SPACEX FALCON 9

First Stage Analysis

Daniel Hoppo

4th June 2024

Outline

Executive Summary

- This project presents a comprehensive framework for estimating the probability of successful Falcon 9 first stage landings through predictive modeling.
- Utilising data collected via web scraping from Wikipedia and the SpaceX REST API, alongside data wrangling techniques, the analysis incorporates exploratory data analysis (EDA) including correlation analysis to reveal interesting patterns and relationships.
- Interactive visual analytics using Folium and Plotly Dash provide intuitive insights into the data.
- Predictive models, including SVM, Classification Trees, and Logistic Regression, were developed and evaluated using metrics such as Confusion Matrix and Accuracy Scores.
- By providing robust predictive models and actionable insights, this project contributes to enhancing SpaceX's cost estimation processes and enables other companies to competitively bid for launches.

Introduction

Project Background

- In this project we take a dive into SpaceX's reusable rocket technology that has revolutionised space travel, specifically focusing on the Falcon 9. SpaceX has the capability to reuse the first stage of the rocket if a successful return to Earth is accomplished
- This reusability relies on the first stage landing successfully with the help of cutting-edge technologies such as landing legs and grid fins. By analysing this information, it can be used to determine the cost of a launch and provide insights for alternative companies to competitively bid against a SpaceX launch

Problem Statement

- SpaceX's ability to reuse the Falcon 9's first stage means significant reduction in launch costs compared to expendable rockets. The problem is not every landing is successful leading to variability in mission costs. Predicting the likelihood of a successful landing is important for improving cost estimations and planning future missions more effectively

Introduction

Nature of the Analysis

In this data science project, we aim to develop a predictive model to determine the likelihood of successful landings for the Falcon 9 first stage. This quantitative analysis seeks to address the problem of uncertainty in landing outcomes, which significantly impacts cost estimation and strategic planning for space missions

Research Questions for Analysis

1. What are the key factors that influence the success of a Falcon 9 first stage landing?
2. How accurately can we predict the landing outcome based on historical data?
3. What patterns or trends can be identified from past launches that correlate with successful landings?

METHODOLOGY

Methodology

Data Collection

Web scraping of Falcon 9 launches from Wikipedia using BeautifulSoup

Gathered data using the SpaceX REST API

Data Wrangling

Cleansed the data

Mean imputation technique

Class label creation

Performed one-hot encoding

Standardisation of data

Exploratory Data Analysis

Bivariate analysis using scatterplots and line charts

Correlation analysis using Seaborn and SQL

Visual Analytics

Folium Library to explore geospatial data

Interactive dashboard created with Plotly Dash

Predictive Analysis

Binary classification models built

Support Vector Machines, Decision Trees, K-Nearest Neighbours and Logistic Regression

Evaluation metrics using Confusion Matrices and Accuracy Scores

Data Collection

SpaceX REST API

METHOD:
Performed GET request on endpoint api.spacexdata.com/v4/launches/past

LAUNCH DATA:
rocket name, payload, launch & landing specifications, landing outcome

BASE URL:
api.spacexdata.com/v4/

ENDPOINTS:
/launchpad /cores /payloads /rocket

ADDITIONAL DATA:
orbit type, grid fins, landing pad, legs used, launch site name

Web Scraping Wikipedia

TOOLS:
BeautifulSoup in Python used to parse and extract the relevant launch records **data**

URL:
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

PROCESS:
Parsed the launch HTML tables and created a data frame with relevant launch records

Data Collection: SpaceX API

SpaceX REST API

GET Request
api.spacexdata.com/v4/launches/past

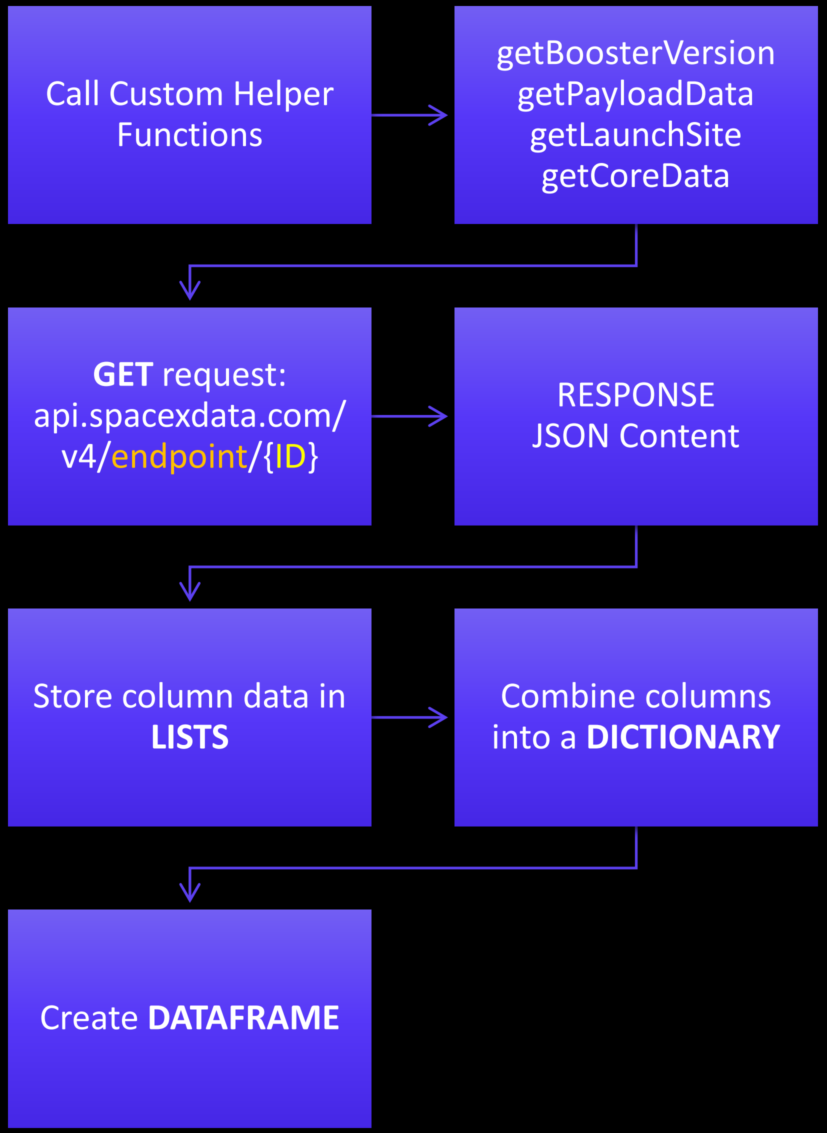
RESPONSE
STATUS 200

Decode RESPONSE
Content to **JSON** Data

Normalise to
Pandas **Dataframe**

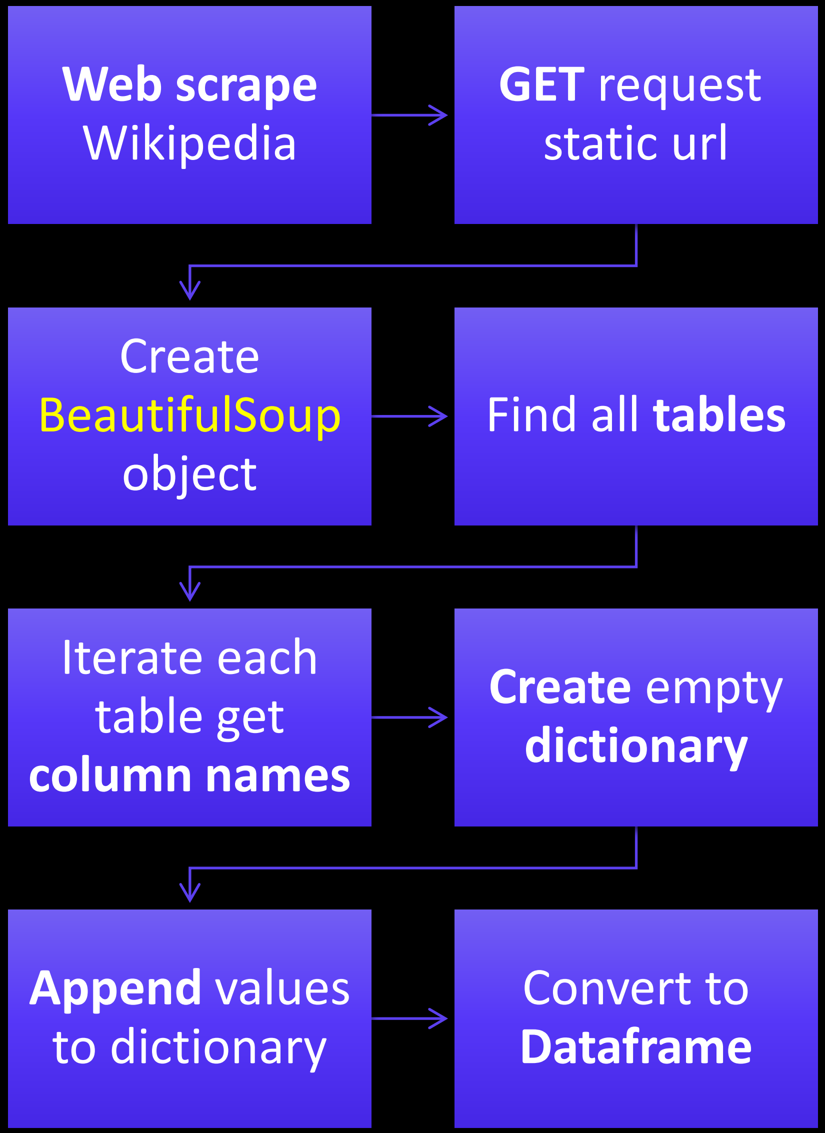
Data Collection:
Custom Functions

- **Custom Helper Functions** were created to assist in extracting additional valuable information using the API
- Performed GET request on each endpoint to extract additional launch data for each ID
- **Endpoints include**:
  - /launchpad
  - /cores
  - /payloads
  - /rockets



