IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

KHOR WEI GENE

October 10, 2021

Outline

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

Executive Summary

- **Summary of methodologies:**
  - ✓ Data collection
  - ✓ Data wrangling
  - ✓ Exploratory Data Analysis (EDA) with Data Visualization
  - ✓ EDA with SQL
  - ✓ Building interactive maps with Folium
  - ✓ Building dashboard with Plotly Dash
  - ✓ Predictive analysis with Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbor (KNN)
- **Summary of all results**
  - ✓ EDA results
  - ✓ Interactive analytics demo in screenshots
  - ✓ Predictive analysis results

Introduction

- **Project background and context**

SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars on its website, which is a whopping 103 million dollars below the cost reported by its competitors (165 million dollars). According to SpaceX, this substantial cost reduction comes from the reusability of rocket's first stage. Therefore, if we can determine whether the rocket's first stage will fly back to Earth safely or not, we can determine the cost of a launch. This project applies data science using Python programming to predict the returnability of the rocket's first stage to predict the cost of a launch.

- **Problems to address**

i. How variables like payload mass, launch site, number of flights, and destinated orbits affect the success of the first stage landing?

ii. Does the rate of successful landings increase over the years?

iii. What is the best machine learning algorithm that can be used for the binary classification in this data science project?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX REST API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Organizing the data
  - Dealing with missing values
  - Using One-Hot Encoding to convert categorical value into binary value
- Perform EDA using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology (Continued)

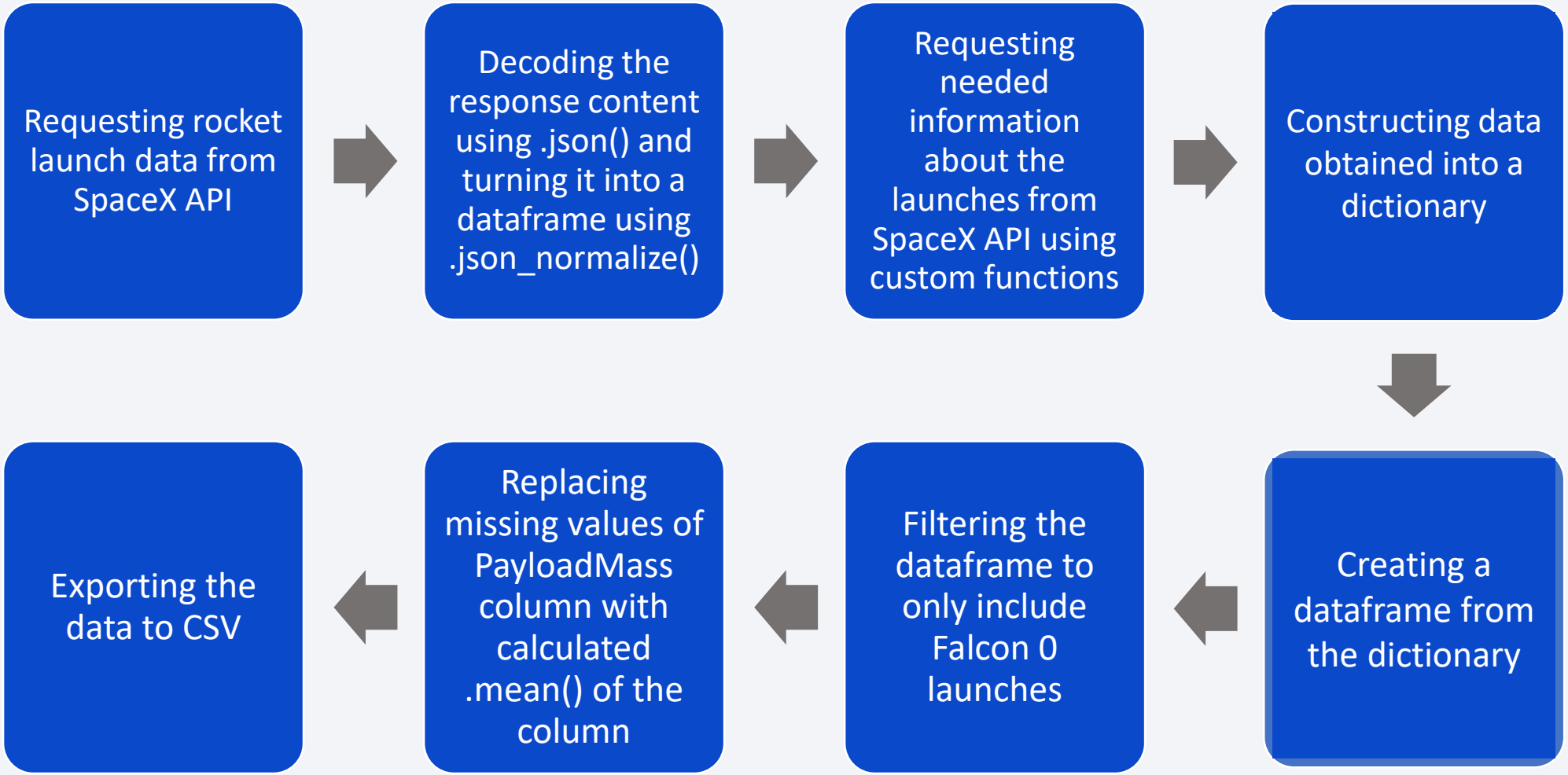
Executive Summary (Continued)

- Perform predictive analysis using classification models
    - Building, tuning, and evaluating classification models to ensure the highest accuracy.

