IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SARA F.-DANS
OCTOBER 2024

Outline

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

Executive Summary

- SpaceX launch data analysis was conducted to predict first stage landing success. The project combined data from multiple sources, including the SpaceX API and Wikipedia, to create a comprehensive dataset. The analysis involved various techniques:
  - Data Processing and Exploration
  - Machine Learning Approach
  - Multiple algorithms were employed, including: Support Vector Machines (SVM), Classification Trees, Logistic Regression and K Nearest Neighbors.
  - Model Evaluation: the performance of each model was assessed using accuracy scores, allowing for comparison of their predictive capabilities. This comprehensive approach aimed to develop a robust prediction system for SpaceX's first stage landing outcomes
- The machine learning models demonstrated consistent performance in predicting SpaceX's first stage landing outcomes.
  - Model Performance
    - Peak Accuracy: The highest success rate observed across models was approximately 83.33%.
    - Test Set Validation: This accuracy level was maintained when evaluating predictions on the designated test set.
  - Prediction Bias
    - Overestimation Tendency: All models showed a propensity to overpredict successful landings.
    - Balanced Interpretation: While the 83.3% accuracy is impressive, the consistent overestimation of successes suggests room for model refinement.

Introduction

- SpaceX is a rocket launch company that charges $62 million per launch, significantly lower than its competitors, who typically charge $165 million or more. This substantial cost advantage is largely due to SpaceX's practice of landing and reusing the first stage of its rockets for future launches.

- We aim to determine whether we can predict the success of a launch by training a machine learning model to forecast successful Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected the data from SpaceXAPI and SpaceX Wikipedia Page
- Perform data wrangling
  - The data was stored in a pandas DataFrame. Falcon 1 data was excluded from the analysis. NULL values were replaced with the corresponding column means. Additionally, a binary 'Outcome' column was generated to represent launch success.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree, and K nearest neighbor models were built and tuned using a grid search technique and a training set of data. Models were evaluated using a test data set.

Data Collection

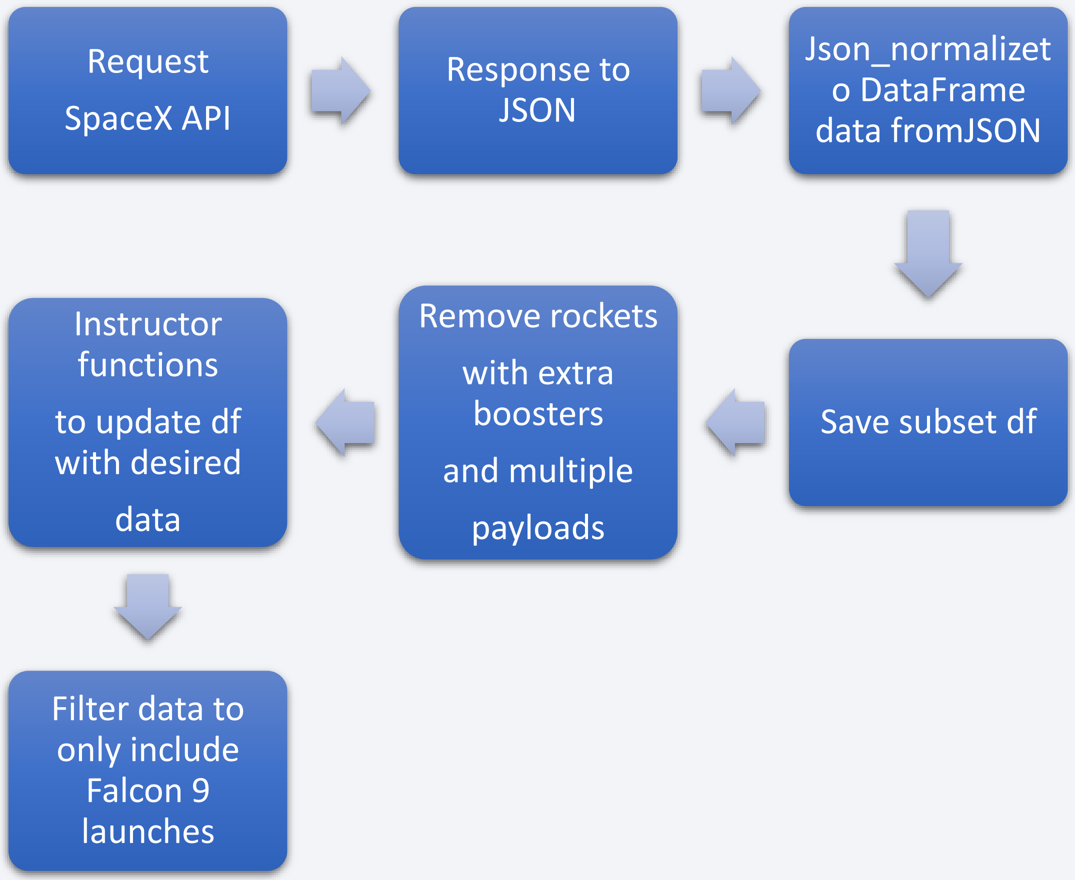
The data collection and processing for the SpaceX launch analysis involved a comprehensive approach combining API requests and web scraping techniques. Here's an overview of the process:

- API Data Collection
  - The SpaceX API was queried using the Python requests library
  - JSON responses were parsed and normalized into pandas DataFrames
  - Data was filtered to focus on specific attributes: Rockets, Launchpads, Payloads, Cores, Flight Numbers, and Dates
  - Exclusions were made for rockets with extra boosters and multiple payloads
  - Custom functions were applied to enrich the dataset with additional columns:
    - Booster Version, Payload Mass, Orbit, Launch Site, Outcome
    - Technical details: Flights, Grid Fins, Reused status, Legs, Block, Reused Count, Serial
    - Geographical data: Longitude, Latitude
  - The final dataset was narrowed down to Falcon 9 launches exclusively
- Web Scraping
  - SpaceX's Wikipedia entry was scraped to complement API data
  - BeautifulSoup with html5lib parser was used to extract information from HTML tables
  - Extracted data was structured into dictionaries and then converted to DataFrames

Data Collection – SpaceX API

Laboratory 1.A: Collecting the data (API)

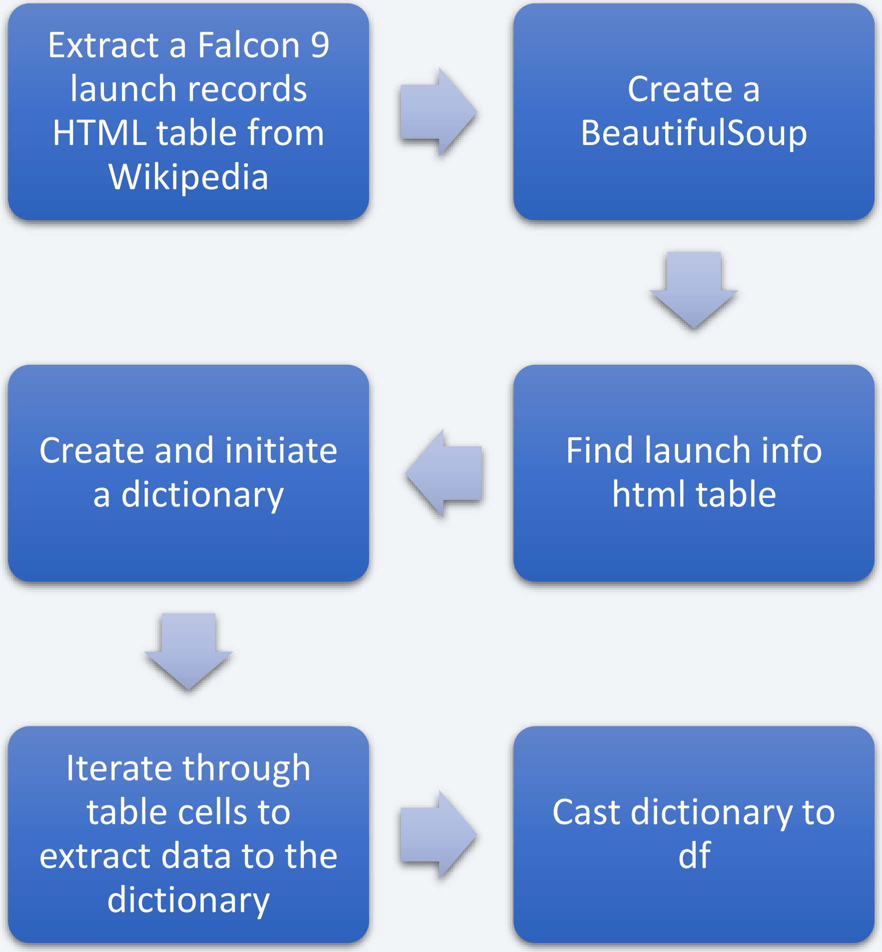
https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

Laboratory 1.B: Web scraping

https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/jupyter-labs-webscraping.ipynb



Data Wrangling

1. The SpaceX Falcon 9 launch data was loaded from a CSV file.
2. Exploratory data analysis was performed:
   - The percentage of missing values in each attribute was calculated.
   - The numerical and categorical columns were identified.
3. The number of launches at each launch site was calculated using the value_counts() method on the LaunchSite column.
4. The number and occurrence of each orbit type were determined by applying value_counts() on the Orbit column.
5. The number and occurrence of mission outcomes were analyzed by using value_counts() on the Outcome column.
6. A new column Class was created, where the value is 0 if the landing was unsuccessful and 1 if the landing was successful, based on the Outcome column.
7. The processed data was then exported to a new CSV file named dataset_part_2.csv.

