IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Nguyen Thai Huy

Tuesday 7 March 2023

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API and web scraping
  - Exploratory Data Analysis (EDA), including data wrangling, data visualisation and interactive visual analytics
  - Machine Learning Prediction
- Summary of all results
  - It was possible to collect useful data from public sources
  - EDA enabled us to determine which features best predict the success of launches
  - Using all collected data, Machine Learning Prediction revealed the best model for predicting which characteristics are important to drive this opportunity

# Introduction

- Background
  - an alternate company wants to bid against SpaceX for a rocket launch
- Objectives
  - predict if the Falcon 9 first stage will land successfully
  - determine the cost of a launch
- Findings
  - the best place to make launches
  - the best method for estimating total launch costs by predicting successful first-stage rocket landings

Section 1

# Methodology

# Methodology

## Executive Summary

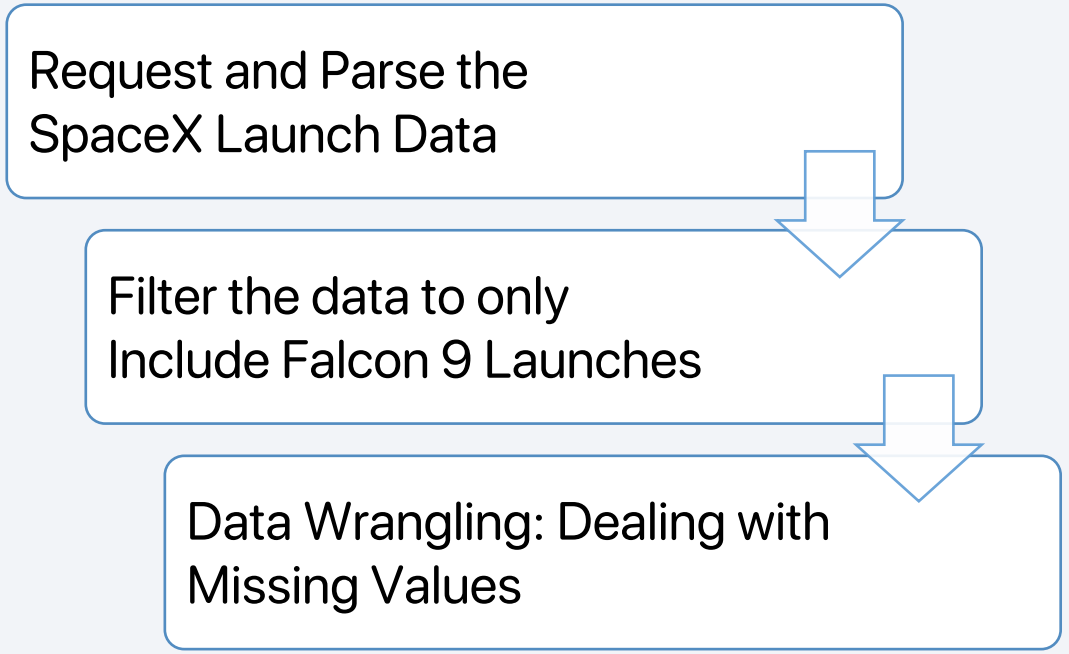
- Data collection methodology
  - SpaceX data was obtained from two sources: the Space X API and Wikipedia
- Perform data wrangling
  - Perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determined what would be the label for training supervised models
- Perform exploratory data analysis (EDA) using visualisation and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalise the data collected up to this point
  - Train and test the dataset by four different classification models
  - Evaluate the accuracy of each model using different parameter combinations

# Data Collection

- SpaceX data was collected from two sources:
  - the Space X API (https://api.spacexdata.com/v4/rockets/)
  - web scraping (https://en.wikipedia.org/wiki/List of Falcon/ 9/ and Falcon Heavy launches/)

# Data Collection – SpaceX API

1. Use the GET request and parse the SpaceX launch data
2. Filter data using the BoosterVersion column to only keep Falcon 9 launches
3. Do some basic data wrangling and formatting for missing values

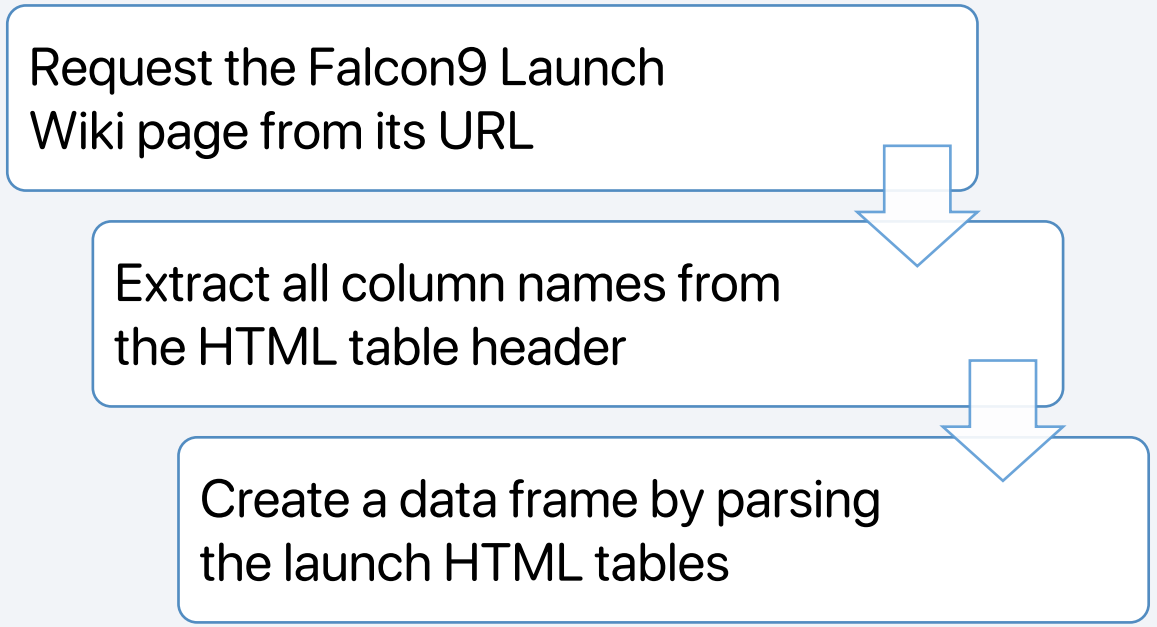



The SpaceX API Data Collection Process

HERE is the GitHub URL of the completed interactive map with Folium notebook, as an external reference.

# Data Collection - Scraping

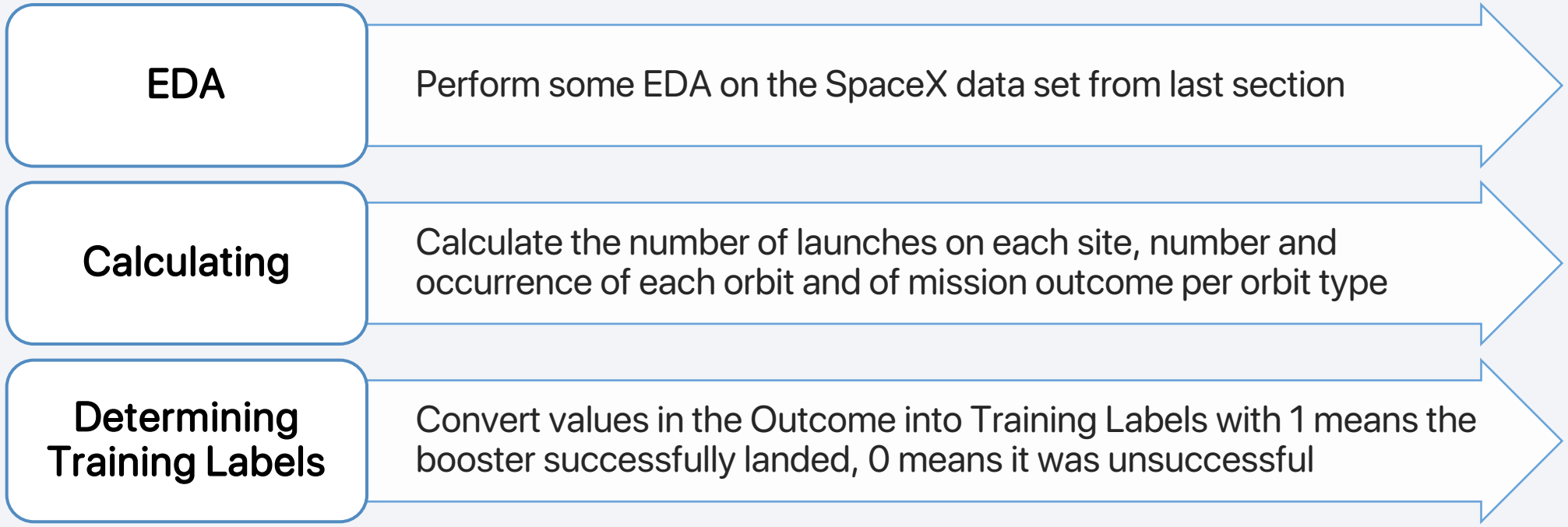
1. Perform an HTTP GET method to request the Falcon9 Launch HTML Wiki page, as an HTTP response
2. Extract all column/variable names from the HTML table header
3. Parse the launch record values into a dictionary and created a data frame from it




The Web Scraping Data Collection Process

HERE is the GitHub URL of the completed interactive map with Folium notebook, as an external reference.

