IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Analyzed SpaceX Falcon 9 launch data using data wrangling, exploratory data analysis, and interactive visual analytics.
- Geographic analysis showed launch sites are close to coastlines and infrastructure while maintaining safe distances from cities.
- Built and evaluated classification models including Logistic Regression, SVM, Decision Tree, and KNN using GridSearchCV.
- Support Vector Machine (RBF kernel) achieved the best prediction accuracy (~85%).
- Interactive Dash dashboard revealed KSC LC-39A as the most successful launch site, optimal payload range between 4,000–8,000 kg, and FT/B5 as the best booster versions.

Introduction

Introduction

• SpaceX aims to reduce launch costs by reusing the Falcon 9 first-stage booster, making launch success a critical factor in mission planning.

- Understanding historical launch performance helps identify key factors that influence successful landings and reusability.
- This project analyzes past SpaceX launch data to uncover patterns related to payload mass, launch site, orbit type, and booster version.

Key Questions Addressed:

- Which launch sites have the highest success rates?
- How does payload mass affect launch success?
- Which booster versions perform best?
- Can machine learning models accurately predict launch success?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected SpaceX launch data from APIs and CSV datasets provided in the project.
  - Included launch site, payload mass, orbit type, booster version, and launch outcome.
- Perform data wrangling
  - Cleaned and preprocessed data by handling missing values and encoding categorical eatures and Prepared the dataset for analysis and machine learning modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built classification models including Logistic Regression, SVM, Decision Tree, and KNN.
  - Tuned hyperparameters using GridSearchCV and evaluated models using test accuracy.

Data Collection

- Launch data was collected from publicly available SpaceX APIs and provided CSV files.
- The datasets were extracted, merged, and prepared for further analysis and modeling.

SpaceX Launch Data

↓

REST APIs / CSV Files

↓

Data Extraction (Python)

↓

Data Integration

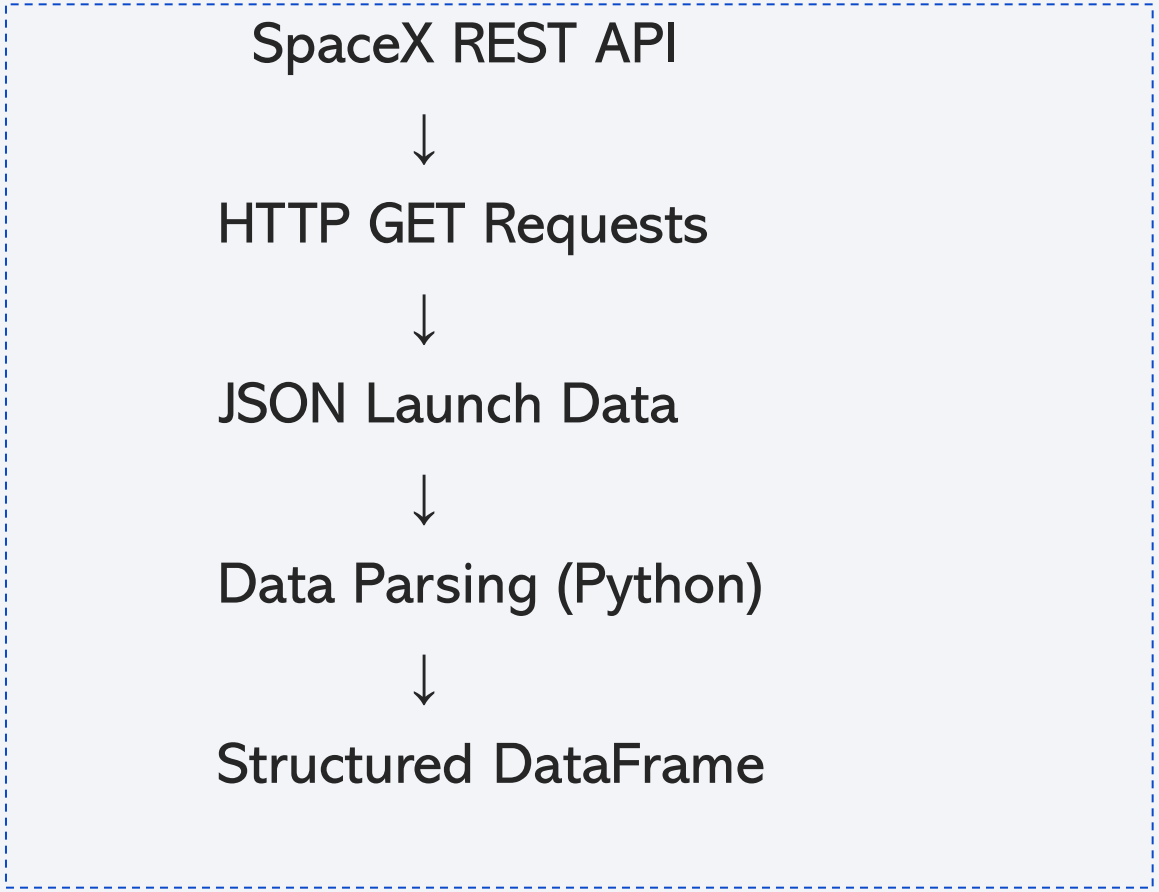
↓

Cleaned Dataset for Analysis

Data Collection – SpaceX API

Key Phrases

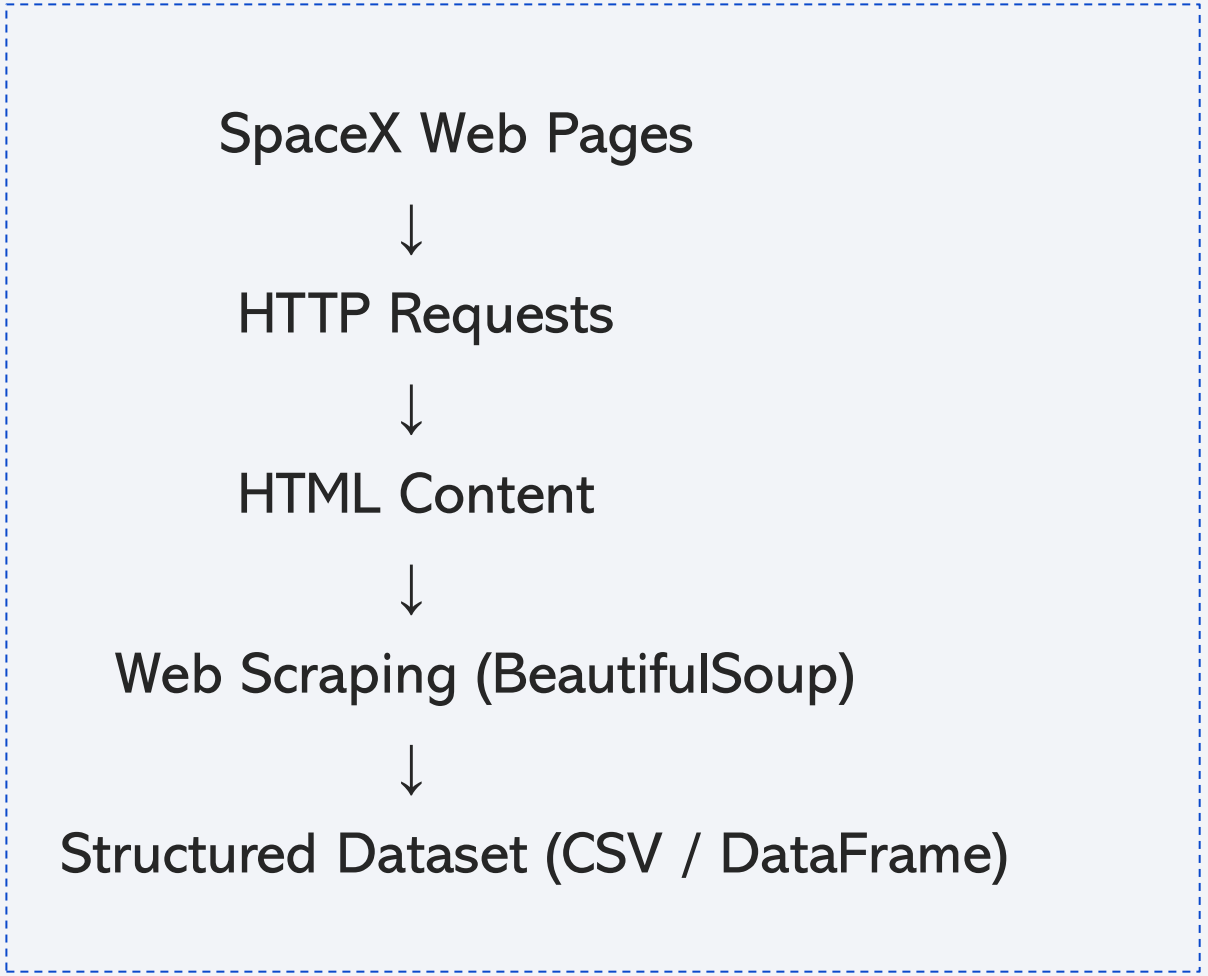
- SpaceX REST API
- Falcon 9 launch records
- JSON data retrieval
- Automated data extraction
- Python API requests
- https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

