IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Markella Lyra
6/5/2025

Outline

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

Executive Summary

- **Summary of methodologies**

1. Data Collection through API
2. Data Collection with Web Scraping
3. Data Wrangling
4. Exploratory Data Analysis with SQL
5. Exploratory Data Analysis with Data Visualization
6. Interactive Visual Analytics with Folium
7. Machine Learning Prediction

- **Summary of all results**

1. Exploratory Data Analysis result
2. Interactive analytics in screenshots
3. Predictive Analytics result from Machine Learning Lab

Introduction

- SpaceX has revolutionized the aerospace sector by drastically lowering launch costs, offering missions with the Falcon 9 rocket at just $62 million, compared to traditional providers charging over $165 million. This price drop is largely due to their innovation in reusing the first-stage booster.

- As data scientists in a startup competing with SpaceX, our mission is to construct a machine learning model that can forecast the success of first-stage landings. Doing so will allow us to better strategize pricing and improve competitive positioning.

- Problem-questions:

- What factors influence landing outcomes?
- How do these variables interact?
- What conditions maximize landing success?

Section 1

Methodology

Methodology

Executive Summary

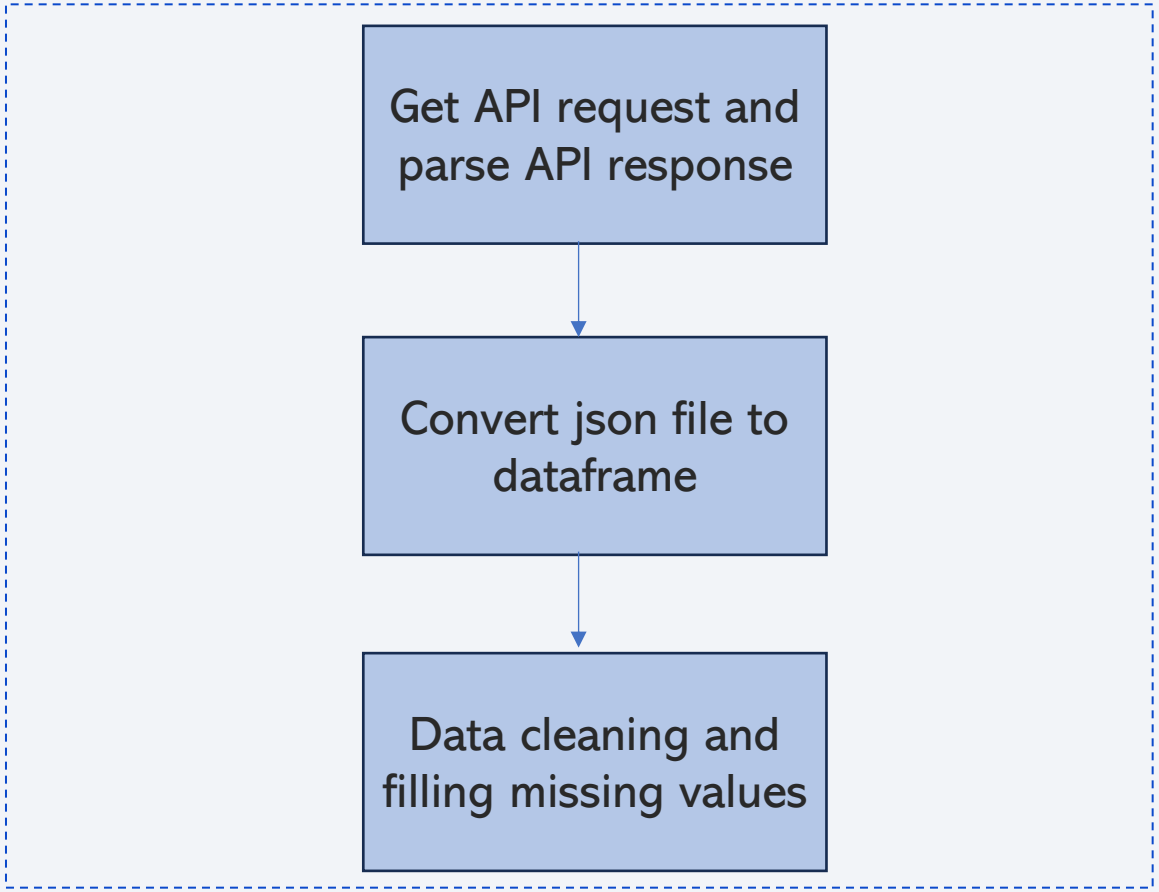
- Data collection methodology:
  - Data was collected from SpaceX REST API and web scrapping from Wikipedia
- Perform data wrangling
  - Data was processed using one-hot encoding for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built, tuned and evaluated the classification models in order to achieve the best results possible.

Data Collection

- Data collection involves systematically gathering and measuring information related to specific variables, enabling meaningful analysis and informed decision-making. In this project, data was sourced from the SpaceX REST API and Wikipedia through web scraping.
- For the REST API, we initiated the process with a GET request. The JSON response was parsed and converted into a Pandas DataFrame using the `json_normalize()` function. Afterward, the dataset was cleaned and inspected for missing values, which were addressed accordingly.
- For web scraping, we used the BeautifulSoup library to extract launch data from an HTML table on the Falcon 9 Wikipedia page. The extracted content was then parsed and transformed into a Pandas DataFrame for further analysis.

Data Collection – SpaceX API

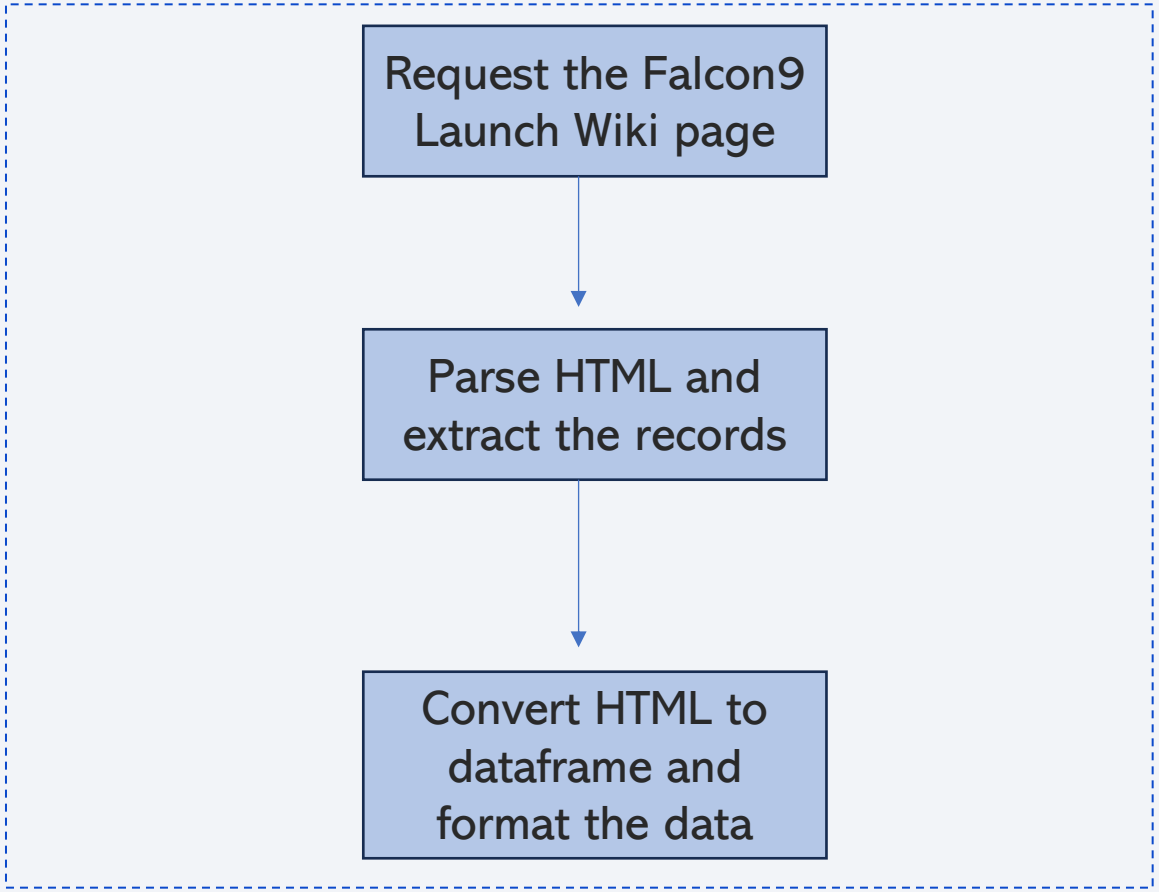
- The data collection for SpaceX API procedure followed the methodology presented in the flowchart.



- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- The data collection for scraping procedure followed the methodology presented in the flowchart.



- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/jupyter-labs-webscraping.ipynb

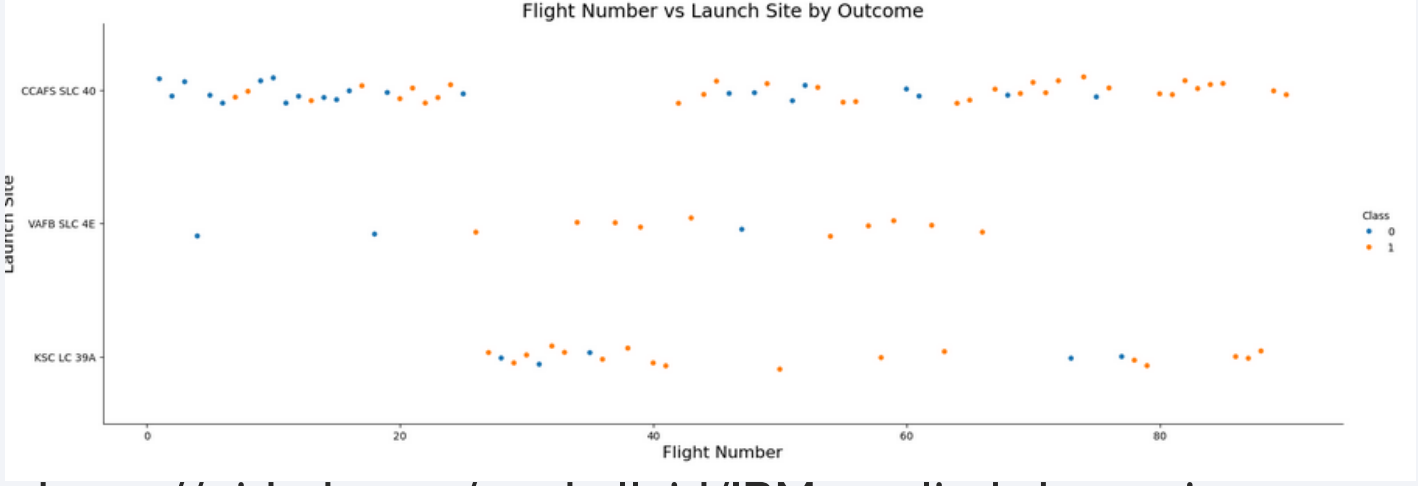
Data Wrangling

- **Data wrangling** is the process of cleaning, organizing, and transforming raw data into a structured format suitable for analysis and exploration.
- Our first step involves calculating the total number of launches per launch site, followed by analyzing the frequency of mission outcomes for each orbit type.
- Next, we generate a new column to categorize landing outcomes, which simplifies future visualizations and modeling tasks.
- Finally, the refined dataset is exported to a CSV file for use in further stages of the project.

- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We began our analysis by creating scatter plots to investigate the relationships between various attributes. This included examining how payload correlates with flight number, how flight number relates to launch site, the connection between payload and launch site, the relationship between flight number and orbit type, and how payload varies across different orbit types. These visualizations helped uncover potential patterns and dependencies that could influence the success of rocket landings.



- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/edadataviz.ipynb

EDA with SQL

- The SQL queries performed are as follow:

- Displaying the names of the launch sites.

- Displaying 5 records where launch sites begin with the string 'CCA'.

- Displaying the total payload mass carried by booster launched by NASA (CRS).

- Displaying the average payload mass carried by booster version F9 v1.1.

- Listing the date when the first successful landing outcome in ground pad was achieved.

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

- Listing the total number of successful and failure mission outcomes.

- Listing the names of the booster_versions which have carried the maximum payload mass.

- Listing the failed landing_outcomes in drone ship, their booster versions, and launch sites names for in year 2015.

- Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In our use of Folium maps, several visual elements were employed to enhance interactivity and clarity. Markers were placed to represent specific points of interest, such as launch sites. Circles were used to highlight particular geographic areas, for instance, around the NASA Johnson Space Center. Marker clusters grouped multiple events occurring at the same coordinates, such as repeated launches from the same site. Lastly, lines were drawn to illustrate distances between two points, helping to contextualize spatial relationships.

- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was developed using Plotly Dash, enabling users to explore and interact with the data dynamically. The dashboard includes pie charts that display the total number of launches across various sites. Additionally, scatter plots were created to illustrate the relationship between launch outcomes and payload mass (in kilograms) for different booster versions, providing deeper insights into performance trends.

- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Building the Model

- Load the dataset using NumPy and Pandas
- Clean and transform the data
- Split the data into training and testing sets
- Choose a suitable machine learning algorithm
- Apply GridSearchCV for hyperparameter tuning and model fitting

Evaluating the Model

- Check accuracy scores for each model
- Tune hyperparameters to improve performance
- Generate and analyze confusion matrices

Improving the Model

- Apply feature engineering techniques
- Perform additional algorithm tuning

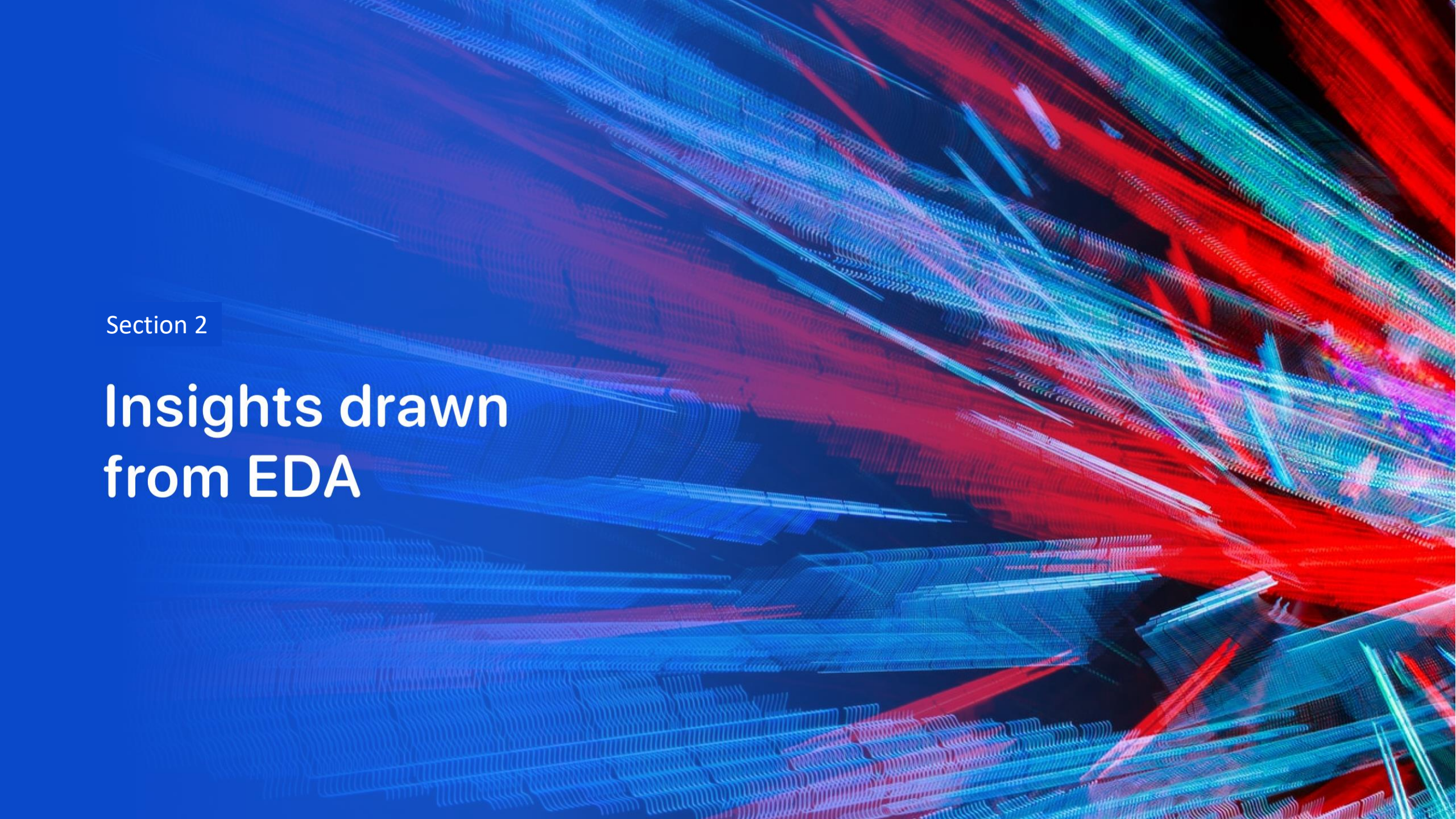
Finding the Best Model

- Select the model with the highest accuracy score
- Use this model as the final predictive tool

