IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

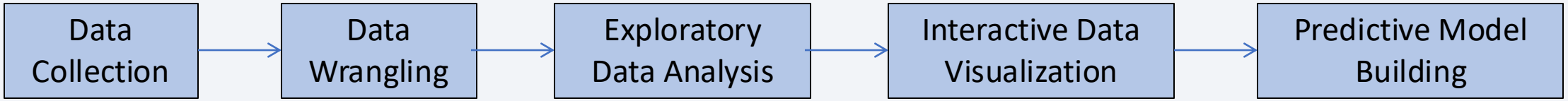
<Shaowen Feng>
<June 2025>

Outline

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

Executive Summary

- Summary of methodologies



- Summary of Results

  - Best model chosen: Decision Tree model
  - Model performance: Accuracy score: 87.7%

Introduction

- **Project background and context**
  - SpaceX advertised Falcon 9 rocket launches with a much less cost than other providers because it can reuse the first stage. In order to determine the cost of each launch, we need to determine if the first stage will land successfully.

- **Problems I want to find answers**
  - What factors affect the successful landing rate?
  - How do I gather the necessary information?
  - What method should I choose to analyze the data?
  - How do I present the findings to my audience?
  - What model should I choose to predict the outcome of landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Making a get request to the SpaceX API
  - Performing web scraping to collect Falcon 9 historical launch data from a Wikipedia page
- Perform data wrangling
  - Creating a Class variable to convert outcome into training label with 1 as success and 0 as failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a cross-validation object finding the best hyperparameters for the following classfication models: SVM, classification trees, Logistic Regression and KNN

Data Collection

- Method 1: Collecting SpaceX Falcon 9 Launch data by SpaceX API request with the following URL: “https://api.spacexdata.com/v4/launches/past”

- Method 2: Collecting SpaceX Falcon 9 Launch data by web scraping process from Wikipedia page with the title “List of Falcon 9 and Falcon Heavy launches” https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

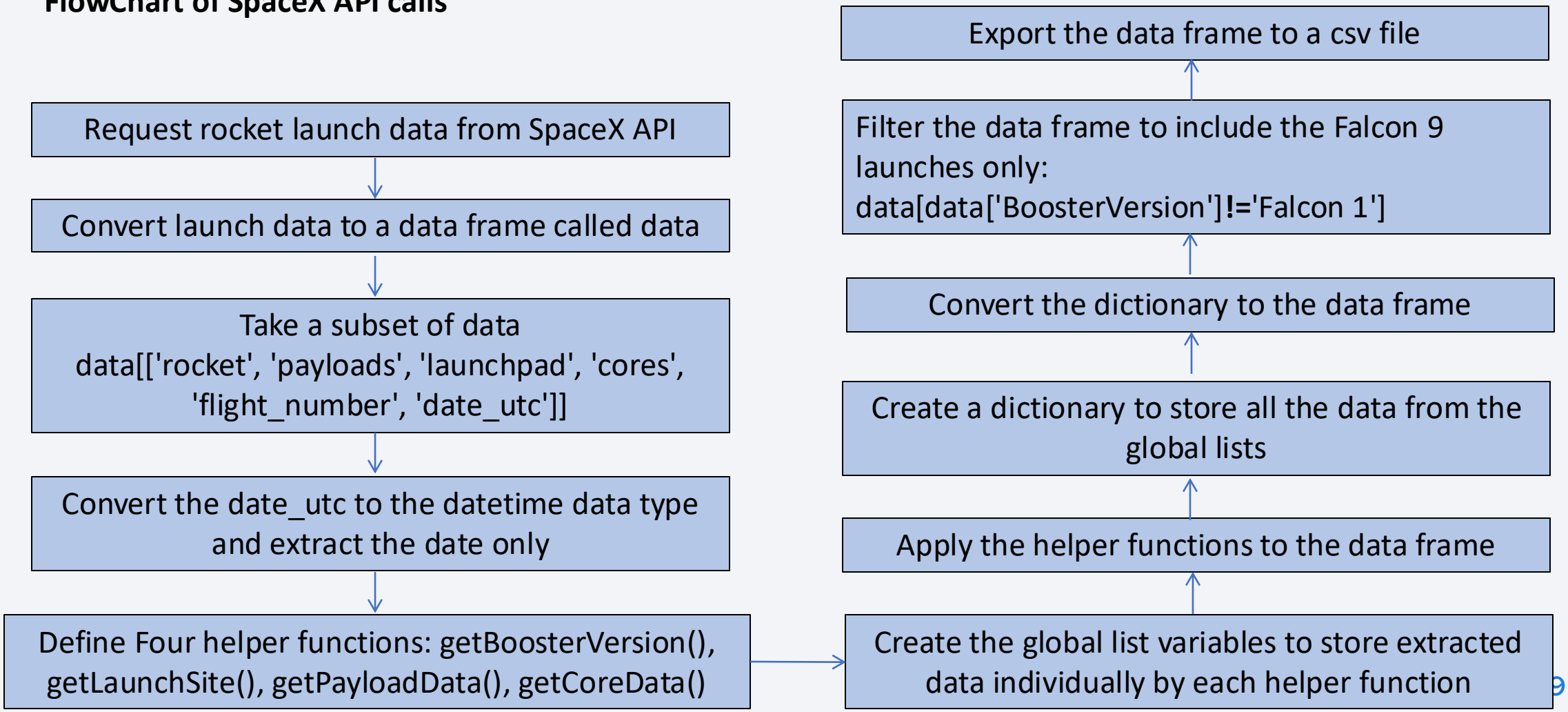
Key data collection phrases

- Request the SpaceX launch data using the GET request and convert it to a data frame
- Take a subset of the original data frame keeping only the necessary features
- Use a series of of helper functions to parse information using ID numbers in the launch data
- Filter the data frame to only include the Falcon 9 launches data

Here is the GitHub URL of the completed SpaceX API calls notebook:
https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – SpaceX API

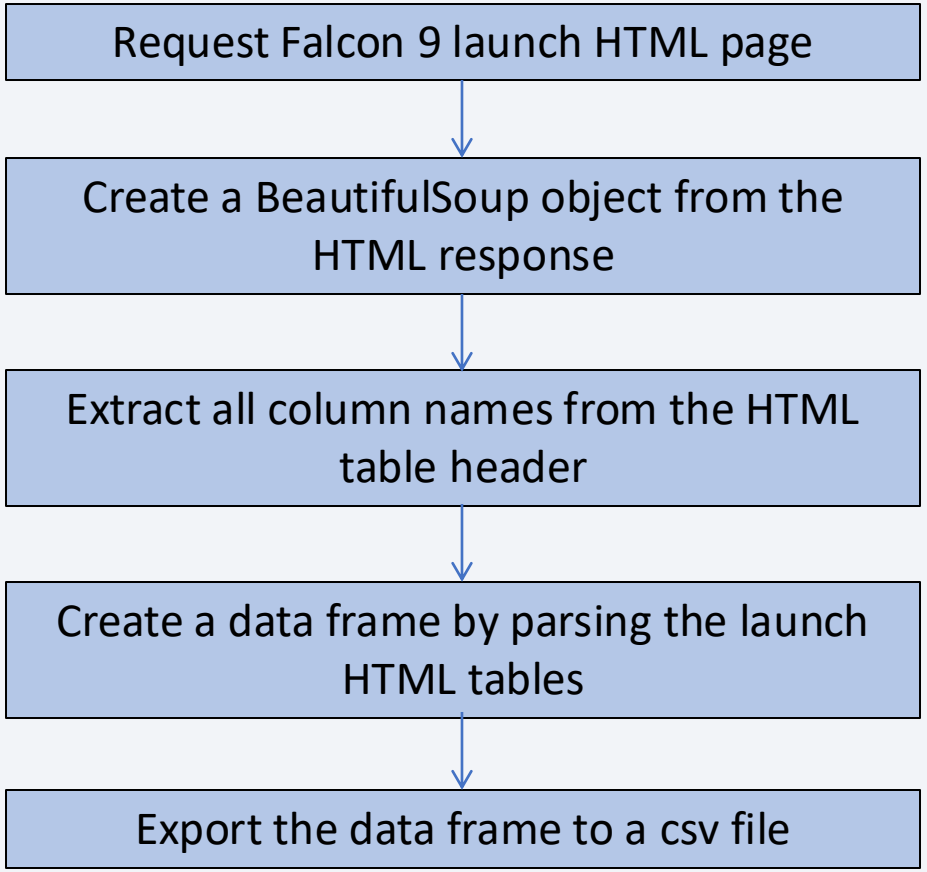
FlowChart of SpaceX API calls



Data Collection - Scraping

Key data collection phrases by web scraping

- Web scrape Falcon 9 launches data with BeautifulSoup
  - Extract a Falcon9 launch records HTML table from Wikipedia
  - Parse the table and convert it to a Pandas data frame
- Here is GitHub URL of the completed web scraping notebook: https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/jupyter-labs-webscraping.ipynb



