IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

M. Romanenko
02-OCT-2023

Presentation Structure

Topic	Reference
Uploaded the URL of your GitHub repository including all the completed notebooks and Python files (1 pt)	GitHub
Uploaded your completed presentation in PDF format (1 pt)	
Executive Summary slide (1 pt)	Slide 4
Introduction slide (1 pt)	Slide 5
Data collection and data wrangling methodology related slides (1 pt)	Slide 8
EDA and interactive visual analytics methodology related slides (3 pts)	Slide 11
Predictive analysis methodology related slides (1 pt)	Slide 15
EDA with visualization results slides (6 pts)	Slide 18
EDA with SQL results slides (10 pts)	Slide 24
Interactive map with Folium results slides (3 pts)	Slide 35
Plotly Dash dashboard results slides (3 pts)	Slide 39
Predictive analysis (classification) results slides (6 pts)	Slide 43
Conclusion slide (1 pts)	Slide 45
Applied your creativity to improve the presentation beyond the template (1 pts)	
Displayed any innovative insights (1 pts)	

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Data Collection - Aggregated comprehensive data on Falcon rocket launches from multiple sources, including APIs and online databases.

Data Processing & Analysis

1. Exploratory Data Analysis (EDA)
   - SQL queries for data extraction, cleansing, and initial analysis.
   - Visualized key metrics like mission outcomes and launch sites.
2. Interactive Visual Analytics
   - Folium maps to geo-locate launch sites.
   - Plotly Dash for real-time, interactive dashboards.
3. Machine Learning
   - Built and tuned four classification models: Logistic Regression, SVM, Decision Trees, and K-NN.
   - Utilized 10-fold cross-validation for hyperparameter tuning via GridSearchCV.

Key Findings

- High accuracy achieved in Logistic model, but debugging required due to uniform accuracy across models.
- Uncovered insights on influential variables like launch sites, payload, and reused parts.

Introduction

Objective

To predict the success of first-stage landings of SpaceX's Falcon rockets, thereby estimating launch costs for competitive bidding.

Next Steps

- Investigate the issue of identical accuracies across models.
- Potentially include additional variables like weather conditions for more nuanced predictions.

Section 1

Methodology

Methodology Summary

Data Collection

- Data collected for multiple SpaceX launches, including features like payload mass, orbit type, launch site, and outcome.
- Data collected from SpaceX's rocket launches, sourced from online databases and APIs.

Data Processing

1. Exploratory Data Analysis (EDA)
    - Utilized SQL for data querying and aggregation.
    - Visualized data using histograms, bar charts and scatter plots to uncover patterns and insights.
2. Interactive Visual Analytics
    - Utilized Folium for geospatial mapping of launch sites.
    - Developed a Plotly Dash app for real-time analytics, offering interactive data filtering.
3. Predictive Analysis
    - Employed classification models like Logistic Regression, SVM, Decision Trees, and K-NN.
    - Tuned models using GridSearchCV for hyperparameter optimization.

Data Collection – SpaceX API

- The code shows how to make API calls to SpaceX endpoints.
- Markdown cells provide explanations about SpaceX, REST calls, and data collection.
- [GitHub URL of the completed SpaceX API calls notebook]

Flowchart:

1. Initialize REST API Call
2. Receive API Response
3. Parse JSON Data
4. Populate DataFrame
5. Save Data

Data Collection - Scraping

- The code initializes and uses BeautifulSoup for parsing HTML.
- Text content is extracted from the HTML elements.
- Various HTML elements are identified and extracted.
- Markdown cells provide additional explanations about web scraping.
- GitHub URL of the completed web scraping notebook

Flowchart:

1. Initialize BeautifulSoup
2. Fetch HTML Page
3. Parse HTML
4. Extract Text Content
5. Populate DataFrame
6. Save Data

Data Wrangling

- Objective: Clean and Transform SpaceX Launch Data for Analysis
- Key Operations: Handling missing values, data type conversion, feature extraction
- [GitHub URL of the completed data wrangling notebook]

Flowchart:

1. Load Data
2. Identify Missing Values
3. Fill or Drop Missing Values
4. Data Type Conversion
5. Feature Extraction
6. Data Normalization
7. Save Data

EDA with Data Visualization

- Histogram of Launch Years: To show the frequency of SpaceX launches over time.
- Boxplot of Payload Mass by Year: To visualize the range and spread of payload masses across different years.
- Scatter Plot of Payload Mass vs. Orbit: To understand how the payload mass varies with the type of orbit.
- Heatmap of Mission Outcome vs. Orbit Type: To identify any patterns between the mission outcome and the type of orbit.
- Scatter Plot of Launch Site vs. Mission Outcome: To see if certain launch sites have higher success rates.
- GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

- **Count of Launches by Site** to identify which launch sites are most frequently used.
- **Success Rate by Launch Site** to calculate the success rate for each launch site.
- **Success Rate by Year** to understand how the success rate has evolved over time.
- **Payload Mass Statistics by Orbit Type** to summarize the payload masses for different orbit types.
- **Missions with Heaviest Payloads** to list the missions that have carried the heaviest payloads.
- **Success Rate by Orbit Type** to calculate the success rate for each orbit type.
- **Average Payload Mass by Customer** to find out which customers typically require heavier payloads.
- GitHub URL of the completed EDA with SQL notebook

Build an Interactive Map with Folium

Map Objects Created:

1. Launch Site Markers:
    - Type: Marker
    - Objective: To pinpoint the location of each launch site.
2. Launch Success Circles:
    - Type: Circle
    - Objective: To visualize the success rate of each launch site.
3. Distance Lines to Nearby Railways, Highways and Cities:
    - Type: PolyLine
    - Objective: To show the distance from each launch site to the nearest object.

