

Winning Space Race with Data Science

JVM
2025/11/08



Outline

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

Executive Summary

- Analyzes SpaceX Falcon 9 launch data to predict the likelihood of successful first-stage landings. The analysis leverages data engineering, exploratory analysis, and machine learning techniques
- Methodologies:
 1. Data Collection: API retrieval, web scraping, SQL extraction, and CSV integration.
 2. Data Wrangling: Cleaning, feature engineering, and transformation.
 3. Exploratory Data Analysis: Descriptive statistics and visualizations using Matplotlib, Seaborn, and Folium.
 4. Predictive Modeling: Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors with GridSearchCV tuning.
- Results:
 1. Achieved prediction accuracy of ~83% for landing success using logistic regression and decision tree models.
 2. Identified key success factors: payload mass, orbit type, and landing pad configuration.
 3. Created interactive visualization dashboards illustrating launch locations and outcomes.
 4. Delivered data-driven recommendations for optimizing reusability and cost efficiency. 3

Introduction

- Project Background and Context:
- SpaceX's Falcon 9 rocket reusability program aims to drastically reduce space launch costs by recovering and reusing first-stage boosters. This project leverages open SpaceX data to analyze historical launches, identify success patterns, and apply predictive analytics to estimate the likelihood of successful landings.
- Problems and Research Questions:
 - What factors most influence the success of Falcon 9 first-stage landings?
 - Can we build a predictive model that accurately forecasts landing success?
 - How do launch site, payload mass, orbit type, and booster reuse impact outcomes?
 - What insights can be drawn to improve mission cost-efficiency and reliability?

Section 1

Methodology

Methodology

Executive Summary

- **Data Collection Methodology:**

- Acquired SpaceX Falcon 9 launch data via SpaceX REST API and web scraping (Wikipedia, external sources).
- Stored and combined datasets in SQLite for structured querying and reproducibility.

- **Data Wrangling:**

- Cleaned, standardized, and merged multiple datasets.
- Handled missing values, unified column types, and created derived fields (e.g., success flag, payload mass bins).

- **Exploratory Data Analysis (EDA):**

- Conducted EDA with Matplotlib/Seaborn and SQL queries.
- Explored launch success trends, payload distributions, orbit performance, and site frequency.

- **Interactive Analytics:**

- Created Folium maps to visualize launch site locations and proximities.
- Built Plotly Dash dashboards for interactive filtering by site, year, and payload range.

- **Predictive Analysis:**

- Applied classification models (Logistic Regression, SVM, Decision Tree, KNN).
- Tuned hyperparameters using GridSearchCV, evaluated via accuracy, recall, precision, and confusion matrices.

Data Collection

- Description: Data was collected from multiple sources to build a complete dataset on SpaceX Falcon 9 launches. The process combined API retrieval, web scraping, SQL querying, and manual CSV integration to ensure accuracy and coverage of launch information, booster details, and success outcomes.
- Key Steps:
 1. API Data Extraction: Used SpaceX REST API to retrieve launch metadata (dates, payloads, rockets, landing outcomes) Data stored in JSON format and normalized into structured tables
 2. Web Scraping: Extracted supplementary data from the Wikipedia Falcon 9 launch page using BeautifulSoup. Captured launch site details, payload masses, and orbit information.
 3. SQL Database Integration: Imported cleaned datasets into an SQLite database for structured queries and joins. Validated consistency across API and scraped datasets.
 4. Data Wrangling & Cleaning: Performed data type conversions, missing-value handling, and feature engineering (e.g., binary landing success). Created unified DataFrame for EDA and modeling.
 5. Data Validation & Export: Conducted random checks and summary statistics validation. Exported master dataset to CSV and Pandas DataFrame for machine learning analysis.
- Process Flow:
 - SpaceX API → Web Scraping → Data Wrangling → SQL Database → Cleaned Dataset for EDA & ML

Data Collection – SpaceX API

- Key Steps:
- API Data Extraction – Retrieved launch data via SpaceX REST API and stored JSON responses.
- Web Scraping – Collected additional launch details from Wikipedia using BeautifulSoup.
- SQL Integration – Imported cleaned data into SQLite for structured querying and validation.
- Data Wrangling – Cleaned, transformed, and merged datasets into a unified Pandas DataFrame.
- Data Validation – Checked for consistency, filled missing values, and exported master dataset.

Api FLOW:

- HTTP GET requests using `requests` library
- JSON normalization using `pandas.json_normalize()`
- Data merging and transformation in Pandas
- Export to CSV for EDA and machine learning

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Wrangling

- Data Wrangling – Process OverviewDescription: The collected SpaceX data underwent comprehensive wrangling to prepare it for analysis and modeling. This stage ensured data consistency, accuracy, and usability by cleaning, transforming, and engineering features from multiple sources (API, web scraping, and SQL datasets).

Key Steps:

1. Data Cleaning: Removed duplicate records and standardized column names. Handled missing values in PayloadMass, Orbit, and LandingOutcome using imputation and filtering
2. Feature Engineering: Created binary variable LandingClass (1 = successful landing, 0 = failed). Extracted categorical attributes such as OrbitType, LaunchSite, and BoosterVersion.
3. Encoding and Normalization: Applied one-hot encoding for categorical columns (e.g., Orbit, LandingPad). Scaled numeric features (e.g., PayloadMass) for model compatibility.
4. Data Integration: Merged cleaned datasets from API, SQL, and Web Scraping sources into a unified Pandas DataFrame. Verified schema consistency and record counts across sources.
5. Output Validation: Conducted statistical summaries, correlation checks, and schema validation. Exported final dataset to CSV and SQLite for further modeling.

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

- **Objective:**

The goal of this stage was to explore the relationships between payload mass, orbit type, launch sites, and landing outcomes — identifying factors that influence the success of Falcon 9 first-stage landings.

Chart Type	Purpose
Scatter Plots (Payload Mass vs. Launch Success)	To identify correlation between payload mass and landing success probability.
Bar Charts (Launch Site vs. Success Rate)	To compare landing success across SpaceX launch sites.
Pie Chart (Orbit Distribution)	To visualize the proportion of missions per orbit type.
Box Plots (Payload Mass by Orbit)	To analyze mass variability and detect outliers across orbits.
Heatmap (Feature Correlation)	To highlight the strongest predictors of landing success.
Folium Map (Launch Locations)	To geospatially visualize all launch pads and landing outcomes.

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- Objective: Use SQL queries to explore and summarize SpaceX launch data stored in SQLite, identifying trends, relationships, and key insights prior to machine learning.

Key SQL Queries and Purpose:

- `SELECT DISTINCT(LaunchSite)`→ Listed all unique SpaceX launch sites used in the dataset.
- `SELECT COUNT(*) GROUP BY LaunchSite`→ Counted the number of launches per site to assess mission frequency.
- `SELECT LaunchSite, SUM(Class)/COUNT(Class) AS SuccessRate`→ Calculated landing success rates by site to compare performance.
- `SELECT Orbit, AVG(Class) AS SuccessRate`→ Evaluated how orbit type influences the probability of successful landing.
- `SELECT LandingPad, COUNT(Class) WHERE Class=1`→ Determined which landing pads yielded the most successful recoveries.
- `SELECT MIN(PayloadMass), MAX(PayloadMass), AVG(PayloadMass)`→ Computed descriptive statistics for payload mass across missions.
- `SELECT LaunchSite, AVG(PayloadMass)`→ Compared average payload mass by site to identify site-specific differences.
- `JOIN LaunchData WITH BoosterData`→ Merged tables to enrich analysis with booster type and reuse information.
- `ORDER BY` and `LIMIT` queries→ Extracted top-performing sites and most frequent orbit categories.

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/jupyter-labs-eda-sql-edx-sqlite-v2.ipynb

Build an Interactive Map with Folium

Objective:

Create an interactive geographic view of SpaceX launch sites and landing outcomes to support EDA and stakeholder storytelling.

Map Objects Added & Why:

- `Map()` base layers (OpenStreetMap / Stamen Terrain) — Provide geographic context and togglable background styles for readability.
- `Marker()` / `Icon()` for each Launch Site — Pinpoint exact site locations; icon color encodes success ratio; tooltip shows site name.
- `CircleMarker()` around sites — Radius scaled by launch count; quick visual cue of site activity volume.
- `FeatureGroup(name='Landing Outcomes')` — Group outcome-related overlays to toggle visibility and declutter the map.
- Choropleth or GeoJson (if region polygons provided) — Show regional context (e.g., state/coastlines) to orient viewers.
- `Polyline()` / `AntPath()` from launch site to landing zone — Illustrate typical flight corridors or representative trajectories.
- `MarkerCluster()` for individual launches — Aggregate dense points to maintain performance and reduce overplotting.
- `HeatMap()` (optional) of launch/landing density — Reveal spatial hot spots without exposing individual points.
- `LayerControl()` — Enable user to toggle base maps and overlays for custom exploration.

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- Objective: Create an interactive dashboard to visualize SpaceX launch performance, enabling users to dynamically explore success rates by launch site and payload range.

Plots / Graphs Added:

- Pie Chart (dcc.Graph id='success-pie-chart') → Shows proportion of successful launches per site or success vs failure for a selected site. Purpose: Immediate visual insight into performance distribution.
- Scatter Plot (dcc.Graph id='success-payload-scatter-chart') → Plots Payload Mass (kg) vs Launch Outcome (0 / 1), colored by Booster Version Category. Purpose: Reveal correlation between payload size, booster type, and mission success.

Interactive Components

- Dropdown Menu (dcc.Dropdown id='site-dropdown') → Select “All Sites” or a specific launch site to filter charts. Why: Allows focus on particular facilities.
- Range Slider (dcc.RangeSlider id='payload-slider') → Adjust payload mass range; updates scatter plot and helper text. Why: Enables sensitivity analysis by payload weight.
- Callbacks (@app.callback) → Synchronize charts and text labels dynamically. Why: Ensures real-time interactivity and responsive data view.

