IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Iyad Mahdy Ibrahim
21/9/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis
  - Interactive Visualizations
  - Predictive Analysis

- Summary of all results
  - EDA Results
  - Insights from Visualizations
  - Prediction Results

Introduction

- Project background and context
  - In this project, we analyzed SpaceX's Falcon 9 rocket launches, which cost $62 million—significantly lower than competitors charging up to $165 million. The key to this cost advantage is the reuse of the first stage. Using machine learning models, we predicted whether the first stage would successfully land based on historical data. Our findings will help Space Y, a new competitor, estimate launch costs and improve decision-making for future launches.
- Problems we want to find answers to
  - What factors influence the success of a first-stage landing?
  - How often does SpaceX successfully recover the first stage, and how does this trend over time?
  - What is the impact of mission-specific parameters (e.g., payload, orbit) on first-stage recovery?

Section 1

Methodology

Methodology

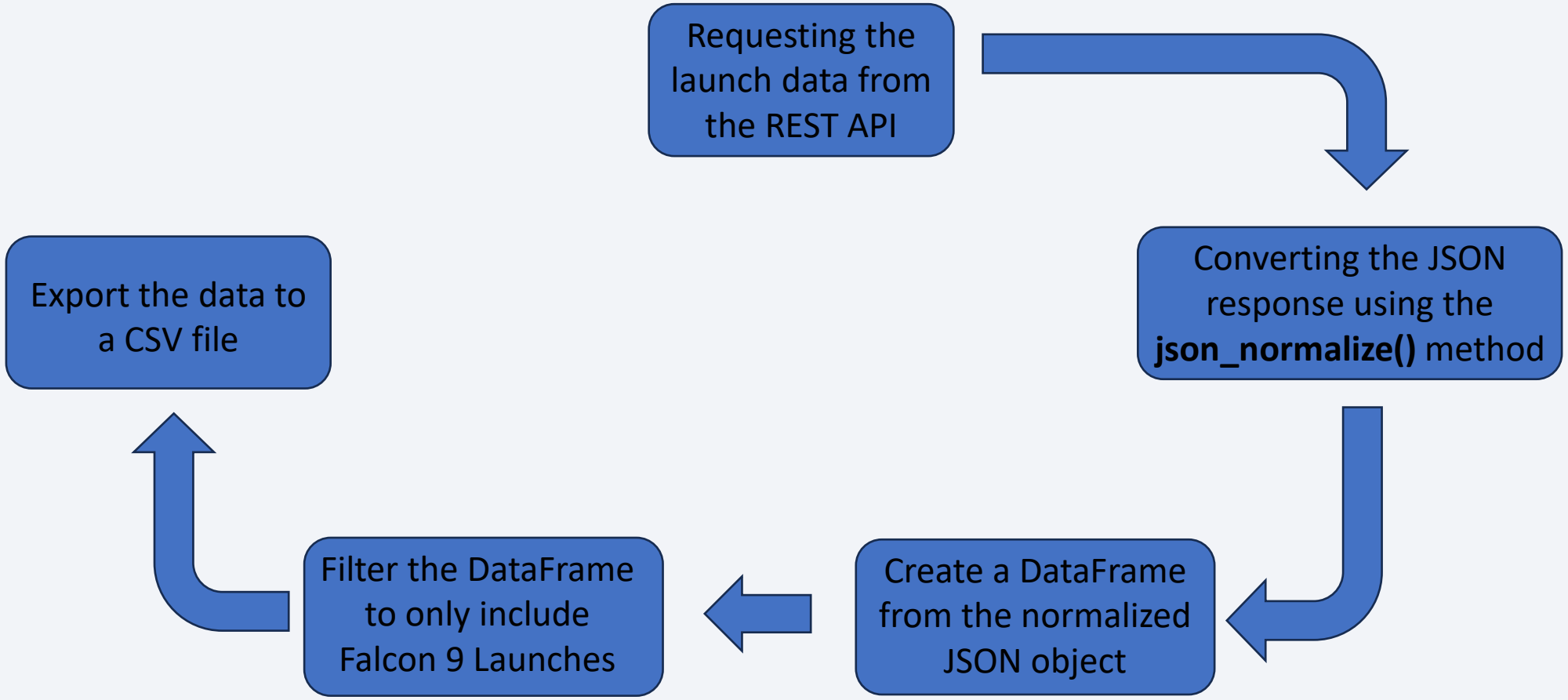
Executive Summary

- Data collection methodology:
  - SpaceX's REST API
  - Web Scraping
- Perform data wrangling
  - Handling Missing Values
  - One-Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluating classification models

Data Collection

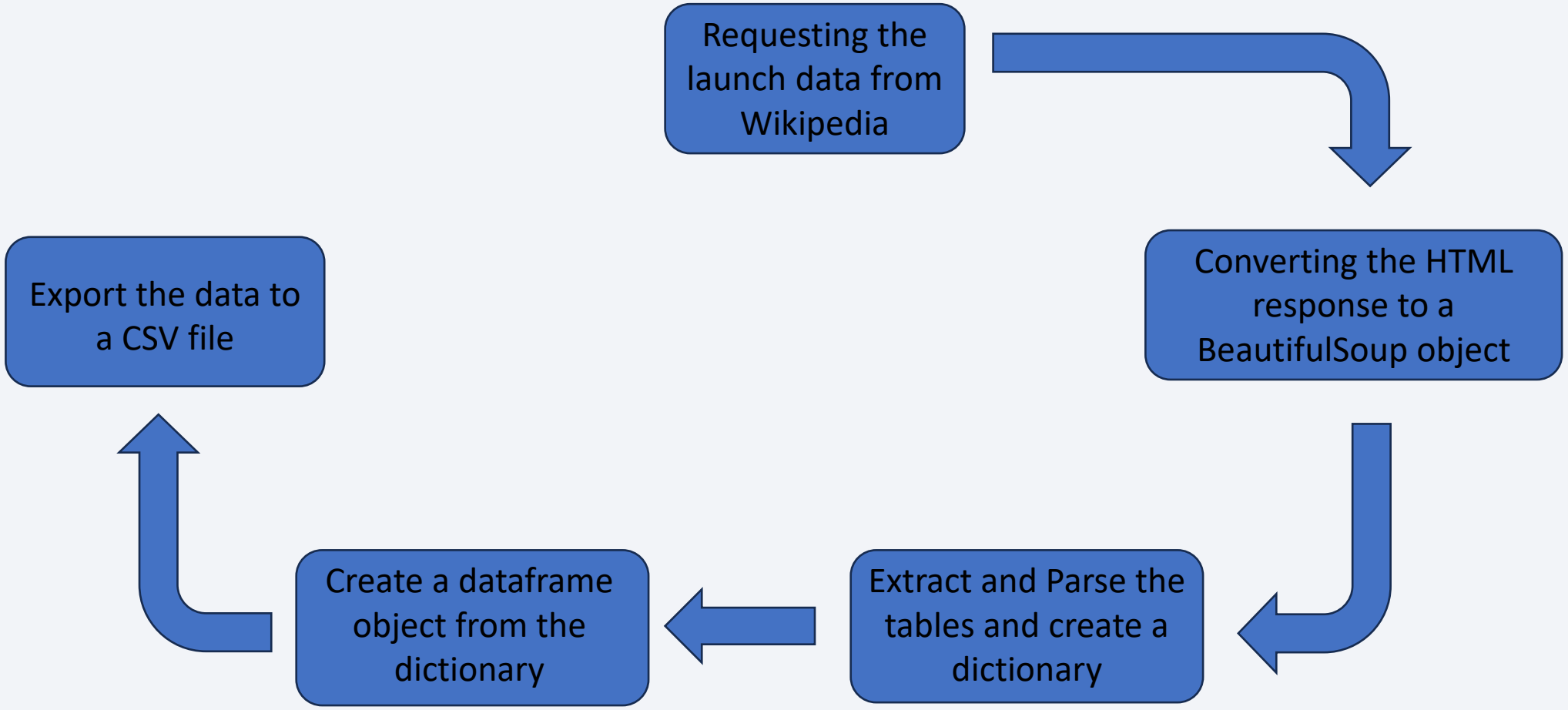
- Data collection for this project involved acquiring comprehensive SpaceX launch data to analyze rocket launches and predict landing success. The data collection process was divided into two main methods: using the SpaceX REST API and web scraping.
  - REST API:
    - The primary data source was the SpaceX REST API, specifically the /v4/launches/past endpoint. This API provided detailed information about past launches, including rocket types, payloads, launch specifications, and landing outcomes. The data, retrieved in JSON format, was converted into a structured table using the **json_normalize** function to facilitate analysis.
  - Web Scraping:
    - To supplement the API data, additional Falcon 9 launch records were collected through web scraping of HTML tables on Wikipedia using the BeautifulSoup package. This data was parsed and transformed into a Pandas DataFrame, providing a more comprehensive dataset for analysis and visualization.

