IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jay Nadz G. Libardo
May 14, 2025

Outline

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

Executive Summary

- This capstone project aims to assess and predict the likelihood of successful landings of the Falcon 9 first stage, a key cost-saving factor in SpaceX's launch strategy.
- SpaceX offers orbital rocket launches at approximately $62 million, a significantly lower cost than traditional providers (upward of $165 million per launch), primarily due to its ability to recycle first-stage boosters.
- By analyzing historical launch data, this project offers strategic insights that can aid potential competitors in evaluating SpaceX's performance and formulating competitive bids.

Executive Summary

Methodologies

- Data Collection & Preparation: Historical SpaceX launch data was cleaned and preprocessed, with particular focus on variables such as Launch Site, Payload Mass, Booster Version, and Mission Outcome (Success or Failure).
- Exploratory Analysis: Visualization techniques were employed to identify patterns and correlations affecting landing success. We examined success rates by site, payload weight, and booster type.
- Interactive Dashboard Development: An interactive dashboard was created using Plotly Dash to allow dynamic exploration of:

1. Launch success rates by site
2. Impact of payload mass on mission outcome
3. Booster version influence on landing success

- Real-Time Filtering: The dashboard features dropdowns and sliders to filter results by site and payload, enabling business users and analysts to visualize performance scenarios instantly.

Executive Summary

Key Results

- Launch Site Analysis: While CCAFS LC-40 had the highest number of launches, KSC LC-39A showed a consistently higher success rate for landings — a strategic consideration for assessing future launch reliability.
- Payload Impact: Launches with payloads in the 2000–4000 kg range tended to have higher landing success, suggesting a potential operational sweet spot.
- Orbit ISS has the highest success rate, followed by Orbit GTO, making them the best options for launch paths.
- Booster Version Performance: Newer booster versions (FT) showed significantly higher success rates, reinforcing the importance of technological iteration in reusability.
- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.
- Cost Implication: Since landing success is closely tied to cost efficiency, being able to predict landing success based on mission parameters provides a valuable forecasting tool. Competing aerospace providers can use these insights to estimate effective pricing models when bidding against SpaceX.

Introduction

Project Background and Context

- In the rapidly evolving commercial space industry, cost-efficiency and reliability are critical drivers of competitive advantage. SpaceX has revolutionized orbital launch services with its reusable Falcon 9 rocket, claiming a launch price of $62 million—a fraction of the traditional market rate of over $165 million. This cost reduction is primarily due to the successful recovery and reuse of the Falcon 9's first stage.
- The ability to predict whether a first stage will land successfully has significant implications not only for SpaceX's internal operations and mission planning, but also for external stakeholders, including satellite companies, insurers, investors, and competing launch providers. Understanding the patterns and factors that influence landing success can help these parties make informed strategic, financial, and risk management decisions.

Introduction

Problems We Aim to Solve

This project seeks to explore and answer the following key questions:

- What are the major factors influencing the success of Falcon 9 first stage landings?
    - Do specific launch sites, payload weights, or booster versions have a measurable impact on mission outcomes?
- Is there a correlation between payload mass and landing success?
    - Can we identify an optimal payload range for successful first stage recovery?
- How do different booster versions perform over time?
    - Are newer versions more reliable, and does this contribute to increased mission cost-efficiency?
- Can we build a tool that allows stakeholders to interactively explore these patterns?
    - What insights can an interactive dashboard provide to business users, such as alternate launch providers considering bidding against SpaceX?

By addressing these questions, this project will not only uncover key performance drivers but also provide a decision-support tool through a visual dashboard that can guide competitive analysis and investment considerations in the space launch market.

Section 1

Methodology

Data Collection

- This project utilized the official SpaceX v4 REST API as the primary source of data.
- The API provides structured, real-time access to all historical SpaceX launches, rockets, payloads, cores, and launchpads.
- This ensured that the data was both accurate and up-to-date, eliminating the need for manual web scraping or third-party datasets.

Web Scraping

Process Overview

1. Import libraries and set display options for better visibility.
2. Use requests.get() to pull historical SpaceX launch data from the API.
3. Normalize the JSON to flatten nested structures into tabular format.
4. Filter data:
   - Include only rows with one core and one payload.
   - Restrict launches to those before November 13, 2020.
5. Extract nested details through separate functions:
   - getBoosterVersion(): Retrieves rocket name from rocket ID.
   - getLaunchSite(): Gets launch site name and coordinates.
   - getPayloadData(): Extracts payload mass and orbit.
   - getCoreData(): Gathers details about core block, reuse, landing outcome, etc.
6. Construct the final dataset using a dictionary and convert it into a pandas DataFrame.
7. Filter only Falcon 9 launches and reassign flight numbers sequentially.
8. Exported the collected data to dataset_part_1.csv for further use in data scraping.

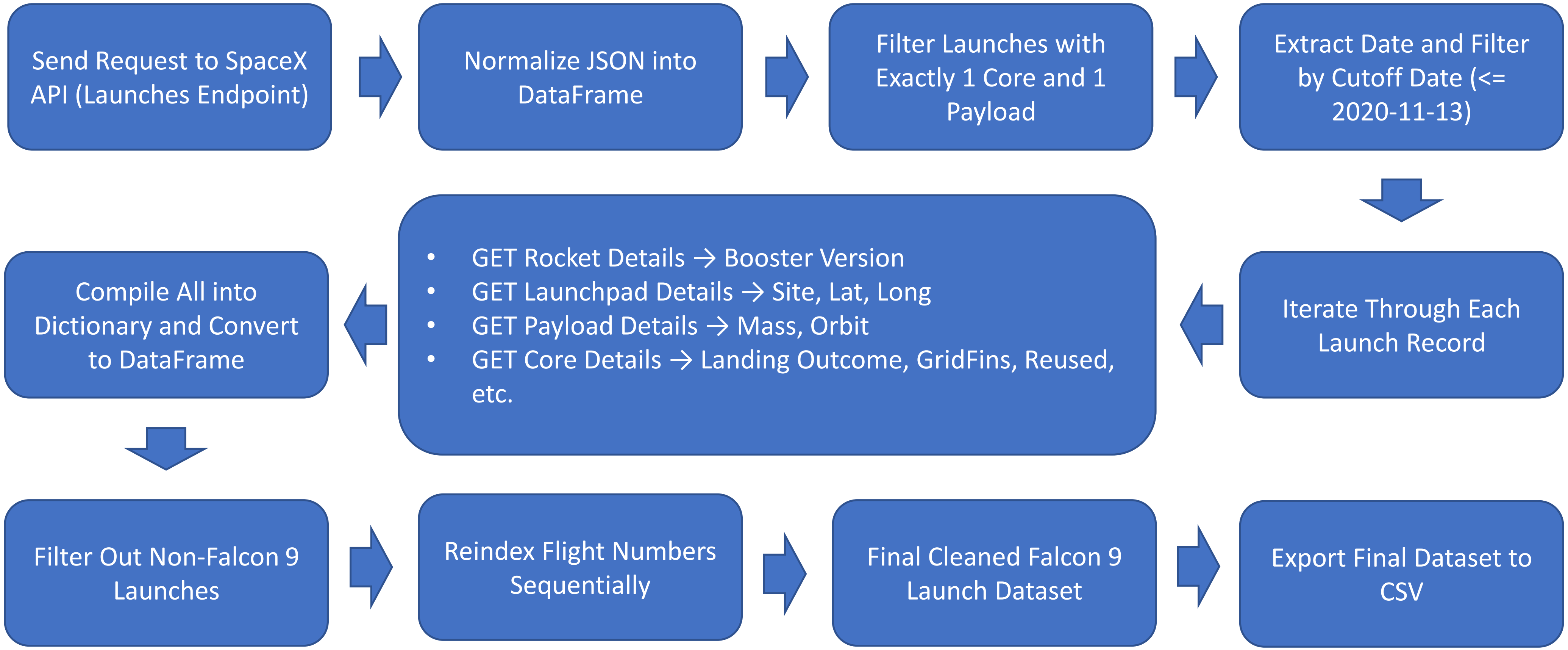
Web Scraping

Key Phrases:

- API Endpoint: Accessed data from https://api.spacexdata.com/v4/launches/past.
- HTTP Request: Retrieved JSON data using the requests library.
- JSON Normalization: Flattened nested JSON into a pandas DataFrame.
- Data Filtering: Selected relevant launches and ensured 1:1 mapping for cores and payloads.
- Iterative API Calls: Fetched additional details from nested endpoints (rockets, payloads, cores, launchpads).
- Data Aggregation: Compiled features like BoosterVersion, LaunchSite, Orbit, PayloadMass, and landing Outcome.
- Final Dataset: Cleaned and structured dataset for Falcon 9 launches up to November 13, 2020.

Web Scraping

Flowchart:



Web Scraping

GitHub URL of the completed Web Scraping notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/111_SpaceX_Data_Collection_API.ipynb

Data Wrangling

Process Overview

1. Imported the dataset from a CSV URL.
2. Checked missing values to ensure the dataset was complete and ready for analysis.
3. Identified data types for proper feature handling.
4. Explored key categorical variables to understand launch distributions and outcomes.
5. Manually defined “bad” landing outcomes based on initial analysis.
6. Engineered a binary target column Class, where 1 indicates success and 0 indicates failure.
7. Exported the wrangled data to dataset_part_2.csv for further use in modeling and analysis.

