IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SpaceX Falcon 9 First
Stage Landing Prediction

Date :17-12-2025

Outline

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

Executive Summary

- The goal of this project is to predict whether SpaceX will successfully land and reuse the Falcon 9 first stage.

- Public SpaceX launch data was collected using REST APIs and web scraping.

- Exploratory data analysis, interactive maps, dashboards, and machine learning models were developed.

- The best performing classification model achieved the highest accuracy in predicting launch success

Introduction

- SpaceX significantly reduces launch costs by reusing rocket first stages.
- Space Y aims to compete by estimating launch prices using data-driven methods.
- Predicting first-stage landing success helps estimate reusability and cost.
- This project uses historical launch data and machine learning to solve this problem

Section 1

Methodology

Methodology

Executive Summary

- Data Collection using SpaceX REST API
- Data Collection using Web Scraping (Wikipedia)
- Data Wrangling and Cleaning
- Exploratory Data Analysis using Visualization and SQL
- Interactive Visual Analytics using Folium
- Dashboard Development using Plotly Dash
- Predictive Analysis using Classification Models

Data Collection

- Launch data was collected using SpaceX REST APIs.
- Additional payload and booster details were obtained via web scraping.
- Data was converted from JSON and HTML tables into Pandas DataFrames.

Data Collection – SpaceX API

- SpaceX API endpoints were used to retrieve launch records.
- Relevant attributes such as payload, orbit, landing outcome were extracted.
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/jupyter-labs-webscraping.ipynb

```
[ ]: Start
       ↓
     Send HTTP Request to SpaceX REST API
       (requests.get)
       ↓
     Receive JSON Response
       ↓
     Parse JSON Data
       ↓
     Extract Required Fields
       (Flight No, Launch Site, Payload, Orbit, Outcome, etc.)
       ↓
     Convert to Pandas DataFrame
       ↓
     Data Ready for Wrangling & Analysis
     End
```

Data Collection - Scraping

- Scraped SpaceX data from Wikipedia.
- Parsed HTML tables using BeautifulSoup.
- Cleaned and stored data in a DataFrame.

- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/jupyter-labs-webscraping.ipynb

```
[ ]: Start
       ↓
     Send HTTP Request to Wikipedia Page
       ↓
     Receive HTML Content
       ↓
     Parse HTML using BeautifulSoup
       ↓
     Locate Launch Record Table (wikitable)
       ↓
     Extract Table Rows and Columns
       ↓
     Clean and Format Extracted Data
       ↓
     Convert Data into Pandas DataFrame
       ↓
     Web Scraped Data Ready for Integration
     End
```

Data Wrangling

- Raw data from API and web scraping were combined.
- Missing and inconsistent values were cleaned.
- Categorical features were encoded.
- A binary landing success variable was created.
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

```
[ ]: Start
       ↓
     Load Raw Dataset
       ↓
     Handle Missing Values
       ↓
     Filter Relevant Columns
       ↓
     Convert Categorical Values to Numerical
       ↓
     Create Landing Success Target Variable
       ↓
     Final Clean Dataset Ready for Analysis
     End
```

EDA with Data Visualization

- Scatter plots were used to analyze payload and launch success.
- Bar charts were used to compare success rates across orbit types.
- Line charts were used to observe yearly success trends.
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/edadataviz.ipynb

EDA with SQL

- Identified unique launch sites
- Calculated total and average payload masses
- Analyzed success and failure outcomes
- Ranked landing outcomes over time
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Launch sites were marked using map markers.
- Success and failure outcomes were color-coded.
- Distances to highways, railways, and coastlines were calculated.
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

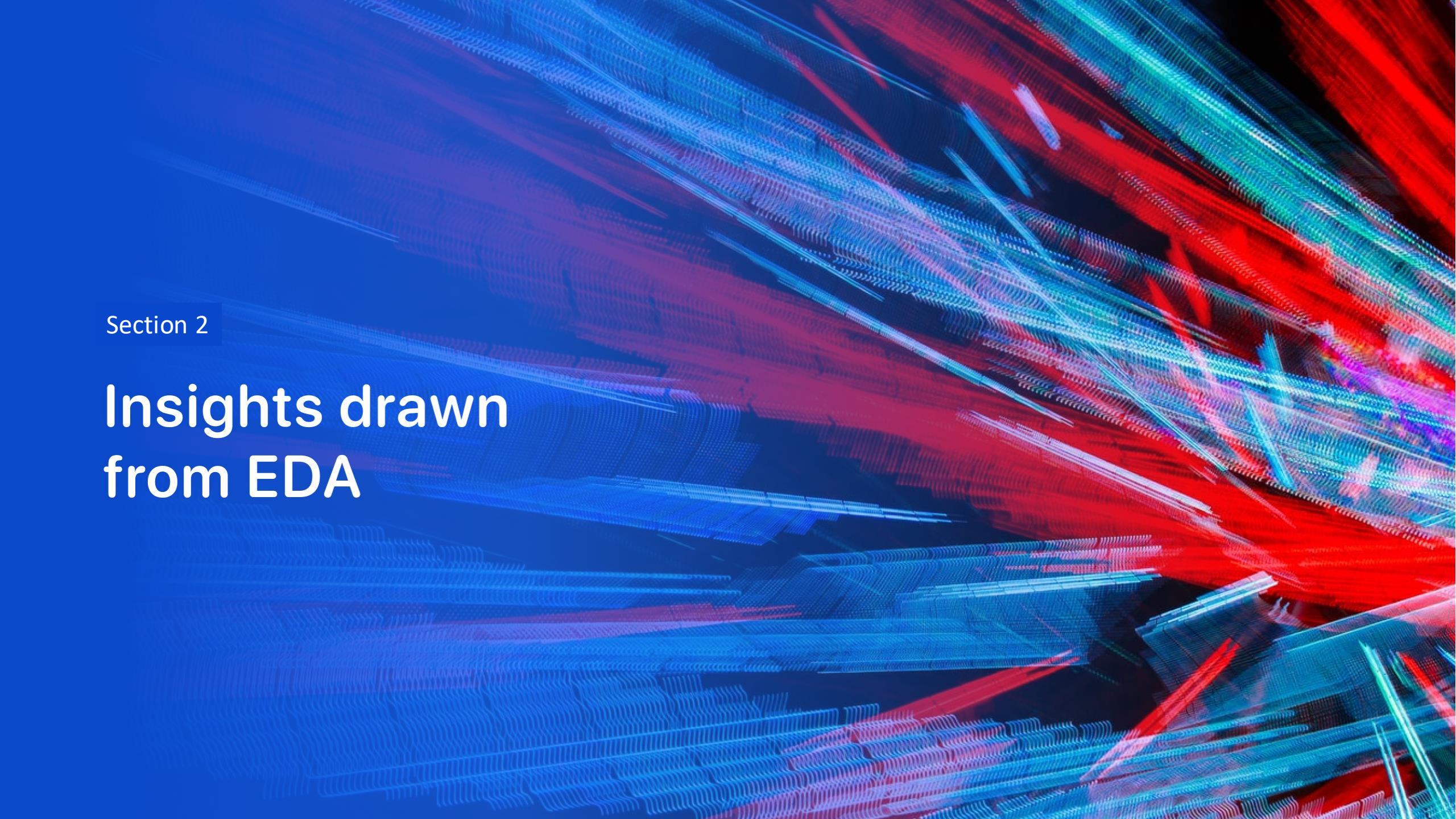
- A dropdown allows users to select launch sites.
- Pie charts show launch success distribution.
- A payload slider enables interactive filtering.
- Scatter plots show correlation between payload and success.
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Selected relevant launch features for prediction.
- Split data into training and testing sets.
- Trained multiple classification models.
- Tuned hyperparameters to improve performance.
- Selected the best model based on accuracy.
- https://github.com/RaghumanKhan02/Data-Science-Final-project-/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

