IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohammad Aljelo

29 September 2023

Outline

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

Executive Summary

In this project, our primary aim is to forecast the successful landing of the SpaceX Falcon 9 first stage. The ability to ascertain the first stage's landing outcome is pivotal in determining the launch cost. This objective will be accomplished through the application of diverse machine-learning classification algorithms.

The methodology we will adhere to encompasses several crucial stages, commencing with data collection, followed by data wrangling and preprocessing, proceeding to exploratory data analysis and data visualization, and culminating in the utilization of machine learning for predictive modeling.

In the course of our inquiry, our analytical findings reveal that certain characteristics of rocket launches exhibit correlations with their success or failure.

Ultimately, our investigation leads us to the conclusion that the Decision Tree algorithm may offer the most suitable machine-learning approach for this particular challenge.

Introduction

In this project, we will predict whether the SpaceX Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. This will be achieved with the use of different machine-learning classification algorithms.

During our investigation, the results of our analysis indicate that there are some features of rocket launches that have a correlation with the success or failure launches.

In the end we conclude that the Decision Tree may be the best machine learning algorithm to for this problem.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Initially, data acquisition commenced by utilizing the SpaceX API, a RESTful API, through the execution of a GET request directed at the SpaceX API endpoint. To streamline this process, a set of auxiliary functions were meticulously defined to facilitate the extraction of information from the launch data via identification numbers. Subsequently, rocket launch data was retrieved from the SpaceX API URL.
- To enhance the uniformity of the requested data in JSON format, the SpaceX launch data underwent retrieval and parsing through a GET request. Subsequently, the response content was decoded, resulting in a JSON output, which was subsequently transformed into a Pandas data frame.
- Additionally, a web scraping operation was conducted to compile historical Falcon 9 launch records from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches." These launch records were embedded in HTML format on the Wikipedia page. By employing the BeautifulSoup and requests libraries, the Falcon 9 launch HTML table records were extracted, parsed, and subsequently converted into a Pandas data frame for further analysis.

Data Collection – SpaceX API

To collect data for predicting the successful landing of SpaceX Falcon 9 first stages, we follow a structured process using SpaceX's REST API. This involves API initialization, sending GET requests, handling responses, and converting data into a usable format. The following flowchart illustrates this step-by-step data collection journey.

GitHub URL: [Data Collection SpaceX API]

- Start
- Obtain SpaceX API Key
- Initialize API Request
    - -Specify API Endpoint
    - -Add API Key to Headers
    - -Set Query Parameters
- Send GET Request to SpaceX API
- Check Response Status Code
- Decode Response Content as JSON
- Convert JSON to Pandas DataFrame
- Export the dataframe to a CSV file

Data Collection - Scraping

In this Stage, we will perform web scraping to collect historical launch records of Falcon 9 rockets from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches." Our objective is to extract data such as flight numbers, dates, launch sites, payloads, payload masses, orbits, customers, launch outcomes, booster versions, booster landing status, and more.

To accomplish this, we will utilize Python packages like BeautifulSoup and requests to parse the HTML content of the Wikipedia page, extract relevant information from HTML tables, and convert it into a Pandas data frame. The collected data will be used for further analysis and prediction.

Now, let's proceed with the flowchart illustrating the web scraping process.

GitHub URL: [Web Scraping]

- Start
- Request the Falcon9 Launch Wiki page from its URL
  - -Perform an HTTP GET request to fetch the Wikipedia page
  - -Create a BeautifulSoup object from the HTML response
  - -Verify if the BeautifulSoup object was created properly
- Extract all column/variable names from the HTML table header
  - -Find all tables on the Wikipedia page
  - -Identify the target table containing launch records
  - -Extract column names from the table header
- Create a data frame by parsing the launch HTML tables
  - -Initialize an empty dictionary with column names
  - -Fill the dictionary with launch records from table rows
  - -Extract flight number, date, time, booster version, etc.
  - -Create a Pandas dataframe from the dictionary
  - -Export the dataframe to a CSV file

Data Wrangling

This data wrangling process involves loading and preparing the dataset, performing exploratory data analysis, calculating statistics for launches, orbits, and mission outcomes, creating a landing outcome label, calculating the success rate, and finally exporting the processed data for further analysis.

Now, let's proceed with the flowchart illustrating the Data wrangling process.

GitHub URL: [Data wrangling]

Data Wrangling Process:

- Import libraries and load the dataset.
- Handle missing values and review data types.

Launch Site Analysis:

- Calculate the number of launches from each site.

Orbit Analysis:

- Calculate the number of launches for each orbit type.

Mission Outcome Analysis:

- Calculate the number of mission outcomes for different types.

Create Landing Outcome Label:

- Assign 1 for successful landings and 0 for unsuccessful ones.

Calculate Success Rate:

- Find the success rate of Falcon 9 first stage landings.

Export Processed Data:

- Save the processed data, including the landing class, to a CSV file.

EDA with Data Visualization

1. **FlightNumber vs. PayloadMass (Catplot):** Visualizes the impact of flight number and payload mass on launch success.
2. **Launch Site vs. FlightNumber (Relplot and Catplot)**: Identifies launch site preferences over time and their effect on success rates.
3. **PayloadMass vs. Launch Site (Relplot and Catplot):** Examines the relationship between launch sites and payload mass.
4. **Success Rate vs. Orbit Type (Catplot - Bar chart):** Reveals success rates for different orbit types.
5. **FlightNumber vs. Orbit Type (Lineplot and Catplot):** Explores how flight number relates to success in various orbit types.
6. **PayloadMass vs. Orbit Type (Catplot):** Investigates how payload mass influences successful landings in different orbits.
7. **Launch Success Yearly Trend (Lineplot):** Shows the trend in launch success rates over the years.

GitHub URL: [EDA with Data Visualization]

EDA with SQL

Here is a summary of the SQL queries performed:

1. Display unique launch site names.
2. Show 5 records of launch sites starting with 'CCA'.
3. Calculate the total payload mass for NASA (CRS) missions.
4. Find the average payload mass for booster version F9 v1.1.
5. Determine the date of the first successful ground pad landing.
6. List boosters with drone ship success and payload mass between 4000 and 6000.
7. Count successful and failed mission outcomes.
8. Identify boosters with the maximum payload mass.
9. List records for failed drone ship landings in 2015.
10. Rank landing outcomes between specific dates.

GitHub URL: [EDA with SQL]

Build an Interactive Map with Folium

In this Folium analysis, I added the following map objects:

1. Markers: Used to represent launch site locations, aiding in visualizing their distribution.
2. Circles: Highlighted specific locations, e.g., NASA Johnson Space Center, for visual emphasis.
3. PolyLines: Drew lines connecting launch sites to nearby points like coastlines, cities, railways, and highways to measure and visualize distances.
4. Distance Markers: Displayed distances between launch sites and proximity points in kilometers for reference.

I used these objects for:

1. Marking Launch Sites: Employed markers to pinpoint launch site locations for easy identification.
2. Highlighting Launch Outcomes: Used colored markers (green for success, red for failure) to show launch results, aiding in spotting high success rate sites.
3. Calculating Distances: Employed polylines and distance markers to compute and visualize distances to nearby points (coastlines, cities, railways, highways), offering insights into spatial relationships.

