IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muktar Arubayi
arubayimuktar@gmail.com

July 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Data-Driven Insights for Space Y's Commercial Space Ventures

- Critical factors influencing launch success and reliability analyzed to aid Space Y's strategic decisions in entering the commercial space industry.
- Data sourced from SpaceX API and Wikipedia, cleaned for effective analysis and explored using SQL and visualization.
- Interactive visual analysis was performed with Folium and Plotly Dash, while predictive analysis utilized various classification methods.

Key Findings:

- SpaceX Success: Consistent upward trend in launch success since 2013, with a slight dip in 2018.
- Optimal Payload: 2000-4000 kg payload yields the highest launch success rates.
- Orbit Success: ES-L1, GEO, HEO, and SSO exhibit the most successful outcomes.
- Promising Sites: KSC LC-39A boasts a remarkable 76.9% launch success rate.
- Reliable Predictions: The decision three model offers 90% accuracy in launch outcome predictions.

By harnessing data-driven insights, Space Y can navigate the competitive landscape, optimize launches, and achieve success in the commercial space industry.

Thank you for considering our data analysis. Let's propel Space Y's ventures to new heights with the power of data-driven decisions!

Introduction

Commercial Space: Exploring the Potential of Space Y's Ventures

- Commercial Space Industry: Fast-growing with Space X as a top competitor.
- Space X's Competitive Edge: Utilizing Falcon 9 rocket launcher with first-stage reusability.
- Cost Advantage: Reducing space travel cost to $62 million (others at $165 million) through successful first-stage landings.

Problem Statement:

- Assessing the feasibility of Space Y's entry into the space industry.
- Understanding the impact of first-stage landing success on launch reliability.
- Leveraging data science to make informed business choices.

Key Questions for Analysis:

- Factors Influencing Reliability: What factors significantly influence the reliability of first-stage landings?
- Data-Driven Insights: How can data-driven insights shape Space Y's strategic direction?
- Predictive Models: How reliable are our predictive models in evaluating space exploration outcomes?

We aim to provide valuable insights that will shape Space Y's future by analysing these critical aspects. Join us as we leverage data science to unlock the possibilities of space exploration and make informed strategic decisions for Space Y's ventures!

Section 1

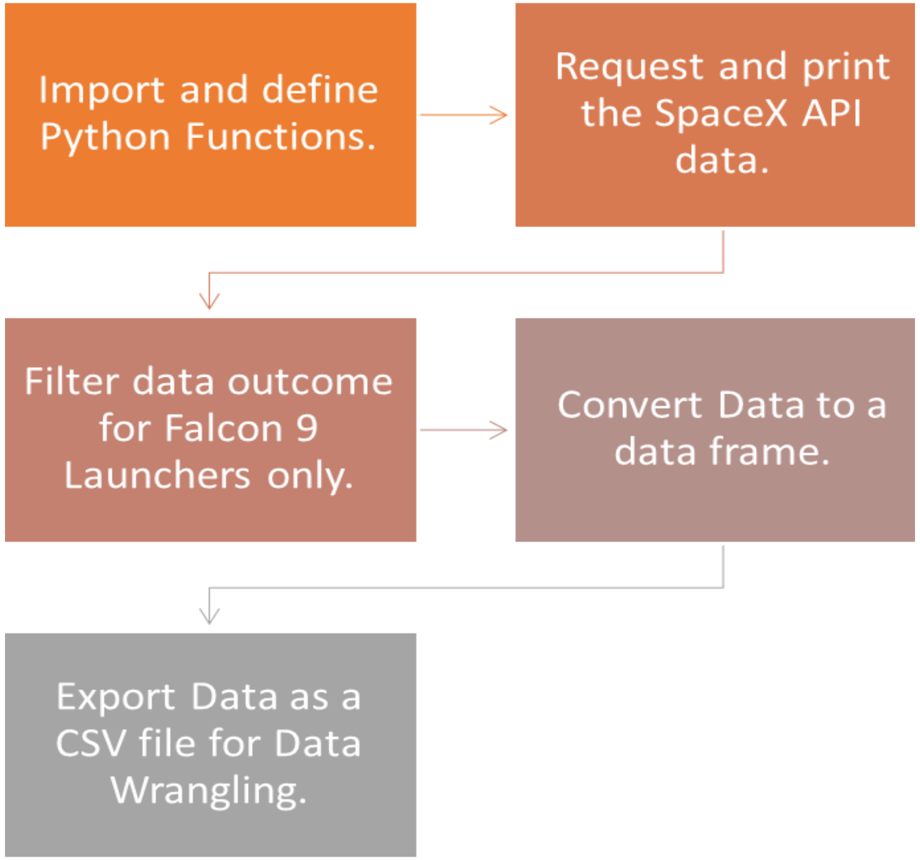
Methodology

Methodology

- Data collection methodology:
  - Data was collected by a combination of requests from Space X API and Web scrapping from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
- Perform data wrangling
  - The collected data was loaded as a data frame, checked for missing data and manipulated to ensure that data formats are correct for the analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was loaded as a data frame, standardized and split into test and train data. Also, the objects were fitted to the best parameters before classification models were tested for their accuracy.

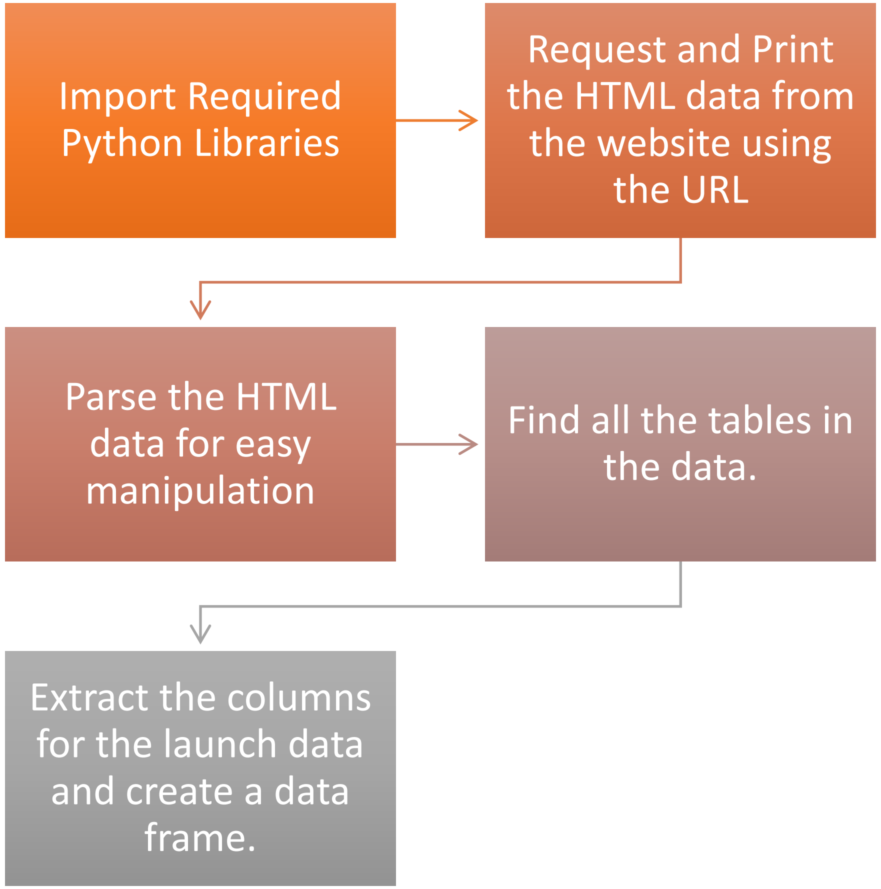
Data Collection - Space X API

- The useful Python libraries required for API data collection were imported and defined.
- Next, using the request and response function, data were obtained from the SpaceX API.
- The collected data were converted to a data frame using the JSON and normalized function.
- Finally, the data were filtered for only Falcon 9 launchers, leaving out other launchers.
- Click here for the github URL that shows the code and outcome of the data collection.



