Winning Space Race with Data Science

Jhermien Paul Alejandria
Data Scientist

github.com/jhermienpaul

Outline

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

Executive Summary

This capstone project delivers a full-stack data science workflow to analyze and predict SpaceX launch success, using techniques from data wrangling and SQL exploration to geospatial visualization and machine learning. Data from multiple sources were cleaned and explored for hidden patterns, with interactive dashboards and maps revealing key trends in launch sites, payloads, and mission outcomes.

Results show that launch success is consistently high across major SpaceX sites, with no strong link between payload mass and failures. All four classification models (Logistic Regression, SVM, Decision Tree, and KNN) achieved the same strong test accuracy (83%), so model selection comes down to interpretability and transparency. This project provides both actionable insights and a technical template for predictive analytics in aerospace.

Introduction

- Project Background
  - SpaceX has pioneered commercial spaceflight, launching hundreds of missions with diverse configurations and outcomes.
- Research/Business Question
  - What are the key drivers and spatial/temporal patterns of SpaceX launch success, and how can we predict future mission outcomes?
- Objectives
  - Identify features and locations most correlated with launch success.
  - Explore data through visual and spatial analytics.
  - Build an interactive dashboard for stakeholders.
  - Construct and evaluate a predictive model for mission outcomes.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - REST API and Web Scraping
- Perform data wrangling
  - Data Cleaning, Data Transformation, and Feature Engineering
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, Decision Trees, K-Nearest Neighbors, and Support Vector Machine

Data Collection

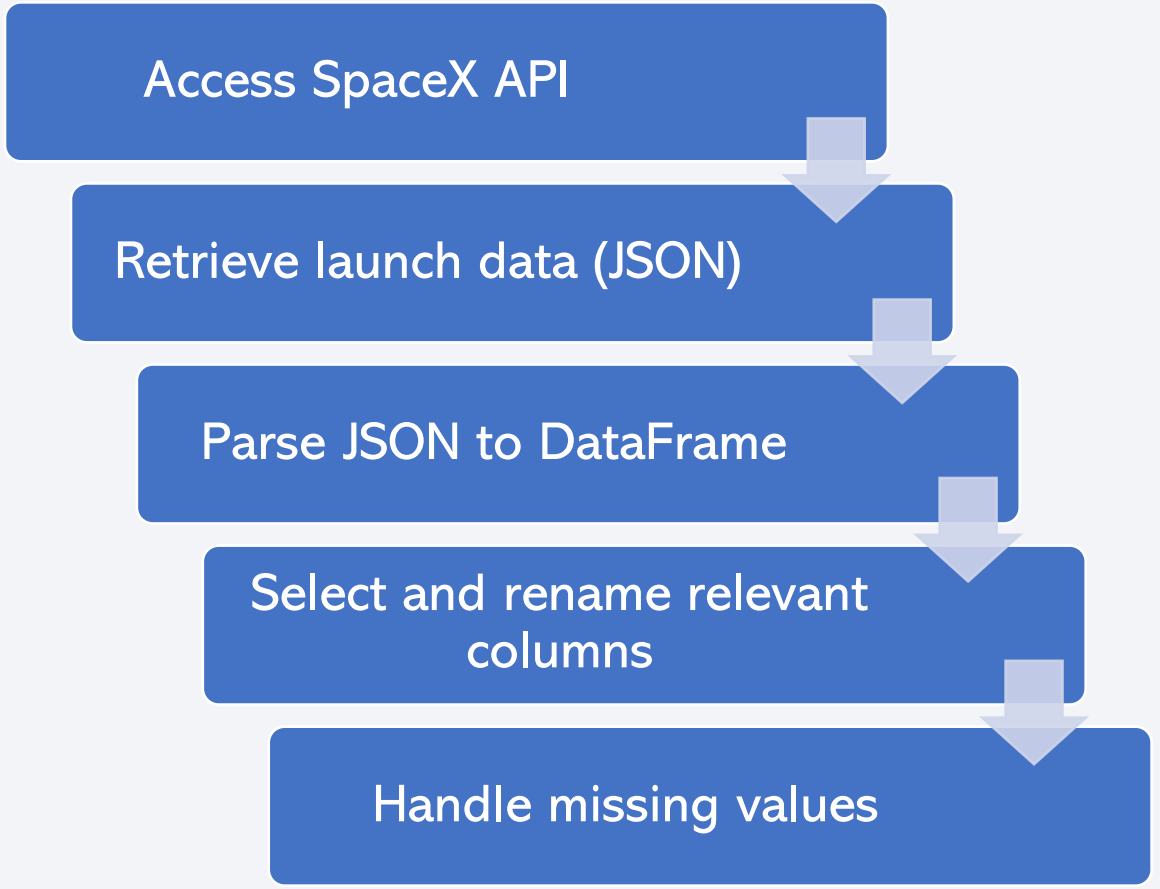
- REST API Extraction
  - Queried the official SpaceX API to obtain structured, up-to-date launch records in JSON format.
  - Parsed key attributes such as flight number, launch date, launch site, payload mass, orbit, booster version, customer, and mission outcome. Handled missing data by filtering nulls and verifying field presence.
- Web Scraping (Wikipedia)
  - Supplemented API data by scraping SpaceX launch history tables from Wikipedia using requests and BeautifulSoup.
  - Extracted additional fields and corrected inconsistencies (e.g., launch site aliases, payload mass, missing customers). Removed annotations/references, harmonized column types, resolved duplicates, and merged scraped data with API dataset.

Data Collection – SpaceX API

- Key Steps / Phases
  - I connected to the SpaceX API endpoint, downloaded launch data in JSON format, parsed it into a pandas DataFrame, selected and renamed relevant columns, and cleaned up missing values for consistency.

- Jupyter notebook:

  Data Collection with API.ipynb

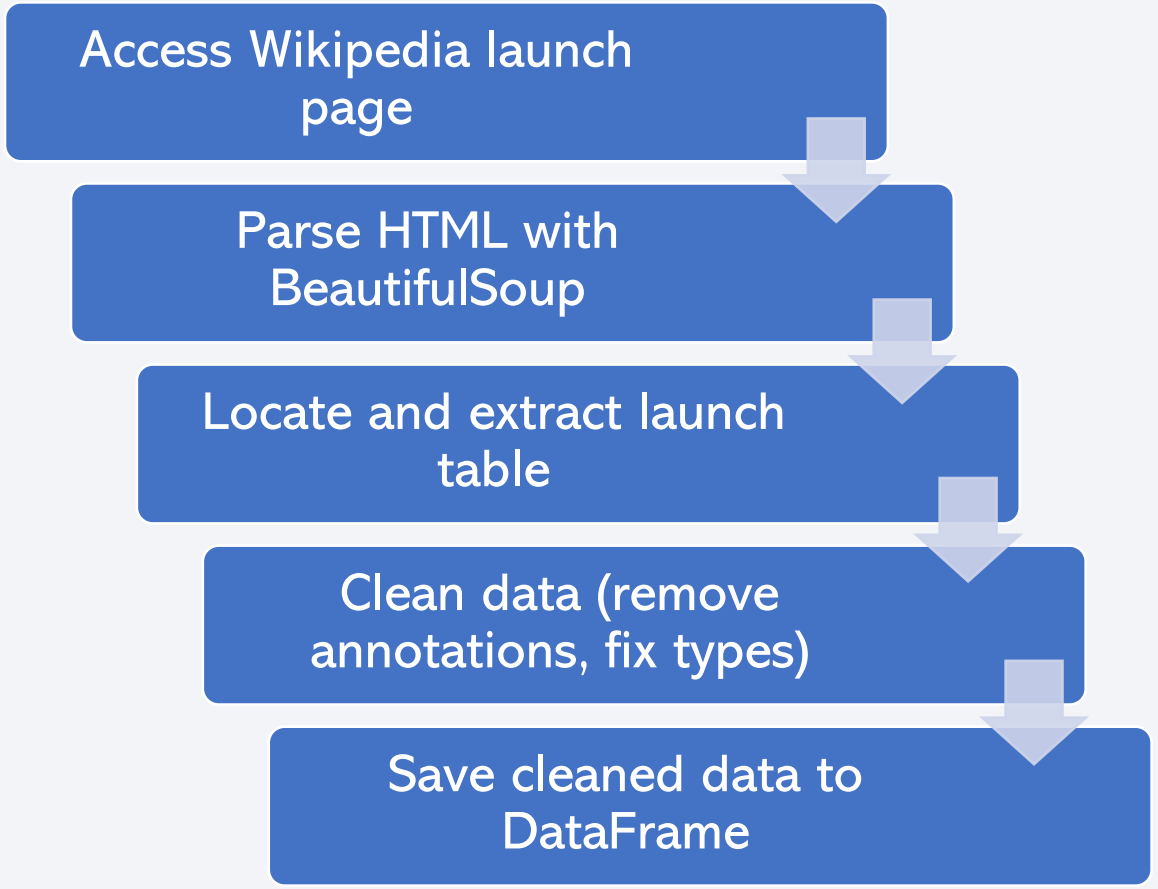


Data Collection – Scraping

- Key Steps / Phases
  - I accessed the Wikipedia SpaceX launches page, parsed the HTML with BeautifulSoup to find and extract launch tables, removed reference links and fixed messy entries, converted the cleaned data to a DataFrame, and ensured consistent formatting.