Why These Objects Were Added:

- Markers: To make it easy to identify each launch site.
- Circles: To offer a quick visual cue on the performance of each launch site.
- Lines: To give context on how far each launch site is from closest railways, highways and cities
- GitHub URL of the completed interactive map with Folium map

Build a Dashboard with Plotly Dash

Added:

- A dropdown list to enable Launch Site selection
- A pie chart to show the total successful launches count for all sites
- A slider to select payload range
- A scatter chart to show the correlation between payload and launch success
- *A callback function for `site-dropdown` as input, `success-pie-chart` as output*
- *A callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output*
- GitHub URL of the completed Plotly Dash lab
- GitHub URL of the Plotly Dash screenshot

Predictive Analysis (Classification)

Key Phrases:

- Data Preprocessing
- Train-Test Split
- Hyperparameter Tuning
- Model Fitting
- Model Evaluation
- Confusion Matrix
- Select Best Model

I employed various machine learning algorithms such as Logistic Regression, SVM, Decision Trees, and KNN. These models were tuned using GridSearchCV for hyperparameter optimization. The models' performance was evaluated using accuracy as the metric and visually inspected through confusion matrices. Finally, the model with the highest accuracy was selected as the best performing model.

Flowchart:

1. Standardize the features using sklearn's preprocessing.
2. Divide the dataset into training and test sets.
3. Use GridSearchCV to find the best parameters for different models.
4. Fit the models using the best parameters obtained from GridSearchCV.
5. Evaluate the performance of the models on the test dataset.
6. Plot the confusion matrix to understand the performance of the models.
7. Compare the accuracy of the different models and select the best one.

