IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noureddine Lakouari
12/02/2024

Outline

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

Executive Summary

Summary of Methodologies

Data Collection Techniques:

- Through API: Harnessing APIs for direct and efficient structured data acquisition.
- With Web Scraping: Automated extraction of data from websites, transforming unstructured or semi-structured data into a structured format.

Data Preparation:

- Data Wrangling: Cleansing, structuring, and enriching raw data for improved analysis efficiency.

Exploratory Data Analysis (EDA):

- With SQL: Using SQL for querying and summarizing data to uncover initial insights.
- With Data Visualization: Employing graphical representations to identify data patterns, relationships, and outliers.

Advanced Analytics:

- Interactive Visual Analytics with Folium: Creating interactive geospatial visualizations to enhance data exploration.
- Machine Learning Prediction: Analyzing historical data with machine learning to forecast future trends or outcomes.

Executive Summary

Summary of All Results

- Exploratory Data Analysis Results: A comprehensive overview of insights and patterns identified through EDA.
- Interactive Analytics: Documentation of dynamic data exploration, supported by screenshots of interactive visualizations.
- Predictive Analytics Results: Presentation of predictive modeling outcomes, highlighting key predictions and their decision-making implications.

Introduction

- Project background and context

SpaceX promotes its Falcon 9 rocket launches at a competitive rate of 62 million dollars on its official site, in contrast to other service providers whose charges start at 165 million dollars. A significant portion of this cost efficiency stems from SpaceX's innovative approach to refurbishing and reusing the rocket's first stage. Understanding the likelihood of the first stage landing successfully is pivotal in estimating the overall cost of a launch. This analysis becomes crucial for any competing firms aiming to submit a competitive bid against SpaceX for launch services. The project's primary objective is to develop a predictive model through machine learning that can accurately forecast the success of the first stage landing.

- Problems you want to find answers

The project seeks to unravel:

- The key factors influencing the likelihood of a successful rocket landing.

- How different variables interact to affect the probability of a successful landing.

- The operational conditions necessary for a successful landing program.

By addressing these questions, the project aims to shed light on the complexities of space launch economics and the technological prowess needed for reusable rocket components, potentially revolutionizing cost structures in the space industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.

Our approach to gathering the necessary data encompassed a range of techniques:

- We initiated the process by sending GET requests to the SpaceX API, a crucial step for accessing detailed launch data.
- Upon receiving the API response, we employed the ".json()" function to decode the content into JSON format, subsequently utilizing "pandas.json_normalize()" to seamlessly transition this data into a structured pandas dataframe.
- The data underwent a thorough cleaning process, during which we identified and rectified missing values, ensuring a comprehensive dataset for analysis.
- To complement our dataset, we also engaged in web scraping endeavors targeting Wikipedia for Falcon 9 launch records. Leveraging the BeautifulSoup library, we extracted launch records presented in HTML tables, parsed these tables, and converted the information into a pandas dataframe for in-depth analysis.

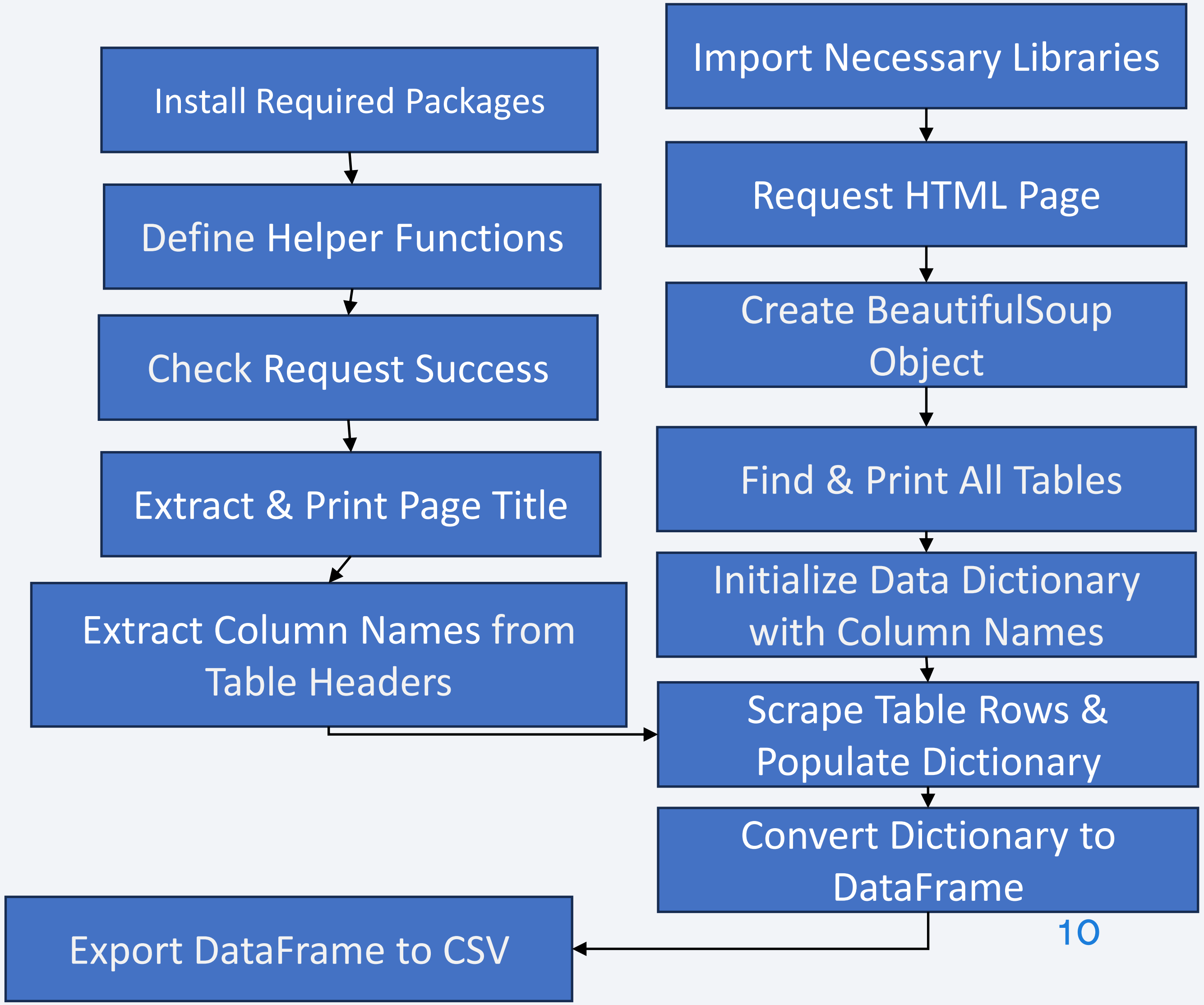
Data Collection – SpaceX API

- We initiated data collection by issuing GET requests to the SpaceX API, followed by the cleansing and preliminary structuring of the retrieved data.
- (https://github.com/lakouari/IBM/blob/main/jupyter-labs-spacex-data-collection-api.ipynb), as an external reference and peer-review purpose

