IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vy Vu
Oct 24, 2024

Outline

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

Executive Summary

In this project, a competitor to SpaceX analyzes Falcon 9 rocket data to assess the success rate of first-stage landings and the cost per launch. The following sections provide a summary of the methods used and the results obtained.

Summary of methodologies

- Data Collection
- Data wrangling
- Exploratory data analysis with data visualization and SQL
- Building an interactive map with Folium
- Building Dashboard with Plotly Dash
- Predictive analysis (classification)

Summary of all results

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

Introduction

- This capstone project, part of the IBM Data Science Professional Certificate, showcases skills in data science and machine learning by applying them to a real-world scenario, with the results documented in a report.
- The project revolves around a competitor, SpaceY, analyzing SpaceX's Falcon 9 rocket data to assess the success rate of first-stage landings and estimate launch costs. SpaceY uses these insights to submit competitive bids against SpaceX. While SpaceX promotes a launch cost of $62 million for the Falcon 9, other providers charge upwards of $165 million per launch.
- All data processing and analysis were conducted using Python within Jupyter notebooks, with both the notebooks and the final report hosted on my GitHub repository.
- The report covers key aspects such as data collection, cleaning, exploratory analysis (EDA), interactive visualizations, and the development and evaluation of machine learning (ML) models. It concludes by comparing the predictive performance of various ML algorithms in forecasting the outcome of future Falcon 9 first-stage landings.

Section 1

Methodology

Methodology

Executive Summary

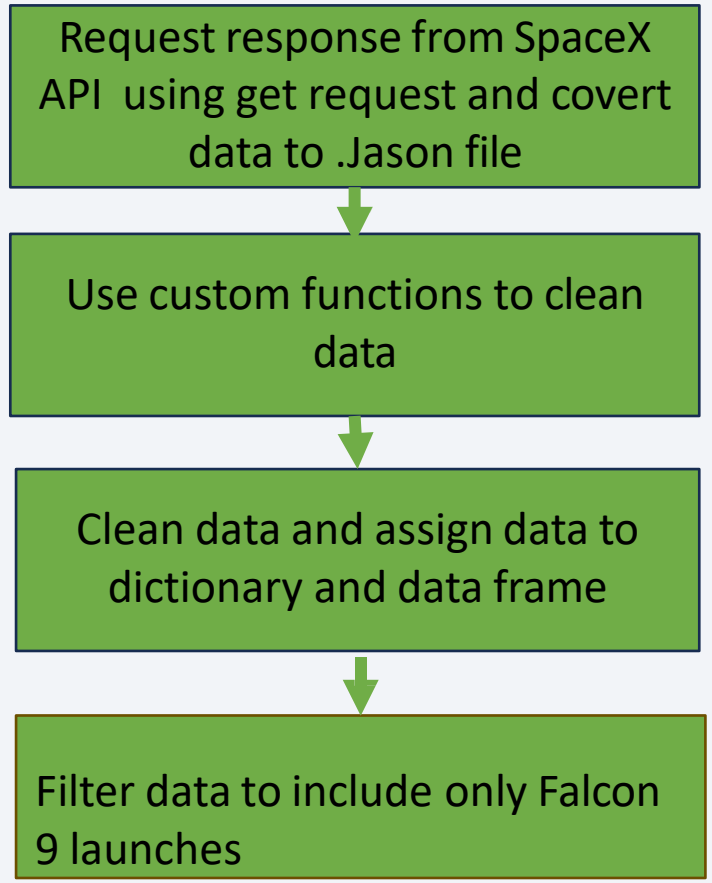
- This project utilized data sourced from the SpaceX REST API and the Wikipedia launch table. The data wrangling process involved cleaning, organizing, and preparing the data for visualization, as well as extracting key insights for machine learning models, including logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbors (KNN).
- Exploratory data analysis (EDA) was performed using both visual tools and SQL queries. The Python libraries Folium and Plotly Dash were employed to enhance data presentation and provide interactive visual analytics.
- For predictive analysis, classification models were developed to determine the likelihood of a successful Falcon 9 first-stage landing. The models were implemented using Scikit-learn, and their performance was evaluated based on accuracy metrics.

Data Collection

- Step 1: Gather Data from SpaceX API and convert data to .jason file
- Step 2: Scrap and filter data to include Falcon 9 data, assign data to dataframe and dictionary, and export data to a csv file
- Step 3: Plot and visualize the data

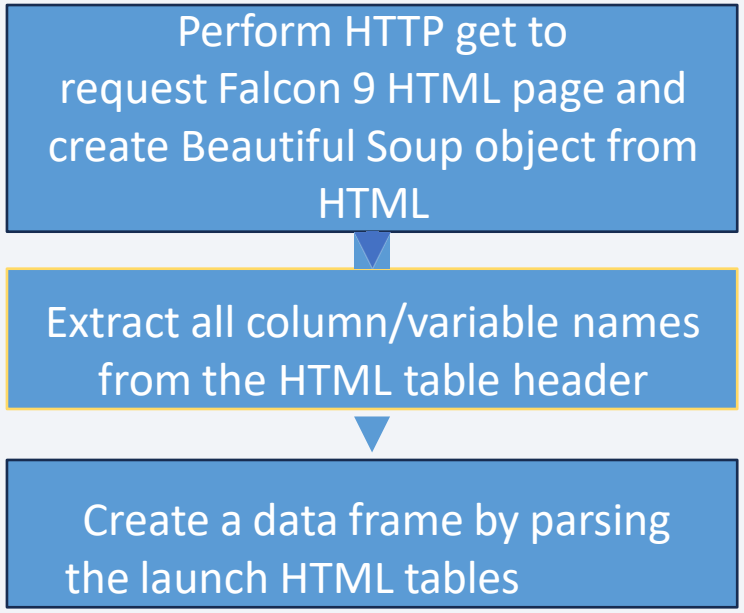
Data Collection – SpaceX API

https://github.com/SereneB2099/ibm-capstone/blob/main/1.%20Data%20Collection%20API.ipynb



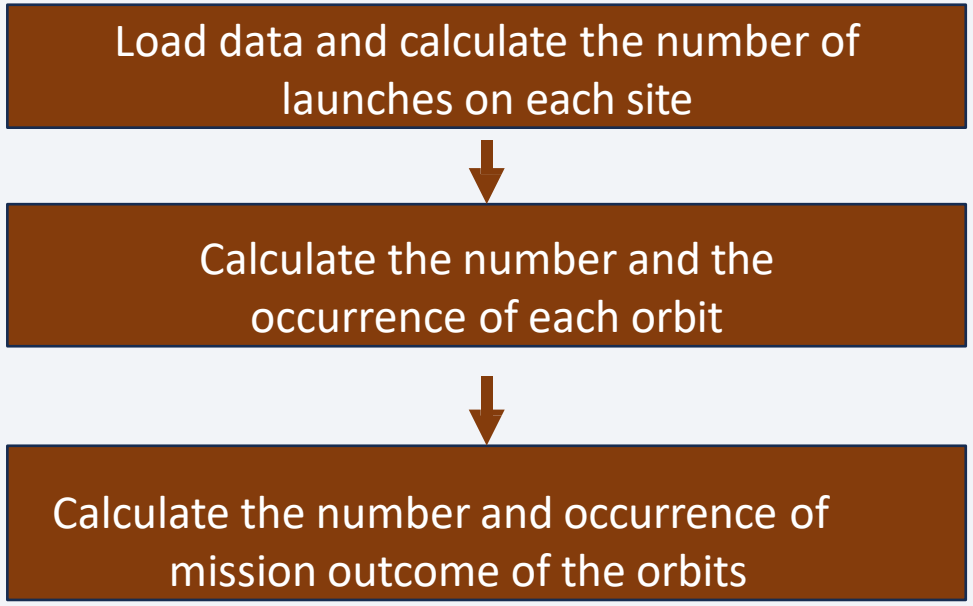
Data Collection - Scraping

https://github.com/SereneB2099/ibm-capstone/blob/main/2.%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

https://github.com/SereneB2099/ibm-capstone/blob/main/3.%20Data%20Wrangling.ipynb