Key Phrases

- Web scraping
- SpaceX launch webpages
- HTML content extraction
- BeautifulSoup
- Automated data collection
- https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Key Phrases

- Data cleaning
- Handling missing values
- Feature selection
- Categorical encoding
- Data transformation

https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/edadataviz.ipynb

EDA with Data Visualization

- **Bar Charts**
  Used to compare launch success rates across different launch sites and orbit types.
- **Scatter Plots**
  Used to analyze the relationship between payload mass and launch success outcomes.
- **Box Plots**
  Used to understand the distribution of payload mass for successful and failed launches.
- **Interactive Maps (Folium)**
  Used to visualize launch site locations and analyze proximity to coastlines, highways, railways, and cities.
- https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

EDA with SQL

EDA Using SQL

- Queried total number of launches and successful launches.
- Calculated success rates grouped by launch site and orbit type.
- Analyzed average payload mass for successful and failed launches.
- Filtered launches based on booster versions and mission outcomes.
- Identified trends and patterns using grouped and aggregated SQL queries.

- https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers & Circle Markers

  Used to display SpaceX launch site locations and differentiate successful and failed launches.

- Marker Clusters

  Applied to group nearby launch points and improve map readability at different zoom levels.

- Circles

  Added to visualize proximity zones around launch sites for geographic analysis.

- Polylines

  Used to show distances between launch sites and nearby coastlines, highways, railways, and cities.

- Popups & Labels

  Added to provide interactive information such as launch site names and distance values.

https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Launch Site Dropdown

  Added to allow users to filter visualizations by individual launch sites or view all sites together.

- Success Pie Chart

  Used to visualize launch success versus failure and compare success rates across launch sites.

- Payload Range Slider

  Added to dynamically filter launches based on payload mass and analyze performance across payload ranges.

- Payload vs Success Scatter Plot

  Used to examine the relationship between payload mass and launch success, colored by booster version.

- Interactive Callbacks

  Implemented to update charts in real time based on user selections, enabling interactive data exploration.

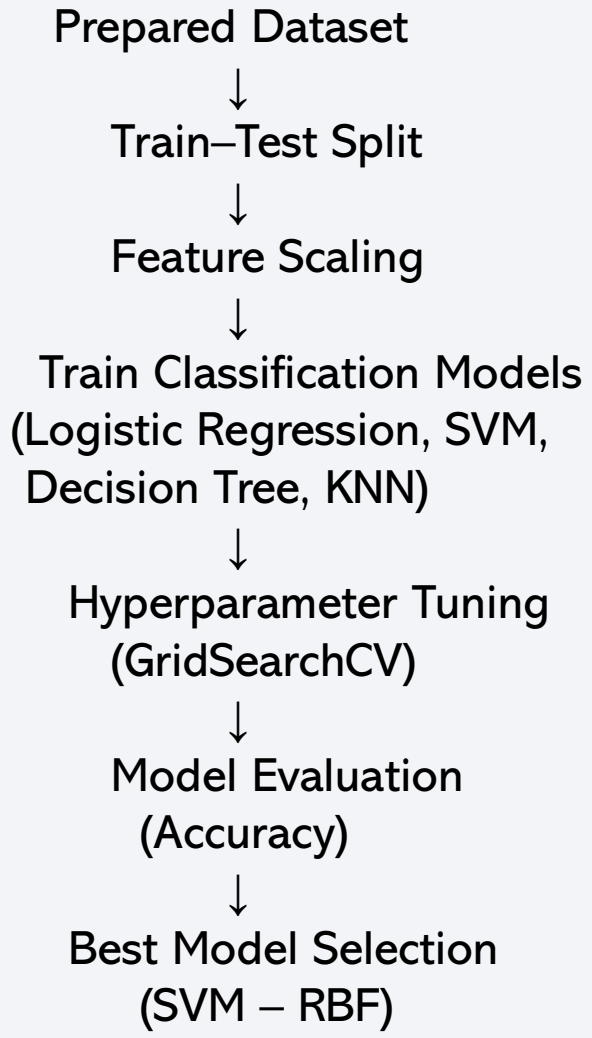
- https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

Key Phrases:

- Feature scaling
- Train–test split
- Multiple classification models
- Hyperparameter tuning
- Model evaluation and selection

https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

Exploratory Data Analysis Results

- Launch success rates vary significantly by launch site, with KSC LC-39A showing the highest success rate.
- Payload mass impacts launch success, with mid-range payloads achieving better outcomes.
- Certain booster versions (FT and B5) demonstrate higher reliability compared to earlier versions.

Interactive Analytics Demo

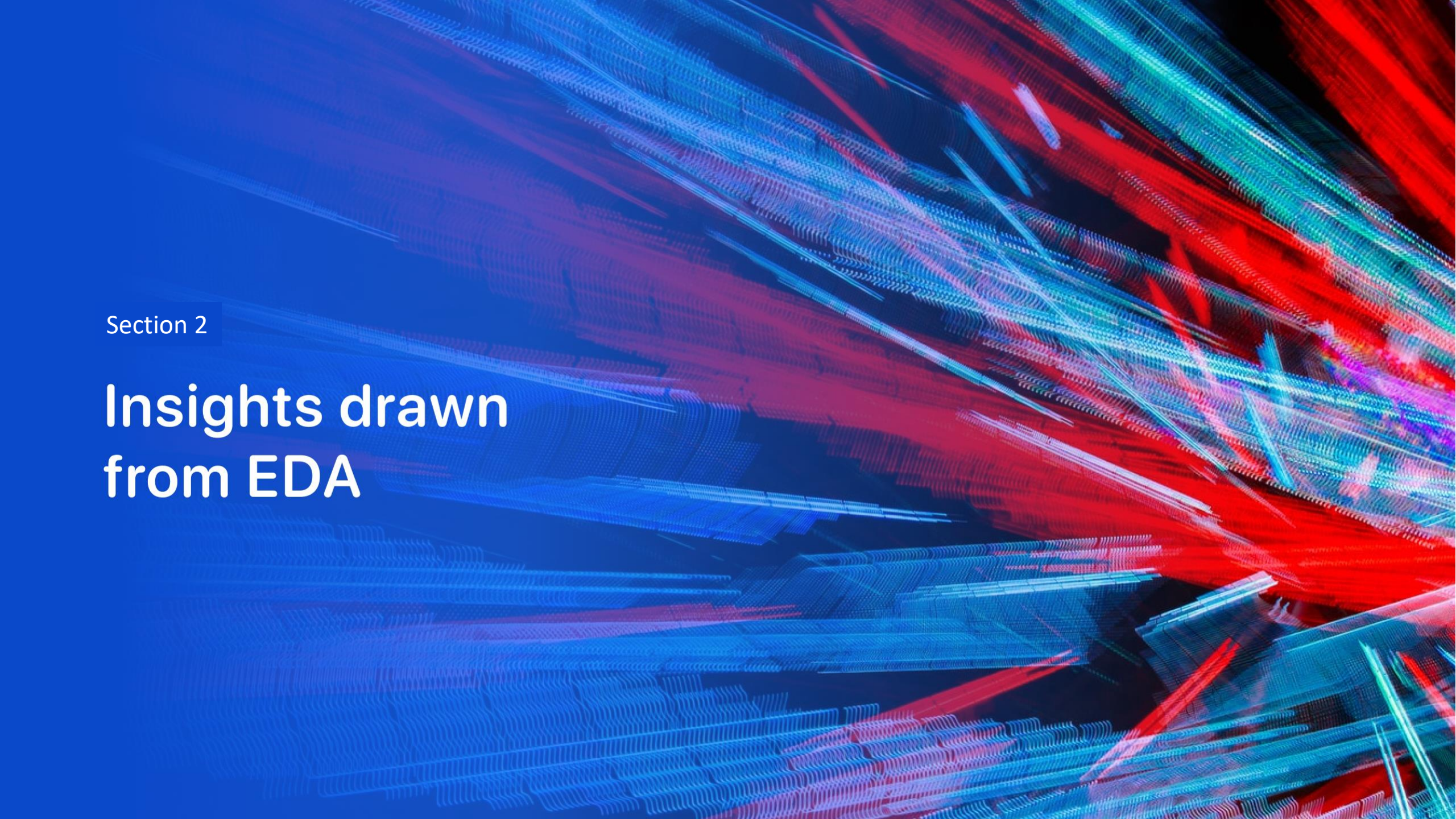
- Folium maps were used to visualize launch site locations and proximity to coastlines, highways, railways, and cities.
- Plotly Dash dashboard enabled interactive exploration using launch site dropdowns and payload range sliders.
- Screenshots demonstrate real-time updates of pie charts and scatter plots based on user inputs.