Data Collection

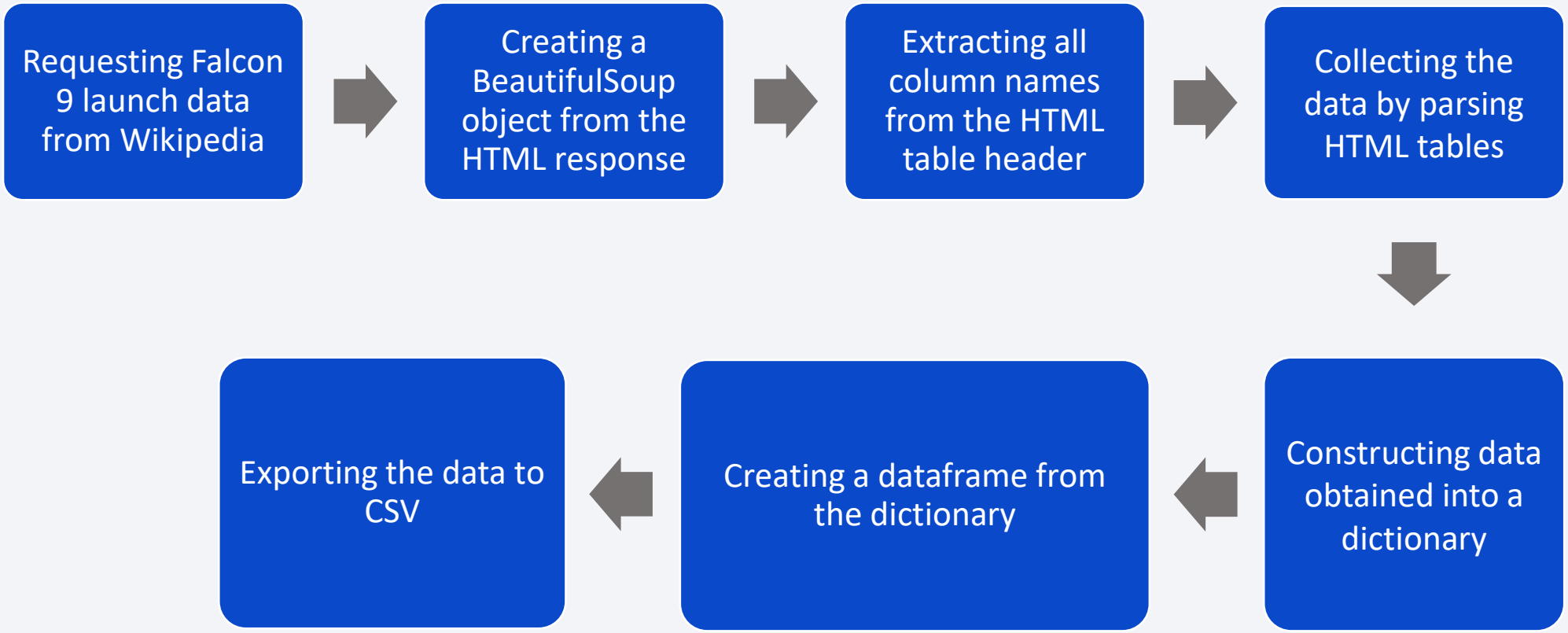
- In this project, data collection process involved a combination of API requests using SpaceX API and Web Scraping of data from SpaceX's Wikipedia entry. These methods complement each other and ensure complete information and adequate data are extracted.
- Data extracted using SpaceX REST API were (by their column names): **FlightNumber**, **Date**, **BoosterVersion**, **PayloadMass**, **Orbit, LaunchSite**, **Outcome**, **Flights**, **GridFins**, **Reused**, **Legs**, **LandingPad**, **Block**, **ReusedCount**, **Serial**, **Longtitude**, and **Latitude**.
- Data extracted using Web Scraping were (by their column names): **FlightNo.**, **Launch site**, **Payload**, **PayloadMass**, **Orbit**, **Customer**, **Launch Outcome**, **Version Booster**, **Booster Landing**, **Date**, and **Time**.

Data Collection – SpaceX API



GitHub URL: Data Collection Notebook

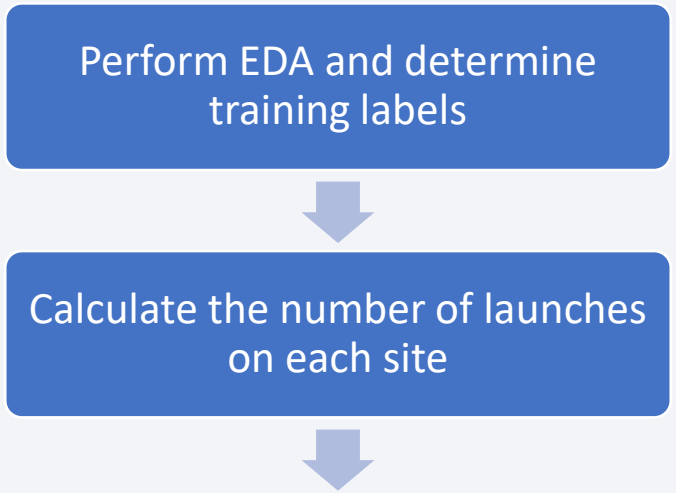
Data Collection – Web Scraping API



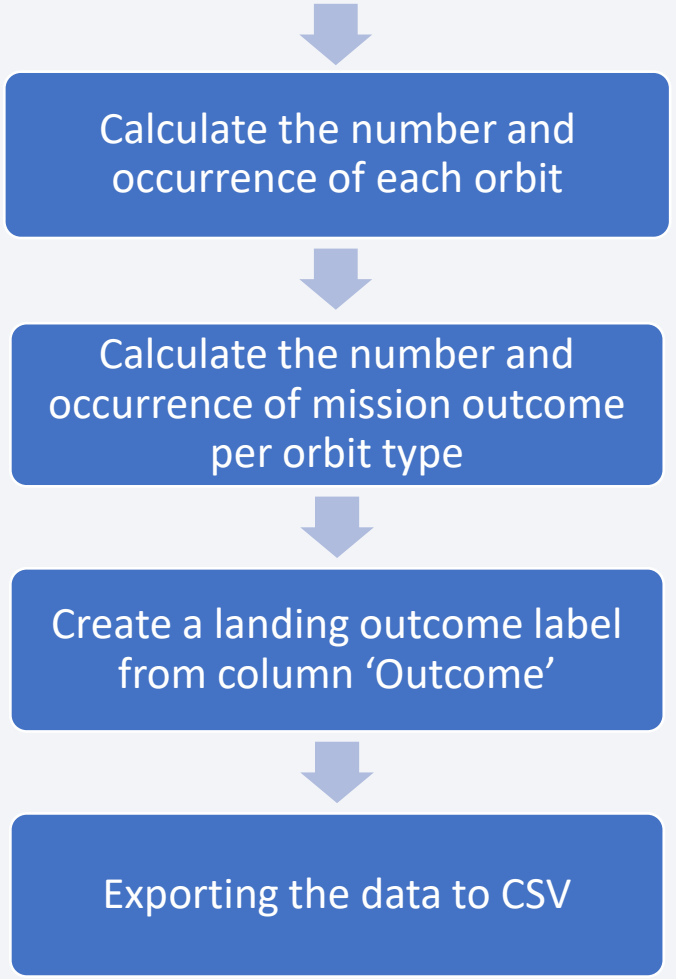
GitHub URL: Web Scraping Notebook

Data Wrangling

- Some EDA were performed to identify insights and patterns in the date and to determine the label for training supervised classification models.
- The outcomes of the launches, either successful or unsuccessful, were converted binarily (1 for successful landing and 0 for unsuccessful landing) into training labels for the supervised classification models.



Data Wrangling (Continued)



GitHub URL: Data Wrangling Notebook

EDA with Data Visualization

- Scatter plots were plotted to illustrate the correlation between two variables:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit
  - Payload Mass vs. Orbit
- Bar plots were plotted to compare the data from different groups:
  - Orbit vs. Success Rate
- Line plot was plotted to show success rate over time:
  - Year vs. Success Rate

GitHub URL: EDA with Data Visualization Notebook

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2020-06-04 and 2017-03-20 in descending order

GitHub URL: EDA with SQL Notebook

Build an Interactive Map with Folium

- Markers of all launch sites:
  - Added circular marker, popup label, and text label for NASA Johnson Space Center using its coordinates.
  - Added circular markers, popup labels, and text labels for all launch sites using their coordinates to show their geographical locations on the map and their relative proximities to the Earth's equator and coasts.
- Coloured markers to indicate launch outcomes for each launch site:
  - Added coloured markers (green for successes landing and red for unsuccessful landing) using Marker Cluster to identify and compare the success rates between different launch sites graphically.
- Distances between a launch site and its proximities:
  - Added coloured lines to show distances between the launch site, for instance, KSC LC-39A, and its proximities like railway, highway, coastline, and closest city.

