IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fernando Gutiérrez Nieto
December 29,2024

Outline

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

Executive Summary

SpaceX has emerged as a leader in the aerospace industry, revolutionizing the economics of space exploration through innovative technologies. While traditional rocket launches cost approximately $165 million, SpaceX achieves comparable objectives at a significantly lower cost of $60 million, thanks to its reusable rocket systems.

This report evaluates the feasibility of competing with SpaceX by analyzing historical launch data. By leveraging SpaceX's REST API and supplementary web-scraped datasets, we performed data preprocessing, exploratory data analysis (EDA), and predictive modeling to uncover critical factors driving SpaceX's success.

To enhance our findings, interactive visual analytics were created using Folium and Plotly Dash, providing detailed insights into launch site performance, payload characteristics, and success trends. Furthermore, machine learning models, such as Decision Trees and Logistic Regression, were utilized to predict the success of first-stage landings.

The insights derived from this analysis offer actionable strategies for new market entrants, like Space Y, to optimize their operations and improve competitiveness in the space industry.

Introduction

- Since the launch of Sputnik in 1957, the space race has driven nations and private entities to innovate in aerospace technologies. However, the high costs associated with space missions have historically posed significant barriers, with traditional rocket launches averaging $165 million per mission. SpaceX has revolutionized this landscape through groundbreaking advancements, such as reusable first-stage rockets, reducing costs to approximately $60 million. This innovative approach has not only made space exploration more attainable but also set a new standard for efficiency within the industry.

- This project examines the critical factors contributing to successful first-stage rocket landings. By applying data science methodologies, we analyze variables such as payload mass, launch site, orbit type, and launch trends. The insights derived aim to inform strategies for competing in the rapidly expanding private aerospace sector.

Section 1

Methodology

Methodology

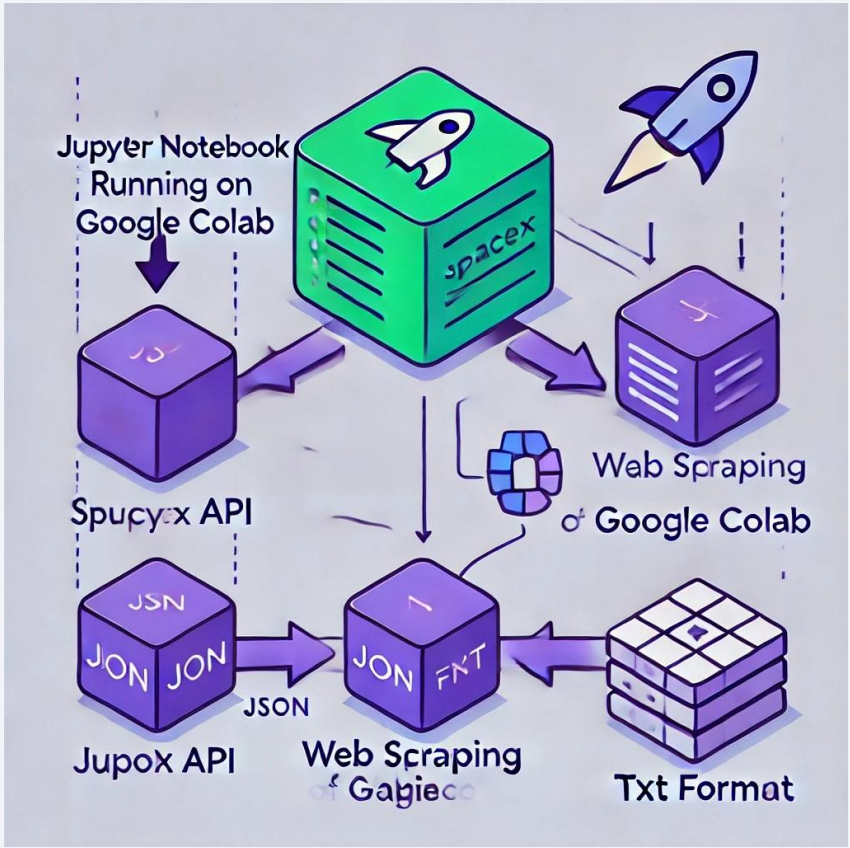
Executive Summary

Data Collection Methodology: Data was collected using the SpaceX REST API and enriched with additional information obtained through web scraping techniques from relevant Wikipedia pages.
Data Wrangling: The dataset was preprocessed using Pandas and NumPy. Key steps included one-hot encoding, removal of irrelevant columns, and normalization/standardization of data to ensure consistency and readiness for analysis.
Exploratory Data Analysis (EDA): Comprehensive visualization and statistical analysis were performed using libraries such as Seaborn and Matplotlib. Additionally, SQL queries were utilized to conduct more granular investigations of the dataset.
Interactive Visual Analytics: Dynamic and interactive visualizations were developed using Folium and Plotly Dash. These tools facilitated deeper insights and provided an intuitive platform for exploring the data interactively.
Predictive Analysis with Classification Models: The dataset was split into training and testing subsets. Classification models were optimized using Grid Search to identify the best-performing algorithms and hyperparameters. The final model was deployed to generate effective predictions.

Data Collection

SpaceX API: A publicly accessible REST API offering detailed information on launches, rockets, cores, capsules, Starlink satellites, launchpads, and landing pads.

