Data Science Capstone Project

Neel Pandey

31/12/2022

Outline

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

Executive Summary

We predicted if the Falcon 9 first stage will land successfully. Much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

The following methodologies were used to analyze data:

- Data collection using web scraping and SpaceX Api
- Exploratory Data Analysis (EDA) including data wrangling, data visualization and interactive visual analytics
- Machine Learning Prediction

Results Summary

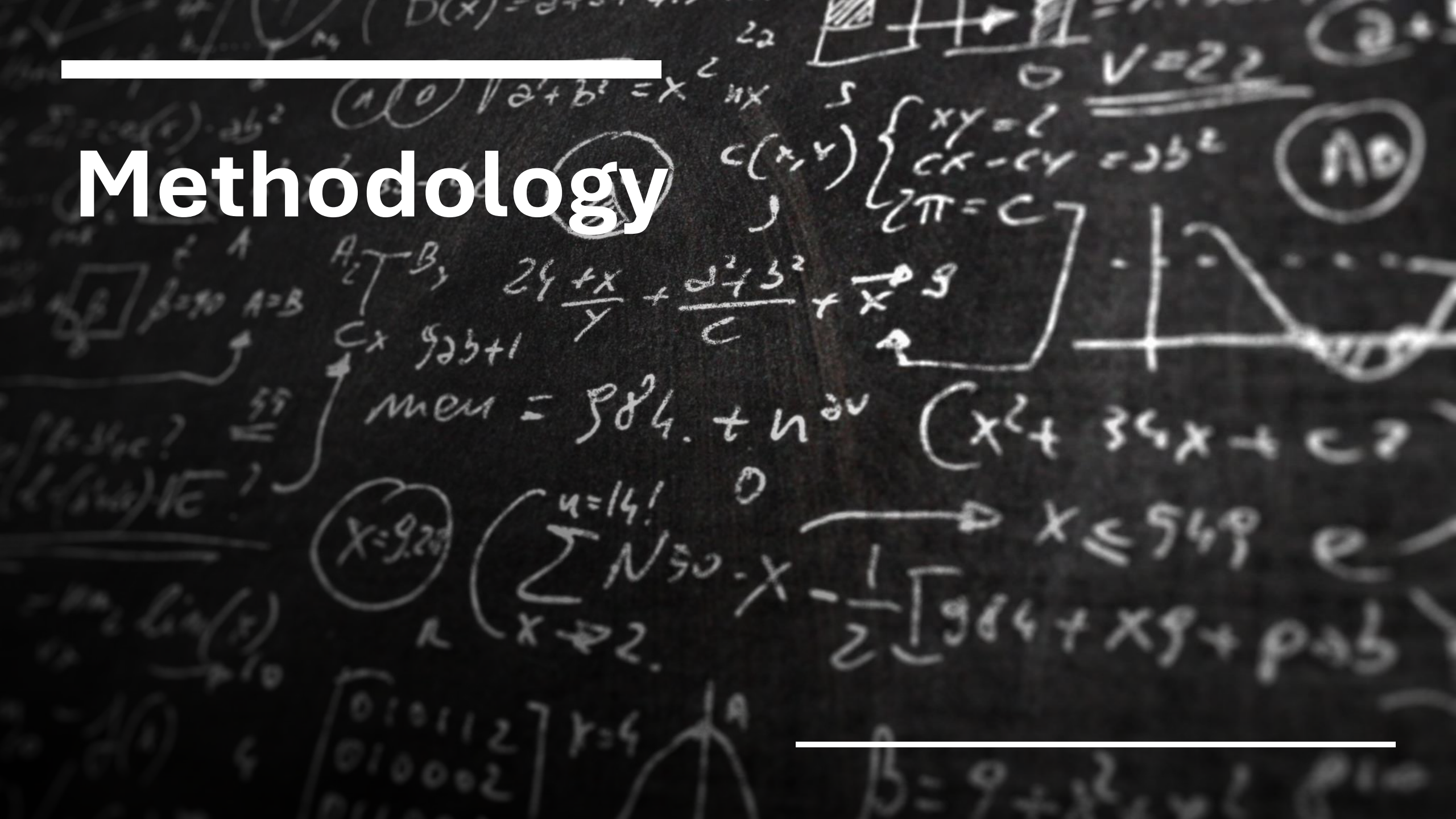
- With methods used, it was possible to collect valuable data from public sources
- EDA allowed to select best features to predict successful landing
- Machine Learning Prediction showed the best model and its characteristics selected for successful landing

Introduction

A new rocket company Space Y that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. We need to determine the following:

- the price of each launch, by gathering information about Space X and creating dashboards to extract information
- if SpaceX will reuse the first stage by training a machine learning model and use public information to predict if SpaceX will reuse the first stage.
- Best location to make launches

Methodology



Methodology

Executive Summary

- Data collection methodology:

  Data from Space X was obtained from 2 sources:

  - Space X API (https://api.spacexdata.com/v4/rockets/)

  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling

  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

Data sets were collected from:

- Space X API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

using web scraping technics.

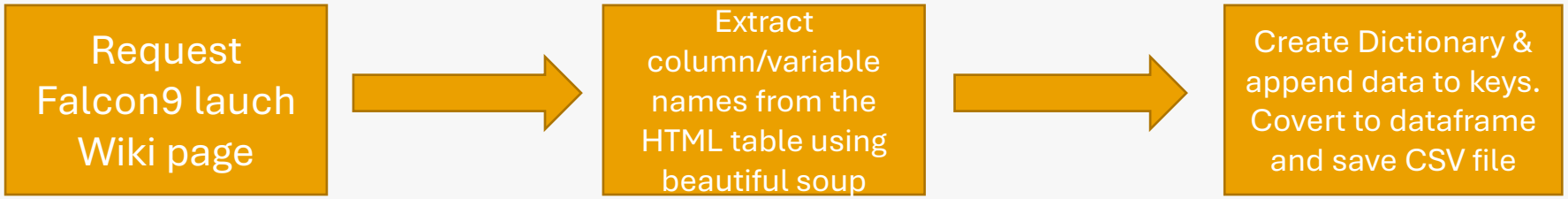
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and used for this project
- This API was used according to the flowchart beside and then data is persisted

- Source code: https://github.com/KaivitiBa/Applied-Data-Science-Capstone-Project

Data Collection – Web Scraping

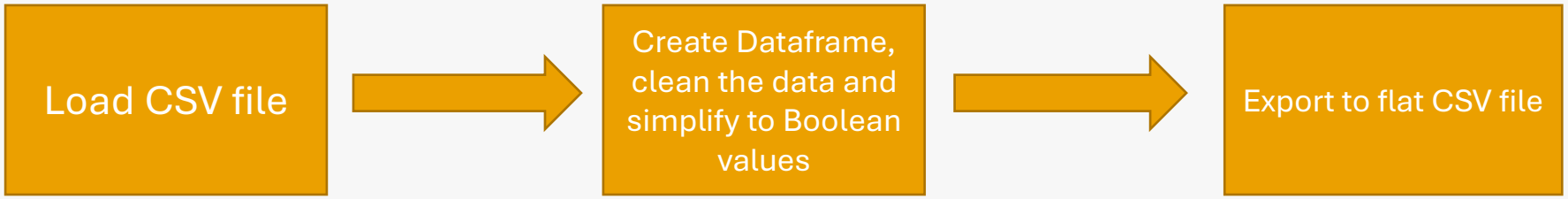
- SpaceX offers a public API from where data can be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and parsed



- Source code: https://github.com/KaivitiBa/Applied-Data-Science-Capstone-Project

