IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<NATTWUT ONTO>
<8/5/2025>

Outline

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

Executive Summary

- Analyzed SpaceX launch records to predict mission success.
- Data collected via SpaceX REST API and web scraping from Wikipedia.
- Performed exploratory data analysis using SQL and visualizations.
- Built an interactive dashboard with Plotly Dash and geospatial maps with Folium.
- Applied classification models (Decision Tree and KNN) to predict mission outcomes. The Decision Tree model performed best.

Introduction

- SpaceX is pioneering commercial space travel by emphasizing rocket reusability.
- Our project aims to analyze historical SpaceX launch data to determine key factors behind successful missions.
- The main question: Can we predict whether a mission will succeed based on historical data?

Section 1

Methodology

Methodology

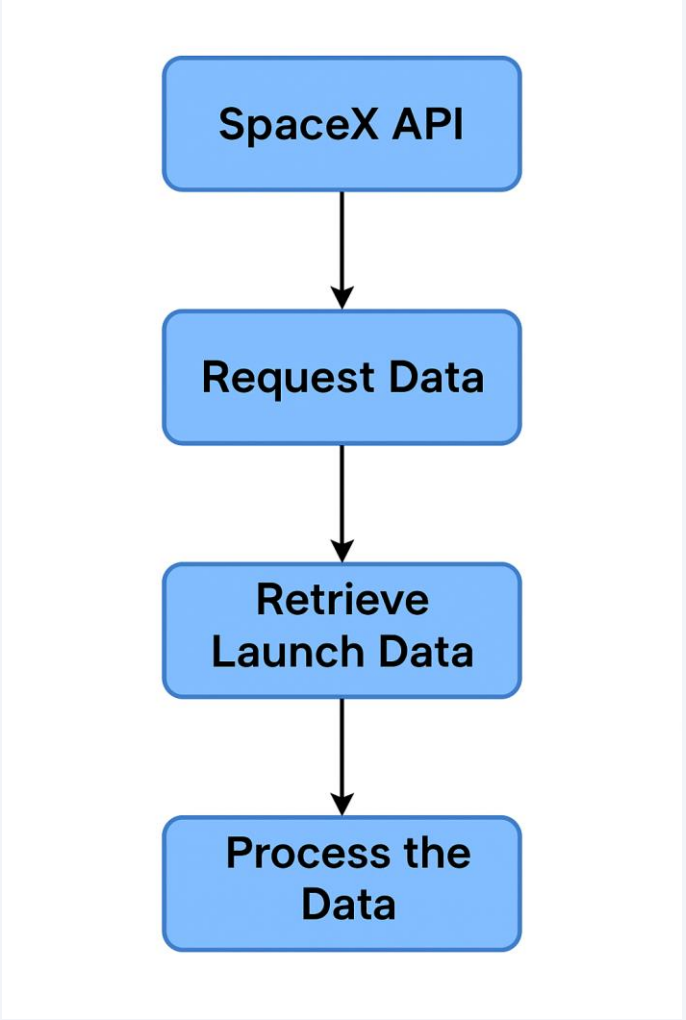
•**Data Collection**: Launch records were retrieved using the SpaceX REST API, and supplementary launch metadata was scraped from Wikipedia.
•**Data Wrangling**: The collected data was cleaned, merged, and transformed to ensure consistency across fields like payload mass, launch outcome, and site name.
•**Exploratory Data Analysis (EDA)**:
•Visualizations (scatter plots, bar charts) were used to uncover trends and correlations.
•SQL queries were used to extract statistical summaries and answer business-related questions.
•**Interactive Analytics**:
•**Folium Maps** were created to visualize launch site geography and proximity to infrastructure.
•**Plotly Dash Dashboard** enabled interactive exploration of launch outcomes, payloads, and booster types.
•**Predictive Modeling**:
•Applied supervised machine learning (Decision Tree, KNN) to predict launch success.
•Models were trained and optimized using GridSearchCV with 10-fold cross-validation.
•Final evaluation used accuracy scores and confusion matrices.

Data Collection

- Data was collected from the SpaceX REST API to retrieve structured launch information.
- Additional data such as launch outcomes and booster information were scraped from Wikipedia.
- All sources were consolidated into a Pandas dataframe for further analysis and modeling.

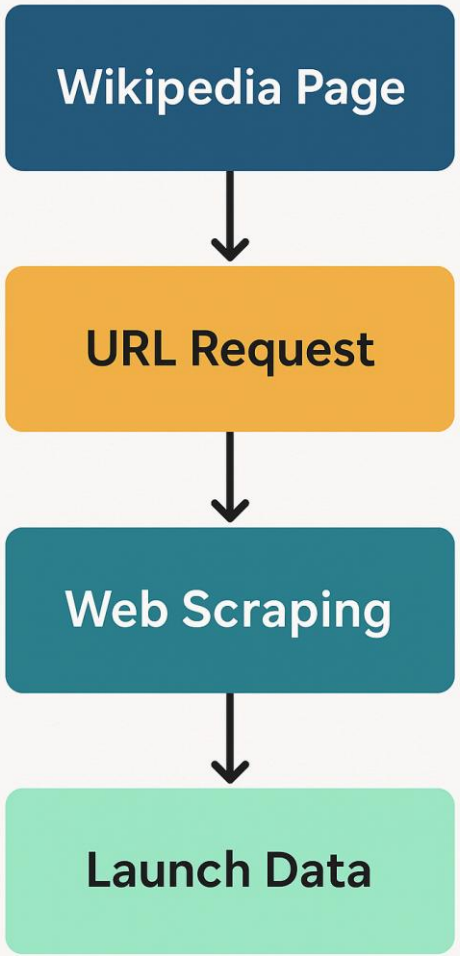
Data Collection – SpaceX API

- The SpaceX API was used to collect structured data on historical launches, including fields such as launch date, payload mass, orbit type, booster version, and mission outcome.
- The data retrieval process involved sending a GET request to the /v4/launches/past endpoint of the API. This returned a JSON object with details on every completed SpaceX launch.
- The JSON response was loaded into a Pandas DataFrame for further processing. This allowed for easy manipulation, cleaning, and integration with other data sources.
- The use of a direct API ensured that we worked with authoritative and up-to-date data, removing the need for manual entry or data cleaning from unstructured sources.
- The API workflow followed these main steps:
  - Access SpaceX
  - APIRequest data using HTTP GET
  - Retrieve and load launch records
  - Convert and preprocess the data using Pandas
- https://github.com/Doldaysd/IBM-DS_CAPSTONE_WINNING_SPACERACE_WITH_DS/blob/main/jupyter-labs-spacex-data-collection-api-bak-2025-08-04-21-00-27Z.ipynb



Data Collection - Scraping

- In addition to the SpaceX API, web scraping was used to collect supplementary launch data from a relevant Wikipedia page.
- The scraping process involved sending an HTTP request to the URL and retrieving the HTML content of the launch tables.
- The tables were parsed using BeautifulSoup and pandas.read_html(), which allowed for automated extraction of structured tables into DataFrames.
- Fields such as launch site, landing outcome, payload description, and booster version were extracted and merged with API data.
- This approach enabled the inclusion of additional fields not provided by the API, especially those concerning landing status and detailed mission descriptions.
- Scraped data was cleaned and normalized to ensure consistency with the API dataset, including renaming columns and converting date formats.
- Final output: a complete and unified dataset ready for wrangling and analysis.
- https://github.com/Doldaysd/IBM-DS_CAPSTONE_WINNING_SPACERACE_WITH_DS/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Removed missing values, corrected inconsistent formats, and renamed columns for clarity.
- Created new features such as Launch Outcome (binary) and normalized payload fields.
- Joined data from different sources (API + scraped content) into a single dataset.

EDA with Data Visualization

- Visualized relationships between **payload mass**, **orbit type**, and **launch outcome** using scatter plots and bar charts.
- Analyzed how **flight number**, **orbit type**, and **launch site** affected mission success rates.
- Used scatter plots to observe trends such as:
  - Higher payloads are often associated with more advanced booster versions.
  - Certain orbit types (e.g., GTO) have lower success rates compared to LEO.
- Line chart used to illustrate the **yearly trend** of launch success, showing improvements over time.
- Plots helped guide feature selection for machine learning by highlighting key predictors of launch success.