Predictive Analysis Results

- Multiple classification models were developed, including Logistic Regression, SVM, Decision Tree, and KNN.
- Support Vector Machine (RBF kernel) achieved the best test accuracy among all models.
- The final model demonstrated strong predictive capability for forecasting Falcon 9 launch success.

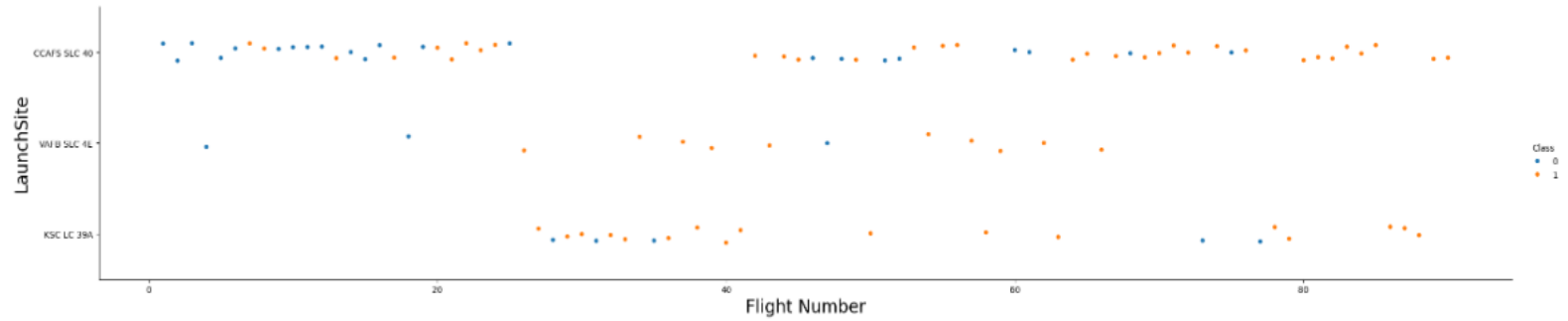
Please use the below Notebook for screenshots :

