IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Imran Chowdhury Dipto
28-02-2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API calls and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using Visualization and SQL
  - Interactive Visual Analytics using Folium and Dash
  - Predictive Analysis using Classification
- Summary of all results
  - Visualization of the key findings

Introduction

- Background
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars
  - Other companies cost upward of 165 million dollars each
  - SpaceX can save a high amount because they can reuse the first stage.
- Aims
  - To determine if the first stage stage will land which will help us determine the cos of a launch
  - Help a different company bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Called the SpaceX API and stored the response
  - Converted the response into a JSON format to decode the data
  - Chosen relevant features of the dataset
  - Created several helper function to extract the required data from the API
  - Created lists and applied the function to store the data in them
  - Cleaned the dataframe by replacing missing values
  - Created a dataframe from the lists and saved it as a CSV file for further analysis

Methodology

Executive Summary

- Data Wrangling
  - Performed Exploratory Data Analysis (EDA) to identify missing values and column types
  - Used EDA to Identify
    - Number of Launches on Each Site
    - Number of Occurrence of Each Orbit
    - Number and Occurrence of mission outcome of the orbits

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization
  - Used visualisations to identify the relationships between
    - Payload Mass and Flight Number
    - Launch Site and Flight Number
    - Success Rate by Type of Orbit
    - Flight Number Vs Orbit Type
    - Payload Mass and Orbit
    - Success Rate over the years

Methodology

Executive Summary

- **Exploratory Data Analysis (EDA) using SQL to perform 10 tasks and some of them are as follows**;
- Performed exploratory data analysis (EDA) using SQL to perfor 10 tasks and some of them are as follows:
- Display the names of the unique launch sites
- Display 5 launch sites starting with ‘CCA’
- Total Payload Mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved

Methodology

Executive Summary

- Interactive Visual Analytics using Folium
    - Marked the Launch Sites on a Map
    - Marked the success/failed launches for each site on the map
    - Calculated the distances between a launch site to it’s proximities and displayed the outcomes on the map

Methodology

Executive Summary

- Interactive Visual Analytics using Plotly Dash
  - Created an interactive Visual analytics dashboard
  - Allowed users to see visuals based on successful launches on each sites by a pie chart
  - Has a slider that allows users to select payload mass which acts as a filter for a scatterplot displayed
  - The scatterplot shows correlation between Payload Mass and Success Rate of the launch based on Booster Version Categories
  - Users have the option of selecting a specific site to filter the visuals

Methodology

Executive Summary

- Interactive Visual Analytics using Plotly Dash
  - Created an interactive Visual analytics dashboard
  - Allowed users to see visuals based on successful launches on each sites by a pie chart
  - Has a slider that allows users to select payload mass which acts as a filter for a scatterplot displayed
  - The scatterplot shows correlation between Payload Mass and Success Rate of the launch based on Booster Version Categories
  - Users have the option of selecting a specific site to filter the visuals

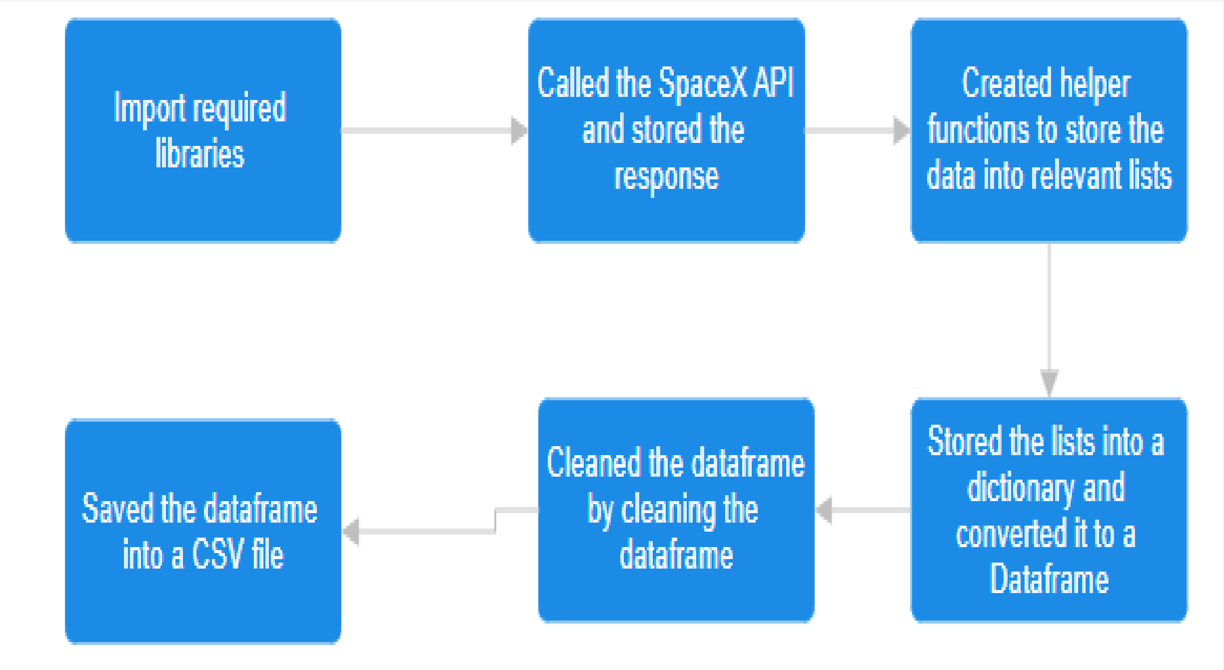
Methodology

Executive Summary

- Predictive Analysis using Classification Models
  - Stored relevant parts of the data into feature and target sets
  - Standardized the feature sets
  - Split the feature and targets sets into training and test sets
  - Trained 4 different classifiers on the train set by building the models and trained them using extensive hyperparameter tuning using GridSearchCV to identify the best parameters for the models
  - Tested the trained models on the test set and produced results by means of tables and charts