GitHub URL: [Interactive Map with Folium]()

Build a Dashboard with Plotly Dash

In the dashboard, I incorporated various interactive elements to enhance the analysis of SpaceX launch records

1. Dropdown List: Users can choose specific launch sites for focused analysis or opt for an overview of all sites.
2. Pie Chart: This chart dynamically updates to display the total successful launch counts for all sites when "All Sites" is selected, or the success vs. failure counts for the chosen site. It provides an at-a-glance summary of launch outcomes.
3. Payload Range Slider: A slider enables users to filter payload data within a specified range, offering more precise control over the data subset they want to explore.
4. Scatter Chart: This chart showcases the relationship between payload mass and launch success. Users can visualize how different payload masses impact launch outcomes, with color coding for Booster Version Categories. It provides valuable insights into the correlation between payload and success.

I added these plots and interactions to provide users with a comprehensive and interactive dashboard for analyzing SpaceX launch records. These features allow users to explore launch success rates, payload mass relationships, and specific launch site performance easily.

GitHub URL: [Dashboard with Plotly Dash]

Predictive Analysis (Classification)

1. **Data Preprocessing**: Created a binary class column, standardized data, and split it for training and testing.
2. **Model Selection and Tuning**: Evaluated Logistic Regression, SVM, Decision Tree, and KNN, using GridSearchCV for hyperparameter tuning.
3. **Model Evaluation**: Trained models and evaluated accuracy on test data, visualized results with confusion matrices.
4. **Best Model**: Decision Tree achieved the highest accuracy (88.89%) on test data, making it the preferred choice for predicting Falcon 9 landings.

Here's a flowchart summarizing the process:

- Data Preprocessing
- Model Selection (Logistic Regression, SVM, Decision Tree, KNN)
- Hyperparameter Tuning (GridSearchCV)
- Model Training
- Model Evaluation (Test Data Accuracy)
- Model Comparison (Decision Tree: 88.89%)

Results

Exploratory Data Analysis Results:

- Conducted exploratory data analysis on SpaceX Falcon 9 launch data.
- Explored features like launch outcomes, payload mass, launch sites, and more.
- Identified trends, patterns, and relationships in the data.
- Visualized key findings to gain insights.

Interactive Analytics Demo in Screenshots:

- Created an interactive analytics demo to explore the data.
- Included screenshots showcasing the demo's functionality and features.
- Demonstrated how users can interact with the data through the dashboard.

Predictive Analysis Results:

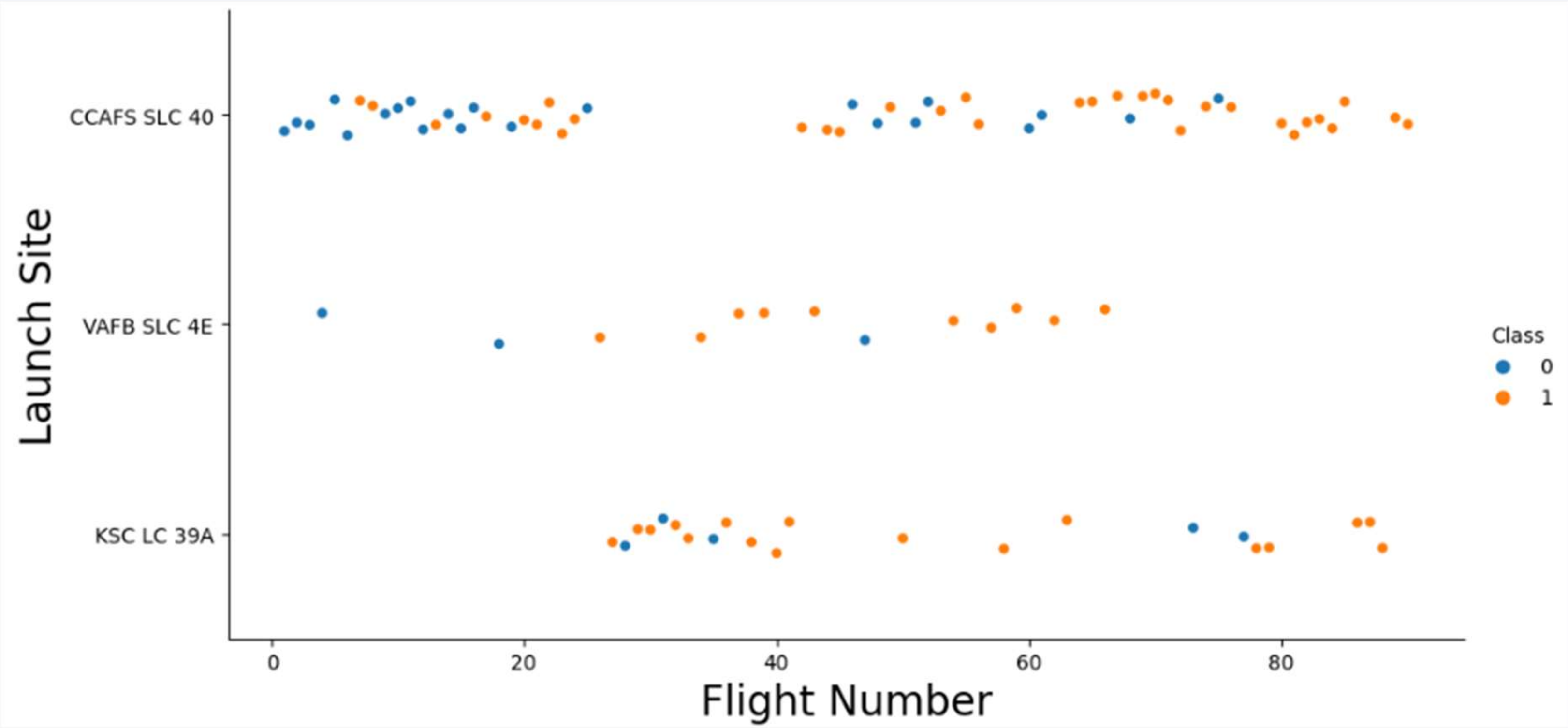
- Developed machine learning models (Logistic Regression, SVM, Decision Tree, and KNN) to predict Falcon 9 first-stage landings.
- Tuned model hyperparameters using GridSearchCV.
- Evaluated model performance on test data.
- Identified the Decision Tree model as the best-performing, achieving an accuracy of 88.89% on test data.

