IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Charity Orhoridiohwo
04/12/2025

Outline

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

Executive Summary

This project predicts the likelihood of a SpaceX Falcon 9 first stage landing successfully using historical launch data.

The goal is to help estimate mission reliability and support cost-saving decisions based on rocket reusability.

Methodology Summary

- Data was collected from the SpaceX API and web scraping.
- Data wrangling included cleaning, handling missing values, and one-hot encoding categorical features.
- Exploratory Data Analysis (EDA) revealed that payload mass, orbit type, launch site, and booster reuse strongly affect landing success.
- Visualization tools (Matplotlib, Folium, Plotly Dash) were used to explore geographic patterns and build an interactive dashboard.
- Machine learning models such as Logistic Regression, SVM, Decision Tree, and KNN were trained with GridSearchCV and 10-fold cross-validation.

Executive Summary Contd.

Results Summary

- Logistic Regression performed best with an accuracy of about 0.833 and the fewest misclassifications, correctly predicting most successes and failures. .
- Lighter payloads, certain orbits (e.g., LEO), and specific launch sites show higher success rates.
- Reused boosters tend to have improved landing success.
- The interactive dashboard helps to easily explore launch factors and outcomes.

Introduction

SpaceX aims to reduce the cost of space travel by reusing Falcon 9 rockets, making the success of first-stage landings extremely important. This project analyzes historical SpaceX data to understand which factors influence landing success and to build models that can predict whether a landing will succeed.

The key questions the project aims to answer are:

- Which variables (payload, orbit, launch site, booster reuse, etc.) most affect landing success?
- Can we accurately predict landing outcomes using machine learning?
- How do launch site and orbit type influence success rates?
- Do reused boosters perform differently from new ones?
- Can interactive visualizations help stakeholders understand these patterns?

Section 1

Methodology

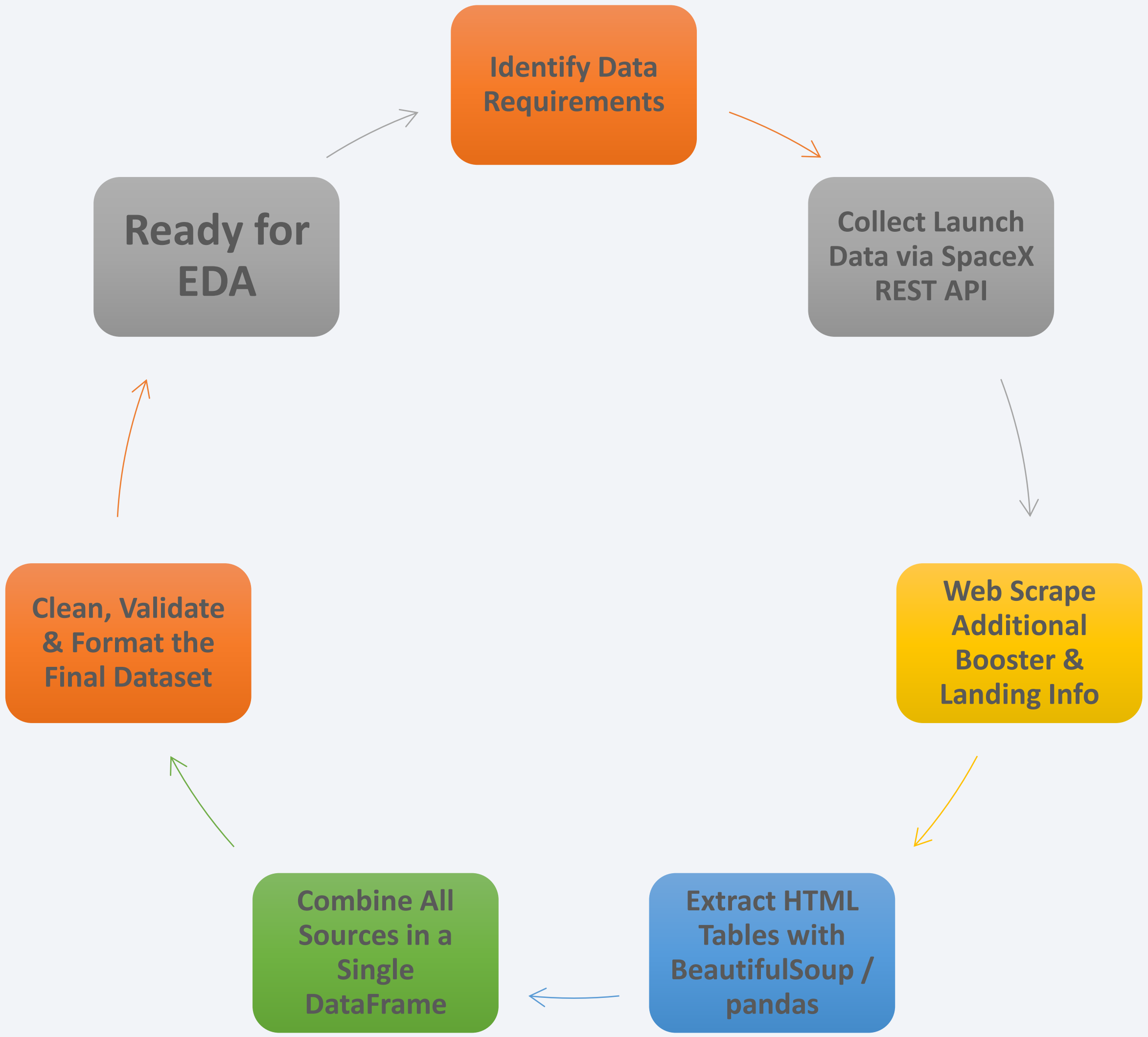
Methodology

Executive Summary

- Data collection methodology:
  - ➢ Data was collected from the SpaceX API and web scraping.
- Perform data wrangling

  This included cleaning, handling missing values, and one-hot encoding categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - ➢ Build models: Split data into train/test sets and train classifiers (Logistic Regression, SVM, Decision Tree, KNN) on the prepared features.
  - ➢ Tune models: Use GridSearchCV with 10-fold cross-validation to find the best hyperparameters for each model.
  - ➢ Evaluate models: Compare test accuracy, review confusion matrices, and select the model with the strongest and most stable performance.

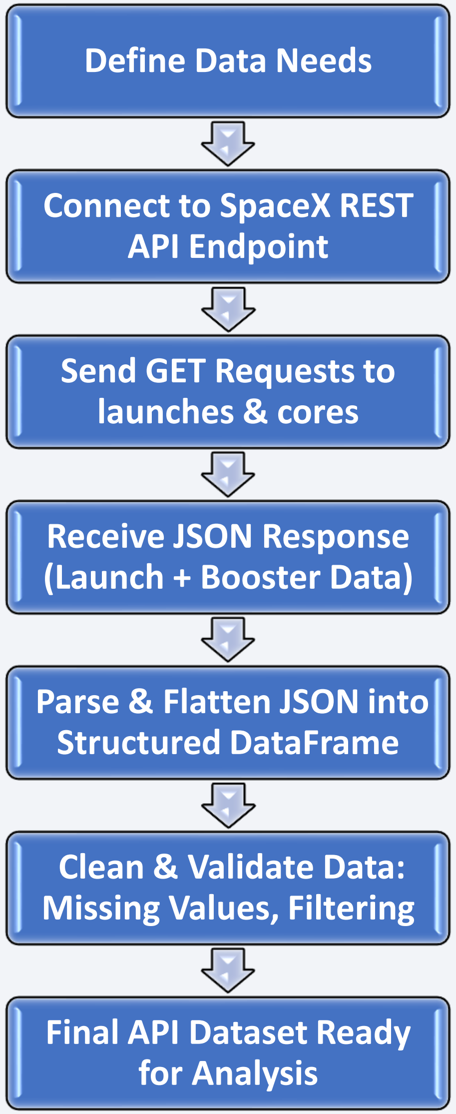
Data Collection



Data Collection – SpaceX API

- The SpaceX REST API was used as the primary source for retrieving structured launch data. The process involved making HTTP GET requests to specific API endpoints, parsing JSON responses, and converting them into a clean tabular format for analysis.

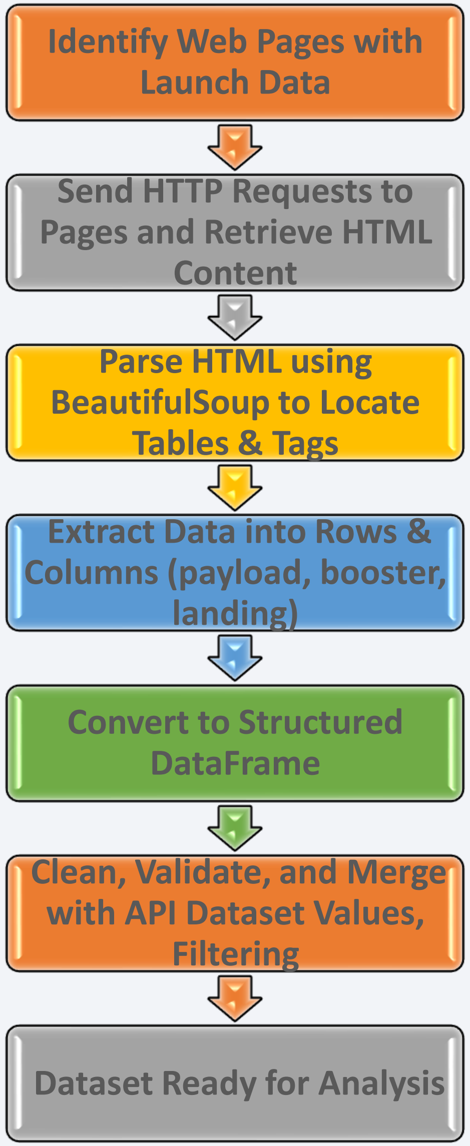
- A reference for this process can be seen in GitHub via the link: https://github.com/CharityO3/DataScienceEcosystem/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- In addition to the SpaceX REST API, web scraping was used to collect supplemental data such as landing outcomes, booster details, and launch notes from public SpaceX web pages. The process ensured that the dataset was complete and accurate.