[Laboratory 2](https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/labs-jupyter-spacex-Data%20wrangling%20-LAB2.ipynb): Data wrangling

https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/labs-jupyter-spacex-Data%20wrangling%20-LAB2.ipynb

Data Wrangling

Load SpaceX Falcon9 data

↓

Exploratory data analysis

↓

Calculate launches per site

↓

Calculate orbits and occurrences

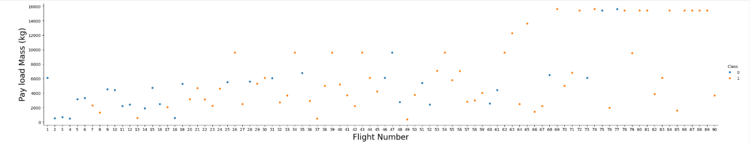
↓

Create landing success label

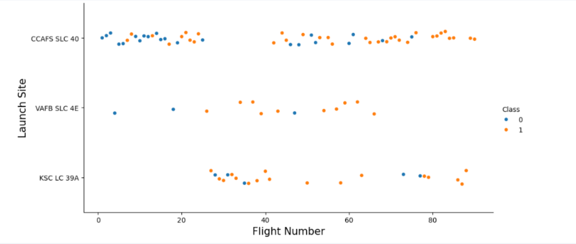
↓

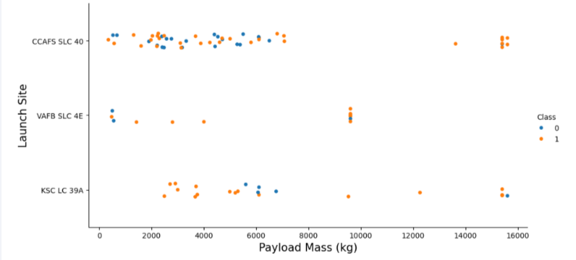
Export dataq to .csv

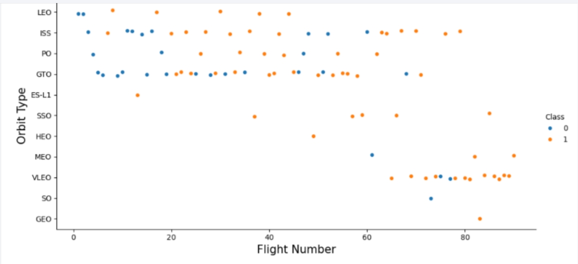
EDA with Data Visualization

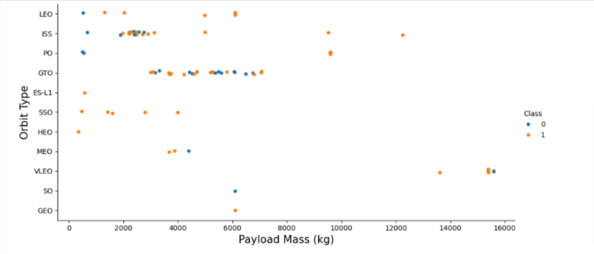


- Scatter plot of:
  - **Payload Mass vs Flight Number** to overlay the launch outcome.
  - **The relationship between Flight Number and Launch Site.** Examining the scatter plot a few patterns might stand out:
    1. Popular Launch Sites: KSC LC-39A, CCAFS SLC-40, and VAFB SLC-4E are frequently used by SpaceX for multiple launches.
    2. Success vs. Failure: The color coding (hue) shows successful (1) and failed (0) missions. This may reveal if certain sites have better success rates than others.
    3. Improving Over Time: Higher flight numbers might show increased success rates, suggesting SpaceX's technological improvements and learning from experience
  - **Relationship between Payload Mass and Launch Site.** We observe in the Payload Mass vs. Launch Site scatter plot that there are no rockets launched from the VAFB-SLC launch site for heavy payloads (greater than 10,000).
  - **Relationship between FlightNumber and Orbit type.** We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.
  - **Relationship between Payload Mass and Orbit type.** With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.







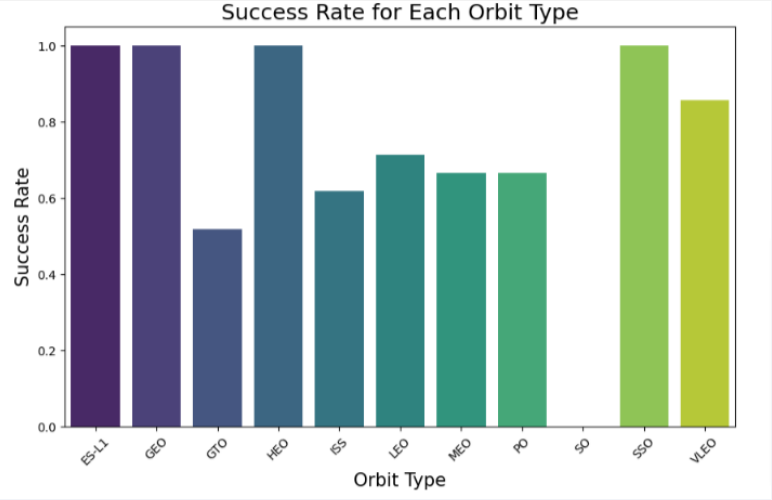


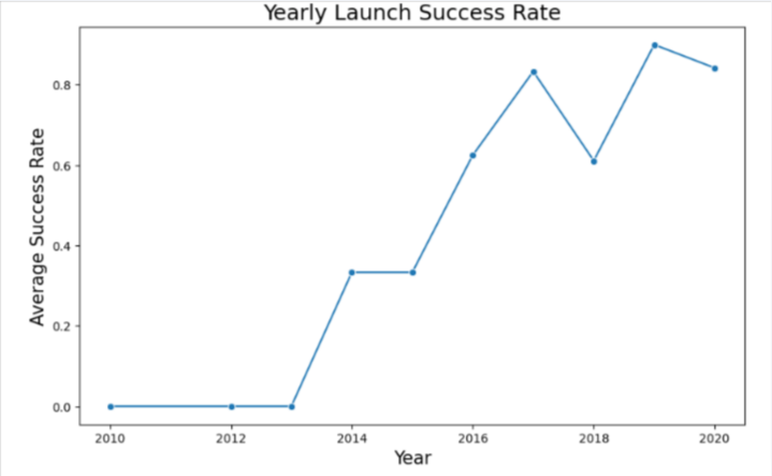
EDA with Data Visualization

- Bar Chart to **Visualize the relationship between success rate of each orbit type.** The data shows that GTO and LEO are among the more frequent orbit types for SpaceX launches. However, further analysis would be needed to determine if these orbits have higher success rates compared to other mission types.
- Line plot to **Visualize the launch success yearly trend.** We can observe that the success rate since 2013 kept increasing till 2020

- Laboratory 4: EDA with Data Visualization

https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/edadataviz.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 kg but less than 6000 kg.
- 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.¶

Laboratory 3: EDA with SQL

https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Each of these objects contributes to a better visualization and understanding of the map, providing additional details and a visual structure that facilitates the analysis of locations and distances:

- Markers: Added markers at launch site locations. Each marker includes a "popup" with the name of the launch site, allowing users to visually identify each location on the map. The reason for adding these markers is to highlight specific points of interest (in this case, launch sites) and make it easier to identify them on the interactive map.

- Circles (Circle and CircleMarker): Circles were placed around the launch sites with different radii, indicating areas of influence or interest around each site. Circles can also reflect proximity to landmarks or risk areas, which can be important in security or logistics studies. In addition, the "CircleMarkers" act as stylized markers that emphasize the exact location with an eye-catching visual design.

- Lines (Polylines): Lines were drawn from launch sites to specific reference locations. These lines are used to measure distances or represent routes, which is useful for assessing access or connection between key points in the context of trajectory analysis or launch logistics.

- Laboratory 5*: Interactive map with Folium

https://github.com/Yeira-black/Curso-IBM/blob/8bc459ff40ba103b50930e5d907ac9e6f060b4d3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The dashboard in spacex_dash_app.py includes the following plots and interactive components:

- Dropdown Menu for Launch Site Selection:
  - Allows users to select a specific SpaceX launch site or view data for all sites collectively.
  - Purpose: Provides flexibility in data visualization, enabling users to drill down into individual launch sites or analyze the data in aggregate.
- Pie Chart for Successful Launches:
  - Displays the total count of successful launches for each site when "All Sites" is selected. If a specific site is chosen, the pie chart shows the success versus failure counts for that site.
  - Purpose: This visualisation offers a quick, intuitive understanding of launch success rates across different sites or for a particular site, helping users evaluate SpaceX's operational performance at each location.

Build a Dashboard with Plotly Dash

- Payload Range Slider:
  - Allows users to filter data based on the payload mass range (in kilograms).
  - Purpose: This interactive filter enables users to examine how launch success correlates with payload mass, giving insights into the payload capacity performance across different missions.
- Scatter Plot for Payload vs. Launch Success:
  - Plots payload mass against launch success, with different colours representing booster versions. This chart updates based on the selected launch site and payload range.
  - Purpose: Highlights any correlation between payload mass and launch success, which can reveal operational patterns or limitations related to payload capacity.

