is approximately 0.86 km from the coastline, making it strategically close to the ocean.
  - This proximity is significant for launch operations, as ocean-based recovery methods for rockets are often employed to avoid land hazards.

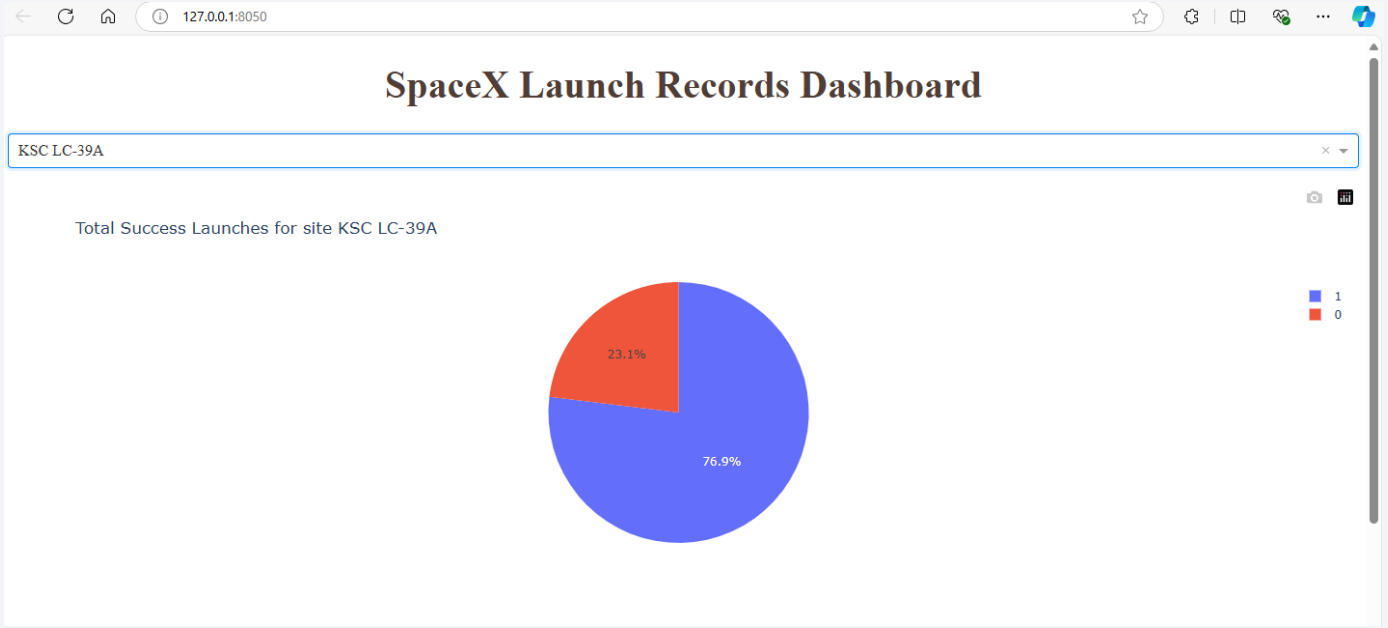

Section 4
Build a Dashboard with Plotly Dash

< Total Success Launches by Site >



- The pie chart displays the distribution of successful SpaceX launches across four different launch sites.
- The chart provides an easy-to-understand visual breakdown of launch success distribution, showing which sites are most frequently used for successful launches.

< Launch Success Ratio by Site >



- The pie chart represents the ratio of successful and failed launches for the selected site (in this case, KSC LC-39A).
- The dashboard allows filtering by different launch sites using a dropdown menu. In the screenshot, the site “KSC LC-39A” is selected, allowing users to focus on launch outcomes s
- The legend on the right explains the color coding in the pie chart:
  - "1" refers to a successful launch.
  - "0" refers to a failed launch.pecific to that site.