Data Wrangling

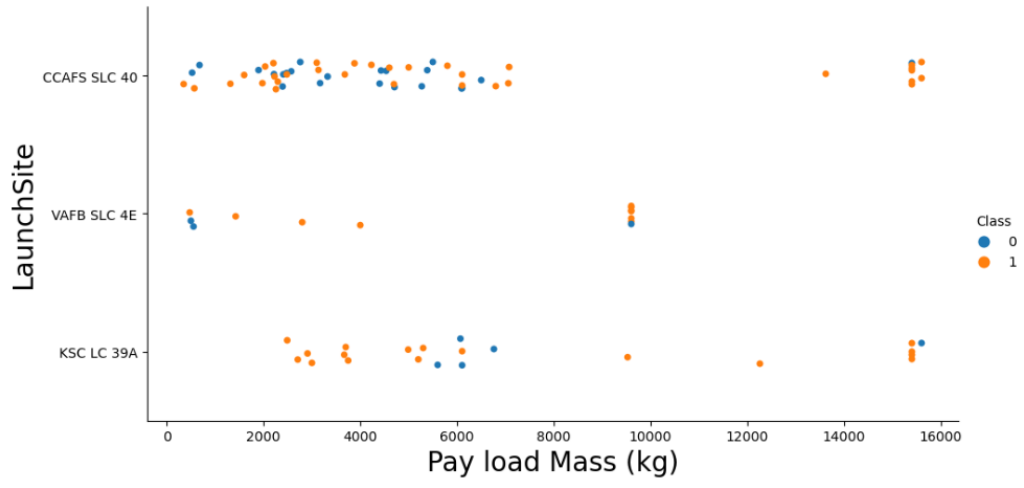
- Data Wrangling process cleans and simplifies messy, complex data sets for easy access and analysis
- Here we converted those outcomes into training labels with 1 meaning successful booster landing and 0 unsuccessful



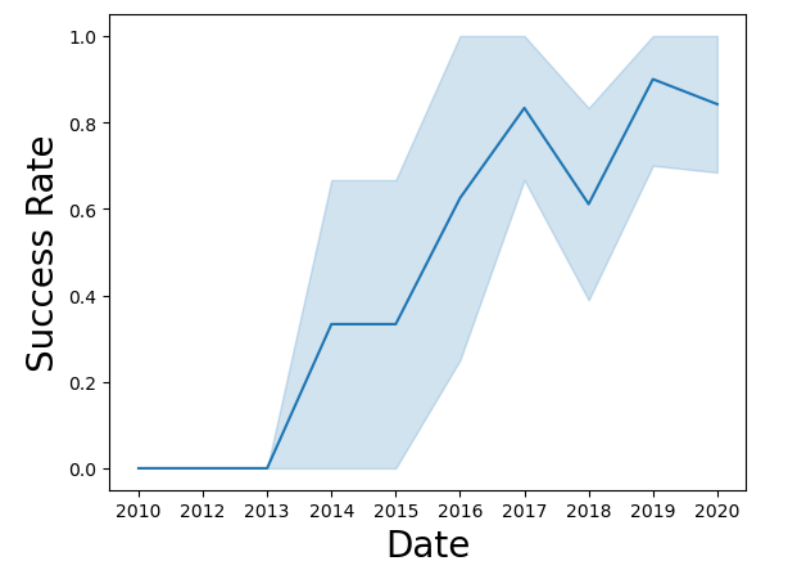
- Source code: https://github.com/KaivitiBa/Applied-Data-Science-Capstone-Project

EDA with Data Visualisation

- To explore data, scatterplots, line graph and barplots were used to visualise the relationship between pair of features
  - Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit vs Flight Number, Payload vs Orbit
- Scatter Plot

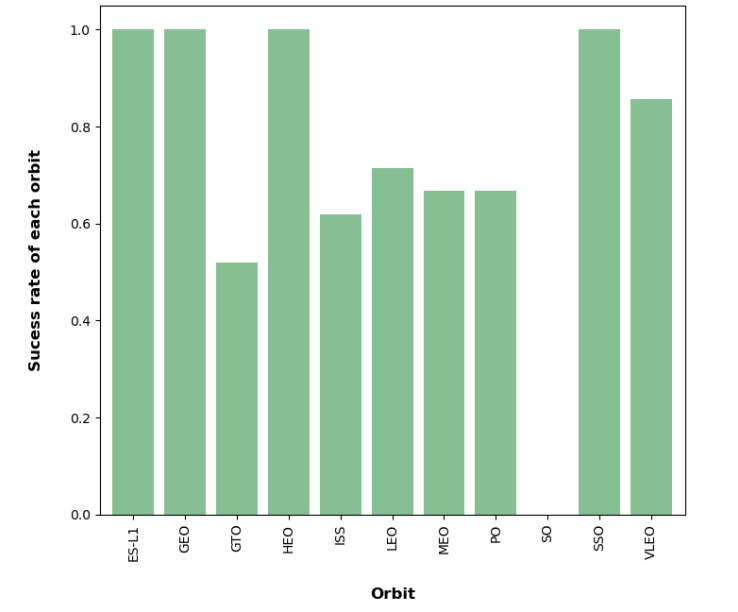


Line Graph



EDA with Data Visualisation

- Bar Graph



- Source code: https://github.com/KaivitiBa/Applied-Data-Science-Capstone-Project

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were used with Folium Maps:

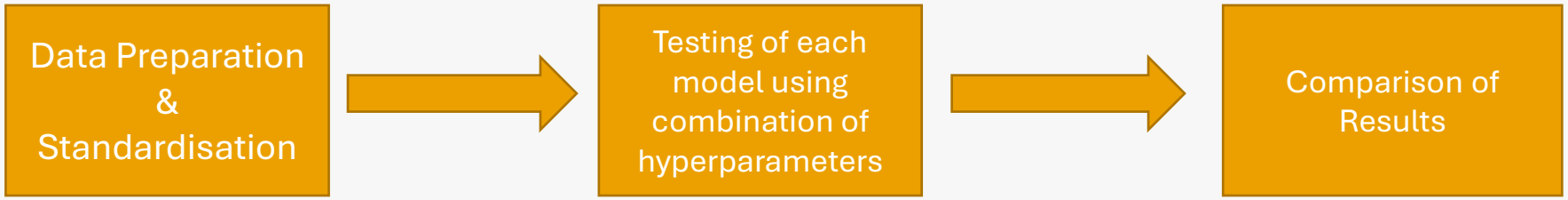
- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines are used to indicate distances between two coordinates.

