IBM Data Science Capstone Project

German Fuentes

31/01/2024

OUTLINE

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

EXECUTIVE SUMMARY

- **Methodologies:**
- **Data Collection:** Gathered relevant datasets for analysis.
- **Data Wrangling:** Cleaned and processed raw data for quality.
- **EDA:** Explored data patterns for insights.
- **Visual Analytics:** Used interactive tools for data representation.
- **Predictive Analysis:** Applied classification models for forecasting.
- **Results:**
- **EDA Insights:** Revealed data structure and key characteristics.
- **Geospatial Analytics:** Mapped geographical patterns for enhanced understanding.
- **Interactive Dashboard:** User-friendly interface for dynamic data exploration.
- **Predictive Classification:** Models for informed decision-making.
- The project successfully executed each step, providing actionable insights through data exploration, visualization, and predictive analysis.

INTRODUCTION

This project focuses on predicting the successful landing of the SpaceX Falcon 9 first stage, a pivotal factor in the company's cost-efficient rocket launches priced at $62 million.

The ability to forecast landing outcomes enables an estimation of launch costs, aiding in the assessment of competitiveness against other providers charging significantly higher prices. By deploying predictive analysis, decision-makers can strategically evaluate whether alternate companies should bid against SpaceX for rocket launches.

Ultimately, this initiative aims to enhance decision-making in the space industry, offering valuable insights for cost-effective and reliable launch options.

METHODOLOGY

Data Collection Approach aggregated data from the SpaceX public API and SpaceX Wikipedia page Conducted data wrangling procedures.

Categorized successful and unsuccessful landings through classification.

Executed exploratory data analysis (EDA) employing visualization techniques and SQL queries.

Implemented interactive visual analytics using Folium and Plotly Dash.

Conducted predictive analysis using classification models Optimized models through parameter tuning using GridSearchCV

DATA COLLECTION ANALYSIS

Engaged in data collection through a dual approach, utilizing API requests from SpaceX's public API and web scraping data from a table within SpaceX's Wikipedia entry.

SpaceX API Data Fields: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
Wikipedia Web Scraped Data Fields: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Wrangling

- **Generate Training Labels:** Established a training label based on landing outcomes, designating 'successful' as 1 and 'failure' as 0 within the 'Outcome' column, which comprises 'Mission Outcome' and 'Landing Location.' Introduced a new training label column named 'class,' assigning a value of 1 when 'Mission Outcome' is true, and 0 otherwise. Value mapping includes:
- True ASDS, True RTLS, & True Ocean, set to 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS, set to 0

Exploratory Data Analysis (EDA)

Conducted EDA on key variables including Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. Utilized various plots for analysis, such as Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs. Orbit, and Success Yearly Trend.

Implemented scatter plots, line charts, and bar plots to discern relationships between variables, facilitating the identification of significant patterns for incorporation into the machine learning model training process.

Data Integration and Analysis

Imported the dataset into an IBM DB2 Database for efficient storage and retrieval. Utilized SQL Python integration to execute queries, extracting valuable insights to enhance understanding of the dataset.

Formulated queries to retrieve information on launch site names, mission outcomes, payload sizes of customers, booster versions, and landing outcomes. This approach facilitated a comprehensive exploration of the dataset, aiding in the extraction of pertinent details for further analysis and decision-making.

Folium Mapping Analysis

Implemented Folium maps to visually represent Launch Sites, successful and unsuccessful landings, along with proximity markers for key locations: Railway, Highway, Coast, and City.

This visualization strategy provides insights onto the rationale behind the selection of launch site locations, offering a comprehensive understanding of their proximity to crucial infrastructure.

The mapping also effectively illustrates the distribution of successful landings in relation to geographical features, enhancing the interpretation of spatial patterns and informing future launch site decisions.

Dashboard Overview

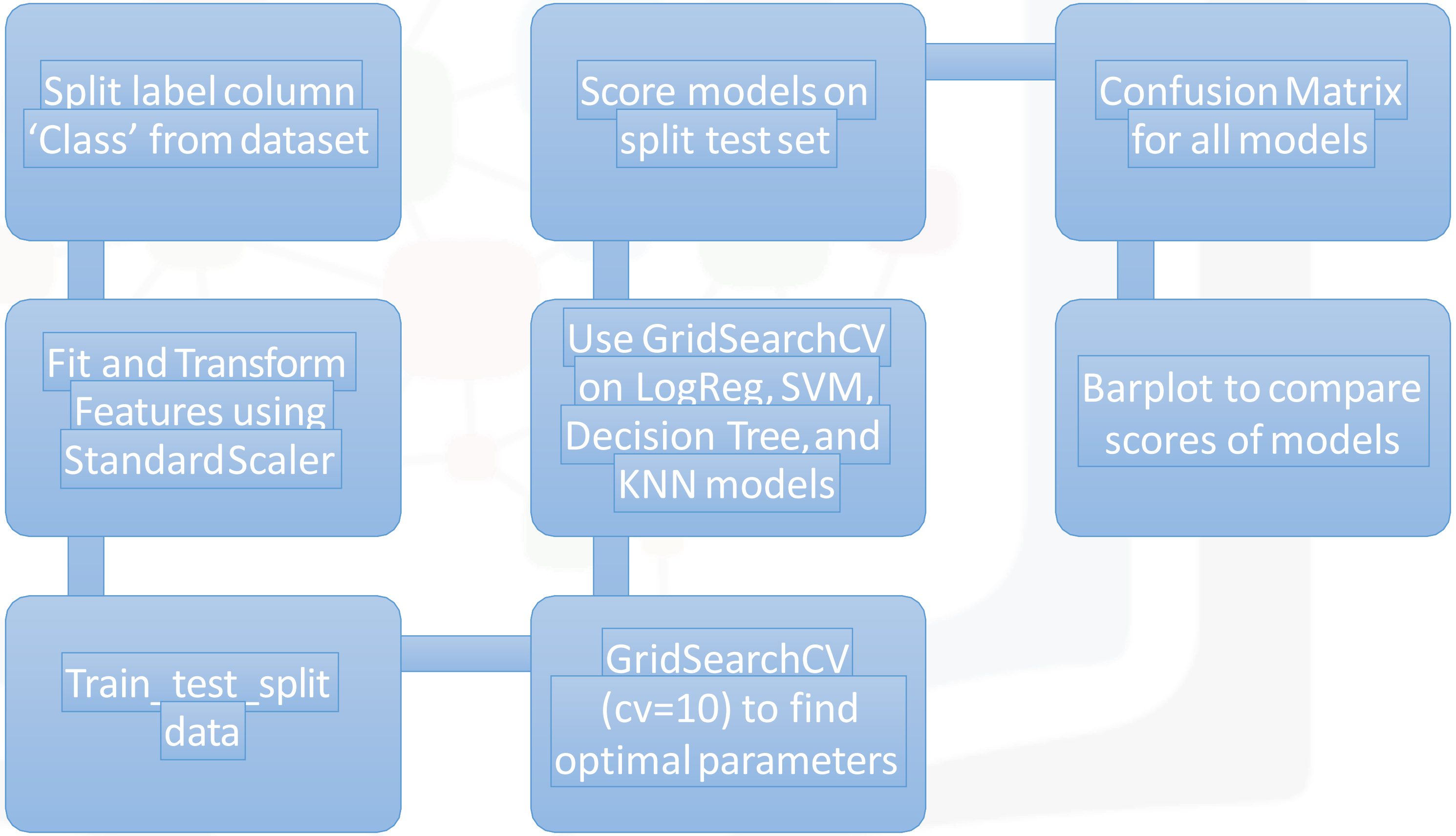
Incorporated a dynamic dashboard featuring a pie chart and a scatter plot to enhance data visualization.

The pie chart allows users to toggle between displaying the distribution of successful landings across all launch sites and exploring individual launch site success rates.

The scatter plot is designed with two inputs: the option to view data for all sites or an individual site, and a payload mass slider ranging from 0 to 10000 kg.

The pie chart serves as an intuitive tool for visualizing launch site success rates, while the scatter plot offers insights into the variability of success across launch sites, payload masses, and booster version categories. This interactive dashboard facilitates a comprehensive exploration of key factors influencing successful landings.

Predictive Analysis



Split label column ‘Class’ from dataset

Fit and Transform Features using StandardScaler

Train_test_split data

GridSearchCV (cv=10) to find optimal parameters

Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models on split test set

Confusion Matrix for all models

Barplot to compare scores of models

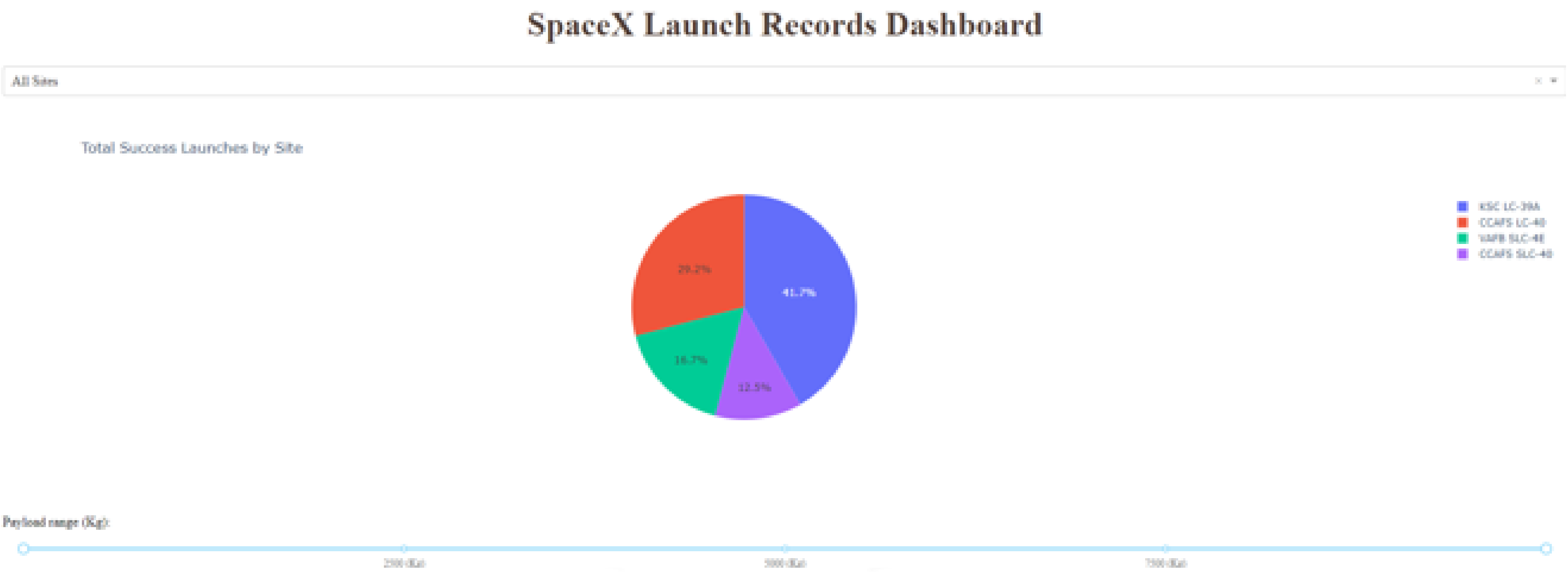
Results

1) Exploratory Data Analysis
2) Interactive Analysis
3) Predictive Analysis (Classification)