- Jupyter notebook:

  Data Collection with Web Scraping.ipynb

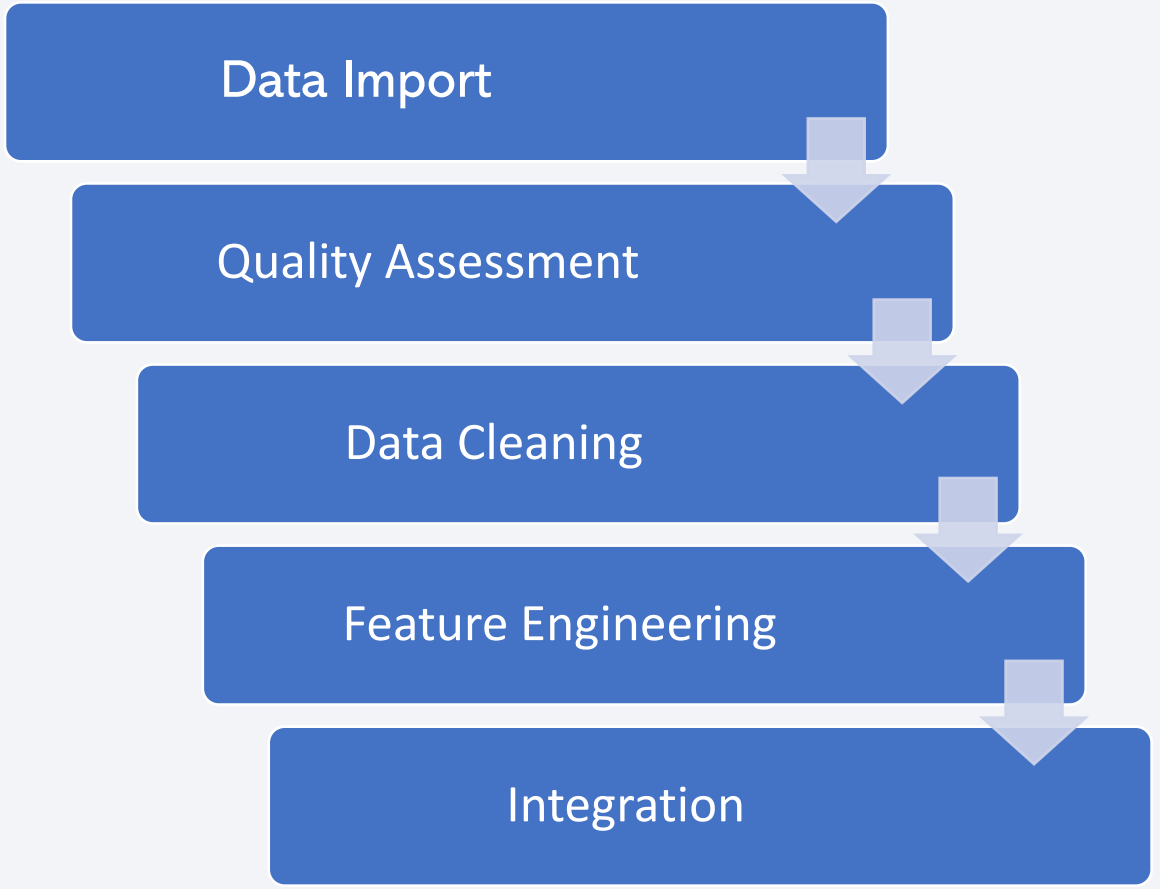


Data Wrangling

- Key Steps / Phases
  - Raw data from multiple sources were imported, assessed for quality, cleaned for consistency, engineered to create useful features, and merged into a unified dataset ready for analysis.

- Jupyter notebook:

  Data Wrangling.ipynb



EDA with Data Visualization

- Summary of Charts Plotted:
  - Scatter Plots: Explored relationships between launch sites, orbit types, flight numbers, payload mass, and mission outcomes. This helped identify patterns and anomalies in launch success and failure rates.
  - Bar Chart: Compared launch success rates across different orbit types to quickly pinpoint which orbits have the highest/lowest reliability.
  - Line Chart: Visualized the average annual success rate, highlighting SpaceX's dramatic improvements in launch reliability over time.

- Jupyter notebook: Exploratory Data Analysis with Visualization.ipynb

EDA with SQL

- Summary of SQL Queries Performed:
  - Queried and summarized key mission stats (unique launch sites, mission outcomes, payload totals/averages) to understand overall launch patterns.
  - Filtered records to answer specific business questions (e.g., booster performance, landing outcomes, milestone dates) using flexible SQL conditions and aggregations.
  - Used subqueries and grouping to benchmark technical achievements (e.g., max payloads, monthly/yearly trends, recovery methods).

- Jupyter notebook: Exploratory Data Analysis with SQL.ipynb

Build an Interactive Map with Folium

- Summary of Map Objects and Rationale
  - Markers: Plotted each launch site as an interactive marker to visualize SpaceX's geographic distribution and enable site-specific insights.
  - Marker Clusters: Used marker clusters to avoid overlapping points and make dense regions (like Florida) easier to explore.
  - Distance Lines: Drew lines from launch pads to the nearest coastline to analyze site safety and proximity to the ocean.

- Jupyter notebook: Interactive Mapping with Folium.ipynb

Build a Dashboard with Plotly Dash

- Plots, Graphs, and Interactions Added:
  - Pie charts to visualize launch success counts by site and by outcome (success vs. failure).
  - Dropdown menu for selecting specific launch sites, instantly updating all charts.
  - Scatter plot showing the relationship between payload mass, launch outcome, and booster version.
  - Payload range slider that lets users filter data by payload mass, updating the scatter plot in real time.

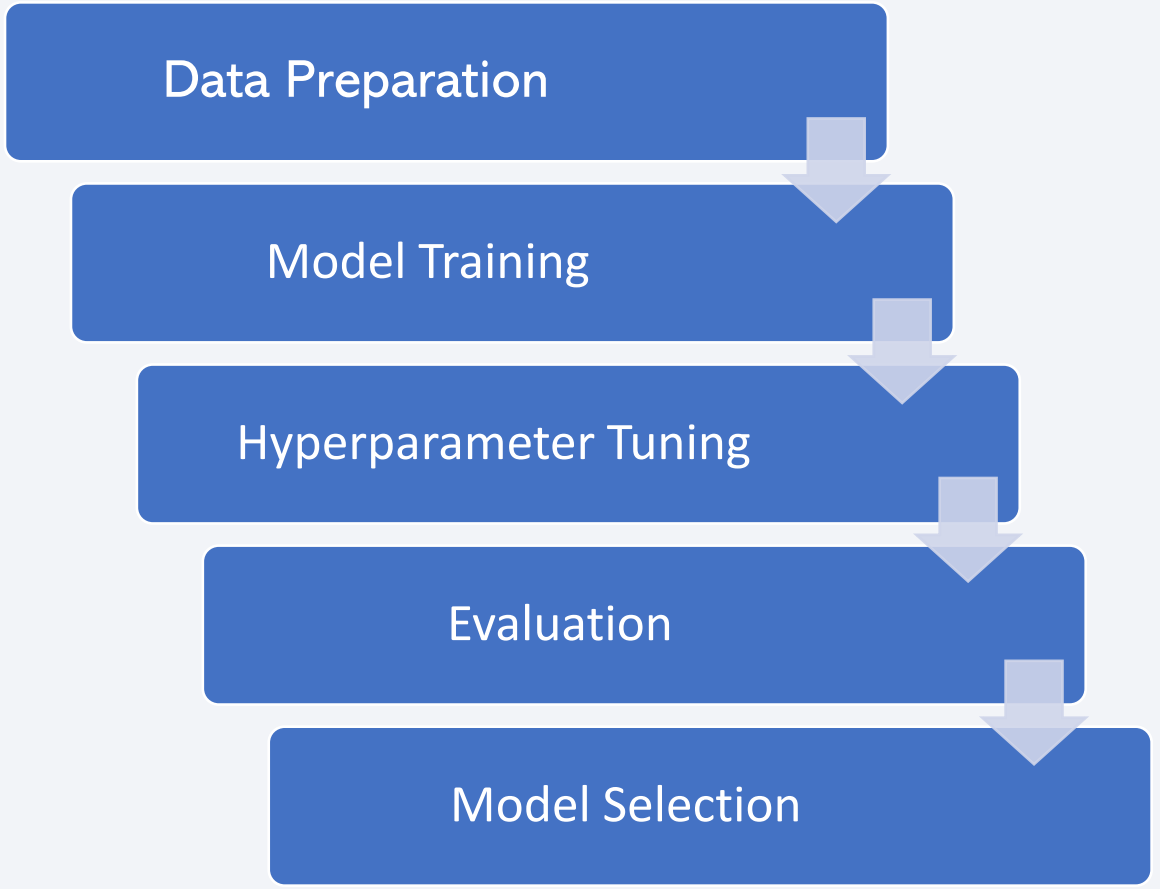
- Plotly Dash: Interactive Dashboard with Plotly Dash.py

Predictive Analysis (Classification)

- Key Steps / Phases
  - Built, tuned, and evaluated multiple classification models (Logistic Regression, SVM, Decision Tree, KNN) using a consistent pipeline with cross-validation, and selected the best performer based on test accuracy.

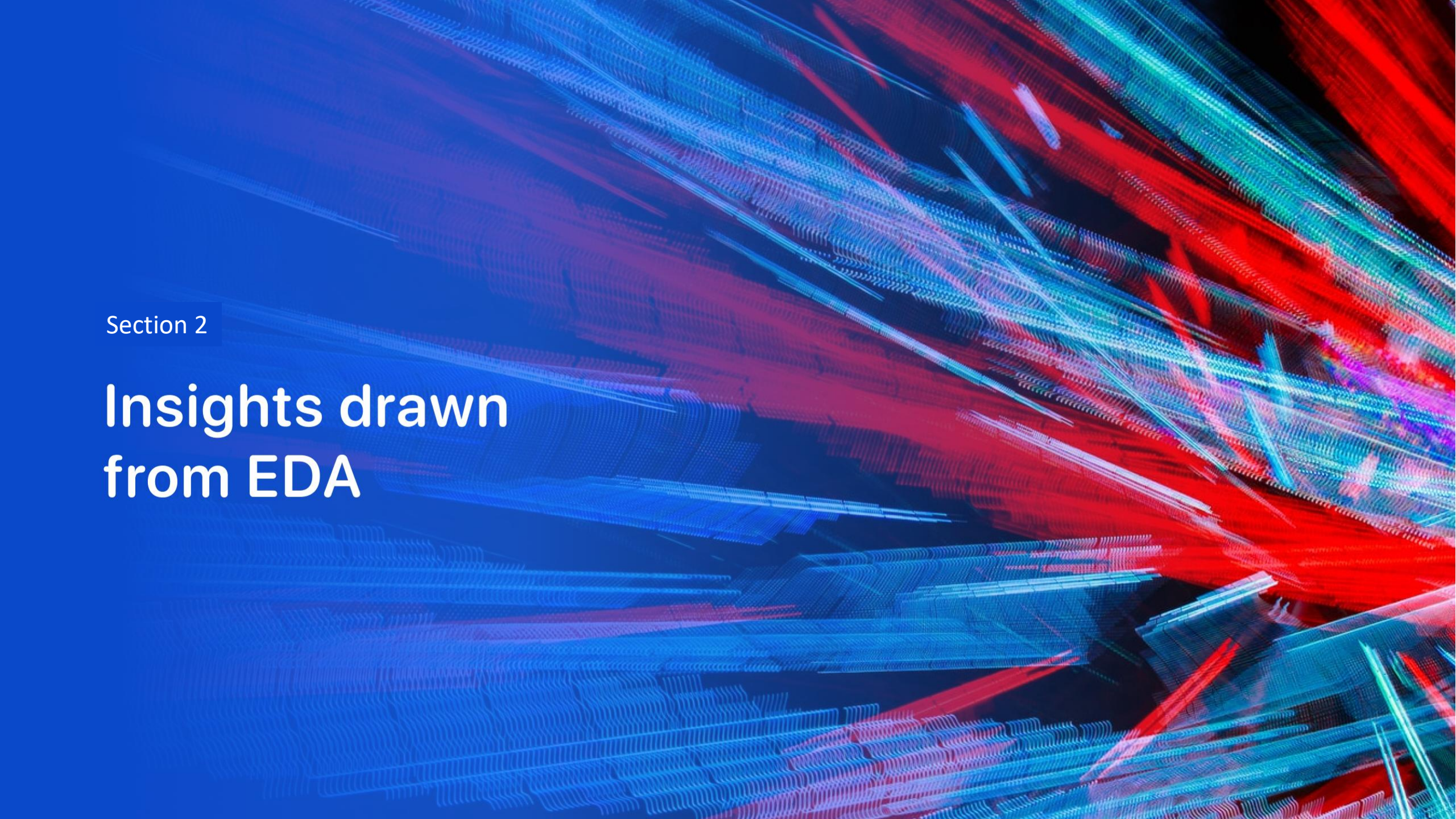
- Jupyter notebook:

  Machine Learning Model Pipeline.ipynb



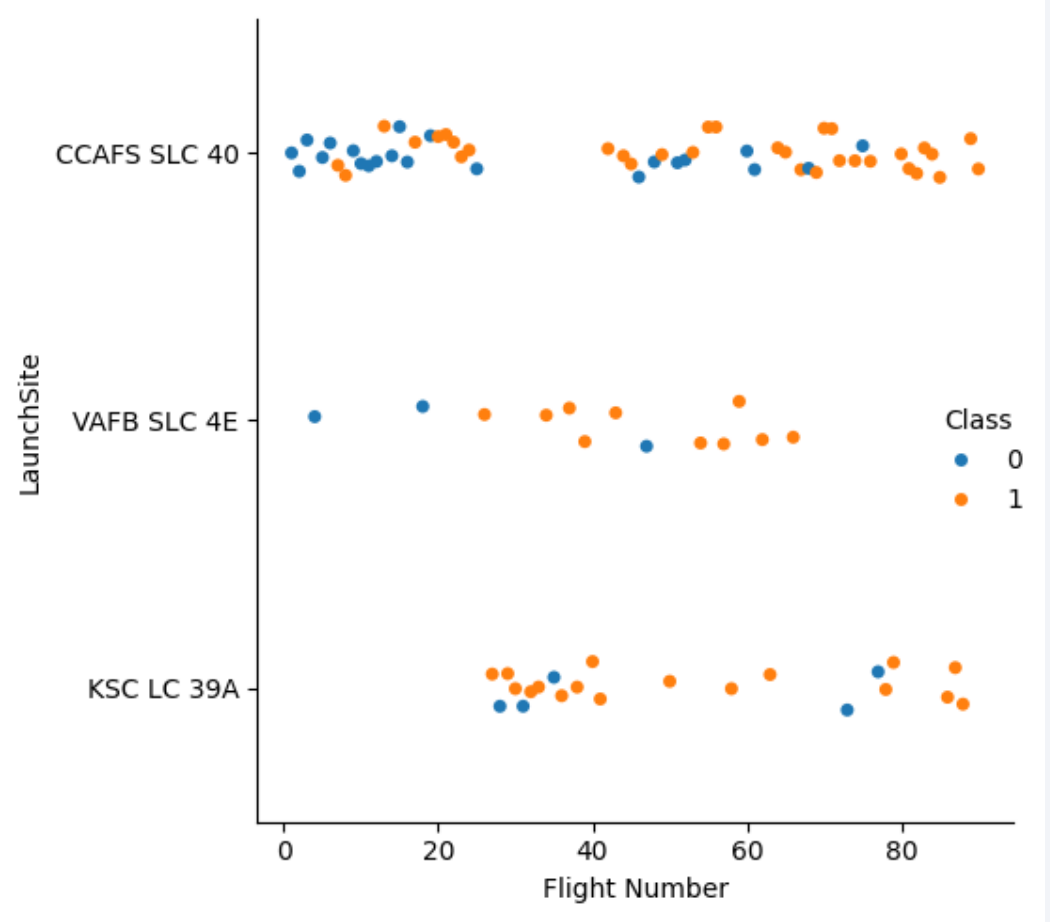
Results

- Exploratory Data Analysis (EDA):
  - Found that most launches happened at two main sites in Florida and California, with launch frequencies and payload masses visualized through charts and interactive maps.
- Descriptive Analytics:
  - Showcased dynamic dashboards and maps where users can explore launch success rates by site, booster versions, and payload mass. Some launch sites consistently performed better, but larger payloads didn't always mean lower success. The dashboards made it easy to spot trends and outliers at a glance.
- Predictive Analysis:
  - Multiple machine learning models (Logistic Regression, SVM, Decision Tree, KNN) were built to predict landing success, all achieving 83% accuracy—solid performance, no clear "winner."


Section 2
Insights drawn from EDA

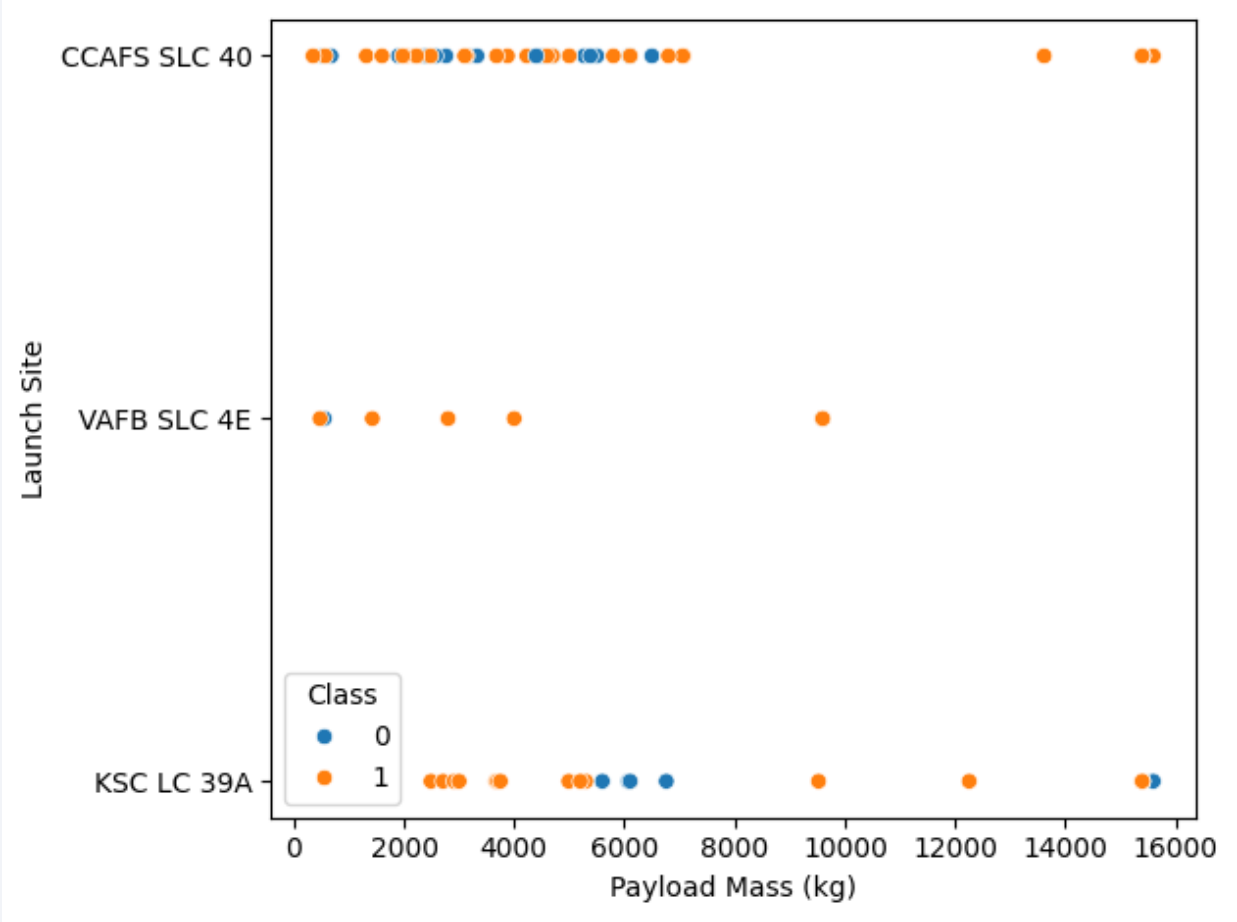
Flight Number vs. Launch Site

- In this chart, Class 1 indicates a successful launch, while Class 0 represents a failed launch.
- CCAFS SLC 40 had the highest launch frequency and showed a clear improvement in success rates (more orange) as flight numbers increased.
- All sites experienced both failures and successes but later launches at each site generally shifted toward more successful outcomes.
- There is strong evidence of operational learning, with launch success improving over time regardless of site.