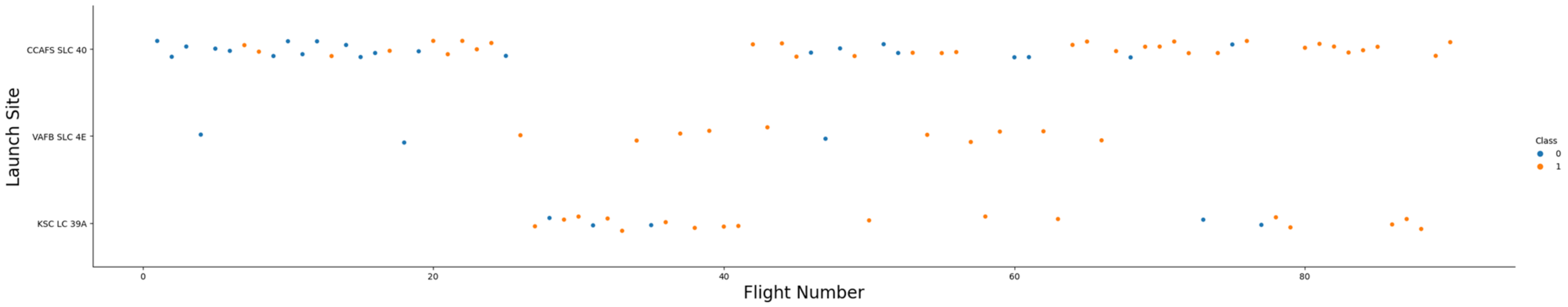
Results

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

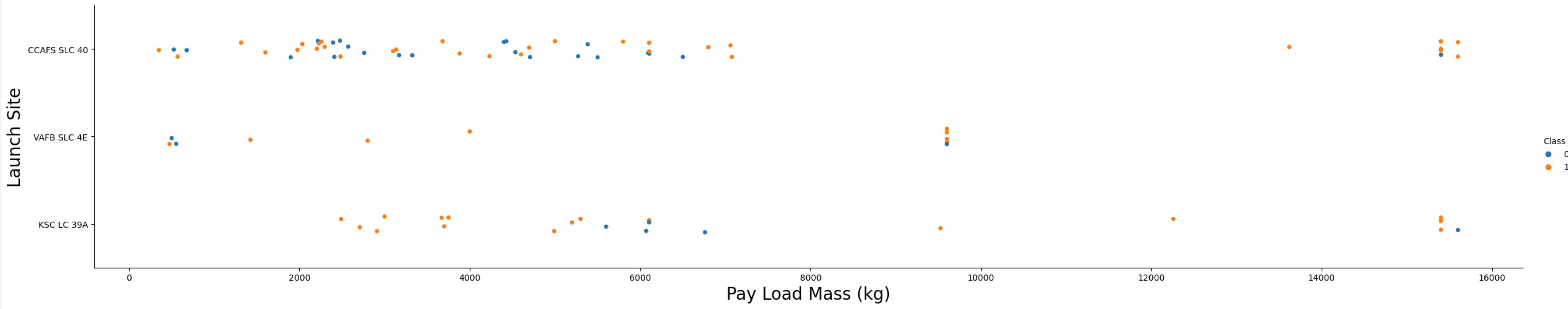
Section 2

Insights drawn from EDA

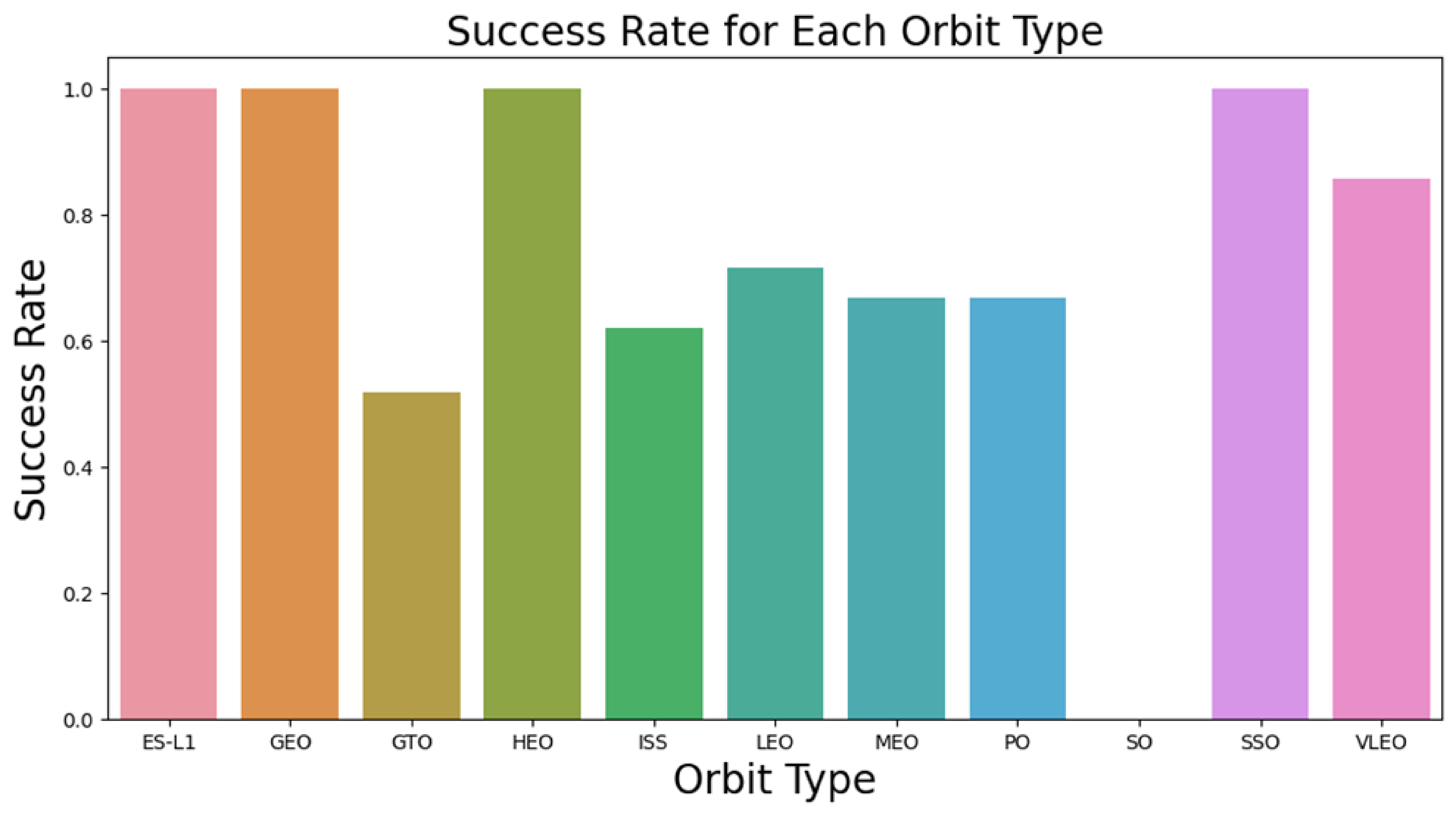
Flight Number vs. Launch Site



