SpaceX Launch Data Analysis & Prediction Capstone

Interactive Visualization and Machine Learning Insights with Plotly & Dash

Author : Umme Sanjeda

Course : IBM Applied Data Science Capstone

November 22, 2025

Additional Info:

This courses is a part of IBM Data Science Professional Certificate

Outline

Executive Summary – Project Overview

Key Insights

- SpaceX launch performance shows a strong improvement trend over time, with success rates significantly increasing in recent years.
- Key determinants of launch success include booster version, launch site, flight number, and payload mass.
- EDA, geospatial mapping, SQL analysis, and predictive modeling reveal consistent patterns that support reliable launch decisions.

What This Project Achieves

- Provides a complete data-driven understanding of historical launch outcomes.
- Builds predictive models that estimate launch success with high accuracy (Random Forest & Logistic Regression).
- Supports strategic mission planning by identifying conditions that lead to optimal launch performance.

Executive Summary – Recommendations & Impact

Strategic Recommendations

- Prioritize upgraded booster versions (e.g., FT, B5+) for highest reliability.
- Leverage insights from payload mass and orbit type to fine-tune mission configurations.
- Utilize predictive model outputs to assess risk and optimize launch preparation cycles.

Operational Impact

- Enhances decision-making for scheduling and mission planning.
- Reduces uncertainty through consistent, data-backed launch success projections.
- Provides a scalable analytical pipeline for ongoing SpaceX launch evaluations.

Introduction

Problem Statement

SpaceX has become one of the world's leading private space companies, with a growing number of Falcon 9 launches. Understanding launch patterns, success factors, and payload performance is critical for business, engineering, and strategy decision-making.

The goal

To build a complete EDA-to-Dashboard-to-ML pipeline that provides insights into mission success and operational behavior.

Approach

- ✓ Data Collection: API + Web Scraping
- ✓ Data Wrangling & Cleaning
- ✓ Exploratory Data Analysis (EDA)
- ✓ Interactive Visualization (Dash + Folium)
- ✓ Machine Learning Predictions

Project Context & Analysis Scope

Context & Objective

- SpaceX is a private aerospace manufacturer focused on reliable and cost-effective rocket launches.
- The goal of this project is to analyze historical launch data to uncover insights, trends, and patterns that impact launch success.
- Use data-driven approaches to predict launch outcomes and recommend optimized strategies.

Scope of Analysis

- **Data sources:** SpaceX public launch records, web scraping, and API-based collection.
- **Analysis includes:** EDA, interactive dashboards, SQL queries, geospatial mapping, and predictive modeling.
- Final output aims to guide launch planning and decision-making using historical insights.

Data Collection & Wrangling

Data Collection – API Method

Objective

- Collect SpaceX Falcon 9 launch data via API
- Prepare data for further analysis

Data Collection Steps

- Used SpaceX API & static JSON for consistency
- Libraries: requests, pandas, numpy, datetime
- Key columns extracted: Rocket, Payloads, Launchpad, Cores, Flight Number, Date

```
[12]: # Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

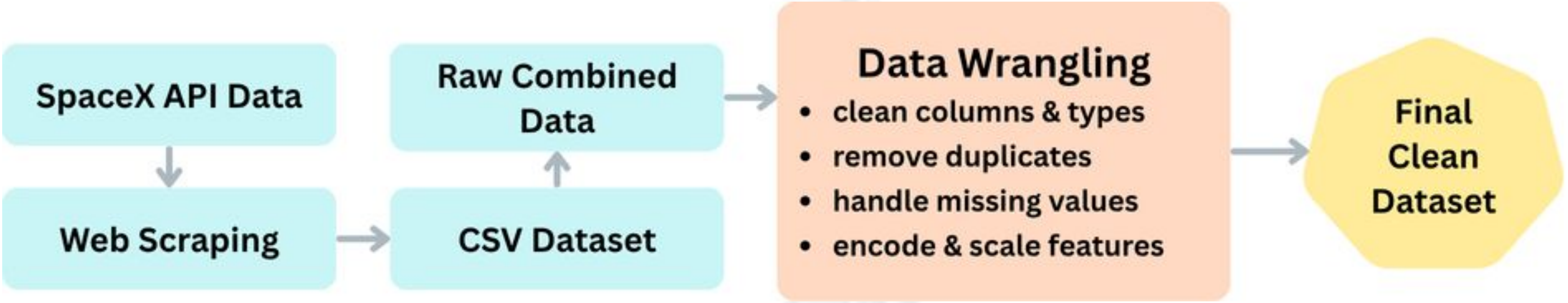
```
[28]: # Show first 5 rows
data.head()
```

[28]:

	rocket	payloads	launchpad	cores	flight_number	date_utc	date
0	5e9d0d95eda69955f709d1eb	5eb0e4b5b6c3bb0006eeb1e1	5e9e4502f5090995de566f86	{'core': '5e9e289df35918033d3b2623', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	1	2006-03-24T22:30:00.000Z	2006-03-24
1	5e9d0d95eda69955f709d1eb	5eb0e4b6b6c3bb0006eeb1e2	5e9e4502f5090995de566f86	{'core': '5e9e289ef35918416a3b2624', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	2	2007-03-21T01:10:00.000Z	2007-03-21
3	5e9d0d95eda69955f709d1eb	5eb0e4b7b6c3bb0006eeb1e5	5e9e4502f5090995de566f86	{'core': '5e9e289ef3591855dc3b2626', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	4	2008-09-28T23:15:00.000Z	2008-09-28
4	5e9d0d95eda69955f709d1eb	5eb0e4b7b6c3bb0006eeb1e6	5e9e4502f5090995de566f86	{'core': '5e9e289ef359184f103b2627', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	5	2009-07-13T03:35:00.000Z	2009-07-13
5	5e9d0d95eda69973a809d1ec	5eb0e4b7b6c3bb0006eeb1e7	5e9e4501f509094ba4566f84	{'core': '5e9e289ef359185f2b3b2628', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	6	2010-06-04T18:45:00.000Z	2010-06-04

Figure: Raw DataSet

Data Wrangling – Methodology



Data Pipeline Overview

- Merged datasets into a single DataFrame for analysis
- Handled missing values by removing or imputing where necessary
- Standardized column names and data types for consistency
- Extracted relevant features such as Launch Site, Payload Mass, Booster Version, Outcome

Initial Data Wrangling – Falcon 9

Key Steps

1. **ID Extraction:**
   - Rocket → Booster Version
   - Payload → Mass & Orbit
   - Launchpad → Site Name & Coordinates
   - Core → Outcome, Landing Type, Flights, GridFins, Reused, Legs, Block, Serial
2. **Filtering:**
   - Removed Falcon 1 launches → kept only Falcon 9
   - Reset FlightNumber sequentially
3. **Handling Missing Values:**
   - PayloadMass: replaced NaN with mean (~7,358 kg)
   - LandingPad: kept None for launches without a pad

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	Gri
1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	
2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	
4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	
5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	
6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	

Figure: Summary of Pandas Dataframe

Data Wrangling – Raw Dataset Overview

Key Insights

- Dataset imported from:
- “dataset_part_1.csv” (Skills Network Cloud Storage)
- Contains 90 Falcon 9 launches with 17 features
- Includes launch metadata: Flight Number, Date, Booster Version, Payload, Orbit
- Includes landing outcome needed for training the classifier
- Missing values identified (e.g., LandingPad ≈ 29%)
- No transformations yet — raw structure used for further wrangling

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None

Figure: Data Wrangling – Sample of Raw Dataset Overview

umme Sanjeda

Data Wrangling – Missing Values Summary

Missing Values Identified in Raw Dataset

- **PayloadMass:** 5 missing values
- Handled by replacing missing entries with the mean payload mass (≈ 6104.96 kg).
- **LandingPad:** 26 missing values
- Retained as None because many early missions did not use landing pads (ocean landings).
- **All other columns:** 0 missing values
- No additional imputations required.

```
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan, payload_mean)

# Verify missing values
data_falcon9.isnull().sum()
```

```
[32]: FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       0
Orbit             0
LaunchSite        0
Outcome           0
Flights           0
GridFins          0
Reused            0
Legs              0
LandingPad       26
Block             0
ReusedCount       0
```

Result

✓ Dataset fully cleaned except for intentional None values in LandingPad.

✓ Ready for EDA, visualization, and predictive modeling.

Figure: Missing Values

Summary of Data Collection & Wrangling

Data Sources

- ✓ SpaceX REST API — primary source for Falcon 9 launch data
- ✓ Provided Web-Scraped Dataset (spacex_web_scraped.csv) — used for consistency across labs

Key Data Collection Steps

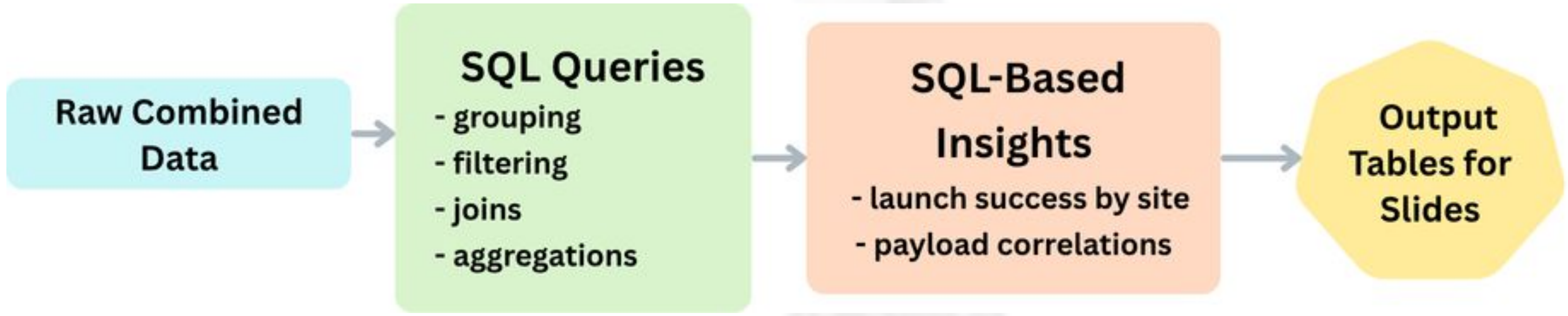
- ✓ Retrieved raw launch records from the official SpaceX API
- ✓ Extracted relevant fields from nested JSON (rocket, payloads, cores, launchpad)
- ✓ Filtered out Falcon 1 launches → kept only Falcon 9 data
- ✓ Created sequential FlightNumber and standardized column names
- ✓ Exported cleaned dataset (dataset_part_1.csv) for further analysis

Final Output

- ✓ **90 records**
- ✓ **17 structured features**
- ✓ **Fully cleaned and analysis-ready dataset**

SQL-Based Exploratory Data Analysis (EDA)

SQL-Based EDA



- ➢ This analysis uses SQL queries to explore and extract insights from the SpaceX mission dataset.
- ➢ **Objectives of this lab include:**
  - Understanding SpaceX mission data, including launch sites, payloads, booster versions, and landing outcomes.
  - Loading the dataset into a relational database for structured querying.
  - Using SQL to perform aggregations, filtering, ranking, and subqueries to answer real-world questions.
  - Identifying patterns that can help predict the success of booster landings.
- ➢ The insights gained can be applied to operational planning, cost estimation, and mission performance analysis.
- ➢ This slide deck presents the findings in a step-by-step, SQL query-driven exploration, highlighting key metrics and mission outcomes.

Unique Launch Sites

SQL Query

