Winning Space Race with Data Science

Wening Dyah Locitaresmi

October 2024

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Executive Summary - Overview

This project aims to predict the successful landing of SpaceX's Falcon 9 rocket's first stage, a critical factor in reducing launch costs through reuse. By pulling data from the SpaceX API, we gathered historical records of Falcon 9 launches, cleaned the data, and developed a predictive model. Additional data was also collected through web scraping from Wikipedia to enrich the dataset.

Key Objectives:

1. Gather and clean data on Falcon 9 launches
2. Extract key information such as booster version, payload mass, launch site, and landing outcomes
1. Build a model to predict the probability of a successful first-stage landing

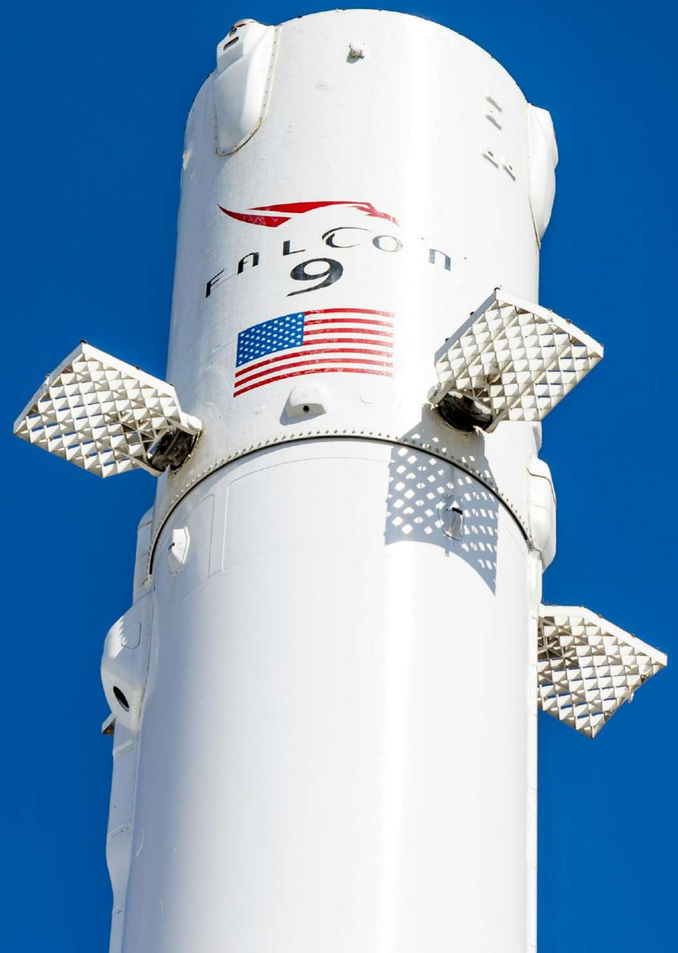
Executive Summary - Key Insights

1. The project successfully gathered and cleaned SpaceX Falcon 9 launch data.
2. The analysis identified that variables such as booster version, payload mass, and launch site played significant roles in determining landing success.
3. The model provides valuable predictions on landing outcomes.
4. The ability to predict successful landings supports SpaceX’s cost-saving strategy by maximising reusability of the Falcon 9 first stage.

Introduction

SpaceX has transformed the space industry by cutting launch costs through the reuse of the Falcon 9 rocket's first stage. Each successful landing saves millions, with SpaceX launches costing $62 million compared to competitors at $165 million.

This project aims to predict the success of these landings using data from the SpaceX API, supplemented with additional records. The insights gained will help improve cost-efficiency and optimise launch strategies for both SpaceX and other providers.



Section 1

Methodology

Summary - Methodologies

1: Data Collection

Data on Falcon 9 launches was gathered using the SpaceX API. This included key information such as booster version, launch site, payload mass, and landing outcomes. Web scraping from Wikipedia was also used to supplement the dataset with additional launch records.

2: Data Cleaning & Processing

The raw data was processed and cleaned, with missing values handled and irrelevant data filtered out.

3: Data Wrangling

Various functions were applied to extract booster versions, launch site coordinates, payload mass, orbit, and landing outcomes, ensuring all data was structured for analysis.

4: Prediction Model

The cleaned data was used to build a model that predicts whether the Falcon 9 first stage will successfully land, based on historical factors.

Data Collection - Overview

The project involved collecting data from multiple sources to predict Falcon 9 first-stage landing outcomes.

Two main methods were used:

SpaceX API

for retrieving detailed launch data.

Web Scraping

for extracting historical launch records from Wikipedia.

Data Collection - SpaceX API

Data was gathered directly from the SpaceX API, which offers a structured way to access historical information on Falcon 9 launches.

Booster version

From the rocket API endpoint

Launch site details

Long, lat, site name

Payload information

Mass and orbit type

Core details

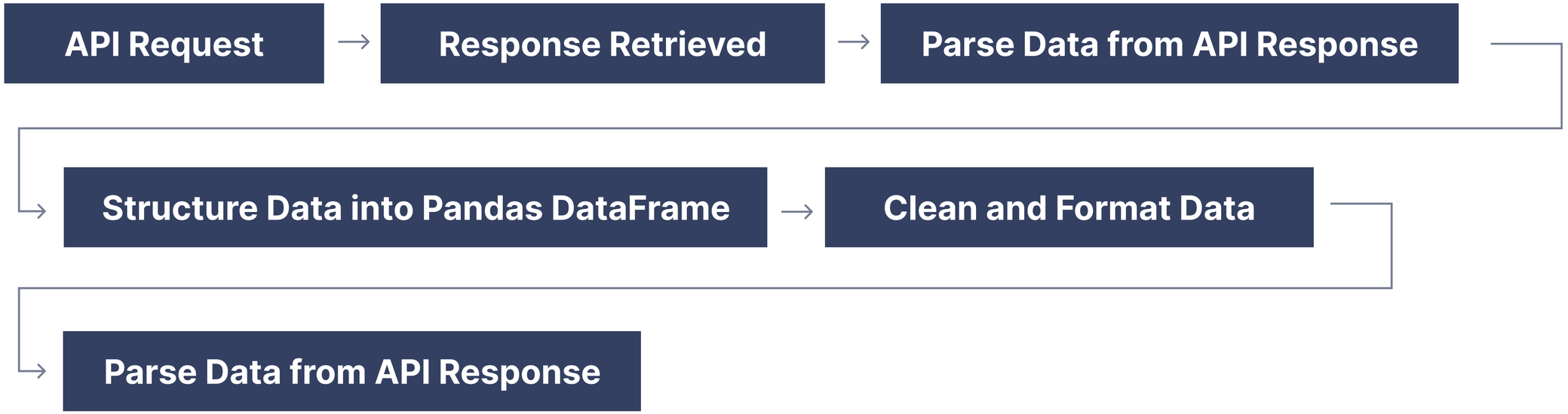
Landing outcome, flight number, and reuse details

Each launch record was retrieved using unique IDs and parsed into a structured format using Pandas for further analysis

Github URL

Data Collection - SpaceX API (cont'd)

The process:

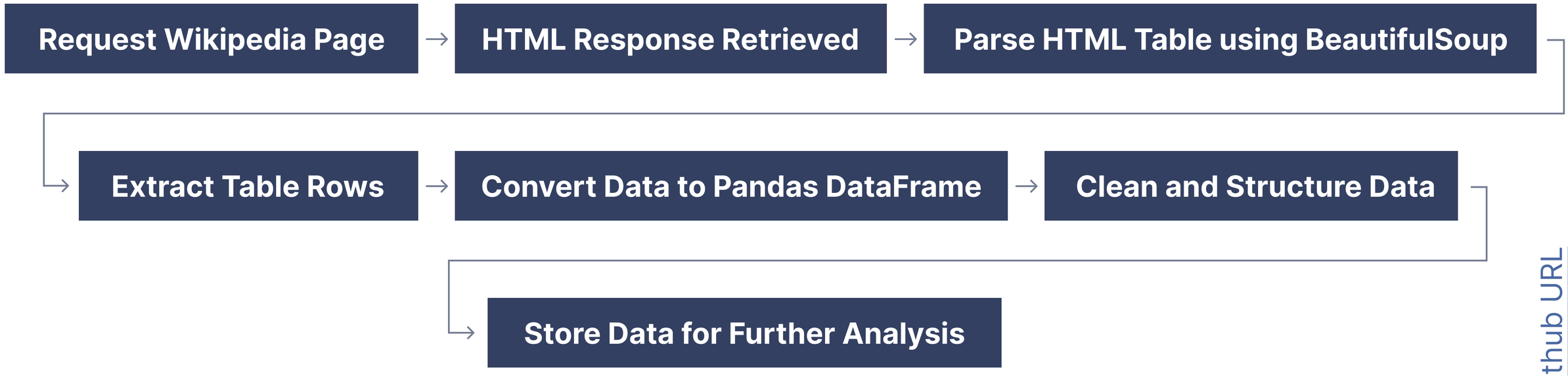


Github URL

Data Collection - Web Scraping

In addition to the API, web scraping was utilised to gather any additional data points that were not available via the API to enrich the dataset and to fill in potential gaps.