# Data Wrangling




The Data Wrangling Process

HERE is the GitHub URL of the completed interactive map with Folium notebook, as an external reference.

# EDA with Data Visualisation

- Scatter plots were used to visualise the correlation between numerical variables.
- Bar plot illustrates the relationship between numerical and categorical variables.
- Line graph was used to track changes over a short or long periods of time. It can also make prediction for unseen data.

- Scatter plots:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit type
  - Payload and Orbit type
- Bar plot:
  - Success Rate of Each Orbit Type
- Line graph:
  - Launch Success Yearly Trend

HERE is the GitHub URL of the completed EDA with data visualisation notebook, as an external reference.

# EDA with SQL

The following SQL queries were performed to gather and understand data:

- List the names of the unique launch sites in space mission
- Display the 5 records where launch sites begin with ‘CCA’
- Show the total payload mass carried by boosters launched by NASA (CRS)
- Find the average payload mass carried by booster version F9 v1.1
- Find the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and its payload mass in between 4000 and 6000 kg
- Find the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- Display the records which will display the failure landing_outcomes in drone ship, booster versions, launch site for the months in year 2015
- Rank the count of successful landing_outcomes between the date 4 Jun 2016 and 20 Mar 2017 in descending order

HERE is the GitHub URL of the completed EDA with SQL notebook, as an external reference.

# Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were created and added to the folium maps to better understand the problem and the data

- The red circles at several location display theirs coordinate, and the labels reveal theirs name
- The cluster of points show multiple and distinct information for the same coordinates
- Green markers represent the successful landings, whilst red markers represent the failed landings
- The coloured lines indicate the distance between the launch site and key locations (railway, motorway, coastline, and city)

HERE is the GitHub URL of the completed interactive map with Folium notebook, as an external reference.

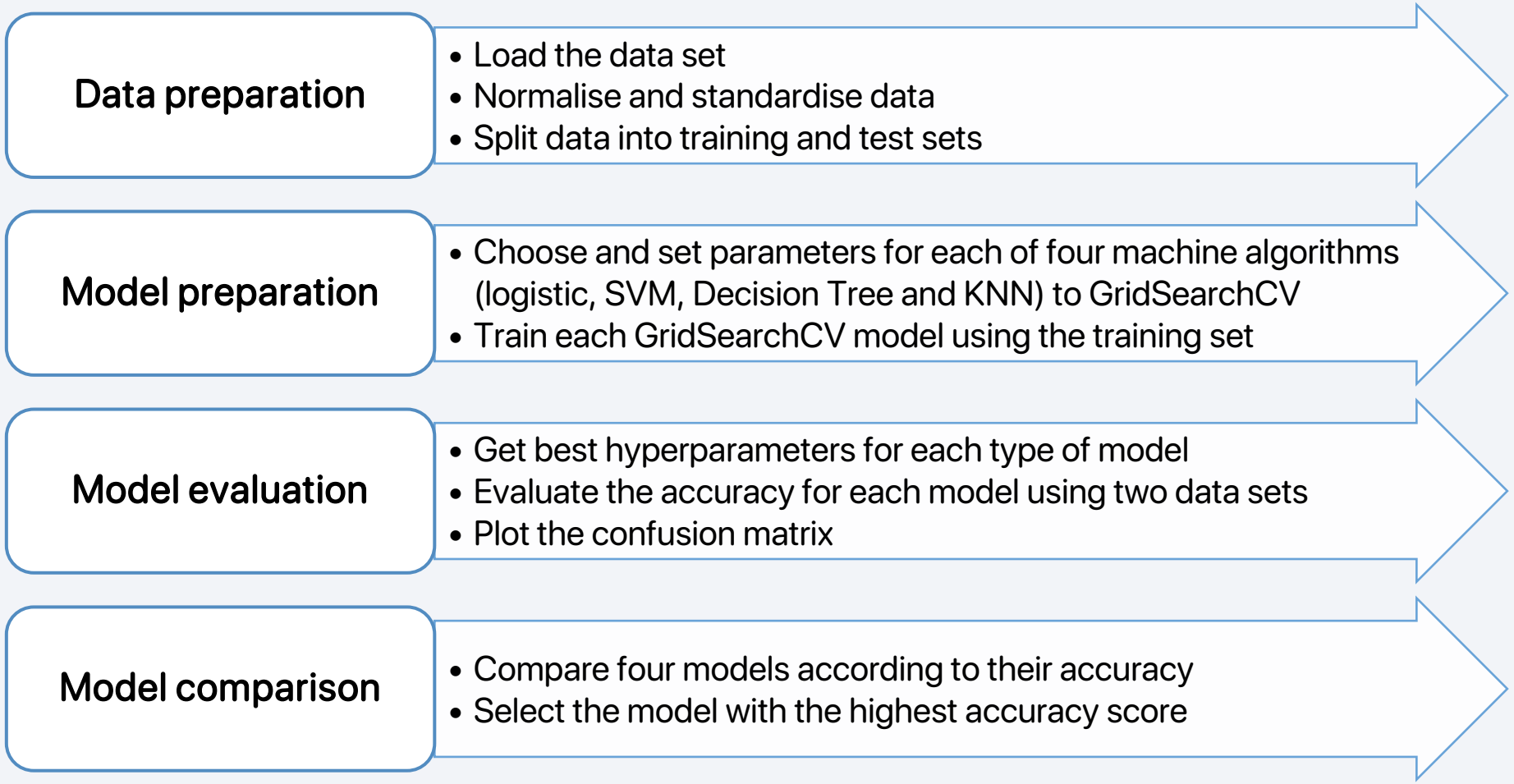
# Build a Dashboard with Plotly Dash

The dashboard includes dropdown, pie chart, range slider and scatter plot elements

- The Dropdown allows a user to select a specific launch site or all launch sites
- The pie chart displays both the total success and total failure for the launch site selected
- Range slider is used to select a payload mass in a fixed range
- Scatter chart illustrates the relationship between two variables, in particular Success vs Payload Mass

HERE is the GitHub URL and HERE are the screenshots of the completed Plotly Dash lab, as an external reference.

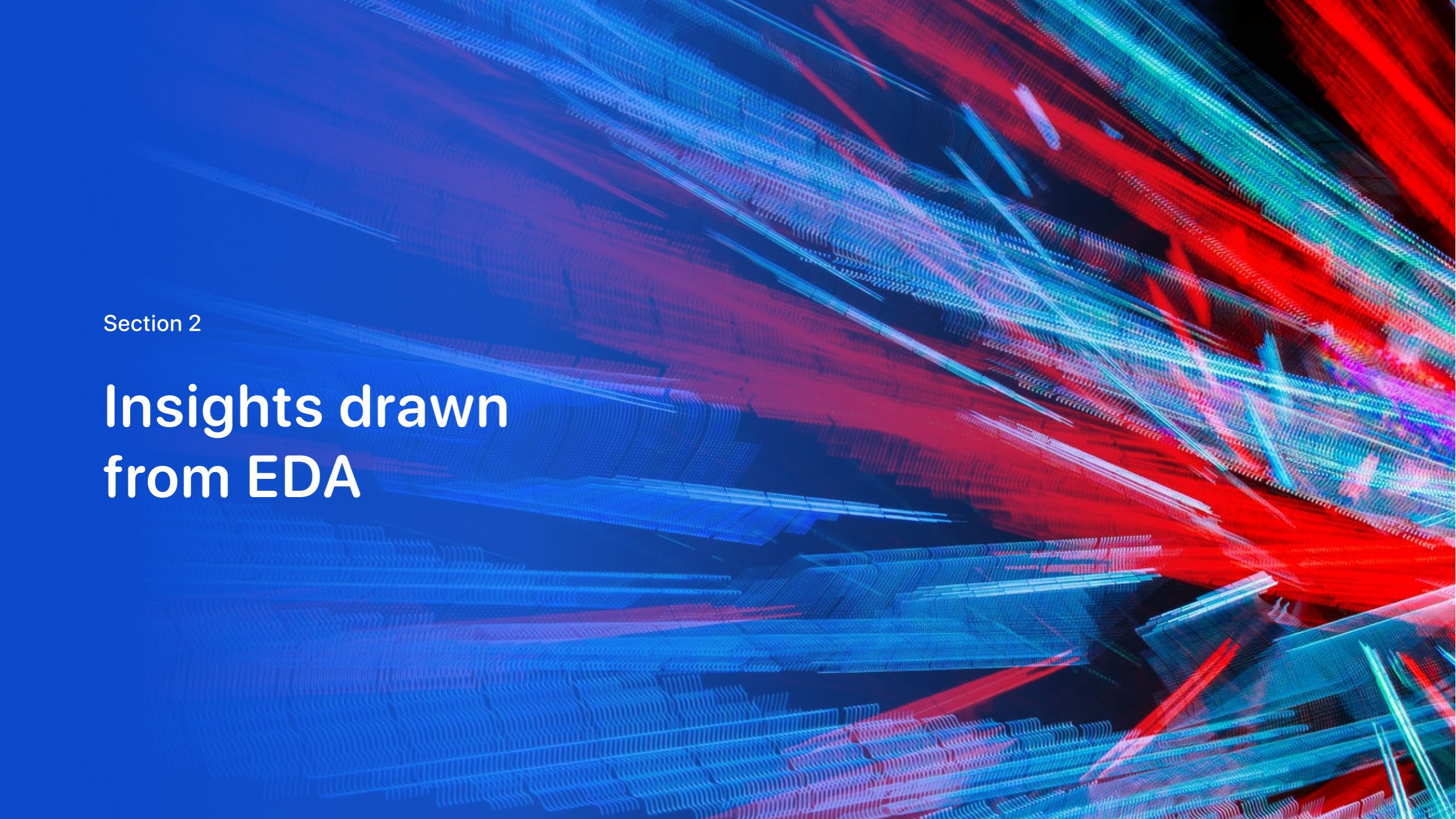
# Predictive Analysis (Classification)




HERE is the GitHub URL of the completed interactive map with Folium notebook, as an external reference.