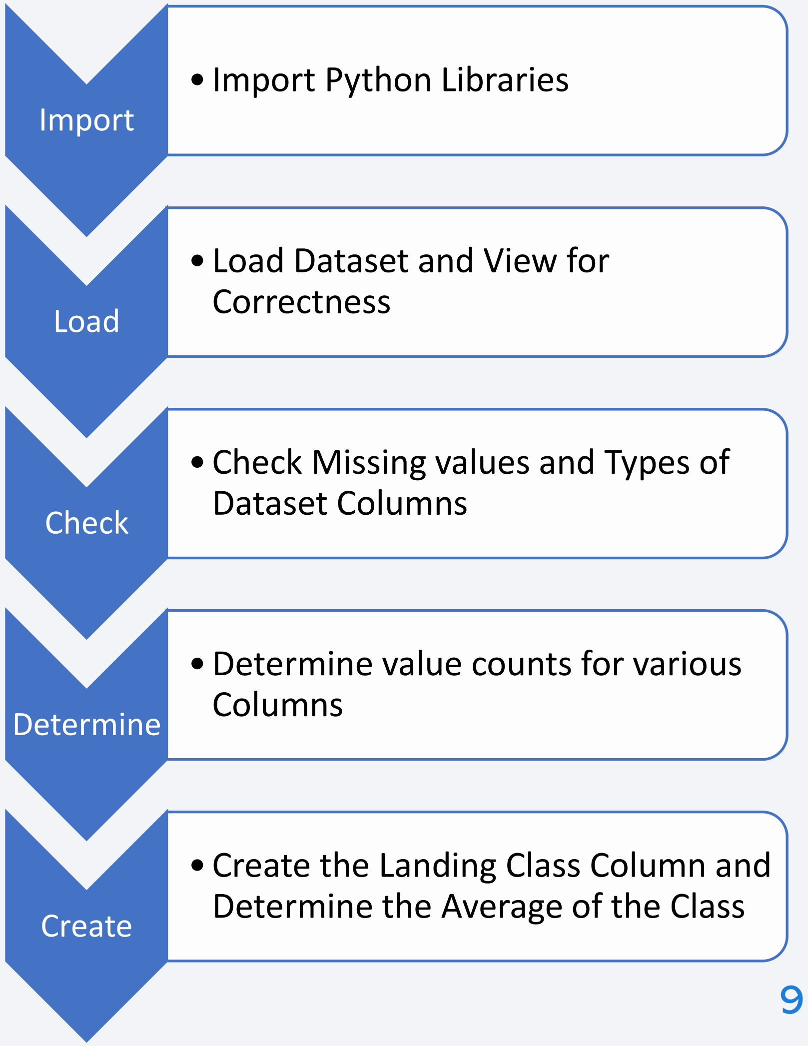
Data Collection - Web Scaping

- Web scraping allows us to obtain valuable datasets from websites.
- The request function was used to obtain the HTML data from the website, using the URL.
- The beautiful soup function was used to parse the HTML data for easy manipulation.
- The find-all function of the beautiful soup was used to extract all tables in the URL.
- Columns in the tables that have the Falcon 9 launch details were extracted and printed.
- The data frame of the launch records was created and extracted as a CSV file.
- Click here for the github URL that shows the code and outcome of the web scraping.



Data Wrangling

- Data wrangling involves processing data for further analysis.
- Dataset was loaded as a data frame by using the read_csv function. The dataset was viewed using the head function.
- Missing values were identified, and the data types were also checked for correctness using the df.types function.
- The value_count function was used to determine the number of launches for each site and the frequency of occurrences for each orbit and landing outcome.
- The class column was created from the outcome columns, with 0 for failed and 1 for successful outcomes.
- The mean value of class outcome was determined as 66.67%. Then, the dataset was exported as CSV for further analysis.
- Click here for the github URL that shows the code and outcome of data wrangling.



EDA with Data Visualization

- Categorical Plot (Cat Plot): This compares the distribution of data across different groups. For instance, the Cat plot was used to visualize how increases in flight numbers affect launch outcomes for different launch sites.
- Bar Chats: This shows the frequency of occurrence for a particular data group. Here, the chart was used to display success rates of launches per orbit.
- Line Plot: It displays a trend in the dataset across different periods. The plot was used to display the yearly success rate of launches.
- Click here for the github URL that shows the code and outcome of data visualizations.

EDA with SQL

- SQL query was performed using the select and distinct function to obtain unique launch sites.
- Other SQL queries used to obtain valuable data insights include the “where and like” function to obtain launch sites which begins with “CCA”. The wide card function was used to match the string.
- The arithmetic function was also used to obtain values like average, sum and count.
- Also, the “group by” function was used along with an arithmetic function in SQL to obtain the total of each category outcome.
- Click here for the github URL that shows the code and outcome of the SQL.

Build an Interactive Map with Folium

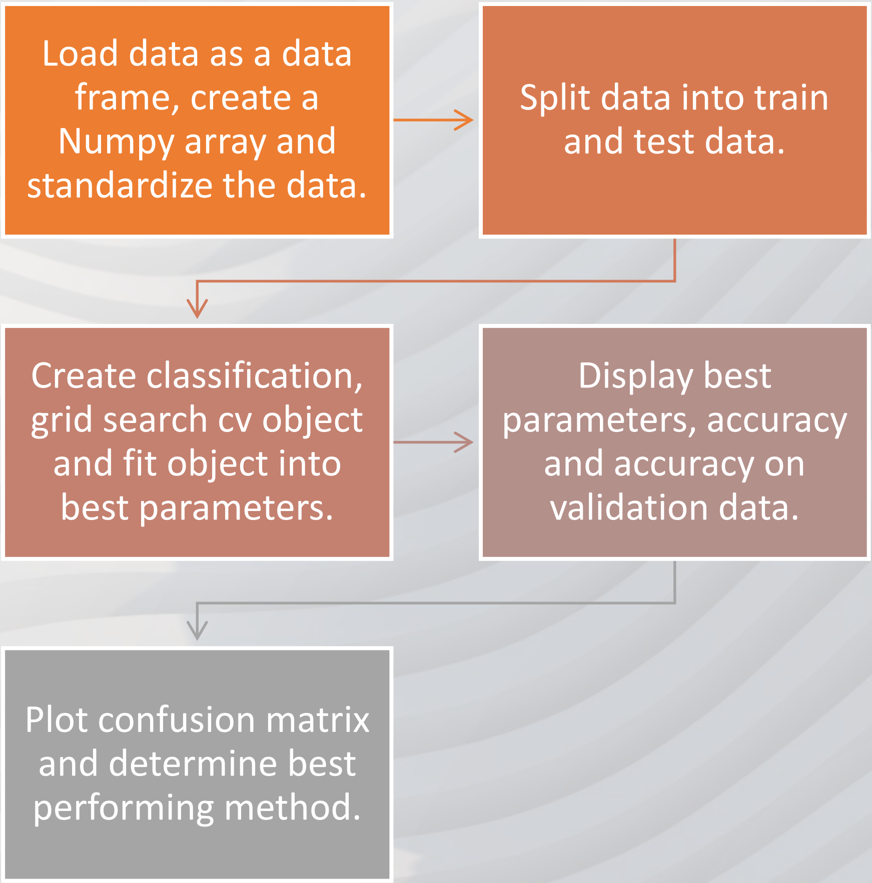
- Markers, circles, marker clusters and mouse positions were added to make the map interactive and more functional.
- The circle and the marker were used to identify the launch site locations on the map using the launch site coordinates.
- Marker cluster was used to display the outcomes of launches for each launch site on the map.
- Mouse position allows identification of coordinates on the map once a location is pointed. The distance of launch sites to coastlines, railways, highways etc. can be easily determined.
- Click here for the github URL that shows the code and outcome of the Map.
- Also, Click here to view the code and outcome of the Map in nbviewer because of the limitation of Github in displaying the interactive Map.

Build a Dashboard with Plotly Dash

- Pie chart and Scatter plots were added to the dashboard. Also, dropdown and range sliders were added as interactive tools to the dashboards.
- The pie chart displays the success rate of all rockets launched for all the launch sites and the success rate for each launch site.
- Scatter plot showed the relationships between the payload mass and the success rates.
- A dropdown was added to provide options for selecting the launch site to display by the pie chart while the range slider is used to adjust the payload masses.
- Click here for the github URL that shows the code of the dashboard.

Predictive Analysis (Classification)

- Import Python Libraries and define plot confusion matrix.
- Load data as a data frame, create a Numpy array and standardize the data.
- Split data into train and test data by using the train test split.
- Create classification objects, gridsearch_cv and fit objects into best parameters. Classification objects used are logistic regression, support vector machine, decision tree and k nearest neighbors.
- Display the best parameter using best_params_ and accuracy using best_score. Also, determine the accuracy on validation data using score.
- Plot the confusion matrix and determine the best-performing method.
- [Click here]() for the github URL that shows the code of the dashboard.



