IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Presenter: Atsushi Kobayashi
Date: 8 April 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- This analysis explored the key factors influencing launch success probability by plotting and analysing the relationships between multiple variables using EDR. The findings revealed unique trends across the different variables.
- All launch sites were located along the east and west coasts of the United States and near the equator.
- A successful visualization of launch success probabilities was achieved through a Folium map.
- The Decision Tree Model emerged as the most effective model for this project, achieving an accuracy rate of 83.33%.
- Looking ahead, integrating launch sites and optimising flight numbers based on relationships with total Payload Mass and success probability could lead to significant cost savings. Furthermore, the Decision Tree Model holds promise for enhancing launch success rates in the future.

Introduction

This project analyses SpaceX and aims to understand its launch operations and cost structure. The goal is to predict whether the first stage of the Falcon 9 rocket will land successfully and be reused. Additionally, the project seeks to explore opportunities for reducing launch costs, improving success rates, and developing efficient competitive strategies. As the demand for rocket launches continues to grow, optimising costs and success rates is key to remaining competitive in the commercial space industry. Therefore, this analysis is of significant importance.

Section 1

Methodology

Methodology

Executive Summary

- Data was collected from the SpaceX API using a GET request and filtered to include only Falcon 9 launches. The dataset was cleaned and missing values were addressed.
- Exploratory Data Analysis (EDA) was performed using Folium for geographic mapping and Plotly Dash for interactive visual analytics. SQL queries helped explore relationships between key variables.
- For predictive analysis, Logistic Regression, SVR, Decision Tree, and KNN models were built and tuned. Each model was evaluated using accuracy and validated with Confusion Matrices to assess performance and determine the best model for predicting launch success.

Data Collection

- Data Collection Process
  In this capstone project, data was collected from SpaceX's public API, which provides detailed information about Falcon 9 rocket launches, including launch dates, payload, launch sites, landing outcomes, and more.
- Data Preparation
  The collected data was cleaned and formatted to ensure consistency and accuracy. Missing values were handled appropriately, and data types were adjusted to support analysis.
- Purpose of Data Collection
  The data is used to predict the success of the Falcon 9 first-stage landing. Successful landings significantly reduce launch costs by enabling reuse, offering SpaceX a competitive advantage.

Data Collection – SpaceX API

Request and parse the SpaceX launch data using the GET request

Data collection	Data construction

↓

Filter the dataframe to only include Falcon 9 launches

Data wrangling

↓

Dealing with Missing Values

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/jupyter_labs_spacex_data_collection_api.ipynb

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Missing values were examined with .isnull().sum().
- The mean was calculated with .mean().
- The missing values were replaced by the mean with relace().

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/jupyter_labs_spacex_data_collection_api.ipynb

EDA with Data Visualization

The relationships between each of the following variables were plotted and visualised to explore the relationships between them.

- Flight Number and Launch Site
- Payload Mass and Launch Site
- Success rate of each Orbit type
- Flight Number and Orbit type
- Payload Mass and Orbit type

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/Exploring%20and%20Preparing%20Data.ipynb

EDA with SQL

If it can be determined whether the first stage of Space X's rocket will land, the cost of the launch can be determined. For this purpose, the researcher entered SQL codes to answer the following queries:

- List Boosters with Successful Drone Ship Landings and Specific Payload Mass
- Count of Successful and Failure Mission Outcomes
- List Booster Versions with Maximum Payload Mass
- Records for Failure Landing Outcomes in 2015
- Rank Landing Outcomes by Count in a Date Range

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/SQL%20Notebook%20for%20Peer%20Assignment.ipynb

Build an Interactive Map with Folium

The success rate of a launch depends on many factors and may also depend on the initial position of the rocket's trajectory. The following tools were therefore used in Folium to analyse the location of existing launch sites in order to find the best locations for launch site construction:

- **folium.Map** was used to determine the location of NASA Johnson Space Centre.
- **folium.Circle** was used to add highlighted circular areas with text labels at specific coordinates.
- **folium.Circle** and **folium.Marker** were used and the location of the launch site was added.
- **folium.Polyline** were drawn from the launch site to the selected coastline.

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

The Plotly Dash application was built to enable real-time, interactive visual analysis of SpaceX launch data. This dashboard application includes pie charts and scatter plots.

- By creating a pie chart, it is possible to know the probability of success for each launch site.
- The scatter plots allow a visual observation of how the payload correlates with the mission results for the selected sites.

Predictive Analysis (Classification)

The following machine learning models were built to predict successful launches.

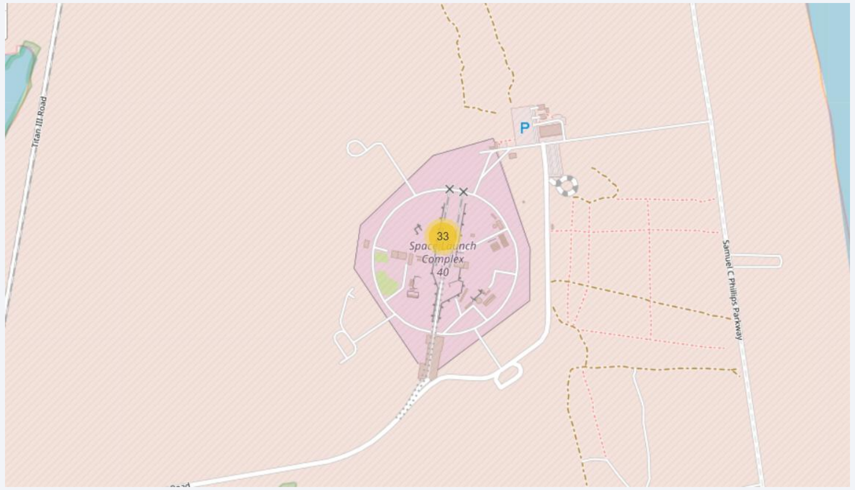
- **Logistic Regression**
- **Support Vector Machine**
- **Decision Tree**
- **K-Nearest Neighbors**

The accuracy of these parameters was tested to verify which was the best model.

https://github.com/Atsushi-DataScientist/Data-Science-project/blob/main/Machine%20Learning%20Prediction_accuracy%20rate%20of%2083.33%25.ipynb