Providing a sneak peek into the upcoming Plotly dashboard, the subsequent slides will showcase the results of Exploratory Data Analysis (EDA) with visualizations, EDA utilizing SQL, an Interactive Map with Folium, and culminate with the presentation of our model results, boasting an impressive accuracy rate of approximately 83%.

These sections collectively illustrate the comprehensive journey from data exploration to predictive modeling, highlighting key insights gained and the successful application of analytical methodologies throughout the project.

Results



SpaceX Launch Records Dashboard

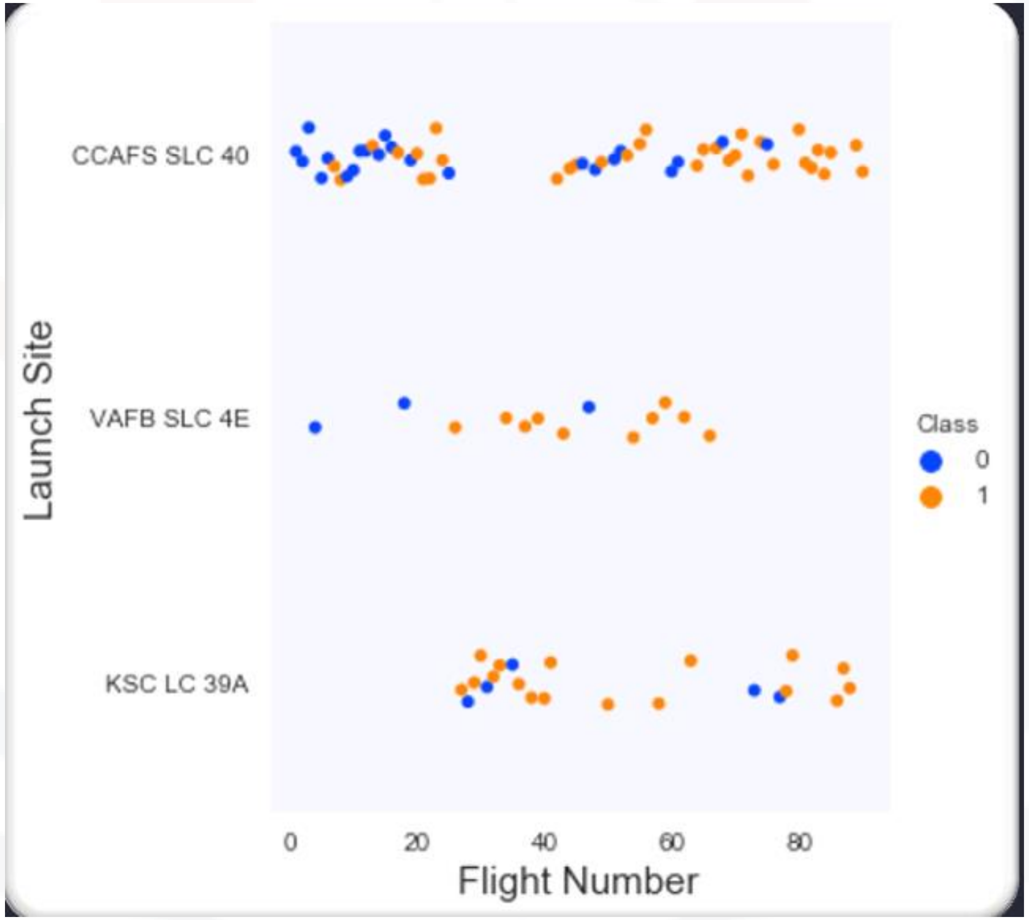
All Sites

Total Success Launches by Site

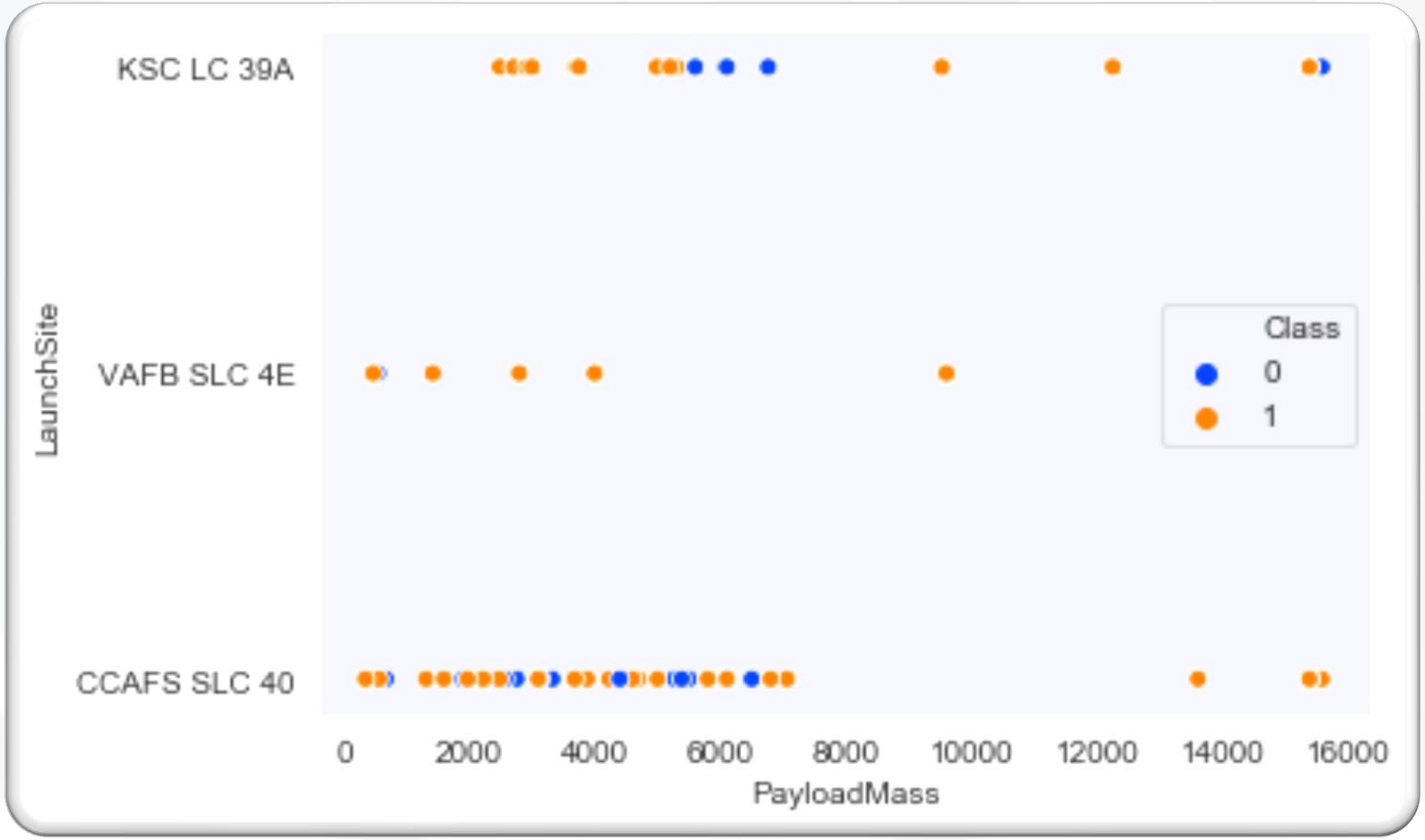
Payload range (Kg):