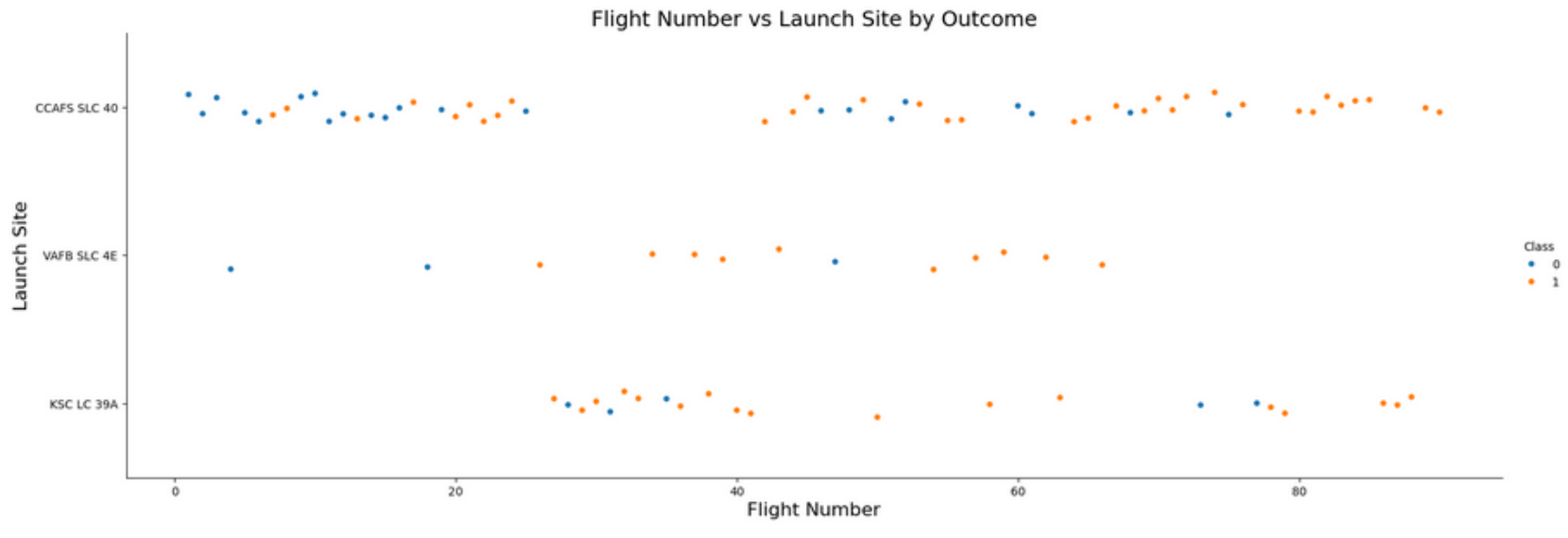
- https://github.com/markellajd/IBM-applied-data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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


Section 2
Insights drawn from EDA

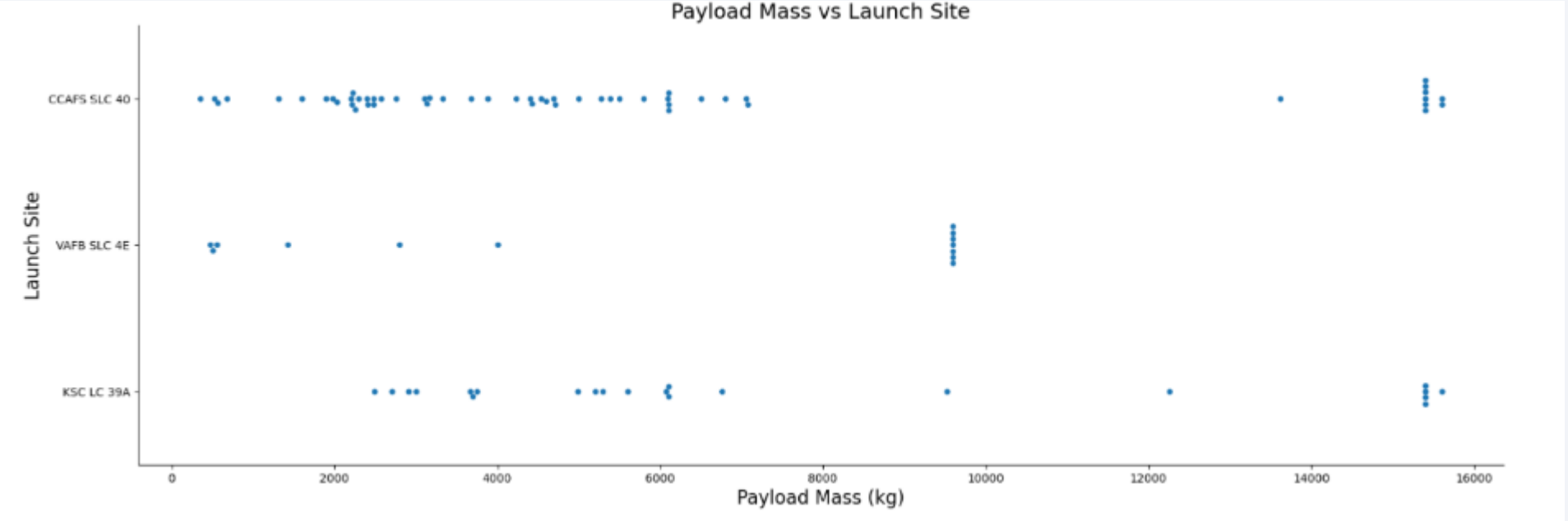
Flight Number vs. Launch Site



- CCAFS SLC 40 and KSC LC 39A show both successful and failed landings, suggesting that landing success depends on more than just location.

- Launches occur consistently across flight numbers, indicating steady activity without a clear trend in landing outcomes.