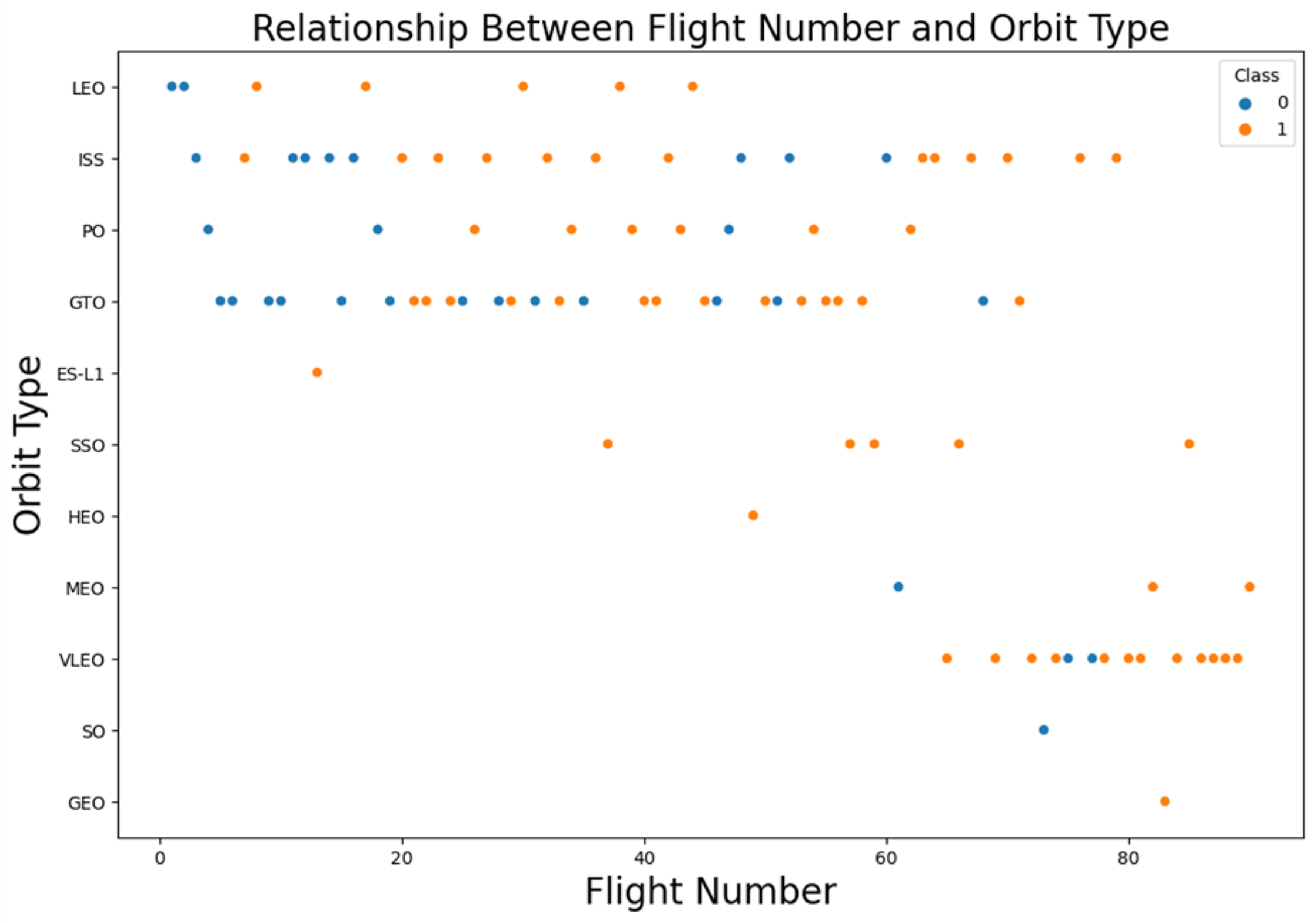
Payload vs. Launch Site



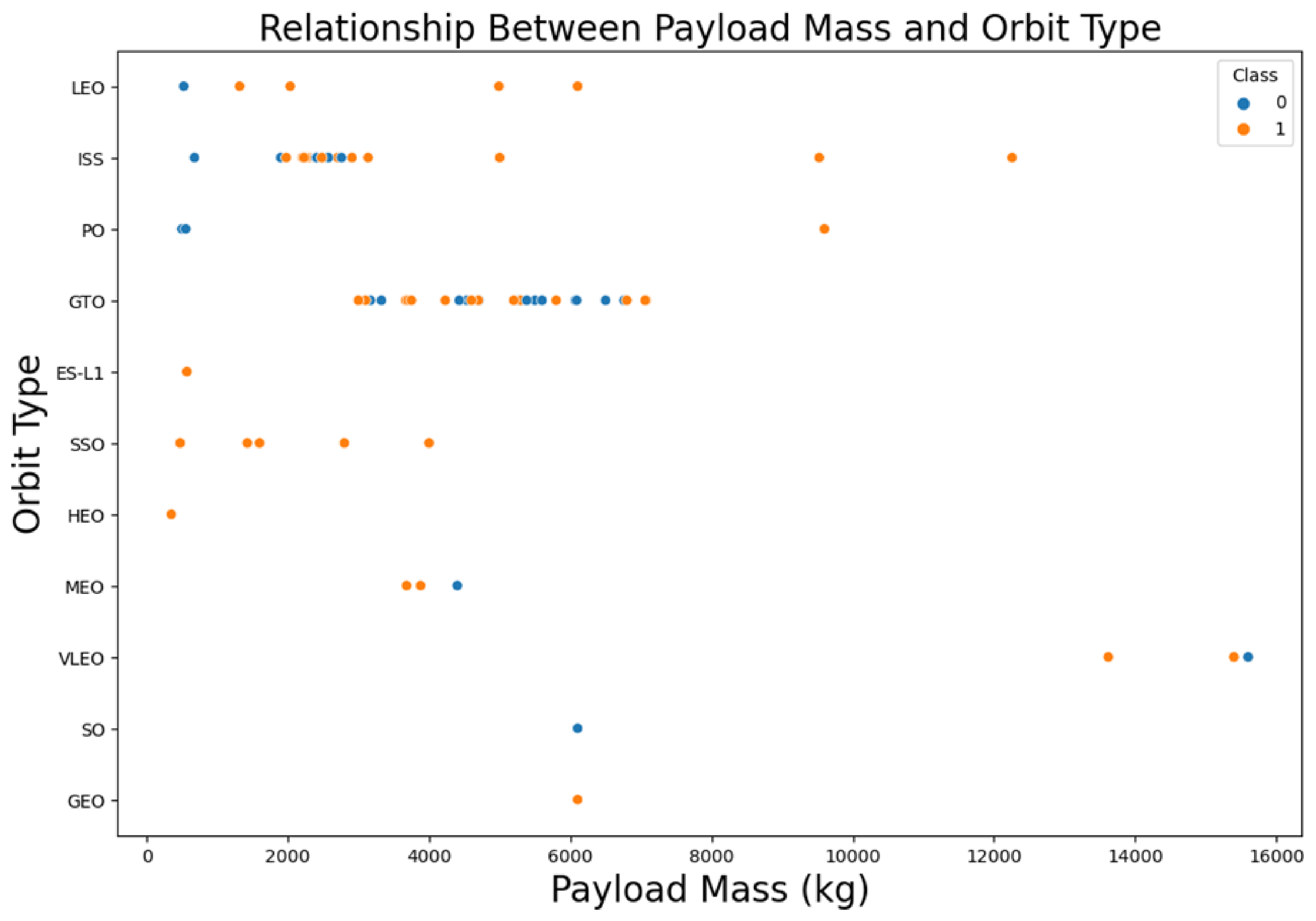
Success Rate vs. Orbit Type



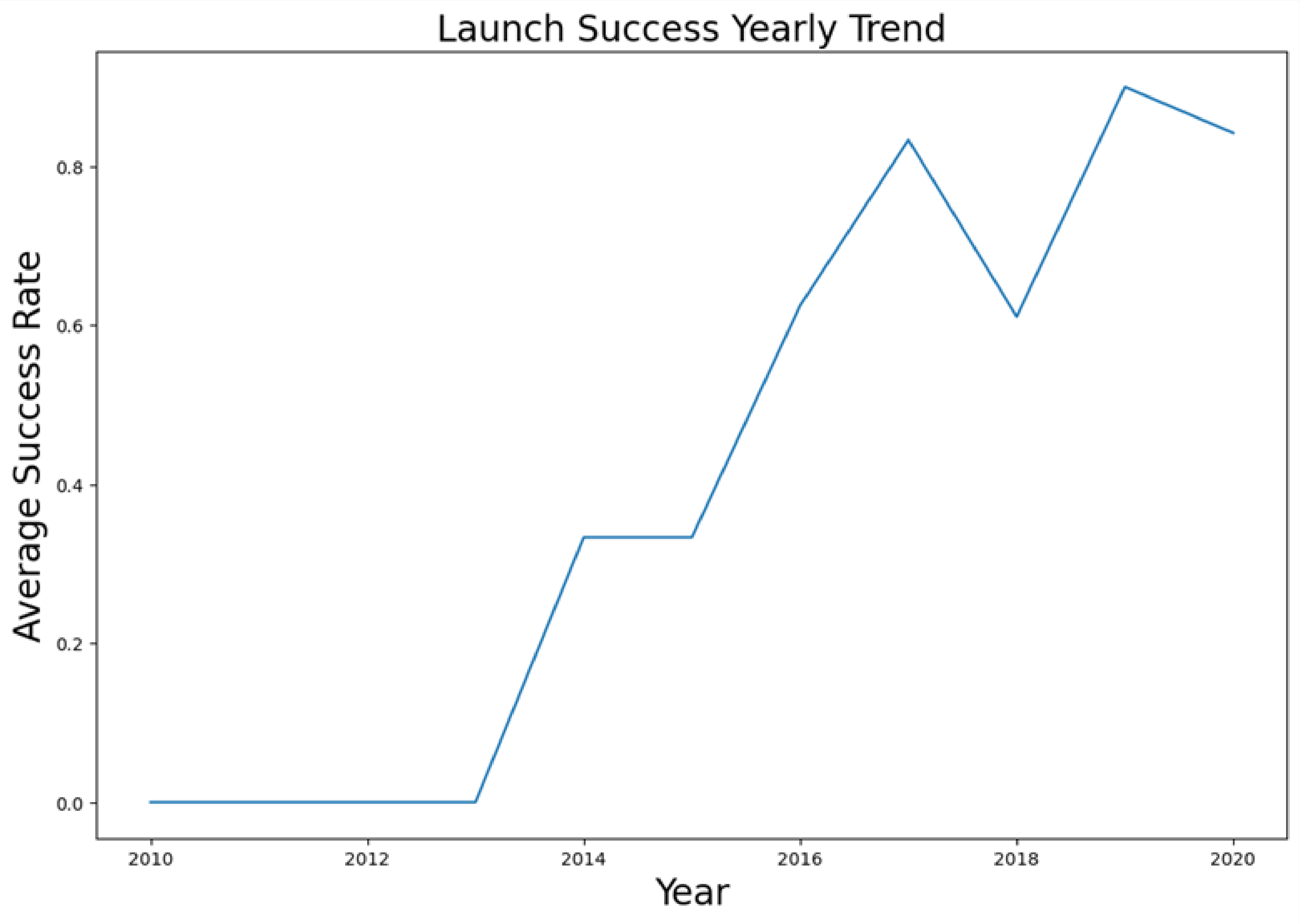
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

