IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeff Hawkins
11/26/2024

Outline

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

Executive Summary

Overview of Methods:

- Collecting and preparing data
- Cleaning and organizing datasets
- Exploring data through visualizations
- Performing SQL-based data analysis
- Creating interactive maps with Folium
- Developing dashboards with Plotly Dash
- Building predictive models using classification techniques

Key Outcomes:

- nsights from exploratory data analysis
- Interactive analytics demonstrated through visuals
- Results of predictive modeling efforts

Introduction

Project Overview:

- SpaceX is a leader in the commercial space industry, revolutionizing space travel by making it more affordable. Their Falcon 9 rocket launches are listed at $62 million, significantly less than the $165 million charged by other providers. This cost advantage stems from SpaceX's ability to reuse the rocket's first stage. By predicting whether the first stage will successfully land, we can estimate the cost of a launch. Using publicly available data and machine learning models, this project aims to determine the likelihood of SpaceX reusing the first stage.

Key Questions:

- How do factors like payload mass, launch site, number of flights, and orbit type influence the success of first-stage landings?
- Is the rate of successful landings improving over time?
- Which binary classification algorithm performs best for this prediction?

Section 1

Methodology

Methodology

Executive Summary

Data Collection:

- Leveraged the SpaceX REST API
- Extracted additional information using web scraping from Wikipedia

Data Wrangling:

- Applied data filtering techniques
- Handled missing values effectively
- Used One-Hot Encoding to prepare data for binary classification

Exploratory Data Analysis (EDA):

- Conducted EDA using visualizations and SQL

Interactive Visual Analytics:

- Built interactive visualizations with Folium and Plotly Dash

Predictive Analysis:

- Developed, fine-tuned, and evaluated classification models to achieve optimal results

Data Collection

The data collection process utilized a combination of API requests from the SpaceX REST API and web scraping from a Wikipedia table about SpaceX launches. Both methods were essential to gather a comprehensive dataset for detailed analysis.

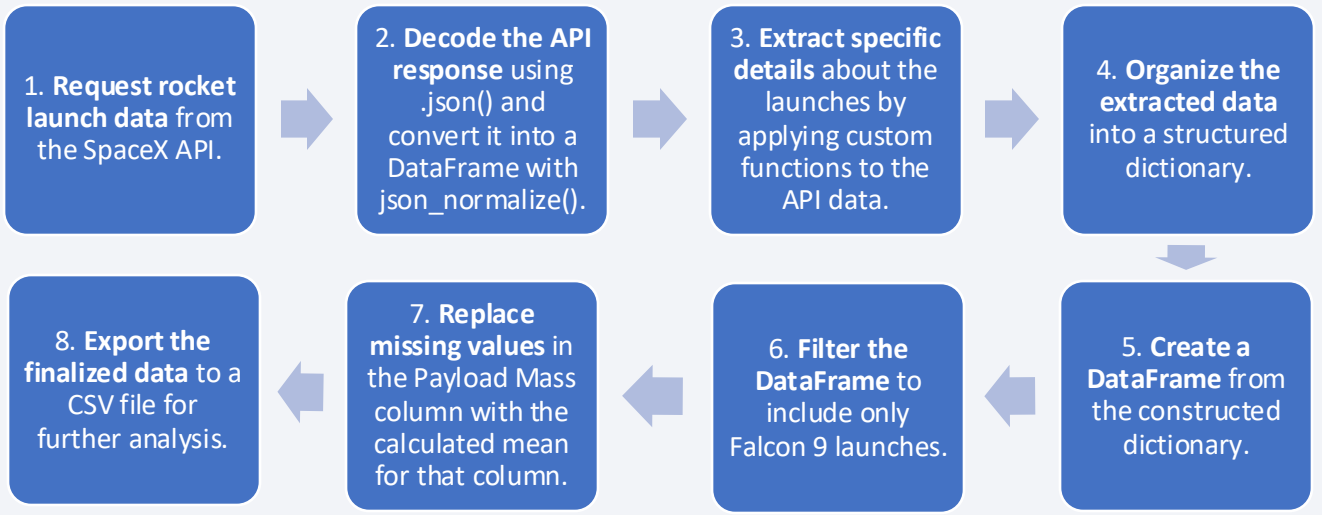
Data Columns Retrieved via SpaceX REST API:

- Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome
- Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

Data Columns Retrieved via Wikipedia Web Scraping:

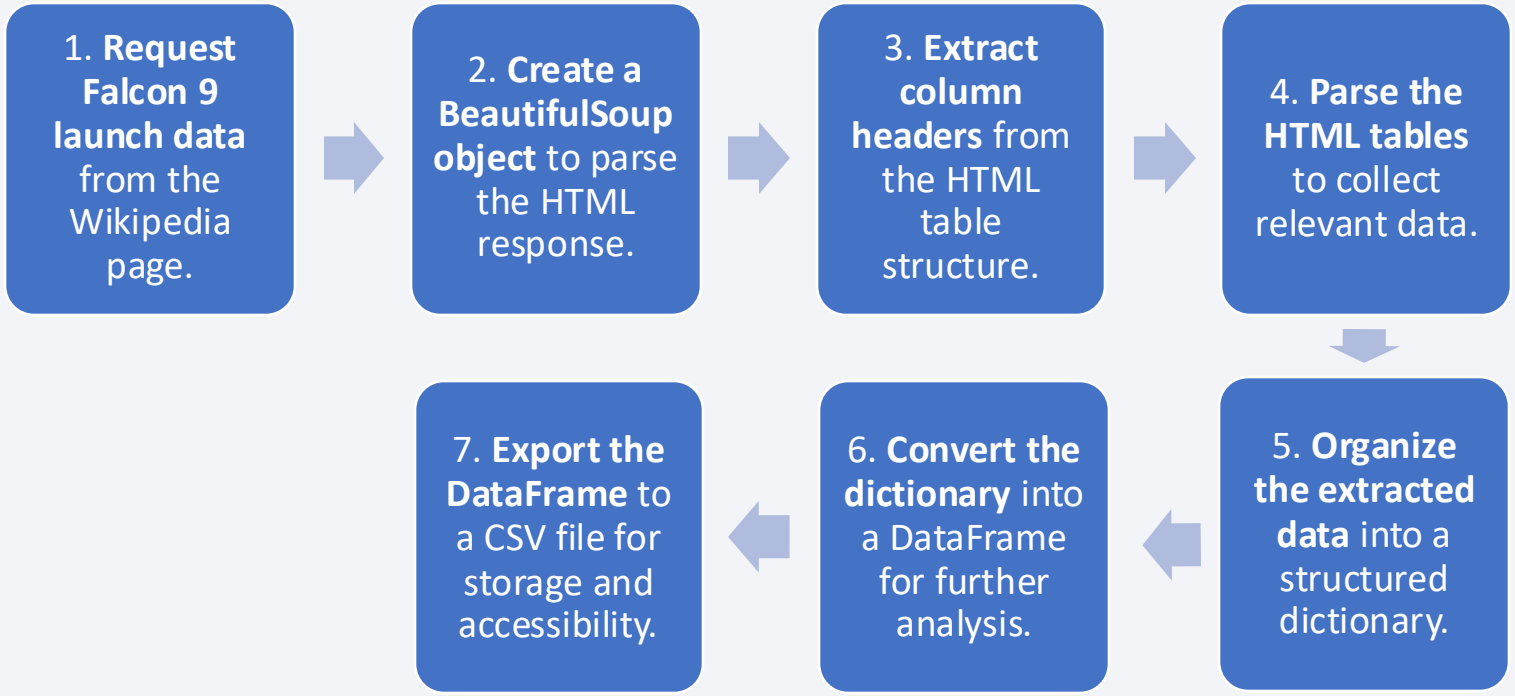
- Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome
- Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API



Data Collection GitHub

Data Collection - Scraping



Data Collection GitHub

Data Wrangling

- The dataset includes various scenarios where the booster did not land successfully. For instance:
- • **True Ocean:** The booster successfully landed in a specific ocean region.
- • **False Ocean:** The booster attempted to land in the ocean but failed.
- • **True RTLS (Return to Launch Site):** The booster successfully landed on a ground pad.
- • **False RTLS:** The booster attempted to land on a ground pad but failed.
- • **True ASDS (Autonomous Spaceport Drone Ship):** The booster successfully landed on a drone ship.
- • **False ASDS:** The booster attempted to land on a drone ship but failed.
- These outcomes were converted into training labels:
- • **1:** Booster successfully landed.
- • **0:** Booster landing was unsuccessful.