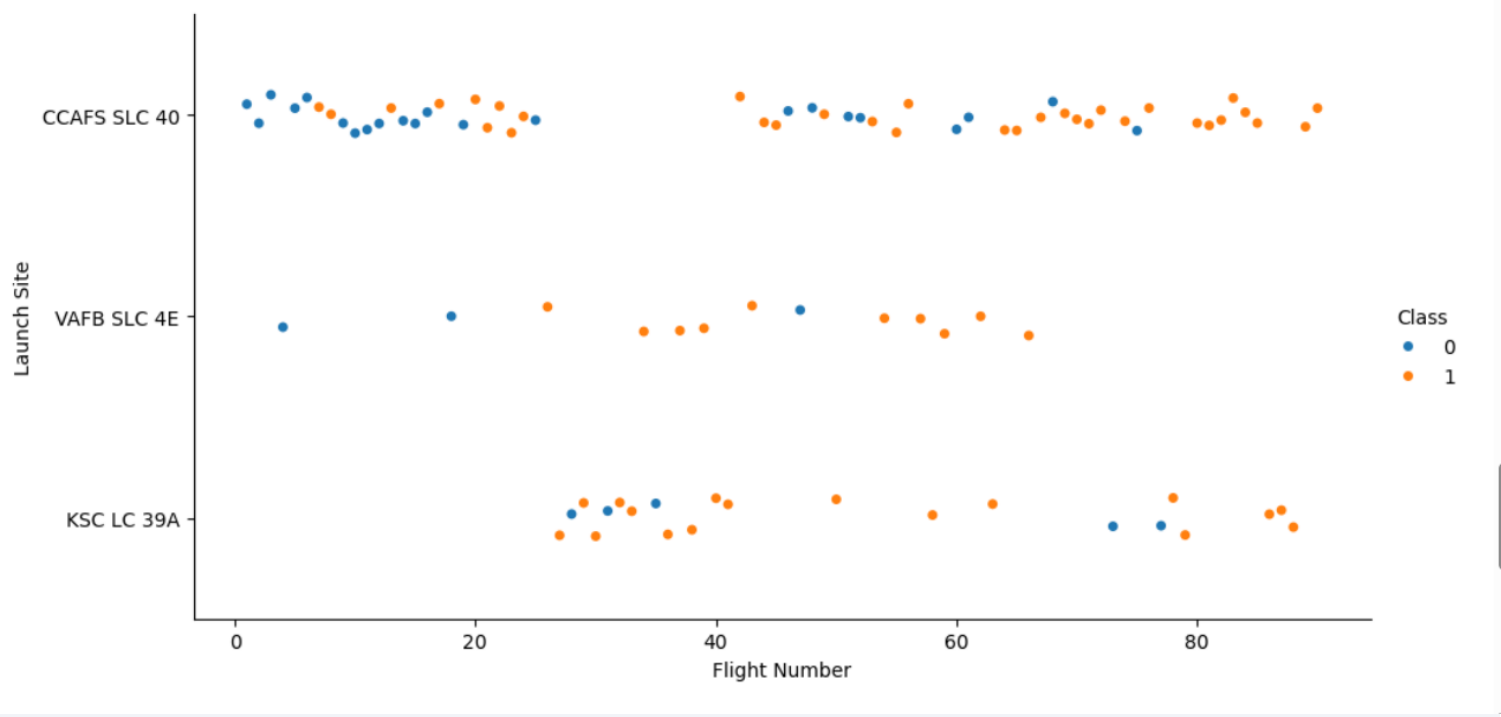
Results

- Exploratory data analysis results
  - Launch success varies by payload, orbit type, and launch site.
  - Certain orbits and sites show higher success rates.
  - Launch success has improved over time.
- Interactive analytics demo in screenshots
  - Interactive maps show launch site locations and outcomes.
  - Dashboards allow filtering by launch site and payload range.
  - Visual interactions help identify successful launch patterns.
- Predictive analysis results
  - Multiple classification models were trained and evaluated.
  - The best model achieved the highest prediction accuracy.
  - The model can effectively predict first-stage landing success.


Section 2
Insights drawn from EDA

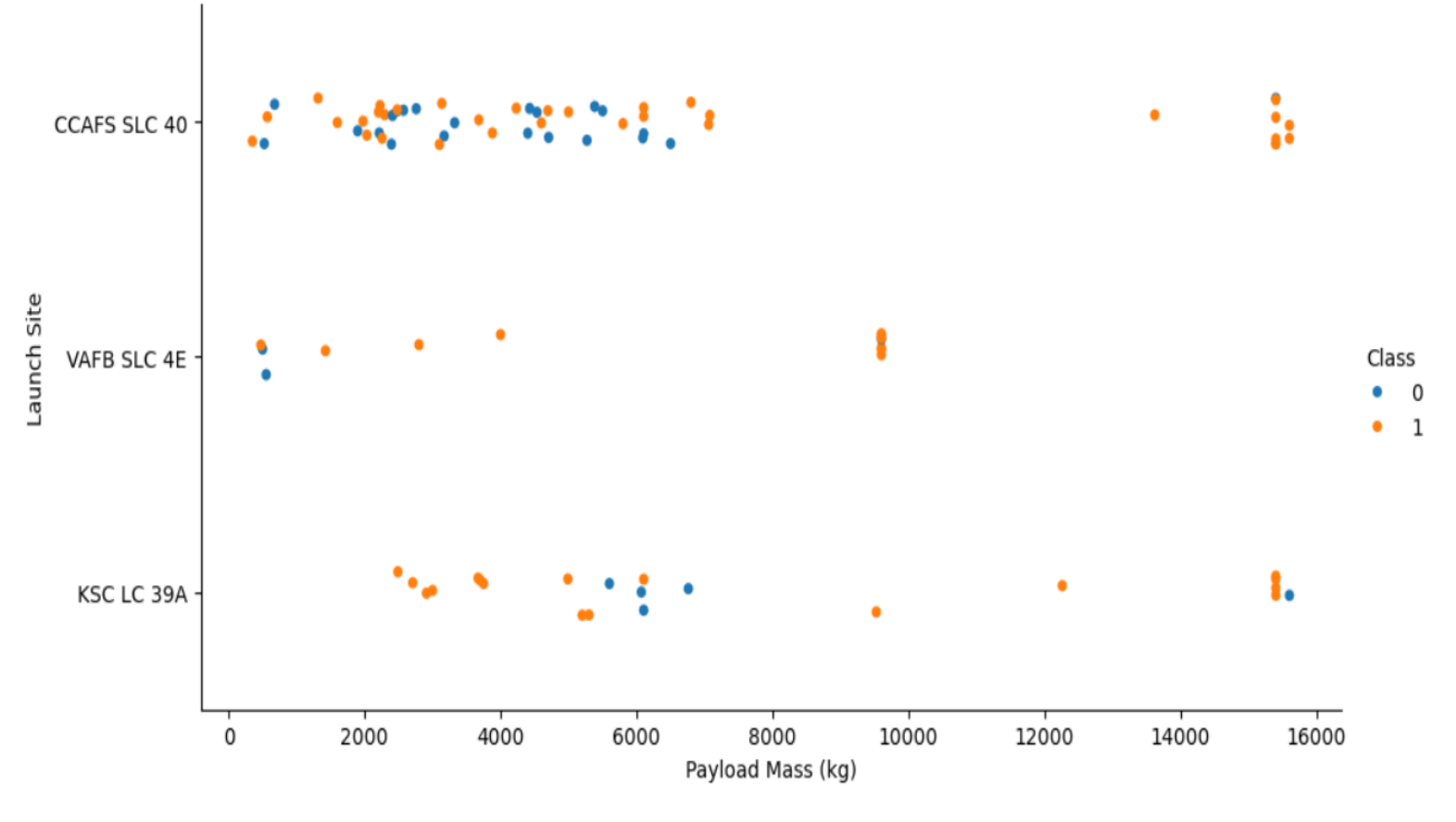
Flight Number vs. Launch Site

- The plot shows how launches are distributed across different launch sites over time.
- CCAFS LC-40 has the highest number of launches, especially in later flight numbers.
- This indicates that SpaceX increasingly relied on specific launch sites as experience and launch frequency increased.