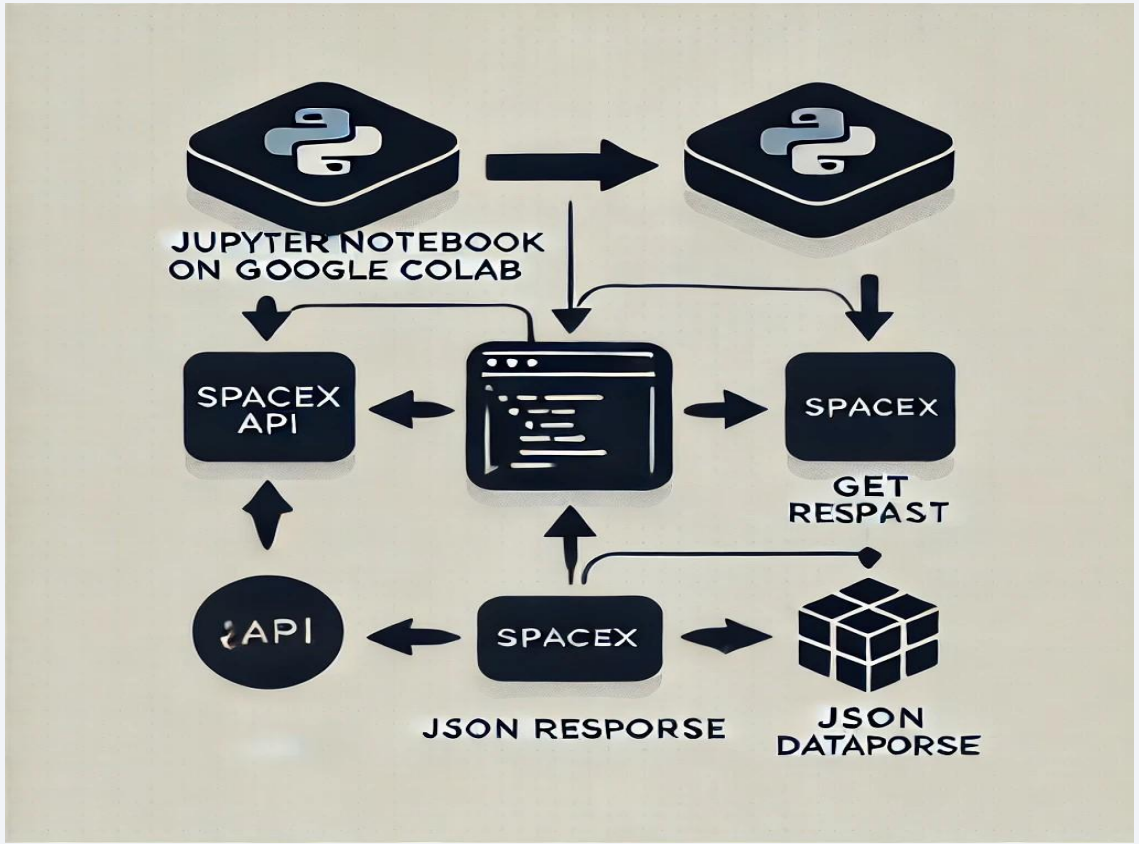
Wikipedia: A free, collaborative online encyclopedia maintained by volunteers and hosted by the Wikimedia Foundation, serving as a valuable source of comprehensive information.



Data Collection – SpaceX API

Process Overview

- **Initiation**: Used Jupyter Notebook on IBM Watson to import key libraries like Pandas, NumPy, and Requests.
- **API Request**: Sent a GET request to the SpaceX API and extracted relevant data such as geospatial details, rocket types, orbit information, and flight numbers.
- **Data Conversion**: Converted the JSON response into a Pandas DataFrame for analysis and visualization.



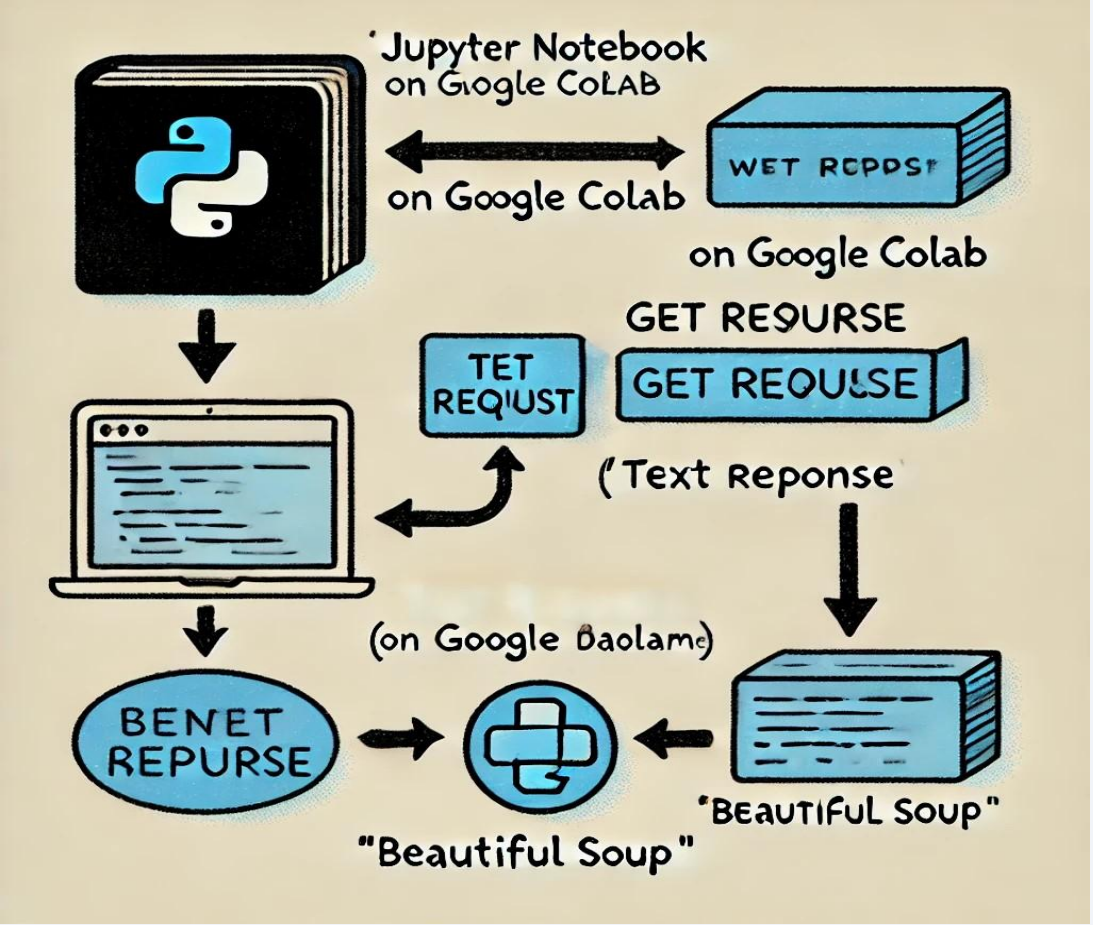
Link GitHub Repository:

https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/001_jupyter_labs_spacex_data_collection_api_page8.ipynb

Data Collection - Scraping

Process Overview

- To complete this task, we used Python libraries like **BeautifulSoup** and **Requests** to scrape data from the Wikipedia page *"SpaceX Falcon 9 First Stage Landing Prediction."*
- An HTTP GET request retrieved the webpage content in text format.
- Using **BeautifulSoup**, we extracted the relevant tables and columns.
- The extracted data was structured and stored in a **Pandas DataFrame** for analysis.