1. Perform exploratory data analysis to define training labels.
2. Calculate the number of launches for each site.
3. Determine the count and frequency of each orbit type.
4. Analyze mission outcomes by orbit type.
5. Create a landing outcome label from the “Outcome” column.
6. Export the processed data to a CSV file.

Data Wrangling GitHub

EDA with Data Visualization

Visualizations Created:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Success Rate Yearly Trends

Chart Types and Their Purpose:

- **Scatter Plots:** Highlight relationships between variables, which can be leveraged in machine learning models.
- **Bar Charts:** Compare discrete categories, illustrating the relationships between specific groups and their values.
- **Line Charts:** Showcase trends over time (time series data).

Data Visualization GitHub

EDA with SQL

SQL Queries Performed:

- Retrieve the unique launch site names from the dataset.
- Display five records where the launch site starts with “CCA”.
- Calculate the total payload mass carried by boosters launched by NASA (CRS).
- Find the average payload mass carried by booster version F9 v1.1.
- Identify the date of the first successful landing on a ground pad.
- List boosters that successfully landed on a drone ship with payload mass between 4000 and 6000.
- Count the total number of successful and failed mission outcomes.
- Find the booster versions that carried the maximum payload mass.
- List failed landing outcomes on drone ships, including booster versions and launch sites, for 2015.
- Rank landing outcomes (e.g., failures on drone ships or successes on ground pads) between 2010-06-04 and 2017-03-20 in descending order.

Data SQL EDA GitHub

Build an Interactive Map with Folium

Launch Site Markers:

- Added a marker with a circle, popup label, and text label for NASA's Johnson Space Center, using its latitude and longitude as a starting point.
- Plotted markers with circles, popup labels, and text labels for all launch sites to visualize their geographical locations and proximity to the equator and coastlines.

Colored Markers for Launch Outcomes:

- Included green markers for successful launches and red markers for failed launches using Marker Clusters to highlight sites with high success rates.

Distance from Launch Sites to Proximities:

- Added colored lines to illustrate distances between launch sites (e.g., KSC LC-39A) and nearby landmarks such as railways, highways, coastlines, and cities.

Data Folium GitHub

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Implemented a dropdown menu to allow users to select specific launch sites.

Pie Chart for Launch Success Rates (All Sites/Selected Site):

- Added a pie chart to display the total number of successful launches across all sites, along with success vs. failure counts for a selected launch site.

Payload Mass Range Slider:

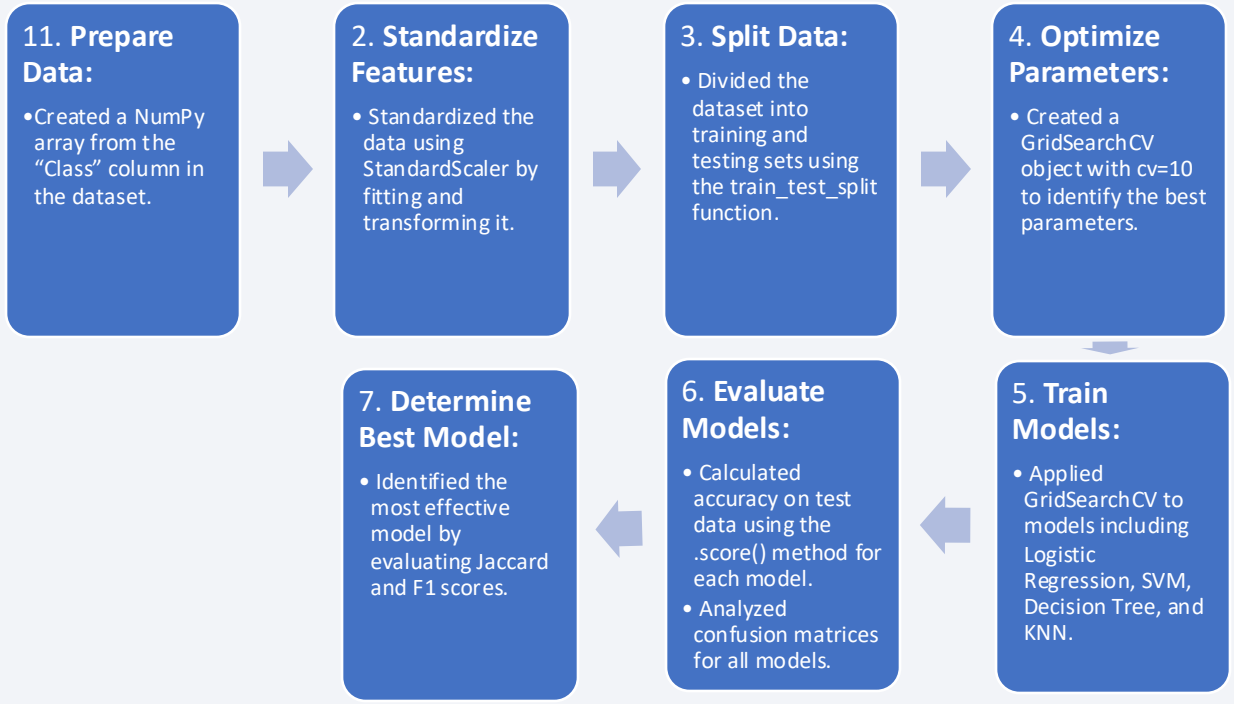
- Included a slider to let users filter data based on a selected payload mass range.

Scatter Plot: Payload Mass vs. Success Rate by Booster Versions:

- Designed a scatter plot to visualize the correlation between payload mass and launch success rates across different booster versions.

Data Plotly GitHub

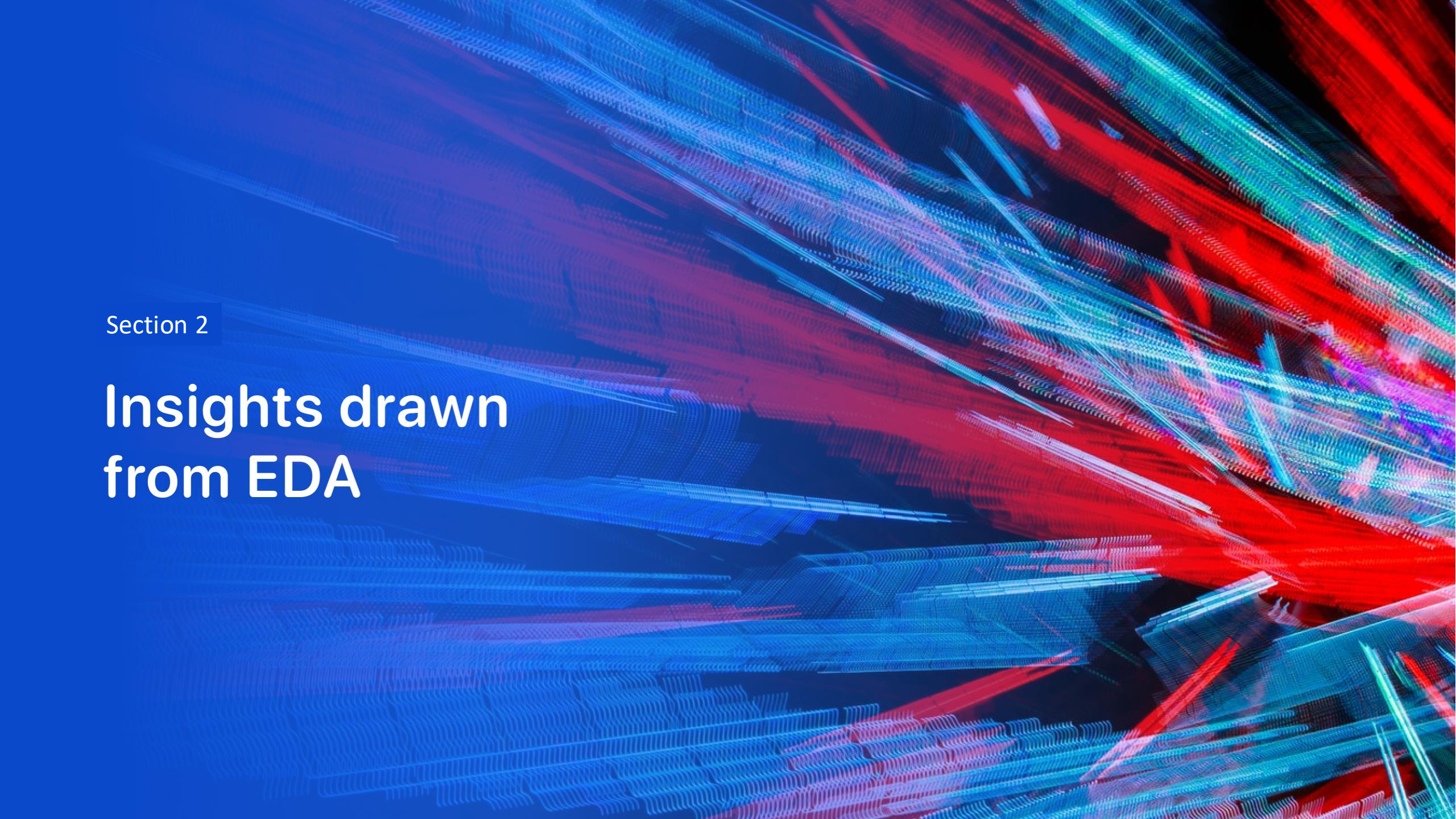
Predictive Analysis (Classification)



