IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Han, Sijia
July 26, 2024
[GitHub]

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collections
  - Data Wrangling
  - Exploratory Data Analysis
  - Interactive Visual Analytics
  - Machine Learning Prediction Analysis
- Summary of all results
  - 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.
  - Most launch sites are near the equator, and all are close to the coast.
  - All models performed similarly on the test set. The decision tree model slightly outperformed.

Introduction

- Project background and context
  - SpaceX, a trailblazer in the space industry, is committed to making space travel accessible and affordable for everyone. Its remarkable achievements include sending spacecraft to the International Space Station, launching a satellite constellation to provide global internet access, and conducting manned missions to space. SpaceX's success largely stems from its innovative reuse of the first stage of its Falcon 9 rocket, which significantly reduces launch costs to approximately $62 million per launch. In contrast, other providers who do not have the capability to reuse the first stage face costs upwards of $165 million per launch. By determining the likelihood of a successful first-stage landing, we can estimate the launch price. Public data and machine learning models can be utilized to predict whether SpaceX or a competing company can reuse the first stage, thereby impacting overall costs.
- Problems we want to find answers
  - How payload mass, launch site, number of flights, and orbits affect first-stage landing success
  - Rate of successful landings over time
  - Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX REST API and web scraping techniques
- Perform data wrangling
  - Filtering the data, handling missing values and applying one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predict landing outcomes using classification models.
  - Tune and evaluate models to find the best model and parameters.

Data Collection – SpaceX REST API

- Using the SpaceX REST API to retrieve data about launches, including information about the rocket used, payload, launch specifications, landing specifications and landing outcomes.
- Steps
  - Request the SpaceX launch data using the GET request.
  - Decode response content as a JSON and convert it to a Pandas dataframe.
  - Use the API to request information about launches and retrieve data on rockets, payloads, launchpads, and cores.
  - Create a Pandas dataframe to store all the data.

Data Collection – Web Scraping

- Web scrapping to collect Falcon 9 historical launch records from a Wikipedia page titled *List of Falcon 9 and Falcon Heavy launches.*
- Steps
  - Request Falcon 9 launch Wiki page from its URL using the HTTP GET method.
  - Create a BeautifulSoup object from the HTML response.
  - Extract all column/variable names from the HTML table header.
  - Create a Pandas dataframe by parsing the launch HTML tables.

Data Wrangling

- Data obtained from APIs or web scraping is often imperfect and requires thorough cleaning before further exploration. Additionally, performing an initial data analysis is crucial to gain insight into the gathered data, allowing us to better understand the variables and their relationships.
- Steps
  - Dealing with the missing values – using mean() to replace them
  - Get the insight into the variables using value_counts() to
    - Calculate the number of launches on each site
    - Calculate the number and occurrence of each orbit
    - Calculate the number and occurrence of mission outcomes of the orbits
  - Create a binary landing outcome label from the Outcome column
    - 0 means the first stage did not land successfully
    - 1 means the first stage landed successfully

EDA with SQL

- To gain deeper insights into the data, we use SQL queries to calculate and display key metrics and relevant information.
- The info includes:
  1. Names of the unique launch sites in the space mission.
  2. 5 records where the launch sites begin with 'CCA'
  3. Total payload mass carried by boosters launched by NASA (CRS).
  4. Average payload mass carried by booster version F9 v1.1.
  5. The date of the first successful landing outcome in the ground pad was achieved.
  6. Names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000.
  7. Total number of successful and failed mission outcomes.
  8. Names of the booster versions which have carried the maximum payload mass.
  9. Records which will display the month, failure landing outcomes in drone ships, booster versions, and launch sites for the months in the year 2015.
  10. Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

EDA with Data Visualization

- Scatter Charts
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Flight Number vs. Orbit type
  - Payload Mass vs. Launch Site
  - Payload Mass vs. Orbit type
- Bar Charts
  - Success rate of each orbit type
- Line Charts
  - Launch success yearly trend

Build an Interactive Map with Folium

- Mark all launch sites on a map
  - Added blue circles at **NASA Johnson Space Center's coordinates** with a popup label showing its name using its latitude and longitude coordinates.
  - Added red circles at **all launch sites' coordinates**, with a popup label showing the site's name using its latitude and longitude coordinates.
- Mark the success/failed launches for each site on the map
  - Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- Calculate the distances between a launch site and its proximities
  - Added colored lines to show the distance between launch site **CCAFS SLC – 40** and its proximity to the **nearest coastline, railway, highway and city**.

Build a Dashboard with Plotly Dash

- Dropdown **list with launch sites**
  - Allow users to select all launch sites or a certain launch site
- Pie chart **showing successful launches**
  - Allow users to view successful and unsuccessful launches as a percentage
- Slider **of payload mass range**
  - Allow users to select payload mass range
- Scatter chart **showing payload mass vs success rate by booster version**
  - Allow users to view the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- After reviewing the data, we will develop a machine-learning pipeline to predict whether the first stage will land, utilizing the data from the previous results.
- Steps
  - Create a Numpy array from the Class column in the data.
  - Standardize the data with StandardScaler.
  - Fit and transform the data.
  - Split the data into training data and test data.
  - Create a GridSearchCV object with cv = 10 for parameter optimization.
  - Apply GridSearchCV on different algorithms: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbor.
  - Find the best hyperparameters for all models.
  - Calculate the accuracy on the test data for all models.
  - Assess the confusion matrix for all models.
  - Find the method that performs best.

Results

- 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 analysis
  - 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 analysis
  - The Support Vector Machine, Logistic Regression, and K-Nearest Neighbor all achieved the same accuracy score of 83.33%, outperforming the Decision Tree.