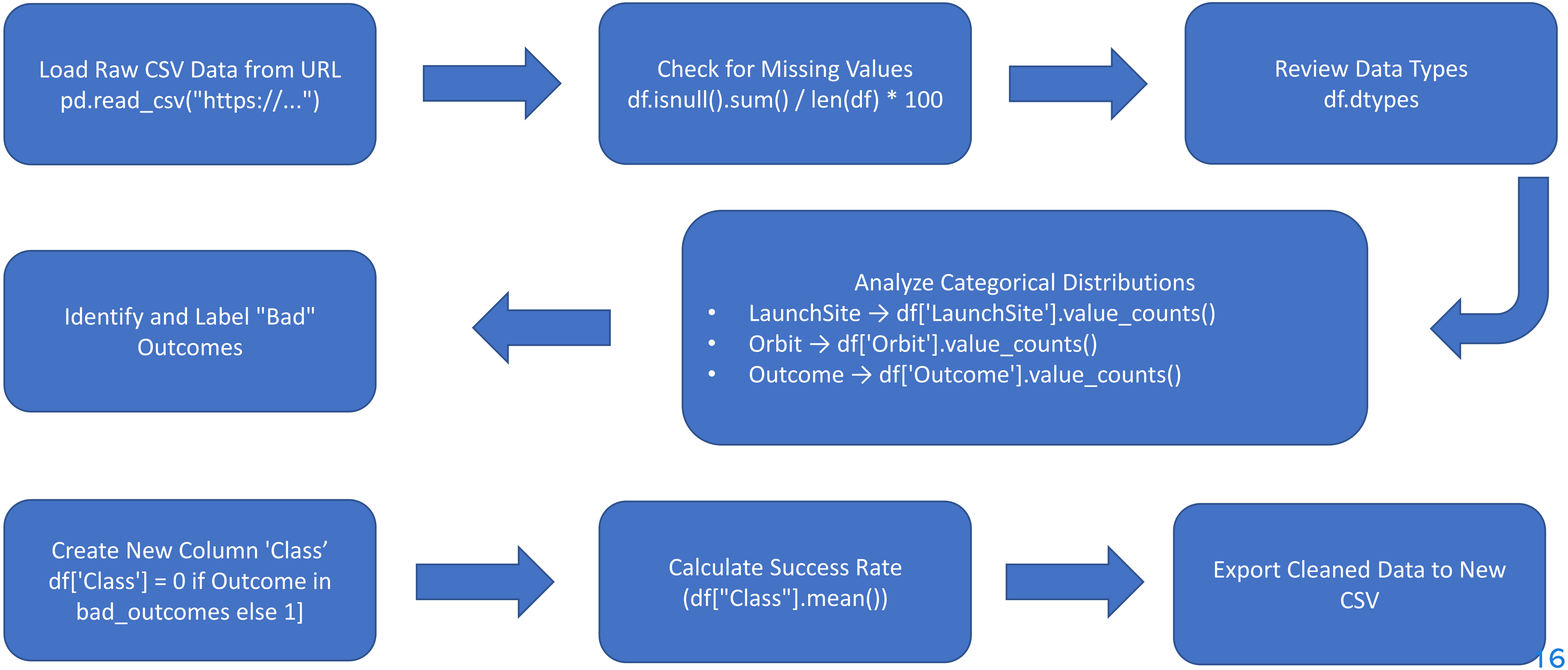
Data Wrangling

Key Phrases:

- Data Loading: Loaded raw data from a public IBM-hosted URL using pandas.read_csv().
- Missing Value Inspection: Assessed the percentage of missing data for each column to evaluate data quality.
- Data Type Identification: Checked data types of each column to understand the structure.
- Feature Analysis: Used value counts to understand the distribution of LaunchSite, Orbit, and Outcome.
- Target Label Engineering: Defined a new binary column Class to categorize launch outcomes as either successful (1) or failed (0).
- Manual Labeling: Labeled specific Outcome types as "bad" based on domain understanding and frequency.
- Data Export: Saved the wrangled and labeled dataset as a new CSV file for downstream tasks (e.g., modeling).

Data Wrangling

Flowchart:



Data Wrangling

GitHub URL of the completed Data Wrangling notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/112_SpaceX_Data_Wrangling.ipynb

EDA with SQL

Process Overview

1. Loaded data from a SpaceX mission dataset CSV using pandas.

2. Connected to SQLite using %sql magic in Jupyter Notebook.

3. Saved the dataset into a table SPACEXTBL, and filtered it to create a cleaner version SPACEXTABLE.

4. Performed SQL-based EDA:
    - Identified distinct launch sites.
    - Queried launch details from specific locations.
    - Aggregated payload data by Customer and Booster Version.
    - Investigated landing outcomes and their timing.
    - Tracked booster performance for large payloads.
    - Summarized mission success/failure trends.

5. All insights were obtained using concise, well-targeted SQL queries.

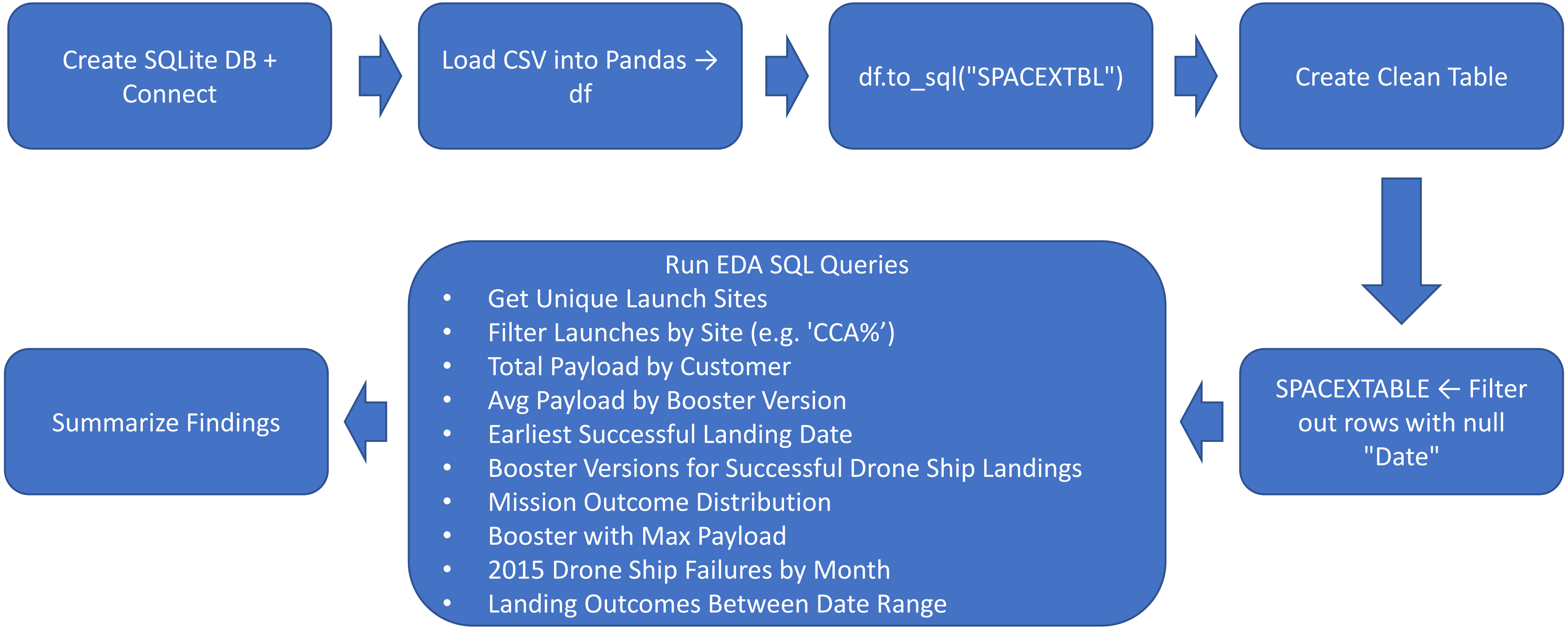
EDA with SQL

Key Phrases:

- Database Creation: Established a local SQLite database (my_data1.db) and loaded data from a SpaceX CSV file.
- Data Ingestion: Imported data using pandas.to sql() into a working table (SPACEXTBL).
- Data Cleaning: Created a clean table (SPACEXTABLE) by filtering out rows with null dates.
- Exploratory Queries:
  - Extracted unique launch sites
  - Filtered launches by site name patterns
  - Aggregated payload mass by customer and booster version
  - Identified launch dates and booster performance
  - Investigated success/failure outcomes by date, customer, and year

EDA with SQL

Flowchart:



EDA with SQL

GitHub URL of the completed EDA with SQL notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/113_EDA_SQL.ipynb

EDA with Data Visualization

Process Overview

1. Visualized Payload Trends
Used sns.catplot and scatter plots to explore how payload mass varies by flight number and how success (Class) differs across flight attempts.

2. Launch Site Performance
Created scatter plots of flight numbers vs. launch sites, revealing success patterns per location.

3. Orbit Analysis
Aggregated launch success rate per orbit and presented findings with a bar chart. This highlights which orbit types are more likely to result in successful missions.

4. Temporal Trends
Extracted the launch year from the date field, then visualized how the success rate improved over time using a line plot.

5. Feature Engineering
Prepared the dataset for modeling by applying one-hot encoding to categorical variables and exporting the feature matrix.

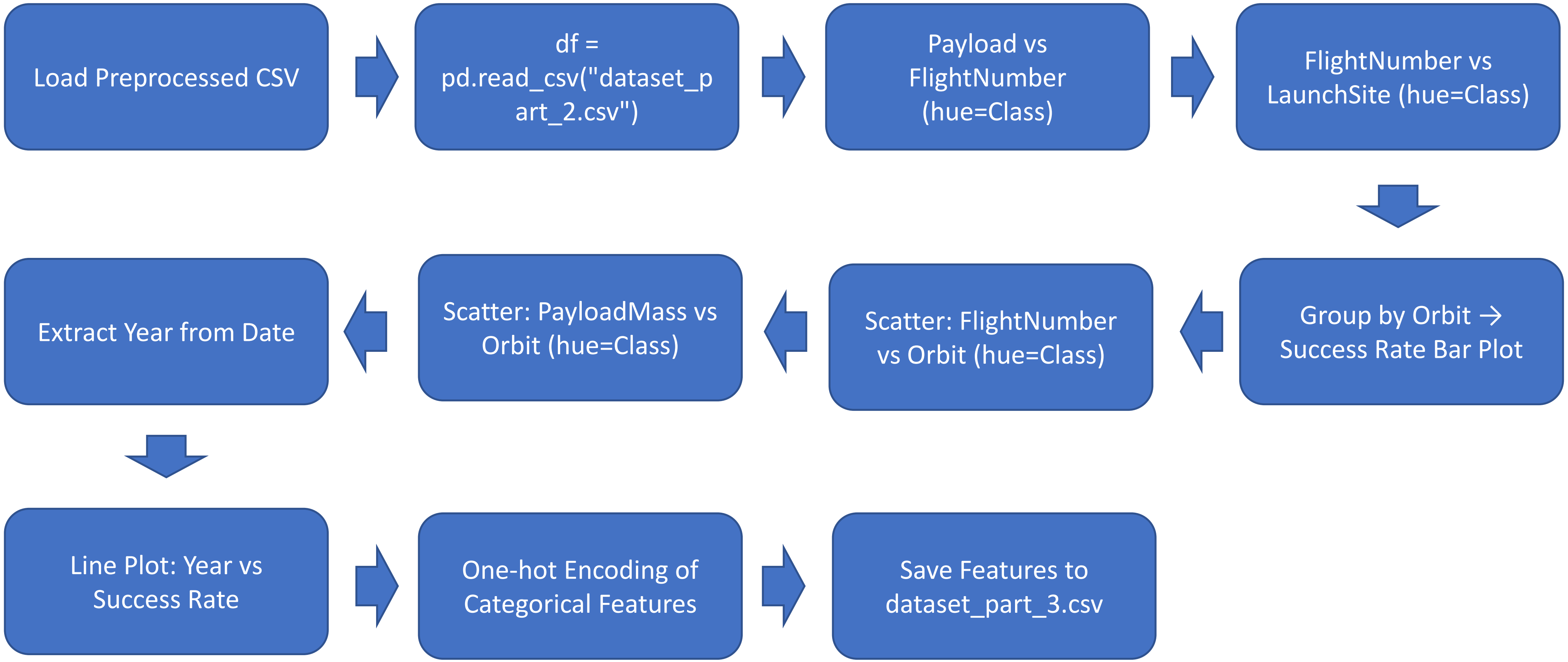
EDA with Data Visualization

Key Phrases:

- Dataset Overview: Loaded a preprocessed dataset (dataset_part_2.csv) containing information on Falcon 9 launches.
- Categorical Visual Analysis:
  - Visualized payload mass vs. flight number, colored by launch success (Class).
  - Analyzed launch site effectiveness using scatter plots.
- Orbit-Specific Analysis:
  - Compared success rate across different orbit types using bar plots.
  - Explored the relationship between flight number/orbit and launch outcomes.
- Payload-Based Analysis:
  - Investigated payload mass vs. orbit type, revealing patterns in success rates.
- Temporal Trends:
  - Extracted year from Date and plotted year-wise success rate trend using line plots.
- Feature Preparation:
  - Applied one-hot encoding to categorical variables (Orbit, LaunchSite, LandingPad, Serial) for future machine learning tasks.
  - Saved the final feature matrix as dataset_part_3.csv.

EDA with Data Visualization

Flowchart:



EDA with Data Visualization

GitHub URL of the completed EDA with Data Visualization notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/114_EDA_with_Visualization.ipynb

Interactive Map with Folium

Process Overview

1. Map Initialization
Created an interactive folium.Map centered around NASA Johnson Space Center with a zoom level appropriate for analysis.

2. Launch Site Marking
Extracted unique launch sites from the dataset, added circular overlays and labels for visual identification.

3. Launch Outcome Mapping
Color-coded markers using success/failure classification:

- Successes: green markers.
- Failures: red markers.
  These were grouped using MarkerCluster for cleaner visualization.

4. Mouse Interaction
Integrated MousePosition plugin to dynamically show lat-long coordinates on hover—useful for site exploration and further manual analysis.

5. Distance Calculations
Applied the Haversine formula to compute real-world distances from a launch site to:

- Coastline (~0.78 km)
- Railway (~1.0 km)
- Highway (~0.55 km)
- Nearest City (~54.5 km)
  Annotated distances using DivIcon labels and PolyLine visual paths.

6. Insights Gained

- Launch sites are strategically placed near coasts (for safety and trajectory).
- Proximity to infrastructure (rail, road) is essential for logistics.
- Launches are relatively remote from populated cities to reduce risk.

Interactive Map with Folium

Key Phrases:

- Objective: Analyze the geographic distribution and environmental proximity of SpaceX launch sites using folium, an interactive mapping library.
- Dataset: Used spacex_launch_geo.csv containing geolocation data (latitude, longitude, launch outcome).
- Key Visuals:
  - Plotted NASA Johnson Space Center with a labeled marker and circular highlight.
  - Mapped All Launch Sites with markers and circles indicating their locations.
  - Colored Launch Outcome Markers:
    - Green for successful launches.
    - Red for failed launches.
  - Clustered Markers to avoid clutter.
- Proximity Analysis:
  - Measured distances from a selected launch site to:
    - Coastline
    - Railway
    - Highway
    - Nearest City
  - Used great-circle distance formula (Haversine) to compute distances.
  - Drew lines and midpoint labels with distance annotations on the map.

Interactive Map with Folium

Flowchart:

1. Load Launch Site Data from CSV
2. Create Folium Map Centered on NASA JSC
3. Add Launch Site Markers & Circles
4. Color Markers Based on Launch Success/Failure
5. Add Marker Clustering
6. Enable Mouse Position Plugin for Coordinates
7. Calculate Distances to Environmental Landmarks
8. - Coastline
   - Railway
   - Highway
   - City
9. Draw Lines Between Points + Add Distance Labels
10. Render Interactive Map

Interactive Map with Folium

GitHub URL of the completed Interactive Map with Folium notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/115_Launch_Sites_Locations_Analysis_with_Folium.ipynb

Dashboard with Plotly Dash

Process Overview

1. This dashboard allows users to interactively analyze success patterns of SpaceX missions. Two key variables influence the outcomes:
    - The launch site, which may differ in equipment or geography.
    - The payload mass, which affects the rocket's performance.

2. By filtering payload ranges and choosing different launch sites, stakeholders can:
    - Investigate site-specific performance.
    - Compare booster categories.
    - Explore if heavier/lighter payloads influence launch success.

3. Visual Output Summary
    - Pie Chart
        - Shows success count breakdown per site or per outcome.
        - Dynamically changes based on:
            - Dropdown (site).
            - Slider (payload).

- Scatter Chart (Right Panel)
    - Displays each launch as a data point with:
        - X-axis: Payload mass.
        - Y-axis: Launch class (success = 1, failure = 0).
        - Color: Booster version category.

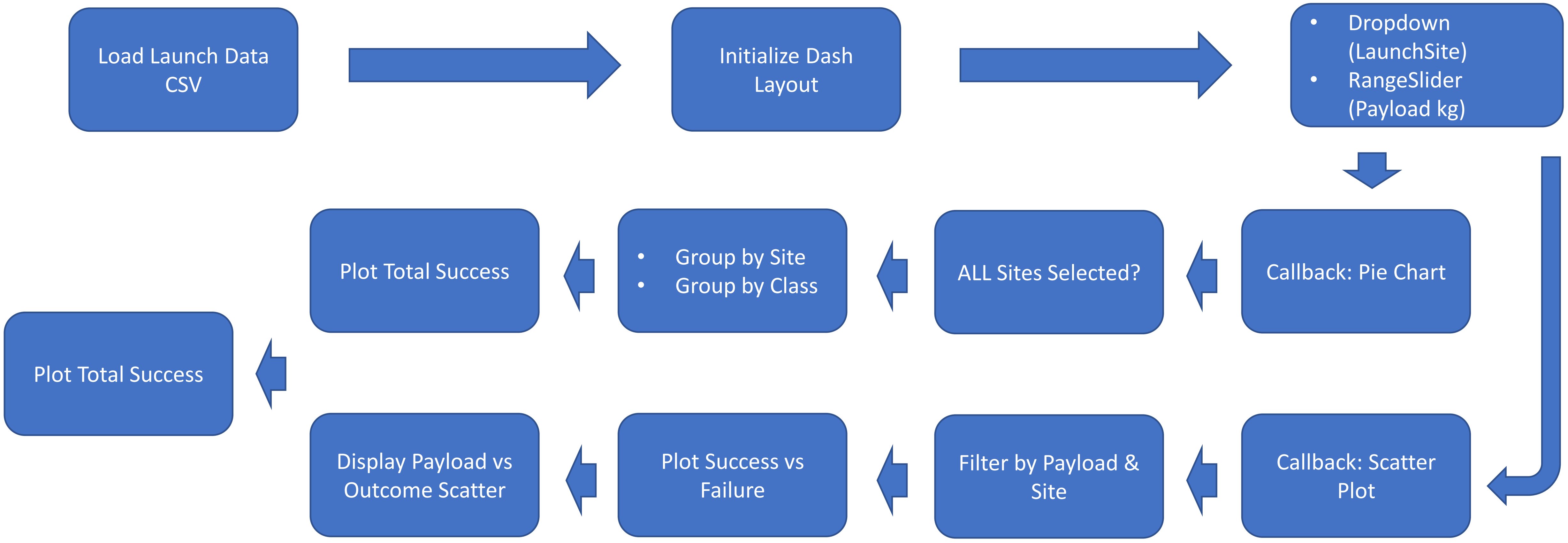
Dashboard with Plotly Dash

Key Phrases:

- Objective: Create an interactive dashboard to explore SpaceX launch performance based on launch sites and payload range.
- Technology Stack:
  - Plotly for data visualization
  - Dash for interactive components
  - Pandas for data wrangling
  - Python & HTML for layout design
- Key Components:
  - Dropdown Menu – Select individual launch sites or view all.
  - Payload Mass Slider – Filter data within a customizable payload range.
  - Pie Chart – Visualize launch success distribution:
    - If "All Sites" selected → Shows total successful launches by site.
    - If a specific site selected → Compares success vs failure.
  - Scatter Plot – Shows Payload vs Outcome, color-coded by Booster Version.
    - Allows site-specific or overall trends to be examined.
- Live Interactivity: Dynamic updates of both plots in real-time based on user input.

Dashboard with Plotly Dash

Flowchart:



Dashboard with Plotly Dash

GitHub URL of the completed Dashboard with Plotly Dash notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/116_Interactive_Dashboard_with_Ploty_Dash.ipynb

Predictive Analysis (Classification)

Process Overview

The mission was to build a predictive system for SpaceX's Falcon 9 boosters using historical launch data. The system classifies whether a booster successfully lands or not using supervised learning models. A set of features including payload mass, booster version, orbit, and more were used. Data was cleaned and standardized before being passed into several machine learning models with hyperparameter tuning via grid search.

After training and evaluation:

- The best-performing model during training (based on cross-validation) was Decision Tree with 90.17% accuracy.
- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.
- On your test set (20% of the data not seen during training), all models perform the same — 83.33% accuracy.
  - The dataset is small or not complex enough to distinguish model performance clearly.
  - The test set does not contain sufficient variability.
  - All models are performing similarly well, but no model has a clear edge.

