IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prakash R
20-Nov-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Business Understanding and Analysis Approach
  - Data Requirements and Data Collection
  - Data Understanding and Data Preparation
  - Modeling and Evaluation
- Summary of all results
  - EDA
  - Visualization
  - Model Evaluation Score

Introduction

- Project background and context
  - The Falcon 9 first stage will land successfully. SpaceX 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 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Which launch sites have high success rate?
  - Which payload has highest success rate?
  - Analyze the various features to reduce cost and increase the success rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using Python Requests and BeautifulSoup libraries capture the data from API end points and websites using Web Scraping methods.
- Perform data wrangling
  - Using Numpy and Pandas packages handle the missing value, null value and standardize the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Using Pandas and Matplotlib libraries performs EDA and Feature Engineering.

Methodology

Executive Summary

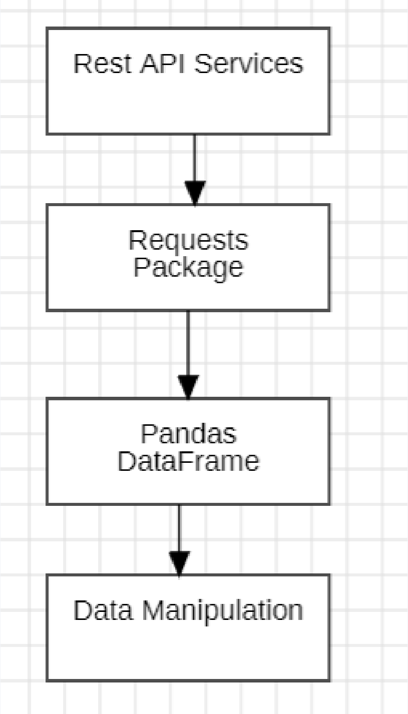
- Perform predictive analysis using classification models
  - Using scikit-learn library, it provides various machine learning algorithms perform the predictive analysis.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Plots various charts and graphs to explore the data as visual.
- How to build, tune, evaluate classification models
  - Using machine learning evaluation methods in the scikit-learn package, we have evaluate the different models.

Data Collection

- Describe how data sets were collected.
    - Using python Requests package, call the API endpoints, convert the json response data to data frame object.
    - Using BeautifulSoup library get the data from the websites.
- You need to present your data collection process use key phrases and flowcharts

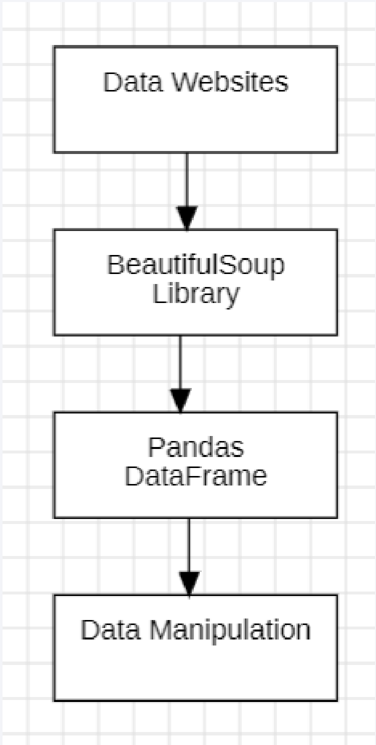
Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
  - Call the Rest API using Requests library
  - Converts the json response data to DataFrame object using Pandas.
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/87b0cba1722cf54c8487d27a6b2c7ad25a27841f/Capstone/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
  - Using Requests object get the html raw data from website.
  - Using BeautifulSoup library, parse the html content and extract the data from table cells.
  - Create a DataFrame object to store all the extracted data in the tabular format.
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/87b0cba1722cf54c8487d27a6b2c7ad25a27841f/Capstone/jupyter-labs-webscraping.ipynb



Data Websites

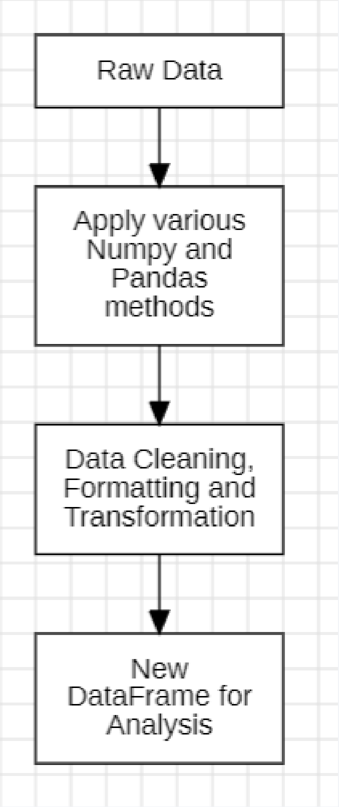
BeautifulSoup Library

Pandas DataFrame

Data Manipulation

Data Wrangling

- Describe how data were processed
  - Check the data type, missing data and null values in each row and column.
  - Transform the data into meaning format and remove unwanted data.
  - Make the data that suitable for analysis.
- You need to present your data wrangling process using key phrases and flowcharts
  - Using Numpy and Pandas packages, these libraries provide various methods to process the data cleaning and manipulation.
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/87b0cba1722cf54c8487d27a6b2c7ad25a27841f/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb


Raw Data
Apply various Numpy and Pandas methods
Data Cleaning, Formatting and Transformation
New DataFrame for Analysis

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Line Chart - changes in value across continuous measurements
  - Bar Chart - values for different categorical groupings
  - Scatter Chart - versatile demonstration of the relationship between the plotted variables—whether that correlation is strong or weak
  - Histogram Chart - bar chart are actually continuous numeric ranges.
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/4e3c9ada6ab92b9df030ef93bed4cdca1cd40774/Capstone/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Load the data from database
  - Summarize the data easily and fast using SQL queries
  - Fetch the data from more that one table and join the tables.
  - Manipulate the data using filtering, aggregate the summary data.
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/4e3c9ada6ab92b9df030ef93bed4cdca1cd40774/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

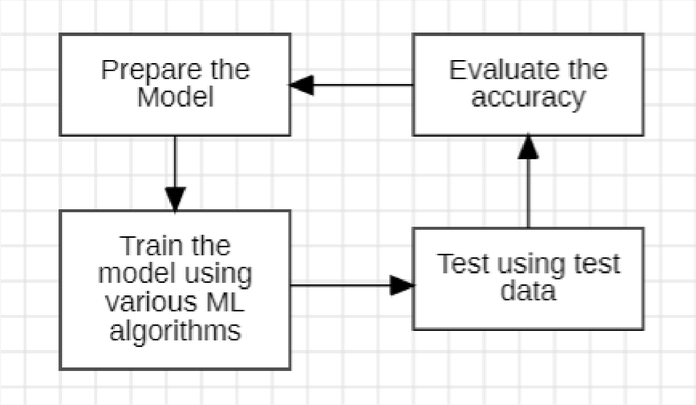
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Circle – Added to identify the location of launch pad area.
  - Markers – Points the locate the coordinate of place and tooltip for description.
  - Lines – Helps to find the distance of two places.
- Explain why you added those objects
  - Folium provides real-time maps and quickly visualize the data on the map.
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/4e3c9ada6ab92b9df030ef93bed4cdca1cd40774/Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Line charts, bar charts, scatter charts and pie charts.
- Explain why you added those plots and interactions
  - Interactive – Charts are plotted dynamically based on user inputs.
  - Customization – Charts are customized for easy understanding and interaction
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/4e3c9ada6ab92b9df030ef93bed4cdca1cd40774/Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Applied KNN, Logistic Regression, Decision Tree and SVM algorithms to find the best prediction method.
- You need present your model development process using key phrases and flowchart
  - Initialize the parameters and train the training dataset using various classification algorithms.
  - Test the prediction using test data and evaluate the model.
  - Repeat these steps until more accuracy.
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
  - https://github.com/prakashare/Ibm-DataScience/blob/4e3c9ada6ab92b9df030ef93bed4cdca1cd40774/Capstone/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Prepare the Model

Train the model using various ML algorithms

Test using test data

Evaluate the accuracy

Results

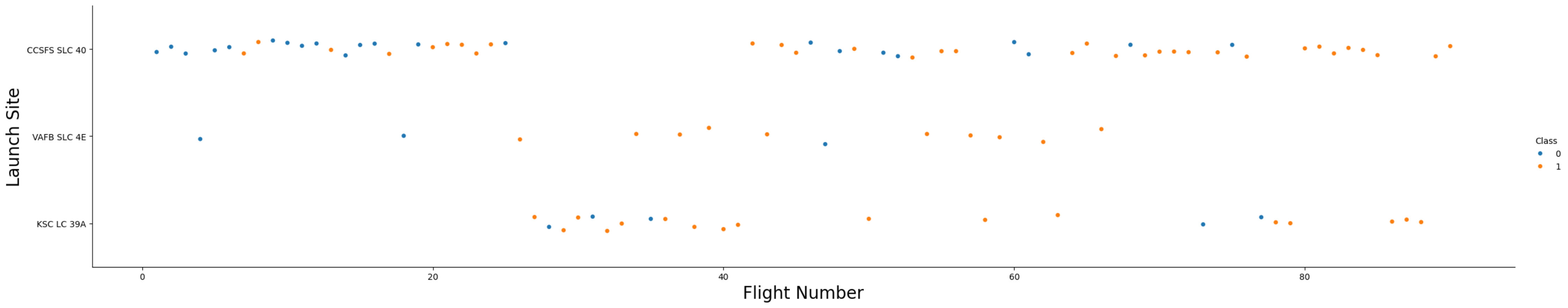
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

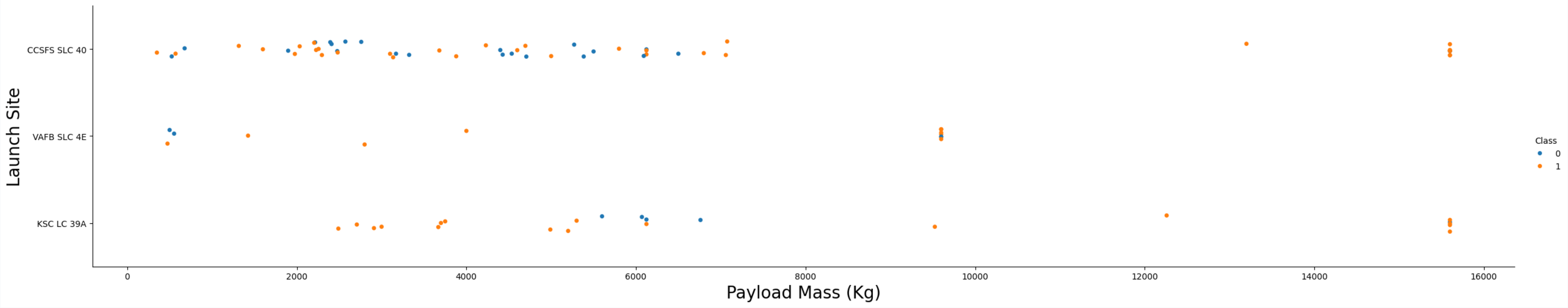
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations
  - CCSFS SLC 40 launch site handles all types of flights and KSC LC 39A site handle middle ranges of flight only.