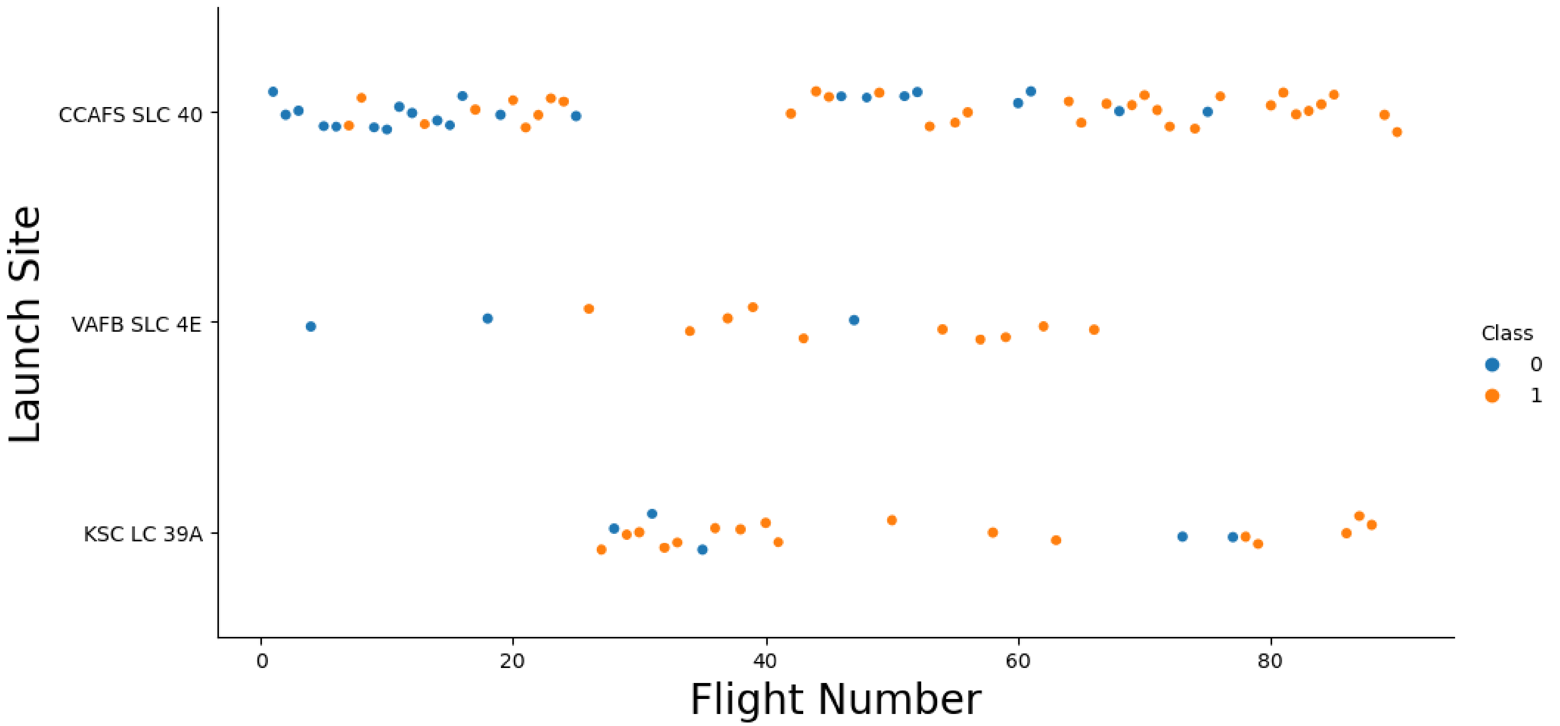
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Insights from the plot

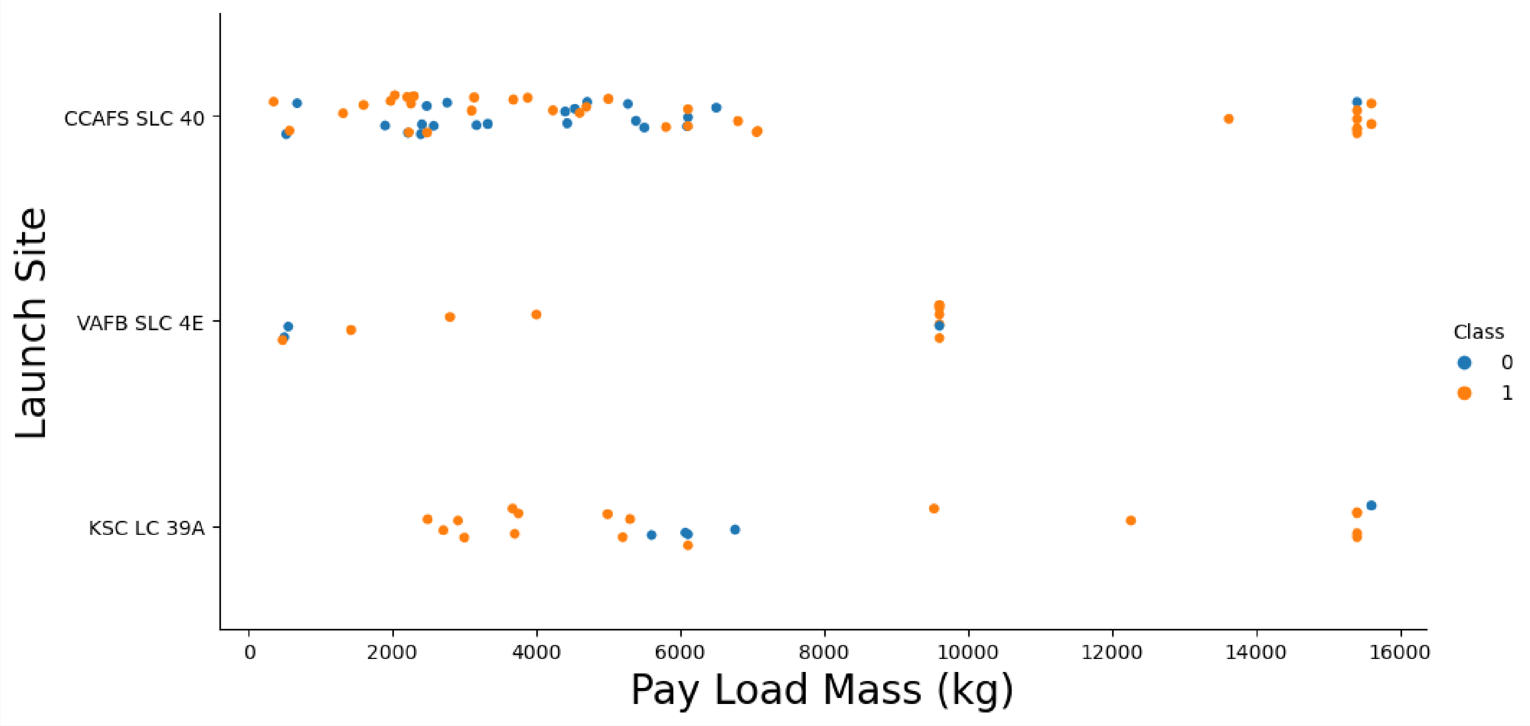
- Earlier flights had a lower success rate (blue = fail).
- Later flights had a higher success rate (orange = success)
- CCAFS SLC 40 launch site has the most launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.