Data Collection – SpaceX API



Data Collection – API Notebook

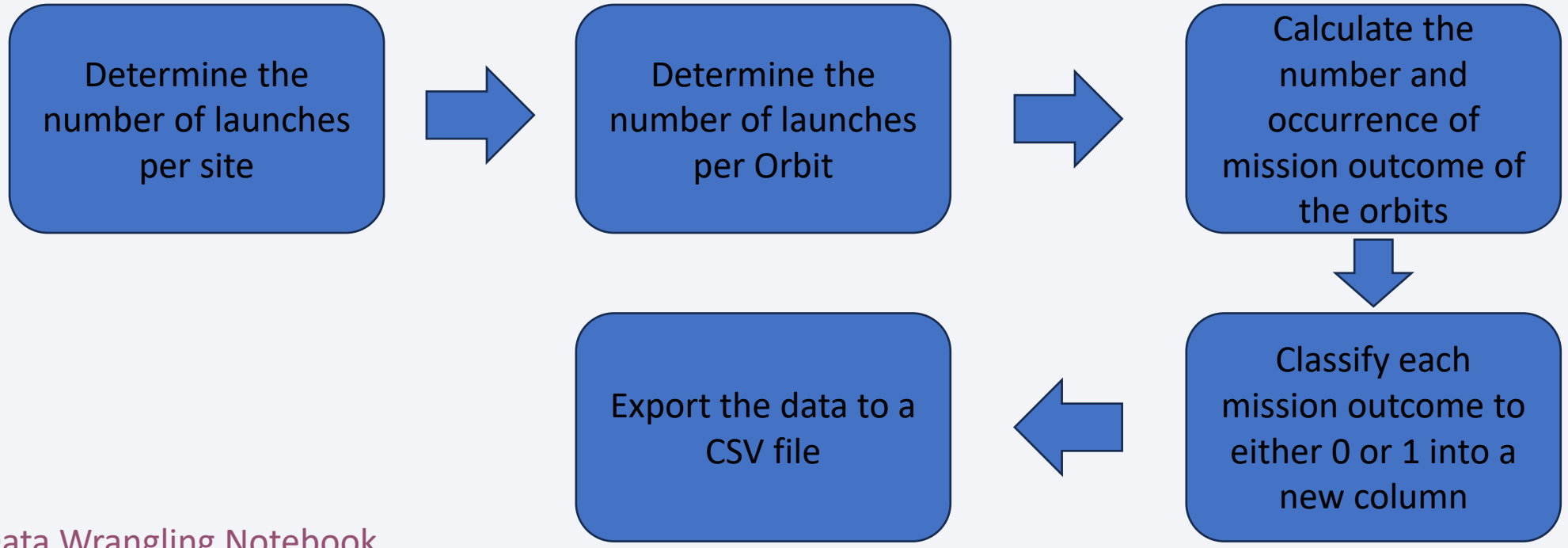
Data Collection - Scraping



Data Collection – Scraping Notebook

Data Wrangling

- The mission outcomes are represented as strings (e.g., True RTLS, False Ocean) that all signify either success or failure. Therefore, we need to encode those string variables into either 1 or 0.



Data Wrangling Notebook

EDA with Data Visualization

- Scatter Plots 
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Flight Number vs Orbit
  - Payload Mass vs Launch Site
  - Payload Mass vs Orbit

Scatter plots are used to visualize the relationship between two numerical variables.

- Bar Graph 
  - Orbit vs Success Rate

Bar graphs are used to compare the values of different categories.

- Line Plot 
  - Date vs Success Rate

Line plots are used to show trends or changes over time.

EDA – Data Visualization Notebook

EDA with SQL

- Using SQL queries, we:
  - Retrieved distinct launch sites.
  - Examined the first 5 launches from sites starting with 'CCA'.
  - Calculated the total payload mass for NASA (CRS).
  - Found the average payload mass for Booster Version 'F9 v1.1'.
  - Identified the first 5 dates with successful ground pad landings.
  - Listed Booster Versions with successful drone ship landings and payload mass between 4000 and 6000 kg.
  - Counted the occurrences of each mission outcome.
  - Identified distinct Booster Versions with the maximum payload mass.
  - Retrieved month, landing outcome, booster version, and launch site for failures on drone ships in 2015.
  - Counted occurrences of each landing outcome and ordered them by count.

EDA – SQL Notebook

Build an Interactive Map with Folium

- The Folium Map is centered on NASA Johnson Space Center at Houston, Texas.
- Red circles are added on the map at the location of each launch site, with a marker that indicates their respective names.
- Each launch site has been assigned a marker cluster to display multiple information for the same coordinate.
- Each marker cluster is assigned a green marker if it held a successful landing, or a red marker otherwise.

Interactive Visualization – Folium Notebook

Build a Dashboard with Plotly Dash

The Dashboard contains Interactive and Graphical Components

Interactive Components:

- A Drop-down menu for filtering based on the desired Launch Site
- A Range Slider to select the Payload Range

Graphical Components:

- A Pie Chart showing the success rate for the selected Launch Site.
- A Scatter Plot of success rate vs payload mass.

