IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Stephen Tyler Williams
16 June 2024

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection & Wrangling:
    - Web scraping (BeautifulSoup)
    - API requests (Space X)
    - Data normalization from json
    - Missing value Imputation
    - Preliminary data profiling
  - EDA
    - Visualization (Scatter, Bar, etc charts)
    - SQL queries
    - Descriptive stats
    - Time Series Visualization
  - Geo-spatial EDA
    - Inspection of geo-spatially potentially relevant landmarks influencing launch & landing success
  - Dashboard development to probe success by launch site and payload mass
  - Classification Model (Base —> Optimized Model)

- Summary of all results
  - EDA revealed possible correlations between a number of variables which stood to inform the variables selected for the final classification model.
  - Visualization were used to investigate largely categorical variables.
  - In total the following columns were chosen as features:
    - ['FlightNumber', 'Date', 'BoosterVersion', 'PayloadMass', 'Orbit', 'LaunchSite', 'Outcome', 'Flights', 'GridFins', 'Reused', 'Legs','LandingPad', 'Block', 'ReusedCount', 'Serial', 'Longitude', 'Latitude', 'Class'].
  - Four classification models were fit and optimized. The decision tree produces the best results on both the train and test set showing little to no overfitting with an out-of-dataset accuracy of 88.93%

# Introduction

- Space Y aims to be the US's leading commercialized space agency.
- Facing strong competition from the first-to-market vendor Space X.
- Using Space X data to build a model that predicts rocket landing success.
- This prediction will help forecast costs, as first-stage landing significantly impacts overall mission expenses.

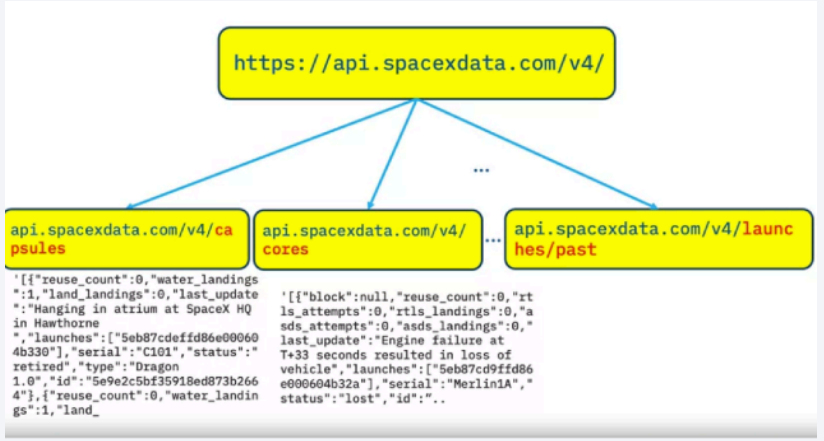
Section 1

# Methodology

# Methodology

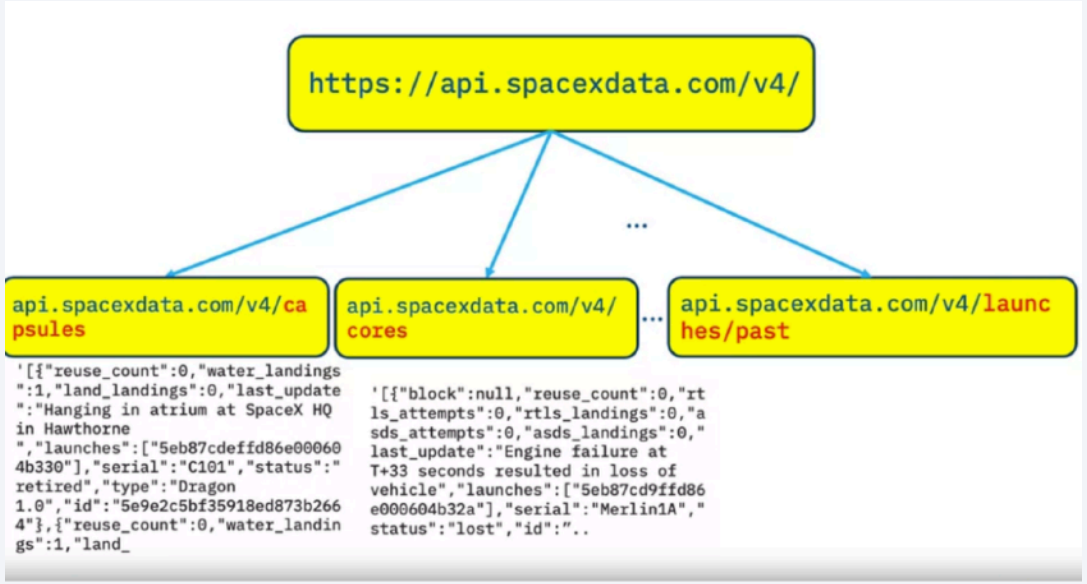
## Executive Summary

- Data collection methodology:
  - Data were collected using publicly available websites and APIs
    - APIs: url = “https://api.spacexdata.com/v4/launches/past" with subsequent requests for submodules
    - Web scraping: Publicly available wiki tables via BeutifulSoup
- Perform data wrangling
  - In order to examine the relationship between when a booster did and did not land successfully
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Feature selection
  - Model selection
  - GridSearch Cross-Validation Parameter Selection
  - Model Fitting
  - Scoring on a test dataset.




# Data Collection

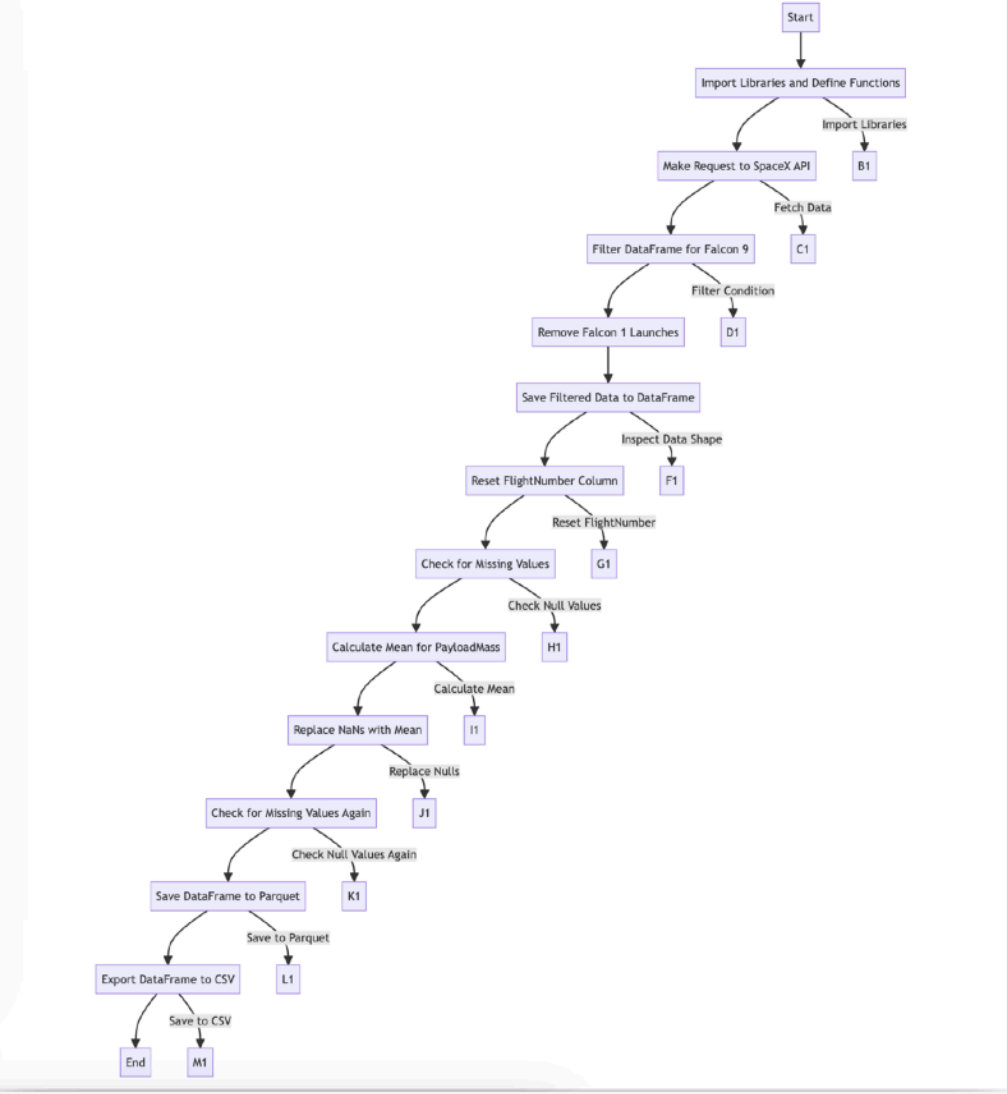
- Data were collected through multiple API requests made to the module
- Primary keys were then used to gather further information, e.g., capsules, cores, and past launch information
- Data were subsequently cleaned to produce a total of more than 100 historical launches.



# Data Collection – SpaceX API

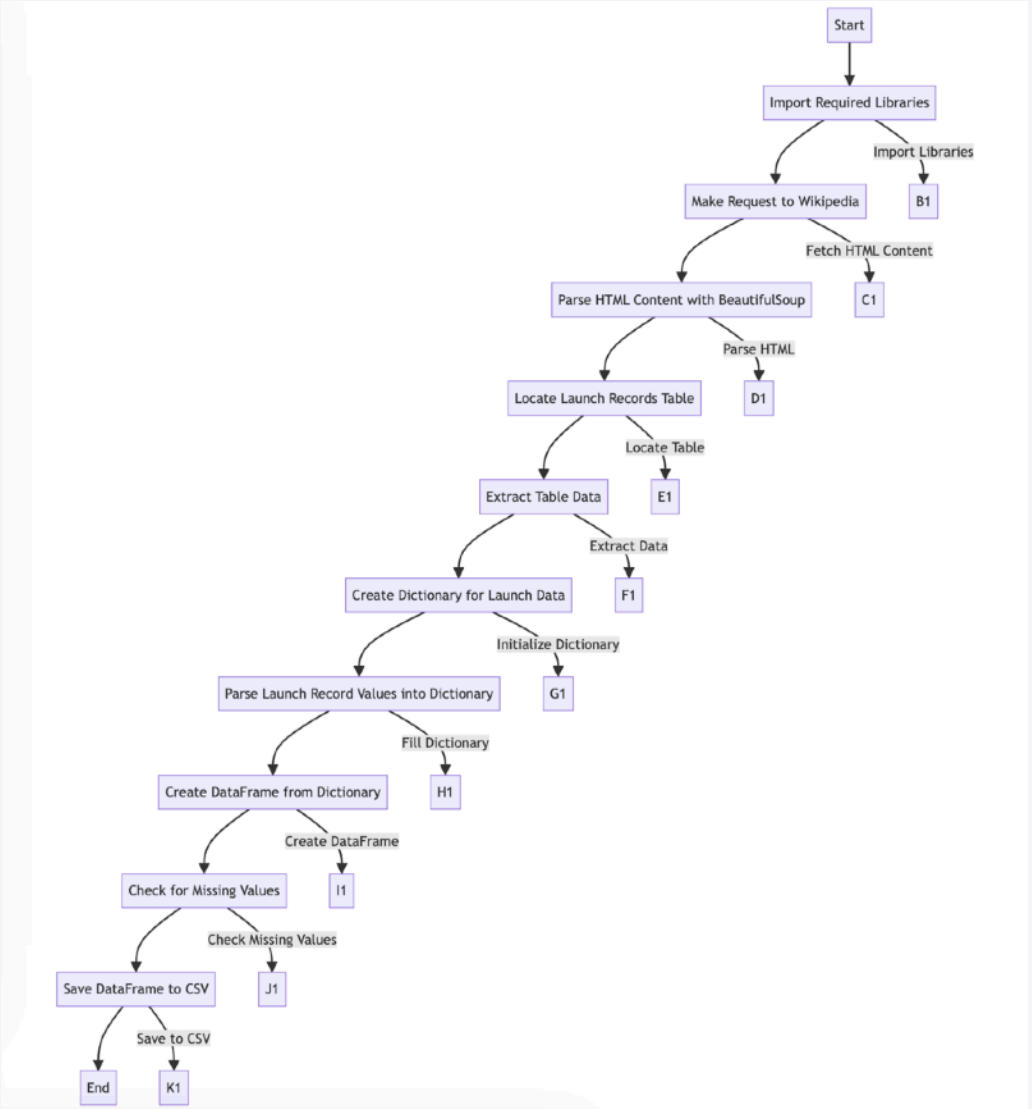
1. **Start**: Begin the data processing workflow.
2. **Import Libraries and Define Functions**: Load necessary libraries and define auxiliary functions to be used throughout the notebook.
3. **Make Request to SpaceX API**: Fetch data from the SpaceX API to obtain launch data.
4. **Filter DataFrame for Falcon 9**: Filter the data to include only Falcon 9 launches.
5. **Remove Falcon 1 Launches**: Exclude Falcon 1 launches from the dataset.
6. **Save Filtered Data to DataFrame**: Store the filtered data in a new DataFrame called `data_falcon9`.
7. **Reset FlightNumber Column**: Reset the `FlightNumber` column to ensure it is sequential.
8. **Check for Missing Values**: Identify any missing values in the DataFrame.
9. **Calculate Mean for PayloadMass**: Compute the mean value of the `PayloadMass` column.
10. **Replace NaNs with Mean**: Replace any missing values in the `PayloadMass` column with the calculated mean.
11. **Check for Missing Values Again**: Verify that there are no remaining missing values in the DataFrame.
12. **Save DataFrame to Parquet**: Export the cleaned DataFrame to a Parquet file for further analysis.
13. **Export DataFrame to CSV**: Save the DataFrame to a CSV file for consistency and future use.
14. **End**: End the data processing workflow.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git



# Data Collection - Scraping

1. **Start**: Begin the web scraping workflow.
2. **Import Required Libraries**: Load necessary libraries for web scraping and data manipulation (requests, BeautifulSoup, pandas).
3. **Make Request to Wikipedia**: Fetch the HTML content of the Wikipedia page containing Falcon 9 launch records.
4. **Parse HTML Content with BeautifulSoup**: Parse the fetched HTML content using BeautifulSoup.
5. **Locate Launch Records Table**: Locate the HTML table containing the launch records.
6. **Extract Table Data**: Extract data from the table by iterating over its rows and columns.
7. **Create Dictionary for Launch Data**: Initialize a dictionary to store the parsed launch data.
8. **Parse Launch Record Values into Dictionary**: Fill the dictionary with parsed launch record values.
9. **Create DataFrame from Dictionary**: Convert the dictionary to a pandas DataFrame.
10. **Check for Missing Values**: Check the DataFrame for any missing values.
11. **Save DataFrame to CSV**: Save the cleaned DataFrame to a CSV file.
12. **End**: End the web scraping workflow.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Data Wrangling

**Start**: Begin the data wrangling workflow.
**Import Required Libraries**: Load necessary libraries for data manipulation and visualization.
**Load Data**: Load the SpaceX launch data from a CSV file.
**Data Exploration**: Perform exploratory data analysis (EDA) to understand the structure and content of the data.