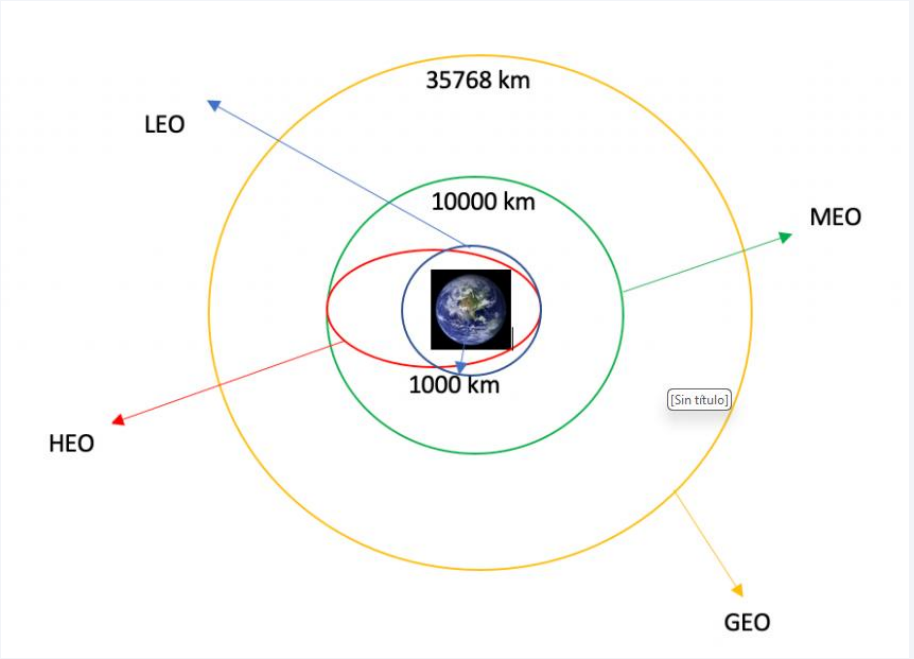


Link GitHub Repository:

https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/002_jupyter_labs_webscraping_page9.ipynb

Data Wrangling

- Loading the collected dataset
- Identifying and calculating the percentage of missing values in each attribute
- Identifying which columns are numerical and categorical
- Calculating the number of launches on each site
- Calculating the number and occurrence of each orbit
- Creating a landing outcome label from the Outcome column
- Determining the success rate of returning the first stage of the rocket



Link GitHub Repository:

https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/003_labs_jupyter_spacex_Data_wrangling_page10.ipynb

EDA with Data Visualization

At this stage, we finalized the Exploratory Data Analysis (EDA) by identifying correlations between features and the target variable. Visualization tools like Seaborn and Matplotlib were employed to extract meaningful insights from the data. Furthermore, we performed feature engineering by converting categorical variables into dummy variables, ensuring the dataset was properly prepared for machine learning models.

- Visualization of the relationship between Flight Number and Launch Site.
- Visualization of the relationship between Payload and Launch Site.
- Visualization of the relationship between Success Rate for each Orbit Type.
- Visualization of the relationship between Flight Number and Orbit Type.
- Visualization of the relationship between Payload and Orbit Type.
- Visualization of the Launch Success yearly trend.

Link GitHub Repository:
https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/005_edadataviz_page11.ipynb

EDA with SQL

- Display the unique launch site names from the space mission.
- Show 5 records where the launch site names start with "CCA".
- Calculate the total payload mass carried by boosters launched by NASA (CRS).Identify the date of the first successful ground pad landing.
- List the booster names that succeeded on drone ships and carried payloads between 4000 and 6000.Count the total number of successful and failed mission outcomes.
- Find the booster versions that carried the maximum payload mass using a subquery.
- List failed drone ship landings, including booster versions and launch site names, for the year 2015.Rank the count of landing outcomes (e.g., "Failure (drone ship)" or "Success (ground pad)") between June 4, 2010, and March 20, 2017, in descending order.

Link GitHub Repository:

https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/004_jupyter-labs-eda-sql-coursera_sqllite_page12.ipynb

Build an Interactive Map with Folium

Process OverviewWe utilized the Folium library to represent our analysis as geospatial data. This was achieved by adding markers, such as circles and lines, to an interactive map.

Launch Site MarkersWe began by placing circular markers on the map to highlight the four Falcon 9 rocket launch sites. The coordinates for these sites are as follows:

Launch Site	Latitude	Longitude
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610746

Success/Failure Markers
Markers were placed at these locations to indicate whether the rocket's first stage successfully returned or failed.
Distance Calculation
We computed the distances from CCAFS LC-40 to three important points:
•The nearest city.
•The coastline.
•A nearby highway.
Using this data, we added polylines to the map to visually display these distances.

Link GitHub Repository:
https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/006_lab_jupyter_launch_site_location_page13.ipynb

Build a Dashboard with Plotly Dash

Process Overview

- We used the Plotly library to create an interactive dashboard for our analysis. Here are the main tasks performed:
- Added a dropdown menu to select a launch site, with options such as "All Sites," "CCAFS LC-40," "CCAFS SLC-40," "VAFB SLC-4E," and "KSC LC-39A."Included a pie chart to display the total count of successful launches for all sites.
- Added a slider to filter payload values within the range of 0 to 10,000.
- Created a scatter chart to illustrate the correlation between payload and launch success.

