Winning Space Race with Data Science

Furkan Önal
19.03.2025

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Executive Summary

Methodology

- Collected and processed SpaceX launch data using API integration and web scraping.
- Conducted Exploratory Data Analysis (EDA) to identify key success factors.
- Developed machine learning models (Logistic Regression, SVM, Decision Tree, KNN) for landing success prediction.
- Built an interactive dashboard using Plotly Dash & Folium for real-time data analysis.

Results

- Launch success rates improved over time, with LEO, ISS, and Polar orbits showing the highest success.
- Payloads >10,000 kg significantly increased success rates at certain launch sites.
- KSC LC-39A recorded the highest launch success rate (~77%).
- Logistic Regression was the most effective model for predicting landing outcomes.

Introduction

Background & Context

The commercialization of space travel has driven significant advancements in reusable rocket technology. SpaceX's Falcon 9 is a leader in this field, offering cost-efficient launches by successfully landing and reusing its first stage. Understanding the factors influencing landing success is critical for cost reduction and competitive advantage in the space industry.

Key Problems to Address

- What factors contribute to a successful Falcon 9 first-stage landing?
- How do payload mass, launch site, and orbit type impact landing success?
- Can machine learning accurately predict landing success?
- How can an interactive dashboard provide actionable insights for future launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used SpaceX REST API to collect data
  - Used web-scrapping to scraped data from Wikipedia (BeautifulSoup)
- Performed data wrangling
  - Standardization of the data and their labels
  - Dealing with missing values
  - Transformation of the data with one-hot-encoding to prepare data for prediction
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Building, tuning, and evaluating classification models involves selecting an appropriate algorithm, optimizing its hyperparameters to improve performance, and assessing its effectiveness

Data Collection

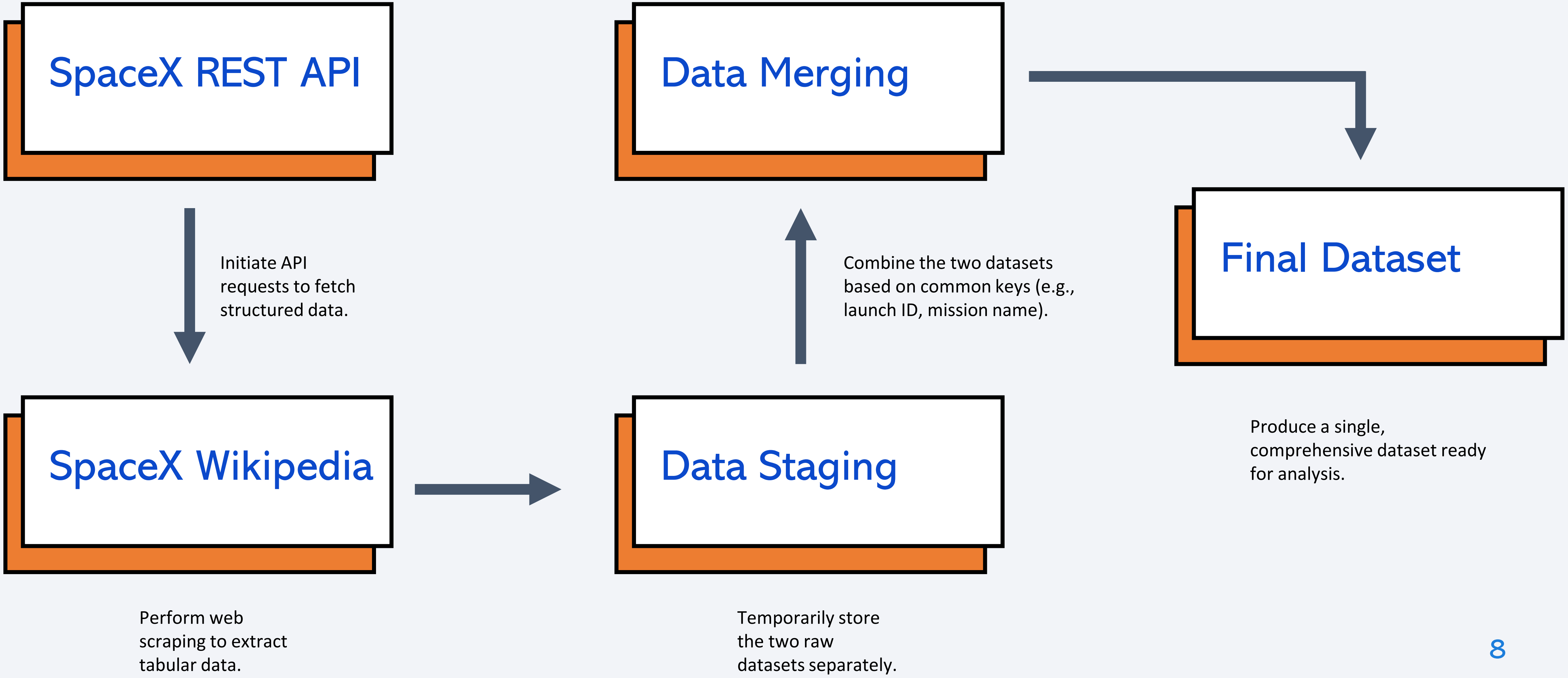
The data collection process was a multi-faceted approach designed to ensure a complete and accurate dataset for analysis. It involved two primary methods:

1. **API Requests to the SpaceX REST API:** The official SpaceX API was a key source of data. This method allowed for the retrieval of structured, real-time information about launches, rockets, capsules, and other relevant details. API requests are generally reliable and provide data in a clean, machine-readable format like JSON.

2. **Web Scraping from a Wikipedia Entry:** To supplement the API data, web scraping was performed on a specific Wikipedia table. This was done to capture information that may not have been available or as detailed in the API, or to fill in historical gaps. Web scraping involves using a script to extract data directly from the HTML of a web page, which in this case was the "List of Falcon 9 and Falcon Heavy launches" page.

By combining these two distinct data collection methods, the project was able to create a comprehensive dataset, leveraging the strengths of each approach to overcome the limitations of a single source.

Data Collection – Overview



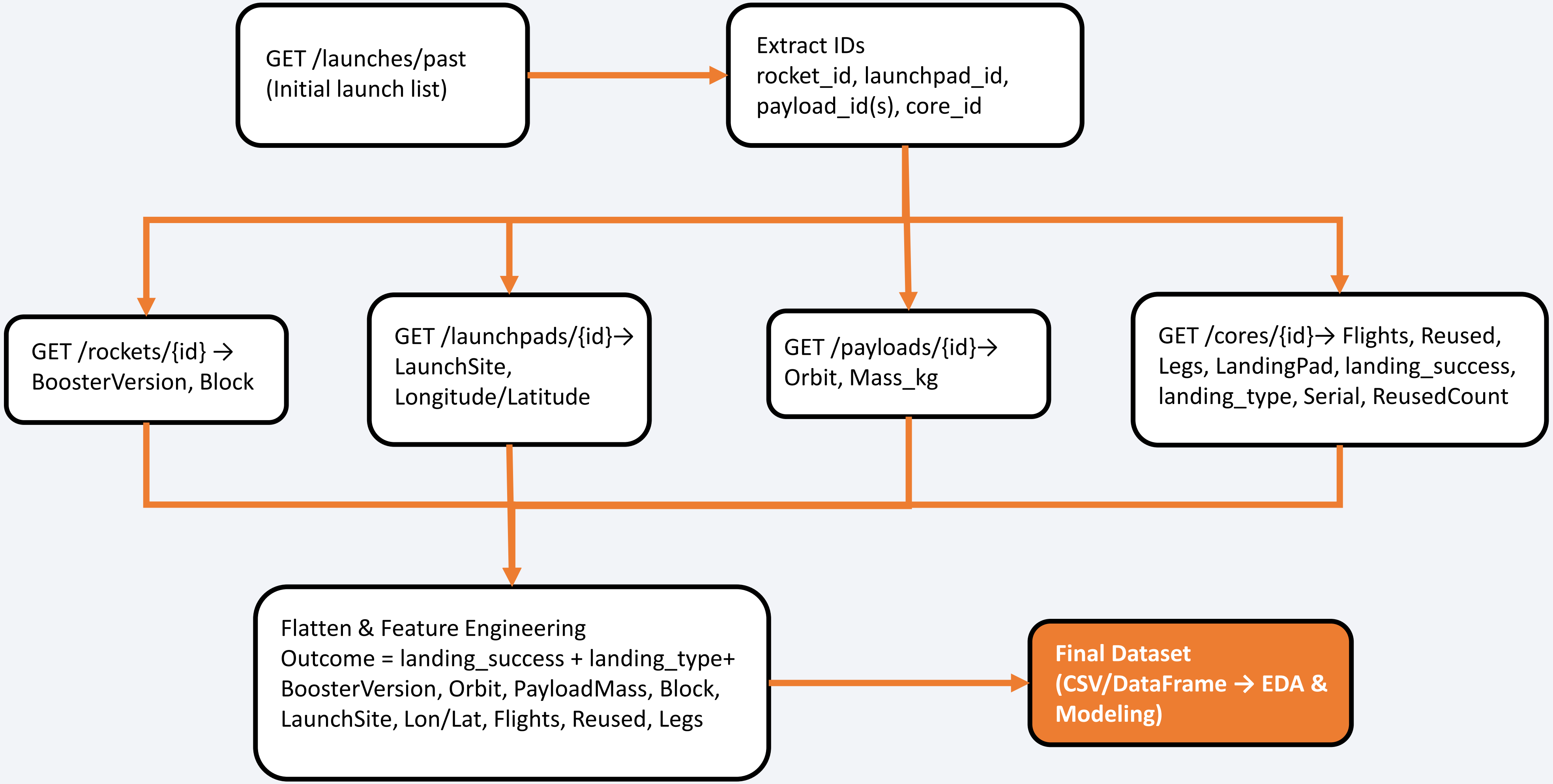
Data Collection – SpaceX API

[Click Here to Access the Notebook]

- The SpaceX API v4 was queried using `GET/launches/past` to retrieve historical launches.
- From each launch, IDs were dereferenced to enrich fields via secondary calls to `/rockets, /launchpads, /payloads`, and `/cores`.
- Nested JSON was flattened and analysis-ready columns were engineered: BoosterVersion, PayloadMass (kg), Orbit, LaunchSite, Longitude/Latitude, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, and Outcome (a combination of `landing_success` and `landing_type`).
- A raw JSON snapshot (with a static fallback URL in the notebook) was retained and a cleaned table was prepared to ensure reproducibility.
- Through this pipeline, API objects were transformed into a single, model-ready dataset focused on landing outcomes.

