IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gabriel Santiago Murillo Barragan
– BSc Biomedical Engineer (Universidad de los Andes)
_ Medical Student (Universidad Nacional de Colombia)
< 03/14/2025>

Outline

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

Executive Summary

- This report presents a thorough analysis of SpaceX Falcon 9 launch data with the objective of predicting whether the first stage will land successfully. By collecting and cleaning data from multiple sources (including the official SpaceX API), performing exploratory data analysis (EDA), and applying various predictive models, we identify the critical factors that affect landing success.
- • Predictive analysis was performed using the following machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN)
- The findings highlight:
  - The role of launch site characteristics in determining success rates.
  - The impact of payload mass on fuel consumption and structural stability.
  - The iterative improvements of the booster versions that correlate with higher landing success.
  - Logistic Regression, SVM, and KNN performed equally well on this dataset for predictive purposes. After fine – tuning and adjustment of hyperparameters the best model was Decision Tree.

Introduction

- Reusable rockets has marked a paradigm shift in the aerospace industry, significantly reducing launch costs and improving the sustainability of space exploration.
- The introduction of Space X's Falcon 9 revolutionized the industry demonstrating that a reusable first stage could land successfully and be relaunched multiple times
- First-stage landing success is influenced by multiple variables, including fuel efficiency, atmospheric conditions, thrust vector control, and landing pad precision.
- Advances in deep learning and reinforcement learning have improved trajectory optimization and landing stability, making automated rocket landings more reliable than ever before.

Objectives

- **General Objective:** To identify and characterize the key factors influencing the successful landing of the Falcon 9 first stage through data-driven analysis, enabling the development of predictive models that enhance landing reliability and contribute to cost-effective space operations.
- **Specific Objectives:**
  - To determine the influence of technical, environmental, and operational variables on the success rate of Falcon 9 first-stage landings.} By analyzing historical launch data, this objective aims to establish statistical relationships between independent factors (e.g., weather conditions, rocket thrust, payload weight) and landing outcomes.
  - To assess the predictive power of machine learning models in forecasting landing success.} This includes evaluating the accuracy, interpretability, and applicability of different algorithms in predicting first-stage recovery, contributing to improved decision-making in launch operations.

Section 1

Methodology

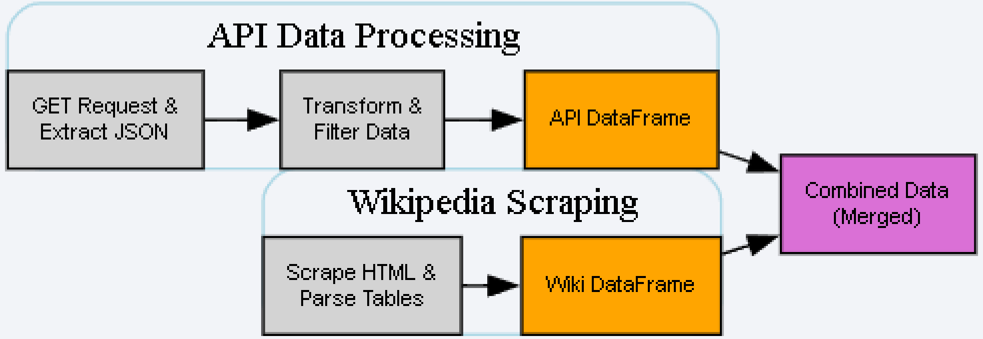
Methodology

Executive Summary

- Data collection methodology: Data on the SpaceX Falcon 9 first stage landings was collected from a public API and through Web Scrapping in Wikipedia. Additional datasets were provided in the IBM Data Scientist Coursera – Capstone in CVS file format.
- Performed data wrangling in preparation further data analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models such as Logistic Regression, k-Nearest Neighbor (kNN), Decision Trees and Support Vector Machine (SVM).

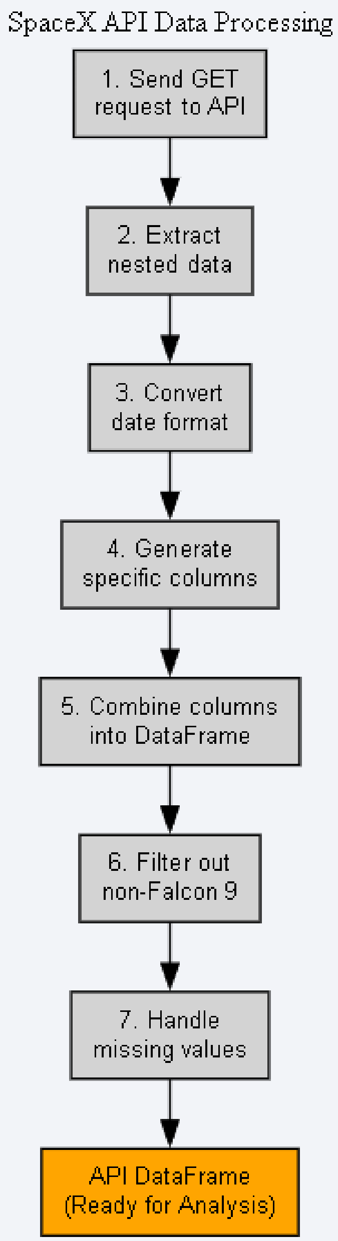
Data Collection

- The data sets were collected from:
  - An open-source accessible API with launch data in JSON format.
  - A Wikipedia page with launch data in HTML tables.
  - Additional datasets were provided in the IBM Data Scientist Coursera – Capstone in CVS file format.



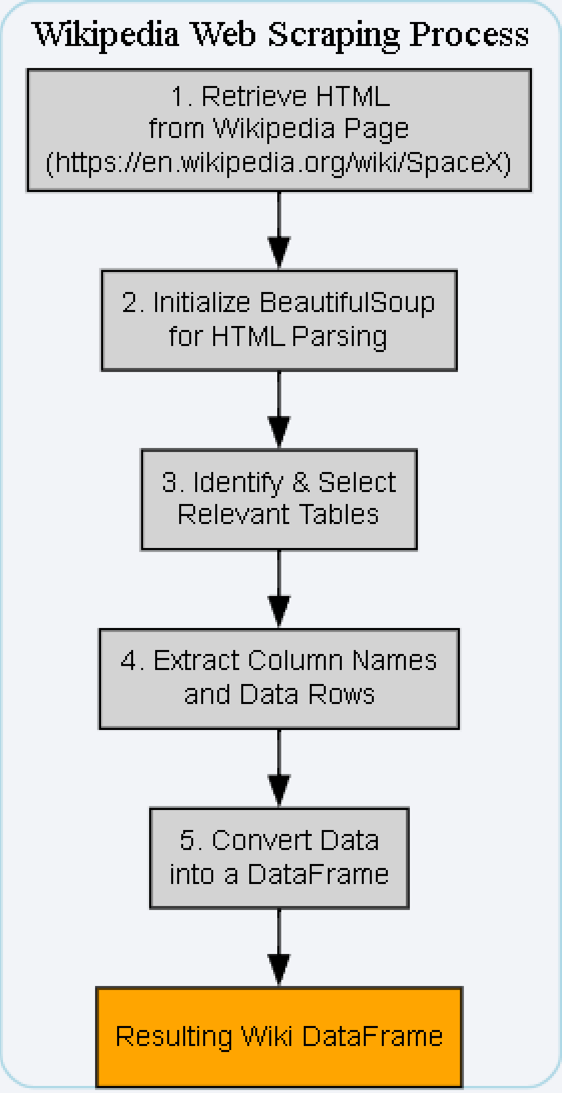
Data Collection – SpaceX API

- Space X data was collected from the API endpoint: https://api.spacexdata.com/
- Data was extracted from the API and loaded into a Pandas DataFrame to further analysis.
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/44f9837fc0bce4a531160e5cefb7a70c239a5f8a/code/jupyter-labs-spacex-data-collection-api.ipynb


SpaceX API Data Processing
1. Send GET request to API
2. Extract nested data
3. Convert date format
4. Generate specific columns
5. Combine columns into DataFrame
6. Filter out non-Falcon 9
7. Handle missing values
API DataFrame (Ready for Analysis)

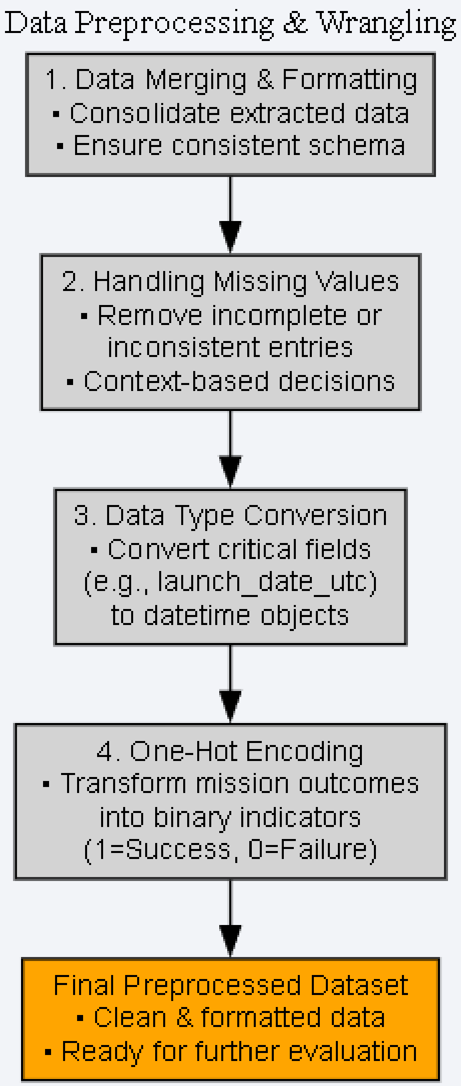
Data Collection - Scraping

- SpaceX launch data was scraped from HTML tables from a link of SpaceX Wikipedia page. (https://en.wikipedia.org/wiki/SpaceX)
- Launch data was extracted from these tables and loaded into a Pandas DataFrame for further analysis.
- GitHub URL of the completed web scraping notebook: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/44f9837fc0bce4a531160e5cefb7a70c239a5f8a/code/jupyter-labs-webscraping.ipynb


Wikipedia Web Scraping Process
1. Retrieve HTML from Wikipedia Page (https://en.wikipedia.org/wiki/SpaceX)
2. Initialize BeautifulSoup for HTML Parsing
3. Identify & Select Relevant Tables
4. Extract Column Names and Data Rows
5. Convert Data into a DataFrame
Resulting Wiki DataFrame

Data Wrangling

- Data was preprocessed to ensure dataset's consistency and usability. Some preprocessing steps we made:
  - **Handling Missing Values:** Entries with incomplete or inconsistent data were removed based on context.
  - **Data Type Conversion:** Critical fields, particularly "launch_date_utc" were converted into datetime objects to facilitate precise temporal analysis.
  - **Data Merging and Formatting:** Data was extracted, scrapped and then consolidated into a unified dataset and then processed and formatted for further evaluation.
  - One-Hot encoding: The mission outcome types were converted to a binary classification where 1 represented the Falcon 9 first stage landing being a success and 0 represented a failure.
- GitHub URL Data Wrangling: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/44f9837fc0bce4a531160e5cefb7a70c239a5f8a/code/labs-jupyter-spacex-Data%20wrangling.ipynb


Data Preprocessing & Wrangling
1. Data Merging & Formatting
• Consolidate extracted data
• Ensure consistent schema
2. Handling Missing Values
• Remove incomplete or inconsistent entries
• Context-based decisions
3. Data Type Conversion
• Convert critical fields (e.g., launch_date_utc) to datetime objects
4. One-Hot Encoding
• Transform mission outcomes into binary indicators (1=Success, 0=Failure)
Final Preprocessed Dataset
• Clean & formatted data
• Ready for further evaluation

EDA with Data Visualization

1. Launch Site Trends

- **Scatter Plots**
  - **Launch Site vs. Flight Number:** Helps determine how mission outcomes vary by launch site across different flight counts.
  - **Launch Site vs. Payload:** Shows whether payload size correlates with success or failure at specific launch sites.