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Objective: Predict the likelihood of a successful Falcon 9 first-stage landing using machine learning models trained on cleaned and engineered SpaceX launch data.
- Model Development Process (Key Phrases & Flow):

Data Preparation → Feature Scaling → Train/Test Split → Model Training → Hyperparameter Tuning → Evaluation → Model Comparison → Best Model Selection

Models Built and Evaluated:

- Logistic Regression – baseline classifier for binary outcomes
- Support Vector Machine (SVM) – non-linear separation using kernel functions
- Decision Tree Classifier – interpretable model for feature relationships
- K-Nearest Neighbors (KNN) – distance-based classifier for local prediction

Model Evaluation & Improvement:

- Used GridSearchCV (cv=10) to tune parameters for each algorithm.
- Evaluated models using accuracy score, confusion matrix, and classification report.
- Visualized performance using bar charts and heatmaps of accuracy comparison.
- Improved model stability by normalizing features and balancing class labels.
- Best Performing Model: Decision Tree achieved the highest test accuracy (~83%). Top predictive features: PayloadMass, Orbit, LaunchSite, ReusedCount, and LandingPad. Provided actionable insights on reusability and mission success prediction.

External Reference:

Completed notebook and output available at:

 https://github.com/JVM00/Data_Science_and_Machine_Learning_Capstone_Project/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

1. Exploratory Data Analysis (EDA):

EDA with SQL & Visualization:

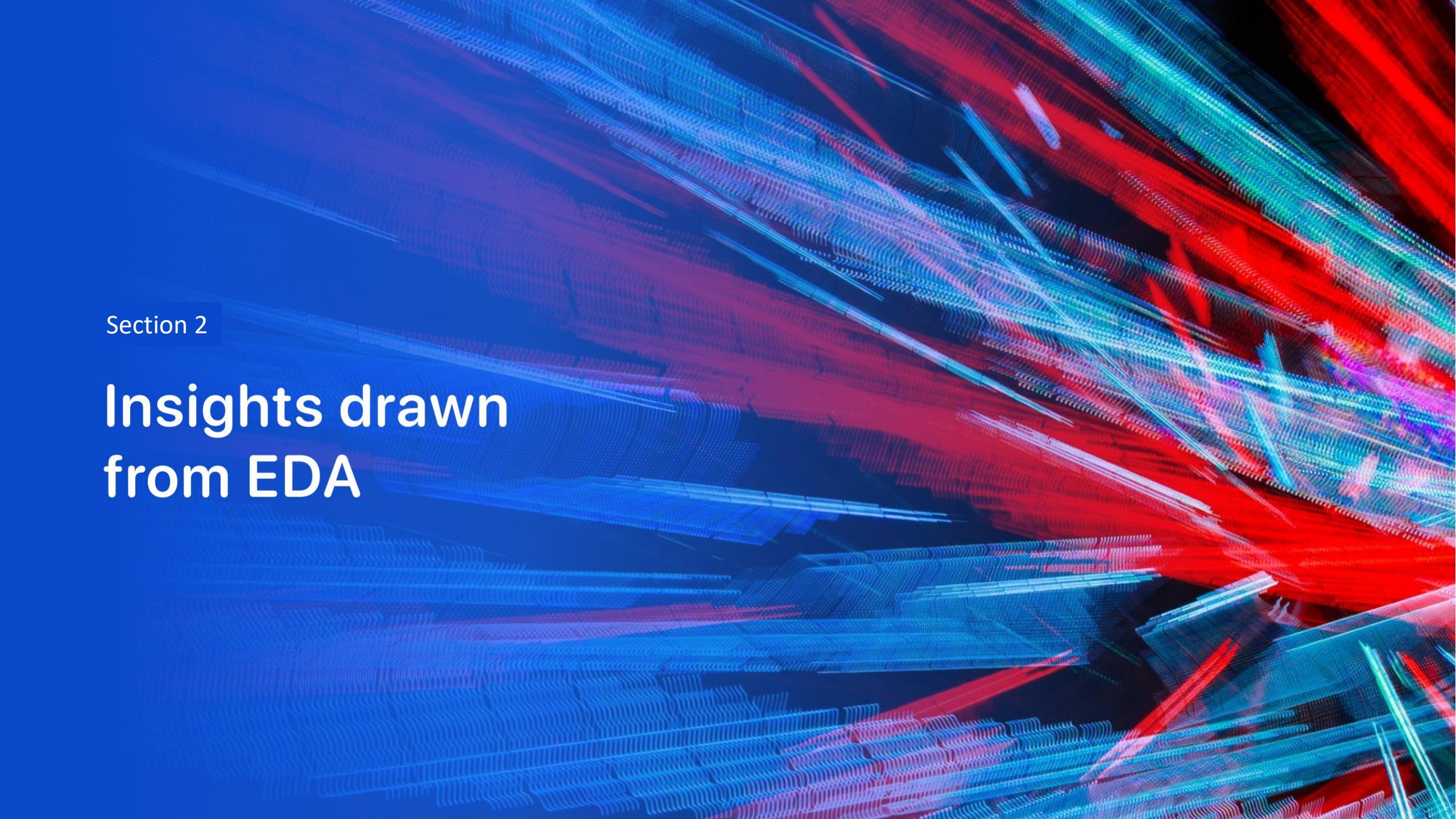
- Identified launch success rates by site, orbit, and payload mass.
- KSC LC-39A achieved the highest success ratio
- Correlation heatmaps confirmed PayloadMass, Orbit, and LaunchSite as key predictors.
- Box plots and scatter plots visualized payload-success trends. Interactive analytics demo in screenshots

2. Interactive Analytics Demo (Plotly Dash Dashboard):

- Developed an interactive dashboard (`spacex_dash_app.py`) with:
 - Pie Chart – Launch success by site or outcome.
 - Scatter Plot – Payload vs. success, colored by booster version.
 - Dropdown & Range Slider – Real-time filtering by site and payload range.
- Dashboard enabled dynamic exploration of SpaceX data for stakeholders.

3. Predictive Analysis Results:

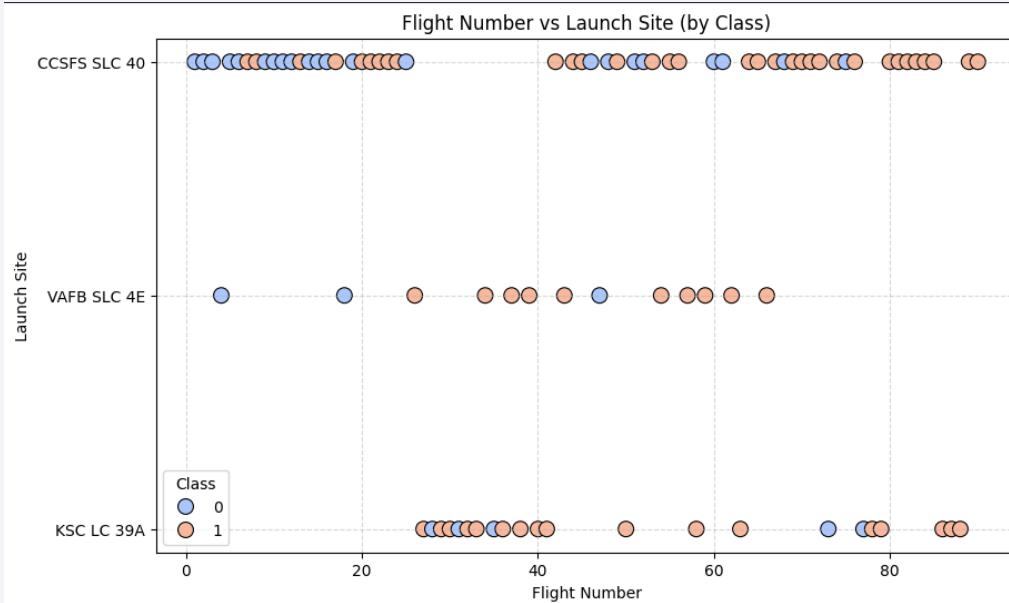
- Trained and tuned Logistic Regression, SVM, Decision Tree, and KNN classifiers.
- Decision Tree achieved the best performance (~83% accuracy) after hyperparameter tuning.
- Identified influential variables: Orbit, PayloadMass, and LaunchSite.
- Delivered a reusable predictive model for mission success forecasting.