Payload vs. Launch Site

Insights from the plot

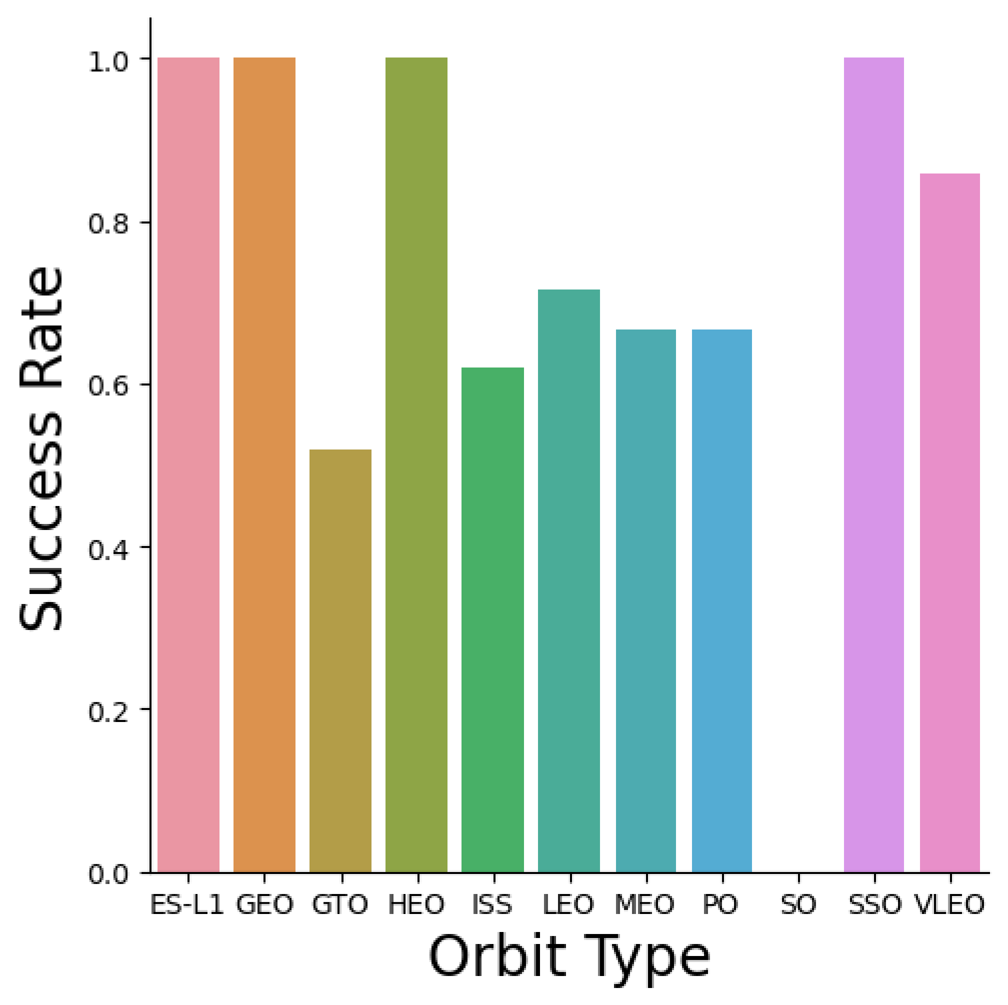
- The heavier payload mass of launches seems to have a higher success rate.
- KSC LC 39A has a 100% success rate for launches less than 5500 kg.
- There is no clear correlation between payload mass and success rate.



Success Rate vs. Orbit Type

Insights from the plot

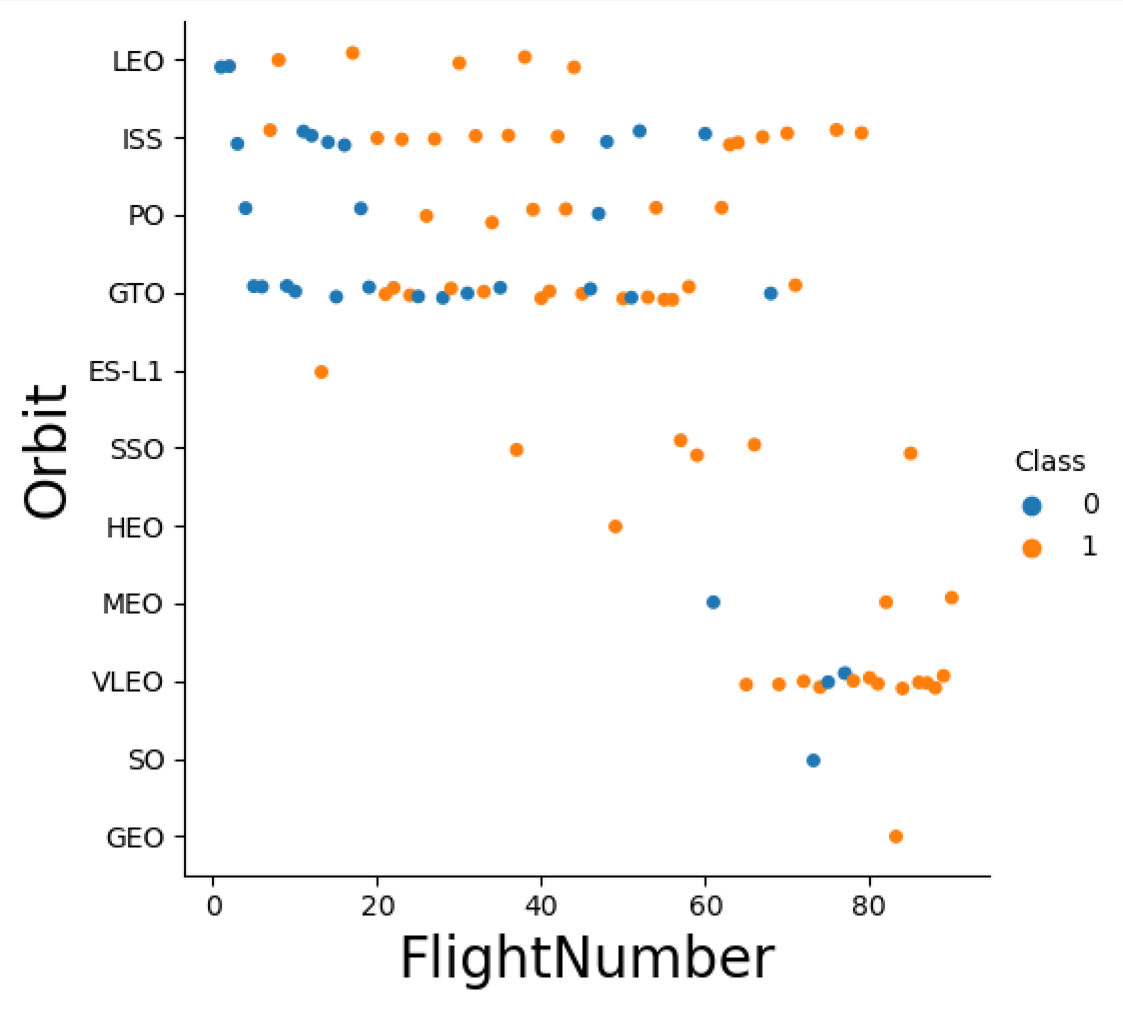
- 100% success rate
    - ES-L1
    - GEO
    - HEO
    - SSO
- 0% Success Rate
    - SO