Task 1

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

```
 * 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'

Task 2

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

```
%sql SELECT * FROM SPACEXTABLE 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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Task 3

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)%';
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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 DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 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

Task 7

List the total number of successful and failure mission outcomes

```
%sql ALTER TABLE SPACEXTABLE ADD COLUMN Simplified_Outcome TEXT;
%sql UPDATE SPACEXTABLE SET "Simplified_Outcome" = CASE WHEN "Mission_Outcome" LIKE 'Success%' THEN 'Success' ELSE 'Failure' END;
```

```
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
101 rows affected.
[]
```

```
%sql SELECT Simplified_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Simplified_Outcome;
```

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

Simplified_Outcome	COUNT(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

Task 8

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

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

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

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

Task 9

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT Landing_Outcome, substr(Date, 6, 2) as Month, substr(Date, 1, 4) as Year, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date, 1, 4)='2015' AND Landing_Outcome LIKE 'Failure
```

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

Landing_Outcome	Month	Year	Booster_Version	Launch_Site
Failure (drone ship)	10	2015	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	04	2015	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

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(*) as Outcome_Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Outcome_Count DESC;
```

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

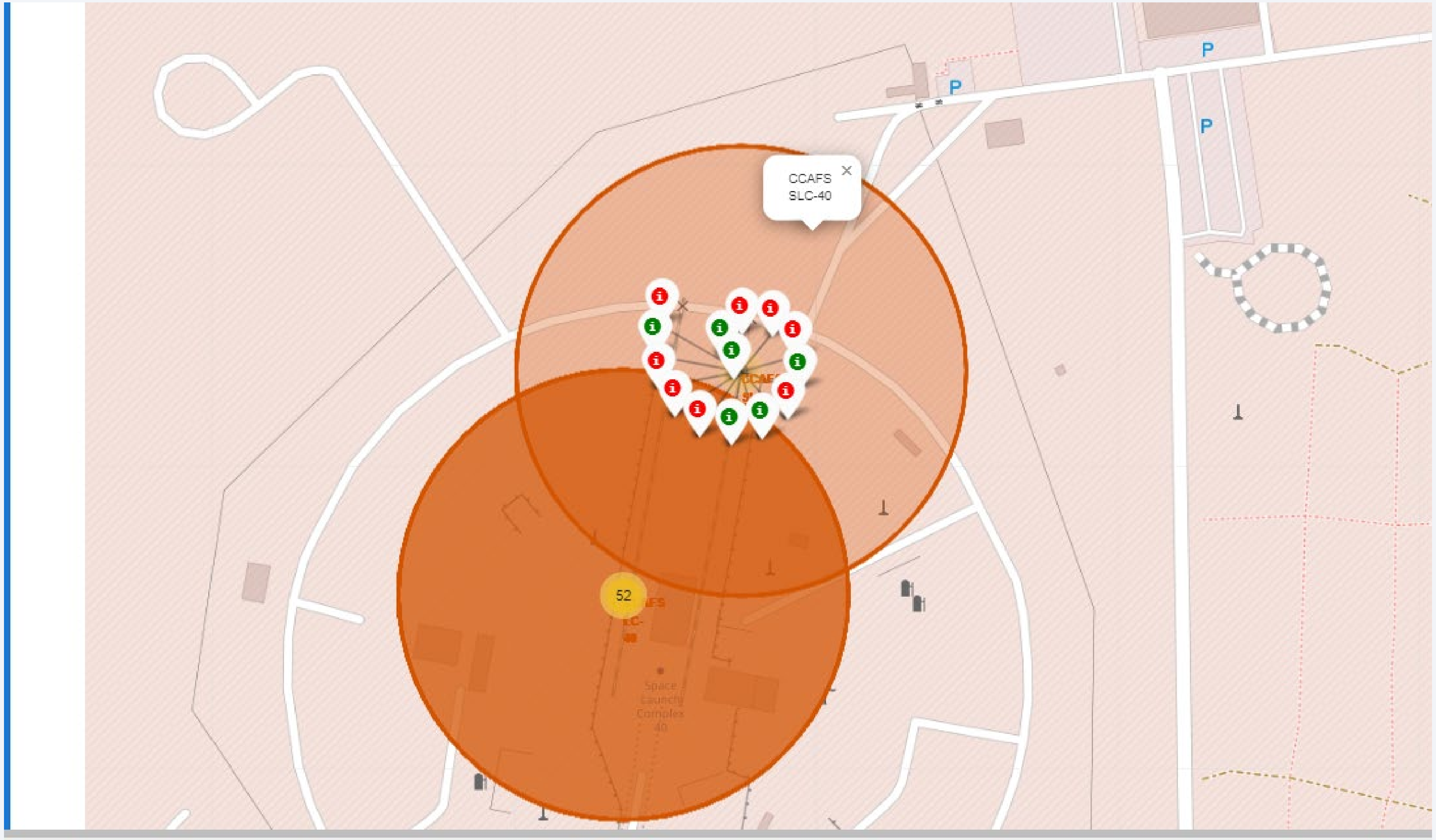
Landing_Outcome	Outcome_Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