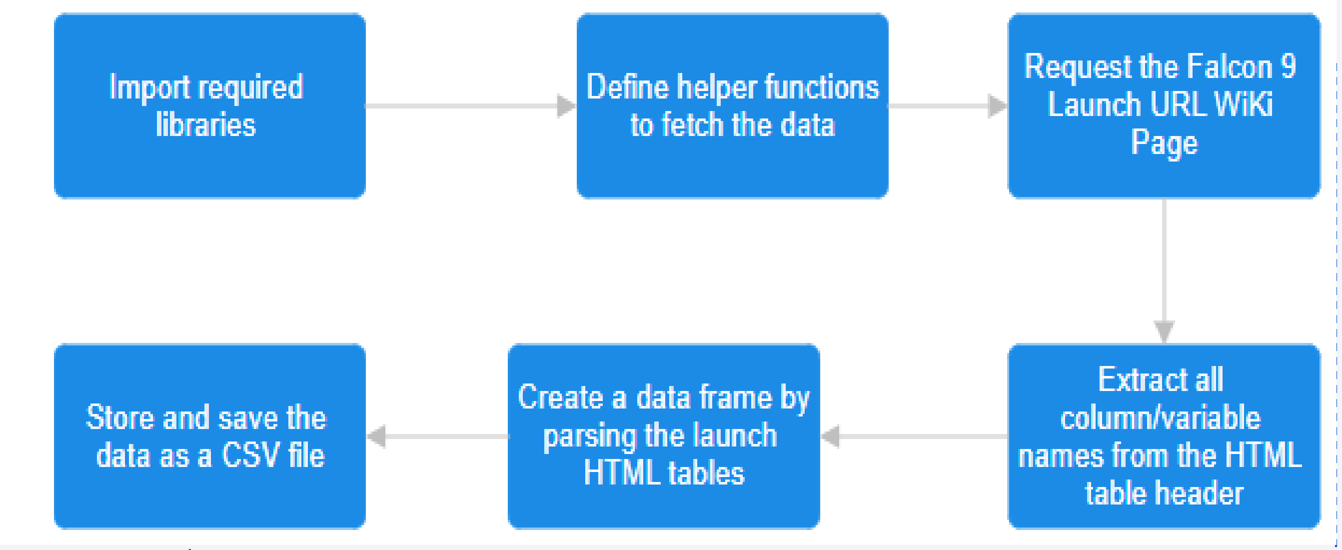
Data Collection – SpaceX API

- Data Collection – SpaceX API Process
- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/SpaceX%20Data%20Collection%20API.ipynb

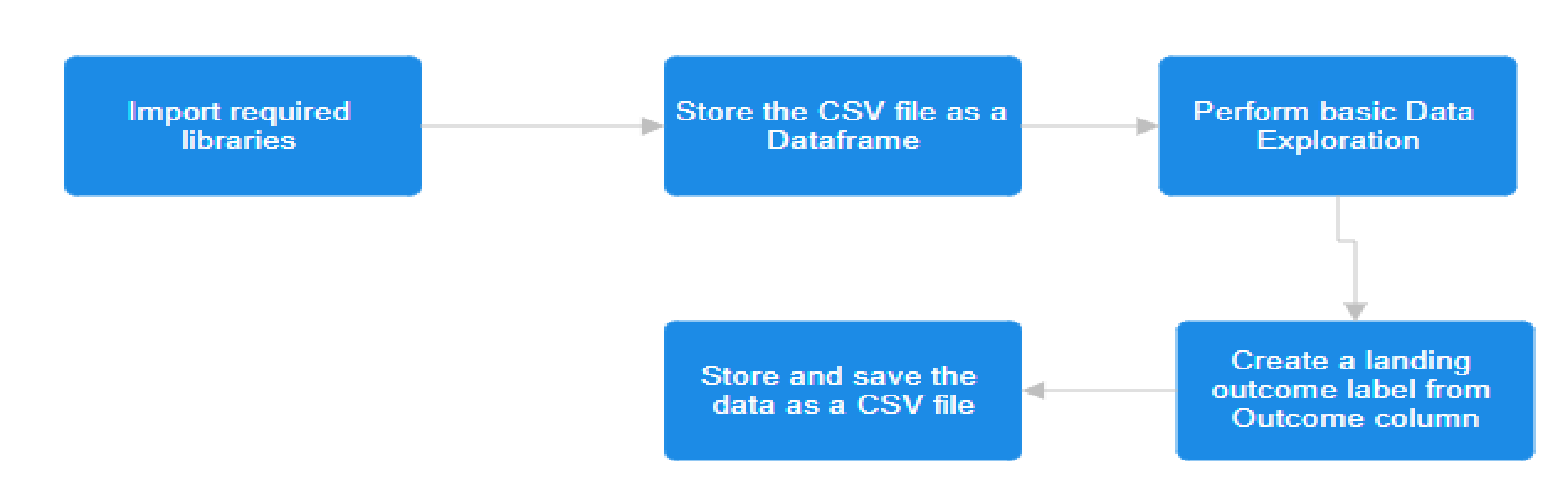


Data Collection - Scraping

- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/Data%20Collection%20with%20Webscraping.ipynb



Data Wrangling



- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/SpaceX%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

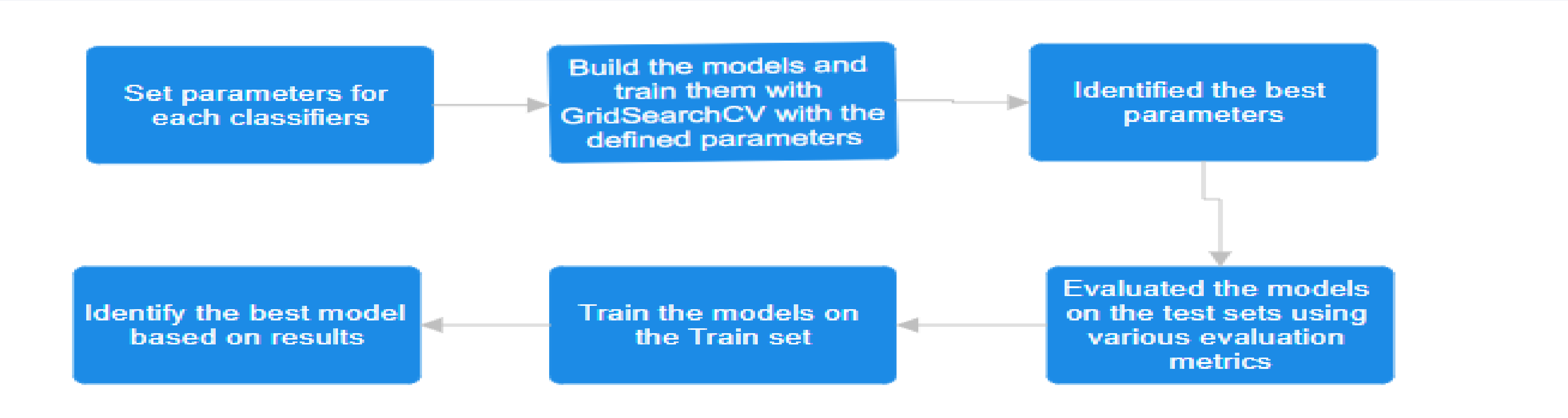
- Marker: To mark all launch sites on the map
- Circle: To highlight circle area with a text label on a specific coordinate
- Poly Line: To show a launch site to the selected coastline point

- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/Launch%20Site%20Location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/SpaceX%20Launch%20Records%20Dashboard.py

Predictive Analysis (Classification)



- Link: https://github.com/imranchy/IBM_DataScience_Capstone_Project/blob/main/Predictive%20Analytics%20(Classification).ipynb

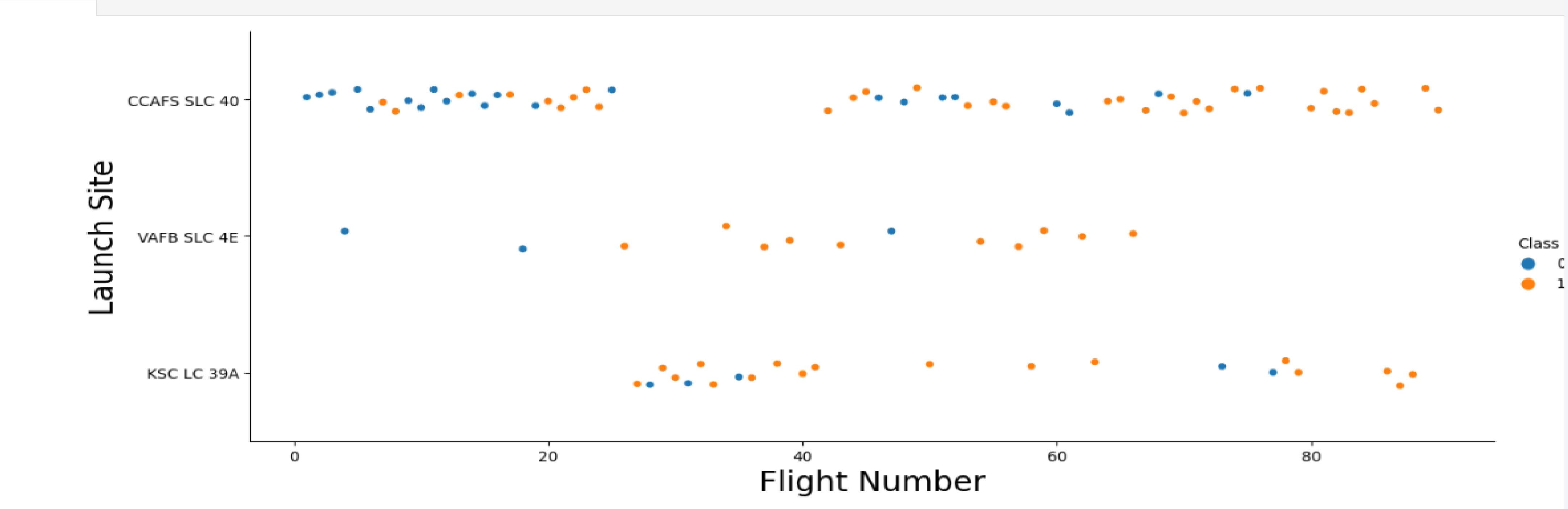
Results

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