Section 2

Insights drawn from EDA

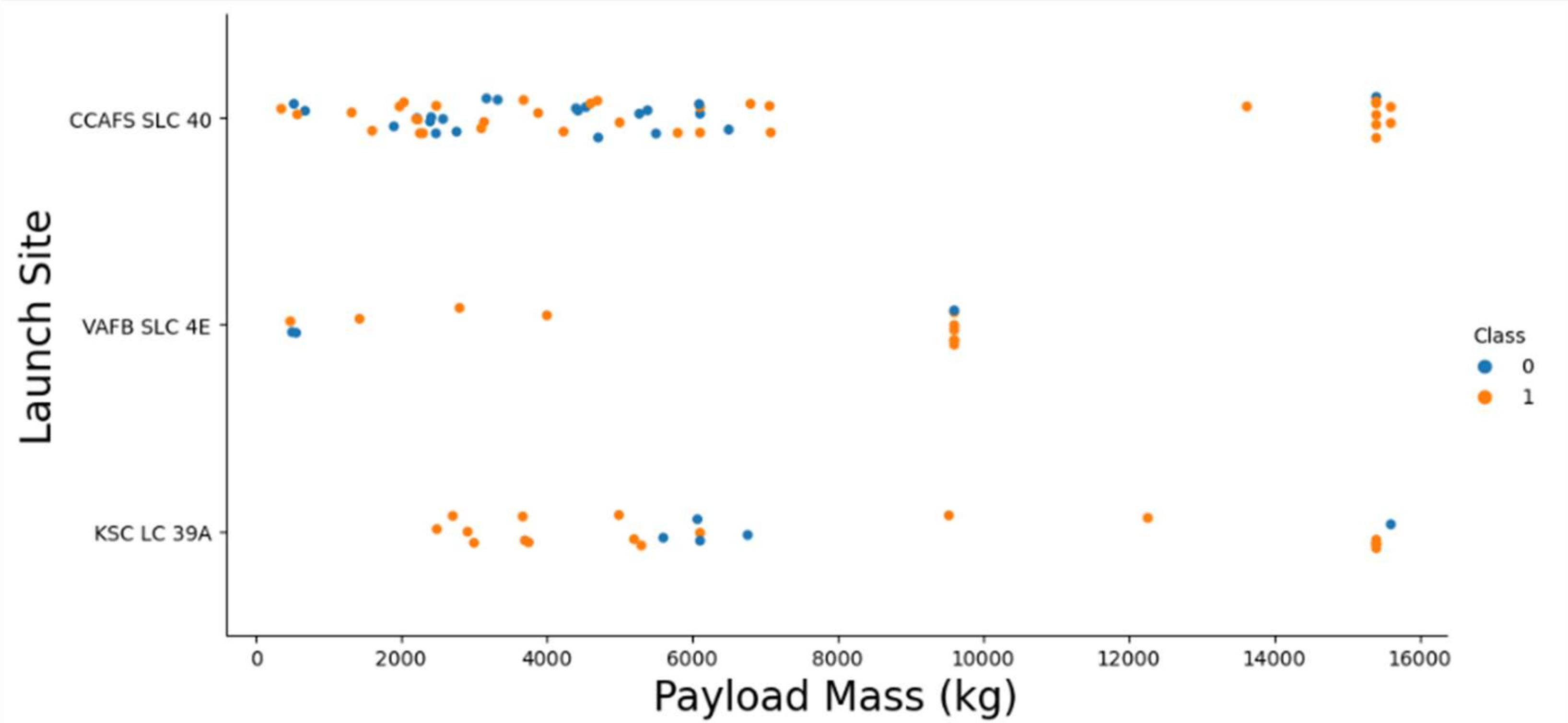
Flight Number vs. Launch Site

In this illustration, we can observe a rise in the success rate corresponding to an increase in the number of flights conducted. The successful launches are denoted by blue dots, while the unsuccessful ones are represented by a red dot. Notably, there appears to be a noticeable uptick in successful flights following the 40th launch.

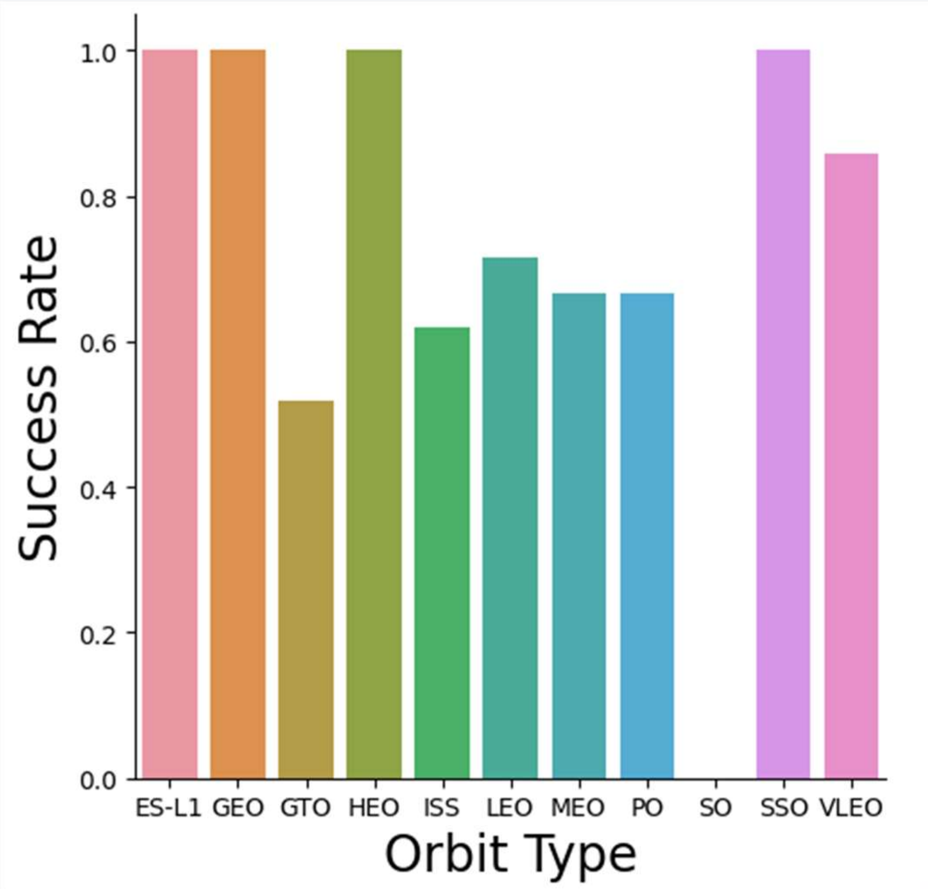


Payload vs. Launch Site

- In this illustration, The blue dots symbolize successful launches, whereas the red dots depict unsuccessful ones. Interestingly, the VAFB-SLC launch site has not hosted any rocket launches with heavy payload masses.
- It appears that there exists a limited correlation between Payload and Launch Site. Consequently, this metric alone may not provide sufficient information for making decisions.



