Falcon 9 Launch Success Prediction

Diogo Miranda
06/03/2024

Table of Contents

Executive Summary

Purpose

The goal of this project is to accurately determine the outcome of Falcon 9 rocket launches to yield financial benefits.

Summary of Methodologies

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

Summary of Results

- EDA results
- Interactive analytics demo
- Predictive analysis results

Introduction

Background and context

SpaceX, a leader in the space industry, advertises Falcon 9 rocket launches on its website, priced at $62 million. In contrast, other providers offer solutions upwards of $165 million each. A significant portion of the cost savings stems from SpaceX's innovative ability to reuse the first stage of its rockets. Hence, accurately predicting whether the first stage will land successfully holds immense financial implications, directly influencing the cost of a launch. Such insights become invaluable for potential competitors seeking to bid against SpaceX for rocket launch contracts, allowing for informed decision-making and strategic positioning in the market.

Common challenges to address

- Determining variables that influence launch outcome.
- Significance of each feature.
- Impact of relationships within rocket features.
- External conditions and their effect.

Methodology

Methodology

Data collection

- Utilize the SpaceX REST API for direct data retrieval.
- Employ web scraping techniques to extract data from Wikipedia.
- Conduct data wrangling to clean and organize the collected data.
- Apply one-hot encoding to categorical features for modeling.

Exploratory Data Analysis

- Utilize visualization techniques to explore and understand the data.
- Utilize SQL queries to extract insights from the dataset.

Interactive Visual Analytics

- Utilize Folium for interactive mapping and geographical visualization.
- Utilize Plotly Dash for interactive and dynamic visual analytics.

Predictive Analysis

- Employ classification models for predictive analysis.
- Train, tune, and test several classification models to identify the most effective approach.

Data Collection

The primary dataset was obtained through the SpaceX API, where data on launches from 2010 to the present was requested and retrieved. Subsequently, the raw data underwent transformation into a more manageable format and was organized into a dataframe for analysis.

Supplementary information was gathered from Wikipedia using web scraping techniques facilitated by BeautifulSoup.

Data Collection

SpaceX API

- Request data from SpaceX API
- Decode response and normalize it into a structured flat table format
- Use custom functions to extract and save specific data of interest
- Create a dataframe from the extracted data
- Filter the dataframe to include only Falcon 9 rocket launches
- Replace missing values of payload mass with the mean value

GitHub URL to SpaceX API Notebook

Web Scraping

- Get an HTTP response from Wikipedia
- Create a BeautifulSoup object
- Extract column names from HTML table
- Create a dataframe from the extracted data

GitHub URL to Web Scraping Notebook

Data Wrangling

A binary classifier column was introduced to categorize the success or failure of booster landings.

This additional column condenses the categorical data from the "Outcome" column into two distinct categories, simplifying the dataset for training our predictive model.

GitHub URL to Data Wrangling Notebook

EDA with Visualization

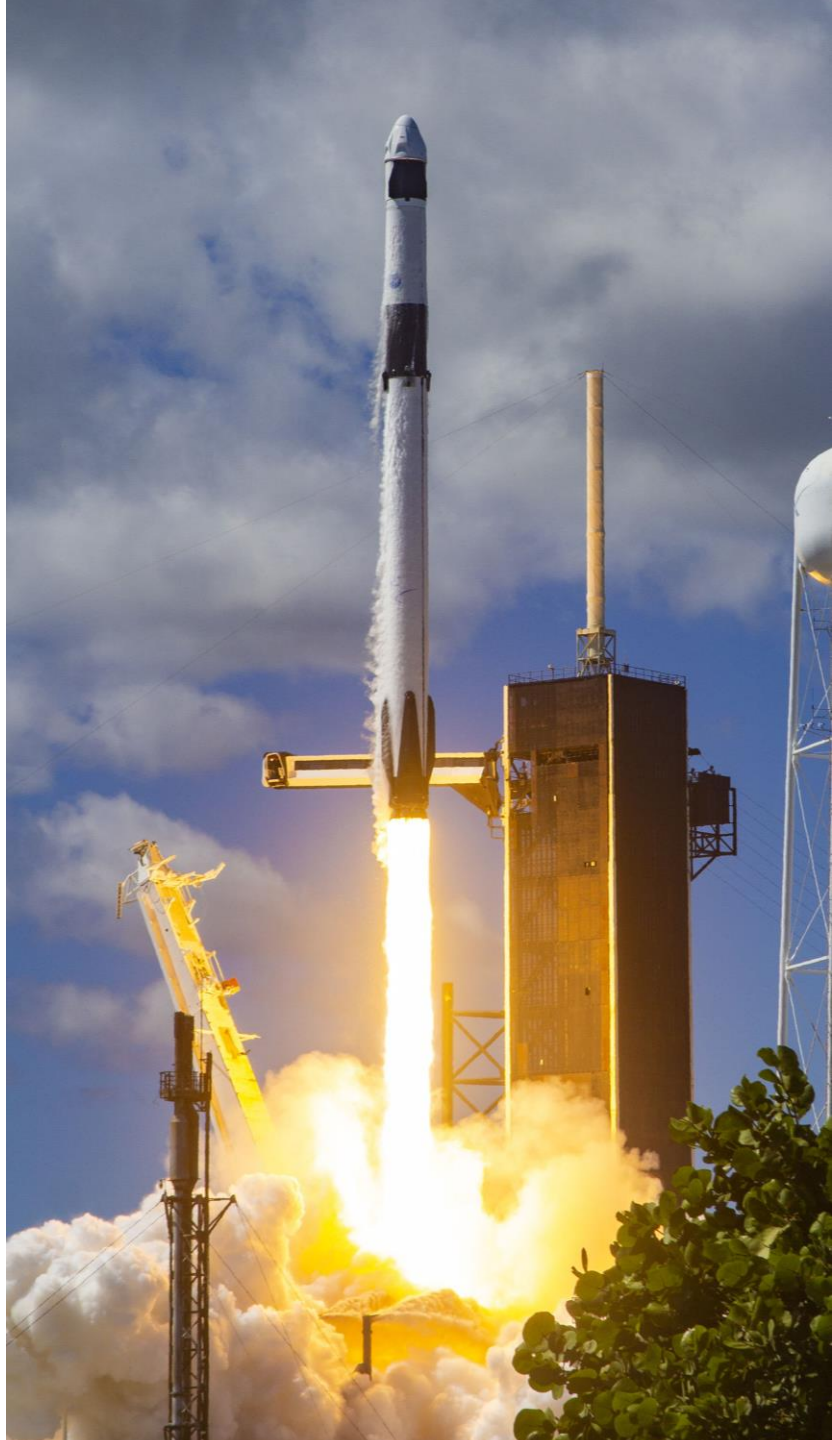
The charts that were plotted include:

- **Scatter charts** (display the relationship between two features)
  - Launch Site vs. Flight Number
  - Payload Mass vs. Flight Number
  - Orbit vs. Flight Number
  - Orbit vs. Payload Mass

- **Bar chart** (displays the magnitude of each feature)
  - Success Rate vs. Orbit

- **Line chart** (displays a trend over time)
  - Success Rate vs. Date

GitHub URL to EDA with Visualization Notebook



EDA with SQL

The following queries were executed to perform EDA:

- Display the names of the unique launch sites
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL to EDA with SQL Notebook

Folium Interactive Map

The following components were generated for visualization:

- **Individual circles and markers for each launch site** to accurately pinpoint their locations on the map.
- **Marker clusters to group launches at each site together**, streamlining visualization especially when coordinates are closely positioned at a larger scale. This feature also aids in quickly identifying successful rocket landings by their color (green for success and red for failure).
- **Lines from CCAFS SLC-40 Launch Site to nearest coastline, city, railway, and highway**. This analysis provides insights into the proximity of launch sites to different landmarks, facilitating a comprehensive understanding of their geographic positioning.

GitHub URL to Interactive Map Notebook

Plotly Dash Dashboard

Dropdown menu

- Allows users to select data pertaining to a specific launch site or from all launch sites.

Pie chart

- Shows the total number of successful launches, offering insights for either all sites or the selected site.

Scatter plot

- Demonstrates the correlation between payload and launch success.

Payload slider

- Allows users to select specific payload ranges, providing them with the capability to manipulate the displayed data withing the scatter plot window.

GitHub URL to Dashboard Script

Predictive Analysis

Model Building

- Load and transform the data.
- Split data into train and test sets.
- Initialize the model.
- Perform cross-validation for parameter tuning.
- Fit GridSearchCV object to the data.

