IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Kamonpurn Songjalern>
<31 12 2023>

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection methodology
  - Perform data wrangling
  - Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models
- Summary of all results
  - EDA with visualization results
  - EDA with SQL results
  - interactive map with Folium results
  - Plotly Dash dashboard results
  - predictive analysis (classification)

Introduction

- SpaceX has revolutionized the economics of rocket launches with its Falcon 9 rocket. Priced at a mere 62 million dollars per launch, SpaceX's offering significantly undercuts its competitors, whose prices soar upward of 165 million dollars. The secret behind this cost efficiency lies in SpaceX's groundbreaking approach to rocket reusability. By ingeniously designing the Falcon 9's first stage to be reusable, SpaceX has transformed the traditionally expendable nature of space launches. This innovation not only slashes production costs but also paves the way for rapid turnarounds, in-house manufacturing efficiencies, and a strategic edge in capturing a diverse market of commercial and governmental clients. As a result, SpaceX's disruptive model has redefined the economics of space access, making it a pivotal player in shaping the future of space exploration and satellite deployment.
- **Problems to Address:**
    1. **Identifying Factors for Landing Outcome**
    2. **Analyzing Relationships between Variables**
    3. **Determining Best Conditions for Successful Landing**

Section 1

Methodology

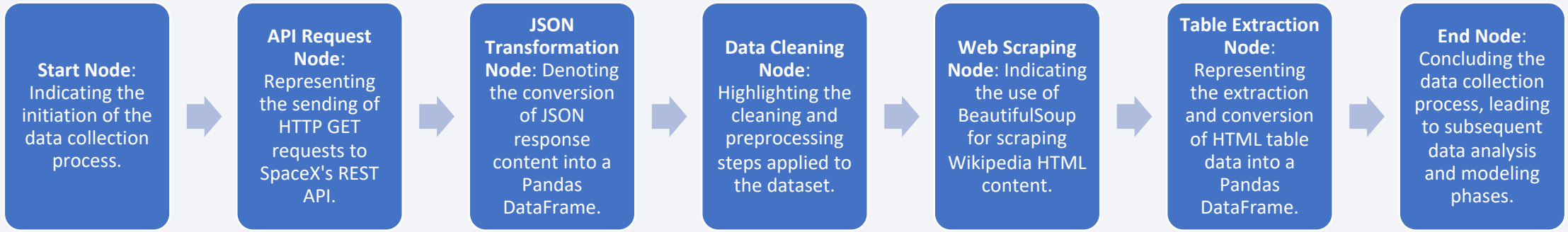
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by API from the SpaceX REST API and web scraping to collect Falcon 9 historical launch records from a Wikipedia
- Perform data wrangling
  - data was processed using one-hot encoding for categorical landing outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - build, tune, evaluate classification models

Data Collection

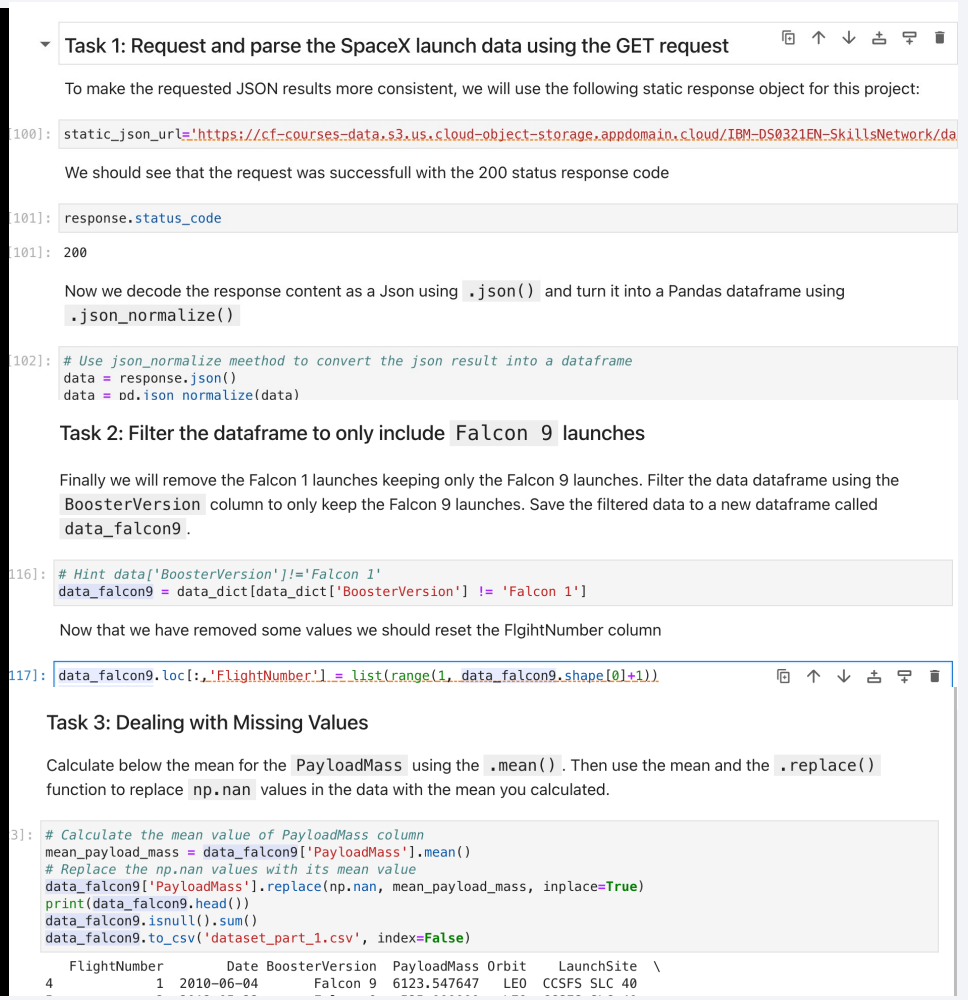
1. **REST API Data Collection:**
    1. Initiated the data collection process by sending GET requests to SpaceX's REST API endpoint to retrieve relevant Falcon 9 launch data.
    2. Decoded the API response content formatted as JSON and transformed it into a structured Pandas DataFrame using the json_normalize function.
    3. Conducted data cleaning processes, including handling missing values, standardizing formats, and ensuring data consistency to prepare the dataset for further analysis.
2. **Web Scraping from Wikipedia:**
    1. Utilized BeautifulSoup, a Python library, to scrape Falcon 9 historical launch records presented in HTML tables on specific Wikipedia pages.
    2. Parsed the HTML tables to extract relevant data fields, such as launch dates, mission outcomes, payload details, etc.
    3. Converted the scraped data into a structured Pandas DataFrame to facilitate exploratory data analysis, visualization, and integration with other datasets.



Data Collection – SpaceX API

- **API Endpoint**: Specific URL provided by SpaceX to access Falcon 9 launch data.
- **HTTP GET Request**: Method used to retrieve data from the SpaceX REST API.
- **JSON Response**: Data format returned by the SpaceX API containing Falcon 9 launch records.
- **Data Transformation**: Conversion of JSON response into a structured Pandas DataFrame using json_normalize.
- **Data Cleaning**: Process of refining and preparing the collected data for analysis.

https://github.com/quiisleepyhead/My-submission/blob/2e6c0542cd03dcdae9fdb678c59fbfbf8deadb7b/jupyter-labs-spacex-data-collection-api.ipynb



Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
[100]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/da
```

We should see that the request was successfull with the 200 status response code

```
[101]: response.status_code
[101]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
[102]: # Use json_normalize meethod to convert the json result into a dataframe
data = response.json()
data = pd.json_normalize(data)
```

Task 2: Filter the dataframe to only include `Falcon 9` launches

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

```
[116]: # Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = data_dict[data_dict['BoosterVersion'] != 'Falcon 1']
```

Now that we have removed some values we should reset the FlgihtNumber column

```
[117]: data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
```

Task 3: Dealing with Missing Values

Calculate below the mean for the `PayloadMass` using the `.mean()`. Then use the mean and the `.replace()` function to replace `np.nan` values in the data with the mean you calculated.

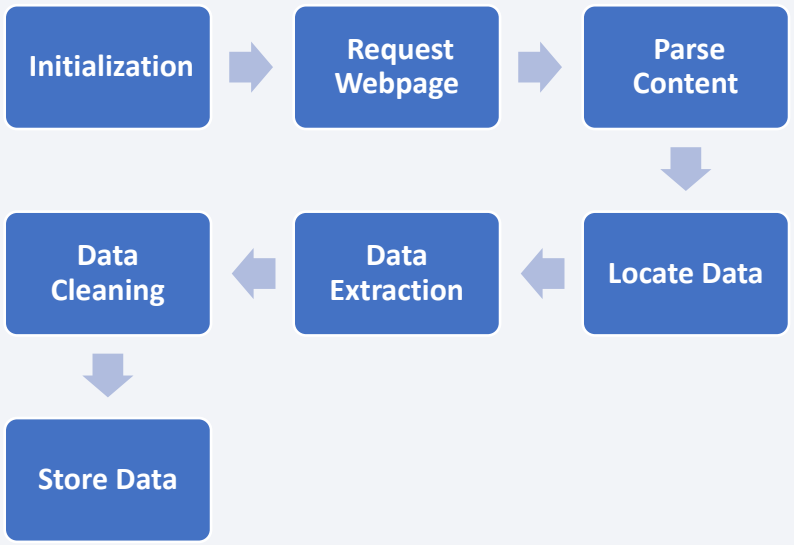
```
[3]: # Calculate the mean value of PayloadMass column
mean_payload_mass = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, mean_payload_mass, inplace=True)
print(data_falcon9.head())
data_falcon9.isnull().sum()
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

```
   FlightNumber        Date BoosterVersion  PayloadMass Orbit    LaunchSite  \
4             1  2010-06-04       Falcon 9  6123.547647   LEO  CCSFS SLC 40
```