EDA - WITH VISUALIZATION

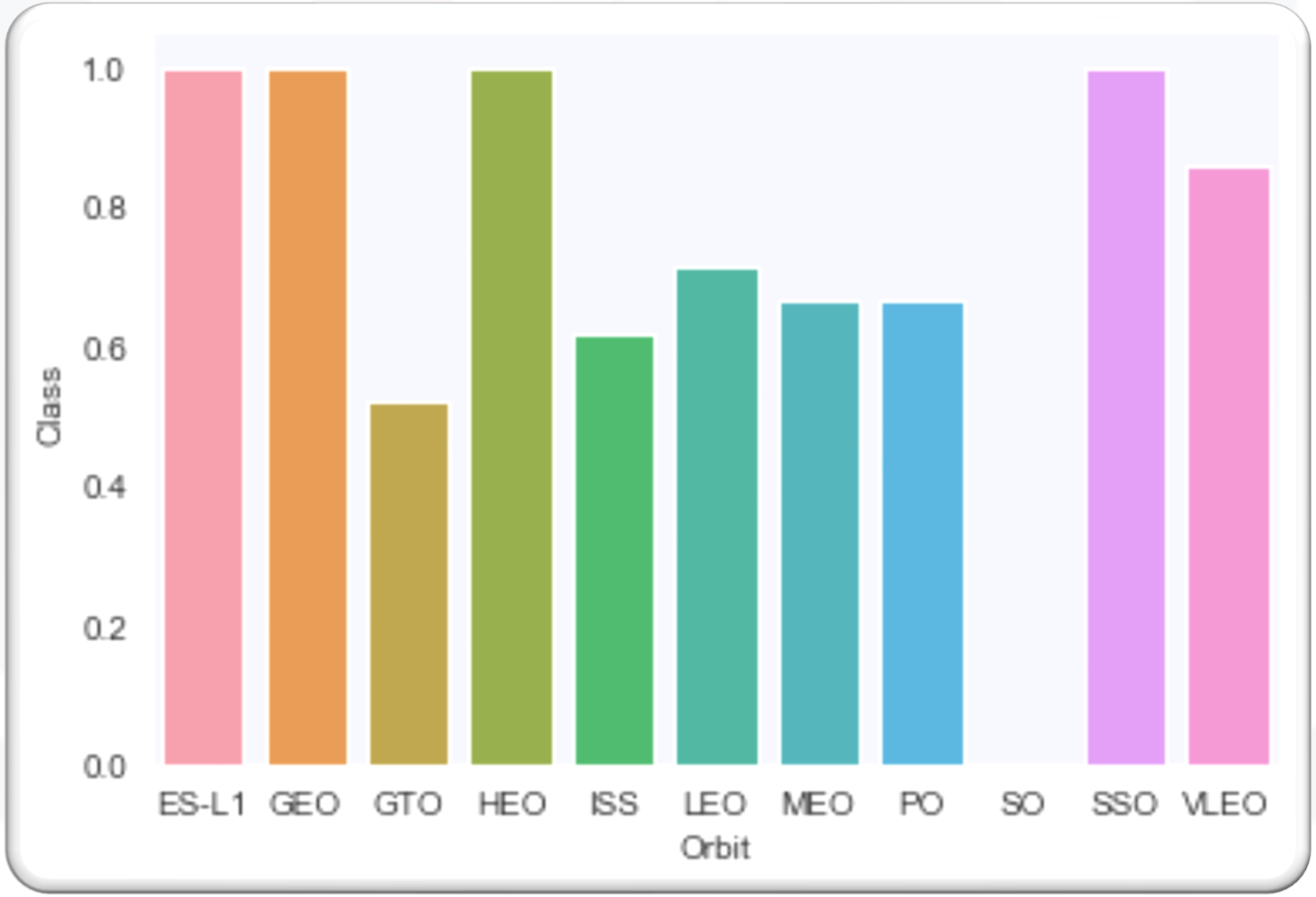
Flight Number vs. Launch Site



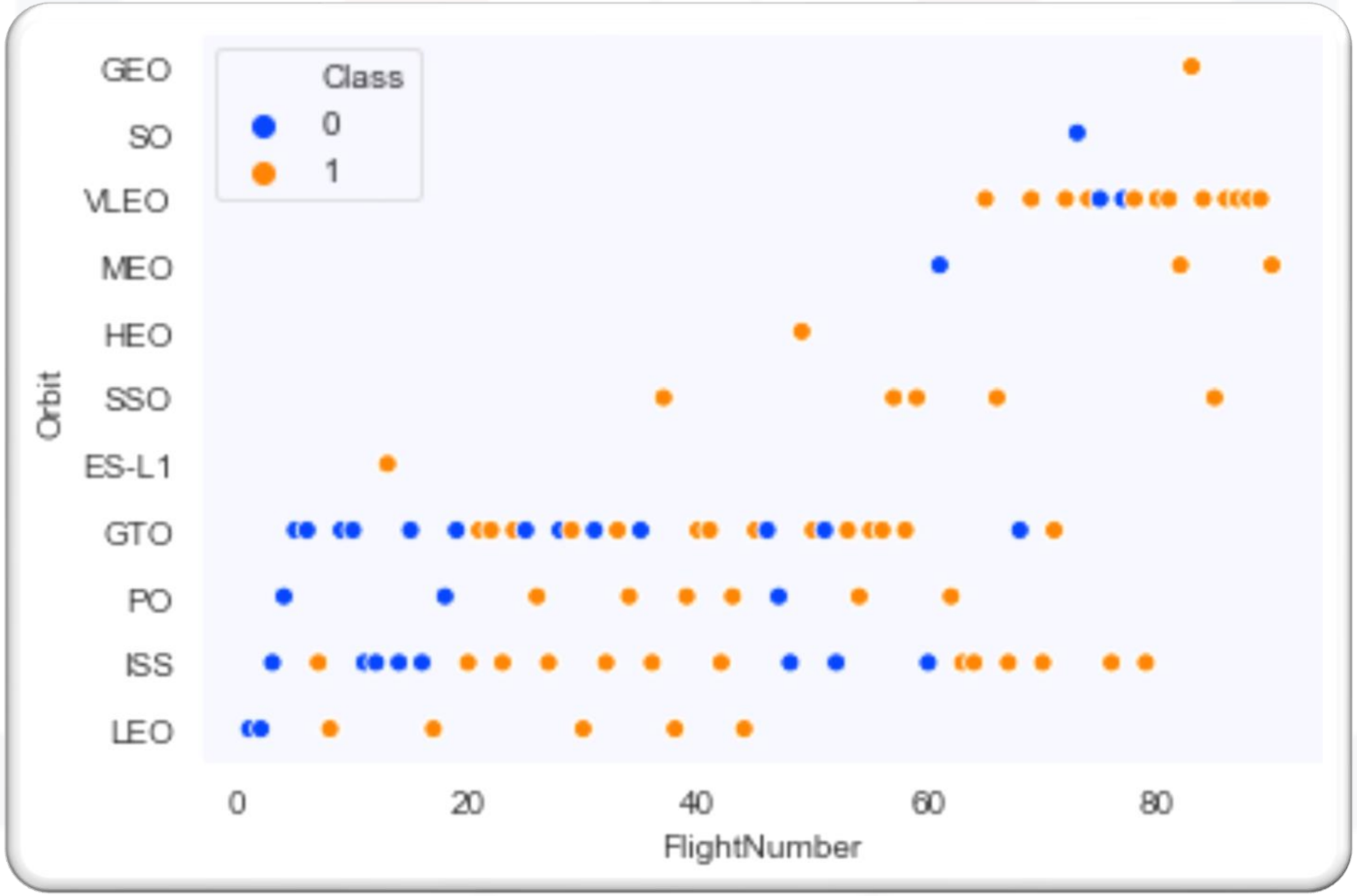
Payload vs. Launch Site



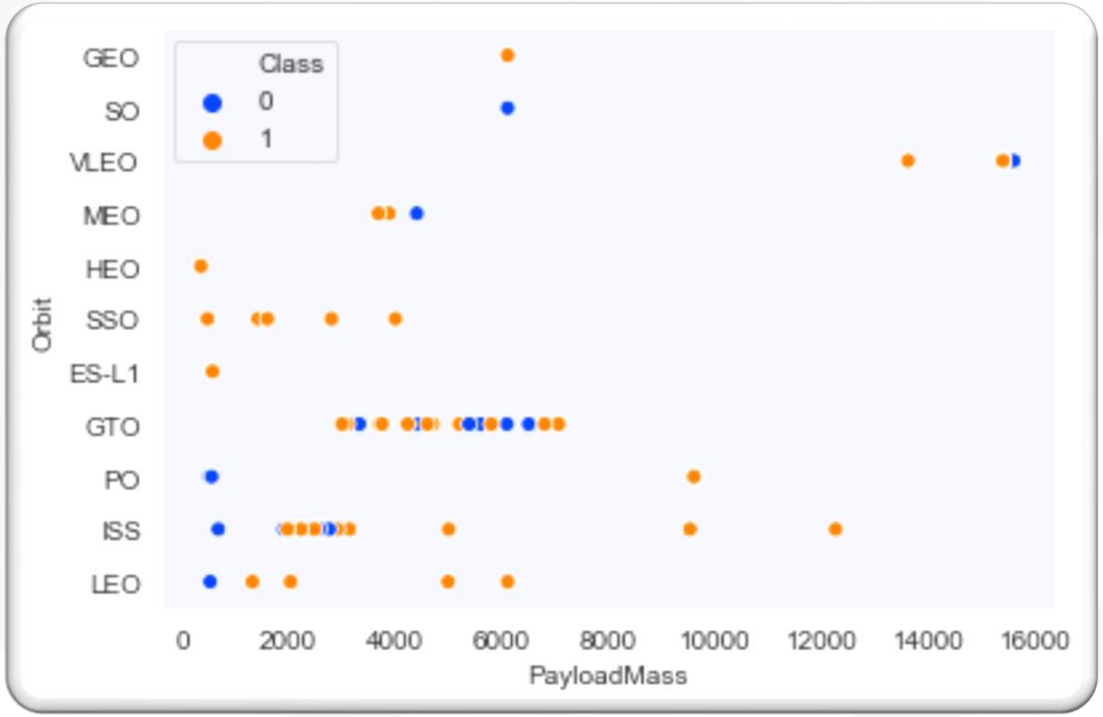
Success rate vs. Orbit type



Flight Number vs. Orbit type



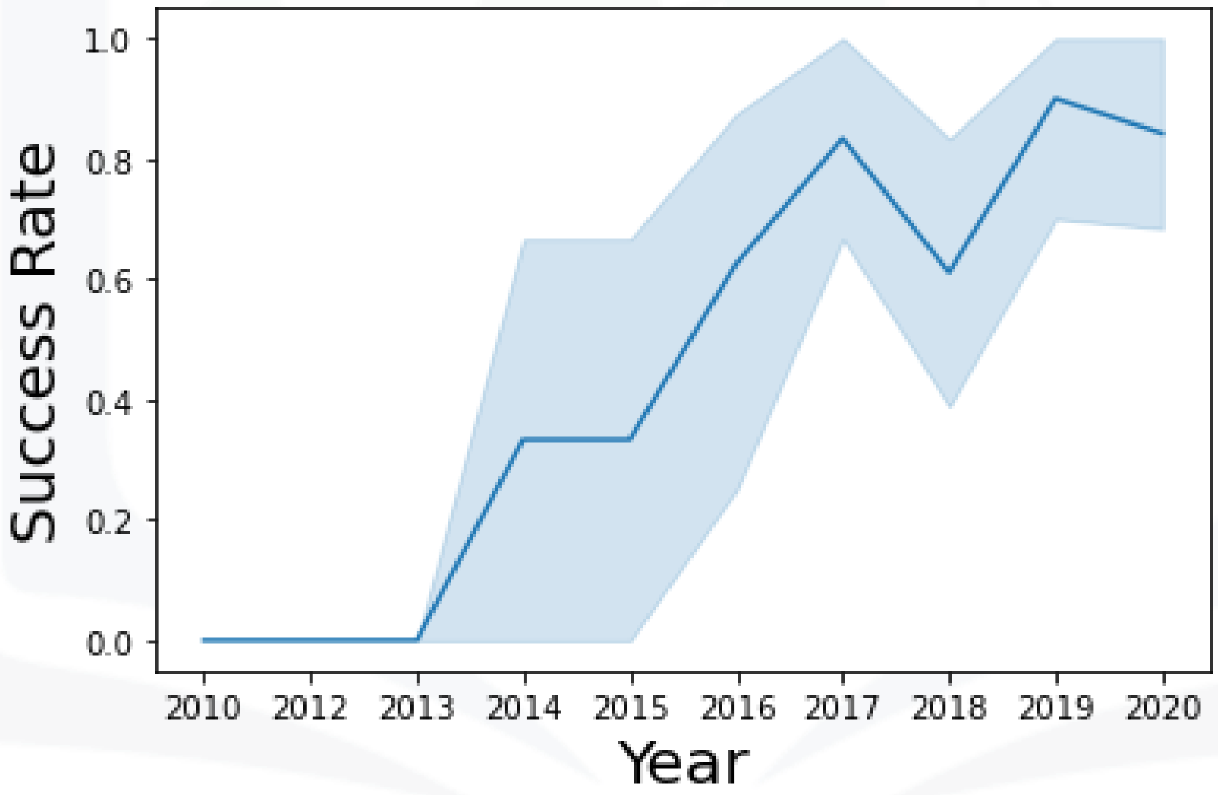
Payload vs. Orbit type



IBM Developer

SKILLS NETWORK

Launch Success Yearly Trend

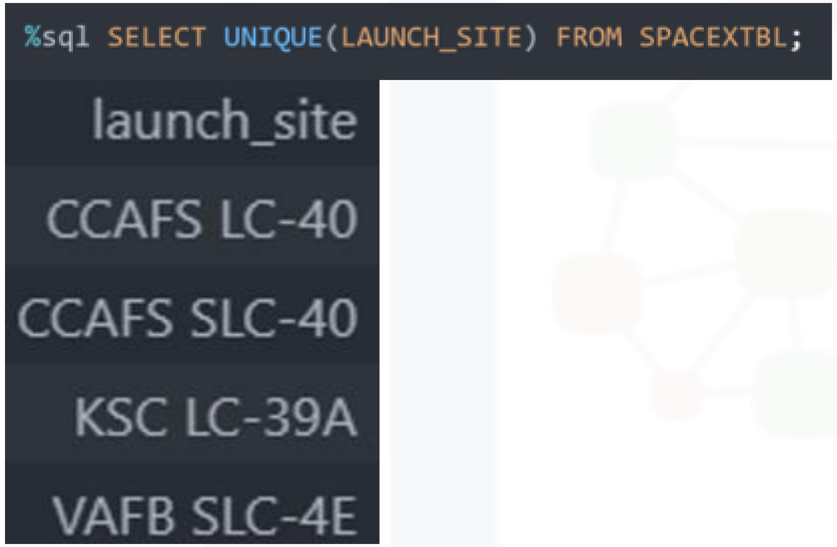


EDA - WITH SQL

IBM Developer

SKILLS NETWORK

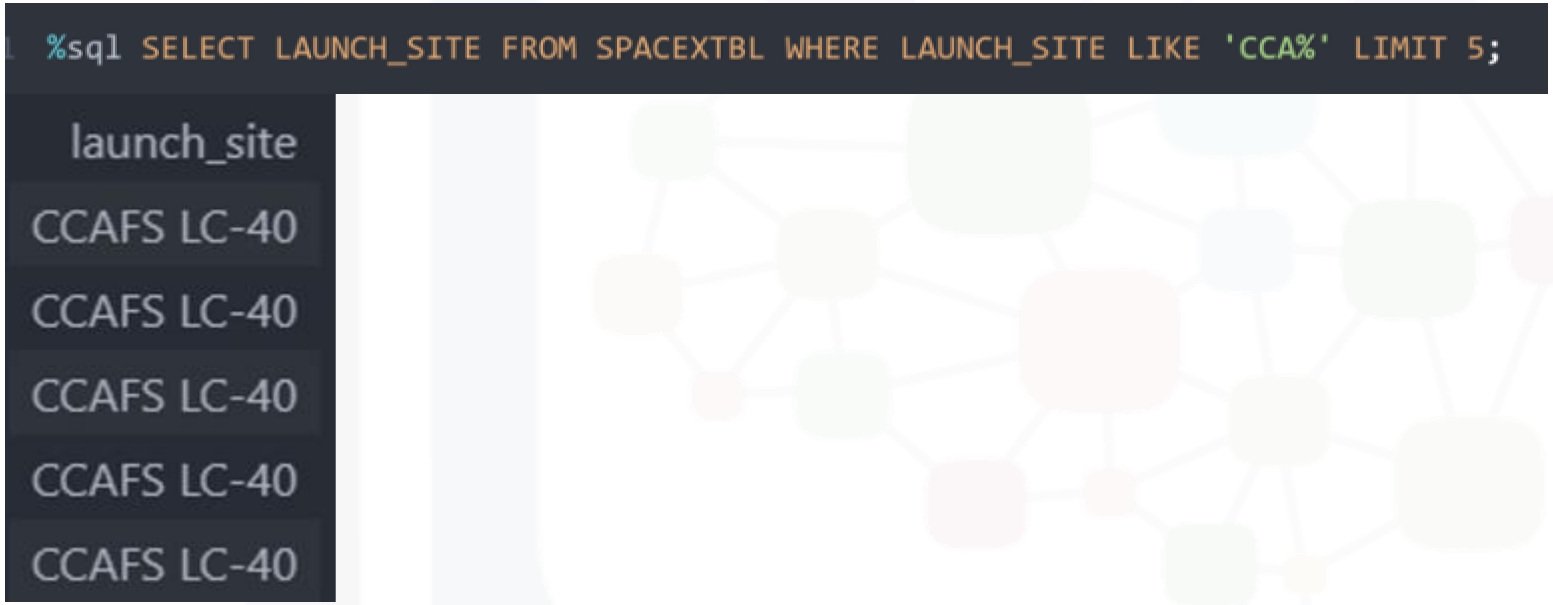