Results

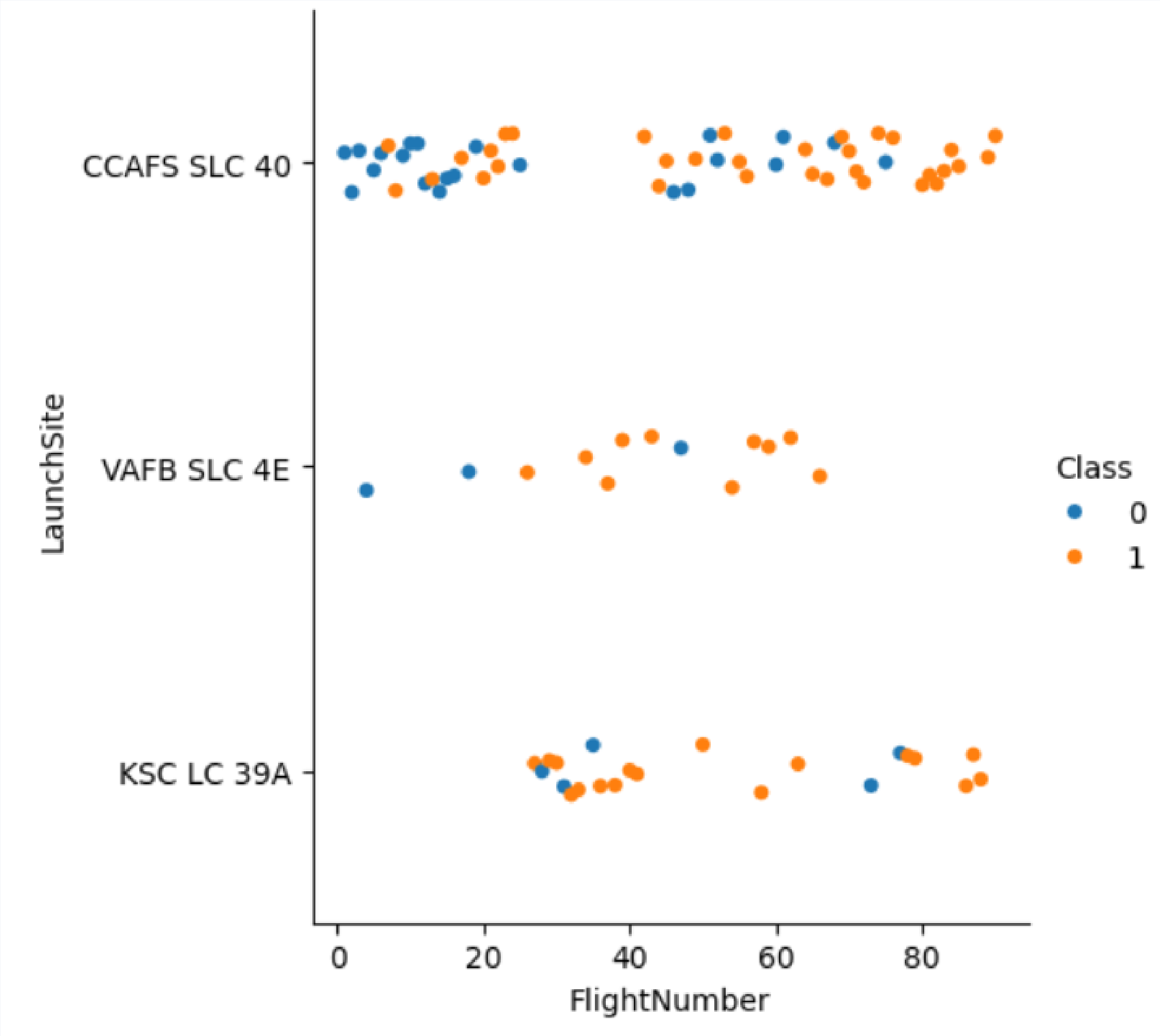
- Exploratory data analysis (EDA) performed the necessary work to understand the data, visualising it and analysing correlations. As a result, it made it possible to search for launch sites with a high success rate.
- It is now easier to zoom in or out on a favourite area of data and also on a map to analyse the location of successful launch sites.
- Machine learning models were built and tested for accuracy, and the decision tree model was the best model with an accuracy of 83.33%.



