IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wei Wang
Feb 17, 2025

Outline

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

Executive Summary

- Today, SpaceX is considered on of the best companies over the world in terms of space rockets launches, due to the rocket science advancements that achieved in making space missions more affordable and practical, generally space rockets companies offer a rocket launch with a cost of 165 million dollars, whereas SpaceX offers by the same service for only 62 million dollars which considered a huge savings led organizations such NASA to sign contracts with SpaceX.
- In this report, I am taking the role of a data scientist working for a new rocket company, called SpaceY that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. my job is to use data science instead of rocket science to discover the possibility of competing with SpaceX. I am doing this by gathering information about SpaceX, performing data analytics, modeling ML algorithms and creating dashboards for my team

Introduction

- Since 1957, the countries around the world are competing to expand beyond Earth, whether through satellites launches or space exploration, these missions require huge amounts of money where a launch of one space rocket costs in average 165 million dollars, a company like Space X changes the equation by reducing this amount of money massively to only 60 million dollars due to its unique and advanced technologies in returning the first stage of rocket structure
- As mentioned above one of the key factor of SpaceX's success in the space race is the first stage of rockets return through a safe landing, from this point we will discover together what are the main attributes or variables that control these successful landings, through asking questions about the nature of the launch process , the payload mass of rocket, launch location, rocket orbit, and more by analyzing and visualizing these information through the data science methodology provided by IBM

Section 1

Methodology

Methodology

Executive Summary

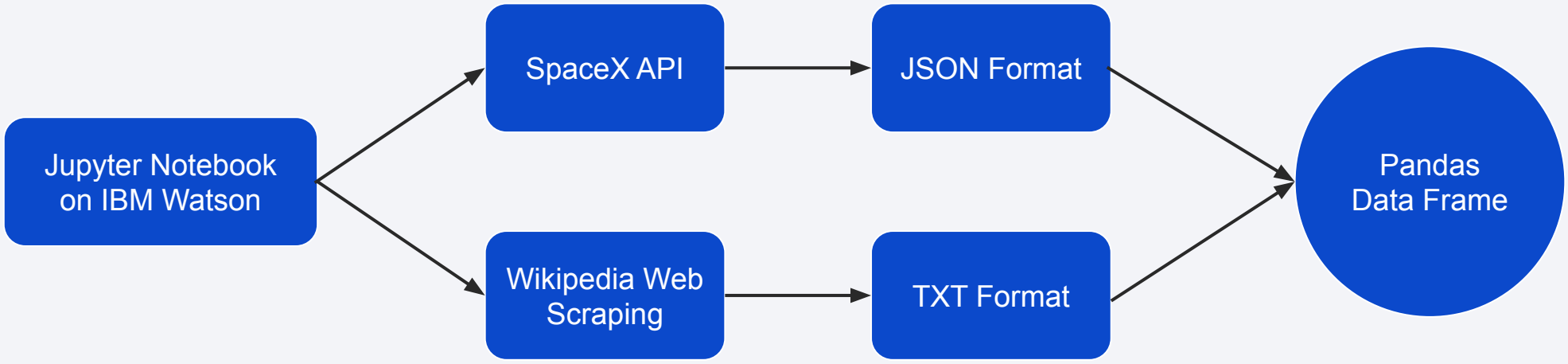
- Data collection methodology:
  - The data was collected using SpaceX rest API in addition of using data web scraping on Wikipedia webpages
- Perform data wrangling
  - The data was preprocessed using Panadas and NumPy, some of main technique are used: OneHot encoding,
  - unnecessary columns removal, data normalization and standardization.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Starting with splitting the data into train and test sets
  - Identifying the best algorithm and parameters through hyperparameters tuning using Grid Search
  - Adopting the best algorithm and parameters for the purpose of model deployment.

Data Collection

We have collected the data from two main sources:

- SpaceX API: Open Source REST API for launch, rocket, core, capsule, starlink, launchpad, and landing pad data.
- Wikipedia: is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation

The Process of data collection:



Data Collection – SpaceX API

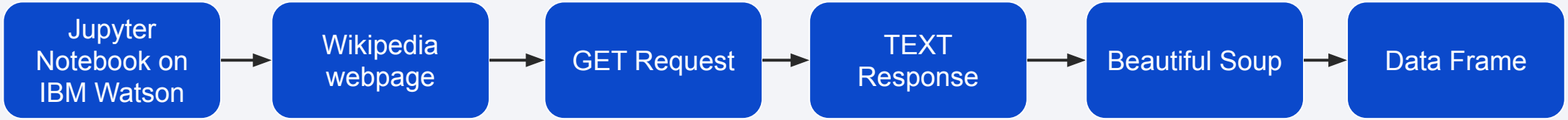
- We started the data collection from SpcaeX API by importing the required libraries such as pandas, NumPy and Request, then we established a URL GET request, this request is raised as JSON file to be finally converted to a data frame through choosing the required information like the geospatial info, rocket type, orbit, flight number and more
- GitHub URL of the completed SpaceX API calls notebook [Link]

Data Collection Process from SpaceX API

Data Collection – Scraping

- As we have done before we start by importing the required Python libraries beautiful soup and request to perform our task, and this time, we have used a webpage on Wikipedia called “Space X Falcon 9 First Stage Landing Prediction” as a data source, then we initialized an HTTP Get Request and the response was as a text format, then we used the beautiful soup library to extract the tables and columns effectively from the text response to be converted later to a pandas' data frame.
- GitHub URL of the completed web scrapping task notebook Link

Data Collection via Wikipedia webpage scrapping



Data Wrangling

- In this stage we started by importing pandas and NumPy, loading our collected data in the previous stage to perform our exploratory data analysis which aimed to clean the data and choose the valid features for training a machine learning model

- GitHub URL of the completed Data wrangling notebook Link

Data Wrangling Stages

1.Loading the collected dataset.

2.Identifying and calculating the percentage of the missing values in each attribute.

3.Identifying which columns are numerical and categorical.

4.Calculating the number of launches on each site.

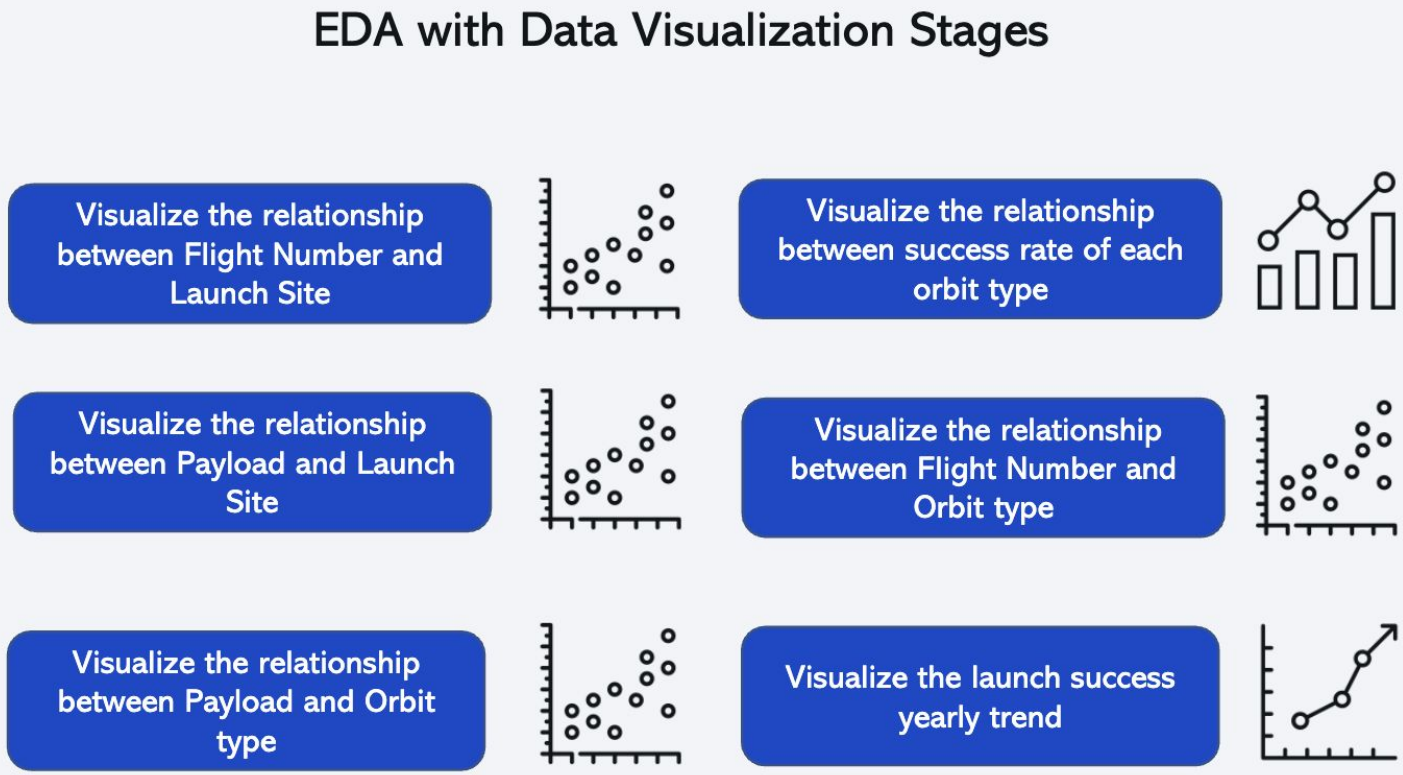
5.Calculating the number and occurrence of each orbit.

6.Creating a landing outcome label from the Outcome column.

7.Determining the success rate of returning the first stage of the rocket.

