IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rooshikesh Bhatt
December, 2025

Outline

- **Executive Summary**
- **Introduction** (Business Problem & Objectives)
- **Methodology**
  - Data Collection (API + Web Scraping)
  - Data Wrangling & Feature Engineering
  - Exploratory Data Analysis (EDA) + SQL
  - Interactive Visual Analytics (Folium + Dash)
  - Predictive Modeling (Classification + Tuning)
- **Results**
  - Key EDA Findings
  - Launch Site Geography Insights
  - Dashboard Insights
  - Best Model Performance + Confusion Matrix
- **Conclusion**
- **Appendix** (Extra Charts / Tables / Code References)

Executive Summary

Summary of methodologies

- Data collection via API
- Data collection with Web Scraping
- Data wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Plotly Dash
- Machine Learning Prediction

Summary of all results

- **Success Rate over Time:** An improvement in landing success rate was observed over time (especially in later years).
- **Success Rate by Launch Site: KSC LC-39A** shows the highest landing success rate among the launch sites in this dataset.
- **Success Rate by Orbit: ES-L1, GEO, HEO, and SSO** have the highest observed success rates (≈100%) in this dataset (while **SO** shows the lowest observed success rate).
- **Payload:** Heavier payload missions show more failures earlier in the program, with a noticeable improvement in outcomes over time.
- **Predictive Analysis: Logistic Regression, SVM, and KNN tied for the best test accuracy (83.33%)** after GridSearchCV; **Decision Tree** achieved **77.78%**.

Introduction

Project background and context

- SpaceX reduces launch cost by **reusing Falcon 9 first-stage boosters**.
- The **key cost driver** is whether the first stage **lands successfully** (enabling recovery/reuse).
- Goal: use historical launch data to **analyze drivers of landing success** and **build a predictive model**.

Problems / questions we want to answer

- How has the **landing success rate changed over time**?
- Which **launch sites** show higher success rates?
- Which **orbits** are associated with higher/lower success rates?
- How does **payload mass** relate to landing success?
- Can we build a model to **predict landing outcome (Class: land vs not land)**, and which algorithm performs best?

Section 1

Methodology

Methodology

Executive Summary

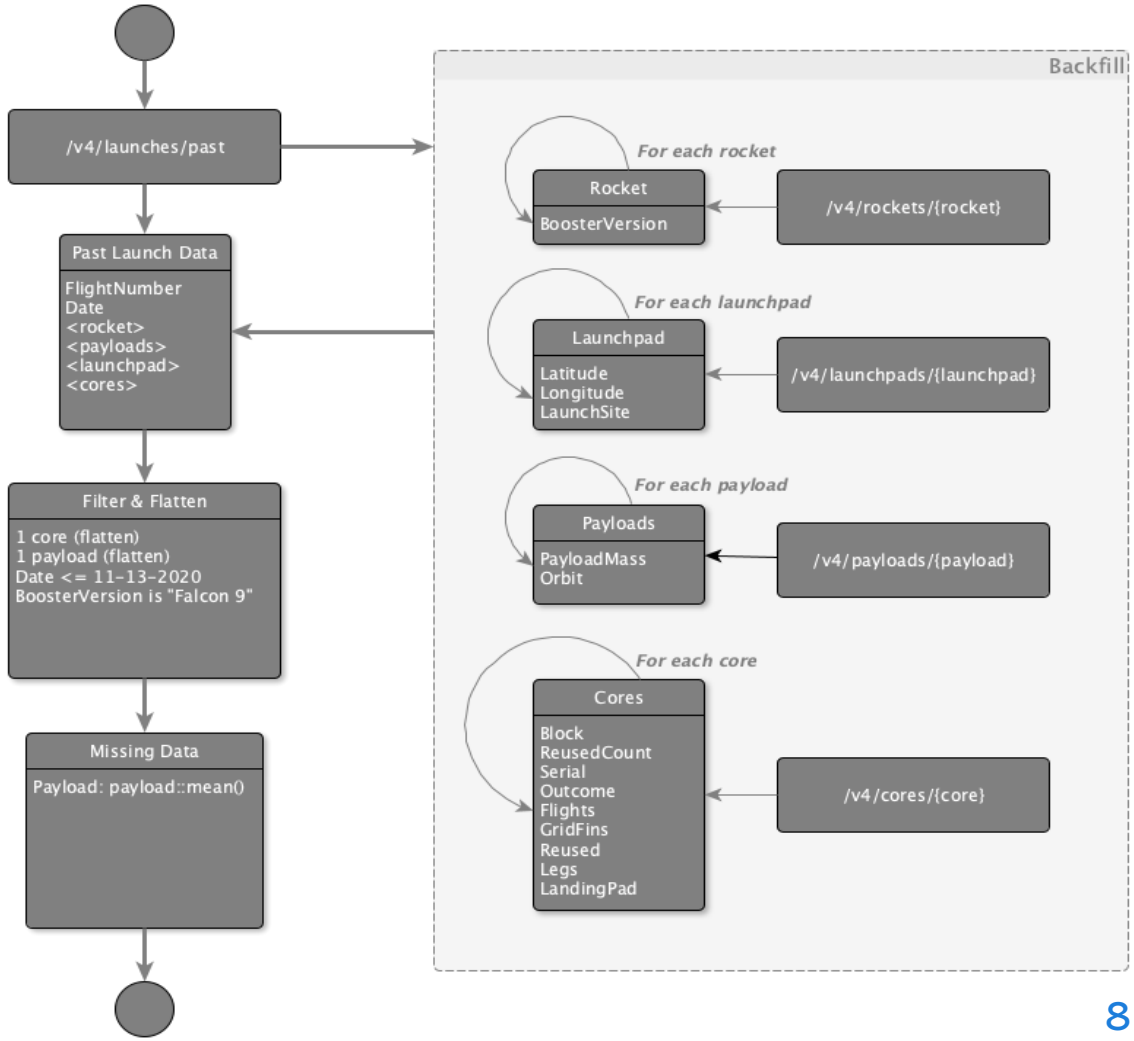
- Data collection methodology
  - Retrieval and consolidation from multiple SpaceX API endpoints
  - Web scaping tabular data from Wikipedia
- Perform data wrangling
  - Extracted relevant records
  - Flattened fields and resolved missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Visualize variable relationships
  - Look at the data in aggregate
- Perform interactive visual analytics using Folium and Plotly Dash
  - Mark all launch sites on a map
  - Mark successful and failed launches
  - Calculate distances to proximate locations
  - Provide for interactive exploration of the data
- Perform predictive analysis using classification models
  - Build, evaluate, and compare several predictive classification models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

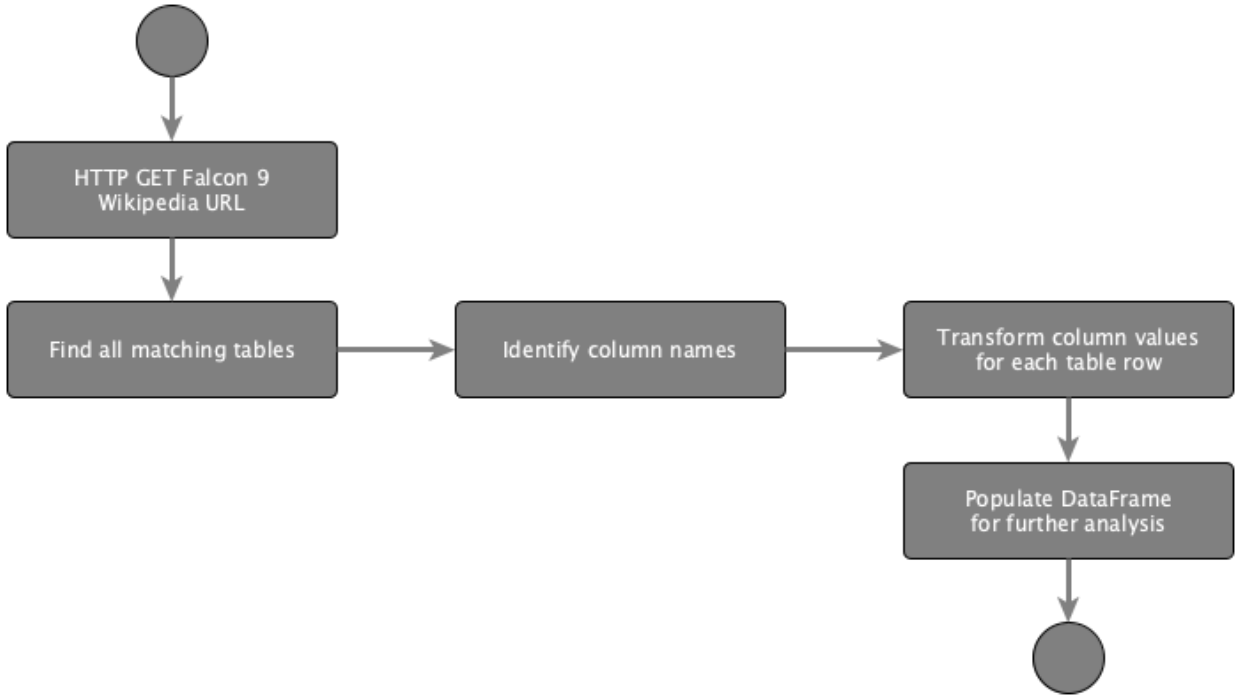
- The initial dataset was retrieved from the /v4/launches/past API endpoint.
- For records with valid linked IDs, missing details were supplemented using the rocket, launchpad, payloads, and cores API endpoints.
- https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

Web scraping workflow (Wikipedia)

- Send an **HTTP GET** request to the **Falcon 9 launches** Wikipedia page to retrieve the HTML content.
- Parse the page with **BeautifulSoup** and extract the launch tables.
- Clean/standardize the extracted fields and load them into a **Pandas DataFrame** for downstream analysis.
- https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/jupyter-labs-webscraping.ipynb



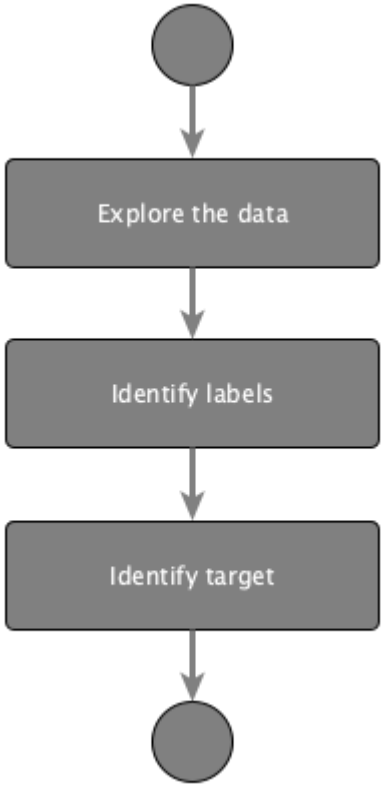
Data Wrangling

Goals

- Apply Exploratory Data Analysis (EDA) to uncover trends and relationships within the dataset.
- Define appropriate labels required for training supervised machine learning models.

Steps

- Calculate the percentage of missing values for each feature to assess data completeness.
- Classify dataset columns as numerical or categorical variables.
- Analyze and visualize the number of launches conducted at each launch site.
- Examine the distribution of different orbit types across the dataset.
- Investigate launch outcomes and categorize them into binary classes (success or failure).
- Create a binary target variable named **Class**, which is used as the prediction label for model training.