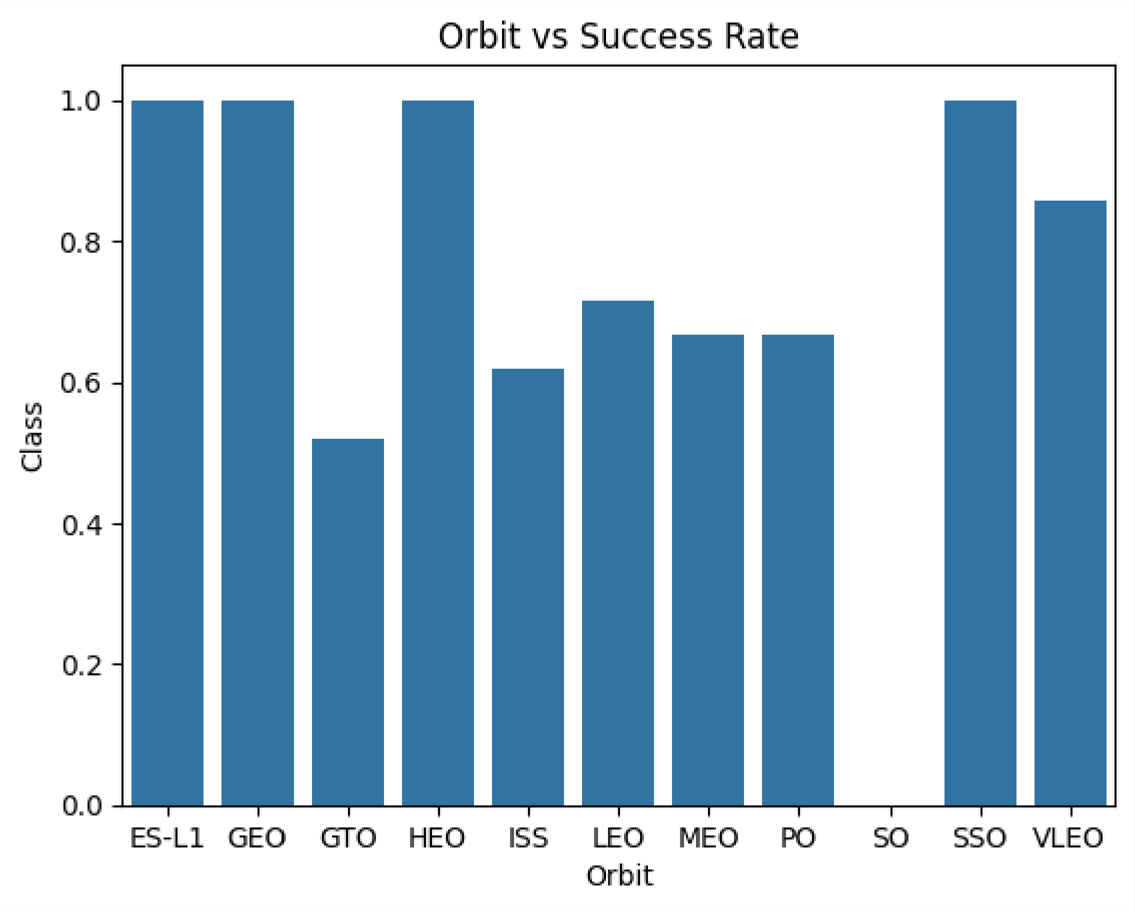
Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations
    - CCSFS SLC 40 launch site handles more low to middle payloads and VAFB handles less no. of payloads.



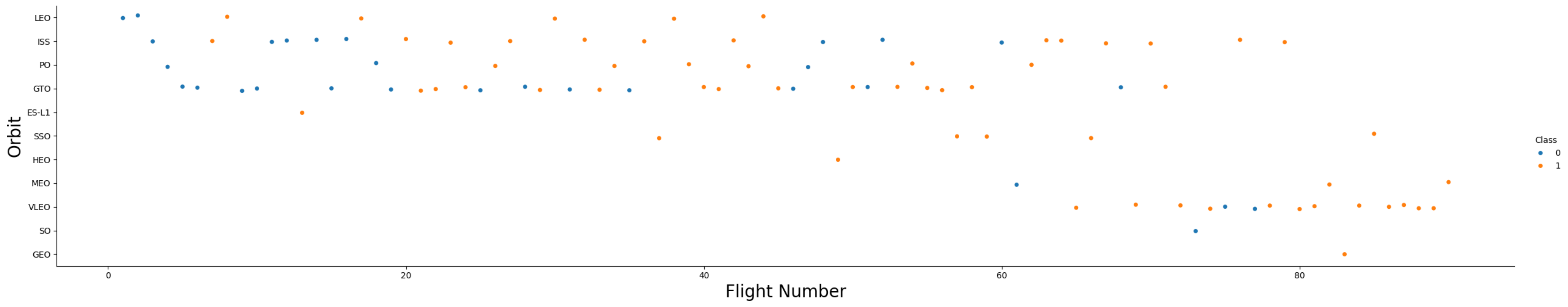
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations
    - ESL1, GEO HEO and SSO orbits have high success rates.
    - No success rate from the SO orbit.
    - All others have 40 to 75% of success rates.



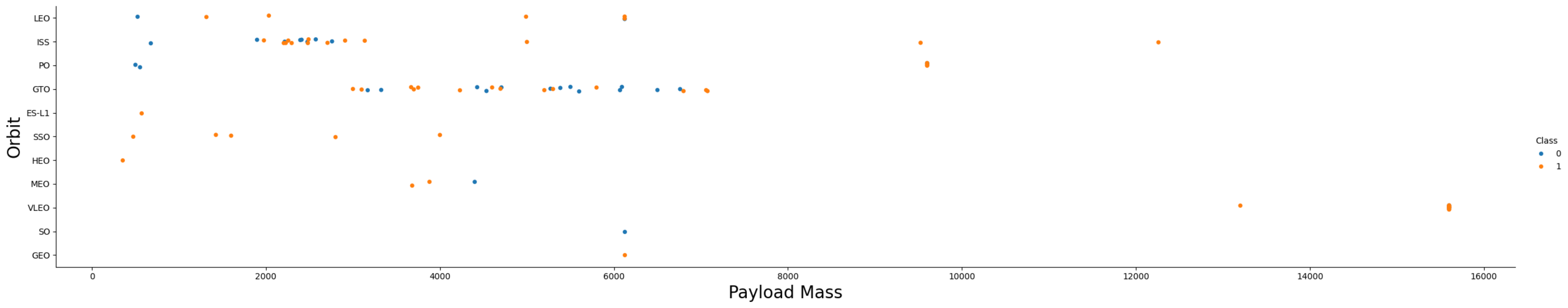
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations
  - Most of the flight placed in the LEO, ISS, PO and GTO orbits.



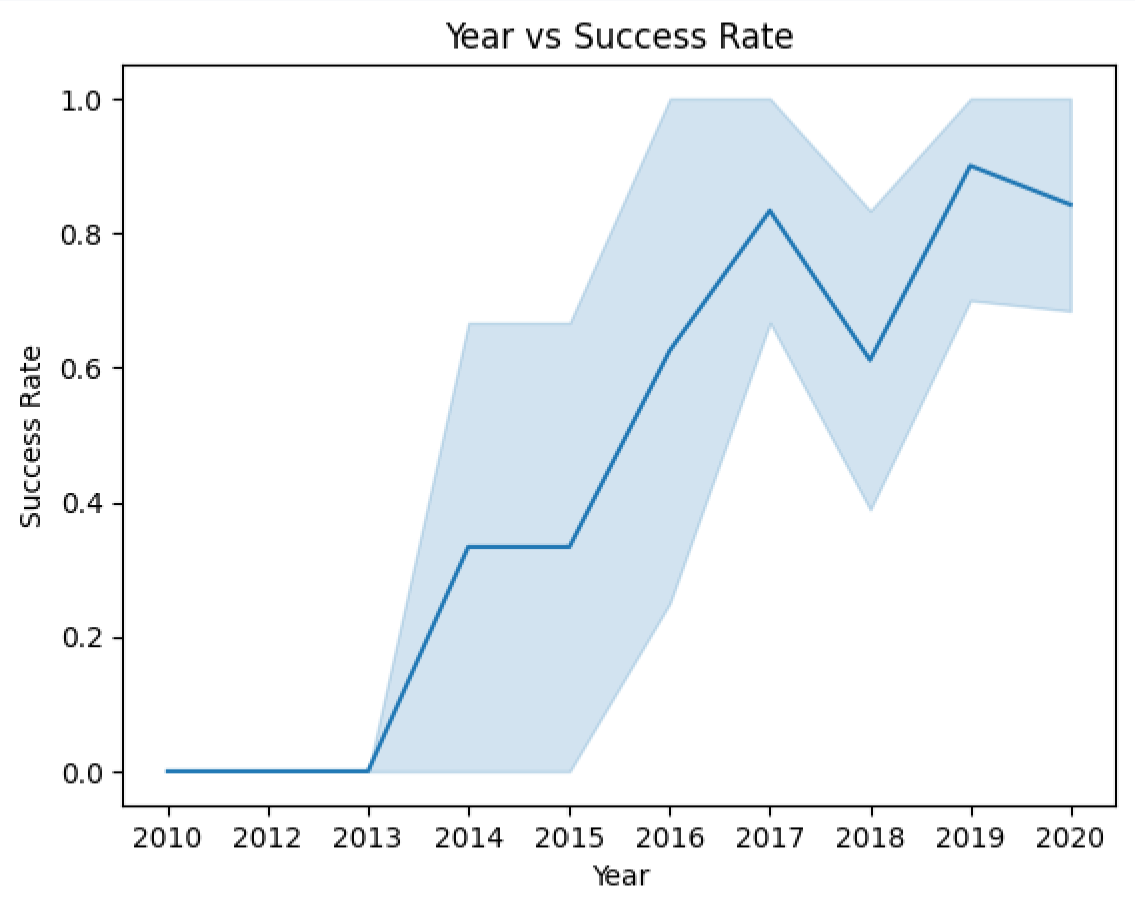
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations
  - More low to middle range of payloads are placed in the GTO orbit.



