IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

DAVID OGU
02-SEP-2024

# Outline

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

# Executive Summary

This report examines factors influencing the success of SpaceX rocket landings, crucial for reducing launch costs and enhancing bid competitiveness. Data were sourced from the SpaceX website and Wikipedia through web scraping, with exploratory data analysis identifying key factors such as launch site, orbit type, payload mass, and coastal proximity.

Predictive models, including Decision Tree Classifier, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression, were employed to determine landing success probabilities, with the Decision Tree Classifier showing the highest accuracy.

Findings reveal that inland launch sites and certain orbits (ES-L1, GEO, HEO, SSO) have higher success rates, while proximity to coastal lines reduces success likelihood. The overall landing success rate is 66%, with recent improvements. These insights aim to guide our company's strategic positioning against SpaceX by focusing on key factors that influence landing success and launch costs.

# Introduction

**Project Background and Context**

SpaceX's innovative approach to rocket reusability has disrupted the space launch industry, offering significantly lower launch costs ($62 million) compared to other providers ($165 million). The ability to land and reuse the first stage of rockets is a key factor in their cost advantage. This project analyzes SpaceX launch data to identify factors influencing the success of first-stage landings, providing insights that can help our company enhance its bid strategies.

**Problems Addressed**

The analysis seeks to answer the following questions:

- **Launch Site Success:** Which launch sites have the highest success rates?
- **Orbit Influence:** How does orbit type affect landing success?
- **Launch Site Proximity:** Does launch site proximities influence landing success?
- **Predictive Models:** Which models best predict landing success—Decision Tree, K-Nearest Neighbors, SVM, or Logistic Regression?

These findings aim to equip our company with strategic insights for competitive positioning against SpaceX.

Section 1

# Methodology

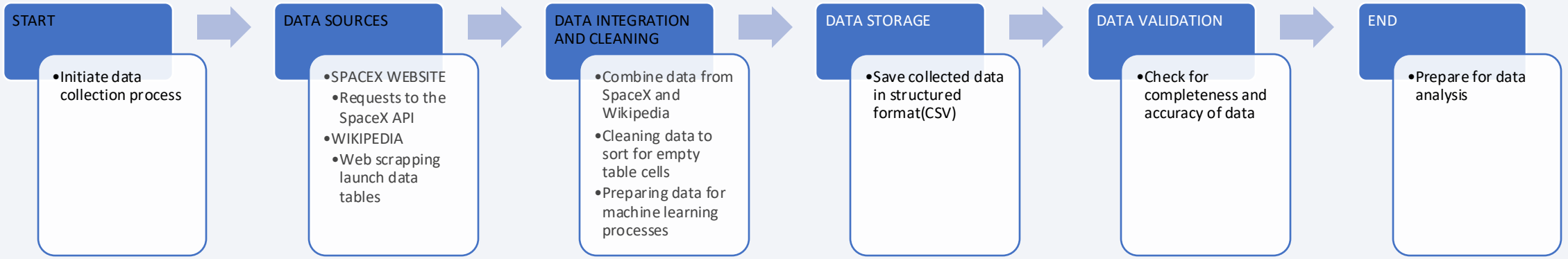
# Methodology

## Executive Summary

- Data collection methodology:
  - Data were sourced from the SpaceX website and supplemented by web scraping Wikipedia for comprehensive launch information.
- Perform data wrangling
  - The collected data underwent cleaning and preprocessing to handle missing values and inconsistencies.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

# Data Collection

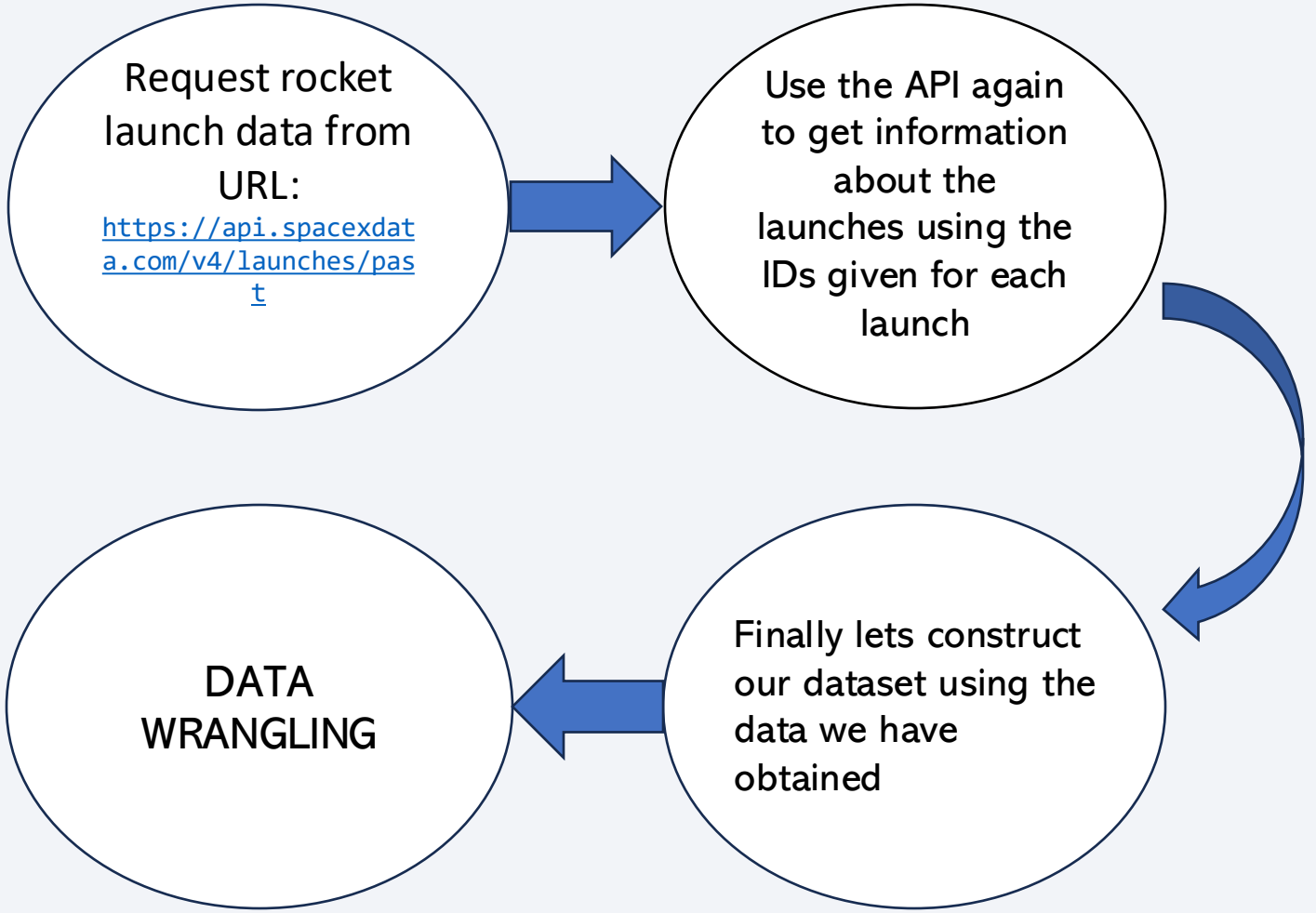
Visual representation of the data collection process:



# Data Collection – SpaceX API

- The flow diagram presented as follows represents the data collection process using SpaceX REST APIs

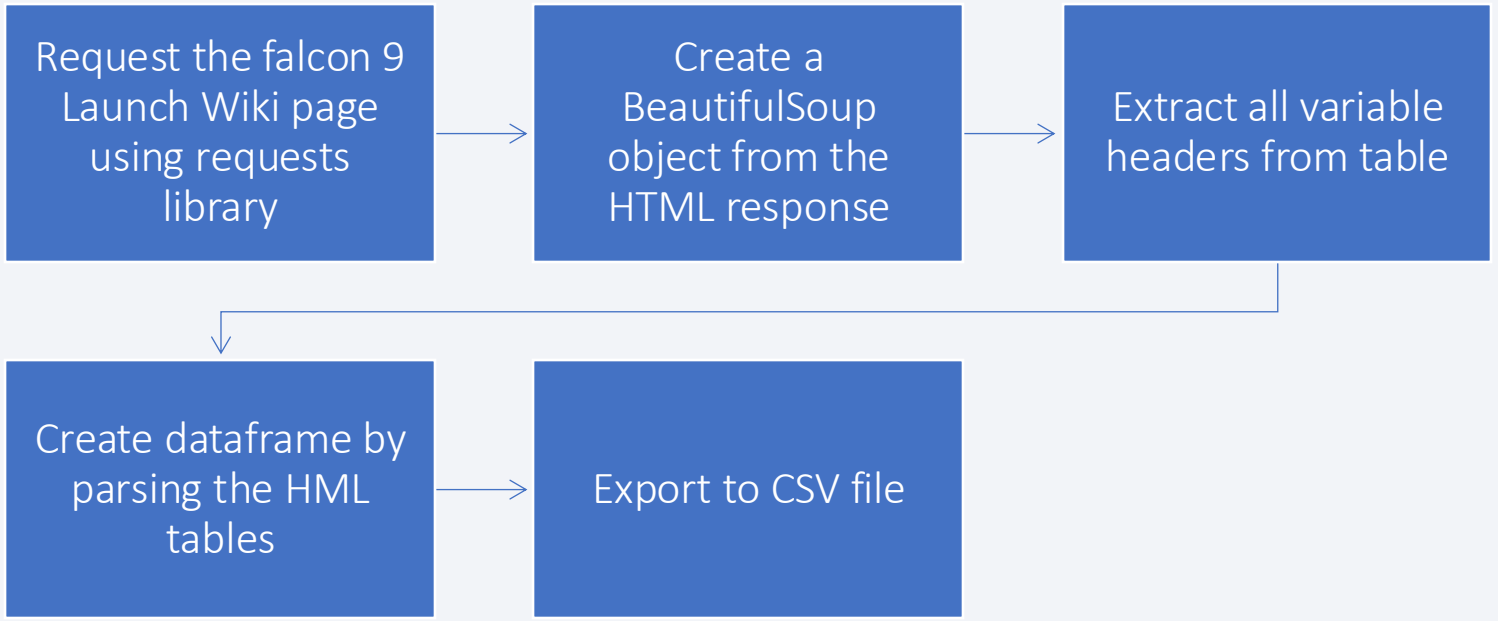
- GitHub link to data collection notebook: Github/data collection notebook