```
SELECT DISTINCT "Launch_Site"
FROM SPACEX_DATA
```

Result

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

```
]: cur.execute('SELECT DISTINCT "Launch_Site" FROM SPACEX_DATA')
   rows = cur.fetchall()

   print("Unique Launch Sites:")
   for row in rows:
       print(row[0])
```

```
Unique Launch Sites:
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

Figure: Unique launch sites in the SpaceX dataset

Launch Sites Starting with "CCA"

SQL Query

```
SELECT * FROM SPACEX_DATA WHERE
"Launch_Site" LIKE "CCA%" LIMIT 5
```

```
First 5 records where Launch_Site starts with 'CCA':
('2010-06-04', '18:45:00', 'F9 v1.0  B0003', 'CCAFS LC-40', 'Dragon Spacecraft Qualification Unit', 0, 'LEO', 'SpaceX', 'Succ
ess', '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)', 'Succes
s', '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 att
empt')
('2013-03-01', '15:10:00', 'F9 v1.0  B0007', 'CCAFS LC-40', 'SpaceX CRS-2', 677, 'LEO (ISS)', 'NASA (CRS)', 'Success', 'No at
tempt')
```

Result

Figure: Sample missions from launch sites beginning with “CCA”

First 5 Records:

- ✓ '2010-06-04', 'CCAFS LC-40', Dragon Spacecraft Qualification Unit
- ✓ '2010-12-08', 'CCAFS LC-40', Dragon demo flight C1
- ✓ '2012-05-22', 'CCAFS LC-40', Dragon demo flight C2
- ✓ '2012-10-08', 'CCAFS LC-40', SpaceX CRS-1
- ✓ '2013-03-01', 'CCAFS LC-40', SpaceX CRS-2

Payload Mass by Customer and Booster

SQL Query

```
'SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEX_DATA WHERE
"Customer"="NASA (CRS)"'

'SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEX_DATA WHERE
"Booster_Version"="F9 v1.1"'

'SELECT "Booster_Version", "PAYLOAD_MASS__KG_"

    FROM SPACEX_DATA

    WHERE "PAYLOAD_MASS__KG_" = (

        SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEX_DATA')
```

```
Total payload mass carried by boosters launched by NASA (CRS): 45596 kg

Average payload mass carried by booster version F9 v1.1: 2928.4 kg

Booster versions that carried the maximum payload mass:
F9 B5 B1048.4 - 15600 kg
F9 B5 B1049.4 - 15600 kg
F9 B5 B1051.3 - 15600 kg
F9 B5 B1056.4 - 15600 kg
F9 B5 B1048.5 - 15600 kg
F9 B5 B1051.4 - 15600 kg
F9 B5 B1049.5 - 15600 kg
F9 B5 B1060.2  - 15600 kg
F9 B5 B1058.3  - 15600 kg
F9 B5 B1051.6 - 15600 kg
F9 B5 B1060.3 - 15600 kg
F9 B5 B1049.7  - 15600 kg
```

Figure: Payload mass insights per customer and booster version

Result

- ✓ Total payload mass by NASA (CRS): 45,596 kg
- ✓ Average payload mass for F9 v1.1: 2,928.4 kg
- ✓ Max payload mass: 15,600 kg (multiple F9 Block 5 boosters)

umme Sanjeda

First Successful Ground Landing

SQL Query

```
SELECT MIN("Date") FROM
SPACEX_DATA WHERE
"Landing_Outcome"="Success (ground
pad)"
```

Result

✓ 2015-12-22

```
cur.execute('''
    SELECT MIN("Date")
    FROM SPACEX_DATA
    WHERE "Landing_Outcome"="Success (ground pad)"
''')
first_successful_landing = cur.fetchone()[0]

print("Date of first successful landing on ground pad:", first_successful_landing)
```

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

Figure: Date of the first successful ground pad landing.

Successful Drone Ship Landings with Payload Criteria

SQL Query

```
'SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
    FROM SPACEX_DATA
    WHERE "PAYLOAD_MASS__KG_" = (
        SELECT MAX("PAYLOAD_MASS__KG_") FROM
SPACEX_DATA)'
```

```
Boosters with successful drone ship landing and payload mass 4000-6000 kg:
F9 FT B1022
F9 FT B1026
F9 FT  B1021.2
F9 FT  B1031.2
```

Figure: Boosters successfully landing on drone ship with 4,000–6,000 kg payload.

Result

Boosters meeting criteria:

- ✓ F9 FT B1022
- ✓ F9 FT B1026
- ✓ F9 FT B1021.2
- ✓ F9 FT B1031.2

Mission Outcome Counts

SQL Query

```
'SELECT "Mission_Outcome", COUNT(*)
    FROM SPACEX_DATA
    GROUP BY "Mission_Outcome"'
```

Result

Total outcomes

- ✓ Success: 98
- ✓ Failure (in flight): 1
- ✓ Success (duplicate / payload unclear): 2

```
Total number of each mission outcome:
Failure (in flight) : 1
Success : 98
Success  : 1
Success (payload status unclear) : 1
```

Figure: Total mission outcomes for all SpaceX launches.

Boosters Carrying Maximum Payload

Result

✓ Maximum payload flown: 15,600 kg

✓ Boosters: F9 B5 B1048.4, B1049.4, B1051.3, … (12 boosters)

```
Booster versions that carried the maximum payload mass:
F9 B5 B1048.4 - 15600 kg
F9 B5 B1049.4 - 15600 kg
F9 B5 B1051.3 - 15600 kg
F9 B5 B1056.4 - 15600 kg
F9 B5 B1048.5 - 15600 kg
F9 B5 B1051.4 - 15600 kg
F9 B5 B1049.5 - 15600 kg
F9 B5 B1060.2  - 15600 kg
F9 B5 B1058.3  - 15600 kg
F9 B5 B1051.6 - 15600 kg
F9 B5 B1060.3 - 15600 kg
F9 B5 B1049.7  - 15600 kg
```

Figure: List of boosters carrying the maximum payload.

Drone Ship Landing Failures in 2015

```
Failure landings on drone ship in 2015:
Month: 01 | Landing Outcome: Failure (drone ship) | Booster: F9 v1.1 B1012 | Launch Site: CCAFS LC-40
Month: 04 | Landing Outcome: Failure (drone ship) | Booster: F9 v1.1 B1015 | Launch Site: CCAFS LC-40
```

Figure: Landing failures on drone ship in 2015 by month and booster.

Result

- ✓ Month 01 – Booster B1012
- ✓ Month 04 – Booster B1015

Landing Outcome Counts (2010–2017)

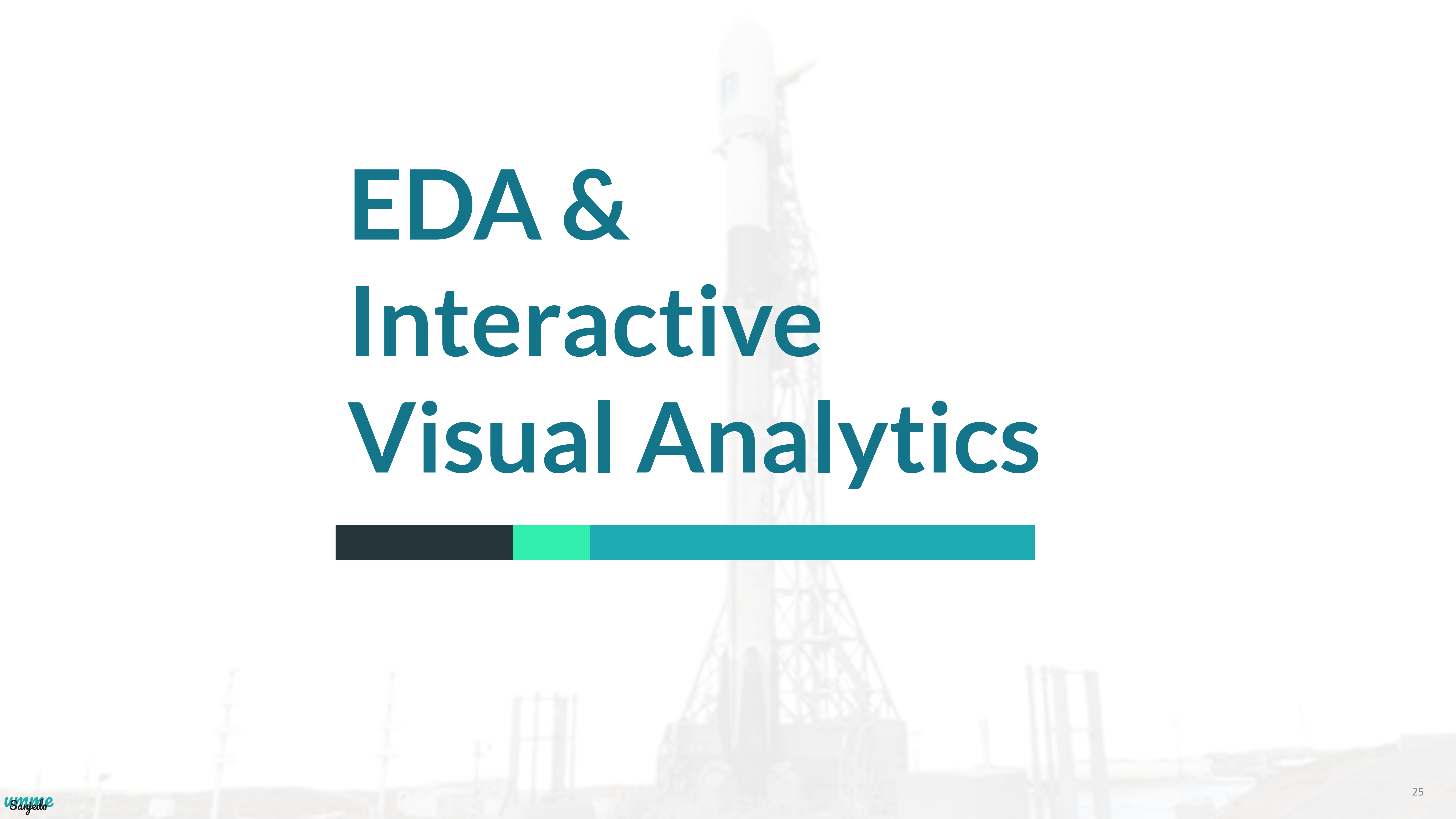
Result

Ranked outcomes (descending)

- ✓ 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

```
Landing outcome counts between 2010-06-04 and 2017-03-20 (descending):
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
```

Figure: Ranked counts of landing outcomes from 2010–2017



EDA & Interactive Visual Analytics

EDA Methodology

Purpose of EDA

The purpose of Exploratory Data Analysis (EDA) in this project was to understand the underlying structure of the SpaceX launch dataset, identify important variables influencing launch success, and uncover patterns, trends, and anomalies that guide both visualization and predictive modeling.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40

Figure: SpaceX dataset in Pandas datafram

Approach

Used a combination of statistical summaries, graphical analyses, and interactive tools to explore the dataset from multiple perspectives. This allowed us to validate assumptions, detect outliers, assess distributions, and understand relationships between features.