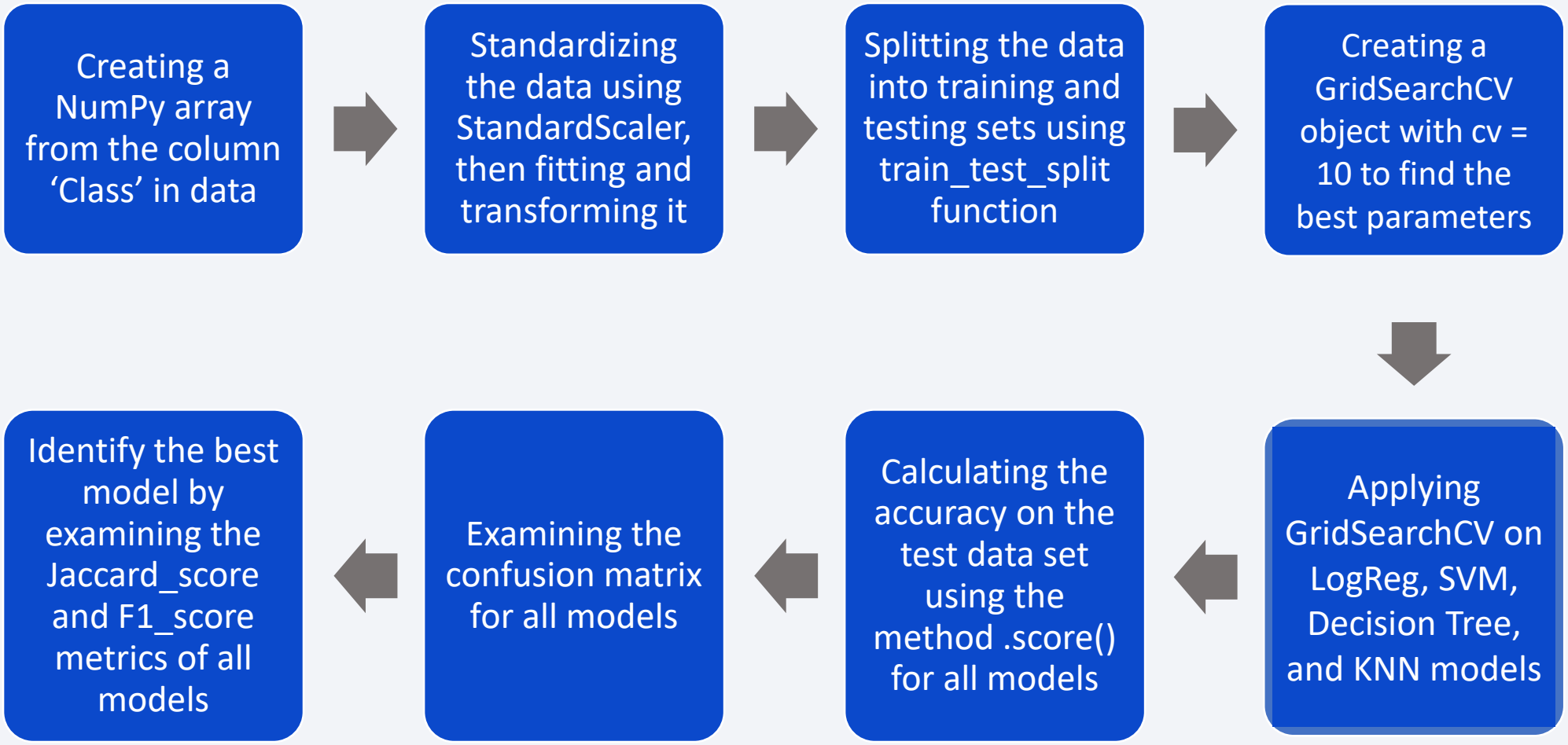
GitHub URL: Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

- Launch sites dropdown list:
  - Added a dropdown list to enable launch site option selection.
- Pie chart showing success launches:
  - Added a pie chart to visualize the total successful launches count for all sites or a specific site.
- Range slider of payload mass to access the corresponding launch outcomes:
  - Added a slider to determine the range of payload mass to access the corresponding launch outcomes.
- Scatter plot of payload mass versus success rate for different booster versions:
  - Added a scatter plot to illustrate the correlations between payload mass and success rate for different boosters.

GitHub URL: Dash App Notebook

Predictive Analysis (Classification)



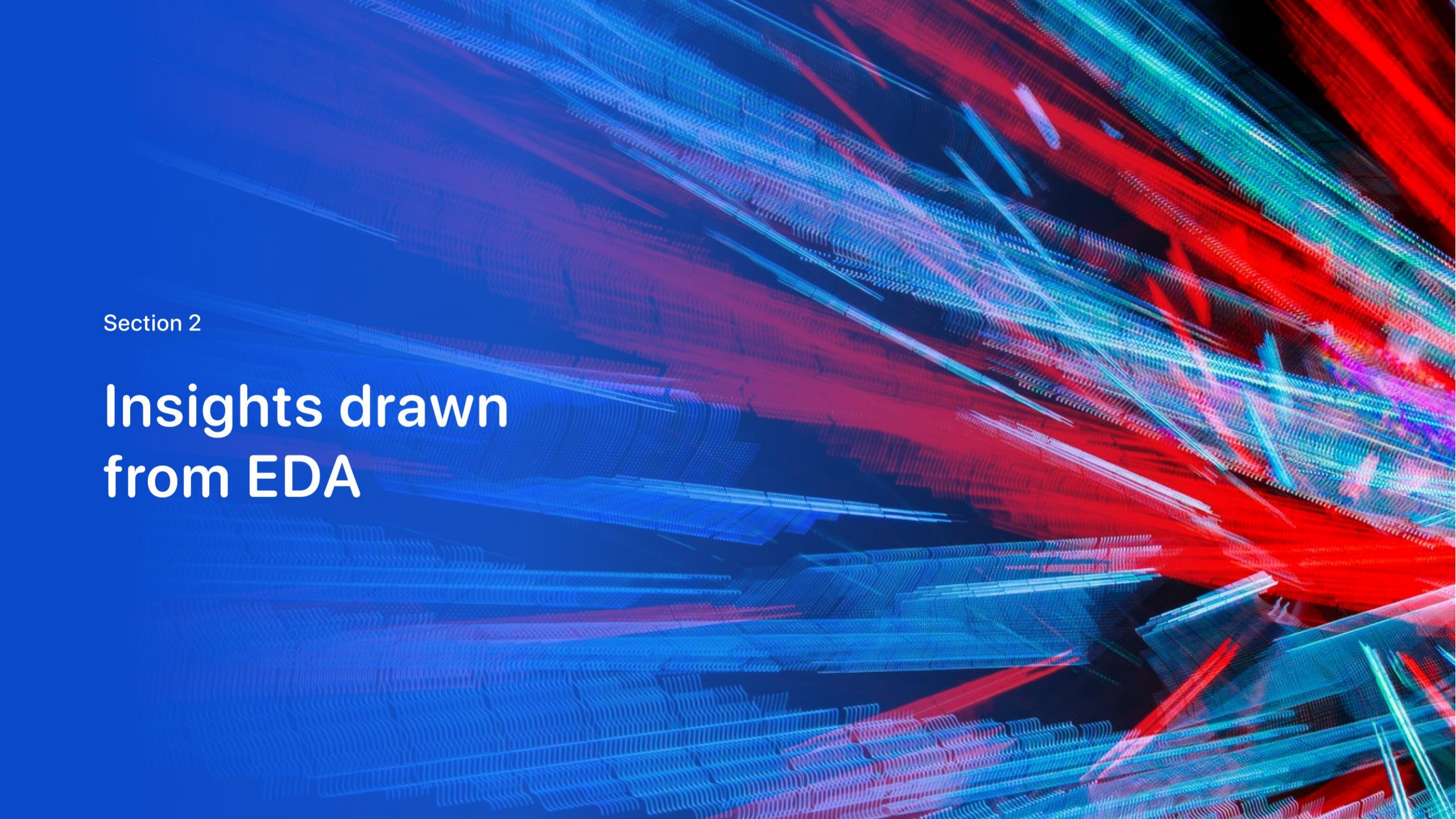
GitHub URL: [Machine Learning Prediction Notebook]