Section 2

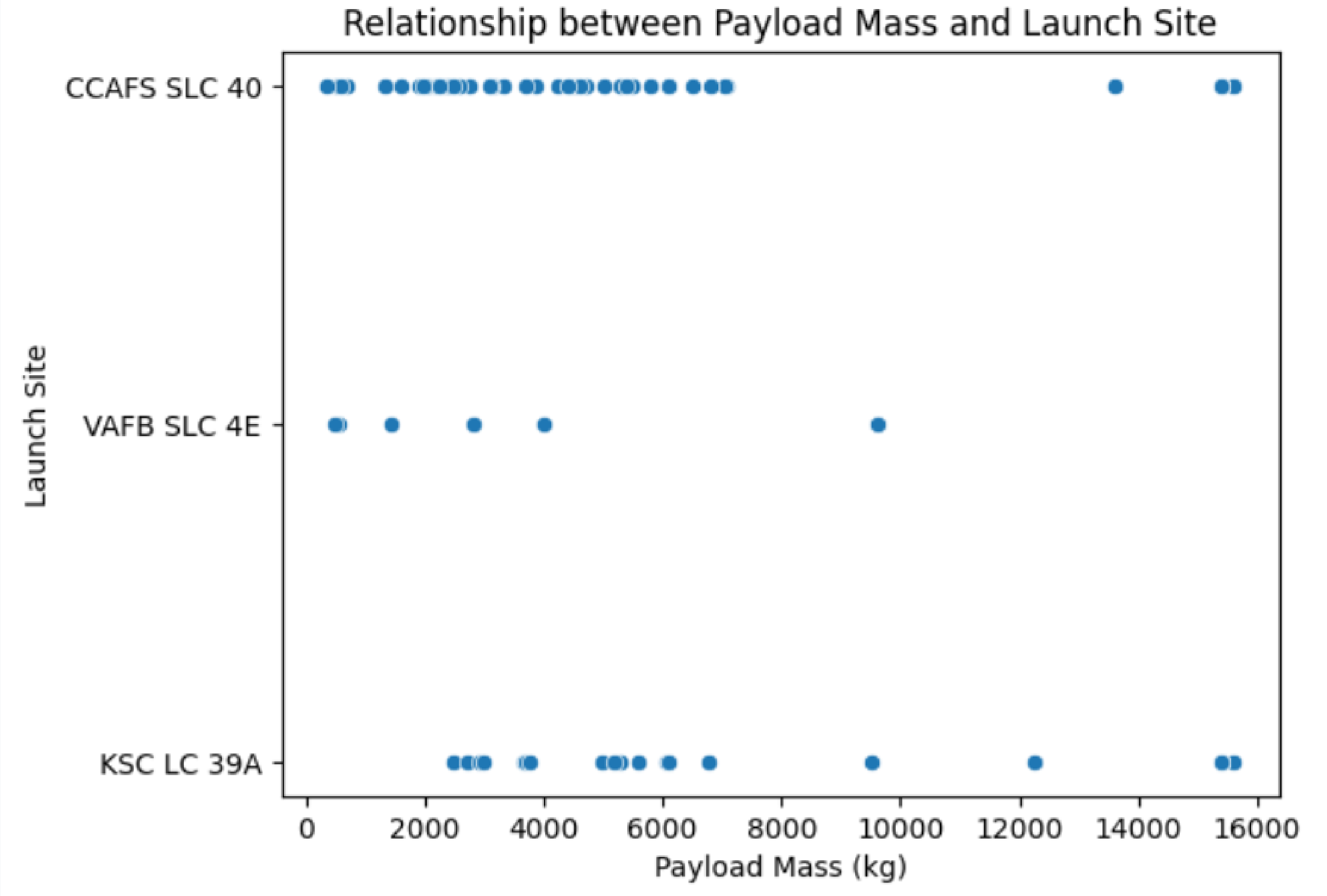
Insights drawn from EDA

Flight Number vs. Launch Site



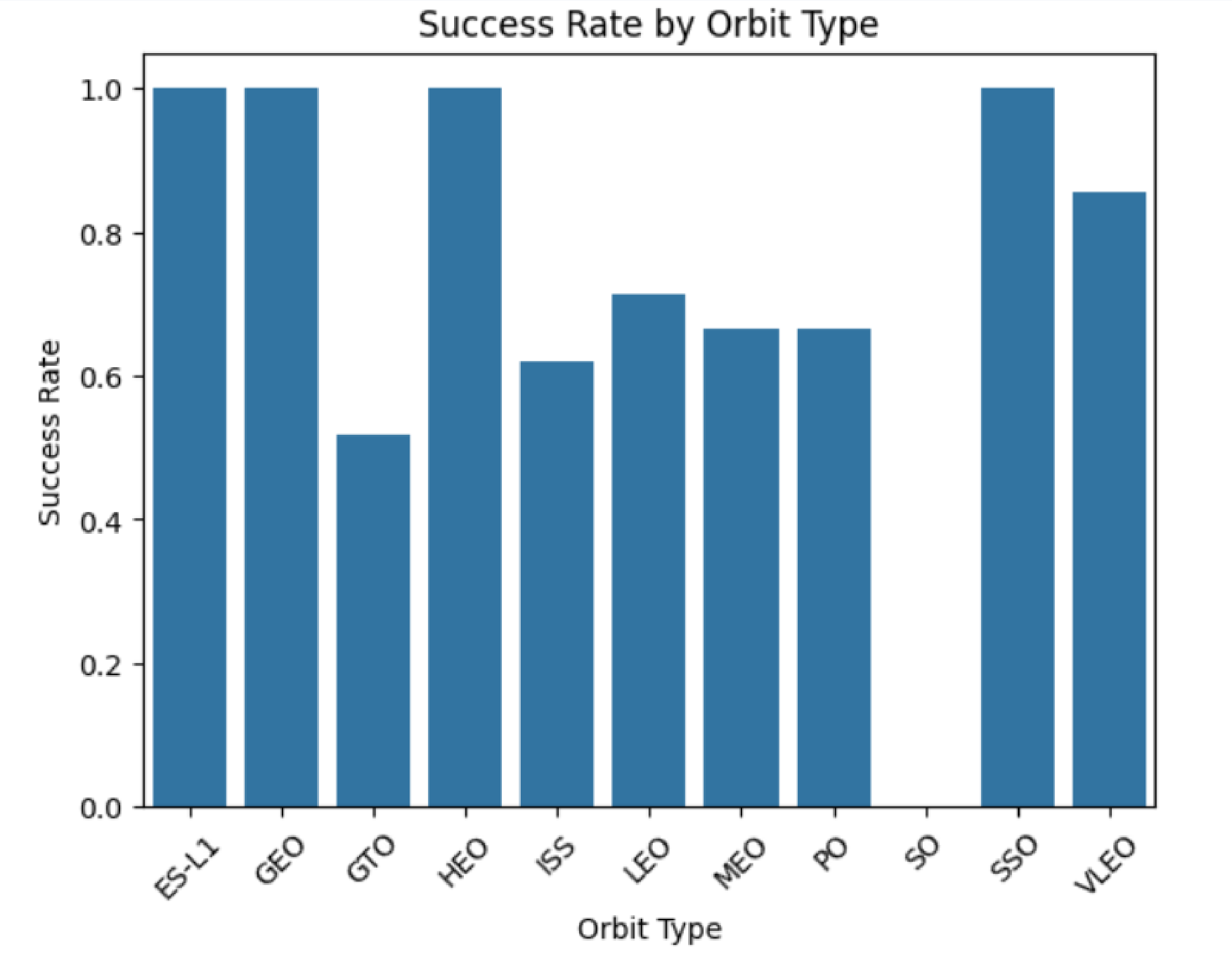
As the FlightNumber increases, the frequency of successful landings (class = 1) appears to increase across all sites. This suggests that SpaceX improved over time in landing their first stage.