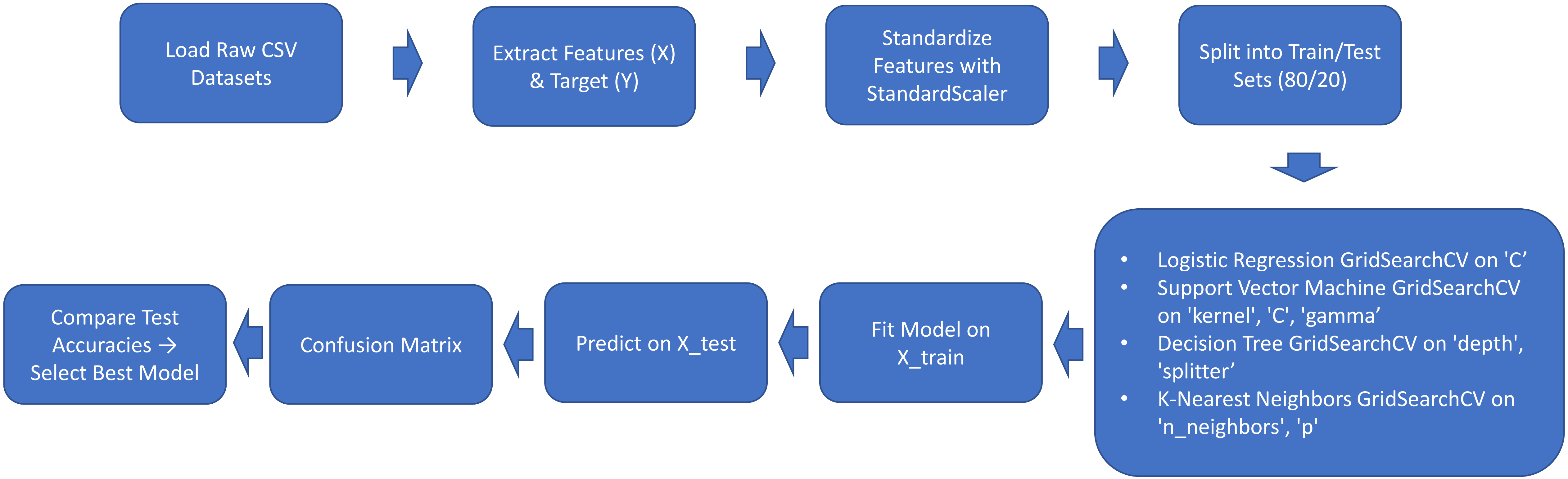
Predictive Analysis (Classification)

Key Phrases:

- Goal: Build a predictive model to determine if the Falcon 9 first stage lands successfully.
- Dataset: Sourced from IBM SpaceX dataset.
- Target Variable:
    - Class = 1: Rocket landed successfully
    - Class = 0: Rocket did not land
- Features: Engine type, booster version, payload mass, orbit, launch site, etc.
- Preprocessing:
    - Standardized features with StandardScaler.
    - Split into training (80%) and testing (20%) datasets.
- Algorithms Applied:
    - Logistic Regression
    - Support Vector Machine
    - Decision Tree
    - K-Nearest Neighbors
- Evaluation:
    - Used GridSearchCV for hyperparameter tuning.
    - Evaluated with accuracy scores and confusion matrices.
- Best Performer: Identified model with the highest test accuracy.

Predictive Analysis (Classification)

Flowchart:



Predictive Analysis (Classification)

GitHub URL of the completed Predictive Analysis (Classification) notebook:

https://github.com/jaynadzlibardo/applied_data_science_capstone/blob/main/117_Space_X_Falcon_9_First_Stage_Landing_Prediction.ipynb

Results

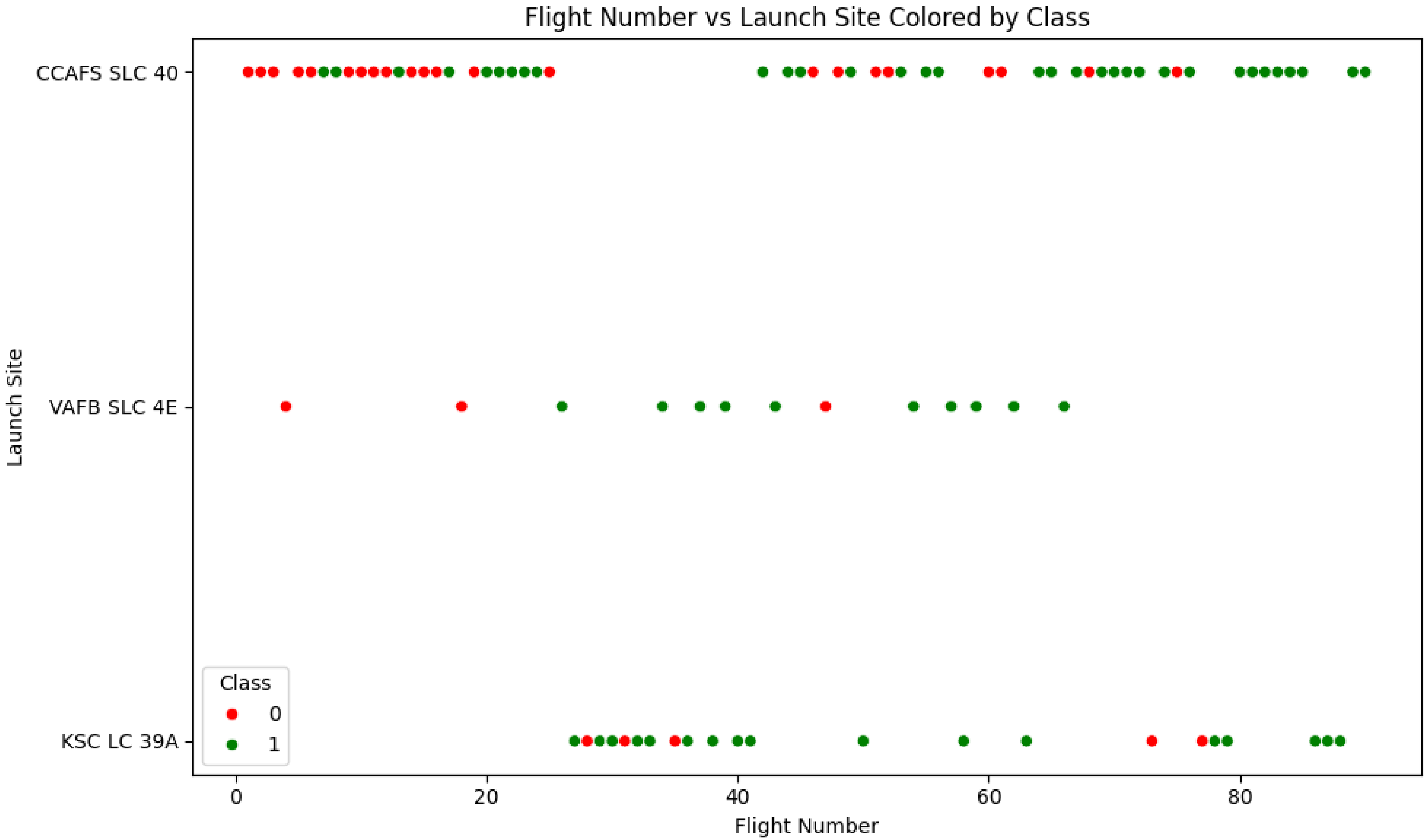
<table>
<tr><td>Exploratory Data Analysis Results</td><td colspan="4">1. Launch Sites - KSC LC-39A not only conducted the most successful missions (41.7%) but also showed the highest success rate (76.9%), making it the most reliable SpaceX launch pad.
2. Payload Mass - A U-shaped trend was observed:
• Low payloads (0–2000kg): more failures.
• Medium payloads (2000–4000kg): highest success rate.
• Very high payloads: fewer samples but decent performance.
3. Orbit ISS has the highest success rate at 61.90%, followed by Orbit GTO with 51.85%.
4. Booster Version Performance
• The FT version (Full Thrust) consistently showed high reliability.
• Earlier versions like v1.0 and v1.1 had more variability in outcomes.
5. Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.</td></tr>
<tr><td rowspan="5">Predictive Analysis Results</td><td>Model</td><td>Best Parameters Found</td><td>Cross-Validation Accuracy (Train Set)</td><td>Test Accuracy (Test Set)</td></tr>
<tr><td>Logistic Regression</td><td>C=0.01, penalty='l2’, solver='lbfgs'</td><td>84.64%</td><td>83.33%</td></tr>
<tr><td>Support Vector Machine</td><td>kernel=’sigmoid', C=1, gamma=0.03162277660168379</td><td>84.82%</td><td>83.33%</td></tr>
<tr><td>Decision Tree</td><td>criterion='entropy', max_depth=4, max_features='sqrt’, min_samples_leaf=2, min_samples_split=10, splitter='random'</td><td>90.17%</td><td>83.33%</td></tr>
<tr><td>K-Nearest Neighbors</td><td>n_neighbors=10, algorithm='auto’, p=1 (Manhattan)</td><td>84.82%</td><td>83.33%</td></tr>
</table>

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- CCAFS SLC-40 has the highest number of flights compared to VAFB SLC-4E and KSC LC-39A.
- Most of the successful flights occurred between Flight Numbers 30 and 90.
- For the three launch sites, success appears to be related to the number of flights.