# Data Collection - Scraping

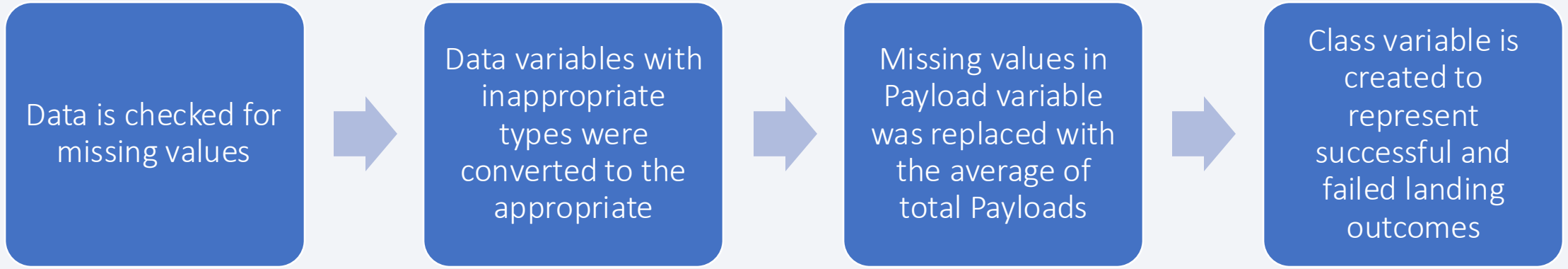
- The following flow diagram represents the web scrapping process from Wikipedia

- Github repository link



# Data Wrangling

At this stage the data was cleaned, and appropriate measures were taken on variables with missing values




[GitHub repository link to data wrangling notebook]

# EDA with Data Visualization

- Flight number was plotted against Payload: We saw that with increase in flight number, first stage is more likely to land successfully, also the more massive the payload, the less likely first stage will return
- Flight number was plotted against Launch site: We observed with increasing flight number, the first stage was more likely to return and also, the CCAFS SLC 40 launch site has more flight attempts than other launch sites
- Payload VS Launch site: We observed for the VAFB-SLC launch site there was no rocket launch for heavy payload, also the CCAFS SLC 40 has higher success rate than other launch sites.
- Success rate VS Orbit type: We observed the ES-L1, GEO, HEO and SSO orbits have higher success rates than the GTO, ISS, LEO, MEO, PO and VLEO orbits

# EDA with Data Visualization (cont.)

- Flight number VS Orbit type: We observed, more earlier flights were done on the LEO, ISS, PO, and GTO orbits, also observed that with increasing flight number the stage one was more likely to return
- Payload VS Orbit type: We observed that with increasing payload the LEO, ISS and PO orbits, the first stage was more likely to return.
- Launch success vs Yearly trend: We observed that the success rate since 2013 kept increasing till 2017 (stable in 2014 and after 2015 started incresing), having a dip in 2017 and increased further since 2018

Github repository link to data visualization notebook

# EDA with SQL

From the dataset we observed:

- There are 4 Launch sites used in the space mission, namely CCAFS LC 40, VAFB SLC 4E, KSC LC 39A, CCAFS SLC 40
- Total payload mass carried by boosters launched by NASA is 99,980KG
- Average payload mass carried by booster version F9 v1.1 is 1,986.1kg
- First successful landing on a ground pad was in 2015-DEC-22
- Booster versions which have success in drone ship with payload between 4000 and 6000 include F9 FT(B1022, B1026, B1021.2, B1031.2)
- Drone ship has the highest count of success and failed landings (5 each), ground pad has 3 successful landings and there was also 3 controlled ocean landings

Github repository link to SQL EDA notebook

# Build an Interactive Map with Folium

- Location data analysis was done on the launch sites and outcome of each launches,
- We observed areas with more launch performed (CCAFS SLC 40) and areas with higher success rate (VAFB SLC 4E)
- Markers were added to display the distance of launch site to the closest cities, railway, highway and other amenities

Github repository link to location analysis notebook

# Build a Dashboard with Plotly Dash

A pie chart and scatter plot was added to my dashboard and I observed:

- From the overall success chart that the KSC LC 39A launch site contributed more success overall, CCAFS SLC 40 having the least count of successful landings
- CCAFS SLC 40 has more failed landings (57.1%) than successful landings
- KSC LC 39A has a (76.9%) success rate from it's total number of missions.

Github repository link to dashboard source code

# Predictive Analysis (Classification)

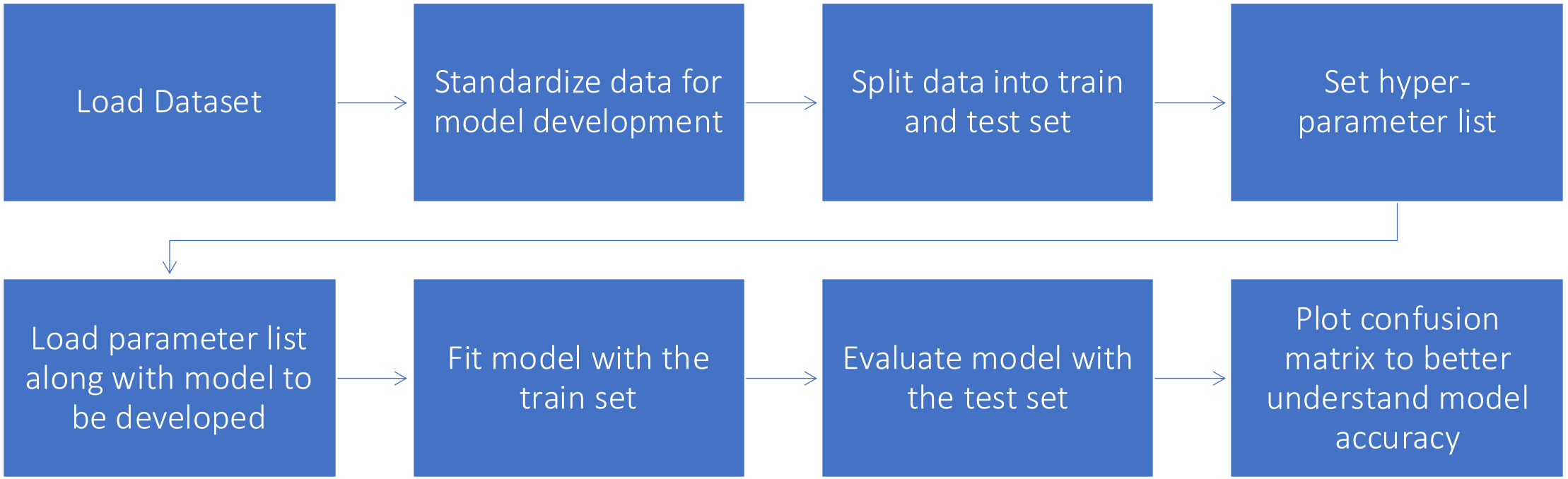
MODEL DEVELOPMENT: The following classification models were implemented to find the best classifier. The models include: Logistic regression, K-Nearest neighbors, SVM and Decision tree classifier.

The dataset was split into a train and test set and for all models, the GridSearchCV algorithm was implemented set to 10 folds to find the best hyper-parameters and solving methods to be used

For all models, the confusion matrix was plotted to observe accuracy per class and to identify points for model improvements.

See following page for visual representation of the model development phase.