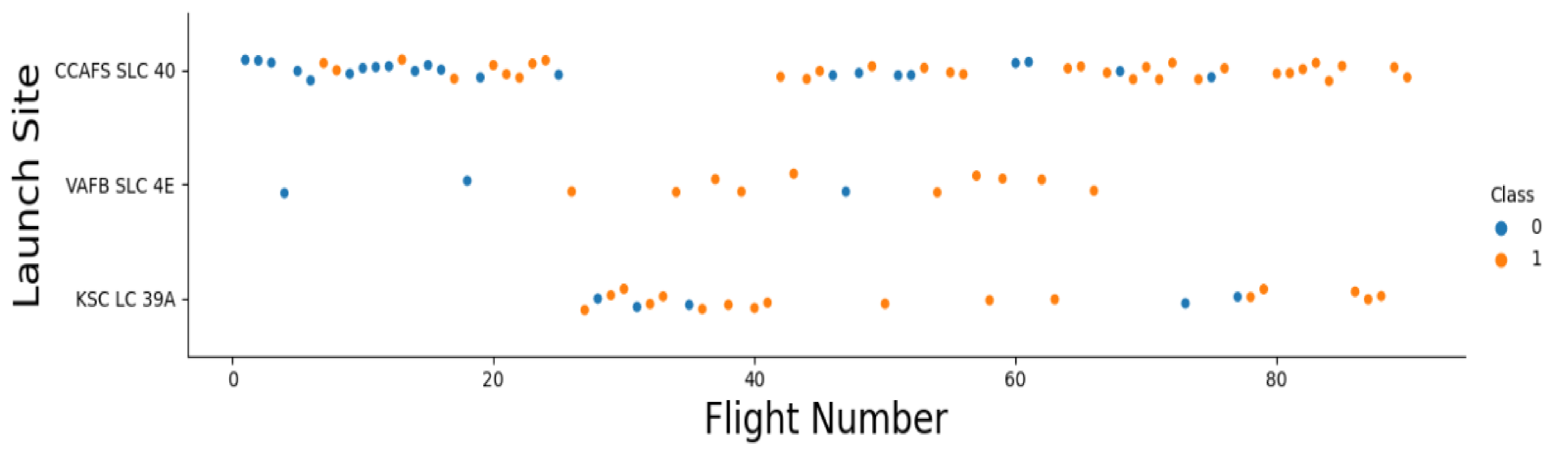
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



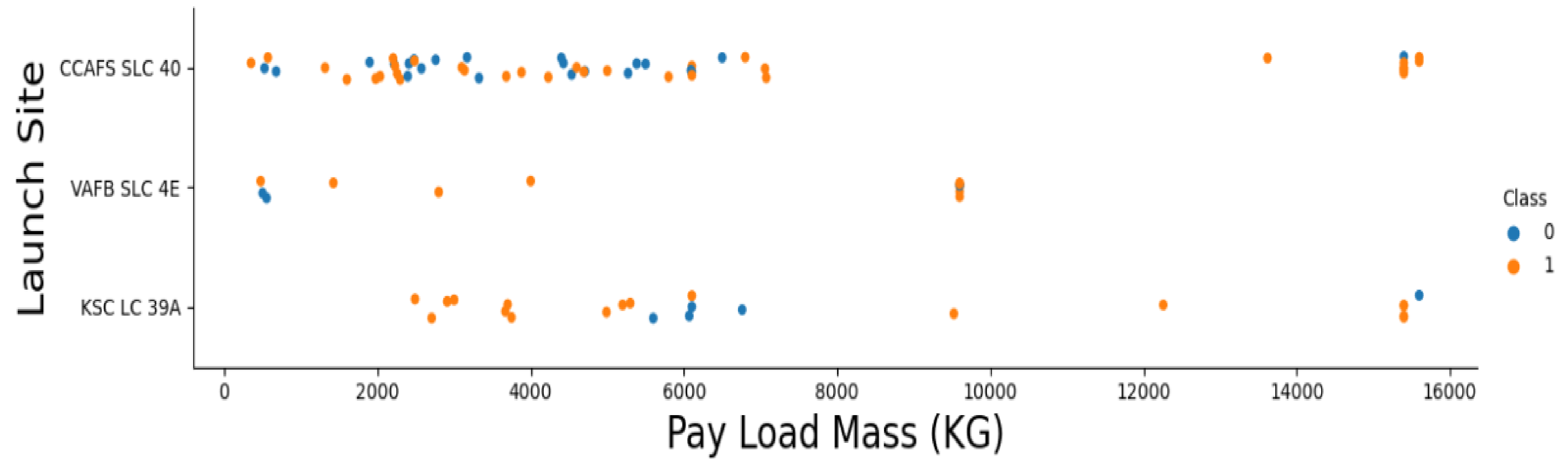
"0" represents a failed outcome, while "1" indicates a successful outcome.

The launches become more successful as the number of flights increases.

VAFB SLC-4E shows higher success rates after 20 flights but no launches beyond 70.

KSC LC 39A and CCAFS SLC-40: success rates of rockets increase when the number of flights is 80 or higher.

Payload vs. Launch Site



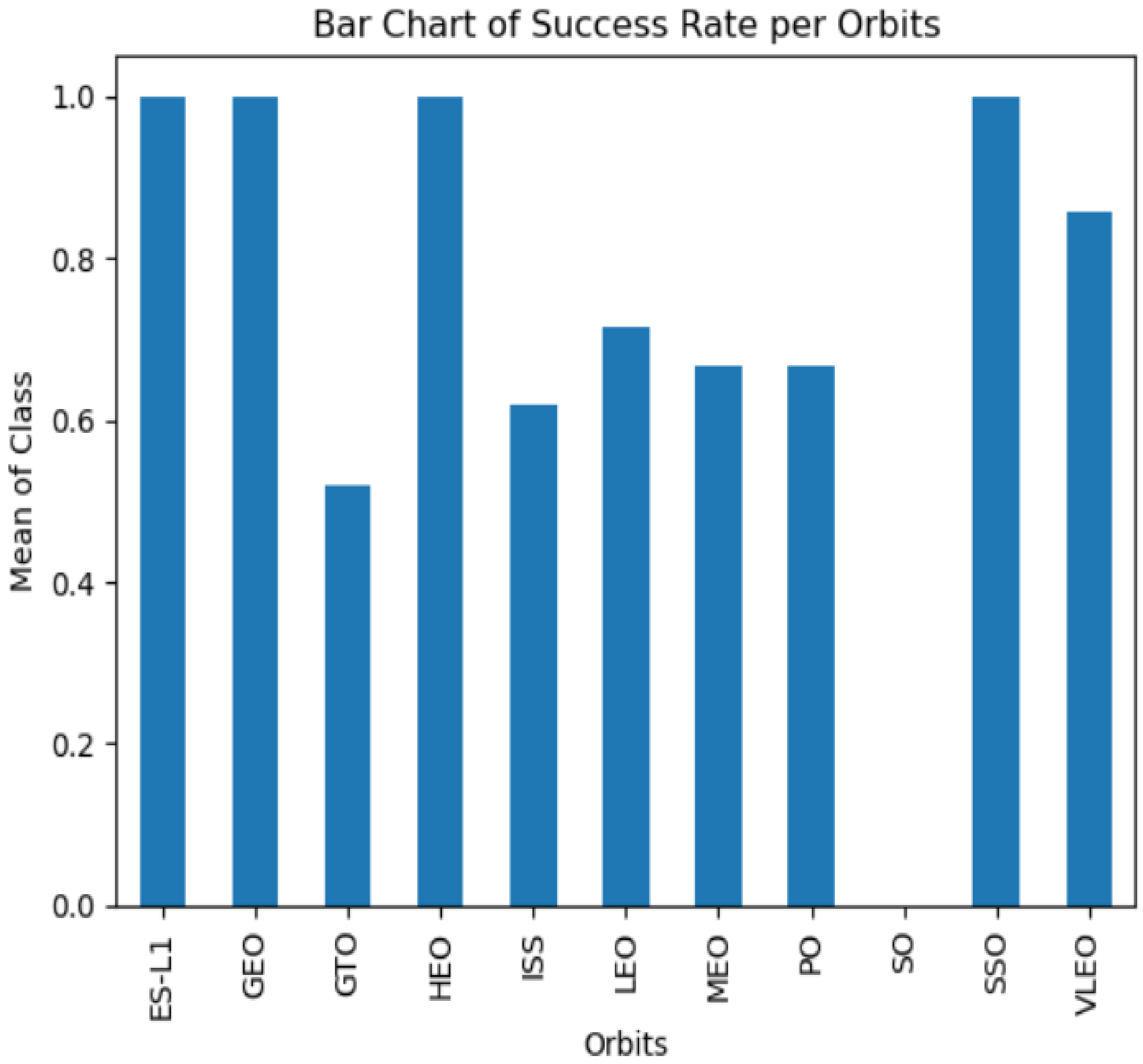
VAFB SLC-4E: No rocket launches have been observed with a payload mass beyond 10,000 kg.

