IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason Street
2/6/2025

Outline

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

Executive Summary

- Summary of Methodologies:
  - Data was collected using SpaceX API calls and web scraping.
  - Data was cleaned and processed through wrangling techniques.
  - Exploratory Data Analysis (EDA) was performed using visualization and SQL.
  - Interactive visual analytics was conducted using Folium and Plotly Dash.
  - Machine learning models (SVM, Logistic Regression, Decision Trees, KNN) were developed, tuned, and evaluated.
- Summary of All Results:
  - The analysis identified key factors influencing successful rocket landings.
  - Interactive maps highlighted geographical launch patterns.
  - Predictive models provided insights into successful and failed launches.Summary of methodologies

Introduction

- Project Background and Context:
  - The SpaceX Falcon 9 project aims to make space travel more cost-effective with reusable rockets.
  - Understanding launch success factors can help optimize future launches.
- Problems You Want to Find Answers:
  - What factors contribute to a successful rocket landing?
  - Does payload mass or orbit type influence success rates?
  - What launch site has the highest success rate?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API was used to retrieve launch details.
  - Web scraping was conducted to obtain additional launch records.
  - Data wrangling was performed to clean and preprocess data.
  - EDA and SQL were used for data insights.
  - Interactive dashboards were created using Folium and Plotly Dash.
  - Machine learning models were trained for predictive analysis.

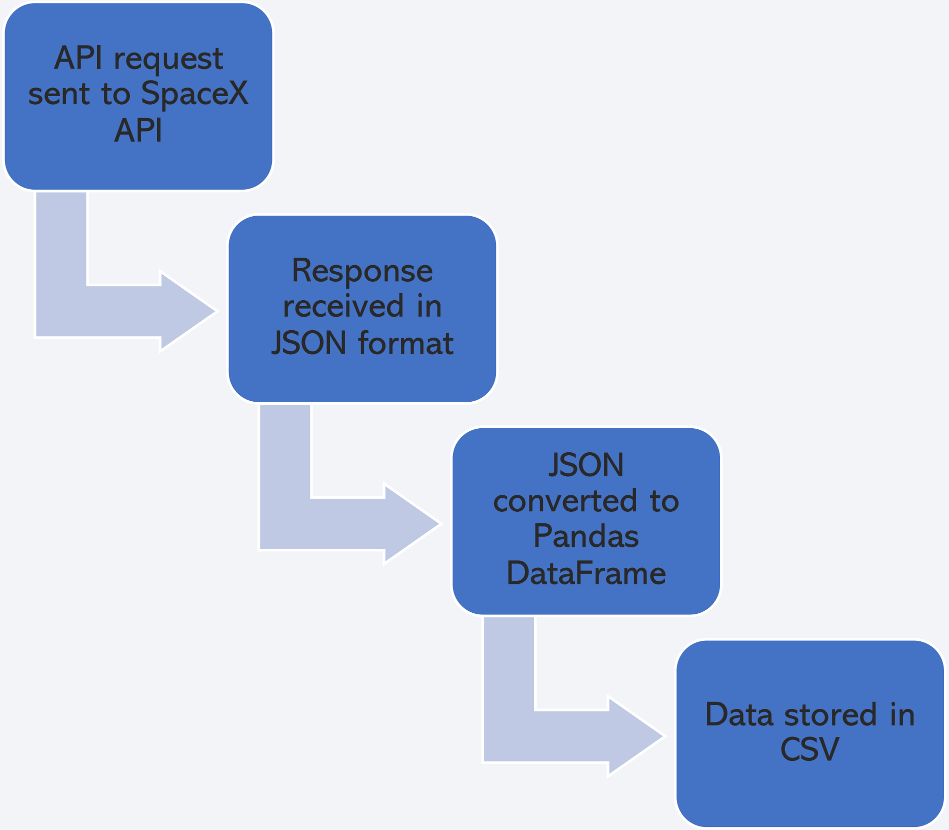
Data Collection

- Data Sources:
  - SpaceX REST API
  - Wikipedia (web scraping)
  - Preprocessed datasets for predictive analysis
- Data Collection Process:
  - API requests retrieved launch history data.
  - Web scraping extracted launch records.
  - Data was stored in CSV format for further analysis.

Data Collection – SpaceX API

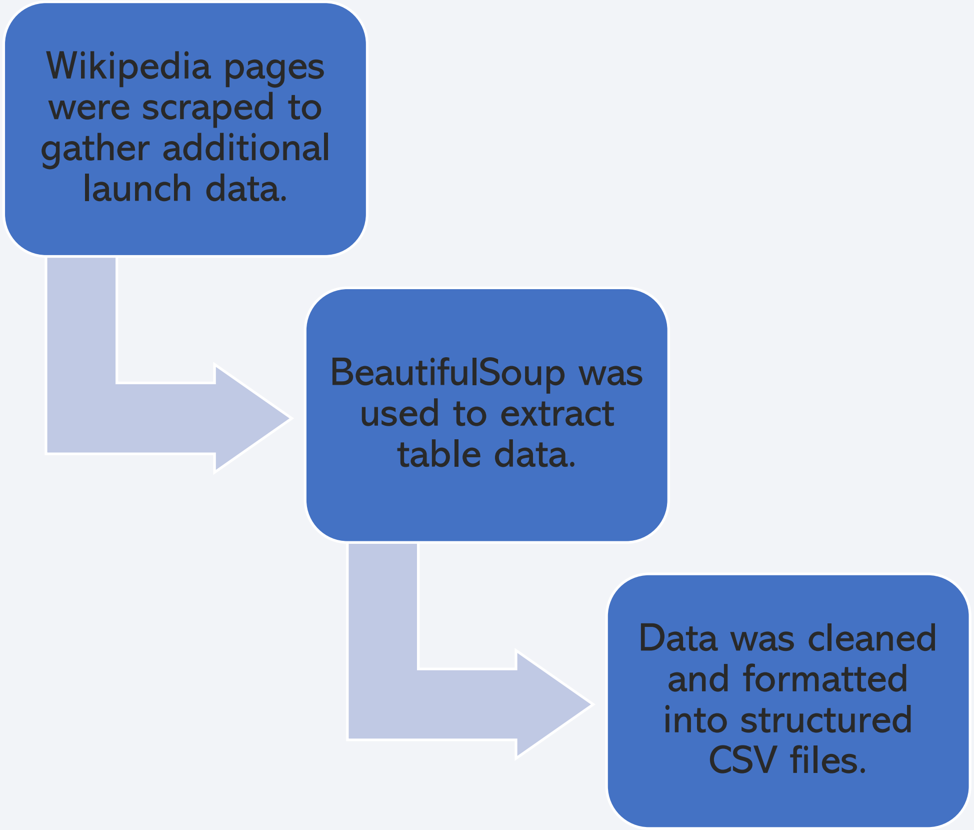
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- GitHub URL for Notebook: https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- Processing Steps:
  - Removing irrelevant columns.
  - Handling missing values.
  - Standardizing data formats.
  - Transforming categorical variables into numerical values.
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts and Visualizations:
  - Scatter plots for Flight Number vs. Launch Site.
  - Bar charts for launch success rates by orbit type.
  - Line charts for yearly success trends.
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/edadataviz.ipynb