Step Number	Process Step	Description
1	Start	Begin the data collection and processing sequence
2	Make GET Request to SpaceX API	Send a GET request to the SpaceX API endpoint to retrieve data
3	Receive API Response	Obtain the response from the SpaceX API, which includes data in JSON format
4	Check Response Status	Verify if the API call was successful (e.g., HTTP 200 OK). Handle errors if the call was unsuccessful
5	Decode JSON Response	Convert the JSON response into a structured data format for easier manipulation
6	Data Cleaning	Identify and rectify any missing or incomplete data fields within the dataset
7	Data Wrangling and Formatting	Transform and restructure the data as needed for analysis, such as normalizing date formats
8	Store or Output Processed Data	Save the cleaned and structured data for subsequent analysis or visualization tasks.
9	End	Conclude the data collection and processing workflow.

Data Collection - Scraping

- The flowchart outlines the sequential steps involved in the web scraping project, from initial setup and data collection to data extraction, transformation, and finally exporting the structured data to a CSV file. Each step is designed to progressively move from setting up the environment and gathering data to processing and storing that data in a usable format. https://github.com/lakouari/IBM/blob/main/jupyter-labs-webscraping.ipynb

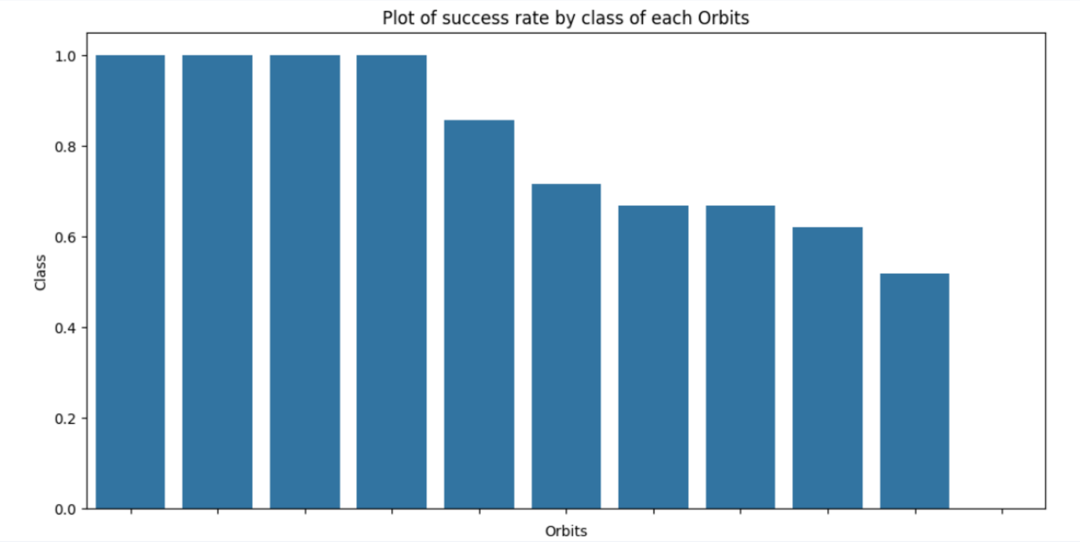


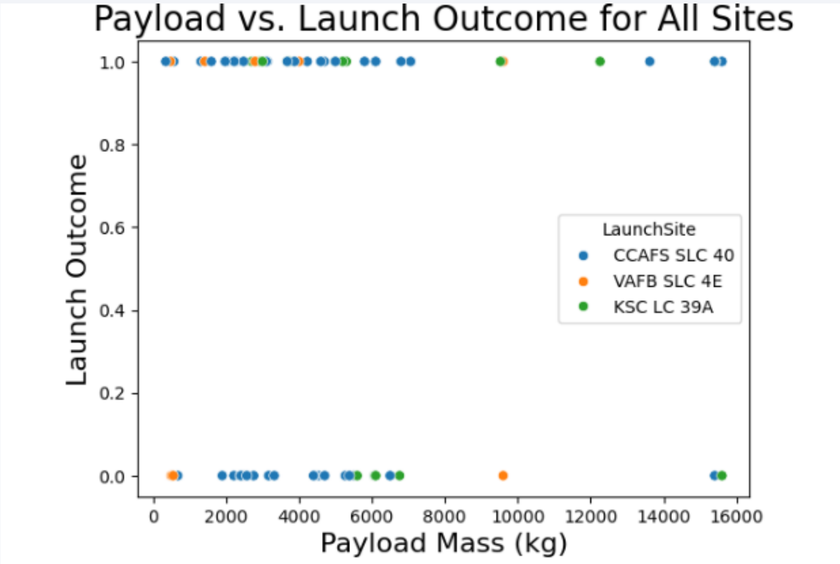
Data Wrangling

- The lab involved performing Exploratory Data Analysis (EDA) on Space X Falcon 9 launch data to understand patterns and prepare the data for supervised machine learning models. The key task was to process the data to create a label indicating whether the booster landing was successful (1) or not (0). Various outcomes like True Ocean, False RTLS, and others were mapped to these binary outcomes. The lab began with importing necessary Python libraries, followed by loading and inspecting the dataset for missing values and understanding data types. Further steps included analyzing launch sites, orbits, and mission outcomes, leading to the creation of a binary classification column for landing success.
- https://github.com/chuksoo/IBM-Data-Science-Capstone-SpaceX/blob/main/Data%20Wrangling.ipynb