CCAFS SLC-40: The landing outcome is independent of the payload mass.

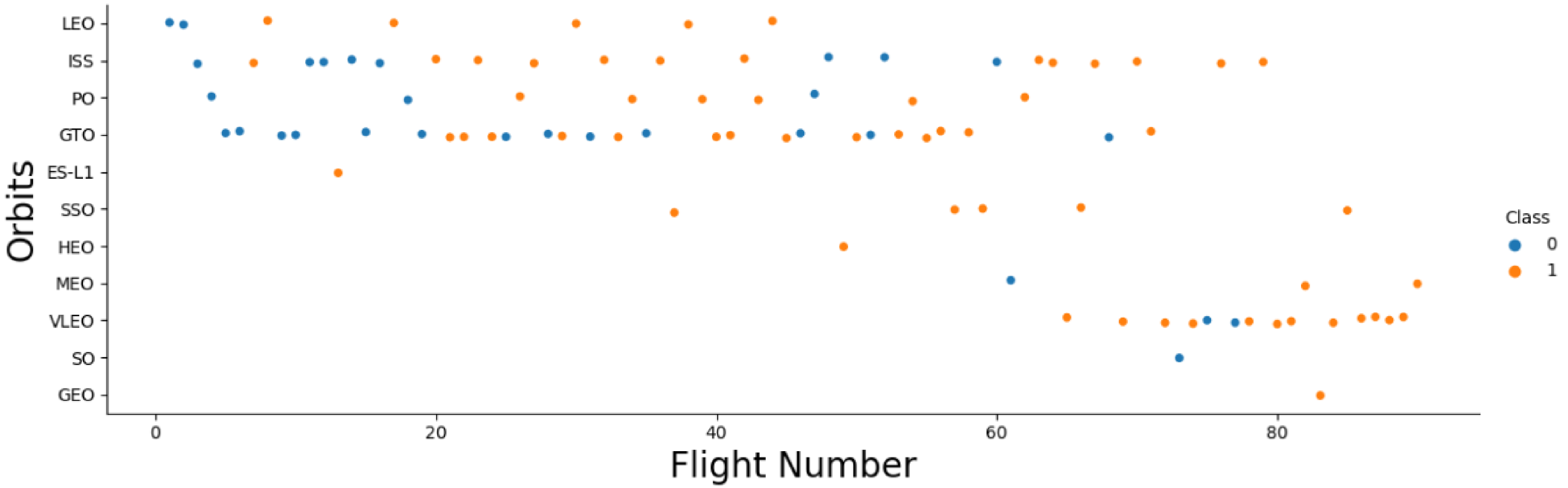
KSC LC-39A: Payloads with a mass below 5,000 kg are more likely to succeed at this site.

Success Rate vs Orbit Type

- Rockets launched at Orbit ES-L1, GEO (Geostationary Orbit), HEO (Highly Elliptical Orbit), and SSO (Sun-Synchronous Orbit) demonstrate the highest success rates.
- Rockets launched at Orbit SO (Orbit Standard) have experienced a 100% failure rate.



Flight Number vs. Orbits

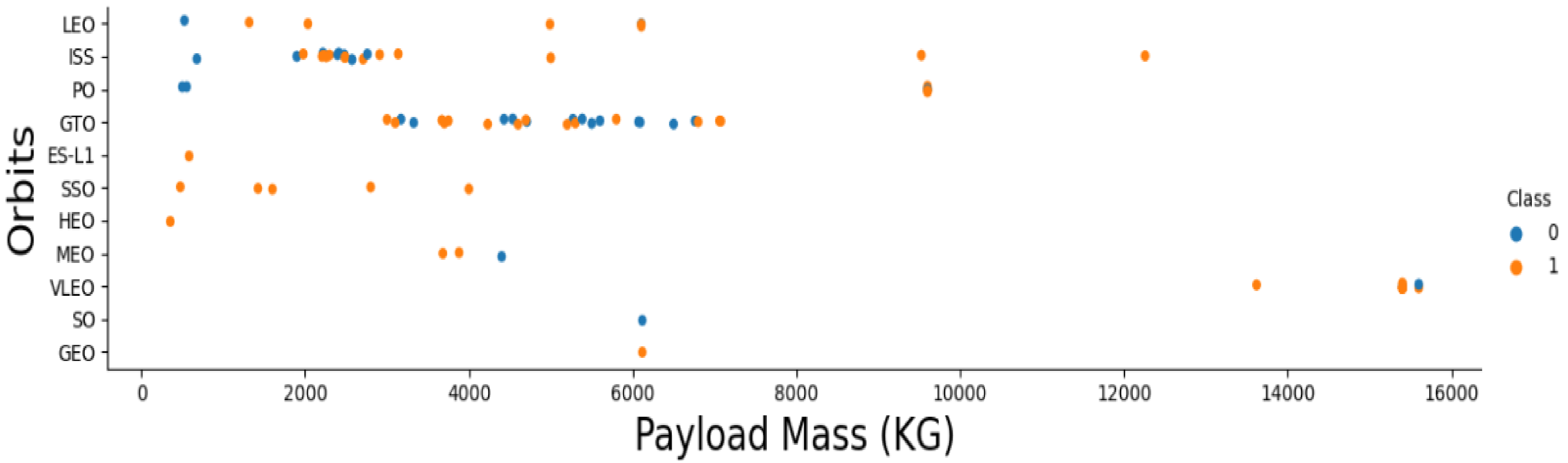


Orbit-type ISS (International Space Station) is more likely to succeed when the flight number is between 20 and 40 and above 60.

Orbits LEO (Low Earth Orbit), VLEO (Very Low Earth Orbit), and PO (Polar Orbit) show an increasing likelihood of success as the flight number increases.

Orbit SSO (Sun-Synchronous Orbit) has not experienced any failed outcomes regardless of the number of flights.

Payload Mass vs. Orbits



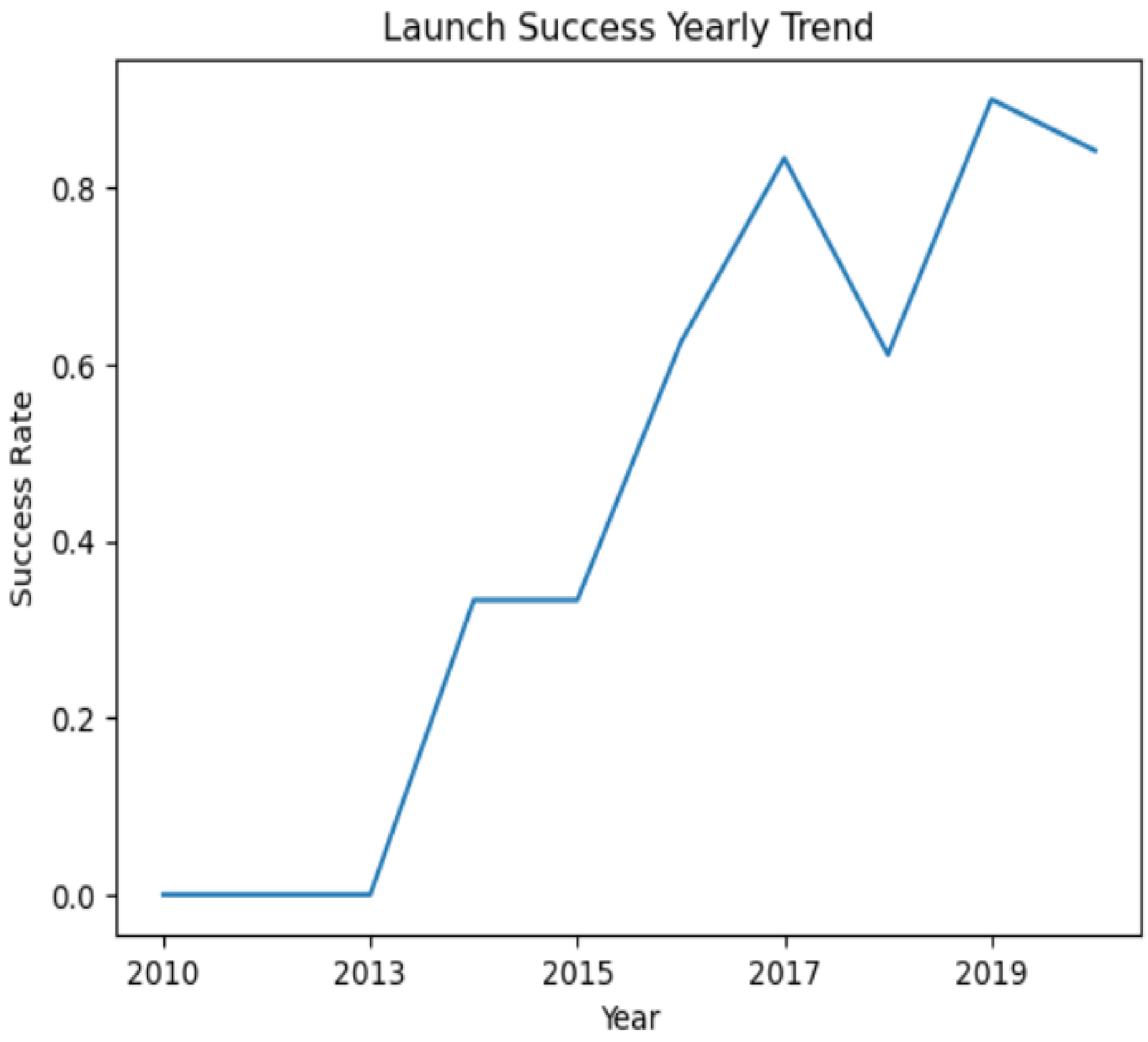
Rockets launched in Orbits LEO (Low Earth Orbit) and ISS (International Space Station)and SSO demonstrate a positive correlation between payload mass and success rate. Higher payload masses are associated with more successful outcomes.