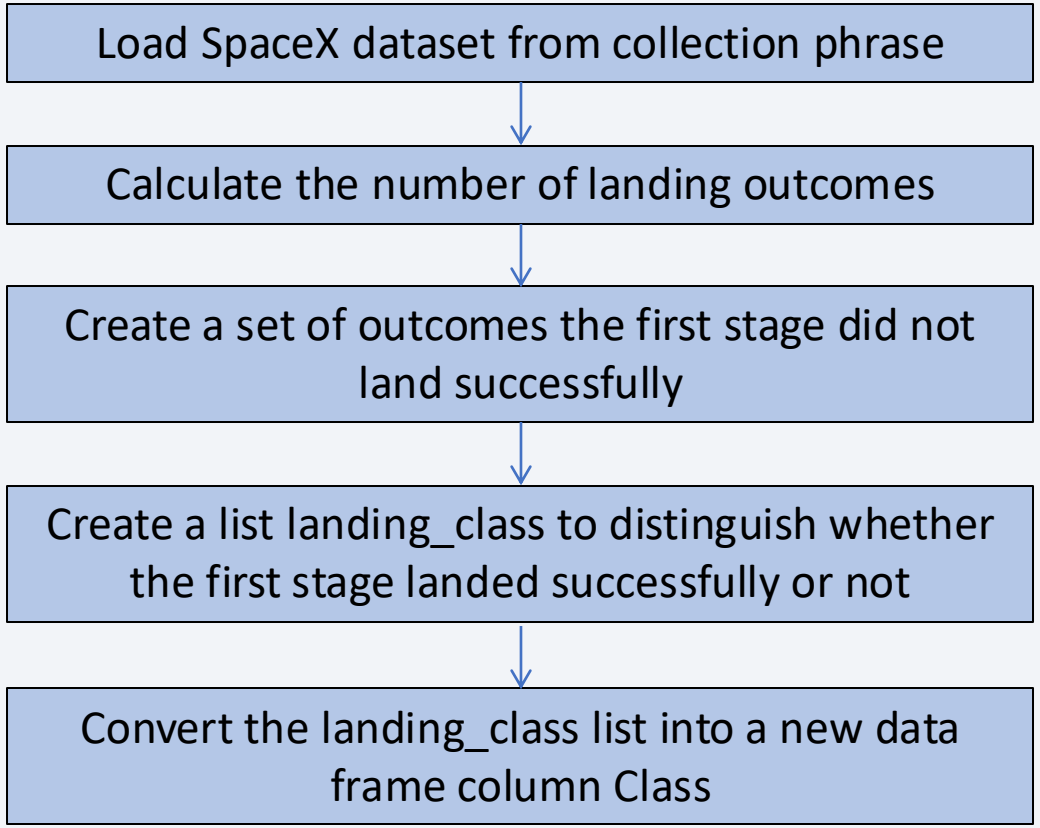
Data Wrangling

Key phrases of data wrangling process

- Replace missing value with the mean
- Convert outcomes into training class labels

Here is the GitHub URL of my completed data wrangling related notebooks: https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/labs-jupyter-spacex-Data%20wrangling.ipynb

Flow chart of converting outcomes into training class



EDA with Data Visualization

- Scatter plot of the FlightNumber vs. PayloadMass to indicate how they affect the launch outcome
- Scatter chart of the FlightNumber vs. Launch Site to visualize how they affect the launch outcome
- Scatter chart of the Payload Mass vs. Launch Site to indicate how they affect the launch outcome
- Bar chart for the success rate of each orbit site
- Scatter plot of the FlightNumber vs. Orbit to visualize how they affect the launch outcome
- Scatter chart of the Payload Mass vs. Orbit to indicate how they affect the launch outcome
- Line chart of the success rate over years trend

This is the GitHub URL of my completed EDA with data visualization notebook:
https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/edadataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display five records where launch sites begin with the string “CCA”
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived
- 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 all the booster_versions that have carried the maximum payload mass
- List 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 between the date 2010-06-04 and 2017-03-20, in descending order

Here is the GitHub URL of my completed EDA with SQL notebook:
https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Create *an orange circle* using folium.Circle to highlight *NASA Johnson Space Center* and a*dd a marker object as a popup label showing its name*
- *Generate circles to highlight each SpaceX launch site and add markers as popup labels for them*
- *Create a marker cluster object to simplify a map containing many markers at the same coordinate*
- *Create marker object for each launch to indicate if the launch was successful or failed*
- *Add a mouse position object to get coordinate for a mouse over a point on the map*
- *Create a polyline object to draw a line between a launch site and a selected point*
- *Add a distance marker to indicate the distance between a launch site and a selected point*

Here is the GitHub URL of my completed interactive map with Folium map:
https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Create an interactive pie chart to visualize the successful launch rates both among all the launch sites and for a specific launch site
- Create an interactive scatter chart to indicate the correlation between Payload Mass and success for all launch sites or the choosing launch site

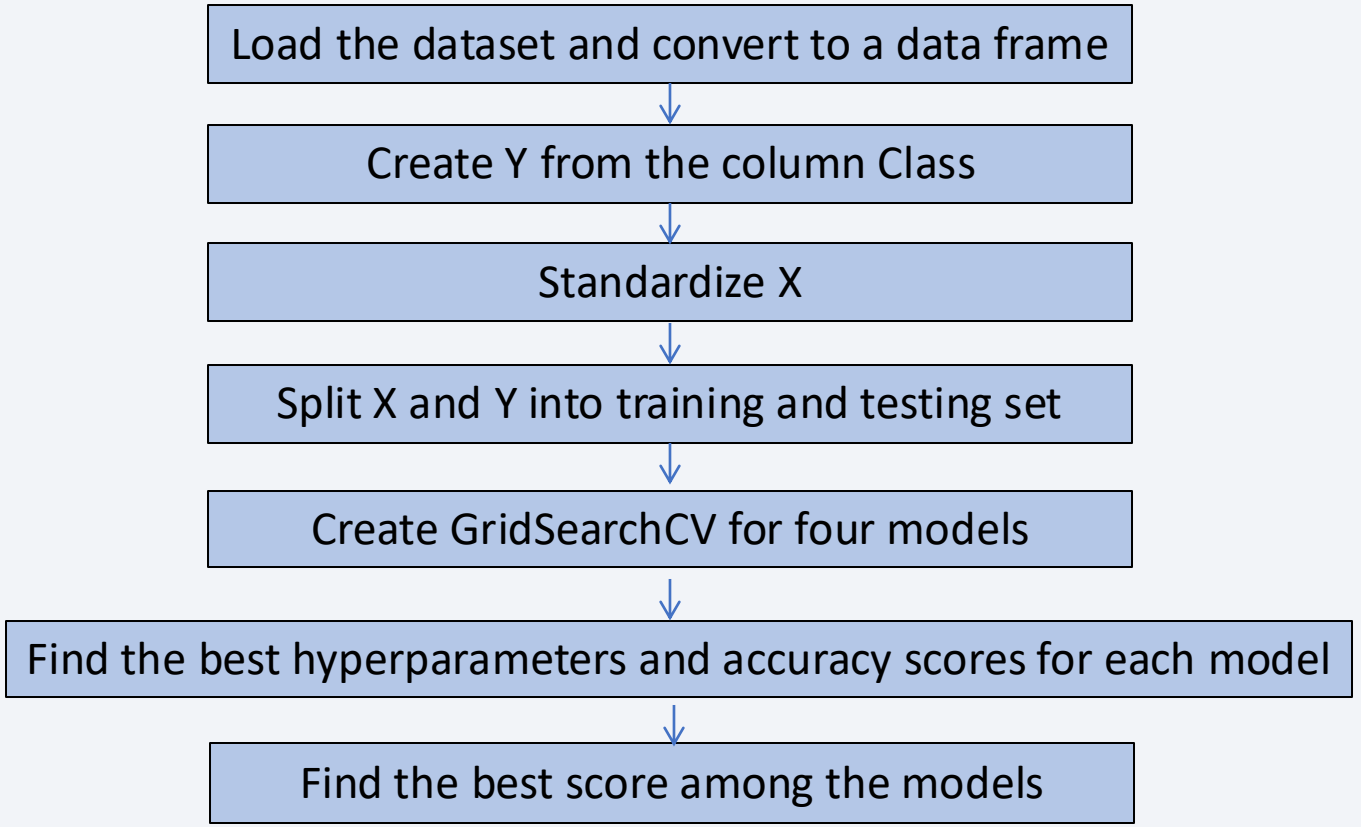
Here is the GitHub URL of my completed Plotly Dash lab:
https://github.com/Mimicorn/Space-Repo/blob/main/python/spacex_dash_app.py

Predictive Analysis (Classification)

Key phrases of model development

- Create a column for the class y
- Standardize the data X
- Split the data into training set and testing set
- Create a cross-validation GridSearchCV object for different machine learning models
- Fit the GridSearchCV object with the training data and find the best hyperparameters and the accuracy score for each model
- Compare the accuracy score among the four classification models and find the best one