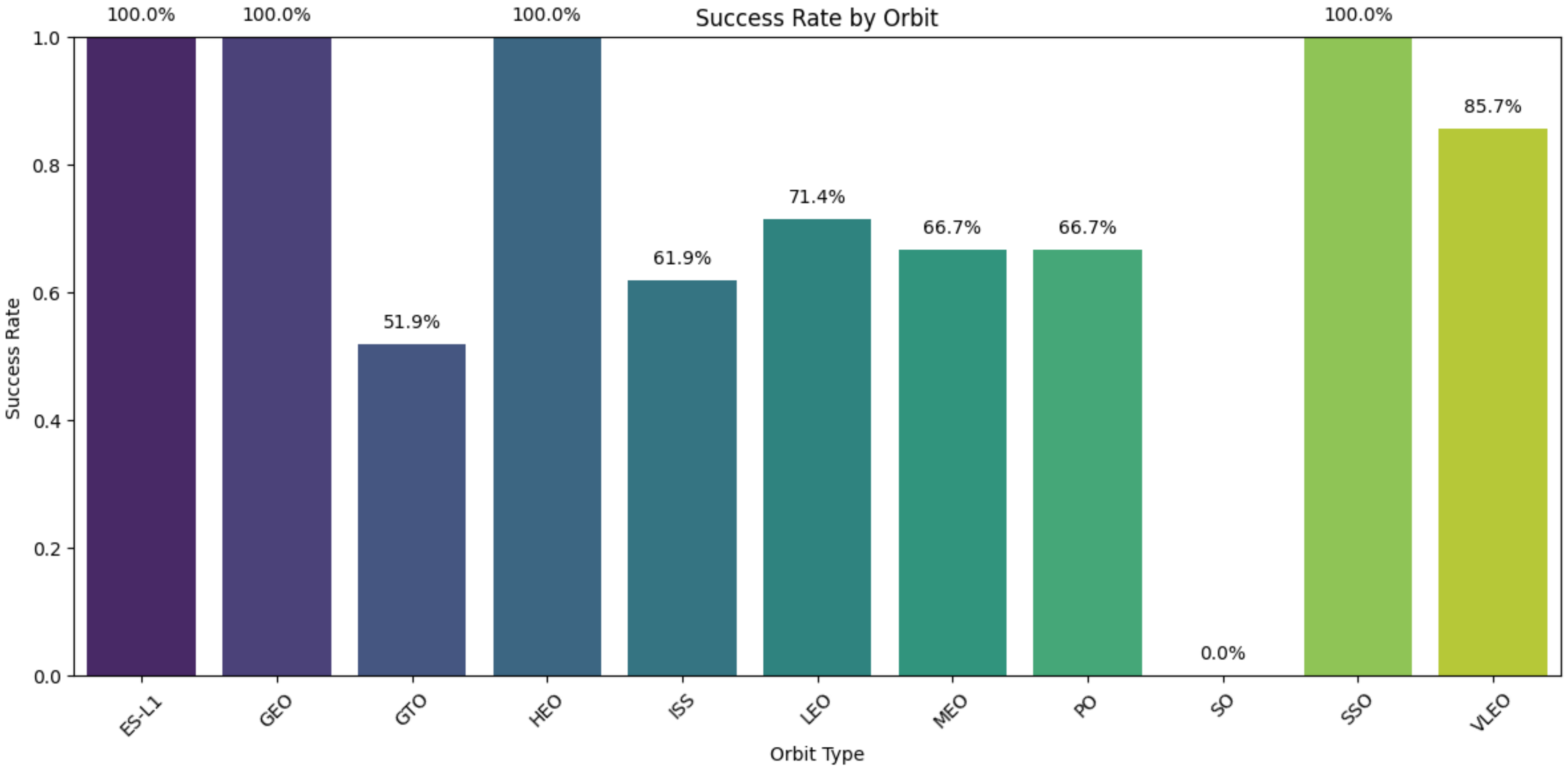


Payload vs. Launch Site

- For the VAFB SLC-4E launch site, there are no rockets launched with a heavy payload mass (greater than 10,000 kg).
- CCAFS SLC-40 flights have operated between 1000 kg and 7000 kg.
- Most of the flights occurred with a payload mass between 1000 kg and 7000 kg.
- Most successful flights occurred with a payload mass between 2000 kg and 4000 kg.

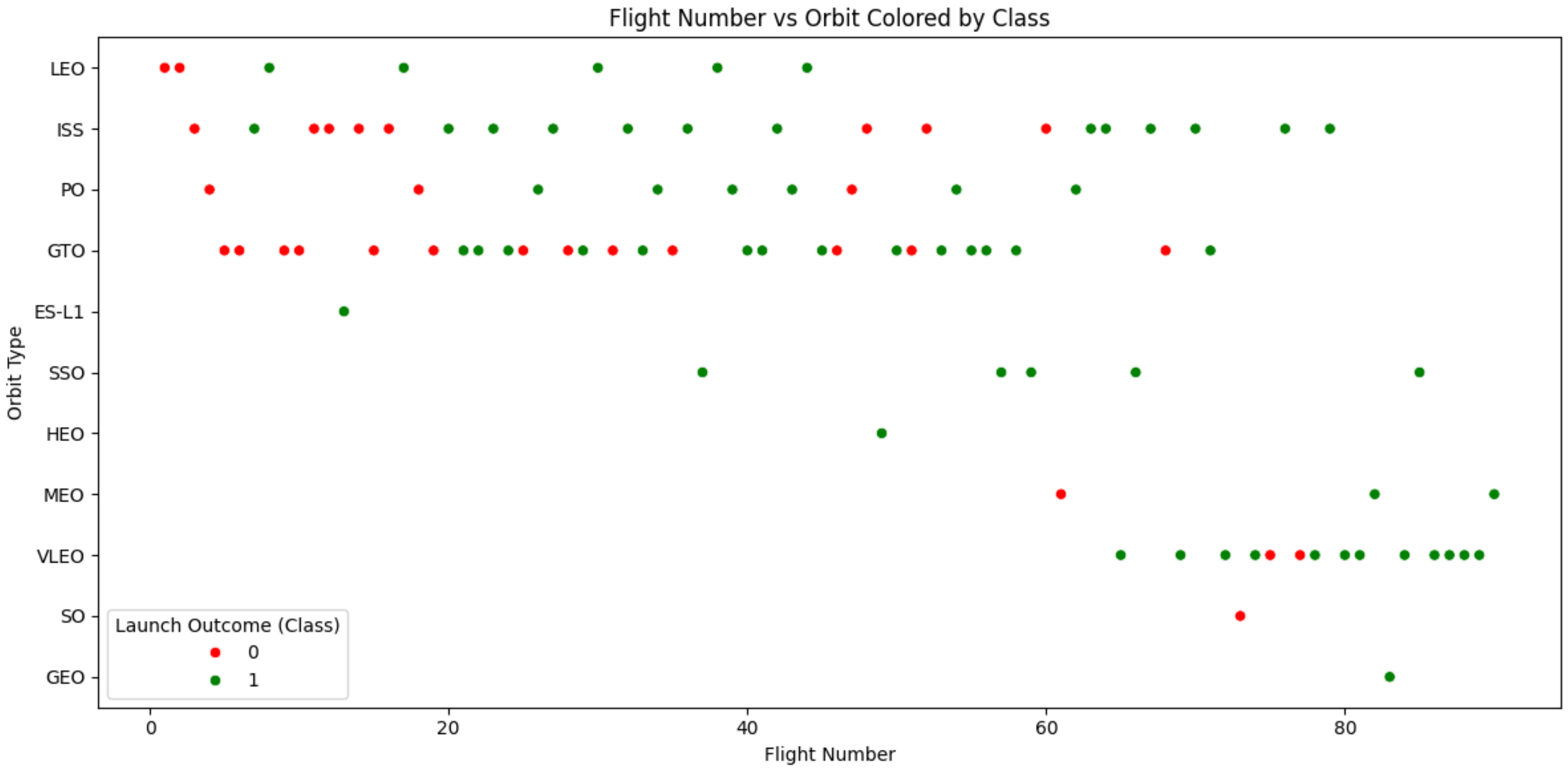
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.
- Other orbits show a success rate between 50% and 85%.
- The SO orbit has a 0% success rate, being the only orbit with no successful launches.



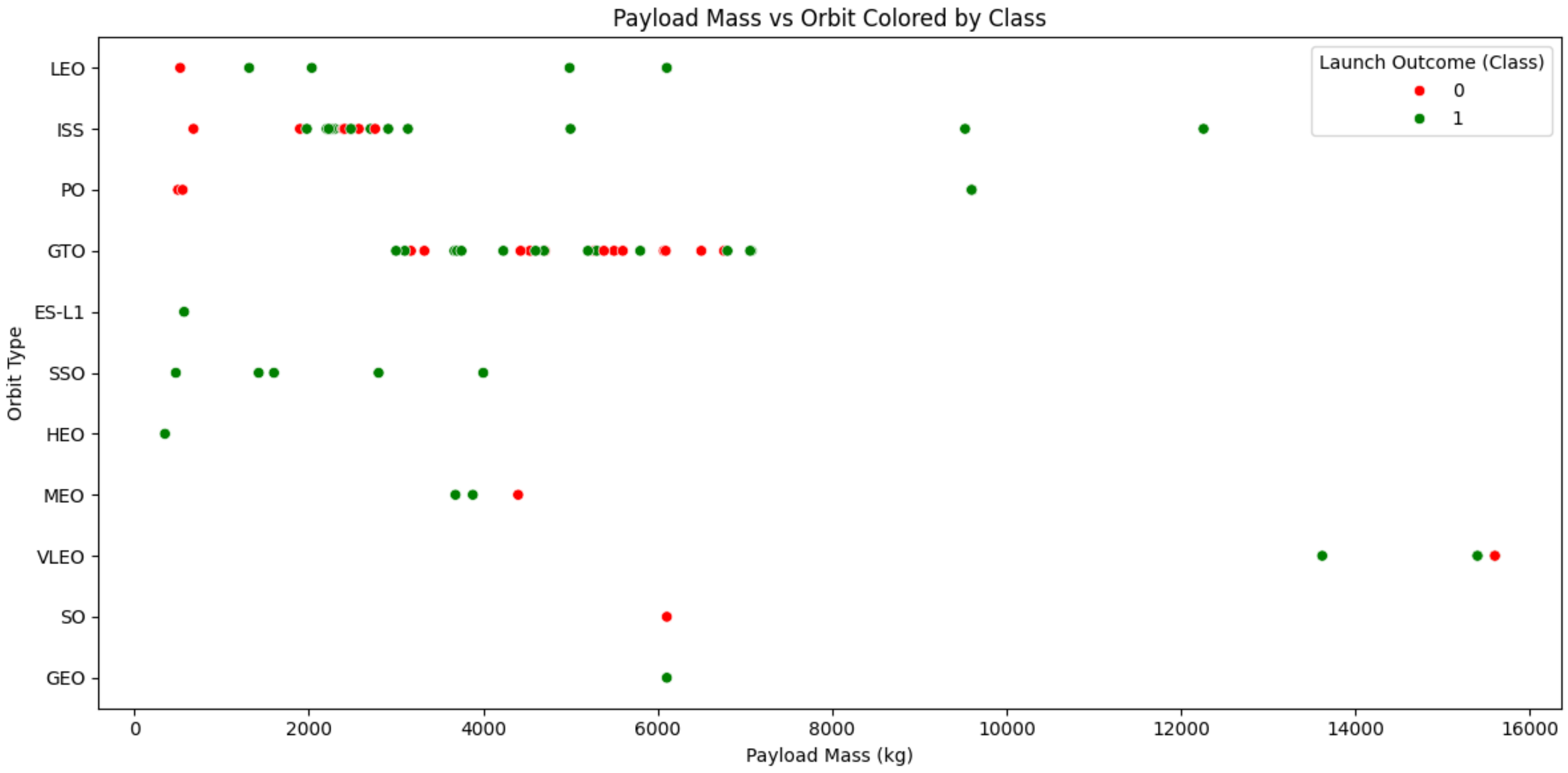
Flight Number vs. Orbit Type

- In the LEO, SSO, VLEO and PO orbit the success appears related to the number of flights.
- Orbit ISS and GTO have the highest number of flights.
- Orbits LEO, PO, and VLEO have a moderate number of flights.
- Orbit ISS has the highest success rate at 61.90%, followed by Orbit GTO with 51.85%.