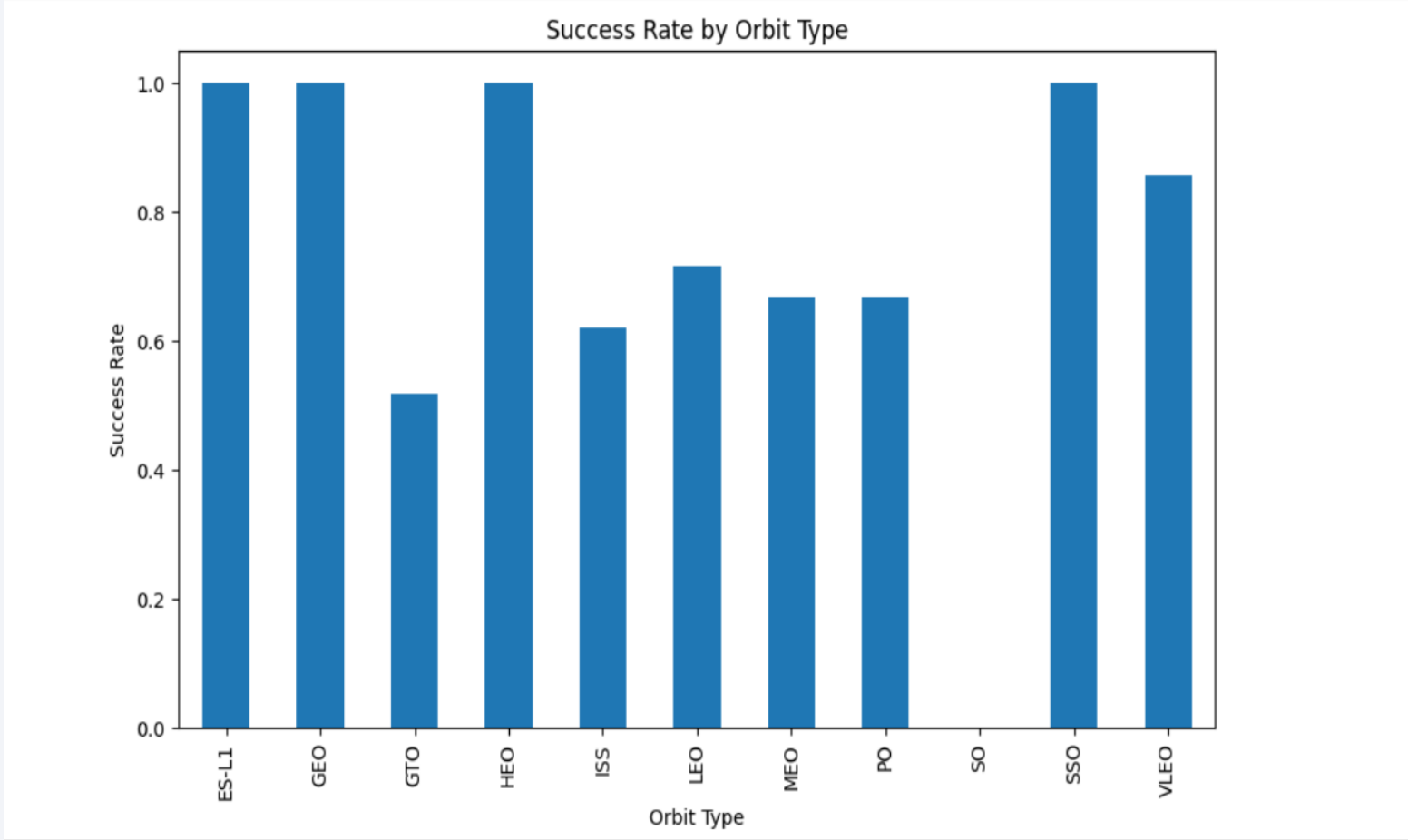
Payload vs. Launch Site

- The plot shows that payload mass varies across different launch sites.
- Higher payload missions are more frequently launched from CCAFS LC-40 and KSC LC-39A.
- This suggests that certain launch sites are better suited for heavier payload launches.