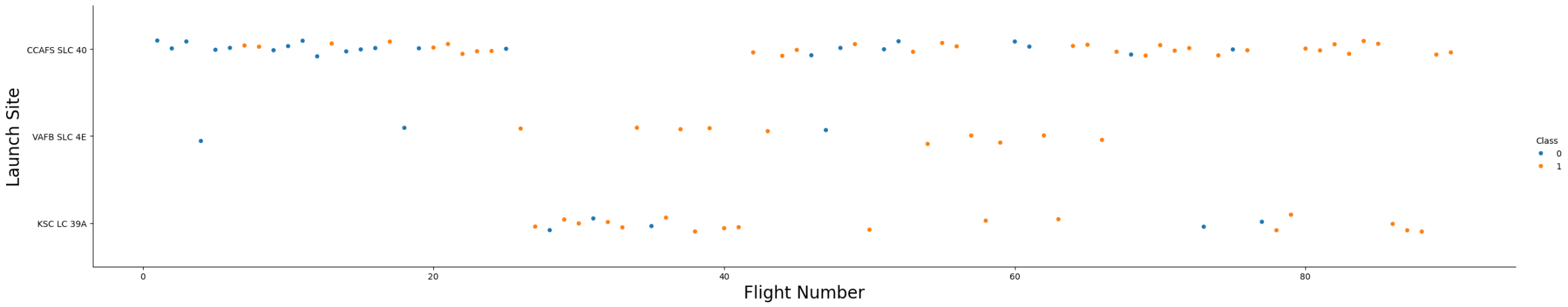
Data Predictions GitHub

Results

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


Section 2
Insights drawn from EDA

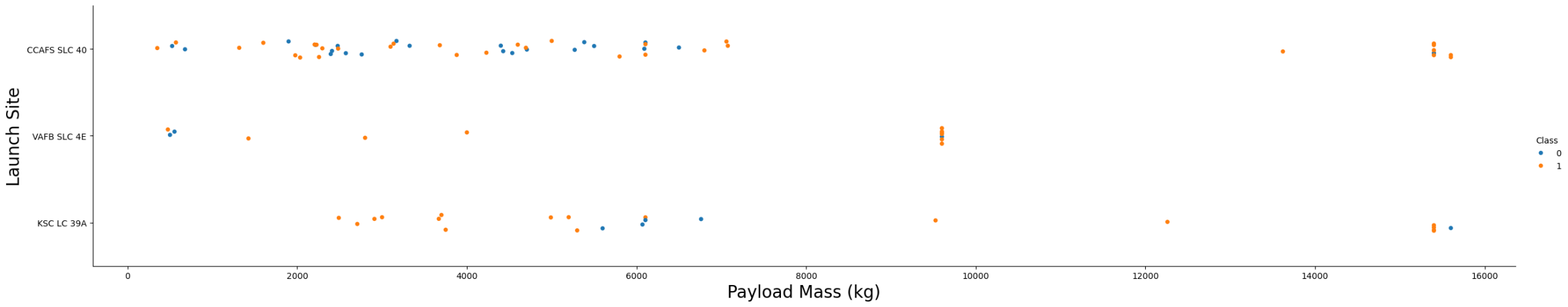
Flight Number vs. Launch Site



Explanation:

- The chart showcases a clear trajectory of improvement in launch success over time, underscoring the iterative development and increasing reliability of the booster versions and operational practices.
- Failures (Class 0) are concentrated in earlier flight numbers and primarily at high-use sites like CCAFS LC-40, which were likely used for testing and development phases.
- Later flights, especially at VAFB SLC-4E, exhibit strong reliability, potentially due to the use of more mature technology or focused mission planning.

Payload vs. Launch Site



Explanation:

• **Payload and Success Trends**: Heavier payloads (above ~10,000 kg) tend to correlate with a higher proportion of successful flights (Class 1), indicating robust capability for large payloads.

• **Launch Site Utilization**: CCAFS LC-40 and KSC LC-39A handle a broader range of payload masses, with successes across varying weights, while VAFB SLC-4E primarily manages lighter payloads, likely due to specialized mission profiles.

• **Failure Patterns**: Failures (Class 0) are more prevalent for payloads under ~6,000 kg, possibly reflecting testing or higher risk associated with smaller-scale missions.

• **Operational Excellence**: The successful handling of both light and heavy payloads at key sites reflects growing reliability and versatility in the launch program over time.

Success Rate vs. Orbit Type

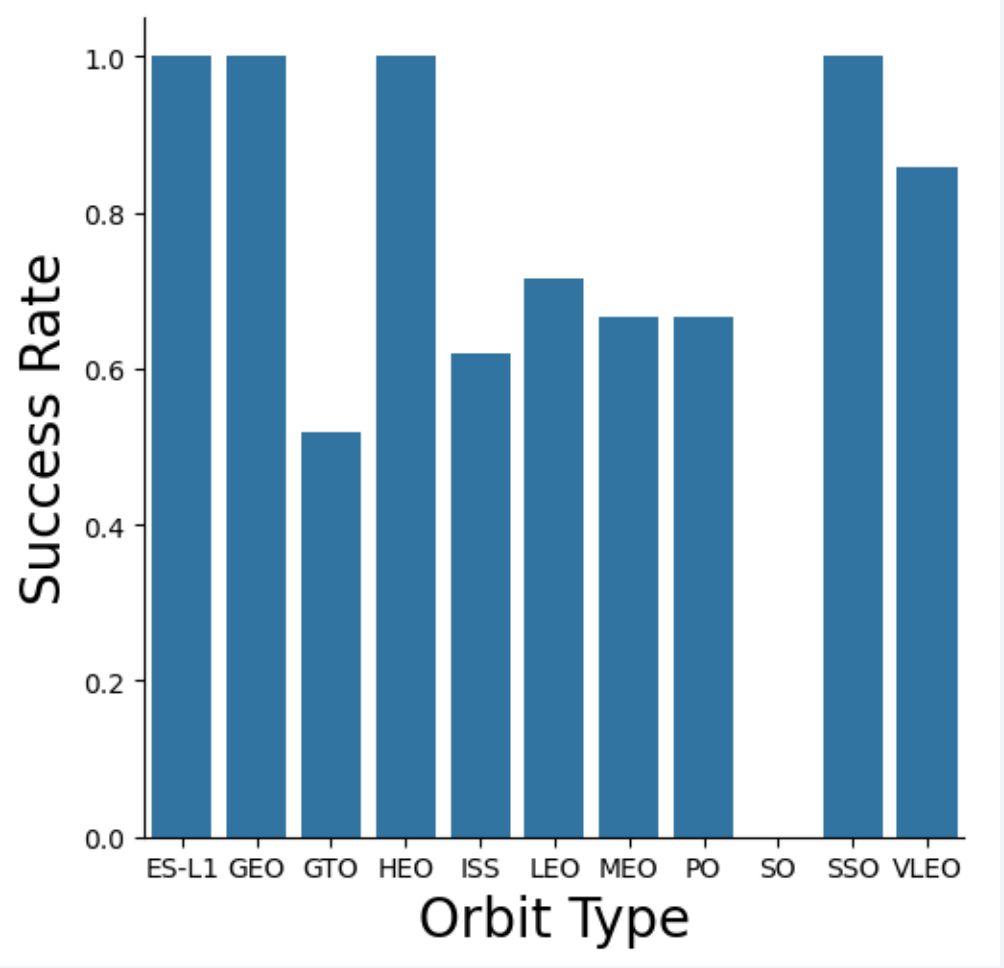
Explanation:

• **High Success Rates**: ES-L1, GEO, and SSO orbits exhibit a 100% success rate, indicating strong reliability for these mission types.