Results

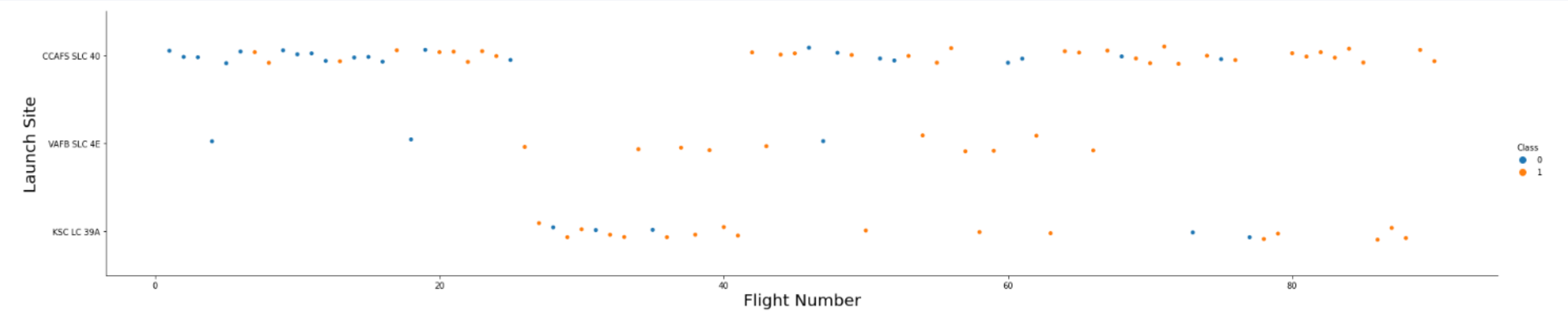
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

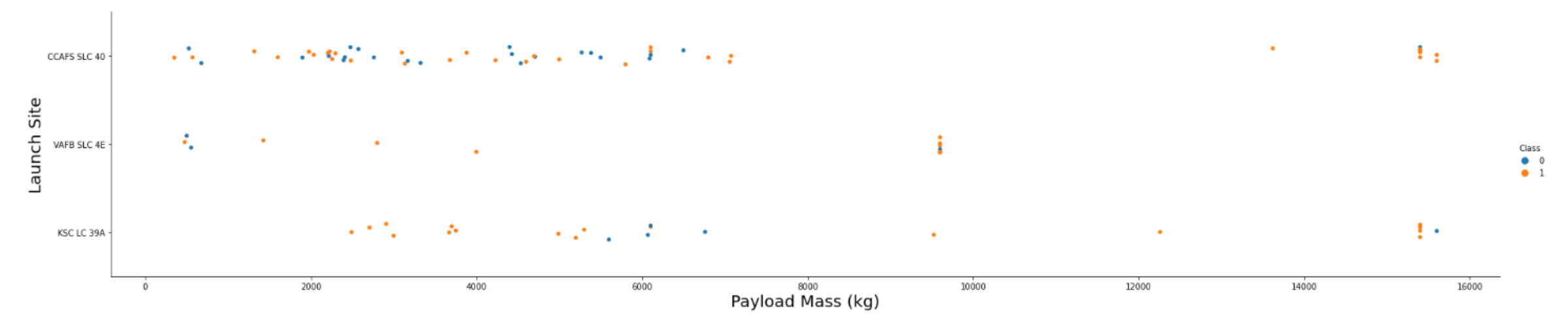


Flight Number vs. Launch Site



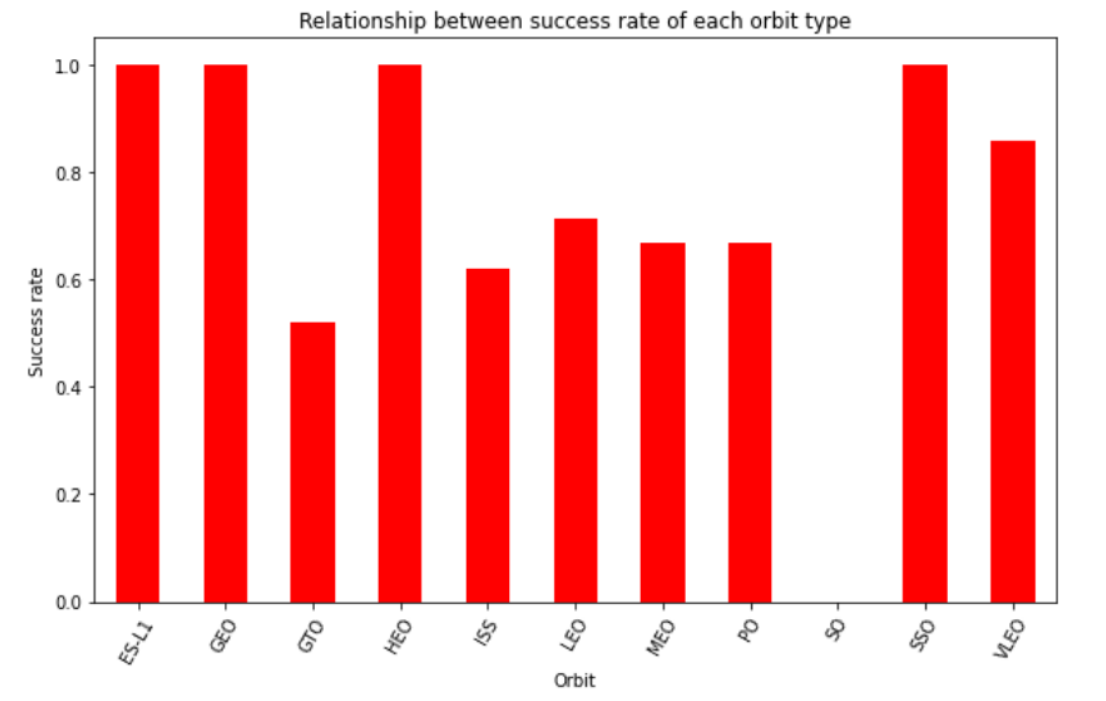
- Insights:
  - The CCAFS SLC-40 launch site has recorded most launches.
  - VAFB SLC-4E and KSC LC-39A have higher success rates.
  - The newer the launch, the higher the rate of success.

