Space Y
Data Science Capstone Project

Sanad Fakhoury

8/26/2024

Presentation Content

Executive Summary

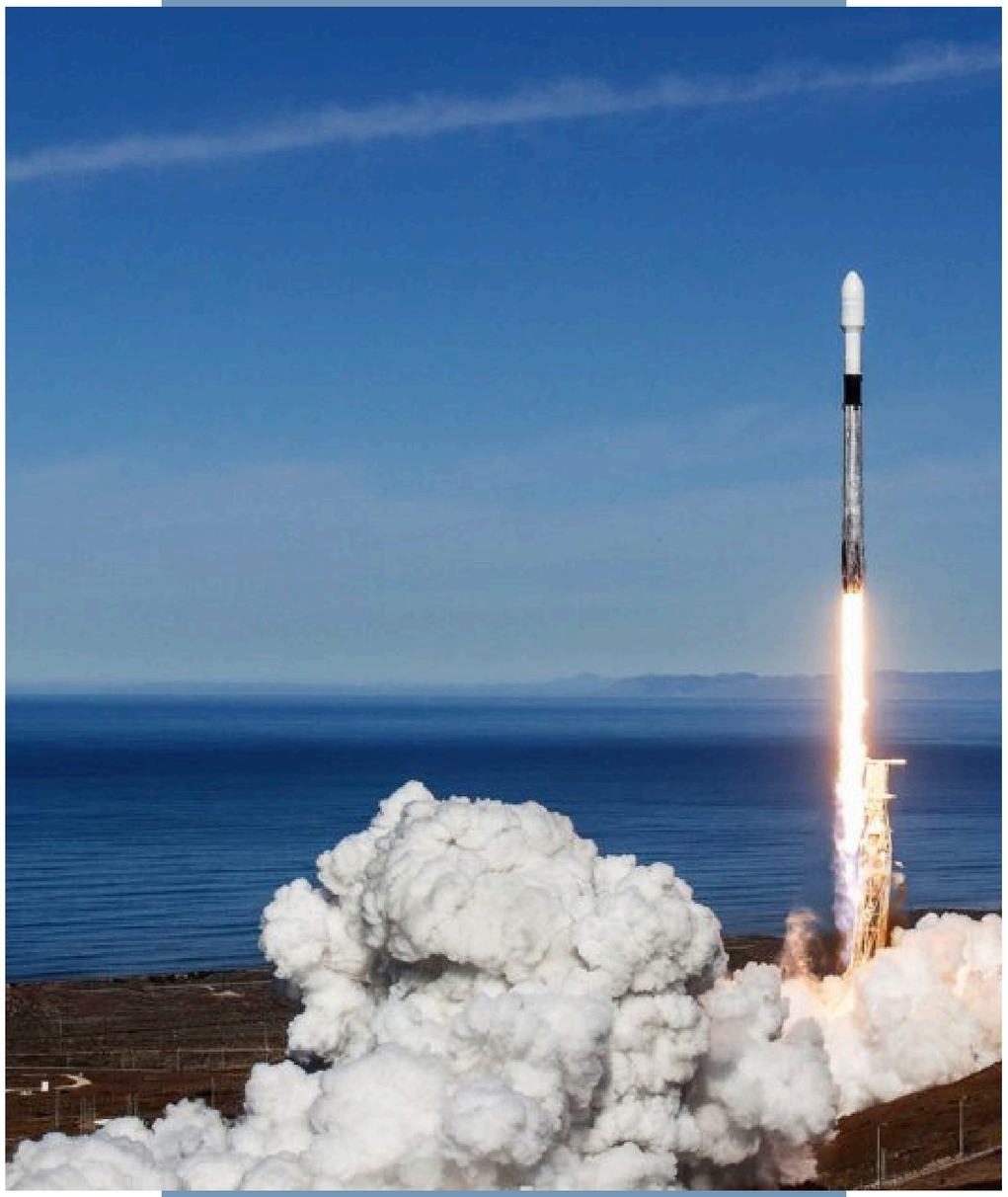
Introduction

Methodology

Results

- EDA with visualization
- EDA with SQL
- interactive map with Folium
- Plotly Dash dashboard
- predictive analysis

Conclusion



Executive Summary

- **Data Collection**
- **Data Wrangling**
- **Exploratory Data Analysis (EDA)** with Data Visualization techniques taking into account the following data: Payload launch site, the flight number and yearly trends
- **Analyze** the data with SQL, calculating the following data: Total Payload, Payload Range for successful launch and total of successful and failed outcomes
- **Build Models** to predict landing outcomes using logistic regression, K-nearest neighbor (KNN), support vector machine(SVM) and decision tree.
- **Building** an interactive map with folium.
- **Building** a Dashboard with Plotly Dash.

Results

predictive Analytics:

- All models demonstrated comparable performance on the test set, with the decision tree model slightly outperforming the others.

Visualization/Analytics:

- Most launch sites are located close to the coast, with many situated near the equator.

Exploratory Data Analysis:

- The KSC LC-39A landing site has the highest success rate among all locations.
- Over time, the success rate of the launches has improved.
- The ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Introduction

SpaceX has long dominated the affordable rocket launch market, operating without direct competition for years. However, SpaceY is poised to enter the space race and bring competition to this market. To succeed, SpaceY must focus on keeping launch costs down and maximizing the reusability of rocket stages, much like SpaceX has done. The reuse of the first stage of SpaceX's Falcon 9 rocket has been a key factor in their ability to offer launches at an average cost of $62 million, significantly lower than other providers who charge around $165 million per launch. By leveraging publicly available data from SpaceX, we aim to analyze the factors that influence the successful landing of the first sage, which directly impacts launch costs.

In this analysis, we will explore how payload mass, launch site, number of flights, and orbit types affect the success rate of first-stage landings. We will also examine the rate of successful landing over time identify the best predictive model for determining whether a first stage will successfully land. Key questions include: What features determine the success of a first-stage landing? What conditions are required for a successful landing? These insights will be crucial for understanding how to maintain low launch costs and ensure the reusability of rocket stages, a vital aspect of competing in the space launch market.

Methodology

Methodology

Steps

Data Collection Methodology:

- Data extracted from SpaceX via REST API and web scraping Wikipedia launch tables.

Perform data wrangling:

- Filter data, handle missing values, and apply one-hot encoding.
- Convert categorical data regarding launch success to continuous values and append to the dataframe.

- Conduct exploratory data analysis (EDA) using SQL and data Visualization techniques.
- Visualize data using interactive tools like Folium and Plotly Dash.