Section 2

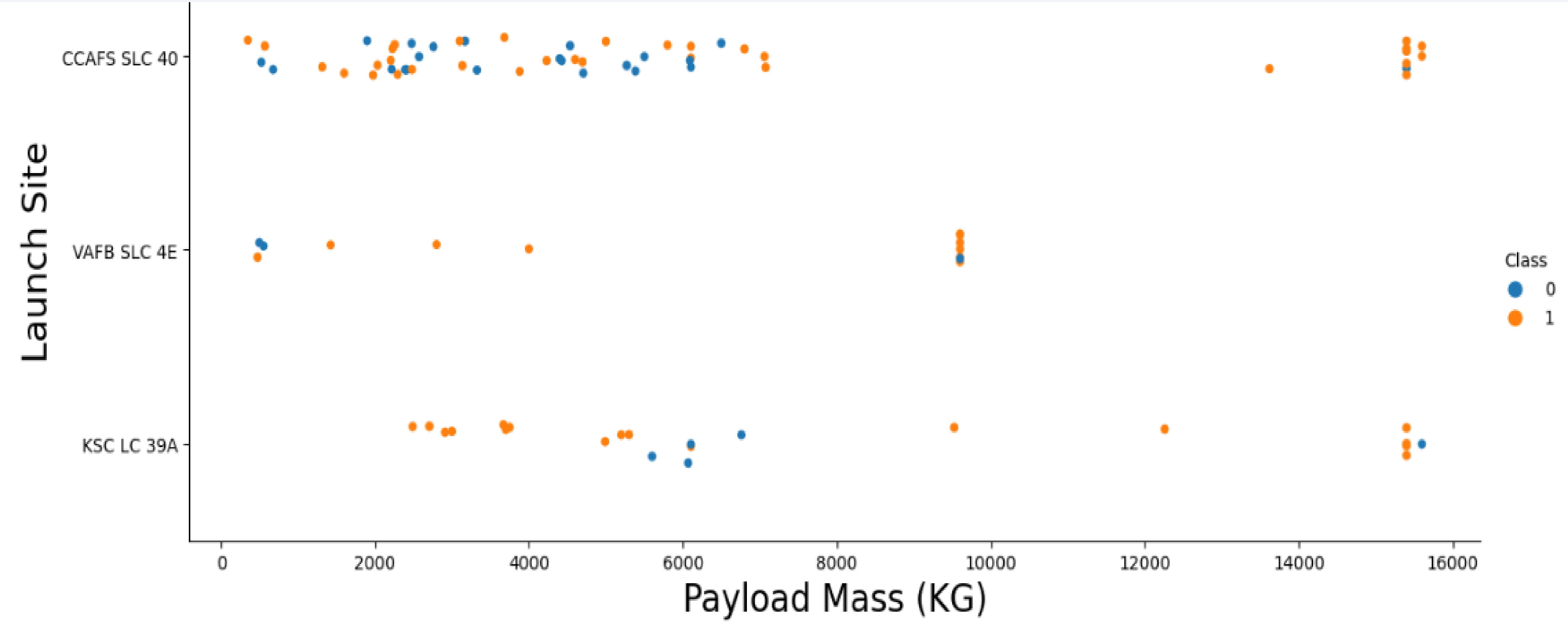
Insights drawn from EDA

Flight Number vs. Launch Site



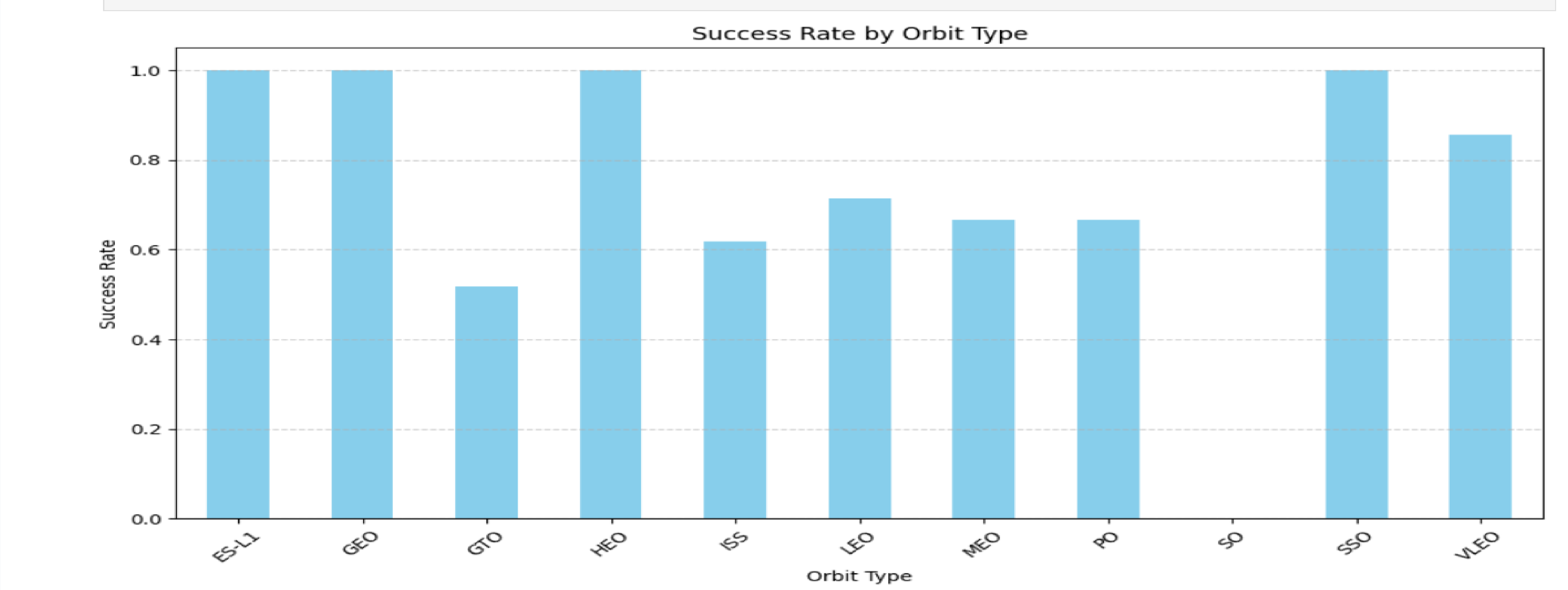
- As the flight number increases we can see a higher number of success rate for CCAFS SLC 40

Payload vs. Launch Site



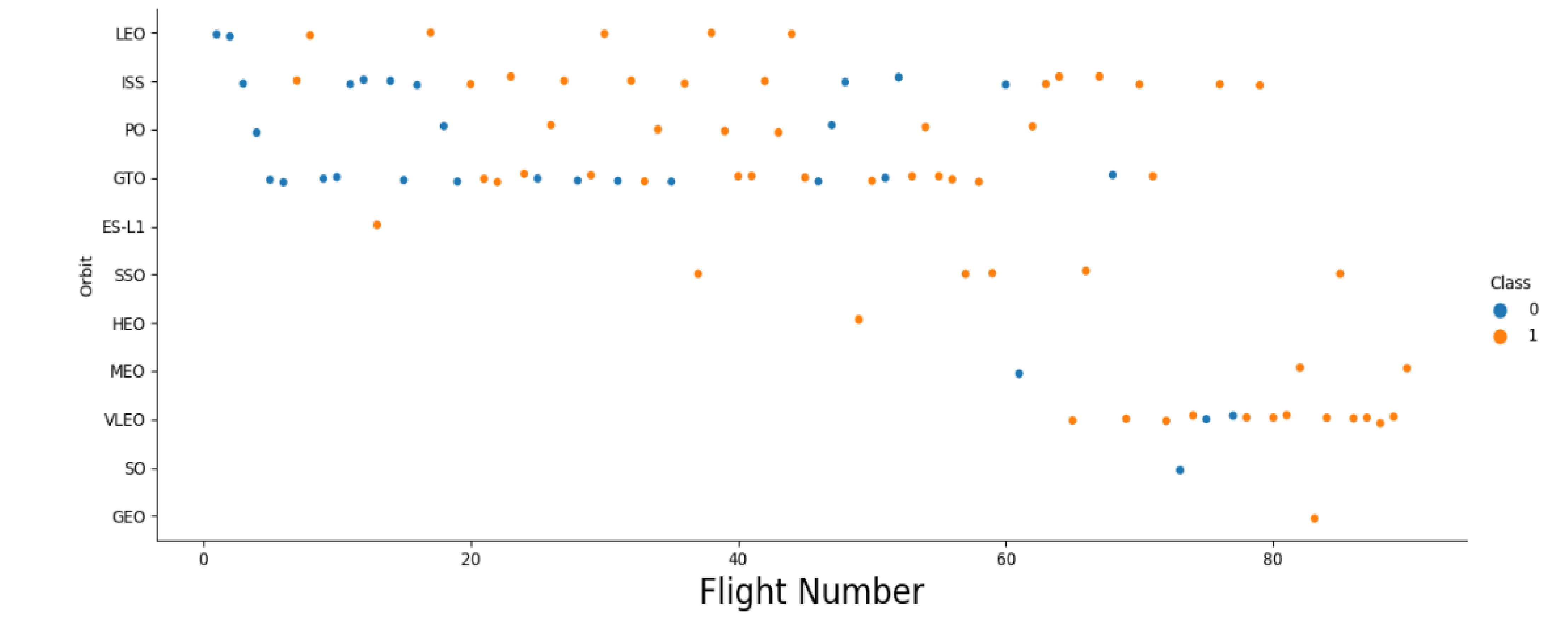
- With Payload Mass greater than 10,000 we see that VAFB SLC 4E has no rockets launched