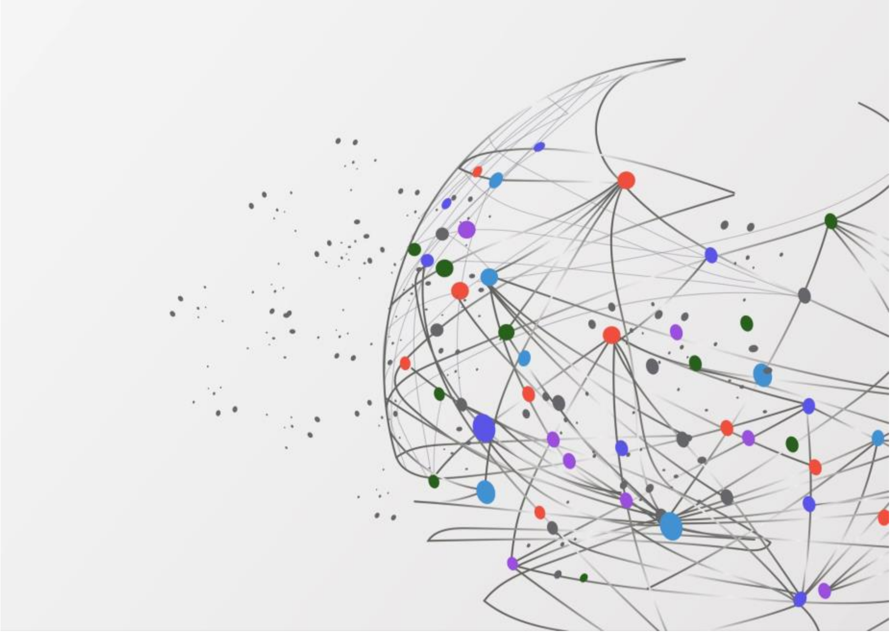
EDA with Data Visualization

- In this assignment, the objective was to prepare data for predicting the success of SpaceX Falcon 9 first-stage landings. The work involved Exploratory Data Analysis (EDA) and Feature Engineering using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn.
- https://github.com/lakouari/IBM/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb





EDA with SQL



The SQL queries performed in the peer assignment:

- Task 1: Queried the database to list unique launch sites.
- Task 2: Retrieved 5 records where launch sites begin with 'CCA'.
- Task 3: Calculated the total payload mass carried by boosters launched by NASA (CRS).
- Task 4: Determined the average payload mass carried by the booster version F9 v1.1.
- Task 5: Identified the date of the first successful landing outcome at a ground pad.
- Task 6: Listed booster names that had a successful landing on a drone ship with a payload mass between 4000 and 6000 kg.
- Task 7: Counted the total number of successful and failed mission outcomes.
- Task 8: Found the booster versions that carried the maximum payload mass.
- Task 9: Displayed records with failure landing outcomes on drone ships, along with booster versions and launch sites for the year 2015.
- Task 10: Ranked the count of various landing outcomes between specified dates in descending order.

https://github.com/lakouari/IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

In the Folium map analysis for the SpaceX launch site locations, the following map objects were created and added:

- Markers: To pinpoint the exact locations of the SpaceX launch sites on the map.
- Circles: To highlight the area around the launch sites for better visibility and to represent a region of interest.
- Lines (Polylines): To illustrate the distance from each launch site to nearby points of interest, such as coastlines, railways, highways, and cities.

These objects were added to provide interactive visual aids that can help analyze geographical patterns related to launch sites. The markers and circles allow users to easily identify the location of launch sites. The lines offer a simple way to measure distances from launch sites to key infrastructure and geographic features, which might influence launch site selection and success rates due to logistical, economic, and safety considerations.

- https://github.com/lakouari/IBM/blob/main/lab_jupyter_launch_site_location_jupyterlite.ipynb

Build a Dashboard with Plotly Dash

In the Dash application code, several interactive components and visualizations have been included to create a dashboard for analyzing SpaceX launch data:

Dropdown for Launch Site Selection:

- A dropdown component allows users to select a specific launch site or all launch sites.
- This interaction enables users to filter the data according to the launch site they are interested in.

Pie Chart for Launch Success Rates:

- A pie chart visualization displays the total number of successful launches per site or the success and failure counts when a specific site is selected.
- The pie chart is dynamically updated based on the launch site selected from the dropdown, providing a visual comparison of launch outcomes.

Range Slider for Payload Mass:

- A range slider lets users select a range for the payload mass to analyze.
- This interaction helps in narrowing down the data to payload masses that are within the user's specific interest.

Scatter Plot for Payload vs. Launch Success:

- A scatter plot shows the correlation between payload mass and launch success. This chart also uses the booster version category as a color dimension.
- The scatter plot updates based on both the selected launch site and the range of payload mass chosen. It is useful for identifying trends and patterns in how payload mass might relate to launch success rates

https://github.com/lakouari/IBM/blob/main/Dash_lab.ipynb

Predictive Analysis (Classification)

To summarize the process of building, evaluating, improving, and finding the best-performing classification model, the following key steps were taken:

Data Loading and Preprocessing:

- Data was loaded from a CSV file.
- Features and labels were extracted.
- The data was standardized using StandardScaler.

Splitting Data:

- The dataset was split into training and testing sets using train_test_split.

Model Training and Hyperparameter Tuning:

- Various classification models were instantiated: Logistic Regression, SVM, Decision Tree, and KNN.
- GridSearchCV was used to find the best hyperparameters for each model by searching through a predefined grid of parameters.
- Each model was trained using cross-validation to ensure the model's generalizability.

Model Evaluation:

- The best hyperparameters were selected based on the highest cross-validation accuracy.
- Models were evaluated on the test set to determine their performance.
- Accuracy scores and confusion matrices were generated for each model.

Selection of the Best Model:

- Comparing the test set accuracies, the model with the highest accuracy was selected as the best classifier.
- A bar chart was used to visually compare the performance of each model.

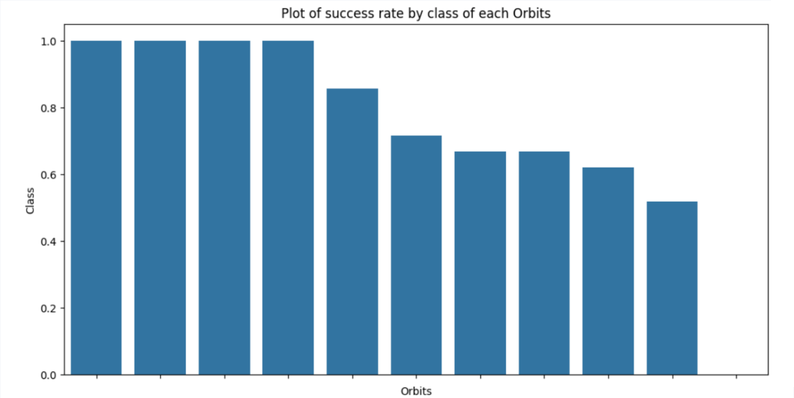
Best Model Output:

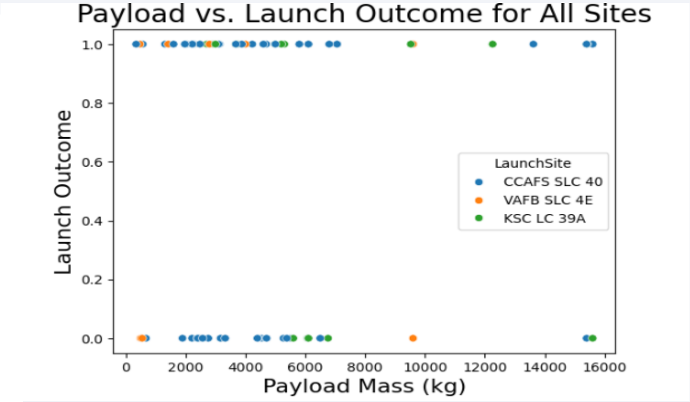
- The best model's parameters and test set accuracy were reported.

