IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aimboon Wiratsin
26/03/2025

Outline

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

Executive Summary

- Collected data from SpaceX API and Wikipedia to classify successful landings.
- Explored data using SQL, visualizations, Folium maps, and dashboards.
- Selected relevant features and converted categorical variables to binary with one-hot encoding.
- Standardized the data and used GridSearchCV to optimize machine learning models.
- Visualized accuracy scores of all models.

Models Used:

- Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- All models achieved around 83.33% accuracy.
- More data is needed to improve model performance.

Introduction

Project Background & Context

- **SpaceX's Impact**: SpaceX is a pioneering private company that has reduced the cost of rocket launches through innovative reusability (e.g., Falcon 9's reusable first stage).
- **Cost Reduction**: SpaceX charges $62 million per launch, far less than competitors, due to the ability to reuse key components.
- **Significance**: Understanding whether the first stage of a Falcon 9 rocket will successfully land could significantly impact the cost prediction of future missions and provide a competitive advantage.

Problems to Find Answers

- **Predict Landing Success**: Can we predict whether the first stage of a Falcon 9 rocket will successfully land based on mission data?
- **Cost Implications**: How does predicting landing success help in estimating launch costs?
- **Competitive Bidding**: Can this data be used to create a model that competitors could use when bidding against SpaceX for launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

-

Data Collection

Data Sources

- **SpaceX API**: Public API accessed for mission-related data.
- **Wikipedia Web Scraping**: Scraped data from SpaceX's Wikipedia entry, focusing on a table containing launch details.

Data Columns from SpaceX API

- Key columns include **FlightNumber**, **BoosterVersion**, **PayloadMass**, **Orbit**, **LaunchSite**, **Outcome**, **Latitude**, **Longitude**, and others related to mission success and vehicle details.

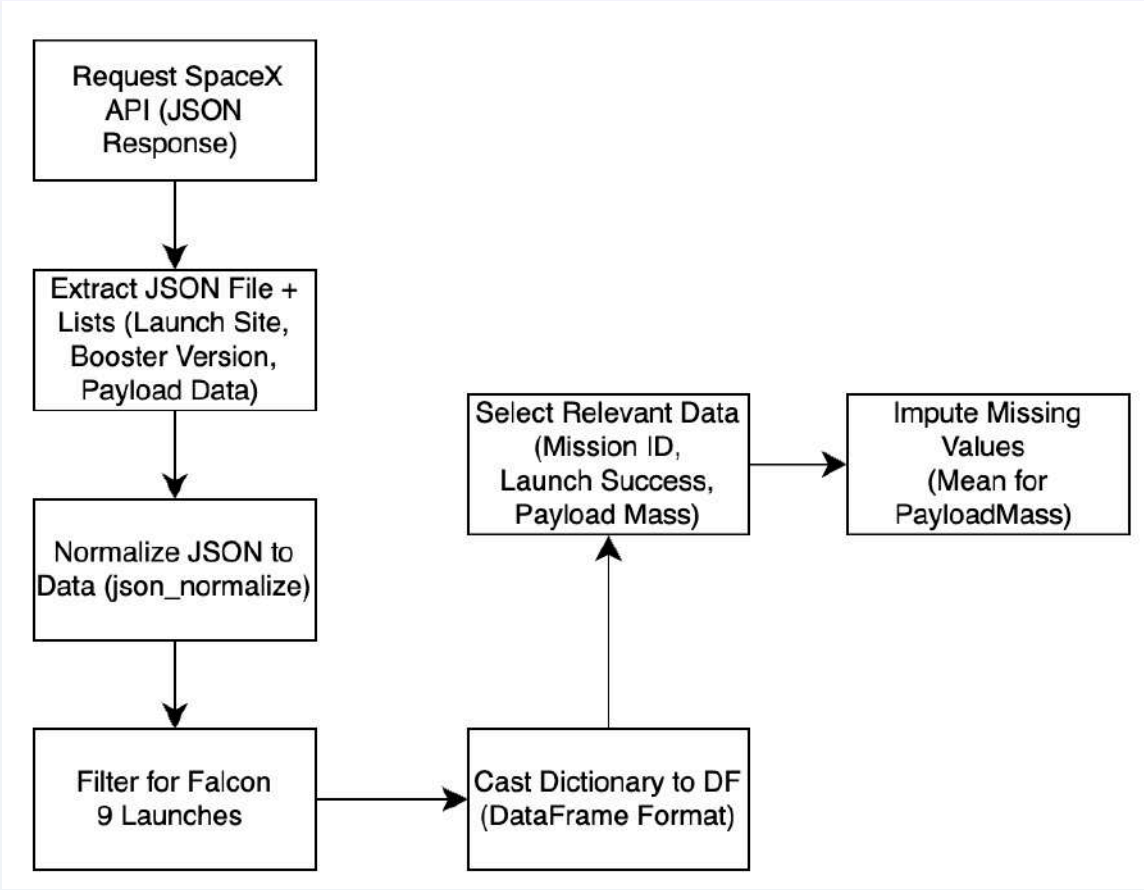
Data Columns from Wikipedia Web Scraping

- Key columns include **Flight No.**, **Launch site**, **Payload**, **PayloadMass**, **Orbit**, **Customer**, **Launch outcome**, and **Booster landing**.

Data Wrangling

- **API Data**: Used **json_normalize** to structure the nested JSON data.
- **Web Scraped Data**: Cleaned and structured the data from the table into a usable format.

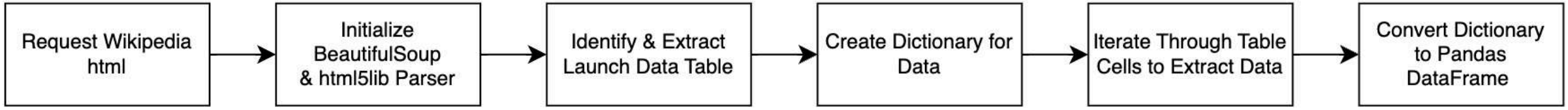
Data Collection – SpaceX API



->Github Data Collection

https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/Data%20Collection%20Api%20.ipynb

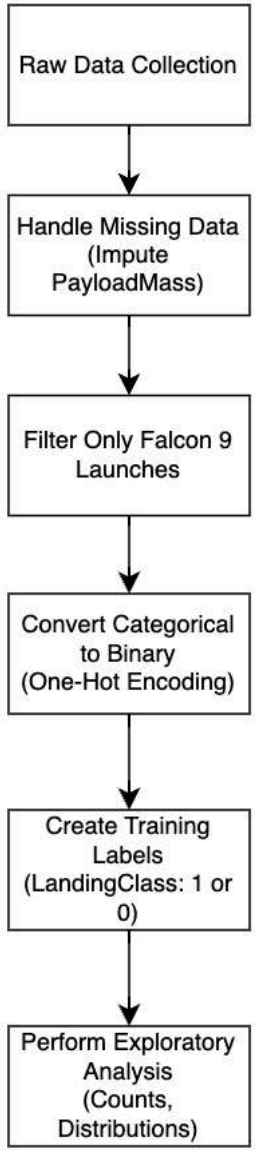
Data Collection - Scraping



- GitHub URL
- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/jupyter-labs-webscraping.ipynb

Data Wrangling

1. **Handling Missing Data**
   - Identified and imputed missing values in **PayloadMass** using the mean.
2. **Filtering Data**
   - Removed irrelevant launches (e.g., non-Falcon 9 missions).
3. **Converting Categorical Data to Numerical Format**
   - Applied **one-hot encoding** to categorical variables such as **LaunchSite** and **Orbit**.
4. **Generating Training Labels**
   - Created a new column **LandingClass**:
     - **1** → Successful landings (**True Ocean, True RTLS, True ASDS**)
     - **0** → Unsuccessful landings (**False Ocean, False RTLS, False ASDS**)
5. **Exploratory Data Analysis (EDA)**
   - Counted the number of launches per site.
   - Analyzed **orbit types** and **landing outcomes**.
   - Visualized distributions using **matplotlib & seaborn**.



- GitHub URL
- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- GitHub URL
- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/edadataviz.ipynb

Flight Number vs. Launch Site (Categorical Plot)

- Shows the frequency of launches at different sites. Helps determine which sites are used most often and if flight experience impacts launch success.

Payload Mass vs. Launch Site (Scatter Plot)

- Examines whether different launch sites handle different payload sizes, revealing any site-specific payload capacity preferences.

Success Rate of Each Orbit Type (Bar Chart)

- Displays how successful each orbit type has been, helping identify which orbits have the highest probability of a successful landing.

Flight Number vs. Orbit Type (Categorical Plot)

- Helps assess whether certain orbits are chosen more frequently over time and whether flight experience impacts orbit selection.

Payload Mass vs. Orbit Type (Scatter Plot)

- Explores the relationship between payload mass and orbit type, identifying if specific orbits accommodate heavier or lighter payloads.

Launch Success Yearly Trend (Line Chart)

- Shows how SpaceX's launch success rate has changed over time, providing insights into improvements in landing technology.

EDA with SQL

- summarize the SQL queries performed

- Loaded data set into IBM DB2 Database.

- Queried using SQL Python integration.

- Queries were made to get a better understanding of the dataset.

- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

- GitHub URL

- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium maps highlight launch sites, depict successful and failed landings, and showcase proximity to key locations like railways, highways, coastlines, and cities.
- This provides insight into the reasoning behind launch site selection and visualizes landing success in relation to geographical factors.
- GitHub URL
- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The dashboard features both a pie chart and a scatter plot.

- The pie chart displays the distribution of successful landings across all launch sites or, alternatively, the success rate of a specific site.
- The scatter plot allows users to select either all launch sites or a specific site while adjusting payload mass using a slider ranging from 0 to 10,000 kg.
- The pie chart provides a visual representation of launch site success rates.
- The scatter plot helps analyze how success is influenced by launch site, payload mass, and booster version category.

- GitHub URL
- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

Separate Target Variable – Extract the 'Class' column as the label from the dataset.

Feature Scaling – Apply StandardScaler to normalize the feature values.

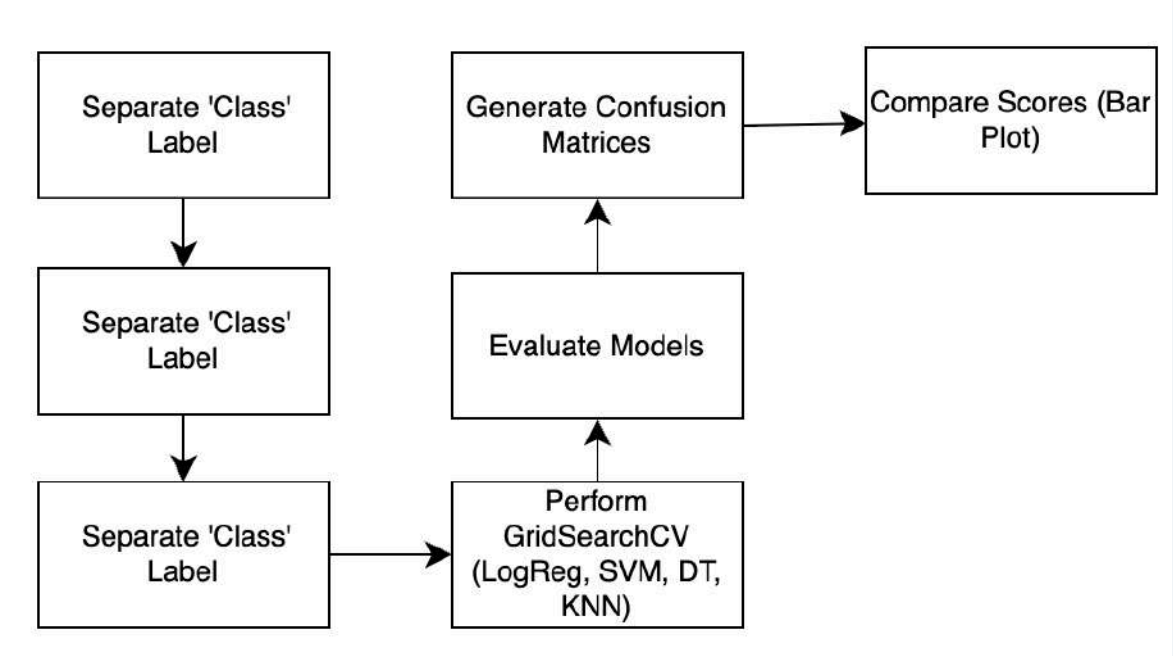
Split Data – Use train_test_split to divide the dataset into training and testing sets.

Hyperparameter Tuning – Perform GridSearchCV on Logistic Regression, SVM, Decision Tree, and KNN models to find the best parameters.

Model Evaluation – Assess model performance using the test set.

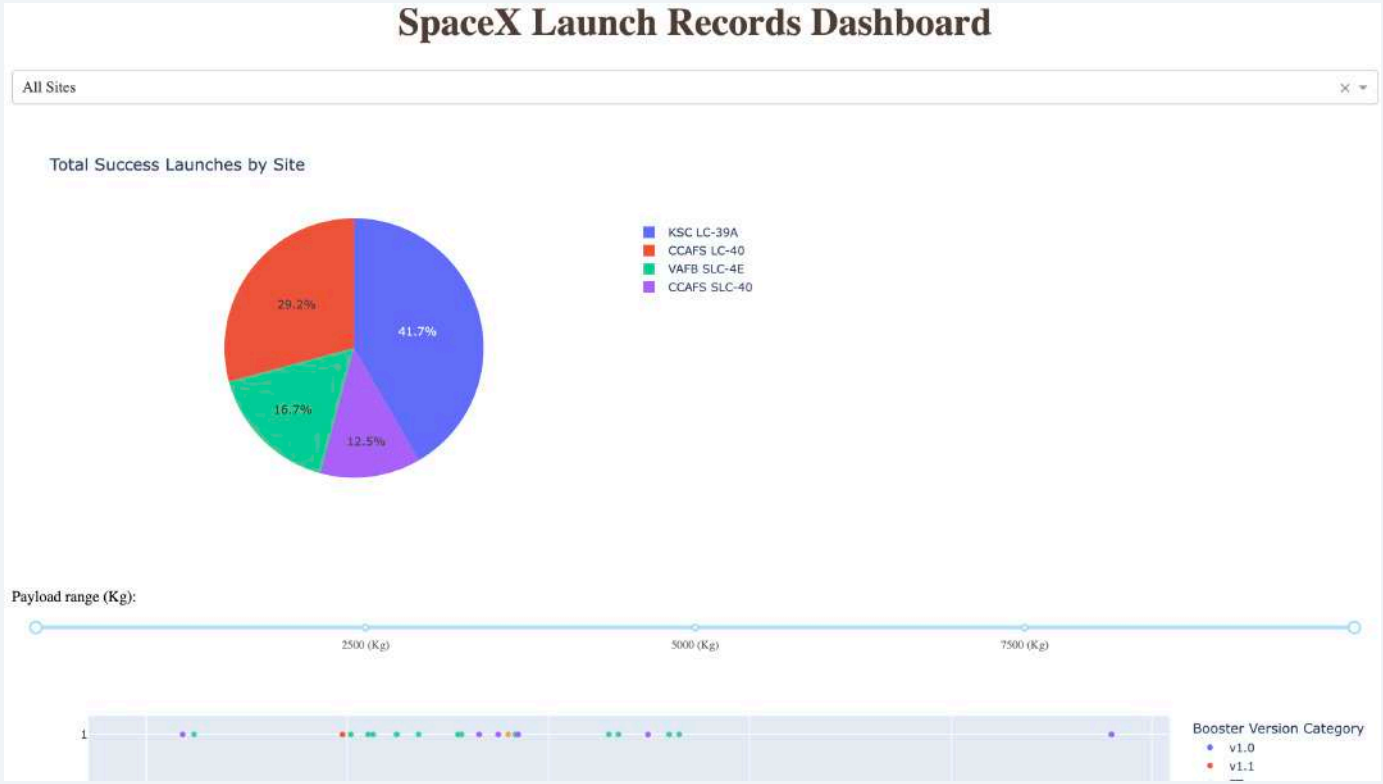
Confusion Matrix Analysis – Generate confusion matrices for all models to evaluate prediction accuracy.

Performance Comparison – Create a bar plot to compare model scores.



- Add the GitHub URL
- https://github.com/AimboonWir/data-science-notebook/blob/main/10.Applied_Data_Science_Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

