# Exploring SpaceX Launches: Predictions and Insights

Project Report

Martins Nnamchi

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

# Executive Summary

This report details a comprehensive analysis of **Space X's** rocket launches with the objective of predicting the likelihood of the first stage landing successfully. I employ Web Scraping and a REST API for data collection, followed by data wrangling, exploratory data analysis with static and interactive visualizations, and machine learning predictions.

Four machine learning models were built for the prediction, namely **Decision Tree**, **Logistic Regression**, **K-Nearest Neighbors**, **and Support Vector Machine**. The Decision Tree model delivered the best performance with an accuracy of 87%. The models developed in this project will provide insights that will aid **Space Y** in strategizing its market entry and operational planning, ultimately providing a competitive edge in the burgeoning commercial space industry.

# Introduction

The commercial space age has arrived, with private companies revolutionizing space travel. Among them, *SpaceX* stands out with its achievements in reducing launch costs through the reuse of the Falcon 9 rocket's first stage. This project, part of the **IBM Professional Data Science Certificate Program**, involves predicting the likelihood of successful landings for these first stages.

In this capstone, I act as a data scientist for *SpaceY*, a new competitor founded by billionaire industrialist *Allon Musk*. I analyze *SpaceX*'s launch data and develop machine learning models to forecast first-stage landing successes. My predictions will help determine launch costs and provide strategic insights for *SpaceY*'s market entry and operational planning. Through the application of advanced data science techniques to this real-world problem, this project demonstrates the significant potential of machine learning in the rapidly evolving field of space exploration.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Describing how data was collected
- Perform data wrangling
  - Describing how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Describing the building, tuning, and evaluation of classification models

# Data Collection

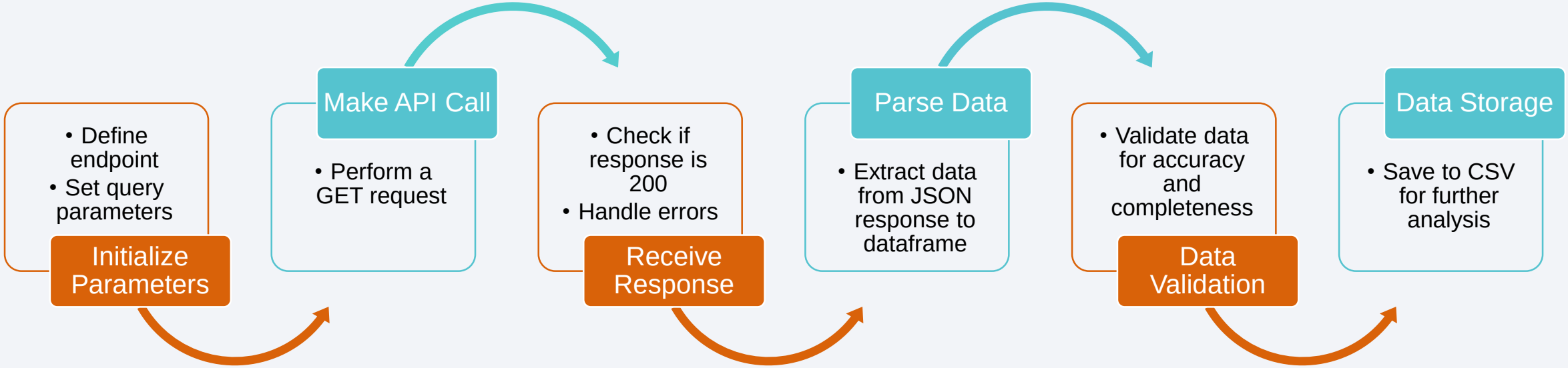
In this capstone project, I will be working with SpaceX launch data obtained from the following two sources:

- SpaceX REST API
  - This API provides data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Web scrapping related Wiki pages Using the Python BeautifulSoup package
  - The Wiki pages provide some HTML tables that contain valuable Falcon 9 launch records.

In the next two slides, the details of the two data-collection methods are provided.

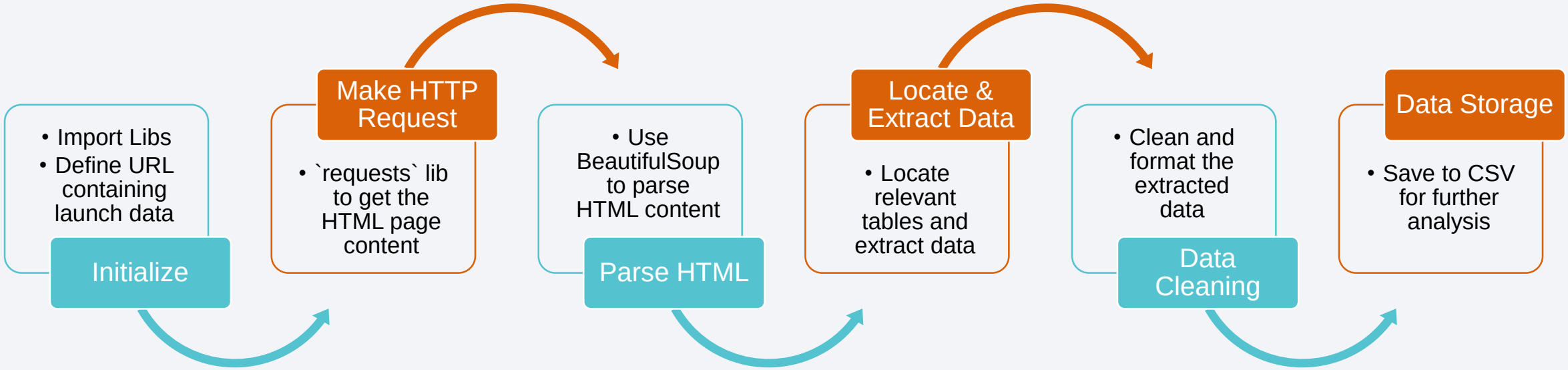
# Data Collection – SpaceX API

- The SpaceX REST API endpoints, or URL, start with api.spacexdata.com/v4/. There are different end points, e.g., :/capsules and /cores. For this project, I will be working primarily with the endpoint api.spacexdata.com/v4/launches/past
- GitHub URL of the completed SpaceX API calls notebook: Click here



# Data Collection – Web Scraping

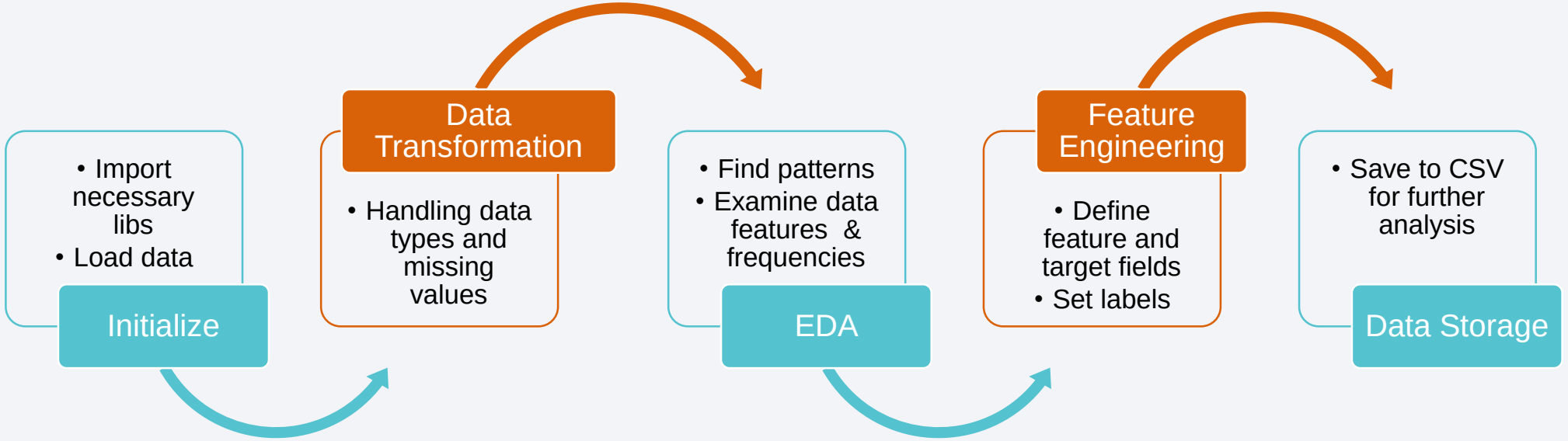
- To obtain further data, I employed the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records from the following Wiki page: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- GitHub URL of the Web Scraping notebook: Click here