Source: Wikipedia page listing Falcon 9 and Falcon Heavy launches



Github URL

Data Wrangling

The data wrangling process focused on cleaning, transforming, and preparing the raw data collected from the SpaceX API and web scraping for further analysis.

1

Exploratory Data Analysis (EDA)

Conducted to understand the data structure and identify any missing values or irregularities. The outcome types and the number of launches from various launch sites were examined.

2

Handling Missing Values

Missing values in certain attributes such as LandingPad were identified, and steps were taken to address them appropriately.

3

Feature Engineering

A new feature Class was created from the Outcome column to simplify landing success categorisation, where Class = 1 indicates a successful landing and Class = 0 indicates a failure.

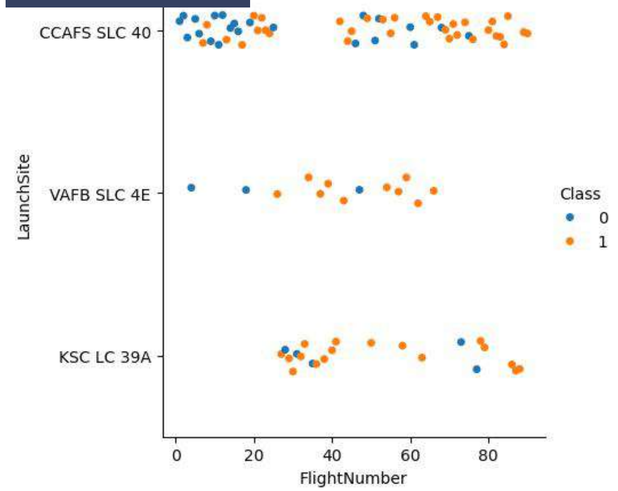
Github URL

EDA with Data Visualisation

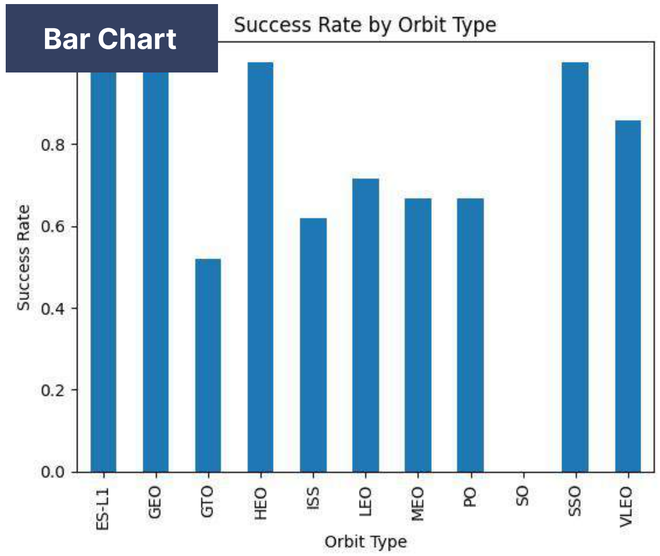
Exploratory Data Analysis (EDA) was conducted using visualisation techniques to better understand the relationships and trends in the dataset that could impact Falcon 9 first-stage landing success.

Key Charts Used:

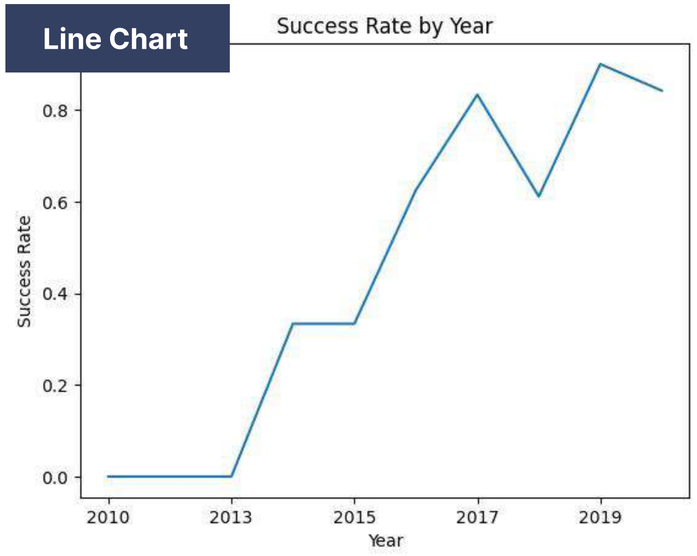
Scatter Plot



Bar Chart



Line Chart



Github URL

EDA with Data Visualisation (cont’d)

1 Flight Number vs. Payload Mass (Categorical Plot)

To understand the relationship between flight experience (represented by flight numbers) and payload mass, and how these factors influence the success of the first stage landing.

Insight: Higher flight numbers showed an increased likelihood of successful landings, even with heavier payloads.

2 Flight Number vs. Launch Site (Scatter Plot)

To explore how launch sites impact landing success across different flight numbers.

Insight: Some sites, such as KSC LC 39A, demonstrated a higher success rate as the number of launches increased.

Github URL

EDA with Data Visualisation (cont’d)

3 Payload Mass vs. Launch Site (Scatter Plot)

To investigate whether launch sites affect payload mass capabilities and landing success.

Insight: VAFB-SLC site handled lower payload masses, while other sites could support larger payloads with successful landings.

4 Flight Number vs. Launch Site (Scatter Plot)

To determine which orbit types are most likely to result in successful landings.

Insight: Orbits like ES-L1, GEO, HEO, and SSO had the highest success rates.

Github URL

EDA with Data Visualisation (cont’d)

5 Flight Number vs. Orbit Type (Scatter Plot)

To examine the connection between flight experience (flight numbers) and orbit type in relation to landing success.

Insight: LEO orbit showed a clear relationship between flight number and success, while GTO did not.

6 Payload Mass vs. Orbit Type (Scatter Plot)

To see if there is any correlation between payload mass and orbit type affecting landing success.

Insight: Heavier payloads were more likely to land successfully in Polar, LEO, and ISS orbits, while GTO showed mixed outcomes.

Github URL

EDA with Data Visualisation (cont’d)

7 Flight Number vs. Orbit Type (Scatter Plot)

To track the overall success rate of Falcon 9 landings over the years.

Insight: The success rate consistently increased from 2013 to 2020, indicating improvements over time.

Github URL

EDA with SQL

SQL queries were used to further explore the dataset, allowing for specific, structured queries that provided valuable insights on the data.

1 Display unique launch sites

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

2 Retrieve 5 records where launch sites start with 'CCA'

```
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

3 Calculate the total payload mass carried by NASA (CRS) boosters

```
SELECT SUM(PAYLOAD_MASS__KG_) AS 'Total Payload Mass'
FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

Github URL

EDA with SQL (cont’d)

4 **Display average payload mass carried by F9 v1.1 boosters**

```
SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average Payload Mass'
FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
```

5 **List the date of the first successful landing on a ground pad**

```
SELECT MIN(Date) AS 'First Successful Landing in Ground Pad'
FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

6 **List boosters with successful drone ship landings and payloads between 4000–6000 kg**

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

Github URL

EDA with SQL (cont’d)

7 **Count total successful and failed mission outcomes**

```
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS 'Total' FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success' OR Landing_Outcome = 'Failure' GROUP BY Landing_Outcome;
```

8 **List booster versions that carried the maximum payload mass**

```
SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

9 **Display month names and launch details for 2015 drone ship failures**

```
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';
```

Github URL

EDA with SQL (cont’d)

10 **Rank landing outcomes between 2010–2017 in descending order**

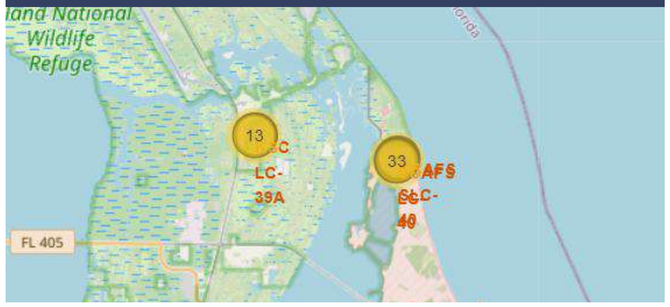
```
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS 'Total' FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC;
```

Github URL

Building an Interactive Map with Folium

The goal of using Folium was to create an interactive map displaying key geographical data related to SpaceX launch sites. This visualisation helps in understanding the proximity of launch sites to coastlines, cities, railways, and highways, which are all essential for mission logistics and safety.