< Payload vs. Launch Outcome by Site >

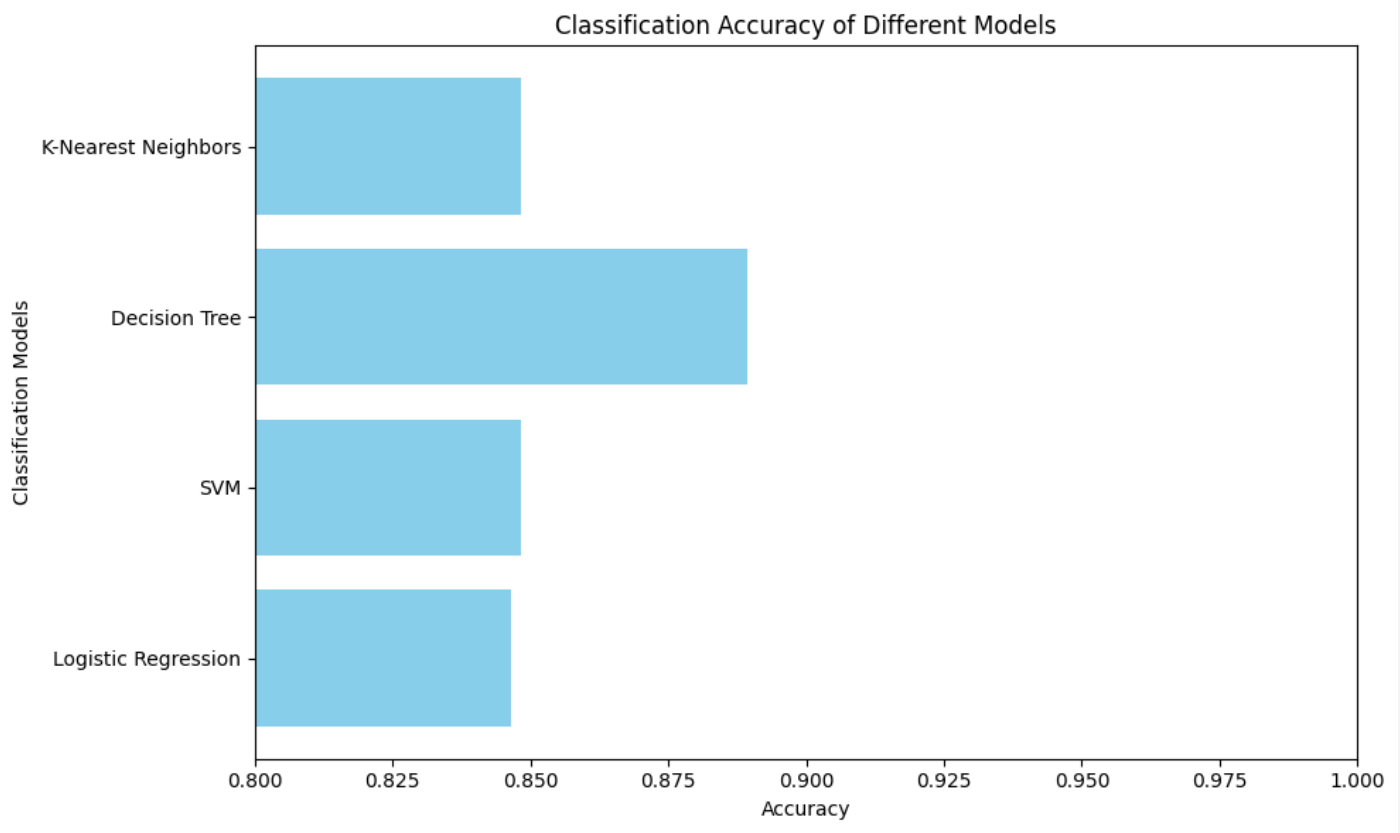
- Launches with payloads around 5000 kg show a higher success rate.
- The **FT** booster version has a higher representation across various payload masses.
- Larger payloads tend to have a mix of outcomes, but payloads over 6000 kg seem to have fewer data points and slightly more variance in outcomes.