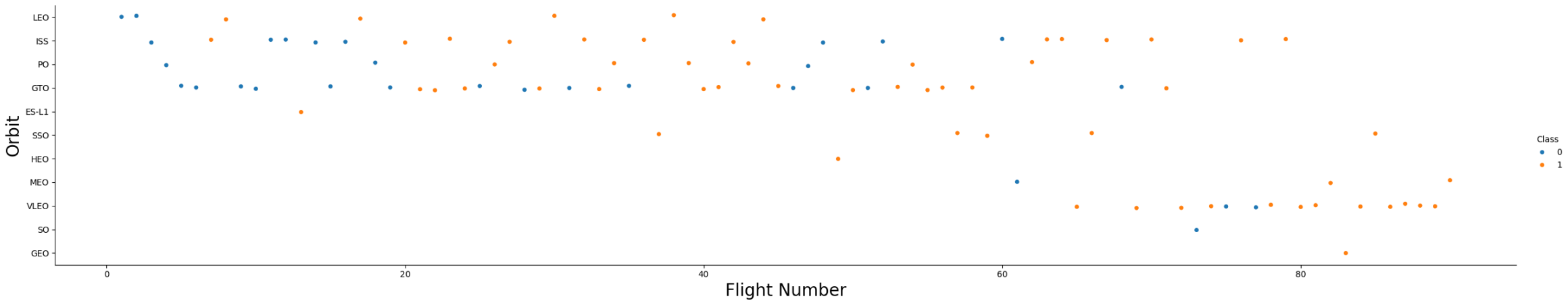
• **Moderate Reliability**: LEO, MEO, and VLEO show success rates around 70%, suggesting these orbits may involve more challenging or experimental missions.

• **Lowest Success Rate**: GTO has the lowest success rate, possibly reflecting higher complexity or risks associated with geostationary transfer missions.

• **Mission Trends**: The consistently high success for certain orbits highlights operational strengths, while areas with lower rates may indicate opportunities for improvement in mission planning or technology.



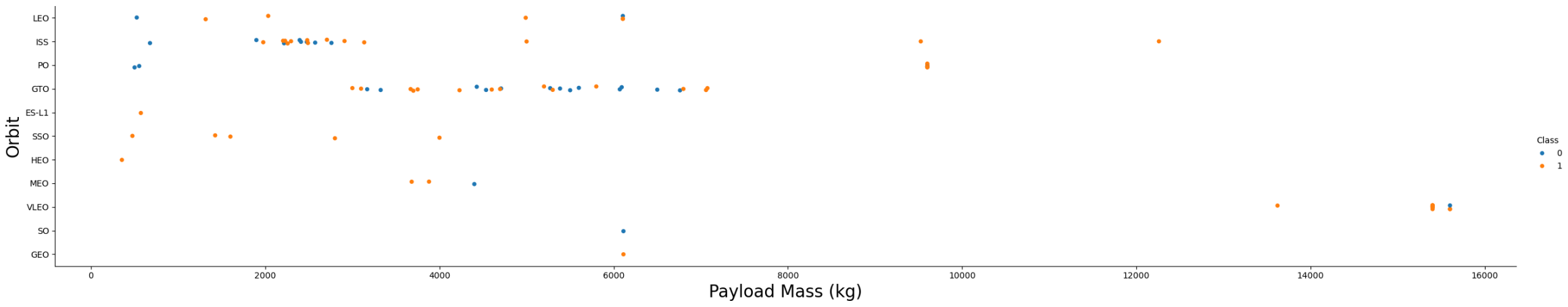
Flight Number vs. Orbit Type



Explanation:

- **Flight Success Over Time**: As flight numbers increase, successful launches (Class 1) become more frequent, reflecting improvements in operational reliability and technology.
- **Orbit Distribution Trends**: Early missions primarily targeted lower orbits (LEO), while later flights diversified into higher orbits (e.g., GEO, SSO).
- **Failure Patterns**: Failures (Class 0) are more common in early flights and lower orbits, suggesting developmental challenges during initial phases.
- **Mature Operations**: The consistent success in higher flight numbers and diverse orbit types indicates the program's growth and adaptability to various mission profiles.

Payload vs. Orbit Type



Explanation:

- **Payload Mass and Orbit Trends**: Lighter payloads (<6,000 kg) dominate lower orbits (LEO), while heavier payloads (>10,000 kg) are almost exclusively placed into higher orbits (GTO, SSO).
- **Success Patterns**: Successful launches (Class 1) are more frequent across all payload masses, with failures (Class 0) more common in lighter payloads and lower orbits.
- **Mission Specialization**: Heavier payloads are successfully deployed into higher orbits, showcasing the program's capability for complex, high-stakes missions.
- **Reliability Growth**: Consistent successes across varying payload masses suggest technological advancements and refined operational procedures over time.

Launch Success Yearly Trend

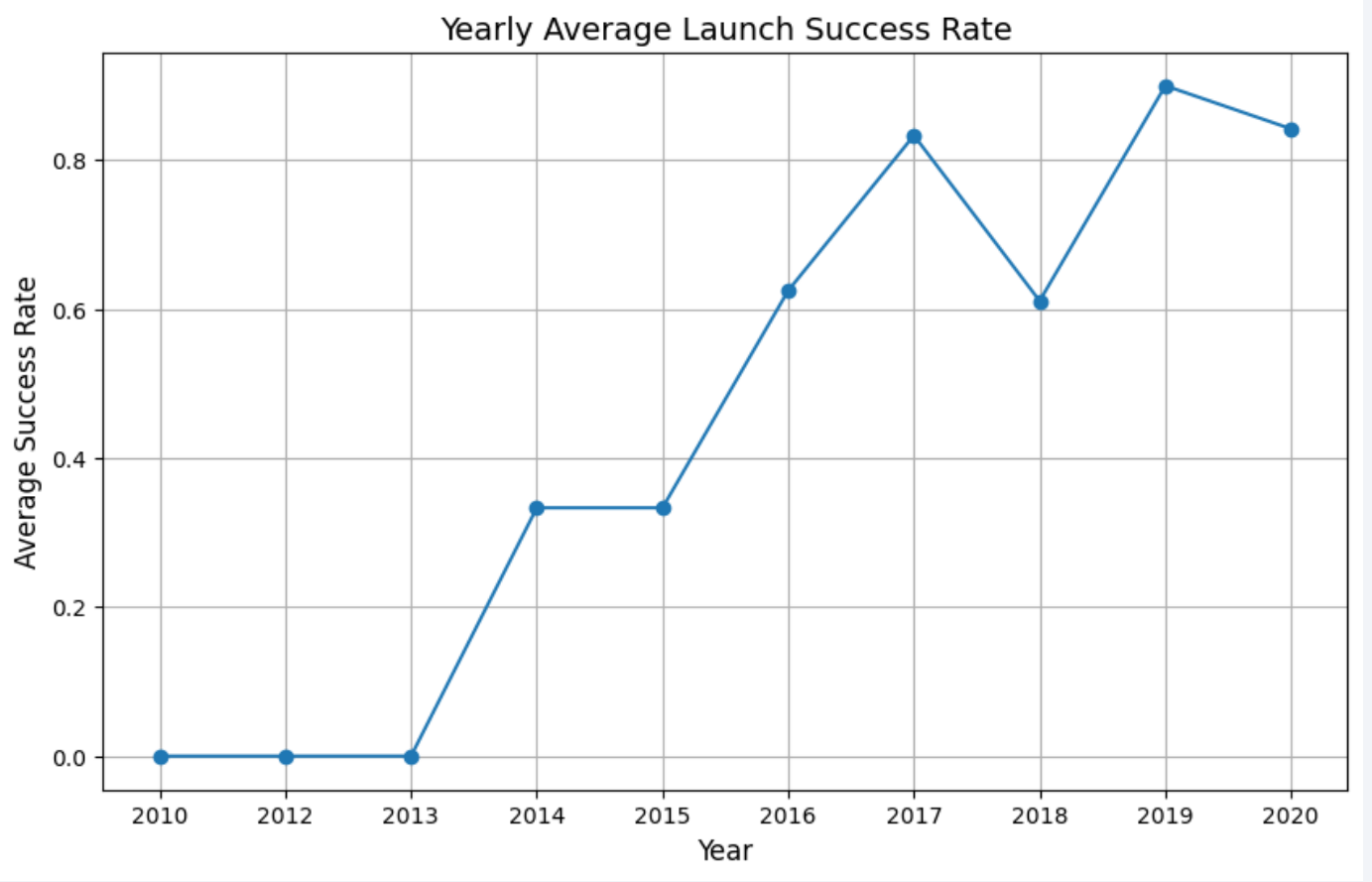
Explanation:

• **Consistent Improvement**: The average launch success rate shows significant growth from 2013 onward, reflecting advancements in technology and operational expertise.

• **Stabilization**: After 2018, the success rate consistently remains above 80%, indicating a mature and reliable launch program.

• **Early Challenges**: The low success rates before 2013 highlight initial developmental hurdles that were effectively overcome in later years.

• **Peak Performance**: The peak in 2018 demonstrates the culmination of iterative improvements and optimization efforts.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
[11]: %sql select distinct launch_site from SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

[11]:

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

Explanation:

• The query successfully identifies four unique launch sites: **CCAFS LC-40**, **VAFB SLC-4E**, **CCAFS SLC-40** and **KSC LC-39A**.

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

[12]:
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

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

[12]:

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

Explanation:

- The query retrieves five launches from **CCAFS LC-40**, showcasing a mix of payloads for SpaceX and NASA customers, including ISS missions.

Total Payload Mass

Task 3

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

```
[13]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

[13]:

total_payload_mass
45596

Explanation:

- The query shows the total payload mass carried by boosters for NASA's CRS missions is **45,696 kg**.
- This demonstrates a significant contribution to NASA's resupply efforts, emphasizing SpaceX's role in supporting ISS logistics.

Average Payload Mass by F9 v1.1

Task 4

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

```
[14]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';

 * sqlite:///my_data1.db
Done.
[14]: average_payload_mass
      2534.6666666666665
```

Explanation:

- The query shows the **average payload mass** carried by the F9 v1.1 booster version is approximately **2,534.67 kg**.
- This reflects the typical capacity of the F9 v1.1 for medium payload missions, highlighting its role in versatile launch operations.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
[16]: %sql select min(date) as first_successful_landing from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
[16]: first_successful_landing
2015-12-22
```

Explanation:

- The query shows the **first successful ground pad landing** occurred on **2015-12-22**.
- This milestone marked a significant achievement in reusable rocket technology, demonstrating SpaceX's advancements in landing precision and recovery.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[17]: %sql select booster_version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6
```

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

[17]:

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

Explanation:

- The query shows the boosters **F9 FT B1022, B1026, B1021.2, and B1031.2** successfully landed on a drone ship while carrying payloads between **4,000 and 6,000 kg**.
- These missions demonstrate the Falcon 9's reliability in handling medium payloads with precision landings at sea.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[18]: %sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
 * sqlite:///my_data1.db
Done.
```

[18]:

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

Explanation:

- The query shows the dataset records **98 successful missions** and **1 in-flight failure**, indicating a high success rate for launches.
- An additional **1 mission** succeeded with an unclear payload status, reflecting minor data discrepancies.
- These results highlight the overall reliability and effectiveness of the launch program.

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[19]: %sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
 * sqlite:///my_data1.db
Done.
```

[19]:

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

Explanation:

- The query shows multiple booster versions, including **F9 B5 B1048.4**, **F9 B5 B1049.4**, and others, have carried the **maximum payload mass**.
- These boosters highlight the Falcon 9 Block 5's capability to handle the heaviest payloads, showcasing its advanced engineering and reliability.

2015 Launch Records

Explanation:

- The query shows in **2015**, there were **two failures on drone ships**: one in **January** (F9 v1.1 B1012) and another in **April** (F9 v1.1 B1013).
- Both launches occurred from **CCAFS LC-40**, reflecting early challenges in achieving successful drone ship landings during this period.

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.

[25]:
```
%%sql

SELECT
    CASE substr(Date, 6, 2)
        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,
    Date,
    booster_version,
    launch_site,
    Landing_Outcome
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date, 0, 5) = '2015';
```

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

[25]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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.

[26]:
```
%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTBL
      where date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

[26]:

Landing_Outcome	count_outcomes
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

Explanation:

- The query shows between **2010-06-04** and **2017-03-28**, the most frequent landing outcome was **No attempt** (15), followed by **Success (ground pad)** (10).
- **Failures on drone ships** (5) highlight early challenges in sea landings, whereas controlled recoveries and parachute landings occurred less frequently (2 each).
- The data reflects SpaceX's gradual improvement in landing precision and technology over this period.