Data Collection – SpaceX API

Click Here to Access the Notebook



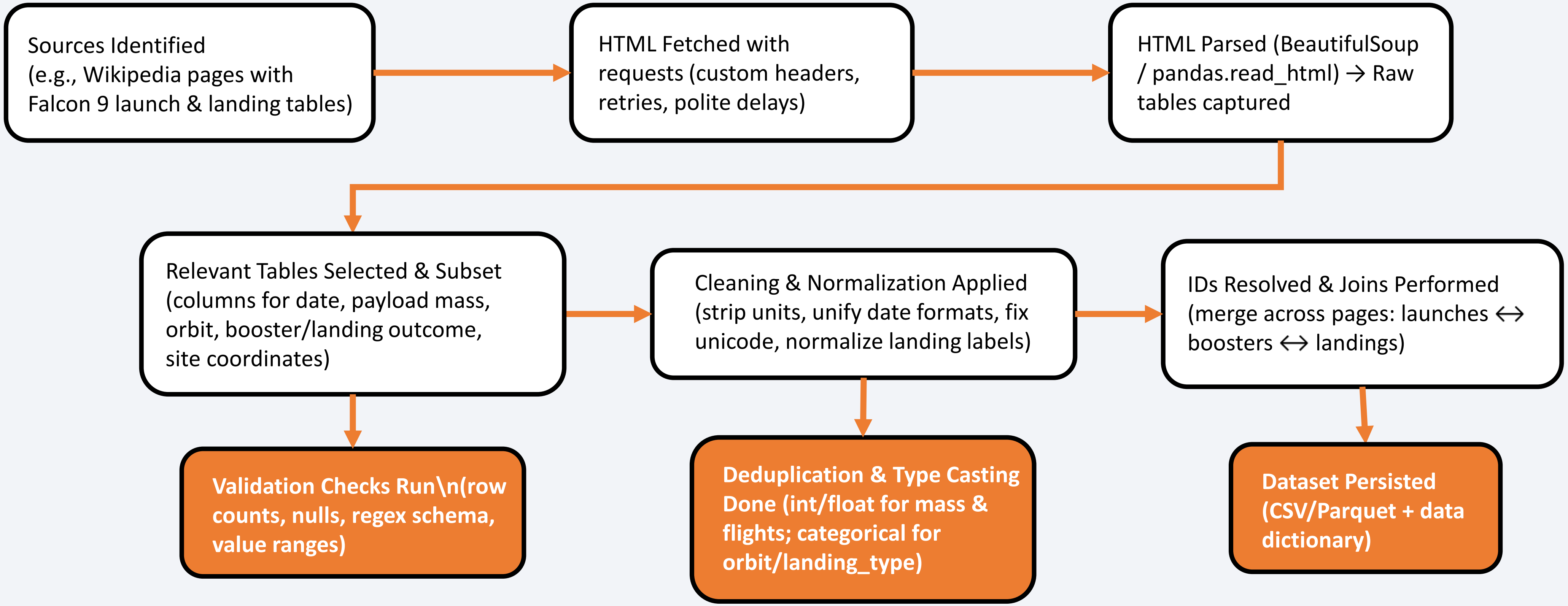
Data Collection - Scraping

[Click Here to Access the Notebook]

- Target pages were identified (e.g., Falcon 9 launch and landing tables on reference sites).
- HTML was fetched with `requests` using polite headers, retries, and delays.
- Tables were parsed via `BeautifulSoup / pandas.read_html`, and relevant columns were selected.
- Data was cleaned and normalized (units stripped, dates unified, labels standardized).
- Records were joined across pages to resolve IDs (launches ↔ boosters ↔ landings).
- Validation checks were executed (row counts, nulls, schema/regex, value ranges).
- Deduplication and type casting were performed (numeric for masses/flights; categorical for orbit/landing_type).
- The final dataset was persisted as CSV/Parquet with a brief data dictionary.”

Data Collection – Scraping

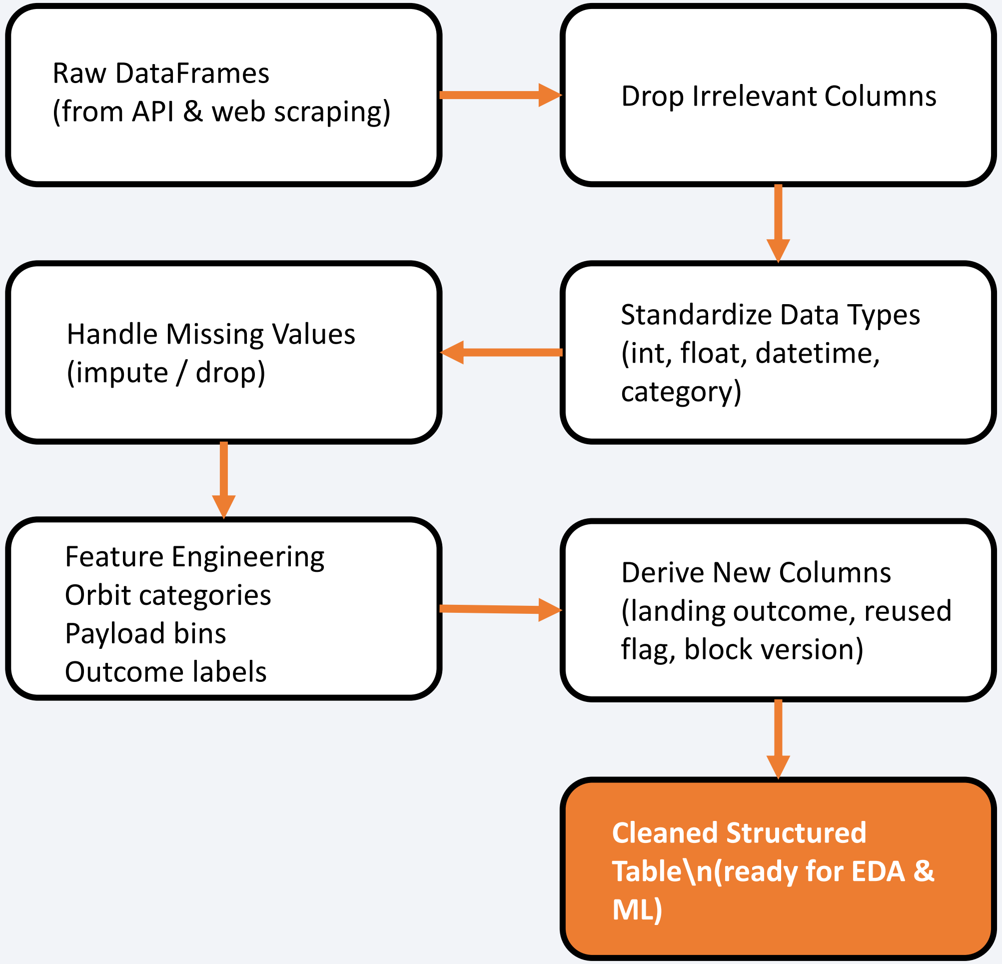
[Click Here to Access the Notebook]



Data Wrangling

[Click Here to Access the Notebook]

- The collected SpaceX data were converted into pandas DataFrames for processing.
- Irrelevant columns were dropped and missing values were handled (imputed or removed as appropriate).
- Data types were standardized (numeric, categorical, datetime).
- Feature engineering was applied: orbit categories, payload mass bins, success/failure outcomes.
- New columns were derived (e.g., landing outcome consolidated from multiple flags).
- Data were normalized and prepared as an analysis-ready structured table.”



EDA with Data Visualization

[Click Here to Access the Notebook]

Flight Number vs. Payload Mass (Catplot, hue = Class)

Why chosen: To test whether more flights (experience) and payload mass influenced success.
Key insight: Later flights generally had higher success rates, showing improvement with experience; payload mass was not a limiting factor once reliability improved.

1

Flight Number vs. Launch Site (Catplot)

Why chosen: To see if launch site location impacted outcomes.
Key insight: Success was achieved at all sites, but higher consistency was visible at specific sites (e.g., Cape Canaveral).

2

Payload Mass vs. Launch Site (Catplot)

Why chosen: To analyze if heavier payloads were riskier depending on site.
Key insight: Sites handled a wide range of payloads; no strong evidence of payload mass restricting success at a given site.

3

Orbit vs. Success Rate (Bar Plot)

Why chosen: To compare orbital destinations (LEO, GTO, Polar, etc.) with landing outcomes.
Key insight: Higher success rates were observed for LEO and SSO missions, while GTO missions were more challenging with lower average success.

4

Flight Number vs. Orbit (Scatter / Swarm)

Why chosen: To investigate if mission type (orbit) influenced landing outcomes over flight history.
Key insight: Earlier GTO missions had lower success, but reliability improved across all orbit types with later flights.

5

Payload Mass vs. Orbit (Swarm Plot)

Why chosen: To see if payload mass varied systematically by orbit and if that affected outcomes.
Key insight: Heavier payloads were concentrated in GTO missions, partially explaining their lower success rate.

6

Success Rate Over Time (Line Chart)

Why chosen: To identify overall performance improvement trend.
Key insight: A clear upward trajectory in success rate was seen, stabilizing near 90–100% in recent years which confirms SpaceX's operational maturity.

7

EDA with SQL

Click Here to Access the Notebook

Query	Purpose	Key Insight
`CREATE TABLE SPACEXTABLE … WHERE Date IS NOT NULL;`	Remove rows with missing dates	Ensured reliable time-based analysis
`SELECT DISTINCT Launch_Site …`	List unique launch sites	Established canonical set of sites
`SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;`	Inspect Cape Canaveral records	Quick check of site-specific data
`SELECT SUM(PAYLOAD_MASS__KG_) … WHERE Customer='NASA (CRS)';`	Aggregate payload mass by customer	NASA CRS accounted for significant payload
`SELECT AVG(PAYLOAD_MASS__KG_) … WHERE Booster_Version LIKE 'F9 V1.1%';`	Compute average payload for version	Compared payload capacity across booster versions
`SELECT DISTINCT Landing_Outcome …`	Retrieve unique landing labels	Confirmed categorical outcomes
`SELECT MIN(Date) … WHERE Landing_Outcome='Success (ground pad)';`	Identify first successful landing	Established milestone for timelines
`SELECT DISTINCT Booster_Version … WHERE Landing_Outcome='Success (drone ship)' AND Payload 4000-6000;`	Profile heavy-payload successes	Linked booster versions to heavier missions
`SELECT Mission_Outcome, COUNT(*) … GROUP BY Mission_Outcome;`	Summarize outcomes by category	Overall mission distribution observed
`SELECT DISTINCT Booster_Version … WHERE Payload=MAX(Payload);`	Retrieve booster at max payload	Identified peak performance
`SELECT CASE substr(Date,6,2)… WHERE Landing_Outcome LIKE 'Failure (drone ship)' AND Date LIKE '2015%';`	Extract months for failures	Analyzed seasonality/context of failures
`SELECT Landing_Outcome, COUNT(*) … BETWEEN 2010-06-04 AND 2017-03-20;`	Count outcomes in date window	Ranked outcomes by frequency

Build an Interactive Map with Folium

[Click Here to Access the Notebook]

Circles

- Circles were added at NASA Johnson Space Center and at each launch site coordinate.
- **Why:** to highlight the geographic footprint of each site with a clear, visible radius.