EDA with Data Visualization

- FlightNumber vs. PayloadMass (Scatter Plot)

Purpose: Analyze the impact of flight number and payload mass on launch success

- FlightNumber vs. LaunchSite (Scatter Plot)

Purpose: Explore the influence of launch site on success rate over time

- PayloadMass vs. LaunchSite (Scatter Plot)

Purpose: Examine how payload mass affects success rates at different sites

- Orbit vs. Success Rate (Bar Plot)

Purpose: Evaluate success rate for different orbits

https://github.com/SereneB2099/ibm-capstone/blob/main/5.%20EDA%20with%20Visualization.ipynb

EDA with SQL

Key SQL Queries:

- **Create Table:** Generated a table containing only non-null dates.
- **Distinct Launch Sites:** Extracted a list of unique launch sites.
- **Total Payload for NASA:** Computed the total payload carried for NASA's CRS missions.
- **First Successful Landing:** Identified the date of the first successful ground pad landing.
- **Mission Outcomes:** Counted and grouped various mission outcomes

https://github.com/SereneB2099/ibm-capstone/blob/main/4.%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- **Summary of Map Objects in Folium:**
- **Markers:** Placed to mark the locations of launch sites.
- **Circles:** Added to emphasize areas around launch sites, representing regions of interest.
- **Lines:** Drawn to link various launch sites, showing routes or trajectories.
- **Purpose of Map Objects:**
- **Markers:** Pinpoint the exact positions of launch sites on the map.
- **Circles:** Visually represent the surrounding areas, potentially indicating safety zones or zones of influence.
- **Lines:** Depict connections or paths between launch sites, helping to understand spatial relationships and logistics.

https://github.com/SereneB2099/ibm-capstone/blob/main/6.%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Dropdown Menu:

Lets users choose a specific launch site or view data across all sites.

- Pie Chart:

Visualizes the total successful launches per site, dynamically updating based on the dropdown selection.

- Range Slider:

Allows users to filter data by adjusting the range of payload mass.

- Scatter Plot:

Displays the relationship between payload mass and launch success, updating according to the selected site and payload range.

Predictive Analysis (Classification)

- **Create NumPy Array:** Extracted from the 'Class' column.
- **Standardize Data:** Used `StandardScaler` to fit and transform the data for consistency.
- **Train-Test Split:** Divided the dataset using `train_test_split` for model training and evaluation.
- **GridSearchCV Setup:** Configured a `GridSearchCV` object with `cv=10` for hyperparameter tuning.
- **Model Application:** Applied `GridSearchCV` on multiple algorithms, including:
- Logistic Regression (`LogisticRegression()`)
- Support Vector Machine (`SVC()`)
- Decision Tree (`DecisionTreeClassifier()`)
- K-Nearest Neighbors (`KNeighborsClassifier()`)
- **Model Accuracy:** Calculated performance on the test set using `.score()` for all models.
- **Confusion Matrix:** Evaluated the confusion matrix for each model to analyze predictions.
- **Best Model Selection:** Identified the top-performing model using Jaccard Score, F1 Score, and Accuracy.

https://github.com/SereneB2099/ibm-capstone/blob/main/7.%20Machine%20Learning%20Prediction.ipynb

Results

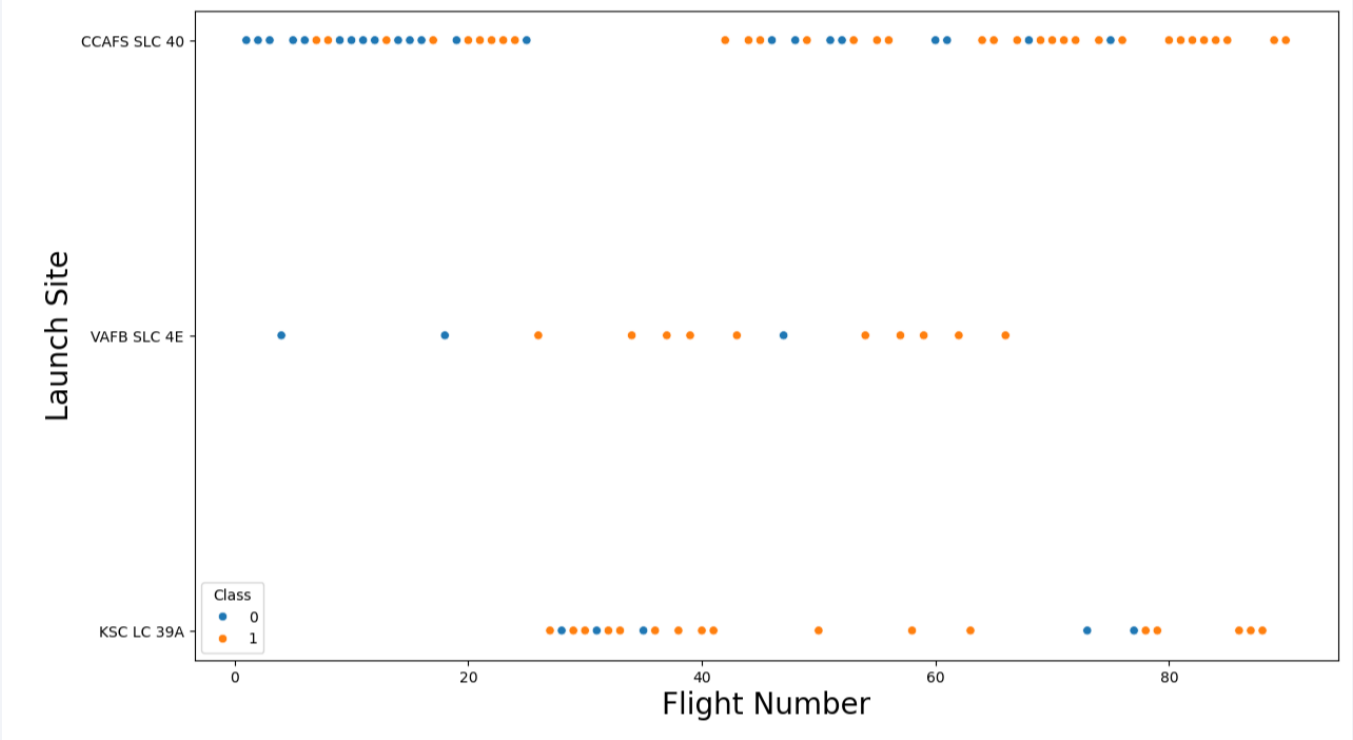
- **Exploratory Data Analysis (EDA):** Shows a rising trend in the success rate of SpaceX missions over time, reflecting ongoing technological improvements.
- **Interactive Analytics:** Highlight launch sites located near safety zones, coastal areas, and regions with robust logisical support.
- **Predictive Analysis:** Multiple models exhibit similar test accuracy, indicating that the dataset may be too limited to clearly identify the most effective model.