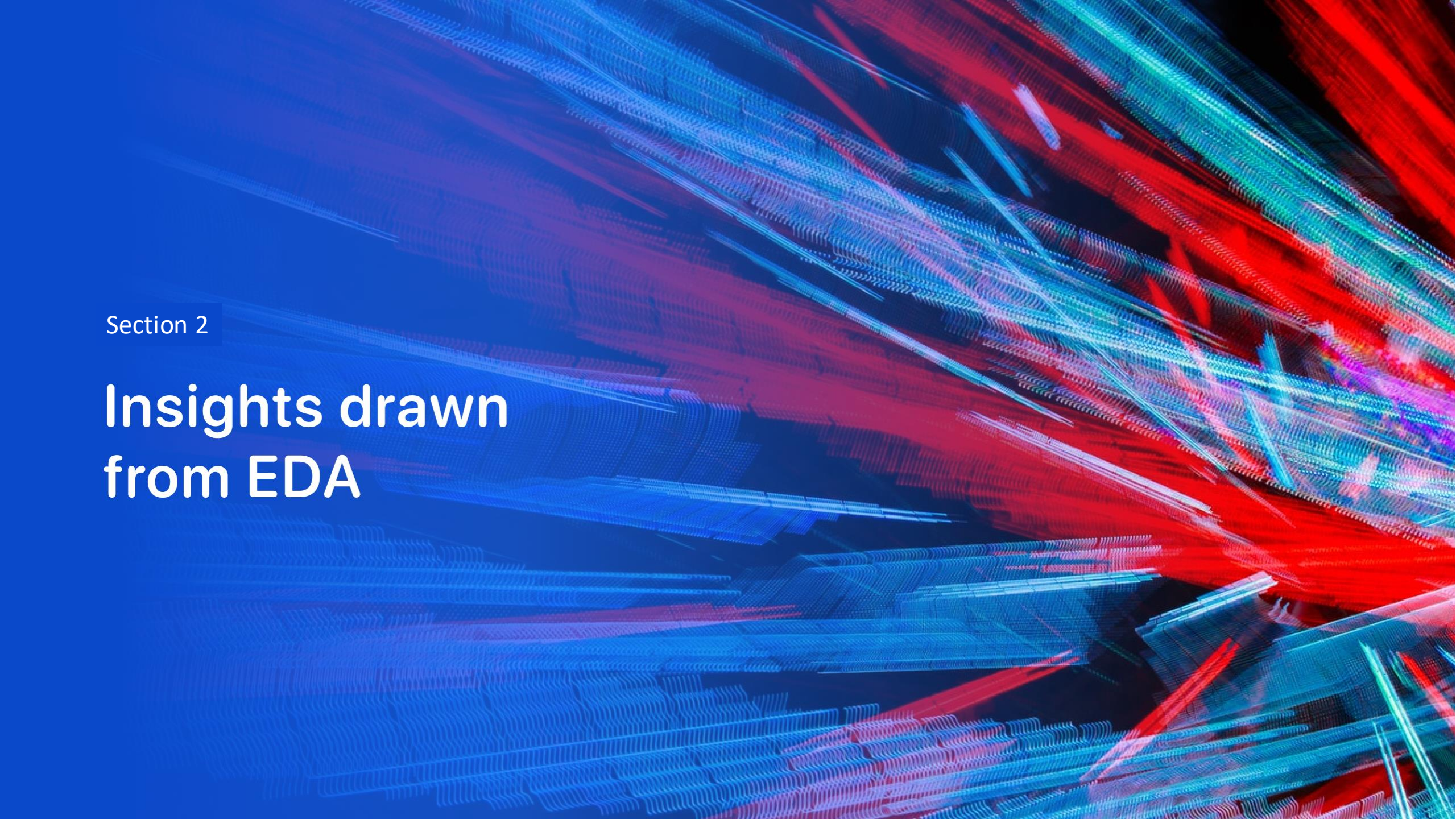
Flow chart of model development



Here is the GitHub URL of my completed predictive analysis lab:
https://github.com/Mimicorn/Space-Repo/blob/main/notebooks/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

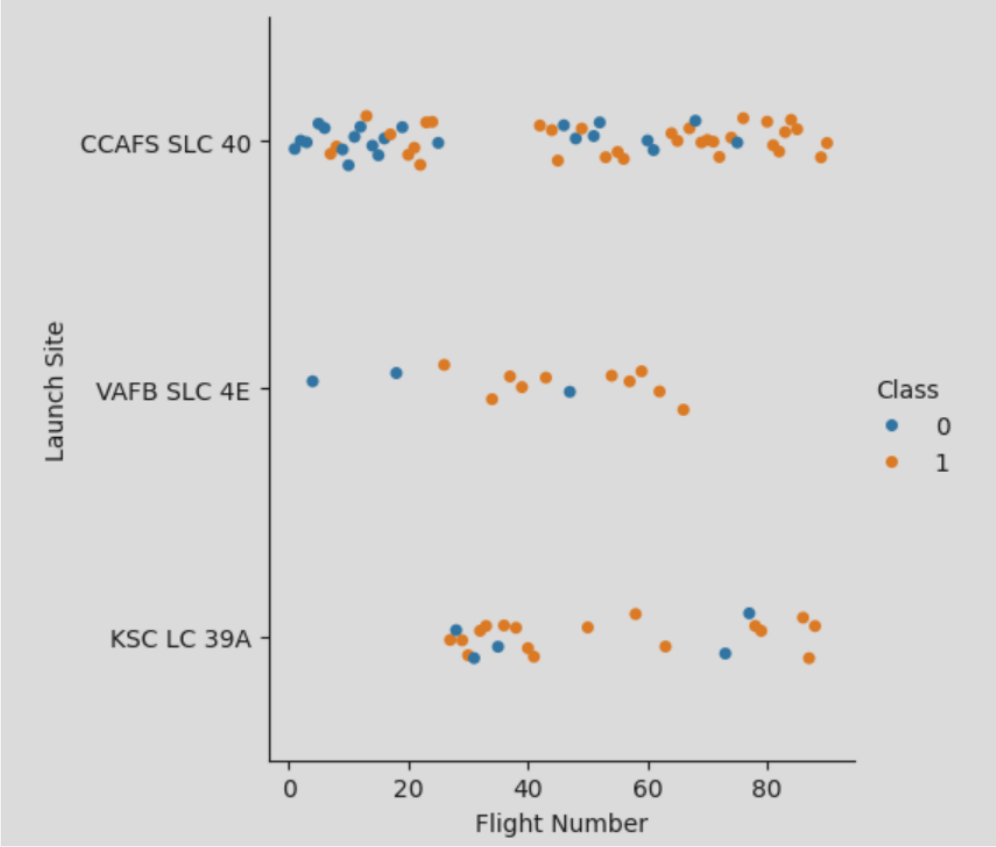
Results

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


Section 2
Insights drawn from EDA

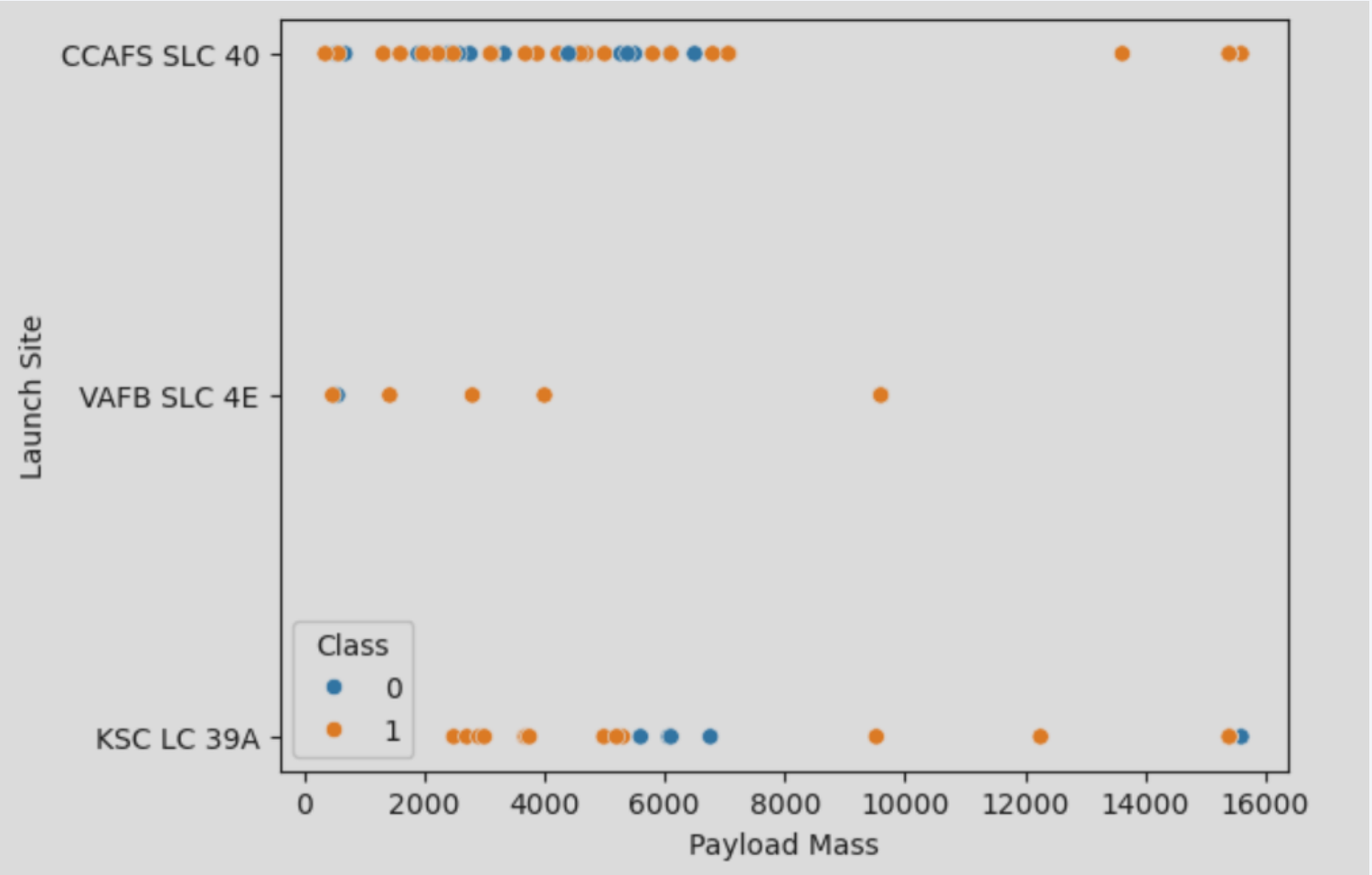
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site with Class value 1 as a successful landing and 0 as a failed landing



