IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roberta Farris
December 4th, 2025

Outline

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

Executive Summary

Summary of Methodologies

- **Data Preparation:** Collected, cleaned, and structured SpaceX launch data.
- **Exploratory Data Analysis:** Visualized launch success by site, payload, orbit, and booster version.
- **Dashboard:** Built interactive Plotly Dash dashboard for visual analytics.
- **Machine Learning:** Trained Logistic Regression, SVM, Decision Tree, and KNN models; tuned hyperparameters with GridSearchCV.

Summary of Results

- **Most reliable launch sites:** CCAFS SLC-40 & KSC LC-39A.
- **Payload impact:** Moderate payloads show higher success rates.
- **Best models:** Logistic Regression, SVM, KNN (~83% test accuracy).
- **Decision Tree:** Overfits training data, lower test accuracy.
- **Actionable Insight:** Models and dashboard enable **mission planning and success prediction**.

Introduction

Project Background and Context

- SpaceX conducts frequent launches with different boosters, payloads, and orbit types.
- Understanding factors affecting launch success can improve planning, reduce failures, and optimize resources.

Problems to Find Answers

- Which launch sites have the highest success rates?
- How does **payload mass** affect launch success?
- Which **booster versions** are most reliable?
- Can we **predict launch success** using historical data?

Section 1

Methodology

Methodology

Data Collection

Data Wrangling & Processing

Exploratory Data Analysis (EDA)

Interactive Visual Analytics

Predictive Modeling

Data Collection

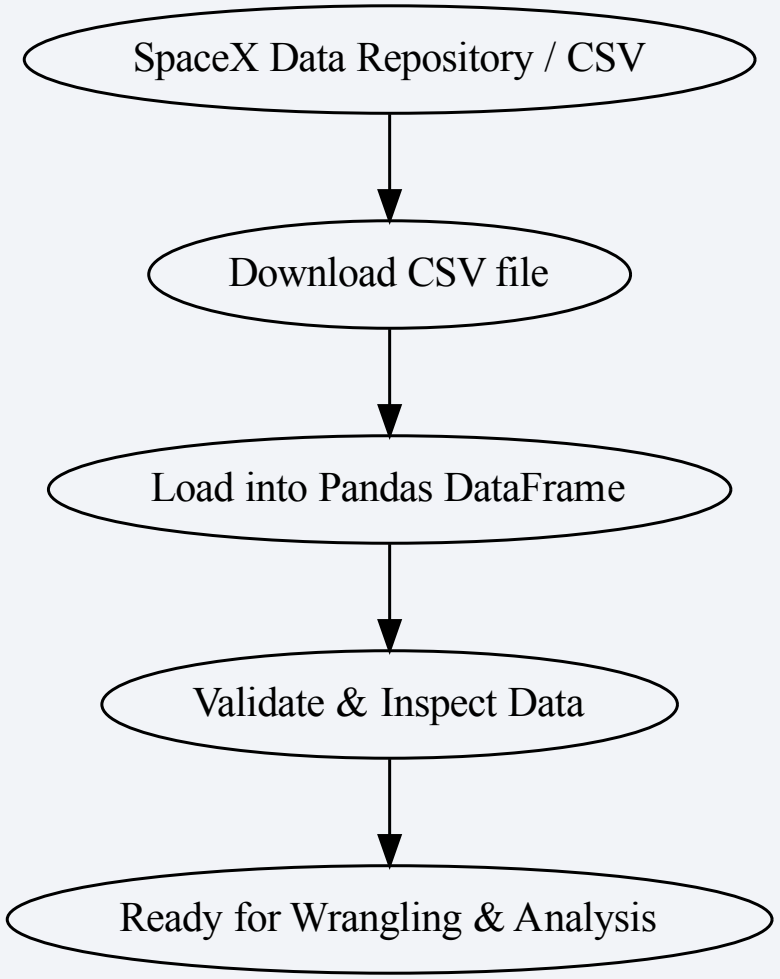
Source: Public SpaceX launch records CSV files

Download: Using `wget` or direct download from IBM/ SpaceX data repository

Content: Includes `Launch Site`, `Date`, `Payload Mass`, `Booster Version`, `Orbit`, `Outcome (class)`

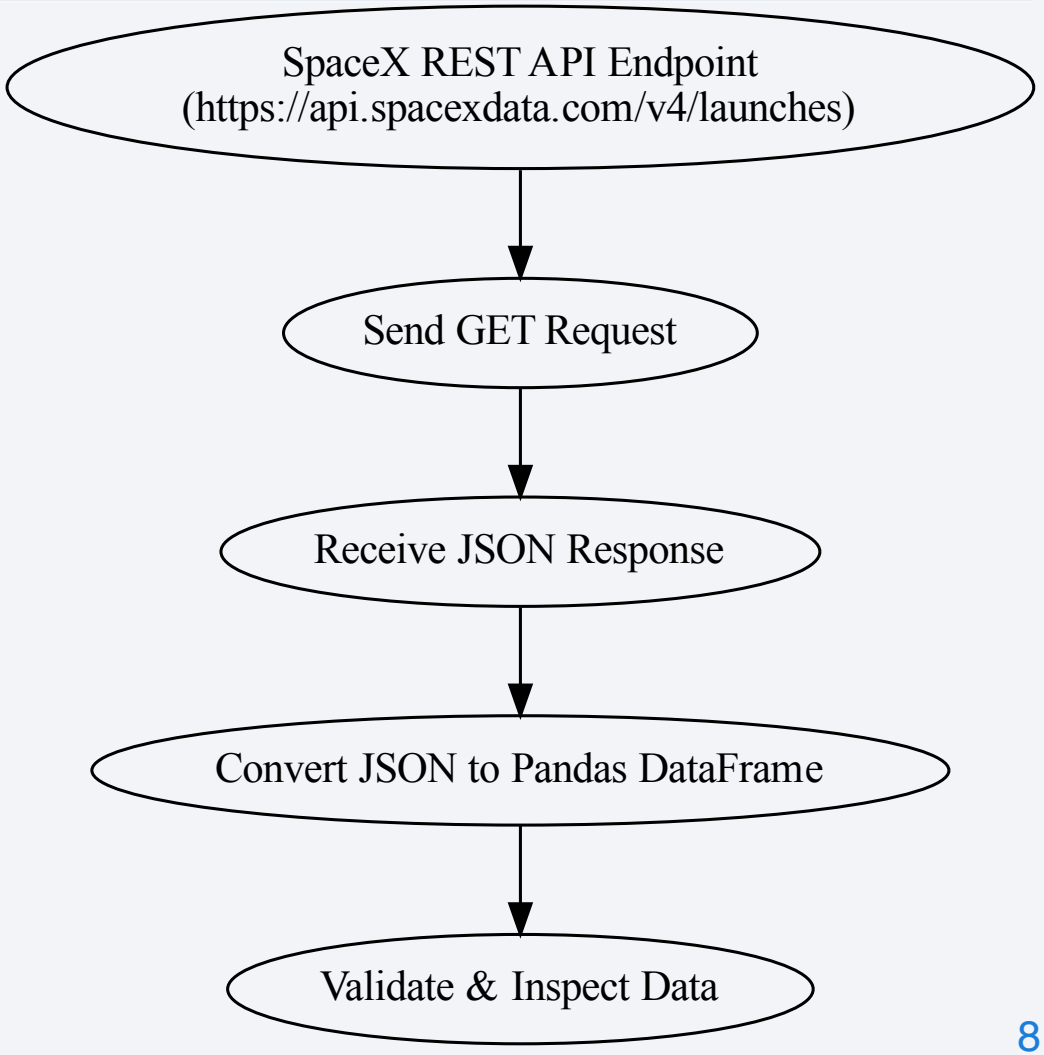
Storage: Imported into **Pandas DataFrame** for processing