Data Collection - Scraping

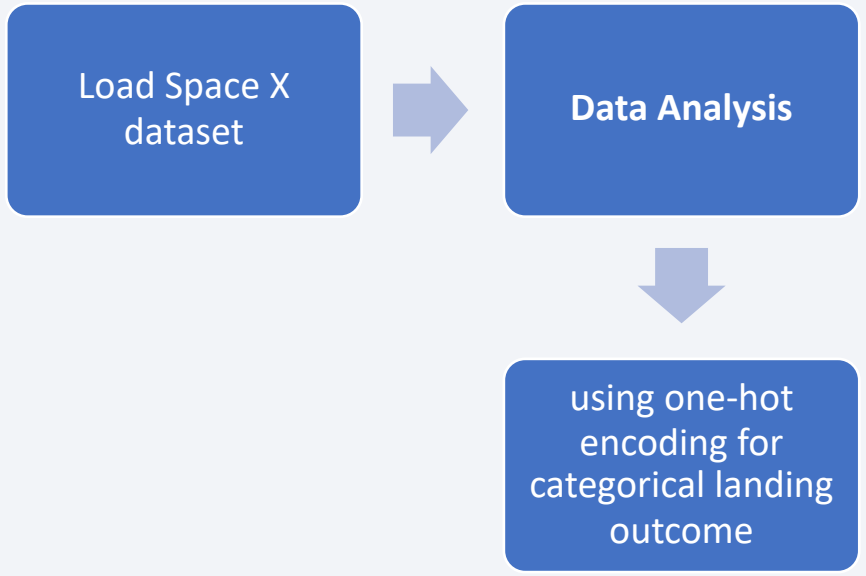
1. **Initialization:** Import required libraries and define the target URL.
2. **Request Webpage:** Fetch webpage content using requests.
3. **Parse Content:** Use BeautifulSoup to navigate and parse HTML structure.
4. **Locate Data:** Identify HTML elements containing the target data.
5. **Data Extraction:** Extract required data, handling inconsistencies.
6. **Data Cleaning:** Clean and format extracted data.
7. **Store Data:** Convert data to desired format (e.g., DataFrame, CSV).
8. **Error Handling:** Implement error handling and logging mechanisms.

https://github.com/quiisleepyhead/My-submission/blob/0a3fe40c3a4d4ebba490cf4b8d5dde85a22eb4c3/jupyter-labs-webscraping.ipynb



Data Wrangling

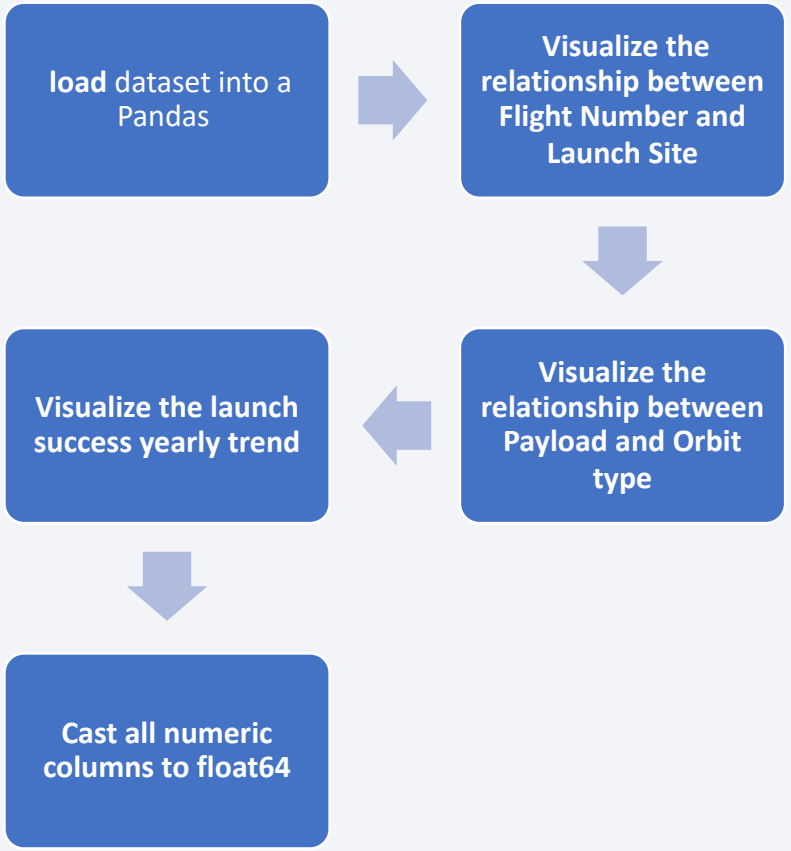
- Data was processed using one-hot encoding for categorical landing outcome ; convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



https://github.com/quiisleepyhead/My-submission/blob/c0e08442ac99533f0fb463686240d2560dc65e8e/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Perform exploratory Data Analysis and Feature Engineering
using Pandas and Matplotlib
•Exploratory Data Analysis
•Preparing Data Feature Engineering



- https://github.com/quiisleepyhead/My-submission/blob/8172fb5d9dcd4f8735fb91ecc6e44fca5716488f/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- **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)**
- **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 the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
- **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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order**

- https://github.com/quiisleepyhead/My-submission/blob/8172fb5d9dcd4f8735fb91ecc6e44fca5716488f/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Task 1: Mark All Launch Sites on a Map**
- **Map Objects Created:**
  - **Markers**: Positioned on each launch site location.
- **Why Added**:
  - To provide a visual representation of existing launch site locations.
  - Enables stakeholders to identify geographical distribution and clustering of launch sites.
- **Task 2: Mark the Success/Failed Launches for Each Site**
- **Map Objects Created:**
  - **Differentiated Markers**: Using distinct colors or icons to represent successful and failed launches at each site.
- **Why Added**:
  - To correlate launch outcomes with specific site locations.
  - Facilitates analysis of site-specific success rates and potential influencing factors.
- **Task 3: Calculate Distances Between Launch Sites and Proximities**
- **Map Objects Created:**
  - **Lines/Circles**: Representing trajectories or proximity zones around each launch site.
- **Why Added**:
  - To evaluate the spatial relationship between launch sites and their surrounding areas.
  - Assists in determining optimal locations based on factors such as safety, accessibility, and environmental considerations.

https://github.com/quiisleepyhead/My-submission/blob/ee978eca45e8f14b4436151a75cff4087eef7432/lab_jupyter_launch_site_location.jupyterlite.ipynb

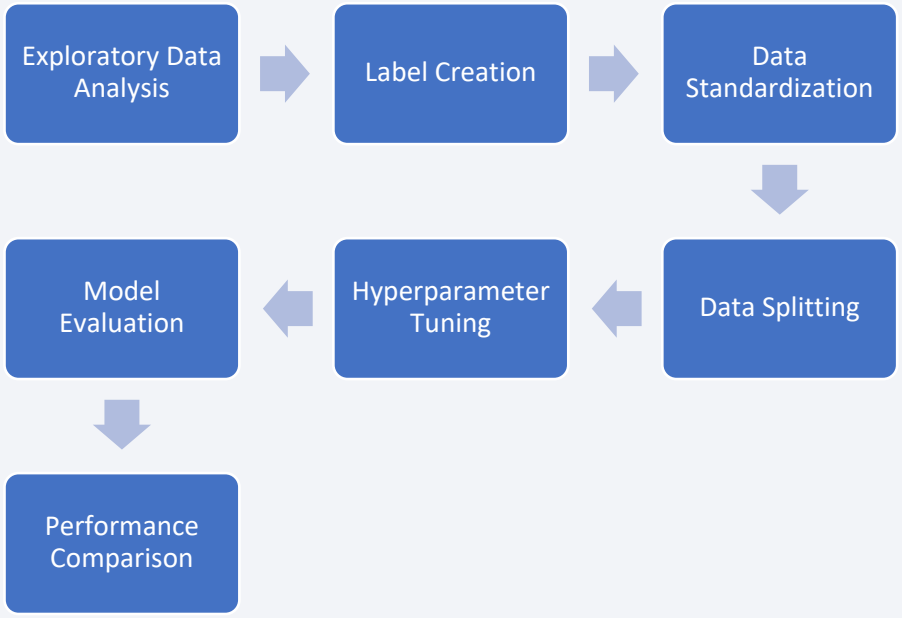
Build a Dashboard with Plotly Dash

- **Plots/Graphs:**

1. **Pie Chart - Success vs. Failed Launches**:
    1. **Interactivity**: Dropdown selection for individual launch sites or all sites.
    2. **Why Added**:
        1. To provide an overview of the success rate for launches across different sites or for all sites combined.
        2. Enables stakeholders to quickly assess the overall success rate and identify potential variations among different launch sites.
2. **Scatter Chart - Payload vs. Outcome**:
    1. **Interactivity**:
        1. Dropdown selection for individual launch sites or all sites.
        2. Range slider to filter payloads within a specific range.
    2. **Why Added**:
        1. To visualize the relationship between the payload mass and launch outcomes (success or failure).
        2. Facilitates analysis of how varying payload masses influence the success rate across different launch sites or specific payload ranges.

https://github.com/quiisleepyhead/My-submission/blob/ee978eca45e8f14b4436151a75cff4087eef7432/spacex_dash_app.py

Predictive Analysis (Classification)



1.Exploratory Data Analysis (EDA):
1. Understand data distributions, relationships, and patterns.

2.Label Creation:
1. Create a column representing the class for classification.

3.Data Standardization:
1. Standardize features to have mean=0 and variance=1.

4.Data Splitting:
1. Divide data into training and test sets

5.Hyperparameter Tuning:
1. Find optimal hyperparameters for SVM, Classification Trees, and Logistic Regression using techniques like Grid Search or Randomized Search.