- **Check Data Types and Missing Values**: Inspect the data types of columns and identify any missing values.
- **Statistical Summary**: Generate a statistical summary of the numerical columns.

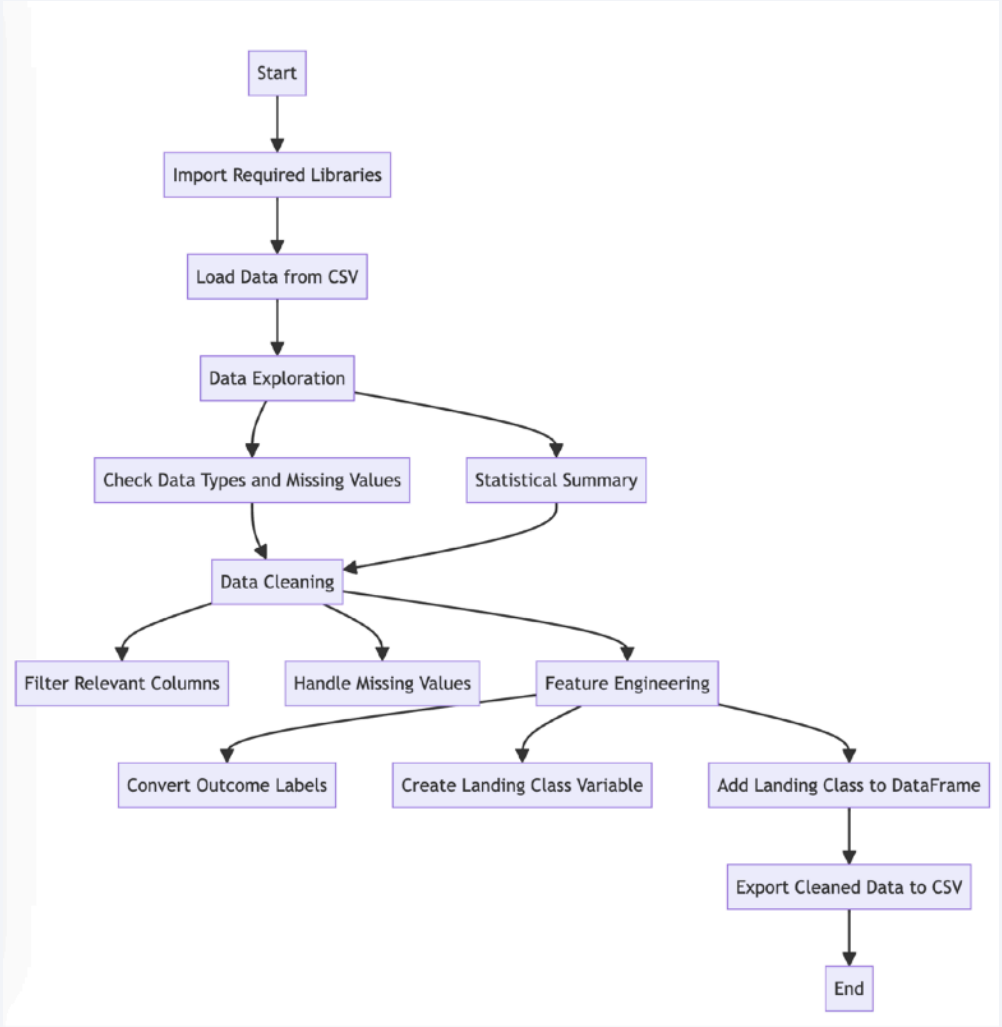
**Data Cleaning**: Clean the data to prepare it for analysis.

- **Filter Relevant Columns**: Select columns relevant to the analysis.
- **Handle Missing Values**: Impute or remove missing values in the dataset.

**Feature Engineering**: Create new features from existing data.

- **Convert Outcome Labels**: Convert mission outcome labels to binary format for classification.
- **Create Landing Class Variable**: Define a new variable `landing_class` to represent the landing outcome.
- **Add Landing Class to DataFrame**: Add the `landing_class` variable to the DataFrame.

**Export Cleaned Data**: Save the cleaned and processed data to a CSV file for further analysis.

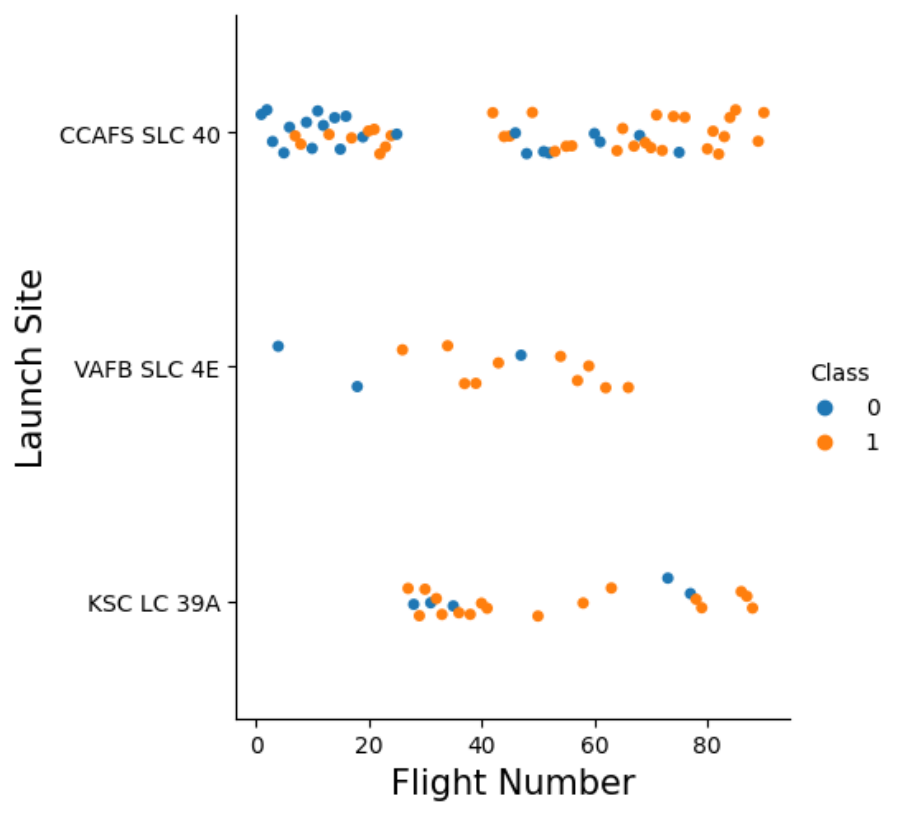



**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# EDA with Data Visualization

**Visualization Descriptions**

1. **Histogram of Launch Site Distribution**
   - **Visualization**: Histogram
   - **Description**: This histogram displays the distribution of launches across different launch sites. It is important to examine this distribution to understand which launch sites are most frequently used and to identify any potential biases in the data.
2. **Pie Chart of Landing Outcomes**
   - **Visualization**: Pie Chart
   - **Description**: The pie chart illustrates the proportion of different landing outcomes (e.g., success, failure, no attempt). This visualization is crucial for understanding the overall success rate of the Falcon 9 first stage landings and identifying areas for improvement.
3. **Box Plot of Payload Mass for Each Orbit**
   - **Visualization**: Box Plot
   - **Description**: This box plot shows the distribution of payload mass for each type of orbit (e.g., LEO, GTO). It highlights the variations in payload mass and helps identify any outliers. This visualization is important for understanding the payload capacity for different orbits and optimizing launch strategies.
4. **Scatter Plot of Payload Mass vs. Launch Outcome**
   - **Visualization**: Scatter Plot
   - **Description**: The scatter plot displays the relationship between payload mass and launch outcome (e.g., success, failure). It helps to identify if there is a correlation between the payload mass and the likelihood of a successful landing. This is important for risk assessment and mission planning.
5. **Bar Chart of Launch Outcomes by Year**
   - **Visualization**: Bar Chart
   - **Description**: This bar chart shows the number of launches and their outcomes (e.g., success, failure) by year. It helps to identify trends over time and evaluate the performance improvements or setbacks in Falcon 9 launches. This visualization is essential for assessing the progress of SpaceX's launch programs.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# EDA with SQL

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

### Task 1: Query to Count Launch Outcomes by Site

- **Description**: This query counts the number of launches for each launch site.
- **Importance**: It helps in understanding the distribution of launches across different sites, which is important for logistical and operational planning.

### Task 2: Query to Retrieve Launch Sites with Maximum Payload Mass

- **Description**: This query retrieves the launch sites and the associated maximum payload mass for each site.
- **Importance**: It helps identify which sites are used for heavier payloads, which is useful for site capacity planning and optimization.

### Task 3: Query to Count Successful and Failed Landings

- **Description**: This query counts the number of successful and failed landings for each booster version.
- **Importance**: It provides insights into the reliability and performance of different booster versions, which is crucial for improving future designs and launches.

### Task 4: Query to Find the Total Payload Mass Carried by Each Booster Version

- **Description**: This query sums the payload mass carried by each booster version.
- **Importance**: It helps assess the load capacity and performance of each booster version, which is important for mission planning and booster selection.

### Task 5: Query to Calculate the Average Payload Mass for Successful Landings

- **Description**: This query calculates the average payload mass for missions that had successful landings.
- **Importance**: It provides an understanding of the payload capacity associated with successful landings, which can inform future payload and mission planning.

### Task 6: Query to Retrieve Launch Outcomes Grouped by Year

- **Description**: This query groups the launch outcomes by year and counts the number of each outcome.
- **Importance**: It helps identify trends and patterns in launch outcomes over time, which is important for evaluating progress and improvements in launch success rates.

### Task 7: Query to Find the Most Common Launch Site

- **Description**: This query identifies the launch site with the highest number of launches.
- **Importance**: It highlights the most frequently used launch site, which is useful for resource allocation and infrastructure development.

### Task 8: Query to Retrieve the Number of Landings for Each Landing Outcome

- **Description**: This query counts the number of each type of landing outcome (e.g., success, failure).
- **Importance**: It provides an overview of the landing success rate, which is crucial for evaluating the effectiveness of landing techniques and technologies.

### Task 9: Query to Calculate the Success Rate of Landings

- **Description**: This query calculates the percentage of successful landings out of the total number of landings.
- **Importance**: It helps quantify the overall success rate of landings, which is important for performance evaluation and reporting.

### Task 10: Query to Retrieve the Number of Successful Landings by Booster Version

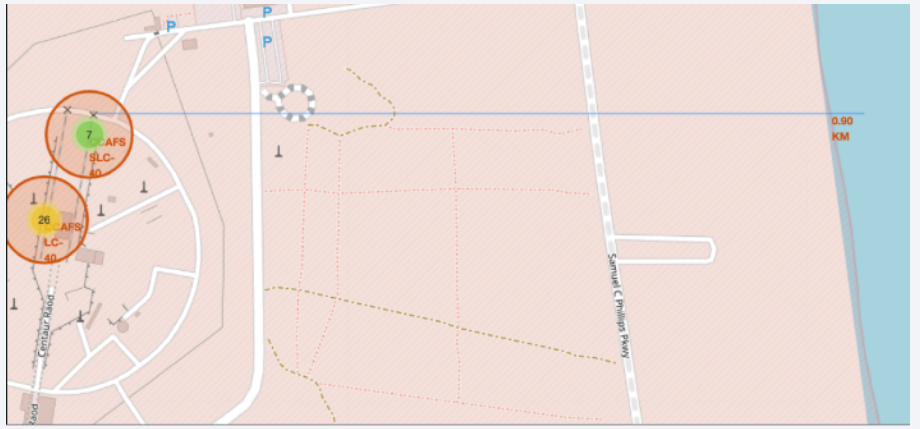
- **Description**: This query counts the number of successful landings for each booster version.
- **Importance**: It provides insights into the reliability of different booster versions, which is essential for decision-making in future booster development and deployment.

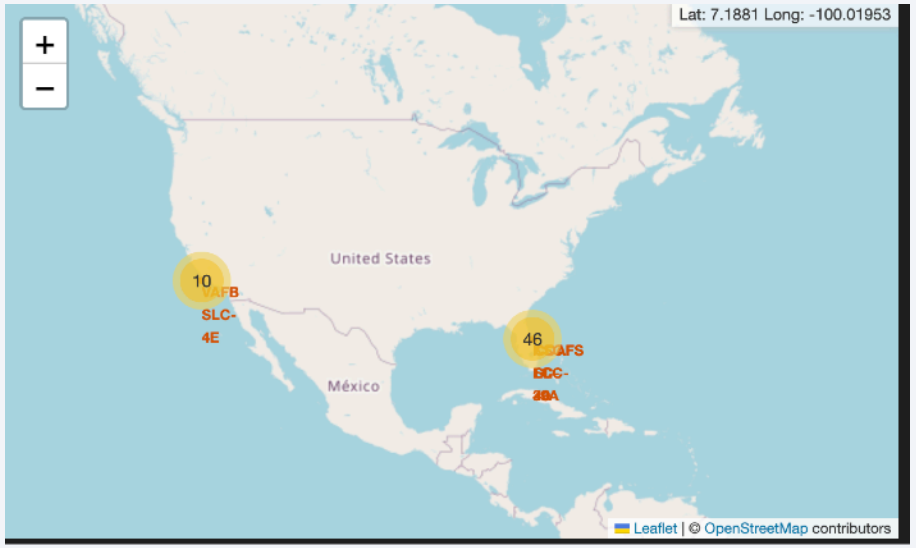
# Build an Interactive Map with Folium

**Map Objects Created and Added to the Folium Map**

1. **Markers for Launch Sites**
   - **Description**: Markers were added to the map to represent the locations of different SpaceX launch sites.
   - **Reason**: These markers help visualize the geographical distribution of launch sites, making it easier to analyze their spatial relationships and potential impact on launch outcomes.
2. **Circle Markers for Launch Success and Failures**
   - **Description**: Circle markers were added to indicate the success or failure of each launch at the respective launch sites. Green circles indicate successful launches, while red circles indicate failed launches.
   - **Reason**: This visualization helps in quickly identifying the success rates of launches at different sites and allows for a more detailed examination of patterns and trends in launch outcomes.
3. **Lines for Distances Between Launch Sites and Proximities**
   - **Description**: Lines were drawn to represent the distances between each launch site and its proximities (e.g., nearby cities, airports, or other significant landmarks).
   - **Reason**: These lines help in analyzing the potential influence of nearby geographical features on launch outcomes, such as the effect of proximity to populated areas or critical infrastructure.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git



# Build a Dashboard with Plotly Dash

**Dropdown for Launch Site Selection**

- **Description**: A dropdown menu that allows users to select a specific launch site or view data for all sites.
- **Importance**: This interaction enables users to filter the data based on the launch site, providing a customized view of the launch records for more detailed analysis.