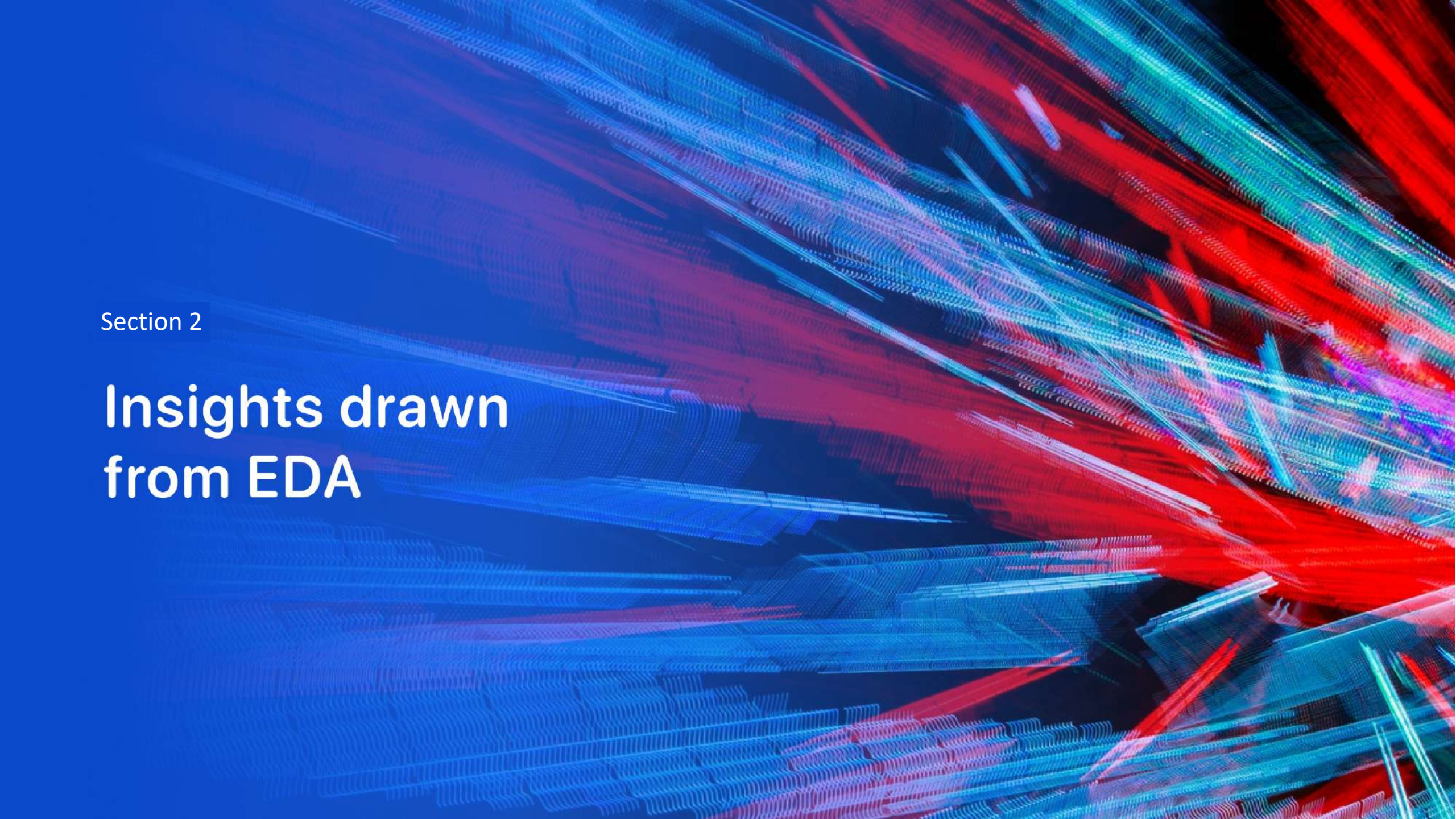
Results



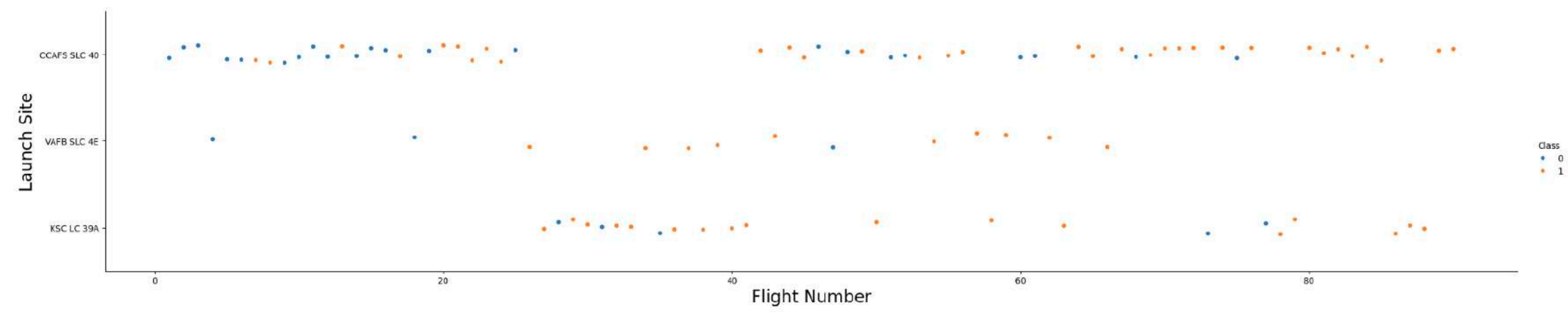
- This is a preview of the Plotly dashboard. The upcoming slides will present the results of Exploratory Data Analysis (EDA) using visualizations, EDA with SQL, an interactive map created with Folium, and the final model results, which achieve approximately 83% accuracy.

Section 2

Insights drawn from EDA



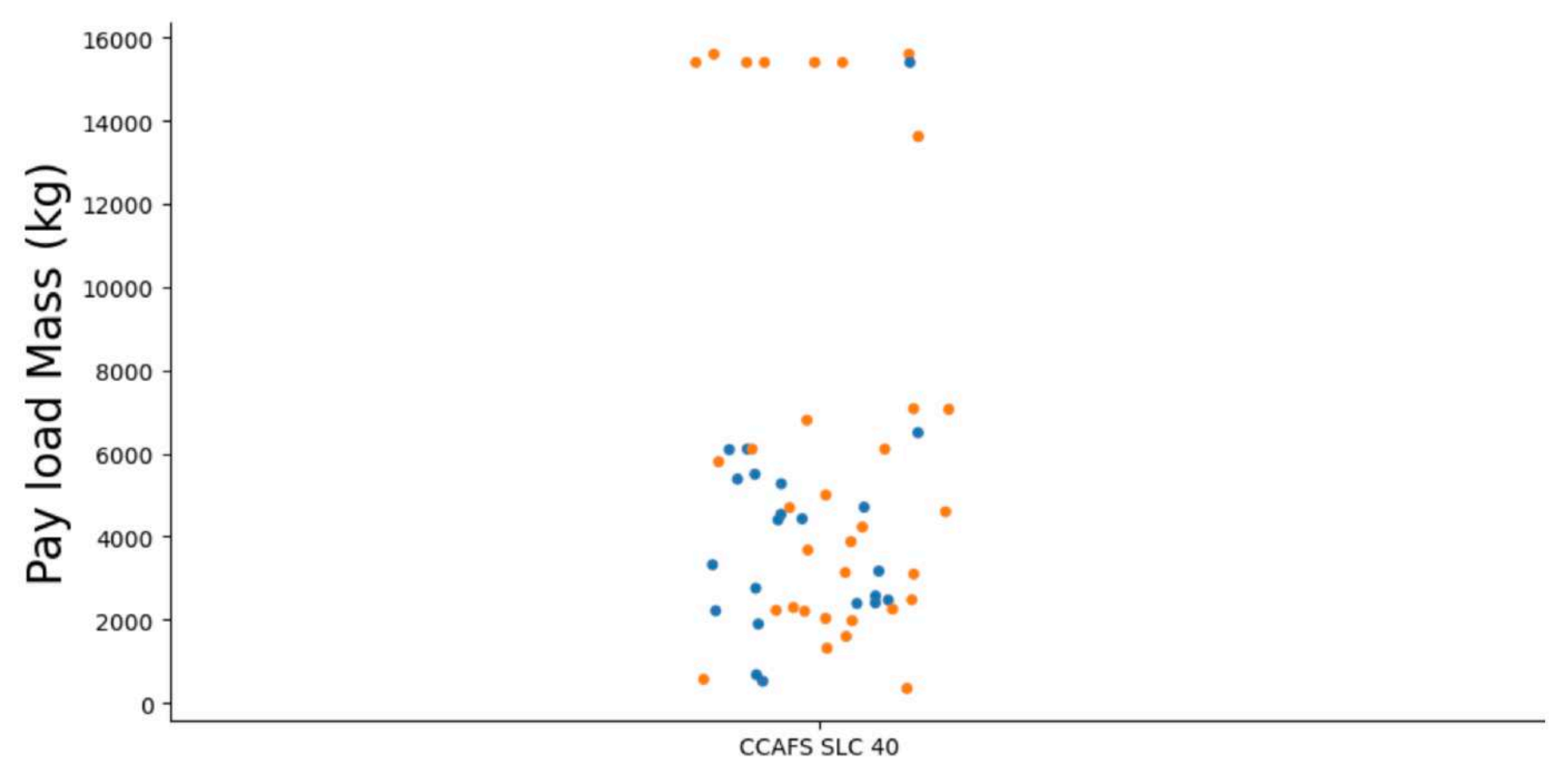
Flight Number vs. Launch Site



- Orange : successful launch
- Blue : unsuccessful launch

The graph indicates a rising success rate over time, as shown by the Flight Number. A major breakthrough seems to have occurred around the 20th flight, leading to a significant improvement in success rates. Additionally, CCAFS appears to be the primary launch site, handling the highest number of launches.