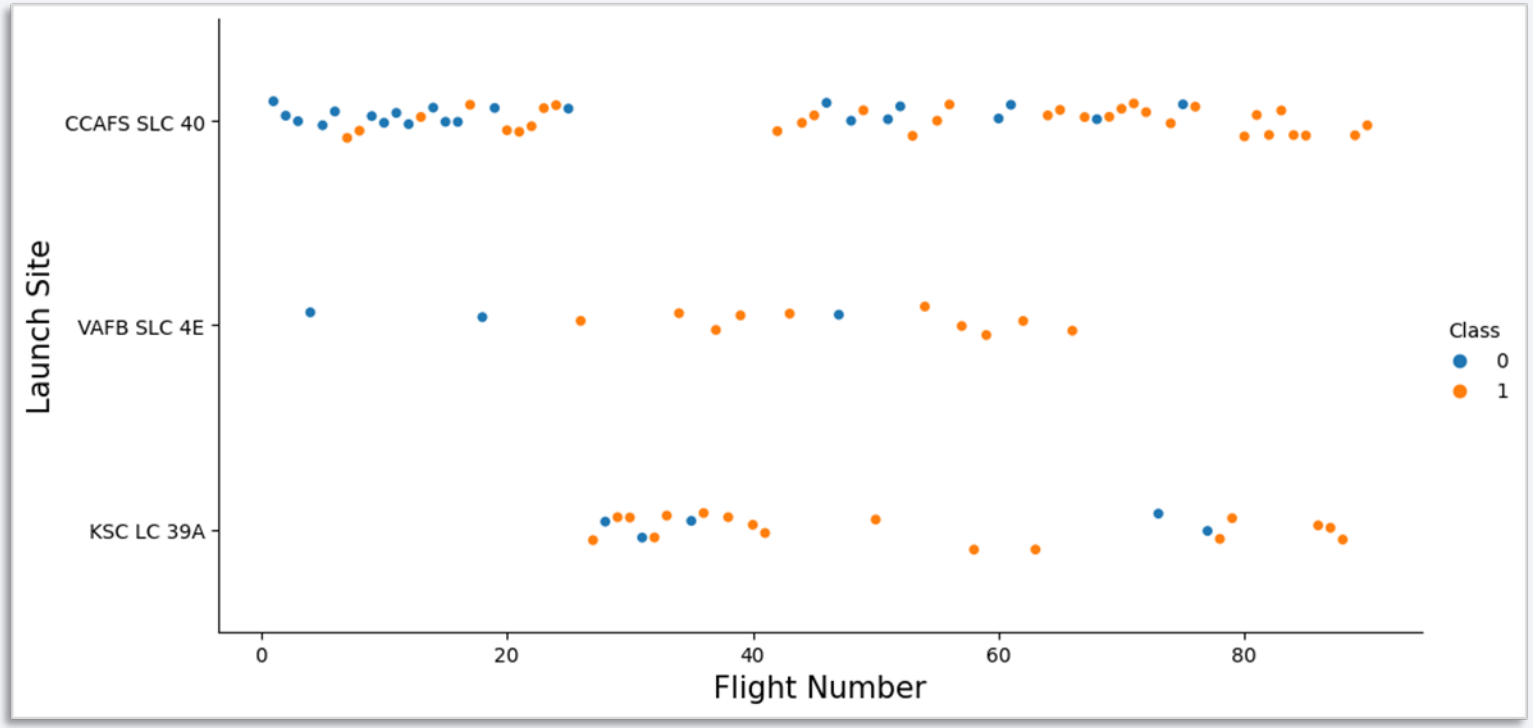
# Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn
from EDA

# Flight Number vs Launch Site

The greater the number of flights on each launch site, the higher the success rate

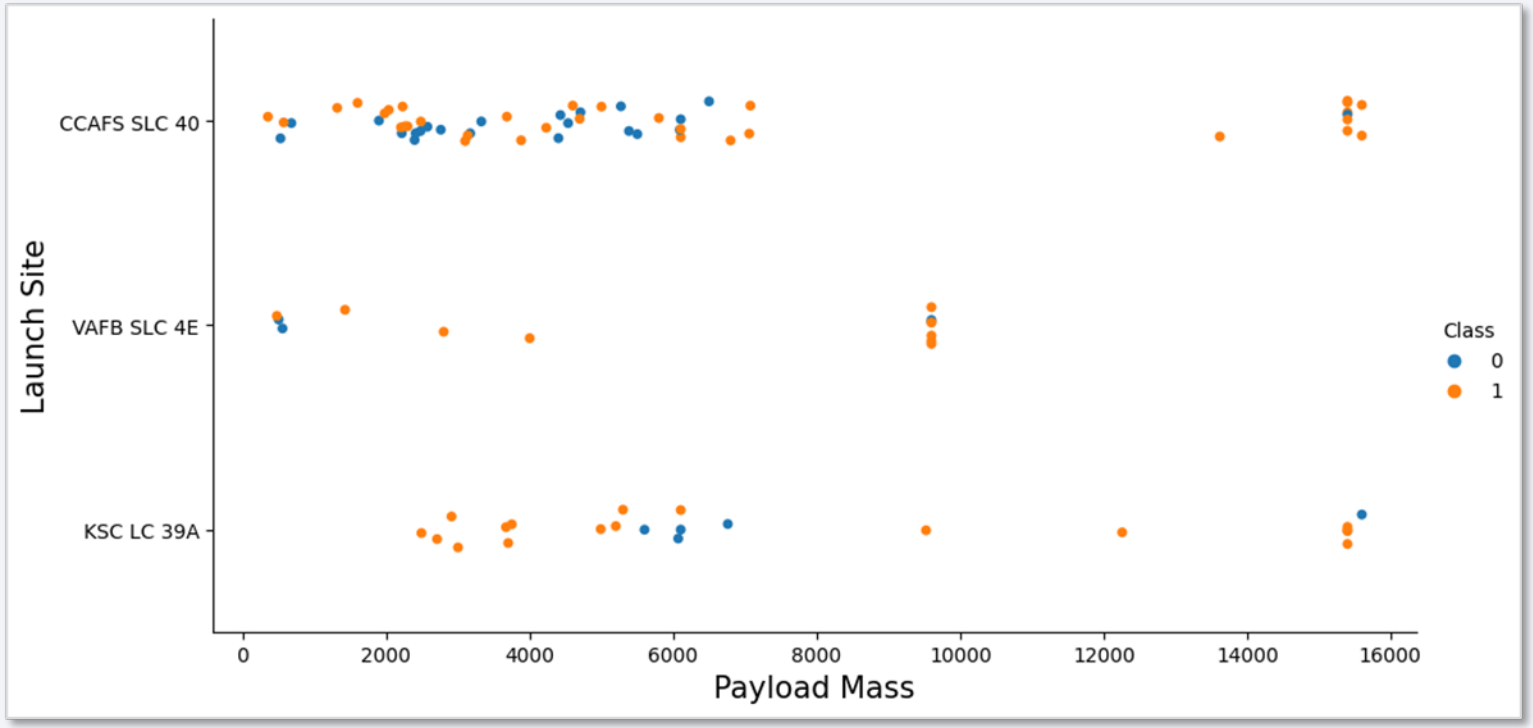



Scatter plot of Flight Number vs Launch Site

The best launch site is KSC LC 39A, where the most past and recent launches were successful

# Payload vs Launch Site

Depending on the launch site, a heavier payload may be necessary for a successful launch




Scatter plot of Payload Mass vs Launch Site

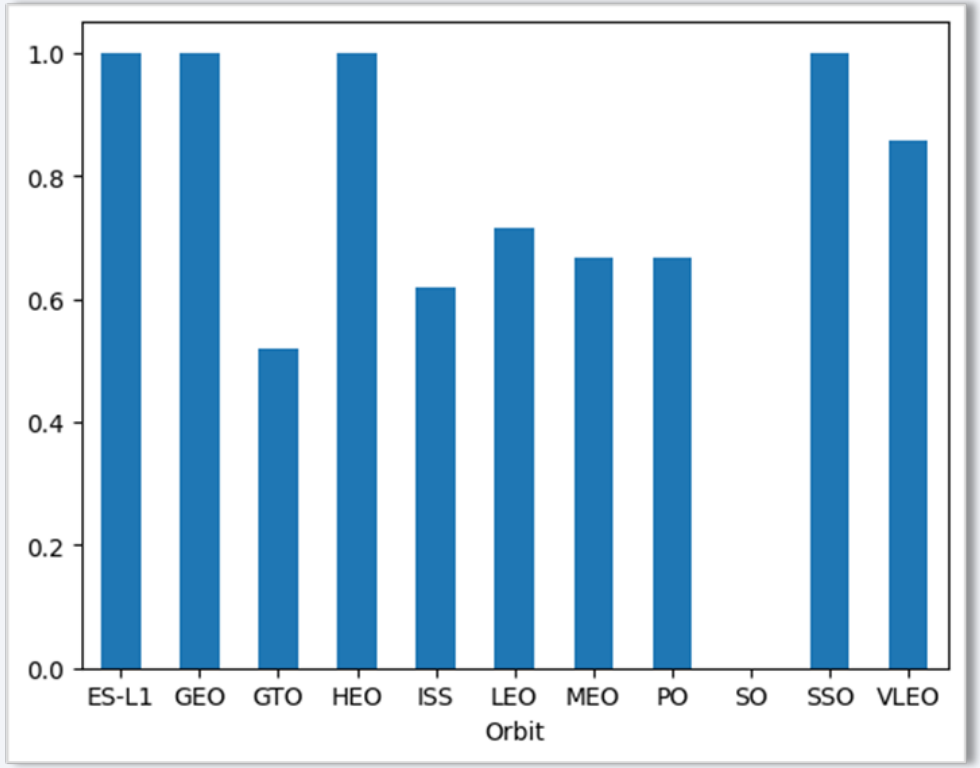
Payloads weighing more than 7,000 kg have a high success rate.