Validation: Checked for missing values and duplicates



Data Collection – SpaceX API

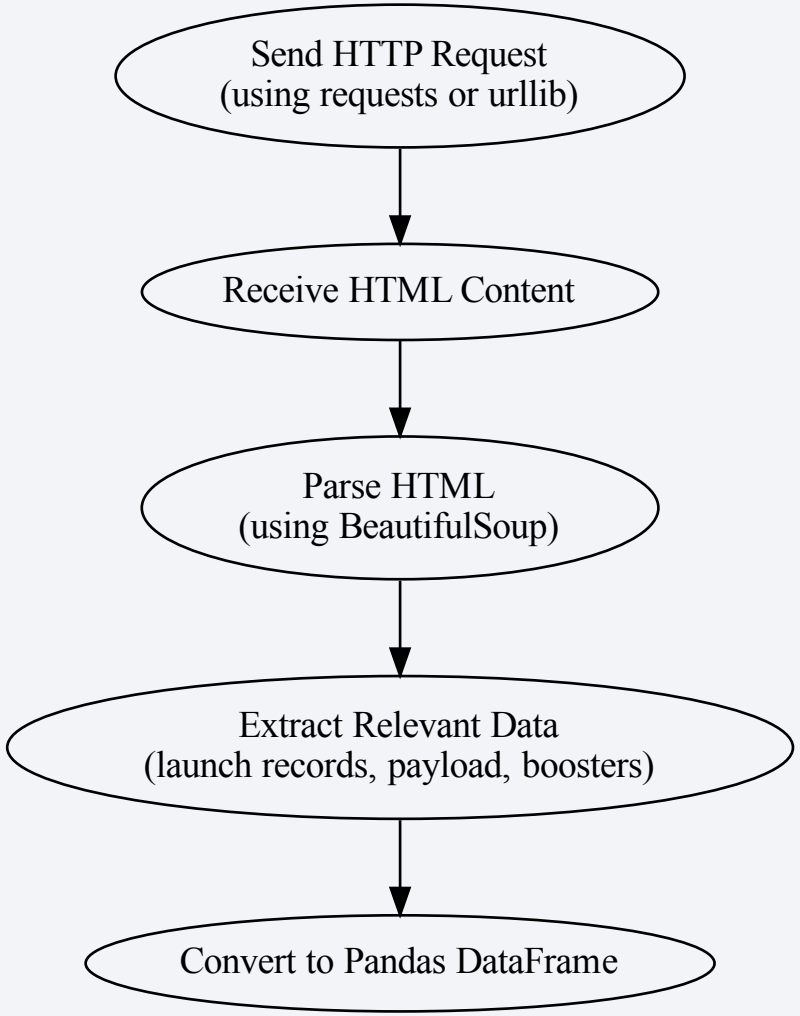
1. **API Source:** SpaceX public REST API (`https://api.spacexdata.com/v4/launches`)
2. **Request:** Use `requests.get()` to fetch launch data in JSON format
3. **Response:** JSON contains launch metadata, payloads, boosters, launch sites, outcomes
4. **Data Transformation:** Convert JSON to **Pandas DataFrame** for analysis
5. **Validation:** Inspect for missing values, duplicate entries, and correct data types
6. **Storage:** Data is ready for **EDA, visualization, and predictive modeling**



https://github.com/RobertaFarris93/Datascience_capstone/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

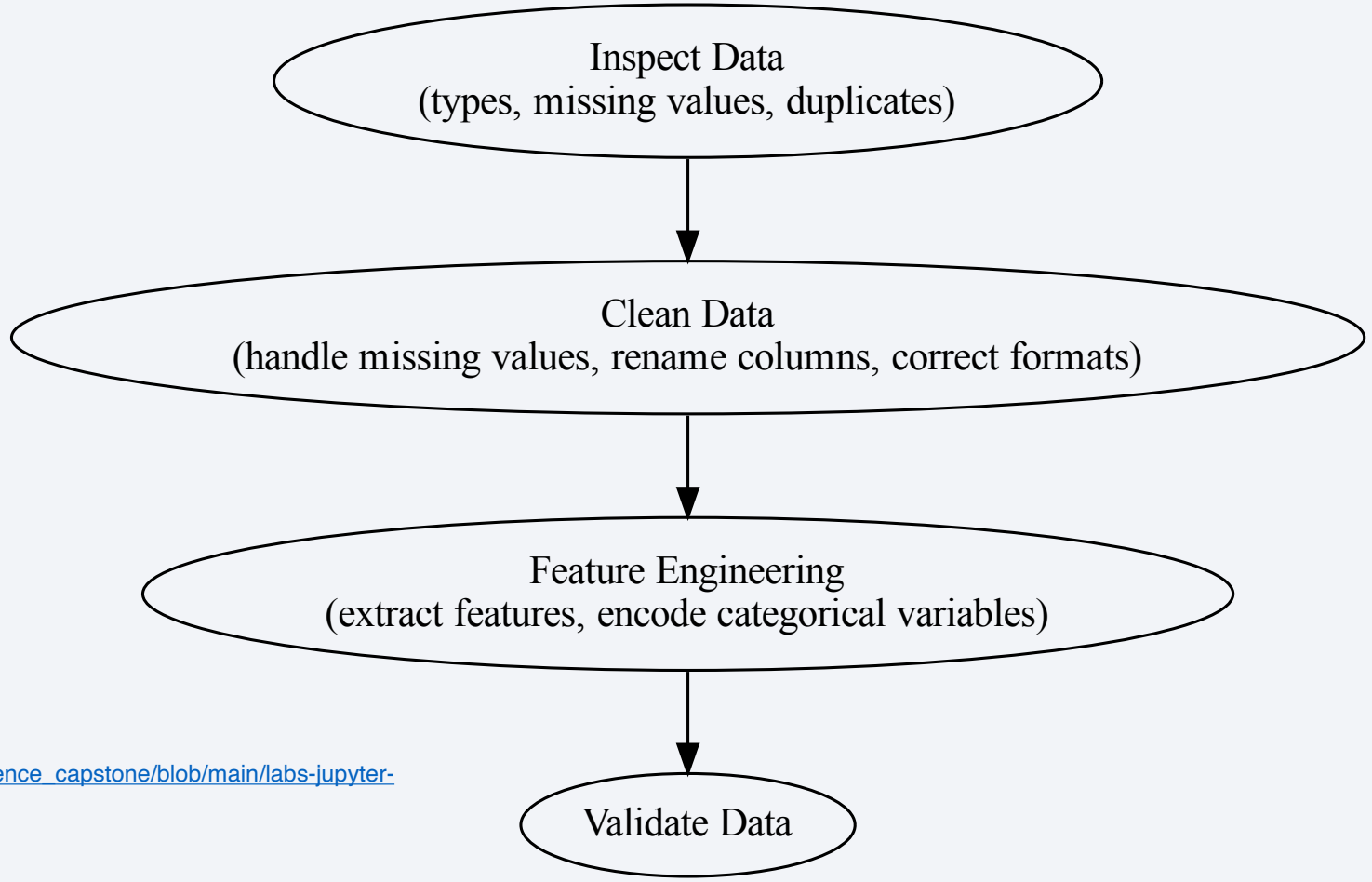
Launch data was collected from public sources using Python scripts, ensuring a clean and structured dataset for analysis.



For more details check out:

https://github.com/RobertaFarris93/Datascience_capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



For more information check:

https://github.com/RobertaFarris93/Datascience_capstone/blob/main/labs-jupyter-spacex-Data_wrangling-v2.ipynb

EDA with Data Visualization

- **Flight Number vs Launch Site (catplot)**
  Shows how experience and launch site influence success.
- **Payload Mass vs Launch Site (scatter)**
  Examines whether heavier payloads reduce success at specific sites.
- **Success Rate by Orbit Type (bar chart)**
  Compares mission difficulty across orbits such as LEO, GEO, GTO, SSO.
- **Flight Number vs Orbit (scatter)**
  Evaluates how growing flight experience affects outcomes for each orbit.
- **Payload Mass vs Orbit (scatter)**
  Reveals payload distribution and risk patterns within each orbit type.
- **Yearly Success Trend (line chart)**
  Shows long-term improvement in SpaceX reliability over time.

Overall Goal

These plots were selected to uncover relationships, trends, and patterns that directly guide **feature engineering** and the **predictive model design** in the ML phase.

https://github.com/RobertaFarris93/Datascience_capstone/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- Queried **unique launch sites** to understand mission locations.
- Filtered mission records by **date range**, **orbit type**, **launch site**, and **landing outcome**.
- Calculated mission counts and performance metrics using **GROUP BY** with `COUNT()`, `AVG()`.
- Ranked **landing outcomes (2010–2017)** in descending order to identify the most common results.
- Extracted **payload**, **orbit**, and **landing outcome** patterns to support EDA and modeling.
- Used SQL to generate **success rates** and **trend summaries** aligned with visualization steps.
- Retrieved date components to support **yearly success trend analysis**.

https://github.com/RobertaFarris93/Datascience_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Map Objects Created

- **Markers** → Identify where launch sites are and where each mission occurred.
- **Circles** → Visualize the "influence radius" around launch sites and highlight spatial boundaries.
- **Clusters** → Prevent overlapping markers in areas with many launches.
- **MousePosition tool** → Quickly capture coordinates of features of interest while exploring the map.
- **PolyLines** → Visually show proximity relationships and compute distances to validate safety constraints.

https://github.com/RobertaFarris93/Datascience_capstone/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

Interactive Components

- **Dropdown (Launch Site Selection)**
- **Range Slider (Payload Mass Selection)**