EDA with Data Visualization

- In this stage we completed our EDA process through finding the correlation between the features and the target using different visualization tools via seaborn and matplotlib furthermore we have performed feature engineering by converting categorical features into dummy values
- GitHub URL of the completed EDA with data visualization notebook, Link



EDA with SQL

- In this stage we used SQL quires as shown on the right to complete our EDA on our collected datase
- GitHub URL of the completed EDA with SQL notebook, Link

SQL Quires

- 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).
- List the date when the first successful landing outcome in ground pad was achieved.
- 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. Use a subquery.
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for the in year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06- 04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

- In this stage we used folium library to represent our work as geospatial data by drawing markers circles and lines on an interactive map.
- GitHub URL of the completed interactive map with Folium map notebook, Link

Building an Interactive Map with Folium

- We started our interactive map by drawing 4 circles on 4 different sites belongs to Falcon 9 rockets lunches have the following information:

Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.57682
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610746

- We put markers to on the same sites to represent the successful/failed first stage of rockets return using marker objects :
- Finally, we calculated the distances between the launch site (CCAFS LC-40) to its proximities 1-the closest city, 2-coastline, and 3-highway. Then we drew polylines to represent these distances using PolyLine object

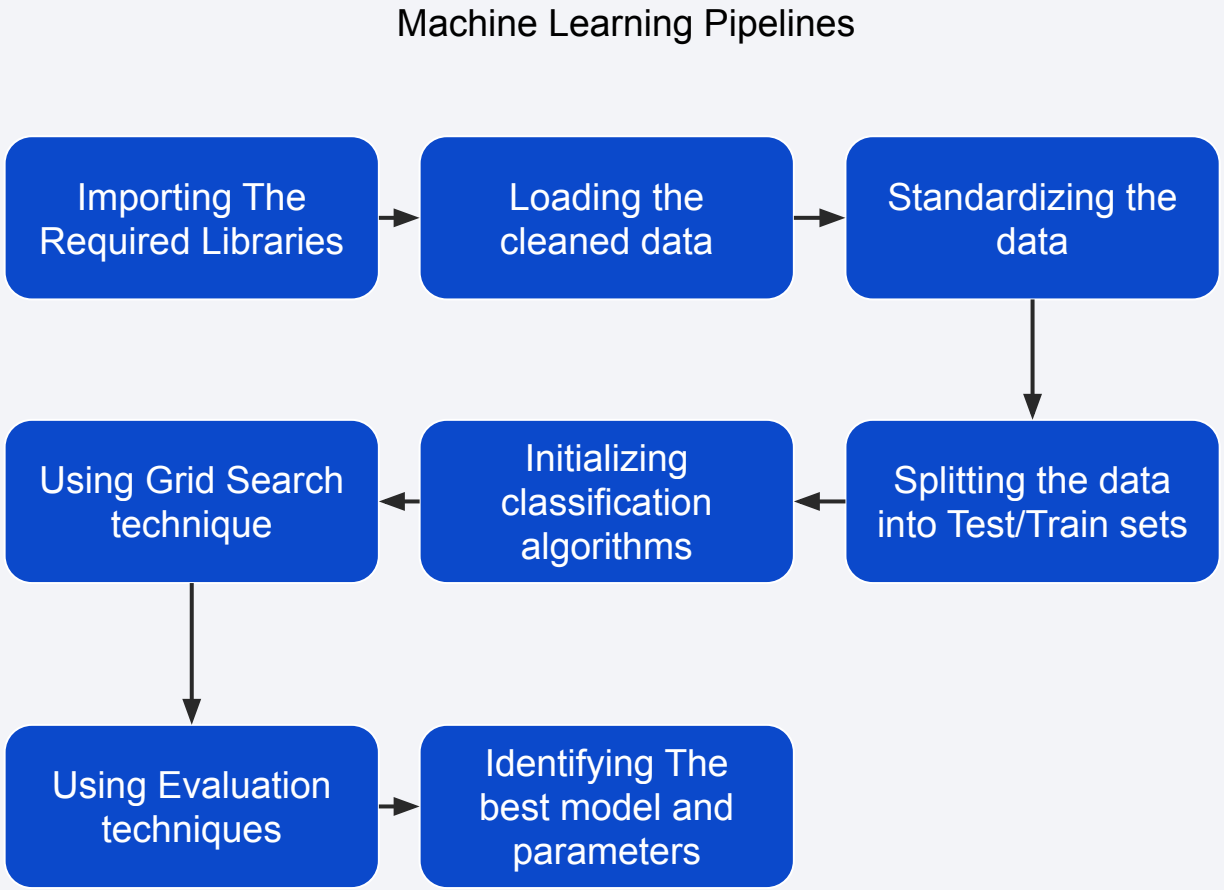
Build a Dashboard with Plotly Dash

- Building an Interactive Dashboard with Plotly Dash
- 1- We added a dropdown list to enable Launch Site selection including the following options:
  - All Sites, CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E,   KSC LC-39A
- 2- we added a pie chart to show the total successful launches count for all sites
- 3- we added a slider to select payload which ranges from 0 -10000
- 4- finally we added a scatter chart to show the correlation between payload and launch succes

GitHub URL of the completed Building an Interactive Dashboard with Plotly Dash notebook, Link

Predictive Analysis (Classification)

- Machine Learning Stages:
- 1- Importing the required libraries.
- 2- Loading the cleaned data.
- 3- Standardizing the data to prevent the bias.
- 4- splitting the data into 20% for testing data and 80% training data.
- 5- Initializing 4 different classification algorithms:
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree (DT)
  - K nearest neighbors (KNN)
- 6- Using Grid Search technique to find the best parameters
- 7- Using Evaluation techniques including, Confusion matrix ,
- F1 score, Jaccard Score for the purpose of using the best model among the algorithms abov
- Link

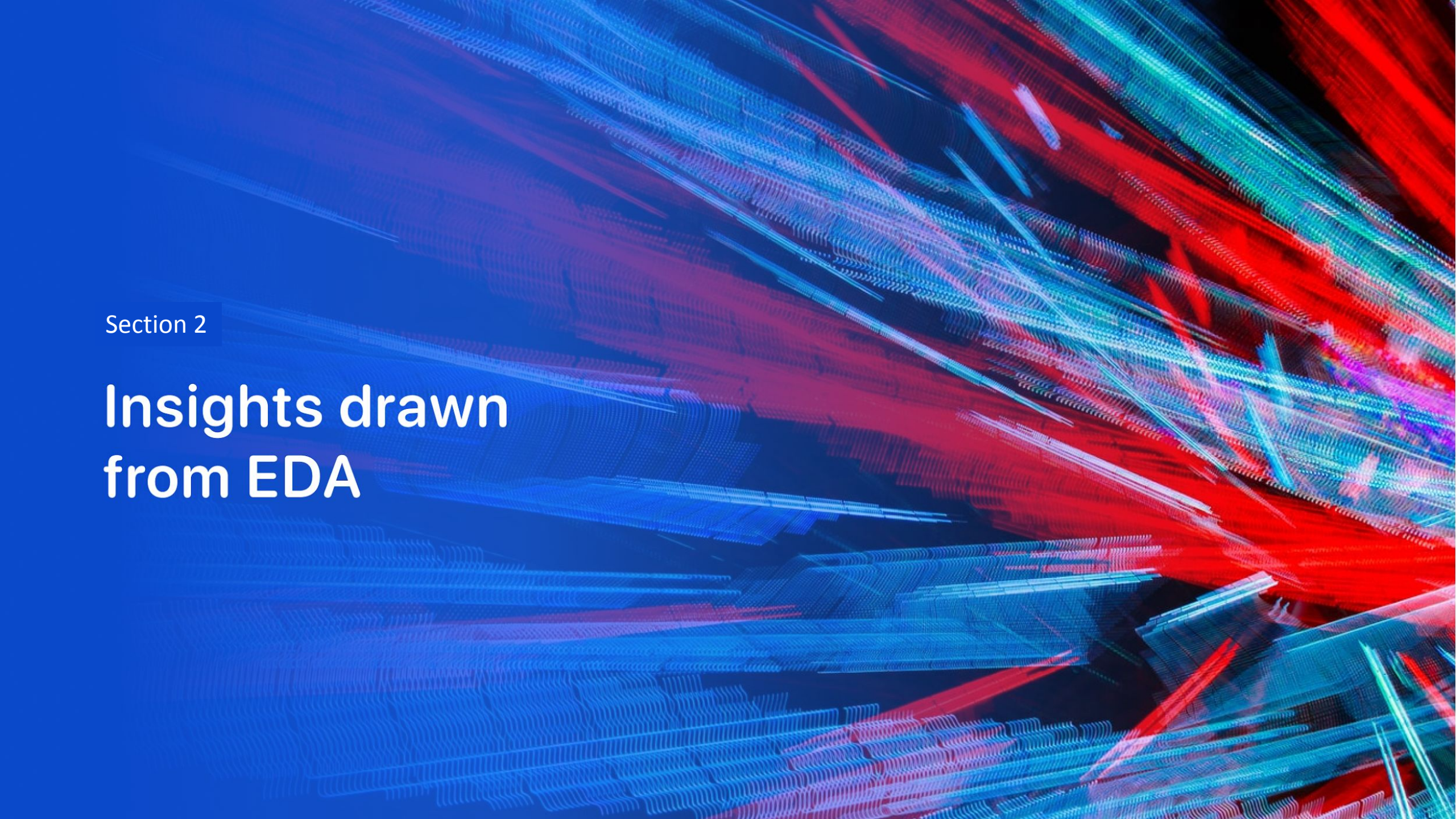


Results

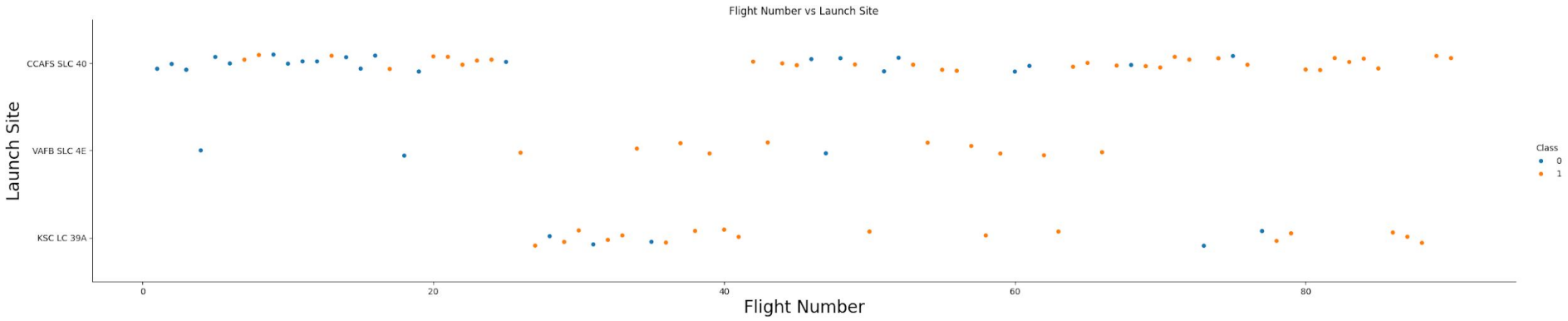
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