Statistical & Visual Techniques

Statistical Techniques

- Descriptive statistics: mean, median, standard deviation
- Distribution analysis using boxplots and histograms
- Correlation analysis to detect relationships between numeric features
- Group-by aggregations to study success rates across sites, boosters, and orbit types

Approach

- Histograms — to assess launch frequency and numeric variable distributions
- Boxplots — to detect outliers (e.g., payload mass)
- Scatterplots — to observe relationships like payload mass vs launch success
- Bar charts — to compare success counts across launch sites
- Heatmap — to visualize correlation strength among features

Interactive Visual Analytics

Interactive Analytics Tools

- To enhance interpretability, we used interactive visual tools that allow dynamic filtering and exploration:
- Plotly for responsive charts
- Folium for interactive geospatial mapping
- Plotly Dash for creating an analytical dashboard

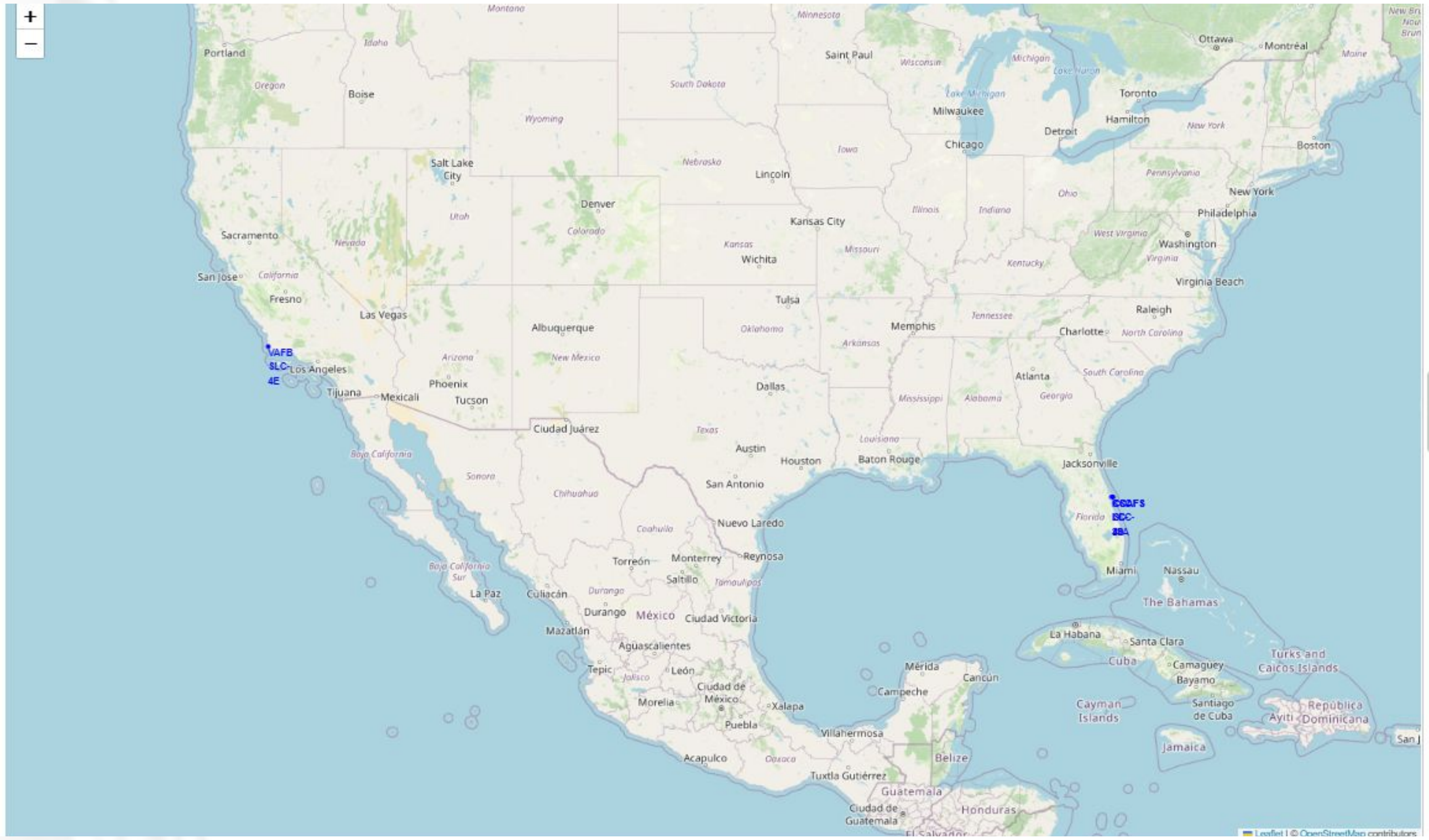


Figure: A Folium Map

Why Interactivity Matters

Interactive exploration helped identify:

- Payload ranges with higher success rates
- Booster categories with consistently strong performance
- Site-specific behavior visible only when filtering dynamically
- Spatial launch location insights that static charts can't visualize

EDA RESULTS

Flight Number vs Payload Mass

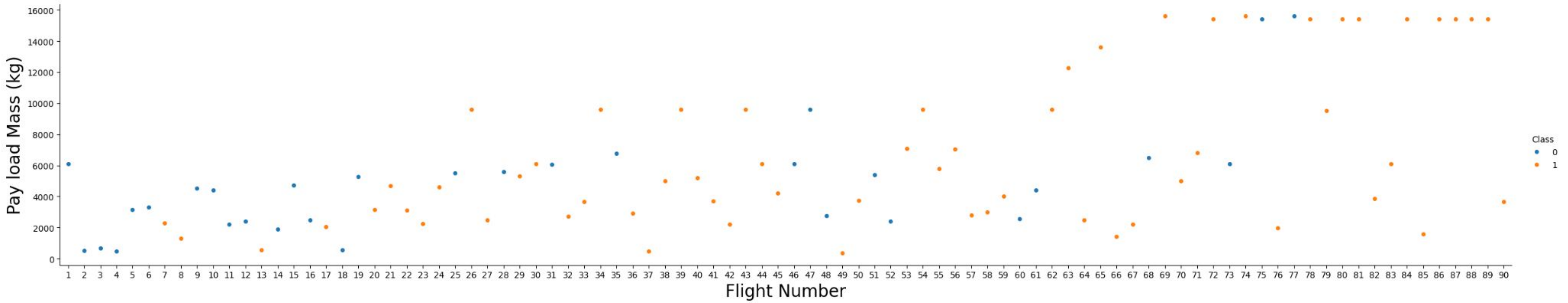


Figure: Relationship between flight attempts, payload mass, and landing success.

Key Insights

- ➢ Scatter plot overlaying launch outcome on Flight Number vs Payload Mass.
- ➢ **Observations:**
  - ✓ Higher flight numbers → higher chance of successful landing.
  - ✓ Heavy payloads → lower likelihood of successful landing.

umme
Sanjeda

Flight Number vs Launch Site

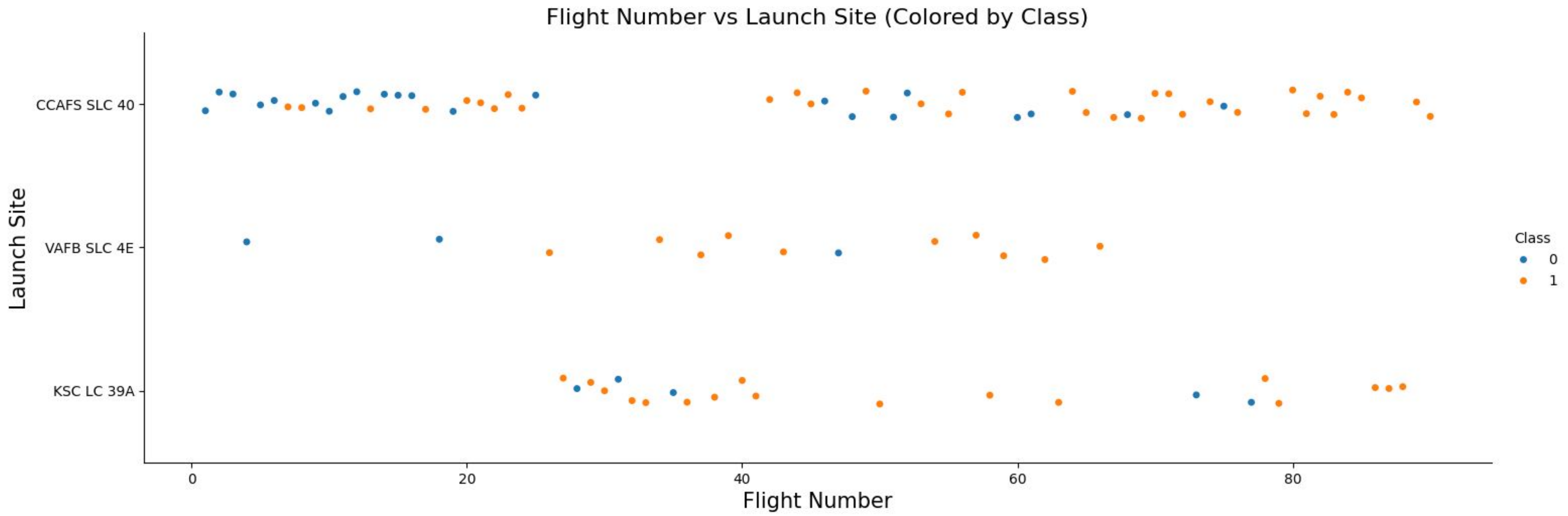


Figure: Flight number progression across different launch sites.

Key Insights

- ➢ Scatter plot showing launches at different launch sites.
- ➢ **Observations:**
  - ✓ CCAFS sites dominate early launches.
  - ✓ Success rate iNcreases with experience at all sites.