EDA with SQL

- Queried for:
  - Total number of successful vs. failed missions.
  - Average payload mass by booster version (F9 v1.1).
  - Dates of first successful ground landings.
  - Launches on drone ships with payloads between 4000–6000 kg.
  - All launch sites starting with 'CCA'.
  - Landing outcomes between 2010–2017, ranked by frequency.
  - Used SQLite to perform relational queries for deep insight.

Build an Interactive Map with Folium

- A Dropdown menu allows filtering data by launch site.
- A Pie chart displays launch success rate per site (or overall if "All Sites" is selected).
- A Range slider adjusts payload mass range dynamically.
- A Scatter plot shows correlation between payload mass and mission outcome across different booster versions.

Build a Dashboard with Plotly Dash

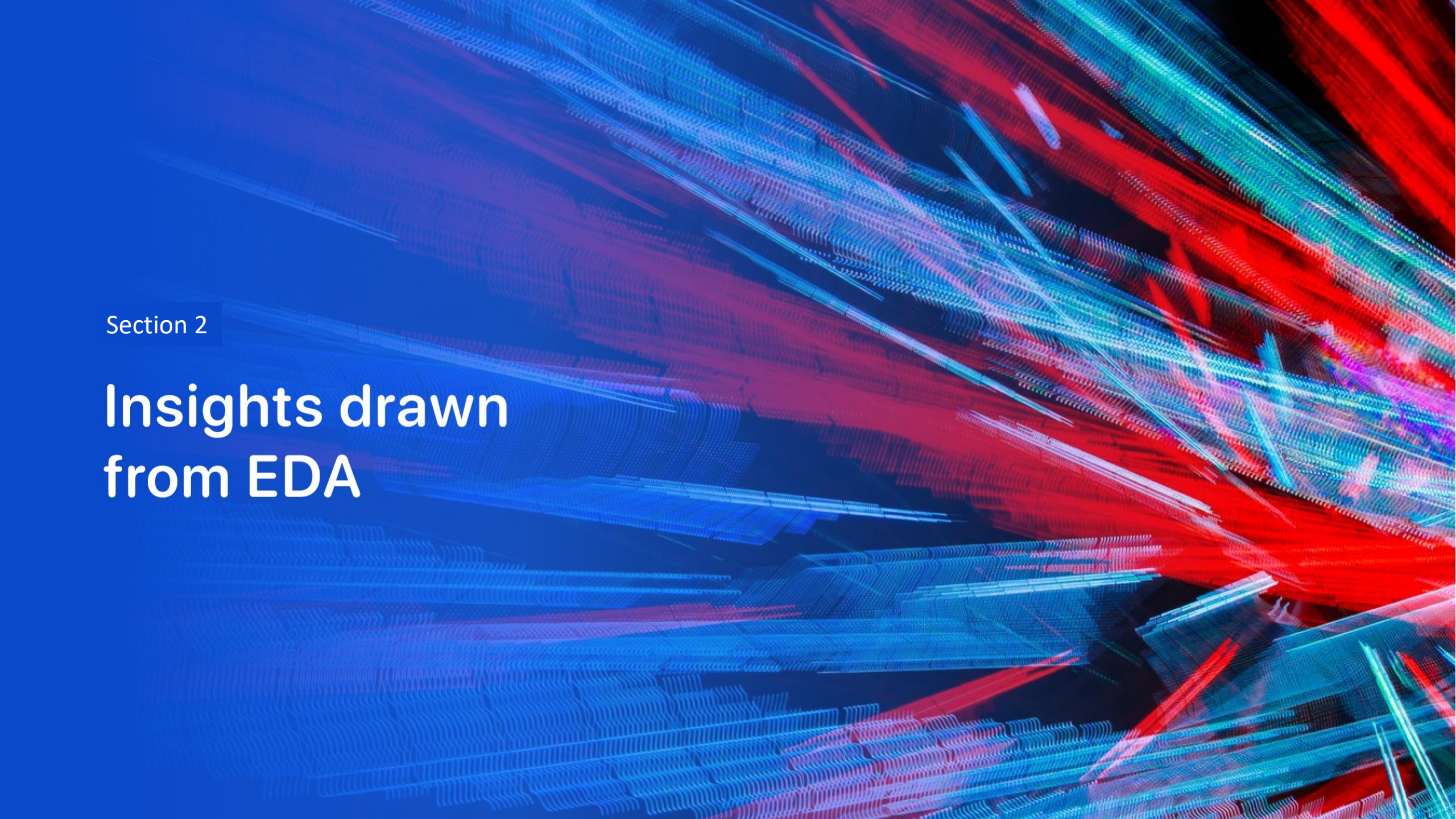
- A Dropdown menu allows filtering data by launch site.
- A Pie chart displays launch success rate per site (or overall if "All Sites" is selected).
- A Range slider adjusts payload mass range dynamically.
- A Scatter plot shows correlation between payload mass and mission outcome across different booster versions.

Predictive Analysis (Classification)

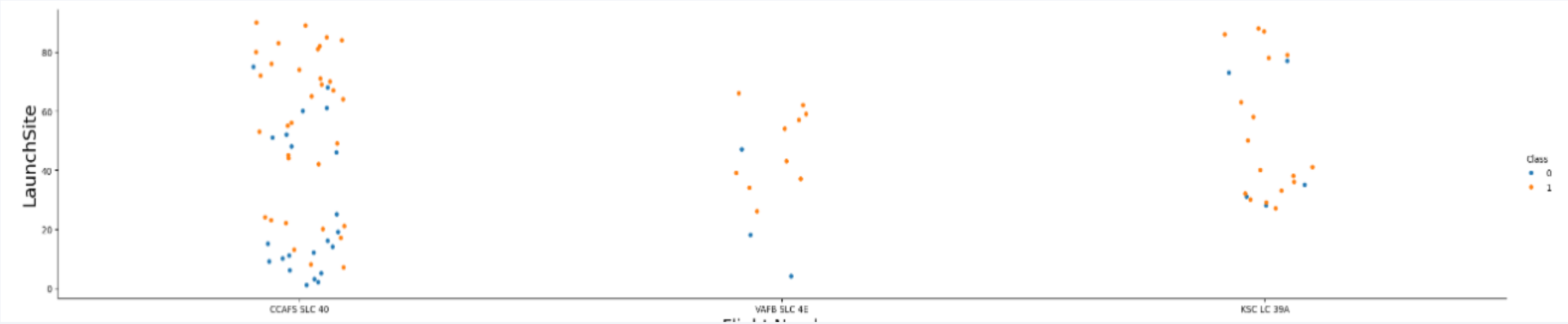
- Built and compared Decision Tree and K-Nearest Neighbors (KNN) models.
- Used GridSearchCV with 10-fold cross-validation to optimize hyperparameters.
- Best performing model: Decision Tree, tuned with:
  - criterion='entropy', max_depth=4, splitter='best', min_samples_split=2, min_samples_leaf=1, max_features='sqrt'
- Model evaluated using accuracy and confusion matrix.

Results

- Flight Number vs. Launch Site: Revealed consistent increase in successful launches across all sites, especially at CCAFS and KSC.
- Payload vs. Launch Site: Heavier payloads were concentrated at specific sites like VAFB, hinting at specialized missions.
- Success Rate by Orbit Type: Low Earth Orbit (LEO) had the highest success rate, while GTO showed more frequent failures.
- Flight Number vs. Orbit Type: Certain orbits became more common over time, reflecting SpaceX's evolving mission focus.
- Payload vs. Orbit Type: GTO missions tended to carry heavier payloads, often correlating with greater risk.
- Yearly Launch Success Trend: Demonstrated steady improvement, indicating SpaceX's growing reliability over time.
- SQL Analysis Results:
    - Identified booster versions with highest payloads and success rates.
    - Found earliest successful landings and key patterns in failed drone ship recoveries.
    - Calculated mission counts, payload averages, and ranked landing outcomes.