EDA with SQL

- SQL Queries Performed:
  - Identified unique launch sites.
  - Filtered data based on specific conditions (e.g., NASA launches).
  - Aggregated payload masses for different booster versions.
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- Map Objects Added:
  - Launch Site Markers (folium.Marker) – Identify SpaceX launch sites.
  - Circular Zones (folium.Circle) – Show area influence (1 km radius).
  - Success/Failure Markers (folium.MarkerCluster) – Green for success, red for failure.
  - Mouse Position Tool (MousePosition) – Displays real-time coordinates.
  - Distance Measurement Lines (folium.PolyLine) – Connect sites to coastlines, highways, cities.
- Purpose:
  - Visualize launch site distribution.
  - Analyze proximity to infrastructure.
  - Identify geographic patterns in success rates.
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dashboard Components Added:
  - Dropdown Menu (dcc.Dropdown) – Select launch sites.
  - Success Rate Pie Chart (dcc.Graph) – Displays launch success distribution.
  - Payload Slider (dcc.RangeSlider) – Filters payload mass range.
  - Scatter Plot (dcc.Graph) – Shows payload vs. launch success with booster version color-coded.
- Purpose:
  - Allow interactive exploration of launch success rates.
  - Identify key factors influencing successful landings.
  - Analyze payload impact on mission success.
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data Preparation
  - Standardized features using StandardScaler
  - Split data into training (80%) and testing (20%) sets
- Model Training & Hyperparameter Tuning
  - Used GridSearchCV to find the best parameters for:
    - Logistic Regression (C, penalty, solver)
    - Support Vector Machine (SVM) (kernel, C, gamma)
    - Decision Tree (criterion, max_depth, min_samples_split)
    - K-Nearest Neighbors (KNN) (n_neighbors, algorithm, p)
- Model Evaluation
  - Compared models based on accuracy scores on test data
  - Used confusion matrices to analyze misclassifications
- Best Performing Model
  - The model with highest test accuracy was selected
  - If multiple models had the same accuracy, interpretability & computation efficiency were considered
- GitHub URL for Notebook:

https://github.com/JasonTaylorStreet/PublicCoursera/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

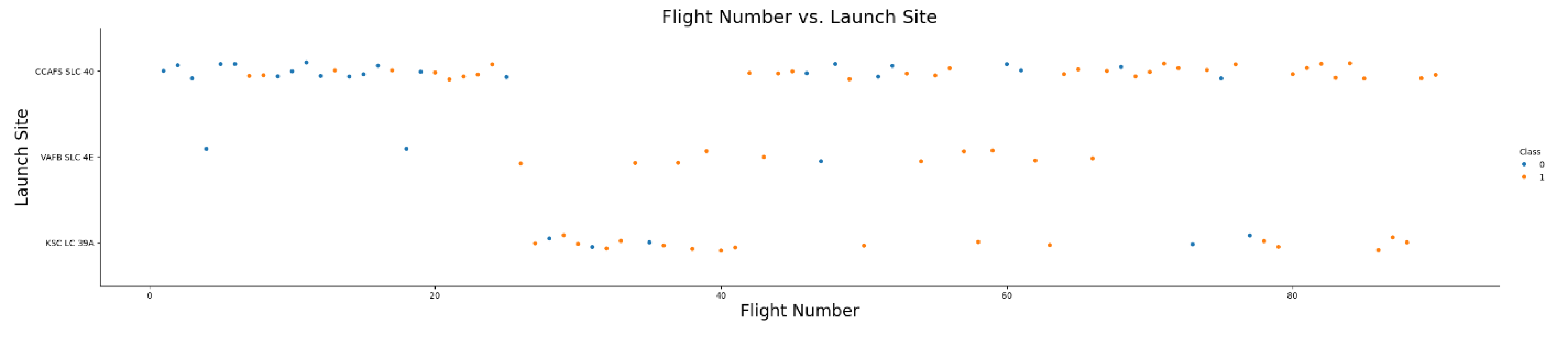
Results

- Exploratory Data Analysis Findings:
  - Certain launch sites had higher success rates.
  - Orbit type influenced landing success.
- Machine Learning Findings:
  - SVM and Logistic Regression models performed best.
  - Payload mass was a key determinant of success.

Section 2

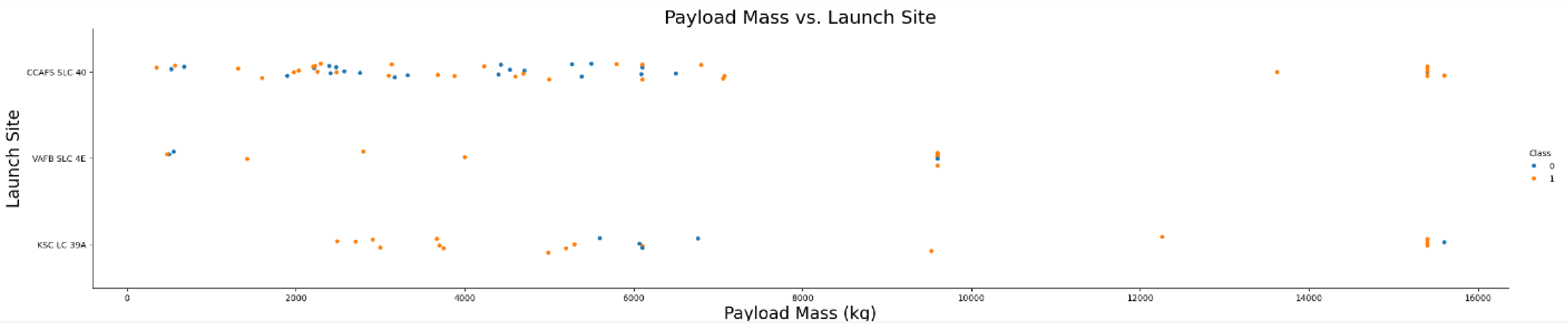
Insights drawn from EDA

Flight Number vs. Launch Site



- Most flights took place at site CCAFS SLC 40. There appears to be a range of flight numbers that did not occur at this site but rather KSC LC 39A. Further investigation required into probable causes such possible site maintenance