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a three-dimensional space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

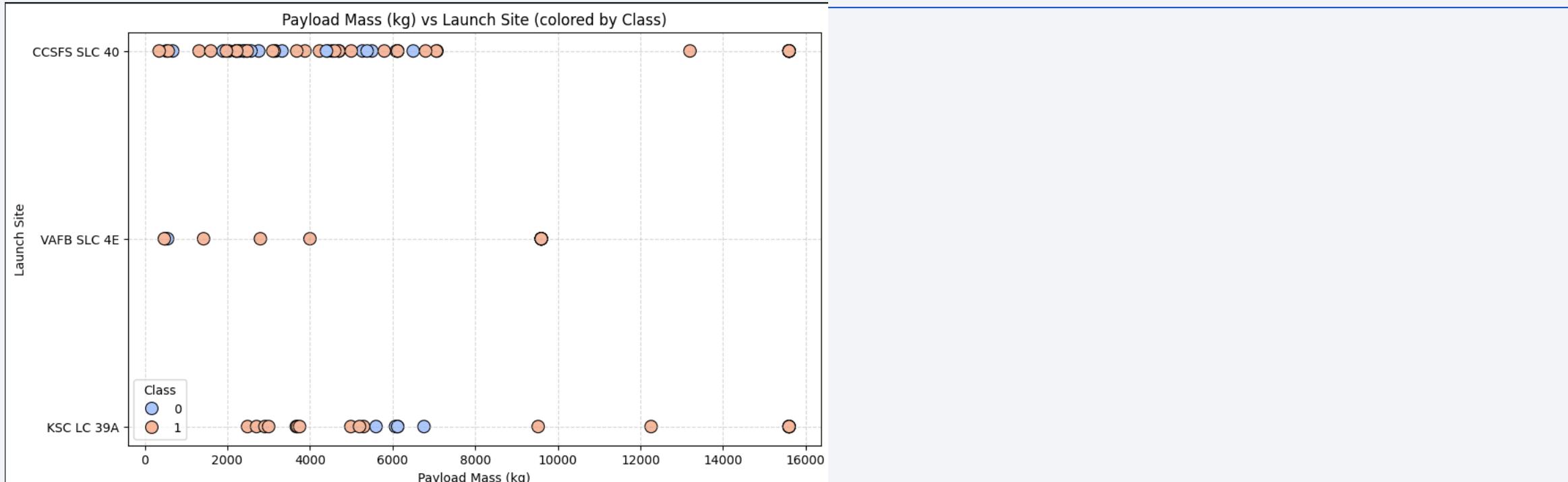
Insights drawn from EDA

Flight Number vs. Launch Site



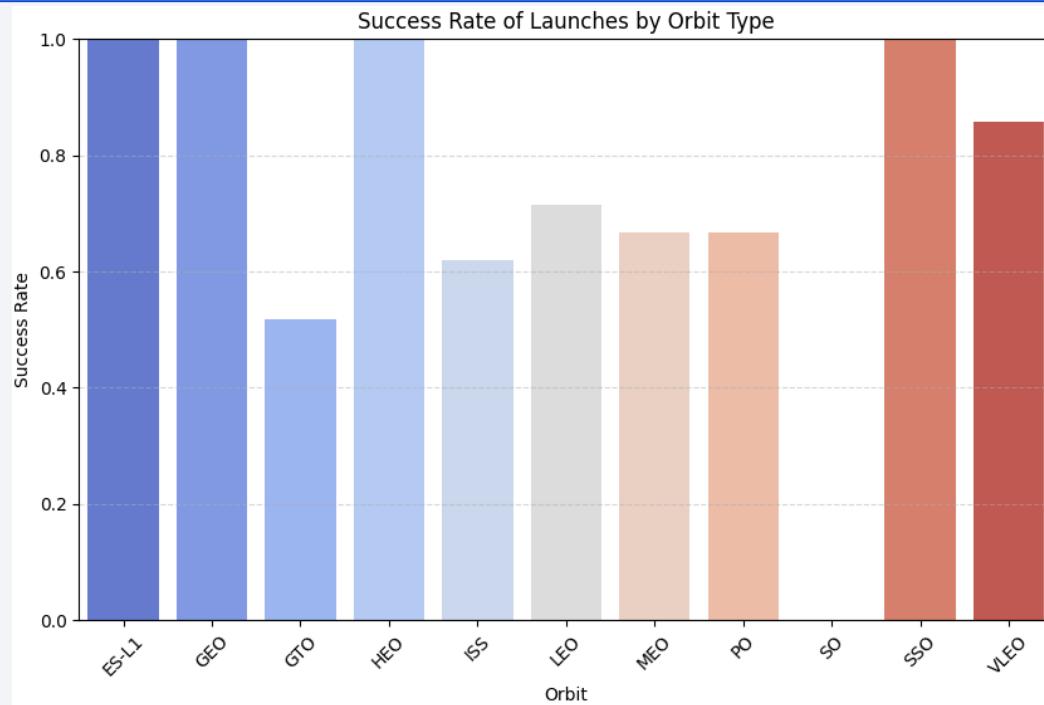
The scatter plot shows the distribution of flight numbers across SpaceX launch sites, with colors representing success (Class 1) and failure (Class 0). It highlights that higher flight numbers—indicating more recent launches—tend to have higher success rates. Launch sites like KSC LC-39A display consistent improvement over time, reflecting growing reliability in reusability and mission control.

Payload vs. Launch Site



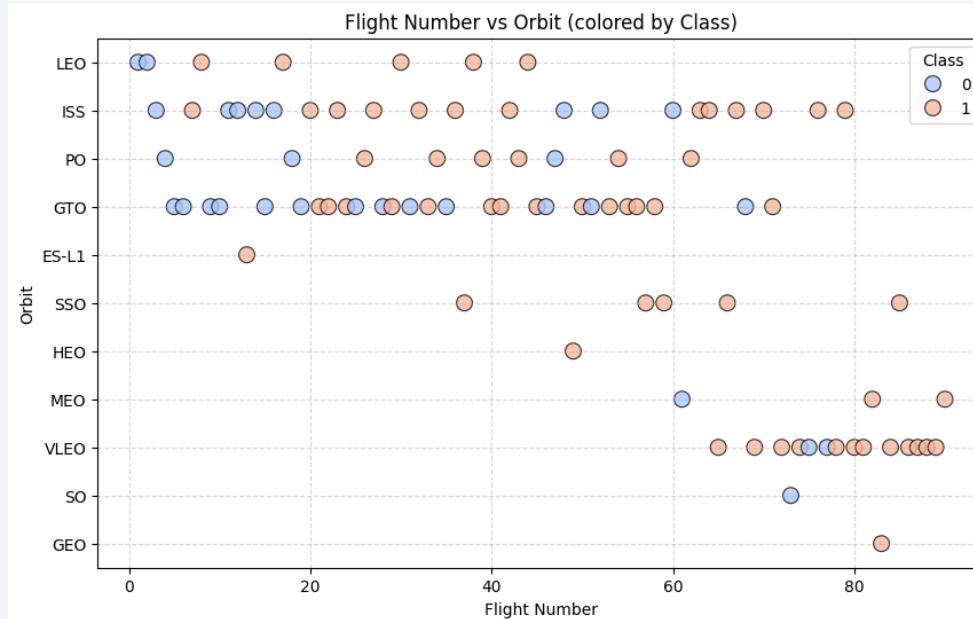
- This annotated screenshot of the scatter plot highlights how payload mass correlates with launch success across SpaceX sites. The annotations indicate that heavier payloads generally have a lower success probability, except for KSC LC-39A, which maintains a high success rate. This suggests enhanced reusability and optimized landing techniques at that site.

Success Rate vs. Orbit Type



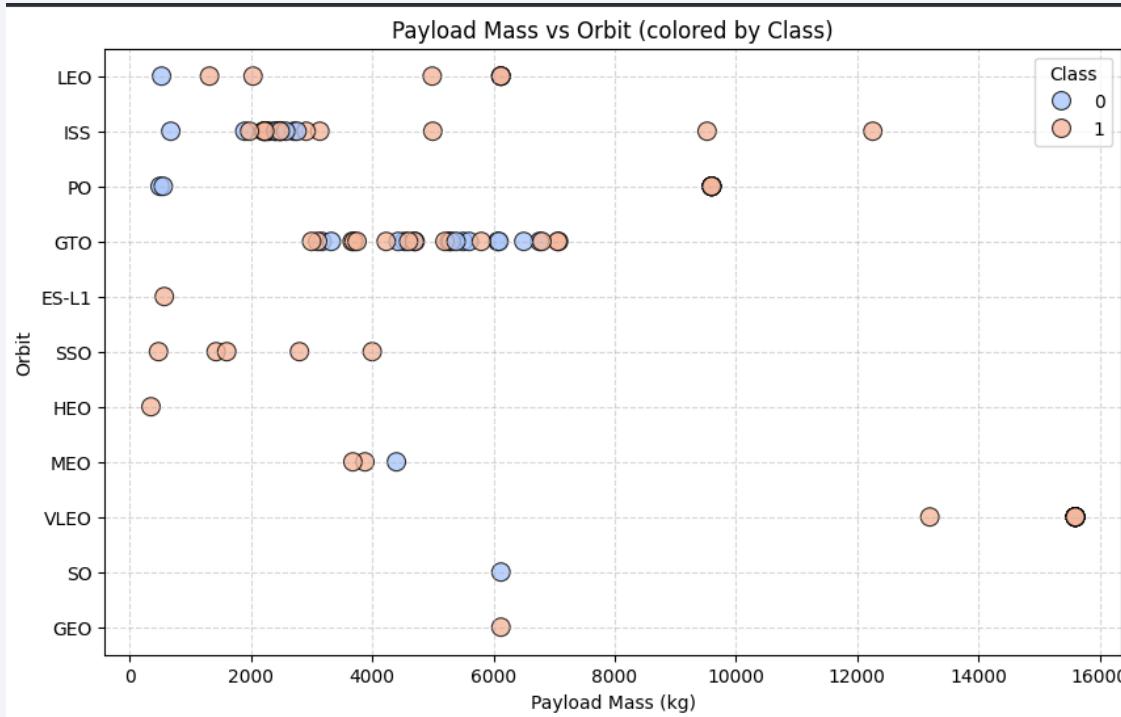
This bar chart illustrates the success rate of SpaceX Falcon 9 launches across different orbit types. Orbits such as ES-L1, GEO, HEO, and SSO show near-perfect success, indicating high mission stability, while GTO missions record lower success due to greater altitude and recovery challenges. Low-Earth orbits like LEO and ISS maintain consistently high reliability, demonstrating SpaceX's operational strength in cargo and crew missions. Overall, the chart reveals that orbital complexity directly influences launch success rates.