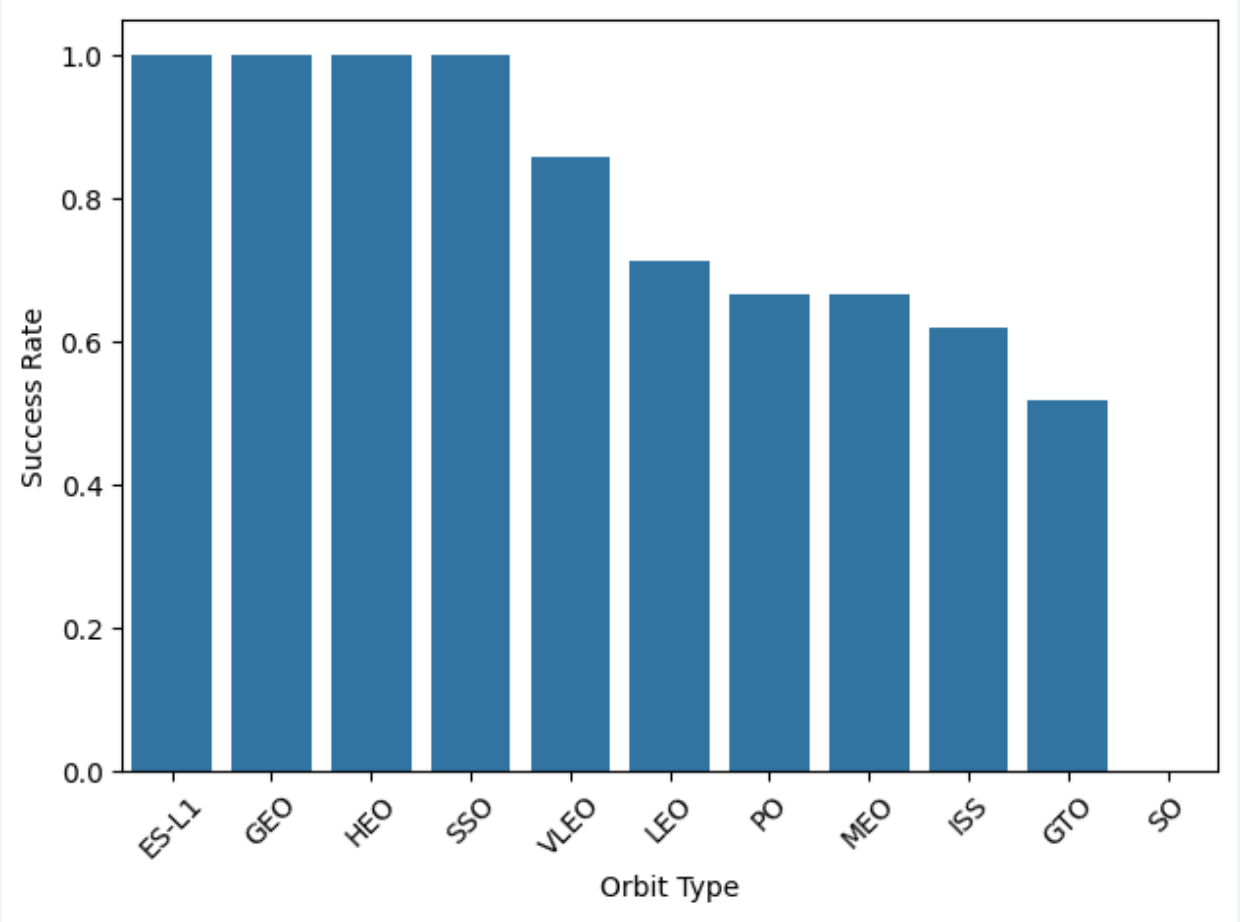
Payload vs. Launch Site

- Successful launches (Class 1) occurred across a wide range of payload masses at all launch sites.
- Failures (Class 0) are more frequent with lower to mid-range payloads, especially at CCAFS SLC 40 and KSC LC 39A.
- Very high payload masses (above 10,000 kg) are almost always associated with successful launches.
- There is no clear payload mass threshold below which launches consistently fail, indicating other factors also play a role in launch outcome.



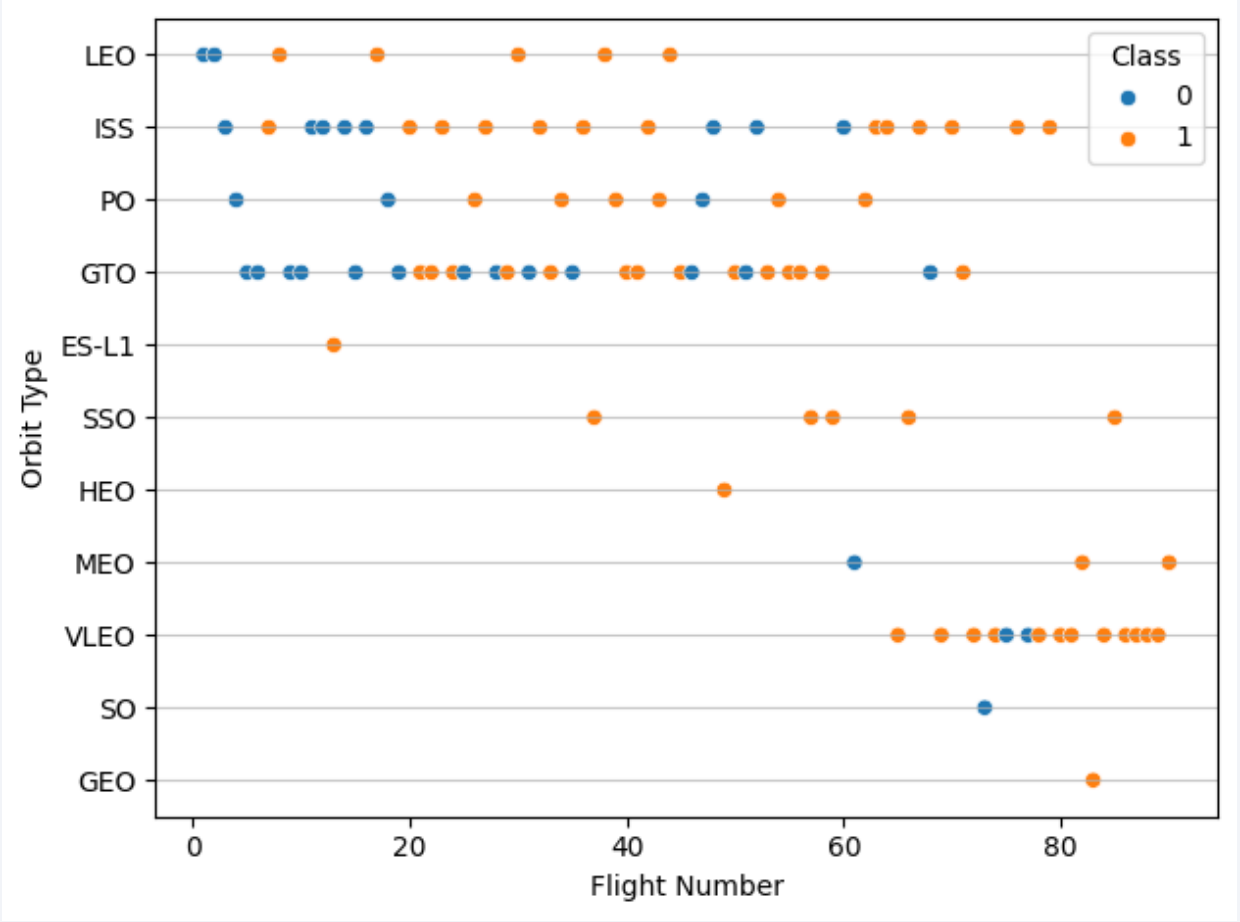
Success Rate vs. Orbit Type

- Launches targeting ES-L1, GEO, HEO, and SSO orbits had a perfect (100%) success rate.
- Success rates drop for lower orbits, with LEO, PO, and MEO orbits showing moderate reliability.
- The lowest success rates were observed for ISS, GTO, and SO orbits, indicating higher mission risk.
- Orbit type is a strong predictor of launch success, with deep space and geostationary missions being the most consistently successful.



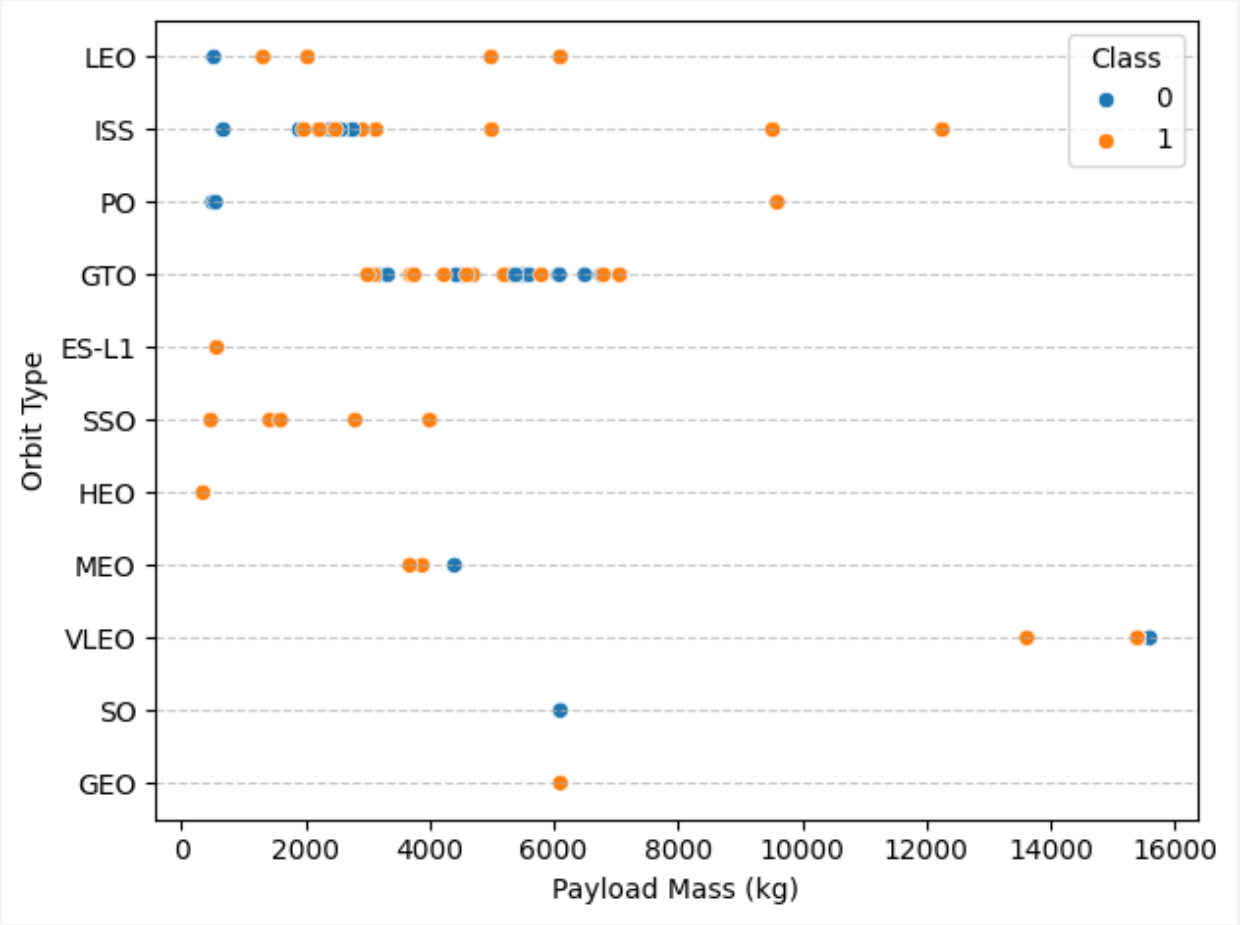
Flight Number vs. Orbit Type

- Successes (Class 1) and failures (Class 0) are observed for most orbit types, especially in LEO, ISS, PO, and GTO, indicating varying reliability.
- Missions to higher orbits like VLEO, SSO, GEO, and HEO show mostly successful outcomes with few or no failures.
- For some orbits (e.g., LEO, ISS, GTO), the rate of success improves as flight number increases, suggesting learning and operational improvement over time.
- Some orbit types, such as SO, ES-L1, and GEO, have very limited launches but all were successful.