However, there is no definite pattern to suggest that the success rate of a launch depends on the payload mass or the launch site.

# Success Rate vs Orbit Type

The ES-L1, GEO, HEO and SSO orbits have a significant high success rate (nearly 100%), whereas the SO orbit produced 0% rate of success.

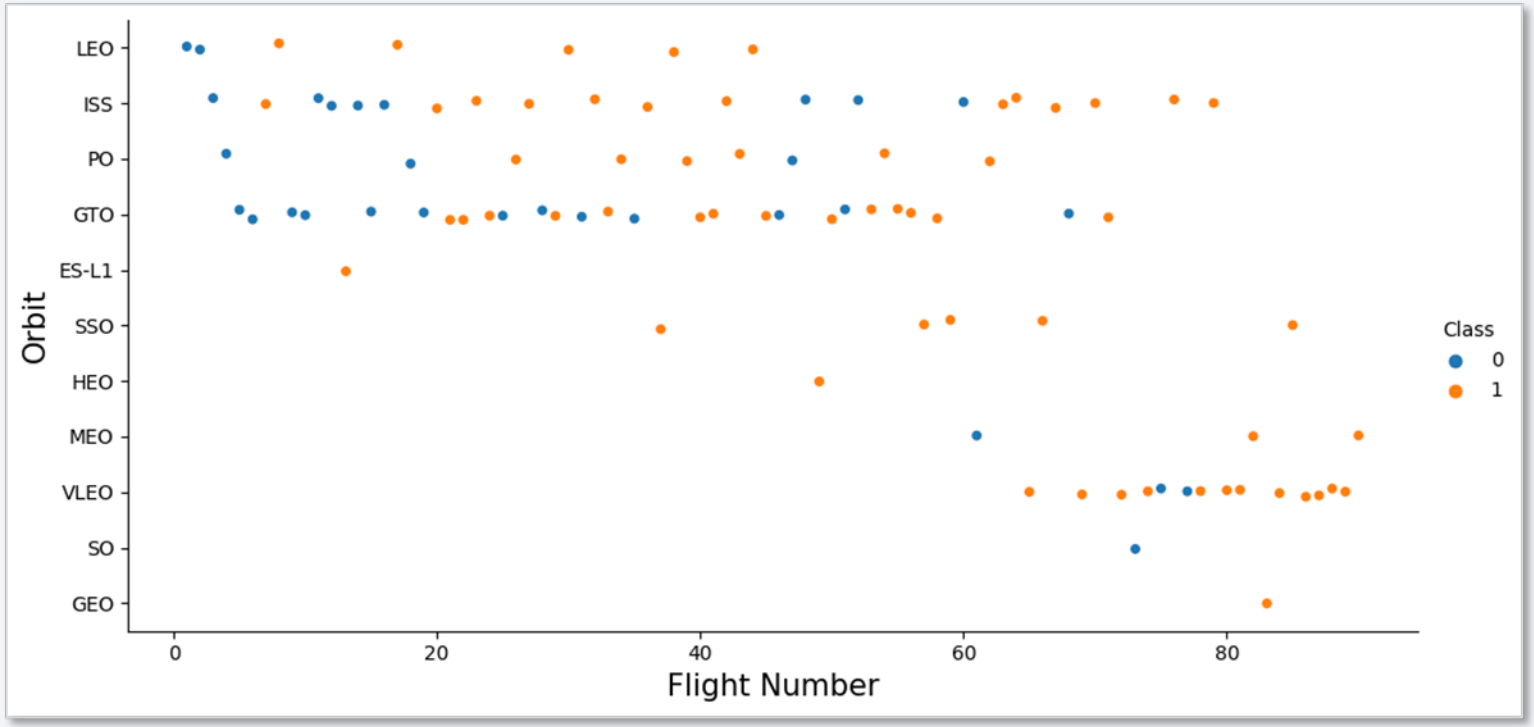
However, deeper analysis revealed that some of these orbits, like GEO, SO, HEO, and ES-L1, only occurred once, so more datasets are required to see certain patterns or trends before we can draw any conclusion.




Bar chart for the success rate of each orbit

# Flight Number vs Orbit Type

As the number of flights increased, so did the success rates in certain orbits



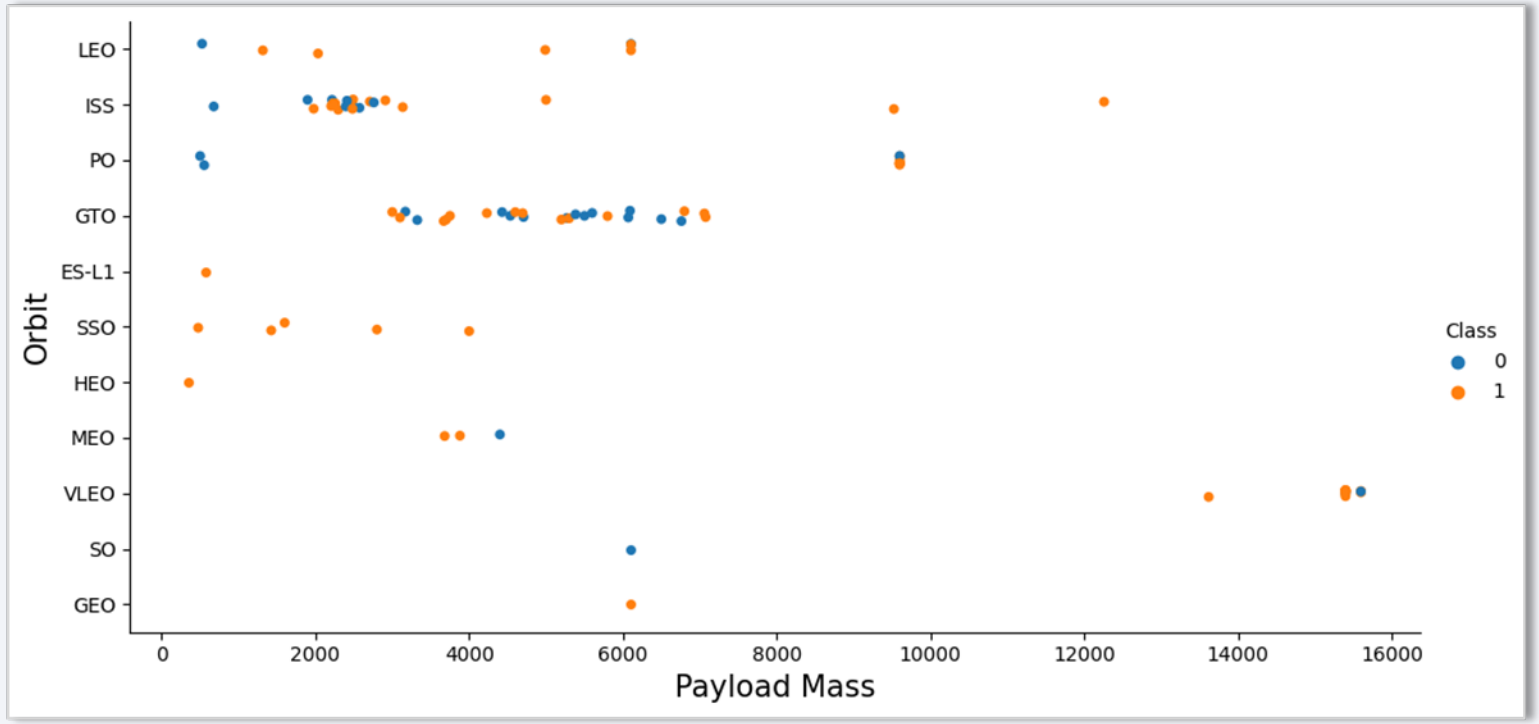

Scatter plot of Flight Number vs Orbit Type

Except for GTO, the greater the number of flights on each orbit, the higher the success rate (especially on LEO orbit).

As mentioned above, orbit that only occurred once should be excluded from the analysis.