Payload vs. Launch Site



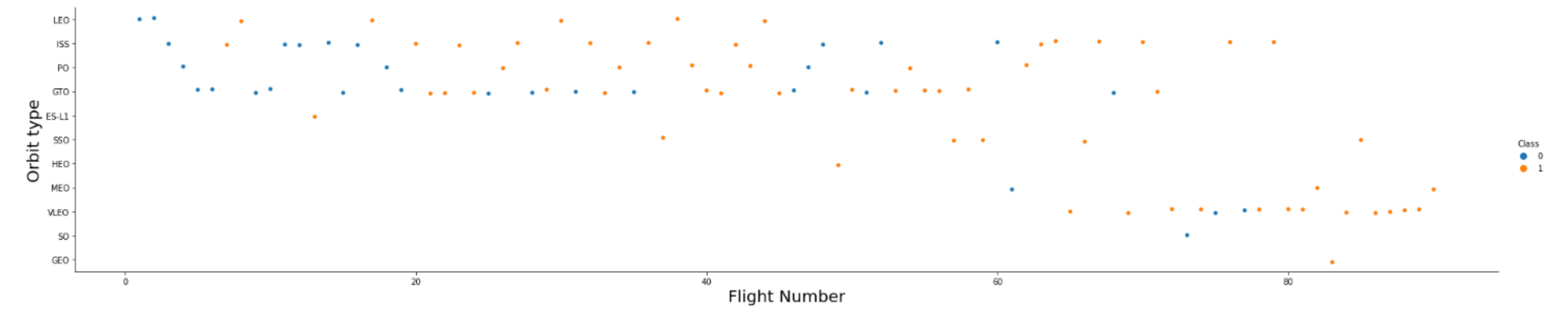
- Insights:
  - The higher the payload mass, the higher the success rate (for each launch site).
  - Most launches with payload mass over 7000 kg were successful.
  - KSC LC-39A has a 100% success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type



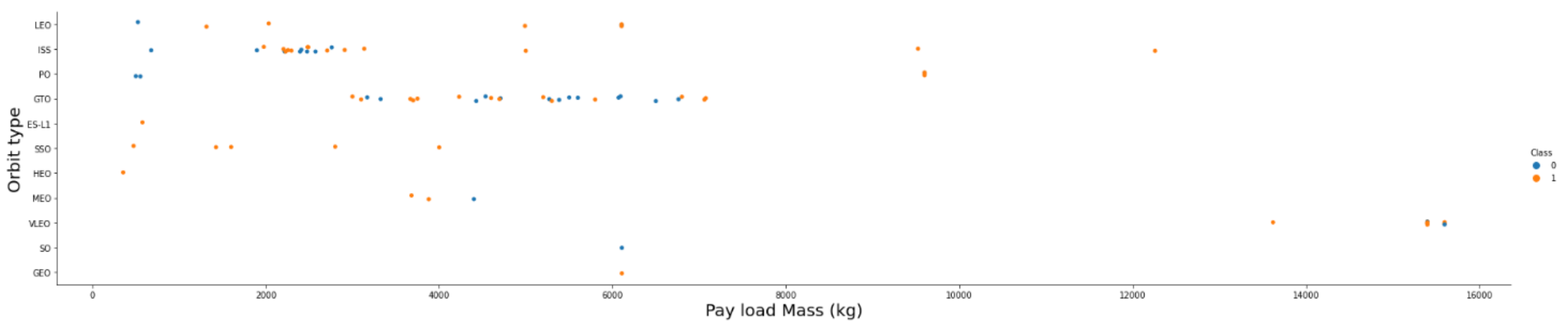
- Insights:
  - ES-L1, GEO, GEO, and SSO are orbits with 100% success rate.
  - SO is the orbit with 0% success rate.
  - GTO, ISS, LEO, MEP, and PO are orbits with success rate between 50% and 85%.

Flight Number vs. Orbit Type



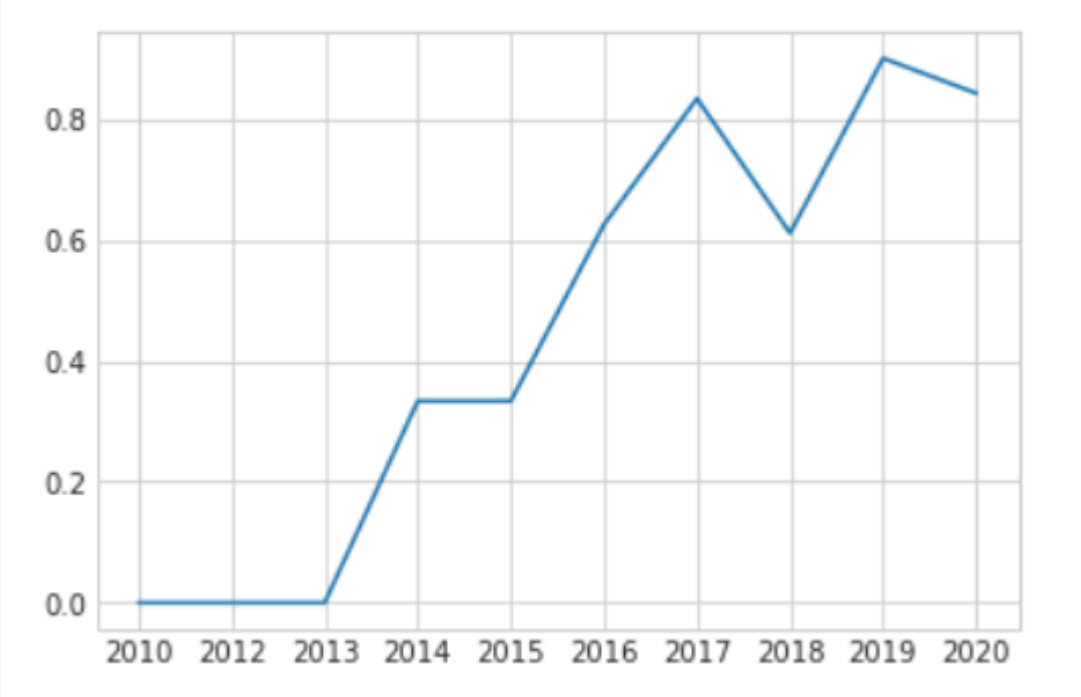
- Insights:
  - Flight number appears to be highly correlated with success in LEO orbit
  - Flight number appears to be uncorrelated with success in GTO orbit.

Payload vs. Orbit Type



- Insights:
  - Payload mass appears to be highly correlated with success in LEO and ISS orbit.
  - Payload mass appears to be uncorrelated with success in GTO, MEO, and VLEO orbits.

Launch Success Yearly Trend



- Insights:
  - The success rate since 2014 kept increasing till 2020.

All Launch Site Names

```
%sql SELECT DISTINCT launch_site FROM SPACEXDATASET
```

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

- Explanation:
  - Displaying the names of the unique (DISTINCT in the query yields only unique values) launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
%%sql SELECT * FROM SPACEXDATASET
    WHERE launch_site LIKE 'CCA%'
    LIMIT 5;
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Explanation:
  - Displaying 5 records where the launch sites begin with the string ‘CCA’. (Note the LIKE ‘CCA%’)

Total Payload Mass

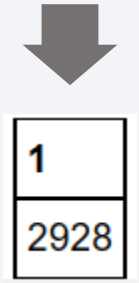
```
%%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET
    WHERE customer = 'NASA (CRS)'
```

1
45596

- Explanation:
  - Displaying the total payload mass carried by boosters launched by NASA (CRS). Note the SUM function that calculates the total in the payload mass column and the WHERE clause that filters the dataset to include only the customer NASA (CRS).

Average Payload Mass by F9 v1.1

```
%%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET
    WHERE booster_version = 'F9 v1.1'
```



1
2928

- Explanation:
  - Displaying the average payload mass carried by booster version F9 v1.1. Note the AVG function that calculates the average in the payload mass column and the WHERE clause that filters the dataset to include only the F9 v1.1 booster version.