Success Rate vs. Orbit Type



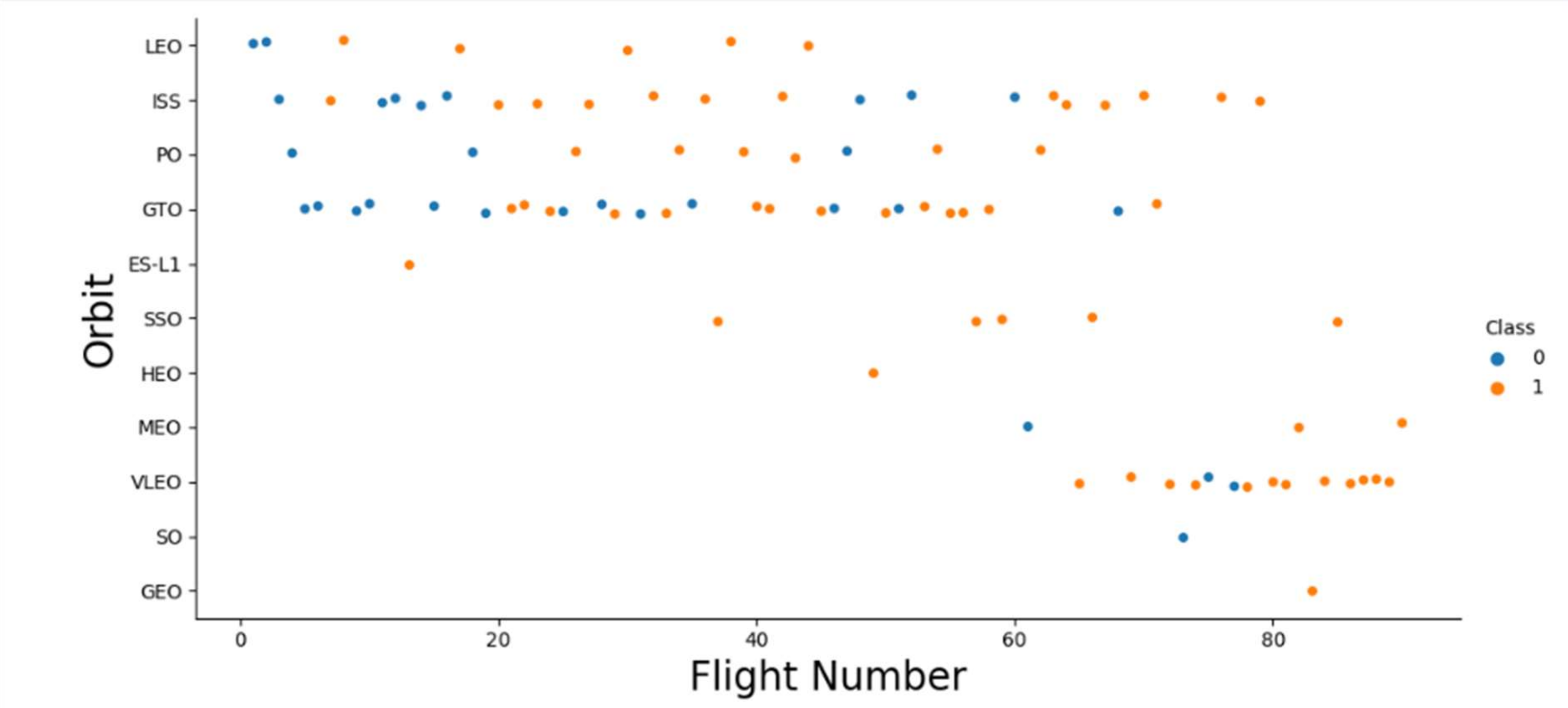
- The SSO, HEO, GEO, and ES-L1 orbits have all achieved a flawless 100% success rate. However, it's worth noting that the SO orbit has not seen any successful launches, resulting in a 0% success rate.

I want to rephrase the following text

M

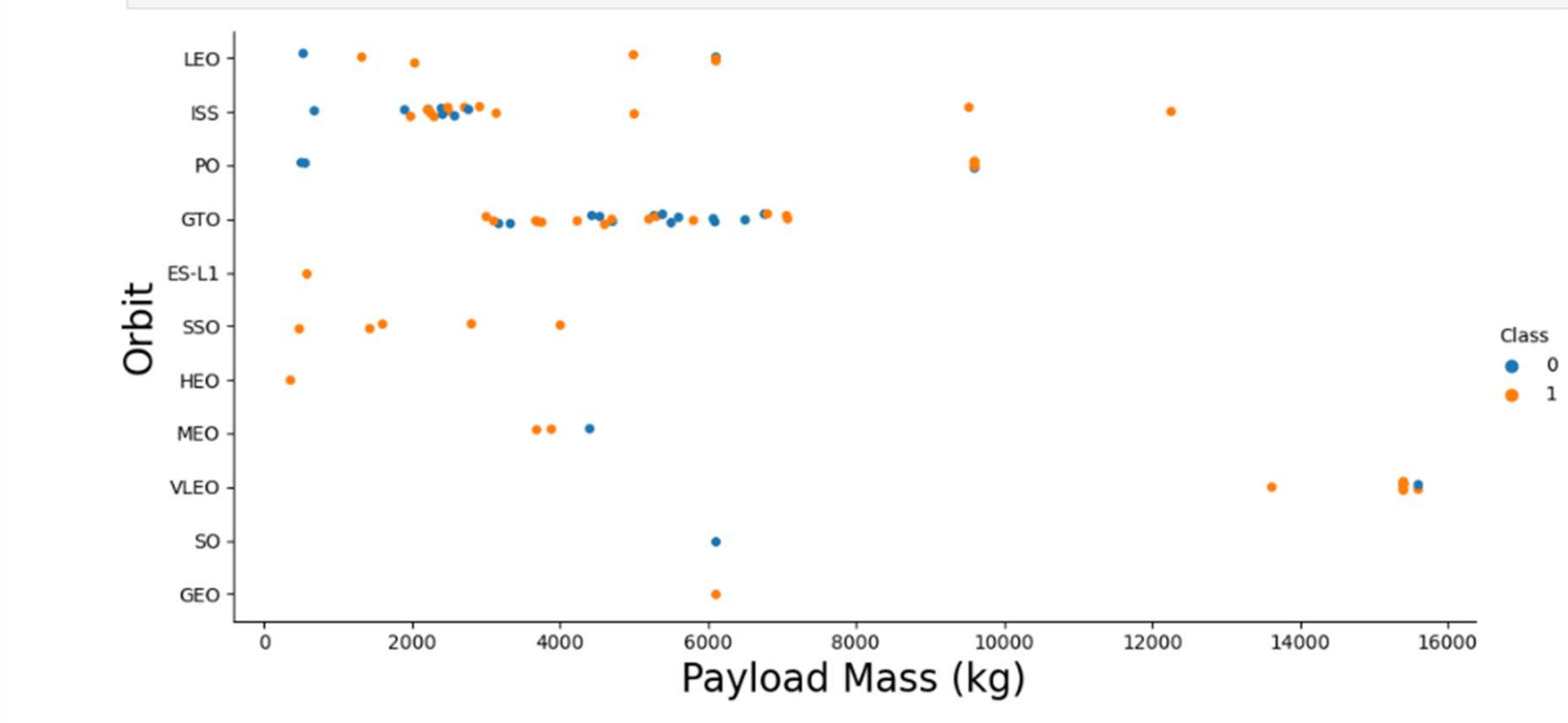
Flight Number vs. Orbit Type

- Within the LEO orbit, there is a noticeable positive correlation between success and the number of flights.
- In contrast, there appears to be no discernible relationship between flight number and success in the GTO orbit.
- Despite having fewer flights compared to other orbits, the SSO orbit boasts a flawless 100% success rate.