# Payload vs Orbit Type

The weight of payloads may have positive or negative impacts on the success rate of launches depending on specific orbit type




Scatter plot of Payload vs Orbit Type

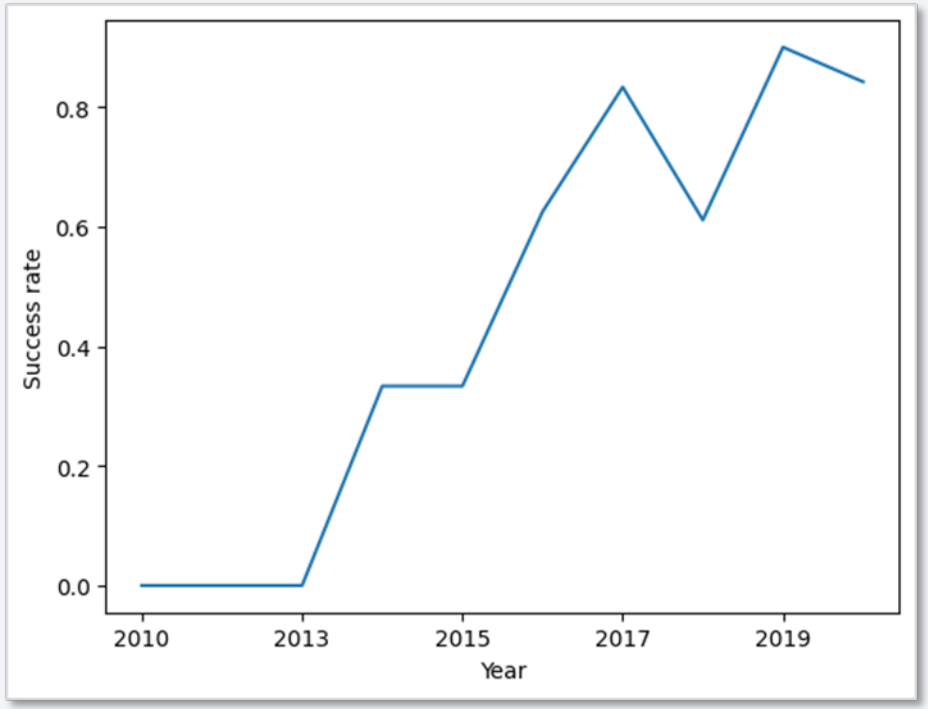
The attributes appear to have no relationship in the GTO orbit.

Again, more datasets are required to observe patterns or trends in the SO, GEO, ES-L1 and HEO orbits.

# Launch Success Yearly Trend

From 2013 to 2020, the SpaceX success rate has increased rapidly

There have been no changes in the first three years, it may be a period of adjustment and technological advancement




Line chart for the average launch success yearly trend

# All Launch Site Names

Find the names of the unique launch sites

```
SELECT DISTINCT launch_site FROM spacex
```

The use of DISTINCT in the query was used to select the unique ‘launch_site’ names from the data set

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

# Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with ‘CCA’

```
SELECT * FROM spacex WHERE launch_site LIKE 'CCA%' LIMIT 5
```

The WHERE clause, followed by the LIKE clause, filtered out launch sites with the prefix ‘CCA’. LIMIT 5 displays the first 5 records from the filtered set

| 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

Calculate the total payloads carried by booster from NASA (CRS)

```
SELECT SUM(payload_mass_kg_) as total_payload_mass FROM spacex
WHERE customer = 'NASA (CRS)'
```

| total_payload_mass |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

```
SELECT AVG(payload_mass_kg_) FROM spacex
WHERE booster_version LIKE '%F9 v1.1%'
```

This query returned the average of all payload masses where the booster version contains the substring ‘F9 v1.1’, which is 2534 kg

| avg_payload |
| --- |
| 2534 |

# First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

```
SELECT MIN(DATE) FROM spacex WHERE landing_outcome = 'Success (ground pad)'
```

| first_successful_date |
| --- |
| 2015-12-22 |

The WHERE clause filtered the data set in order to keep only records where landing on ground pad was successful.

The MIN function returned the record with the oldest date.

# Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
SELECT booster_version, payload_mass_kg_ FROM spacex
WHERE payload_mass_kg_ BETWEEN 4000 AND 6000
AND landing_outcome = 'Success (drone ship)'
```

| booster_version | payload_mass_kg_ |
|---|---|
| F9 FT B1022 | 4696 |
| F9 FT B1026 | 4600 |
| F9 FT B1021.2 | 5300 |
| F9 FT B1031.2 | 5200 |

# Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
SELECT mission_outcome, COUNT(*) as occurences FROM spacex
GROUP BY mission_outcome
```

Grouping mission outcomes and counting records for each group led us to the summary below

| mission_outcome | occurences |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

```
SELECT DISTINCT booster_version, payload_mass_kg_
FROM spacex
WHERE payload_mass_kg_ = (SELECT
                          MAX(payload_mass_kg_) FROM spacex)
```

| 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 |

The subquery filtered the data by returning only the heaviest payload mass with the MAX function. The main query used the subquery results and returned unique booster version (SELECT DISTINCT) with the heaviest payload mass

# 2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
SELECT booster_version, launch_site,
landing_outcome FROM spacex
WHERE YEAR(DATE)=2015
AND landing_outcome = 'Failure (drone ship)'
```

| booster_version | launch_site |
|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

This query returned booster versions and launch sites where landing was failed within 2015.

YEAR(DATE) extracted the year in DATE column.

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