- https://github.com/lakouari/IBM/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

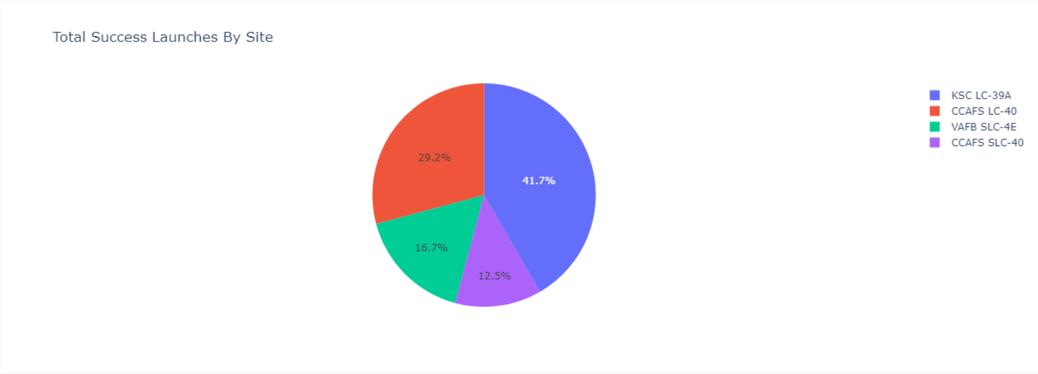
Results

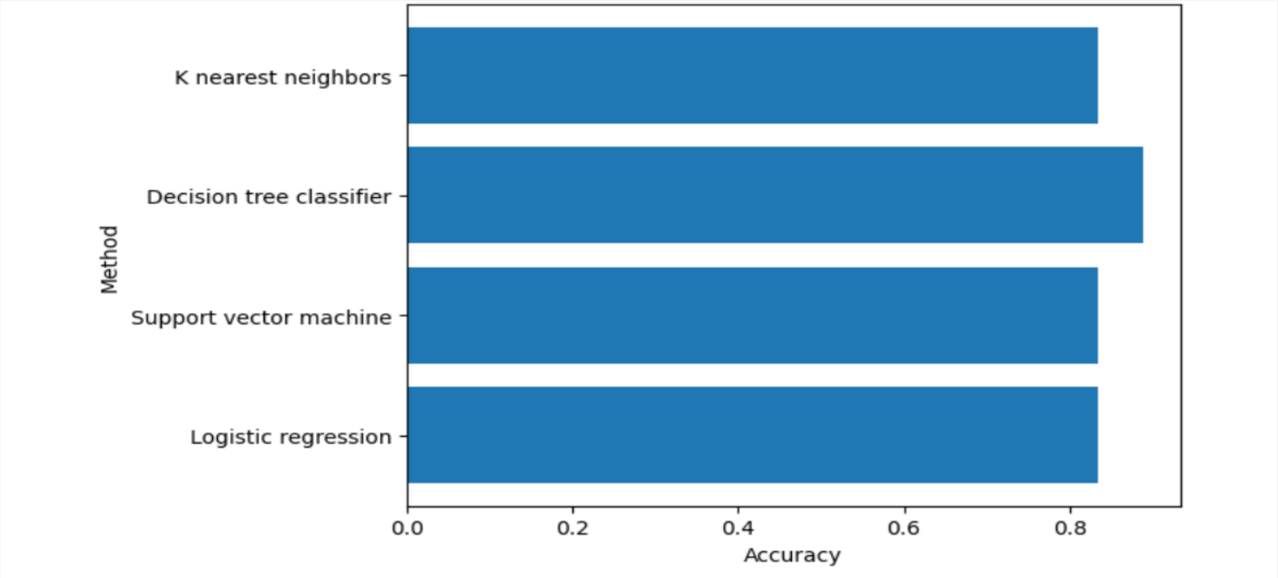
- Exploratory data analysis results





- Interactive analytics demo in screenshots





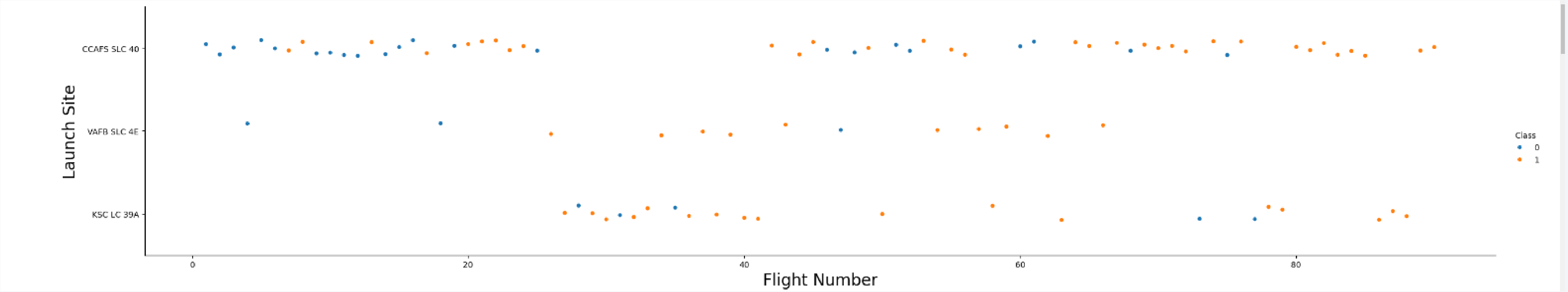
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

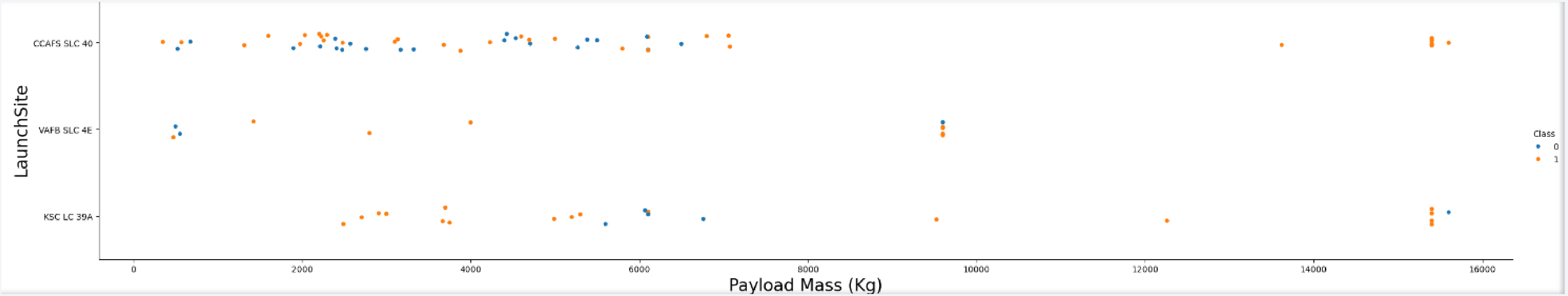
- Show a scatter plot of Flight Number vs. Launch Site



This figure shows that the success rate increased as the number of flights increased (`Class 0` (blue) represents unsuccessful launch while `Class 1` (orange) represents successful launch).