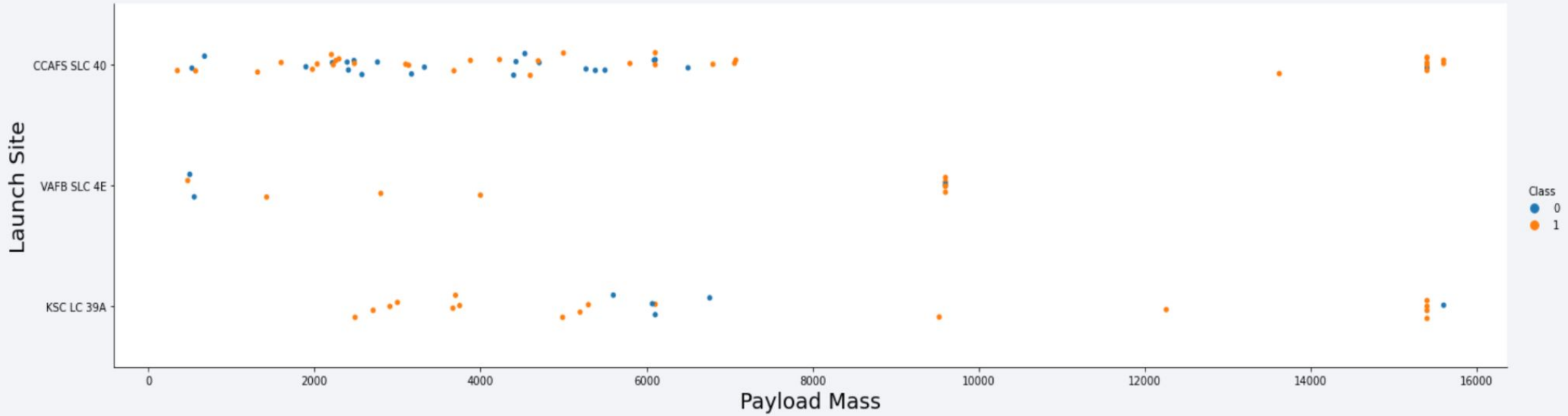


Flight Number vs. Launch Site



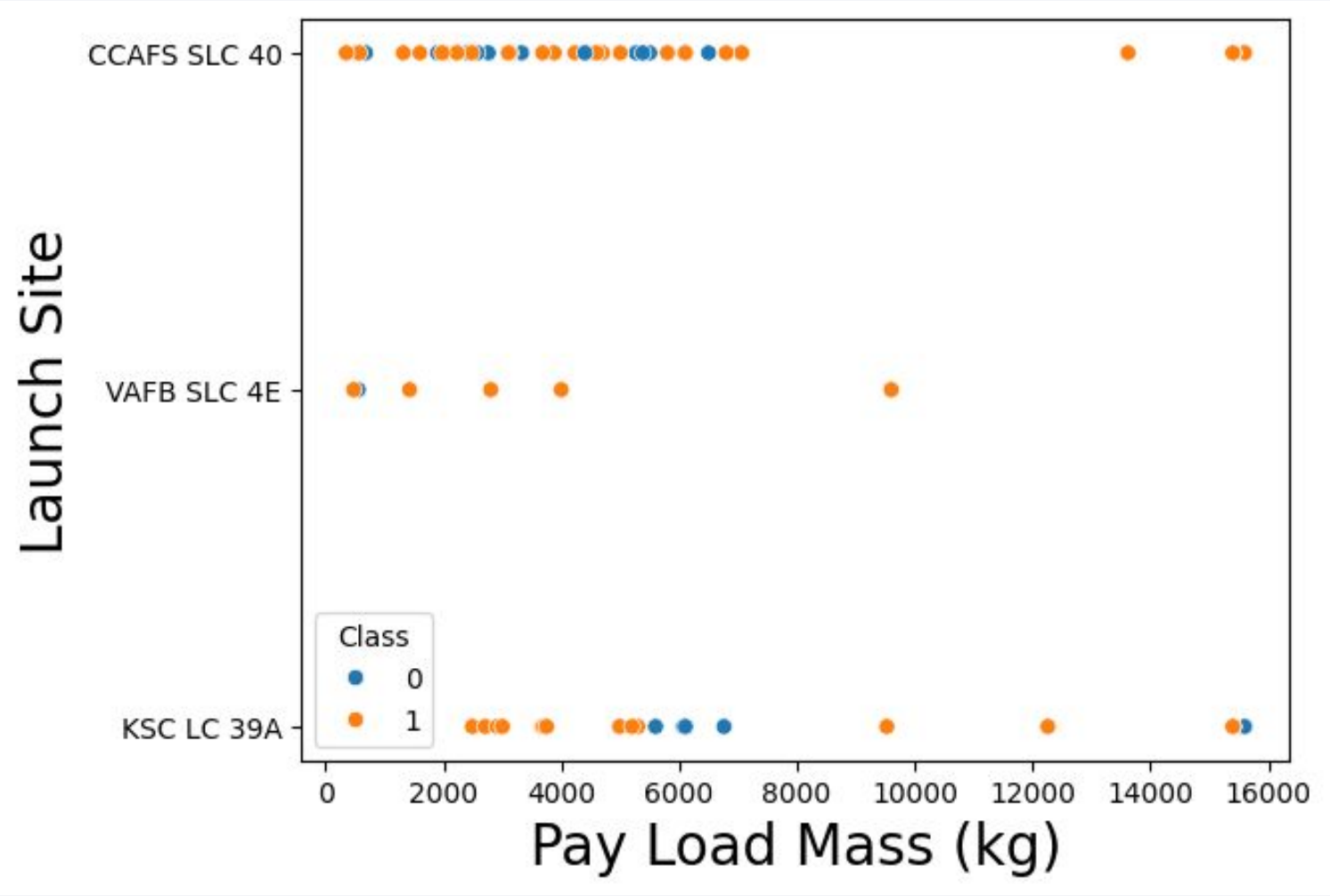
- 1- CCAFS SLC 40 : is the most usable site for launching SpaceX's rockets and it has 55 trials, 33 of them are successful and 22 of them are failed # 60% success rate
- 2- VAFB SLC 4E : is the least usable site for launching SpaceX's rockets and it has 13 trials, 10 of them are successful and 03 of them are failed # 77% success rate
- 3- VAFB SLC 4E : is a moderate site in terms of launching SpaceX's rockets and it has 22 trials, 17 of them are successful and 05 of them are failed # 77% success rate

Payload vs. Launch Site



- According to the plot above:
- there is no strong relationship between the payload mass and the success of first stage return since there are approximately equivalent numbers of failed and successful trials.