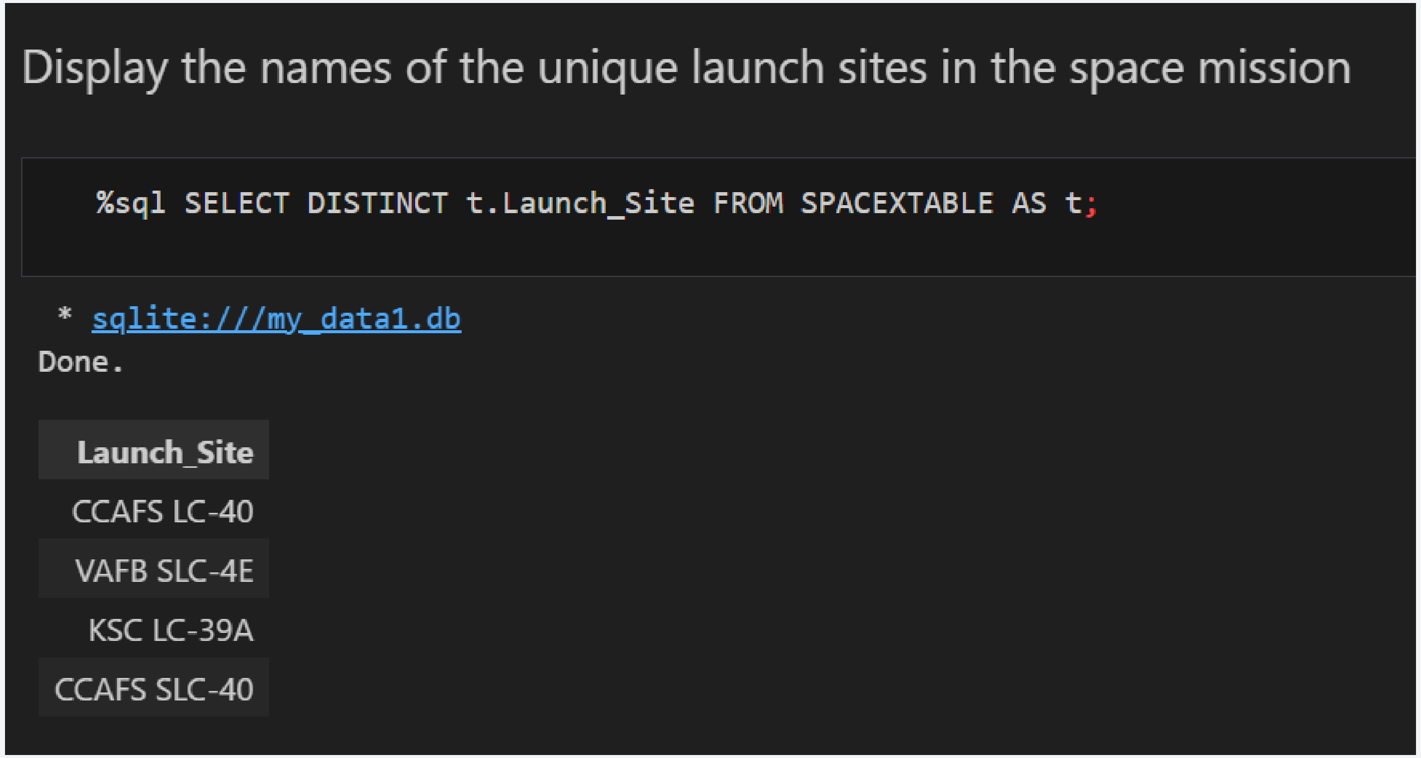
Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations
  - The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here
  - The DISTINCT statement used to fetch the unique data from the column.



Display the names of the unique launch sites in the space mission

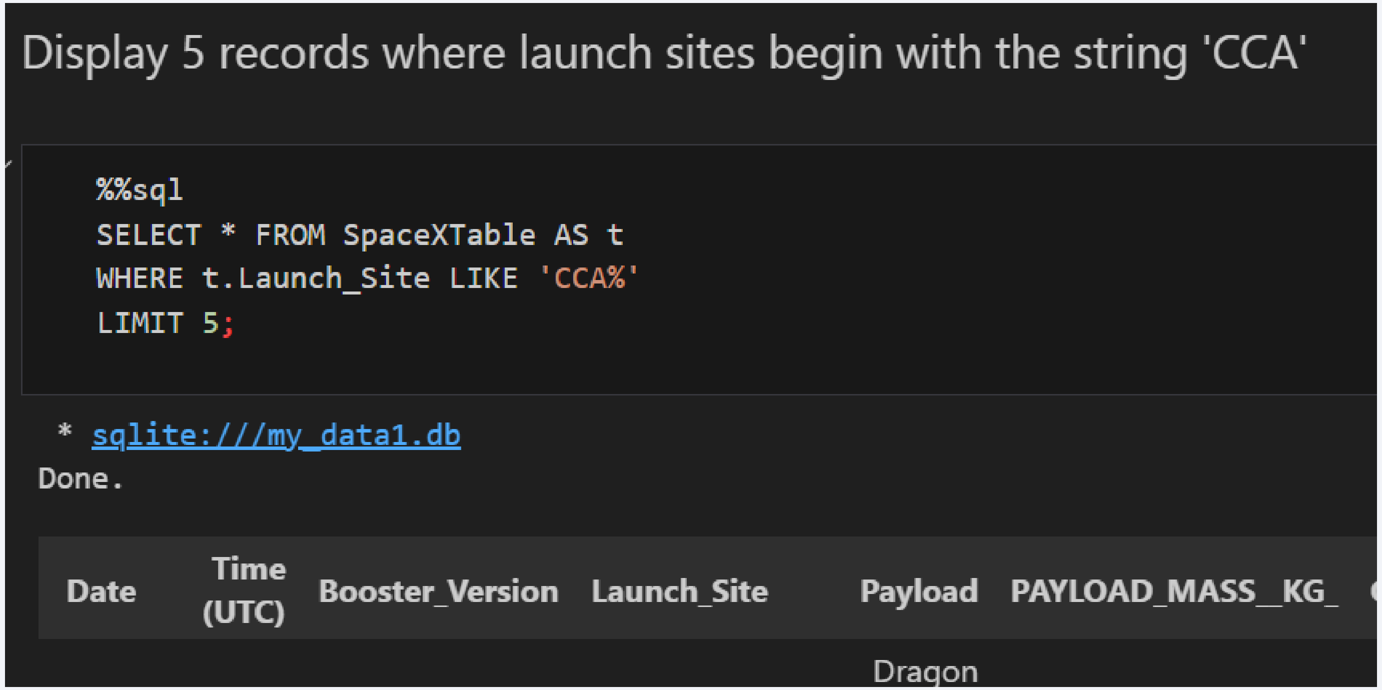
```
%sql SELECT DISTINCT t.Launch_Site FROM SPACEXTABLE AS t;
```

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here
  - The WHERE clause and LIKE operator used to filter the rows.
  - The LIMIT statement used to fetch the rows with specified number.



Display 5 records where launch sites begin with the string 'CCA'

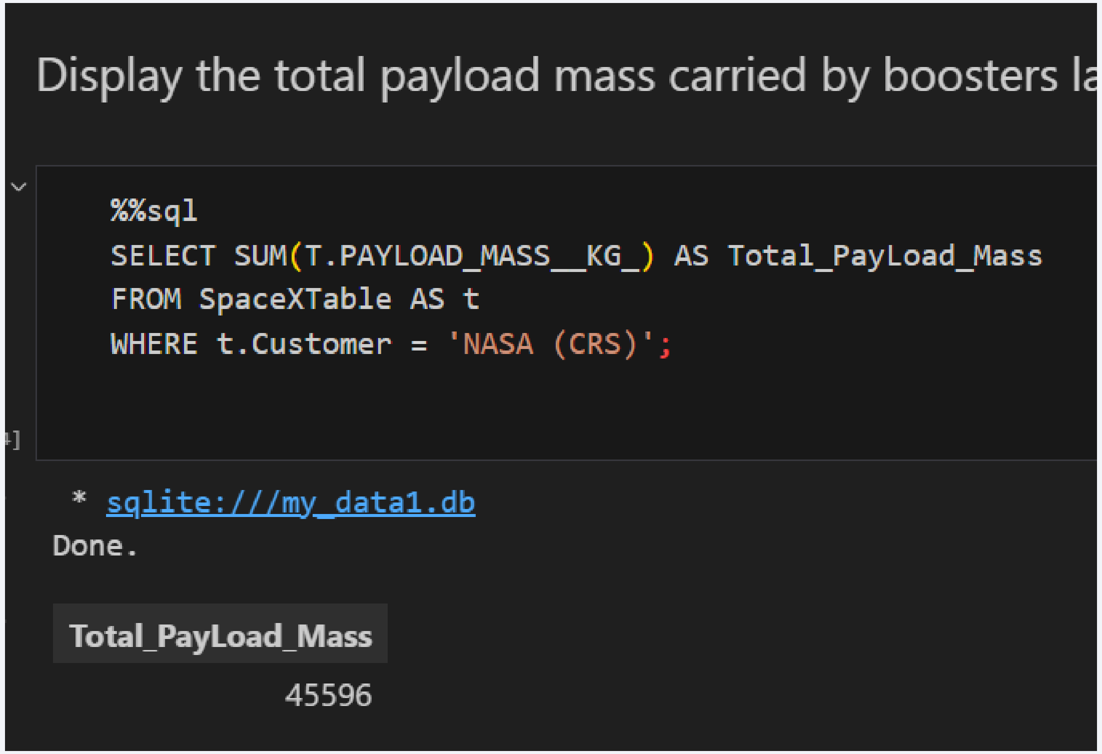
```
%%sql
SELECT * FROM SpaceXTable AS t
WHERE t.Launch_Site LIKE 'CCA%'
LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_
				Dragon	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here
  - The SUM function is used to summarized the data in the column.



Display the total payload mass carried by boosters la

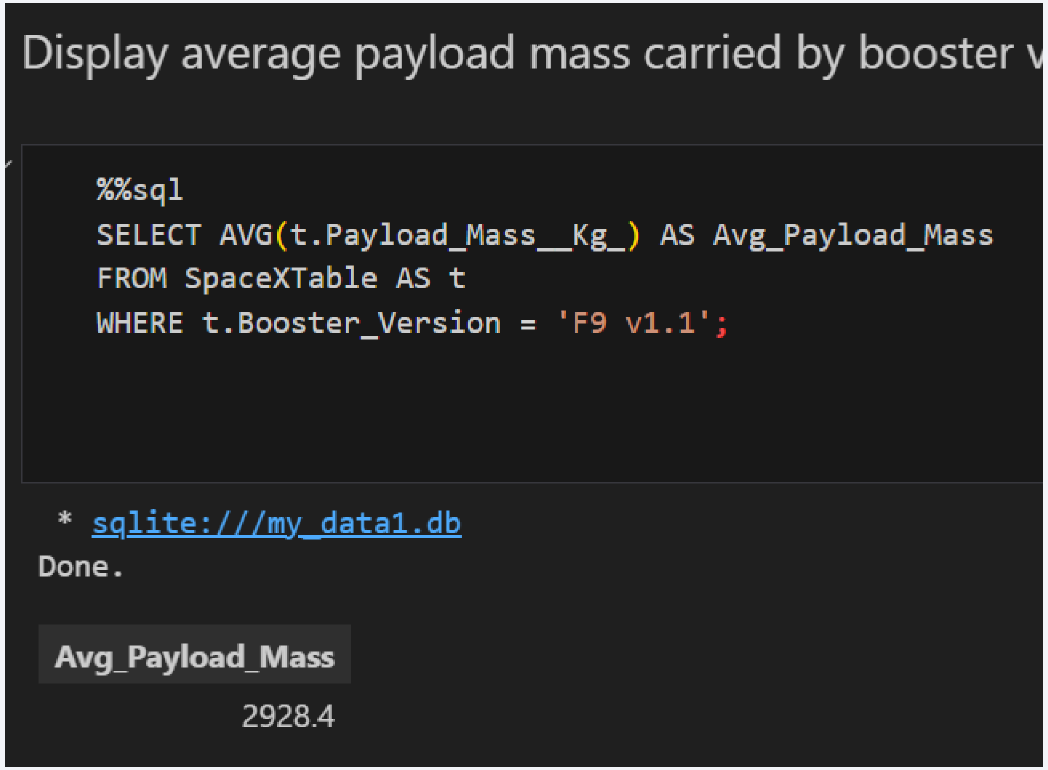
```
%%sql
SELECT SUM(T.PAYLOAD_MASS__KG_) AS Total_PayLoad_Mass
FROM SpaceXTable AS t
WHERE t.Customer = 'NASA (CRS)';
```

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

Total_PayLoad_Mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here
    - The AVG function is used to calculate the average value of the column data.



Display average payload mass carried by booster v

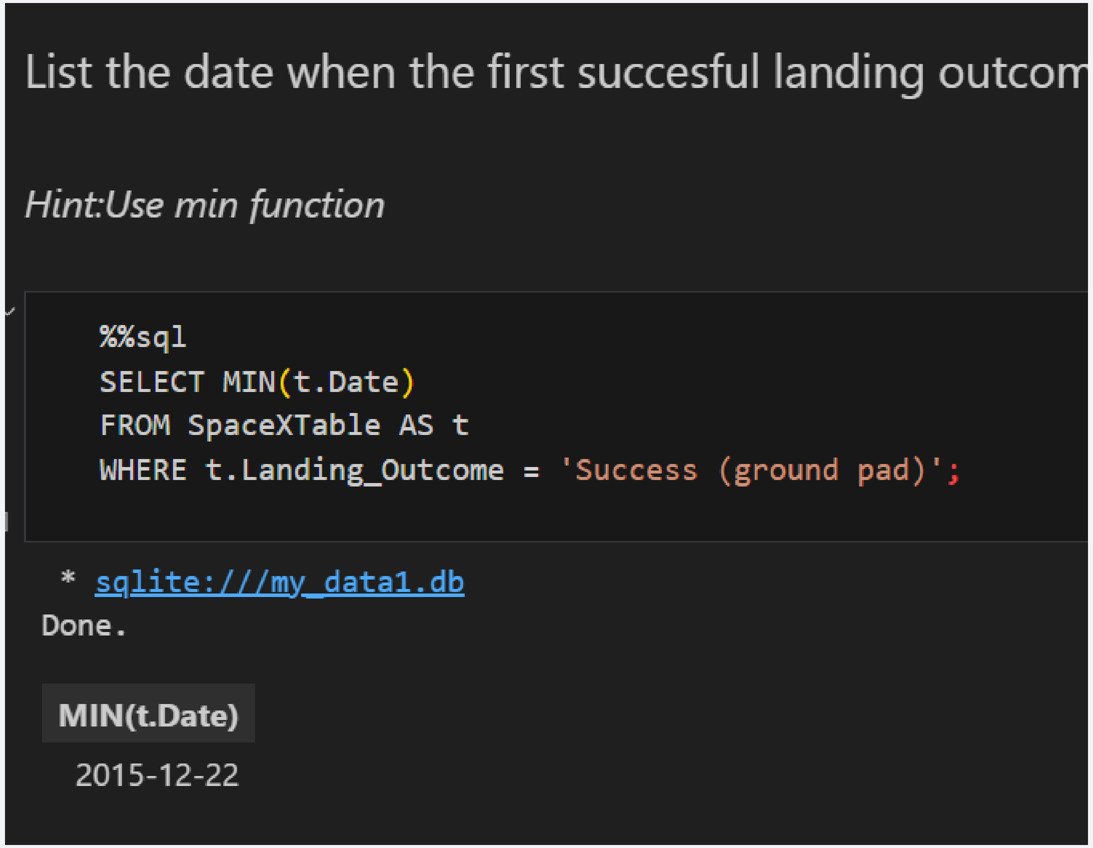
```
%%sql
SELECT AVG(t.Payload_Mass__Kg_) AS Avg_Payload_Mass
FROM SpaceXTable AS t
WHERE t.Booster_Version = 'F9 v1.1';
```

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

Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here
  - The MIN function used to calculate the first date or minimum value of the data in the column.



List the date when the first succesful landing outcom

Hint:Use min function