Model Evaluation

- Compute accuracy scores.
- Analyze the confusion matrix.

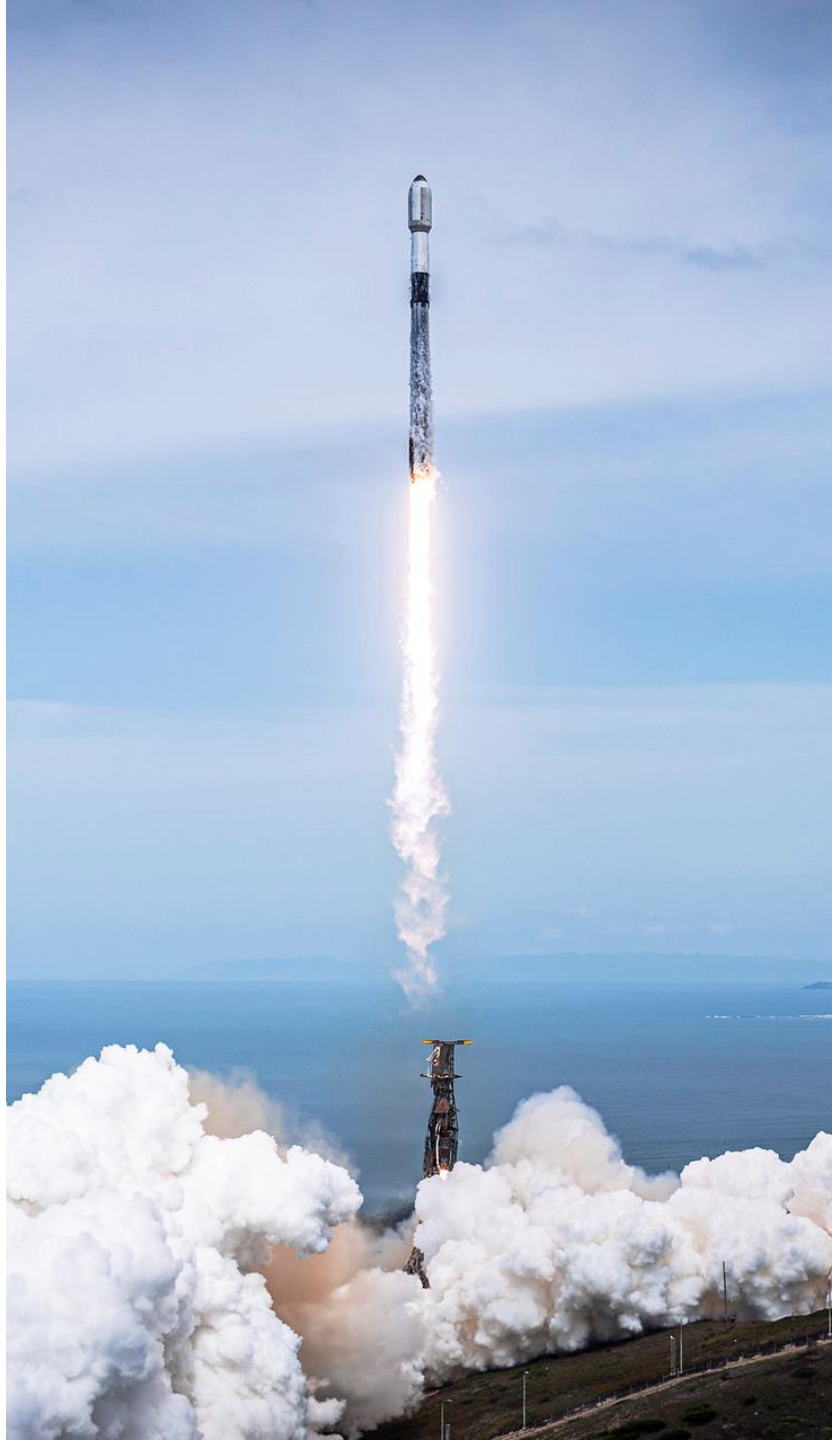
Model Selection

- Compare test score accuracies and choose the model with the highest score.

Model Improvement

- Feature Engineering.
- Parameter Tuning.

GitHub URL to Machine Learning Notebook



Results

Exploratory data analysis results

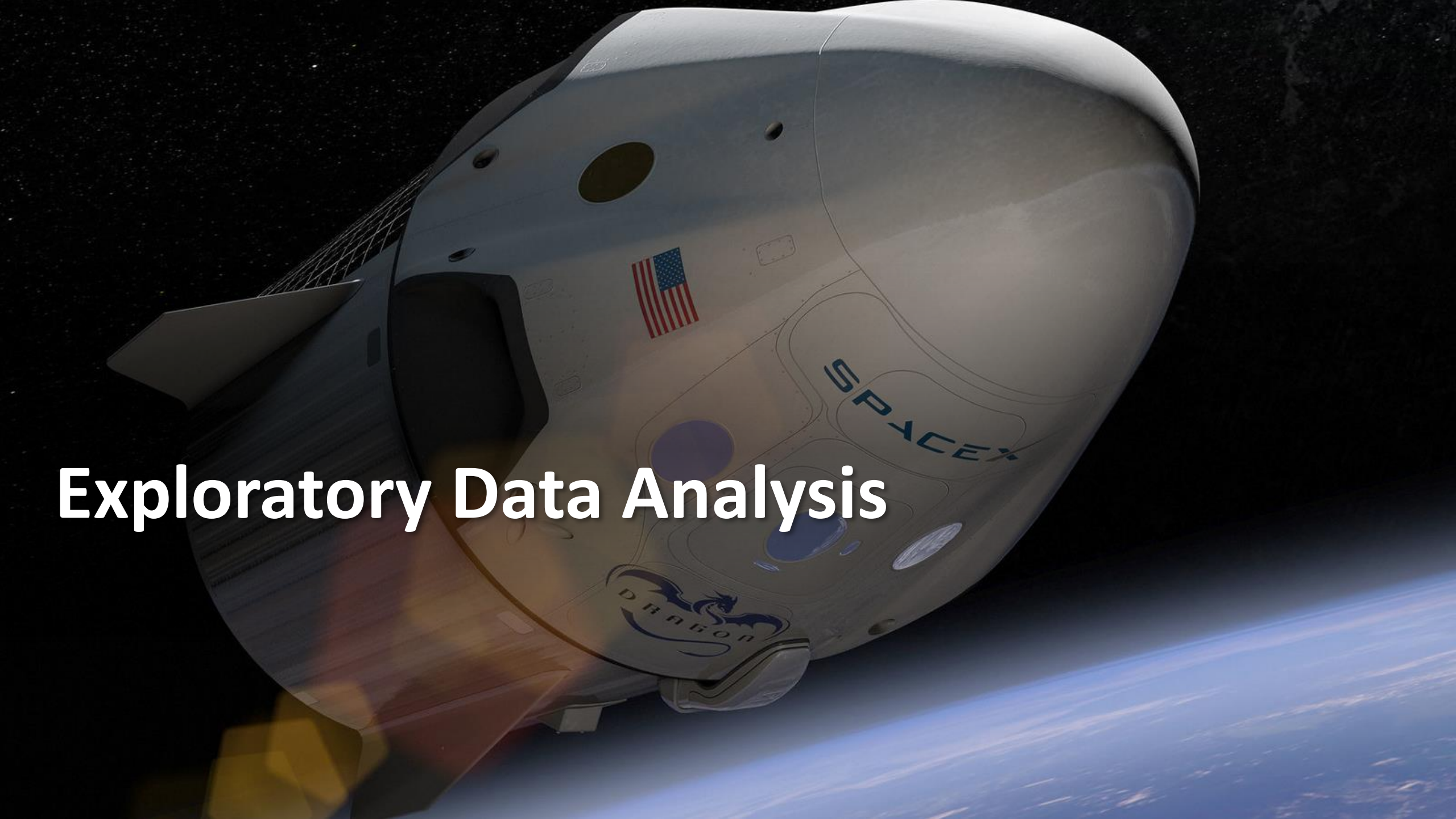
- Landing success has improved over time.
- Orbits ES-L1, GEO, HEO and SSO demonstrate 100% success rate

Interactive analytics demo in screenshots

- KSC LC-39A is the launch site with the highest number of successful launches.
- Launch sites exhibit proximity to coastlines and distance from urban centers, railways and highways.

Predictive analysis results

- The logistic regression model achieves an accuracy of 83.3% in correctly predicting the outcome of a rocket launch.