Flight Number vs. Orbit Type

Insights from the plot

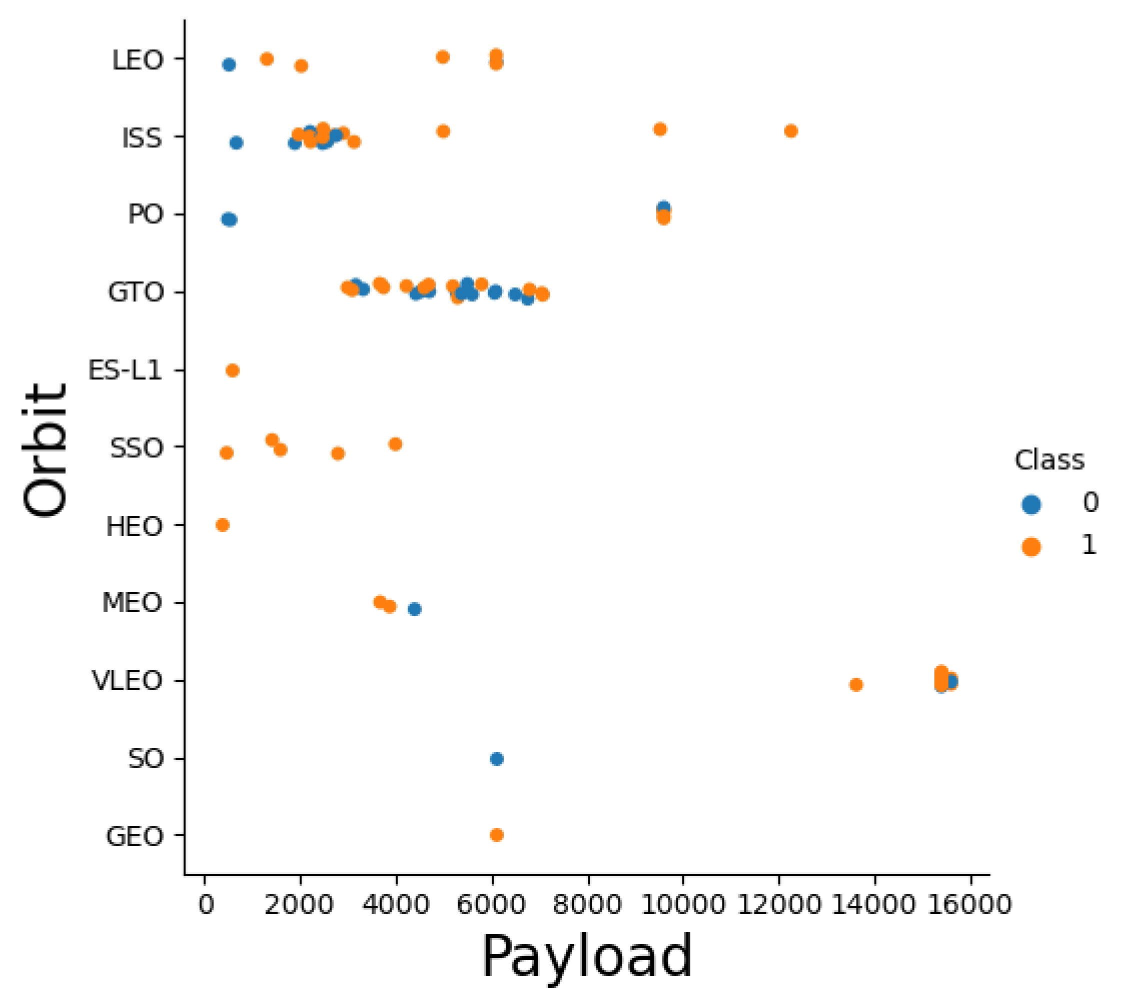
- The high success rate of the GEO, HEO, and ES-L1 orbits can be explained by only one flight into the respective orbits.
- Generally, the success rate increases with the number of flights for each orbit.



Payload vs. Orbit Type

Insights from the plot

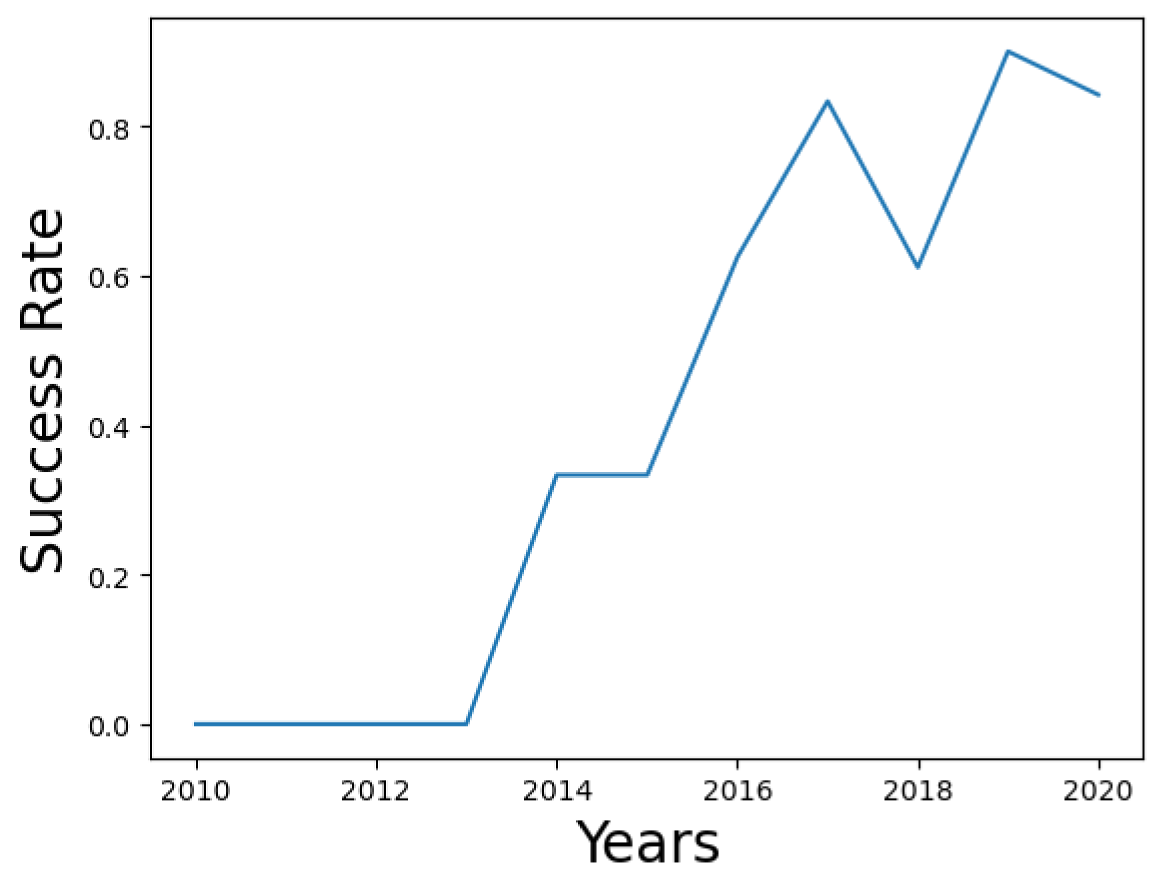
- Heavy payloads are better with LEO, ISS and PO orbits.
- The GTO orbit has an unclear relationship between payload mass and success rate.