These plots and interactions make the dashboard highly interactive and user-friendly, allowing users to explore and analyze SpaceX launch data from multiple perspectives

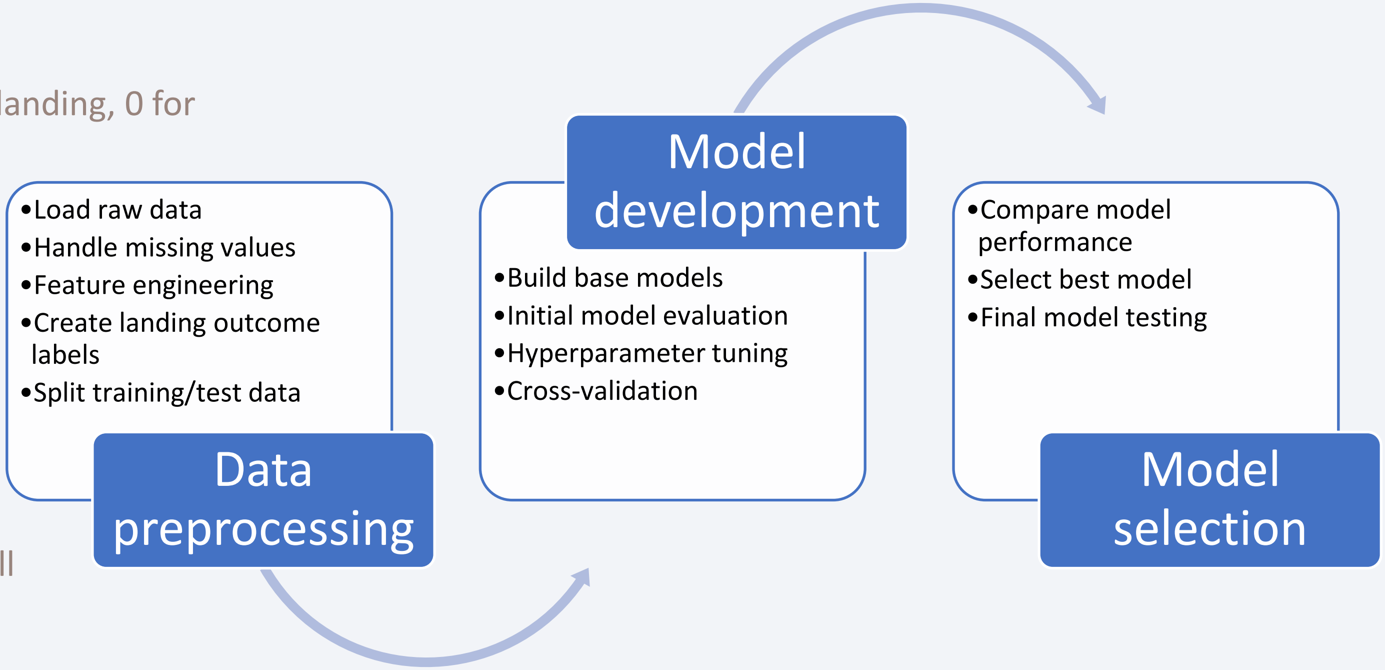
Laboratory 6: Dashboard with Plotly Dash

https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/spacex_dash_app.py

Predictive Analysis (Classification)

Key Phrases in the Model Development Process:

1. Data Preprocessing Phase:
    - Data loading and cleaning
    - Missing value analysis and handling
    - Feature engineering from launch data
    - Binary classification label creation (1 for successful landing, 0 for unsuccessful)
    - Dataset splitting into training and testing sets
2. Model Development Phase:
    - Initial model building with multiple algorithms
    - Base model evaluation using standard metrics
    - Hyperparameter optimization for each model type
    - Cross-validation to ensure model robustness
3. Model Selection Phase:
    - Performance comparison across models
    - Selection criteria based on accuracy, precision, recall
    - Final model validation on test set
    - Performance documentation and analysis



Predictive Analysis (Classification)

The process focused on creating a reliable binary classification model to predict whether the Falcon 9 first stage would land successfully. The key improvement steps included:

- Feature engineering to capture launch characteristics
- Iterative model testing and refinement
- Hyperparameter tuning for optimization
- Cross-validation to ensure reliability
- Final validation against held-out test data

This systematic approach ensures both model performance and reliability for predicting landing outcomes of future launches.

Laboratory 7: **Space X Falcon 9 First Stage Landing Prediction**

https://github.com/Yeira-black/Curso-IBM/blob/3a77db0d3d2a75f7bbd3b9ab44cd7312825fa42d/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory Data Analysis (EDA) Insights

1. Temporal Trend in Landing Success:
   The analysis revealed that the likelihood of successful landings increased over time, suggesting a maturation of the technology used in SpaceX launches.
2. Payload Quantification:
   EDA allowed for the quantification of the total payload successfully delivered to space across all launches.
3. Launch Site Analysis:
   Visualizations were created to show the distribution of launches across different sites and their respective success rates. This provided additional insights into site-specific performance.
4. Interactive Analytics:
   - Success rates were plotted, likely showing trends over time or across different variables.
   - A scatter plot was created to visualize the relationship between payload mass and landing success, both overall and for individual launch sites.

Predictive Analysis

The predictive analysis aimed to forecast landing success. Key points include:Multiple machine learning models were employed, likely including SVM, Classification Trees, and Logistic Regression (based on earlier mentions in the document).

- No single model significantly outperformed the others in predicting landing success.
- The models achieved an accuracy rate of approximately 83.33% on the test set.
- All models showed a tendency to overpredict successful landings, indicating a potential bias in the predictions.

Results

Additional Observations

1. Launch Sites: The data included launches from multiple sites, with CCAFS SLC 40 being the most frequent (55 launches), followed by KSC LC 39A (22 launches) and VAFB SLC 4E (13 launches).
2. Orbit Types: Various orbit types were targeted, with GTO (27 launches) and ISS (21 launches) being the most common.
3. Landing Outcomes: The most frequent successful outcome was "True ASDS" (41 instances), indicating successful landings on drone ships.
4. Success Rate: The overall success rate for landings was calculated to be approximately 66.67% (df["Class"].mean()).

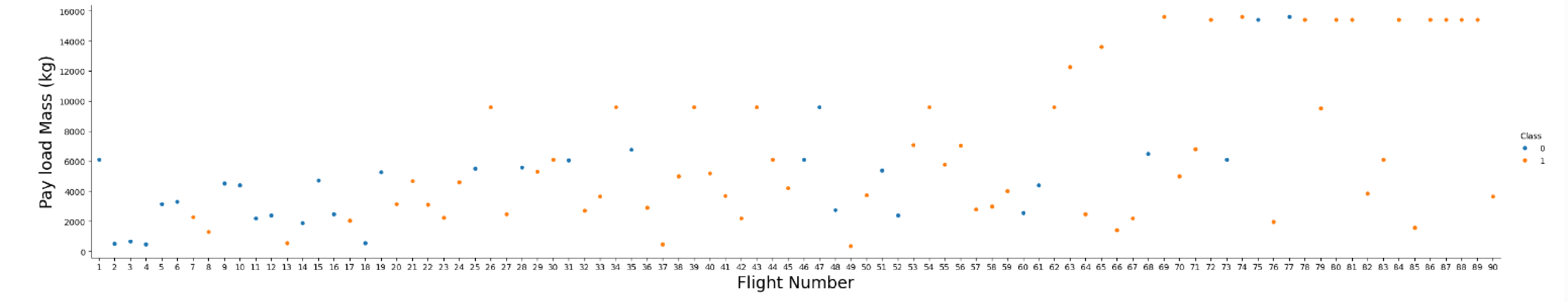
This comprehensive analysis provided valuable insights into SpaceX's launch operations, technological progress, and the factors influencing landing success. The predictive models, while not perfect, offer a useful tool for estimating the likelihood of successful landings for future missions.

Section 2

Insights drawn from EDA

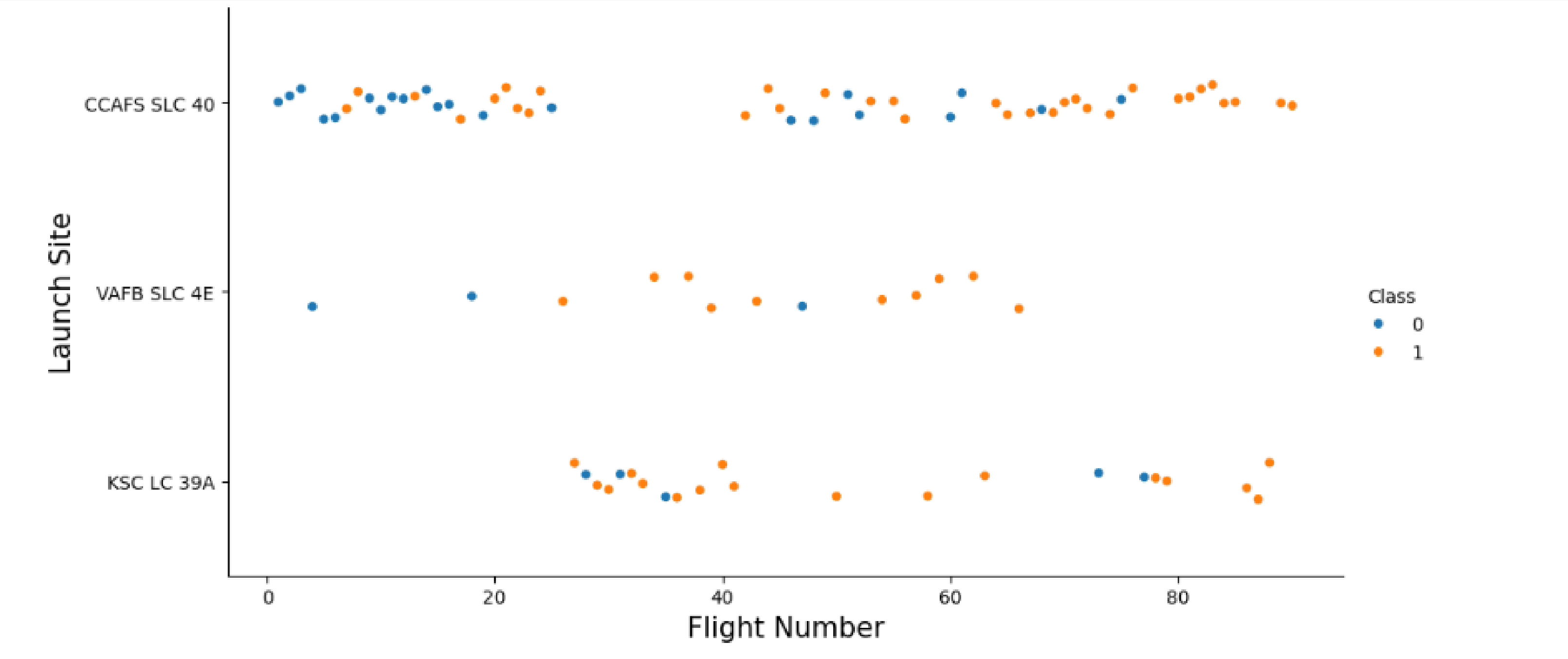
Flight Number vs. Launch Site

Early flight numbers experienced a higher rate of failures in landing the first stage. As the flight numbers increased, there was a corresponding rise in average payloads and a greater number of successful landings.