Markers

- Standard markers were placed at each launch site, with popup labels showing site names.
- Text-label markers (DivIcon) were added to annotate NASA JSC.
- **Why:** to provide clickable details and names for easy identification of locations.

Marker Clusters

- A MarkerCluster object was created to group individual launch event markers (success/failure).
- **Why:** to prevent clutter and allow zoom-in exploration of multiple launches at the same site.

Color-coded Markers

- Markers were colored according to launch outcome (marker_color derived from dataset).
- **Why:** to visually separate successful vs. failed launches, making patterns obvious.

Popups

- Each marker's popup displayed attributes such as launch site and class (success/failure).
- **Why:** to allow interactive inspection of individual launches.

Build a Dashboard with Plotly Dash

Click Here to Access the Notebook

PLOTS/GRAPHS ADDED

Pie Chart (Success by Site)

- Displays either total success launches across all sites or success vs. failure for a specific site
- Helps compare performance across sites and inspect detailed outcomes per site

Scatter Plot (Payload vs. Outcome)

- Shows relationship between payload mass and launch outcome, with point color representing booster version.
- Enables visual analysis of how payload and booster type affect success

INTERACTIONS ADDED

Pie Chart (Success by Site)

- Displays either total success launches across all sites or success vs. failure for a specific site
- Helps compare performance across sites and inspect detailed outcomes per site

Scatter Plot (Payload vs. Outcome)

- Shows relationship between payload mass and launch outcome, with point color representing booster version.
- Enables visual analysis of how payload and booster type affect success

Predictive Analysis (Classification)

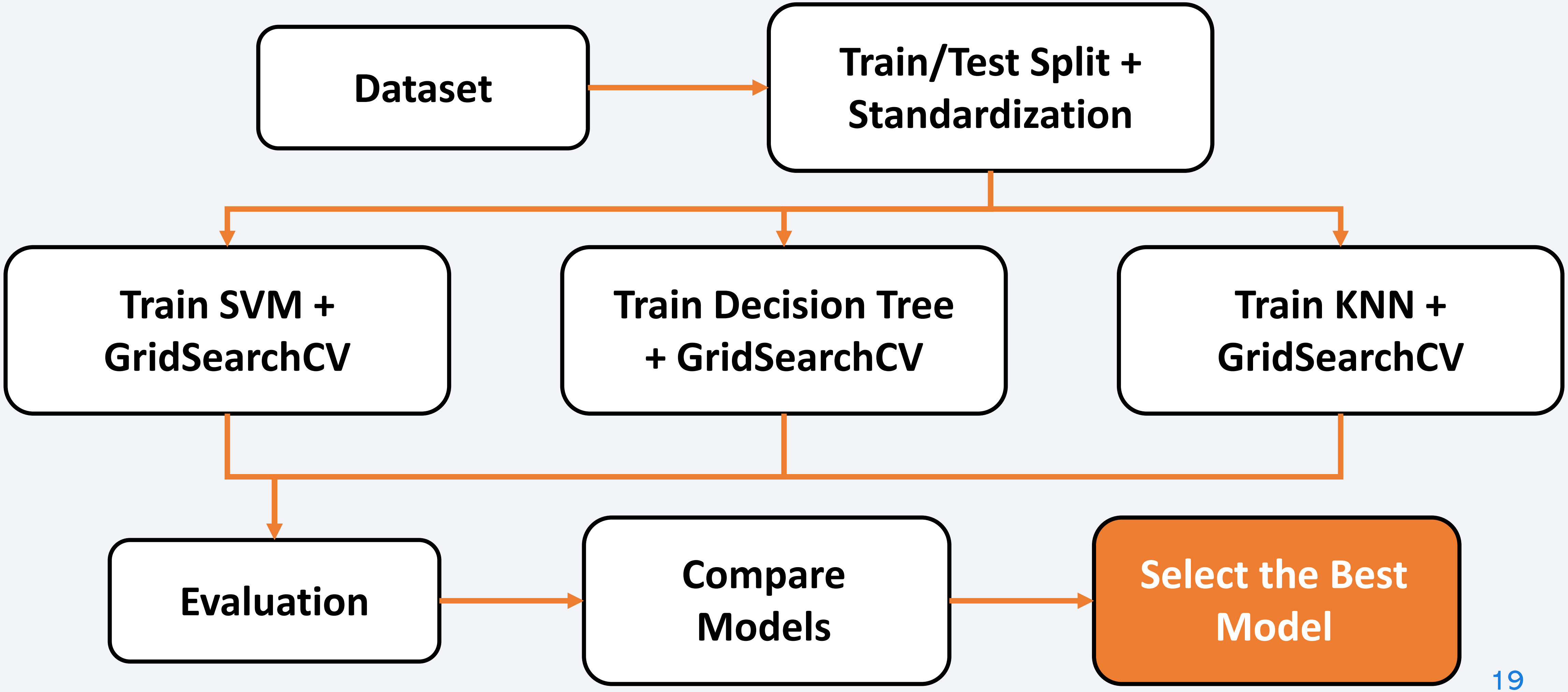
[Click Here to Access the Notebook]

Model Development Process

- **Data Preparation**
  - Features and labels were defined, and the dataset was split into training and testing sets.
  - Standardization was applied to ensure features were on comparable scales.
- **Model Building**
  - Multiple classification algorithms were tested: Logistic Regression, SVM, Decision Tree, and KNN.
  - Hyperparameters were tuned using GridSearchCV with cross-validation.
- **Evaluation**
  - Models were evaluated on the test set using accuracy score and confusion matrices.
  - Results were compared side by side to identify strengths and weaknesses of each classifier.
- **Improvement**
  - Hyperparameter optimization improved model performance (e.g., best penalty and solver for Logistic Regression, kernel/C/gamma for SVM).
  - Decision boundaries and misclassifications were analyzed via confusion matrix visualization.
- **Best Model**
  - Among all tested algorithms, the best-performing classifier (highest test accuracy with balanced results) was identified and chosen for final prediction tasks.

Predictive Analysis (Classification)