Section 2
Insights drawn from EDA

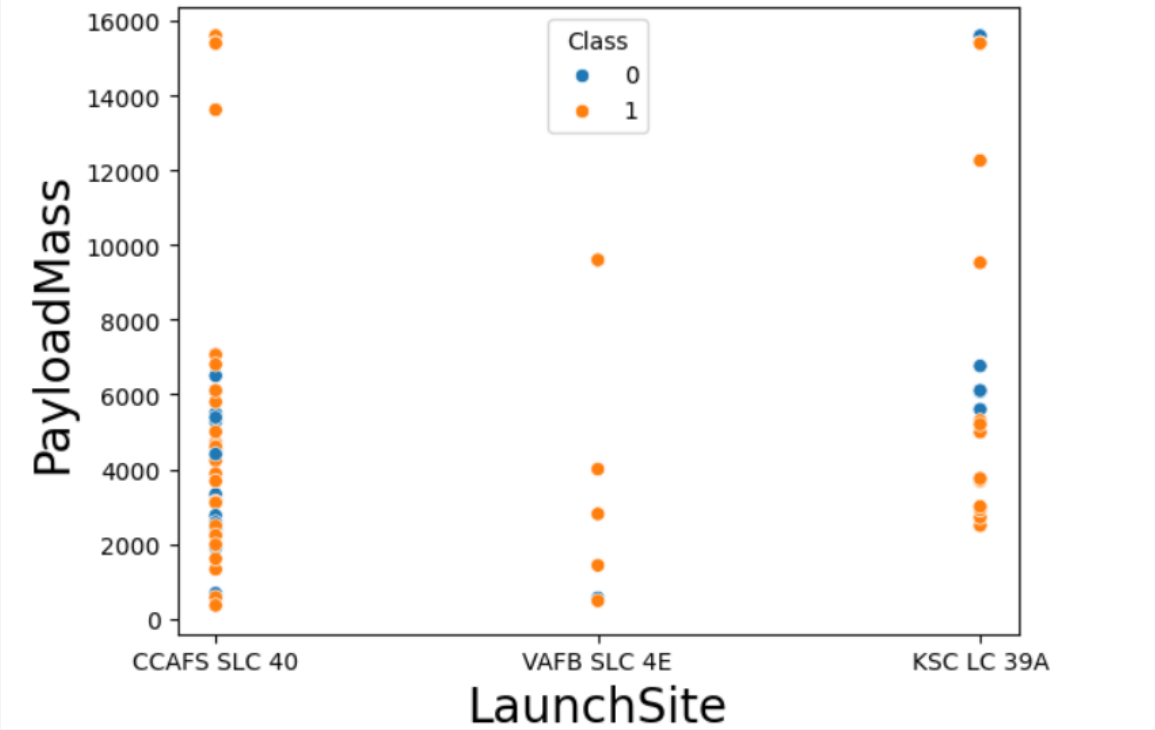
Flight Number vs. Launch Site



- Orange dots (1) indicate successful launches, while blue dots (0) indicate failures.
- CCAFS SLC 40 has the highest number of launches, with a clear trend toward increasing success in later missions.
- VAFB SLC 4E has fewer launches, but a notable portion are successful.
- KSC LC 39A shows mostly successful launches, often in more recent flight numbers, reflecting its role in high-profile missions.

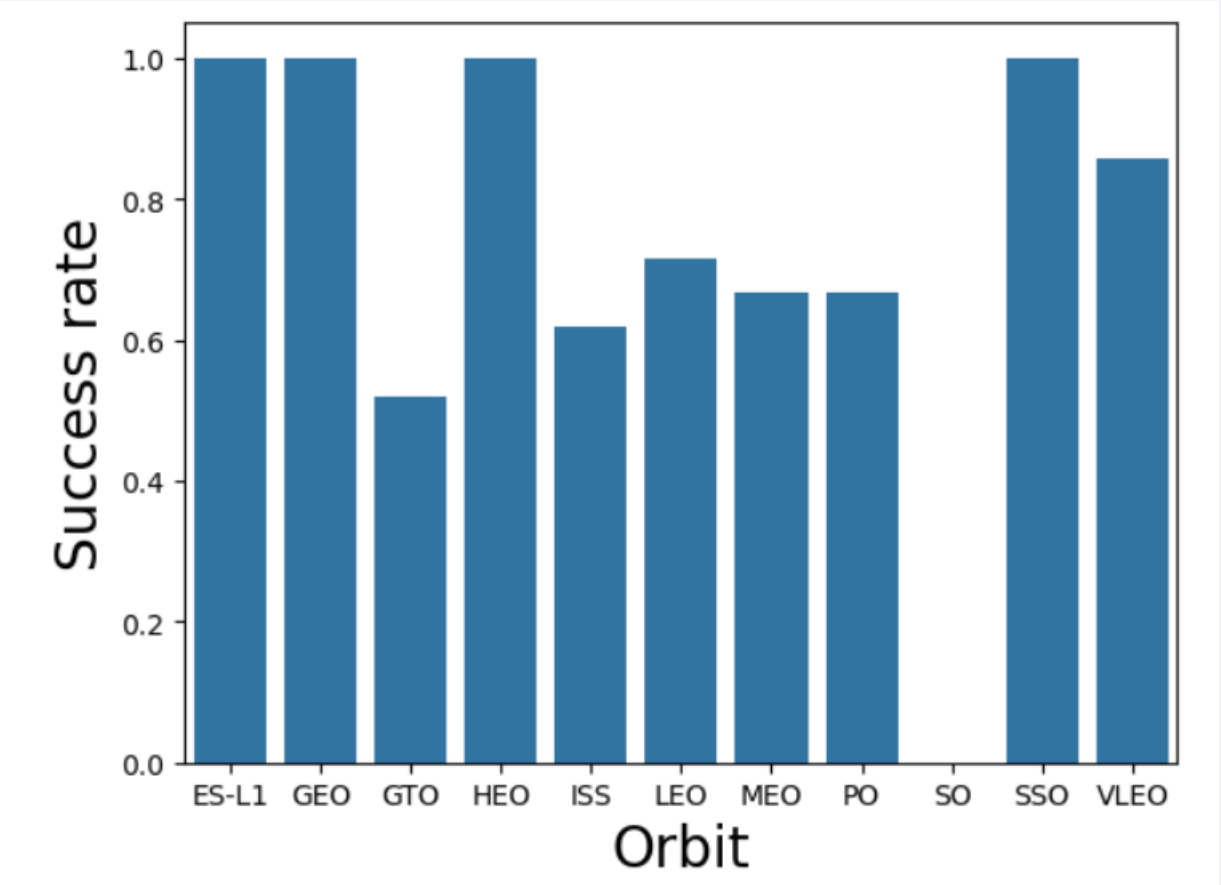
Payload vs. Launch Site

- CCAFS SLC 40 handled a wide range of payloads, including the heaviest ones (>15,000 kg), with many successes.
- KSC LC 39A had several high-payload missions, most of which were successful, suggesting its role in critical launches.
- VAFB SLC 4E shows a narrower payload range and fewer launches overall, but most were successful.
- Success is generally higher for medium-to-heavy payloads across all sites.



