Winning Space Race with Data Science

Oleksii Maliovanyi
September 2024

Outline

Executive Summary

Predicting Falcon 9 Landing Success

Problem: Binary classification to predict Falcon 9 landing success.

Data: Compiled from SpaceX API and web scraping.

Methodology: Data cleaning, EDA, machine learning modeling (KNN).

Key Findings:

- **Launch Sites:** CCSF SLC 40, KSC LC 39A, VAFB SLC 4E.
- **Payload Mass:** Most launches under 8,000 kg, FT, B4, B5 models used for heavier payloads.
- **Orbit:** Success rate correlates with launch frequency. Heavier payloads for closer orbits.
- **Model:** K-Nearest Neighbor achieved 83.34% accuracy.

Introduction

Project background and context

- **Space Travel Costs:** Space exploration is a costly endeavor, with SpaceX charging approximately $62 million per Falcon 9 launch.
- **Financial Implications:** Accurate prediction of Falcon 9 landing success can significantly impact launch costs. A successful prediction could result in savings of up to tens of millions of dollars.

Problems you want to find answers

- This project aims to investigate whether publicly available data on historical SpaceX Falcon 9 launches can be used to develop a machine learning model that accurately predicts landing success. By leveraging this data, we can potentially provide valuable insights for future launch decisions and help reduce overall costs.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two primary sources: the SpaceX REST API and web scraping of the SpaceX Wikipedia page using BeautifulSoup.
- Perform data wrangling
  - Missing values in payload mass were replaced with the mean.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were employed to predict Falcon 9 landing success. The process involved feature selection, model training using K-Nearest Neighbors, hyperparameter tuning, and evaluation using appropriate metrics

Data Collection

Data Collection Methods

- **SpaceX REST API:** The SpaceX REST API was used to gather data on Falcon 9 launches, including booster version, launch site coordinates, payload information, flight number, and date.
- **Web Scraping:** Web scraping techniques were employed to extract critical tables from the Falcon 9 Wikipedia page, capturing data points such as launch time, payload mass, orbit, customer, and launch outcome.
- **Technical Implementation:** GET requests, BeautifulSoup, and html.parser were utilized for efficient data extraction.

Data Collection – SpaceX API

SpaceX REST API:

- Utilized the SpaceX REST API to retrieve relevant data.
- Collected data points such as:
  - Booster version
  - Longitude and latitude of launch sites
  - Payload information
  - Cores
  - Flight number
  - Date
  - Filtered the data to focus exclusively on Falcon 9 launches.
- GitHub URL of the completed SpaceX API calls and web scraping notebook: Link

```
import requests
import pandas as pd

# Perform the API request
url = "https://api.spacexdata.com/v4/launches"
response = requests.get(url)
data = response.json()

# Convert the response to a DataFrame
df = pd.json_normalize(data)

# Check the first row of the column 'static_fire_date_utc'
first_date = df['static_fire_date_utc'].iloc[0]
year = pd.to_datetime(first_date).year
print(year)
```

```
2006
```

Data Collection - Scraping

Web Scraping

- Employed web scraping techniques to extract critical tables from the Falcon 9 Wikipedia page.
- Used GET requests, BeautifulSoup, and html.parser for efficient data extraction.
- GitHub URL of the completed SpaceX API calls and web scraping notebook: Link

```
display(df.head(10))
display(df.tail(10))
df.to_csv('spacex_web_scraped.csv', index=False)
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10
5	6	VAFB	CASSIOPE	500 kg	Polar orbit	MDA	Success	F9 v1.17B10038	Uncontrolled	29 September 2013	16:00
6	7	CCAFS	SES-8	3,170 kg	GTO	SES	Success	F9 v1.1	No attempt	3 December 2013	22:41
7	8	CCAFS	Thaicom 6	3,325 kg	GTO	Thaicom	Success	F9 v1.1	No attempt	6 January 2014	22:06
8	9	Cape Canaveral	SpaceX CRS-3	2,296 kg	LEO	NASA	Success\n	F9 v1.1	Controlled	18 April 2014	19:25
9	10	Cape Canaveral	Orbcomm-OG2	1,316 kg	LEO	Orbcomm	Success	F9 v1.1	Controlled	14 July 2014	15:15

Data Wrangling

- **Missing Value Handling**: Checked for missing data in each variable and expressed the percentage of missing values.
- **Data Type Conversion:** Verified data types using the dtypes method to ensure compatibility with subsequent analysis and modeling.
- **Variable Exploration:** Conducted value_counts() on key variables to understand their distributions and identify potential outliers or imbalances.
- **Feature Engineering:** Created a new binary "Class" variable based on the "Outcome" variable to categorize successful and unsuccessful landings.
- **GitHub URL** of data wrangling related notebook Link

EDA with Data Visualization

Summary of Charts:

- **Scatterplot of Flight Number, Payload Mass, and Class:**
  - Assessed the relationship between flight number, payload mass, and landing success.
  - Identified any trends or patterns in the data.
- **Scatterplot of Flight Number, Launch Site, and Class:**
  - Examined the distribution of launch sites and their impact on landing success.
  - Looked for any notable outliers or correlations between launch site and outcome.
- **Bar Chart Showing Success Rate of Each Type of Orbit:**
  - Compared the success rates of different orbits.
  - Identified orbits with significantly higher or lower success rates.
- **Scatterplot of Orbit, Flight Number, and Class:**
  - Analyzed the relationship between orbit, flight number, and landing success.
  - Discovered any patterns or dependencies.
- **Scatterplot of Orbit, Payload Mass, and Class:**
  - Assessed the relationship between orbit, payload mass, and landing success.
  - Determined if there were any specific payload mass ranges suitable for certain orbits.
- **Line Plot of Mean Annual Success Rate:**
  - Visualized the trend of landing success over time.
  - Assessed whether the program has been improving in terms of landing success.

- **GitHub URL** of completed EDA with data visualization notebook Link

EDA with SQL

Identify Unique Launch Sites:

```
SELECT DISTINCT Launch_Site FROM launch_data;
```

Filter Launch Sites:

```
SELECT * FROM launch_data WHERE Launch_Site LIKE '%CCA%';
```

Calculate Payload Mass for NASA Missions:

```
SELECT SUM(Payload Mass kg) FROM launch_data WHERE
Customer = 'NASA (CRS)';
```

Calculate Mean Payload Mass for F9 v1.1

```
SELECT AVG(Payload_Mass_kg) FROM launch_data WHERE
Booster_Version = 'F9 v1.1';
```

Find Date of First Successful Ground Pad Landing:

```
SELECT MIN(Date) FROM launch_data WHERE Launch_Outcome =
'Success (ground pad)';
```

List Boosters with Success on Drone Ship and Specific Payload Mass: `SELECT Booster_Version FROM launch_data WHERE Launch Outcome = 'Success (drone ship)' AND Payload_Mass_kg BETWEEN 4000 AND 6000;`

Count Successful and Failed Missions: `SELECT Launch Outcome, COUNT(*) FROM launch_data GROUP BY Launch_Outcome;`

Find Booster Versions with Maximum Payload Mass: `SELECT Booster_Version FROM launch_data WHERE Payload_Mass_kg = (SELECT MAX(Payload_Mass_kg) FROM launch_data);`

Extract Month Names, Failure Outcomes, Booster Versions, and Launch Sites for 2015: `SELECT SUBSTR(Date, 6, 2) AS Month, Booster_Version, Launch_Site FROM launch_data WHERE Launch Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 1, 4) = '2015';`

Rank Landing Outcomes Within a Date Range: `SELECT Launch Outcome, COUNT(*) AS Outcome Count FROM launch_data WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Launch_Outcome ORDER BY Outcome_Count DESC;`

GitHub URL of completed EDA with SQL notebook Link

Build an Interactive Map with Folium

- **Launch Site Visualization:** Circles and markers were created to represent launch sites on the map.
- **Outcome Differentiation:** Green markers were used for successful launches, while red markers indicated unsuccessful launches.
- **Proximity Analysis:** A marker was placed on the nearest coastline, and the distance between the launch site and the coastline was calculated.
- **Infrastructure and Safety:** Markers were added to nearby highway, railroad, and city, along with distance calculations to assess logistical considerations and potential safety hazards.
- **Polyline Connection:** A polyline was drawn between the launch site and the coastline to visualize the distance.
- **GitHub URL** of completed interactive map with Folium map Link

Build a Dashboard with Plotly Dash

- **Launch Site Selection:** A dropdown input component was added to allow users to select specific launch sites.
- **Success Pie Chart:** A pie chart was created to visualize the distribution of successful and unsuccessful launches for the selected launch site.
- **Payload Range Slider:** A range slider was implemented to enable users to select a specific payload range.