Markers for Launch Sites



Each launch site was marked on the map using its latitude and longitude to visualise its exact location.

Circles around Launch Sites

A circle with a fixed radius was added around each launch site to highlight the areas and mark the proximity visually.

Mouse Position

A tool that allows users to see the coordinates of any point on the map by hovering over it. This feature helps in marking additional proximities.

Github URL

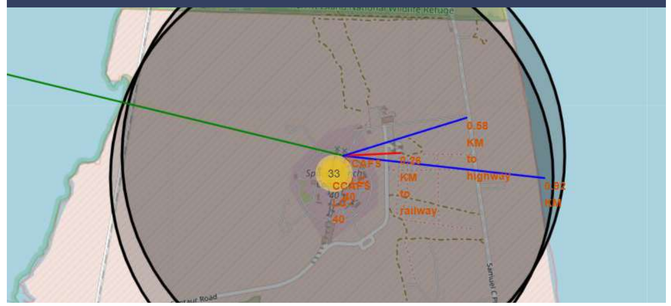
Building an Interactive Map with Folium (cont'd)

More Features:

Distance Markers

Markers showing the distance between launch sites and key proximities, such as coastlines or highways, were added to understand the logistical aspects of launch site selection.

Polylines



Lines were drawn between the launch sites and important proximities (e.g., closest railway, coastline, etc.) to show exact distances visually.

Github URL

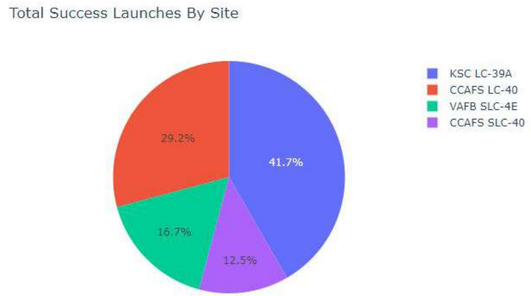
Build a Dashboard with Plotly Dash

This dashboard application provides an interactive way to analyse SpaceX launch data in real time. Users can explore the success rates of launches by selecting different sites, payload ranges, and booster versions.

Launch Site Dropdown

Allows users to select specific launch sites or view data from all sites.

Success Pie Chart



Allows users to select specific launch sites or view data from all sites.

Payload Range Slider

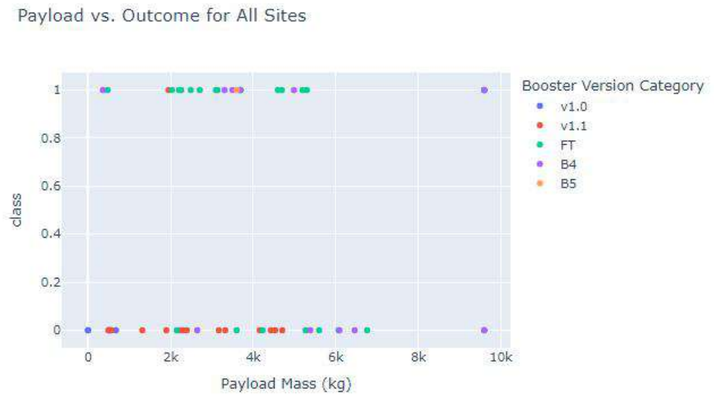
Allows users to select specific launch sites or view data from all sites.

Github URL

Build a Dashboard with Plotly Dash (cont'd)

More Features:

Success-Payload Scatter Plot



Plots payload mass against the launch success outcome, with points coloured by booster version.

Github URL

Predictive Analysis (Classification)

The goal of this step was to predict the success of the Falcon 9 rocket's first-stage landing using machine learning classification models. We aimed to find the best-performing model by evaluating and improving various algorithms.

1 Data Preprocessing

- Data was standardised using StandardScaler to ensure that features had similar scales.
- The dataset was split into training and test sets with an 80-20 ratio.

2 Algorithms Used

- Logistic Regression
- K-Nearest Neighbours (KNN)
- Decision Trees
- Support Vector Machine (SVM)

Github URL

Predictive Analysis (Classification) (cont'd)

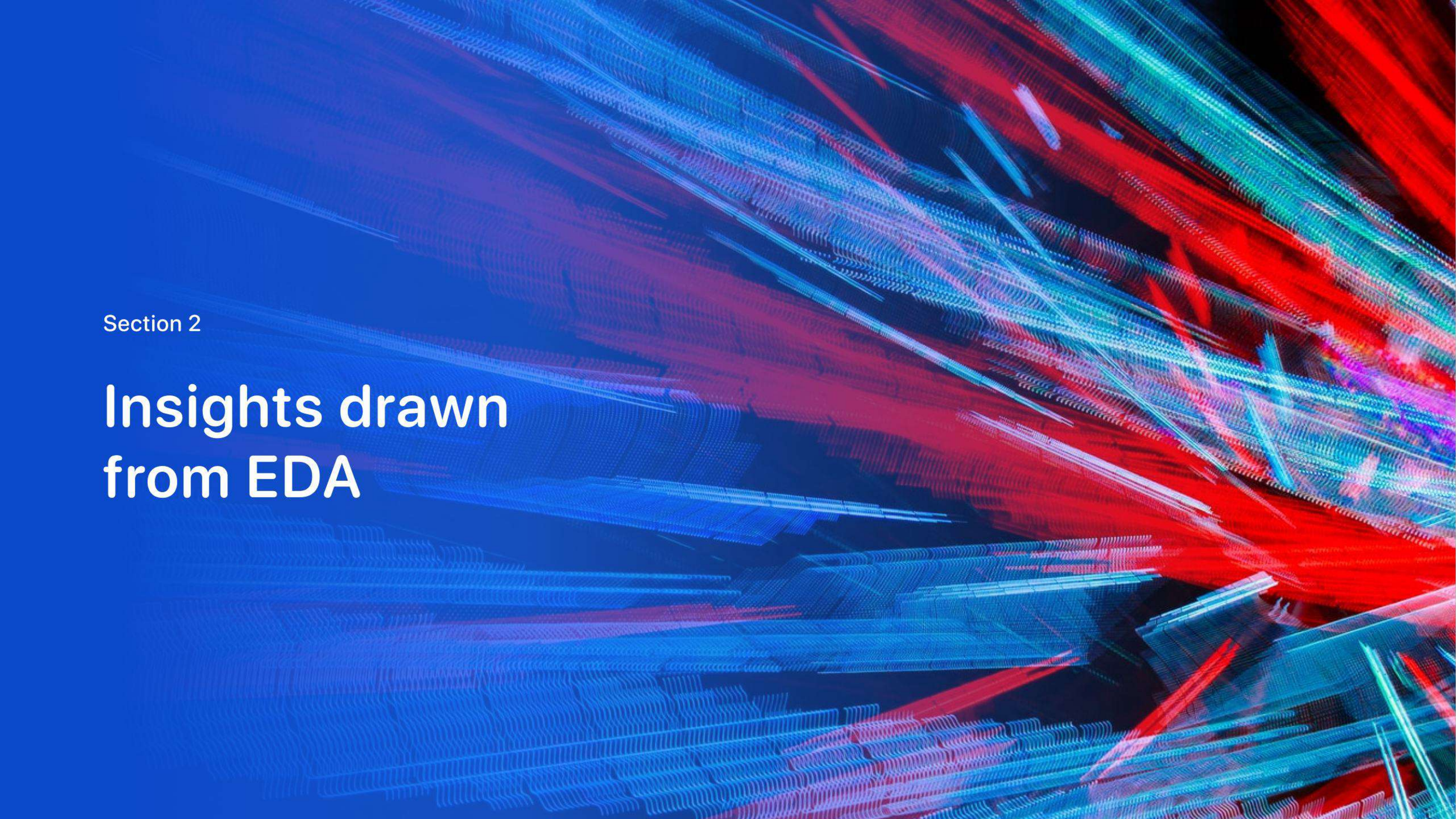
For each algorithm, a **GridSearchCV** was used to optimise hyperparameters by testing various parameter combinations across **10-fold cross-validation**.

