IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mduduzi S Nkomo
5/12/2025

Outline

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

Executive Summary

- **Methodologies** :
  - Data Acquisition and Web Scraping **(**Python libraries (requests, BeautifulSoup)**)**
  - Data Wrangling and Preparation (Python libraries (pandas, numpy), SQL (SQLite))
  - Exploratory Data Analysis (EDA) & Visualization (**Tools:** Python (pandas, plotly.express, folium), SQL)
  - Feature Engineering and Preparation for ML ()

- **Overall Results** :
  - The **Support Vector Classifier (SVC)** emerged as the most effective model for this dataset, achieving the highest overall accuracy of approximately **85%** and strong performance across all metrics.
  - The **Logistic Regression** and **Decision Tree** models performed comparably well, both reaching an accuracy of approximately **81%** and balanced F1 scores.
  - The **K-Nearest Neighbors (KNN)** model demonstrated significantly lower performance, particularly in precision and F1-score, indicating it was the least suitable model for predicting SpaceX launch outcomes among those tested.
- The success rate of each launch site differed from one to another with KSC LC-39A (Kennedy Space Center Launch Complex 39A) for falcon 9

-)

Introduction

- **Project background and context:**
- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launc
- **Problems you want to find answers:**
- Predict if the first stage of falcon 9 will land successfully
- Success rate of Falcon 9 per launch site
- Predict if the landing will be successful
- Features that have an effect on a successful landing

Section 1

Methodology

Methodology

Executive Summary	
Data collection methodology:	Collected the data with an API and webscrapping
Perform data wrangling	Performed EDA to identify necessary data and filtered the data
Perform exploratory data analysis (EDA) using visualization and SQL	
Perform interactive visual analytics using Folium and Plotly Dash	
Perform predictive analysis using classification models	Used DecisionTree , KNeighborsClassifier , SVC & Logistic regression to select the best model

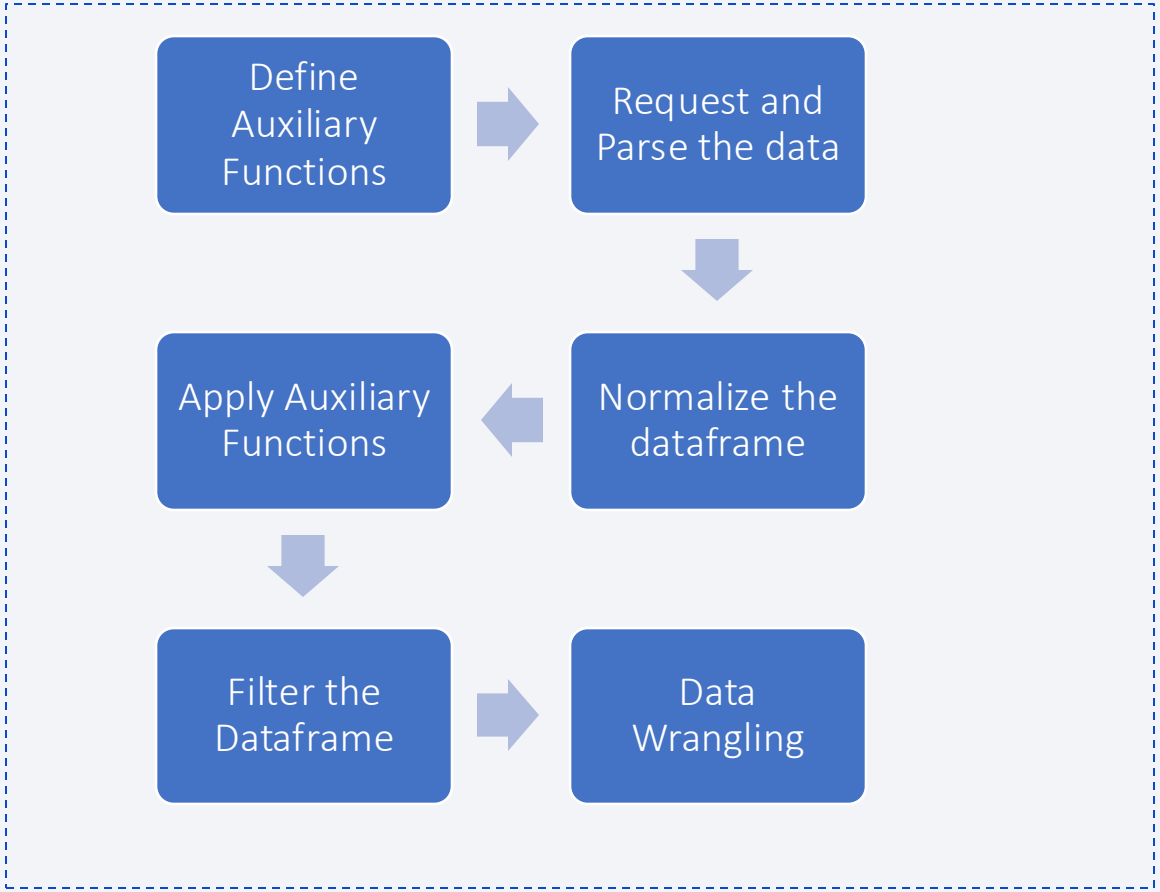
Data Collection

- I collected the Data for Space X using 2 methods :
  - Web-scrapping : Collected the data from the Internet
  - Application Programming Interface
- After extracting the data from web-scrapping l transformed it into a form by removing all the tags from the data and using python code to appropriately format it into a readable format.
- When I used an API l started by defining Auxiliary Functions which extracted the data and store it in a variable l normalized and stored it in a data frame for further processing

Data Collection – SpaceX API

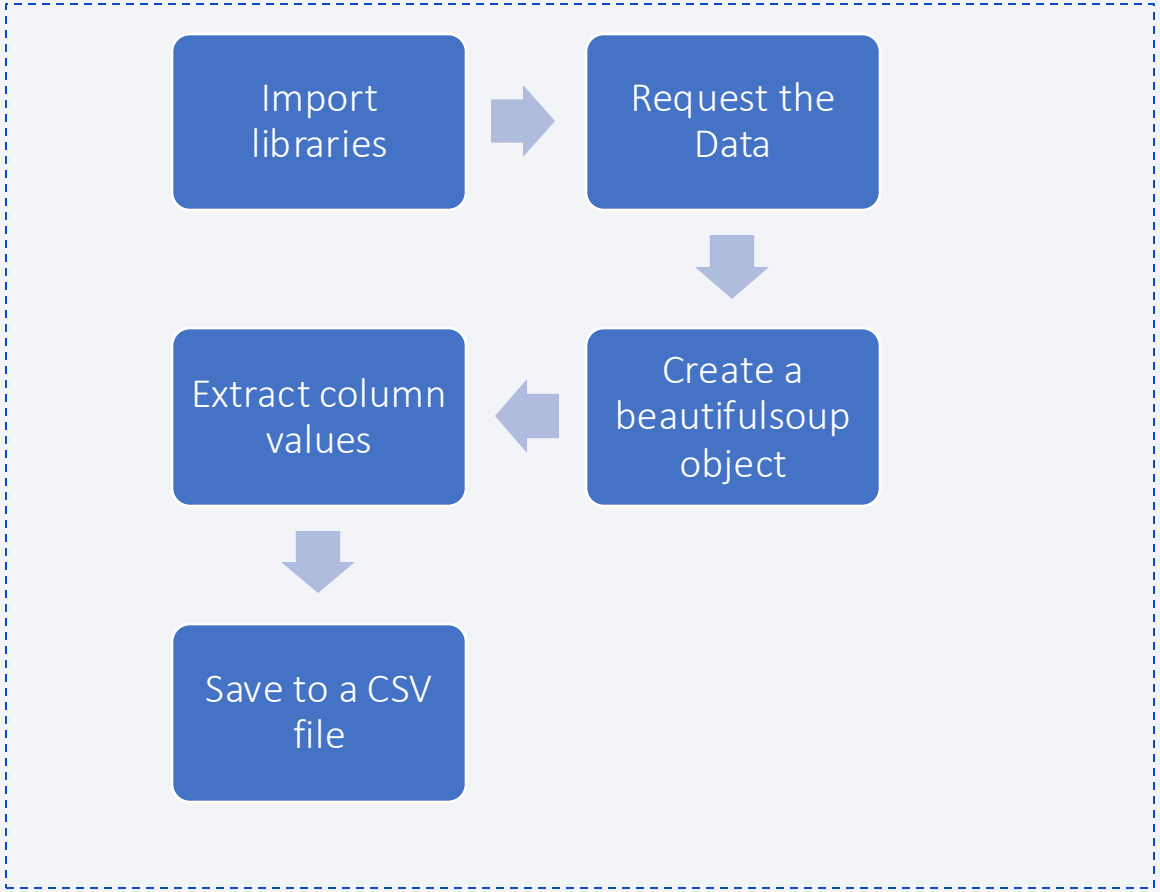
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/Emitrisia/Space_Y/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- https://github.com/Emitrisia/Emitrisia/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Data Wrangling is the process of cleaning the data .