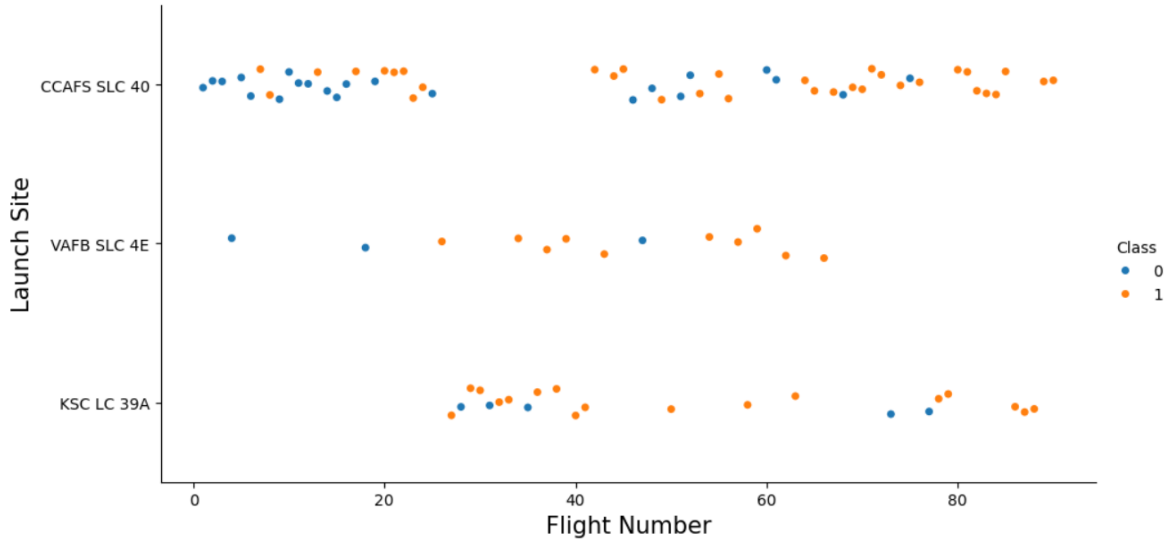


Exploratory Data Analysis

EDA with Visualization

Launch Site vs. Flight Number

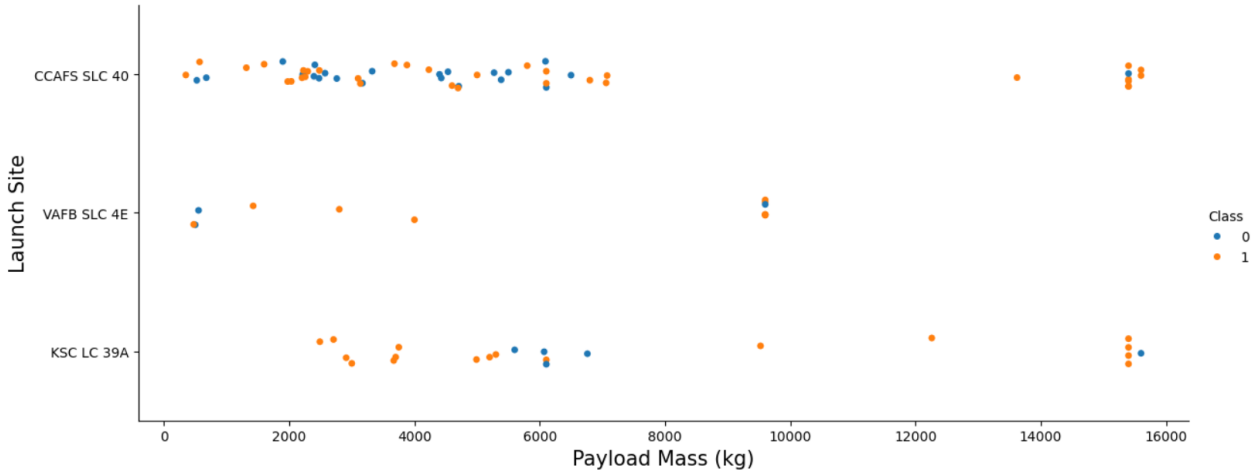
- Success Rate increases with flight frequency.
- CCAFS SLC 40 has the highest number of launches.
- VAFB SLC 4E and KSC LC 39A show higher success rate.

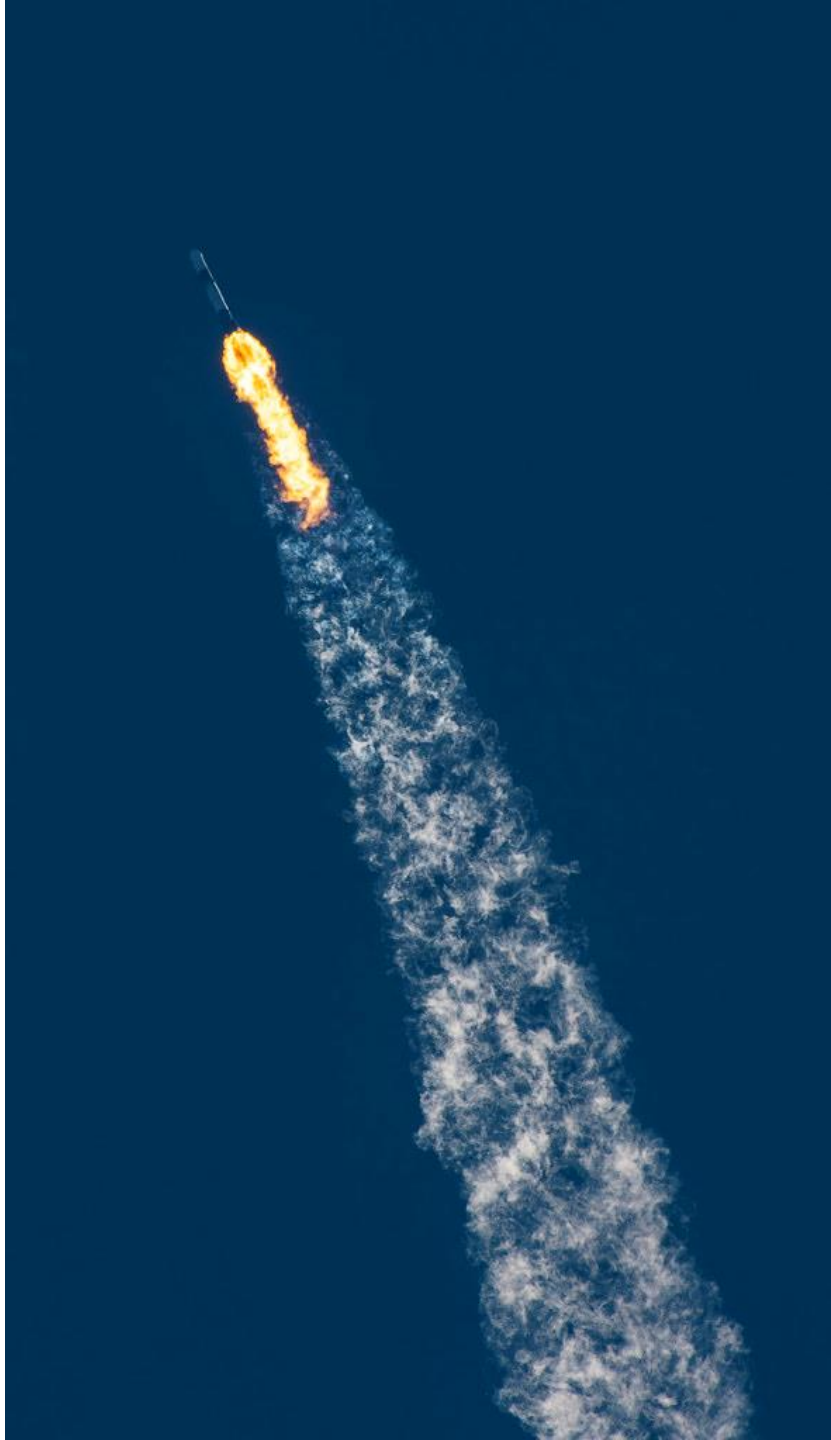


EDA with Visualization

Payload Mass vs. Launch Site

- Success Rate typically rises with the payload mass.
- VAFB SLC 4E hasn't launched rockets with payloads over than 10000kg.
- KSC LC 39A achieves 100% success rate for launches below 5500 kg.

