IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ali Salman

March 4, 2024

Outline

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

Executive Summary

- We collected data from the Space X API and Wikipedia, creating a 'class' column for categorizing successful landings. Utilizing SQL, visualizations, Folium maps, and dashboards, we explored the data, identified key features, and converted categorical variables to binary through one-hot encoding. Standardizing the data, we employed GridSearchCV to optimize machine learning model parameters, visualizing accuracy scores.
- The machine learning models (Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors) all exhibited approximately 83.33% accuracy, consistently over-predicting successful landings. To enhance accuracy, obtaining additional data is crucial.

Introduction

- We examined Space X and its emerging competitor, Blue Origin, in the context of the commercial space age. Space X's pricing dominance, offering launches at $62 million compared to Blue Origin's $165 million, is notable. The recovery of rocket components, especially in Stage 1, is a key factor contributing to Space X's cost advantage. As Blue Origin seeks to challenge Space X, a detailed exploration of their strategies and innovations is warranted.
- Blue Origin has commissioned us to develop a machine learning model aimed at predicting the success of Space X's Stage 1 recovery.

Section 1

Methodology

Methodology

Executive Summary

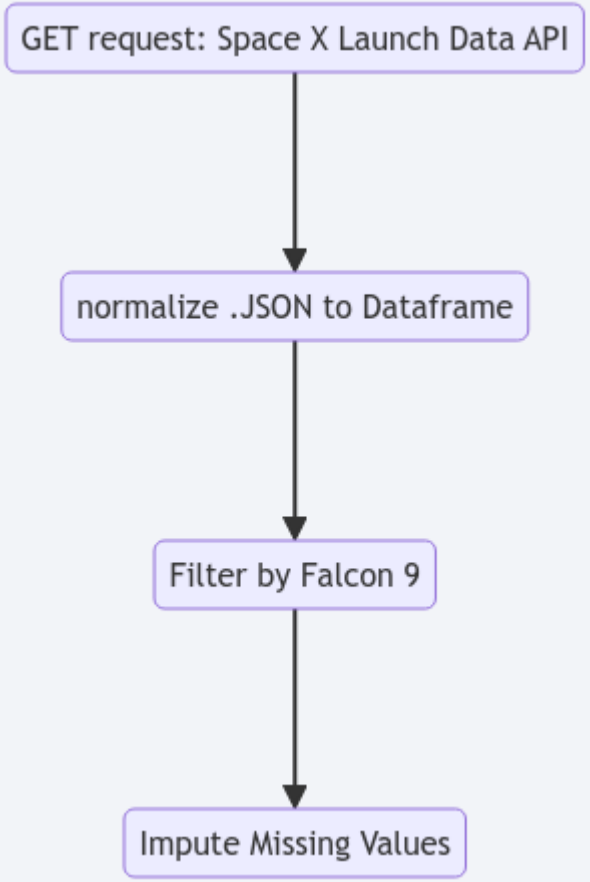
- Data collection methodology:
  - Merged information sourced from both the Space X public API and the Space X Wikipedia page.
- Perform data wrangling
  - Categorizing landings as either successful or unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Refined models through the utilization of GridSearchCV.

Data Collection

- The data collection process involved a dual approach, comprising API requests from the Space X public API and web scraping data from a table within Space X's Wikipedia entry.
- The upcoming slide will present the flowchart detailing the data collection process through API, while the subsequent one will illustrate the flowchart for data collection via web scraping.

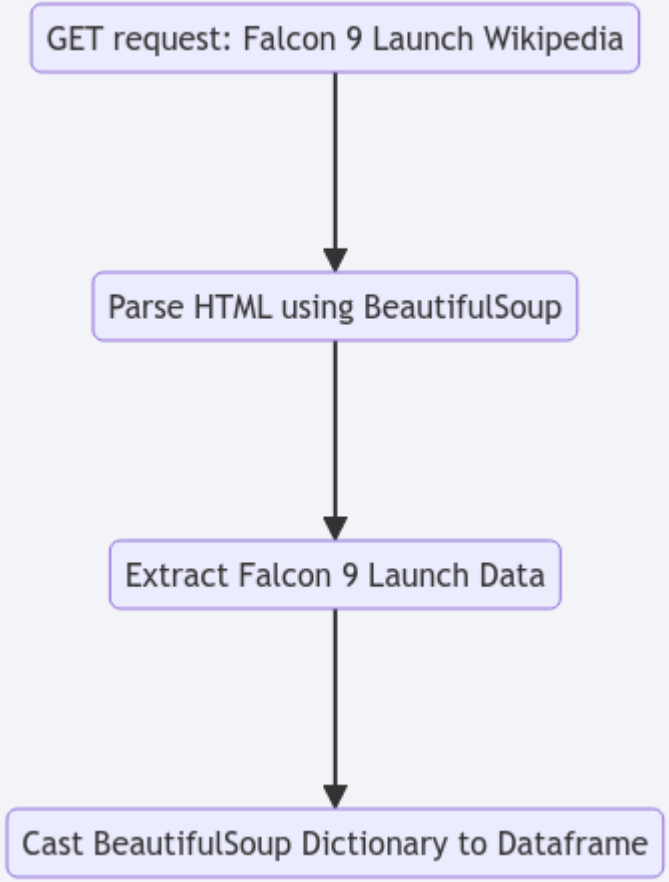
Data Collection – SpaceX API

- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%201/1.%20SpaceX_Data_Collection_API.ipynb



Data Collection - Scraping

- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%201/2.%20SpaceX_Webscraping.ipynb



Data Wrangling

- We introduce a new training label column named 'class', which takes a value of 1 when the 'Mission Outcome' is deemed successful, and 0 otherwise. Here's the breakdown for value mapping:
  - If 'Mission Outcome' is True and corresponds to ASDS, RTLS, or Ocean, we assign the class value as 1.
  - If 'Mission Outcome' is None for both components, or if it's False for ASDS, or if it's False for Ocean, or if it's False for RTLS, we assign the class value as 0.
- This classification scheme simplifies the representation of landing outcomes while enabling effective training and analysis.
- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%201/3.%20SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

- Conducted Exploratory Data Analysis (EDA) on variables including Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year.
- Employed scatter plots, line charts, and bar plots to assess relationships.
- Explored Flight Number vs. Payload Mass, Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs. Orbit, and Success Yearly Trend.
- The objective was to ascertain the presence of any discernible relationships among the variables, thereby determining their suitability for inclusion in the machine learning model training process.
- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%202/2.%20SpaceX_EDA_DataViz.ipynb

EDA with SQL

- Utilized SQL Python integration for querying purposes, aiming to enhance comprehension of the dataset.
- Extracted information on launch site names, mission outcomes, diverse payload sizes for customers, booster versions, and landing outcomes.
- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%202/1.%20SpaceX_EDA_SQL.ipynb

Build an Interactive Map with Folium

- Folium maps are employed to pinpoint Launch Sites, marking both successful and unsuccessful landings, along with illustrating proximity to key locations such as Railway, Highway, Coast, and City.
- This visualization aids in comprehending the rationale behind the selection of launch site locations. Additionally, it provides a visual representation of successful landings in relation to these key locations.
- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%203/1.%20SpaceX_LaunchSite_Location.ipynb