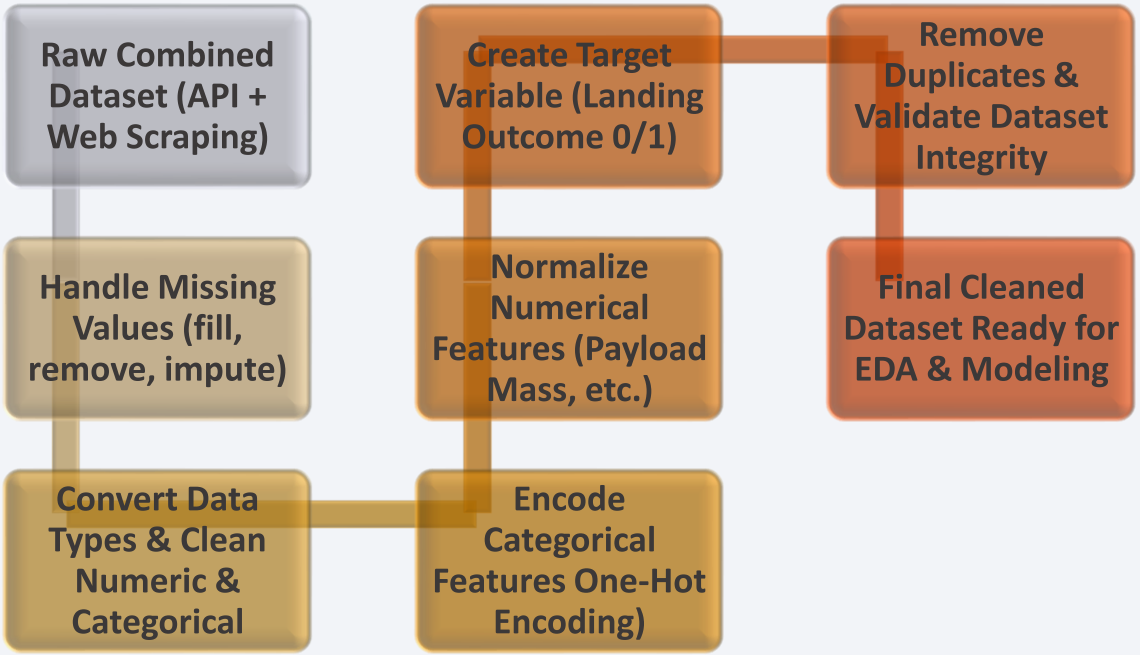
- A reference for this process can be seen in GitHub via the link: https://github.com/Charity03/DataScienceEcosystem/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- After collecting data from the SpaceX REST API and web scraping, the raw dataset underwent extensive cleaning and transformation to prepare it for analysis and modeling.

- A reference for this process can be seen in GitHub via the link: https://github.com/CharityO3/DataScienceEcosystem/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Charts used and its purpose:
    - ➢ Scatter plots – Show relationships between numerical features (e.g., payload vs. success) and detect trends or outliers.
    - ➢ Bar charts – Compare categorical groups (e.g., launch sites, orbit types) on success rates.
    - ➢ Line plots – Visualize trends over time (e.g., landing success improvement by year).
    - ➢ Histograms – Show distributions of numerical variables (e.g., payload mass) to identify skew or outliers.
- Purpose: To explore relationships, compare categories, track trends, and understand distributions to inform modeling.
- A reference for this process can be seen in GitHub via the link: https://github.com/CharityO3/DataScienceEcosystem/blob/main/edadataviz.ipynb

EDA with SQL

- Summary of SQL queries performed:
  - ➢ View all records: SELECT * FROM table
  - ➢ Identify unique values: SELECT DISTINCT column
  - ➢ Count and group by categories: SELECT COUNT(*), GROUP BY column
  - ➢ Aggregate numerical features: AVG, MAX, MIN on payload, etc.
  - ➢ Filter records: WHERE clauses for site, payload range, success, booster reuse
  - ➢ Combine filters and aggregates: e.g., successful landings by site or payload
  - ➢ Sort results: ORDER BY to identify trends over time.

- A reference for this process can be seen in GitHub via the link: https://github.com/CharityO3/DataScienceEcosystem/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Maps objects added:
  - ➢ Base Map – main geographic canvas.
  - ➢ Circles – highlight launch site locations.
  - ➢ Markers with popups/labels – show launch site names.
  - ➢ Marker clusters – group multiple launch events at the same site.
  - ➢ Color-coded markers – indicate landing success (green) or failure (red).
  - ➢ Lines (Polylines) – show spatial relationships or distances to nearby features.

- Purpose: To visualize launch site locations, identify sites and events interactively, show success/failure patterns, and explore spatial relationships.
- A reference for this process can be seen in GitHub via the link: https://github.com/CharityO3/DataScienceEcosystem/blob/main/lab_jupyter_launch_site_location.ipynb

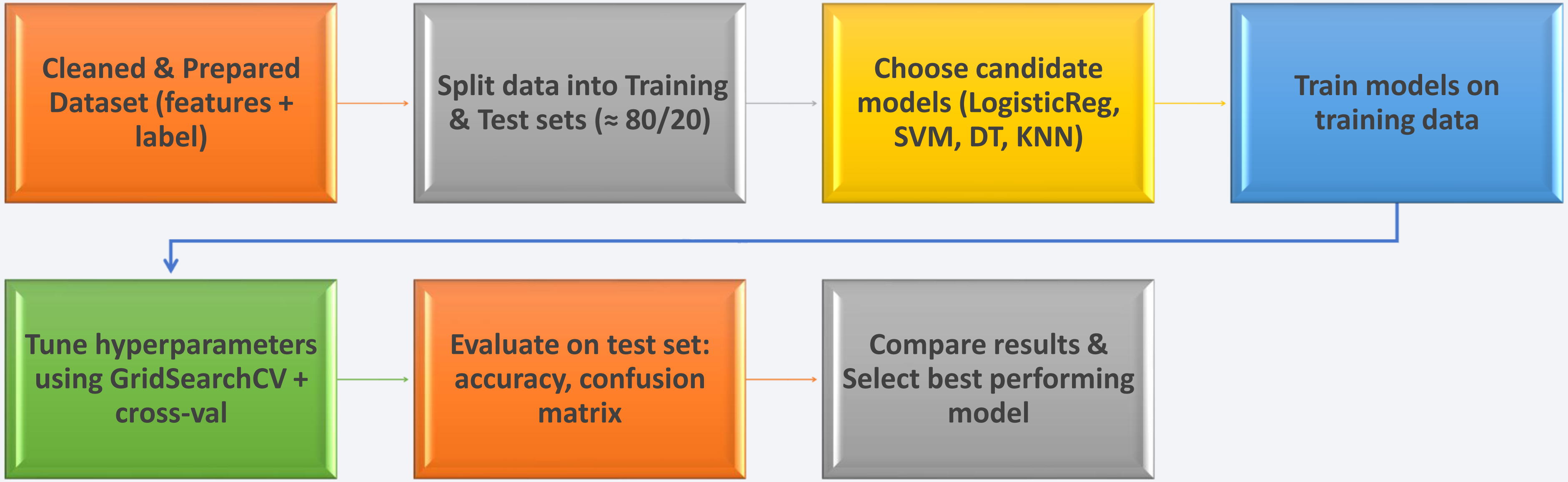
Build a Dashboard with Plotly Dash

- Dashboard Plots & Interactions:
    - ➢ Dropdown menu – filter launches by site.
    - ➢ Payload range slider – filter launches by payload mass.
    - ➢ Pie chart – shows success vs. failure distribution for selected site or all sites.
    - ➢ Scatter plot – payload vs. success, colored by booster version.
    - ➢ Interactive filtering – charts update dynamically when dropdown or slider values change.

