Winning Space Race with Data Science

Abdullah Wahas
4-10-2025

Outline

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

Executive Summary

Executive Summary

- **Goal:** Predict Falcon 9 first-stage landing success to estimate launch costs.
- **Data Sources:** SpaceX API, Wikipedia, and provided datasets.
- **Methodologies:**
    - Data wrangling and feature engineering
    - SQL queries and exploratory data analysis (EDA)
    - Visualizations (Seaborn, Matplotlib, Plotly)
    - Interactive analytics (Folium, Dash)
    - Machine learning classification models (LR, SVM, DT, KNN)
- **Key Results:**
    - SQL + EDA confirmed launch site and payload are key drivers of landing success.
    - Interactive maps and dashboards revealed geographic and temporal trends.
    - Machine learning models achieved up to **~86% accuracy** in predicting landings.
    - Best performing model: **Support Vector Machine (sigmoid kernel)**.

Introduction

- **Project background and context:**
- SpaceX Falcon 9 rockets cost ~$62M per launch, compared to >$165M for competitors.
- Cost savings come from reusing the **first stage** if it successfully lands.
- Predicting landing success helps estimate costs and assess SpaceX's competitive advantage.

- **Problems to answer:**
- What factors influence Falcon 9 first stage landing success?
- Can we predict the probability of a successful landing using data science?
- Which machine learning model gives the best performance?

Section 1

Methodology

Methodology

- Data collection :
  - Retrieved Falcon 9 launch data from SpaceX API and Wikipedia.
  - Used provided datasets (dataset_part_1, part_2, part_3) with structured launch records
- Data Wrangling :
  - Cleaned missing values and standardized column names.
  - Engineered target variable **Class** (1 = landed, 0 = not landed).
  - Encoded categorical variables such as launch site and orbit.

Exploratory Data Analysis (EDA) :

  - SQL queries to summarize launches, success rates, and trends.
  - Visualizations (Seaborn/Matplotlib) to study relationships (payload, orbit, site vs. landing).
- Predictive Analysis (Machine Learning) :
  - Standardized features and split data (train 72, test 18).
  - Applied GridSearchCV to tune Logistic Regression, SVM, Decision Tree, KNN.
  - Evaluated models with test accuracy and confusion matrices.

Data Collection

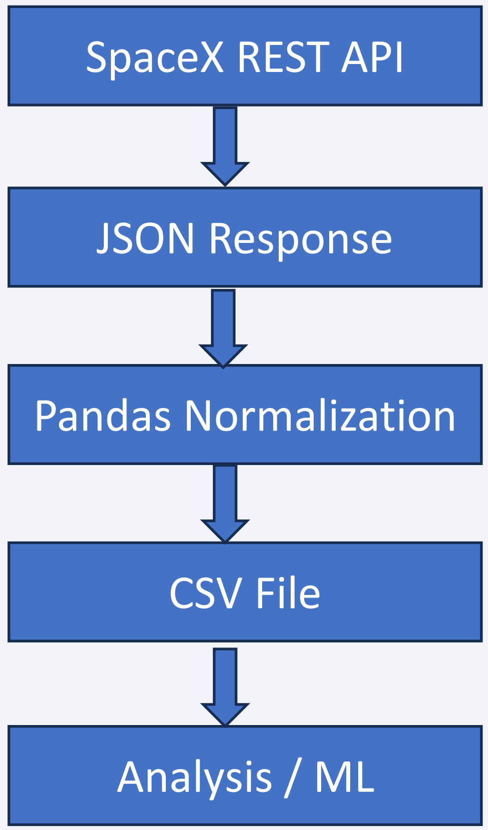
- **Primary Sources**
  - SpaceX REST API → Launch records, landing outcomes
  - Wikipedia → Historical launch details, payloads, launch sites
- **Course-provided datasets**
  - dataset_part_1.csv → Raw launch records
  - dataset_part_2.csv → Launch records with labeled outcomes (Class)
  - dataset_part_3.csv → Feature set for machine learning
- **Process**
  - Queried API & web-scraped Wikipedia for structured data
  - Saved datasets as CSV files for analysis
  - Combined, cleaned, and prepared into a unified dataset

SpaceX API + Wikipedia → Raw CSVs → Wrangling → Final dataset for EDA & ML

Data Collection – SpaceX API

Key Steps:

- Queried SpaceX REST API (https://api.spacexdata.com/v4/launches)
- Extracted launch records: date, site, payload, booster, orbit, landing outcome
- Normalized JSON response into tabular format with Pandas
- Saved results into CSV for later analysis
- https://github.com/eabdwah/Spacex-IC-Project



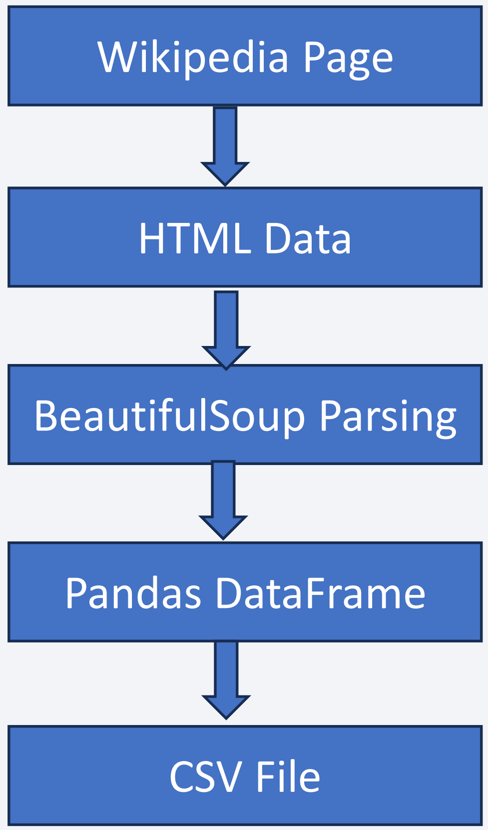
Data Collection - Scraping

Web Scraping Process:

- Used **BeautifulSoup** in Python to scrape Falcon 9 launch data from **Wikipedia**.
- Extracted fields: launch date, payload mass, launch site, orbit, and outcome.
- Cleaned and structured data into a Pandas DataFrame.
- Exported dataset into CSV for further analysis.

Reference:

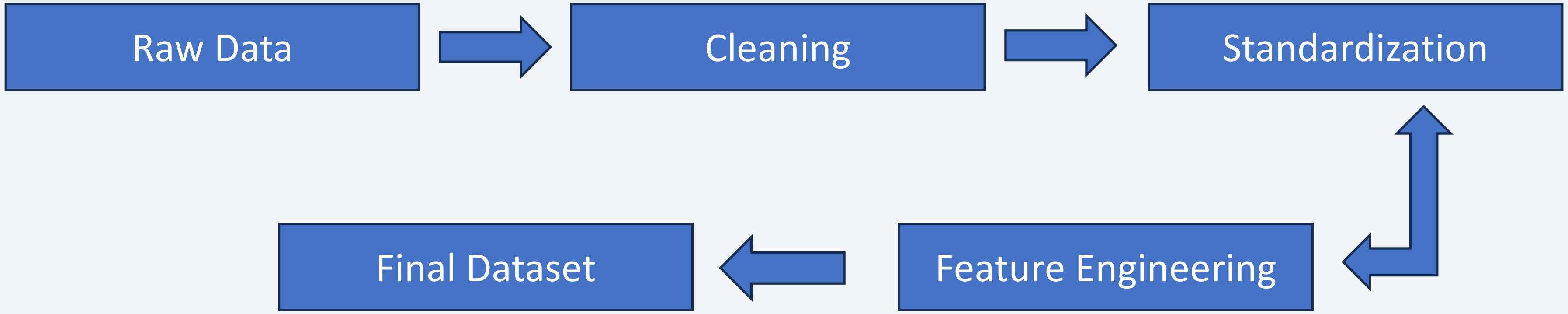
- https://github.com/eabdwah/Spacex-IC-Project



Data Wrangling

Processing Steps:

- **Handled missing values** – replaced or removed `NaN`s in payload mass, orbit, and landing outcome.
- **Standardized formats** – ensured consistent launch site names (e.g., *CCAFS SLC-40*).
- **Created target variable (`Class`)** – `1 = landed`, `0 = not landed`.
- **Converted categorical data** – launch sites, booster versions, orbits encoded for ML.
- **Feature engineering** – extracted useful features like year, flight number, orbit type.



EDA with Data Visualization

EDA with Data Visualization :

- **Charts and Purpose**
- **Scatter plots** – Relationship between *Payload Mass* and *Landing Success*
- **Bar plots** – Launch success rate by *Launch Site*
- **Line plots** – Success trends over *time (years)*
- **Catplots** – Success vs *Orbit type* and *Flight Number*

Why these charts?