This interactivity was made possible with callback functions, applying filtering of the dataset based on the values of the input components

Interactive Visualization – Dash Code

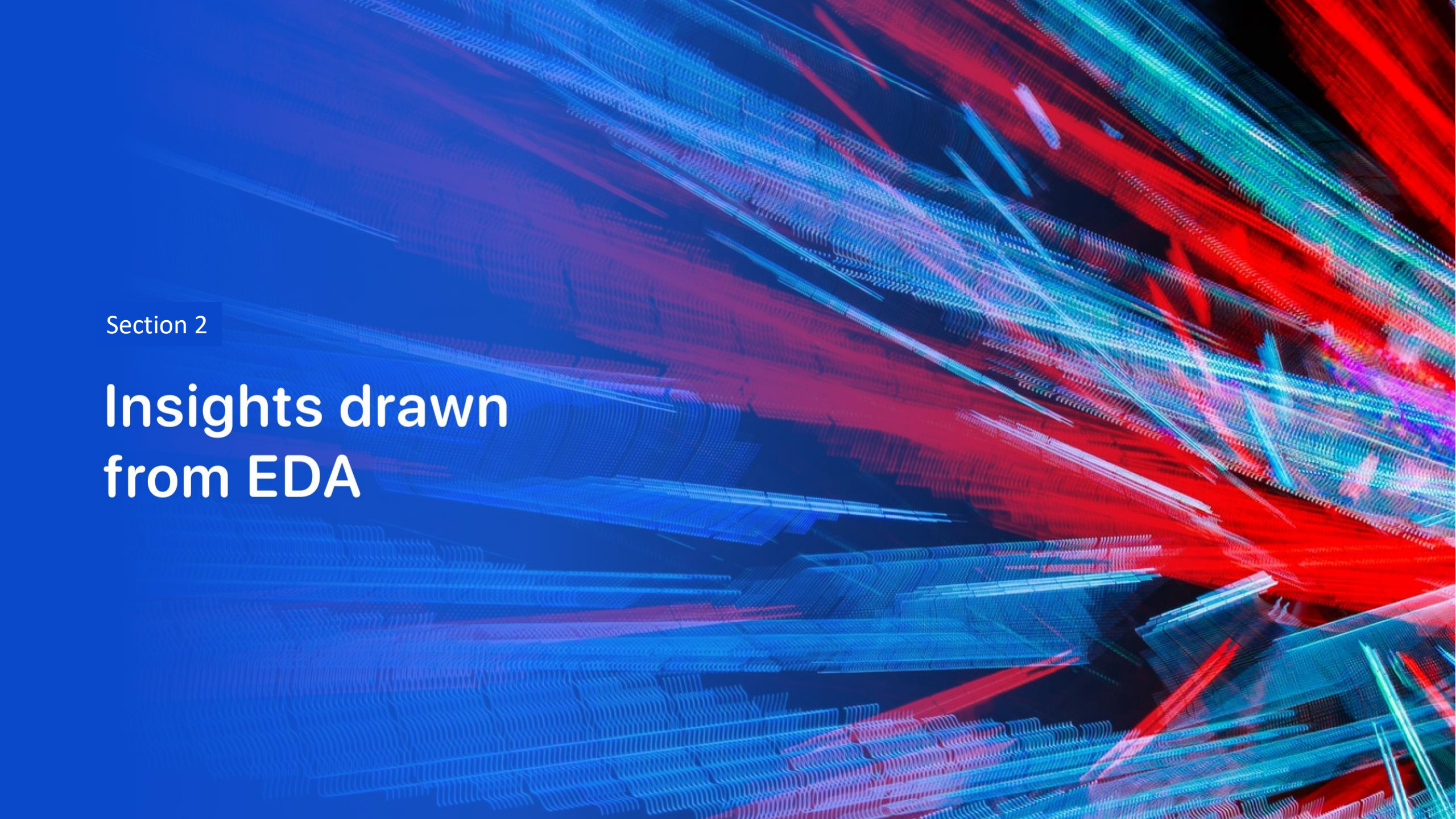
Predictive Analysis (Classification)

- Data Preparation
  - Loading the data
  - Standardizing the features
  - Splitting the data into training and testing sets
- Model Preparation
  - Choosing the machine learning models
  - Selecting the optimal parameters using Grid Search
  - Training the model with the optimal parameters using Cross Validation on the train set
- Model Evaluation
  - Compute each model's accuracy
  - Plot the Confusion Matrix
- Model Comparison
  - Select the model with the highest accuracy