Payload vs. Launch Site

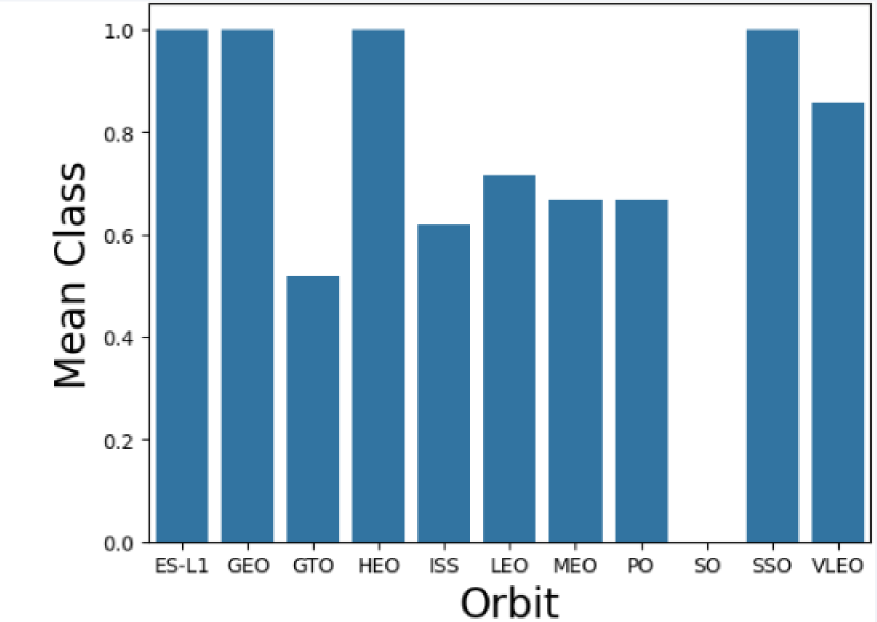
- Show a scatter plot of Payload vs. Launch Site



Class 0, depicted in blue, indicates a failed launch, whereas Class 1, shown in orange, signifies a successful one. Initial observations suggest that an increased payload mass could potentially correlate with a heightened likelihood of a successful launch. However, discerning a definitive pattern or relationship between launch success and payload mass from the graph alone proves challenging due to the absence of an immediately discernible trend.

Success Rate vs. Orbit Type

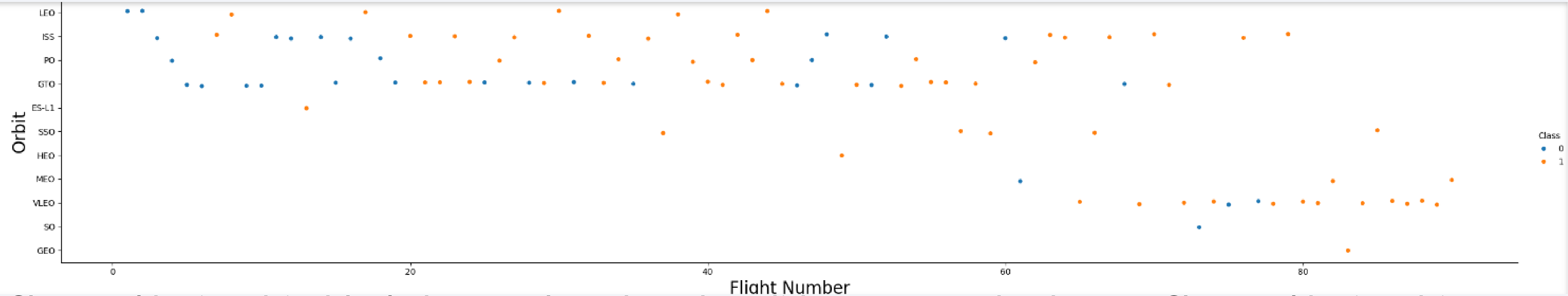
- Show a bar chart for the success rate of each orbit type



- Orbit classifications such as SSO, HEO, GEO, and ES-L1 are observed to achieve the maximum success rates, hitting the 100% mark. In contrast, the GTO orbit type exhibits a relatively modest success rate of 50%, which stands as the lowest among the categories, aside from the SO type, which experienced a single unsuccessful attempt.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

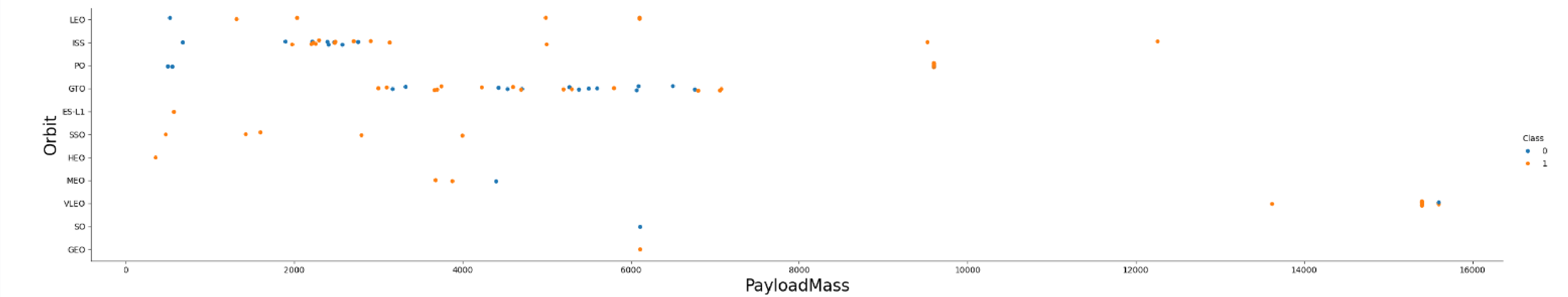


- Class 0 (depicted in blue) denotes launches that did not succeed, whereas Class 1 (depicted in orange) signifies launches that were successful.

- Generally, there appears to be a link between the outcome of the launch and the sequential number of the flight. However, for launches into GTO orbit, this correlation does not seem to hold, suggesting flight numbers do not predict success rates in this context. Initially, SpaceX's launches to LEO achieved moderate success, but more recent launches have increasingly targeted VLEO, which has shown a higher success rate.