Build a Dashboard with Plotly Dash

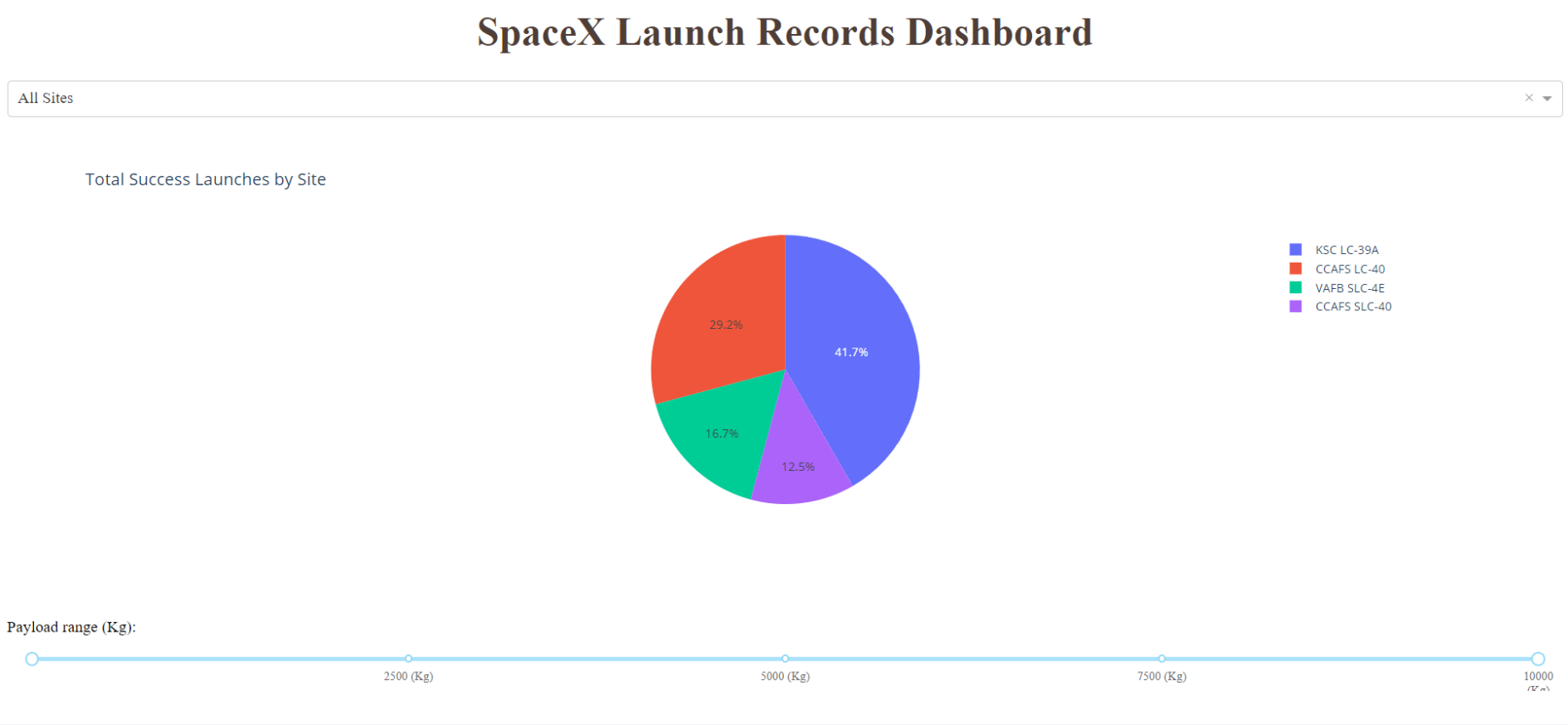
- The dashboard features a pie chart for viewing successful landing distribution across all or individual launch sites, highlighting success rates.
- The scatter plot, with adjustable payload mass (0 to 10,000 kg) and site selection, enables the exploration of success variations across launch sites, payload mass, and booster version categories.
- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%203/SpaceX_Dashboard.py

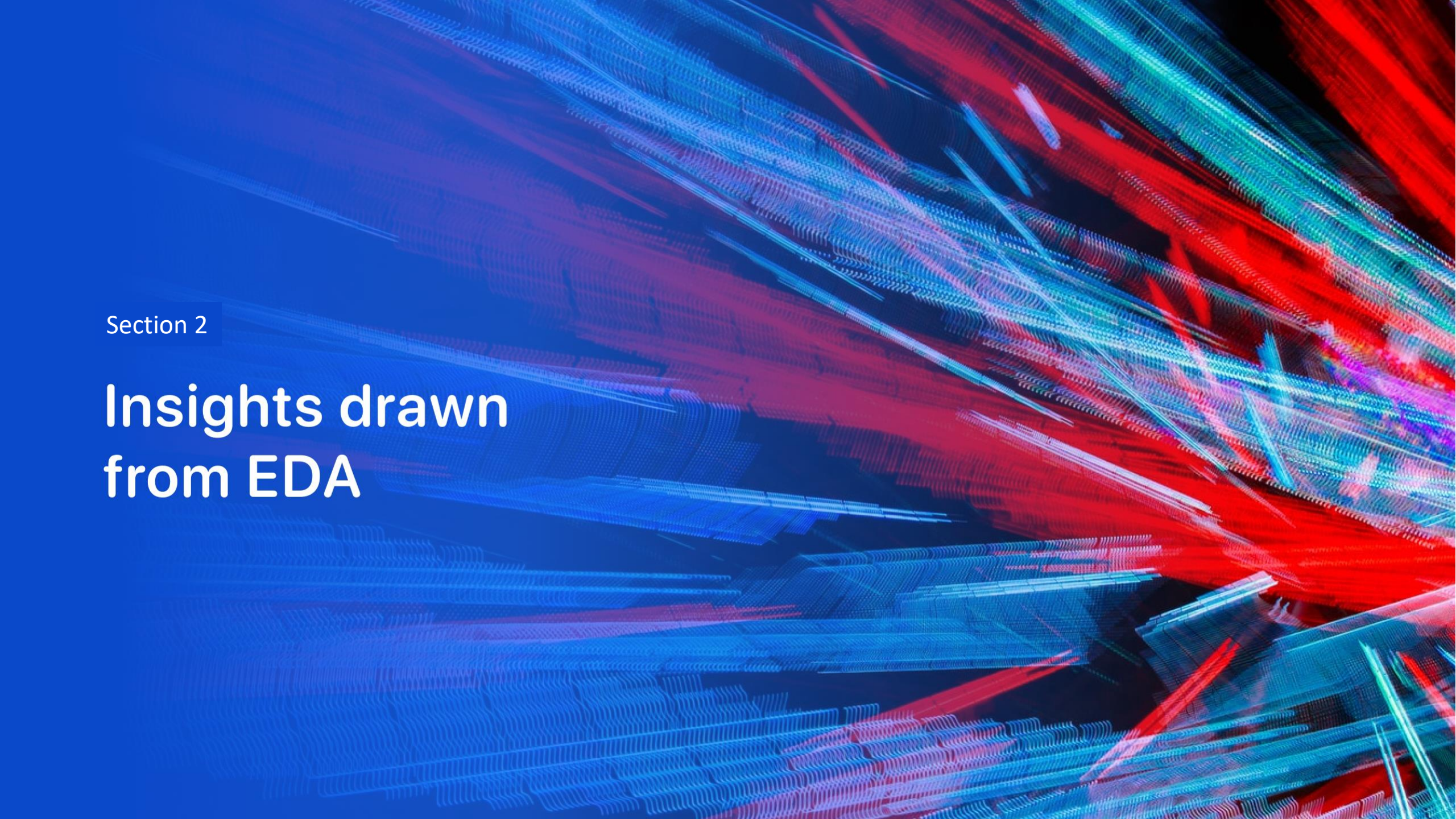
Predictive Analysis (Classification)

- The initial phase involved preprocessing the dataset by isolating the "Class" label and standardizing features with the standard scaler.
- Subsequent steps included dividing the data into training and testing sets, utilizing GridSearchCV for optimal parameter selection across logistic regression, support vector machine, decision tree, and KNN algorithms. Following this, the models were trained and assessed on the test set.
- Evaluation extended to generating confusion matrices for each model, and a concise view of their relative scores.
- GitHub URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/blob/main/Capstone/Week%204/1.%20SpaceX_Machine_Learning.ipynb