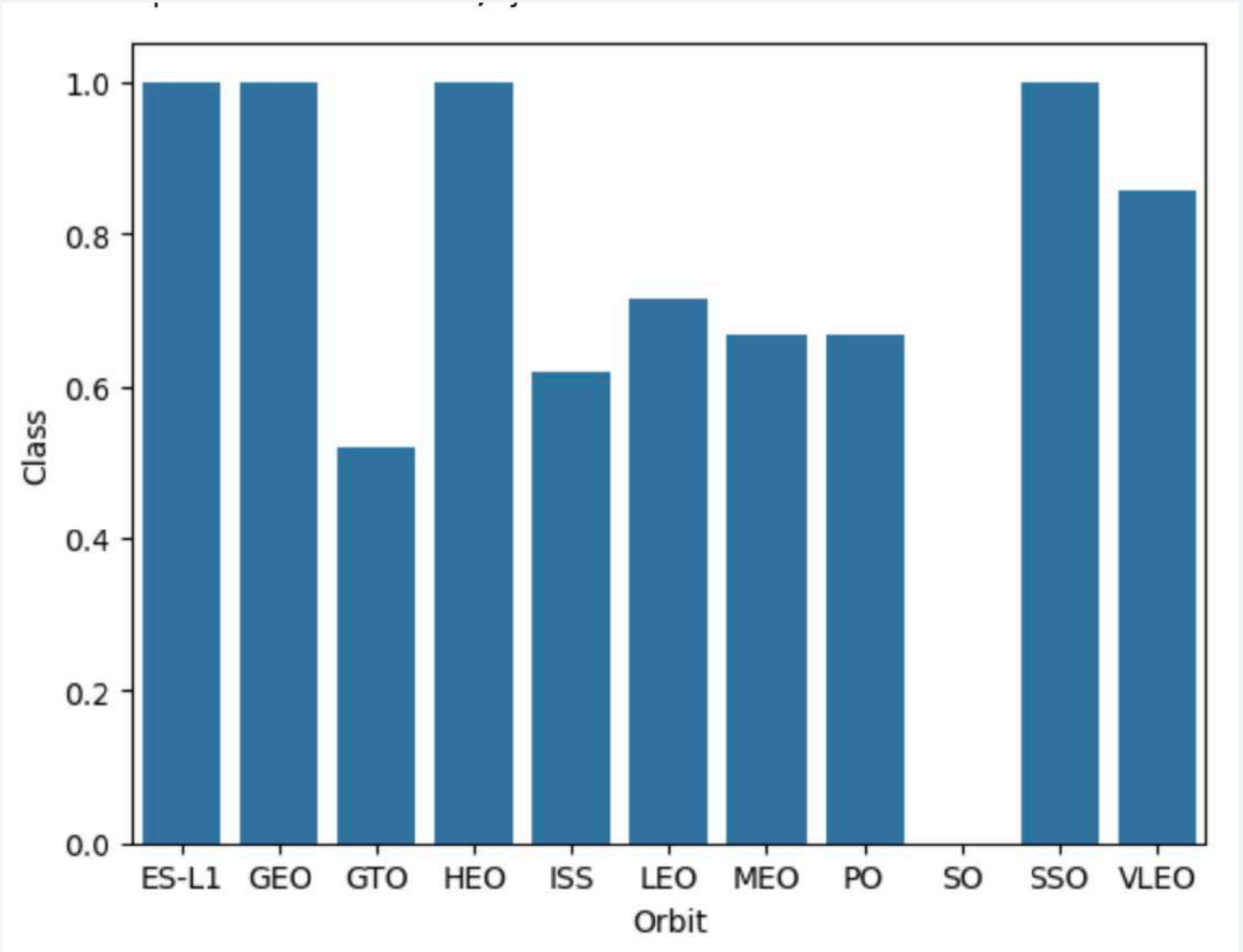
Payload vs. Launch Site



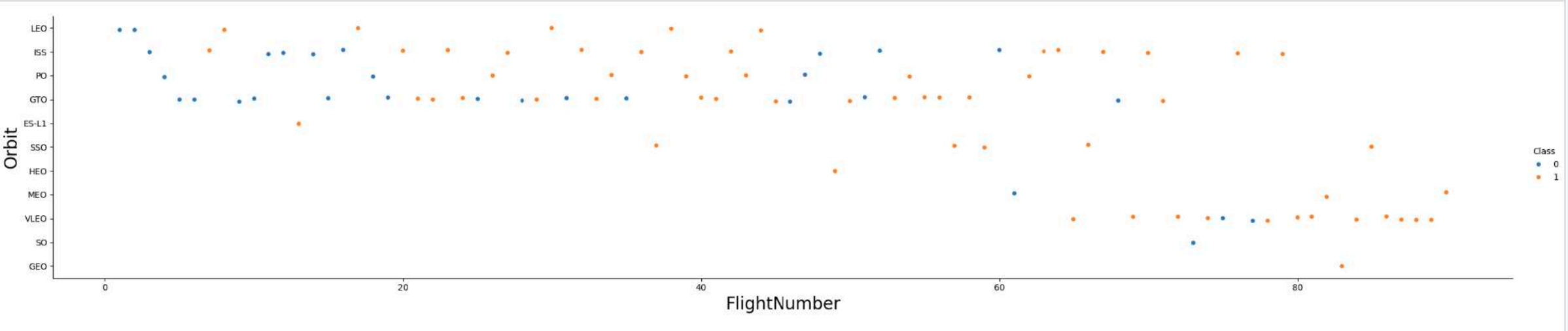
Different launch sites also seem to use different payload mass.

Success Rate vs. Orbit Type

- ES-L1 (1), GEO (1), and HEO (1) achieved a 100% success rate (sample sizes in parentheses).
- SSO (5) also maintained a perfect success rate.
- VLEO (14) demonstrated a relatively strong success rate with multiple attempts.
- SO (1) had a 0% success rate, while GTO (27) had the highest number of attempts but only around a 50% success rate.



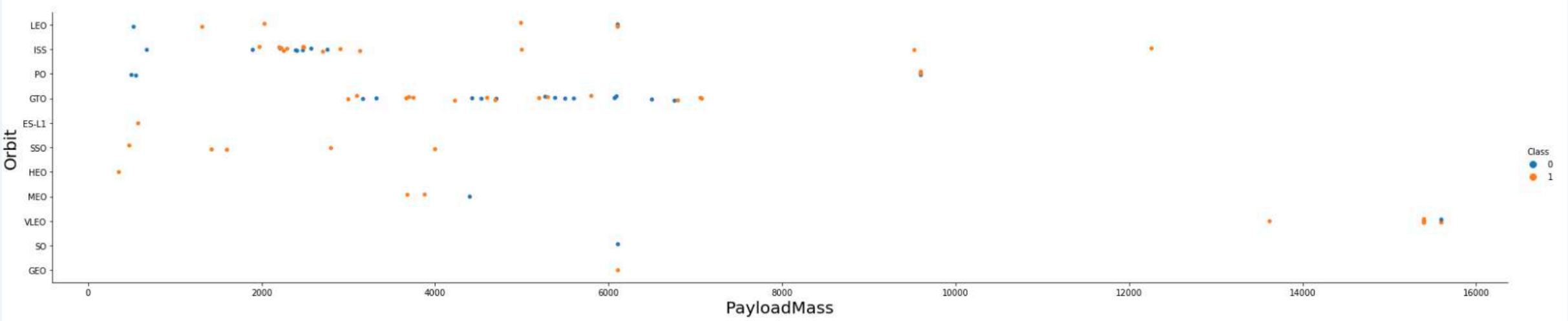
Flight Number vs. Orbit Type



- Orange : successful launch
- Blue : unsuccessful launch

The preferred launch orbits have evolved over the course of SpaceX's flight history, with launch outcomes appearing to be influenced by these changes. Initially, SpaceX focused on LEO orbits, achieving moderate success. In more recent missions, they have shifted back to VLEO. Overall, SpaceX seems to achieve better performance in lower orbits and Sun-synchronous orbits.