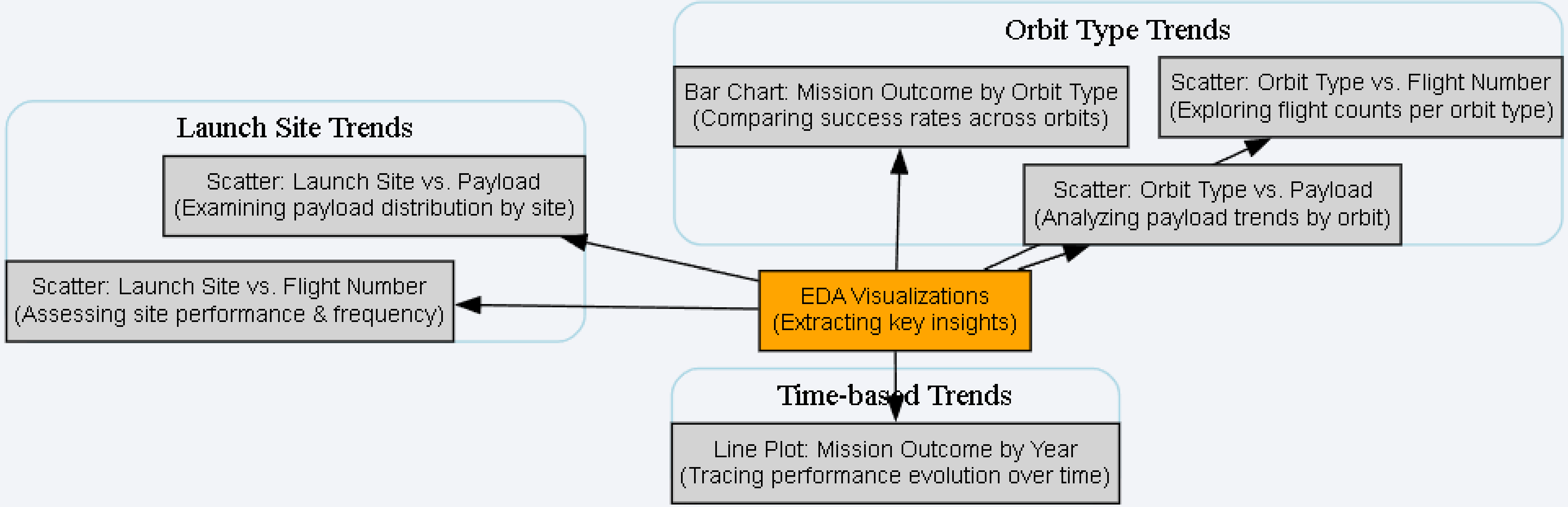
2. Orbit Type Trends

- **Bar Chart**
  - **Mission Outcome by Orbit Type:** Offers a straightforward comparison of success rates across different orbital categories.
- **Scatter Plots**
  - **Orbit Type vs. Flight Number:** Investigates how flight count and orbit type jointly influence the likelihood of a successful mission.
  - **Orbit Type vs. Payload:** Reveals any relationship between payload size, orbit type, and mission outcome.

EDA with Data Visualization

3. Time-Based Trends

- **Line Plot**
  - **Mission Outcome Trend by Year:** Highlights the evolution of success/failure rates over time, allowing for temporal analysis and identification of performance improvements.



Orbit Type Trends
Bar Chart: Mission Outcome by Orbit Type (Comparing success rates across orbits)
Scatter: Orbit Type vs. Flight Number (Exploring flight counts per orbit type)
Scatter: Orbit Type vs. Payload (Analyzing payload trends by orbit)
Launch Site Trends
Scatter: Launch Site vs. Payload (Examining payload distribution by site)
Scatter: Launch Site vs. Flight Number (Assessing site performance & frequency)
EDA Visualizations (Extracting key insights)
Time-based Trends
Line Plot: Mission Outcome by Year (Tracing performance evolution over time)

- GitHub URL EDA Visualization: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/b77dfe332aaa432a892409bd0da9eec25d243353/code/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

EDA with SQL

- SQL queries were written to extract information about:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes
- GitHub URL (EDA with SQL): https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/44f9837fc0bce4a531160e5cefb7a70c239a5f8a/code/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Interactive Map with Folium

- Certain **map objects** were created and added to the Folium Map
  - **Markers:**
    - Placed at key locations, including each launch site and the **NASA Johnson Space Center**. These markers serve as precise location indicators, enabling viewers to easily identify critical facilities on the map.
  - **Circles:**
    - Drawn around the launch sites to visually emphasize their spatial extent and significance. They help in quickly identifying the areas directly associated with each launch facility.
  - **Lines:**
    - Added to illustrate the distances from **CCAFS LC-40** to nearby geographic features:
      - Coastline: Indicates proximity to water bodies, which is essential for understanding environmental and logistical considerations.
      - Rail Line: Highlights connectivity and accessibility for transportation.
      - Perimeter Road: Demonstrates infrastructure availability around the launch site, relevant for operational planning.
- GitHub URL Folium Map: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/44f9837fc0bce4a531160e5cefb7a70c239a5f8a/code/lab_jupyter_launch_site_location.ipynb

Interactive Dashboard with Plotly Dash

Some elements were added with Plotly Dash to visualize properly the information.

- Pie Chart:
  - **Functionality:** Displays the distribution of successful vs. failed Falcon 9 first stage landings for a selected launch site (“one”) or across all sites (“all”).
  - **Purpose:** Offers a quick visual comparison of landing outcomes, allowing users to identify performance variations between different sites.
- Scatter Plot
  - **Functionality:** Shows the distribution of Falcon 9 first stage landings split by payload mass, mission outcome, and booster version category.
  - **Purpose:** Enables detailed exploration of how payload mass correlates with landing success, while highlighting the influence of booster versions.

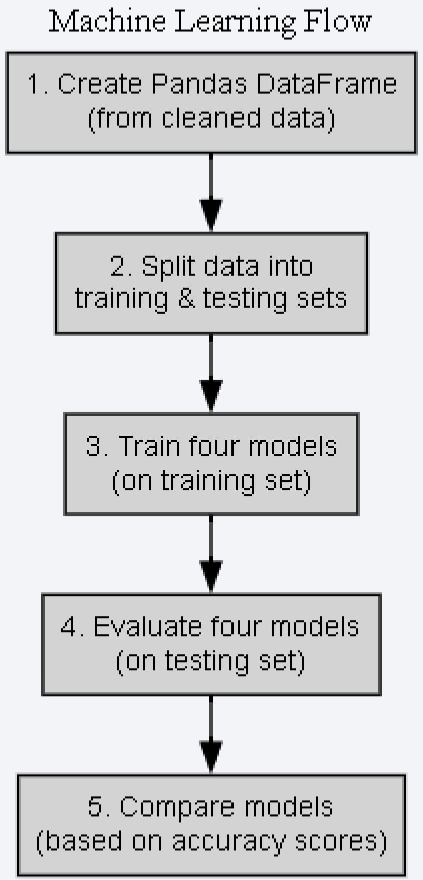
Interactive Dashboard with Plotly Dash

Also, there were some interactive elements added to enhance user engagement and analytical depth. By allowing the user to dynamically refine the visualizations, these controls support both broad and focused insights into Falcon 9 landing performance.

- **Dropdown Input**
  - **Functionality:** Allows selection between analyzing a single launch site or aggregating data from all sites for both the pie chart and scatter plot.
  - **Purpose:** Facilitates comparative analysis by letting the user switch between site-specific metrics and a holistic view of all sites.
- **Slider**
  - **Functionality:** Filters the scatter plot by a range of payload masses.
  - **Purpose:** Provides granular control over the data displayed, enabling users to focus on specific payload intervals and examine their impact on mission outcomes.
- GitHub URL Plotly Dash lab: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/44f9837fc0bce4a531160e5cefb7a70c239a5f8a/code/Interactive%20Dashboard%20with%20Ploty%20Dash.pdf

Predictive Analysis (Classification)

- The dataset was split into training and testing sets (80/20 respectively).
- The following machine learning models were trained on the training data set:
  - Logistic Regression
  - SVM (Support Vector Machine)
  - Decision Tree
  - kNN (k-Nearest Neighbors)
- Hyper-parameters were evaluated using **GridSearchCV()** and the best was selected using the **best_params method**.
- GitHub URL (Machine Learning Prediction Part 5): https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/b77dfe332aaa432a892409bd0da9eec25d243353/code/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb


Machine Learning Flow
1. Create Pandas DataFrame (from cleaned data)
2. Split data into training & testing sets
3. Train four models (on training set)
4. Evaluate four models (on testing set)
5. Compare models (based on accuracy scores)

Predictive Analysis (Classification)

- Some metrics were considered to evaluate the performance of the models.
  - TP (True Positives): Cases where successful landings were correctly predicted.
  - TN (True Negatives): Cases where landing failures were correctly predicted.
  - FP (False Positives): Instances where the model erroneously predicted a successful landing.
  - FN (False Negatives): Instances where the model erroneously predicted a landing failure when the landing was in fact successful.
- GitHub URL (Machine Learning Prediction Part 5): https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/blob/b77dfe332aaa432a892409bd0da9eec25d243353/code/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

Section 2

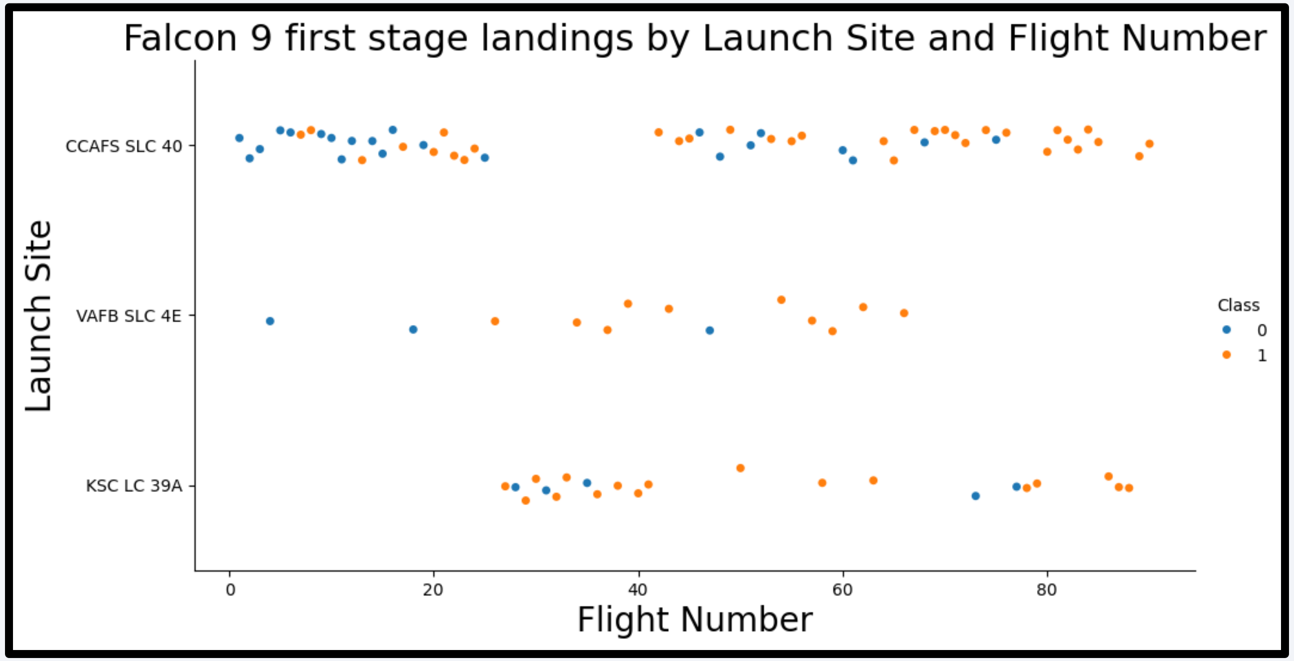
Insights drawn from EDA

Results

- Insights Drawn from EDA (Exploratory Data Analysis)
    - Exploratory Data Analysis – Data Visualizations
    - Exploratory Data Analysis – SQL Queries
- Launch Sites Proximities Analysis
    - Interactive Folium Maps (Screenshots)
- Build a Dashboard with Plotly Dash
    - Interactive Plotly Dash Dashboard (Screenshots)
- Predictive Analysis (Classification)
    - Predictive Analysis (Classification) – Machine Learning and Metrics Comparison