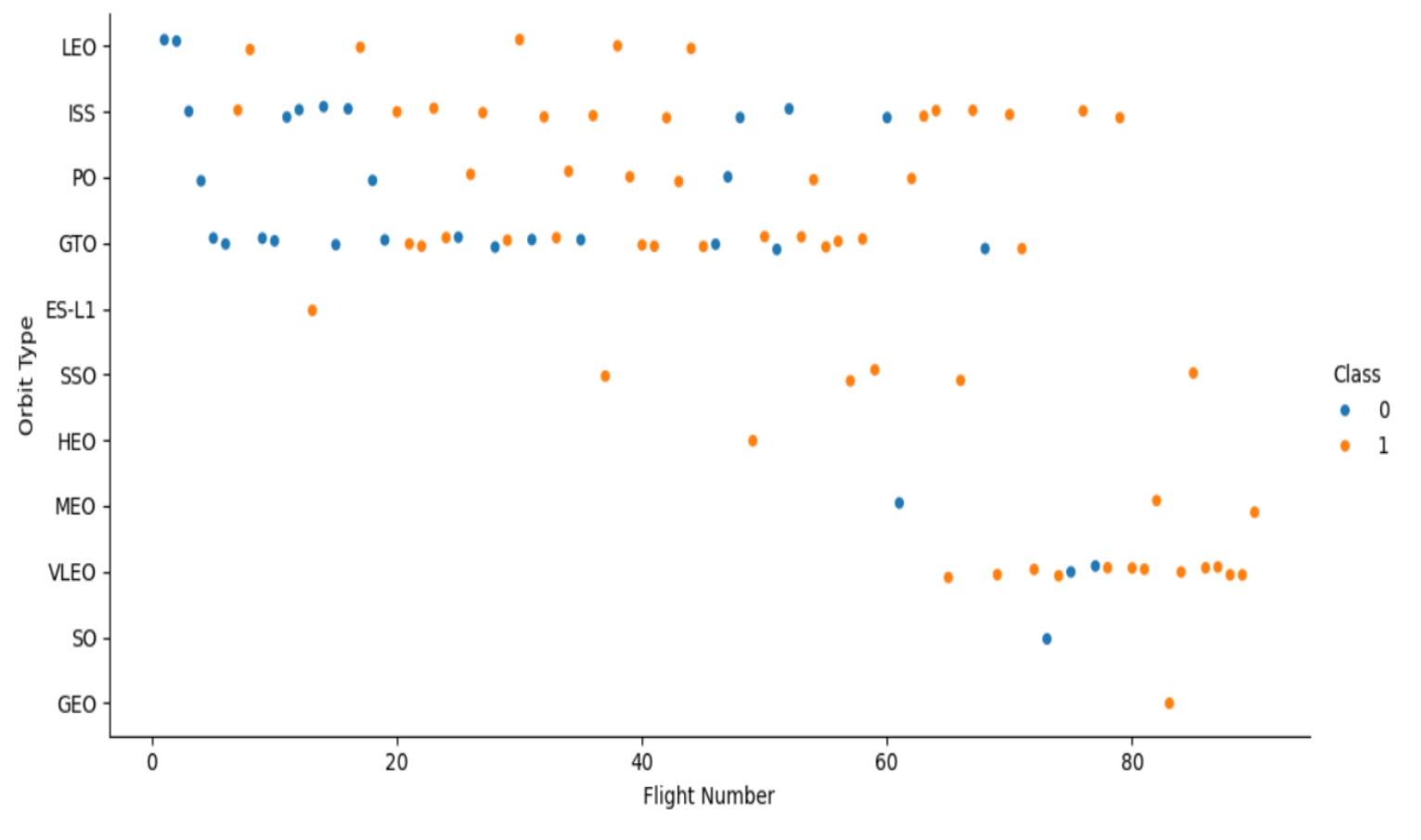
Success Rate vs. Orbit Type

- The chart shows that launch success rates vary across different orbit types.
- Orbits such as GTO and LEO have higher success rates compared to others.
- This indicates that mission orbit type influences launch and landing success.



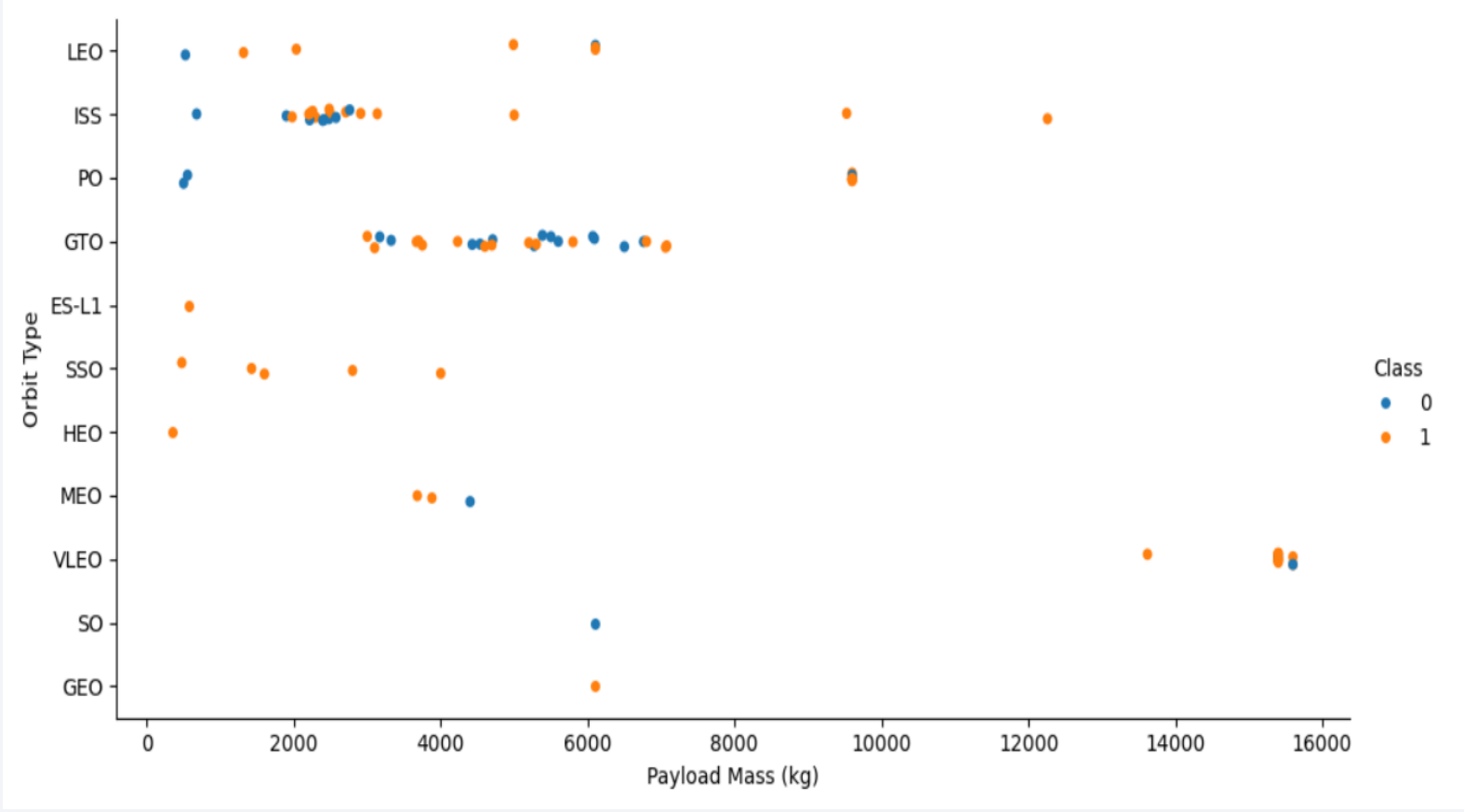
Flight Number vs. Orbit Type

- The plot shows how different orbit types are used across flight numbers.
- Early flights are concentrated on fewer orbit types, while later flights cover a wider range of orbits.
- This indicates that SpaceX expanded mission complexity and orbit diversity as flight experience increased.