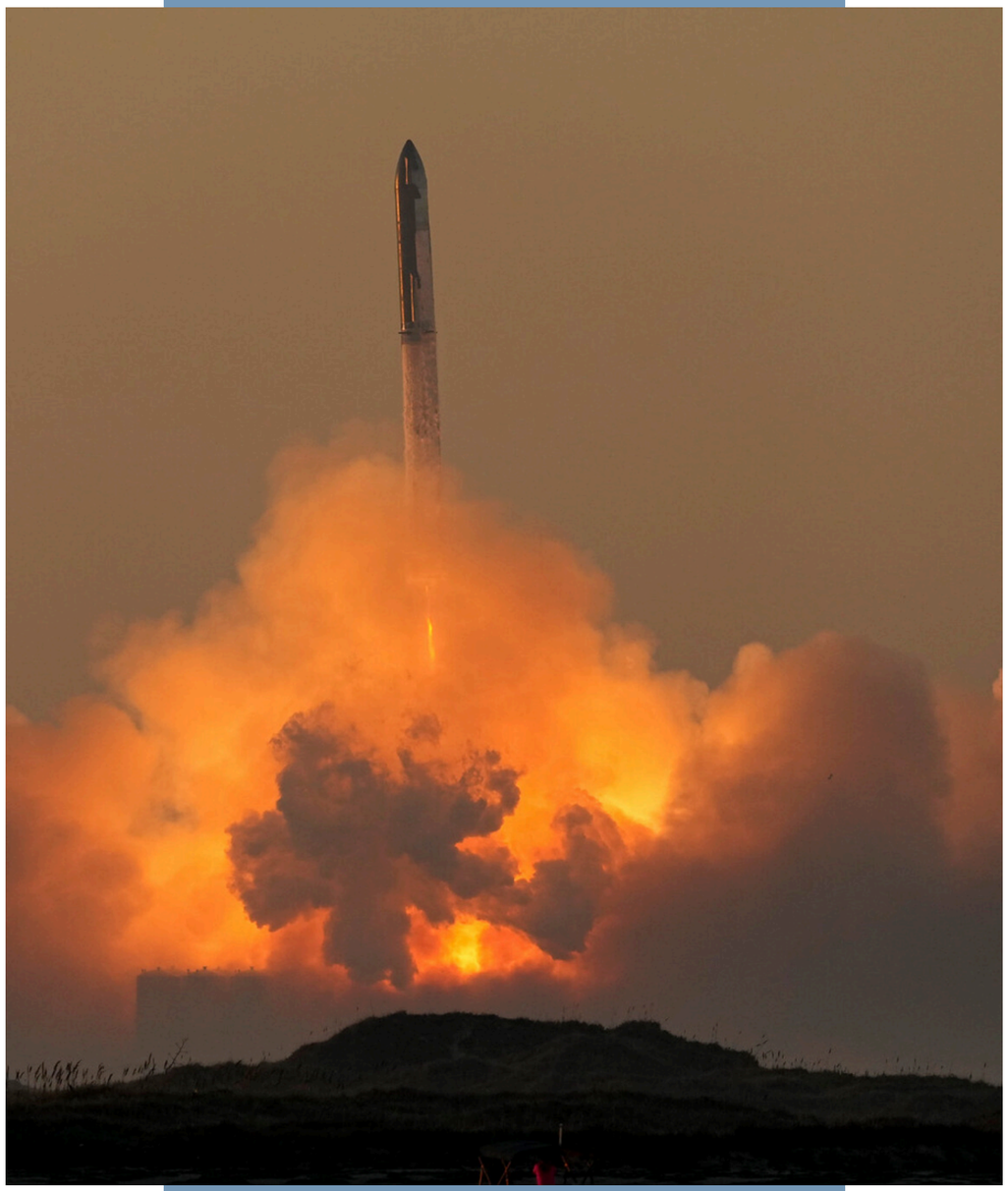
Build classification models to predict landing outcomes:

- Tune and evaluate models to identify the best model and parameters.

Data Collection - API

Steps:

- retrieve rocket launch launch data by sending a request to the SpaceX API.
- Parse the response with **.json()** and transform it into a dataframe using **.json_normalize().**
- Obtain additional launch details from the SpaceX API by implementing custom functions.
- Organize the retrieved data into a dictionary.
- Convert the dictionary into a dataframe for further processing.
- Narrow down the dataframe to include Falcon 9 launches.
- Fill in missing values in the Payload Mass column by using the **.mean()** of the data.
- Save the processed data to a CSV file subsequent analysis.



Data Collection - Web Scraping

Steps:

- Retrieve Falcon 9 launch data by making a request to Wikipedia.
- Construct a BeautifulSoup object from HTML response.
- Extract the column headers from the HTML table.
- Parse the HTML tables to gather the necessary data.
- Organize the extracted data into a dictionary.
- Convert the dictionary into a dataframe for easier manipulation.
- Save the final dataframe to a CSV file for further use.

Data Wrangling

- **Perform exploratory Data analysis** and determine Training Labels

Calculate:

- Number of launches for each site.
- Number and of occurrence of orbit
- Number and occurrence of mission outcome of the orbits

Landing Outcome

- Landing did not always succeed
- **outcome converted** into 1 for a successful landing and 0 for an unsuccessful landing..

Landing Outcome Context:

- **True Ocean**: means the mission outcome was successfully landed to a specific region of the ocean.
- **False Ocean**: means the mission outcome was unsuccessfully landed to a specific region of the ocean.
- **True RTLS:** means the mission outcome successfully landed to a ground pad.
- **False RTLS**: means the mission outcome unsuccessfully landed to a ground pad.
- **True ASDS**: means the mission outcome successfully landed on a drone ship.
- **False ASDS**: means the mission outcome successfully landed on a drone ship.

EDA with Visualization

Charts:

- Flight Number compared to Launch Site
- Flight Number compared Payload Mass (kg)
- Payload Mass (kg) compared to Launch Site
- Paload Mass (kg) compared to Orbit Type

EDA with Visualization:

Analysis:

- Explore relationships by utilizing scatter plots. These variables could be beneficial for machine learning if a correlation is present.
- Display comparisons across different categories using bar charts. Bar charts highlight the relationships between categories and the corresponding measured values.