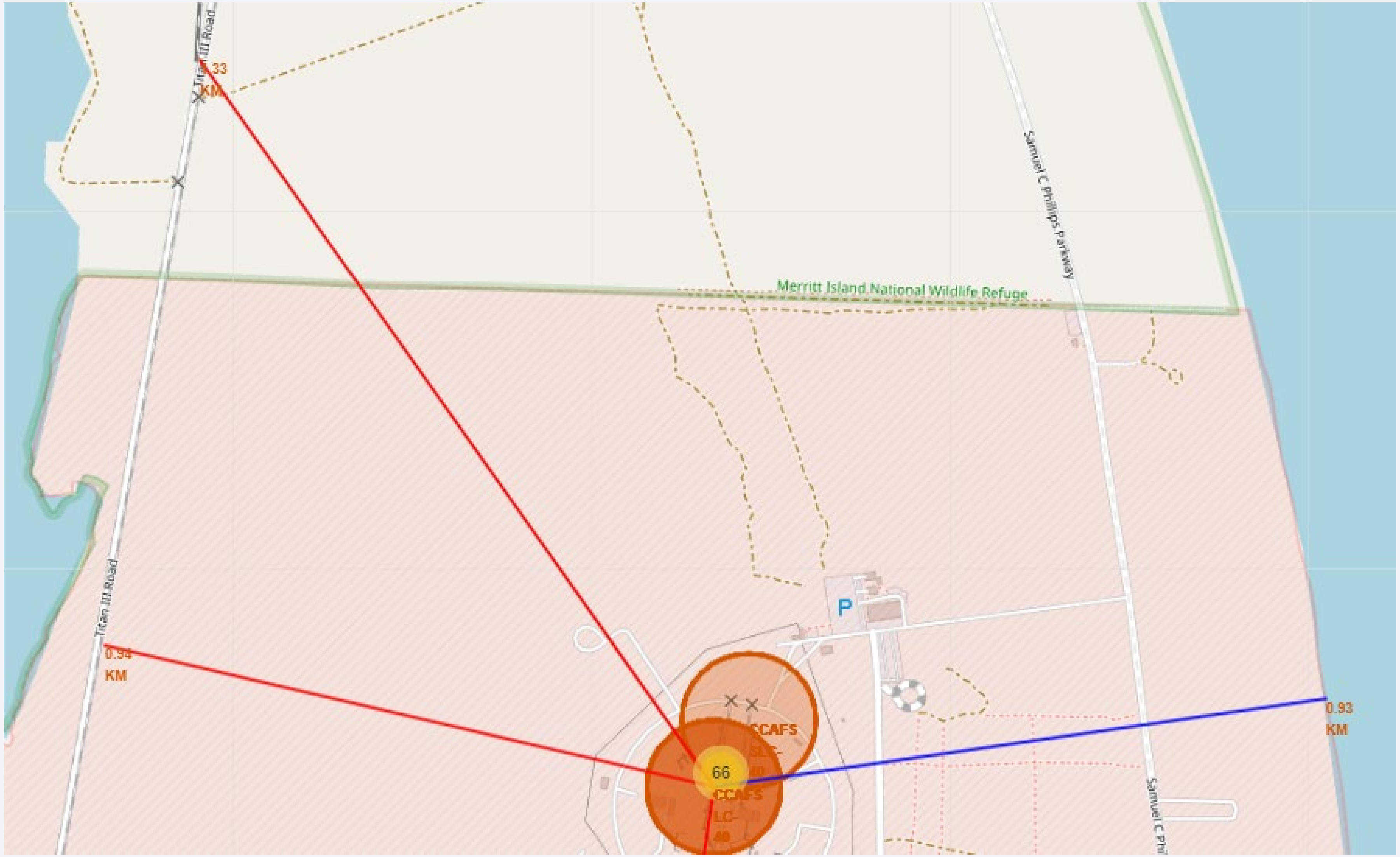
Launch Sites Proximities Analysis

Launch Sites Locations

Launch Outcomes Map CCAFS SLC-40



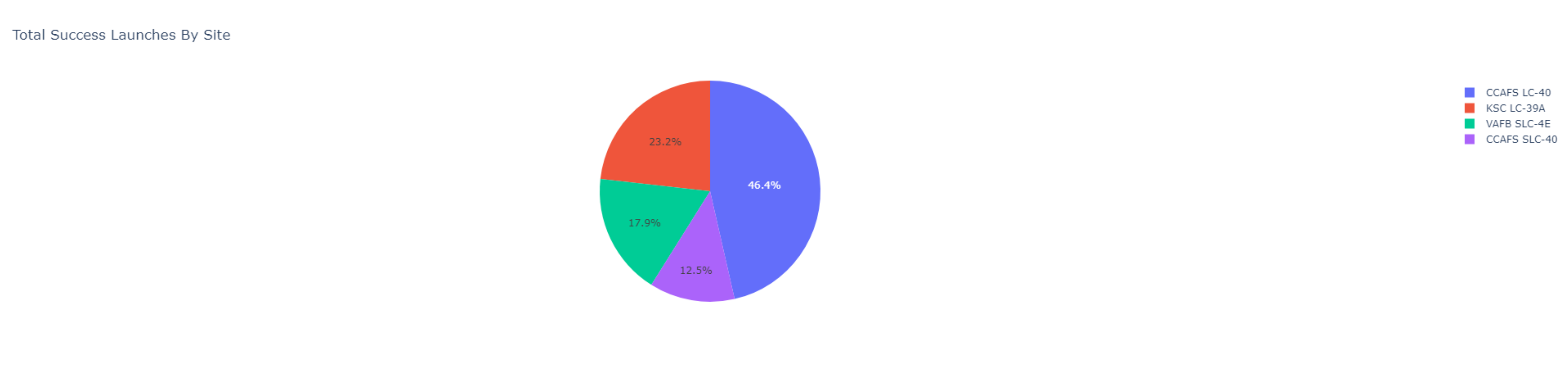
Selected Launch Site To Its Proximities Such As Railway, Highway, Coastline