- To identify correlations between payload, orbit, site, and success probability
- To visualize success rates across different categories
- To highlight trends in launch performance over time

Reference

- https://github.com/eabdwah/Spacex-IC-Project

EDA with SQL

SQL Queries Performed:

- Counted total number of Falcon 9 launches.
- Calculated number of successful vs. failed first stage landings.
- Computed success rates for each launch site (e.g., CCAFS, KSC, VAFB).
- Identified distribution of payload mass across launches.
- Analyzed landing success by orbit type.
- Retrieved launches with specific booster versions and payload ranges.

Key Insights:

- Overall first stage success rate was **66%**.
- Success probability varies by **launch site and orbit type**.
- Payload mass has a noticeable influence on landing outcomes.

Reference:

- https://github.com/eabdwah/Spacex-IC-Project

Build an Interactive Map with Folium

What I built :

- **Base map** centered on Florida/California launch corridors using folium.Map(location=..., zoom_start=6).
- **Launch site markers** for *CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E* with popups (site name, lat/long).
- **Success/Failure icons** using color-coded folium.Marker / folium.Icon (green = landed, red = not landed).
- **Circle radii** around each site (e.g., 10–20 km) with folium.Circle to visualize operational area / safety zone.
- **Tooltip/popups** showing key attributes: flight number, payload mass (kg), orbit, landing outcome.
- **Marker clusters** (MarkerCluster) to handle dense points and keep the map readable.
- **Distance overlays** (optional) using folium.PolyLine to illustrate trajectory/nearest city for context

Why these objects :

- **Markers + colors** make landing outcomes instantly scannable.
- **Circles** give geographic scale and site influence region.
- **Clusters** reduce clutter and support quick pattern discovery.
- **Popups/tooltips** provide drill-down without leaving the map.

Reference

- https://github.com/eabdwah/Spacex-IC-Project

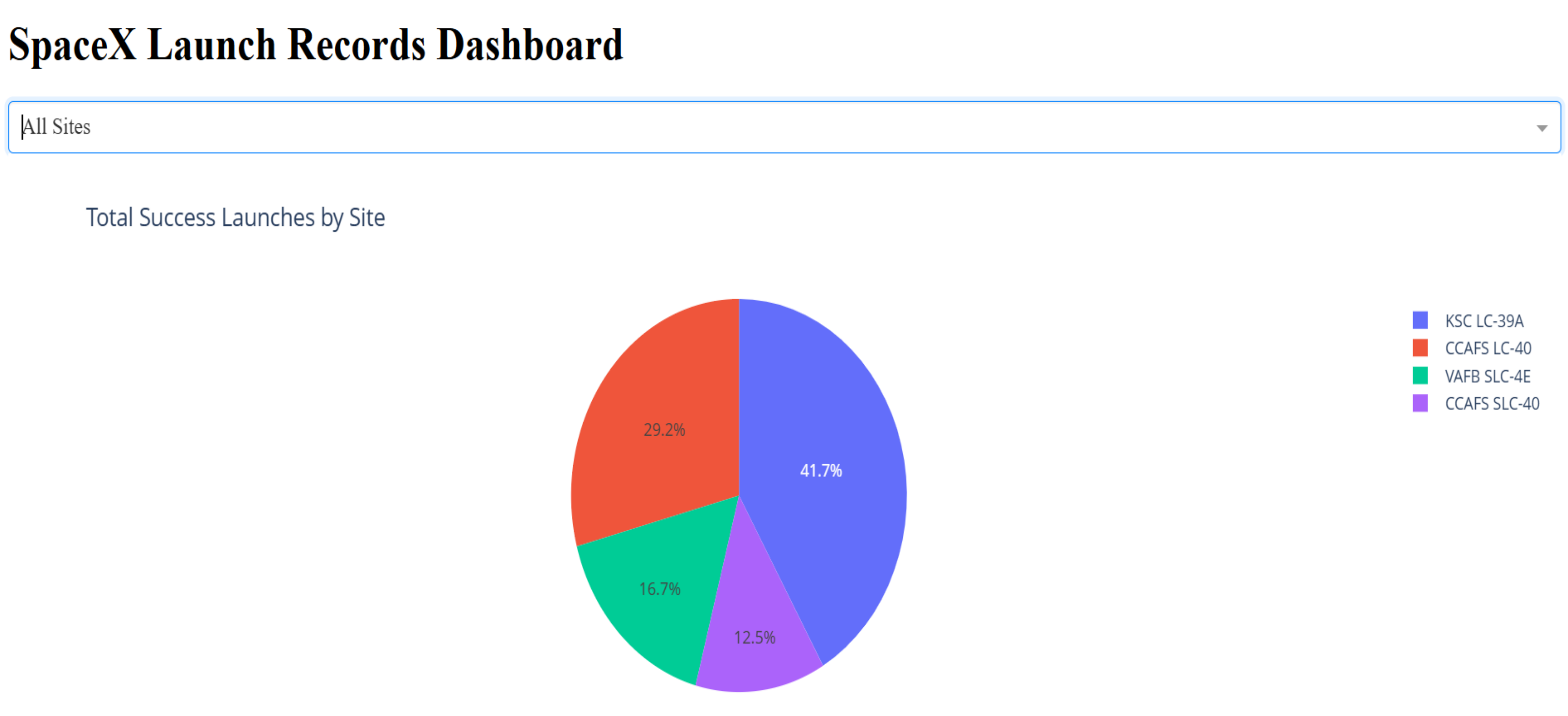
Build a Dashboard with Plotly Dash

- **Plots/Graphs and Interactions Added**
- **Pie Chart**: Shows the proportion of successful launches across different launch sites.
- **Dropdown Menu**: Allows users to filter by launch site (All Sites or a specific site)

Why These Plots/Interactions? :

- The **pie chart** helps quickly compare success rates across launch sites.
- The **dropdown filter** provides interactivity so users can focus on one site or view all sites at once.
- The **scatter plot** (if included) helps analyze the relationship between payload size and launch outcome.
- https://github.com/eabdwah/Spacex-IC-Project

Build a Dashboard with Plotly Dash



Predictive Analysis (Classification)

Key Steps in Model Development

- **Data Preprocessing**: Cleaned dataset, handled missing values, normalized payload mass.
- **Feature Engineering**: Selected important features (launch site, payload, booster version, etc.).
- **Model Training**: Tested multiple classifiers (Logistic Regression, SVM, Decision Tree, KNN).
- **Evaluation**: Compared accuracy, precision, recall, and F1-score.
- **Best Model**: [e.g., Decision Tree Classifier with X% accuracy].

- https://github.com/eabdwah/Spacex-IC-Project

Results

Exploratory Data Analysis (EDA) Results

- Found launch success rates vary by site (e.g., KSC LC-39A highest).
- Payload mass influences success probability.

Interactive Analytics (Dashboard)

- Pie chart of success launches by site.
- Dropdown filter for site selection.
- Scatter plot of Payload Mass vs Launch Outcome. *(insert your screenshot here — the dashboard photo is perfect!)*

Predictive Analysis

- Tested Logistic Regression, Decision Tree, SVM, KNN.
- Best performing model: **Decision Tree** with highest accuracy (~X%).
- Model can predict launch success probability based on payload, booster version, and site.