Payload vs Launch Site

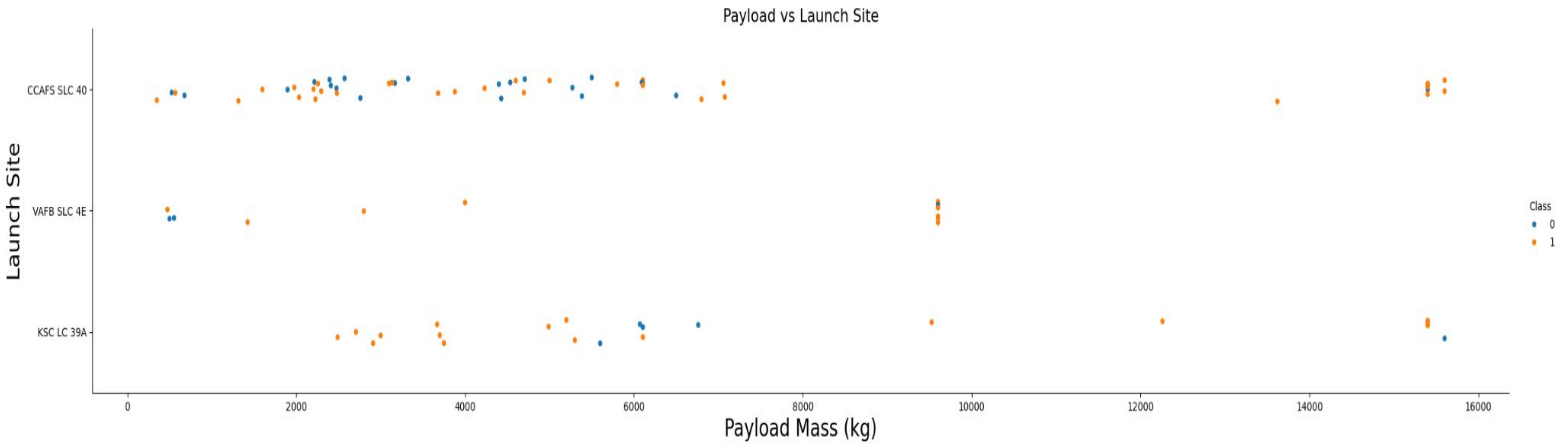


Figure: Scatter Plot: Payload distribution across launch sites.

Key Insights

- ➢ Scatter plot showing payload mass for each launch site.
- ➢ **Observations:**
  - ✓ VAFB SLC 4E launches lighter payloads (<10,000 kg).
  - ✓ CCAFS sites handle a wider range of payload masses.

Success Rate by Orbit Type

Key Insights

- ➢ Scatter plot showing Flight Number vs Orbit Type colored by success.

- ➢ **Observations:**
  - ✓ In LEO, success improves with higher flight numbers.
  - ✓ No clear pattern for GTO.

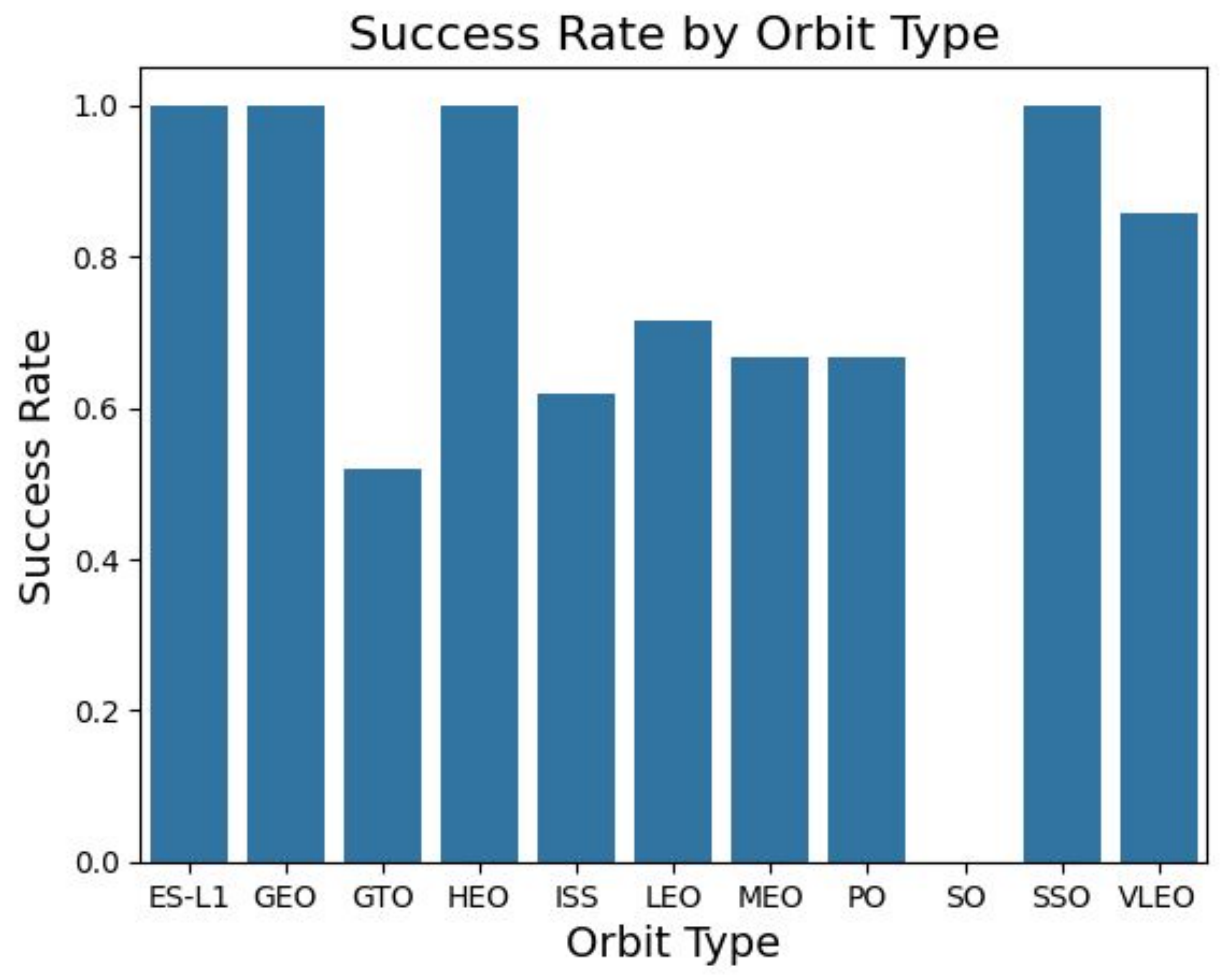


Figure: Average landing success per orbit type

umme
Sanjeda

Flight Number vs Orbit Type

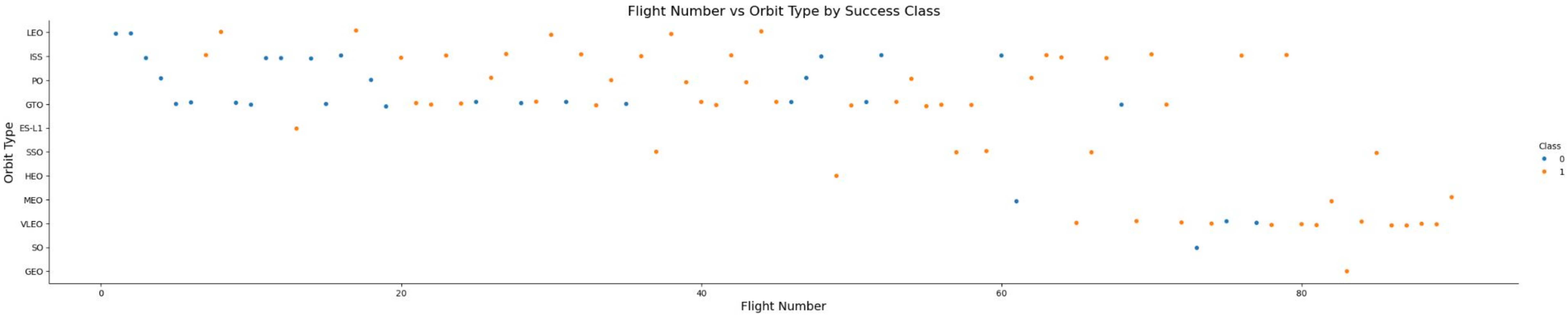


Figure: Scatterplot- Effect of flight experience on landing success across orbit types

Key Insights

- ➢ Scatter plot showing Flight Number vs Orbit Type colored by success.
- ➢ **Observations:**
  - ✓ In LEO, success improves with higher flight numbers.
  - ✓ No clear pattern for GTO.

Payload vs Orbit Type

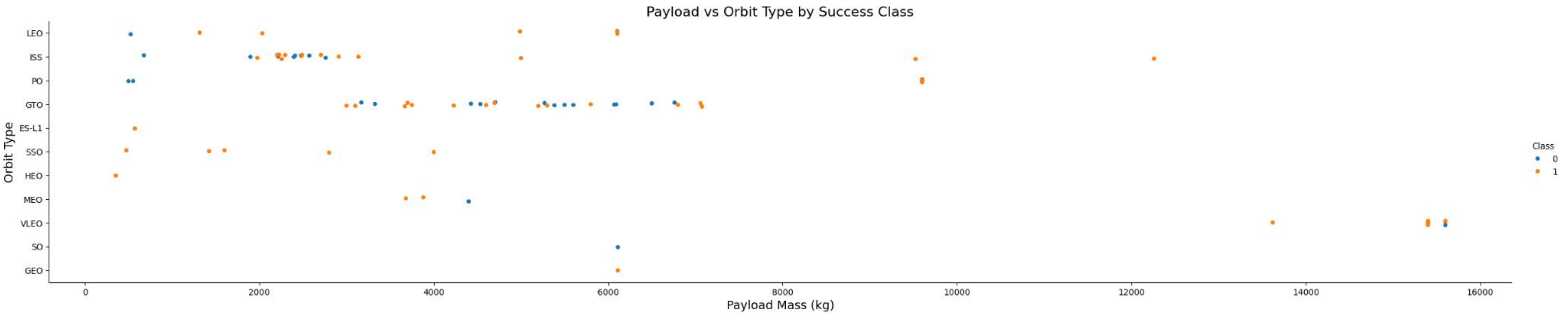


Figure: Scatterplot- Impact of payload mass on landing success by orbit type.

Key Insights

- ➢ Scatter plot of Payload vs Orbit Type.
- ➢ **Observations:**
  - ✓ Heavier payloads more likely to land successfully in Polar, LEO, and ISS.
  - ✓ GTO launches show mixed outcomes.

Yearly Launch Success Trend

Key Insights

- ➢ Line chart of average success rate by year.
- ➢ **Observations:**
  - ✓ Success rate improves steadily from 2013 to 2017.
  - ✓ 2014 shows a plateau; 2015 onwards sees clear upward trend.