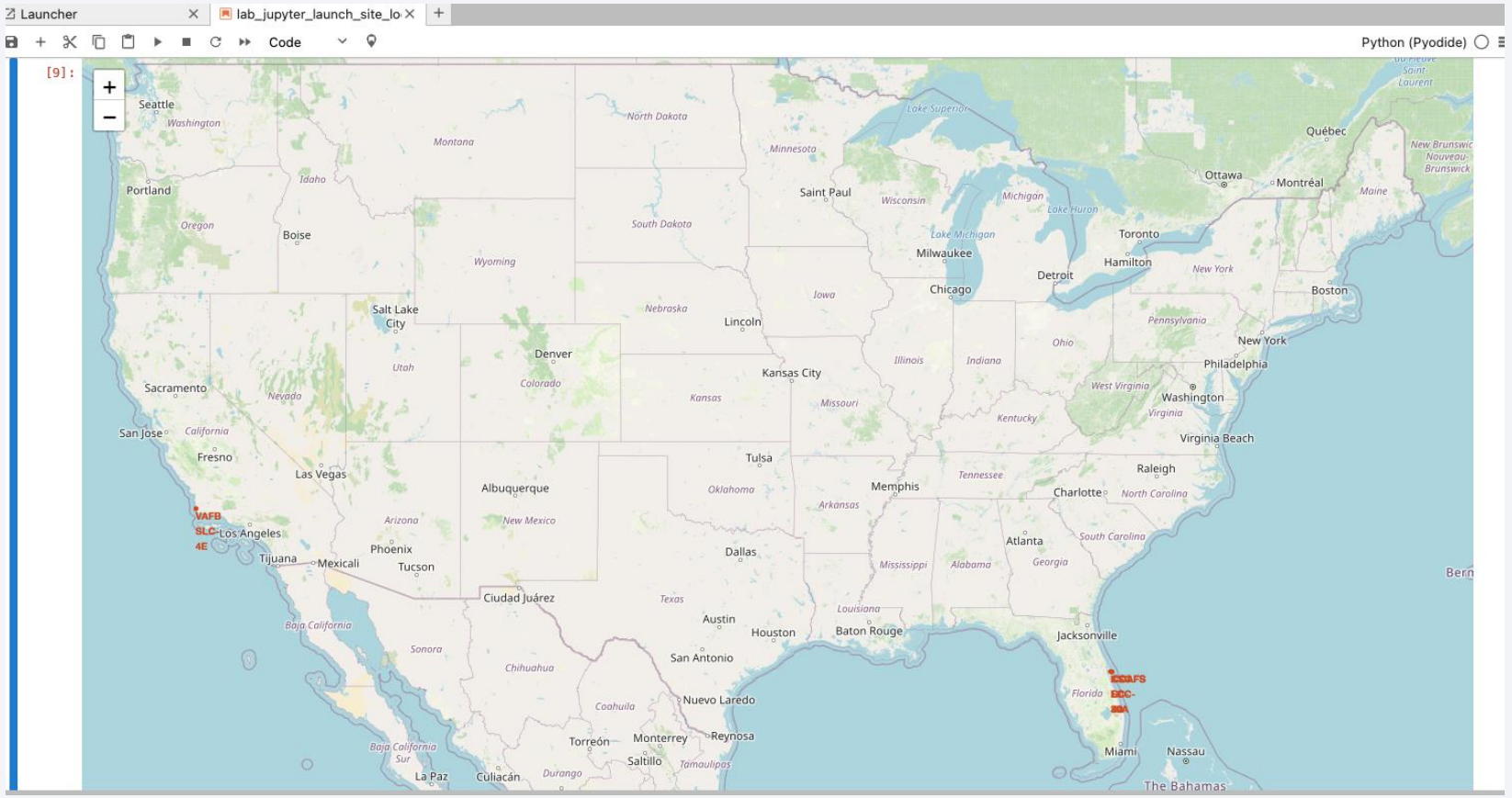
Section 3

Launch Sites Proximities Analysis

Folium Launch Sites

Explanation:

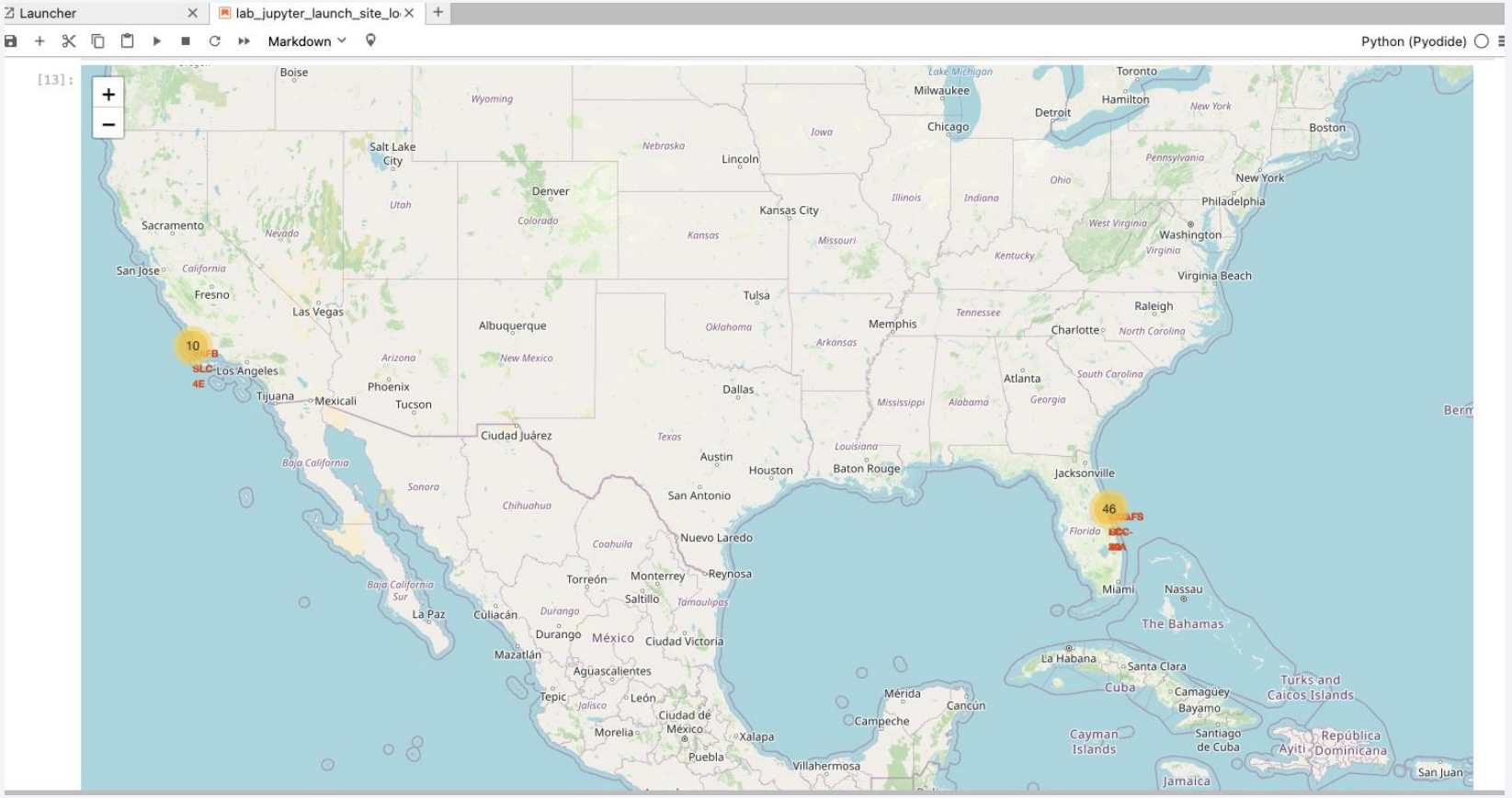
- The map shows all launch sites on



Folium Launch Results

Explanation:

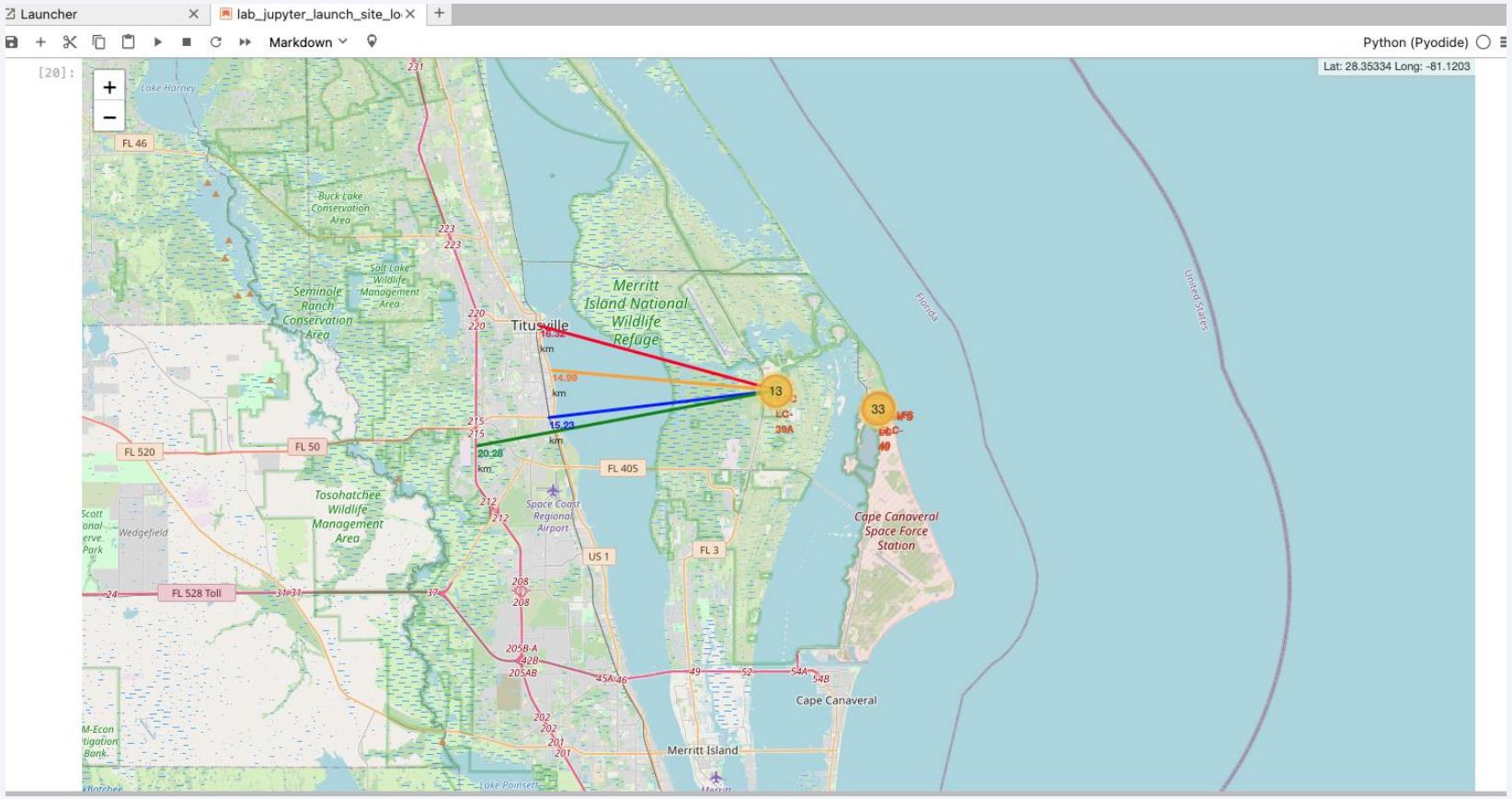
- The map shows marks for the success/failed launches for each site as well as the amount