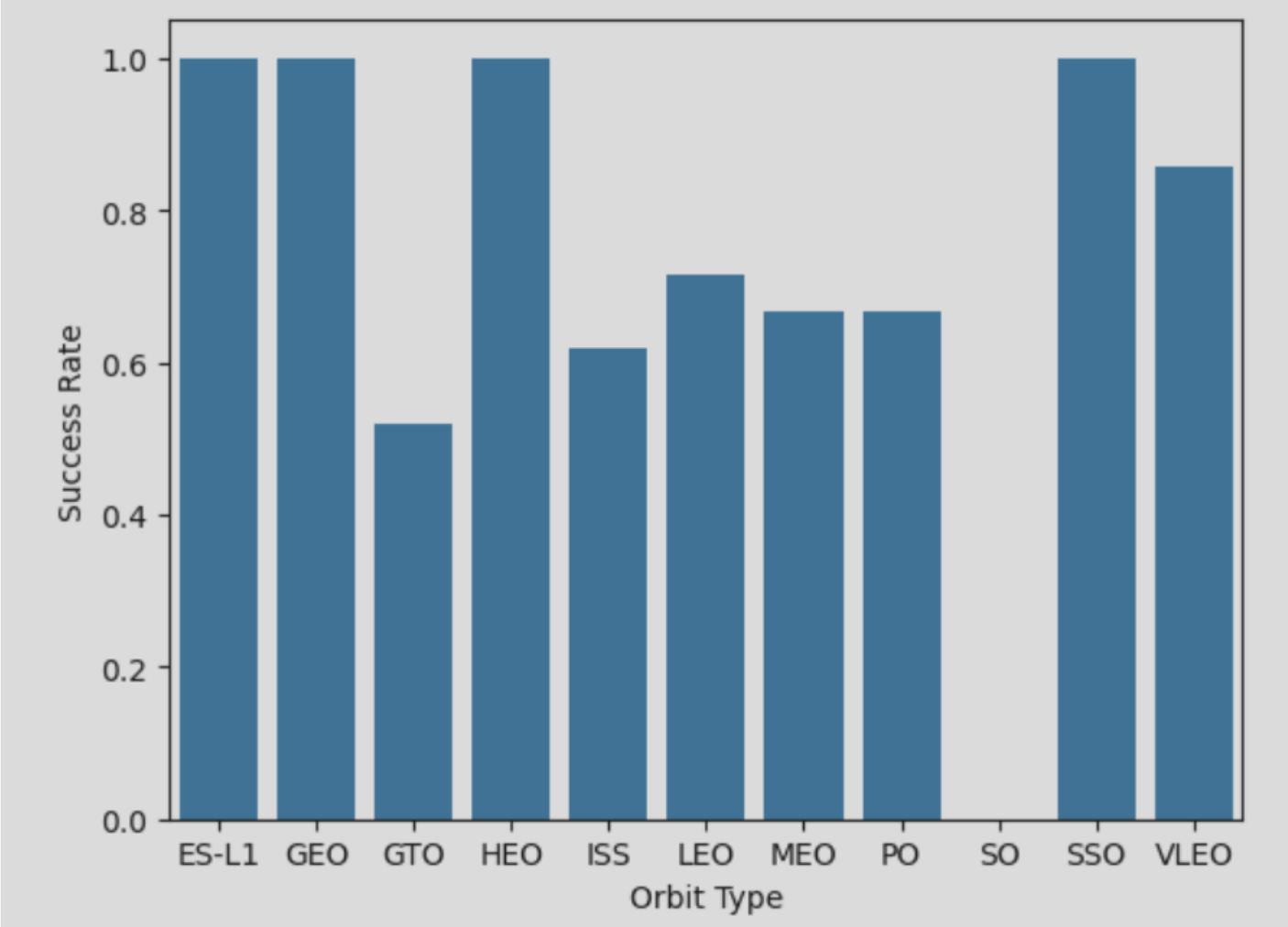
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site with Class value 1 indicating a successful landing and 0 indicating a failed landing



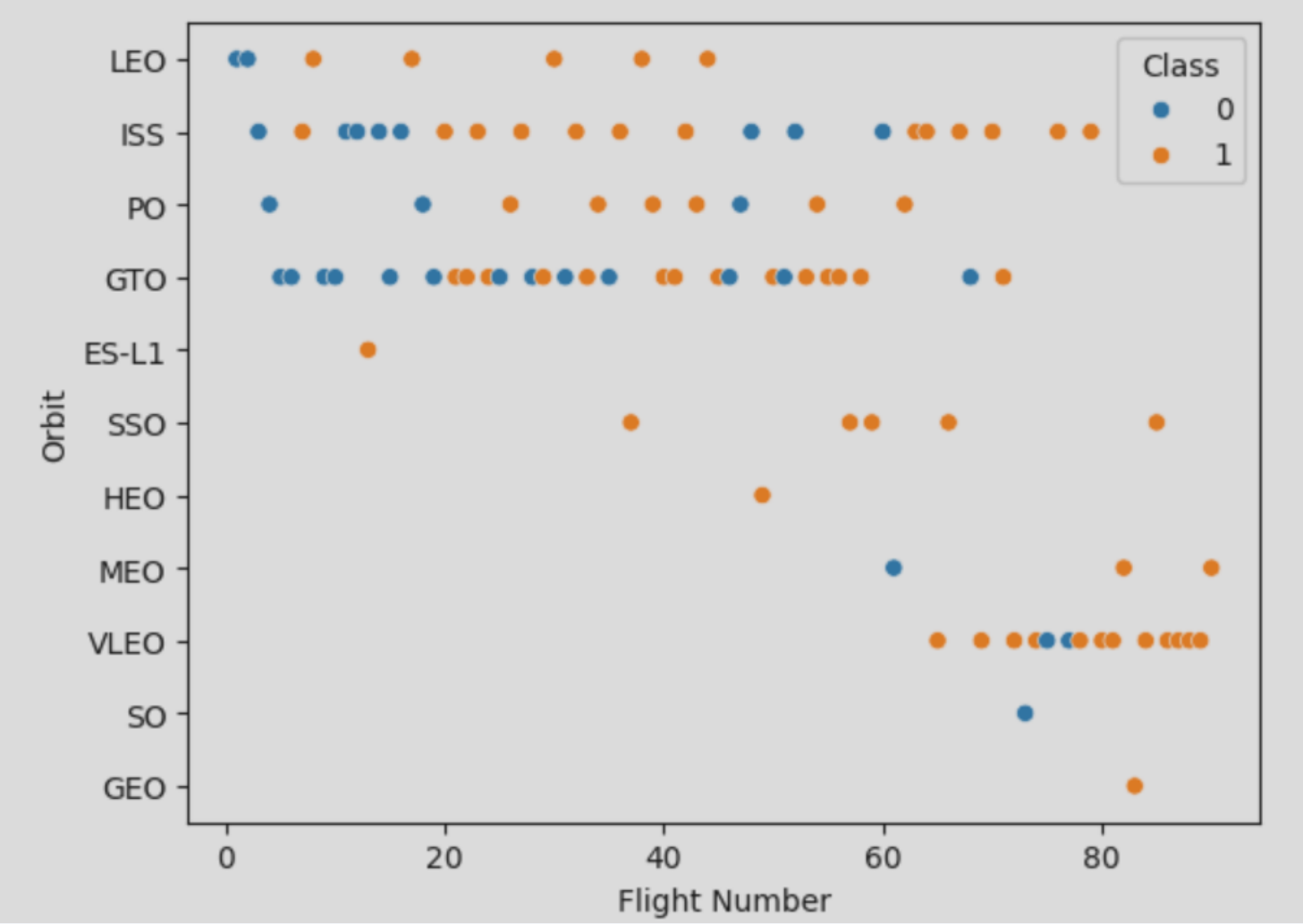
Success Rate vs. Orbit Type

- Bar chart visualizing the relationship between success rate of each orbit type



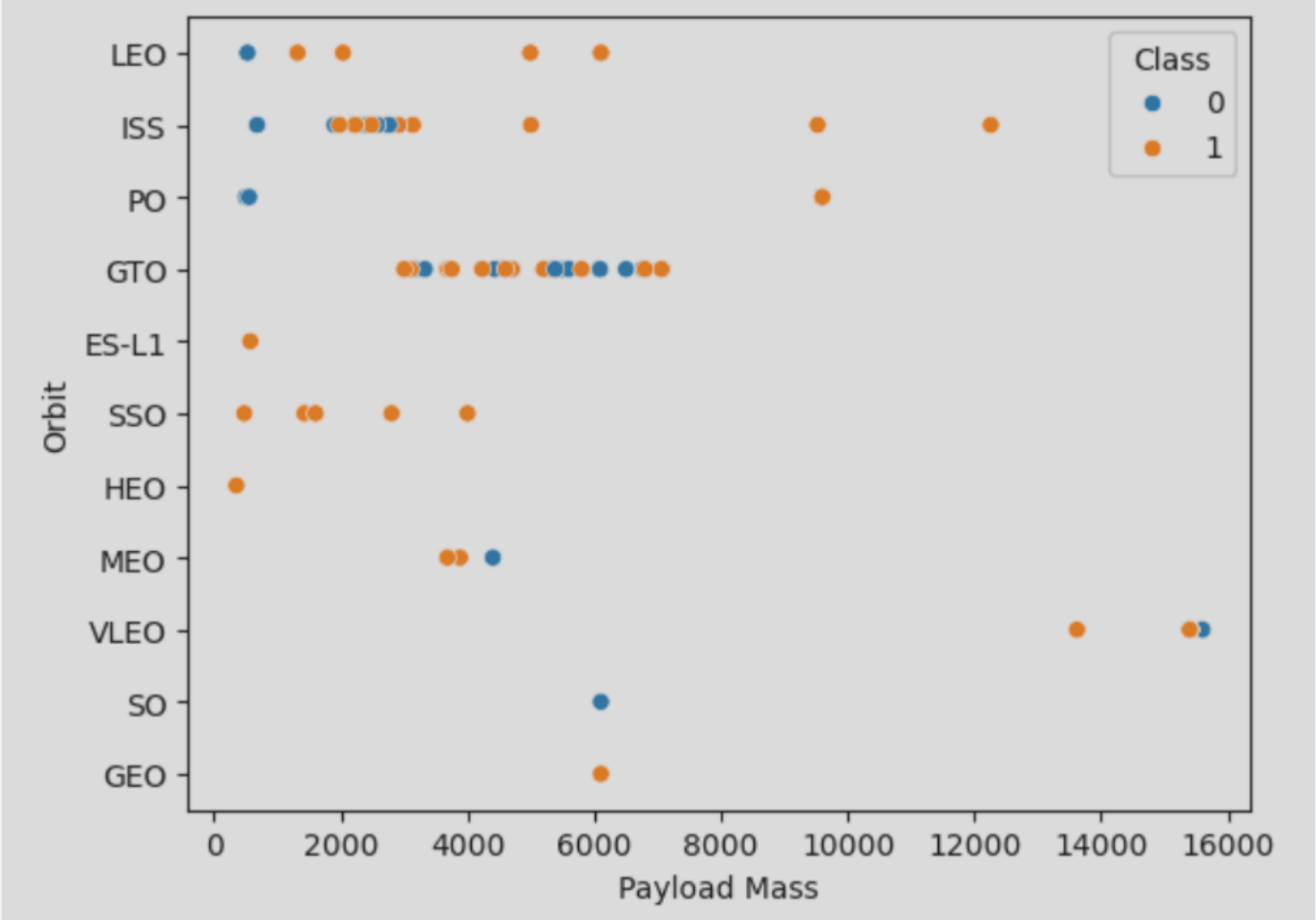
Flight Number vs. Orbit Type

- Scatter chart to visualize the relationship between Flight number and Orbit type



