IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Myriam Hamed
03/18/2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

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

❑Steps involved in collecting and preparing the data:

- **Data collection:**

- Data was collected using a GET request to the SpaceX API.
- The response content was decoded as JSON using the .json() function call, transforming it into a pandas DataFrame using .json_normalize()

- **Data cleaning and preprocessing:**

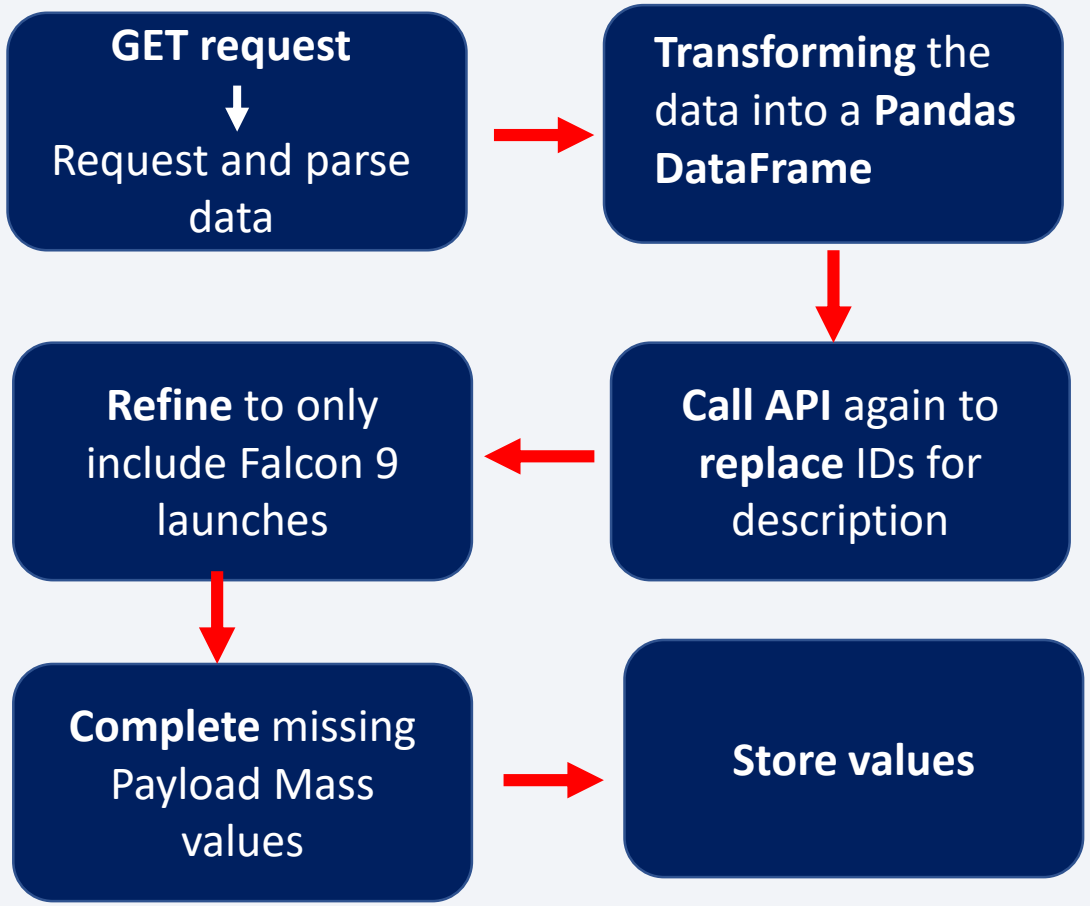
- The collected data was cleaned, checking for missing values, and filling them where necessary

- **Web scraping from Wikipedia:**

- Web scraping was performed to extract Falcon 9 launch records from a Wikipedia page using BeautifulSoup.
- The HTML table containing the launch records was parsed, and the data was converted into a pandas DataFrame.

Data Collection – SpaceX API

- Data was collected, converted, cleaned and filtered



1. Get request for rocket lauch using API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```

2. Use json_normalize methid to convert json to dataframe

```
# Use json_normalize meethod to convert the json result into d
respjson = response.json()
data = pd.json_normalize(respjson)
```

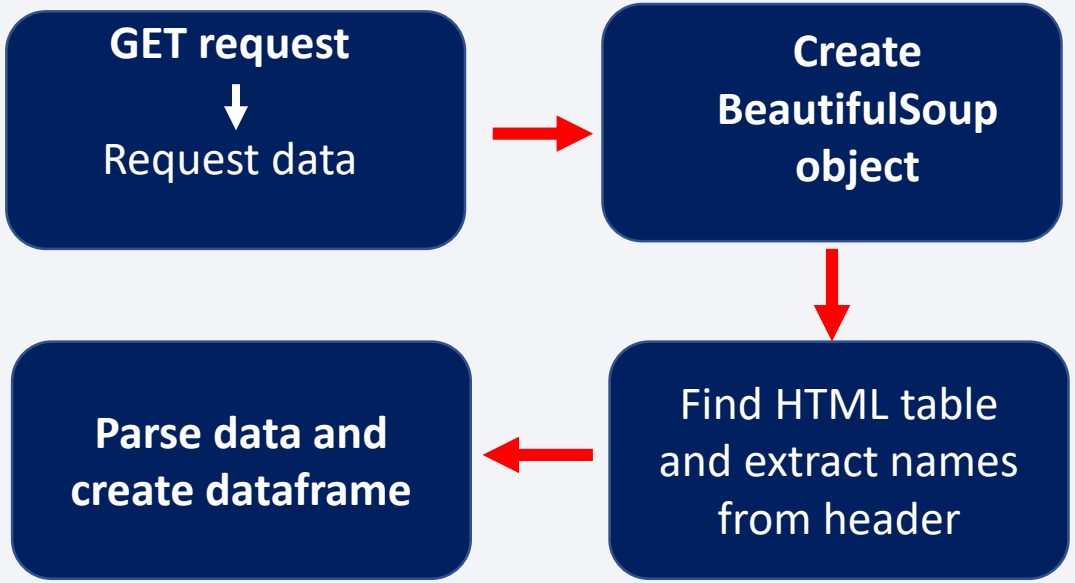
3. Data cleaning and filing in the missing values

```
# Calculate the mean value of PayloadMass column
plm_mean = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, plm_mean, inplace=True)
# Check PayloadMass column for missing values
data_falcon9['PayloadMass'].isnull().sum()
```

GitHub URL (Data Collection): https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Data was obtained through web scraping from the Wikipedia page dedicated to Falcon 9 launches



Github URL (Web Scraping):
https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-webscraping.ipynb

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text co
soup = BeautifulSoup(response.content, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

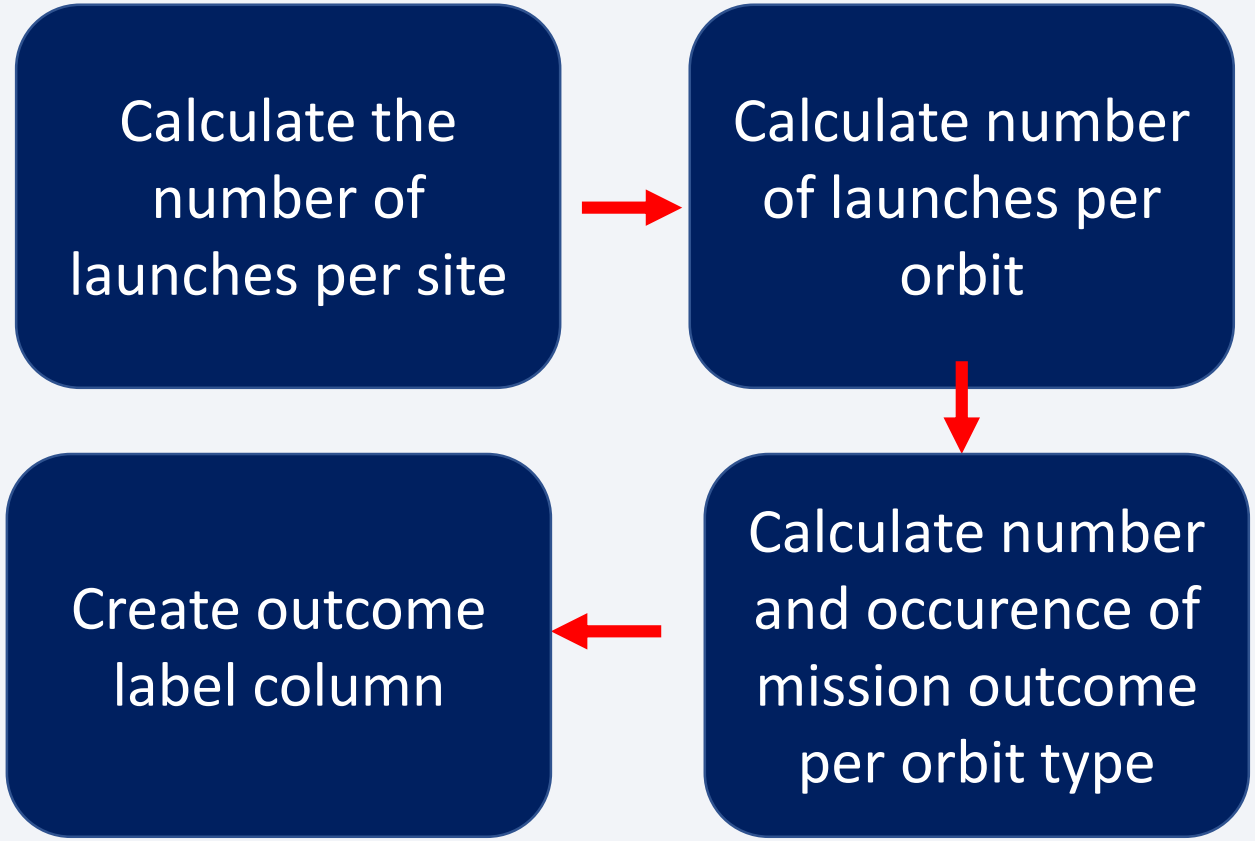
```
# Use soup.title attribute
soup.title
```

