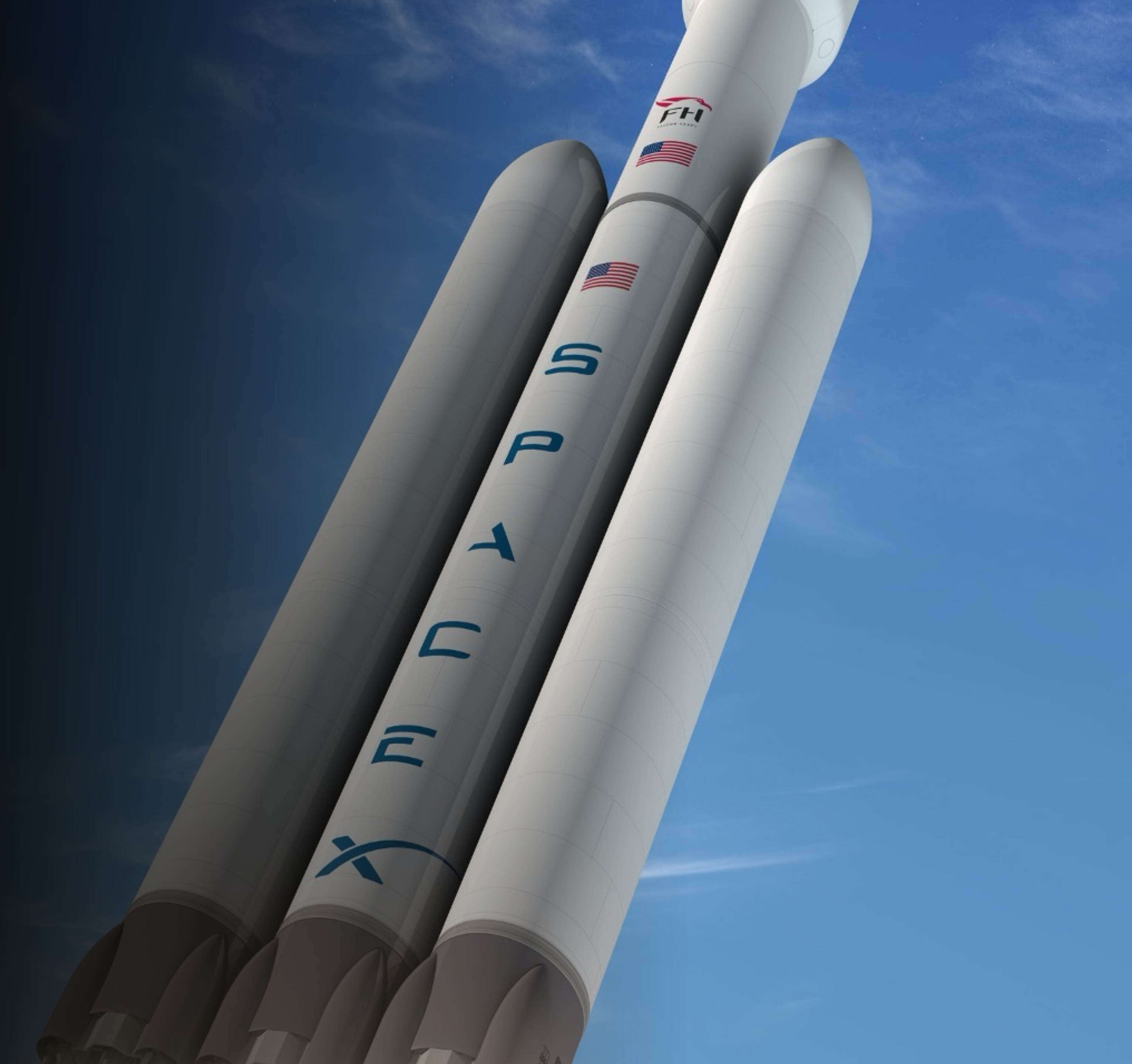
SpaceY

IBM Applied Data Science Capstone Final Project by

Mohamad Nour Alhendi



Outline

- Executive Summary: **Slide 3**
- Introduction: **Slide 4**
- Methodology: **Slides 5 - 13**
- Results: Slides **14 - 34**
- Conclusion: **Slide 35**

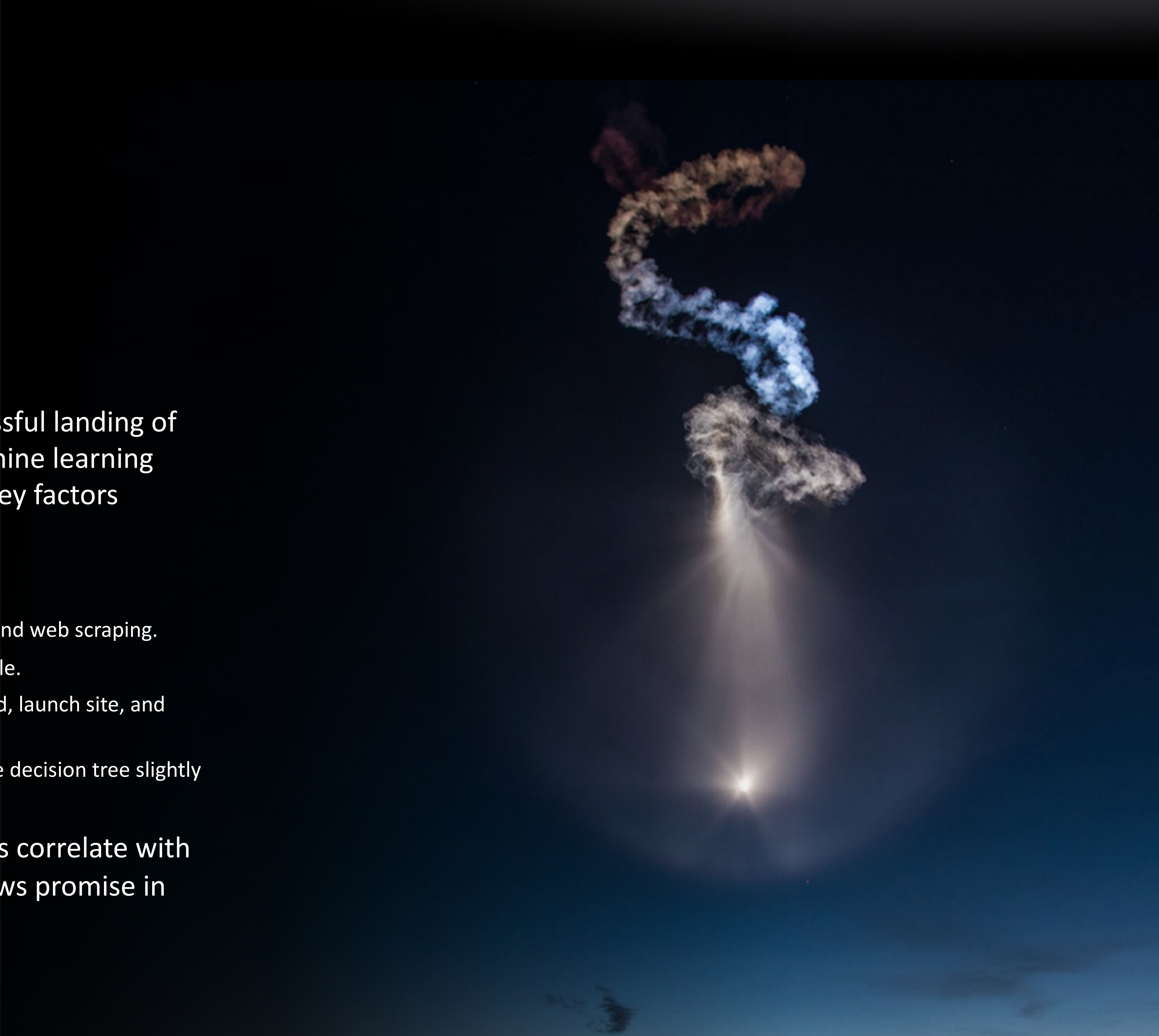
Executive Summary

This capstone project aimed to predict the successful landing of the SpaceX Falcon 9 first stage using various machine learning algorithms. The research focused on identifying key factors influencing landing success.