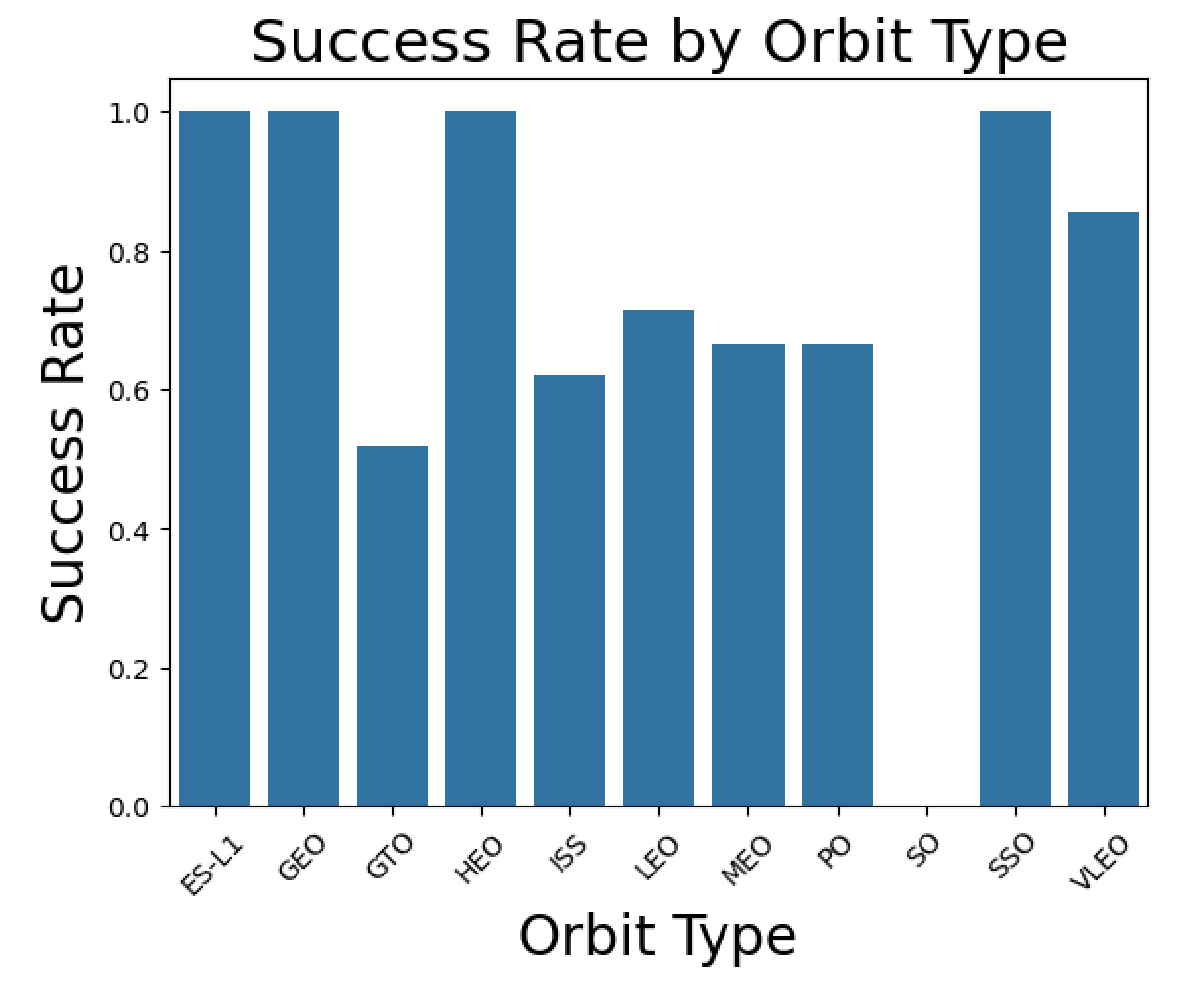
Payload vs. Launch Site



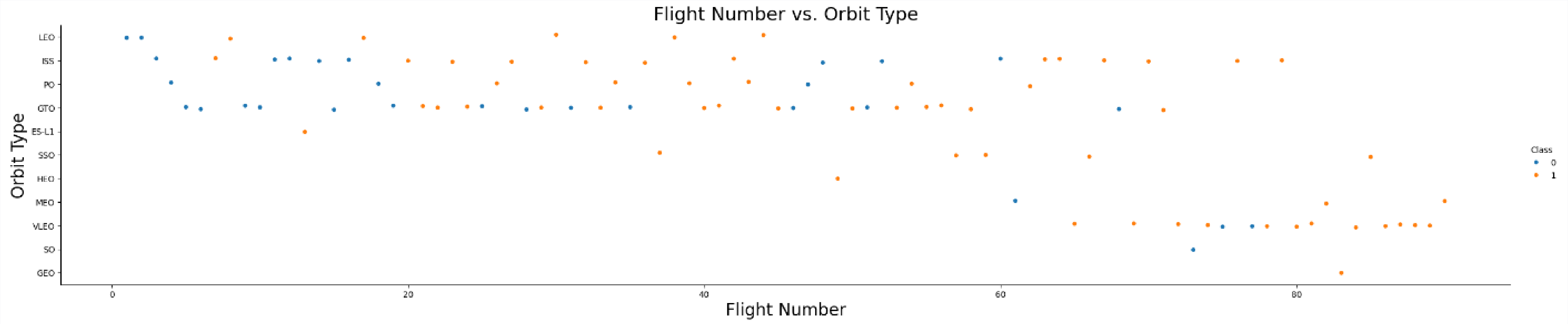
- The CCAFS SLC 40 site had the most variety in payload mass which trends with the number of flights from the previous slide.

Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, SSO, and VLEO orbit types had the greatest success
- Next slide will provide perspective at the displayed rates. For example, GEO had only 1 flight number assigned