Data Collection:
Web Scraping

- Web scraping of Falcon 9 launch records was performed with BeautifulSoup
- We extracted the Falcon 9 launch records only from the HTML table on Wikipedia site
- Parsed the table and converted it into a Pandas data frame



Data Wrangling

- Filtered the original data frame to only Falcon 9 launches
- Calculated the number of launches per site
- Calculated the number and occurrence of each orbit
- Determined the different types of landing outcomes
- Created a landing outcome label from Outcome column
- Mean imputation technique on missing values
- One-hot encoded categorical variables

EDA with Visualisation

Performed exploration data analysis utilising **Seaborn** library:

- **Categorical & Scatter plots** - created for comparing relationships involving categorical variables. Such as Launch Site vs Flight Numbers
- **Line plots** - created to evaluate the successful yearly trend from 2010 to 2020
- **Bar plots** - utilised for analysing the average Success Rate by Orbit type

EDA with SQL

- We executed SQL queries within Jupyter Notebook to derive valuable insights
- **Queries performed include:**
  - Display the **names** of the **unique launch sites** in the space mission
  - Display the **total payload mass** carried by boosters launched by NASA (CRS)
  - Display **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 > 4000 and < 6000 kgs
  - 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

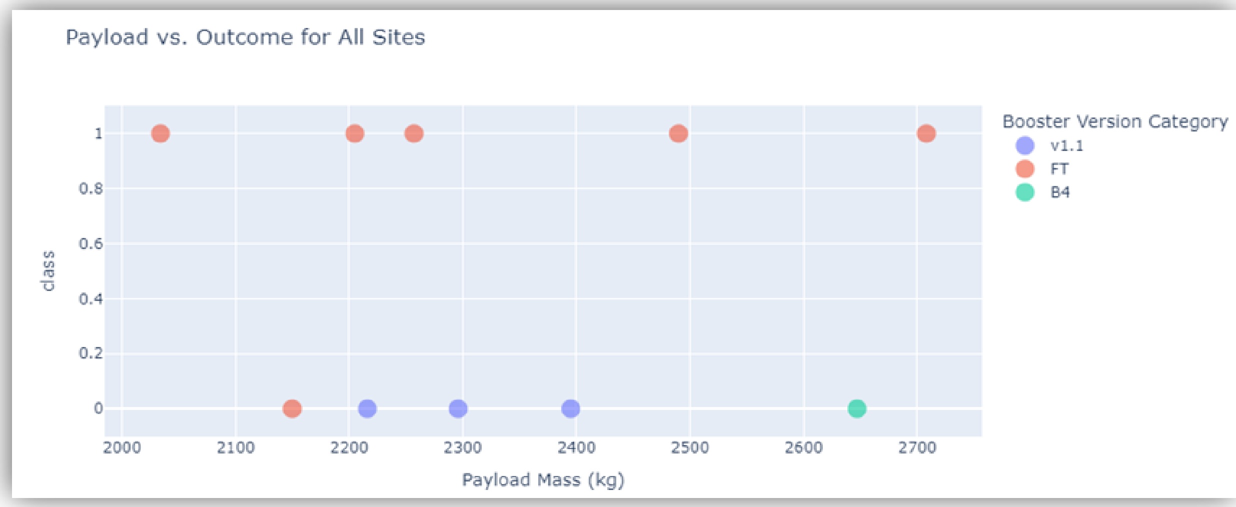
Interactive Map with Folium

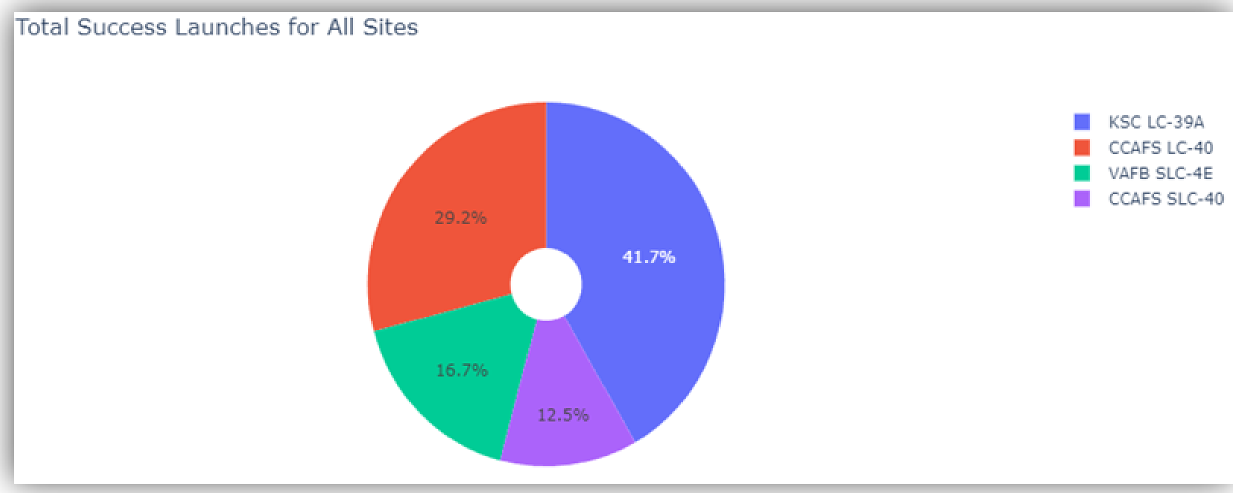
- We marked **all launch sites** on an interactive map using markers, popup labels for site names, and circles to highlight the locations
- Blue circle was used for NASA Johnson Space Center as the starting center location
- Marked the success/failed launch outcomes for each site on the map using colour-coded **marker clusters**
- Calculated the **distances** between a launch site to its proximities, such as railway, cities, highway, and coastlines
- We drew a line between launch site and proximities with a distance label in kilometers

Dashboard with Plotly Dash

The following visuals were added to the interactive Dashboard:

- **Pie** charts illustrate effectively the total success launches with sites selected by a dropdown list
- **Scatter** plots display the relationship between weighted Payload Mass in kilograms and their outcomes for respective **Booster Versions** (Hue semantic)
  - Payload amounts made adjustable using an interactive slider





Predictive Analysis

- We created a **machine learning pipeline** to predict if the first stage will land successfully
  - Standardised the data using scikit-learn preprocessing
  - Split the data into 80% training and 20% test data
  - Built Logistic Regression, KNN, SVM and a Decision Tree model
  - Utilised GridSearchCV to identify the optimum hyperparameters
  - Evaluated the performance of each model using accuracy scores and assessment of the Confusion Matrix
- We found the best performing classification model by using the max function on the list of accuracy score tuples

Results

Exploratory Data Analysis

- Most successful launch site is KSC LC-39A
- Yearly upward trend in successful launches
- FT Booster Version achieves very high success rate on low payloads
- The sun-synchronous orbit has the highest success rate with 100% for all 5 launches
- ES-L1, GEO, and HEO orbits also have 100% success but only performed one launch

Visual Insights

- All launch sites are situated near coastlines and maintain a safe distance from cities
- Railways & Highways are in proximity to the launch sites but are not directly adjacent