Section 5

Predictive Analysis (Classification)

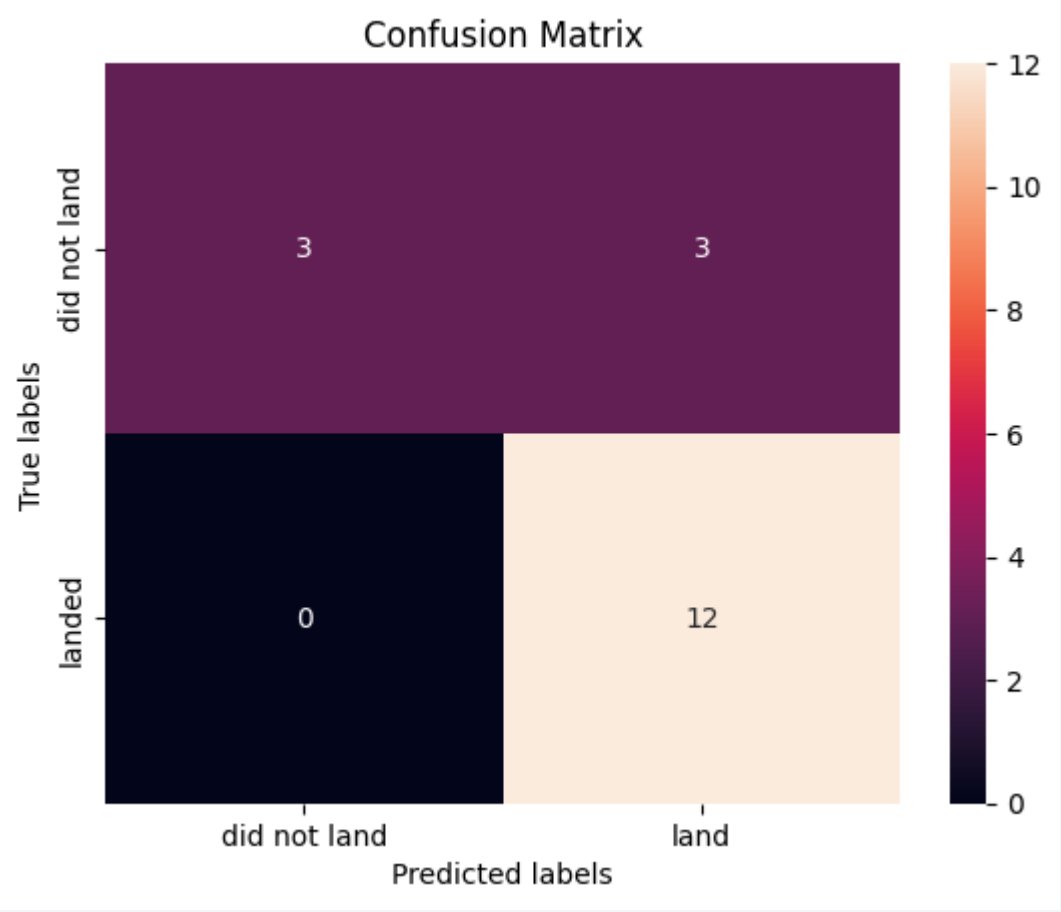
Classification Accuracy

- Chart Explanation: The bar chart visualizes the accuracy of four different classification models: K-Nearest Neighbors, Decision Tree, SVM, and Logistic Regression.
- Highest Accuracy: The Decision Tree model has the highest classification accuracy at around 90%. It outperforms the other models such as K-Nearest Neighbors, SVM, and Logistic Regression, which hover around the 85-87% accuracy mark.
- Insights: This suggests that the Decision Tree model is the most reliable for this classification task, while the other models, though competitive, are slightly less accurate.



Confusion Matrix

- The confusion matrix shown is from the best-performing classification model, where the true labels (actual results) are plotted against the predicted labels (predicted outcomes by the model).
- The matrix indicates that the model is quite accurate in predicting landings, with no false negatives and a relatively small number of false positives.



Conclusions

- The Decision Tree model emerged as the best-performing model with the highest accuracy at approximately 90%, followed by the other models such as Logistic Regression, K-Nearest Neighbors, and SVM.
- Through Exploratory Data Analysis (EDA), factors like payload mass, launch site, and orbit type were found to have a significant influence on the success of SpaceX launches.
- The interactive visualizations provided a clear understanding of launch outcomes based on different filters such as launch site, payload mass, and booster versions.
- Recommendations for future improvements: More data should be collected to improve the accuracy of classification model on the test set.

Appendix

- For more info, refer the repo in GitHub.
- https://github.com/WeebOppa/FP-DS-Capstone-by-IBM

Thank you!