```
SELECT landing_outcome, COUNT(*) as occurences
FROM spacex
WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY landing_outcome
ORDER BY occurences DESC
```

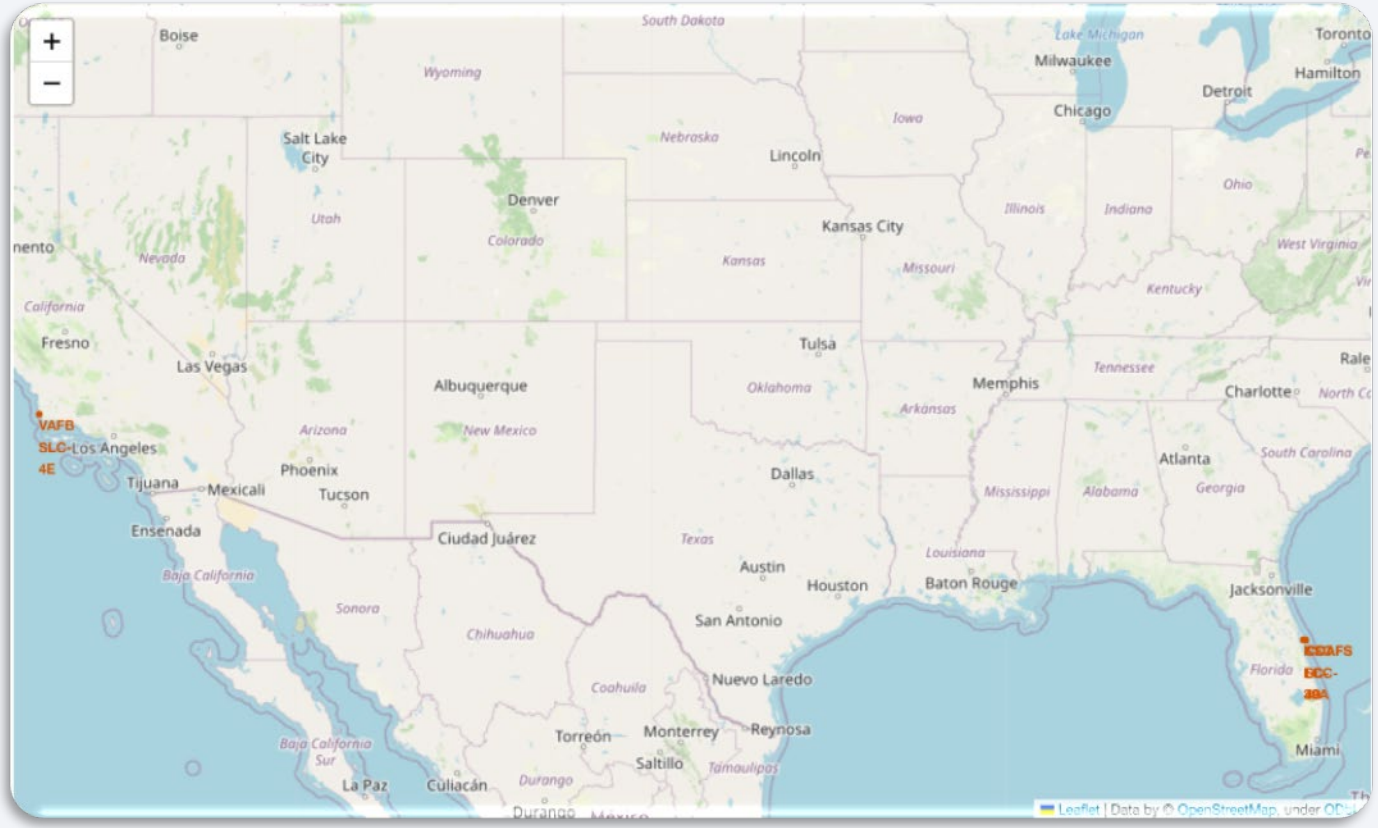
This query ranked landing outcomes by their number of occurrences from 4 Jun 2010 to 20 Mar 2017. The GROUP BY clause grouped the results by landing outcome and ORDER BY DESC displayed the results in descending order of occurrences.

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

Section 3

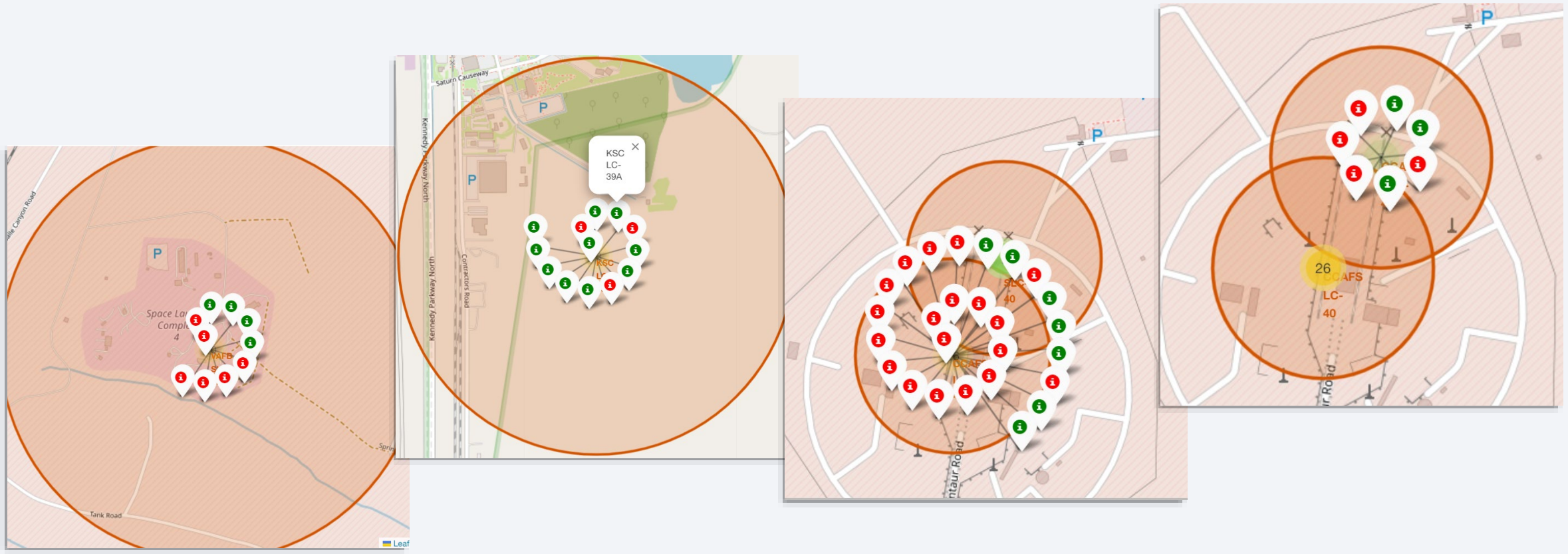
# Launch Sites Proximities Analysis

# All launch sites



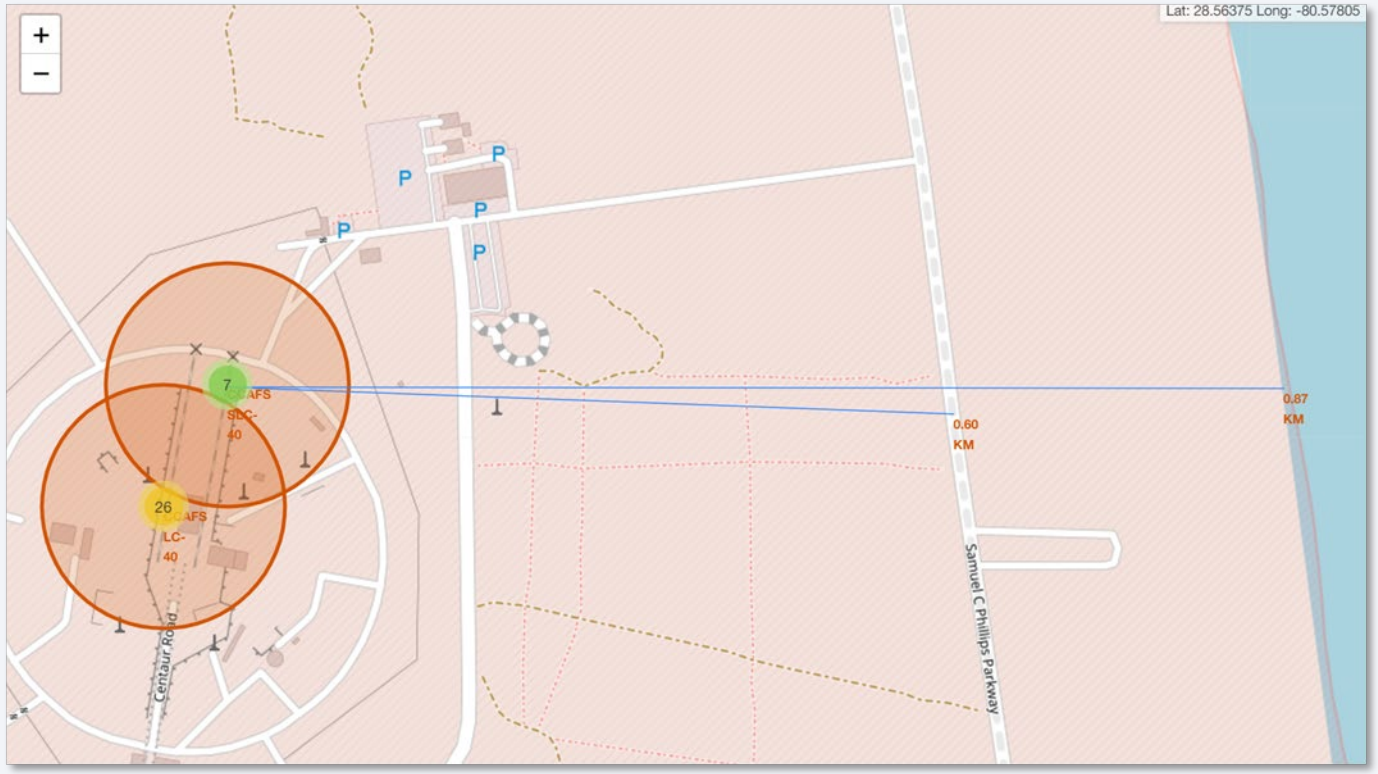

Launch sites are located on the coast of the United States, probably for safety, but not too far from roads and railways.

# The colour-labelled launch site

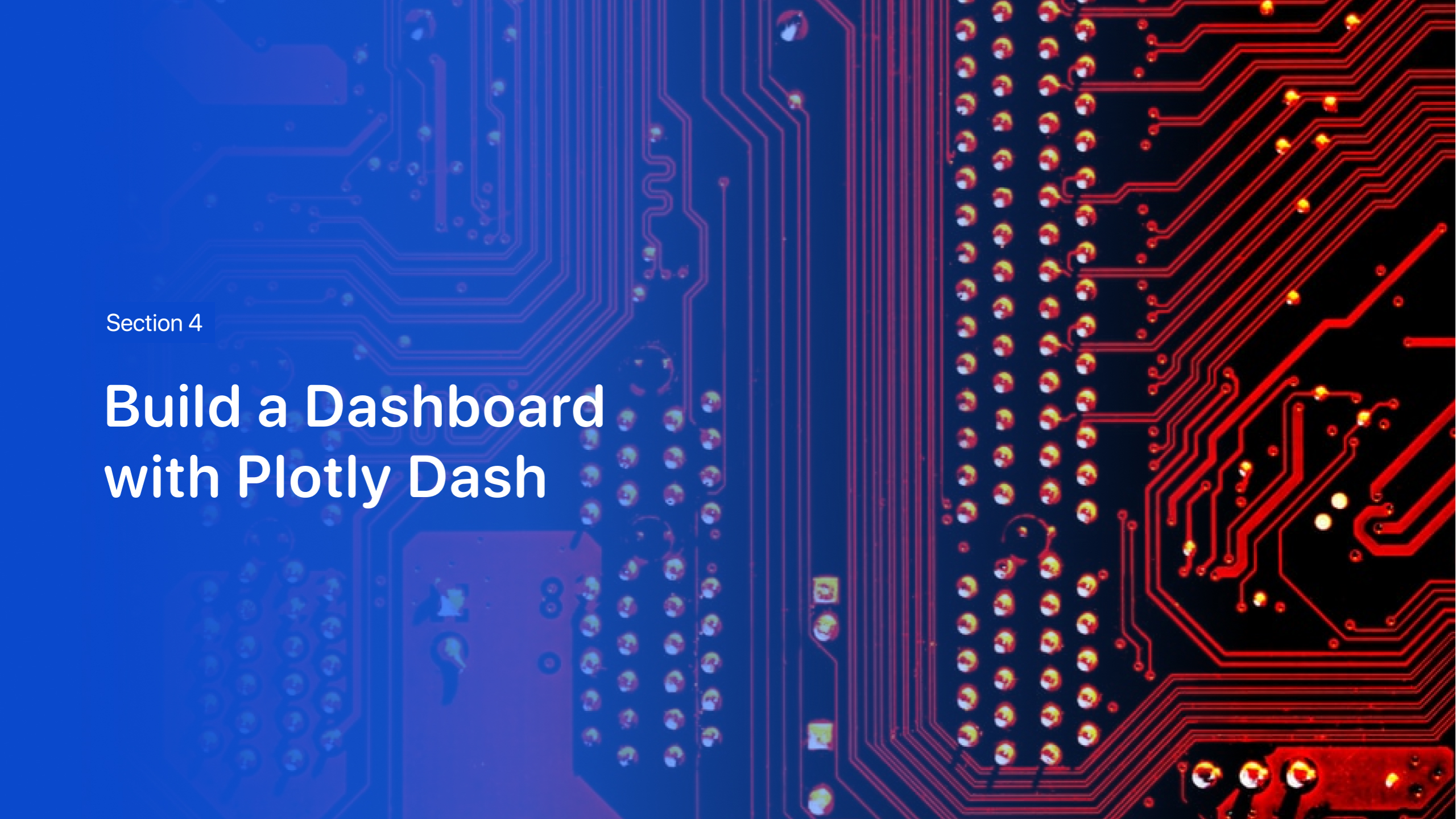



The green marker denotes successful launches. The red marker denotes failed launches. We notice that the KSC LC-39A has a higher success rate.

# Logistics and Safety

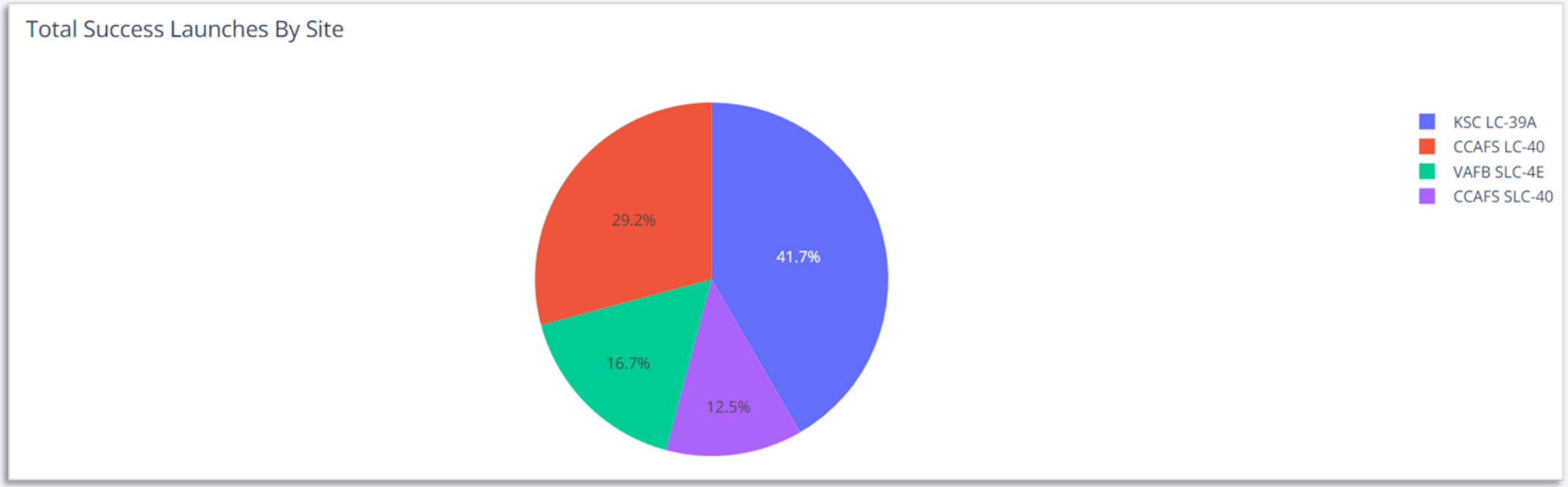



The launch site CCAFS SLC-40 has good logistics infrastructure, as it located in close proximity to railways, highways, coastline and relatively far from cities