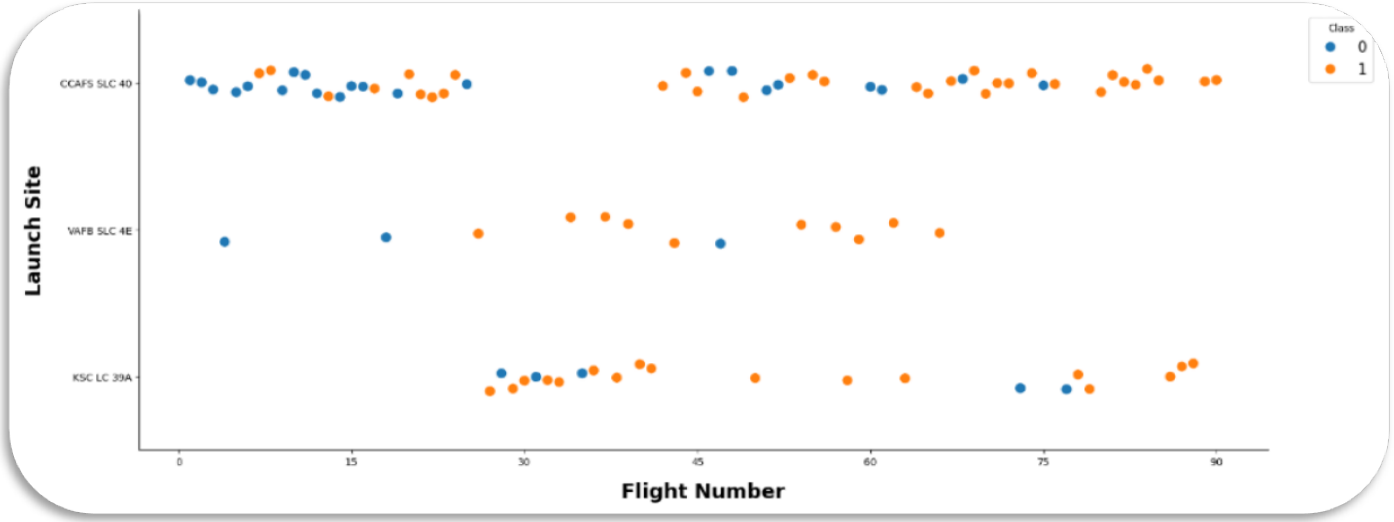
Predictive Highlights

- Best performing model was the Decision Tree classifier
- Logistic Regression, SVM, & KNN out-of-sample accuracy achieved 83.33%
- Decision Tree achieved an F1-Score of 96%
- Models correctly identified all positive outcomes with Recall metric reporting 100.0%

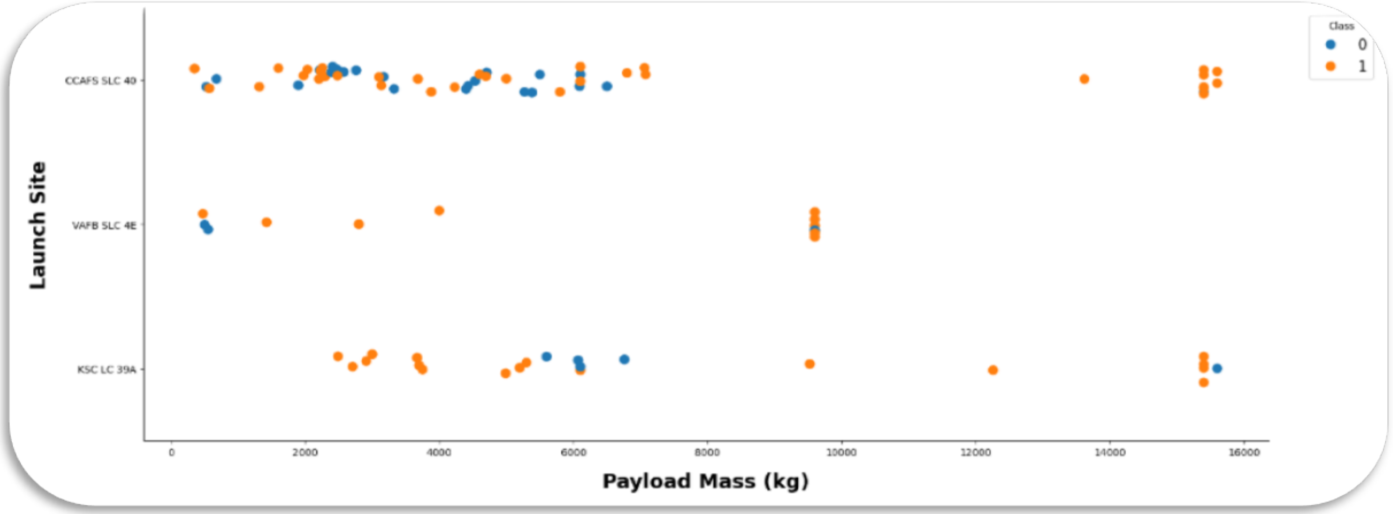
KEY INSIGHTS FROM **EXPLORATORY DATA ANALYSIS**

Flight Number vs. Launch Site

- More **frequent launch usage** at the CCAFS SLC 40 site
- Noticeable **increase** in **success rate** as flight numbers increase
- The last five launches for each site achieved a successful outcome