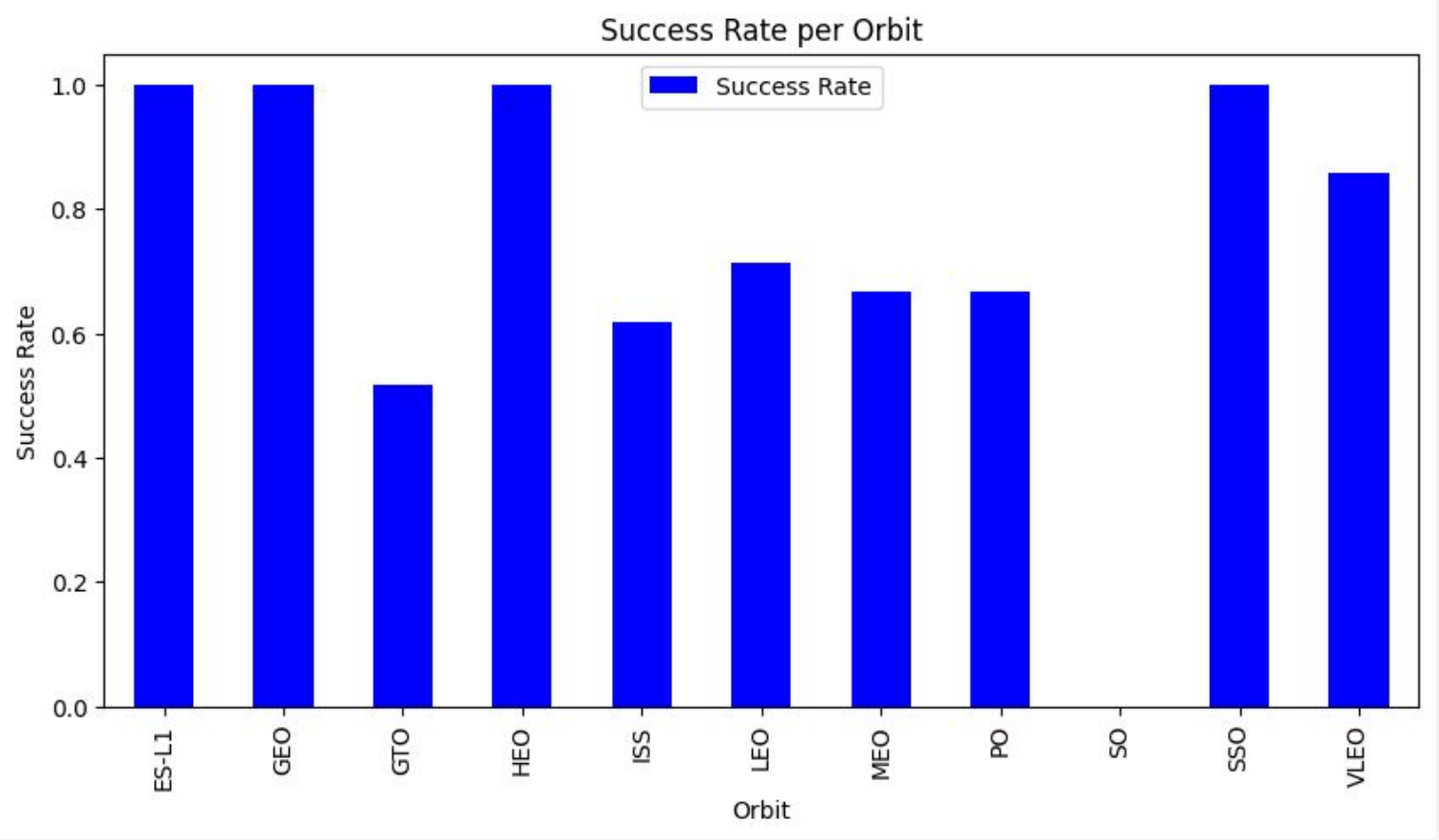
Payload vs. Launch Site



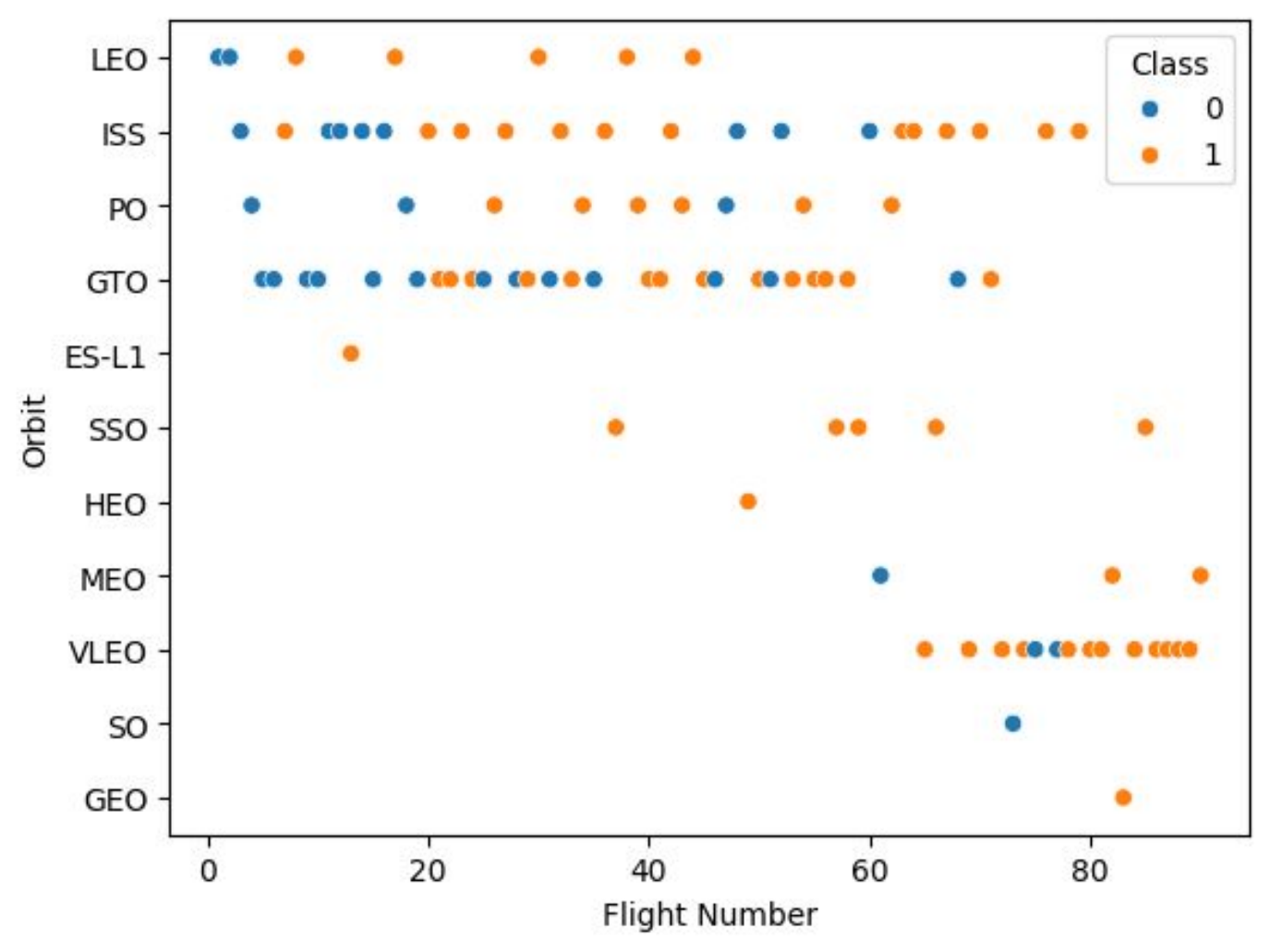
- CCAFS SLC 40 and KSC LC 39A have a wider range of payload masses compared to VAFB SLC 4E.
- VAFB SLC 4E has fewer launches and generally handles lower payload masses
- Successful landings (Class 1, orange dots) occur across all payload masses at CCAFS SLC 40 and KSC LC 39A.
- VAFB SLC 4E shows fewer data points, but successful landings are present, with no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type

- According to the bar plot :
- The best orbits in terms of successful first stage returns are ['ES-L1', 'GEO', HEO, SSO]
- Where the worst orbit is 'GTO' , therefore we need to understand why it is the worst to avoid the failure of first stage return

