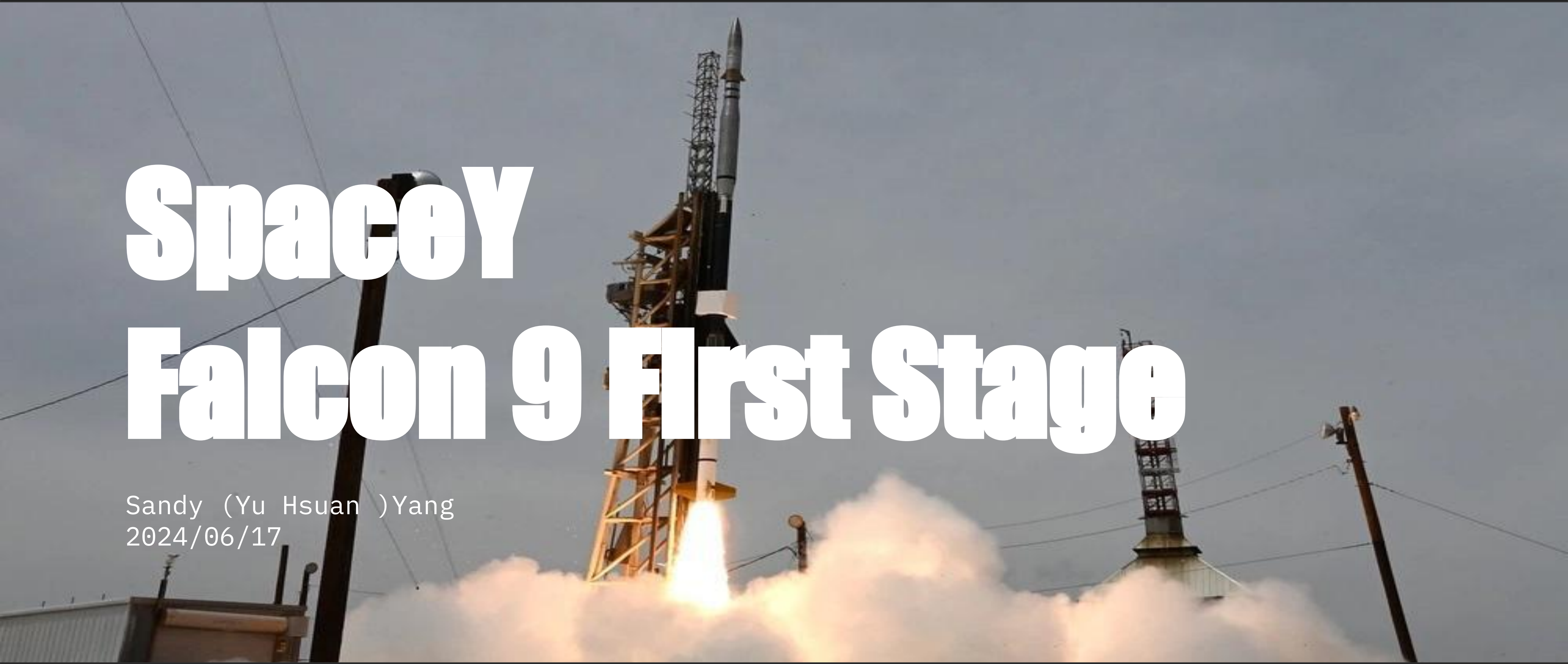


SpaceY
Falcon 9 First Stage

Sandy (Yu Hsuan)Yang
2024/06/17

Table of Content

Executive Summary

The objective of this research is to delineate the key determinants contributing to the **success of rocket landings**.

The study integrates a comprehensive analytical approach, utilizing data acquisition, processing techniques, data exploration and predictive modeling to provide a analysis of the factors influencing rocket landing outcomes.

Executive Summary - Key Finding

Data Visualization Insights:

- The geographical analysis reveals that most launch sites are strategically positioned near the equator and always in close proximity to coastlines, optimizing trajectory and fuel efficiency.

Exploratory Data Analysis Outcomes:

- We've observed a positive trend in launch success over the years, indicating advancements in technology and operational strategies.
- Among the various launch sites, KSC LC-39A stands out with the highest success rates.
- Certain orbits like ES-L1, GEO, HEO, and SSO have demonstrated a 100% success rate, highlighting their reliability in mission planning.

Predictive Modeling Results:

- Comparative analysis of various predictive models shows uniform performance with the decision tree model exhibiting a slight edge in accuracy on the test set, suggesting its effectiveness in handling the dataset's features.

Introduction

Background

SpaceX is pioneering advancements in space travel, dramatically reducing costs through the innovative reuse of rocket components. Unlike traditional launches, which can exceed $165 million, SpaceX has managed to cut this cost to approximately $62 million per launch, primarily due to its successful reuse of the Falcon 9 rocket first stage. This substantial reduction in expense is not only a competitive edge but also a step towards more sustainable space exploration.

This study harnesses publicly available data along with machine learning techniques to predict the reusability of rocket first stages. By doing so, it aims to refine cost estimations and enhance operational efficiencies, potentially benefiting SpaceX and other entities within the space industry.

The project focuses on analyzing how various factors—such as payload mass, launch site, number of flights, and orbit trajectories—impact the success of first-stage landings. Additionally, it evaluates different predictive models to determine which is most effective for forecasting successful landings, a key component in the pursuit of reusable rocket technology.

Methodology

Methodology

1. **Data Collection:** Using API and web scraping with SpaceX REST API alongside advanced web scraping techniques to gather comprehensive launch data.
2. **Data Preparation**: Integrating the dataset with filtering, handling missing values and one-hot encoding:
3. **Data Exploration**:
   a. Conduct Exploratory Data Analysis (EDA) using SQL.
   b. Employed visualization tools such as Folium and Plotly Dash for a dynamic view of the data and preliminary insights.
4. **Model Development:** Developing several classification models to predict the outcome of rocket landings. Models tested included Logistic Regression, Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN).
5. **Model Tuning and Evaluation:** Applied techniques like cross-validation and grid search to fine-tune the parameters for each model. Assessed model performance using metrics such as accuracy, precision, recall, and F1-score to determine the most effective model.

Data Collection
SpaceX API

1. **Request** data using the GET request
2. **Parsed** the JSON output to extract relevant details for our analysis.
3. **Filter** the dataframe to only include Falcon 9 launches
4. **Dealing** with Missing Values with following:
    a. Imputation
    b. Categorical Data
    c. Dropping Data

Data Collection:
Data Scraping

1. **API Request:** Sending an HTTP GET request to the Falcon 9 Launch Wiki page using its UR
2. **Extract Header:** Utilize BeautifulSoup in Python to navigate through the HTML structure. Focus on extracting the column and variable names from the HTML table headers.
3. **Parse HTML Table**: Proceed to parse the HTML table to extract all relevant data about the Falcon 9 launches. Organize and Convert the HTML table data into a structured pandas DataFrame

Data Collection:
Data Wrangling

- **Launches by Site:** Calculated the total number of launches at each site to assess the operational load and identify high-activity locations.
- **Orbit Occurrences**: Determined the frequency of each orbit type used in launches.
- **Mission Outcomes by Orbit**: Calculated the number and occurrence of mission outcomes (such as success or failure) per orbit type.
- **Creating Landing Outcome Labels**: Transformed the 'Outcome' column in our dataset to create a clear, categorical label indicating the landing outcome (e.g., success, failure).

EDA with Data Visualization

With EDA, we utilized data visualization techniques to explore and illustrate the relationships between various launch parameters and outcomes.

Launch Dynamics Visualizations:

Flight Number and Launch Site: Visualized the distribution of flights across different launch sites to assess site-specific launch activity.

Payload vs. Launch Site: Examined how payload capacities vary across different launch sites, potentially influencing launch success.