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

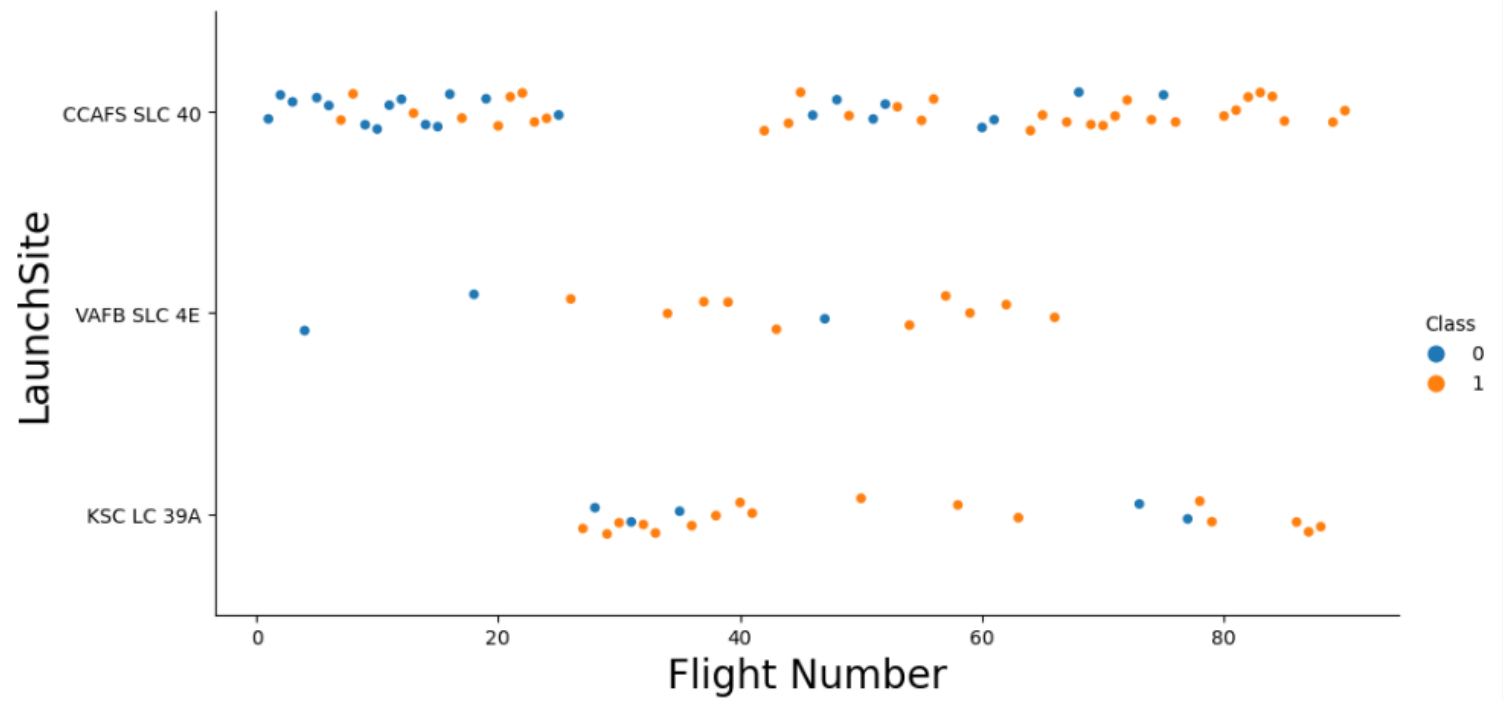


- Source code: https://github.com/KaivitiBa/Applied-Data-Science-Capstone-Project

Results
EDA with Data Visualisation

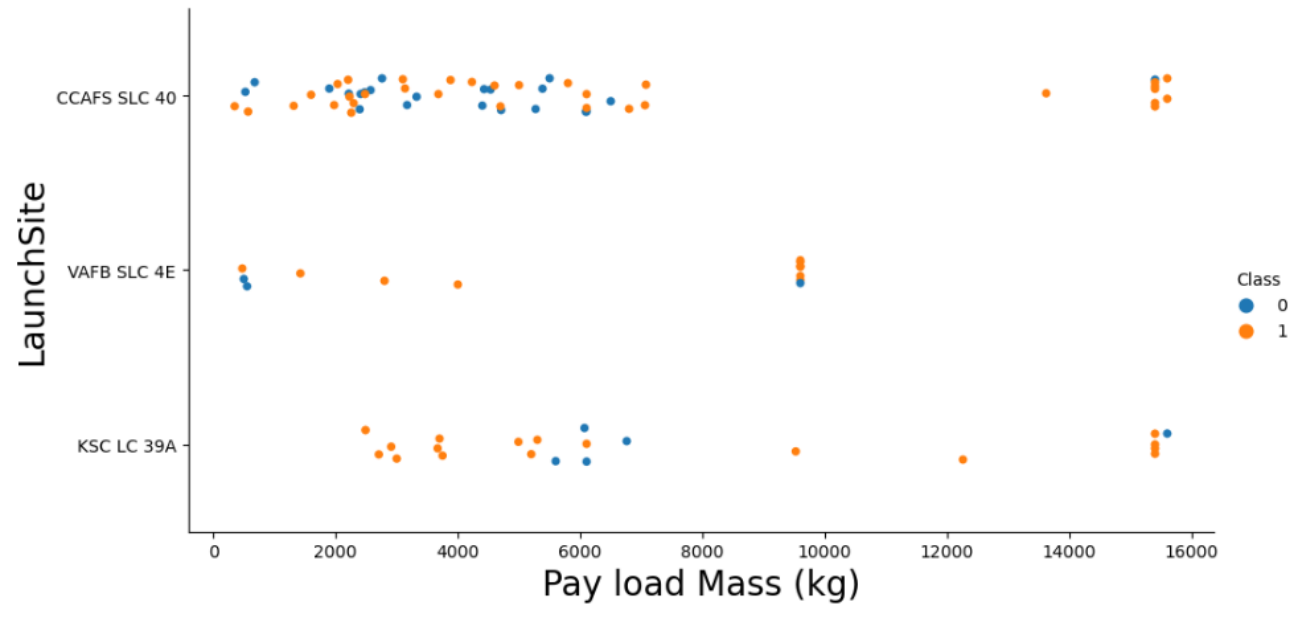
EDA with Data Visualisation

- Flight Number vs Launch Site:
  - Higher number of flights (>25) indicated increase in success rate for the Rockets



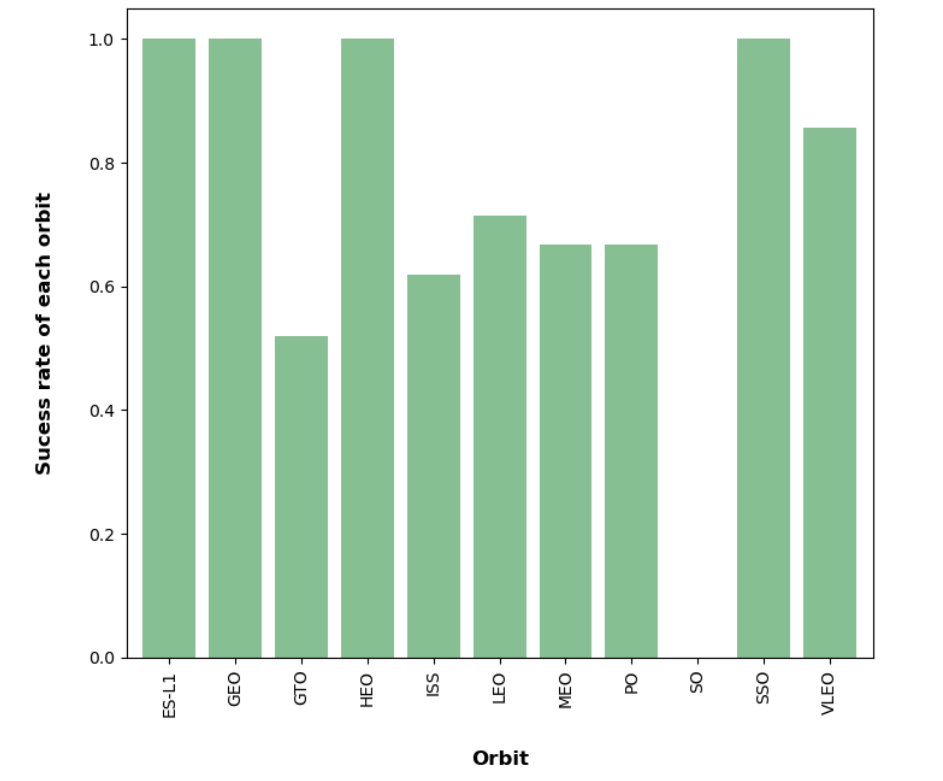
EDA with Data Visualisation

- Payload vs Launch Site:
  - Higher the payload mass (>7000Kg) higher the success rate for the Rockets. However no clear pattern to make decision if the launch site of dependent on Payload Mass only for successful launch.