Link GitHub Repository:
https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/007_spacex_dash_app.py

Predictive Analysis (Classification)

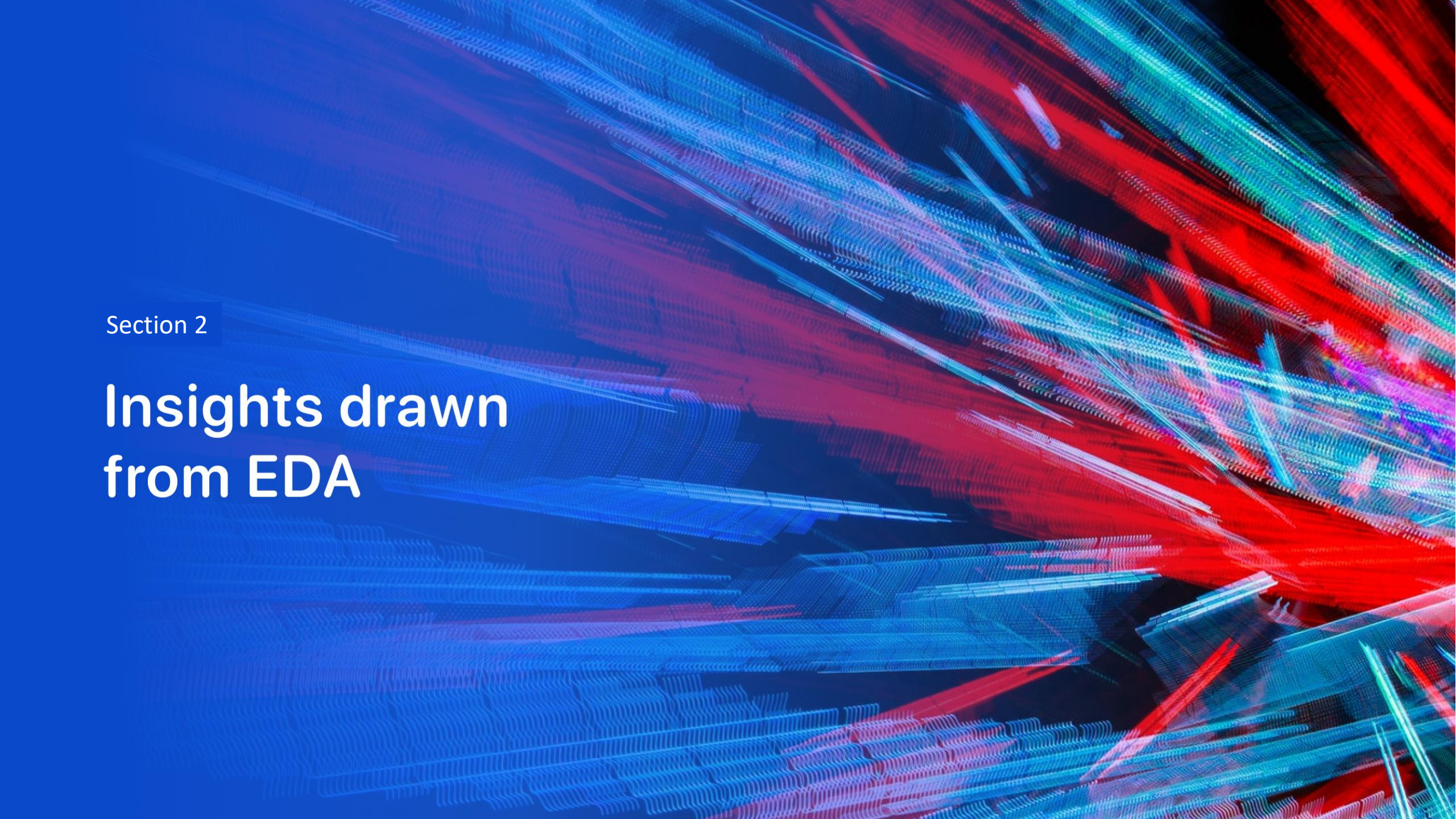
- **Machine Learning Stages**
- **Import Libraries:** Start by importing the necessary libraries for data manipulation and model development.
- **Load Pre-Processed Data:** Use the cleaned dataset to ensure it is ready for analysis.
- **Standardize Data:** Normalize the dataset to reduce bias during model training.
- **Split the Dataset:** Separate the data into training (80%) and testing (20%) sets to evaluate model performance.
- **Initialize Models:** Configure four classification algorithms for comparison: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN).
- **Optimize Hyperparameters:** Use Grid Search to test various parameter combinations and identify the best-performing model.
- **Evaluate Performance:** Measure model performance using metrics such as the Confusion Matrix, F1 Score, and Jaccard Score.

Link GitHub Repository:

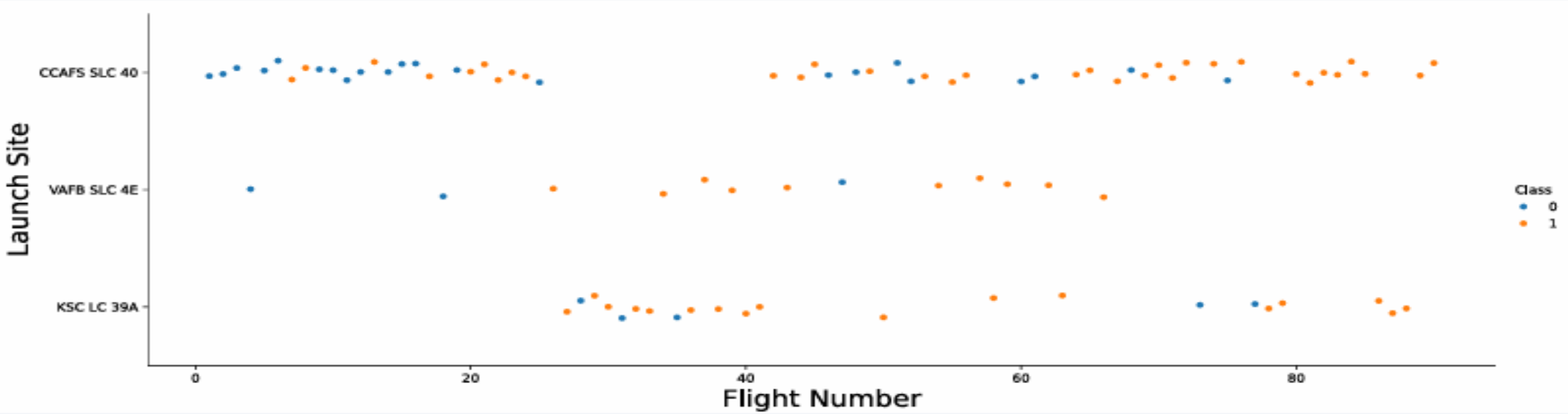
https://github.com/fernandogn72/Repfgn/blob/main/IBM_Curso10_Capastone/008_SpaceX_Machine_Learning_Prediction_Part_5_page15.ipynb

Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

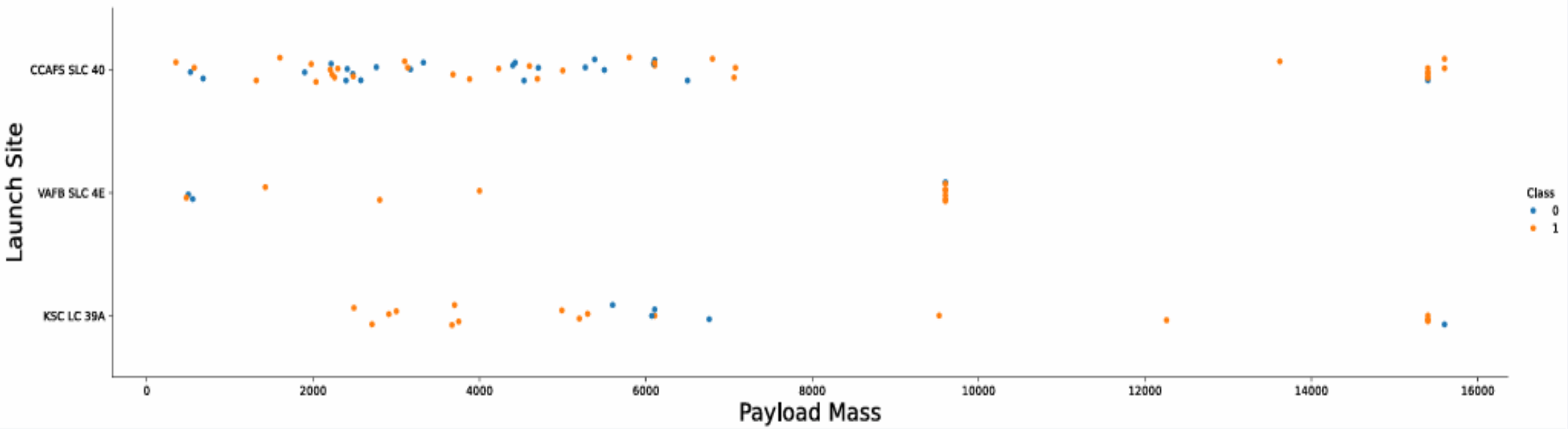



```
[ ]  # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
     sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 3)
     plt.xlabel("Flight Number",fontsize=20)
     plt.ylabel("Launch Site",fontsize=20)
     plt.show()
```

Comments

- CAFS SLC-40: The most frequently used SpaceX launch site, with 55 trials—33 successes and 22 failures, resulting in a 60% success rate.
- VAFB SLC-4E: The least used site, with 13 trials—10 successes and 3 failures, achieving a 77% success rate.
- KSC LC-39A: A moderately used site, with 22 trials—17 successes and 5 failures, also achieving a 77% success rate.

Payload vs. Launch Site

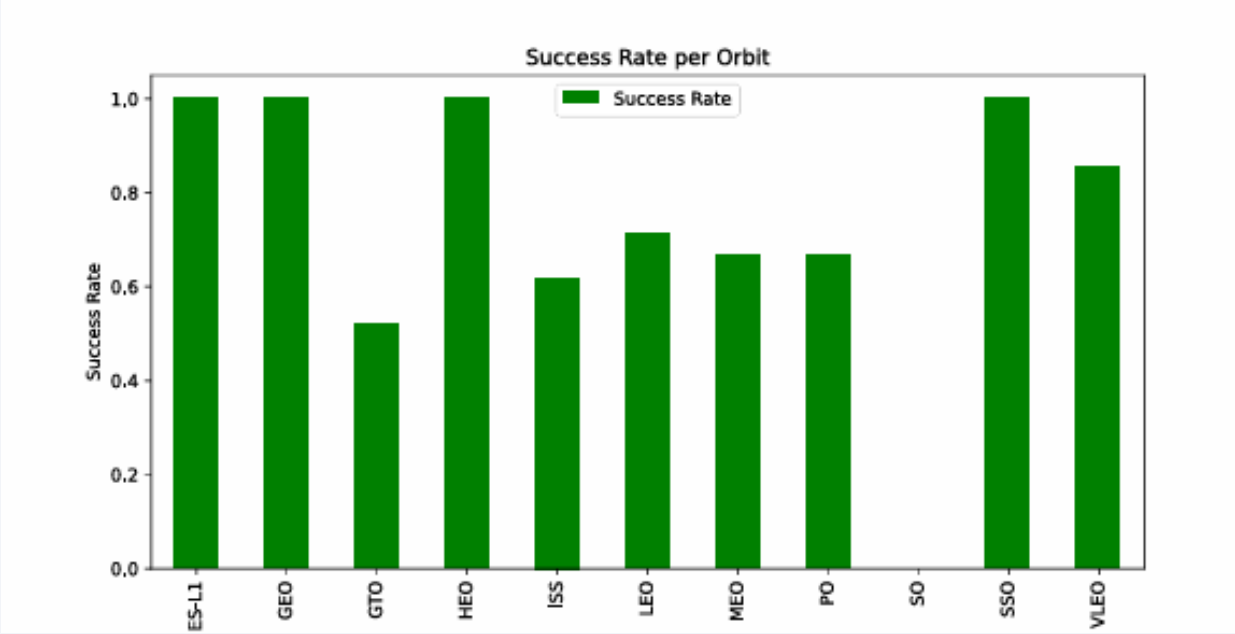



```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 4)
plt.xlabel("Payload Mass",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

Comments

Based on the plot above, there appears to be no strong correlation between payload mass and the success of the first-stage return, as the number of successful and failed trials is approximately equal.

Success Rate vs. Orbit Type

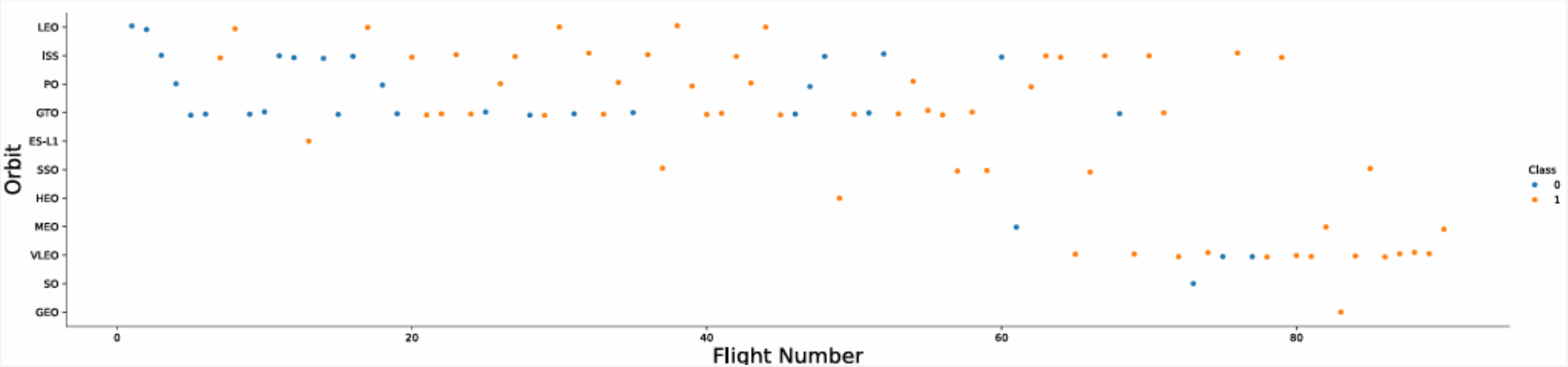