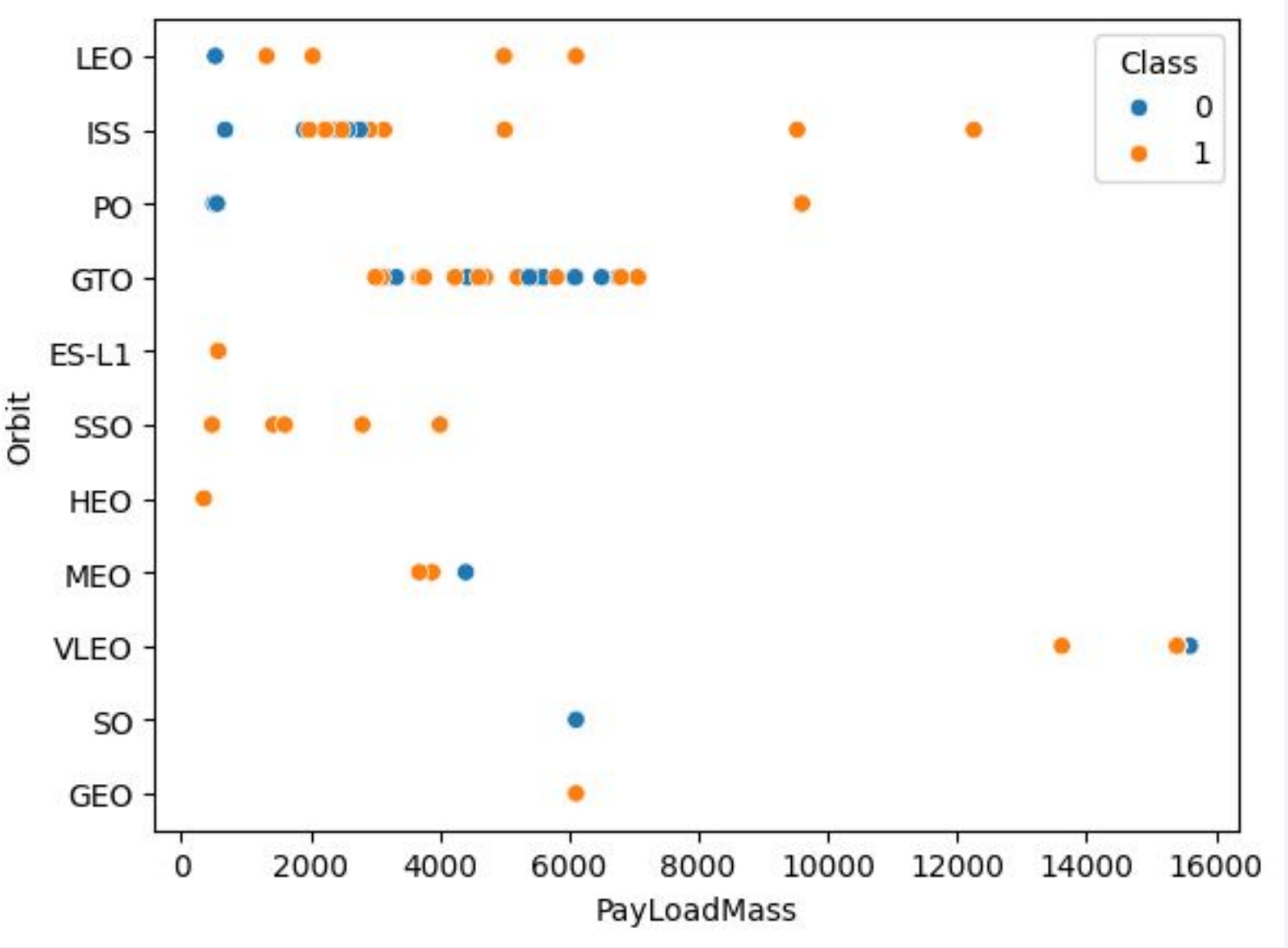


Flight Number vs. Orbit Type



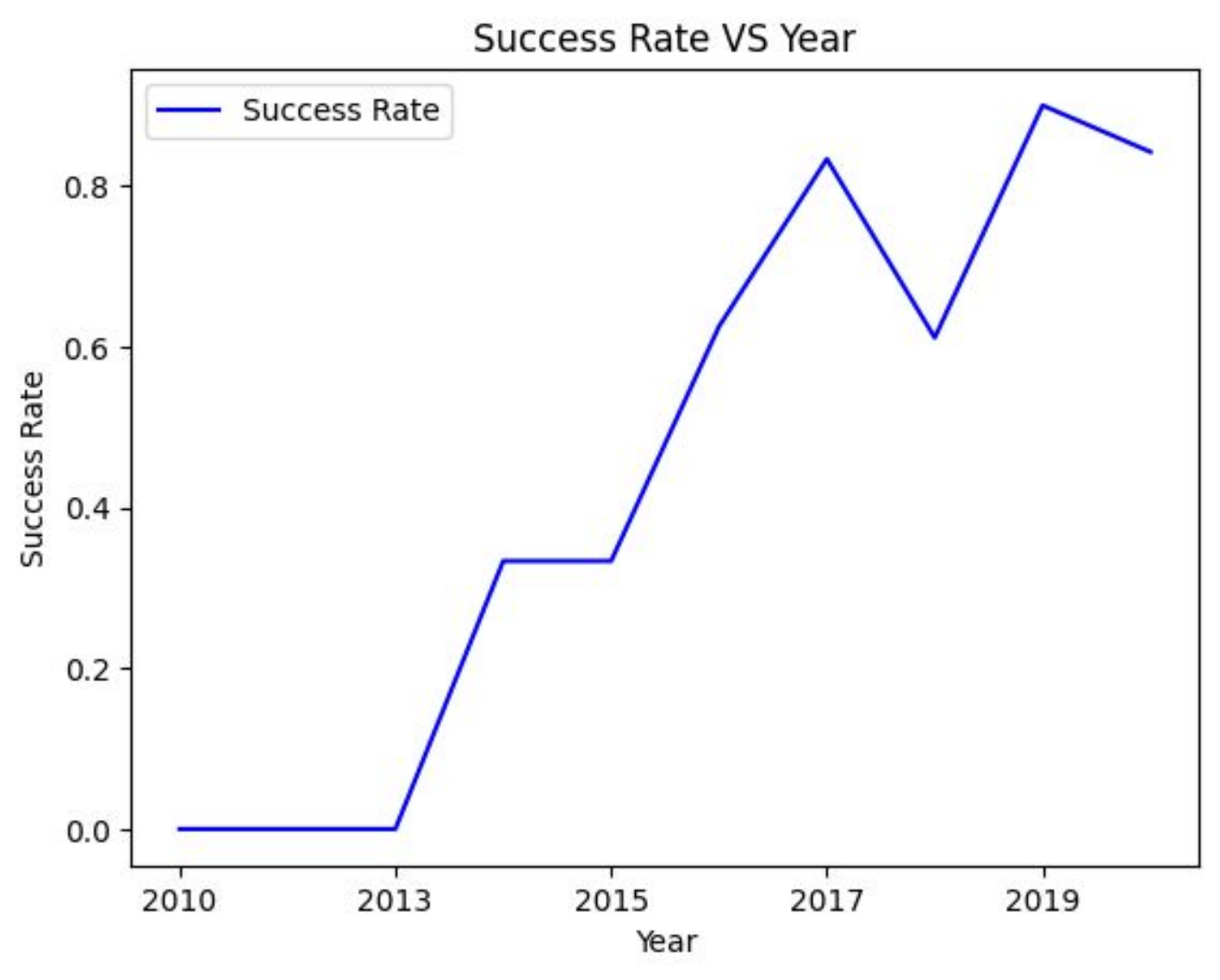
- As the flight number increases, indicating more recent launches, there is a trend towards more successful landing s (Class 1, orange dots).
- Some orbits, like ES-L1, SSO, and GEO, show a higher concentration of successful landings.
- LEO and ISS have a mix of successful and unsuccessful landings, with a slight increase in success over time.
- GTO shows a more balanced mix of outcomes, indicating challenges in achieving consistent success.

Payload vs. Orbit Type



- GTO orbit shows a concentration of payloads around 6000 kg, with mixed success.
- LEO and ISS orbits handle a wide range of payload masses, with many successful landings (Class 1, orange dots).
- Higher payload masses, especially in VLEO and GEO, tend to have successful outcomes.

Launch Success Yearly Trend



- The yearly launch success rate since 2013 kept increasing till 2017, then there was a fluctuation in 2018
- From 2019 to 2020, the success rate rebounded to above 80%
- Overall, Lauch success rates has significantly improved over the years.

All Launch Site Names

- The unique launch sites in the SpaceX dataset are:
- **CCAFS SLC 40**: Cape Canaveral Air Force Station Space Launch Complex 40.
- **VAFB SLC 4E**: Vandenberg Air Force Base Space Launch Complex 4E.
- **KSC LC 39A**: Kennedy Space Center Launch Complex 39A.

CCAFS SLC 40 and CCAFS LC 40 are the same. This needs data cleaning to merge the two.

Launch Site Names Begin with 'CCA'

- The Launch Site beginning with "CCA," is CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40).
- All entries use the Falcon 9 v1.0 booster, indicating early missions in the Falcon 9 program.

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS) is 45596 kilograms

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.67 Kilograms

First Successful Ground Landing Date

- The first date of successful landing outcome on ground pad was December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were

Mission Outcome	Total Outcome
Failure	1
Success	100

- With data cleaning, all successes were merged to obtain 100 successes.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were

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

All boosters with sub versions of F9 B5

2015 Launch Records

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

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 launches occurred in January (01) and April (04).
- Both launches took place at CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40).
- The booster versions used were F9 v1.1 B1012 and F9 v1.1 B1015, both part of the Falcon 9 v1.1 series.

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

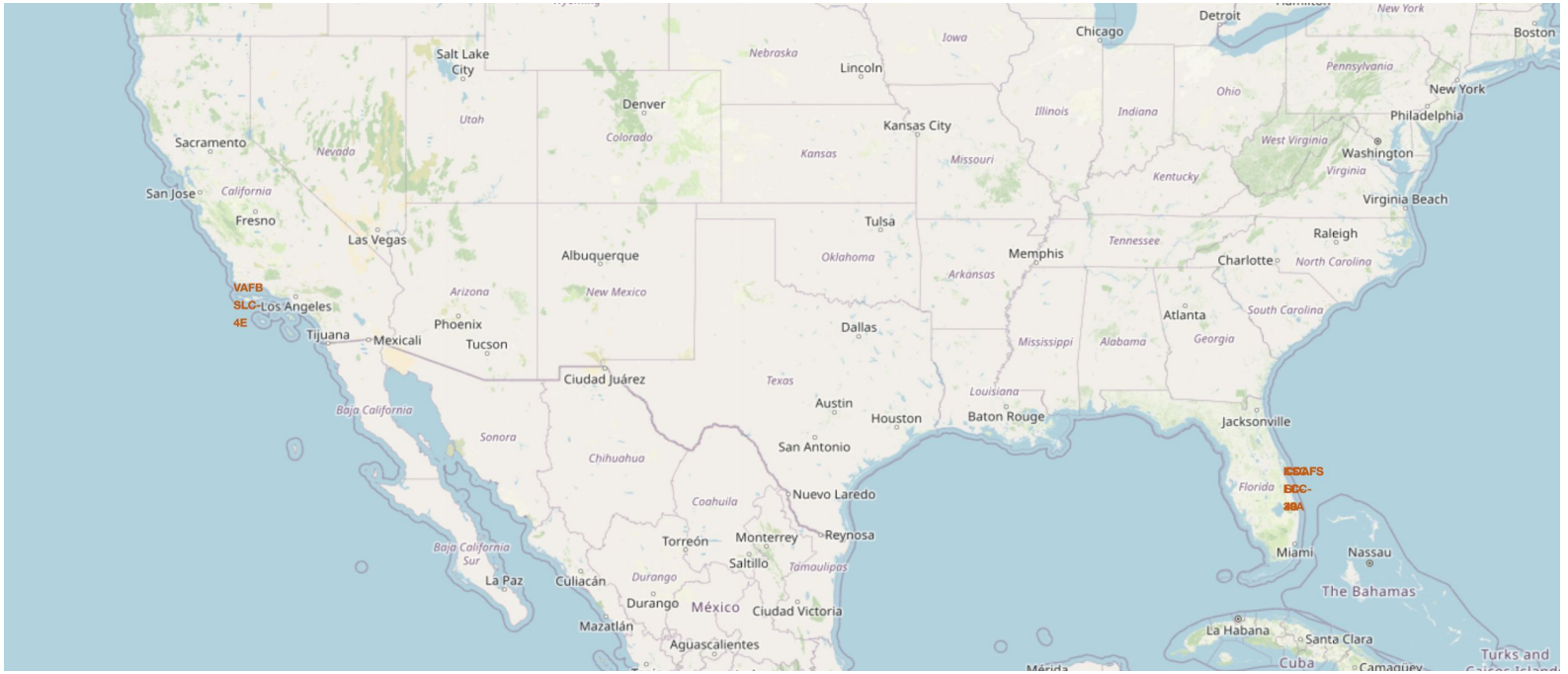
- The Rank of landing outcomes between the date 2010-06-04 and 2017-03-20, are

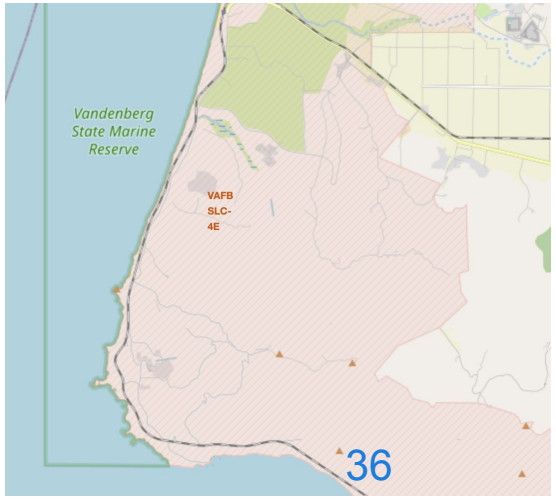
Landing Outcome	Total Outcomes
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2