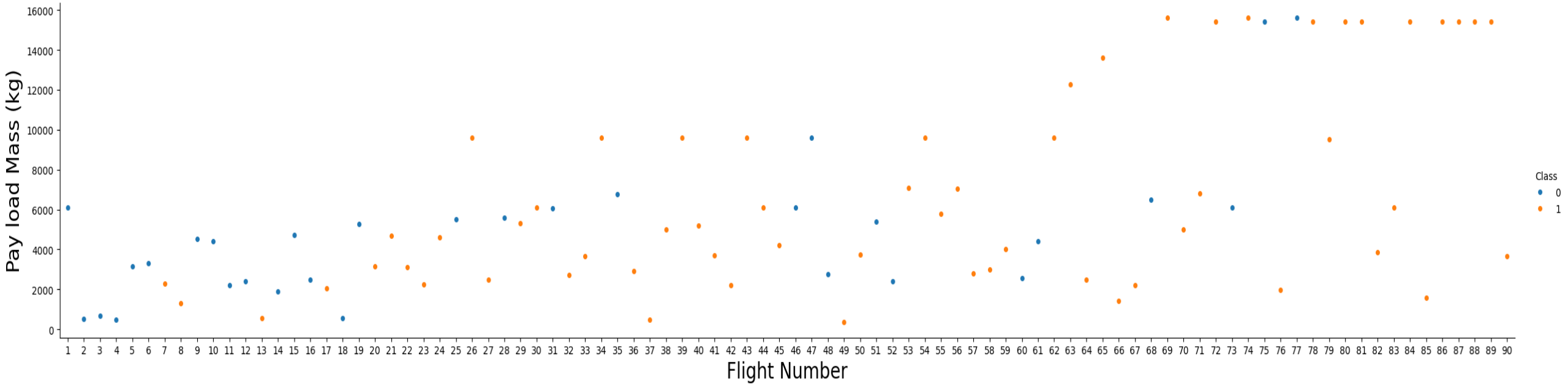


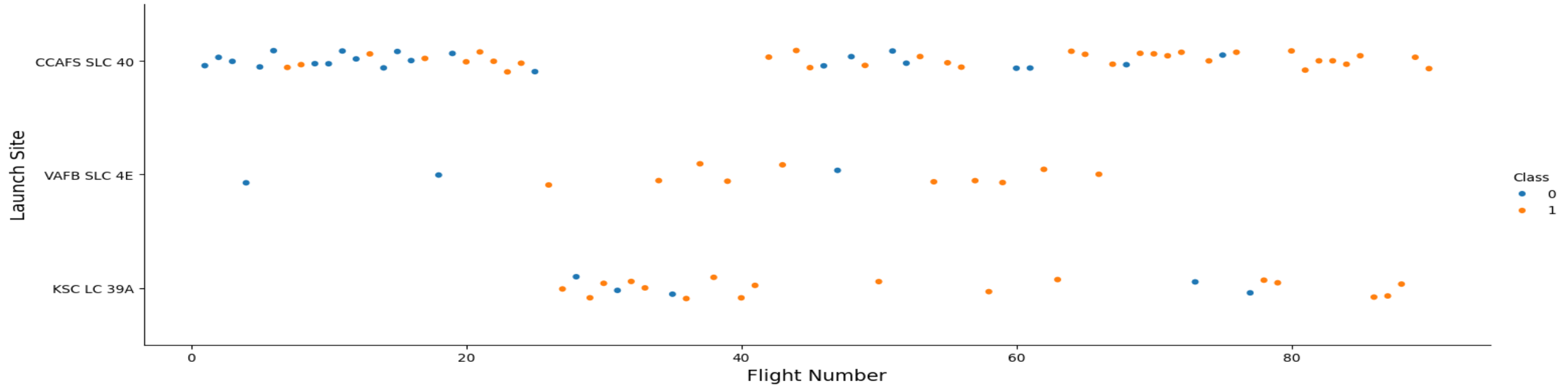
https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

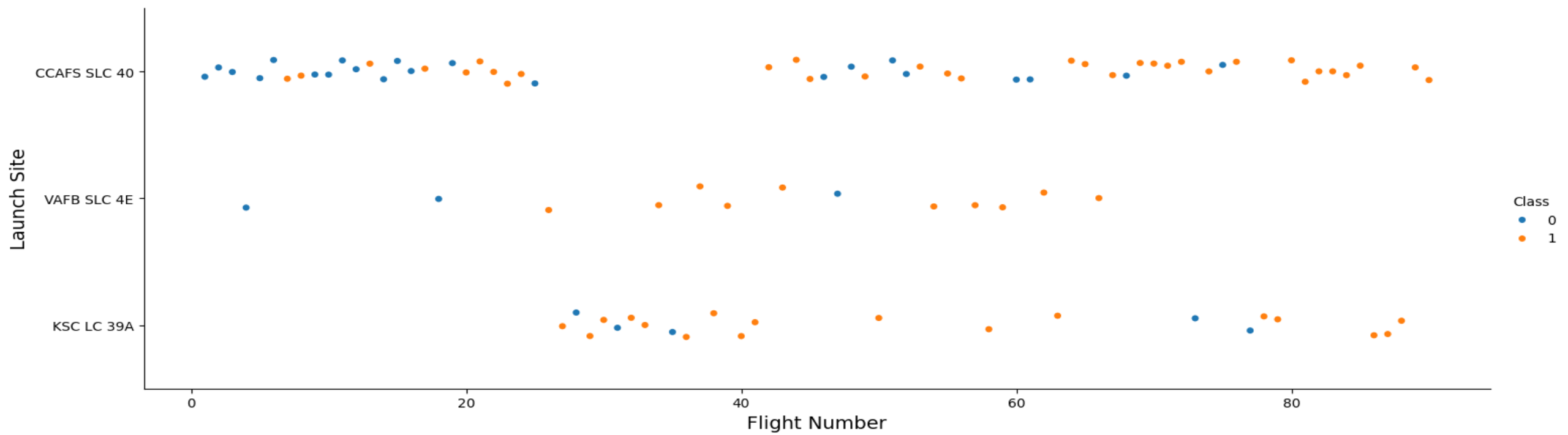
EDA with Data Visualization

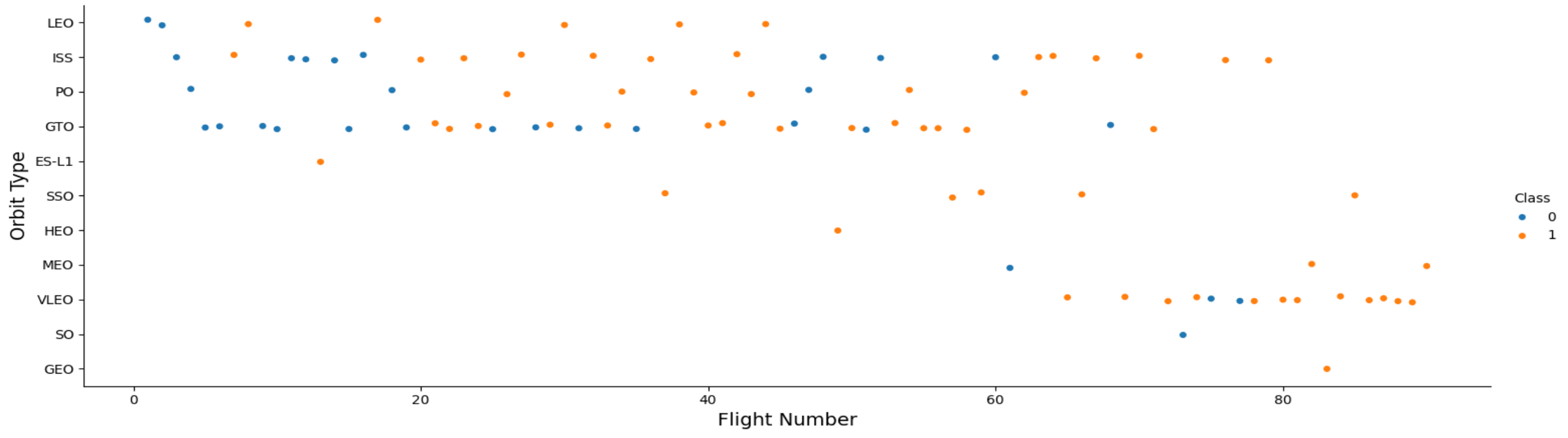
Visualization and Feature Preparation

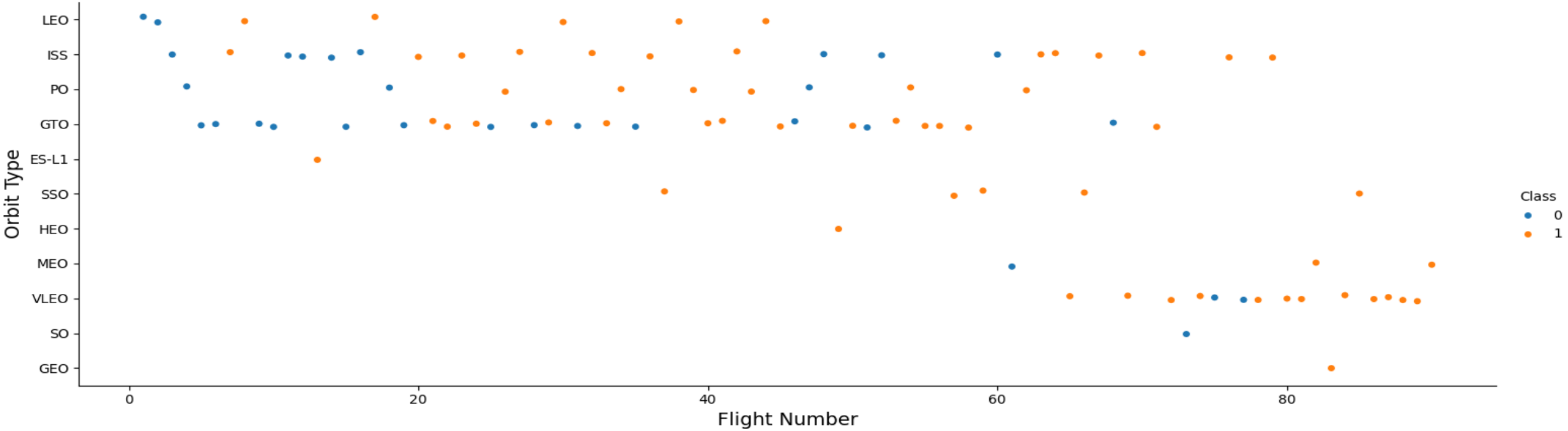
- Visualize key variable relationships to understand their influence on launch outcomes, including:
  - Flight number versus landing outcome
  - Flight number versus launch site
  - Payload mass versus launch site
  - Orbit type versus landing outcome
  - Flight number versus orbit type
  - Payload mass versus orbit type
  - Year-over-year launch success trends
- Transform categorical features into numerical format using one-hot (dummy) encoding.
- Ensure all numerical features are cast to float64 data type to support machine learning model training.
- https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/edadataviz.ipynb