```
# HINT use groupby method on Orbit column and get the mean of Class column
df_bar = df[['Orbit','Class']].groupby('Orbit').mean()
df_bar.columns = ['Success Rate']
df_bar.plot(kind='bar', color='g', figsize=(10, 5))
plt.xlabel('Orbit')
plt.ylabel('Success Rate')
plt.title('Success Rate per Orbit')
plt.show()
```

Comments

The bar plot shows that the most successful orbits for first-stage returns are ES L1, GEO, HEO, and SSO. In contrast, GTO is the worst-performing orbit. Understanding the factors contributing to GTO's poor performance is crucial for reducing failures in first-stage returns.

Flight Number vs. Orbit Type

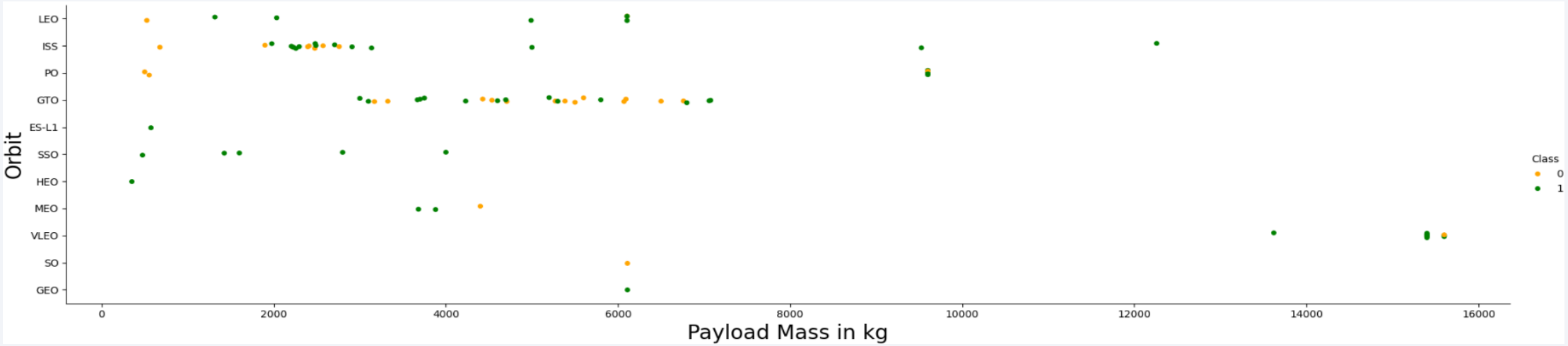



```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 4)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

Comments

In the LEO orbit, success seems to be linked to the number of flights. However, in the GTO orbit, no clear relationship is observed between the number of flights and success.

Payload vs. Orbit Type

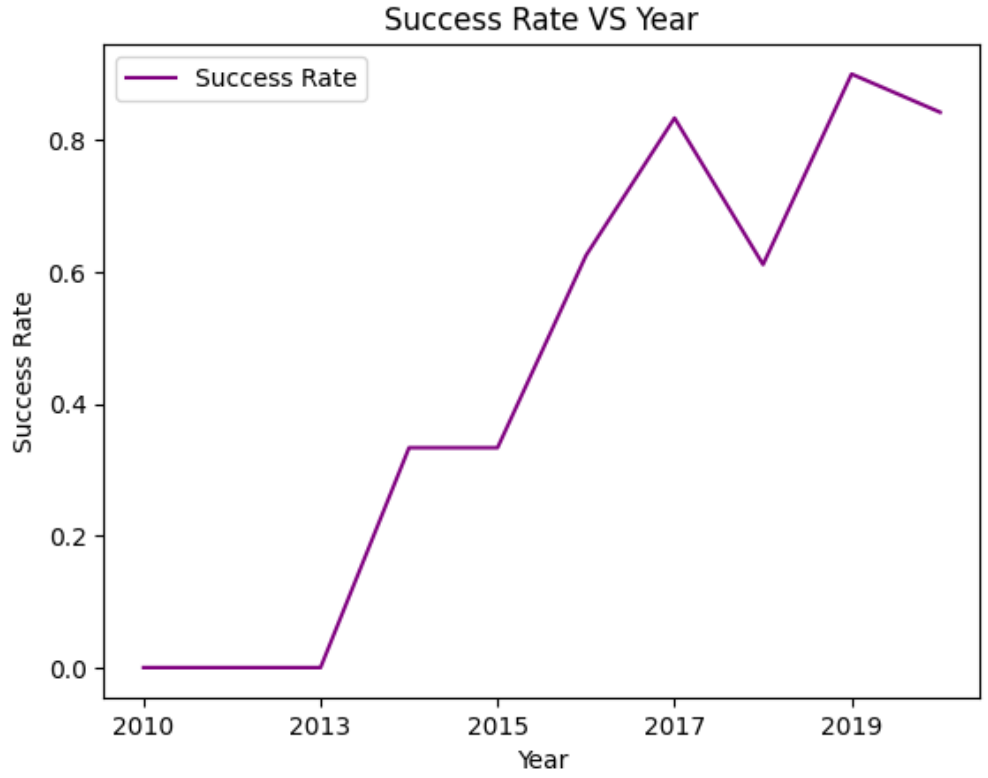



```
# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 4, palette=sns.color_palette(['orange', 'green']))
plt.xlabel("Payload Mass in kg",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

Comments

Heavy payloads negatively affect GTO orbits but have a positive impact on Polar, LEO, and ISS orbits, where the success rate for landings is higher with heavier payloads. However, in GTO orbits, distinguishing between successful and unsuccessful landings is challenging, as both outcomes are observed.

Launch Success Yearly Trend