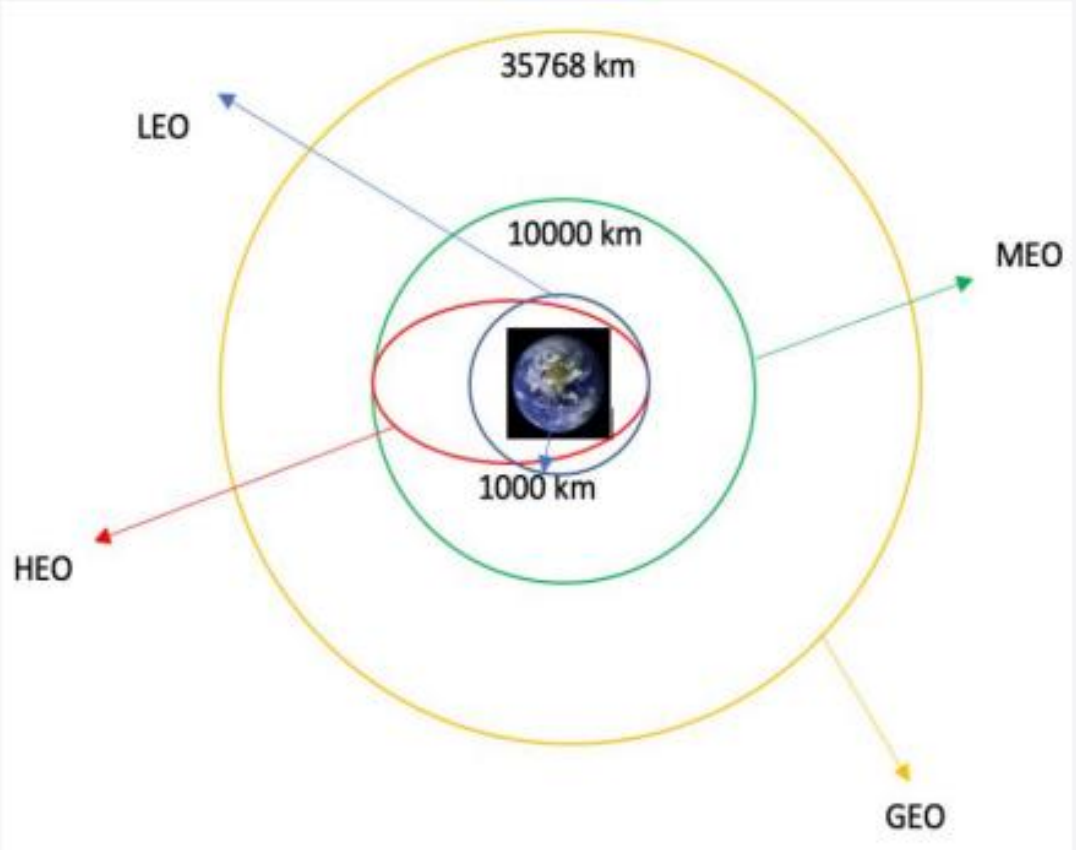
```
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header()
# Append the Non-empty column name (`if name is not None and len(name) > 0`) i

table = first_launch_table.find_all('th')
for row in table:
    name = extract_column_from_header(row)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

Data Wrangling

- Conducted initial EDA to summarize data and categorize outcomes as successful or unsuccessful landings.

GitHub URL (Data Wrangling):
https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb





EDA with Data Visualization

Summary of the visualization tasks:

- **Launch Site Trends:**

✓ Scatterplot: Shows the relationship between mission outcomes, split by Launch Site and Flight Number.

✓ Scatterplot: Illustrates the relationship between mission outcomes, split by Launch Site and Payload.

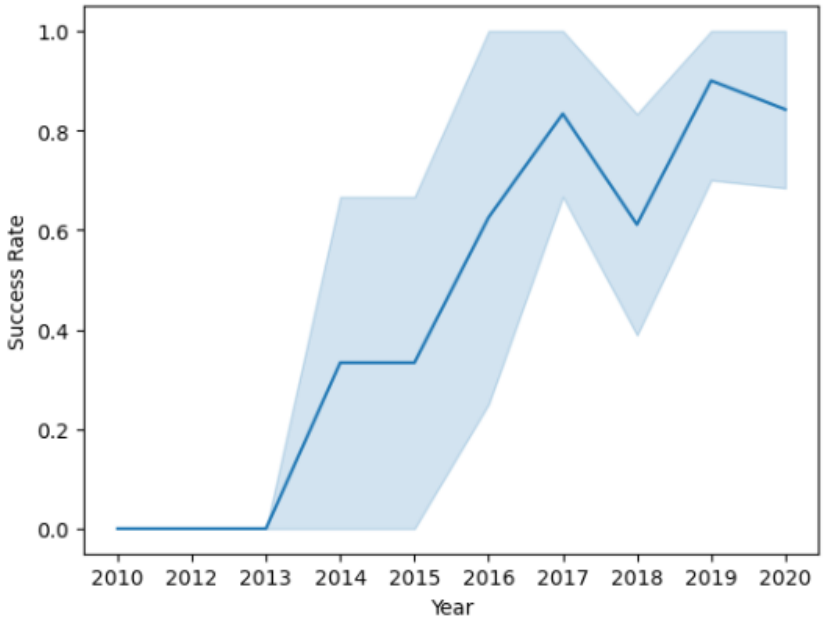
- **Orbit Type Trends:**

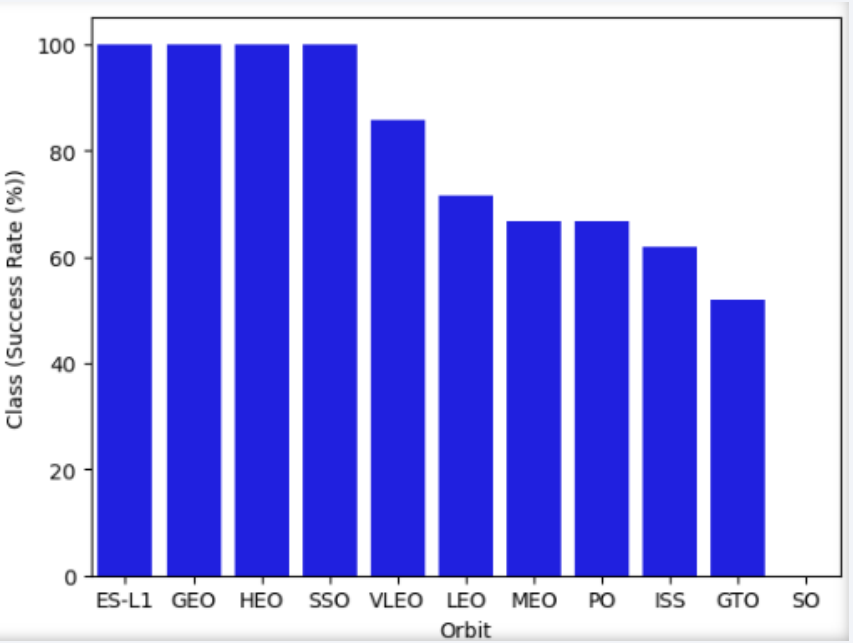
✓ Bar chart: Displays the relationship between mission outcomes and Orbit Type.

✓ Scatterplot: Visualizes the relationship between mission outcomes, split by Orbit Type and Flight Number.

✓ Scatterplot: Represents the relationship between mission outcomes, split by Orbit Type and Payload.

✓ Line plot: Depicts the trend in mission outcomes over the years.





GitHub URL (EDA with SQL): https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-eda-sql.ipynb

EDA with SQL

- **Data Exploration with SQL:**

✓ The SpaceX launch dataset was loaded into a DB2 database.

✓ Exploratory Data Analysis (EDA) techniques were applied using SQL queries to gain insights from the data.

- **Launch Sites:**

✓ Identified the unique launch sites used for SpaceX missions.

- **Payload Mass:**

✓ Calculated the total payload mass carried by boosters launched by NASA (CRS).

✓ Determined the average payload mass carried by a specific booster version

- **Mission Outcomes:**

✓ Determined the total number of successful and failed missions.

✓ Analyzed failed landing attempts on drone ships, including booster version and launch site.

GitHub URL (EDA with Data Visualization):
https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-eda-data_visualization.ipynb

Build an Interactive Map with Folium

❑Summary of Folium Map Analysis:

- **Launch Site Visualization:**

✓A folium map was created to visualize all SpaceX launch sites.