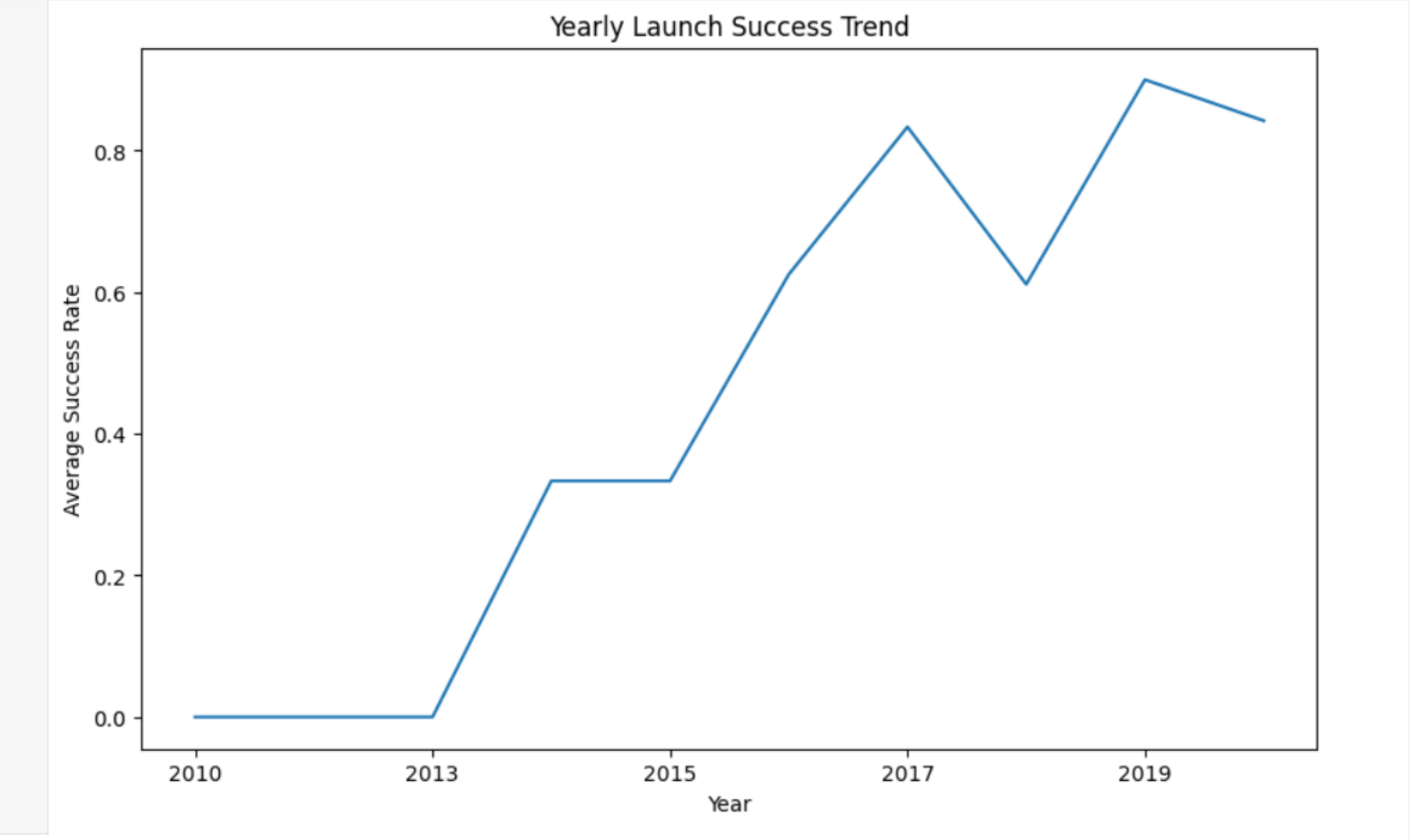
Payload vs. Orbit Type

- The plot shows that payload mass varies significantly across different orbit types.
- Higher payloads are more commonly associated with orbits such as GTO, while lower payloads are seen in LEO missions.
- This indicates that orbit type strongly influences the payload capacity of a launch.



Launch Success Yearly Trend

- The line chart shows an overall increase in launch success rate over the years.
- Earlier years have lower and more variable success rates, while recent years show consistently higher success.
- This trend indicates improvements in launch technology and operational experience over time.



All Launch Site Names

- Unique Launch Sites

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

- The query identifies all unique launch sites used by SpaceX.
- These sites represent different geographic locations used for Falcon 9 launches.

Launch Site Names 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

- This query filters launch records from launch sites beginning with 'CCA'. It shows early and frequent
- Falcon 9 launches from Cape Canaveral facilities.

Total Payload Mass

```
13]: %sql SELECT SUM("PAYLOAD_MASS__KG_") AS TOTAL_PAYLOAD_MASS FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';

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

13]:

TOTAL_PAYLOAD_MASS
45596

- This query calculates the total payload mass carried by boosters used in NASA missions.
- It shows the combined payload contribution of NASA-related launches.

Average Payload Mass by F9 v1.1

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

 * sqlite:///my_data1.db
Done.
14]:
```

AVG_PAYLOAD_MASS
2928.4

- This query calculates the average payload mass carried by the Falcon 9 v1.1 booster.
- It shows the typical payload capacity handled by this booster version.

First Successful Ground Landing Date

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

 * sqlite:///my_data1.db
Done.
15]:
```

FIRST_SUCCESS_DATE
2015-12-22

- This query identifies the first date when a Falcon 9 booster successfully landed on a ground pad.
- It marks an important milestone in SpaceX's reusable rocket development.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "P
```

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

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

- This query identifies boosters that successfully landed on a drone ship with payload mass between 4000 kg and 6000 kg.
- It highlights boosters capable of handling medium-heavy payload missions.

Total Number of Successful and Failure Mission Outcomes

```
17]: %sql SELECT "Mission_Outcome", COUNT(*) AS TOTAL FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

17]:

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

- This query calculates the total number of successful and failed SpaceX missions.
- The results show that most missions were completed successfully.

Boosters Carried Maximum Payload

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

```
 * 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

- This query identifies the booster that carried the maximum payload mass.
- It highlights the booster used for the heaviest SpaceX mission.

2015 Launch Records

```
[9]: %sql SELECT substr("Date", 6, 2) AS MONTH,"Landing_Outcome","Booster_Version","Launch_Site" FROM SPACEXTABLE WHERE substr("Date", 1, 4) = '201
```

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

[9]:

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

- This query identifies failed drone ship landing attempts in the year 2015.
- It shows the booster versions and launch sites associated with these failures.

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

```
]: %sql SELECT "Landing_Outcome", COUNT(*) AS TOTAL_COUNT FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_C
```

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

]:

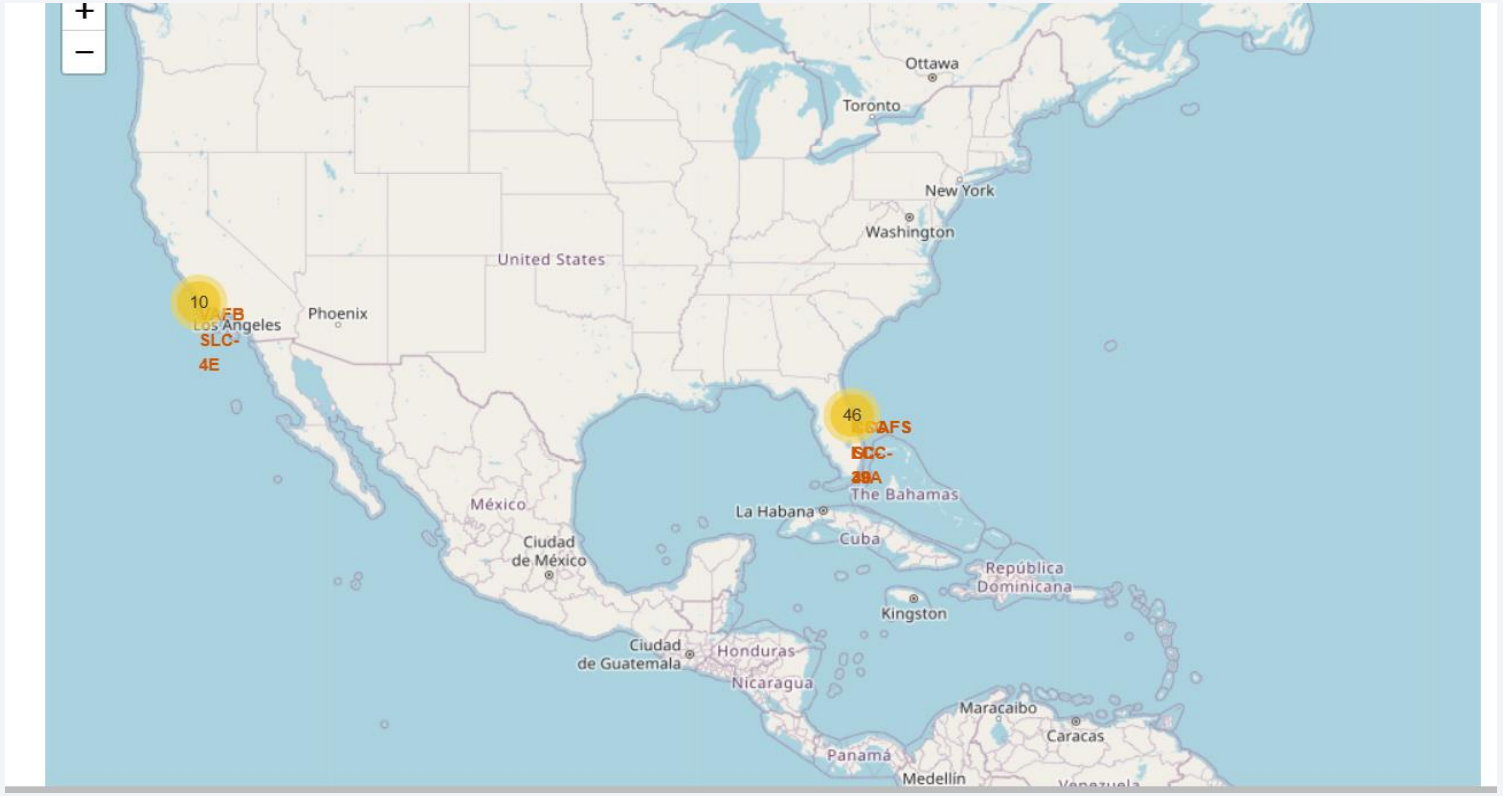
Landing_Outcome	TOTAL_COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This query ranks landing outcomes based on their frequency within the given time period.
- It shows that drone ship landings were the most common outcomes during early SpaceX missions.