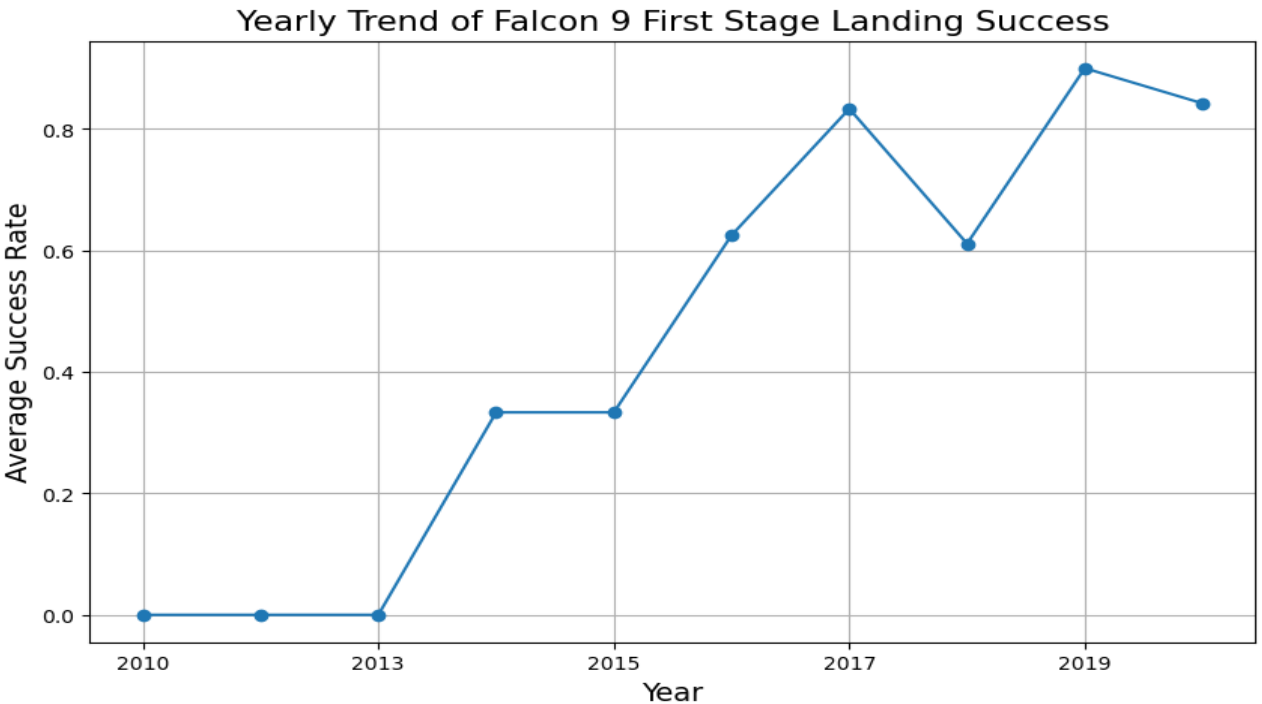

Pay load Mass (kg)
Flight Number
Class
0
1

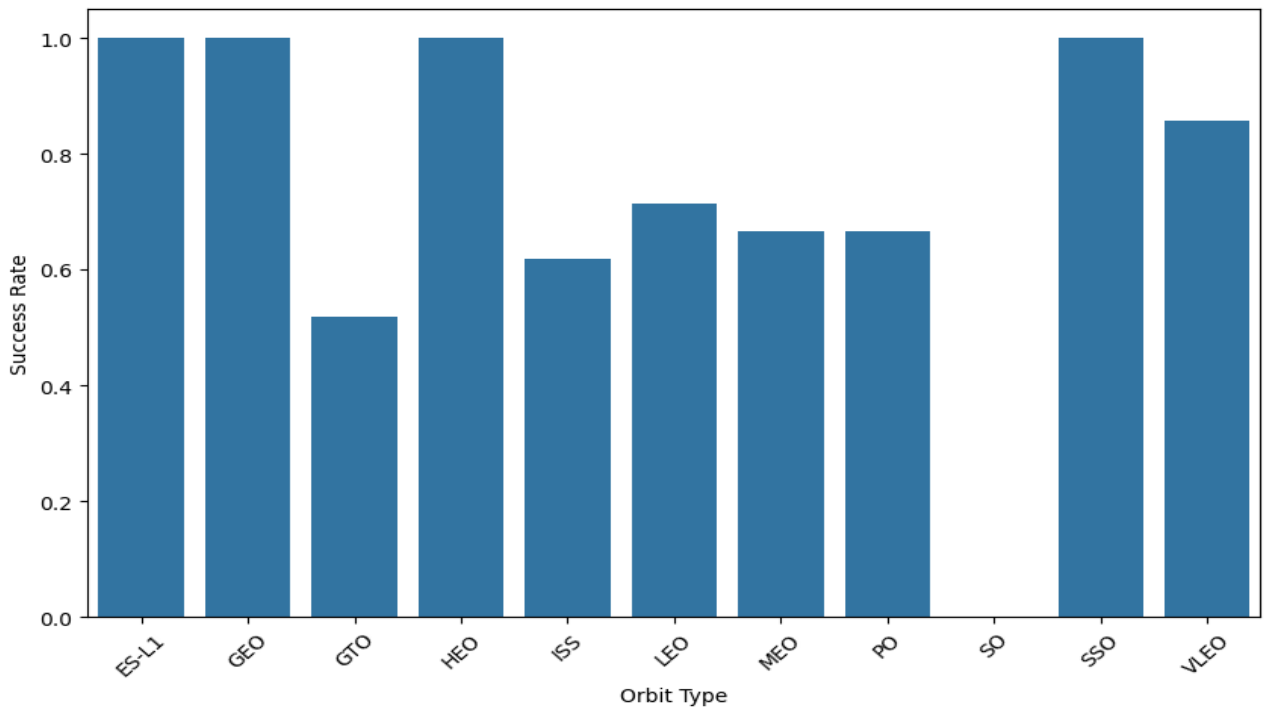

Launch Site
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Flight Number
Class
0
1


Launch Site
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
0
20
40
60
80
Flight Number
Class
0
1


Orbit Type
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
20
40
60
80
Flight Number
Class
0
1


Orbit Type
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
20
40
60
80
Flight Number
Class
0
1


Yearly Trend of Falcon 9 First Stage Landing Success
Average Success Rate
0.0
0.2
0.4
0.6
0.8
2010
2013
2015
2017
2019
Year


Success Rate
0.0
0.2
0.4
0.6
0.8
1.0
ES-L1
GEO
GTO
HEO
ISS
LEO
MEO
PO
SO
SSO
VLEO
Orbit Type

EDA with SQL

SQL Analysis Conducted

- Retrieved all distinct launch site names from the dataset.
- Extracted a sample of five missions where the launch site name begins with **“CCA”**.
- Calculated the total payload mass delivered by boosters used in **NASA (CRS)** missions.
- Computed the average payload mass carried by the **Falcon 9 v1.1** booster variant.
- Identified the date of the first successful **ground-based booster landing**.
- Determined which booster versions successfully landed on a **drone ship** while carrying payloads between **4,000 kg and 6,000 kg**.
- Counted the total number of **successful** and **failed** mission outcomes.
- Listed all booster versions that transported the **maximum payload mass** recorded in the dataset.
- Retrieved the **month**, **landing outcome**, **booster version**, and **launch site** for drone ship landing failures that occurred in **2015**.
- Analyzed the overall distribution of landing outcomes between **June 4, 2010** and **March 20, 2017**.
- https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

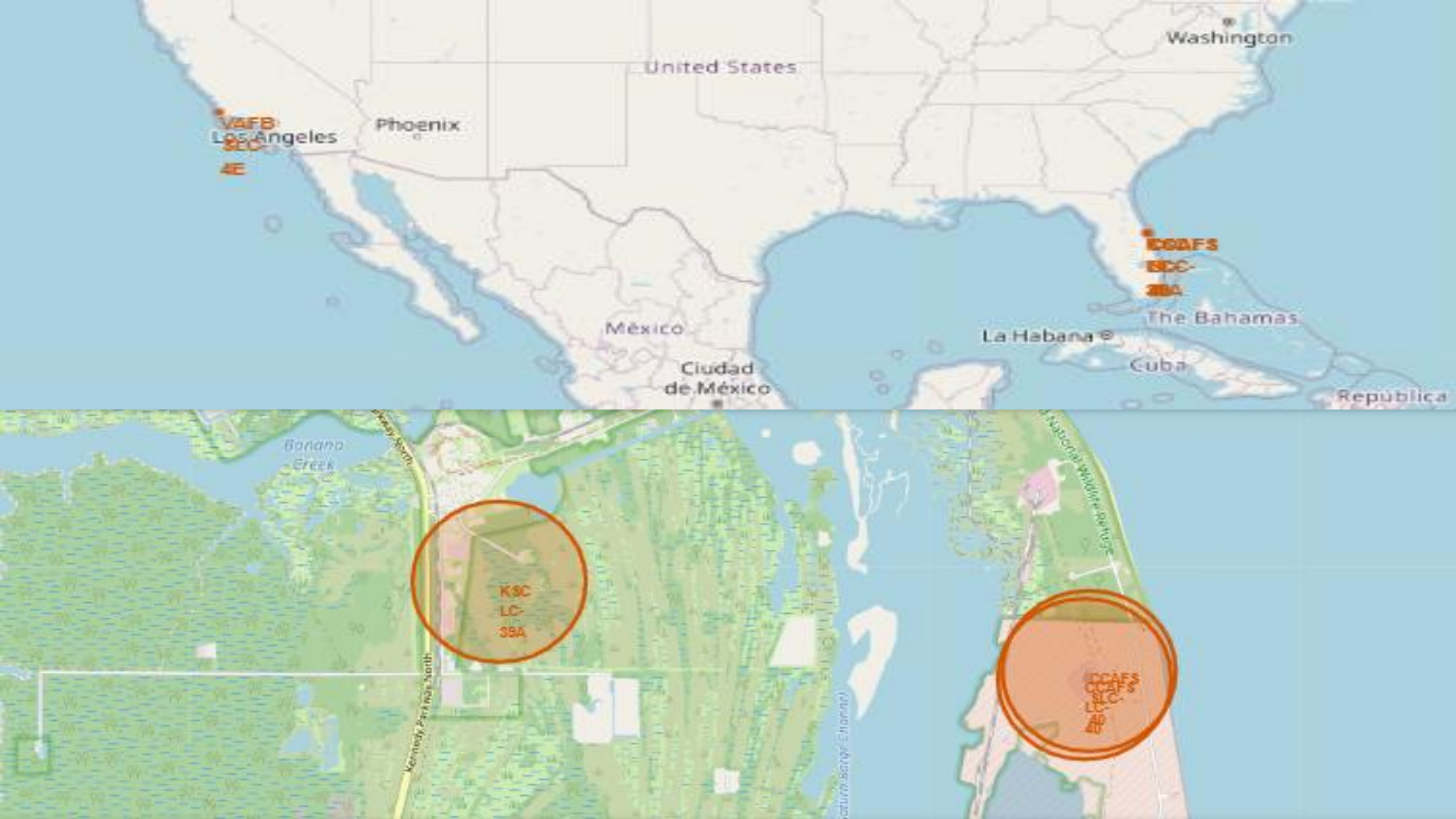
Geospatial Analysis Using Interactive Maps

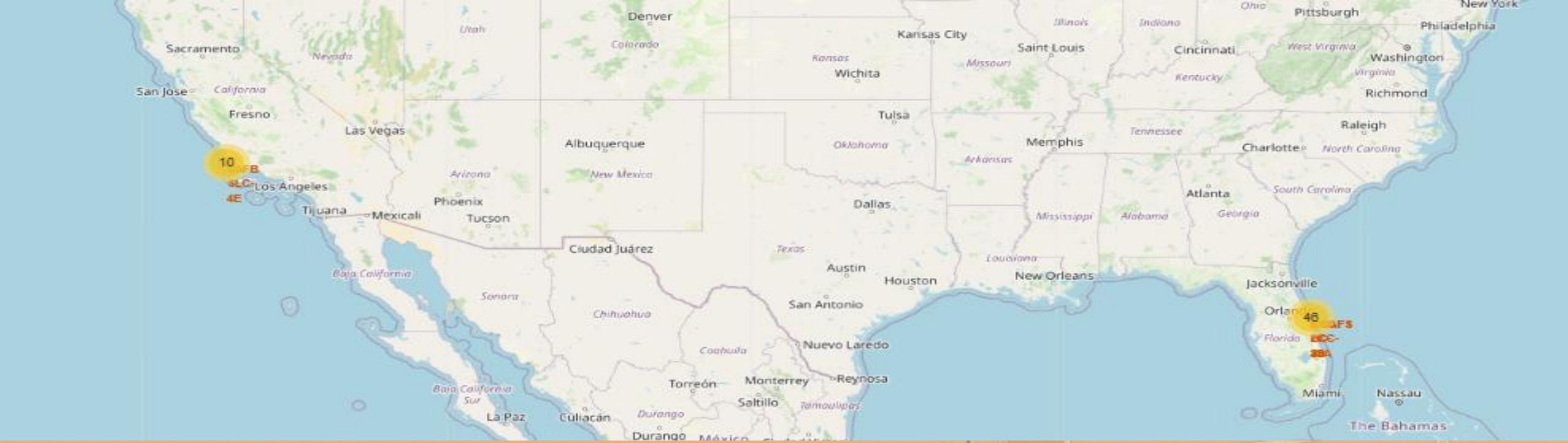
To identify spatial trends and geographic influences on launch outcomes, an interactive map was created to visualize key location-based information:

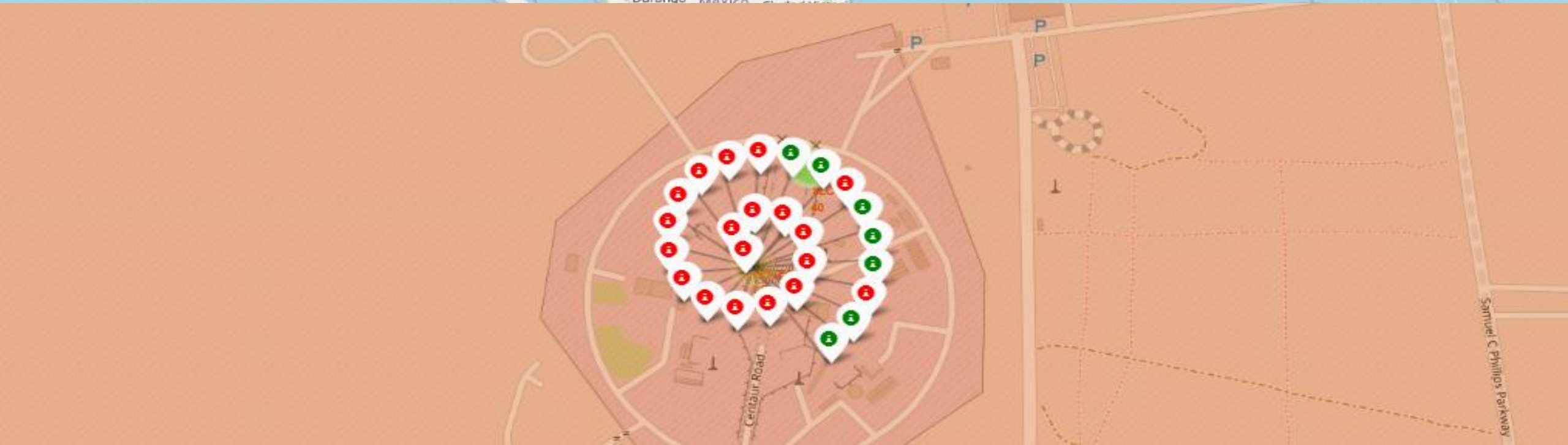
- Plotted the geographic locations of **all Falcon 9 launch sites**.
- Differentiated **successful and unsuccessful launches** using visual markers.
- Measured and displayed the **distances between each launch site and nearby geographic landmarks**, such as coastlines and infrastructure.