```
%%sql
SELECT MIN(t.Date)
FROM SpaceXTable AS t
WHERE t.Landing_Outcome = 'Success (ground pad)';
```

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

MIN(t.Date)
2015-12-22

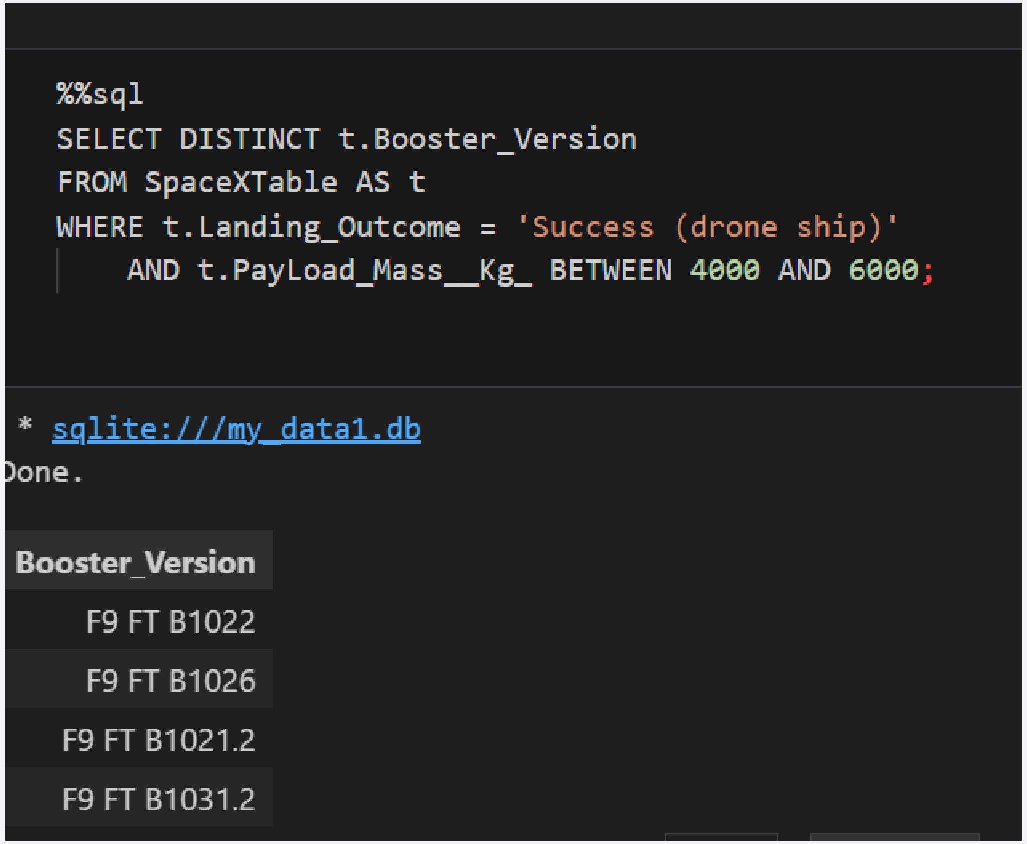
Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Present your query result with a short explanation here

  - The WHERE clause and BETWEEN & AND clause used to filter the row data.

  - The DISTINCT statement used to fetch unique values in the column data.



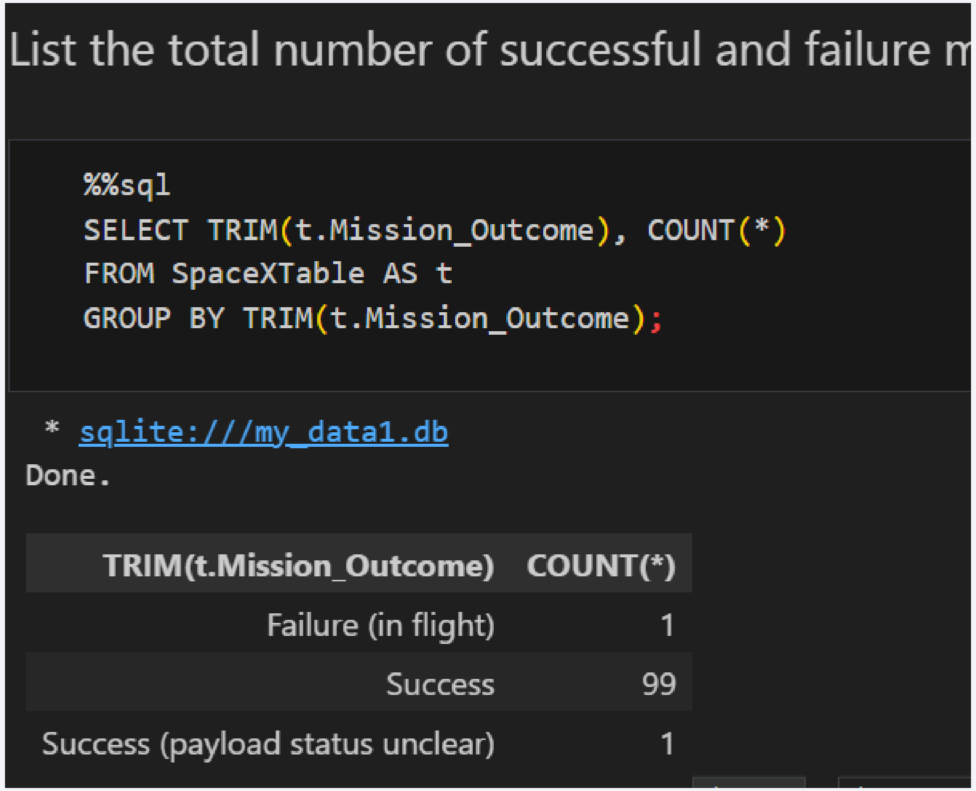
```
%%sql
SELECT DISTINCT t.Booster_Version
FROM SpaceXTable AS t
WHERE t.Landing_Outcome = 'Success (drone ship)'
    AND t.PayLoad_Mass__Kg_ BETWEEN 4000 AND 6000;
```

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

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 outcomes
- Present your query result with a short explanation here
  - The GROUP BY and COUNT aggregate function is used to summarize the data by category wise.



List the total number of successful and failure m

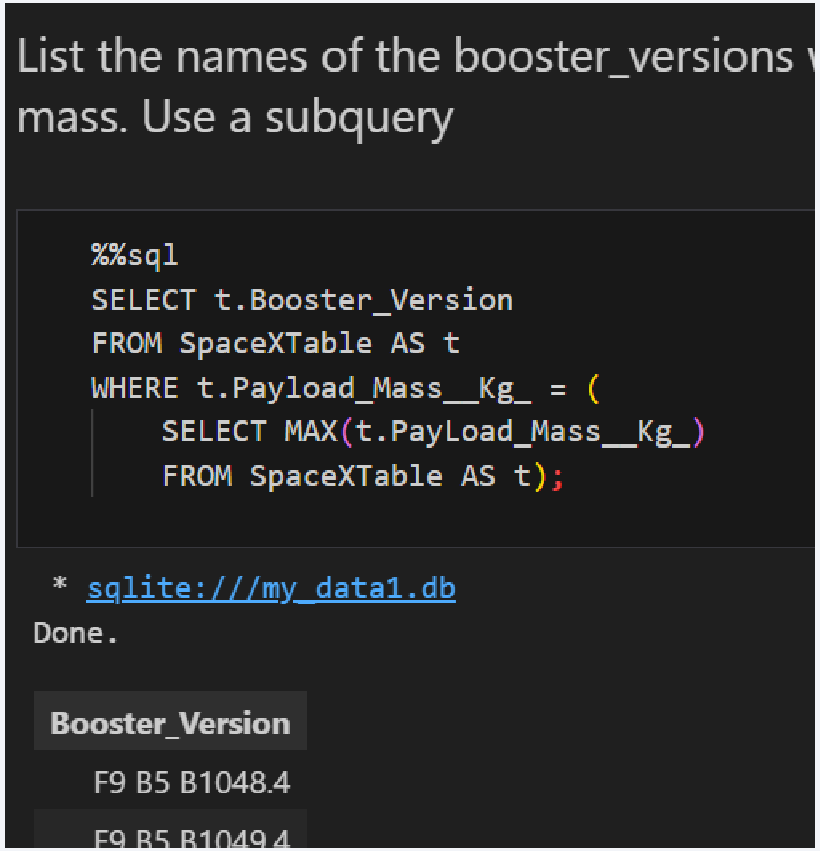
```
%%sql
SELECT TRIM(t.Mission_Outcome), COUNT(*)
FROM SpaceXTable AS t
GROUP BY TRIM(t.Mission_Outcome);
```

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

TRIM(t.Mission_Outcome)	COUNT(*)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here
  - The subquery used by the main query to fetch the data based on subquery returns data.



List the names of the booster_versions

mass. Use a subquery

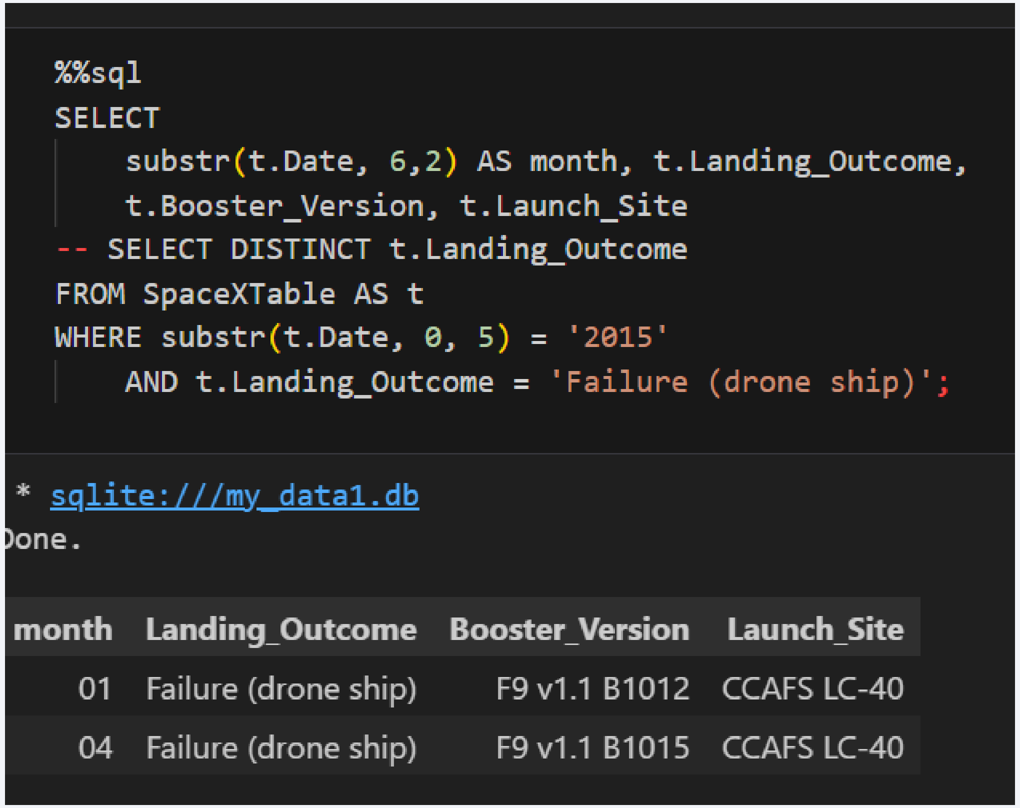
```
%%sql
SELECT t.Booster_Version
FROM SpaceXTable AS t
WHERE t.Payload_Mass__Kg_ = (
    SELECT MAX(t.PayLoad_Mass__Kg_)
    FROM SpaceXTable AS t);
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here
  - The substr function is used to get the year portion from the date data.



```
%%sql
SELECT
    substr(t.Date, 6,2) AS month, t.Landing_Outcome,
    t.Booster_Version, t.Launch_Site
-- SELECT DISTINCT t.Landing_Outcome
FROM SpaceXTable AS t
WHERE substr(t.Date, 0, 5) = '2015'
    AND t.Landing_Outcome = 'Failure (drone ship)';
```