Click Here to Access the Notebook

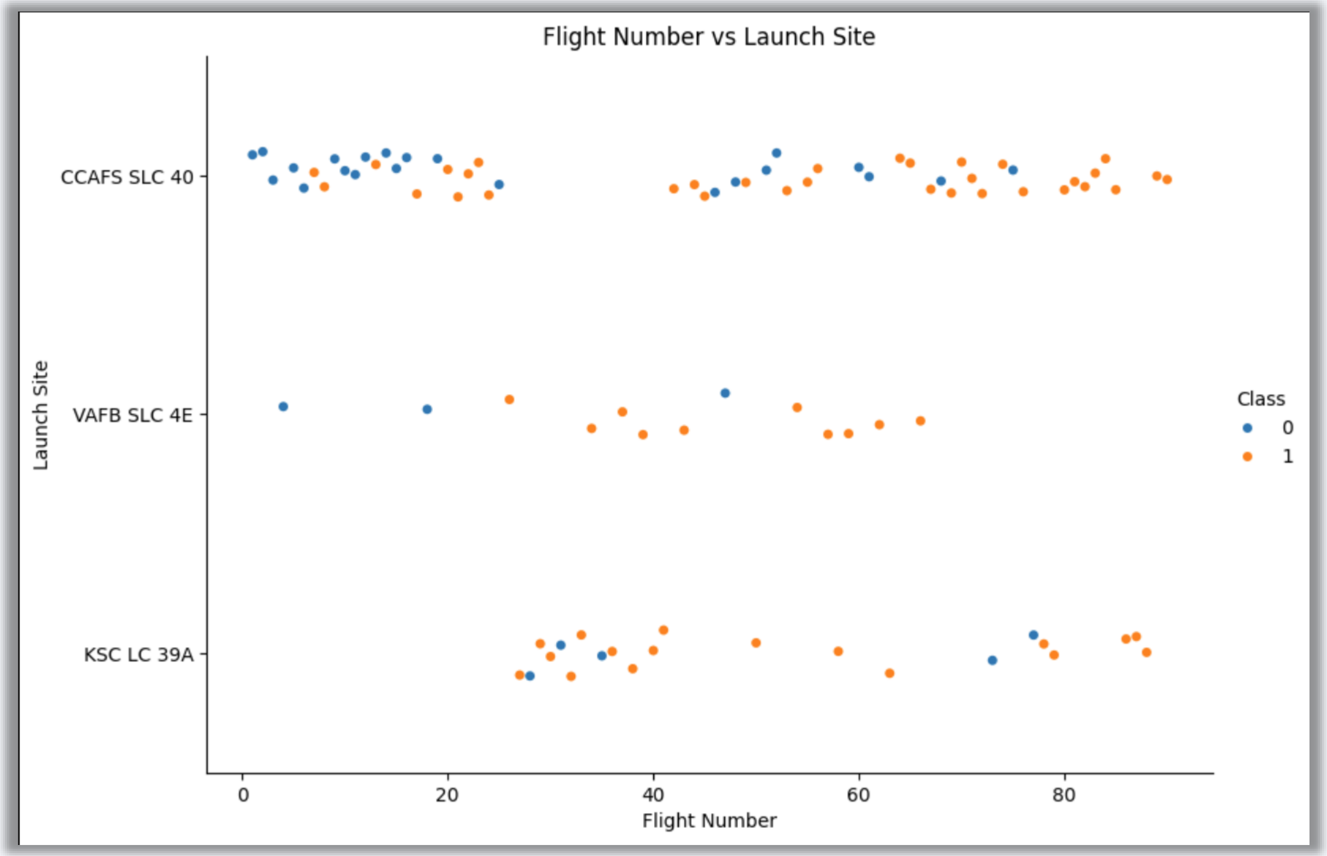


Section 2

Insights drawn from EDA

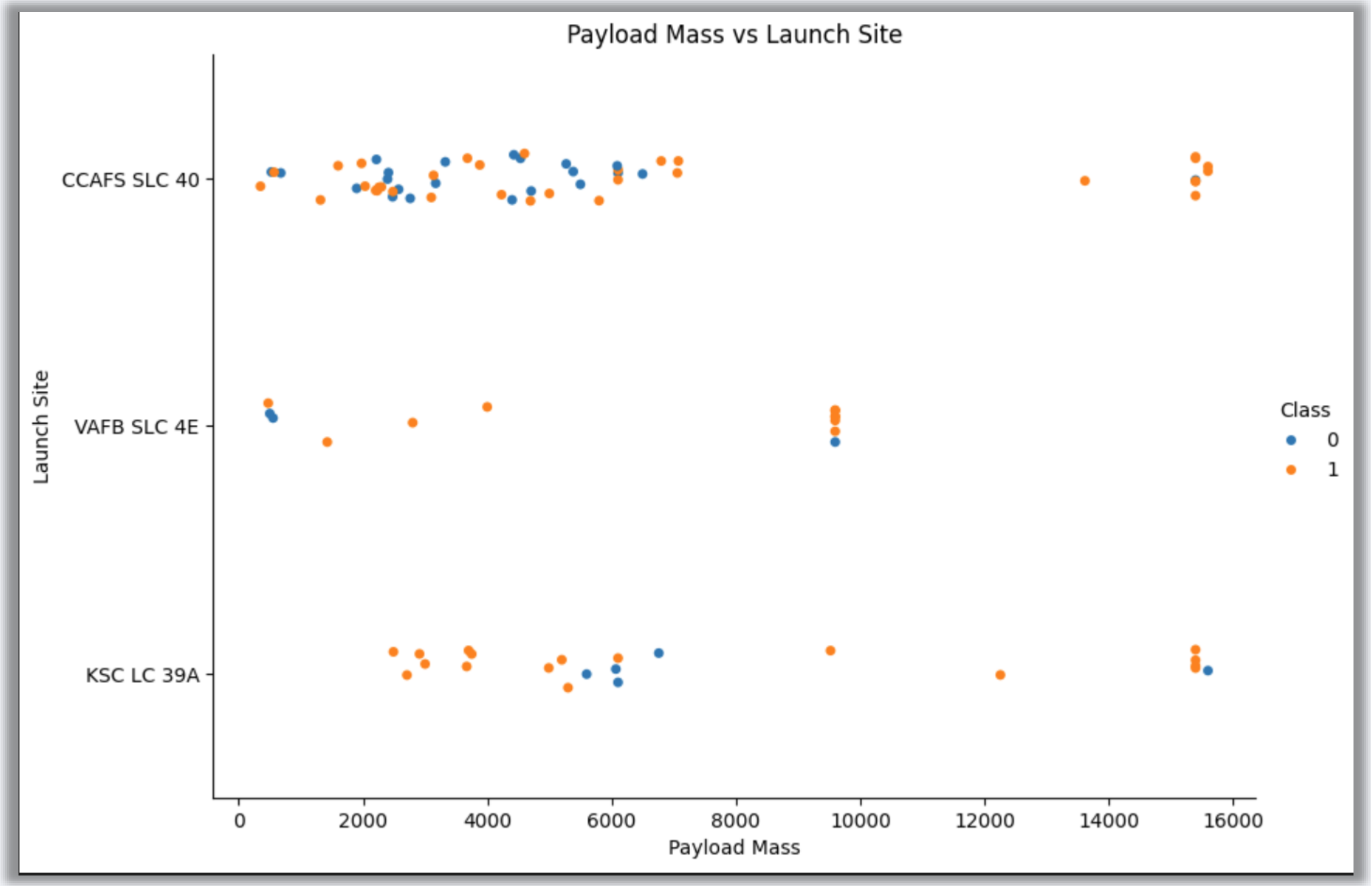
Flight Number vs. Launch Site

- **Purpose:** To analyze how the number of flights correlates with launch success at each site.
- **Observation**
  - Success rates increased with higher flight numbers, especially at Cape Canaveral SLC-40 and KSC LC-39A.
  - Early flights showed inconsistent results, while later flights achieved more consistent success.
- **Insight:** Launch experience and iterative improvements significantly boosted landing reliability.



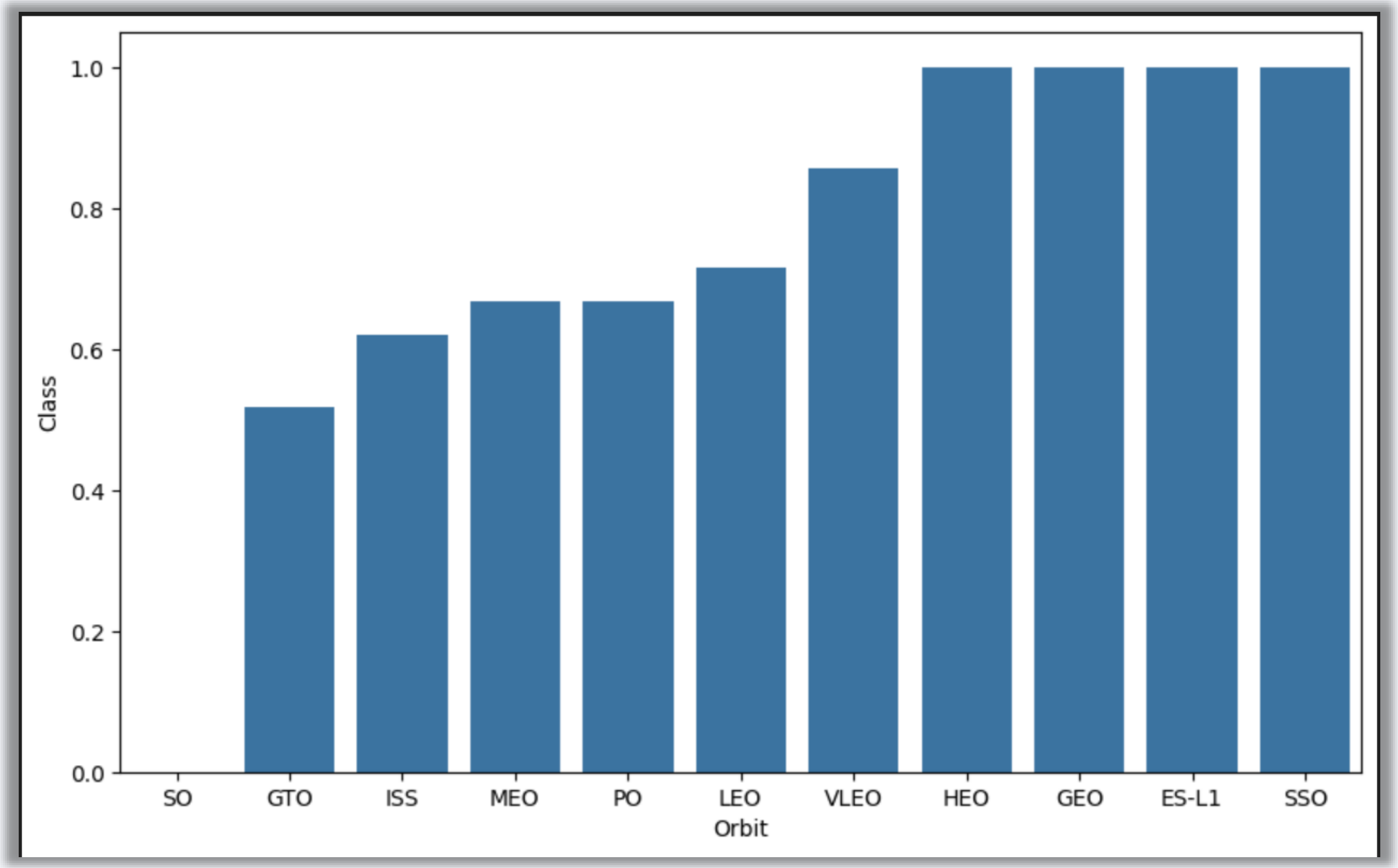
Payload vs. Launch Site

- **Purpose:** To determine if heavier payloads affected success rates across different sites.
- **Observation**
  - Most payloads ranged between 2,000–6,000 kg across sites.
  - Success rates remained high even for heavier payloads, particularly at the main Cape Canaveral sites.
  - KSC LC 39A has a 100% success rate with payloads lower than 5000 kg
- **Insight:** Payload mass did not significantly limit landing success once booster reliability matured.



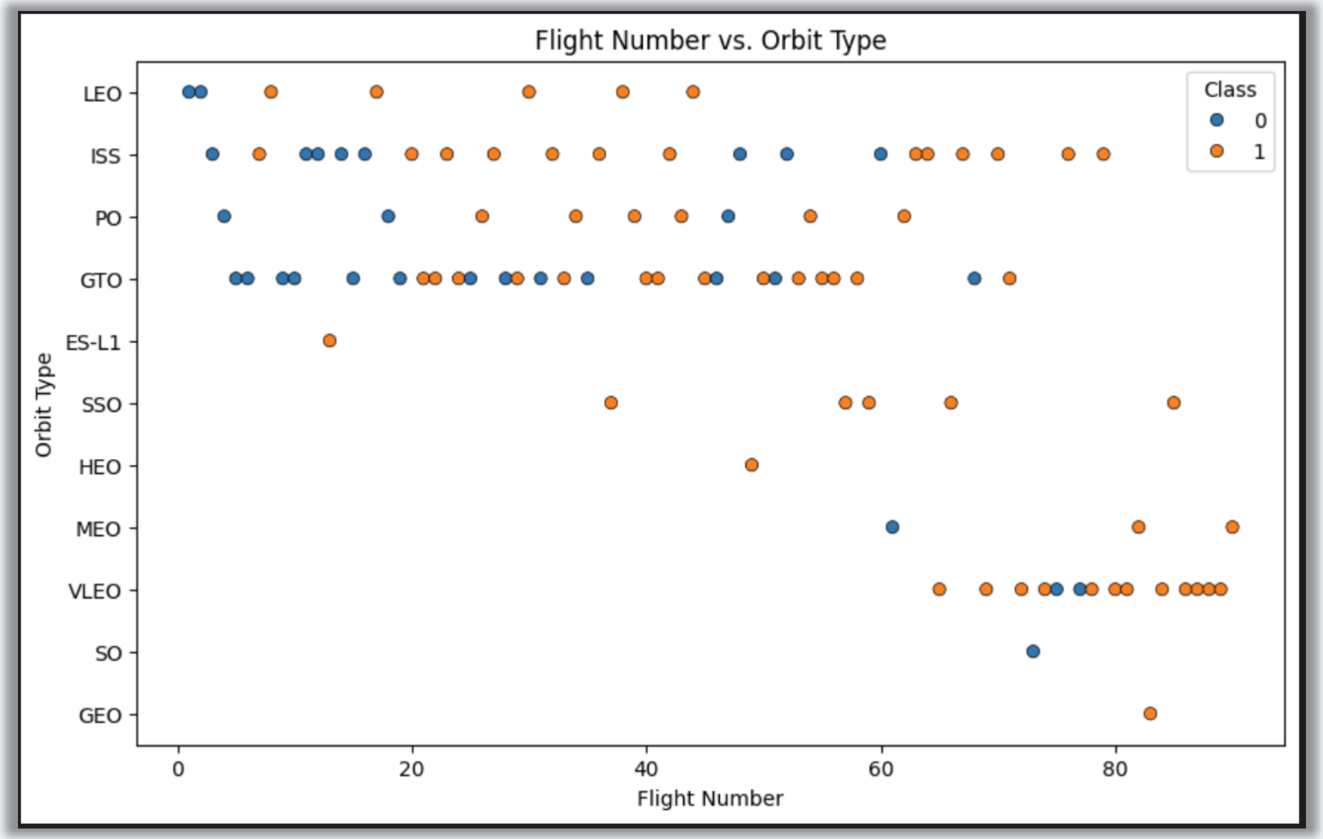
Success Rate vs. Orbit Type

- **Purpose:** To compare landing success across different orbital destinations.
- **Observation**
  - LEO, GEO, ES-L1 and SSO had the highest success rates.
  - SO orbit had no successful
  - GTO, ISS, MEO and PO missions had comparatively lower success rates due to higher energy and distance requirements.
- **Insight:** Orbit type influenced mission difficulty which higher orbits correlated with reduced landing success probability.