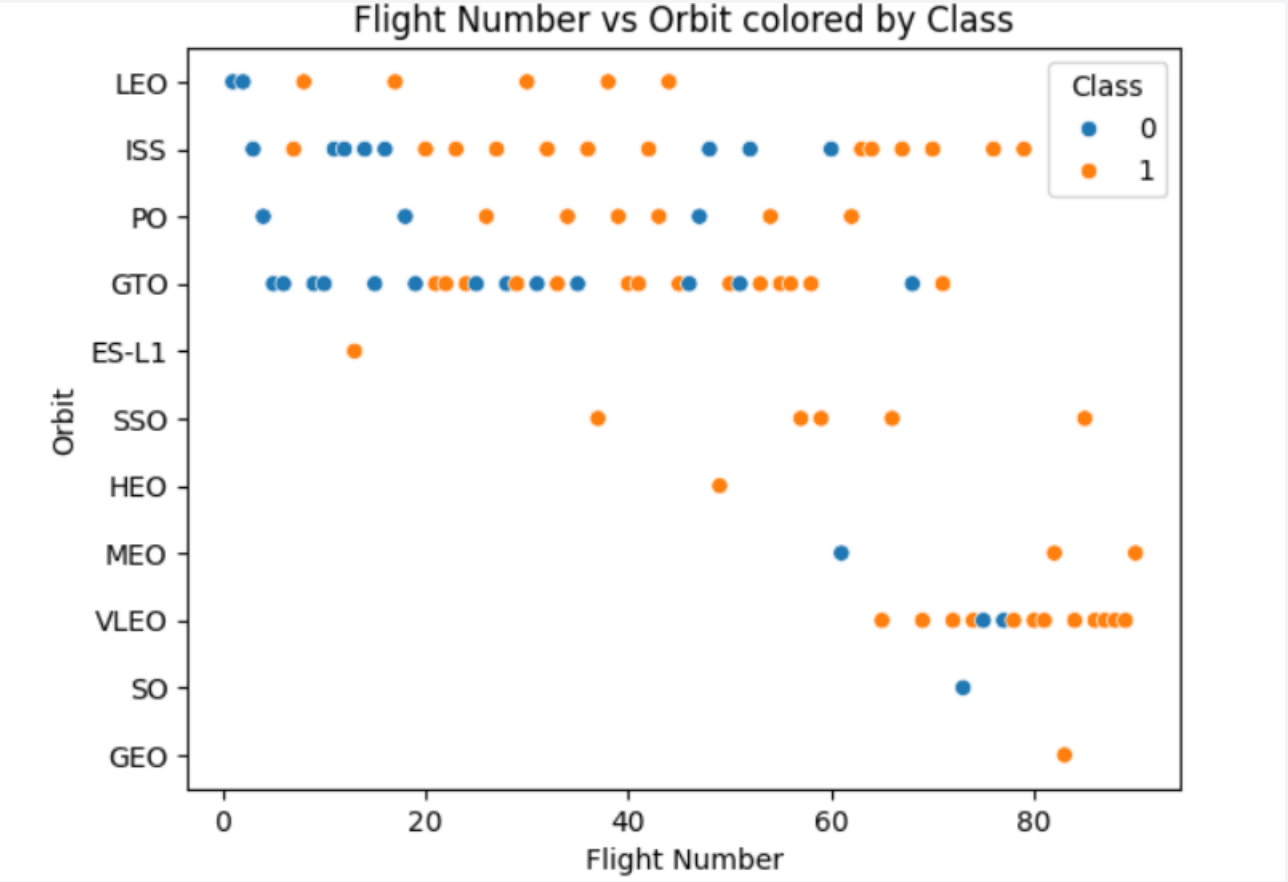
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO all show a perfect success rate (1.0), indicating strong reliability for missions targeting these orbits.
- GTO (Geostationary Transfer Orbit) has the lowest success rate, under 0.55, possibly due to higher mission complexity.
- Common orbits like LEO (Low Earth Orbit) and MEO have moderate success rates between 0.7–0.8.
- ISS missions have a success rate just above 0.6, which may reflect their stringent safety or launch timing requirements.
- VLEO (Very Low Earth Orbit) also performs well, with a success rate above 0.85.



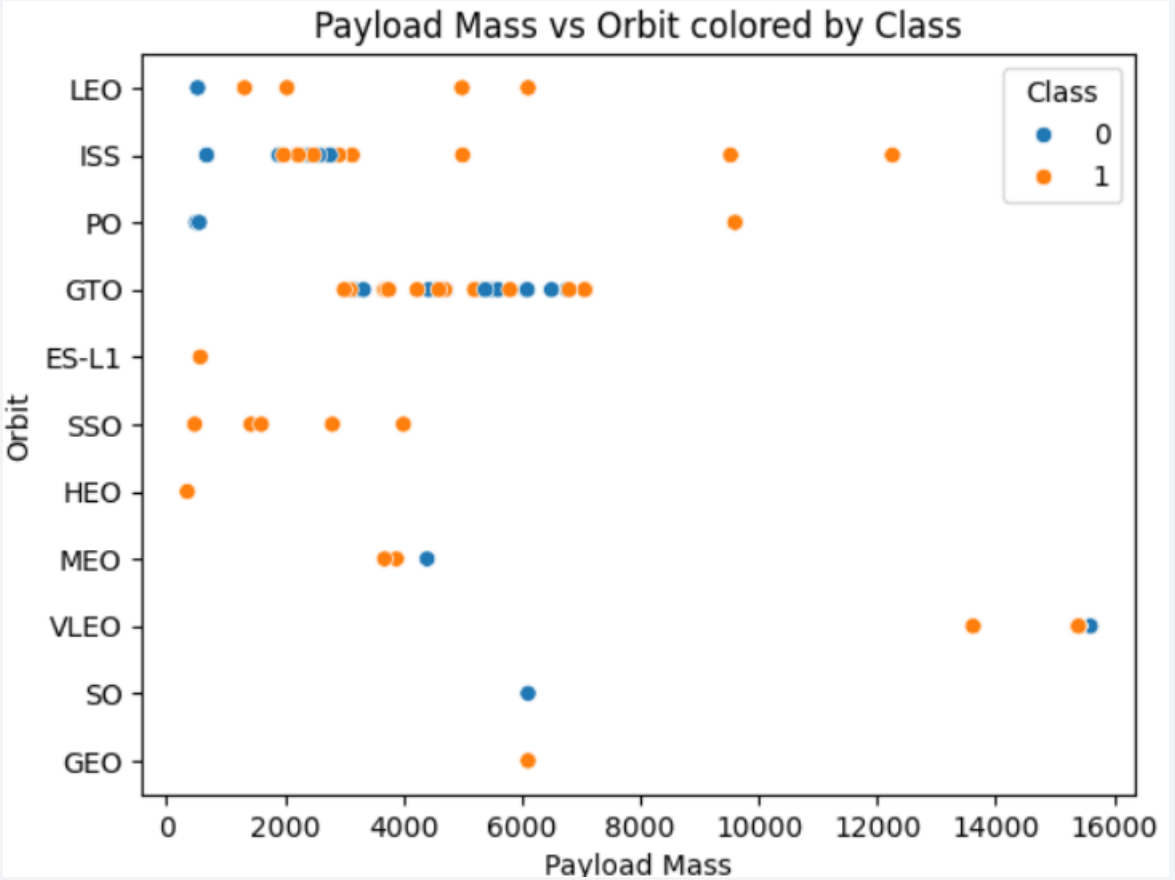
Flight Number vs. Orbit Type

- LEO, ISS, PO, and GTO are the most frequently used orbits across all flight numbers.
- Newer orbits like VLEO and SSO became more prominent in later missions (higher flight numbers).
- Early flights saw more variation in success, while later missions targeting newer orbits had higher consistency in success.
- Failures are more scattered early on, indicating improving reliability over time regardless of orbit.



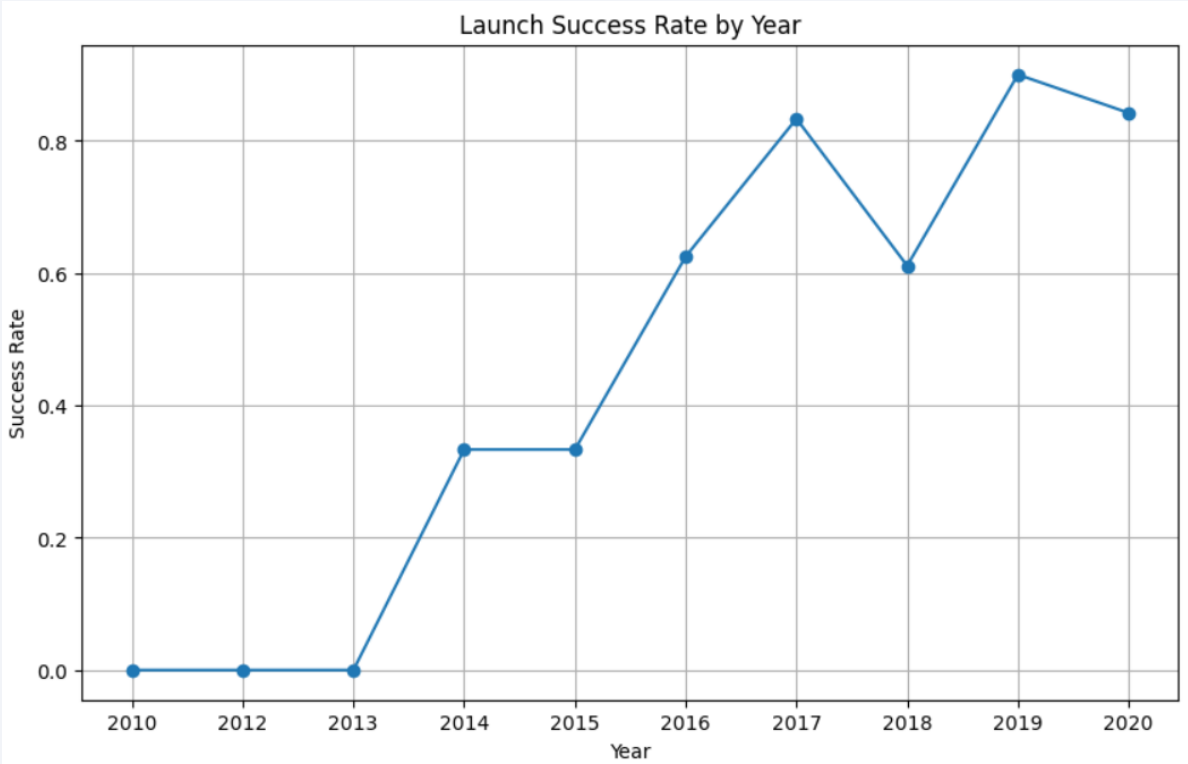
Payload vs. Orbit Type

- GTO (Geostationary Transfer Orbit) missions typically carry heavier payloads, up to ~10,000 kg and beyond, but include both successes and failures.
- LEO and ISS orbits generally involve lighter payloads (<6000 kg), with high success rates.
- SSO, HEO, and VLEO also show a concentration of successful launches in mid-weight ranges.
- The heaviest payloads (~15,000 kg) were directed toward GEO and KSC LC 39A, with mixed outcomes.