In Data Wrangling we started with checking for missing values in the columns

Then I replaced the missing values with the mean of that column

After that I saved the file to a csv for further Analysis

https://github.com/Emitrisia/Emitrisia/blob/main/labs-jupyter-spacex-Data%20wrangling%20(2).ipynb

Checking for null values

Replacing null Values

Saving

EDA with Data Visualization

- We use Exploratory Data Analysis(EDA) with data Visualization to identify how different features affect or have an effect on our target variable . To identify correlation and relationships between variables .
- In this project l used various data visualiztions for EDA which include :
  - Scatterplots
  - Catplots
  - Lineplots
  - Bar Chart

EDA with SQL

- Selecting Unique Launch sites data from the Table
- Selecting Launch Site with string CCA
- Displaying Total Payload Mass launched by Nasa
- Average Payload Mass by F 9 booster version
- Date of the first successful landing on ground pad
- Names of the boosters which have success in drone ship and have payload mass > 4000& < 6000
- Total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass
- the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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

- Added Cirlce markers , Mouse pointer , Launch site Marker popups and LInes to indicate distance to and fr
- Explain why you added those objects
- Circle Markers were used to mark the radius of the Launch sites
- Added Mouse pointer to know the exact coordinates of an infrastructure to a launch site
- Launch site Marker popups were used to help us visualize successful and failed landing attempts
- Lines were used to measure and help us know the distance of the nearest infrastructure to the Launch Site
- https://github.com/Emitrisia/Emitrisia/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

I used a pie chart and a scatter pot

I used those data visualization to output launch site success rate and how different payload mass affected the success rate

Kennedy Space Center () showed high success rate of successful landings from other launch sites

Predictive Analysis (Classification)

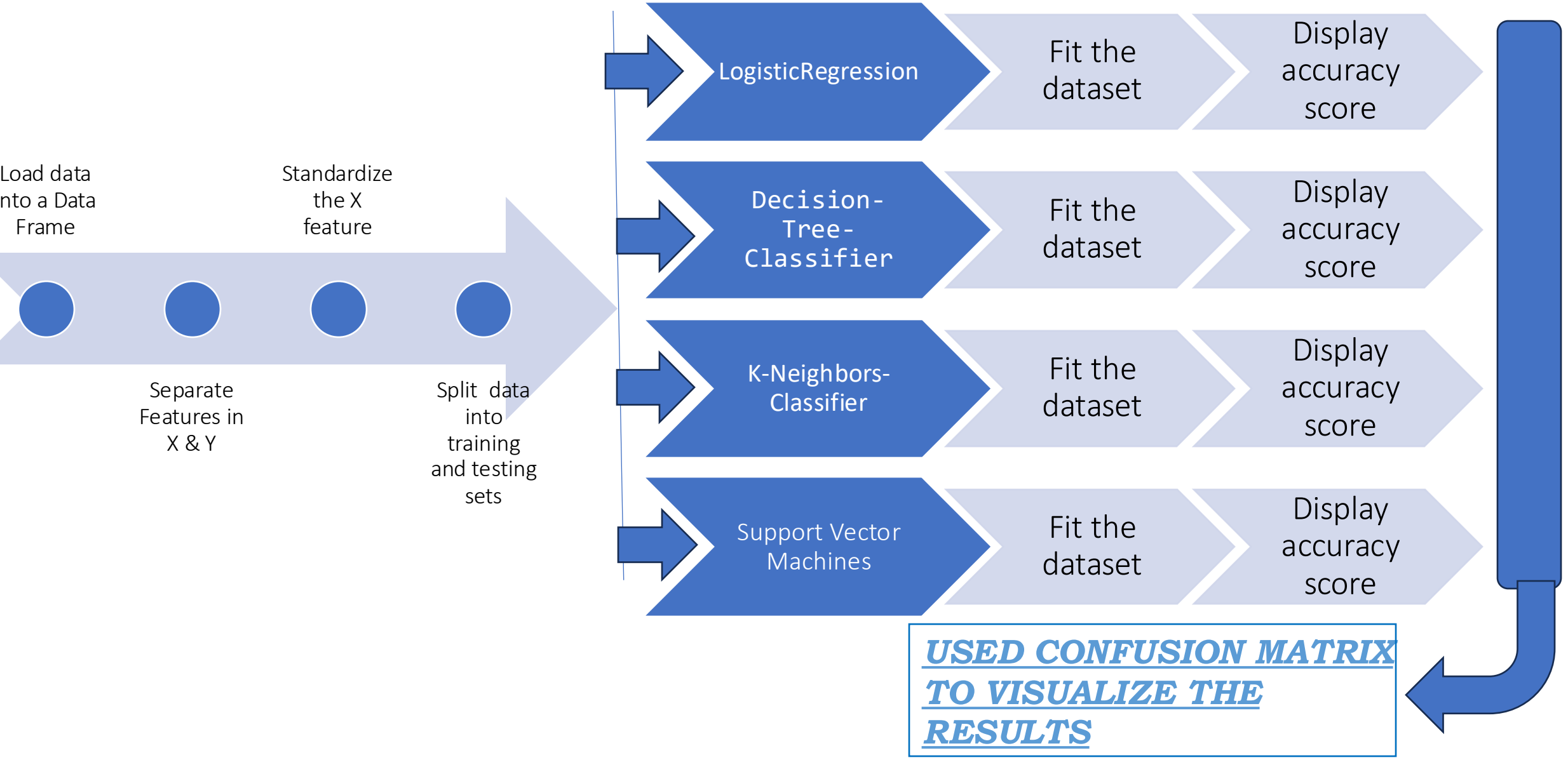
I started by loading the data into a data frame then I separated the data into X (Independent variable) & Y (Dependent variable).

Then I used standard scaler for the feature X to standardize the dataset , after that I split the data into training and testing sets 80% for training .

I then used logistic regression to train and test the model and to see the accuracy of the model , I also used Decision Tree Classifier , Support Vector Machines and K-Neighbors-Classifier

After fitting the data to each method I used a confusion matrix to visualize the results

Predictive Analysis (Classification)



Results

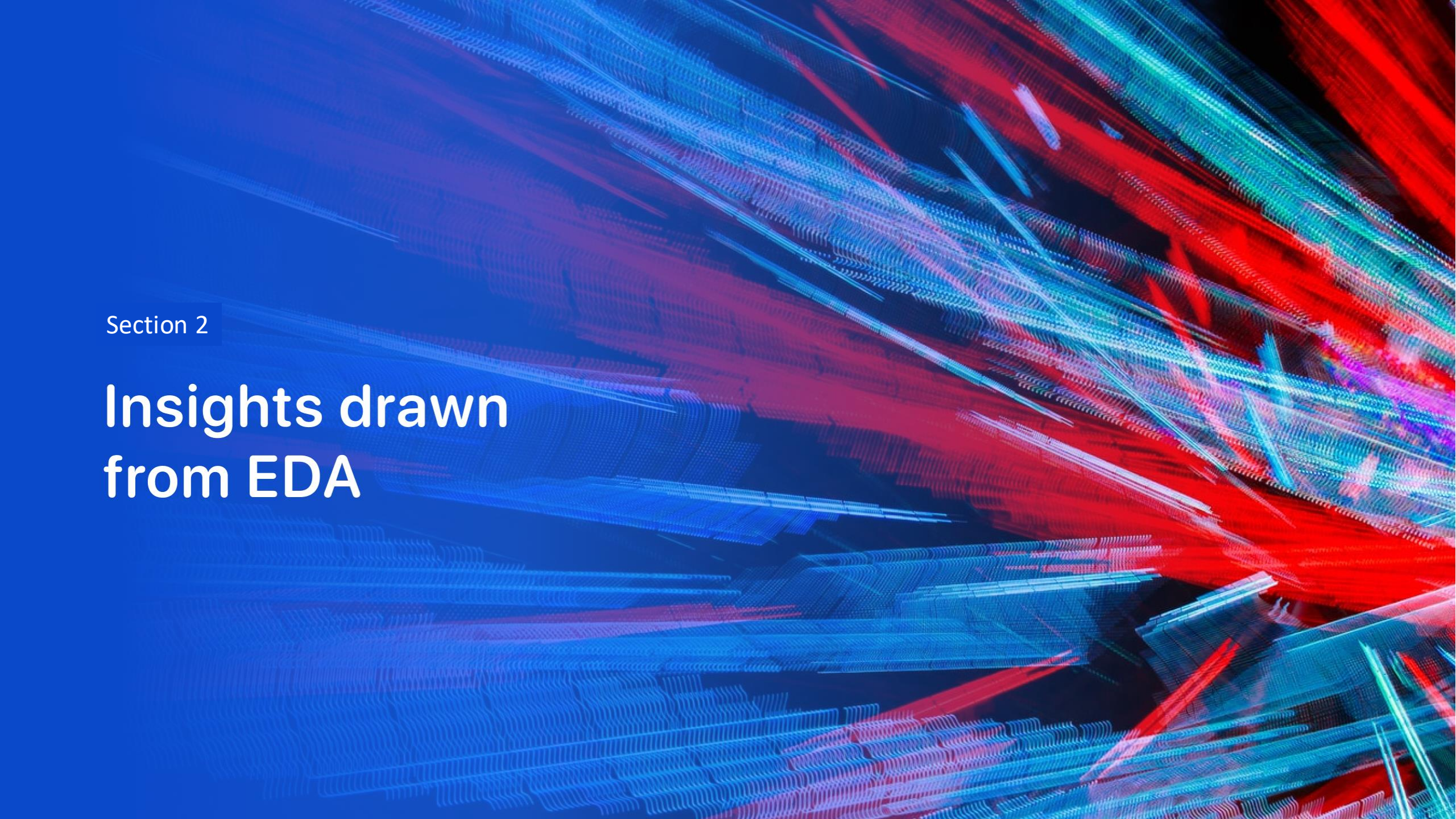
- I used EDA to see the features and effect on the dependent value and l noticed that success rate has been steadily increasing from 2013 until 2020
- I also noticed that Payload and Orbit type have a great effect on the success rate
- VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)
- With heavy payloads the successful landing or positive landing rate are more for Polar ,LEO and ISS
- I also tried to visualize the relationship between orbit type and payload mass

