IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Aibamaya Alvarez
10/29/2022

# Outline

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

GitHub repo: https://github.com/aibamaya/Falcon_9_project

# Executive Summary

- Summary of methodologies
  - Data Collection via API and Web Scraping
  - Exploratory Data Analysis with Data visualizations
  - Exploratory Data Analysis with SQL
  - Interactive map with Folium
  - Dashboards with Plotly Dash
  - Predictive Analysis

- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Maps and Dashboards
  - Predictive results

# Introduction

- This project aims to predict whether the first stage of Falcon 9 will land successfully.
- SpaceX advertises Falcon 9 rocket launches on its website, costing 62 million dollars. Other providers cost upward of 165 million dollars each. Much of the savings are because SpaceX can reuse the first stage.
- By determining if the first stage will land, we can determine the cost of a launch. This information is valuable if an alternate company wants to bid against SpaceX for a rocket launch.

We will be answering these questions in this presentation:

1. What conditions will allow SpaceX to achieve the best landing success rate?
2. What are the main features to determine a successful or failed landing?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX REST API and Web Scraping from Wikipedia
- Perform data wrangling
  - Dealing with missing values, dropping unnecessary columns, and performing one hot encoding to categorical variables.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Application of Grid Search to train, tune and test classification models.

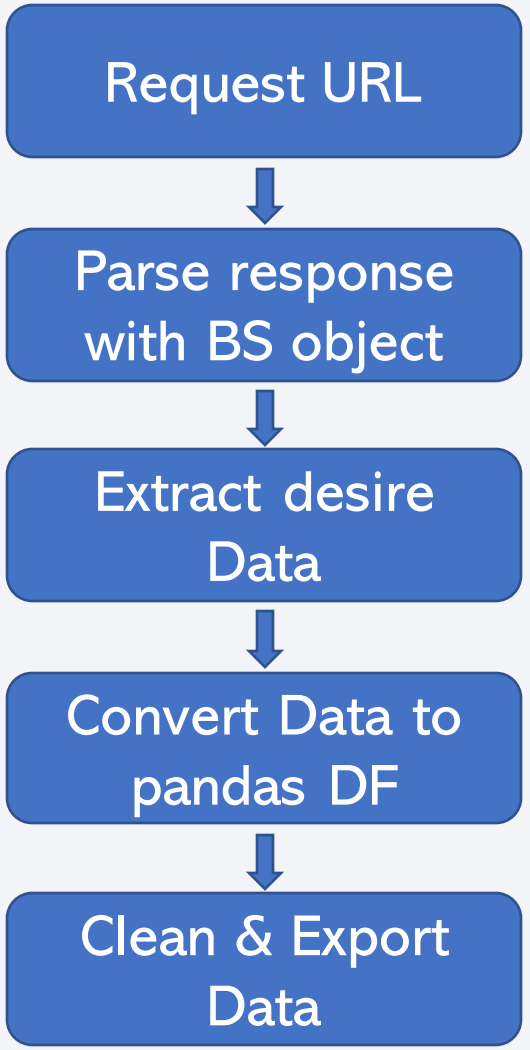
# Data Collection

Dataset were collected from SpaceX Rest API and Wikipedia

- SpaceX offers a free Rest API from where the data was collected:
  - URL: https://api.spacexdata.com/v4/
  - To obtain information about rockets, landing and payload.
- The API data was complemented with data obtained from Wikipedia:
  - URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
  - To obtain information about rockets, landing and payload.

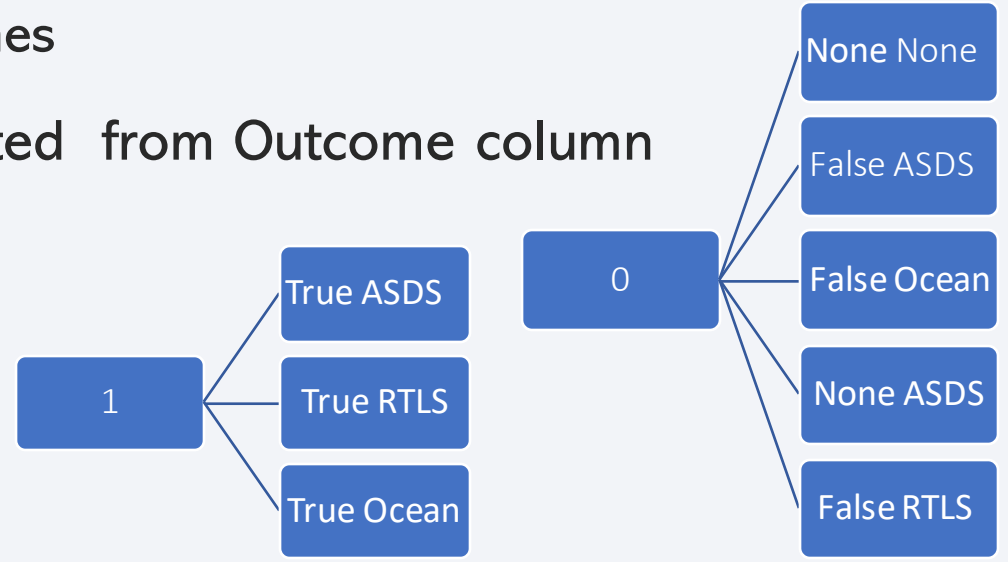
# Data Collection - Scraping

- Request to the Wikipedia URL
- Response object was parsed using the BeautifulSoup library.
- The desire data was extracted from the Wikipedia's tables and cleaned it.
- The data was converted to a pandas data frame and exported.
- Link to the notebook in GitHub



# Data Wrangling

- Some exploratory data analysis was perform to:
  - Identify and calculate the percentage of the missing values in each attribute
  - Identify which columns were numerical and categorical
  - Determine the number of launches on each site
  - Determine the number and occurrence of each orbit
  - Determine the number of landing outcomes
- A landing outcome label (Class) was created from Outcome column




- Link to the notebook in GitHub

# EDA with Data Visualization

- Scatter Graph (to visualize relationships between features)
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit type
  - Payload vs. Orbit type
- Bar Graph (to visualize relationships among categorical variables)
  - Success rate vs. Orbit type
- Line Graph (to visualize variables and their trends)
  - Launch success vs. Year
- Link to the notebook in GitHub

# EDA with SQL

- The fallowing queries were performed to gather and understand the data:
  - Display the names of the unique launch sites in the space mission.
  - Display 5 records where launch sites begin with the string 'CCA'.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display average payload mass carried by booster version F9 v1.1.
  - List the date when the first successful landing outcome in ground pad was archived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the "booster_versions" which have carried the maximum payload mass.
  - 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.
  - Rank the count of successful "landing_outcomes" between the date 04-06-2010 and 20-03-2017 in descending order.



- Link to the notebook in GitHub

# Build an Interactive Map with Folium

- We analyzed the existing launch site locations to find geographical patterns about them.
  - We marked all launch sites on a map.
  - We added launch outcomes for each site to see which had high success rates.
    - If a launch was successful (class=1), we use a green marker; if a launch failed, we use a red one (class=0) .
    - As a launch only happens in one of the four launch sites, we used marker clusters to simplify a map containing many markers having the same coordinate.
  - We explored and analyzed the proximities of launch sites by calculating the distances between a launch site to its proximities.

- Link to the notebook in GitHub

# Build a Dashboard with Plotly Dash

- We use the following components to build a Plotly Dash real-time application for users to perform interactive visual analytics on SpaceX launch data.

    1. A "launch_site" dropdown to see which site has the most considerable success count. We can also select a specific one to check its detailed success rate.