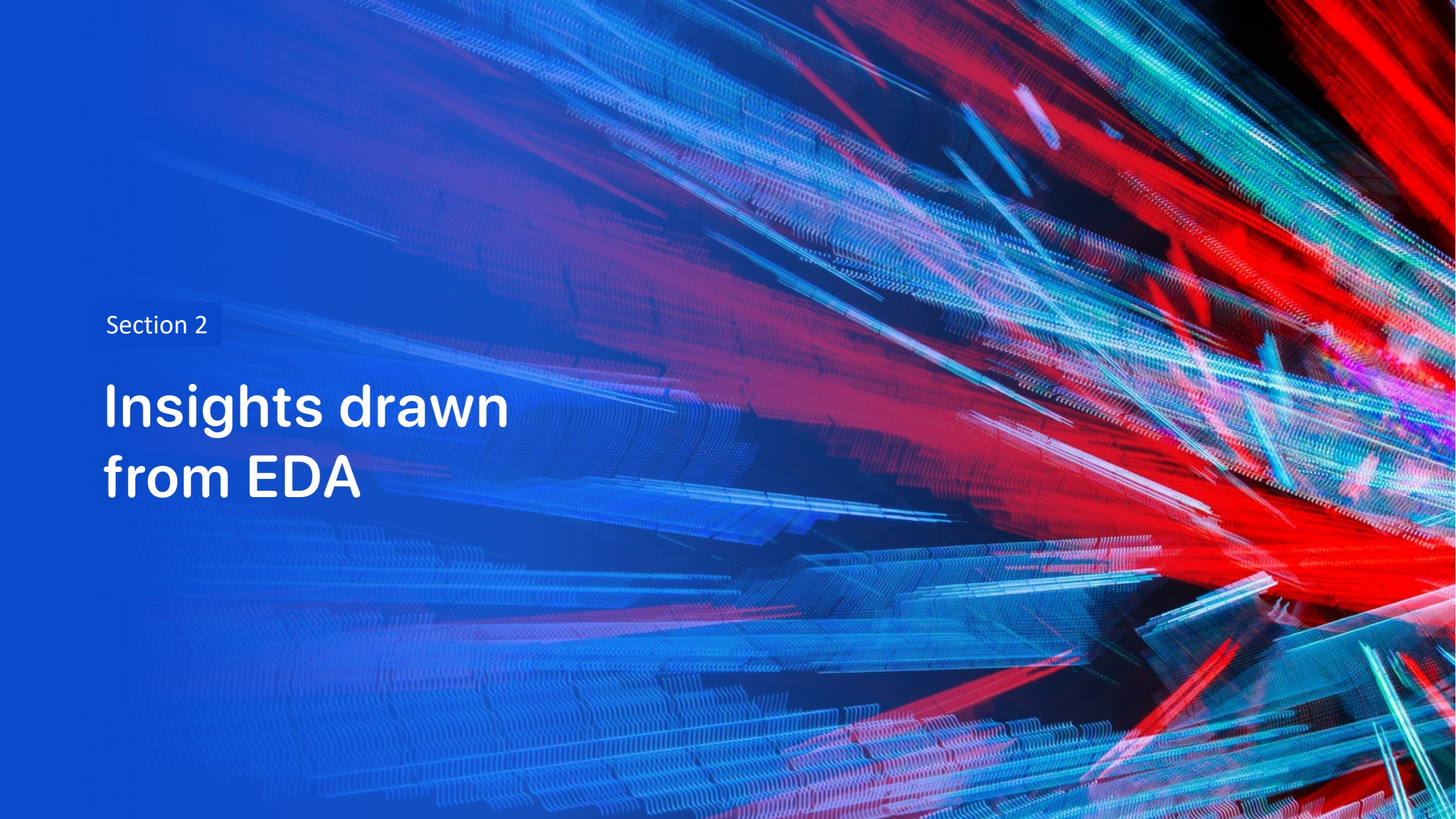
6.Model Training & Evaluation:
1. Train models with the training data.
2. Evaluate models using test data based on chosen metrics (e.g., accuracy, precision, recall).

7.Performance Comparison:
1. Compare the performance of all models to determine the best-performing method.

https://github.com/quiisleepyhead/My-submission/blob/34794eab382ce60c551b168be78c844807fb7486/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

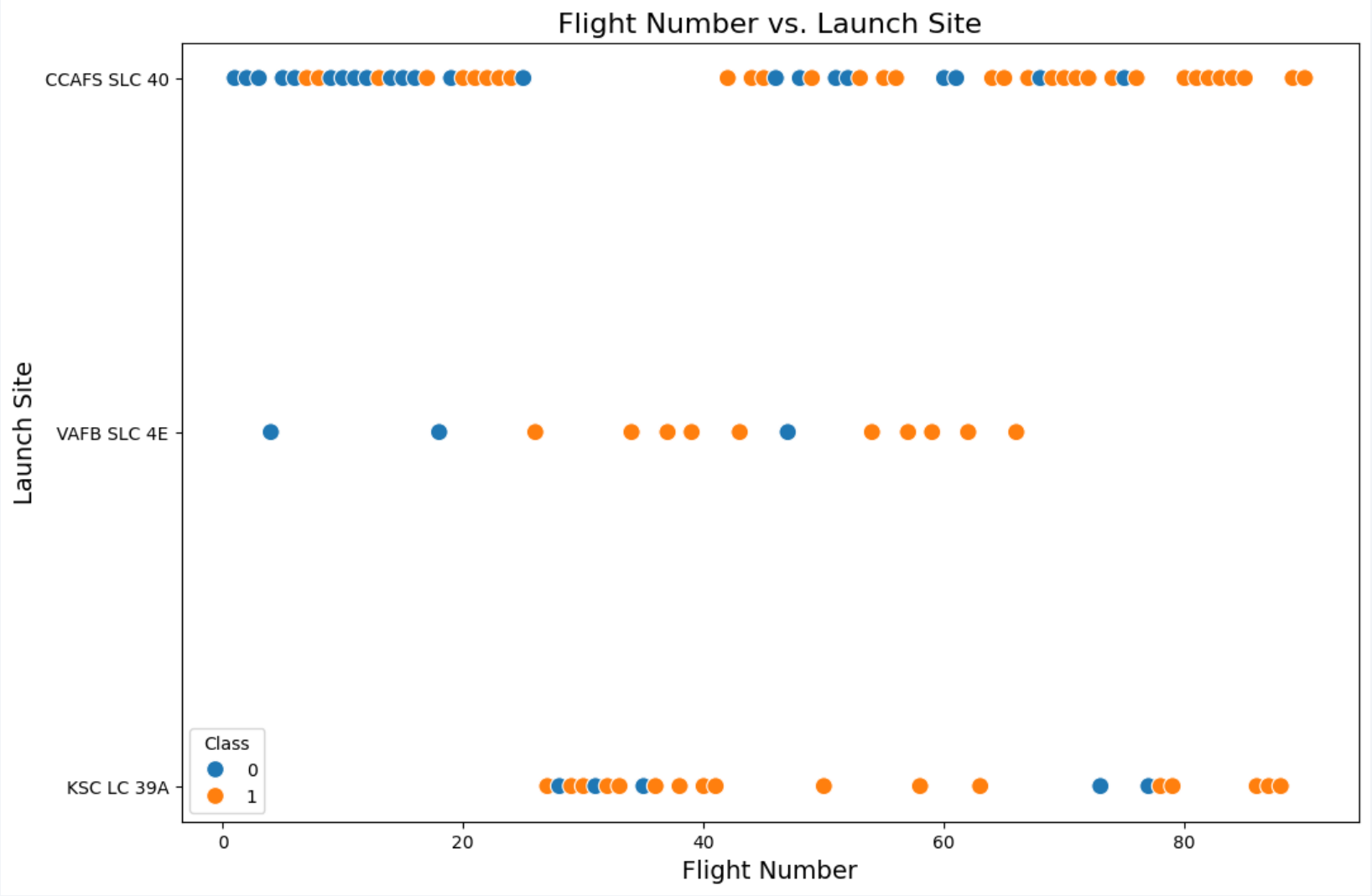
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Insights:

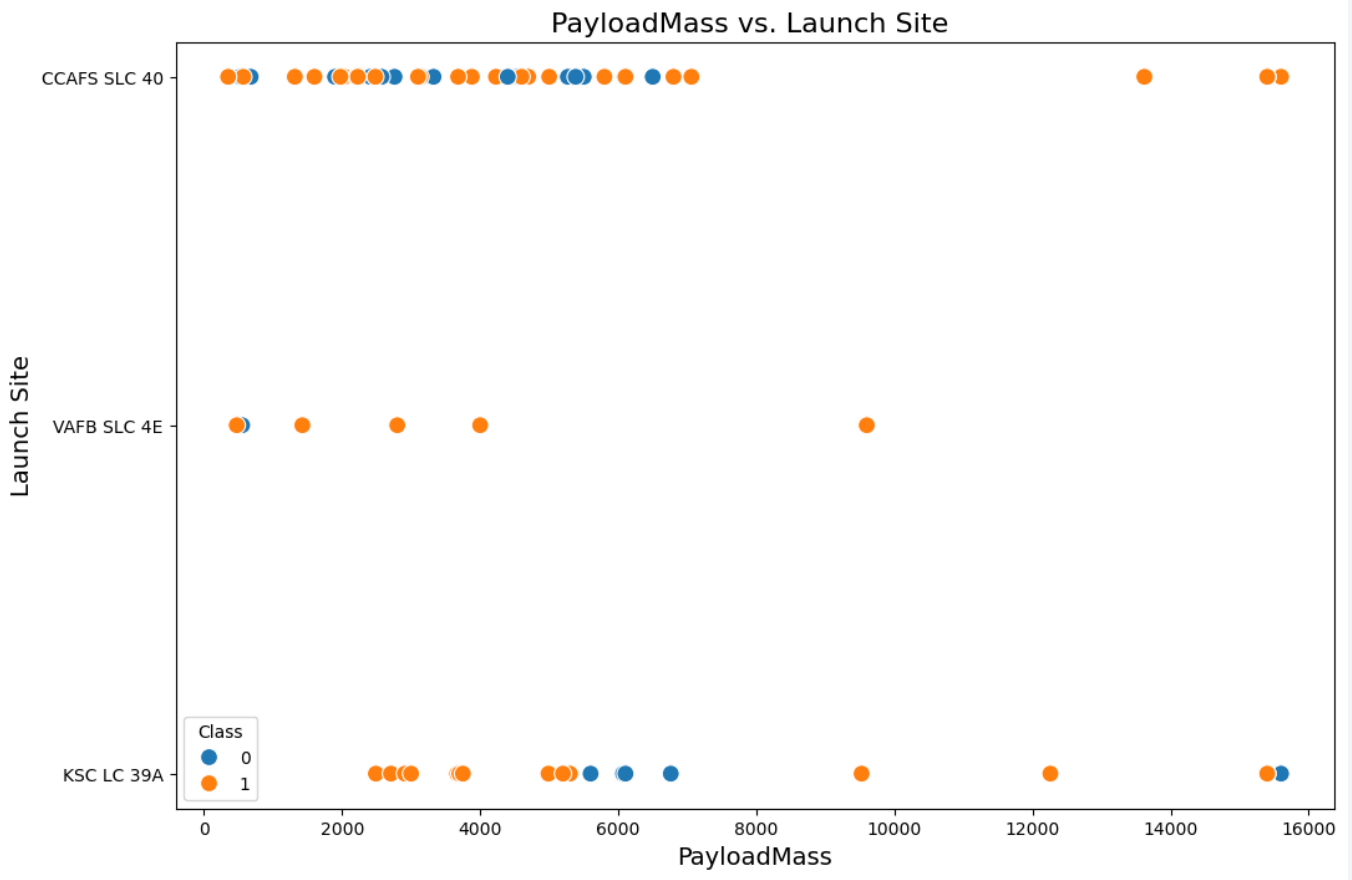
- Scatter plot indicates a general trend: higher flight number correlate with increased success rates across launch sites.
- CCAFS SLC40 shows inconsistent success rates despite significant flight number