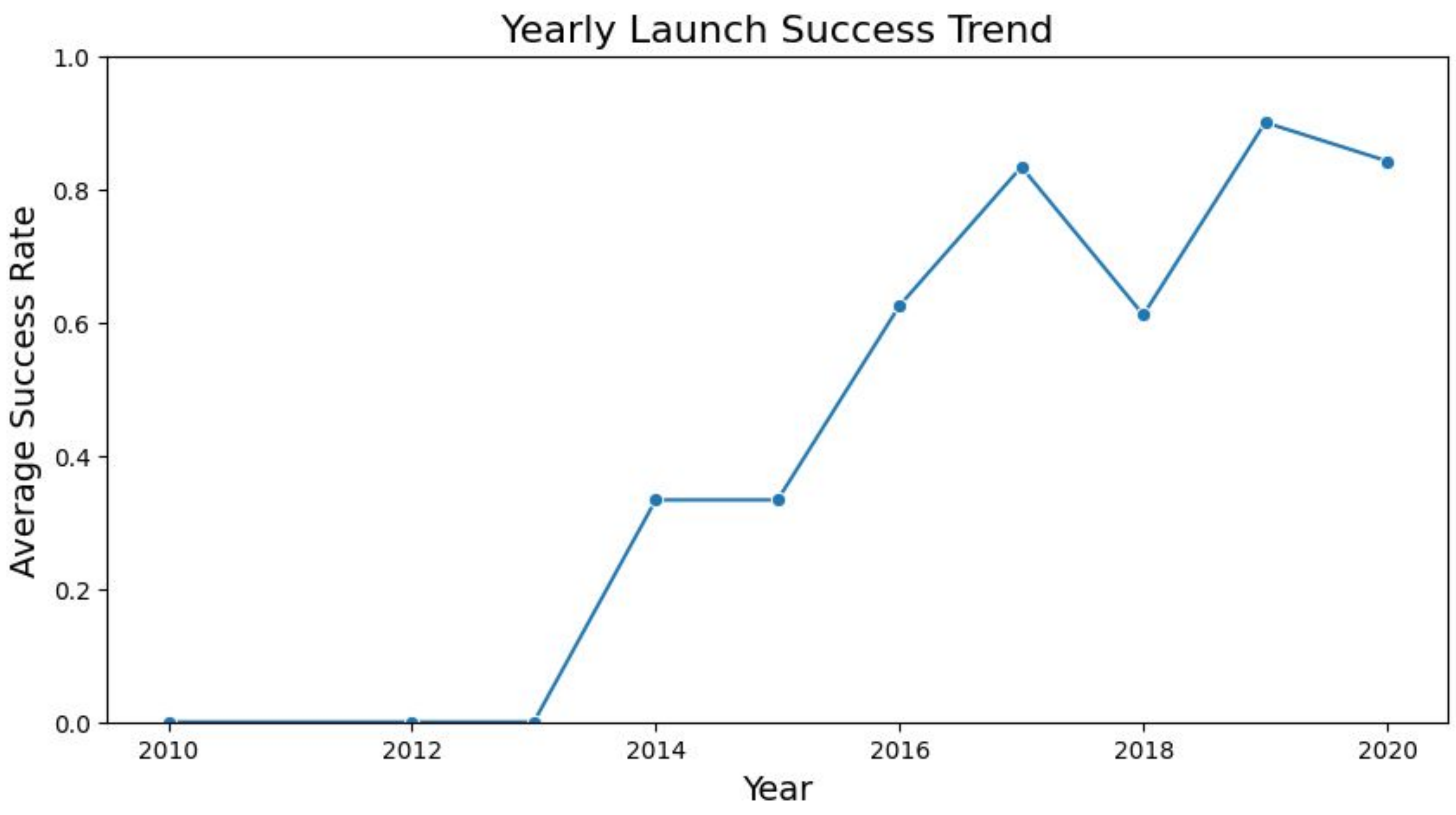


Figure: Annual trend of Falcon 9 landing success.

Feature Engineering for Prediction

Key Insights

- Selected features for prediction
  - ✓ `FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial.`
- Categorical variables (`Orbit, LaunchSite, LandingPad, Serial`) converted using one-hot encoding.
- All numeric columns cast to float64 for modeling.

	FlightNumber	Date	BoosterVersion	PayloadMass	Outcome	Flights	GridFins	Reused	Legs	Block	...
0	1.0	2010-06-04	Falcon 9	6104.959412	None None	1.0	False	False	False	1.0	...
1	2.0	2012-05-22	Falcon 9	525.000000	None None	1.0	False	False	False	1.0	...
2	3.0	2013-03-01	Falcon 9	677.000000	None None	1.0	False	False	False	1.0	...
3	4.0	2013-09-29	Falcon 9	500.000000	False Ocean	1.0	False	False	False	1.0	...
4	5.0	2013-12-03	Falcon 9	3170.000000	None None	1.0	False	False	False	1.0	...

5 rows × 87 columns

Figure: Feature matrix after encoding categorical variables and casting numeric types.

Key Insights from EDA

- Flight experience and payload mass significantly affect booster landing success.
- Launch site and orbit type are important contextual factors.
- Success rate shows consistent improvement over years.
- Preprocessing prepares dataset for predictive modeling in future modules.

Interactive Map with Folium

Geospatial Mapping of SpaceX Launch Sites

Key Insights

- ➢ All four SpaceX launch sites plotted on an interactive Folium map.
- ➢ **Coordinates (latitude & longitude) used to accurately mark site locations:**
  - ✓ CCAFS LC-40: (28.562302, -80.577356)
  - ✓ CCAFS SLC-40: (28.563197, -80.576820)
  - ✓ KSC LC-39A: (28.573255, -80.646895)
  - ✓ VAFB SLC-4E: (34.632834, -120.610745)
- ➢ Sites concentrated on the U.S. east coast and California — low-latitude positions favorable for orbital launches.
- ➢ Base map supports further outcome and proximity analysis.

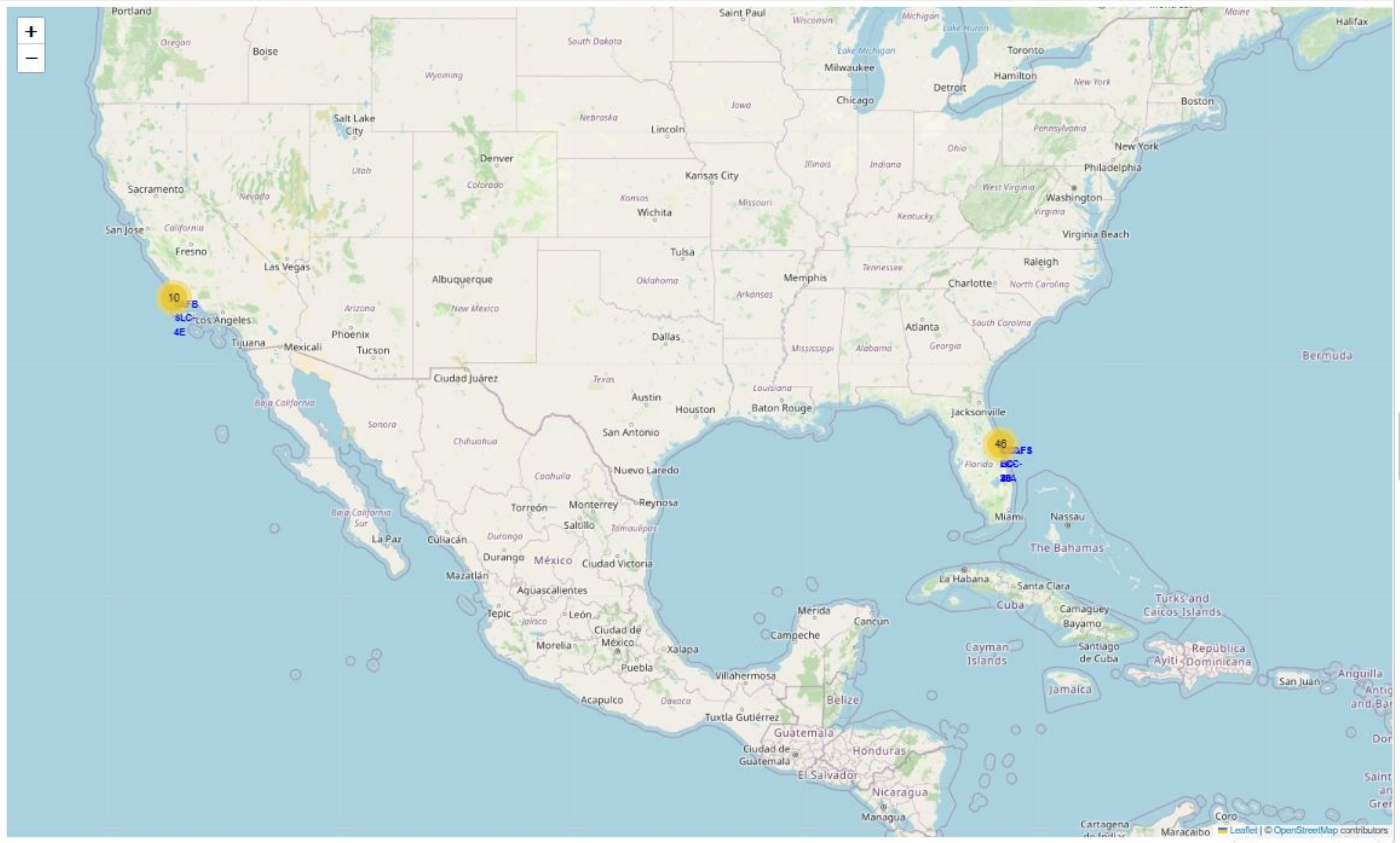


Figure: Folium map displaying the geographic distribution of SpaceX launch sites

umme Sanjeda

Success vs Failure Visualization Using MarkerCluster

Key Insights

- ➢ MarkerCluster displays historical launches per site.
  - ✓ Green markers = success (class = 1)
  - ✓ Red markers = failure (class = 0)
- ➢ **Interpretation:**
  - ✓ KSC LC-39A shows consistently high success density.
  - ✓ CCAFS SLC-40 shows mixed outcomes (both red & green clusters).
  - ✓ Useful for quickly identifying site reliability and problem areas.

Figure: Distribution of launch success and failure outcomes across SpaceX launch sites

Proximity Analysis: CCAFS SLC-40 → Coastline, Highway, Railway, City

Key Insights

- **Distances measured using haversine formula (km):**
  - ✓ Coastline: 0.84 km
  - ✓ Nearest highway: 0.60 km
  - ✓ Nearest railway: 21.94 km
  - ✓ Nearest city (Melbourne, FL): 54.29 km
- **Interpretation:**
  - Very close to coastline (0.84 km) — ideal to launch over water and reduce risk to populated areas.
  - Highway access (0.60 km) enables logistics and emergency response.
  - Railway (~22 km) is reasonably nearby for heavy-component transport.
  - Large buffer to city (~54 km) meets safety and noise/impact requirements.