Launch Success Yearly Trend

- From 2010 to 2013, SpaceX had no successful launches, reflecting its early testing phase.
- A significant increase in success began in 2014, with steady improvements each year.
- 2017 and 2019 marked peak performance, with success rates above 85%.
- 2018 shows a noticeable dip, suggesting a brief setback, followed by a recovery in 2019.
- Overall, the chart demonstrates a clear upward trend, highlighting SpaceX's increasing launch reliability over the decade.



All Launch Site Names

```
df['Launch_Site'].unique()
```

```
array(['CCAFS LC-40', 'VAFB SLC-4E', 'KSC LC-39A', 'CCAFS SLC-40'],
      dtype=object)
```

- It uses .unique() to return an array of all distinct values found in the 'Launch_Site' column.
- The result shows that launches occurred from:CCAFS LC-40VAFB SLC-4EKSC LC-39A

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE "CCA%"
LIMIT 5;
```

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

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

- Purpose: Retrieve launch records specifically from Cape Canaveral (e.g., CCAFS LC-40).
- Result: Shows early SpaceX launches (2010–2013), all from CCAFS LC-40, with mission outcomes marked as success, though some had landing failures or no attempt.

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
From SPACEXTABLE
WHERE Customer is "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- Result: 45,596 kg
- It shows that NASA (CRS) missions have had a combined payload mass of 45,596 kilograms, indicating SpaceX's contribution to NASA's Commercial Resupply Services.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
From SPACEXTABLE
WHERE Booster_Version is "F9 v1.1"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Result: 2928.4 kg
- This indicates that, on average, missions using the F9 v1.1 booster carried around 2.9 metric tons of payload.

First Successful Ground Landing Date

```
%%sql
SELECT min(Date)
From SPACEXTABLE
WHERE Landing_Outcome LIKE "%Success (ground pad)%"
```

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

min(Date)
2015-12-22

- Result: 2015-12-22
- It identifies December 22, 2015, as the date of SpaceX's first successful ground landing, marking a major milestone in reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Purpose: Identify boosters that:Carried a payload between 4000 and 6000 kg, and
  - Had a successful mission outcome (likely landed on a drone ship).
  - Result: A list of booster versions such as F9 v1.1 B1011, F9 FT B1020, F9 B5 B1062.1, etc.
- This helps analyze which specific booster hardware reliably delivered medium-weight payloads and successfully completed their missions, especially on drone ship landings.

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Result:Failures: 1Successes: 100
- Purpose: To compare how many missions succeeded vs. failed using partial string matching in the Mission_Outcome field.
- This confirms SpaceX's high success rate, with only one recorded mission failure in the dataset.

```
%%sql
SELECT Count(Mission_Outcome) as failure, (SELECT Count(Mission_Outcome)
From SPACEXTABLE
WHERE Mission_Outcome LIKE "%Success%") as success
From SPACEXTABLE
WHERE Mission_Outcome LIKE "%Failure%"
```

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

failure	success
1	100

Boosters Carried Maximum Payload

- Result: Lists multiple F9 B5 boosters (e.g., B1048.4, B1049.4, B1051.3, etc.).
- Purpose: Identify all boosters that launched the heaviest payload ever carried by SpaceX.
- This helps analyze which booster configurations were trusted with the most demanding missions.

```
%%sql

SELECT Booster_Version
From SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_)
From SPACEXTABLE)
```

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

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

2015 Launch Records

```
%%sql
SELECT * ,substr(Date, 6,2) as month
FROM SPACEXTABLE
WHERE substr(Date,0,5)='2015' AND Mission_Outcome LIKE "%Failure%"
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	month
2015-06-28	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)	06

- Result:One mission on 2015-06-28 with Booster F9 v1.1 B1018
- Launch Site: CCAFS LC-40
- Mission failed in flight, and landing was precluded (drone ship)
- Purpose: To identify failures in 2015 specifically involving drone ship landing attempts, showing when and where they occurred.
- This query provides insight into SpaceX's early challenges with landing technology.

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

- Result:Most common outcome: "No attempt" (10 times)
- Followed by:
  - "Success (drone ship)" — 5
  - "Failure (drone ship)" — 5
  - "Success (ground pad)" — 3
- Purpose: To analyze how frequently each landing strategy was used and how successful it was during early SpaceX missions.

This helps assess SpaceX's progress in landing attempts during its experimental period.

```
%%sql
SELECT "Landing_Outcome", COUNT(*) as Times
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY Times DESC;
```

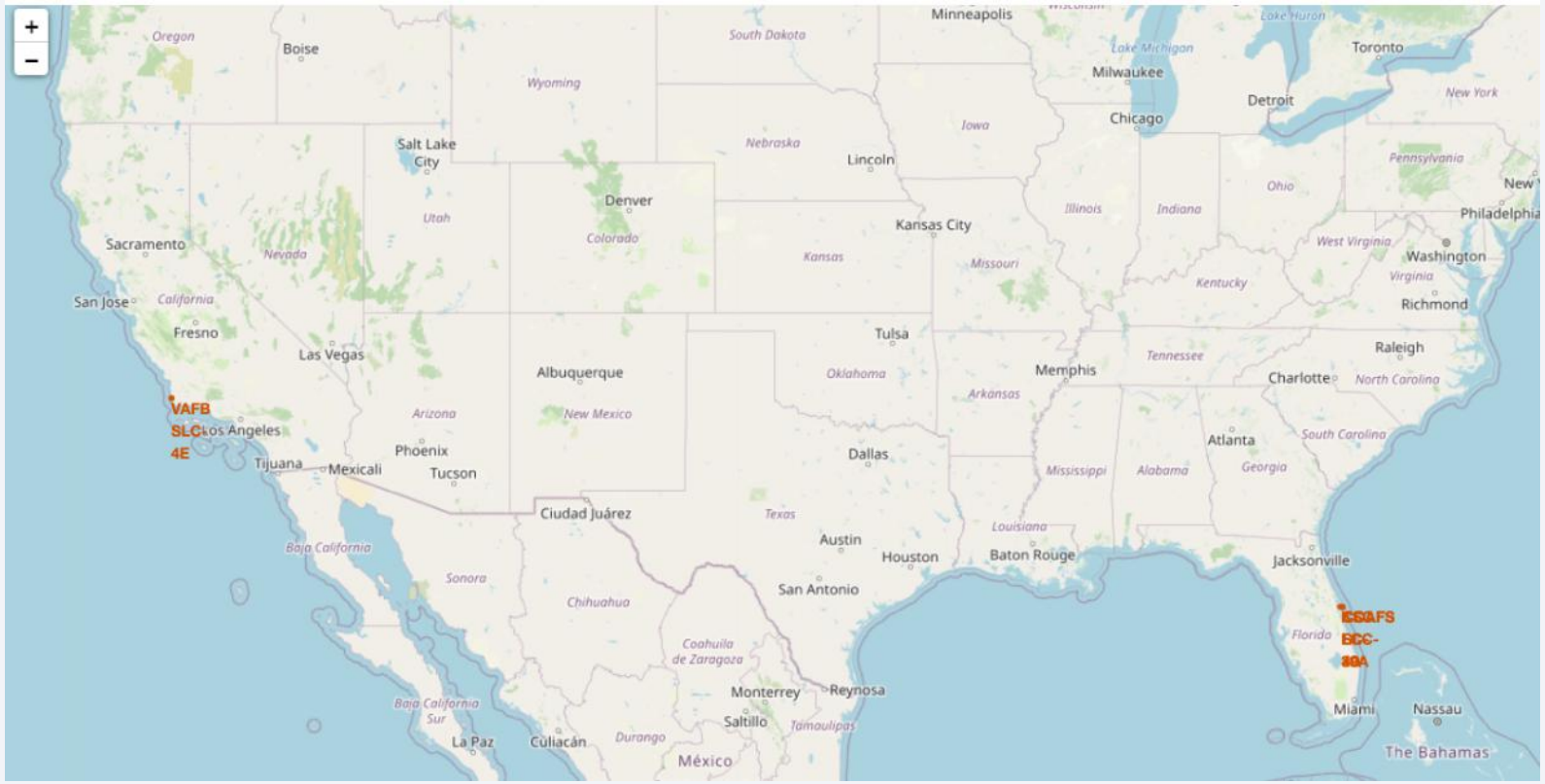
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Times
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