EDA with SQL

Queries:

Display:

- Names of the Unique Land Sites.
- 5 records where land sites begin with ‘CCA’.
- Total Payload Mass (kg) carried by boosters Launched by NASA.
- Average Payload Mass carried by booster F9 v1.1

Lists:

- Date when first successful landing outcome in ground pad achieved.
- Name of the boosters which had success in drone ship and payload mass greater than 4,000 but less than 6,000.
- Total number successful and failure mission outcomes.
- The names of booster versions which have carried the max payload.
- the records which will display the month names, failure landing outcomes booster versions and landing sites for the months in the year 2015.
- 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.

Map with Folium

All launch sites were marked, and notations to mark the success or failure of each launch indicated with:

- Markers
- Circles
- Lines

- Class 1 was assigned to success, and class 0 was assigned to failure
- Launch sites with high success rate were denoted with color-labeled marker clusters

Distance from launch sites to nearby entities were measured:

- Closest Coastline
- Closest Railway
- Closest Highway
- Closest City

Added **colored lines** to **show distance between** launch sites **CCAFS SLC - 40.**

Dashboard with Plotly Dash

Dropdown list with launch sites:

- Enables the user to chose either all launch sites or a specific launch site.

Create a pie chart for successful launch sites:

- Allows the user to visualize the successful vs the unsuccessful launch sites as percent of the total.

Create slider for Payload Mass Range:

- Allows user to check payload mass range.

Create a scatter chart to show the correlation between Payload and launch success:

- Allows the user to check the correlation between the payload and the launch success.

Predictive Analytics

- Generate: a NumPy array from the Class column.
- Normalize the data using StandardScaler. Fit and transform the dataset.
- Divide the data with train_test_split.
- Set up a GridSearchCV object with cv = 10 for parameter tuning.
- Implement GridSearchCV across different algorithms: Logistic Regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and K-Nearest Neighbor (KNeighborsClassifier()).
- Measure accuracy on the test set using .score() for each model.
- Evaluate the confusion matrix for each model.
- Determine the best model using Jaccard_Score, F1_Score, and Accuracy.

Results

Result Summary

The analysis reveals the **Payload Mass** and **Orbit Type** are crucial factors influencing success rate. Over time, the overall launch success has improved., with **KSC LC-39A** showing the highest success rate among landing sites. Notably, orbits like **ES-L1**, **GEO**, **HEO** and **SSO** boast a 100% success rate. Most launch sites are strategically located near the equator and close to the coast, far enough from populated areas to minimize risks in case of a failed launch, yet close enough to essential infrastructure to support launch activities.

Result Summary:

- All machine learning models performed similarly, with false positives being the primary issue in predictions.
- the **Decision Tree Model** emerged as the most effective predictive model for the dataset.

Visual Analytics:

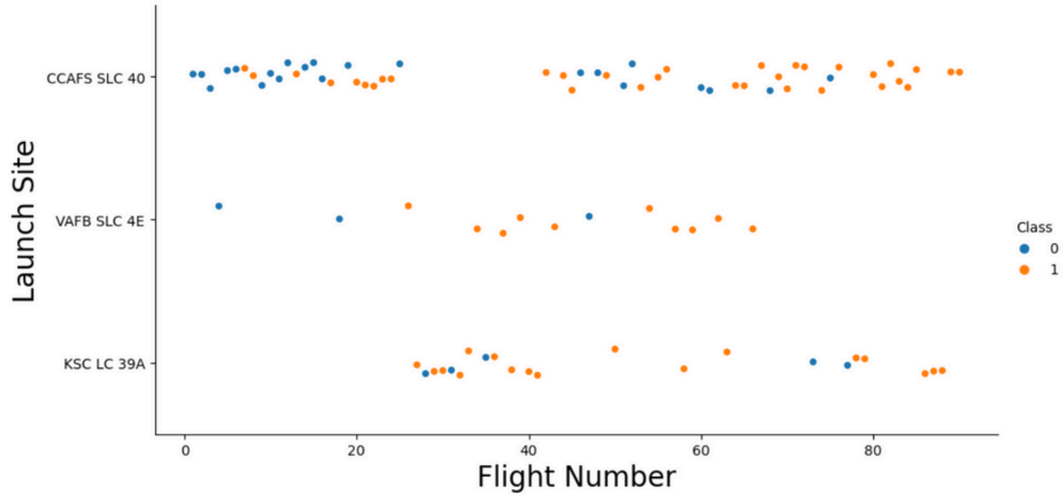
- Launch sites are positioned to balance safety and operational efficiency

Flight Number vs Launch Site

- **KSC LC 39A:** Highest success rate.
- **CCAFS SLC 40 & VAFB SLC 4E:** Mixed outcomes with some success & failures.
- **Flight Number Trend:** Success rate improves with higher flight numbers, especially at **KSC LC 39A** .

Conclusion:

- **KSC LC 39A** is the most reliable site.
- Success rate increases over time, indicating better technology or processes.



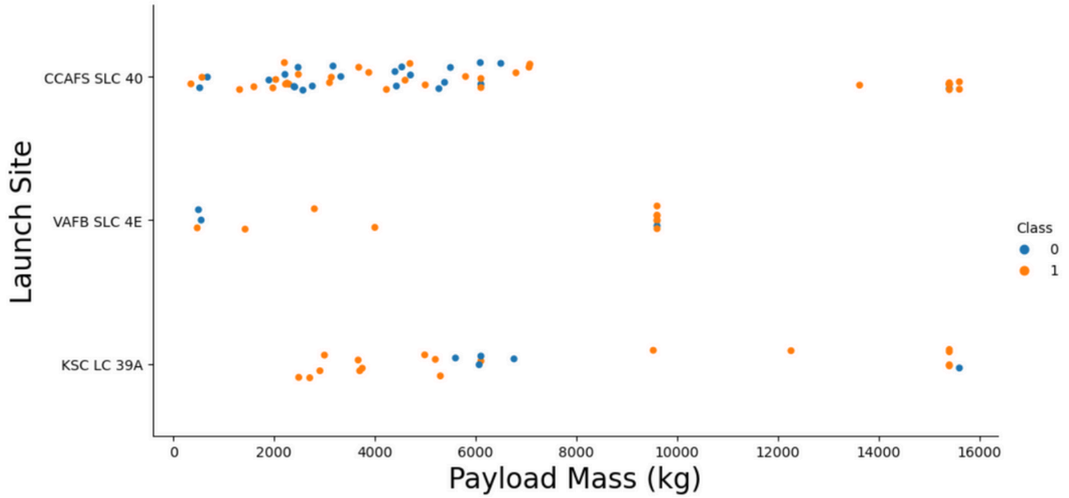
Payload vs Launch Site

Payload Mass Range:

- **CCAFS SLC 40**: Successful launches observed between **2,000** kg and **10,000** kg payloads, with consistent performance.
- **VAFB SLC 4E**: Limited launches, mostly under **4,000** kg, with mixed results.
- **KSC LC 39A**: Successful launches for payloads ranging from **2,000** kg up to **16,000** kg, with few failures.