# Data Wrangling

**Main Points**

- The column Outcome indicates if the first stage successfully landed and is designated as the target field Y.
- I perform some initial Exploratory Data Analysis (EDA) to find some patterns in the data.
- The target field Y is converted into Classes 0 or 1: 0 being bad landing outcome and 1 being good outcome.
- GitHub URL of the data wrangling notebook: Click here



# EDA with Data Visualization

I employed scatterplots, line charts, and bar plots to visualize the data. These tools provide a comprehensive overview of the data, facilitating better understanding and analysis of the complex datasets. Additionally, they are integral to EDA and model development in this project. Further details are provided below:

- Scatter Plots: I employed scatter plots to visualize the relationship between variables and find correlations. For example, FlightNumber vs. PayloadMass and FlightNumber vs LaunchSite.
- Line Chart: Line charts were employed to visualize the data points over a continuous time interval. For example, visualizing the success rate for different years.
- Bar Plots: Bar plots were employed to compare different categories of data. For example, comparing the success rates of different orbits.
- GitHub URL of the EDA with data visualization notebook: Click here

# EDA with SQL

Using `%sql` magic and SQLite3 in a Jupyter notebook, I performed some basic SQL queries to find patterns in the data. Here are some of the information retrieved:

- Names of the unique launch sites and 5 records where launch sites begin with the string 'CCA‘.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by booster version F9 v1.1.
- The date of the first successful landing outcome on ground pad.
- The booster names with successes in drone ship and whose payload masses are greater than 4000 but less than 6000.
- The names of the booster_versions with the maximum payload mass using a subquery.
- Specific information for the months in 2015.
- A ranking of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between two periods, in descending order.
- GitHub URL of EDA with SQL notebook: Click here

# Build an Interactive Map with Folium

I employed Folium to build interactive maps to analyze the launch site geo and proximities. First I created a Map object, to which I added the following objects:

- `folium.Circle()` draws a highlighted circle with a text label and specified radius around a location.
- `folium.Popup()` in the circle marker adds a popup with additional information.
- `folium.map.Marker()` adds a marker at a specific coordinate with an optional Icon as a text label.
- `MarkerCluster()` was employed to group nearby markers into clusters to improve map readability.
- `MousePosition()` was employed to obtain the coordinates for a point over which the mouse is hovering.
- `folium.Polyline()` was employed to draw a line between a launch site and surrounding locations.
- GitHub URL of the Folium interactive maps notebook: Click here

# Build a Dashboard with Plotly Dash

For stakeholders, I build an interactive dashboard that enables users to explore and manipulate data in a real-time way. This was achieved using Python Plotly Dash Package.

## Input components

- `A drop-down list` to filter the visualizations by ‘all’ launch sites or a specific launch site.
- `A range slider` to filter by the payload range.

## Graphs

- Pie chart and scatter plot.

## Interactions

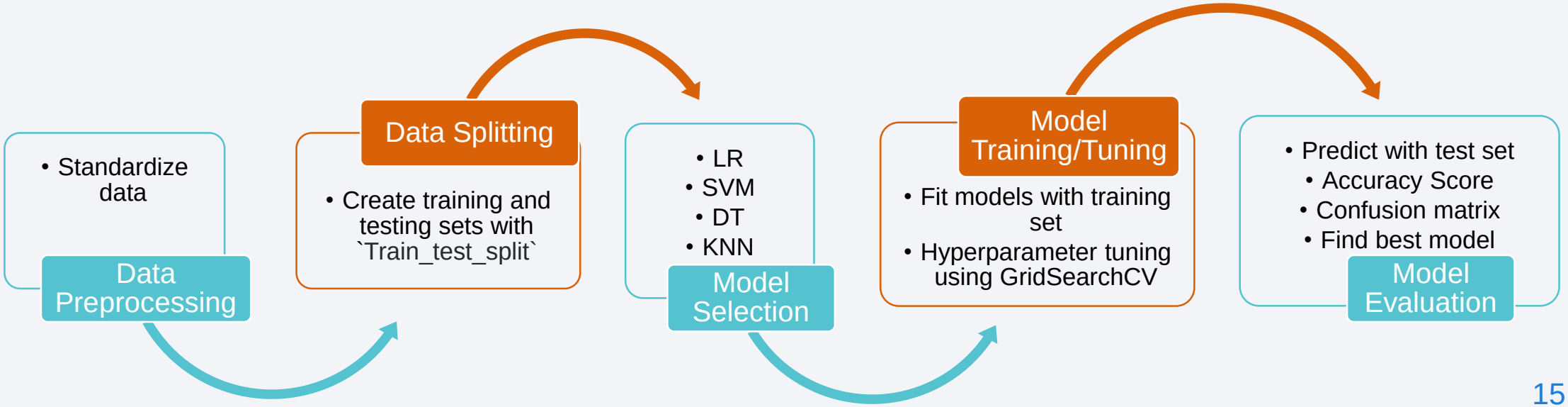
- Zoom-in, zoom-out, pan, filter, download as png, box select, and isolation by clicking.

- GitHub URL of Dash Python Script: Click here

# Predictive Analysis (Classification)

At this stage, I built a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully given the data from the preceding processes. The model-development process included preprocessing, data splitting, model selection/training, model tuning, and model evaluation. Four classification models were built, and their results were compared: Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors.

- GitHub URL of the data wrangling notebook: Click here

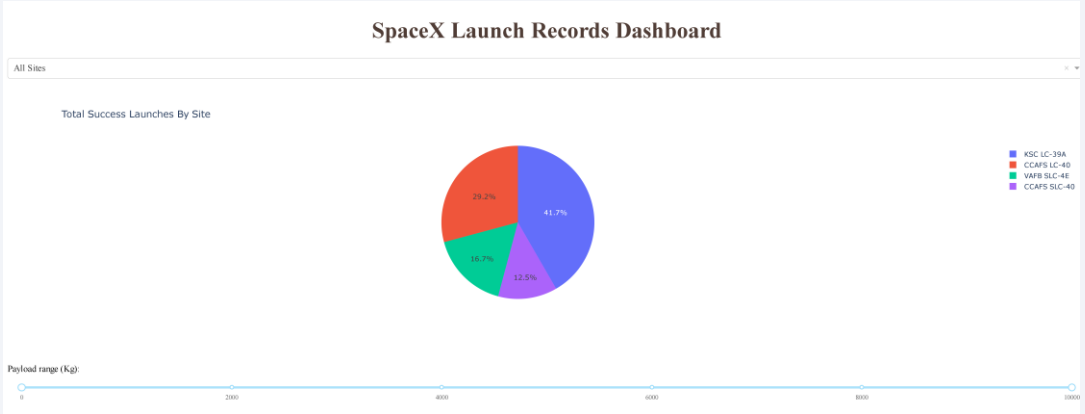


# Results

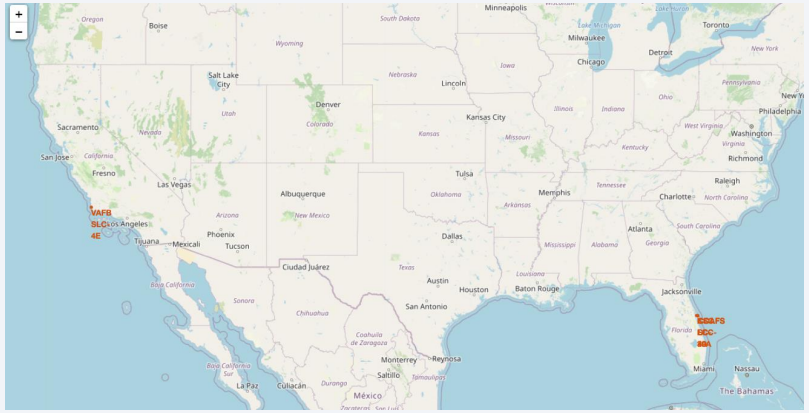
## Results of Exploratory Data Analysis

- 99% the missions were successful, i.e., mission outcomes.
- The first successful landing was recorded five years after the first launch, i.e., in 2015.
- The lowest success rate was recorded in the GTO orbit.
- For the VAFB-SLC launch site, no rockets were launched with payloads greater than 10,000.
- In 2015, two failures were recorded in the CCAFS LC-40 Launch Site in January and April.