Flight Number vs. Launch Site

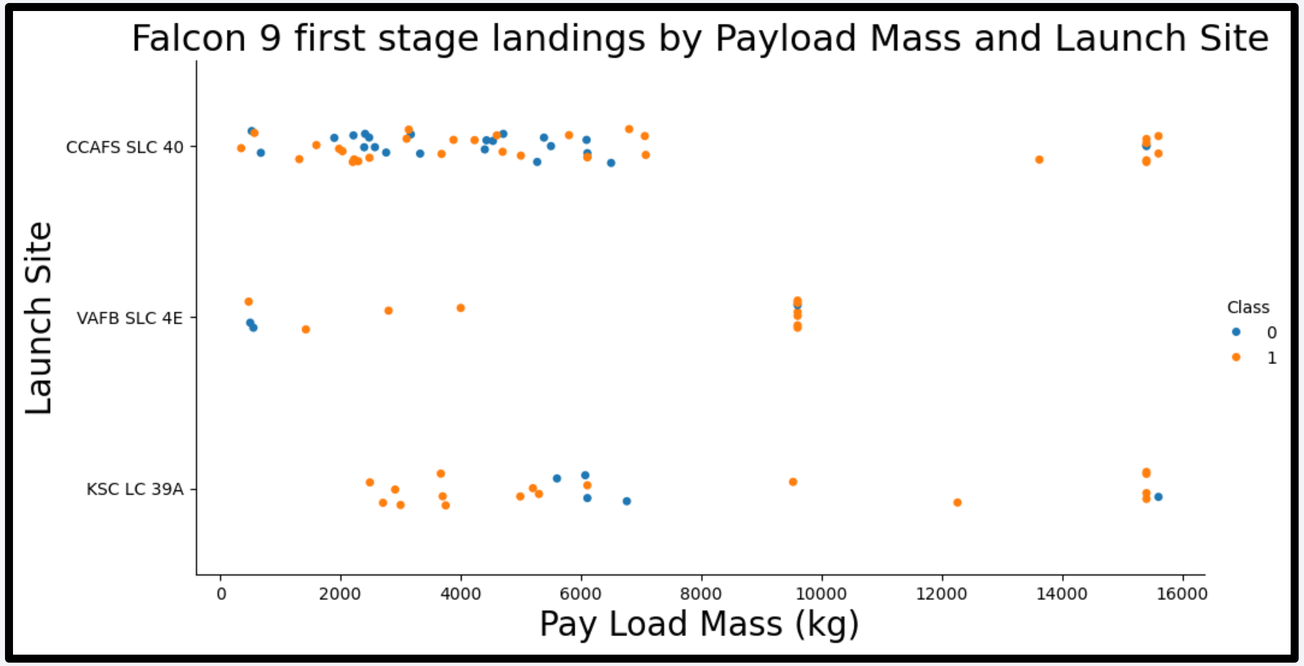
- Growing Success with Experience: Higher flight numbers show more successful (orange) landings, suggesting improved reliability over time.
- Site-Specific Variations: Each launch site has a distinct distribution of successes and failures, hinting at differing operational efficiencies.
- Reduced Failures at Higher Flight Counts: Fewer red markers beyond flight number 60 indicate learning-curve effects and ongoing refinements.
- Clusters by Site: Launch sites form separate horizontal bands, revealing unique patterns in mission outcomes across flight numbers.



- Falcon 9 **first stage failed landings** are indicated by the **'0' Class (● *blue markers*).**
- Falcon 9 **first stage successful landings** are indicated by the **'1' Class (● orange markers)**

Payload vs. Launch Site

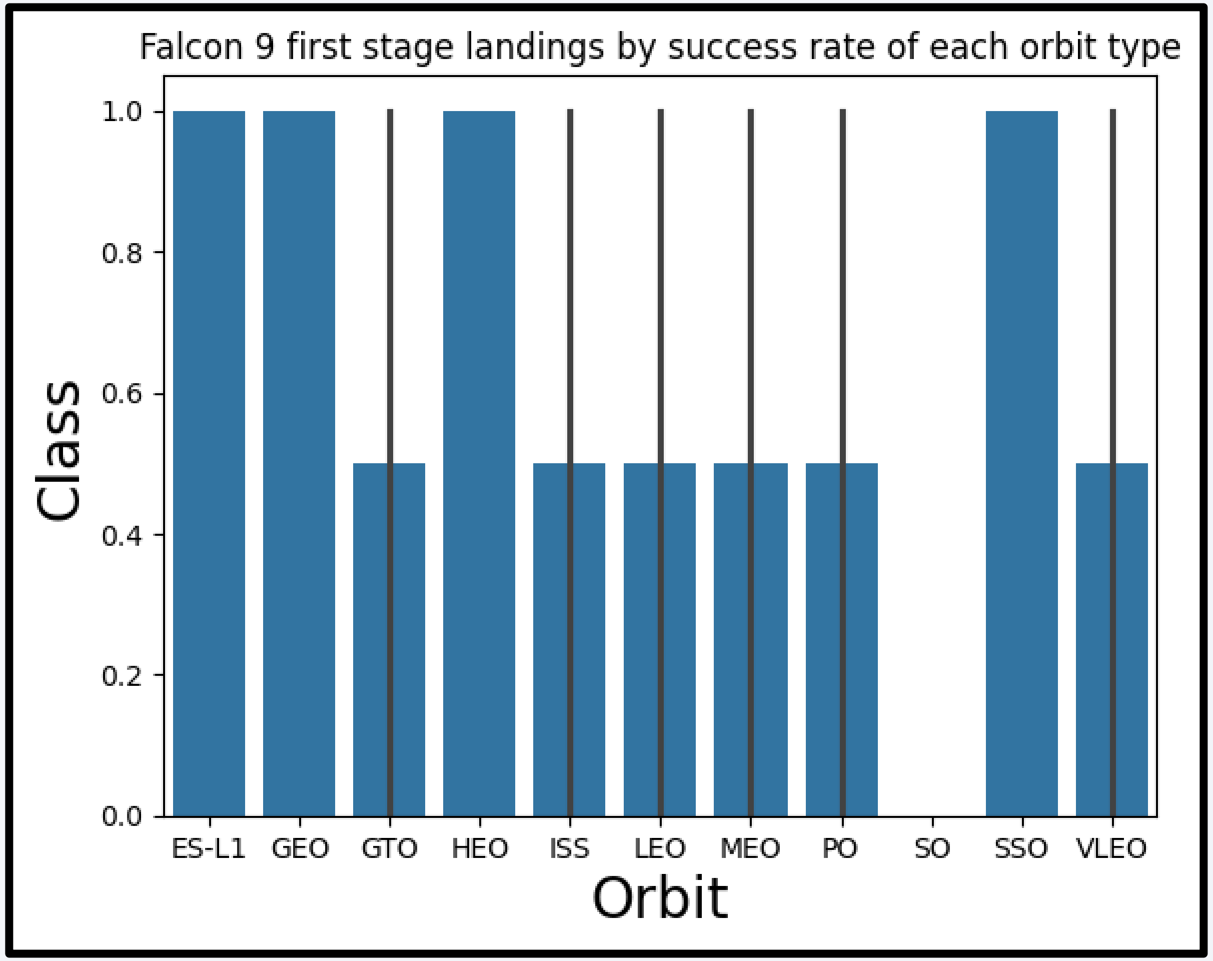
- Higher Payloads at Specific Sites: Some launch sites (CCAFS SLC 40) handle heavier payloads, indicating more robust infrastructure or mission profiles.
- Success Distribution: Orange points (success) spread across various payload ranges, but heavier payloads tend to concentrate at specific sites.
- Site-Specific Payload Capacity: Each horizontal band reveals how different sites accommodate distinct payload masses, potentially influencing landing outcomes.



- Falcon 9 **first stage failed landings** are indicated by the **'0' Class (● *blue markers*).**
- Falcon 9 **first stage successful landings** are indicated by the **'1' Class (● orange markers)**

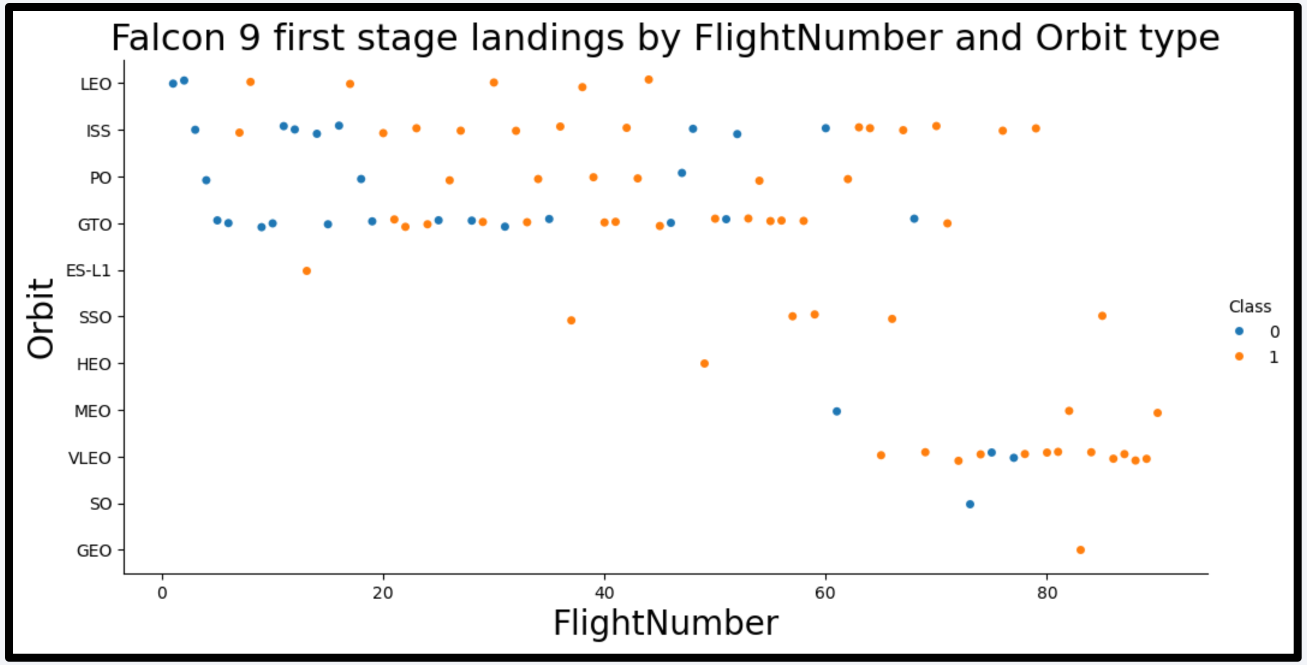
Success Rate vs. Orbit Type

- High Success in Certain Orbits: ES-L1, GEO, HEO, and SSO show near 100% success, indicating reliable performance in these orbital regimes.
- Moderate Success for Others: Orbits like GTO, MEO, PO, and VLEO exhibit mid-range success rates, suggesting varying operational complexities.
- Potential Difficulty with ISS & LEO: Lower success rates may reflect more stringent mission parameters or technical challenges specific to these orbits.