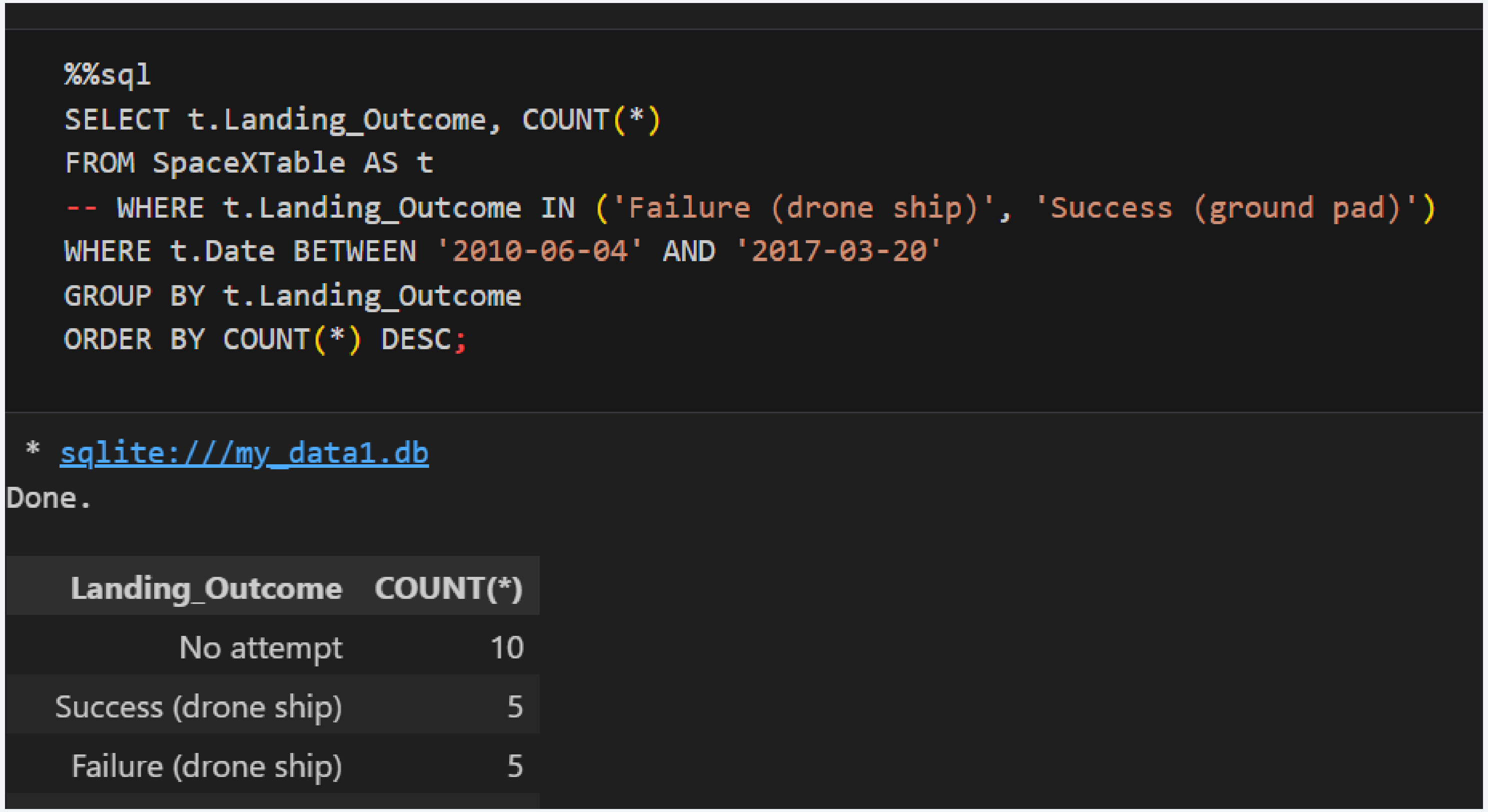
```
 * sqlite:///my_data1.db
Done.
```

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

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

- Present your query result with a short explanation here
  - The WHERE and BETWEEN used to filter the row data.
  - The GROUP BY clause used to group the row data and apply aggregate function on column.



```
%%sql
SELECT t.Landing_Outcome, COUNT(*)
FROM SpaceXTable AS t
-- WHERE t.Landing_Outcome IN ('Failure (drone ship)', 'Success (ground pad)')
WHERE t.Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY t.Landing_Outcome
ORDER BY COUNT(*) DESC;
```