Section 3

Launch Sites Proximities Analysis

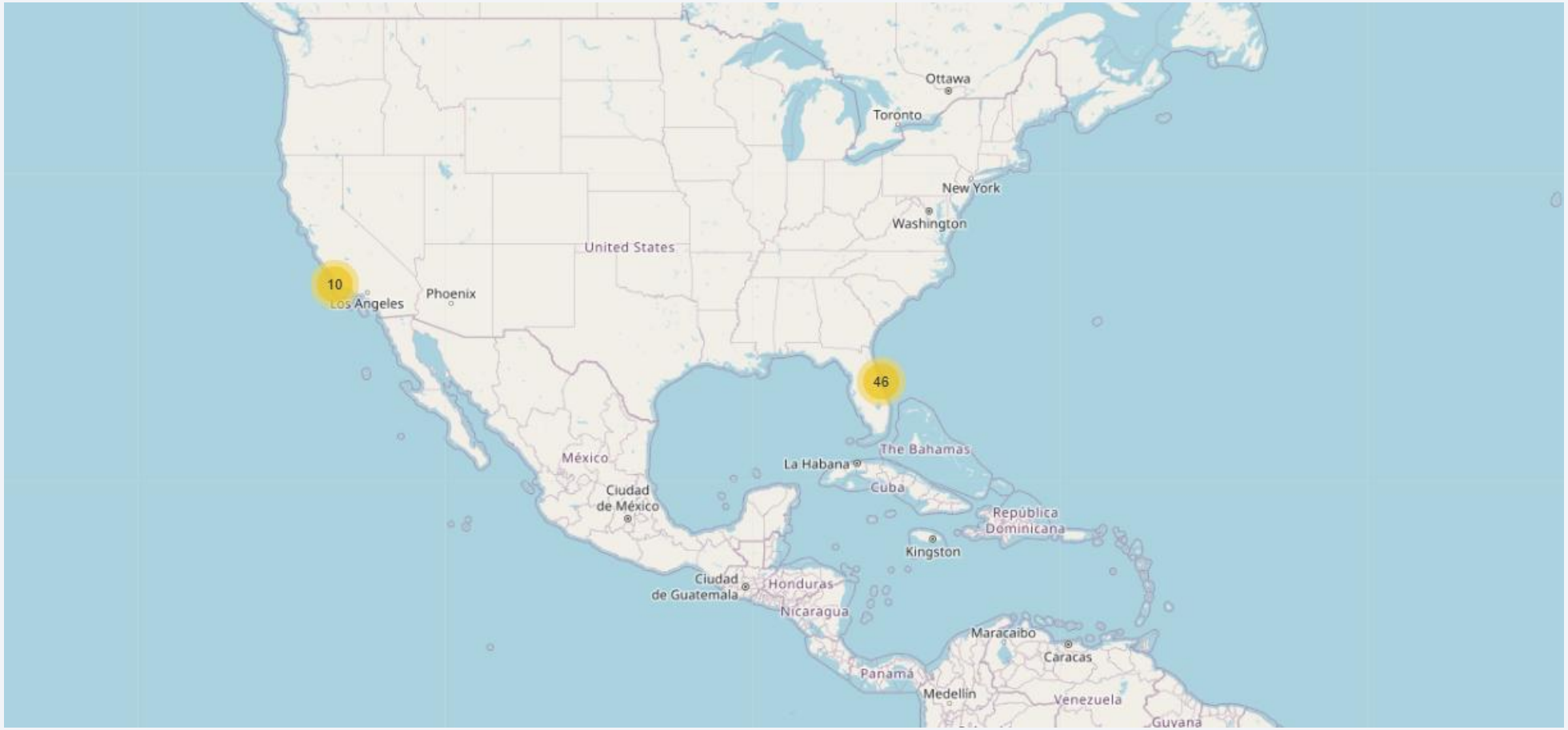
Launch Sites of SpaceX Missions in the United States



Launch Sites of SpaceX Missions in the United States

- East Coast Concentration:Two major launch sites are located in Florida (CCAFS and KSC), reflecting the importance of the Cape Canaveral region for orbital launches and ISS missions.
- West Coast Capability:VAFB SLC-4E in California is primarily used for polar and sun-synchronous orbit missions.
- Geographical Strategy:The map reveals SpaceX's strategic use of both coasts for diverse orbital requirements—east for equatorial and ISS launches, and west for high-inclination orbits.
- Operational Coverage:The map highlights that SpaceX launch coverage is nationwide, supporting a wide range of mission types.

SpaceX Launch Outcomes by Location



SpaceX Launch Outcomes by Location

- Clustered Markers:The yellow circles with numbers (e.g., 10, 46) represent marker clusters—a grouping of multiple launches occurring at close proximity.
- Clicking these clusters in an interactive notebook would expand them to reveal individual launch events.
- Launch Locations:
  - Florida (46 launches): This corresponds to multiple launch pads at Cape Canaveral and Kennedy Space Center.
  - California (10 launches): Represents launches from Vandenberg Air Force Base (VAFB).
- Color-coded Icons (visible when unclustered):
  - Each launch would be marked with a green rocket icon for success or a red rocket icon for failure (this detail is implemented in code even if clusters are shown in the screenshot).