Launch Success Yearly Trend

Insights from the plot

- No landings were successful from 2010 to 2013.
- The success rate improved from 2013 to 2017 and 2018 to 2019.
- The success rate decreased from 2017 to 2018 and 2019 to 2020.
- The success rate has improved overall since 2013.



All Launch Site Names

```
%%sql

SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

```
%%sql

SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

AVERAGE_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

```
%%sql

SELECT MIN(DATE) AS DATE_OF_FIRST_SUCCESSFUL_LANDING_OUTCOME
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

DATE_OF_FIRST_SUCCESSFUL_LANDING_OUTCOME
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql

SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql

SELECT
    'Success' AS Outcome,
    SUM(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 ELSE 0 END) AS Count
FROM
    SPACEXTABLE
GROUP BY
    Outcome

UNION

SELECT
    'Failure' AS Outcome,
    SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) AS Count
FROM
    SPACEXTABLE
GROUP BY
    Outcome;
```

Outcome	Count
Failure	1
Success	100

Boosters Carried Maximum Payload

```
%%sql

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

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

```
%%sql

SELECT SUBSTR(DATE, 6, 2) as Month, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Fail%' AND SUBSTR(DATE, 0, 5) = '2015'
```

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql

SELECT LANDING_OUTCOME, COUNT(*) AS QTY
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY QTY DESC;
```

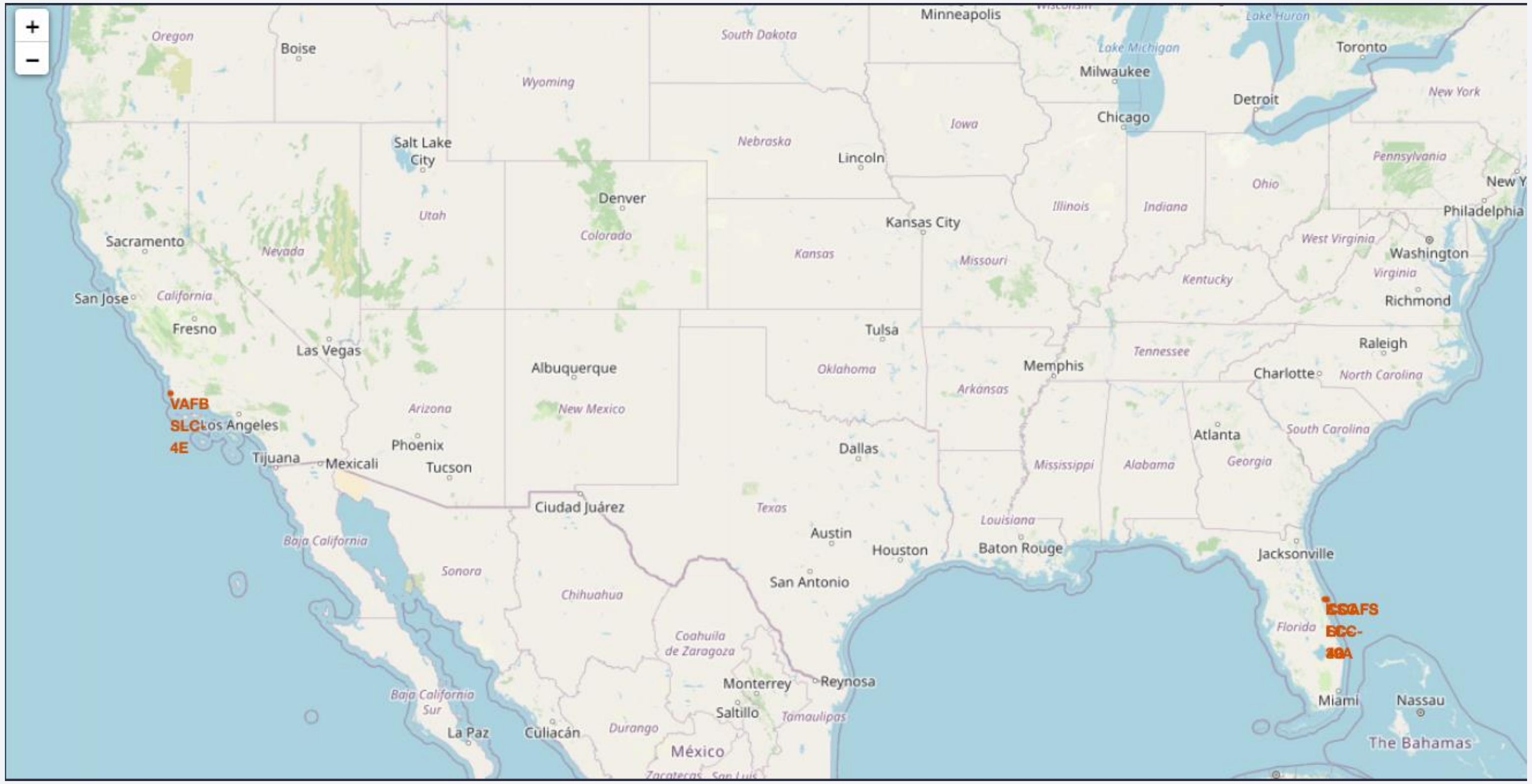
Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