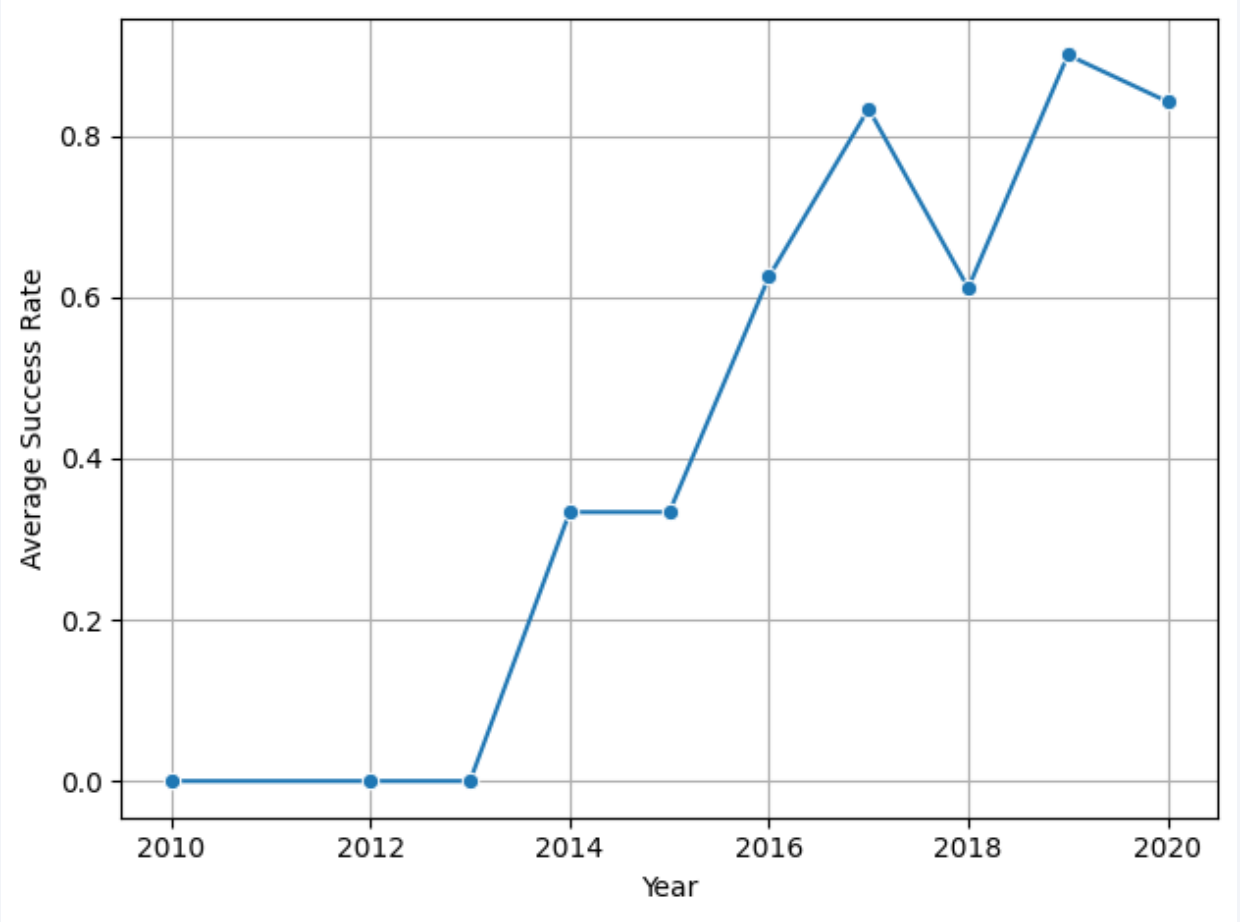
Payload vs. Orbit Type

- Both successes (Class 1) and failures (Class 0) occur across a wide range of payload masses for most orbit types, particularly LEO, ISS, PO, and GTO.
- High payload masses (above 10,000 kg) are rare and mostly associated with successful missions to VLEO and GEO orbits.
- Orbits like SSO, HEO, ES-L1, and SO have only successful launches, regardless of payload mass.
- There is no single payload mass range that guarantees launch success or failure, highlighting the importance of other contributing factors.



Launch Success Yearly Trend

- SpaceX's average launch success rate increased dramatically after 2013, showing rapid operational improvement.
- After 2015, success rates consistently remained above 60%, with several years exceeding 80%.
- The trend confirms strong organizational learning and reliability gains over time.



All Launch Site Names

- This SQL query retrieves the unique names of launch sites recorded in the mission database.
- The result lists all distinct launch site identifiers, confirming that launches occurred at multiple locations: CCAFS LC-40, VAFB SLC-4E, and KSC LC-39A.

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

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

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

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

Launch Site Names Begin with 'CCA'

- This SQL query filters and displays the first five launch records where the launch site name starts with “CCA,” specifically CCAFS LC-40.
- The resulting table shows detailed mission data—including date, booster version, payload, orbit, customer, and mission outcome.

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' Limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- This SQL query calculates the total payload mass (in kilograms) delivered by boosters on missions for NASA (CRS).
- The result shows that NASA (CRS) missions carried a combined payload mass of 45,596 kg, highlighting the substantial cargo capacity delivered by SpaceX for NASA's Commercial Resupply Services.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [12]:
%sql SELECT SUM("Payload_Mass__kg_") AS TOTAL_PAYLOAD
 * sqlite:///my_data1.db
Done.
Out[12]:
```

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

- This SQL query calculates the average payload mass delivered by the F9 v1.1 booster version.
- The result shows that, on average, each F9 v1.1 launch carried 2,928.4 kg of payload, providing a benchmark for the typical mission capacity of this booster model.

Display average payload mass carried by booster version F9 v1.1

In [13]:

```
%sql SELECT AVG("Payload_Mass__kg_") AS Avg_Payload_M
```

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

Out[13]:

Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

- This SQL query identifies the earliest date of a successful landing in the dataset.
- The result (2015-12-22) marks a milestone for SpaceX, representing the company's first ever successful booster landing.

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [15]:

```
%sql SELECT MIN(Date) AS First_Succesful_landing FROM
```

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

Out[15]:

First_Succesful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This SQL query filters boosters that achieved a successful drone ship landing and carried payloads between 4,000 and 6,000 kg.
- The result lists four specific booster versions (F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2) highlighting which boosters met these precise mission criteria.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [16]:

```
%sql SELECT "Booster_version" FROM SPACEXTBL WHERE "Landing_(
```

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

Out[16]:

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

Total Number of Successful and Failure Mission Outcomes

- This SQL query counts the total number of missions by outcome type.
- The results show that the vast majority were successful (98 missions), with only a few failures or unclear outcomes, highlighting SpaceX's high mission success rate in the dataset.

List the total number of successful and failure mission outcomes

```
In [17]:
%sql SELECT "Mission_Outcome", COUNT(*) AS TOTAL FROM SPACEXTE
 * sqlite:///my_data1.db
Done.
Out[17]:
```

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

Boosters Carried Maximum Payload

- This SQL query lists all booster versions that carried the maximum payload mass recorded in the dataset, which is 15,600 kg.
- Multiple boosters, all in the F9 B5 series, achieved this feat, highlighting their capability for maximum cargo delivery in SpaceX launches.

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
Payload_Mass__kg_" FROM SPACEXTBL WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTBL);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- This SQL query extracts records from 2015 that had failed drone ship landings, showing the launch month, booster version, and launch site.
- The result reveals that there were two such failures at CCAFS LC-40, in January (F9 v1.1 B1012) and April (F9 v1.1 B1015), highlighting specific challenges SpaceX faced with drone ship recoveries that year.

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
In [19]:
%sql SELECT SUBSTR(Date, 6, 2) AS Month, "Landing_Outcome", "Booster_V
 * sqlite:///my_data1.db
Done.
Out[19]:
```

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

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

- This query ranks all landing outcomes between June 2010 and March 2017.
- Most launches made no landing attempt (10), but there were an equal number of successes and failures on drone ships (5 each).
- Ground pad landings had 3 successes, while ocean landings were a mix of controlled (3), uncontrolled (2), and parachute failures (2).

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

```
In [20]:
%sql SELECT "Landing_Outcome", Count(*) AS Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN '
 * sqlite:///my_data1.db
Done.
Out[20]:
```