- Purpose: To allow users explore launch success patterns by site, payload, and booster interactively.
- A reference for this process can be seen in GitHub via the link: https://github.com/CharityO3/DataScienceEcosystem/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- Steps in Model Development:



- A reference for this process can be seen in GitHub via the link: https://github.com/Charity03/DataScienceEcosystem/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

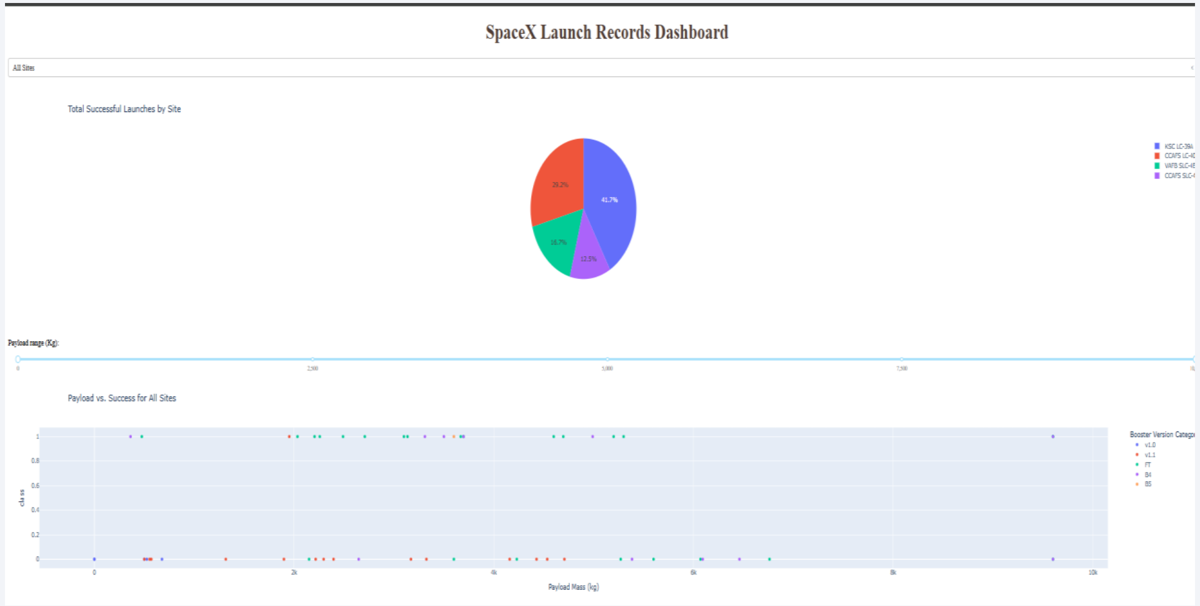
Results

Exploratory data analysis results:

- ➢ Launch sites: Some sites have higher launch counts; success rates vary by site.
- ➢ Payload mass: Most launches 2,000–8,000 kg; heavier payloads slightly reduce success.
- ➢ Orbit type: LEO launches have higher success; GTO and SSO show more failures.
- ➢ Booster reuse: Reused boosters perform similarly or slightly better; drone ship landings are harder.
- ➢ Flight experience: Success improves with more previous flights.
- ➢ Visualizations: Scatter plots, bar charts, histograms, pie charts, and Folium maps highlight patterns.

- Insight: Success depends on payload, orbit, launch site, booster reuse, and flight experience

Interactive analytics demo in screenshots



Results Contd

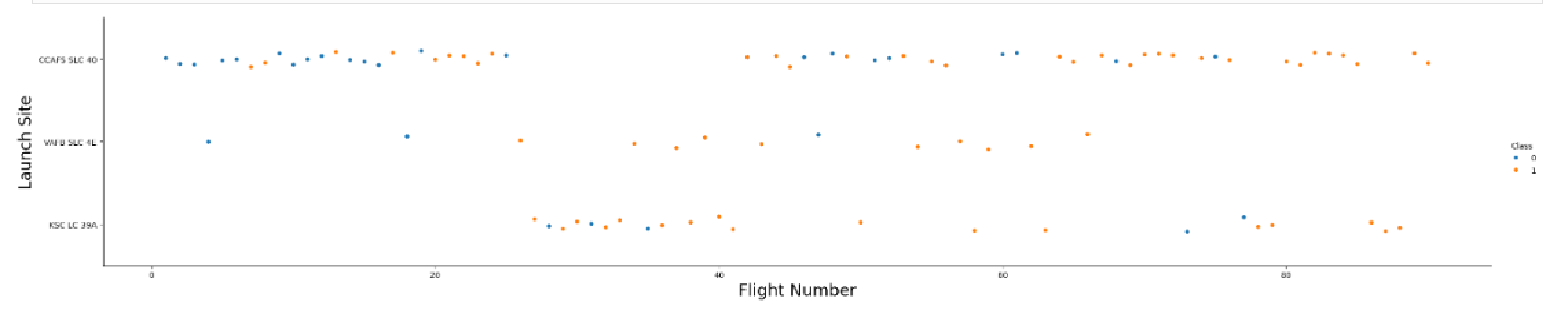
- Predictive analysis results:
- Models evaluated: Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN).
- Performance:
  - ➢ Logistic Regression: highest accuracy (~83%)
  - ➢ SVM and KNN: same accuracy as LogReg (~83%)
  - ➢ Decision Tree: lowest (~76%)
- Logistic Regression, KNN & SVM were the best-performing model (~83% accuracy), showing that landing success can be reliably predicted using payload, launch site, orbit, booster reuse, and flight number.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

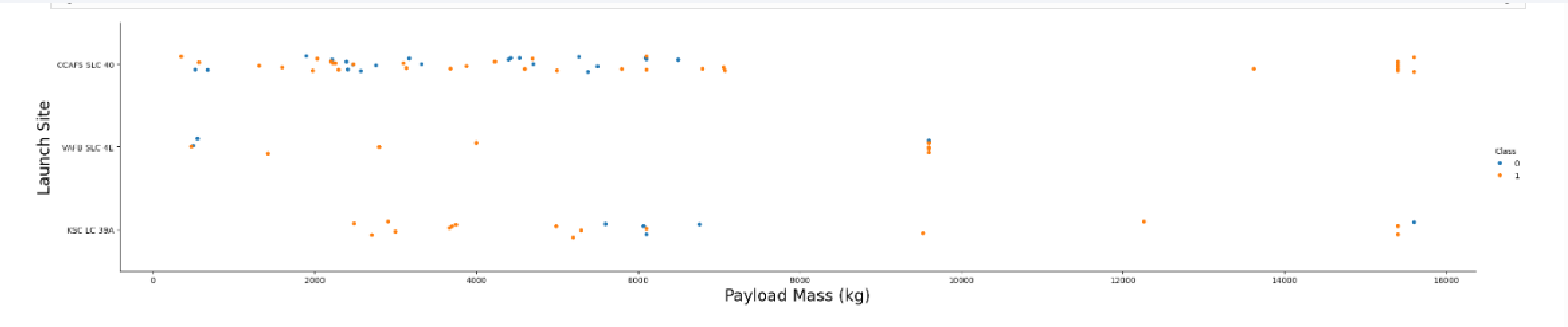
- Scatter plot of Flight Number vs. Launch Site



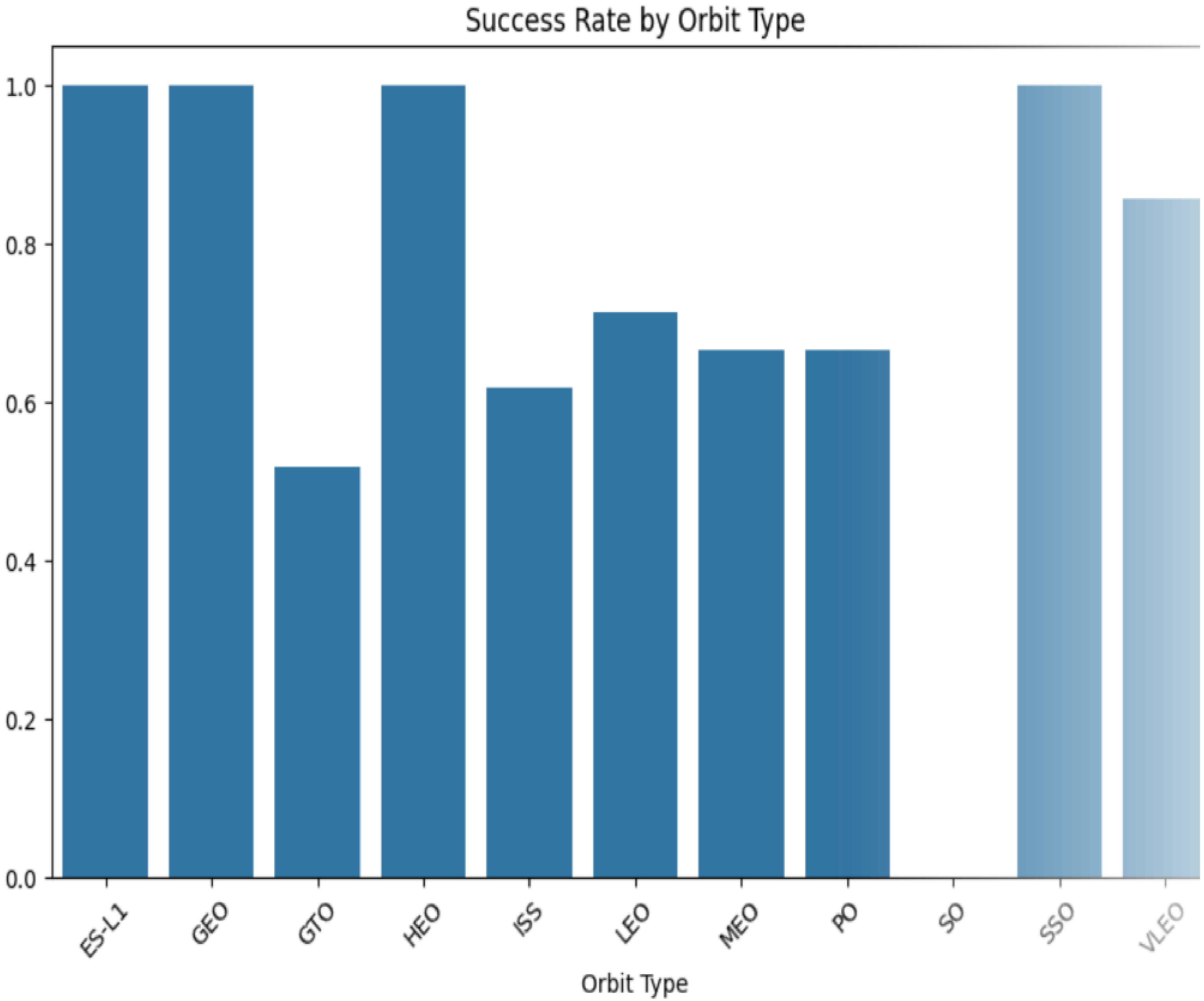
- Key findings:
  - ➢ Launch frequency: CCAFS SLC-40 has the most launches, followed by KSC LC-39A and VAFB SLC-4E.
  - ➢ Temporal trends: Later flights (higher flight numbers) tend to show more successful landings, indicating improvement over time.
  - ➢ Site performance: While all sites show both successes and failures, patterns suggest some sites (e.g., KSC LC-39A) have slightly higher early success rates than others.
  - ➢ Data distribution: Launches are unevenly distributed across sites, which is important to consider when modeling landing success.
- Insight: Success improves over time due to experience and operational improvements, and launch site heavily influences the number of launches and early landing outcomes.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