Section 3
Launch Sites
Proximities Analysis

Launch Sites on the US Map

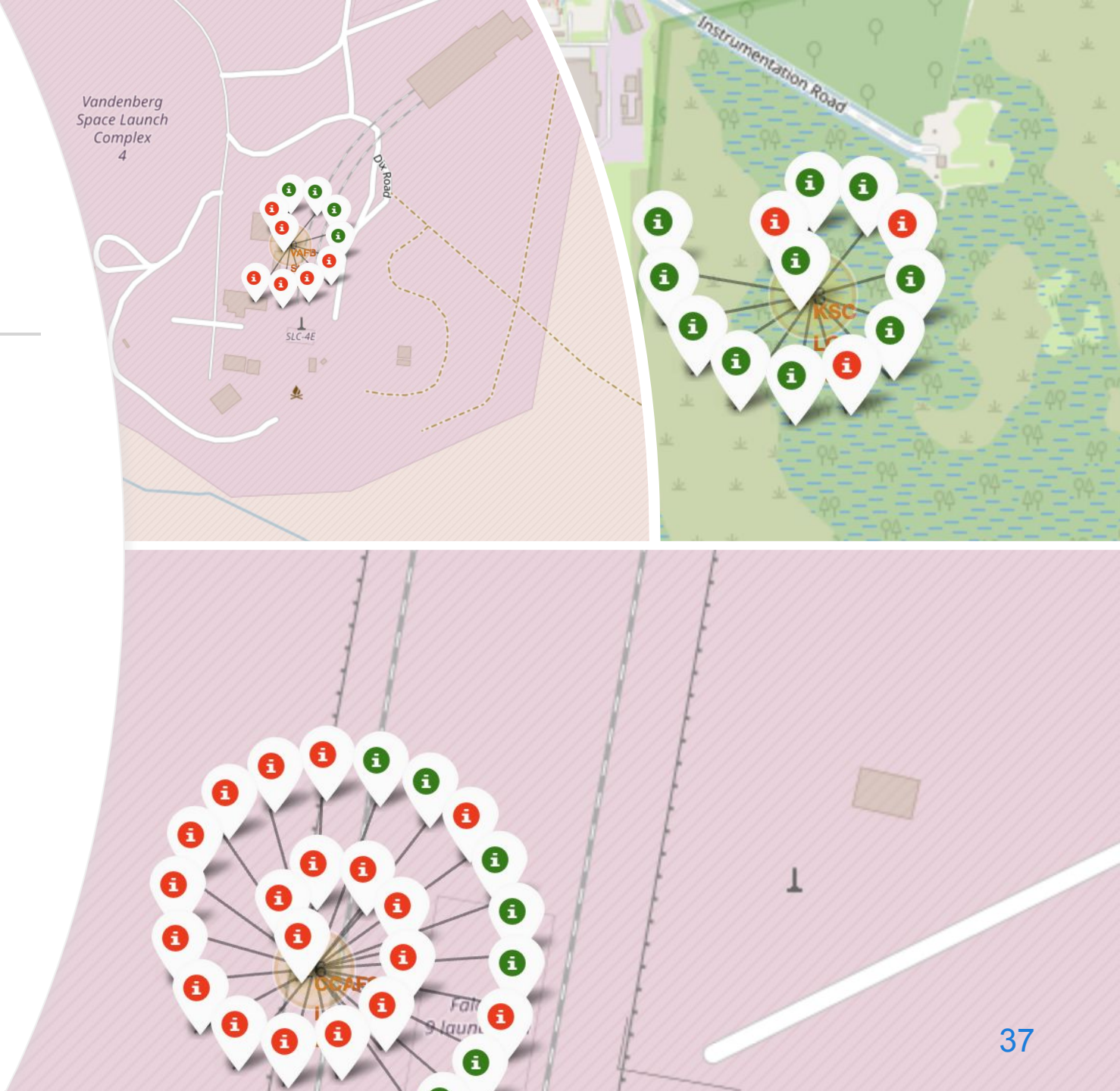
- Labeling of Launch sites on the east and west coasts of the USA





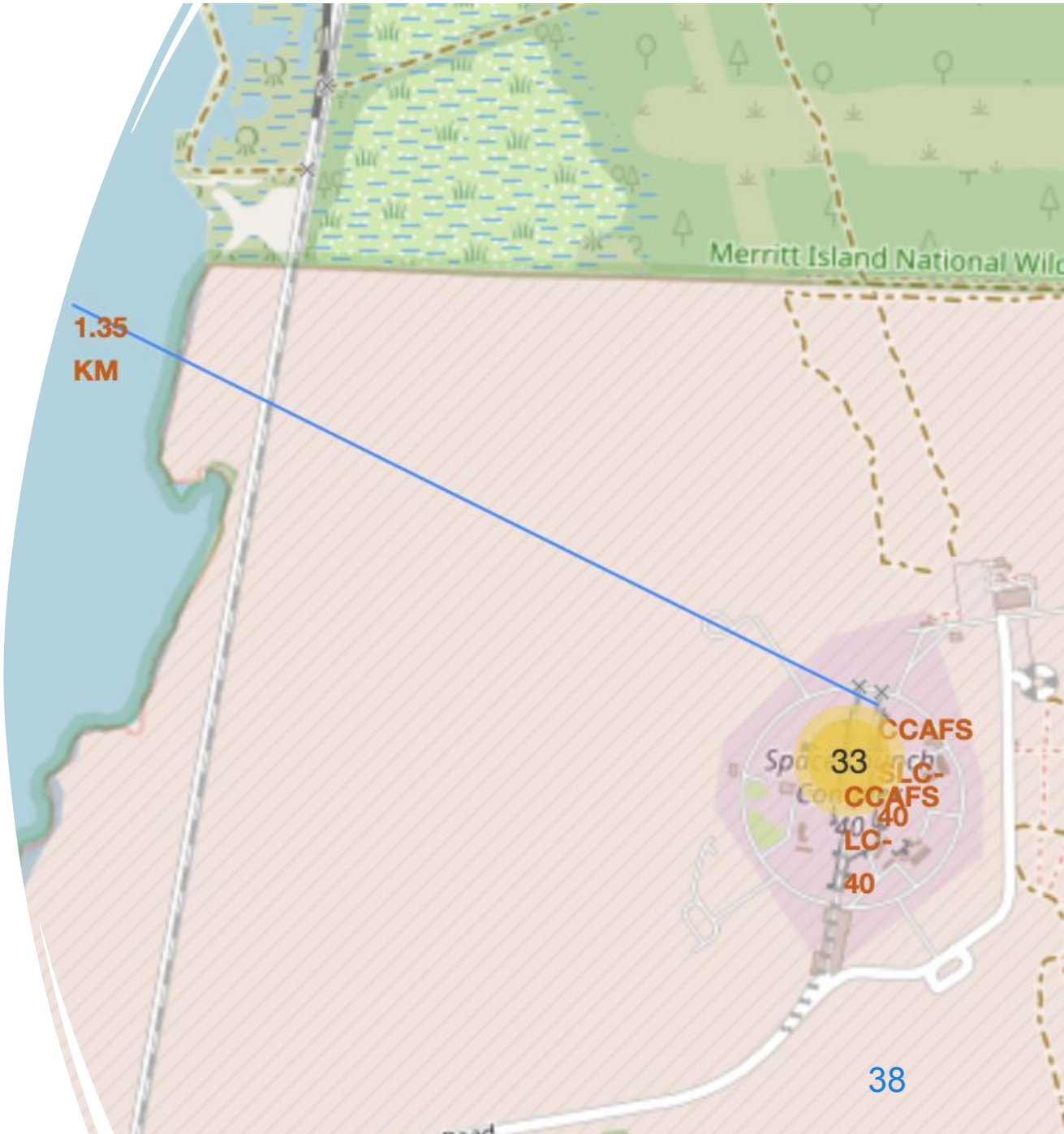
Launch site outcomes

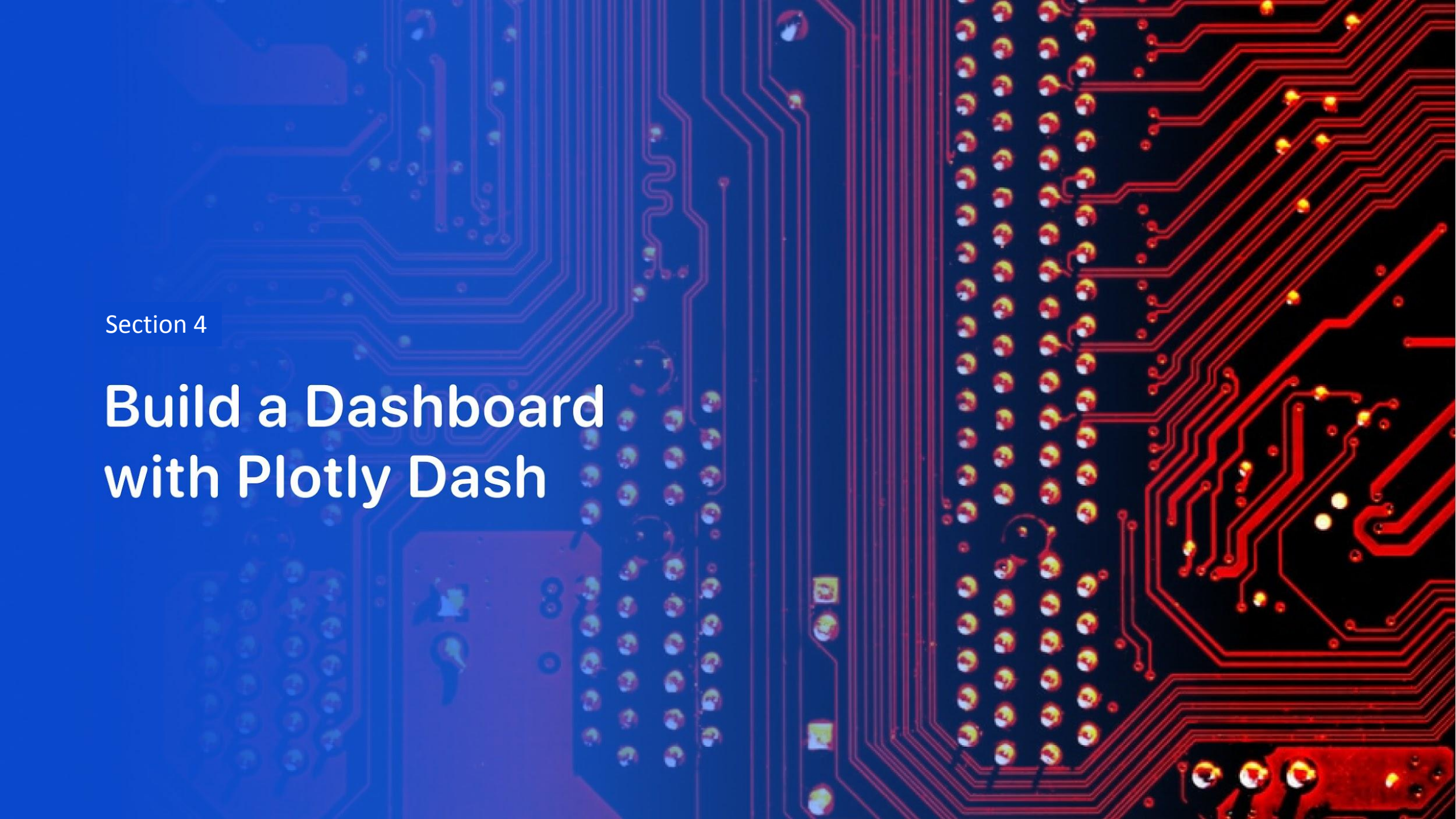
- Successful outcome at the various launch sites are labeled in green with a marker, or red for failure outcomes.
- This help identifies which specific spots at the Launch sites are more likely to have a good successful outcome, all other things being equal.