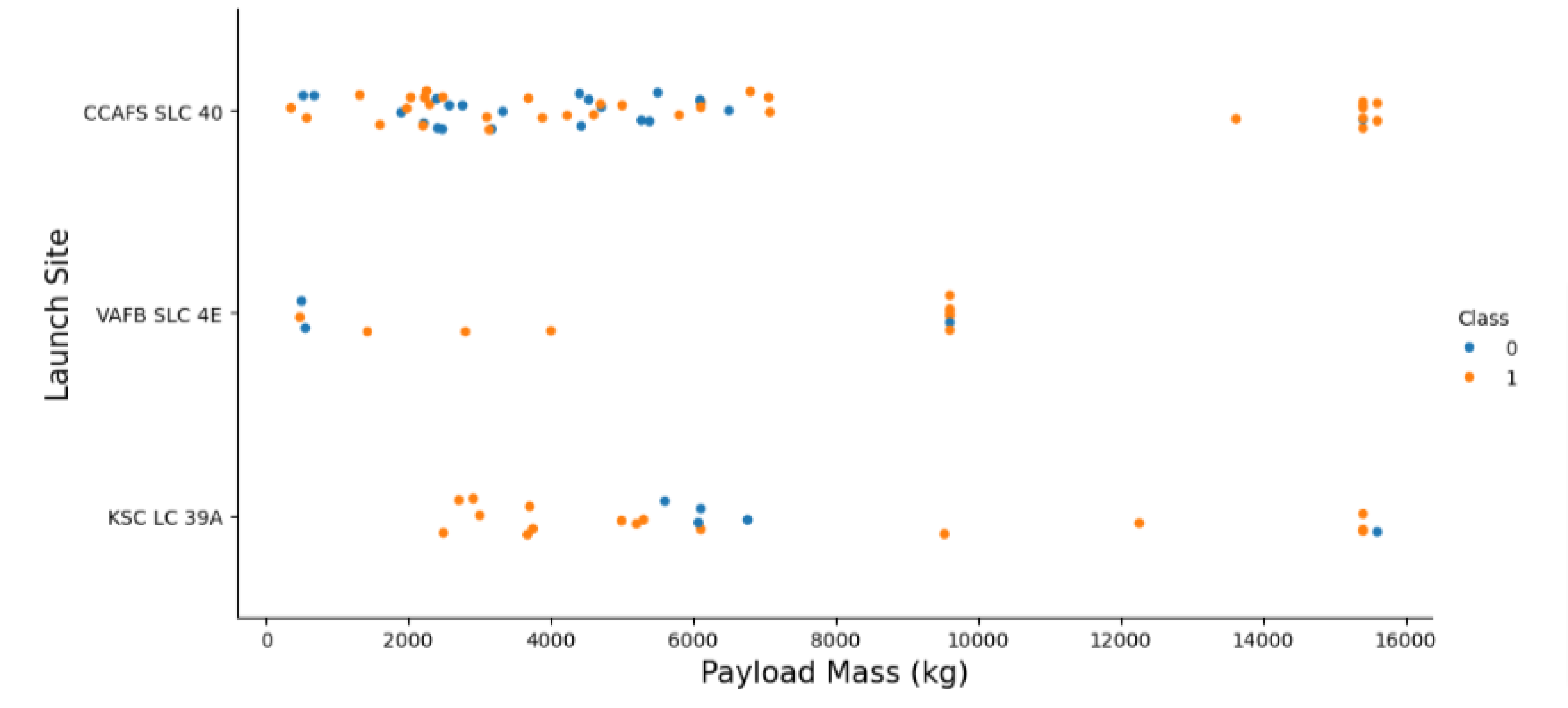
Flight Number vs. Launch Site

- Clusters at Specific Sites: Certain launch sites like KSC LC-39A, CCAFS SLC-40, and VAFB SLC-4E seem to host multiple launches, indicating these are frequently used by SpaceX.
- Class Distribution: The hue (Class) indicates whether the mission was a success (1) or failure (0). The pattern of successes and failures across different launch sites and flight numbers might reveal if some sites are historically more successful than others.
- Flight Progression: As the flight numbers increase, the launches might show a progression in the success rate, potentially indicating SpaceX's improvement and learning curve.



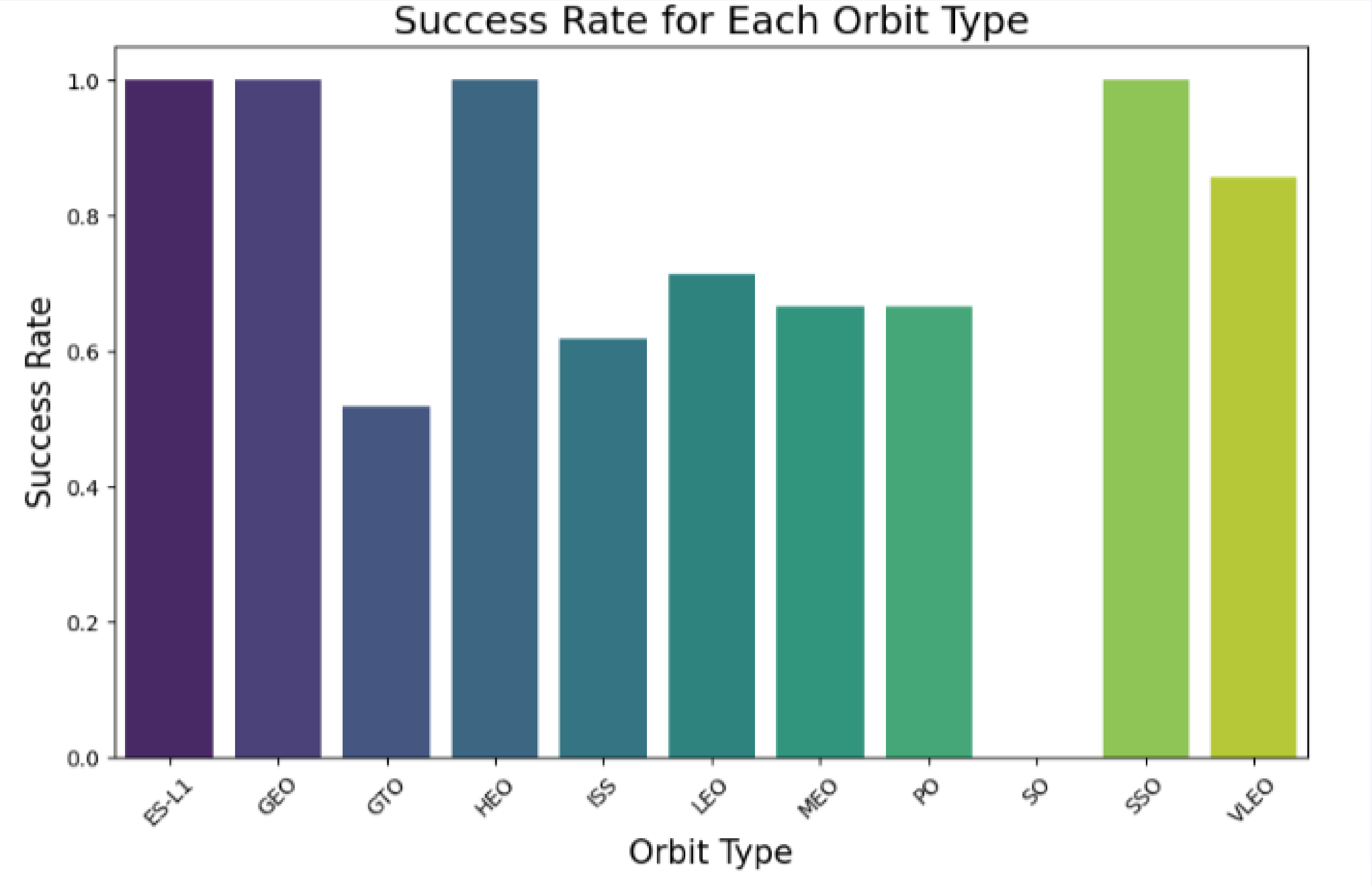
Payload vs. Launch Site

- There is no real trends between payload and launch sites.
- KSC had no launches below 2,000 kg
- VAFB had no launches above 10,000 kg
- CCAFS had no launches between 8,000 and 12,000 kg