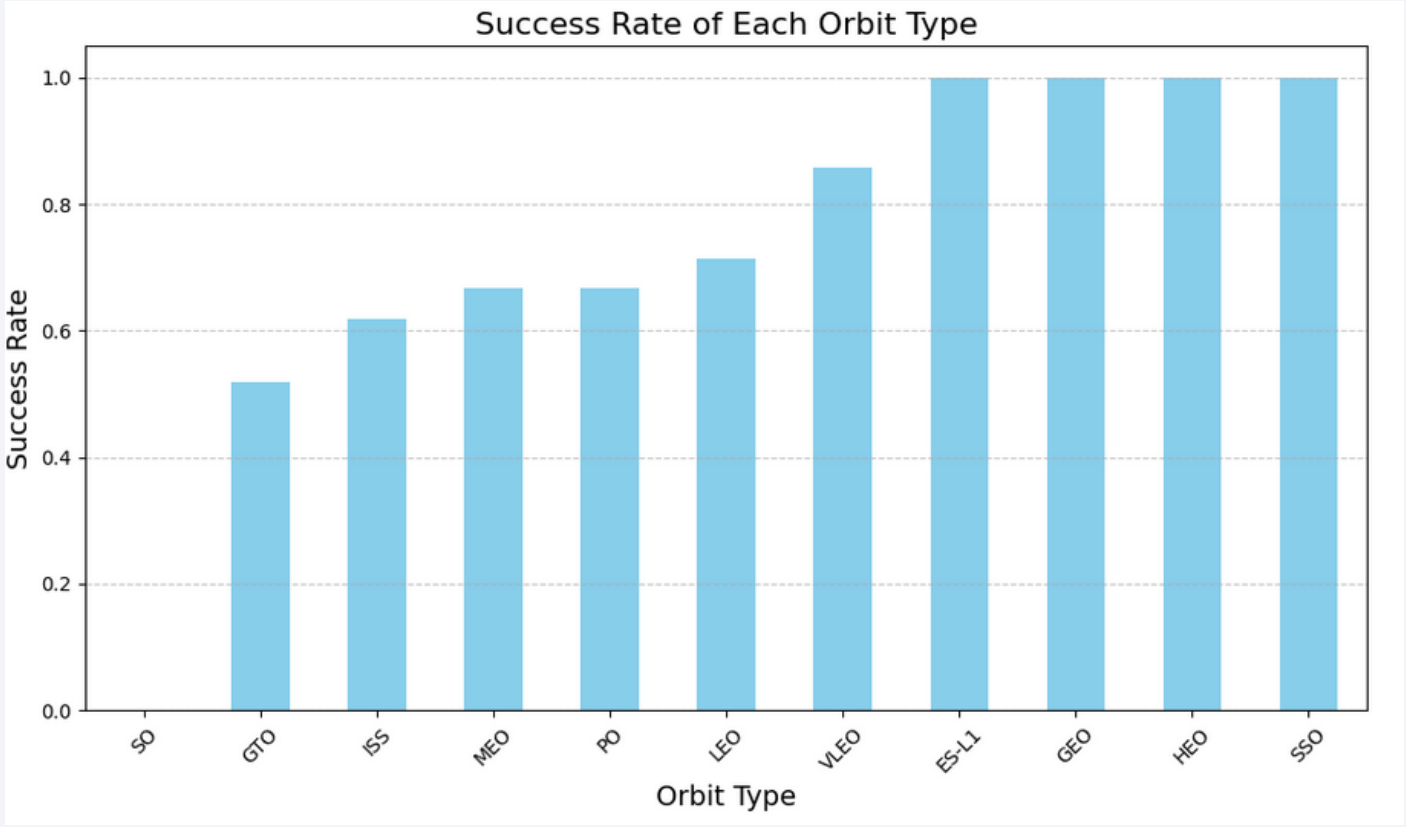
Payload vs. Launch Site



- CCAFS SLC 40 typically launches payloads under 10,000 kg, while VAFB SLC 4E and KSC LC 39A handle a broader range of masses, reflecting diverse mission types.

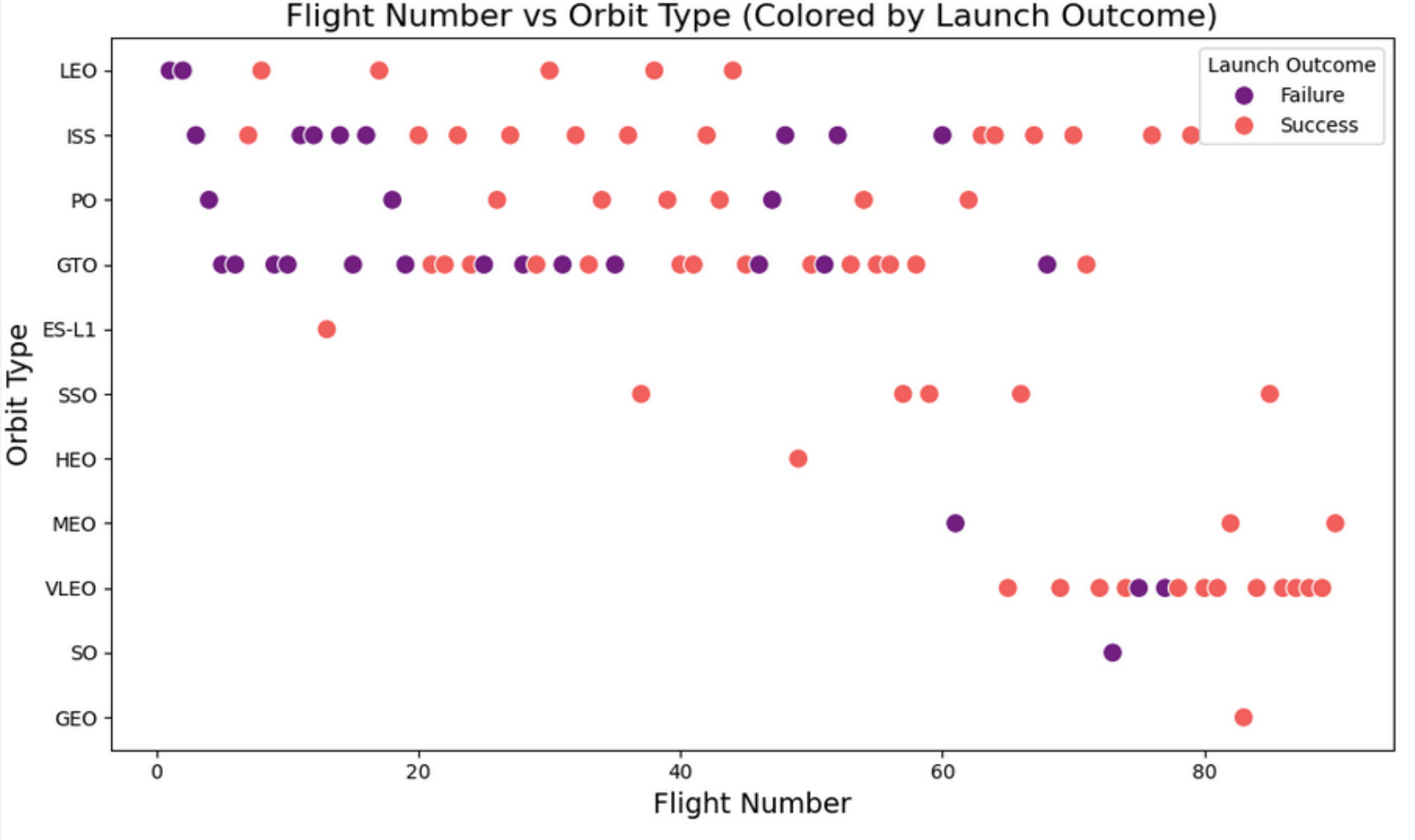
- KSC LC 39A often supports high-mass launches, with several exceeding 15,000 kg, highlighting its role in heavy-capacity missions.

Success Rate vs. Orbit Type



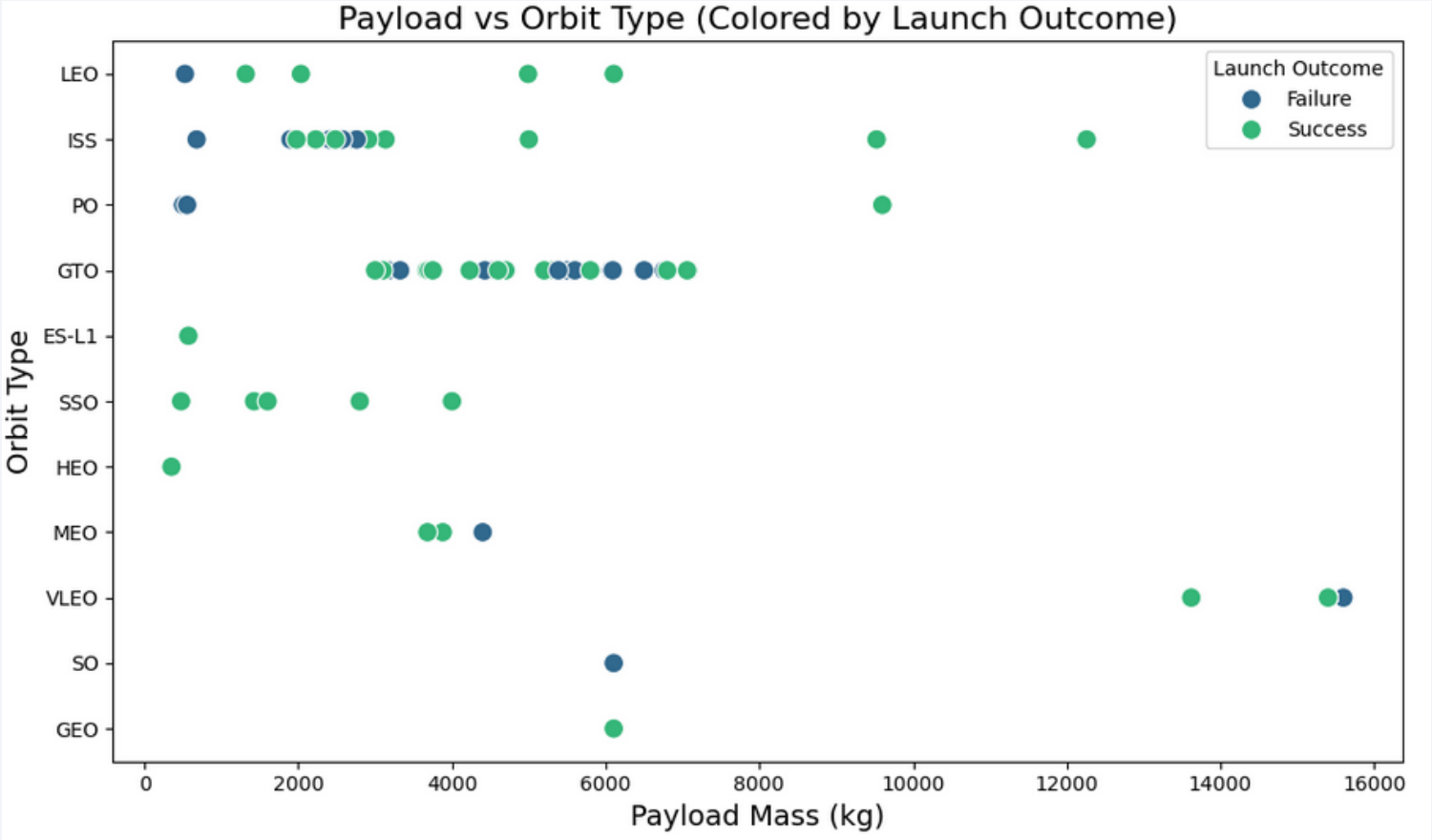
- ES-L1, GEO, HEO and SSO orbits have shown a 100% success rate, indicating strong reliability for first-stage landings.
- The GTO orbit has a notably lower success rate, suggesting it poses greater technical challenges compared to other orbits.