**Pie Chart for Total Successful Launches Count**

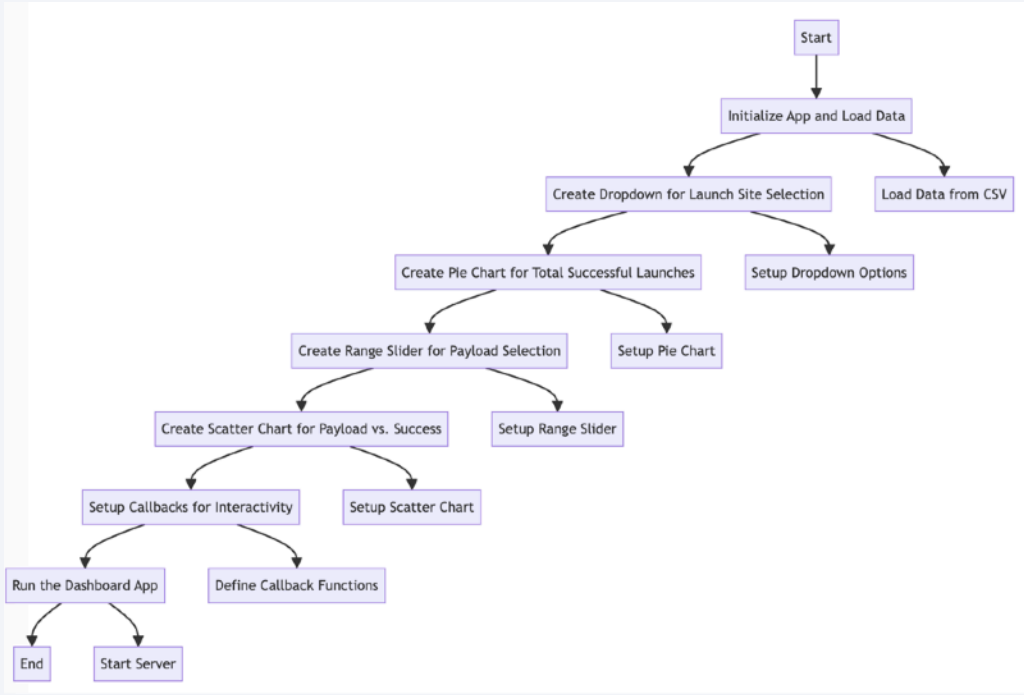
- **Description**: A pie chart that displays the proportion of successful launches.
- **Importance**: The pie chart provides a quick visual summary of the success rate of launches, helping users to assess the overall performance at a glance.

**Range Slider for Payload Selection**

- **Description**: A range slider that allows users to select a payload mass range.
- **Importance**: This interaction helps users to filter the data based on payload mass, enabling the analysis of launch outcomes for different payload ranges and identifying trends related to payload size.

**Scatter Chart for Correlation Between Payload and Launch Success**

- **Description**: A scatter chart that shows the relationship between payload mass and launch success.
- **Importance**: The scatter plot helps in understanding if there is any correlation between the payload mass and the success rate of launches. It is crucial for identifying potential factors that affect launch outcomes.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Predictive Analysis (Classification)

**Data Preprocessing**:

- **Description**: The data was cleaned, missing values were handled, and features were selected for the model.
- **Importance**: Ensures the data is in the correct format and all necessary information is available for the model to learn effectively.

**Feature Engineering**:

- **Description**: New features were created from the existing data to improve model performance.
- **Importance**: Helps in providing the model with more relevant information, potentially improving its accuracy.

**Splitting the Data**:

- **Description**: The data was split into training and testing sets.
- **Importance**: Allows the evaluation of the model on unseen data to assess its generalization ability.

**Model Building**:

- **Description**: Several classification models were built using algorithms like Logistic Regression, SVM, KNN, and Decision Trees.
- **Importance**: Provides a variety of models to compare and select the best performing one.

**Hyperparameter Tuning**:

- **Description**: Hyperparameters of the models were tuned using GridSearchCV.
- **Importance**: Finds the optimal parameters for each model to maximize performance.

**Model Evaluation**:

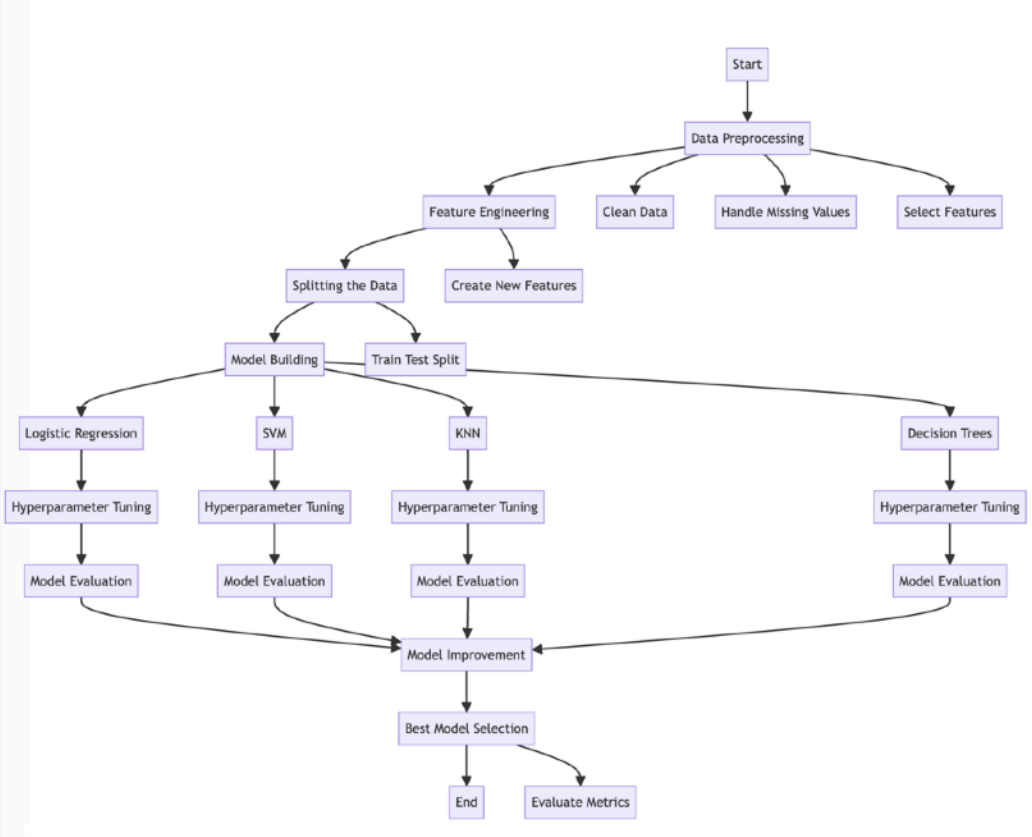
- **Description**: Models were evaluated using metrics .score() based on the _best_estimates_ attribute
- **Importance**: Allows comparison of models to determine which one performs best on the test set.

**Model Improvement**:

- **Description**: The models were iteratively improved based on the evaluation metrics.
- **Importance**: Ensures that the best performing model is as accurate and reliable as possible.

**Best Model Selection**:

- **Description**: The model with the highest evaluation metrics was selected as the best performing model.
- **Importance**: Provides a final model that is ready for deployment and can make accurate predictions.
-




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Results

### Exploratory Data Analysis (EDA) Results

- **Launch Site Distribution**:
  - Histogram showing the number of launches at each site.
  - **Result**: Most launches occurred at CCAFS SLC-40.
- **Landing Outcomes**:
  - Pie chart of successful vs. failed landings.
  - **Result**: Approximately 60% success rate for landings.
- **Payload Mass vs. Launch Outcome**:
  - Scatter plot correlating payload mass with launch success.
  - **Result**: No clear correlation observed between payload mass and launch success.

### Interactive Analytics Demo

1. **Launch Site Dropdown**:
   - Filter launches by site.
   - **Result**: Users can analyze data specific to each launch site.
2. **Success Pie Chart**:
   - Visualize proportion of successful launches.
   - **Result**: Immediate understanding of overall launch success rates.
3. **Payload Range Slider**:
   - Filter data by payload mass.
   - **Result**: Examine launch success trends for different payload sizes.
4. **Payload vs. Success Scatter Chart**:
   - Show relationship between payload and success.
   - **Result**: Analyze potential factors affecting launch outcomes.

### Predictive Analysis Results

1. **Model Building**:
   - Logistic Regression, SVM, KNN, Decision Trees.
   - **Result**: Developed and compared multiple models.
2. **Hyperparameter Tuning**:
   - Used GridSearchCV.
   - **Result**: Found optimal parameters to enhance model performance.
3. **Model Evaluation**:
   - Accuracy on test set using `.score()`.
   - **Result**: Decision Tree Model achieved the highest accuracy on the test set
4. **Best Model Selection**:
   - Selected the Decision Tree Model.
   - **Result**: Decision Tree Mode was the best performing model with an accuracy of > 88% on the test set.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

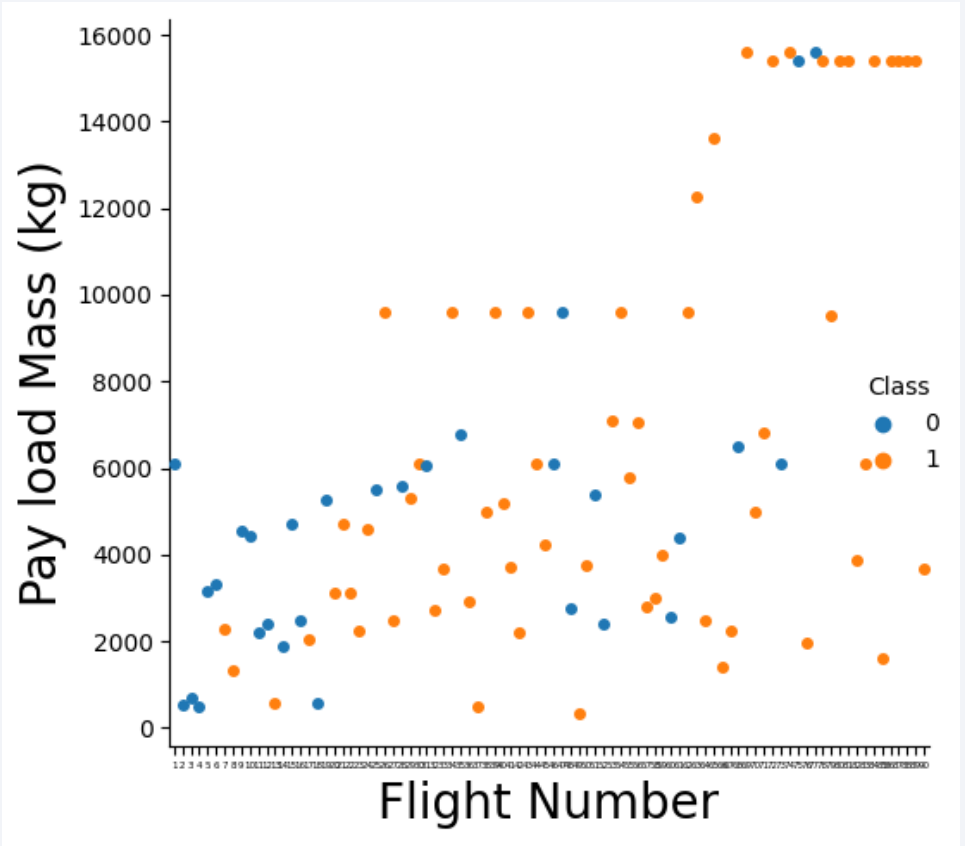
Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

**Key Observations**

1. **Trend Over Flight Numbers**:
   - Earlier flight numbers (lower values) show a mix of successful and unsuccessful landings with lower payload masses.
   - As flight numbers increase, there is a trend toward higher payload masses and more successful landings (orange dots).
2. **Payload Mass Distribution**:
   - Successful landings (orange) are observed across a wide range of payload masses, from low to very high values (up to 16,000 kg).
   - Unsuccessful landings (blue) are primarily clustered around lower to mid-range payload masses, with fewer occurrences at higher payloads.
3. **High Payload Mass Region**:
   - At the higher end of payload masses (above 10,000 kg), successful landings are more frequent than unsuccessful ones.
   - There is a significant cluster of successful landings around the 16,000 kg payload mass, suggesting a possible improvement in handling higher payloads over time.
4. **Overall Success Rate**:
   - The distribution suggests an overall improvement in landing success rates over time, as indicated by the increasing frequency of orange dots in later flights.
   - Earlier flights show more variability in outcomes, indicating potential initial challenges that were overcome in subsequent missions.
5. **Potential Correlation**:
   - There seems to be a potential positive correlation between flight number and the likelihood of a successful landing, possibly due to technological advancements and experience gained over time.

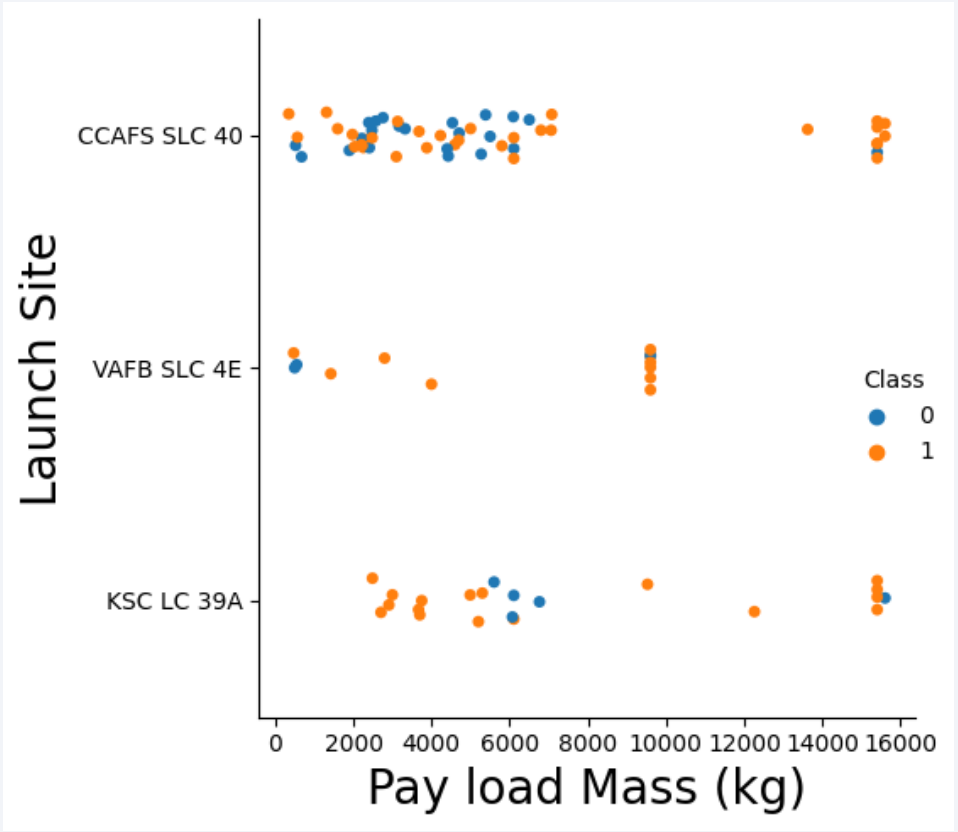



**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Payload vs. Launch Site

**Key Observations**

1. **CCAFS SLC 40**:
    - Most launches are clustered around lower payload masses (below 7000 kg) with a mix of successes and failures.
    - Successful landings are more frequent, indicated by a higher number of orange dots.
    - Notable outliers of successful landings around 16,000 kg payload mass.
2. **VAFB SLC 4E**:
    - Launches are more spread out across different payload masses.
    - A significant number of launches with higher payloads (up to 10,000 kg) show successful landings.
    - Fewer overall data points compared to other sites, but higher success rate is visible.
3. **KSC LC 39A**:
    - Shows a diverse range of payload masses, including very high payloads (up to 16,000 kg).
    - Successful landings dominate, especially at lower and mid-range payload masses.
    - Some unsuccessful landings are still present at lower payloads, indicating ongoing challenges.
    -