Results

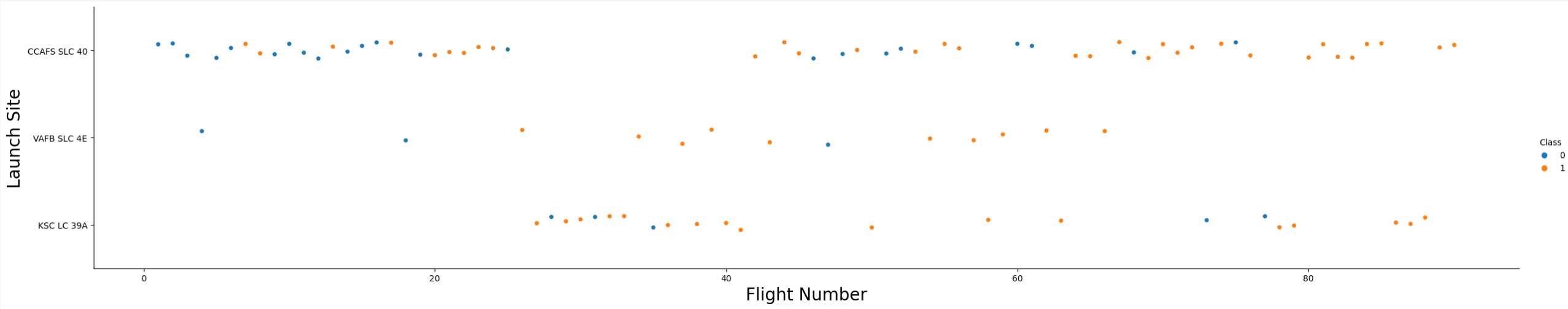
- Here's a sneak peek at the Plotly dashboard, providing insights into the exploratory data analysis (EDA) results through visualization and SQL.




Section 2
Insights drawn from EDA

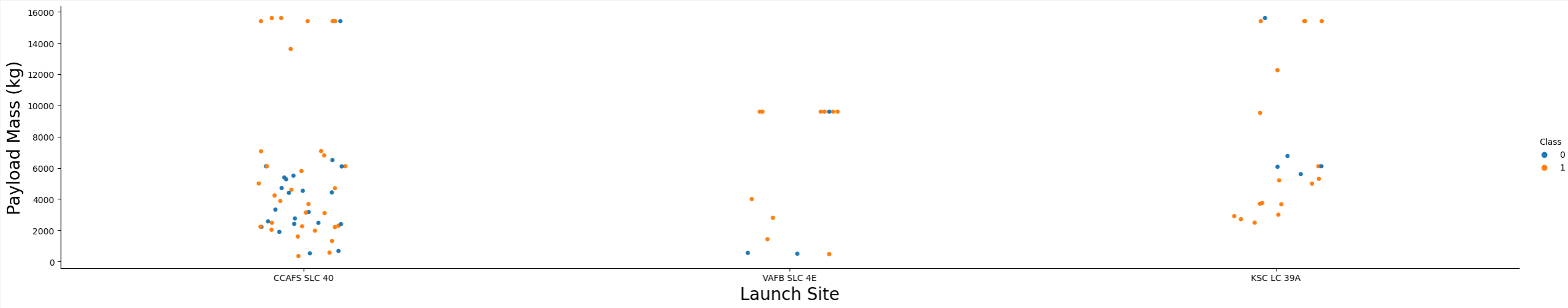
Flight Number vs. Launch Site

- The color scheme distinguishes successful (orange) and unsuccessful (blue) launches in the graphic.
- Notably, there's a discernible upward trend in success rates over time, particularly evident around flight number 20, suggesting a potential breakthrough that significantly boosted success. CCAFS emerges as the primary launch site, evident from its substantial launch volume.



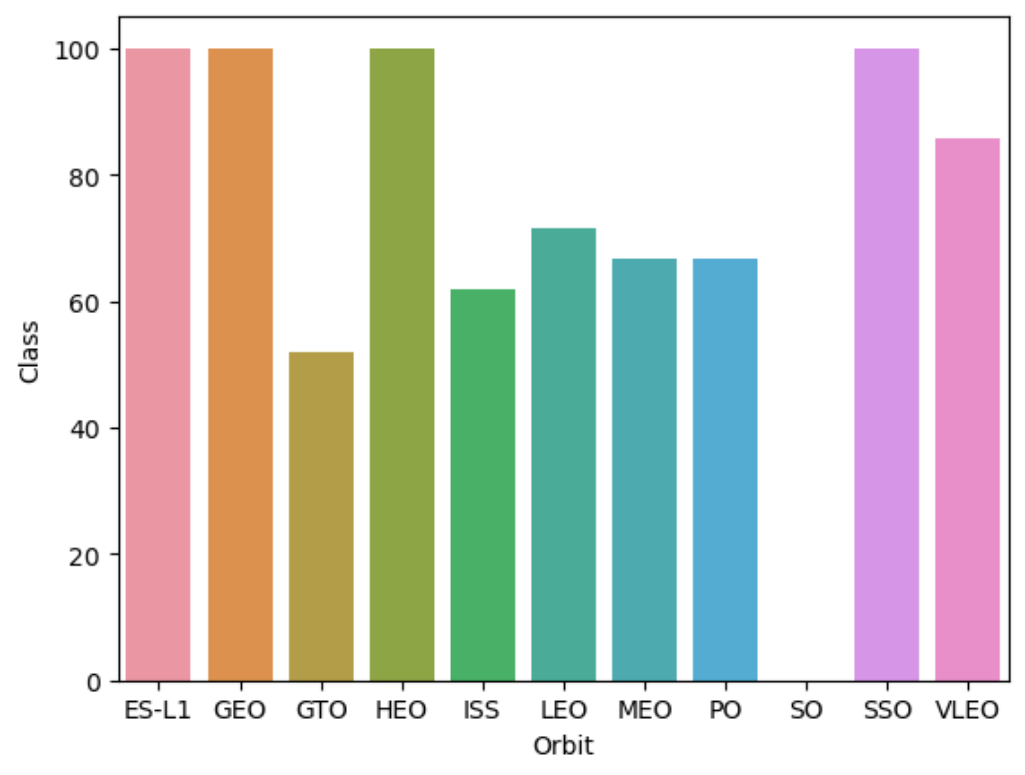
Payload vs. Launch Site

- The color scheme distinguishes successful (orange) and unsuccessful (blue) launches in the graphic.
- The payload mass is predominantly within the 0-6000 kg range, with varying usage across different launch sites.