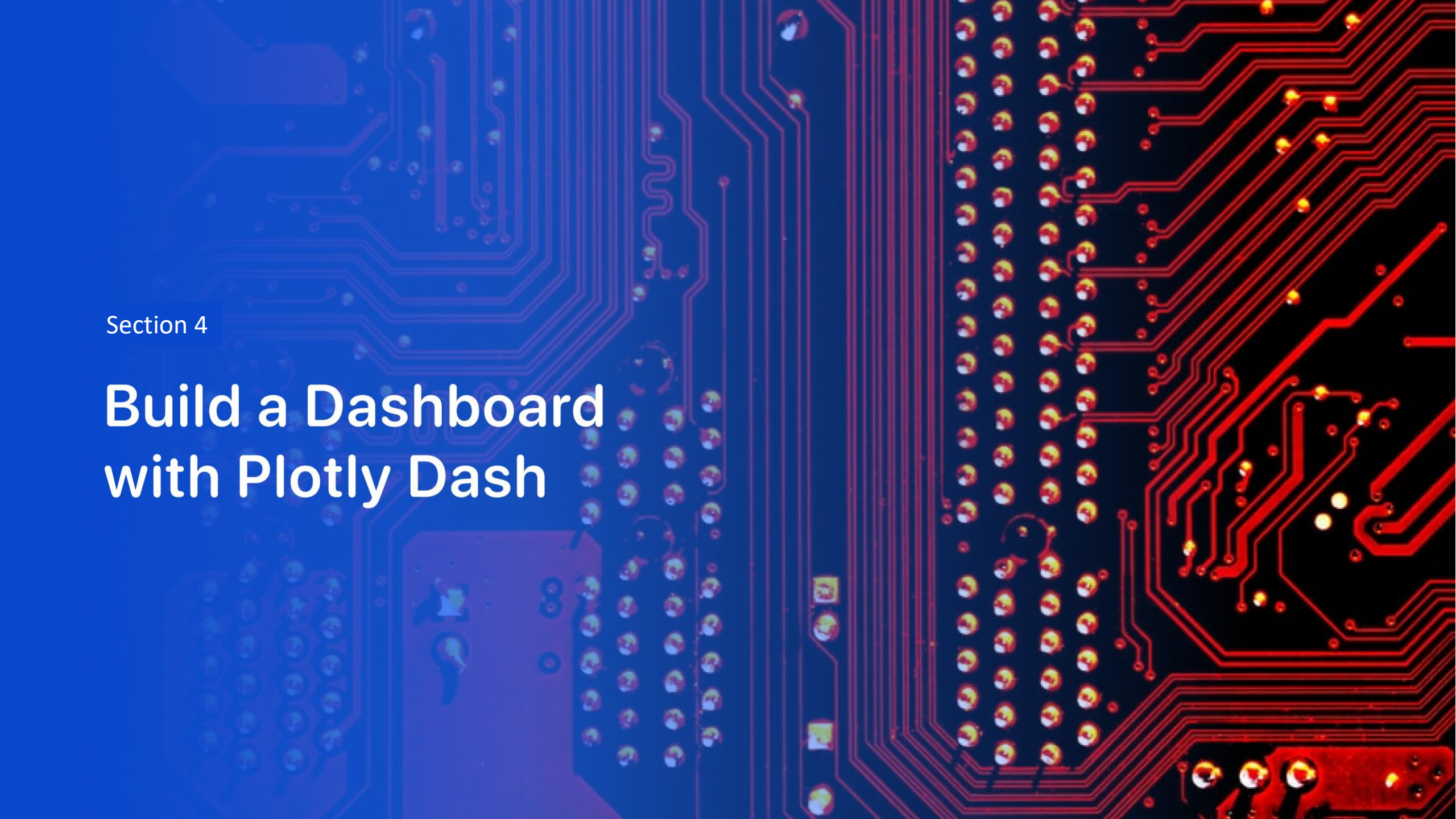
SpaceX Launch Outcomes by Location

- The majority of launches occurred from Florida (46 vs. 10), highlighting its significance as the main SpaceX launch hub.
- The use of marker clusters ensures performance and clarity when displaying dense geographic data points.
- Although success/failure isn't visible directly in the screenshot due to clustering, the underlying code supports color-labeling based on outcome—useful for analysis once zoomed in.

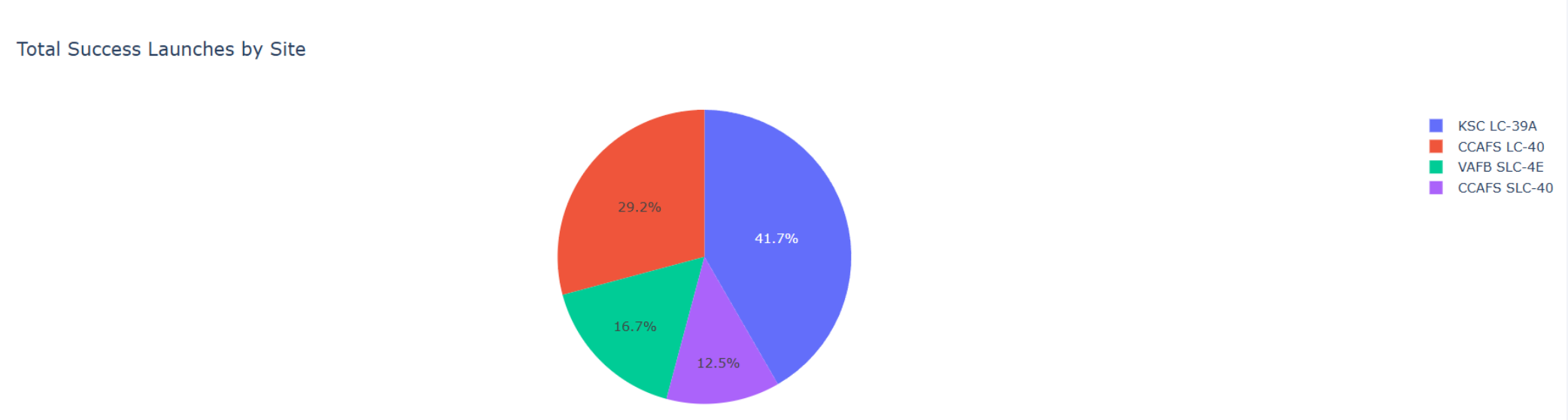
Proximity Analysis of CCAFS SLC-40 Launch Site

Proximity Analysis of CCAFS SLC-40 Launch Site

- The **coastline is ~23.07 km away**, making it feasible for drone ship operations and booster recovery.
- There is a **railway (or similar infrastructure)** within **0.85 km**, which is optimal for ground transport.
- The **launch site is well-placed** logistically for supporting SpaceX's launch and recovery operations.

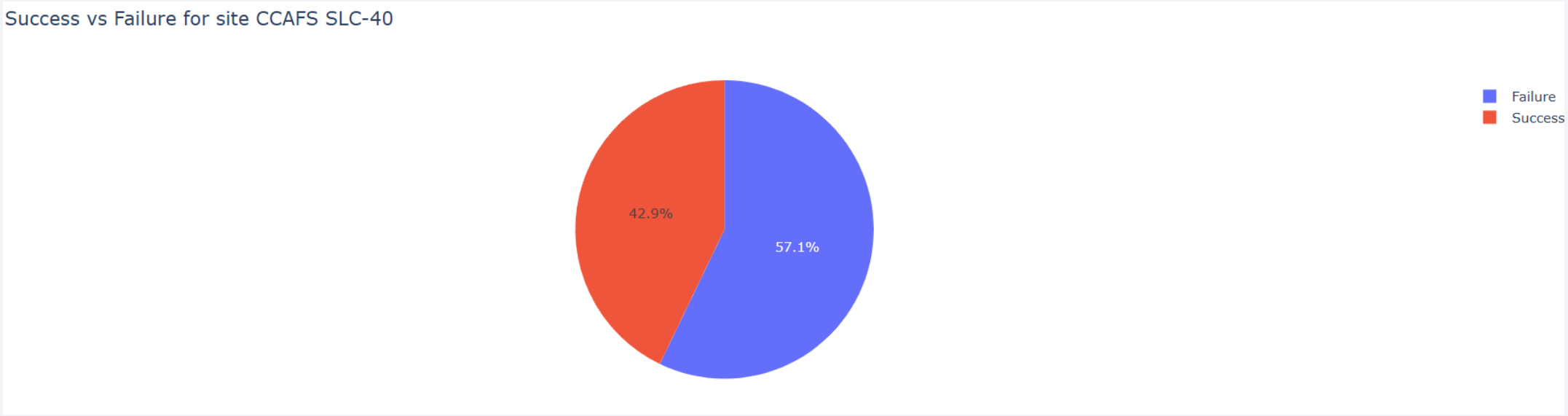

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Launch Site



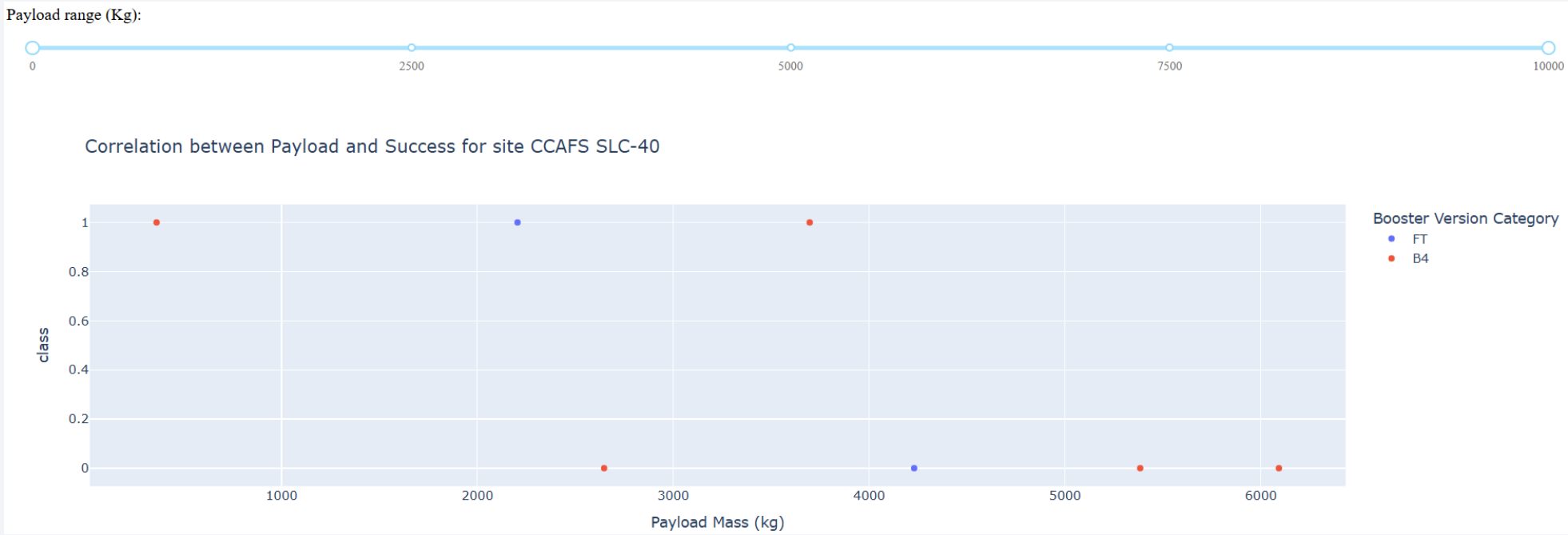
- **Kennedy Space Center (LC-39A)** is the most successful and frequently used launch site.
- The **Florida-based launch pads (KSC & CCAFS)** together account for **over 80% of all successful launches**, reinforcing Florida's dominance in SpaceX's launch operations.
- **VAFB (California)** supports a smaller, though significant, portion of launches—likely polar or sun-synchronous orbits.