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

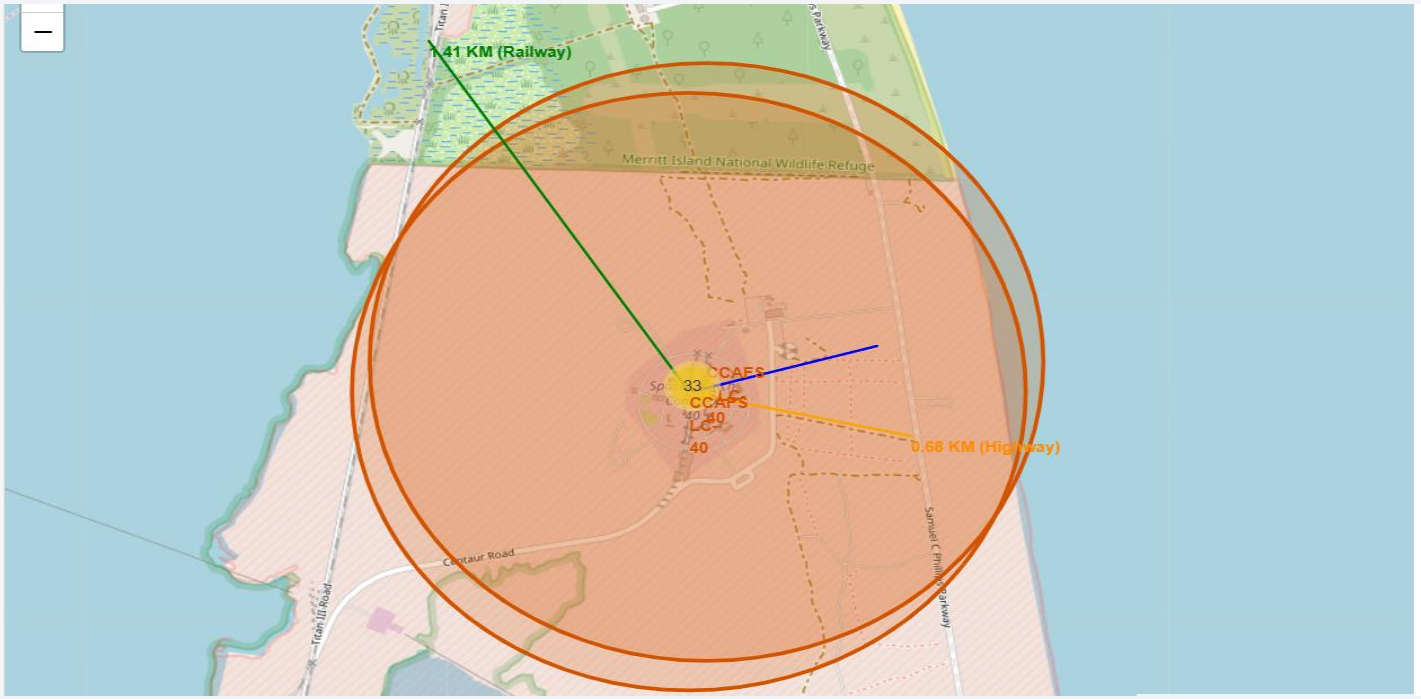


- Launch sites are located near coastlines for safety and launch efficiency.
- Multiple launches are concentrated at specific locations.
- Geographic location plays an important role in launch operations.

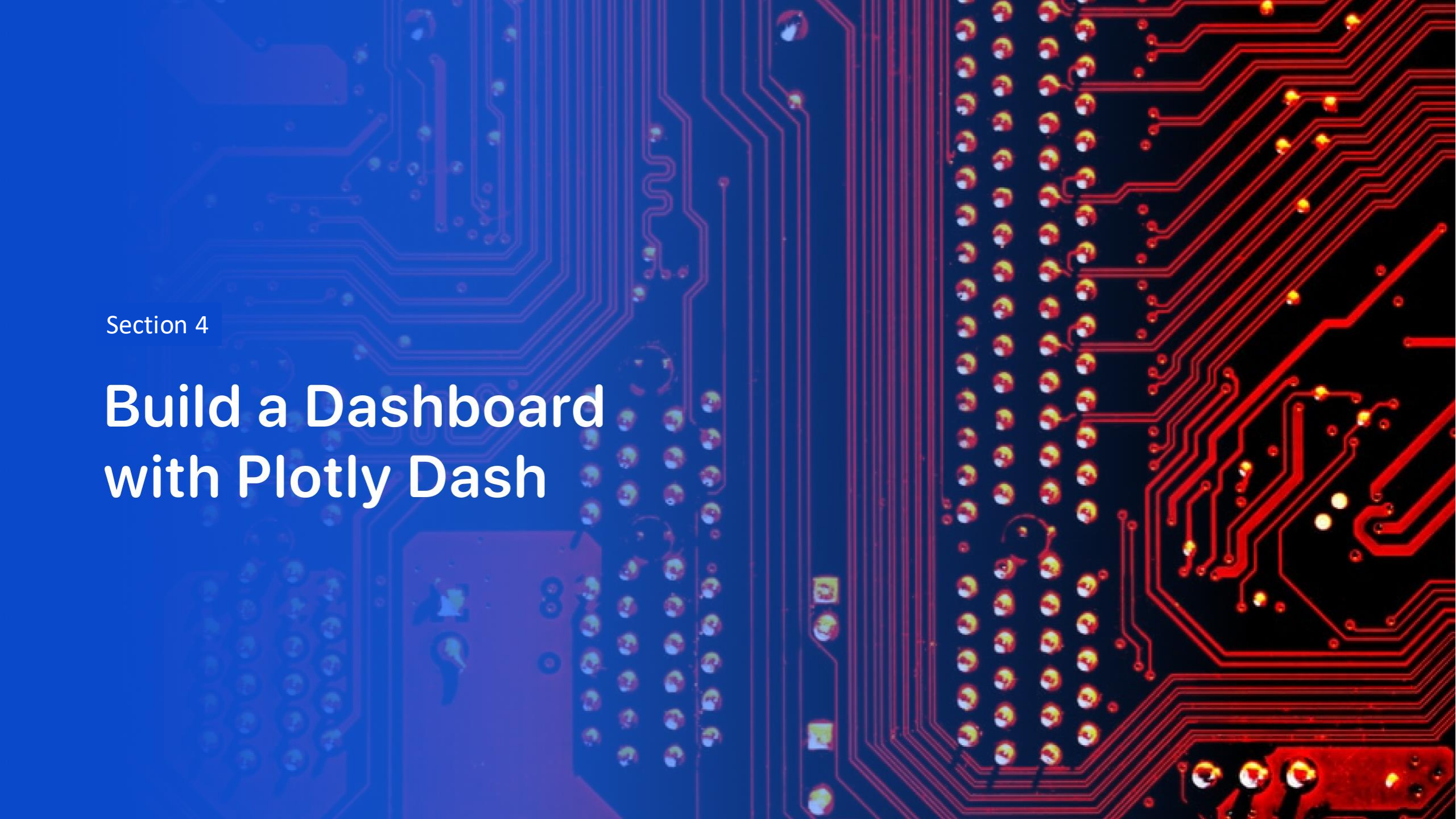
<Folium Map Screenshot 2>

- Successful launches are more frequent than failures.
- Certain launch sites show higher success rates.
- Visual clustering helps identify reliable launch locations.