✓Markers were used to represent each launch location.

✓Circles around the markers potentially indicated launch areas.

- **Mission Outcome Representation:**

✓Launch outcomes (success or failure) were assigned numerical classes (0 for failure, 1 for success).

✓Marker clusters were employed, with colors likely used to differentiate successful and unsuccessful launches at each site. This facilitated identifying launch sites with high success rates.

- **Proximity Analysis:**

✓Distances between launch sites and nearby features were calculated. These features might include, closest city, railways, highways, coastlines

GitHub URL (Folium Maps):
https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

❑Summary of Plotly Dash Analysis:

- **Interactive Dashboard:**

✓Created using Plotly Dash for user *interaction* and *exploration*.

- **Data Visualization:**

✓**Pie Charts:** Displayed the total successful launches for each launch site.

✓**Scatter Plot:** Examined the relationship between mission outcome (success/failure) and payload mass (kg) for different booster versions.

- **Filtering and Controls:**

✓A dropdown menu allowed users to filter data in both charts by selecting a specific launch site.

✓A range slider enabled users to control the range of payload mass displayed in the scatter plot.

GitHub URL (Dashboad File):
https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/SpacexBash.py

Predictive Analysis (Classification)

❑**Summary of Predictive Modeling:**

- **Data Preparation:**

✓The data was transformed and split into training and testing sets for model development and evaluation.

- **Model Selection and Tuning:**

✓Four machine learning models were employed:

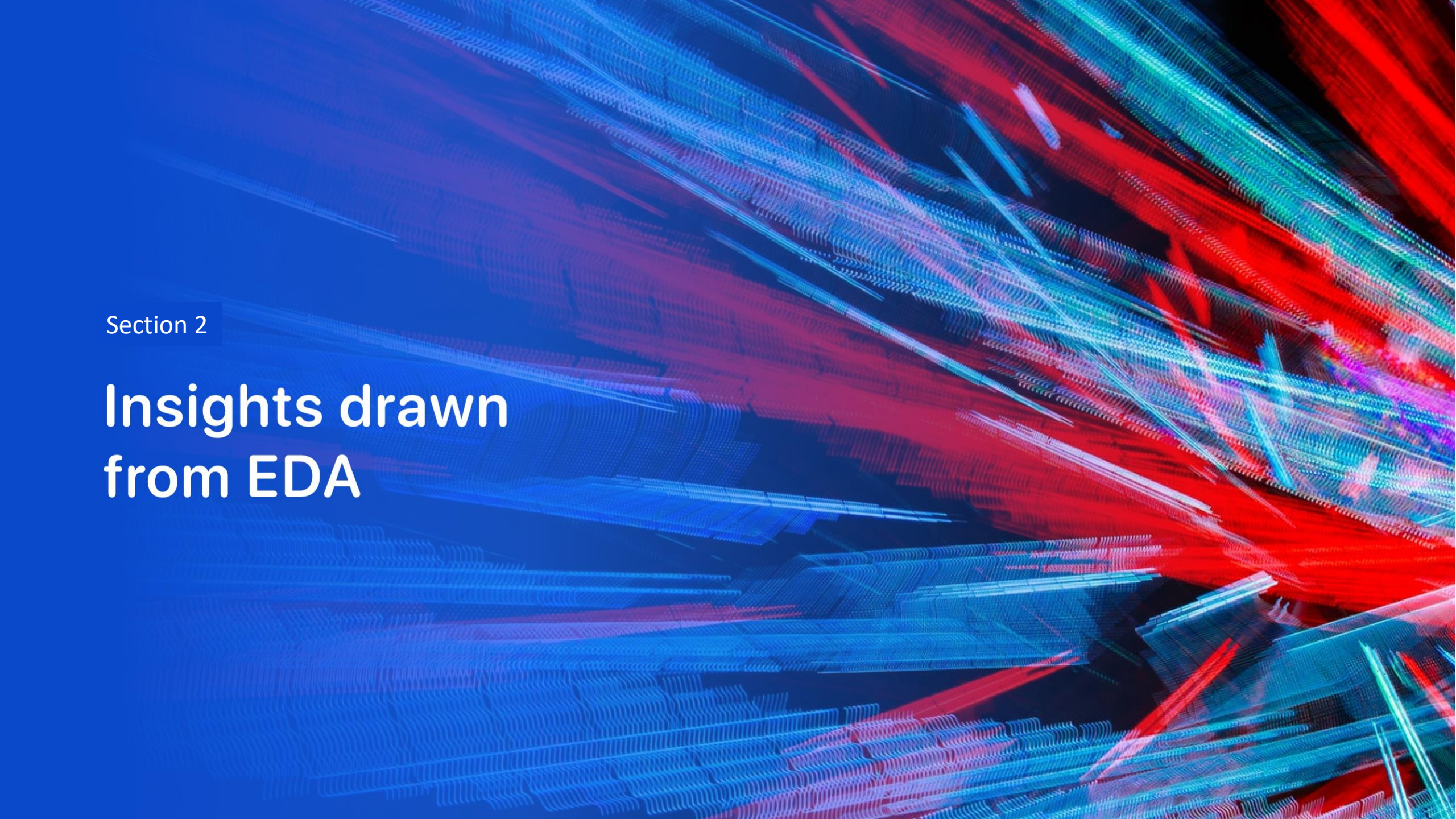
  - Support Vector Machine
  - Classification Tree
  - Logistic Regression
  - K-Nearest Neighbor

- Hyperparameter tuning was performed for each model to optimize its performance.
- Found the best classification model

GitHub URL (Machine Learning):
https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

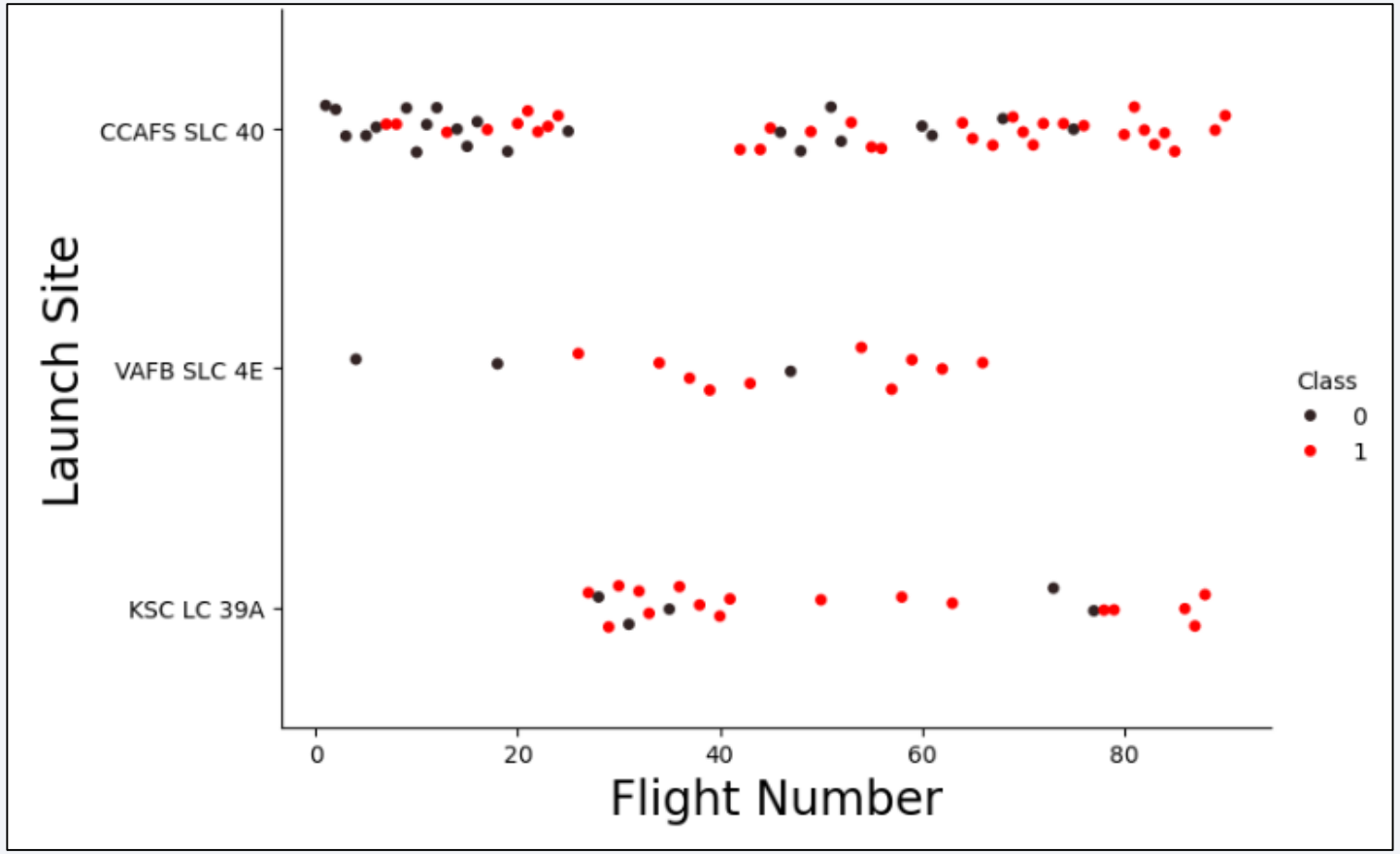

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Analysis shows:

Launch sites with more flights tend to have higher success rates.

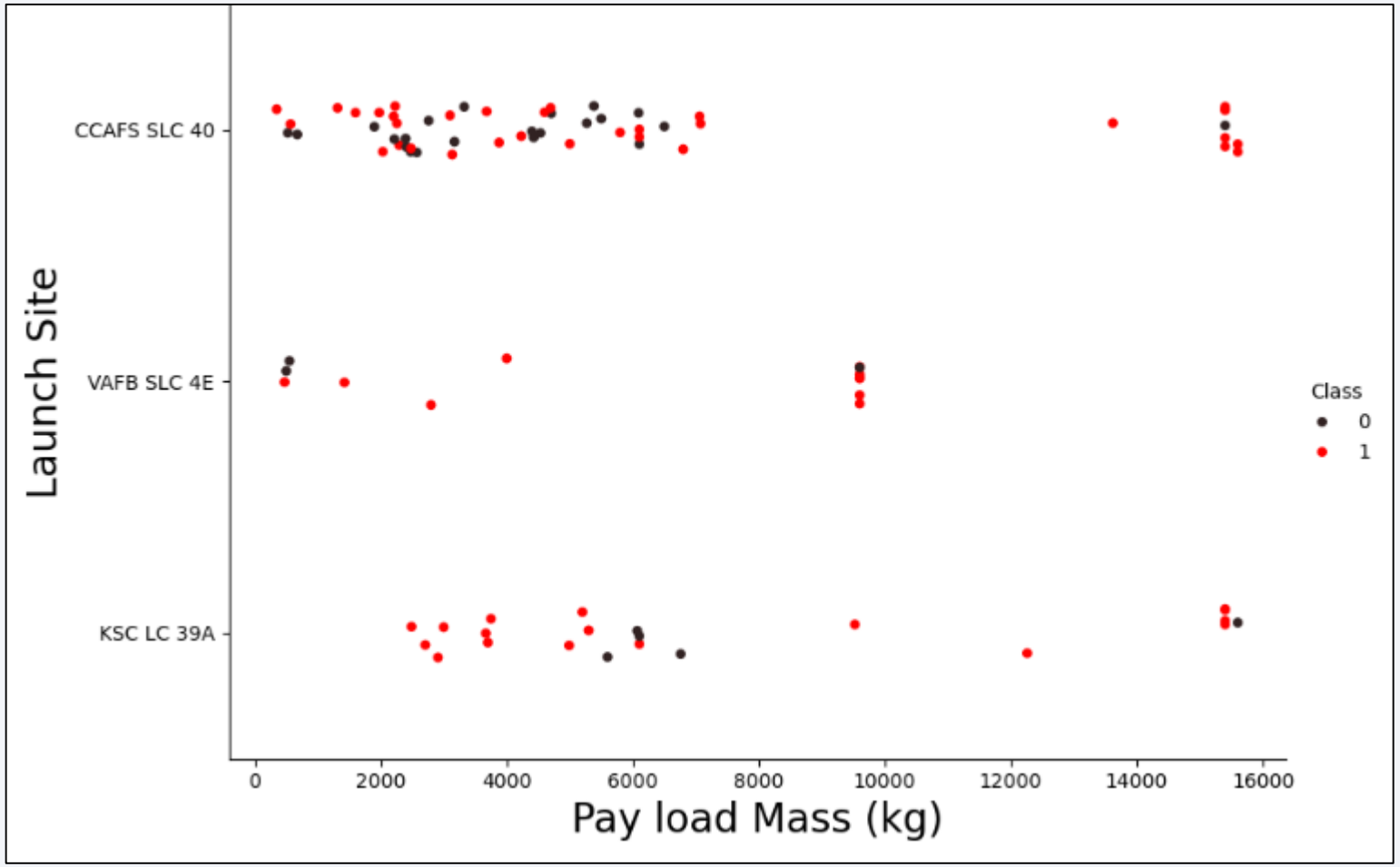
Successful Falcon 9 landings become more likely with more flights