Payload vs. Launch Site

Insights:

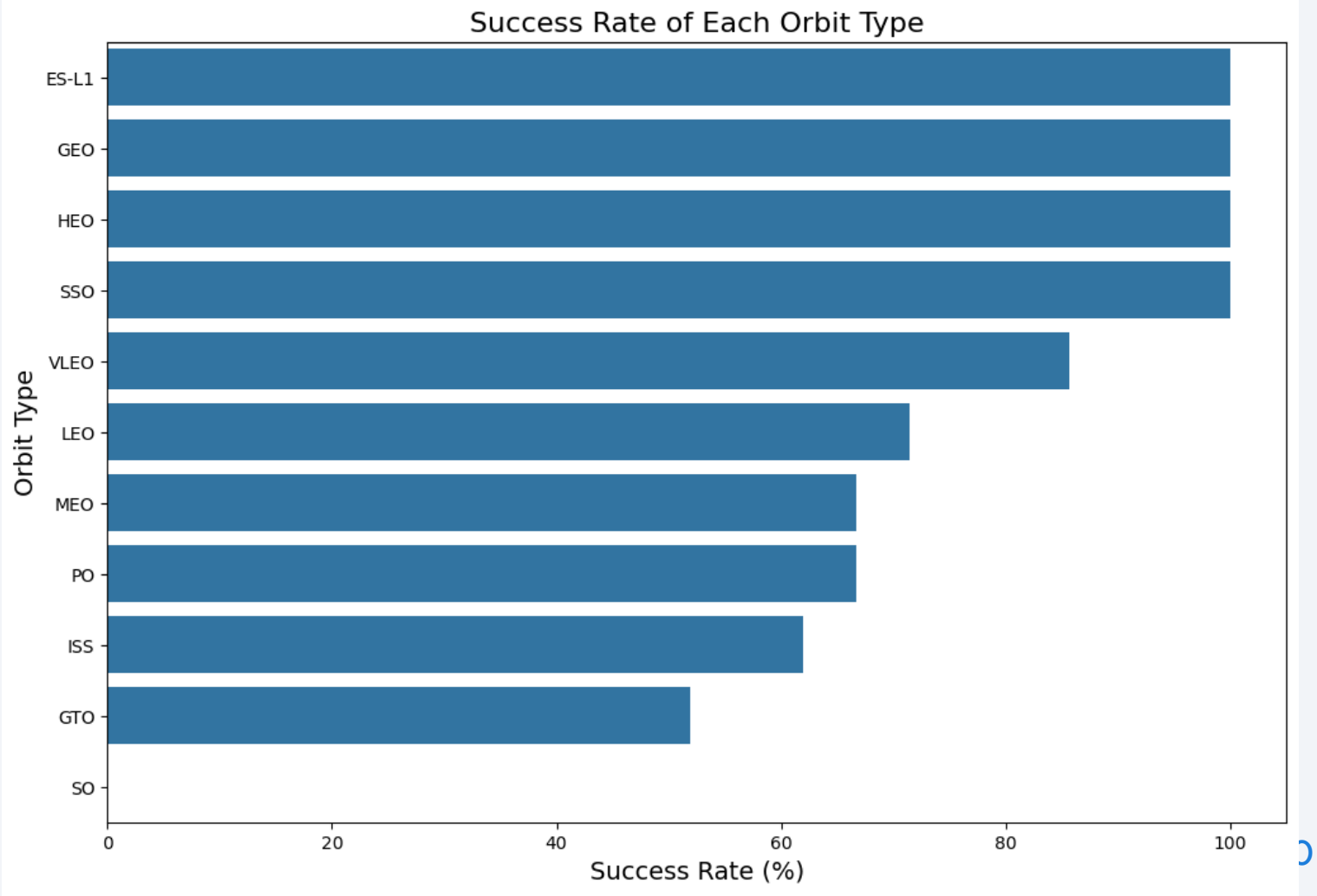
- Scatter plot reveals a significant trend: payloads exceeding 7000kg exhibit notably higher success rates.
- No definitive correlation observed between launch site and payload mass concerning success rates.
- VAFB-SLC launch site notably lacks launches for heavy payloads (>10000kg).



Success Rate vs. Orbit Type

Insights:

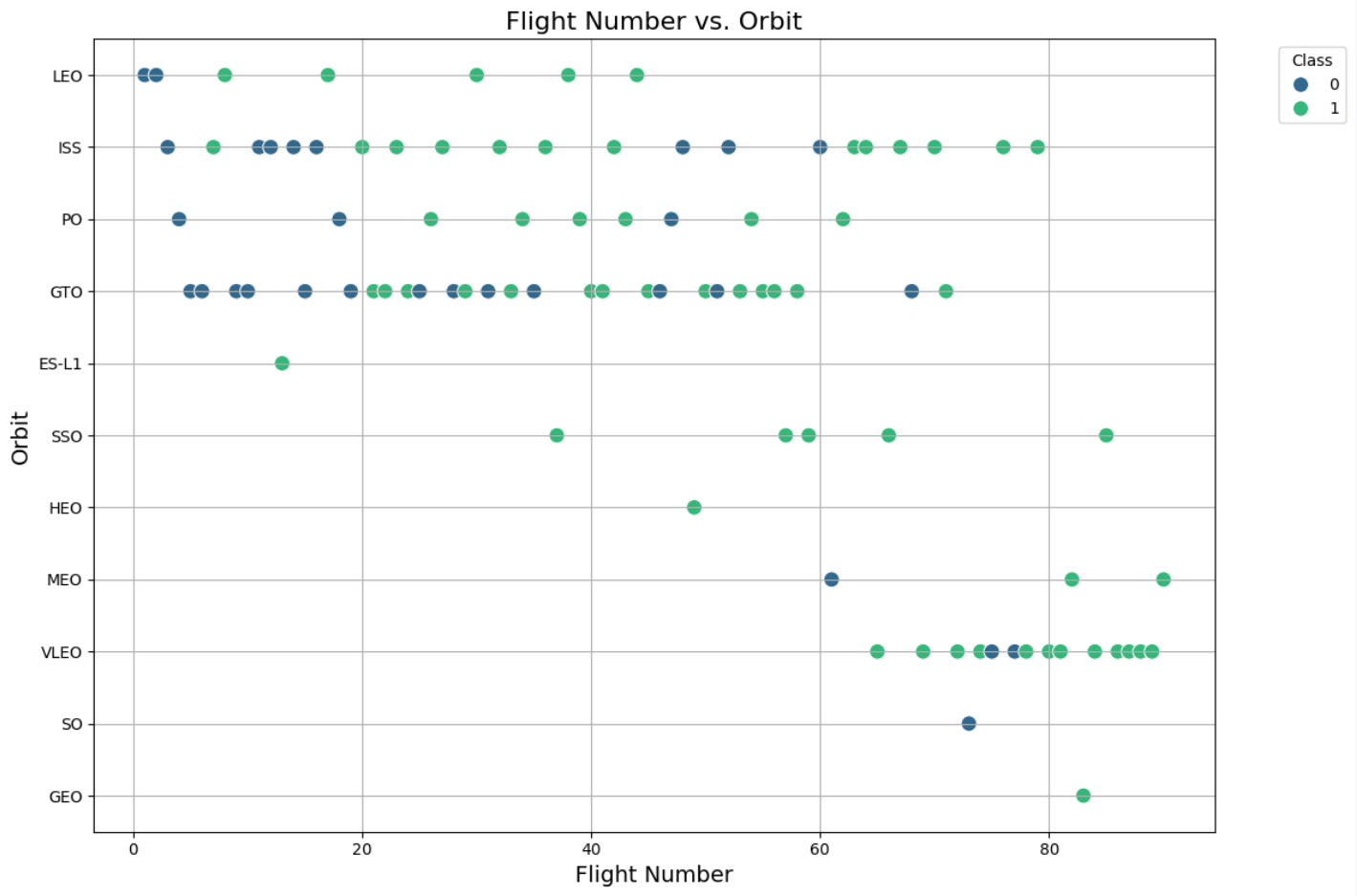
- Orbits like SSO, HEO, GEO, and ES-L1 demonstrate a 100% success rate.
- Conversely, the SO orbit registers a 0% success rate.