## Interactive Analytics Demo Screenshots




Dash App




Launch Site Markers on Map

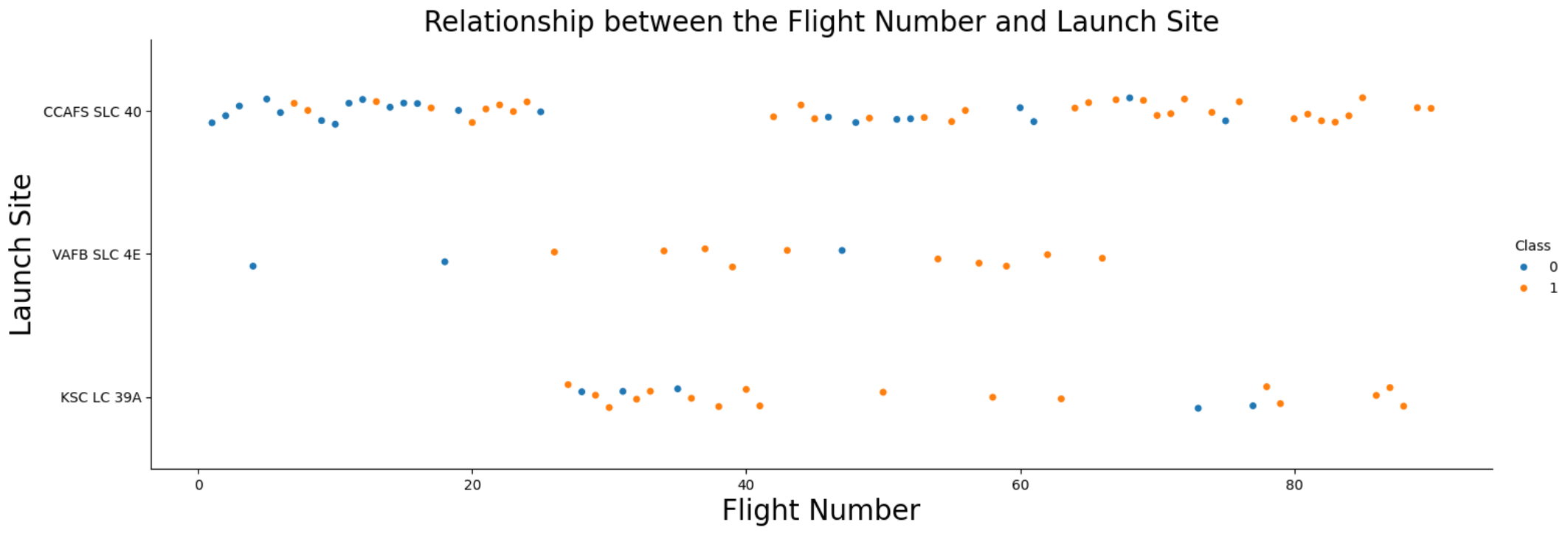
## Model Prediction Results

- Except the DT model, all models exhibited similar performances, with an accuracy score of ~83%.
- The DT model exhibited the best performance, with an accuracy score of ~88%.

Section 2

# Insights drawn from EDA

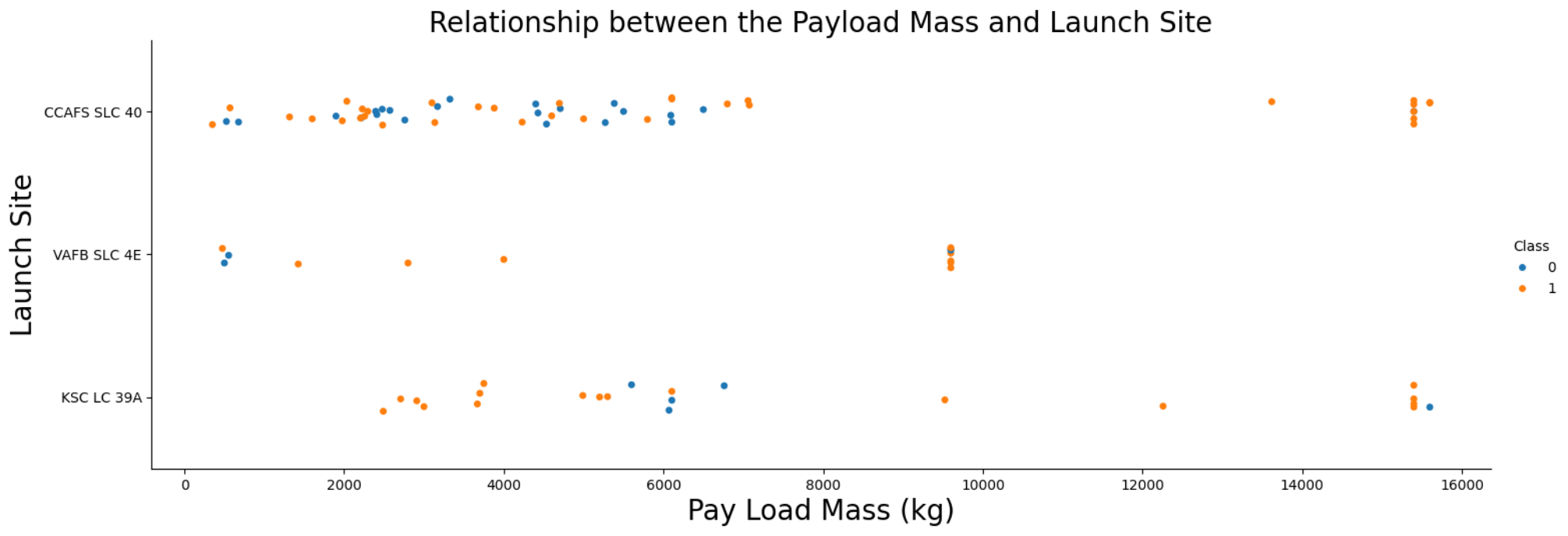
# Flight Number vs. Launch Site




From the figure above, we can make the following observations:

- The landing success rate on all three sites increases as the flight number increases.
- The highest proportion of successful landings was achieved on the CCAFS SLC 40 launch site.

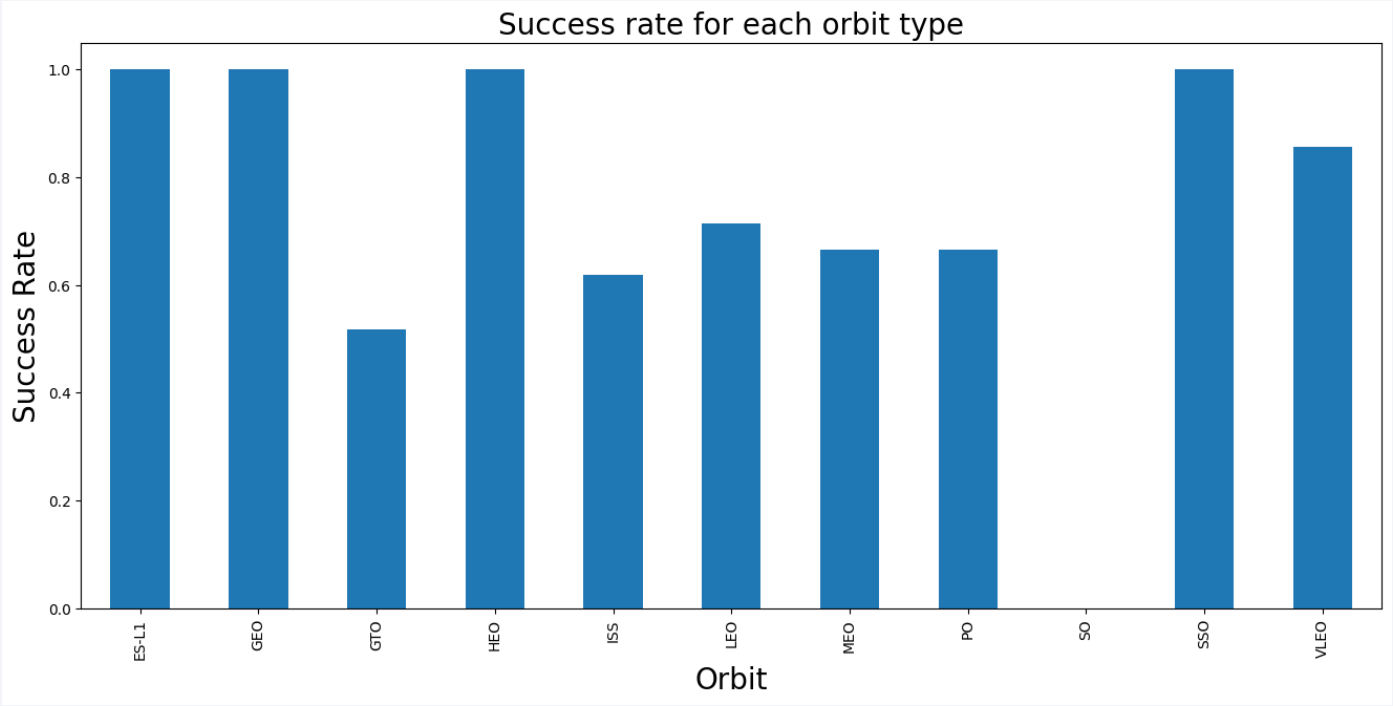
# Payload vs. Launch Site




From the figure above, we can make the following observations:

- Rockets with payload masses greater than 10,000 kg were not launched on the VAFB SLC 4E site.
- A 100% landing success rate was achieved on the CCAFS SLC 40 site for rockets with payload masses greater than12,000 kg.
- Generally, rockets with heavy payloads achieve high landing success rates on all sites for the first stage.