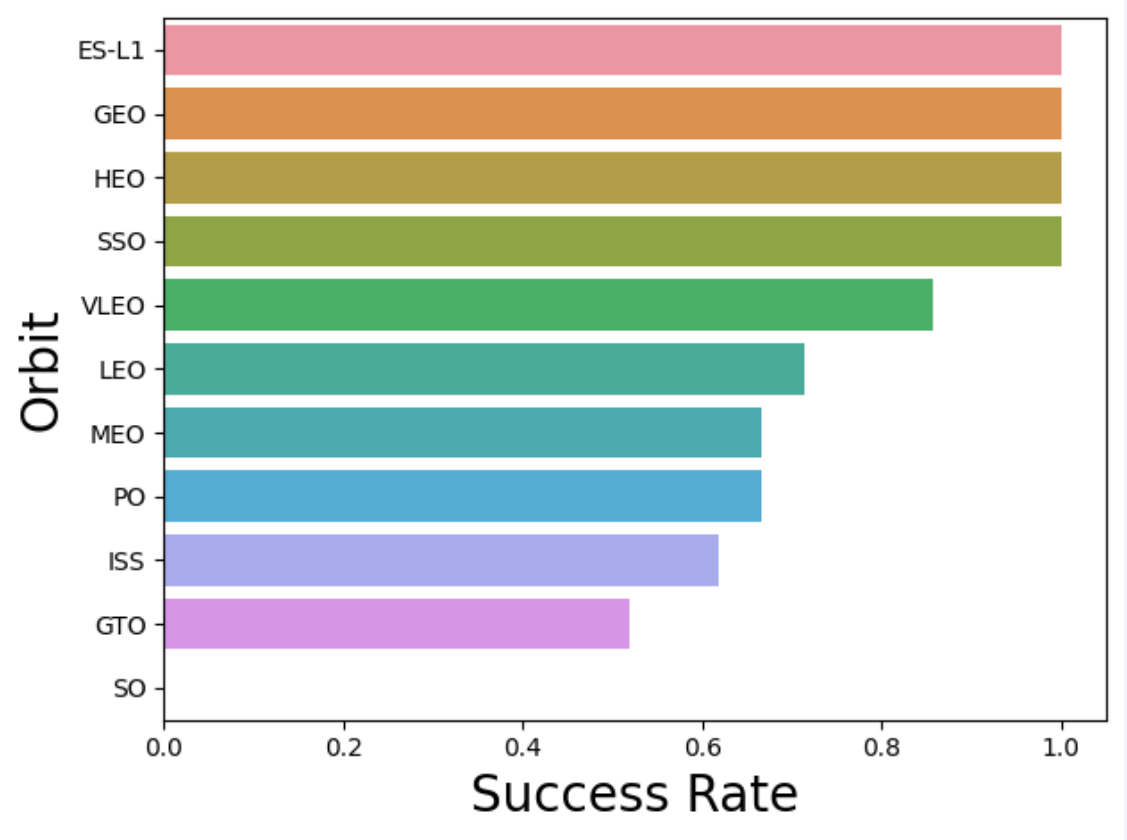



**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Success Rate vs. Orbit Type

**Key Observations from the Bar Chart**

- **ES-L1, GEO, HEO, SSO**: All have a 100% success rate.
- **VLEO**: High success rate, slightly below 100%.
- **LEO, MEO, PO**: Moderate success rates, around 70%-80%.
- **ISS**: Lower success rate, approximately 50%.
- **GTO**: Success rate around 30%.
- **SO**: Lowest success rate, close to 20%.




**Conclusion**

The chart highlights that ES-L1, GEO, HEO, and SSO orbits have perfect success rates, while GTO and SO orbits present more significant challenges, reflected in their lower success rates.

# Flight Number vs. Orbit Type

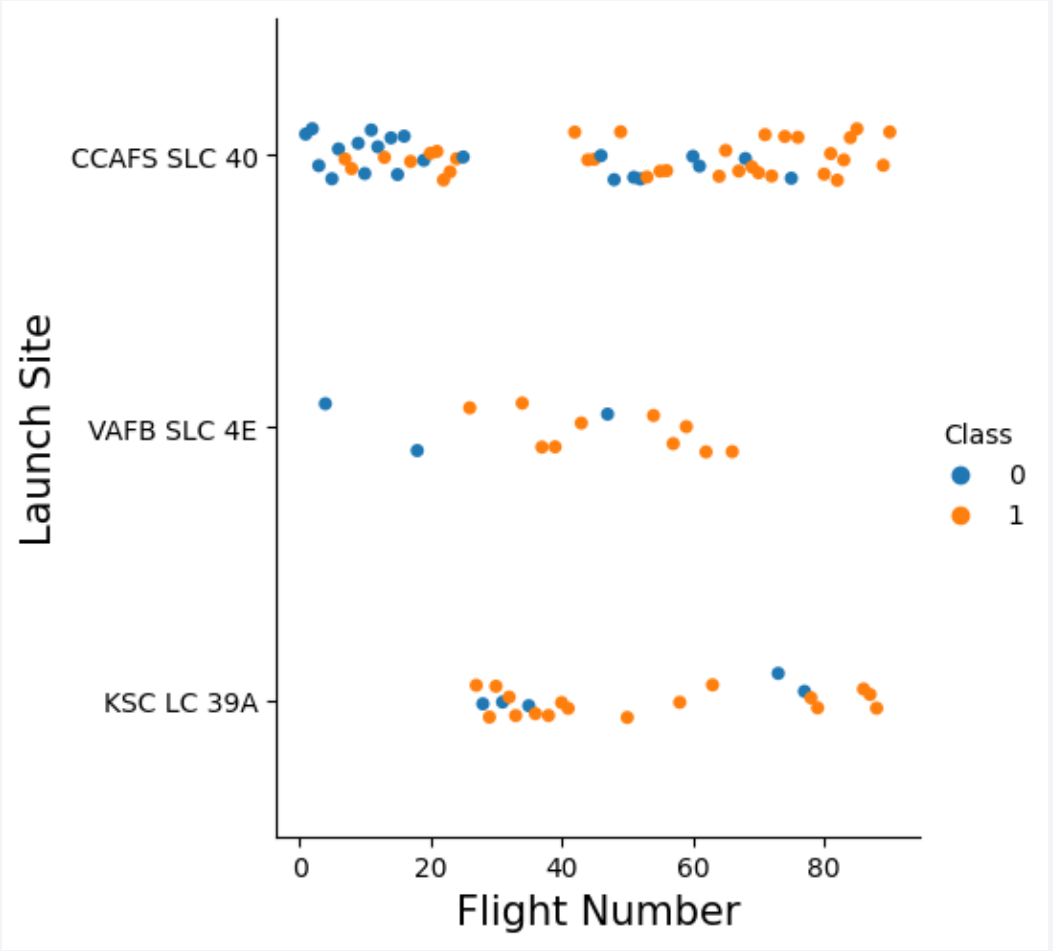
**Key Observations**

- **CCAFS SLC 40**:
  - High concentration of launches between Flight Numbers 0 to 40 and 60 to 90.
  - Early flights (0-20) show a mix of successes (orange) and failures (blue).
  - Later flights (60-90) show more successful landings (orange).
- **VAFB SLC 4E**:
  - Fewer launches compared to CCAFS SLC 40.
  - Consistent success (orange) observed from Flight Numbers 20 to 80.
  - Overall, a higher success rate (more orange) is visible.
- **KSC LC 39A**:
  - Launches start around Flight Number 20.
  - Mix of successes (orange) and failures (blue) throughout the flight numbers.
  - Notable clusters of successes around Flight Numbers 60 to 80.

**Conclusion**

- **CCAFS SLC 40**: High frequency of launches with improved success rates in later flights.
- **VAFB SLC 4E**: Fewer launches but higher consistency in successful landings.
- **KSC LC 39A**: Balanced mix of successes and failures, with visible improvement in later flights.

These observations indicate that over time, the success rate of launches improves, with each site showing its own trend and pattern of successes and failures.



# Payload vs. Orbit Type

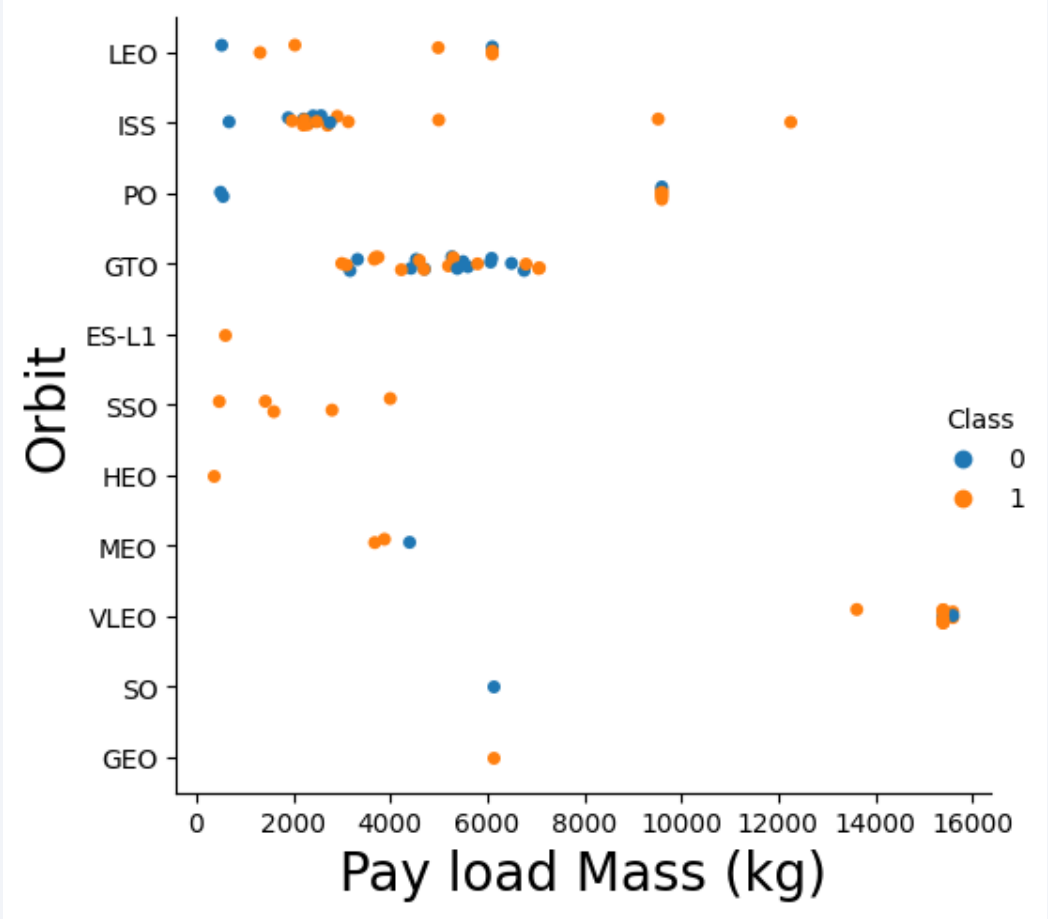
**Key Observations from the Scatter Plot**

- **LEO (Low Earth Orbit)**:
  - Wide range of payload masses with a mix of successes (orange) and failures (blue).
  - Successful landings (orange) are spread across low to high payload masses.
- **ISS (International Space Station)**:
  - Clustered around lower payload masses (up to 6000 kg).
  - Higher proportion of successful landings (orange) compared to failures.
- **PO (Polar Orbit)**:
  - Mixed success (orange) and failure (blue) rates, primarily around lower payload masses.
- **GTO (Geostationary Transfer Orbit)**:
  - Significant number of launches with payloads around 4000 to 6000 kg.
  - Mixed success and failure rates.
- **ES-L1, SSO, HEO, MEO**:
  - Higher success rates (more orange dots) across a range of payload masses.
  - Notably consistent success in SSO (Sun-Synchronous Orbit).
- **VLEO (Very Low Earth Orbit), SO (Sub-Orbital), GEO (Geostationary Orbit)**:
  - Fewer data points with varied payload masses.
  - High success rate (more orange) in GEO and VLEO.

**Conclusion**

- **LEO and ISS**: Show a broad range of payloads with relatively high success rates.
- **GTO**: Exhibits mixed results, indicating potential challenges with these missions.
- **SSO and GEO**: Demonstrate high success rates, suggesting effective handling of these orbits.
- **Overall**: Success rates vary significantly across different orbits and payload masses, with some orbits (e.g., SSO, GEO) showing consistently better outcomes.

These observations highlight the varying challenges and successes associated with different orbital missions, providing insights into where SpaceX has been most and least successful.



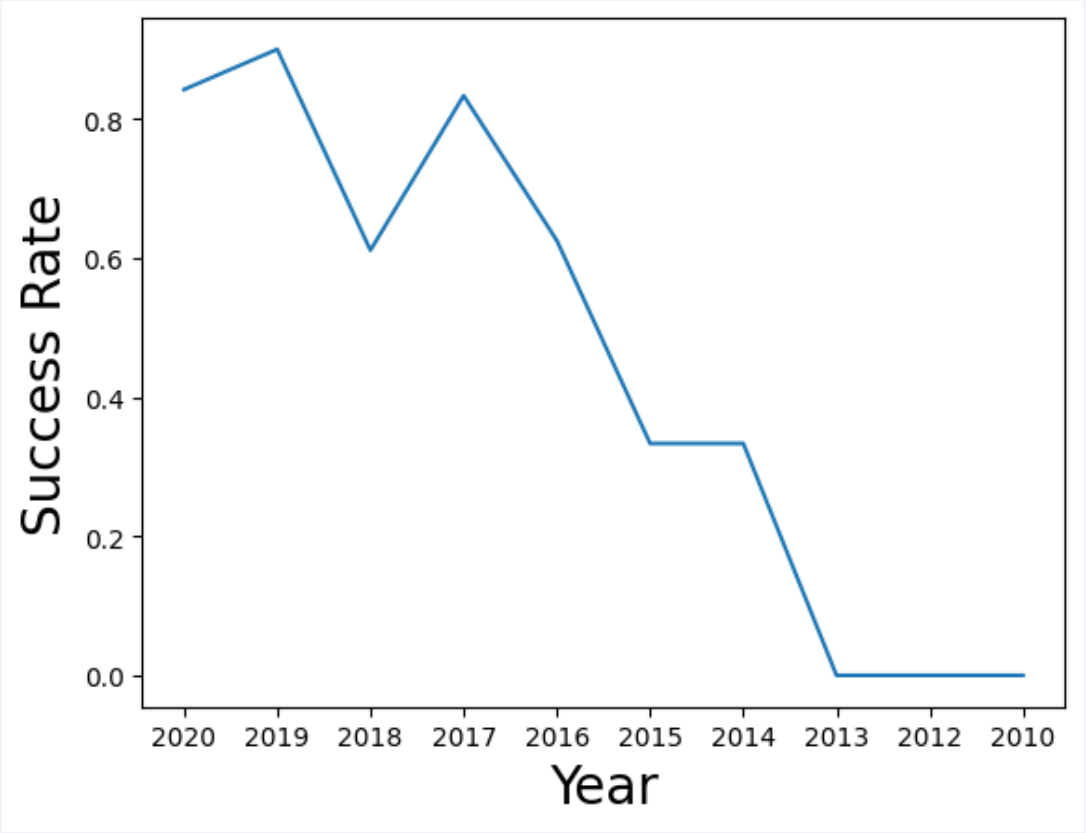
# Launch Success Yearly Trend

**Key Observations from the Line Chart**

- **Trend Over Years**:
  - **2016-2020**: High success rates, mostly above 60%, peaking near 90% in some years.
  - **2015-2016**: Gradual decline in success rates.
  - **2014**: Sharp drop to approximately 20%.
  - **2013 and Earlier**: Success rates approach zero.