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

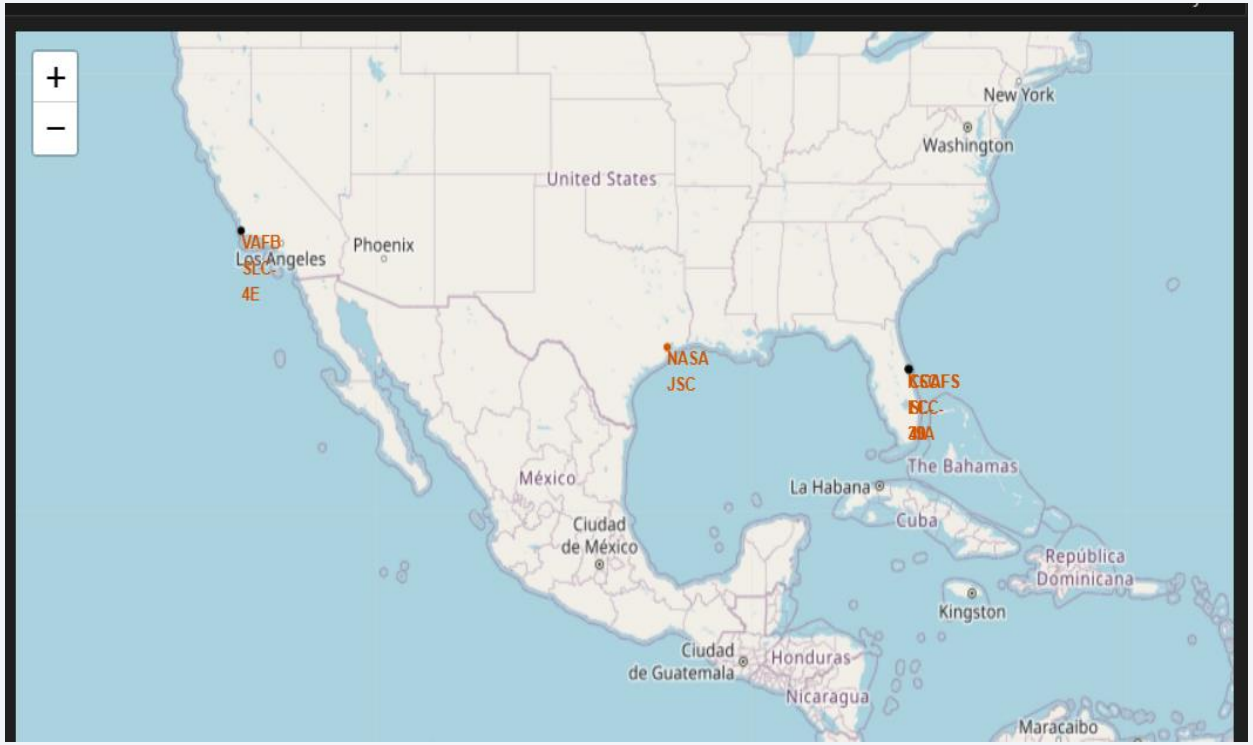
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5

Section 3

Launch Sites Proximities Analysis

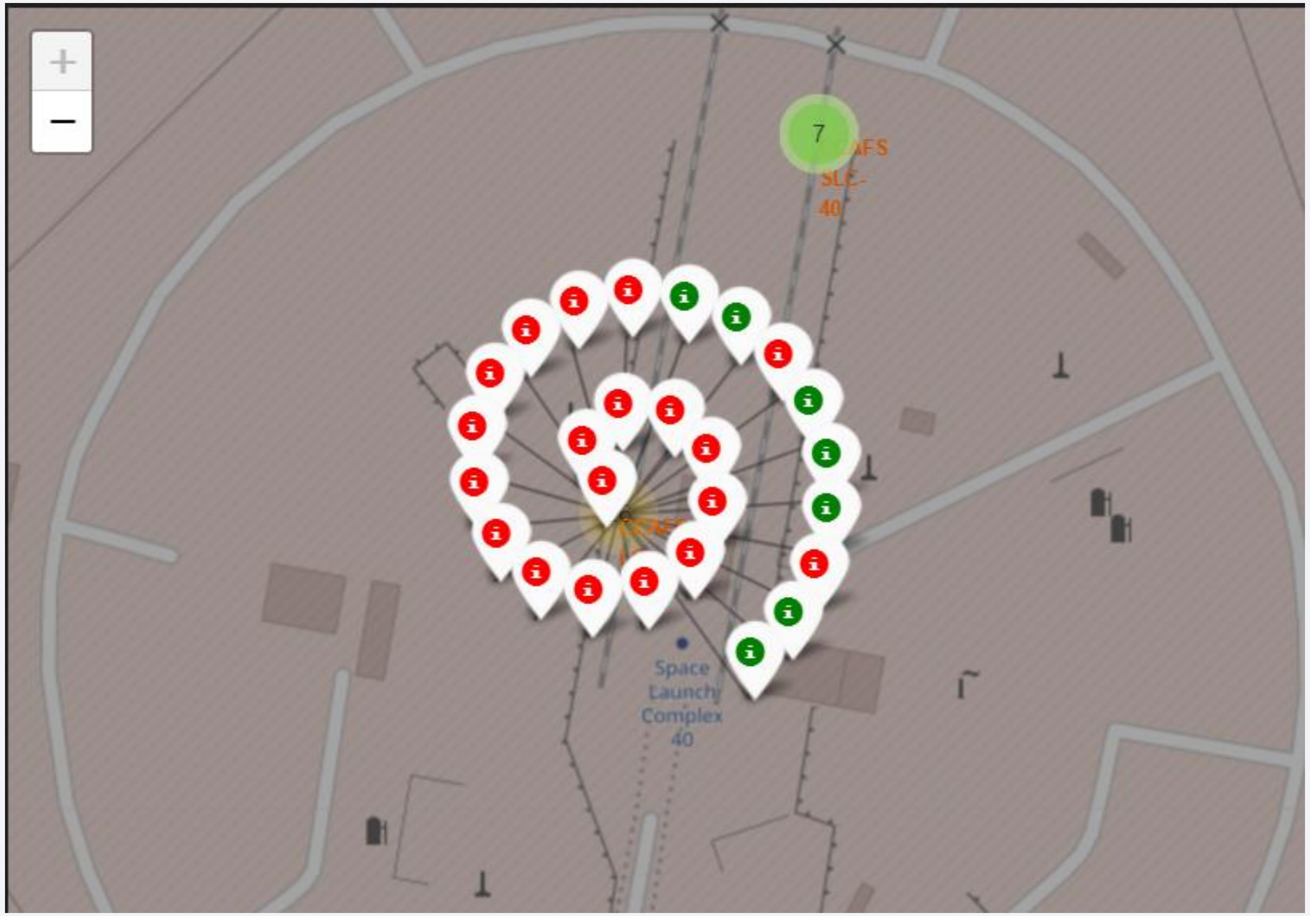
Launch Sites Location

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot
  - Easily locate and identify the launch sites using Map chart.
  - All sites locations are near to coastal.