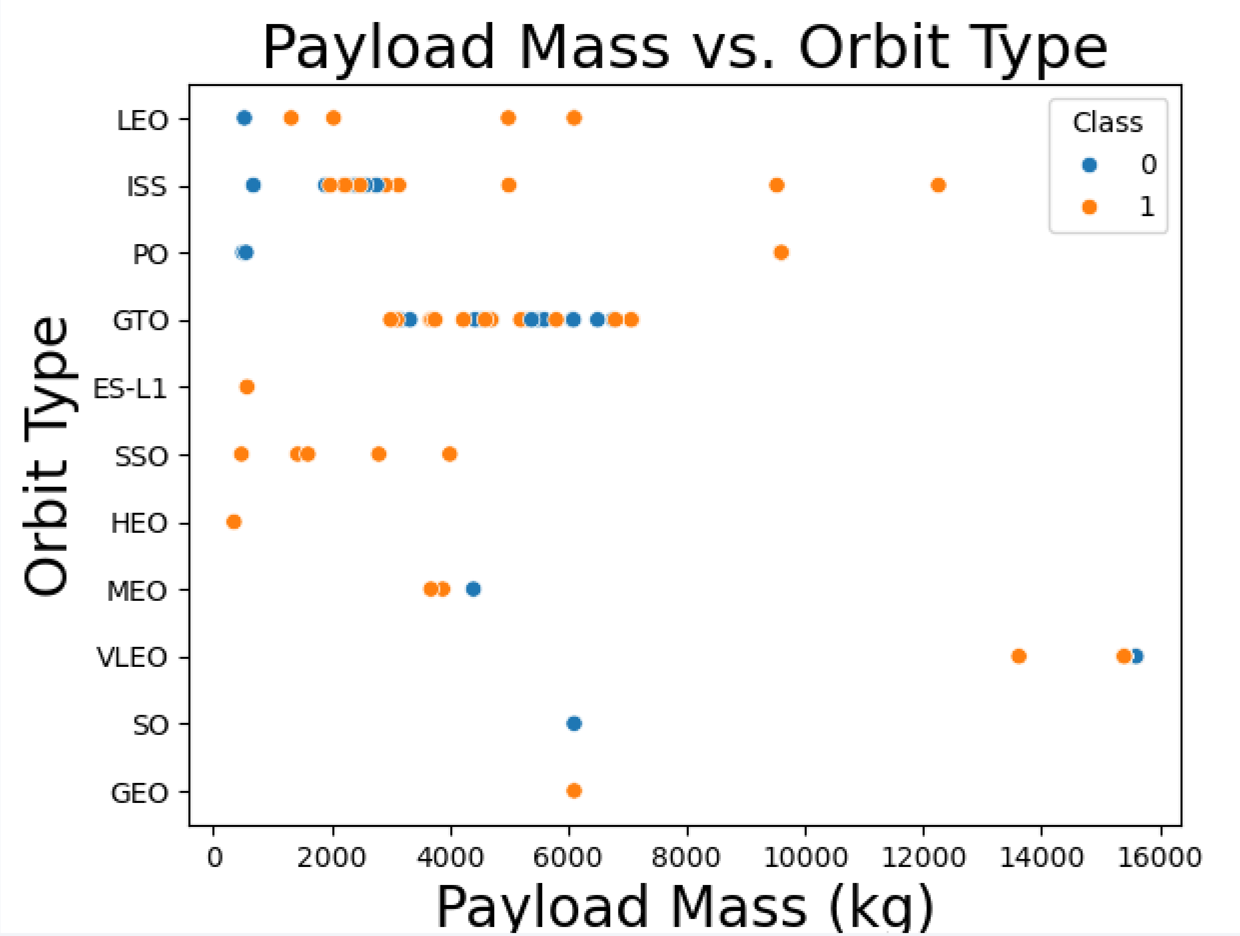
Flight Number vs. Orbit Type



- The orbit types ES-L1, SSO, HEO, MEO, SO, and GEO had relatively fewer flight numbers which in which case a single failure will have a greater impact on the success rate of the previous slide.

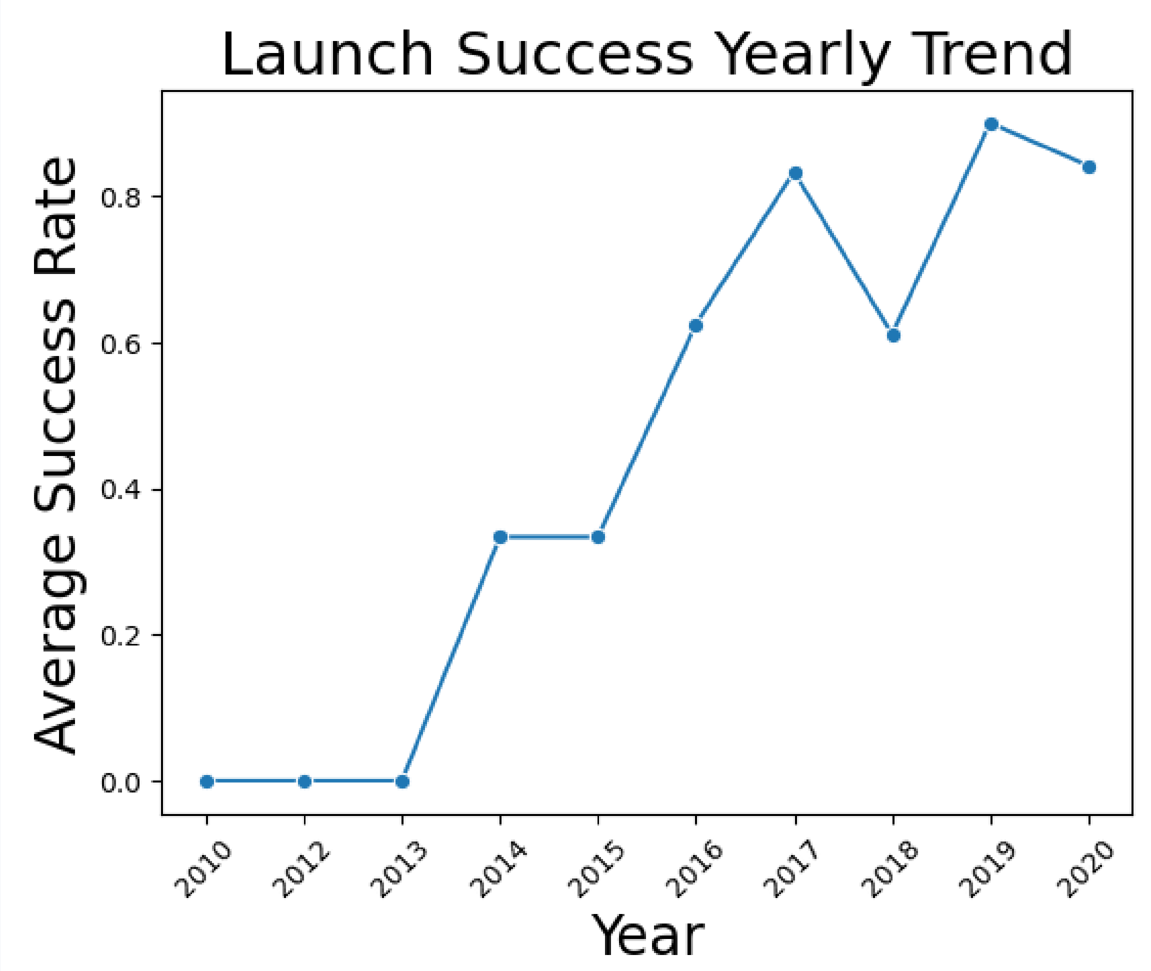
Payload vs. Orbit Type

- Specific orbit types have specific ranges of payload mass with most below 8000kg (except 2 ISS, 1 PO, and 3 VLEO)



Launch Success Yearly Trend

- There is a positive correlation for success as time progresses. Further investigation is required as to lowered success in 2018.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

Utilized DISTINCT to return the 4 unique launch sites.