Payload vs. Orbit Type

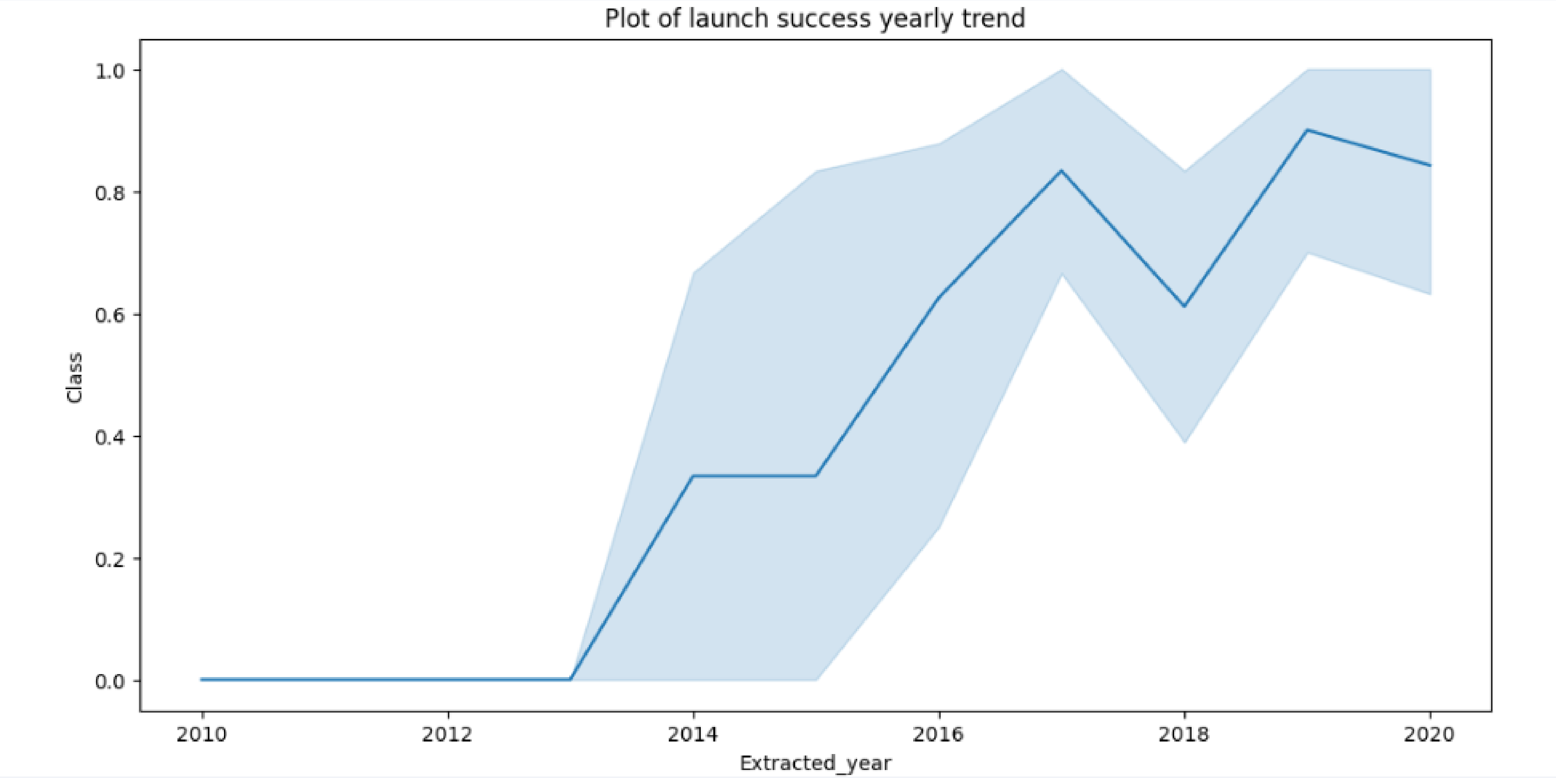
- Show a scatter point of payload vs. orbit type



- In the classification scheme, Class 0 (visualized in blue) stands for launches that did not end successfully, while Class 1 (visualized in orange) indicates those that did. Observations suggest that launches carrying heavier payloads tend to have a higher likelihood of successful landings, particularly for missions directed towards Low Earth Orbit (LEO) and the International Space Station (ISS). Conversely, for Geostationary Transfer Orbit (GTO) missions, it is challenging to differentiate between successful and unsuccessful landings, as the data points are closely clustered without a clear pattern of success.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- Beginning in 2013, there has been an upward trend in the rate of successful launches, peaking in 2017. A minor dip in this trend was observed in 2018.

All Launch Site Names

- Find the names of the unique launch sites

```
[8]: %%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

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

[8]:

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

- Utilizing the DISTINCT keyword in SQL queries ensures that only non-duplicate values are presented from the Launch_Site column of the SpaceX dataset. This operation reveals that there exist four distinct launch locations: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
[9]: %%sql
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

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

[9]:

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

- Calculate the total payload carried by boosters from NASA

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

```
[10]: %%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

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

[10]:

total_payload_mass_kg
45596

Using the `SUM()` function to calculate the sum of column `PAYLOAD_MASS__KG_`.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

Task 4

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

```
[11]: %%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

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

[11]:

avg_payload_mass_kg
2928.4

Using the AVG() function to calculate the average value of column PAYLOAD_MASS__KG_.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
[16]: %%sql
SELECT MIN(DATE) AS first_successful_landing_date
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'

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

[16]:

first_successful_landing_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

```
[17]: %%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)

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

[17]:

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

[15]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

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

[15]:

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

To obtain a summary of mission outcomes, the dataset is aggregated based on the mission result categories using the GROUP BY clause. The COUNT() function is then applied to tally the occurrences within each category. The derived data suggests that SpaceX has achieved a remarkable mission success rate, with approximately 98% of its missions culminating successfully.

Boosters Carried Maximum Payload

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

[19]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

* sqlite:///my_data1.db
Done.

[19]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

A subquery is utilized to first determine the heaviest payload carried by employing the MAX() function on the payload mass column. Then, the main query filters the dataset to identify records where the payload mass equals this maximum value. The outcome reveals that the version F9 B5 B10xx.x boosters have the capability to transport the heaviest payload as per the available data.