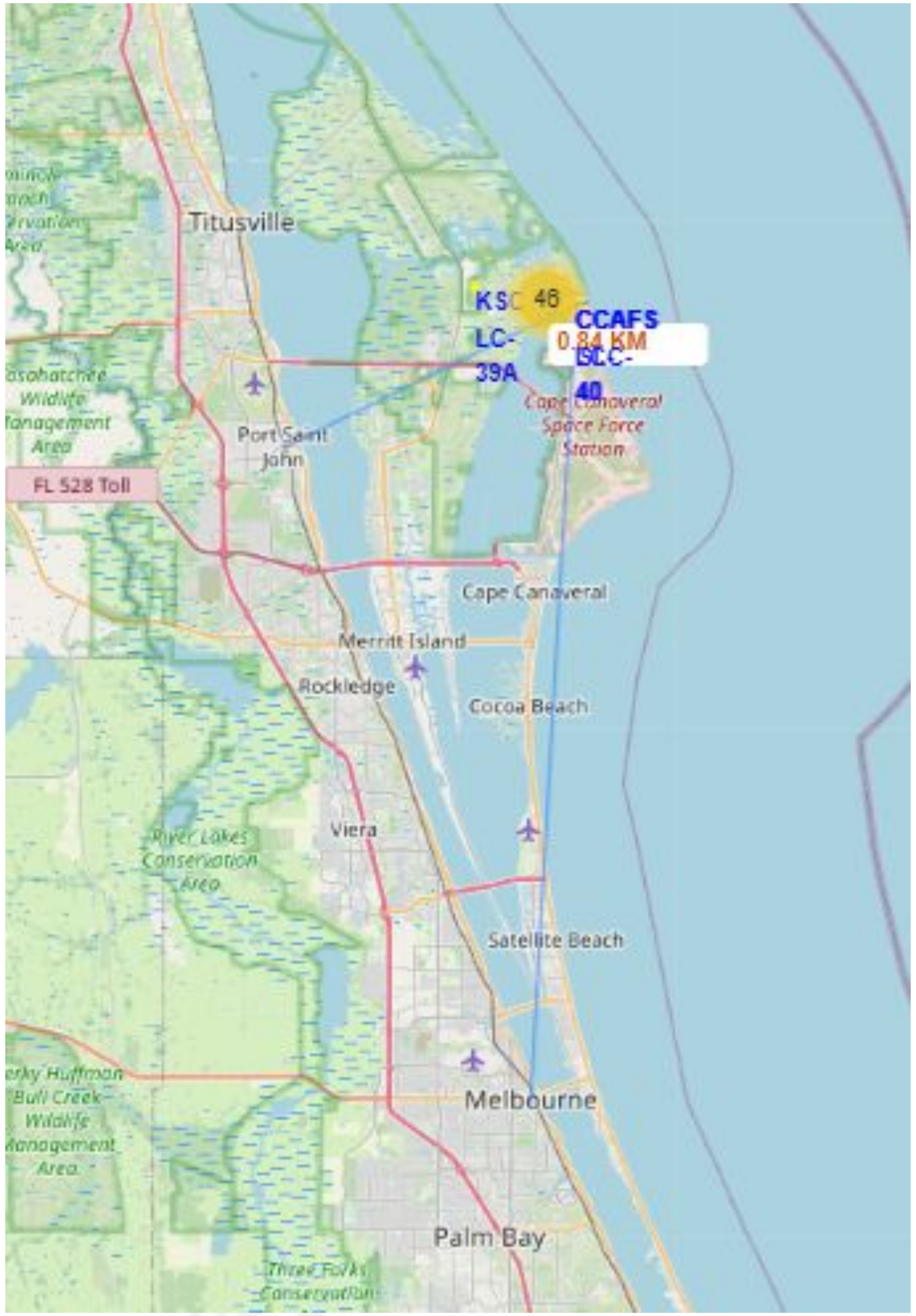


Figure: Distance lines from CCAFS SLC-40 to nearby infrastructure

Overall Summary of SpaceX Launch Site Geospatial Insights

- Folium mapping exposes geographic advantages: coastal placement & low-latitude positioning.
- MarkerCluster outcome mapping highlights site-specific reliability differences.
- Proximity metrics (e.g., coastline = 0.84 km) confirm sites are intentionally near water and transport routes while remaining well-separated from major population centers.
- These spatial findings inform safety planning, logistics, and site selection considerations.

Predictive Analysis & Machine Learning

Machine Learning Prediction Framework

Overview

- The goal is to predict Falcon 9 first-stage landing success (binary classification).
- **Dataset split using `train_test_split` with:**
  - 80% training, 20% testing, `random_state = 2`
  - Test set contains 18 samples
- Hyperparameter tuning performed using GridSearchCV (cv = 10).
- **Four models evaluated:**
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)

Key Method Steps

1. **Train/Test Split**
2. **10-fold Cross-Validation**
3. **Grid Search Hyperparameter Tuning**
4. **Model Selection Based on Test Accuracy**

umme Sanjeda

Logistic Regression Performance

Key Insights

- **Best Hyperparameters:**
  - ✓ C = 0.01
  - ✓ Penalty = L2
  - ✓ Solver = lbfgs
- **Validation Performance (CV = 10):**
  - ✓ Accuracy: 0.8464
- **Test Set Performance:**
  - ✓ Accuracy: 0.83

Interpretation

- Model correctly predicts most landings.
- True Positives = 12
- False Positives = 3
- Model tends to over-predict “landing success”.

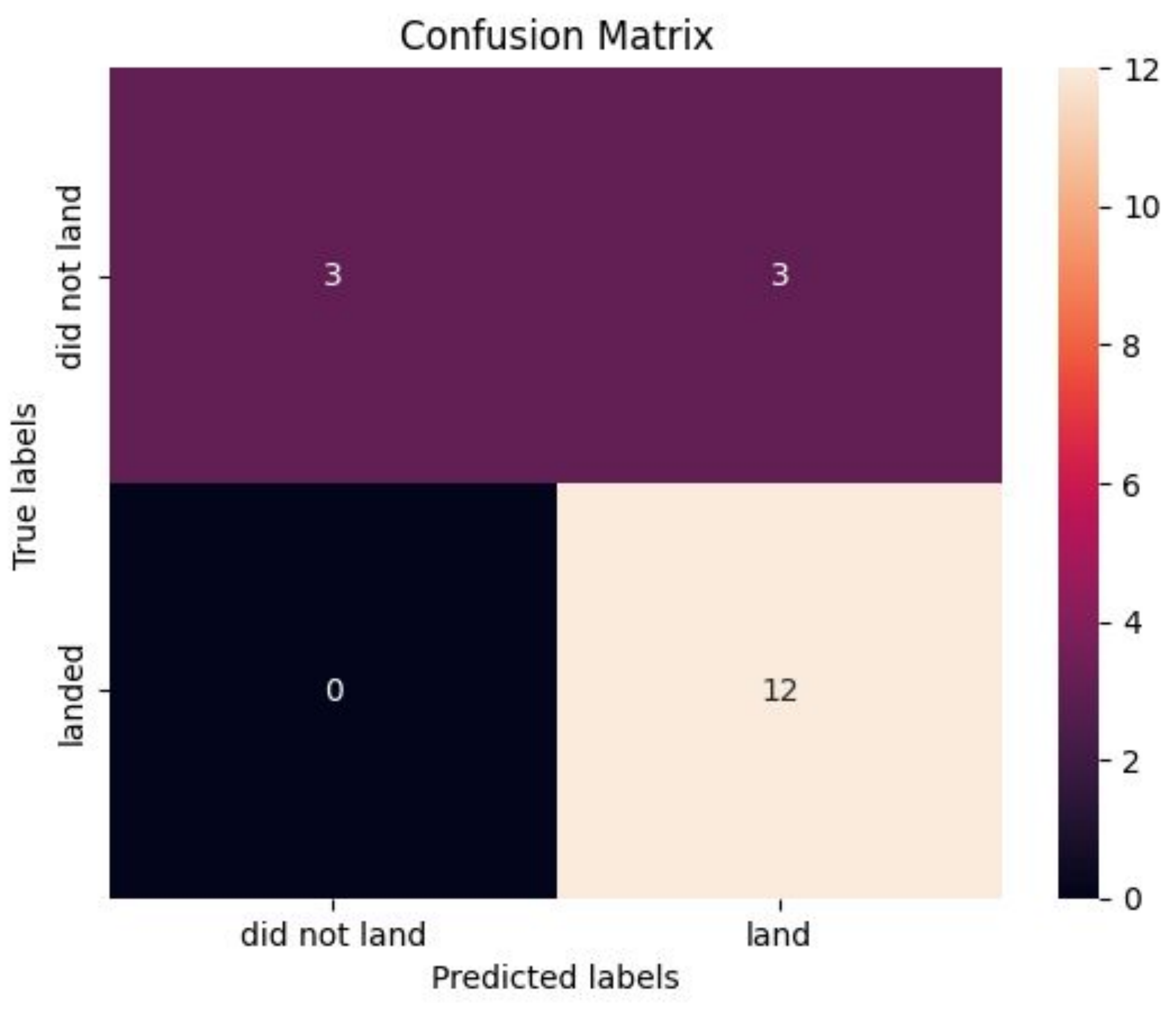


Figure: Confusion Matrix - Logistic Regression

Support Vector Machine (SVM) Performance

Key Insights

- ➢ **Best Hyperparameters:**
  - ✓ Kernel: sigmoid
  - ✓ C = 1.0
  - ✓ Gamma = 0.0316
- ➢ **Validation Performance (CV = 10):**
  - ✓ Accuracy: 0.8482
- ➢ **Test Set Performance:**
  - ✓ Accuracy: 0.83

Interpretation

- ➢ Model correctly predicts most landings.
- ➢ True Positives = 12
- ➢ False Positives = 3
- ➢ Model tends to over-predict “landing success”.

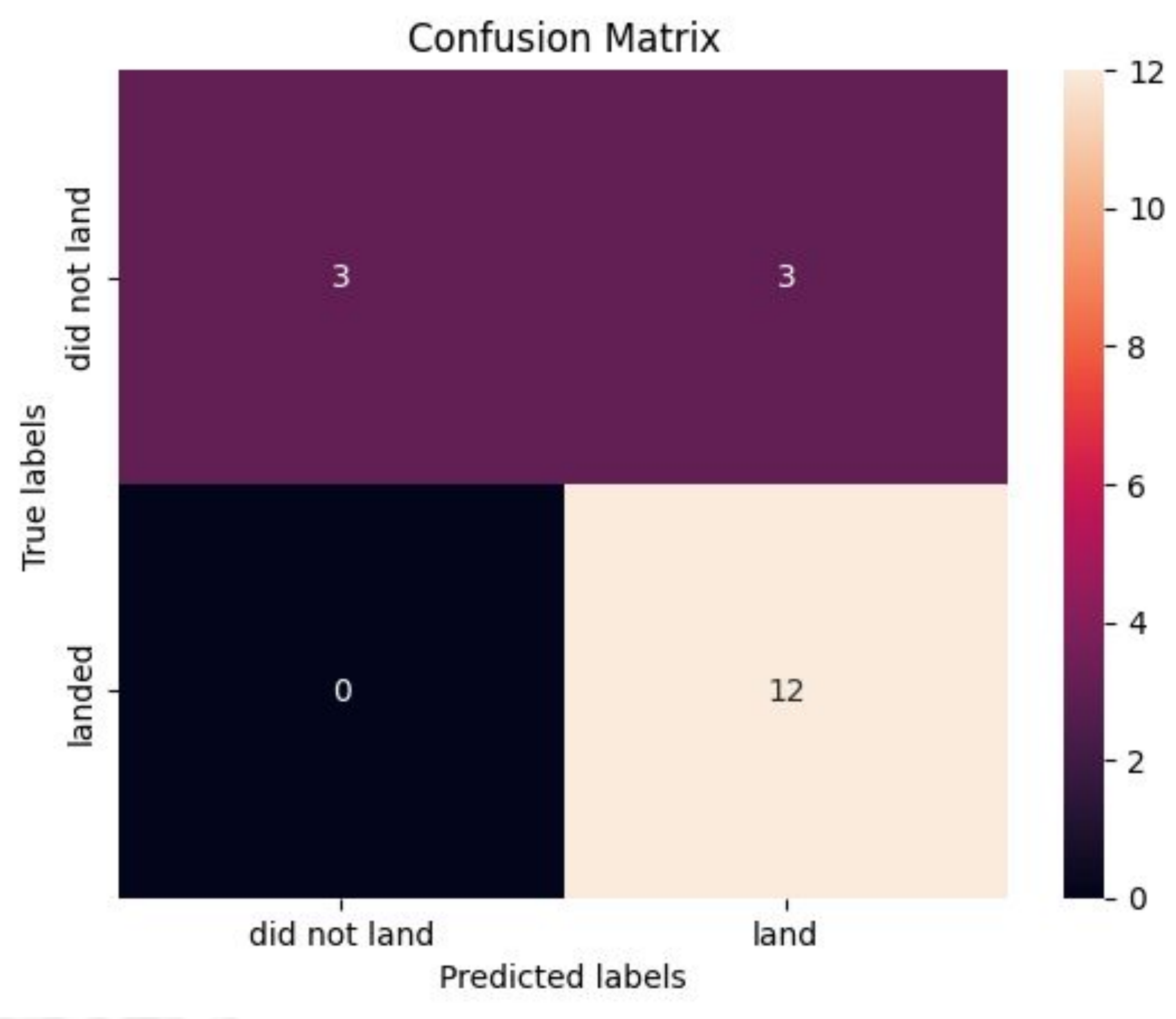


Figure: Confusion Matrix -SVM

Decision Tree Performance

Key Insights

- **Best Hyperparameters:**
  - ✓ Criterion: entropy
  - ✓ Splitter: random
  - ✓ Max Depth: 10
  - ✓ Max Features: sqrt
  - ✓ Min Samples Leaf: 1
  - ✓ Min Samples Split: 10
- **Validation Performance (CV = 10):**
  - ✓ Accuracy: 0.8768 (best among models)
- **Test Set Performance:**
  - ✓ Accuracy: 0.78 (lowest among models)