```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df_line.plot(kind = 'line', color='purple')
plt.title('Success Rate VS Year')
plt.xlabel('Year')
plt.ylabel('Success Rate')
```

Comments

The success rate shows a steady increase from 2013 to 2020.

All Launch Site Names

We use the following query to retrieve the names of the unique launch sites.

```
[ ]  %sql select distinct launch_site from SPACEXTBL
```

There are 4 sites for rockets launches:

```
 * sqlite:///my_data1.db
Done.
  Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

We used the following query to retrieve 5 records where the launch site names start with:

```
[ ]  %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

Output:

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

We used the following query to calculate the total payload carried by NASA boosters.

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

Following output:

```
 * sqlite:///my_data1.db
Done.
 sum(payload_mass__kg_)
 45596
```

NASA has transported a total payload of 45,596 kg to outer space using SpaceX rockets, equivalent to approximately 50.261 US tons.

Average Payload Mass by F9 v1.1

We used the following query to calculate the average payload mass carried by the F9 v1.1 booster version.

```
%sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1'
```

Following output:

```
 * sqlite:///my_data1.db
Done.
avg_mass_F9
2928.4
```

The Falcon 9 booster version v1.1 has an average payload mass of 2,928.4 kg.

First Successful Ground Landing Date

We used the following query to determine the dates of the first successful landing on a ground pad.

```
[ ]  %sql select min(DATE) from SPACEXTBL where landing_outcome = 'Success (ground pad)'
```

Following output:

```
 * sqlite:///my_data1.db
Done.
 min(DATE)
2015-12-22
```

The Falcon 9 achieved its first successful ground pad landing on December 22, 2015, during flight 20. This milestone at Cape Canaveral marked a breakthrough in SpaceX's reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the following query to list boosters that successfully landed on a drone ship with payloads between 4000 and 6000 kg.

```
[ ]  %sql select booster_version from SPACEXTBL where (landing_outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and  payload_mass__kg_ < 6000));
```

Following output:

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

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

This indicates that only four boosters successfully landed on a drone ship with payloads between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

We used the following query to determine the total number of successful and failed mission outcomes.

```
%sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
```

Following output:

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

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

The mission outcomes clearly showcase an exceptional success rate, with 99 successful missions and only 1 failure. This highlights a remarkable level of reliability and achievement.

Boosters Carried Maximum Payload

We used the following query to display the list of boosters that carried the maximum payload mass.

```
%sql select distinct booster_version from SPACEXTBL\
where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL);
```

Following output:

```
 * 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

The boosters with the highest payload capacity belong to the Falcon 9 Block 5 series, specifically models ranging from B1048 to B1060. Designed for improved performance and reusability, these boosters excel in delivering significant payloads to Low Earth Orbit (LEO) and beyond.

2015 Launch Records

We used the following query to list the failed drone ship landing outcomes in 2015, along with their booster versions and launch site names.

```
%sql select landing_outcome, booster_version, launch_site from SPACEXTBL\
where (landing_outcome = 'Failure (drone ship)' and date like '2015%')
```

Following output:

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

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

In 2015, two failed drone ship landings occurred at CCAFS LC-40 using the F9 v1.1 booster, reflecting the challenges SpaceX faced in refining their landing techniques.

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

Finally, we used the following query to rank landing outcomes (e.g., "Failure (drone ship)" or "Success (ground pad)") between June 4, 2010, and March 20, 2017, in descending order.

```
[ ] %sql select landing_outcome, count(*) as counts_of_landing_outcomes from SPACEXTBL where DATE between '2010-06-04' and  '2017-03-20' group by landing_outcome order by count(landing_outcome) desc
```

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

Landing_Outcome	counts_of_landing_outcomes
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

Between June 4, 2010, and March 20, 2017, SpaceX's landing outcomes reflect notable progress in recovery technology, with equal successes and failures on drone ships, three successful ground pad landings, and several missions without landing attempts.

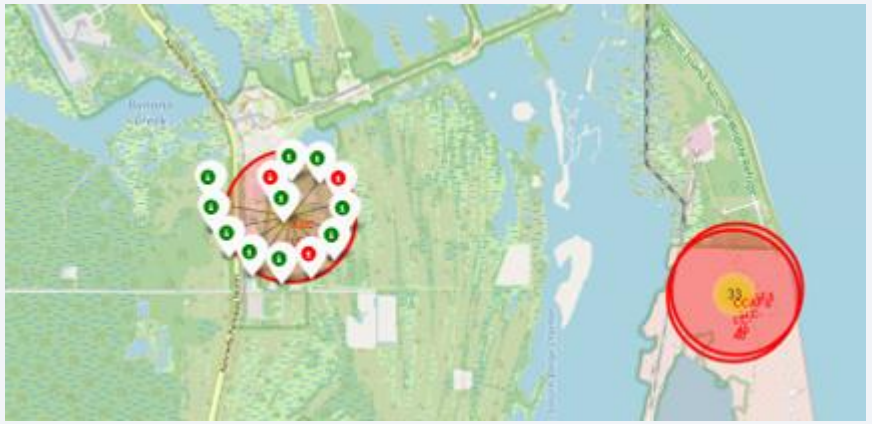
Section 3

Launch Sites Proximities Analysis

Folium Map: Launch sites overview

All SpaceX launch sites are strategically positioned near the coast and close to the equator, leveraging proximity to water and the zero-latitude line to reduce launch risks. These facilities are located across two states: California and Florida.