Success Trend:

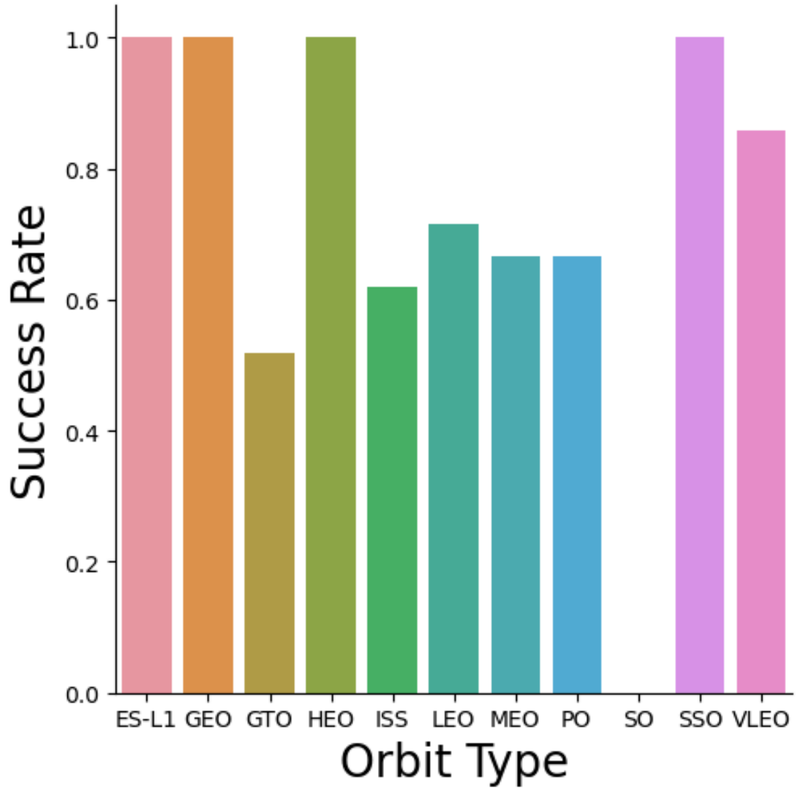
- Higher payload masses (over **10,000** kg) generally show more successes, particularly at **KSC LC 39A** and **CCAFS SLC 40**.



Success Rate of each Orbit

EDA:

- 100% Success Rate: **ES-L1**, **SSO**, **HEO** and **GEO.**
- 50%-90% Success Rate: **VLEO, PO, ISS, MEO,LEO AND GTO.**
- 0% Success Rate: **SO.**

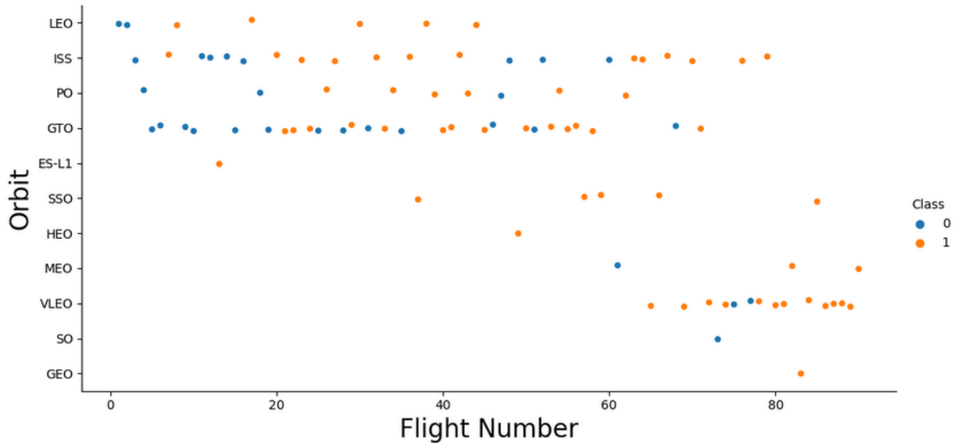


Flight Number vs Orbit

- High Success Orbits: **SSO**, **ES-L1**, **GEO** show near-perfect success rates.
- Mixed Outcomes: **LEO**, **GTO**, **ISS** have variable success rates across different flights.
- Trend: Higher flight numbers generally correlate with increased success.

Conclusion:

- **SSO**, **ES-L1**, **GEO** are the most reliable orbits.
- **LEO** and **GTO** show improvement in recent flights

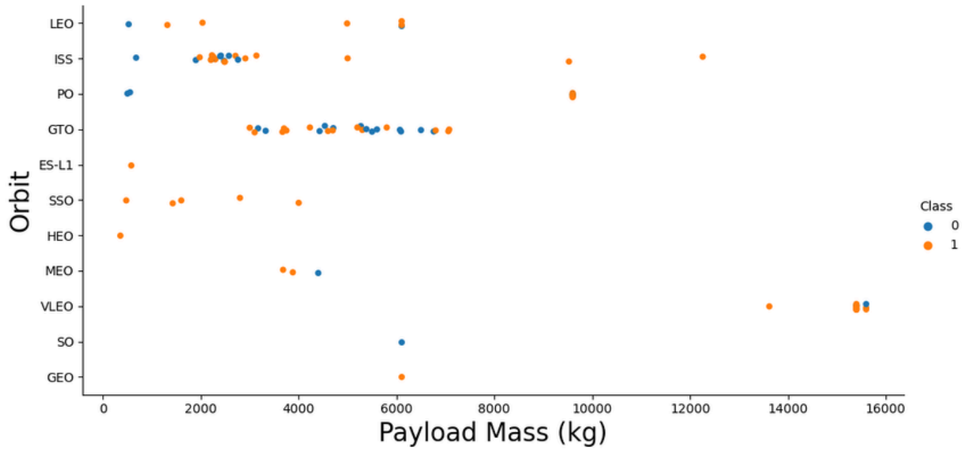


Orbits vs Payload

- High Success Orbits: **SSO**, **ES-L1**, **GEO** maintain high success rates across various payloads.
- Mixed Success: **LEO**, **GTO**, **ISS** show variable outcomes, especially with payloads under **6,000** kg.
- Large Payloads: High payloads (above **10,000** kg) tend to be successful, particularly in **GTO** and **GEO** orbits.