Predictive Analysis Results

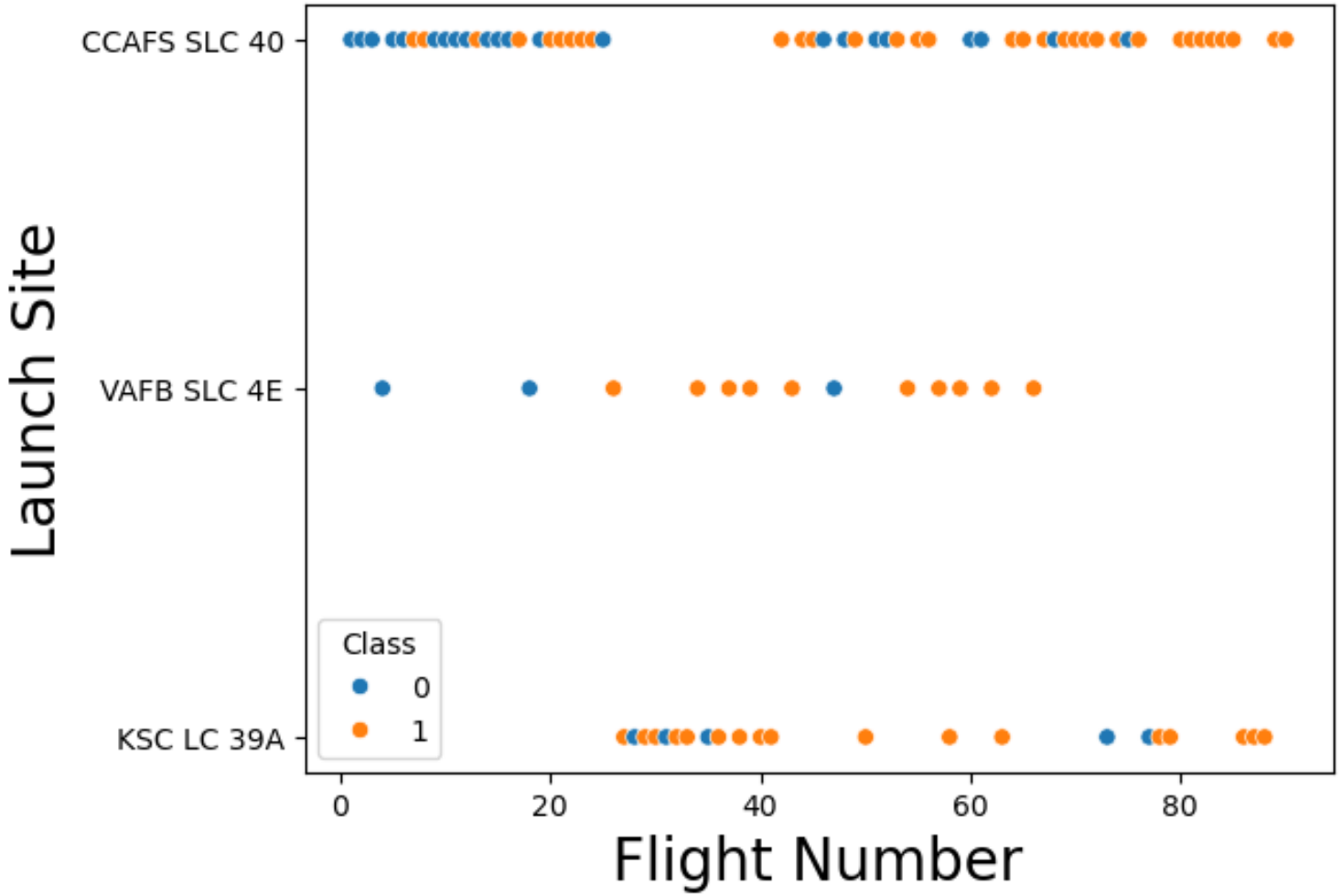
The **Support Vector Classifier (SVC)** emerged as the most effective model for this dataset, achieving the highest overall accuracy of approximately **85%** and strong performance across all metrics.

The **Logistic Regression** and **Decision Tree** models performed comparably well, both reaching an accuracy of approximately **81%** and balanced F1 scores.

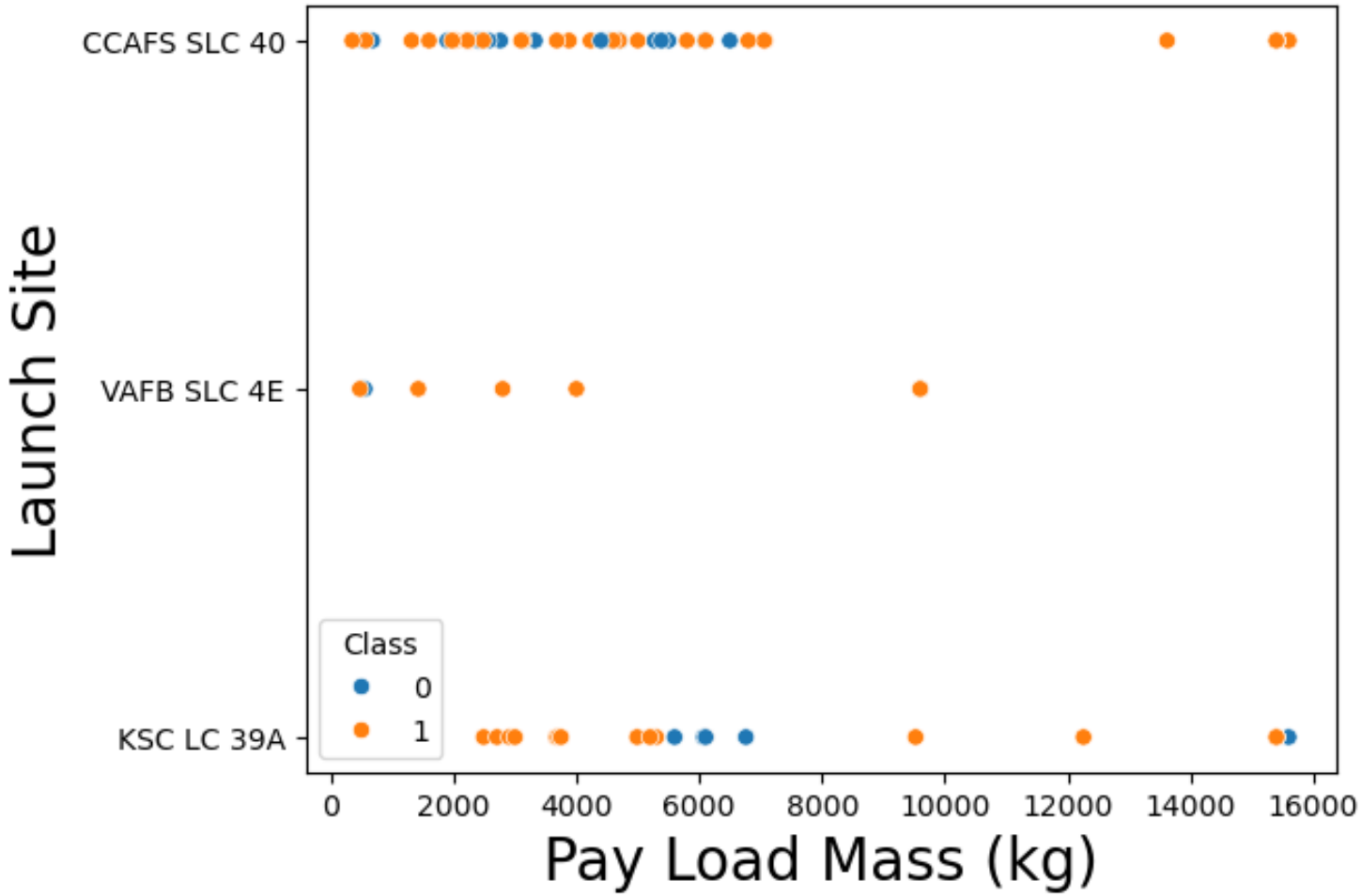
The **K-Nearest Neighbors (KNN)** model demonstrated significantly lower performance, particularly in precision and F1-score, indicating it was the least suitable model for predicting SpaceX launch outcomes among those tested.