Success Rate vs. Orbit Type



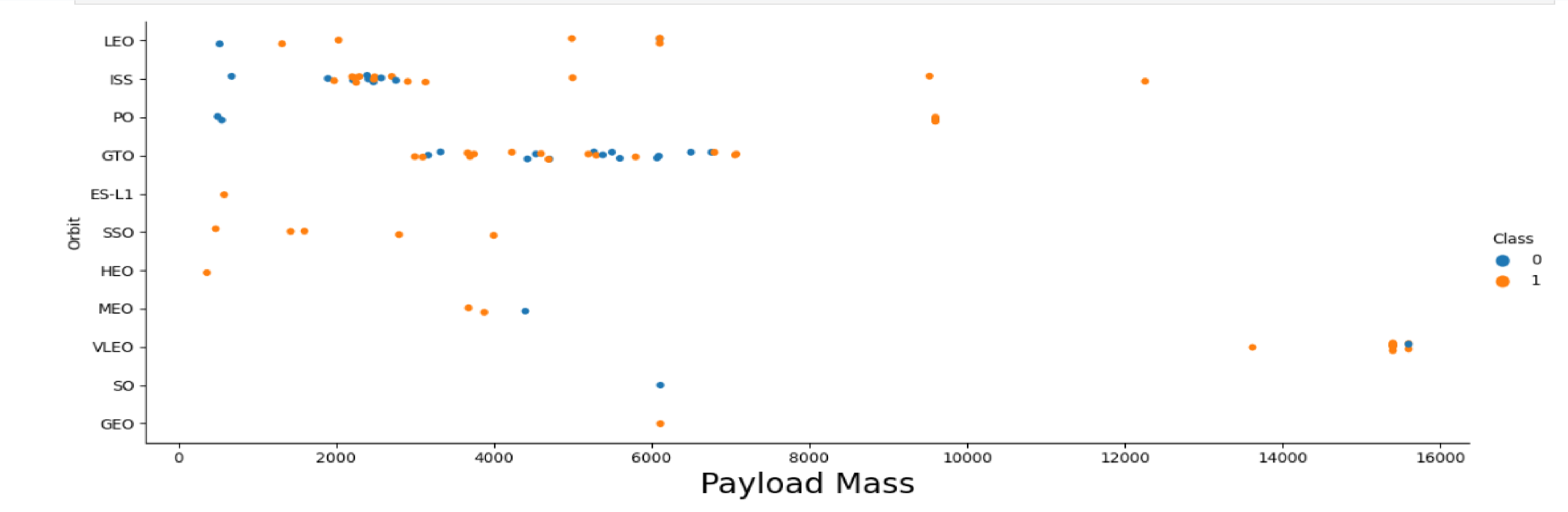
- Orbit types ES-L1, GEO, HEO and SSO have the highest success rates

Flight Number vs. Orbit Type



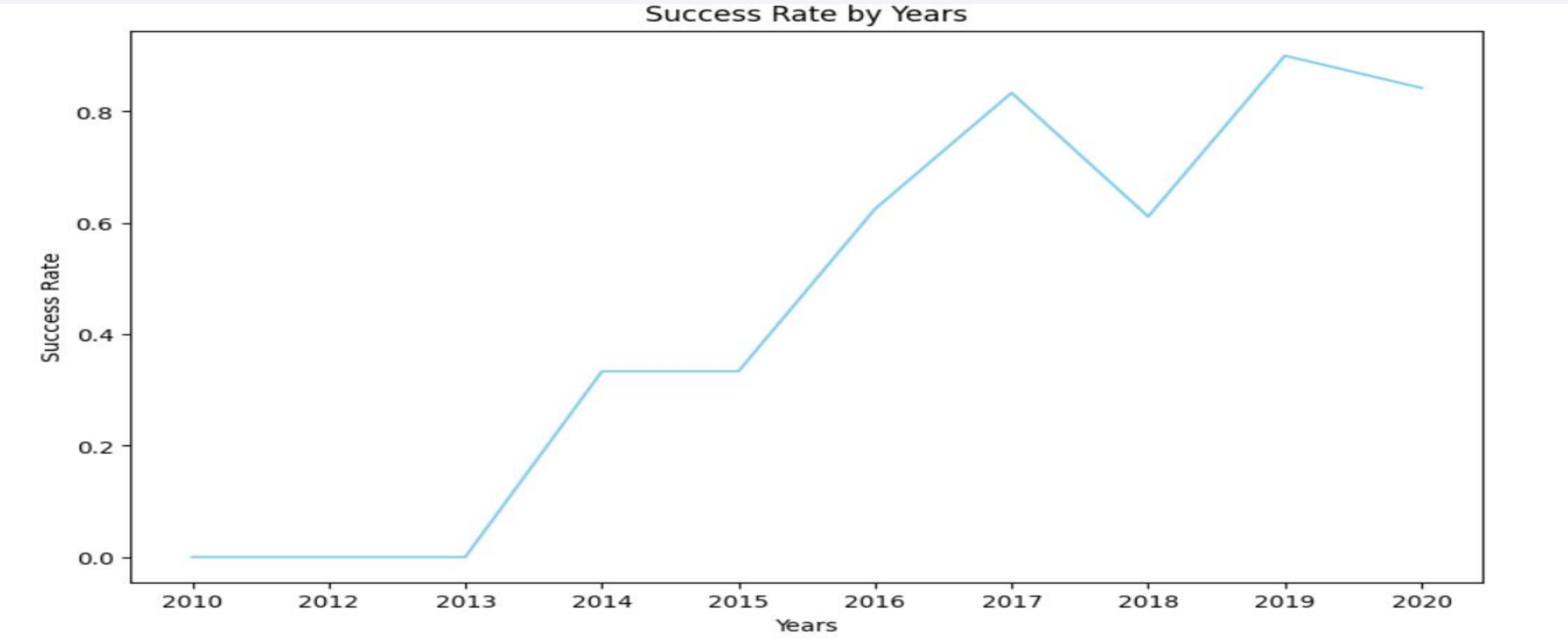
- With LEO success rate increases with number of flights

Payload vs. Orbit Type



- With heavy payloads the success rate for landing are more for Polar, LEO and ISS