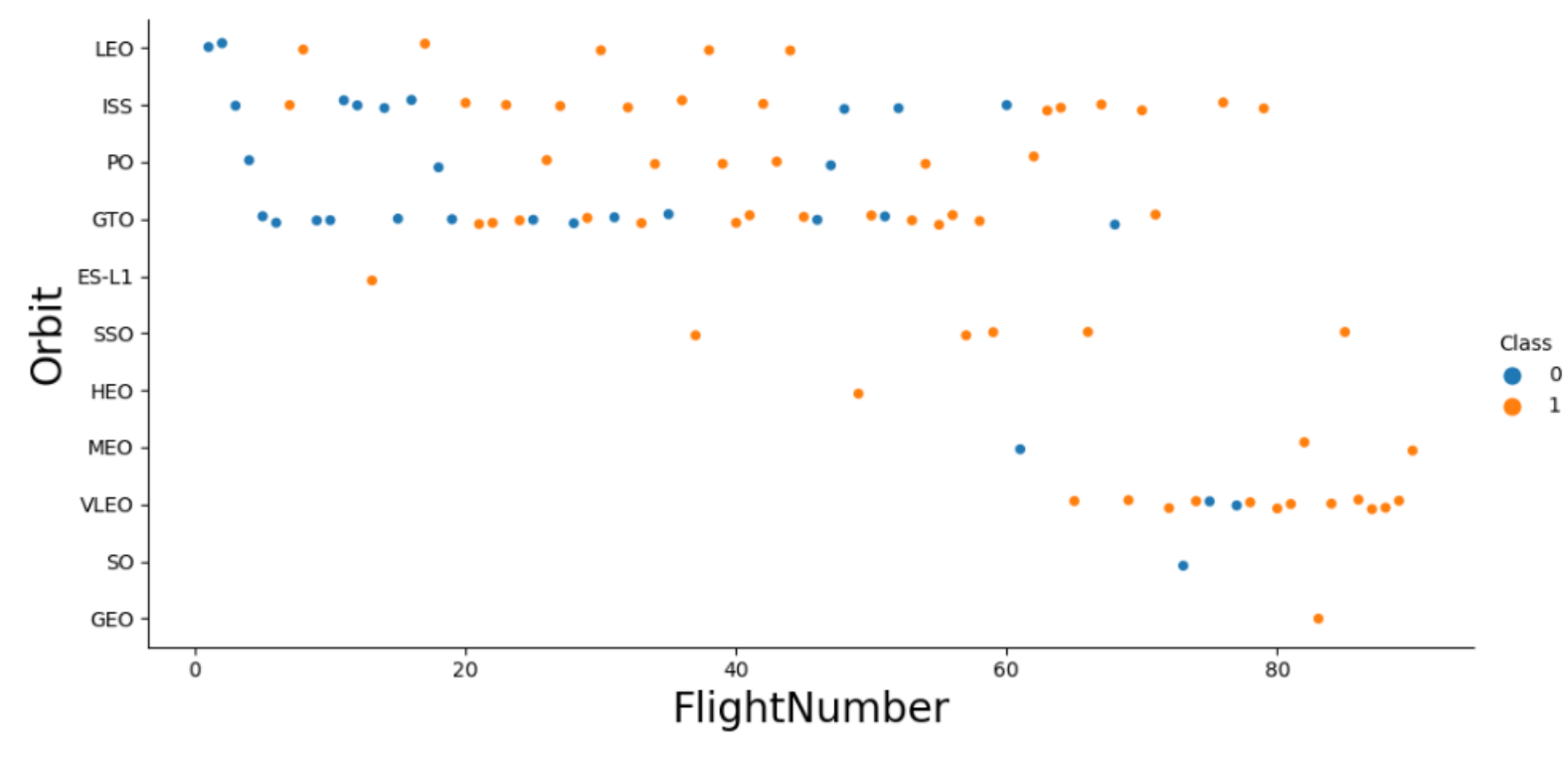
EDA with Data Visualisation

- Success Rate vs Orbit:
  - ES-L1, GEO, HEO, SSO has highest Success Rates



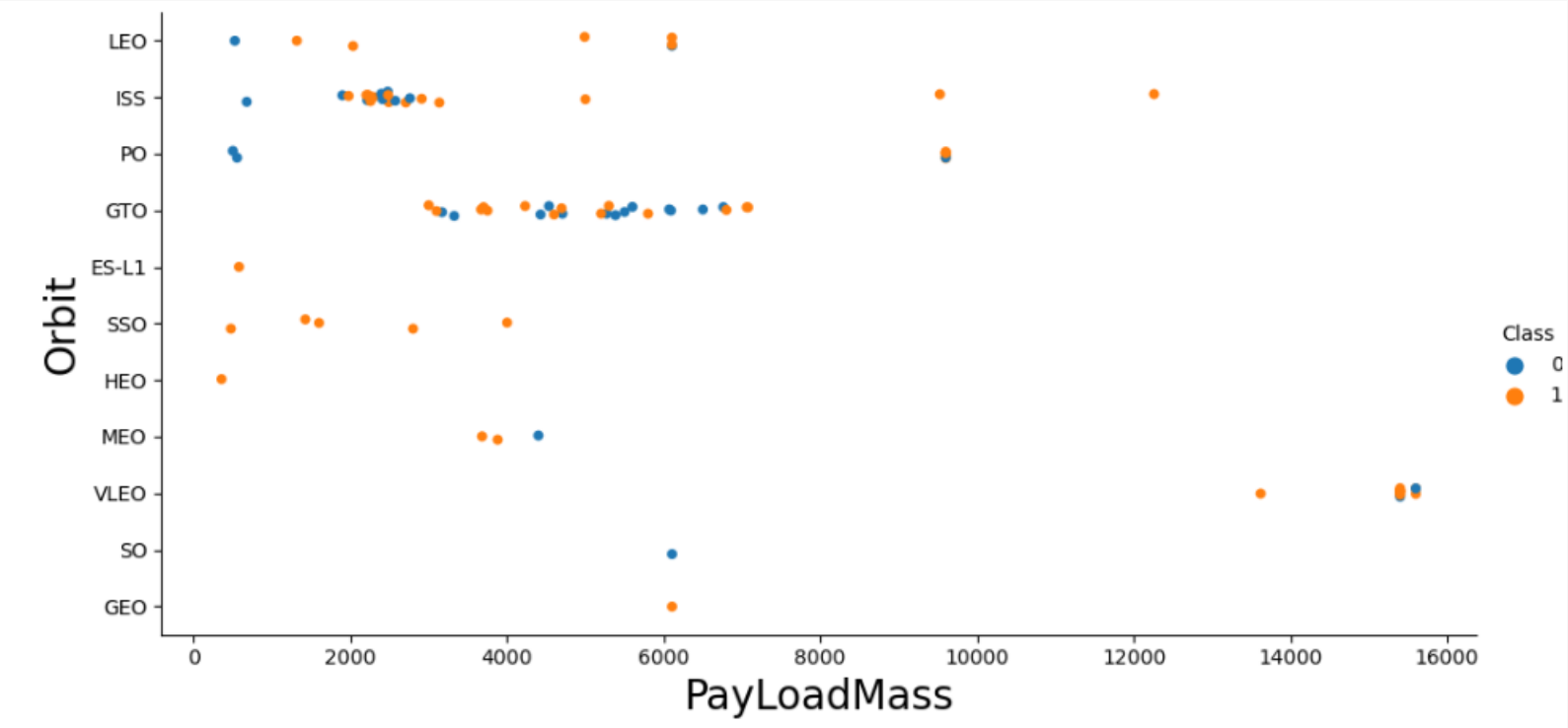
EDA with Data Visualisation

- Flight Number vs Orbit:
  - For LEO orbit, the success increases with the number of flights, however no relationship between GTO and flight number.



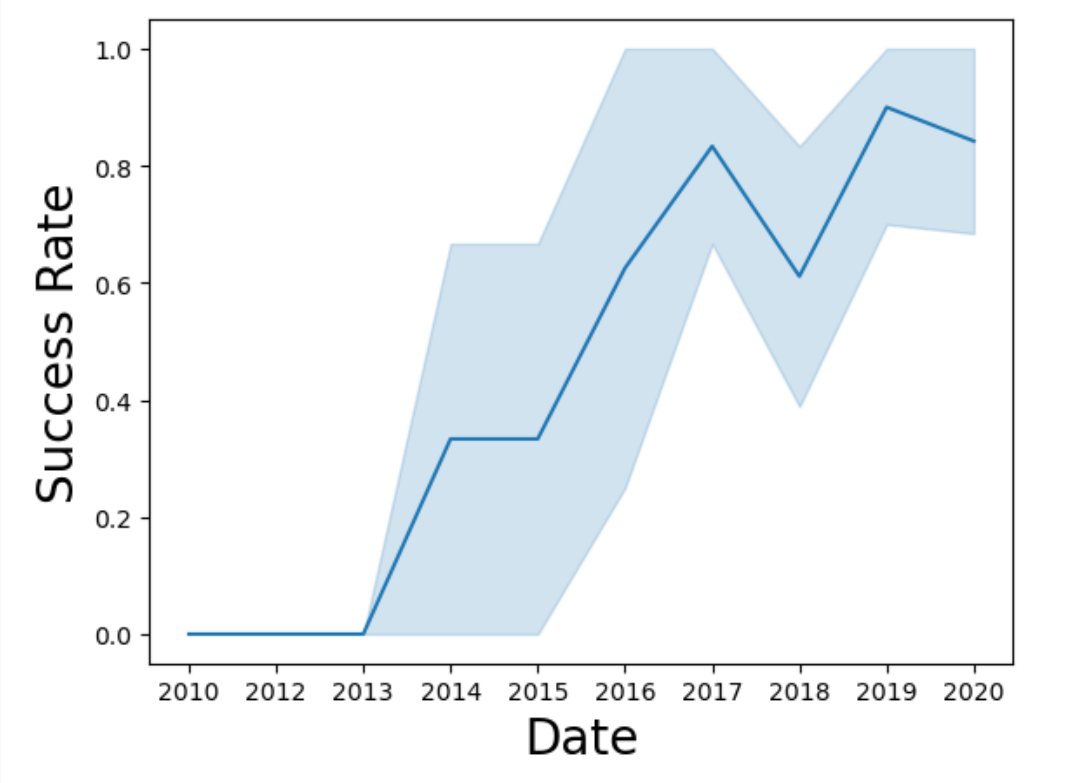
EDA with Data Visualisation

- Payload vs Orbit:
  - LEO & ISS orbit has better success with increase in Payload
  - GTO, VLEO & MEO relationship is inconclusive