Section 2
Insights drawn from EDA

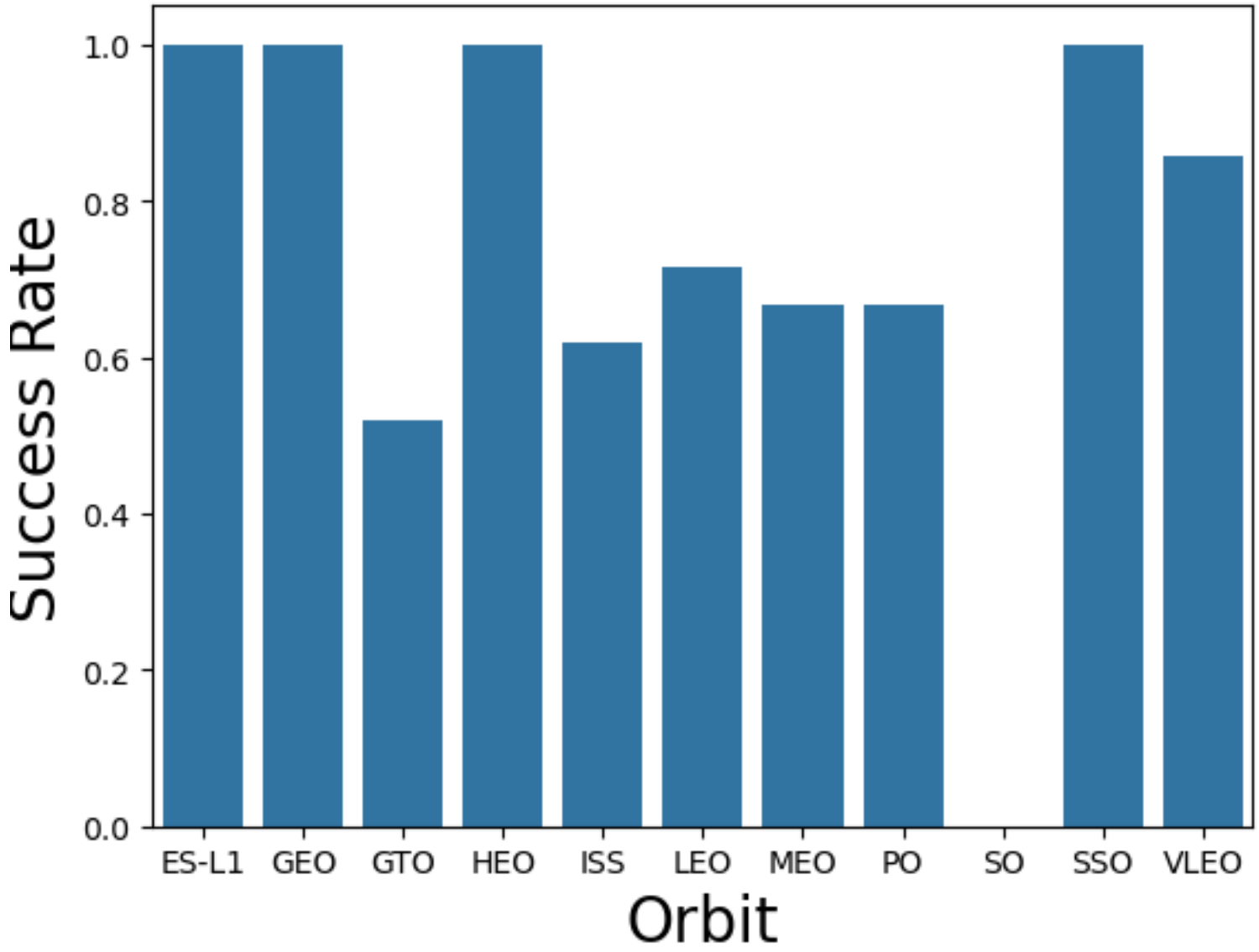
Flight Number vs. Launch Site



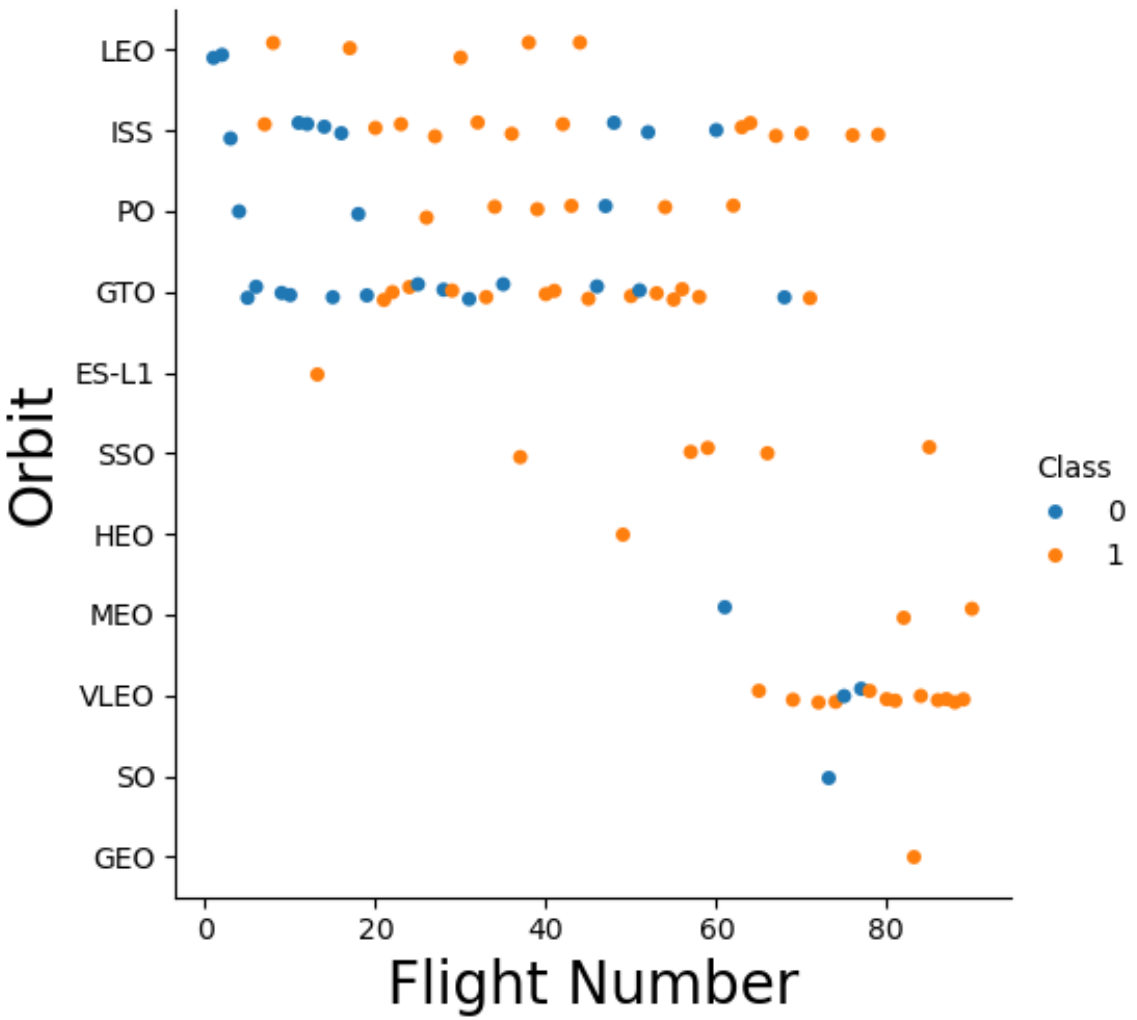
Payload vs. Launch Site



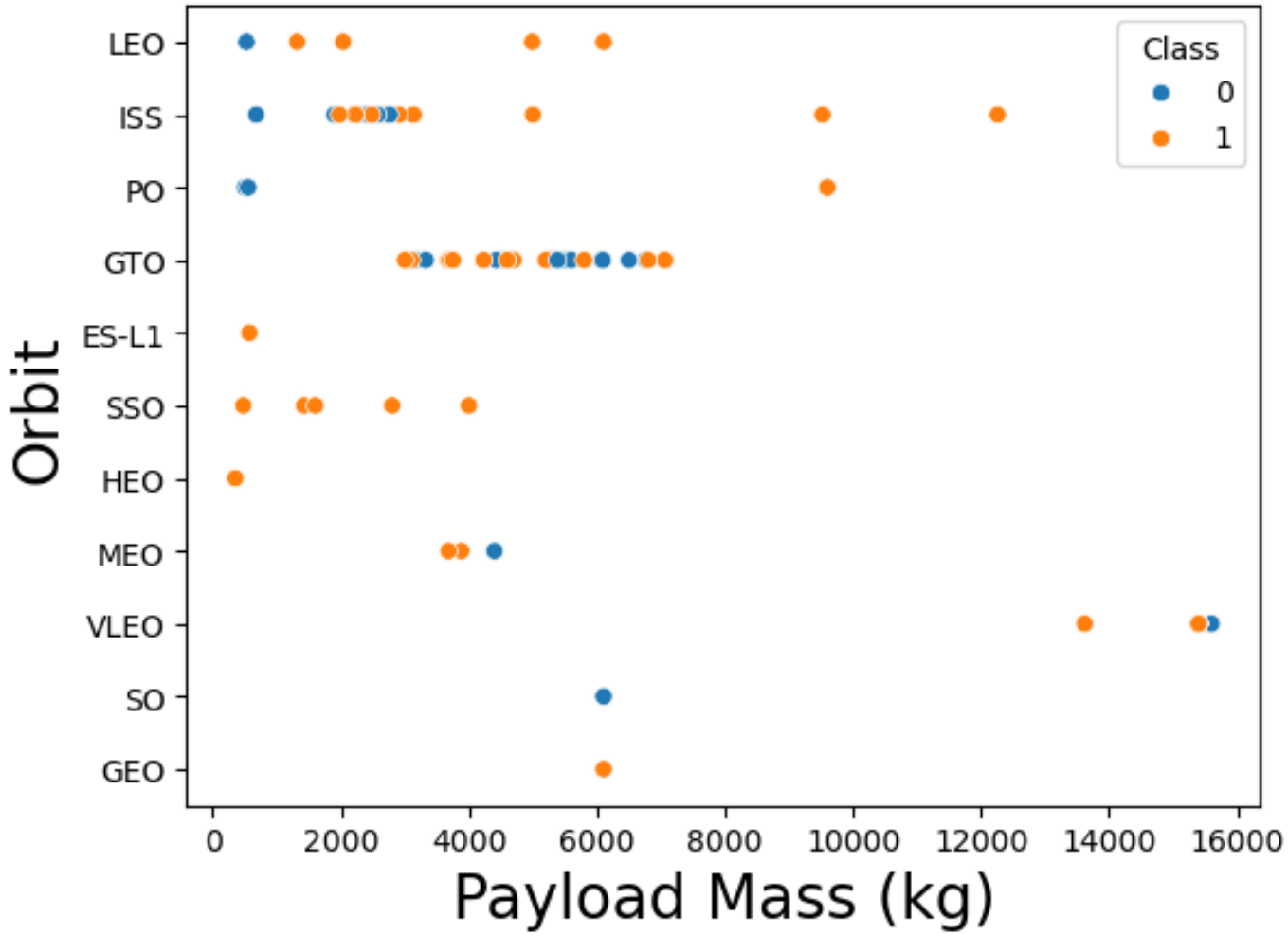
Success Rate vs. Orbit Type



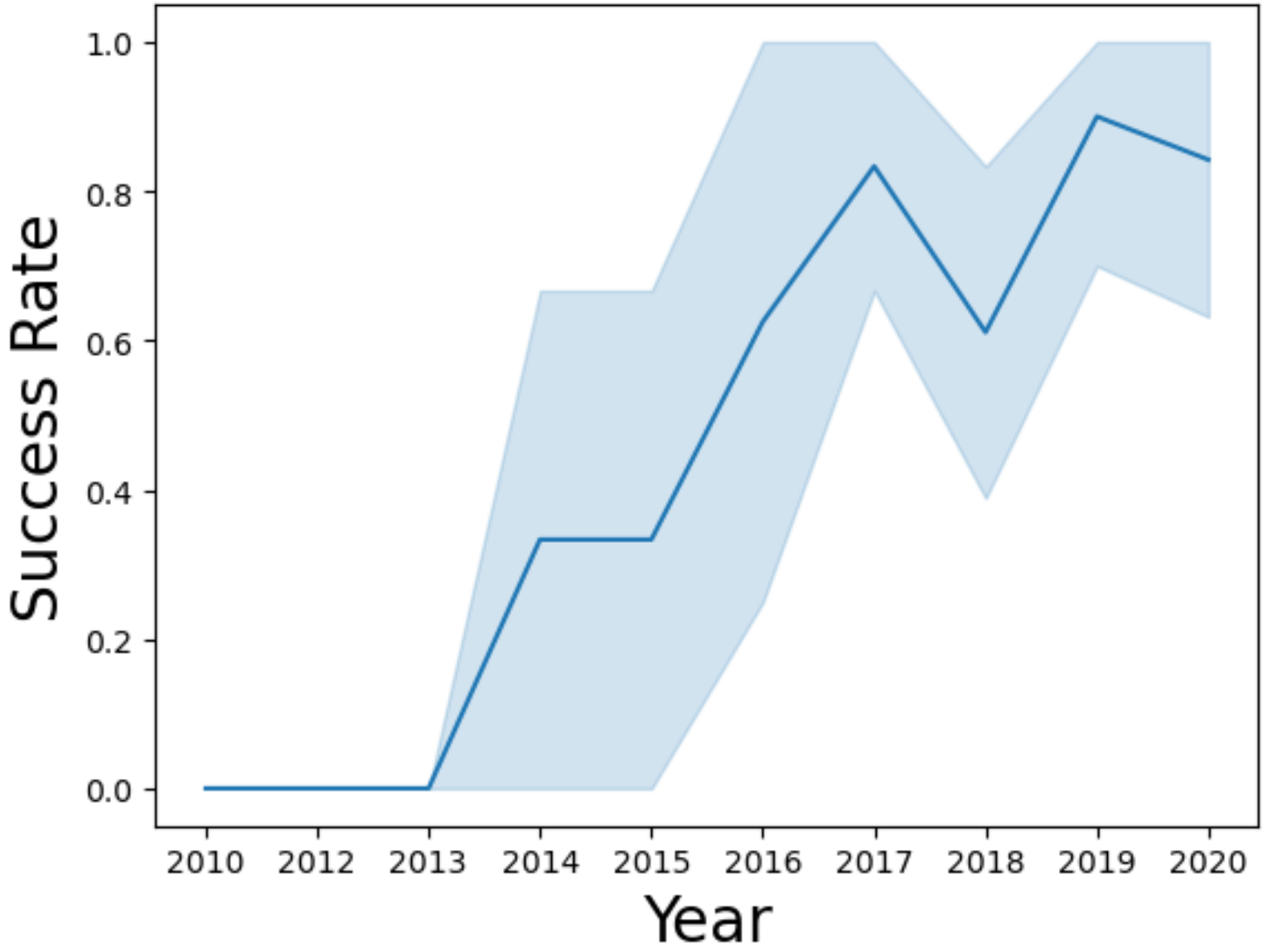
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

All Launch Site Names

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- This selects all the columns from the table and outputs the site names that begin with CCA and limits to the first 5 rows of the queried dataset

Total Payload Mass

- The total payload carried by boosters from NASA:
- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)';

- This adds up all the vales of payload mass column when the customer is NASA

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1:
- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';

- This displays the mean for Average Mass when the booster version column value starts with F9 v1.1

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad
- %sql SELECT Date FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%ground%' LIMIT 1;

- Outputs the date of the first successful ground landing

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 :
- %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%drone%' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;

- I first selected the column l want as output then l queried the data set to only include data where landing outcome is Drone and payload mass is > 4000 and < 600

Total Number of Successful and Failure Missi on Outcomes

- The total number of successful and failure mission outcomes
- %sql SELECT Mission_Outcome, COUNT(*) AS Count FROM SPACEXTABLE GROUP BY Mission_Outcome;

- This outputs the count of the mission outcome byt grouping data using the mission outcome column

Boosters Carried Maximum Payload

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

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
);
```

- I used a sub query to find the booster which had the highest payload mass in the data set

F9 B5 B1048.4

F9 B5 B1049.4

F9 B5 B1051.3

F9 B5 B1056.4

F9 B5 B1048.5

F9 B5 B1051.4

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 :