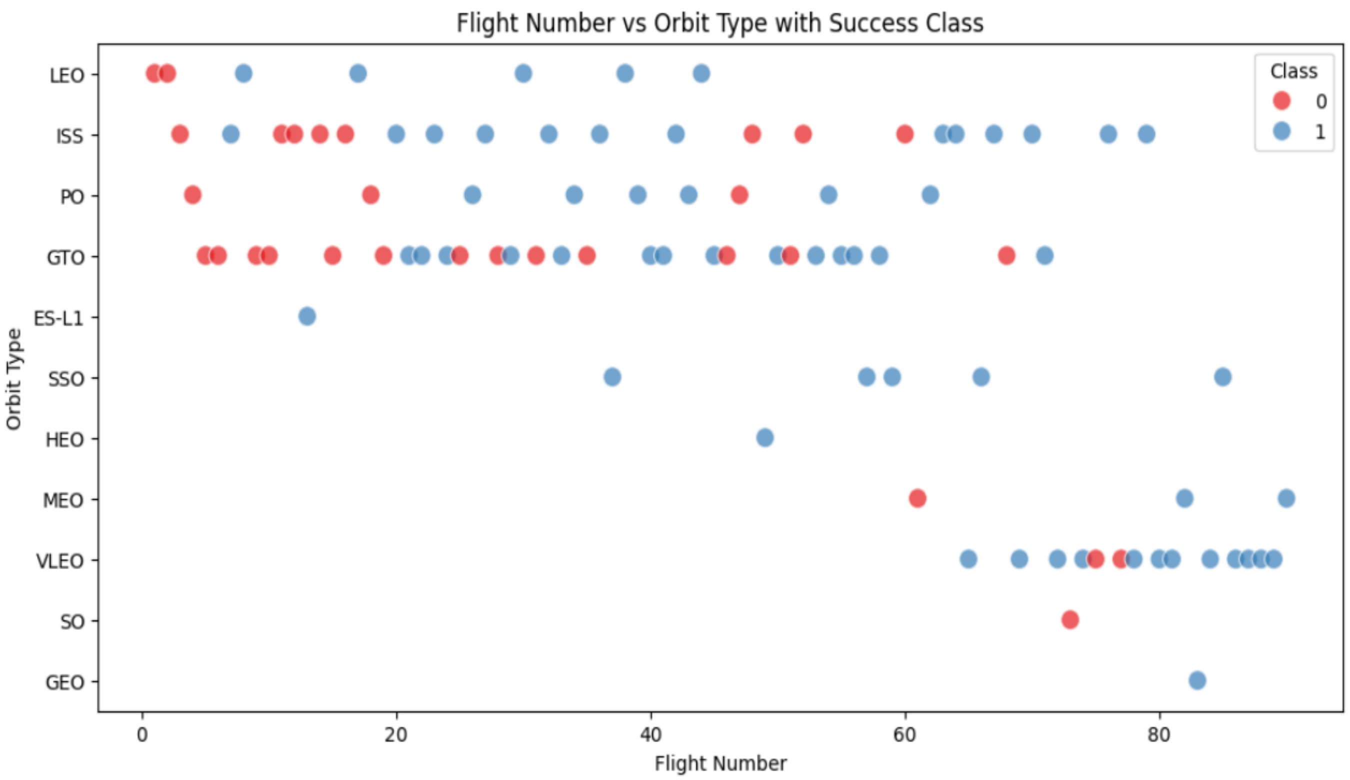
- Key findings:
  - ➢ Payload distribution: CCAFS SLC-40 and KSC LC-39A handle a wider range of payloads, while VAFB SLC-4E generally handles lighter payloads.
  - ➢ Success correlation: Heavier payloads tend to have slightly lower landing success, while lighter payloads show higher success rates.
  - ➢ Site performance: Some sites (e.g., KSC LC-39A) show more consistent success across payload ranges.
  - ➢ Data distribution: Payloads are unevenly distributed across sites, which affects modeling of landing success.
- Insight: Landing success is influenced by payload mass, with lighter payloads achieving higher success, and launch site selection affects payload handling and outcomes.



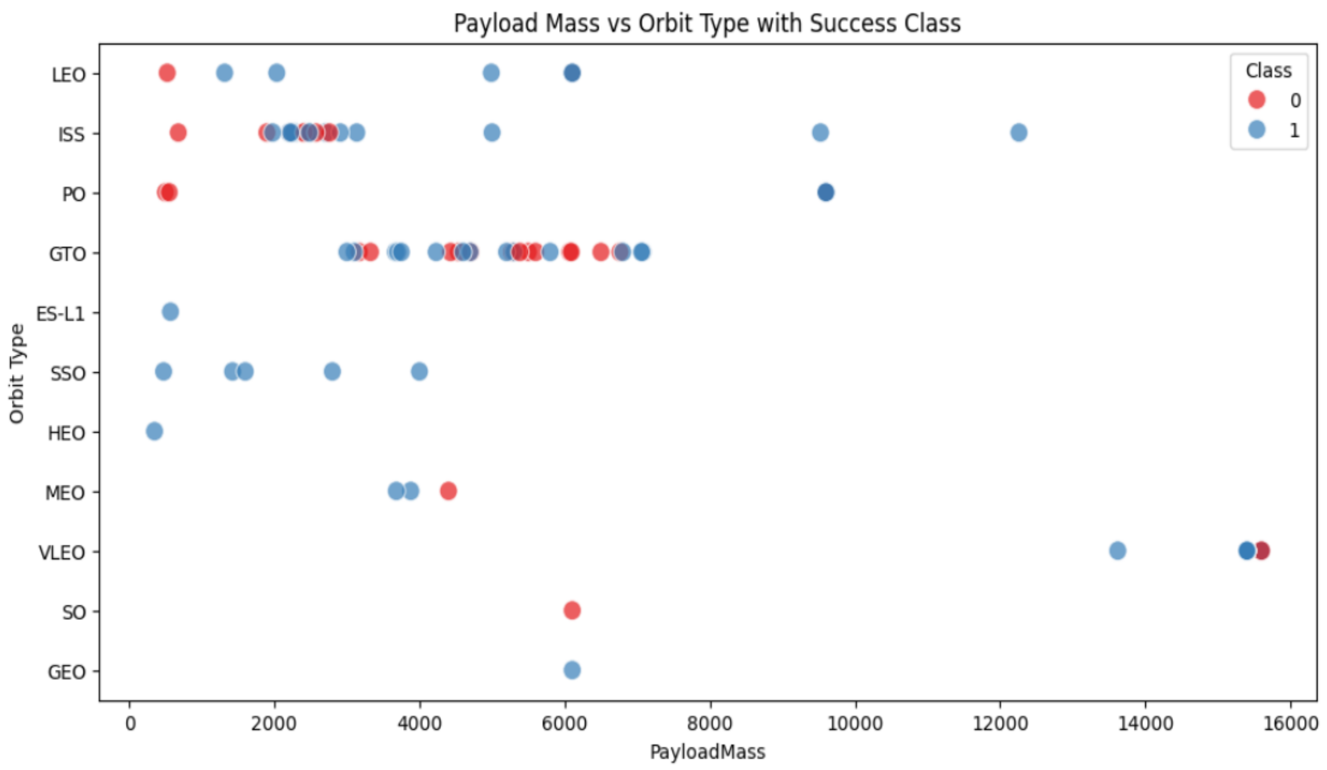
Success Rate VS. Orbit Type

- Key findings:
  - Highest success rates: ES-L1, GEO, HEO, SSO often show 100 % success in the dataset.
  - Moderate/Lower success rates: LEO, GTO, ISS, VLEO exhibit more variability and lower overall success.
  - Orbit performance matters: The recovery success of first-stage boosters depends strongly on the target orbit — missions to some orbits are more favorable for landing success.
  - Uneven data distribution: The number of launches to high-success orbits (e.g. GEO, ES-L1) is often small, which may bias the success rate calculation; common orbits dominate the dataset.
- Insight: Orbit choice appears to significantly affect landing success — rarer, high-energy orbits aren't necessarily worse: in some analyses, orbits like ES-L1, GEO, HEO and SSO show perfect or near-perfect success rates, suggesting that mission parameters and recovery strategy (rather than just orbit energy) strongly influence landing outcomes.