Key Methodologies:

- Data Collection: Gathered data via SpaceX's REST API and web scraping.
- Data Wrangling: Created a success/fail outcome variable.
- Exploratory Data Analysis: Analyzed factors like payload, launch site, and flight trends using data visualization.
- Model Building: Developed predictive models, with the decision tree slightly outperforming other algorithms.

Our findings highlight that certain launch features correlate with landing success, and the decision tree model shows promise in predicting outcomes.



Introduction

Background:

SpaceX, a pioneer in the space industry, has revolutionized space travel by making it more affordable, largely due to its ability to reuse the first stage of its Falcon 9 rocket. While SpaceX charges $62 million per launch, other providers, unable to reuse the first stage, charge over $165 million. Predicting whether the Falcon 9's first stage will land successfully allows us to estimate launch costs, which is valuable for both SpaceX and potential competitors. **Exploration**

Focus:

- How factors like payload mass, launch site, and orbit type affect first-stage landing success.
- Success rates of landings over time.
- Identifying the best predictive model for landing success (binary classification).

Methodology

Our approach involved several key steps to predict the landing success of the Falcon 9 first stage:

1. Data Collection & Preparation:
   - Gathered data using SpaceX's REST API and web scraping.
   - Wrangled the data by filtering, handling missing values, and applying one-hot encoding for analysis and modeling.
2. Exploratory Data Analysis (EDA):
   - Analyzed data using Pandas, NumPy, and SQL.
   - Visualized key insights with Matplotlib, Seaborn, Folium, and Plotly Dash.
3. Machine Learning Prediction:
   - Built and evaluated models using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbors (KNN) to determine the best predictive model.

Methodology: Data Collection - API

To gather and prepare the data for analysis, the following steps were taken:

- **API Requests**: Retrieved rocket launch data from the SpaceX API, specifically filtering for Falcon 9 launches using custom functions.
- **Data Processing**: Decoded API responses with .json(), converted them into dataframes, and created a dictionary to organize the data.
- **Data Cleaning**: Replaced missing values in the Payload Mass column with the column's mean.
- **Exporting**: The cleaned and filtered data was exported to a CSV file for further analysis.

Methodology: Data Collection – Web Scraping

- **Data Source**: Retrieved Falcon 9 launch data from Wikipedia.
- **Data Extraction**: Created a BeautifulSoup object from the HTML response to parse tables and extract relevant information.
- **Data Processing**: Extracted column names, collected data, and created a dictionary to organize the information. The data was then converted into a dataframe.
- **Data Cleaning**: Ensured no missing entries were present. Categorical features were encoded using one-hot encoding. An additional column, ‘Class,’ was introduced to indicate the outcome (0 for failure, 1 for success).
- **Final Dataset**: The processed data resulted in a dataset with 90 rows (instances) and 83 columns (features), which was exported to a CSV file.

Methodology: Data Wrangling

- **Exploratory Data Analysis (EDA):** Conducted EDA to identify key data labels and calculate metrics such as the number of launches per site, occurrences of different orbits, and mission outcomes per orbit type.
- **Landing Outcome Classification**:
    - True Ocean: Successful landing in a specific ocean region.
    - False Ocean: Unsuccessful landing in a specific ocean region.
    - True RTLS: Successful landing on a ground pad.
    - False RTLS: Unsuccessful landing on a ground pad.
    - True ASDS: Successful landing on a drone ship.
    - False ASDS: Unsuccessful landing on a drone ship.
- **Binary Classification**: Converted the landing outcomes into a binary format (1 for successful landing, 0 for unsuccessful) and added it as a dependent variable.

The cleaned and processed data, including the binary landing outcome, was exported for further analysis and modeling.

Methodology: Exploratory Data Analysis (EDA) with Visualization and SQL 1/2

EDA with Pandas, NumPy, and SQL:

- **Data Analysis:**
  - Used Pandas and NumPy to derive key insights such as the number of launches at each site, the frequency of each orbit, and the occurrence of mission outcomes.
- **Queried the data** using SQL to answer specific questions:
  - Identified unique launch sites.
  - Calculated total payload mass for NASA (CRS) missions and average payload mass for booster version F9 v1.1.
  - Listed important details such as the first successful ground pad landing, boosters with successful drone ship landings, and missions carrying the maximum payload.

Methodology: Exploratory Data Analysis (EDA) with Visualization and SQL 2/2

Visualization with Matplotlib and Seaborn:

- **Charts and Plots**:
  - Scatter Plots: Explored relationships between variables like flight number and payload mass across different launch sites and orbit types.
  - Bar Charts: Compared discrete categories to visualize relationships between launch sites, payload mass, and success rates.
- **Key Visualizations:**
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass (kg) vs. Launch Site
  - Payload Mass (kg) vs. Orbit Type

These analyses and visualizations provided valuable insights into the data, revealing relationships that could be leveraged in machine learning models.

Methodology: Map Visualization with Folium

Interactive Maps:

- **Launch Sites:**
  - Markers: Added blue circles for NASA Johnson Space Center and red circles for other launch sites, with popup labels showing their names.
- **Launch Outcomes:**
  - Colored Markers: Used green markers for successful launches and red markers for unsuccessful ones to highlight high-success-rate sites.
- **Distances:**
  - Colored Lines: Indicated distances between the CCAFS SLC40 launch site and its nearest coastline, railway, highway, and city.

Folium was employed to create interactive maps that visualize launch sites, outcomes, and proximities, enhancing the understanding of launch site locations and their success rates.

Methodology: Plotly Dash

Interactive Features:

- **Dropdown List**: Allows users to select specific launch sites or view data for all sites.
- **Payload Mass Range Slider**: Lets users filter data based on payload mass ranges.
- **Pie Chart**: Displays the percentage of successful versus unsuccessful launches.
- **Scatter Chart**: Shows the correlation between payload mass and launch success rate by booster version.

Using Dash, we created an interactive dashboard that enables users to explore launch site data, visualize success rates, and analyze the impact of payload mass on launch outcomes.

Methodology: Predictive Analytics

Machine Learning Models:

- **Data Preparation:**
  - Created a NumPy array from the Class column.
  - Standardized the data using StandardScaler and split it into training and test sets with train, test, split.
- **Model Building and Optimization:**
  - Used GridSearchCV with 10-fold cross-validation to optimize parameters for:
    - Logistic Regression (LogisticRegression())
    - Support Vector Machine (SVC())
    - Decision Tree (DecisionTreeClassifier())
    - K-Nearest Neighbor (KNeighborsClassifier())
- **Evaluation:**
  - Calculated accuracy for each model using .score().
  - Assessed model performance with confusion matrices.
  - Identified the best model using metrics such as Jaccard Score, F1 Score, and Accuracy.

Scikit-learn functions facilitated the creation, tuning, and evaluation of machine learning models, enabling us to determine the most effective model for predicting Falcon 9 landing success.

Results

Exploratory Data Analysis:

- Trend: Launch success rates have improved over time.
- Top Site: KSC LC-39A shows the highest success rate.
- Orbits: ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Visual Analytics:

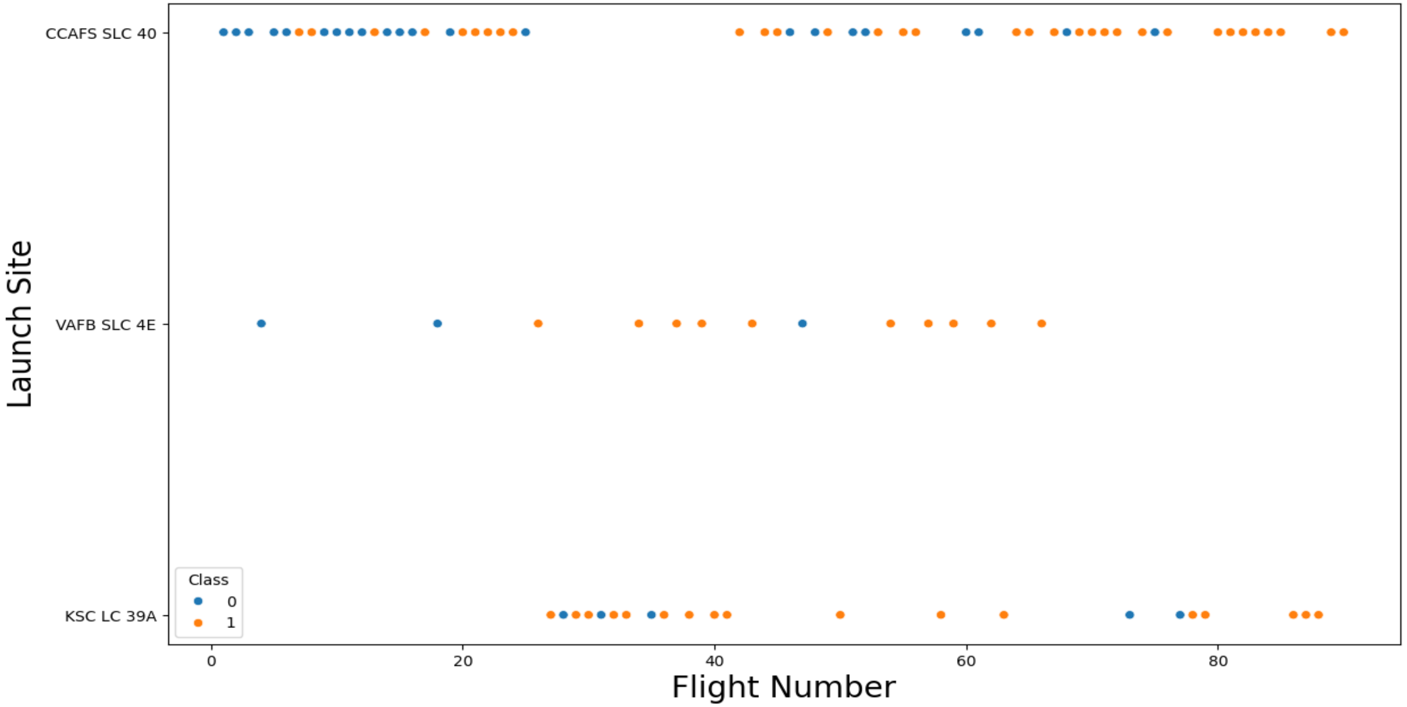
- Geographic Insights: Most launch sites are near the equator and close to the coast.
- Safety: Launch sites are strategically positioned away from urban areas, highways, and railways to mitigate damage from failed launches while maintaining accessibility for support activities.

Predictive Analytics:

- Best Model: The Decision Tree model performs best for predicting Falcon 9 landing success.

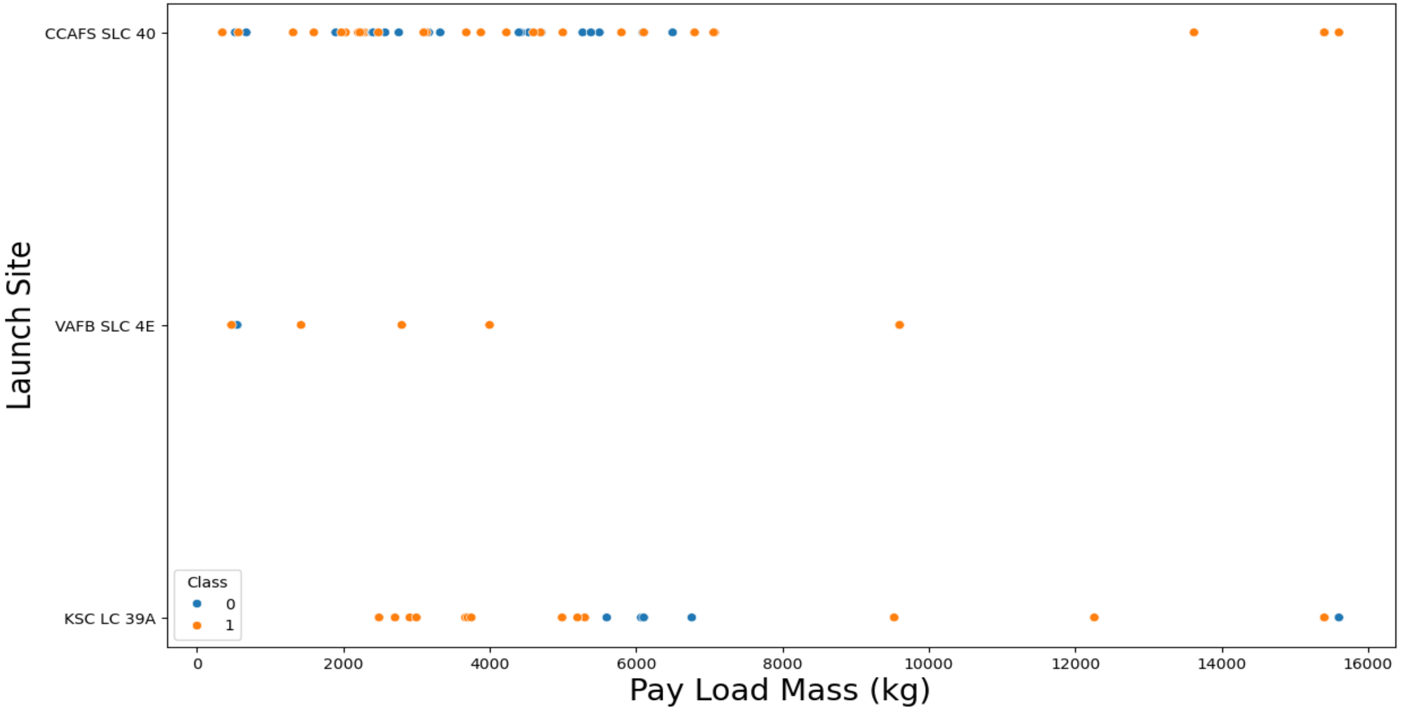
Results: Flight Number vs. Launch Site

- Here is the relation ship between flight number and launch site:



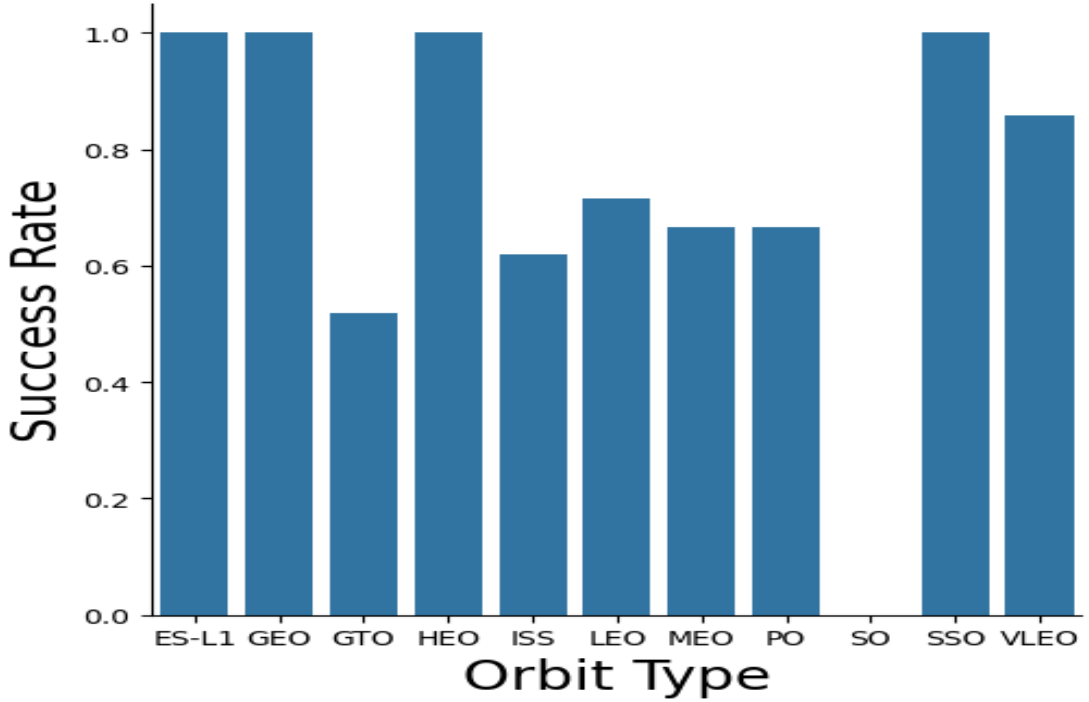
Results: Payload vs. Launch Site

- **Here is the relation ship between payload and launch site:**



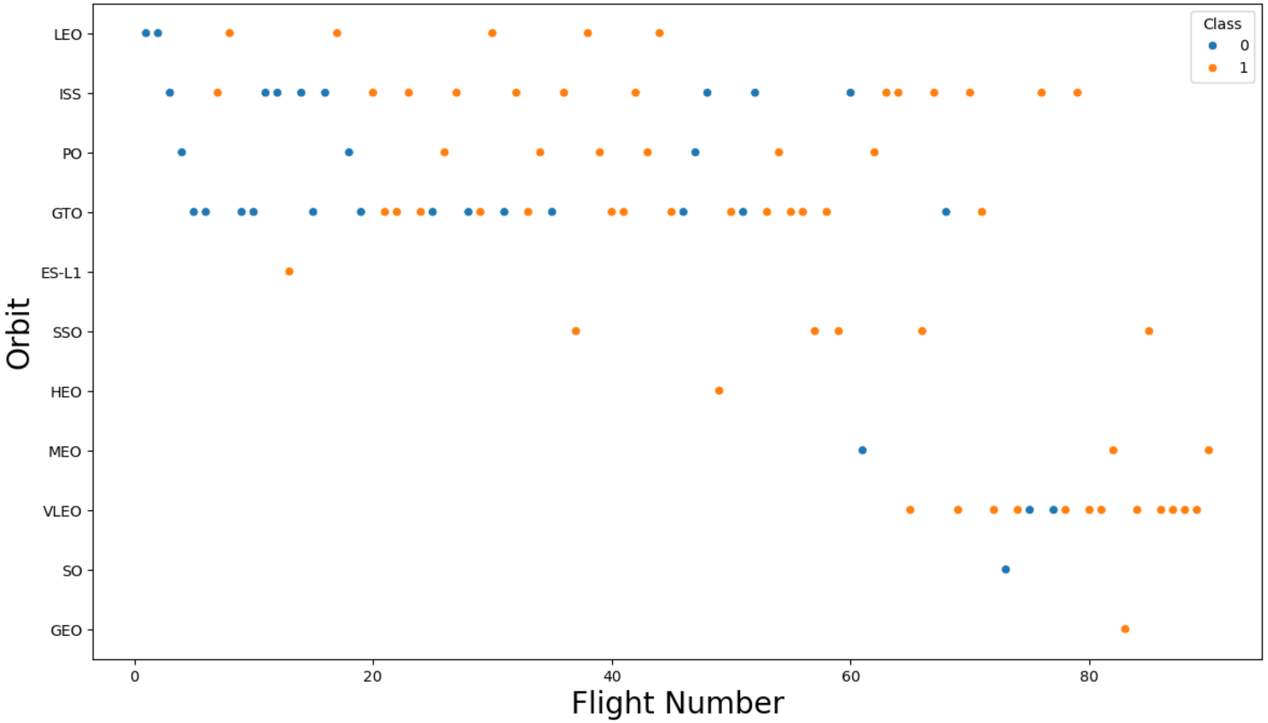
Results: Success Rate by Orbit

- Here is the Success Rate by Orbit:



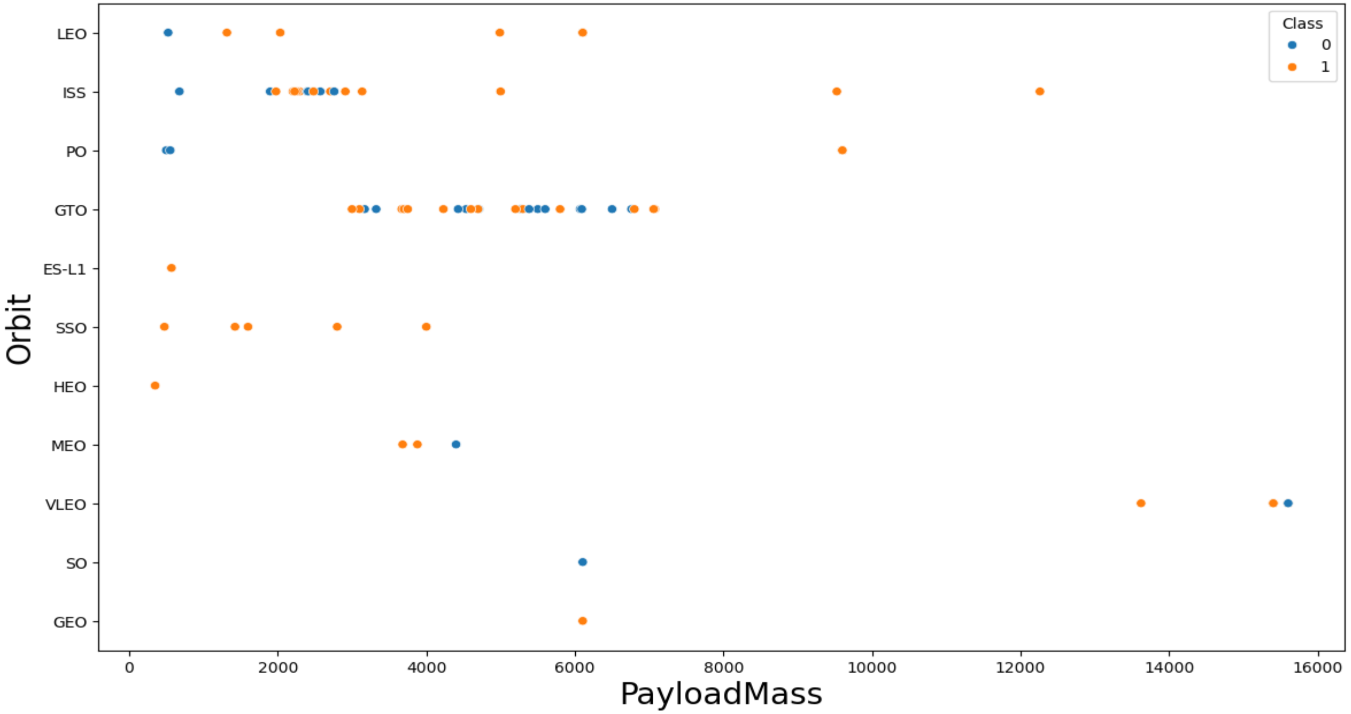
Results: Flight Number vs. Orbit

- Here is the relation ship between flight number and orbit:



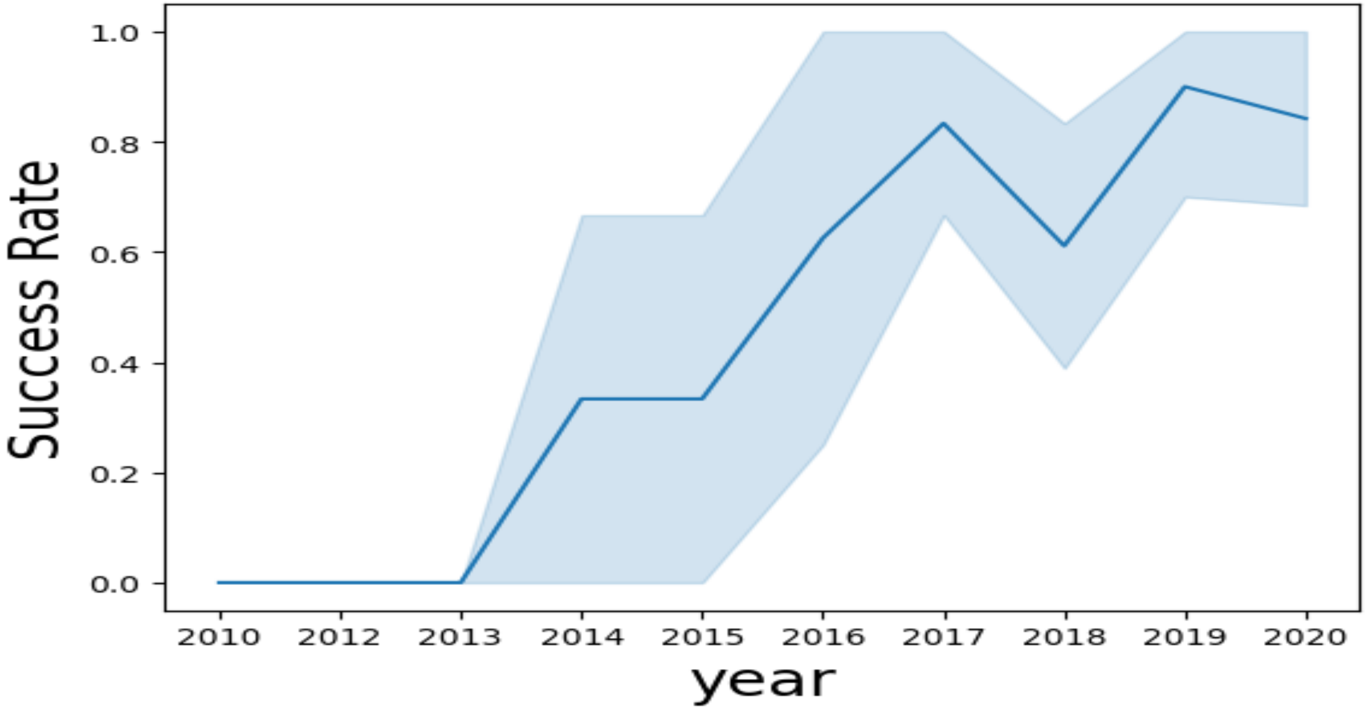
Results: Payload vs. Orbit

- Here is the relation ship between Payload and orbit:



Results: Launch Success over Time

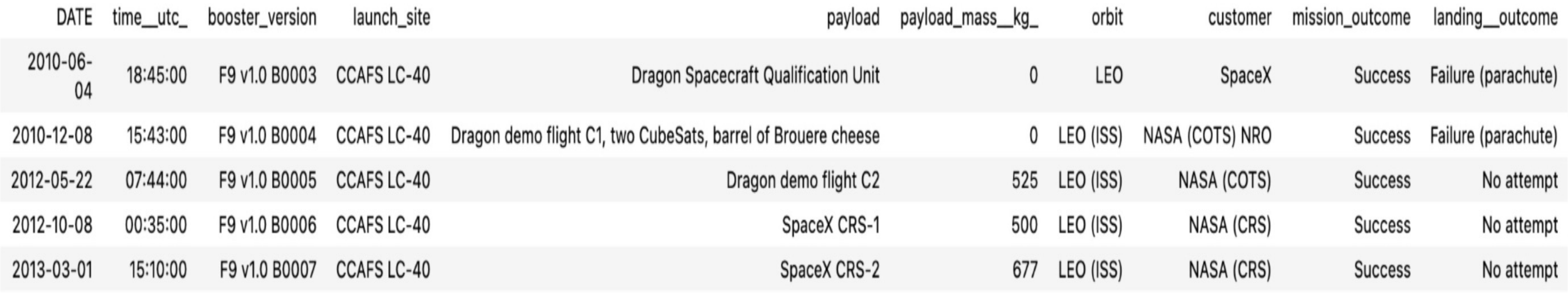
- Here is the launch success rate yearly trend:



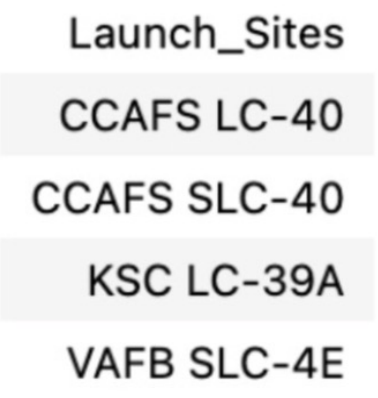
Results: Launch Site Information

Here are the:

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with ‘CCA’



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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

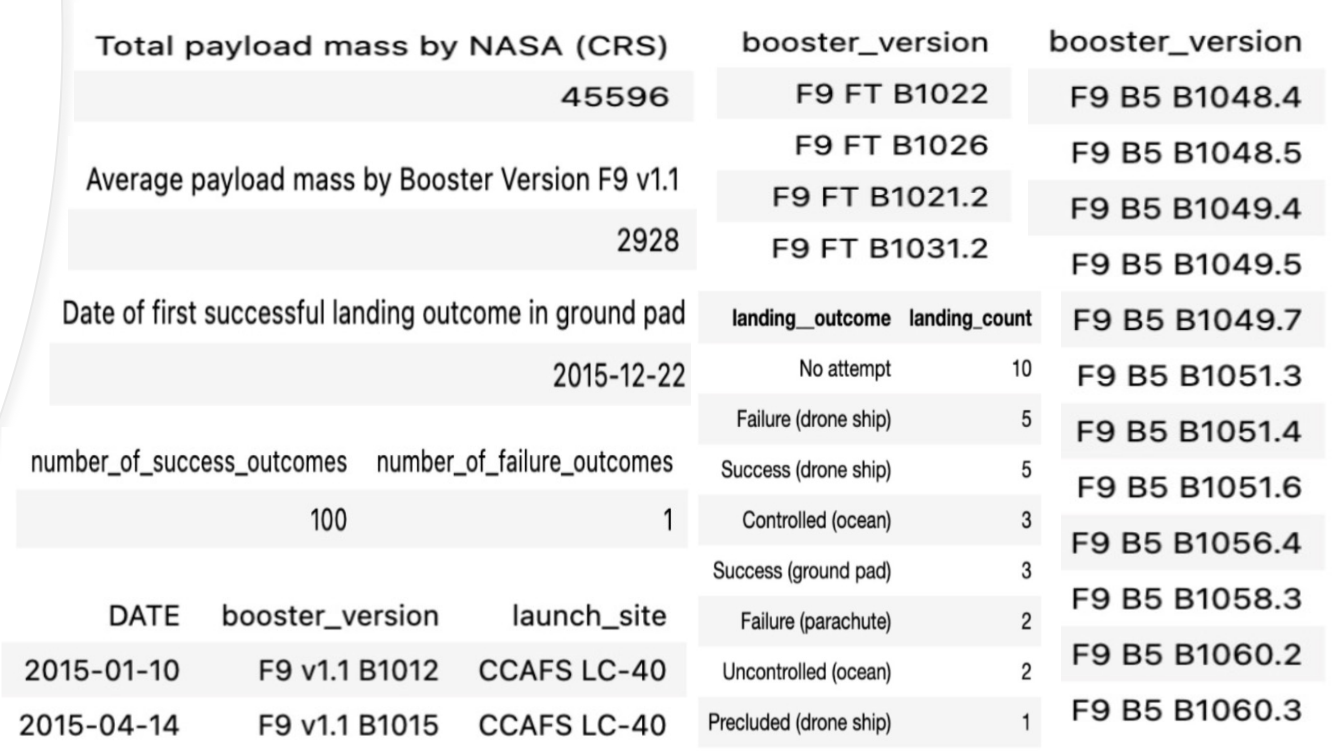


Launch_Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Results: Payload Mass, Landing & Mission Info, Boosters, Failed Landings on Drone Ship, Count of Successful Landings

Here are the:

- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The names of the booster versions which have carried the maximum payload mass
- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.



Total payload mass by NASA (CRS)
45596

Average payload mass by Booster Version F9 v1.1
2928

Date of first successful landing outcome in ground pad
2015-12-22

number_of_success_outcomes	number_of_failure_outcomes
100	1

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

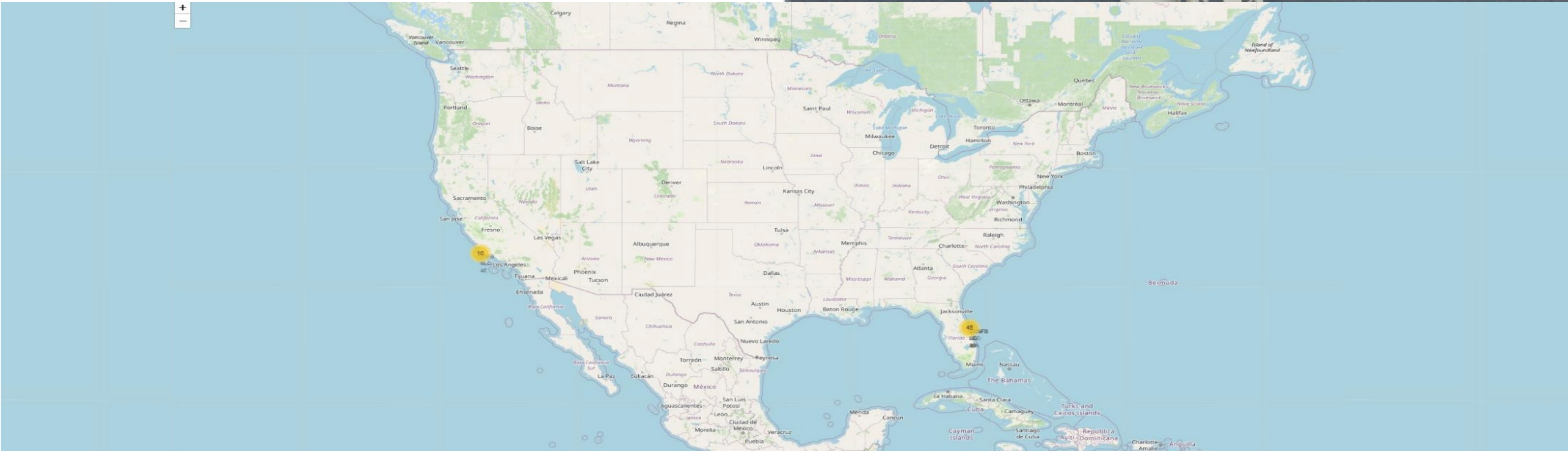
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