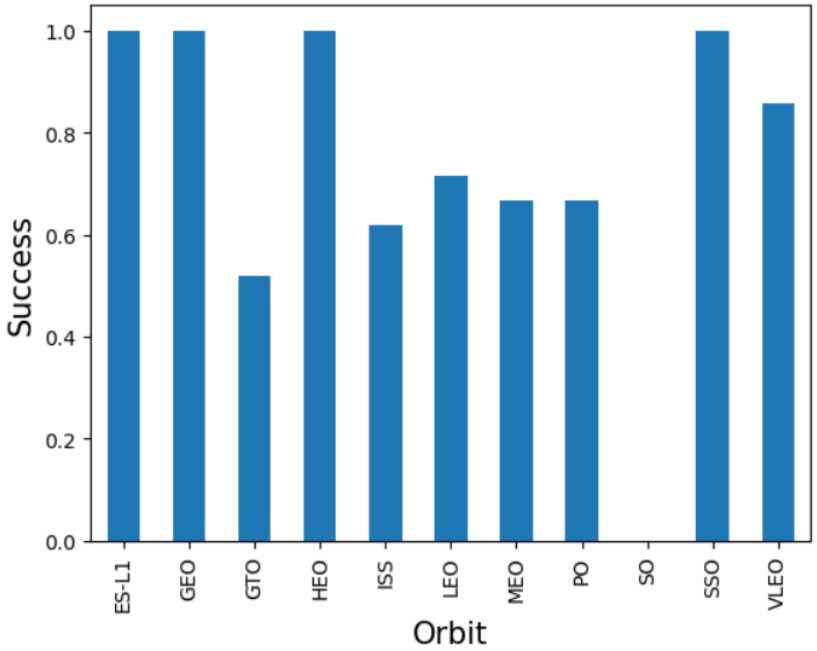




EDA with Visualization

Success Rate vs. Orbit

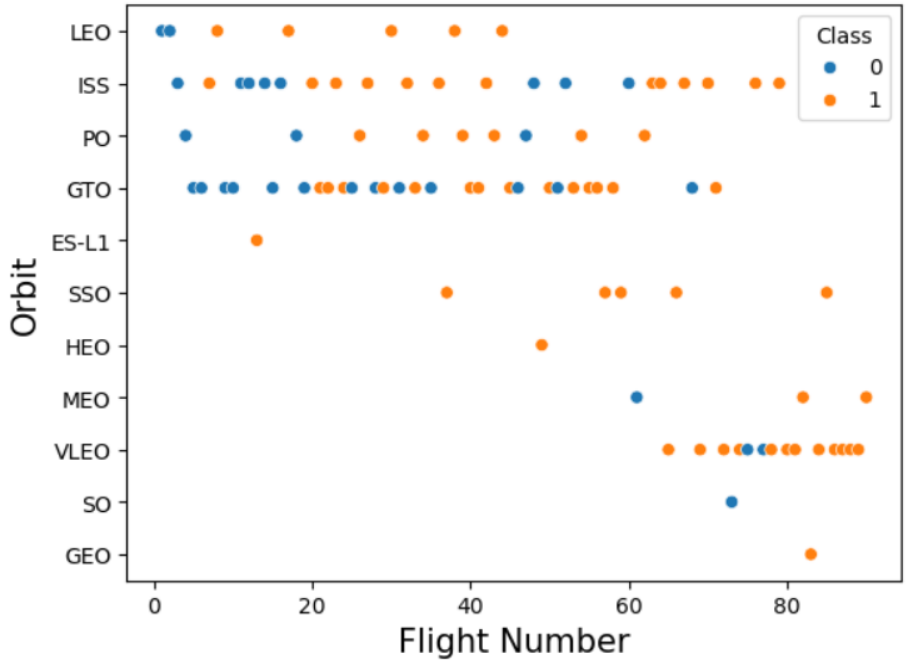
- ES-L1, GEO, HEO and SSO demonstrate 100% success rate.
- GTO exhibits the lowest success rate.



EDA with Visualization

Orbit vs. Flight Number

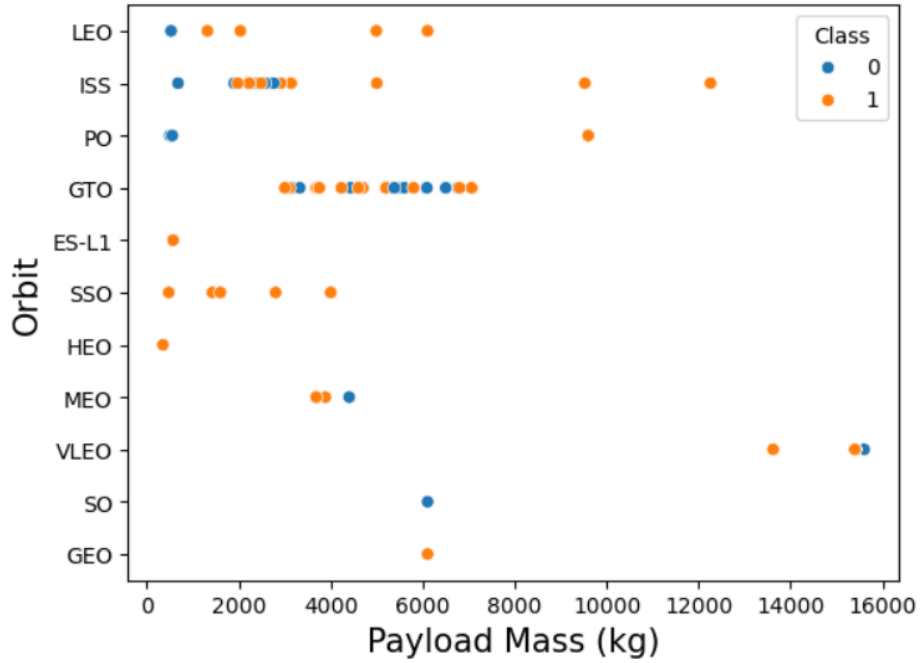
- Success rate increases with flight frequency.
- GTO is an exception to this trend.



EDA with Visualization

Orbit vs. Payload Mass

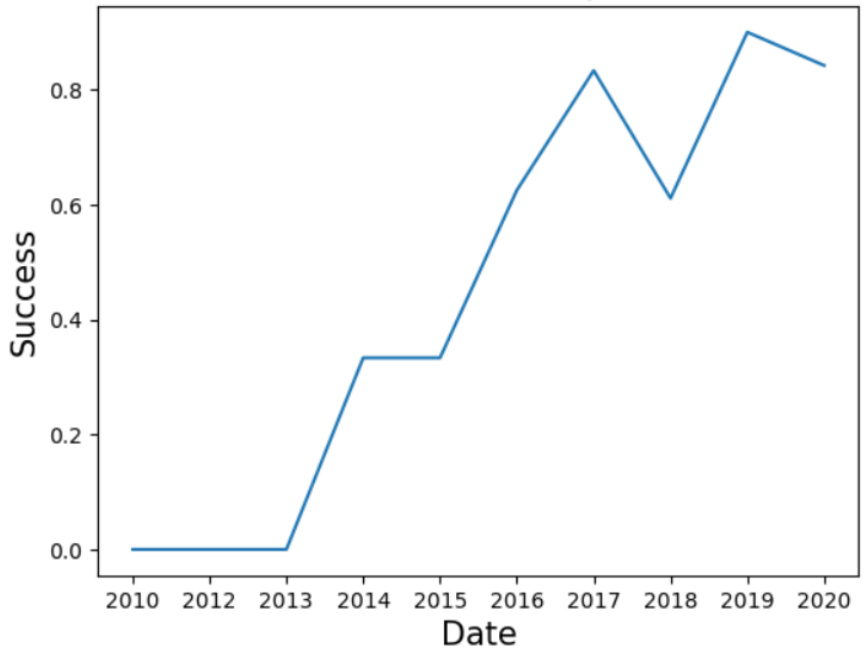
- Success rate increases with payload for LEO, ISS and PO.
- GTO is an exception to this trend.



EDA with Visualization

Launch Success Yearly Trend

- Success rate demonstrates an upward trend over time.
- Recent launches achieve an approximate 80% success rate.
- A special decline in success rate is noted between 2017 and 2018.



EDA with SQL

Launch Site Names

SQL QUERY

- SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE

QUERY EXPLANATION

- The query retrieves unique launch sites within the “SPACEXTABLE” table.

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

EDA with SQL

Launch Sites beginning with ‘CCA’

SQL QUERY

- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE ‘CCA%’ LIMIT 5

QUERY EXPLANATION

- The query retrieves the records from the “SPACEXTABLE” table where “Launch_Site” starts with ‘CCA’. It limits the output to the first 5 records that match this condition.