Flight Number and Orbit Type: Explored the relationship between the sequence of flights and their designated orbits.

Outcome-Oriented Visualizations:

Success Rate by Orbit Type: Displayed the success rates for each orbit type.

Payload and Orbit Type: Analyzed how payload variations correlate with different orbit types to determine if certain payloads are more common for specific missions.

Yearly Launch Success Trends: Illustrated changes in launch success rates over time, highlighting improvements in operations or persisting challenges.

EDA with Data SQL

In this phase of our project, we leveraged SQL to conduct a thorough exploratory analysis of the Falcon 9 launch data.

- **Data Overview**: Initial queries were executed to get a basic understanding of the data structure and volume. This involved SELECT statements to view top rows, COUNT queries for total entries, and descriptive statistics like MIN, MAX, and AVG for key numeric columns.
- **Aggregation Queries**: Aggregate functions to calculate summaries of important metrics, such as the average payload size and the total number of launches per site.
- **Comparison Queries**: Compared various factors across different conditions, such as launch outcomes between different sites or under different weather conditions, using conditional SQL statements like WHERE.

Query Index::

a. SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL
b. SELECT FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
c. SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
d. SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
e. SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'
f. SELECT PAYLOAD FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
g. SELECT MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME
h. SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
i. SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] FROM SPACEXTBL where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
j. SELECT [Landing _Outcome], count(*) as count_outcomes \
k. FROM SPACEXTBL WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC

Build an Interactive Map with Folium

1. **Initialize Map**: Begin by creating a base map using Folium to visualize geographic details. Set initial zoom levels and center coordinates based on the geographical spread of the launch sites.
2. **Mark Launch Sites**:
- Utilize distinct icons to differentiate between launch sites.
- Integrate tooltips or popups to display the name of each site and a summary of launch statistics .
- Categorize launch outcomes by marking successful and failed launches site with different color.

3. **Calculate Distances:** Compute the distance from each launch site to key proximities (e.g., nearby cities, airports, or other relevant geographic features).

4. **Visualize Distances:**

- Draw lines between launch sites and points of interest to visually represent distances.
- Incorporate a feature to display distance information on the map when users interact with these lines or markers.

Build a Dashboard with Plotly Dash

- **Initialize Dashboard**: Set up the basic structure of the dashboard using Plotly Dash.
- **Utilizing the callback function** so when a user selects a site, the dashboard will update other components accordingly to reflect data specific to the **chosen site** and **payload Range**.

Input Component:

- Launch site dropdown list
- Payload Range Selection Slider

Output Component:

- Pie Chart for Launch Success Overview
- Scatter Chart for Payload vs. Launch Success

Predictive Analysis (Classification)

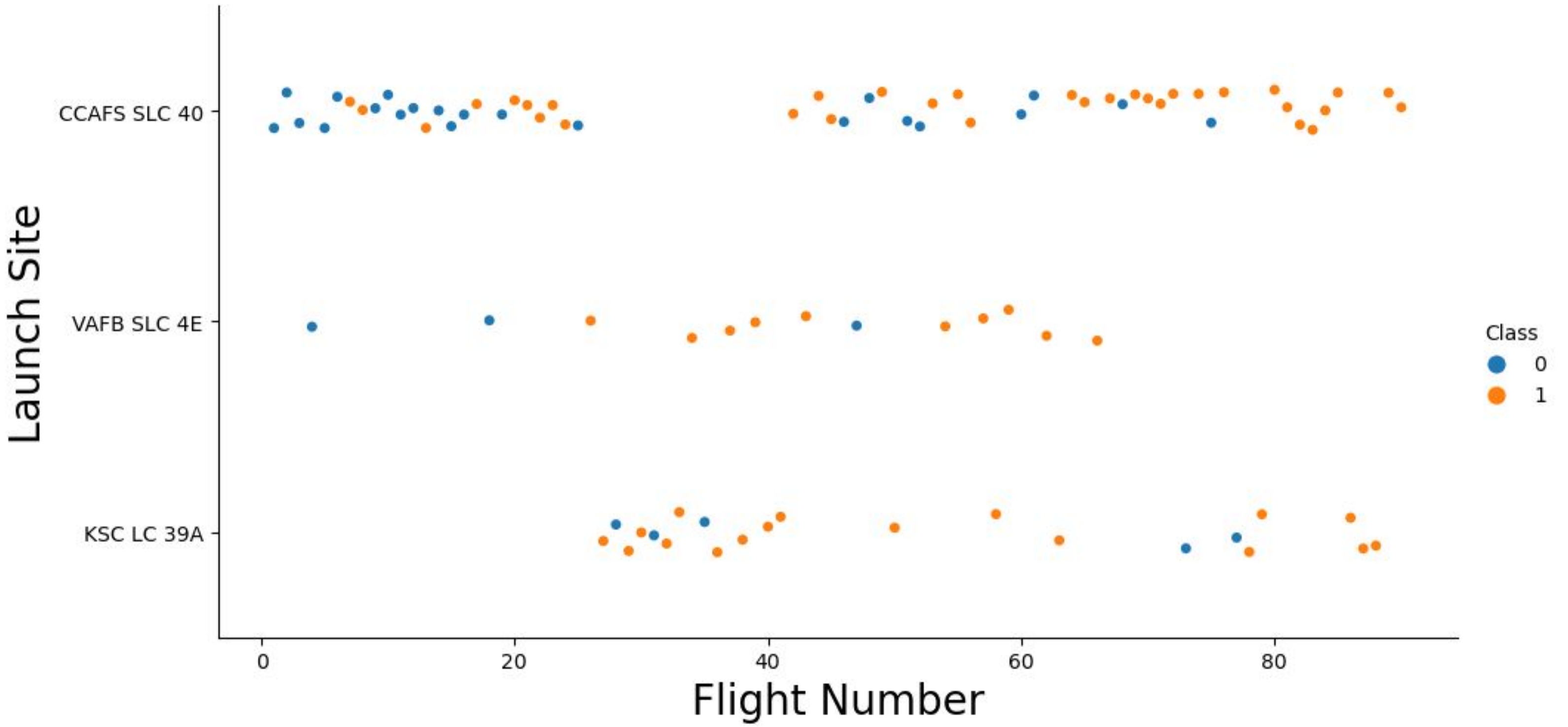
- **Define Variables**: Create the dependent variable (Y) and independent variable (X)
- **Standardize** the data with StandardScaler.
- **Dataset Splitting**: Split the dataset into a training set and a testing set with test size =0.2
- **Create Predictive Models**: Develop several models to find the best performer:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - KNN Classification
- **Model Refinement**: Apply Grid Search CV (Cross-Validation) techniques to optimize the hyperparameters of each model.
- **Model Evaluation**: Plot performance metrics and calculate each model scores to identify the one that performs the best on the testing dataset.