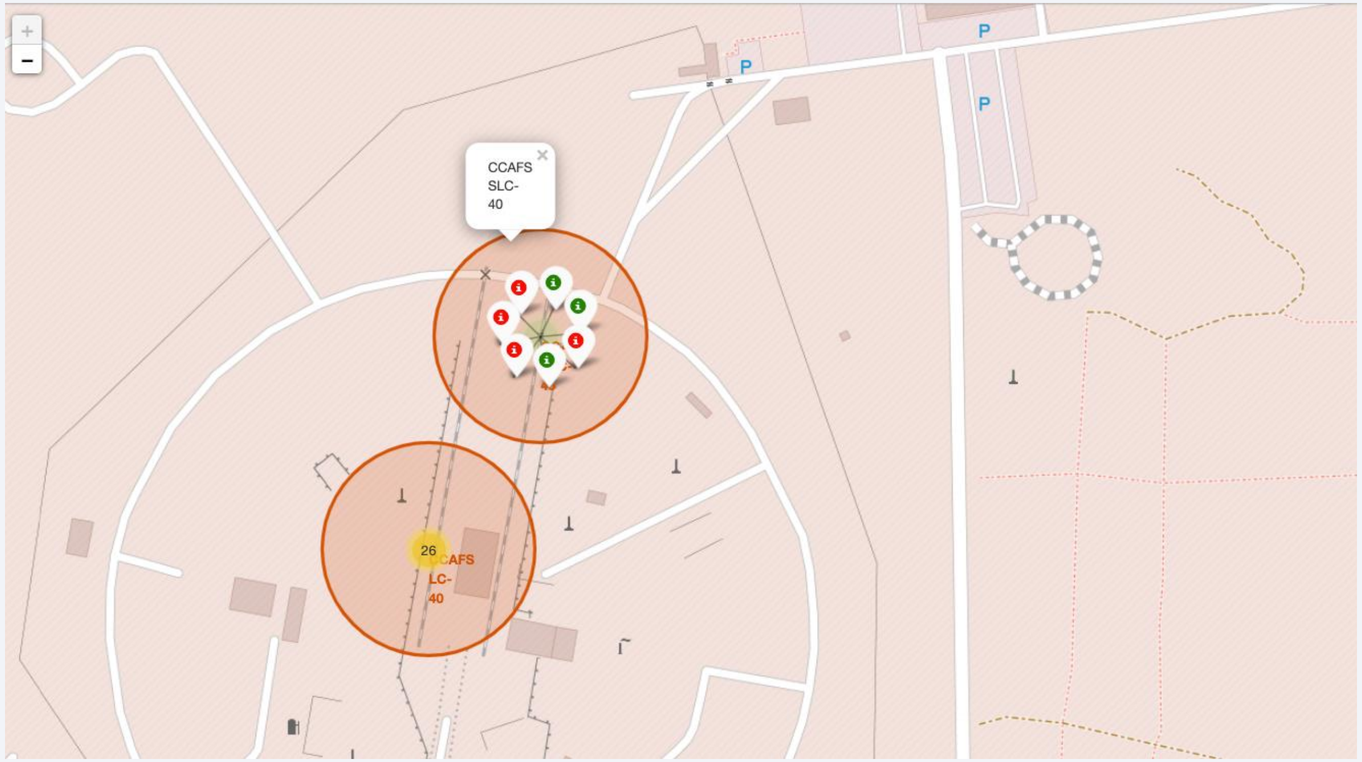
Launch Sites

- All SpaceX launch sites are on the coasts of the United States of America.



Launch Outcomes

- Green markers for successful launches
- Red markers for failed launches
- Launch site CCAFS SLC-40 has a 42.9% success rate.