Conclusion:

- **SSO**, **ES-L1**, **GEO** are consistently reliable across all payload masses.
- Larger payloads generally succeed, especially in higher orbits like **GTO** and **GEO**.

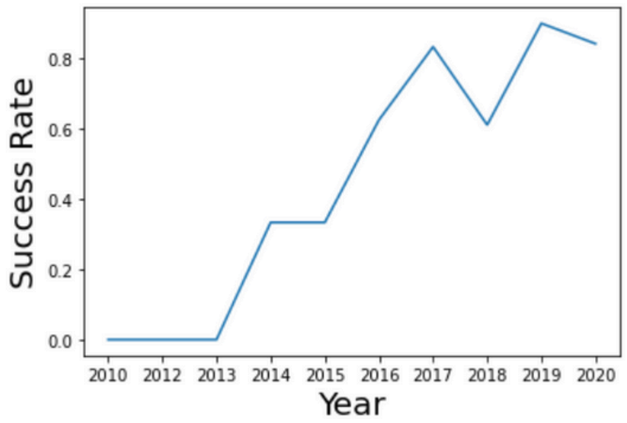


Launch Success over time

- Initial Growth: Success rate remained low until **2014**, after which it started increasing.
- Steady Improvement: Noticeable improvement from **2015** to **2018**, with the success rate peaking in **2019**.
- Recent Trends: Slight decline in success rate after **2019**, but still remains high.

Conclusion:

- The success rate has significantly improved over the years, especially from **2014** onward, showing the impact of refined processes and technologies.



Launch Site Information

Landing Outcome Content

```
%sql select distinct launch_site from SPACEXTBL;

[22] ✓ 0.0s

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

...
```

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

Launch Site Names

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

Records with Launch Site Starting with CCA

- Displaying 5 records below

```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

[12]                                                                    Python

... * 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

Total Payload Mass

Total Payload Mass:

- 45596 Kgs total were carried by boosters launched by NASA

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

[12] ✓ 0.0s

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass:

- 2928.4 Kgs average were carried by boosters launched by NASA

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

[13] ✓ 0.0s

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

AVG(PAYLOAD_MASS__KG_)
2928.4

Landing & Mission Info

First successful landing in Ground Pad:

- 2015-12-22

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

[14] ✓ 0.0s

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

MIN(DATE)
2015-12-22

Booster Drone ship Landing

- Booster mass greater than 4,000 but less than 6,000
- JSCAT-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;
```

[23] ✓ 0.0s

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

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

Total Number of Successful and Failed Mission Outcomes:

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_MISSION_1_0 \
    FROM SPACEXTBL \
    GROUP BY MISSION_OUTCOME;
```

[16] ✓ 0.0s

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

Mission_Outcome	TOTAL_MISSION_1_0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters

Booster Versions with Maximum Payload Mass:

- 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

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

[17] ✓ 0.0s

```
 * 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

Failed Landings on Drone Ships

- Showing month, date, Booster versions, launch site and landing outcome:

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [LANDING_OUTCOME] \
    FROM SPACEXTBL \
    WHERE [LANDING_OUTCOME] = 'Failure (drone ship)' and substr(Date,0,5) = '2015';
```

[18] ✓ 0.0s

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

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

Count of successful Landings

- Count of landing outcomes from 2010-06-04 and 2017-03-20 in descending order:

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

✓ 0.0s

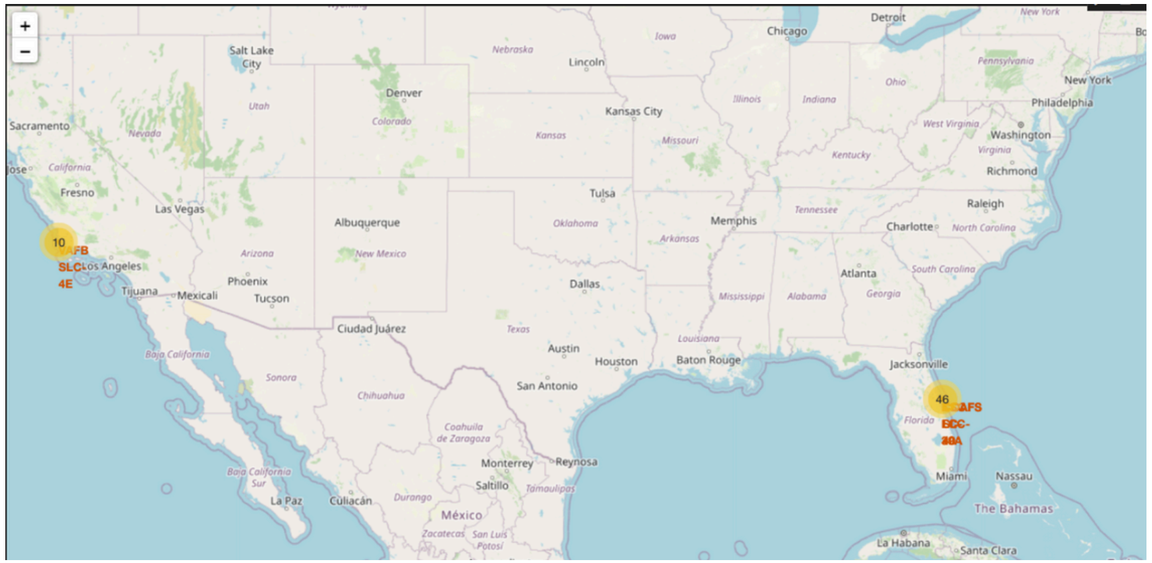
* sqlite:///my_data1.db
Done.

Landing_Outcome	COUNT_OUTCOMES
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

Launch Site Analysis

Launch Site

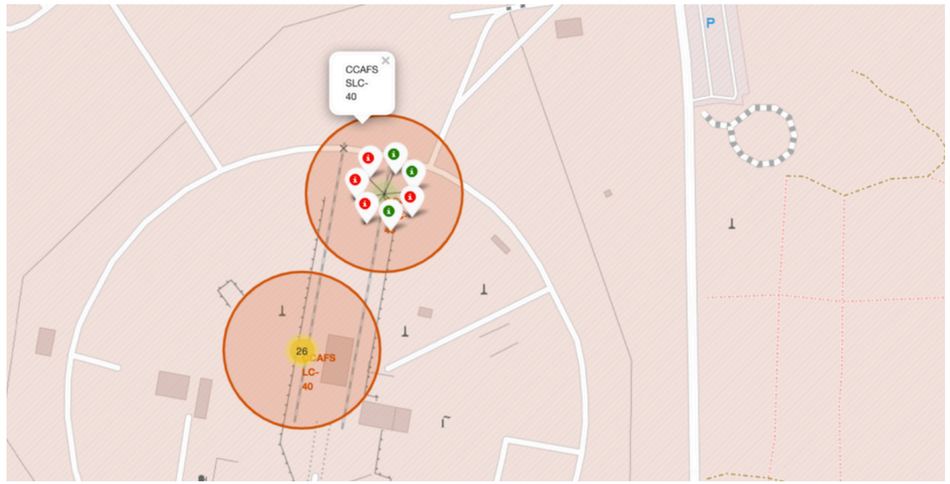
- Launch sites located **near the equator** benefit significantly when it comes to launching rockets into equatorial orbits. The closer a site is to the equator, the more it can take advantage of Earth's rotational speed, providing a natural boost for prograde orbits. This advantage reduces the need for **additional fuel** and boosters, making launches more efficient and **cost-effective**



Launch Outcomes

Each Launch Site:

- Outcomes:
- Green markers for successful landings
- Red Markers for unsuccessful landings
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Distance to Proximities

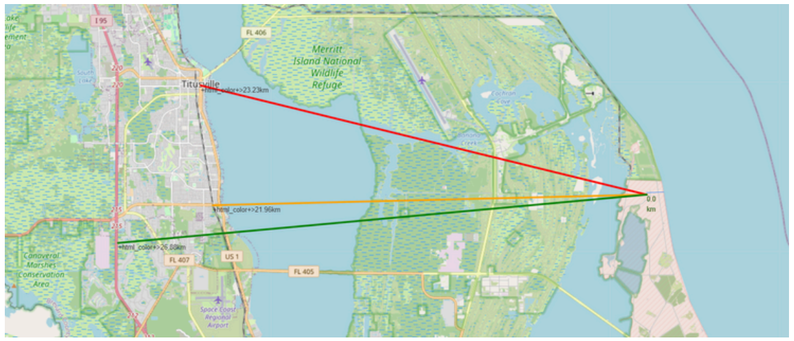