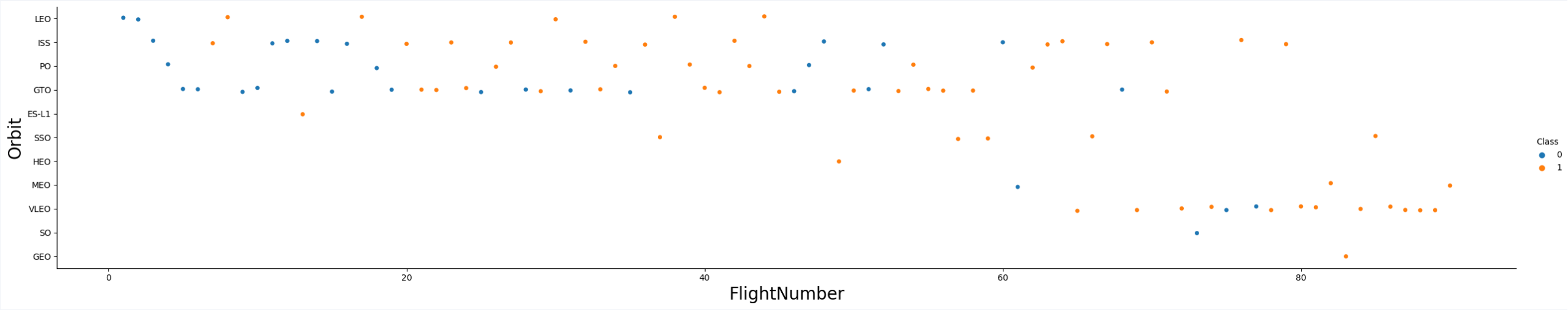
Success Rate vs. Orbit Type

- ES-L1, GEO, and HEO, each with a sample size of one, boast a 100% success rate.
- Similarly, SSO, with a sample size of five, achieves a perfect success rate.
- VLEO, with 14 attempts, shows a decent success rate, while GTO, despite its largest sample size of 27, achieves around a 50% success rate.
- Conversely, SO, with a single attempt, records a 0% success rate.



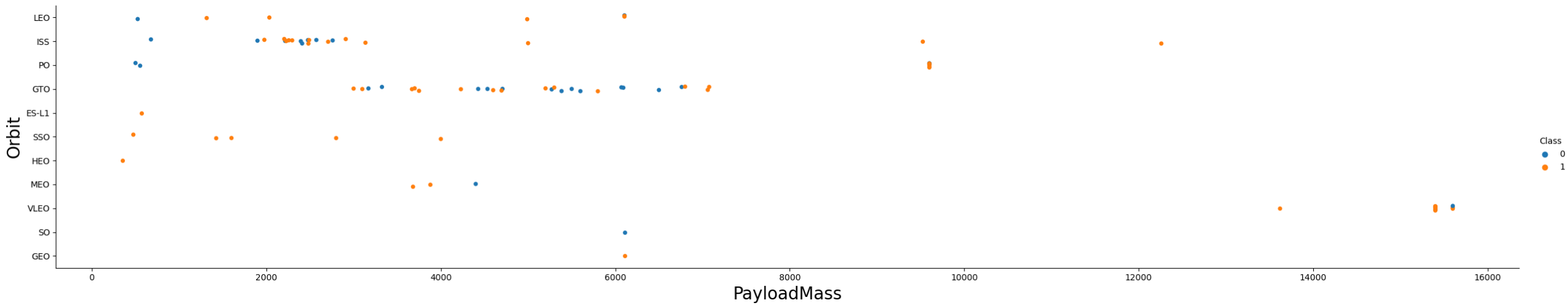
Flight Number vs. Orbit Type

- The color scheme distinguishes successful (orange) and unsuccessful (blue) launches in the graphic.
- The preference for launch orbits shifted with the progression of Flight Number, and there appears to be a correlation between Launch Outcome and this preference. SpaceX initially focused on LEO orbits, achieving moderate success, then shifted back to VLEO in recent launches. Notably, SpaceX exhibits better performance in lower orbits or Sun-synchronous orbits.



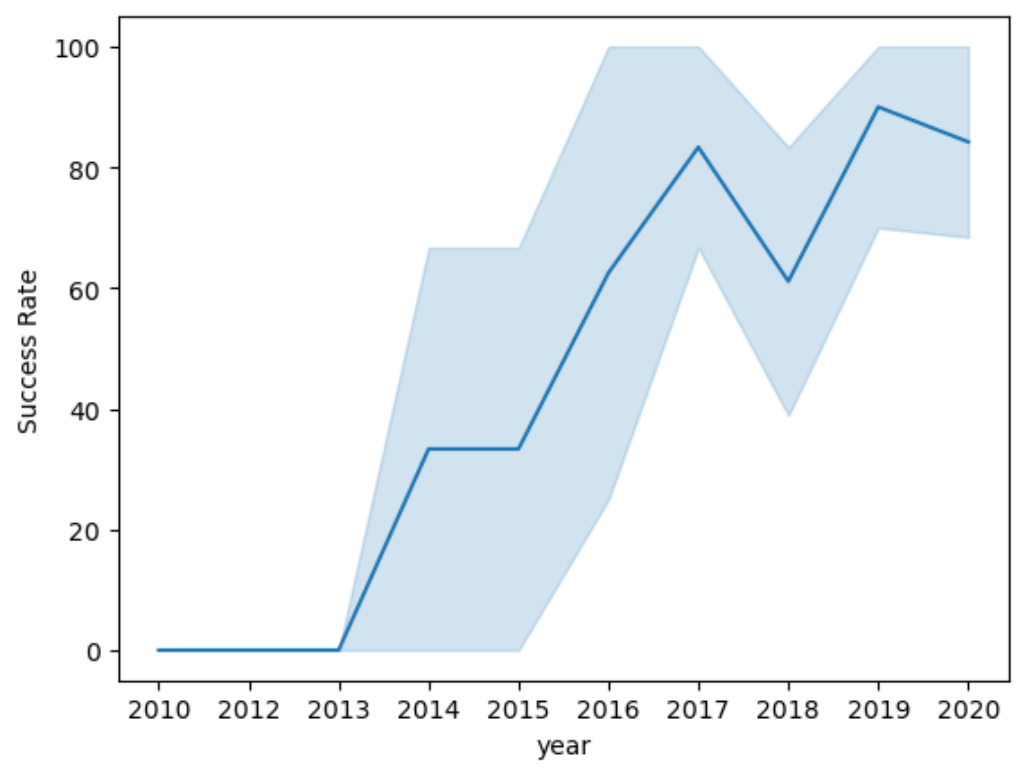
Payload vs. Orbit Type

- The color scheme distinguishes successful (orange) and unsuccessful (blue) launches in the graphic.
- There's a noticeable correlation between payload mass and orbit. Lower Earth Orbit (LEO) and Sun-synchronous Orbit (SSO) tend to have relatively lower payload masses, while Very Low Earth Orbit (VLEO), the other highly successful orbit, primarily features payload mass values at the higher end of the range.



Launch Success Yearly Trend

- The blue ribbon denotes the 95% confidence interval
- Overall, success rates have shown a general upward trend since 2013, with a minor dip observed in 2018.
- In recent years, success rates have consistently hovered around 80%.



All Launch Site Names

- Retrieved unique launch site names from the database and identified potential data entry errors.
- It appears that CCAFS SLC-40 and CCAFSSLC-40 likely represent the same launch site, possibly due to errors in data entry.
- The previous name, CCAFS LC-40, suggests that there are likely only three unique launch site values: CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

```
In [7]: %sql select DISTINCT LAUNCH_SITE from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
Out[7]:
```

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

Launch Site Names Begin with 'CCA'

- Retrieved 5 records where launch sites begin with `CCA`

```
In [8]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

Total Payload Mass

- This query calculates the cumulative payload mass, measured in kilograms, for missions where NASA served as the customer.
- Notably, the designation CRS (Commercial Resupply Services) signifies that these payloads were specifically intended for delivery to the International Space Station (ISS).

Task 3

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

In [9]:
```
%sql select sum(payload_mass__kg_) as sum from SPACEXTABLE where customer like 'NASA (CRS)'
```

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

Out[9]:

sum
45596

Average Payload Mass by F9 v1.1

- This query computes the average payload mass for launches employing the booster version F9 v1.1.
- It's worth noting that the average payload mass for F9 1.1 tends to be on the lower side of our payload mass range.

Task 4

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

In [10]:
```
%sql select avg(payload_mass__kg_) as Average from SPACEXTABLE where booster_version like 'F9 v1.1%'
```

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

Out[10]:

Average
2534.6666666666665

First Successful Ground Landing Date

- This query provides the date of the initial successful ground pad landing, which occurred towards the close of 2015
- Successful landings, in general, seem to commence around 2014.

Task 5

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

Hint:Use min function

In [11]:
```
%sql select min(date) as Date from SPACEXTABLE where mission_outcome like 'Success'
```

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

Out[11]:

Date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query identifies the four booster versions that achieved successful drone ship landings with a payload mass falling between 4000 and 6000 kg (non-inclusive).

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 [18]:
```
%sql select booster_version from SPACEXTABLE where (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND
```

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

Out[18]:

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

Total Number of Successful and Failure Mission Outcomes

- This query provides a count of each mission outcome, indicating that SpaceX achieves its mission objectives nearly 99% of the time.
- Notably, the data suggests that most landing failures are intentional. Intriguingly, there is one launch with an unclear payload status, and regrettably, one experienced an in-flight failure.

Task 7

List the total number of successful and failure mission outcomes

In [19]:
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE GROUP by mission_outcome ORDER BY mission_outcome
```