Section 2

Insights drawn from EDA

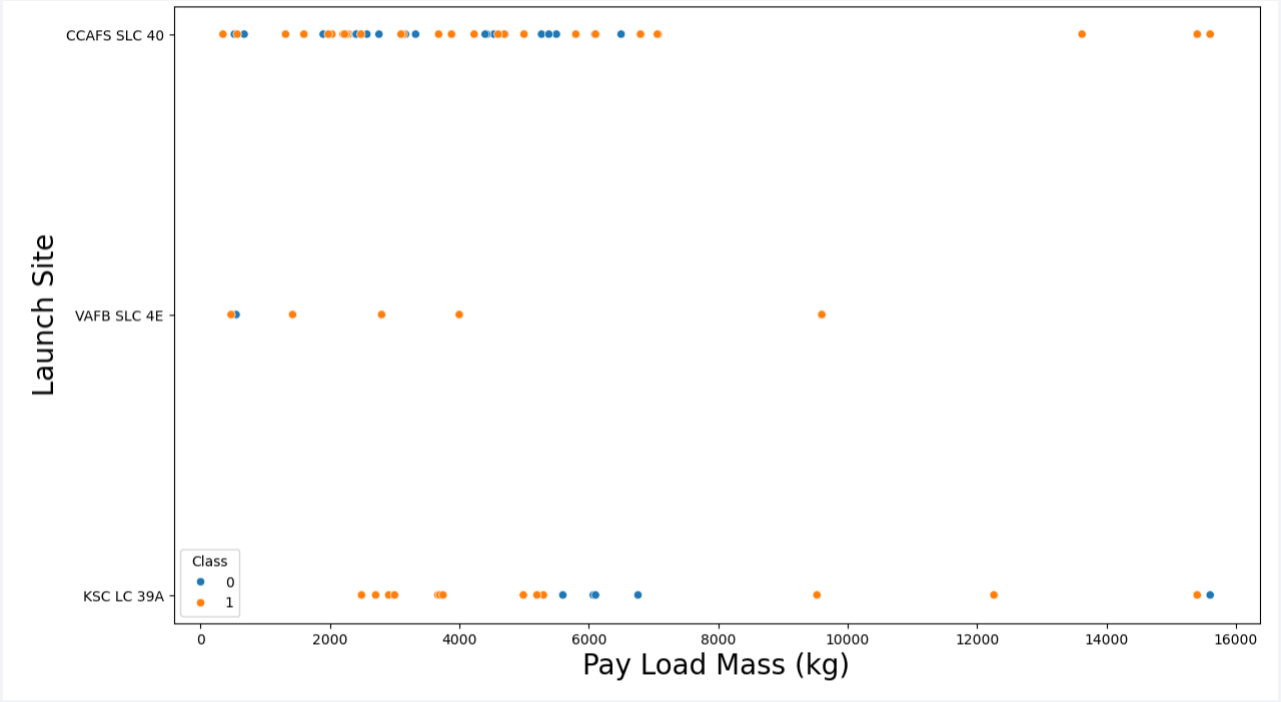
Flight Number vs. Launch Site

- **Launch Frequency:** Most flights were conducted from the CCAFS SLC-40 site.
- **Success Rates:** The VAFB SLC-4E and KSC LC-39A sites boast higher success rates compared to other locations.
- **Flight Trends:** Recent launches show greater success rates than earlier missions.



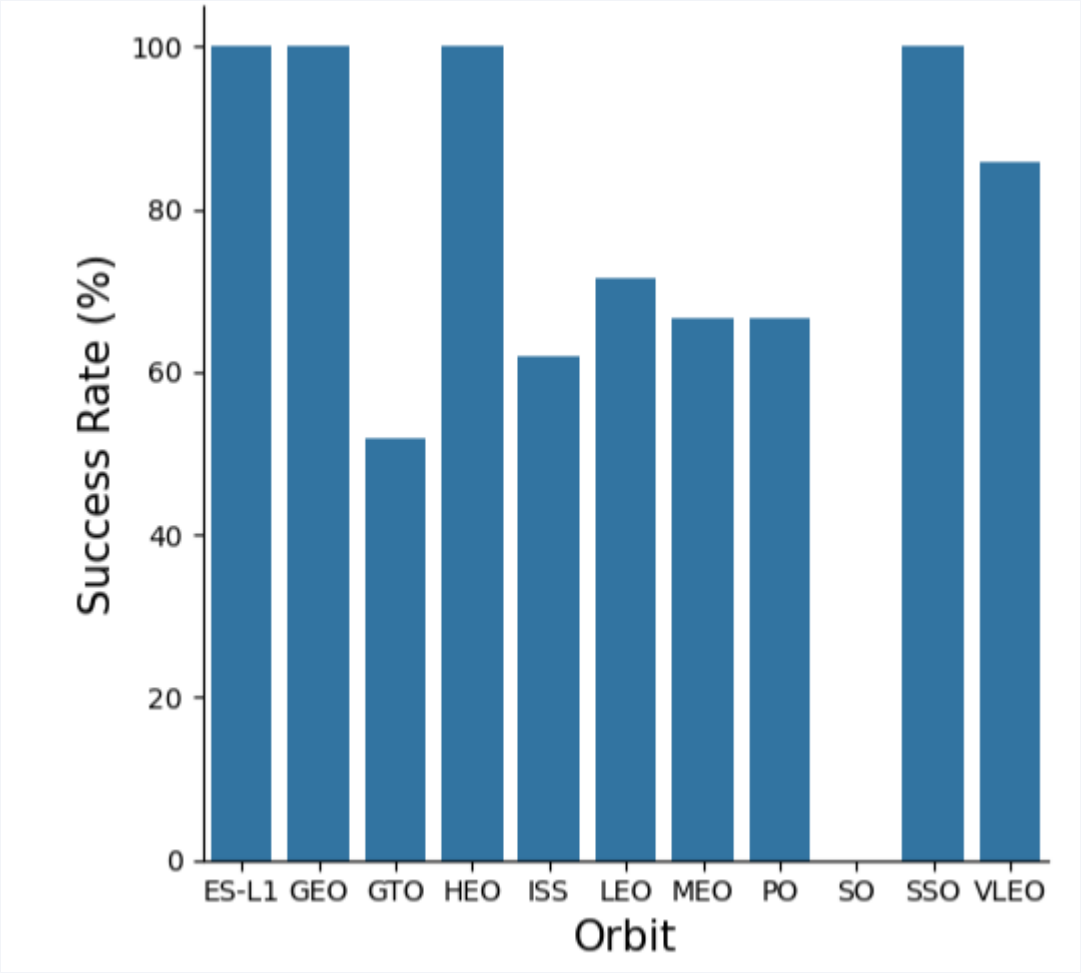
Payload vs. Launch Site

- **High Payload Success:** Most flights carrying payloads over 7000 kg were successful.
- **KSC LC-39A Performance:** Achieved a 100% success rate for payloads under 5500 kg.
- **Payload-Performance Relationship:** Across all launch sites, success rates increase in proportion to payload mass.