Distance to Proximities

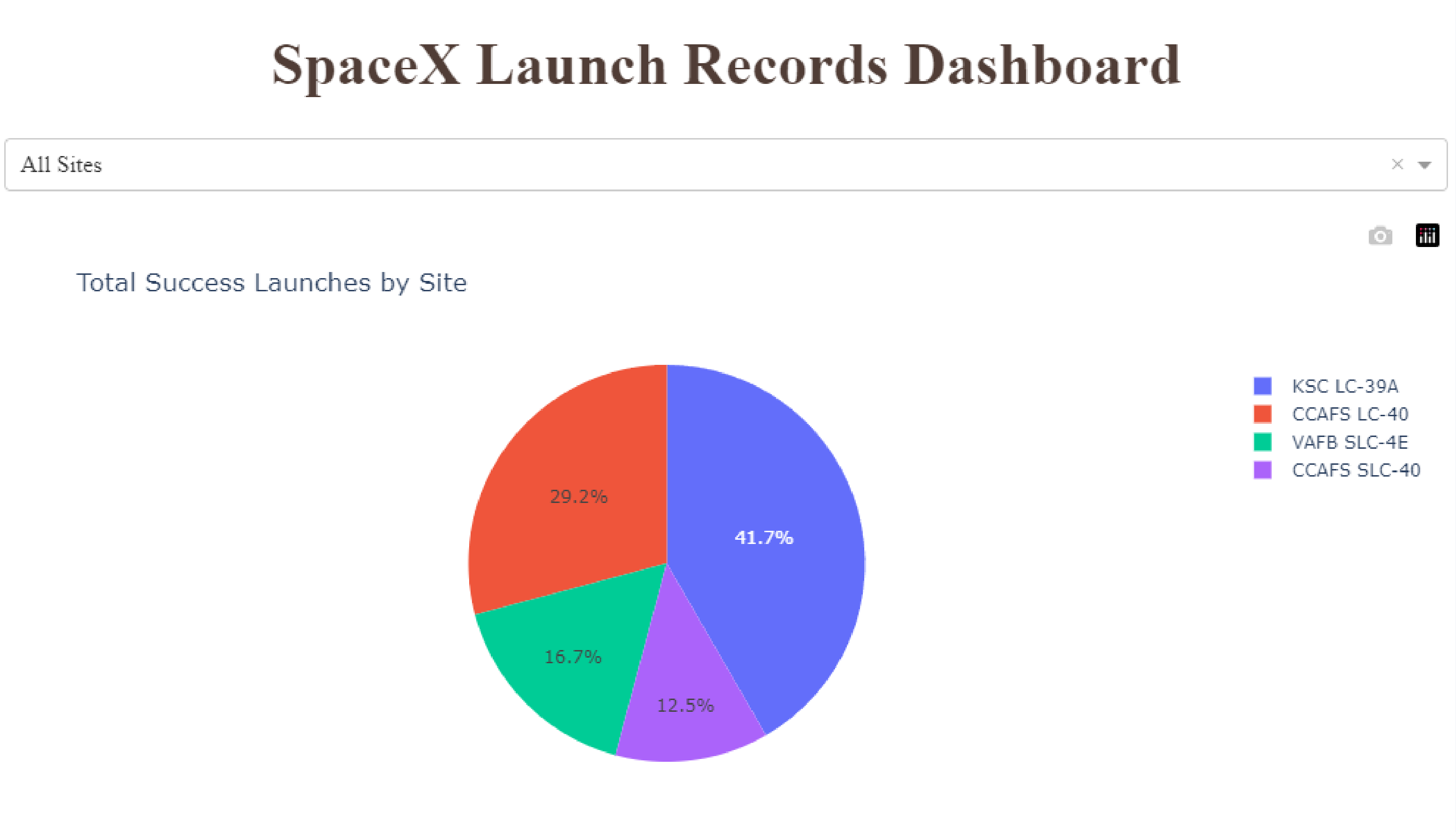
For CCAFS SLC-40

- 0.90 km from the nearest coastline
- 21.96 km from the nearest railway
- 23.23 km from the nearest city
- 0.59 km from the nearest highway