Payload vs. Orbit Type

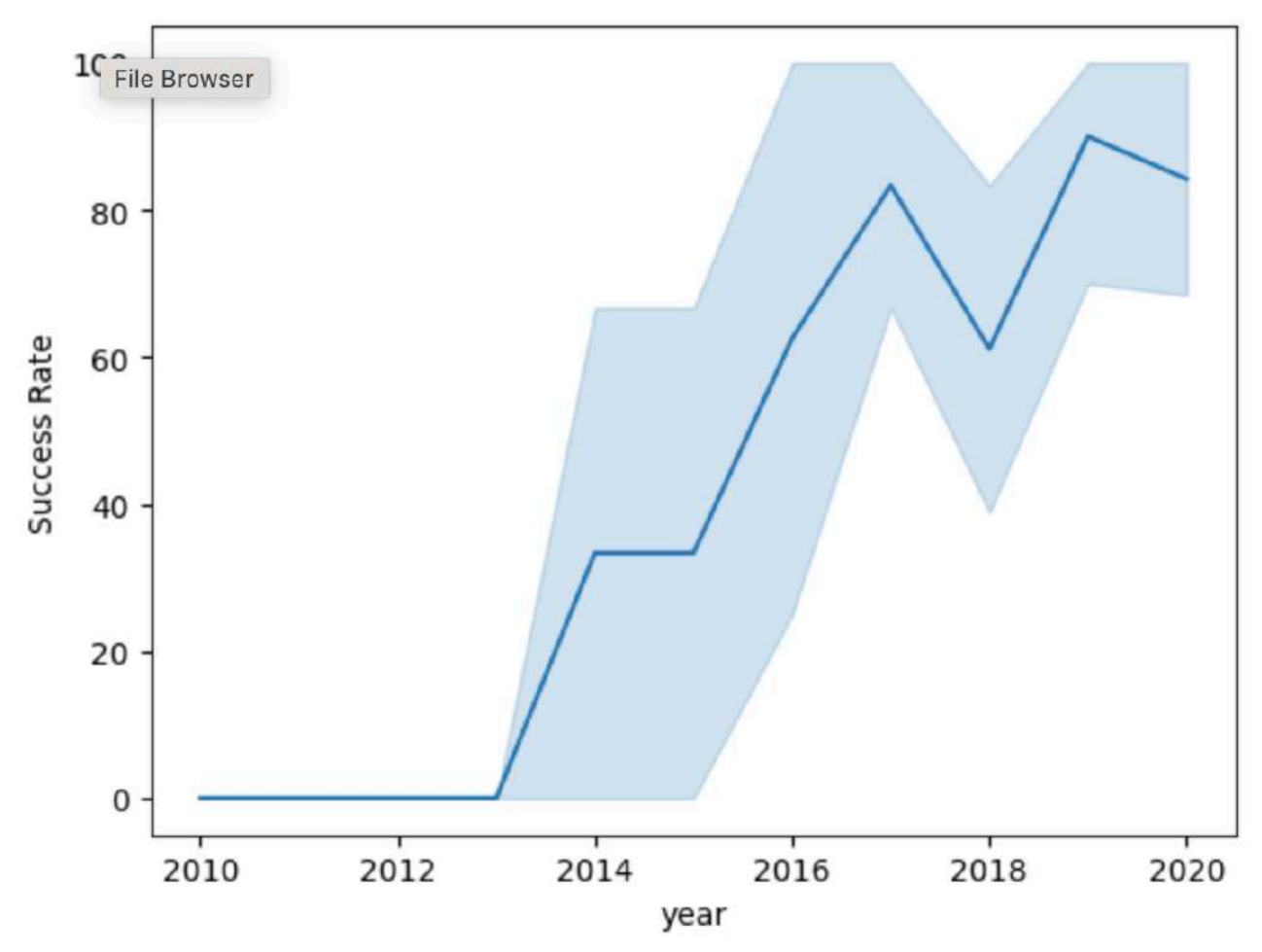


- Orange : successful launch
- Blue : unsuccessful launch

  - There appears to be a relationship between payload mass and orbit type. LEO and SSO generally have lower payload masses, while VLEO, one of the most successful orbits, only features payload masses at the higher end of the range.

Launch Success Yearly Trend

- Overall, success has gradually improved since 2013, with a small decline in 2018. In recent years, the success rate has stabilized around 80%.



All Launch Site Names

```
[12]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[12]:

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

- Retrieve distinct launch site names from the database.
- The entries "CCAFS SLC-40" and "CCAFSSLC-40" are probably referring to the same launch site, with potential data entry errors.
- "CCAFS LC-40" was its former name.

Launch Site Names Begin with 'CCA'

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

[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcor
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachu
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 (parachu
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atten
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atten
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atten

Total Payload Mass

```
[14]: %sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
[14]:  sum
       45596
```

- This query calculates the total payload mass in kilograms for payloads where NASA was the customer. "CRS" refers to Commercial Resupply Services, which means these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
[15]: %sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
[15]:
```

Average
2534.6666666666665

- This query computes the average payload mass for launches that used the F9 v1.1 booster version. The average payload mass of the F9 v1.1 is on the lower end of the payload mass range.

First Successful Ground Landing Date

```
[16]: %sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
 * sqlite:///my_data1.db
Done.
[16]:
```

Date
2010-06-04

- This query retrieves the date of the first successful ground pad landing. The first successful ground pad landing occurred at the end of 2015, while successful landings in general began around 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[28]: %%sql
SELECT booster_version
FROM SPACEXTBL
WHERE mission_outcome = 'Success'
AND payload_mass__kg_ BETWEEN 4000 AND 6000
AND landing_outcome = 'Success (drone ship)';
```

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

[28]:

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

- This query retrieves the four booster versions that successfully landed on a drone ship and had a payload mass between 4000 and 6000 kg, excluding 6000.

Total Number of Successful and Failure Mission Outcomes

```
[36]: %sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
 * sqlite:///my_data1.db
Done.
```

[36]:

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

This query provides a count of each mission outcome. SpaceX seems to successfully complete its missions approximately 99% of the time, with most landing failures being intentional. Notably, one launch has an unclear payload status, and unfortunately, one failed during flight.

Boosters Carried Maximum Payload