All Launch Site Names



```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL;
```

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

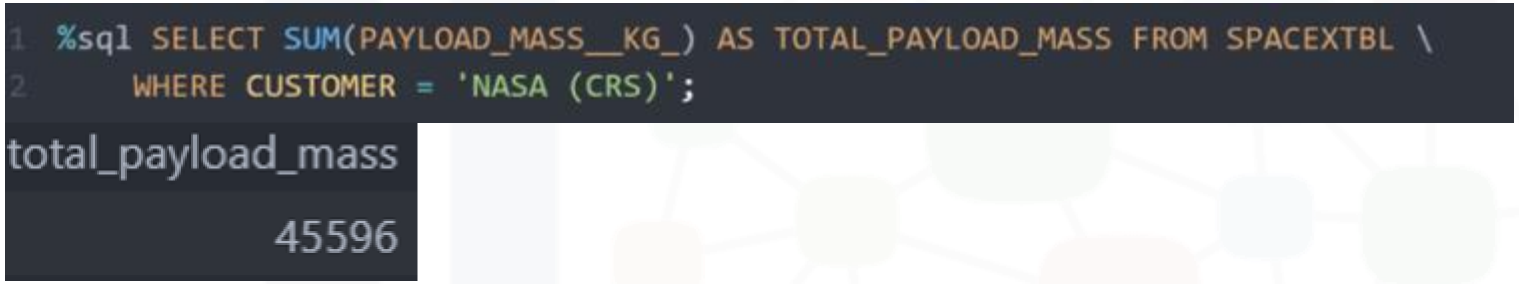
Launch Site Names Begin with 'CCA'



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

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

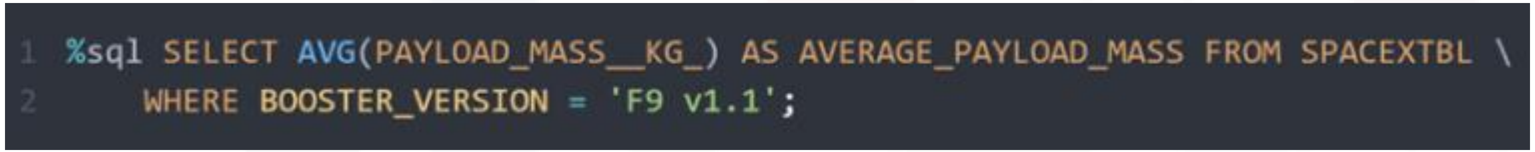
Total Payload Mass from NASA



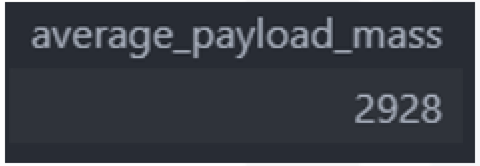
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