    2. A pie chart to visualize launch success counts based on selected site dropdown.

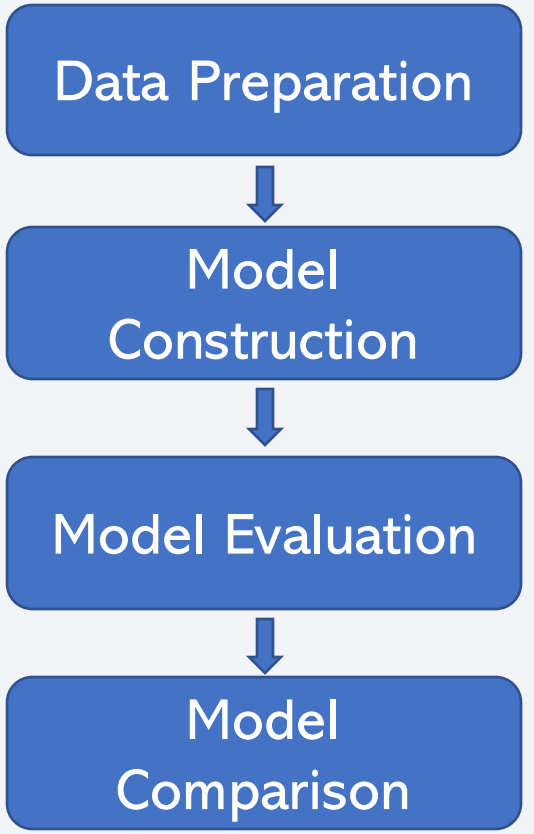
    3. A range slider to select between different payloads.

    4. A scatter chart to visually observe how payload may be correlated with mission outcomes for the selected site(s). In addition, we color-label the Booster version on each scatter point so that we may observe mission outcomes with different boosters.

- Link to the notebook in GitHub

# Predictive Analysis (Classification)

- Data Preparation
  - Select the feature matrix and the response vector
  - Standardize the feature matrix
  - Split the data into train and test sets
- Model Construction
  - Select the algorithms (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors)
  - Set the hyperparameters for each algorithm
  - Train a GridSearchCV model with the training dataset
- Model Evaluation
  - Test the model on the test dataset
  - Compute the accuracy score
  - Plot the Confusion Matrix
- Model Comparison
  - Compare the models based on their accuracy score
- Link to the notebook in GitHub