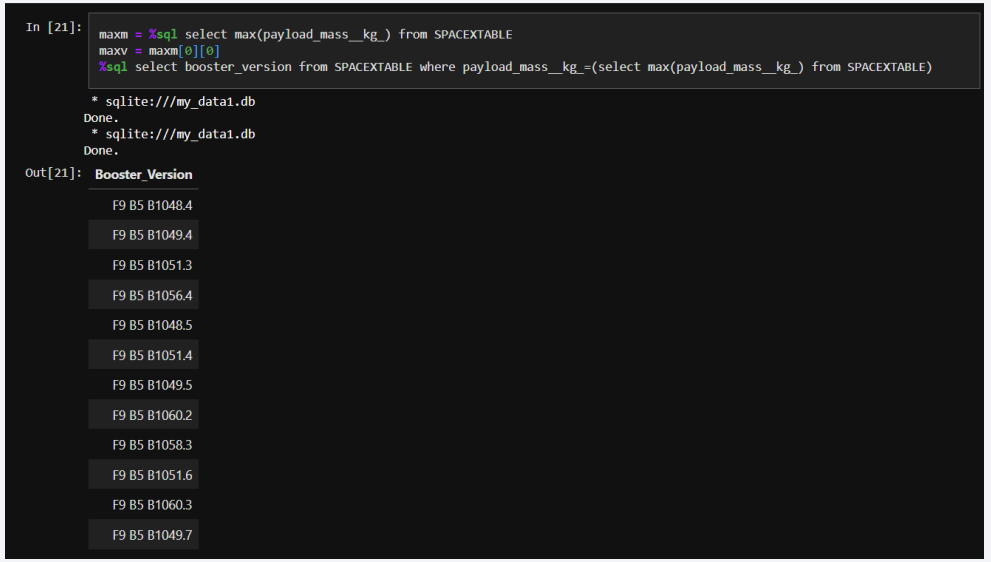
```
 * sqlite:///my_data1.db
Done.
```

Out[19]:

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

Boosters Carried Maximum Payload

- This query identifies the booster versions that carried the maximum payload mass of 15600 kg, revealing that they are all closely related, belonging to the F9 B5 B10xx.x variety.
- This observation suggests a correlation between payload mass and the specific booster version employed.



```
In [21]: maxm = %sql select max(payload_mass__kg_) from SPACEXTABLE
         maxv = maxm[0][0]
         %sql select booster_version from SPACEXTABLE where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
Out[21]:
```

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

2015 Launch Records

- This query provides details on launches from 2015 where the first stage failed to land on a drone ship. The information includes the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch Site.
- Notably, there were two instances meeting these criteria during that year.

```
In [41]: %sql SELECT strftime('%m', DATE) AS Month, landing_outcome, booster_version, launch_site FROM SPACEXTABLE WHERE DATE LIKE '2
```

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

Out[41]:

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 compiles a list of successful landings between June 4, 2010, and March 20, 2017, inclusive.
- These successful landings encompass two types: drone ship and ground pad landings. In total, there were eight successful landings recorded during this specified time period.

```
In [42]: %sql select landing_outcome, count(*) as count from SPACEXTABLE where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP by
 * sqlite:///my_data1.db
Done.
```

Out[42]:

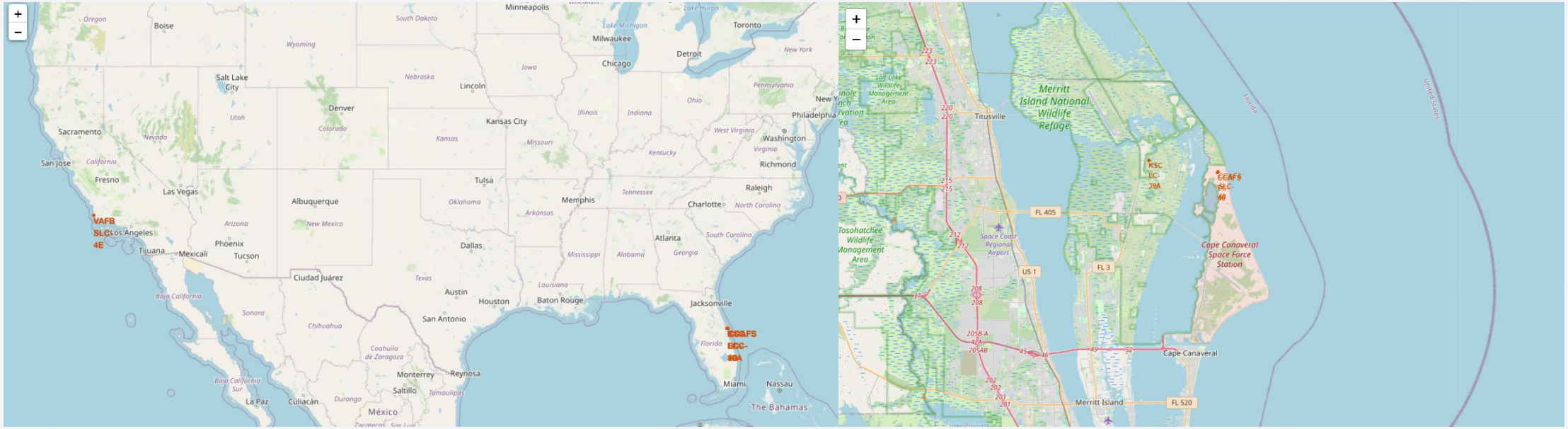
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