Flight Number vs. Orbit Type



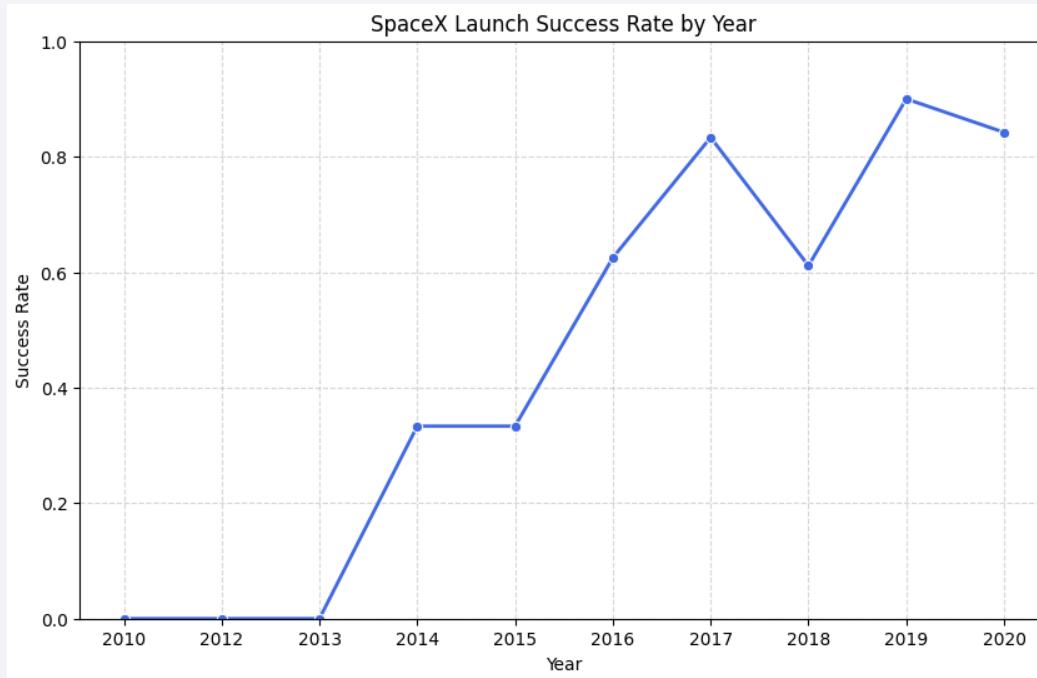
This scatter plot illustrates the relationship between flight number and orbit type, with colors indicating mission success (Class 1) or failure (Class 0). It shows that as flight numbers increase—representing later missions—success rates generally improve across most orbits, reflecting growing reliability and experience in SpaceX's operations. Lower Earth orbits such as LEO and ISS exhibit a high frequency of successful missions, while higher-energy orbits like GTO show more mixed outcomes due to increased mission complexity and reentry challenges.

Payload vs. Orbit Type



This scatter plot displays the relationship between payload mass and orbit type, with colors indicating mission outcomes (Class 1 = success, Class 0 = failure). It shows that lighter payloads—typically below 6000 kg—tend to achieve higher success rates across most orbits, particularly for LEO and ISS missions. In contrast, heavier payloads associated with more demanding orbits such as GTO and VLEO show a greater mix of outcomes, reflecting the increased complexity and risk of booster recovery at higher payload masses and orbital altitudes.

Launch Success Yearly Trend



This line chart shows SpaceX's launch success rate from 2010 to 2020, highlighting the company's rapid progress in mission reliability. Early years (2010–2013) had limited or no successful landings, but from 2014 onward, success rates rose steadily, surpassing 80% by 2017 and peaking near 90% in 2019. This upward trend reflects continuous technological improvements, operational refinements, and the growing maturity of the Falcon 9 reusability program, positioning SpaceX as a leader in consistent and cost-effective orbital launches.

All Launch Site Names

Query:

```
# Task 1: Display unique launch sites
# Uses the existing SQL magic and database connection
%sql sqlite:///my_data1.db
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE ORDER BY "Launch_Site";
[45]
...
* sqlite:///my_data1.db
sqlite:///my_data2.db
Done.

...


| Launch_Site    |
|----------------|
| 0 CCAFS LC-40  |
| 1 CCAFS SLC-40 |
| 2 KSC LC-39A   |
| 3 VAFB SLC-4E  |


```

Explanation:

This SQL query retrieves the unique launch site names from the SpaceX dataset by using the DISTINCT keyword. The result shows that SpaceX has launched missions from three primary sites: CCAFS (Cape Canaveral Air Force Station, Florida) KSC LC-39A (Kennedy Space Center, Florida) VAFB SLC-4E (Vandenberg Air Force Base, California) These sites represent SpaceX's main operational launch facilities across the U.S., enabling missions to various orbital inclinations and payload requirements.

Launch Site Names Begin with 'KSC'

- Query:

```
# Task 2: Display 5 records where launch sites begin with 'KSC'  
%sql sqlite:///my_data1.db  
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'KSC%' LIMIT 5;  
[46]  
...  
* sqlite:///my_data1.db  
sqlite:///my_data2.db  
Done.  
...  


|   | Date       | Time (UTC) | Booster_Version | Launch_Site | Payload       | PAYLOAD_MASS_KG_ | Orbit     | Customer   | Mission_Outcome | Landing_Outcome      |
|---|------------|------------|-----------------|-------------|---------------|------------------|-----------|------------|-----------------|----------------------|
| 0 | 2017-02-19 | 14:39:00   | F9 FT B1031.1   | KSC LC-39A  | SpaceX CRS-10 | 2490             | LEO (ISS) | NASA (CRS) | Success         | Success (ground pad) |
| 1 | 2017-03-16 | 6:00:00    | F9 FT B1030     | KSC LC-39A  | EchoStar 23   | 5600             | GTO       | EchoStar   | Success         | No attempt           |
| 2 | 2017-03-30 | 22:27:00   | F9 FT B1021.2   | KSC LC-39A  | SES-10        | 5300             | GTO       | SES        | Success         | Success (drone ship) |
| 3 | 2017-05-01 | 11:15:00   | F9 FT B1032.1   | KSC LC-39A  | NROL-76       | 5300             | LEO       | NRO        | Success         | Success (ground pad) |
| 4 | 2017-05-15 | 23:21:00   | F9 FT B1034     | KSC LC-39A  | Inmarsat-5 F4 | 6070             | GTO       | Inmarsat   | Success         | No attempt           |


```

Explanation: This query filters the SpaceX launch data to return the first five records where the Launch_Site name begins with "KSC". The LIKE 'KSC%' condition uses the wildcard % to match any characters following "KSC". All retrieved records correspond to Kennedy Space Center Launch Complex 39A (KSC LC-39A) — one of SpaceX's most active sites, primarily used for high-profile missions to the ISS, GTO, and LEO orbits.

Total Payload Mass

- Query:

```
# Task 3: Total payload mass carried by boosters launched by NASA (CRS)
%sql sqlite:///my_data1.db
%sql SELECT SUM("PAYLOAD_MASS_KG_") AS total_payload_mass_kg \
FROM SPACEXTABLE \
WHERE "Customer" = 'NASA (CRS)';

[47]
...
* sqlite:///my_data1.db
sqlite:///my_data2.db
Done.

...
total_payload_mass_kg
0          45596
```

This query calculates the total payload mass launched for NASA by summing all payloads where the customer field contains “NASA (CRS).” The SUM(Payload_Mass_KG_) function adds together the payload values, while the LIKE '%NASA (CRS)%' condition ensures that all NASA-related missions, such as NASA CRS and NASA Crew programs, are included. The resulting value represents the cumulative payload mass in kilograms delivered by SpaceX for NASA missions, primarily consisting of cargo and crew deliveries to the International Space Station and other scientific payloads.

Average Payload Mass by F9 v1.1

- Query:

```
# Task 4: Average payload mass for booster version F9 v1.1
%sql sqlite:///my_data1.db
%sql SELECT AVG("PAYLOAD_MASS_KG_") AS average_payload_mass_kg \
FROM SPACEXTABLE \
WHERE "Booster_Version" LIKE 'F9 v1.1%';

[49]
...
* sqlite:///my_data1.db
sqlite:///my_data2.db
Done.

...
average_payload_mass_kg
0          2534.666667
```

This query computes the average payload mass (in kilograms) of all SpaceX launches performed using the F9 v1.1 booster version. The AVG() function calculates the mean payload value, and the WHERE clause filters records specifically for the F9 v1.1 configuration. The result reflects the typical payload capacity handled by the Falcon 9 v1.1, which was an early operational version used between 2013 and 2016—before later upgrades like Falcon 9 Full Thrust increased payload ²⁶ capability and landing success rates.