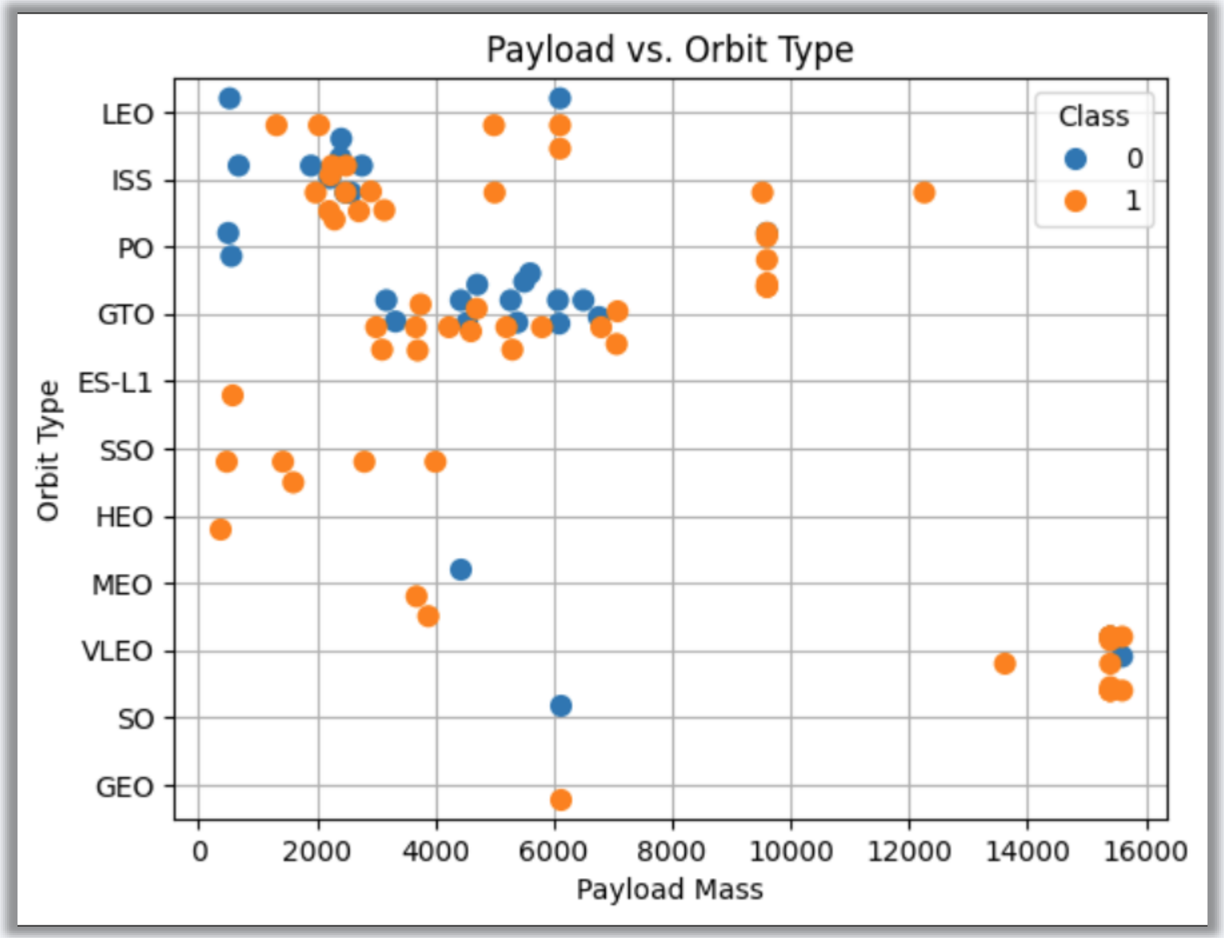
Flight Number vs. Orbit Type

- **Purpose:** To explore how mission types evolved with flight experience.
- **Observation**
  - Earlier missions were primarily LEO, ISS, PO and GTO and had mixed outcomes.
  - Later missions showed more successful orbits across all categories, reflecting improved technology and experience.
- **Insight:** Mission diversity increased with experience, and success stabilized across all orbit types.



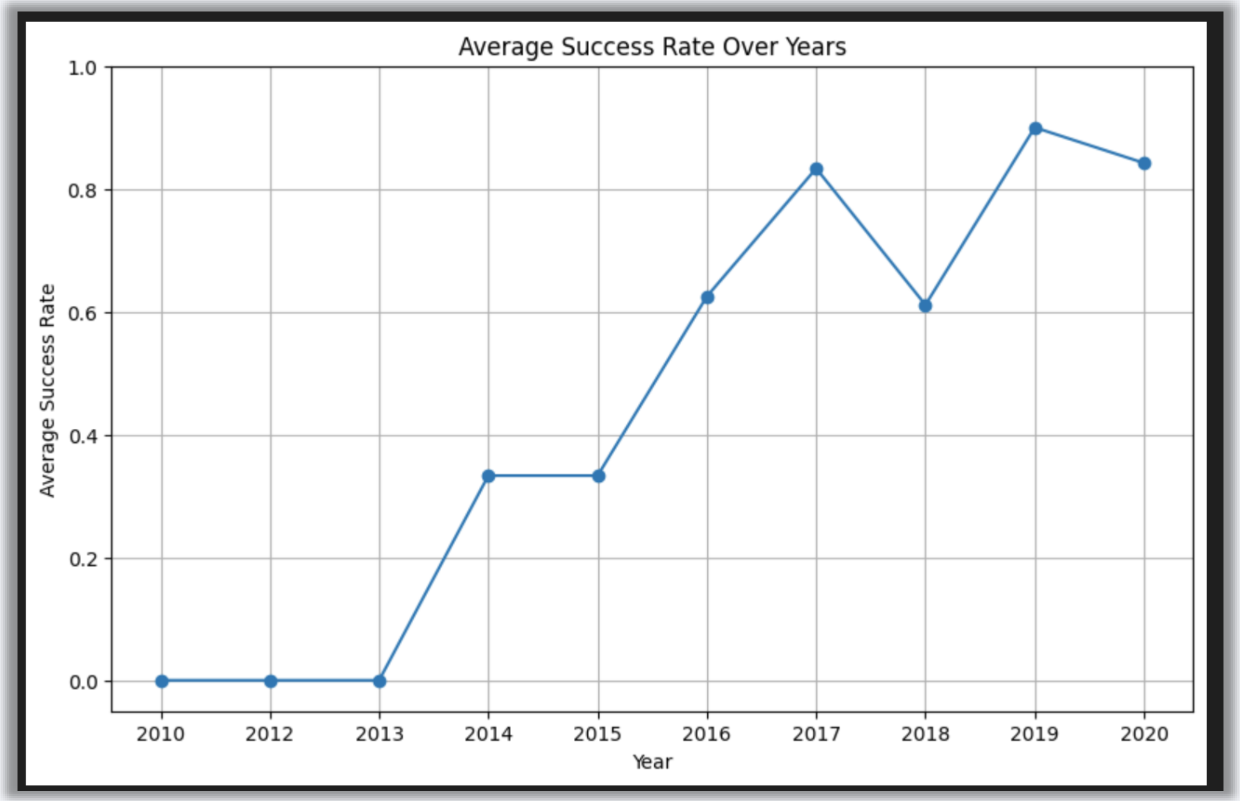
Payload vs. Orbit Type

- **Purpose:** To examine if payload mass varied with orbit and its relationship to success.
- **Observation**
  - Heavier payloads were generally sent to GTO, while lighter payloads went to LEO and SSO.
  - Success rates remained higher for VLEO missions, even with varying payloads.
- **Insight:** Orbit type was a more critical factor than payload mass in determining landing success



Launch Success Yearly Trend

- **Purpose:** To observe how SpaceX's overall landing success evolved over time
- **Observation**
  - Early years showed failures until 2013, from there it showed fluctuated results until 2017.
  - From 2017 onward, success rates climbed sharply and stabilized around 90–100%.
- **Insight:** Continuous technological refinement and operational experience led to exponential improvement in reliability — marking SpaceX's transition to consistent booster recovery.



All Launch Site Names

```
[28]: %%sql

SELECT DISTINCT "Launch_Site"
FROM SPACEXTBL

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

[28]:

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

- All distinct SpaceX launch sites were retrieved from the dataset.
- **Finding:** Four unique launch sites were identified — mainly at Cape Canaveral, Kennedy Space Center, and Vandenberg.
- **Insight:** SpaceX primarily operates from a small, specialized set of coastal sites suitable for orbital launches.

Launch Site Names Begin with 'CCA'

- Records were filtered where the site name begins with ‘CCA’.
- **Finding:** All matching records correspond to Cape Canaveral Air Force Station (SLC-40).
- **Insight:** This confirmed that multiple missions have launched from the Cape Canaveral site, showing its central role in SpaceX operations.

[31]:
```
%%sql

SElECT * FROM SPACEXTBL
WHERE "Launch_Site" LIKE "CCA%"
LIMIT 5;
```

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

[31]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
[34]: %%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

- The total payload mass launched for customer NASA (CRS) was calculated.
- **Finding:** The aggregated payload mass amounted to 45.596 kg.
- **Insight:** NASA's CRS (Commercial Resupply Service) missions constitute a major share of SpaceX's cargo to the ISS, highlighting its strategic partnership.

Average Payload Mass by F9 v1.1

- The mean payload mass for booster version F9 v1.1 was computed.
- **Finding:** The average payload was around 2.534,66 kg, reflecting the vehicle's mid-range capacity.
- **Insight:** The F9 v1.1 variant marked a transition phase between early and modern high-capacity Falcon 9 boosters.