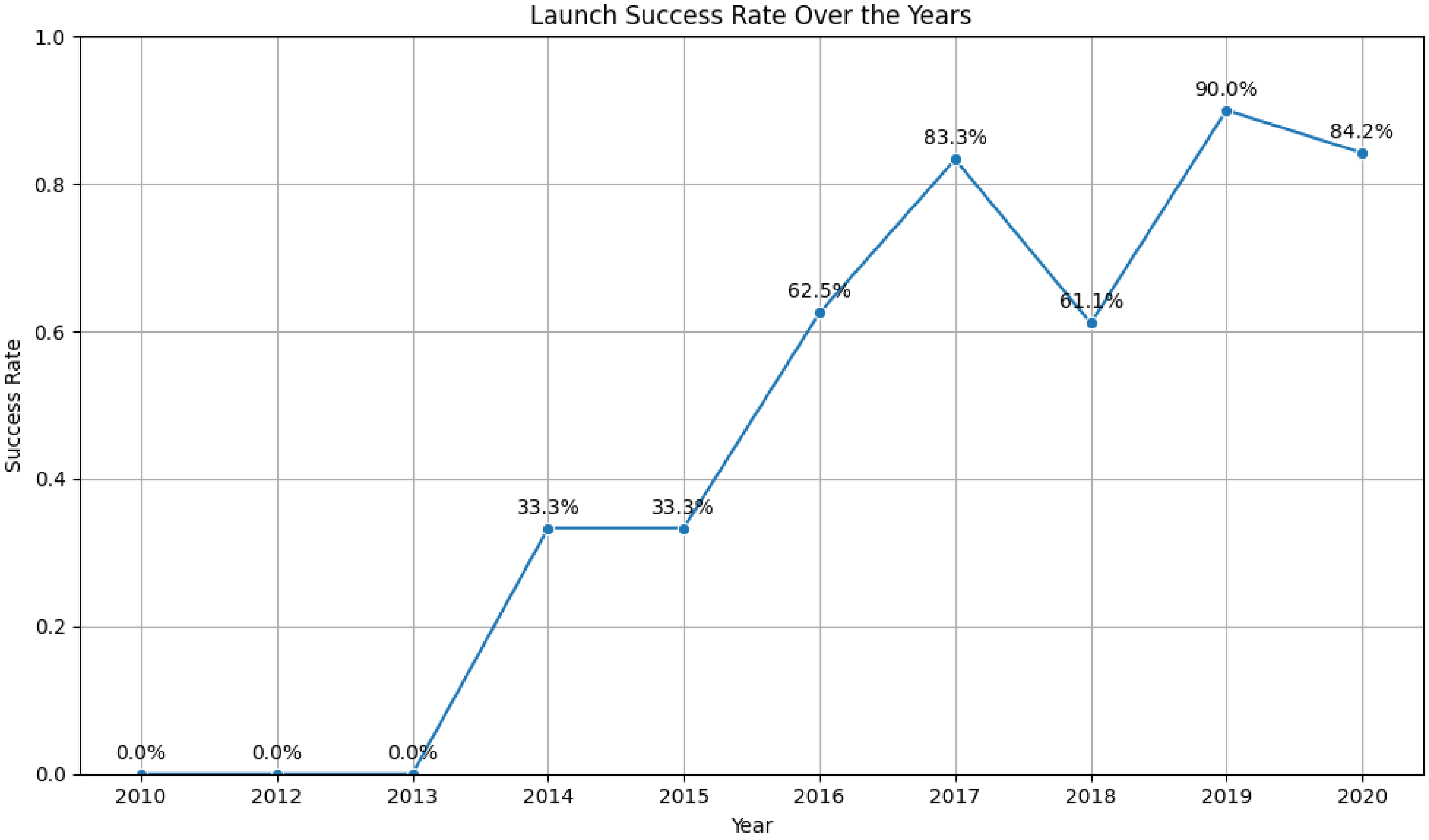
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.
- Most of the successful flights operate with payloads between 2000 kg and 4000 kg for the LEO, ISS, GTO, SSO, and VLEO orbits.



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.
- The success rate steadily increased from 2013 to 2017 due to significant improvements in technology, particularly the introduction of Falcon 9 Full Thrust in 2015.
- SpaceX's ability to rapidly iterate, learn from failures, and enhance quality control also played a key role. After a stable period in 2014, the success rate began rising sharply post-2015, reflecting these operational advancements.



All Launch Site Names

```
[7] %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

- This query returns a list of unique launch site names from the SPACEXTABLE.
- If SPACEXTABLE has multiple entries from the same launch site, DISTINCT ensures each site only appears once in the result.

Launch Site Names Begin with '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
2010-06-04	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

- This query retrieves the first 5 rows from the SPACEXTABLE where the launch site name starts with 'CCA'.
- It returns all columns (*) for those matching records.

Total Payload Mass

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

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

total_payload_mass
45596

- This query calculates the total payload mass (in kilograms) from the SPACEXTABLE for missions where the customer is 'NASA (CRS)’.
- The result is shown under the alias total_payload_mass.

Average Payload Mass by F9 v1.1

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

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

average_payload_mass
2928.4

- This query calculates the average payload mass (in kilograms) from the SPACEXTABLE for missions that used the 'F9 v1.1' booster version.
- The result is labeled as average_payload_mass.

First Successful Ground Landing Date

```
%sql SELECT "Date" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)' ORDER BY "Date" LIMIT 1;
```

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

Date
2015-12-22

This query retrieves the earliest date when a landing with the outcome 'Success (ground pad)' occurred, by:

- Filtering rows where "Landing_Outcome" is 'Success (ground pad)’
- Ordering them by "Date" in ascending order
- Returning just the first (earliest) date

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = '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

- This query returns a list of unique booster versions that successfully landed on a drone ship, carrying payloads between 4000 and 6000 kg.
- It uses DISTINCT to ensure each booster version appears only once.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS total_count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

- This query returns a count of missions grouped by each unique mission outcome.
- It shows how many times each "Mission_Outcome" occurred, labeling the count as total_count.

Boosters Carried Maximum Payload

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

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

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006
F9 v1.0 B0007

This query returns the booster version(s) that carried the heaviest payload recorded in the SPACEXTABLE. It:

- Finds the maximum payload mass using a subquery.
- Retrieves the distinct booster version(s) that match that maximum payload mass.

2015 Launch Records

```
%sql SELECT  substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr("Date", 0, 5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)';

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

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

This query retrieves information about failed drone ship landings that occurred in the year 2015. Specifically, it returns:

- The month extracted from the "Date" column,
- The landing outcome (which will be 'Failure (drone ship)'),
- The booster version, and
- The launch site.

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

```
%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.
```

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

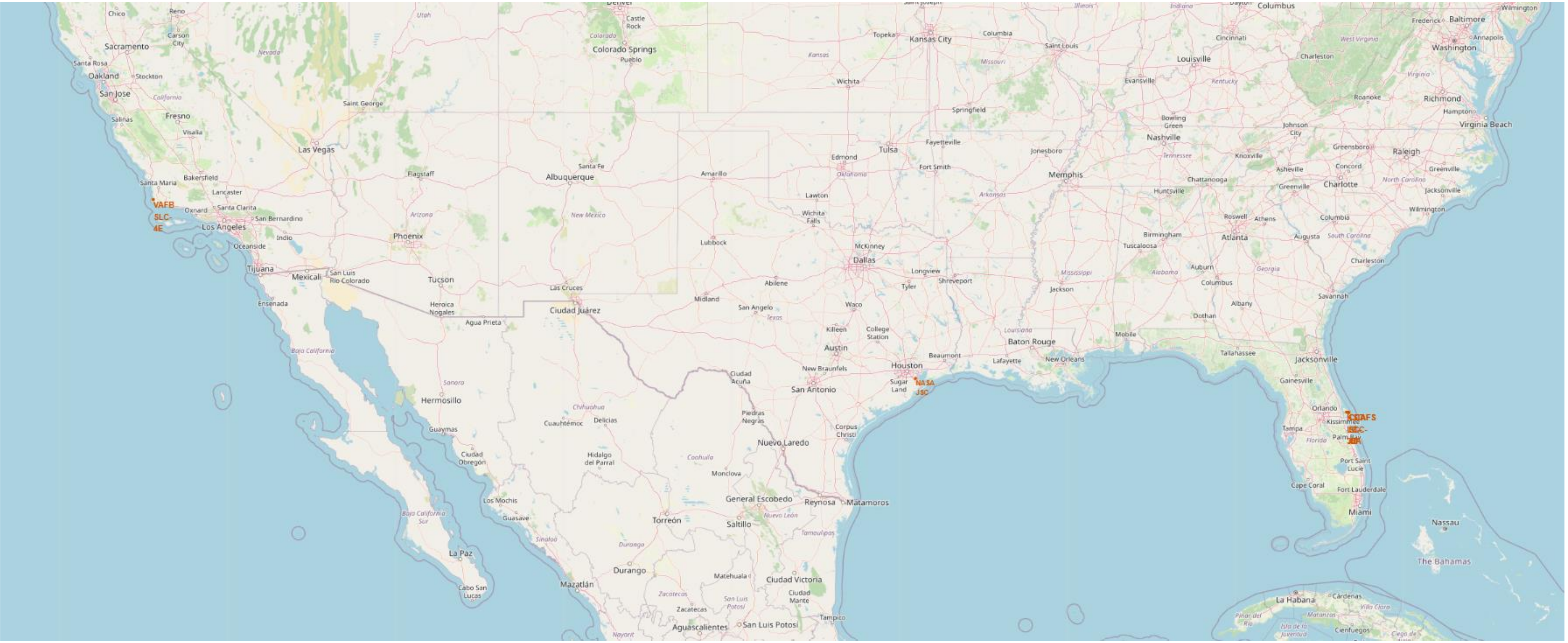
This query returns the number of missions per landing outcome that occurred between June 4, 2010, and March 20, 2017. It:

- Groups the results by "Landing_Outcome",
- Counts how many times each outcome occurred (COUNT(*)),
- Orders the results by the count in descending order (most frequent outcomes first).