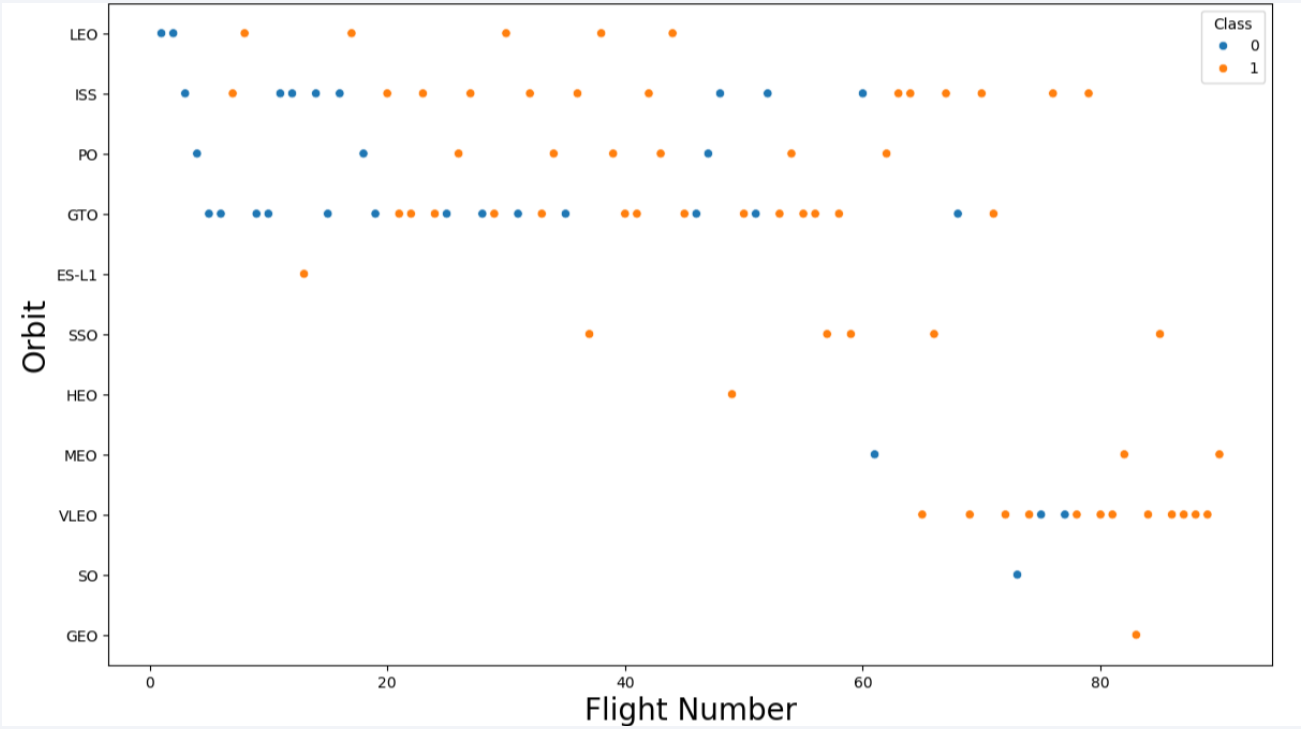
Success Rate vs. Orbit Type

- **3OS Orbit:** Has a 0% success rate.
- **ELS-1, GEO, HEO, and SSO Orbits:** Achieved a 100% success rate.
- **GTO, ISS, LEO, MEO, and PO Orbits:** Maintain success rates between 50% and 75%.



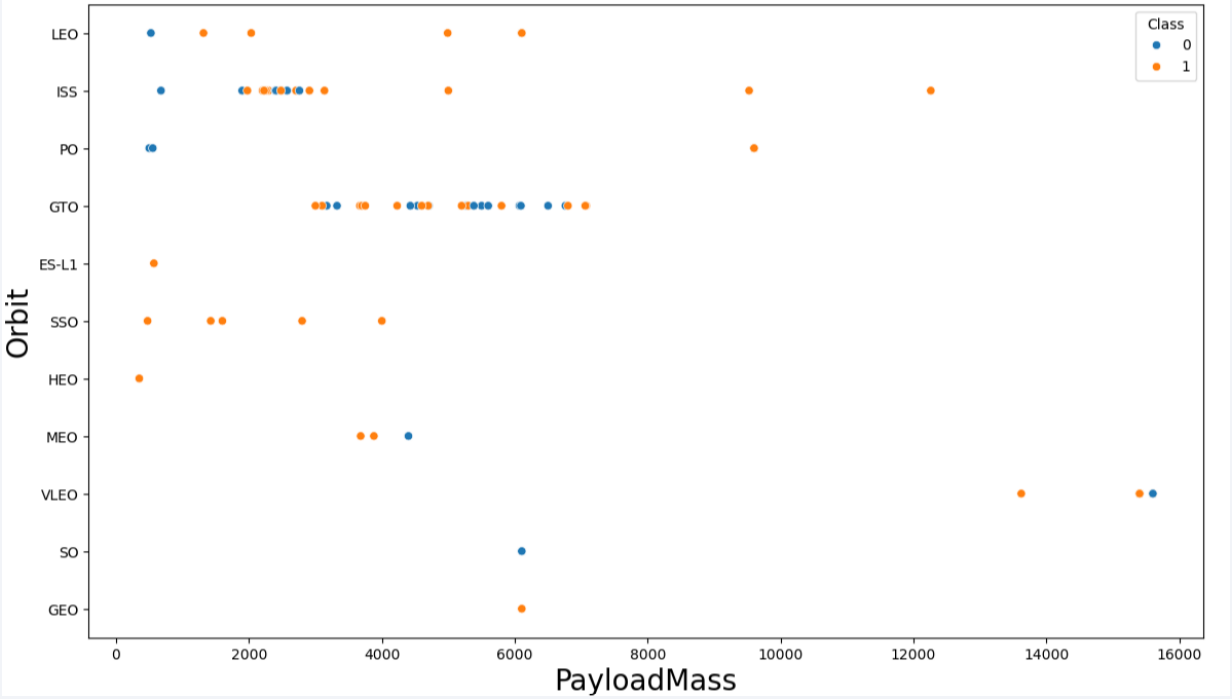
Flight Number vs. Orbit Type

- **Frequent Destinations:** Most flights were directed to the ISS and GTO orbits.
- **Orbit Type vs. Flight Number:** The data indicates no correlation between flight numbers and orbit types.