For rockets in Orbit GTO (Geostationary Transfer Orbit), there is no clear relationship between payload mass and success rate. The success rate does not show a significant dependence on the payload mass.

In the case of SSO (Sun-Synchronous Orbit), there have been no failed outcomes regardless of the payload mass.

Launch Year Success Trend

- The success rate of SpaceX launches has shown a consistent progression since 2013. Over the years, there has been an overall increase in the success of launches.
- However, there was a slight drop in the success rate in 2018 compared to the previous year, 2017. This indicates a temporary decline in the success rate during that specific period.



All Launch Site Names

Query to Obtain Unique Launch Site Names

I utilized the power of Jupyter Notebook and SQL magic code to connect to the database and retrieve valuable information.

The unique names of launch sites were obtained by executing an SQL query. These include CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

The result of the query revealed four distinct launch sites within the dataset. These unique names provide important insights into the analysis's diversity and distribution of launch sites.

```
%sql select distinct (Launch_Site) from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

Five Records of Launch Sites with Names Beginning with ‘CCA’

```
%%sql SELECT *
FROM SPACEXTBL
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_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

This is five records obtained from launch sites with names beginning with CCA using the wide cat function with SQL magic in jupyter notebook.

Total Payload Mass Carried by NASA (CRS)

- The total payload mass carried by NASA (CRS) was obtained at 45,590kg.

- The power of Jupyter Notebook and SQL magic code was used to obtain this result. The notebook environment was connected to a database and valuable information was retrieved.

```
%%sql
SELECT SUM (PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

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

SUM (PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

The average payload mass carried by F9 v 1.1 was obtained at 2,928kg.

The power of Jupyter Notebook and SQL magic code was used to obtain this result. The notebook environment was connected to a database and valuable information was retrieved.

```
%%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

The first successful ground pad landing occurred on 01-08-2018. Although the year trend analysis shows that space x rockets became successful after 2013. However, the ground landing was not successful until 2018.

This insightful information was obtained through Jupyter Notebook's power and SQL magic code's utilisation. We retrieved and analysed valuable data by connecting the notebook environment to a database, uncovering significant events and patterns within SpaceX's journey.

```
%%sql
SELECT MIN(Date) as first_successful_ground_pad_landing_date
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'
```

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

first_successful_ground_pad_landing_date
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

The successful outcome of the drone ship landing with payload masses between 4000 and 6000kg are F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2.

The distinct function was used in the SQL magic query to ensure that the booster version display unique outcomes.

The “where” function was used to specify two conditions to realise the desired outcome. One condition factored in the landing outcome and the other, the payload mass. The two conditions must be met to obtain the outcome.

```
%%sql
SELECT distinct (Booster_Version)
FROM SPACEXTBL
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

- The query shows the classifications of mission outcomes as failure or success. The “group by” function in SQL magic was used to achieve this outcome.

- Also, to obtain the total number of each group, the count function was used. The output displays two columns, the first, showing unique categories under mission outcome, while the second column shows the total count for each category.

List the total number of successful and failure mission outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) as total_count
FROM SPACEXTBL
GROUP BY mission_outcome;
```

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

Mission_Outcome	total_count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The query results in boosters' versions that have carried the maximum payload mass.

The outcomes are 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 and F9 B5 B1049.7 .

To obtain the outcome, a sub-query was used to specify the maximum payload mass with the “where” condition.

“Distinct was also used in the select function to limit the list to unique booster versions.

```
%%sql
SELECT distinct (booster_version)
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
);
```

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

2015 Launch Records

The query shows the record for failed landing outcomes in the drone ship, booster versions, launch sites and the months of occurrence in 2015.

The failure happened in April and October. The booster versions are F9 v1.1 B1012 and F9 v1.1 B1015. The launch site is CCAFS LC-40.

The SQL magic query specifies the condition of the landing outcome as “Failure (drone ship).

```
%%sql
SELECT
CASE substr(Date, 4, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS month_name,
Landing_Outcome as Failure_Landing_Outcome,
Booster_Version,
Launch_Site
FROM SPACEXTBL
WHERE substr(Date,7,4)='2015' AND
Landing_Outcome = 'Failure (drone ship)'
```

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

month_name	Failure_Landing_Outcome	Booster_Version	Launch_Site
October	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 query outcome shows three categories. Success with 20 outcomes, Success (drone ship) with 8 and Success (Ground pad) with 7.

- The SQL magic query specifies two “Where” conditions, one with date intervals and the other with landing outcome.

- An additional condition is specified for the landing outcome, using “like’ and the wide cat to identify items that have success.

- Finally, the group by and order by functions are used to refine the outcome.

```
%%sql
SELECT Landing_Outcome as Successful_La
FROM SPACEXTBL
WHERE Date Between '04-06-2010' and '20
AND Landing_Outcome like '%Success%'
GROUP BY Landing_Outcome
ORDER BY Rank DESC
```

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

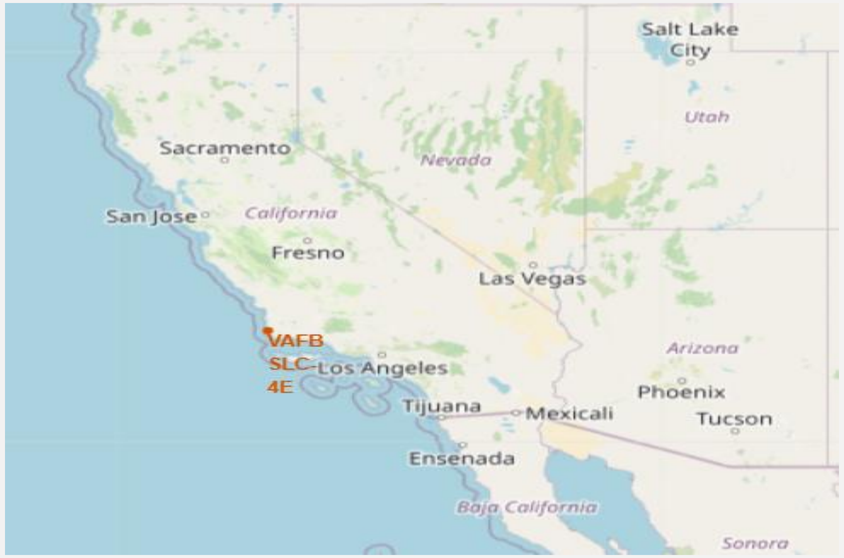
Successful_Landing_Outcomes	Rank
Success	20
Success (drone ship)	8
Success (ground pad)	7