Section 3

Launch Sites Proximities Analysis

All Launch Sites

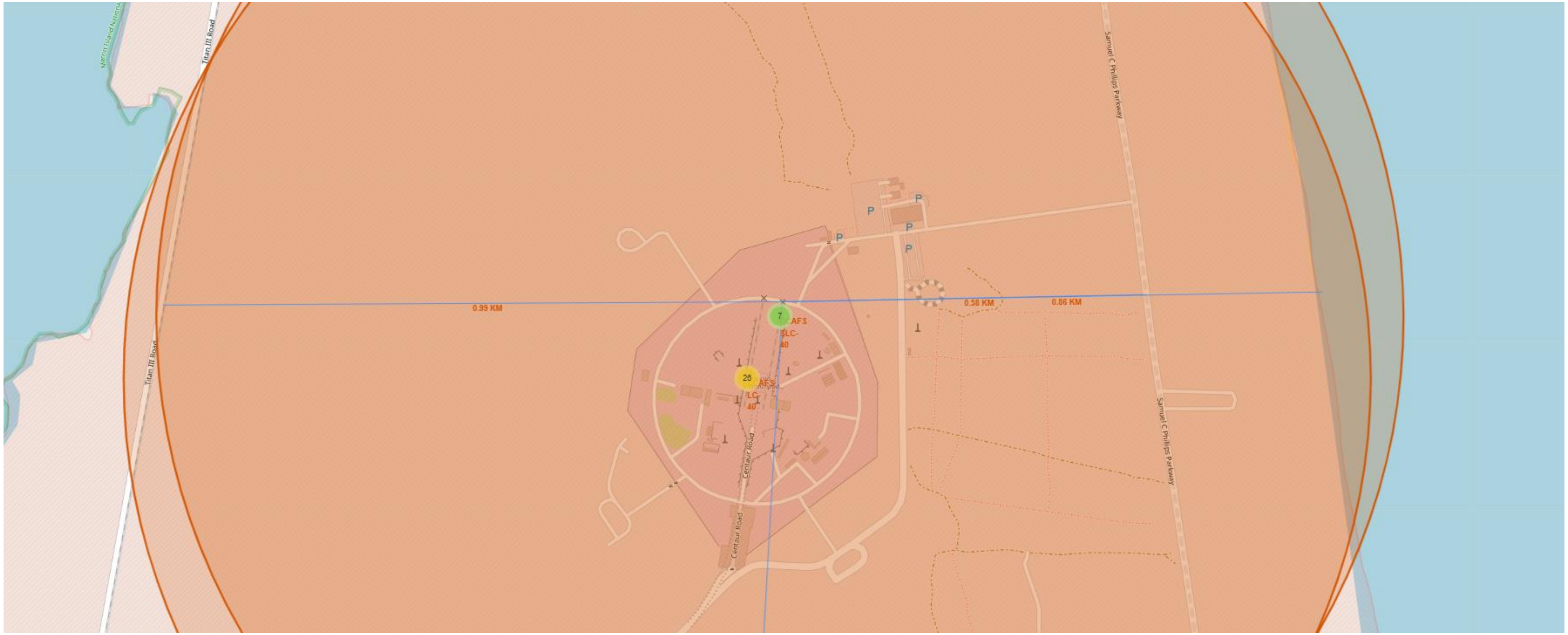


- All launch sites are near coastlines.
- We have launch sites on both the east and west coasts, respectively.

Launch Outcomes for Each Site

- From the color-labeled markers in marker clusters, we should be able to easily identify which launch sites have relatively high success rates.

Launch Site to its Proximities

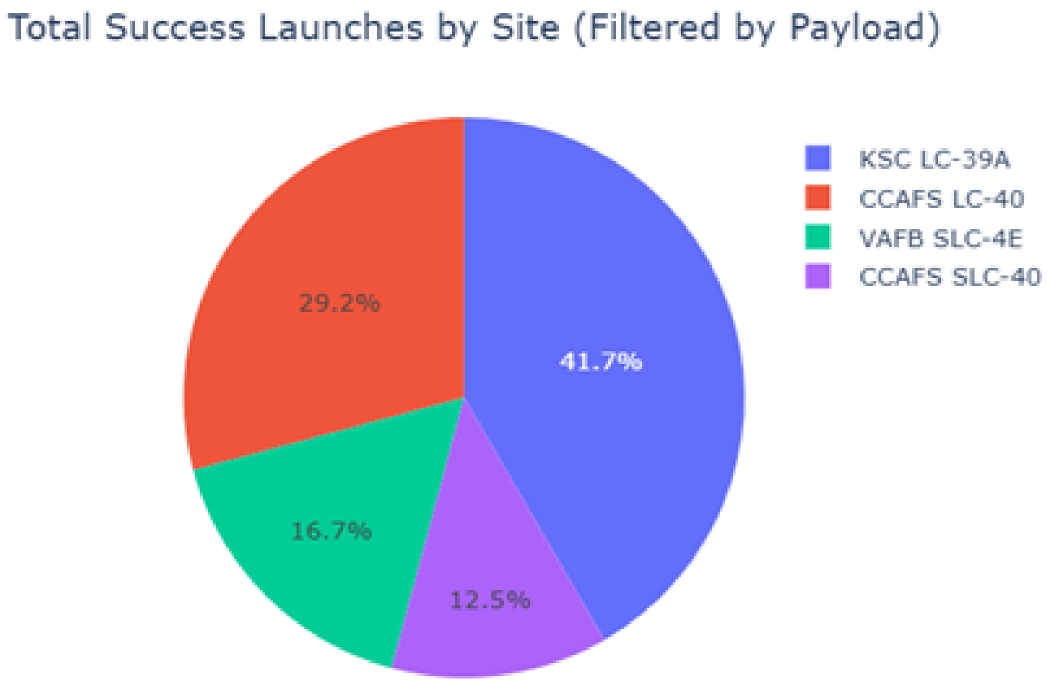


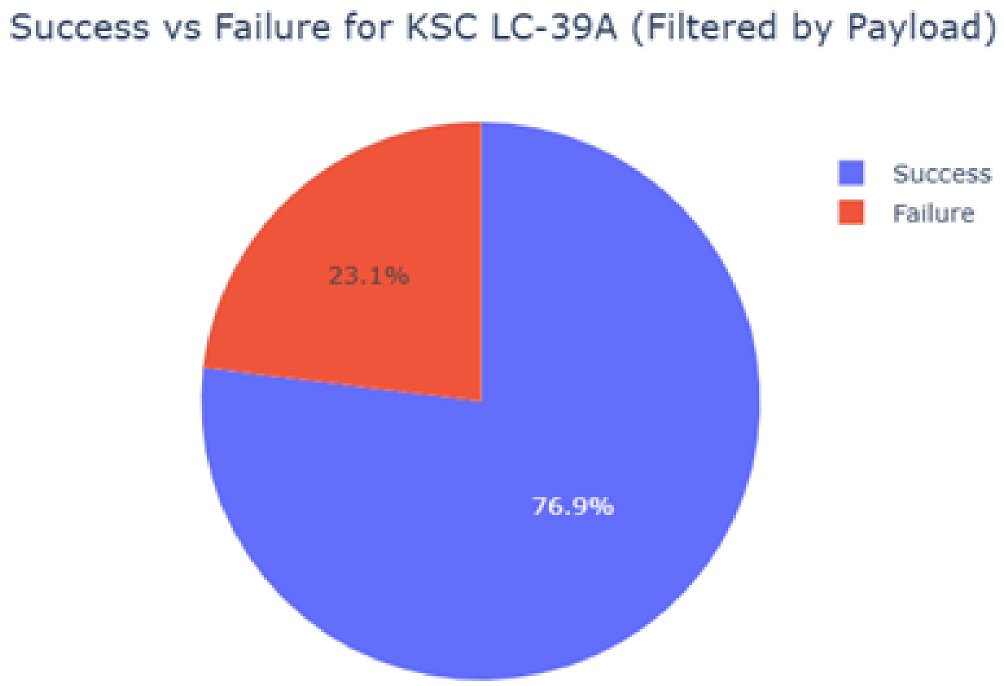
- Distance to Titan III (railway): 0.99 km, SC Phillips Parkway (highway): 0.58 km, Florida Coastline: 0.86 km and Melbourne City: 53.92 km

Section 4

Build a Dashboard with Plotly Dash

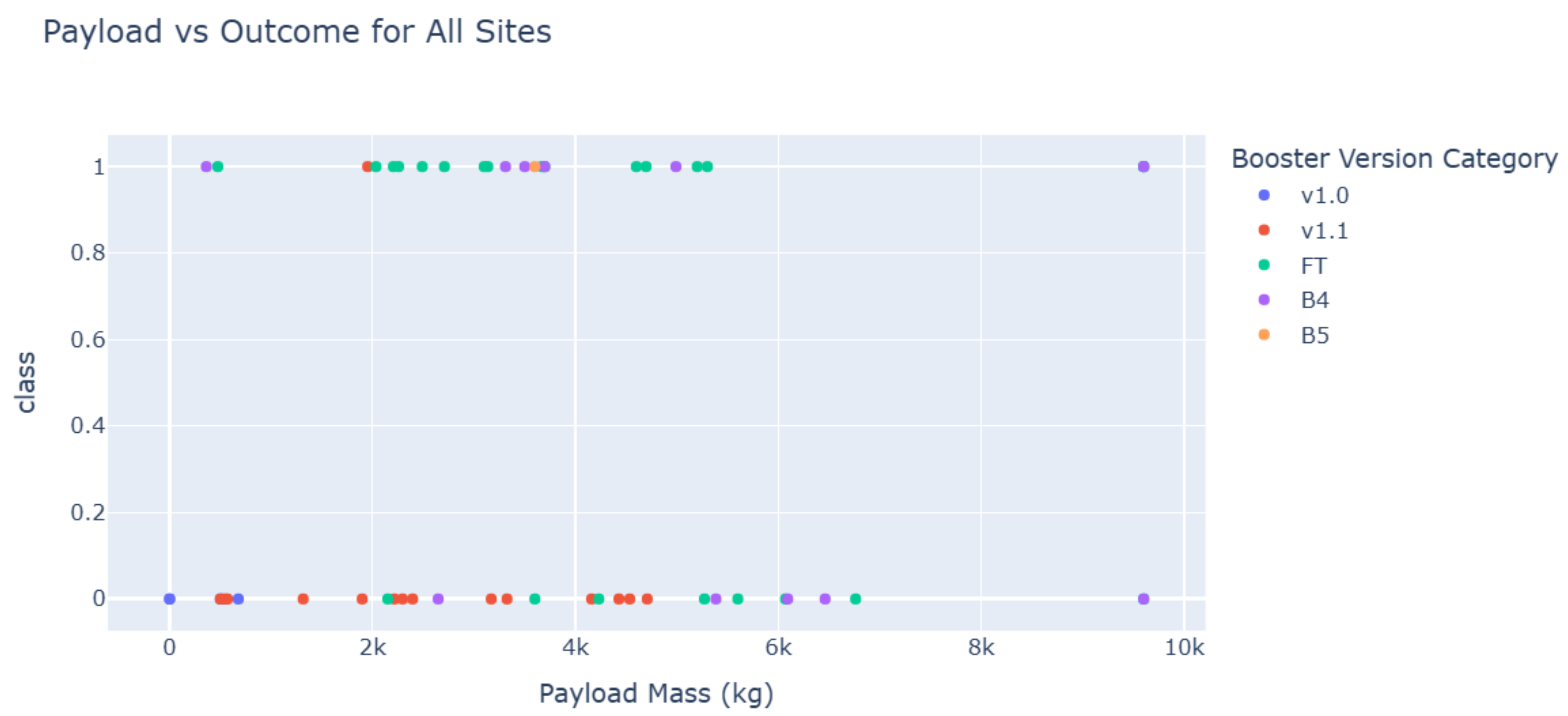
Total Success Launches by Site Pie-Chart





- KSC LC-39A has the highest successful launch rate among all sites, at 41.7%.
- It also has the highest success-to-failure ratio, at 76.9%.