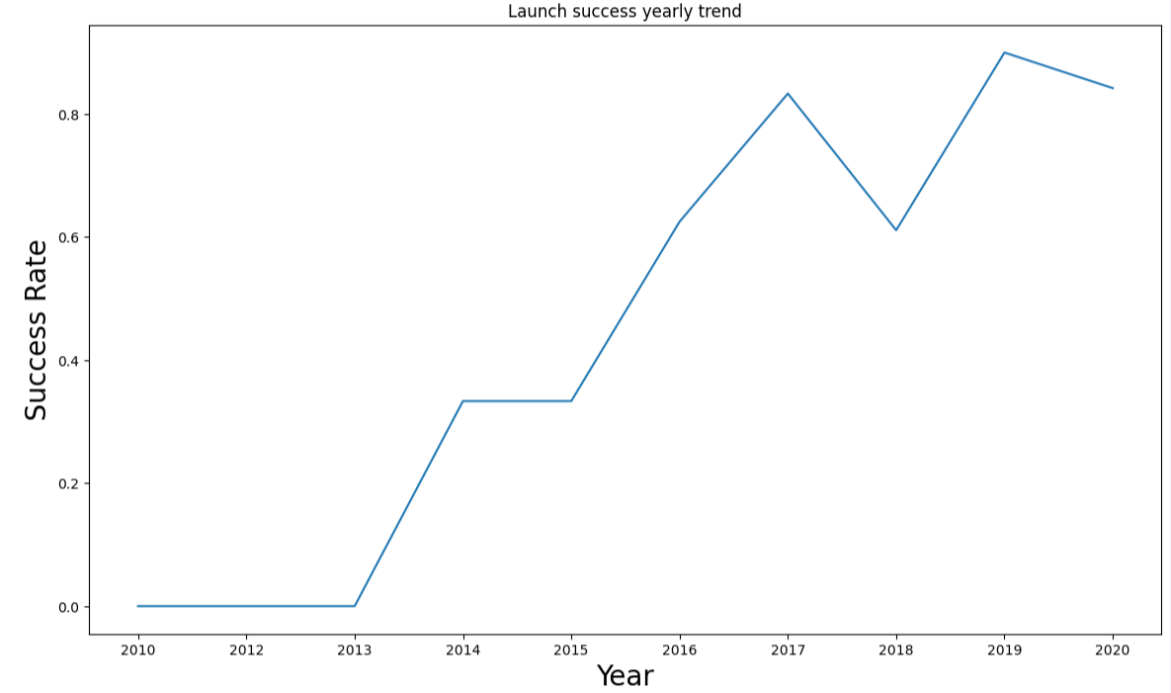
Payload vs. Orbit Type

- **High Payload Success:** Most flights carrying payloads over 7000 kg were successful.
- **KSC LC-39A Performance:** Achieved a 100% success rate for payloads below 5500 kg.
- **Payload-Performance Correlation:** Success rates at all launch sites increase with payload mass.



Launch Success Yearly Trend

- **Steady Improvement:** Launch success rates have grown consistently since 2013.
- **Linear Growth:** From 2013 to 2017, the success rate increased at a steady, linear pace.
- **2018 Decline:** A dip in the success rate occurred during 2018.



All Launch Site Names

The unique launch site names and the query used to retrieve them are provided below.

```
[10]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

[10]:

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

Launch Site Names Begin with 'CCA'

5 records for launch sites begin with the string 'CCA' and the query used for obtaining the information is shown below.

```
[11]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[11]:

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

Total Payload Mass

The total payload mass carried by boosters from NASA site =45596 Kg.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

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

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1=2928.4 Kg.

Task 4

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

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

[13]:

Average payload mass by Booster Version F9 v1.1
2928.4

First Successful Ground Landing Date

The first successful landing outcome on a ground pad was in 2015-12-22.

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

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

First_Succesful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Below is a list of boosters that successfully landed on a drone ship and carried payloads between 4000 and 6000 kg.

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
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

In total, there are 100 successful flights and 1 failed flight.

```
%sql SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM SPACEXTBL WHERE MISS
```

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

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

The following is a list of boosters that carried the highest payload mass.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

2015 Launch Records

The list below shows the failed "landing_outcomes" on drone ships in 2015, along with their booster versions and launch site names.

```
%%sql
SELECT
    CASE
        WHEN substr(Date, 6, 2) = '01' THEN 'January'
        WHEN substr(Date, 6, 2) = '02' THEN 'February'
        WHEN substr(Date, 6, 2) = '03' THEN 'March'
        WHEN substr(Date, 6, 2) = '04' THEN 'April'
        WHEN substr(Date, 6, 2) = '05' THEN 'May'
        WHEN substr(Date, 6, 2) = '06' THEN 'June'
        WHEN substr(Date, 6, 2) = '07' THEN 'July'
        WHEN substr(Date, 6, 2) = '08' THEN 'August'
        WHEN substr(Date, 6, 2) = '09' THEN 'September'
        WHEN substr(Date, 6, 2) = '10' THEN 'October'
        WHEN substr(Date, 6, 2) = '11' THEN 'November'
        WHEN substr(Date, 6, 2) = '12' THEN 'December'
    END AS Month_Name,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
    AND substr(Date, 0, 5) = '2015';
```

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

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The following is a ranking of landing outcome counts (e.g., Failure on drone ship or Success on ground pad) between June 4, 2010, and March 20, 2017, listed in descending order.

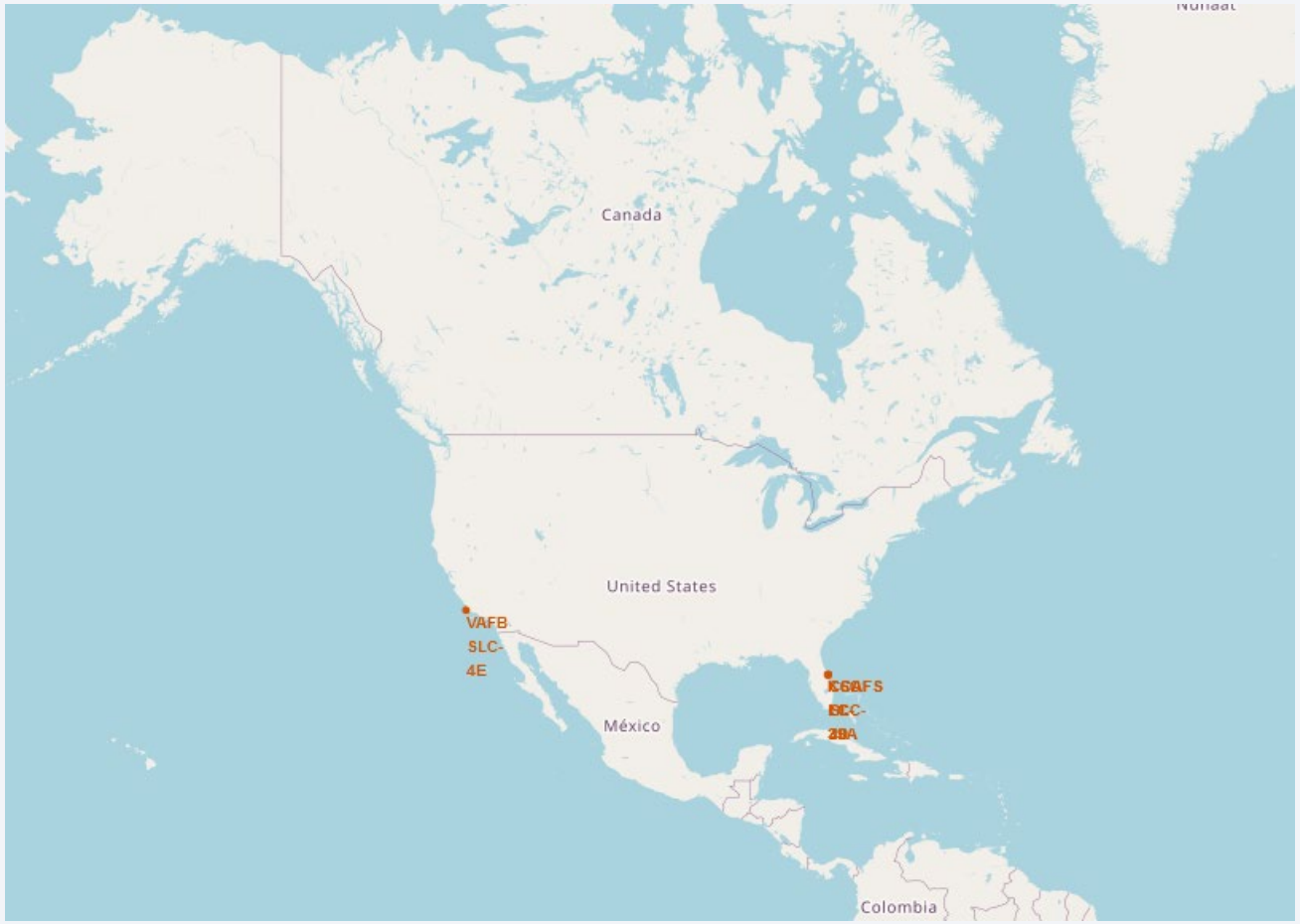
```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS "Count"
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC;
```