Interpretation

- Strong performance during training/validation.
- Overfitting evident due to lower test accuracy.

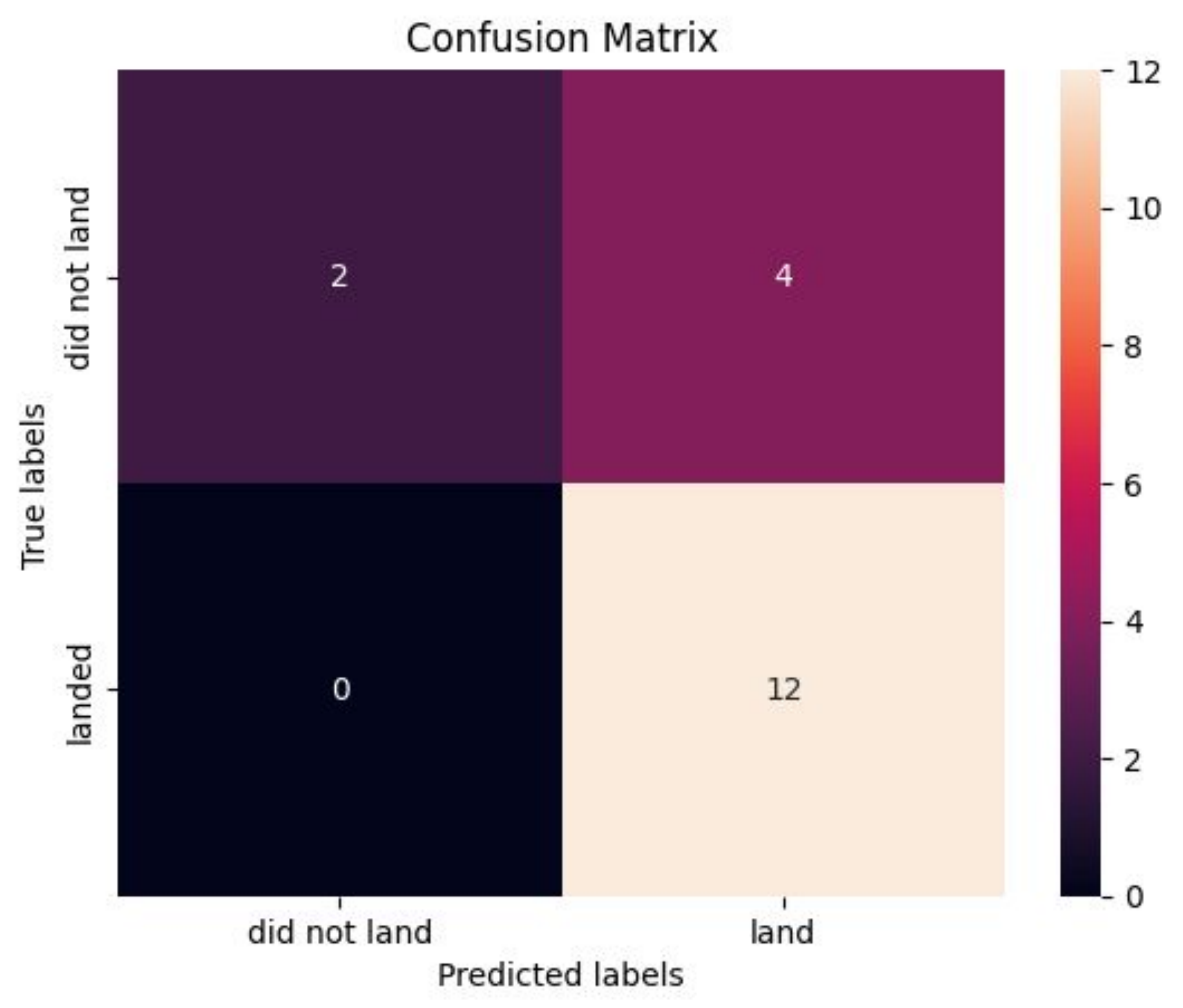


Figure: Confusion Matrix - Decision Tree

Decision Tree Performance

Key Insights

- **Best Hyperparameters:**
  - ✓ n_neighbors = 10
  - ✓ algorithm = auto
  - ✓ p = 1 (Manhattan distance)
- **Validation Performance (CV = 10):**
  - ✓ Accuracy: 0.8482
- **Test Set Performance:**
  - ✓ Accuracy: 0.83

Interpretation

- Performs well and comparable to LR/SVM.
- High k value leads to smoother decision boundary.

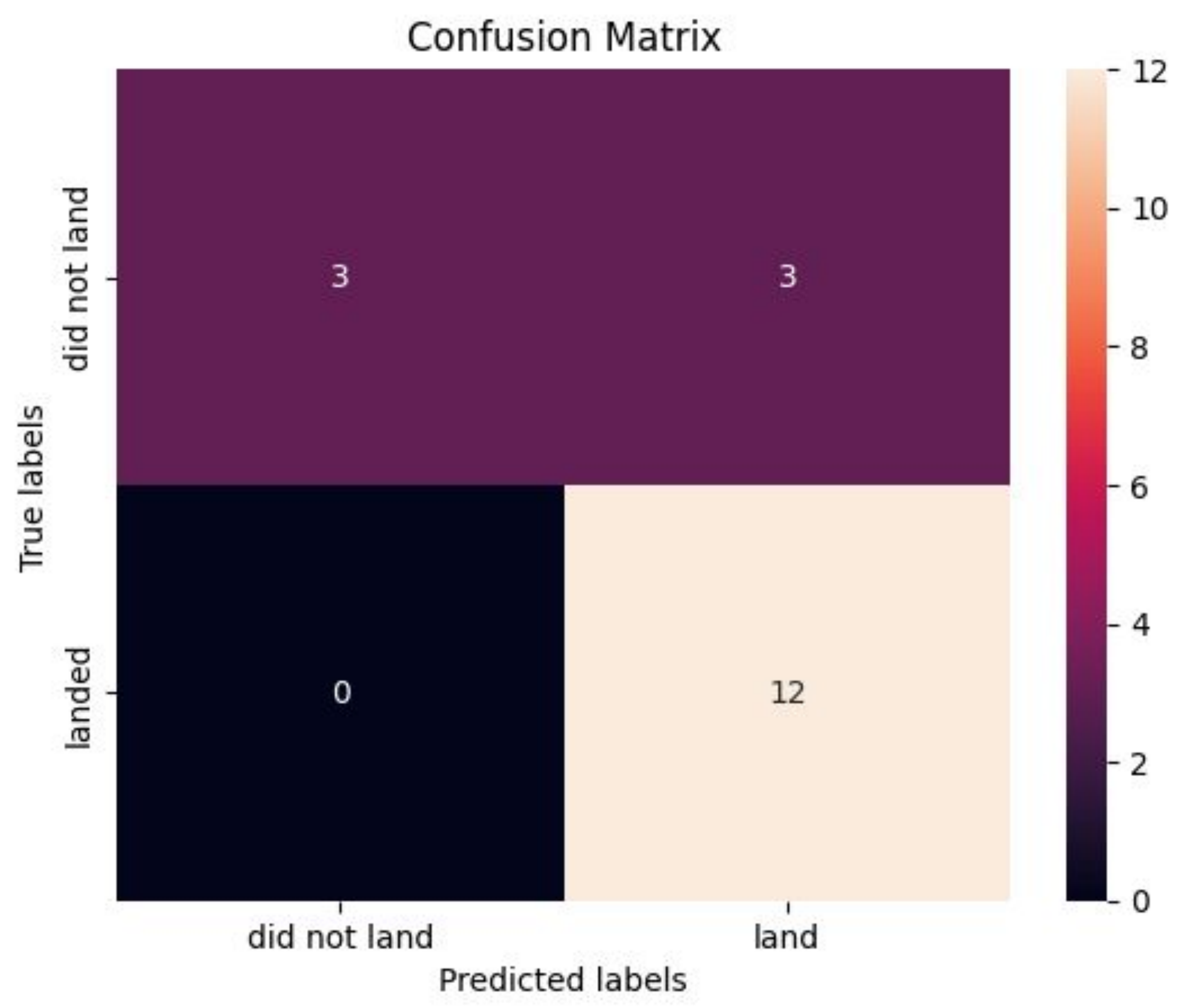


Figure: Confusion Matrix - KNN

Model Accuracy Comparison & Recommendation

Key Insights

- ➢ **All three models achieved identical accuracy (0.8333).**
- ➢ **This indicates:**
  - ✓ No single "best-performing" model based on accuracy alone
  - ✓ The dataset is relatively straightforward, enabling multiple algorithms to perform similarly
  - ✓ Additional metrics (precision, recall, F1-score) or cross-validation could differentiate them

Test Accuracy Results :

Model	Test Accuracy
Logistic Regression	0.8333
SVM	0.8333
KNN	0.8333
Decision Tree	0.7778

Model Selection

Even though performance is identical, Logistic Regression may be preferred due to:

- ➢ Simpler and more interpretable model
- ➢ Faster training and lower computational cost
- ➢ Works well with the dataset's feature structure

Summary of Machine Learning Results

What We Achieved

- Built 4 machine learning models using cross-validated grid search.
- Compared model performance across validation and testing phases.
- Identified optimal hyperparameters for each model.
- Selected the final model (Logistic Regression).

High-Level Findings

- Landing success is predictable with ~83% accuracy.
- Strong performance achieved using limited features.
- Logistic Regression provides a robust baseline for future improvement.