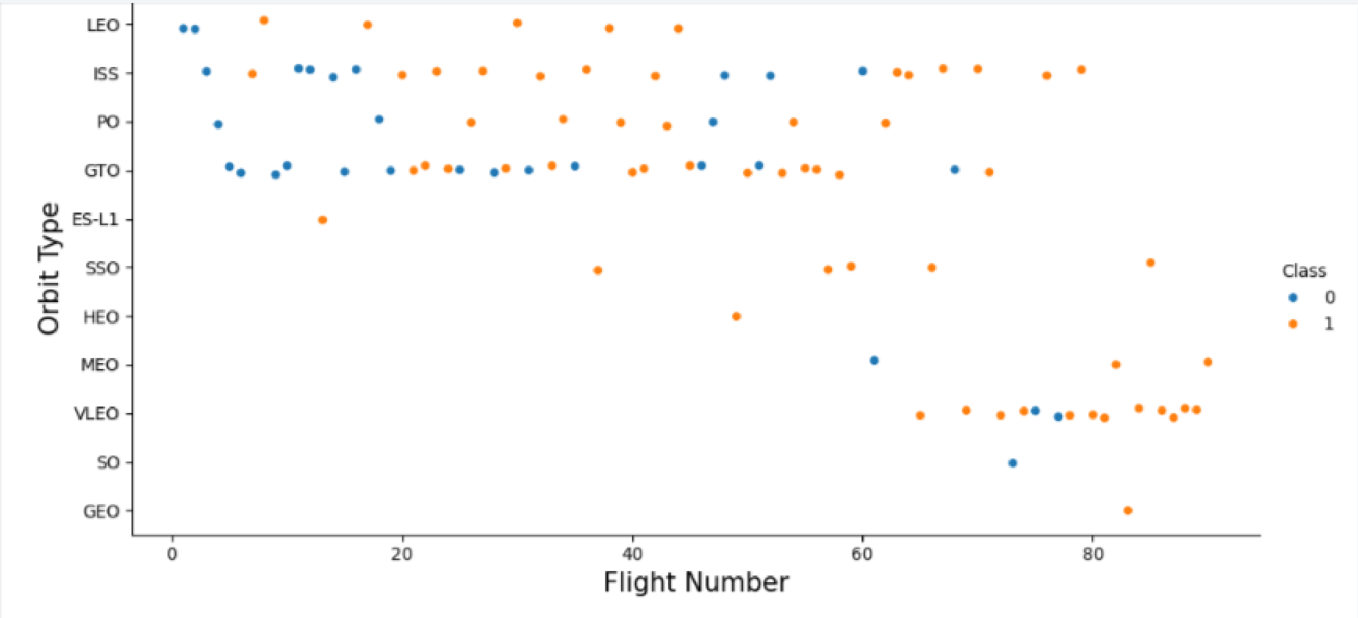
Success Rate vs. Orbit Type

The data shows that GTO and LEO are among the more frequent orbit types for SpaceX launches. However, further analysis would be needed to determine if these orbits have higher success rates compared to other mission types.

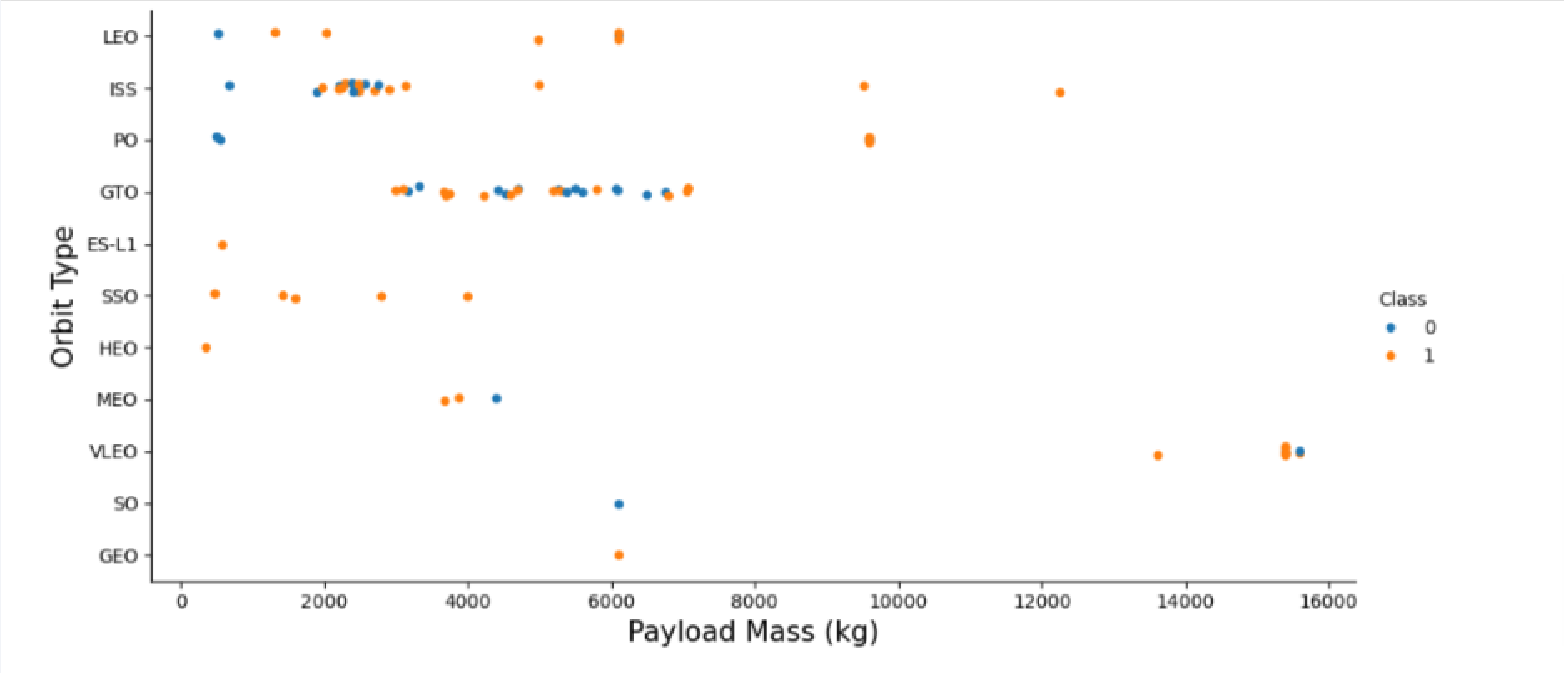


Flight Number vs. Orbit Type

We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



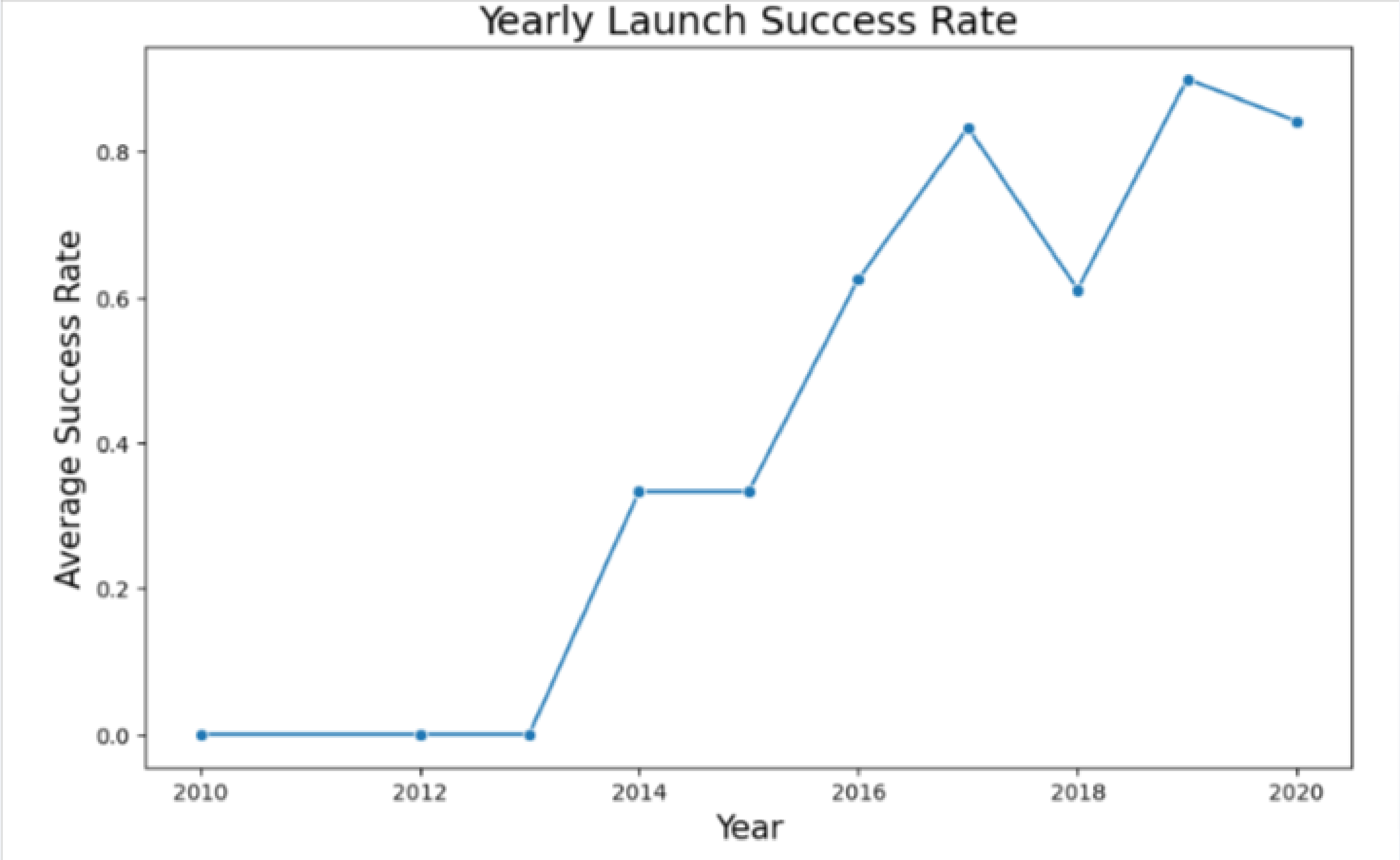
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites:

```
In [11]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE

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

Out[11]:

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

- Cape Canaveral is split into LC-40 and SLC-40
- VAFB stands for Vandenberg Air Force Base
- KSC stands for Kennedy Space Center

Launch Site Names Begin with 'CCA'

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

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

```
[13]: %sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass" \
          FROM SPACEXTABLE \
          WHERE Customer LIKE 'NASA (CRS)%'

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

[13]:

Total Payload Mass
48213

Average Payload Mass by F9 v1.1

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

```
[14]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as "Average Payload Mass" \
          FROM SPACEXTABLE \
          WHERE Booster_Version = 'F9 v1.1'

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

[14]:

Average Payload Mass
2928.4

Average payload mass of F9 v1.1 is on the low end of our payload mass range.

First Successful Ground Landing Date

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

```
[15]: %sql SELECT MIN(Date) as "First Successful Landing Date" \
          FROM SPACEXTABLE \
          WHERE "Landing_Outcome" = 'Success (ground pad)'

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

[15]:

First Successful Landing Date
2015-12-22

First successful landing outcome on ground pad was not until the end of 2015.