- Coasts: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.

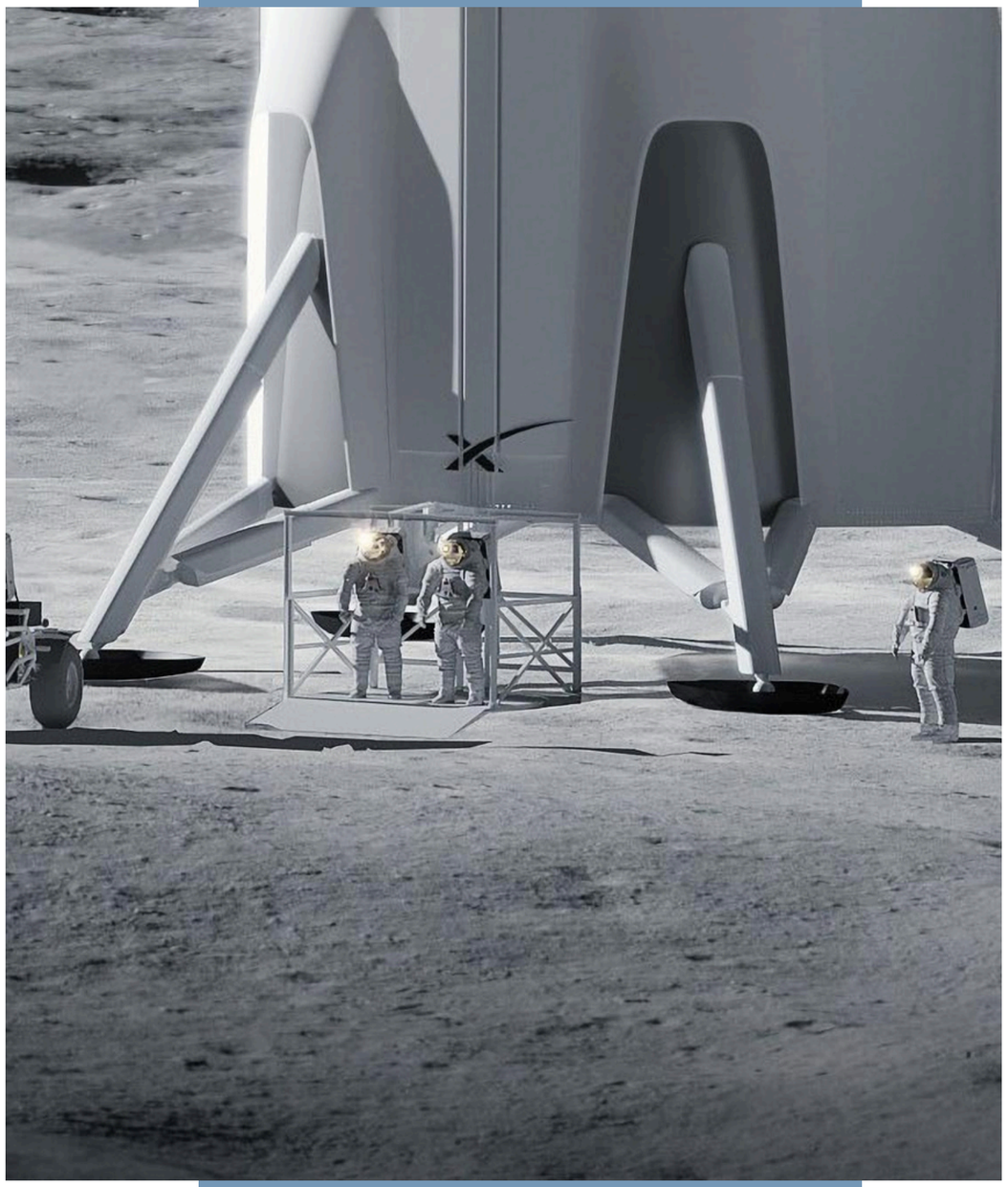
- Safety / Security: needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.

- Transportation/Infrastructure and Cities: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.

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

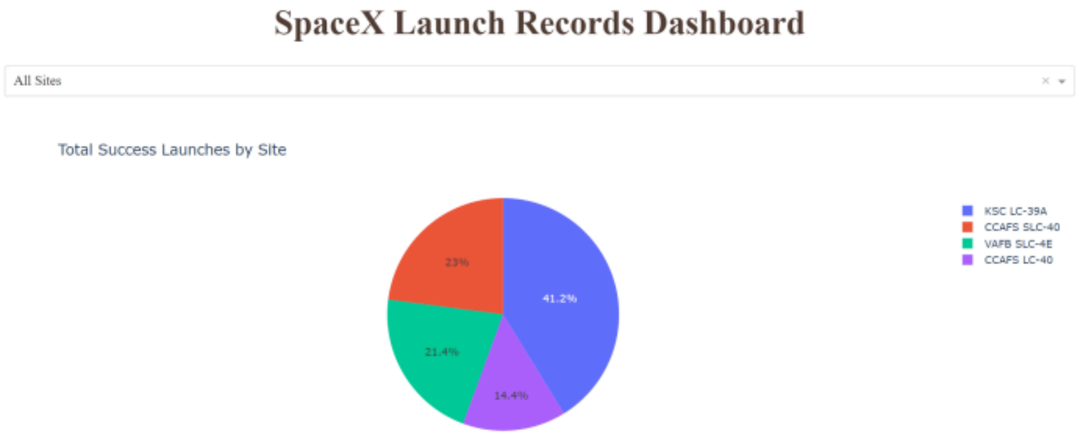





Dashboard
with plotly

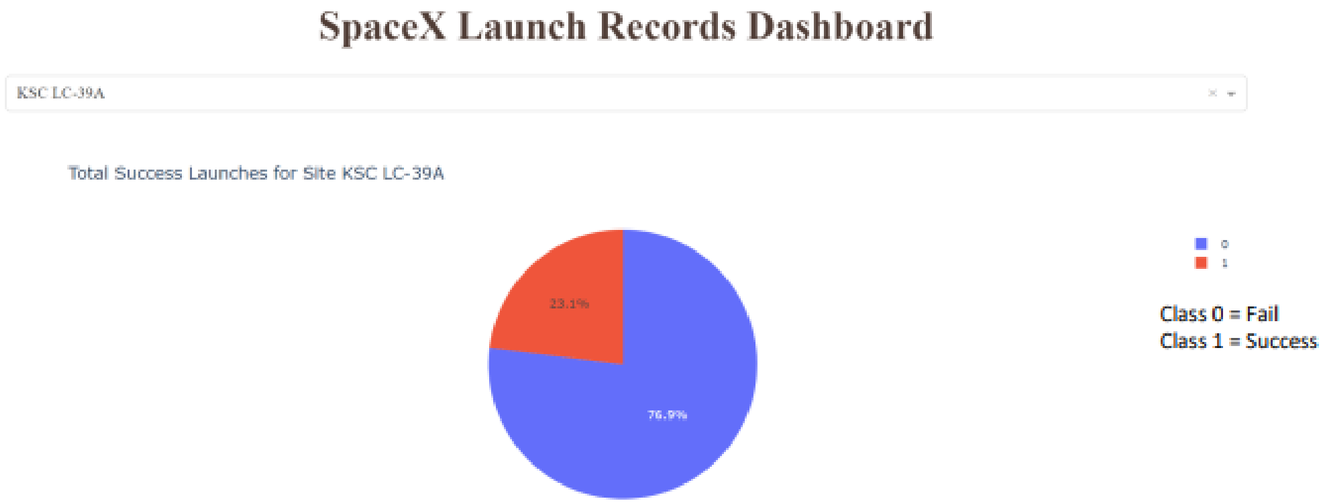
Launch Success by Site

- **KSC LC-39A** has the most successful launches amongst launch sites



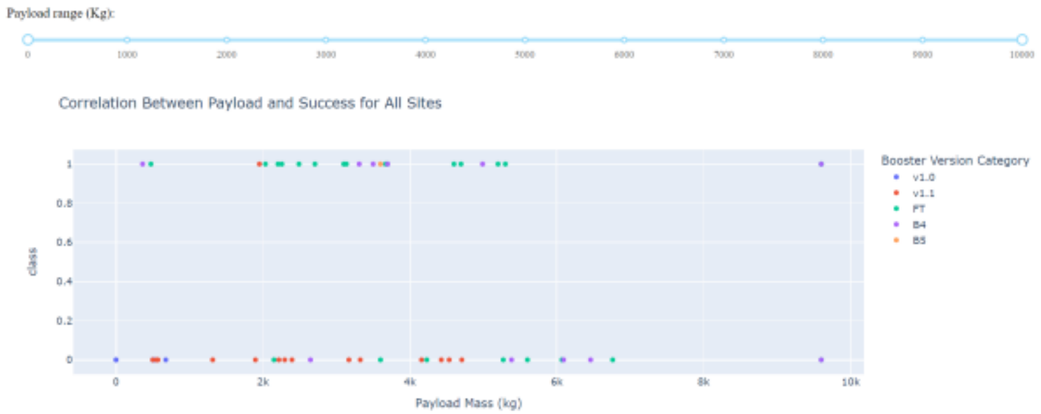
Launch Success (KSC LC-29A)

- **KSC LC-39A** has the highest success rate amongst launch sites **(76.9%)**
- **10** successful launches and **3** failed launches



Payload Mass and Success

- Payloads between **2,000 kg** and **5,000 kg** have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome



Predictive Analytics

Classification

Accuracy

- All the models delivered comparable performance, achieving similar scores and accuracy levels, which can likely be attributed to the limited size of the dataset. However, the Decision Tree model stood out slightly, with a marginally better . **best_score_.**
- This **.best_score_** represents the average score across all cross-validation folds for a specific set of parameters.

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