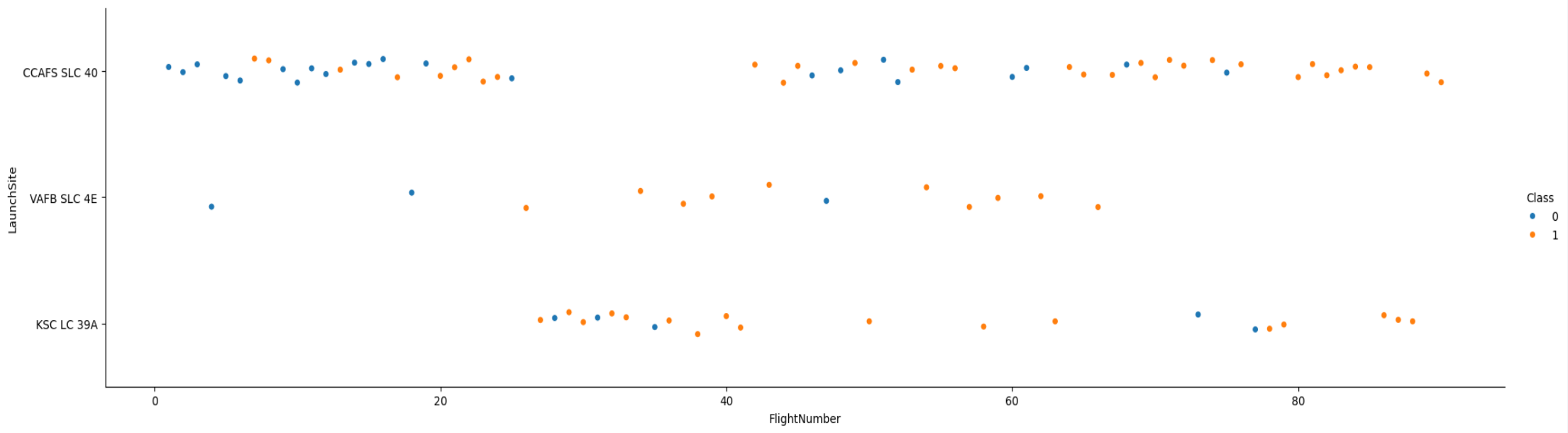
Machine Learning Notebook

Results

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

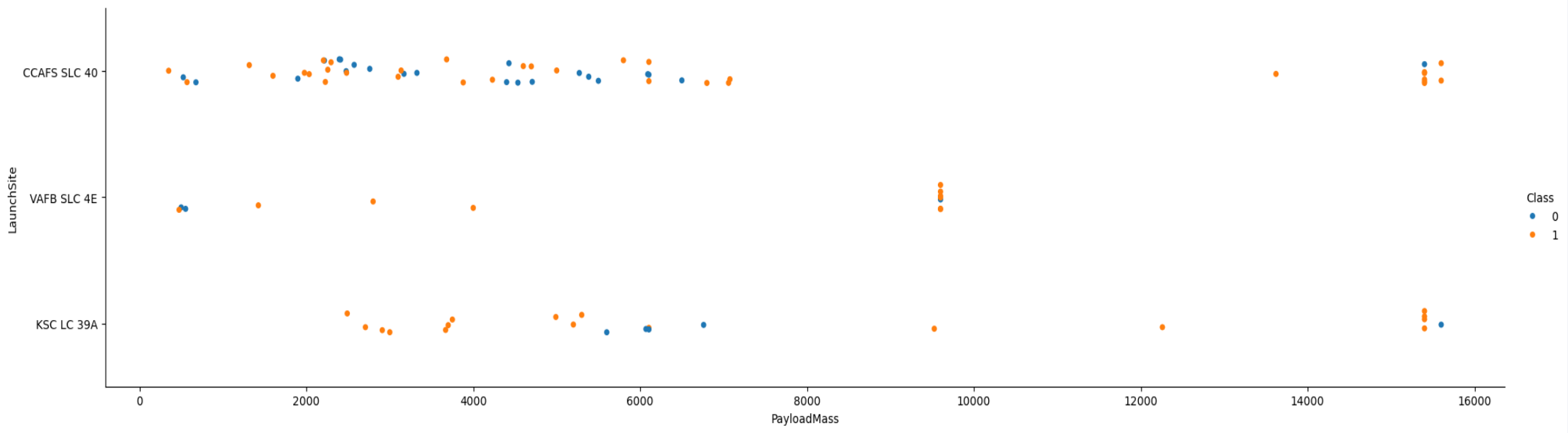

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



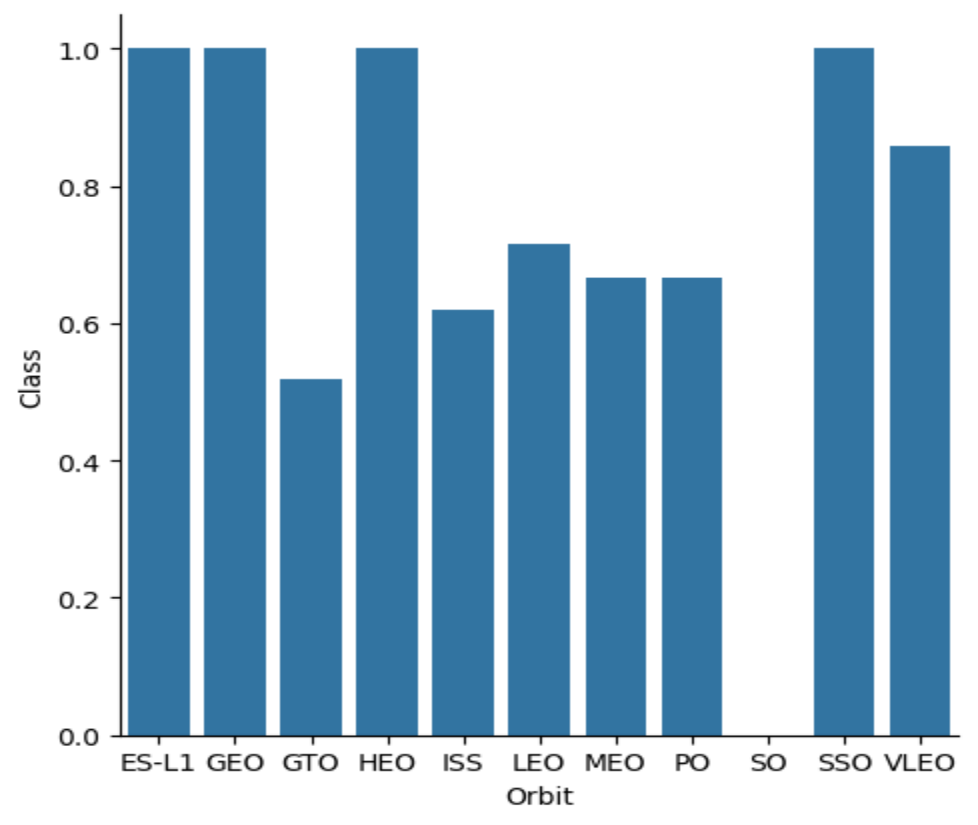
- Most of the attempts at **VAFB SLC 4E** and **KSC LC 39A** were successful
- The **CCAFS SLC 40** site is the least successful site, although it shows some improvement

Payload vs. Launch Site



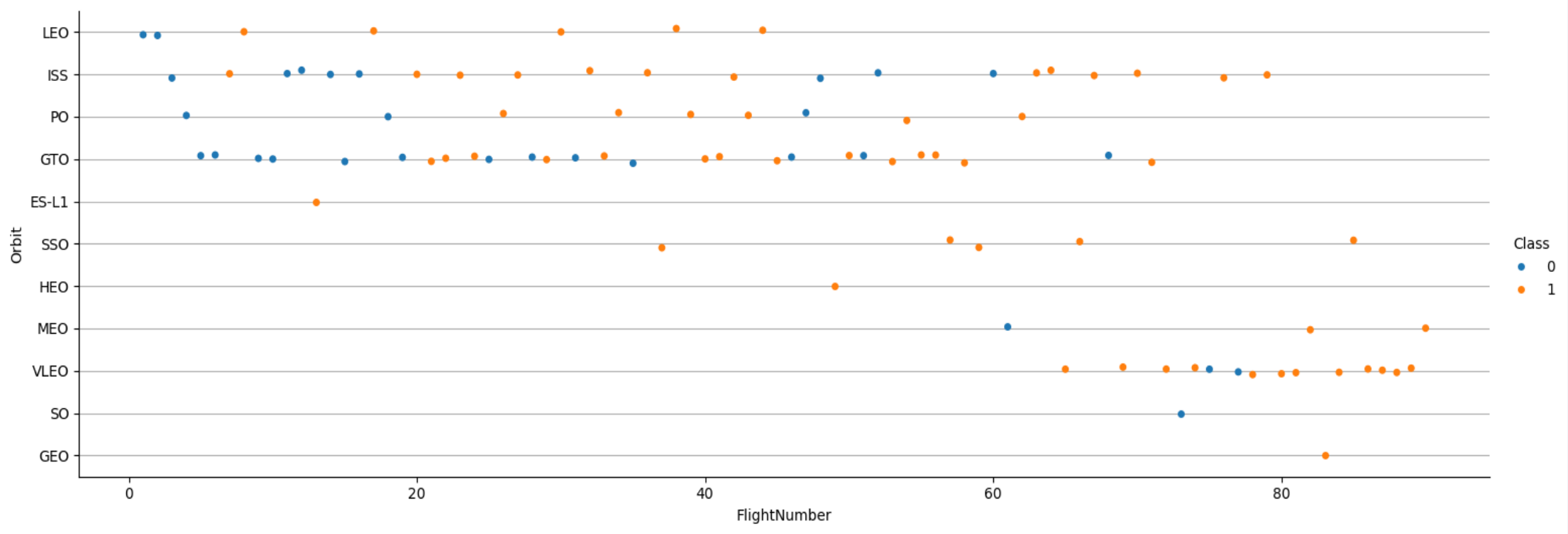
Launches with low payload masses are more common, however, higher payload masses show more success rates