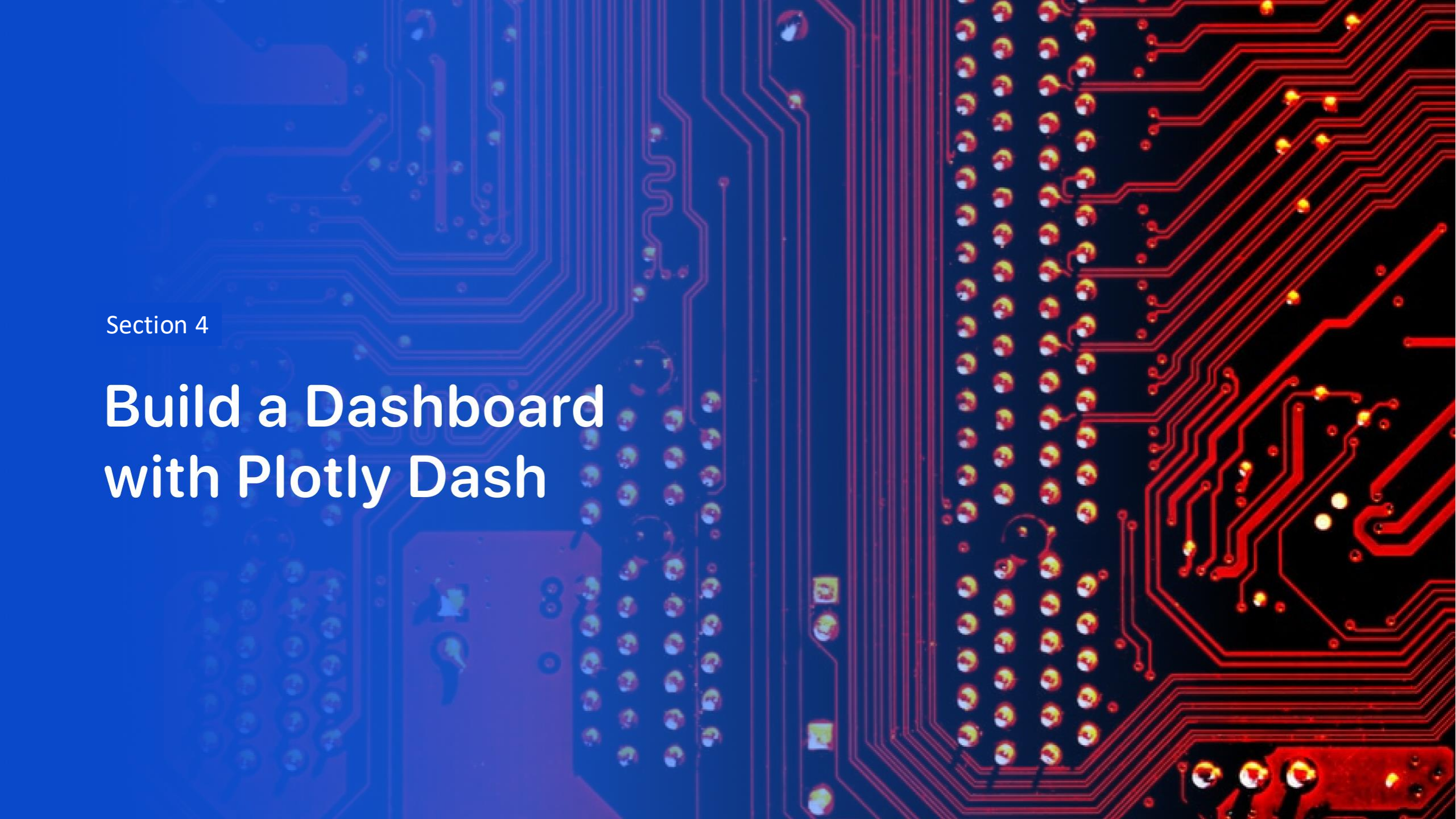


Folium Proximity Map

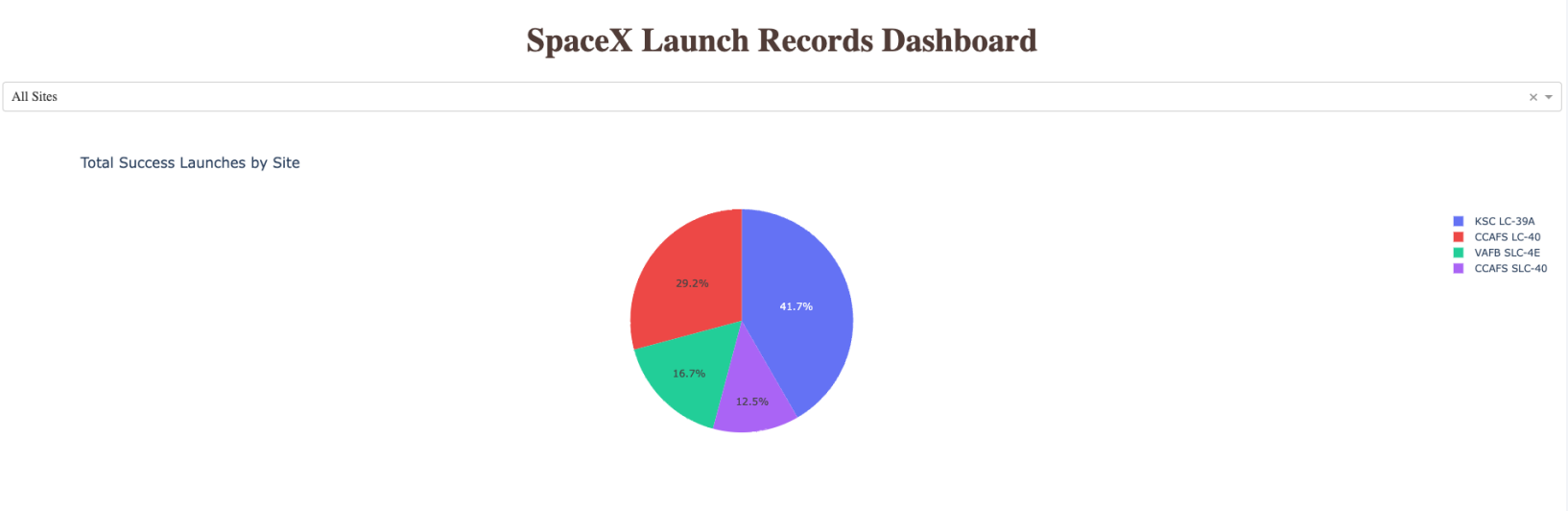
Explanation:

- The map shows the distances between a launch site to its proximities including railways, highways, and coastlines.