- **Failed landings** are indicated by the 0 class (**black marker**)
- **Successfu landings** by the 1 class (**red marker**)

Payload vs. Launch Site

- **Summary of Launch Site and Payload Analysis:**
- **CCAFS SLC-40:** No strong correlation observed between payload mass and landing outcome at this launch site. Landings seem independent of payload weight.
- **KSC LC-39A:** Failed landings here appear concentrated around a specific payload mass range, suggesting a potential issue with handling launches in that weight category.

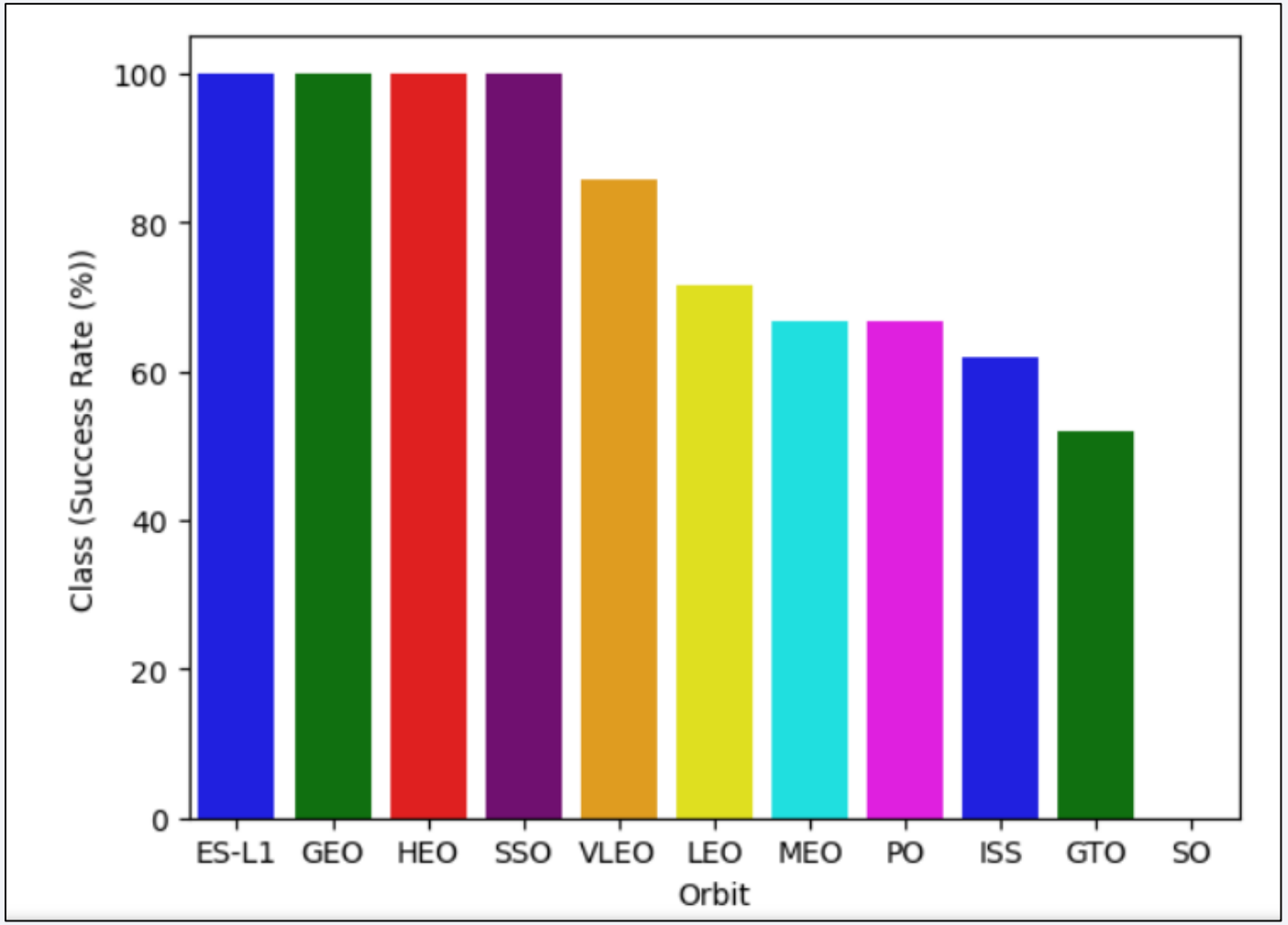


- **Failed landings** are indicated by the 0 class (**black marker**)
- **Successfu landings** by the 1 class (**red marker**)

Success Rate vs. Orbit Type

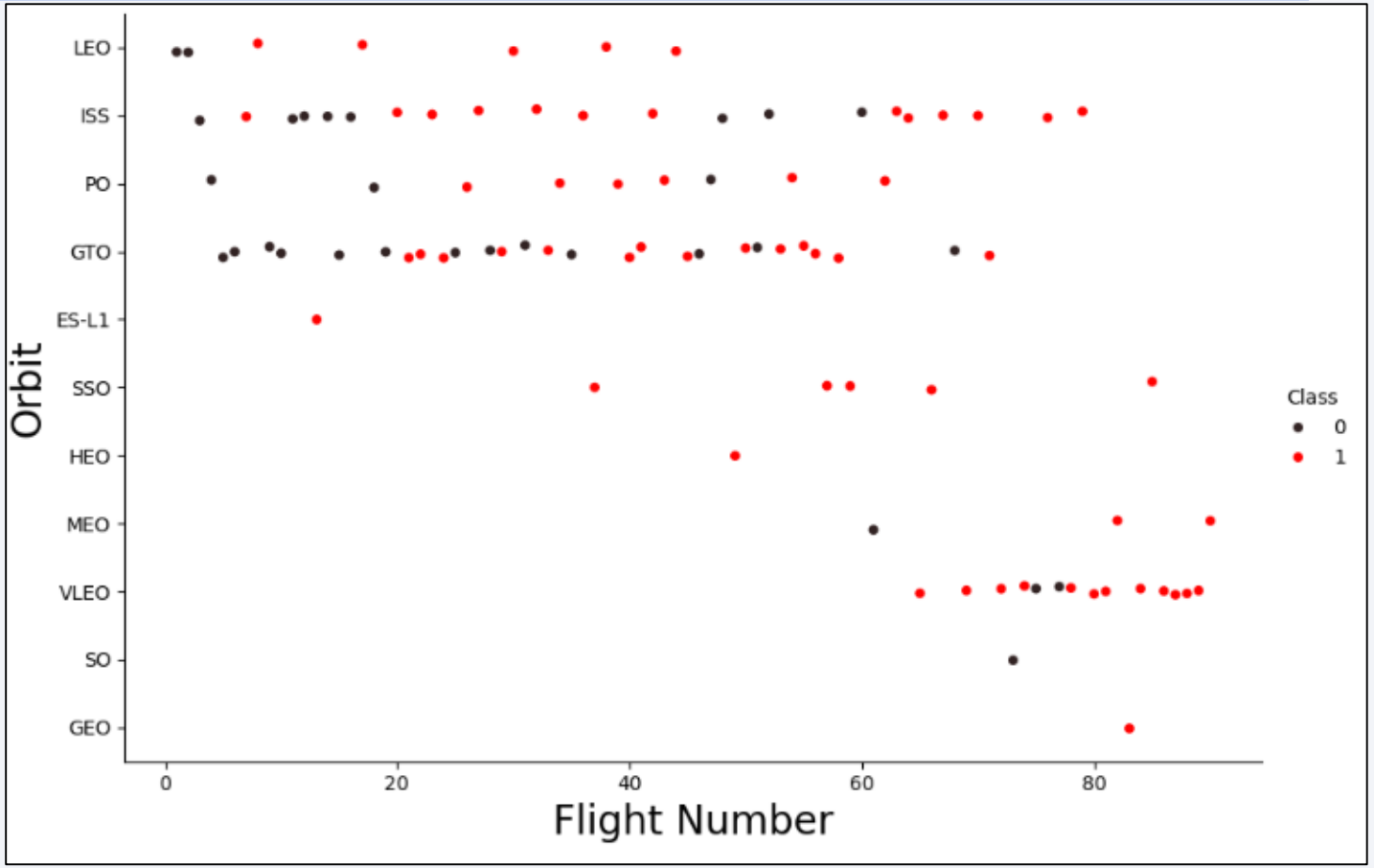
- **High Success Rates:** Orbits like ES-L1, GEO, HEO, and SSO exhibited a 100% success rate for Falcon 9 first-stage landings. SpaceX appears to have achieved high landing success rates for certain orbit types

- **Low Success Rate:** SO (Sun-Synchronous) orbit had a 0% success rate for first-stage landings. This indicates a potential challenge SpaceX needs to address for landings when targeting SO orbits.



Flight Number vs. Orbit Type

- From the plot, we can observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.
- Higher flight numbers generally correlate with higher success rates.

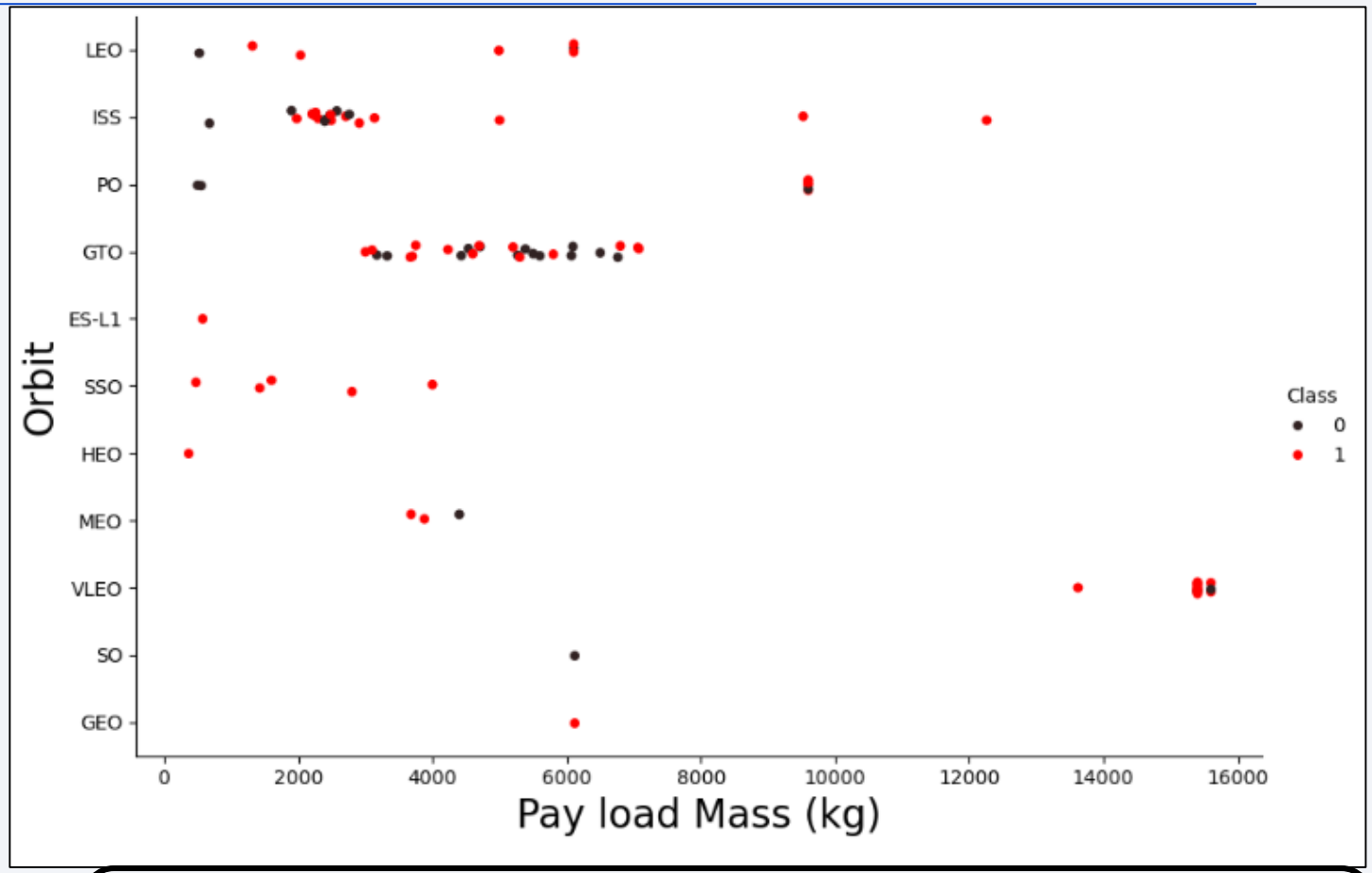


- **Failed landings** are indicated by the 0 class (**black marker**)
- **Successfu landings** by the 1 class (**red marker**)

Payload vs. Orbit Type

- **Payload Mass Not the Main Factor:**

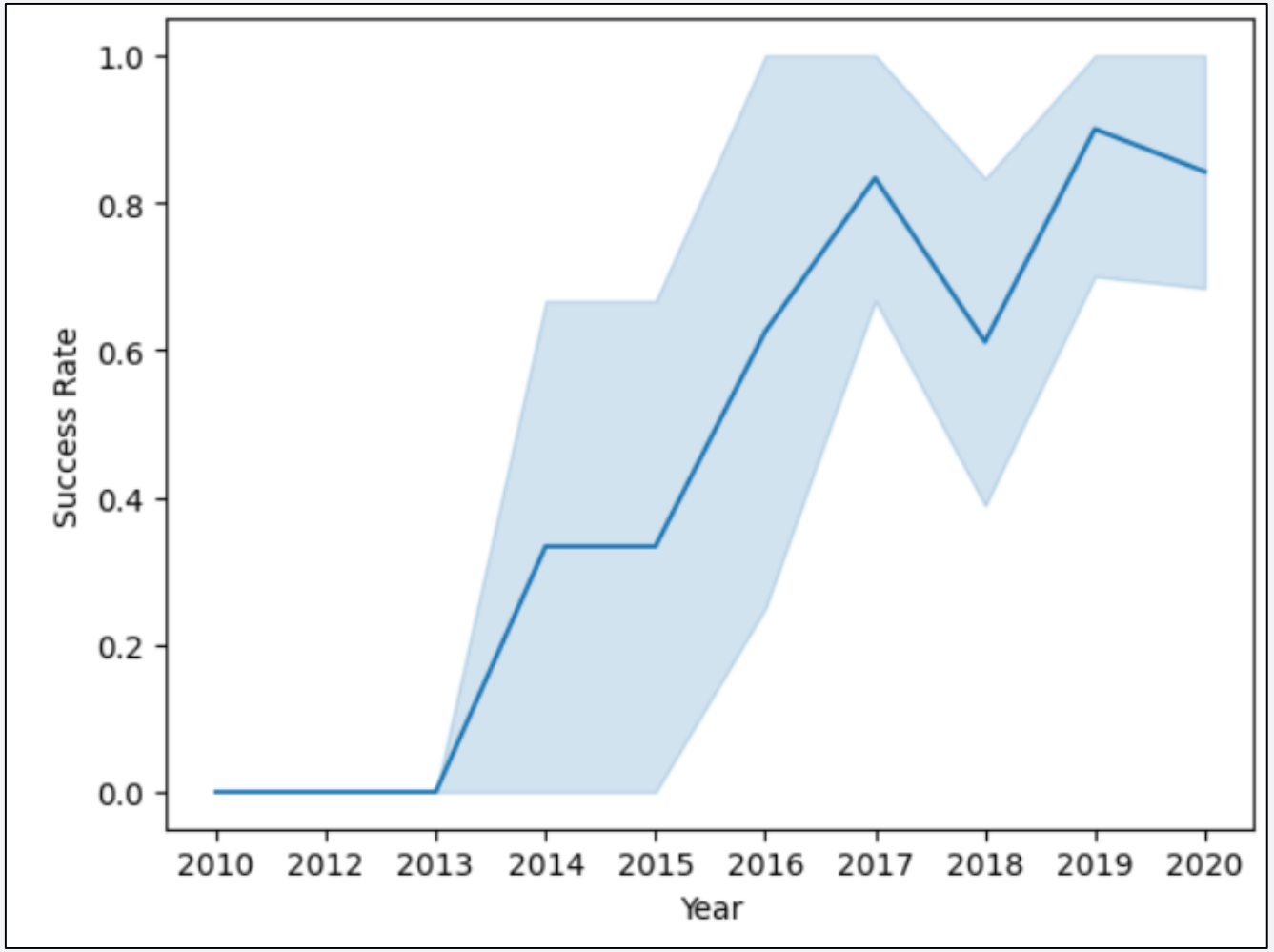
Overall landing success doesn't seem to be strongly influenced by payload weight.



- **Failed landings** are indicated by the 0 class (**black marker**)
- **Successfu landings** by the 1 class (**red marker**)

Launch Success Yearly Trend

- The success rate increase significantly over the years
- SpaceX's Falcon 9 first-stage landing success rate has shown a significant improvement over times



All Launch Site Names

- Find the names of the unique launch sites
- Queried the SpaceX data to find a list of distinct launch sites.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- **Utilized the LIKE keyword to identify launch sites starting with 'CCA', and employed the LIMIT keyword to exhibit the first 5 rows.**