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

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

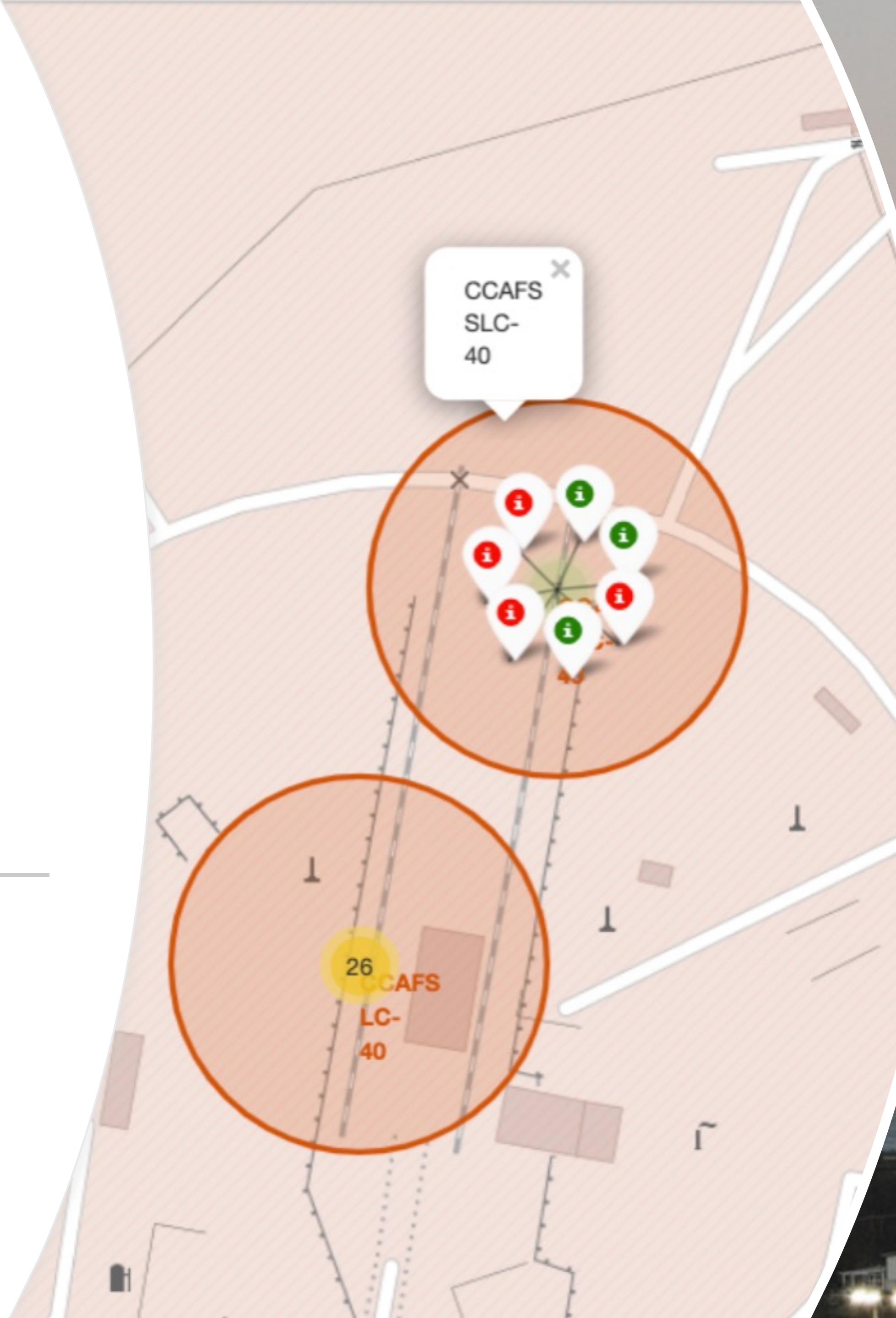
Results: Launch Sites

- **Here are all the launch sites on the map:**



Results: Launch Outcomes

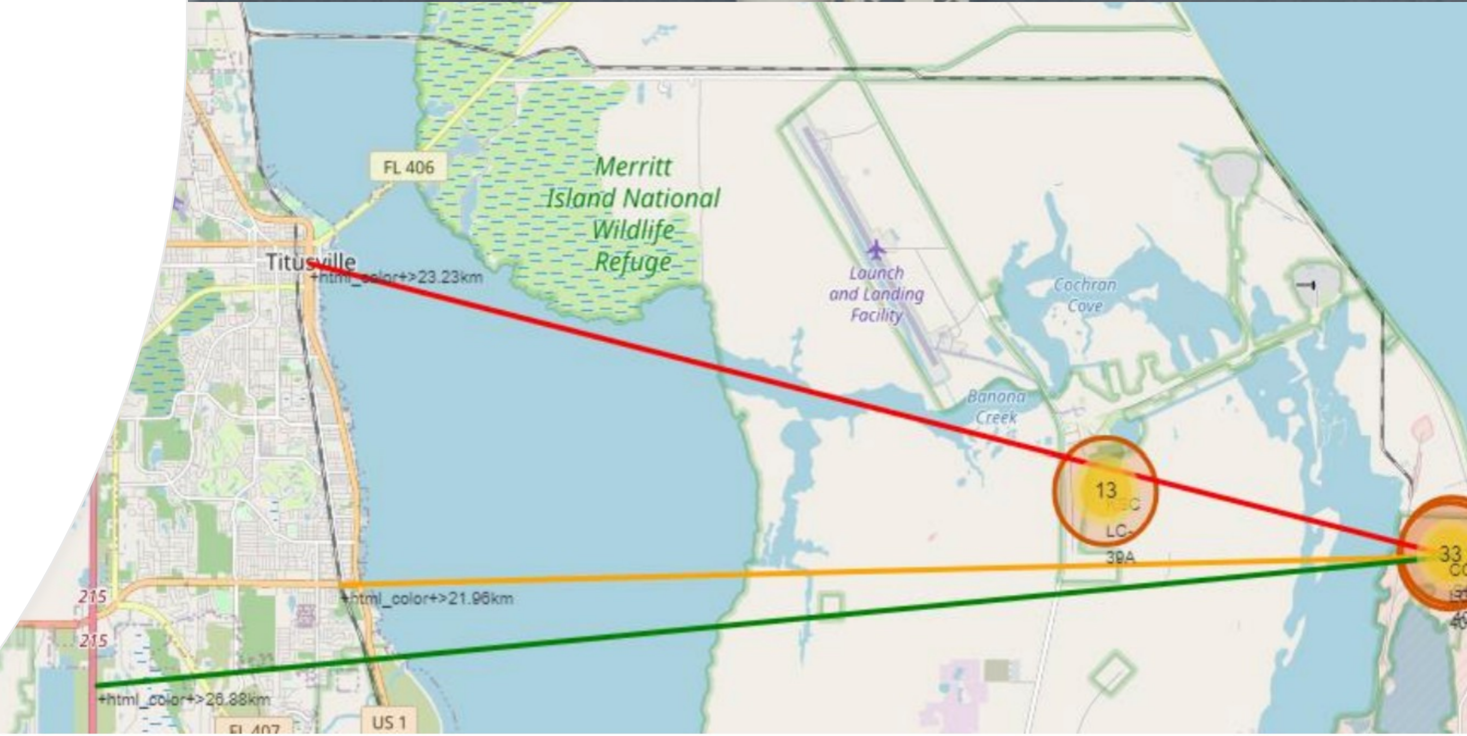
- Here are the launch outcomes:
- 3/7 Success rate (43%)