```
[37]: %%sql

      SELECT AVG("PAYLOAD_MASS__KG_") as average_f9v1_payload
      FROM SPACEXTBL
      WHERE "Booster_Version" LIKE "F9 V1.1%"

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

[37]:

average_f9v1_payload
2534.6666666666665

First Successful Ground Landing Date

```
[39]: %%sql

      SELECT MIN("Date") AS first_successful_ground_landing
      FROM SPACEXTBL
      WHERE "Landing_Outcome" = "Success (ground pad)";

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

[39]:

first_successful_ground_landing
2015-12-22

- The earliest date with a “Success (ground pad)” landing outcome was identified
- **Finding:** The first successful ground landing occurred in December 2015.
- **Insight:** This event represented a historic milestone as the first vertical booster recovery on land.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters were filtered for “Success (drone ship)” landings within that payload range.
- **Finding:** Several Block 3/4 boosters achieved successful landings under these heavy-payload conditions.
- **Insight:** Demonstrated that SpaceX could successfully recover boosters even on high-energy missions, validating drone-ship landing reliability.

```
[42]: %%sql

SELECT DISTINCT "Booster_Version"
FROM SPACEXTBL
WHERE "Landing_Outcome" = "Success (drone ship)"
AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000

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

[42]:

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

Total Number of Successful and Failure Mission Outcomes

- Mission outcomes were grouped and counted by success and failure categories.
- **Finding:** Successful missions significantly outnumbered failures, especially after 2017.
- **Insight:** Reinforces the rapid improvement of SpaceX's launch reliability over time.

```
[45]: %%sql
SELECT "Mission_Outcome", COUNT(*) AS total_missions
FROM SPACEXTBL
GROUP BY "Mission_Outcome";

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

[45]:

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

Boosters Carried Maximum Payload

```
[48]: %%sql

SELECT DISTINCT "Booster_Version"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL
);

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

[48]:

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

- The boosters associated with the maximum recorded payload mass were listed.
- **Finding:** The heaviest payloads were launched by later-generation Falcon 9 Block 5 boosters.
- **Insight:** Confirms technological progress and enhanced lifting capacity in newer booster designs.

2015 Launch Records

- 2015 records were filtered for "Failure (drone ship)" outcomes.
- **Finding:** Multiple failed drone-ship attempts were identified from early booster versions (e.g., F9 v1.1).
- **Insight:** Drone-ship recoveries initially had a learning curve; later missions achieved consistent success.

[51]:
```
%%sql
SELECT
    CASE substr("Date", 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS Month,
    "Booster_Version",
    "Launch_Site",
    "Landing_Outcome"
FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE 'Failure (drone ship)'
AND substr("Date", 0, 5) = '2015';
```

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

[51]:

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

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

[54]:
```
%%sql

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

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

[54]:

Landing_Outcome	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

- Landing outcomes were counted and ranked in descending order within the time range.
- **Finding:** “Success (drone ship)” and “Success (ground pad)” outcomes dominated the ranking.
- **Insight:** Over this period, SpaceX transitioned from experimental recoveries to routine, repeatable booster landings.

Section 3

Launch Sites Proximities Analysis

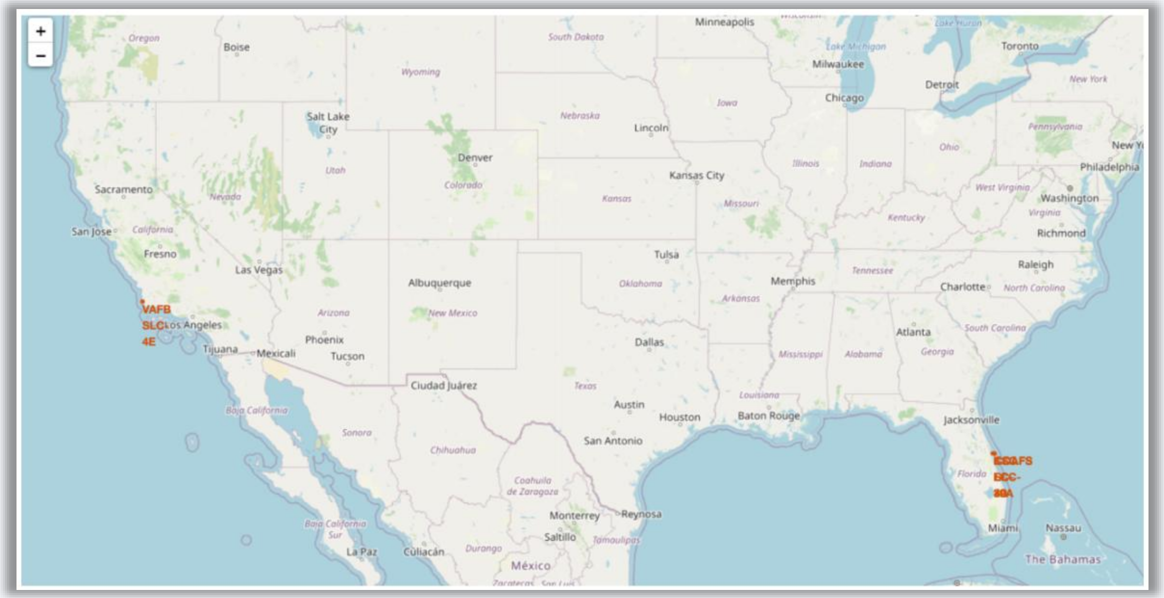
Global Map with All Launch Site Names

Key Elements

- Each launch site was represented on a world map using folium.
- Marker and Circle objects.
- Popups and text labels were added to display site names (e.g., CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E, BoCA CHICA).
- The map provided a global overview of all active SpaceX launch facilities.

Finding

- All sites were found to be in the Northern Hemisphere and positioned relatively close to the equator (latitudes between ~25°N and ~35°N).
- Launch sites being closer to the equator provides a rotational velocity advantage since the Earth's rotation speed is highest near the equator, allowing rockets to gain additional velocity, thereby reducing the fuel required to reach orbit.
- Geographically, all SpaceX launch sites were situated near coastal regions, not inland.



Insight: Launch sites are strategically placed near the equator and along coastlines to optimize rocket efficiency and ensure safety.

Color Labeled Launch Outcomes

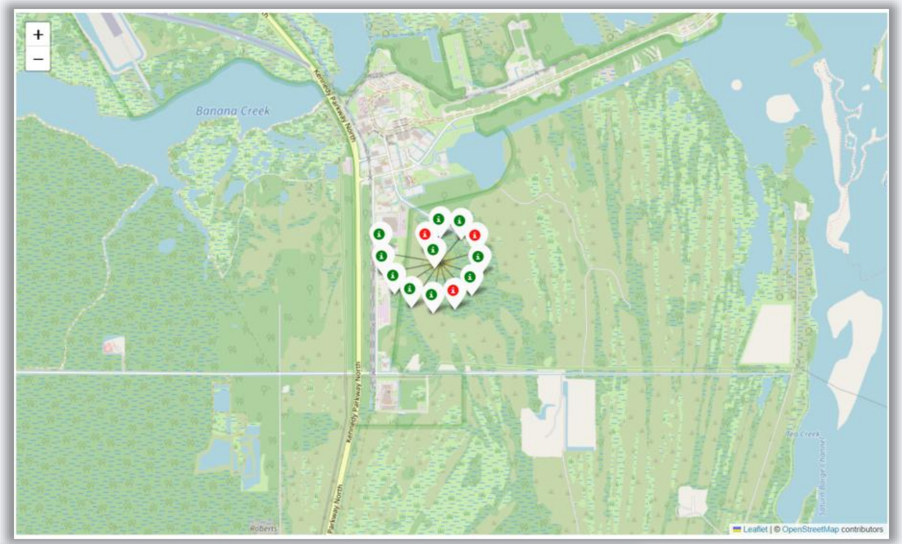
Key Elements

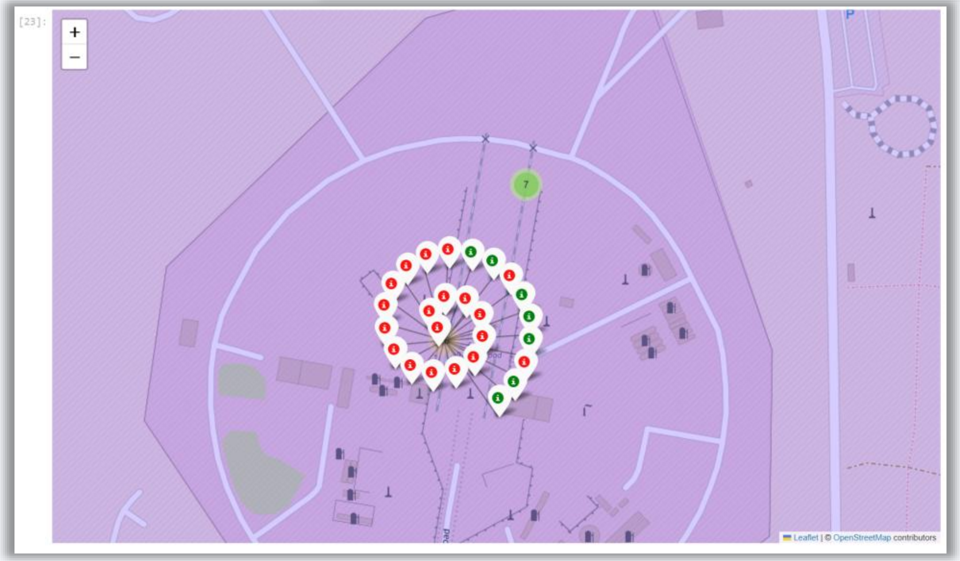
- Markers on the Folium map were color-coded based on launch outcomes using the marker_color field.
    - Green markers represented successful launches.
    - Red markers represented failed launches.
- Popups were used to display details such as launch site name, payload mass, orbit type, and landing outcome.

Finding

- Most markers were green, indicating a high proportion of successful launches across all sites.
- A few red markers (failures) appeared near earlier launch coordinates, showing that failures were primarily from early missions.
- Clustering of successful markers was observed around Cape Canaveral and Kennedy Space Center, confirming these sites as the most frequently used and operationally mature.

Insight: The color-coded outcomes visually demonstrated SpaceX's improving success rate and the geographical consistency of reliable launches across its major sites..





Distance from the Selected Launch Site to its Proximities

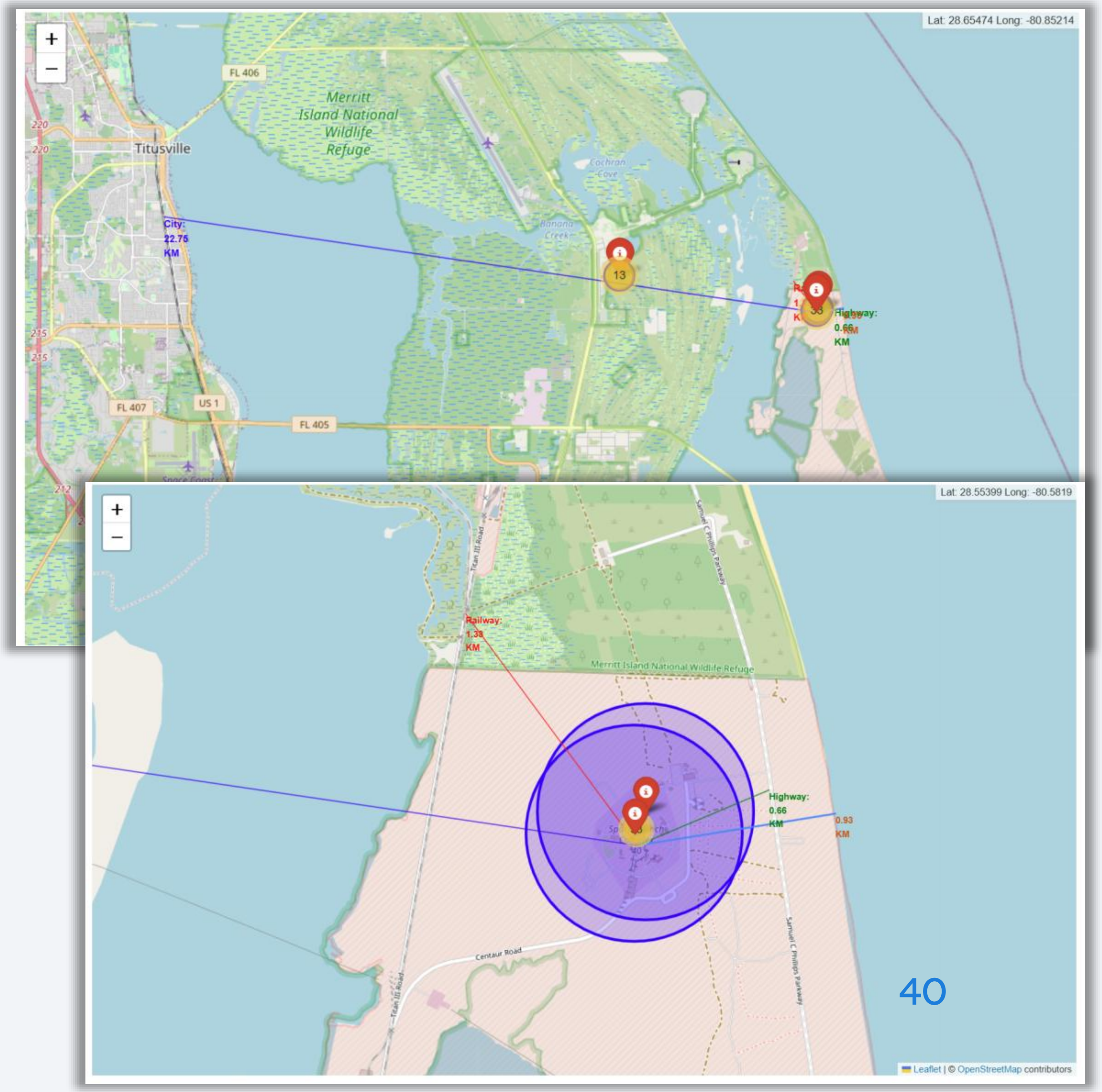
Key Elements

- A selected launch site (e.g., CCAFS SLC-40) was connected to nearby locations such as:
  - Coastline
  - Railway
  - Highway
  - City center
- Folium.PolyLine objects were used to draw distance lines between the launch site and each proximity point.
- Popup boxes displayed the measured distances (in kilometers).

Finding

- The launch site was found very close to the coastline, confirming its ocean-facing orientation.
- Shortest distances were to the coastline (few km), while the longest were to the nearest city center.
- The presence of nearby highways and railways indicates logistical convenience for transporting rocket components and fuel.
- These proximity measurements validated that SpaceX's launch sites were strategically located for both safety (coastal launch path) and infrastructure access (transport connectivity).

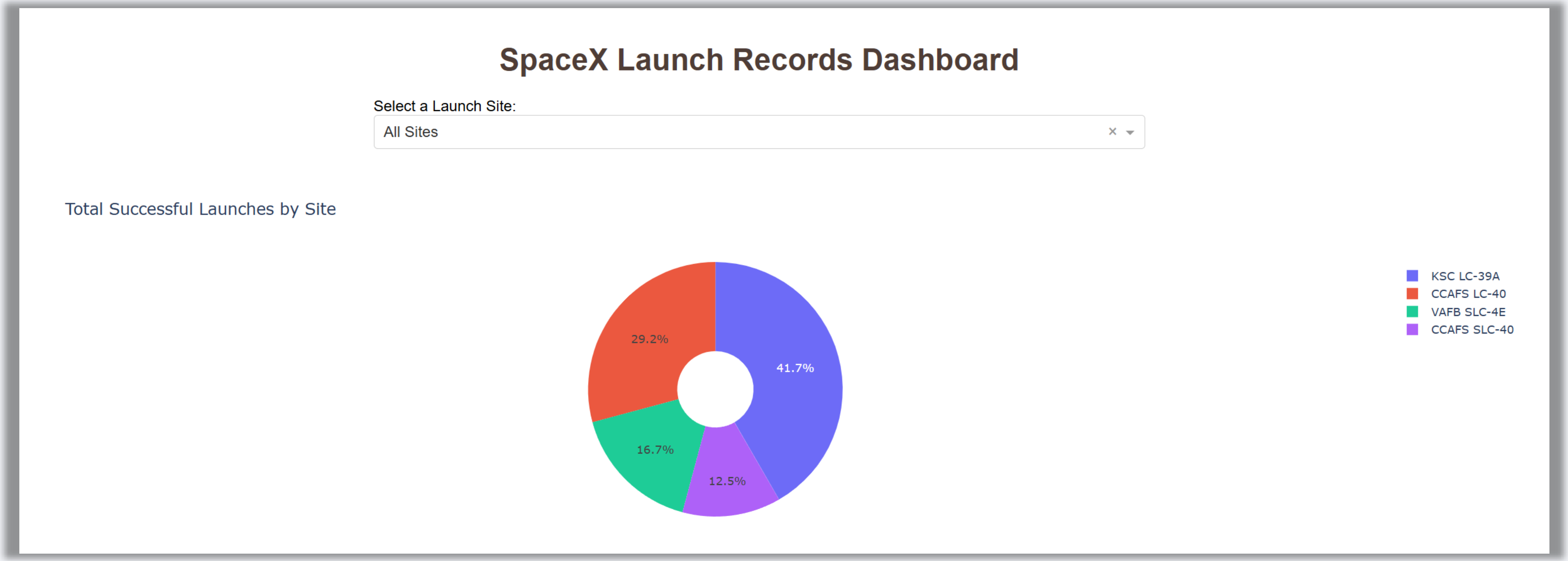