Plots / Graphs

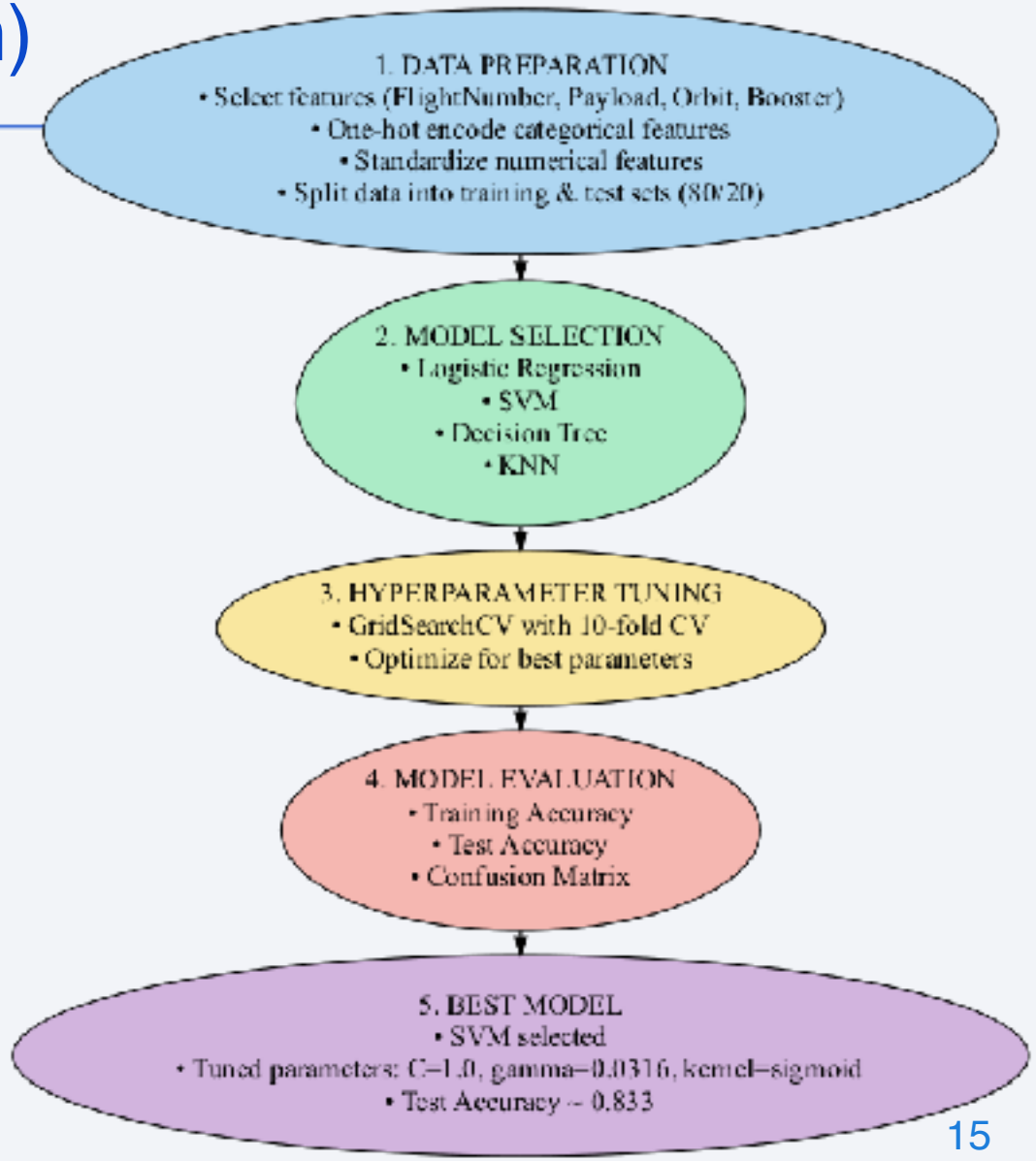
- **Success Pie Chart** (`success-pie-chart`)
- **Payload vs. Launch Success Scatter Chart** (`success-payload-scatter-chart`)

https://github.com/RobertaFarris93/Datascience_capstone/blob/main/Plotly_dashboard.ipynb

Predictive Analysis (Classification)

Four classification models were built and evaluated to predict launch success.

Both **training accuracy** and **test accuracy** were calculated for each model to assess performance and generalization, with test accuracy serving as the primary metric for predictive reliability.

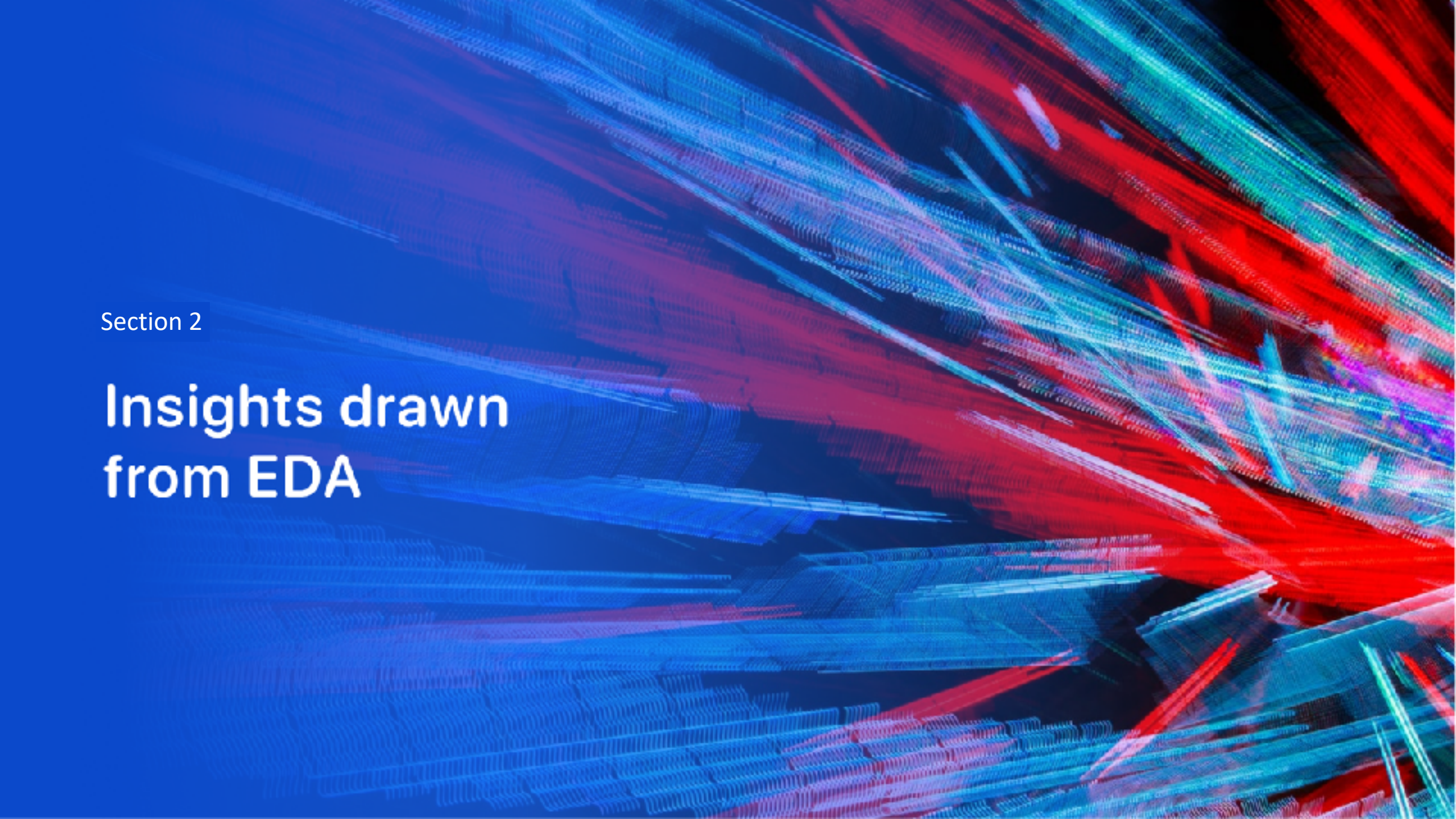


More details can be found here:

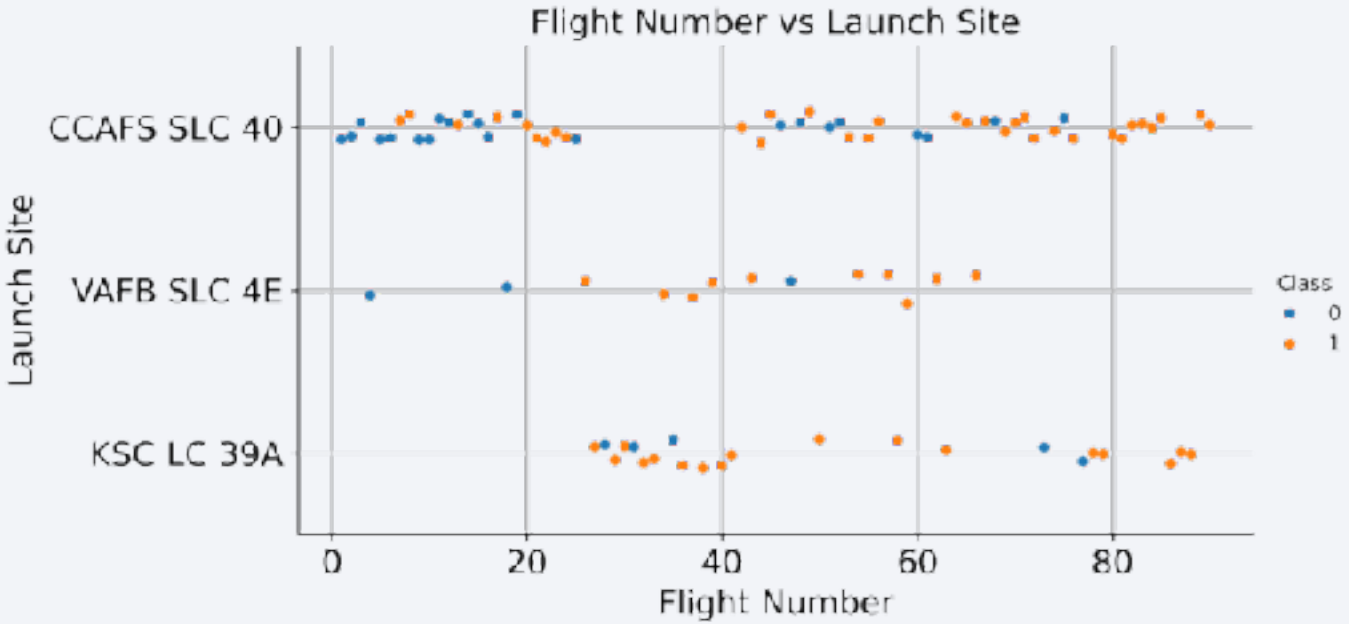
https://github.com/RobertaFarris93/Datascience_capstone/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



`class = 1` → **Success** (booster landed successfully)

`class = 0` → **Failure** (booster did *not* land successfully)

KSC LC-39A was introduced at higher flight numbers, which coincide with increased launch success.