Result - EDA

Results Summary

- **Exploratory Data Analysis**
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- **Visual Analytics**
  - Most launch sites are near the equator, and all are close to the coast
  - Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities
- **Predictive Analytics**
  - Decision Tree model is the best predictive model for the dataset, even though the accuracy just slightly better

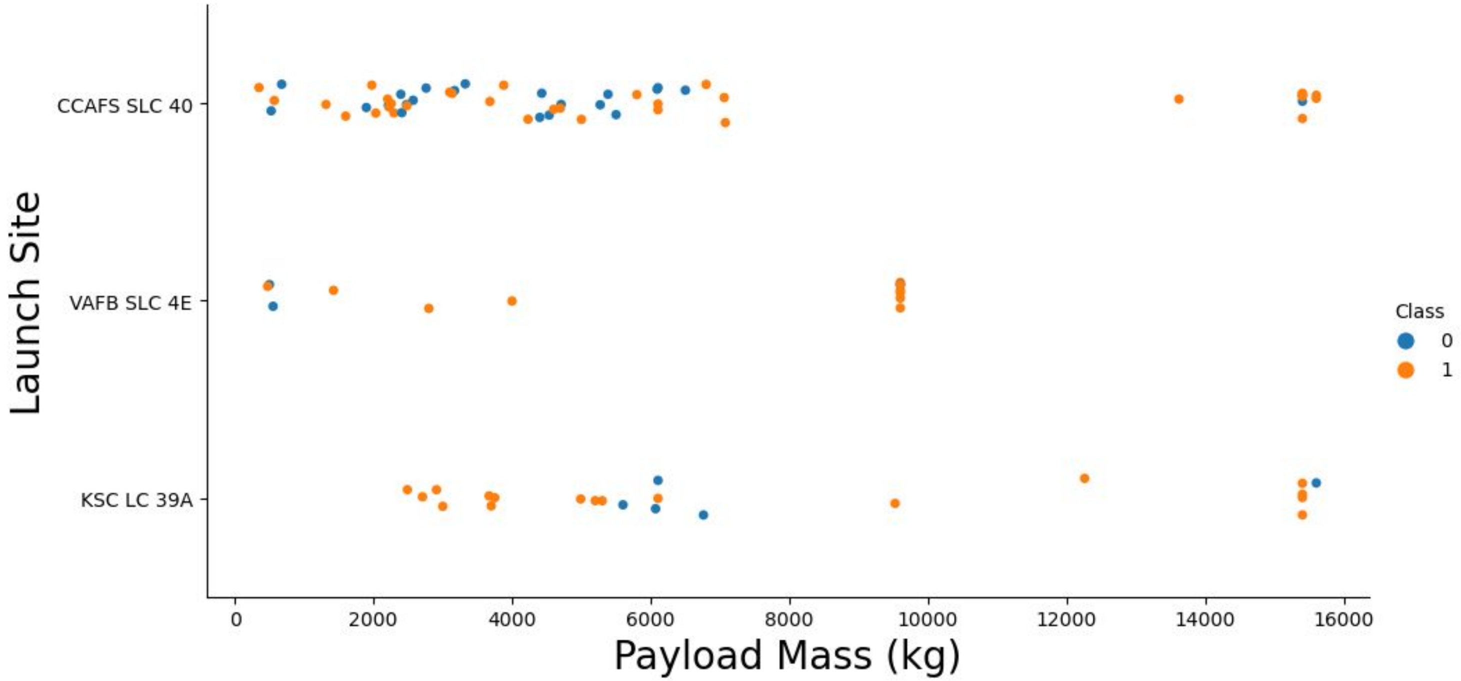
Flight Number vs. Launch Site



Exploratory Data Analysis

- As blue = failure, earlier flight typically has lower success rate
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates, yet there are fewer launches from this site.
- Recent launches tend to have a higher success rate

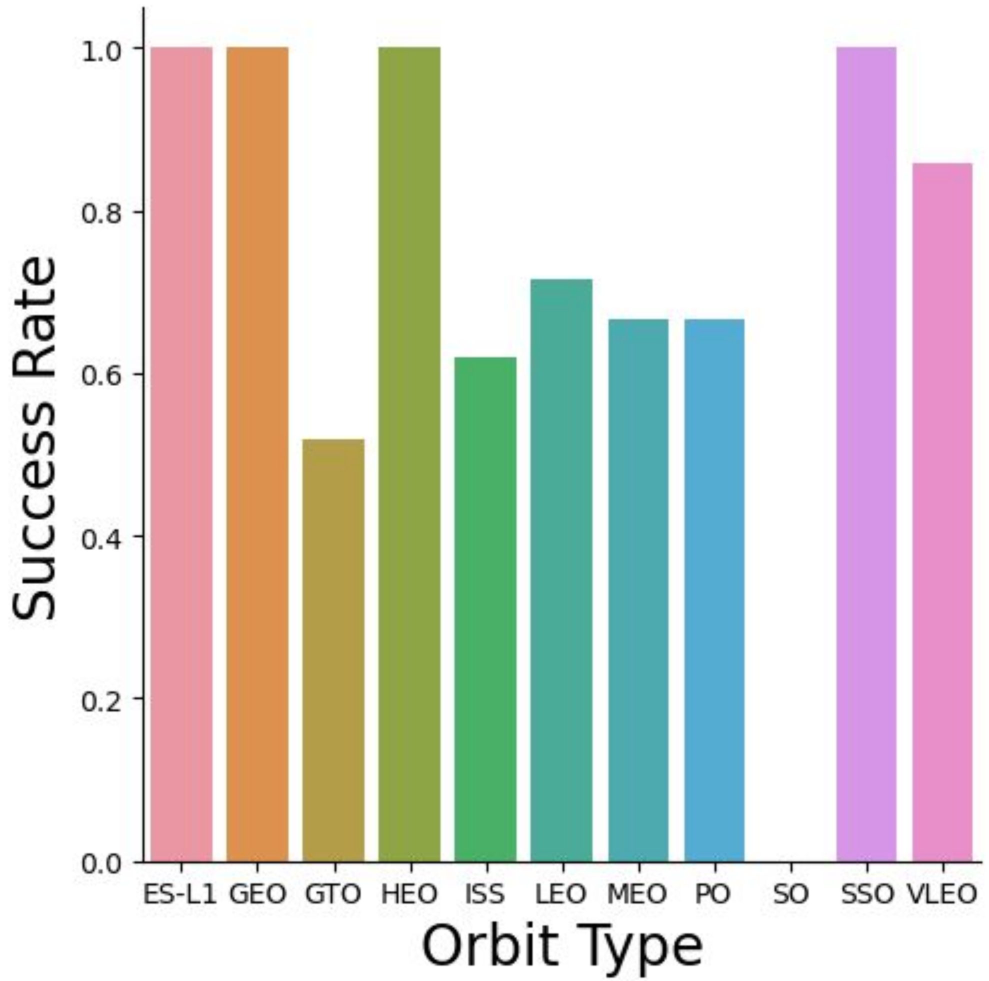
Payload vs. Launch Site



Exploratory Data Analysis

- The higher the payload mass(kg), typically the success rate is high
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SLC 4E has not launched anything greater than ~10,000 kg