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

EDA with SQL

Total Payload Mass carried by Nasa (CRS)

SQL QUERY

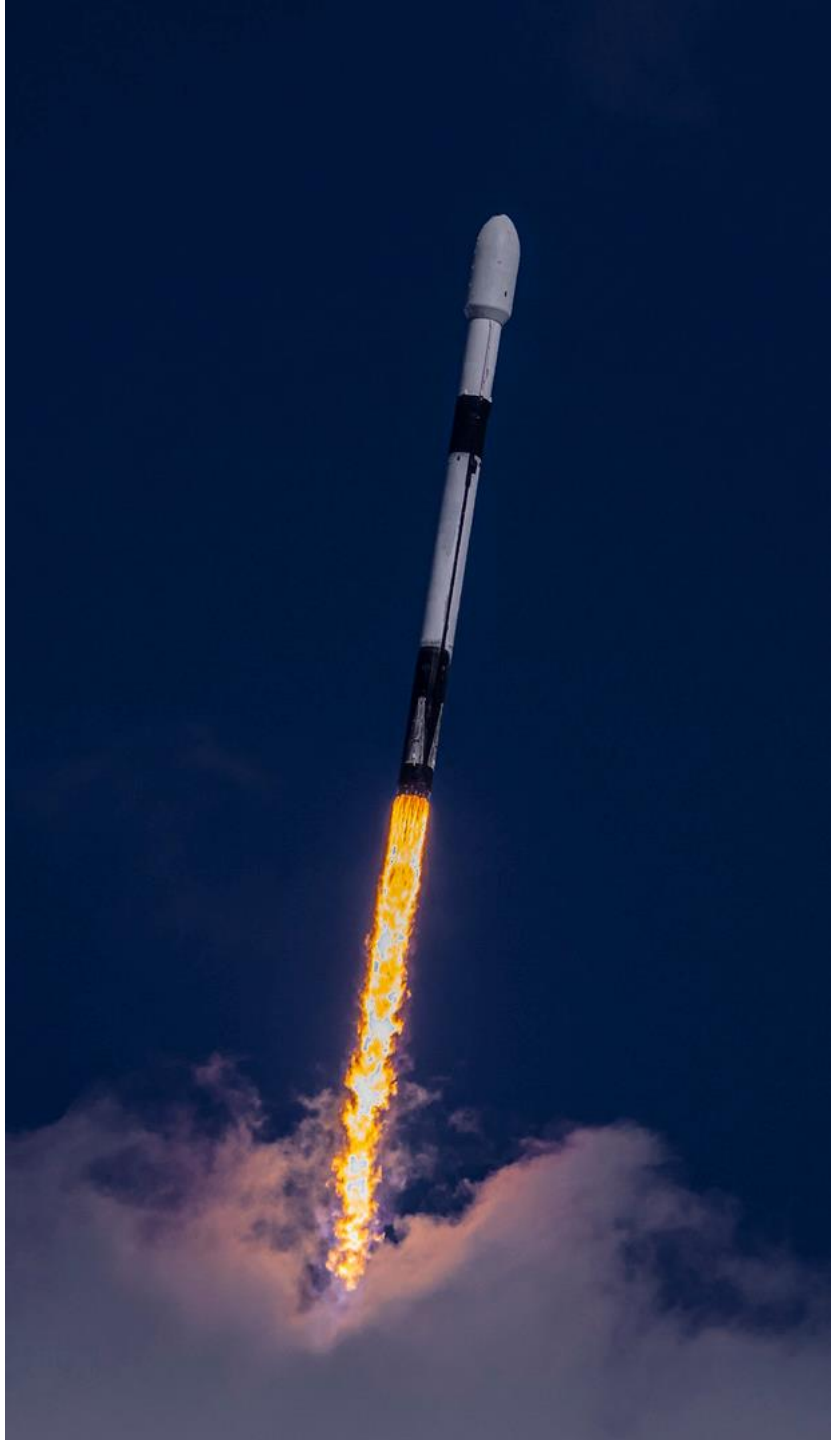
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE

WHERE Customer == ‘NASA (CRS)’

QUERY EXPLANATION

- The query retrieves the total payload mass from the “SPACEXTABLE” table for records where “Customer” is ‘NASA (CRS)’.

SUM(PAYLOAD_MASS__KG_)
45596



EDA with SQL

Average Payload Mass carried by F9 v1.1

SQL QUERY

- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Booster_Version LIKE ‘F9 v1.1’

QUERY EXPLANATION

- The query computes the average payload mass in the “SPACEXTABLE” table for records where “Booster_Version” contains ‘F9 v1.1’.

AVG(PAYLOAD_MASS__KG_)
2928.4

EDA with SQL

Date of the First Successful Landing on Ground Pad

SQL QUERY

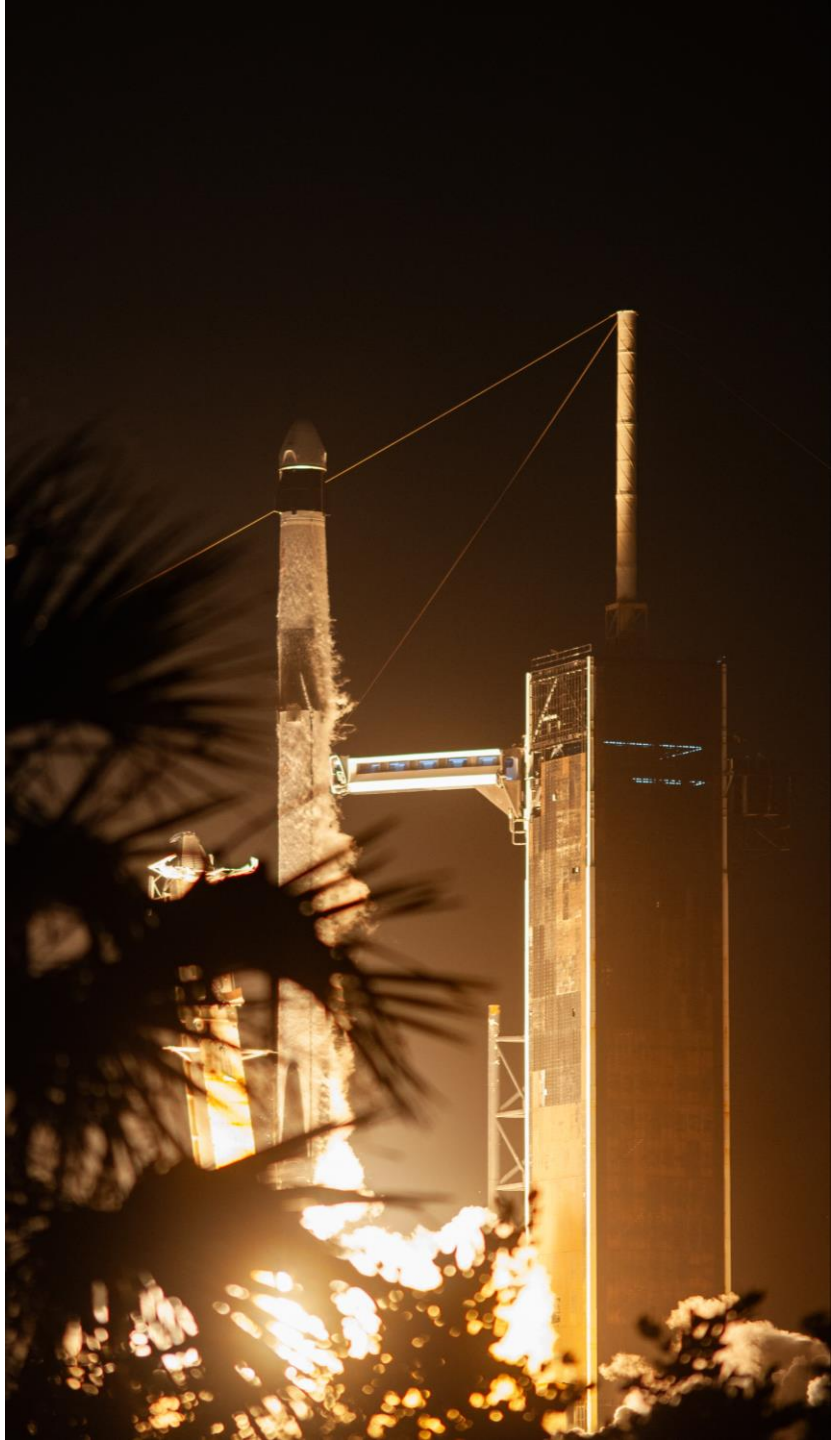
- SELECT MIN(Date) FROM SPACEXTABLE

WHERE Landing_Outcome == ‘Success (ground pad)’

QUERY EXPLANATION

- The query retrieves the minimum date value from the “SPACEXTABLE” table for the records where “Landing_Outcome” is ‘Success (ground pad)’.