Flight Number vs. Orbit Type

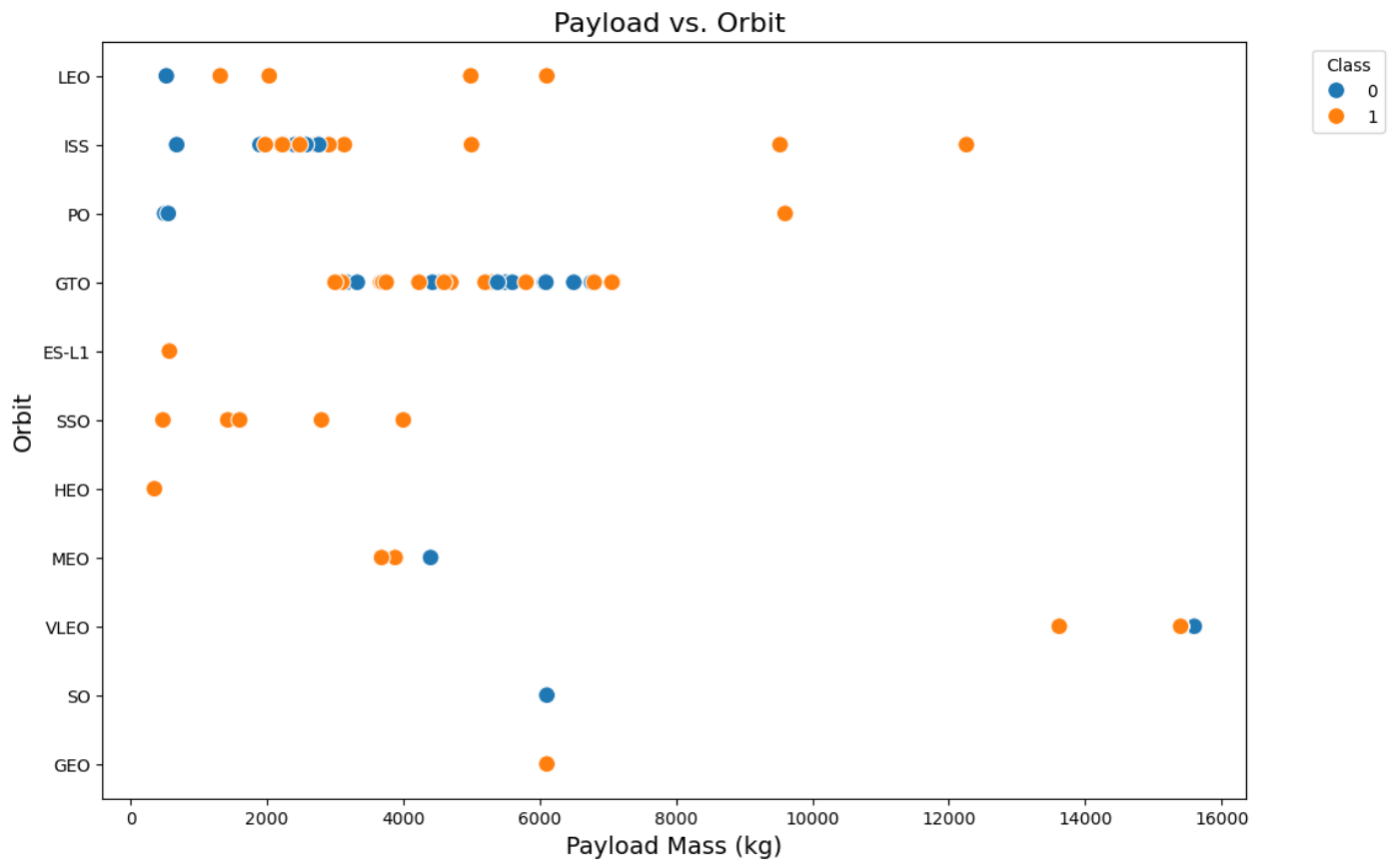
Insights:

- The scatter plot highlights distinct relationships between flight number and success rates across different orbit types.
- While LEO missions suggest enhanced success with increased flights, GTO missions present complexities requiring deeper analysis to understand influencing factors and optimize outcomes effectively.



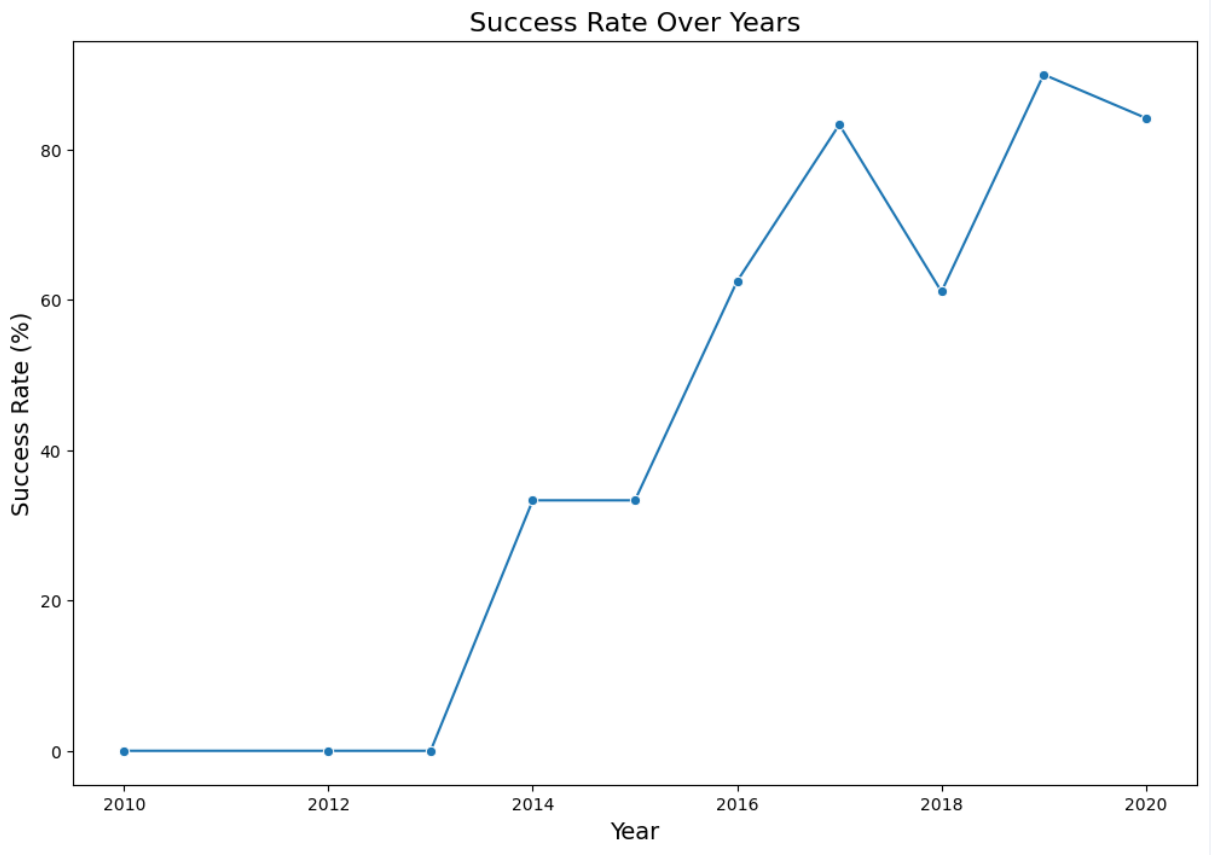
Payload vs. Orbit Type

- While Polar, LEO, and ISS orbits demonstrate a favorable correlation between heavy payloads and successful landings, GTO missions exhibit a less predictable relationship.
- Understanding these orbit-specific nuances is crucial for optimizing mission planning, payload selection, and operational strategies to enhance overall mission success and efficiency.



Launch Success Yearly Trend

- **Line Chart Insight:**
- **Trend Analysis:**
  - Line chart illustrates the average success rate of launches from 2013 to 2020.
  - Notably, the data reveals a consistent upward trend, indicating an increasing average success rate over the specified period.
- **Observation:**
  - Specifically, the success rate shows a continuous improvement from 2013, reaching its peak around 2020.



All Launch Site Names

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

```
]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

```
 * 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'

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

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%'LIMIT 5;
```

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

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

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_KG FROM SPA
```

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

Total_Payload_Mass_KG
48213

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_Payload_Mass_KG FROM SPACE
```

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

AVG_Payload_Mass_KG
2534.6666666666665

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) AS First_Successful_Landing_On_Ground_Pad_Date
```

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

First_Successful_Landing_On_Ground_Pad_Date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIK
```

```
 * 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

List the total number of successful and failure mission outcomes

```
]: %sql SELECT Mission_Outcome, COUNT (*) AS Total_Count FROM SPACEXTBL
```

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

]:

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

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[5]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
```

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

[5]:

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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT CASE WHEN substr(Date, 6, 2) = '01' THEN 'January' WHEN s
```

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

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Done.