EDA with Data Visualisation

- Launch Success Yearly Trend:
  - Steady increase in success rate since 2013



Results
EDA with SQL

EDA with SQL

- Names of the unique launch sites in the space mission
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Launch Site names Beginning with ‘CCA”, 5 samples.

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

EDA with SQL

- Total Payload Mass carried by boosters launched by NASA (CRS)
  - 45596
- Average Payload Mass by F9 v1.1
  - 2928
- Date when the first successful landing outcome in ground pad was achieved
  - 2015-12-22
- 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

EDA with SQL

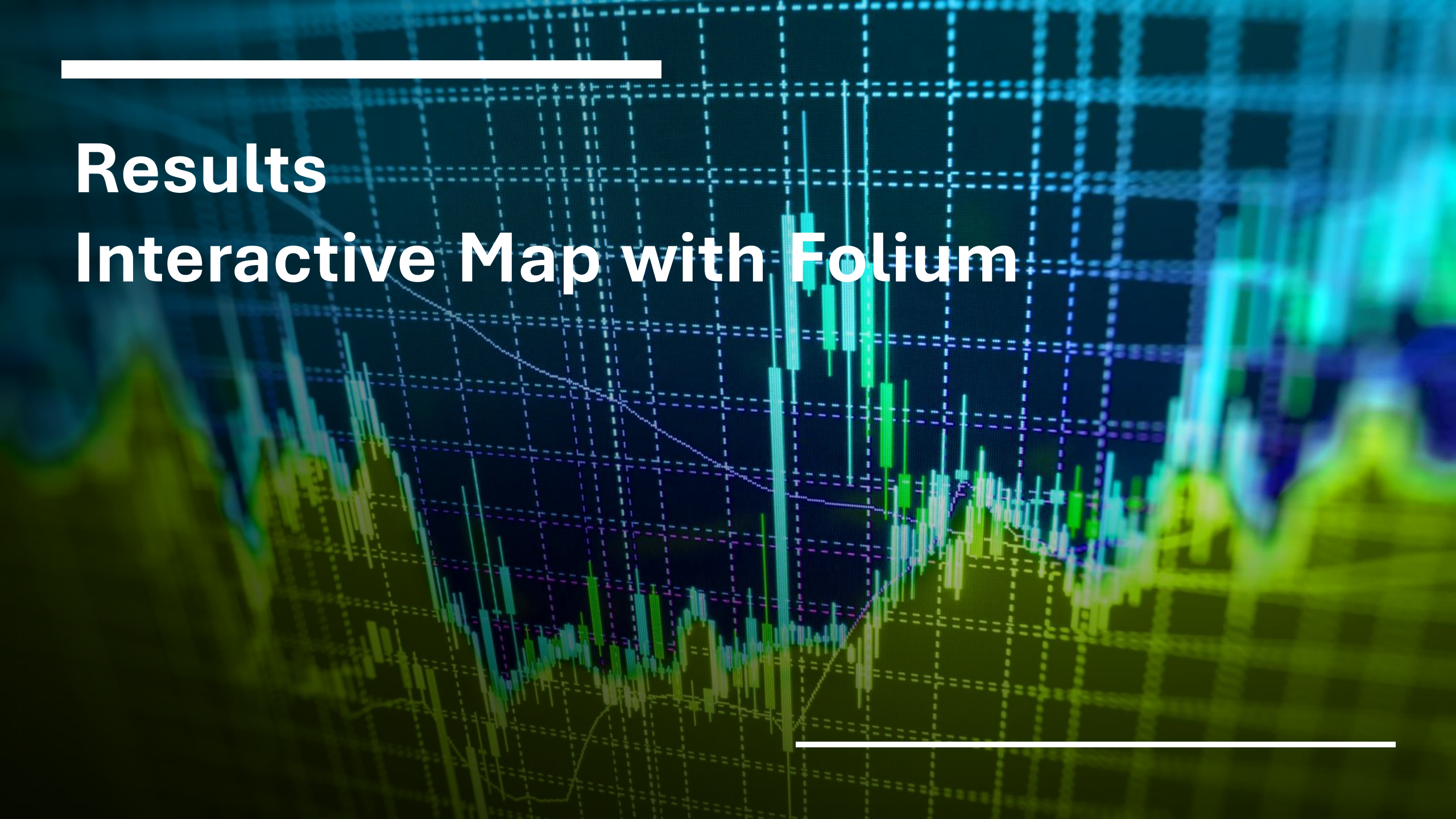
- Total number of successful and failure mission outcomes
  - 100
- The names of the booster_versions which have carried the maximum payload mass
  - booster_version
    - F9 B5 B1048.4
    - F9 B5 B1049.4
    - F9 B5 B1051.3
    - F9 B5 B1056.4
    - F9 B5 B1048.5
    - F9 B5 B1051.4
    - F9 B5 B1049.5
    - F9 B5 B1060.2
    - F9 B5 B1058.3
    - F9 B5 B1051.6
    - F9 B5 B1060.3
    - F9 B5 B1049.7

EDA with SQL

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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