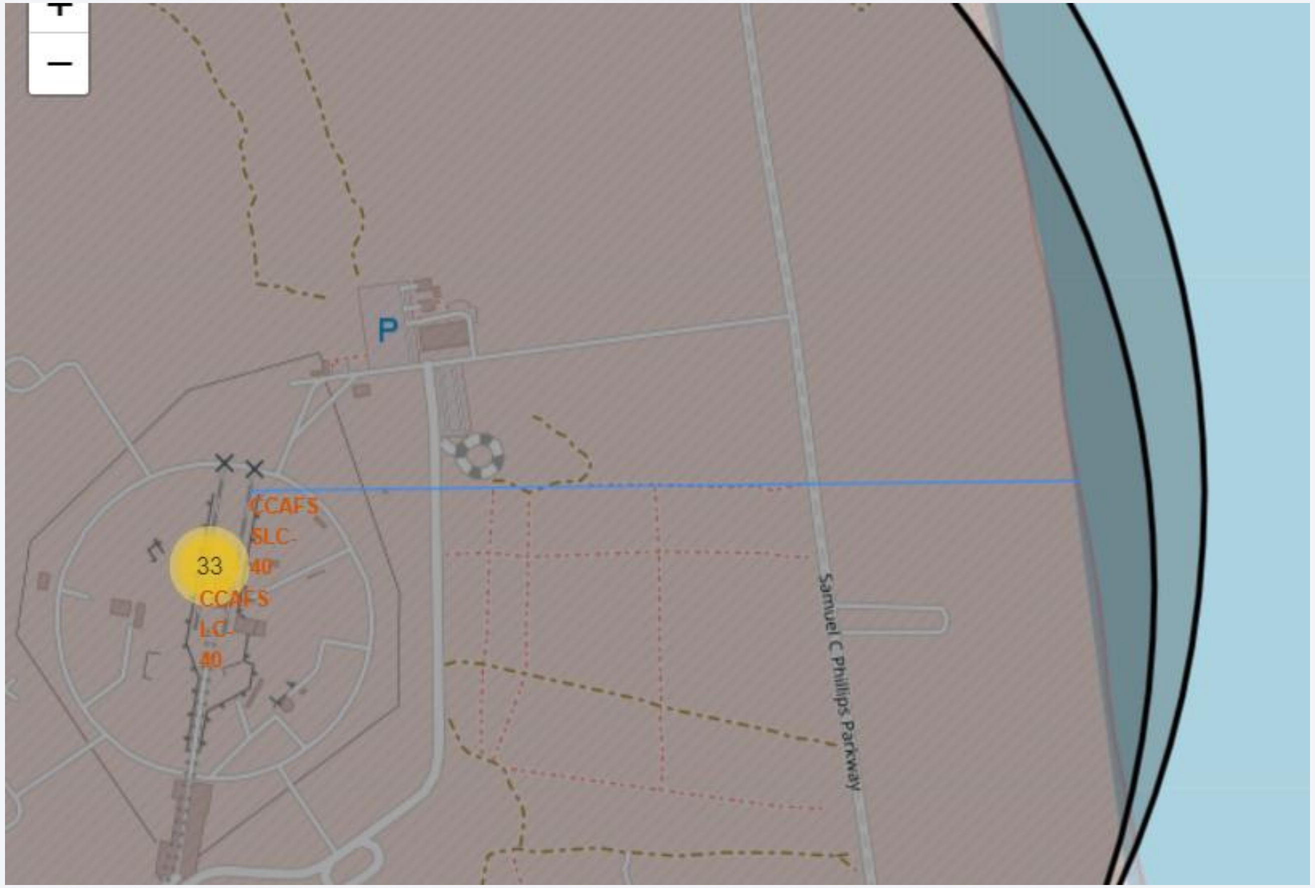
Launch Outcomes Sites

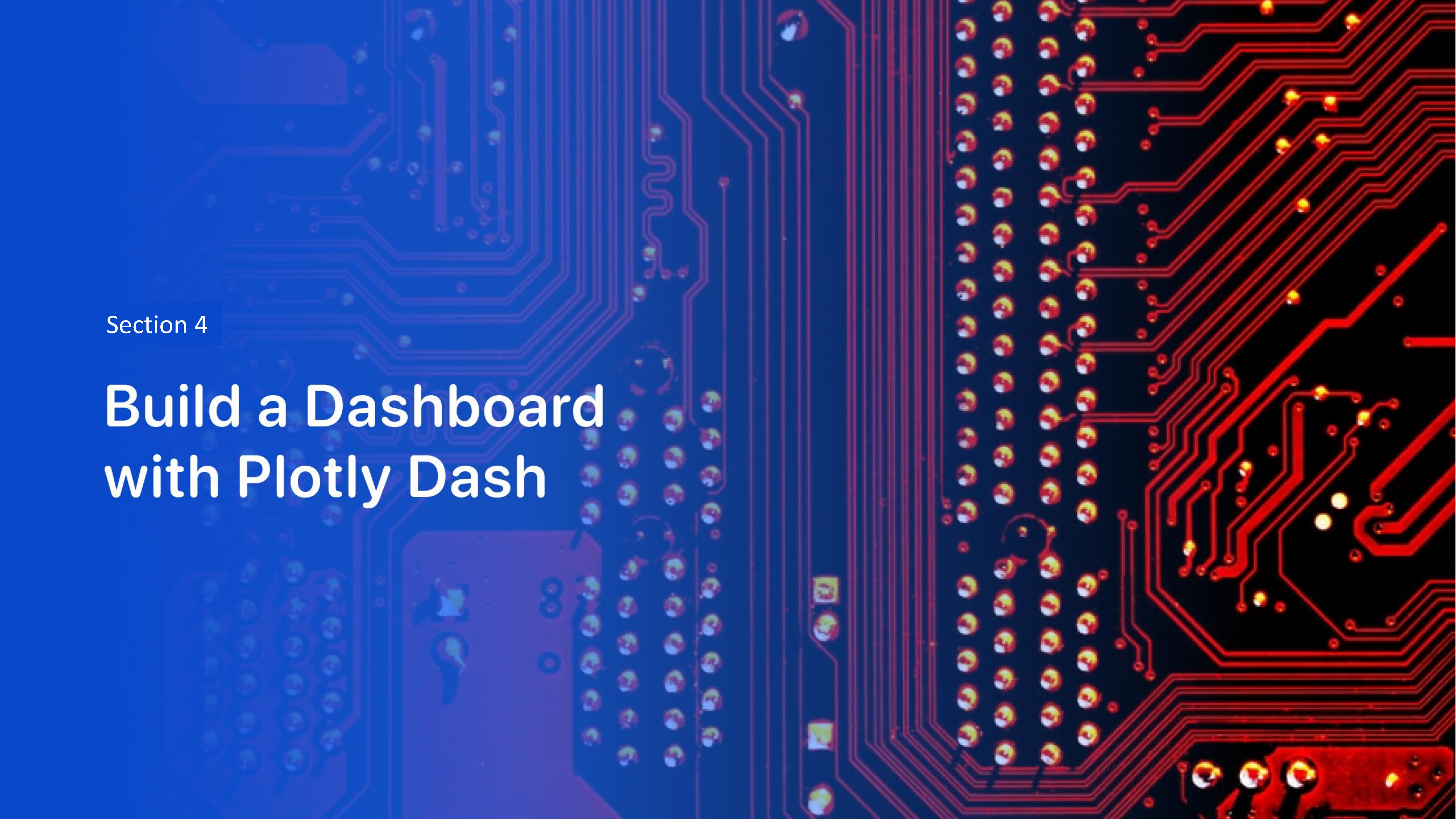
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot
  - From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.



Launch Site and its Proximities

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot
  - This map exposes the distance between site and its proximities.



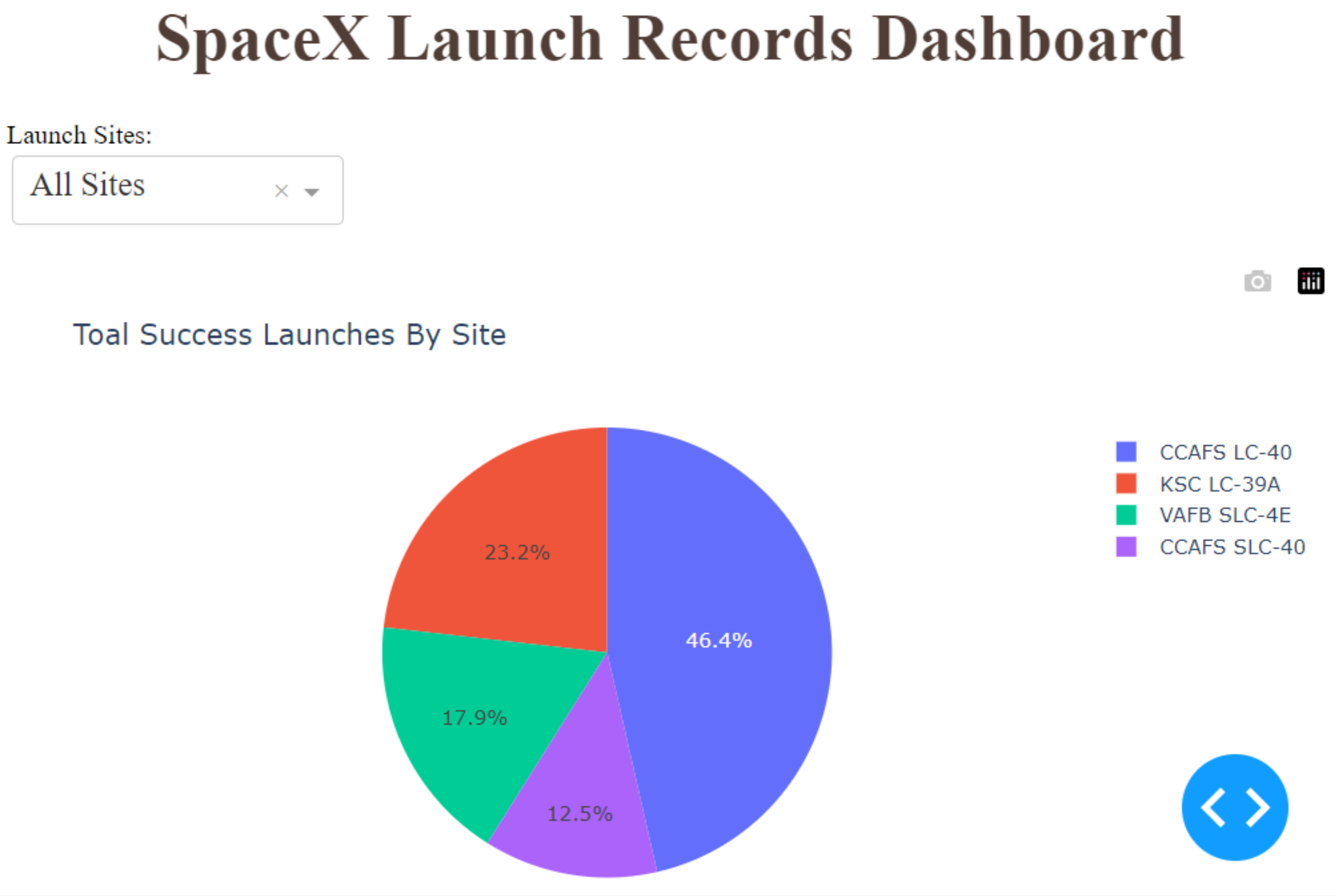


Section 4

Build a Dashboard with Plotly Dash

Launch Outcomes for All Sites

- Show the screenshot of launch success count for all sites, in a pie chart
- Explain the important elements and findings on the screenshot
  - The CCAFS LC-40 site has high success rate.

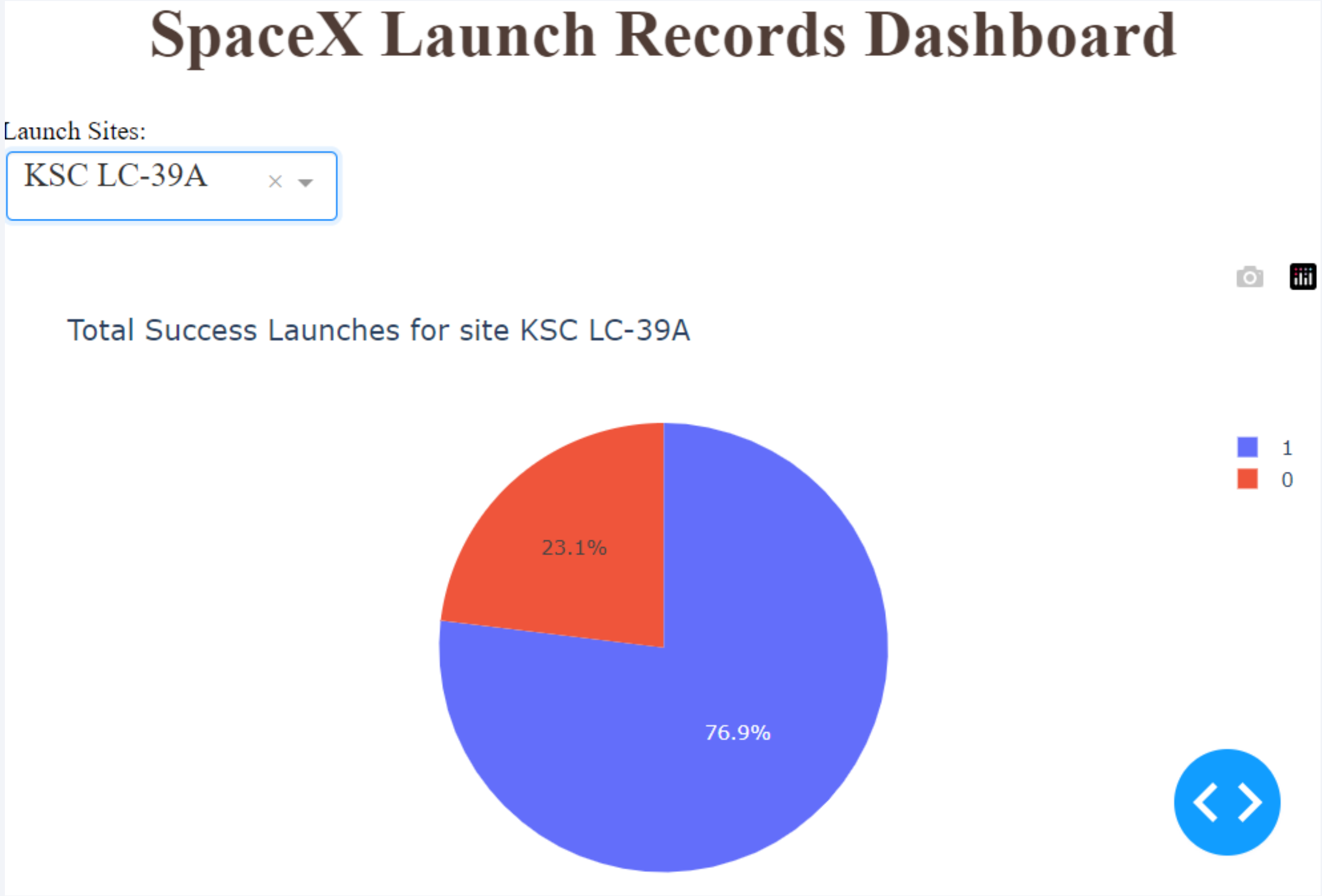


SpaceX Launch Records Dashboard

Launch Sites: All Sites

Highest Success Rate Site

- Show the screenshot of the pie chart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot
  - The KSC LC-39A site have the highest success rate.