Success Rate vs. Orbit Type



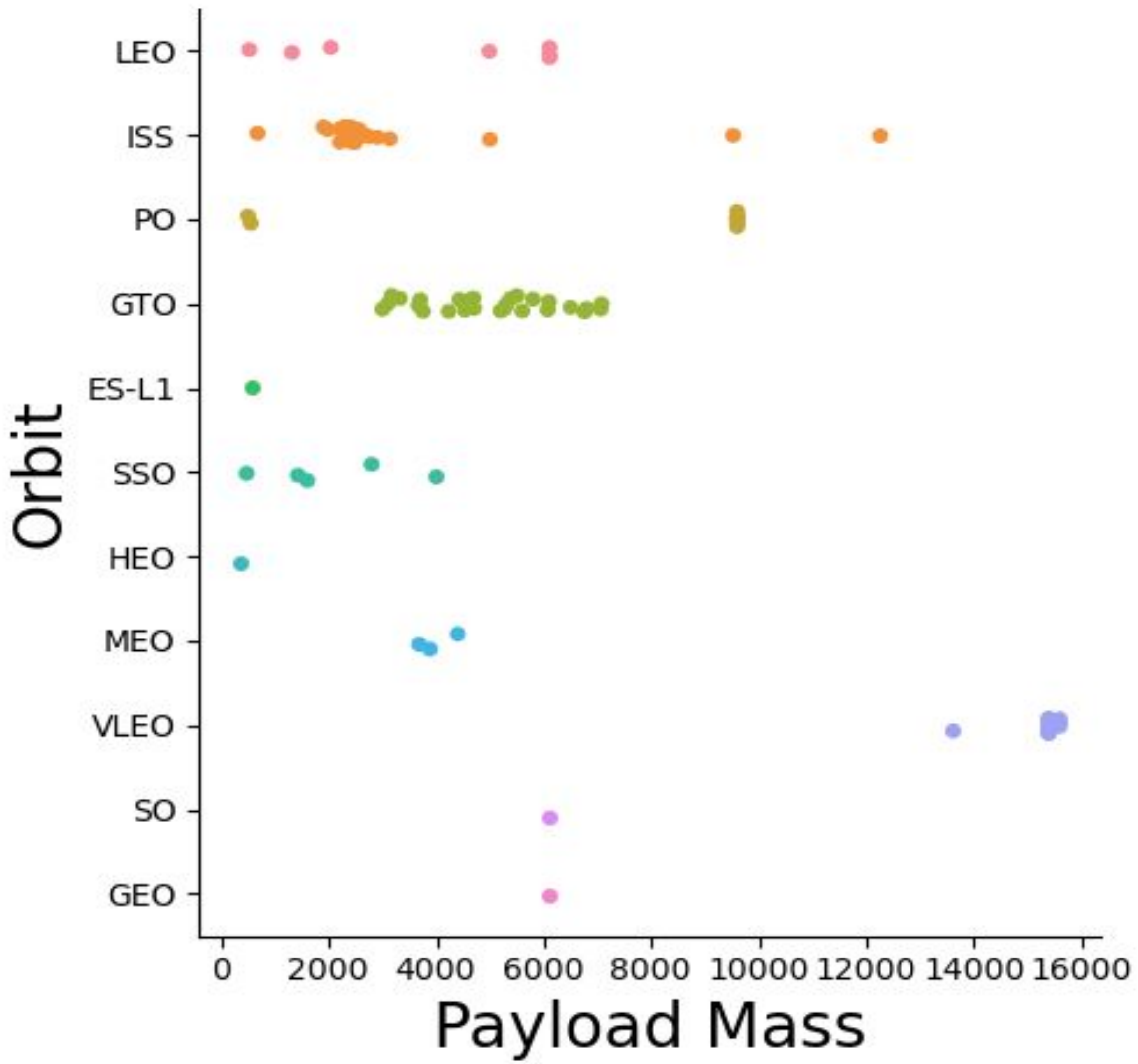
Exploratory Data Analysis

- Orbit Type with 100% Success Rate:
    - ES-L1
    - GEO
    - HEO
    - SSO

- Orbit Type with 50%-80%Success Rate
    - GTO
    - ISS
    - LEO
    - MEO
    - PO

- 0%Success Rate:
    - SO

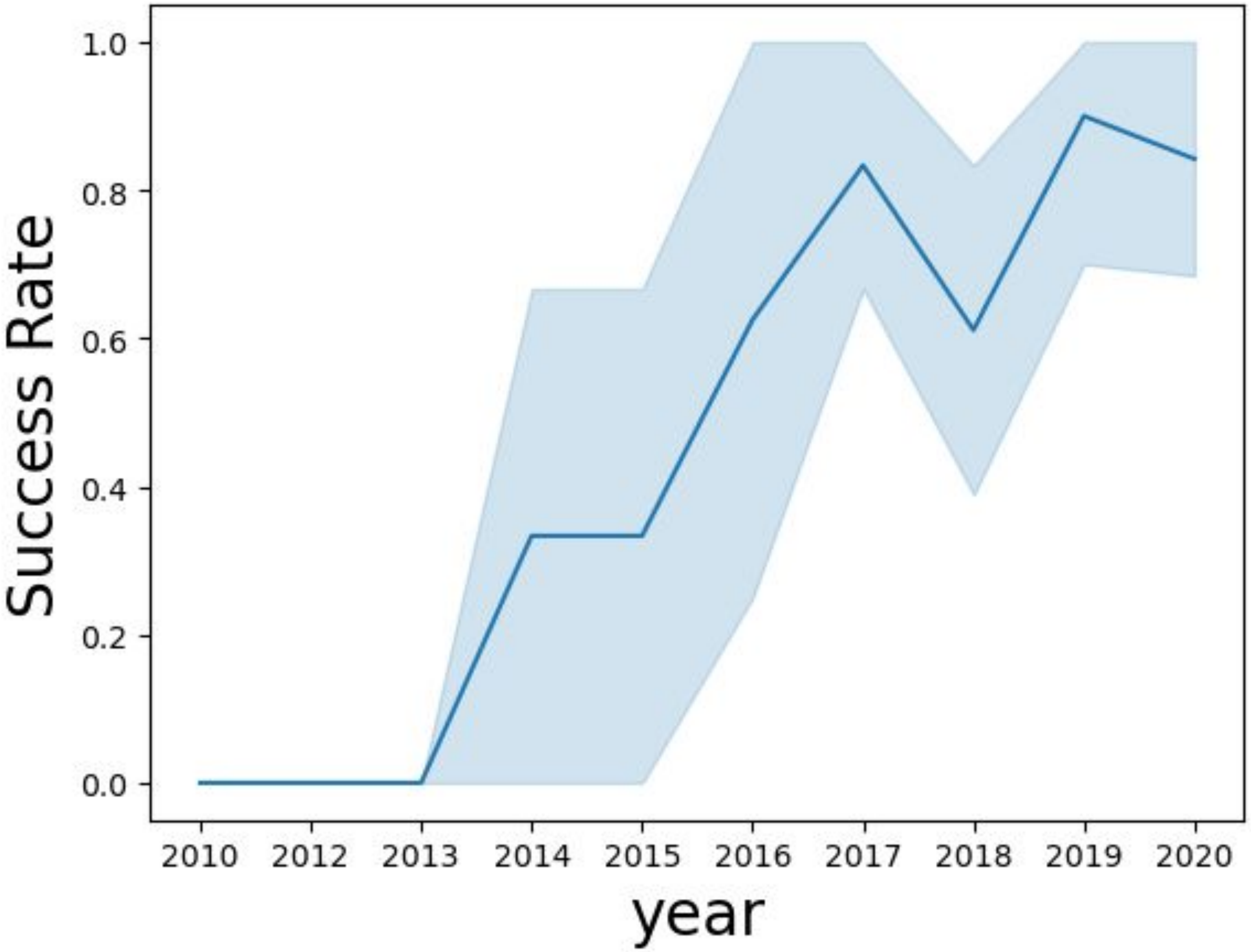
Payload vs. Orbit Type



Exploratory Data Analysis

- VLEO tend to have high Payload Mass, between 12,000 to 16,000
- Most of the ISS does not contain high Payload Mass
- GTO has Payload around 2,000 to 8,000
- ISS has more wilder distribution of Payload Mass, while majority of ISS is around 2,000 to 3,000
- SO and GEO does not have many data points
- HEO has the least Payload Mass while VELO has the highest

Launch Success Yearly Trend



Exploratory Data Analysis

- The success rate increased from 2013 to 2014, 2015 to 2017, and 2018 to 2019
- The success rate declined from 2017 to 2018 and from 2019 to 2020
- From overall 2013-2020, the launch success rate is continuously improving

About - Launch Site

All Launch Site Names

- CCAFSLC-40
- CCAFSSLC-40
- KSCLC-39A
- VAFBSLC-4E

Query:

```
[30]: %sql ibm_db_sa://yyy33800:dwNKg8J3L0IBd6CP@1bbf73c5
      %sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL;

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
   sqlite:///my_data1.db
Done.
```

[30]:

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

Records with Launch Site Starting with CCA

```
Done.
```

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total payload mass:

45,596 KG total carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

On average, alaunch with booster F9 v1.1 has carried payload Mass 2,928.4KG

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

2015/12/22

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SCAT-14
- JSCAT-16
- SES-10
- SES-11 / EchoStar 105

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

Failure (in flight): 1

Success: 99

Success (payload status unclear): 1

In [14]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

Out[14]:

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

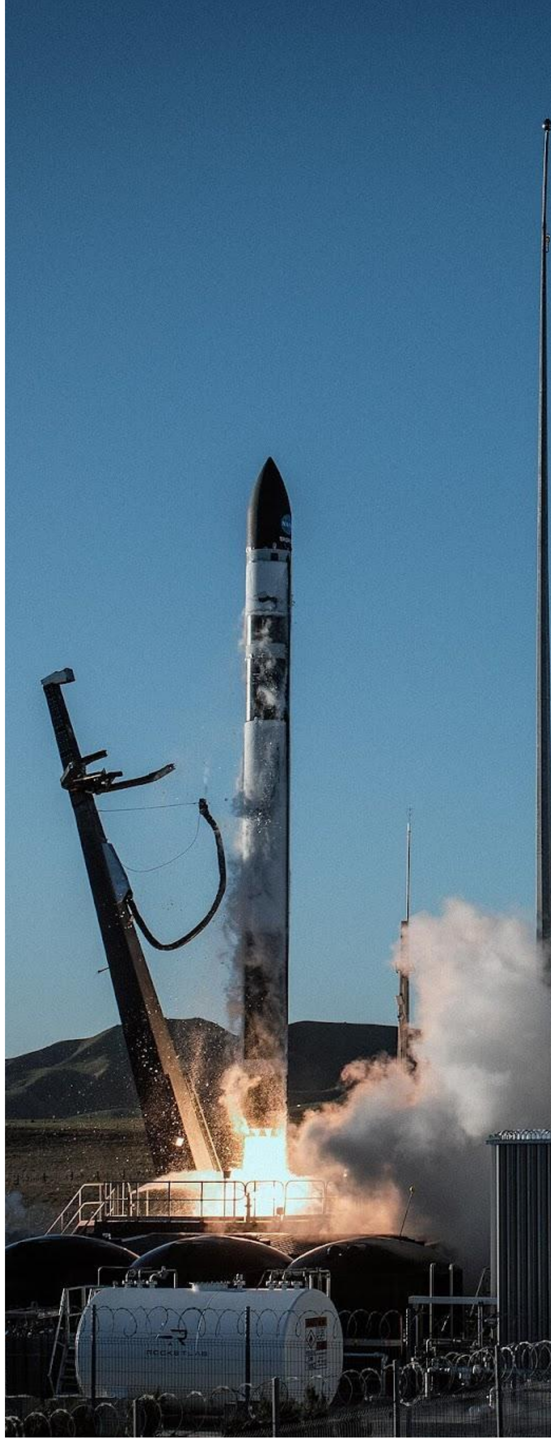
Boosters Carried Maximum Payload

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 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

```
%sql SELECT 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 Failure Launch Records

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='
```

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Records of the month names, booster versions, and launch site for the months in year 2015