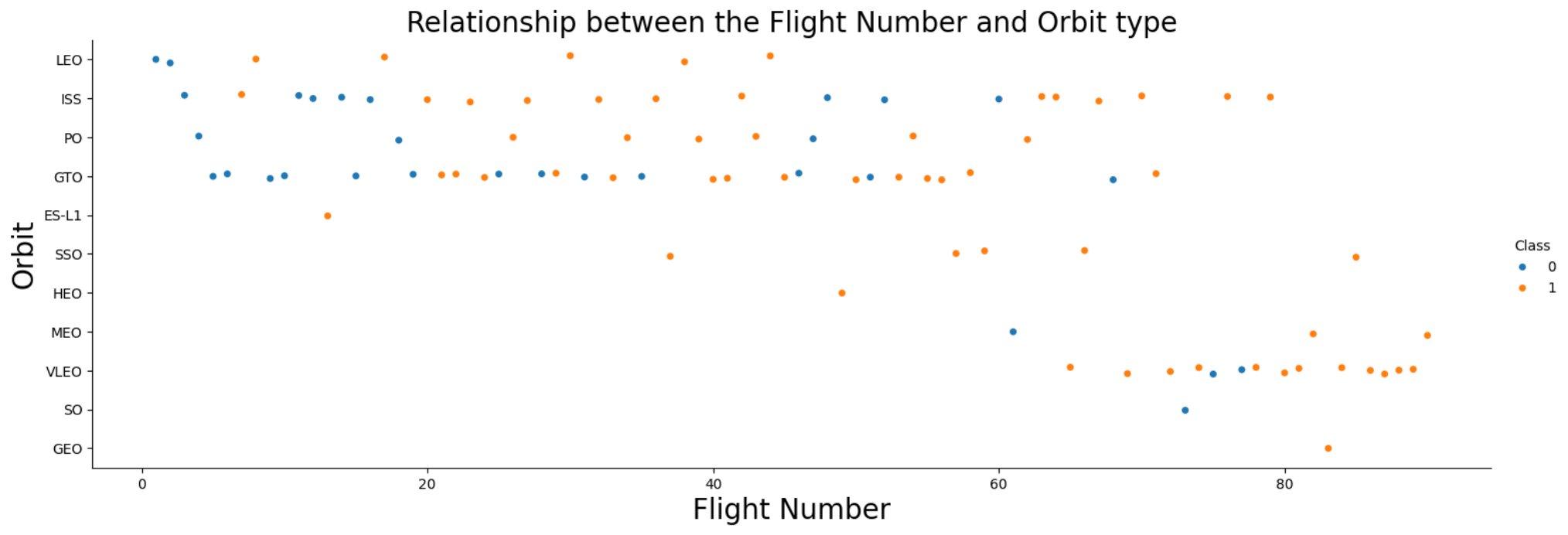
# Success Rate vs. Orbit Type




From the figure above, we can make the following observations:

- The lowest success rate was recorded in the GTO orbit.
- Similarly high success rates were recorded in the ES-L1, GEO, HEO, and SSO orbits.
- The success rate in all orbits, except the SO orbit, is at least 50%.

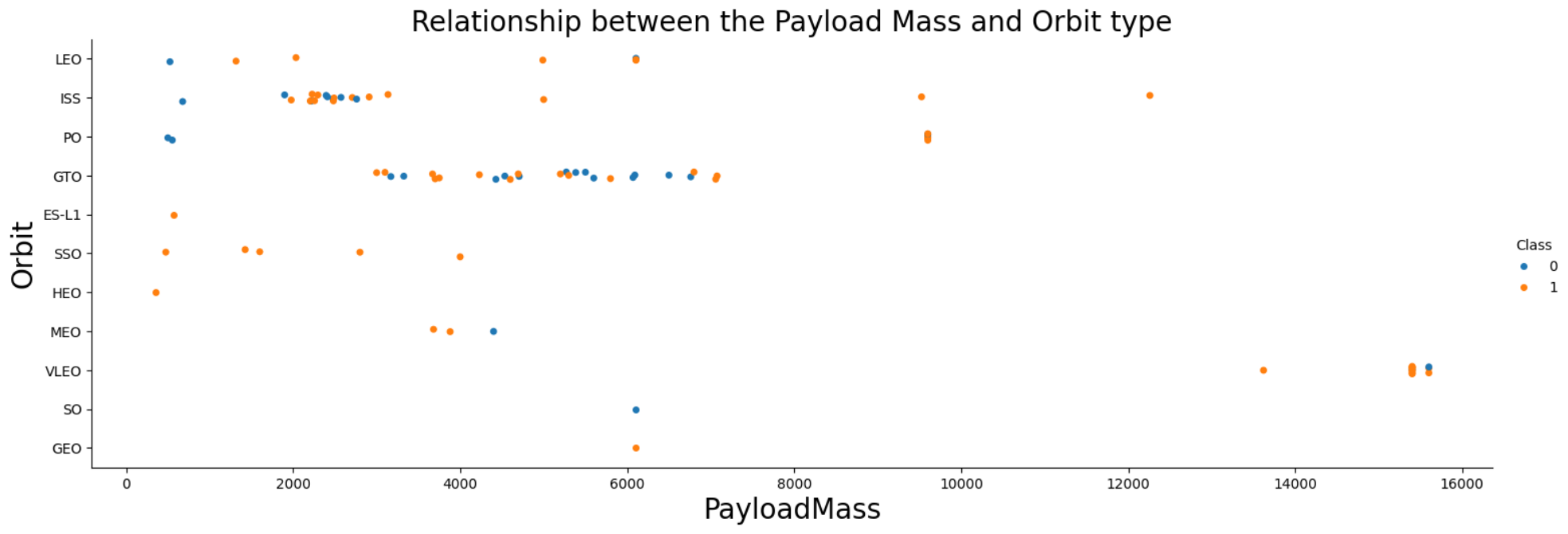
# Flight Number vs. Orbit Type




From the figure above, we can make the following observations:

- The percentage of successful landings for all orbits increased with the flight number, except for the SO orbit.
- It isn't until 60 flights that a rocket is launched into the VLEO orbit, following which a high success rate is achieved.
- It appears the landing protocols were tested heavily in the LEO, ISS, PO, and GTO orbits before launches into other orbits were explored.