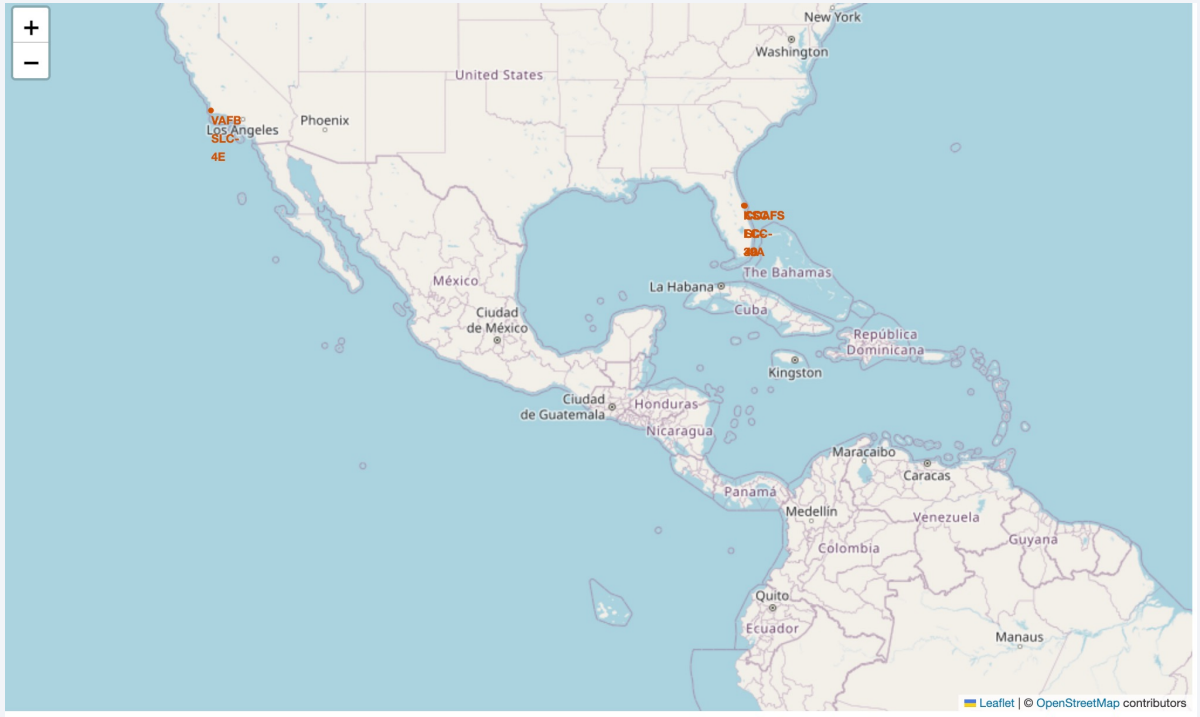
]:

Landing_Outcome	Outcome_Count
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

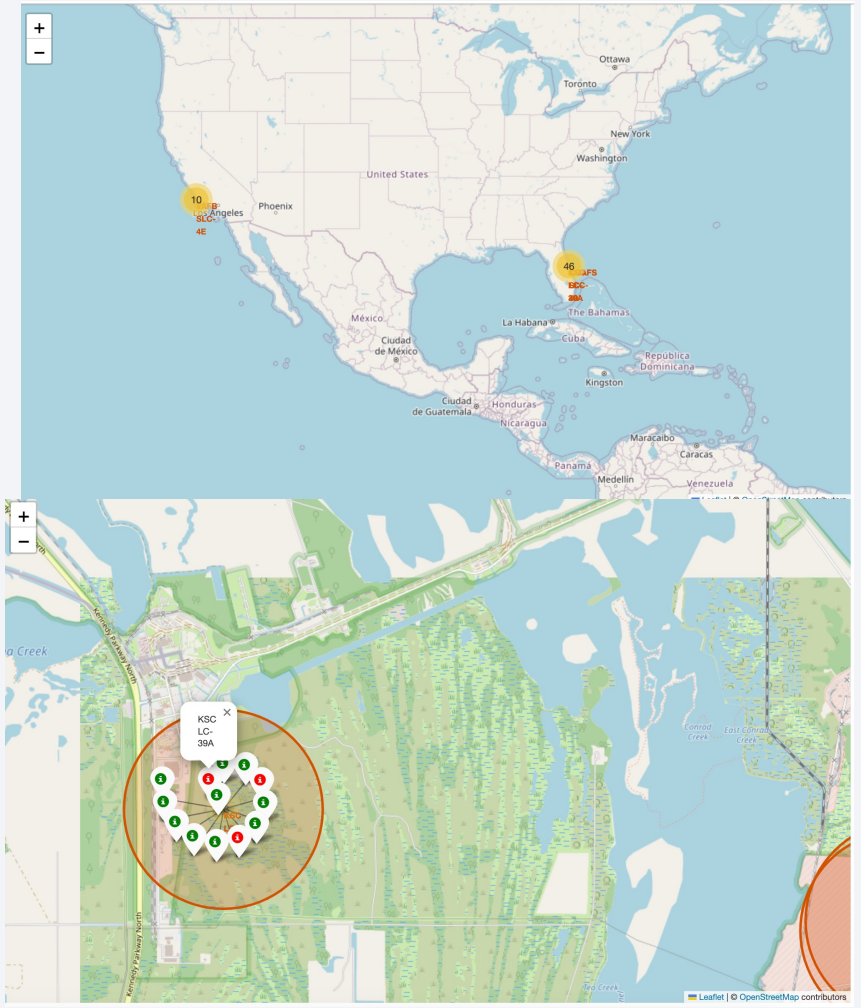
Mark all launch sites on a map

- **Launch Sites Marked:**
  - The map showcases SpaceX launch sites globally, each with a distinct marker and label for identification.
- **Geographic Proximity:**
  - **Equator:** Not all sites are near the Equator, impacting mission parameters.
  - **Coastline:** Sites vary in proximity to coastlines, influencing logistical and operational considerations.



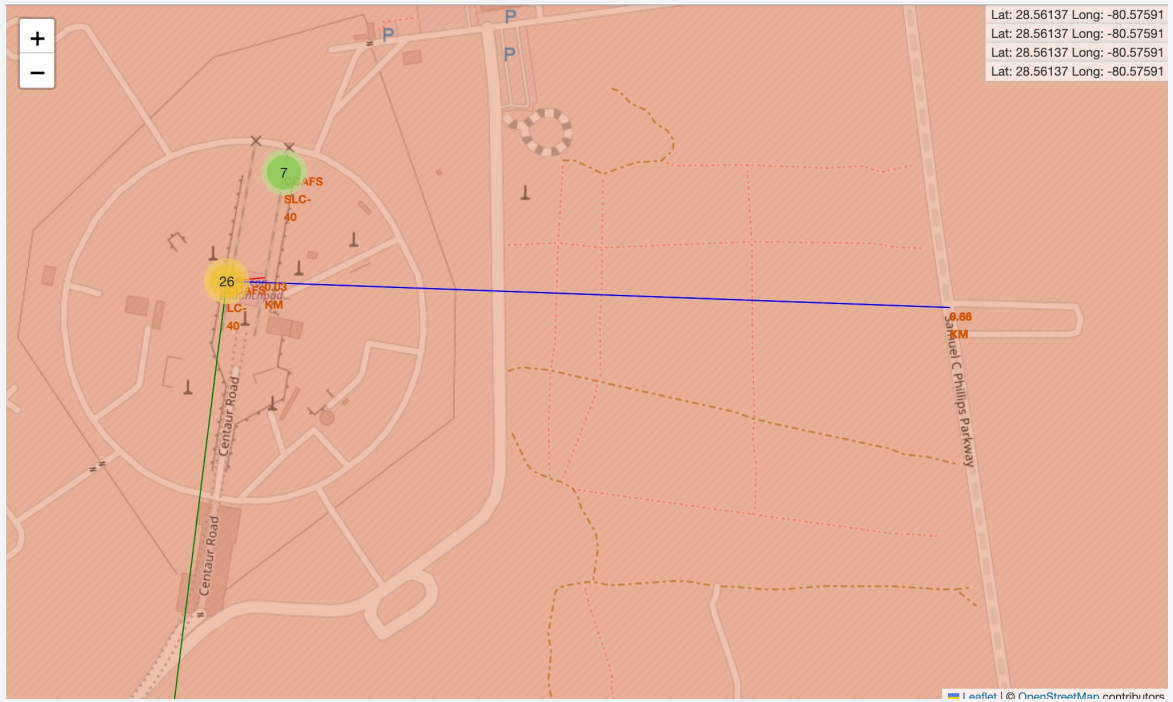
Mark the success/failed launches for each site on the map