Flight Number vs. Orbit Type

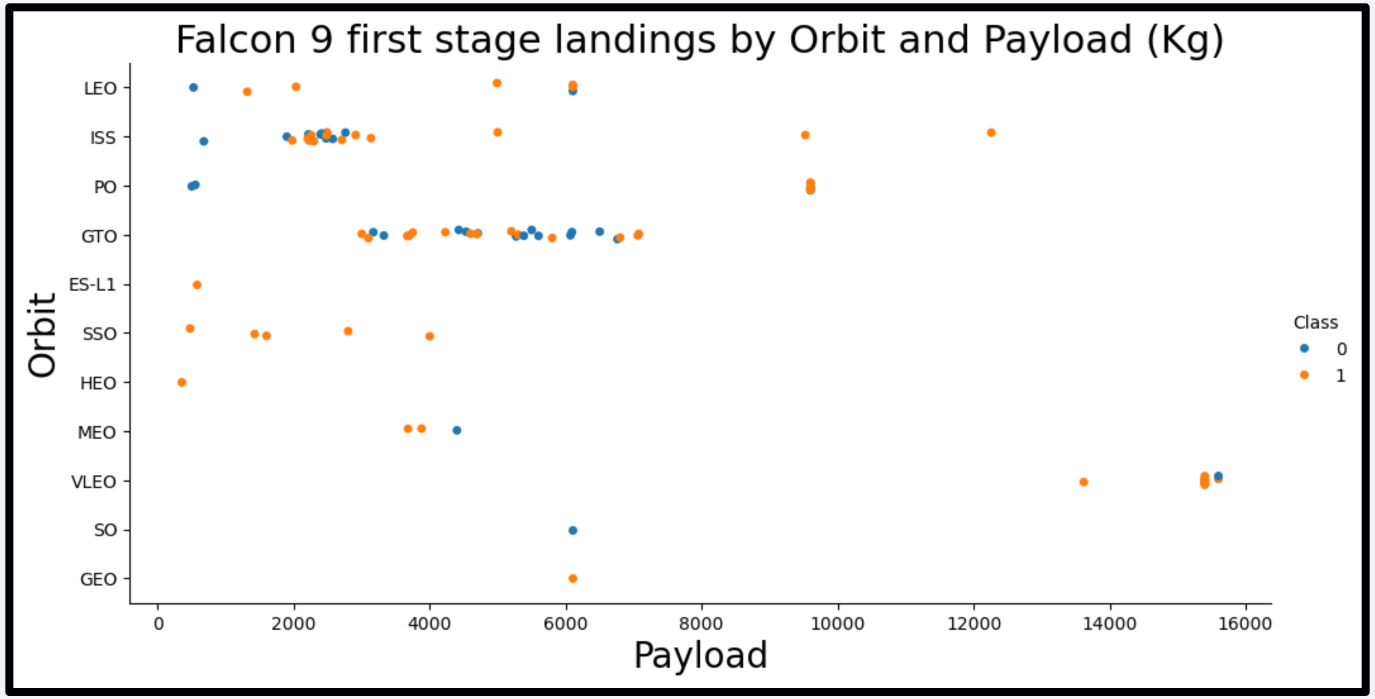
- Orbit-Specific Trends
  - Certain orbits (e.g., LEO, ISS) show more dense clusters at lower flight numbers, while others (e.g., VLEO, GEO) appear later in the sequence.
- Increasing Success with Experience
  - Higher flight numbers (beyond ~60) generally exhibit a larger proportion of successful (orange) landings across various orbits.
- Variation by Orbit Complexity
  - Some orbits with greater technical challenges (e.g., GTO, HEO) may show a mix of successes and failures, highlighting the impact of mission complexity on outcomes.
- Potential for Further Analysis
  - Future work could investigate how payload mass or booster version interacts with orbit type to influence success rates.



- Falcon 9 **first stage** failed **landings** are indicated by the **'0' Class (● *blue markers*).**
- Falcon 9 **first stage** successful **landings** are indicated by the **'1' Class (● orange markers)**

Payload vs. Orbit Type

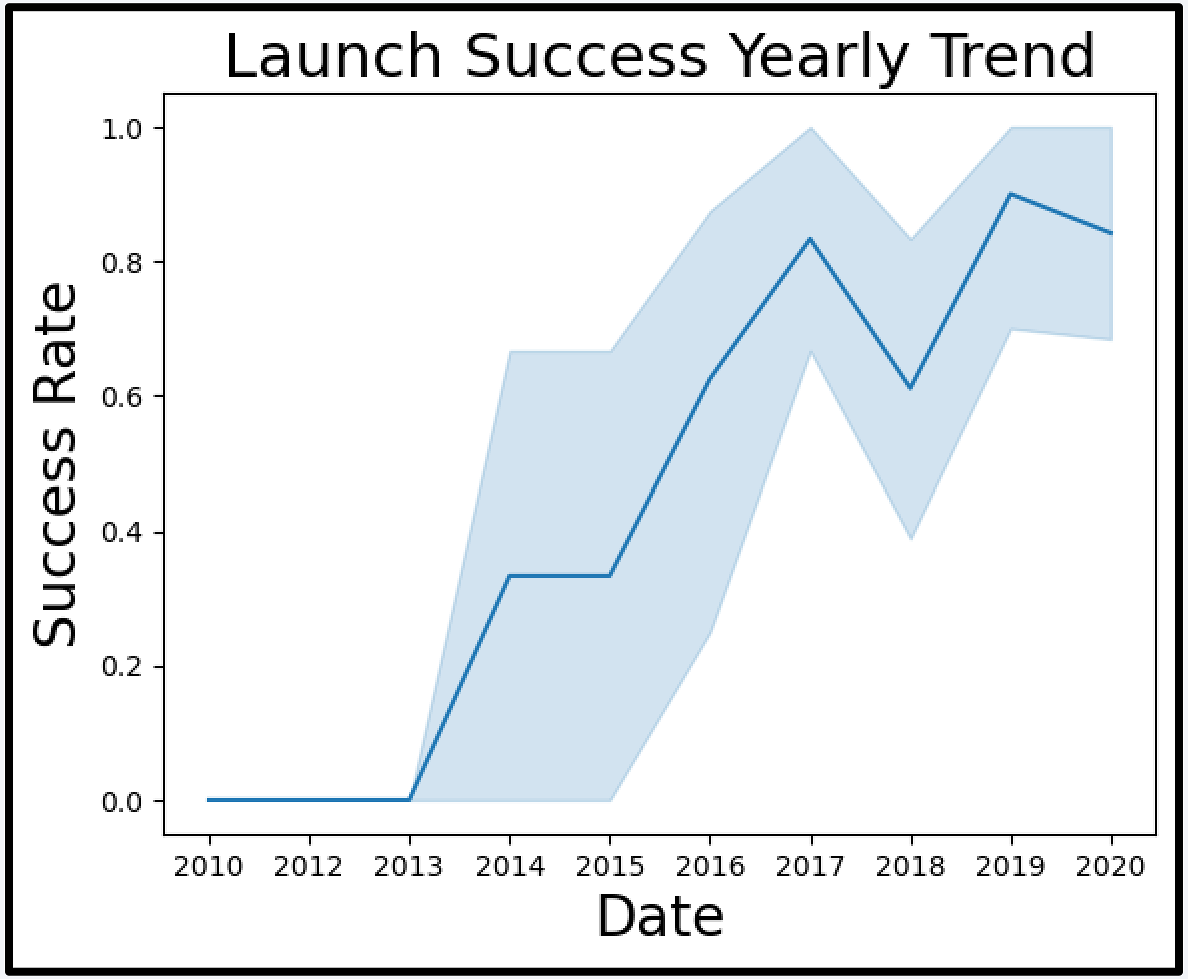
- Heavier Payloads with Higher Success: Orbits like Polar, LEO, and ISS often support larger payloads with a relatively high success rate.
- Mixed Outcomes in GTO: Both successful and unsuccessful landings appear at similar payload ranges, making it harder to distinguish a clear trend.
- Orbit-Specific Payload Ranges: Each orbit displays a unique payload span, highlighting different mission profiles and complexities.



- Falcon 9 **first stage** **failed** **landings** are indicated by the **'0' Class (● *blue markers*).**
- Falcon 9 **first stage** **successful** **landings** are indicated by the **'1' Class (● orange markers)**

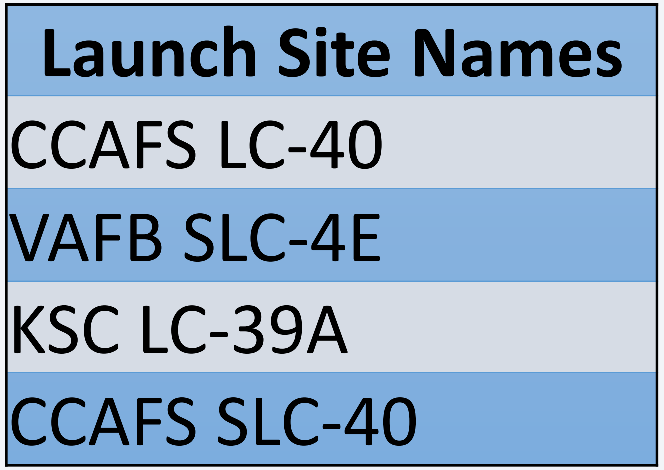
Launch Success Yearly Trend

- Steady Improvement Over Time: Success rate shows a marked climb from near zero in 2013 to consistently high values by 2020, reflecting growing operational experience.
- Near-Perfect Performance in Recent Years: By 2020, success rates hover around 90–100%, suggesting mature launch processes and continuous refinements.



All Launch Site Names

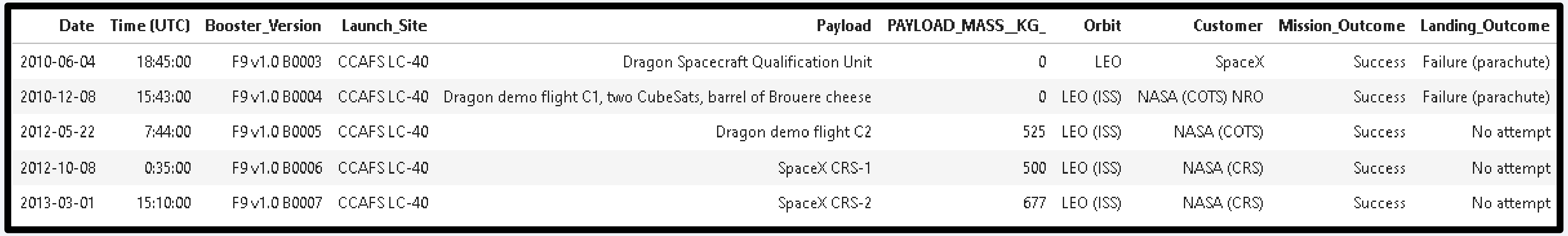
- Task: Display the names of the unique launch sites in the space mission
- Query:
  - SELECT DISTINCT(LAUNCH_SITE)
  - FROM SPACEXTBL;
- Result: By using the DISTINCT keyword on the LAUNCH_SITE column, the query retrieves only the unique site names.
- The result shows four distinct launch sites in the dataset.



Launch Site Names
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Task: Find 5 records where launch sites begin with `CCA`
- Query:
  - SELECT *
  - FROM SPACEXTBL
  - WHERE LAUNCH_SITE LIKE 'CCA%'
  - LIMIT 5;
- Result:

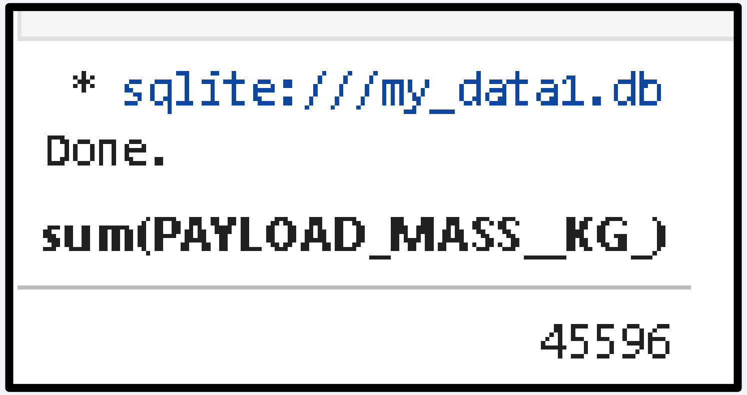


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- The LIKE 'CCA%' condition ensures only records whose launch site names start with “CCA” are returned. LIMIT 5 restricts the output to the first five matching rows. This query shows five missions that launched from CCAFS LC-40.