Section 4

Build a Dashboard with Plotly Dash

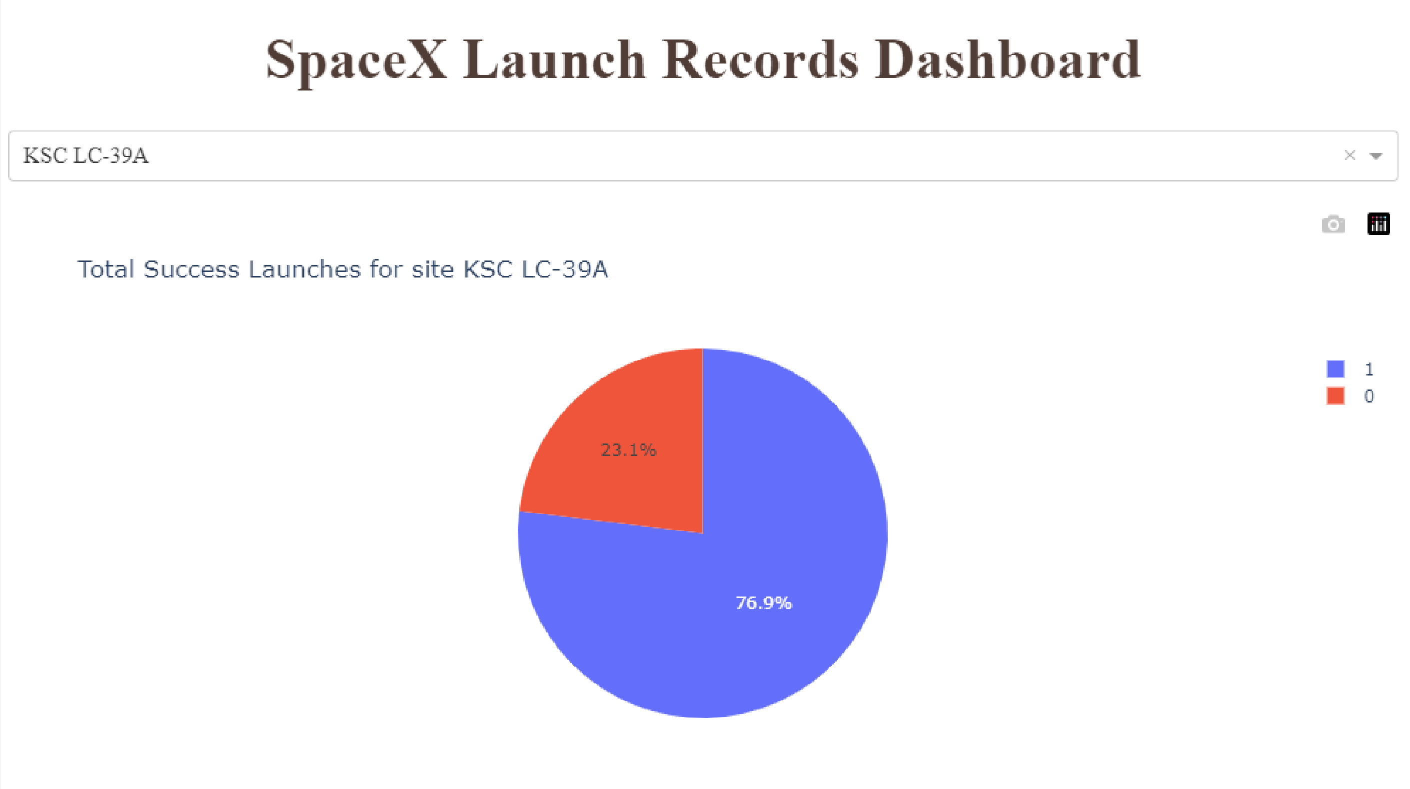
Launch Success by Site



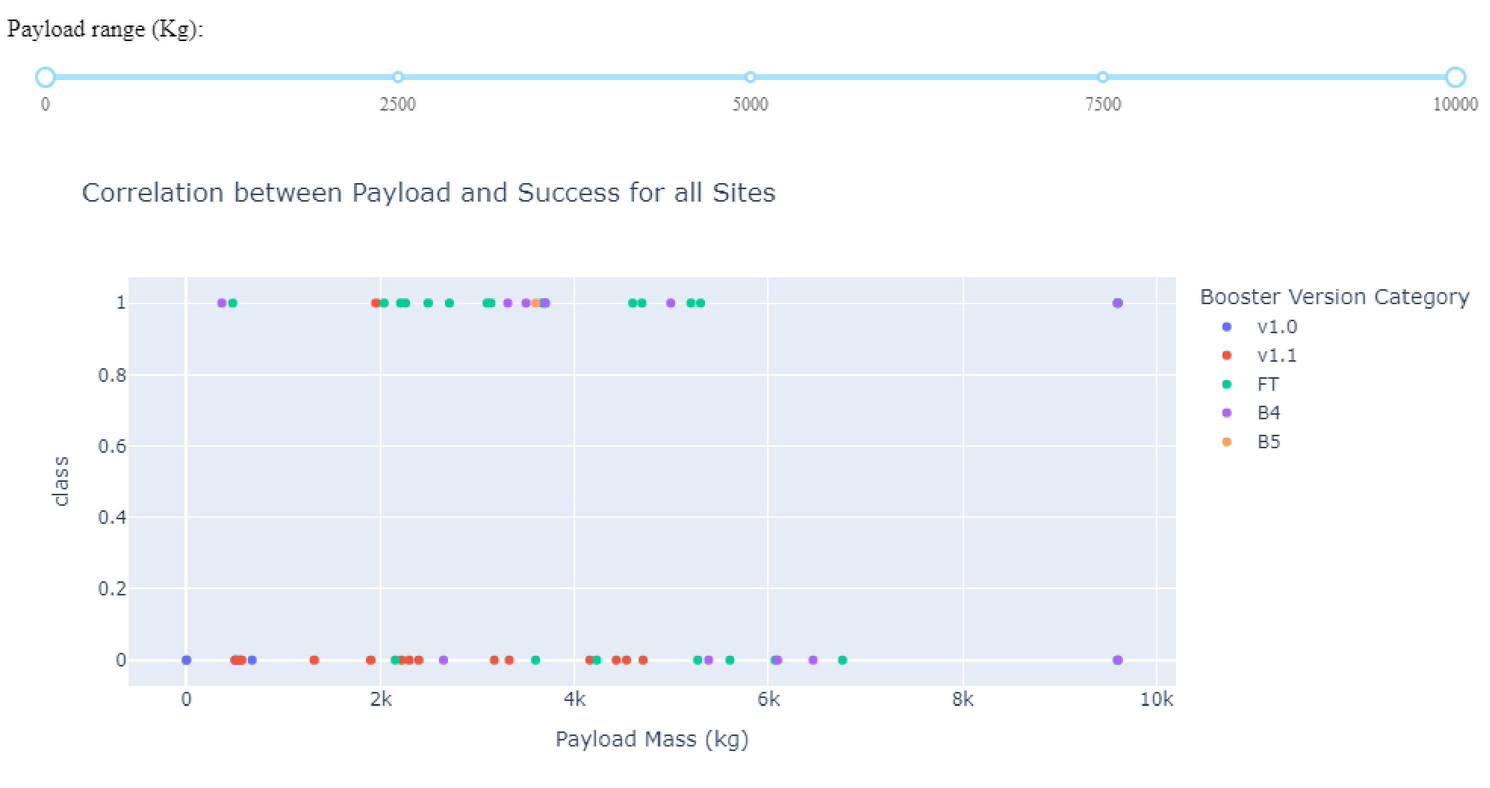
The launch site KSC LC-39A had the most successful launches, with 41.7% of the total successful launches.

Launch Success for KSC LC-39A

- KSC LC-39A has the highest success rate of successful launches, with a 76.9% success rate.



Payload VS. Launch Outcome



- 1 means success.
- 0 means failed.
- Payloads between 2000 kg and 5000 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

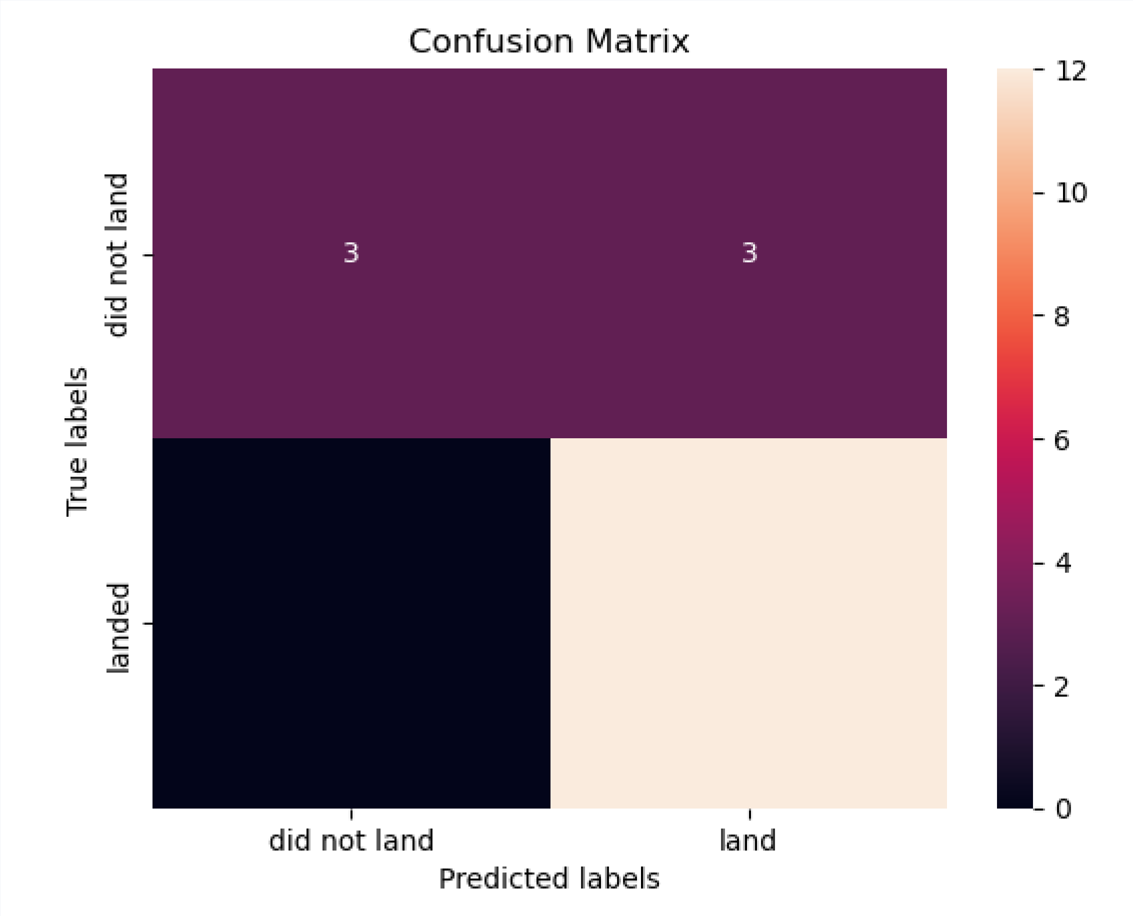
Classification Accuracy

- The SVM, Logistic Regression, and KNN models performed at the same level and had the same accuracy score of 83.33%.
- The Decision Tree model was outperformed and only achieved a score of 77.78%.

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	77.777778

Confusion Matrix

- Confusion Matrix Outputs
  - 12 True Positive
  - 3 True Negative
  - 3 False Positive
  - 0 False Negative



Conclusions

- The SVM, Logistic Regression, and KNN models performed similarly on the test data, while the Decision Tree model performed slightly less.
- All the launch sites are close to the coast.
- Launch success rate increases over time.
- KSC LC-39A has the highest success rate among launch sites. It has a 100% success rate for launches less than 5,500 kg.
- Across all launch sites, the higher the payload mass (kg), the higher the success rate.


Thank you!