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

```
[16]: %sql SELECT DISTINCT Booster_Version \
          FROM SPACEXTABLE \
          WHERE "Landing_Outcome" = 'Success (drone ship)' \
          AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

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

[16]:

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

- This query returns a count of each misión outcome.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.

```
[17]: %sql SELECT Mission_Outcome, COUNT(*) as Count \
          FROM SPACEXTABLE \
          GROUP BY Mission_Outcome

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

[17]:

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

Boosters Carried Maximum Payload

```
[18]: %sql SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_ \
          FROM SPACEXTABLE \
          WHERE PAYLOAD_MASS__KG_ = ( \
              SELECT MAX(PAYLOAD_MASS__KG_) \
              FROM SPACEXTABLE \
          )
```

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

[18]:

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

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

- All the boosters were derivatives of the F9 B5 booster

2015 Launch Records

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

- There were two currencies in 2015

```
[20]: %sql SELECT \
          CASE substr(Date, 6, 2) \
              WHEN '01' THEN 'January' \
              WHEN '02' THEN 'February' \
              WHEN '03' THEN 'March' \
              WHEN '04' THEN 'April' \
              WHEN '05' THEN 'May' \
              WHEN '06' THEN 'June' \
              WHEN '07' THEN 'July' \
              WHEN '08' THEN 'August' \
              WHEN '09' THEN 'September' \
              WHEN '10' THEN 'October' \
              WHEN '11' THEN 'November' \
              WHEN '12' THEN 'December' \
          END as Month, \
          "Landing_Outcome", \
          Booster_Version, \
          Launch_Site \
          FROM SPACEXTABLE \
          WHERE substr(Date, 0, 5) = '2015' \
          AND "Landing_Outcome" LIKE '%failure%' \
          AND "Landing_Outcome" LIKE '%drone ship%'

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

[20]:

Month	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

[22]:
```
%sql SELECT \
        `Landing_Outcome`, \
        COUNT(*) AS `Count` \
    FROM  \
        SPACEXTABLE  \
    WHERE \
        Date BETWEEN '2010-06-04' AND '2017-03-20'  \
    GROUP BY \
        `Landing_Outcome`  \
    ORDER BY  \
        `Count` DESC;
```

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

[22]:

Landing_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

- 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

- In this time frame, they were mostly successes

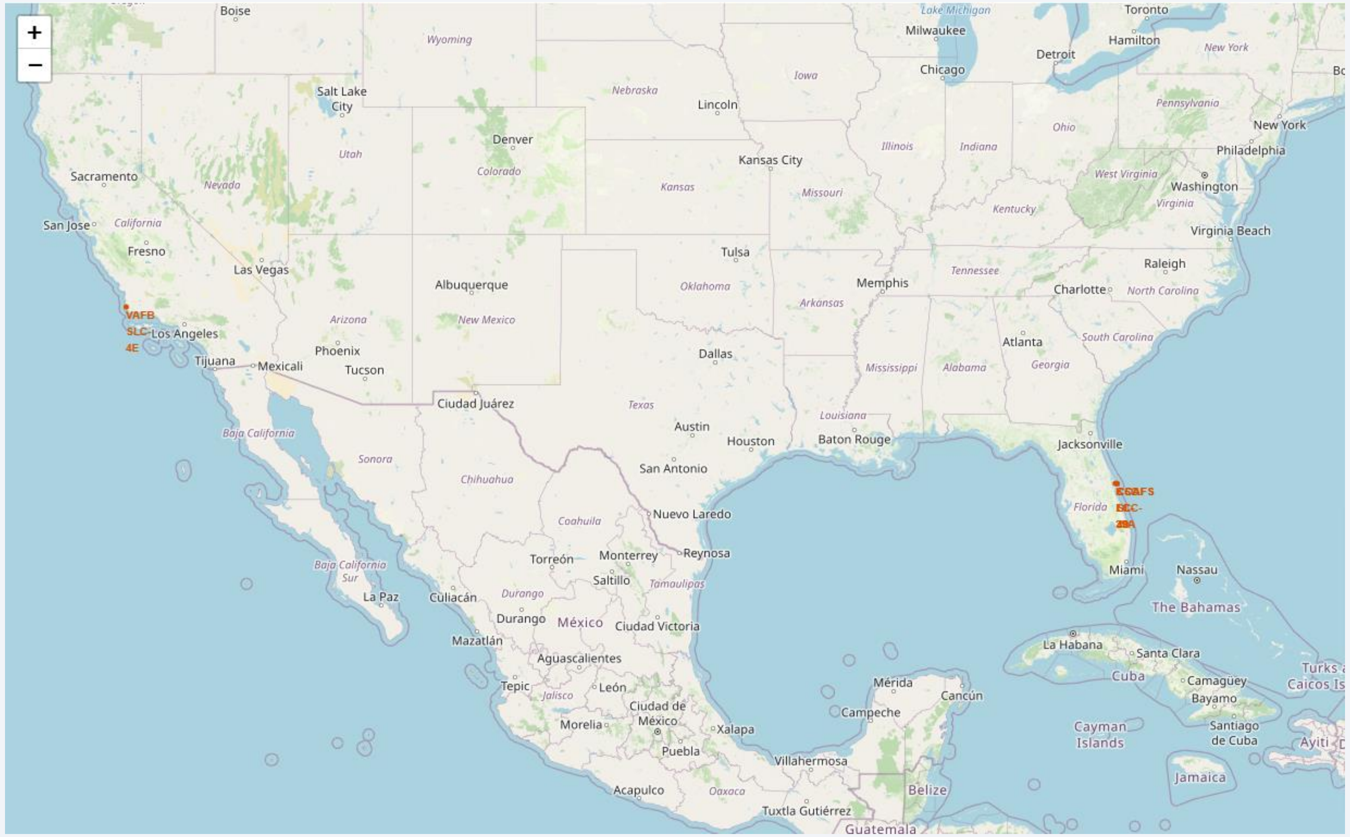
Section 3

Launch Sites Proximities Analysis

Map of all launch sites

By analyzing the data and the map, we can reach the following conclusions:

- In relation to proximity to the Equator:
  - Not all launch sites are near the Equator. Most sites are located between 28°N and 34°N latitude
- In relation to proximity to the coast:
  - Many launch sites are indeed located near coastlines for safety reasons (to avoid flying over populated areas) and to allow expended rocket stages to fall into the ocean.



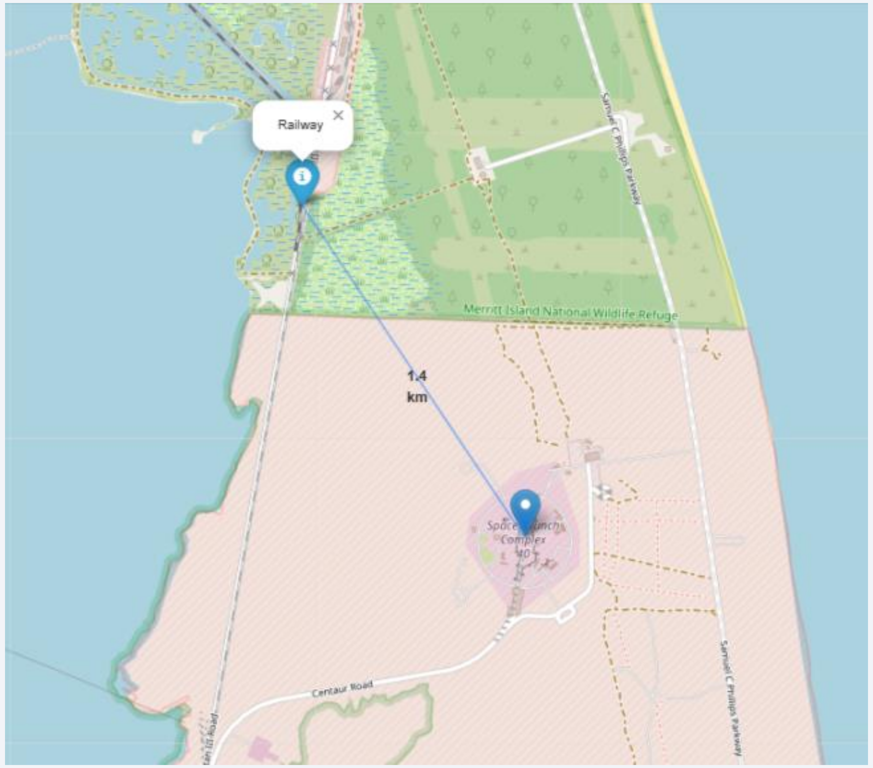
Map of success/failed launches

Map Overview

- Launch Sites
  - All launches have been plotted on the map and are identified with a label for easy reference.
- Clustering
  - Clusters have been created for comparable launch sites to facilitate analysis and comparison.
- Landing Indicators
  - Red Mark: Indicates a failed landing.
  - Green Mark: Indicates a successful landing.

Map of Distances from a Launch Site to Its Proximities

- Proximity to railways:
  Launch sites are typically not in close proximity to railways. Railways are not a critical factor in launch site selection.
- Proximity to highways:
  Launch sites often have some proximity to highways for logistical reasons, but they are usually not immediately adjacent to major roads. A balance is struck between accessibility and safety/security concerns.
- Proximity to coastline:
  Many launch sites are located near coastlines. This is primarily for safety reasons, as it allows rockets to fly over water rather than populated areas during the initial stages of flight.
- Distance from cities:
  Launch sites generally maintain a significant distance from densely populated areas. This is crucial for safety, noise mitigation, and to minimize risks in case of launch failures.



Map of Distances from a Launch Site to Its Proximities

Typical findings for launch site locations:

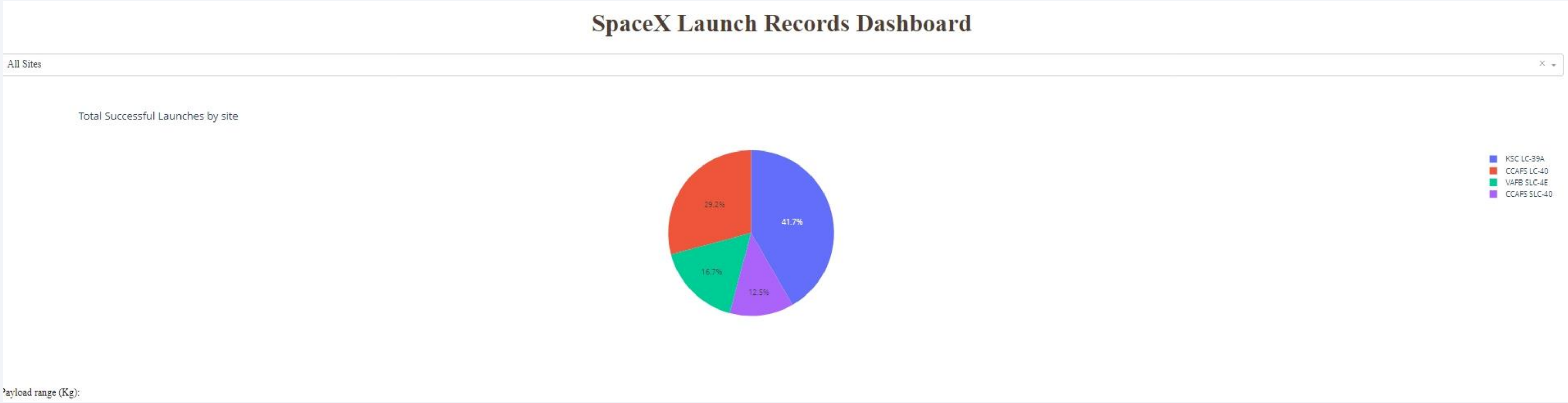
- Coastal preference: Many major launch sites, like Cape Canaveral in Florida or Vandenberg Air Force Base in California, are located on coastlines. This allows for over-water flight paths, reducing risks to populated areas.
- Remote locations: Launch sites are often in relatively remote areas, away from major population centers. This is for safety reasons and to minimize noise and other impacts on communities.
- Large buffer zones: Launch sites typically have extensive buffer zones around them to ensure safety and security.
- Accessibility balance: While not immediately adjacent to major transportation hubs, launch sites need to be reasonably accessible for the transport of personnel, equipment, and rocket components.
- Geographical considerations: Launch sites are often located closer to the equator when possible, as this provides a slight speed boost for launches due to the Earth's rotation

Section 4

Build a Dashboard with Plotly Dash

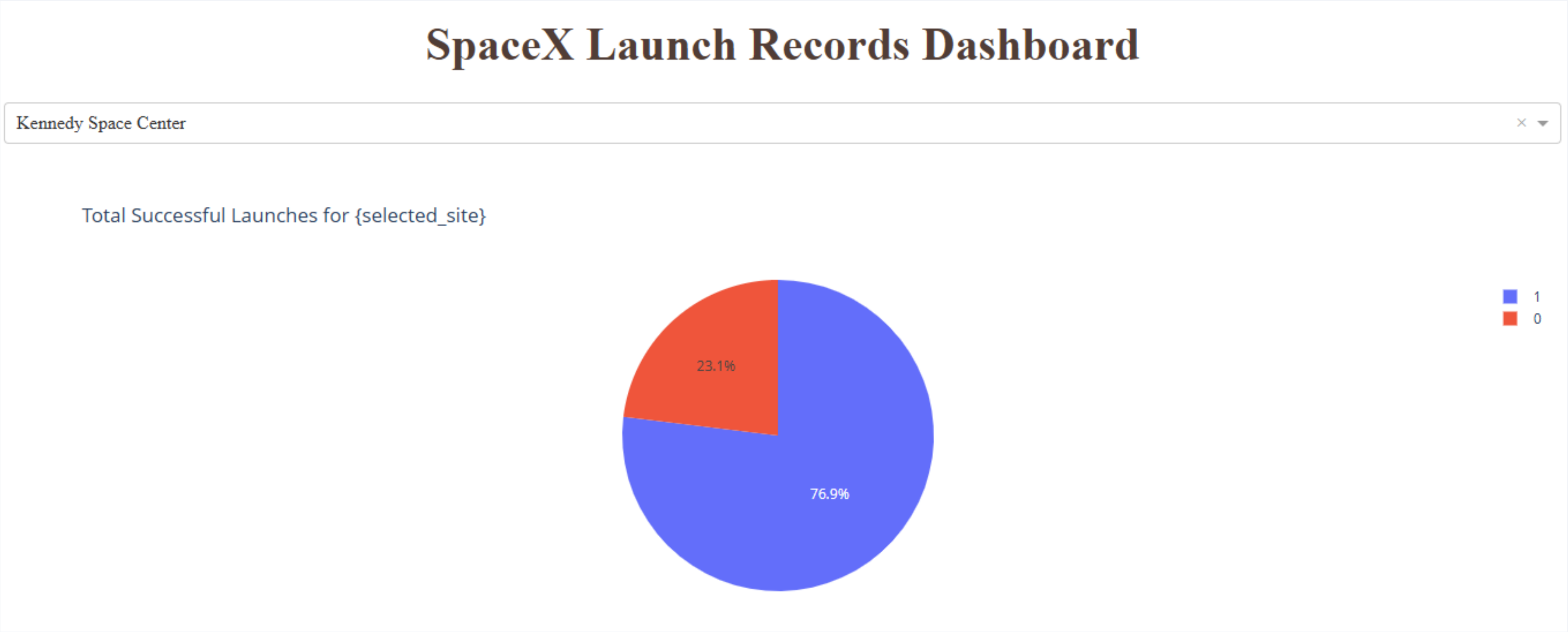
Successful Launches by Site

- Kennedy Space Center had 41.7% of all successful launches, which was the greatest
- Cape Canaveral SLC-40 had 12.5% of all successful launches, which was the least



Highest launch success ratio

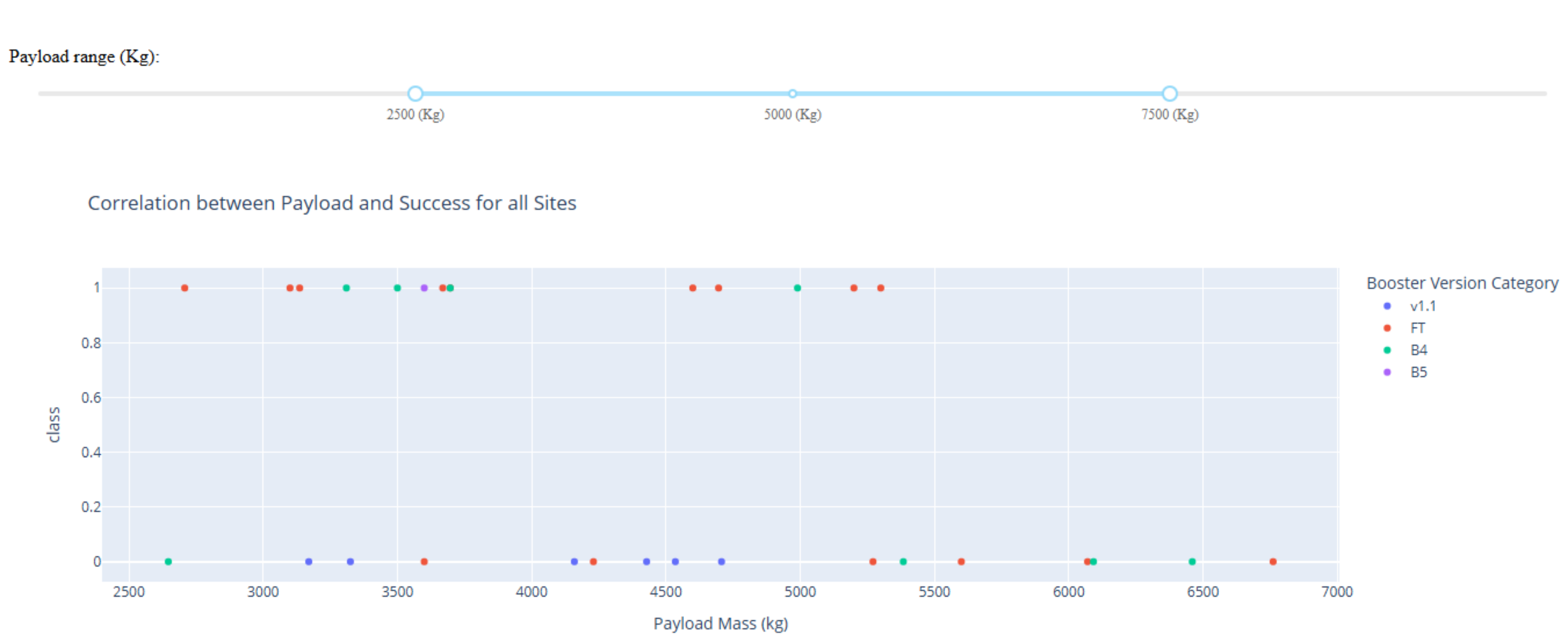
Kennedy Space Center (KSC) boasts the highest success-to-failure rate among all launch sites, with a remarkable 76.9% of its launches being successful.