**Conclusion**

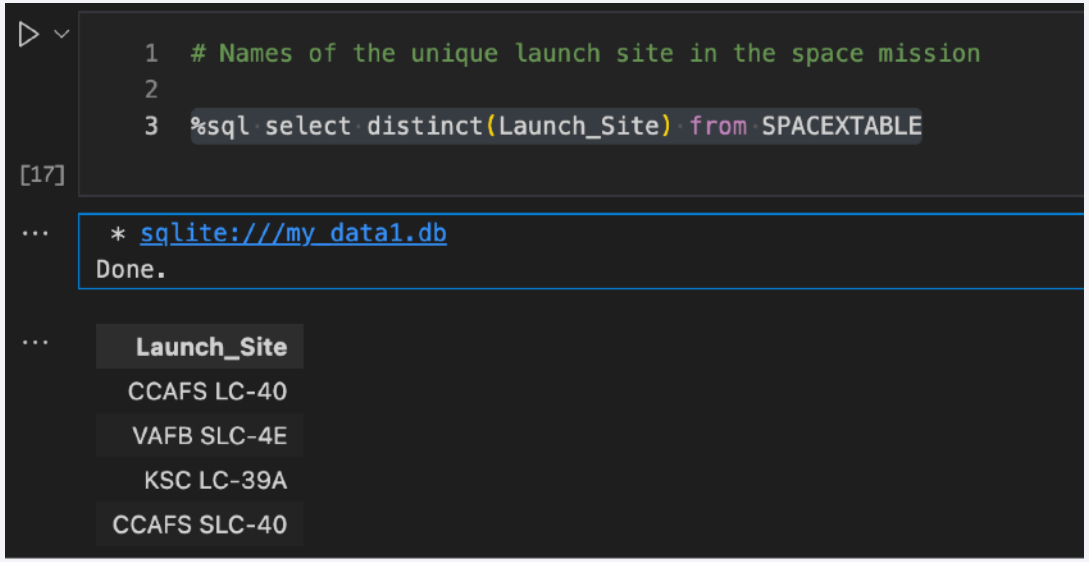
- The success rate of launches significantly improved from 2013 to 2016, peaking around 2018-2019.
- After 2015, there was a notable decline, hitting a low in 2014.
- The general trend shows substantial improvements in recent years compared to earlier years, reflecting advancements in technology and operational efficiency.



# All Launch Site Names

**Results:**

1. **CCAFS LC-40**:
   - Cape Canaveral Air Force Station Launch Complex 40.
   - One of the primary launch sites for SpaceX, used frequently for Falcon 9 launches.
2. **VAFB SLC-4E**:
   - Vandenberg Air Force Base Space Launch Complex 4E.
   - Used for polar orbit launches and other missions requiring a high-inclination trajectory.
3. **KSC LC-39A**:
   - Kennedy Space Center Launch Complex 39A.
   - A historic launch site, originally used for Apollo and Space Shuttle missions, now utilized by SpaceX for Falcon 9 and Falcon Heavy launches.
4. **CCAFS SLC-40**:
   - Cape Canaveral Air Force Station Space Launch Complex 40.
   - Likely a duplicate entry of CCAFS LC-40, indicating potential inconsistency in naming conventions within the dataset.




**Conclusion**

The distinct launch sites identified by the query are key locations used by SpaceX for launching their missions, each serving different types of launch requirements. The presence of a potential duplicate entry (CCAFS LC-40 and CCAFS SLC-40) suggests the need for consistent naming conventions in the dataset.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Launch Site Names Begin with 'CCA'

**Results:**

This query retrieves the first 5 records from the `SPACEXTABLE` where the `Launch_Site` begins with 'CCA'. This pattern matches the launch sites located at Cape Canaveral Air Force Station (CCAFS).

**Brief Explanation:**

- **Purpose**: The query filters the records to include only those launches that took place at Cape Canaveral Air Force Station, specifically at launch sites with names starting with 'CCA'.
- **Results**: The output will show the details of the first 5 launches from CCAFS, providing information such as launch date, booster version, payload, orbit, mission outcome, and landing outcome.