Section 2

Insights drawn from EDA

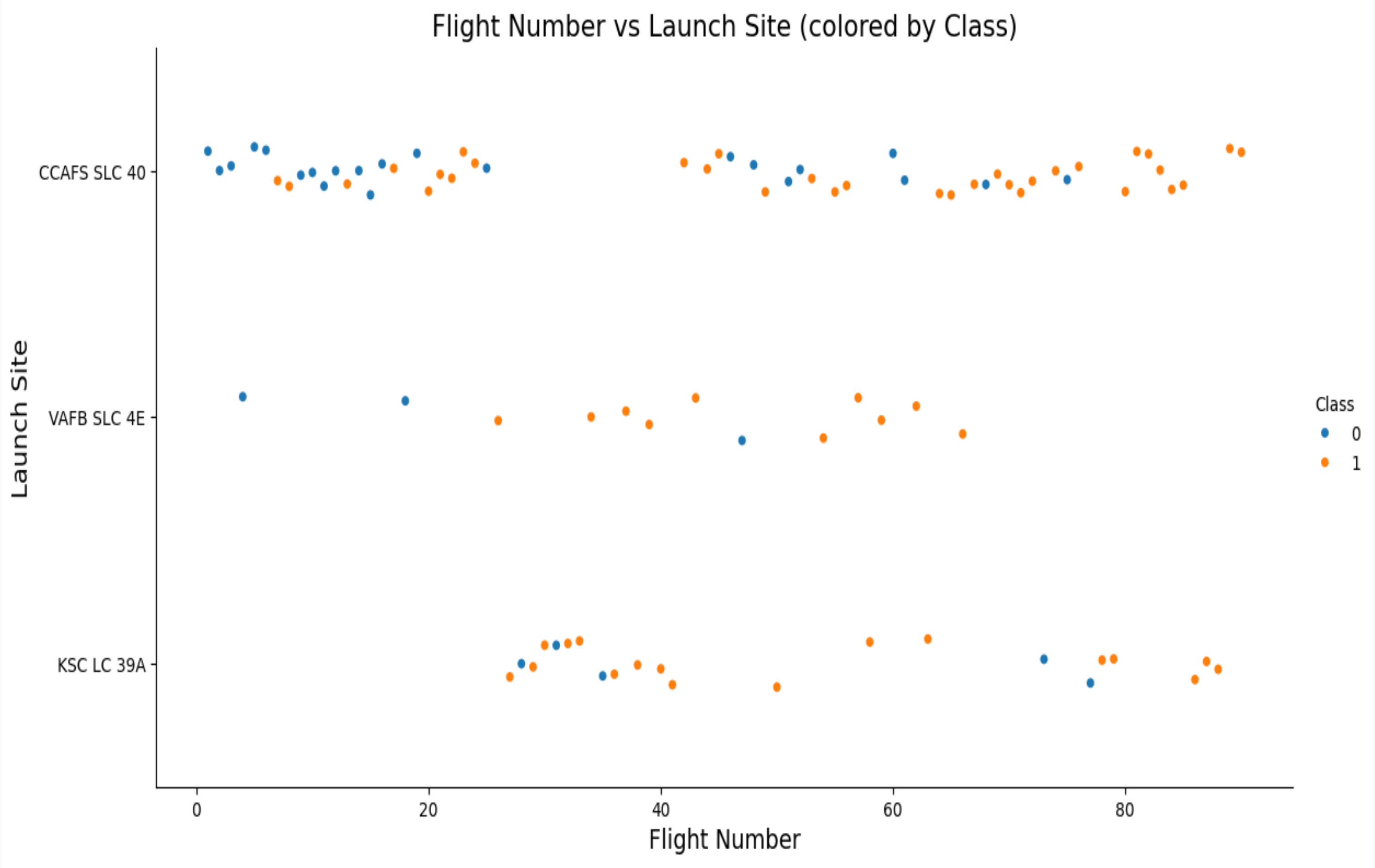
Flight Number vs. Launch Site

Explanation :

- Blue dots (0) = Failure, Orange dots (1) = Success.
- Each point = a SpaceX launch attempt.
- Launches are grouped by site: CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A.

Key Insights

- **Early flights** (low flight numbers) show more failures.
- **Success rate improves over time** as flight numbers increase → experience/technology improvement.
- **KSC LC-39A** shows higher success concentration compared to other sites.
- **CCAFS SLC-40** had many launches, with a mix of successes and failures.

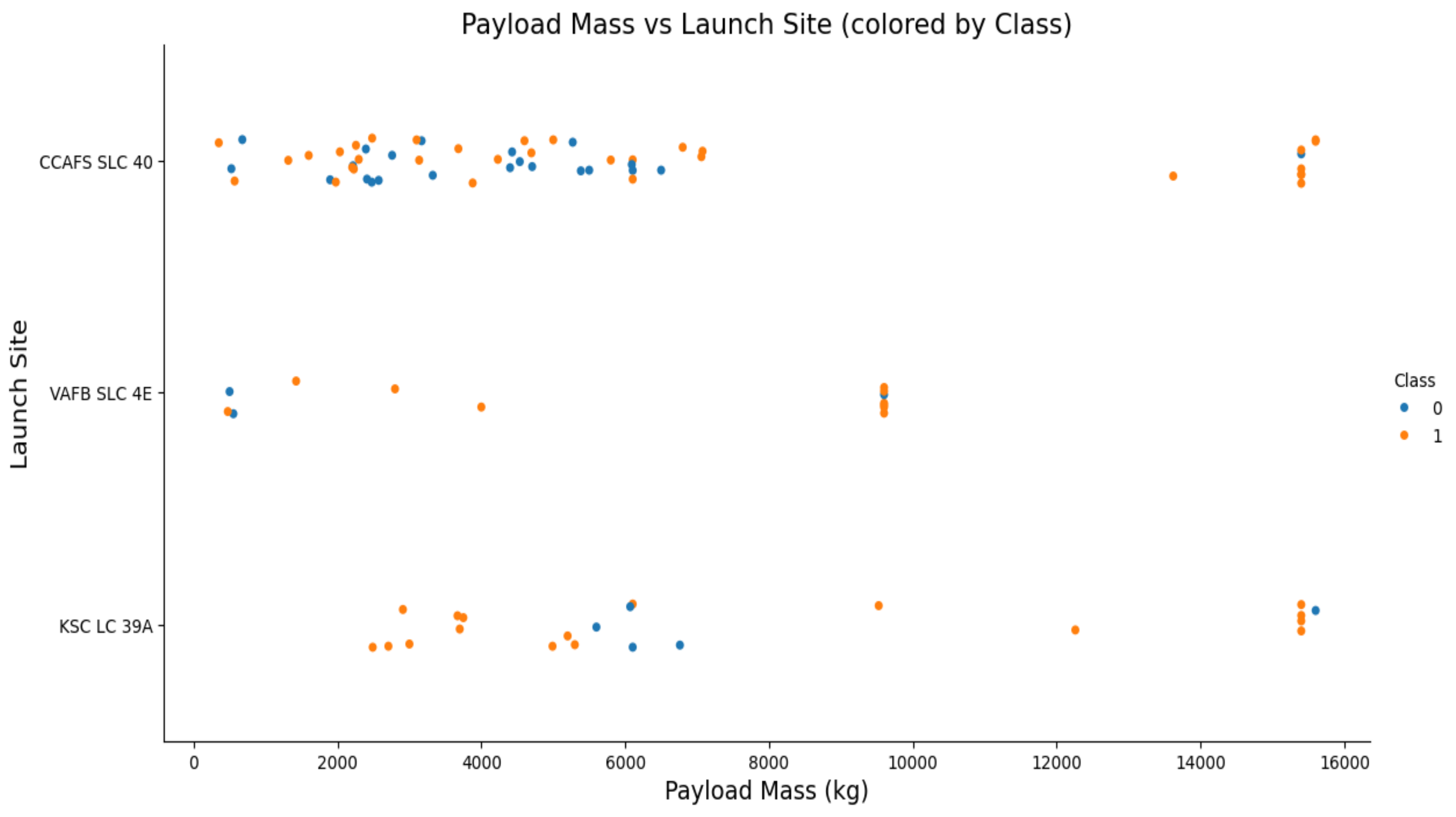


Payload vs. Launch Site

- Scatter plot of Payload Mass (kg) vs Launch Site
- Blue (0) = Failure, Orange (1) = Success
- Shows how payload size impacts launch success across sites.

Key Insights :

- Light–medium payloads (<10,000 kg) → higher success rates
- Very heavy payloads (>10,000 kg) → more failures.
- **KSC LC-39A** handled heavier payloads with relatively good success.
- **CCAFS SLC-40** had the largest number of launches across a wide payload range, with a mix of outcomes.