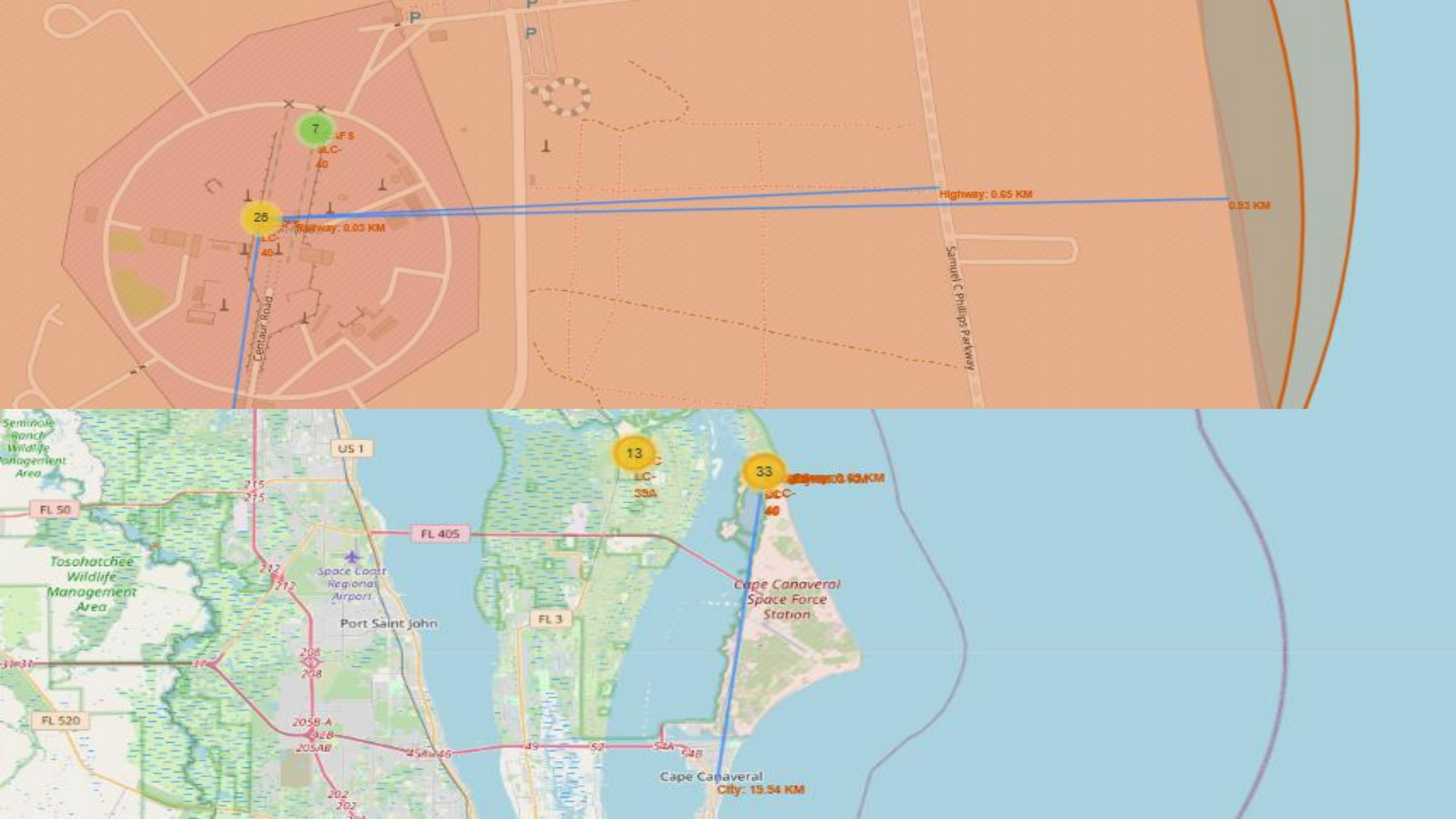
This spatial visualization helped reveal how launch site geography and surrounding features may influence mission outcomes.

https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/lab_jupyter_launch_site_location.ipynb


Washington
United States
VAFB
SLC-
4E
Los Angeles
Phoenix
CCAFS
SLC-
40
México
Ciudad
de México
La Habana
The Bahamas
Cuba
República
Banana
Creek
KSC
LC-
39A
Kennedy Parkway North
CCAFS
SLC-
LC-
40


Denver
Utah
Colorado
Kansas City
Illinois
Indiana
Ohio
Pittsburgh
New York
Philadelphia
Sacramento
Nevada
Saint Louis
Missouri
Cincinnati
West Virginia
Washington
Kansas
Wichita
Kentucky
Virginia
San Jose
California
Richmond
Fresno
Tulsa
Las Vegas
Tennessee
Raleigh
Albuquerque
Oklahoma
Memphis
Charlotte
North Carolina
Arkansas
10
Arizona
New Mexico
Los Angeles
Atlanta
South Carolina
Phoenix
Dallas
Tijuana
Mexicali
Tucson
Mississippi
Alabama
Georgia
Ciudad Juárez
Texas
Austin
Houston
New Orleans
Jacksonville
Baja California
Sonora
San Antonio
Chihuahua
Orlan
46
Florida
Coahuila
Nuevo Laredo
Torreón
Monterrey
Reynosa
Miami
Nassau
Baja California Sur
Saltillo
Tamaulipas
La Paz
Culiacán
Durango
The Bahamas


P
P
Centaur Road
Samuel C. Phillips Parkway


7
LC-40
26
LC-40
Highway: 0.03 KM
Highway: 0.65 KM
0.93 KM
Centaur Road
Samuel C Phillips Parkway
13
LC-39A
33
LCC-40
US 1
FL 50
FL 405
FL 520
FL 3
Seminole Ranch Wildlife Management Area
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
Port Saint John
Cape Canaveral Space Force Station
Cape Canaveral
City: 15.54 KM

Build a Dashboard with Plotly Dash

Interactive Dashboard with Plotly Dash

To support interactive and user-driven analysis, a Plotly Dash dashboard was implemented with the following components:

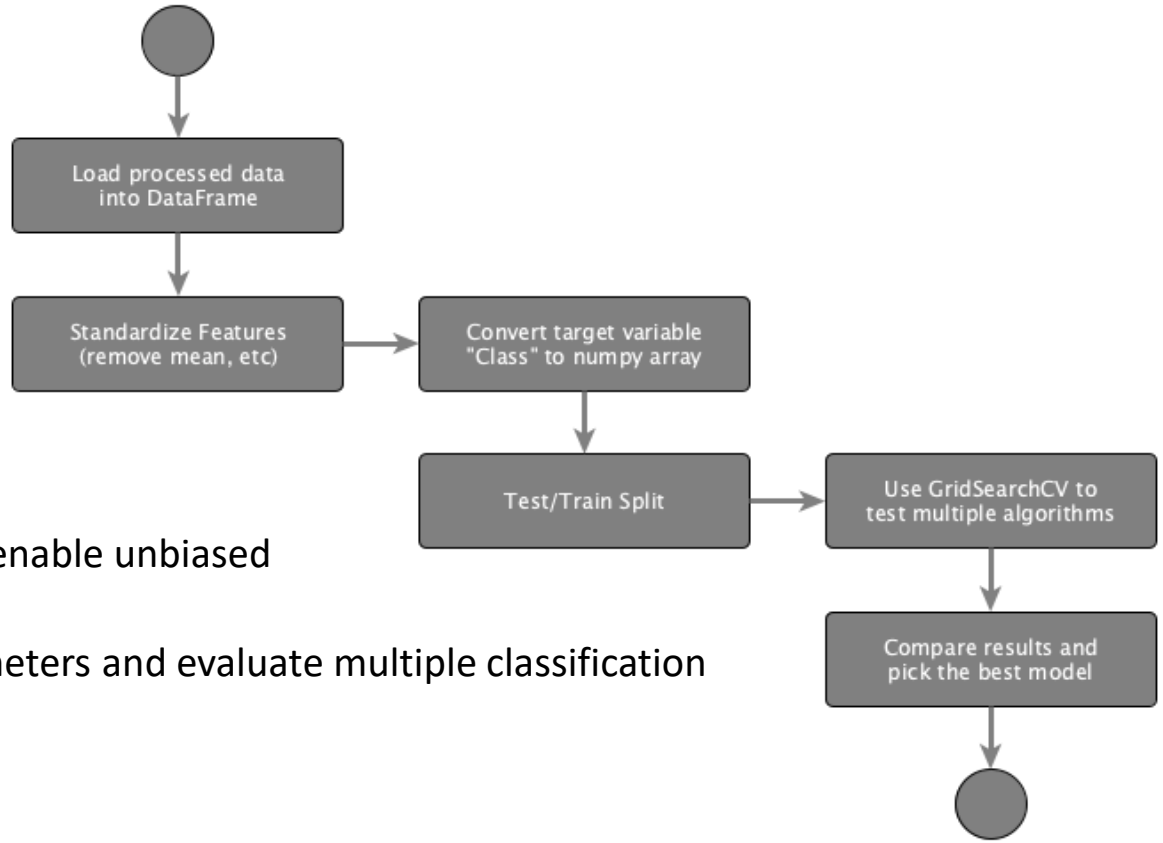
- A **launch site dropdown selector** that dynamically updates all visualizations.
- **Pie Chart Functionality:**
  - When *all launch sites* are selected, the chart displays the distribution of successful landings across all sites.
  - When a *specific launch site* is selected, the chart compares successful versus failed launches for that site.
- **Scatter Plot Functionality:**
  - With *all sites selected*, the scatter plot visualizes landing outcomes based on payload mass and booster version across all launches.
  - With a *single site selected*, the scatter plot filters results to show payload mass versus booster version for that specific location.
- A **payload mass range slider** that allows users to filter scatter plot data points based on payload weight.

This dashboard enables flexible exploration of how launch site and payload characteristics influence mission outcomes.

- https://github.com/rooshikeshbhatt/spacex-falcon9-capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

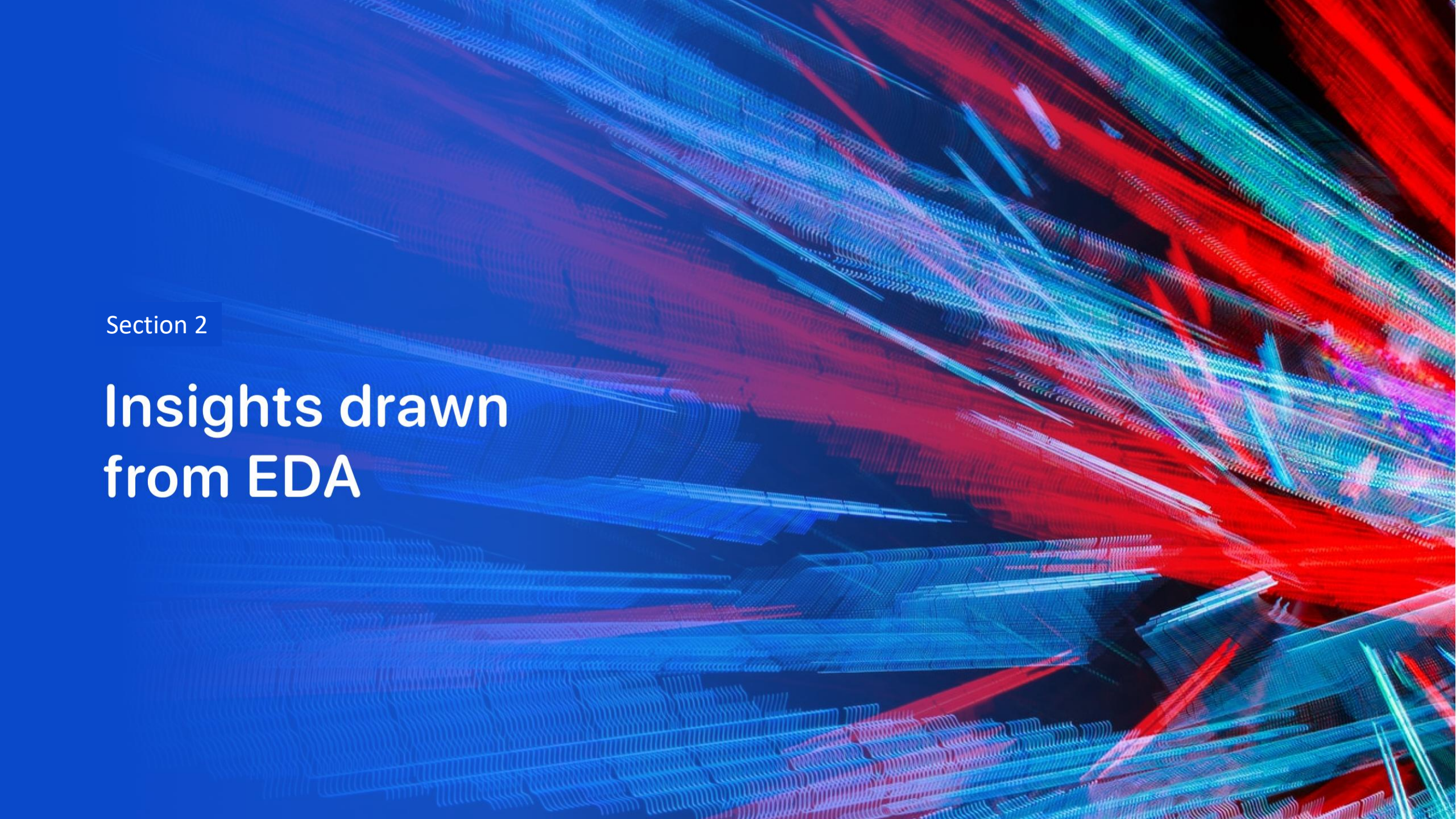
Predictive Analysis (Classification)
The classification workflow was carried out through the following steps:
•The prepared dataset was loaded for model development.
•Feature variables (**X**) were scaled using **StandardScaler** to ensure consistent feature ranges.
•The target variable (**Y**) was converted into a NumPy array for compatibility with machine learning models.
•The dataset was divided into **training and testing subsets** to enable unbiased performance evaluation.