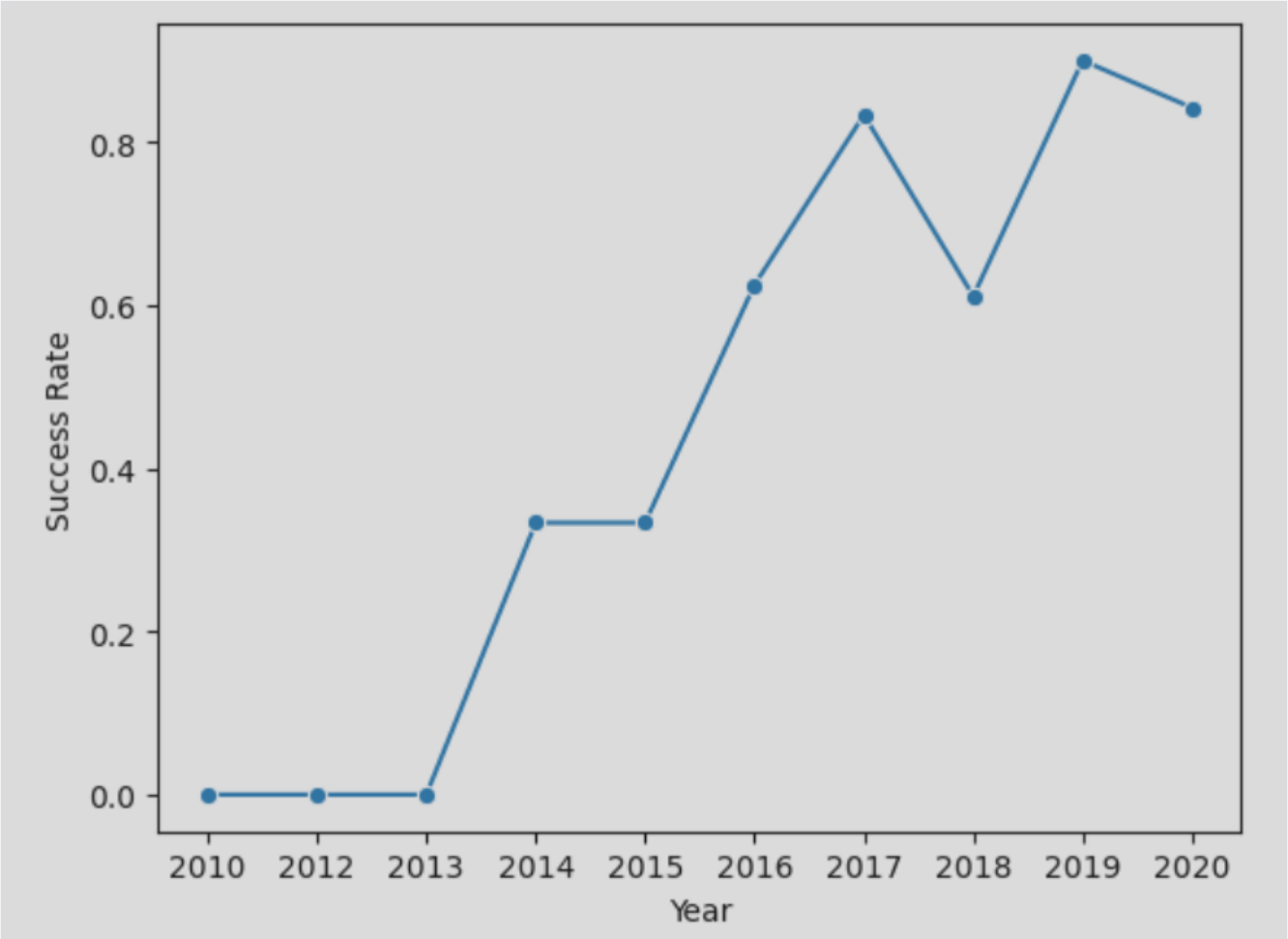
Payload vs. Orbit Type

- Scatter point chart indicating the relationship between Payload Mass and Orbit type



Launch Success Yearly Trend

- Line chart of yearly average success rate trend



All Launch Site Names

- Find the names of the unique launch sites using DISTINCT keyword

```
In [18]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[18]:

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

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`

```
In [19]: %sql Select * from SPACEXTABLE Where "Launch_site" like "CCA%" Limit 5
```

Out[19]:

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

- The total payload carried by boosters from NASA using SUM() function

```
In [28]: %sql Select Sum(PAYLOAD_MASS__KG_) from SPACEXTABLE Where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[28]: Sum(PAYLOAD_MASS__KG_)
---------------------------
                       45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1 using AVG() function

```
In [29]: %sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE Where "Booster_Version" = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

Out[29]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad by MIN() function

```
In [31]: %sql Select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
```

Out[31]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [32]:
```
%%sql Select Booster_Version from SPACEXTABLE Where Landing_Outcome = 'Success (drone ship)'
    and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

Out[32]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcome

```
In [34]: %sql Select count(Mission_Outcome) from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[34]:

count(Mission_Outcome)
101

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
In [38]: %%sql Select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ =
                (Select Max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

Out[38]:

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

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

In [39]:
```
%%sql Select substr(Date, 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site, substr(Date, 0, 5) as Year
        from SPACEXTABLE Where Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0, 5) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[39]:

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

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 **query:**

Select Landing_Outcome, **count**(*) **as** Count_Of_Landing_Outcome **from** SPACEXTABLE **Where** Date **between** '2010-06-04' **and** '2017-03-20' **Group by** Landing_Outcome **Order by** Count_Of_Landing_Outcome **Desc**

• Output **result:**

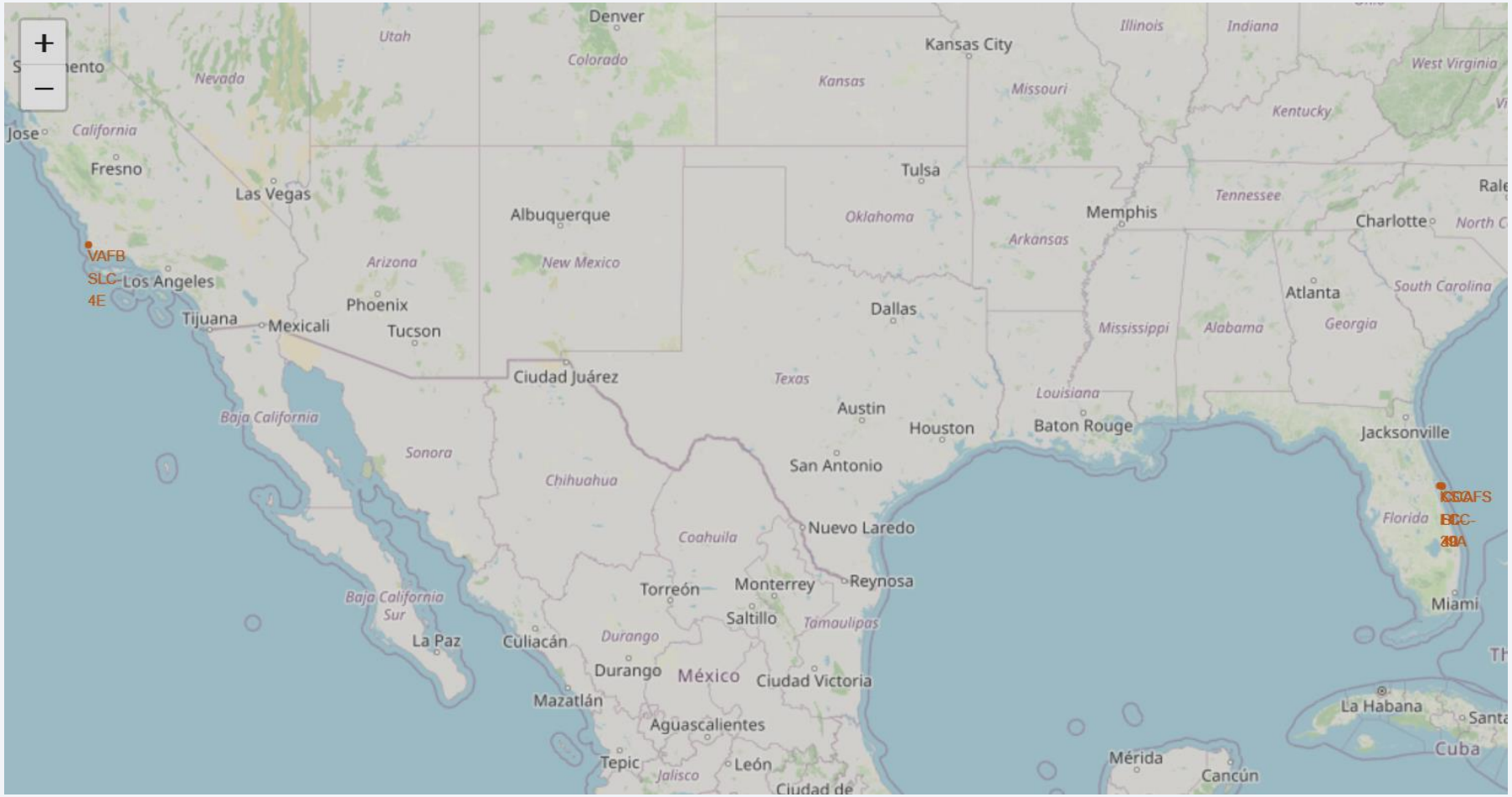
Landing_Outcome	Count_Of_Landing_Outcome
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