Insight: Folium's distance analysis visually confirmed that SpaceX launch sites are coastal, logistically connected, and optimized for operational safety and efficiency.



Section 4

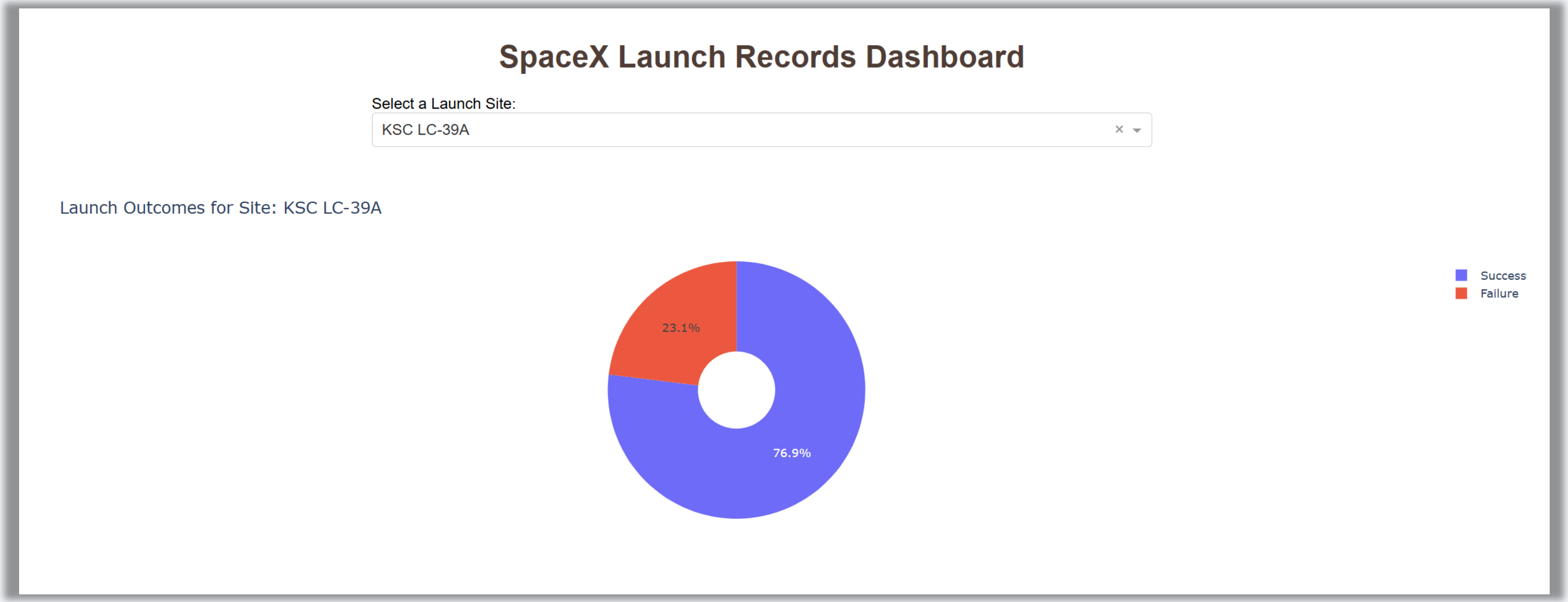
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



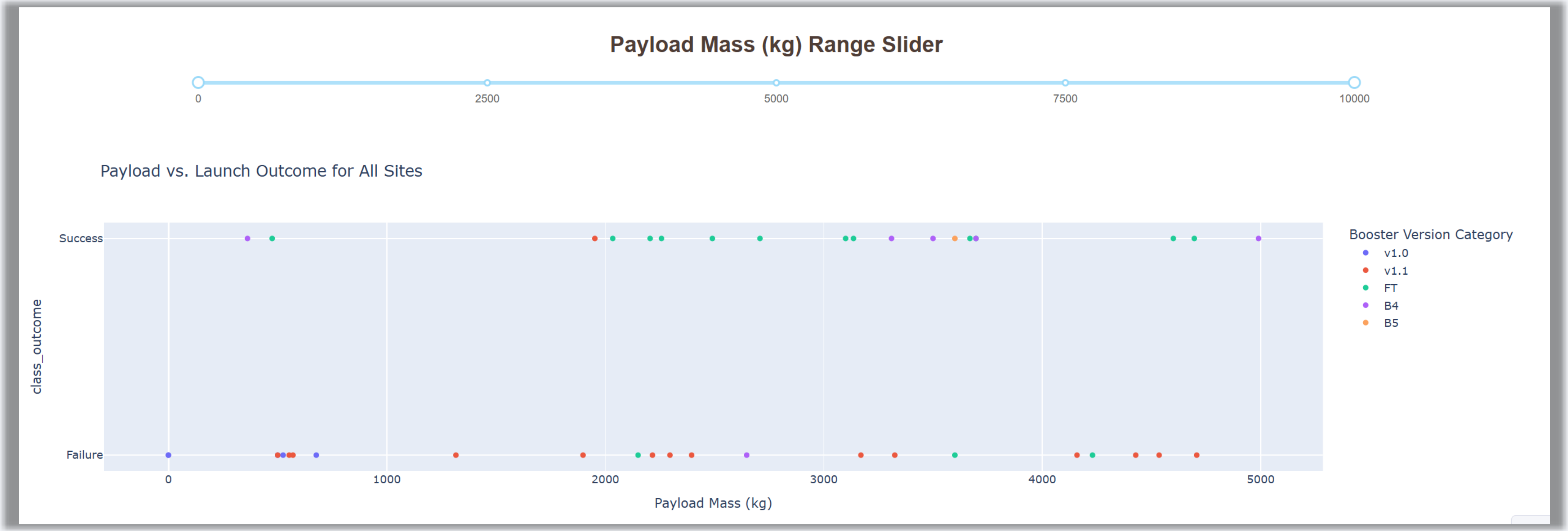
- **KSC LC-39A** accounts for the largest share of successes (~42)
- CCAFS LC-40 is next (~29%), followed by VAFB SLC-4E (~17%) and CCAFS SLC-40 (~12%).

Most Successful Launch Site



- The pie-chart shows that the most successful is KSC LC-39A among the launch sites.

Payload vs. Launch Outcome for All Sites



- Booster version drives success far more than payload
- Payload range does not show a strong monotonic effect.
- Newer boosters are more successful.

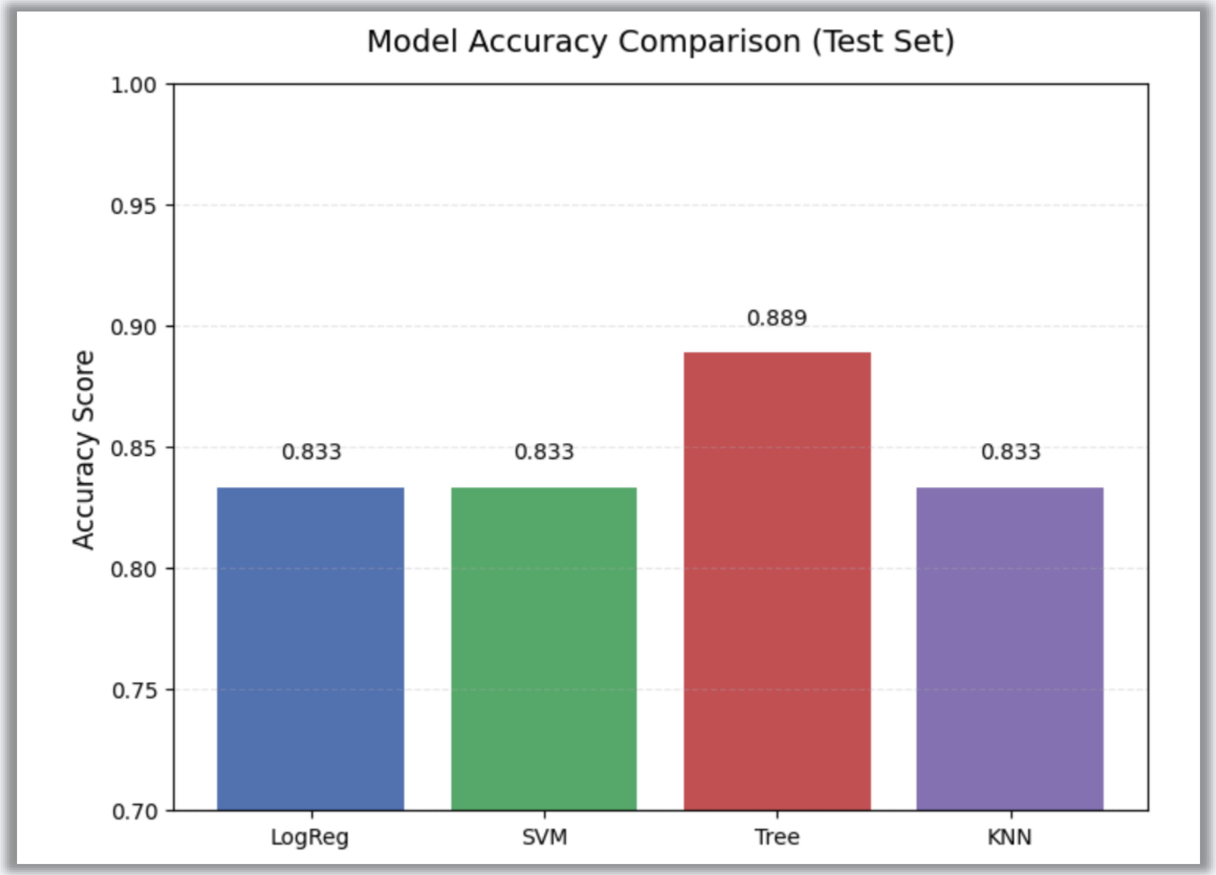
Section 5

Predictive Analysis (Classification)

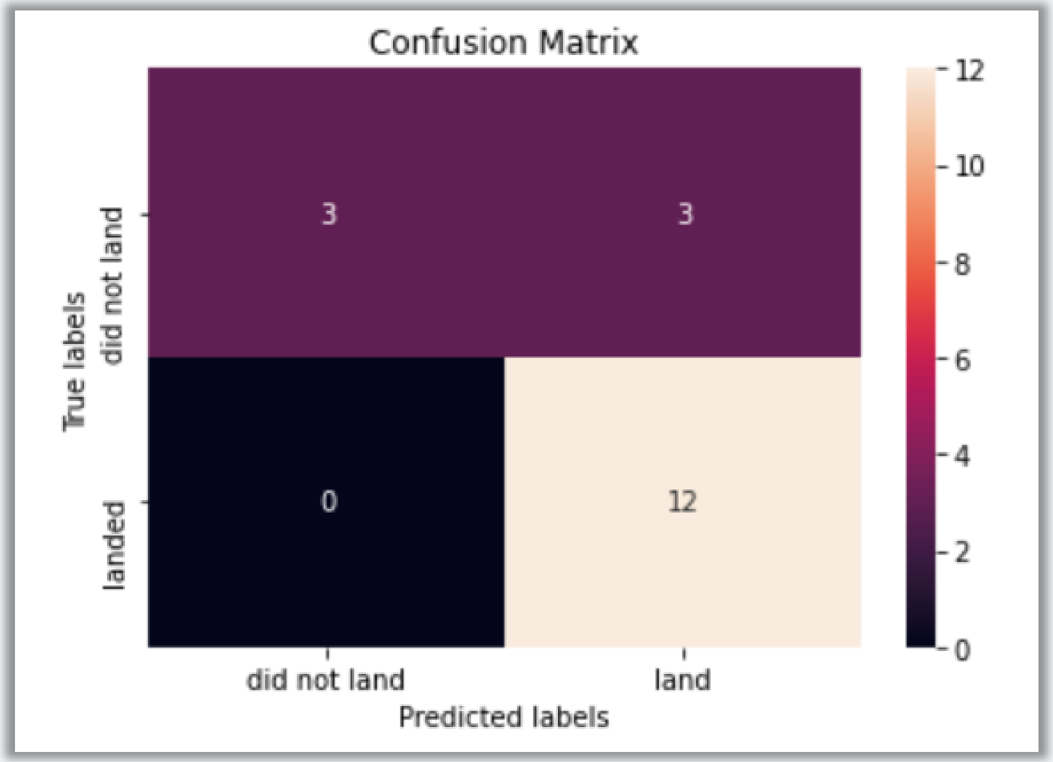
Classification Accuracy

Interpretation

- The Decision Tree model achieved the highest accuracy (0.8889) on the test set, slightly outperforming Logistic Regression, SVM, and KNN (all ≈0.8333).
- All models demonstrated consistent Jaccard (0.8) and F1-scores (0.8889), indicating similar classification performance.
- The Decision Tree may have captured slightly more complex relationships in the dataset, but care must be taken to verify that it does not overfit.
- The Logistic Regression, SVM, and KNN models offered comparable, stable generalization performance with slightly lower but balanced accuracy.



Confusion Matrix



- The absence of false negatives shows the model is very reliable in detecting successful landings — a critical advantage in operational decision-making.
- Some false positives indicate minor overprediction, suggesting potential for refinement (e.g., adjusting decision threshold or adding cost-sensitive penalties).
- Overall, the matrix confirms strong predictive performance, with the model achieving high precision and perfect recall for successful landings.
- The model can accurately forecast reusable booster landings, minimizing the risk of underestimating successful recoveries.
- Occasional overpredictions (false positives) can be managed through operational review rather than impacting mission-critical decisions.

Conclusions

The goal of this project was to predict the success of SpaceX Falcon 9 first-stage landings using real-world data obtained from SpaceX REST API and web-scraped mission records. Accurately predicting landing outcomes provides valuable insights into reusability and cost-efficiency, which are central to SpaceX's competitive advantage.

- SpaceX's success rate has steadily increased with flight experience, highlighting a strong learning curve and engineering improvement cycle.
- Launch sites are strategically positioned near the equator and coastal areas, leveraging Earth's rotation for efficiency and ensuring safe ocean-bound trajectories.
- Lower success in GTO missions compared to LEO/SSO reflects the higher energy demands of distant orbits.
- Machine learning models accurately captured landing success patterns, proving the feasibility of data-driven reliability forecasting.

- A complete data-science pipeline was implemented — from API collection and web scraping to data cleaning, SQL analysis, EDA, and machine learning.
- Predictive models achieved ~85–90% accuracy, with the Decision Tree performing best overall.
- Insights confirm that booster reusability and mission success can be predicted using historical launch data.
- The framework demonstrates how data analytics supports strategic decision-making, optimizing costs, improving reliability, and strengthening SpaceX's competitive advantage in reusable rocket technology.

Thank you!