Algorithm	Best Parameters	Accuracy
Logistic Regression	C = 0.01, penalty = l2, solver = lbfgs	Cross-val: 84.8%, test: 83.3%
KNN	n_neighbors = 10, algorithm = auto	Cross-val: 84.8%, test: 83.3%
Best-Performing **Decision Tree**	max_depth = 8, criterion = entropy	Cross-val: 88.7%, test: 83.3%
SVM	C = 1.0, gamma = 0.0316, kernel = sigmoid	Cross-val: 84.8%, test: 83.3%

Github URL

Section 2

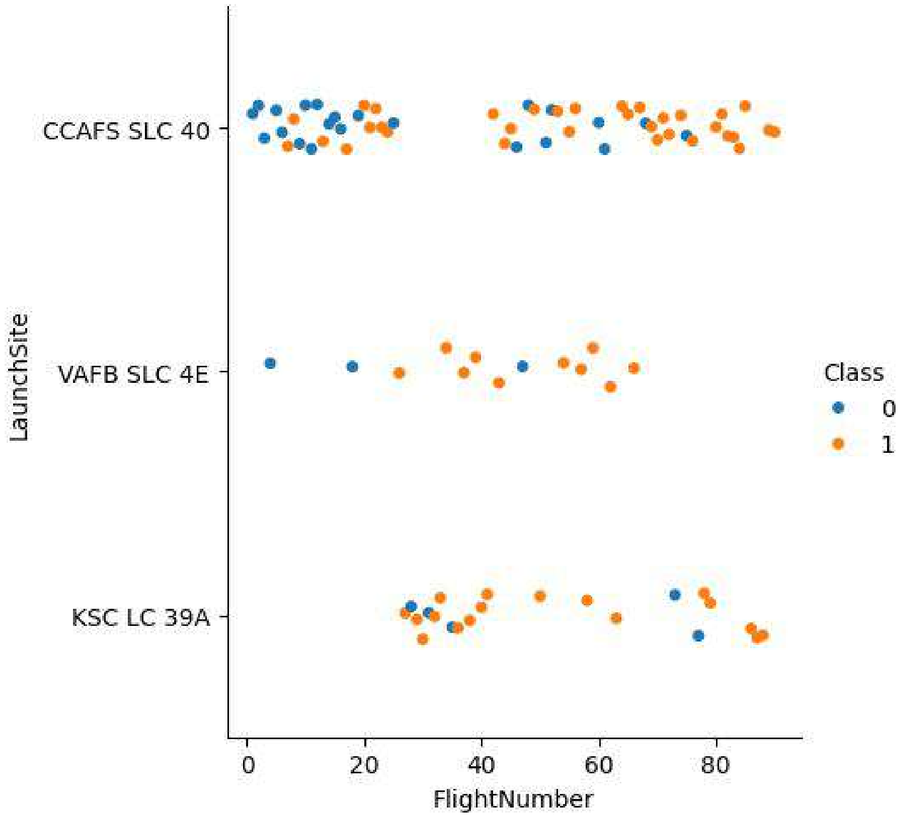
Insights drawn from EDA



Flight Number vs. Launch Site

This graph shows the relationship between launch sites, flight numbers, and landing success.

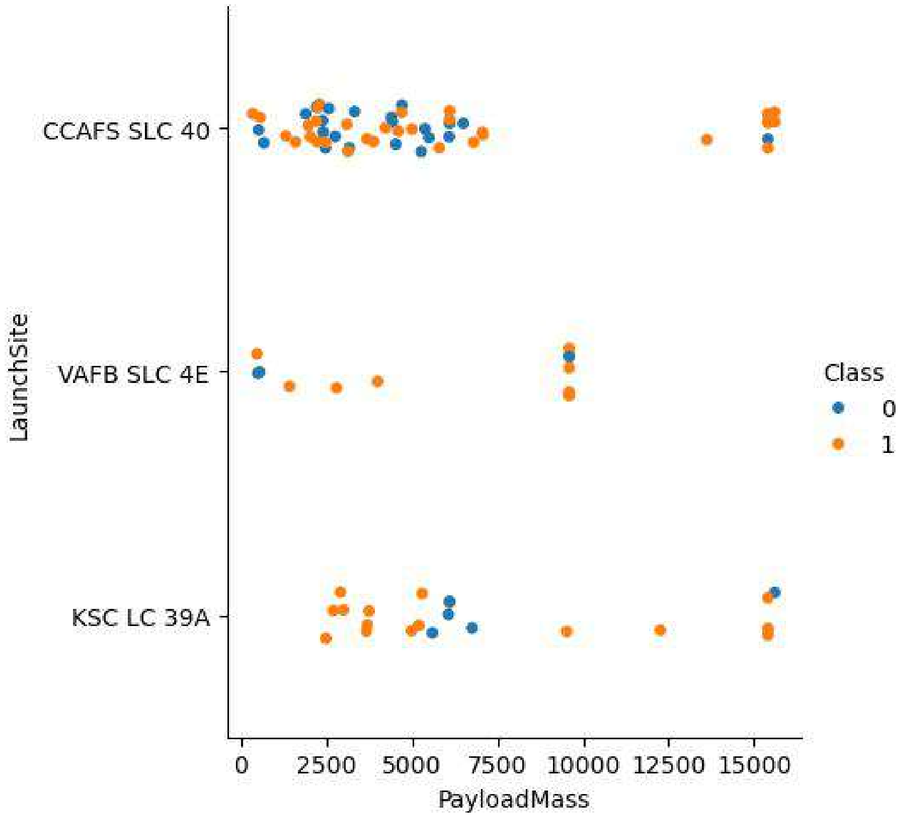
Higher flight numbers, especially from CCAFS SLC 40 and KSC LC 39A, are linked to improved landing success.



Payload vs. Launch Site

This graph shows the relationship between launch sites, payload mass, and landing success.

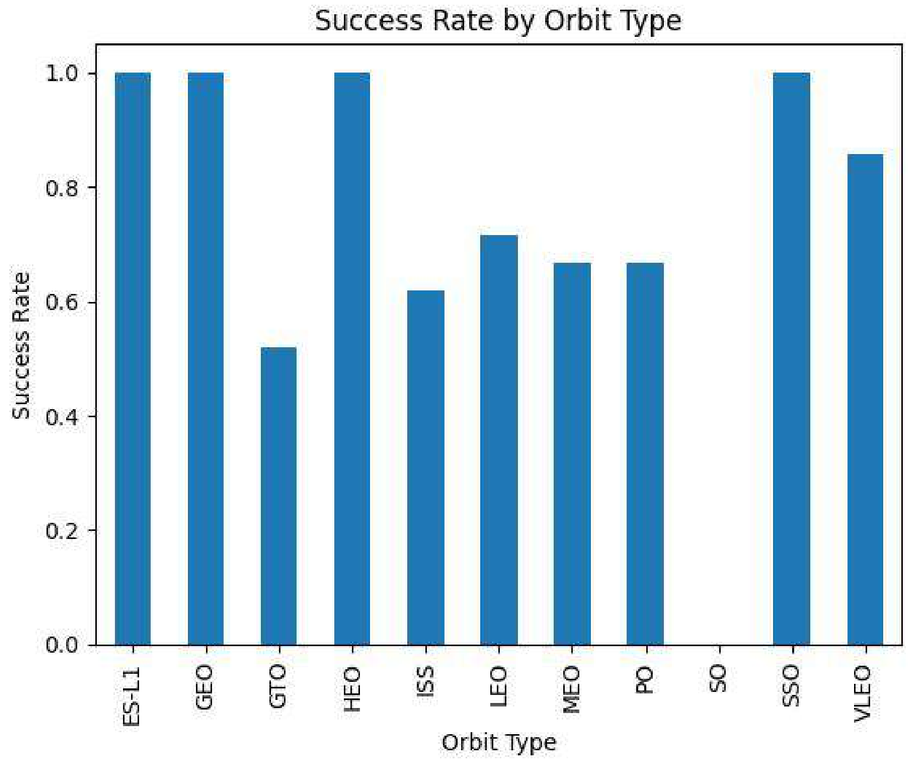
Heavier payloads, especially at CCAFS SLC 40 and KSC LC 39A, tend to have better landing success.



Success Rate vs. Orbit Type

This bar chart shows the Success Rate by Orbit Type and highlights how different orbital missions impact the likelihood of a successful first-stage landing.