VAFB SLC-4E appears primarily in earlier flight numbers and is not present in later launches shown in the dataset.

CCAFS SLC-40 is used consistently across both early and later flight numbers.

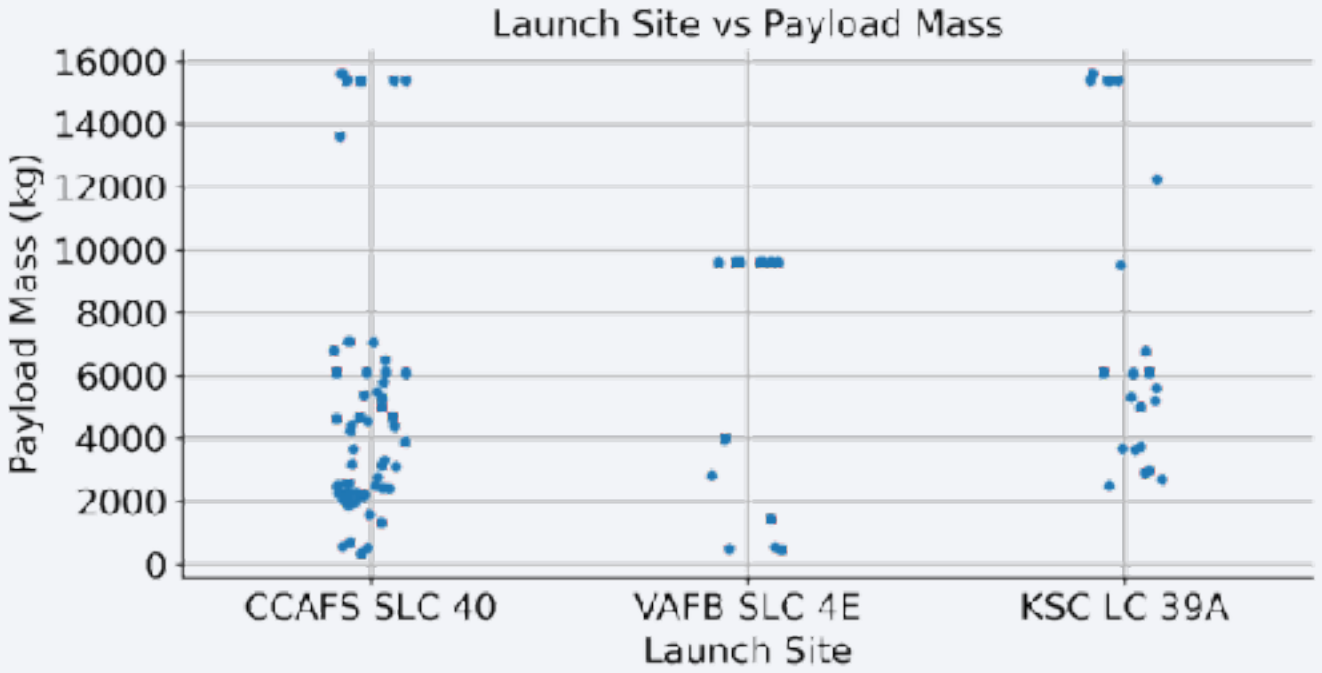
Overall, launch success increases with flight number, indicating improved performance as SpaceX accumulated operational experience, with later launches from all sites showing higher success rates.

Payload vs. Launch Site

🔷 **CCAFS SLC 40 is used for a wide variety of missions, from low to high payload masses, indicating its versatility and frequent use.**

🔷 **VAFB SLC 4E handles fewer missions, with payloads concentrated at specific mass ranges, suggesting more specialized mission profiles.**

🔷 **KSC LC 39A supports heavier payload missions and includes some of the highest payload masses in the dataset.**



Launch frequency

- CCAFS SLC-40 clearly has:
  - The **most launches**
  - The **widest payload range**
- Different launch sites support **different payload mass ranges**
- CCAFS SLC-40 is the most frequently used and most versatile site
- KSC LC-39A supports some of the heaviest payloads

Success Rate vs. Orbit Type

Launch success varies across orbit types, with several orbit categories showing high success rates, while GTO missions exhibit comparatively lower success in the dataset.

🔷 **Very high success rate (≈ 100%)**

- **ES-L1**, **GEO**, **HEO**, **SSO**

⚠️ Some of these orbits have **very few launches**, so a 100% rate does **not necessarily mean** they are easier

🔷 **Moderate success rate (~60–70%)**

- **ISS**, **LEO**, **MEO**, **PO**

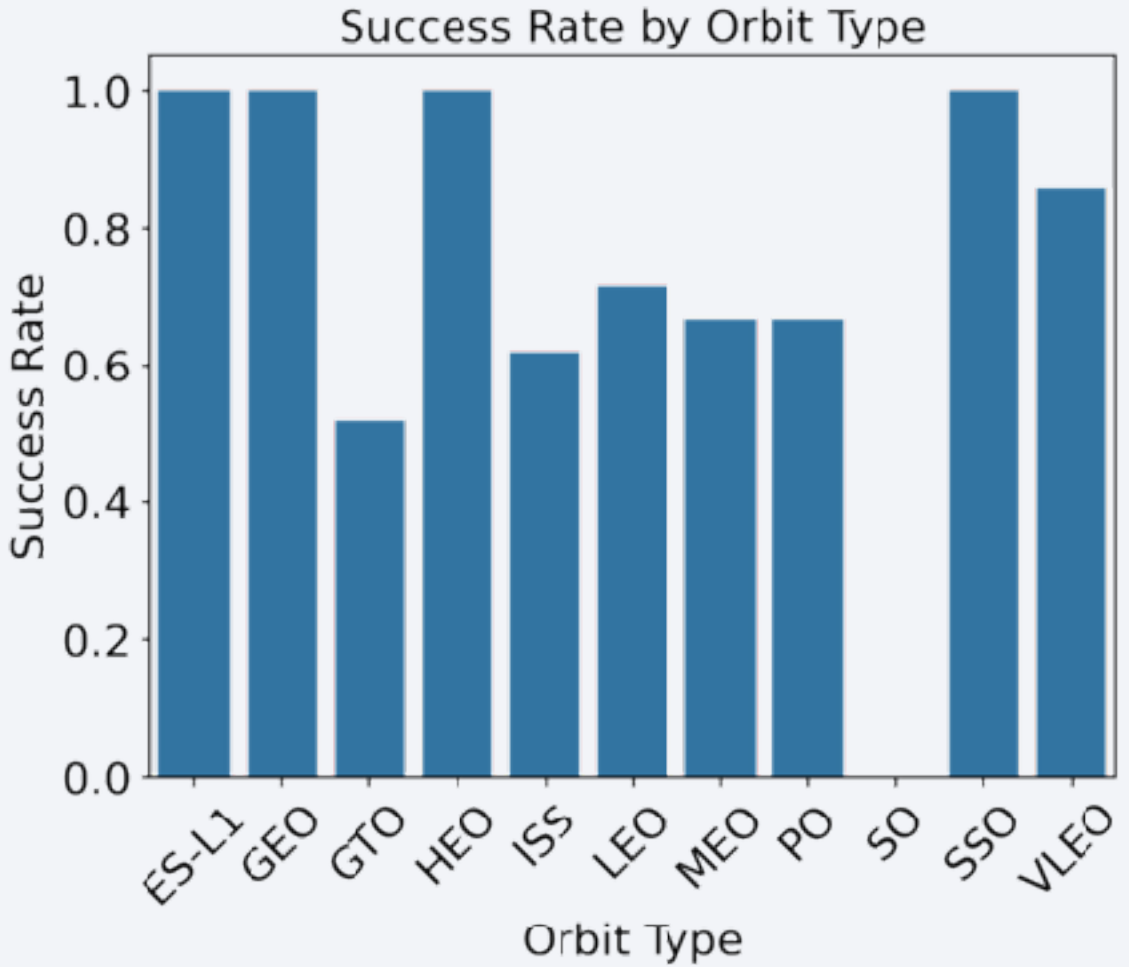
🔷 **Lower success rate**

- **GTO** (~50%)