Flight Number VS. Orbit Type

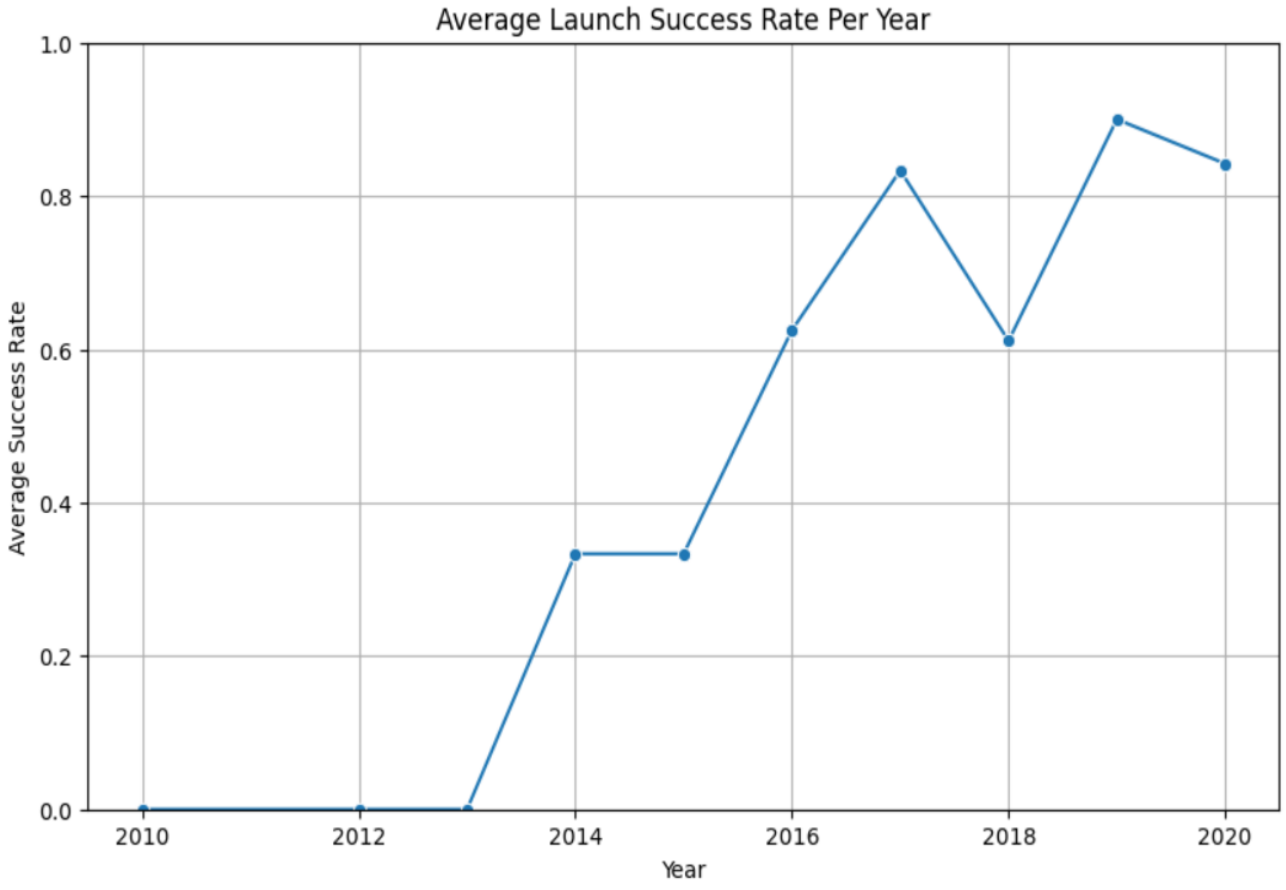


- Key findings:
  - Over time (higher flight numbers), the number of orbit types used increases — newer orbits (e.g. GEO, HEO, others) appear only in later flights.
  - For certain orbits (like LEO), there is a visible trend: later flights (higher flight number) tend to have better success rates — suggesting improvements over time (technology, operations, experience).
  - For some orbits (like GTO), there seems to be no clear improvement trend over time — failures and successes are more scattered across flight numbers, indicating orbit-specific challenges.
  - Many of the less frequent orbits only appear in later flights — this reflects expansion in mission types over time rather than early adoption.
- Insight: The plot reveals that as the number of launches increases, mission complexity and orbit variety increases. Improvements over time (technical and operational) have increased success — particularly for common orbits like LEO — but some orbits remain inherently more challenging (e.g. GTO), with more variable success outcomes.



Payload VS. Orbit Type

- Key findings:
  - Payload distribution: Most orbits (e.g. LEO, ISS, Polar) cover a wide range of payload masses; some less common orbits have fewer data points or narrower payload ranges.
  - Success correlation with orbit and payload: Heavier payloads tend to show successful landings particularly for orbits such as LEO, ISS, and Polar.
  - Orbit-specific differences: For some orbit types (e.g. GTO), both successful and failed landings occur across payload ranges — so no clear payload-success pattern for those orbits.
  - Data unevenness: Some orbits have few launches, which means payload vs. orbit observations — especially for rare orbits — may not reliably reflect general trends.
- Insight: Payload mass interacts with orbit type in affecting landing outcomes — with heavier payloads often landing successfully on more “routine” orbits (LEO, ISS, Polar), but for more complex orbits (e.g. GTO), payload mass alone does not guarantee success; orbit-specific conditions and mission complexity matter.



Launch Success Yearly Trend

- Key findings:
  - ➢ Increasing reliability over time: Success rates generally climb year after year, indicating improving tech, processes, and experience.
  - ➢ High launch volume with high success: Years with many launches still maintain high (or improving) success rates which suggests scalability of operations without sacrificing reliability.
  - ➢ Occasional dips but overall upward trend: Some years may have slight drops (due to failures or mission complexity), but the long-term trend favors improvement.
- Insight: Payload mass interacts with orbit type in affecting landing outcomes — with heavier payloads often landing successfully on more “routine” orbits (LEO, ISS, Polar), but for more complex orbits (e.g. GTO), payload mass alone does not guarantee success; orbit-specific conditions and mission complexity matter.

All Launch Site Names

- **The SpaceX dataset contains launches from multiple sites. This query helps identify all the distinct launch locations used for Falcon 9 missions, which is useful for analyzing site-specific success rates and trends.**

```
12]:  %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;

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

12]:

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

Launch Site Names Begin with 'CCA'

- These records show a subset of launches from Cape Canaveral, allowing quick inspection of payload, orbit, and success outcomes. Limiting to 5 records provides a manageable sample for analysis or visualization.

```
%sql SELECT * FROM SPACEXTBL 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-	15:10:00	F9 v1.0 B0007	CCAFS LC-	SpaceX	677	LEO	NASA	Success	No attempt

Total Payload Mass

The query shows that SpaceX boosters have carried a total of 45,596 kg of payload for NASA CRS missions. This helps quantify the contribution of SpaceX launches to NASA resupply missions and can be used for further analysis of payload trends by customer.

```
]:  %%sql
    SELECT SUM("Payload_Mass__kg_") AS total_payload_mass
    FROM SPACEXTBL
    WHERE "Customer" = 'NASA (CRS)';
```

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

]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

The F9 v1.1 booster has carried an average payload of 2928.4 kg per launch. This provides insight into the payload capacity of this booster version compared to others, which is useful for performance analysis and planning future missions.

```
%%sql
SELECT AVG("Payload_Mass__kg_") AS AVG_payload_mass
FROM SPACEXTBL
WHERE "Booster_Version" = 'F9 v1.1';
```