```
%sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE 'CCA%' AND  LIMIT 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

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

In [12]:

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

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

Out[12]:

SUM(PAYLOAD_MASS__KG_)
45596

The total payload carried by boosters from NASA is 45,596 kg

Average Payload Mass by F9 v1.1

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

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

In [13]:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

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

Out[13]:

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

In [37]:

```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://fhl12841:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

Out[37]:

1
2015-12-22

The first successful landing outcome on ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

The booster versions that have successfully landed on drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg are listed above.

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Observed that 100 missions are successful whereas 1 had failed

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://fhl12841:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
Out[39]:
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum
- payload mass

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

In [40]:

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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

2015 Launch Records

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

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://fhl12841:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
Out[41]:
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

There were two failed landing outcomes with a drone ship in 2015. Both launched from 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

- The most common landing outcome was ‘not attempted

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://fhl12841:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
Out[42]:
```

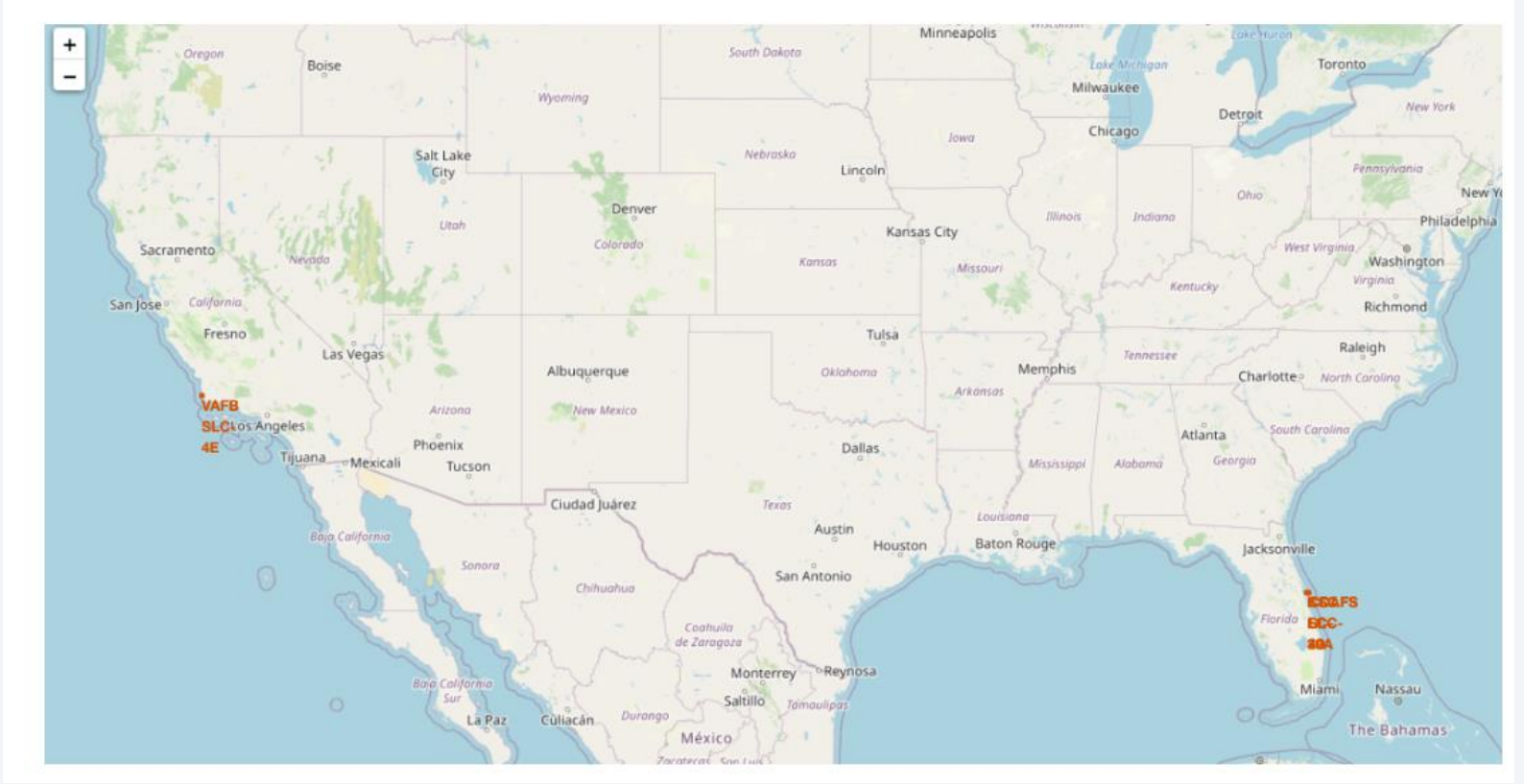
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