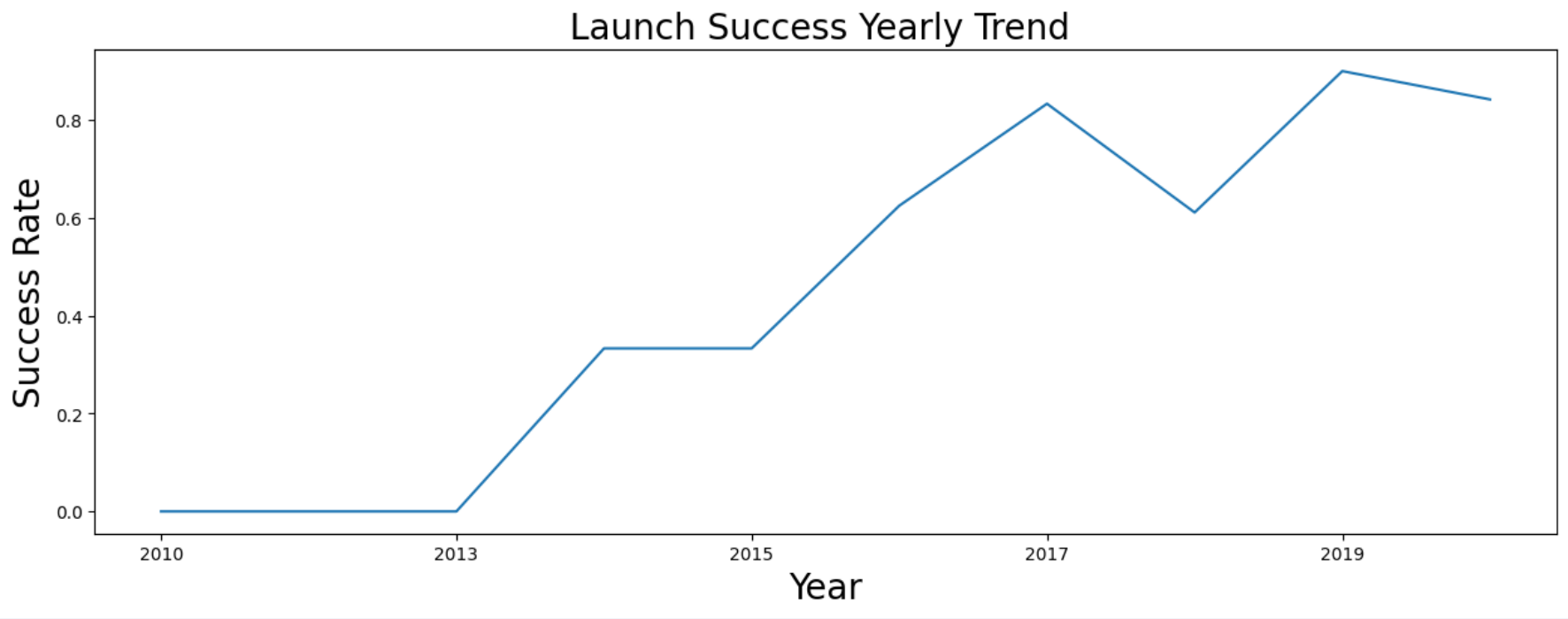
# Payload vs. Orbit Type




From the figure above, we can make the following observations:

- Generally, high success rates are achieved in orbits for rockets with heavy payloads (>10,000 kg).
- No clear trend is observed for the GTO orbit.
- For the SSO orbit, a high success wright is achieved for rockets with light payloads.

# Launch Success Yearly Trend




From the figure above, we can make the following observations:

- The success rate increased with time.
- The highest success rate was recorded in 2019.
- No successful launches were recorded from 2010 to 2013.

# All Launch Site Names

```
%sql select distinct "Launch_Site" from SPACEXTBL;
```

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

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

As shown in the image above, there are four unique launch sites in the dataset.

# Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where "Launch_Site" like "CCA%" LIMIT 5;
```

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

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_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 |

The figure below shows five rocket launches from the CCAFS LC-40 site.

# Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)";
 * sqlite:///my_data1.db
Done.
```

| sum(PAYLOAD_MASS__KG_) |
|---|
| 45596 |

As shown above, the total payload mass carried by boosters launched from NASA (CRS) is 45,596 kg.

# Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like "F9 V1.1%";
```

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

| avg(PAYLOAD_MASS__KG_) |
| --- |
| 2534.6666666666665 |

As shown above, the average payload mass carried by booster version F9 v1.1 is ~2534.7 kg.

# First Successful Ground Landing Date

```
%sql select min(Date) as "Min Date" from SPACEXTBL where Landing_Outcome = "Success (ground pad)";
```

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

| Min Date |
| --- |
| 2015-12-22 |

As shown above, the first successful landing outcome in ground pad was achieved in December 2015.

# Successful Drone Ship Landing with Payload Range

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4001 and 5999;
```

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

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

As shown above, four boosters successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

# Total Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(*) as Frequency from SPACEXTBL group by Mission_Outcome;
```

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

| Mission_Outcome | Frequency |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

As shown above, there were 101 successful missions and 1 failed mission.

# Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * 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 |

As shown above, there were 12 boosters carrying the maximum payload.

# 2015 Launch Records

```
%%sql
select substr("Date", 1, 5) as Year, substr("Date", 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL
    where Landing_Outcome = "Failure (drone ship)" and substr("Date", 1, 4) = "2015";
```

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

| Year | Month | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|---|
| 2015- | 01 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 2015- | 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

The figure above shows two (January and April) launches with failed landing_outcomes in drone ship, the booster versions, and launch site names for year 2015.

# Rank Landing Outcomes Between Fixed Period