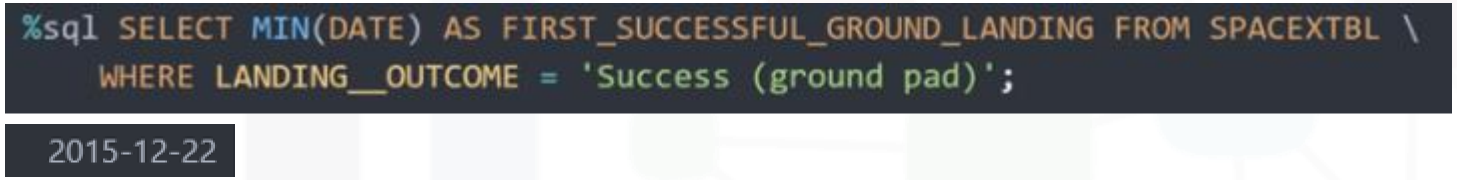


```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```



average_payload_mass
2928

First Successful Ground Pad Landing Date

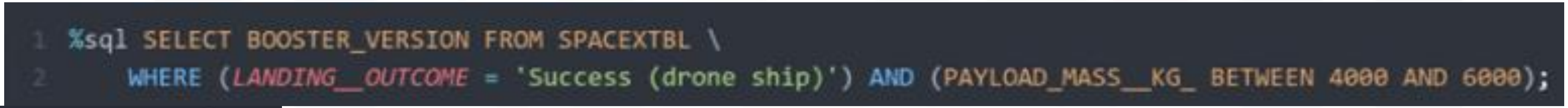


```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
    WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

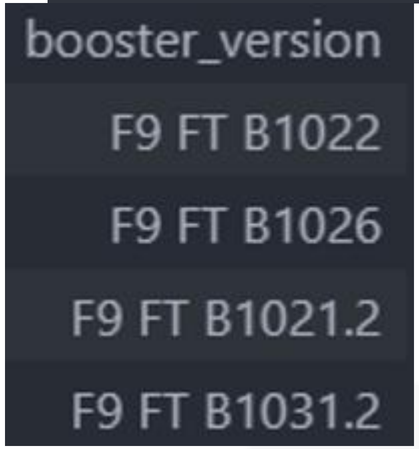
```
2015-12-22
```

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.

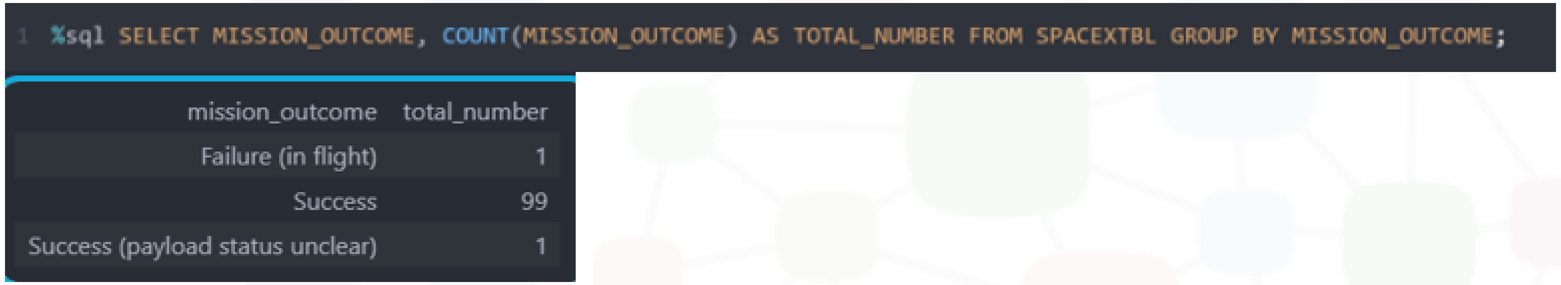


```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```



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

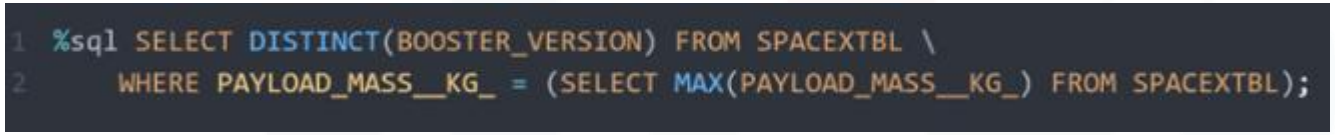
Total Number of Each Mission Outcome



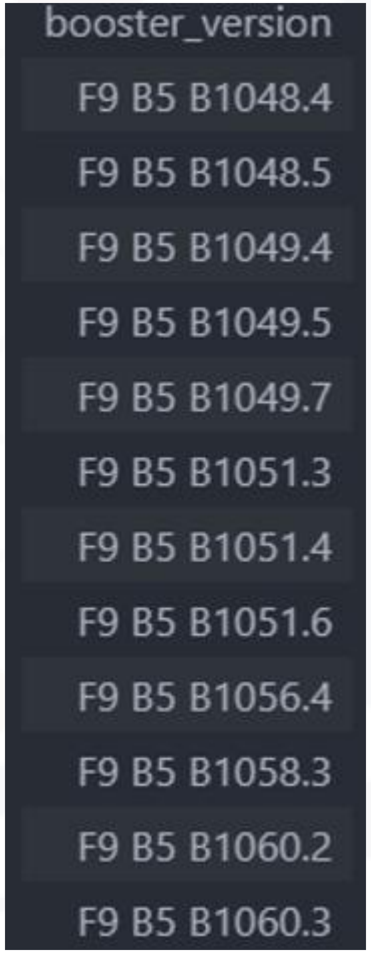
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters that Carried Maximum Payload

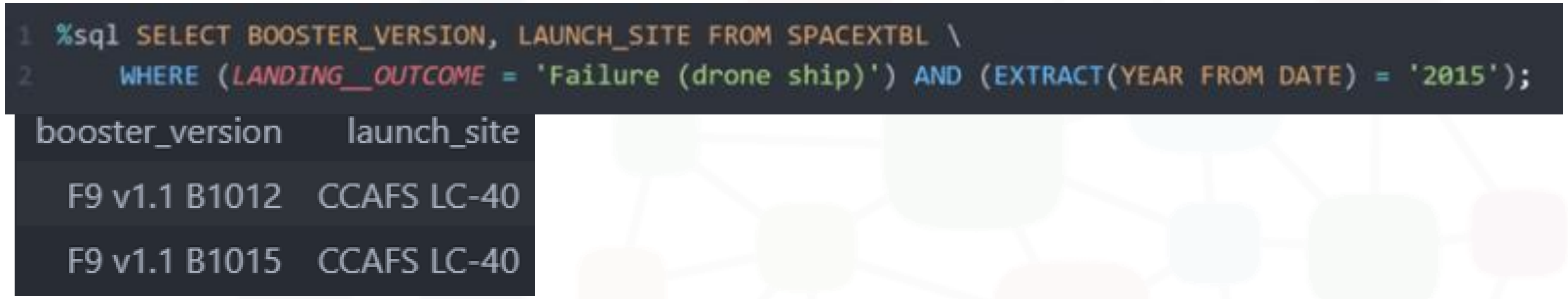