MIN(Date)
2015-12-22



EDA with SQL

Booster Versions with Drone Ship Success and Payload Mass between 4000kg and 6000 kg

SQL QUERY

- SELECT Booster_Version FROM SPACEXTABLE

WHERE Landing_Outcome == 'Success (drone ship)' AND
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

QUERY EXPLANATION

- The query retrieves the booster versions from the "SPACEXTABLE" table for records where the "Landing_Outcome" is 'Success (drone ship)' and the "PAYLOAD_MASS__KG_" is between 4000 and 6000.

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

EDA with SQL

Total Number of Successful and Failure mission outcomes

SQL QUERY

- SELECT MISSION_OUTCOME, COUNT(*) as “TOTAL NUMBER OF LAUNCHES” FROM SPACEXTABLE GROUP BY MISSION_OUTCOME

QUERY EXPLANATION

- The query groups mission outcomes and retrieves the count of records associated with each outcome from the “SPACEXTABLE” table.

Mission_Outcome	TOTAL NUMBER OF LAUNCHES
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

EDA with SQL

Booster Versions which have carried the maximum Payload Mass

SQL QUERY

- SELECT Booster_Version FROM SPACEXTABLE

WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

QUERY EXPLANATION

- The query retrieves the booster versions from the “SPACEXTABLE” table for the records that have the maximum payload mass.

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

EDA with SQL

2015 Launch Records

SQL QUERY

- SELECT SUBSTR(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE SUBSTR(Date,0,5)='2015' AND
Landing_Outcome == 'Failure (drone ship)'

QUERY EXPLANATION

- The query retrieves the month extracted from "Date", along with the values in "Landing_Outcome", "Booster_Version" and "Launch_Site" from the "SPACEXTABLE" table for the records from 2015 and where "Landing_Outcome" is 'Failure (drone ship)'.

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

EDA with SQL

Rank the count of Landing Outcomes between 2010-06-04 and 2017-03-20

SQL QUERY

- SELECT Landing_Outcome, Count(*) as “Number of successful landings”

FROM SPACEXTABLE GROUP BY Landing_Outcome
HAVING Date BETWEEN ‘2010-06-04’ AND ‘2017-03-20’
ORDER BY “Number of successful landings” DESC

QUERY EXPLANATION

- The query groups landing outcomes and retrieves the count of records associated with each landing outcome from the “SPACEXTABLE” table for the records between 2010-06-04 and 2017-03-20. The results are ordered in descending order by the count value.

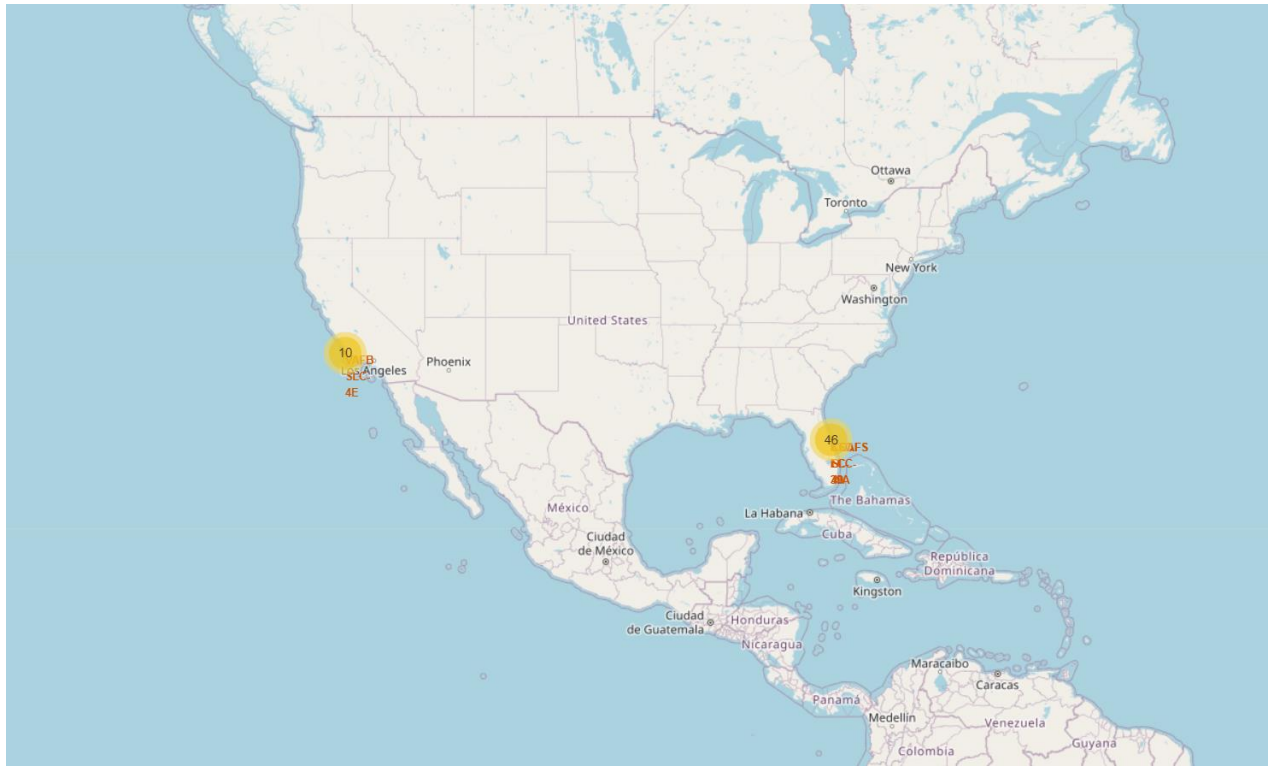
Landing_Outcome	Number of successful landings
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Folium Interactive Map

Folium Interactive Map

Launch Site Circles and Markers

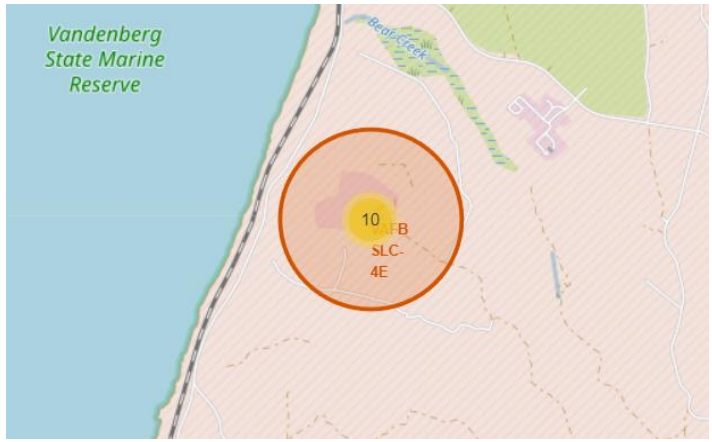
- Launch sites are strategically positioned closer to the Equator to capitalize on the Earth's rotational speed, optimizing rocket launches for efficiency.

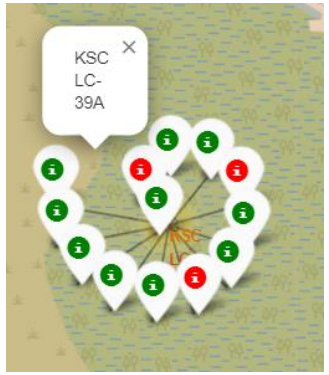


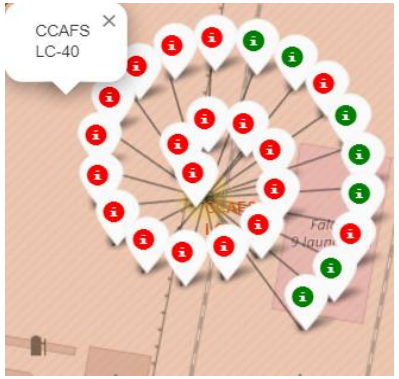
Folium Interactive Map

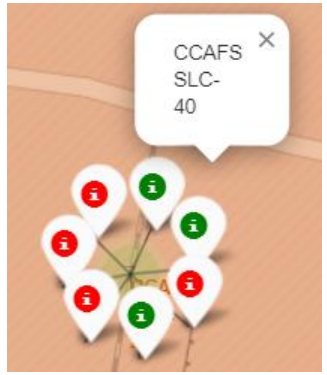
Rocket Launches Color Labeled Markers

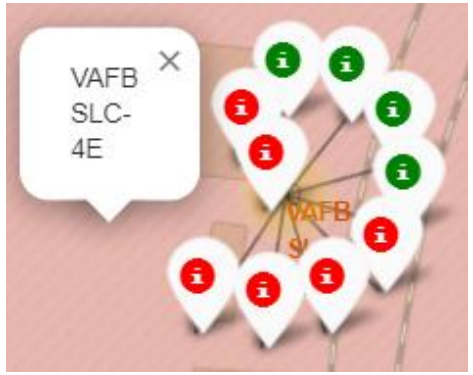
- Successful launches are represented in **Green** and Failures in **Red**.