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

AVG_payload_mass
2928.4

First Successful Ground Landing Date

The first successful ground pad landing occurred on December 22, 2015. This marks a milestone in SpaceX's reusable booster program and can be used as a reference point for analyzing trends in landing success over time.

```
%%sql
SELECT MIN("Date") AS first_successful_landing_date
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

These boosters managed successful drone ship landings within the specified payload range OF 4000 and 6000. This information is useful for analyzing booster performance under moderate payload conditions and can inform operational decisions for similar future missions.

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTBL
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

Total Number of Successful and Failure Mission Outcomes

Out of all recorded SpaceX launches in the dataset, 100 missions were successful and 1 failed. This gives a high-level overview of overall mission reliability and can be used to analyze trends in mission performance over time.

Mission_Outcome	total
Failure	1
Success	100

Boosters Carried Maximum Payload

- **The following booster_Versions:** 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, carried the maximum payload in the dataset.
- This highlights which booster version handled the heaviest mission, useful for analyzing booster capabilities and mission planning.

2015 Launch Records

In 2015, there were two failed drone ship landing in January and April using booster F9 v1.1 from CCAFS SLC-40. This shows the early challenges of drone ship landings and can be used to analyze trends in landing success improvement over subsequent years.

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

Landing_Outcome	total
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

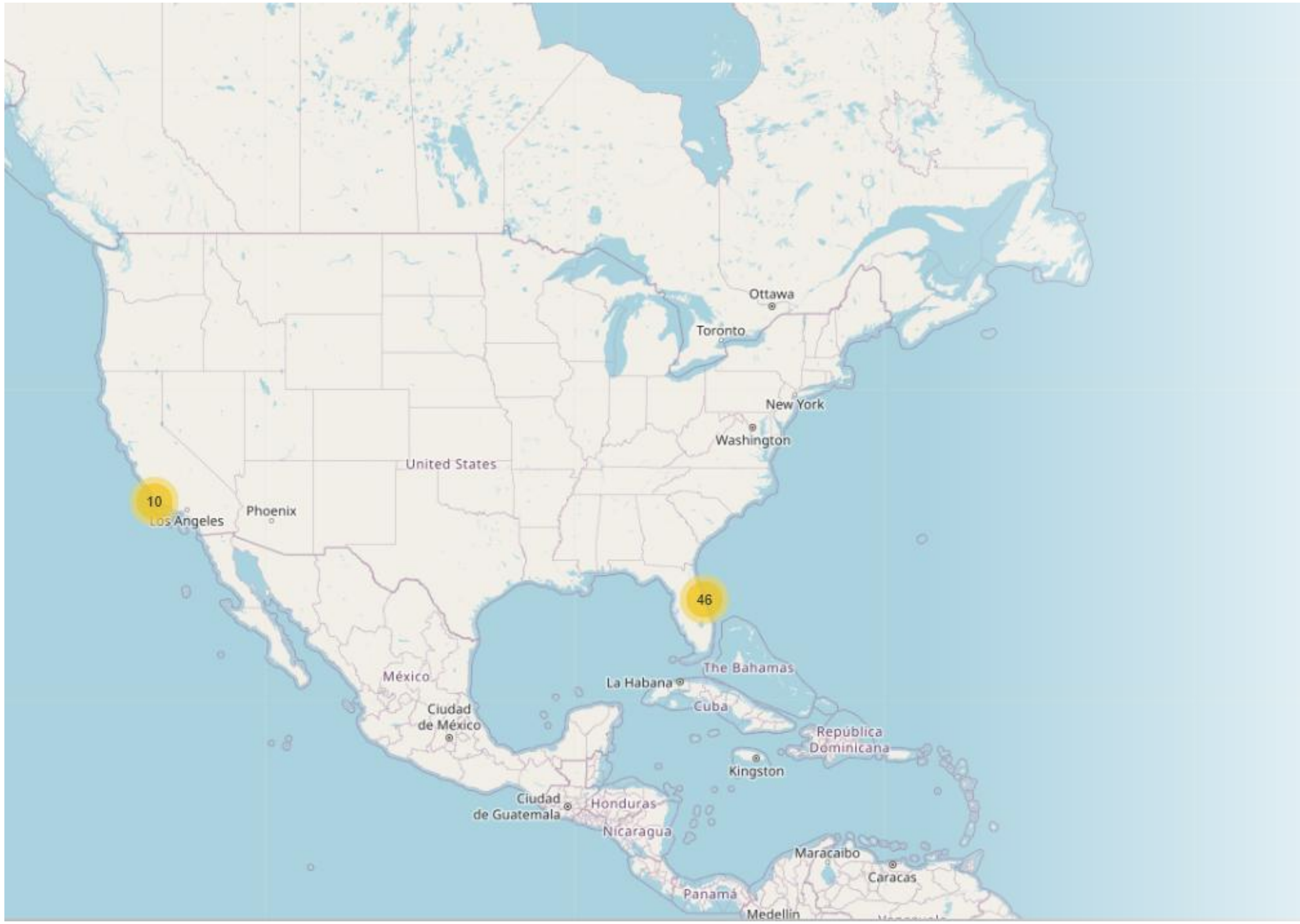
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Most common: No attempt is highest, showing many early missions didn't attempt booster recovery.
- Drone ship: Equal numbers of successes and failures, reflecting testing and refinement.
- Ground pad: Fewer attempts but generally reliable when used.
- Other outcomes: Ocean landings and parachute failures represent early recovery methods.
- Insight: The data shows SpaceX's shift from no recovery attempts to improving success in both drone ship and ground pad landings over time.

Section 3