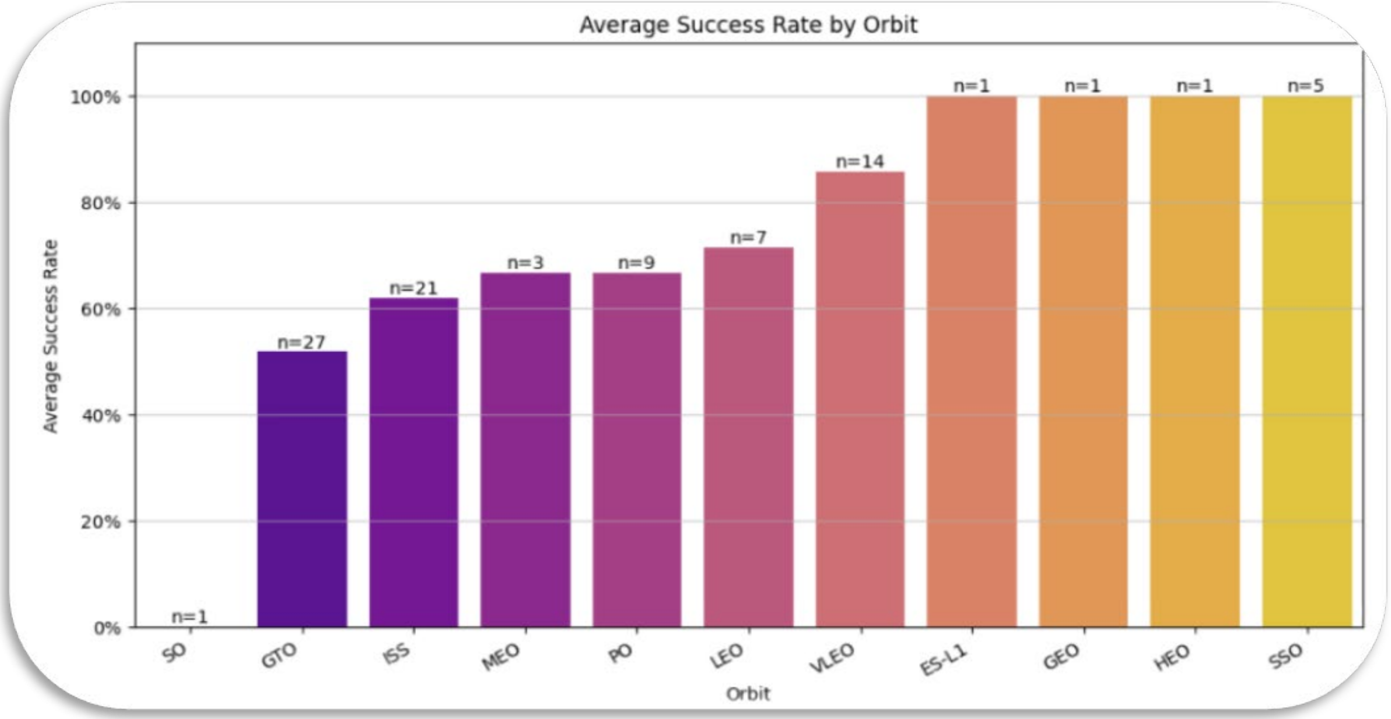


Payload vs. Launch Site



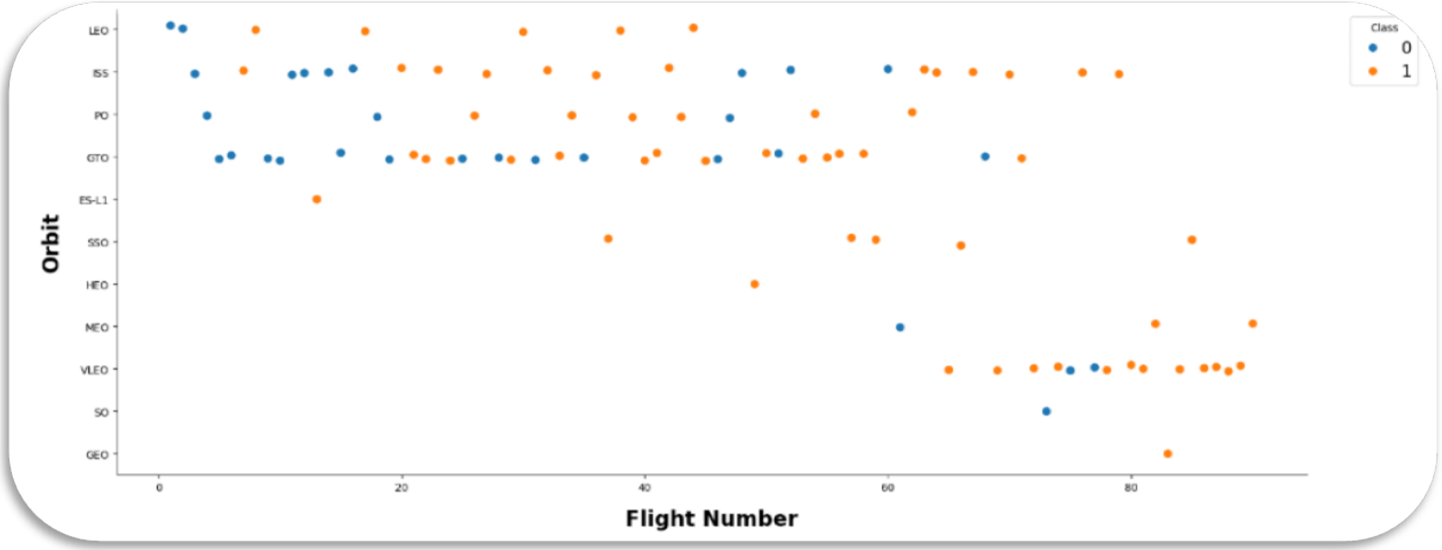
- Heavy payloads typically have higher success rate than lower payload launches
- Launches from VAFB SLC 4E site have never carried payloads exceeding 10,000 kg
- Among launches handling heavy payloads, only the KSC LC 39A site has experienced failures
- KSC LC 39A site experienced greater consistent success rate in the low payload range

Success Rate vs. Orbit Type



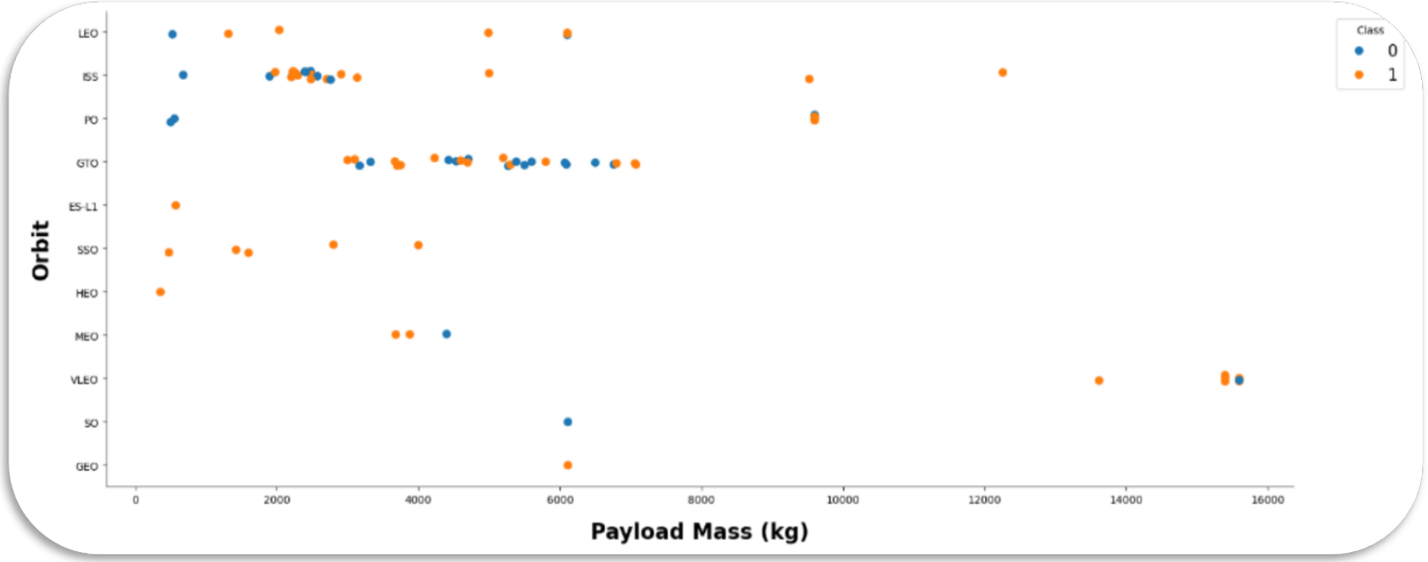
- The sun-synchronous orbit (SSO) has the highest success rate with 100% for all 5 launches
- ES-L1, GEO, and HEO orbits also have 100% success but only performed one launch

Flight Number vs. Orbit Type



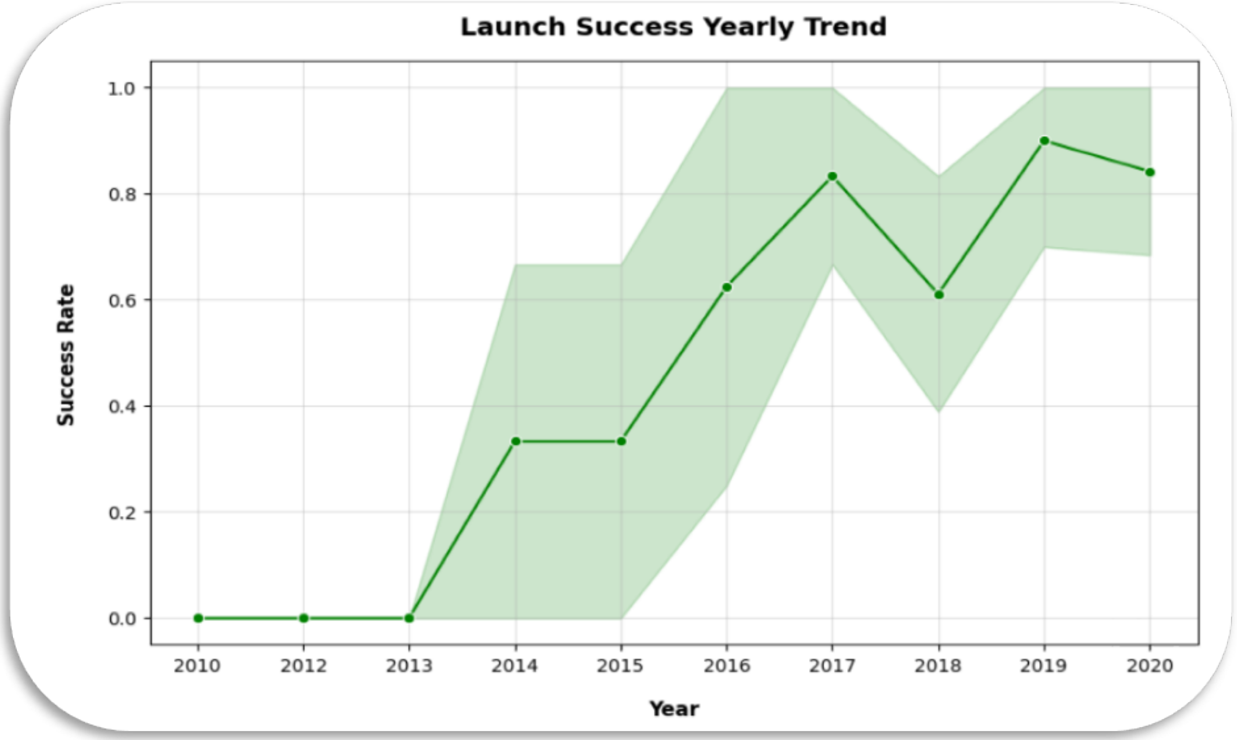
- Success in Low Earth Orbit (LEO) correlates with the frequency of flights
- Recent launches in VLEO and ISS orbits have displayed promisingly consistent high success rates
- Geostationary Transfer Orbit (GTO) shows no correlation between number of flights and success rate

Payload vs. Orbit Type



- SSO orbit is optimised for lower payloads with a perfect success rate
- LEO, ISS, and PO orbits tend to achieve higher success rate with mid-high payloads
- §GTO does not exhibit a clear correlation between success rate and payload mass