2015 Launch Records

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

```
[20]: %%sql
SELECT Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND strftime('%Y', DATE) = '2015'

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

[20]:

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

- Using the `AND` operator to display a record if additional condition `YEAR` is `2015`.
- In 2015, there were two landing failures on drone ships.

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

```
[21]: %%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
```

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

[21]:

Landing_Outcome	total_number
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

- In the query, the data is filtered to select records within the date range of June 4, 2010, to March 20, 2017. The sorted results, arranged in descending order by the total number of landings, indicate a comparable number of both successful and unsuccessful landings within the specified timeframe.

Section 3

Launch Sites Proximities Analysis

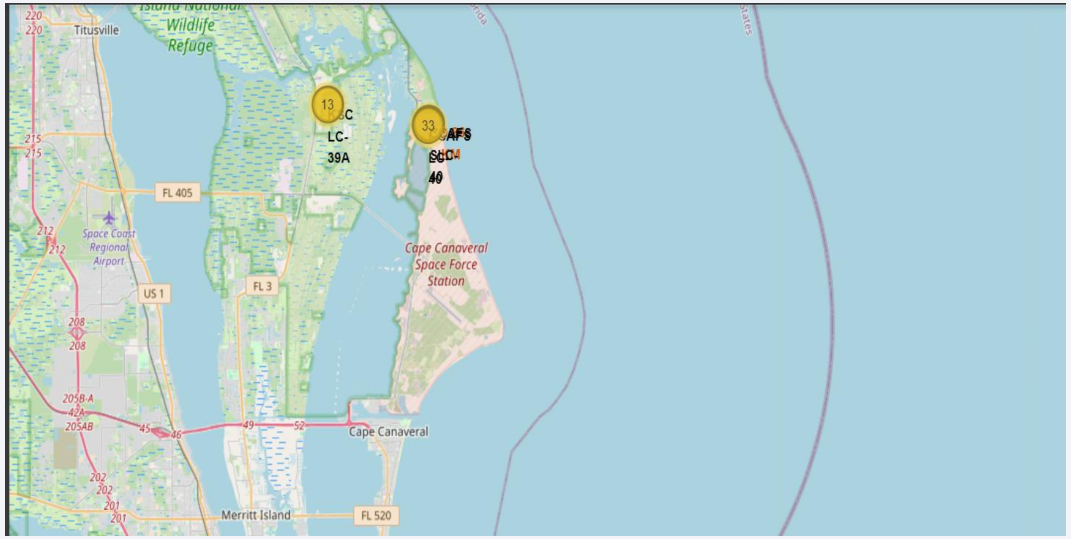
All Launch Sites' Locations

The map shows all *SpaceX* launch sites, and they are in the United States.

Color-labeled Launch Outcomes

By clicking on the marker clusters, successful landing (green) or failed landing (red) are displayed.

Proximities of Launch Sites

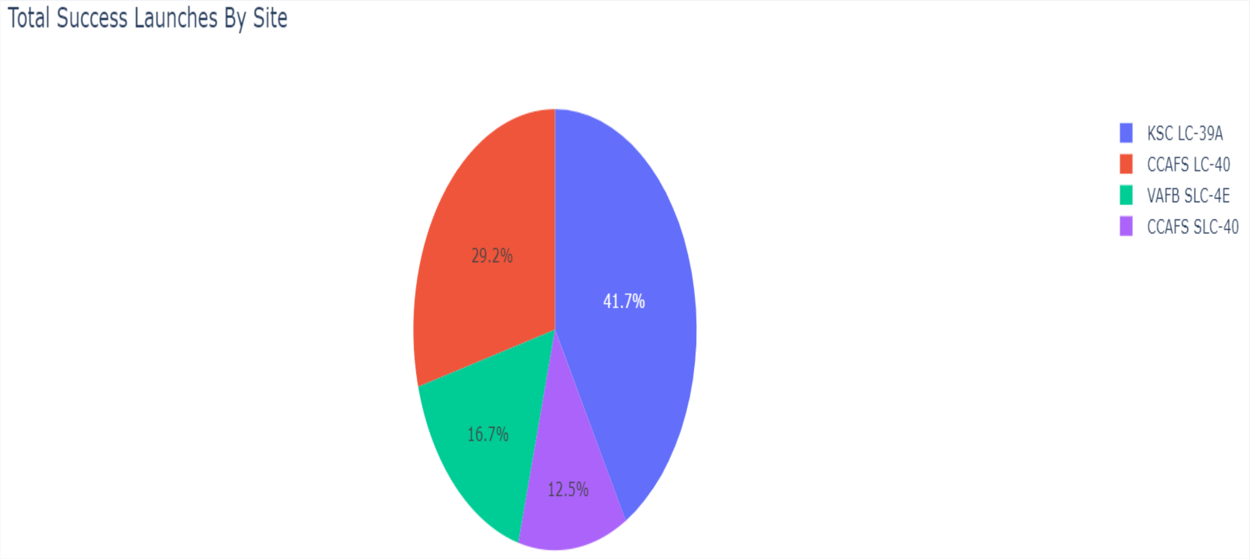


It can be found that the launch site is close to railways and highways for transportation of equipment or personnel, and is also close to coastline and relatively far from the cities so that launch failure does not pose a threat.

Section 4

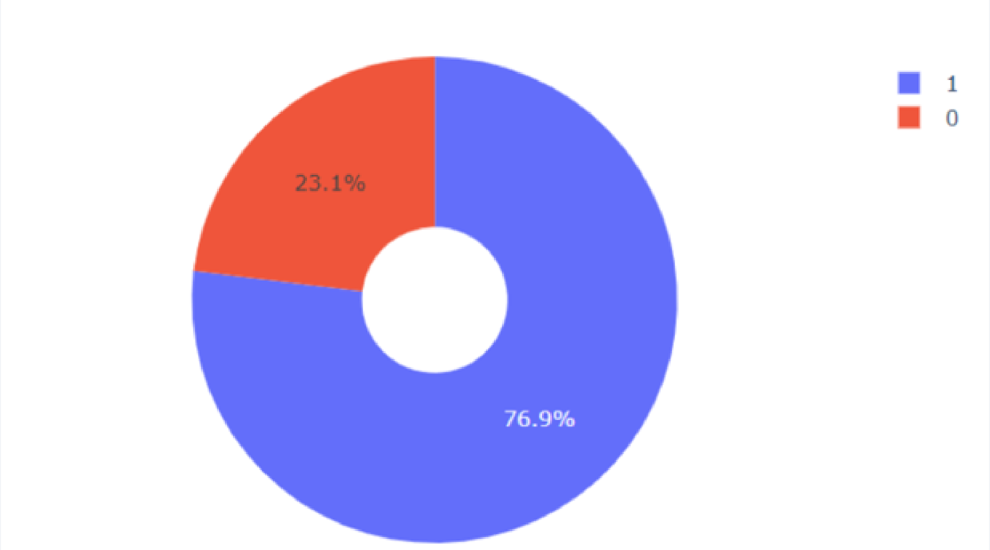
Build a Dashboard with Plotly Dash

Total Success Launches By all sites



The Kennedy Space Center's Launch Complex 39A (KSLC-39A) has the highest count of successful launches in comparison to other sites. Conversely, the Vandenberg Air Force Base's Space Launch Complex 4E (VAFB SLC-4E) exhibits the least number of successful launches, which could be attributed to either a smaller dataset or potentially more challenging launch conditions on the West Coast relative to those on the East Coast.

Launch Site with Highest Launch Success Ratio



- `KSLC-39A` has the highest success rate with 10 landing successes *(76.9%)* and 3 landing failures *(23.1%)*.

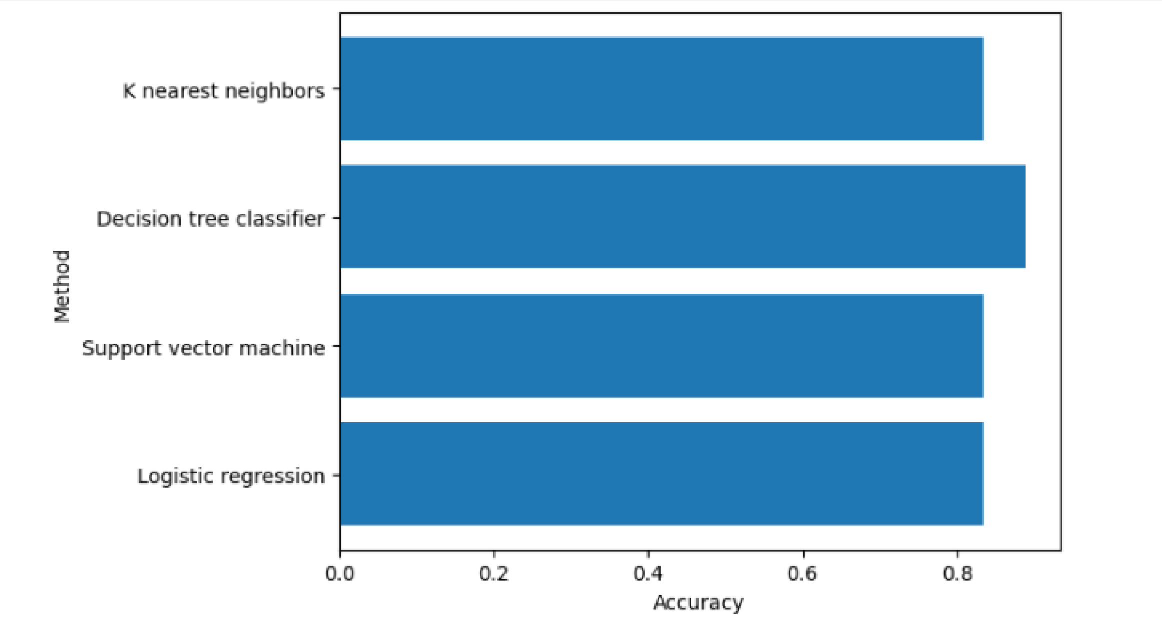
Payload vs. Launch Outcome Scatter Plot for all sites