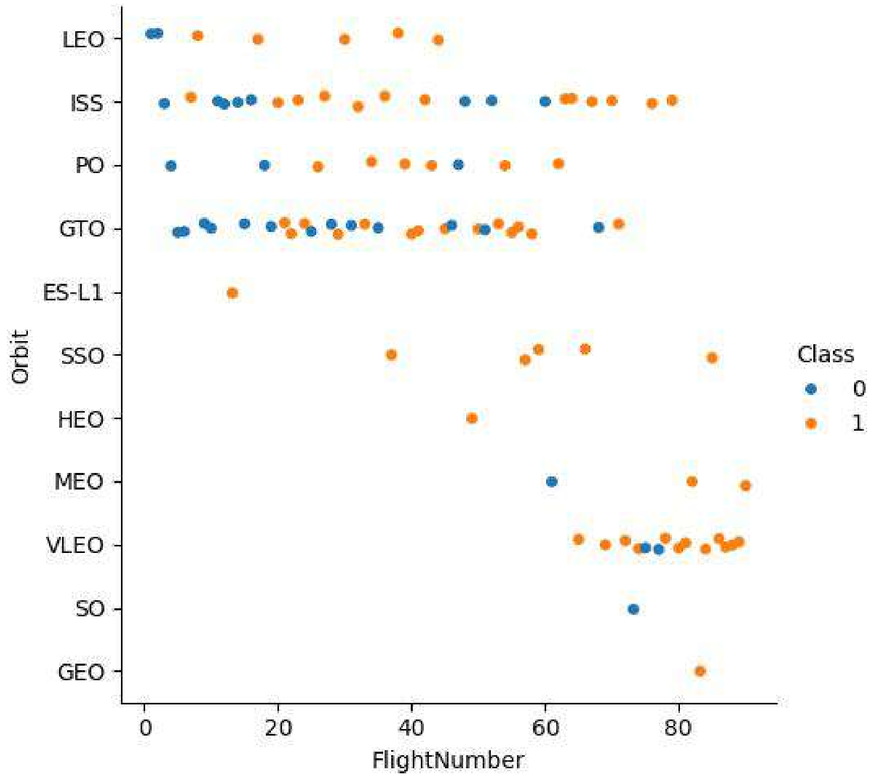
Missions to certain orbits, such as ES-L1 and SSO, have consistently high landing success, while more challenging orbits like GTO have lower success rates.



Flight Number vs. Orbit Type

This graph shows the relationship between FlightNumber, Orbit, and landing success.

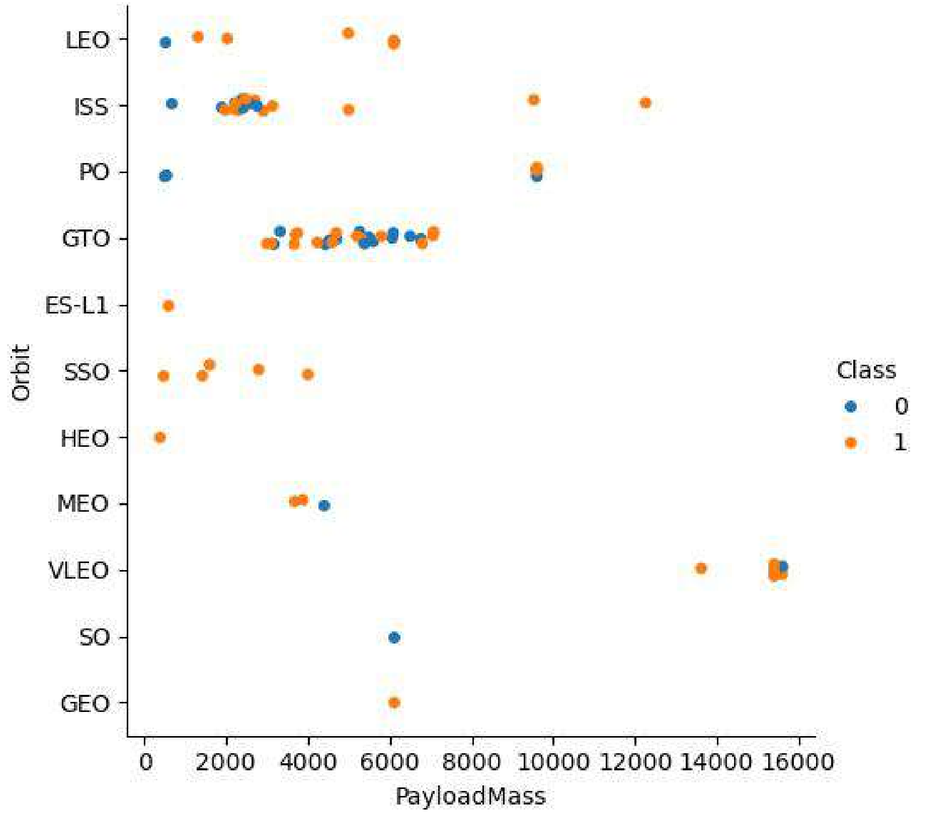
Landing success improves with more flights, especially for orbits like LEO, ISS, and VLEO, likely due to accumulated experience and technology improvements.



Payload vs. Orbit Type

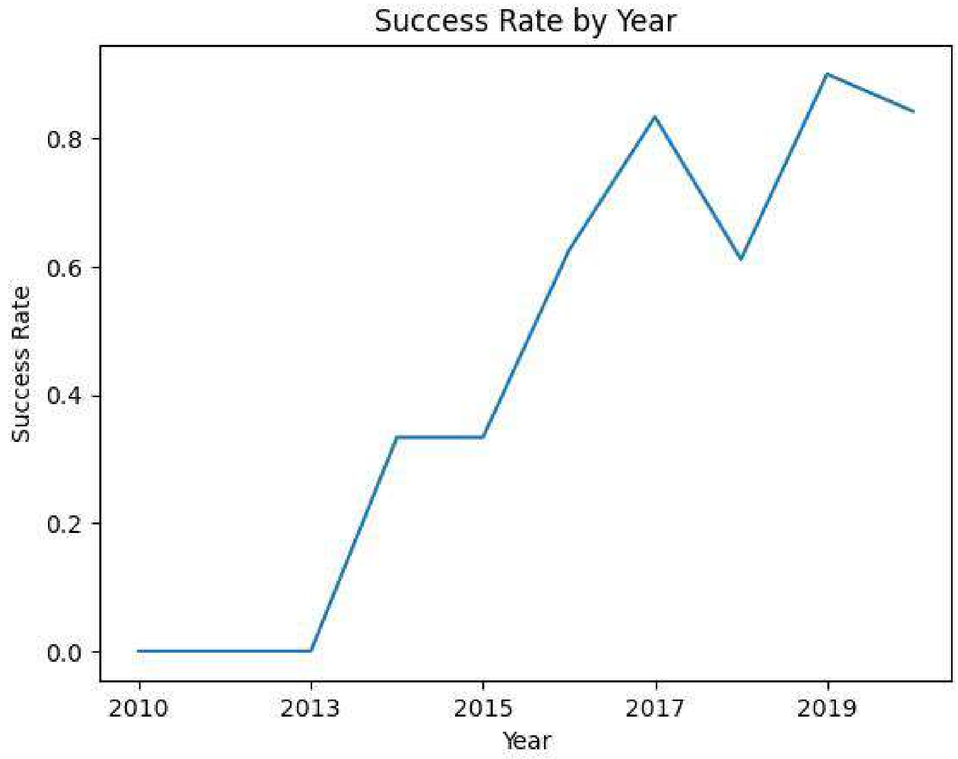
This graph illustrates the relationship between PayloadMass, Orbit, and landing success

Payload mass does not have a strong, clear correlation with landing success, as similar payload sizes often result in both successes and failures, especially in orbits like GTO and VLEO.



Launch Success Yearly Trend

The graph shows a clear trend of improvement in landing success over time, with substantial gains starting in 2015.



All Launch Site Names

The SQL query retrieves the unique names of the launch sites used in the dataset. The result shows four unique launch sites. These sites are key locations where SpaceX Falcon 9 rockets were launched, and each plays a significant role in the analysis conducted in the project.

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

Launch Site Names Begin with 'CCA'

The SQL query retrieves five records where the Launch_Site begins with 'CCA'. All results correspond to the CCAFS LC-40 launch site.

%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Date	Time (UTC)	Booster_Version	Launch_Site	...
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	...
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	...
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	...
2012-10-08	2013-03-01	F9 v1.0 B0006	CCAFS LC-40	...
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	...

Total Payload Mass

The SQL query calculates the total payload mass carried by SpaceX boosters from NASA with a result of 45,596 kg. This represents the cumulative weight of payloads delivered.

Total PayLoad Mass
45596

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

Average Payload Mass by F9 v1.1

The SQL query calculates the average payload mass carried by the F9 v1.1 booster version which results in an average payload of 2534.67 kg. This average represents the typical mass delivered by Falcon 9 v1.1 rockets across multiple launches.

Average PayLoad Mass
2534.6667

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average PayLoad Mass'
FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

First Successful Ground Landing Date

The SQL query finds the date of the first successful ground landing for SpaceX, which occurred on 22 December 2015. This marks the company's first successful landing on a ground pad.