Results
Interactive Map with Folium

An Interactive Map with Folium

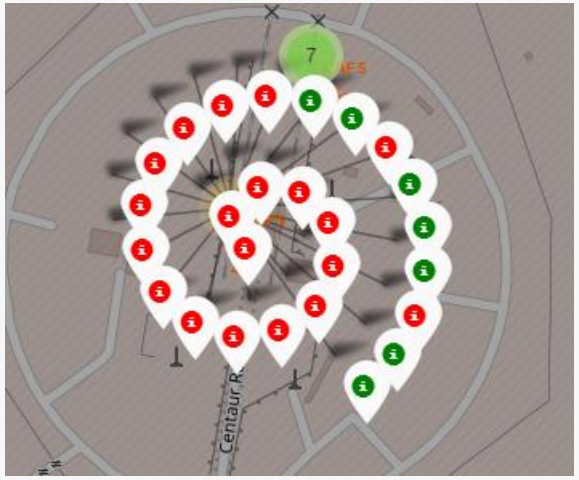
All Lauch Site on Folium Map

An Interactive Map with Folium

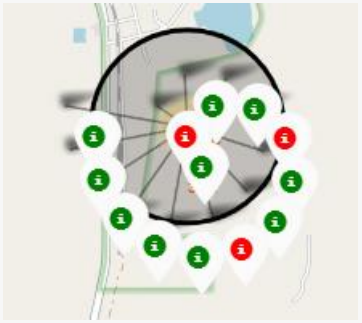
Launch Site color labelled records

CCAKS SLC-40

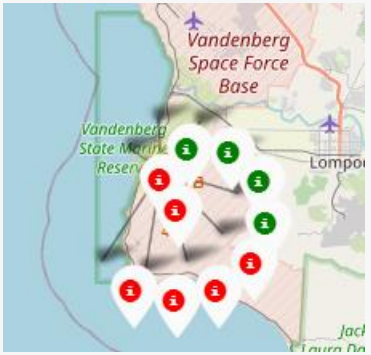
CCAFS LC-40



KSC LC-39A

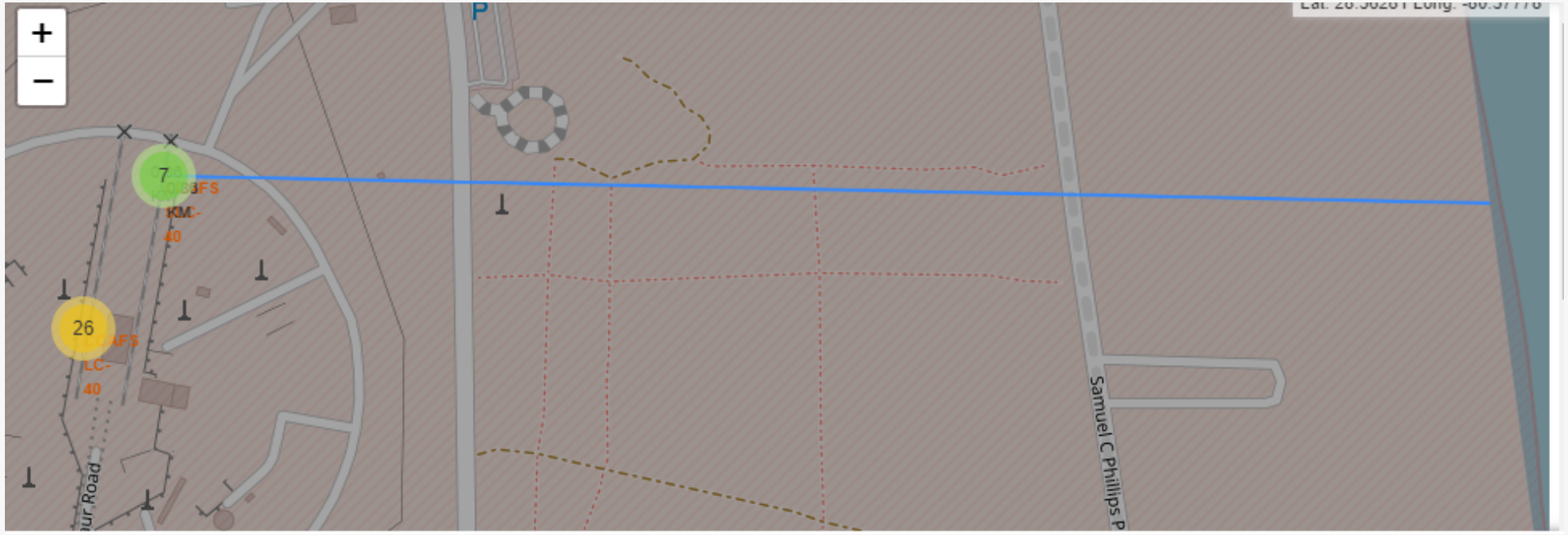


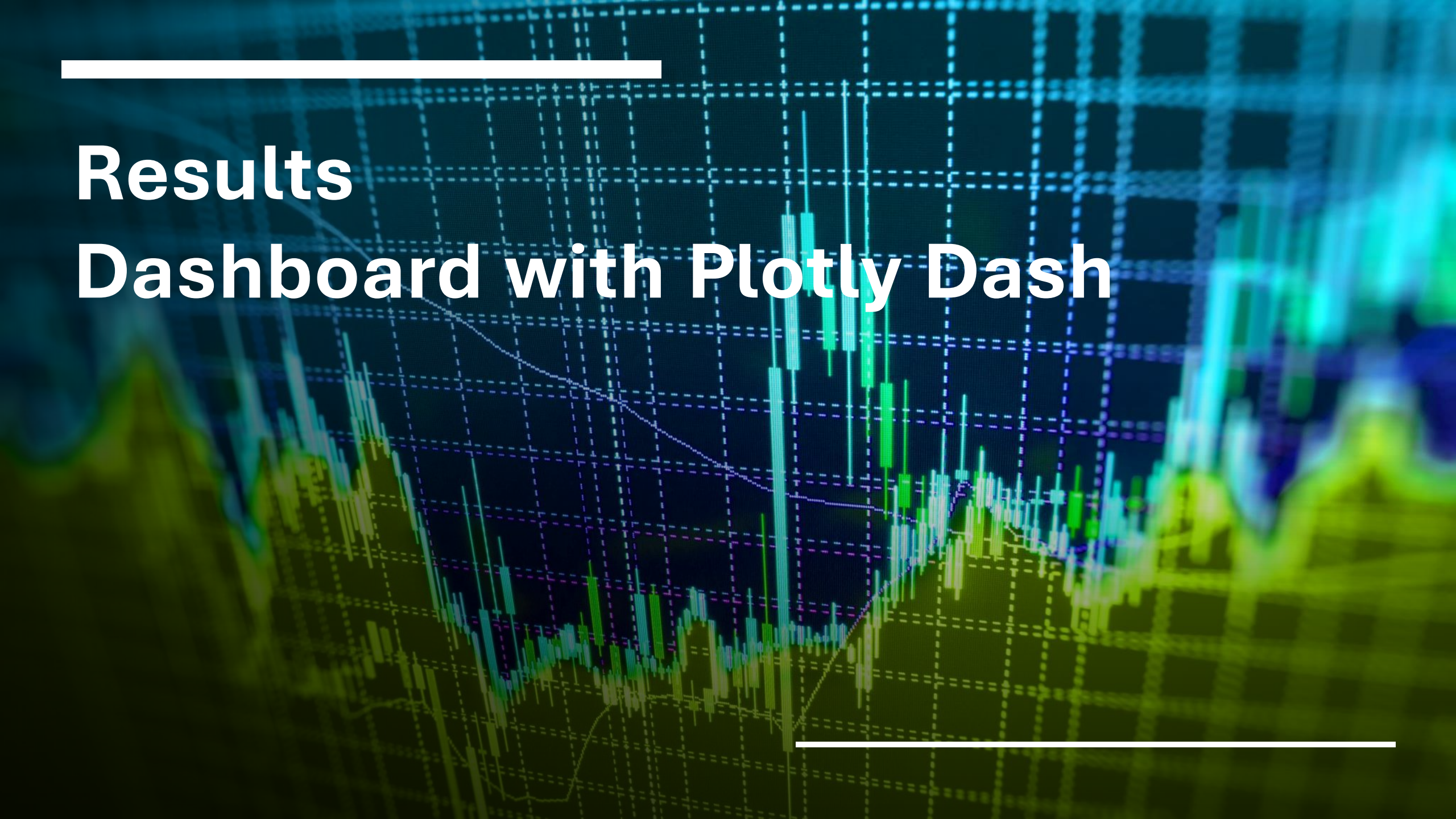
VAFB SLC-4E



An Interactive Map with Folium

Launch Site Distance from Coastlines

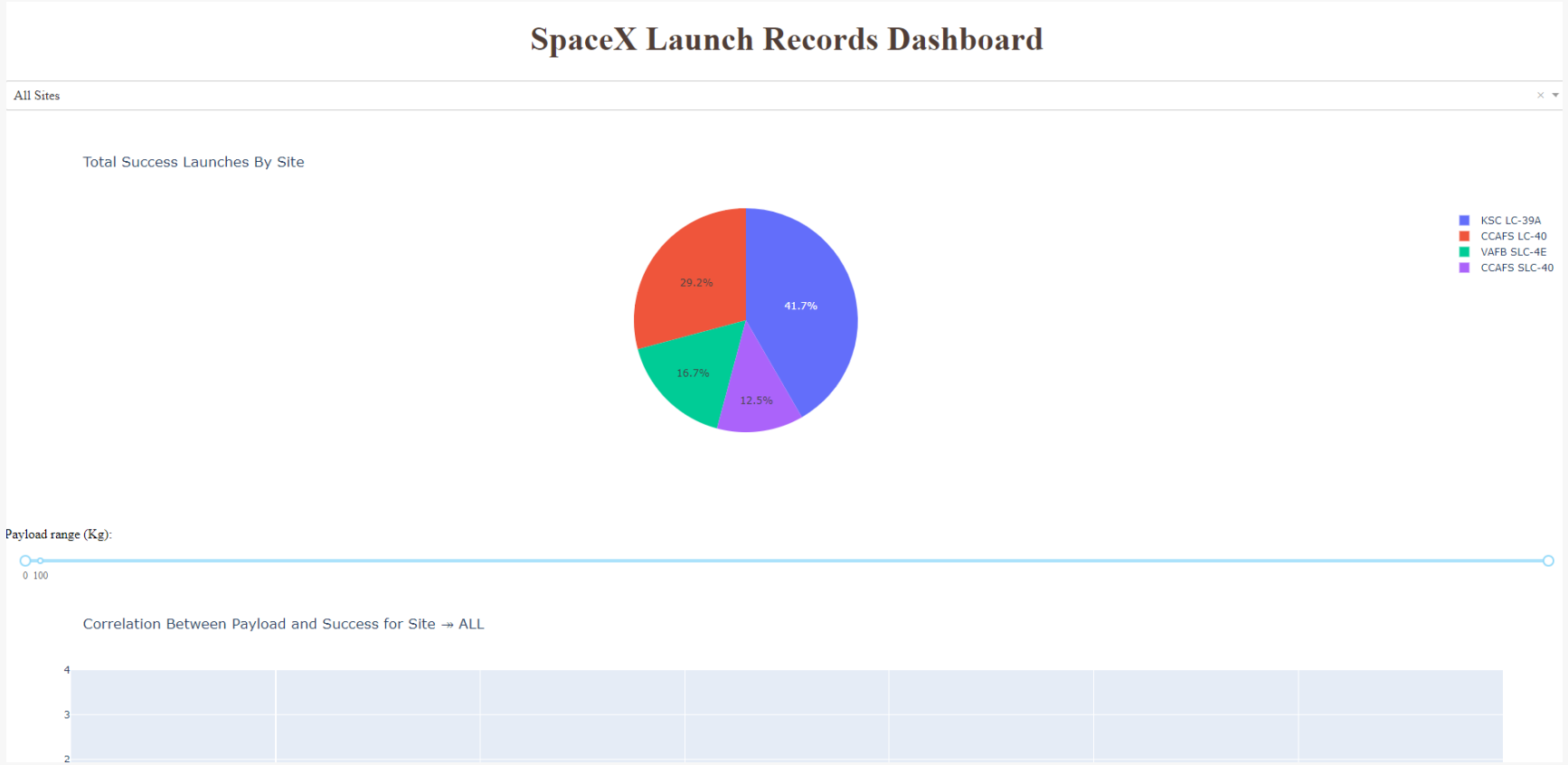




Results
Dashboard with Plotly Dash