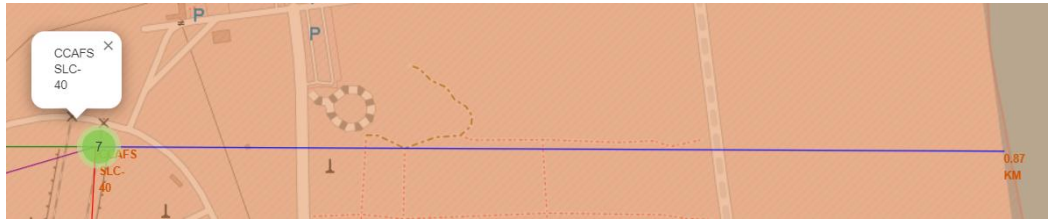


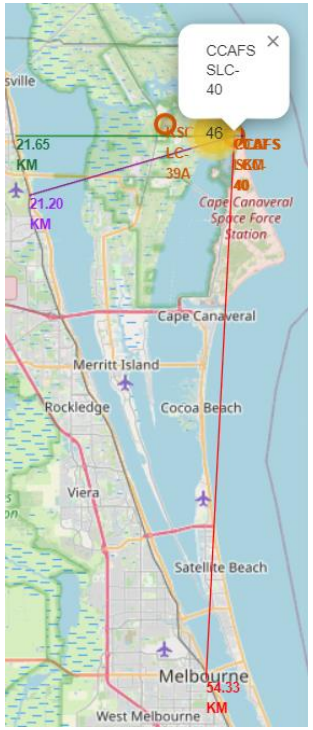
Folium Interactive Map

Distance to Landmarks – CCAFS SLC-40

Each line color on the map corresponds to a specific landmark, with the following colors represented:

- **Coastline:** 0.87km
- **Railway:** 21.20km
- **Highway:** 21.65km
- **City:** 54.33km



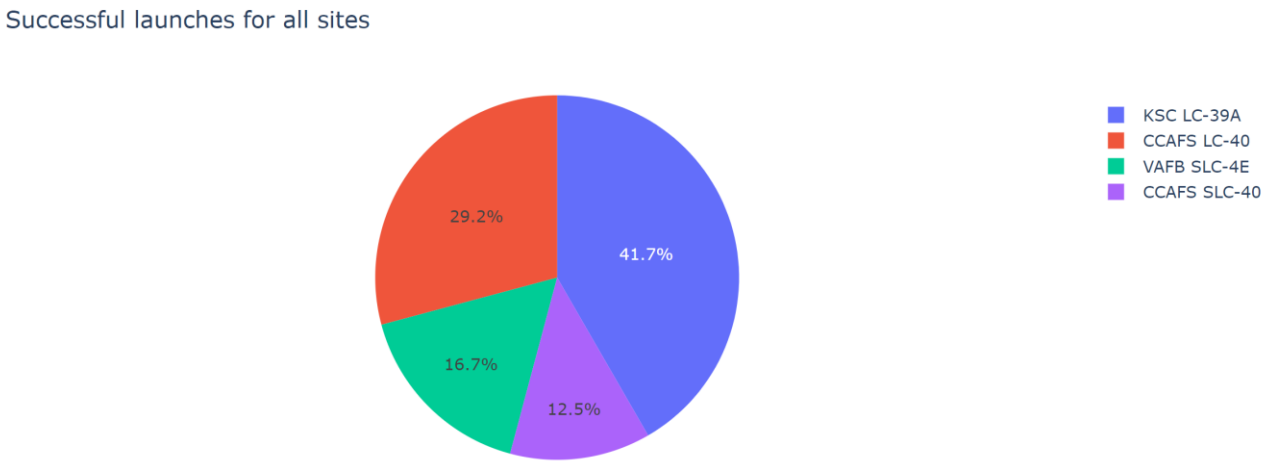


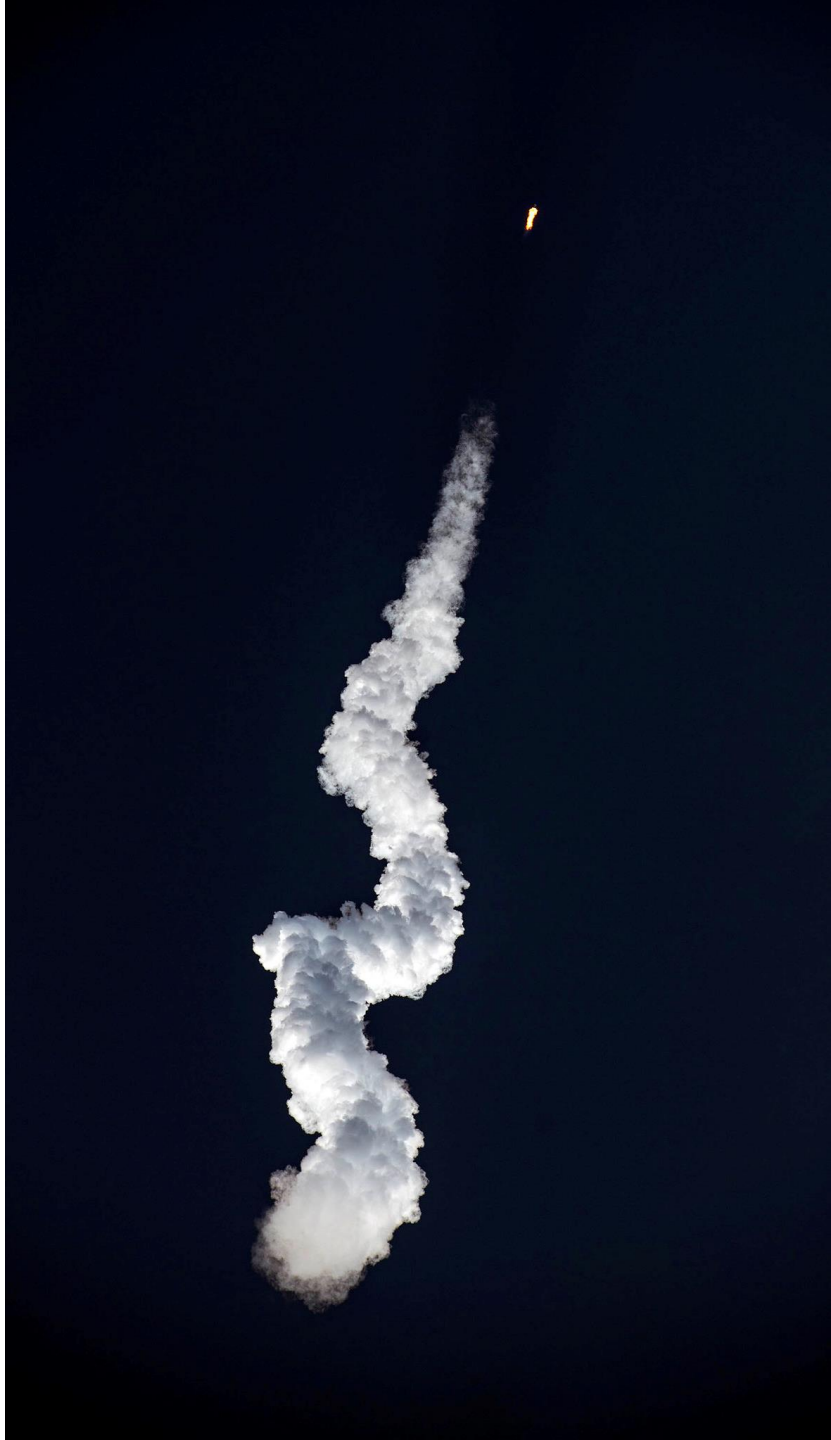

Plotly Dash Dashboard

Plotly Dash Dashboard

Launch Success by Site

- KSC LC-39A is the Launch Site with the highest number of successful launches.