Flight Number vs. Orbit Type



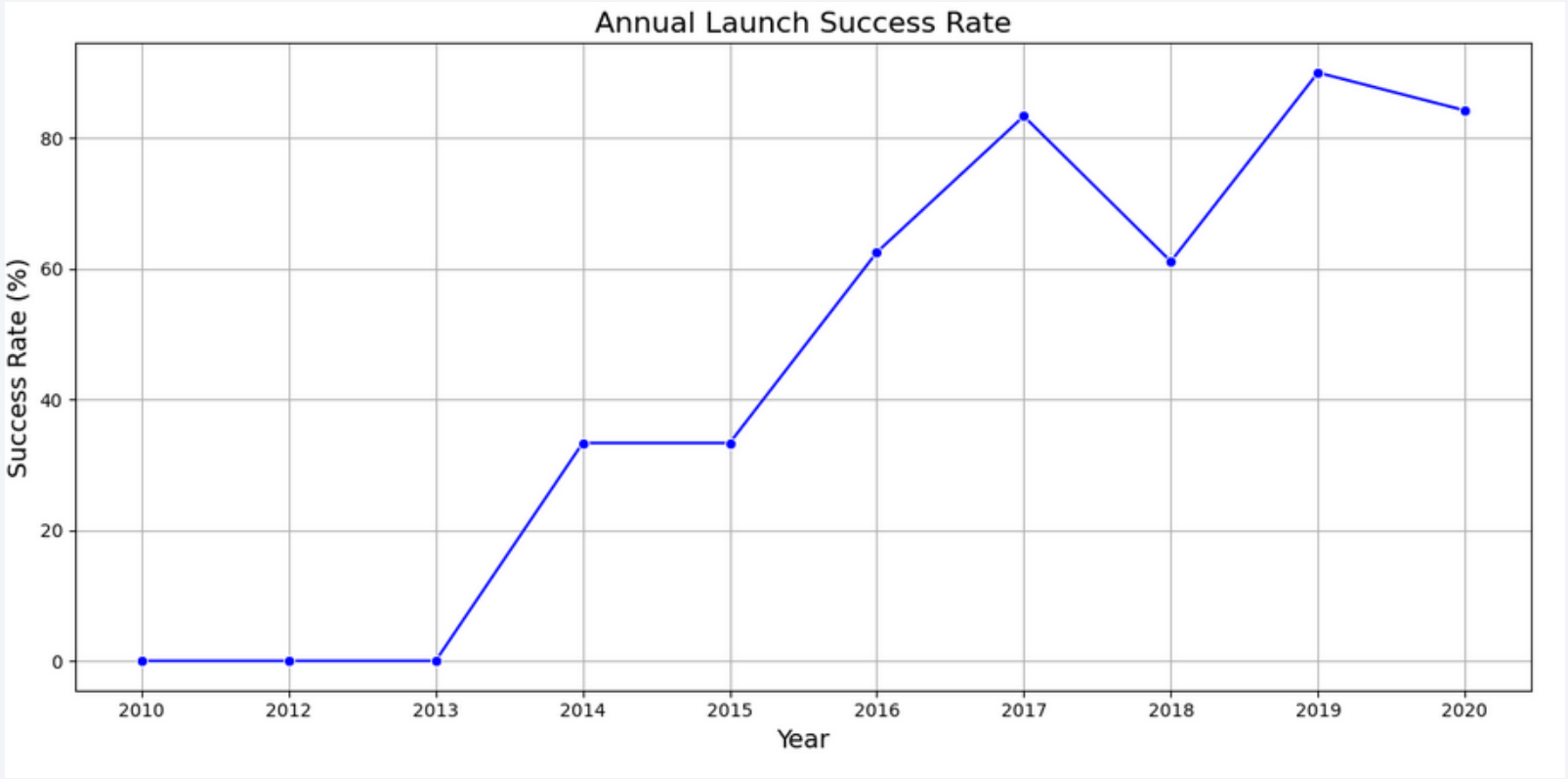
- Falcon 9's success rate increases with flight number, showing that experience and refinements lead to better results.
- While early missions to GTO and ISS had mixed results, recent ones show higher success rates, reflecting progress in mission strategy.

Payload vs. Orbit Type



- Landings are more successful across orbits when payloads are under 6000 kg.
- Payloads over 10,000 kg show mixed results, indicating greater difficulty in achieving successful landings.

Launch Success Yearly Trend



- Since 2013, Falcon 9's success rate has improved significantly, surpassing 80% by 2020.

All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

- We used `DISTINCT` to retrieve unique launch site names from the data..

Launch Site Names Begin with 'CCA'

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

- 5 records where launch sites begin with `CCA`

Total Payload Mass

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

SUM("PAYLOAD_MASS__KG_")
45596

- The total payload was found 45596 using the query above

Average Payload Mass by F9 v1.1

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

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

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

AVG("PAYLOAD_MASS__KG_")
2928.4

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

6]:

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

6]:

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS_
```

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

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM SPACEXTABLE WHERE "Mission_Outcome" IN ('Success', 'Failure') GROUP
```

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

Mission_Outcome	Total
Success	98

Boosters Carried Maximum Payload

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

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

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	CCAFS LC-40

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

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.

```
%%sql

SELECT
    "Landing_Outcome",
    COUNT(*) AS "Count"
FROM
    SPACEXTABLE
WHERE
    "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    COUNT(*) DESC;
```