Total Payload Mass

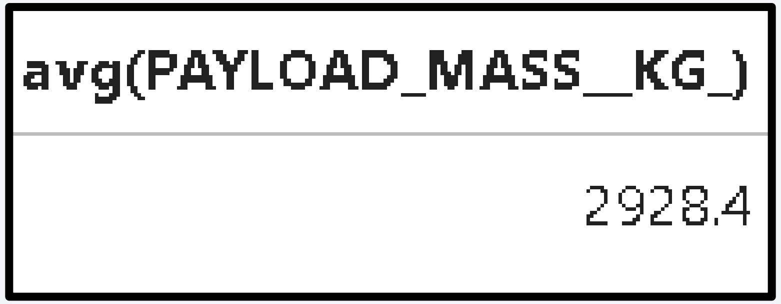
- Task: Display the total payload mass carried by boosters launched by NASA (CRS).
- Query:
  - SELECT SUM (PAYLOAD_MASS__KG_)
  - FROM SPACEXTBL
  - WHERE CUSTOMER = 'NASA (CRS)’;
- Result: The SUM() function aggregates the total payload mass (PAYLOAD_MASS__KG_) for records where the CUSTOMER is NASA (CRS). This yields the combined payload mass of **45596 Kg**



```
 * sqlite:///my_data1.db
Done.
sum(PAYLOAD_MASS__KG_)
                 45596
```

Average Payload Mass by F9 v1.1

- Task: Display the average payload mass carried by booster version F9 v1.1.
- Query:
  - SELECT AVG(PAYLOAD_MASS__KG_)
  - FROM SPACEXTBL
  - WHERE BOOSTER_VERSION = 'F9 v1.1';

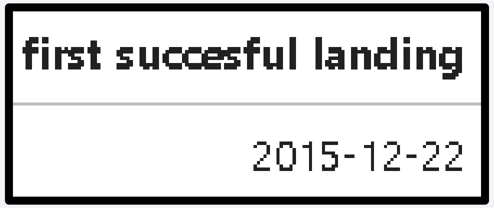


avg(PAYLOAD_MASS__KG_)
2928.4

- Result: The AVG() function calculates the mean payload mass of flights specifically using the F9 v1.1 booster version. This provides a straightforward measure of typical payload size for that booster with an average of **2928.4 Kg.**

First Successful Ground Landing Date

- Task: List the date when the first successful landing outcome on a ground pad was achieved.
- Query:
  - SELECT MIN("Date") AS "first successful landing"
  - FROM SPACEXTBL
  - WHERE "Landing_Outcome" = 'Success (ground pad)’;

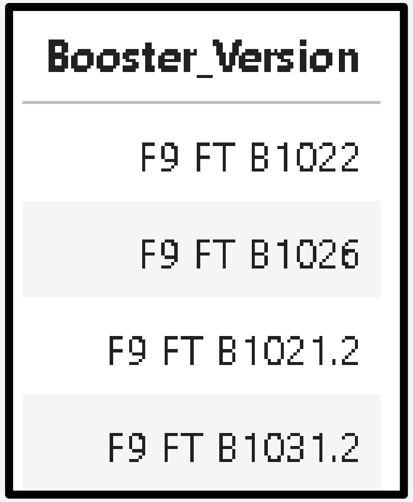


first succesful landing
2015-12-22

- Result: By using the MIN() function on the "Date" column, the query retrieves the earliest recorded date where the Landing Outcome was Success (ground pad). This identifies the first successful ground-pad landing in the dataset:
  - Date: 12- 22- 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

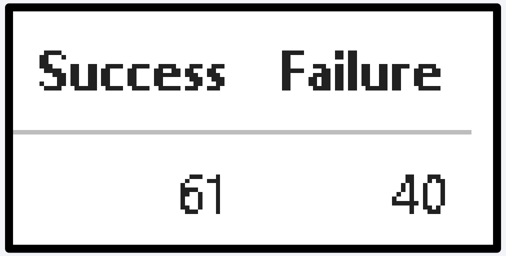
- Task: List the names of the boosters which have a successful landing on a drone ship and carry a payload mass greater than 4000 kg but less than 6000 kg.
- Query:
  - SELECT BOOSTER_VERSION
  - FROM SPACE"Landing_Outcome" XTBL
  - WHERE = 'Success (drone ship)'
  - AND "PAYLOAD_MASS__KG_" > 4000
  - AND "PAYLOAD_MASS__KG_" < 6000;
  - Result: This query filters flights by Landing_Outcome = ‘Success (drone ship)’ and restricts the PAYLOAD_MASS__KG_ to values between 4000 and 6000.



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

Total Number of Successful and Failure Mission Outcomes

- Task: List the total number of successful and failure mission outcomes.
- Query:
  - SELECT
  - (SELECT COUNT(*)
  - FROM SPACEXTBL
  - WHERE LOWER("Landing_Outcome") LIKE '%success%') AS "Success",
  - (SELECT COUNT(*)
  - FROM SPACEXTBL
  - WHERE LOWER("Landing_Outcome") NOT LIKE '%success%') AS "Failure";

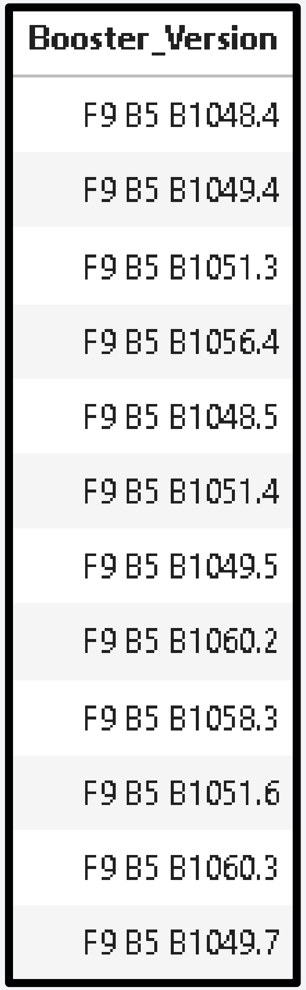


Success	Failure
61	40

- Result: This query counts the rows in which MISSION_OUTCOME is either Success or Failure (in flight).

Boosters Carried Maximum Payload

- Task: List the names of the booster versions which have carried the maximum payload mass. (Use a subquery)
- Query:
  - SELECT BOOSTER_VERSION
  - FROM SPACEXTBL
  - WHERE PAYLOAD_MASS__KG_ = (
  -     SELECT MAX(PAYLOAD_MASS__KG_)
  -     FROM SPACEXTBL
  -);
- Result: The subquery retrieves the highest payload mass (MAX(PAYLOAD_MASS__KG_)) from the table. The outer query selects all booster versions whose PAYLOAD_MASS__KG_ matches that maximum value, showing which boosters carried the heaviest payload(s).



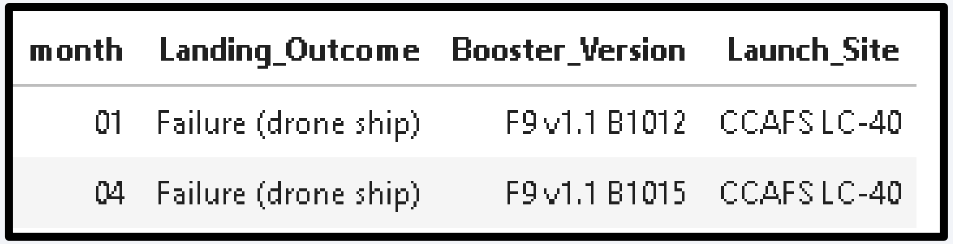
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Task: List the records displaying the month names, failure landing outcomes on drone ships, booster versions, and launch sites for the months in the year 2015.

- Query:
  - SELECT substr("Date", 6, 2) AS month,
  - "Landing_Outcome",
  - "Booster_Version",
  - "Launch_Site"
  - FROM SPACEXTBL
  - WHERE substr("Date", 0, 5) = '2015'
  - AND "Landing_Outcome" = 'Failure (drone ship)’;

- Result:
  - The substr("Date", 6, 2) extracts the month from the date string.
  - substr("Date", 0, 5) = '2015' ensures the record is from year 2015.
  - Landing_Outcome = 'Failure (drone ship)' filters records specifically for drone ship failures.
  - The result shows that in January (01) and April (04) of 2015, there were failures on the drone ship with the given booster versions.



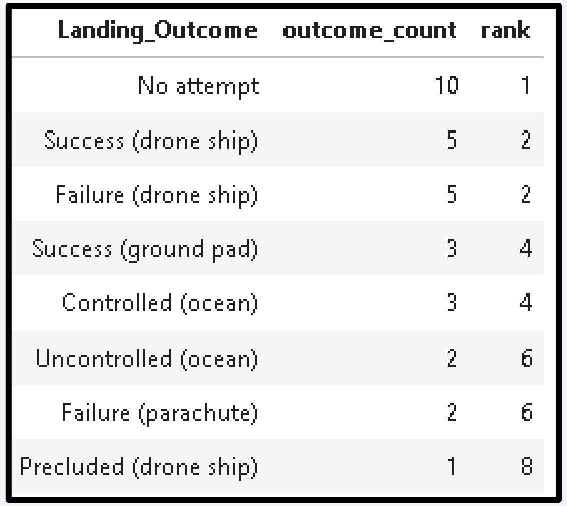
month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Task: Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between 2010-06-04 and 2017-03-20, in descending order.

- Query:
  - SELECT "Landing_Outcome",
  - COUNT(*) AS outcome_count,
  - RANK() OVER (ORDER BY COUNT(*) DESC) AS rank
  - FROM SPACEXTBL
  - WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
  - GROUP BY "Landing_Outcome";

- Results:
  - COUNT(*) calculates how many occurrences each Landing Outcome has in the specified date range.
  - RANK() OVER (ORDER BY COUNT(*) DESC) assigns a ranking based on the descending frequency of each outcome.
  - This reveals which landing outcomes occurred most frequently within the given timeframe.

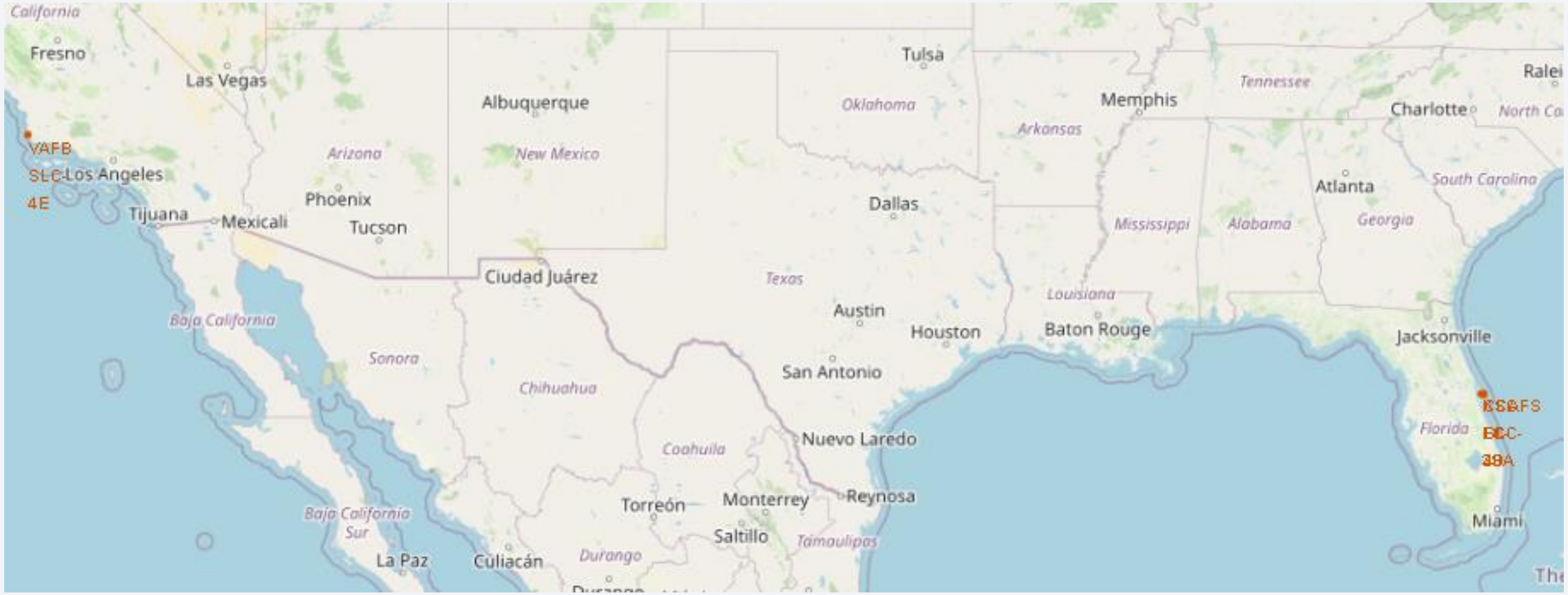


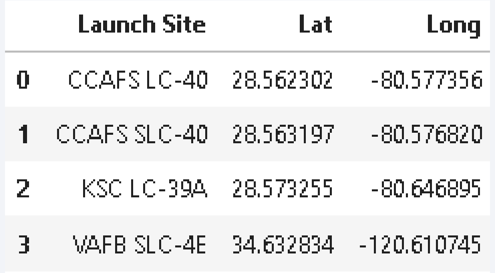
Landing_Outcome	outcome_count	rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