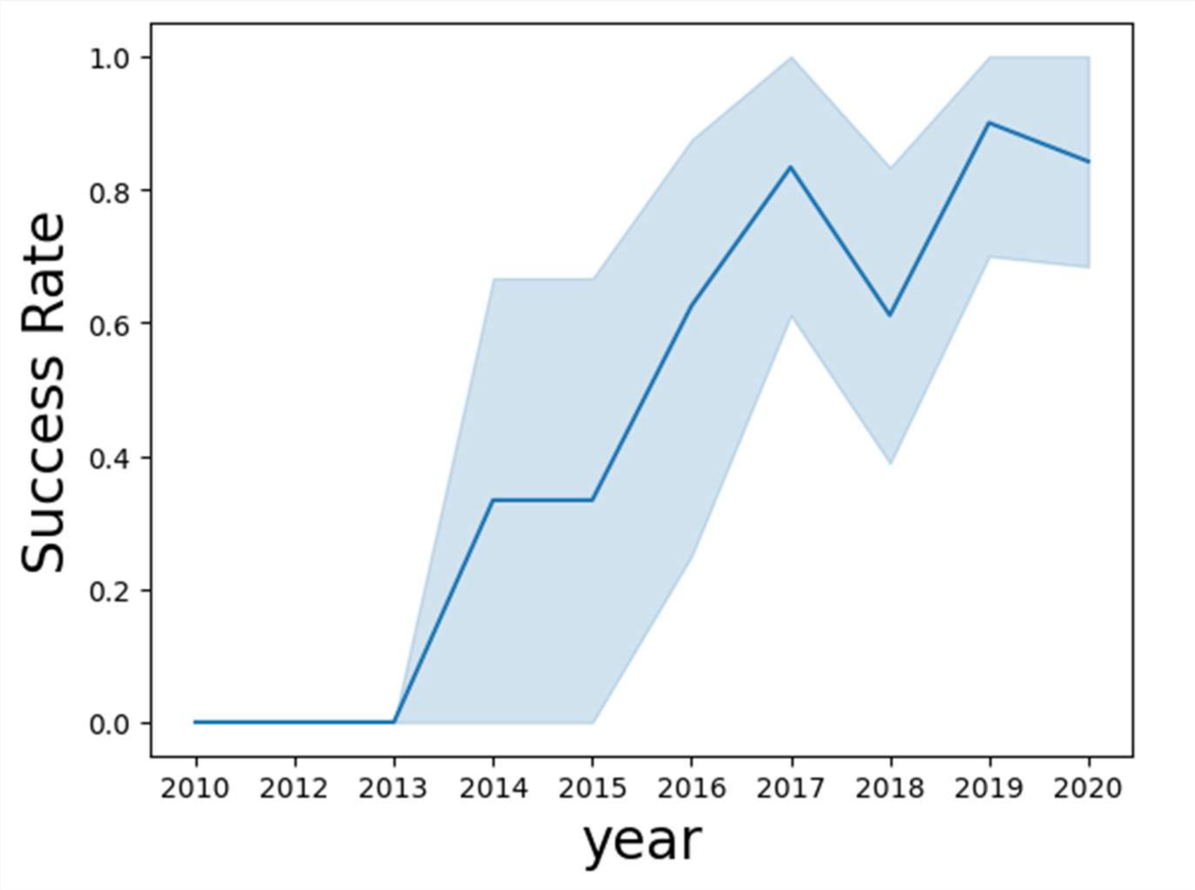


Payload vs. Orbit Type

- In the PO, SSO, LEO, and ISS orbits, a trend emerges: as payloads become heavier, the success rate notably rises.
- However, in the case of the GTO orbit, no clear-cut correlation can be observed between orbit type and payload mass, as both successful and failed launches are evenly distributed



Launch Success Yearly Trend



- The overall pattern depicted in the chart reveals a rising trend in landing success rates over the years. However, there are noticeable declines in both 2018 and 2020.

All Launch Site Names

We utilized the keyword "DISTINCT" to display exclusively the unique launch sites found in the SpaceX data.

In [8]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

- This query retrieves 5 records from the 'SPACEXTBL' table where the 'Launch_Site' column starts with the prefix 'CCA.' The '%' symbol in the 'LIKE' clause is a wildcard character, allowing us to match any characters that follow 'CCA.'

Done.

Out[23]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The query calculates the total payload mass (in kilograms) carried by boosters from NASA (CRS) and presents the result in a table. The result of this query is that boosters from NASA (CRS) have carried a total payload mass of 45,596 kilograms.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The query calculates the average payload mass (in kilograms) carried by booster versions that start with "F9 v1.1" and presents the result in a table.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version
```

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

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

The query identifies the minimum (earliest) date when a successful landing outcome on a ground pad occurred. The result of this query is "2015-12-22," which represents the date of the first successful landing outcome on a ground pad. This information provides insight into the historical timeline of SpaceX's successful ground pad landings.

```
%sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing_Outcome" = "Success (ground pad)";
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query retrieves distinct booster version names from the 'SPACEXTBL' table where the landing outcome was "Success (drone ship)" and the payload mass was greater than 4000 kg but less than 6000 kg. The result of this query provides a list of booster versions that successfully landed on a drone ship within the specified payload mass range.

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- This query counts and categorizes the total number of mission outcomes, grouping them by their respective mission outcomes. The result of this query provides a summary of the total number of missions categorized by their outcomes, including both successful and failed missions.

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

This query retrieves the booster versions and their corresponding payload masses where the payload mass is equal to the maximum payload mass found in the dataset. The result displays the names of boosters that have carried the maximum payload mass, which is 15600 kilograms.

```
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

This query retrieves records from the dataset for the year 2015 where the landing outcome is marked as "Failure (drone ship)." The result displays the year, month, booster version, launch site name, payload, payload mass, mission outcome, and landing outcome for each of these failed drone ship landings in the year 2015.

```
%sql SELECT substr(Date,1,4), substr(Date, 6, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Ou
```

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

substr(Date,1,4)	substr(Date, 6, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing_Outcome
2015	10	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

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

The query groups the landing outcomes, counts the occurrences for each group, and orders them in descending order by count_outcomes. This provides a ranked list of landing outcomes between the specified dates, showing which outcome occurred most frequently during that time period.

```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-04-06' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

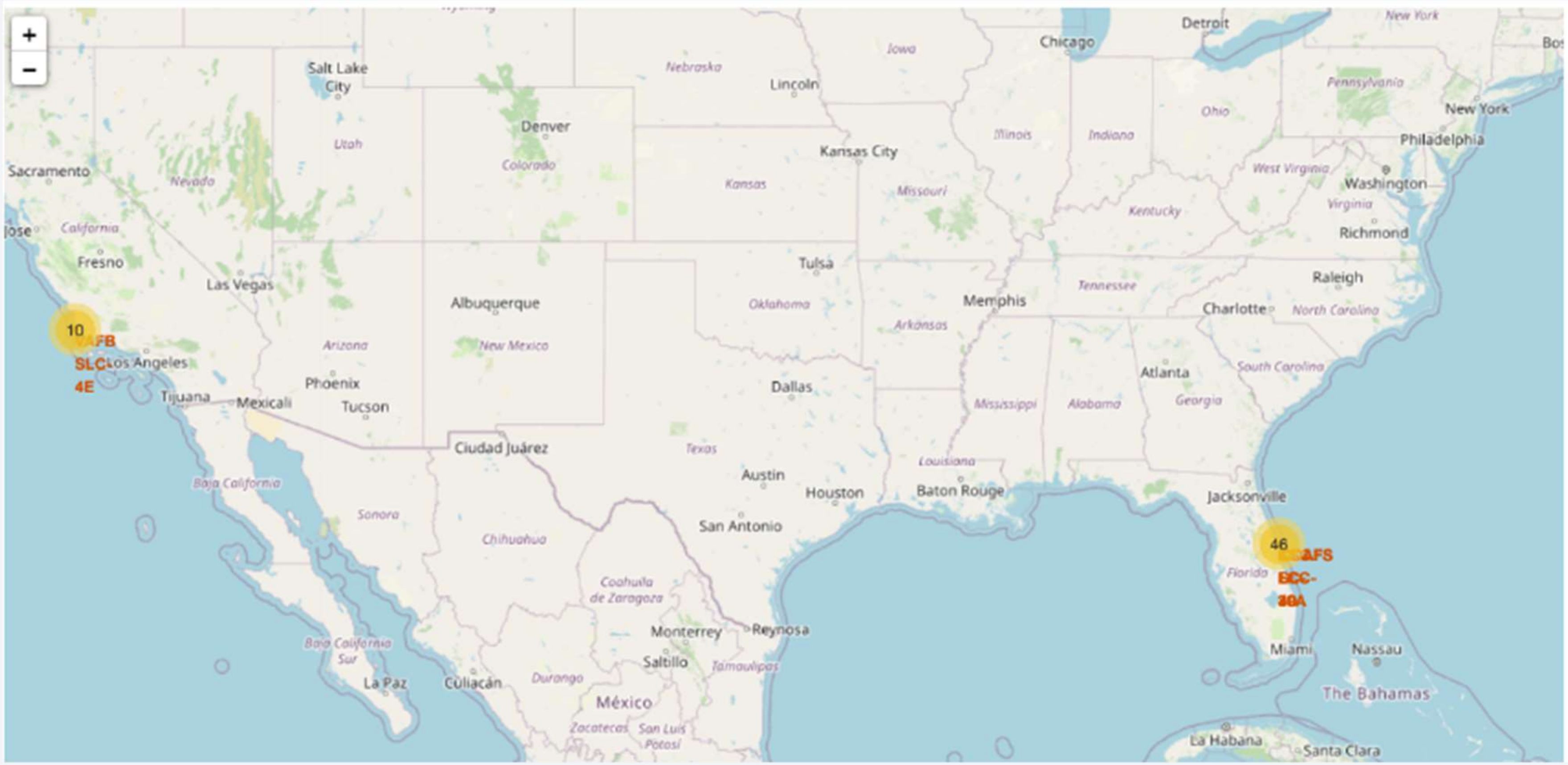
Landing_Outcome	count_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Analyzing SpaceX Launch Site Locations and Success Rates