```
[41]: %%sql
      SELECT booster_version
      FROM SPACEXTBL
      WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL);
```

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

[41]:

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 returns the booster versions that carried the maximum payload mass of 15,600 kg. These booster versions are all quite similar, belonging to the F9 B5 B10xx.x series. This suggests a correlation between payload mass and the specific booster version used.

2015 Launch Records

[44]:
```
%%sql
SELECT substr(DATE, 6, 2) AS Month, landing_outcome, booster_version, launch_site
FROM SPACEXTBL
WHERE substr(DATE, 0, 5) = '2015'
AND landing_outcome LIKE 'Failure (drone ship)';
```

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

[44]:

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 retrieves the month, landing outcome, booster version, payload mass (kg), and launch site for the 2015 launches where stage 1 failed to land on a drone ship. There were two such instances.

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

[47]:
```
%%sql
SELECT landing_outcome, COUNT(*) AS count
FROM SPACEXTBL
WHERE Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP BY landing_outcome
ORDER BY count DESC;
```

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

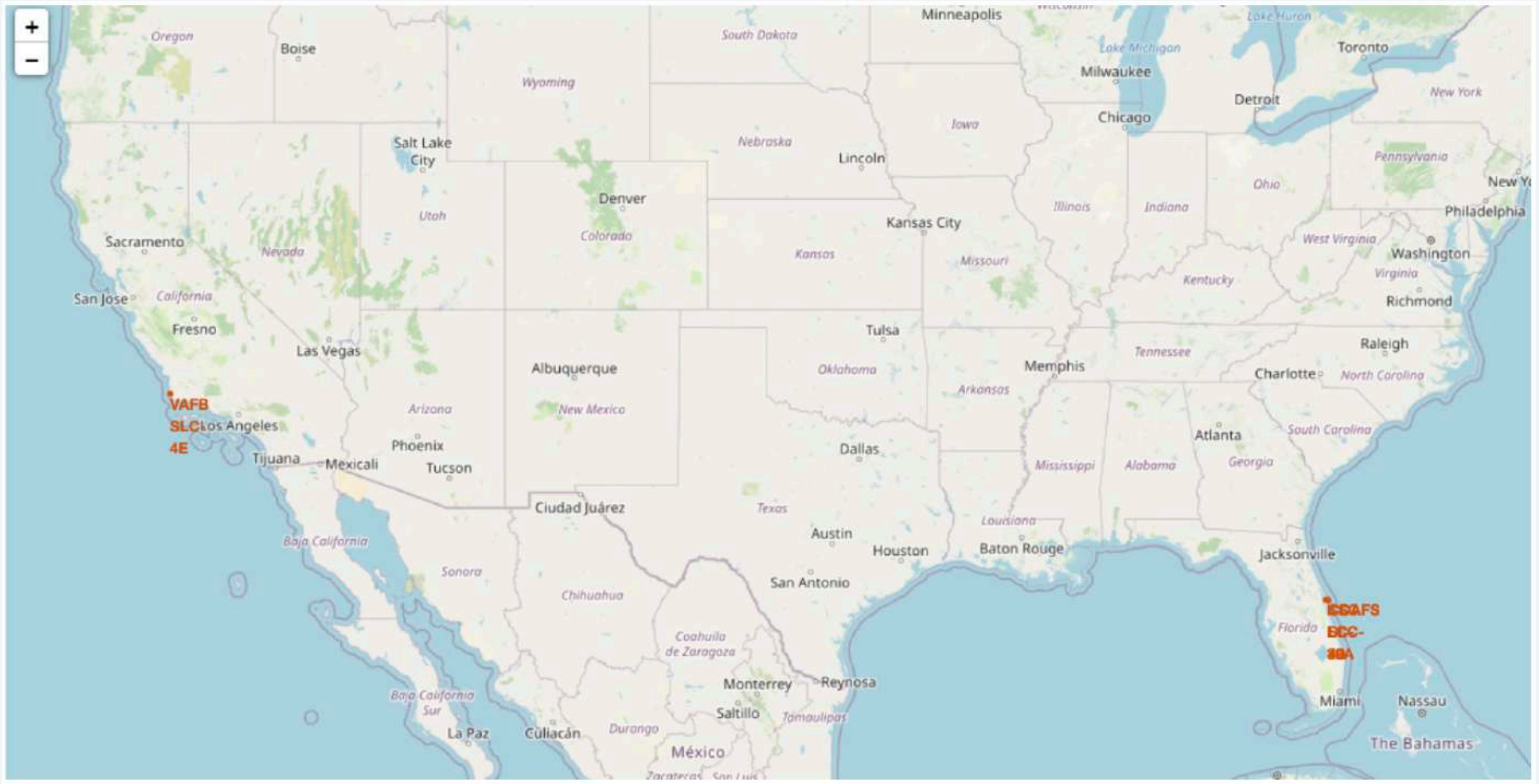
[47]:

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

- This query provides a list of successful landings between 2010-06-04 and 2017-03-20, inclusive. There are two types of successful landing outcomes: drone ship and ground pad landings. A total of 8 successful landings occurred during this period.

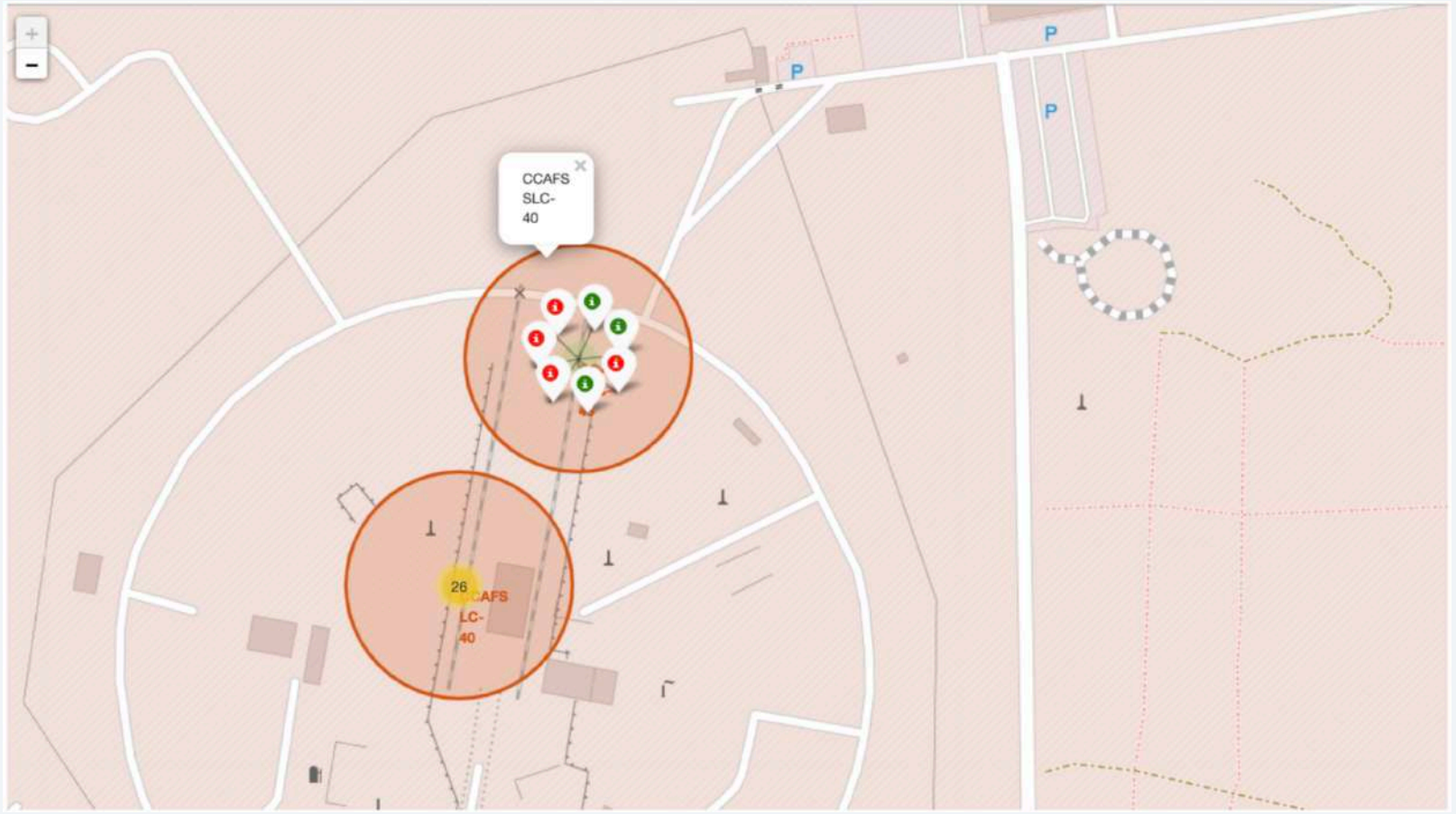