Payload and Launch Outcome

The interactive slider feature enabled users to select specific payload weight ranges for analysis. Upon examination, it became evident that:

1. Launches carrying the heaviest payloads generally demonstrated a higher success rate.
2. For launches in the medium to light payload categories, no clear pattern or trend in success rates could be discerned.

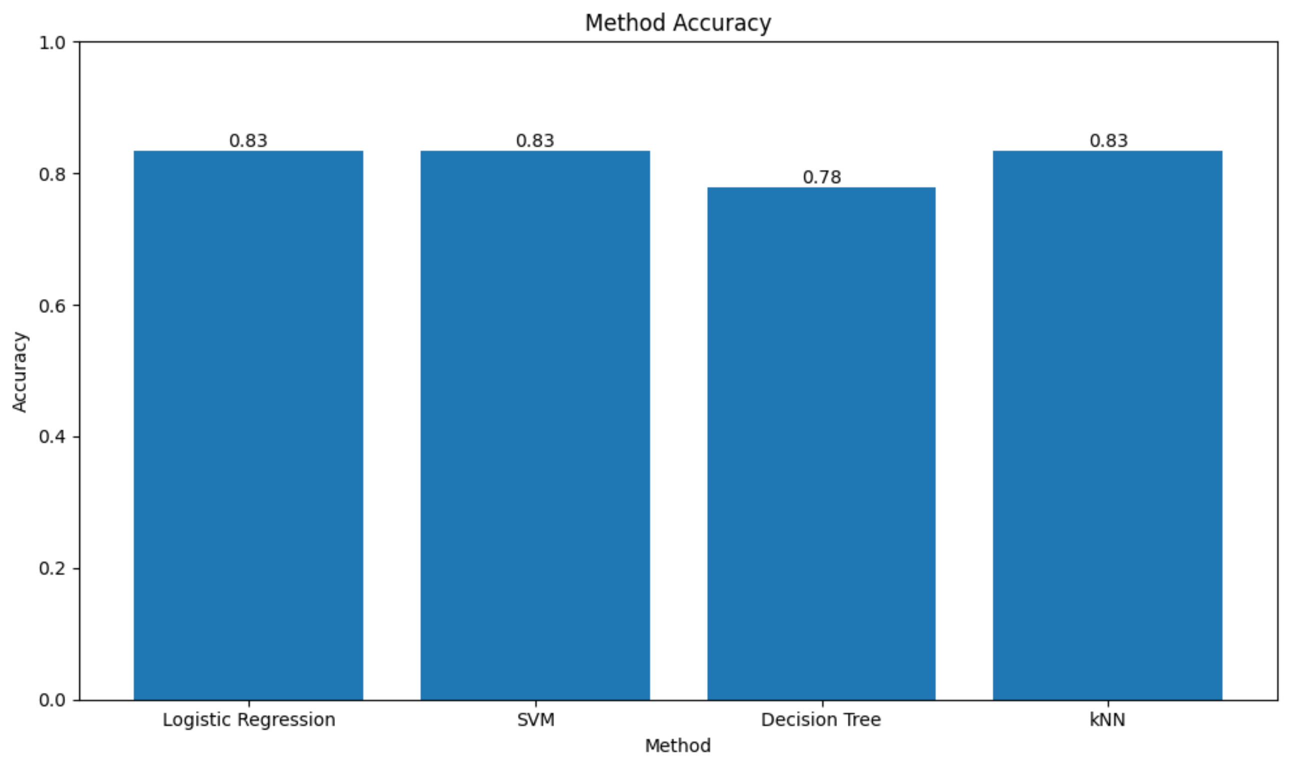


Section 5

Predictive Analysis (Classification)

Classification Accuracy

All models showed a similar accuracy of 83.33% on the test set, although it is important to note that the sample size was small, with only 18 examples. This can lead to significant variability in the results, as observed in the Decision Tree Classifier model, which performed worse than the other methods. It is recommended to reevaluate the models in the future with a larger dataset to obtain more reliable results.

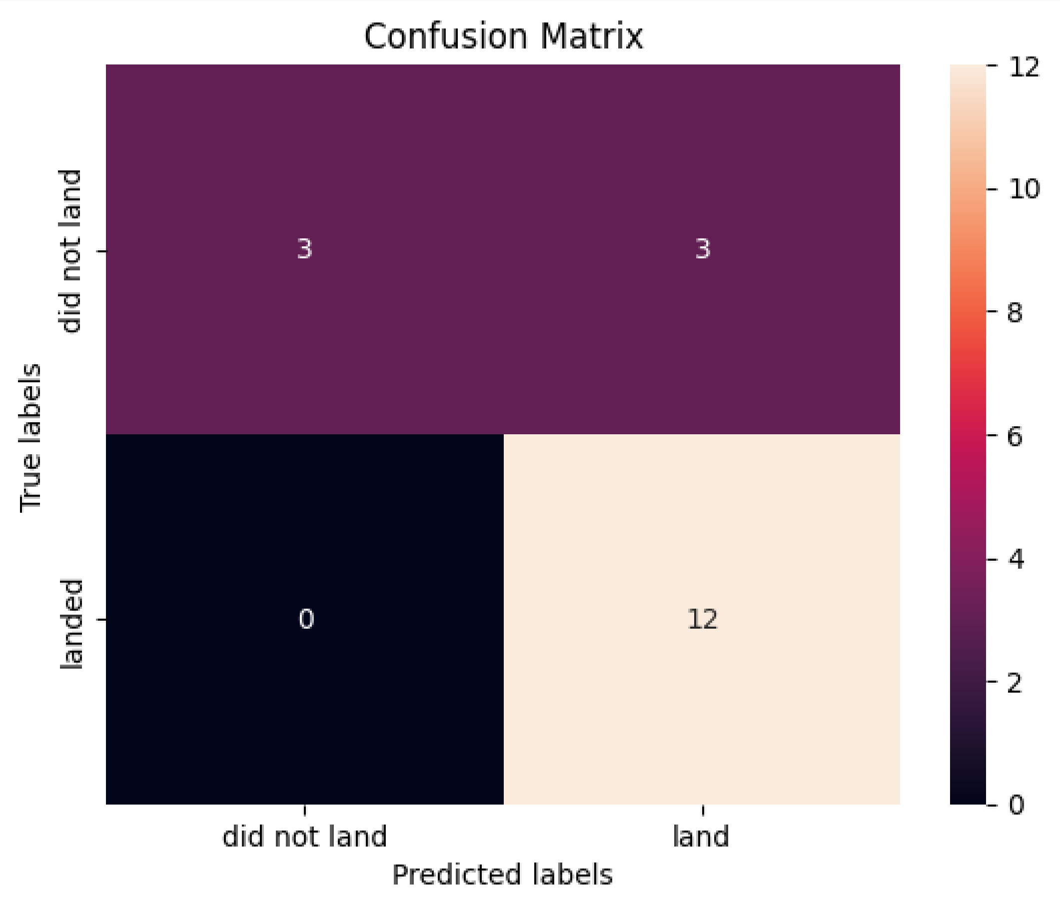


Confusion Matrix

Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms demonstrated identical performance in their confusion matrices:

- Accurate predictions: These models correctly classified 15 out of the 18 test cases.
- Misclassifications: They all made the same error, producing 3 false positives out of the 18 total cases.

This consistency across different classification methods suggests a similar pattern in how these algorithms interpreted and categorized the given data set.



Conclusions

1. SpaceY has commissioned the development of a machine learning model to predict the successful landing of Stage 1 rockets, aiming to save approximately $100 million per launch.

2. The model was developed using data collected from SpaceX's public API and web scraping of the SpaceX Wikipedia page.

3. Labeled data was created and stored in a DB2 SQL database, facilitating efficient data management.

4. A visualization dashboard was created to present the data and model results effectively.

5. The machine learning model achieved an accuracy of 83%.

6. SpaceY's Allon Mask can utilize this model to assess the likelihood of a successful Stage 1 landing before launch, aiding in decision-making regarding whether to proceed with a launch.

Conclusions

7. Historical analysis indicates that SpaceX's launch success rates have improved over time.

8. Heavier payloads tend to have higher success rates and are typically launched later.

9. There is no correlation between launch success and launch site, although sites are strategically located near the equator.

10. Certain orbits demonstrate higher success rates for launches.

11. Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) models all achieved an accuracy of 83% in predicting launch outcomes.

12. To further enhance the model's performance and accuracy, it is advisable to collect additional data for future analysis and refinement.

Appendix

GitHub repository url:

https://github.com/Yeira-black/Curso-IBM.git

Special Thanks to All Instructors

Thank you!