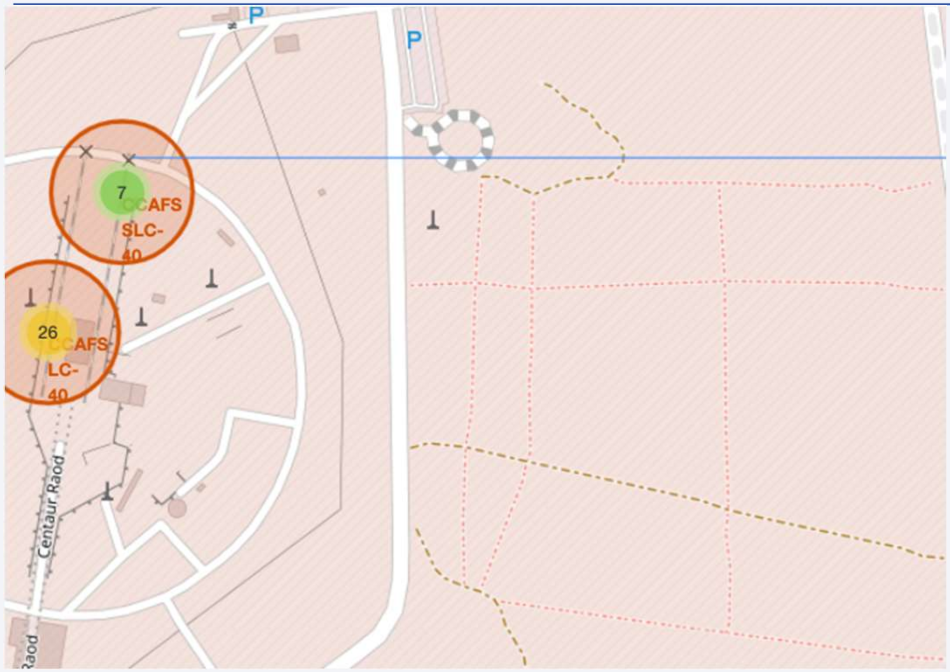
- The screenshot visually represents SpaceX launch sites worldwide.



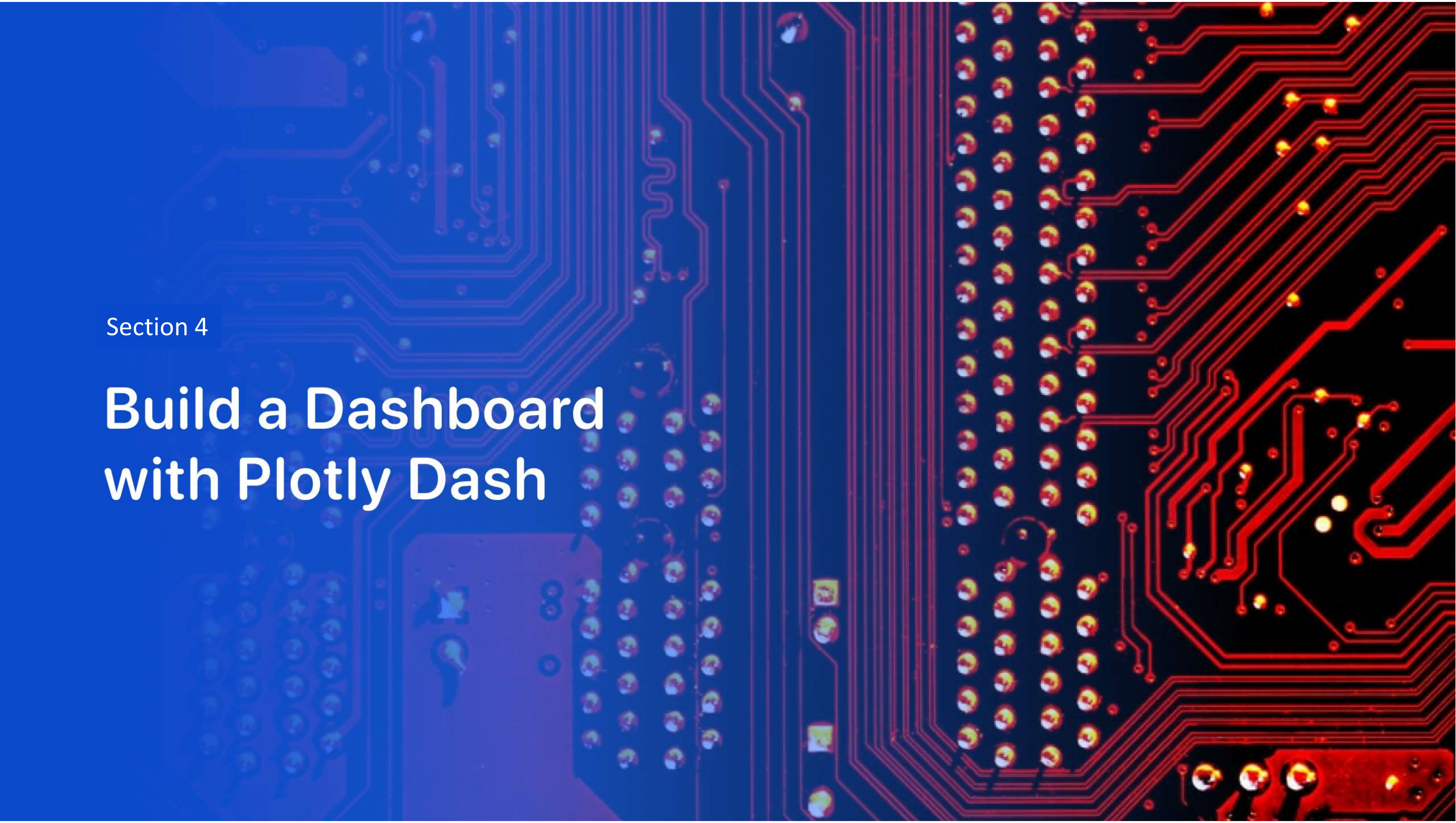
Folium Map Showing Launch Outcomes

- This screenshot shows a Folium map with color-labeled markers indicating the launch outcomes. Green markers represent successful launches, while red markers represent unsuccessful ones.