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

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


Section 3
Launch Sites
Proximities Analysis

All Launch Sites on a Global Map

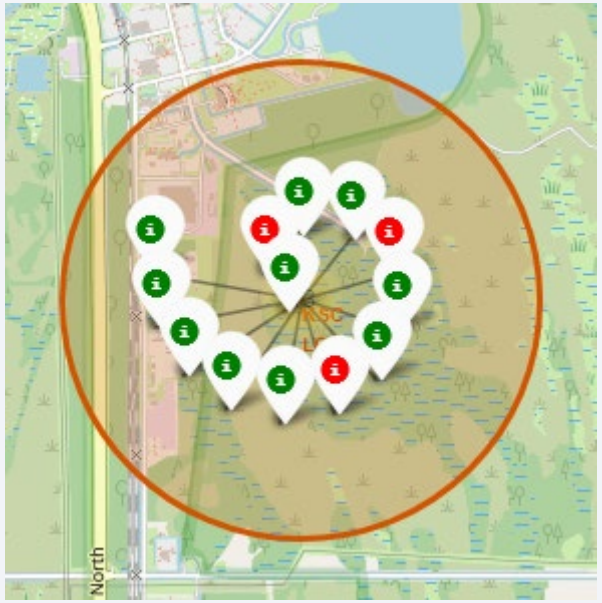


- **Launch Sites:** Vandenberg Air Force Base, Kennedy Space Center, and Cape Canaveral Air Force Station.
- **Coastal Locations:** Positioned near the east or west coasts to reduce risks to populated areas in case of launch failures.
- **Proximity to the Equator:** Their locations provide strategic benefits for accessing specific types of orbits.

Mark the Success/Failed Launches for each Launch Site

Markers display the landing outcomes of SpaceX rocket first stages:

- **Green Markers:** Indicate successful landings.
- **Red Markers:** Indicate failed landings.



Distances between a Launch Site to key Proximities

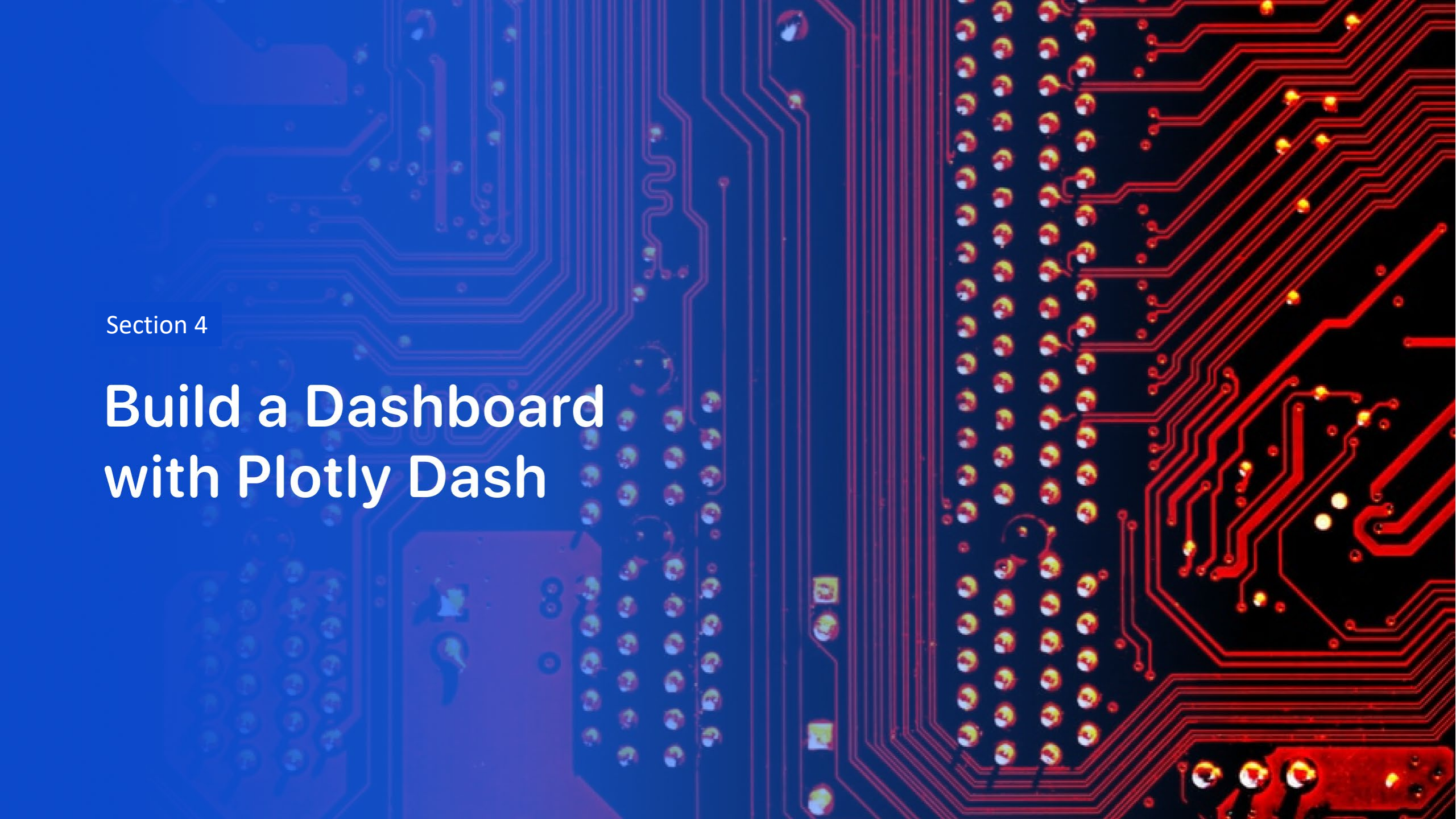
- Distances from the selected rocket launch site to key locations are represented by colored lines:

  Red Line: City center – 23.16 km

  Yellow Line: Coastline – 21.36 km

  Green Line: Highway – 26.83 km

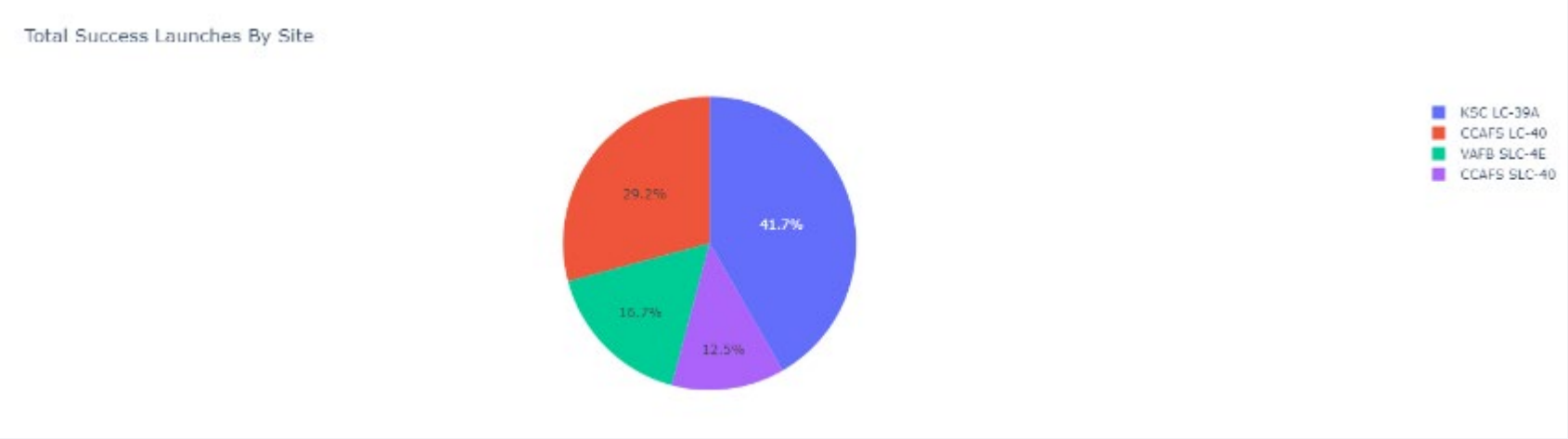
- The site provides strong logistical support and poses a low risk to populated areas.


Section 4
Build a Dashboard with Plotly Dash

Distribution of Successful Space Launches by Site

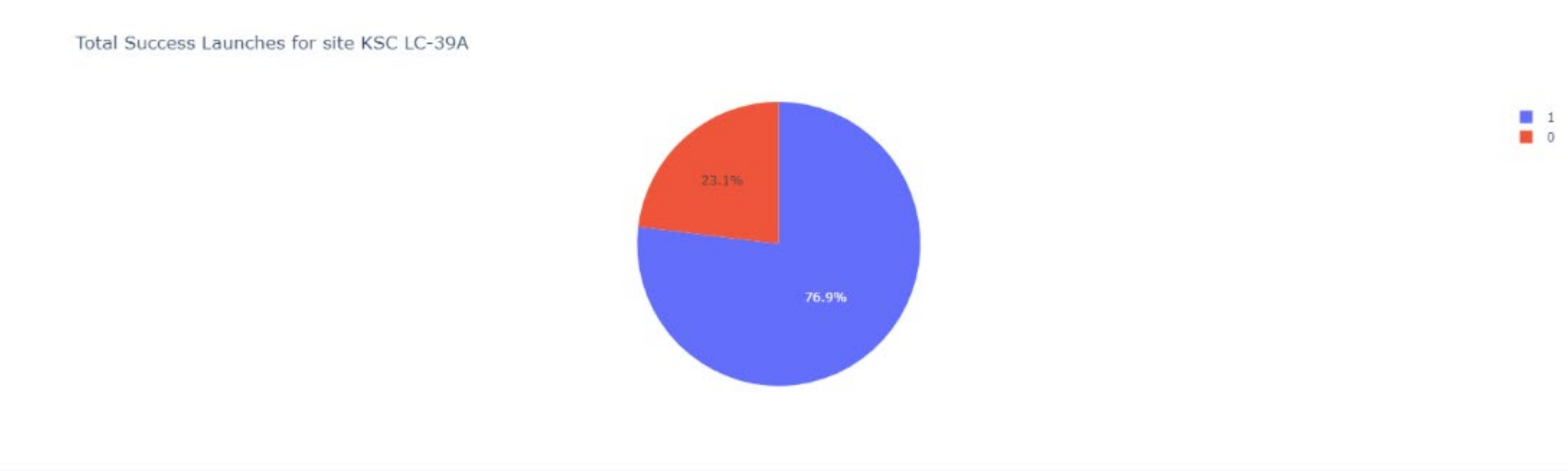
The highest success launch rates were recorded at these sites :

- KSC LC-39A (41.7%)
- CCAFS LC-40 (29.2%)



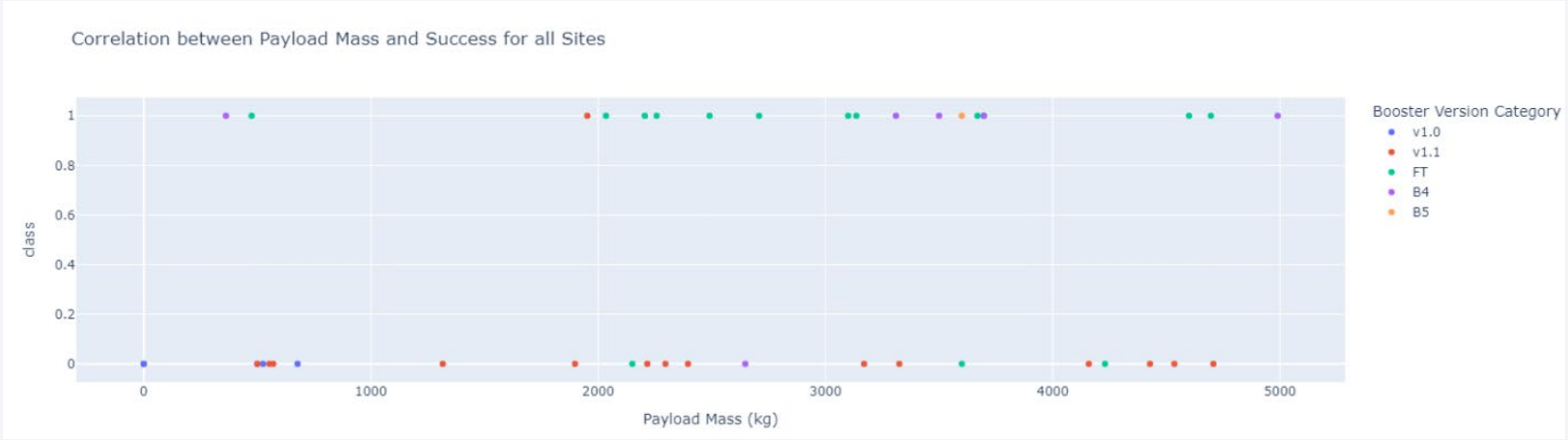
Success Rate of Space Launches at KSC LC-39A