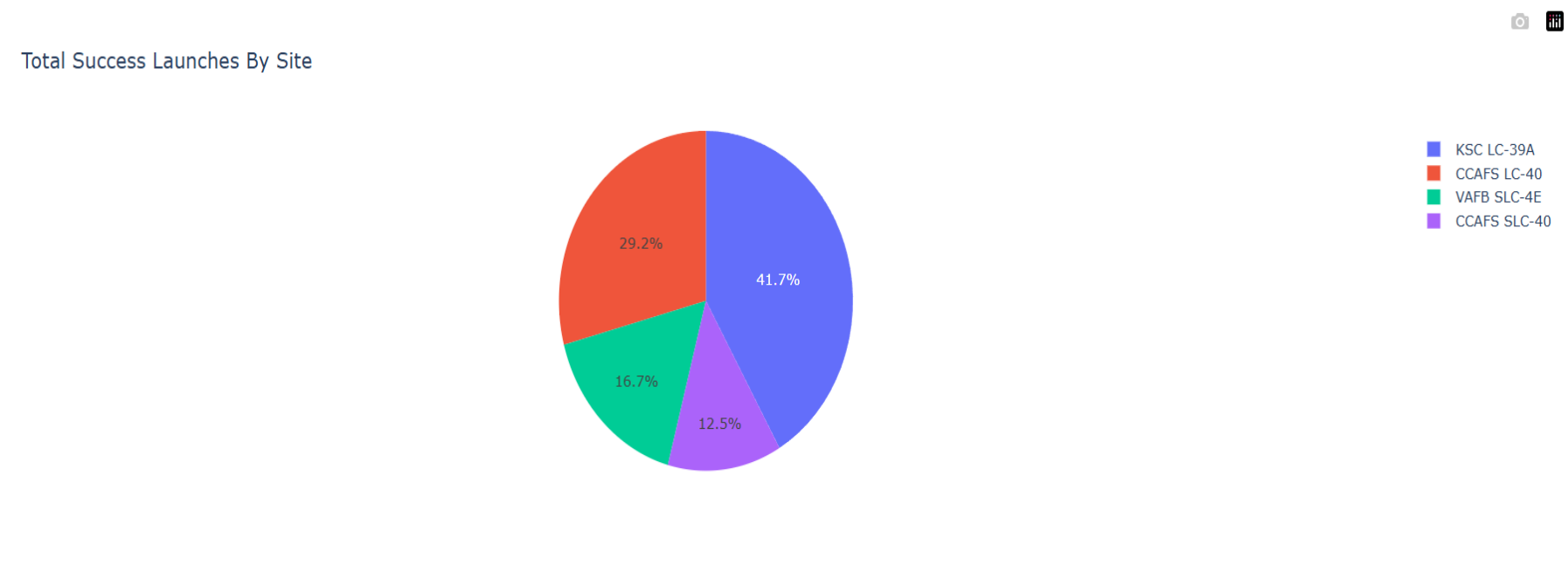
<Folium Map Screenshot 3>



- The launch site is located close to the coastline, improving launch safety.
- Proximity to highways and railways supports efficient transportation and logistics.
- The distances indicate that launch sites are strategically placed for accessibility and risk reduction.

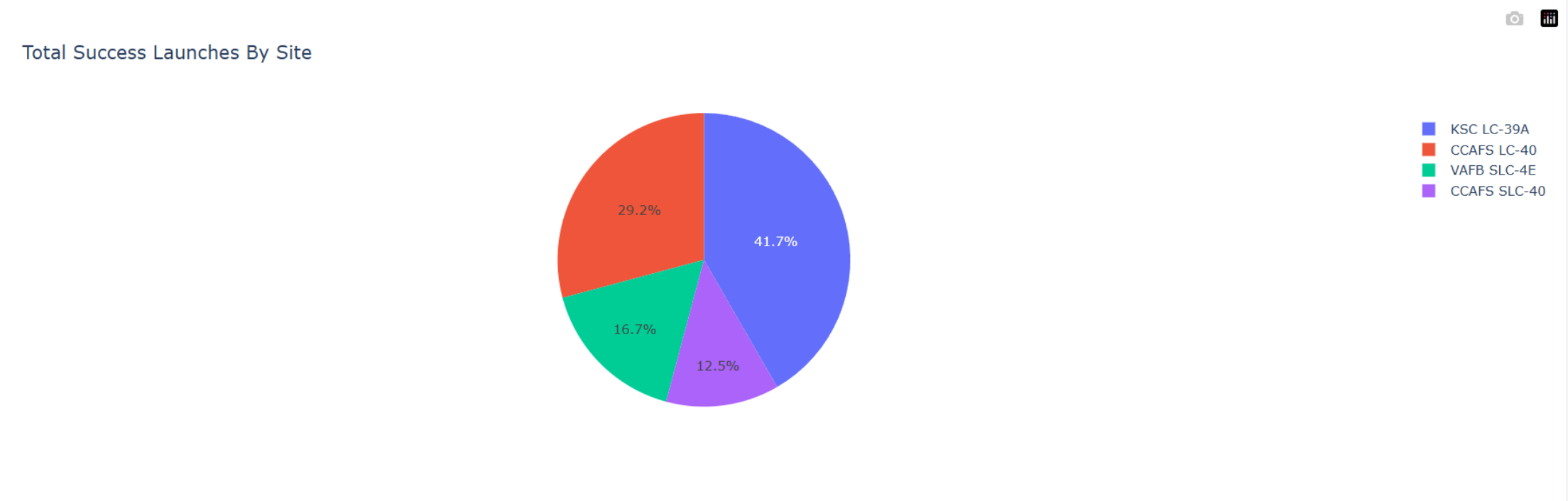

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



- KSC LC-39A has the highest share of successful launches.
- CCAFS LC-40 also contributes a significant portion of successful missions.
- VAFB SLC-4E and CCAFS SLC-40 have fewer successful launches in comparison.
- This indicates that launch success is not evenly distributed across sites.

<Dashboard Screenshot 2>



- KSC LC-39A shows the highest launch success ratio among all sites.
- This indicates a higher proportion of successful launches compared to failures at this location.