Section 3
Launch Sites
Proximities Analysis

<Folium Map Screenshot 1>

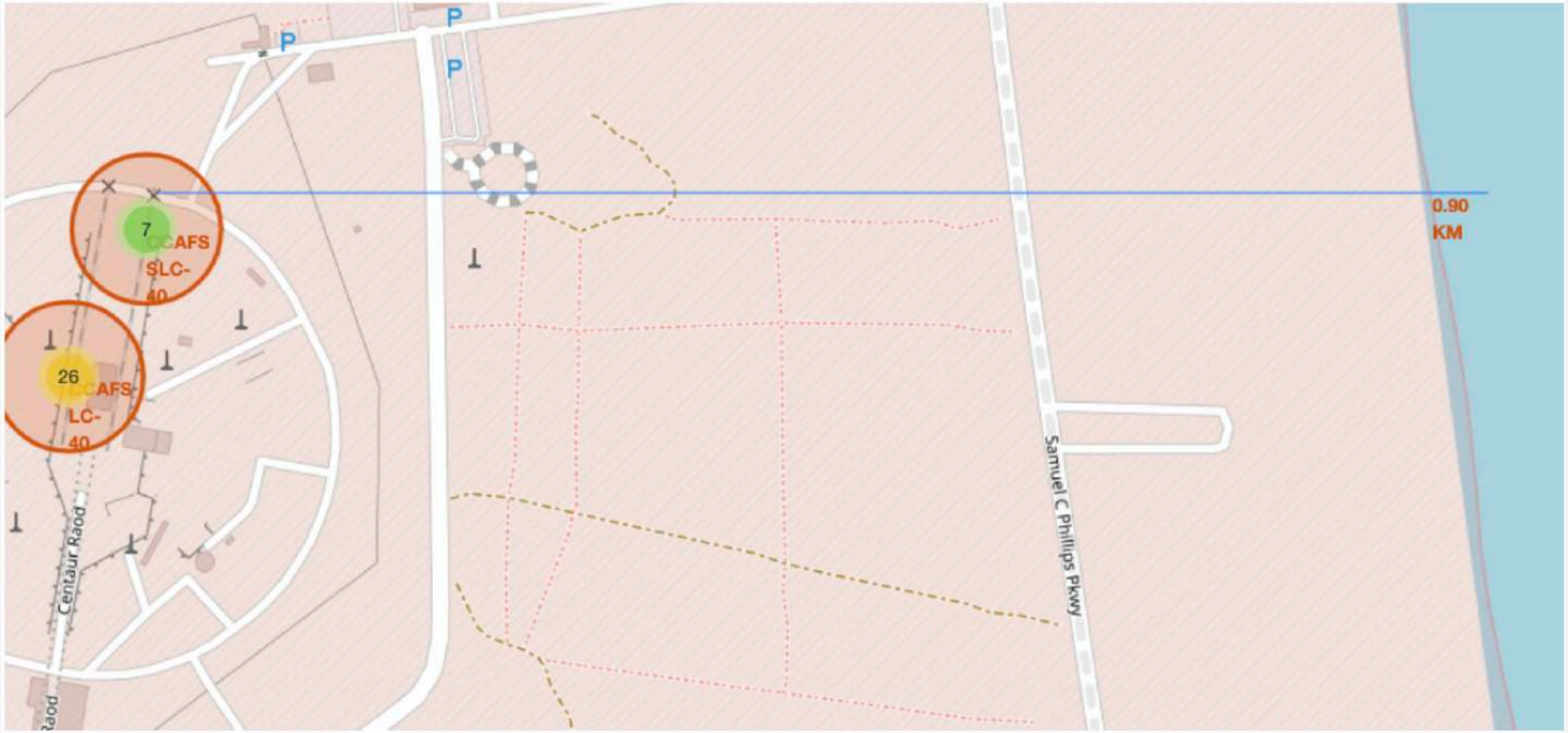


- all launch sites relative US map

<Folium Map Screenshot 2>

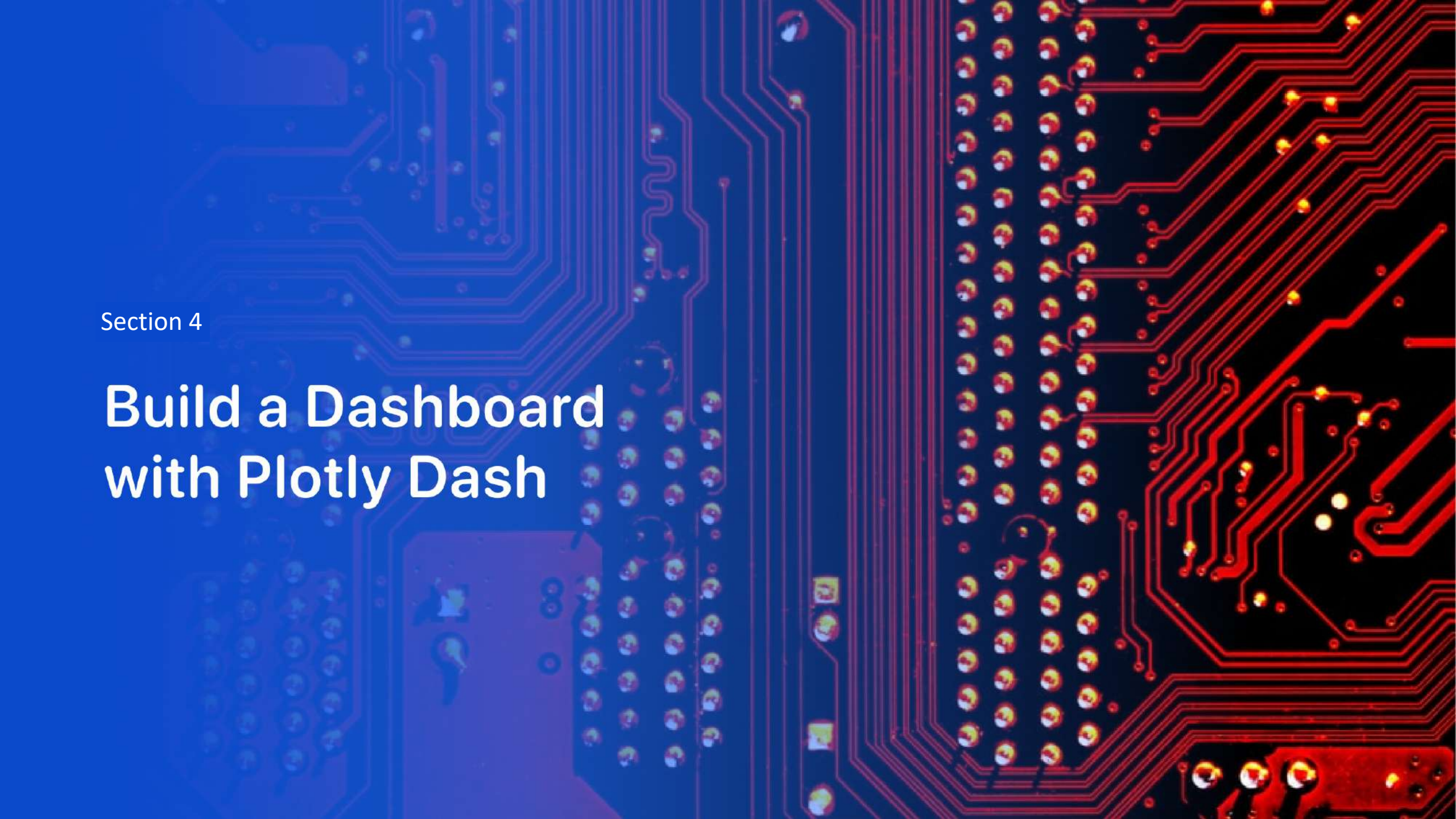


<Folium Map Screenshot 3>

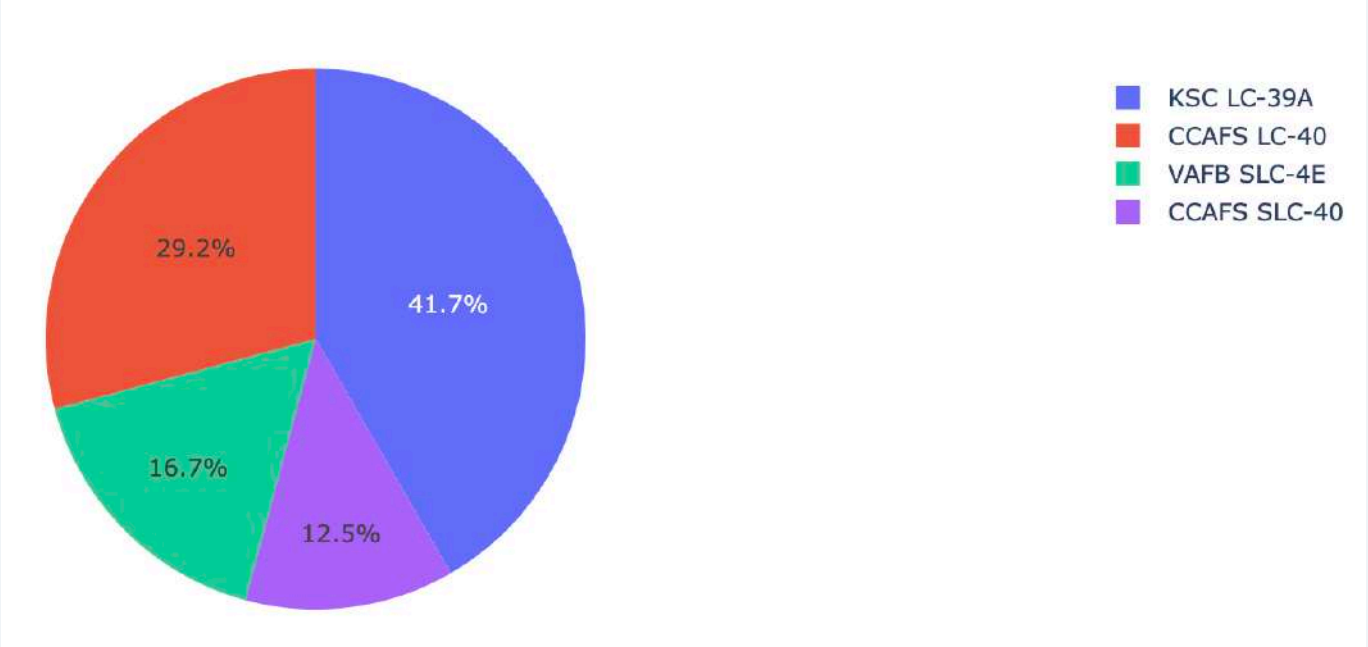


Section 4

Build a Dashboard with Plotly Dash

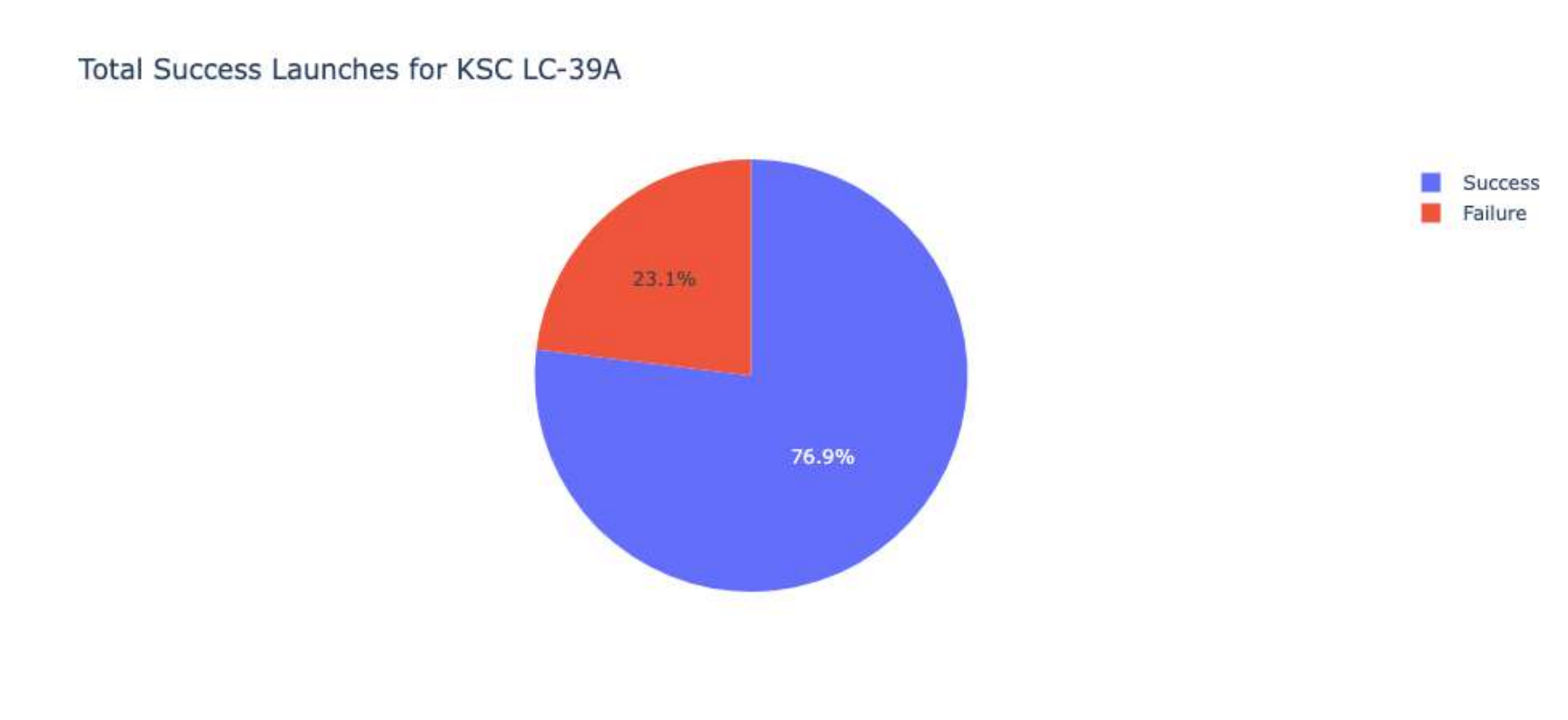


<Dashboard Screenshot 1>



- This shows the distribution of successful landings across all launch sites. CCAFS LC-40, the previous name for CCAFS SLC-40, means that CCAFS and KSC have an equal number of successful landings, though most of the successful landings occurred before the name change. VAFB has the fewest successful landings, likely due to a smaller sample size and the increased difficulty of launching from the West Coast.

<Dashboard Screenshot 2>



- KSC LC-39A boasts the highest success rate, with 10 successful landings and 3 failures.

<Dashboard Screenshot 3>

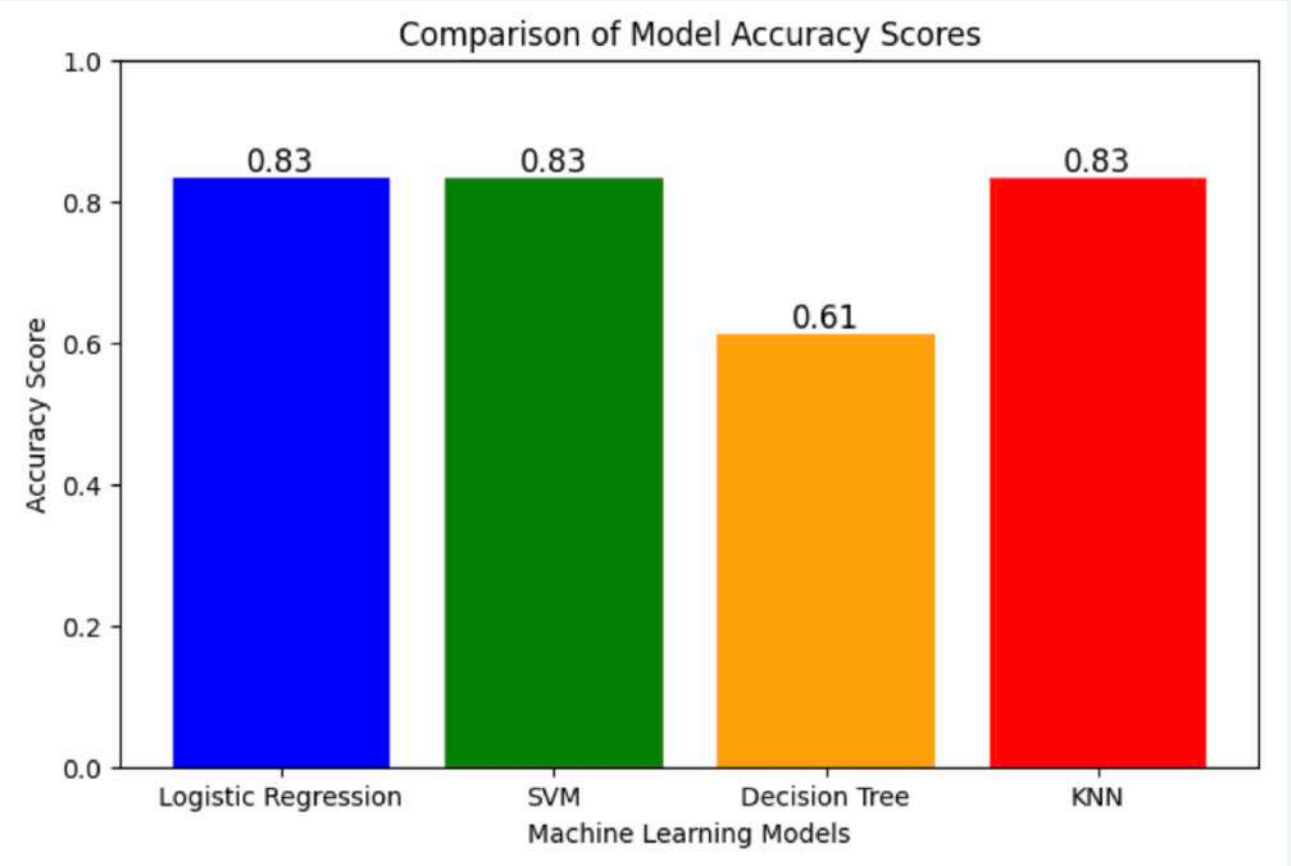
- The Plotly dashboard includes a Payload range selector, but it is set between 0 and 10,000, rather than the maximum payload of 15,600 kg. The 'Class' variable indicates success with 1 and failure with 0. The scatter plot also uses color to represent the booster version category and adjusts the point size based on the number of launches. In the payload range of 0-6,000 kg, it is notable that two failed landings have a payload of 0 kg.

Section 5

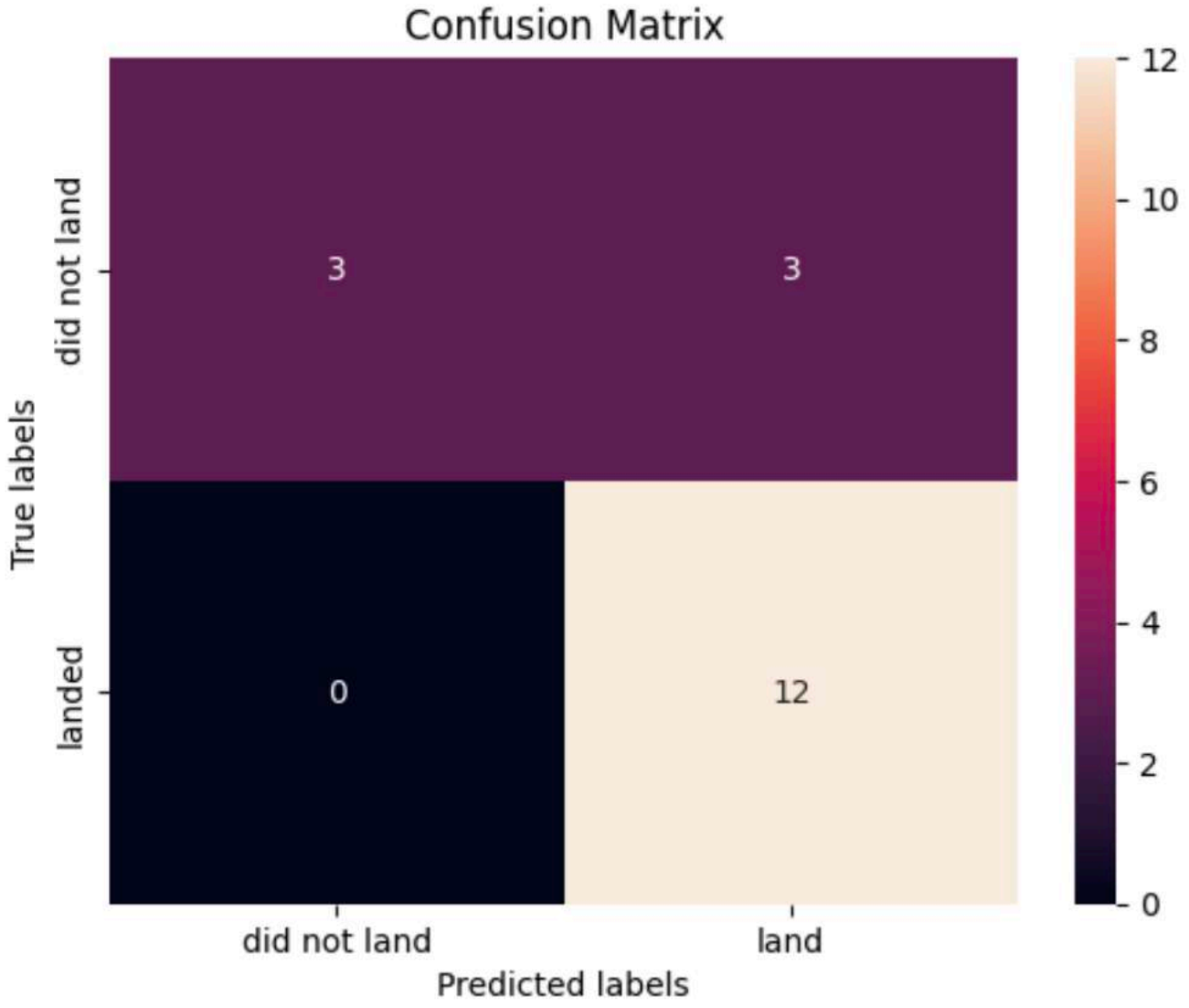
Predictive Analysis (Classification)

Classification Accuracy