Payload vs. Launch Site



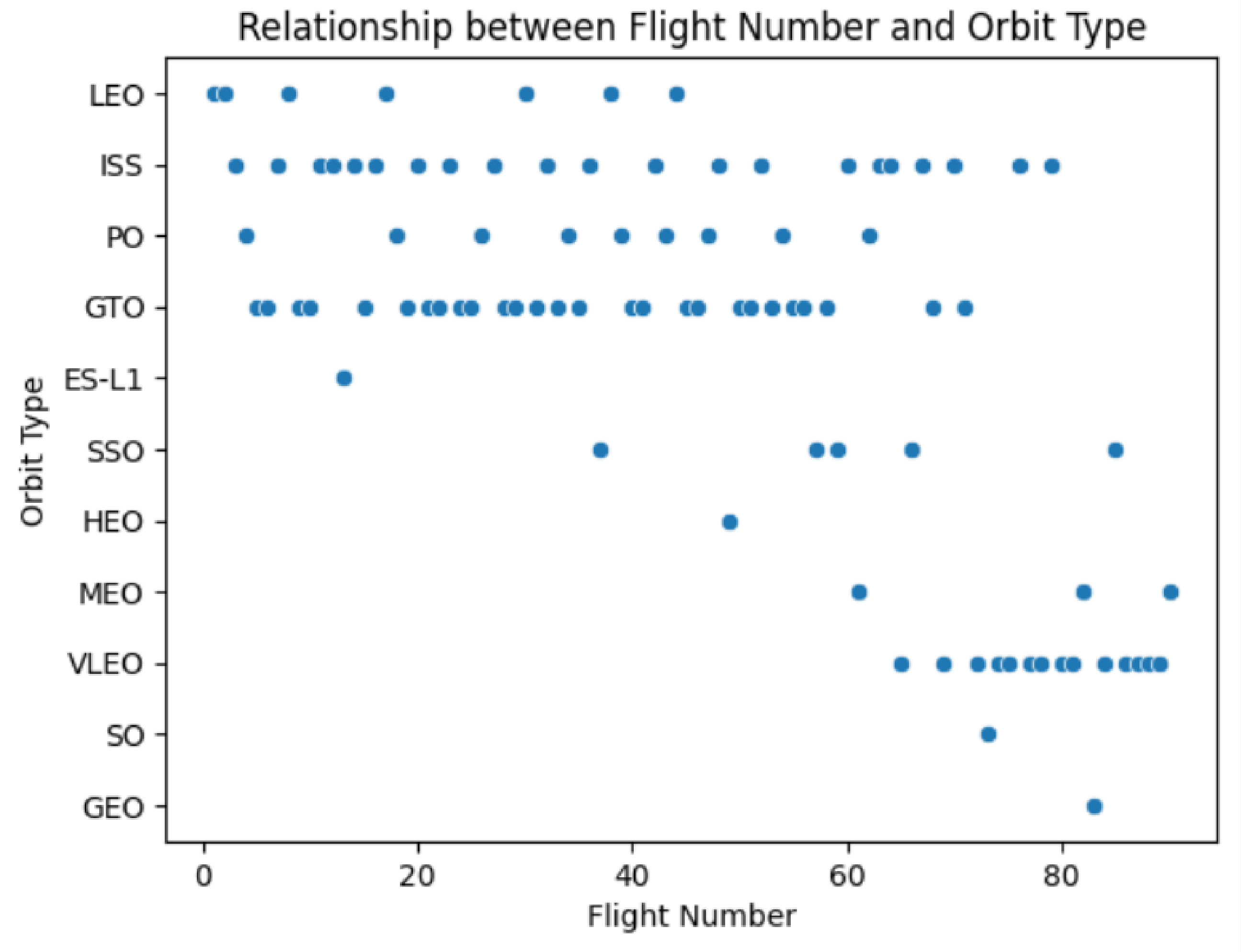
This scatter plot shows that payloads under 7000 kg are the most commonly launched across all sites. CCAFS SLC 40 is used most frequently and supports both light and very heavy payloads, while KSC LC 39A handles a wide range of payload masses. VAFB SLC 4E has fewer launches, typically with lighter payloads.

Success Rate vs. Orbit Type



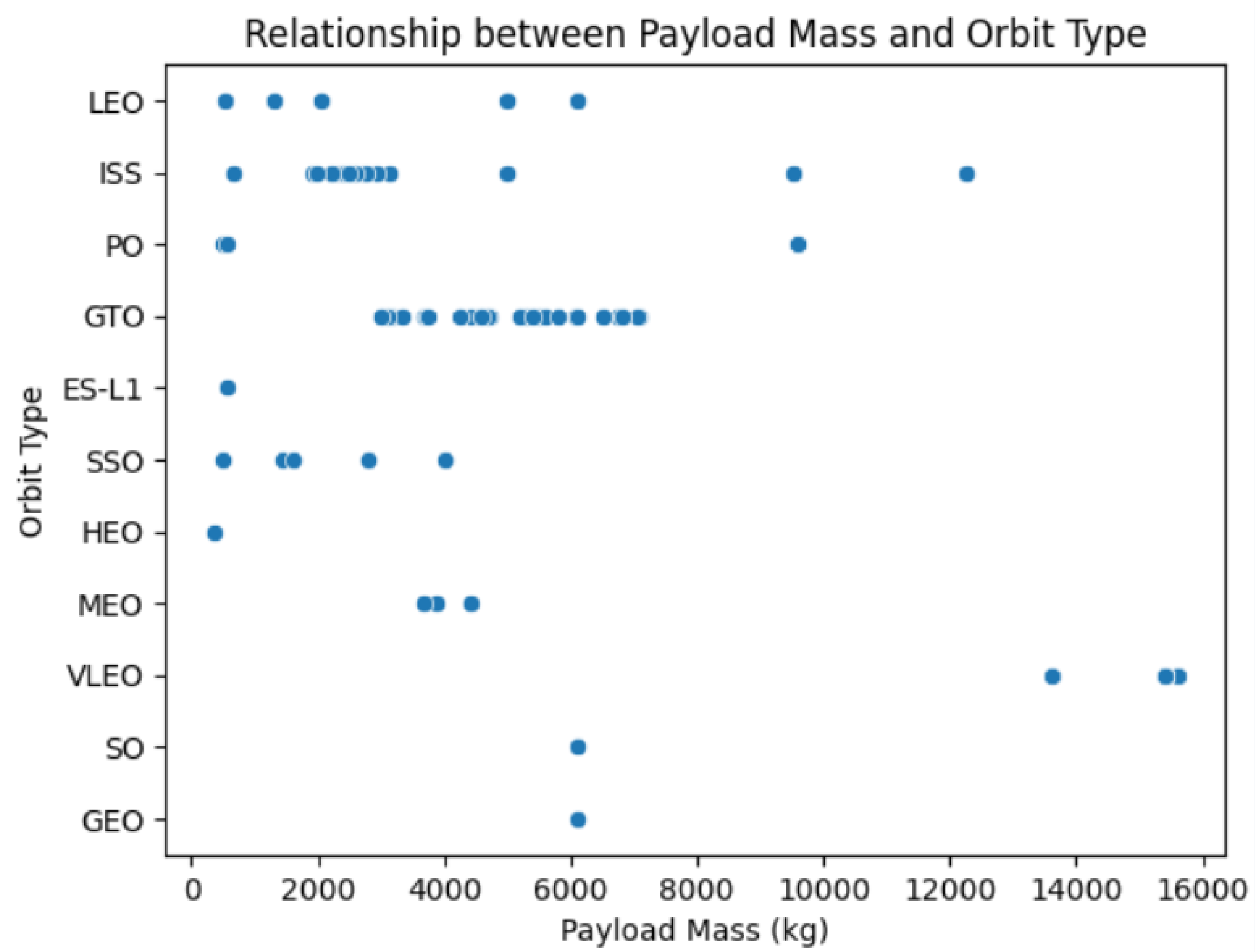
Some orbits (e.g., GEO, HEO, SSO) show consistent success; Orbits like GTO and ISS have more varied success — possibly due to mission complexity or historical launch phases; SO had no successful missions, highlighting a potential challenge or a rare mission case.