```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```



booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

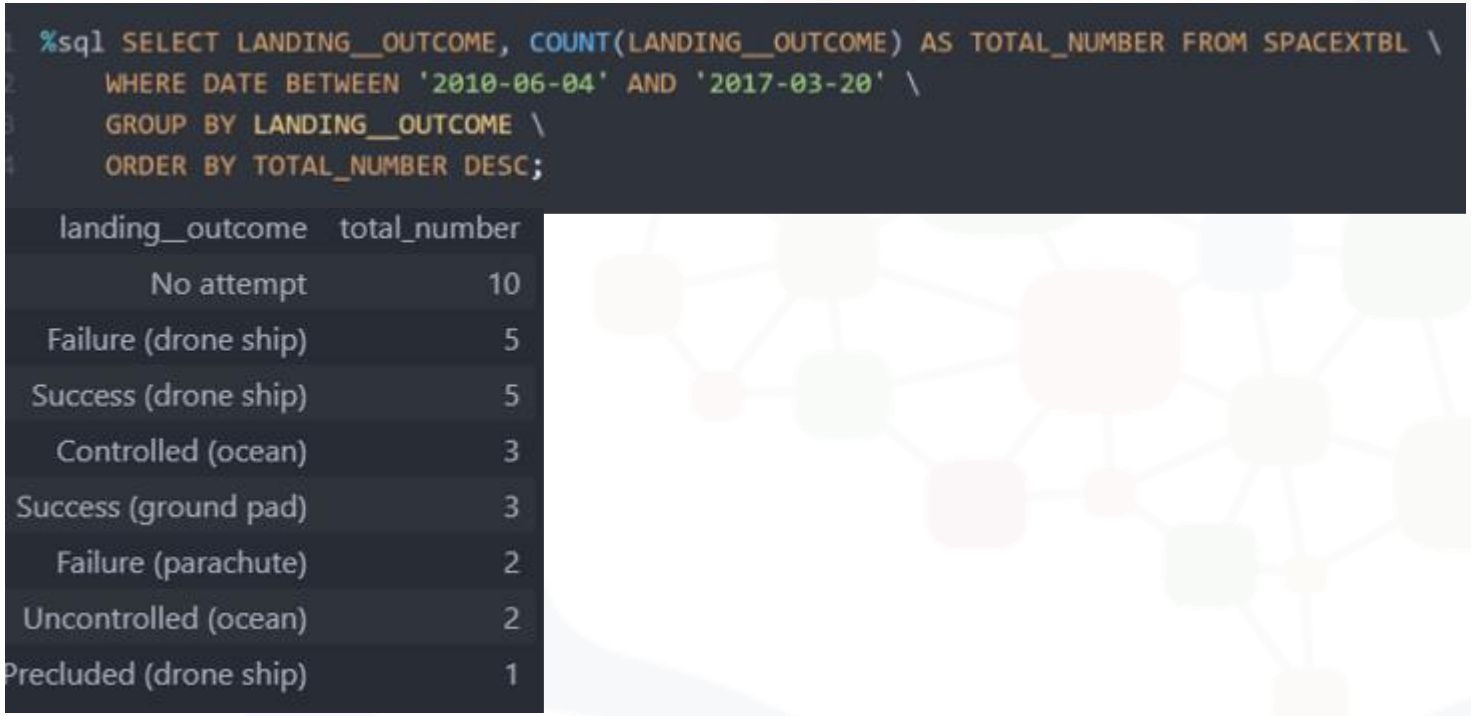
2015 Failed Drone Ship Landing Records



```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (EXTRACT(YEAR FROM DATE) = '2015');
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

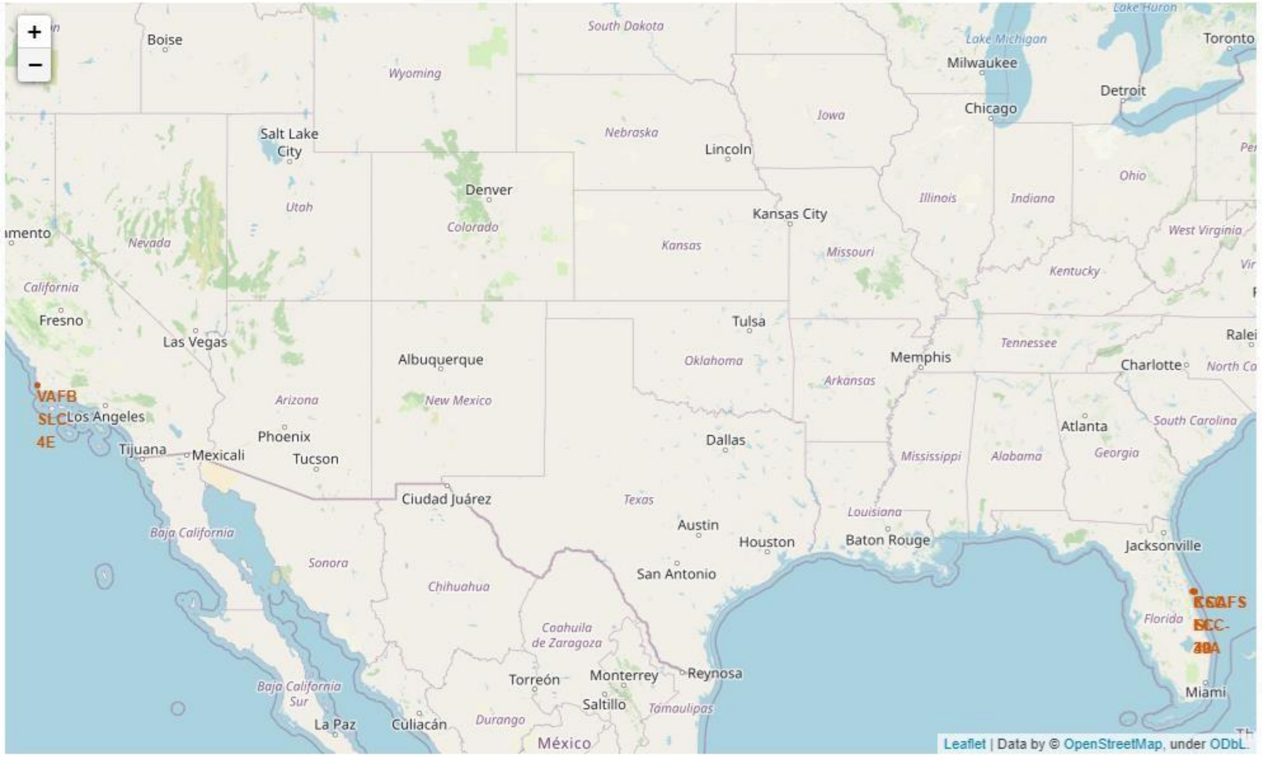


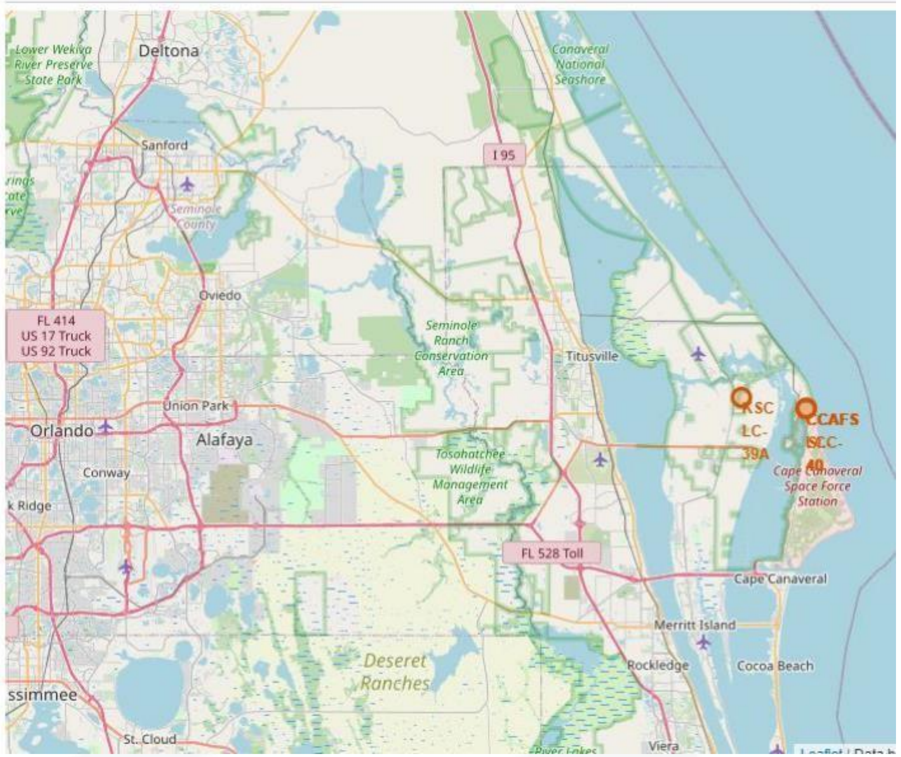
```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING__OUTCOME \
    ORDER BY TOTAL_NUMBER DESC;
```

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

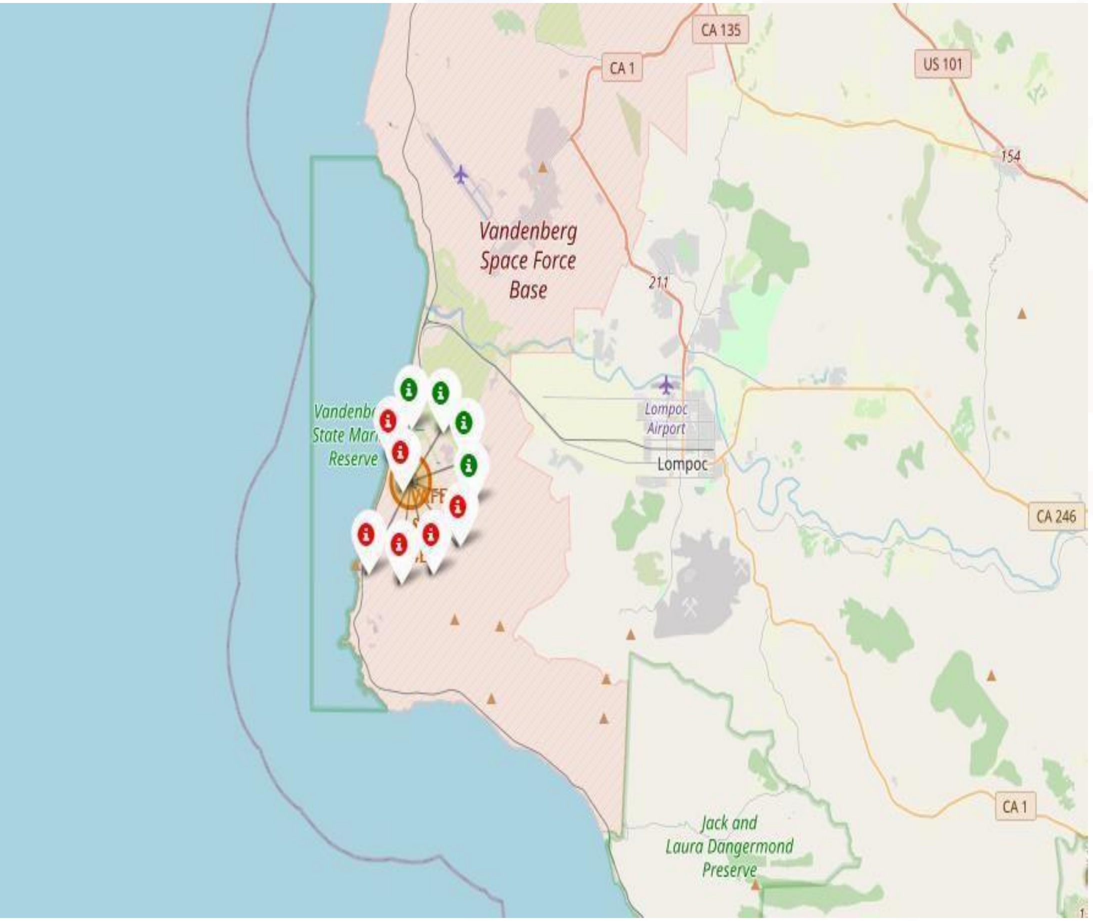
Interactive Map with Folium

Launch Site Locations

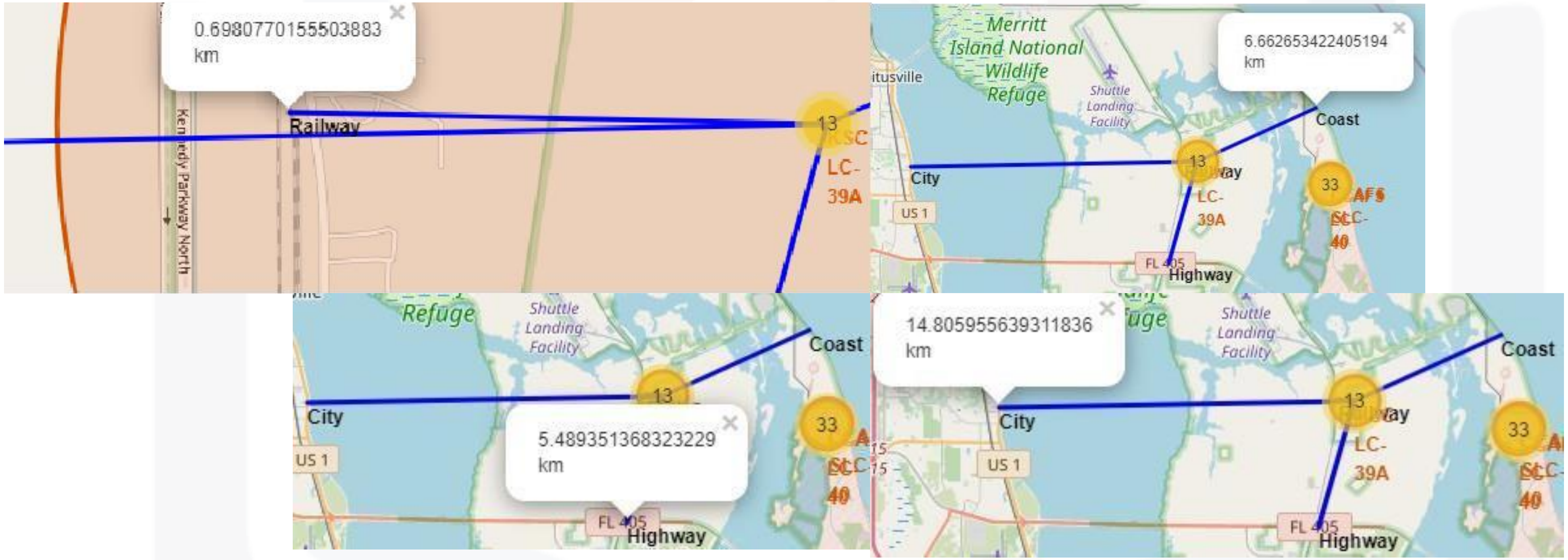




Color-Coded Launch Markers

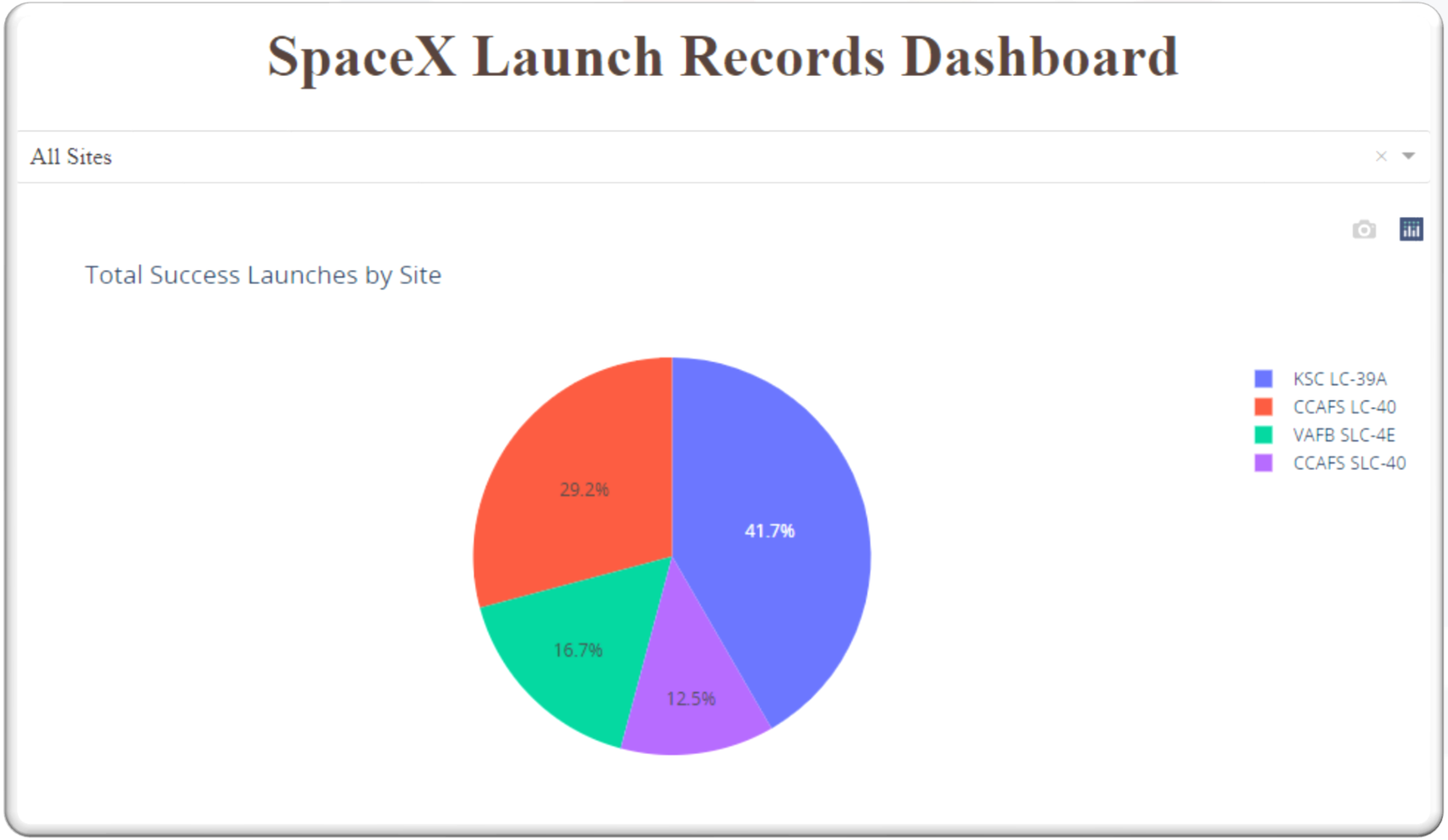

CA 135
CA 1
US 101
154
Vandenberg Space Force Base
211
Lompoc Airport
Lompoc
CA 246
Vandenb State Mar Reserve
CA 1
Jack and Laura Dangermond Preserve

Key Location Proximities



Interactive dashboard   Plotly Dash

Launch success count for all sites

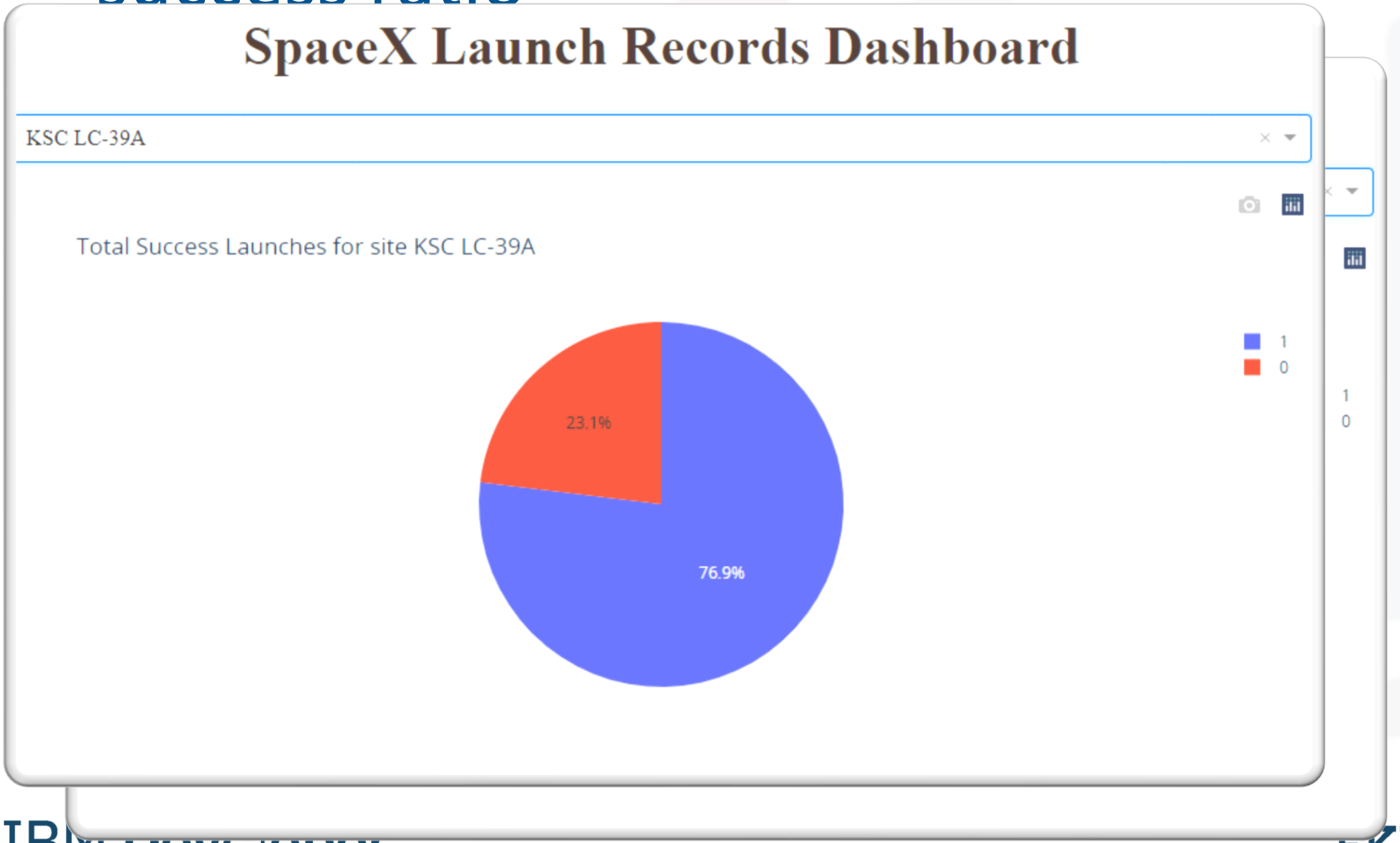


SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

Pie chart for the launch site with highest launch success ratio

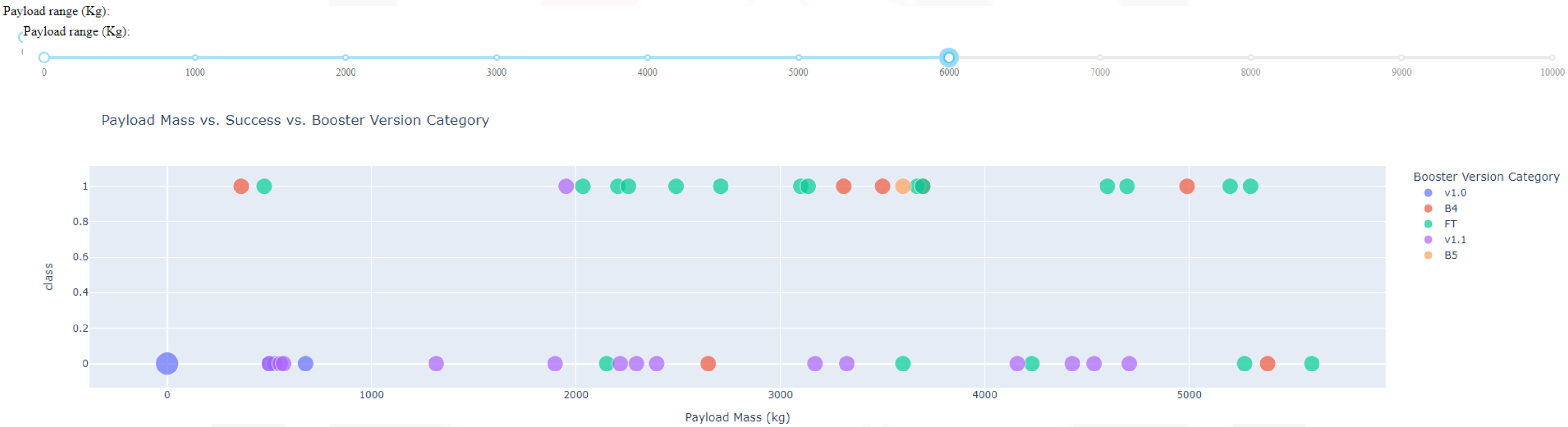


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Payload Mass vs. Success vs. Booster Version Category



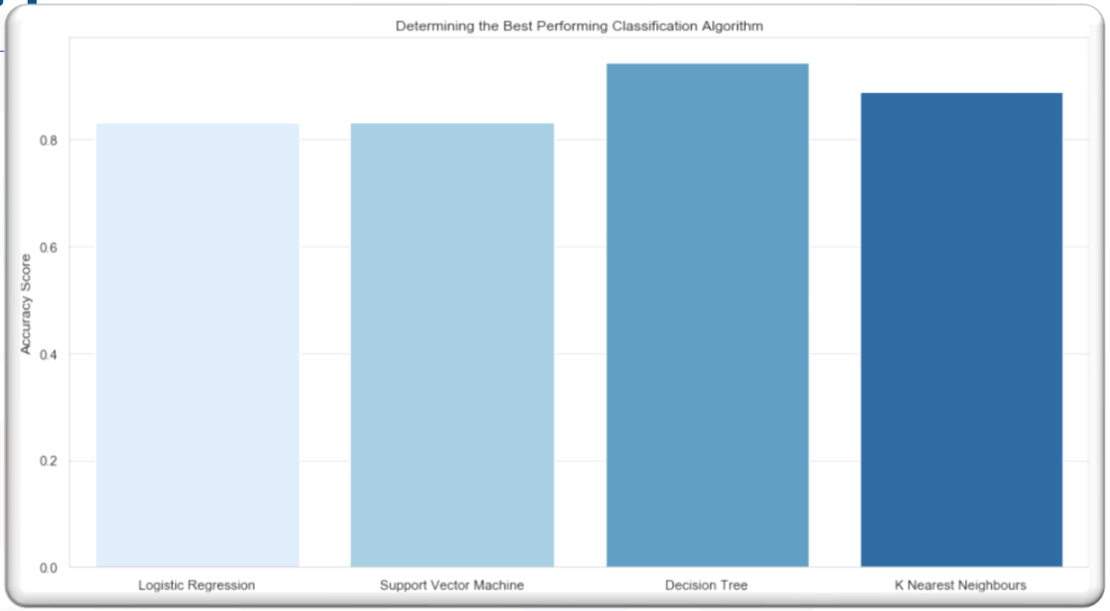
Payload range (Kg):

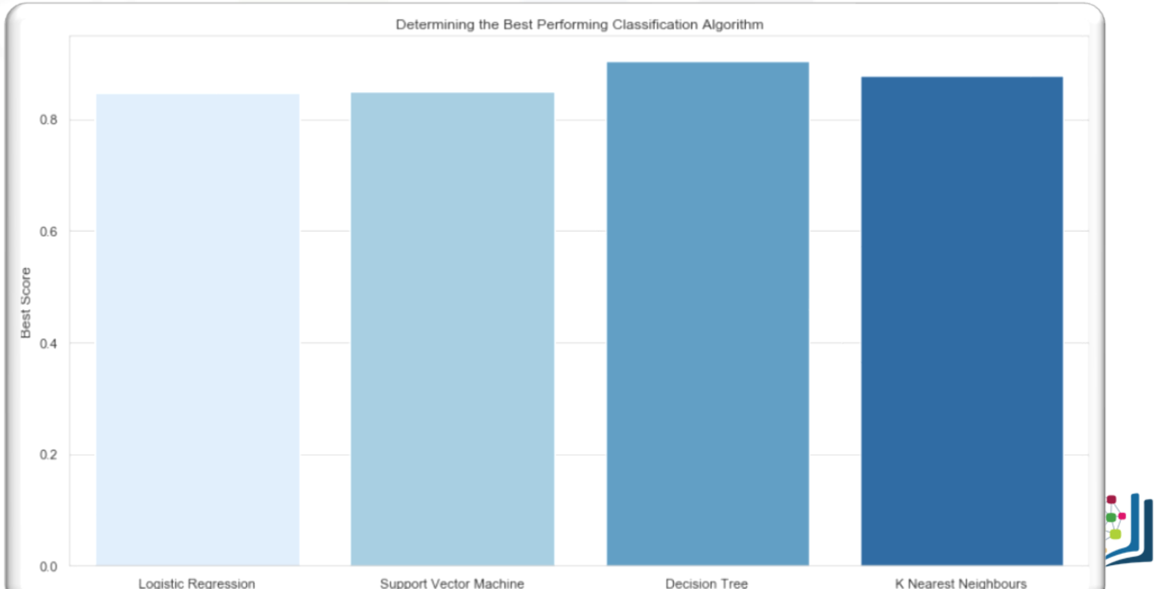
Payload range (Kg):

Predictive Analysis (Classification)

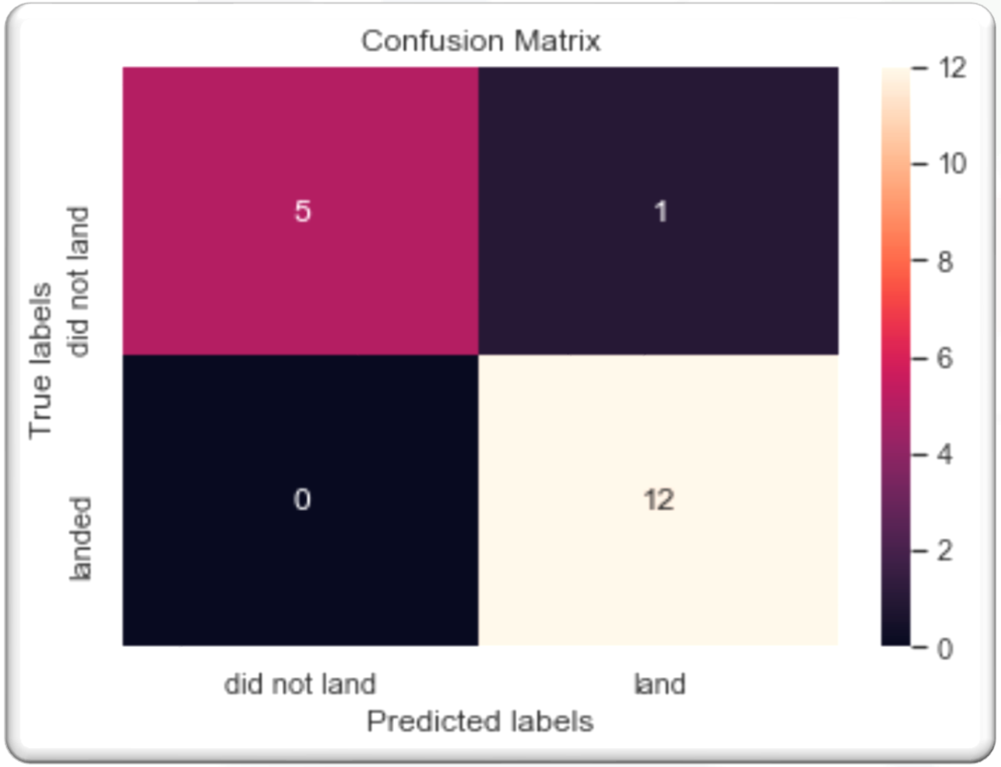
CLASSIFICATION ACCURACY

- Plotting the Accuracy Score and Best Score for each classification algorithm produces the following result:
- The Decision Tree model has the highest classification accuracy
- The Accuracy Score is 94.44%
- The Best Score is 90.36%





Confusion Matrix



- As shown previously, best performing classification model is the Decision Tree model, with an accuracy of 94.44%.
- This is explained by the confusion matrix, which shows only 1 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).
- The other 17 results are correctly classified (5 did not land, 12 did land).

Conclusion

Our mission is to develop a machine learning model for SpaceY, enabling informed bidding against SpaceX by predicting successful Stage 1 landings and potentially saving around $100 million USD per launch. Utilizing data from both a public SpaceX API and web scraping the SpaceX Wikipedia page, we curated data labels and stored the information in a DB2 SQL database. The resulting dashboard offers insightful visualizations.

Model Development:

- Successfully created a machine learning model achieving an accuracy of 83%, providing a reliable tool for SpaceY to predict the success of Stage 1 landings before launch. This predictive capability is crucial for decision-making, allowing SpaceY to assess whether a launch should proceed.

Future Enhancements:

To further optimize the model and improve accuracy, it is recommended to collect additional data. This continuous data collection will facilitate the identification of the best machine learning model, ensuring SpaceY possesses a robust tool for accurate predictions and strategic decision-making in the competitive space launch industry.

Appendix

- Github Link: https://github.com/ShogunGerman/Module-10-/upload/main