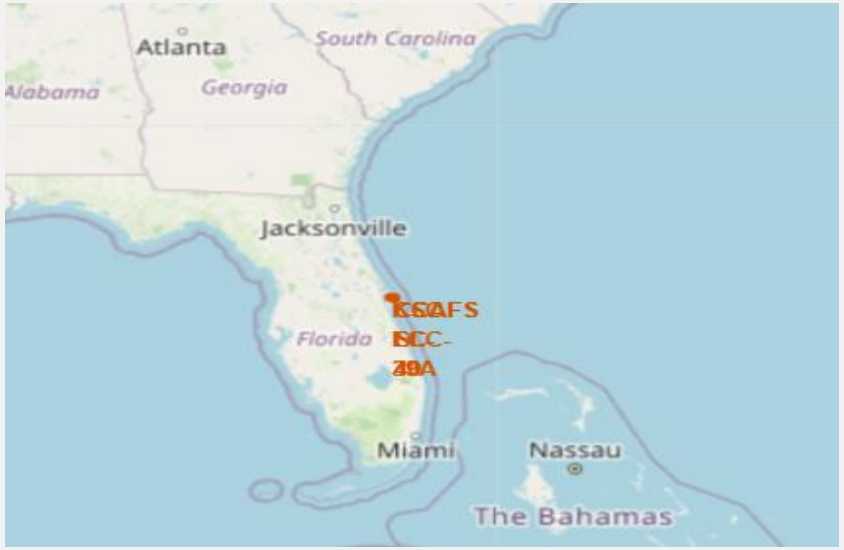
Section 3

Launch Sites Proximities Analysis

Location of Launch Sites on the Global Map

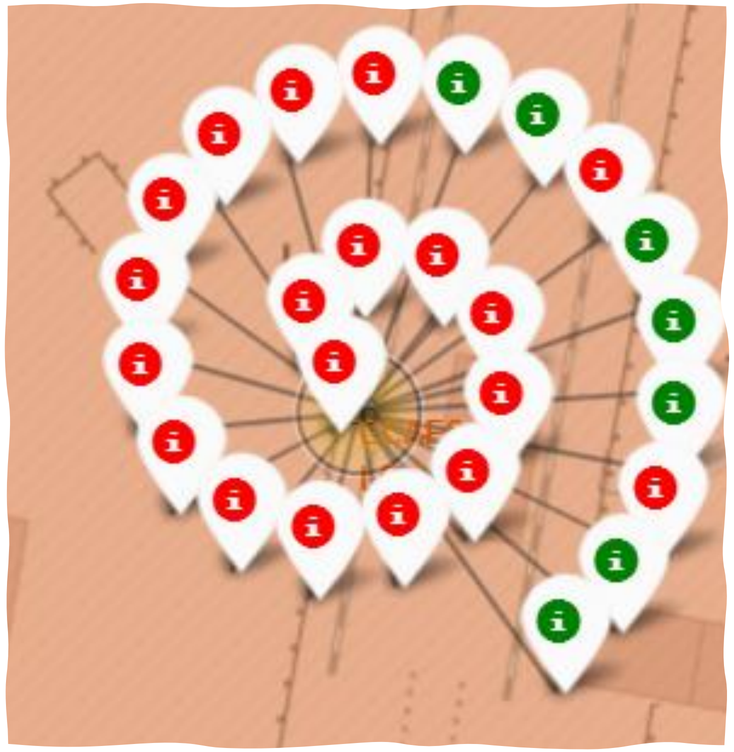
- VAFB SLC-4E: Close to Los Angeles
- CCAFS LC-40, KSC LC-39A, and CCAFS SLC-40: Close to Florida
- All launch sites: Located near the coast.
- Launch sites: Positioned away from cities

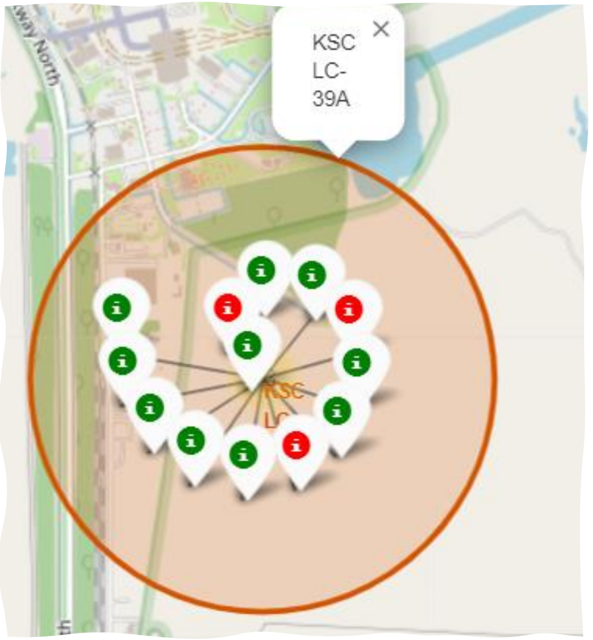


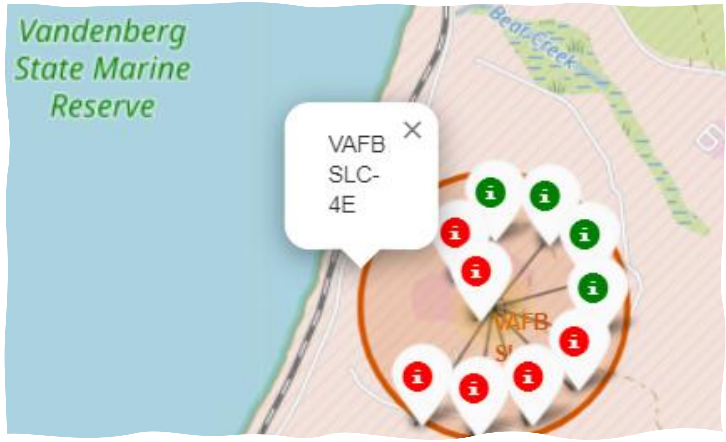


Mark Showing Success/Failure Rate of Launch Sites

- CCAFS LC-40: 26 launches, 7 success outcomes.
- KSC LC-39A: 13 launches, 10 success outcomes.
- VAFB SLC-4E: 10 launches, 4 success outcomes.
- CCAFS SLC-40: 7 launches, 3 success outcomes.







The proximity of Launch Site to a Coastline

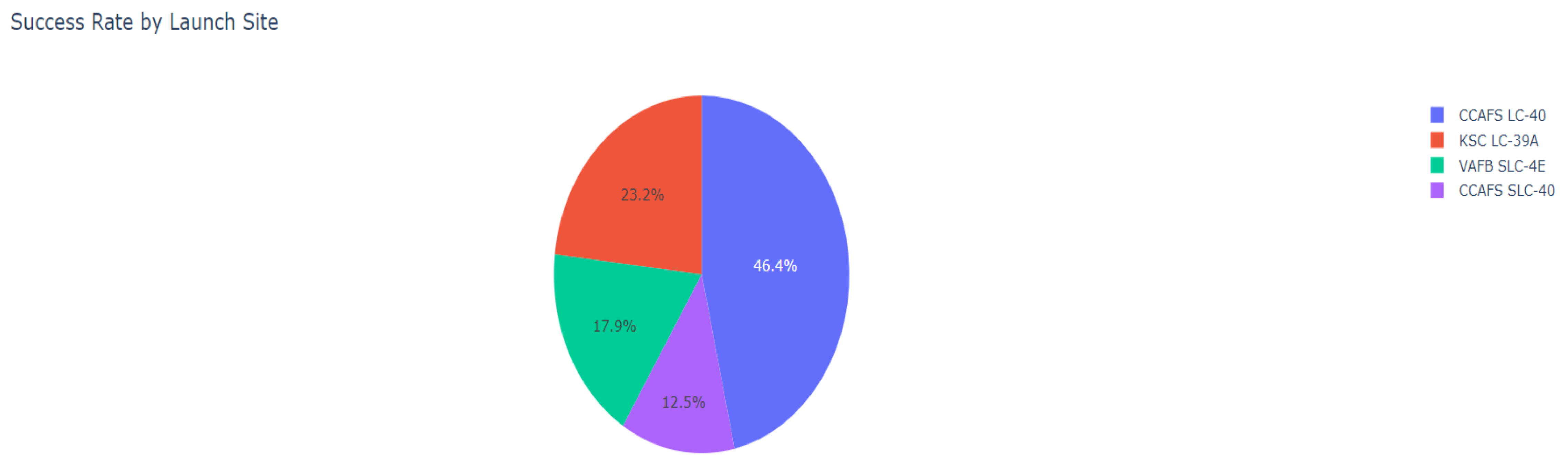
The launch site CCAFS SLC-40 is 0.87 km from the closest coastline. Similarly, all the launch sites show proximity to the coastlines and away from cities.