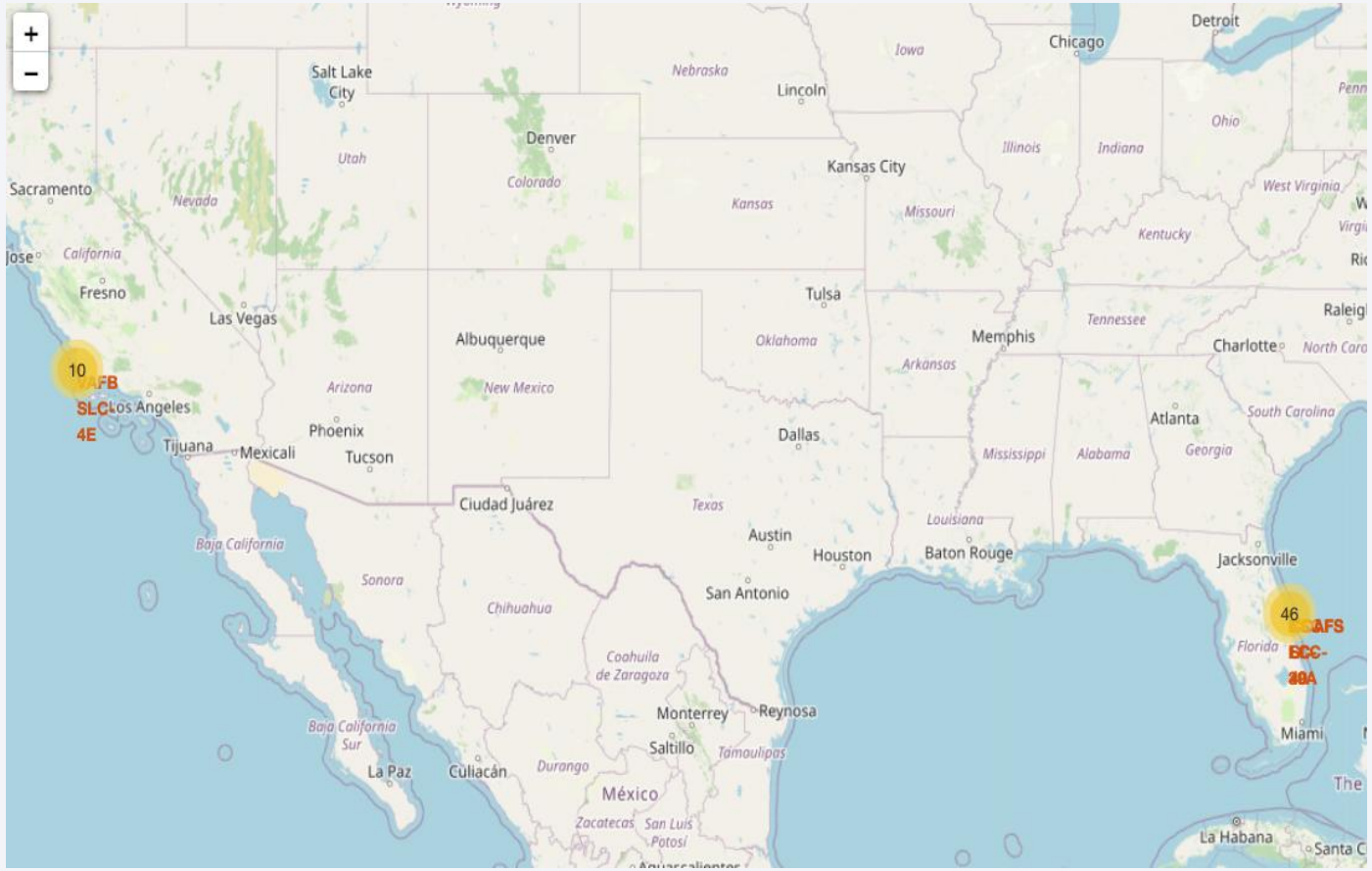
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

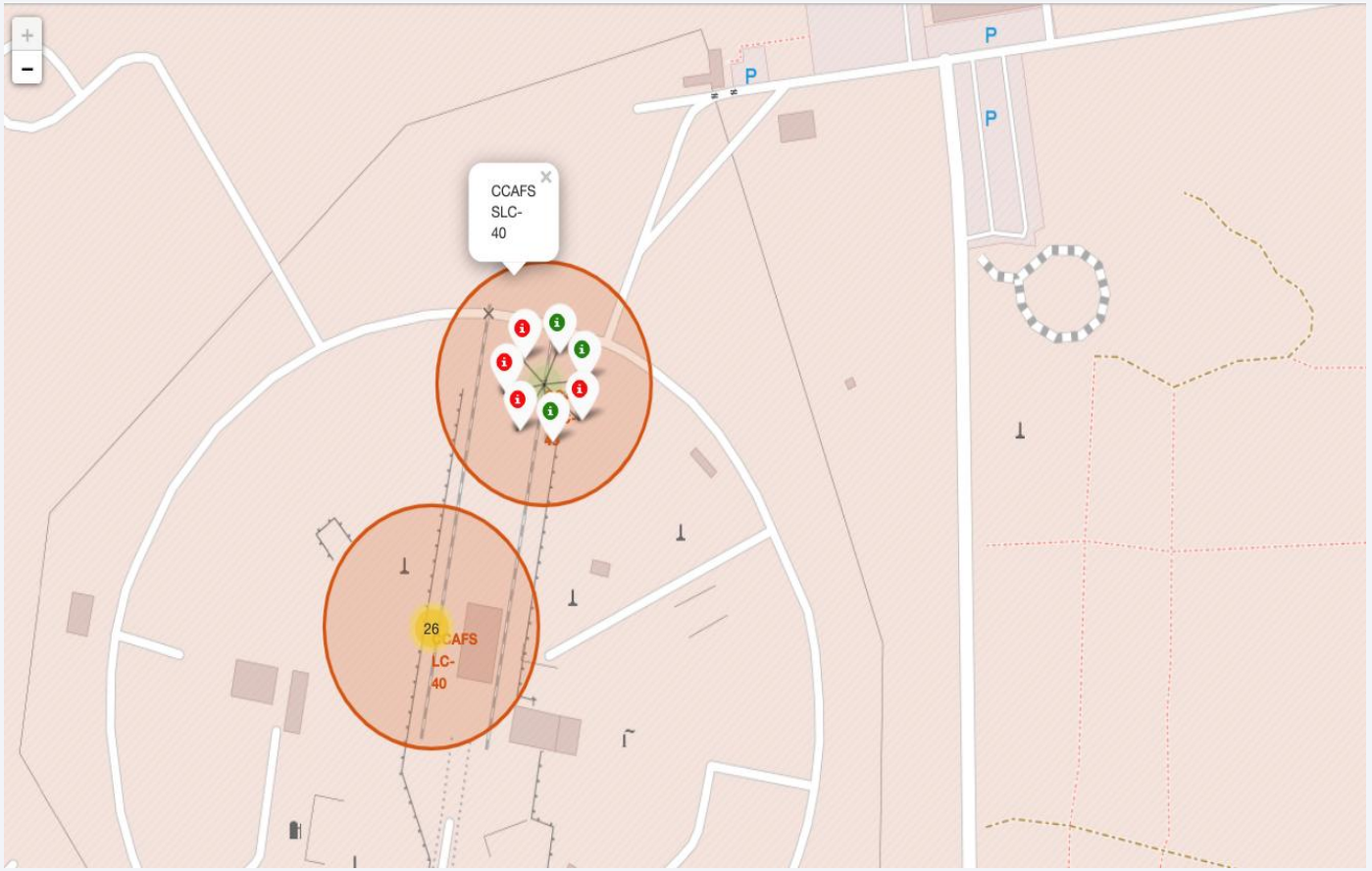
SpaceX Launch Sites Mapped Across the United States

- SpaceX launches are concentrated at three coastal sites, enabling safer rocket trajectories over the ocean and easier booster recovery.
- Florida hosts the majority of launches, reflecting its dominance as the main U.S. spaceport region.
- The wide geographic separation between launch sites (East vs. West coast) demonstrates SpaceX's flexibility in serving different orbital requirements and customer needs.



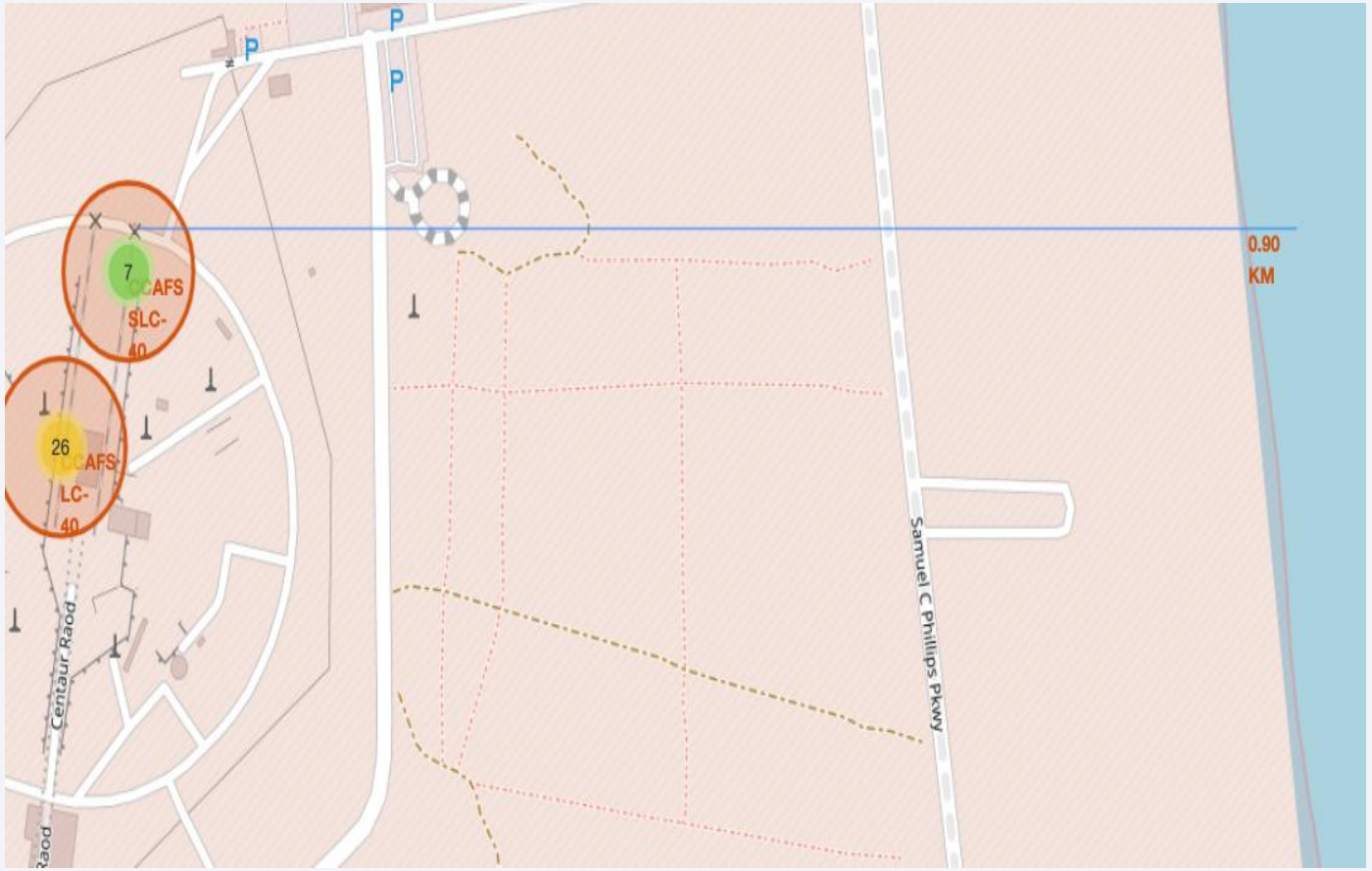
Micro-Cluster Analysis of Launch Outcomes at CCAFS SLC-40