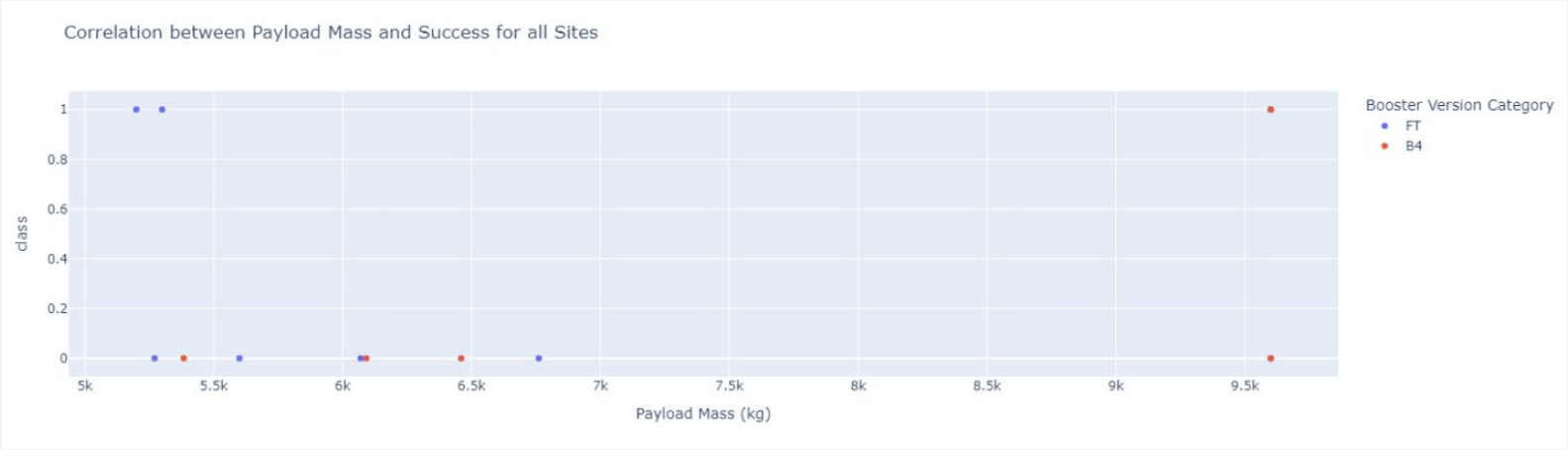
- Site KSC LC-39 success rate is 76.9%



Payload Mass and Launch Success across Booster Versions

- The highest success rate for payloads occurs within the range of 2000 to 5500 kg.





Section 5

Predictive Analysis (Classification)

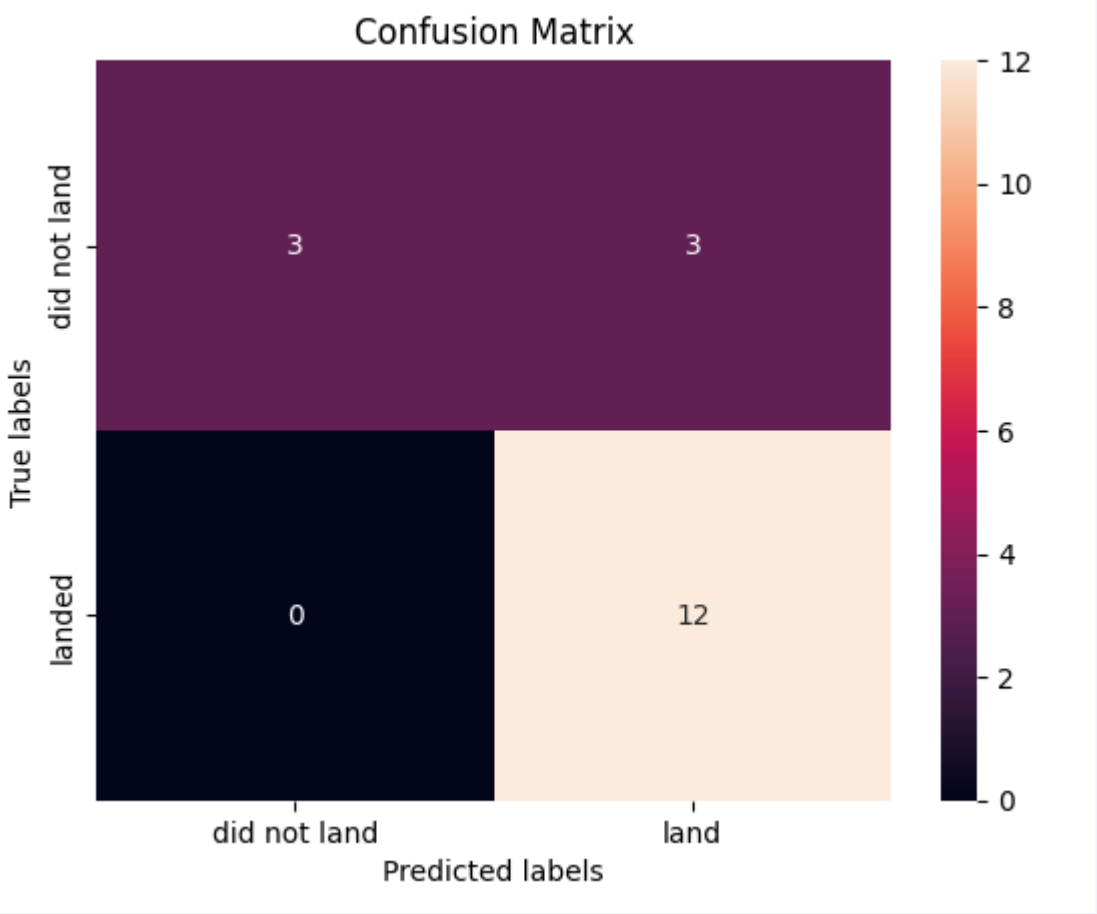
Classification Accuracy

- The accuracy results for the four models were identical when evaluated using the test set.
- Among them, the Tree Model delivered the best accuracy across the entire dataset.

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.887500
KNN	0.848214

Confusion Matrix

- **Confusion Matrix Consistency:** The confusion matrix is the same for the top-performing models: Logistic Regression, SVM, and KNN.
- **Performance Metrics:** These models demonstrate excellent performance in correctly identifying the ‘landed’ class (12 out of 12) and moderate accuracy in predicting the ‘did not land’ class (3 out of 6).
- **Misclassification Imbalance:** There is a notable imbalance in misclassification, with all three misclassified instances being false positives in the ‘did not land’ class, resulting in zero false negatives.



Conclusions

- Rising Success Rate: The success rate for rocket launches improved after 2013.
- Perfect Success Rates: The GEO, HEO, ES-L1, and SSO orbits achieved a 100% launch success rate.
- Top Launch Site: KSC LC-39A recorded the highest success rate.
- Best ML Algorithm: The Decision Tree model emerged as the most effective machine learning algorithm for analyzing the SpaceX dataset, yielding the highest accuracy results.

Appendix

https://github.com/SereneB2099/ibm-capstone

Thank you!