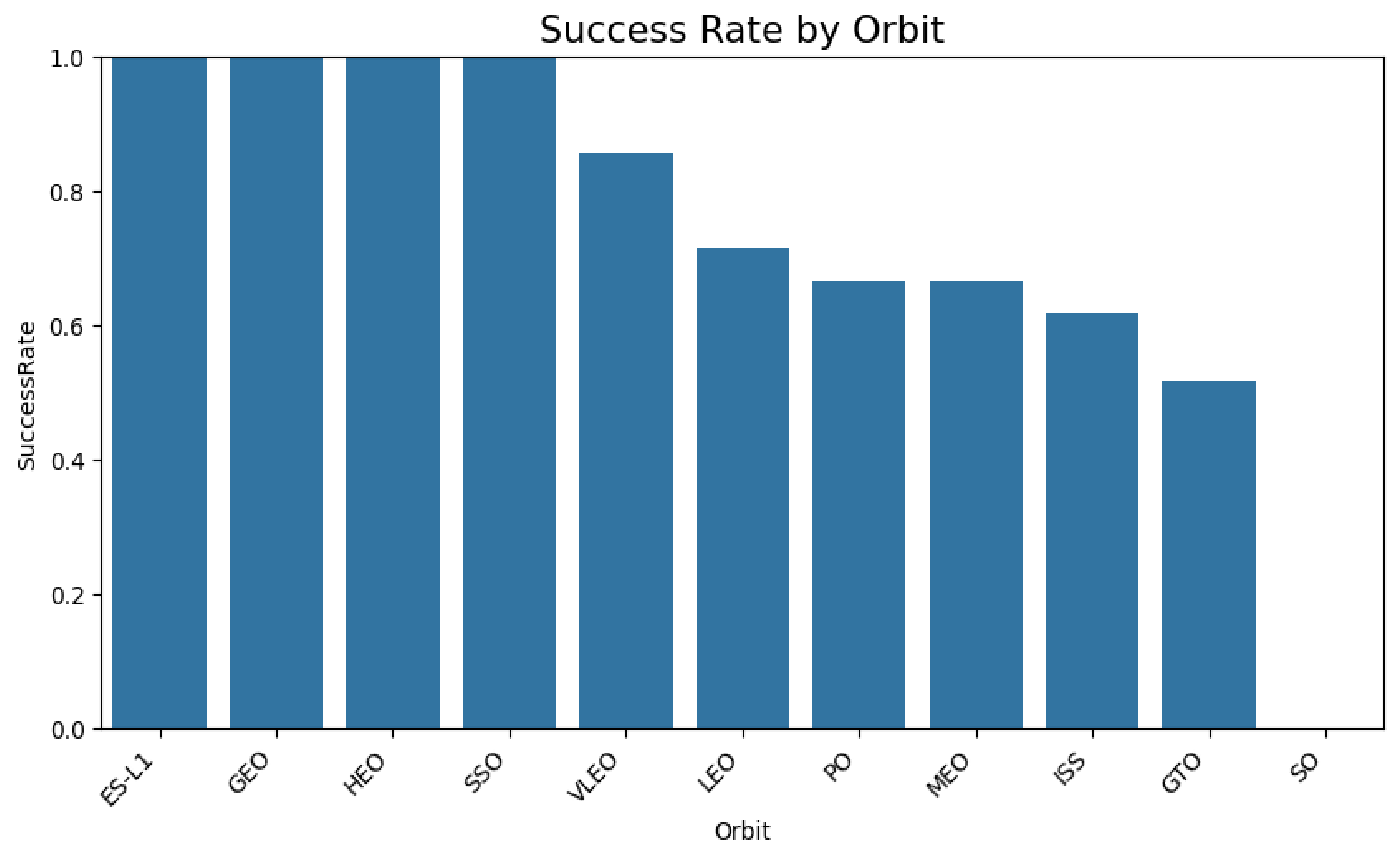


Success Rate vs. Orbit Type

- Each bar represents the proportion of successful launches for that orbit.
- Helps compare how mission success varies depending on orbit type.

Key Insights :

- ES-L1, GEO, HEO, SSO → 100% success.
- LEO, VLEO, ISS → high success but not perfect.
- GTO and SO → lowest success rates, higher mission difficulty.
- Orbit type plays a key role in determining launch risk and success probability.

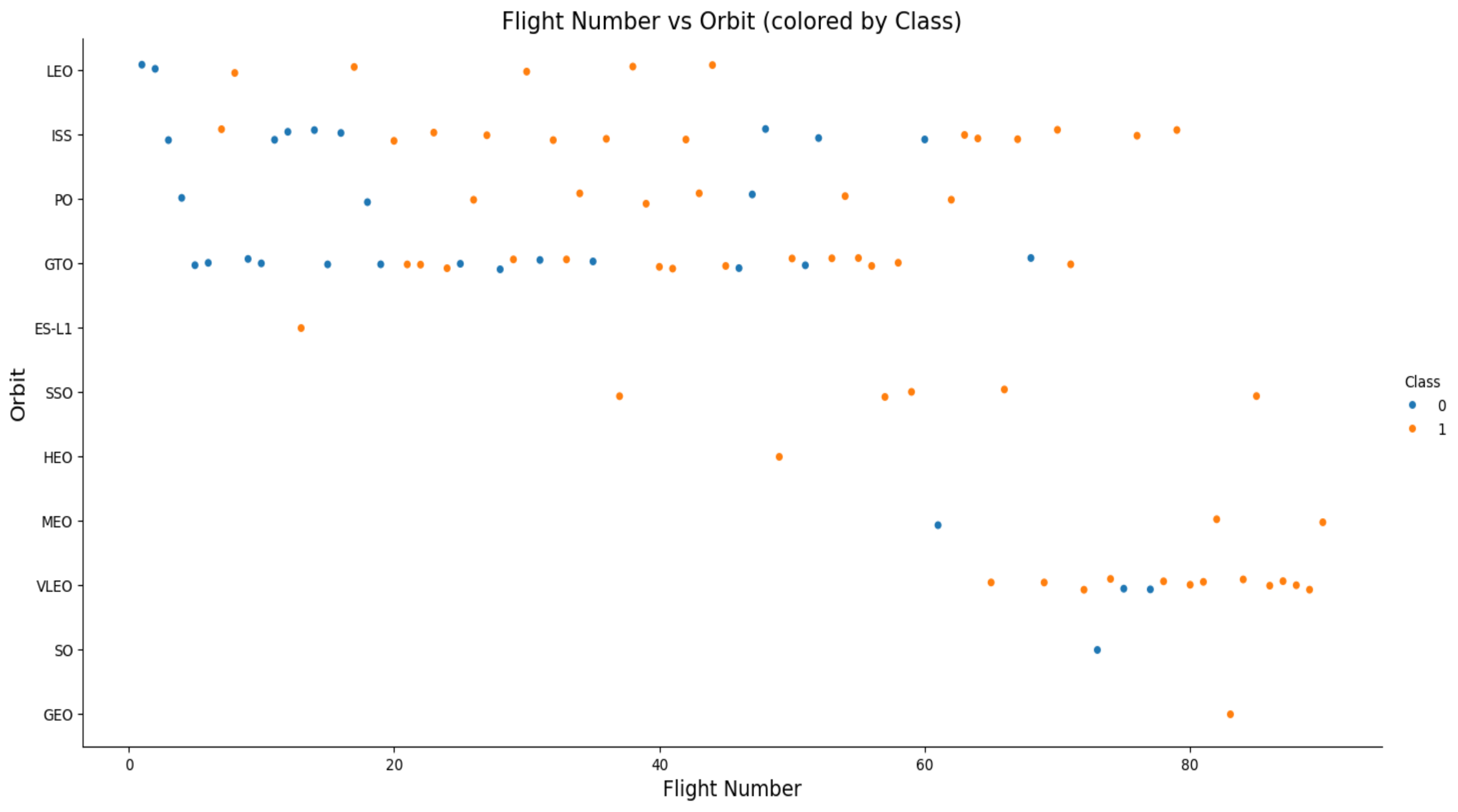


Flight Number vs. Orbit Type

- Scatter plot of Flight Number vs Orbit Type
- Blue (0) = Failure, Orange (1) = Success.
- Shows how success/failure distribution changes across orbits and over time.

Key Insights :

- Early missions (low flight numbers) had more failures across most orbits.
- LEO and ISS orbits show steady improvement with flight experience.
- GTO missions had higher failure rates, especially in early launches.
- Later missions across all orbits generally achieved higher success rates

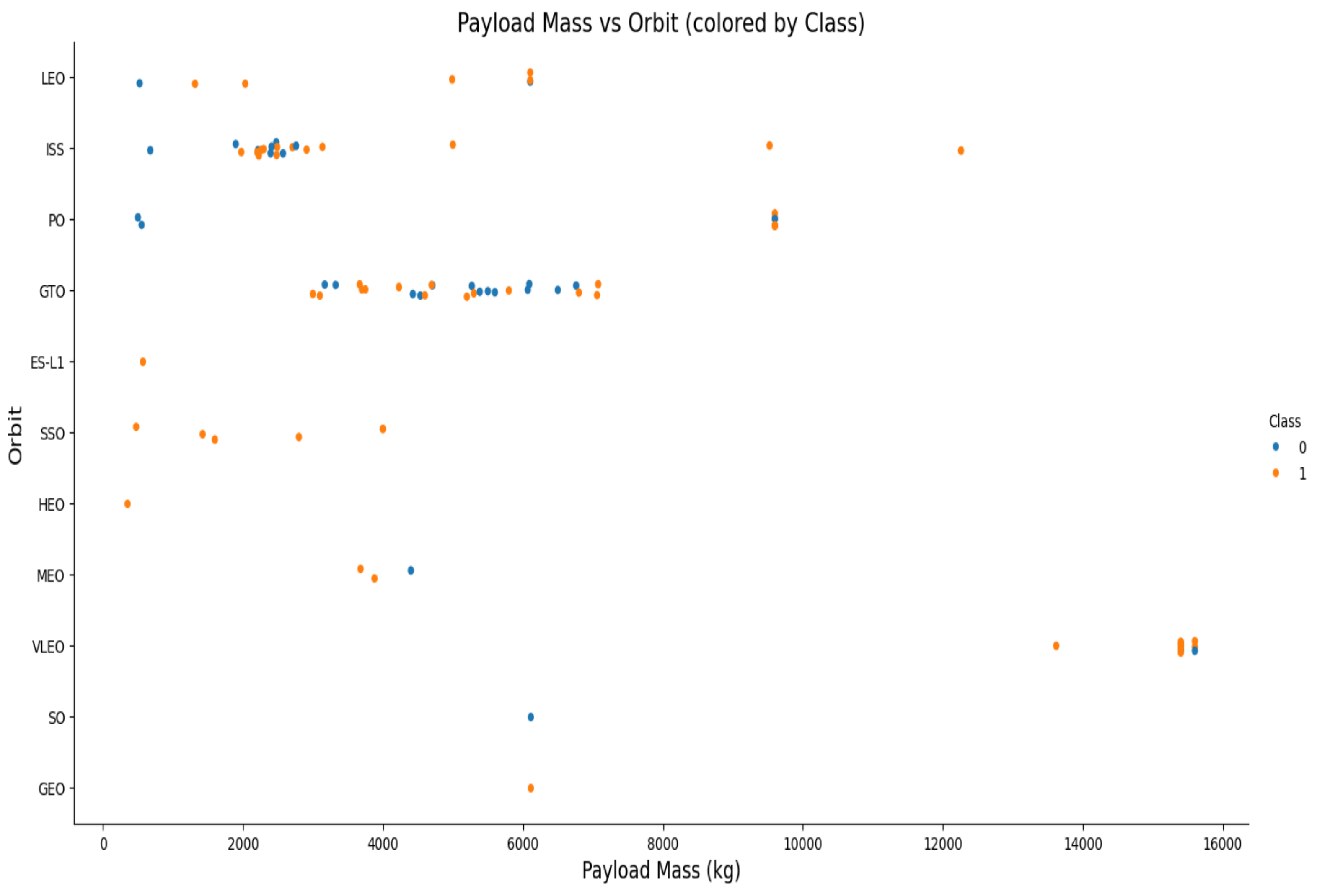


Payload vs. Orbit Type

- Blue (0) = Failure, Orange (1) = Success.
- Shows how payload size affects success rates across different orbits.