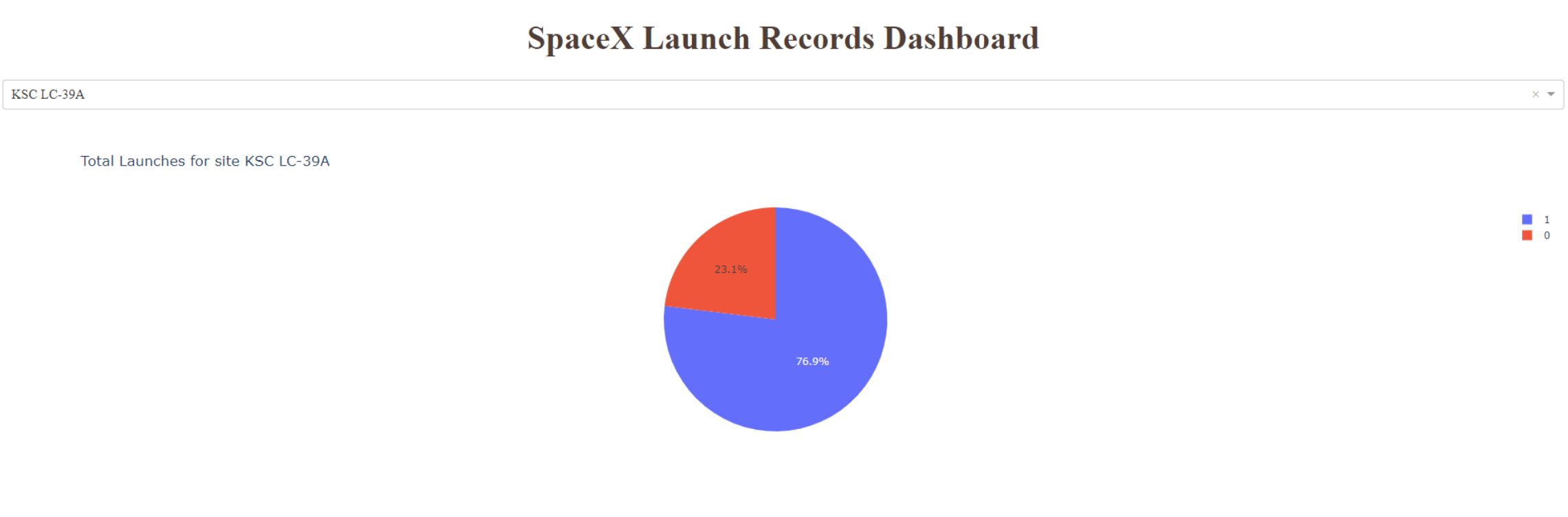
Build a Dashboard with Plotly Dash

Successful Launches by Site



Build a Dashboard with Plotly Dash

Highest Successful Launches Site KSC LC-39A



Results Predictive Analysis (Classification)

Predictive Analysis

Classification Accuracy

- Four classification models were tested, and their accuracies tabulated

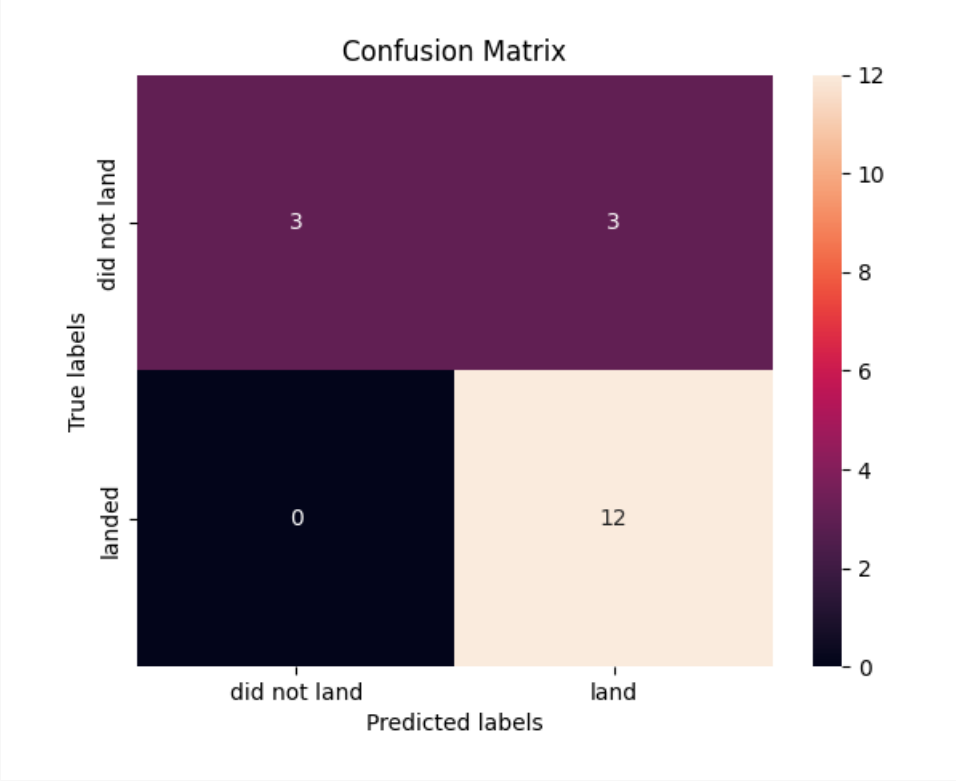
	Accuracy
Logistic Regression	0.846429
SVM	0.848214
KNN	0.848214
Decision Tree	0.875000

- Decision Tree Classifier was the model with the highest classification with an accuracy of 87.5%