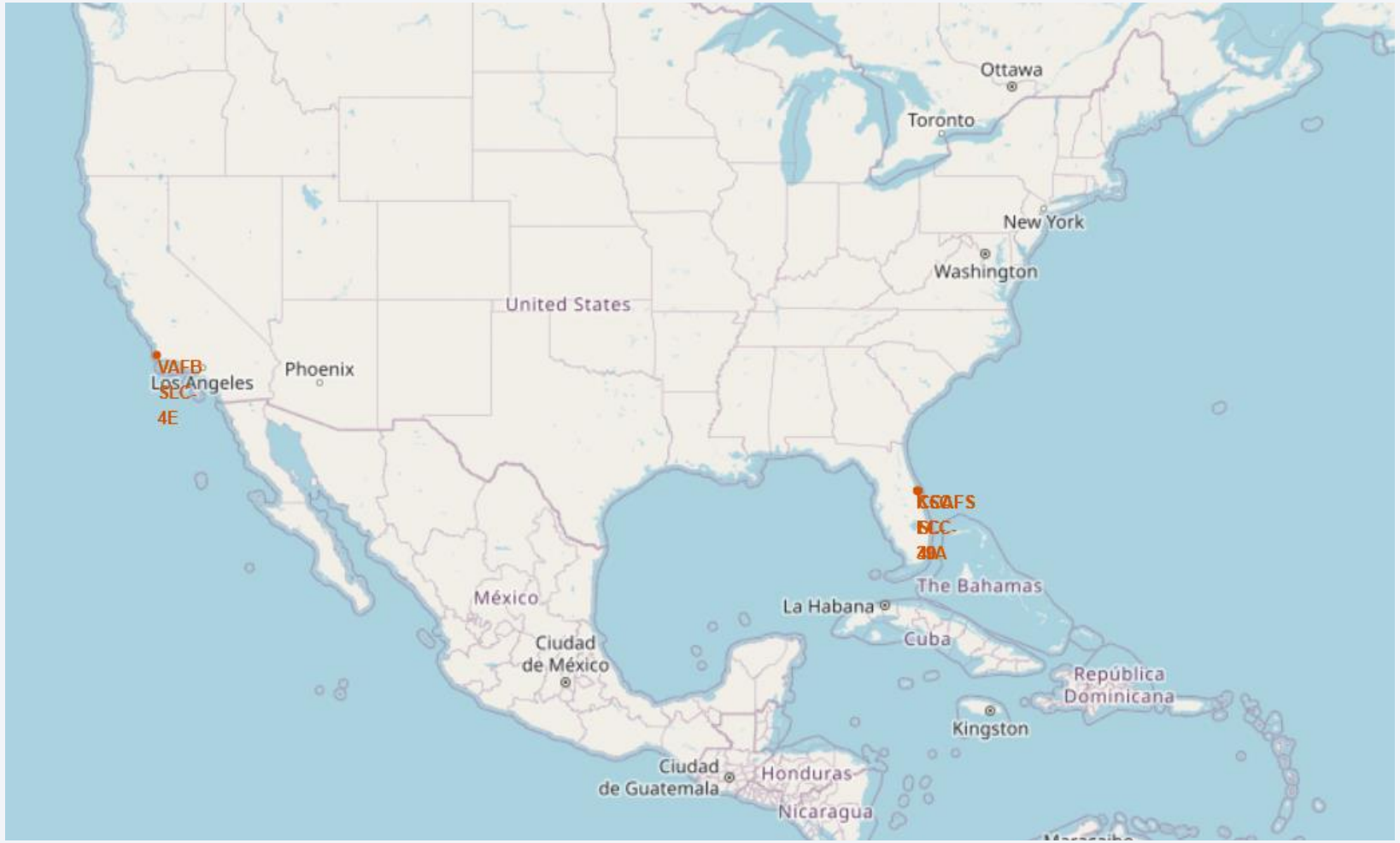
```
 * sqlite:///my_data1.db
Done.
```

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

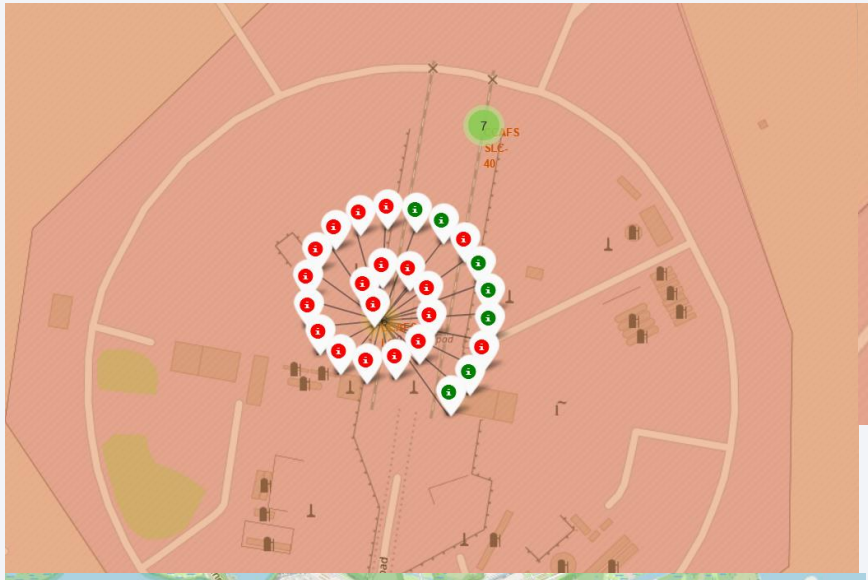
Launch Sites Proximities Analysis

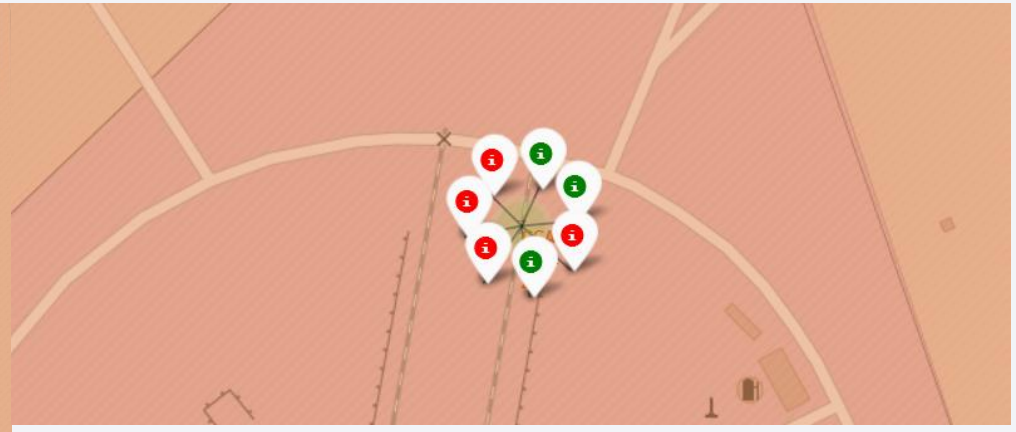
All launch sites on map

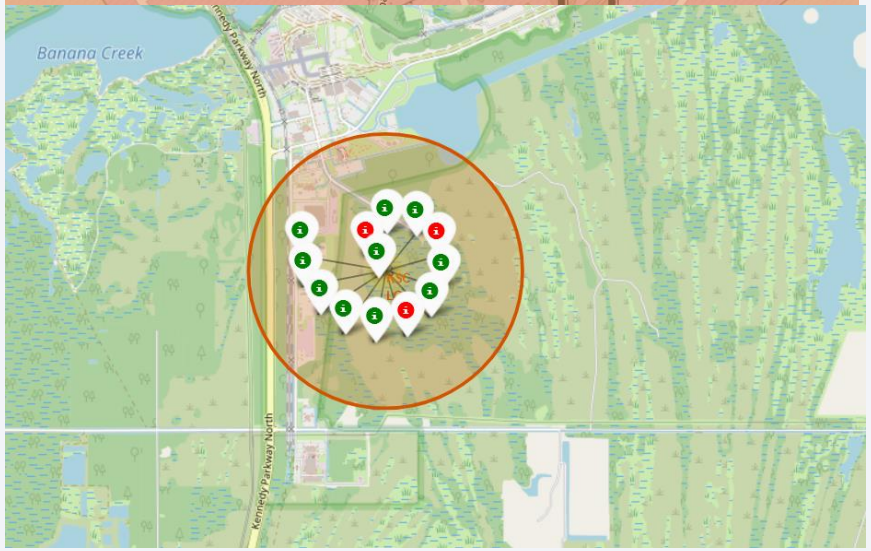


Not all launch sites are close to the Equator, but all are close to the coast.