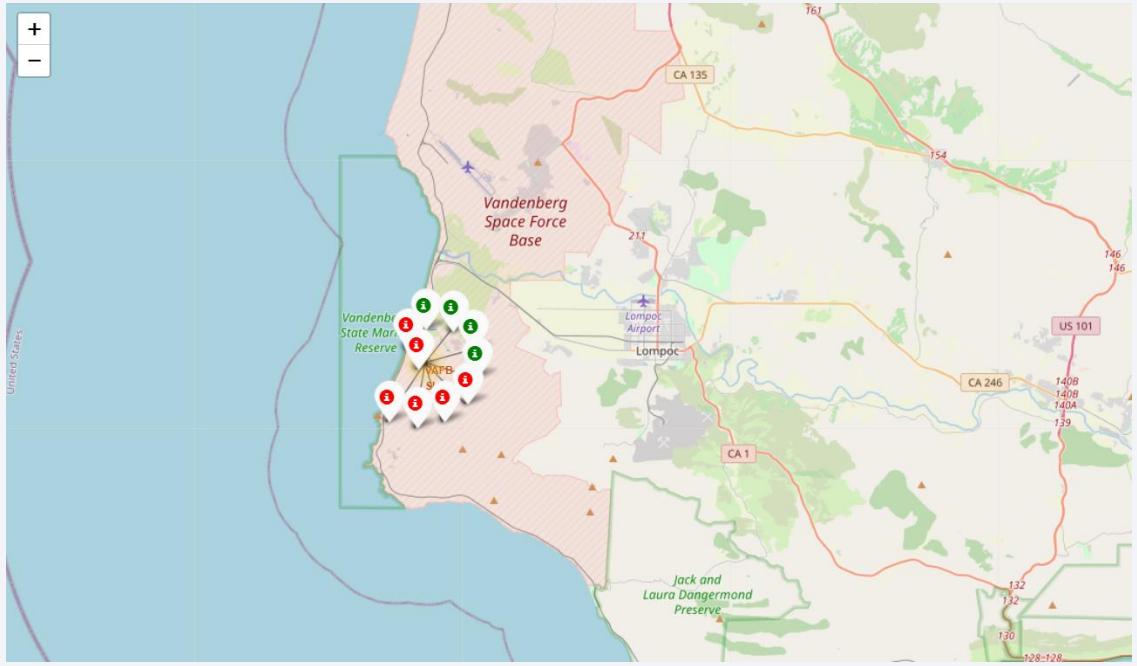
Launch Site: Geography

- The map on the left displays all launch sites in relation to the US map, while the map on the right specifically focuses on the two launch sites in Florida due to their close proximity. Notably, all launch sites are situated near the ocean.



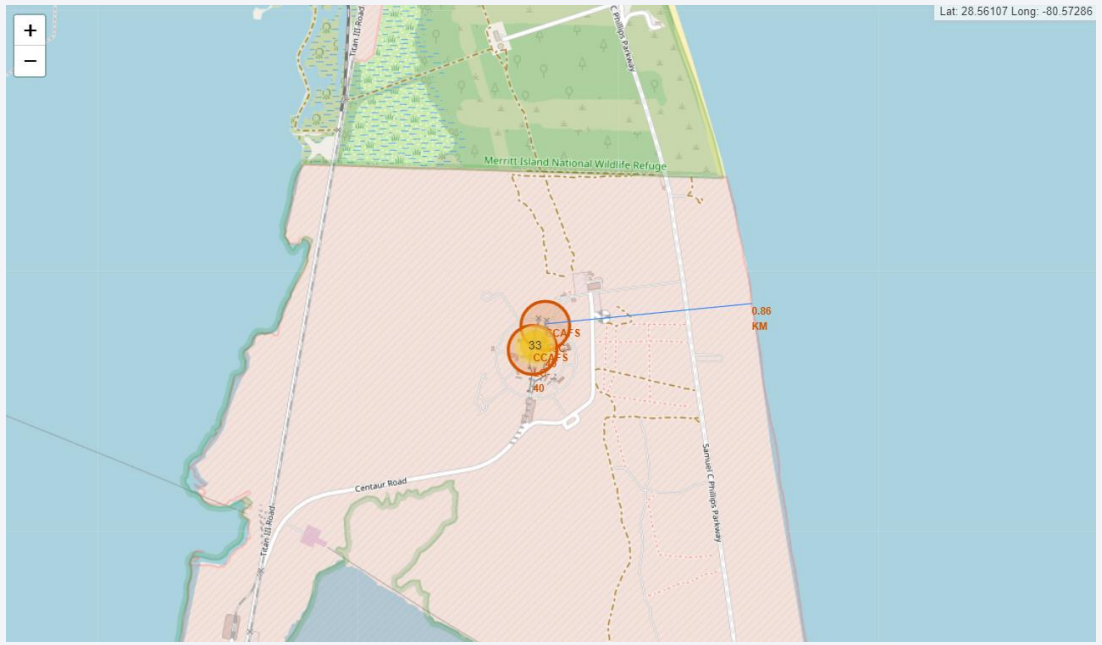
Launch Site: Markers

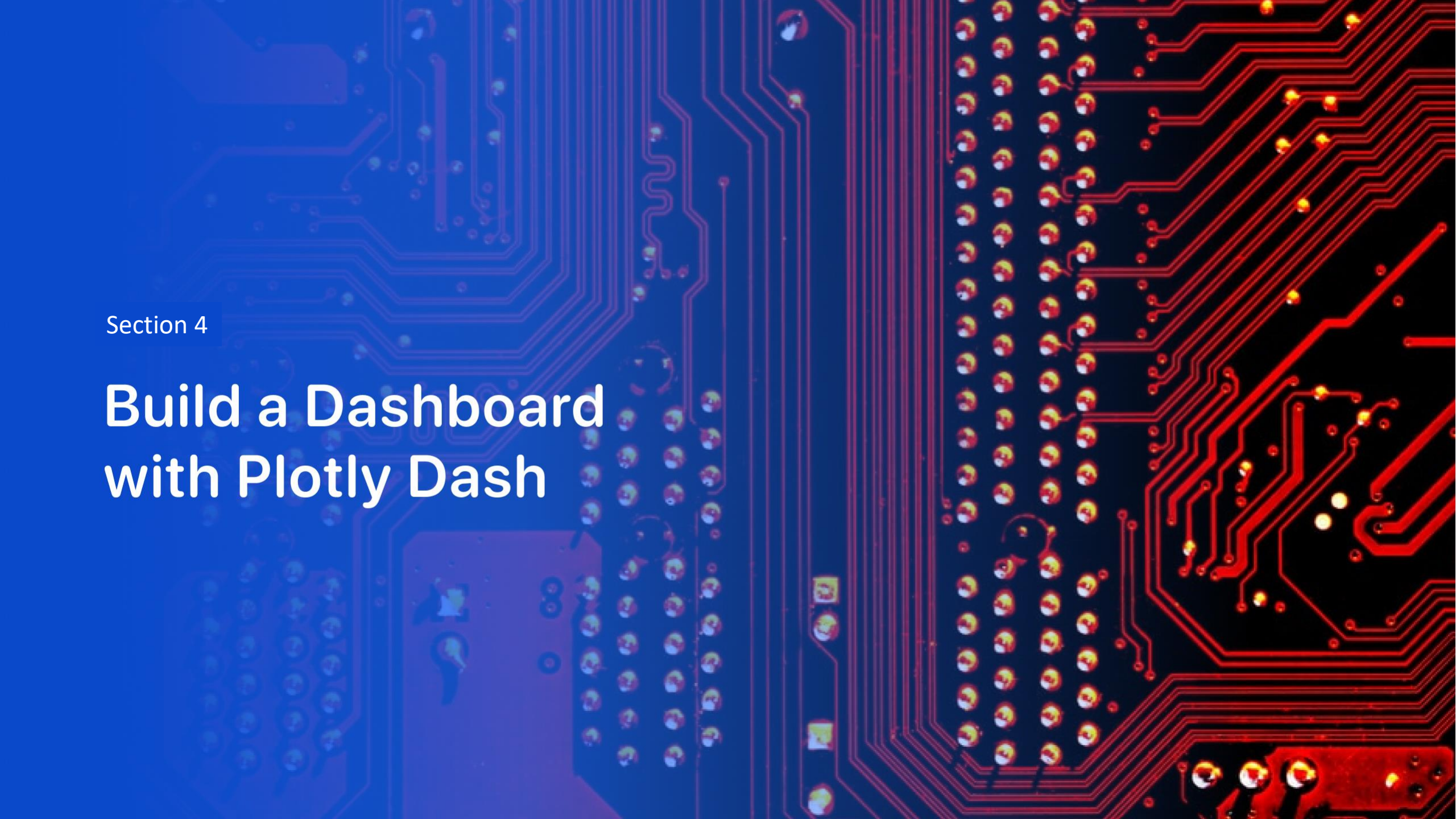
- On the Folium map, clusters can be clicked to reveal individual successful landings (green icon) and unsuccessful landings (red icon).
- As an illustration, clicking on VAFB SLC-4E cluster would disclose 4 successful landings and 6 failed landings associated with that launch site.