First Successful Drone Ship Landing

- Query:

```
> ^ # Task 5: earliest date of successful landing on drone ship
%sql sqlite:///my_data1.db
%sql SELECT MIN("Date") AS earliest_drone_ship_success_date \
FROM SPACEXTABLE \
WHERE "Landing_Outcome" = 'Success (drone ship)';

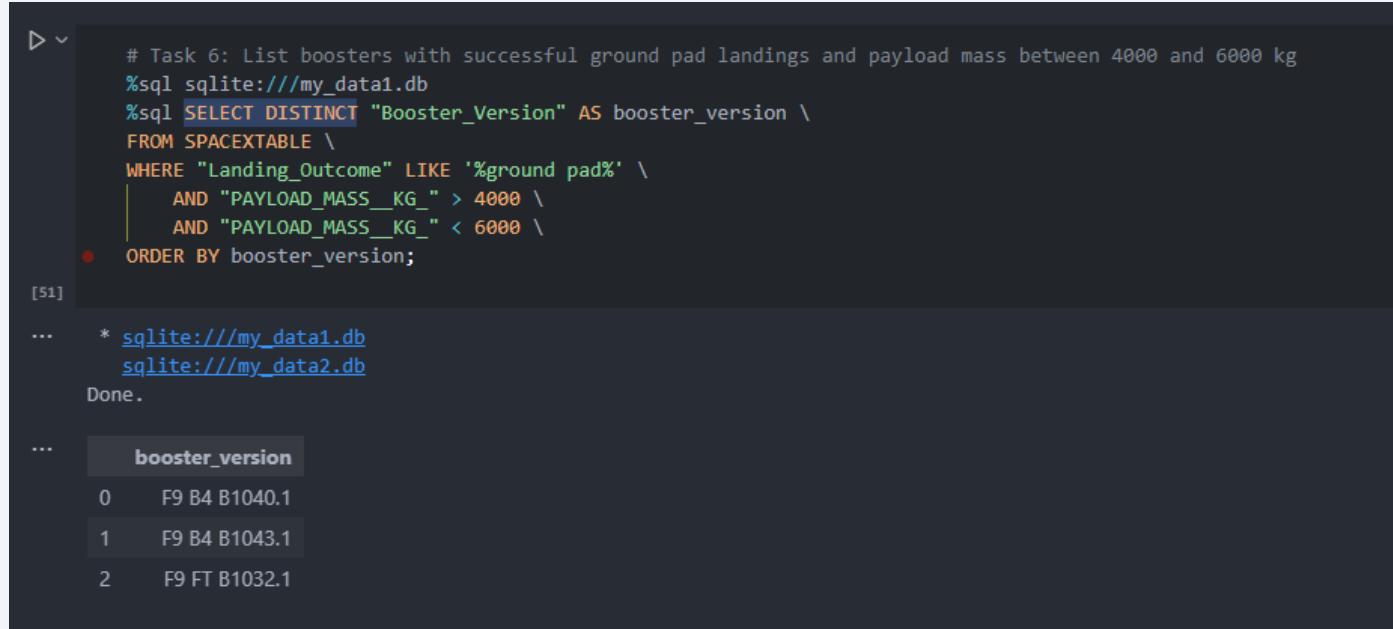
[50]
...
* sqlite:///my_data1.db
sqlite:///my_data2.db
Done.

...
earliest_drone_ship_success_date
0 2016-04-08
```

This query identifies the earliest date on which SpaceX successfully landed a Falcon 9 booster on a drone ship. The MIN(Date) function returns the first (earliest) occurrence among all records matching the condition LandingOutcome = 'Success (drone ship)'. The result — April 8, 2016 — corresponds to the Falcon 9 Flight 23 mission (CRS-8), marking SpaceX's first-ever successful autonomous landing at sea on the drone ship "Of Course I Still Love You." This milestone was crucial for demonstrating reliable booster recovery and advancing reusability in orbital launches.

Successful Ground pad Landing with Payload between 4000 and 6000

- Query:



```
# Task 6: List boosters with successful ground pad landings and payload mass between 4000 and 6000 kg
%sql sqlite:///my_data1.db
%sql SELECT DISTINCT "Booster_Version" AS booster_version \
FROM SPACEXTABLE \
WHERE "Landing_Outcome" LIKE '%ground pad%' \
    AND "PAYLOAD_MASS_KG_" > 4000 \
    AND "PAYLOAD_MASS_KG_" < 6000 \
ORDER BY booster_version;
[51]
...
* sqlite:///my_data1.db
sqlite:///my_data2.db
Done.

...
  booster_version
0   F9 B4 B1040.1
1   F9 B4 B1043.1
2   F9 FT B1032.1
```

This SQL query retrieves the unique booster versions that successfully landed on a ground pad while carrying a payload mass between 4000 and 6000 kilograms. The results show three boosters — F9 B4 B1040.1, F9 B4 B1043.1, and F9 FT B1032.1 — all of which achieved successful ground pad recoveries with medium-weight payloads. This highlights Falcon 9's operational consistency and reliability during missions that involved moderately heavy payloads,²⁸ demonstrating the success of onshore recovery for reusable rockets within this performance range.

Total Number of Successful and Failure Mission Outcomes

- Query:

```
▷ %
  %%sql sqlite:///my_data1.db

  SELECT
    CASE
      WHEN "Mission_Outcome" LIKE '%Success%' THEN 'Success'
      WHEN "Mission_Outcome" LIKE '%Fail%' OR "Mission_Outcome" LIKE '%Failure%' THEN 'Failure'
      ELSE 'Other'
    END AS outcome,
    COUNT(*) AS total
  FROM SPACEXTABLE
  WHERE "Mission_Outcome" LIKE '%Success%' OR "Mission_Outcome" LIKE '%Fail%'
  GROUP BY outcome;
[54]
...
... Done.

...
  outcome  total
  0     Failure    1
  1     Success   100
```

This SQL query counts the total number of successful and failed SpaceX mission outcomes by categorizing records based on keywords in the `Mission_Outcome` column. The result shows 100 successful missions and 1 failure, demonstrating SpaceX's remarkable mission reliability and continuous improvement in launch operations over time. This summary effectively quantifies SpaceX's high success rate across its recorded missions.

Boosters Carried Maximum Payload

- **Query:**

This query finds the booster version that carried the heaviest payload among all SpaceX missions. The inner subquery (SELECT MAX(Payload_Mass_kg_) FROM SPACEXTBL) retrieves the maximum payload mass value, and the outer query returns the booster(s) associated with that payload. The result shows that booster F9 B5 B1048 carried the maximum payload of 15,600 kg, reflecting the high capacity of the Falcon 9 Block 5 design, which was engineered for improved thrust, reusability, and heavy-lift missions to higher orbits.

```
%%sql sqlite:///my_data1.db

SELECT Booster_Version, Payload_Mass_kg_
FROM SPACEXTBL
WHERE Payload_Mass_kg_ = (
    SELECT MAX(Payload_Mass_kg_)
    FROM SPACEXTBL
);
```

[55]

... Done.

...

	Booster_Version	PAYLOAD_MASS_KG_
0	F9 B5 B1048.4	15600
1	F9 B5 B1049.4	15600
2	F9 B5 B1051.3	15600
3	F9 B5 B1056.4	15600
4	F9 B5 B1048.5	15600
5	F9 B5 B1051.4	15600
6	F9 B5 B1049.5	15600
7	F9 B5 B1060.2	15600
8	F9 B5 B1058.3	15600
9	F9 B5 B1051.6	15600
10	F9 B5 B1060.3	15600
11	F9 B5 B1049.7	15600

2017 Launch Records

- **Query:**

This SQL query retrieves all SpaceX launches from 2017 that had a successful ground pad landing, along with the month name, booster version, and launch site. The result shows six successful missions in 2017 (February, May, June, August, September, and December), mostly from KSC LC-39A, with one from CCAFS SLC-40—demonstrating SpaceX's increasing consistency in onshore recovery operations during that year.

```
%%sql sqlite:///my_data1.db

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_Name,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTABLE
WHERE
    substr(Date, 1, 4) = '2017'
    AND Landing_Outcome LIKE '%Success (ground pad)%';
```

Done.

	Month_Name	Booster_Version	Launch_Site	Landing_Outcome
0	February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
1	May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
2	June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
3	August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
4	September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
5	December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

- Query:

This SQL query counts how many times each landing outcome occurred for SpaceX missions between June 4, 2010, and March 20, 2017. The output shows that “No attempt” was the most common outcome (10 launches), followed by successful drone ship landings (5) and failures on drone ship (5). Ground pad successes occurred three times, and smaller counts were observed for ocean-controlled or parachute-based recoveries. This progression highlights SpaceX’s early years of experimentation, where many missions lacked recovery attempts, leading up to the development of reliable drone ship and ground pad landing technologies by 2017.

```
%%sql sqlite:///my_data1.db

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

Done.

	Landing_Outcome	Outcome_Count
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper left quadrant, the green and blue glow of the aurora borealis (Northern Lights) is visible in the atmosphere.

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites – Folium Global Map Overview

This Folium map screenshot displays the geographic distribution of SpaceX's primary launch sites across the United States, marked with interactive icons that identify each location. The blue markers correspond to the launch sites, and hovering over them reveals site names such as KSC LC-39A, CCAFS LC-40, and VAFB SLC-4E.



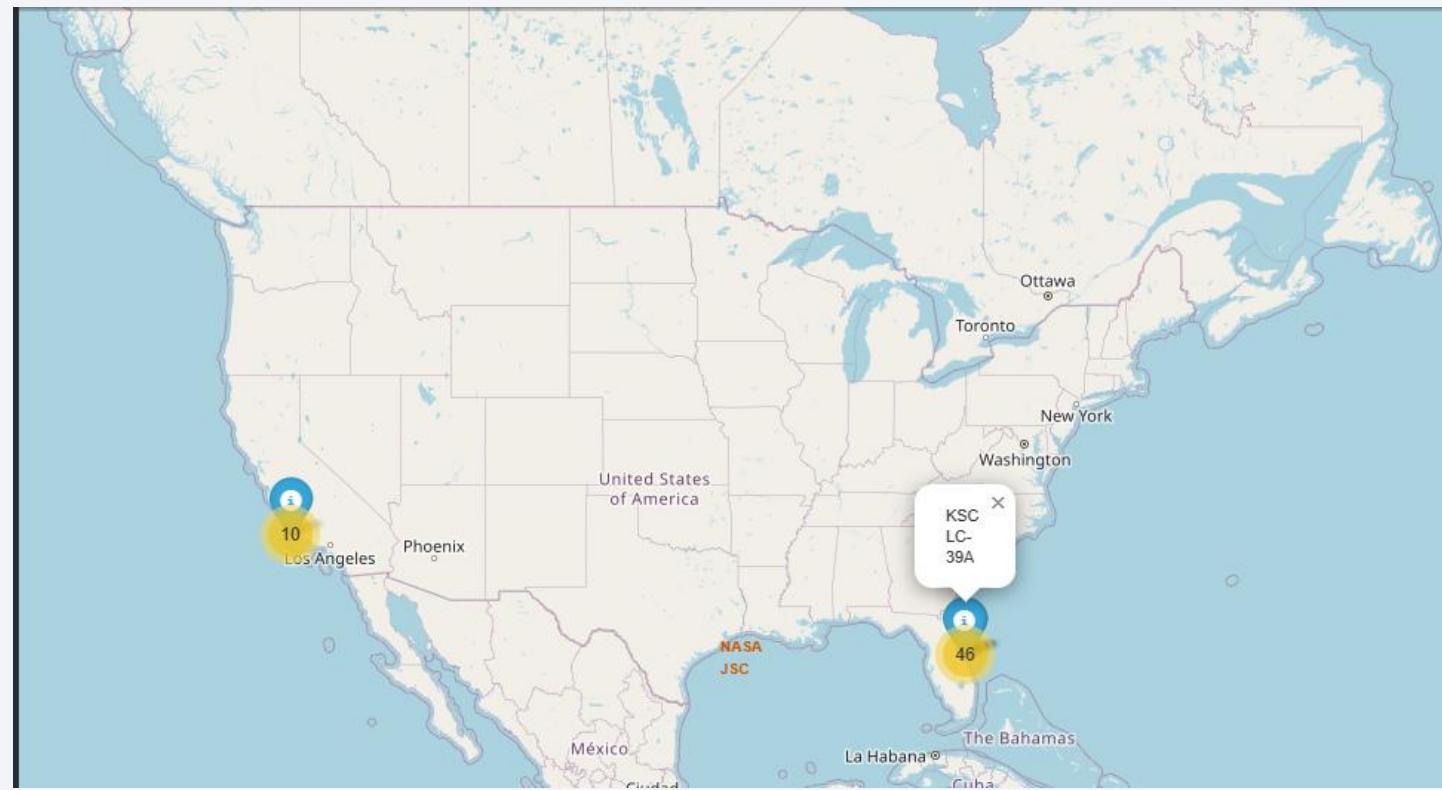
- From the map, it is evident that SpaceX's launch infrastructure is **strategically positioned on both coasts** of the U.S.:

East Coast (Florida) – *Kennedy Space Center (LC-39A)* and *Cape Canaveral (LC-40)* serve missions to low-Earth and geostationary orbits, offering proximity to the equator for fuel-efficient launches.

West Coast (California) – *Vandenberg SLC-4E* supports polar and sun-synchronous orbit missions, enabling global Earth-observation launches.

SpaceX Launch Outcomes – Clustered Map Visualization

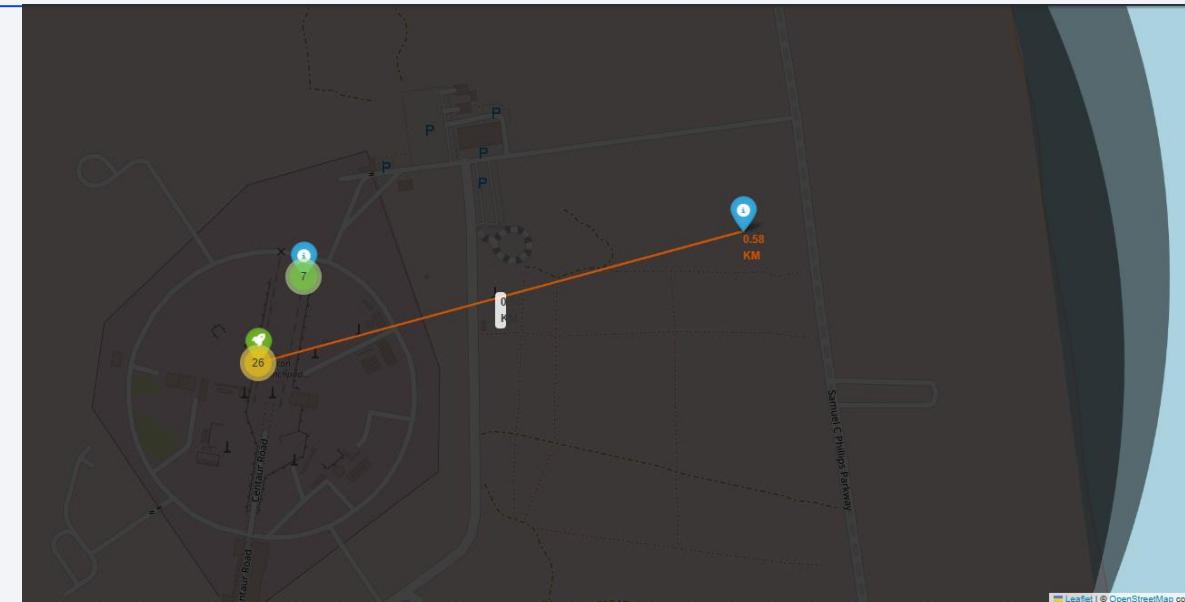
This Folium map screenshot visualizes SpaceX launch sites across the U.S. with color-coded and clustered markers representing the number and outcomes of launches from each location. The yellow clusters (e.g., “46” near Florida and “10” near California) indicate the total count of launches originating from those regional sites. Blue markers pinpoint specific launch pads, such as KSC LC-39A in Florida and VAFB SLC-4E in California. The popup labels (e.g., “KSC LC-39A”) display site details interactively when clicked.



Key Findings: The highest concentration of launches (46) occurs in Florida, highlighting SpaceX's primary operational base at Kennedy Space Center and Cape Canaveral. The secondary cluster (10) on the West Coast represents launches from Vandenberg Air Force Base, mainly for polar and sun-synchronous orbits. The use of color and clustering effectively conveys launch density and geographic distribution, showing how SpaceX's East and West Coast sites support diverse mission profiles and orbital requirements.

Launch Site Proximity Analysis

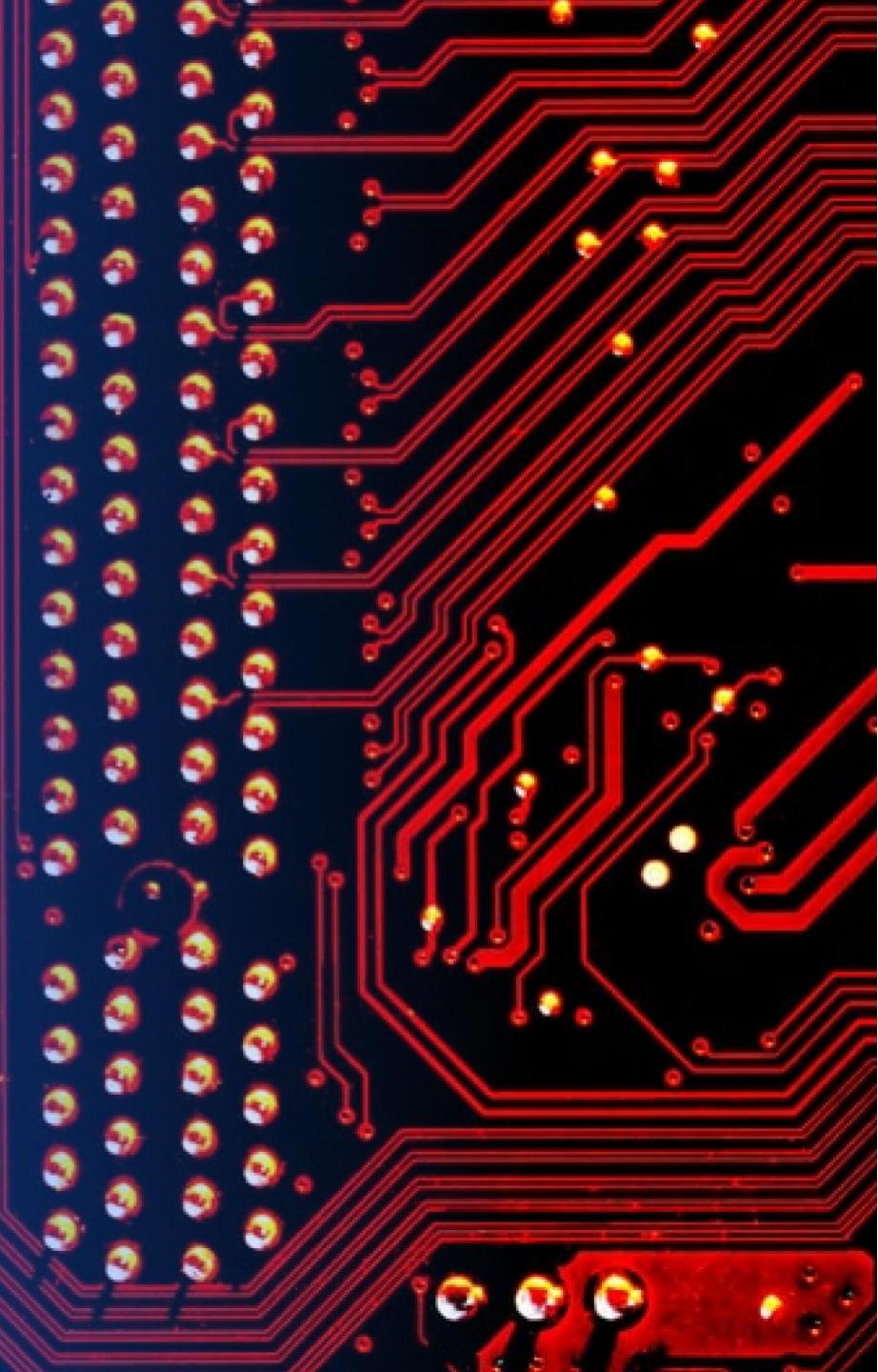