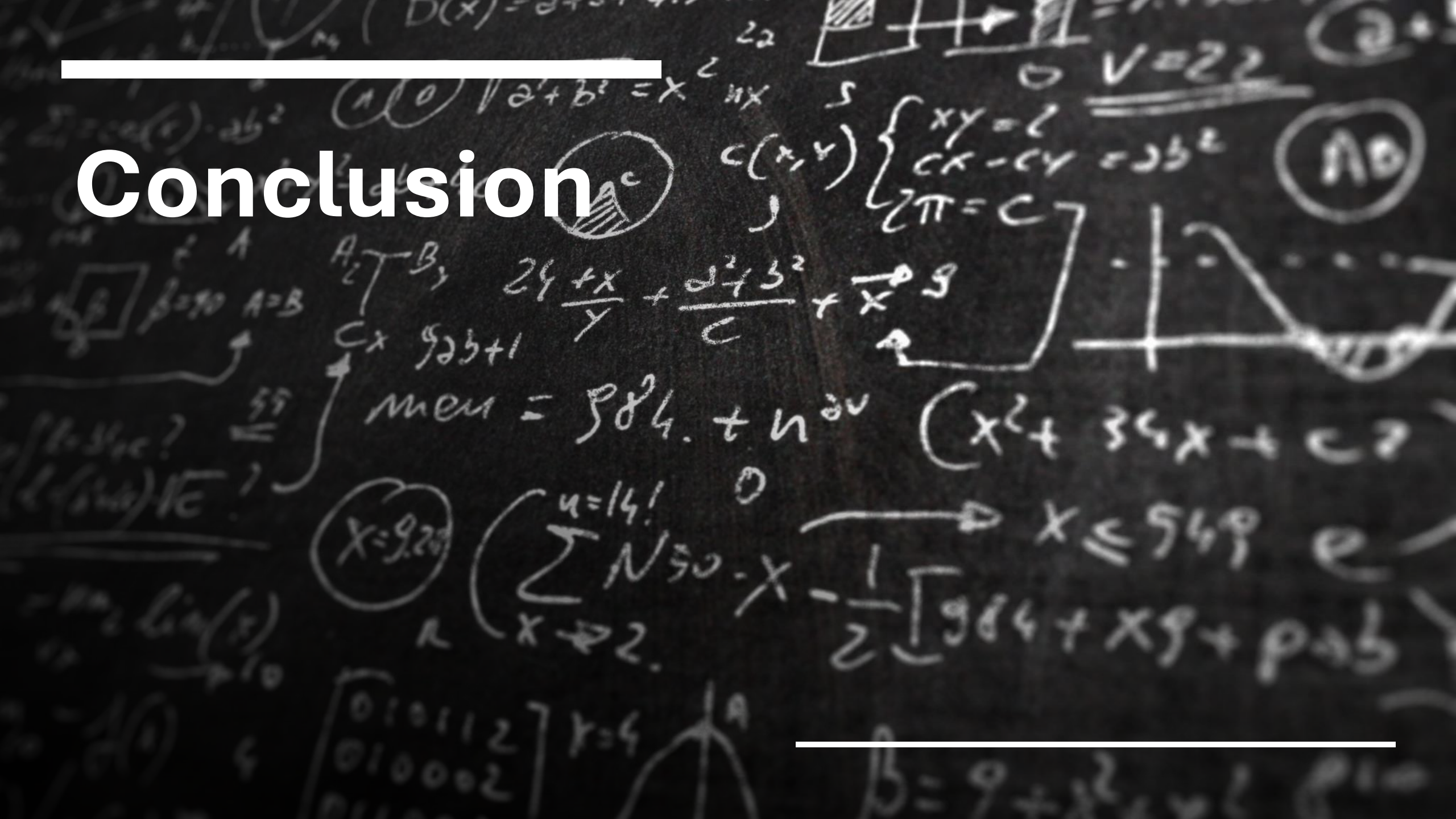
Predictive Analysis

Confusion Matrix of Decision Tree Classifier

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive. However, there are also false positive cases as well.



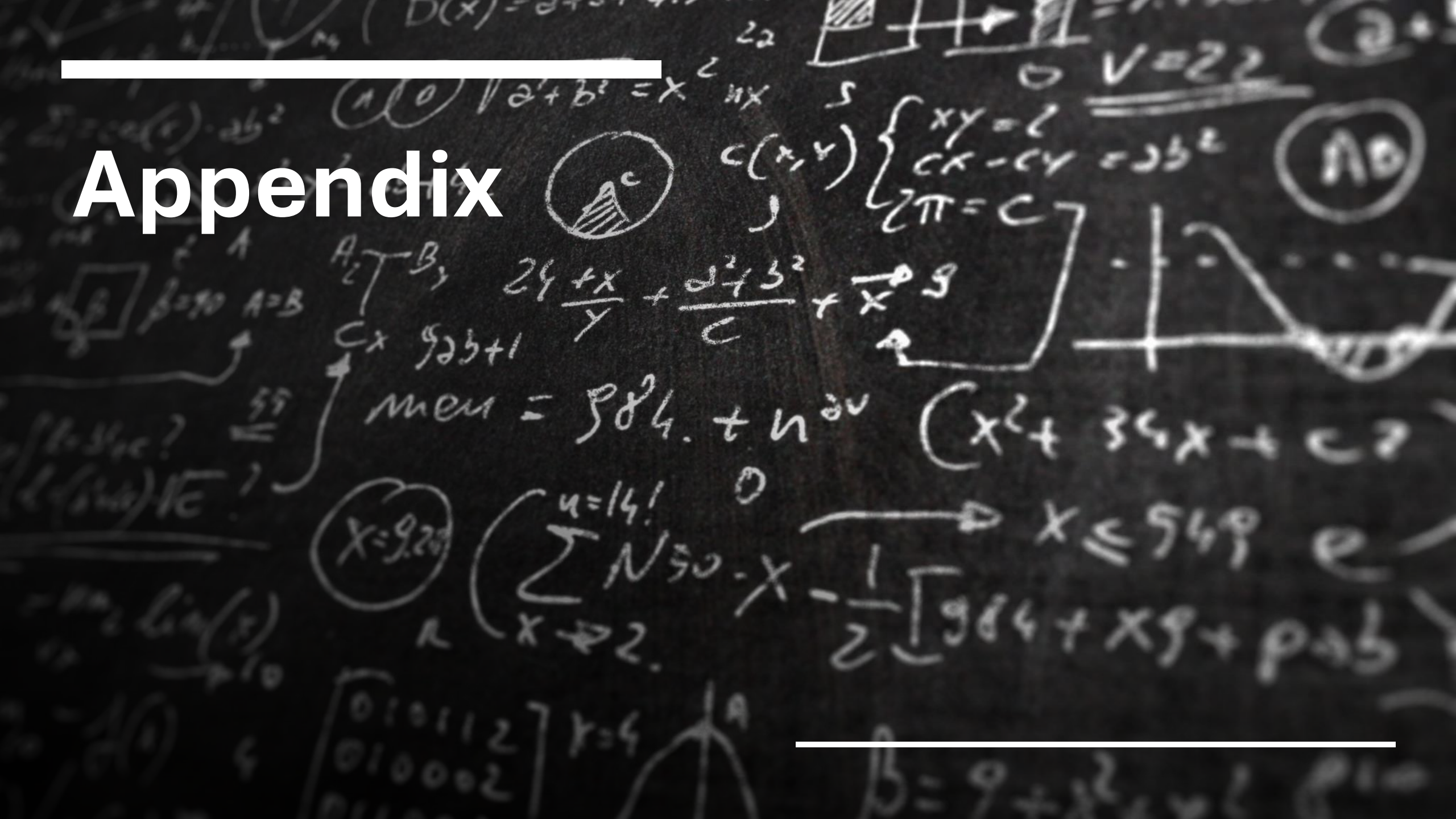
Conclusion



Conclusion

- Different data sources were analyzed. Orbits ES-L1, GEO, HEO, SSO has the highest success rates
- The best launch site is KSC LC-39A. Success rates for SpaceX launches has been increasing with increasing number of flights
- Launches above 7,000kg are less risky
- Decision Tree Classifier algorithm is the best Machine Learning Model for the dataset provided.

Appendix



Appendix

- Great experience in the capstone project.
- It encapsulated the whole courses into one and worked on real world data