Launch Site: Key Proximities

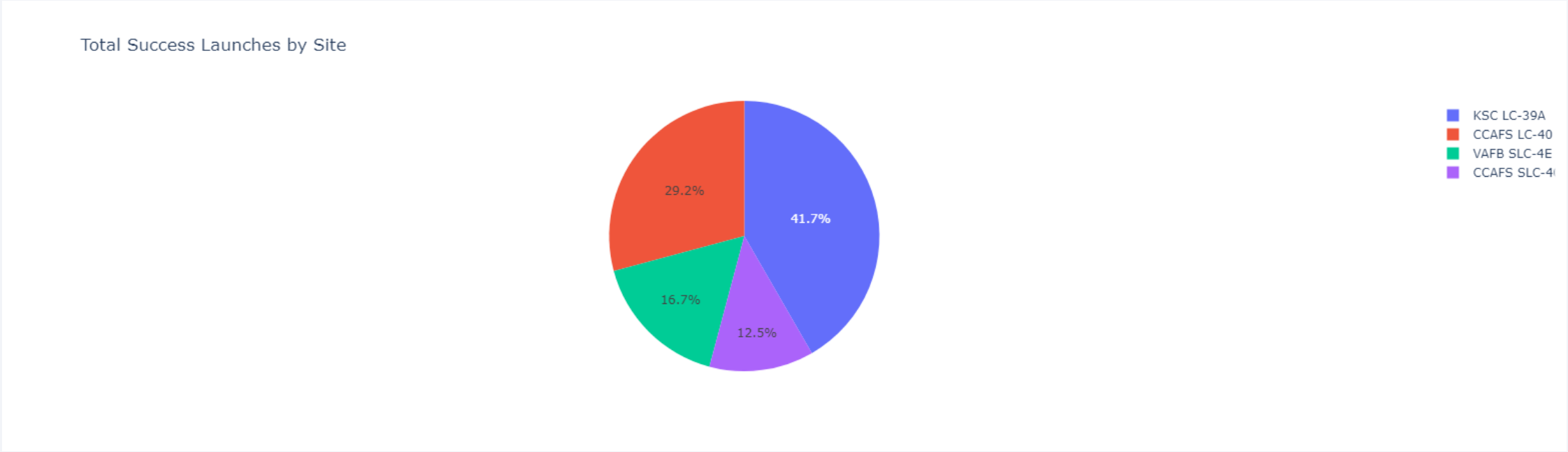
- Taking KSC LC-39A as an illustration, launch sites are strategically positioned in close proximity to railways for efficient large-scale transportation.
- Additionally, they are situated near highways to facilitate human and supply transport. Launch sites are also strategically close to coasts while maintaining a relative distance from densely populated cities. This positioning minimizes the risk of launch failures leading to rockets landing in the sea instead of densely populated areas.




Section 4
Build a Dashboard with Plotly Dash

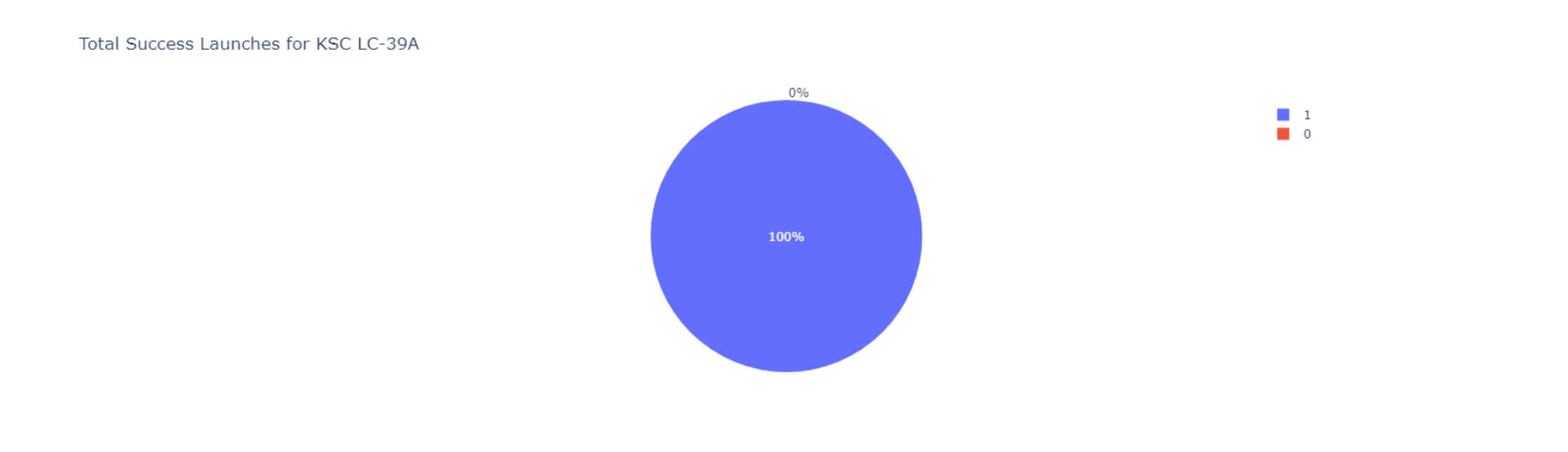
Launch Success Rates Across Facilities

- This outlines the distribution of successful landings across various launch sites. CCAFS and KSC share the same number of successful landings, with a majority occurring before a name change. VAFB has the smallest share, possibly due to its smaller sample size and the increased challenges of launching from the west coast.



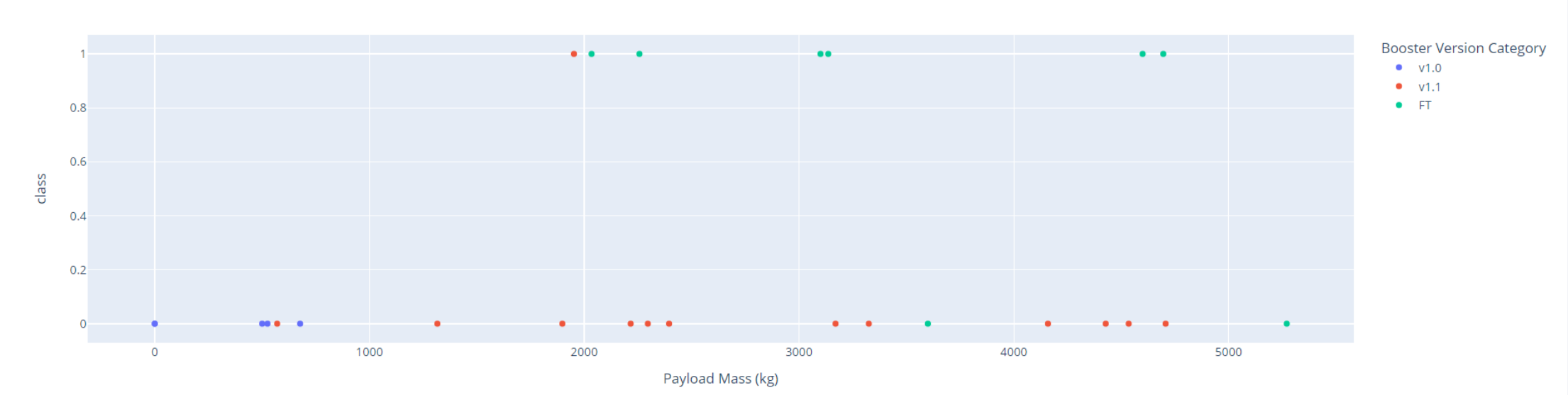
Leading Launch Sites for Mission Success

- KSC LC-39A boasts the highest success rate, with a total of 10 successful landings.