First Successful Ground Landing Date

```
%%sql SELECT MIN(date) FROM SPACEXDATASET
    WHERE landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

- Explanation:
  - Listing the date when the first successful landing outcome in ground pad was achieved. Note the MIN function that yields the earliest date in the column ‘Date’ and the WHERE clause that filters the dataset to include only successful landing outcome on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT booster_version FROM (SELECT * FROM SPACEXDATASET
    WHERE payload_mass__kg_ BETWEEN 4000 AND 6000)
    WHERE landing__outcome = 'Success (drone ship)'
```

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

- Explanation:
  - Listing the boosters with successful drone ship landing and payload between 4000 and 6000. Note the WHERE clause filters the dataset to include only payloads between 4000 and 6000 kg and successful landings on drone ship.

Total Number of Successful and Failure Mission Outcomes

```
%%sql SELECT mission_outcome, COUNT(mission_outcome) AS total F
ROM SPACEXDATASET
    GROUP BY mission_outcome
    ORDER BY total DESC
```

mission_outcome	total
Success	99
Failure (in flight)	1
Success (payload status unclear)	1

- Explanation:
  - Listing the total number of successful and failure mission outcomes. Note that mission outcome is different from landing outcome.

Boosters Carried Maximum Payload

```
%%sql SELECT DISTINCT(booster_version) FROM SPACEXDATASET
    WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FR
OM SPACEXDATASET)
```

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

- Explanation:
  - Listing the names of booster versions that have carried the maximum payload mass. Note that GROUP BY with DESC puts the list on order based on maximum payload mass.

2015 Launch Records

```
%%sql SELECT booster_version, launch_site FROM (SELECT * FROM S
PACEXDATASET
    WHERE DATE >='01-01-2015' AND DATE < '01-01-2016')
    WHERE landing__outcome = 'Failure (drone ship)'
```

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

- Explanation:
  - Listing the failed landing outcomes in drone ship, their boosters' versions and the launch sites names for the months in year 2015.

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

```
%%sql SELECT landing__outcome, COUNT(landing__outcome) AS total
FROM (SELECT * FROM SPACEXDATASET
    WHERE DATE >='06-04-2010' AND DATE <= '03-20-2017')
    GROUP BY landing__outcome
    ORDER BY total DESC
```

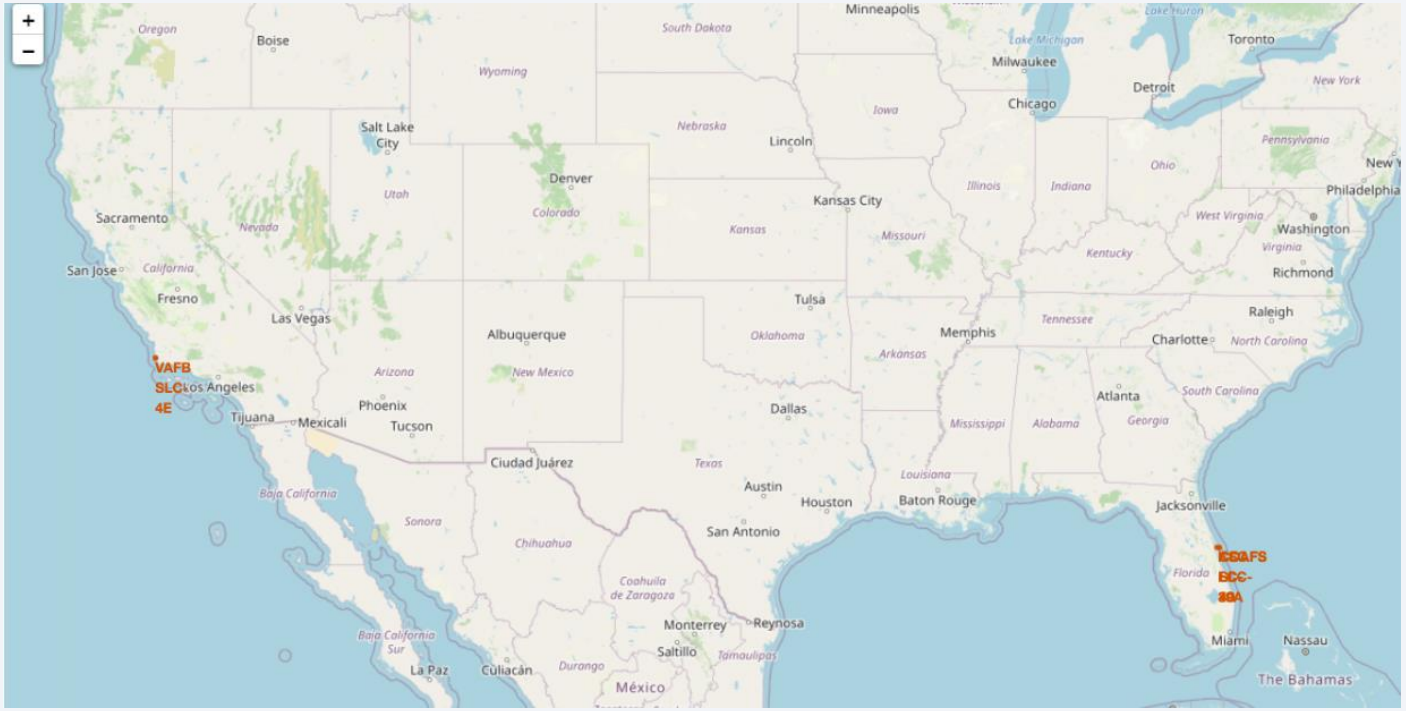
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Explanation:
  - Listing the counts of landing outcomes, both success and failure, between 2010-06-04 and 2017-03-20.

Section 4

Launch Sites Proximities Analysis

Launch Sites on Map

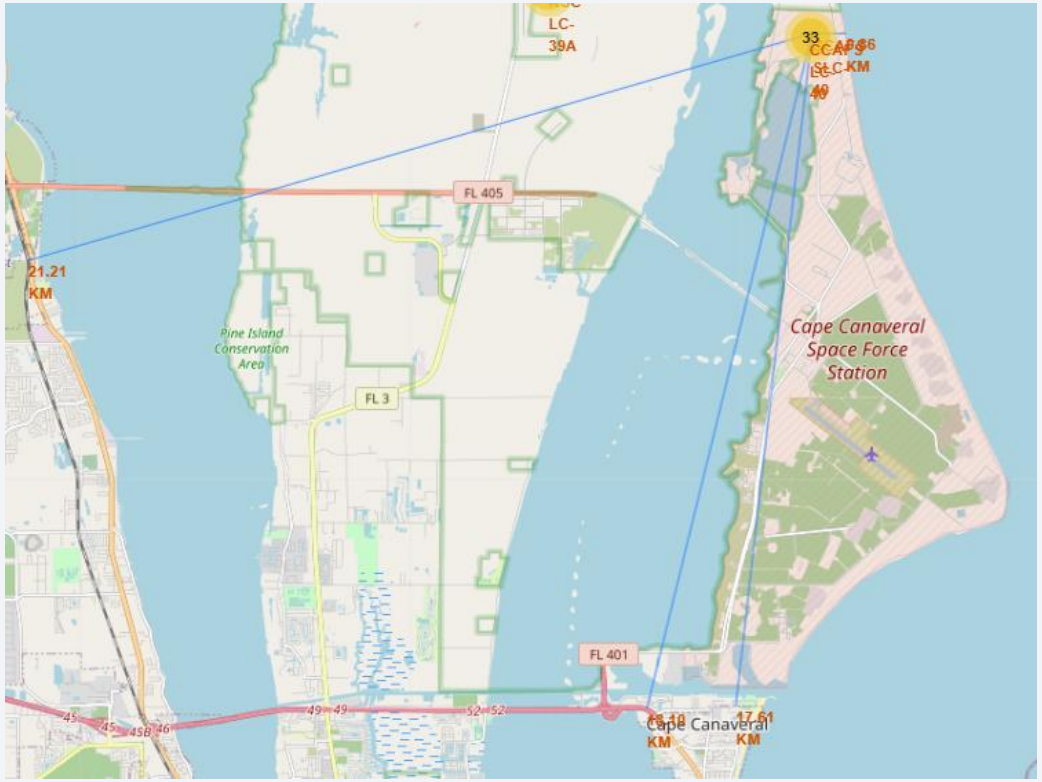


- Explanation:
  - Launch sites are in proximity to the Earth's equator line. The land is moving faster at the equator (larger inertia), if a spaceship is launched from the equator it will go into the space with a larger speed, this is due to the larger inertia.
  - Launch sites are in very close proximity with the coasts, launching rockets towards the ocean minimizes the risk of dropping debris or exploding near crowded places.