Launch Success Yearly Trend



- Overall the launch success rate has **increased** from 2013
- Between 2017-2018 there was an approx. 20% decrease in success rate

```
In [11]:  %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE

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

Out[11]:

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

All Launch Site Names

- **Unique launch site names:**
  - Cape Canaveral Air Force Station Launch Complex 40 (CCAFS LC-40)
  - Vandenberg Air Force Base Space Launch Complex 4 East (VAFB SLC-4E)
  - Kennedy Space Center Launch Complex 39A (KSC LC-39A)
  - Cape Canaveral Air Force Station Space Launch Complex 40 (CCAFS SLC-40)

```
%sql SELECT * FROM SPACEXTABLE 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_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

Launch Sites begin with 'CCA'

- Five launch site records beginning with CCA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

Total_Payload_Mass
45596

Total Payload Mass

- Total payload mass carried by boosters from NASA:

45596 kg

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

Avg_Payload_Mass
2928.4

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

  2928.4 kg

```
%sql SELECT MIN(Date) AS FirstSuccessDate FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

FirstSuccessDate
2015-12-22

First Successful Landing Date

- The first successful landing outcome date on ground pad:

  22nd December 2015

```
%%sql SELECT Booster_Version FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ > 4000
    AND PAYLOAD_MASS__KG_ < 6000
```

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

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

Successful Drone Ship Landing

- List of the Booster names which have successfully landed on drone ship carrying a payload mass > 4000 and < 6000 kg:

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

```
%%sql SELECT Mission_Outcome, COUNT(*) AS Total_Outcomes FROM SPACEXTABLE
    WHERE TRIM(Mission_Outcome) IN ('Success', 'Failure (in flight)', 'Success (payload status unclear)')
    GROUP BY TRIM(Mission_Outcome)
```

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

Mission_Outcome	Total_Outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Total Mission Outcomes

- The total number of successful and failure mission outcomes
- Ensured “Mission_Outcome” values were trimmed using SQL’s TRIM function to remove detected leading spaces in Success outcome values

Boosters Carrying Max Payload

```
%%sql SELECT Booster_Version FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

Booster names which have carried the maximum payload mass

```
%%sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Failure (drone ship)'
    AND substr(Date, 0, 5) = '2015'
```

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

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

2015 Launch Failure Records

- Failed outcomes landing on drone ship, their booster versions, and launch site names for year 2015

Landing Outcomes in Ranking Order

```
%%sql SELECT Landing_Outcome, COUNT(*) AS Total_Outcomes FROM SPACEXTABLE
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY Landing_Outcome
    ORDER BY Total_Outcomes DESC
```

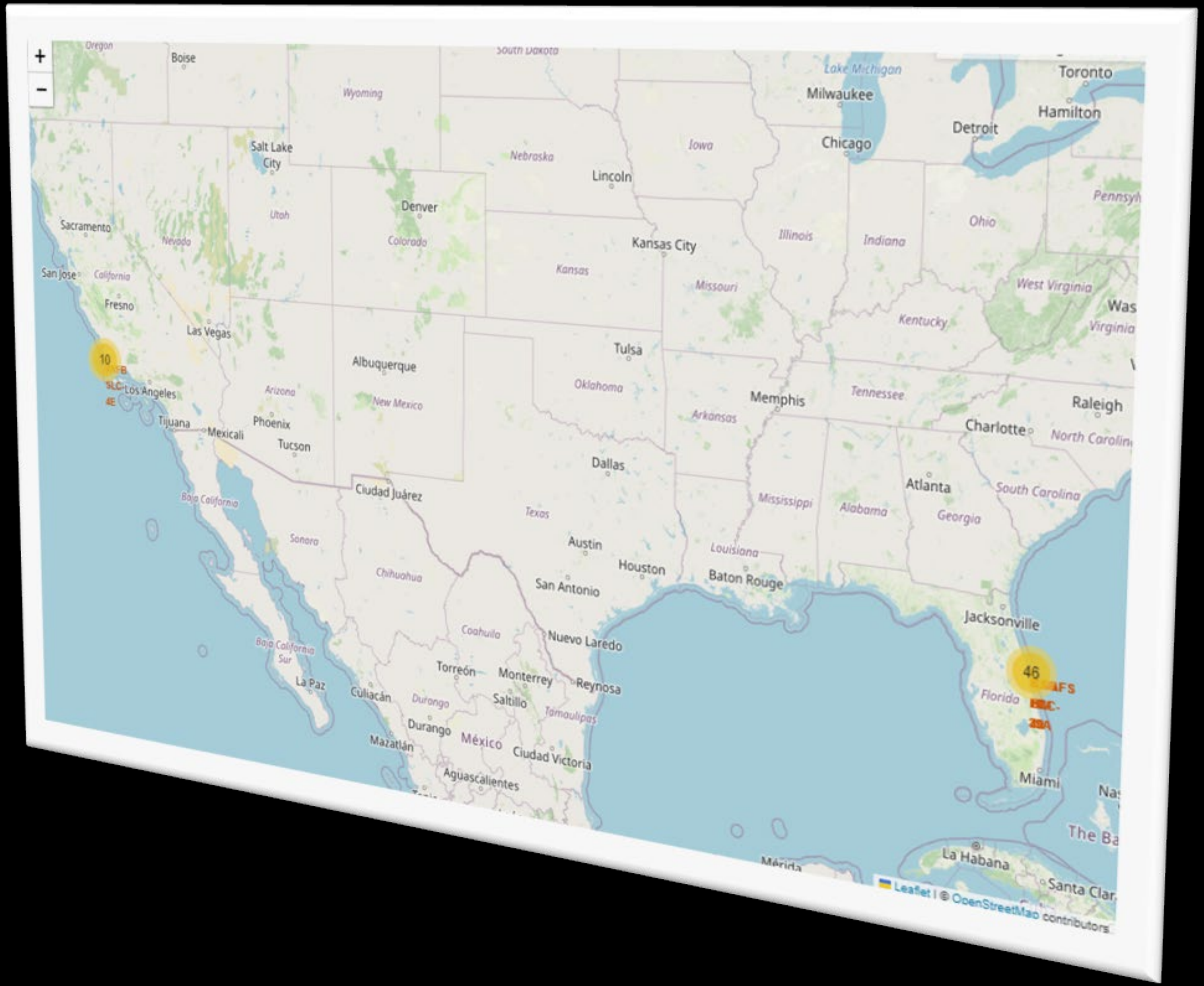
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Total_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

Ranking of the count of landing outcomes (e.g., Failure on a drone ship or Success on a ground pad) between the dates 2010-06-04 and 2017-03-20, in descending order

LAUNCH SITES PROXIMITIES ANALYSIS

All Launch Site Locations



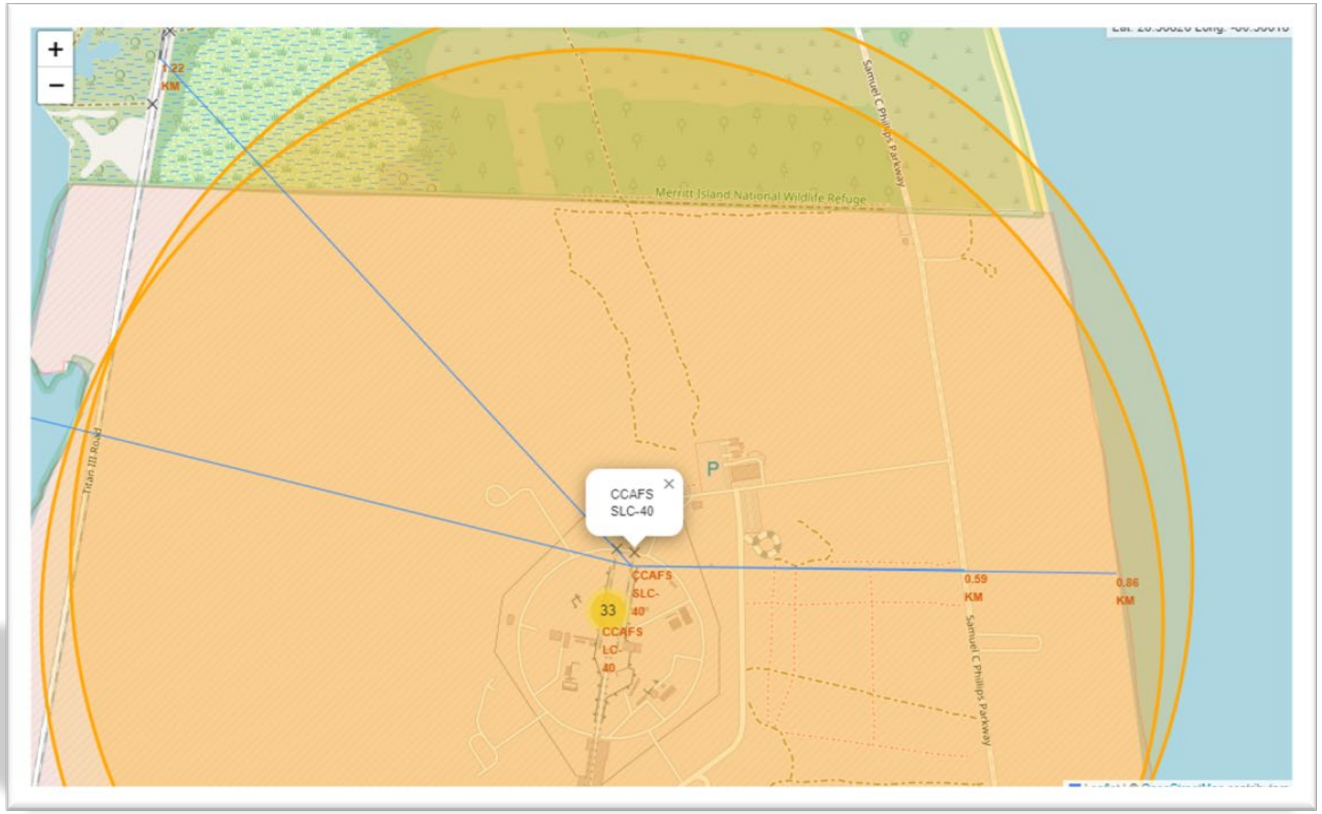
Sites are in **close proximity** to the **coastline**, allowing rockets to fly over the ocean, avoiding populated areas. Such as for the geosynchronous orbit (GTO)