# Predictive Analysis (Classification) Flow Diagram




Girhub repository link to model development notebook

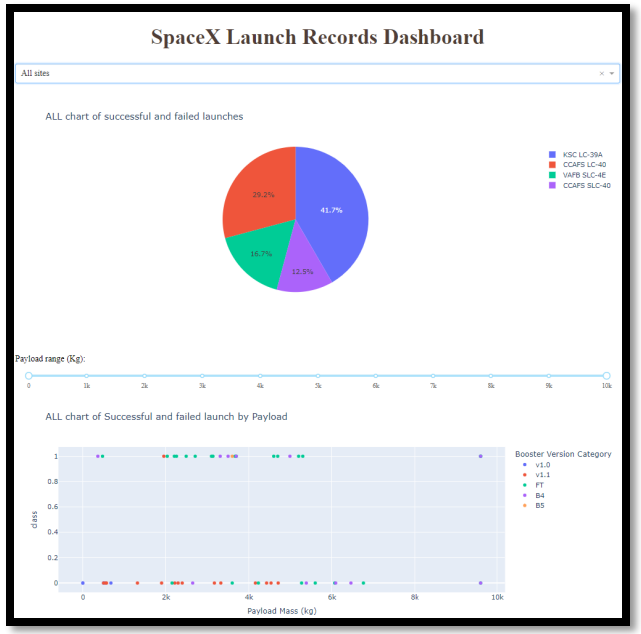
# Results

Summary of all information gotten from Exploratory Data Analysis is as follows:

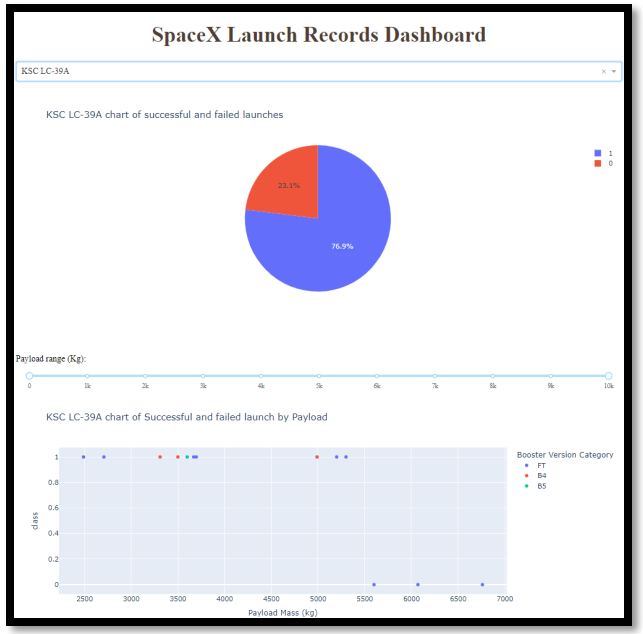
- There are more launch missions carried on the CCAFS SLC 40 launch site compared to other sites
- The VAFB SLC 4E launch site has more successful missions compared to other lanch sites
- Recent launch missions have a higher success rate than earlier launch missions
- Launch missions from the CCAFS SLC 40 site has higher success rate with heavier payloads
- We observed the ES-L1, GEO, HEO and SSO orbits have higher success rates than the GTO, ISS, LEO, MEO, PO and VLEO orbits
- Booster versions which have success in drone ship with payload between 4000 and 6000 include F9 FT(B1022, B1026, B1021.2, B1031.2)

# Results

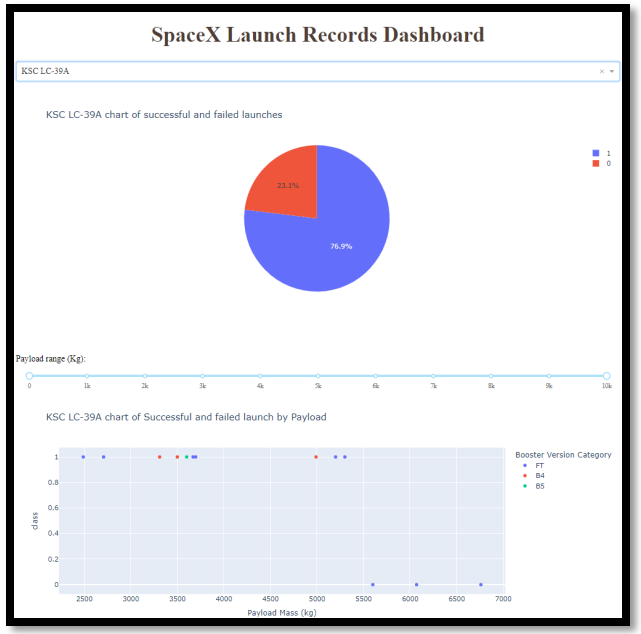
Interactive analytics demo in screenshots




ALL SITES SUCCESSFUL LANDINGS

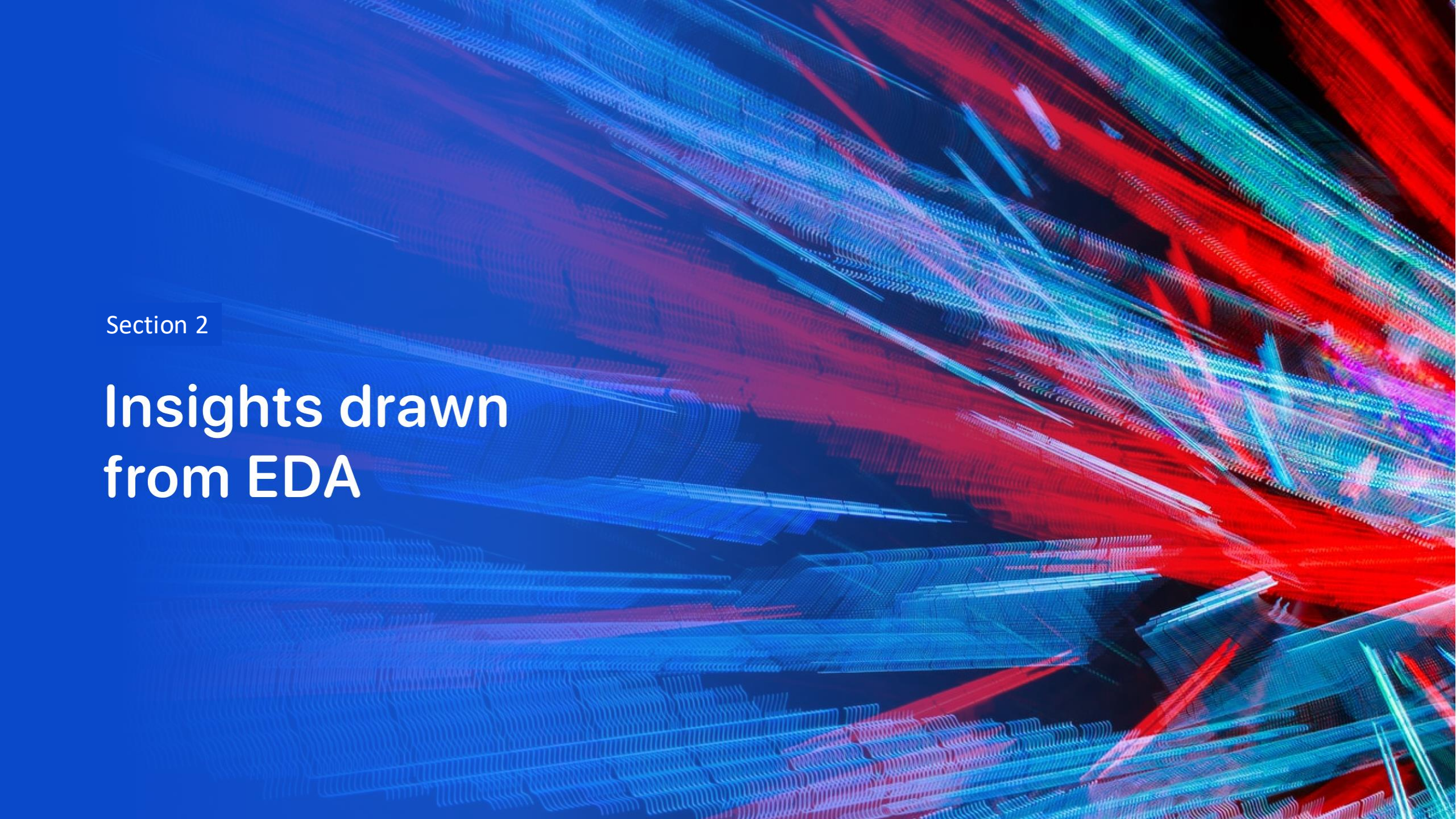



CCAFS LC 40 LAUNCH RESULTS




KSC LC 39A LAUNCH SITE RESULTS

- Predictive analysis results: The Decision tree classifier was made most accurate predictions with an accuracy score of 94%

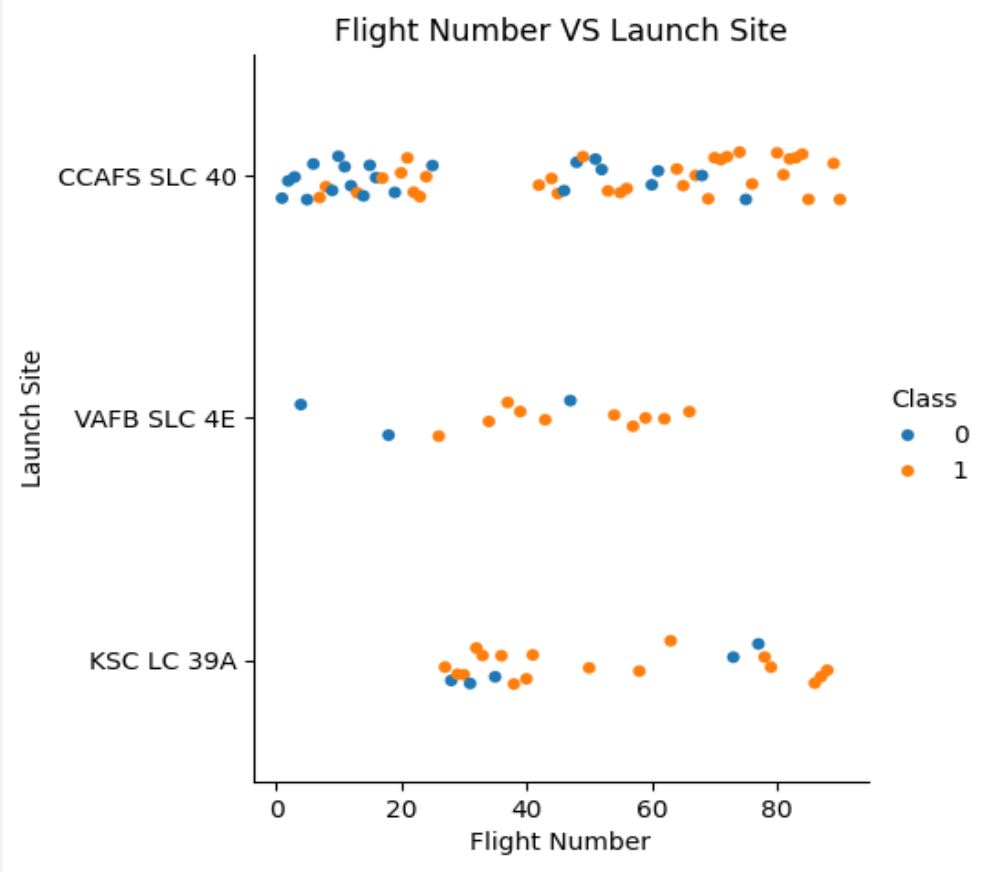

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site

## Scatter plot of Flight Number vs. Launch Site

Explanations:

- Flight number is represented on the x-axis
- Launch site represented on the y-axis
- Blue points represents failed launch missions
- Yellow points represent successful missions

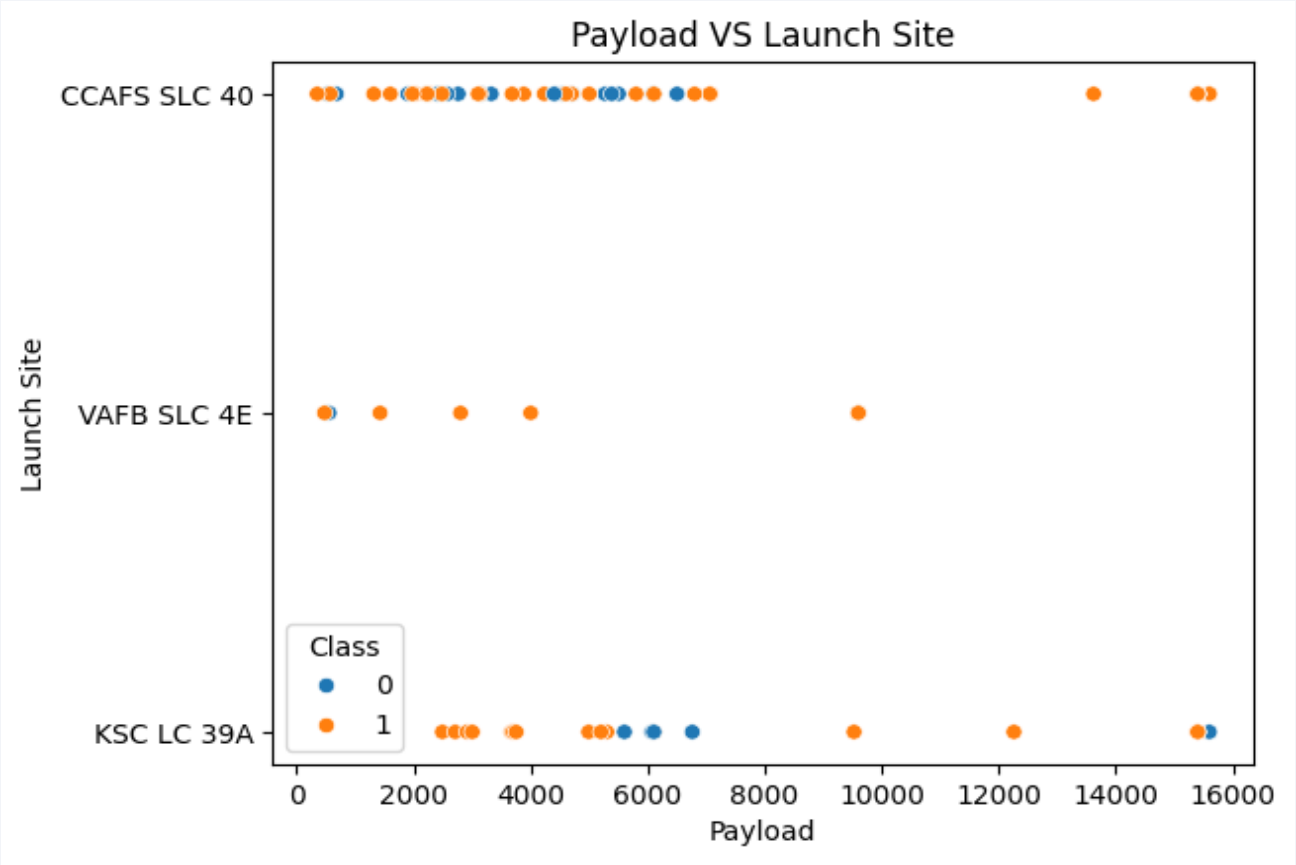


# Payload vs. Launch Site

**Scatter plot of Payload vs. Launch Site**

Explanations:

- Payload is represented on the x-axis
- Launch site represented on the y-axis
- Blue points represents failed launch missions
- Yellow points represent successful missions

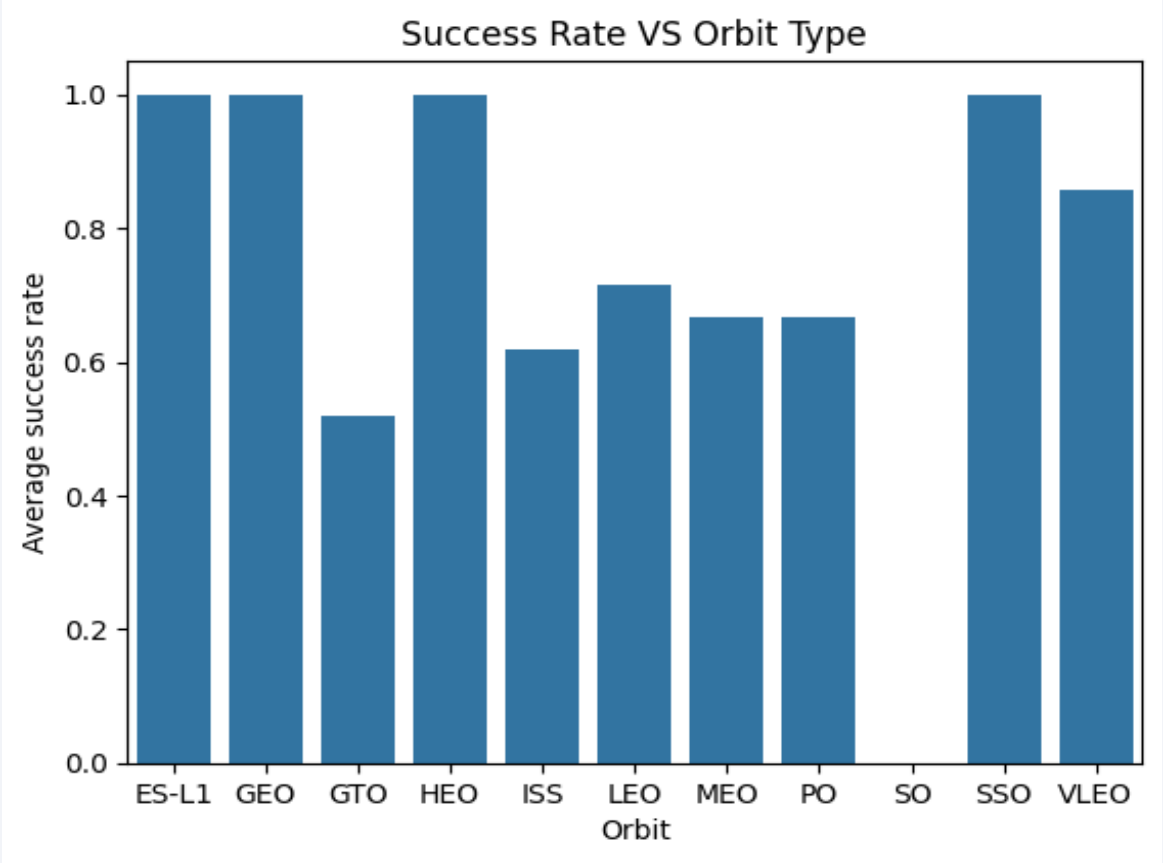


# Success Rate vs. Orbit Type

**Bar chart for the success rate of each orbit type**

**Explanations**:

- Orbit type is represented on the x-axis
- Average success rate is represented on the y-axis
- Height of bars represent amount of success