Launch Site Proximity Map



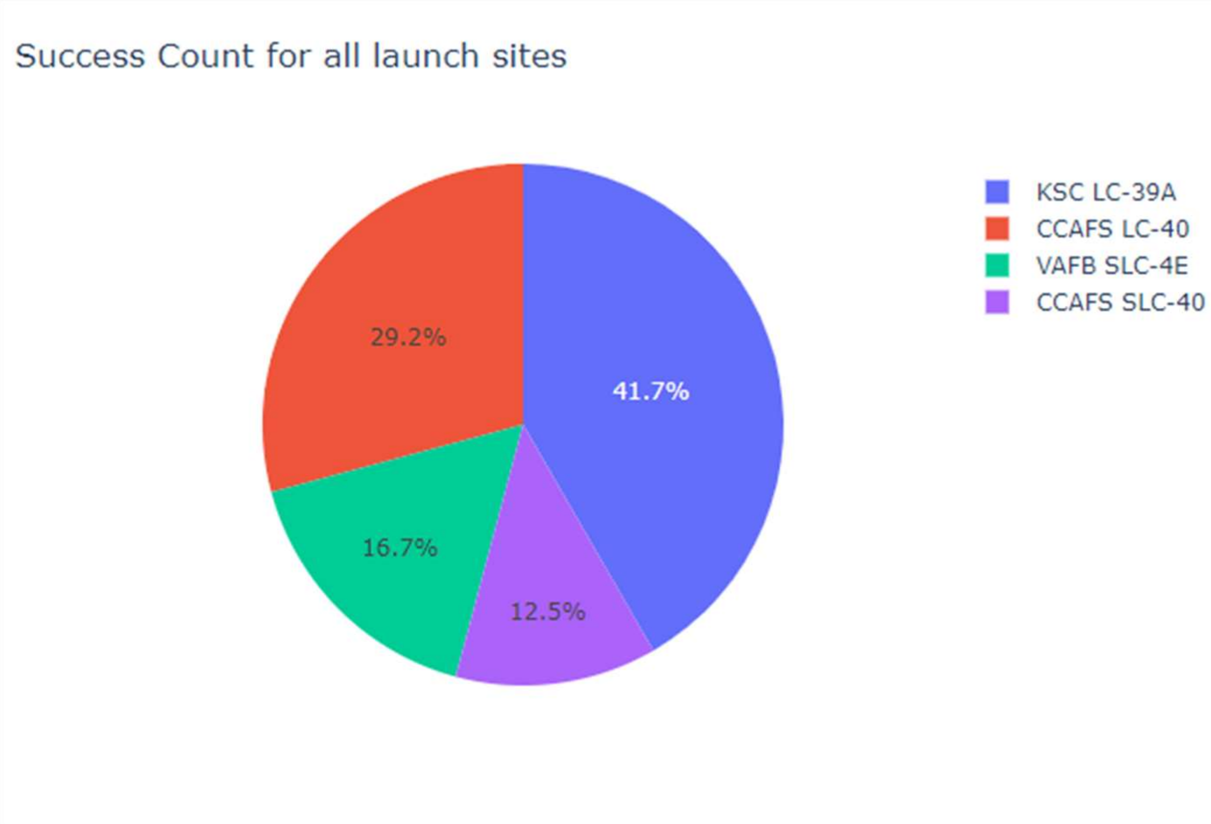
- This screenshot displays a Folium map featuring a selected launch site and its proximity to various geographical features, including a railway, highway, coastline, and distances calculated and displayed. The map utilizes markers and distance lines to illustrate these relationships, providing insights into the launch site's positioning relative to key transportation and geographical elements. This interactive visualization helps analyze the strategic placement of launch sites in relation to infrastructure and natural surroundings.



Section 4

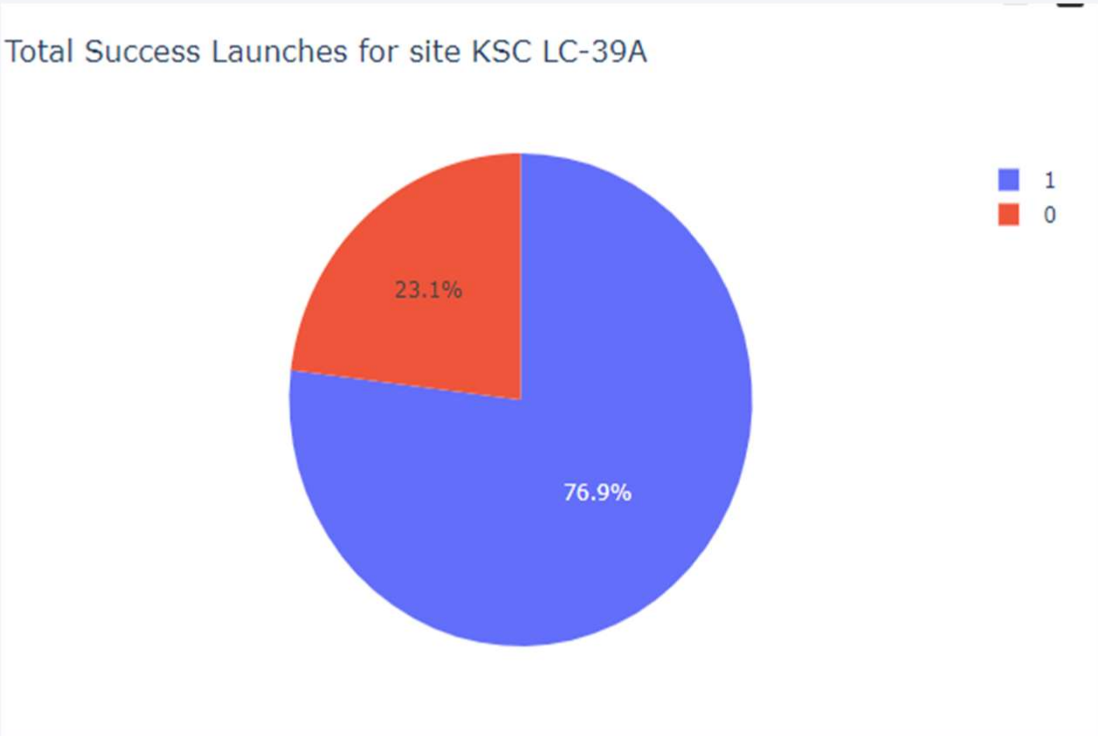
Build a Dashboard with Plotly Dash

Launch Success Count by Site (Pie Chart)