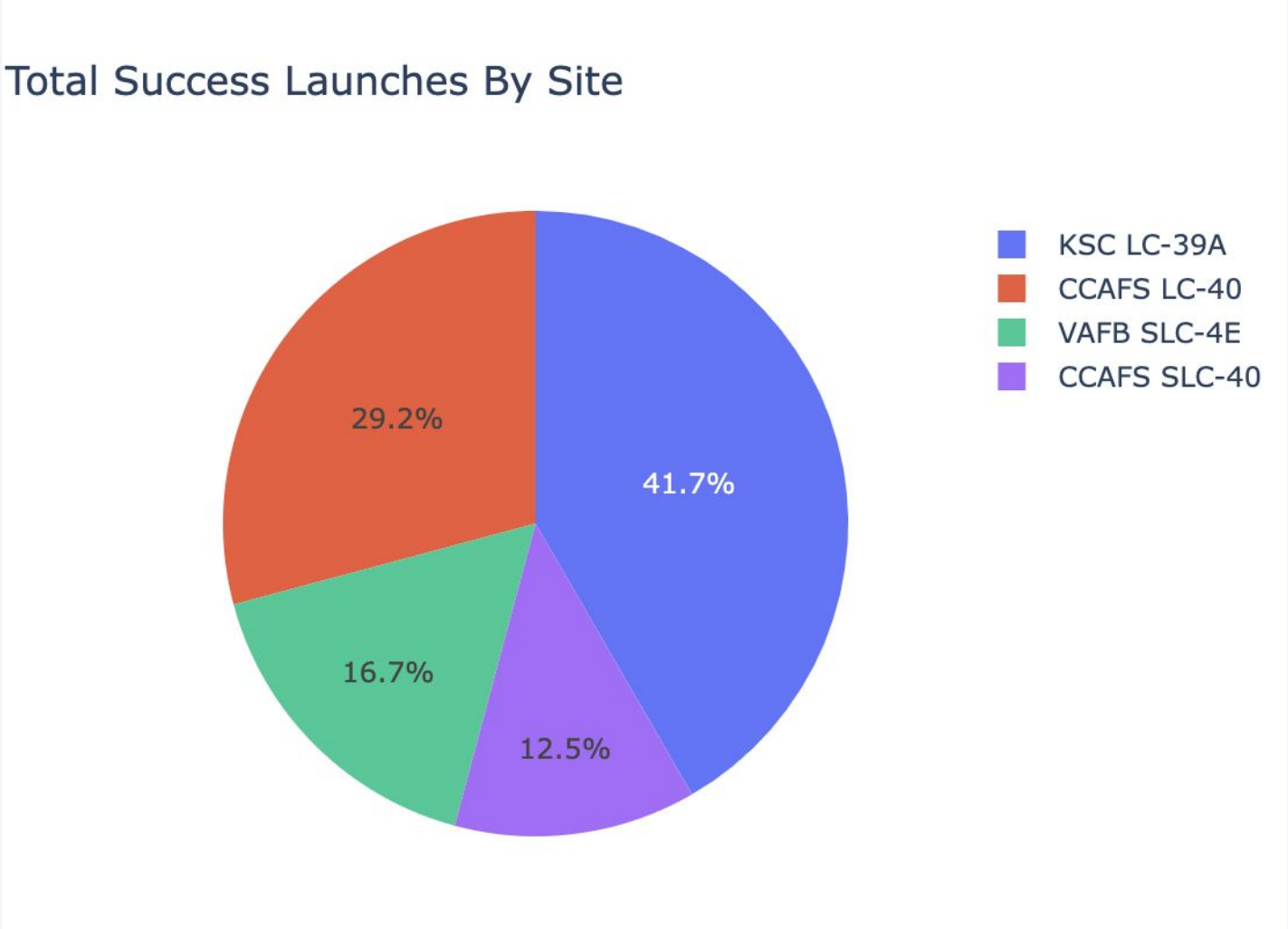
Launch Site proximity to Points of Interest

- A polyline drawn from the launch site to coastal area showing a distance of 1.35 Kilometers