Section 4

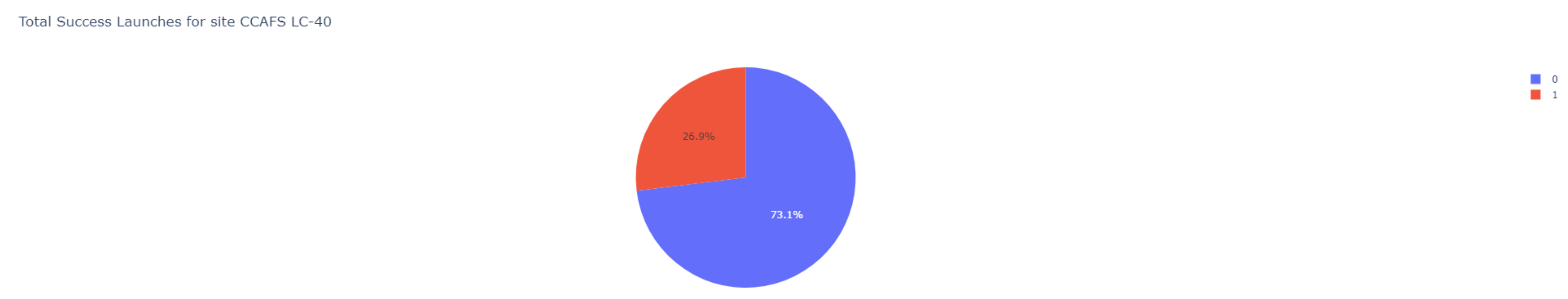
Build a Dashboard with Plotly Dash

SpaceX_Dash



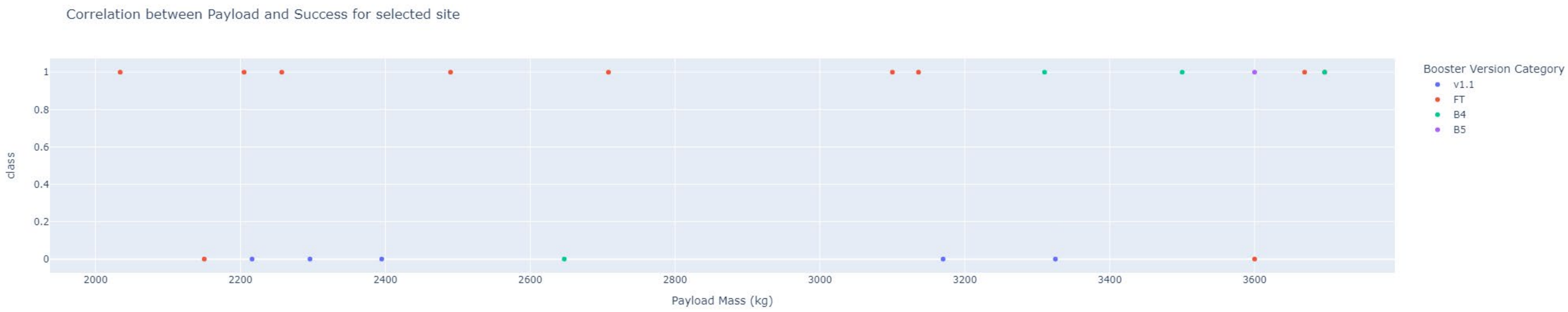
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>



- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

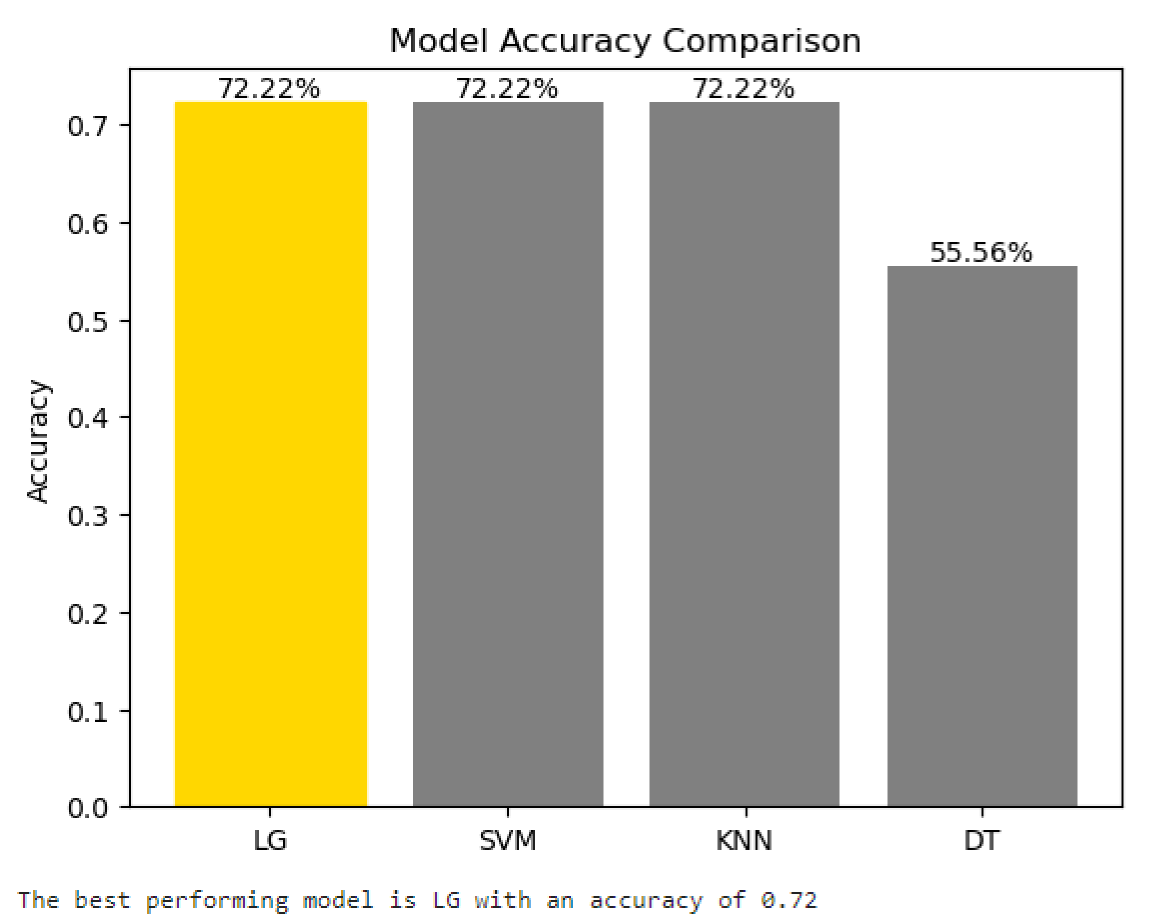


- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

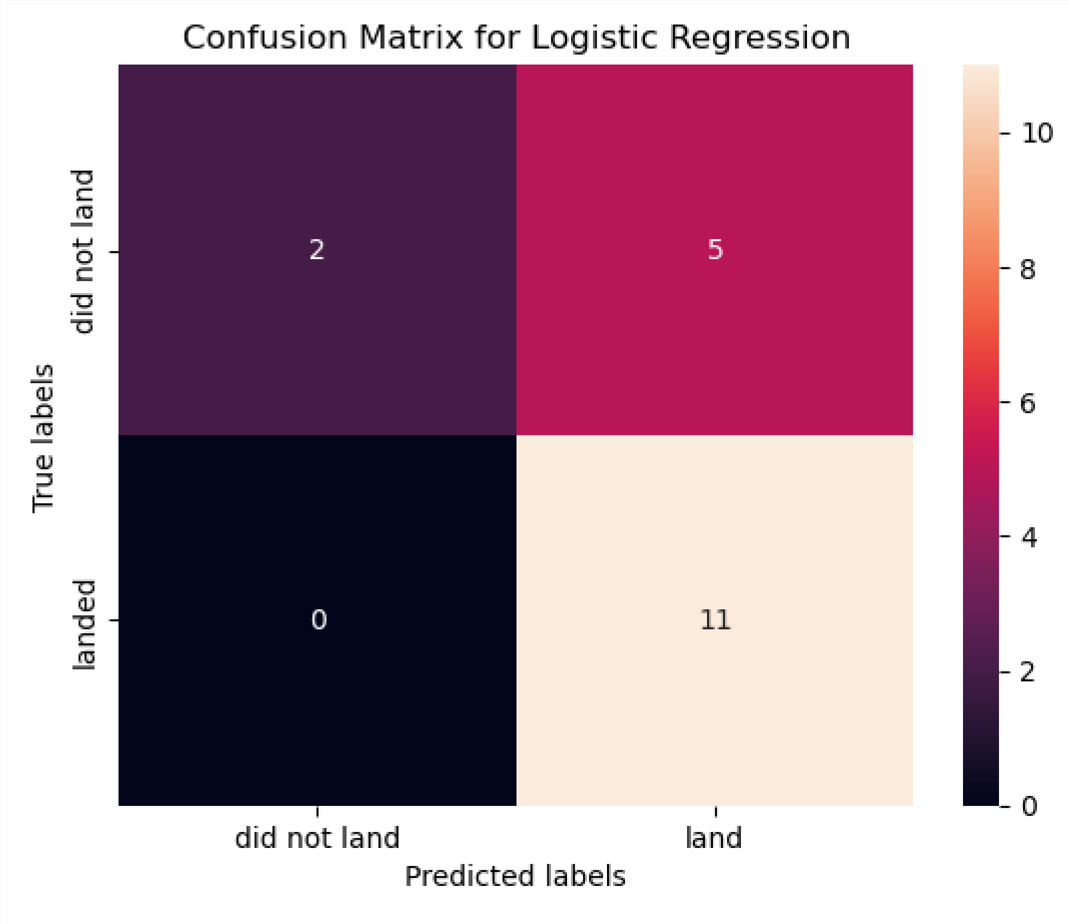
Predictive Analysis (Classification)

Classification Accuracy



- Logistic model has the best accuracy

Confusion Matrix



- False positive is still an issue with this model

Conclusions

- **Data Integrity and Quality**: The dataset was comprehensive and allowed for deep analysis. However, an issue with uniform accuracy across different machine learning models suggests potential data or modeling pitfalls that need investigation.
- **EDA & Visual Analytics**: The exploratory data analysis and interactive visualizations provided valuable insights into launch success factors, such as launch sites and payload mass.
- **Model Performance**: Logistic model showed the highest cross-validated accuracy during the hyperparameter tuning phase. However, the identical test accuracies across SVM and KNN models are confusing and indicate a possible issue that requires further examination.
- **Feature Importance**: Factors like launch site, payload mass, and the number of times parts were reused have been identified as significant predictors for the success of the first-stage landing.
- **Operational Relevance**: Once the model is validated and fine-tuned, it could become a valuable tool for predicting launch costs, thereby aiding in competitive bidding scenarios.
- **Future Scope**: Additional data such as weather conditions, engineering improvements, and more granular details about each flight could further improve the model.
- **Real-time Analytics**: The use of Plotly Dash and Folium maps allows for real-time analytics, which can be beneficial for quick decision-making during actual launches.
- **Business Impact**: Understanding the probabilities of first-stage landing success can be a game-changer in pricing strategies for SpaceX and any competitors.

Thank you!