Section 4
Build a Dashboard with Plotly Dash

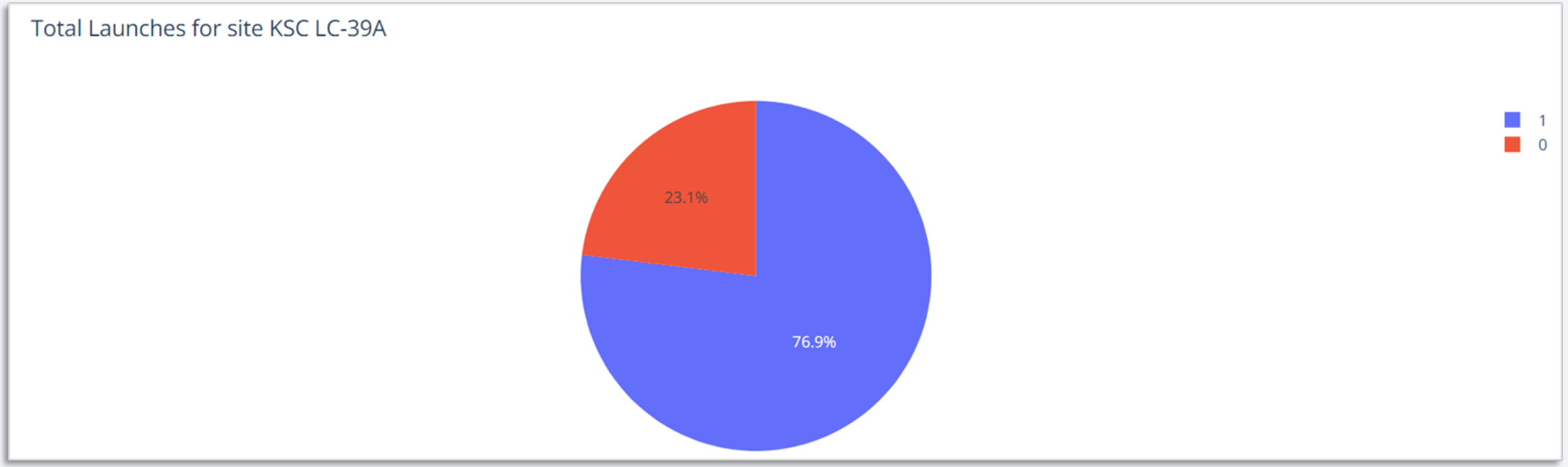
# Total Successful Launches by Site

- Launch site appears to be an important factor of a successful mission
- KSC LC-39A stands out to be the best launch site with the success rate of 41.7%

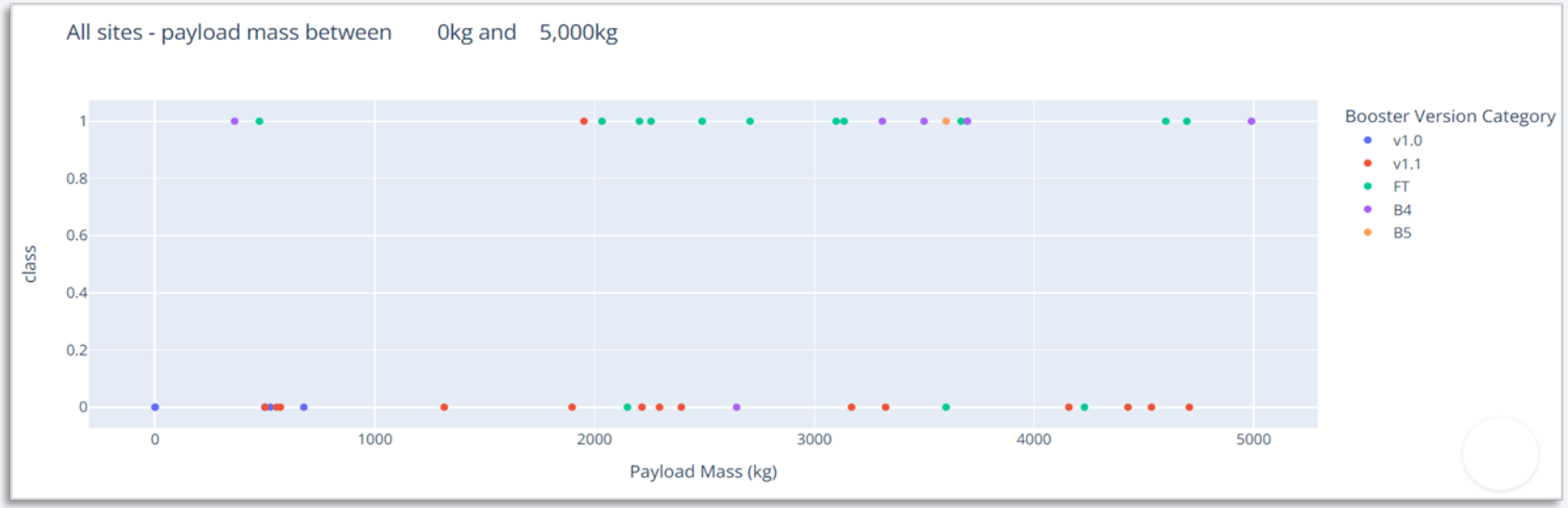


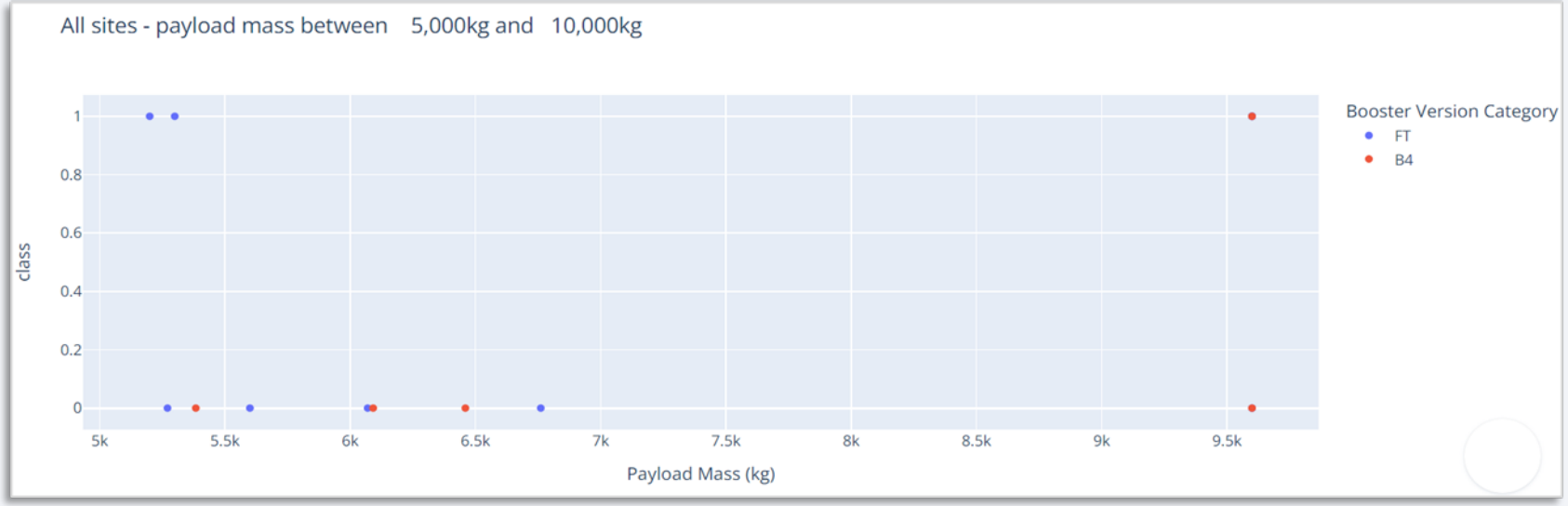
# Launch Success Ratio for KSC LC-39A

76.9% of launches are successful in the launch site KSC LC-39A



# Payload Mass vs Outcome for all sites with different payload mass selected




Low weighted payloads have a higher success rate than heavy weighted payloads.

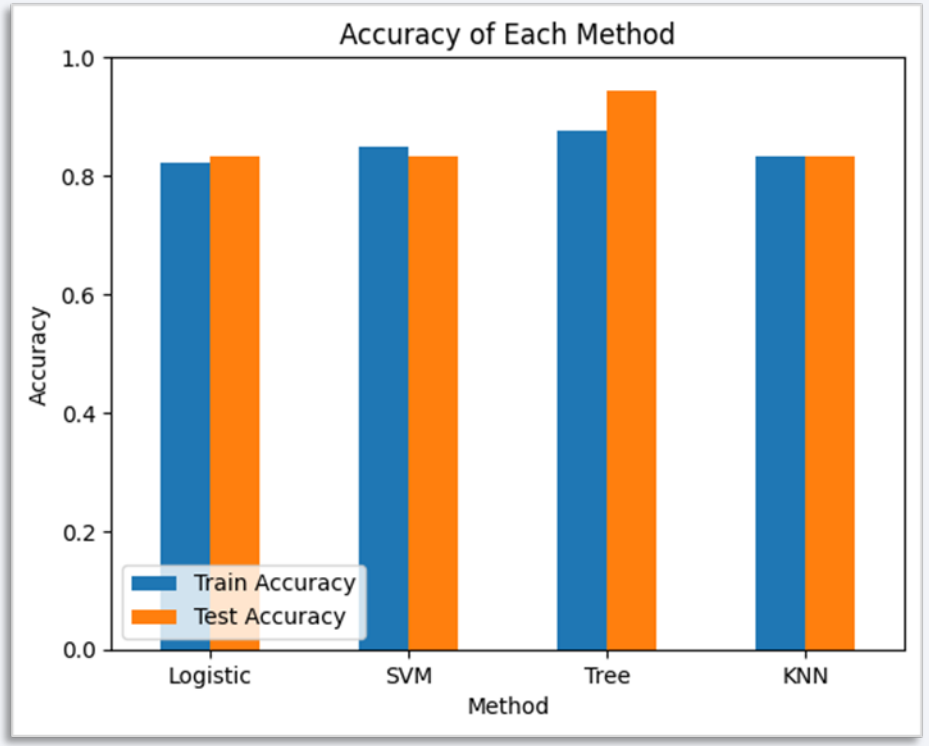
There is insufficient information to comment on payloads ranging from 7,000 to 9,500 kg.

Section 5

# Predictive Analysis (Classification)

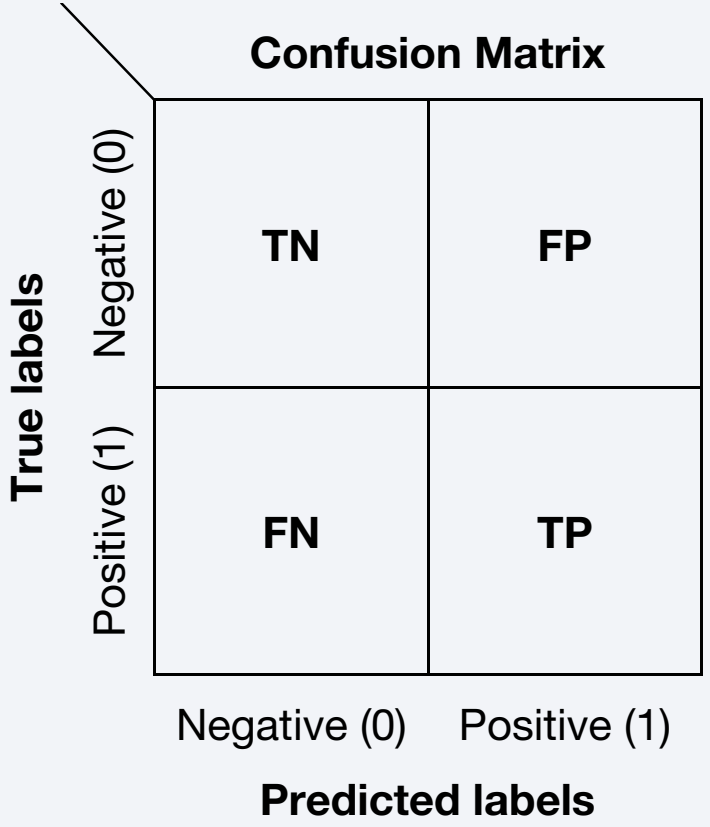
# Classification Accuracy

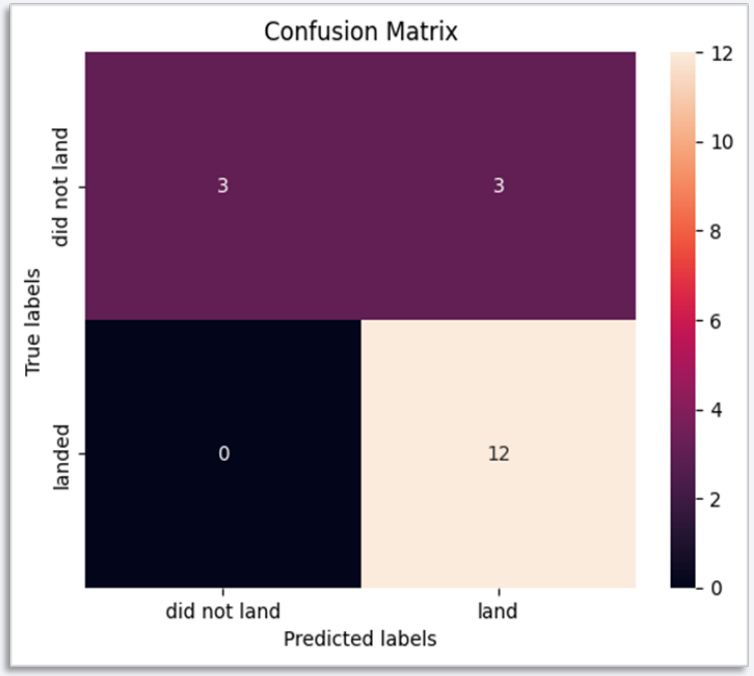
- Four classification models were built, and their accuracies are shown in the accompanying figure.
- Both the train and test accuracy indices of the Decision Tree Classifier are above 87%, which are the highest values in the chart.



# Confusion Matrix of Decision Tree Classifier

The confusion matrix below verifies the accuracy of Decision Tree Classifier due to the large values in true positive and true negative areas compared to false areas



# Conclusions

- In this project, different data sets were analysed to refine conclusions.
- Although GEO, HEO, SSO and ES-L1 orbits have the highest success rates, more data sets are required to analyse patterns or trends before we can draw any conclusion.
- KSC LC-39A is the best launch site according to its success rate of launches.
- Depending on the orbits, payload mass can be a criterion to consider for the success of a mission. Some orbits required a light or heavy payload mass; however, low weighted payloads performed better than the heavy ones in general.
- Successful landing outcomes seem to improve over time.
- Decision Tree Classifier can be used to predict successful landings and increase profits because of its accuracy.

# Appendix

All Python code, Jupyter notebook, SQL queries, charts and data sets that I have created or analysed during this project can be accessed from https://github.com/LeoThaiHuy/Applied-Data-Science-Capstone.git

# Thank you!