Success Rate vs. Orbit Type



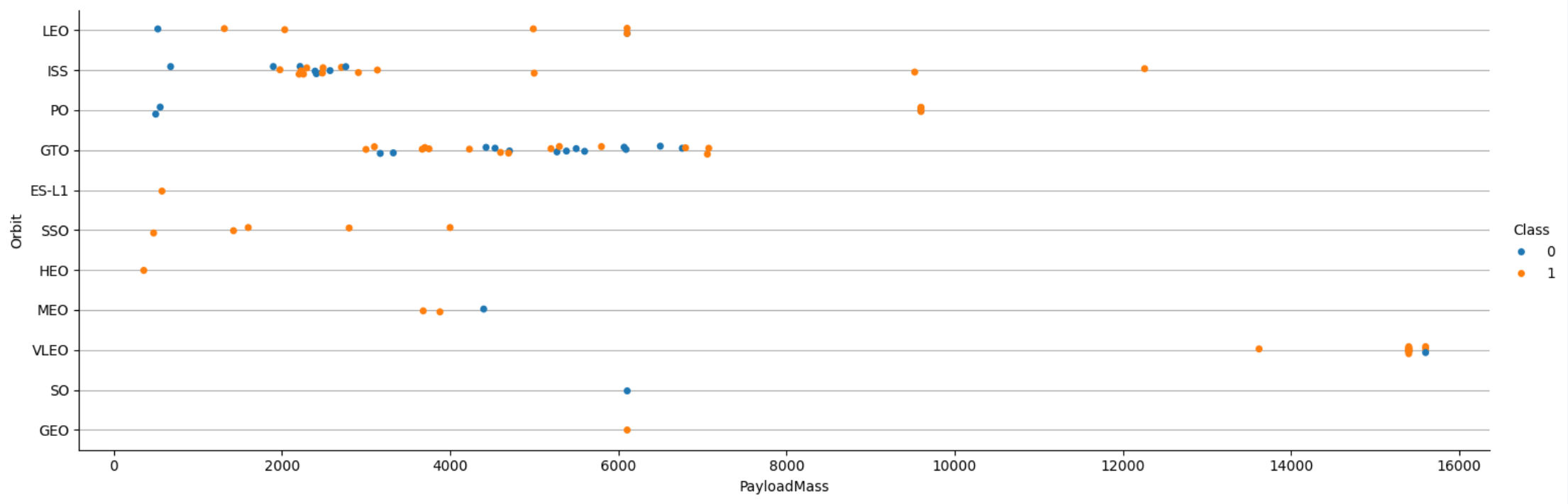
- **ES-L1**, **GEO**, **HEO** and **SSO** orbits have been 100% successful
- **VLEO** orbit has a high success rate
- **SO** orbit hasn't been successful at all

Flight Number vs. Orbit Type



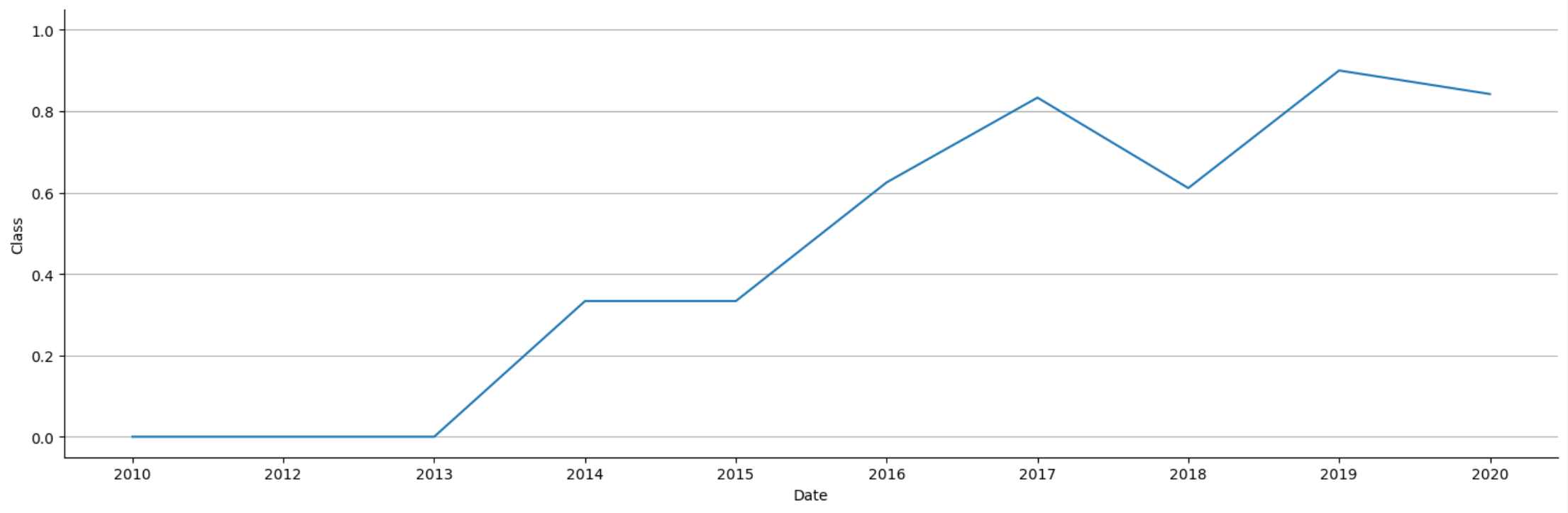
- For some orbits, like **SO**, **GEO**, **HEO** and **ES-L1**, we shouldn't consider their success rates due to their low flight counts
- **LEO** orbit's success rate seems to have improved, while **GTO** orbit doesn't look like it is improving

Payload vs. Orbit Type



SSO orbit is used for relatively low payload mass, while **VLEO** is used for high payload masses

Launch Success Yearly Trend



Success rates have been increasing after 2013, with a slight decrease in 2018

All Launch Site Names

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

These are all the Launch Sites available in the dataset

Launch Site Names Begin with 'CCA'

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

5 Records of Launch Sites with names beginning with ‘CCA’

Total Payload Mass

Total_Payload_Mass
45596

The total Payload Mass recorded in the dataset is 45,596 Kg

Average Payload Mass by F9 v1.1

AVG_PAYLOAD_F9
2928.4

The Average Payload Mass in Kilograms for the Flights containing F9 v1.1 Boosters

First Successful Ground Landing Date

Date
2015-12-22

The date of the first successful landing in ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

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

The Names of the boosters which have success in Drone Ship and have Payload Mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