Section 4

Build a Dashboard with Plotly Dash

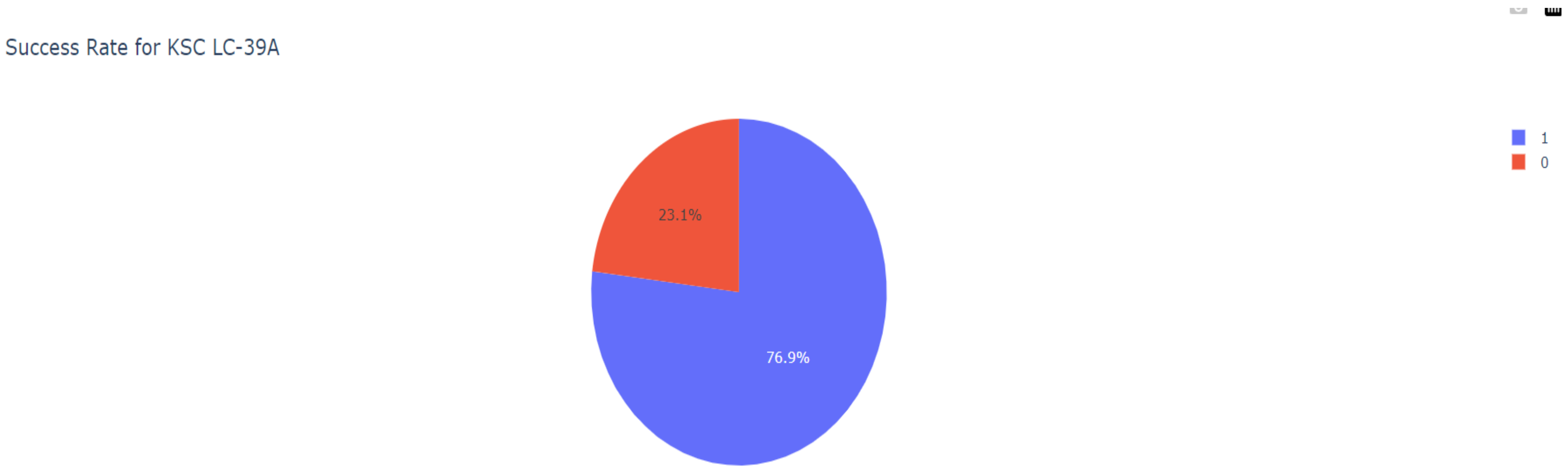
Success Rate by Launch Site

- CCAFS LC-40 has the highest success rate by launch site with a 46.4% success rate. More numbers of launches occurred at the launch site.



Launch Site with the Highest Success Ratio

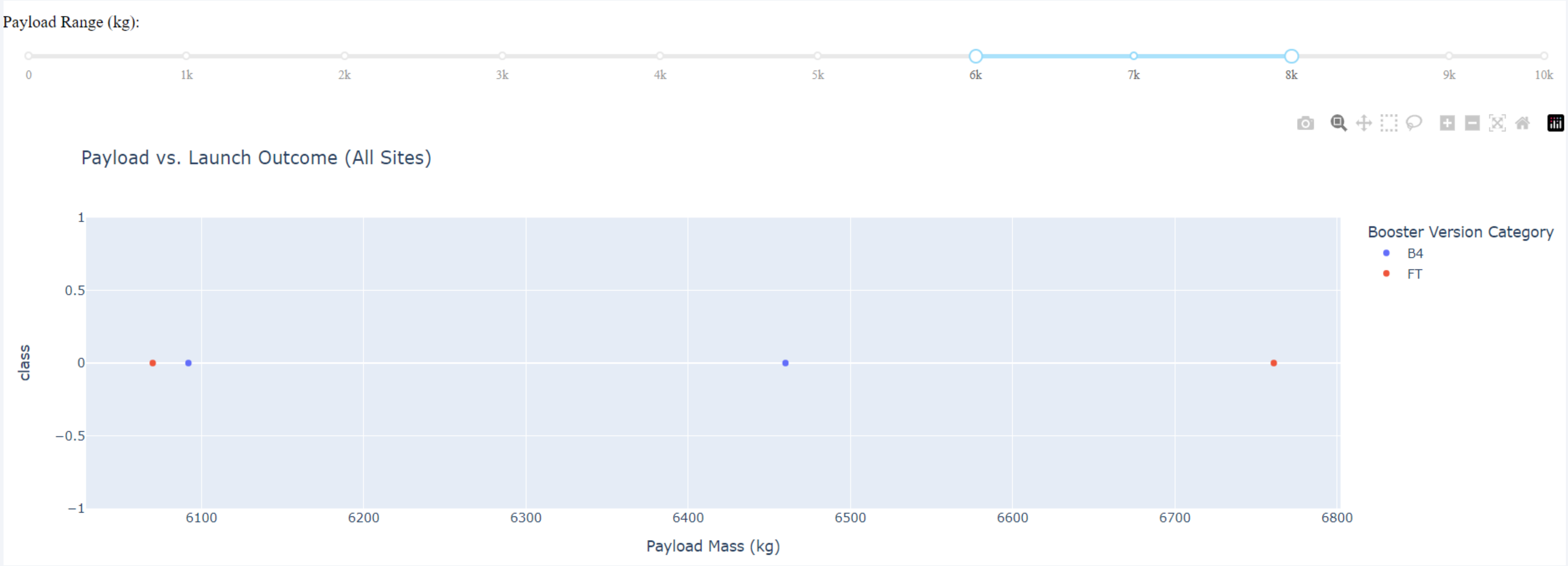
- KSC LC-39A has the highest launch success ratio at a 76.9% success rate. It shows that for every 10 launches, more than 7 are likely to succeed.



Payload Range with the Highest Success Outcome

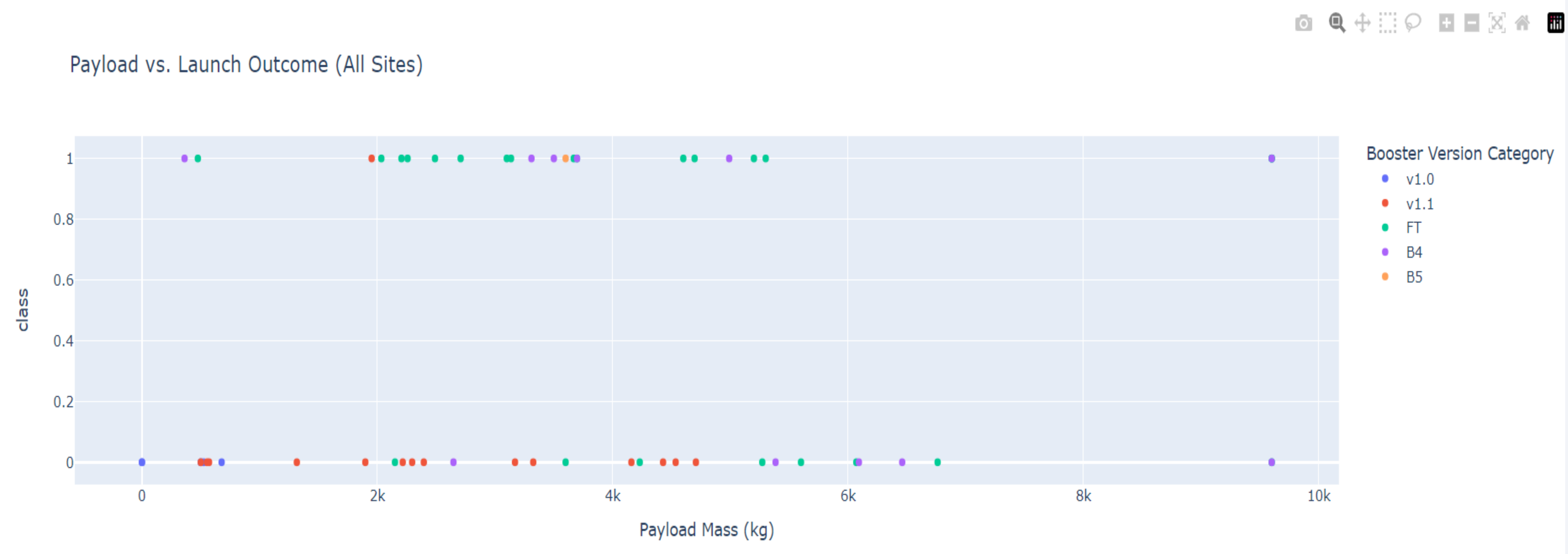
The payload range between 2000 and 4000KG has the highest launch outcome for all launch sites.

Payload Range with the Lowest Success Outcome



The payload range between 6000 and 8000KG has the lowest launch outcome for all launch sites.