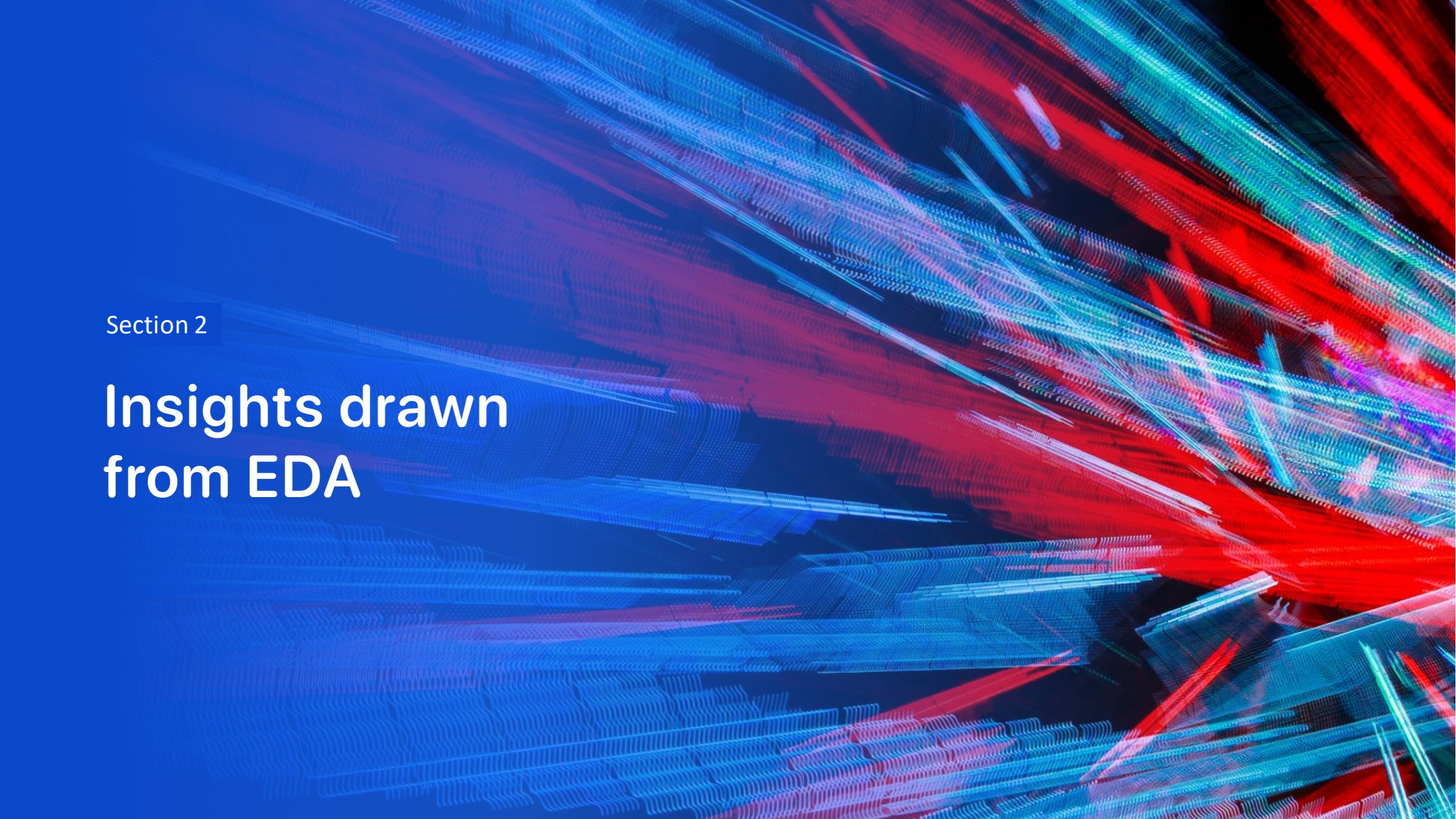


# Results

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

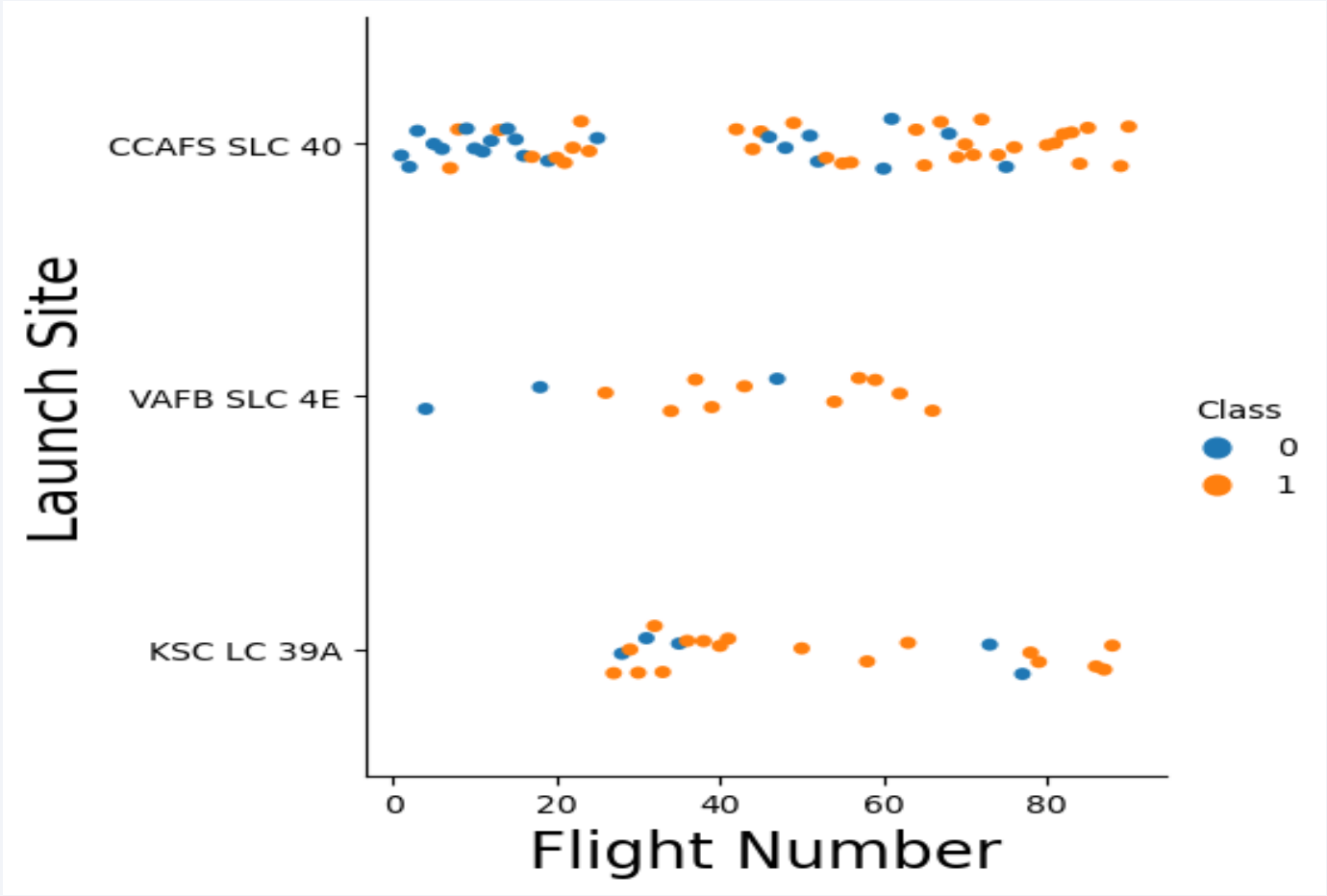
Section 2

# Insights drawn from EDA



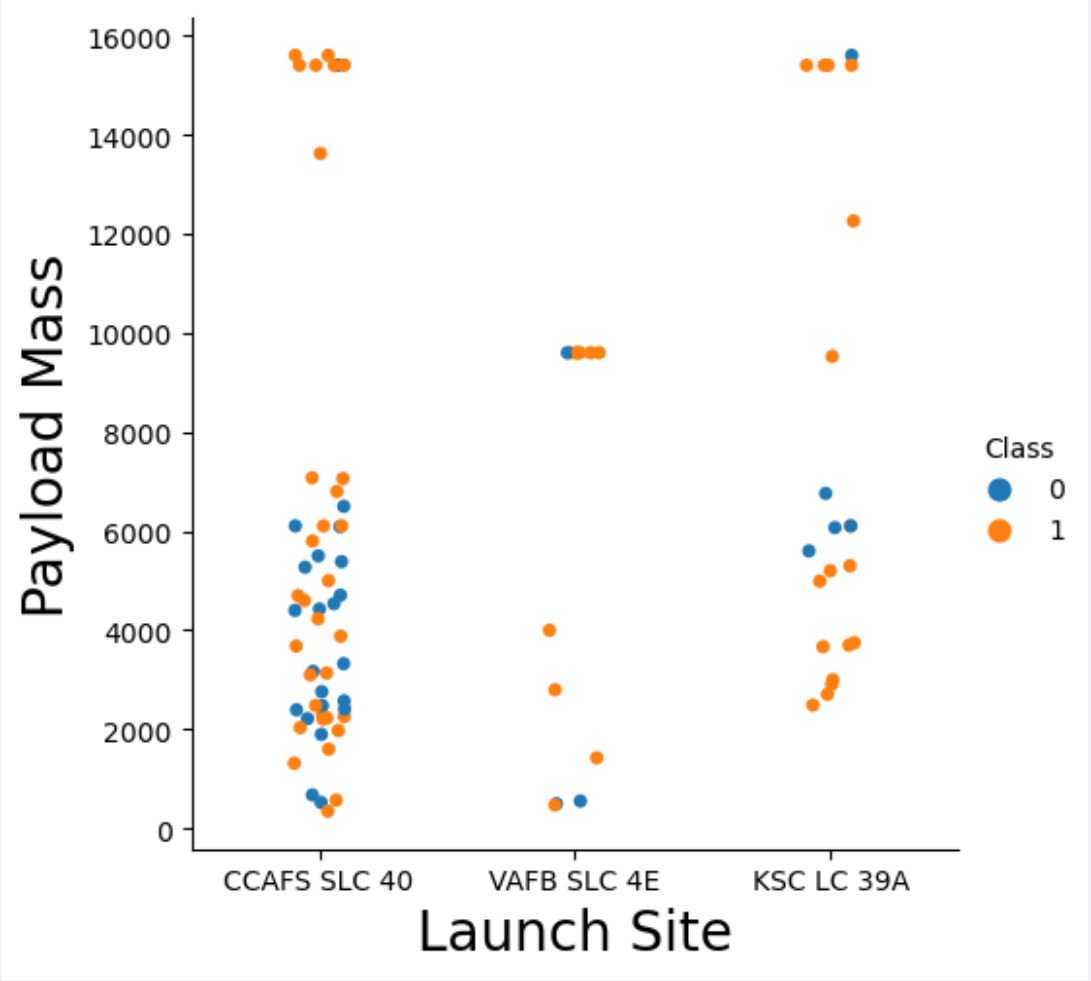
# Flight Number vs. Launch Site

- The launch site CCAFS SLC 40 is where most of the launches have taken place.
- The success rate seems to increase when the flight number increases.



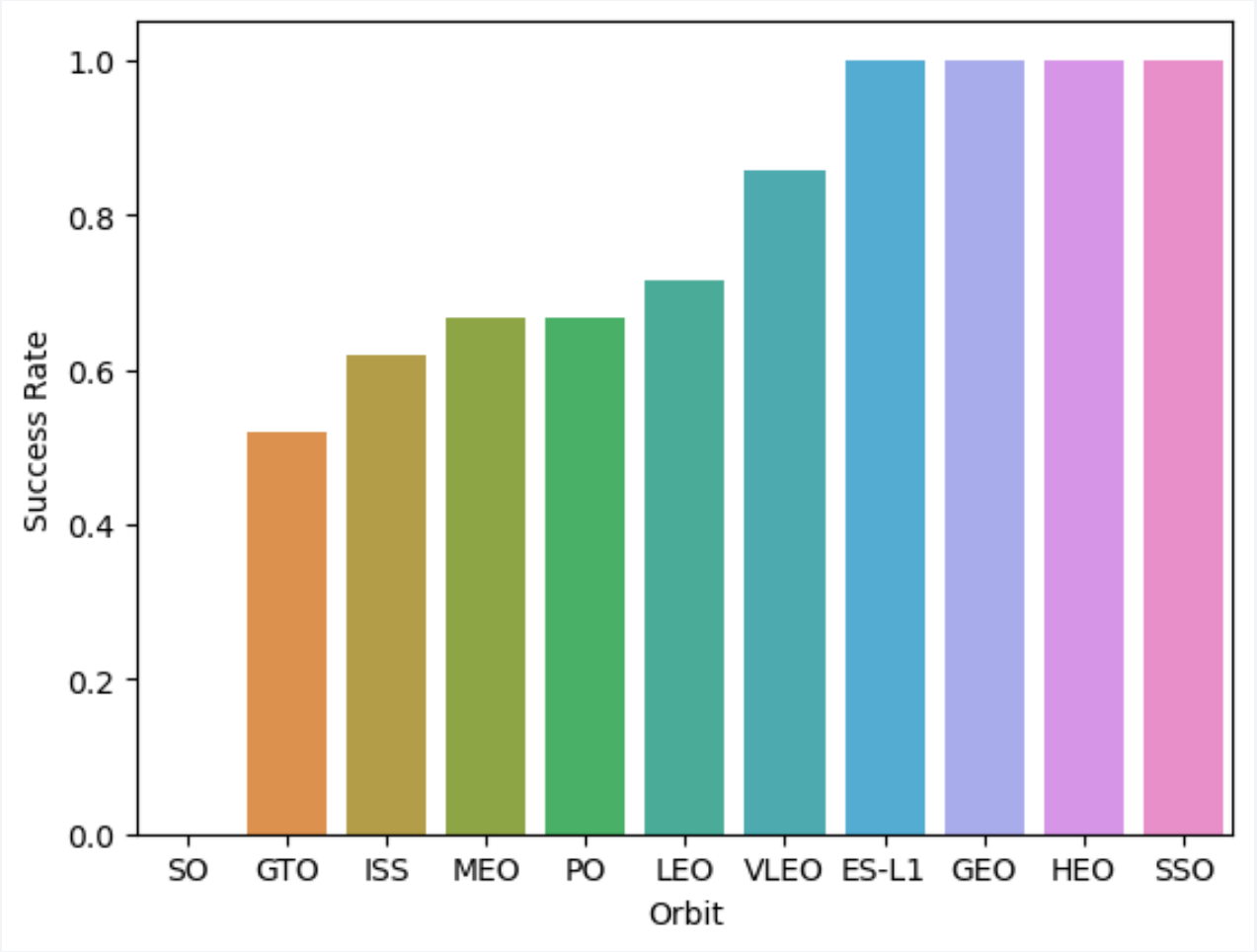
# Payload vs. Launch Site

- Most launches have a payload with a mass of less than 8000 kg.
- The VAFB SLC center doesn't have launches with heavy payloads.