Launch Sites Proximities Analysis

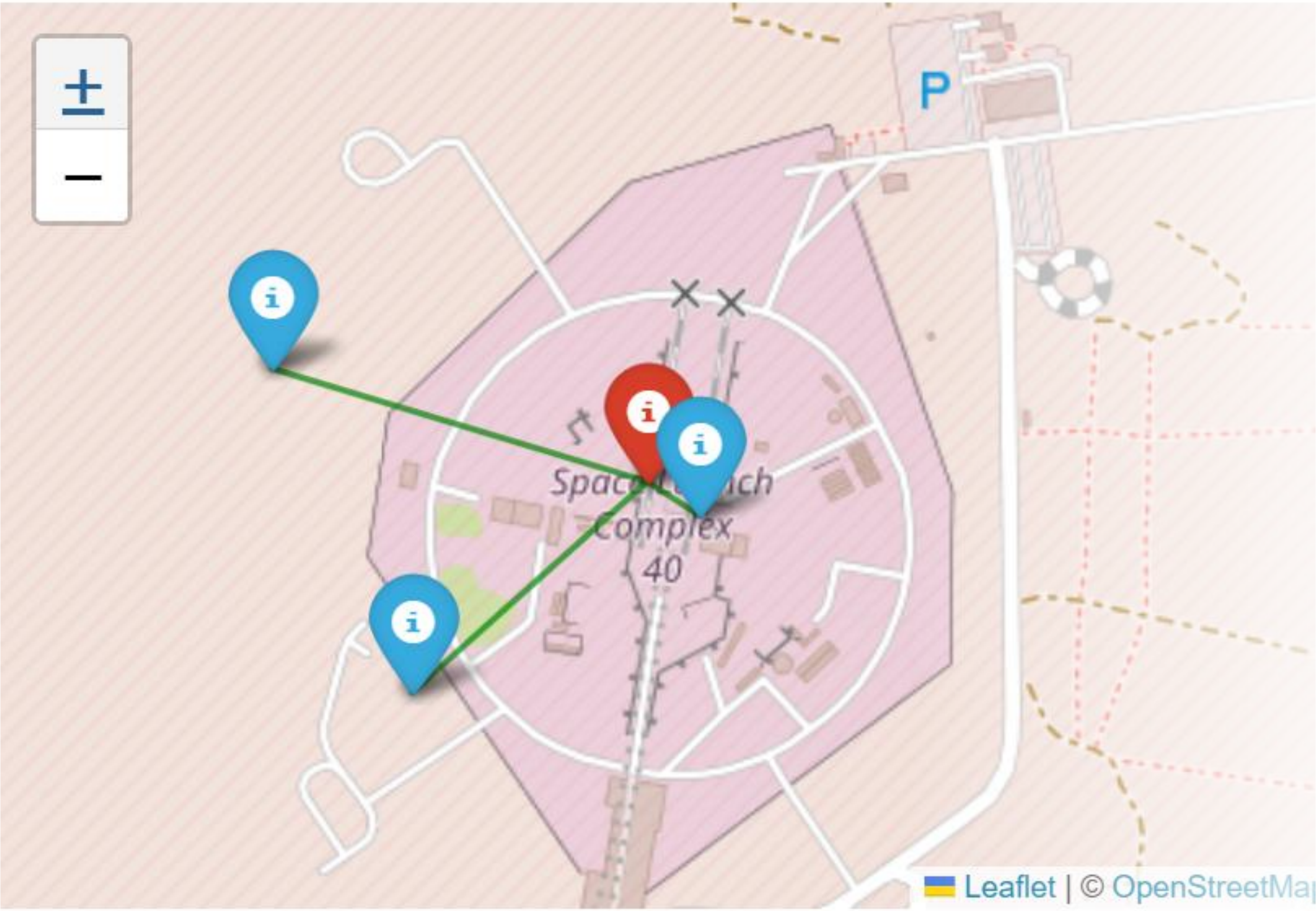
Global Map of SpaceX Launch Site Locations



- The map shows all SpaceX launch site markers, mainly concentrated in the United States.
- Florida has the highest cluster of launches (Cape Canaveral & Kennedy Space Center), shown by the larger marker.
- Vandenberg (California) appears as a single marker on the West Coast, used mostly for polar-orbit missions.
- Boca Chica/Starbase (Texas) is also marked, showing SpaceX's south-central launch activity.
- All sites are located near coastlines, which supports safe launch paths and booster recovery.

SpaceX Launch Sites and Landing Outcomes Map

- **Key Findings**
  - ➢ Florida sites show the highest number of launches, especially CCAFS SLC-40 and KSC LC-39A, indicated by the cluster of markers.
  - ➢ Successful landings are most common, with many markers in green (or the designated success color), reflecting improved reliability of SpaceX rockets.
  - ➢ Drone-ship failures and ocean landings appear less frequently but are visible through distinct colors, showing early-phase recovery challenges.
  - ➢ Vandenberg (California) has fewer launches and mixed outcomes, consistent with its more specialized missions.



Launch Site Proximity to Key Infrastructure and Coastline

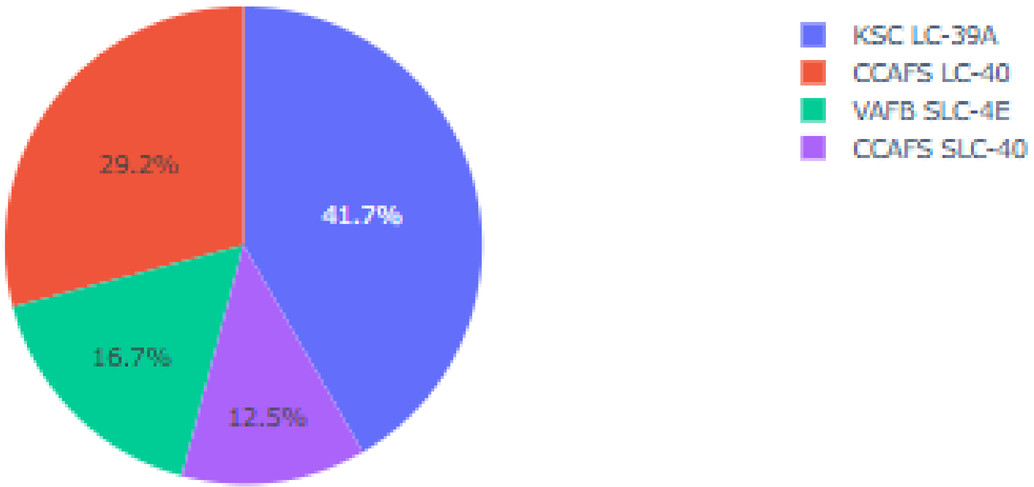
- The selected launch site is located near key infrastructure and natural features. The coastline is approximately 0.5 km away, the highway about 1 km, and the railway around 2 km from the site. These proximities indicate good accessibility for logistics while maintaining safety and operational efficiency, with visual connections clearly shown on the map.

Section 4

Build a Dashboard with Plotly Dash

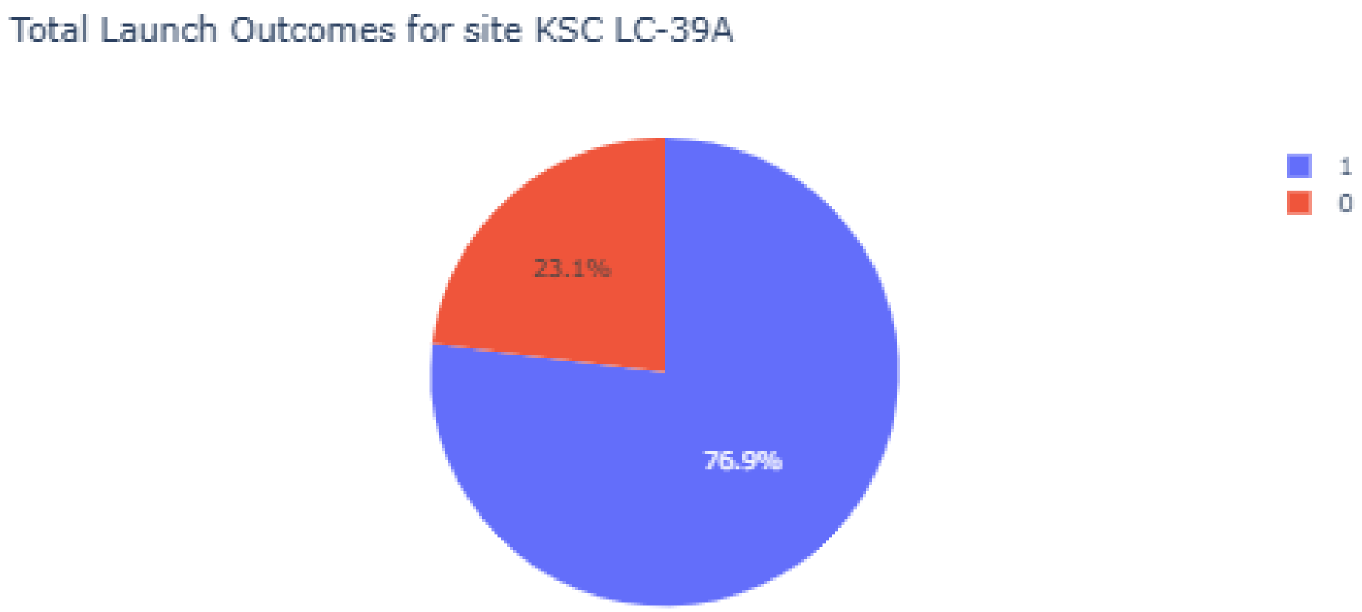
SpaceX Success Launch Dashboard

Total Successful Launches by Site



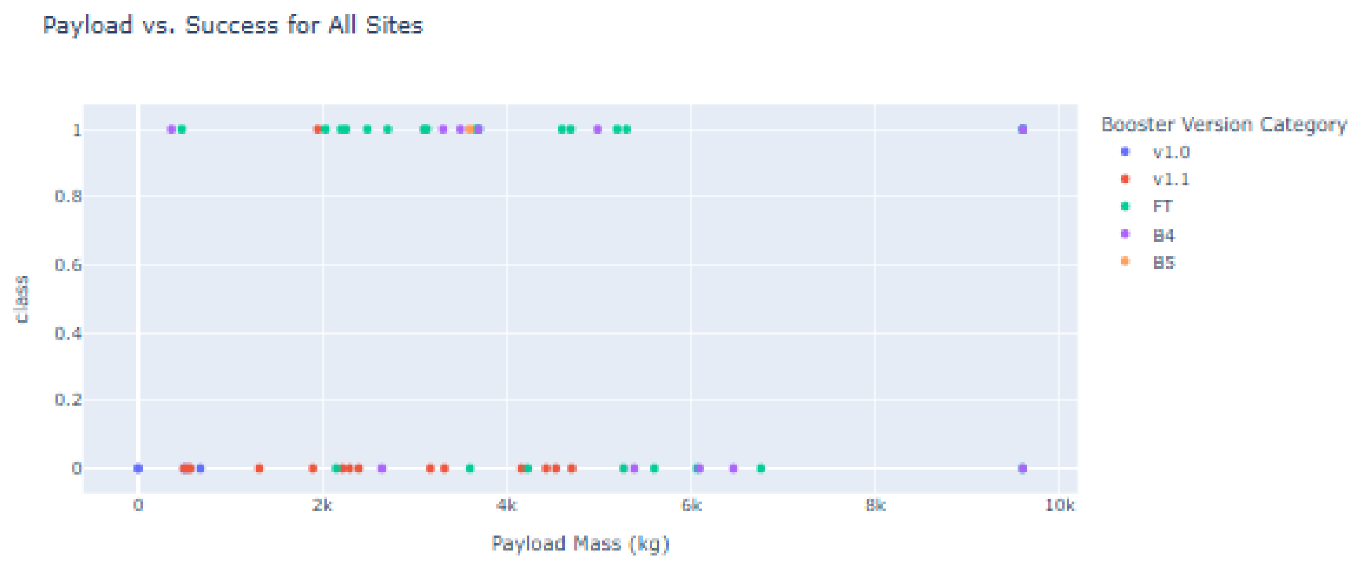
- Key Observations
  - Launch Site Success: Some launch sites contribute more to overall successful launches than others.
  - Site-Specific Reliability: Selecting a single site shows its success vs failure ratio, highlighting reliability.
  - Payload Impact: Heavier or lighter payloads may affect success rates, visible in the scatter plot.
  - Booster Performance: Certain booster versions perform better, especially for specific payload ranges.

KSC C-39A -Launch Site With Highest Success Ratio