<Dashboard Screenshot 2>



- Although CCAFS SLC-40 has a moderate success rate (57.1%), the failure rate is relatively high compared to other launch sites.
- This launch pad has likely been used in earlier stages of Falcon 9 development, contributing to a higher failure rate.
- As one of the major operational pads in Florida, the performance of this site is crucial to SpaceX's launch cadence and reliability history.

<Dashboard Screenshot 3>

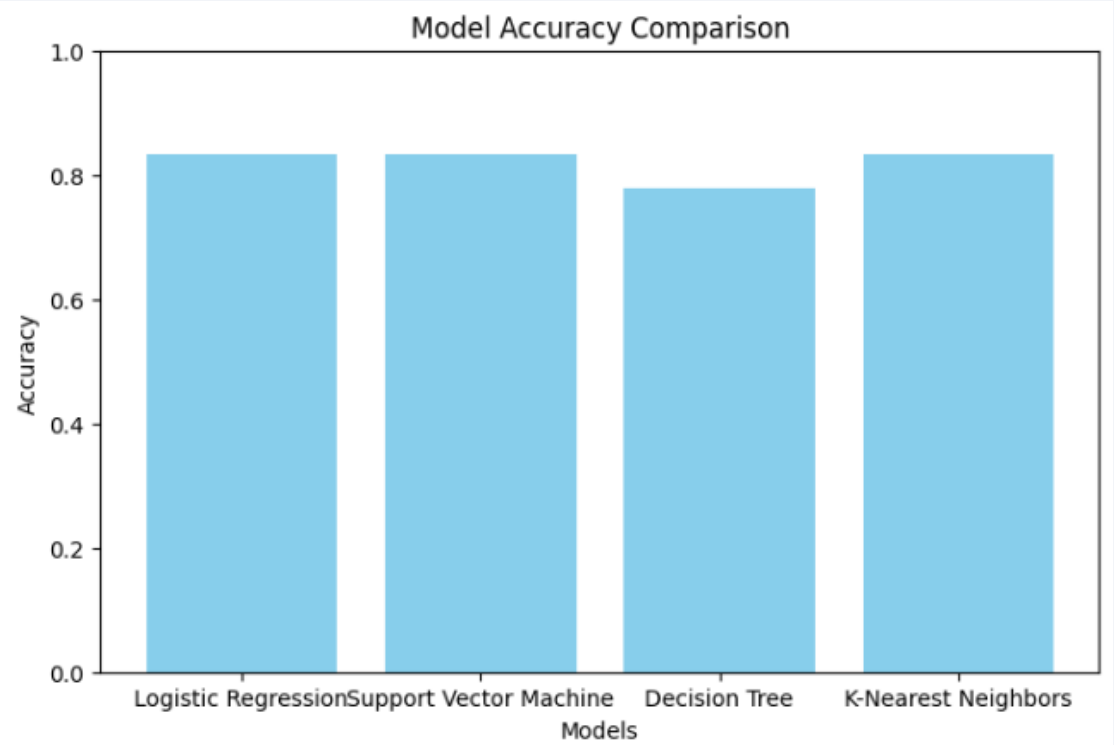


- The successes (y=1) and failures (y=0) are distributed across a variety of payload weights.
- There is no clear linear correlation between higher payload mass and launch outcome, suggesting other factors like booster version or external conditions may play a significant role.
- For lower payloads (~<3000 kg), both successes and failures are observed.FT boosters show both success and failure outcomes, just like B4 boosters, but FT appears less frequently.

Section 5

Predictive Analysis (Classification)

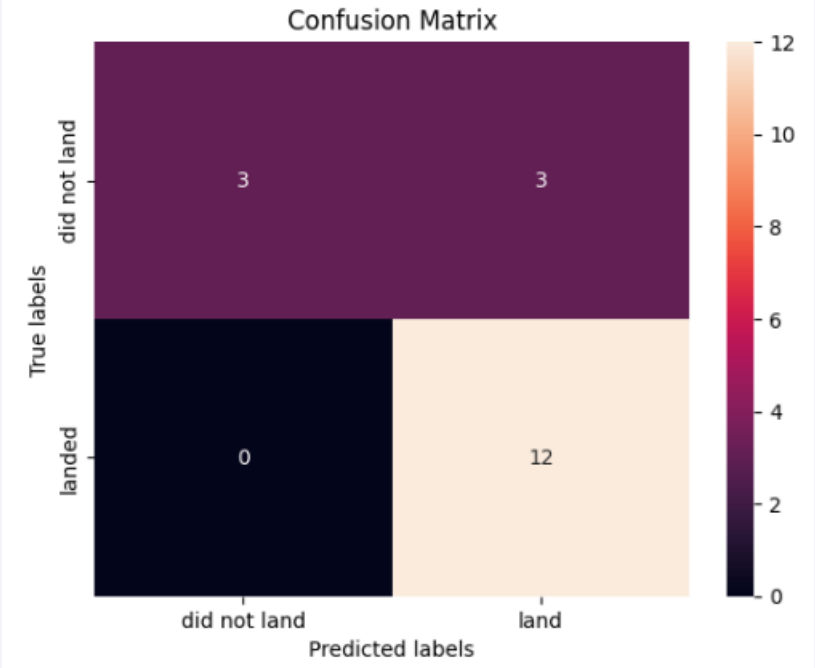
Classification Accuracy



- Best performing model: Logistic Regression with accuracy of 0.83334

Confusion Matrix

- The model is excellent at identifying successful landings (100% recall for landings).
- Some failures are misclassified as landings, which might be critical in real-world applications where safety and cost are involved.
- This behavior suggests a model that leans toward predicting success—good if success is the dominant class, but might need tweaking in a risk-sensitive environment.



Conclusions

- Data-driven insights show that launch success is strongly influenced by launch site, orbit type, and payload mass.
- KSC LC-39A emerged as the most successful launch site, while GTO missions had the highest failure rates.
- Launch success has significantly improved over time, reflecting SpaceX's operational maturity.
- Exploratory analysis via SQL and visualization tools revealed key trends such as orbit usage, payload patterns, and booster reliability.
- Predictive modeling using logistic regression achieved ~83% accuracy, reliably predicting successful landings.
- The confusion matrix indicates strong precision for predicting success, though failure prediction can be further improved.
- This project demonstrates how machine learning and data science can support aerospace operations, improve safety, and optimize launch planning.

Appendix

- Project Repository
  - GitHub LinkKey
- Python Libraries
  - pandas, numpy, matplotlib, seaborn, plotly, dash, folium, sklearn, BeautifulSoup
- Notebooks
  - spacex-data-collection-api-bak.ipynb – API data
  - collectionwebscraping.ipynb – Launch metadata
  - scrapingeda-visualization.ipynb – Data visualization
  - predictive-modeling.ipynb – Machine learning models

Appendix

- SQL Queries
  - Find earliest successful landing
  - Average payload by booster
  - Launches from CCAFS
  - Drone ship recoveries with 4000–6000 kg payload
- Dash App Features
  - Launch site dropdown
  - Payload range slider
  - Pie chart and scatter plot interactivity

Appendix

- Model Parameters (Best Model: Logistic Regression)
  - Regularization: l2
  - Solver: liblinearEvaluation:
  - Accuracy = 0.8333
  - Confusion Matrix Output
  - TP = 12, FN = 0, FP = 3, TN = 3
- High recall for success class, moderate false positive rate
  - Visualization AssetsInteractive
  - launch map with clusteringPayload vs. Outcome scatter plot
  - Site-specific success pie charts

Thank you!