- **Success Payload Scatter Plot:** A scatter plot was created to examine the relationship between payload mass and landing success for the selected launch site.
- **Interactive Analysis:** These interactive components provided a powerful tool for business colleagues to explore the data and gain valuable insights.
- **GitHub URL** of completed Plotly Dash Link

Predictive Analysis (Classification)

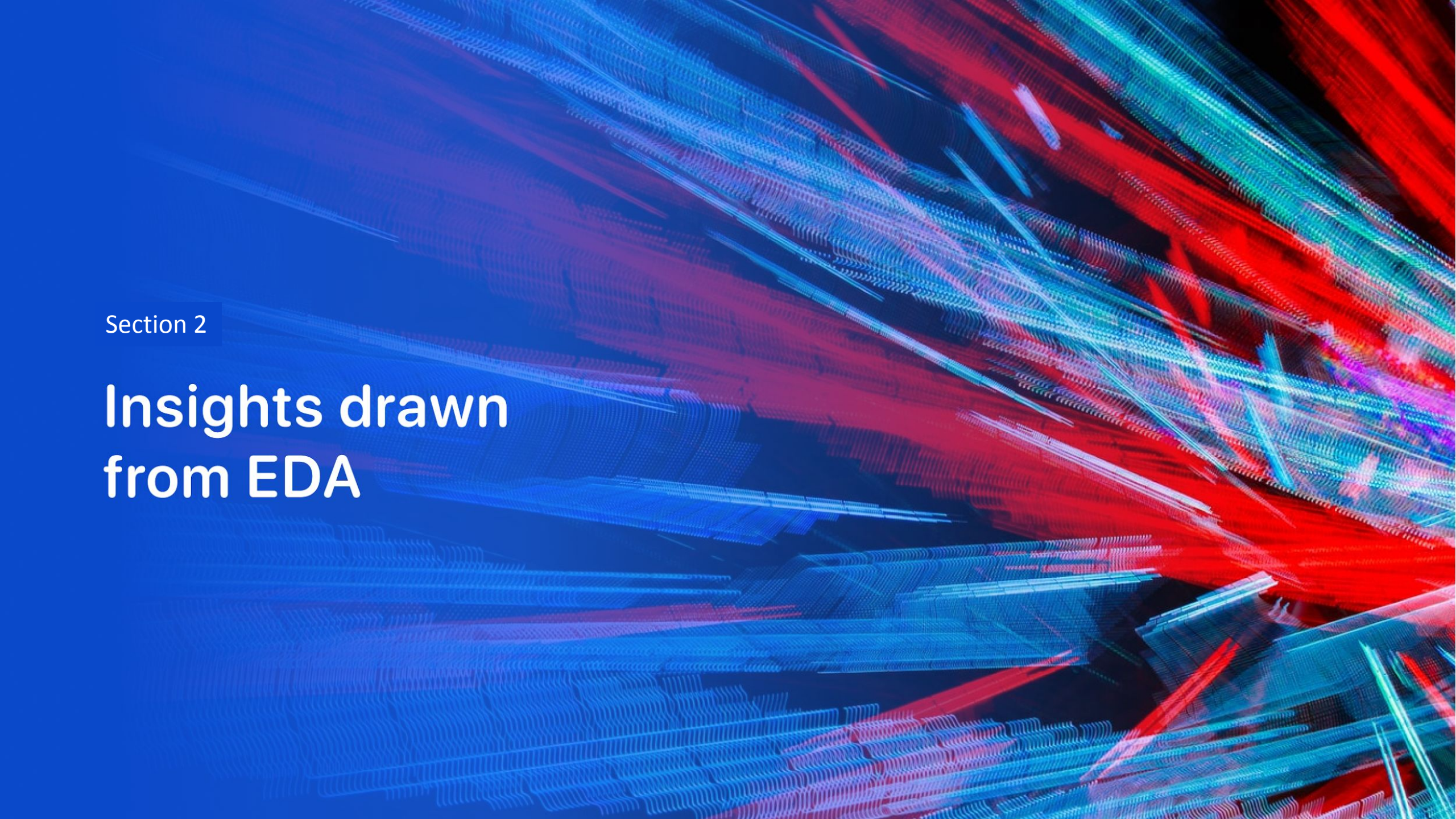
- **Data Preparation:** Created a Numpy array for the target variable (Y) and standardized the features (X) using StandardScaler. Performed a train-test split with an 80:20 ratio.
- **Model Selection:** Considered various classification algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors.
- **Model Training and Evaluation:** Trained each model on the training set and evaluated their performance on the testing set using accuracy, precision, recall, F1-score, and confusion matrices.
- **Hyperparameter Tuning:** Experimentally tuned hyperparameters for each model to optimize performance.
- **Model Comparison:** Compared the performance of different models and selected the best-performing one based on the chosen evaluation metrics.
- **GitHub URL** of completed predictive analysis Link

Results

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

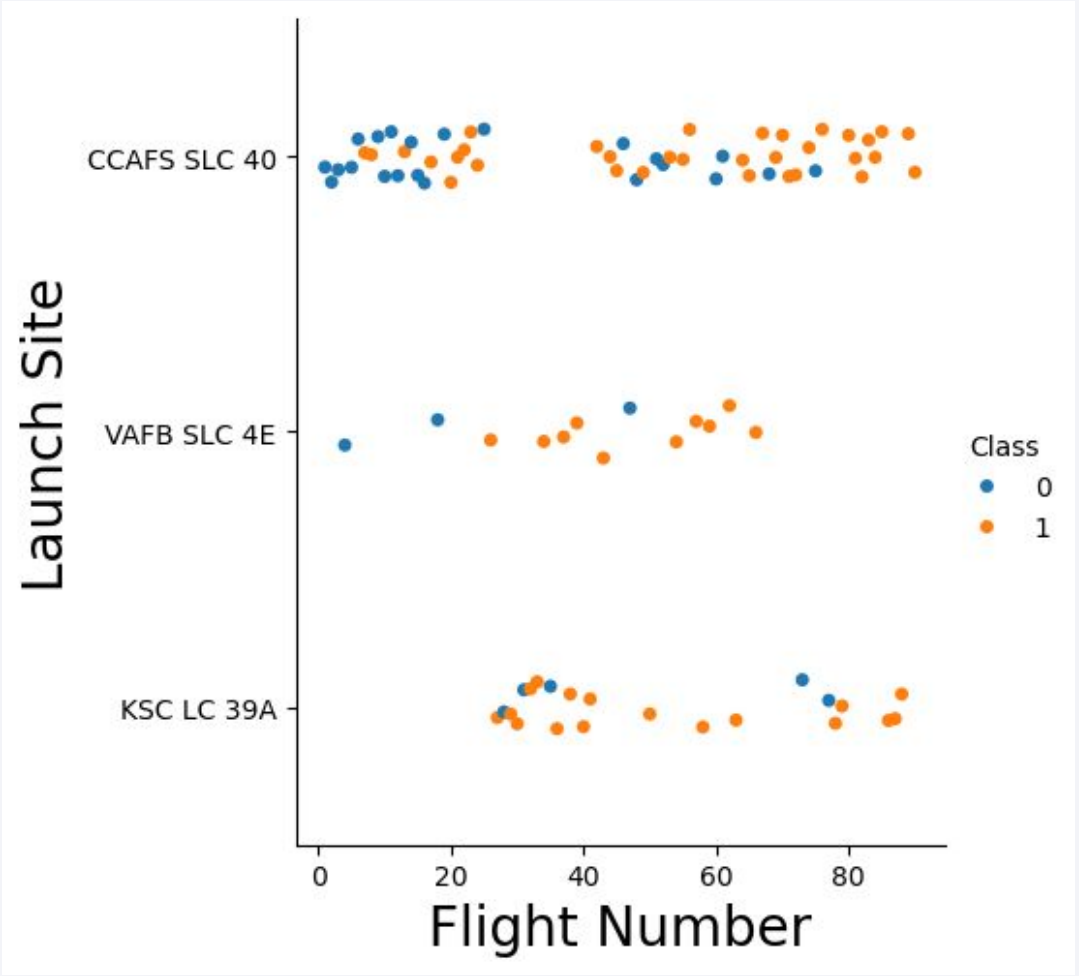
Section 2

Insights drawn from EDA



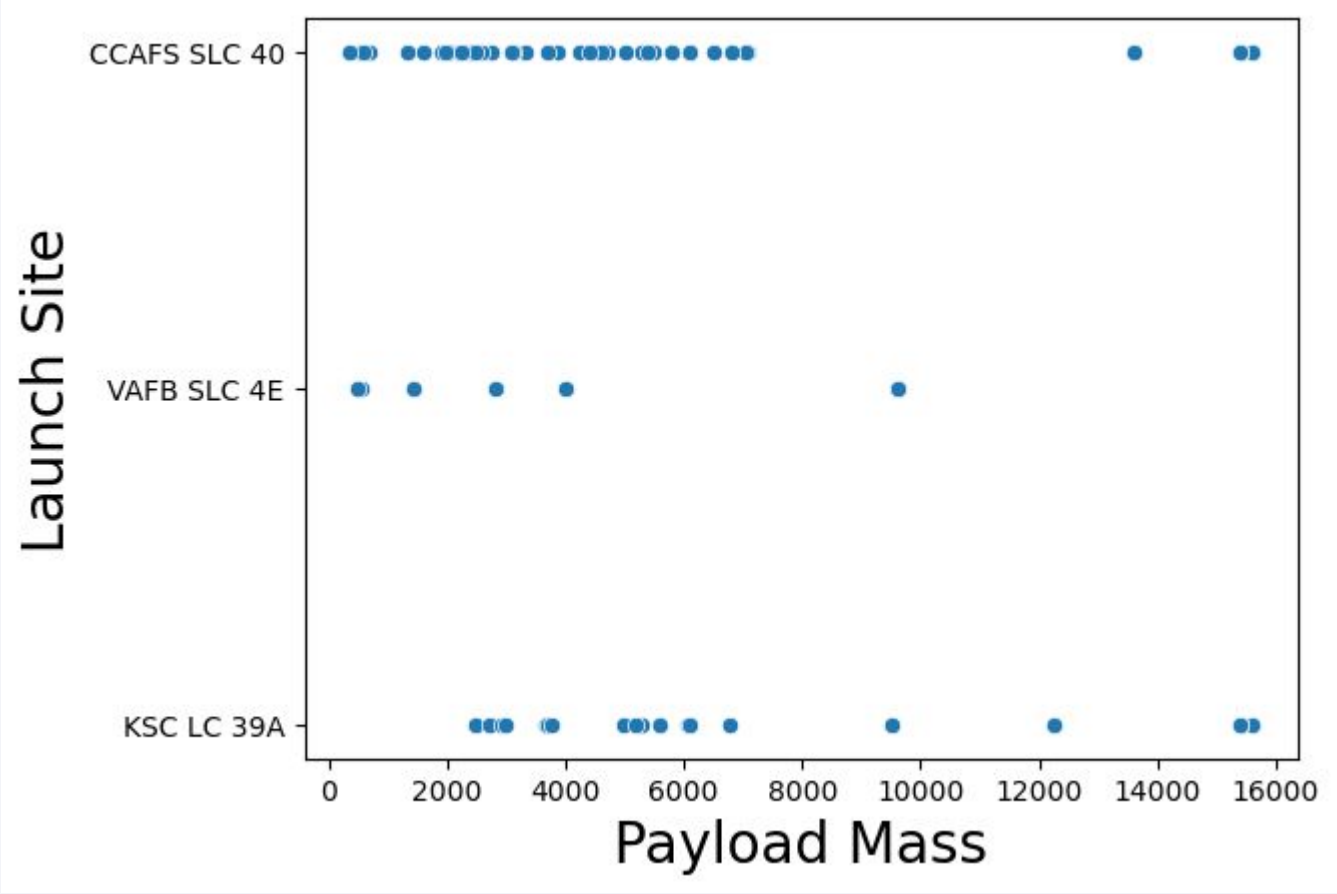
Flight Number vs. Launch Site

- **CCAFS SLC 40** is the most frequently used launch site.
- All three launch sites have experienced both successful and unsuccessful launches.
- **KSC LC 39A** seems to have a higher proportion of unsuccessful launches compared to the other two sites.



Payload vs. Launch Site

- CCAFS SLC 40 is the most versatile launch site, capable of handling a wide range of payload masses.
- VAFB SLC 4E and KSC LC 39A are more specialized, with a focus on lighter payloads.



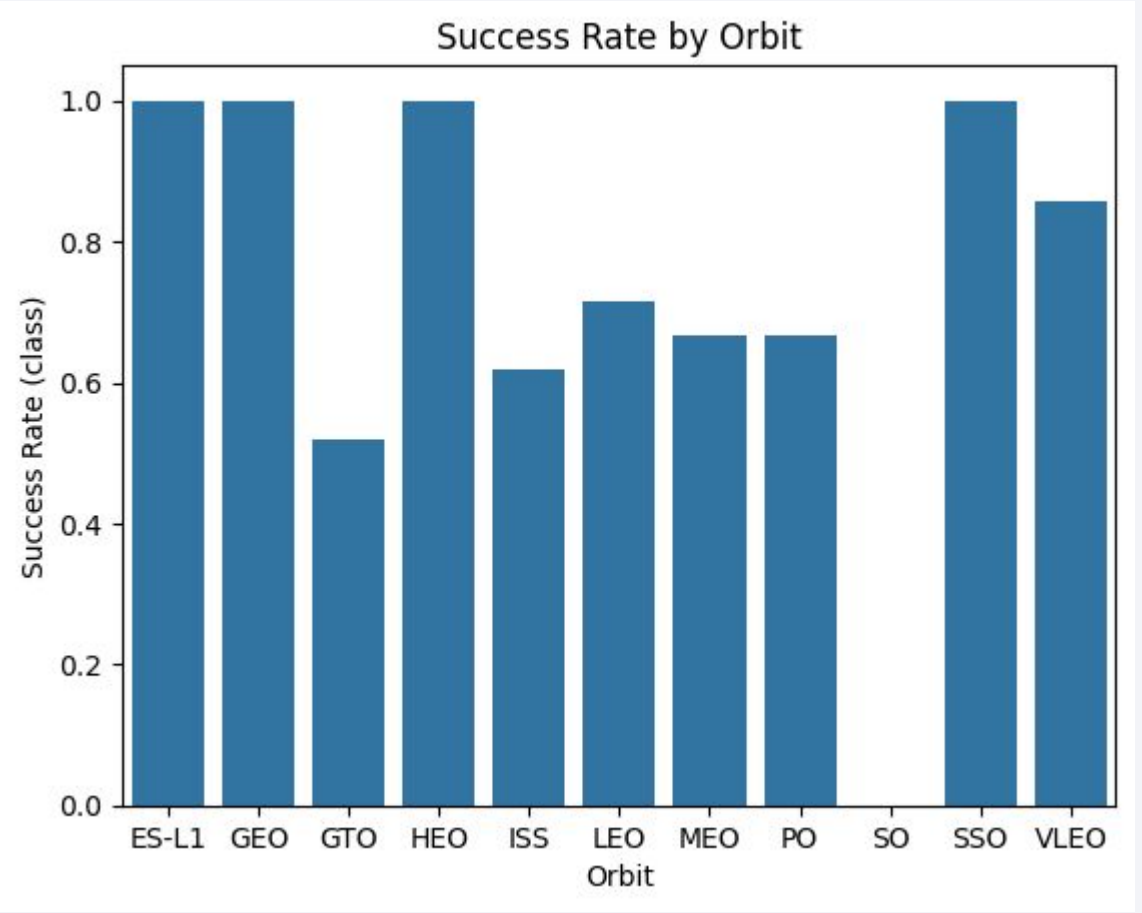
Success Rate vs. Orbit Type

Key Observations

- Highest Success Rates: ES-L1, GEO, GTO, HEO, and SSO orbits have the highest success rates, approaching or reaching 100%.
- Lowest Success Rates: VLEO and MEO orbits have the lowest success rates, below 50%.
- Medium Success Rates: PO, LEO, and ISS orbits have moderate success rates, ranging from approximately 60% to 80%.

Overall

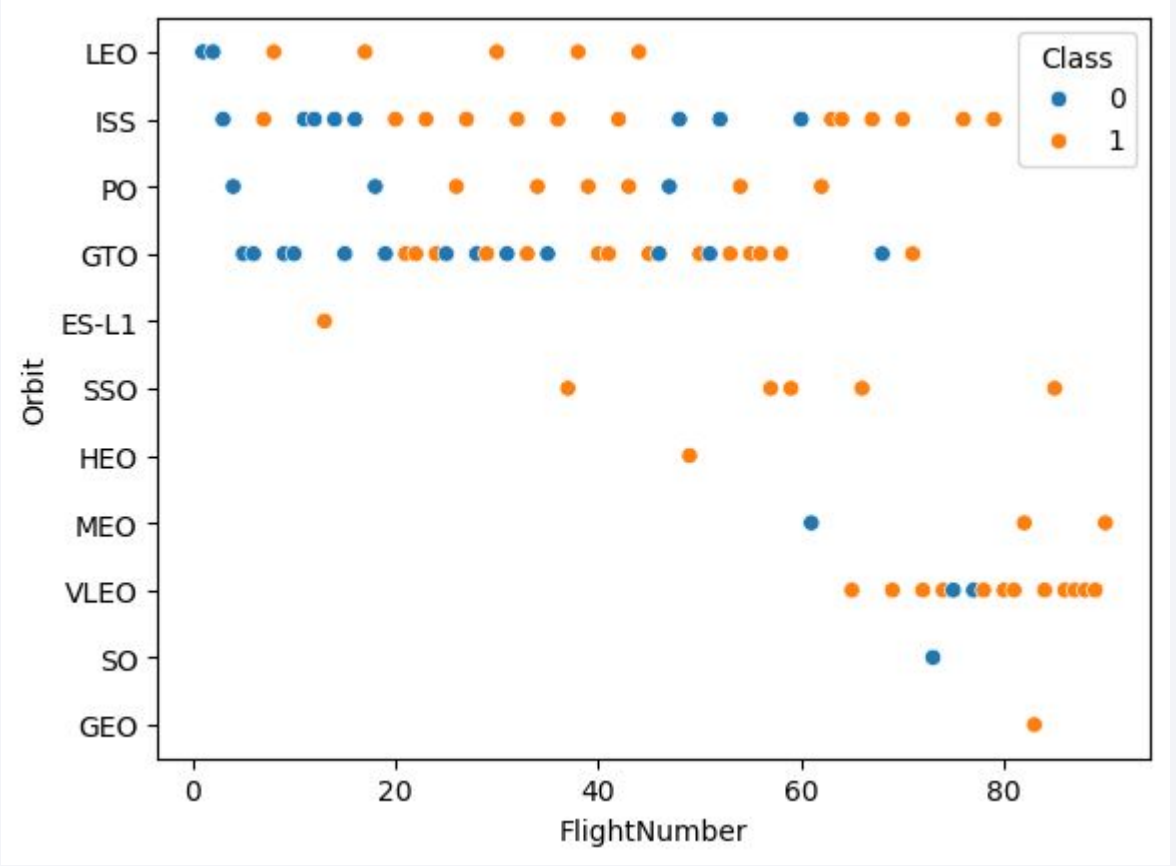
- Certain orbits are more conducive to successful launches than others.