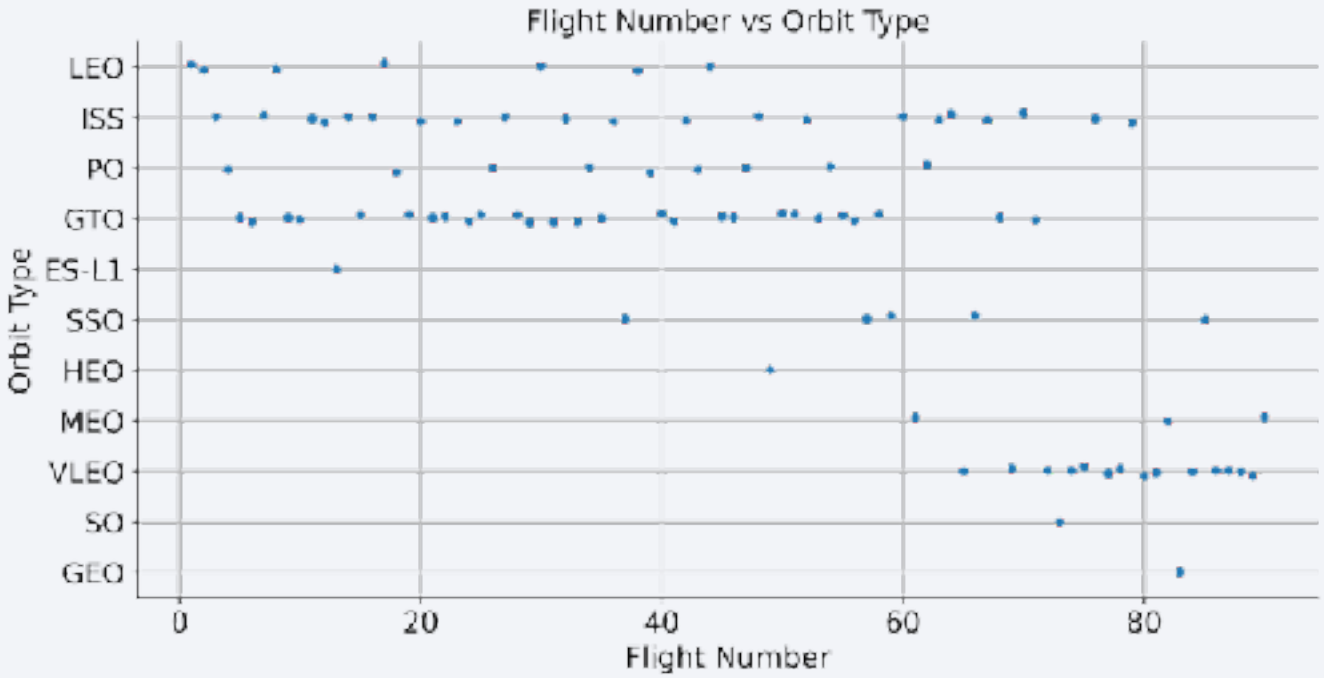
GTO missions show a lower success rate, likely reflecting higher mission complexity or earlier experimental flights.

🔷 **High but not perfect**

- **VLEO** (~85%)

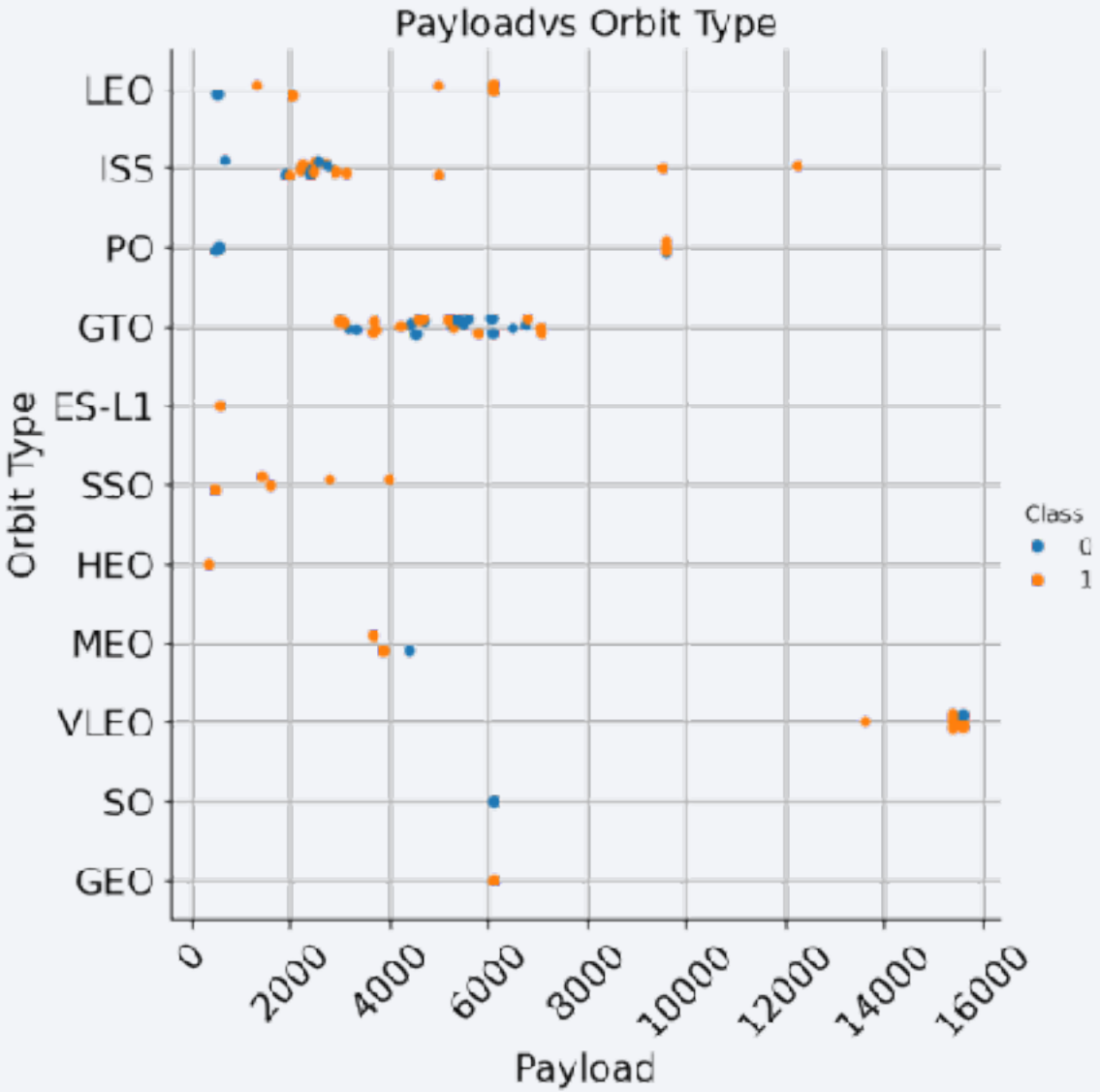


Flight Number vs. Orbit Type



- Some orbit types appear early (at lower flight numbers).
- Other orbit types only appear at higher flight numbers, indicating they were introduced later in the flight program.
- Some orbits are used frequently across many flights.
- Others are used sparingly, appearing in only a few flights.

Payload vs. Orbit Type



Payload suitability:

- Some orbits are more suitable for lower payloads.
- Only Very Low Earth Orbit (VLEO) was used for extremely high payloads.

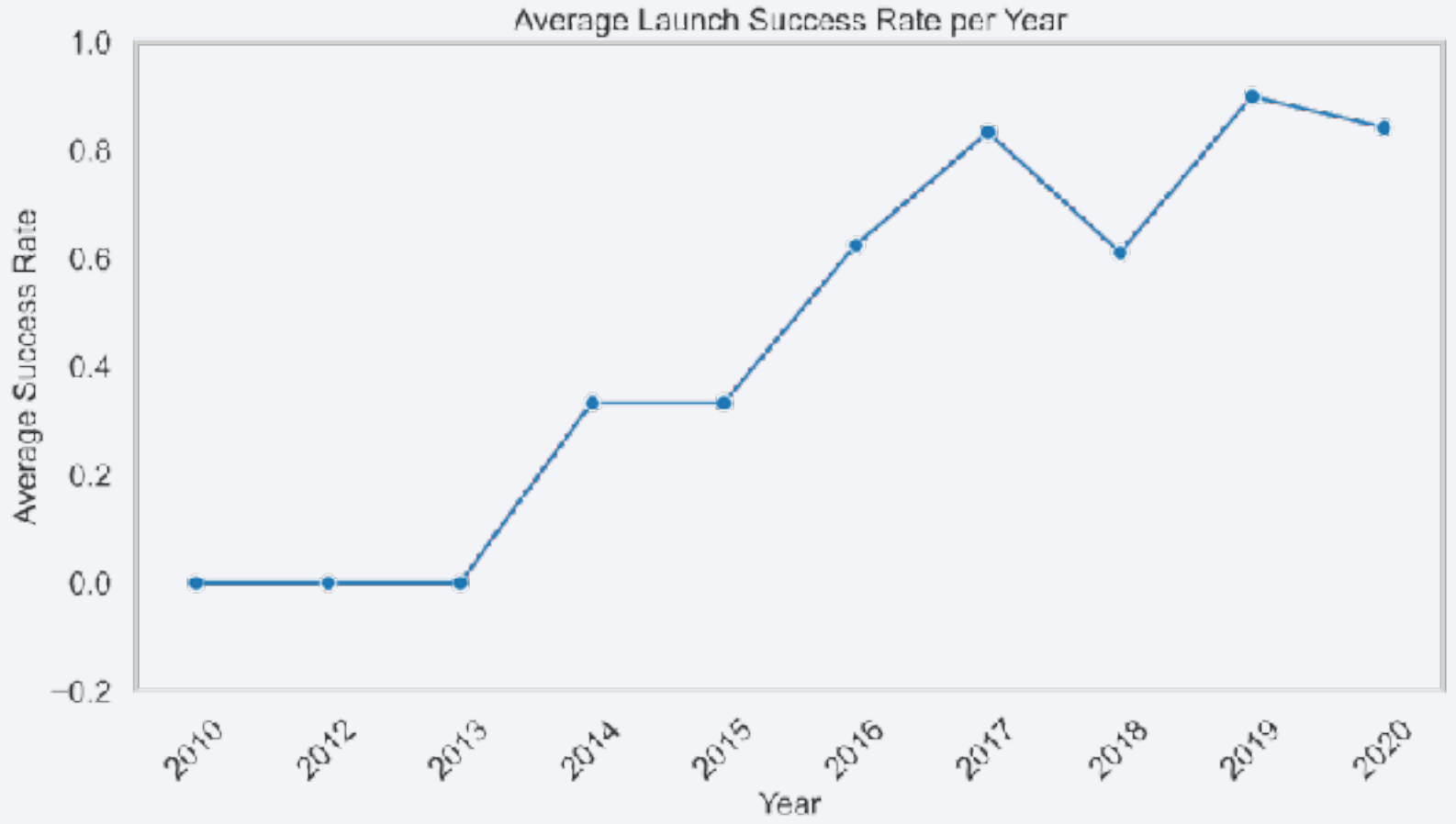
Success rates and reliability:

- Certain orbits show a 100% success rate.
- ISS and GTO required more attempts and had more failures, highlighting their higher complexity.