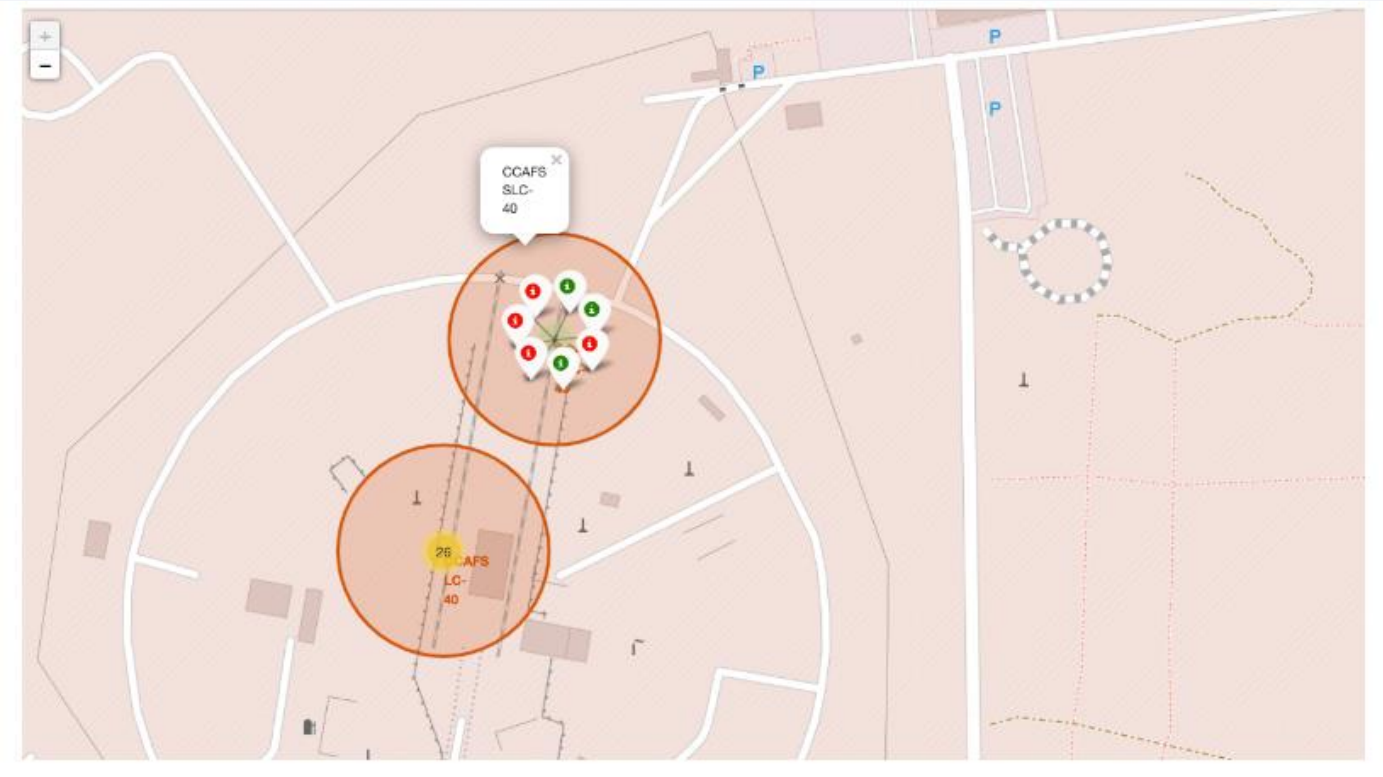
Folium map with all launch sites

- The map reveals that all launch sites are situated along the coastlines of the United States, with one in California and two in Florida.
- This proximity to the coast is likely due to the strategic advantage of launching over water, which minimizes the risk of damage from falling debris.



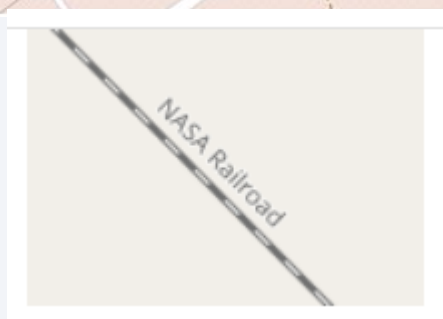
Map with outcome labeled markers

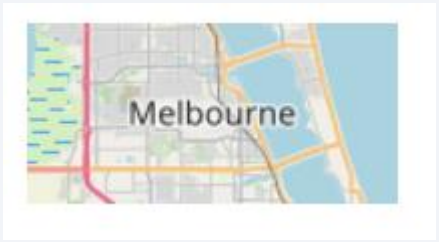
- From the color-labeled markers in the marker cluster, you can easily identify which launch sites have a relatively high success rate.
- To each launch it was placed a marker, with color to distinguish failure (red) from success (green)

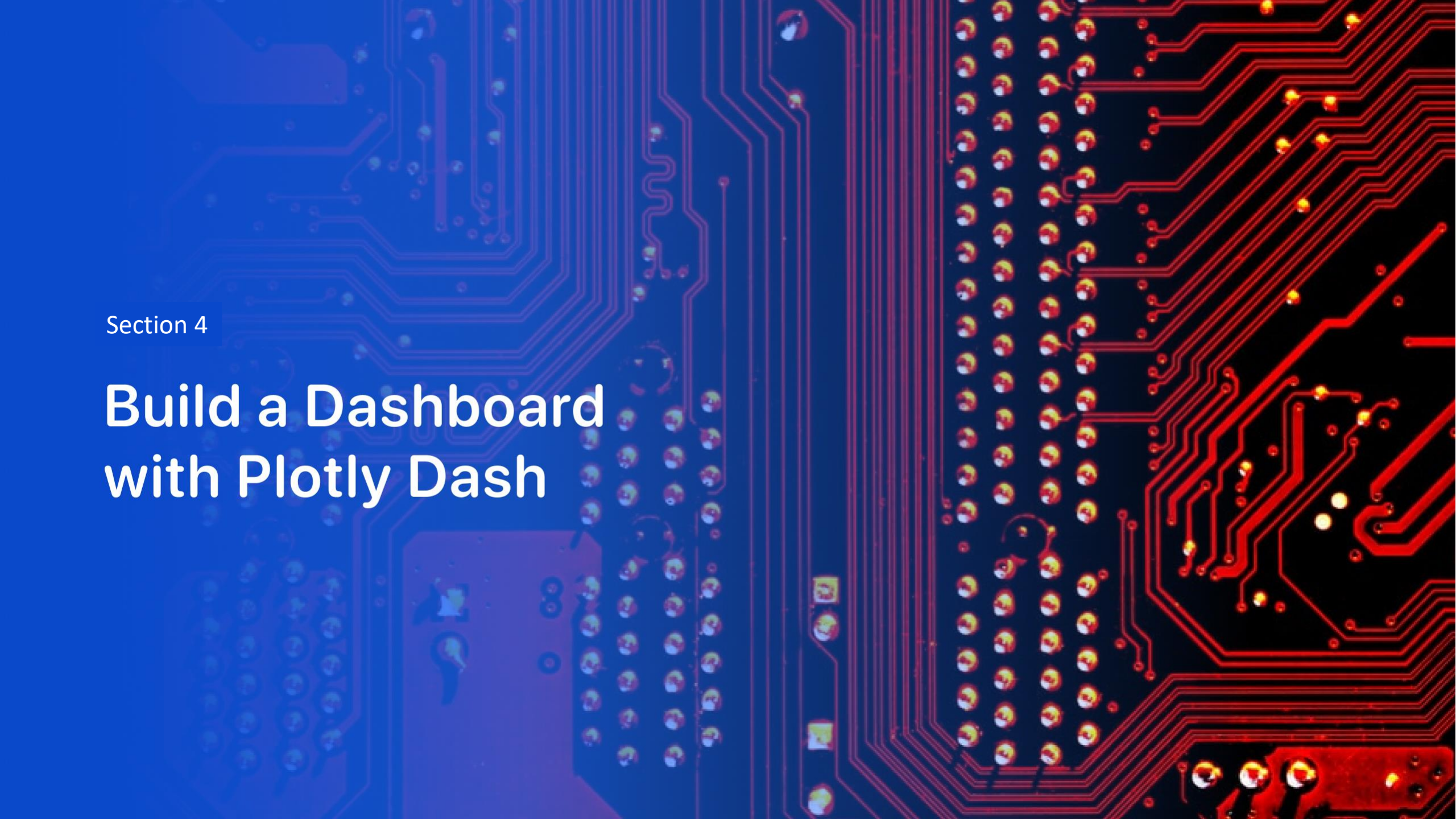


Maps with distance to proximities

- After you plot distance lines to the proximities, you can answer the following questions easily:
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways?No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