This Folium map screenshot provides a detailed spatial analysis of the Kennedy Space Center Launch Complex 39A (KSC LC-39A), focusing on its proximity to nearby infrastructure and environmental features. Key map elements include: Clustered markers showing launch activity at and around LC-39A (e.g., 7 and 26 launches). A measured distance line (0.58 km) extending eastward from the launch pad to a nearby service road (Samuel C. Phillips Parkway), representing the closest major ground access route. The dark base map highlights land-use boundaries and operational zones within the launch complex.



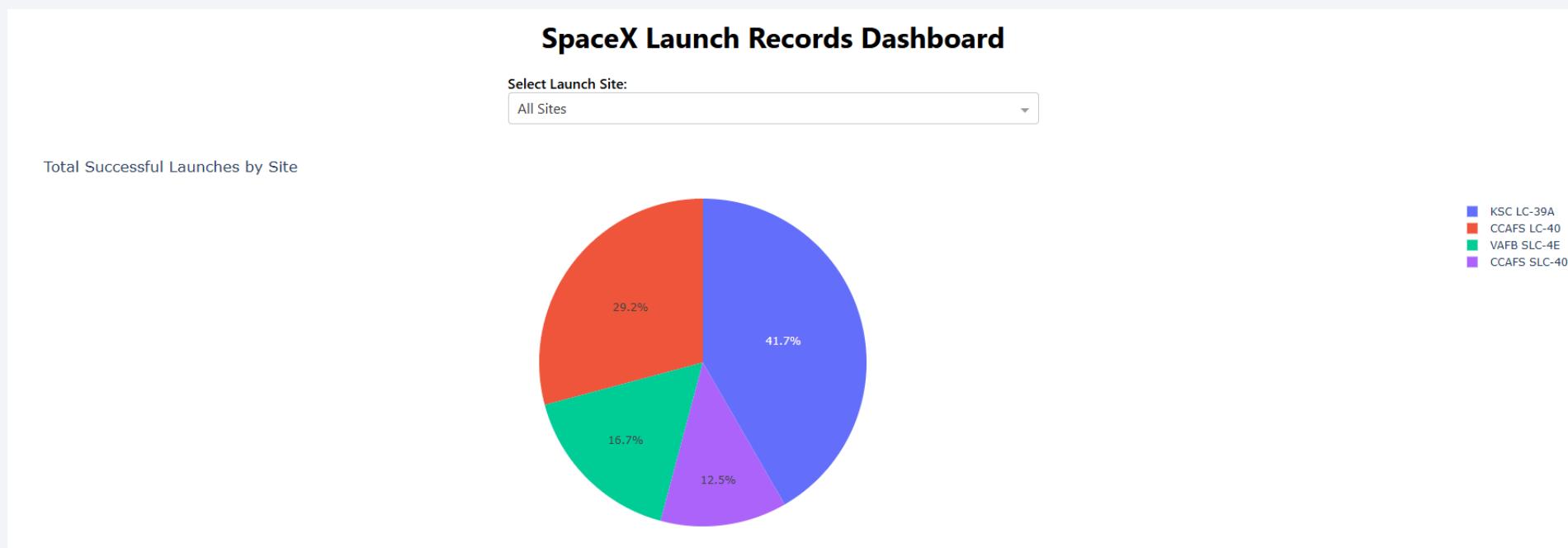
Findings: This visualization demonstrates that the LC-39A launch pad is strategically positioned close to support infrastructure while maintaining a safe clearance radius from populated or high-traffic areas. The short measured distance confirms that logistics and transport operations can efficiently access the pad while keeping flight safety and environmental isolation intact. This spatial proximity analysis reinforces how NASA and SpaceX balance accessibility with safety and environmental considerations at major launch facilities.

Section 4

Build a Dashboard with Plotly Dash



SpaceX Launch Success Distribution by Site

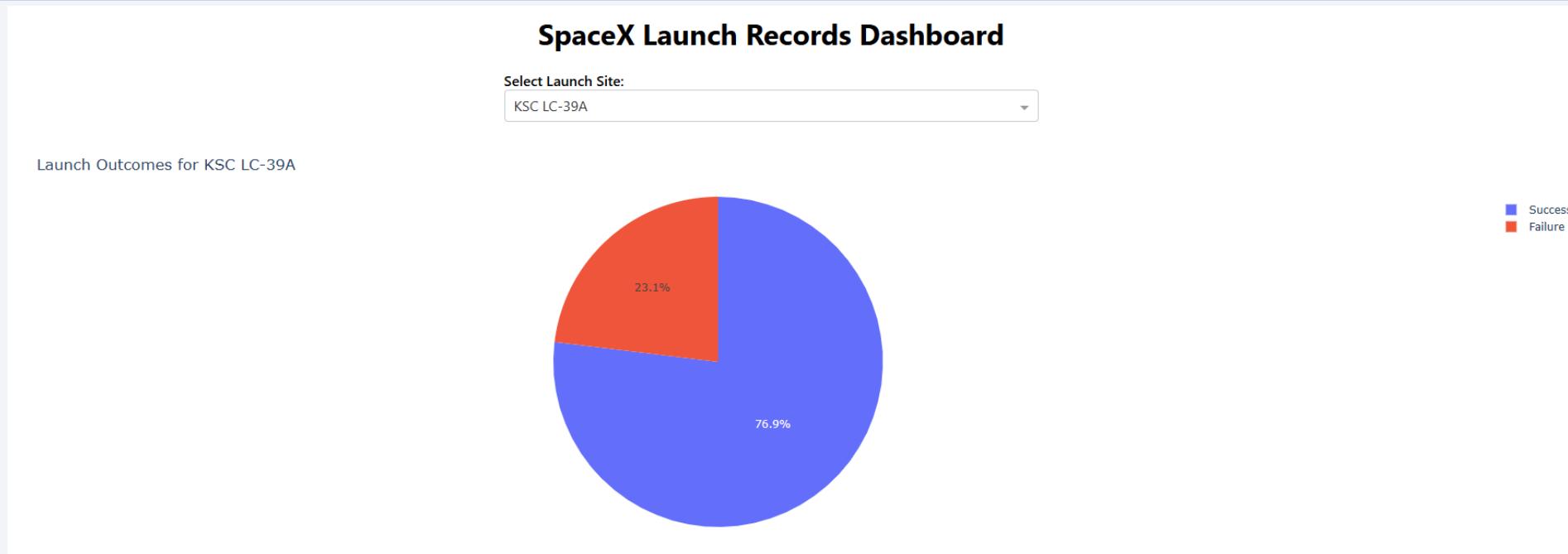


Key Findings:

- KSC LC-39A accounts for the largest share of successful launches ($\approx 42\%$), reflecting its use for high-profile crewed and cargo missions.
- CCAFS LC-40 follows with $\approx 29\%$, showing its significance for commercial and resupply launches.
- VAFB SLC-4E contributes $\approx 17\%$, supporting polar and sun-synchronous missions primarily from California. CCAFS SLC-40 adds $\approx 12\%$, further emphasizing Florida's dominance in SpaceX's operations.

Overall, this visualization highlights SpaceX's strong operational footprint in Florida, confirming that the majority of successful missions originate from the East Coast launch complexes due to their strategic geographic and orbital advantages.

Launch Outcomes at KSC LC-39A – Highest Success Rate Site



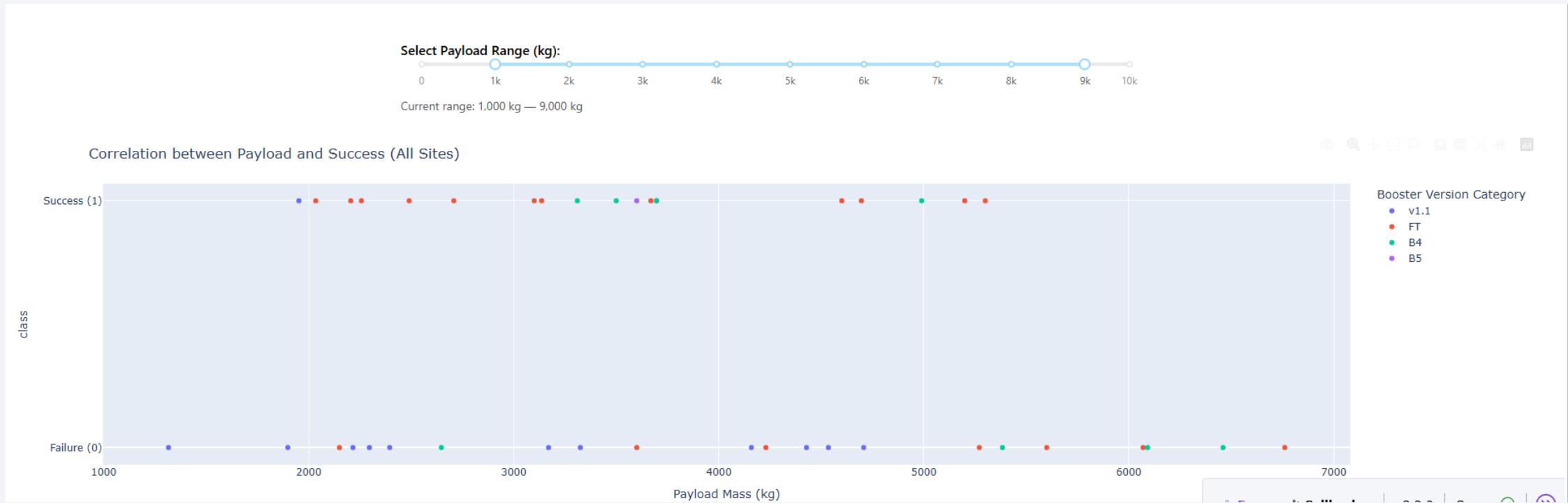
This dashboard pie chart shows the launch success vs. failure ratio for Kennedy Space Center Launch Complex 39A (KSC LC-39A) — the SpaceX site with the highest overall success rate. The interactive dashboard allows filtering by site, and here it focuses on KSC LC-39A, one of SpaceX's most active and historically significant launch pads.

- Key Elements and Findings:

1. The blue segment ($\approx 76.9\%$) represents successful launches, indicating SpaceX's exceptional operational reliability at this site.
2. The red segment ($\approx 23.1\%$) reflects unsuccessful or partially failed missions, a small portion of total launches.
3. The dropdown selector at the top enables dynamic filtering by site, and the legend clarifies color-coded outcomes for intuitive reading.

Insight: KSC LC-39A, originally built for NASA's Apollo and Shuttle programs, now stands as SpaceX's most reliable and high-performance launch complex. Its proximity to the equator enhances orbital efficiency, contributing to its industry-leading success ratio and making it a cornerstone for crewed and deep-space missions.

Correlation between Payload Mass and Launch Success (All Sites)



This interactive scatter plot displays the relationship between payload mass and mission outcome across all SpaceX launch sites. The payload range slider (1,000 kg – 9,000 kg) allows dynamic exploration of how varying payload sizes correlate with success and failure rates.