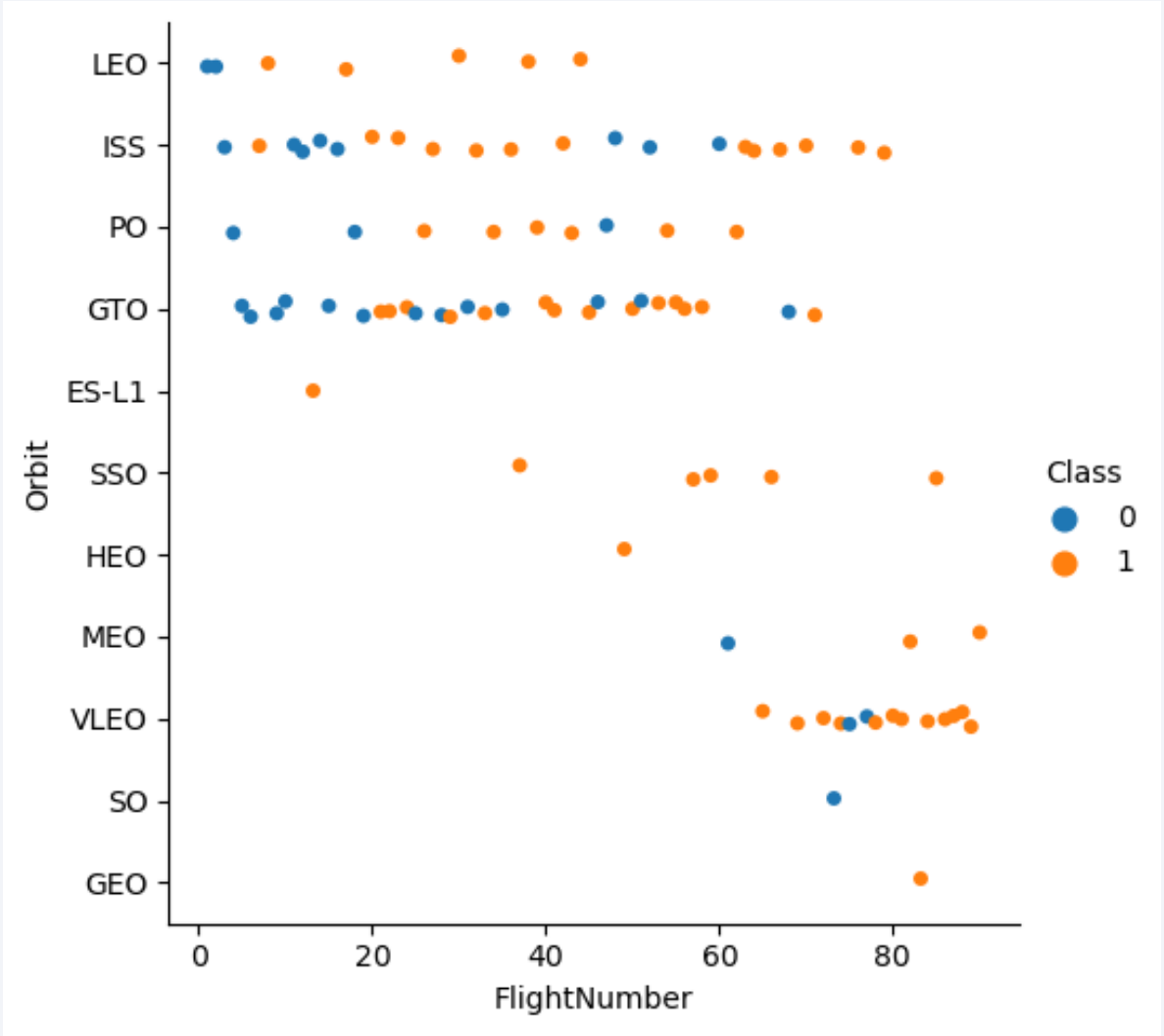
# Success Rate vs. Orbit Type

- The success rate differs between orbits.
- For the ES-L1, GEO, and HEO orbits best success rates are reached
- Note that SSO and SO are the same orbit (heliosynchronous orbit), so its success rate is not 100%.



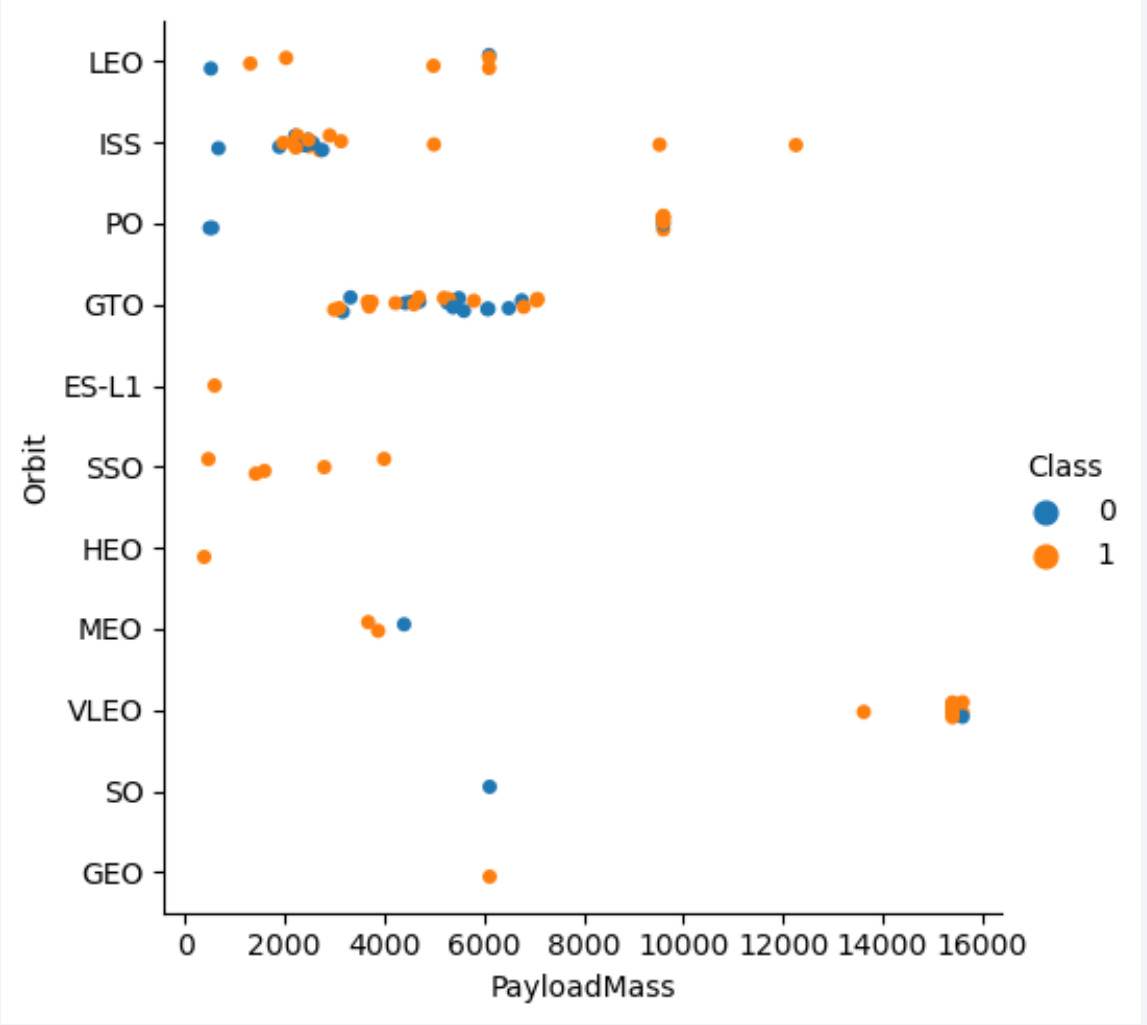
# Flight Number vs. Orbit Type

- The success rate tends to improve with the number of launches across all orbits.