Other orbits:

- Used less frequently, with fewer statistics on failures or successes.

Launch Success Yearly Trend



Success rate trend: Since 2013, the success rate generally increased. It remained stable in 2014, then started rising again after 2015, continuing up to 2017.

All Launch Site Names

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

Unique launch sites: The dataset contains four launch sites:

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

These sites may influence mission success due to their location, infrastructure, or operational differences.

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
```

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

Records from CCAFS launch sites: The first five records from CCAFS SLC 40 show early Falcon 9 missions to LEO, All of these missions in the sample were successful and they occurred between 2010 and 2013.

Total Payload Mass

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

Total_Payload_Mass: 45596

This value represents the sum of all payload masses across every recorded flight for NASA. It gives an overall sense of the amount of material launched and can serve as a baseline when comparing payload distribution by orbit, launch site, or mission type.

Average Payload Mass by F9 v1.1

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

AVG_Payload_Mass
2928.4

Across all recorded missions in the dataset, Falcon 9 boosters carried an average payload of approximately **3000 kg**.

First Successful Ground Landing Date

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

```
Date of first successful landing on a ground pad: 2015-12-22
```

The first recorded successful landing occurred on **December 22, 2015**. This marks a key milestone in booster recovery and reusability efforts.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query result: No booster versions meet the criteria of having a successful drone ship landing with a payload mass between 4,000 kg and 6,000 kg.

Explanation: This indicates that, within the dataset, either successful drone ship landings occurred with payloads outside this range, or missions within this payload range did not achieve successful drone ship landings.

Total Number of Successful and Failure Mission Outcomes

The vast majority of missions were successful.

Only one mission failed during flight, highlighting a high overall mission reliability.

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

Boosters Carried Maximum Payload

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Payload_Mass_(kg)" = (
    SELECT MAX("Payload_Mass_(kg)")
    FROM SPACEXTABLE
);
```

- **F9 v1.0**
  - Early Falcon 9 boosters (e.g., B0003–B0007).
  - Represent the initial development phase with no reusability.
- **F9 v1.1**
  - Includes boosters such as B1003–B1018.
  - Improved performance and reliability compared to v1.0, but limited reuse.
- **F9 Full Thrust (FT)**
  - Boosters like B1019–B1038.
  - Introduced higher thrust and supported early reuse attempts.
- **F9 Block 4 (B4)**
  - Boosters such as B1039–B1045.
  - Transitional version focused on refining reusability.
- **F9 Block 5 (B5)**
  - Boosters from B1046 onward (e.g., B1046–B1063).
  - Designed for rapid and repeated reuse, appearing most frequently in the dataset.

This grouping highlights the **technological progression of Falcon 9 boosters** and explains why later versions dominate the launch records.

2015 Launch Records

:	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)

In **January**, booster **F9 v1.1 B1012** launched from **CCAFS LC-40** and resulted in a **drone ship landing failure**.

In **April**, booster **F9 v1.1 B1015**, also launched from **CCAFS LC-40**, experienced a **drone ship landing failure**.

Both failures occurred during early Falcon 9 v1.1 missions and from the same launch site, reflecting the experimental phase of drone ship landings before landing reliability improved in later booster versions.

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

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

A significant number of missions did not attempt recovery, reflecting early operational constraints. Drone ship landings show an equal number of successes and failures, indicating a learning phase in recovery technology. Ground pad landings, while fewer, demonstrate successful controlled recoveries, and ocean outcomes mostly represent non-recovery or experimental attempts.

Section 3

Launch Sites Proximities Analysis

Global Distribution of SpaceX Launch Sites

The folium map displays the geographical locations of all SpaceX launch sites included in the dataset. Each marker represents a launch site positioned accurately on a global map using latitude and longitude coordinates.

Key observations from the map include:

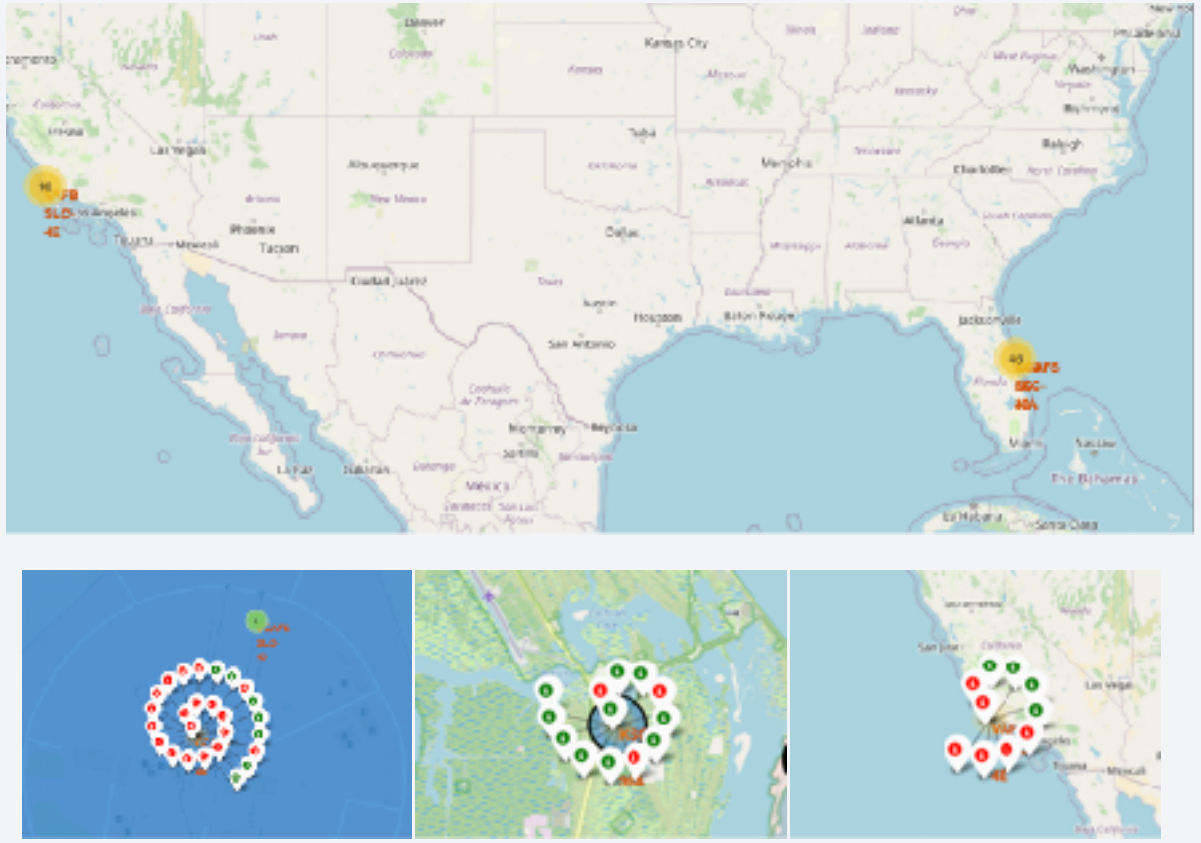
- Launch sites are concentrated in the **United States**, with locations on both the **East Coast** (Florida) and the **West Coast** (California).
- **CCAFS SLC 40** and **KSC LC 39A** are located in Florida, supporting missions to a wide range of orbits.
- **VAFB SLC 4E**, located in California, is primarily used for polar and sun-synchronous orbit launches.
- The spatial distribution of launch sites highlights how geographic location influences mission objectives, orbital inclinations, and launch efficiency.

Launch Outcomes by Location (Color-Coded Map)

The folium map visualizes SpaceX launch locations with **color-coded markers representing launch outcomes**. Each marker corresponds to a specific launch site, with different colors indicating whether the mission was successful or unsuccessful.