Payload vs. Launch Outcome Scatter Plot for all Sites



- Most successful flights carry payloads weighing between 2,000 kg and 4,000 kg.
- The FT booster version has the highest number of successful flights.

Section 5

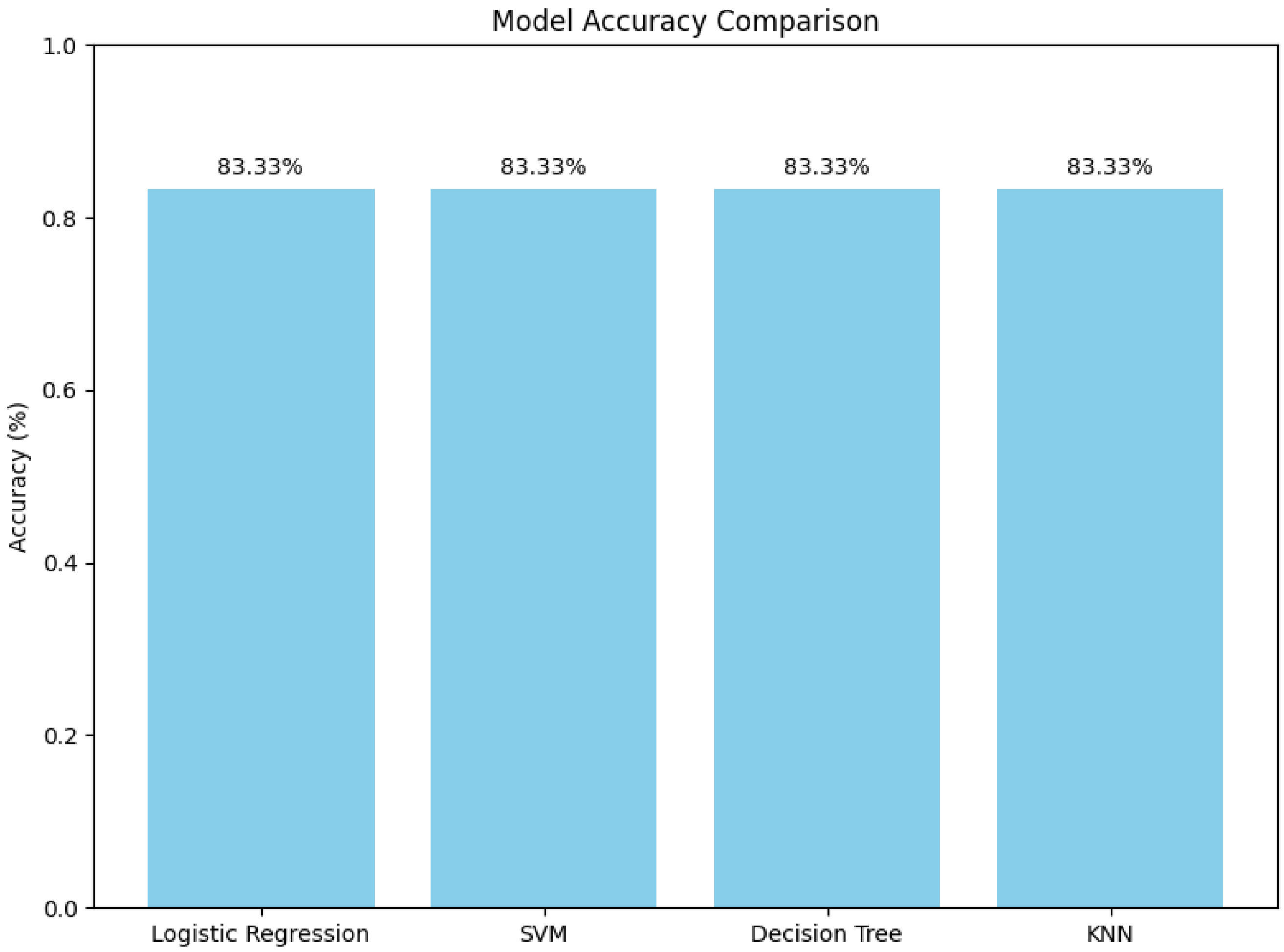
Predictive Analysis (Classification)

Classification Accuracy

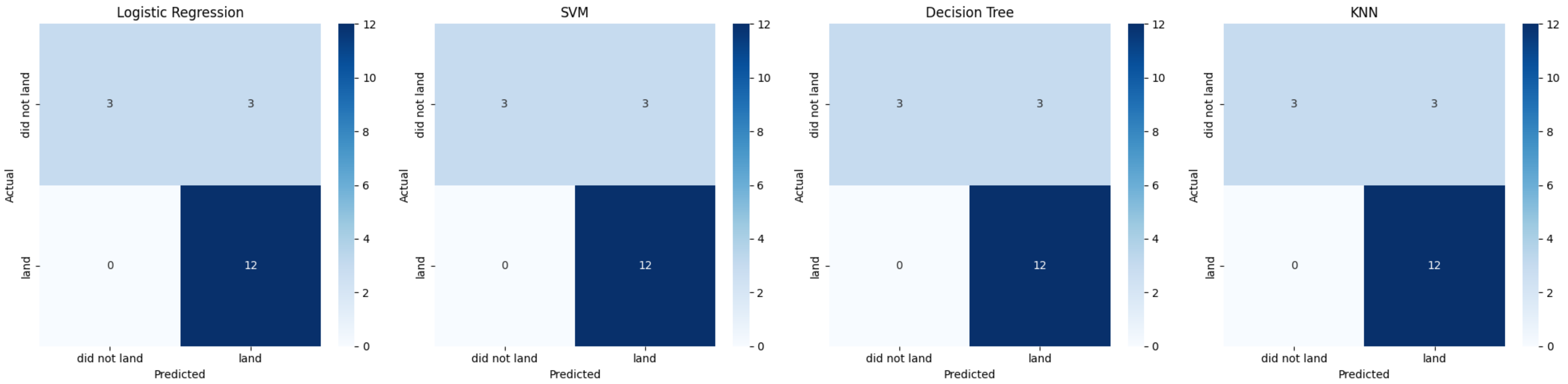
On test set (20% of the data not seen during training), all models perform the same 83.33% accuracy.

This could mean:

- The dataset is small or not complex enough to distinguish model performance clearly.
- The test set does not contain sufficient variability.
- All models are performing similarly well, but no model has a clear edge.



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



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

- True Positive - 12 (True label is landed, Predicted label is also landed)
- False Positive - 3 (True label is not landed, Predicted label is landed)

Conclusions

- This capstone project provided valuable insights into SpaceX's Falcon 9 launch performance through exploratory data analysis and predictive modeling.
- From the EDA, we discovered that KSC LC-39A had the highest number of successful launches and the highest success rate. Launches with payloads between 2000 kg and 4000 kg achieved the highest success rates, and the FT (Full Thrust) booster version proved to be the most reliable. Additionally, orbits such as GEO, ES-L1, HEO, and SSO achieved 100% success rates, highlighting strong mission planning and technological maturity.
- Our machine learning models, including Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors, all performed consistently well, each achieving around 83% accuracy on unseen test data. The Decision Tree classifier showed the highest cross-validated accuracy during training.
- A time series analysis revealed that the success rate improved significantly from 2015 onward, reflecting the impact of engineering refinements and operational experience. This trend underscores SpaceX's continuous innovation and learning from past failures.
- In summary, this project demonstrates how data-driven analysis can uncover operational strengths, highlight performance trends, and support predictive insights in the aerospace industry. These findings can help guide further improvements in launch operations and mission planning.

Appendix

A. Datasets Used

- SpaceX Launch Data
  URL: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv
  Includes details on launch sites, payload mass, booster versions, orbit types, and success classification (Class).
- Falcon 9 Features Dataset
  - dataset_part_2.csv: Contains additional features and target variable (Class).
  - dataset_part_3.csv: Preprocessed input features used for machine learning models.

B. Key Terminologies

- Flight Number - Sequential ID of a SpaceX launch. Higher numbers = later launches.
- Launch Site - SpaceX launch locations (e.g., CCAFS SLC-40, KSC LC-39A).
- Payload Mass - Weight of the payload (in kg) launched into orbit.
- Orbit - Target orbit for the mission (e.g., LEO, GTO, SSO).
- Booster Version - Rocket version used for the launch (e.g., v1.1, FT, B5).
- Class - Binary outcome (1 = Success, 0 = Failure) of the launch's first stage landing.

C. Exploratory Data Analysis (EDA) Summary

- Most launches: CCAFS SLC-40
- Highest success rate: KSC LC-39A (76.9%)
- Best payload range: 2000–4000 kg
- Orbit success: GEO, ES-L1, SSO, and HEO achieved 100% success rates
- Booster performance: FT (Full Thrust) version outperformed other variants

D. Machine Learning Models

- Logistic Regression with an accuracy of 83.33%.
- Support Vector Machine with an accuracy of 83.33%.
- Decision Tree with an accuracy of 83.33%.
- K-Nearest Neighbors with an accuracy of 83.33%.

All models performed equally well on test data, indicating balanced feature representation.

E. Visual Summary

- Pie & scatter plots: Used in Plotly Dash dashboard to visualize launch success by site and payload mass.
- Bar plots: Showed success rate by orbit type.
- Line plot: Revealed steady increase in success rate from 2013–2017.
- Confusion Matrices: Evaluated model predictions vs actual outcomes.

F. Tools and Libraries

- Python Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, plotly, dash
- Environment: Google Colab / Jupyter Notebook
- Visualization: Matplotlib, Seaborn, Plotly

Thank you!