```
%sql SELECT
  SUBSTR(Date, 6, 2) AS Month,
  Landing_Outcome,
  Booster_Version,
  Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR(Date, 0, 5) = '2015' AND Landing_Outcome LIKE 'Failure (drone
   ship)%';
```

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

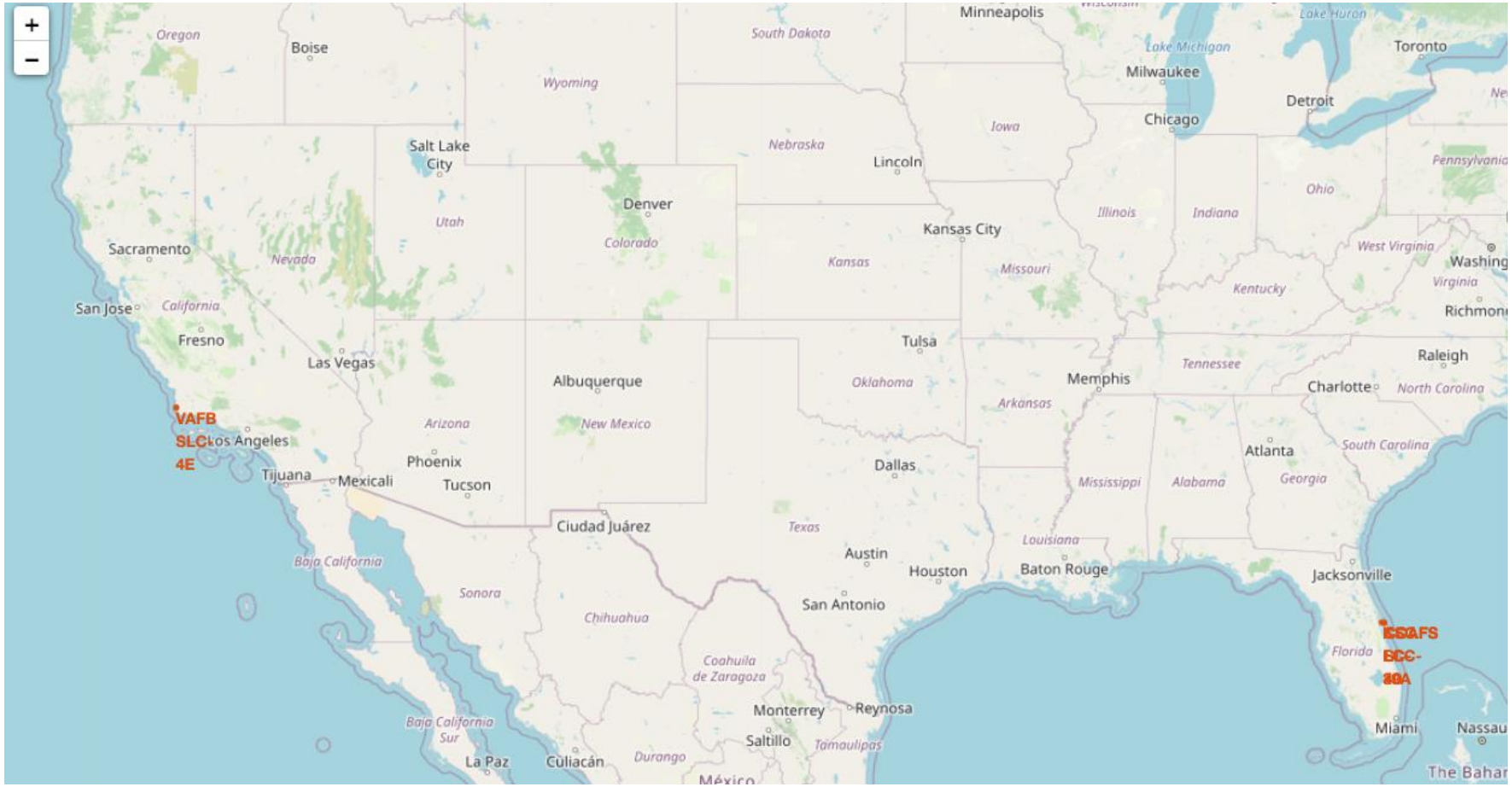
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order :

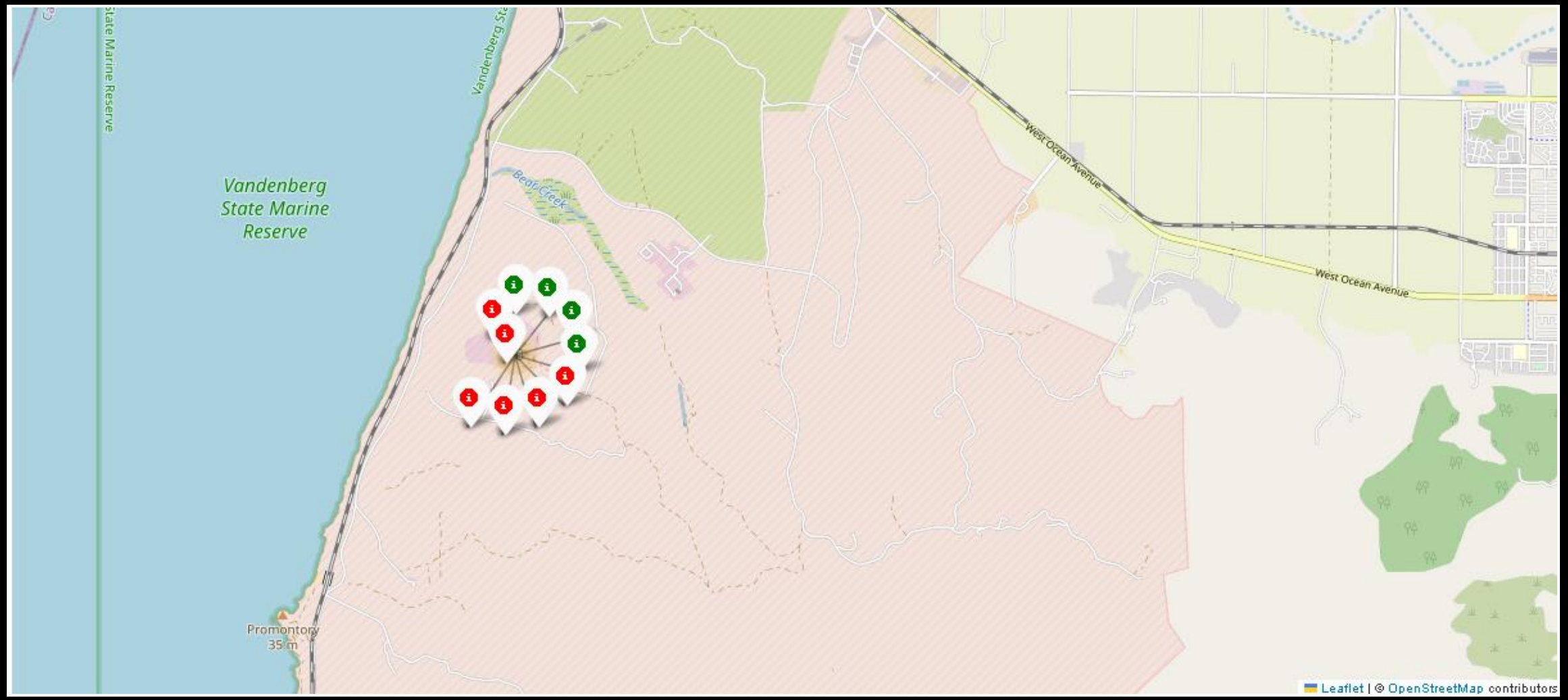
```
%%sql SELECT
    Landing_Outcome,
    COUNT(Landing_Outcome) AS Total_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Total_Count DESC;
```

Section 3

Launch Sites Proximities Analysis

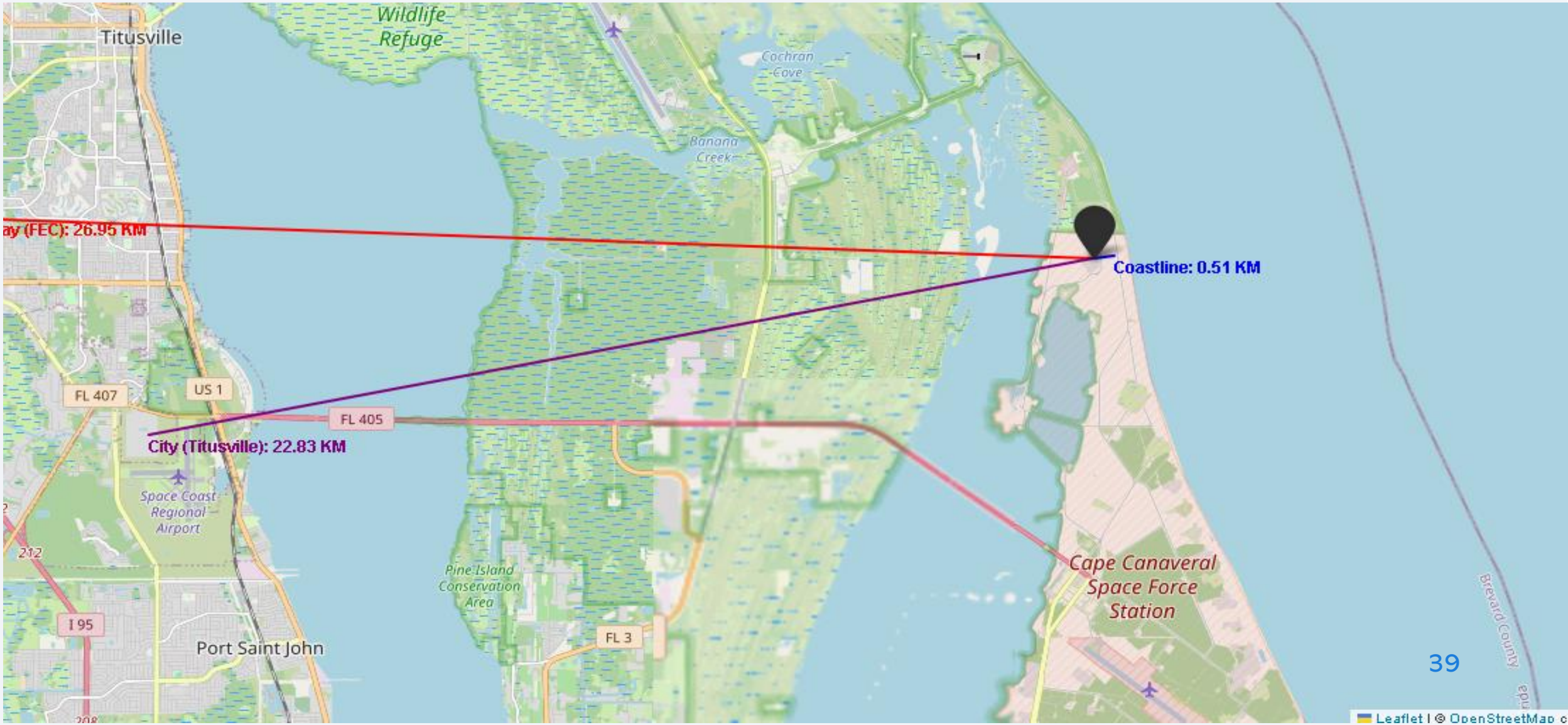
Launch Site Locations

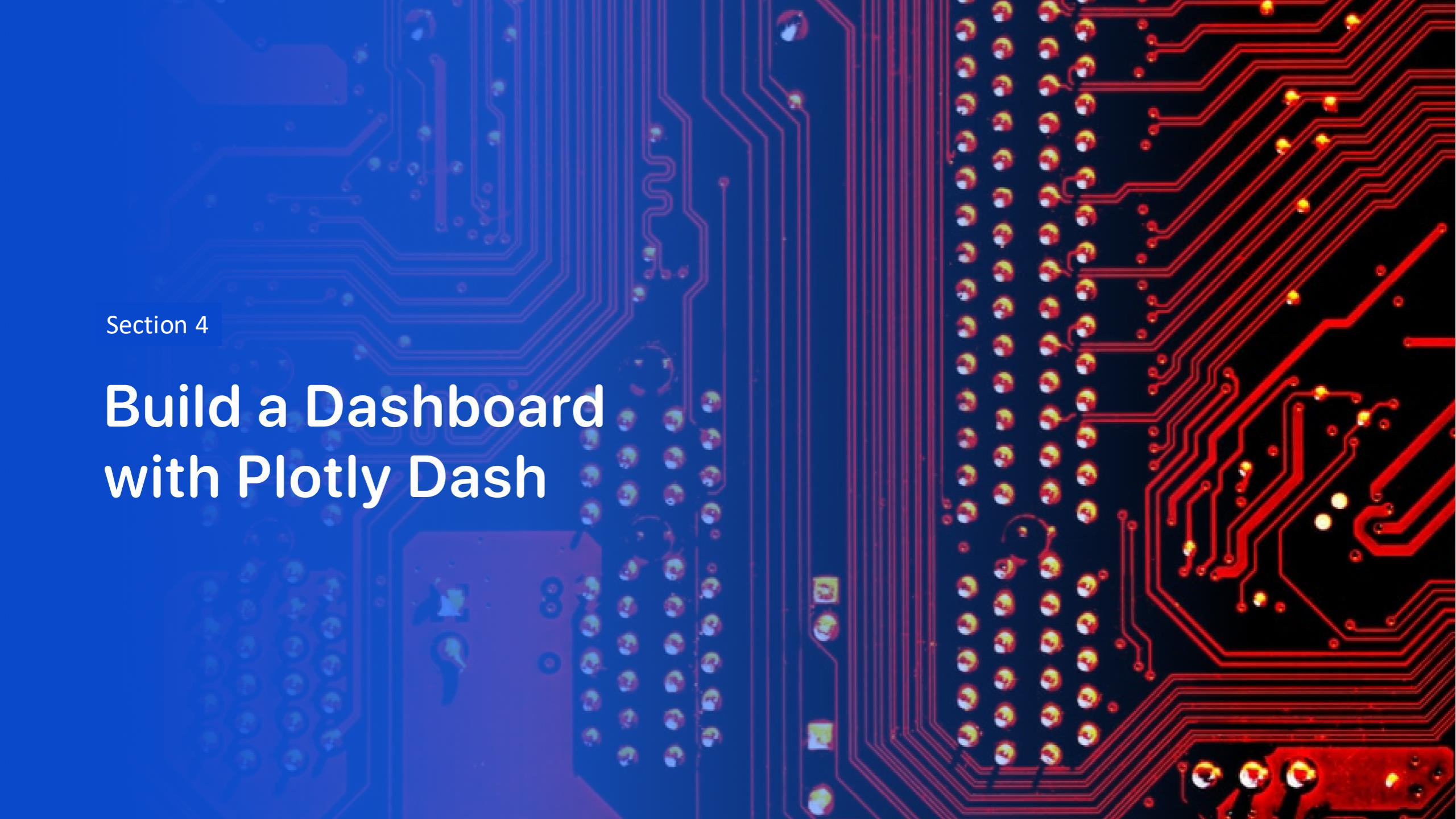




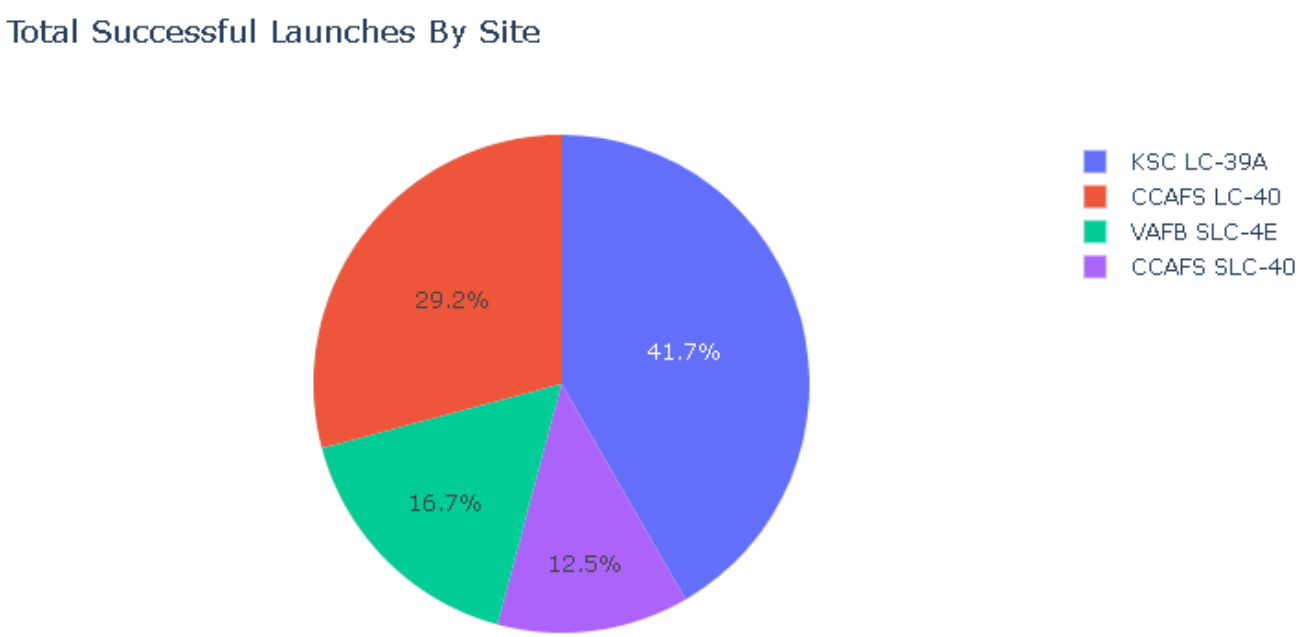
Launch Site Success/Failed markers

Distance comparison

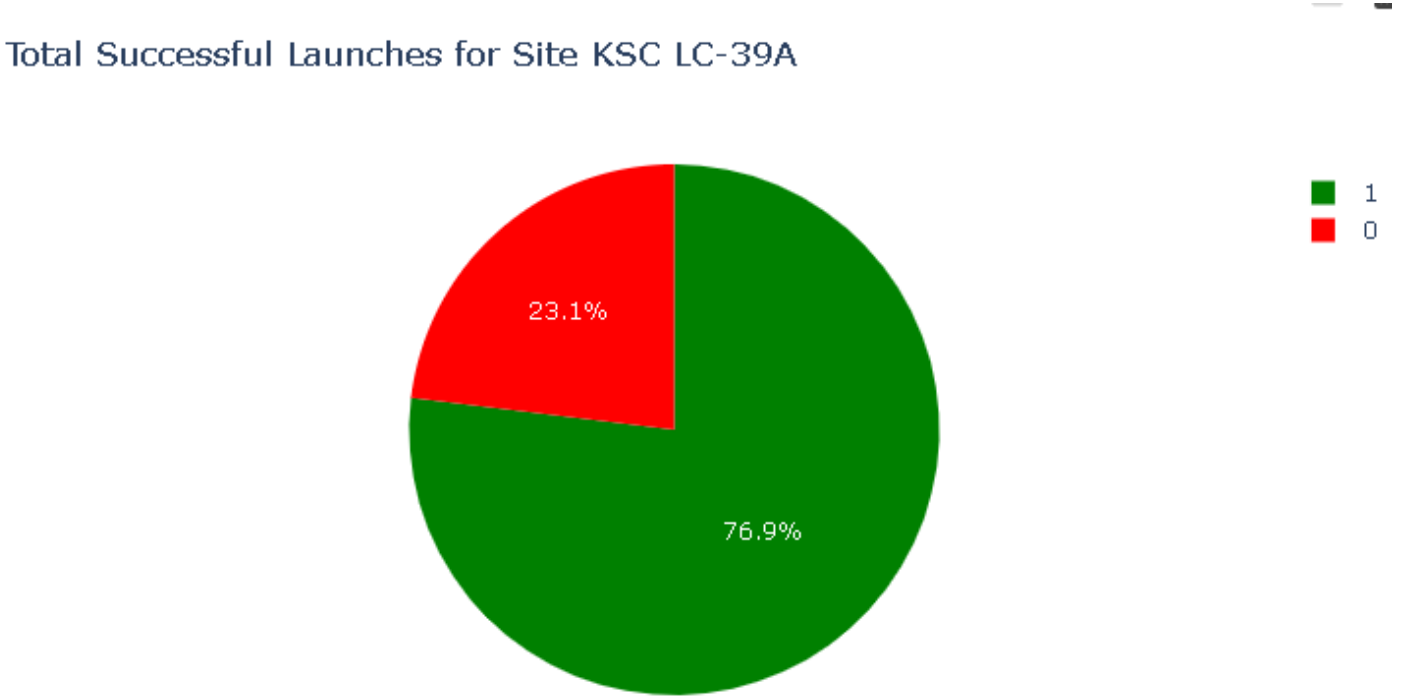



Section 4
Build a Dashboard with Plotly Dash

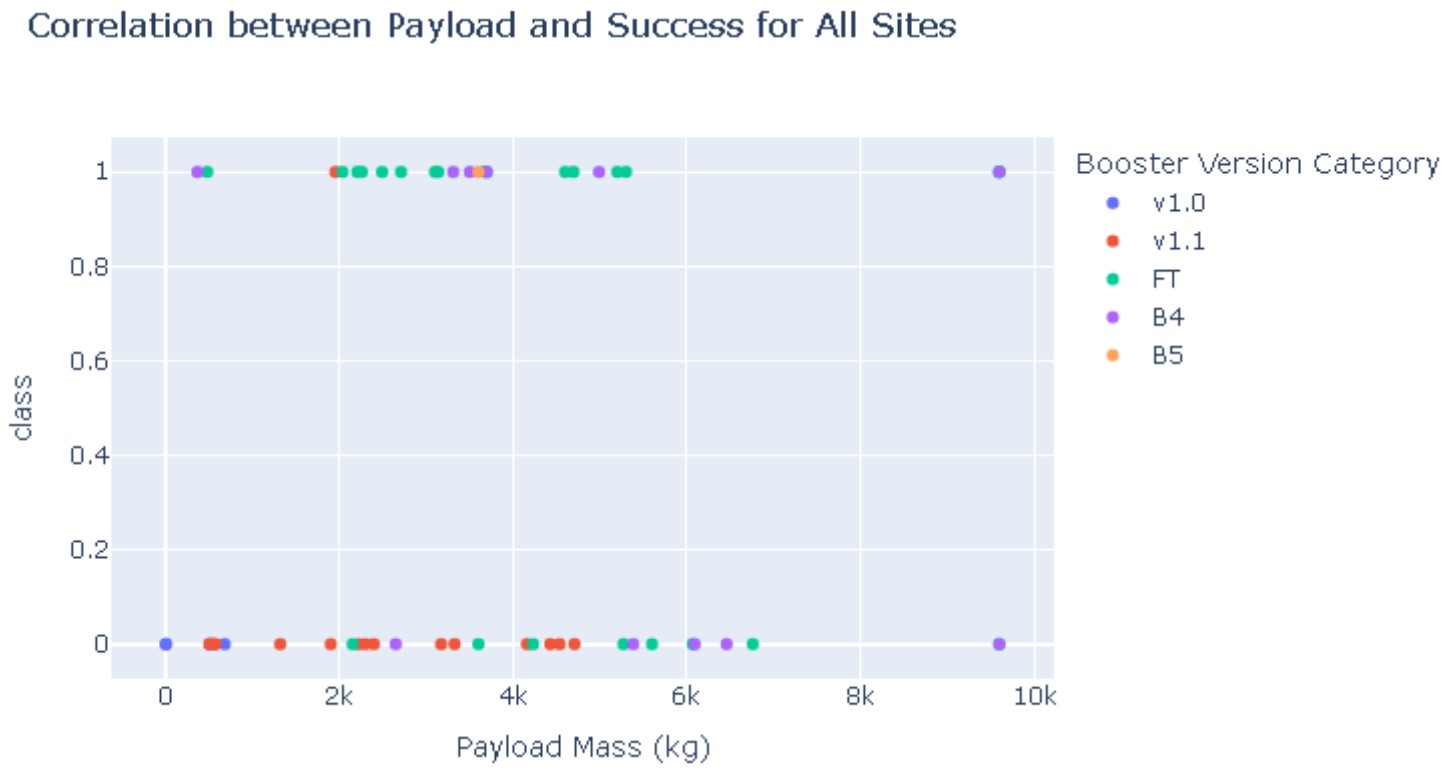
Overall success of each Launch site



Pie chart with the highest success rate



Scatterplot for all launch sites



Section 5

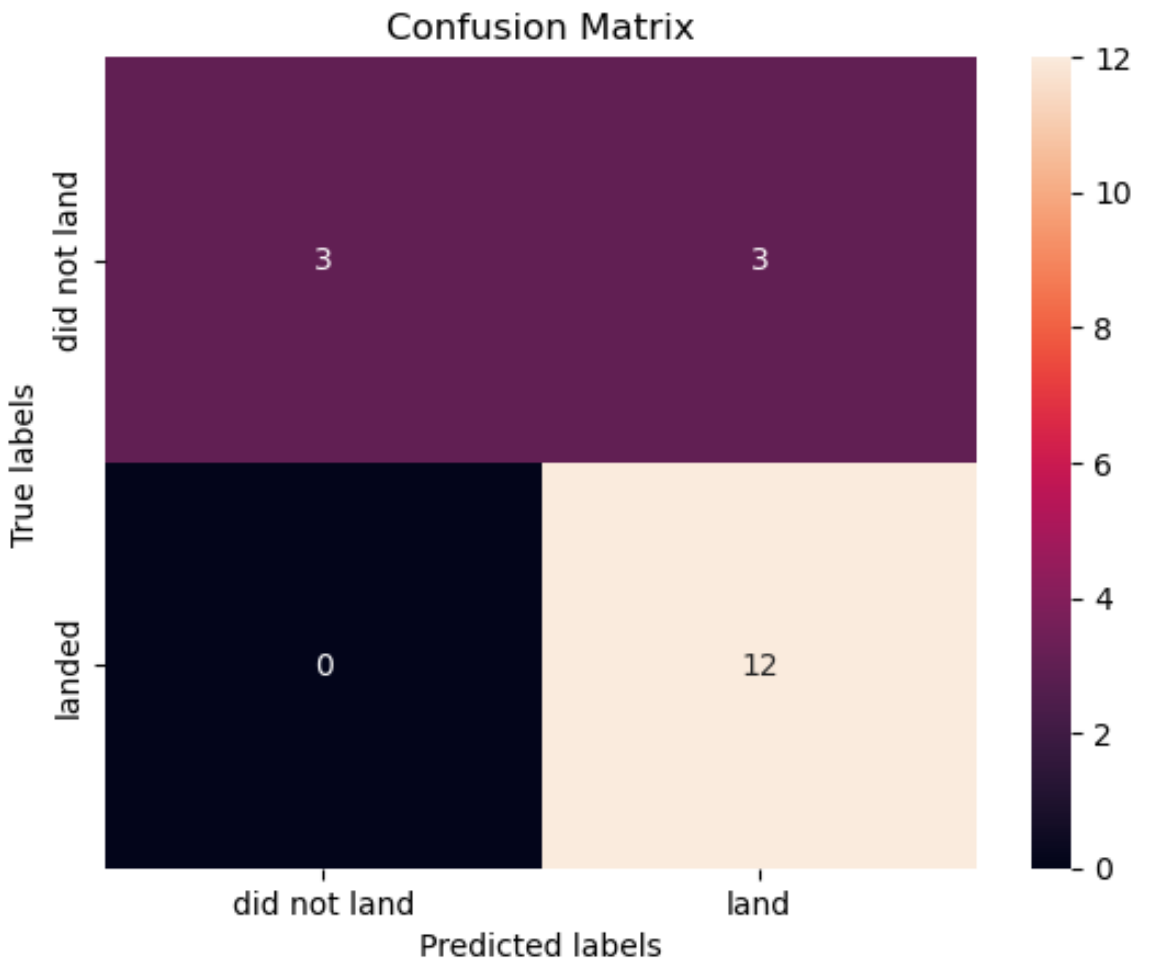