Section 3

Launch Sites Proximities Analysis

Falcon 9 Launch Site Locations



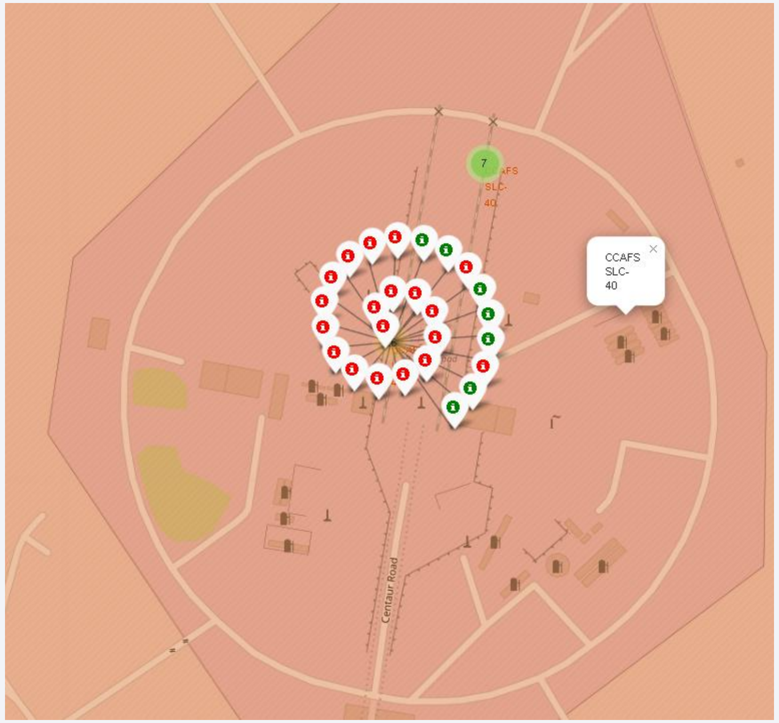


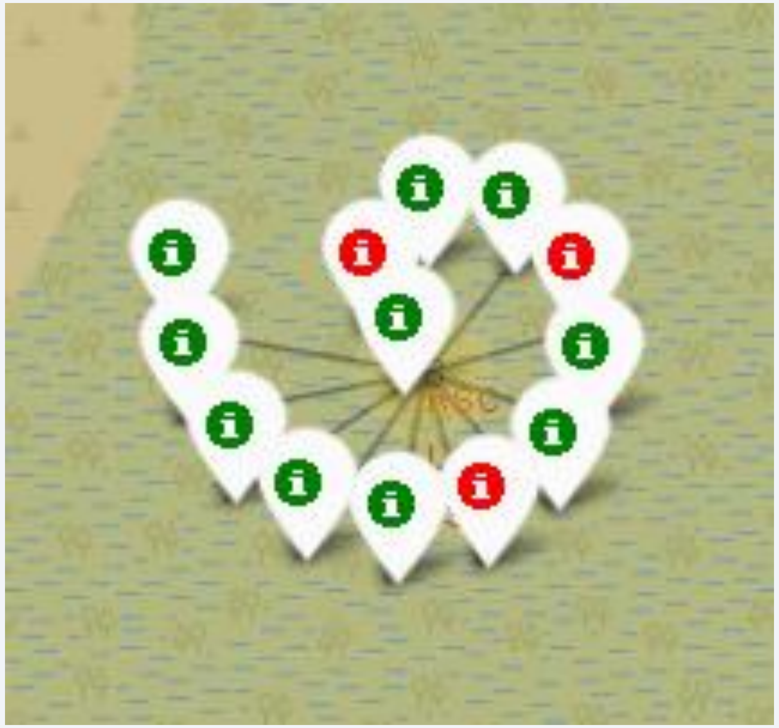
	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

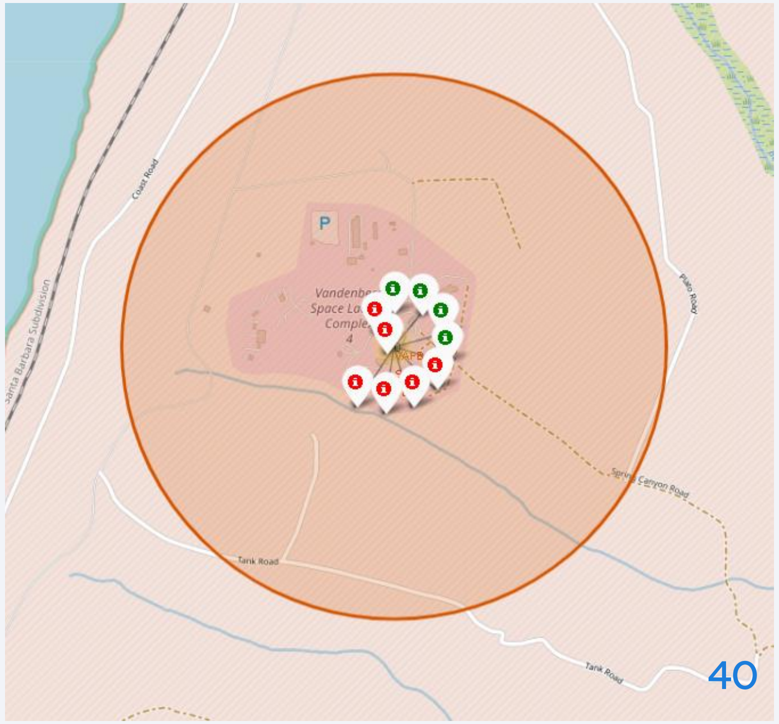
- We can see in the whole map the launch site locations for Falcon 9 in both California, USA and Florida USA.

Map Markers of landings for Falcon 9.

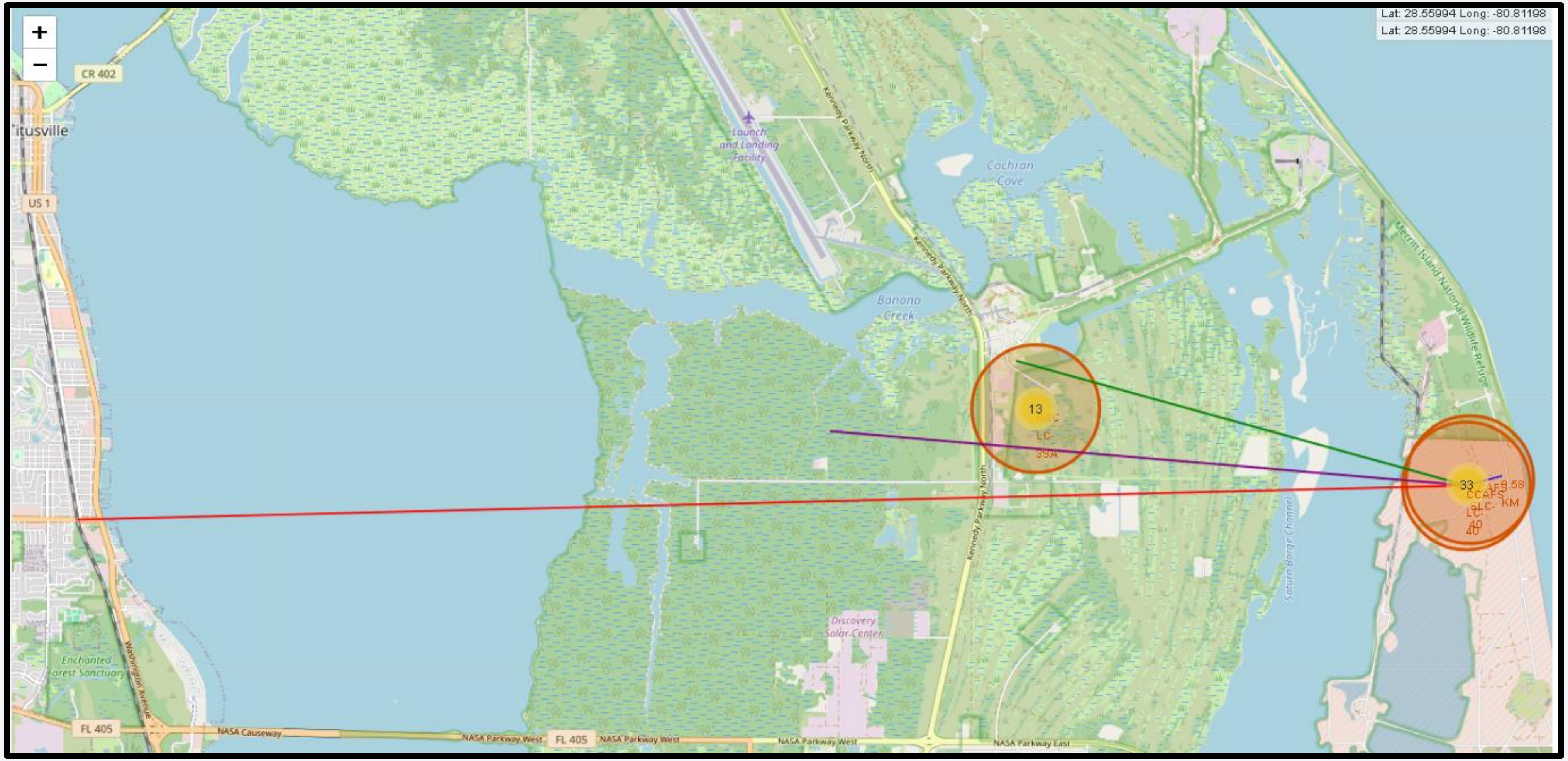
- On the map, each marker denotes the outcome (Success or Failure) of a Falcon 9 first-stage landing, positioned according to the launch site's geographic coordinates. By comparing the relative number of green (successful) markers to red (failed) markers at each location, one can infer the approximate success rate for Falcon 9 first-stage landings at that site. From left to right we can see CCAFS LC-40, KSC LC-39A, VAFB SLC-4E.