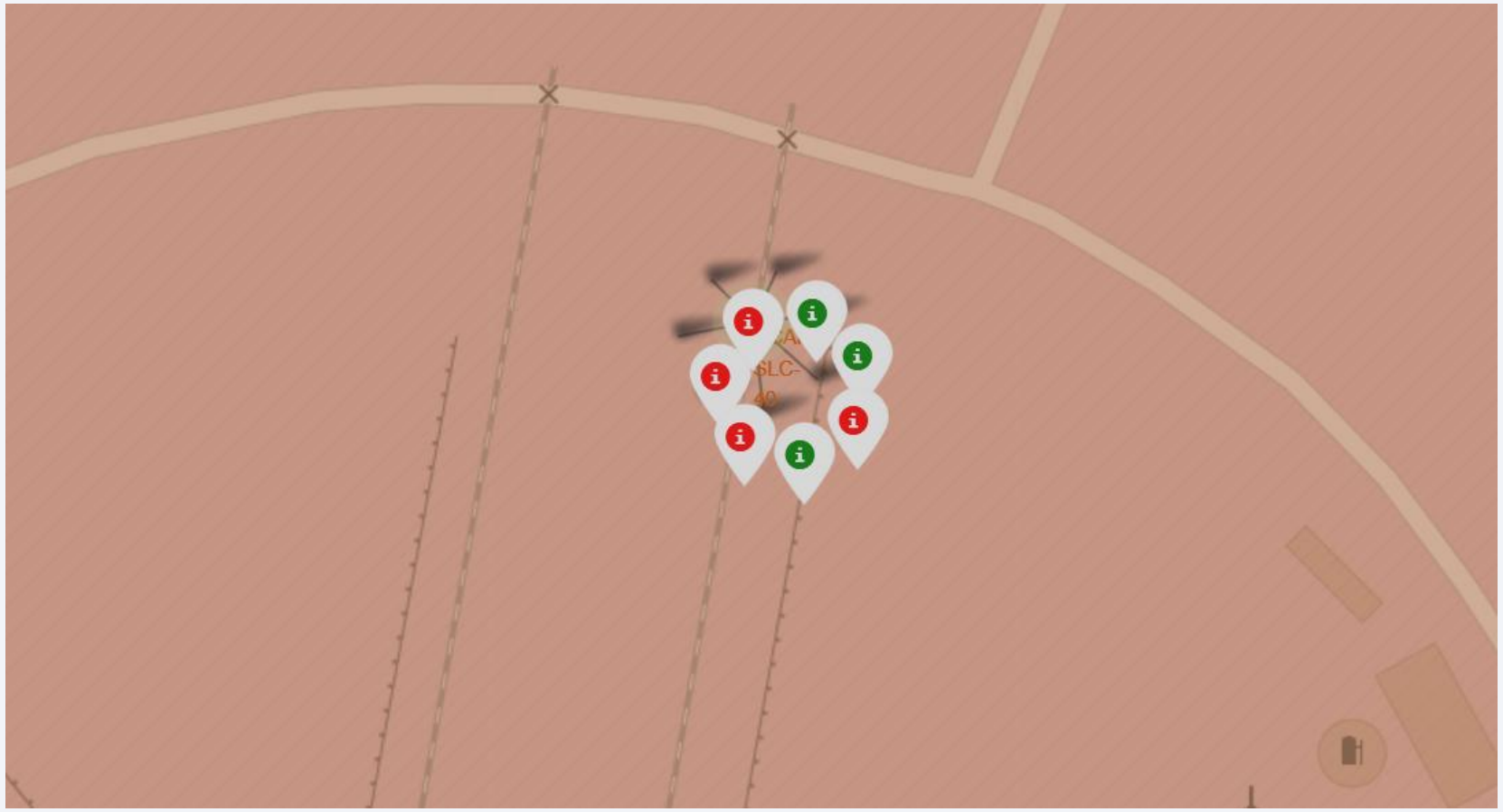
All Launch Sites Folium Map

- The screenshot shows totally four launch sites are marked on a Folium map with three sites located close to each other in Florida



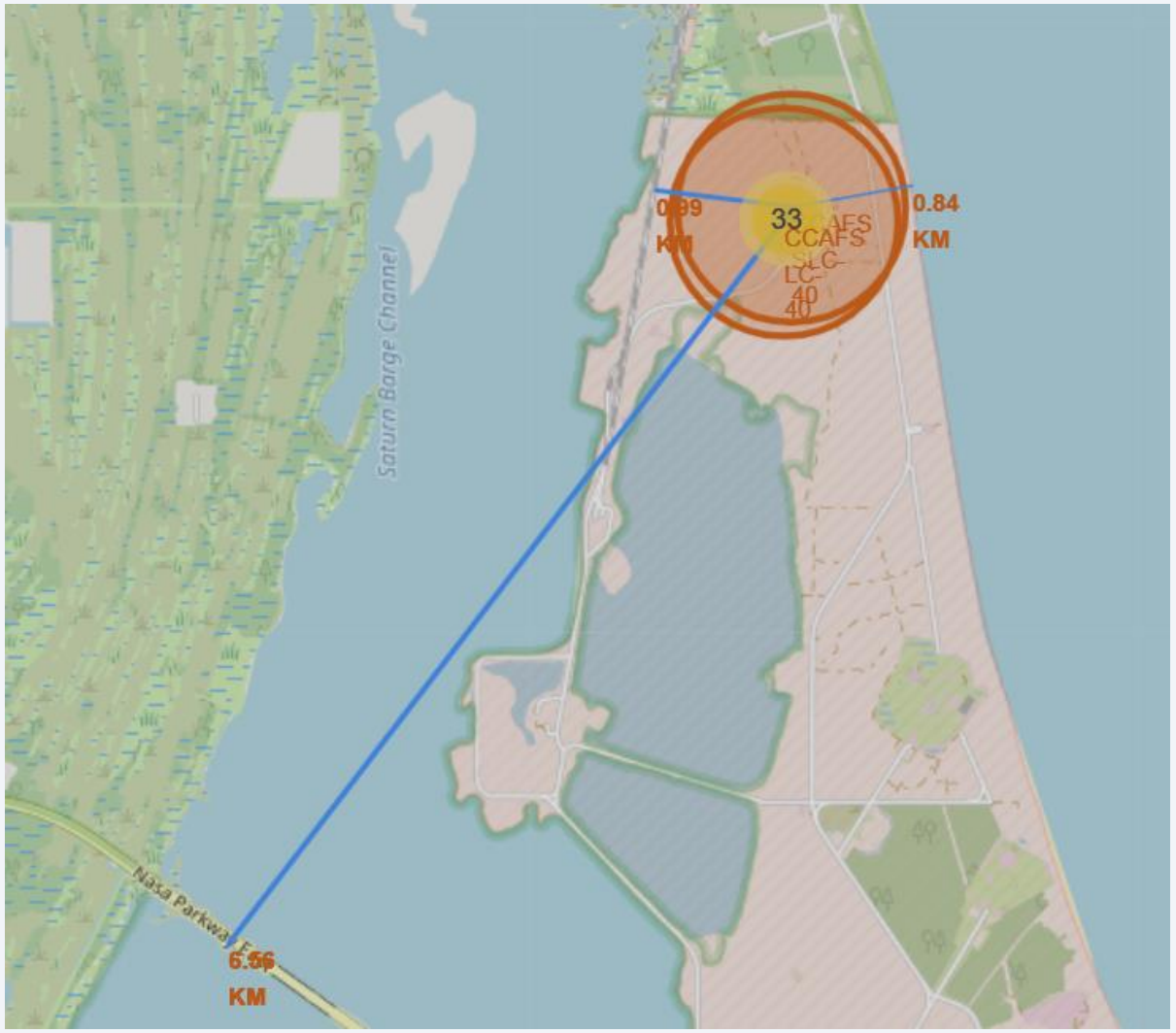
The Successful/Failed Launches for CCAFS SLC-40 on the Map

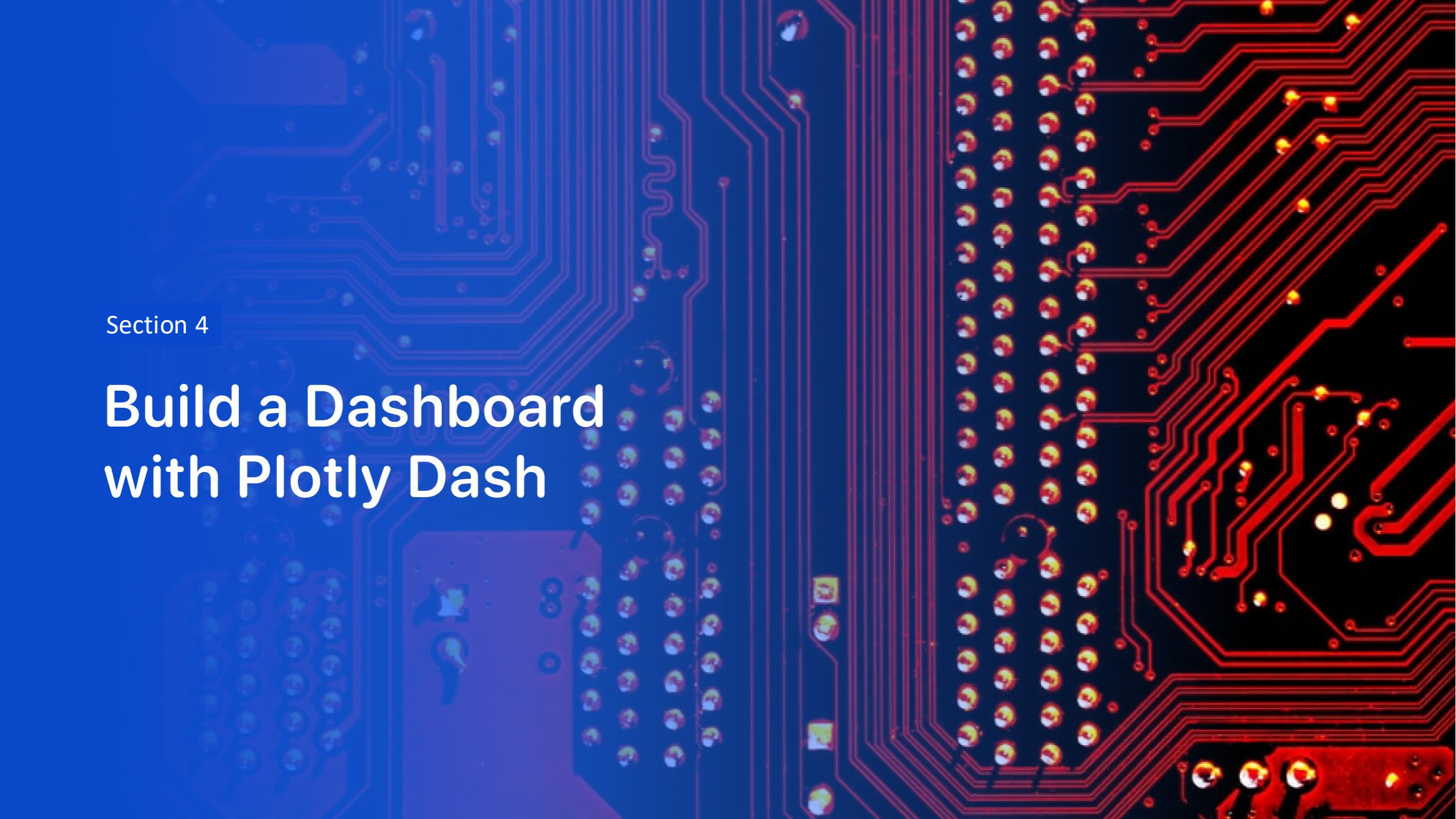
- The following screenshot shows 3 successful launches (green) and 4 failed launches (red) for the site of CCAFS SLC-40 by marker cluster