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

The total number of successful and failed mission outcomes

Boosters Carried Maximum Payload

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

The Booster Versions that Carried the Maximum Payload

2015 Launch Records

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

The Failed Landing Outcomes in the year 2015

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

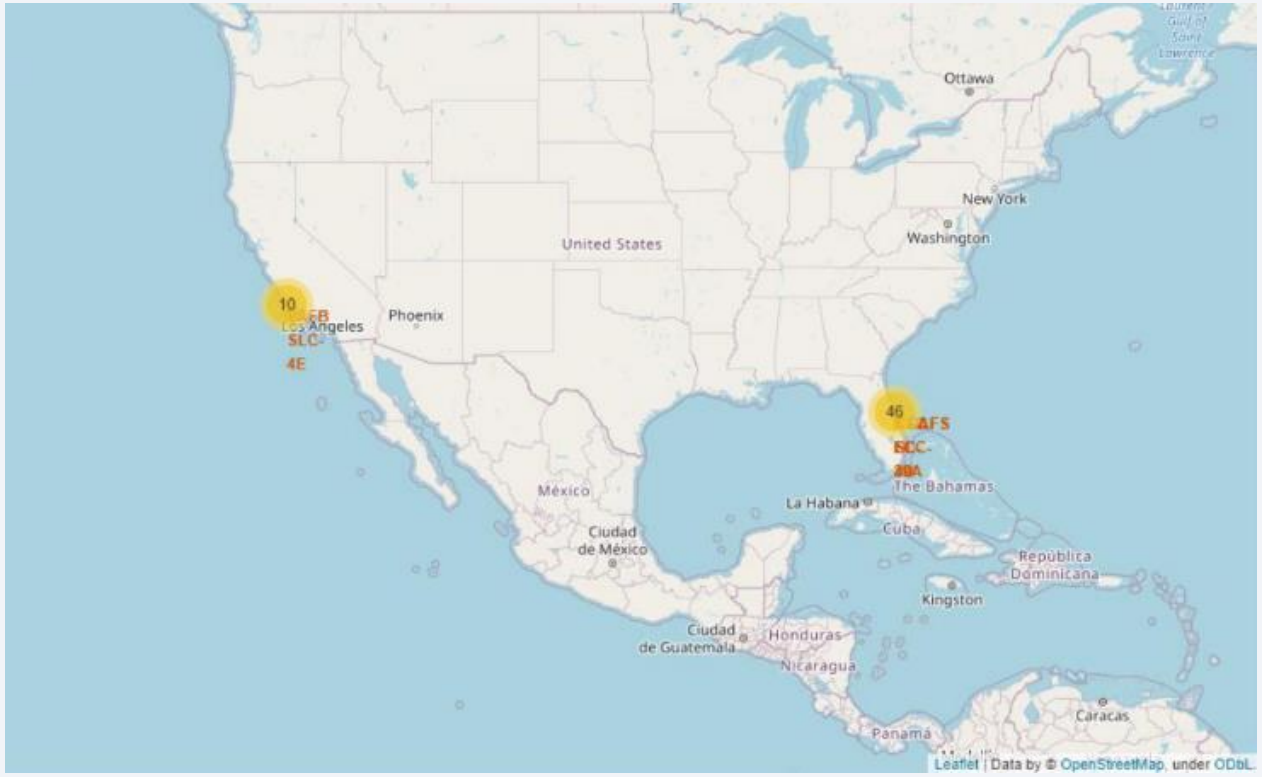
Landing_Outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

The count of the landing outcomes between 2010-06-04 and 2017-03-20, ordered by the count.

Section 3

Launch Sites Proximities Analysis

Launch Sites on the Map

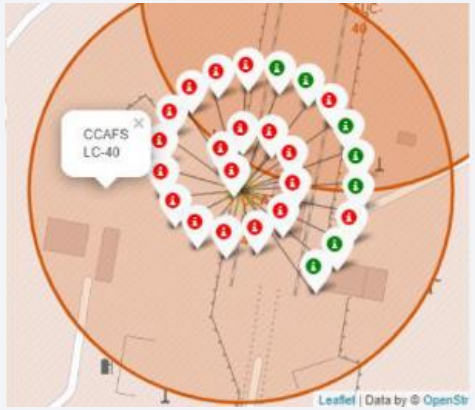


We can see that the launch sites are located on the coasts of the USA

Success Rates for Two Sites

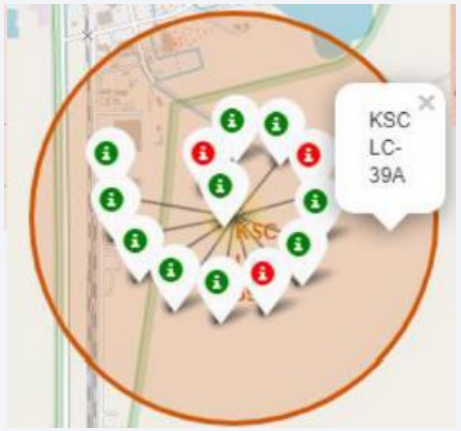
For each site, green markers represent successful landings while red markers represent failed landings