Distance from Launch Site to Proximities

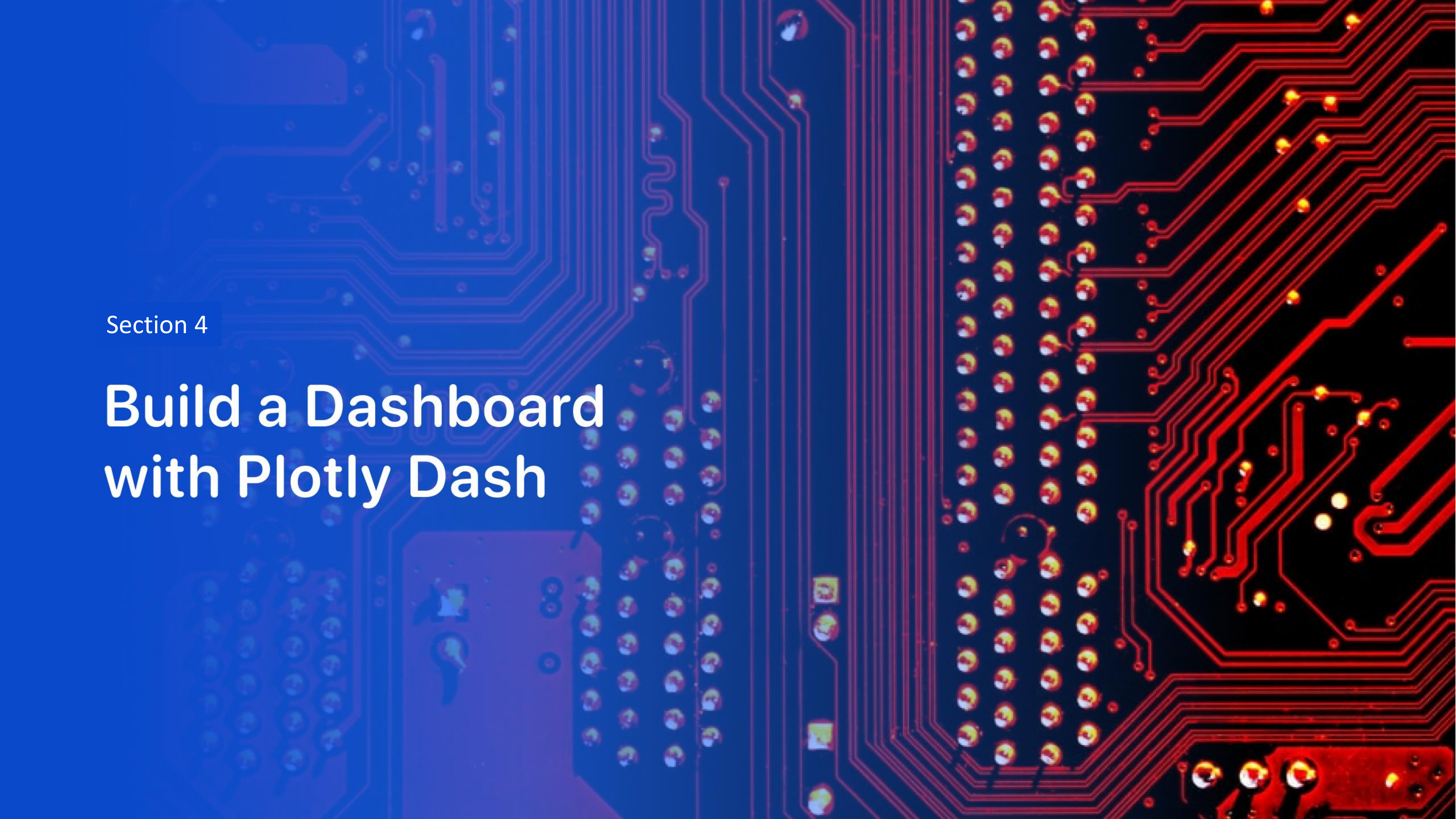


Distance from Launch Site to Proximities

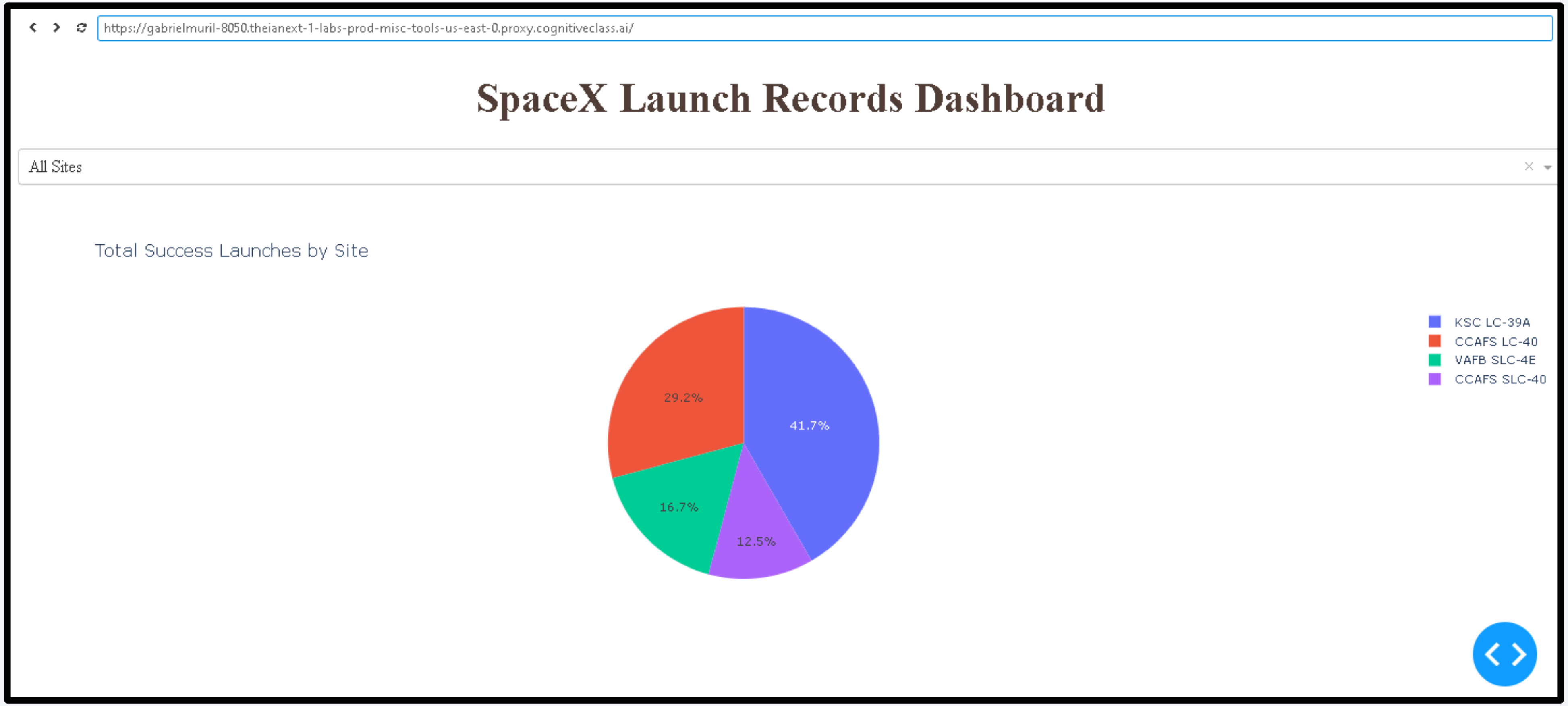
- Are launch sites in close proximity to railways?
  - Yes. The rail line is approximately 1.33 km from CCAFS LC-40. This relatively short distance indicates convenient access to rail transport for heavy equipment and materials.
- Are launch sites in close proximity to highways?
  - Yes. The perimeter road is only 0.19 km away from the launch site, suggesting that ground transportation routes are readily available for personnel and logistical needs.
- Are launch sites in close proximity to coastline?
  - Yes. The coastline is located about 0.92 km from CCAFS LC-40. Such a short distance is advantageous for over-water launch trajectories, minimizing risks to populated areas and simplifying booster recovery operations at sea.
- Do launch sites keep a certain distance away from cities?
  - Although the map primarily focuses on immediate surroundings, it shows a predominantly undeveloped buffer zone near the pad, indicating that major population centers are situated beyond this buffer. This separation reduces potential safety hazards to residents and complies with standard aerospace safety regulations.

Section 4

Build a Dashboard with Plotly Dash



Launch Success Count for All Sites



https://gabrielmuril-8050.theianext-1-labs-prod-misc-tools-us-east-0.proxy.cognitiveclass.ai/

SpaceX Launch Records Dashboard

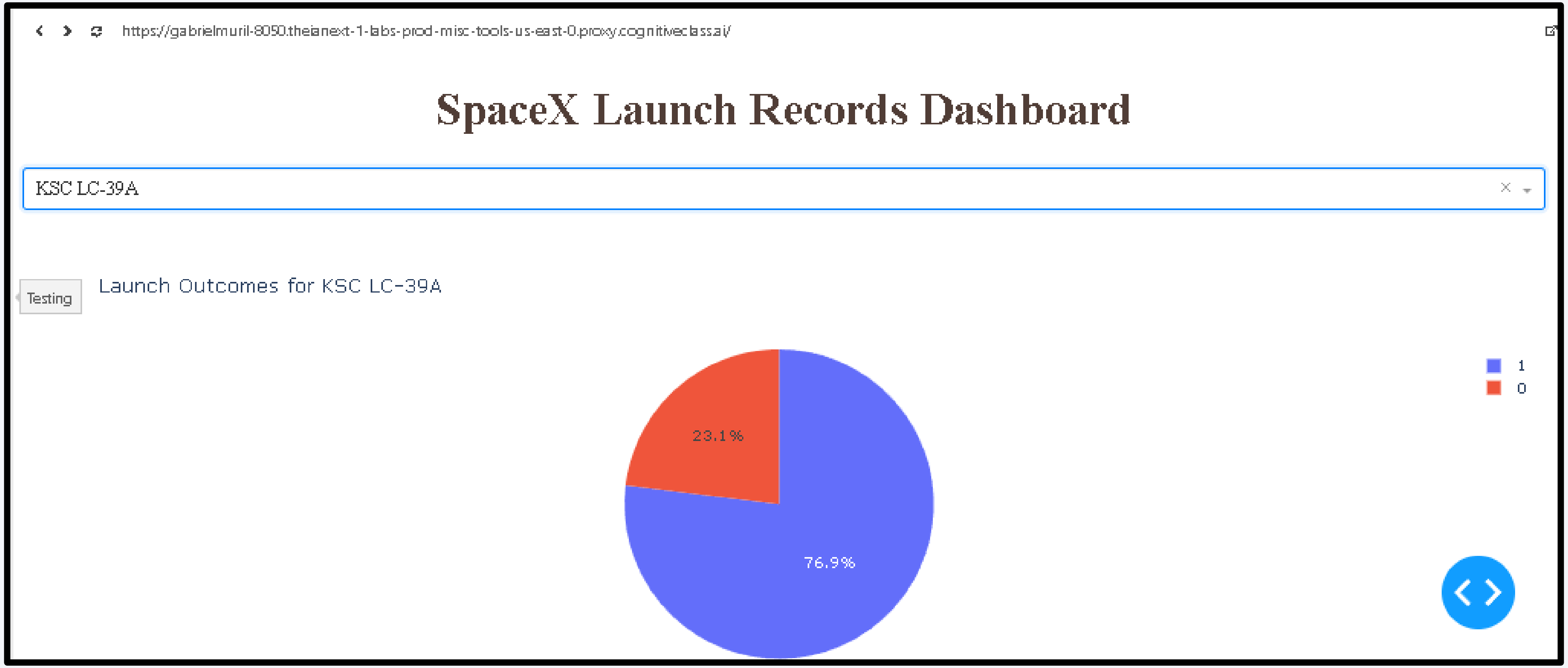
All Sites

Total Success Launches by Site

Launch Success Count for All Sites

- The dropdown menu enabled users to select either a specific launch site or aggregate data from all sites.
- When all launch sites were selected, the pie chart illustrated the distribution of successful Falcon 9 first stage landing outcomes across the different launch locations.
- Notably, KSC LC-39A accounted for the largest proportion of successful first stage landings, representing 41.7% of the total successes.

Launch Site with Highest Launch Success Ratio



https://gabrielmuril-8050.theianext-1-labs-prod-misc-tools-us-east-0.proxy.cognitiveclass.ai/

SpaceX Launch Records Dashboard

KSC LC-39A

Launch Outcomes for KSC LC-39A

Visualization Key:

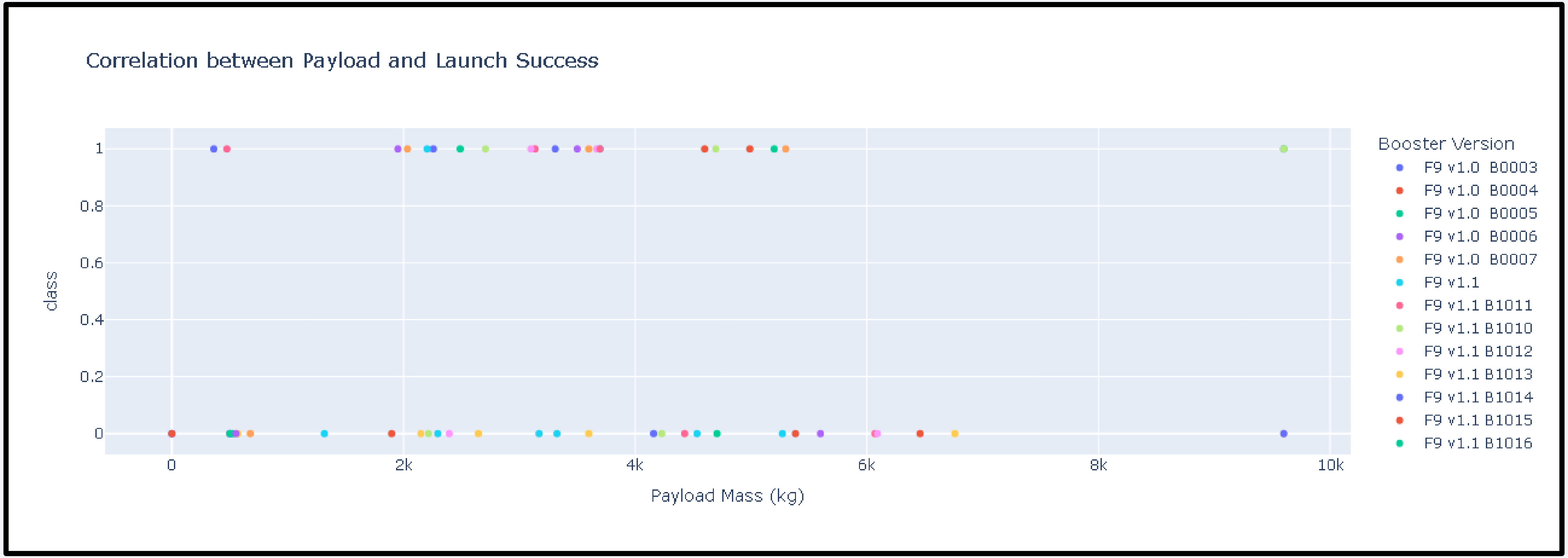
•In the pie chart, Falcon 9 first stage landing outcomes are distinguished by class:

- A **‘0’ Class** (red wedge) represents failed landings.
- A **‘1’ Class** (blue wedge) represents successful landings.