Launch sites with color labels



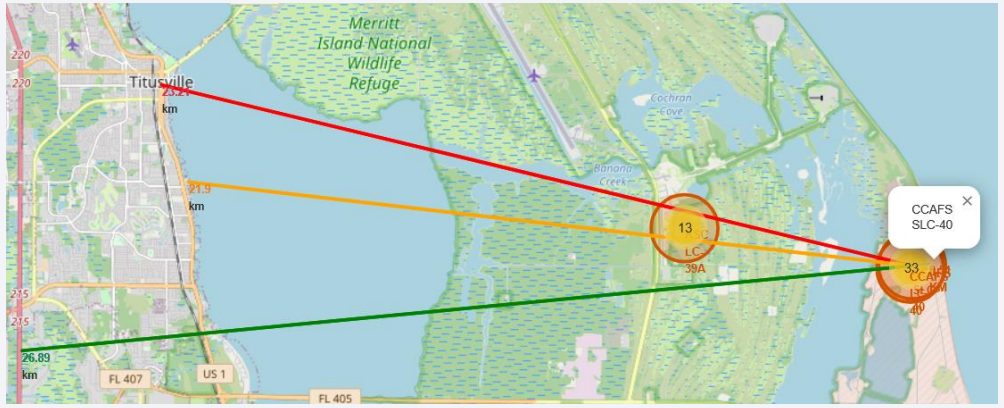


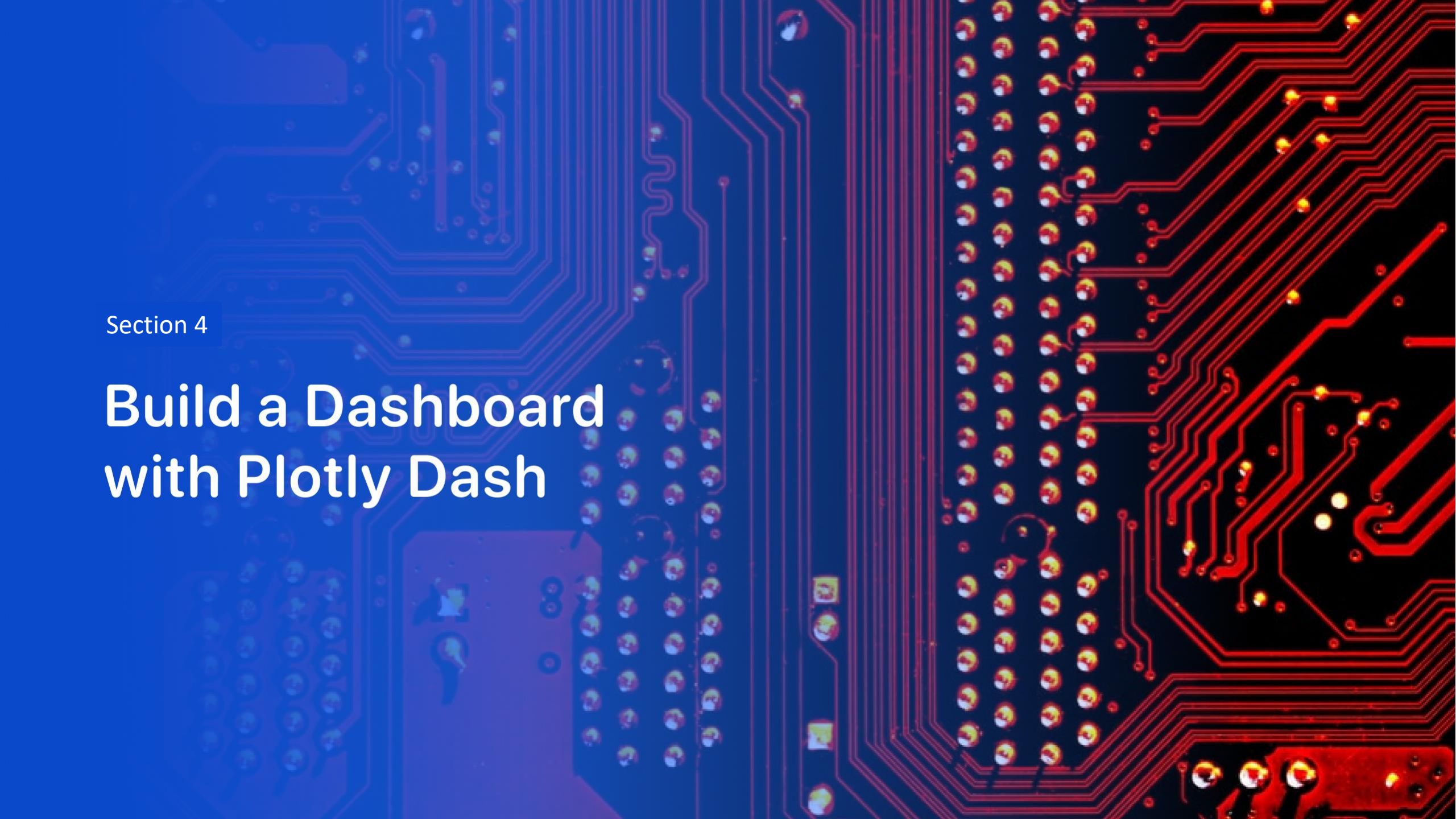


- The green markers are the successful launches and the red are the failed ones.

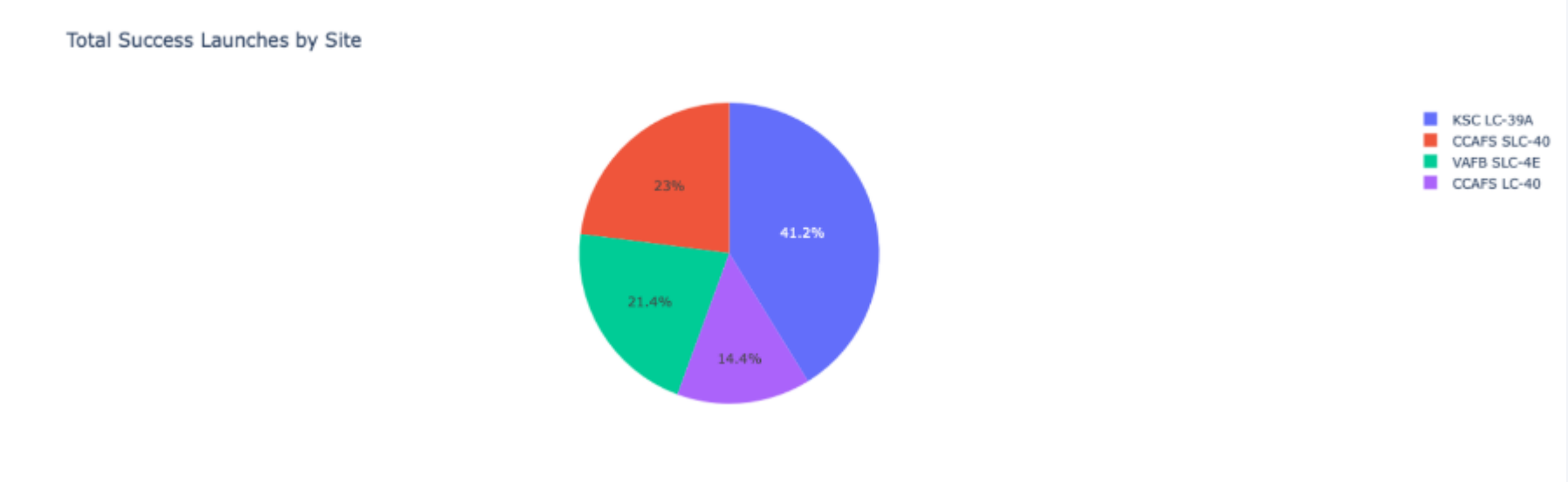
Launch sites Distance to Landmarks

- This plot shows the distance between the CCAFS SLC-40 launch site and the nearest coastline, approximately 0.51 km. The PolyLine highlights the straight-line path, illustrating its strategic placement for safe over-water launches.



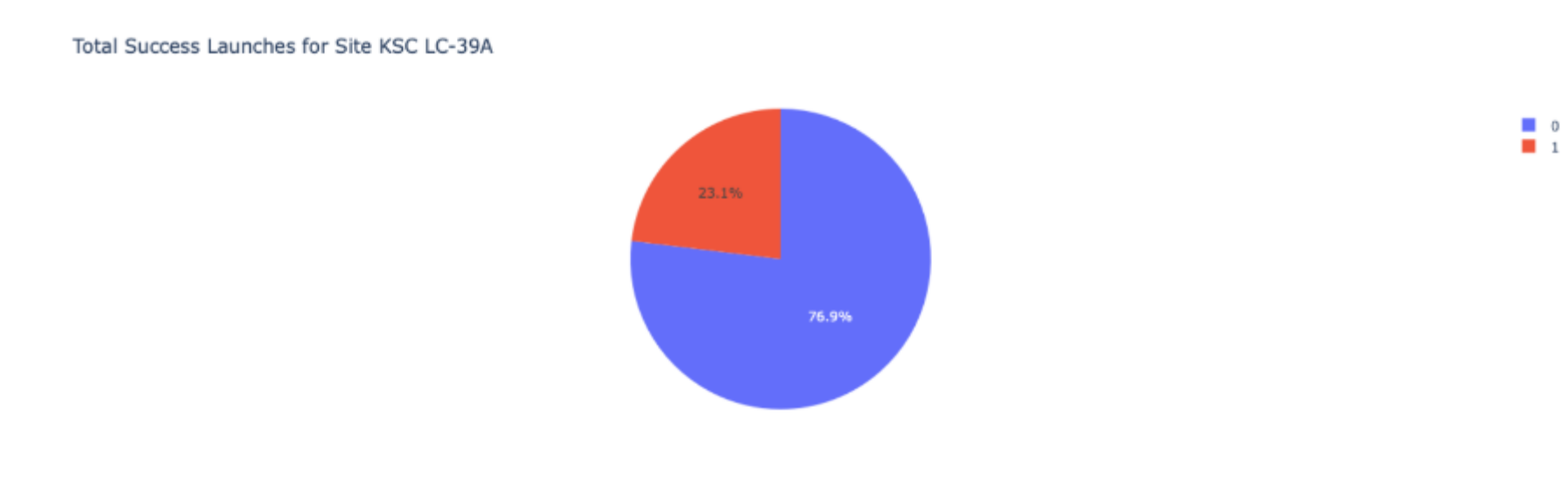

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- The chart shows that KSC LC-39A had the most successful launches

Launch site with highest launch success ratio



- The above Site shown had 76.9% success and 23.1% failed.