Task 2

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

```
1  # sleect launch_site begin with "CCA" LImit 5
2
3  %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_M |
|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | |

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

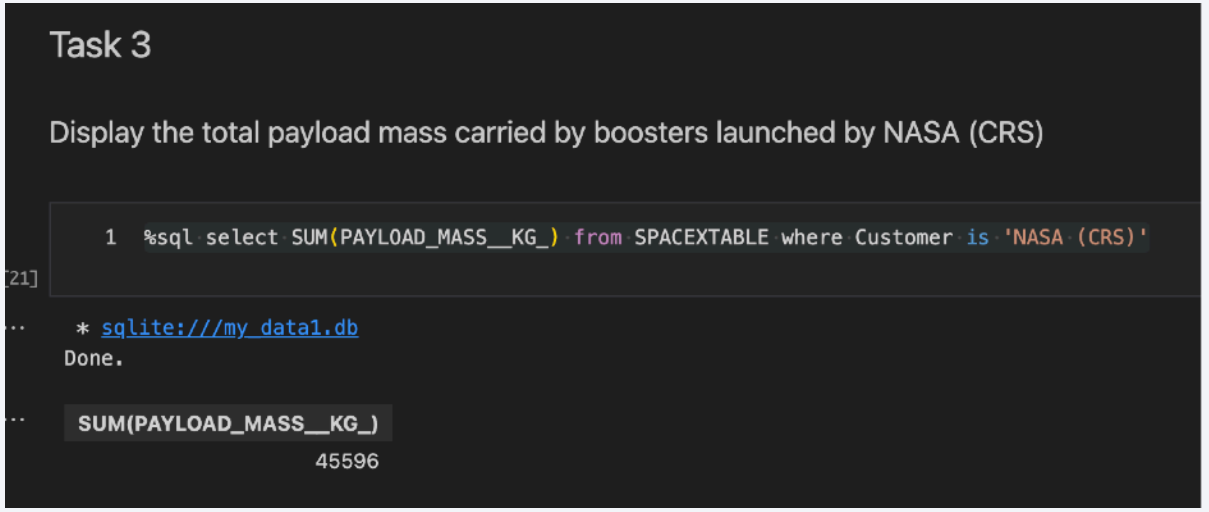
# Total Payload Mass

**Brief Explanation:**

- **Purpose**: The query calculates the total payload mass (in kilograms) for all launches where the customer is 'NASA (CRS)'.
- **Result**: The total payload mass for NASA (CRS) missions is 45,596 kg.

**Conclusion**

The query provides the cumulative payload mass for all missions conducted for NASA under the Commercial Resupply Services (CRS) program, highlighting the total mass of cargo transported to the International Space Station (ISS) by SpaceX for NASA.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

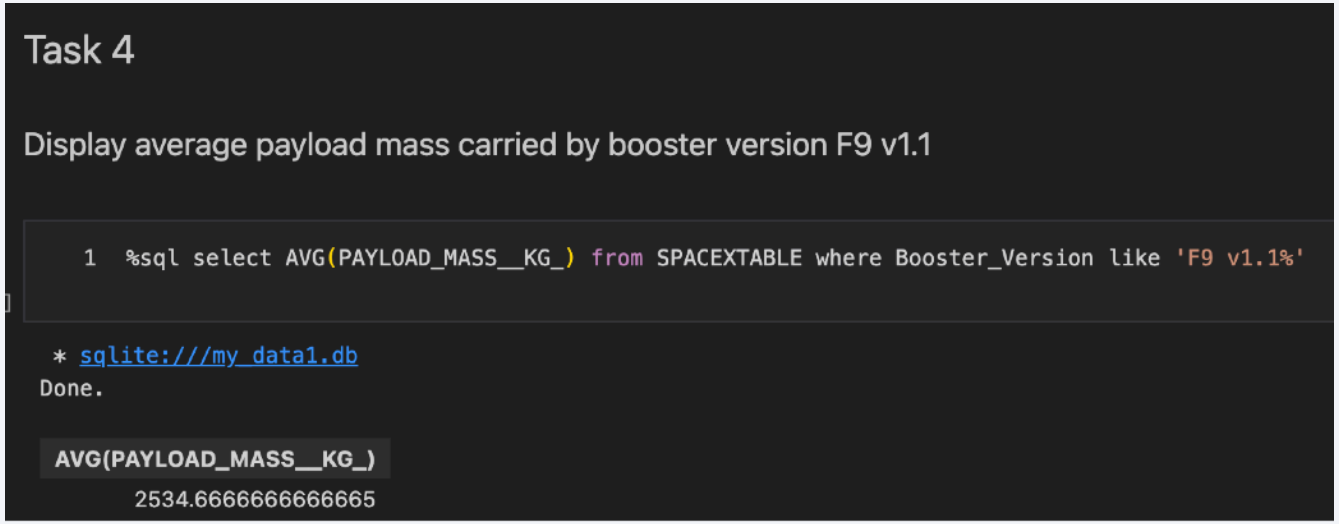
# Average Payload Mass by F9 v1.1

**Brief Explanation:**

- **Purpose**: The query calculates the average payload mass (in kilograms) for all launches that used the booster version starting with 'F9 v1.1'.
- **Result**: The average payload mass for these launches is approximately 2,534.67 kg.

**Conclusion**

The query provides the mean payload mass for launches utilizing the 'F9 v1.1' booster version, giving insights into the typical payload capacity for this specific version of the Falcon 9 rocket. This information is useful for analyzing the performance and capabilities of the 'F9 v1.1' booster in terms of payload delivery.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

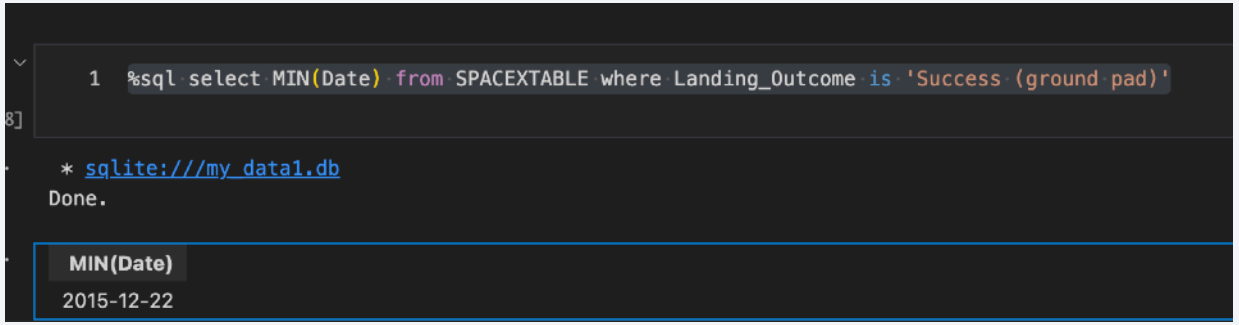
# First Successful Ground Landing Date

**Brief Explanation:**

- **Purpose**: The query retrieves the earliest (minimum) date from the `SPACEXTABLE` where the landing outcome was a successful landing on a ground pad.
- **Result**: The first successful ground landing occurred on December 22, 2015.

**Conclusion**

The query identifies the date of the first successful ground landing by SpaceX, which is a significant milestone in the company's history. This successful landing on December 22, 2015, marks the first time SpaceX achieved a ground pad landing, demonstrating their capability in reusable rocket technology.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

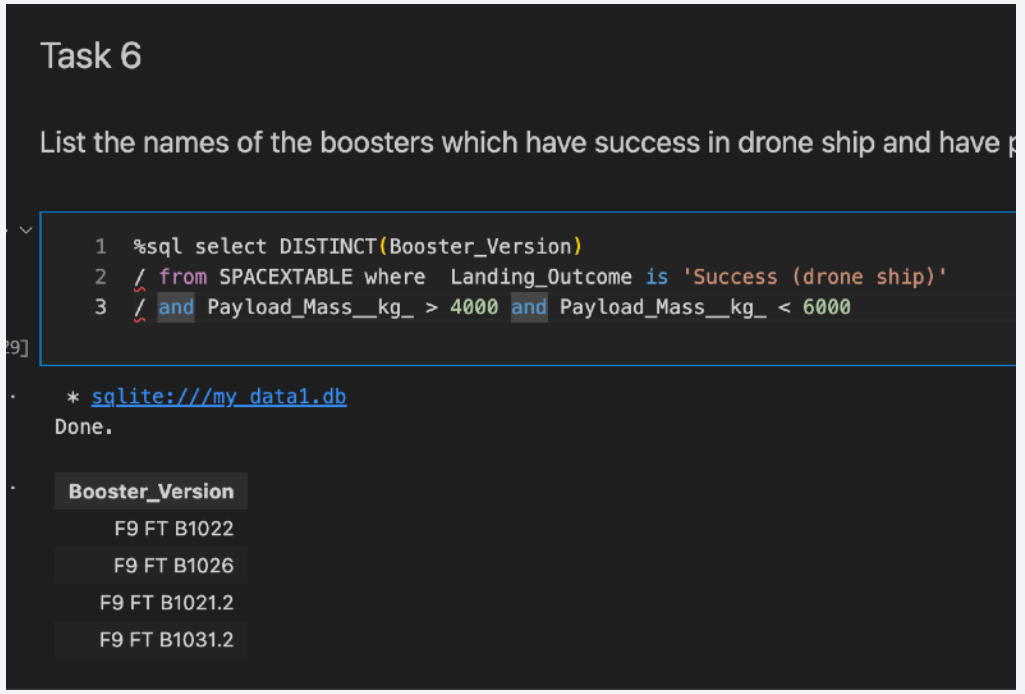
# Successful Drone Ship Landing with Payload between 4000 and 6000

**Brief Explanation:**

- **Purpose**: The query retrieves the distinct booster versions from the `SPACEXTABLE` that have successfully landed on a drone ship and carried a payload mass between 4000 kg and 6000 kg.
- **Result**: The booster versions that meet these criteria are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

**Conclusion**

The query identifies specific booster versions that have demonstrated successful drone ship landings with payloads in the 4000 to 6000 kg range. This information highlights the capabilities and reliability of these particular boosters in handling moderate payloads and achieving successful landings on drone ships.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

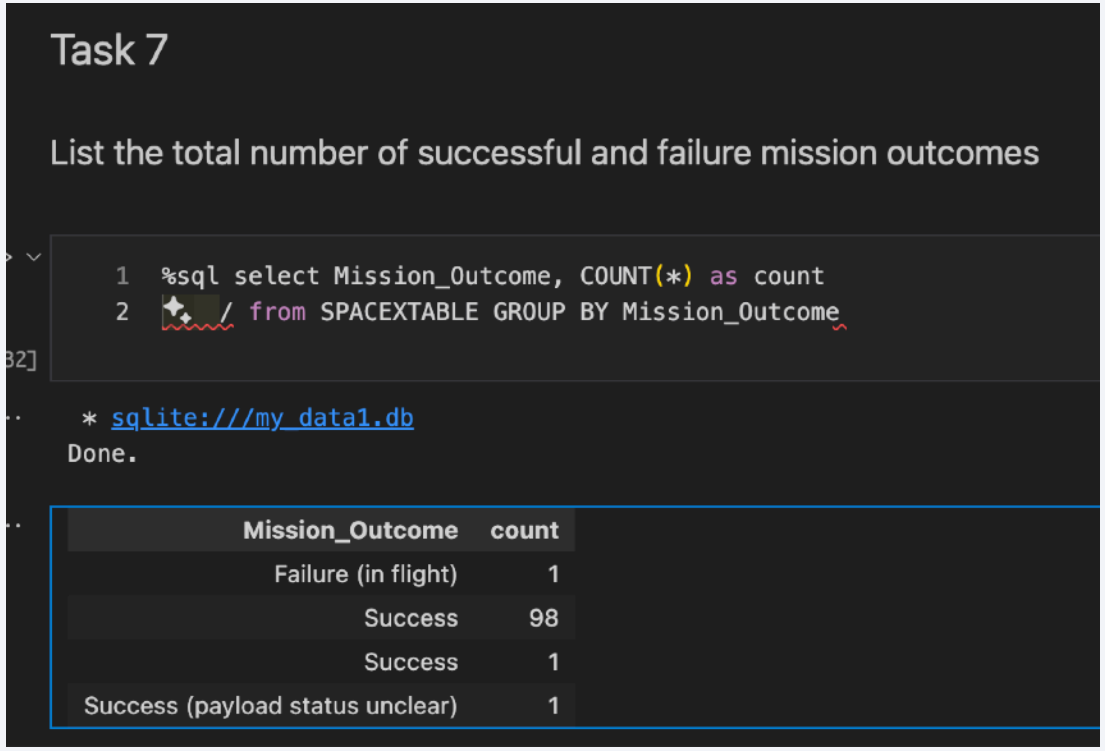
# Total Number of Successful and Failure Mission Outcomes

**Brief Explanation:**

- **Purpose**: The query counts the number of occurrences for each unique `Mission_Outcome` in the `SPACEXTABLE`.
- **Result**:
  - There is 1 mission with the outcome "Failure (in flight)".
  - There are 98 missions with the outcome "Success".
  - There is 1 mission with the outcome "Success".
  - There is 1 mission with the outcome "Success (payload status unclear)".

**Conclusion**

The query provides a summary of mission outcomes and their frequencies. It reveals that the majority of missions (98) were successful. There seems to be a duplicate entry for the "Success" outcome, indicating a potential inconsistency in the dataset. Additionally, there is one mission with an unclear payload status and one failure in flight. This summary helps in understanding the overall success rate and identifying any anomalies or inconsistencies in the mission outcome data.

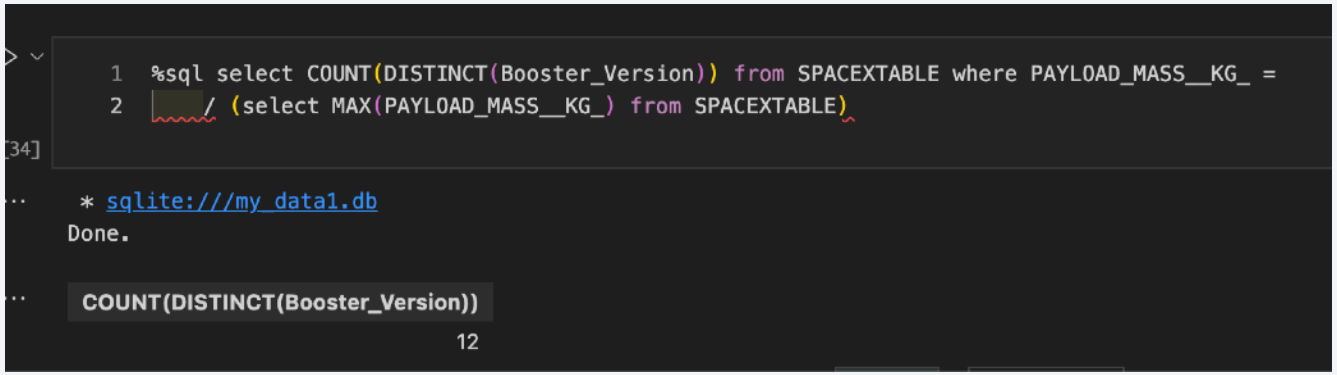



**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Boosters Carried Maximum Payload

**Brief Explanation:**

- **Purpose**: The query counts the number of distinct booster versions that have carried the maximum payload mass recorded in the `SPACEXTABLE`.
- **Result**: There are 12 distinct booster versions that have carried the maximum payload mass.




**Conclusion**

The query identifies that 12 different booster versions have been used to carry the heaviest payloads launched by SpaceX. This information highlights the versatility and capability of multiple booster versions in handling the heaviest payloads, indicating robust performance across various models in the fleet.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

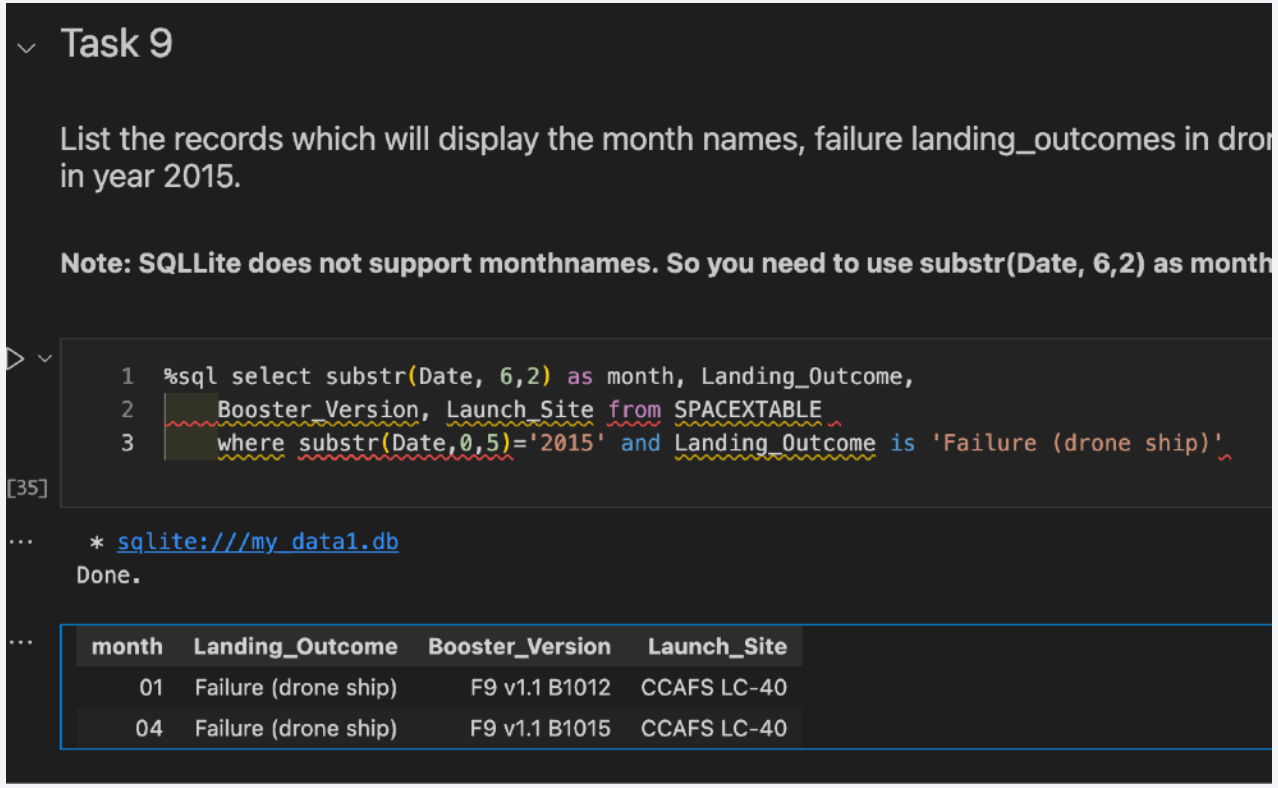
# 2015 Launch Records

**Brief Explanation:**

- **Purpose**: The query retrieves records of launches in 2015 that resulted in a "Failure (drone ship)" landing outcome. It extracts the month, landing outcome, booster version, and launch site for these records.
- **Result**:
  - In January 2015, the booster version F9 v1.1 B1012 launched from CCAFS LC-40 failed to land on a drone ship.
  - In April 2015, the booster version F9 v1.1 B1015 launched from CCAFS LC-40 also failed to land on a drone ship.

**Conclusion**

The query highlights two specific instances in 2015 where SpaceX experienced drone ship landing failures. Both failures occurred with the F9 v1.1 booster version launched from CCAFS LC-40 in January and April. This information can be used to analyze the conditions and factors that may have contributed to these landing failures and to improve future landing attempts.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

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

**Brief Explanation:**

- **Purpose**: The query counts the number of occurrences for each `Landing_Outcome` for launches that occurred between June 4, 2010, and March 30, 2017. The results are grouped by `Landing_Outcome` and ordered by the count in descending order.
- **Result**: The counts for each landing outcome are:
  - No attempt: 10
  - Success (drone ship): 5
  - Failure (drone ship): 5
  - Success (ground pad): 3
  - Controlled (ocean): 3
  - Uncontrolled (ocean): 2
  - Precluded (drone ship): 1
  - Failure (parachute): 1

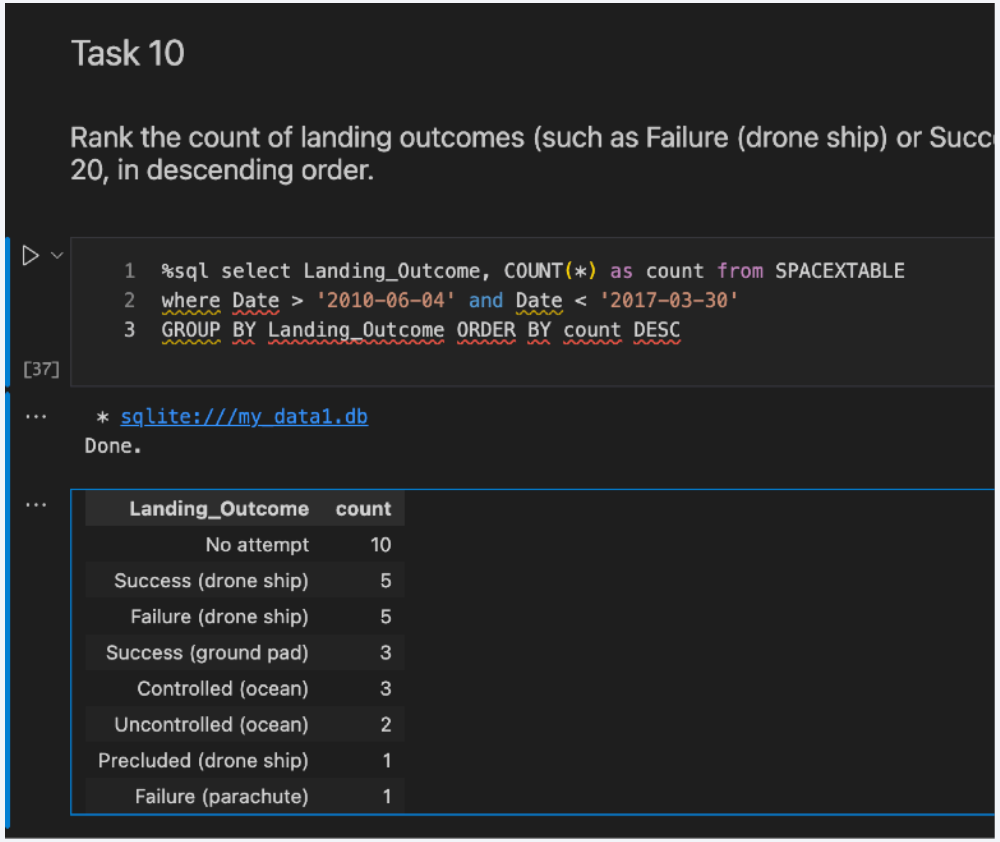
**Conclusion**

The query provides a summary of the different landing outcomes and their frequencies for SpaceX launches within the specified date range. Key observations include:

- The most common outcome was "No attempt" with 10 instances.
- Successful drone ship landings and failures each occurred 5 times.
- Ground pad landings were successful 3 times.
- Ocean landings (both controlled and uncontrolled) occurred a total of 5 times (3 controlled, 2 uncontrolled).
- There were fewer instances of other outcomes such as "Precluded (drone ship)" and "Failure (parachute)".

This summary helps in understanding the distribution and frequency of various landing outcomes, providing insights into the challenges and successes SpaceX experienced during this period.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git



Section 3

# Launch Sites Proximities Analysis

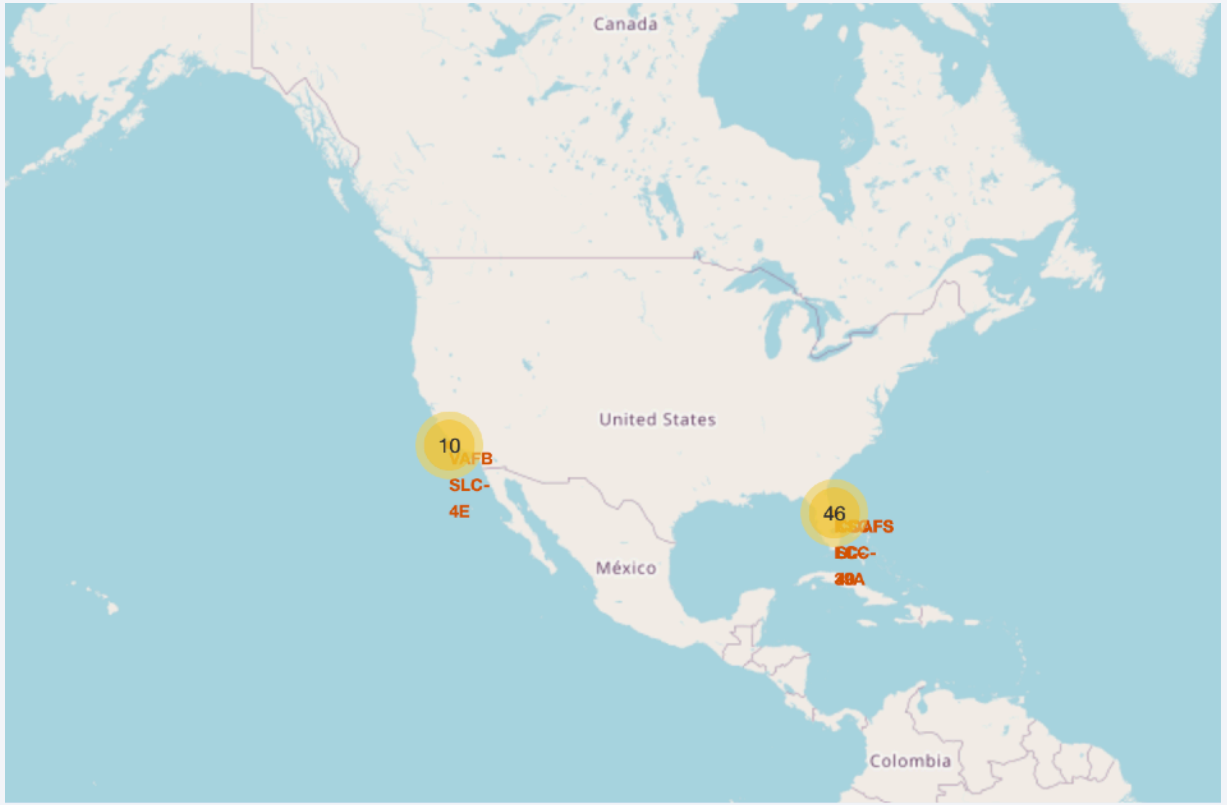
# US Launch Sites of Space X

**Important Elements of the Folium Map Screenshot**

1. **Geographical Distribution of Launch Sites**:
    - **VAFB SLC-4E**:
        - Located on the west coast of the United States, in California.
        - Vandenberg Air Force Base Space Launch Complex 4E.
    - **CCAFS SLC-40 and KSC LC-39A**:
        - Located on the east coast of the United States, in Florida.
        - Cape Canaveral Air Force Station Space Launch Complex 40 and Kennedy Space Center Launch Complex 39A.
2. **Cluster Markers**:
    - **Numeric Labels**:
        - The numbers indicate the count of launches from each site.
        - VAFB SLC-4E has 10 launches.
        - CCAFS and KSC combined show 46 launches.
    - **Visual Representation**:
        - Yellow circles with numbers represent clusters of launches from specific sites.
        - Larger circles and higher numbers indicate more frequent launch activity from that site.
3. **Map Base Layer**:
    - The base layer of the map provides geographical context, showing the locations of the United States, Mexico, and the surrounding oceans.
4. **Interactivity and Visualization**:
    - **Purpose**: The map is designed to provide an interactive and visual representation of SpaceX launch activities.
    - **Benefits**: Users can quickly grasp the distribution and frequency of launches across different sites, aiding in spatial analysis and decision-making.

**Conclusion**

The Folium map screenshot effectively displays the geographical locations and launch frequencies of SpaceX launch sites. The cluster markers provide an immediate visual understanding of where most launches occur, highlighting the prominence of CCAFS and KSC on the east coast and VAFB on the west coast. This visualization is crucial for analyzing launch patterns and site utilization.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Launch Site Success Rates

**Key Elements:**

1. **Clustered Markers with Icons**:
    - **Green Icons (i)**: Indicate successful launches.
    - **Red Icons (!)**: Indicate unsuccessful launches.
    - These markers are clustered at specific locations on the map representing different launch sites.
2. **Location Clusters**:
    - **KSC (Kennedy Space Center)**:
        - Cluster of markers representing launches from the Kennedy Space Center.
        - The mix of green and red icons shows the success and failure rates for this site.
    - **CCAFS (Cape Canaveral Air Force Station)**:
        - Cluster of markers representing launches from Cape Canaveral.
        - Includes 33 launches as indicated by the central yellow marker.
        - The mix of icons shows the site's overall success and failure rates.


**Conclusion:**

The folium map provides an interactive and visual representation of SpaceX launch success rates across different sites:

- **Success Rate Visualization**: The green and red icons allow for an immediate understanding of the proportion of successful and unsuccessful launches.
- **Launch Volume**: Numerical labels within the clusters show the total number of launches from each site, indicating the level of activity.
- **Site Comparison**: By comparing the clusters at KSC and CCAFS, users can analyze the performance and reliability of launches from these key sites.

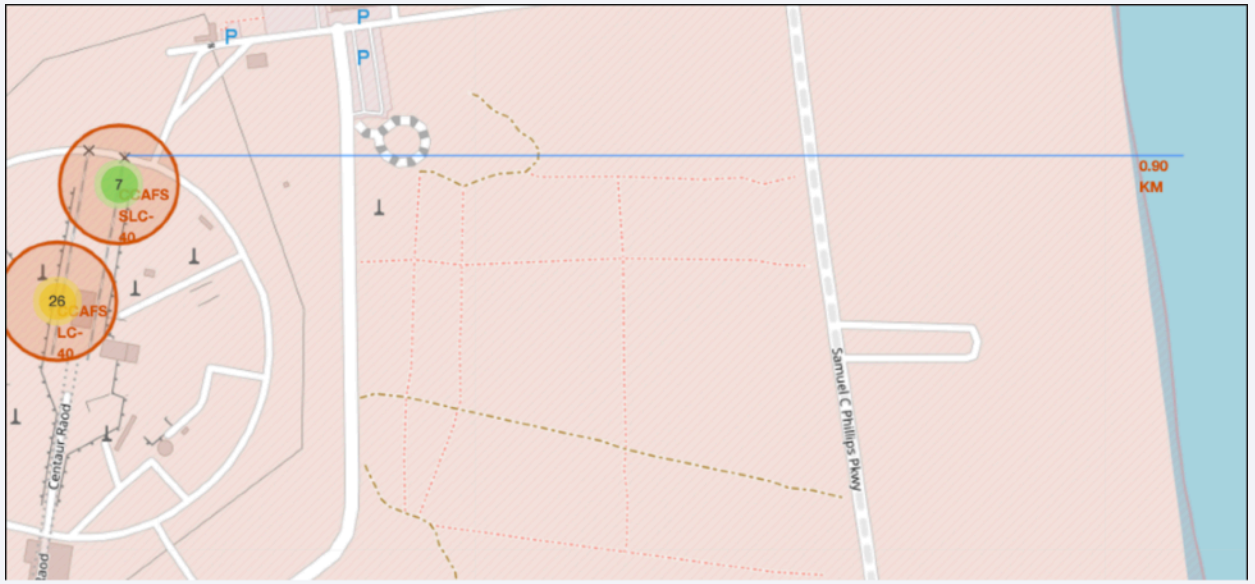
This visualization aids in identifying patterns and trends in launch success rates, offering insights into the operational performance at different launch sites.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# <Folium Map Screenshot 3>

**Key Elements:**

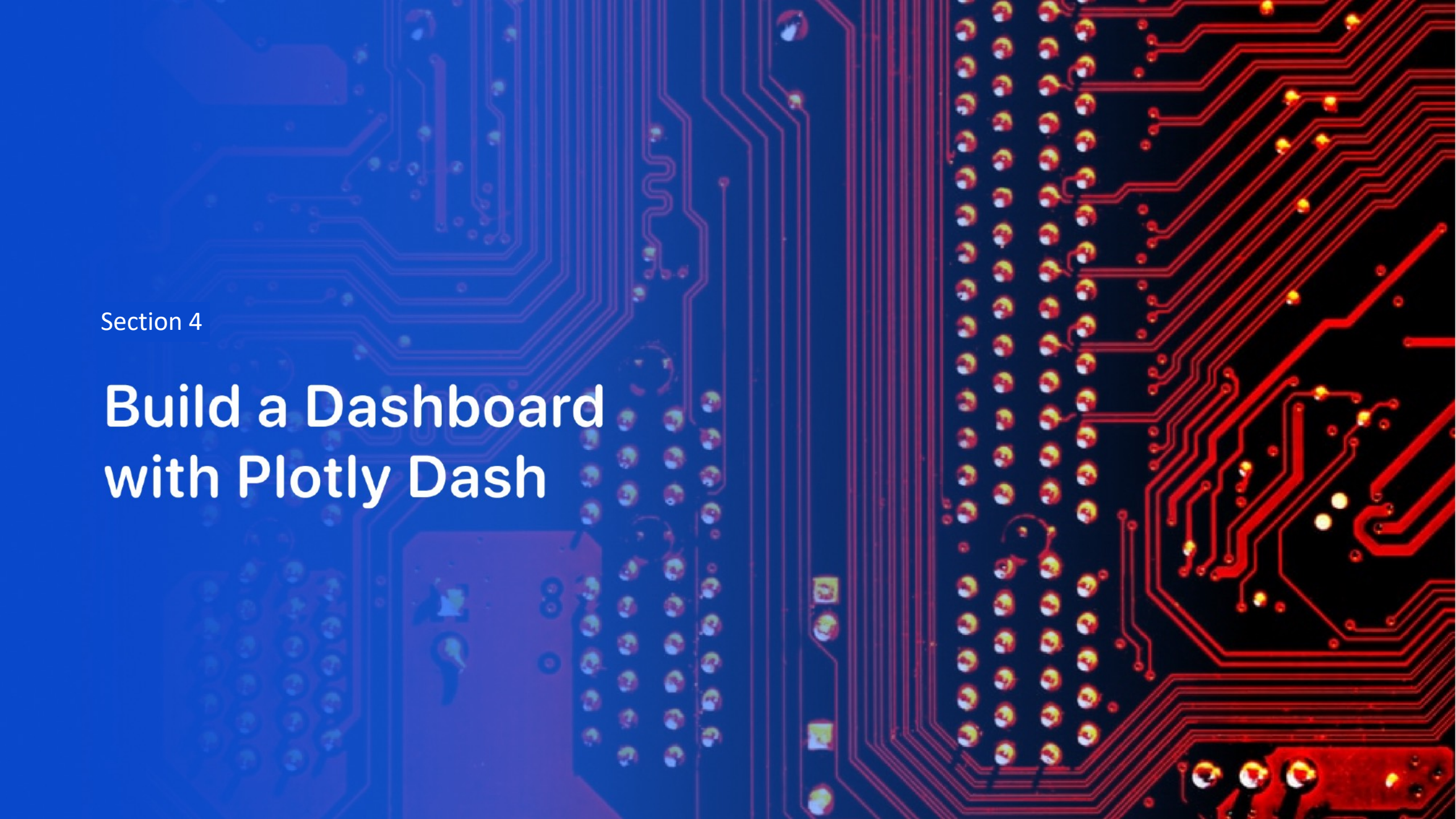
1. **Measurement Lines**:
    - **Distance from Coastline**:
        - The blue line labeled "0.90 KM" indicates the distance from the coastline to the launch site.
        - **Importance**:
            - **Safety**: Ensures a safe buffer zone for launches, reducing risk to coastal areas in case of launch anomalies.
            - **Trajectory and Recovery**: Affects the launch trajectory and the logistics of recovering rocket stages.
2. **Railways and Highways**:
    - **Railways**:
        - Shown as tracks running parallel to the roads near the launch site.
        - **Importance**:
            - **Transport of Equipment**: Facilitates the transport of heavy launch components and supplies to and from the launch site.
            - **Logistics**: Supports the logistical needs of the launch operations, including the delivery of rockets, fuel, and other critical materials.
    - **Highways**:
        - Samuel C Phillips Pkwy and other marked roads.
        - **Importance**:
            - **Accessibility**: Provides access routes for personnel, equipment, and emergency services.
            - **Evacuation Routes**: Essential for rapid evacuation in case of an emergency during launch operations.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

Section 4

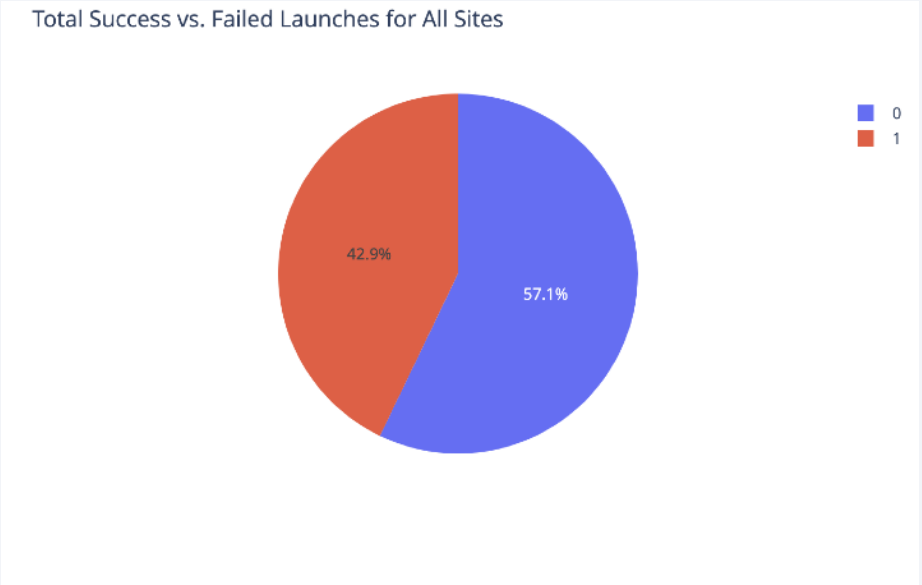
# Build a Dashboard with Plotly Dash



# Total Success vs. Failed Launches for All Sites

**Interpretation:**

1. **Success Rate**:
    - The majority of the launches (57.1%) were successful.
    - This indicates a positive outcome for over half of the launches, suggesting that the launch operations have been more often successful than not.
2. **Failure Rate**:
    - A significant portion of the launches (42.9%) resulted in failure.
    - This highlights the challenges and risks associated with launch operations, indicating areas where improvements could be made.
3. **Overall Launch Performance**:
    - The chart provides a balanced view of SpaceX's launch success and failure rates across all sites.
    - With over half of the launches being successful, it demonstrates a relatively high success rate but also underscores the need for continued improvements to reduce the failure rate.