- Factors such as the complexity of the orbit, distance from Earth, and other variables may influence the success rate.



Flight Number vs. Orbit Type

Key Observations

- **Orbit Clustering:** Certain orbits have a higher concentration of flights. For example, LEO, ISS, PO, and GTO have a larger number of flights compared to other orbits.
- **Success and Failure Patterns:** Some orbits have a higher proportion of successful launches (class 1), while others have a higher proportion of unsuccessful launches (class 0).
- **Flight Number Trends:** Within each orbit, there may be trends in the success or failure of flights based on their flight number.



Overall

- **Orbit Choice:** Orbit choice affects launch success.
- **Orbit Challenges:** Some orbits are more difficult or risky.

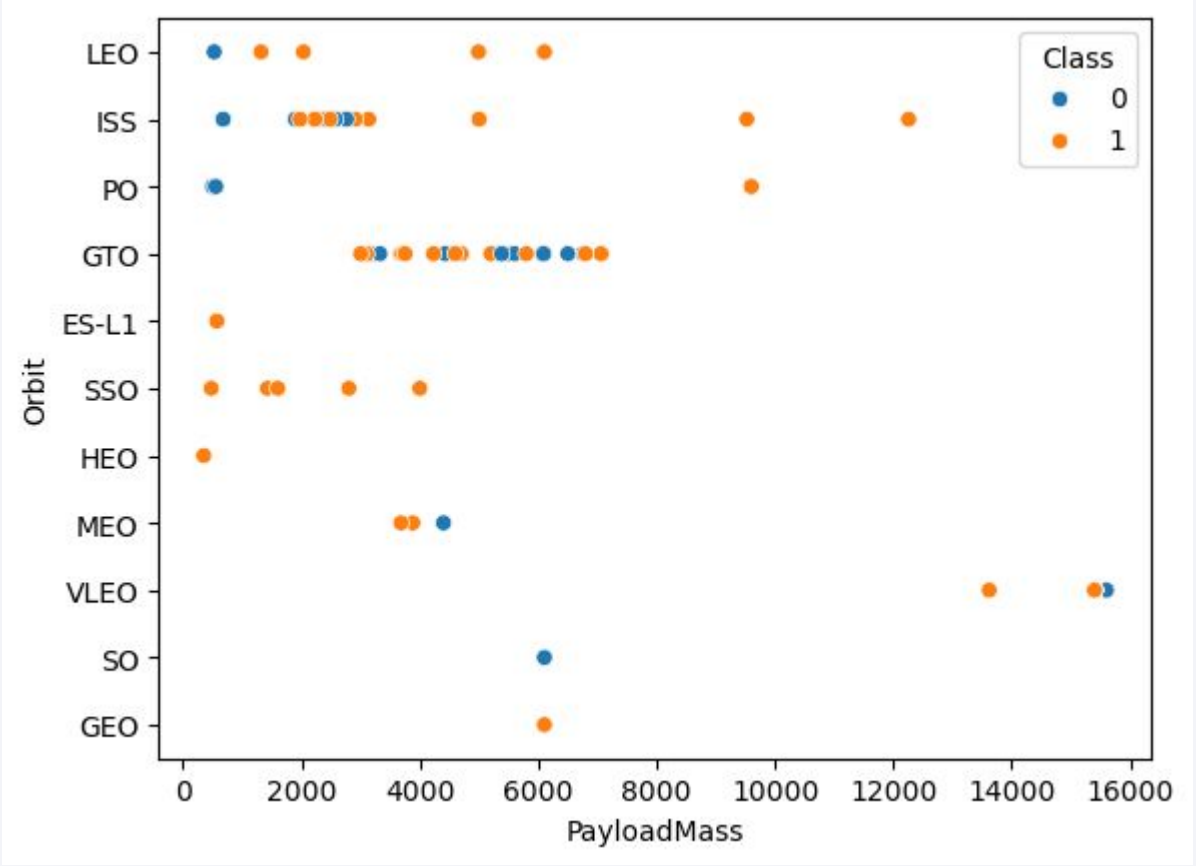
Payload vs. Orbit Type

Key Observations

- **Orbit Concentration:** Some orbits have more frequent launches (LEO, ISS, PO, GTO).
- **Success/Failure Patterns:** Orbit success rates vary.
- **Flight Number Trends:** Success/failure may depend on flight number within orbits.

Overall

- Orbit choice affects launch success.
- Some orbits are more challenging or risky.



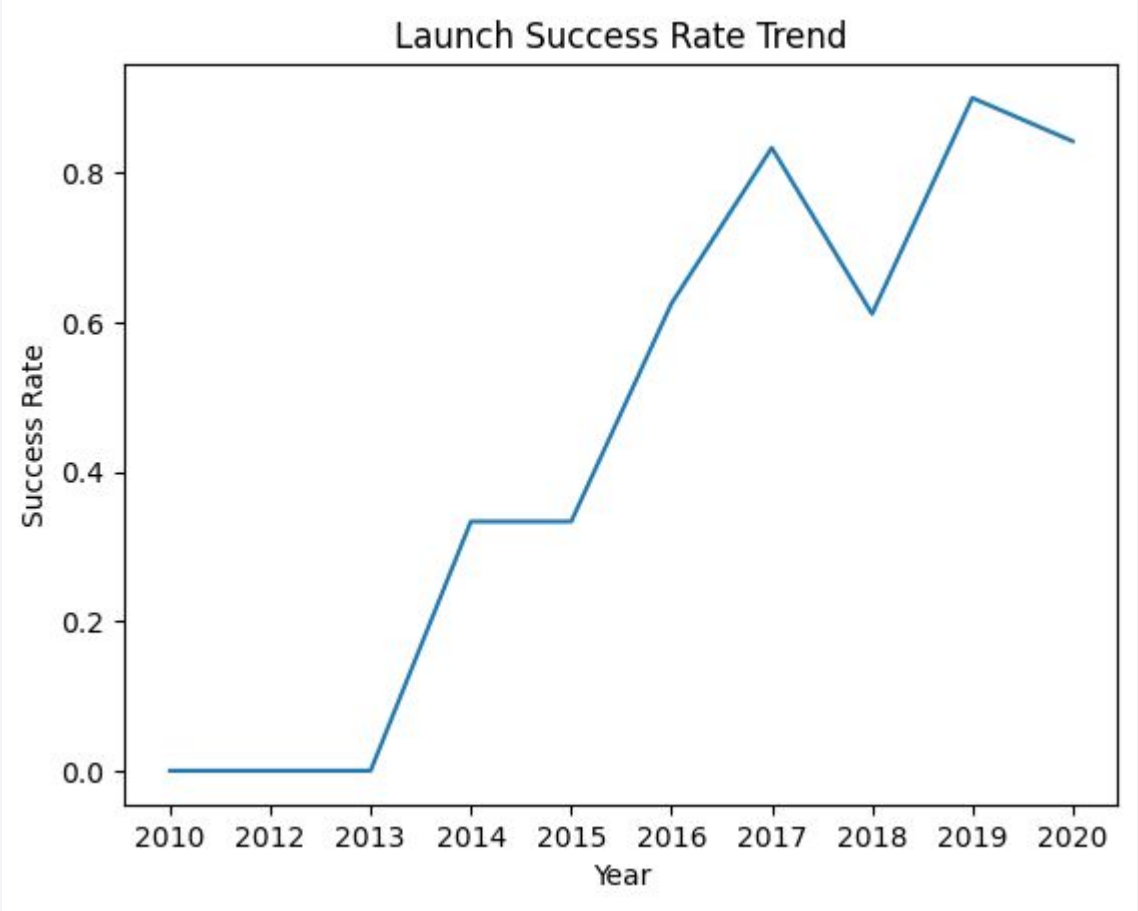
Launch Success Yearly Trend

Key Observations

- Early Struggles: Low success rate from 2010 to 2013.
- Significant Improvement: Success rate increased significantly in 2014.
- Continued Growth: Steady improvement from 2014 to 2019.
- Slight Decline: Minor setback in 2020, but overall success rate remains high.

Overall

- SpaceX has made significant strides in improving launch reliability over the years.
- The trend indicates a positive trajectory for future launches.



All Launch Site Names

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

In [11]:
```
%sql select distinct("LAUNCH_SITE") from SPACEXTBL
```

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

Out[11]:

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where "LAUNCH_SITE" like "CCA%" LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_)  from SPACEXTBL where Booster_Version = "F9 v1.1"
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
## %sql select min(Date) from SPACEXTBL where Landing_Outcome = "Success"
%sql select min(Date) from SPACEXTBL where Landing_Outcome = "Success (ground pad)"
```