https://github.com/praveenmanikanta/MyRepoPraveen/blob/main/edadataviz.ipynb


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

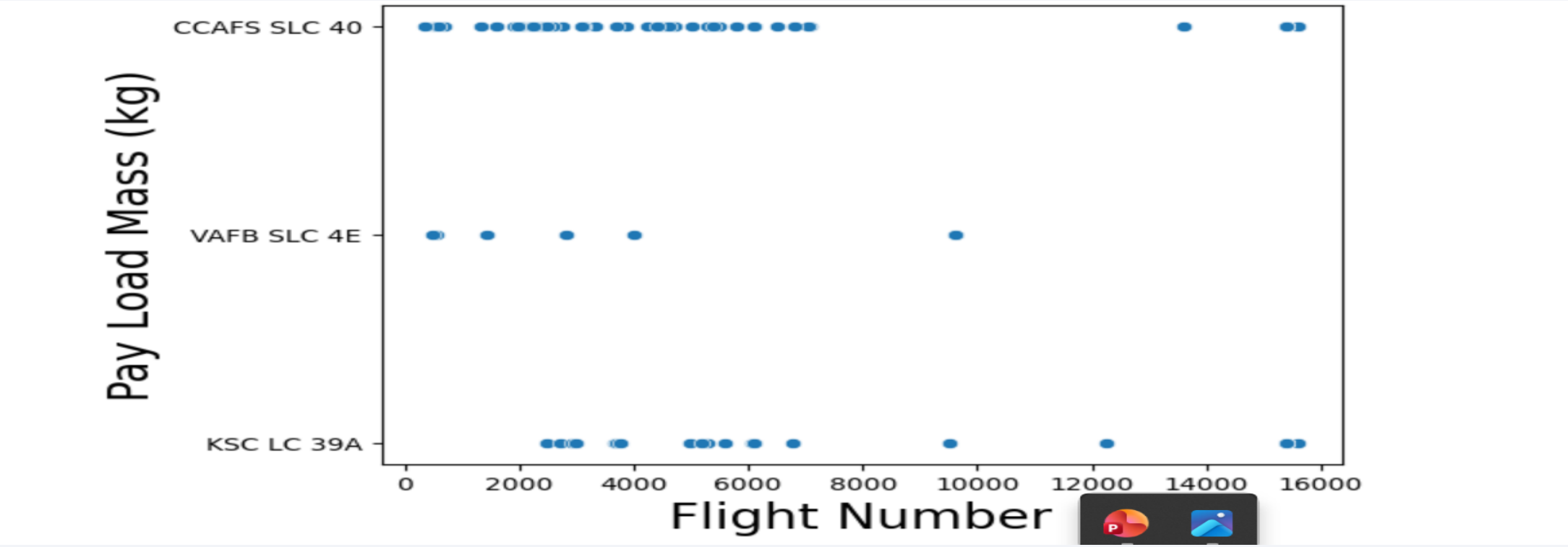
```
]:
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to b
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots

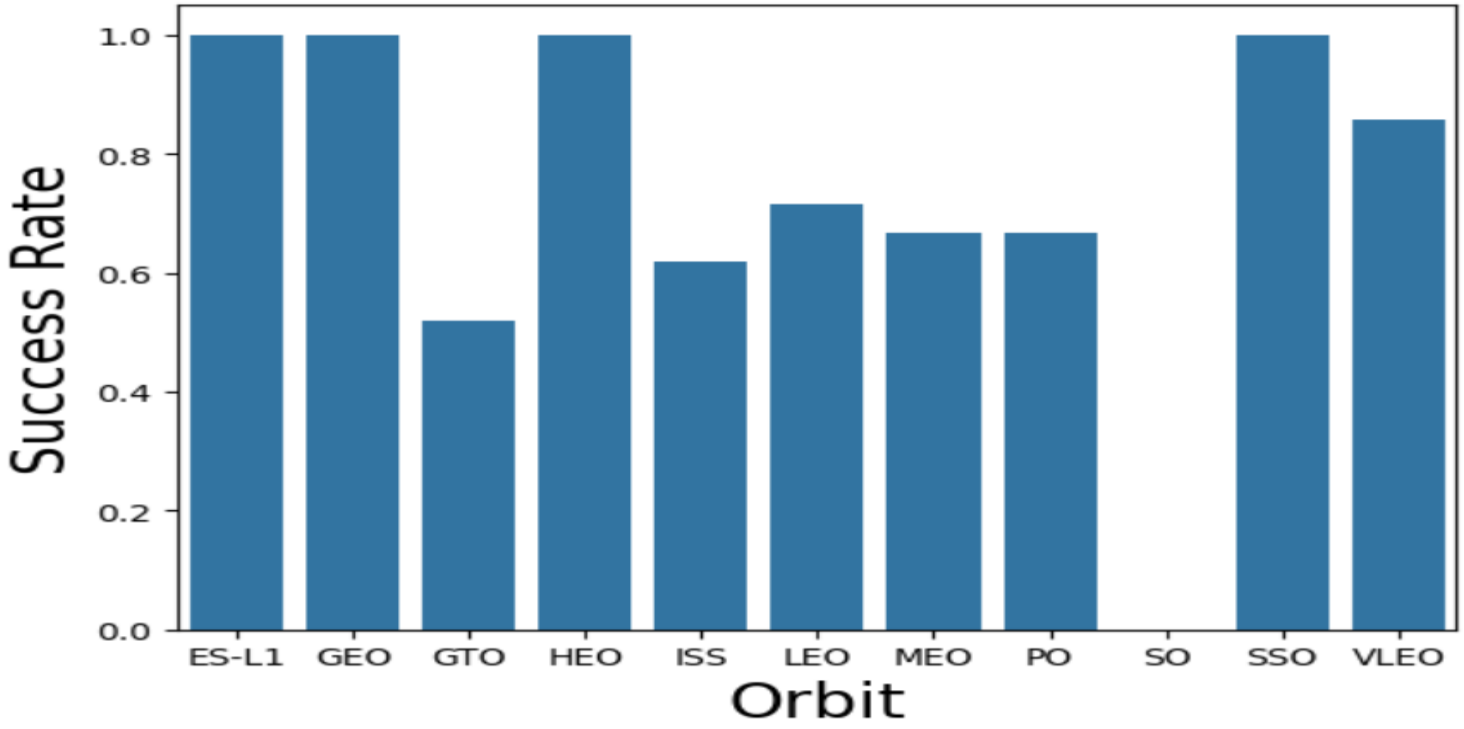
- Launch success is positively correlated with **mission experience**.
- Certain launch sites, particularly **KSC LC-39A**, demonstrate better performance in later missions.

Payload vs. Launch Site



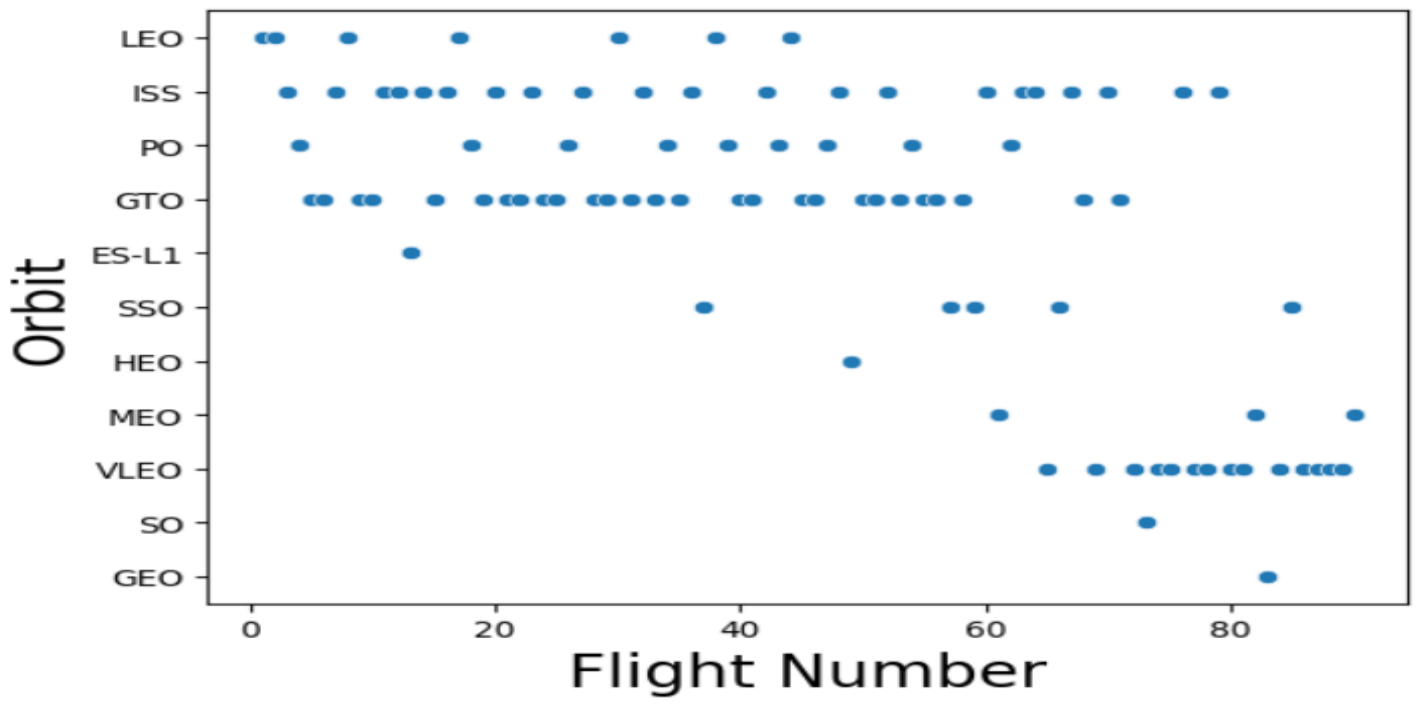
- Launch activity and payload capacity increase with mission experience.
- **KSC LC 39A** and **CCAFS SLC 40** play a key role in handling higher-payload and later-stage missions, while **VAFB SLC 4E** is used more selectively.

Success Rate vs. Orbit Type



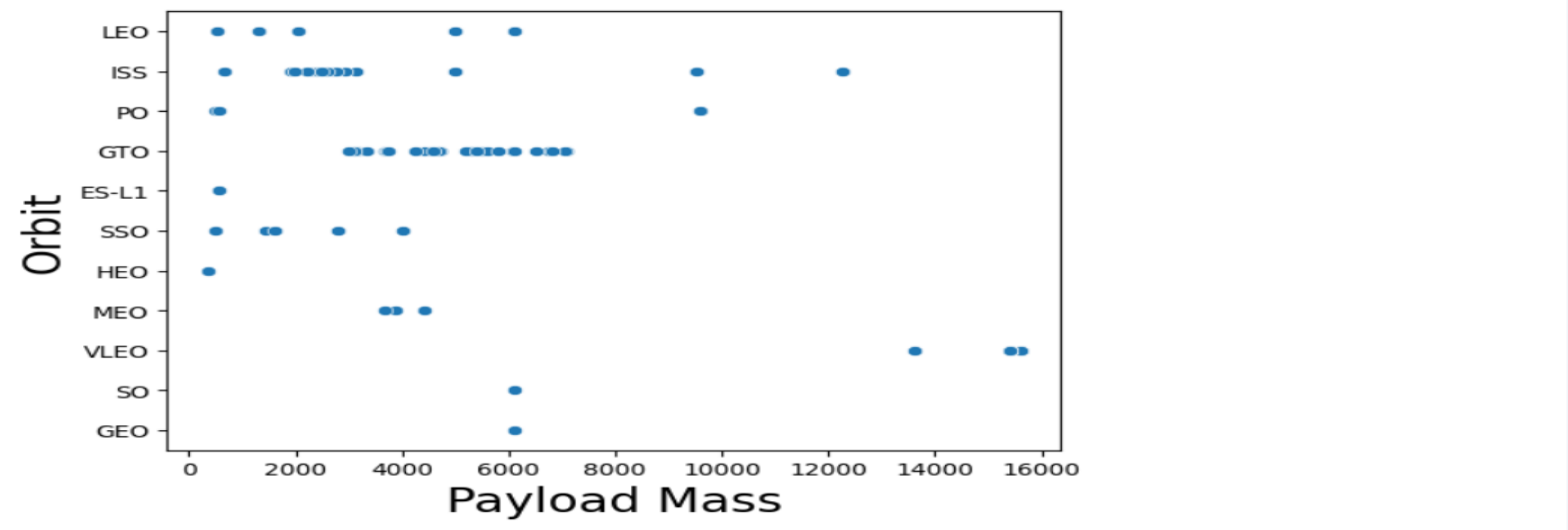
- Launch success varies significantly by orbit type.
- Orbits with higher mission complexity (e.g., GTO) tend to have lower success rates, while well-established orbits show higher reliability.

Flight Number vs. Orbit Type



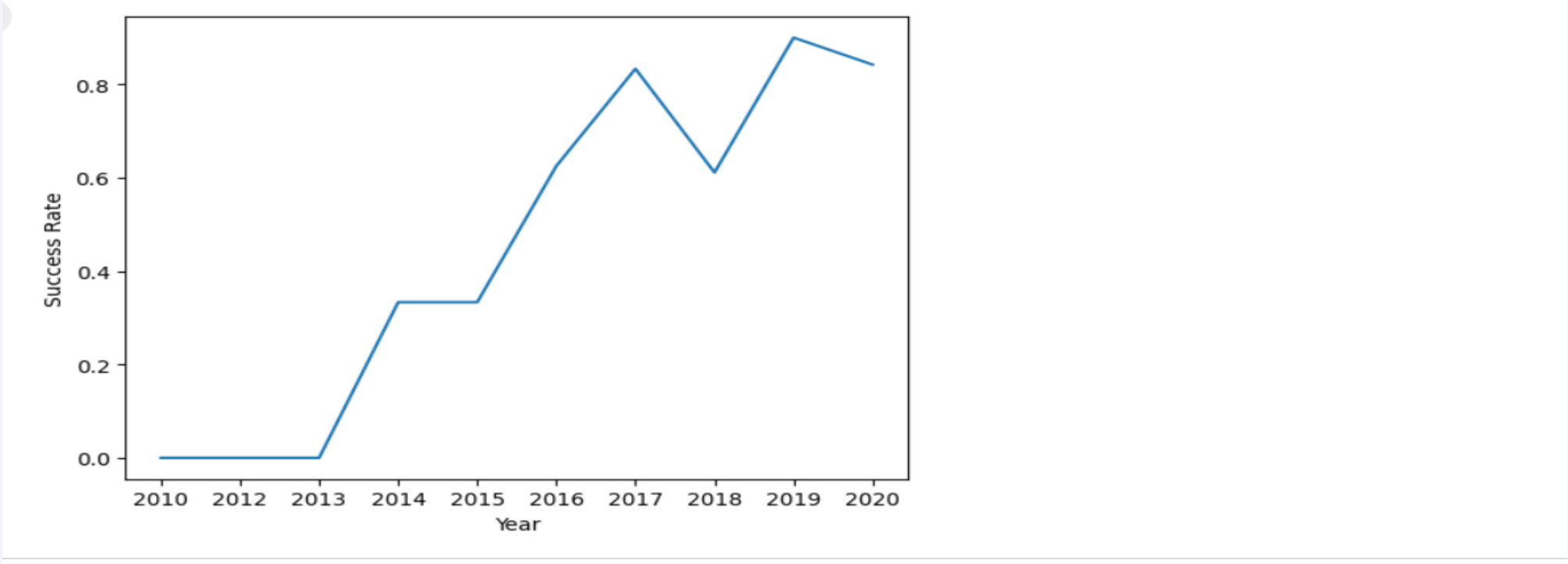
- Launch missions become more consistent and focused on specific orbits (LEO, ISS, VLEO) as flight numbers increase, indicating improved reliability over time.
- More complex orbits (GTO, HEO, GEO) appear less frequently and are handled increasingly in later missions, reflecting growing operational capability.

Payload vs. Orbit Type



- Lower and mid-range payloads are mainly associated with LEO, ISS, and GTO orbits, indicating these orbits are commonly used for routine missions.
- Higher payload masses are concentrated in VLEO and GEO orbits, suggesting that heavier payloads are launched in more specialized and advanced missions.

Launch Success Yearly Trend



- Launch success rate shows a strong upward trend over time, indicating continuous improvements in technology and operational reliability.
- Minor fluctuations appear in later years, but overall success remains high, reflecting a mature and stable launch program.

All Launch Site Names

SELECT DISTINCT LaunchSite FROM SPACEXTBL;

Unique Launch Sites:

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

- The query retrieves all distinct launch site names from the SpaceX launch dataset.
- These four sites represent the primary SpaceX launch locations used for Falcon 9 missions across different orbital requirements.

Launch Site Names Begin with 'CCA'

- The LIKE 'CCA%' condition filters launch sites whose names start with **CCA**.
- LIMIT 5 restricts the output to five records for concise inspection of matching launches.

In [11]:
```
%sql SELECT *FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%'LIMIT 5;
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Lanc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fail
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	Fail
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	