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

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

Total Payload Mass

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

Average Payload Mass
2534.6666666666665

First Successful Ground Landing Date

First Successful Landing
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

Total Number of Successful and Failure Mission Outcomes

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

Boosters Carried Maximum Payload

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

2015 Launch Records

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

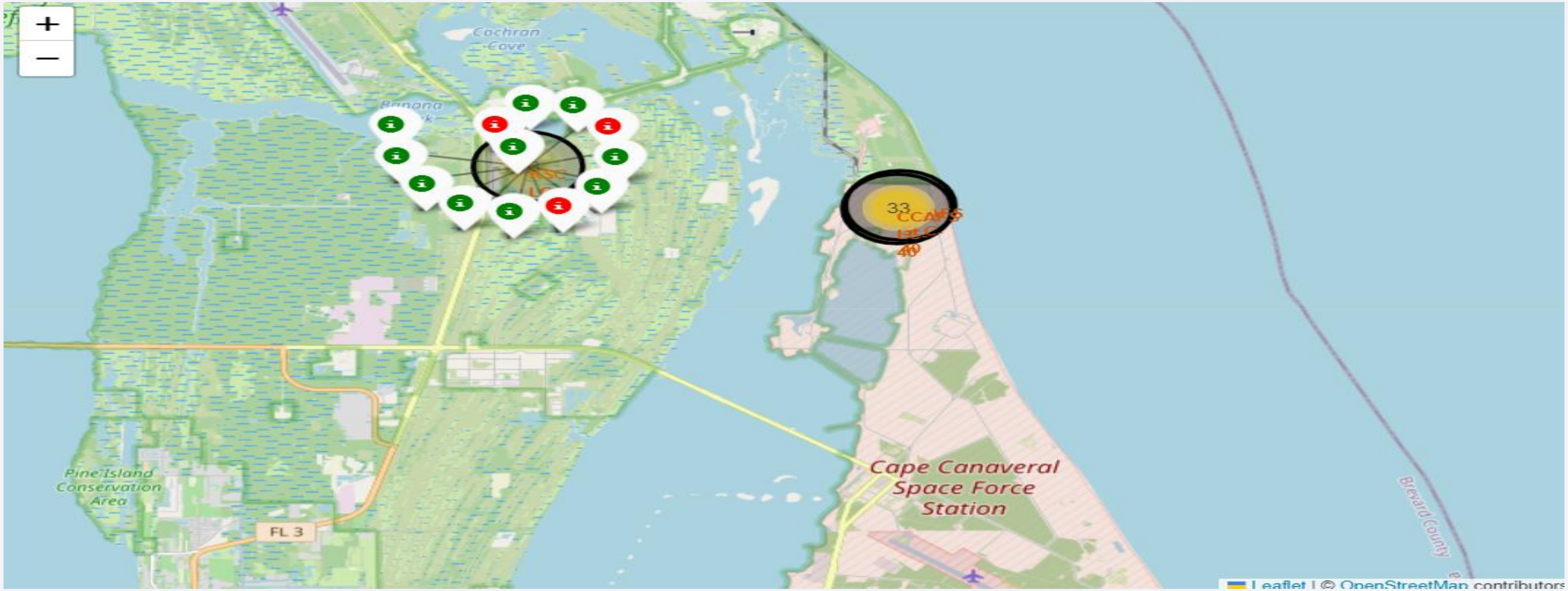
Section 3

Launch Sites Proximities Analysis

All Launch Sites on Map

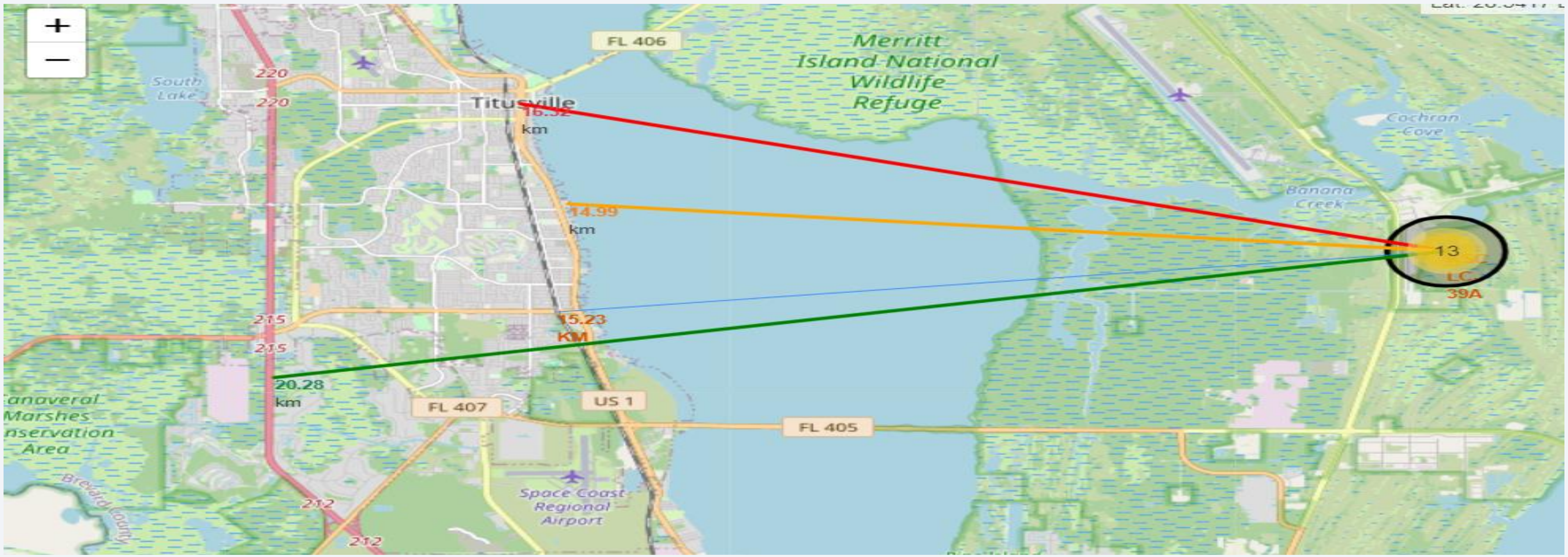
- It seems like the Launch sites are close to the Coast Line and the equator

Landing Outcomes marked on Map



- This visualisation will help us understand which sites has the highest success rates