Flight Number vs. Orbit Type



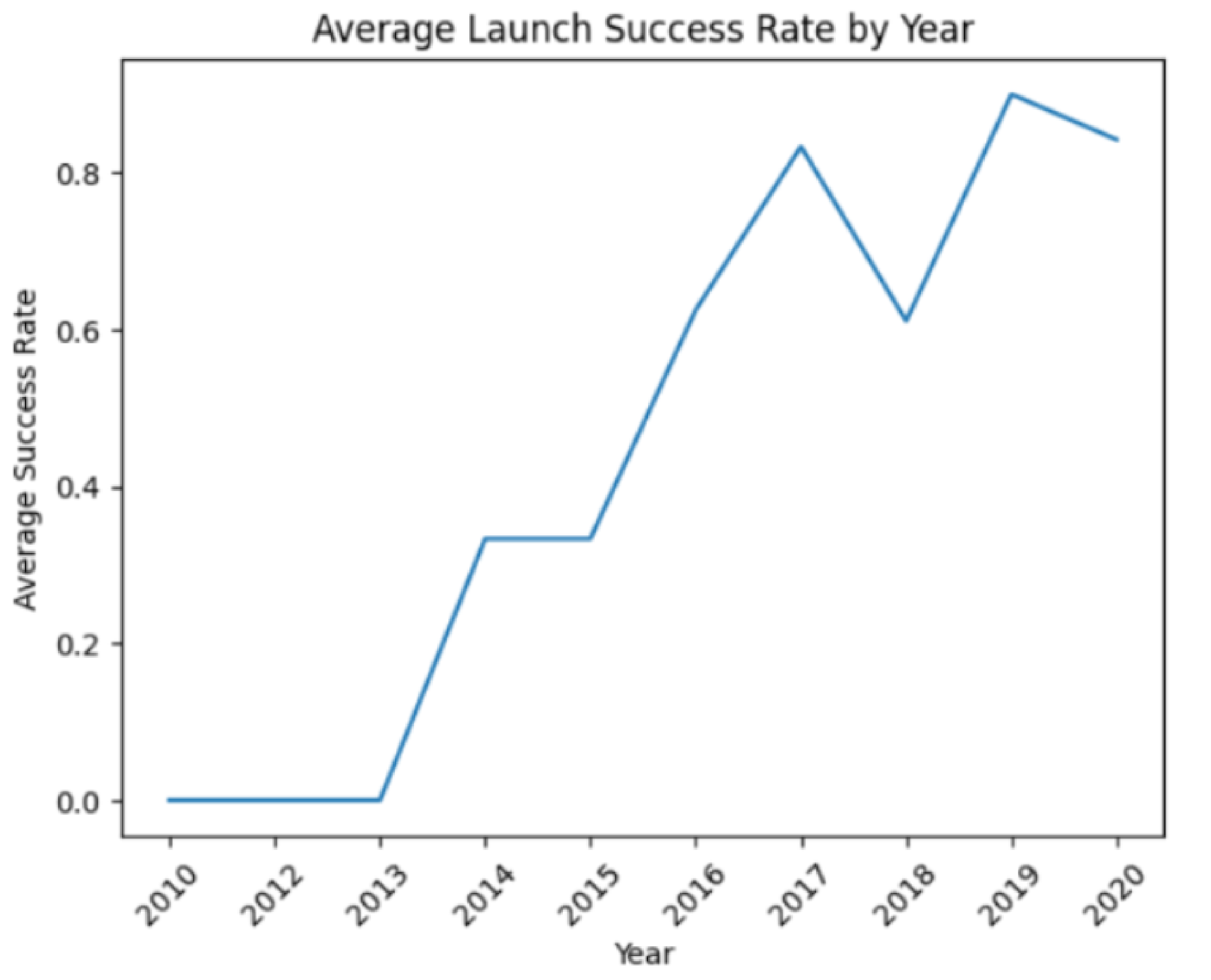
It was found that the orbit types LEO, ISS, PO, and GTO are spread across a wide range of flight numbers, while other orbit types are operated within more limited flight number ranges. Specifically, when looking from ES-L1 to GEO on the Y-axis of the graph, there is a clear trend of increasing flight numbers over time.