SpaceX Launch Records Dashboard

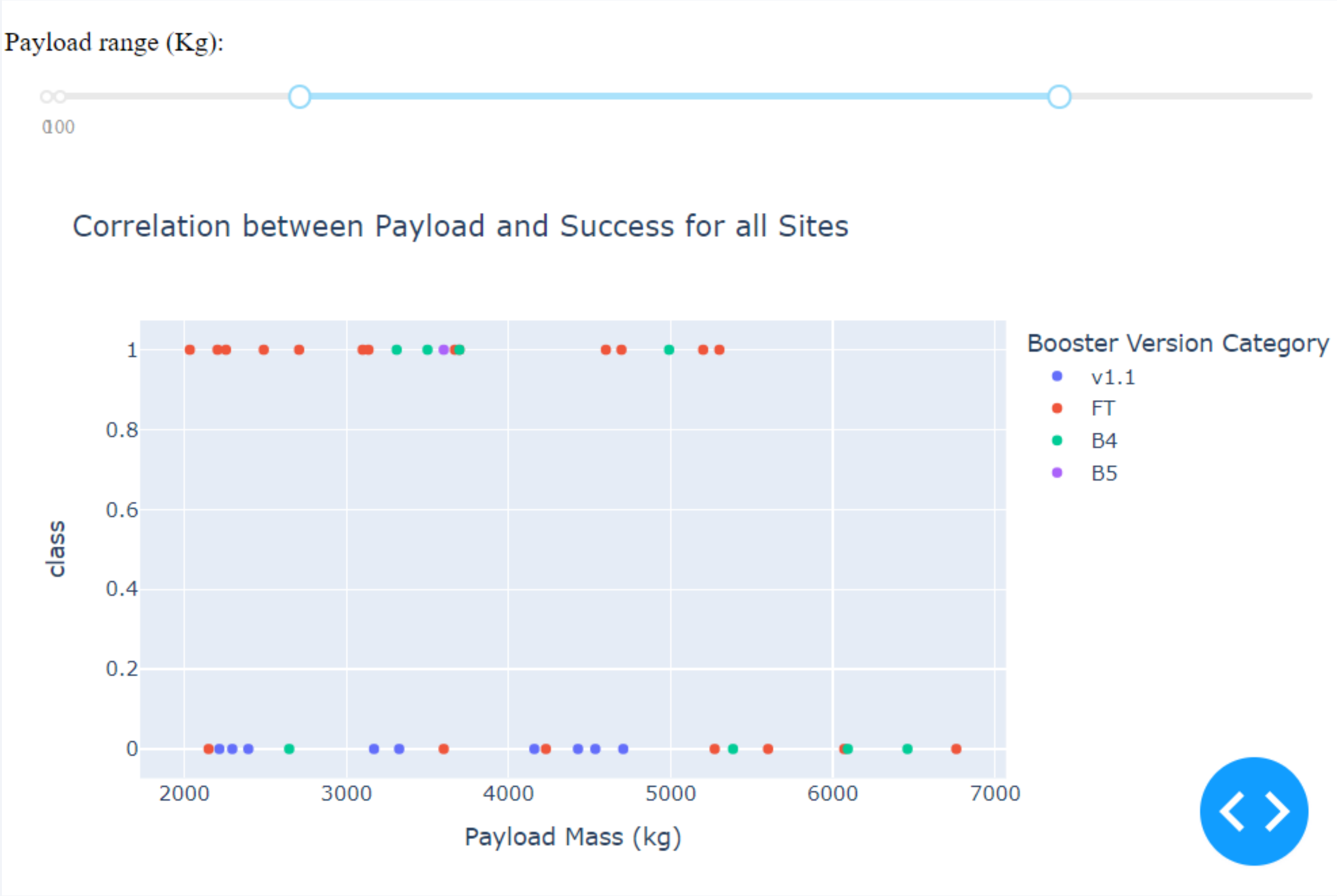
Launch Sites:

KSC LC-39A

Total Success Launches for site KSC LC-39A

Payload vs Launch Outcomes

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.
    - The FT booster version category has largest success rate between 2000 and 7000 payload mass (Kg).



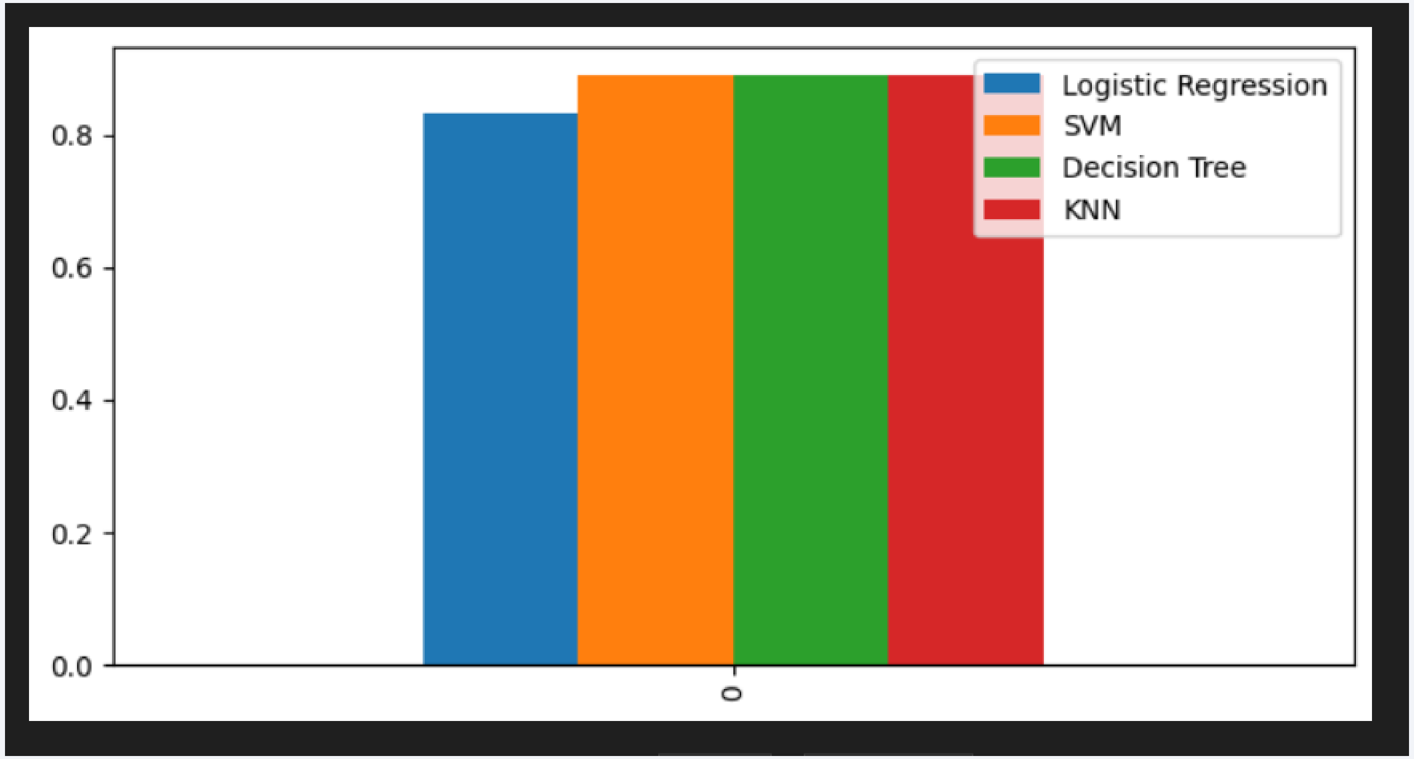
Payload range (Kg):

Section 5

Predictive Analysis (Classification)

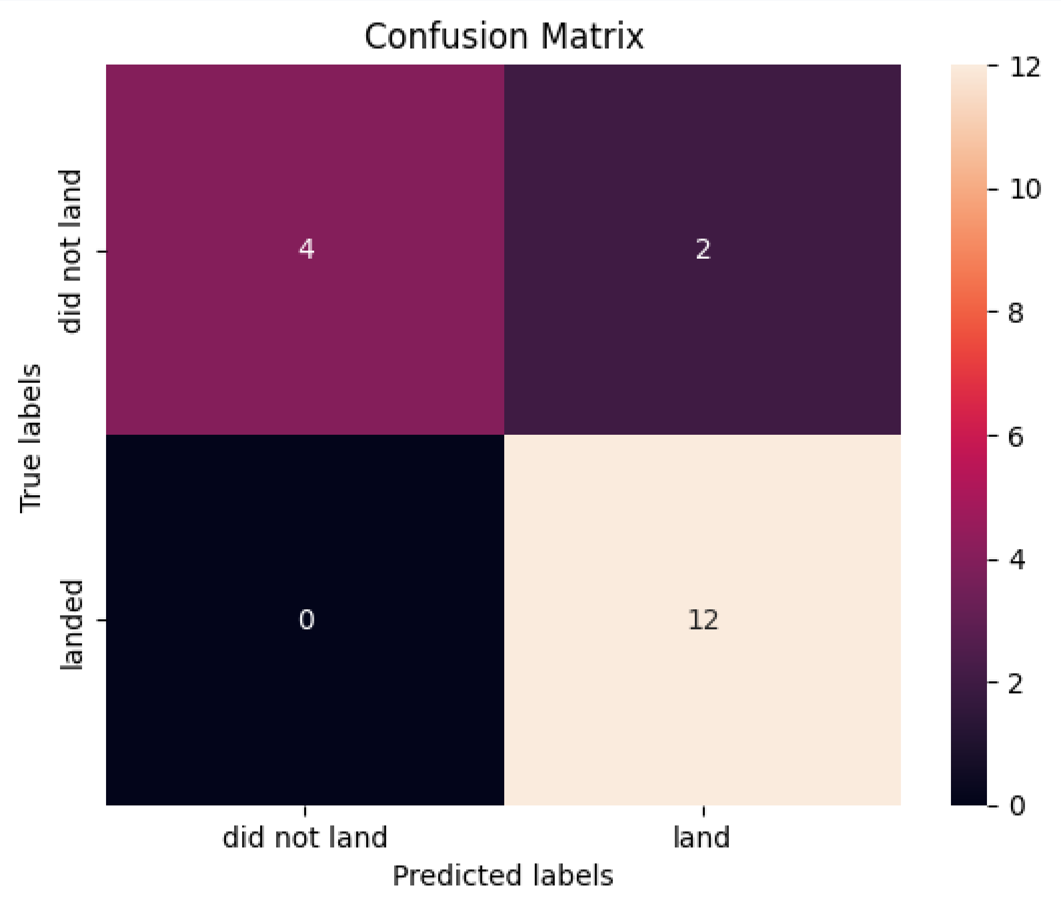
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy
  - All the classification algorithms except Logistic Regression are giving the high same accuracy.



Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation
  - All the models are showing the best prediction rate. Except Logistic Regression other all algorithms have high accuracy prediction score.



Conclusions

- The CCAFS LC-40 and KSC LC-39A sites has high success rate.
- The FT booster version category has largest success rate between 2000 and 7000 payload mass (Kg).
- The SpaceX dashboard provides the launch site and payload mass wise success rates charts. Apply different launch site and payload range to predict from which orbit provides high success rate.
- Applied various classification algorithms and the SVM, KKN and Decision Tree algorithms models provides high accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- Pandas filters examples
  - https://www.listendata.com/2019/07/how-to-filter-pandas-dataframe.html
  - https://sparkbyexamples.com/pandas/pandas-dataframe-filter/
  - https://pandas.pydata.org/docs/user_guide/index.html
- Notebooks and Datasets used in the project are located below
  - https://github.com/prakashare/Ibm-DataScience/tree/main/Capstone

Thank you!