Launch Site Names Begin with 'CCA'

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

```
%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_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

- Utilized % as a wild card to ensure both CCAFS LC-40 and CCAFS SLC-40 were returned

Total Payload Mass

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" Like '%NASA (CRS)%';
```

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

Total_Payload_Mass
48213

- The above is total payload mass specifically for customer NASA (CRS)

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

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

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

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 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

List the total number of successful and failure mission outcomes

```
%sql SELECT "Landing_Outcome", COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY "Landing_Outcome";
```

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

Landing_Outcome	Total_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

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

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

```
 * 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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date, 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, 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

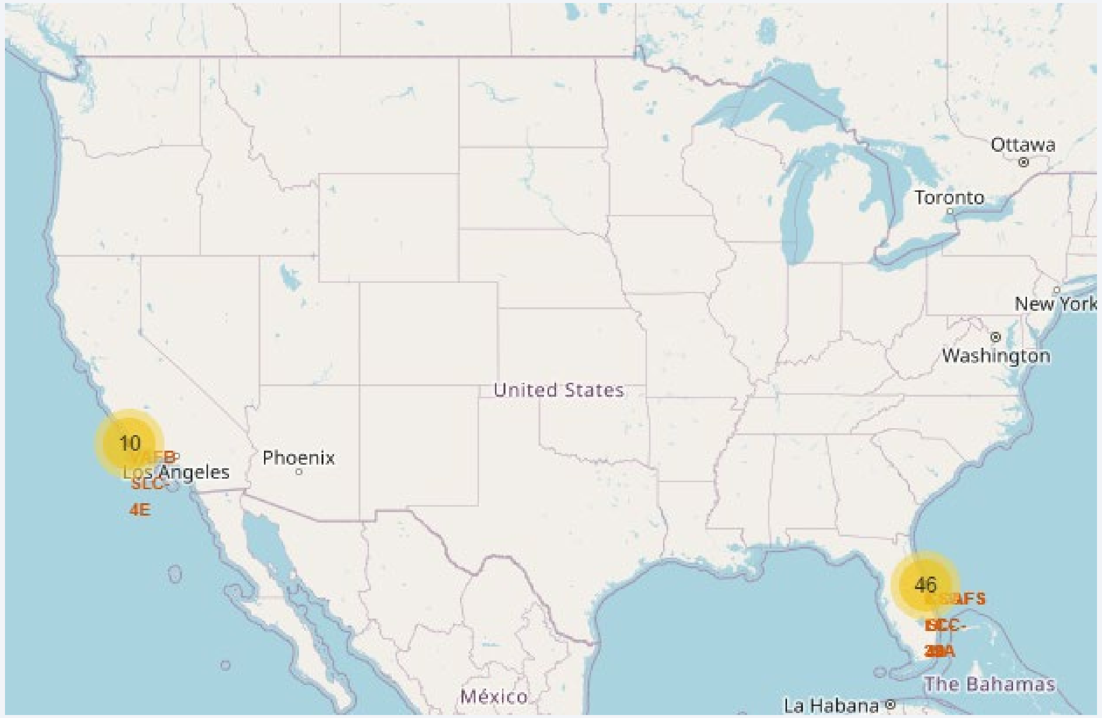
Section 3

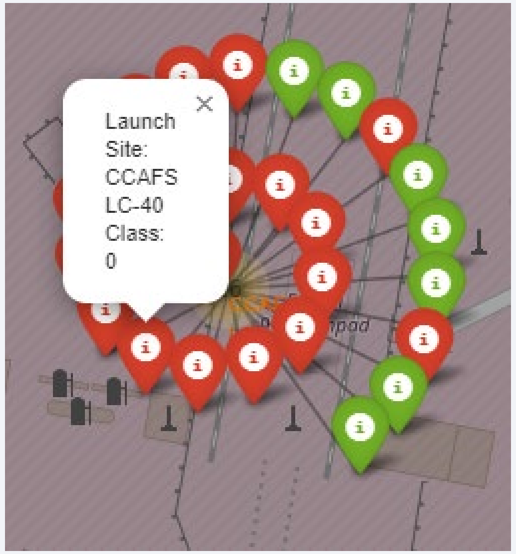
Launch Sites Proximities Analysis

All Launch Sites – Map Marker

Launches Occurred off the coast of California, USA and Florida, USA

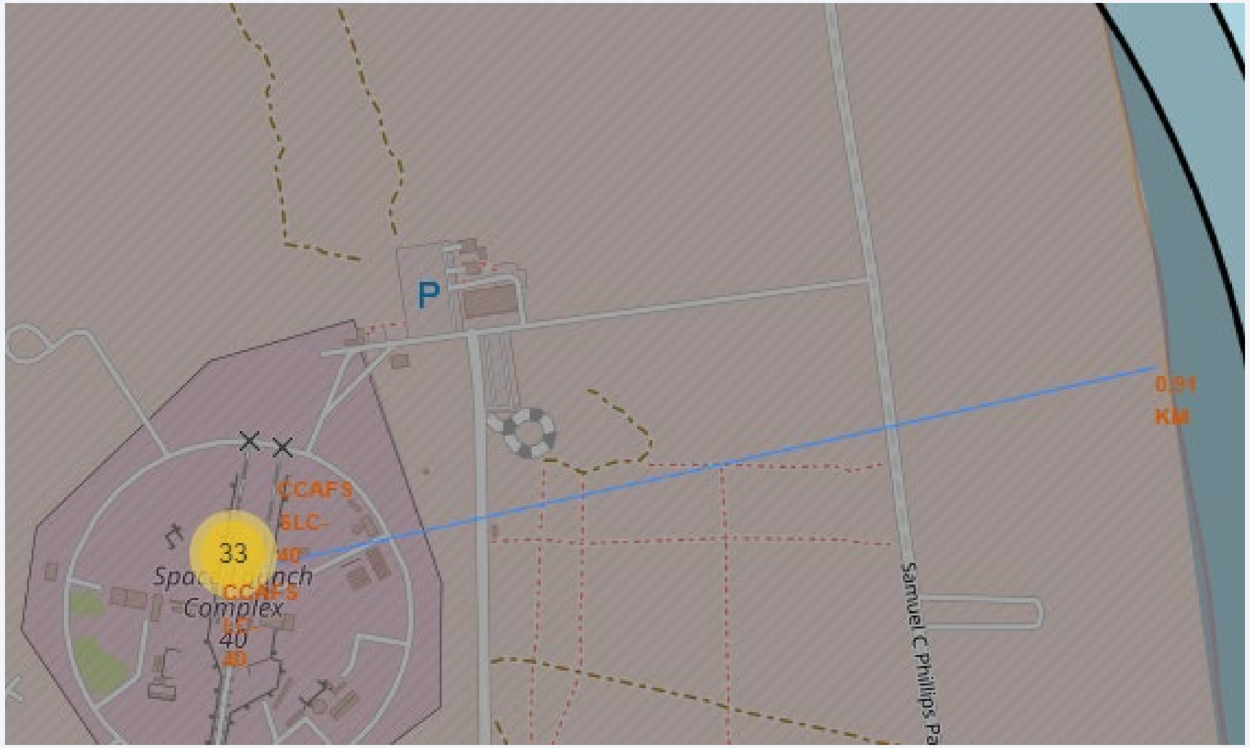
Launch Success & Failure-Detailed Cluster Map





- Cluster with numbers indicate the number of launches from that location
- Green specific launch marker indicates success, Red failure

Distances Between Site and Landmarks

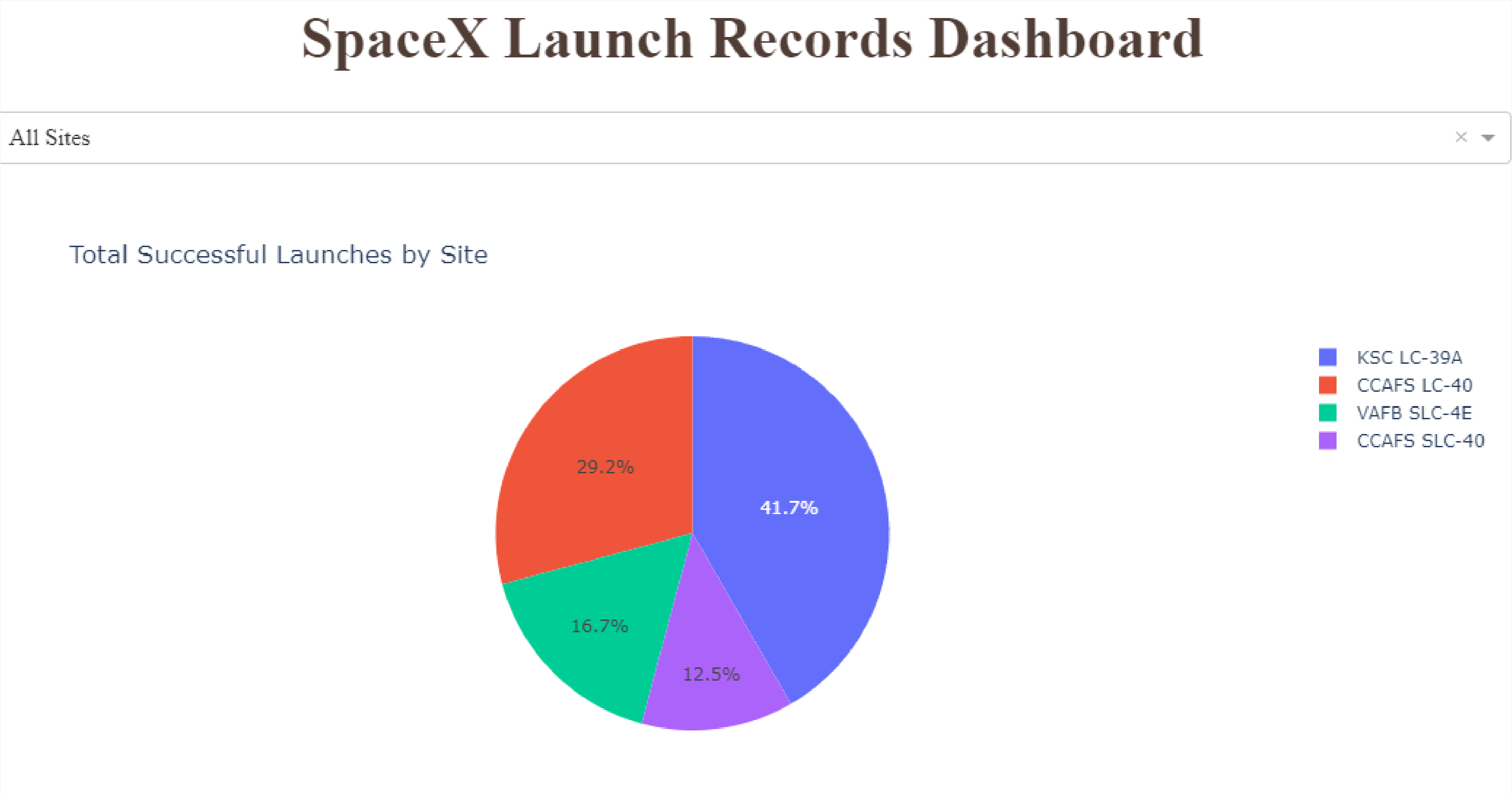


- Map is capable for determining the distance between the selected site and desired landmark. Displayed is example of distance to coastline