<Dashboard Screenshot 3>

<Dashboard Screenshot 4>

<Dashboard Screenshot 5>

<Dashboard Screenshot 6>

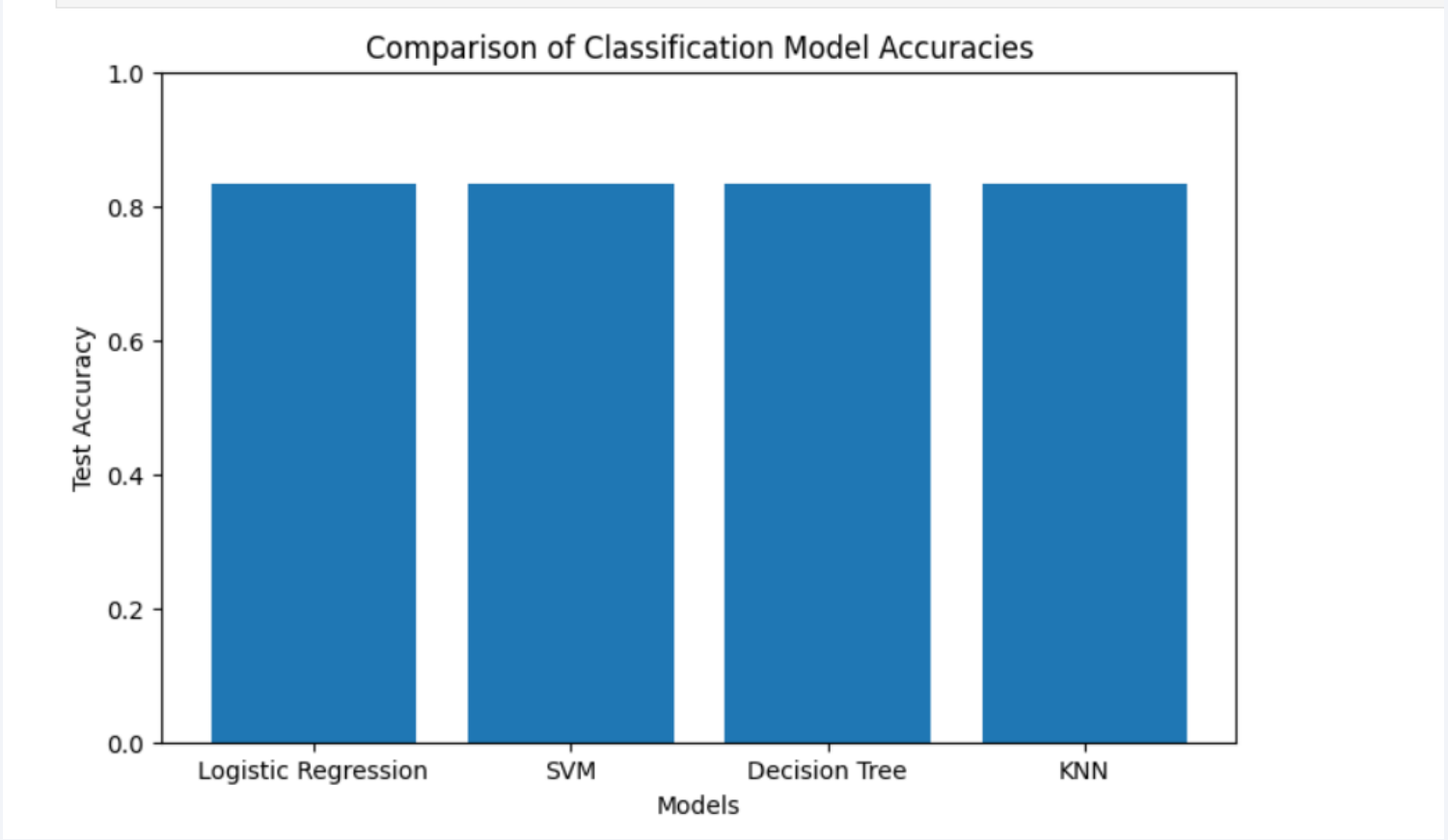
- Mid-range payloads (around 2,500–6,000 kg) show the highest concentration of successful launches.
- Very low payload ranges include more failures, especially for earlier booster versions.
- At higher payloads (above ~6,000 kg), success is still achievable but appears more dependent on newer boosters.
- Newer booster versions (FT, B4, B5) show higher success rates compared to older versions (v1.0, v1.1).
- The heaviest payloads (near ~9,500–10,000 kg) are rare but include successful launches using advanced boosters.

Section 5

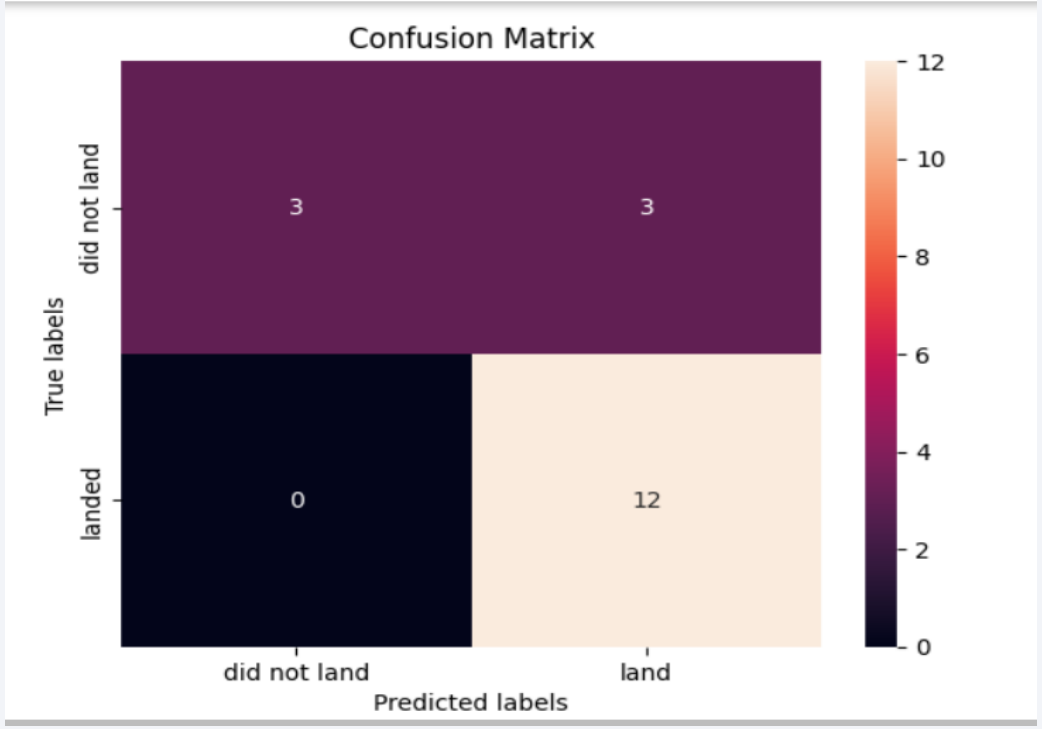
Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree classifier achieved the highest classification accuracy
- during cross-validation (0.8875) and was selected as the best-performing model.
- Although test accuracy is similar across models, Decision Tree performed best during hyperparameter tuning and cross-validation.



Confusion Matrix



- The model correctly predicts most successful landings (high true positives).
- There are zero false negatives, meaning no successful landing was predicted as a failure.
- This is important because predicting a failure when the booster actually lands can lead to wrong cost estimates.
- Although a few failures are misclassified as successes, overall prediction performance is strong.

Conclusions

- SpaceX launch success is influenced by payload mass, orbit type, and launch site.
- Certain launch sites, such as KSC LC-39A, show higher launch success ratios.
- Launch success rates have increased over time, indicating operational and technological improvements.
- Interactive maps and dashboards helped visualize launch locations, outcomes, and payload effects effectively.
- Machine learning models can reliably predict first-stage landing outcomes using historical data.
- The Decision Tree model achieved the highest classification accuracy during model tuning.
- The predictive model can support Space Y in estimating launch costs and planning competitive launch strategies.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!