First Successful Landing in Ground Pad
2015-12-22

```
%sql SELECT MIN(Date) AS 'First Successful Landing in Ground Pad' FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

The SQL query lists the booster versions that successfully landed on a drone ship with payloads between 4,000 and 6,000 kg.

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

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

Total Number of Successful and Failure Mission Outcomes

The SQL query calculates the total number of successful and failed mission outcomes. This highlights the high success rate achieved by SpaceX's landings.

Landing_Outcome	Total
Failure	3
Success	38

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS 'Total' FROM SPACEXTABLE WHERE Landing_Outcome = 'Success' OR Landing_Outcome = 'Failure' GROUP BY Landing_Outcome
```

Boosters Carried Maximum Payload

The SQL query retrieves the names of boosters that have been involved in missions that carried the heaviest payloads recorded in the dataset.

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4

Booster_Version
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2

Booster_Version
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

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

2015 Launch Records

The SQL query shows two failed drone ship landings in 2015, both at CCAFS LC-40 using booster versions F9 v1.1 B1012 (January) and F9 v1.1 B1015 (April).

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

%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'

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

The SQL query ranks the landing outcomes between 2010-06-04 and 2017-03-20.

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

Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3

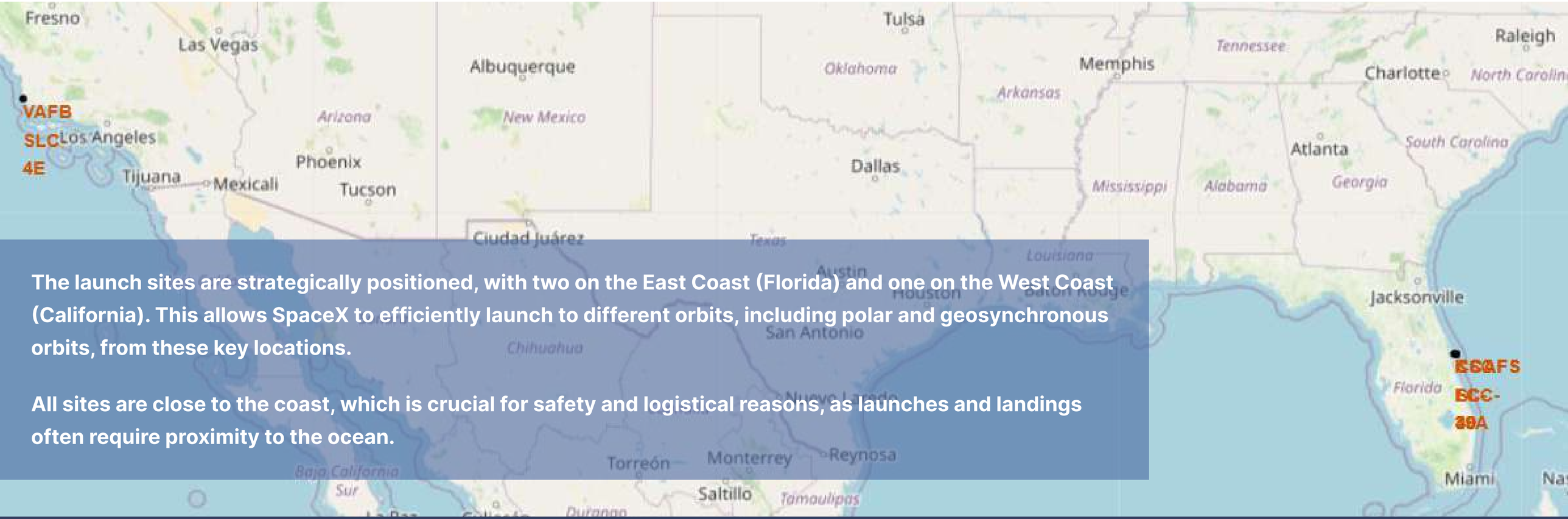
Landing_Outcome	Total
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

SpaceX Launch Sites on the Folium Map

The map displays the marked locations of all SpaceX Falcon 9 launch sites across the United States.



The launch sites are strategically positioned, with two on the East Coast (Florida) and one on the West Coast (California). This allows SpaceX to efficiently launch to different orbits, including polar and geosynchronous orbits, from these key locations.

All sites are close to the coast, which is crucial for safety and logistical reasons, as launches and landings often require proximity to the ocean.

Launch Sites

Colour-Labelled Launch Outcomes

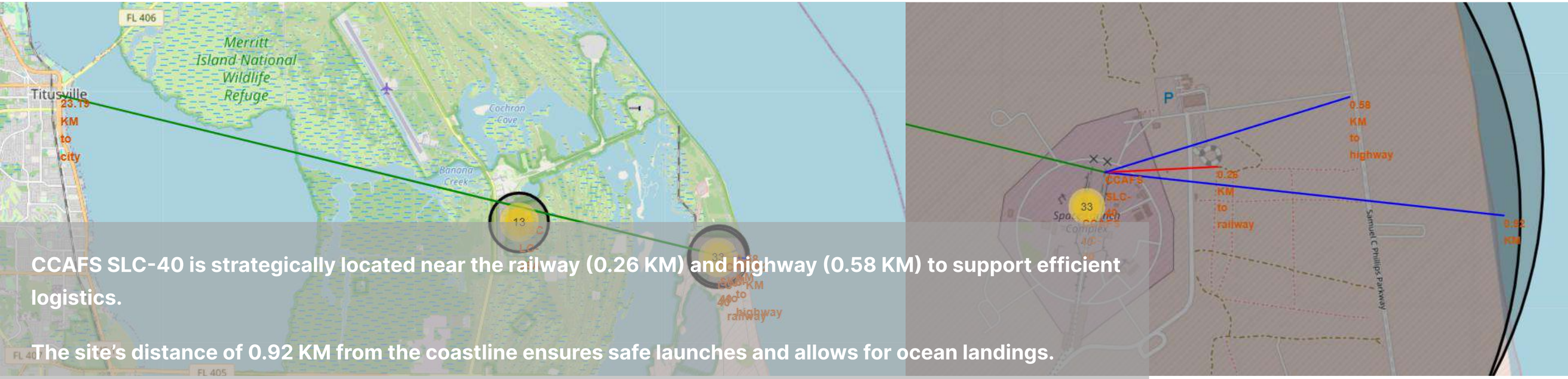
The map uses colour-coded markers to represent the outcomes of launches. Green markers indicate successful landings and yellow markers indicate failed landings.

Florida launch sites (KSC LC-39A and CCAFS SLC-40) show a larger number of launches, with a mix of outcomes but more successful landings. VAFB SLC-4E on the West Coast has fewer launches, with a balanced mix of success and failure.

The use of clusters helps visualise how outcomes vary by location and how many launches have been performed at each site.

Launch Site Proximities and Distances

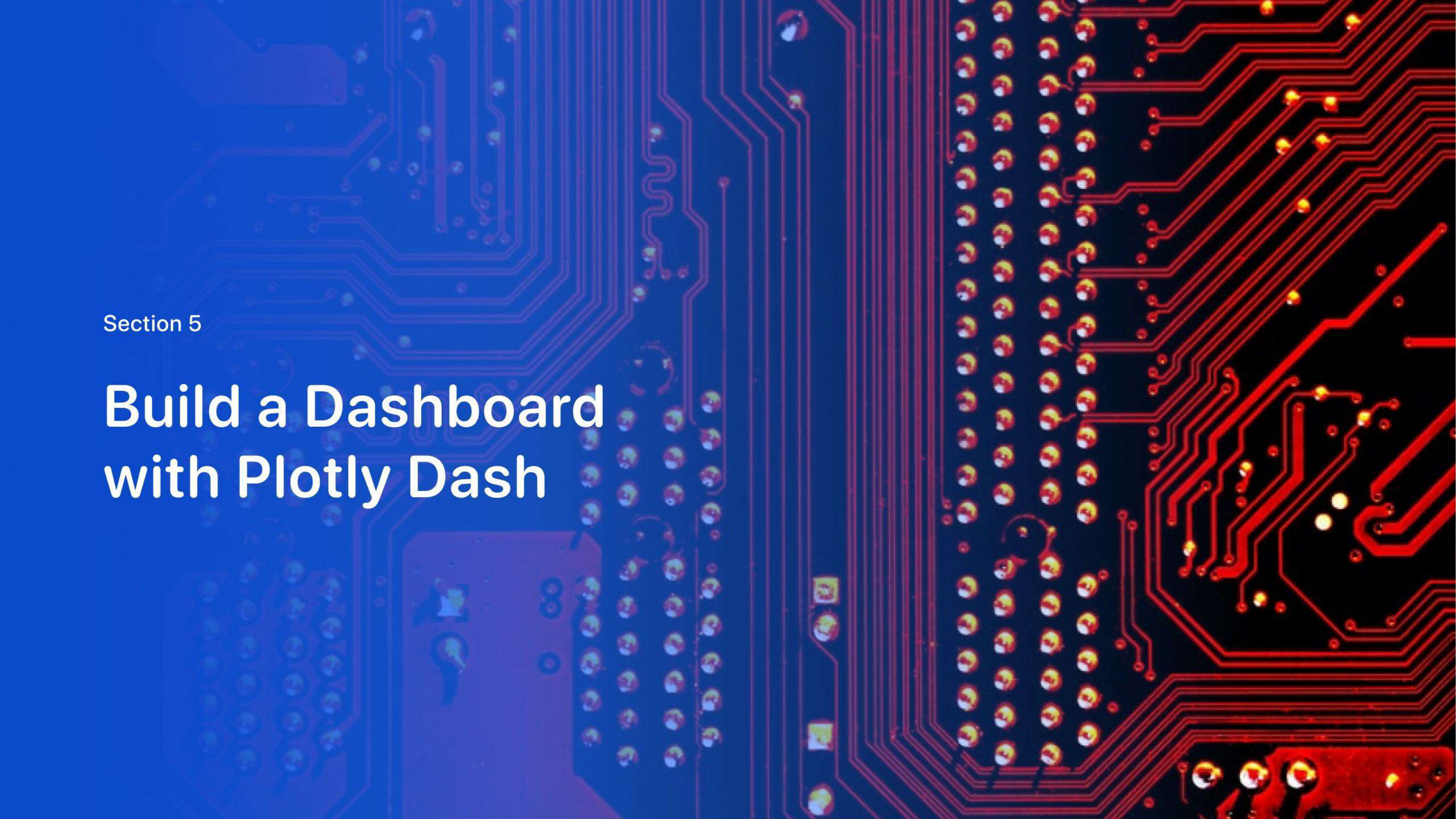
The map calculates and displays the distances between CCAFS SLC-40 and nearby important locations: railway, highway, coastline, and the nearest city.