- In general, launch sites are located near coast lines and distance to other major landmarks is relative

Section 4

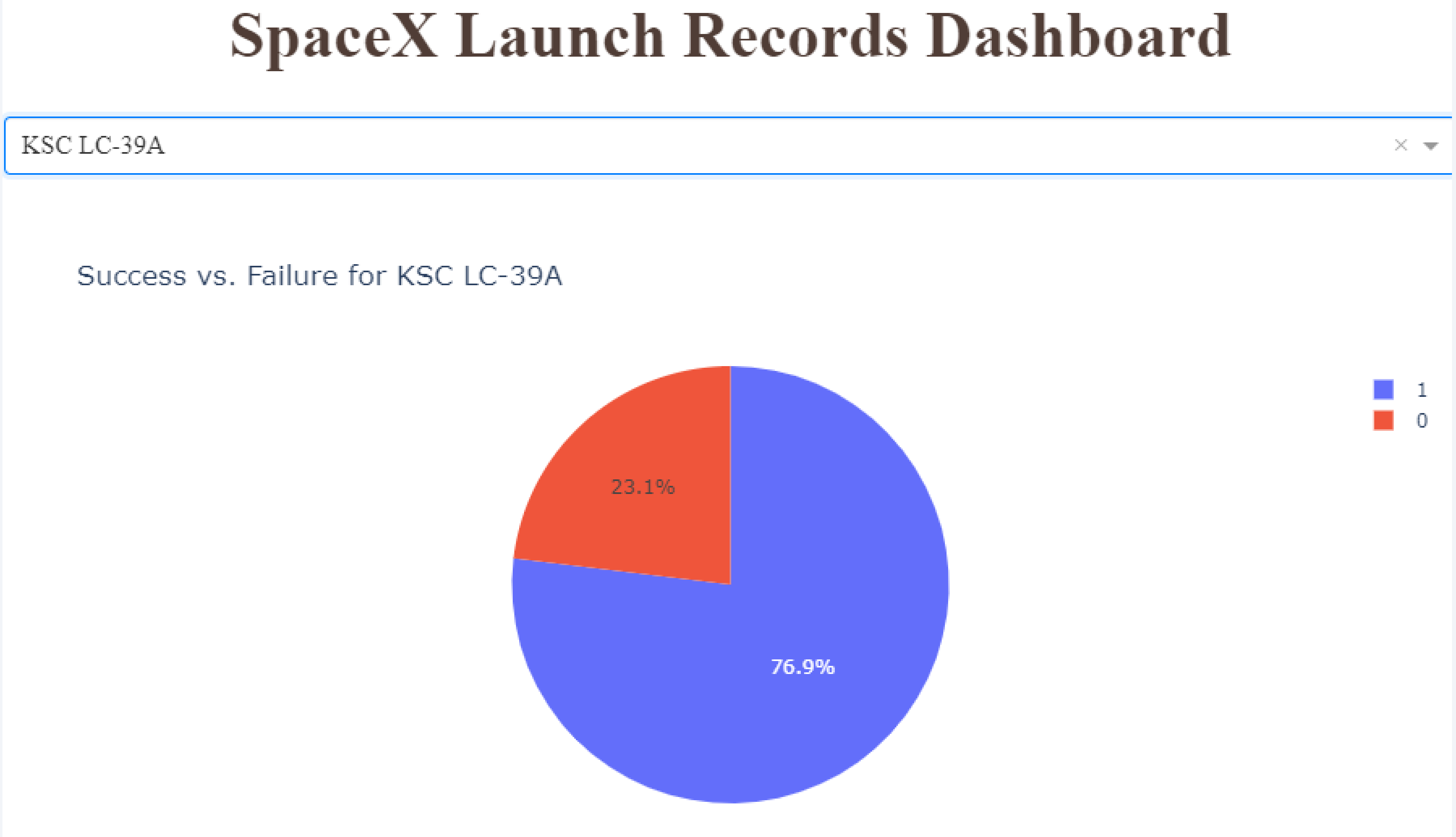
Build a Dashboard with Plotly Dash

Pie chart Success Rate by Launch Site



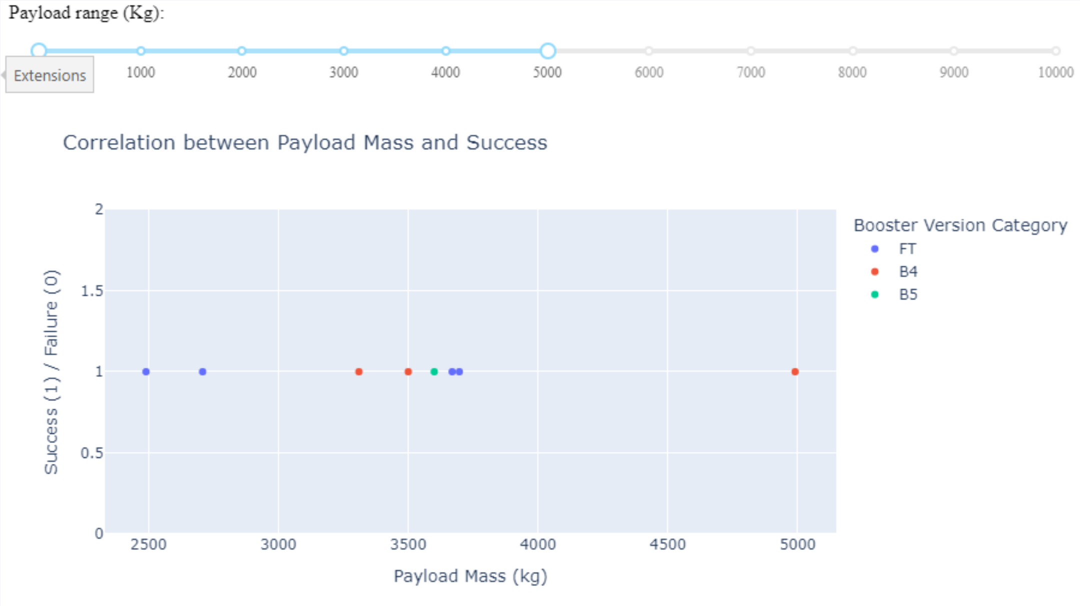
- KSC LC-39A had the most successful launches

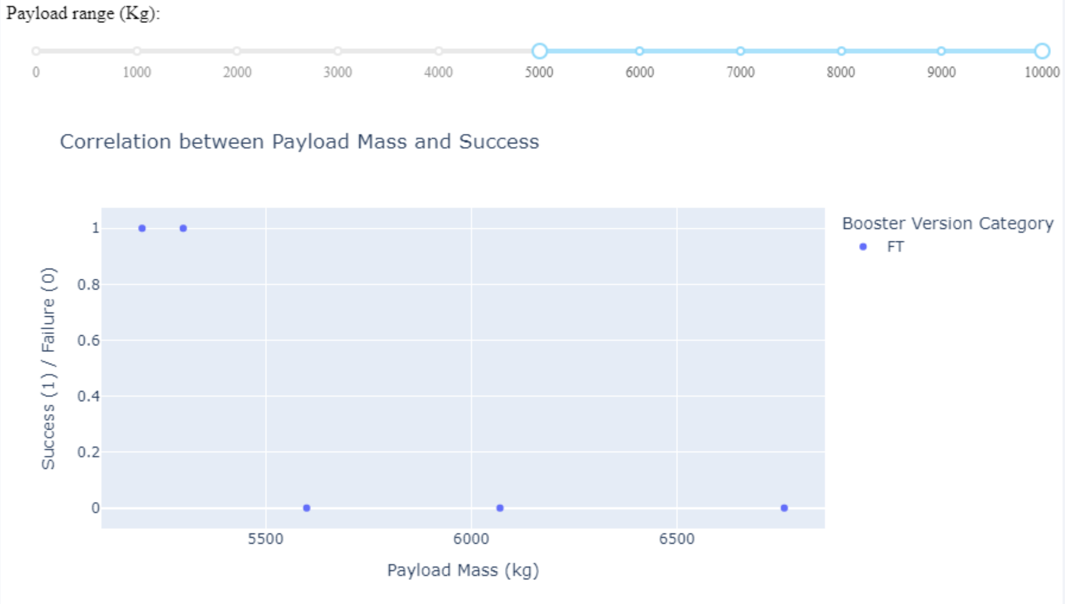
Pie chart Success Rate of the Most Successful Site



Out of the 41.7% Success Rate of all launches, the KSC LC-39A site had a success rate of 76.9% of all of this specific site's launches

Payload Mass vs Success – scatter plot





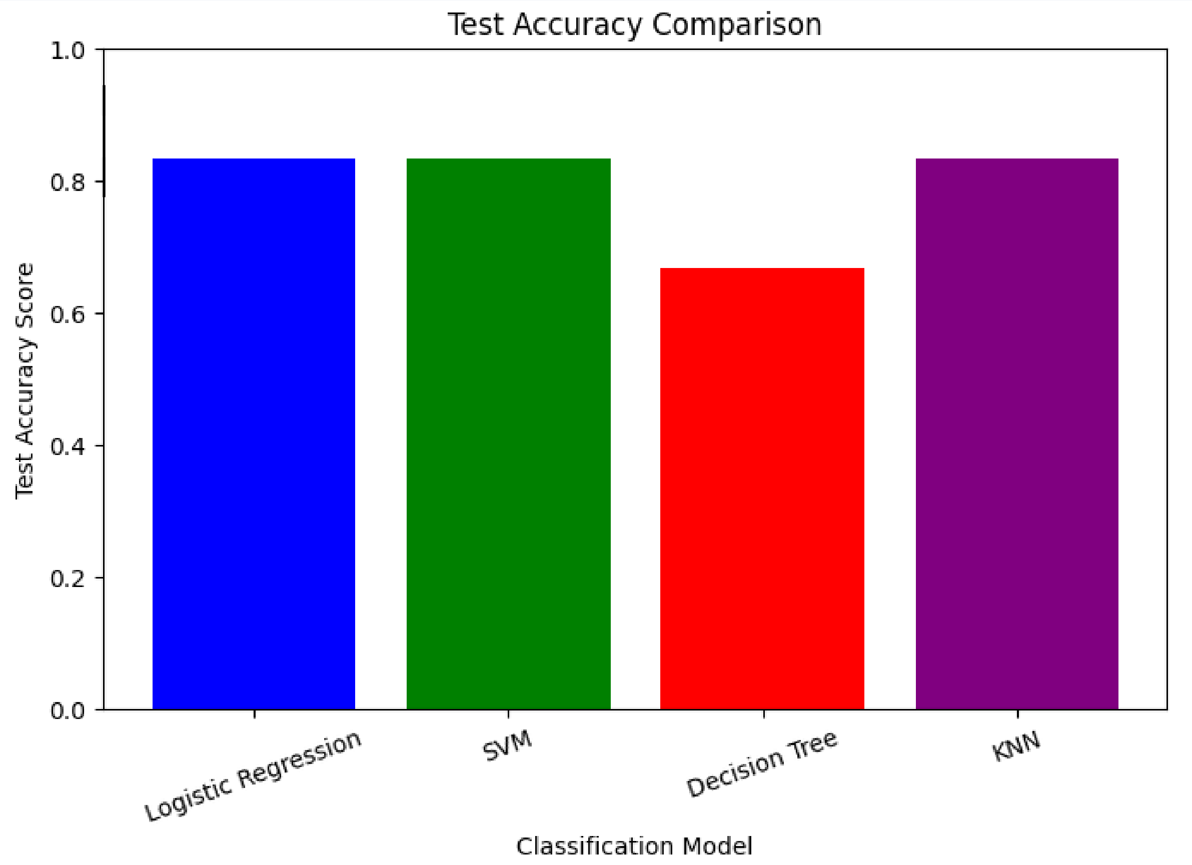
- Left is Payload Mass 0-5000kg and Right is 5000-10000kg
- There were no failures until payload mass was >5500kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

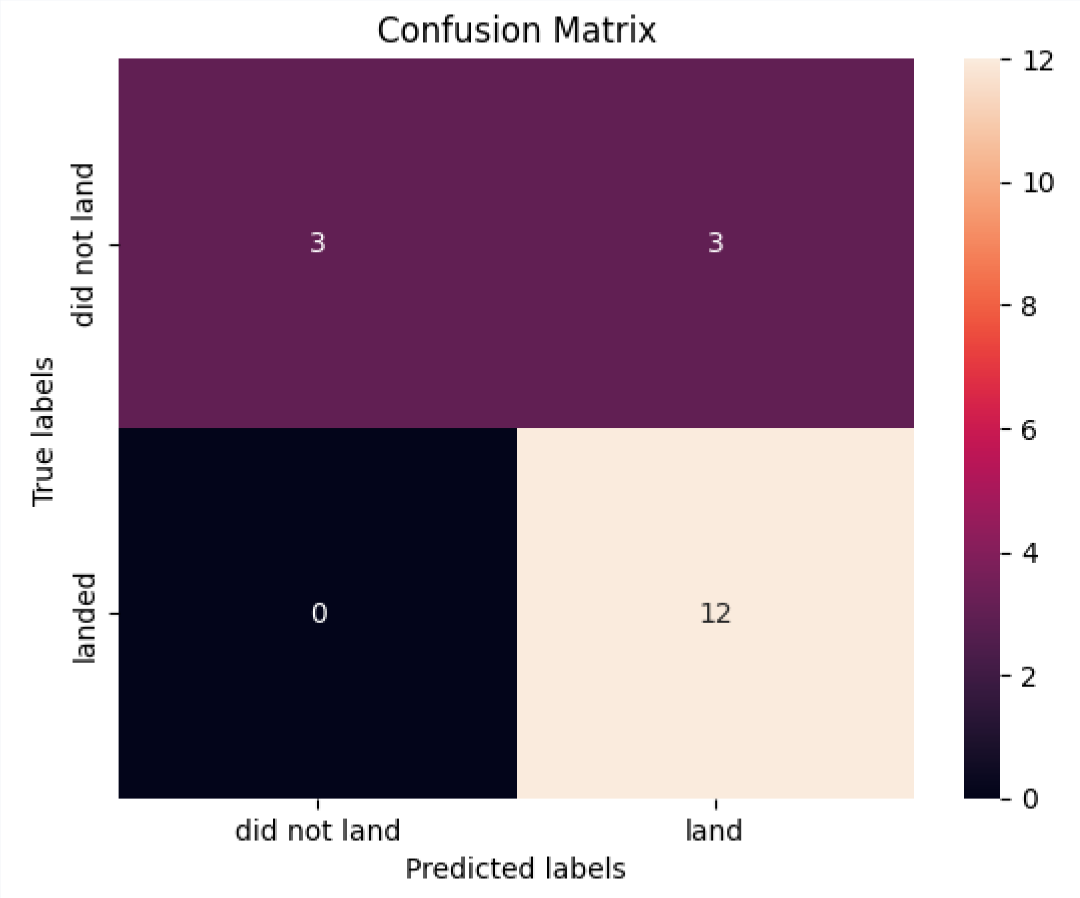
- Test Accuracy (Performance on Unseen Test Data):
  - Logistic Regression: 0.8333
  - SVM: 0.8333
  - Decision Tree: 0.6667
  - KNN: 0.8333
- Cross-Validation Accuracy (Performance During Training):
  - Logistic Regression: 0.8464
  - SVM: 0.8482
  - Decision Tree: 0.8768
  - KNN: 0.8482
- Best Performing Model(s) Based on Test Accuracy: ['Logistic Regression', 'SVM', 'KNN']
- Highest Test Accuracy: 0.8333333333333334
- Best Performing Model(s) Based on Cross-Validation Accuracy: ['Decision Tree']
- Highest Cross-Validation Accuracy: 0.8767857142857143



Decision Tree indicates potential overfitting

Confusion Matrix

- The area of concern is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- Launch Success Trends: KSC LC-39A had the highest success rate. Payload mass and orbit type significantly influenced outcomes. Success rates improved over time.
- Machine Learning Insights:
  - Best test accuracy (83.3%): Logistic Regression, SVM, KNN
  - Highest training accuracy (87.7%): Decision Tree (potential overfitting)
  - Payload mass was a key success factor.
- Interactive Analytics:
- Folium maps revealed launch site proximity patterns.
- Plotly Dash enabled dynamic success rate analysis.

Thank you!