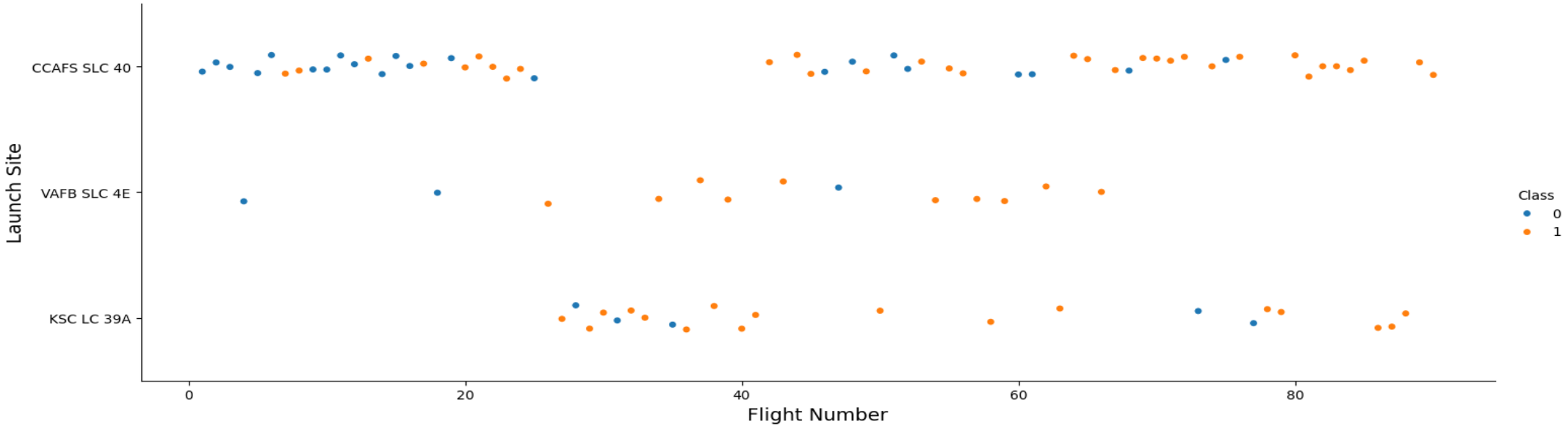
•**GridSearchCV** was applied to systematically tune hyperparameters and evaluate multiple classification algorithms, including:

- Logistic Regression
- Support Vector Classifier (SVC)
- Decision Tree Classifier
- K-Nearest Neighbors (KNN)

This approach ensured fair model comparison and selection of the best-performing classifier based on test accuracy.

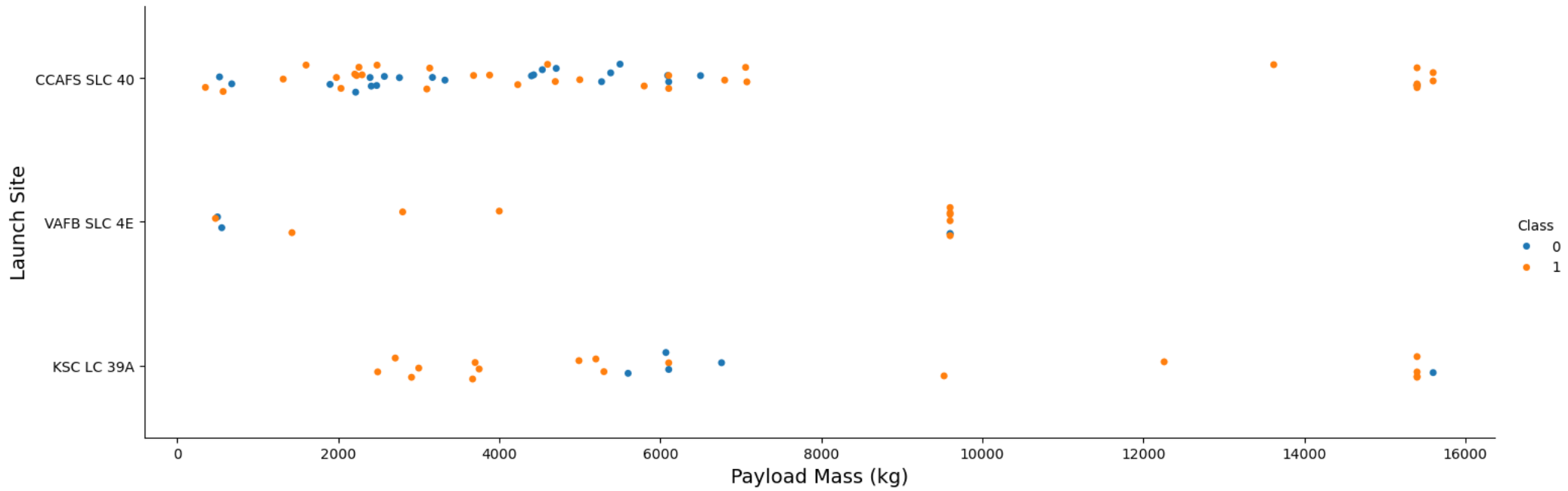

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



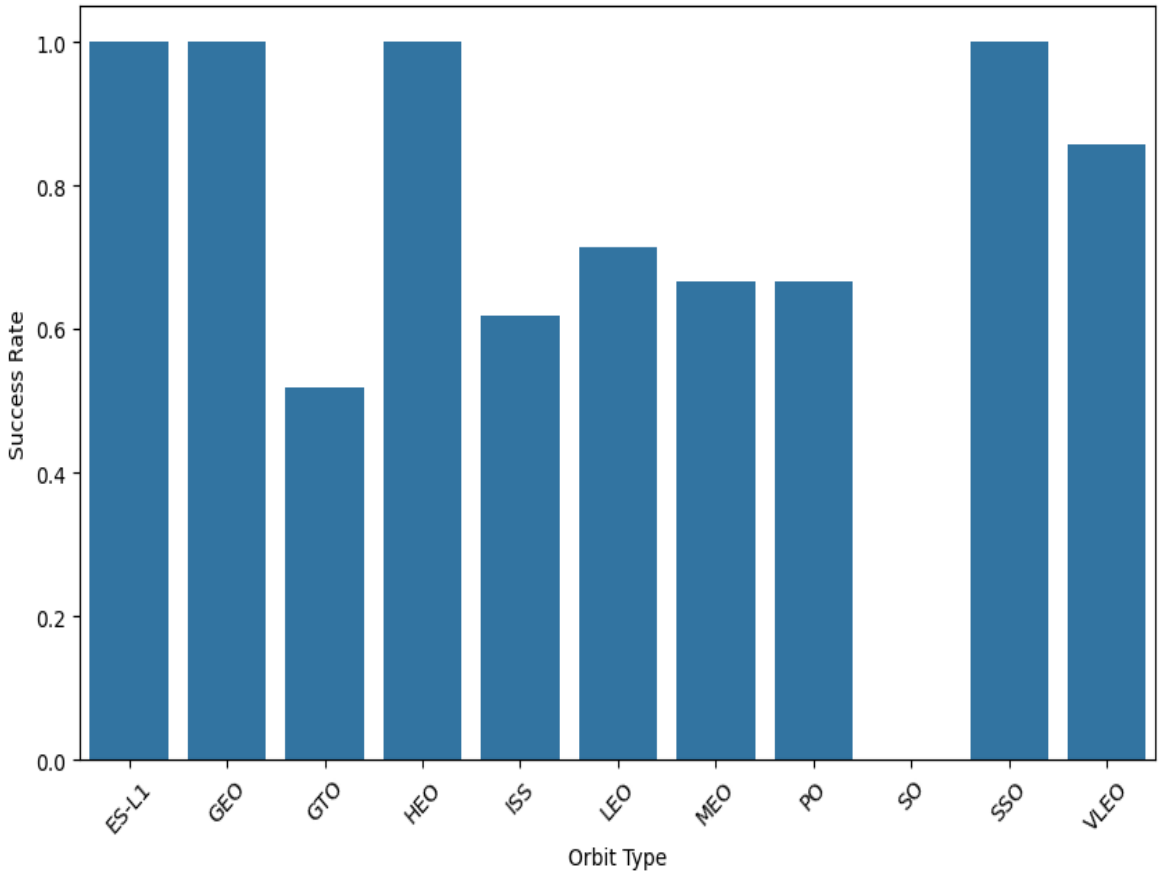
- All launch sites exhibit a combination of successful and unsuccessful first-stage landings, with the frequency of successful recoveries increasing in later missions.
- Initial launches were largely unsuccessful, reflecting early-stage technological and operational limitations that improved over time.
- Although **CCAFS SLC 40** conducted the highest number of launches overall, **VAFB SLC 4E** demonstrates a comparatively higher success rate in first-stage landing outcomes.

Payload vs. Launch Site



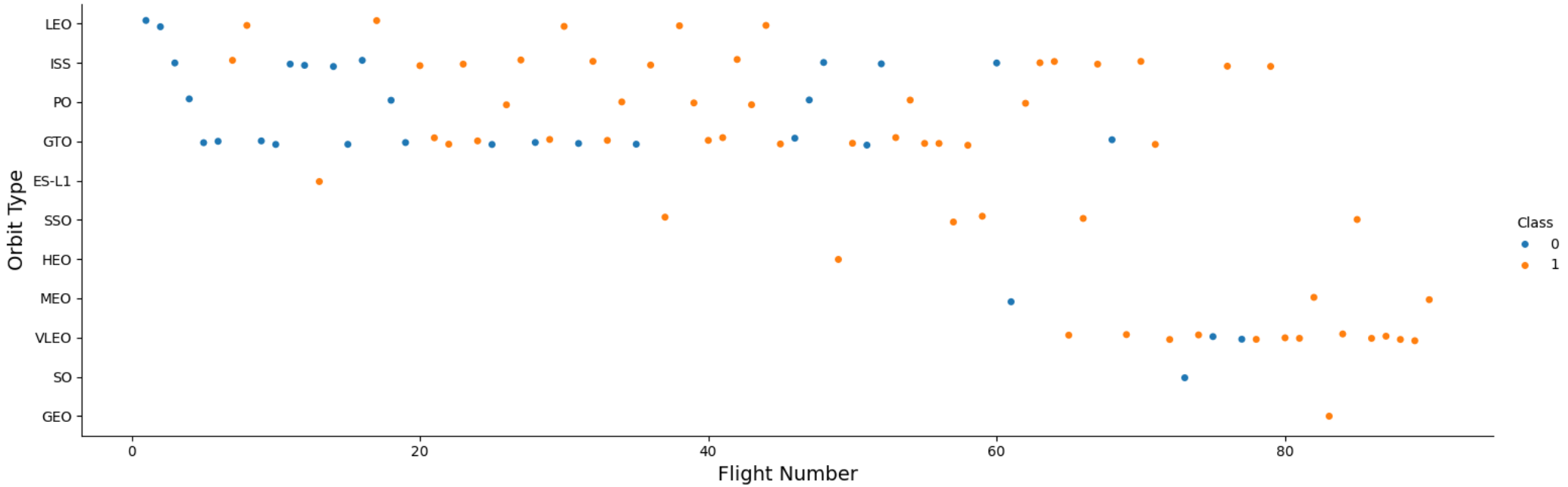
•Launch sites handle a broad spectrum of payload masses, ranging from lightweight to very heavy payloads.
•Earlier missions were primarily associated with lighter payloads and accounted for most first-stage landing failures.
•This pattern indicates that improvements in technology and operational procedures over time enabled more reliable landings, even for heavier payload missions.