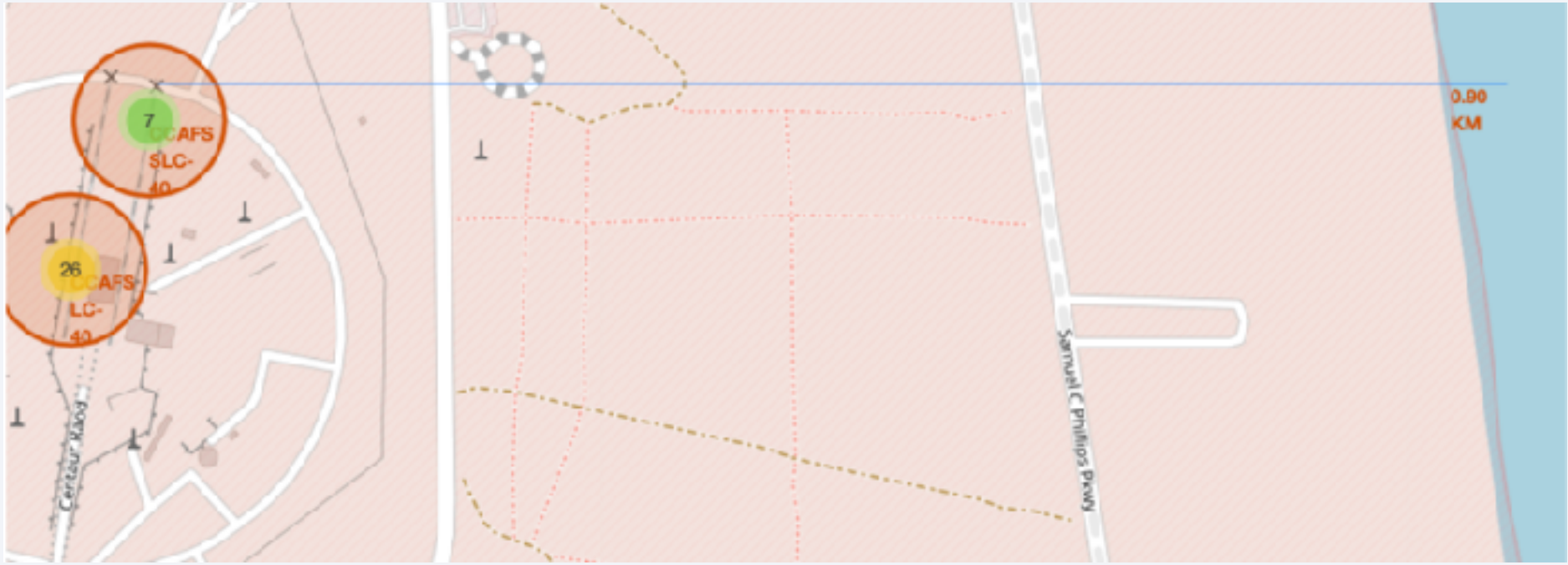
Key elements and findings:

- **Color coding** allows quick visual comparison of launch outcomes across different locations.
- A higher concentration of **successful launches** is visible at **CCAFS SLC 40** and **KSC LC 39A**, indicating improved reliability over time at these sites.
- **VAFB SLC 4E** shows fewer launches, with outcomes reflecting its specialized mission profile.
- The map demonstrates that launch success is not randomly distributed but varies by site, operational experience, and mission complexity.



Launch Site Proximity Analysis: Nearby Infrastructure and Coastline

- Launch sites are typically **very close to coastlines** to reduce risks from failed launches.
- Launch locations maintain **safe distances from cities** to avoid population exposure.
- Sites are often **near highways and railways**, providing logistical access for rockets and equipment.
- Success/Failure markers reveal patterns related to **site suitability and infrastructure support**.

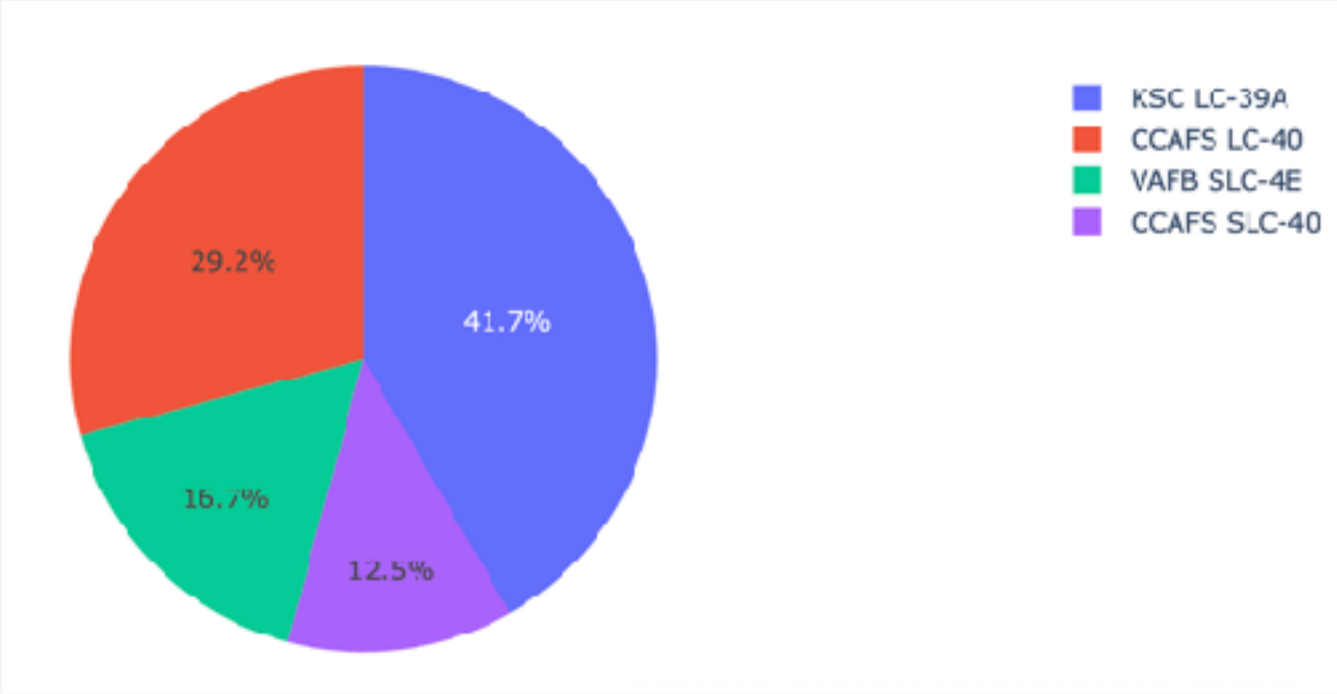


Section 4

Build a Dashboard with Plotly Dash

Launch Success Distribution Across All Sites

- The chart highlights that **CCAFS (S)LC 40** accounts for the largest share of successful launches, reflecting its frequent use and long operational history.
- **KSC LC 39A** also contributes a significant portion of successful missions, indicating its importance in high-profile and recent launches.
- **VAFB SLC 4E** has a smaller slice, consistent with its more specialized and less frequent launch activity.

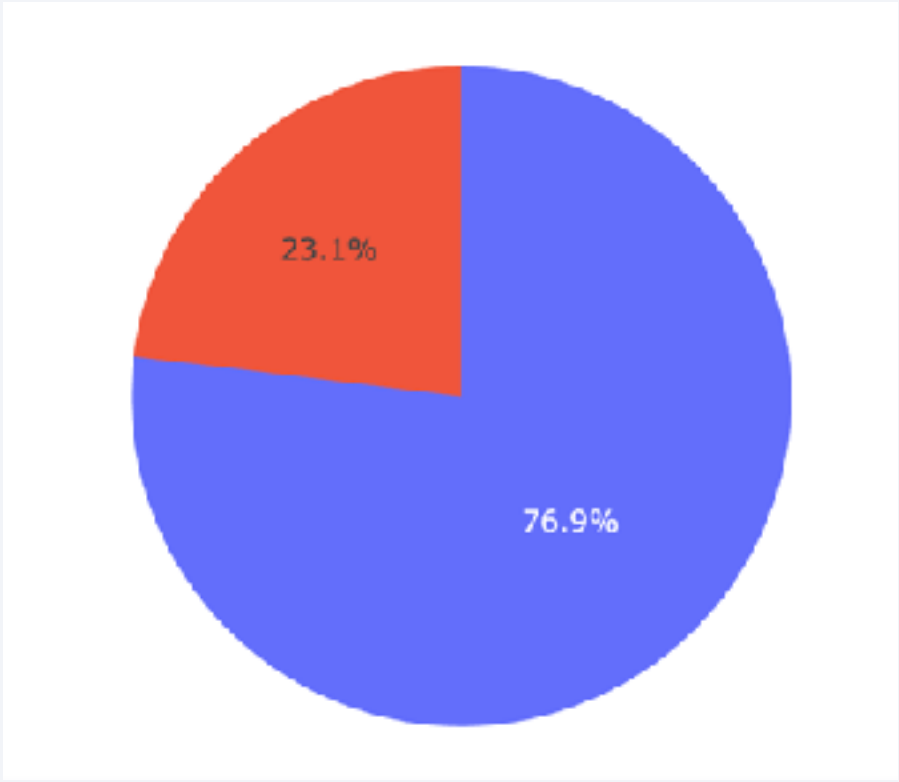


The pie chart provides an intuitive comparison of launch success contributions by site, making it easy to identify the most active and reliable launch locations.

Launch Success Ratio for the Top Performing Site

The chart clearly distinguishes **successes** (e.g., purple slice) from **failures** (e.g., red slice), allowing quick assessment of reliability.

The selected site demonstrates a **very high success ratio**, indicating consistently successful operations.