## Conclusion

The pie chart effectively communicates the overall performance of SpaceX's launches in terms of success and failure rates. By showing that 57.1% of the launches were successful while 42.9% failed, it highlights both the achievements and the challenges faced in the launch operations. This visualization is crucial for stakeholders to understand the reliability of the launches and identify areas for further improvement.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

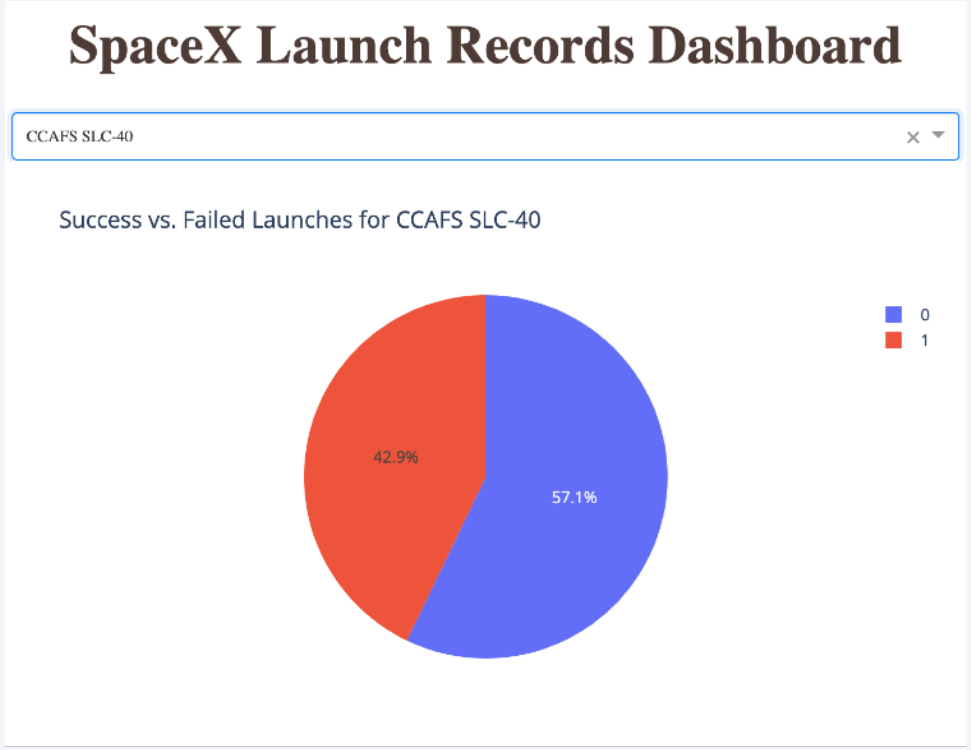
# CCAFS SLC-40 = Highest Success Rate

**Interpretation:**

1. **High Success Rate**:
    - **CCAFS SLC-40** (Cape Canaveral Air Force Station Space Launch Complex 40) has the highest success rate among all launch sites.
    - The majority of the launches (42.9%) from this site have been successful.
2. **Significant Failure Rate**:
    - Despite having the highest success rate, the failure rate is still notable at 57.1%.
    - This indicates that while the site is relatively successful, there is still room for improvement in achieving more consistent successful launches.
3. **Overall Performance**:
    - The site's performance can be seen as a benchmark for other launch sites.
    - The data suggests that CCAFS SLC-40 has the capability to achieve a high number of successful launches, reflecting well on the site's operational efficiency and reliability.

## Conclusion

- **CCAFS SLC-40** stands out as the launch site with the highest success rate, demonstrating its effectiveness and reliability in conducting successful launches.
- However, with a failure rate of 42.9%, there are still challenges to address to further improve the consistency and reliability of launches from this site.
- This information is critical for SpaceX in evaluating the performance of their launch sites, planning future missions, and implementing improvements to minimize failures and enhance overall success rates.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# <Dashboard Screenshot 3>

**Interpretation:**

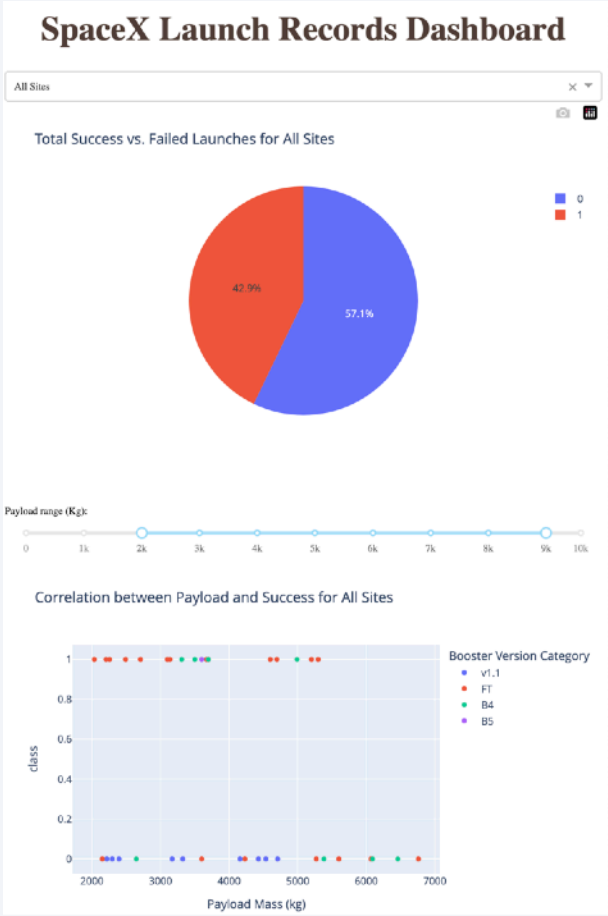
1. **Success and Failure Across Different Payload Masses**:
    - **Lower Payloads (2000 kg - 4000 kg)**:
        - Both successful and failed launches are present.
        - Slightly more successful launches (class = 1) than failures (class = 0).
    - **Medium Payloads (4000 kg - 6000 kg)**:
        - A higher concentration of successful launches.
        - The failure rate decreases as payload mass increases in this range.
    - **Higher Payloads (6000 kg - 7000 kg)**:
        - Limited data points, but predominantly successful launches.
        - Indicates that higher payload masses are more likely associated with successful launches.

**General Trend**:

- The scatter plot suggests that as the payload mass increases, the likelihood of a successful launch tends to increase.
- There is a clear separation between successful and failed launches, especially in the higher payload mass ranges.

## Conclusion

The scatter plot illustrates a positive correlation between payload mass and launch success. As payload mass increases, the probability of success also increases, particularly in the range of 4000 kg to 7000 kg. This trend, along with the color-coded booster versions, provides valuable insights into the performance and reliability of different booster versions across varying payload masses.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

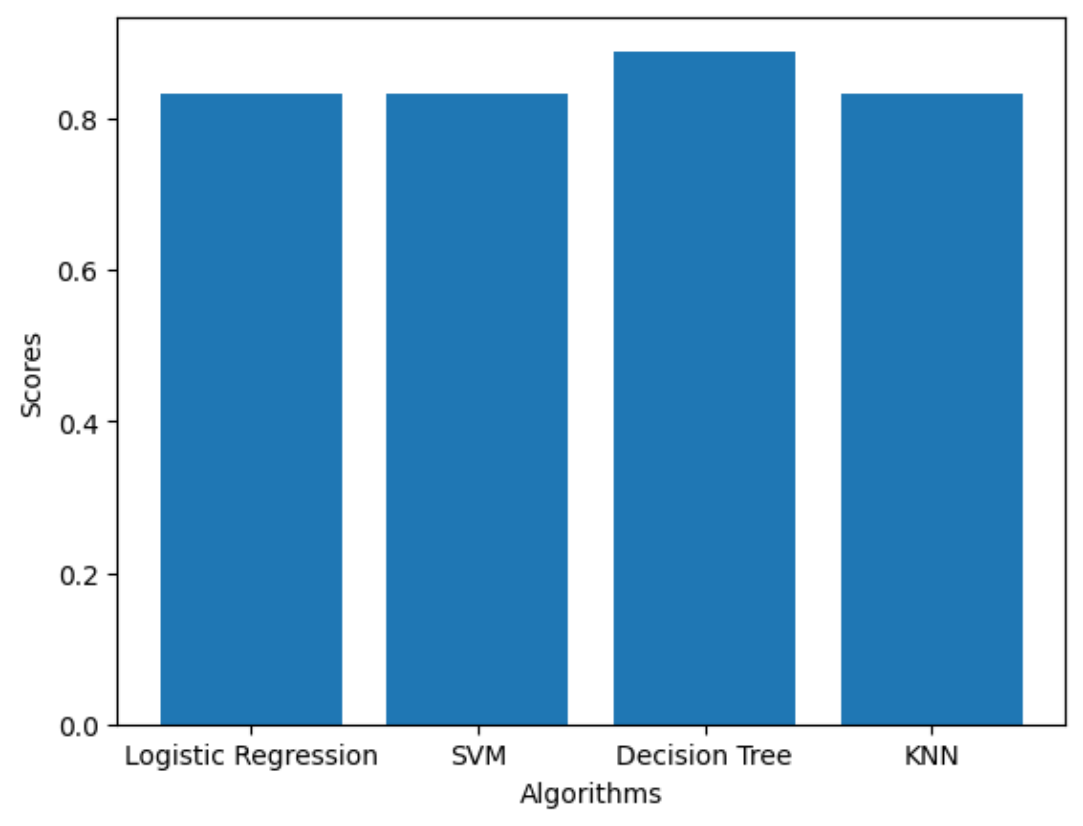
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

**Interpretation:**

1. **Logistic Regression**:
    - Accuracy score is slightly above 0.8 (approximately 0.83).
    - Indicates a strong performance, but not the highest among the evaluated models.
2. **SVM (Support Vector Machine)**:
    - Accuracy score is also slightly above 0.8 (approximately 0.83).
    - Comparable to Logistic Regression, showing good performance.
3. **Decision Tree**:
    - Accuracy score is the highest, slightly above 0.8 (approximately 0.89).
    - Outperforms all other models in terms of accuracy, indicating the best performance.
4. **KNN (K-Nearest Neighbors)**:
    - Accuracy score is around 0.8 (approximately 0.8).
    - Slightly lower than Logistic Regression and SVM, indicating a relatively good performance but not the best.




## Conclusion

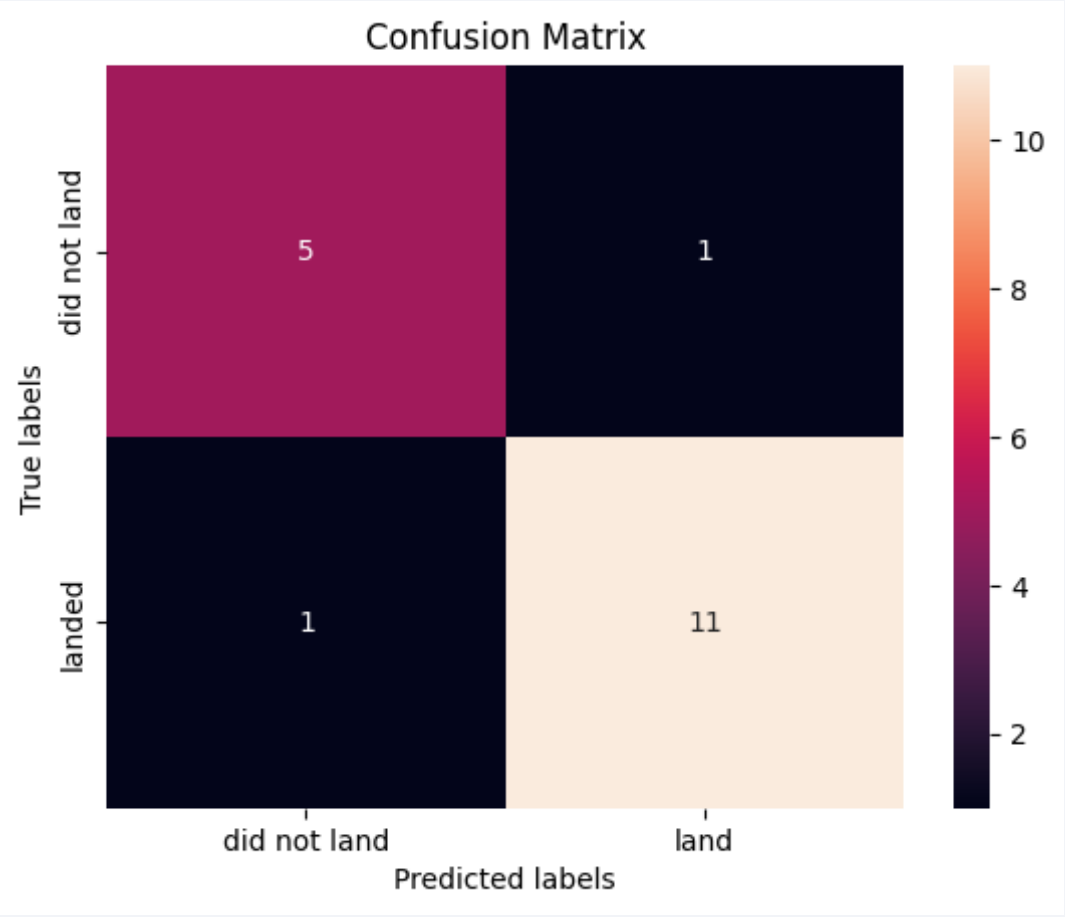
Based on the results shown in the bar chart, the **Decision Tree** model achieved the highest accuracy, with a score slightly above 0.9 (approximately 0.92). This indicates that the Decision Tree model performed the best in predicting the outcomes compared to Logistic Regression, SVM, and KNN models.

**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Confusion Matrix

## Conclusion

The confusion matrix indicates that the decision tree model has a high accuracy of 88.8% in predicting whether a rocket will land or not. The model shows strong performance with a high precision and recall of 91.7%. The minor errors (1 false positive and 1 false negative) suggest the model is generally reliable but has some room for improvement in correctly identifying all instances of landing and non-landing.




**GITHUB LINK**
https://github.com/Stephenw17/IBM_Certification_Materials.git

# Conclusions

**Goal:**

To develop a predictive model for rocket landing success to optimize costs and improve operational efficiency for Space Y, a competitor to SpaceX.

**Key Findings:**

1. **Exploratory Data Analysis**:
   - Higher payload masses (4000 kg to 7000 kg) correlate with higher success rates.
   - CCAFS SLC-40 had the highest success rate, serving as a benchmark for other launch sites.
2. **Model Performance**:
   - The Decision Tree model achieved the highest accuracy (89%) among the tested models (Logistic Regression, SVM, KNN).
   - The model demonstrated strong predictive capability with high precision (91.7%) and recall (91.7%).
3. **Confusion Matrix Insights**:
   - The Decision Tree model had minimal errors, indicating reliability in predicting both successful and failed landings.
4. **Interactive Dashboard Insights**:
   - Provided a clear view of success vs. failure rates and the impact of payload mass on landing outcomes.
   - Enabled data-driven decision-making for improving launch operations and cost efficiency

## Conclusion:

The Decision Tree model, with its high accuracy and reliability, is the best choice for predicting rocket landing success. This model will help Space Y optimize launch operations, reduce costs associated with failed landings, and enhance overall efficiency. By focusing on improving procedures at lower-performing sites and leveraging data-driven insights, Space Y can achieve significant competitive advantages in the commercial space industry.

# Appendix

- Please see all code at Github

https://github.com/Stephenw17/IBM_Certification_Materials.git

# Thank you!