```
%%sql
select Landing_Outcome, count(*) as Frequency from SPACEXTBL
    where Date between "2010-06-04" and "2017-03-20"
    group by Landing_Outcome
    order by count(*) DESC;
```

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

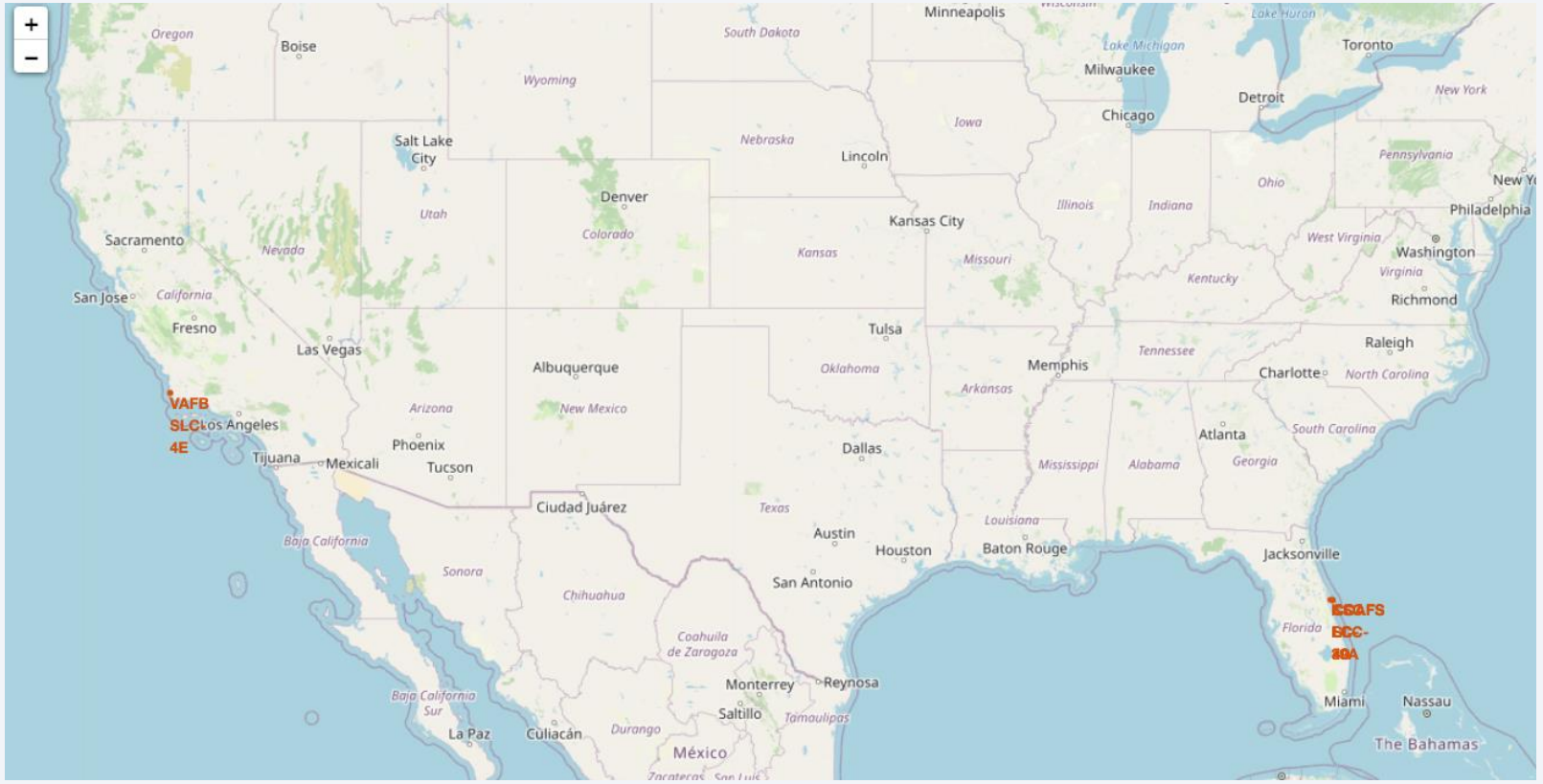
| Landing_Outcome | Frequency |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

As shown in the figure above, “No attempt” has the highest frequency. This should be considered.

Section 3

# Launch Sites Proximities Analysis

# All Launch Sites’ Location Markers




Two launch sites in close proximity

- There are four launch sites.
- Two of the launch sites are in very close proximity.
- The sites are located close to water bodies, probably for safety.

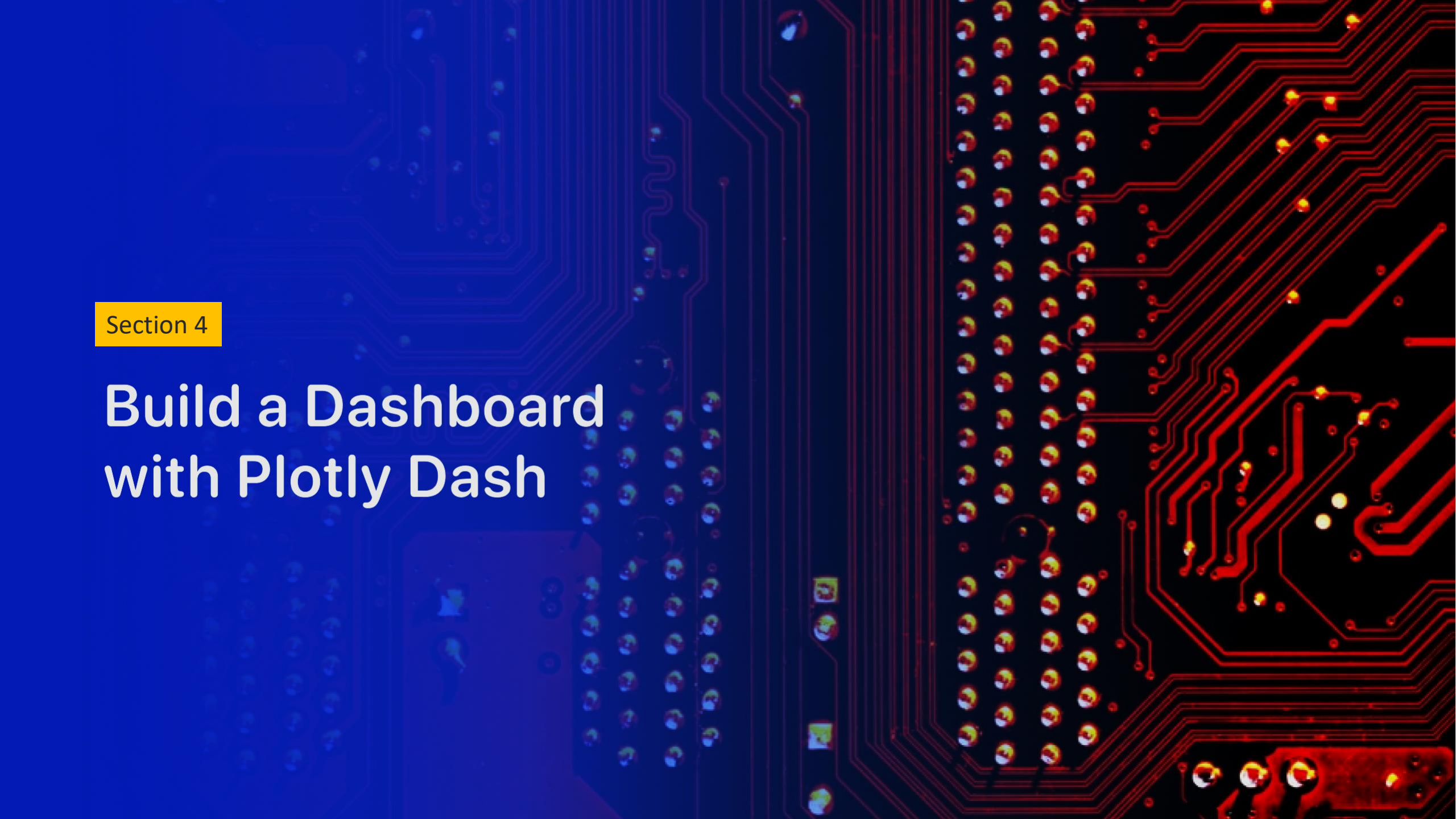
# Color-Labeled Launch Outcomes by Site

Site VAFB SLC-4E | Site KSC LC-39A | Site CCAFS LC-40 | Site CCAFS SLC-40

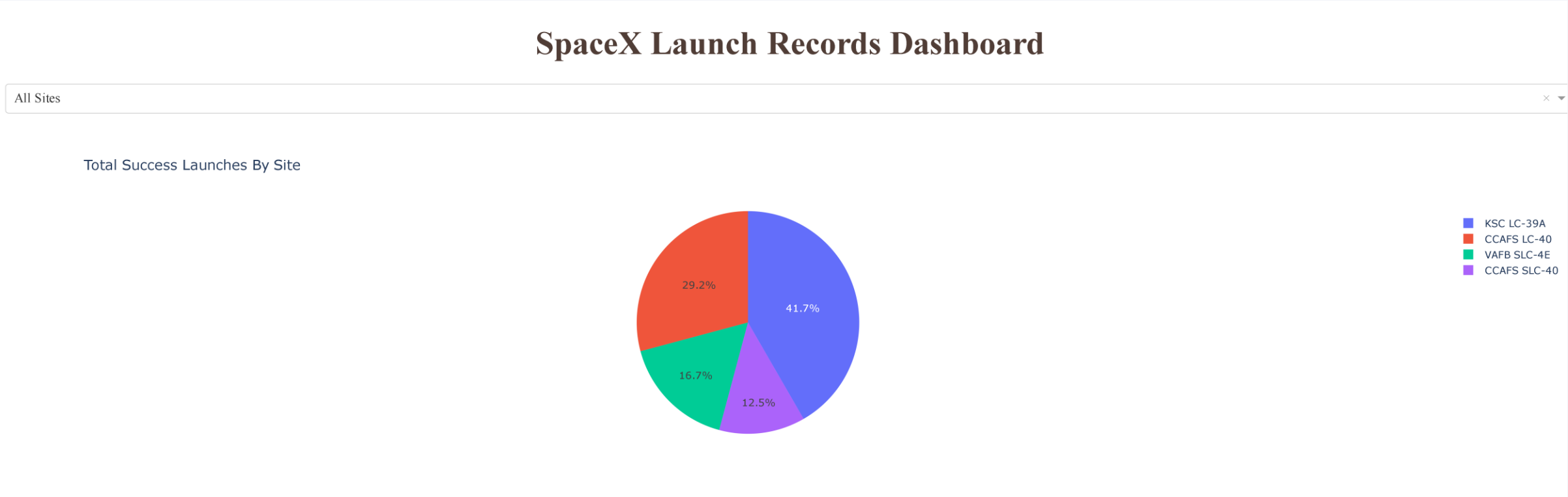

- Site CCAFS LC-40 recorded the highest number of launches and highest number of failed launches.
- Site KSC LC-39A recorded the highest number of successful launches.
- Site CCAFS SLC-40 recorded the lowest number of launches.