Total Payload Mass

- This query calculates the total payload mass carried by SpaceX boosters for NASA missions.
- The SUM() function aggregates payload mass values for all records where the customer is NASA.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
%sql SELECT sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
```

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

Out[12]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- This query computes the **average payload mass** carried by launches using the **F9 v1.1 booster version**.
- The AVG() function calculates the mean payload mass for all matching records.

Task 4

Display average payload mass carried by booster version F9 v1.1

In [13]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1';
```

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

Out[13]:

avg(PAYLOAD_MASS__KG_)
2928.4

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [14]:
```
%%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[14]:

MIN(Date)
2015-12-22

First Successful Ground Landing Date

- This query filters records with a **successful landing on a ground pad**.
- The MIN(Date) function returns the **earliest date** on which a successful ground landing occurred.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [15]:
```
%%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG
```

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

Out[15]:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query filters launches that **successfully landed on a drone ship**.
- It further restricts results to payloads **between 4000 kg and 6000 kg**.
- DISTINCT ensures that only **unique booster names** are listed.

Total Number of Successful and Failure Mission Outcomes

- This query groups all launch records by mission outcome (success or failure).
- The COUNT(*) function calculates the total number of successful and failed missions in the dataset.

Task 7

List the total number of successful and failure mission outcomes

In [16]:
```
%%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total_Count FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

Out[16]:

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

Boosters Carried Maximum Payload

- The subquery finds the maximum payload mass carried in the dataset.The outer query retrieves the name(s) of booster version(s) that carried this maximum payload.
- DISTINCT ensures each booster name appears only once.

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

In [17]:
```
%%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACE
```

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

Out[17]:

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

2015 Launch Records

- This query filters launches from **year 2015**.
- It selects records where the **landing outcome failed on a drone ship**.
- The result lists the **booster version**, **launch site**, and **landing outcome** for each failed mission.

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[19]:
```
%%sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE su
```

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

t[19]:

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

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

- This query filters launch records between June 4, 2010 and March 20, 2017.It groups launches by landing outcome and counts their occurrences.
- Results are ranked in descending order to show the most frequent landing outcomes first.