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

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

In [25]:
```
# %sql select distinct(Landing_Outcome) from SPACEXTBL
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ >
```

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

Out[25]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1038.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(Mission_Outcome) from SPACEXTBL where Mission_Outcome = "Success" or Mission_Outcome = "ailure
```

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

count(Mission_Outcome)
98

```
%sql SELECT Mission_Outcome, COUNT(*) AS Total_Count FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

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

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE Payload_Mass__KG_ = (SELECT MAX(Payload_Mass__KG_) FROM SPACEXTE
```

```
 * 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

```
# %sql select month(Date), Landing_Outcome, Booster_Version Launch_Site FROM SPACEXTBL where year(Date) = 2015
%sql SELECT Date, substr(Date, 6,2) AS Month_Name, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL W
```

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

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

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

In [49]:

```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count \
FROM SPACEXTBL \
WHERE DateTime(DATE) BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY Outcome_Count DESC;
```

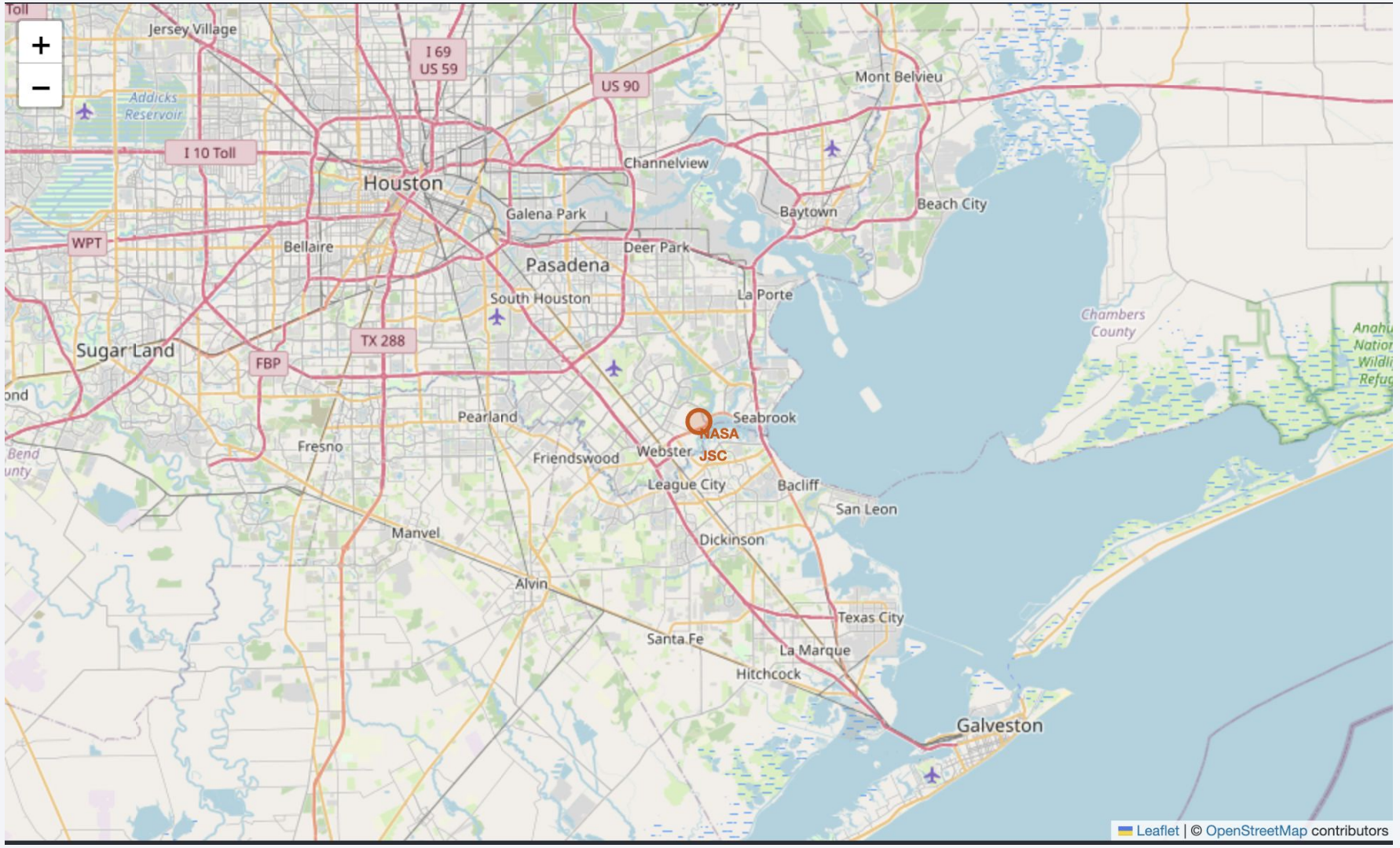
```
 * sqlite:///my_data1.db
Done.
```

Out[49]:

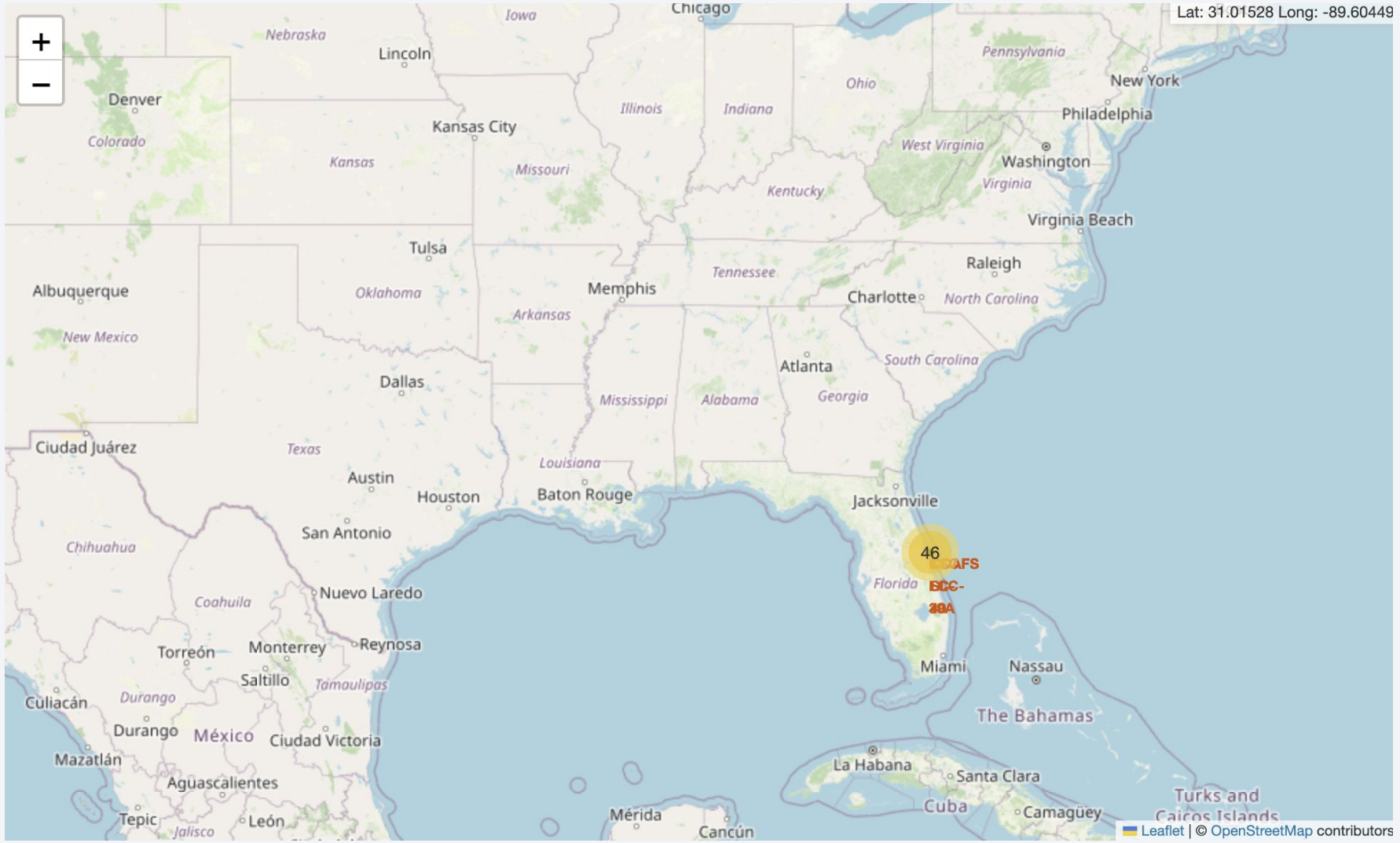
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

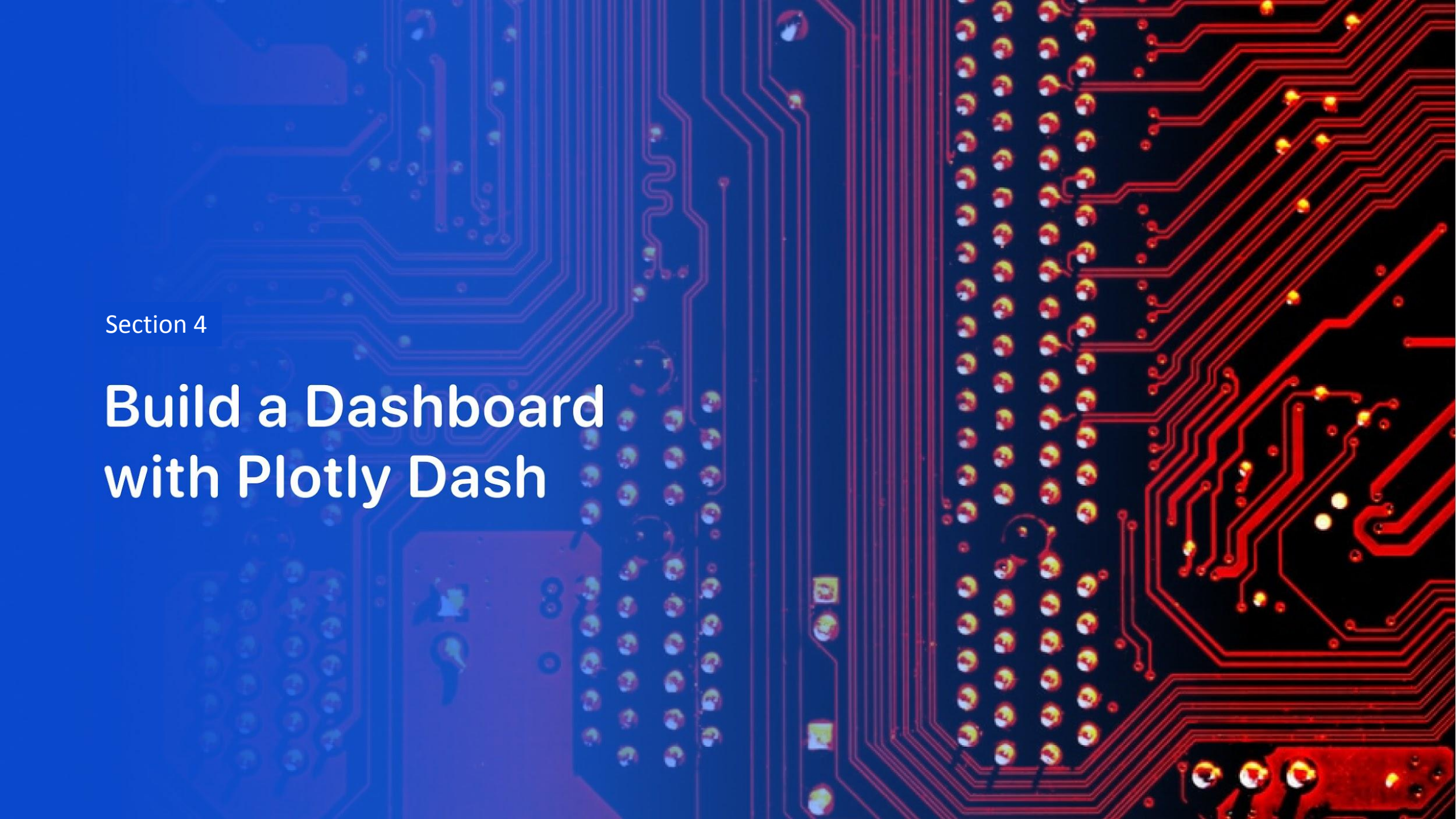
Interactive analysis: Folium leaflet of SpaceX launch sites