Launch sites are situated close to the equator, benefiting from the Earth's rotational velocity and providing an extra boost to the rocket

Color-Labeled Launch Outcomes

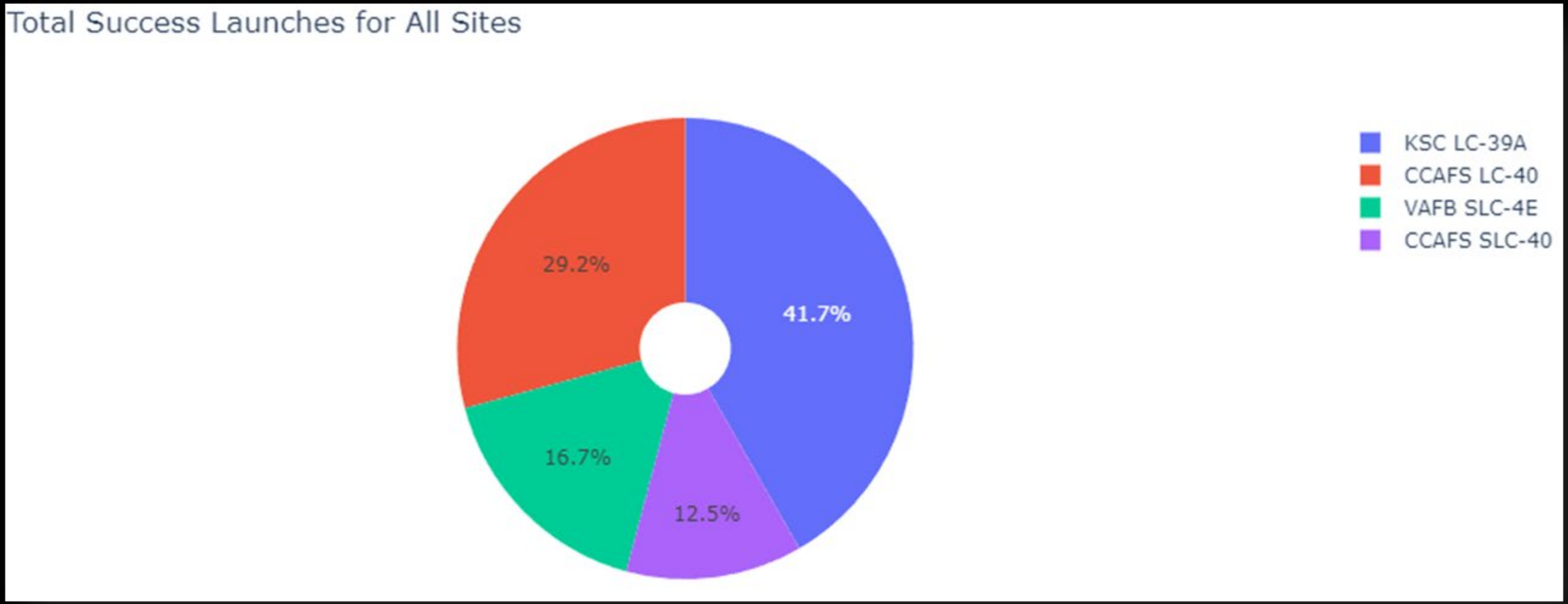
- Launch outcomes are grouped using a marker cluster in Folium, with green markers indicating success and red markers indicating failure

Launch Sites to its Proximities



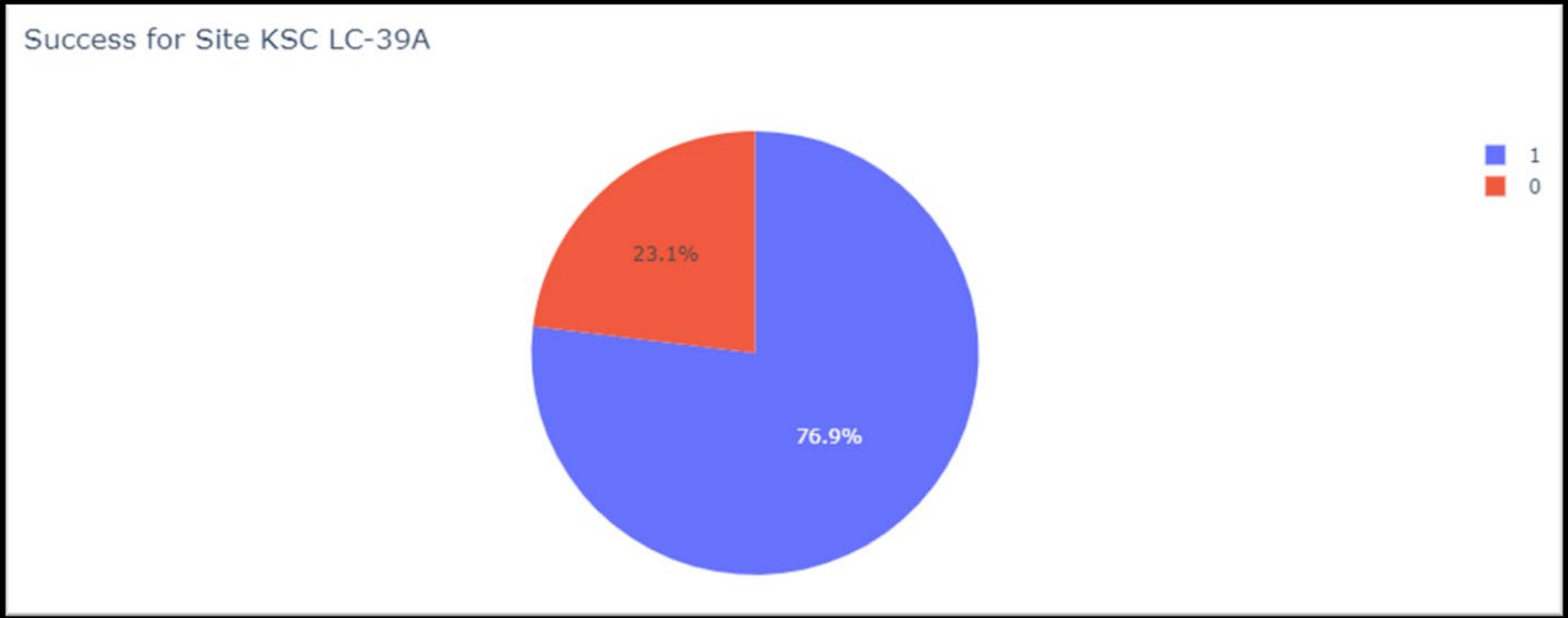
- The CCAFS LC-40 launch site is approx. 0.9 kilometers from the coastline and about 0.59 kilometers from the nearest highway
- Coastal locations often have better access to shipping and transportation routes, which is crucial for transporting large rocket components and supplies
- It is also safely distanced from major population areas, minimising risk to human life and property in the event of a launch anomaly