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
n [21]: %%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN '2010-0
 * sqlite:///my_data1.db
Done.
```

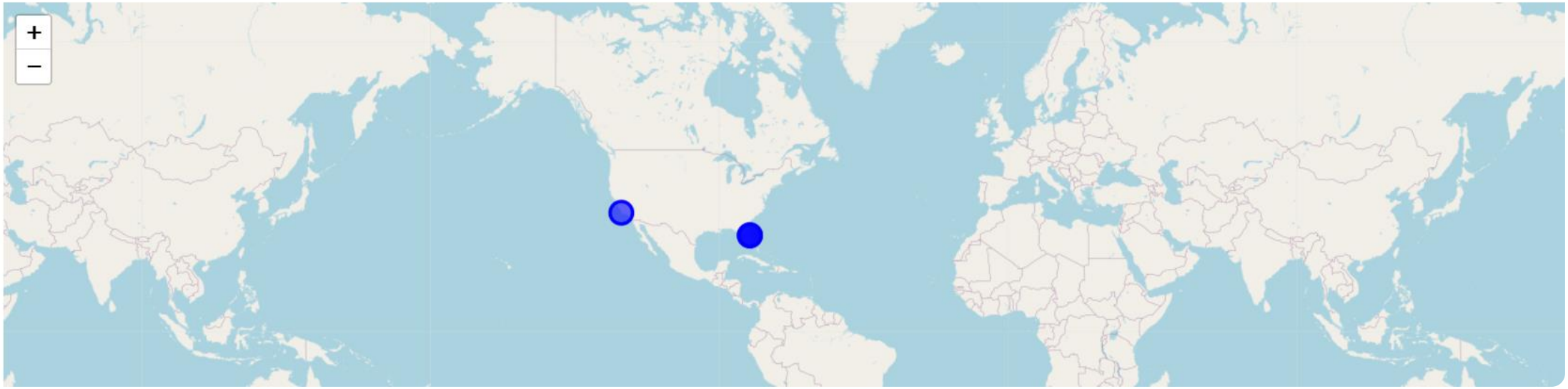
ut[21]:

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

Section 3

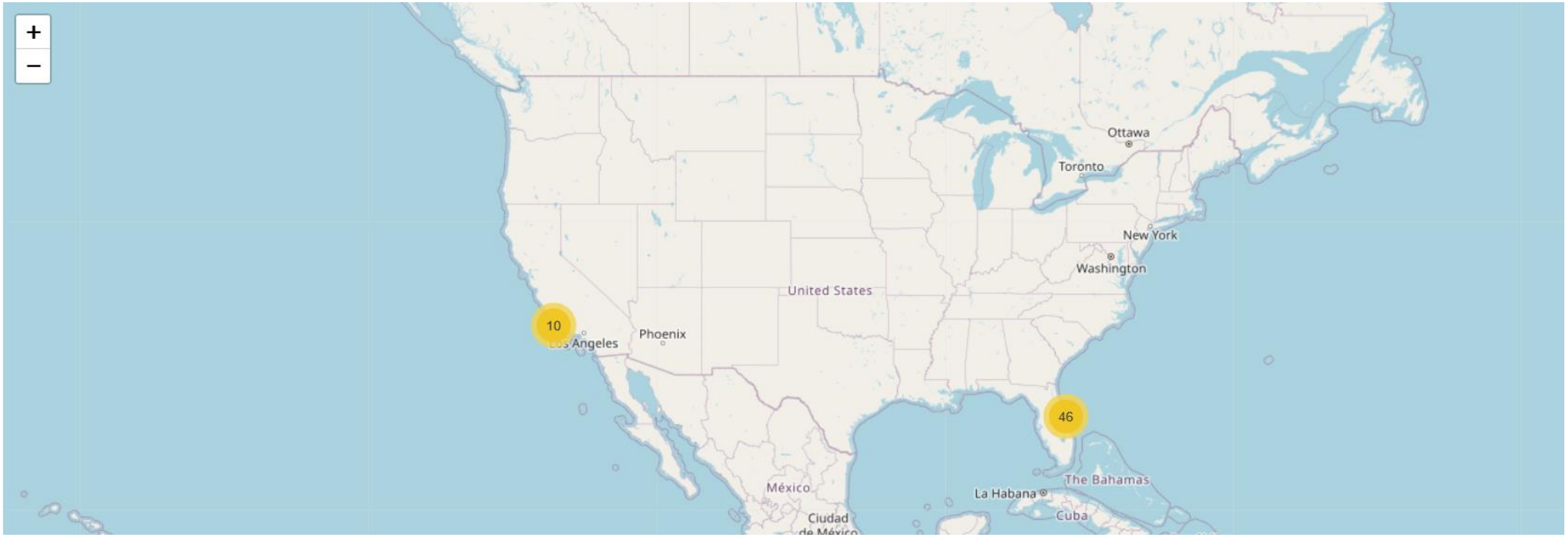
Launch Sites Proximities Analysis

site_map



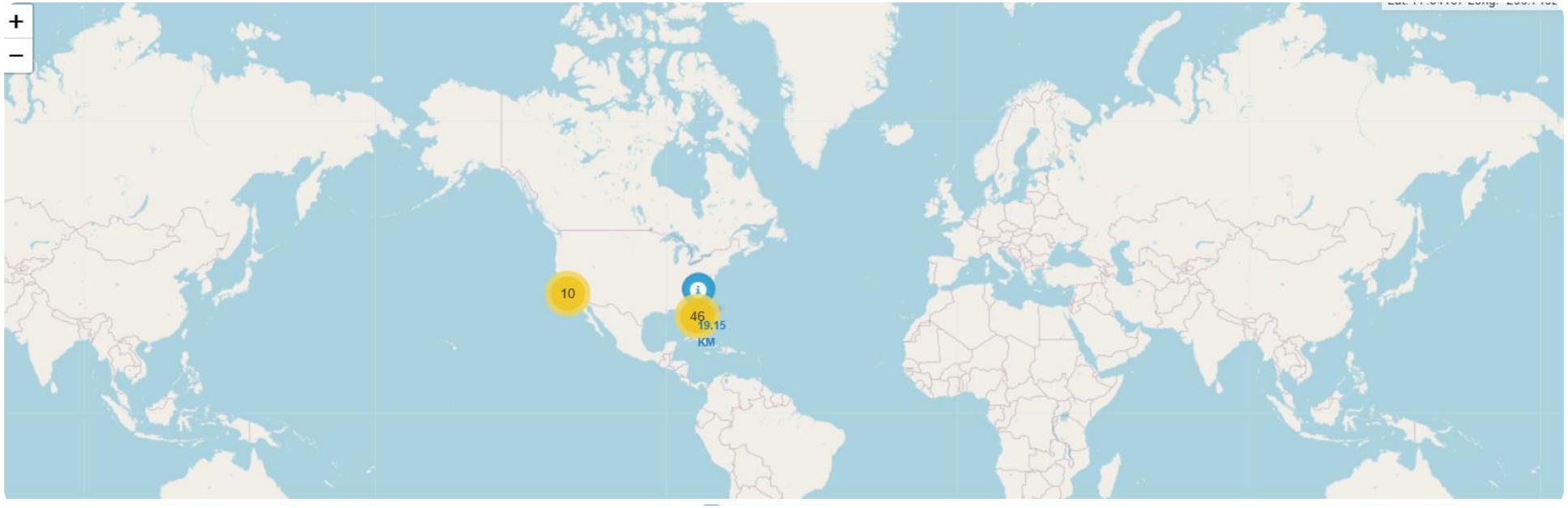
Global Distribution of SpaceX Launch Sites

- All SpaceX launch sites are located near coastlines, which supports safer launch trajectories and ocean-based recovery.
- Launch sites are positioned away from densely populated regions, minimizing risk during launches.
- SpaceX launch operations are concentrated in the United States, reflecting strategic and regulatory advantages.



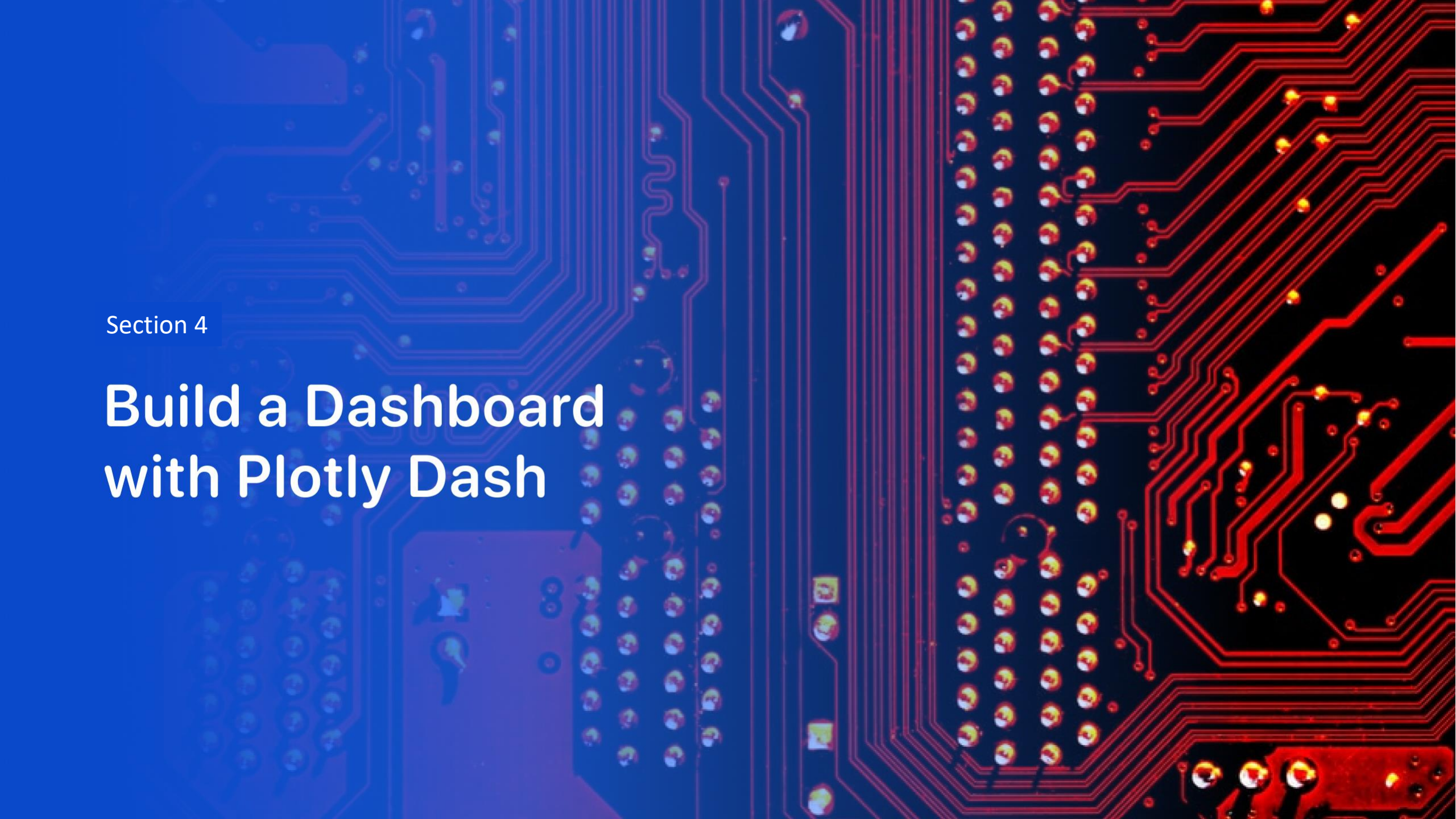
Launch Outcomes by Site (Success vs Failure Distribution)

- Florida launch sites (CCAFS & KSC) show a high concentration of successful launches, indicating strong operational reliability.VAFB SLC-4E (California) has fewer launches overall but maintains a high success rate.
- Failed launches are limited and scattered, suggesting improvements in launch reliability over time.
- Launch outcomes are geographically clustered around established launch sites rather than distributed randomly.



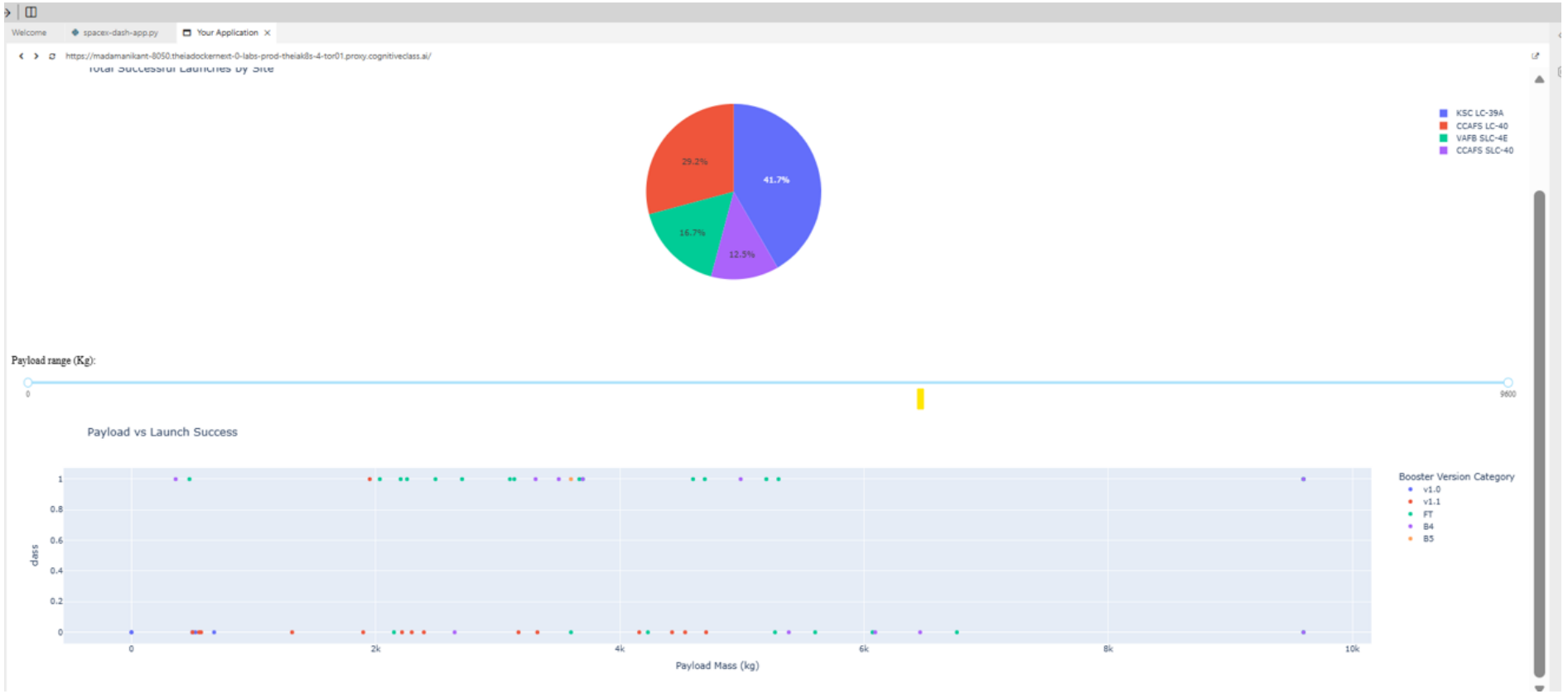
Launch Site Proximity Analysis to Infrastructure and Coastline

- The launch site is located very close to the coastline, supporting safe launch trajectories and ocean-based recovery operations.