```
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_)
```

[33]

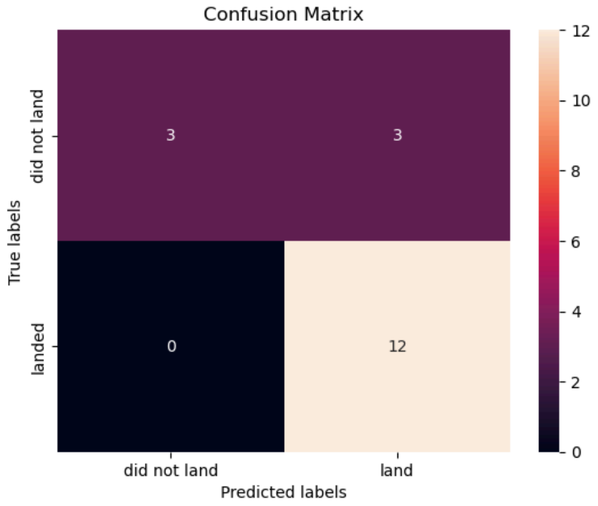
```
Best model is DecisionTree with a score of 0.8892857142857145
Best params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'}
```

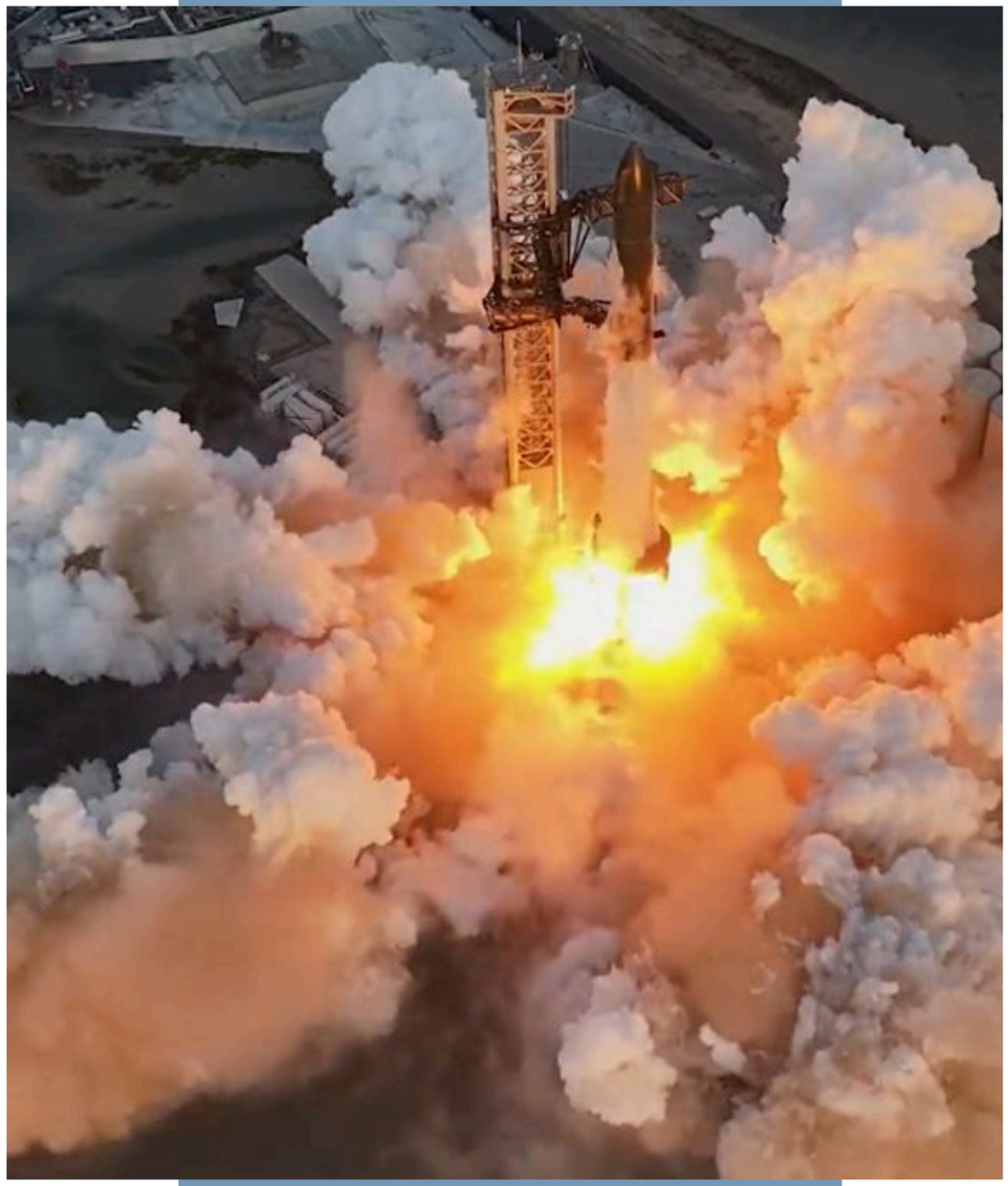
Confusion Matrices

- **True Labels (Y-axis)**: These are the actual outcomes:
  - "did not land": The actual class where the rocket did not land successfully.
  - "landed": The actual class where the rocket landed successfully.
- **Predicted Labels (X-axis)**: These are the predicted outcomes from the model:
  - "did not land": The model predicted that the rocket did not land.
  - "landed": The model predicted that the rocket landed.

Interpretation of the Matrix:

1. **Top-left (3)**:
   - True Negatives (TN): 3 instances where the model correctly predicted that the rocket would not land (it actually didn't land).
2. **Top-right (3)**:
   - False Positives (FP): 3 instances where the model incorrectly predicted that the rocket would land, but it actually did not land.
3. **Bottom-left (0)**:
   - False Negatives (FN): 0 instances where the model incorrectly predicted that the rocket would not land, but it actually did land.
4. **Bottom-right (12)**:
   - True Positives (TP): 12 instances where the model correctly predicted that the rocket would land (it actually landed).





Conclusion

Research Insights:

- **Model Performance:** The models yielded comparable results on the test set, with the **Decision Tree model** showing a slight edge.
- **Equator Advantage:** The majority of launch sites are strategically located near the equator to leverage Earth's rotational speed, providing a natural boost that reduces the need for extra fuel and boosters.
- **Coastal Proximity:** All launch sites are positioned close to the coast.
- **Launch Success:** Success rates have improved over time.
- **KSC LC-39A:** This site boasts the highest success rate among all launch locations, with a flawless record for launches carrying less than **5,500 kg**.
- **Orbits:** ES-L1, GEO, HEO, and SSO orbits have achieved a 100% success rate.
- **Payload Mass:** [Incomplete point; please provide additional context for completion.]
-

Things to Consider

- **Dataset:** Expanding the dataset could enhance the robustness of predictive analytics and determine if the findings are applicable to a broader dataset.
- **Feature Analysis / PCA:** Conducting further feature analysis or implementing principal component analysis (PCA) could potentially improve model accuracy.
- **XGBoost:** Although XGBoost is a highly effective model, it was not utilized in this study. Exploring its performance in comparison to other classification models could yield interesting insights.