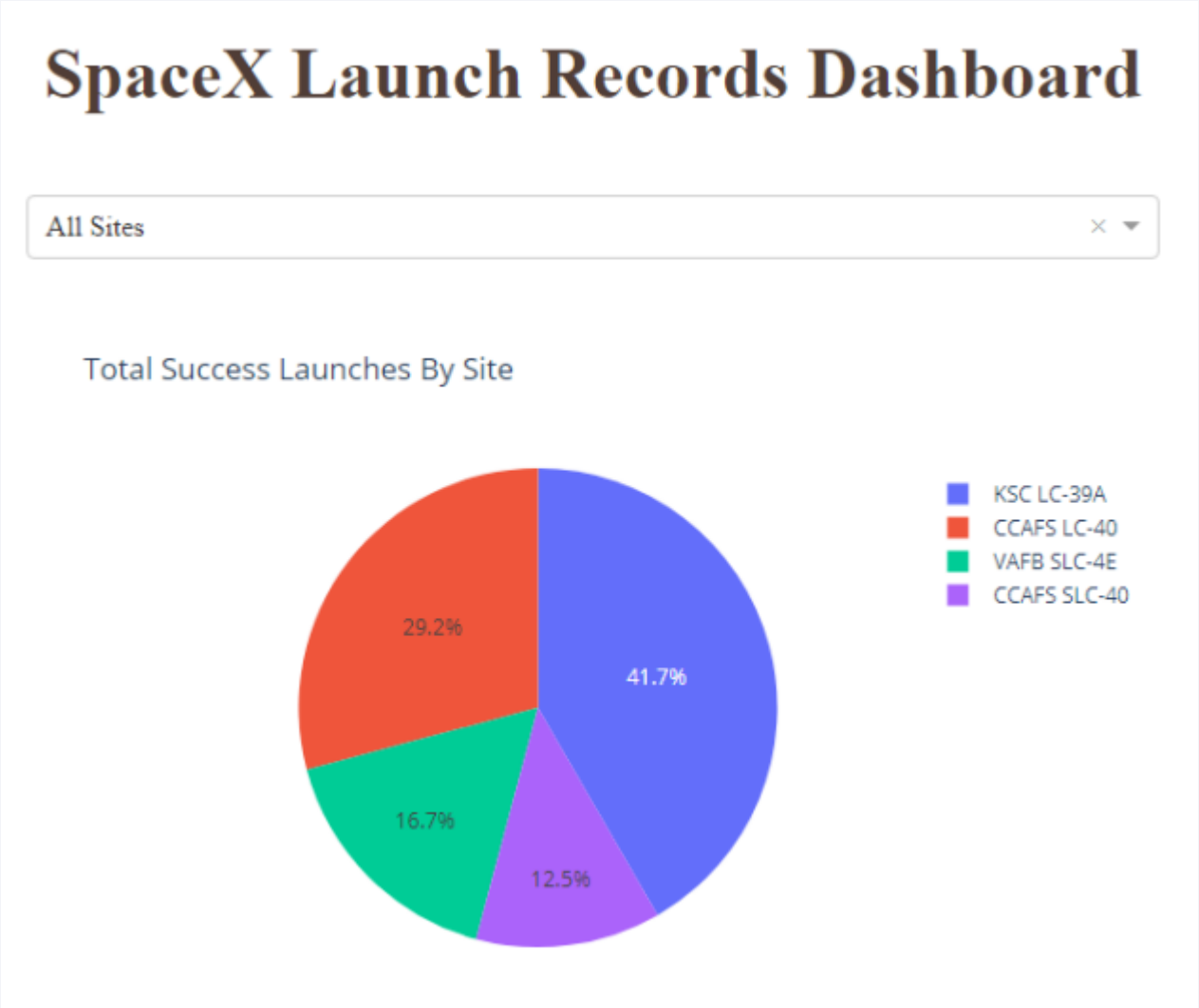





Section 4
Build a Dashboard with Plotly Dash

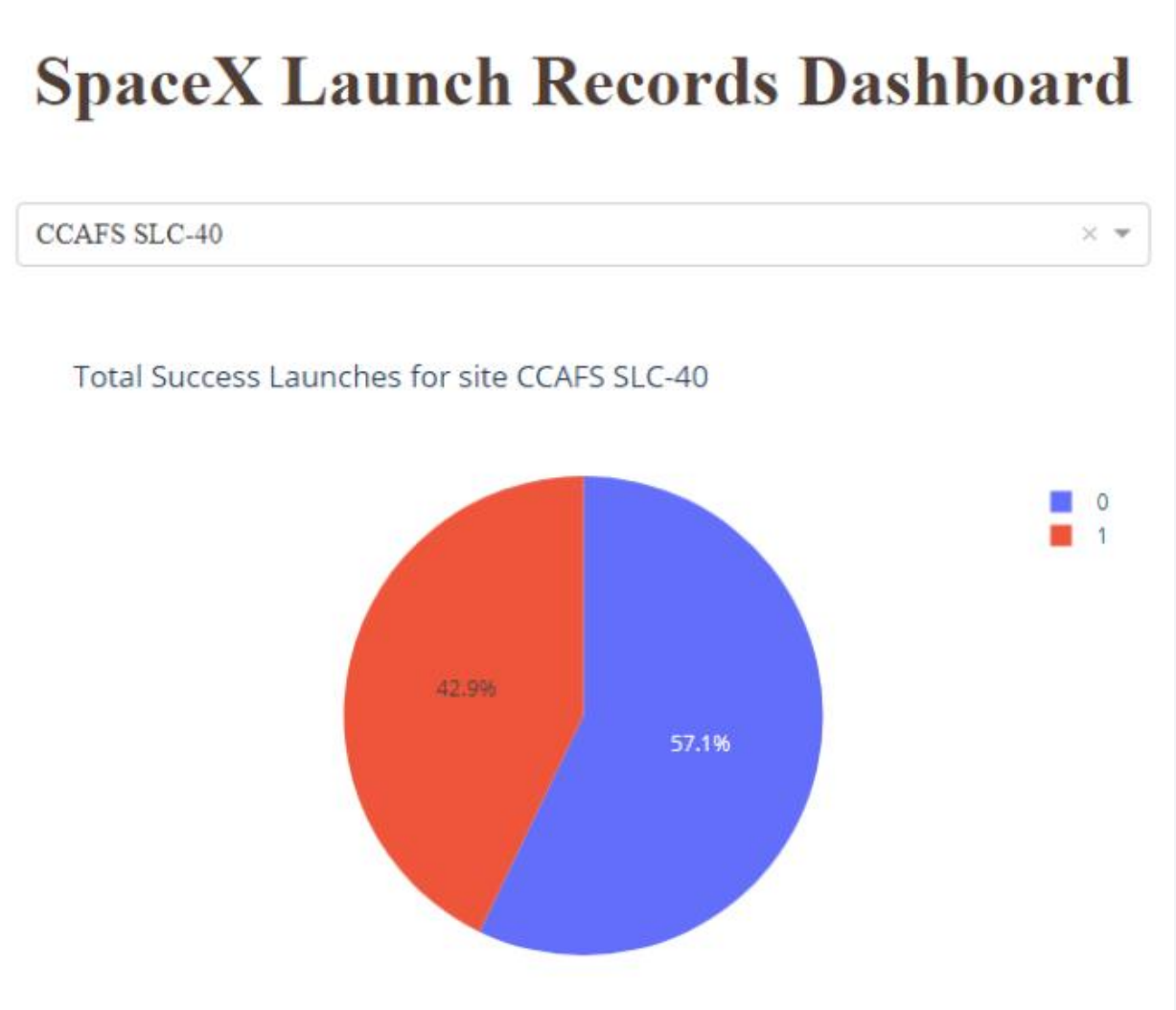
Pie Chart – Success Percentage Achieved by Launch Sites

- Using Plotly Dash, a dashboard was created;
- The first graph is a piechart containing the count of successful launches by site;
- There is also a dropdown where is possible to chose a specific Site;
- From all sites, KSC LC-39A had the most successful launches;



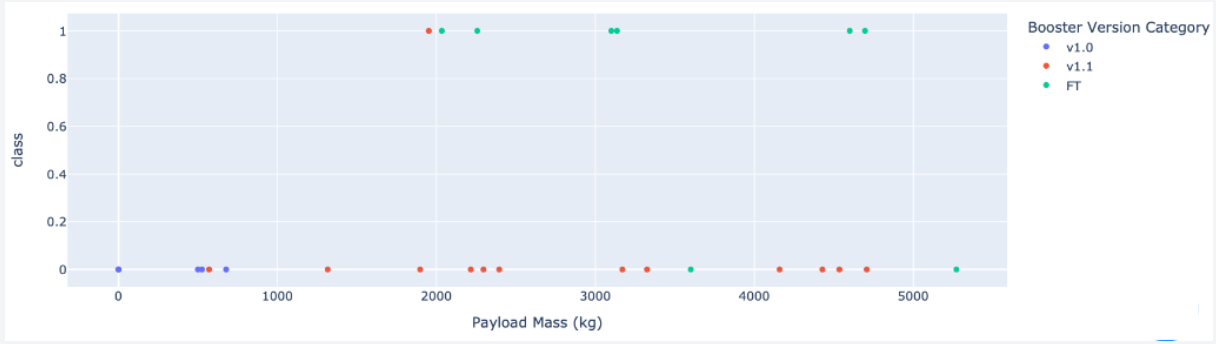
Launch Site with Highest Launch Success Ratio

- Falcon 9 first stage unsuccessful landings are denoted by the '0' Class **(blue wedge in the pie chart)**, while successful landings are represented by the '1' Class **(red wedge).**
- Among the launch sites, CCAFS SLC-40 boasted the highest success rate for Falcon 9 first stage landings at 42.9%.

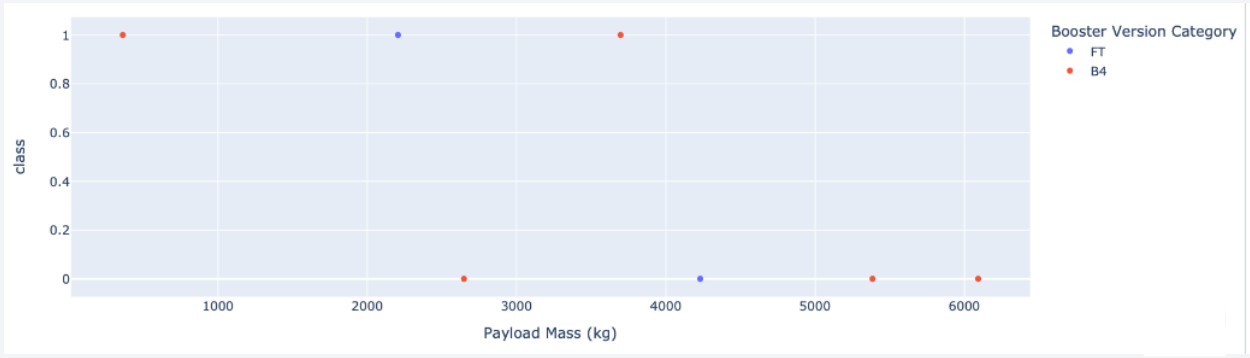


Payload vs. Launch Outcome

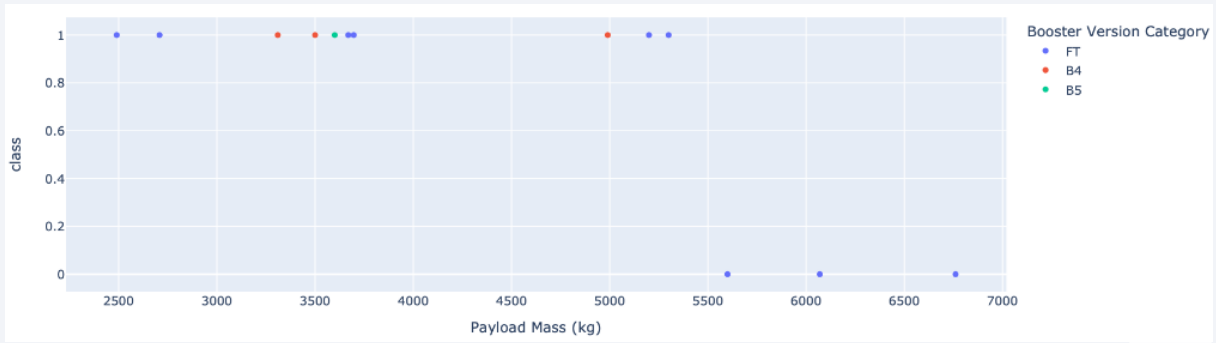
- **Correlation between Payload and Success for CCAFS LC-40**