Payload Ranges for All Outcomes



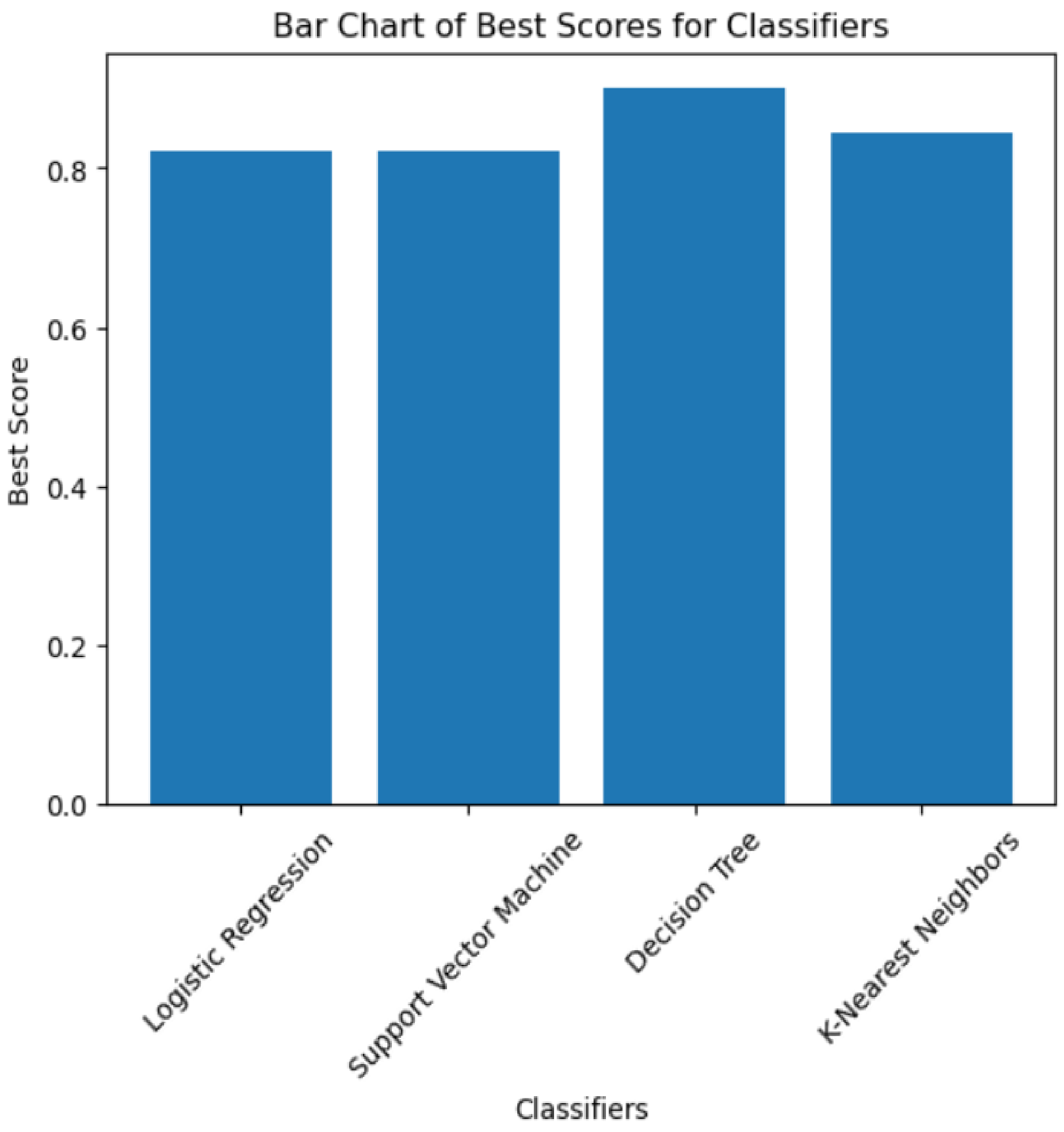
- Payload ranges for all outcomes vary depending on the specific mission and rocket used.
- Booster version FT has the highest success launches among the different versions.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

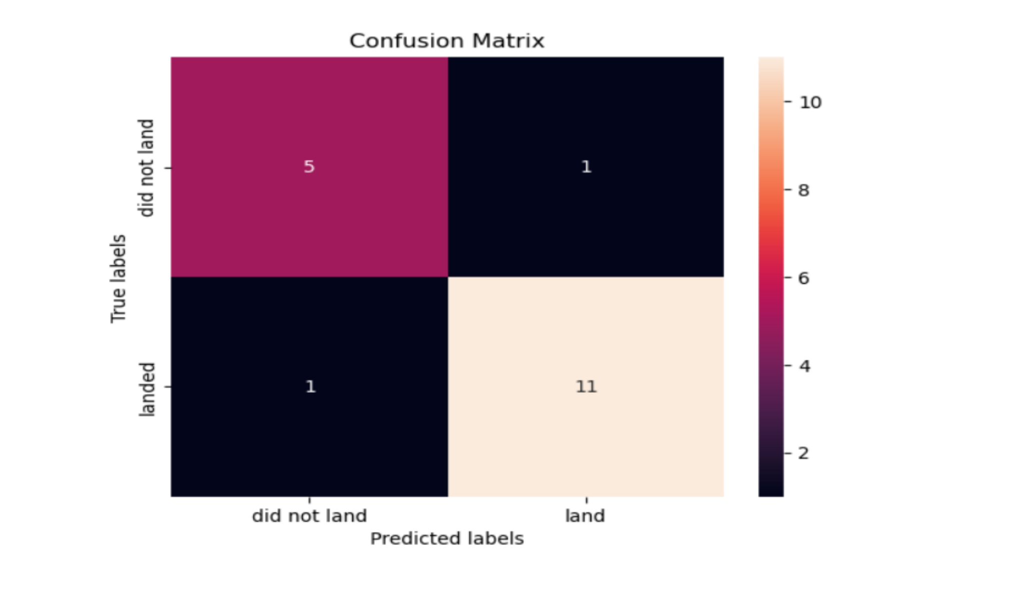
Decision three has the best accuracy score at 0.90.

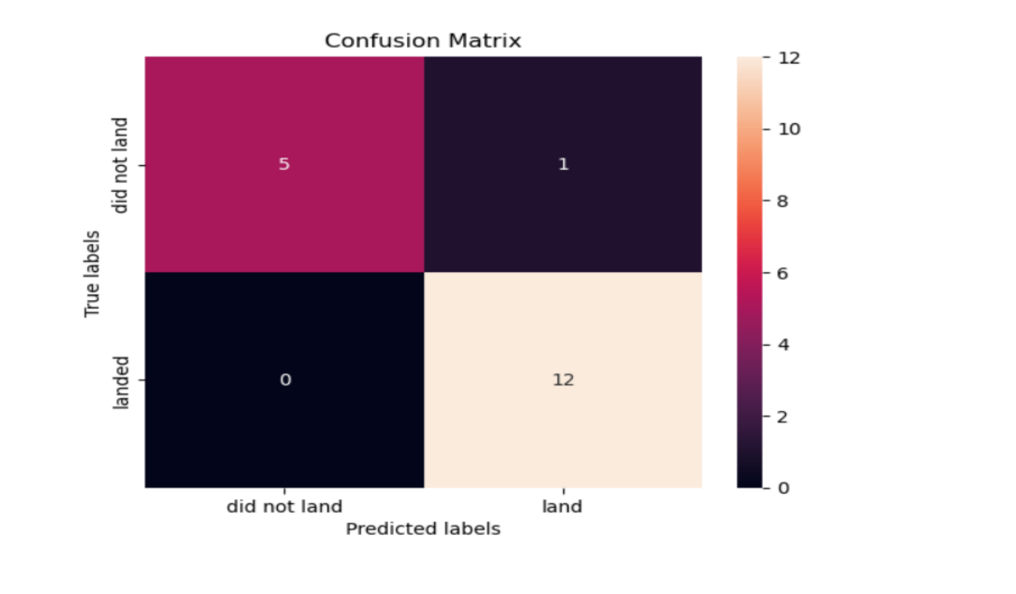


Confusion Matrix

The confusion matrix of log_reg, SVM and KNN performs better with similar results. Although they all predicted an outcome that did not land as land.

The decision tree has both false positives and false negatives of one each. In the case of the false negative, adequate preparation might not have been in place for a rocket that eventually landed.





Conclusions

- SpaceX's launches have shown consistent success since 2013, with a slight decline in 2018.
- Payload masses between 2000-4000 kg demonstrate the highest launch success across all sites.
- Orbits ES-L1, GEO, HEO, and SSO exhibit the highest success rates, while Orbit SO has experienced 100% failure.
- KSC LC-39A achieves a remarkable 76.9% success rate, making it a promising launch site.
- Decision three offers the best accuracy score at 0.90, providing reliable predictive insights.
- Learning from past failures, Space Y can adapt and succeed in the dynamic commercial space industry.
- By leveraging these data-driven insights, Space Y can make informed decisions, optimize launch success, and pave the way for a successful entry into the commercial space industry.
- Thank you for joining us on this enlightening data exploration mission. Together, we aspire to reach for the stars and shape the future of commercial space endeavours.

Thank you!