# Launch Site Proximities


- Site CCAFS SLC-40 is ~0.87 km from the coastline to the East.


Section 4
Build a Dashboard with Plotly Dash

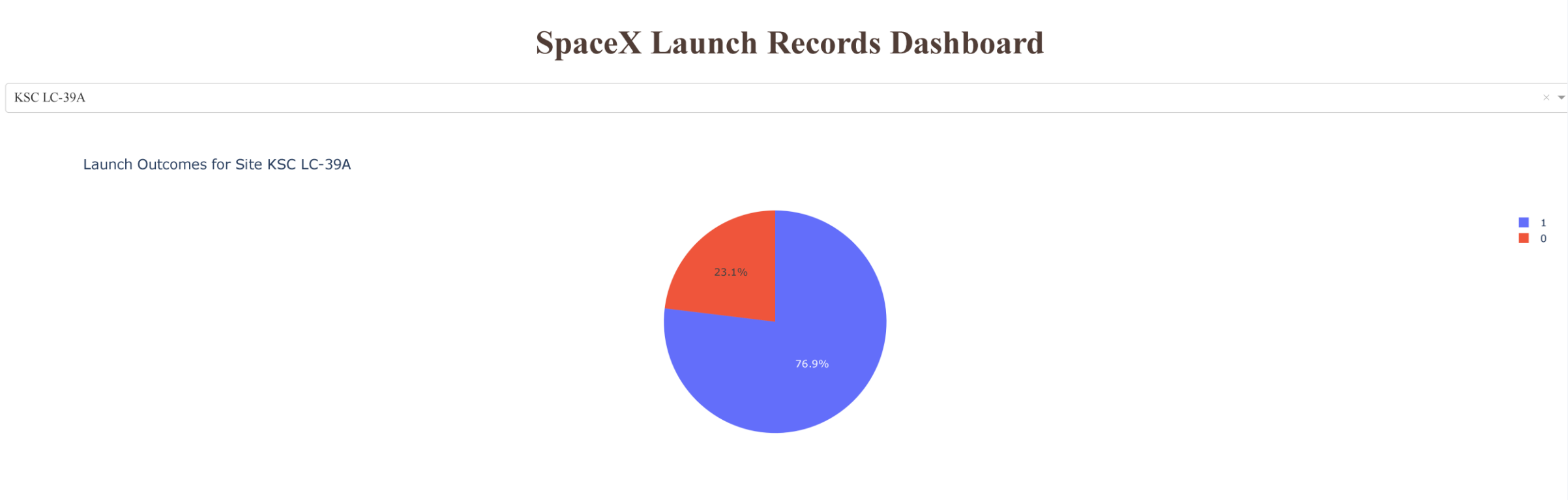
# Success Rate by Site




As shown in the pie chart above,

- Site KSC LC-39A has the highest relative success rate of 41.7%.
- Site CCAFS SLC-40 has the lowest relative success rate of 12.5%.

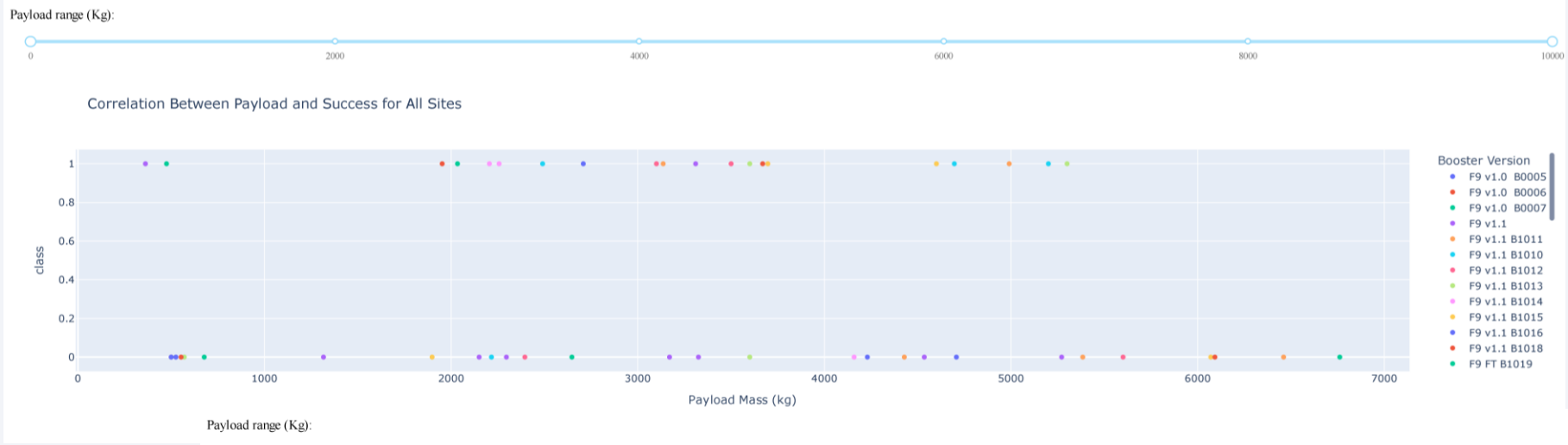
# Launch Outcomes for site KSC LC-39A




As shown in the pie chart above,

- 76.9% of the launches on site KSC LC-39A were successful.
- 23.1% of the launches on site KSC LC-39A were unsuccessful.

# Payload vs. Launch Outcome for All Sites



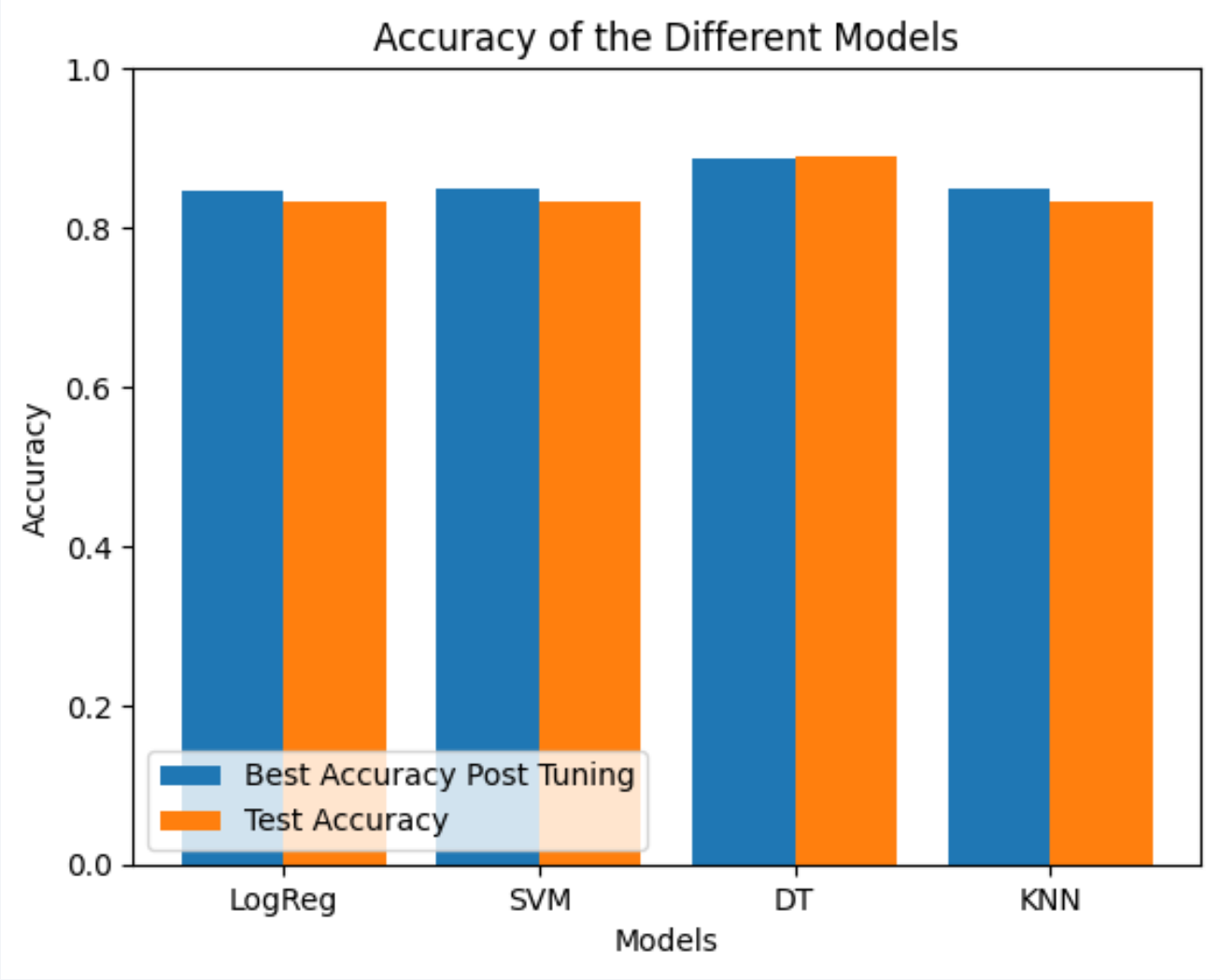


The success rate decreases significantly for payload masses above 5500 kg.

Section 5

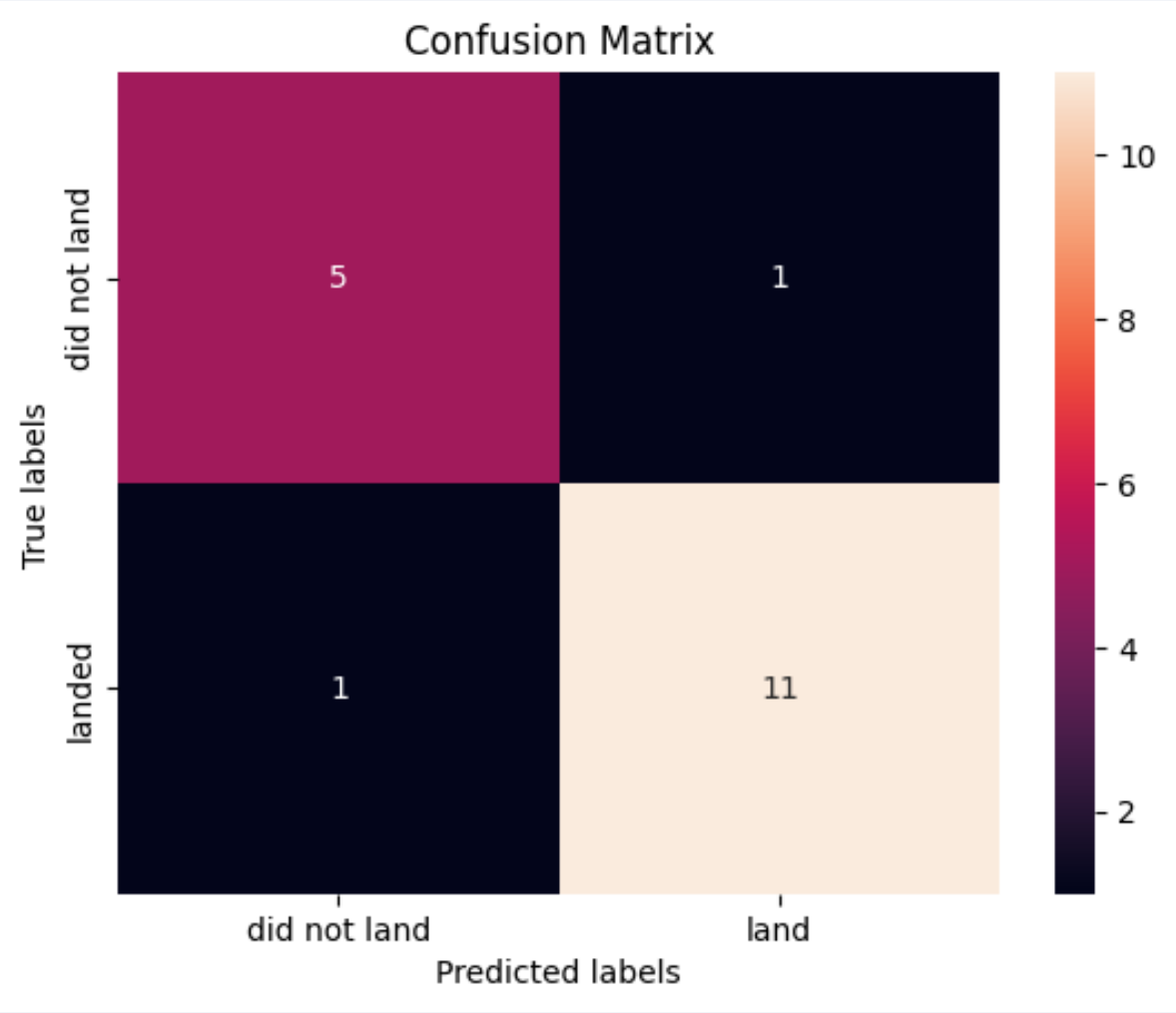
# Predictive Analysis (Classification)

# Classification Accuracy




As shown, the Decision Tree model delivers the best classification accuracy among all models.

# Decision Tree - Confusion Matrix




- The DT model correctly classifies 5 true negatives and 11 true positives.
- Conversely, it gives 1 false positive and 1 false negative.