Section 4
Build a Dashboard with Plotly Dash

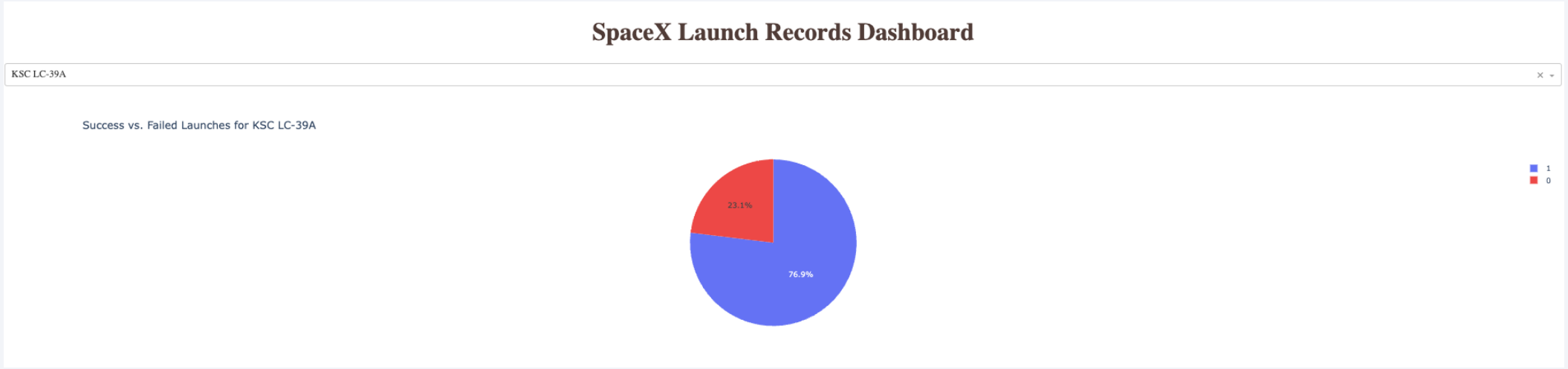
Total Successful Launches by Site



Explanation:

- **KSC LC-39A** leads with 42% of successful launches, highlighting its critical role in SpaceX operations.
- **CCAFS LC-40** and **VAFB SLC-4E** closely follow at 29% and 17%, reflecting balanced utilization across sites.
- SpaceX demonstrates strategic site distribution, optimizing launch capabilities across multiple locations.

KSC LC-39A Launches



Explanation:

- **KSC LC-39A** demonstrates a strong success rate of **76.9%**, reinforcing its reliability as a launch site.
- The **23.1% failure rate** highlights opportunities for further analysis and optimization.
- Overall, the site's performance reflects its critical role in SpaceX's operational strategy

Payload/Success Correlation All Sites

Explanation:

- No clear correlation is observed between payload weight and mission success across all launch sites.
- Success appears consistent across a range of payload weights, suggesting other factors (e.g., rocket type, mission complexity) play a larger role.

Section 5

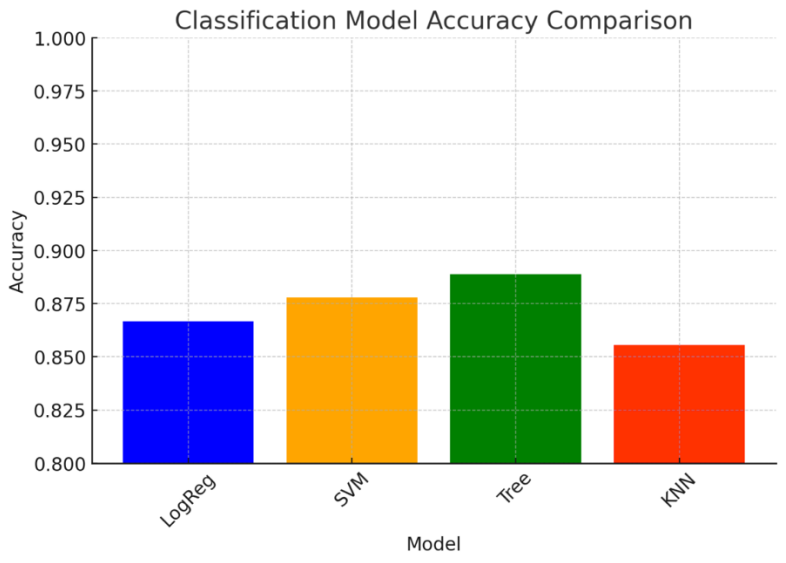
Predictive Analysis (Classification)

Classification Accuracy

The bar chart visualizes the accuracy of all classification models. Among the models, the **Decision Tree (Tree)** has the highest classification accuracy at **88.89%**, indicating it performs best in this comparison.

[33]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.857143	0.819444
F1_Score	0.909091	0.916031	0.923077	0.900763
Accuracy	0.866667	0.877778	0.888889	0.855556



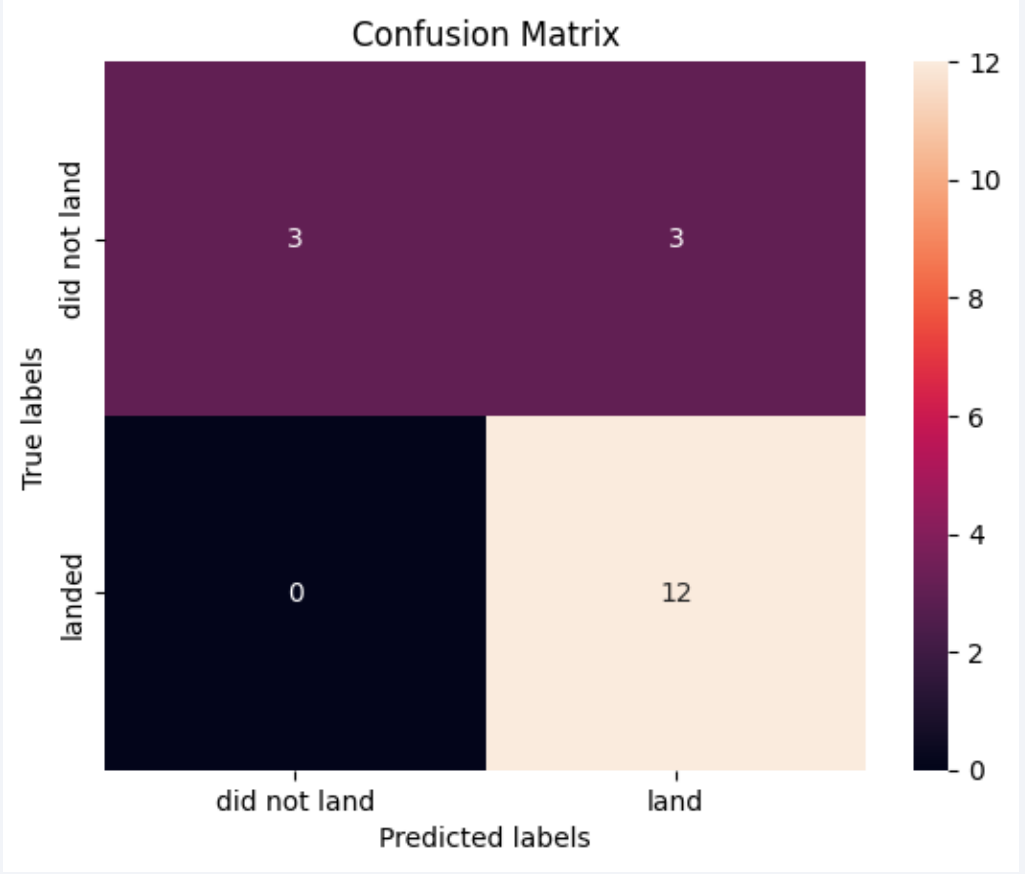
Confusion Matrix

Explanation:

• The **Decision Tree** model correctly classified **12 landed cases** and **3 non-landed cases**, demonstrating strong performance for positive predictions.

• However, it misclassified **3 non-landed cases as landed**, indicating room for improvement in distinguishing failed landings.

• Overall, the model exhibits high precision for successful landings, supporting its utility in predicting mission outcomes.



Conclusions

- The **Decision Tree model** performed best for this dataset, making it the most effective algorithm.
- Launches with **lighter payloads** tend to have higher success rates compared to those with heavier payloads.
- Most launch sites are **located near the Equator and the coast**, optimizing launch efficiency and logistics.
- The **success rate of launches** has steadily improved over time, showing progress in reliability.
- **KSC LC-39A** stands out with the highest success rate among all launch sites.
- Orbits like **ES-L1, GEO, HEO, and SSO** have achieved a **100% success rate**, showcasing their reliability.

Appendix

- Github Main

Thank you!