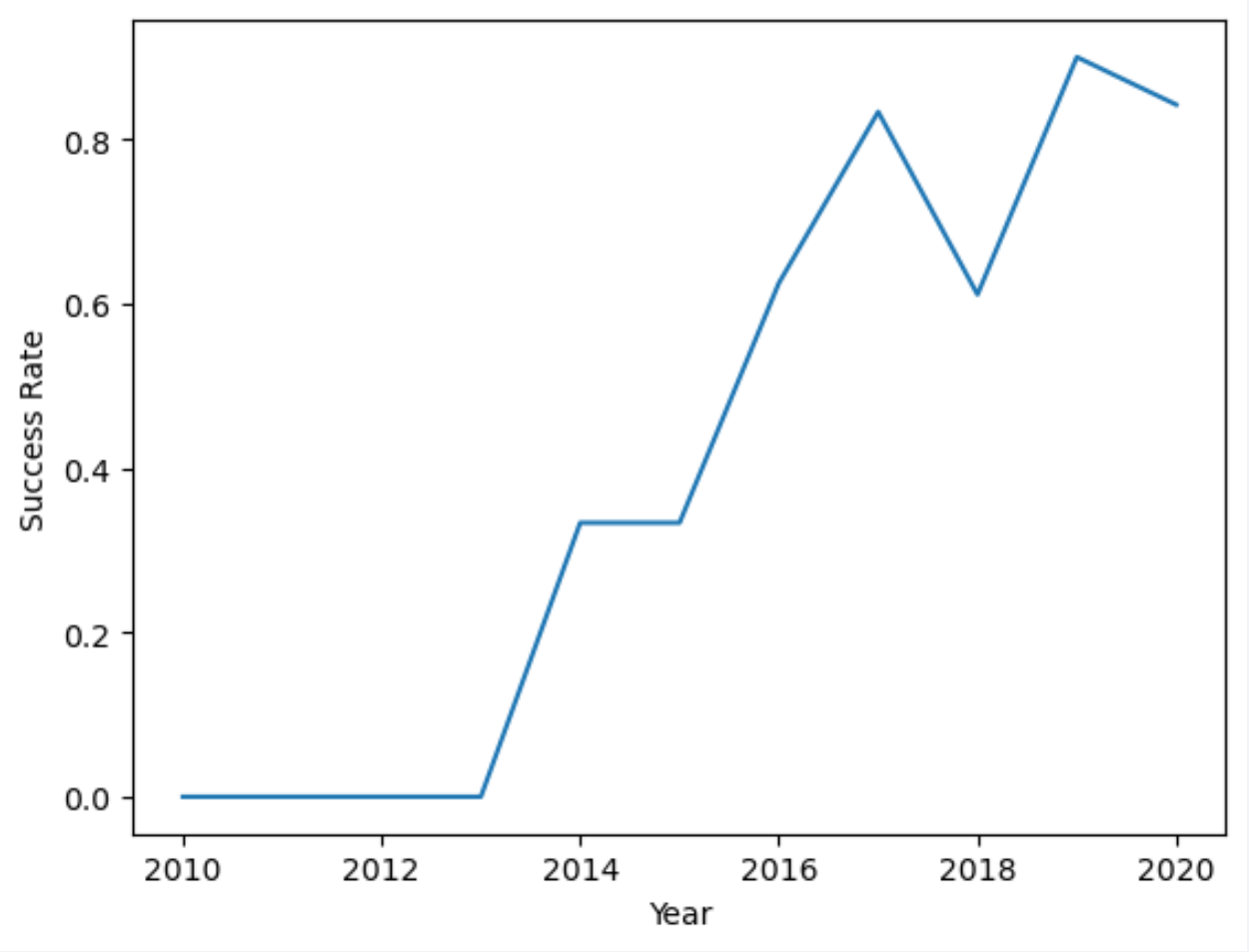
# Payload vs. Orbit Type

- The VLEO orbit has launches with heavy payloads only.
- ISS orbit has the broadest range of payloads mass.



# Launch Success Yearly Trend

- The success rate tends to improve over time



# All Launch Site Names

- Find the names of the unique launch sites

```
SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

- The DISTINCT clause is used to remove duplicates from the results
-
-

| 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 SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

| DATE | TIME__UTC_ | BOOSTER_VERSION | LAUNCH_SITE | PAYLOAD | PAYLOAD_MASS__KG_ | ORBIT | CUSTO |
|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX |
| 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 ( |
| 2012-05-22 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA ( |
| 2012-10-08 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA ( |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA ( |

- The * is used to select all columns
- The WHERE clause, followed by LIKE is used to filter the results.
- LIMIT is used to specify the number of records we want the query to return.

# Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
SELECT SUM(PAYLOAD_MASS__KG_) as "SUM(PAYLOAD_MASS__KG_)"
FROM SPACEXTBL
WHERE customer='NASA (CRS)'
```

| SUM(PAYLOAD_MASS__KG_) |
| --- |
| 45596 |

- SUM function aggregates all payload_mass where the customer was NASA (CRS)

# Average Payload Mass by F9 v1.1

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

```
SELECT AVG(PAYLOAD_MASS__KG_) as "AVG(PAYLOAD_MASS__KG_)"
FROM SPACEXTBL
WHERE Booster_version = 'F9 v1.1'
```

| AVG(PAYLOAD_MASS__KG_) |
| --- |
| 2928 |

- AVG function calculates the mean of all payload_mass where the Booster_version was 'F9 v1.1'

# First Successful Ground Landing Date

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

```
SELECT MIN(Date) as "first_successful_groundpad "
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

| first_successful_groundpad |
| --- |
| 2015-12-22 |

- MIN function find the minimum value of the dates return where the outcome was a successful landing outcome on ground pad

# 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
FROM SPACEXTBL
WHERE landing__outcome ='Success (drone ship)' and
(PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000)
```

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

- WHERE and BETWEEN clauses filter the result set of the query.
- BETWEEN allows us to select values within a range.

# Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
SELECT Mission_Outcome, COUNT(Mission_Outcome) as "Amount"
FROM SPACEXTBL
GROUP BY Mission_Outcome
```

| MISSION_OUTCOME | Amount |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- GROUP BY clause allows to combine the data according to the specified variable

# Boosters Carried Maximum Payload

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

```
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

- The use of subqueries expressions allow us to evaluate aggregate functions in the WHERE clause

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

# 2015 Launch Records

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

```
SELECT MONTH(Date) as "MONTH", Landing__Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE Landing__Outcome='Failure (drone ship)' and YEAR(Date)='2015'
```

| MONTH | LANDING__OUTCOME | BOOSTER_VERSION | LAUNCH_SITE |
|---|---|---|---|
| 1 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 4 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

- The MONTH() and YEAR() functions allow us to work with dates in SQL.

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

- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

```
SELECT Landing__Outcome, COUNT(*) as "Number_Success"
FROM SPACEXTBL
WHERE Landing__Outcome LIKE 'Success%' and
        Date BETWEEN Date('2010-06-04') and Date('2017-03-20')
GROUP BY Landing__Outcome
ORDER BY COUNT(*) DESC;
```

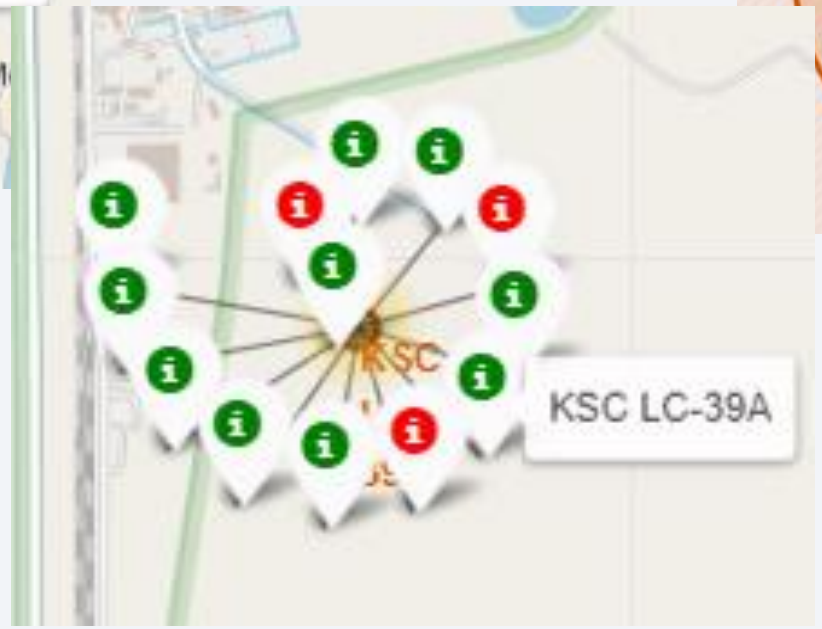
| LANDING__OUTCOME | Number_Success |
|---|---|
| Success (drone ship) | 5 |
| Success (ground pad) | 3 |