Colour-Labeled Launch Records on Map

- Explanation:
  - Green Marker = successful launch, Red Marker = unsuccessful launch.

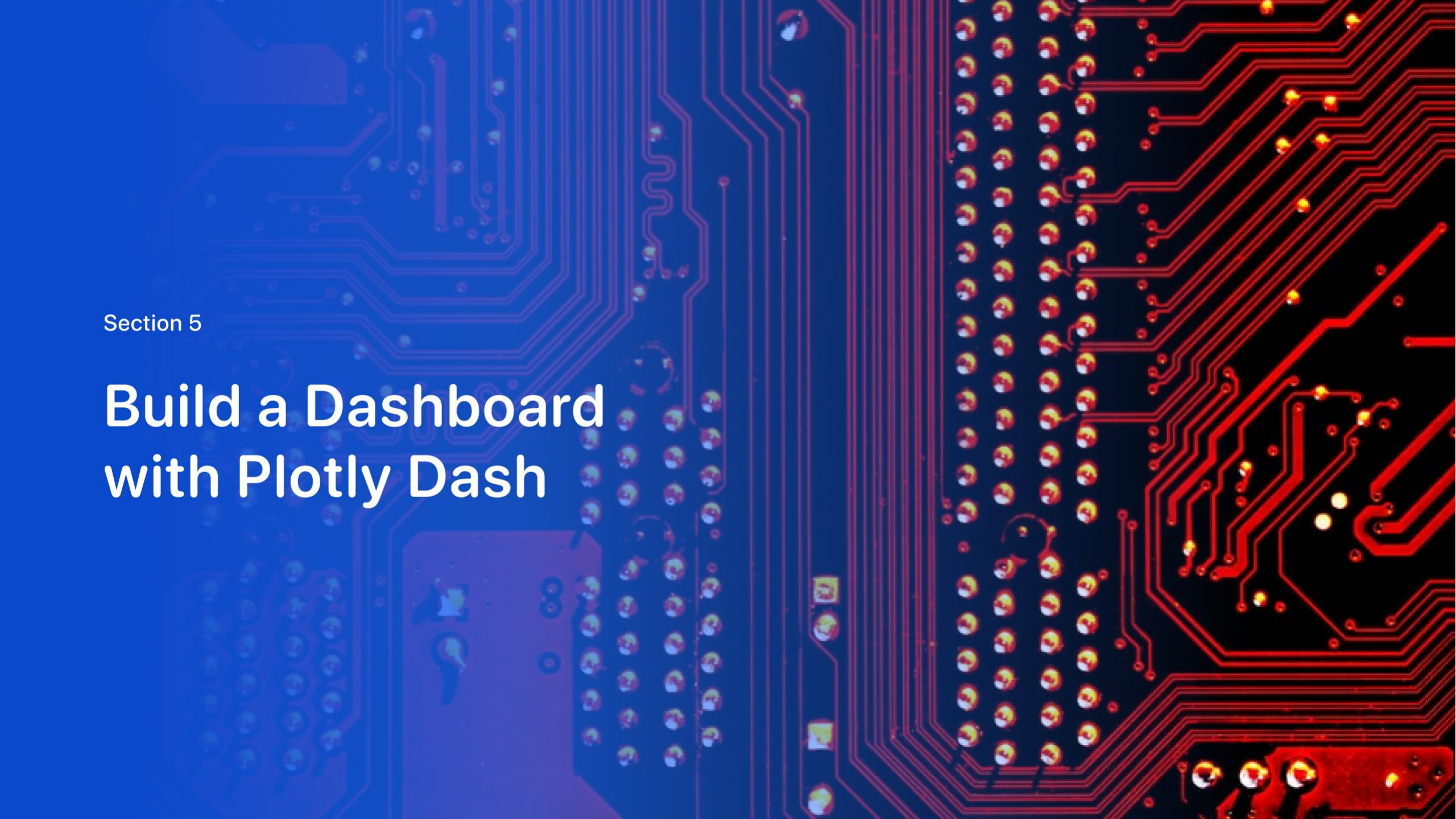
Distances from CCAFS SLC 40 to its Proximities



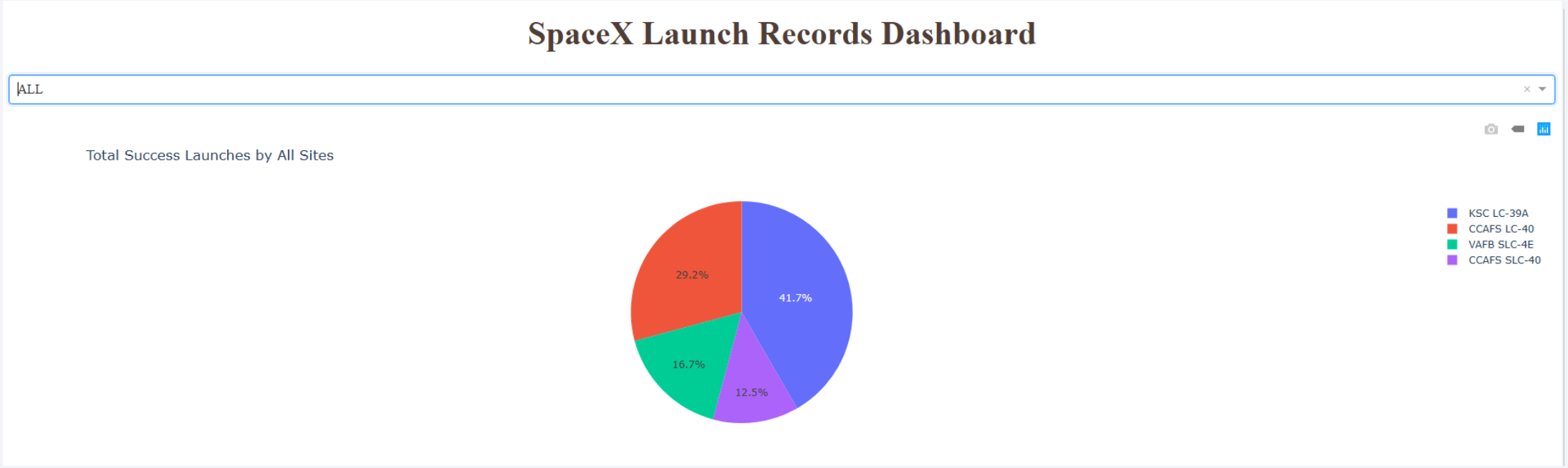
- Explanation:
  - Launch sites are normally close to railway, highway, and coastline but far from populated areas.