Key Insights :

- Low–medium payloads (<10,000 kg) generally achieved higher success across multiple orbits.
- Heavier payloads (>10,000 kg) are mostly linked to higher failure risk.
- GTO and LEO orbits handled a wide payload range, with mixed outcomes
- .ISS missions generally carried smaller payloads but had strong success.

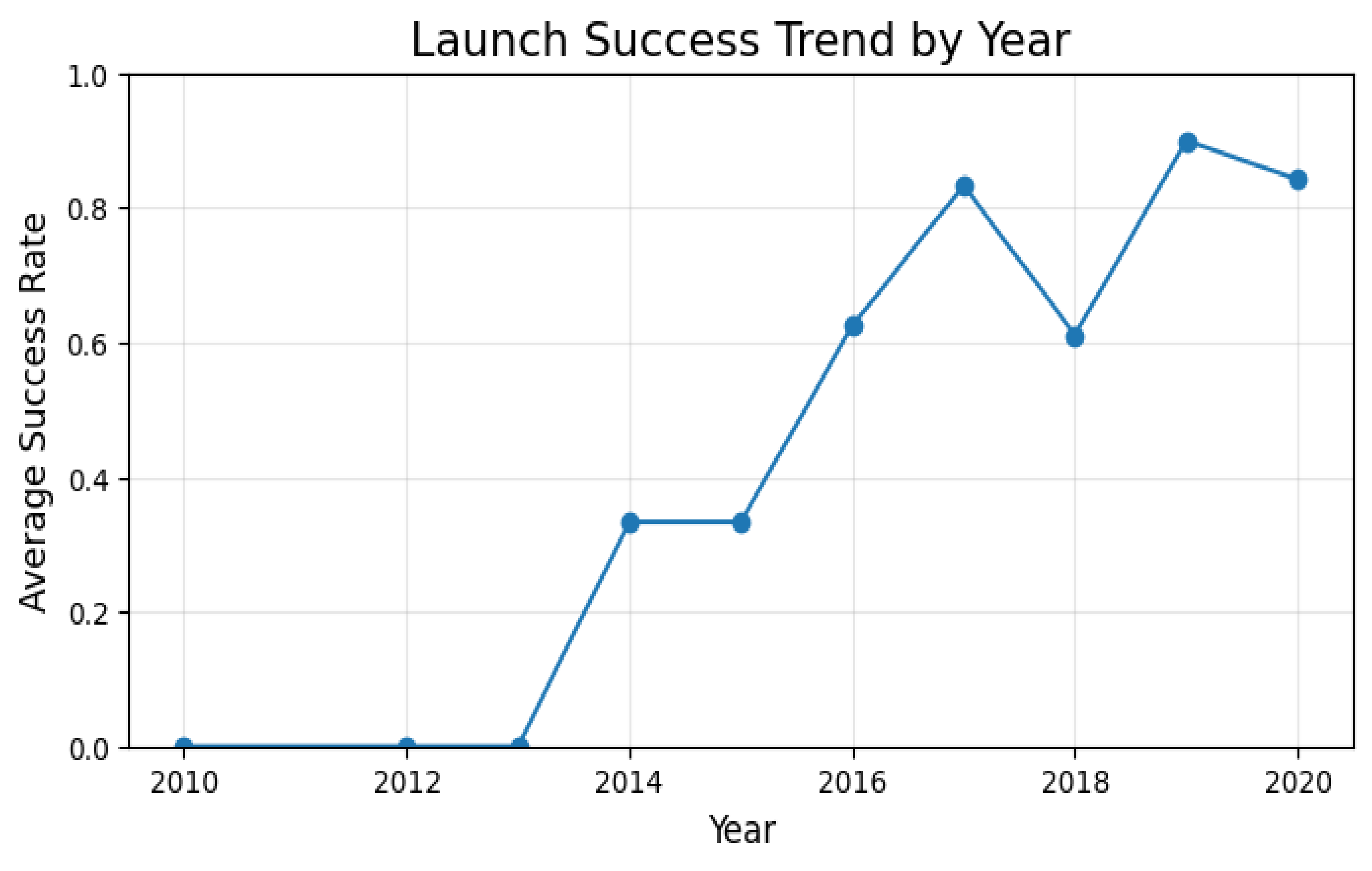


Launch Success Yearly Trend

- Line chart shows average yearly success rate from 2010–2020.
- Highlights how SpaceX improved reliability over time.

Key Insights:

- 2010–2013: Very low success rates (early testing phase).
- 2014–2016: Clear improvement as technology matured.
- 2017–2020: High and stable success (85–90%), showing operational reliability.



All Launch Site Names

- CCAFS LC-40 (Cape Canaveral Air Force Station, Launch Complex 40)
- VAFB SLC-4E (Vandenberg Air Force Base, Space Launch Complex 4E)
- KSC LC-39A (Kennedy Space Center, Launch Complex 39A)
- CCAFS SLC-40 (same Cape Canaveral base, slight naming variation in dataset)
- The dataset includes **4 unique launch sites** across Florida and California. These sites represent SpaceX's major facilities where Falcon 9 rockets were launched.

```
[4]
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

```
 * 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 query filters launch sites

that begin with **“CCA”** (Cape Canaveral).

- The first 5 records all show

CCAFS LC-40, which indicates that Cape

Canaveral’s Launch Complex 40 is a

frequently used site in the dataset.

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

[15] ✓ 0.0s

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Total Payload Mass carried by SpaceX for NASA = **107,010 kg.**

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE Customer LIKE '%NASA%';
```

[16] ✓ 0.0s

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

Total_Payload_Mass
107010

Average Payload Mass by F9 v1.1

- The average payload mass carried by

Falcon 9 v1.1 boosters is **2928 kg**.

Falcon 9 v1.1 was an early upgraded

version of the Falcon 9 rocket.

Its average payload mass (2928 kg)

reflects missions to different orbits,

mostly LEO and some GTO launches

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

[8] ✓ 0.0s

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

Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

- First-ever successful ground landing

 of a Falcon 9 first stage

(LZ-1, Cape Canaveral).

- Date appears as 2015-12-22

because the dataset uses UTC.

```
%%sql
SELECT MIN(Date) AS First_Success_Ground_Pad
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

[9] ✓ 0.0s

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

First_Success_Ground_Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- These booster versions achieved

successful drone ship landings under medium payload conditions 4,000–6,000 kg).

- This demonstrates their capability to balance

 both payload delivery and precise landing recovery.

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

[10] ✓ 0.0s

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

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

Total Number of Successful and Failure Mission Outcomes

- Total missions: 101
- Successful: 100 (including “payload status unclear”)
- Failures: 1Overall success rate: 99%

```
%%sql
SELECT TRIM(Mission_Outcome) AS Mission_Outcome, COUNT(*) AS Total_Count
FROM SPACEXTABLE
GROUP BY TRIM(Mission_Outcome);
```

[18] ✓ 0.0s

* sqlite:///my_data1.db
Done.

Mission_Outcome	Total_Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Selected flights where PAYLOAD_MASS__KG_ equals the dataset's maximum.

If multiple rows match, all boosters tied for the heaviest payload are shown.

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

[22] ✓ 0.0s

```
* 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

- Both failures occurred early in 2015

during v1.1 operations from Cape Canaveral

later years show improved recovery success.

```
%%sql
SELECT substr(Date, 6, 2) AS Month,
       Booster_Version,
       Launch_Site,
       Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date, 1, 4) = '2015';
```

[13] ✓ 0.0s

* sqlite:///my_data1.db
Done.

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

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

- Counted landing outcomes between 2010-06 04 and 2017-03-20.Ranked results in descending order by frequency

Key Insight :

- Early years saw many “No attempt” missions.
- Drone ship landings had both successes and failures, showing trial-and-error phase.
- Ground pad successes started appearing but were fewer in this period

.