Interactive analysis: success and failure of SpaceX launches



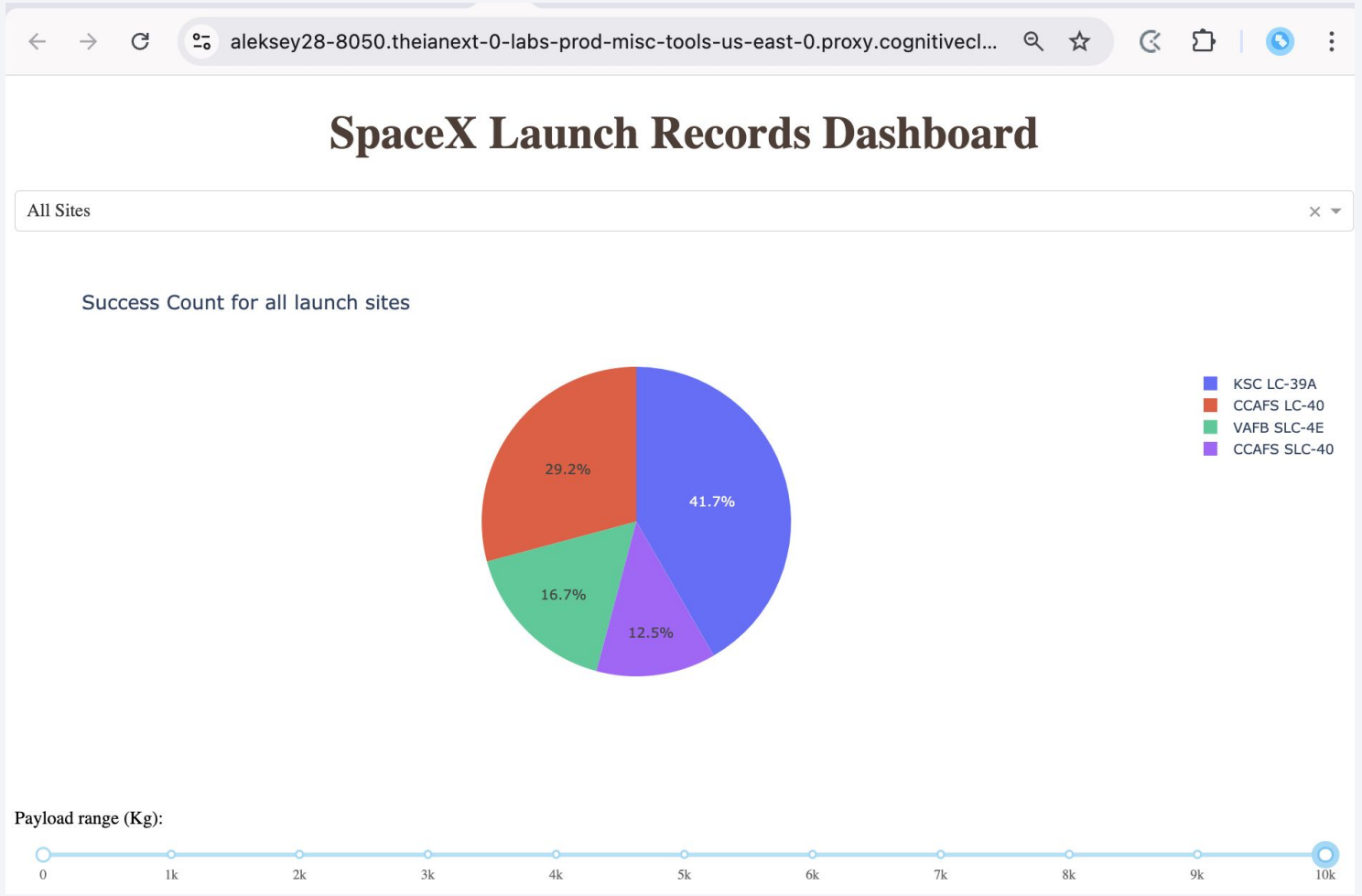
Interactive analysis: Polylines to landmarks


Section 4
Build a Dashboard with Plotly Dash

Dashboard: success count for all sites

Overall

- CCAFS SLC-40 is the most successful launch site.
- KSC LC-39A is also a successful launch site, but not as consistently as CCAFS SLC-40.
- VAFB SLC-4E and CCAFS SLC-40 have lower success rates compared to the other two sites.



Dashboard: launch site with highest launch success ratio

Pie Chart

- 57.1% of launches from CCAFS SLC-40 were successful.
- 42.9% of launches from CCAFS SLC-40 were unsuccessful.

Scatter Plot

- There seems to be a general trend where launches with lower payload mass have a higher success rate.
- Booster version category (FT or B4) may also play a role in launch success, but the sample size is relatively small to draw definitive conclusions.

Dashboard: Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider 3k - 6k

Pie Chart

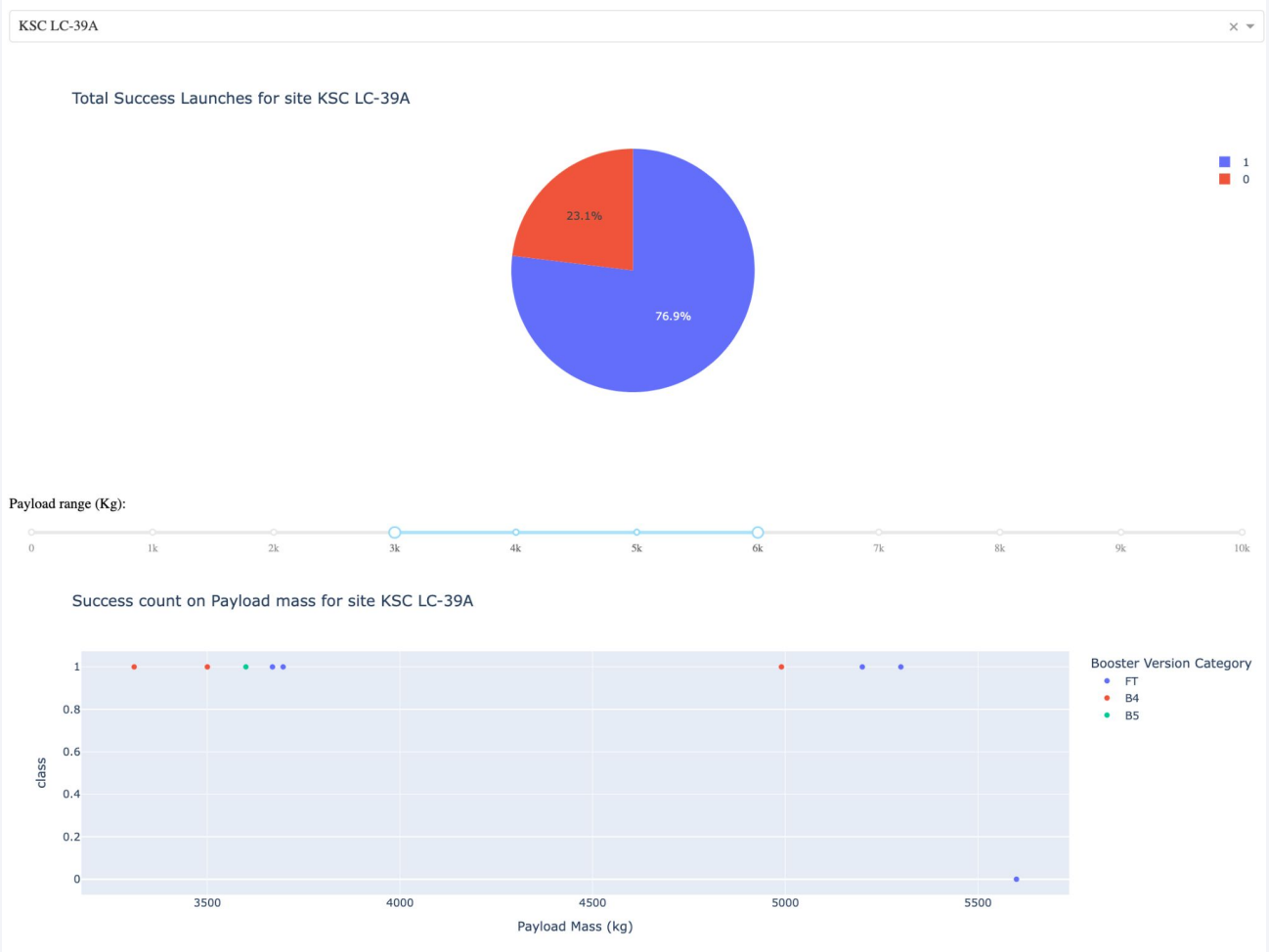
- Shows successful vs. unsuccessful launches from KSC LC-39A. 76.9% successful, 23.1% unsuccessful

Scatter Plot

- Shows payload mass vs. launch success. X-axis: payload mass (kg). Y-axis: success rate (class). Color: booster version (FT, B4, B5).

Key Findings

- **Success:** Most KSC LC-39A launches were successful.
- **Payload Mass:** Lower payload mass may correlate with higher success.
- **Booster Version:** B4 and B5 might have higher success rates than FT.