Success Rate vs. Orbit Type



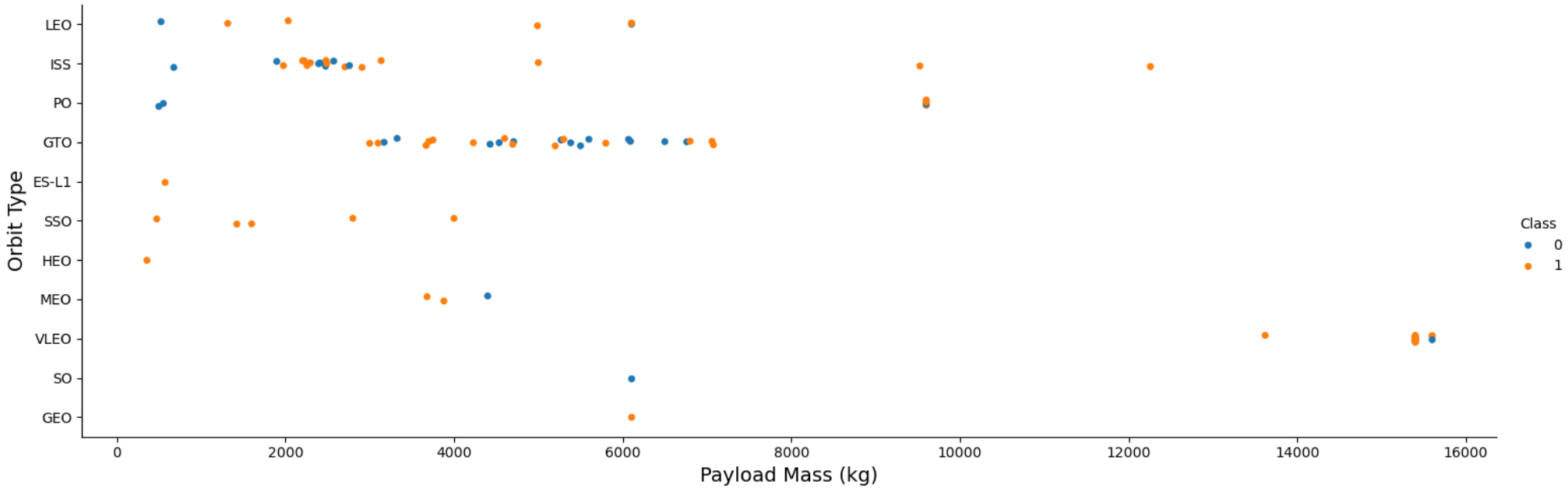
- Certain orbit types, including ES-L1, SSO, HEO, and GEO and demonstrate consistently high landing success rates.
- In contrast, GTO missions exhibit more variable outcomes, indicating that some orbit profiles pose greater operational or technological challenges for booster recovery.
- The SO orbit category contains only a single launch, making the available data insufficient for drawing reliable conclusions.

Flight Number vs. Orbit Type



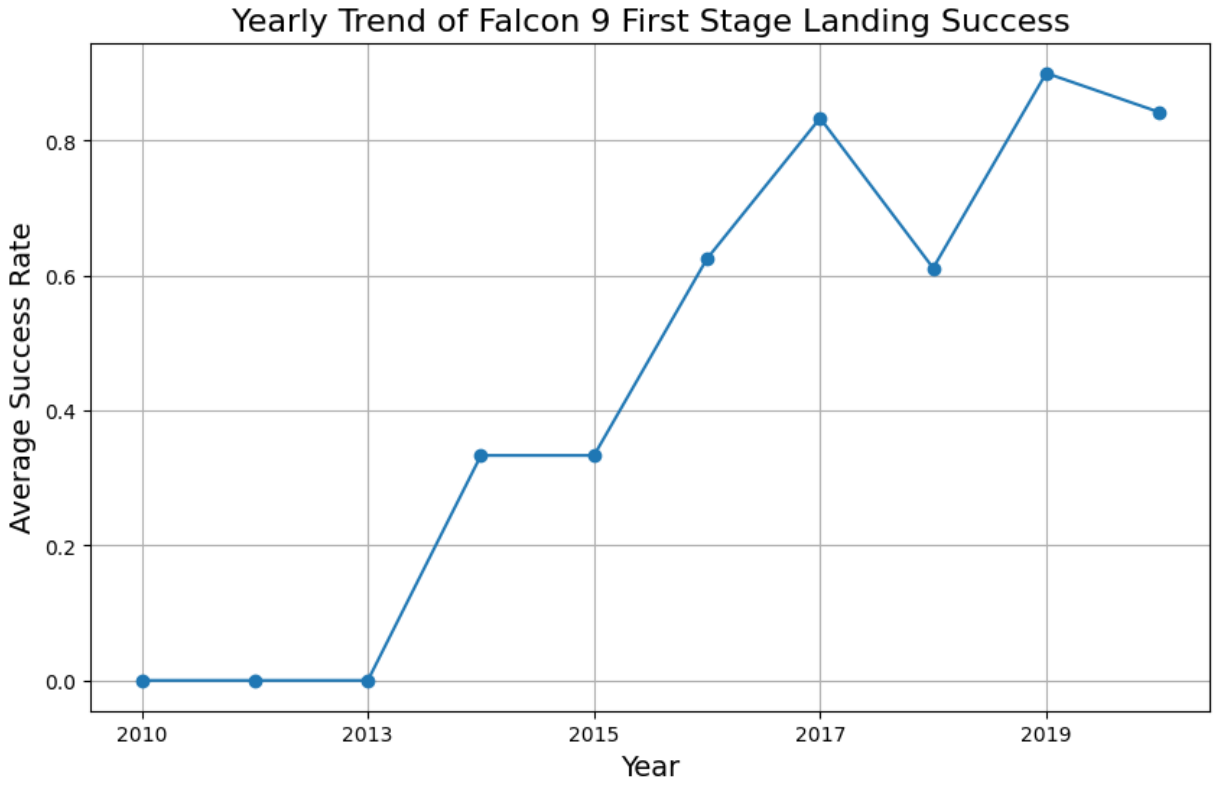
•Multiple orbit types appear across the full range of flight numbers, while certain orbits are only introduced in later missions.
•Landing success rates show a clear upward trend as flight numbers increase, reflecting growing operational experience and continuous technological refinement.

Payload vs. Orbit Type



- Payload mass varies widely across many orbit types, while orbits such as SSO, MEO, HEO, and GEO generally operate within a narrower payload range.
- Orbit categories with more constrained payload limits tend to exhibit higher first-stage landing success rates.
- Although payload mass alone does not directly dictate mission success, its interaction with orbit type indicates a meaningful relationship influencing landing outcomes.

Launch Success Yearly Trend



- The annual trend illustrates a steady transition from early operational difficulties to a high level of reliability in first-stage landings over time.
- Beginning in 2016, SpaceX demonstrated consistent year-over-year improvements in landing success rates, with a brief and limited decline observed in 2018.

All Launch Site Names

There are four unique Launch Sites

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

```
SELECT DISTINCT Launch_Site from SPACEXTABLE;
```

Launch Site Names Begin with 'CCA'

- First five records where launch sites begin with `CCA`

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

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

Total Payload Mass

The total payload carried by boosters from NASA (CRS) is **45,596kg**.

```
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD
  FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is **2,534.67kg**.

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS
  FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
```

First Successful Ground Landing Date

The first successful landing outcome on ground pad occurred on **December 22nd, 2015**.

```
SELECT MIN(Date) as LaunchDate
  FROM SPACEXTABLE
  WHERE Landing_Outcome = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

Booster	Payload Mass
F9 FT B1022	4,696kg
F9 FT B1026	4,600kg
F9 FT B1021.2	5,300kg
F9 FT B1031.2	5,200kg

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Mission Status	Count
Failure	1
Success	100

```
SELECT CASE
    WHEN Mission_Outcome LIKE 'Success%' THEN 'Success'
    WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure'
END as Mission_Status, COUNT(*)
FROM SPACEXTABLE
GROUP BY Mission_Status;
```

Boosters Carried Maximum Payload

- The maximum payload sent was **15,600kg**.
- The boosters that carried the maximum payload are:

```
SELECT
  DISTINCT Booster_Version,
  PAYLOAD_MASS__KG_
FROM SPACEXTABLE
  WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM
  SPACEXTABLE
  )
ORDER BY Booster_Version;
```

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

2015 Launch Records

- List of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

Month	Outcome	Booster	Launch Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
SELECT
    CASE strftime('%m', Date)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as Month,
    Landing_Outcome, Booster_Version, Launch_Site, Date
FROM SPACEXTABLE
WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

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

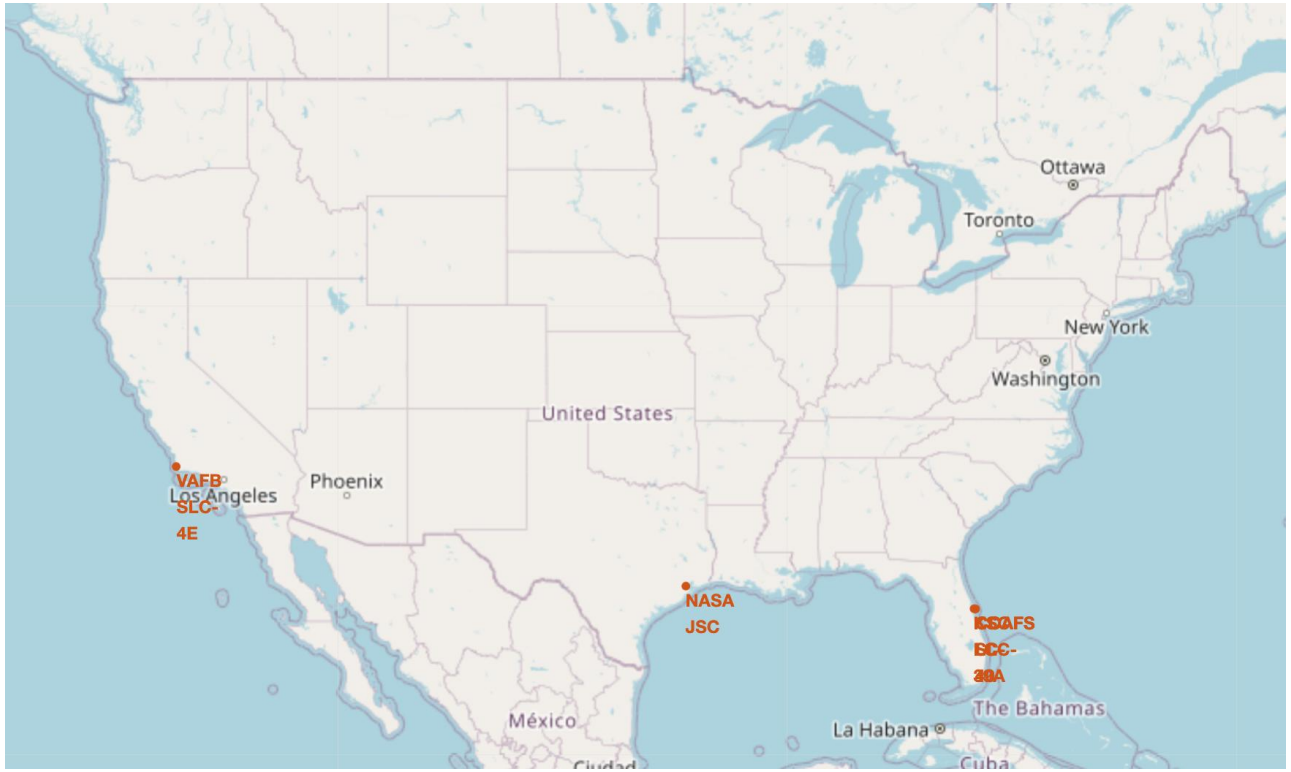
Landing Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

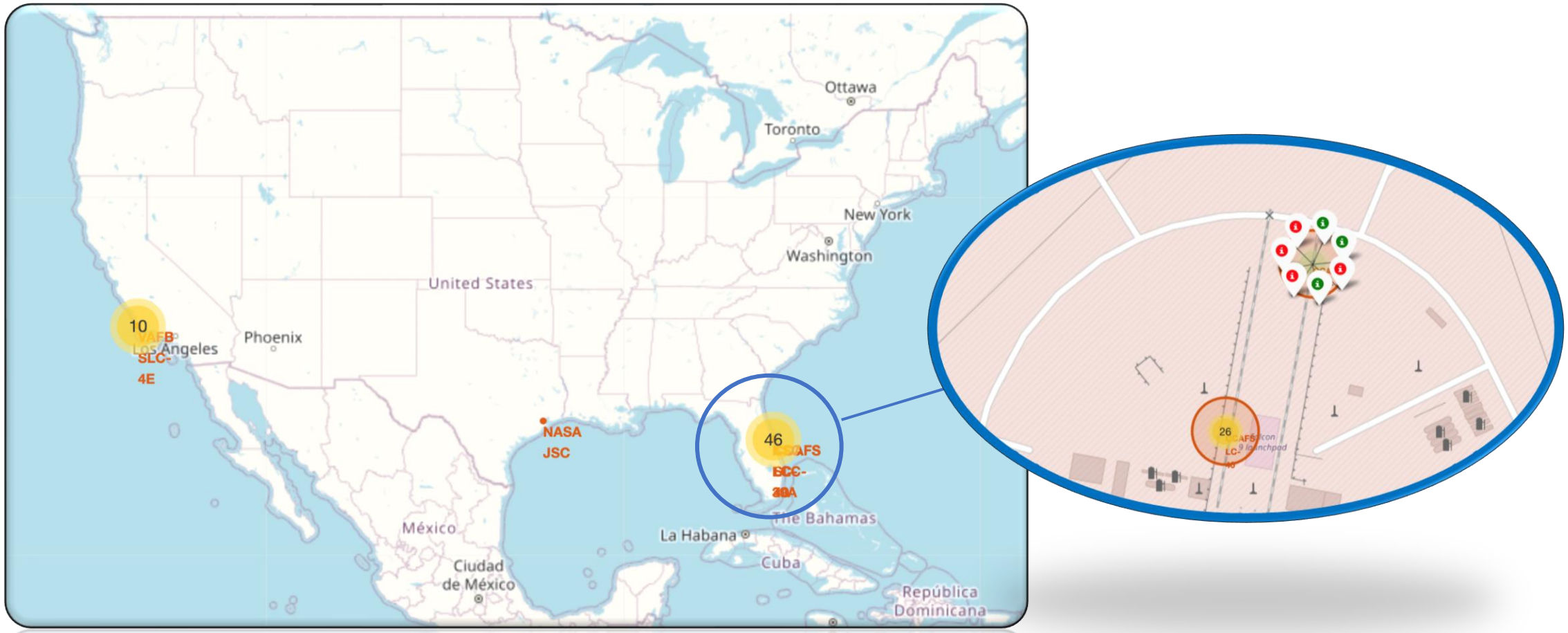
Launch Sites Proximities Analysis

Launch Site Locations

- Launch sites are located near coastal regions in Florida and California to reduce risk of catastrophic failures affecting human activities.

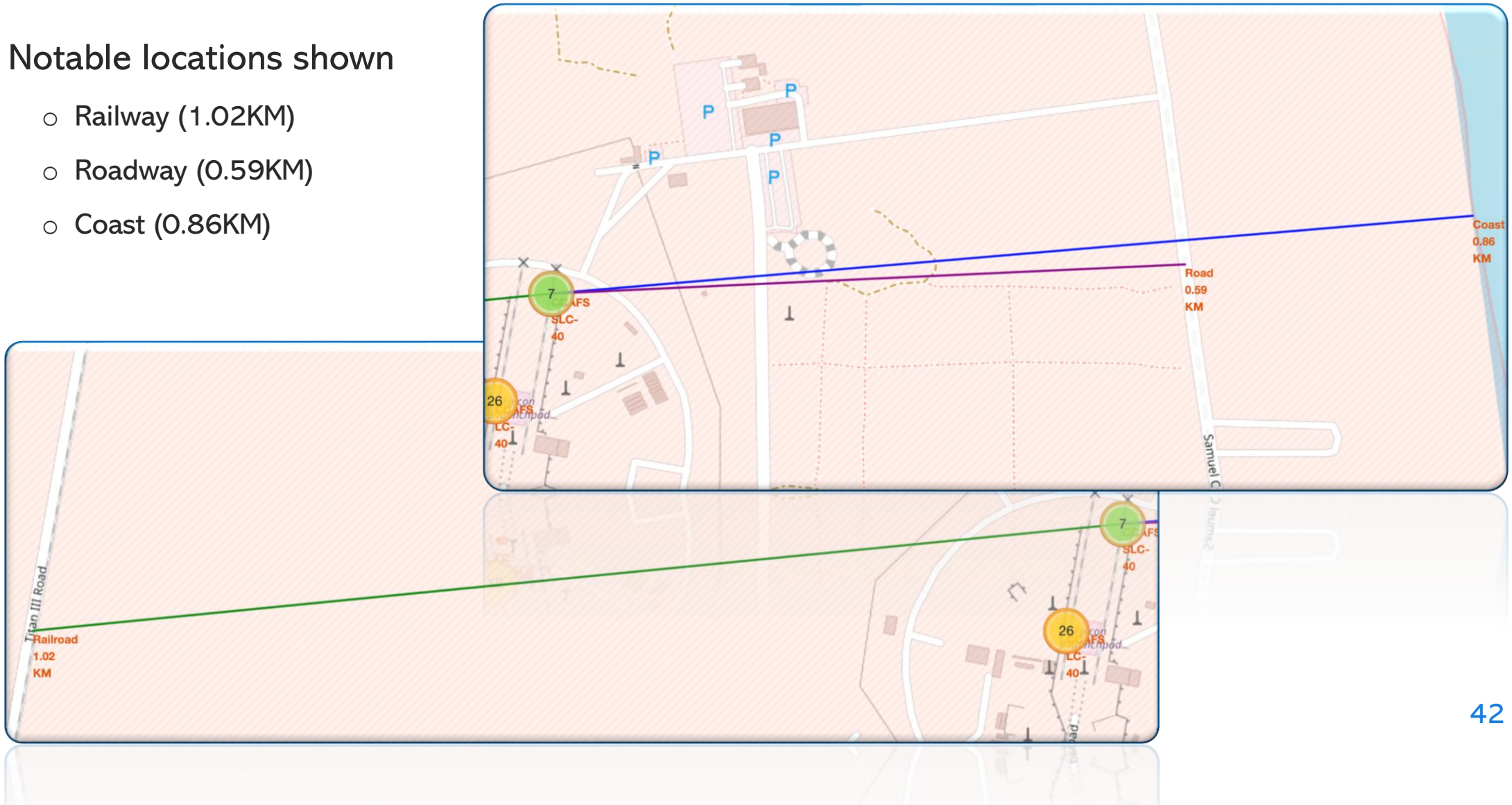


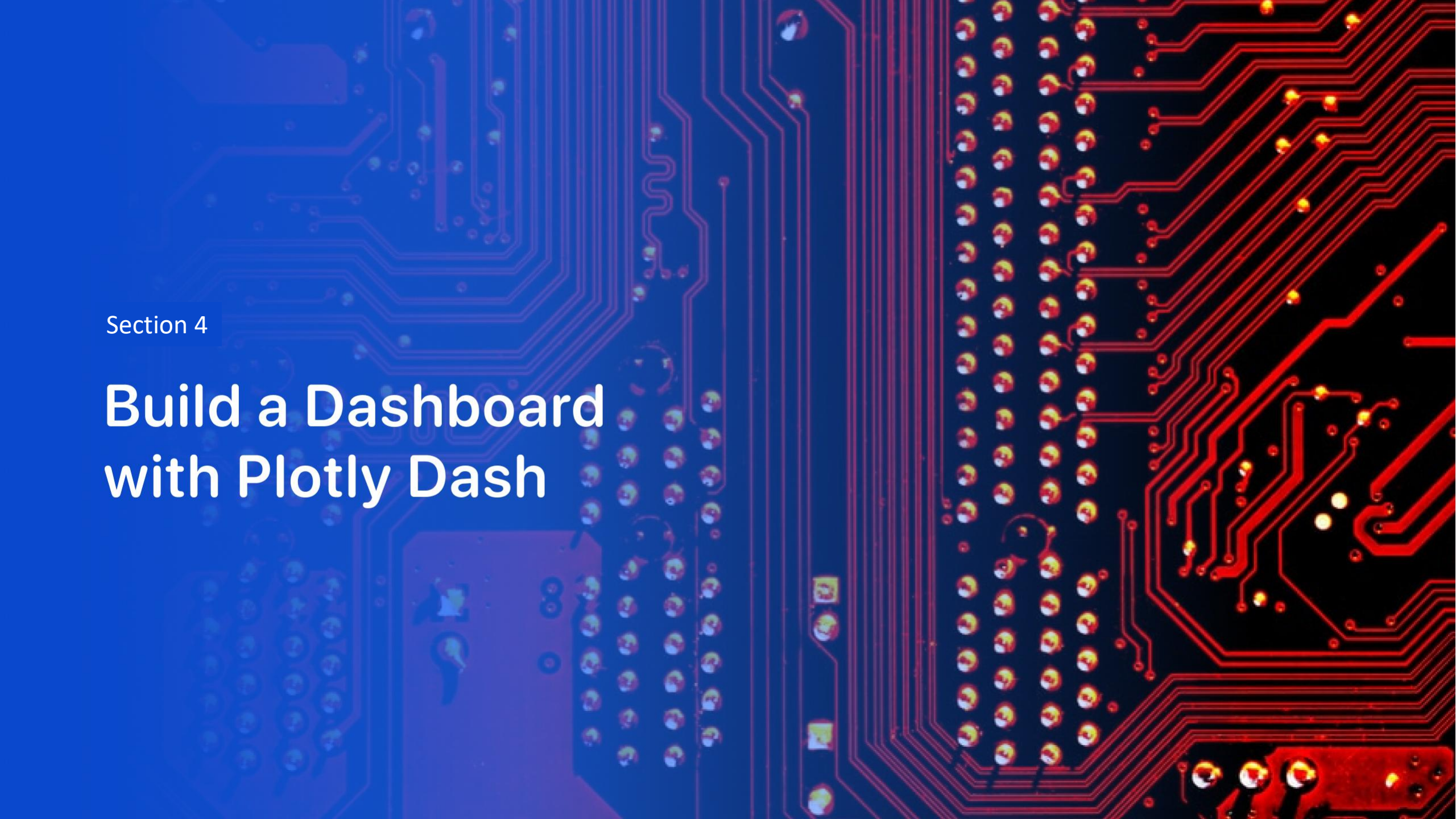
Launch Outcomes



Notable Proximate Locations

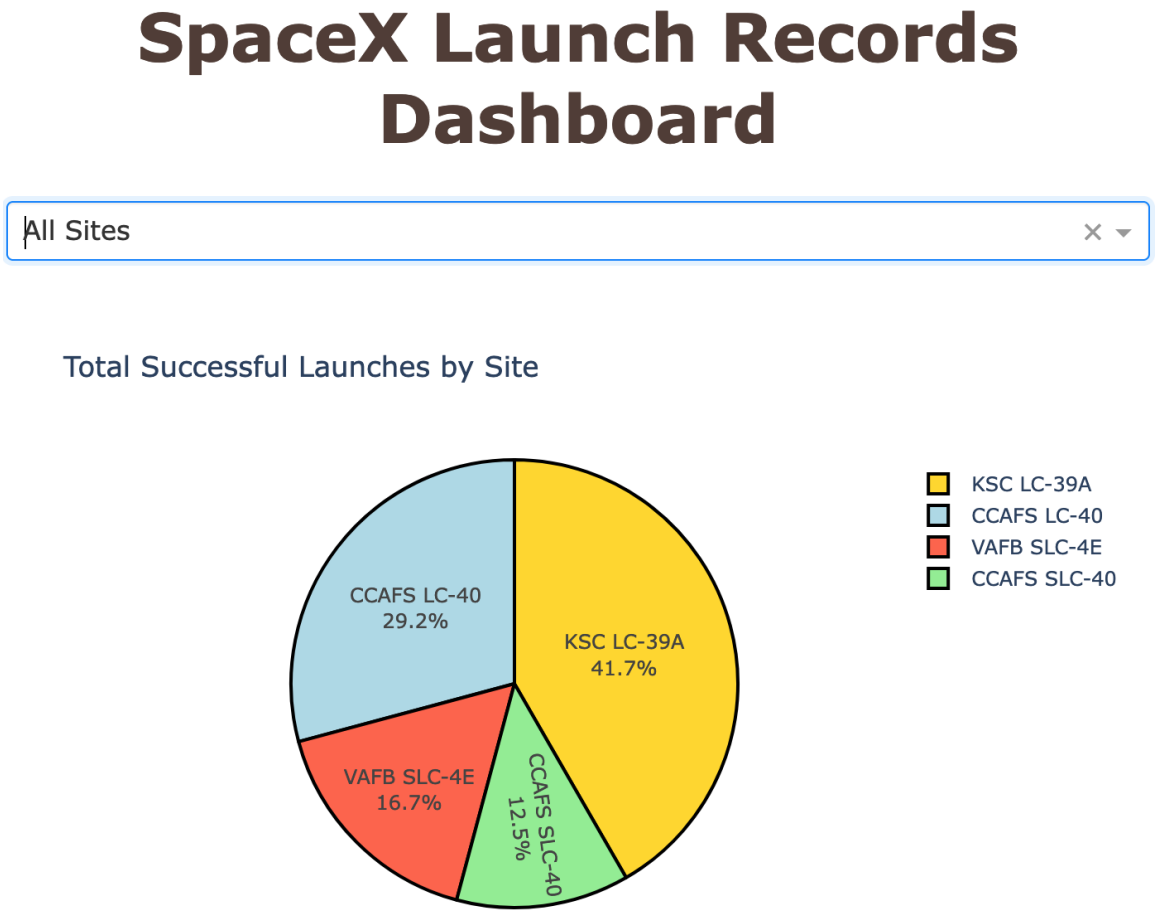
- Notable locations shown
  - Railway (1.02KM)
  - Roadway (0.59KM)
  - Coast (0.86KM)




Section 4
Build a Dashboard with Plotly Dash

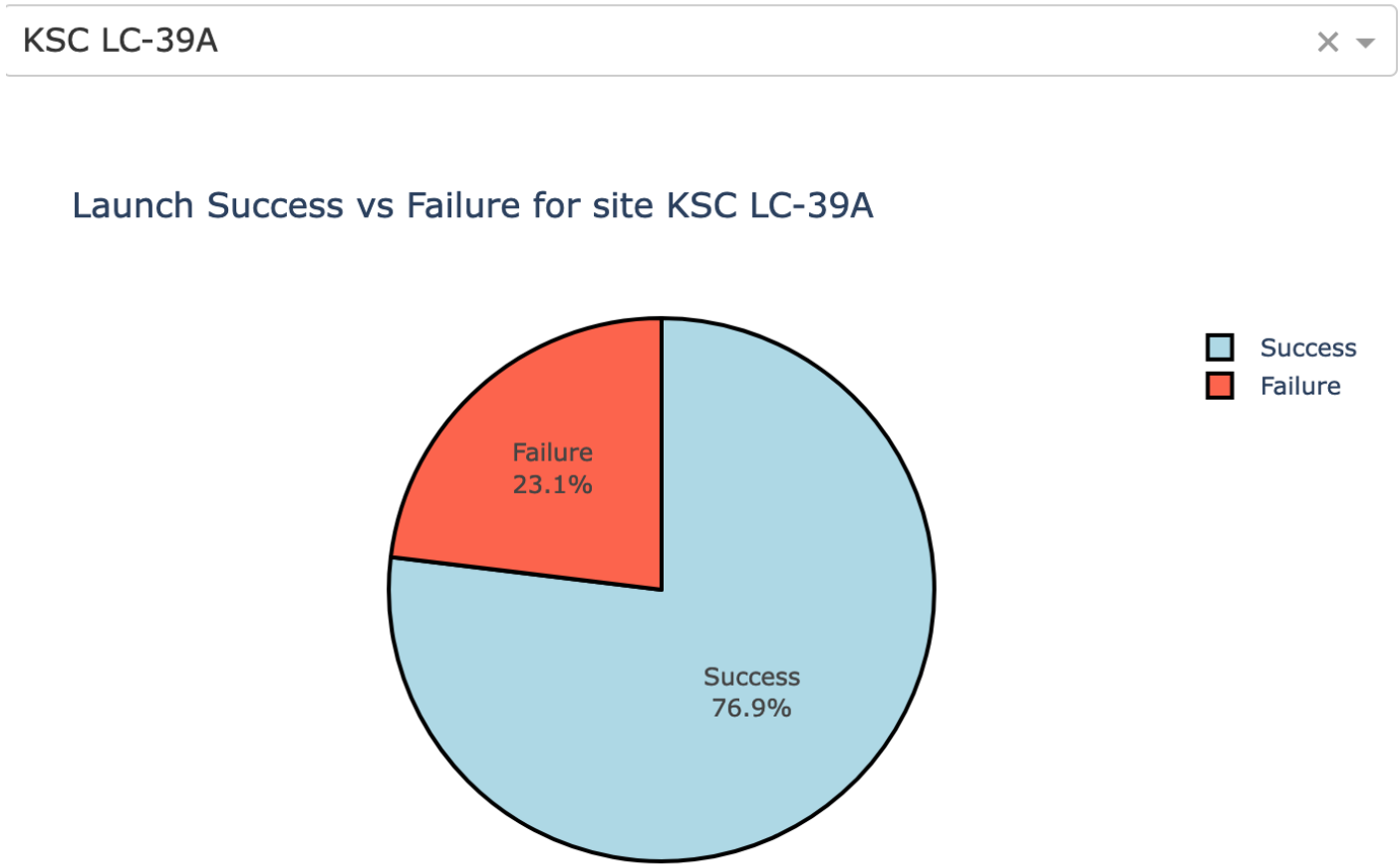
All Launch Sites: Successful Landings

- KSC LC-39A experienced the highest proportion of successful landings, followed by CCAFS LC-40.
- VAFB SLC-4E and CCAFS SLC-40 the lowest.

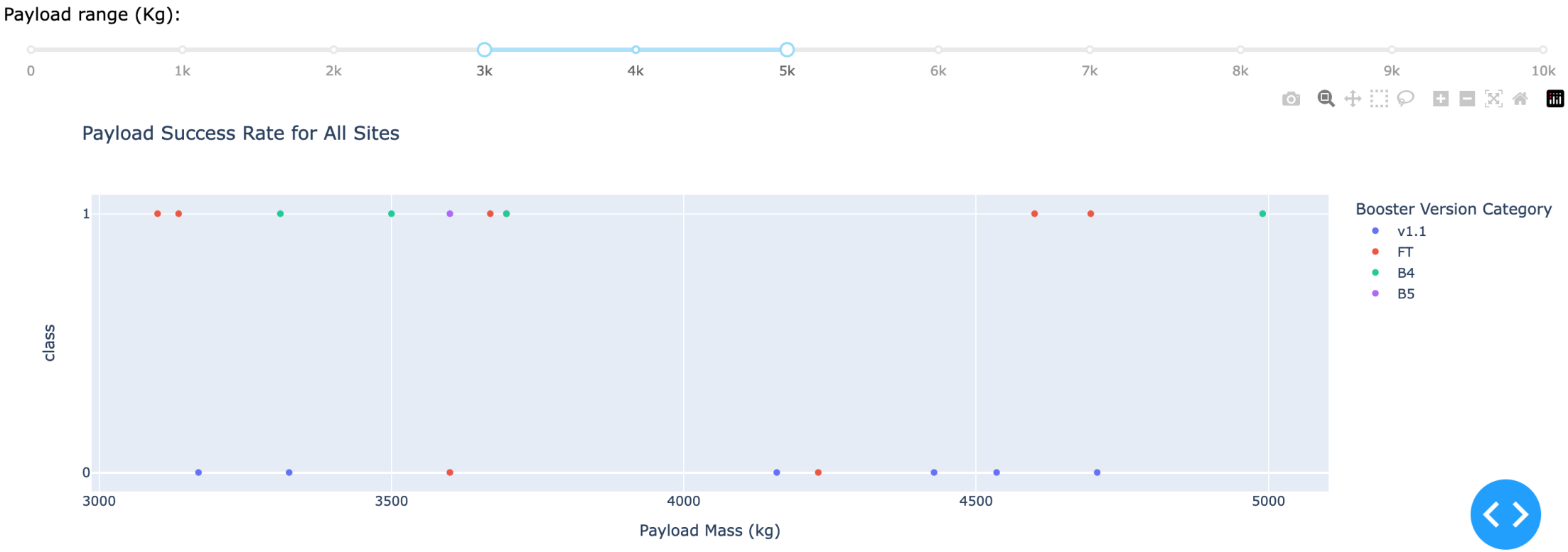


Per-site Launch Success Ratio: High

- KSC LC-39A had the highest ratio of successful landings



Payload Range



- With a payload mass between 3,000kg and 5,000kg, v1.1 boosters performed the worst.
- In the same payload range, B4 and B5 boosters had the best success rate, followed by FT.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

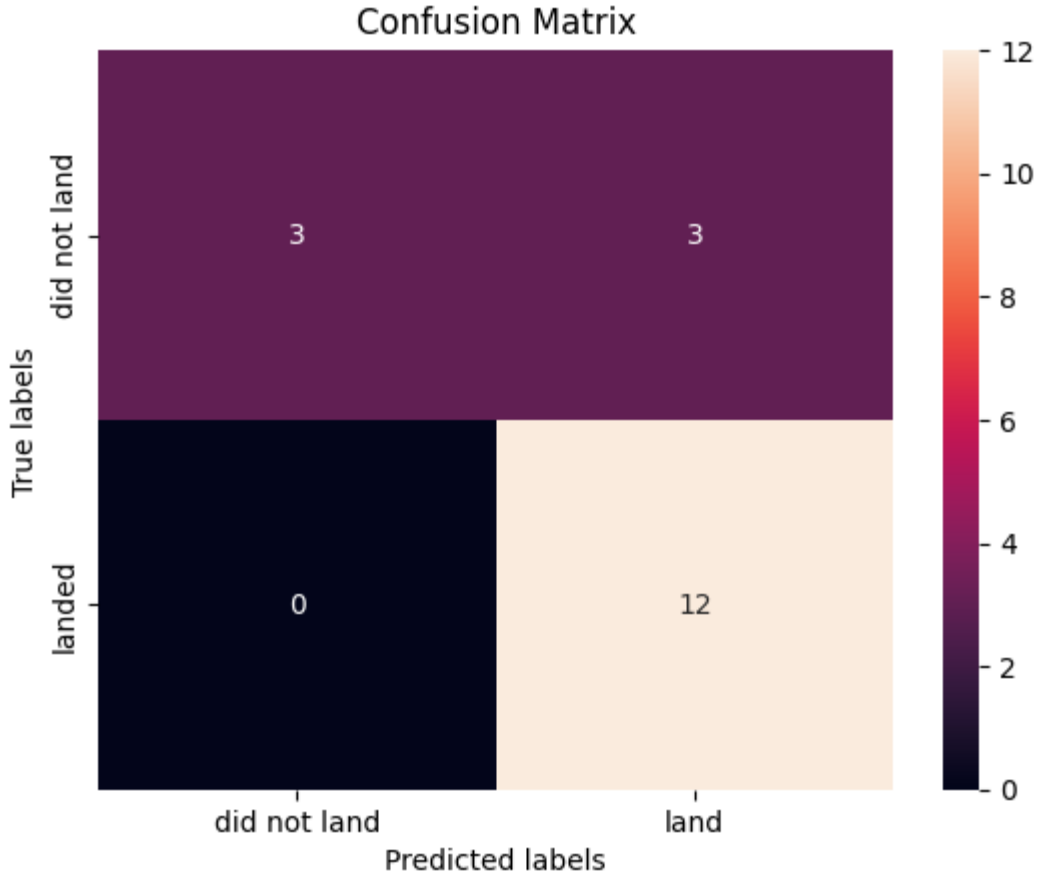
[41]:
```
best_model = max(test_scores, key=test_scores.get)
print("Best model (by test accuracy):", best_model)
print("Best test accuracy:", test_scores[best_model])
```