Launch Site with Highest Launch Success Ratio

- Launch Site Success Rates:
  - CCAFS SLC-40: Achieved a success rate of 73.1%.
  - KSC LC-39A: Demonstrated the highest success rate at 76.9%.
  - VAFB SLC-4E: Recorded a success rate of 60%.
- Conclusions: The blue and red segments in the pie chart clearly distinguish between successes and failures, enabling a visual assessment of each site's performance. Among the analyzed sites, **KSC LC-39A** exhibits the **most reliable performance**, while CCAFS SLC-40 and VAFB SLC-4E show lower but still substantial success rates.
- These variations suggest that operational factors and site-specific conditions may significantly influence landing outcomes, underscoring the need for further investigation into the underlying causes of performance differences.

Payload vs. Launch Outcome



Payload vs. Launch Outcome

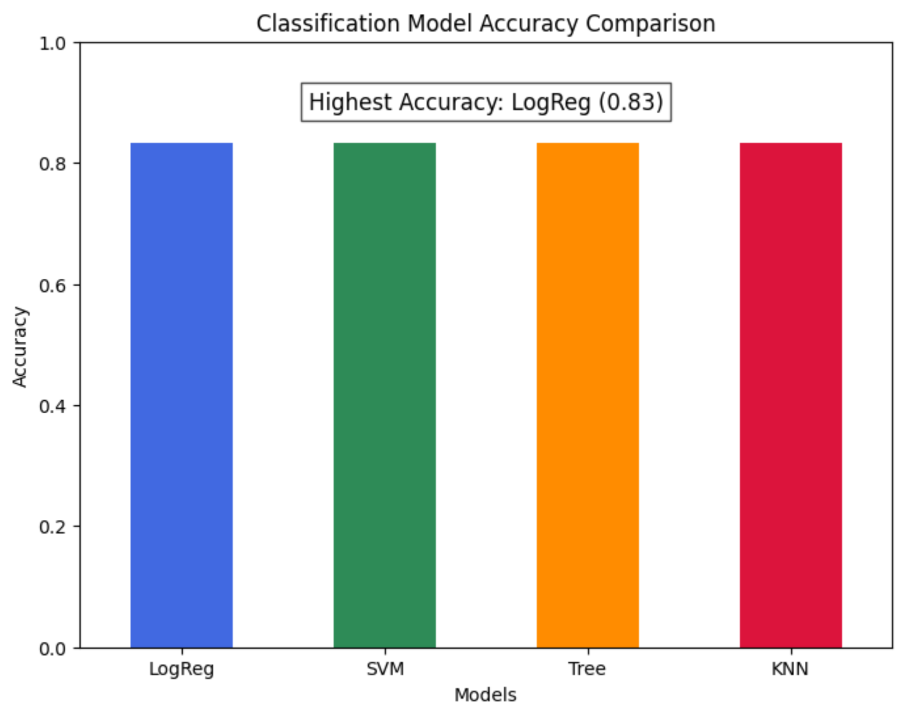
- Key Observations:
  - Wide Success Range: Successful launches (class = 1) are observed across a broad span of payload masses, indicating no strict upper limit beyond which failures dominate.
  - No Strong Negative Correlation: Heavier payloads do not appear to strictly increase failure rates. Instead, successful landings occur at various payload levels, suggesting that other factors (e.g., booster version, flight profile) are more critical to landing success.
  - Booster Version Influence: Multiple booster versions (e.g., Block 5, FT) demonstrate consistent successes over varying payload masses, reinforcing the importance of iterative design improvements on landing reliability.

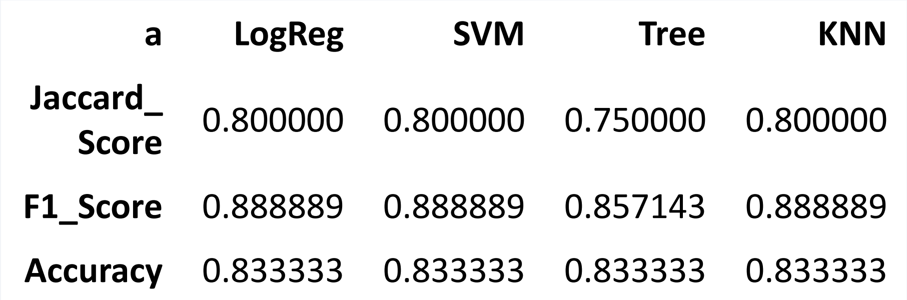
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Uniform Accuracy**: All four models (Logistic Regression, SVM, Decision Tree, KNN) achieved the same overall **accuracy of 83.33%,** indicating that each classifier successfully captures essential patterns in the data.
- **Jaccard and F1 Scores:** While accuracy is identical, the Jaccard and F1 scores reveal small differences. Logistic Regression, SVM, and KNN exhibit slightly higher Jaccard (0.80) and F1 (0.8889) than the Decision Tree (Jaccard: 0.75, F1: 0.8571), suggesting a marginal advantage in balancing precision and recall.
- **Interpretation:** The Decision Tree model, despite matching accuracy, shows a slightly higher rate of misclassifications when considering precision and recall together. Consequently, Logistic Regression, SVM, or KNN may be preferable if misclassification costs are a concern.
- **Overall Implication:** Since all models converge at the same accuracy level, other factors (e.g., interpretability, computational cost, or domain-specific constraints) may guide the final model selection for predicting Falcon 9 first-stage landing success.

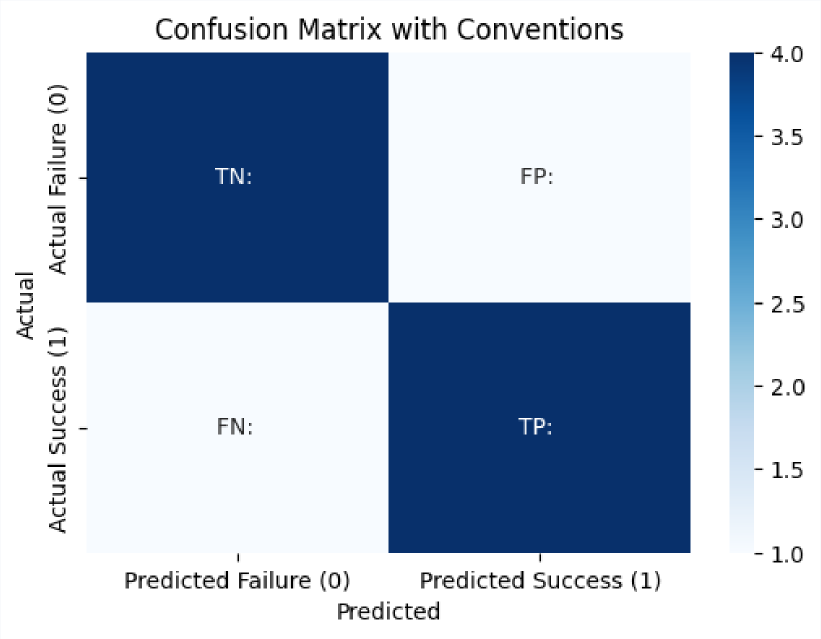


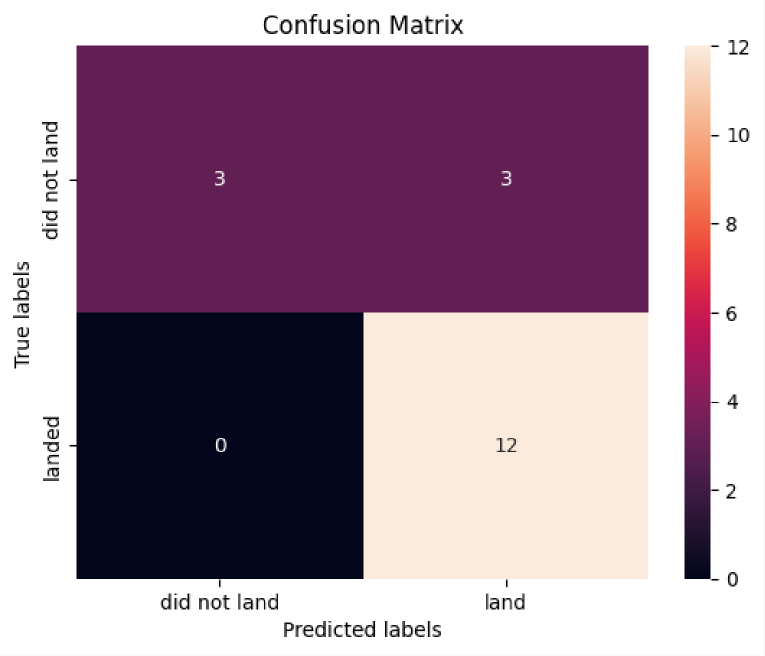


a	LogReg	SVM	Tree	KNN
Jaccard_ Score	0.800000	0.800000	0.750000	0.800000
F1_Score	0.888889	0.888889	0.857143	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Confusion Matrix

- High Recall for Landed Rockets:
  - The model correctly identifies all rockets that actually landed (12 true positives, 0 false negatives), indicating a perfect recall (100%) for the “land” class.
- Moderate Misclassification for Non-Landed Rockets:
  - Three rockets that did not land are incorrectly predicted as landed (3 false positives), suggesting the model is somewhat prone to over-predicting landings.
- Overall Implication:
  - While the model excels at capturing successful landings—ideal if the priority is to avoid missing true positives—it may incorrectly classify some non-landed rockets as landed. Stakeholders should balance this trade-off between high recall for successful landings and the risk of false positives for the “did not land” class.





Conclusions

- **Launch Site Performance Variability:**
    - Significant differences in landing success rates were observed across launch sites, with KSC LC-39A achieving the highest success rate compared to CCAFS SLC-40 and VAFB SLC-4E.
- **Operational Improvement Over Time:**
    - Temporal analysis indicates a steady increase in landing success, suggesting that iterative operational improvements and accumulated launch experience are driving enhanced performance.
- **Payload Impact on Outcomes:**
    - The analysis reveals that certain launch sites are better equipped to handle heavier payloads, demonstrating that payload mass is a critical factor influencing landing reliability.
- **Robust Data Integration:**
    - Integrating data from the SpaceX API, web scraping from Wikipedia, and provided CSV files resulted in a comprehensive dataset, thereby reducing individual data gaps and enhancing overall analysis robustness.

Conclusions

- **Enhanced Data Visualization:**
  - The use of interactive maps, SQL queries, and dashboards (via Folium and Plotly Dash) enabled dynamic exploration of data trends and performance metrics, offering clear insights into complex relationships.
- **Predictive Insights and Model Efficacy:**
  - Predictive modeling identified key factors affecting landing success and achieved competitive accuracy. These insights provide actionable intelligence for refining operational strategies and enhancing risk management in launch planning.
- **Real-World Implications:**
  - The findings support cost optimization strategies by linking successful landings to lower launch costs, thereby offering a competitive edge for operators and informing bid strategies for alternate launch providers.
- **Future Directions:**
  - The analysis underscores the need for incorporating additional variables (e.g., weather conditions, more granular operational metrics) to further improve predictive performance and operational planning.

Appendix

- Initial DataSources
  - SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - SpaceX (CSV):Detailed records of SpaceX launches are provided in this CSV file
  - Launch Geo (CSV):Contains geospatial data about the launch sites: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
  - Launch Dash (CSV):Provides additional launch dashboard metrics in CSV format: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csV
- GitHub repository with the code made in this Project: https://github.com/GabrielSMurillo/spacex-IBM_DataScientist_capstone-project/

Thank you!