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

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-6-04' and '2017-03-20' \
group by [Landing _Outcome] order by count_outcomes DESC;
```

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

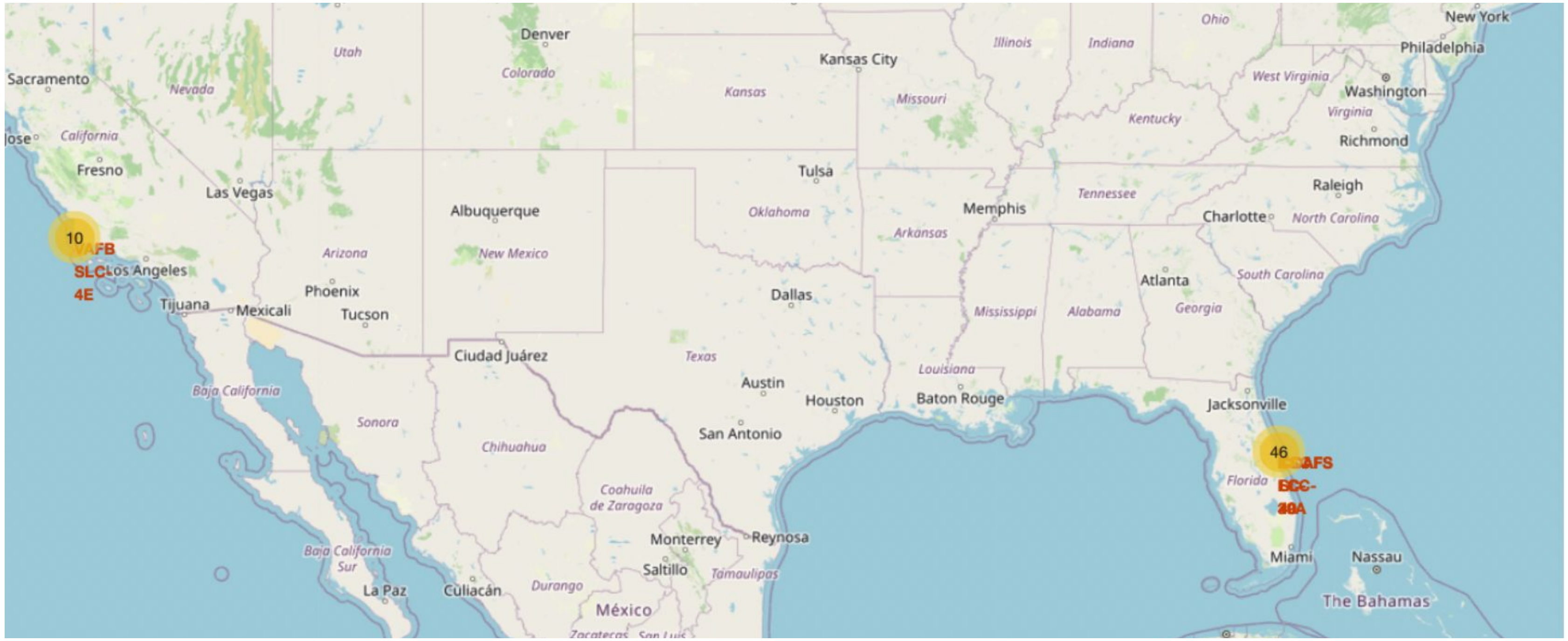
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Rank of successful landing outcomes between the date 04-06-2010 and 20-03-2017

Result - Launch Sites Proximities Analysis

Launch Site

Equatorial Advantage: With the visualization of Folium, we notice that launch sites positioned close to the equator benefit significantly during missions to equatorial orbits. This proximity allows for more efficient launches, as rockets can leverage the Earth's rotational velocity to achieve a prograde orbit more easily. The rotational speed of the Earth near the equator provides a natural boost to rockets, enhancing their velocity.



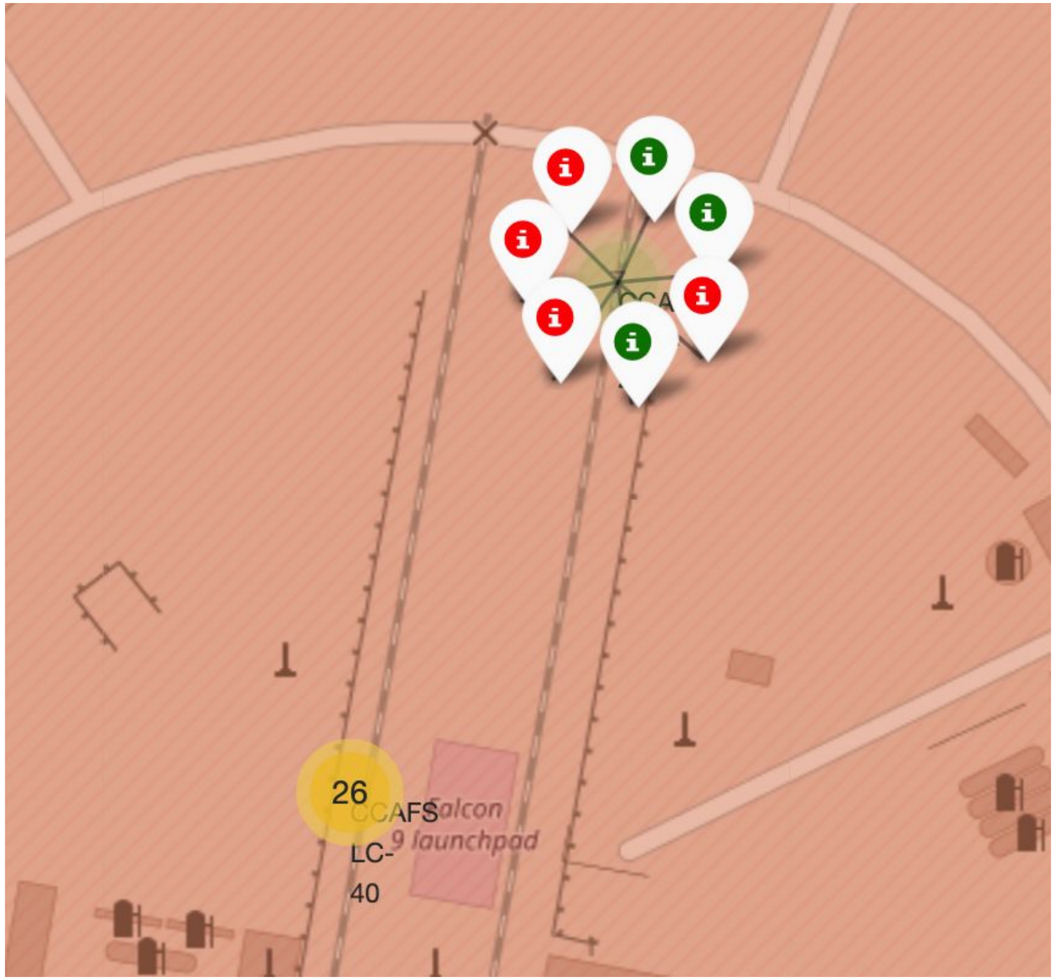
Launch Outcome

For Each Launch Site, we utilize Folium **marker for launch outcome**

Launce Outcomes:

- **Green shown as successful**
- **Red shown as failure**

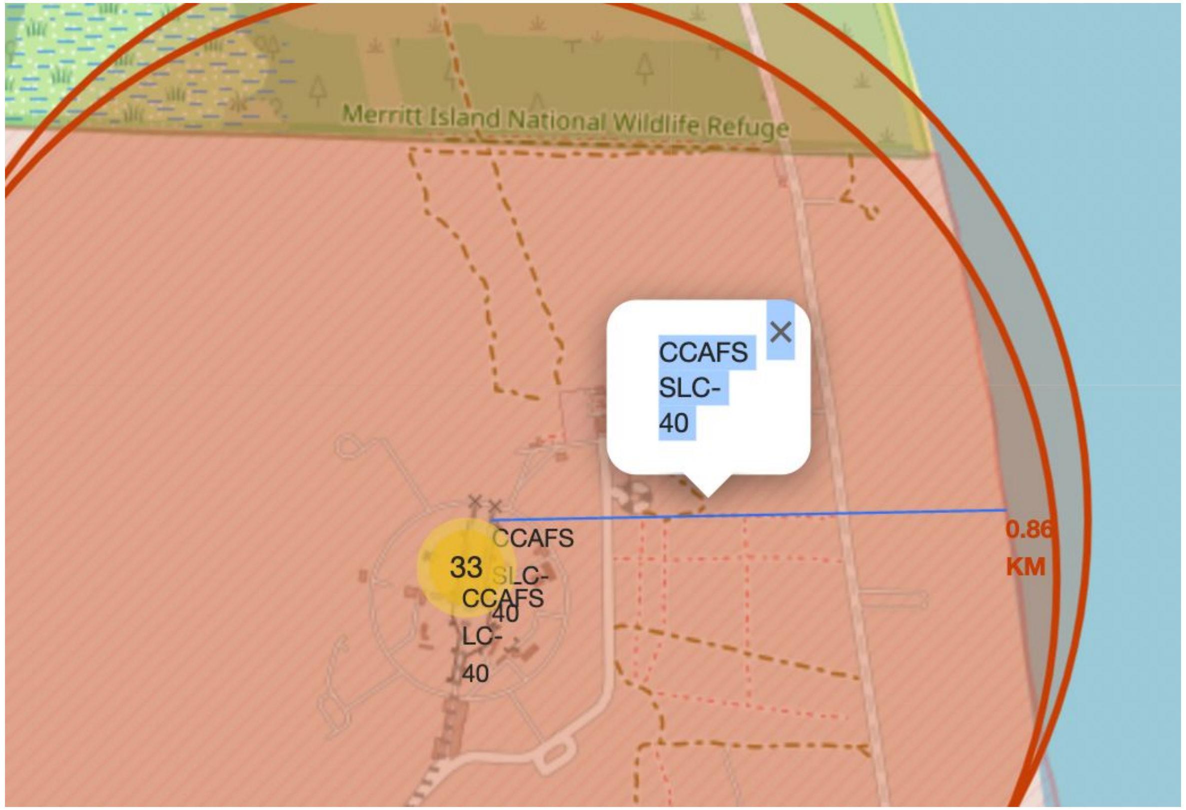
For example:
Launchsite CCAFS SLC-40 has 3 successful launch and 4 unsuccessful launch.



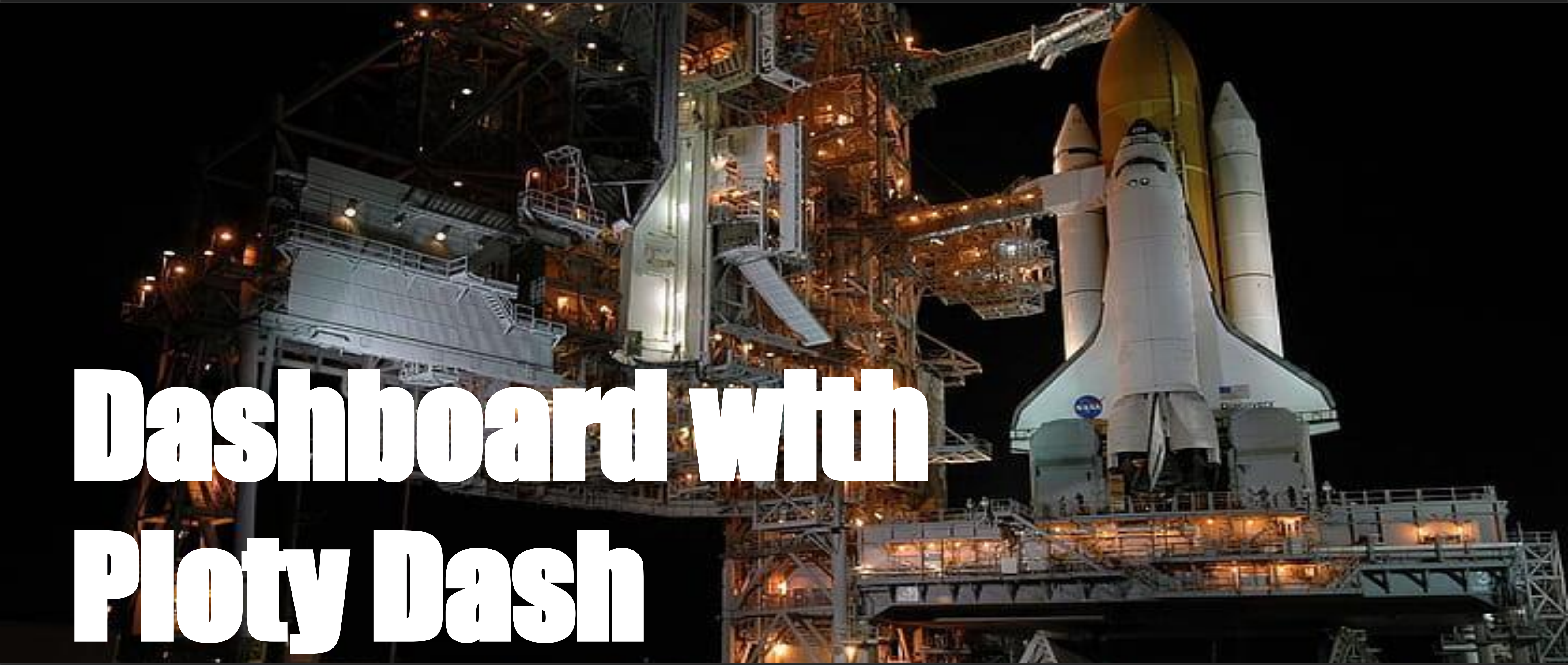
Distance Observation

Coast line position:
Lat: 28.56367 Long: -80.57163

Launch site CCAFS SLC-40 is 0.86KM away from the coast line.



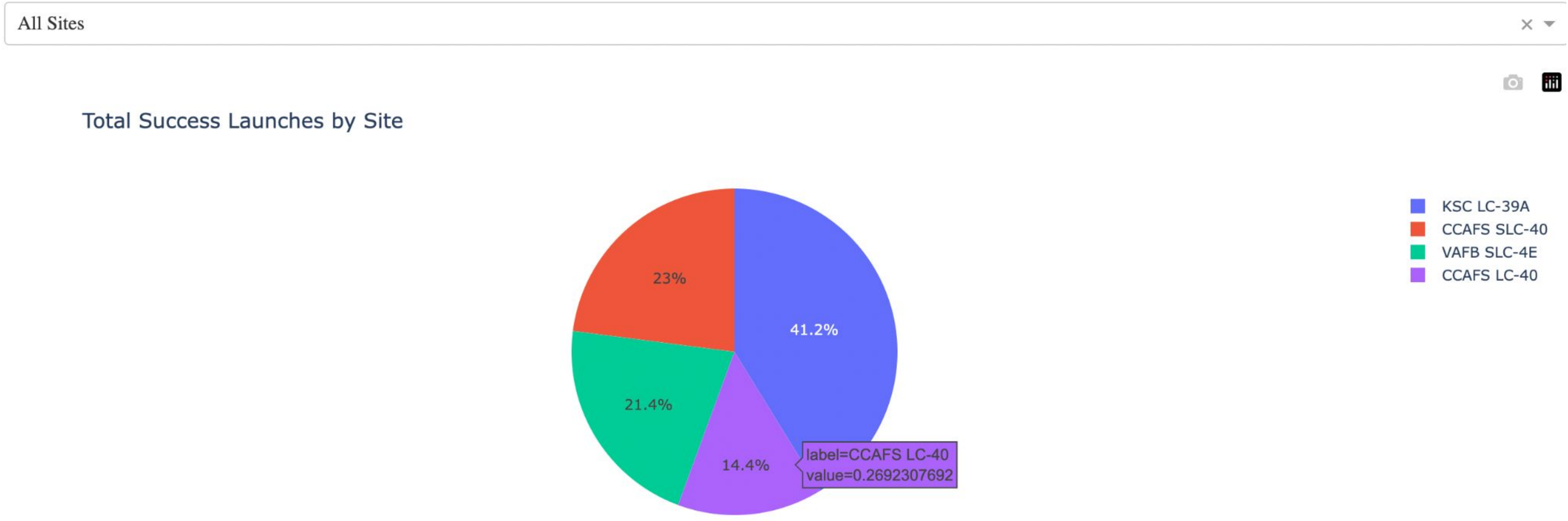