CCAFS SLC-40 is strategically located near the railway (0.26 KM) and highway (0.58 KM) to support efficient logistics.

The site’s distance of 0.92 KM from the coastline ensures safe launches and allows for ocean landings.

At 23.19 KM from Titusville, the site is far enough from populated areas to ensure safety.



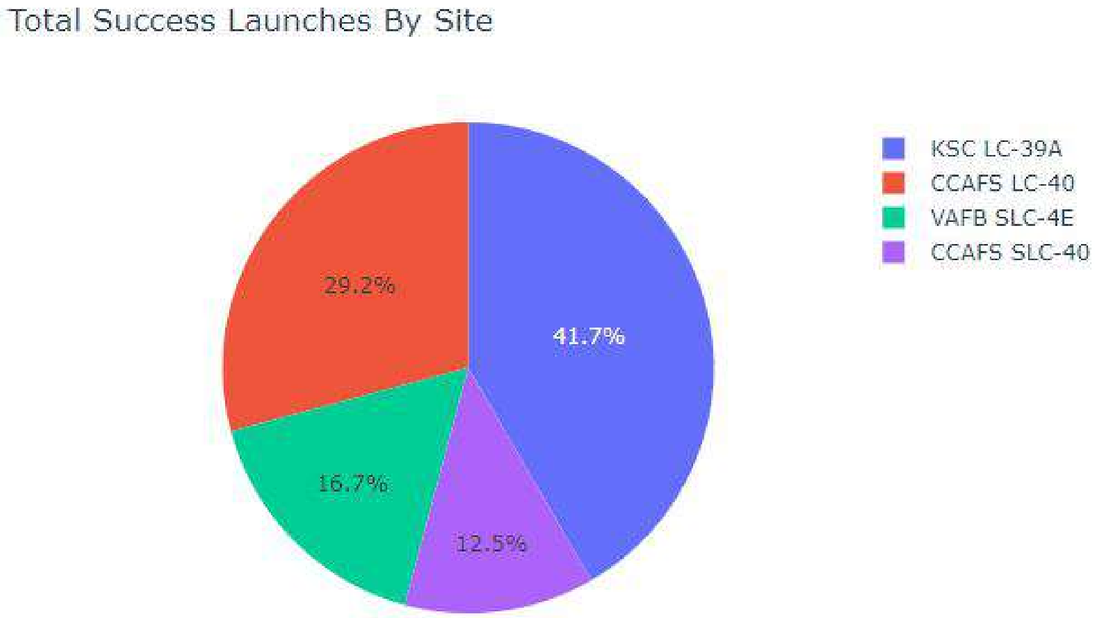
Section 5

Build a Dashboard with Plotly Dash

Total Successful Launches by Site

The pie chart visualises the distribution of successful launches across four launch sites.

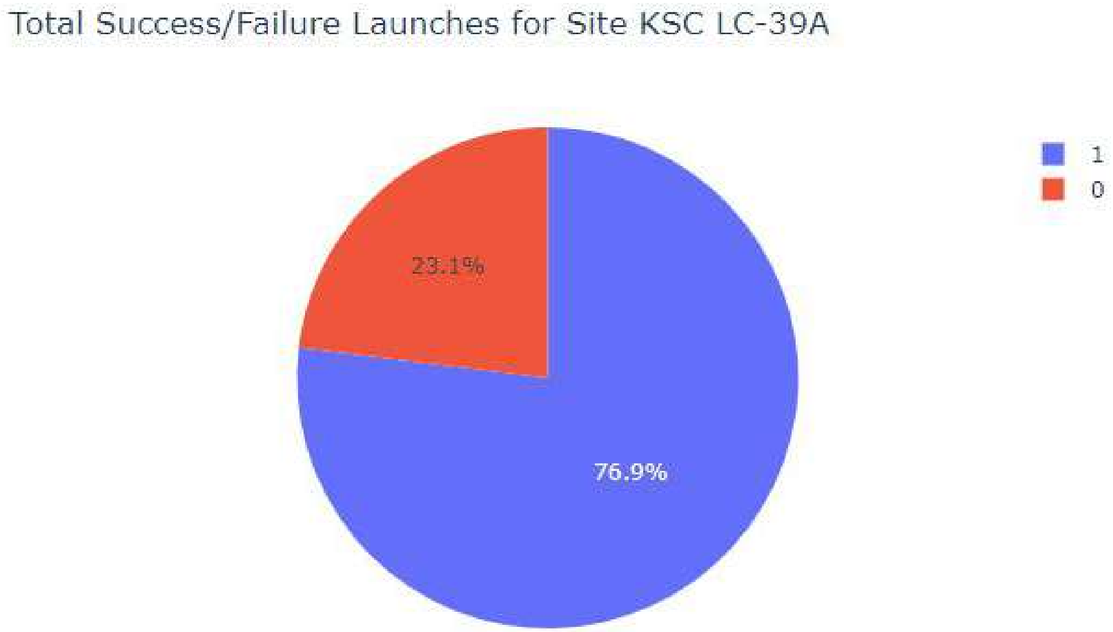
- KSC LC-39A leads with 41.7% of successful launches.
- CCAFS LC-40 accounts for 29.2%, making it another key site.
- VAFB SLC-4E and CCAFS SLC-40 contribute smaller but significant portions.



KSC LC-39A Launch Outcomes

This pie chart represents the success and failure rates of launches from the KSC LC-39A launch site.

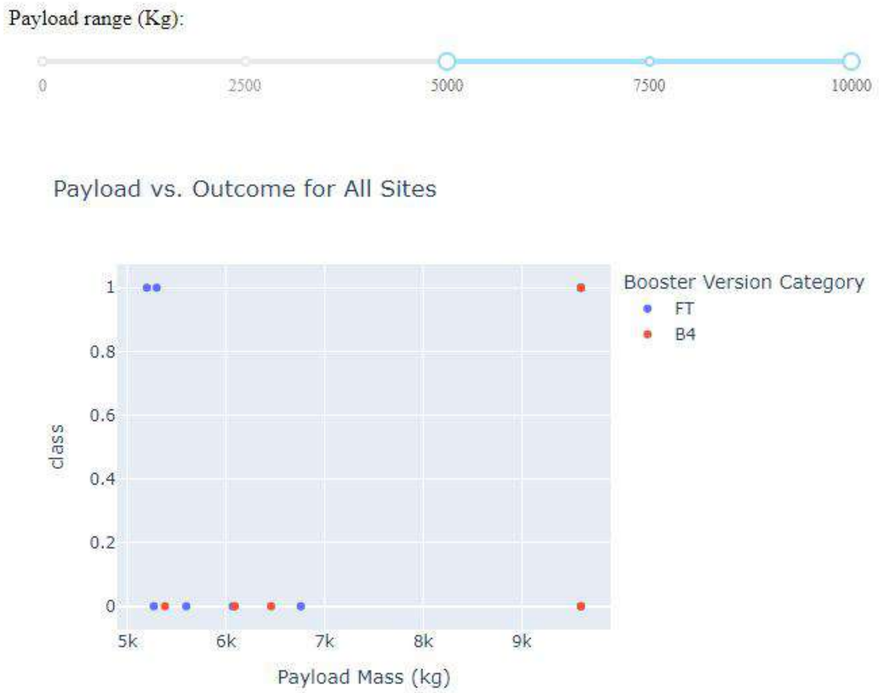
- KSC LC-39A has a strong success rate, with nearly 77% of its launches achieving successful landings.
- The 23.1% failure rate highlights areas for potential improvement but demonstrates the overall reliability of the site.