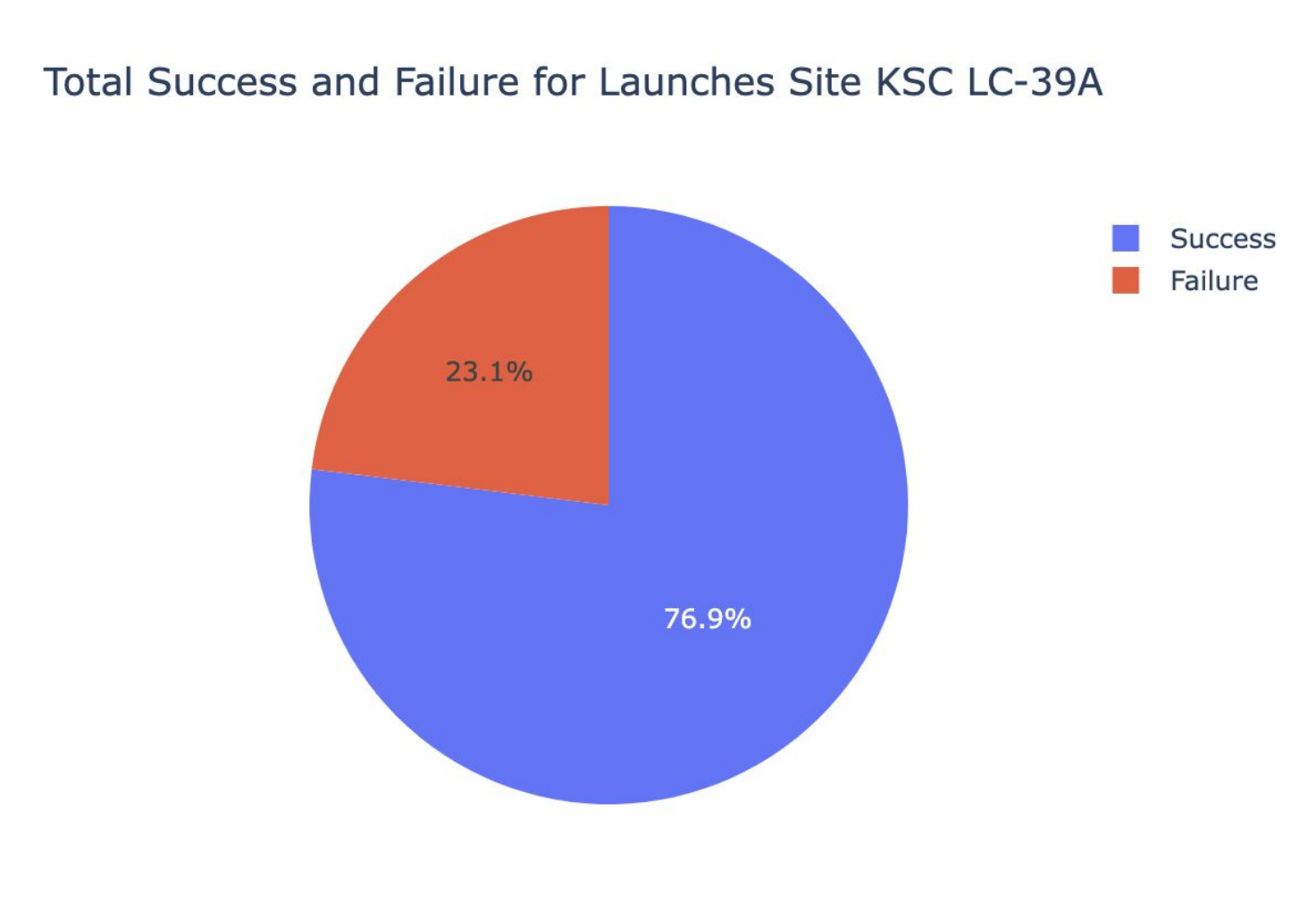

Section 4
Build a Dashboard with Plotly Dash

Launch Success counts per launch site



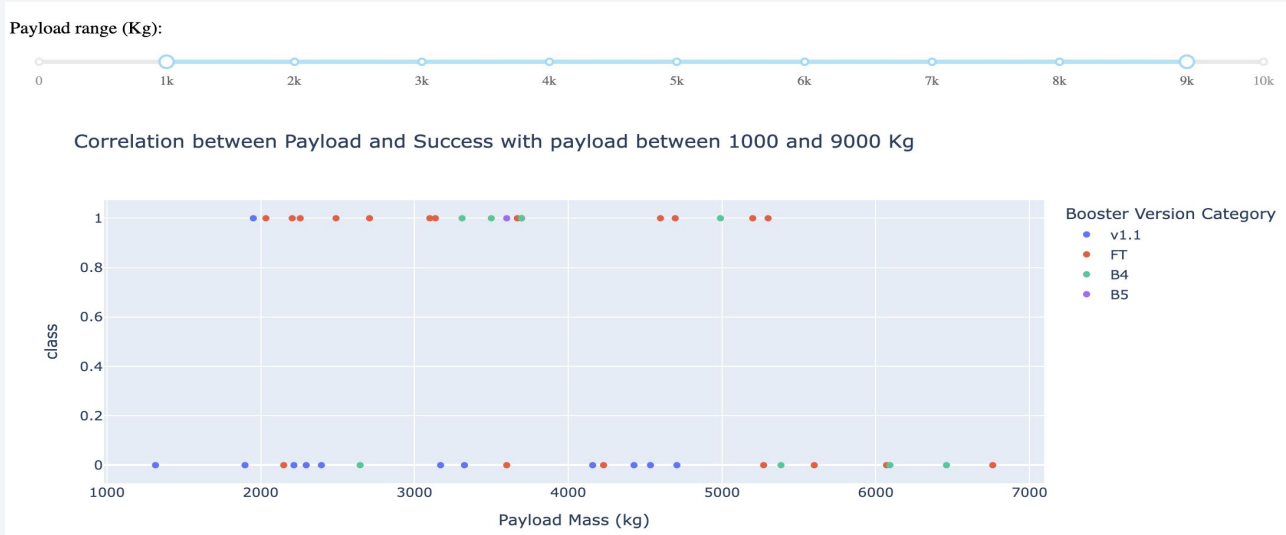
The pie chart shows that KSC LC-39 has the greatest share of successes (41.7%), followed by CCAFS LC-40 with 29.2 %

Outcome Ratio for KSC LC-39 Launch Site



The KSC LC-39 has 76.9% success in all rocket launches

Payload vs Launch Outcomes



Payload range slider allows users to select a range of payload masses (in kg) to filter the data displayed in the scatter plot.

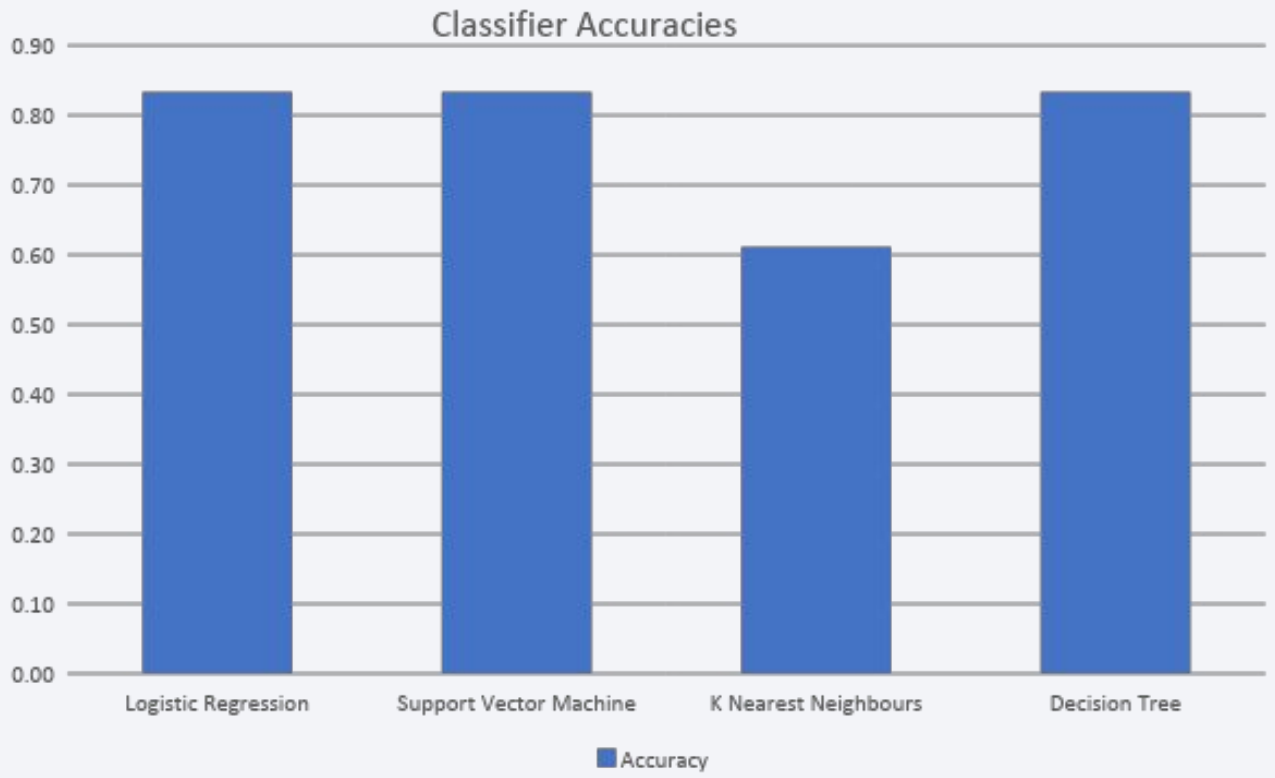
The plots show data for different payload ranges, such as 3000-6000 kg and 1000-9000 kg

The success and failure rate can be observed across all payload ranges

Section 5

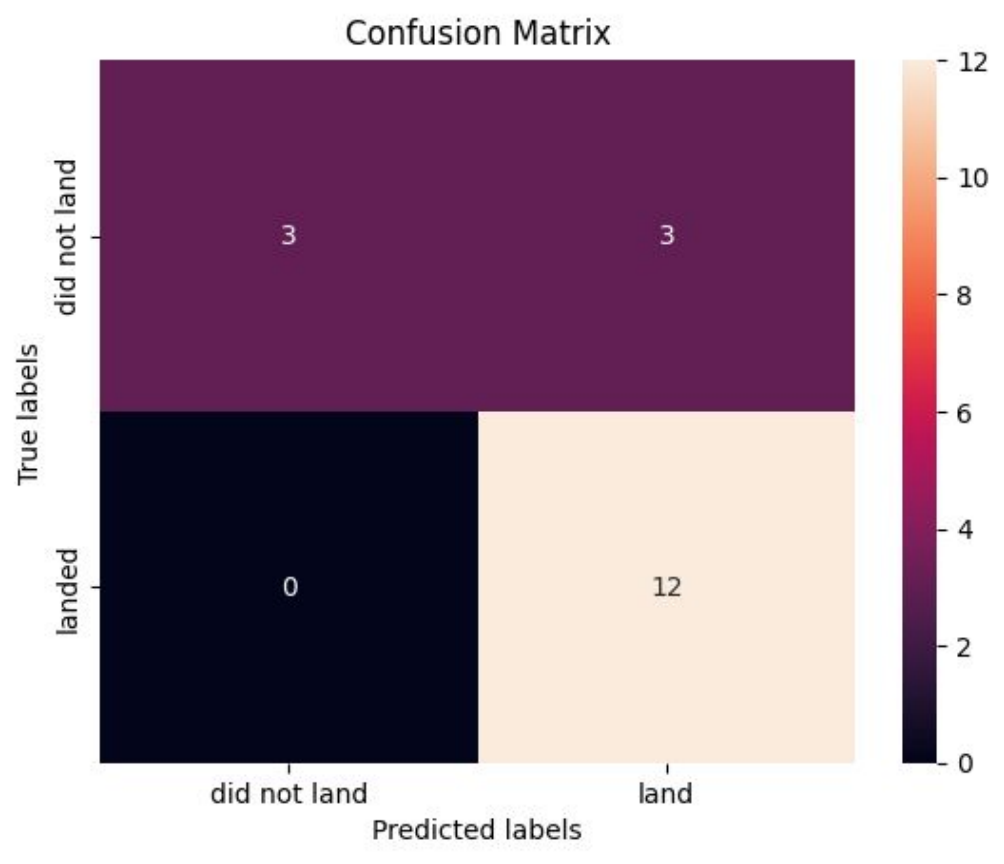
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression, Support Vector Machine, and Decision Tree had the best accuracy of 83.33%
- K-Nearest Neighbors classifier had the least accuracy with 61 %

Confusion Matrix



- True Positives (TP) : The model correctly predicted 12 instances where the first stage landed.
- True Negatives (TN) : The model correctly predicted 3 instances where the first stage did not land.
- False Positives (FP): The model incorrectly predicted 3 instances as landed when they did not land (Type I error).
- False Negatives (FN): The model did not have any instances where it failed to predict a landing when it actually landed (Type II error).
- Overall, the model performs well with an accuracy of approximately 83.3%, but there is room for improvement in reducing false positives.

Conclusions

- SpaceX's Falcon 9 rocket offers launch services at a significantly lower cost ($62 million) compared to traditional providers ($165 million+), primarily due to the reusability of the rocket's first stage, provided it lands successfully.
- The project aimed to identify patterns and trends in landing outcomes through exploratory data analysis (EDA), interactive visualizations, and machine learning techniques. Key findings included:
    - Increased Landing Success Over Time: There has been a notable improvement in successful landings in recent years, especially from launch sites CCAFS SLC 40 and KSC LC 39A.
    - Influence of Target Orbits: Missions to orbits such as ES-L1, GEO, HEO, and SSO exhibited higher landing success rates. In contrast, missions targeting GTO and SO faced more challenges.
    - Predictive Modeling: A machine learning model was developed, achieving an 83% accuracy rate in predicting successful landings. Nonetheless, there is potential to enhance the model's performance, particularly in reducing false positive predictions.
- The find of will help SpaceX in strategizing launches for successful outcomes. Additional research is needed to improve the findings of this project.

Appendix

- The code, notebooks and data for the project can be download on GitHub at https://github.com/randgua/ibm-data-science-professional-certificate

Thank you!