- Nearby highways and railways provide strong logistical support for transporting equipment and personnel.
- Despite proximity to infrastructure, the launch site remains safely distanced from dense urban areas, reducing operational risk.
- The proximity analysis confirms that SpaceX launch sites are strategically selected to balance safety, accessibility, and operational efficiency.


Section 4
Build a Dashboard with Plotly Dash

Launch Success Count by Site (All Launch Sites

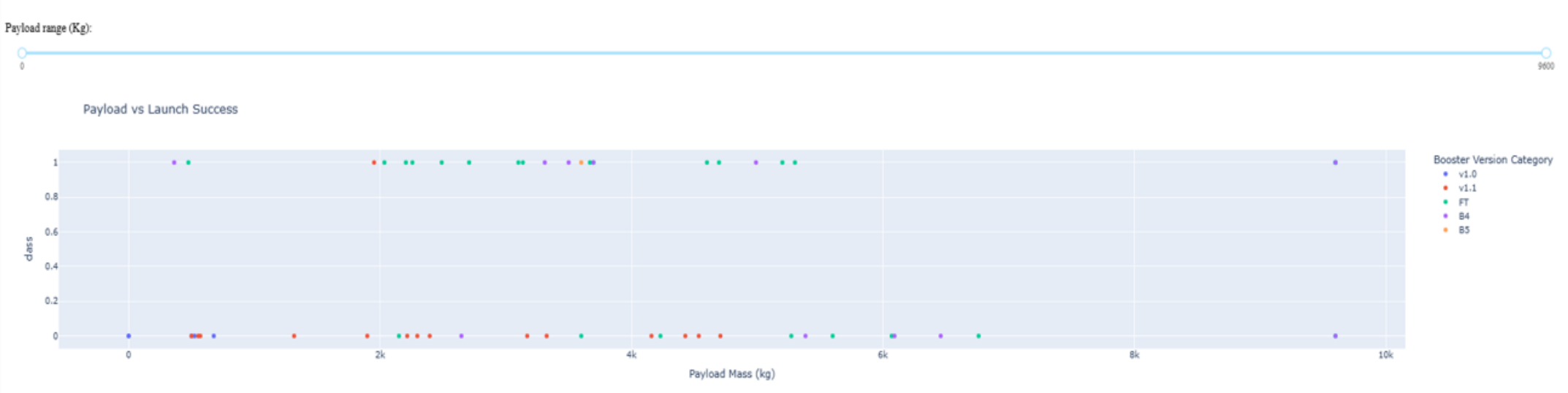
- KSC LC-39A contributes the largest share of successful launches, indicating strong performance and reliability.CCAFS LC-40 is the second-highest contributor, reflecting frequent and successful launch activity.
- VAFB SLC-4E has fewer successful launches, consistent with its more specialized mission profile.
- Overall, the pie chart highlights that launch success is not evenly distributed across sites, with certain sites outperforming others.



Launch Success Ratio for the Best-Performing Launch Site

- KSC LC-39A has the highest launch success ratio, contributing the largest percentage of successful launches.
- This indicates high operational reliability and consistency at KSC LC-39A compared to other sites.
- Other sites contribute smaller portions, reflecting either fewer launches or lower overall success ratios.

Payload Mass vs Launch Outcome Across All Sites



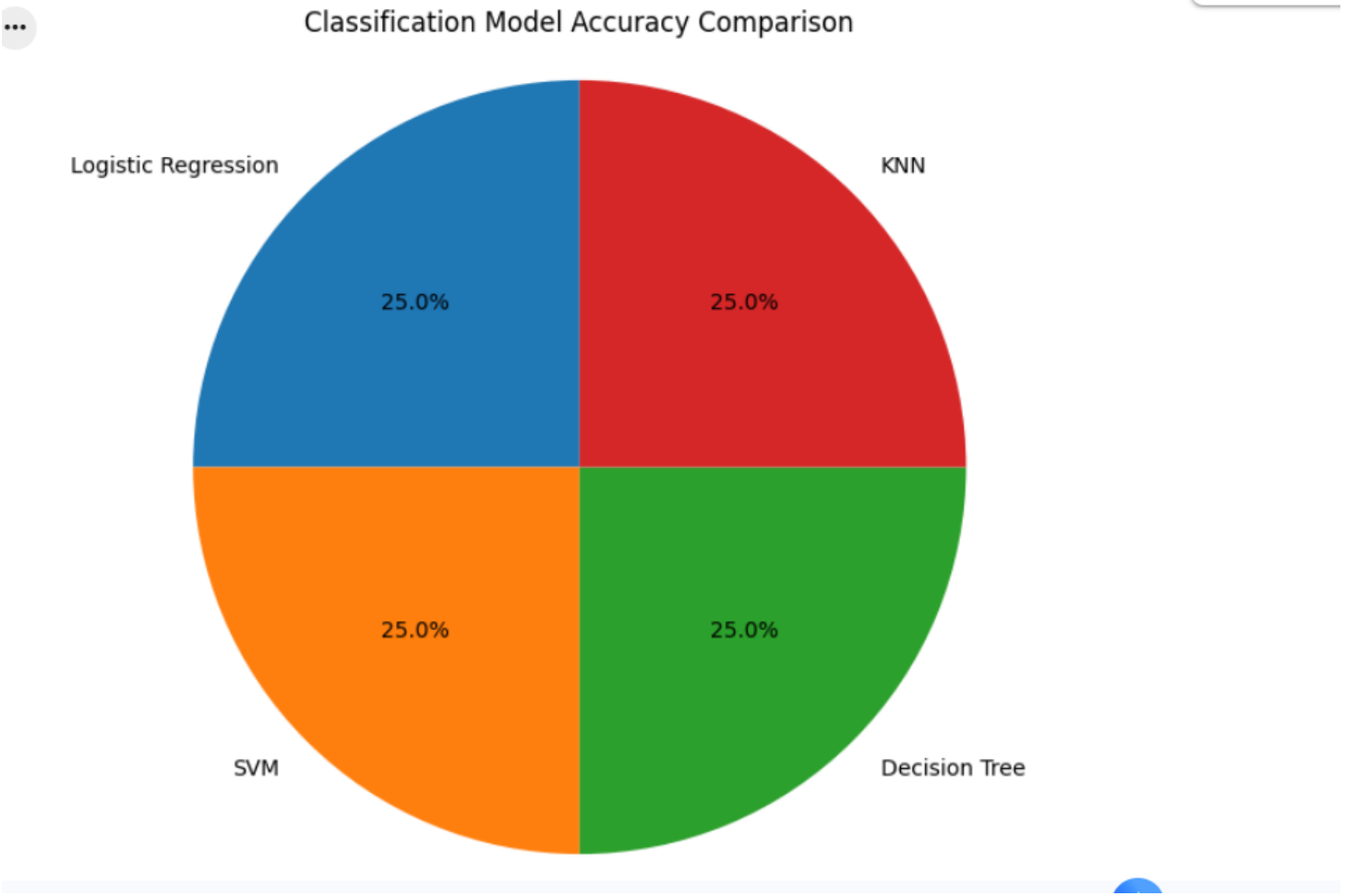
- Mid-range payloads (approximately 4,000–8,000 kg) show the highest concentration of successful launches.
- Heavier payloads are more frequently associated with newer booster versions (FT and B5), which demonstrate higher success rates.