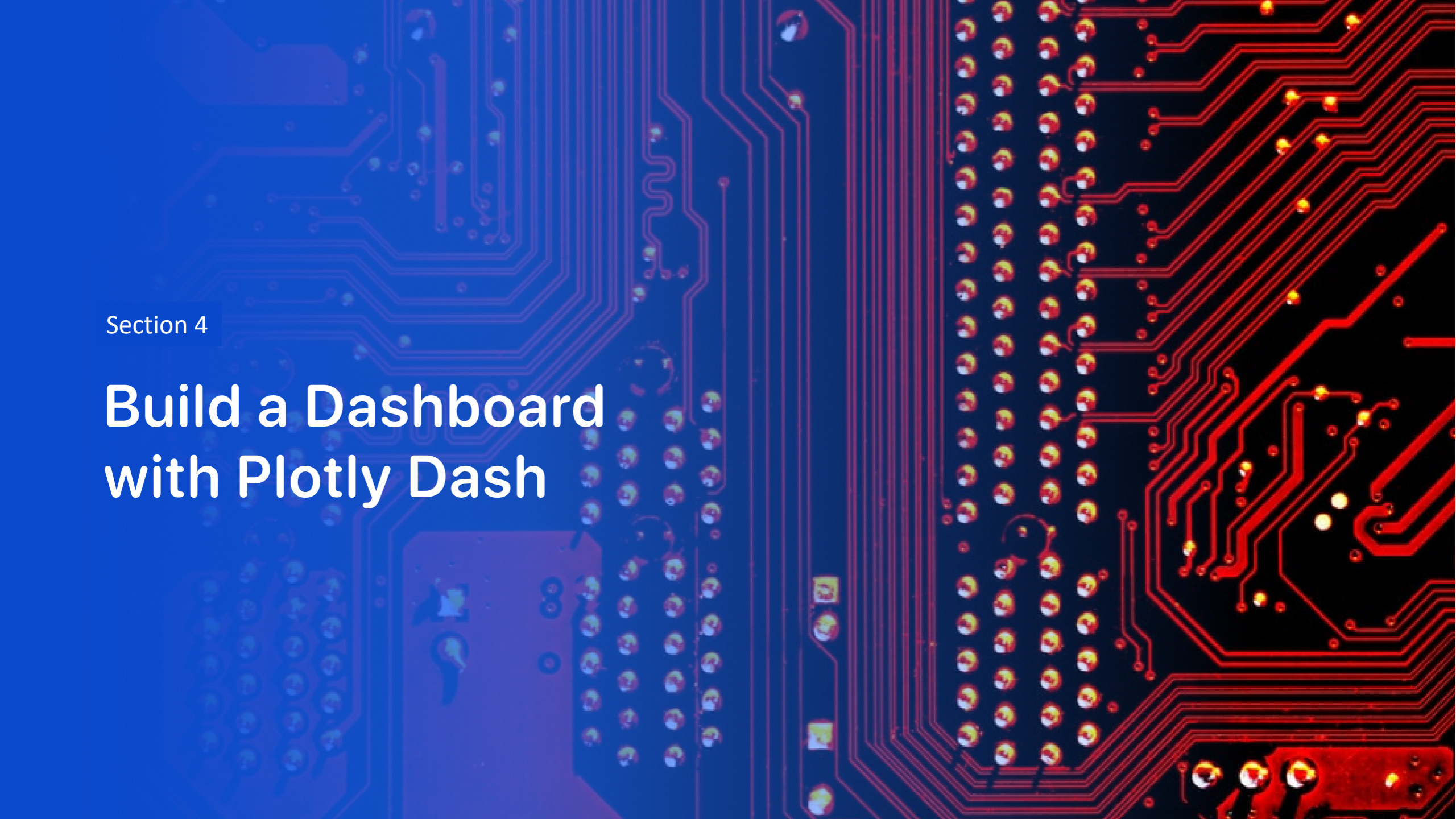
- **Launch Sites with Outcomes:**
  - The map now incorporates markers for each SpaceX launch record, color-coded based on the launch outcome:
    - **Green Marker:** Successful Launch (class=1)
    - **Red Marker:** Failed Launch (class=0)
- **Marker Clusters:**
  - Clusters of markers provide a visual representation of launch outcomes at each site, facilitating comparative analysis of success rates across different locations.
- **Success Rate Analysis:**
- By examining the color-labeled markers, it becomes evident that KSC LC-39A has a notably high success rate, indicated by a predominant presence of green markers relative to red markers.
- Other launch sites may exhibit varying success rates, with distinct patterns and clusters highlighting specific outcomes and operational trends.



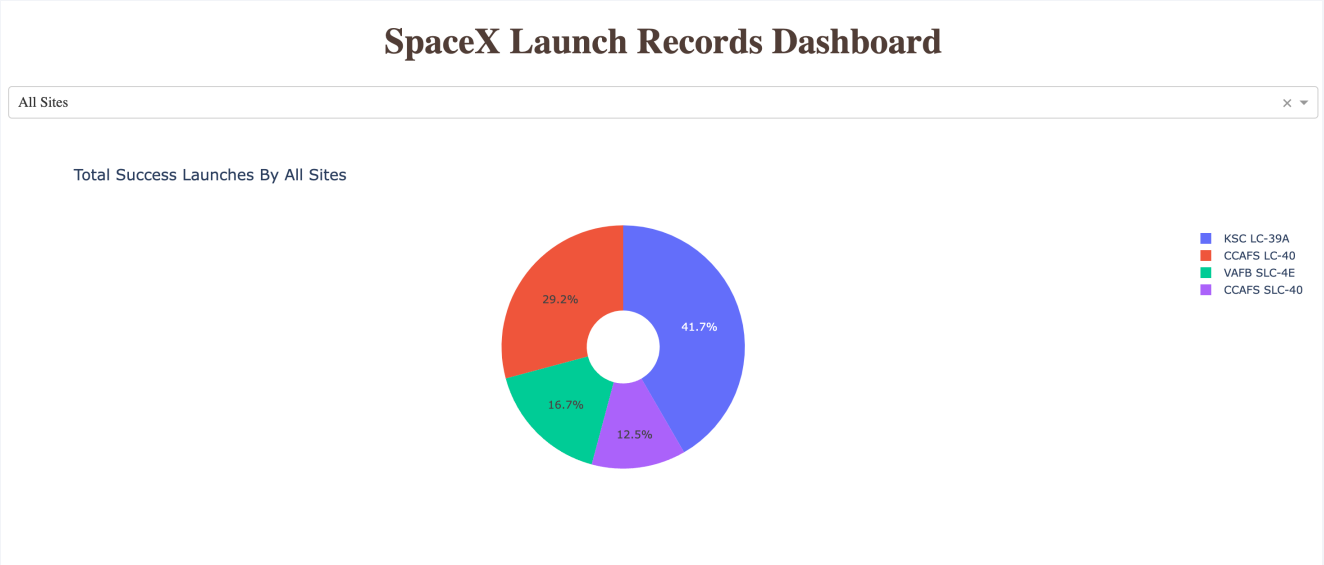
the distances between a launch site to its proximities

- **Proximity Markers & Lines:**The map showcases SpaceX launch sites with added markers and lines, illustrating distances to key proximities such as cities, railways, and highways.
- **Line Color Representation:Green Line:** Indicates the distance from the launch site to a specific city.
- **Red Line:** Represents the proximity of the launch site to a railway.
- **Blue Line:** Highlights the site's closeness to a highway.
- **CCAFS LC-40 Analysis:**The screenshot details CCAFS LC-40's proximity to critical infrastructures.
- A notable observation is the site's close proximity to a railway, indicated by a red line, with a distance of only 0.03 KM.
- Additionally, the map highlights other significant proximities, including highways and cities, offering a comprehensive spatial overview of the launch site's surroundings.



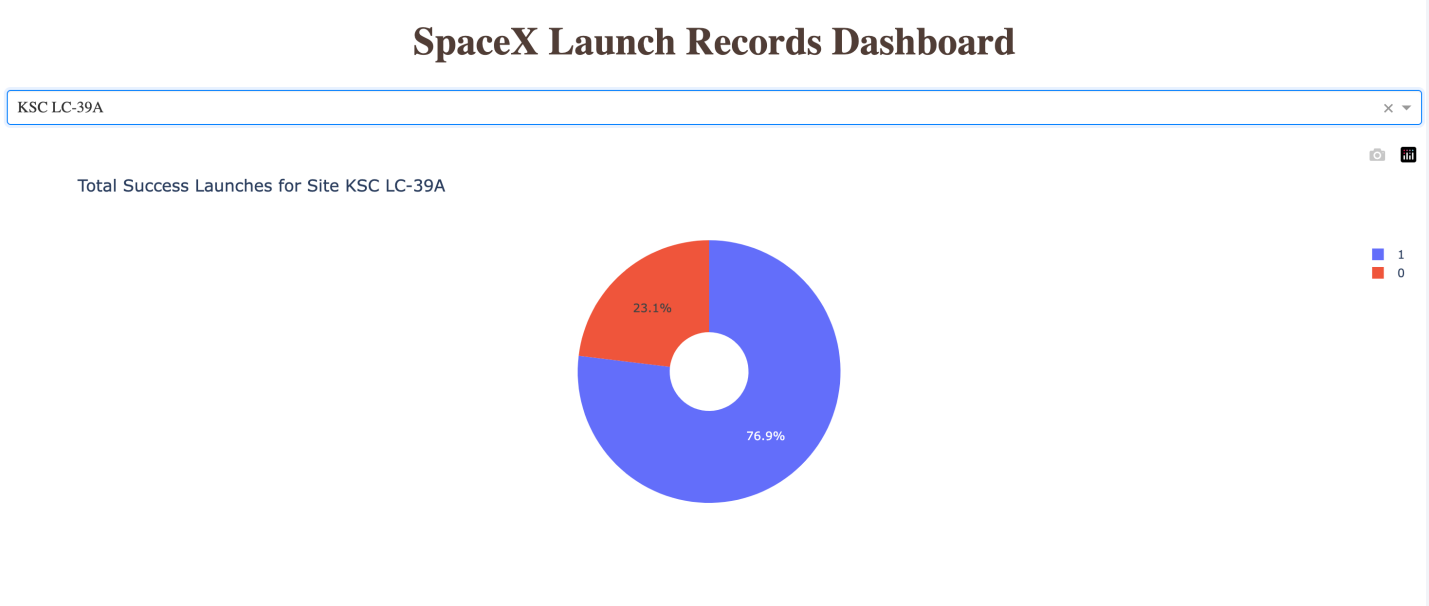

Section 4
Build a Dashboard with Plotly Dash

Launch Success for All Sites - Pie Chart Analysis



- The pie chart provides a visual representation of the success distribution across various SpaceX launch sites.
- Display a screenshot showcasing a pie chart illustrating the distribution of launch success counts for all SpaceX sites.
- **CCAFS SLC-40 :**CCAFS SLC-40 emerges as the top-performing site with a success rate of 41.7%.

Launch Success Ratio at KSC LC-39A Site



showcasing a pie chart representing the launch success ratio specifically for the KSC LC-39A SpaceX site.