- All models achieved nearly the same accuracy on the test set, with an accuracy of 83.33%, except for the decision tree, which had an accuracy of 61%. It's important to note that the test size is small, with only 18 samples, which can lead to significant variability in accuracy results, as seen in the Decision Tree Classifier model across repeated runs. More data is likely needed to accurately determine the best model.



Confusion Matrix



Since all models yielded the same results for the test set, the confusion matrix is identical across them. The models predicted 12 successful landings when the actual outcome was a successful landing. They predicted 3 unsuccessful landings when the actual outcome was unsuccessful. However, the models also predicted 3 successful landings when the true label was unsuccessful (false positives). This indicates that the models tend to over-predict successful landings.

Conclusions

Our task was to develop a machine learning model for Space Y, aiming to compete with SpaceX. The model's goal is to predict when Stage 1 will successfully land, potentially saving around $100 million USD. We used data sourced from a public SpaceX API and web scraping of SpaceX's Wikipedia page. Data labels were created and stored in a DB2 SQL database. Additionally, a dashboard was developed for visualization.

The machine learning model achieved an accuracy of 83%. Allon Mask of SpaceY can utilize this model to predict with reasonable accuracy whether a launch will result in a successful Stage 1 landing before the launch, helping decide whether the launch should proceed.

To further improve accuracy and refine the model, additional data collection is recommended.

Appendix

- GitHub Repository URL:
- https://github.com/AimboonWir/data-science-notebook/tree/main/10.Applied_Data_Science_Capstone

Thank you!