Distance between a Launch Site and it's Proximities

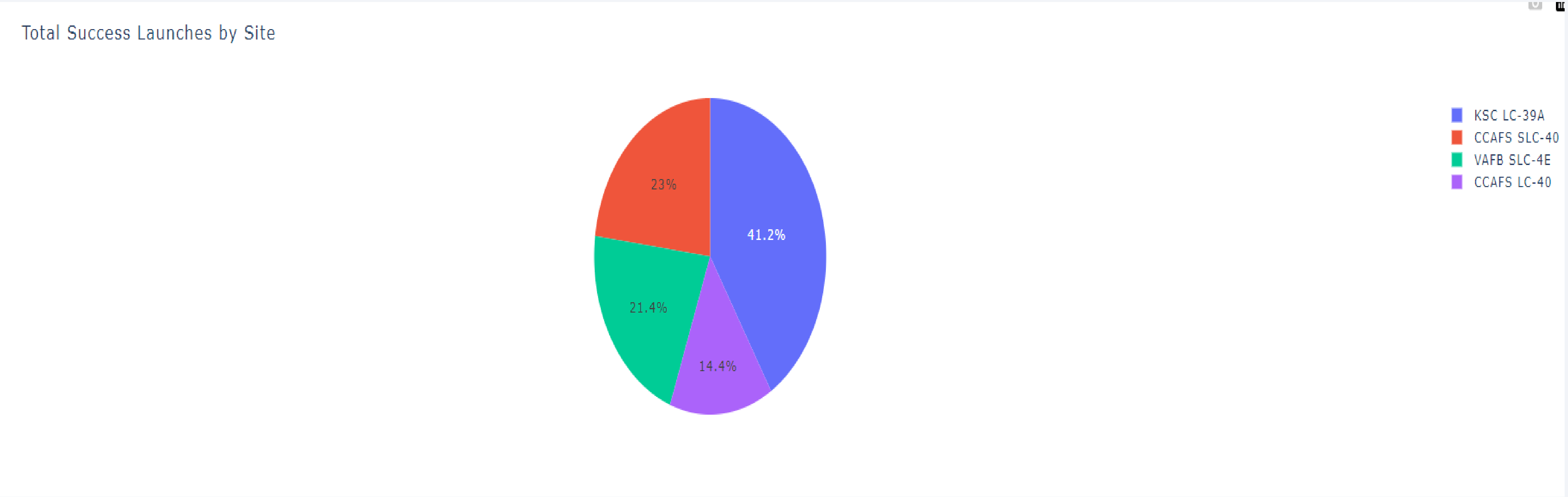


- KSC LC-39A is closer to the coastline

Section 4

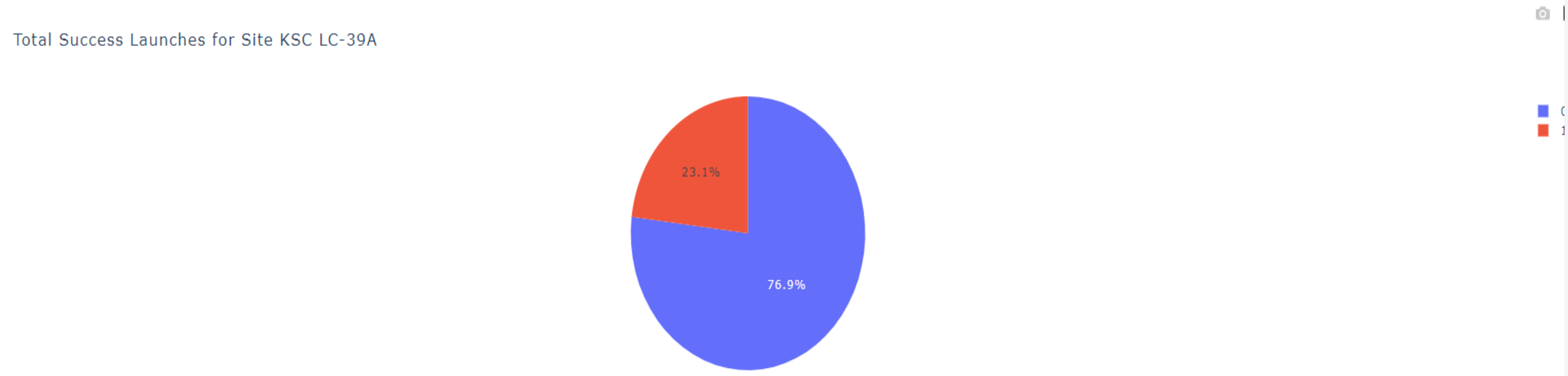
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



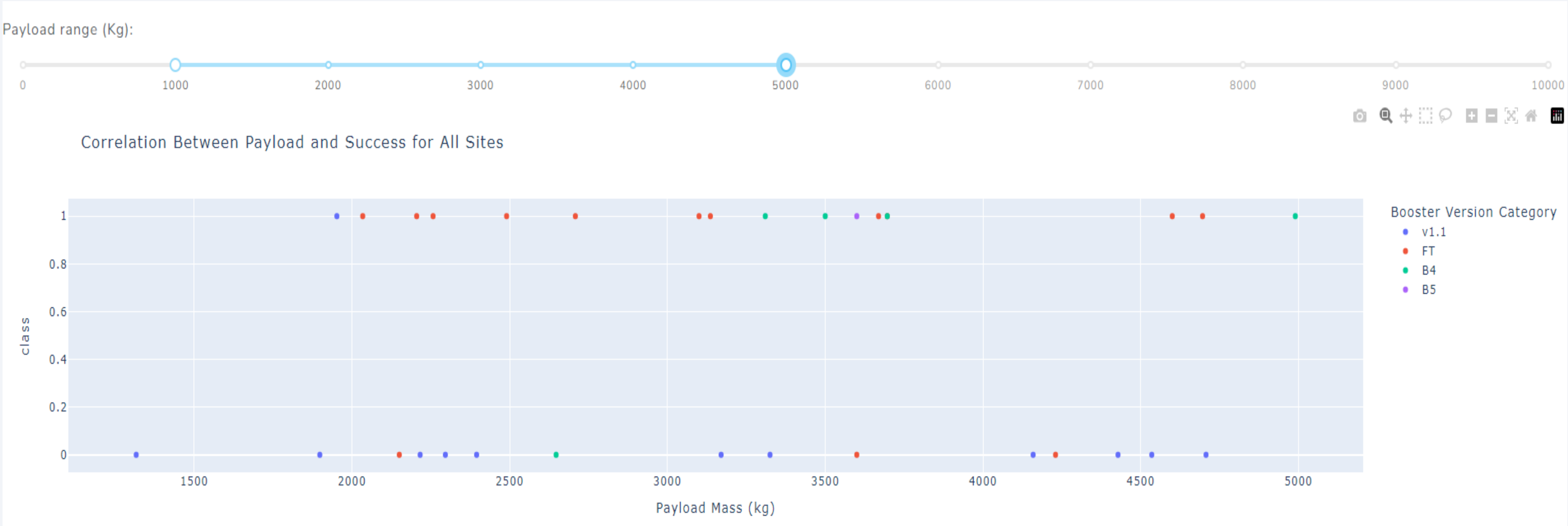
- Highest success rate is seen when rockets were launched from KSC LC-39A