1.Success Dominance at KSC LC-39A:

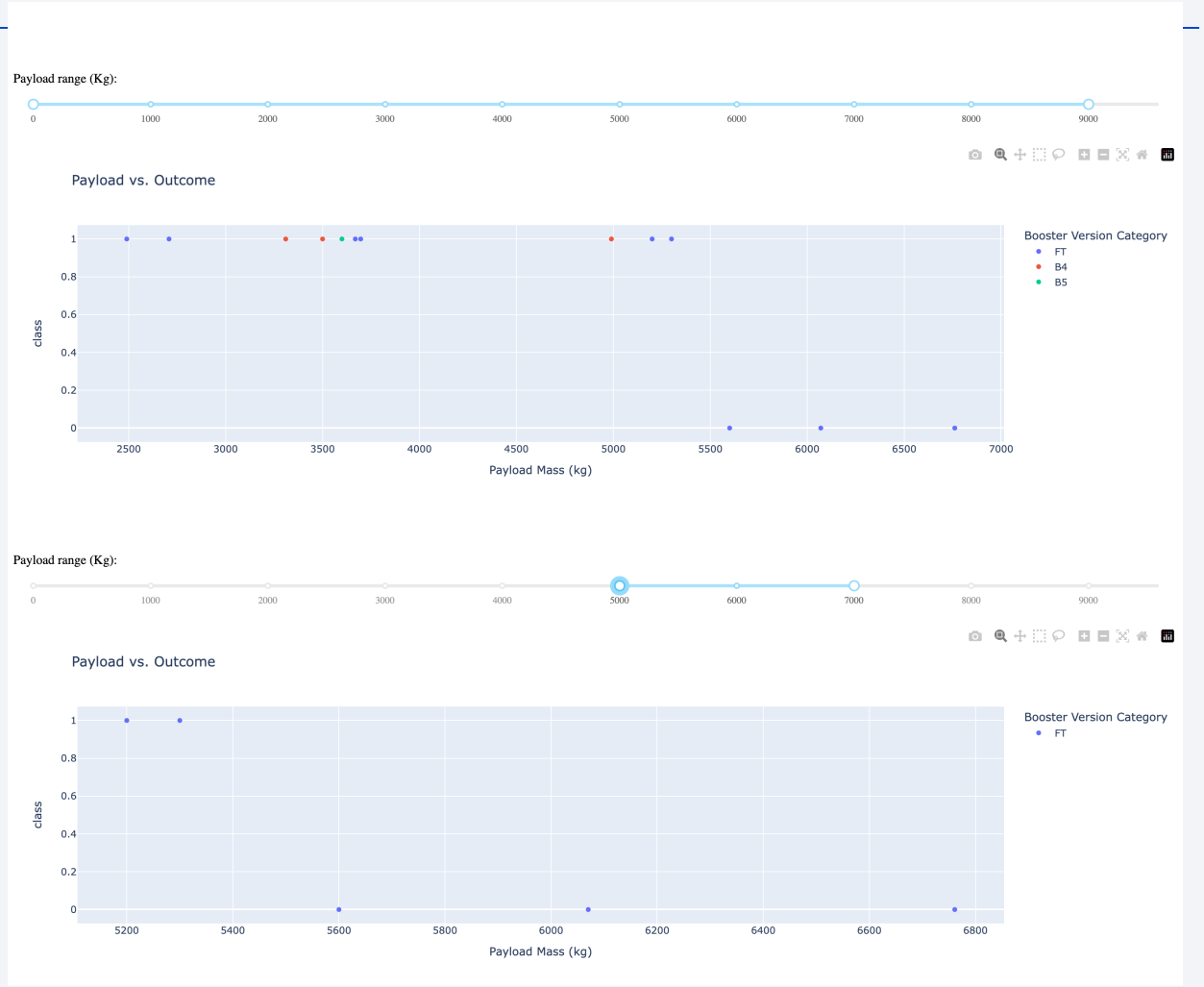
The KSC LC-39A site demonstrates a robust success rate, with a notable 76.9% of launches culminating in success.

2.Failure Rate at KSC LC-39A:

Despite the majority of successful launches, the site records a 23.1% failure rate, depicted in red on the pie chart.

Payload Analysis: Impact on Launch Outcomes Across SpaceX Sites

- the Payload vs. Launch Outcome scatter plot for all SpaceX sites, showcasing variations based on different payload ranges selected via the range slider.

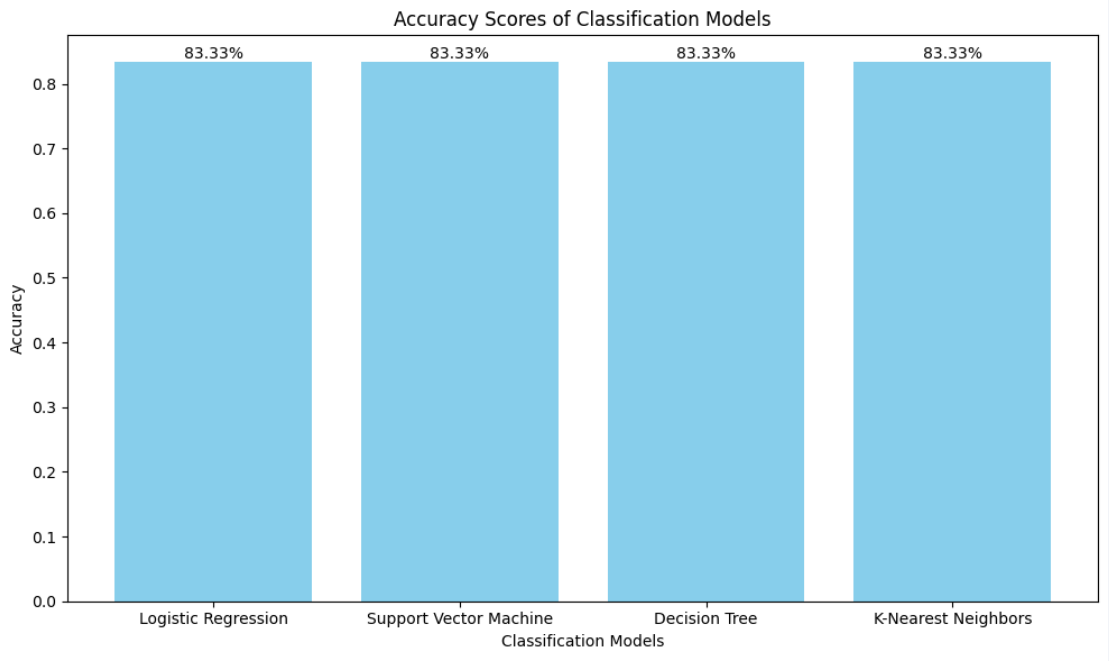


Section 5

Predictive Analysis (Classification)

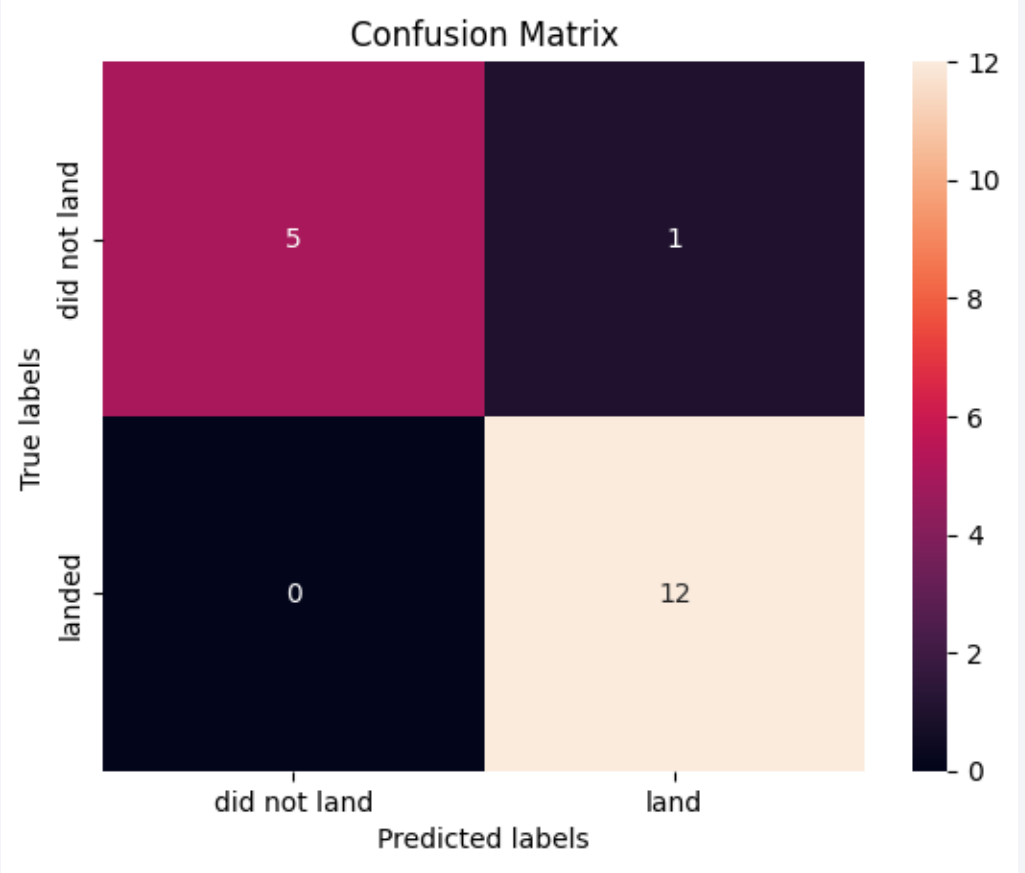
Classification Accuracy

- This slide emphasizes the parity in classification accuracy across the LR, SVM, DT, and KNN models, each achieving an accuracy rate of 83.33%, thereby affirming their consistent and reliable predictive capabilities in the evaluated scenarios.



Confusion Matrix

- The confusion matrix provides a comprehensive overview of the model's performance by presenting True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)



Conclusions

- Launch Site Success Rate: CCAFS SLC-40 stands out with a success rate of 41.7%, followed by CCAFS LC-40 at 29.2%, showcasing their significant contributions to SpaceX's successful launches.
- KSC LC-39A Analysis: The KSC LC-39A site records a dominant success rate of 76.9%, underscoring its operational efficiency in facilitating successful SpaceX missions.
- Payload Impact Analysis: Payload vs. Launch Outcome scatter plots reveal no distinct variation among LR, SVM, DT, and KNN models, all consistently achieving an accuracy rate of 83.33%.
- Model Performance: The confusion matrix of the best-performing model offers a detailed assessment, enabling stakeholders to evaluate precision, recall, accuracy, and F1-score metrics for comprehensive performance insights.

Appendix

- **Data Visualization with Folium**
- **SpaceX Launch Records**
- SpaceX database

Thank you!