```
Best model (by test accuracy): Logistic Regression
Best test accuracy: 0.8333333333333334
```

- Of the algorithms tested, the Logistic Regression was the most accurate.

Confusion Matrix

- **True Negative (TN) = 3**
The model correctly predicted *did not land* when the true outcome was *did not land*.
- **False Positive (FP) = 3**
The model predicted *land* when the true outcome was *did not land*.
These are false alarms.
- **False Negative (FN) = 0**
The model predicted *did not land* when the true outcome was *landed*.
There are no missed landings.
- **True Positive (TP) = 12**
The model correctly predicted *land* when the true outcome was *landed*.



Conclusions

- Success rates increase over time, across all factors, which indicates continuous and incremental operational improvements and technological advancements.
- Different orbits have varying success rates, with ES-L1, SSO, HEO, and GEO showing consistently successful outcomes.
- Launch site was a highly predictive factor, with KSC LC-39A being a top performer, closely followed by CCAFS LC-40.
- Many of the predictive models evaluated were able to predict landing outcome with an acceptable level of accuracy. In the testing performed, Logistic Regression produced best results with high accuracy, precision, and recall.

Appendix

- Data Sources
  - SpaceX API
    - Collected Data: dataset_part_1.csv
  - Wikipedia: List of Falcon 9 and Falcon Heavy launches (June 2021)
  - Data after wrangling: dataset_part_2.csv
  - Geographical data: spacex_launch_geo.csv
  - Interactive data source: spacex_launch_dash.csv

Thank you!