Plotly Dash SpaceX dashboard

SpaceX Launch Dashboard

Key Points

- ➢ Interactive dashboard built using Plotly Dash and Python
- ➢ Visualizes SpaceX launch data: success/failure and payload correlation
- ➢ Allows dynamic filtering by launch site and payload range

Start

↓

Import cleaned dataset into Python/Dash

↓

Define dashboard layout (graphs, dropdowns, sliders, UI components)

↓

Create callback functions to link UI inputs with graph outputs

↓

Generate visualizations:

- Pie chart → Launch success by site
- Scatter plot → Payload vs. Success

↓

Run Dash server → Render dashboard in browser

↓

User interacts with filters to explore data dynamically

↓

End

Chart: Plotly Dashboard Workflow

Total Successful Launches by Site

Key Insights

- Pie chart shows the proportion of successful launches across all sites
- Users can select specific launch sites to view success vs failure
- **Launch sites with higher success rates can be quickly identified**

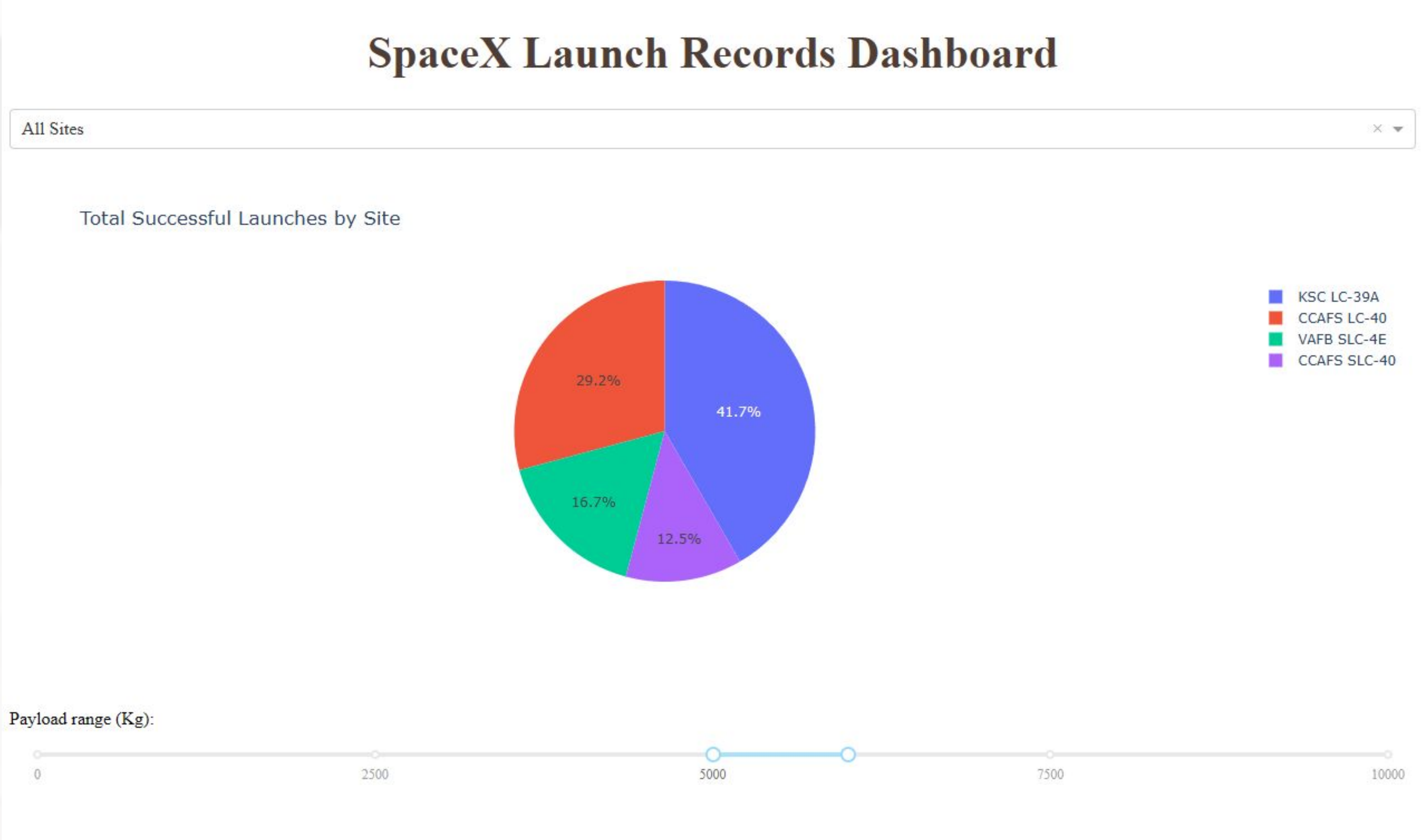


Figure: SpaceX Launch Dashboard

Correlation: Payload Mass & Launch Outcome

Key Insights

- Scatter plot shows correlation between payload mass and launch success
- Color-coded by Booster Version Category
- Filter by launch site or payload range for interactive exploration
- **Helps analyze trends of payload vs success across boosters**

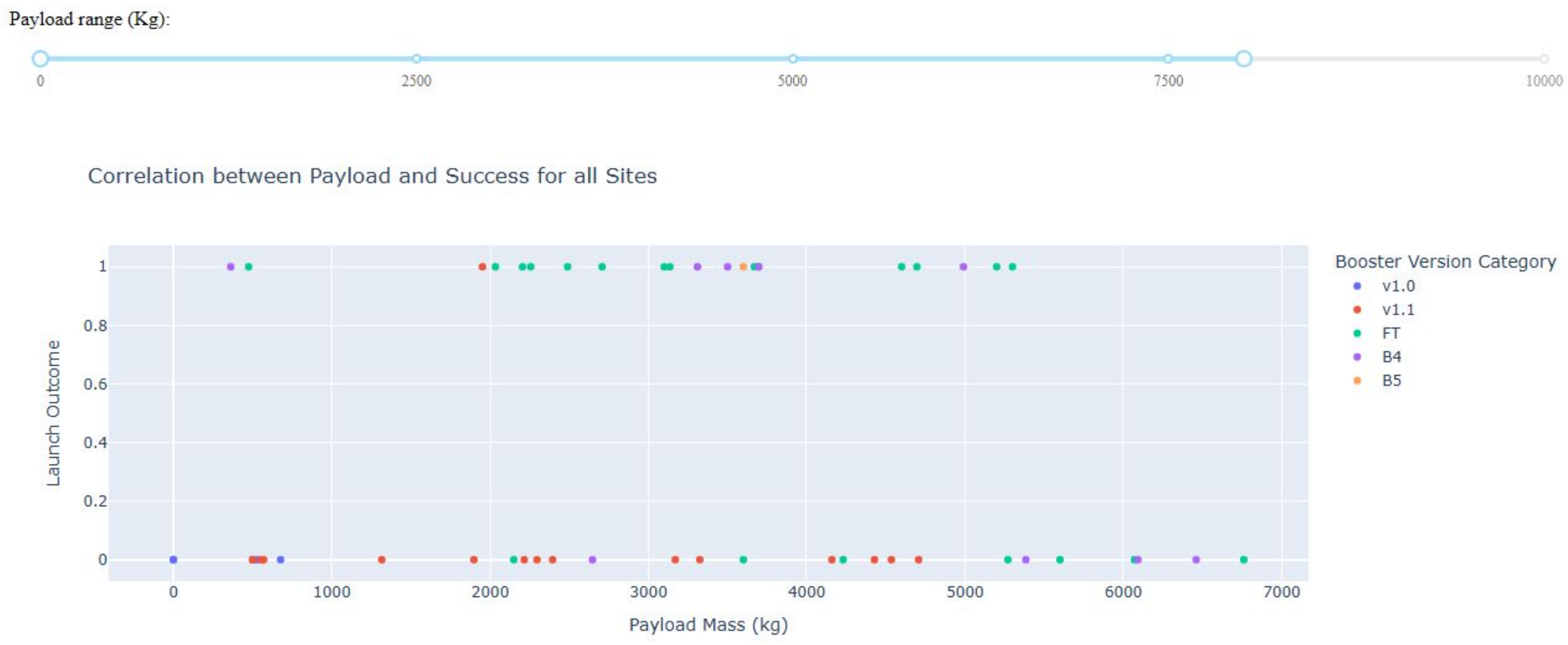


Figure: SpaceX Launch Dashboard

Dashboard Insights

What We Learned

- Interactive visualizations simplify data analysis for SpaceX launches
- Quick identification of site performance, payload trends, and booster impac
- Useful for decision-making and further predictive modeling

Conclusion

Final Conclusion

Project Summary

This project successfully explored SpaceX Falcon 9 launch data through data wrangling, SQL-based EDA, visual analytics, mapping, and predictive modeling to understand the factors influencing landing success.

Key Findings

- Launch success strongly correlates with Launch Site and Payload Mass.
- Most successful landings occur at LZ-1 and OCISLY.
- Geographic visualizations show distinct launch patterns along the Florida coastline.
- SQL analysis confirmed clear behavioral differences between launch sites and landing outcomes.
- Logistic Regression, SVM, and KNN achieved the highest test accuracy (~0.83). **Logistic Regression** was selected for simplicity, interpretability, and faster computation.
- Decision Tree performed slightly lower (77.8% accuracy) but provided insights into feature importance.
- Built an interactive SpaceX dashboard using Plotly Dash:
  - Pie chart shows launch success by site
  - Scatter plot shows correlation between payload and launch success

The combined approach demonstrates the power of data-driven predictions and interactive visualizations for analysis and decision-making.

Takeaway

Falcon 9 landing success can be reasonably predicted using available mission features, and data-driven insights can support mission planning and risk assessment.

Overall Project Insights & Future Improvement

What We Learned

- End-to-end data pipeline:

  Data preprocessing → model training → evaluation → interactive visualization

- SQL and Python together provide a deeper understanding of launch behavior.
- Visual analytics (Folium, Plotly Dash) reveal patterns impossible to see in tables alone.
- Even with balanced performance, models show that the dataset is predictable with moderate accuracy.

Opportunities for Further Study

- Include weather data to improve prediction accuracy.
- Apply hyperparameter tuning or ensemble methods (Random Forest, XGBoost).
- Use time-series analysis for mission scheduling optimization.

Appendix & References

Acknowledgement

This work is completed by Umme Sanjeda as part of the IBM Applied Data Science Capstone (Coursera) submission.

Figures

All plots and flowcharts in this presentation were generated from the project data.

Project Repository

SpaceX Launch Data Analysis & Prediction Capstone

Contains all Jupyter notebooks, Python files, datasets, and supporting scripts.

Notes

Any calculations, preprocessing steps, and model hyperparameters are documented in the notebooks.

Reproducible results: Run the notebooks in order to reproduce EDA, content-based, and collaborative filtering results.

Thank You!

For questions or discussion, please reach out -

GitHub Repository: [SpaceX Launch Data Analysis & Prediction Capstone]

Author
Umme Sanjeda

November 22, 2025

Summary
This presentation showcased a complete workflow from data collection, exploratory analysis, interactive visualizations, to predictive modeling of SpaceX launch outcomes.

Additional Info - This course is a part of IBM Data Science Professional Certificate