Section 5

Build a Dashboard with Plotly Dash

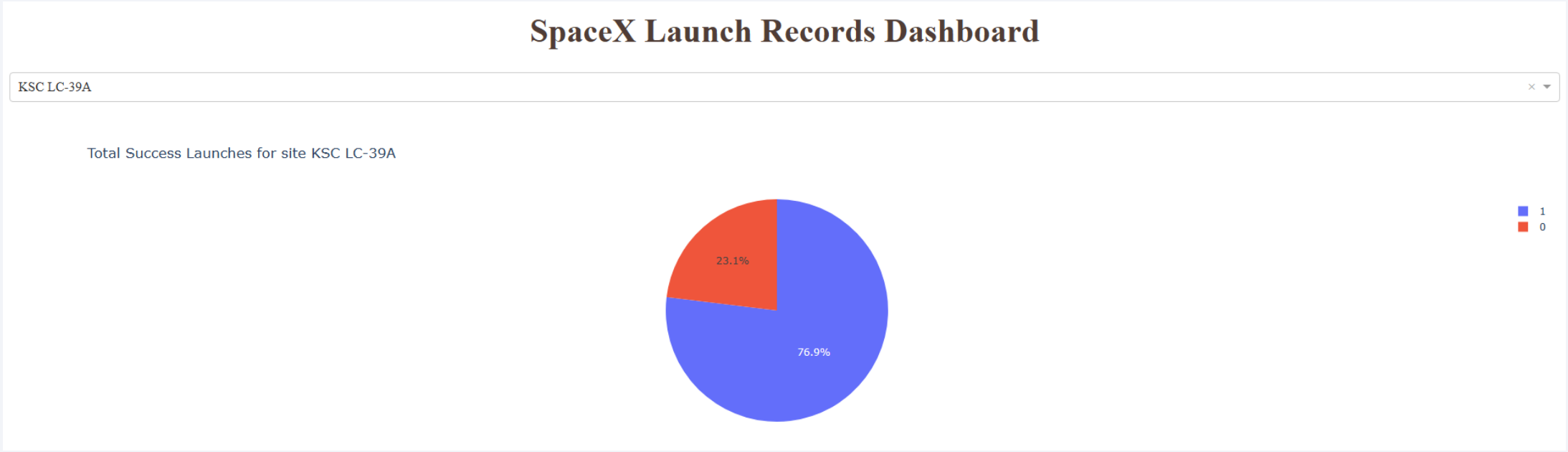


Percentage of Succeeded Launch by Launch Site



- Explanation:
  - KSC LC-39A launch site has the most successful launches.

KSC LC-39A Launch Site



- Explanation:
  - KSC LC-39A launch site has 76.9% of success rate, with 10 successful launches and only 3 unsuccessful landings over the years.

Payload vs. Launch Outcome

- Explanation:
  - The higher the payload mass, the lower the success rate.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

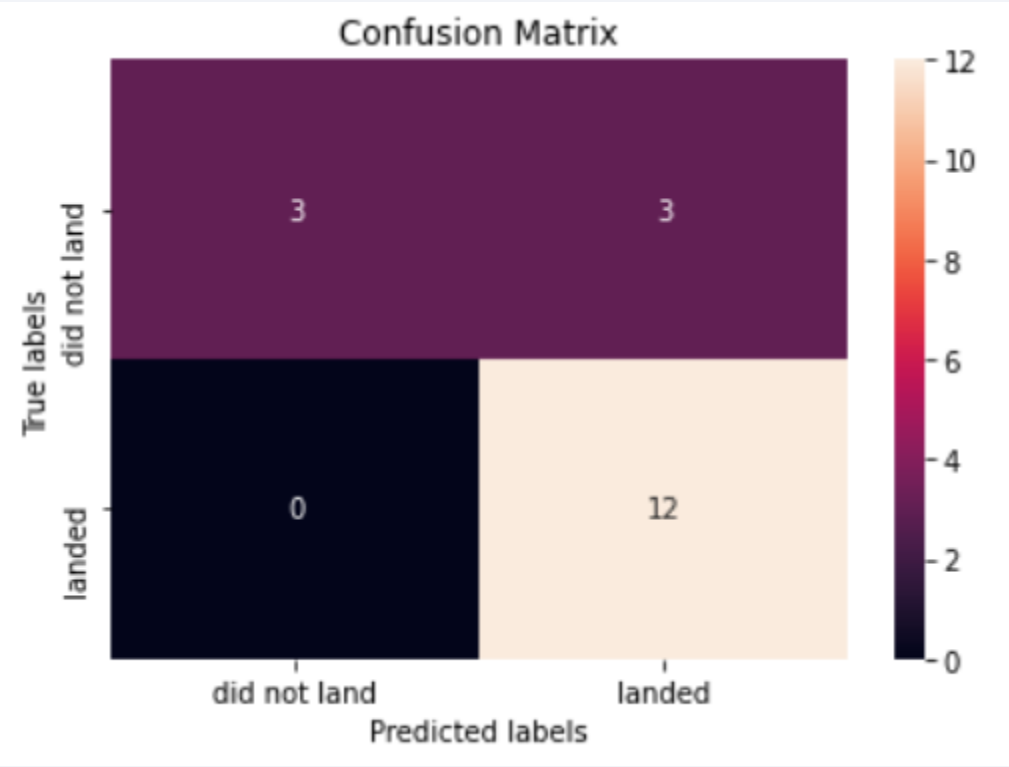
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Explanation: Scores and accuracies of test set. All models performed similarly. (Possible reason: small test sample size)

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.830986	0.819444
F1_Score	0.909091	0.916031	0.907692	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

Explanation: Scores and accuracies of entire data set (train + test). Assessing the models using the entire data set suggests that SVM is the most accurate model.

Confusion Matrix



Explanation: The model performs well in distinguishing classes binarily; false negatives, however, can be observed.

Conclusions

- SVM is the best algorithm for this project.
- Launches with a lower payload mass show better results than launches with a larger payload mass (opposite result is observed when payload mass is plotted against success rate for different launch sites)
- Launch sites are generally in close proximity with the Earth's equator line and coasts.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches as compared to other sites.
- Orbits ES-L1, GEO, HEO, and SSO have 100% success rate up-to-date.

Appendix

```
# Pandas is a software library written for the Python programming language for data manipulation and analysis.
import pandas as pd
# NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays
import numpy as np
# Matplotlib is a plotting library for python and pyplot gives us a MatLab like plotting framework. We will use this in our plotter function to plot data.
import matplotlib.pyplot as plt
#Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics
import seaborn as sns
# Preprocessing allows us to standarsize our data
from sklearn import preprocessing
# Allows us to split our data into training and testing data
from sklearn.model_selection import train_test_split
# Allows us to test parameters of classification algorithms and find the best one
from sklearn.model_selection import GridSearchCV
# Logistic Regression classification algorithm
from sklearn.linear_model import LogisticRegression
# Support Vector Machine classification algorithm
from sklearn.svm import SVC
# Decision Tree classification algorithm
from sklearn.tree import DecisionTreeClassifier
# K Nearest Neighbors classification algorithm
from sklearn.neighbors import KNeighborsClassifier
```

Explanation: Some Python libraries used in this project.

Thank you!