Predictive Analysis (Classification)

Classification Accuracy

Results:

- The **Support Vector Classifier (SVC)** emerged as the most effective model for this dataset, achieving the highest overall accuracy of approximately **85%** and strong performance across all metrics.
- The **Logistic Regression** and **Decision Tree** models performed comparably well, both reaching an accuracy of approximately **81%** and balanced F1 scores.
- The **K-Nearest Neighbors (KNN)** model demonstrated significantly lower performance, particularly in precision and F1-score, indicating it was the least suitable model for predicting SpaceX launch outcomes among those tested.

Confusion Matrix for Support Vector Machines

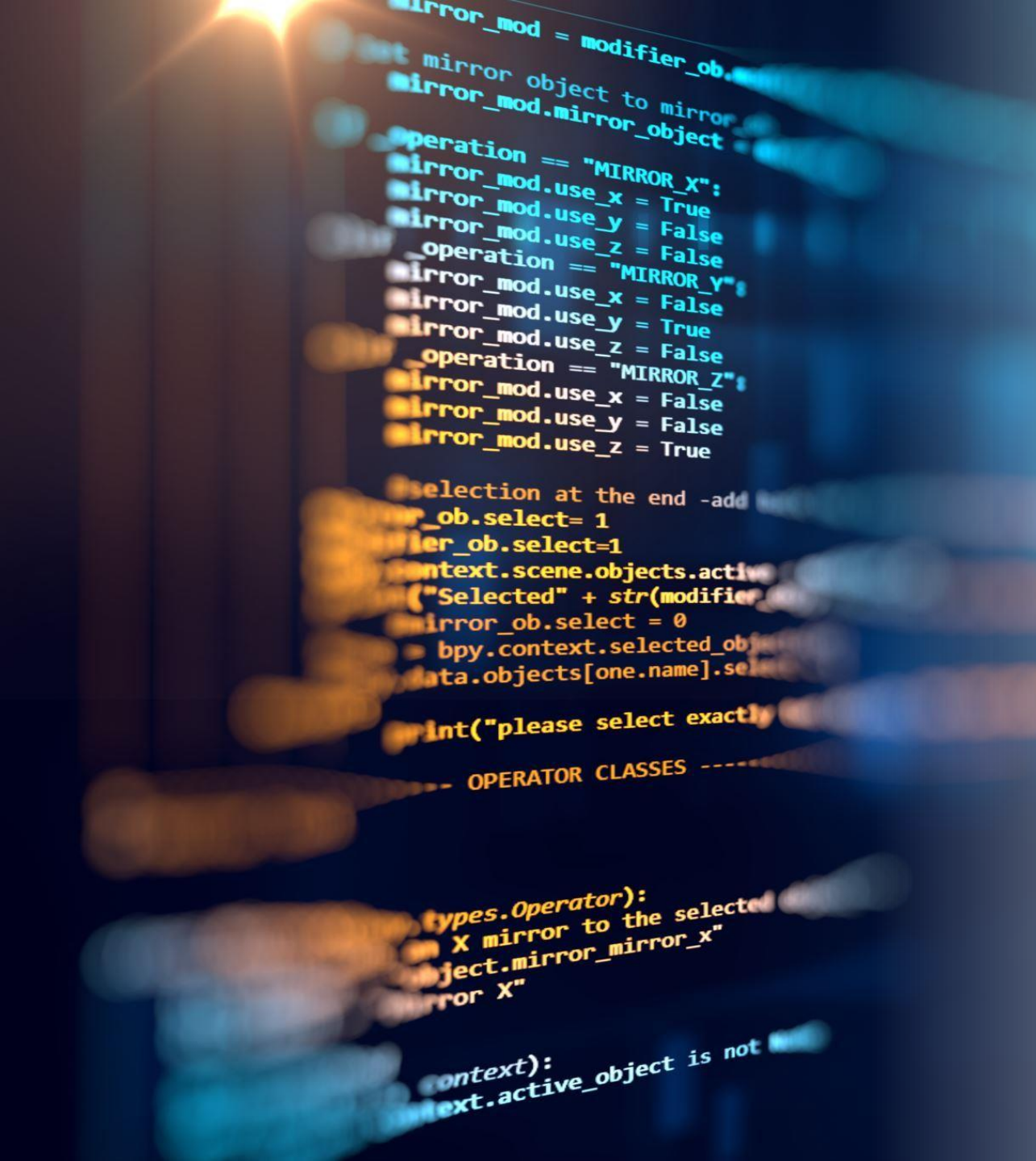


Conclusions

- In this project l drew a couple of conclusions :
  - The success rate started to rise in 2013 up to 2020
  - The success rate was high for the following orbits (ES-L1 ,GEO ,HEO and SSO)
  - Different Launch Sites Differ in the payload of the F 9
  - I also noticed that launch sites are usually located far from public infrastructures
  - Support Vector Machines is the best model for prediction for this data set
  - Kenedy Space Center has the highest success rate amongst all the launch sites

Appendix

- In this project I used various parameters for the predictions models and GridSearch alongside those models



Git-hub File Link

- SQL exploratory data Analysis
  https://github.com/Emitrisia/Emitrisia/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- API data collection
  https://github.com/Emitrisia/Emitrisia/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
- Web Scrapping
  https://github.com/Emitrisia/Emitrisia/blob/main/jupyter-labs-webscraping.ipynb
- Data Wrangling
  https://github.com/Emitrisia/Emitrisia/blob/main/labs-jupyter-spacex-Data%20wrangling%20(2).ipynb
- Dashboards
  https://github.com/Emitrisia/Emitrisia/blob/main/spacex-dash-app.py
- Machine Learning Prediction
  https://github.com/Emitrisia/Emitrisia/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
- Folium map
  https://github.com/Emitrisia/Emitrisia/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Thank you!