Results: Distance to Proximities

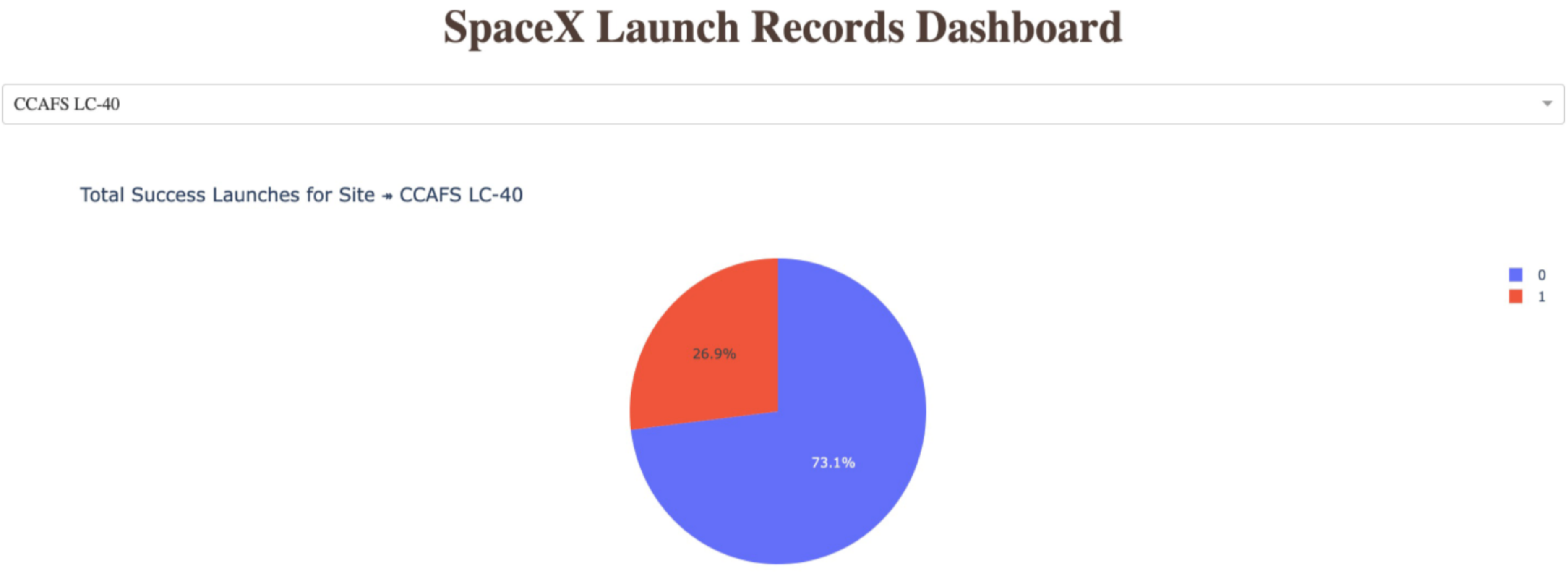
Here are the distance to proximities:

- **CCAFS SLC-40:**
  - .86 km from nearest coastline
  - 21.96 km from nearest railway
  - 23.23 km from nearest city
  - 26.88 km from nearest highway



Results: Launch Success (KSC LC-29A)

- Here is a pie chart when launch site CCAFS LC-40 is chosen obtaining 26.9% success rate: (Red = Success, Blue = Failure)



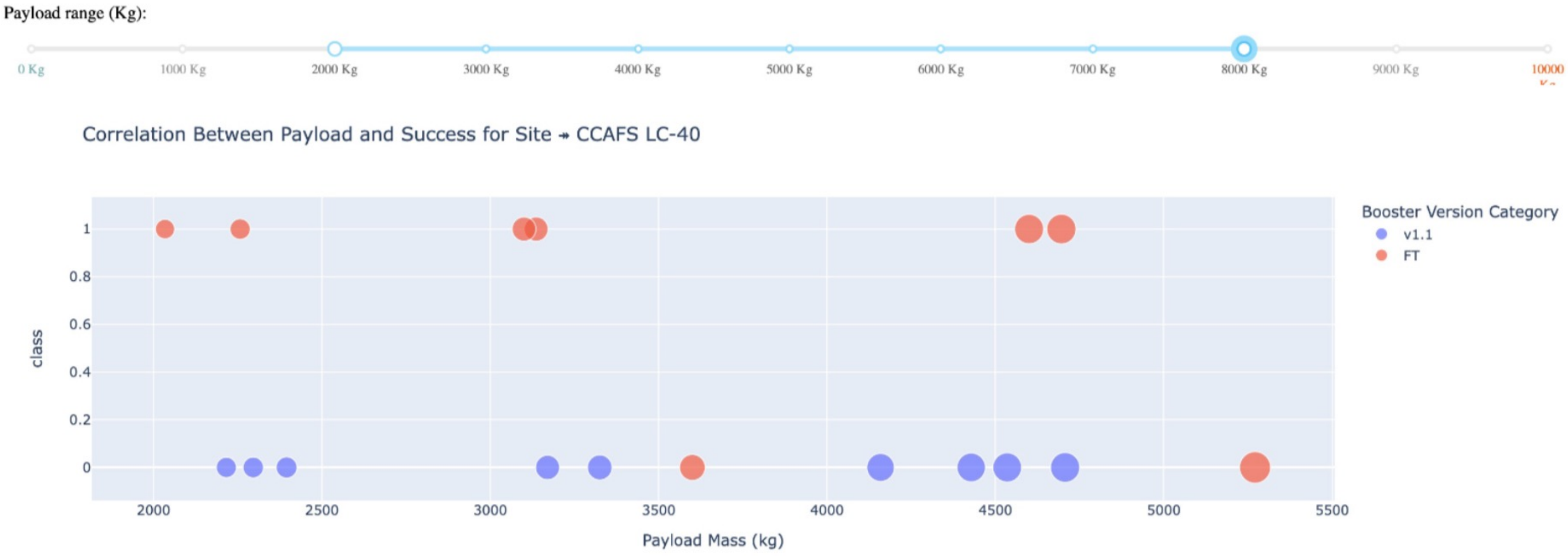
SpaceX Launch Records Dashboard

CCAFS LC-40

Total Success Launches for Site → CCAFS LC-40

Results: Payload Mass and Success

- Here is a scatter plot of when the payload mass range is set to be from 2000kg to 8000kg: (1 = successful, 0 = unsuccessful)



Payload range (Kg):