The Launch Site CCAFS SLC-40 to its Proximities on the Map

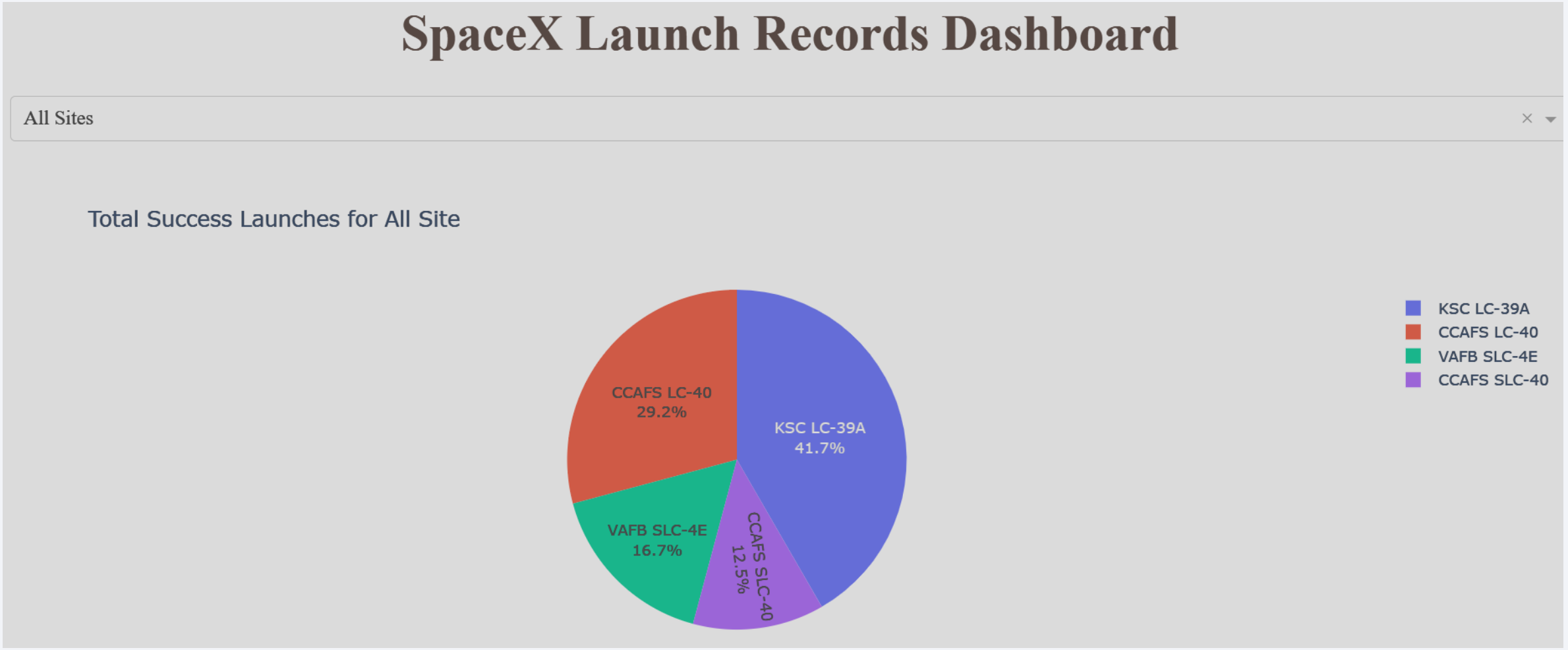
- The screenshot shows the launch site CCAFS SLC-40 to its proximities such as railway, highway, coastline, with distance calculated and displayed by Folium.Polyline and Folium.Marker




Section 4
Build a Dashboard with Plotly Dash

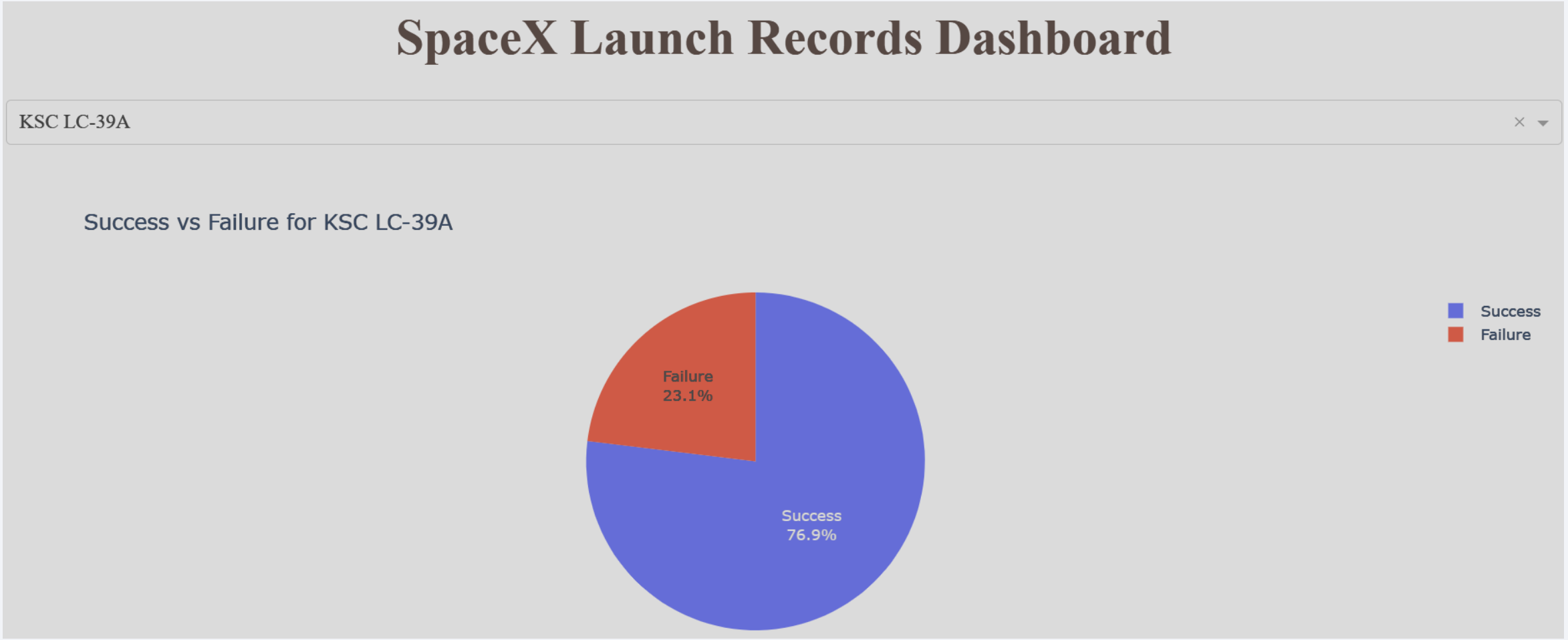
Pie Chart of Total Successful Rates for All Launch Sites

- The screenshot indicates the launch success rates for all four sites in a pie chart. The launch site can be chosen from the dropdown object.



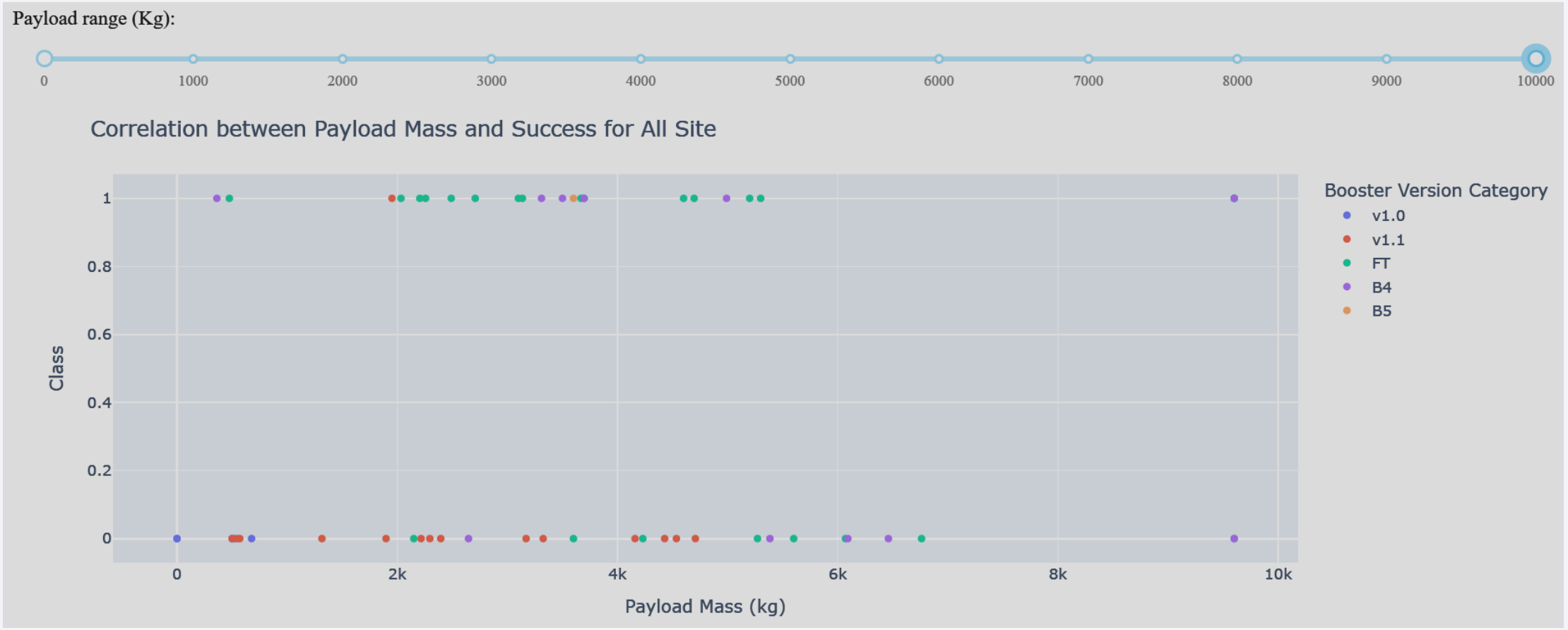
Pie Chart of the Launch Site with the Highest Launch Success Ratio

- The screenshot indicates KSC LC-39A with the highest launch success ratio. The launch site can be chosen from the dropdown object.