Exploring the Impact of Payload Weight, Launch Outcome, and Rocket Configuration

- Examining Payload Weight, Launch Outcome, and Rocket Configuration on the Plotly dashboard reveals a range limitation from 0-10000 instead of the actual max payload of 15600. The "Class" parameter signifies successful (1) or failed (0) landings.

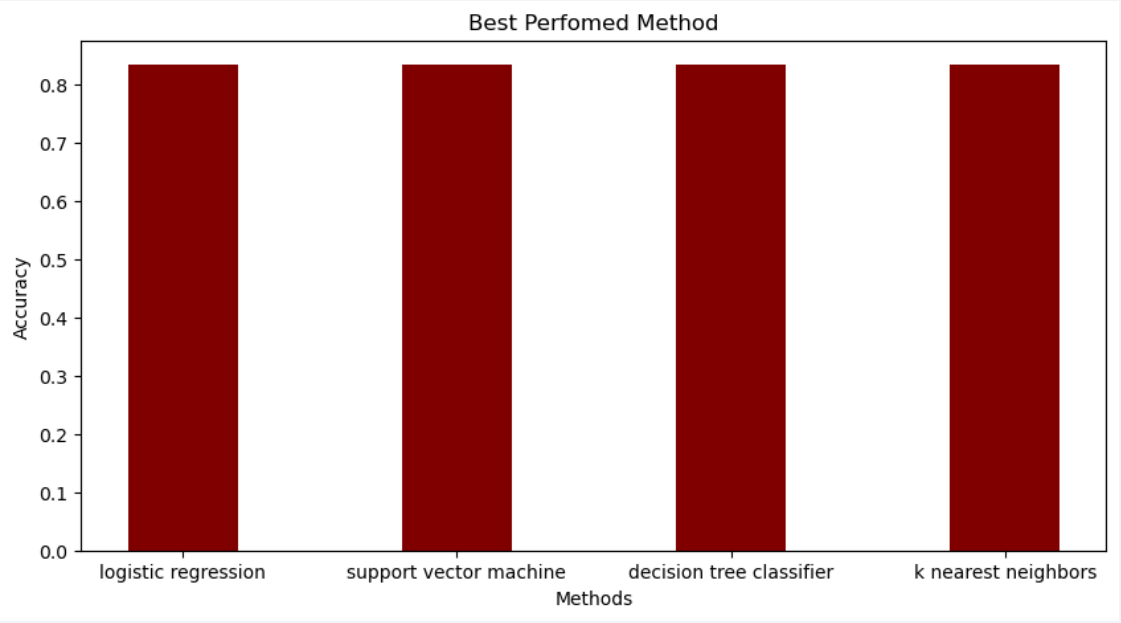


Section 5

Predictive Analysis (Classification)

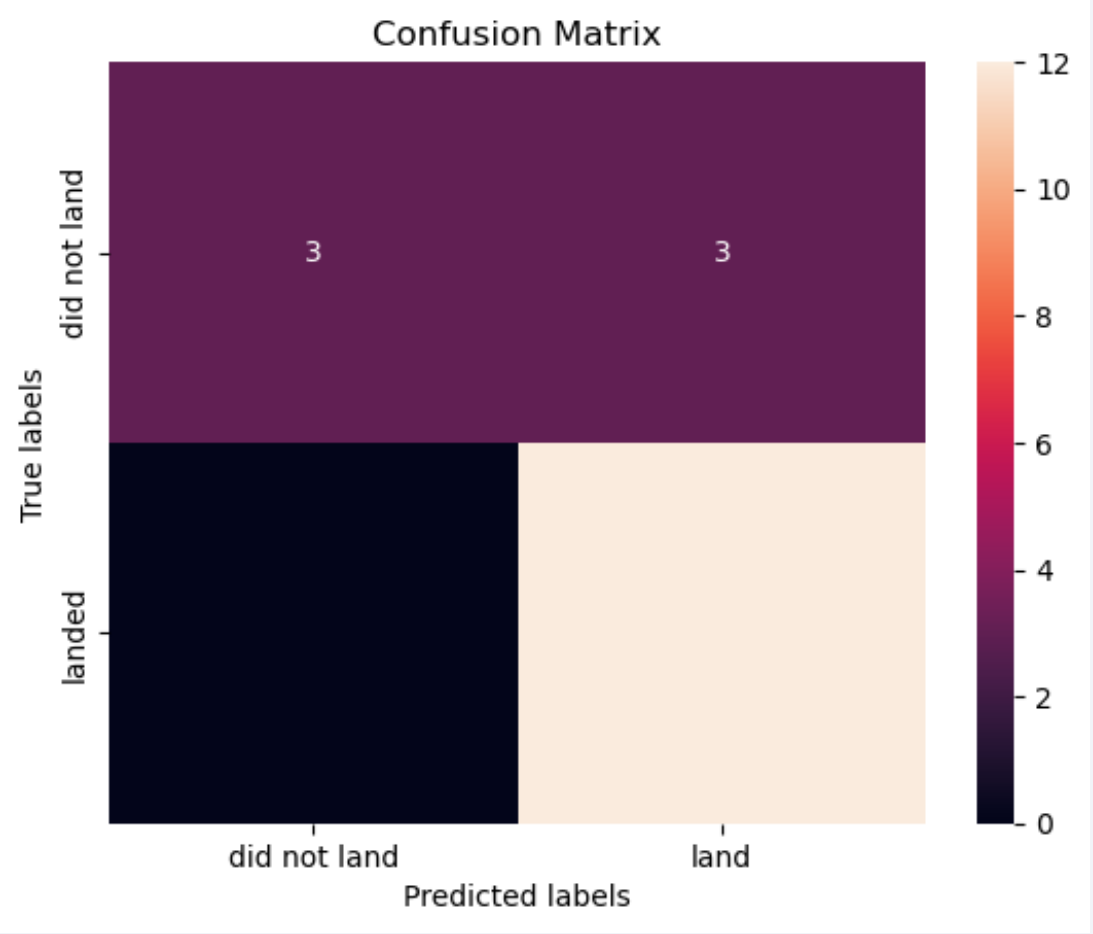
Classification Accuracy

- All models demonstrated nearly identical accuracy on the test set, achieving 83.33%.
- It's crucial to acknowledge the small test size, consisting of only 18 samples, which can contribute to significant variance in accuracy results, as seen in the Decision Tree Classifier model across repeated runs.
- To establish a more definitive determination of the best model, additional data is likely required.



Confusion Matrix

- Given the identical performance of all models on the test set, the confusion matrix is consistent across the board.
- The models accurately predicted 12 successful landings and 3 unsuccessful landings when the true labels matched.
- However, there were instances where the models wrongly predicted 3 successful landings when the actual outcome was unsuccessful (false positives), indicating a tendency for our models to overpredict successful landings.



Conclusions

- Our goal was to create a machine learning model for Blue Origin to predict successful Stage 1 landings, potentially saving around $100 million USD.
- Data from a SpaceX API and Wikipedia were used, labeled, and stored in a SQL database for analysis. A dashboard was developed for visualization, and the resulting machine learning model achieved an 83% accuracy rate.
- This model could assist Jeff Bezos in predicting successful Stage 1 landings before launch, aiding decision-making. However, to refine the model and improve accuracy, collecting more data is advisable for future predictions.

Appendix

- GitHub Capstone Project URL: https://github.com/ScrubsAndStats/IBM-DS-Professional-Labs/tree/main/Capstone

- I am grateful to the instructors for generously sharing their time and knowledge, and to Coursera for providing this valuable opportunity for learning and personal growth.

Thank you!