Highest Success Launch



- The highest success rate is about 77% for this site

<Dashboard Screenshot 3>

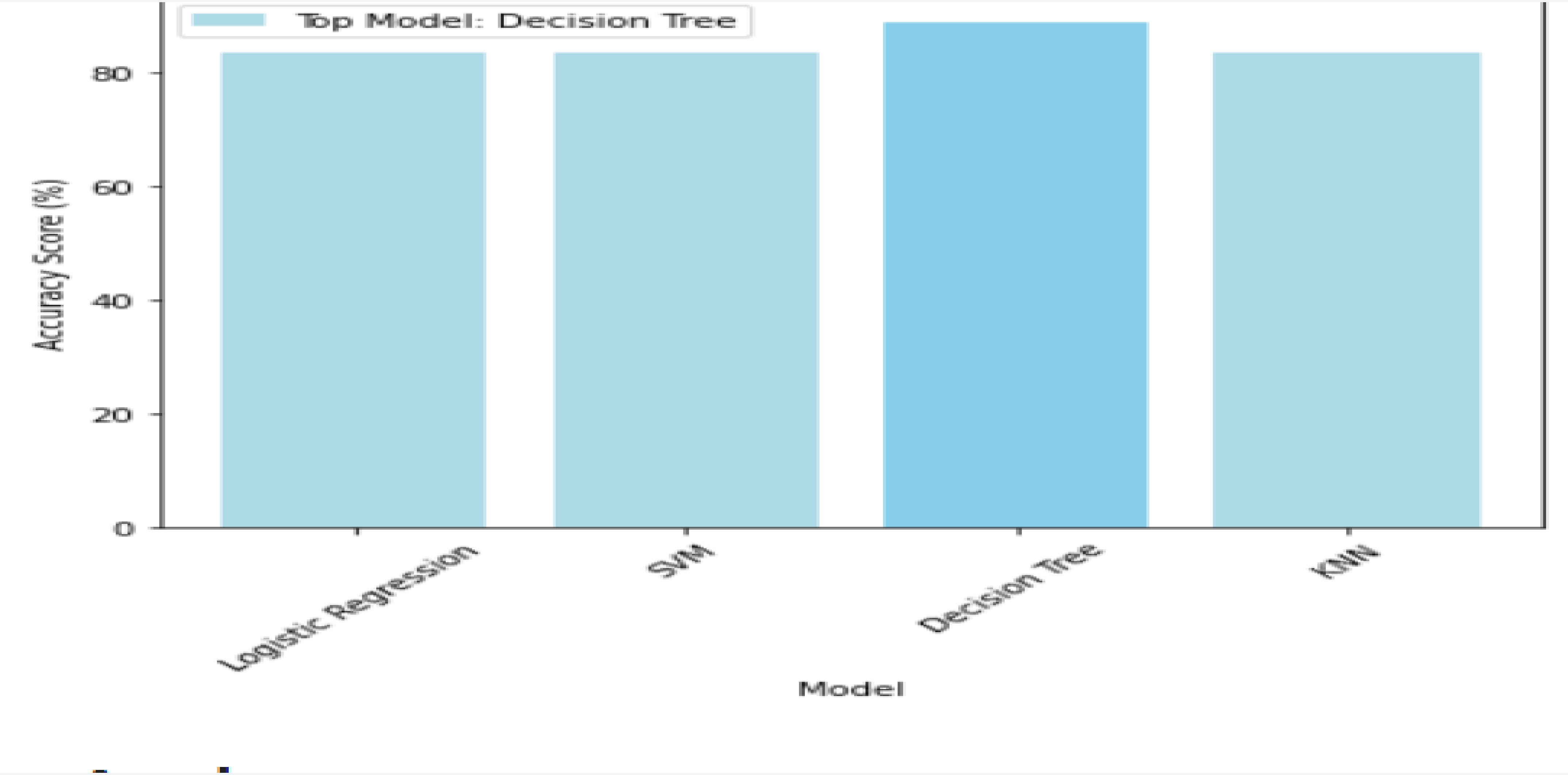


- For the Payload Mass between 1000 to 5000 rockets launched with FT has the highest success rate

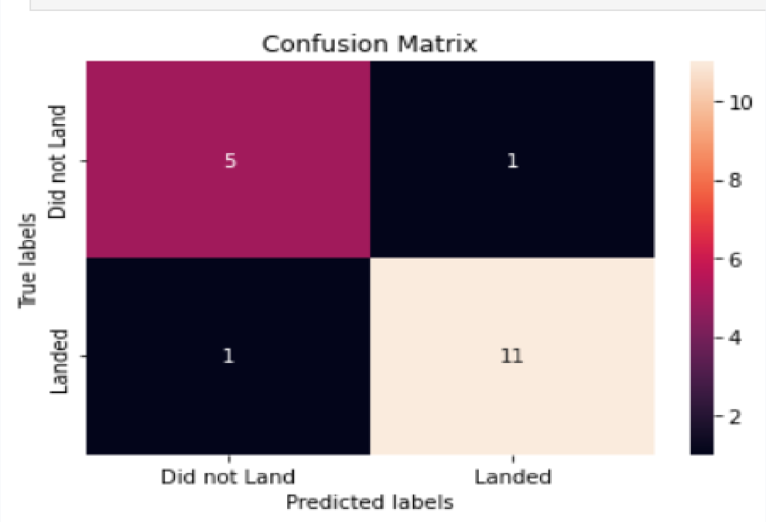
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



- The confusion matrix of Decision Tree shows only 1 record where true label is “Did not Land” where predicted label is “Landed” and vise versa. Whereas, it has been able successfully predict 11 instances of “Landed” and 5 instances of “Did not Land” cases on the test data. This means this model has the highest accuracy among all models and the lowest error of missclassification

Conclusions

- Decision Tree is the ideal model for outcome classification
- The best hyperparameters found are; criterion: gini, max_depth = 12, max_features = log2, min_samples_leaf=2, min_samples_split = 2, splitter = best
- Test accuracy score achieved = 88.89%
- Compared to other models the way Decision Tree model works can be easily explained to the stakeholders

Future Recommendations

- Although the Decision Tree model has the highest accuracy on the test data, Logistic Regression has a higher accuracy score on the train data which leaves for further investigation
- It would be ideal to work with a larger dataset as this could be due to more data being taken for the training of the models
- It would be ideal to save and deploy the ideal model into production for the end users in a web based application

Appendix

- Python (Numpy, Pandas, Matplotlib, Seaborn, Plotly, Folium, BeautifulSoup, Dash, Scikit-Learn)
- SQL for querying the dataset for EDA
- Request for calling the API
- Jupyter Notebook for creating the shareable documents
- Github to store and share the notebook links
- Anaconda: Environment used to run the Dash application on the local machine to display the app functionalities

Thank you!