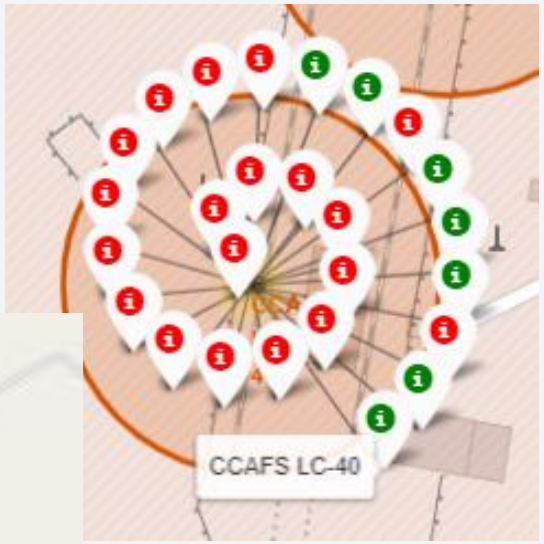

Section 3
Launch Sites
Proximities Analysis

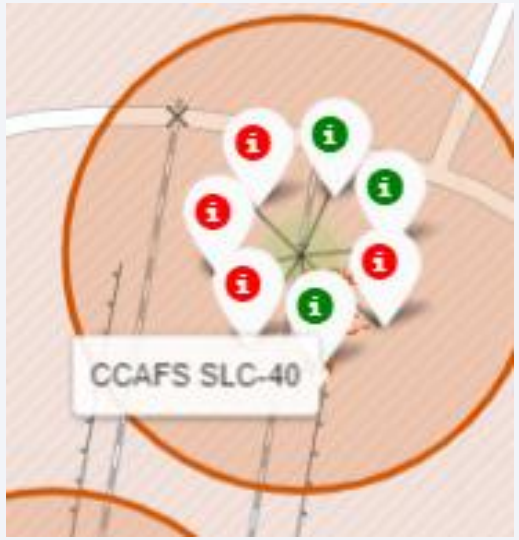
# Folium Map - Ground Stations


- Launch sites are located near the coast

# Folium Map - Launch site Outcomes

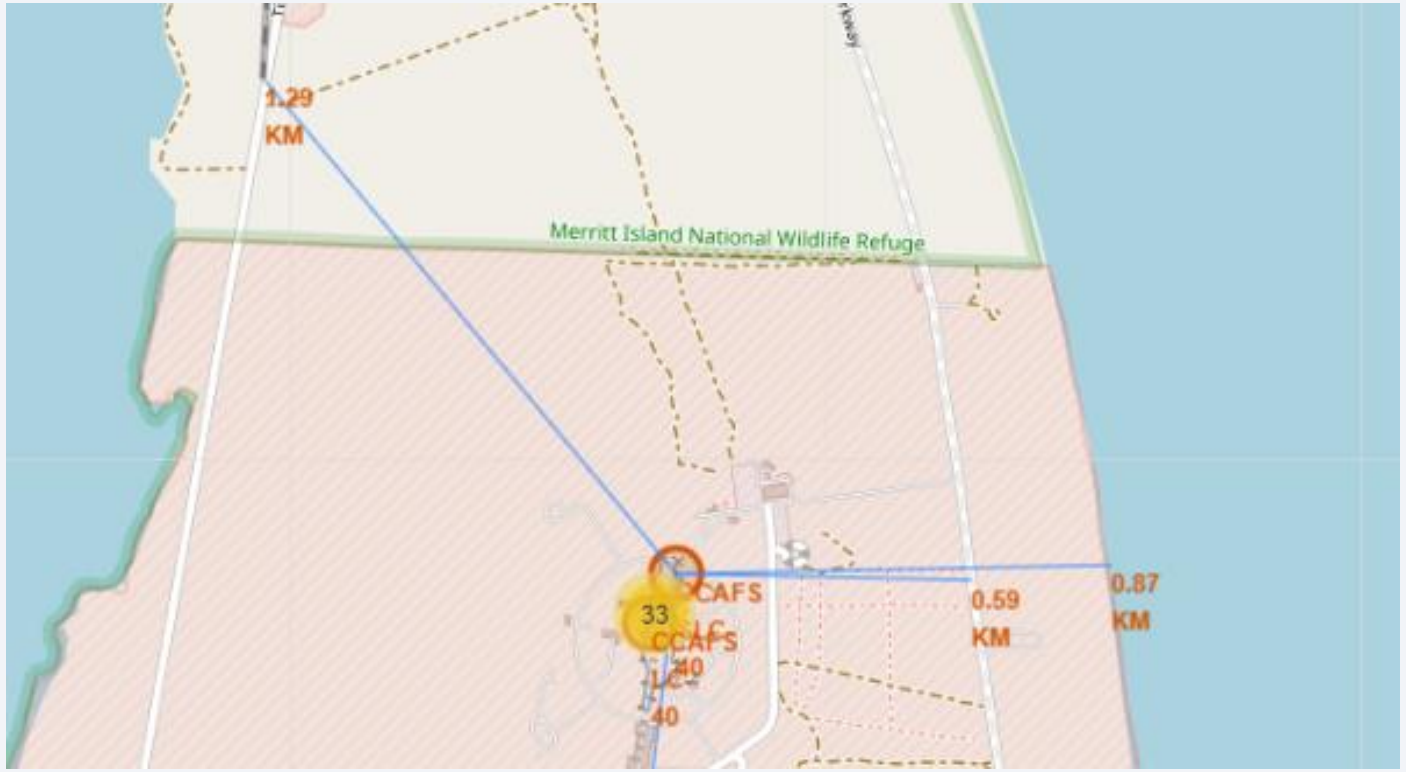



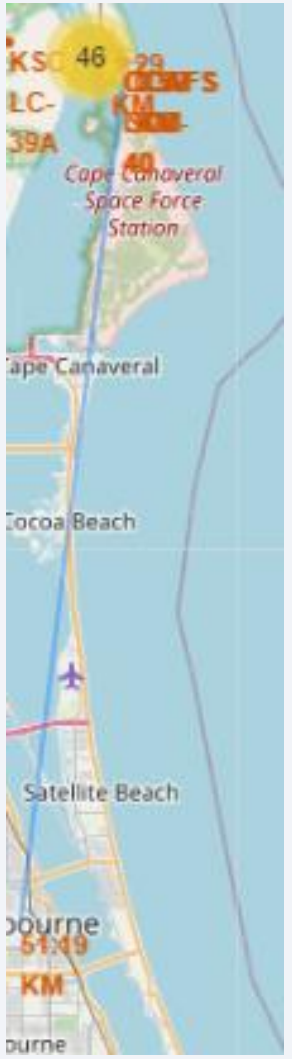



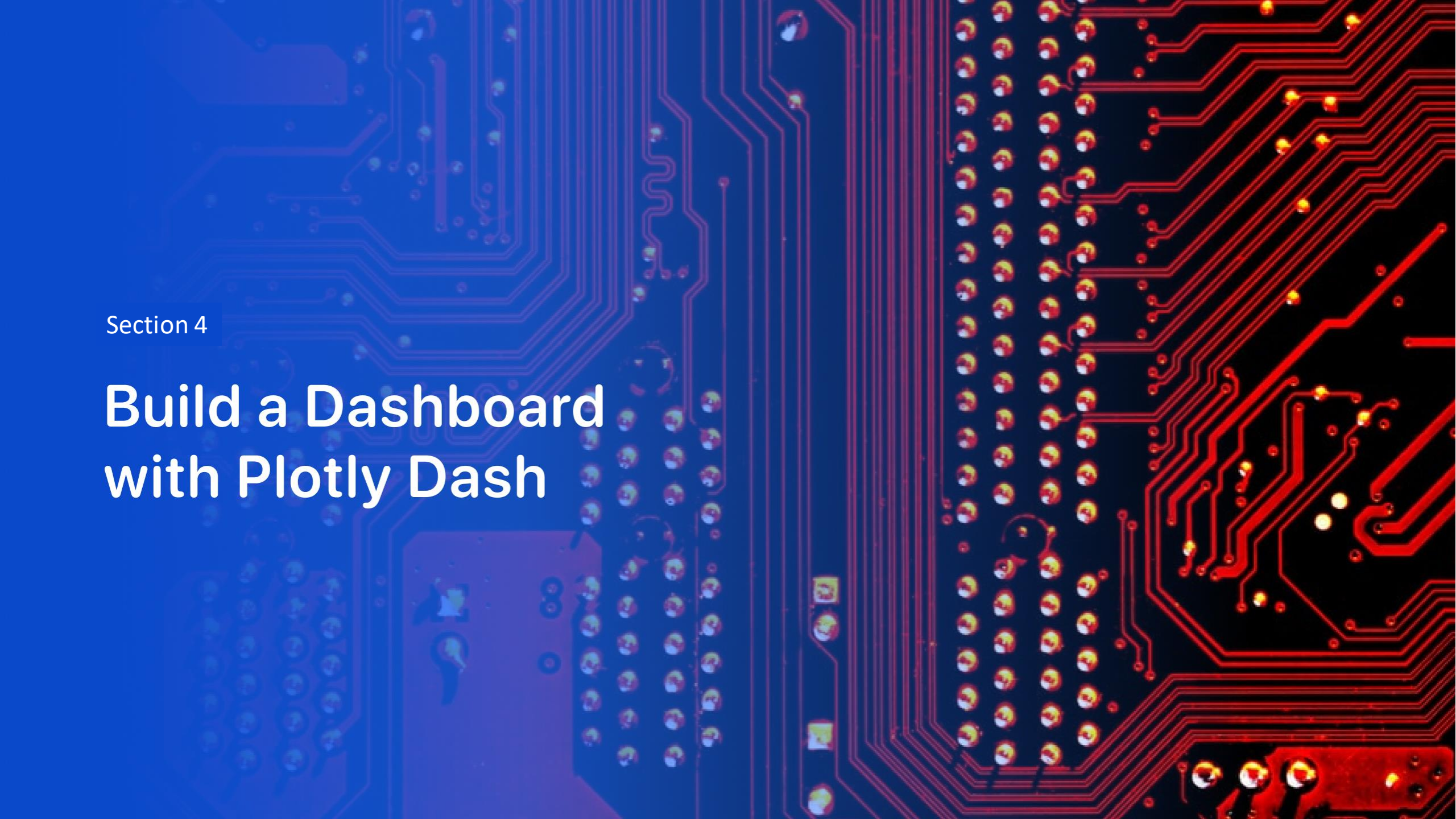

- The launch site with the best success rate is KSC

# Folium Map – Launch sites and proximity areas

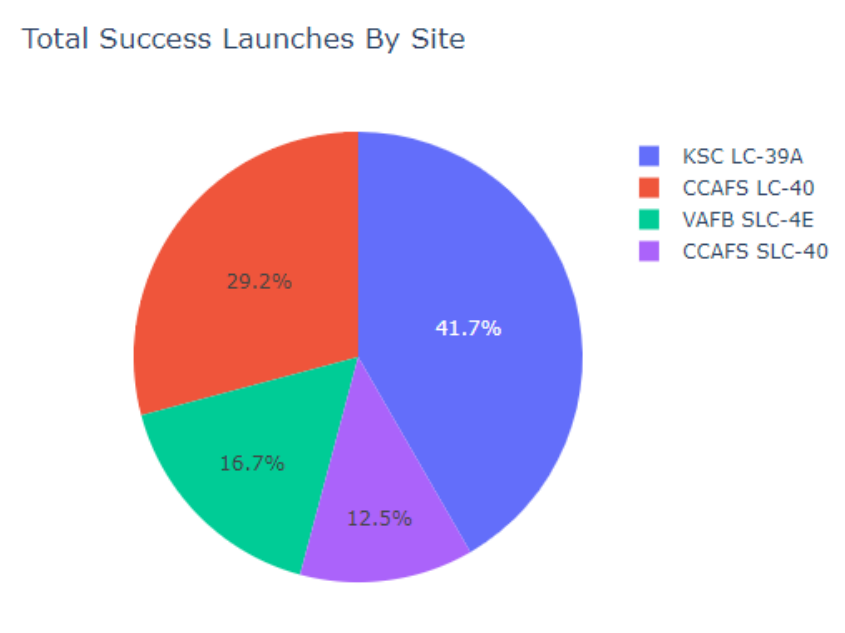