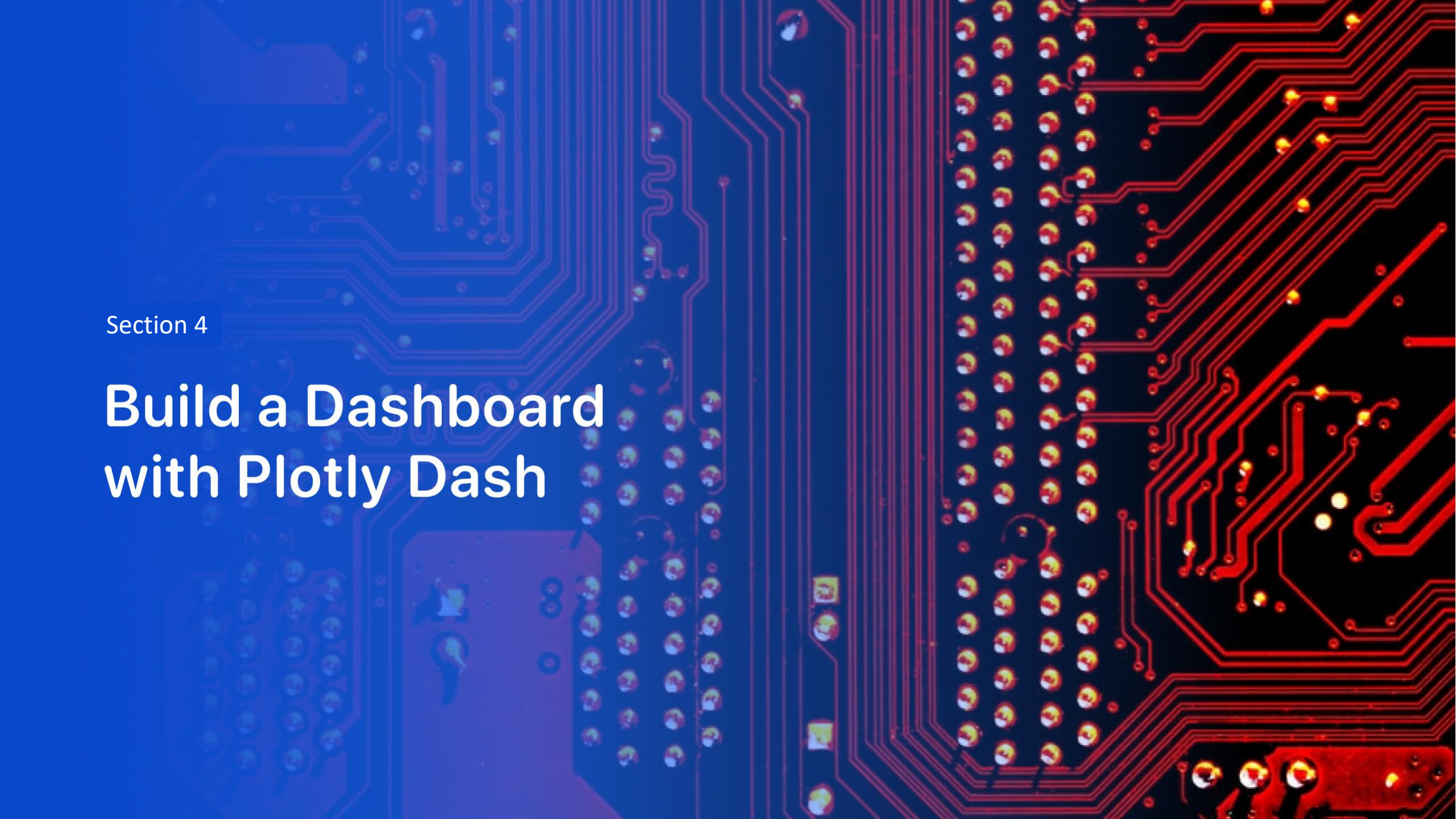
- The map zooms in on CCAFS SLC-40, showing the exact pad location and spatial grouping of launch outcome markers (success in green, failure in red).
- Clusters indicate frequent, closely spaced launches at this site, highlighting operational density.
- Immediate visual access to individual launch outcomes enables fast anomaly detection and site performance evaluation.



Proximity of CCAFS SLC-40 Launch Site to Coastline

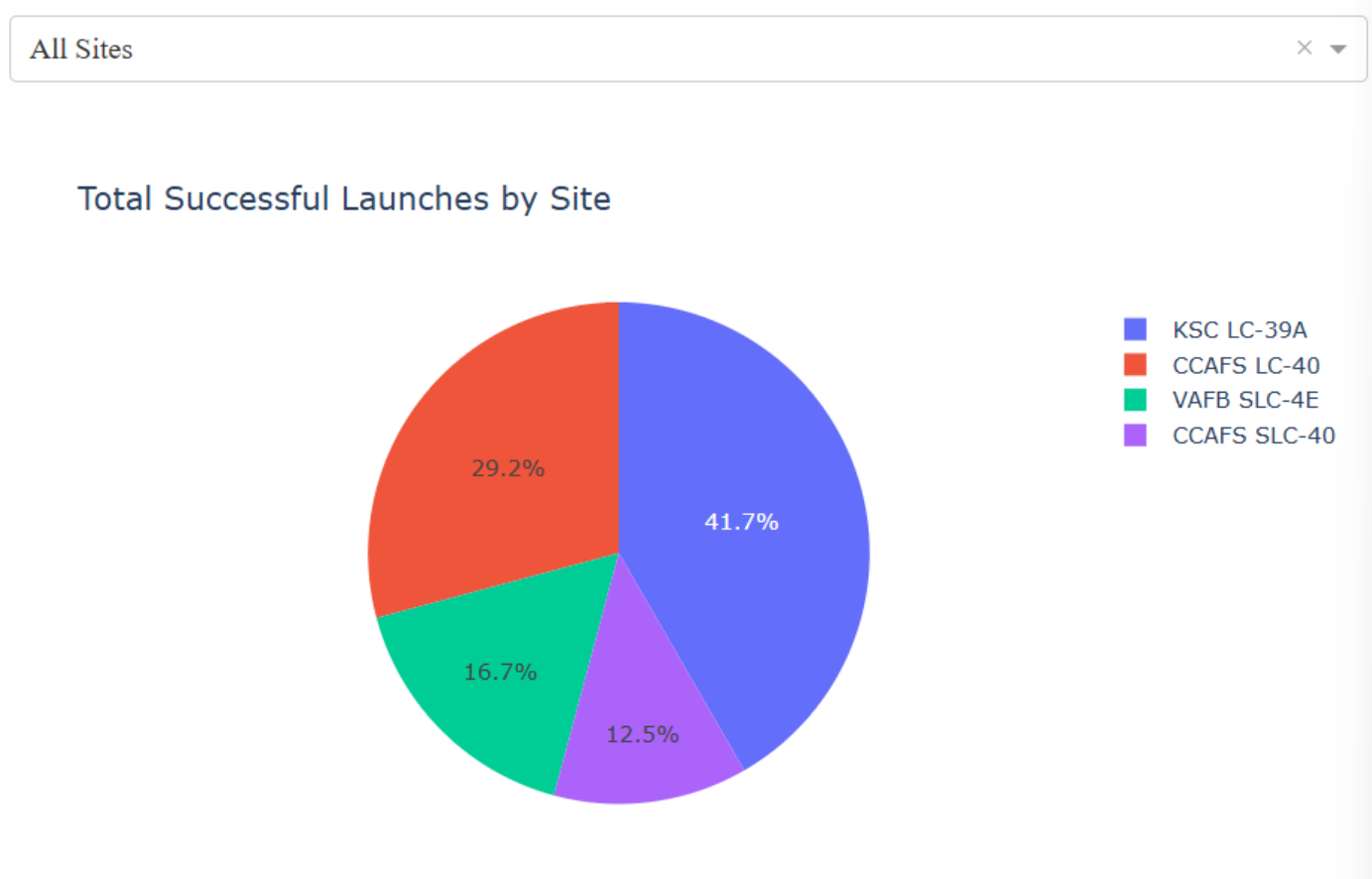
- The blue line highlights the direct distance from the CCAFS SLC-40 launch site to the nearest coastline, measuring just 0.9 km.
- This extremely close proximity to the ocean minimizes risks to populated areas and enables quick rocket stage recovery at sea.
- Visualizing site-to-coast distances helps optimize launch safety protocols and emergency response planning.




Section 4
Build a Dashboard with Plotly Dash

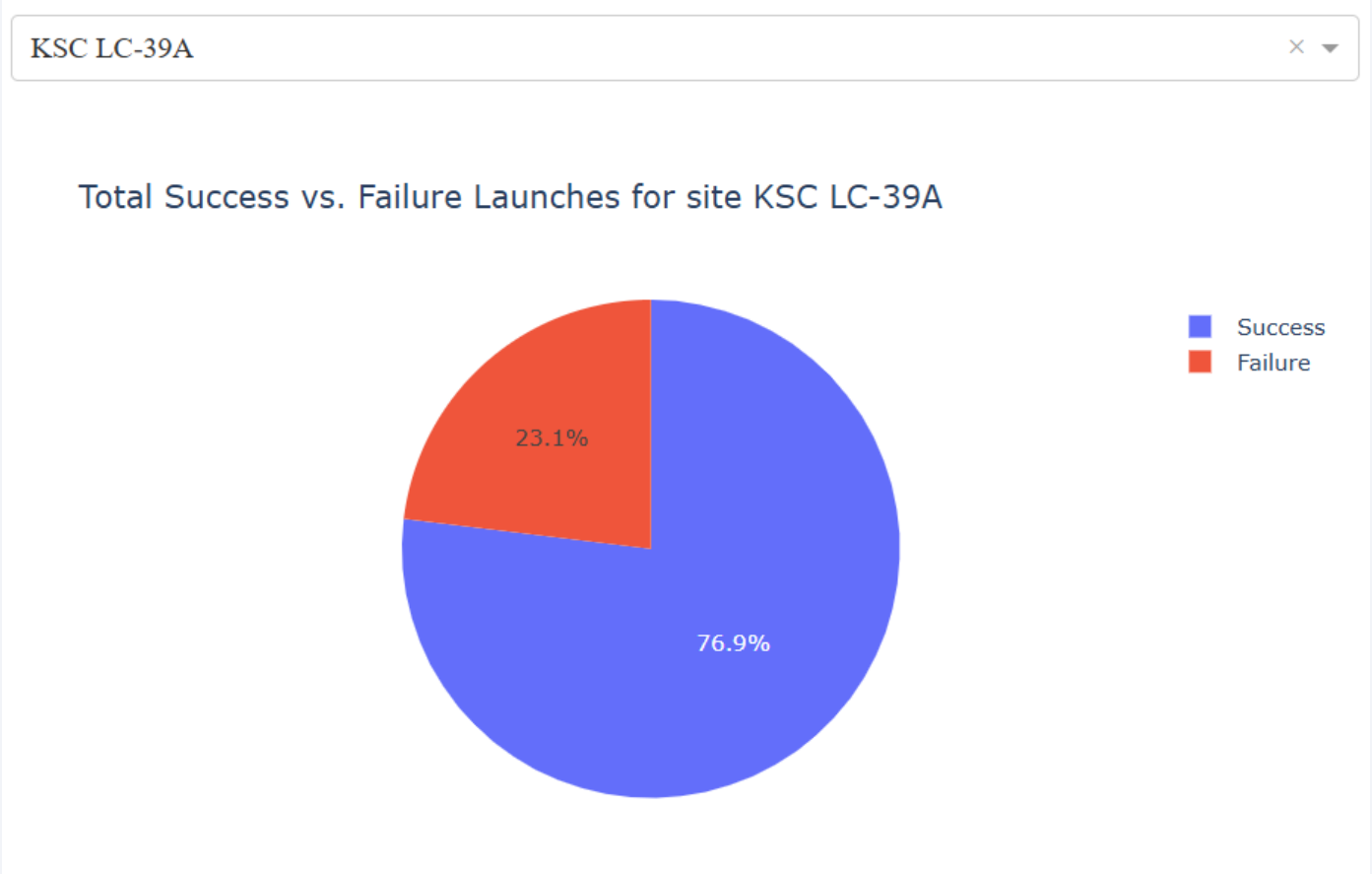
Distribution of Successful Launches by Site