Payload vs. Orbit Type



Overall, it was found that the Orbit type is determined by the total amount of Payload Mass. In particular, GTO launched rockets in the range of 2,000 to 7,000 kg. Additionally, ES-L1 only handles very light Payload Mass, while SO and GEO each launched a single rocket with 6,000 kg. This insight will be helpful in the future for integrating or distributing Payload Mass based on Orbit type.

Launch Success Yearly Trend



Looking at the trend in annual success rates, the average success rate has generally increased from 2013 to 2020. However, there was a drop in the success rate in 2018, and it would be necessary to investigate the cause of this decline.

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

In [14]:

```
import pandas as pd
from sqlalchemy import create_engine

# Create an in-memory SQLite database
engine = create_engine('sqlite://', echo=False)

# Assuming df is your DataFrame
df.to_sql("SPACEXTBL", con=engine, if_exists='replace', index=False)

# Execute the SQL query
query = 'SELECT DISTINCT "Launch_Site" FROM SPACEXTBL'
result = pd.read_sql_query(query, con=engine)

# Display the result
print(result)
```

```
    Launch_Site
0   CCAFS LC-40
1   VAFB SLC-4E
2    KSC LC-39A
3  CCAFS SLC-40
```

Four Launch Sites with unique names were extracted.

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [19]:
```
%%sql
SELECT *
FROM SPACEXTBL
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

Out[19]:

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

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [21]:
```
%%sql
SELECT SUM("Payload_Mass") AS Total_Payload_Mass
FROM SPACEXTBL
WHERE "Launch_Organization" = 'NASA (CRS)';
```

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

Out[21]:

Total_Payload_Mass
None

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [22]:

```
%%sql
SELECT AVG("Payload_Mass") AS Average_Payload_Mass
FROM SPACEXTBL
WHERE "Booster_Version" = 'F9 v1.1';
```

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

Out[22]:

Average_Payload_Mass
0.0

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [23]:

```
%%sql
SELECT MIN("Date") AS First_Successful_Landing
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

Out[23]:

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [24]:

```
%%sql
SELECT "Booster_Name"
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (drone ship)'
AND "Payload_Mass" > 4000
AND "Payload_Mass" < 6000;
```

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

Out[24]:

"Booster_Name"

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [25]:
```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS Total_Count
FROM SPACEXTBL
GROUP BY "Mission_Outcome";
```

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

Out[25]:

Mission_Outcome	Total_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Task 8

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

In [26]:

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTBL
WHERE "Payload_Mass" = (
    SELECT MAX("Payload_Mass")
    FROM SPACEXTBL
);
```

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

Out[26]:

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006
F9 v1.0 B0007
F9 v1.1 B1003
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1010
F9 v1.1 B1012
F9 v1.1 B1013
F9 v1.1 B1014
F9 v1.1 B1015
F9 v1.1 B1016
F9 v1.1 B1018
F9 FT B1019
F9 v1.1 B1017
F9 FT B1020
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1024
F9 FT B1025.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1031.1
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 FT B1034
F9 FT B1035.1
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1037
F9 B4 B1039.1
F9 FT B1038.1
F9 B4 B1040.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 FT B1035.2
F9 FT B1036.2
F9 B4 B1043.1
F9 FT B1032.2
F9 FT B1038.2
F9 B4 B1044
F9 B4 B1041.2
F9 B4 B1039.2
F9 B4 B1045.1
F9 B5 B1046.1
F9 B4 B1043.2
F9 B4 B1040.2
F9 B4 B1045.2
F9 B5B1047.1
F9 B5B1048.1
F9 B5 B1046.2
F9 B5B1049.1
F9 B5 B1048.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1050
F9 B5B1054
F9 B5 B1049.2
F9 B5 B1048.3
F9 B5B1051.1
F9 B5B1056.1
F9 B5 B1049.3
F9 B5 B1051.2
F9 B5 B1056.2
F9 B5 B1047.3
F9 B5 B1048.4
F9 B5B1059.1
F9 B5 B1056.3
F9 B5 B1049.4
F9 B5 B1046.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1059.2
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5B1058.1
F9 B5 B1049.5
F9 B5 B1059.3
F9 B5B1060.1
F9 B5 B1058.2
F9 B5 B1051.5
F9 B5 B1049.6
F9 B5 B1059.4
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5B1062.1
F9 B5B1061.1
F9 B5B1063.1
F9 B5 B1049.7
F9 B5 B1058.4

2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [27]:

```
%%sql
SELECT SUBSTR("Date", 6, 2) AS Month, "Booster_Version", "Launch_Site"
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Failure (drone ship)'
AND SUBSTR("Date", 1, 4) = '2015';
```

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

Out[27]:

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

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

Task 10

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.

In [28]:

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

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

Out[28]:

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

Section 3

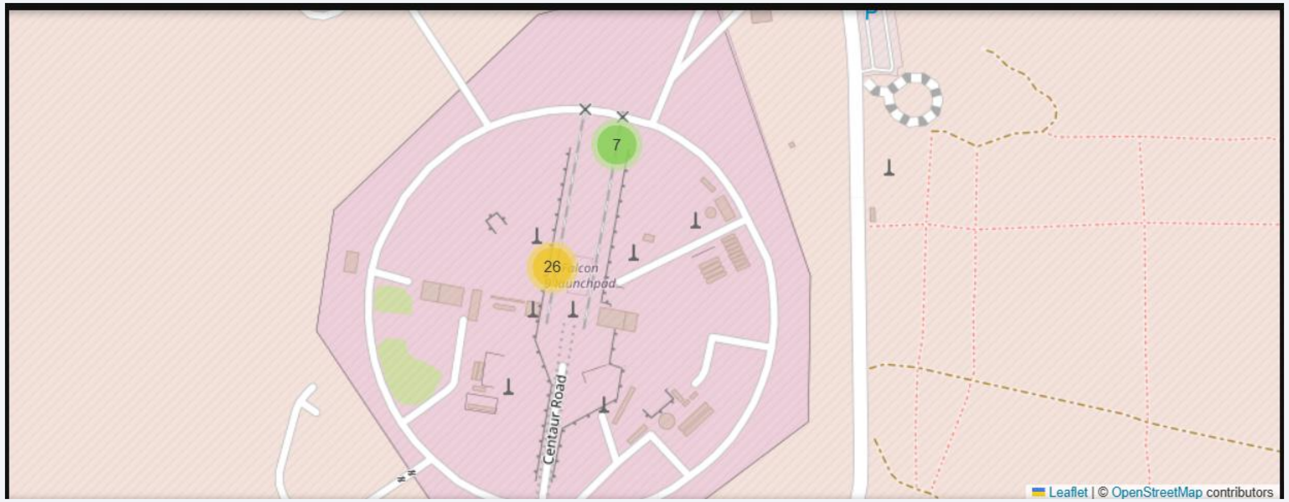
Launch Sites Proximities Analysis

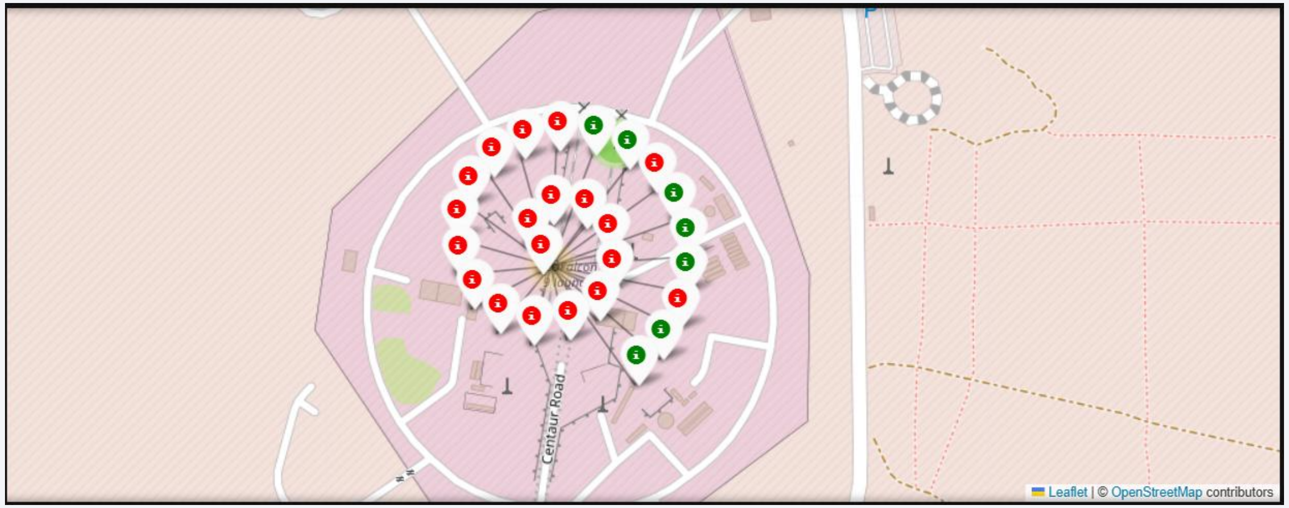
<Location of rocket launch sites>

All launch sites are located near the east and west coasts and, moreover, near the equator.

<Visualisation of launch results by MarkerCluster>

The location of the launch sites and the icons that stand for success or failure can now be visually analysed to determine which launch sites have a higher probability of success.





<The shortest distance from a nearby airfield to the launch site>

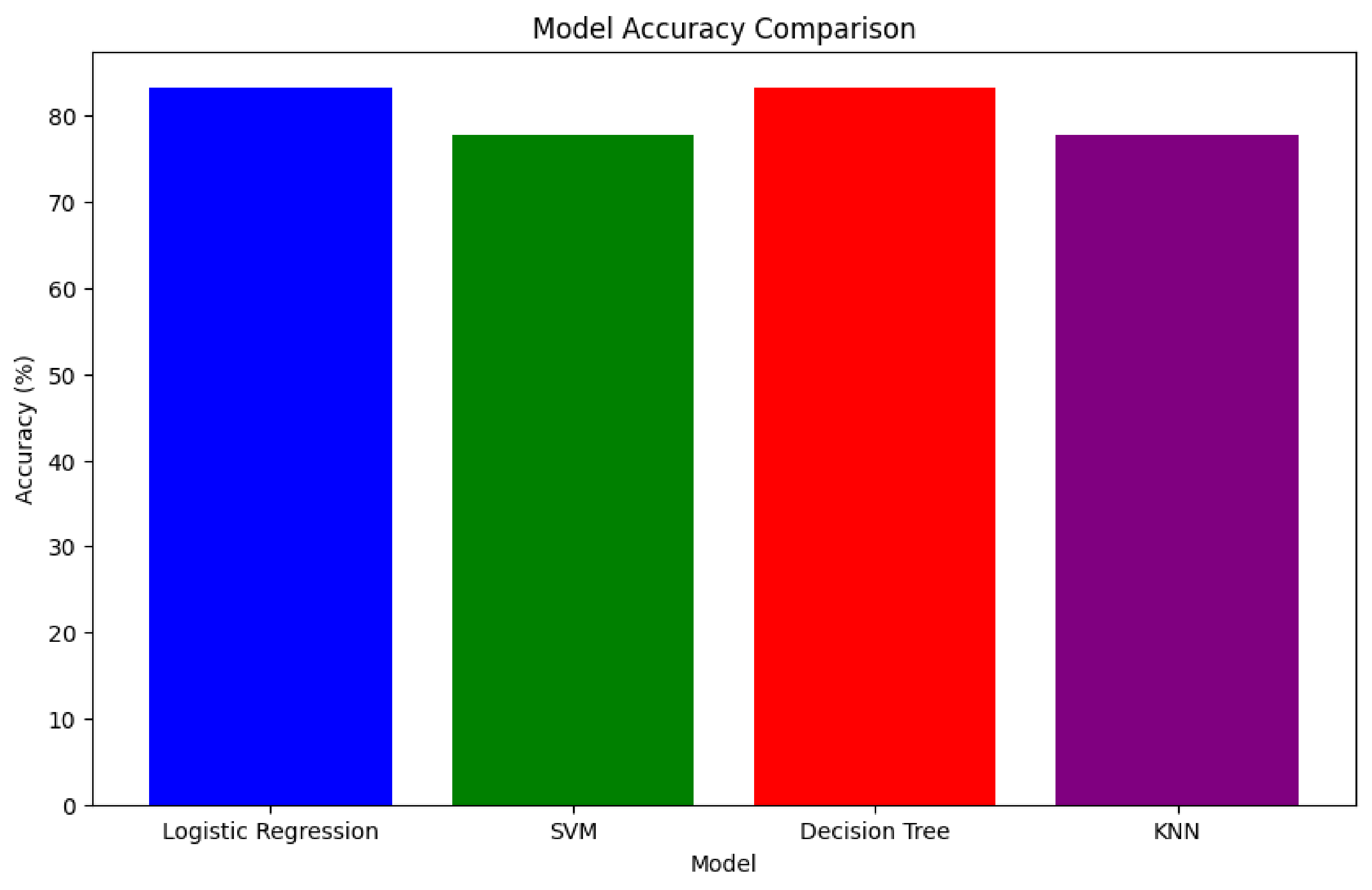
The generated folium map showed the latitude and longitude of the selected launch site and neighbouring airfields, etc., and also calculated the shortest distance between them to be 11.44 km.

Section 4

Predictive Analysis (Classification)

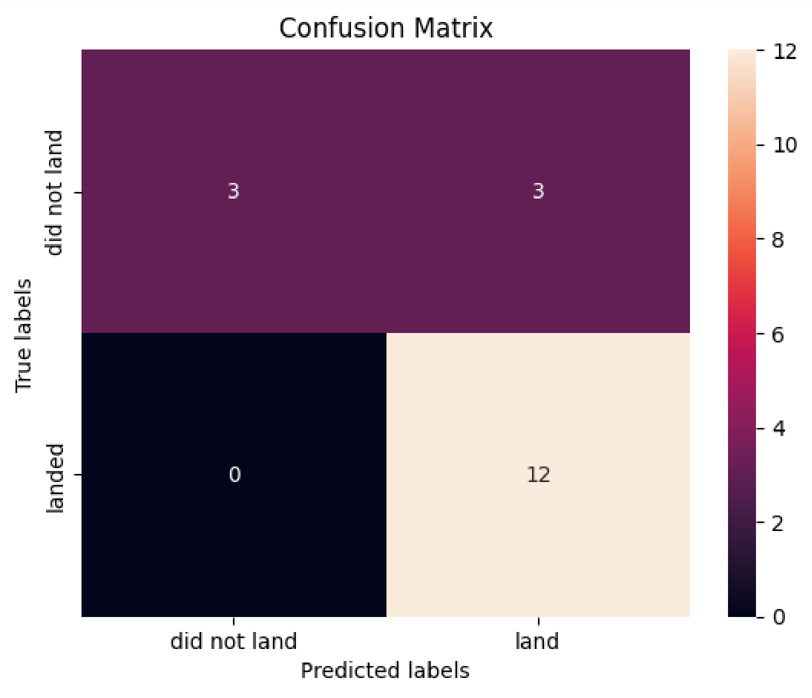
Classification Accuracy