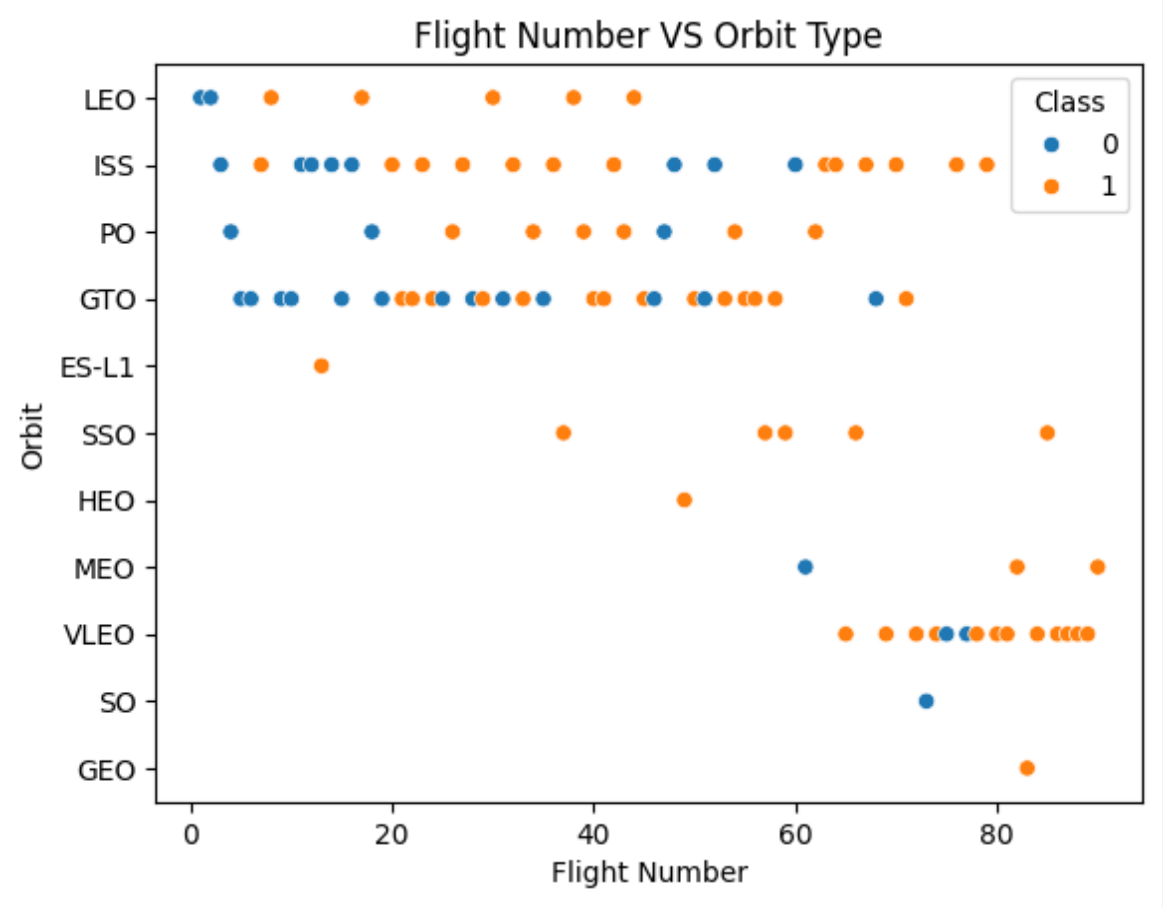


# Flight Number vs. Orbit Type

## Scatter plot of Flight number vs. Orbit type

Explanations:

- Flight number is represented on the x-axis
- Orbit type represented on the y-axis
- Blue points represents failed launch missions
- Yellow points represent successful missions

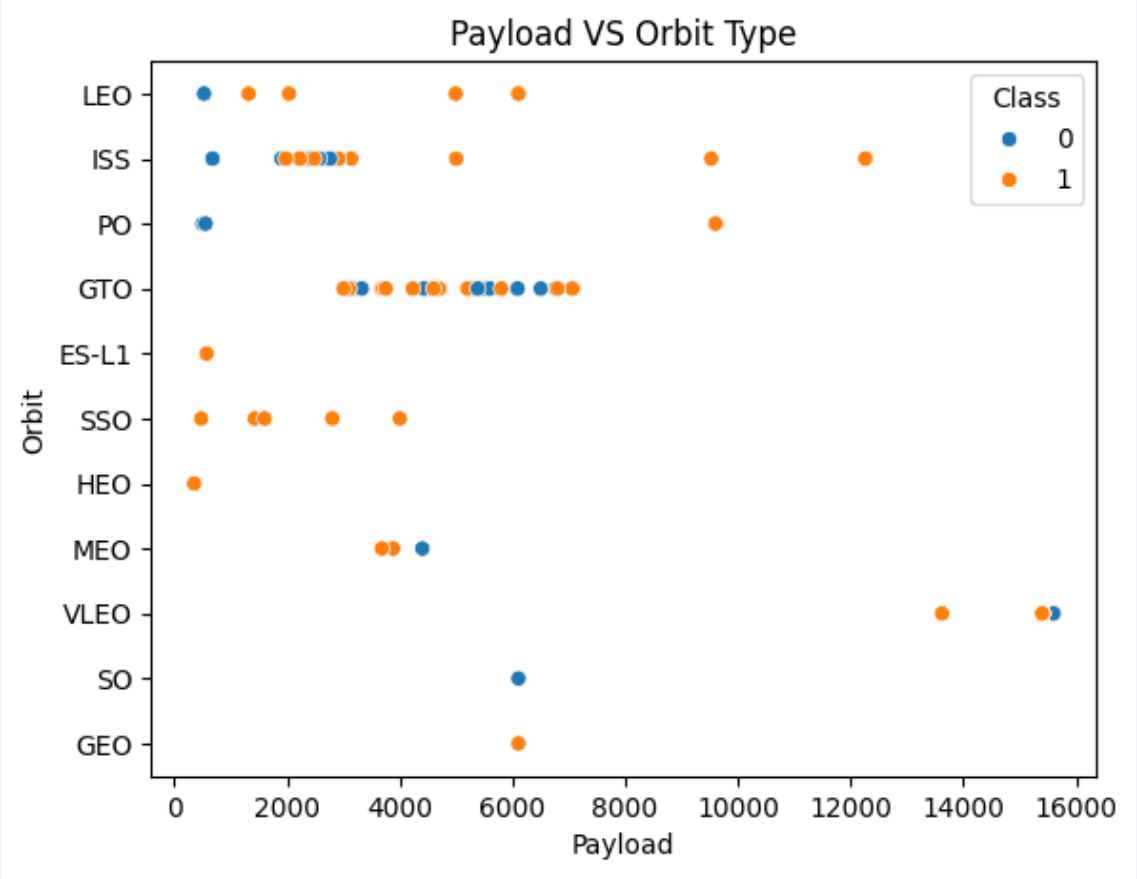


# Payload vs. Orbit Type

**Scatter plot of payload vs. orbit type**

Explanations:

- Payload is represented on the x-axis
- Orbit type represented on the y-axis
- Blue points represents failed launch missions
- Yellow points represent successful missions

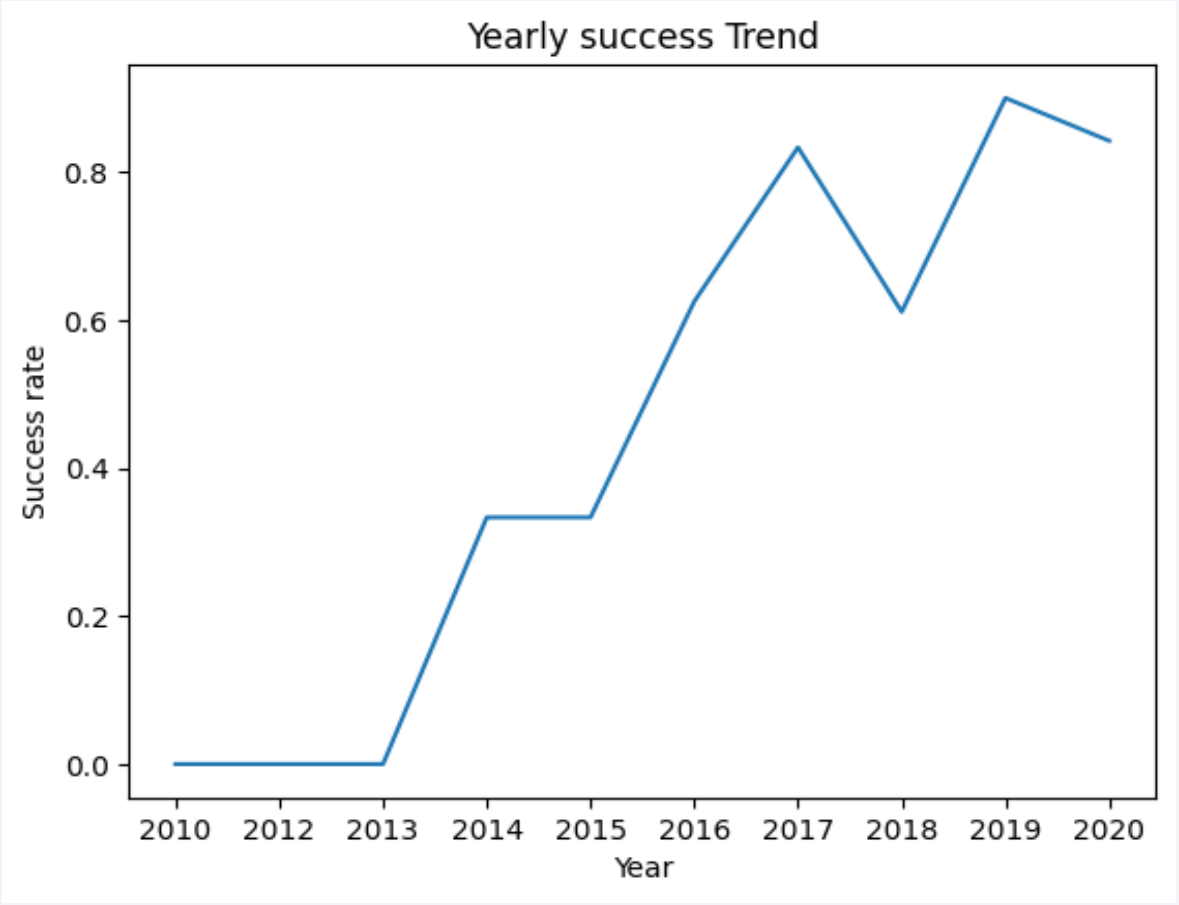


# Launch Success Yearly Trend

**Line chart of yearly average success rate**

Explanations:

- Year is represented on the x-axis
- Success rate represented on the y-axis



# All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTABLE
```

[8] Python

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

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

List of all launch site names

# Launch Site Names Begin with 'CCA'

```
[9] %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
                                                                    Python
 * sqlite:///my_data1.db
Done.
```

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orb |
|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LE |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LE (IS |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LE (IS |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LE (IS |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LE (IS |

**Launch sites that begin with CCA**

# Total Payload Mass

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer like 'NASA
```