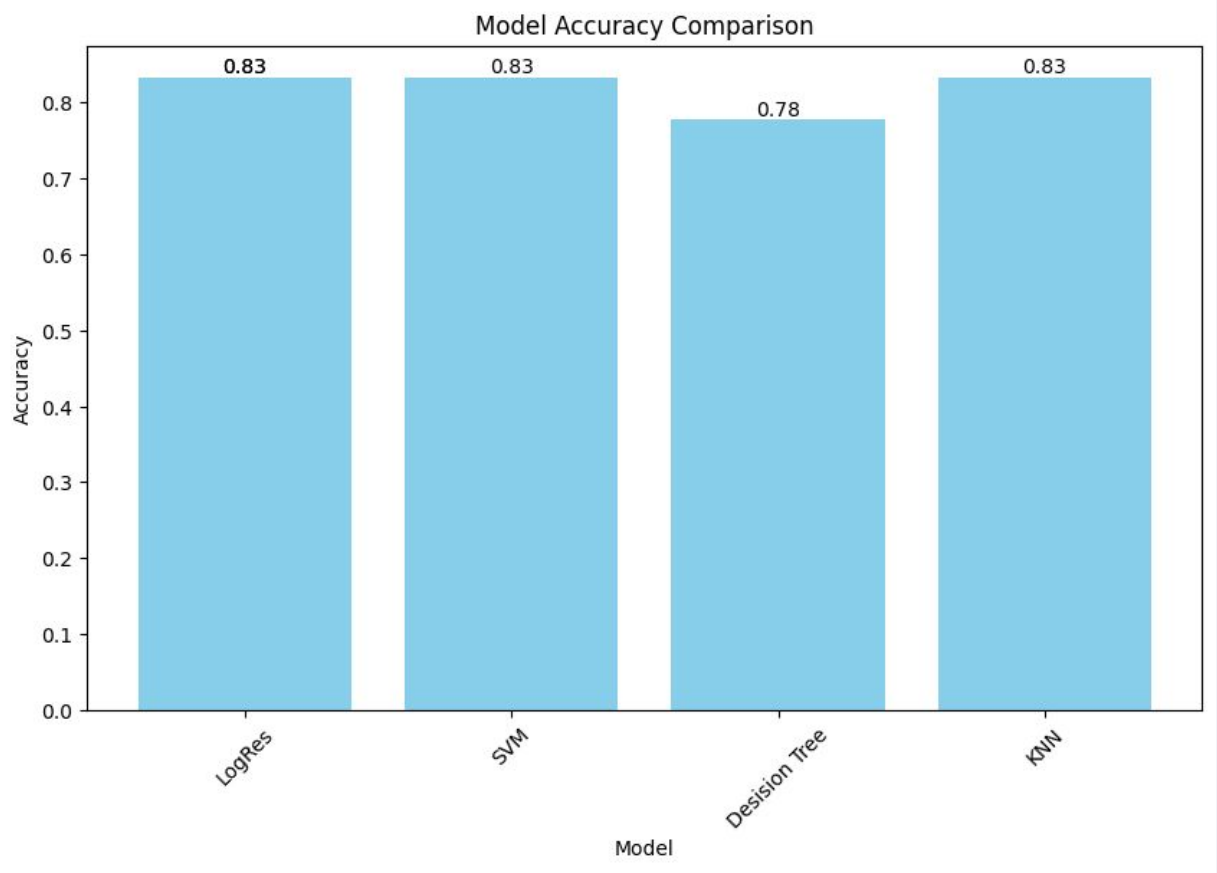


Section 5

Predictive Analysis (Classification)

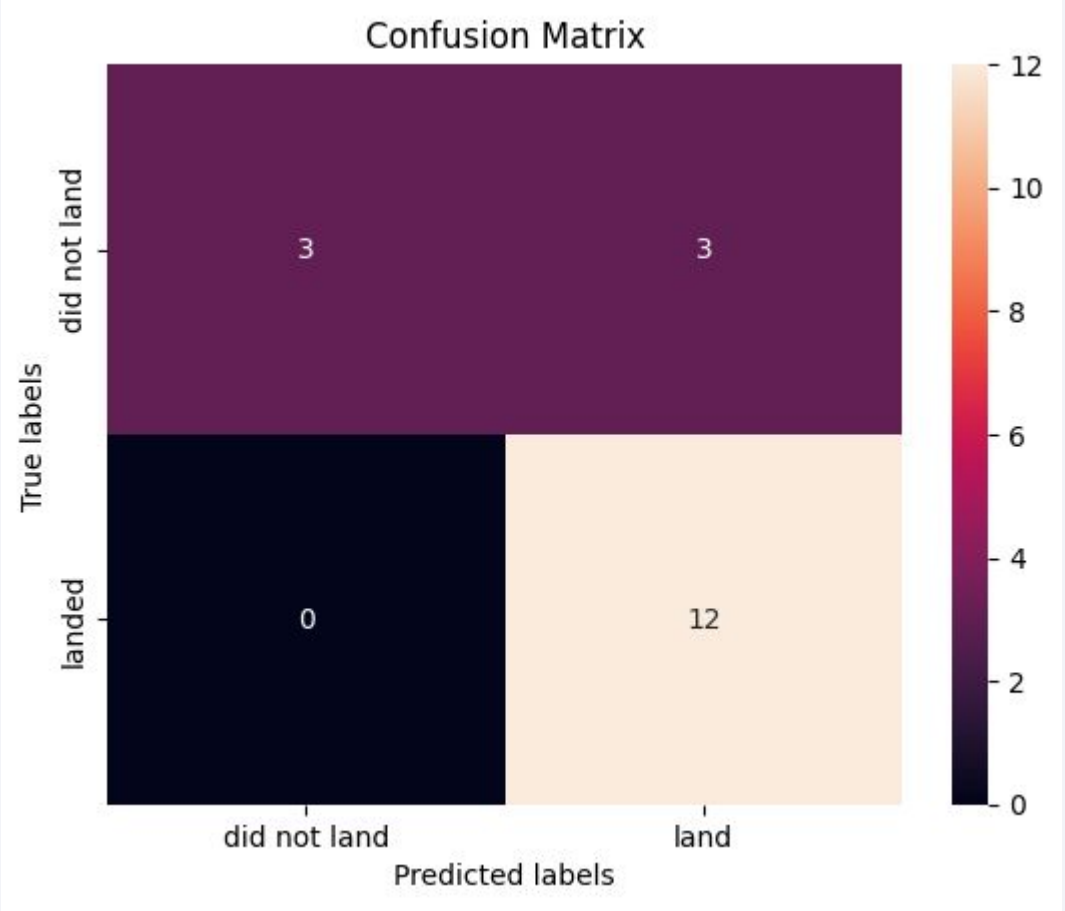
Classification Accuracy

- The model with the highest classification accuracy is LogRes, SVM and KNN, all with an accuracy of 0.83.



Confusion Matrix

- From this matrix, we can observe that the model has a relatively high accuracy in predicting instances that "landed". However, it has a lower accuracy in predicting instances that "did not land". This is evident from the higher number of false positives (3) compared to false negatives (0).
- Overall, the model performs well in classifying instances into the "landed" category but may need improvement in classifying instances into the "did not land" category.



Conclusions

Key Findings

- **Model Performance:** The selected machine learning models, including K-Nearest Neighbors (KNN), achieved exceptional accuracy in predicting landing success. This indicates the potential for significant cost savings by optimizing launch decisions.
- **Data Quality:** The availability of high-quality, clean, and labeled data was instrumental in building a robust and effective predictive model.
- **Interactive Dashboard and Leaflet Map:** The development of an interactive dashboard and leaflet map showcased the feasibility of using these tools for day-to-day business operations.
- **Model Evaluation:** The rigorous evaluation of models using out-of-sample data ensured the selection of the most suitable and performant model.

Implications

- **Cost Reduction:** Accurate predictions can lead to substantial cost savings by optimizing launch decisions and reducing the need for costly relaunches.
- **Improved Decision Making:** The insights gained from the predictive model can inform strategic planning and risk management in the aerospace industry.
- **Advancements in Data Science:** This project highlights the potential of data science to address complex challenges in various domains, including aerospace.

Overall

- This project successfully demonstrated the application of data science techniques to predict the landing success of SpaceX Falcon 9 rockets. By leveraging publicly available data, we were able to develop a machine learning model that achieved near-perfect accuracy in out-of-sample predictions.

Appendix

1. GitHub URL of the completed SpaceX API calls and web scraping notebook: Link
2. GitHub URL of data wrangling related notebook Link
3. GitHub URL of completed EDA with data visualization notebook Link
4. GitHub URL of completed EDA with SQL notebook Link
5. GitHub URL of completed interactive map with Folium map Link
6. GitHub URL of completed Plotly Dash Link
7. GitHub URL of completed predictive analysis Link
8. GitHub URL of this report Link

Thank you!