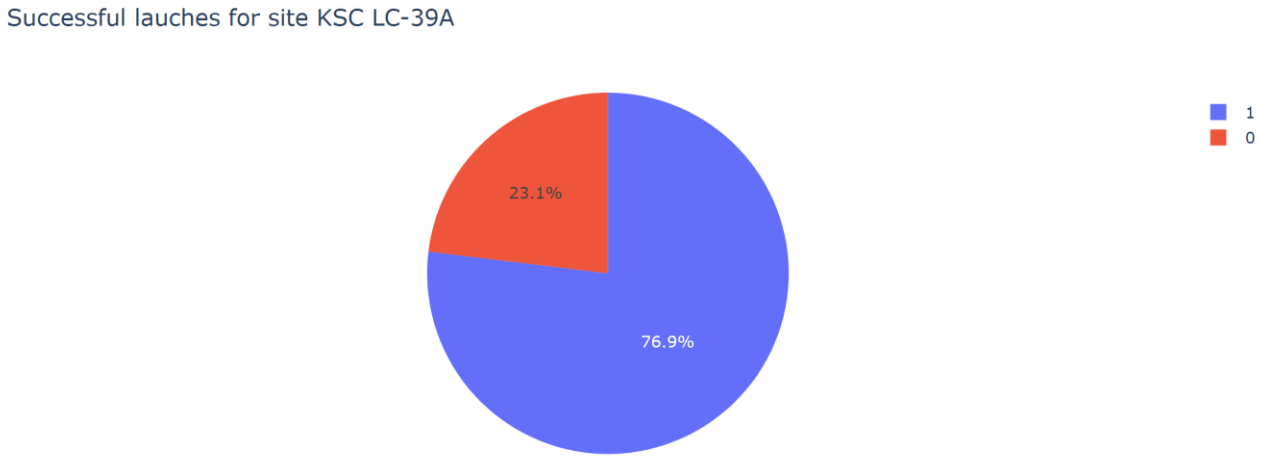




Plotly Dash Dashboard

Launch Success in KSC LC-39A

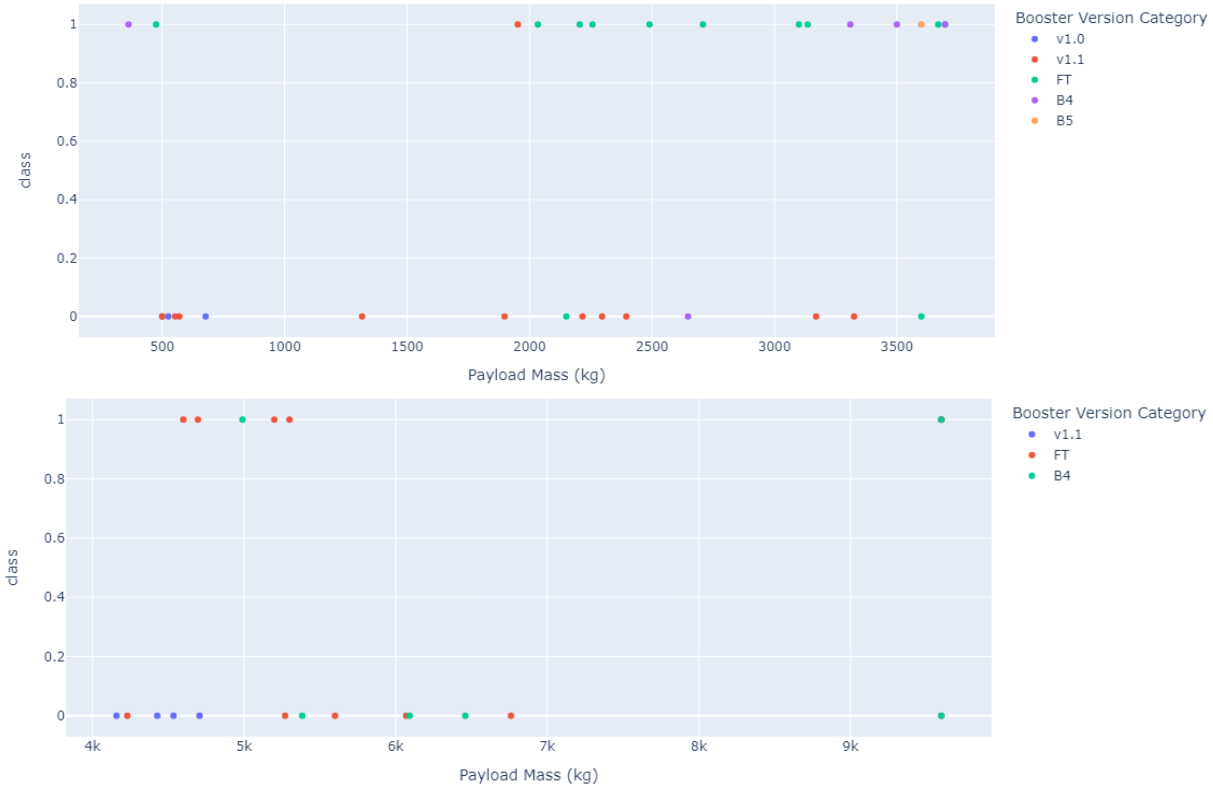
- KSC LC-39A boasts a Success Rate of 76.9%.



Plotly Dash Dashboard

Payload Mass vs. Outcome by Booster Version

- Lightweight payloads (0kg-4000kg) demonstrate a higher success rate compared to heavier payloads (4000kg-10000kg).
- Boosters B5 and FT exhibit higher success rates.

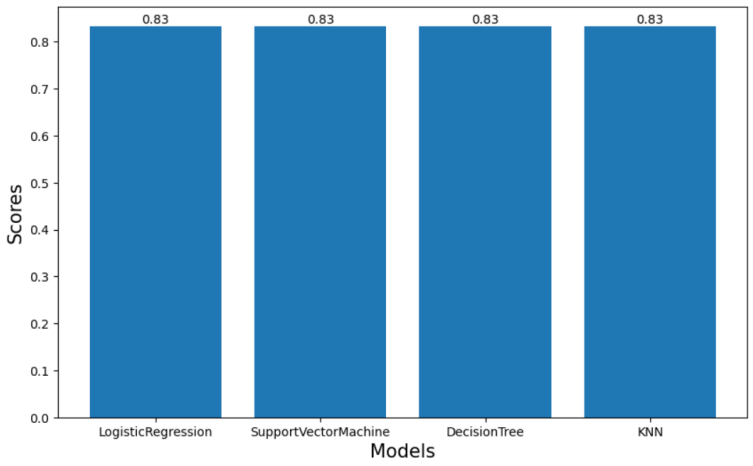


Predictive Analysis

Predictive Analysis

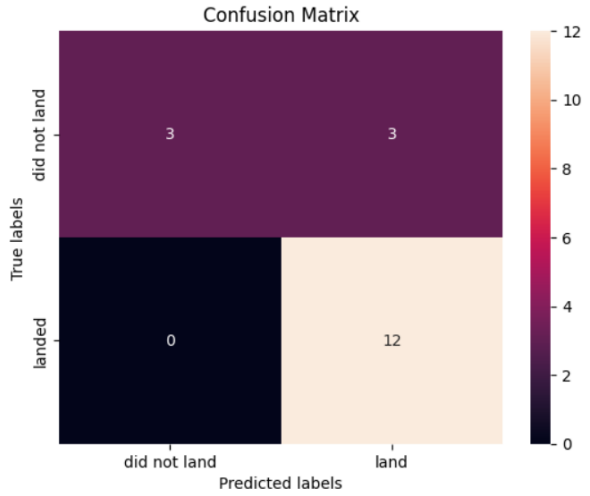
Classification Accuracy

- All models exhibit identical test scores. For this reason, we will opt for the Logistic Regression model due to its simplicity. However, it's advisable to retrain the models with larger datasets to improve their resilience to generalization.



Confusion Matrix

- The model performs well in predicting successful landings but struggles with predicting unsuccessful landings (false positives). Increasing the size of our datasets can address this issue.



Conclusion

Summary

- Our Logistic regression model achieves an accuracy of 83.3% in predicting launch outcomes.
- Light-weighted payloads exhibit higher success rates compared to heavy-weighted payloads.
- KSC LC-39A stands out as the launch site with the highest number of successful launches.
- Orbits ES-L1, GEO, HEO, and SSO boast a 100% success rate.

Considerations

- Improving the accuracy of our predictive model can be achieved by utilizing larger datasets and exploring alternative models such as neural networks.

Complete Dashboard View