CCAFS LC-40

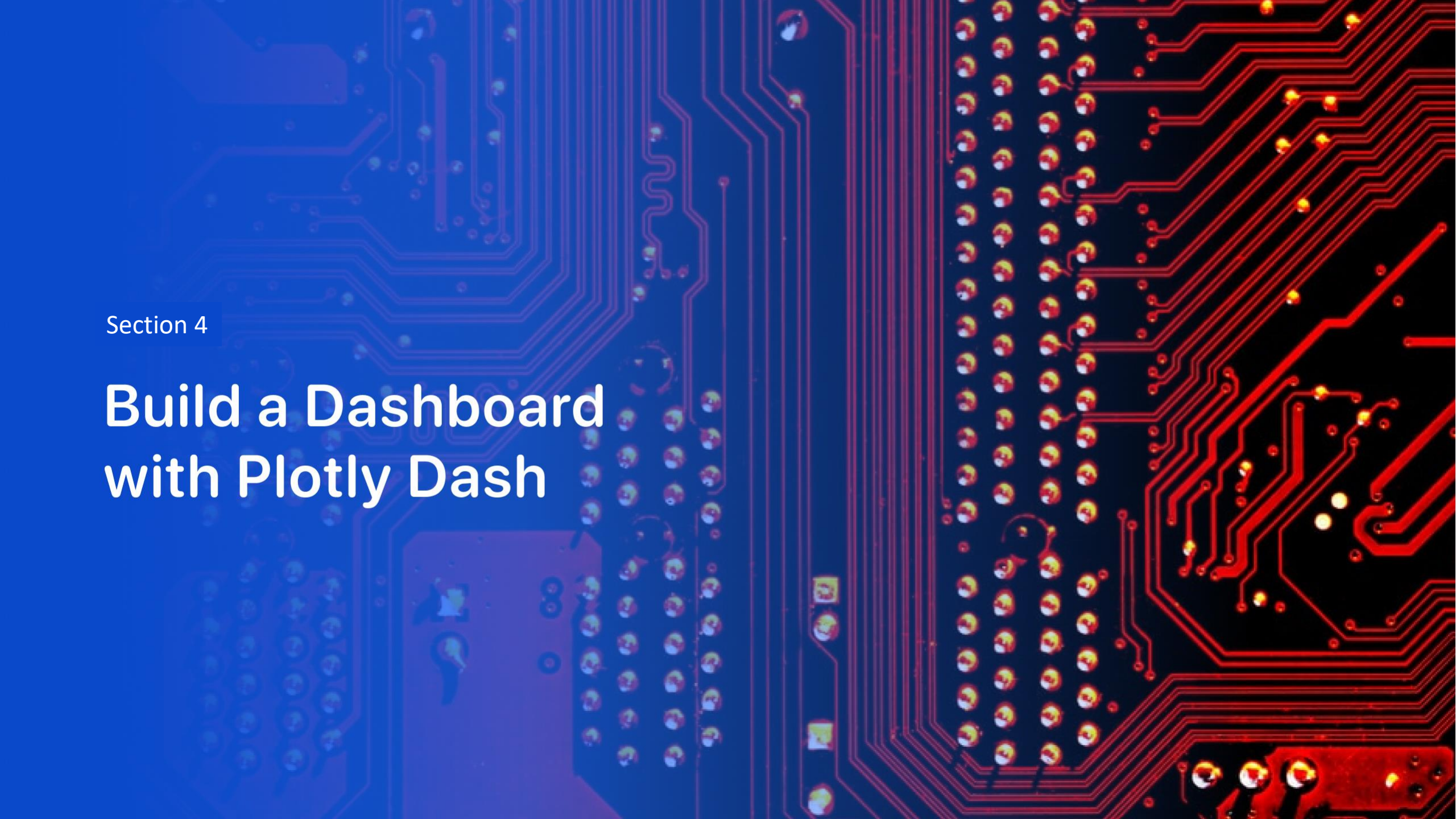


Low Success Rate

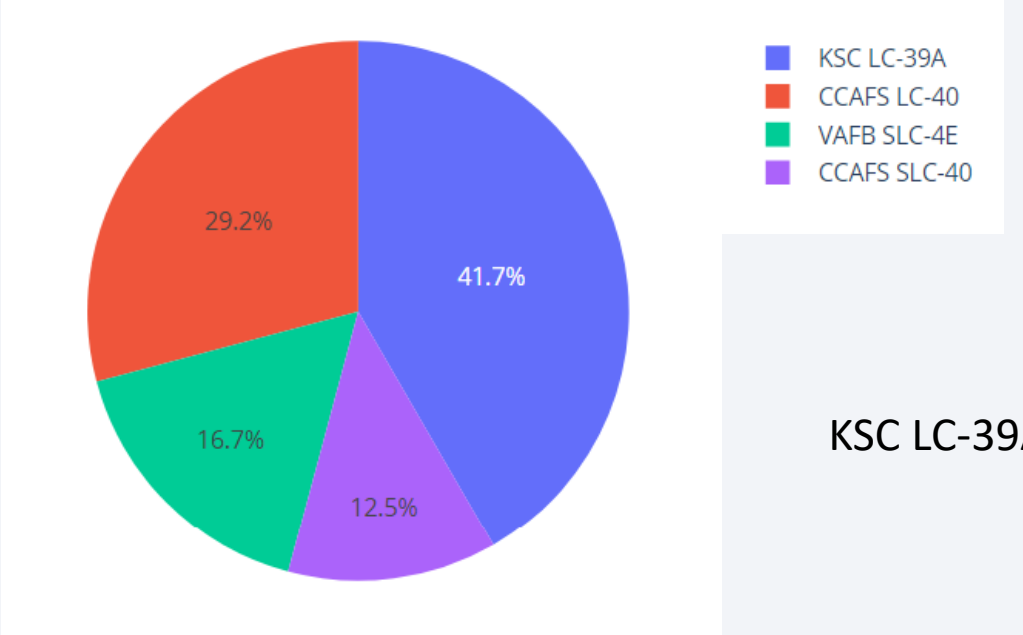
KSC LC-39A



High Success Rate

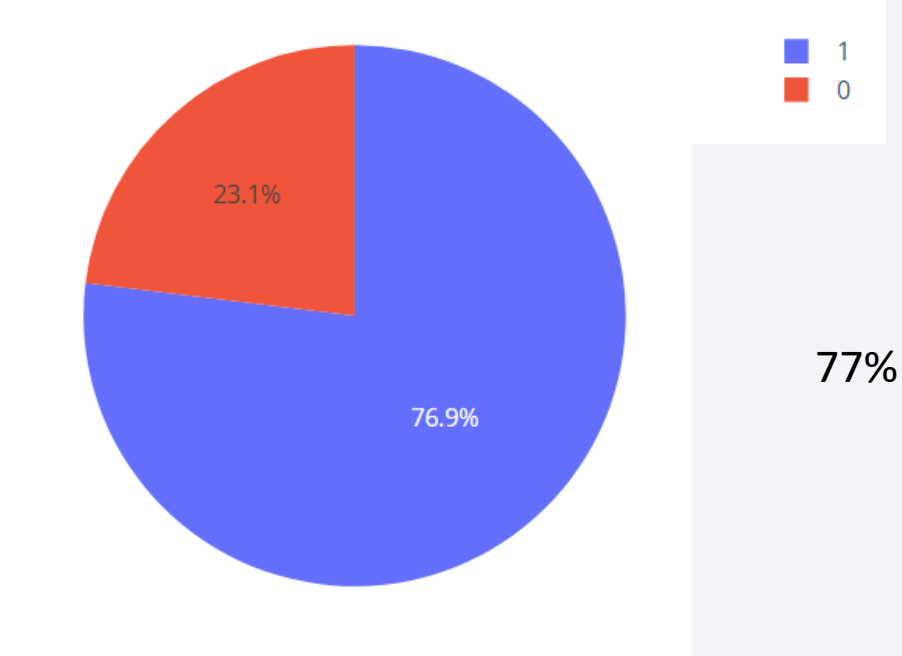

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches per Site



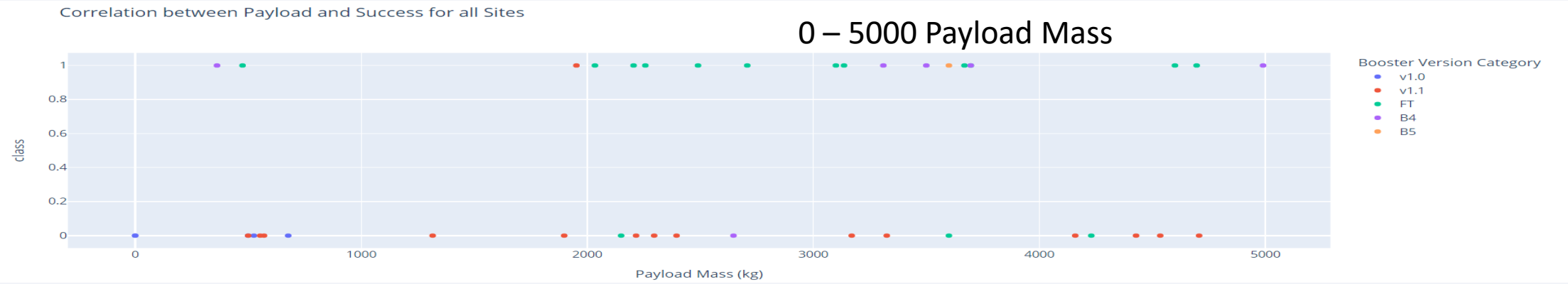
KSC LC-39A is the site with the most successful launches

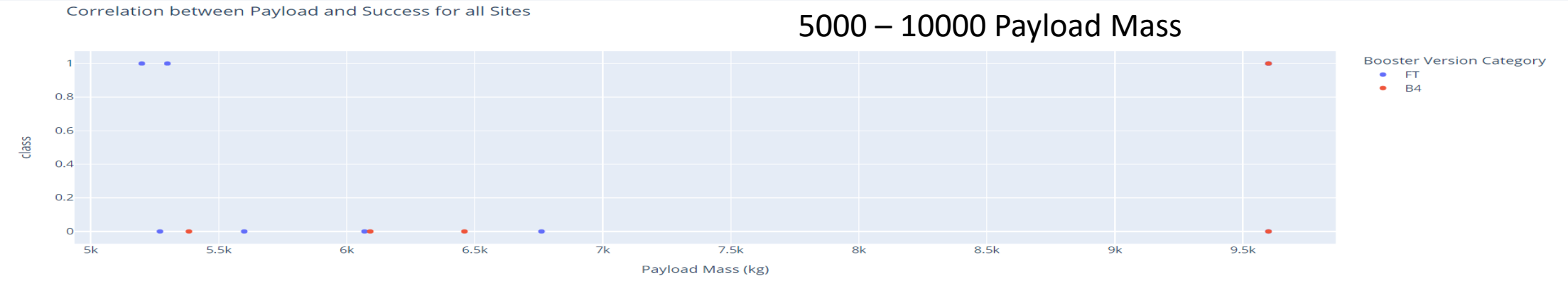
Total Successful Launches for KSC LC-39A