Dashboard with Ploty Dash



Launch Success by Site

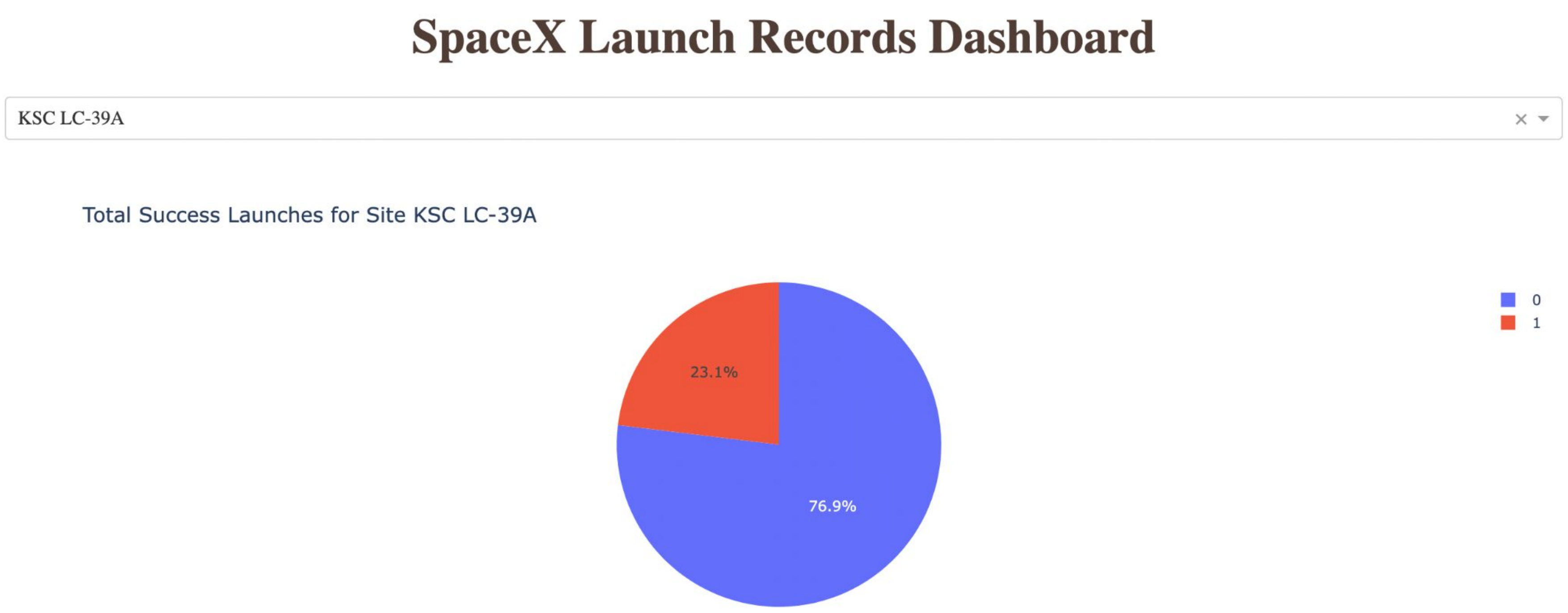
Among all launch site, **KSC LC-39A** has the highest successful launches(41.2%)

Among all launch site, **CCAFS LC-40** has the lowest successful launches amongst launch sites (14.4%)



KSC LC-39A

has the highest success rate amongst launch sites (76.9%)



Payload Mass v.s. Booster Versions v.s. Successful Landing

We can notice that booster version since to have more impact on the successful landing. While 1 represent successful and 0 represent failure, Version v1.1 seems to have high chance of failing no matter the Payload Mass. On the other hand, FT tend to have higher successful rate.

Predictive Analysis

Classification

In classification analysis, all chosen models demonstrated approximately equal performance levels, reflected in similar **scores** and **accuracy** metrics across the board. This uniformity in performance is likely attributed to the constraints imposed by the relatively small size of the dataset.

Among the models, the **Decision Tree** slightly outperformed the others. The metric .best_score_ was used to evaluate its superiority.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

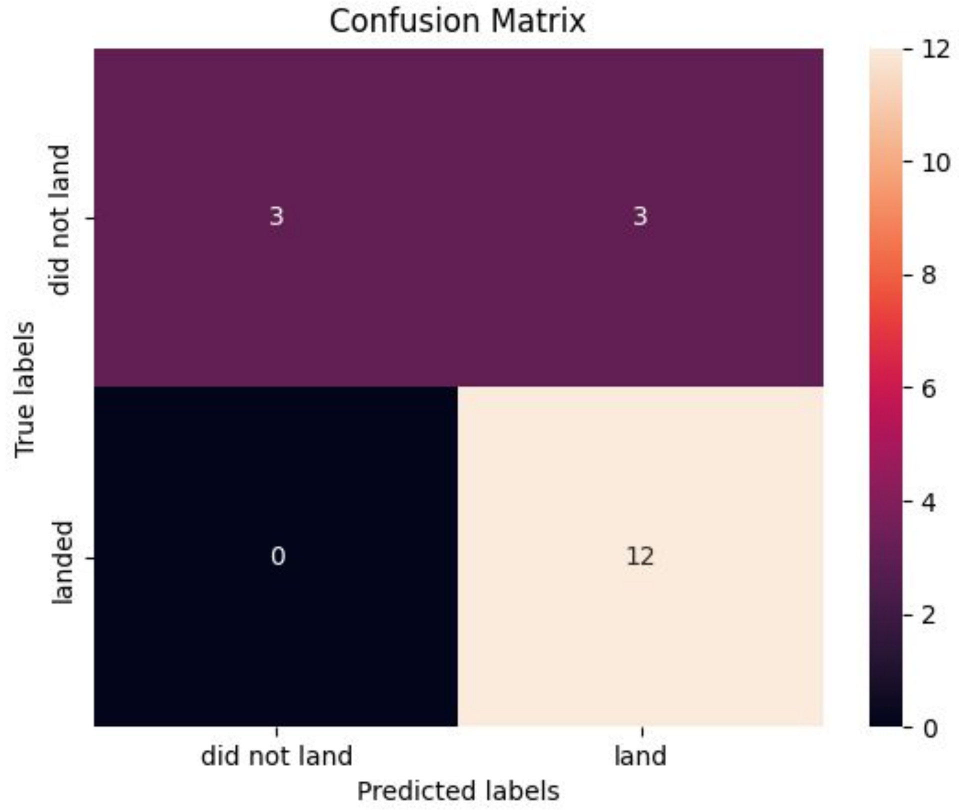
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9017857142857144
Best params is : {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_sample
s_leaf': 2, 'min_samples_split': 2, 'splitter': 'best'}
```