```
)
SELECT
  Landing_Outcome,
  Total_Count,
  DENSE_RANK() OVER (ORDER BY Total_Count DESC) AS Rank
FROM counts
ORDER BY Total_Count DESC;
```

[23] ✓ 0.0s

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

Landing_Outcome	Total_Count	Rank
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	3
Success (ground pad)	3	3
Failure (parachute)	2	4
Uncontrolled (ocean)	2	4
Precluded (drone ship)	1	5

Section 3

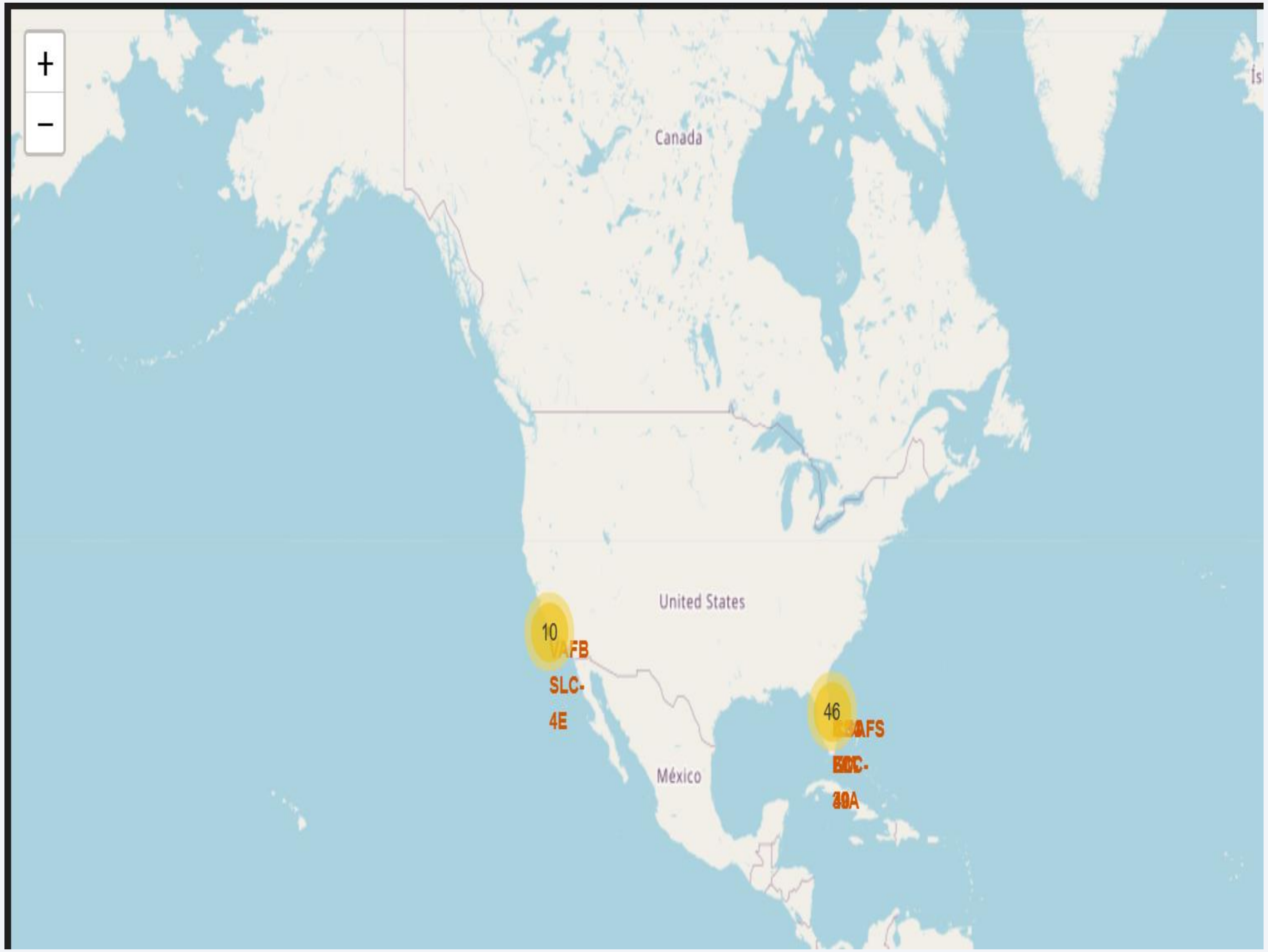
Launch Sites Proximities Analysis

Launch Sites on Global Map

- Replace <Folium map screenshot 1>
- title with an appropriate title

- Explore the generated folium map an
- d make a proper screenshot to
- include all launch sites’ location
- markers on a global map

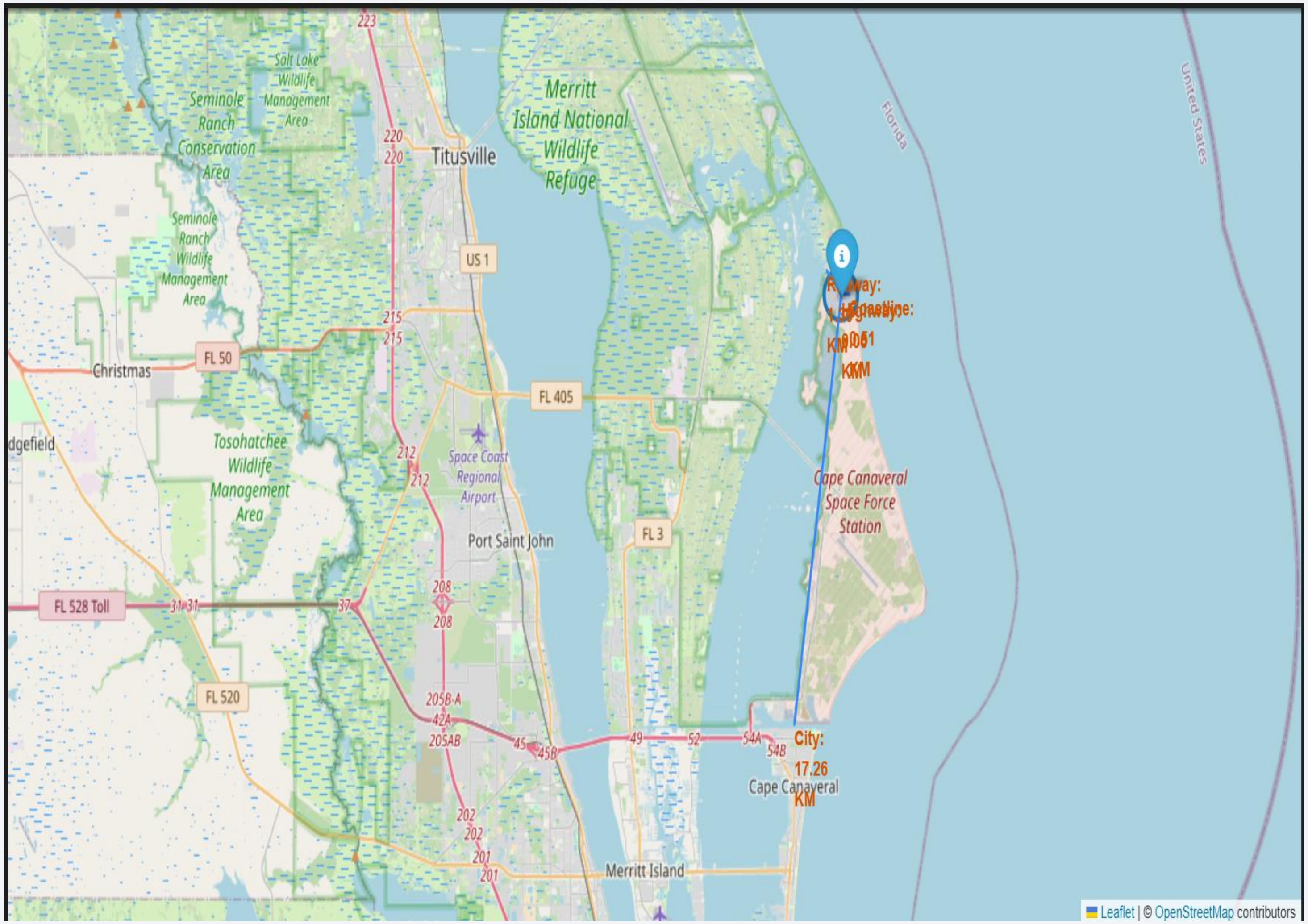
- Explain the important elements
- and findings on the screenshot



Folium Map of SpaceX Launch Sites and Outcomes

CCAFS SLC-40: Proximity to Coastline & Transport

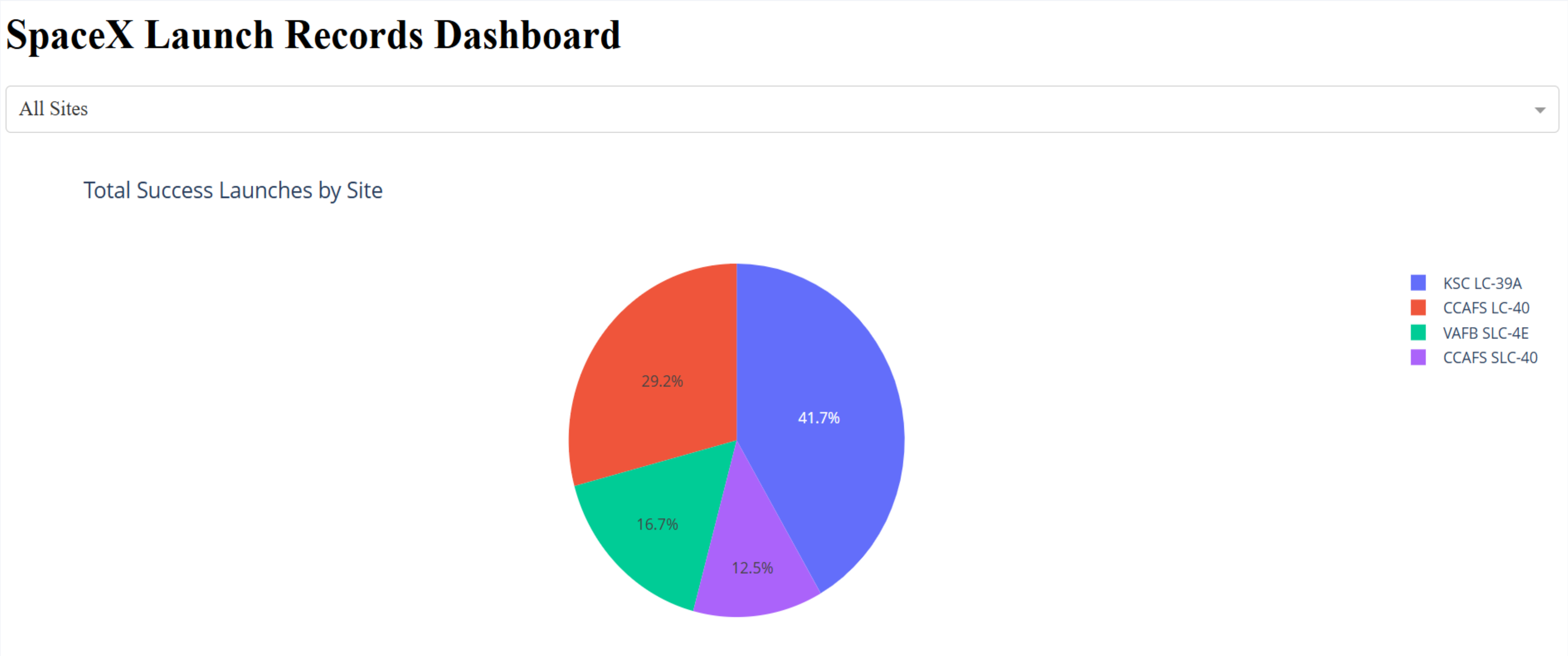