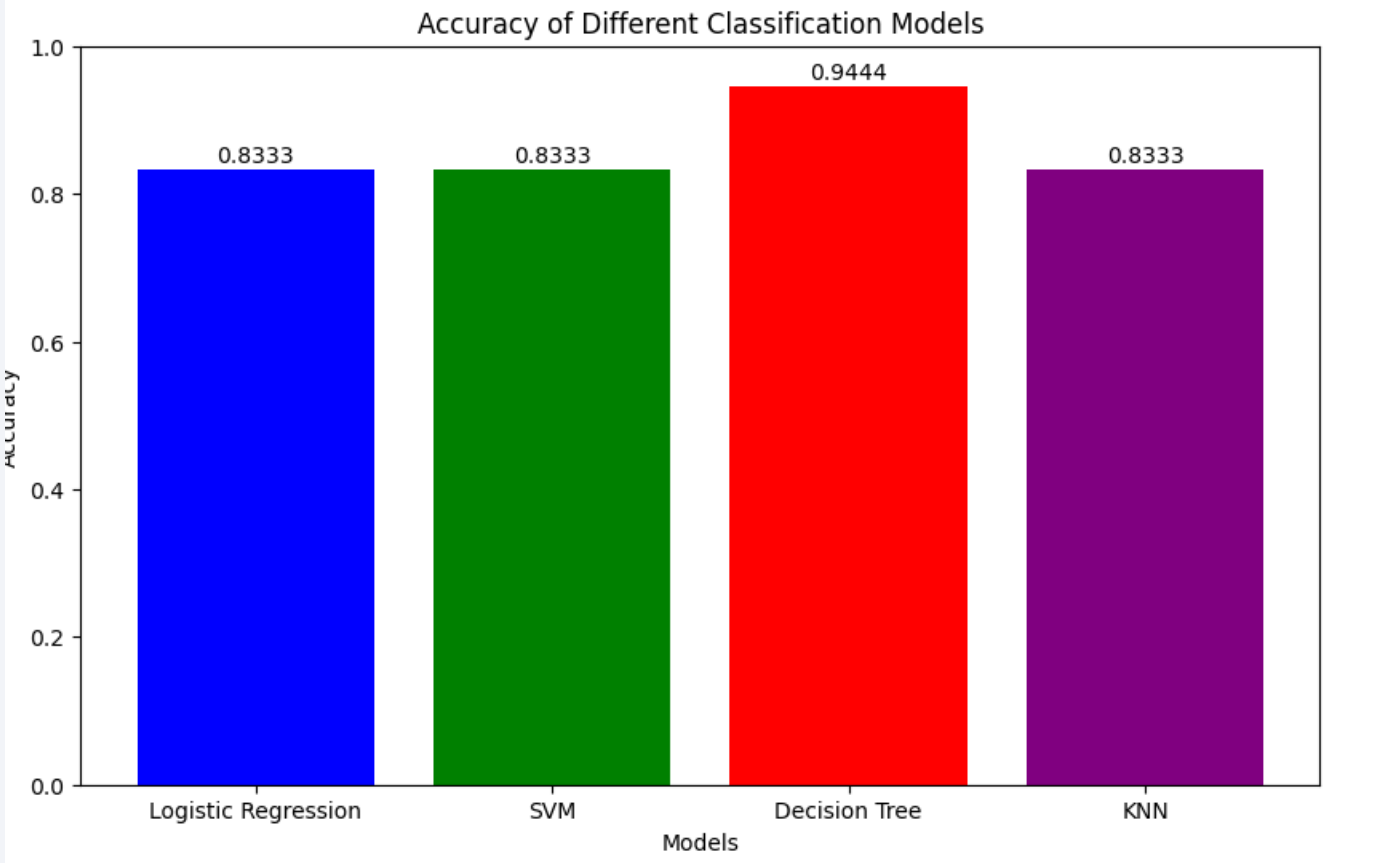
Payload Mass vs. Launch Outcome for all sites

- The charts indicate that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

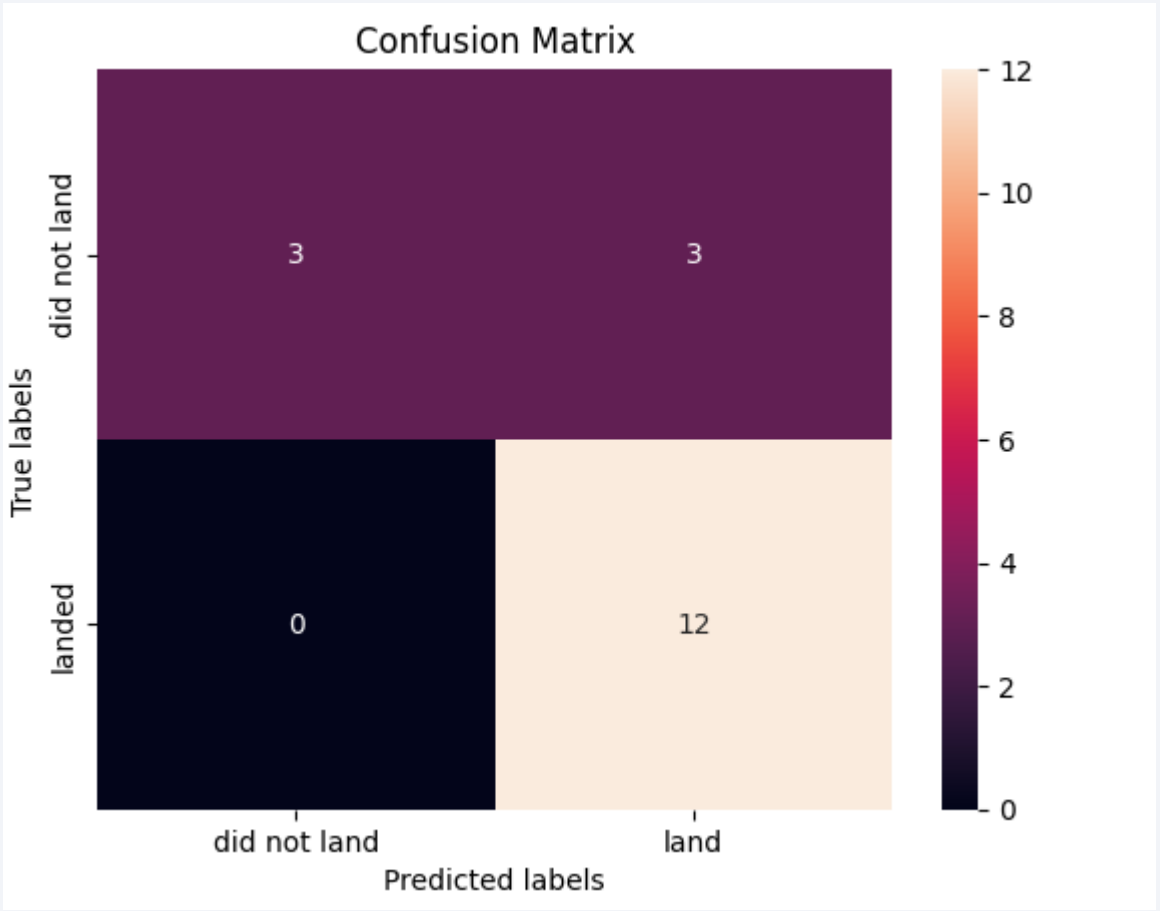
Predictive Analysis (Classification)

Classification Accuracy



- The Decision Tree model achieved the highest accuracy (0.9444), outperforming Logistic Regression, SVM, and KNN, which all scored 0.8333.

Confusion Matrix



- The confusion matrix for the Decision Tree classifier demonstrates its ability to distinguish between classes. However, it exhibits a high number of false positives, where unsuccessful landings are incorrectly predicted as successful.

Conclusions

- In conclusion, the Tree Classifier algorithm proved to be the most effective machine learning model for this dataset.
- Launches with payloads of 4000 kg or less demonstrated higher success rates compared to heavier payloads.
- Since 2013, SpaceX's launch success rate has consistently improved, indicating steady progress toward near-perfect reliability.
- Among the launch sites, KSC LC-39A achieved the highest success rate at 76.9%.
- SSO orbit demonstrated a perfect 100% success rate with multiple occurrences.
- Interactive data visualizations provided insights in a way that the user can understand the data and patterns.

Thank you!