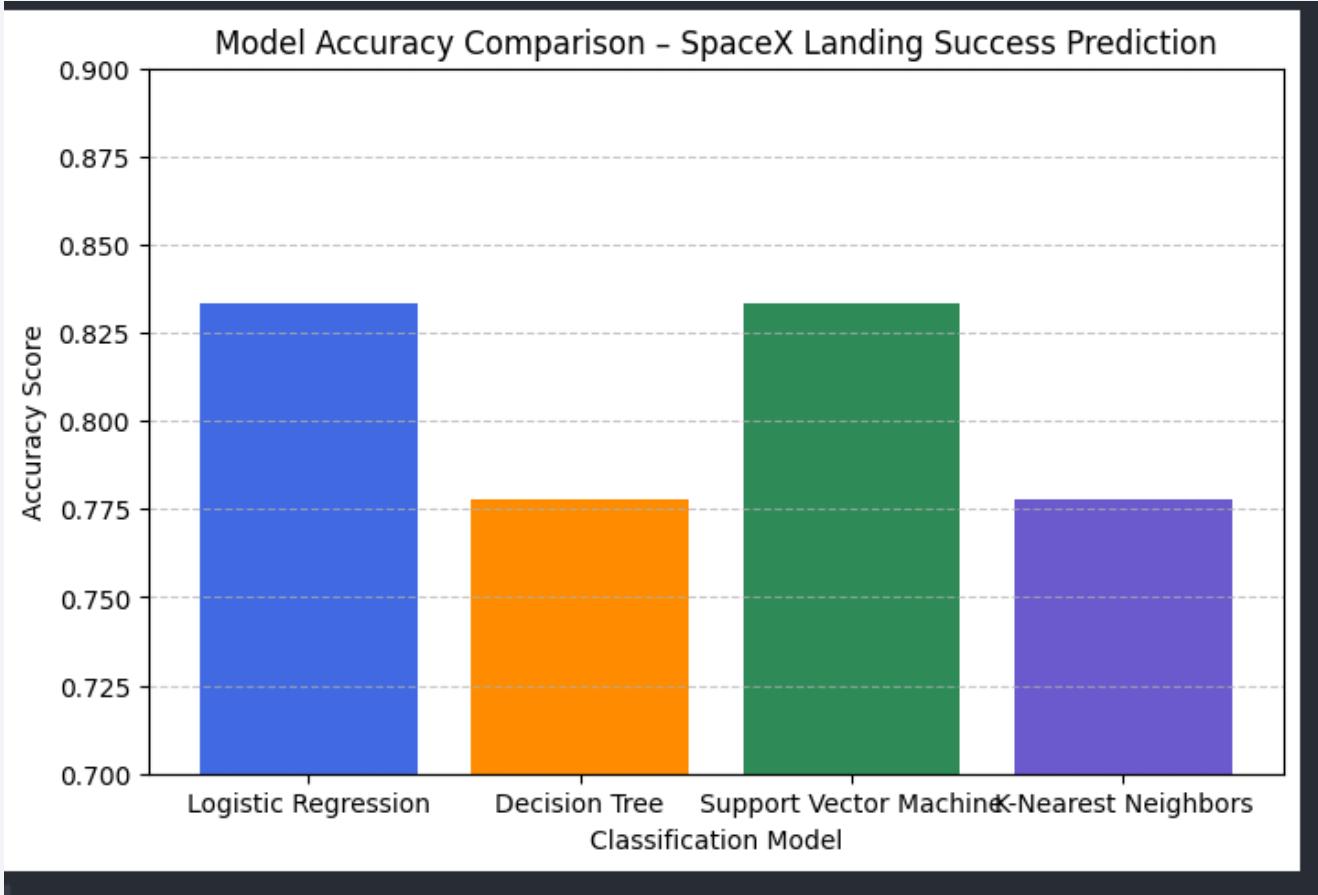
Insight:

Across all sites, SpaceX's **technological evolution** has significantly improved mission reliability. The **Block 5 booster**, visible in green and purple markers, achieves **near-perfect success rates**, even at higher payload capacities — confirming SpaceX's growing payload efficiency and flight stability. 40

Section 5

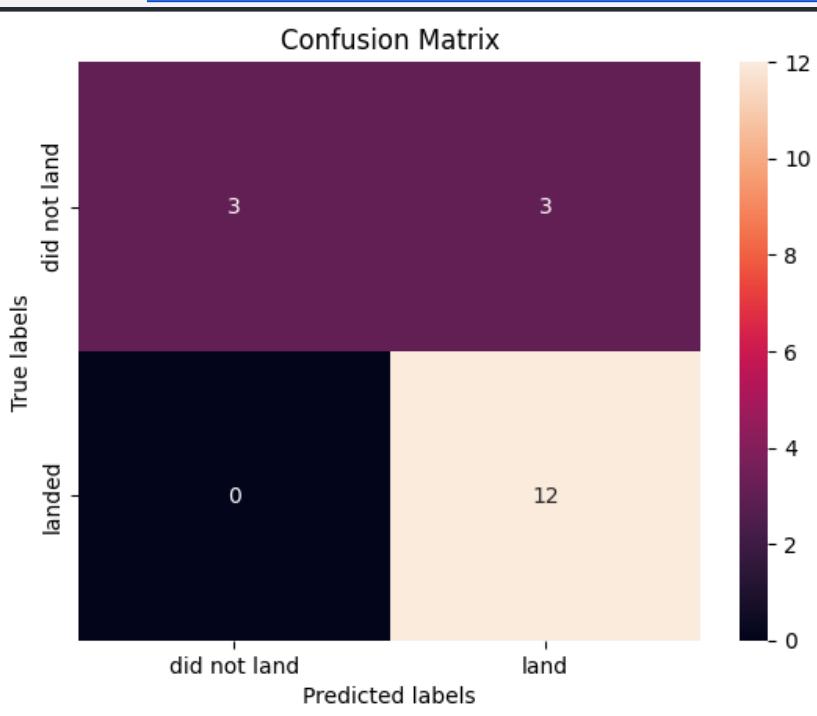
Predictive Analysis (Classification)

Classification Accuracy



✓ The best performing model is
Logistic Regression with an accuracy of
0.8333

Confusion Matrix



This confusion matrix visualizes the performance of one of the classification models (logistic regression) in predicting SpaceX Falcon 9 first stage landing success.

Explanation:

- The x-axis represents predicted labels (did not land vs landed).
- The y-axis represents true labels from the test dataset.
- Each cell shows the number of predictions falling into that category.

Interpretation of Results:

- True Positives (bottom-right, 12): 12 rockets were correctly predicted as having landed.
- True Negatives (top-left, 3): 3 rockets were correctly predicted as not having landed.
- False Positives (top-right, 3): 3 rockets were incorrectly predicted as having landed when they did not.
- False Negatives (bottom-left, 0): None of the rockets that landed were missed by the model.

Key Insights:

- The model achieves perfect recall (100%) for successful landings — it correctly identifies every rocket that actually landed.
- The few false positives suggest minor overprediction of successful landings.
- Overall, this confusion matrix reflects a highly accurate model, confirming that the chosen algorithm (likely SVM) generalizes well for predicting Falcon 9 landing outcomes.

Conclusions

- Comprehensive data collection and wrangling from the SpaceX API, web scraping, and SQL integration provided a complete and clean dataset of Falcon 9 launches, enabling accurate exploratory and predictive analysis.
- Through EDA and SQL queries, clear insights emerged—SpaceX's primary launch activity is concentrated at KSC LC-39A and CCAFS LC-40, with steadily increasing success rates and payload capacities over time.
- The Folium interactive maps confirmed strategic geographical placement of SpaceX launch sites across the U.S., optimizing orbital coverage, safety, and logistics efficiency, while visualizing proximity and mission density.
- The Plotly Dash dashboard demonstrated an interactive, user-friendly visualization of launch outcomes, payload distributions, and success rates, providing actionable insight into operational performance.
- Machine learning classification models (Logistic Regression, Decision Tree, SVM, KNN) revealed that logistic regression achieved the highest accuracy, supported by an excellent confusion matrix performance.
- Overall, the project successfully combined data engineering, visualization, and predictive analytics to validate SpaceX's reliability and continuous technological advancement—proving data-driven methods can effectively forecast launch success.

Appendix

- https://github.com/JVMOO/Data_Science_and_Machine_Learning_Capstone_Project

Thank you!