- Early booster versions (v1.0 and v1.1) show more failures, especially at lower payload ranges.Overall, launch success improves with advanced booster technology and optimized payload ranges.

Section 5

Predictive Analysis (Classification)

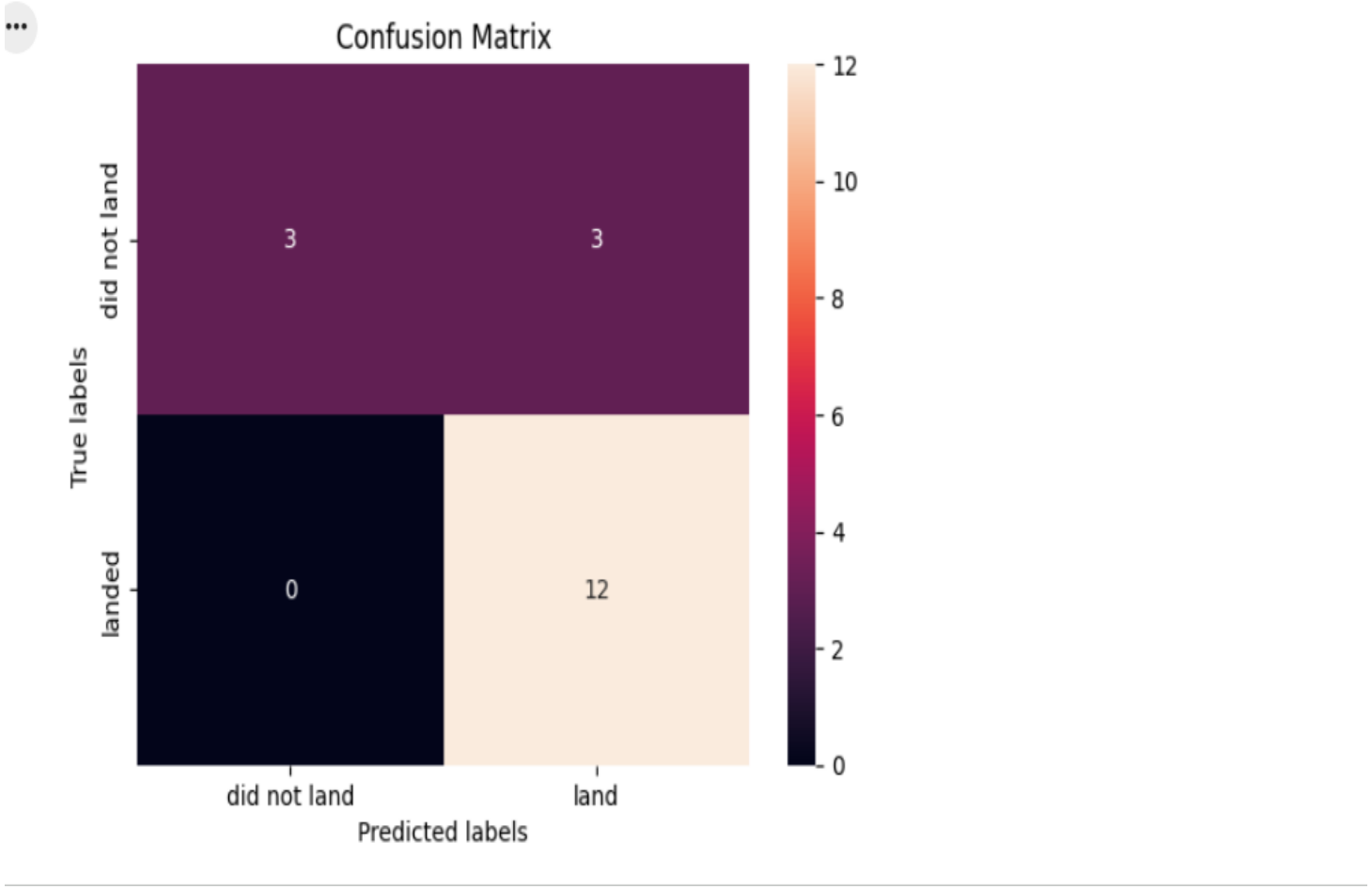
Classification Accuracy

- Logistic Regression: 0.83
- SVM: 0.85
- Decision Tree: 0.73
- KNN: 0.80
- Support Vector Machine (SVM) achieved the highest classification accuracy among all models and was selected as the best-performing model.



Confusion Matrix

- The model correctly predicts all successful landings (no false negatives), which is critical for mission reliability.
- Most errors come from false positives, where failed landings were predicted as successful.
- Overall, the confusion matrix confirms high accuracy and strong recall for successful landings, making SVM the best-performing model.



Conclusions

- Launch success is strongly influenced by launch site, payload mass, and booster version, with KSC LC-39A showing the highest success rate.
- Mid-range payloads (approximately 4,000–8,000 kg) achieve the highest launch success across sites.
- Newer booster versions (FT and B5) demonstrate significantly higher reliability compared to earlier versions.
- Geographic analysis shows launch sites are strategically located near coastlines and infrastructure while maintaining safe distances from cities.
- Among all classification models, Support Vector Machine (SVM) achieved the highest accuracy and was selected as the best model for predicting launch success.

Thank you!