[14] Python

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

| sum(PAYLOAD_MASS__KG_) |
|---|
| 99980 |

**Total payload by NASA (CRS)**

# Average Payload Mass by F9 v1.1

Task 4

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

+ Code + Markdown

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version lik
```

[15] Python

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

| avg(PAYLOAD_MASS__KG_) |
| --- |
| 1986.1 |

**Average Payload Mass by f9 v1.1**

# 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*

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome == 'Success (gr
```

[16] Python

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

| min(Date) |
| --- |
| 2015-12-22 |

**First Successful Landing**

# 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

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome == 'Succes
```

[21] Python

```
 * 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 with Payload between 4000 and 6000**

# Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcc
```

[26] Python

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

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

**Total Number of Successful and Failed Mission Outcomes**

# Boosters Carried Maximum Payload

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD
```

[31] Python

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

| Booster_Version | PAYLOAD_MASS__KG_ |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1060.3 | 15600 |
| F9 B5 B1049.7 | 15600 |

**Boosters Carried Maximum Payload**

# 2015 Launch Records

```
%sql select substr(Date, 6, 2) as Month, Booster_Version, Landing_Outcome, La
```

[33] Python

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

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

**2015 record of failed drone ship landings**

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

```
%sql select Landing_Outcome, count(Landing_Outcome) as num from SPACEXTABLE
```

[38] Python

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

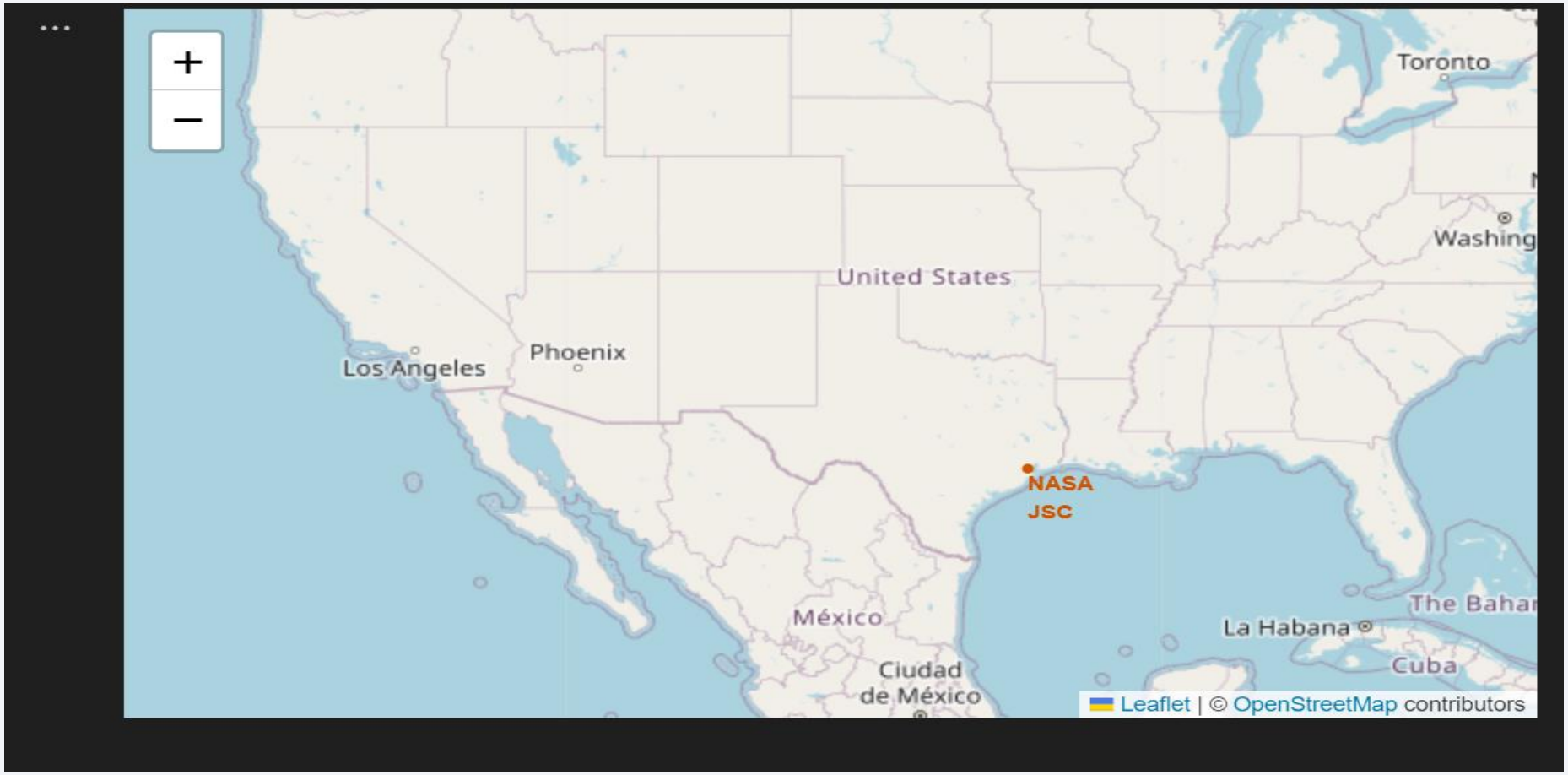
| Landing_Outcome | num |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |
| Failure (parachute) | 1 |