- Launch sites are located relatively close to roads and highways and far from inhabited areas.

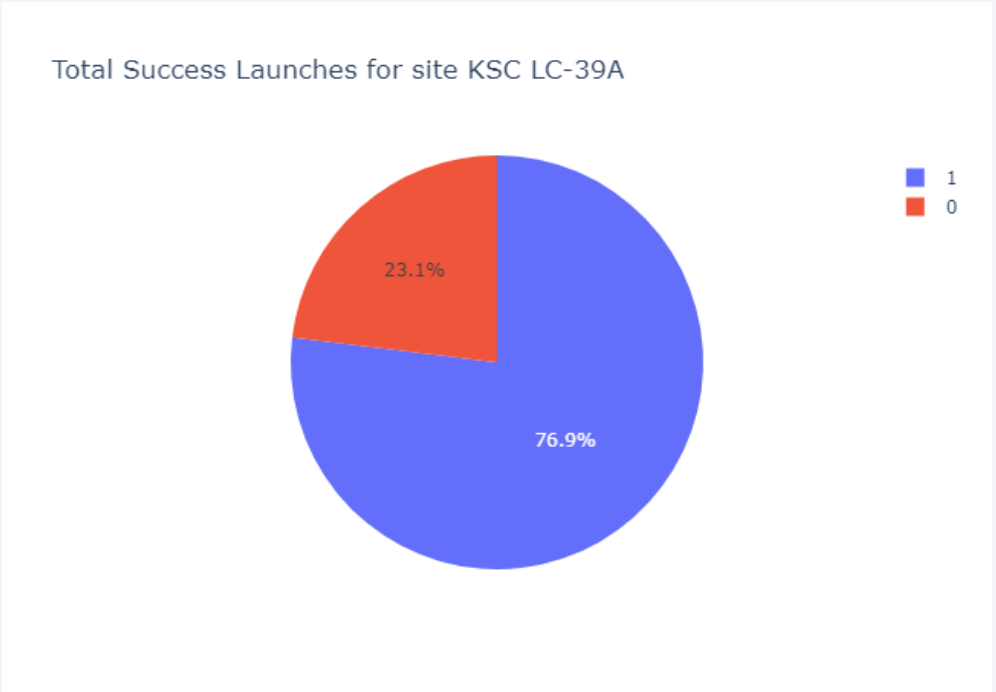

Section 4
Build a Dashboard with Plotly Dash

# Total Success Lauches By Site



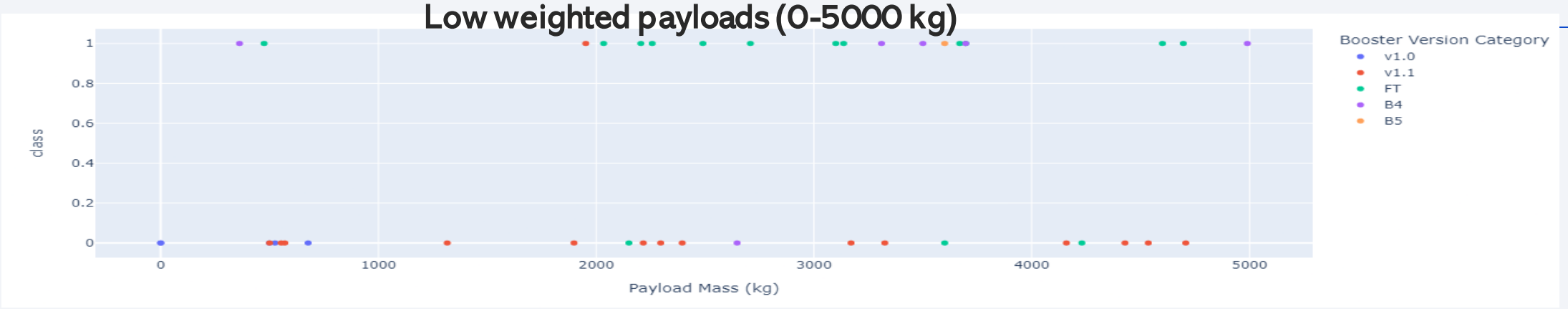

- The KSC center has the higher rate of successful launches, 41.7% of the total.
- On the other hand, CCAFS SLC-40 has the lower rate of successful launches, 12.5%.