77% of the launches in KSC LC-39A were successful

Successful Launches in a range of Payload Mass





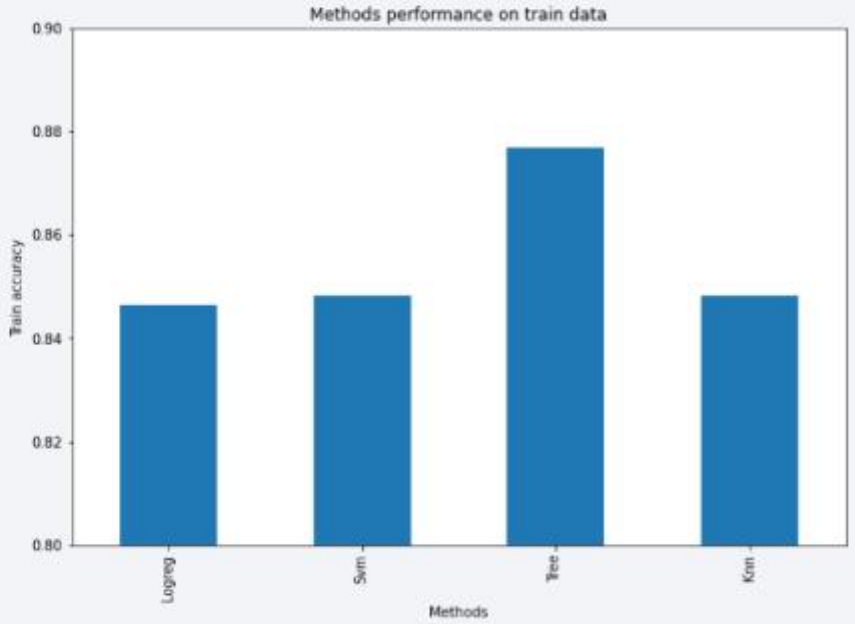
There are more Flights in lower Payload Masses, therefore more Success

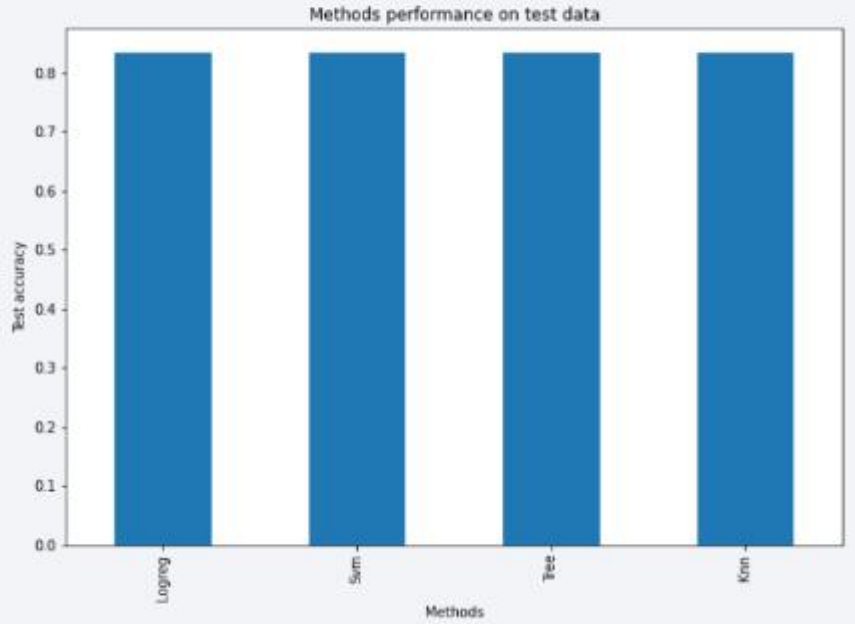
Section 5

Predictive Analysis (Classification)

Classification Accuracy

	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333





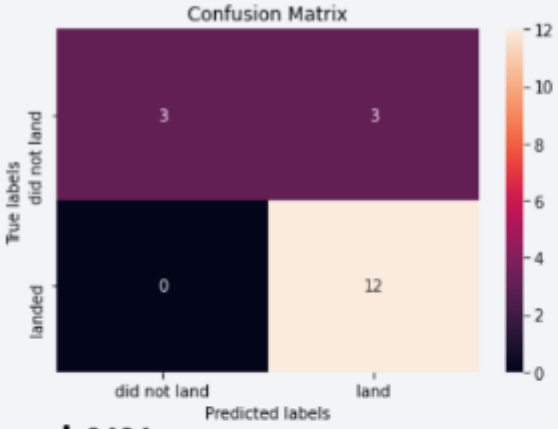
For accuracy test, all methods performed similar. We could get more test data to decide between them. But if we really need to choose one right now, we would take the decision tree.

Decision tree best parameters

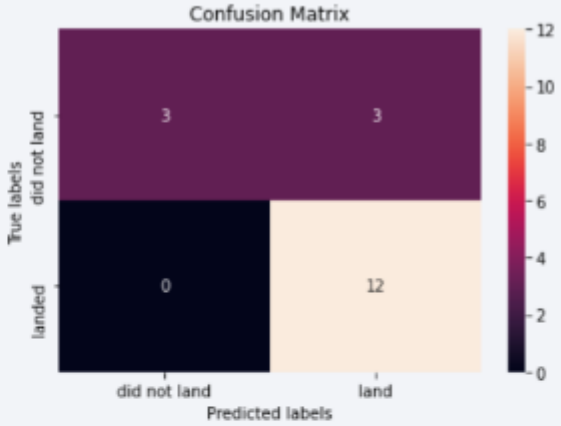
```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 2, 'splitter': 'random'}
```

Confusion Matrix

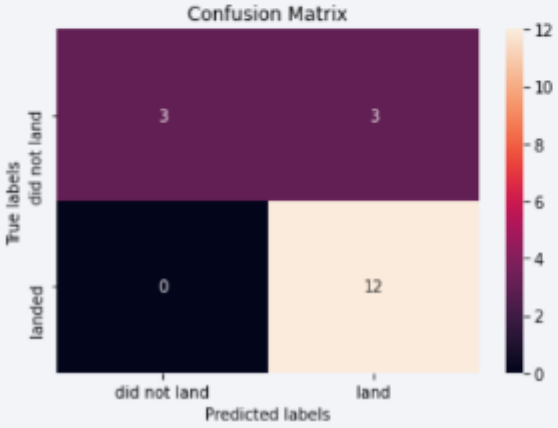
Logistic regression



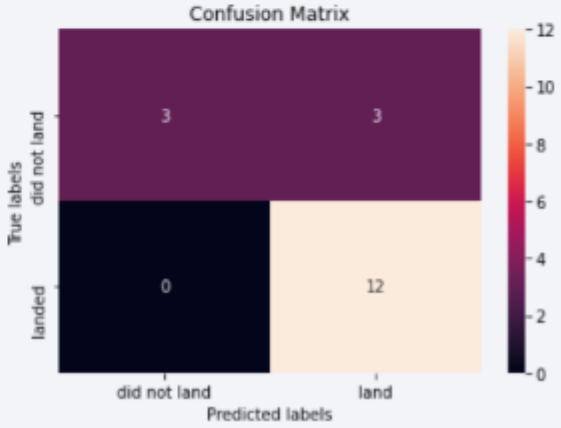
Decision Tree



kNN



SVM



As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.

		Actual values 1	Actual values 0
Predicted values	1	TP	FP
	0	FN	TN

Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

Thank you!