- The successful launch slice dominates, confirming that this site (KSC LC-39A) has the highest launch success ratio among all SpaceX sites.
- The small failure slice shows that only a minor fraction of launches at this site failed.

Payload vs Launch Outcome for All Sites by Booster Version

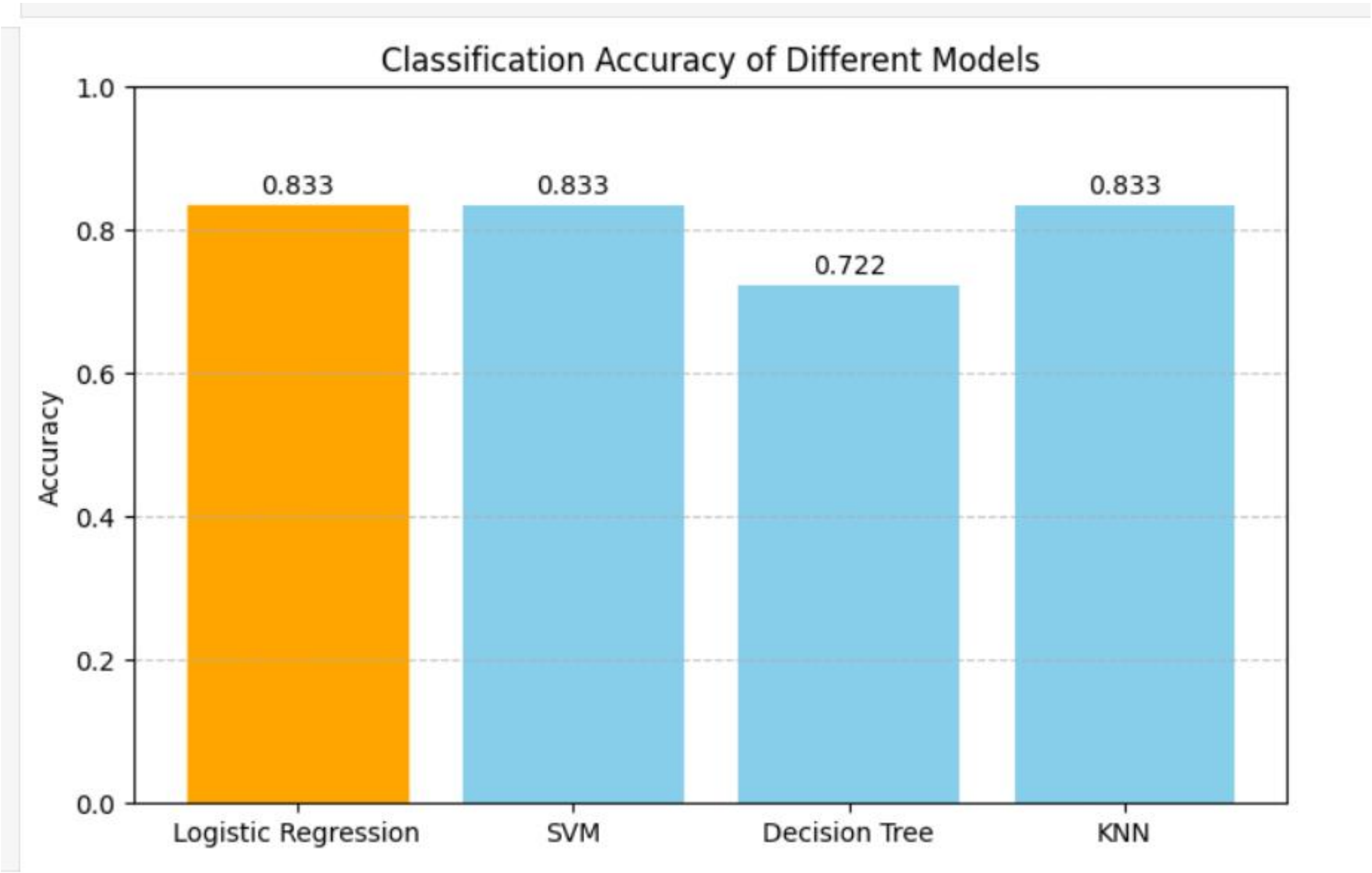


- Key Observations:
  - ➢ High Success Concentration:
  - ➢ Certain payload ranges (for example, 2000–4000 kg) show a high density of successful launches.
  - ➢ Very low or very high payloads may have slightly higher failure occurrences.
  - ➢ Booster Version Performance:
    - ➢ Some boosters (e.g., Falcon 9 Block 5) dominate the successful launches, indicating higher reliability.
    - ➢ Other boosters may appear more frequently in failed launches, suggesting variation in performance.

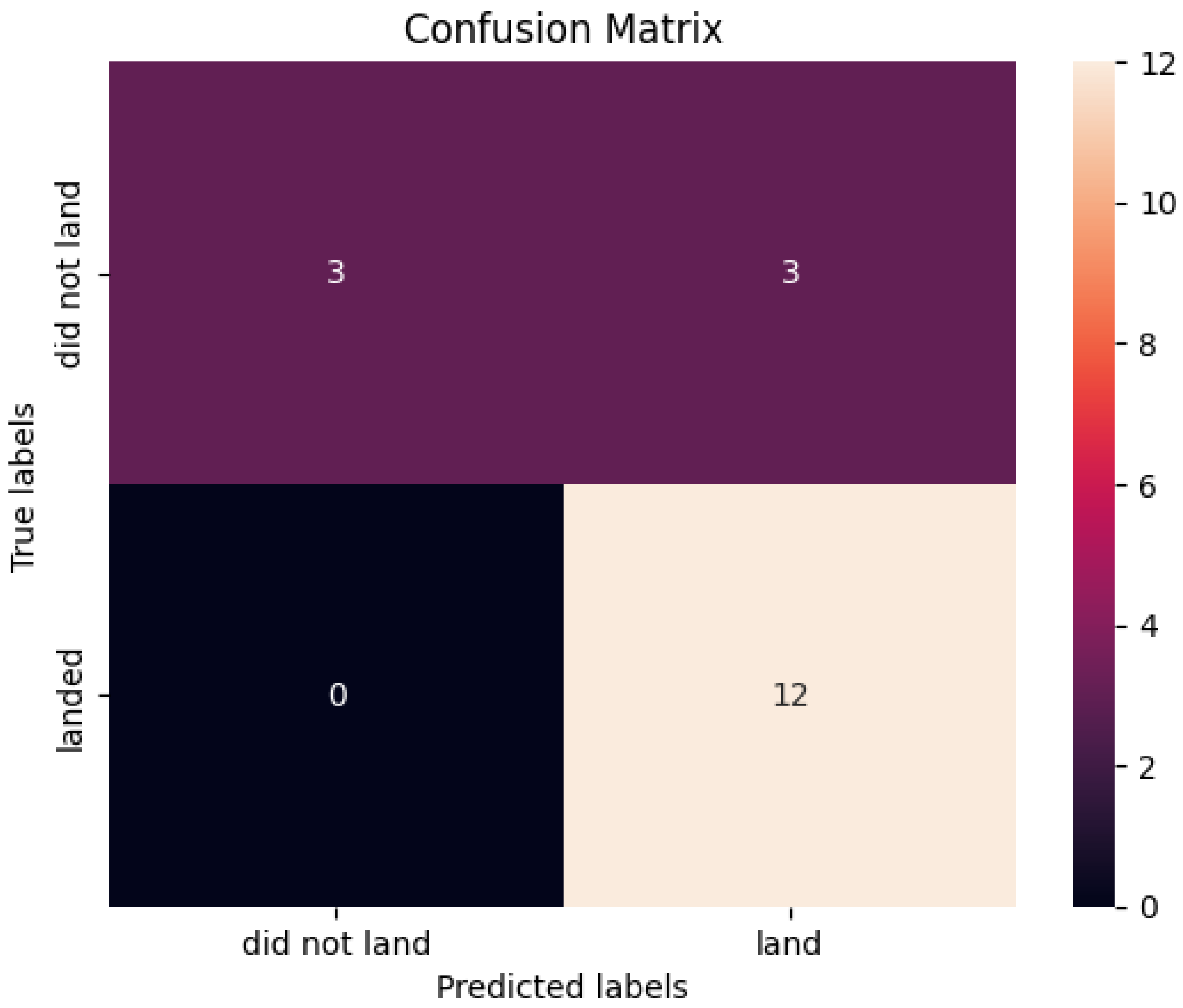
Section 5

Predictive Analysis (Classification)

Classification Accuracy of All Models



- Logistic Regression, KNN & SVM were the best-performing model (~83% accuracy), showing that landing success can be reliably predicted using payload, launch site, orbit, booster reuse, and flight number.



Confusion Matrix For Logistic Regression

- The confusion matrix confirms that Logistic Regression makes the fewest misclassifications, correctly predicting most successes and failures.
- This consistency, combined with its high test accuracy, makes it the best performing model for predicting SpaceX launch outcomes in your dataset.

Conclusions

- In conclusion, the models used to perform predictive analysis can reliably predict Falcon 9 landing success, with Logistic Regression performing best overall.
- Key factors such as payload mass, orbit type, launch site, and booster reuse significantly influence landing outcomes.
- The analysis and visualizations provide actionable insights that help understand SpaceX's mission reliability and reusability strategy.
- The interactive dashboard offers an accessible decision-support tool for exploring launch scenarios and evaluating success probabilities.

Appendix

- To read data:

URL2 = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_3.csv'

resp2 = await fetch(URL2)

text2 = io.BytesIO((await resp2.arrayBuffer()).to_py())

X = pd.read_csv(text2)

- To drop table in SQL:

%sql DROP TABLE IF EXISTS SPACEXTABLE;

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

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

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

Thank you!