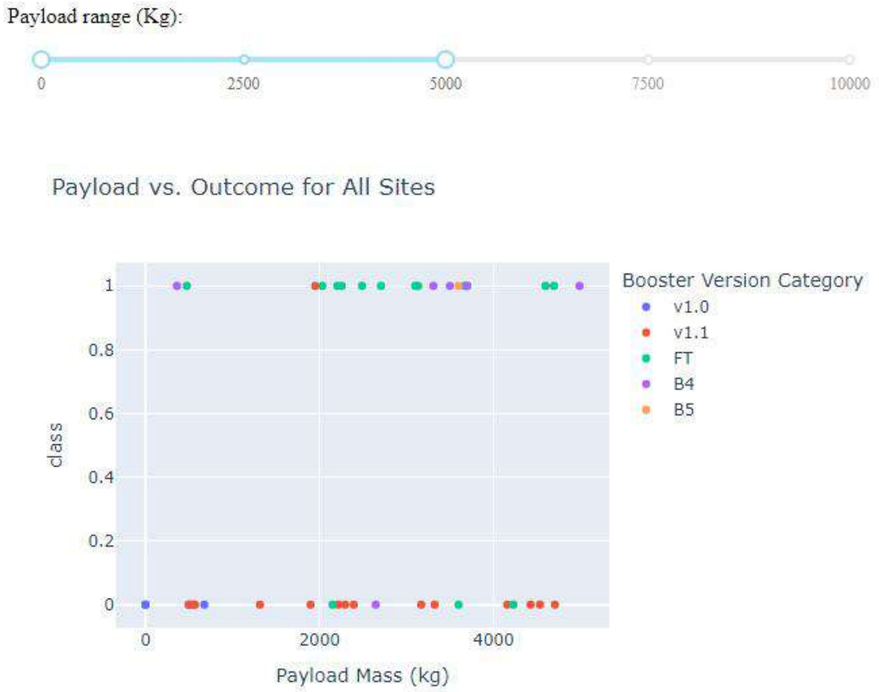


Payload vs. Launch Outcome

These scatter plots show the relationship between payload mass (x-axis) and launch outcome (success/failure) (y-axis) for various booster versions on different payload masses.

- Payloads below 5,000 kg (top) show a wide distribution of successful and failed launches, with all booster versions contributing.
- For payloads over 5,000 kg (bottom), only certain booster versions like FT and B4 are used, and successful landings are more likely as payload increases.





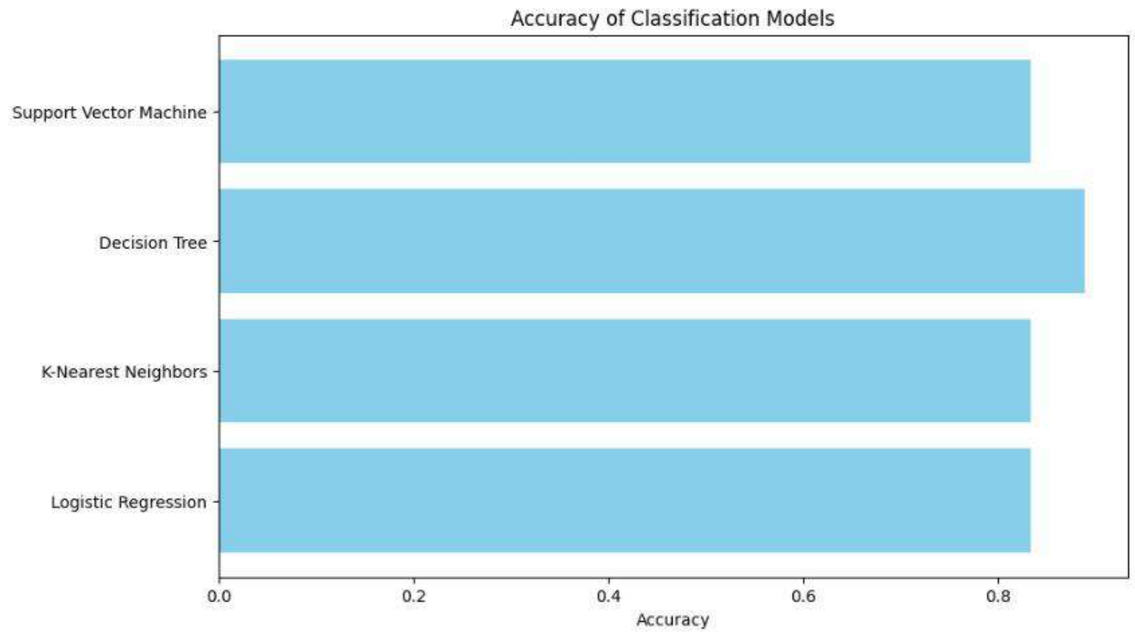
Section 6

Predictive Analysis (Classification)

Classification Accuracy

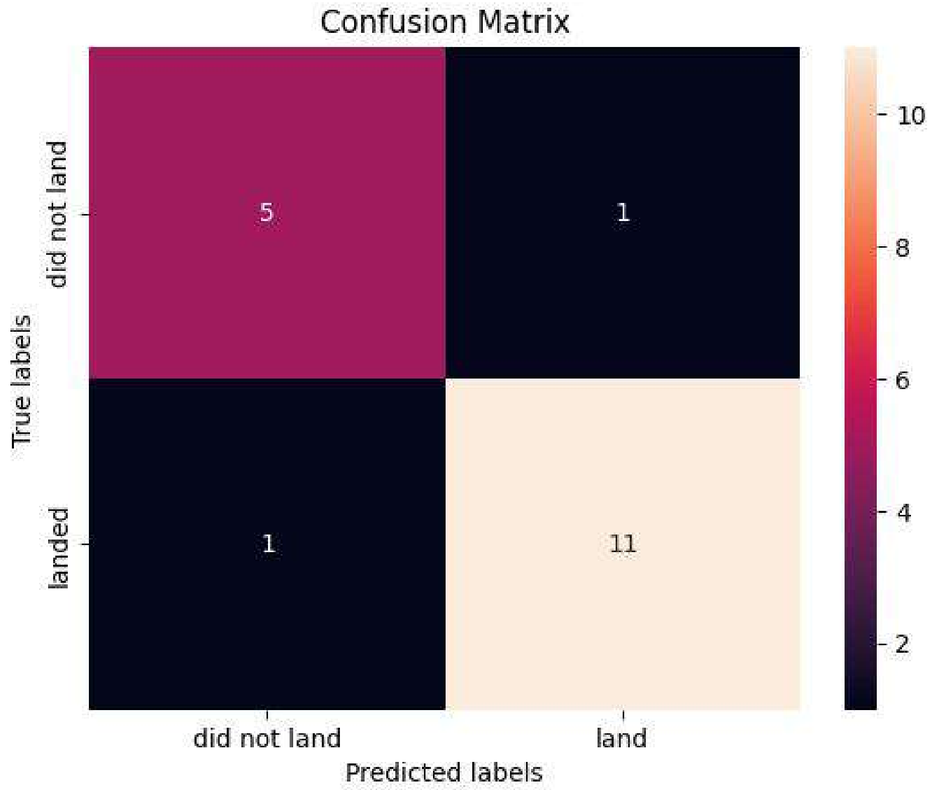
The bar chart illustrates the accuracy performance of the four classification models.

- **The Decision Tree** model outperformed the others with the highest accuracy at around 88.9%.
- All KNN, SVM, and Logistic Regression performed equally well with an accuracy of around 83.4%.



Confusion Matrix - Decision Tree

The model has a strong performance, with only two incorrect predictions out of 18. The majority of the landing outcomes were predicted accurately.



Conclusion

This project successfully gathered and processed SpaceX Falcon 9 launch data to predict the likelihood of successful first-stage landings. Since the reusability of the first stage plays a major role in reducing overall launch costs, being able to predict the success of these landings offers valuable insights for optimising launch operations. These predictions can also give alternative providers a competitive edge when bidding for rocket launches. Future work could involve applying more advanced machine learning techniques to further improve the accuracy of these predictions and enhance the cost-efficiency of space missions.

Thank you!

Wening Dyah Locitaresmi