The data indicates that missions tasked with transporting payloads within the lower weight range (0-5000 kg) have a greater probability of achieving launch success (classified as class 1) when compared to those carrying heavier payloads ranging from 5000 to 10000 kg.

Section 5

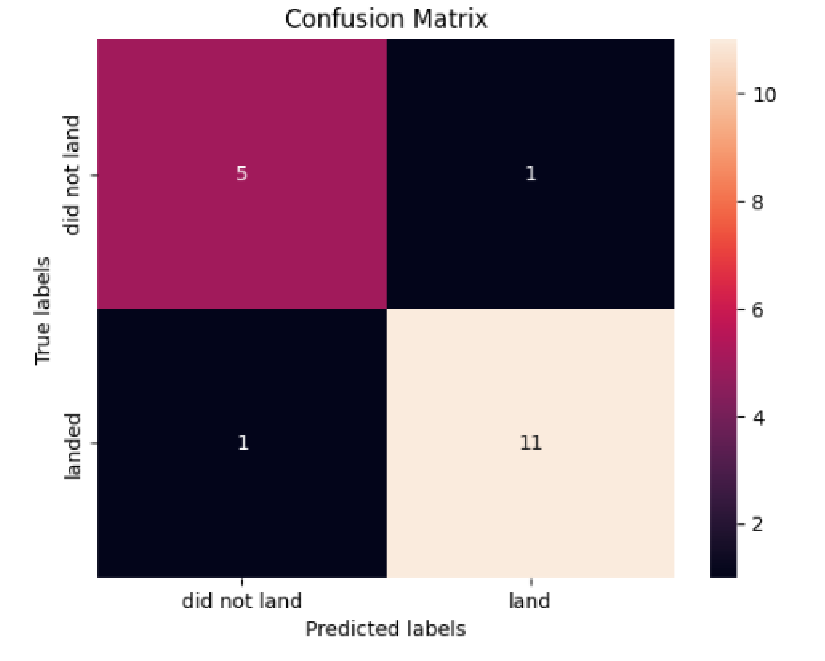
Predictive Analysis (Classification)

Classification Accuracy



- In the evaluation of the test dataset, the decision tree classifier emerged as the best model, achieving an impressive accuracy rate of 88.88%. This performance underscores the model's effectiveness in accurately predicting outcomes based on the provided features.

Confusion Matrix



The matrix is divided into four quadrants:

- True Negatives (Top-Left Quadrant): The number 5 represents the instances where the model correctly predicted the negative class—in this case, that the rocket did not land.
- False Positives (Top-Right Quadrant): The number 1 represents the instances where the model incorrectly predicted the positive class. That means the model predicted the rocket would land when, in fact, it did not.
- False Negatives (Bottom-Left Quadrant): The number 1 here indicates the instances where the model incorrectly predicted the negative class. This means the model predicted the rocket would not land when it actually did.
- True Positives (Bottom-Right Quadrant): The number 11 shows the instances where the model correctly predicted the positive class—meaning it predicted the rocket would land, and it did.

Conclusions

- Analyzing the trend over time, it's apparent that the frequency of successful launches has improved, with recent figures surpassing an 80% success threshold.
- Specific orbital trajectories, namely SSO, HEO, GEO, and ES-L1, demonstrate impeccable launch records, achieving a 100% success metric.
- Proximity assessments reveal launch pads are strategically positioned near essential infrastructure like railways and highways, and within reach of coastal lines, yet maintain a considerable distance from urban developments.
- The KSLC-39A launch complex stands out for its notable record of successful launches, leading the pack in terms of both volume and proportion of successes.
- There's a discernible pattern where launches carrying lighter payloads exhibit a higher likelihood of success compared to those tasked with heavier payloads.
- The decision tree classifier emerged as the best model, achieving an impressive accuracy rate of 88.88%.

Thank you!