Scatter Plot of Payload vs. Launch Outcome for All Sites

- The screenshot indicates the correlation between the Payload and launch Outcome for all sites with color points to distinguish different booster versions. The range slider is used to choose the range of the payload mass. The booster version FT has the highest success rate for the payload range from 0kg to 100000kg.

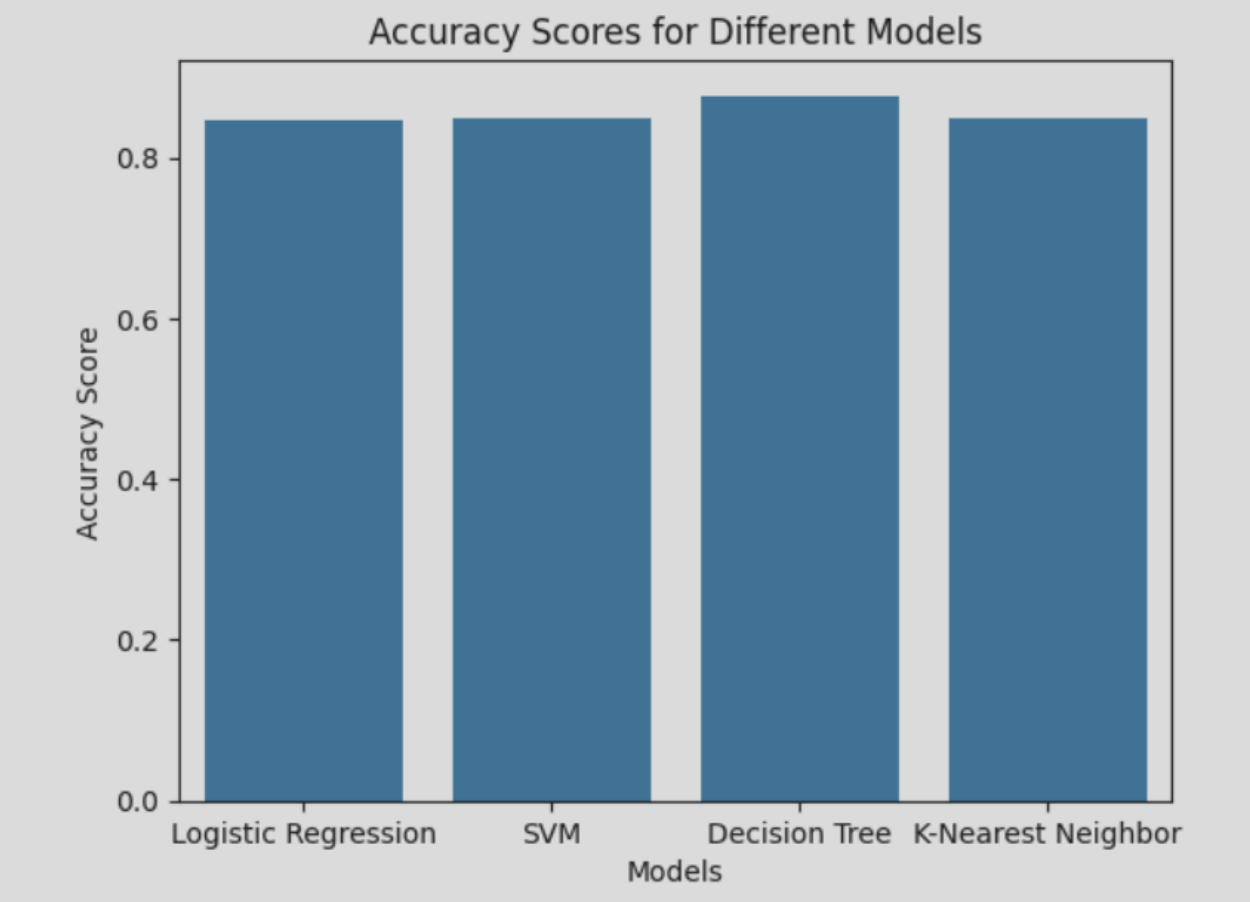


Section 5

Predictive Analysis (Classification)

Classification Accuracy

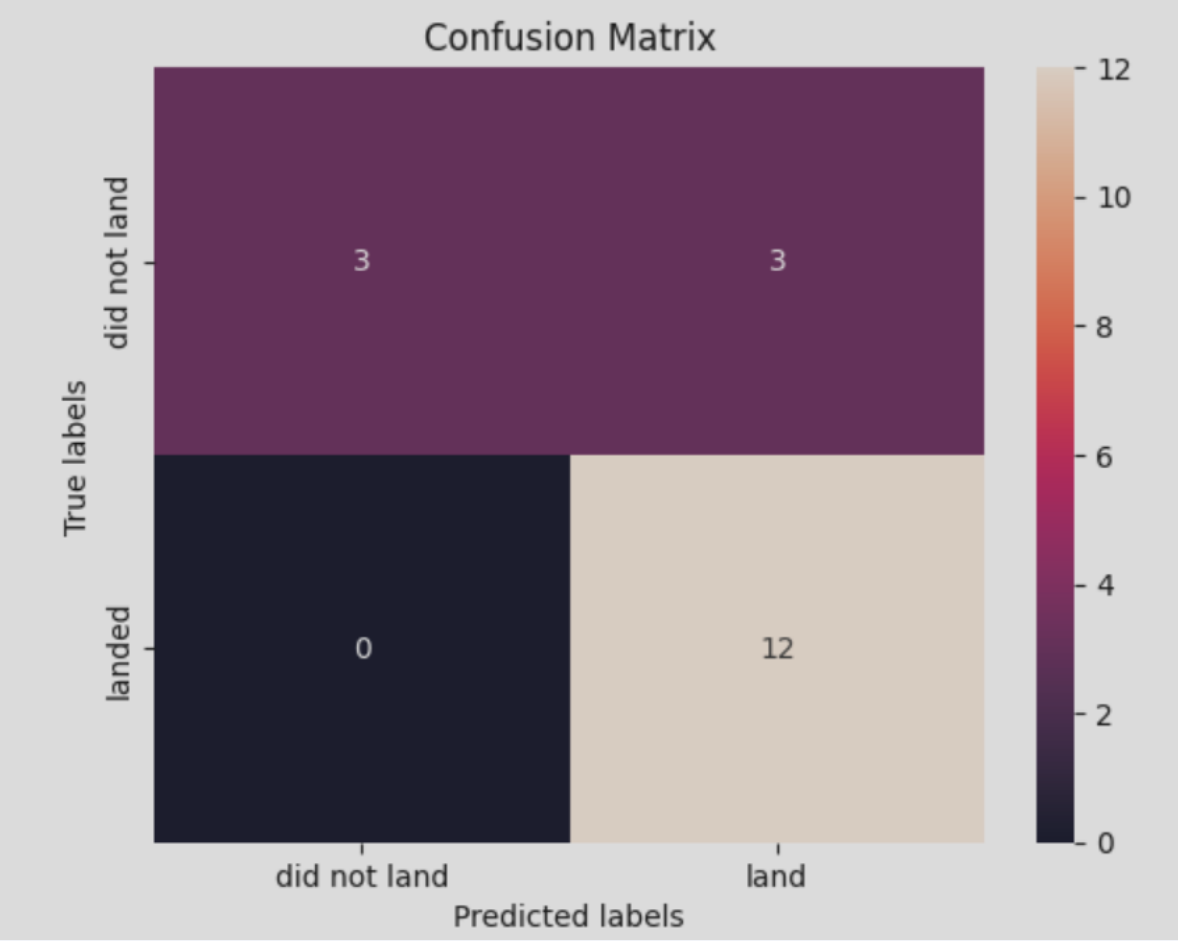
- The screenshot is a bar chart to visualize the built model accuracy for all built classification models



From the bar chart, we can tell the Decision Tree model has the highest accuracy score.

Confusion Matrix

- Confusion matrix of the best performing model: Decision Tree model



True Positive: 12
False Positive: 3
True Negative: 0
False Negative: 3

Conclusions

- The best performing model is decision tree model
- The accuracy score of the decision tree model is 0.8768
- Precision for successful landing outcome is 0.8 (12/(12+3)=0.8)
- Recall for successful landing outcome is 0.8 (12/(12+3)=0.8)

Appendix

- Dataset csv Files:
  - https://github.com/Mimicorn/Space-Repo/blob/main/data/Spacex.csv
  - https://github.com/Mimicorn/Space-Repo/blob/main/data/dataset_part_1.csv
  - https://github.com/Mimicorn/Space-Repo/blob/main/data/dataset_part_2.csv
  - https://github.com/Mimicorn/Space-Repo/blob/main/data/dataset_part_3.csv
  - https://github.com/Mimicorn/Space-Repo/blob/main/data/spacex_launch_dash.csv
  - https://github.com/Mimicorn/Space-Repo/blob/main/data/spacex_web_scraped.csv
- Jupyter Notebooks:
  - https://github.com/Mimicorn/Space-Repo/tree/main/notebooks
- Python Files:
  - https://github.com/Mimicorn/Space-Repo/blob/main/python/spacex_dash_app.py

Thank you!