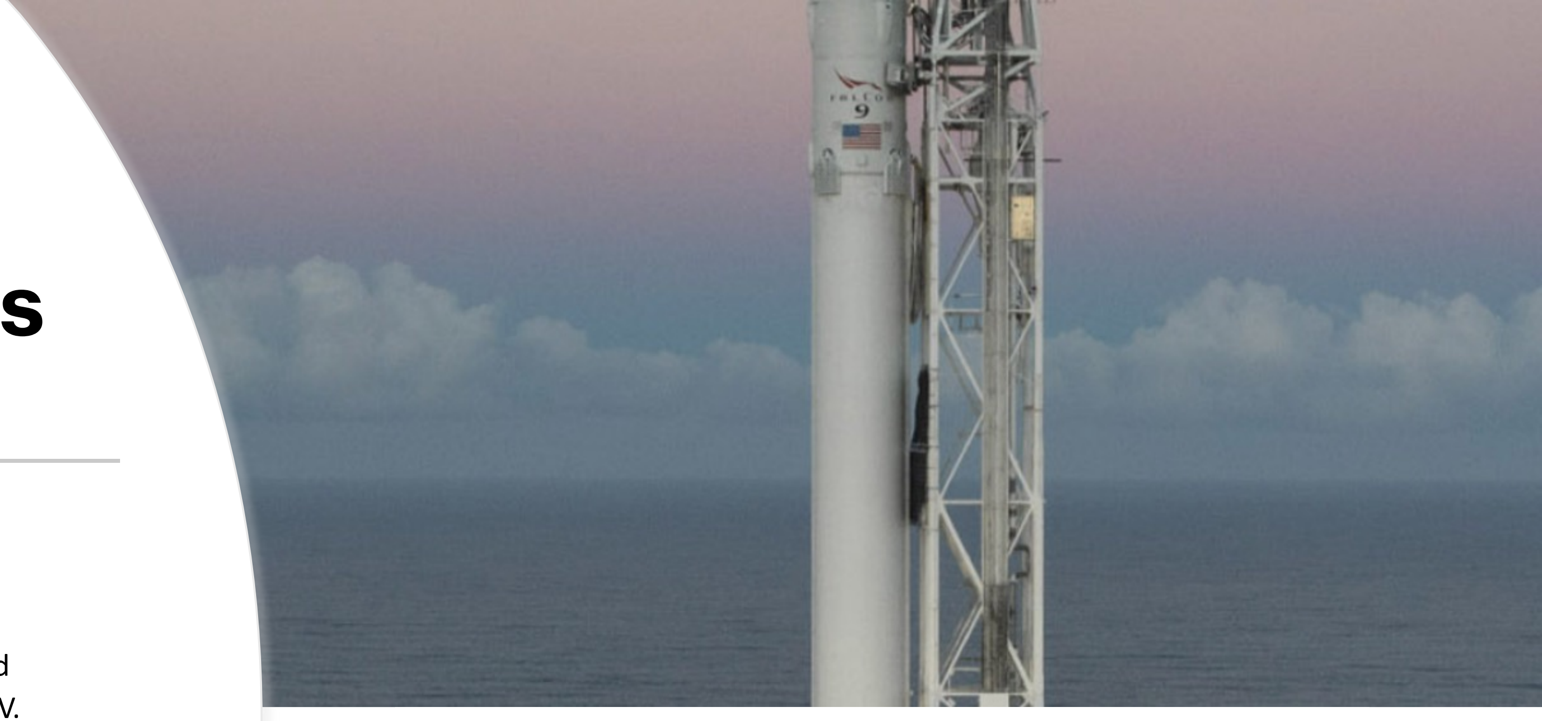
Results: Classification and Confusion Matrices 1/2

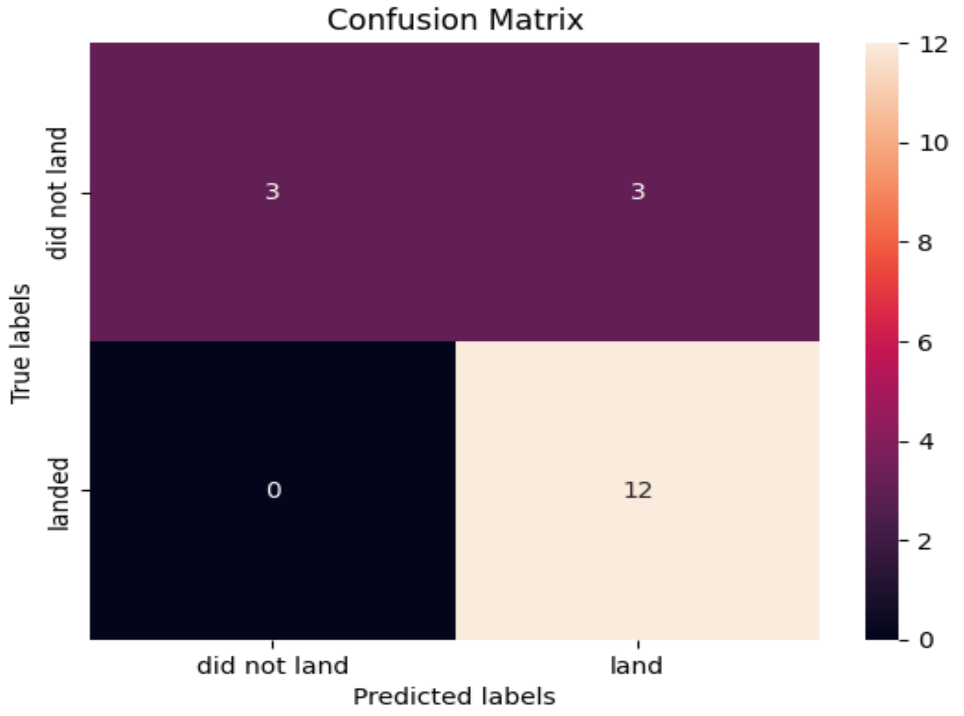
Model Performance:

- Overall Results: All models showed similar accuracy and confusion matrices, likely due to the small dataset.
- Best Model: The Decision Tree model slightly outperformed others based on the .best_score_ metric from GridSearchCV.
- Confusion Matrix: Identical across all models with:
  - True Positives: **12**
  - True Negatives: **3**
  - False Positives: **3**
  - False Negatives: **0**

Metrics:

- Precision: 0.80 (Calculated as TP / (TP + FP))
- Recall: 1.00 (Calculated as TP / (TP + FN))
- F1 Score: 0.89 (Calculated as 2 * (Precision * Recall) / (Precision + Recall))
- Accuracy: 0.83 (Calculated as (TP + TN) / (TP + TN + FP + FN))





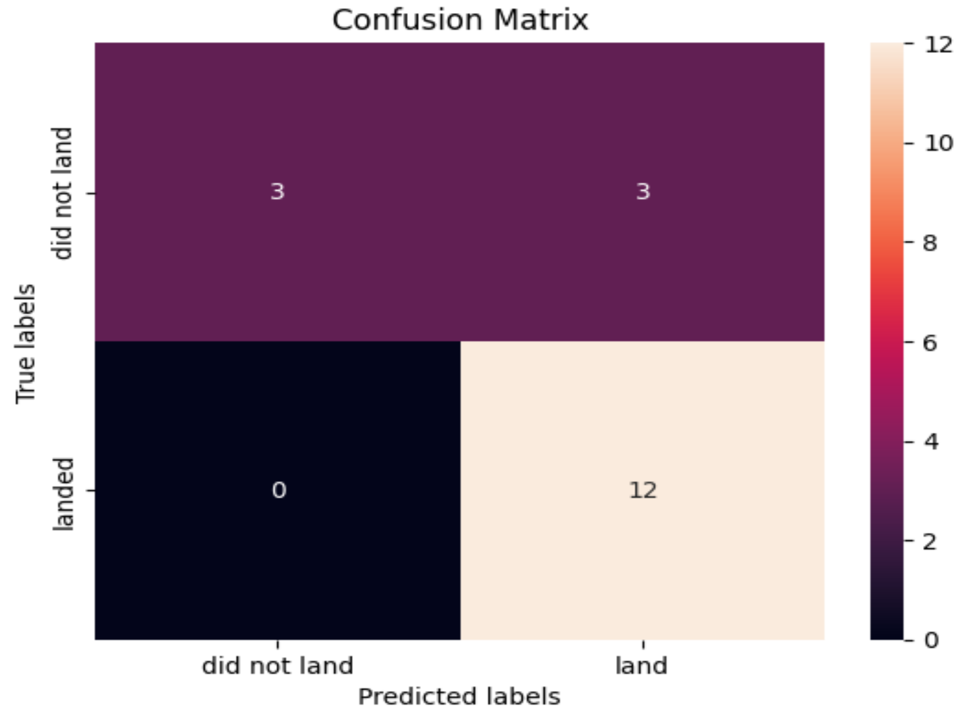
Results: Classification and Confusion Matrices 2/2

Model Ranking:

Based on GridSearchCV best scores:

- Decision Tree: **0.889**
- K-Nearest Neighbors (KNN): 0.848
- Support Vector Machine (SVM): **0.848**
- Logistic Regression: **0.846**

This ranking indicates the Decision Tree model is the most effective for predicting Falcon 9 landing success.



Conclusion

- The study found that **the Decision Tree model** slightly outperformed others in predicting Falcon 9 launch outcomes, though all models had similar performance due to the small dataset. Launch sites near the equator and coastlines, like KSC LC-39A, had higher success rates. Orbits such as ES-L1, GEO, HEO, and SSO were 100% successful, and higher payloads generally correlated with better outcomes.

- Future research could benefit from a larger dataset, additional feature analysis, and exploring models like XGBoost to improve prediction accuracy.