The accuracy of the four machine learning models was tested, and both Logistic Regression and Decision Tree achieved the same accuracy of **83.33%**. At this stage, these two models were considered the best; however, **the Decision Tree was ultimately selected as the best model based on the confusion matrix** shown on the next page.

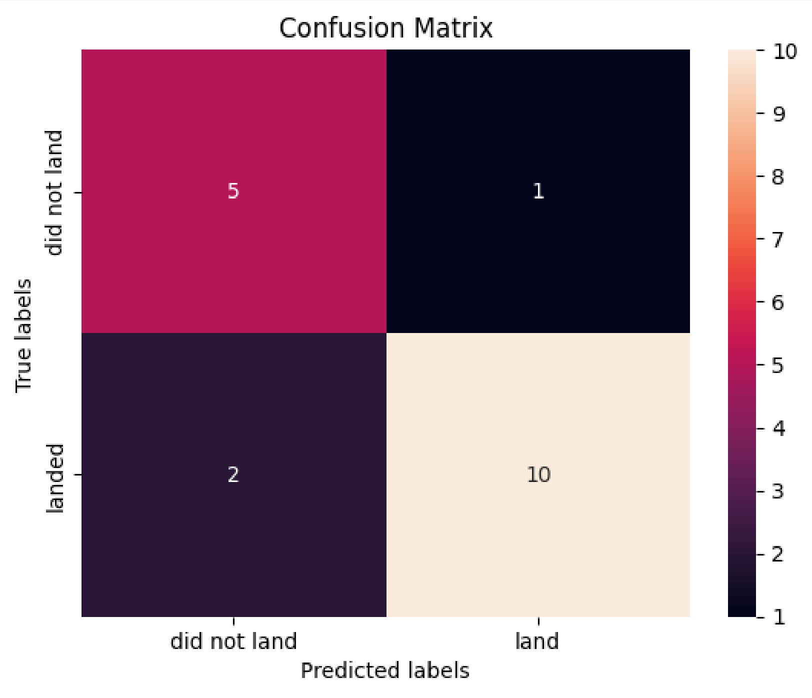


Confusion Matrix

Logistic Regression Model



Decision Tree Model



While the Logistic Regression Model demonstrated a high success prediction rate, it produced three false positives. This poses a problem, as reducing the number of failed launches is crucial given the high cost of SpaceX launches. On the other hand, the Decision Tree Model also showed a high success prediction rate but with only one false positive—fewer than the Logistic Regression Model. Therefore, the **Decision Tree Model is considered the best-suited model for this project.**

Conclusions

- To explore the factors influencing launch success probability, a relationship analysis between multiple variables was performed using EDR, resulting in unique trends observed between each pair of variables.
- All launch sites were located along the east and west coasts of the American continent and near the equator.
- Visualization of launch success probabilities was successfully achieved on a Folium map.
- The best model for this project was the Decision Tree Model, with an accuracy of 83.33%.
- In the future, cost-cutting can be expected by considering the integration of launch sites and the swapping of flight numbers based on relationships with total Payload Mass and success probability. Additionally, the Decision Tree Model is expected to improve launch success probability.

Thank you!