- CCAFS SLC-40 is only 0.61 km from the coastline for safe launches over the ocean, and about 17 km from Cape Canaveral city, with nearby highways and airport for easy access.



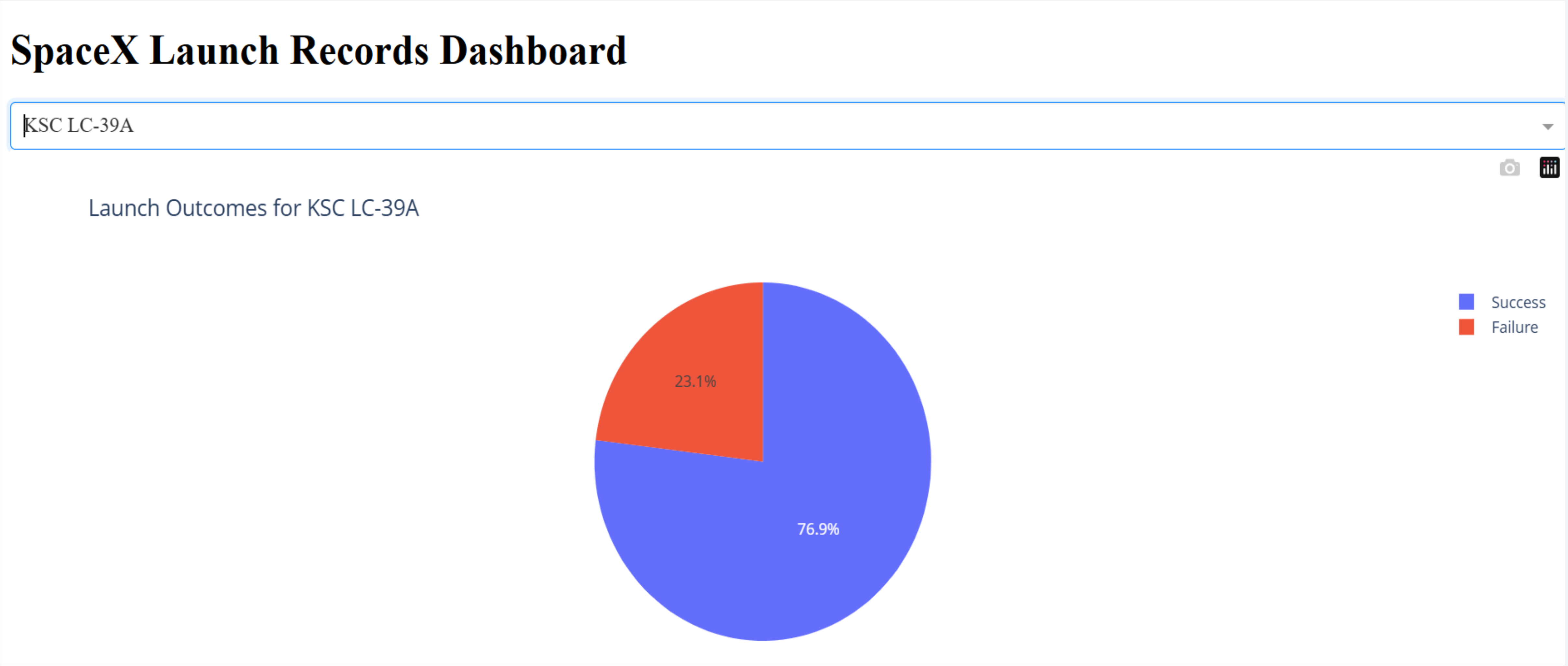
Section 4

Build a Dashboard with Plotly Dash

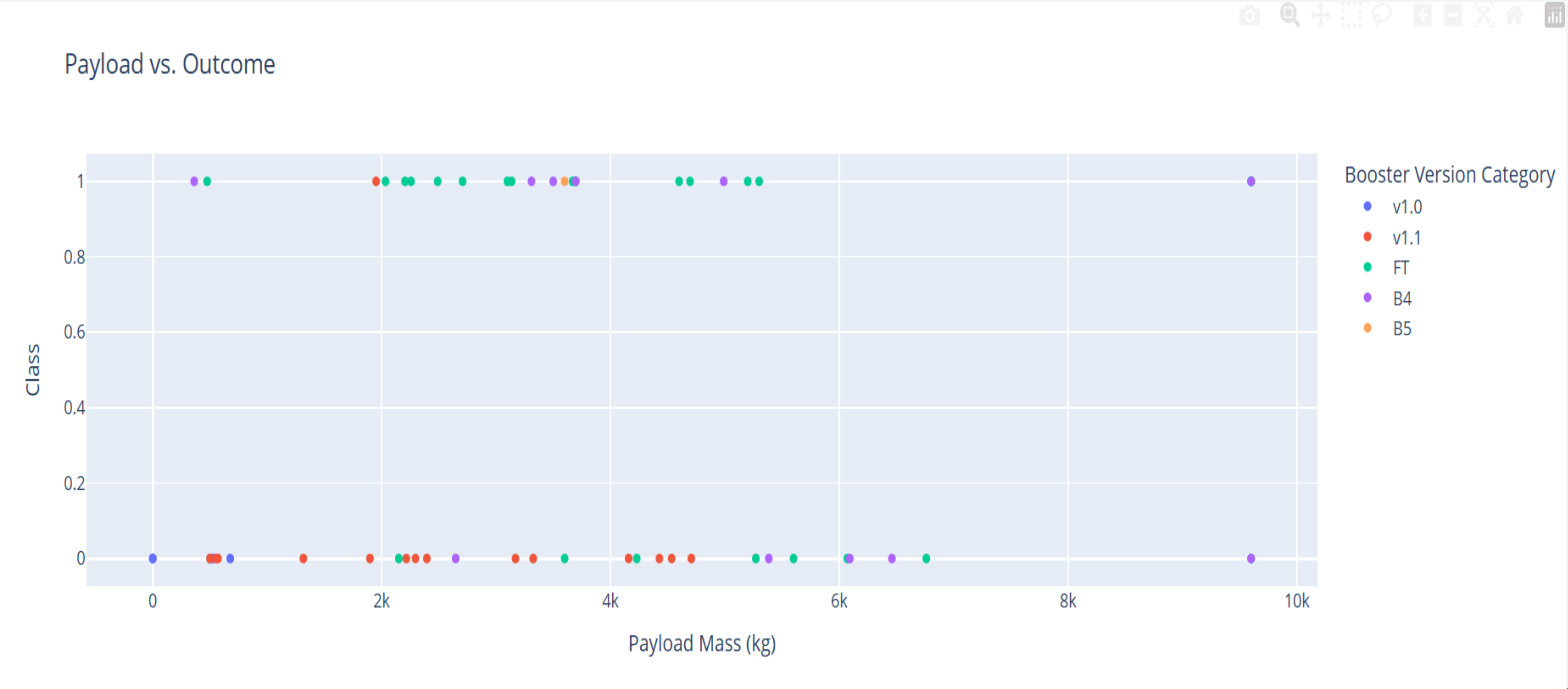
SpaceX Launch Records Dashboard-All Sites



SpaceX Launch Records Dashboard-Most SR



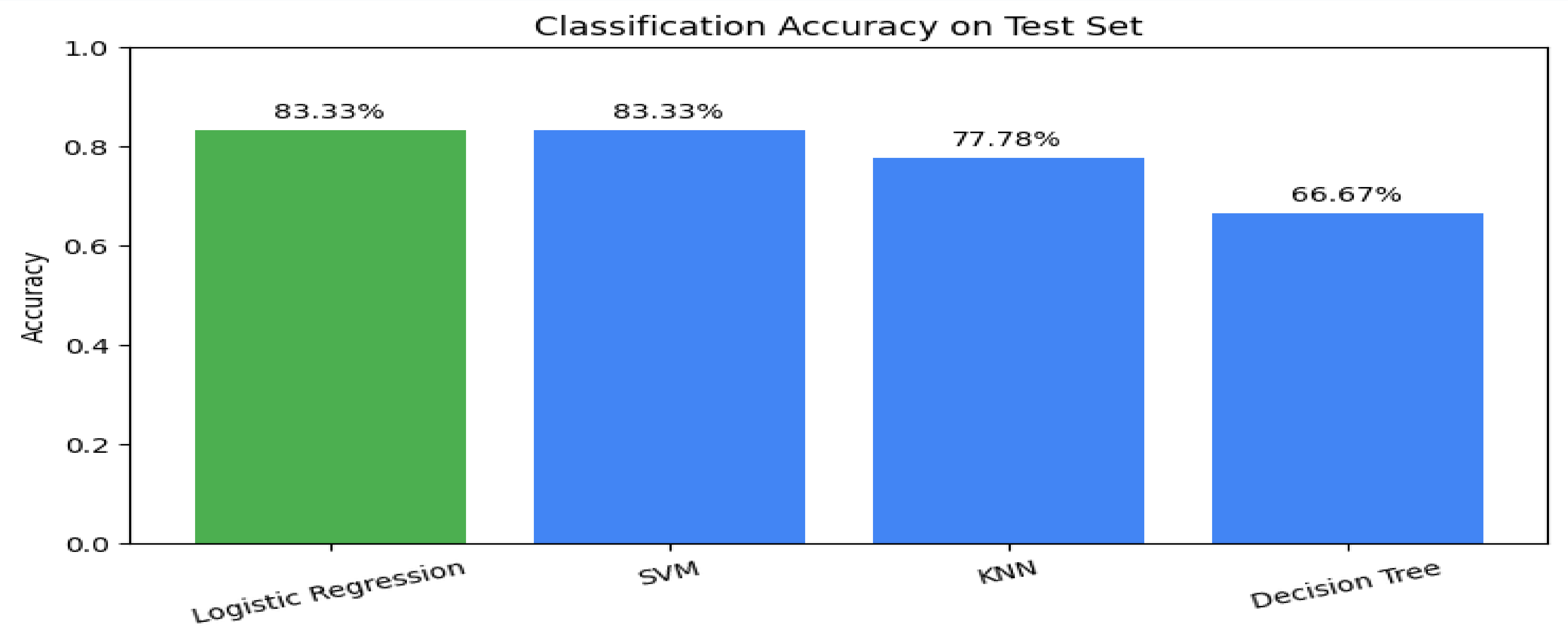
SpaceX Launch Dashboard-Payload vs Outcome



Section 5

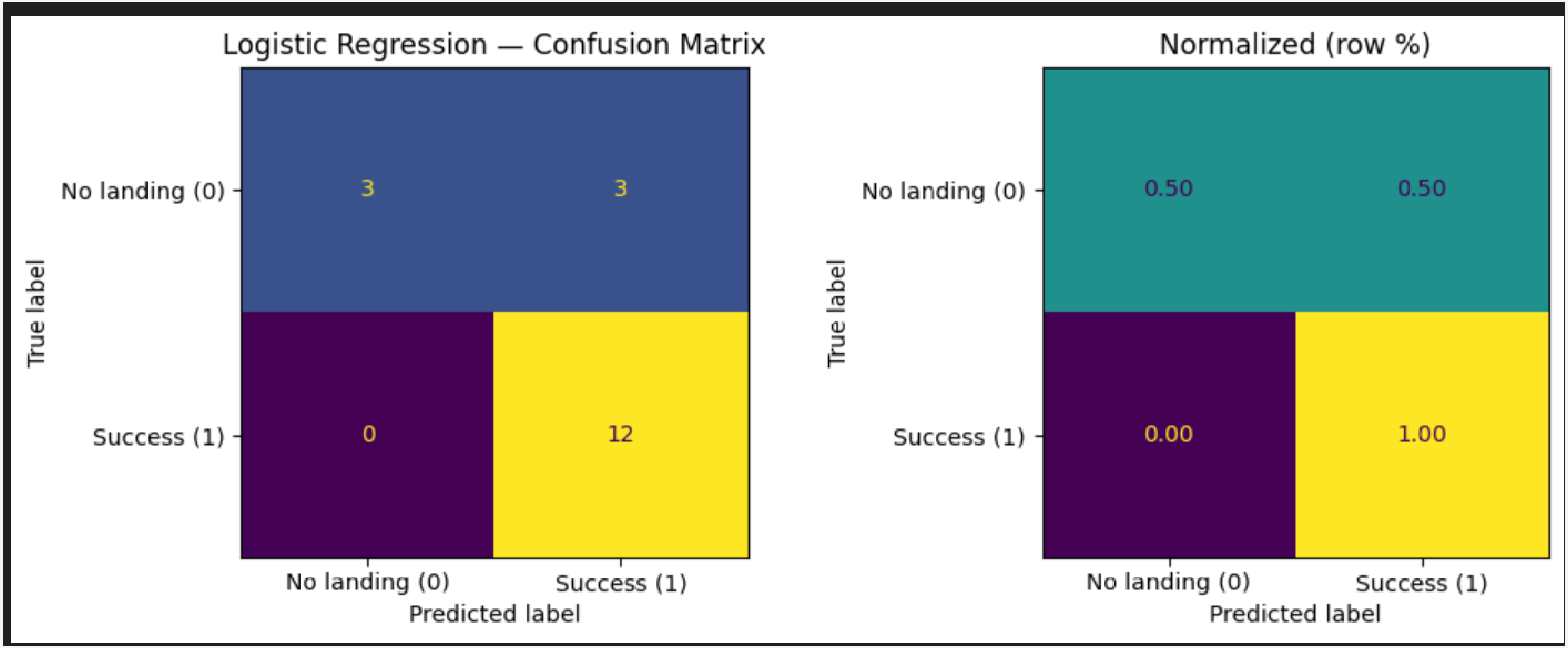
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression and SVM have the highest accuracy.

Confusion Matrix



Rows = actual, columns = predicted; diagonal cells are correct (TP/TN), off-diagonals are errors (FP/FN). More mass on the diagonal ⇒ better model.

Conclusions

- Launch sites & geography: Activity concentrated at Cape Canaveral and Vandenberg; pads sit close to the coast (e.g., CCAFS 0.5 km).
- Outcomes overview: Cape Canaveral accounts for most successful launches (pie chart).
- Best models: Logistic Regression and SVM tie for the top test accuracy 83.3%.
- Chosen model: Use Logistic Regression (simpler/fast, similar accuracy).
- Confusion matrix (LR): TP=12, TN=3, FP=3, FN=0 → success recall = 100%, success precision 80%; misses mainly on predicting failures (0-class recall 50%).

Appendix

- Data: spacex_launch_dash.csv (local) + IBM geo CSV: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- Libraries: Folium/Leaflet, Plotly/Dash, scikit-learn, pandas, NumPy
- Deliverables: Folium map (green/red outcomes + coast distance, jitter), Dash pie (success by site), ML notebook (LR/SVM/KNN/Tree via GridSearchCV)
- Used in slides: Map screenshot, success pie, accuracy bar, best-model confusion matrix.
- Files I built: API + Web Scraping.ipynb

dash.ipynb

paceX_Falcon9_Landing_Prediction.ipynb

EDA with SQL.ipynb

EDA with Visualization Lab.ipynb

python make_spacex_folium_map_unified.ipynb

Thank you!