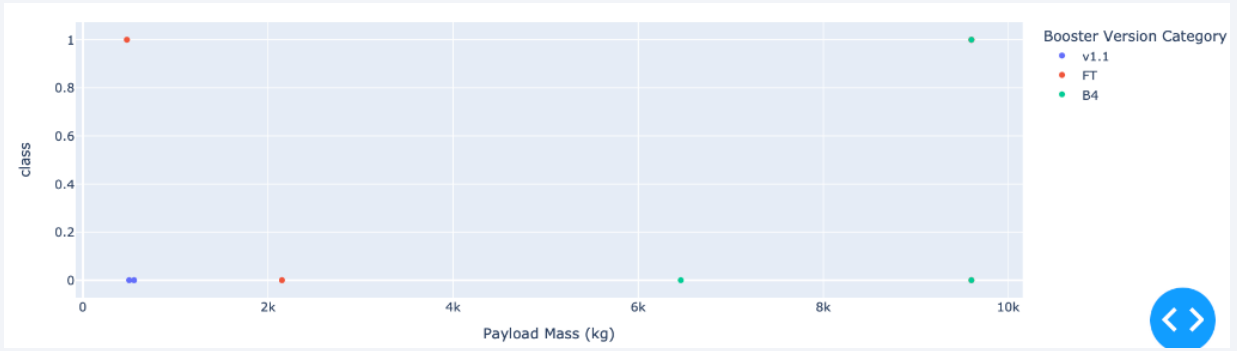
- **Correlation between Payload and Success for CCAFS SLC-40**



- **Correlation between Payload and Success for KSC LC-39A**



- **Correlation between Payload and Success for VAFB SLC-4E**



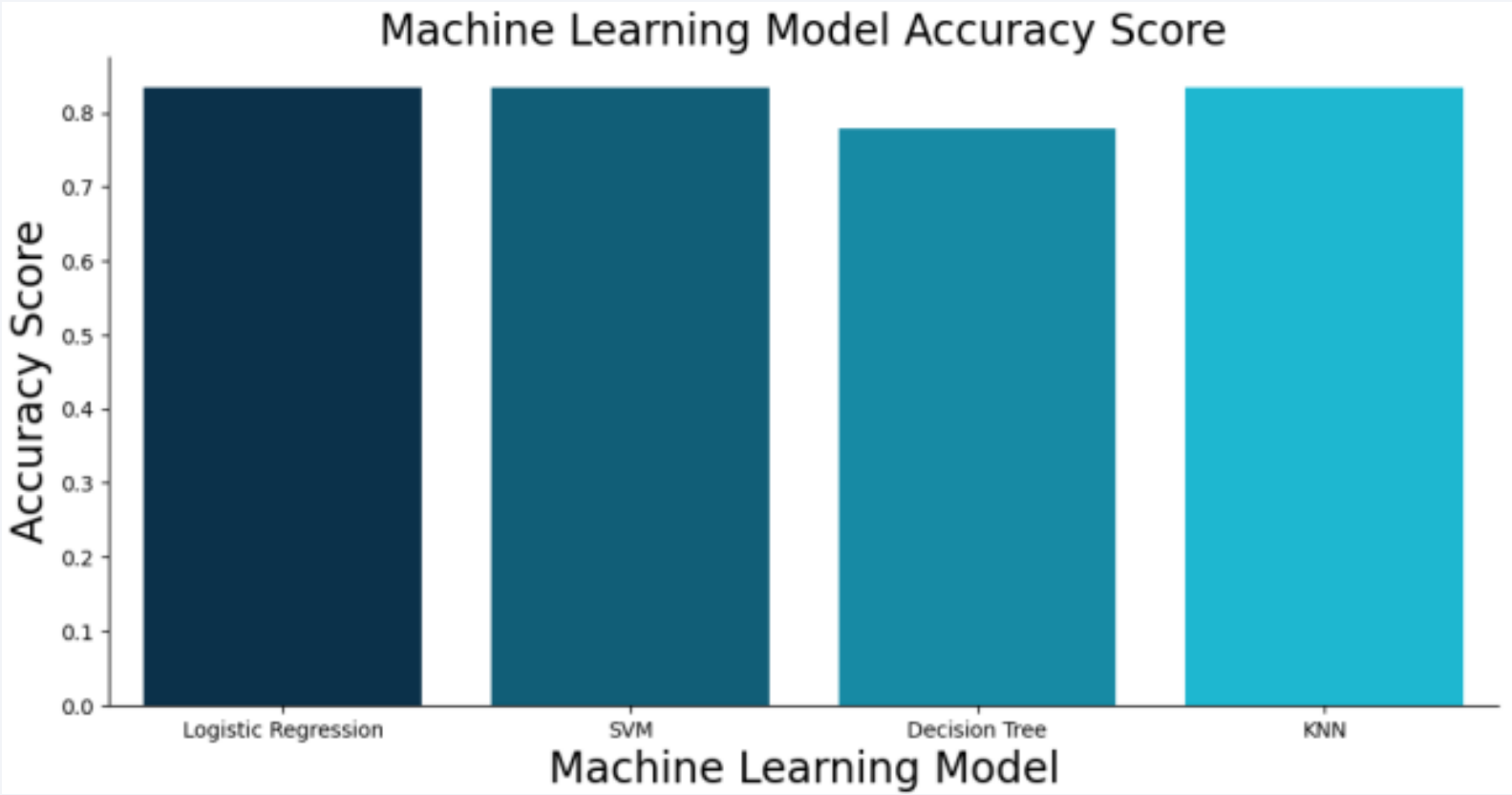
These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider. • The payload range from about 2,000 kg to 5,000 kg has the largest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

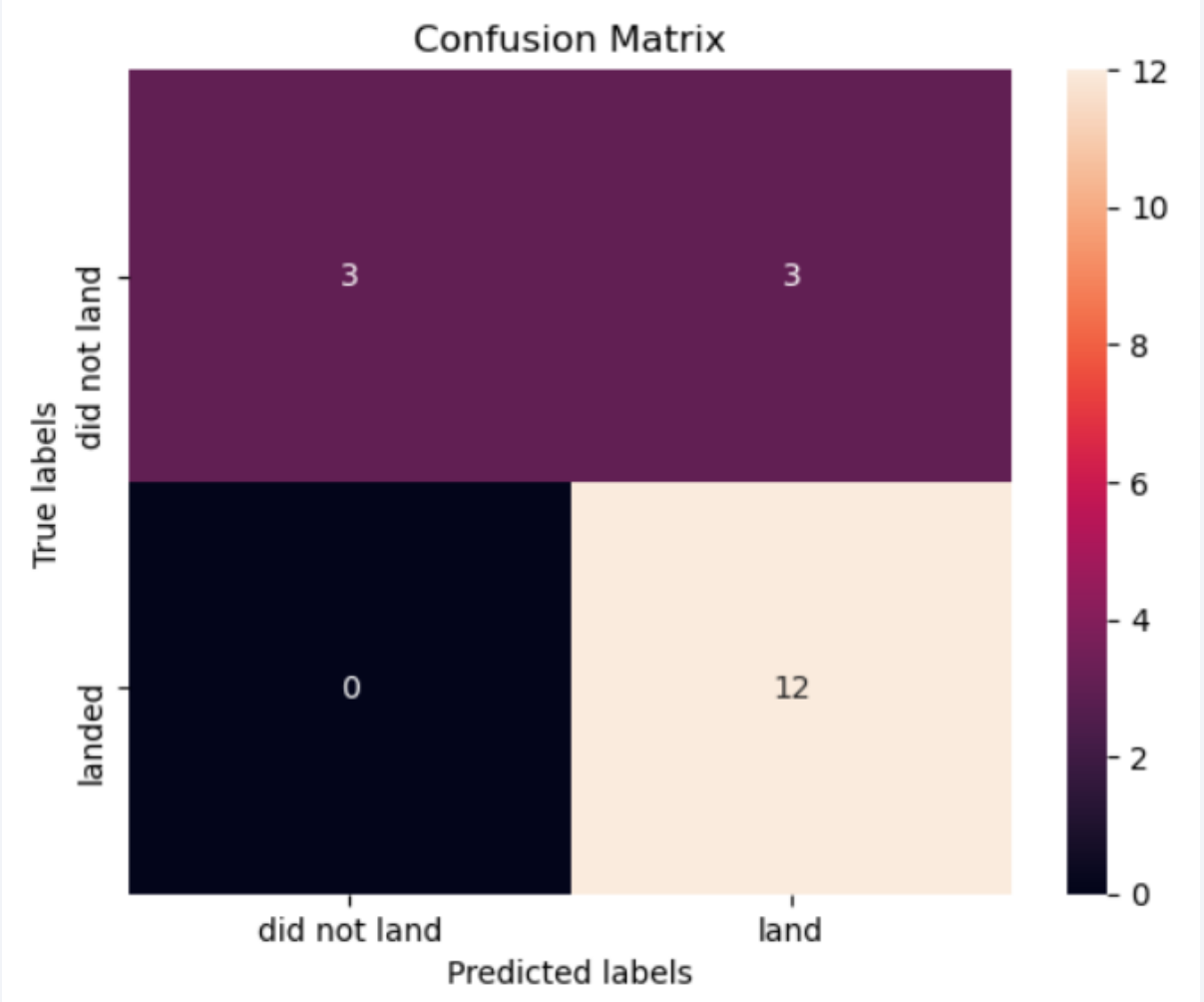
All models exhibited comparable performance, except for the Decision Tree model, which underperformed relative to the others.



Confusion Matrix

- This figure **shows the confusion matric** for the Logistic Regression model

- **Prediction Breakdown:**
- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives



Conclusions

SpaceX's Falcon 9 first-stage landing success rates have demonstrably improved over time. We observed a correlation between success rates and factors like:

- **Launch Site:** Sites with more launches tend to have higher success rates, suggesting potential improvements in infrastructure or experience.
- **Orbit Type:** Orbits like ES-L1, GEO, HEO, and SSO have significantly higher success rates, indicating potentially specialized procedures for these destinations.
- KSC LC-39A had the most successful launches of any sites
- **Overall Launch History:** Landing success rates have been on an upward trend since 2013, showcasing SpaceX's ongoing advancements.
- **Machine learning models**, particularly the decision tree classifier in this case, offer a promising avenue for predicting future landing outcomes.

Appendix

Jupiter Notebooks and Dashboard files:

- **GitHub URL (Data Collection):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
- **Github URL (Web Scraping):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-webscraping.ipynb
- **GitHub URL (Data Wrangling):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- **GitHub URL (EDA with SQL):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-eda-sql.ipynb
- **GitHub URL (EDA with Data Visualization):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/jupyter-labs-eda-data_visualization.ipynb
- **GitHub URL (Folium Maps):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/lab_jupyter_launch_site_location.ipynb
- **GitHub URL (Dashboad File):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/SpacexBash.py
- **GitHub URL (Machine Learning):** https://github.com/myriamh11/Winning_space_race_with_Data_Science_final_proj/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!