DASHBOARD WITH PLOTLY DASH



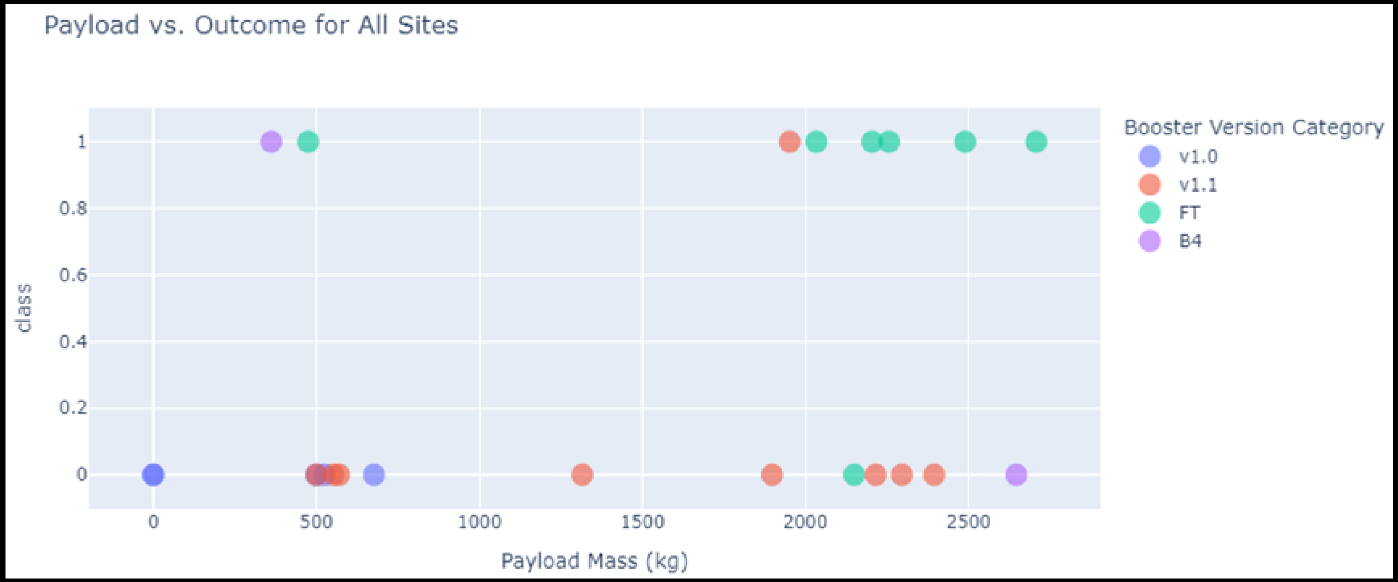
Total Successful Launches for All Sites

- **KSC LC-39A** site has achieved the most successful launches at 41.7% of the total launches
- **CCAFS SLC-40** launch site experienced the lowest amount of successful launches at 12.5%



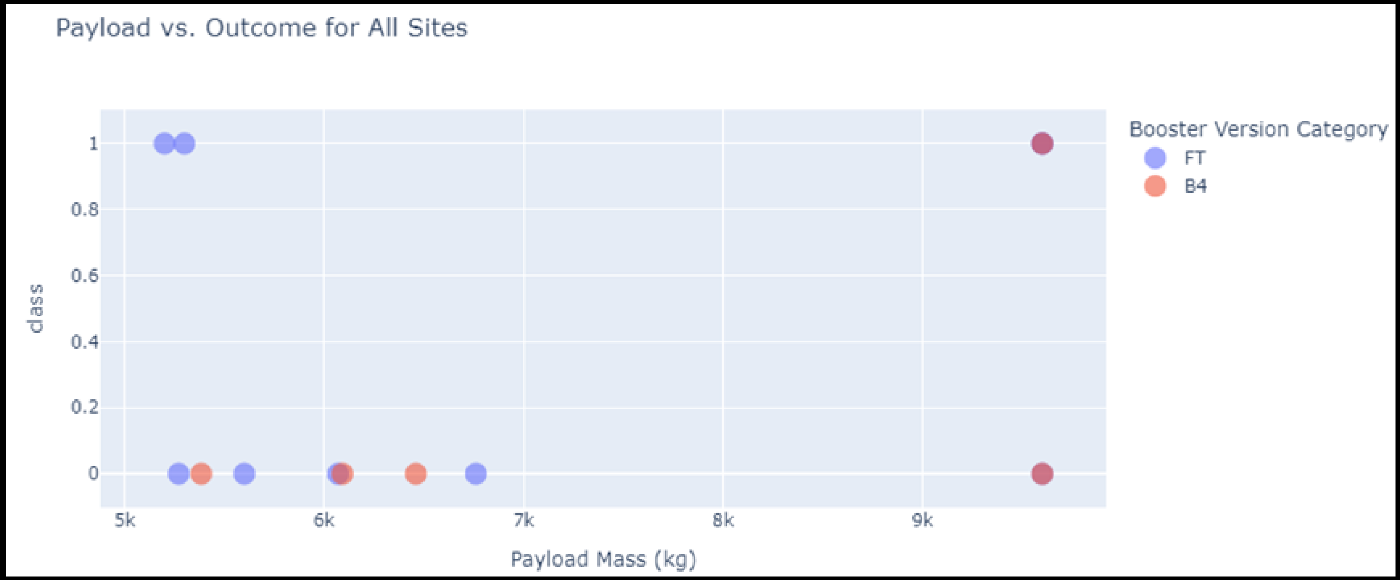
Site with Highest Success Ratio

- Rocket launches at the KSC LC-39A site achieved a **76.9%** success rate with 23.1% launches ended up failing



Payload vs. Launch Outcome

- This plot shows the Booster Version **FT** achieves higher success rate amongst the other Booster Versions in the **low-mid** Payload range 0 - 3000kg

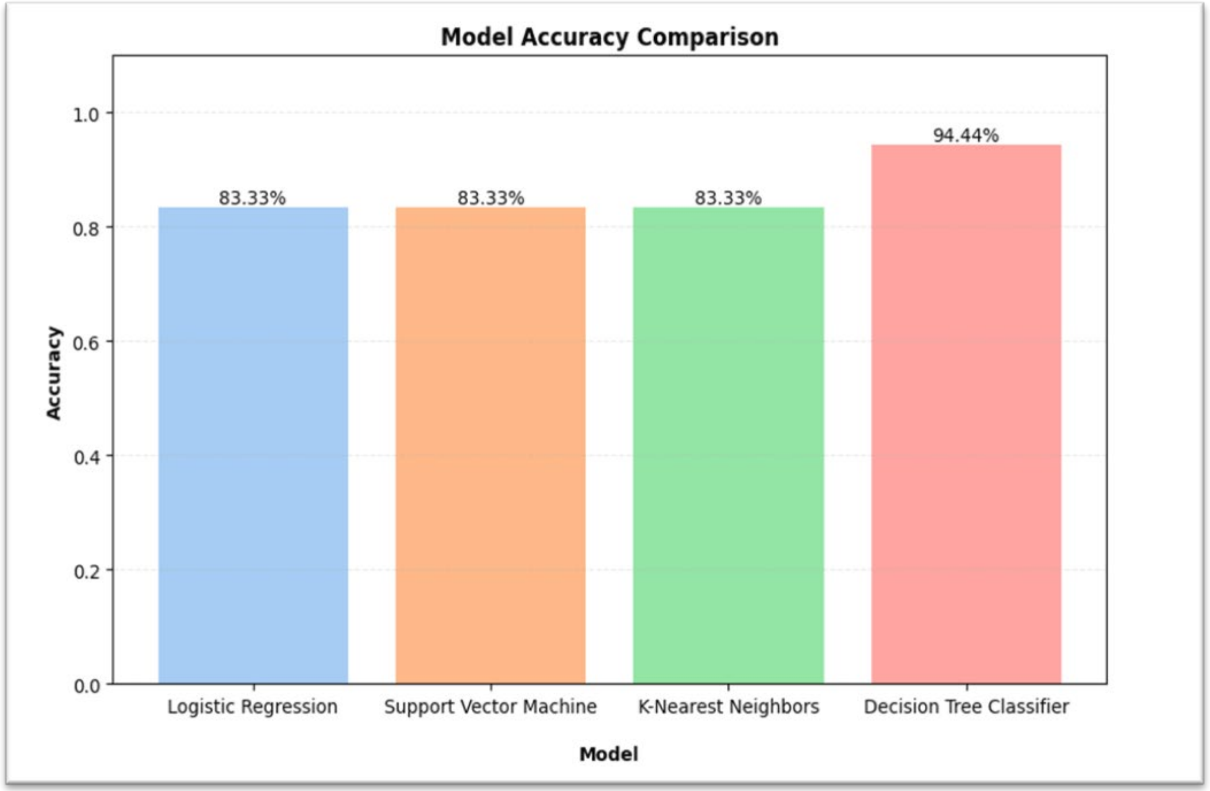


Payload vs. Launch Outcome

- In the **high** payload mass range the Booster Version **B4** is the only launch that has succeeded with a high payload (approx. 9600kg)