Confusion Matrix- Evaluation

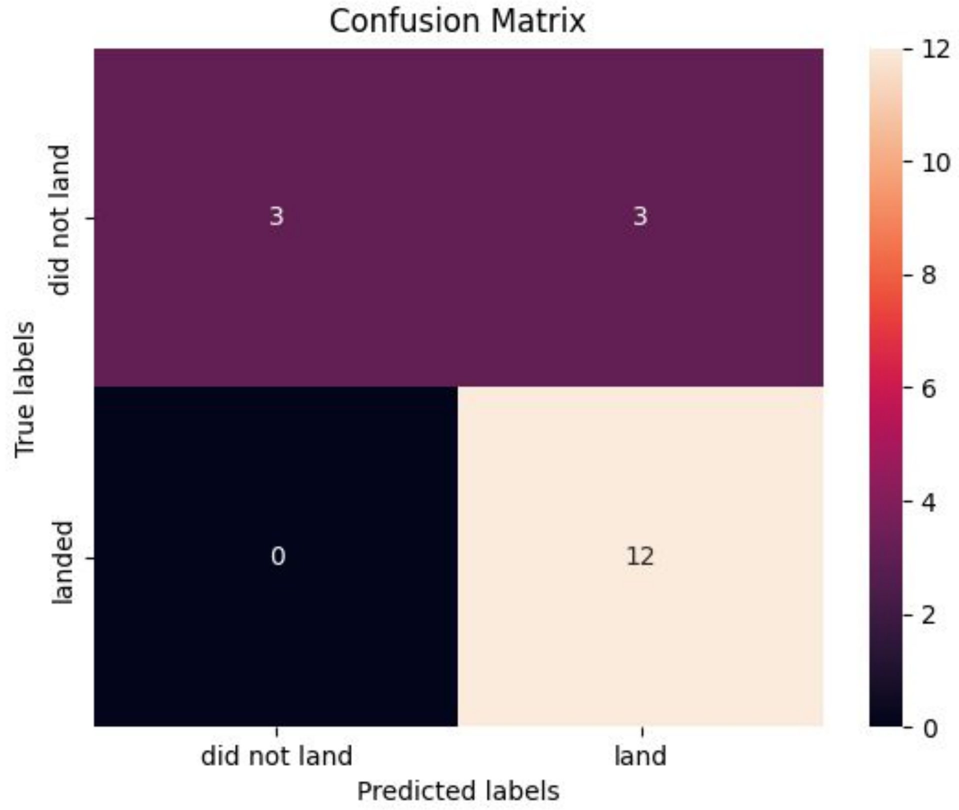
The confusion matrices generated for all models were **identical**, indicating a consistent performance across different algorithms. However, the presence of **false positives** (**Type 1 errors**) is notable and implies that the model incorrectly predicted some negative cases as positive. This is an area for potential improvement.



Confusion Matrix: Score

Key Performance Metrics:

- **Precision:** Calculated as TP / (TP + FP) = 12 / 15 = 0.80. This indicates that 80% of positive predictions were correct.
- **Recall:** Calculated as TP / (TP + FN) = 12 / 12 = 1. This indicates that the model successfully identified all actual positives.
- **F1 Score:** The harmonic mean of precision and recall, given by 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.8 * 1) / (0.8 + 1) = 0.89. This score balances both precision and recall.
- **Accuracy:** Calculated as (TP + TN) / (TP + TN + FP + FN) = (12 + 3) / (12 + 3 + 3 + 0) = 0.833, or 83.3%, indicating the overall rate at which the model correctly predicts outcomes.



Conclusion

Model Performance:

- The decision tree model marginally outperformed other models in predicting launch outcomes, highlighting its robustness in handling the dataset provided.

Geographical Factors:

- **Equatorial Proximity:** Most launch sites are strategically located near the equator to leverage the Earth's rotational speed. This geographical advantage significantly reduces the need for extra fuel and boosters, optimizing launch costs.
- **Coastal Locations:** All launch sites are situated close to coastlines, facilitating safer and more manageable launch and landing operations, particularly for missions requiring oceanic recoveries.

Trends in Launch Success:

- There is a notable improvement in launch success rates over time, reflecting technological advancements, increased experience, and enhanced operational procedures.

Conclusion

Launch Site Performance:

- **KSC LC-39A:** This site stands out with a 100% success rate for launches carrying payloads less than 5,500 kg, underscoring its operational excellence and reliability.

Orbital Missions:

- Missions to **ES-L1, GEO**, **HEO**, and **SSO** orbits have achieved a 100% success rate, indicating high reliability for these specific orbital transfers.

Future Improvement

To enhance the predictive accuracy and operational insights from our analysis of SpaceX Falcon 9 launches, the following improvements are recommended:

Data Enrichment:

- **Increase Dataset Size:** Expand the dataset to include more launches, which could help in capturing a broader range of variability and trends.
- **Incorporate Additional Variables:** Include variables such as weather conditions, technical specifics of the rocket modifications, and crew experience levels to refine the predictive models further.

Model Enhancement:

- **Advanced Algorithms:** Explore more sophisticated machine learning algorithms such as **Random Forests** and **Gradient Boosting Machines**, which may provide better performance due to their ensemble methods.

Analytical Depth:

- Temporal Analysis: Perform time-series analyses to predict trends over time more accurately and to understand how changes in technology or procedures impact success rates.

Thank you!