Launch outcomes (success or failure) versus payload mass

The range slider allows analysis of targeted payload intervals, helping identify **payload thresholds where success rates change**, which is useful for mission planning.

Section 5

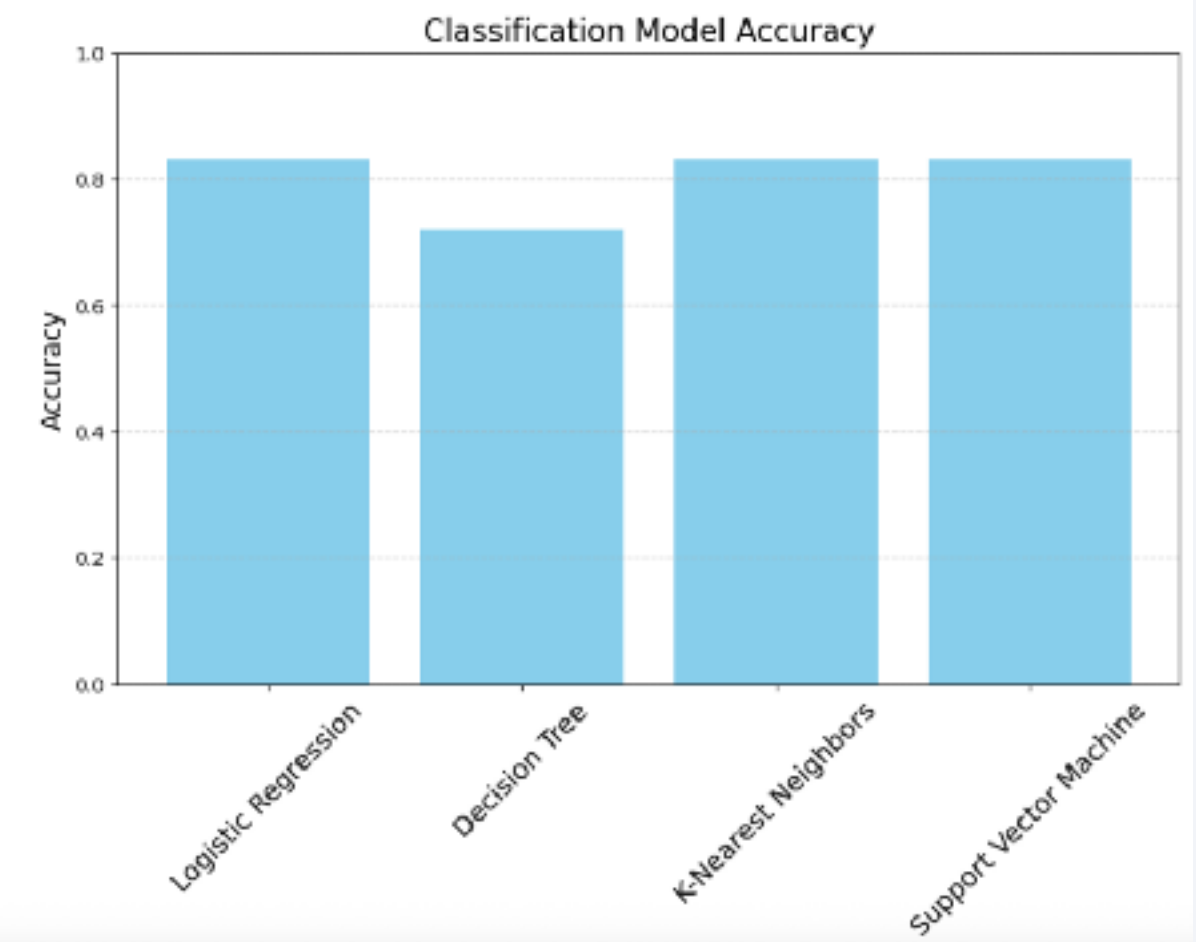
Predictive Analysis (Classification)

Classification Accuracy

The **highest test accuracy** is **83%**, achieved by **KNN, Logistic Regression, and SVM**.

The other model (**Decision Tree**) has a lower accuracy of **72%**.

This indicates that for your dataset, simpler models like KNN, Logistic Regression, and SVM generalize better than the tree-based models.



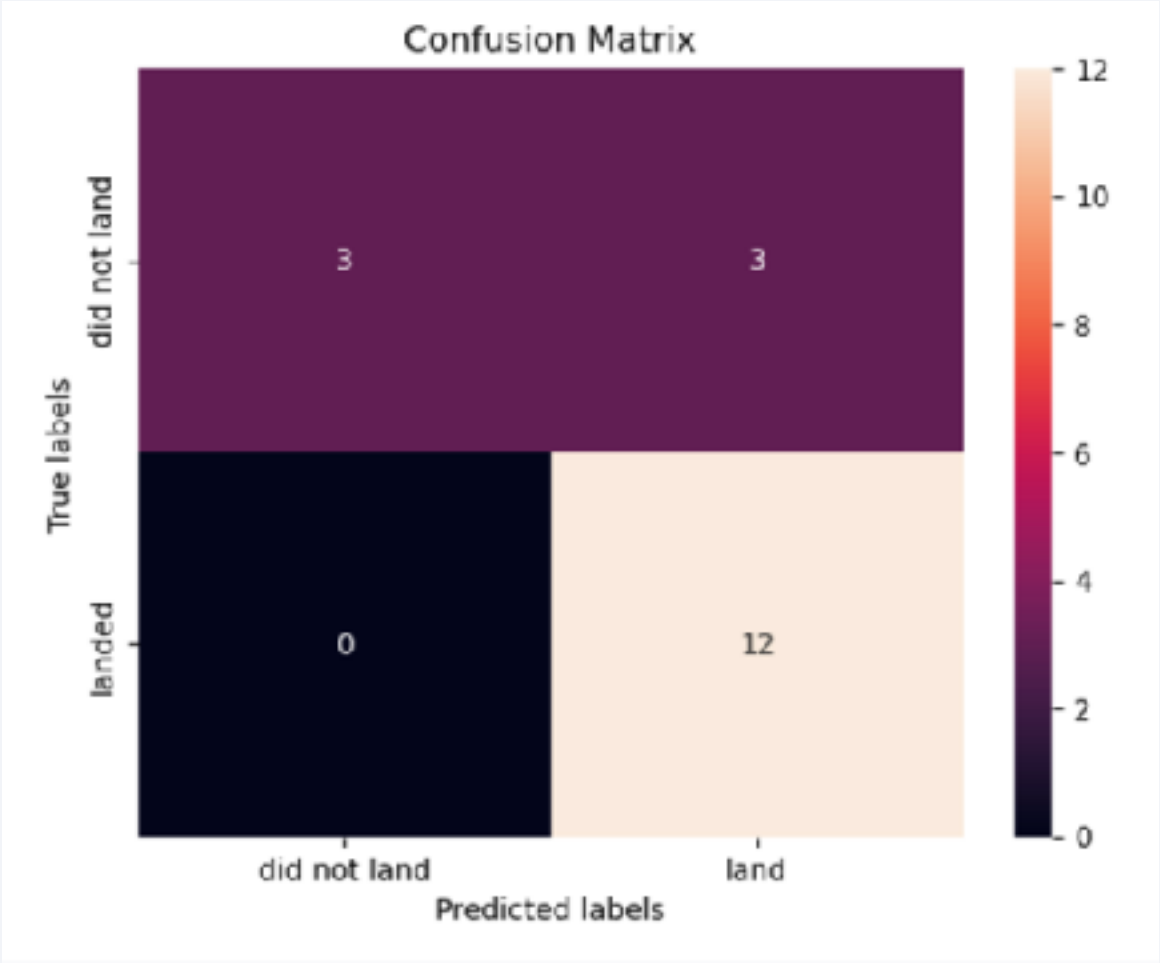
Confusion Matrix - KNN Model

Explanation

- **True Positives (TP)**: 12 → correctly predicted successes.
- **True Negatives (TN)**: 3 → correctly predicted failures.
- **False Positives (FP)**: 3 → failures predicted as successes.
- **False Negatives (FN)**: 0 → successes predicted as failures.

Insights:

- KNN predicts all actual successes correctly except for 3 failures that were misclassified as successes.
- There are **no false negatives**, meaning the model does not miss any successful launches.
- Overall, this confusion matrix aligns with the **83% accuracy**, showing strong performance in identifying successful launches.



Conclusions

- **Experience matters:** Launch success improves as **Flight Number increases**, reflecting operational learning over time.
- **Payload influences outcome:** Very heavy payloads slightly reduce the probability of success, while typical payloads achieve higher reliability.
- **Booster version impact:** Block 5 boosters show the **highest reliability**, highlighting the importance of hardware improvements.
- **Launch site performance differs:** Some sites, like **CCAFS SLC 40** and **KSC LC 39A**, have higher historical success rates, while orbit complexity affects outcomes (e.g., GTO and Polar missions show more failures).
- **Annual success trend rises:** SpaceX has become increasingly reliable in recent years, with success rates improving across all major booster versions.
- **Modeling insights:** Among the classification models, **KNN, Logistic Regression, and SVM** achieved the **highest predictive accuracy (~83%)**, making them the most reliable for predicting launch outcomes.

Appendix

- **Python Code & Notebooks**: https://github.com/RobertaFarris93/Datascience_capstone/blob/main/

- **Datasets**
  - Cleaned SpaceX launch dataset (`spacex_launch_data.csv`)

Thank you!