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

Section 3

# Launch Sites Proximities Analysis

# NASA Position on the Map



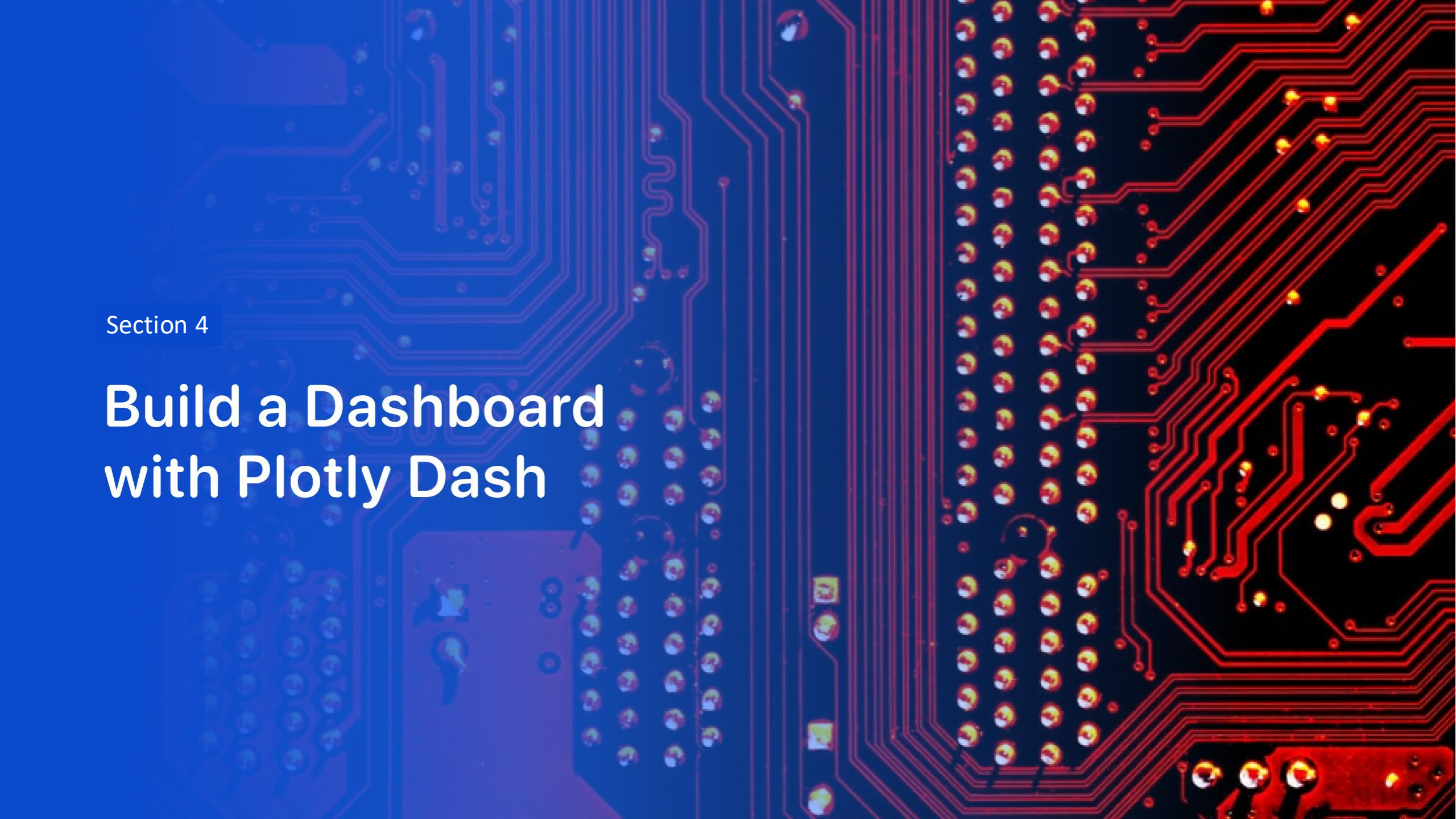

**NASA Position on the map**

# Marked Areas Of Launch Missions

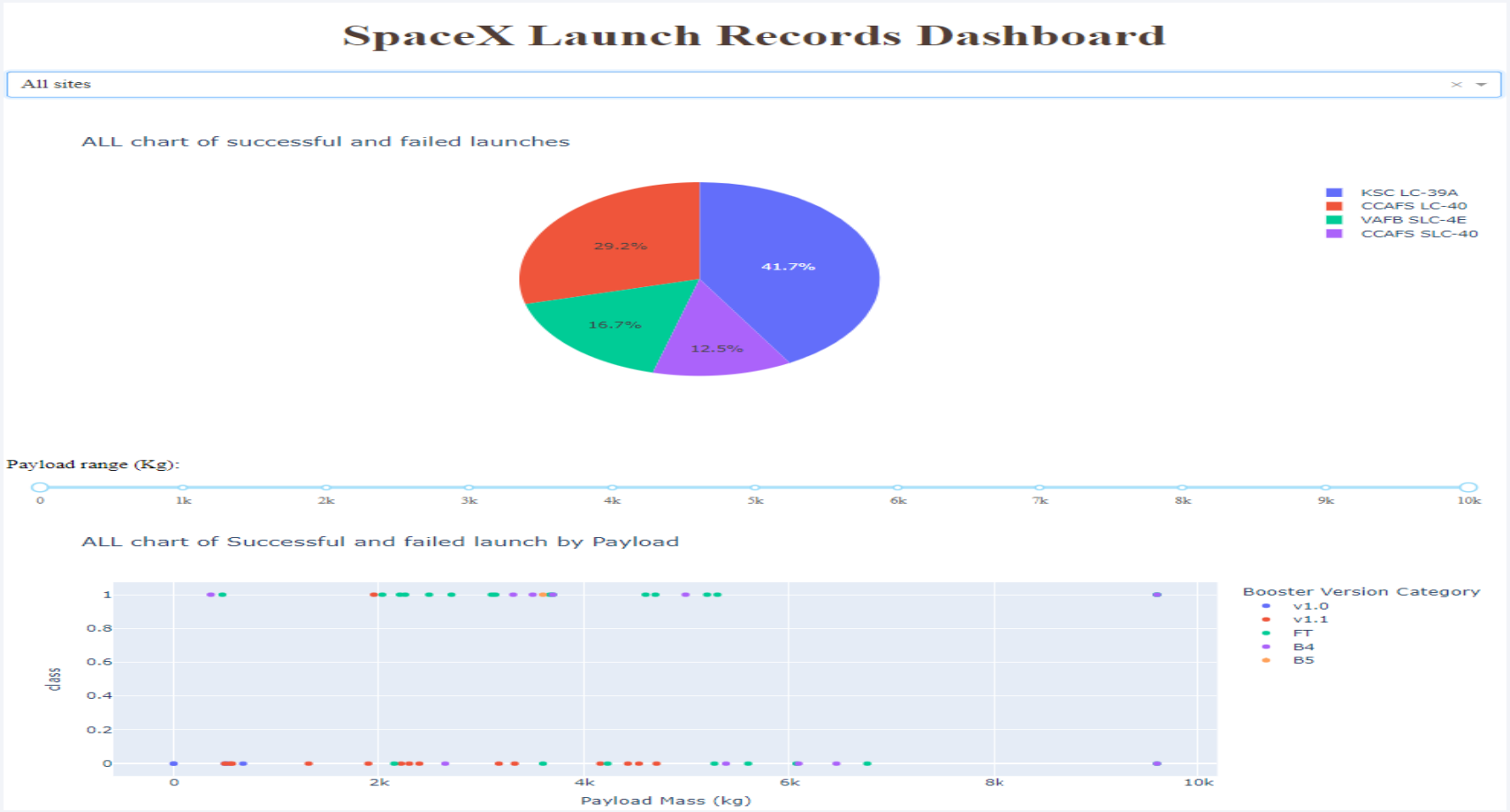

**MARKED AREAS OF LAUNCH MISSIONS**

# Distance of Launch site to Coast Line


**Distance of launch site to coast line**

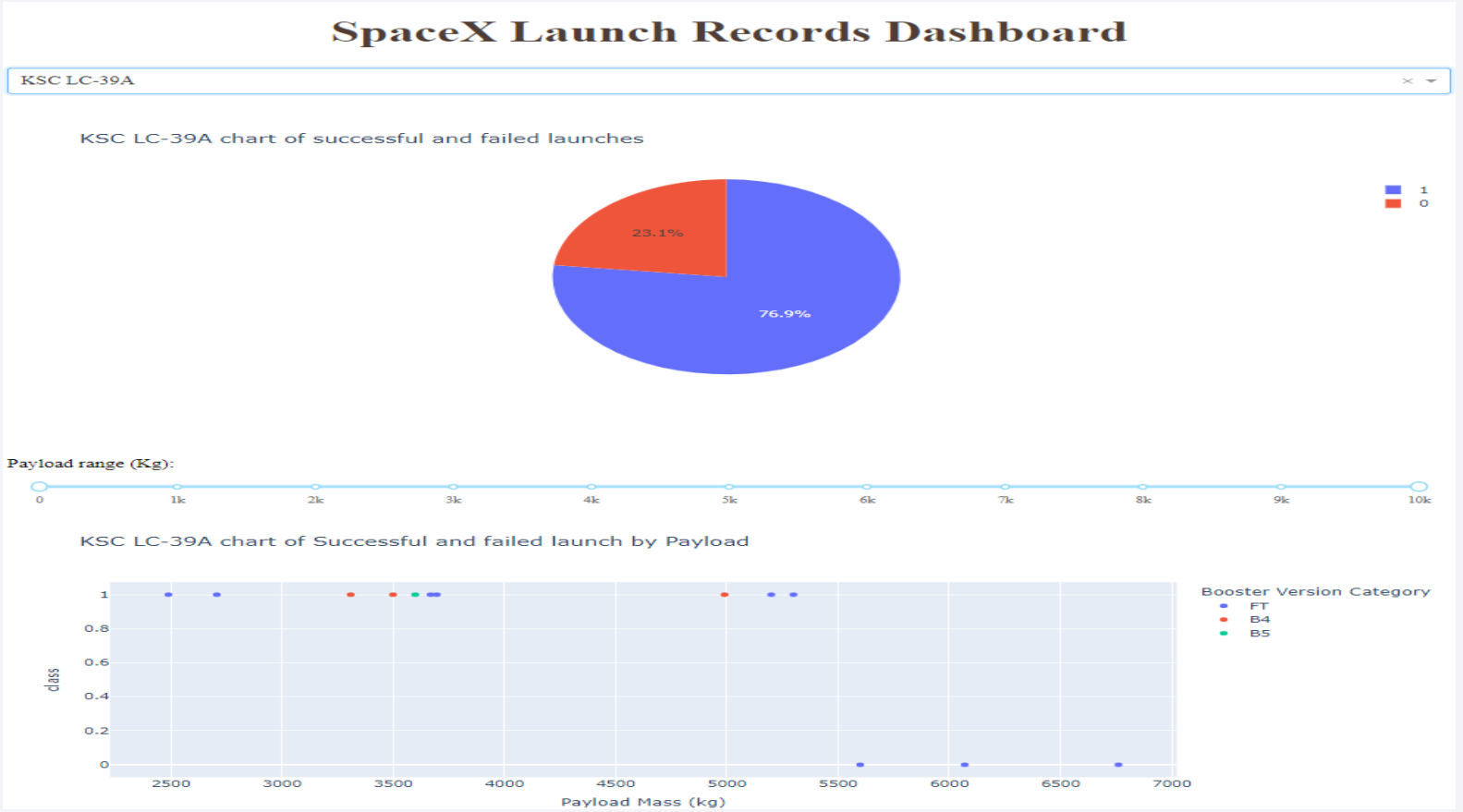

Section 4
Build a Dashboard with Plotly Dash

# Pie Chart of Launch Success For All Sites



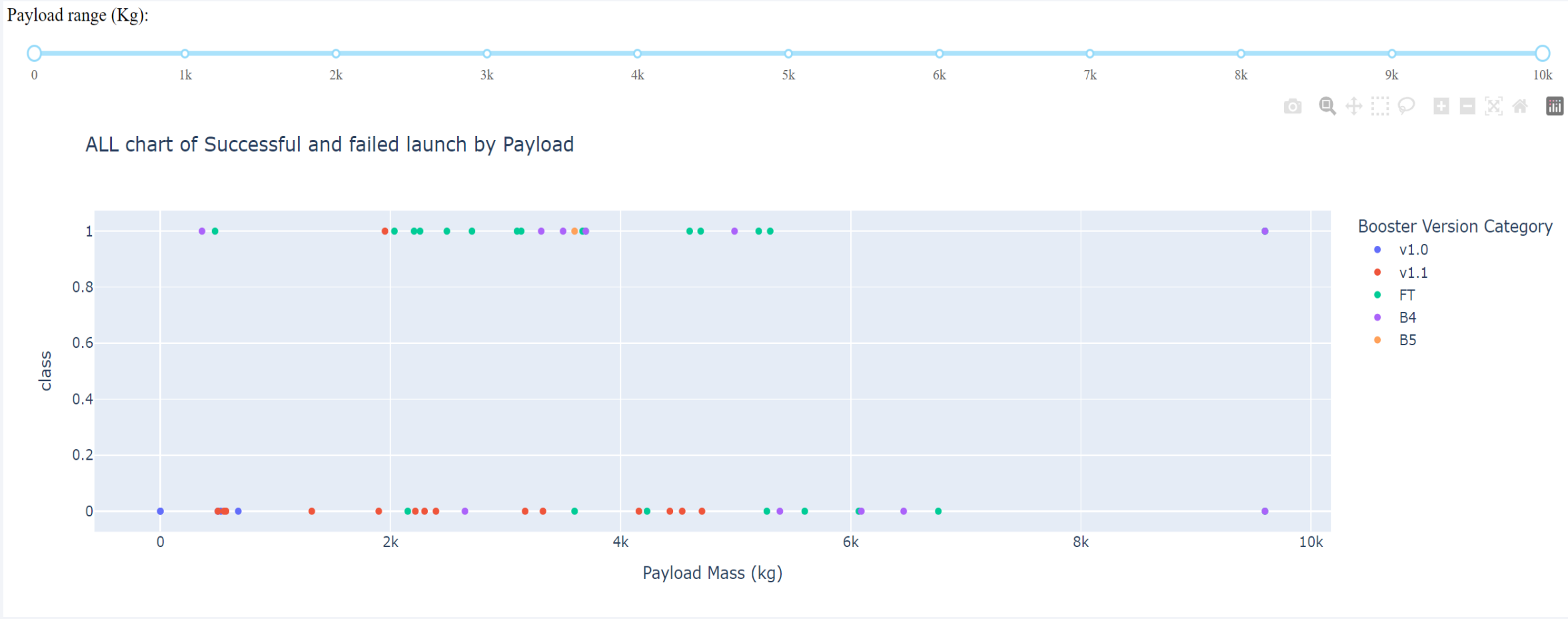

**Pie chart of launch success for all sites**

# Pie Chart For The KSC LC 4E Launch Site




**Pie chart for the KSC LC 4E launch site**

# Payload vs. Launch Outcome Scatter Plot

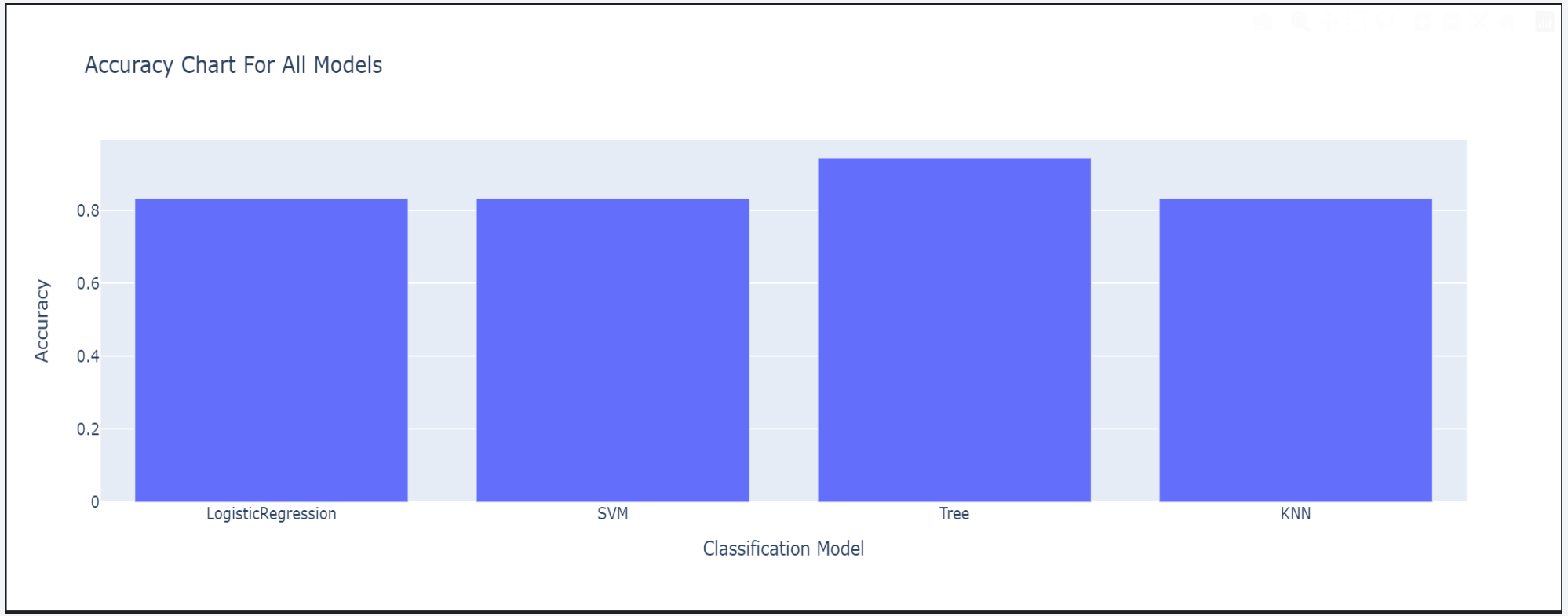



**Scatter plot of Launch Outcome vs. Payload**

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy



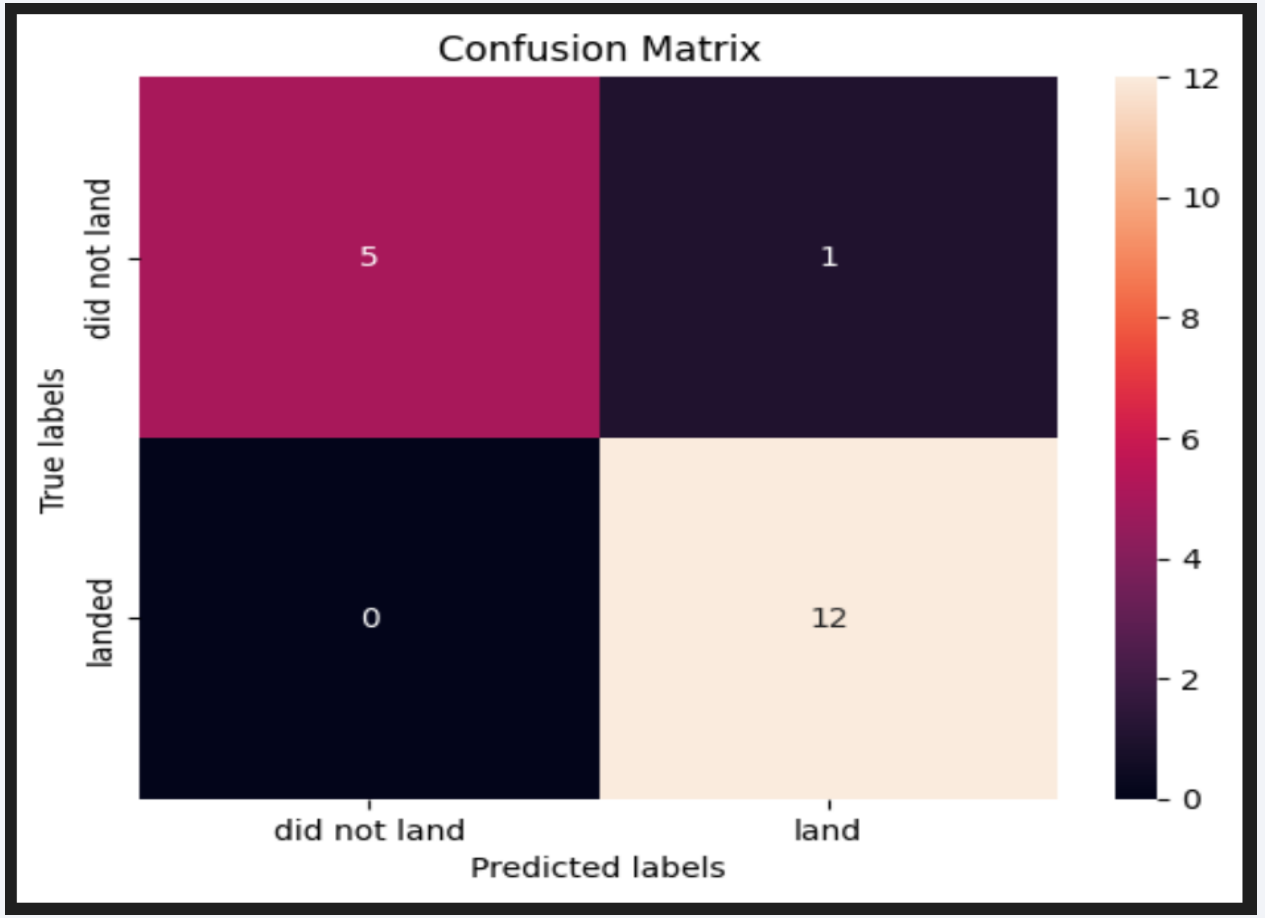

The chart shows the decision tree classifier has best out-of-sample accuracy, making it most suitable for making predictions.

# Confusion Matrix

Confusion Matrix chart for the Decision Tree Classifier:

This chart shows:

- The predictor made all correct prediction for the missions that did land
- The predictor made 5 out of 6 correct predictions for missions that did not land



# Conclusions

- **Launch Site Impact:** Inland launch sites generally have higher landing success rates compared to coastal sites, which are more frequently used but face environmental challenges.
- **Orbit Success Variation:** Rockets launched into ES-L1, GEO, HEO, and SSO orbits exhibit higher success rates for landings, highlighting the importance of orbit selection in mission planning.
- **Payload Mass and Location Significance:** Both payload mass and launch location significantly influence the likelihood of a successful rocket landing, affecting overall mission success.
- **Predictive Model Performance:** Among the predictive models tested, the Decision Tree Classifier showed the highest accuracy in determining landing success, proving valuable for strategic decision-making.

# Appendix

- Github link to data collection notebook
- Github link to web scrapping notebook
- Github link to data Wrangling notebook
- Github link to SQL exploratory data analysis
- Github link to Data exploration with visualization
- Github link to location data analysis
- Github link to dashboard source code
- Github link to model development notebook

# Thank you!