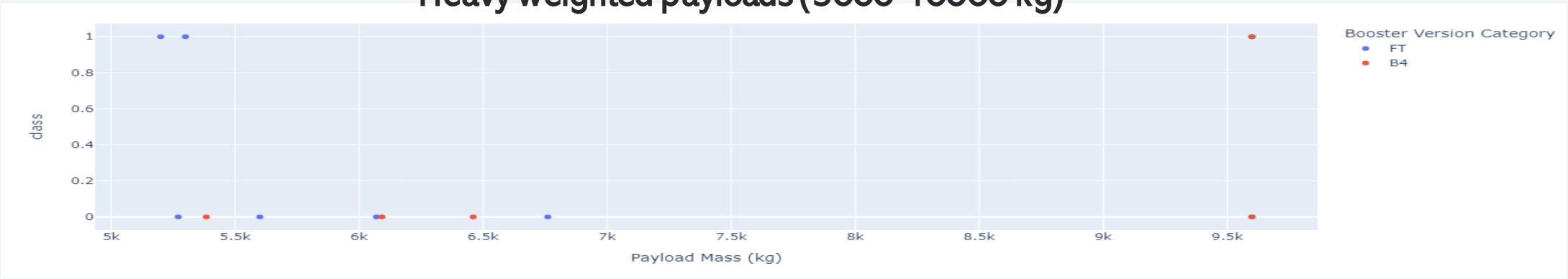
# Total Launches for site: KSC LC-39A




- KSC has the higher rate of successful launches, with a 76.6% of the total launches as successful ones, and 23.9% of them failed.

# Payload Mass vs. Outcome for all sites



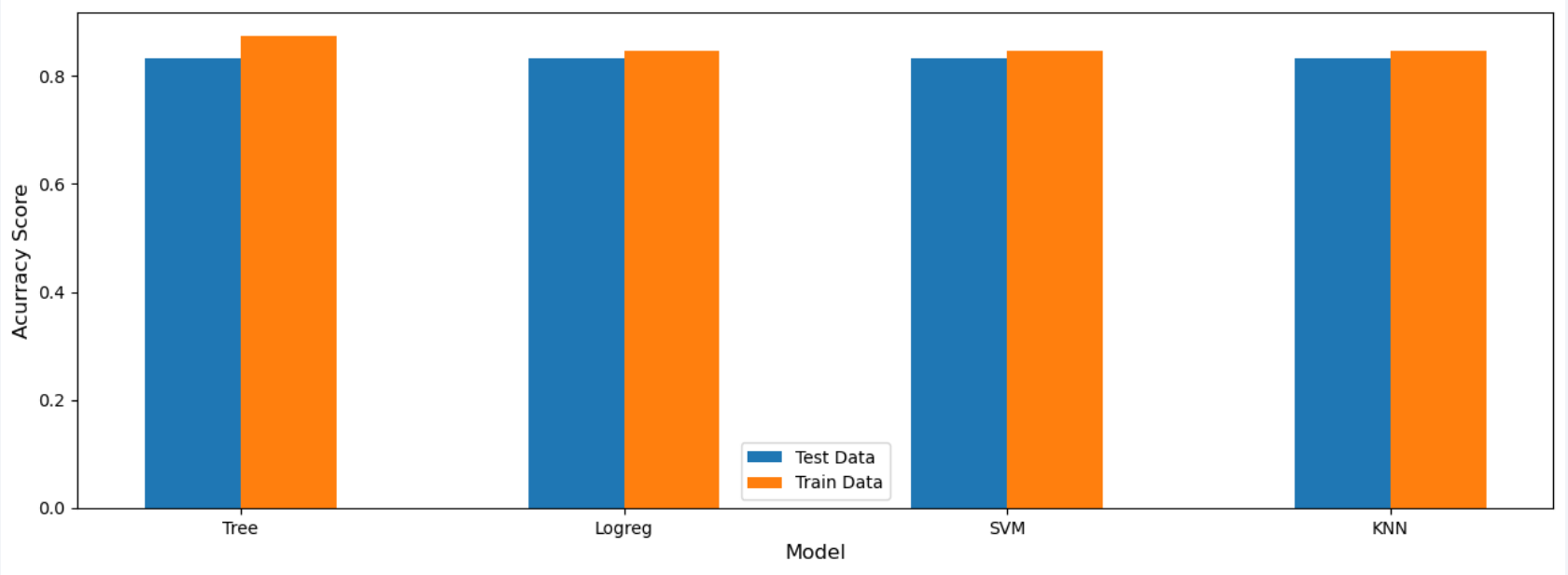

- There are fewer launches with heavy payloads, and there are only two types of booster versions for those launches.
- The success rate for the launches with heavy payloads seems to be lower than the launches made with lighter ones.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

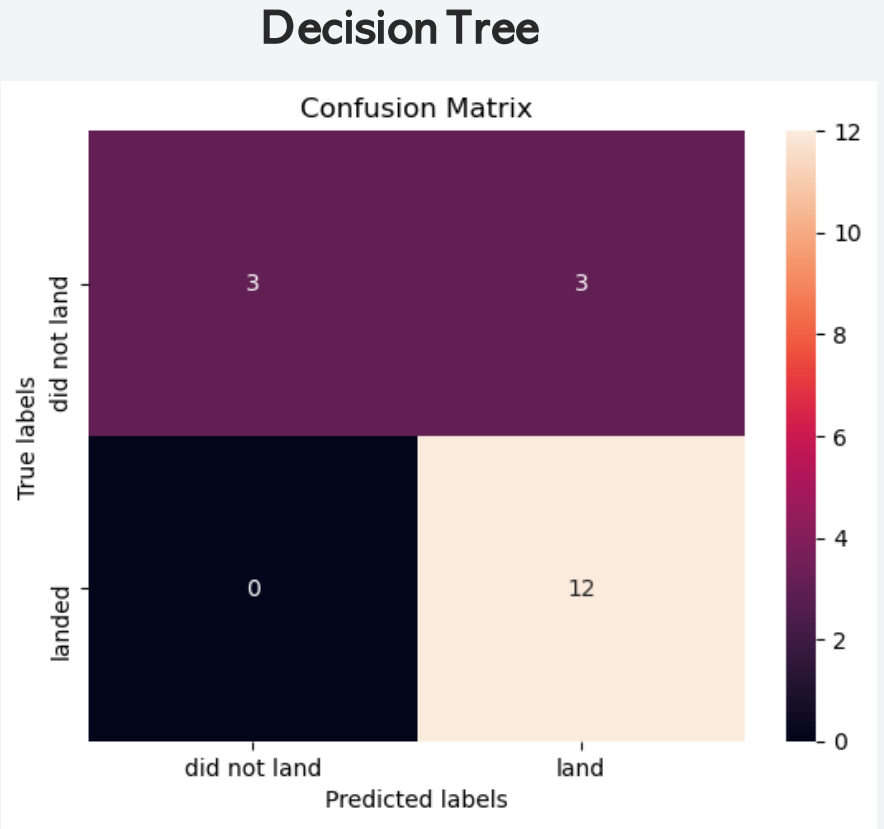
| | Accuracy Test | Accuracy Train |
|---|---|---|
| Tree | 0.833333 | 0.875000 |
| Logreg | 0.833333 | 0.847222 |
| SVM | 0.833333 | 0.847222 |
| KNN | 0.833333 | 0.847222 |




- All methods perform the same in test data. The only difference between them radicates on their performance with the train data.
- The accuracy of train data was used to select the Decision Tree algorithm as the classification method.

# Confusion Matrix

- The Confusion matrix shows how good the model performance is in the negatives, accurately predicting all.
- The main problem of this model is in the positives, wrongly predicting half of them as false positives.



# Conclusions

- The success of a mission can be explained by several factors, such as launch site, payload mass, orbit, and the number of previous flights.
- The orbits with the higher success rates are GEO, HEO, and ES-L1.
- The success rate tends to improve over time.
- The launch site with the better performance is KCS LC-39A.
- We choose the Decision Tree as a classification method for this dataset.

# Appendix

- Link to the GitHub repository with all the notebooks and datasets:
- https://github.com/aibamaya/Falcon_9_project

# Thank you!