- Compared with the other models, the DT classifier exhibits the best performance.

# Conclusions

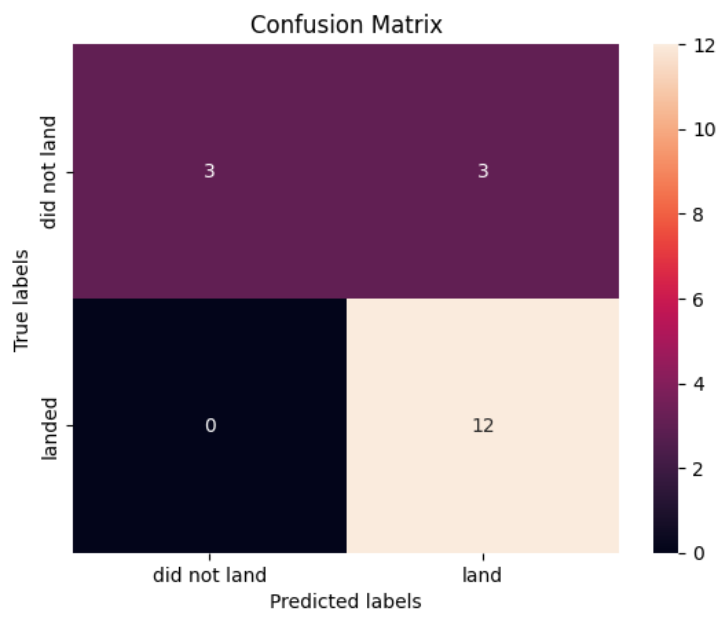
- In this project, I applied machine learning models to predict the landing outcomes of the first stage for rocket launches.
- Four launch sites were analyzed (), most located conveniently close to water bodies. Furthermore, the best landing outcomes were achieved at site KSC LC-39A.
- The launch success rate increased considerably with time.
- There appears to be a correlation between payload mass and launch success rate.
- The Decision Tree Classifier model could correctly predict the launch outcomes with an accuracy of 87%.
- The findings of this project will help determine launch costs and provide strategic insights for *SpaceY*'s market entry and operational planning.

# Appendix

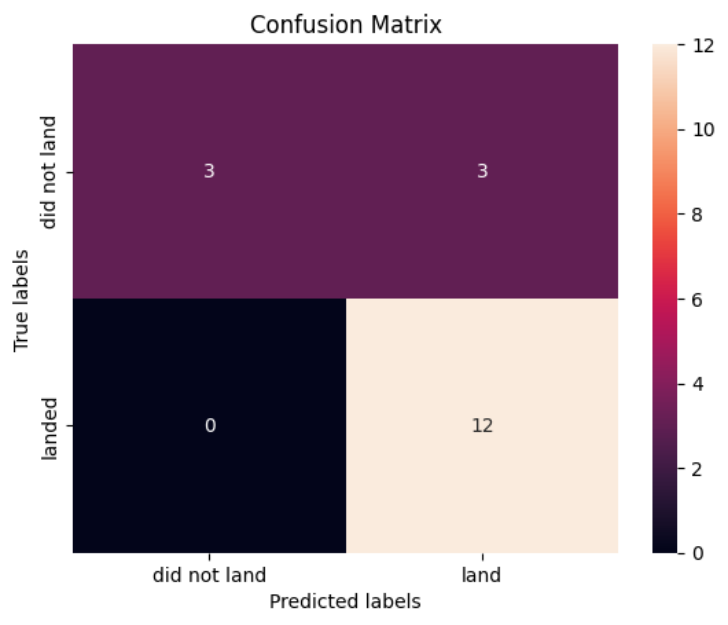
**Project GitHub Repo:** [Click here]

**Confusion matrixes** for the other classification models:

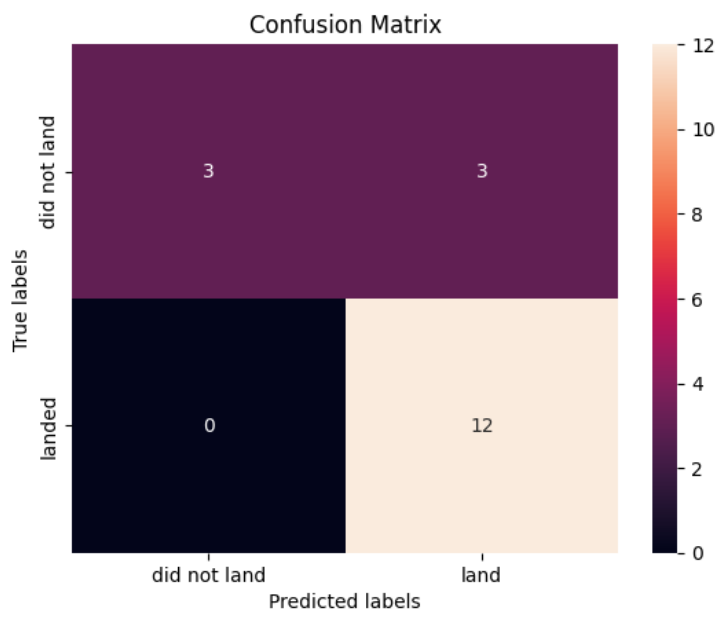
SVM




KNN




LogRes




As you can observe above, all three models exhibited the same classification accuracy.

# Thank you!