PREDICTIVE ANALYSIS

Classification Accuracy



- The Decision Tree classifier is the best performing model with accuracy at 94.44%

Confusion Matrix (CM)

- CM is often good for measuring performance on imbalanced and sensitive datasets
- **Distribution** of classes (*See Appendix 1*):

  0 - failure to land = 33.3%

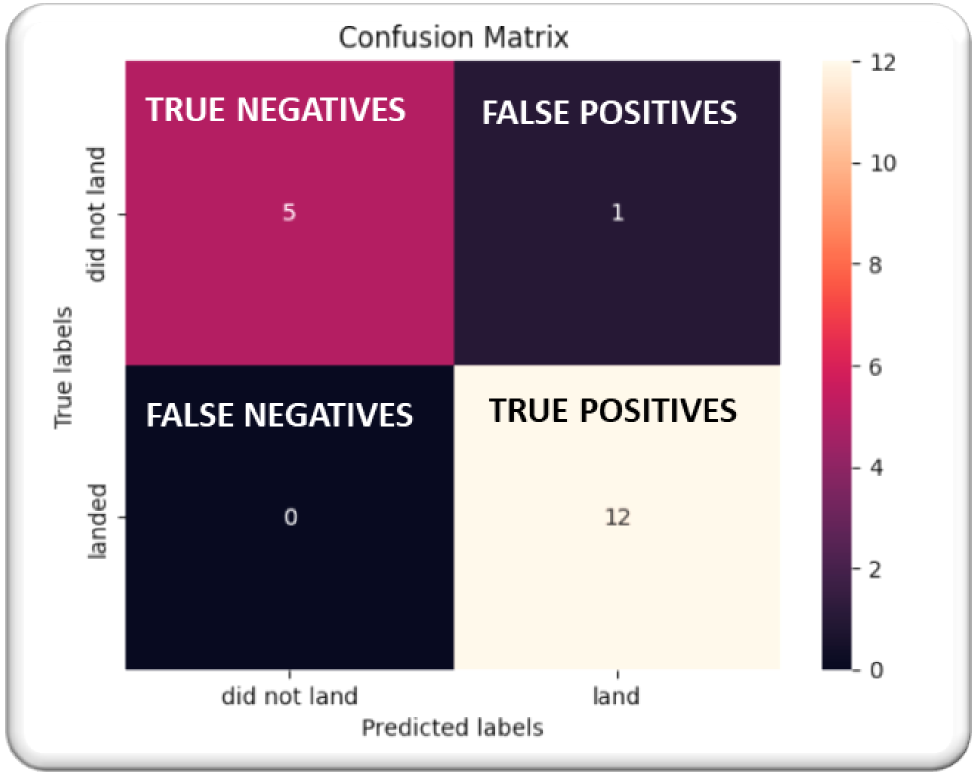
  1 - successful outcome = 66.6%

Precision = 92.3%

- Out of all the launches where the model predicted a successful outcome (positive), 92.3% of those predictions were correct. Model is very good at avoiding false positives (incorrectly predicting success)

Recall = 100%

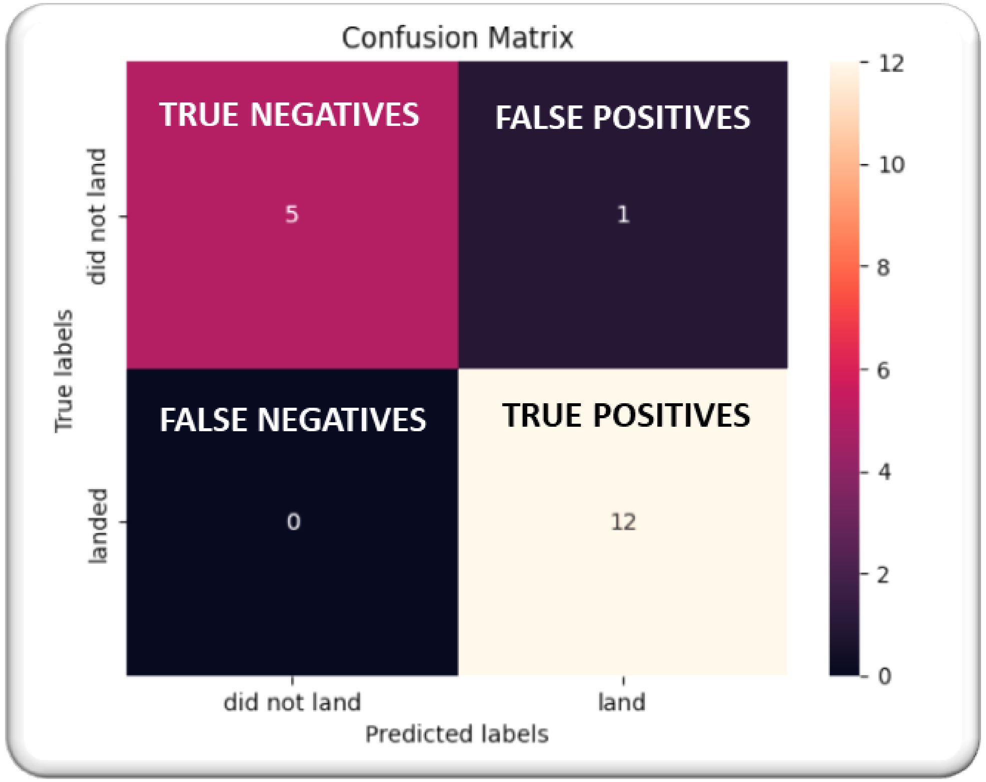
- The model successfully identified all instances of successful outcomes. No false negatives (actual successful landing but predicted failure)



Confusion Matrix (CM)

- **F1-Score = 0.96**

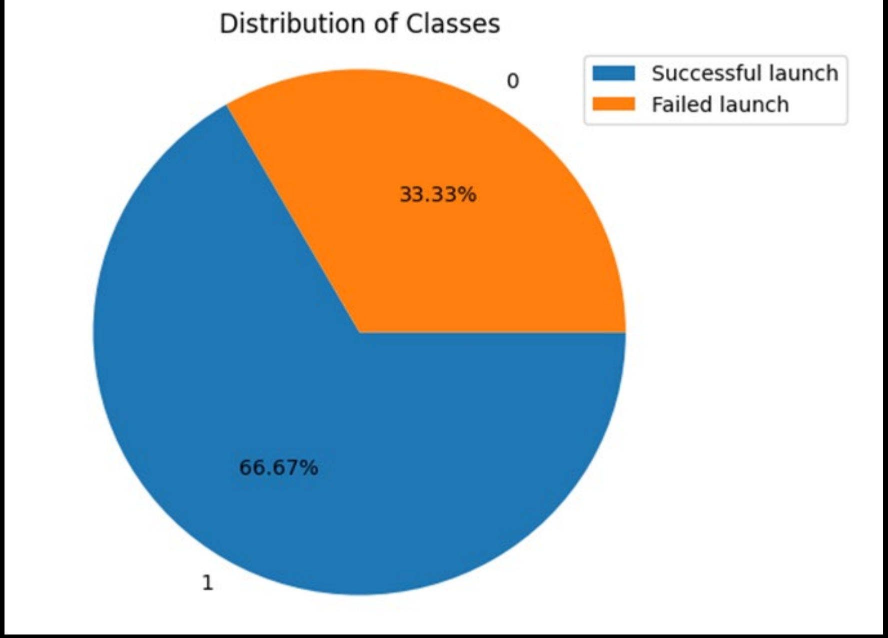
The Decision Tree model shows excellent performance with high precision and perfect recall. This means it is highly reliable in predicting successful outcomes (high precision) and does not miss any successful outcomes (perfect recall)



- **Launch Sites**: The most successful site is **KSC LC-39A**
- **Trending Pattern**: An upward trend in successful launches observed yearly
- **Interesting Insight Extracted**: On a low payload range ~2000-3000kgs the FT Booster Version achieves a very high success rate
- **Orbit findings:** The sun-synchronous orbit has the highest success rate with 100% for all 5 launches. ES-L1, GEO, and HEO orbits also have 100% success but only performed one launch
- **Locations:** All launch sites are situated near coastlines and maintain a safe distance from cities
- **Site Proximities:** Railways & highways are in proximity to the launch sites but are not directly adjacent
- **Predictive Analysis**: The best performing model was the Decision Tree classifier

Conclusion

Appendices




```
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 None ASDS
6 False Ocean
7 False RTLS
```

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]]) # set used to store unique outcomes
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

```
# Using list comprehension
landing_class = [0 if outcome in bad_outcomes else 1 for outcome in df['Outcome']]
```

```
df['Class']=landing_class
```

1. Distribution of Classes


THANK YOU