- KSC LC-39A dominates with the highest share of successful launches (41.7%), highlighting its critical role in SpaceX operations.
- CCAFS LC-40 also contributes significantly (29.2%), indicating it's another major hub for SpaceX.
- VAFB SLC-4E (16.7%) and CCAFS SLC-40 (12.5%) have fewer successful launches, suggesting either fewer attempts or different mission profiles.



Success vs. Failure Rate at KSC LC-39A

- KSC LC-39A achieves a high success rate, with nearly 77% of launches succeeding.
- Only 23% of launches from this site ended in failure, making it the most reliable launch site in this analysis.
- The dashboard filter allows a site-specific breakdown, making it easy to compare launch site performance individually.



Correlation Between Payload Mass and Launch Success Across All Sites

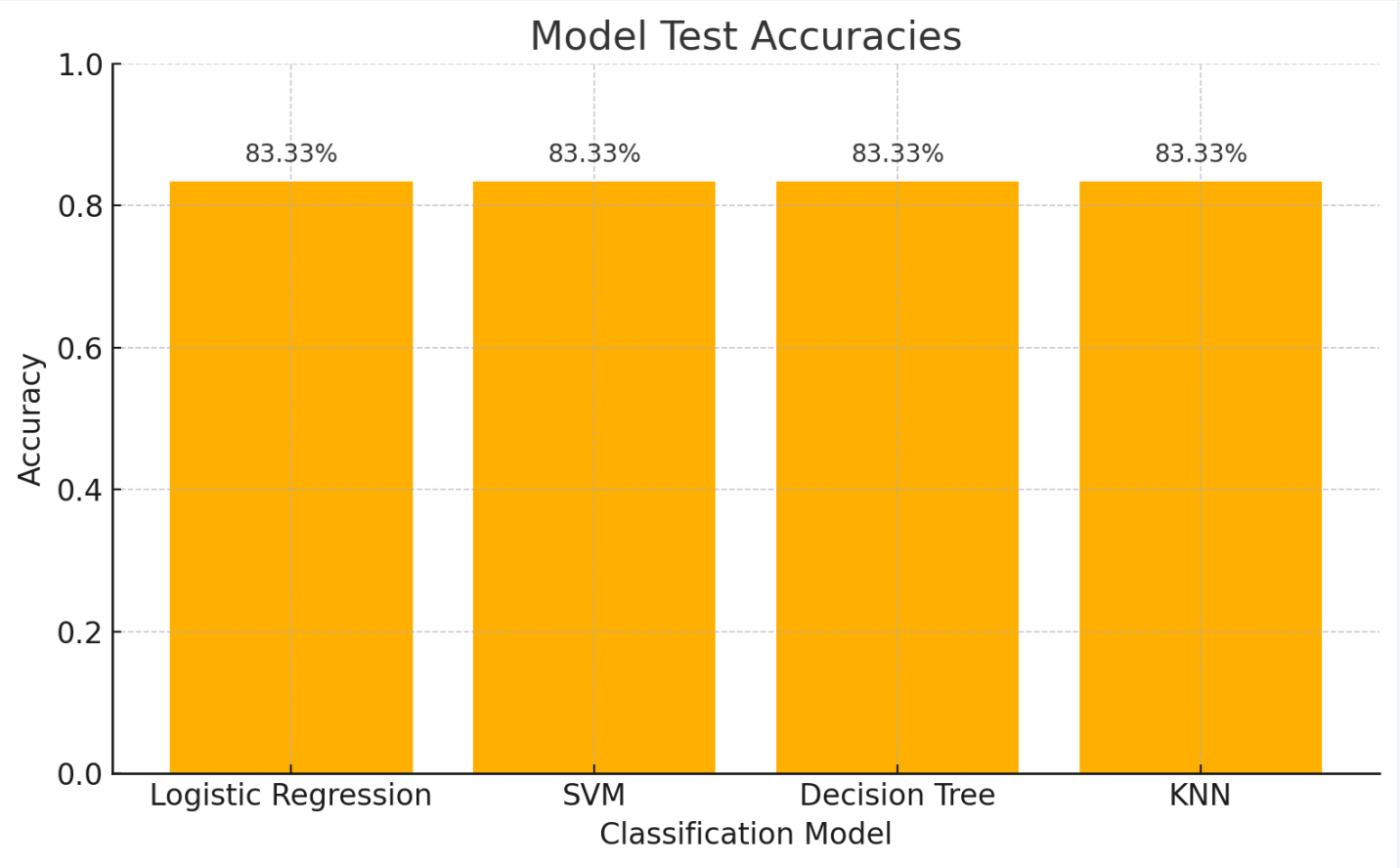
- The scatter plot shows the relationship between payload mass (kg) and launch outcome (1 = Success, 0 = Failure) for different booster versions.
- Most successful launches (class = 1) occur in the lower to mid payload range (up to ~6000 kg), with FT boosters showing the highest concentration of successes.
- Failures (class = 0) are more common with earlier booster versions (v1.0, v1.1) and tend to cluster at lower payloads.

Section 5

Predictive Analysis (Classification)

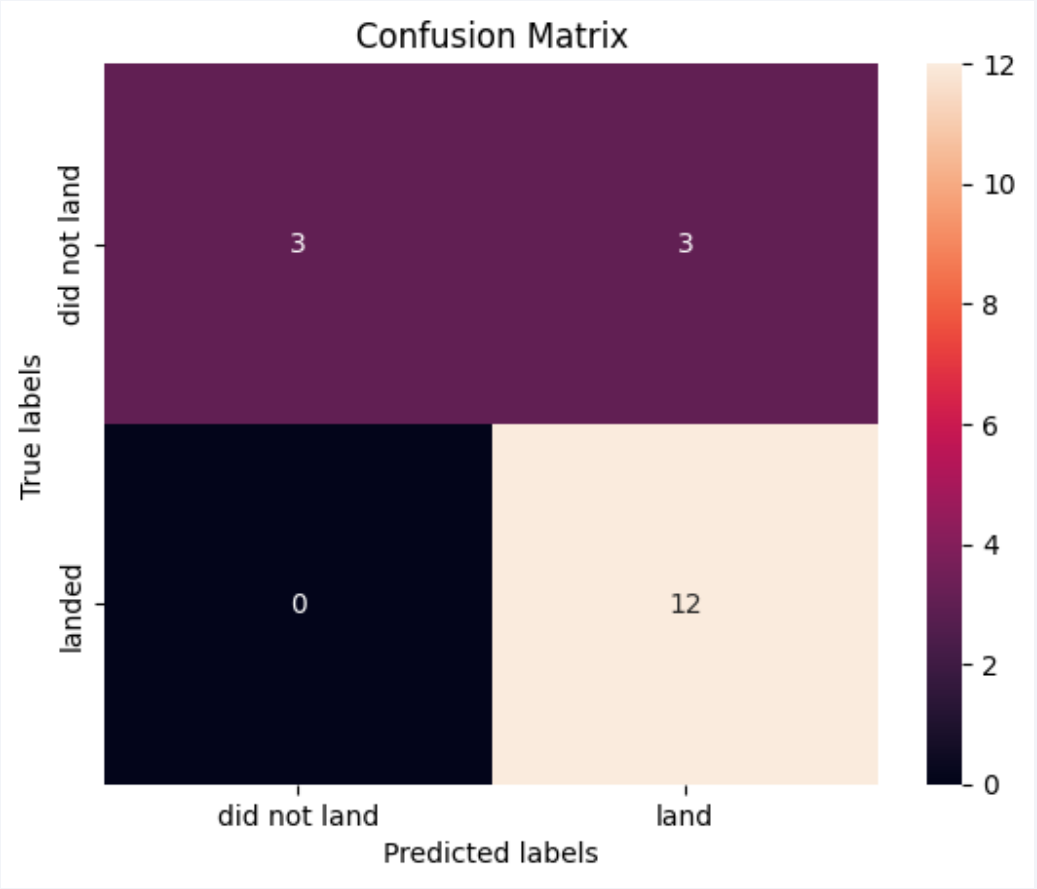
Classification Accuracy

- No single model outperformed the others. All of them reached identical accuracy on the test set, which means they performed equally well on this classification task.



Confusion Matrix

- Zero missed landings: Every actual successful landing was correctly predicted—no false negatives.
- Some overconfident calls: The model falsely predicted "landed" three times when the rocket actually failed.
- High accuracy: 83% overall—most launches are classified correctly.
- Consistent results: All models performed exactly the same, choice comes down to simplicity or explainability—not performance.



Conclusions

- Performance:
  - All four classification models (Logistic Regression, SVM, Decision Tree, and KNN) achieved the same accuracy (83.33%) on the test set, meaning no model outperformed the others in this scenario.
- Prediction Behavior:
  - The confusion matrix reveals zero false negatives (no missed landings), but some false positives (predicting land when it didn't actually happen), indicating the models were a bit “optimistic” but reliably identified successful landings.
- Interpretability Matters:
  - Since all models performed equally well, the best model to choose would be the one with the highest interpretability (Logistic Regression) making it ideal for actionable business insights and easier explanation to stakeholders.

Thank you!