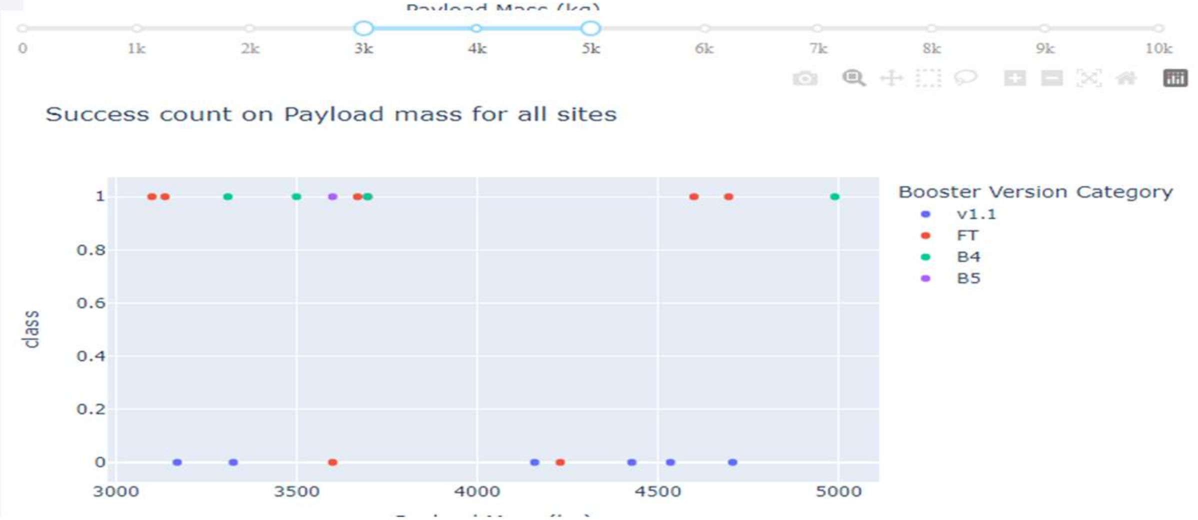
This screenshot showcases a pie chart visualizing the total number of successful launches across all launch sites. The chart provides a clear breakdown of success counts for each launch site, allowing viewers to compare their success rates. It offers valuable insights into which launch sites have been more successful in terms of mission outcomes.

Launch Success Ratio by Site (Pie Chart)



This screenshot displays a pie chart representing the launch success ratio for the launch site with the highest success rate. The chart visually illustrates the distribution of successful and failed launches for this specific site

Payload vs. Launch Outcome Analysis (Scatter Plot)



- These screenshots display scatter plots representing the relationship between payload mass and launch outcomes for all launch sites. The scatter plots allow users to explore how different payload ranges impact the success or failure of launches.
- By interacting with the payload range slider, users can customize the analysis to focus on specific payload mass ranges and gain insights into the impact of payload on launch success.

Section 5

Predictive Analysis (Classification)

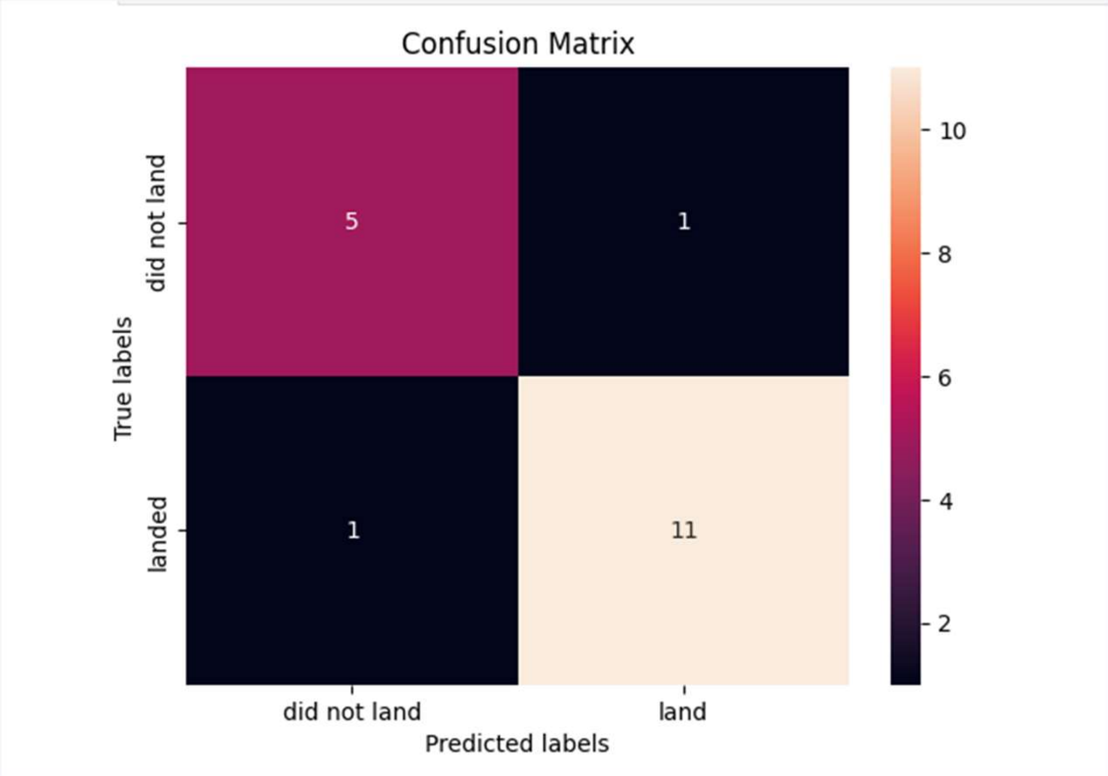
Classification Accuracy

- In this section, we will visualize and compare the accuracy of different classification models using a bar chart. We have evaluated four models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- Logistic Regression and SVM both achieved an accuracy of approximately 83.33% on the test data.
- The Decision Tree model outperformed other models with an accuracy of approximately 88.89% on the test data.
- KNN also achieved an accuracy of approximately 83.33% on the test data.

Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.888889
KNN	0.833333

Confusion Matrix

- True Negatives (TN): There are 11 True Negatives. These are cases where the model correctly predicted unsuccessful landings, and indeed, the Falcon 9 first stage did not land successfully.
- False Positives (FP): There is 1 False Positive. This is a case where the model incorrectly predicted a successful landing, but in reality, the Falcon 9 first stage did not land successfully.
- False Negatives (FN): There is 1 False Negative. This is a case where the model incorrectly predicted an unsuccessful landing, but in reality, the Falcon 9 first stage did land successfully.
- True Positives (TP): There are 5 True Positives. These are cases where the model correctly predicted successful landings, and indeed, the Falcon 9 first stage did land successfully.



Conclusions

- **Point 1**: The machine learning models, including Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors, were trained and evaluated to predict the success of Falcon 9 first stage landings.
- **Point 2**: Among these models, the Decision Tree classifier demonstrated the highest accuracy on the test dataset, achieving an accuracy rate of approximately 88.89%.
- **Point 3**: The confusion matrix analysis of the best-performing model (Decision Tree) revealed that it successfully identified most cases of both successful and unsuccessful landings, with only a small number of misclassifications.
- **Point 4**: The results indicate the potential for using machine learning to predict the outcome of Falcon 9 first stage landings, which could be valuable in optimizing launch operations and cost estimation for space missions.
- **Point 5**: Further research and refinement of machine learning models could enhance their predictive accuracy, ultimately contributing to the cost-efficiency and success rate of SpaceX missions.

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-science-capstone/home/welcome

Thank you!