Folium Map: Success rate for each launch location

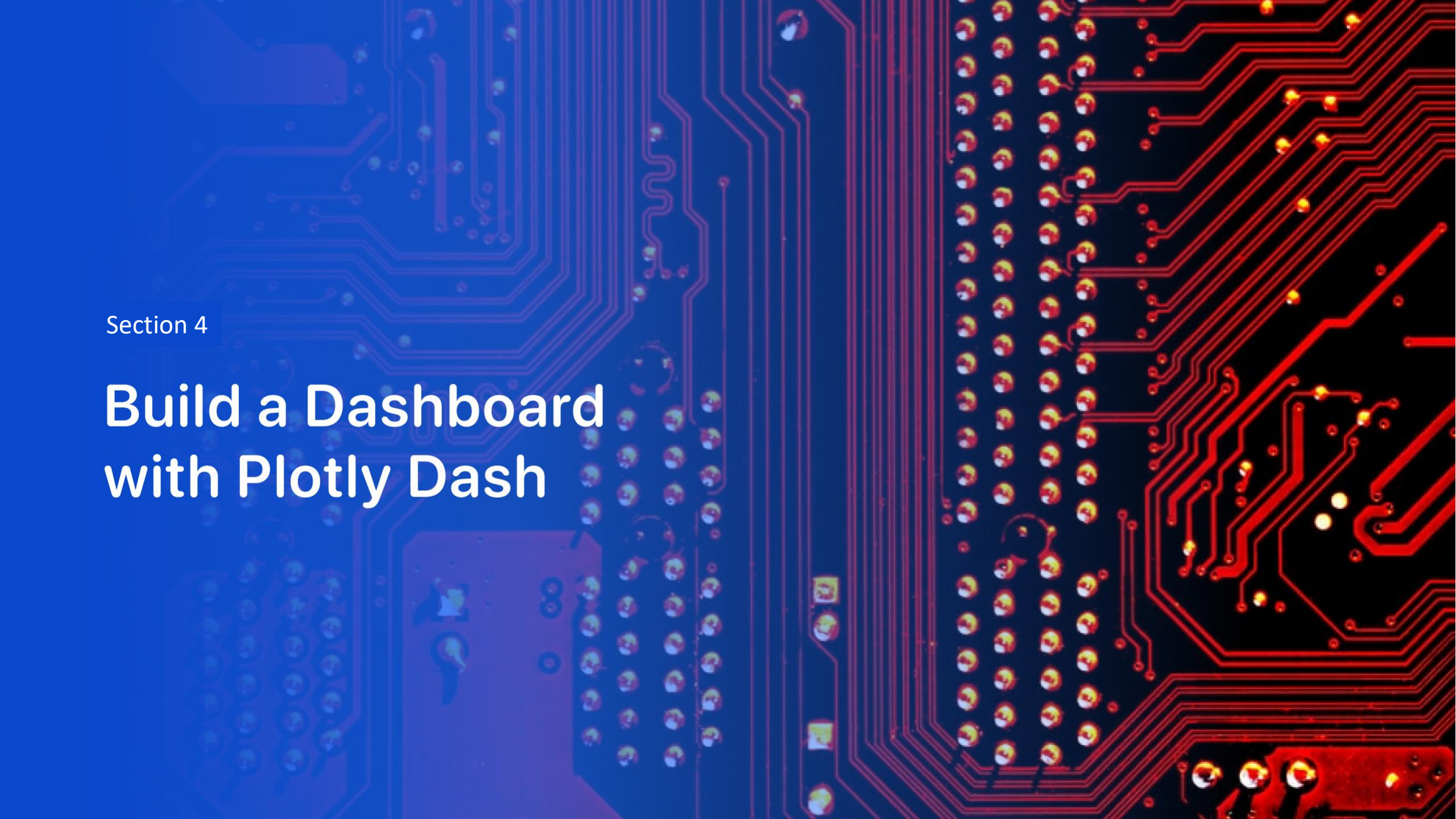


The color-coded markers in the clusters clearly indicate launch site success rates: green for successful returns and red for failed ones.

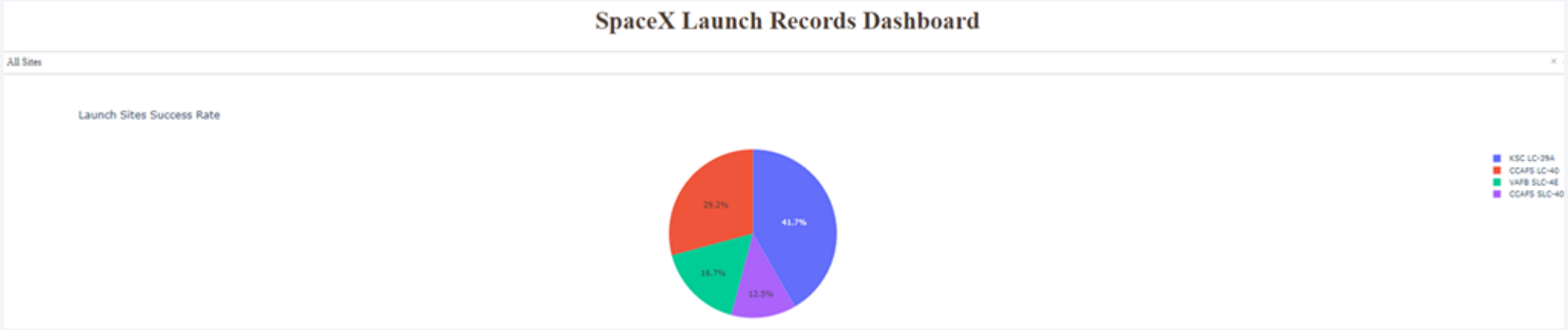
Folium Map: Closest proximities to CCAFS LC-40

	Location	Lat	Long
0	Orlando_Location	28.52300	-81.38260
1	Coastline_Location	28.56146	-80.56746
2	Highway_Location	28.56270	-80.58703

Launch sites are located approximately 78.8 km from Orlando but are less than a kilometer from the coastline and highways.

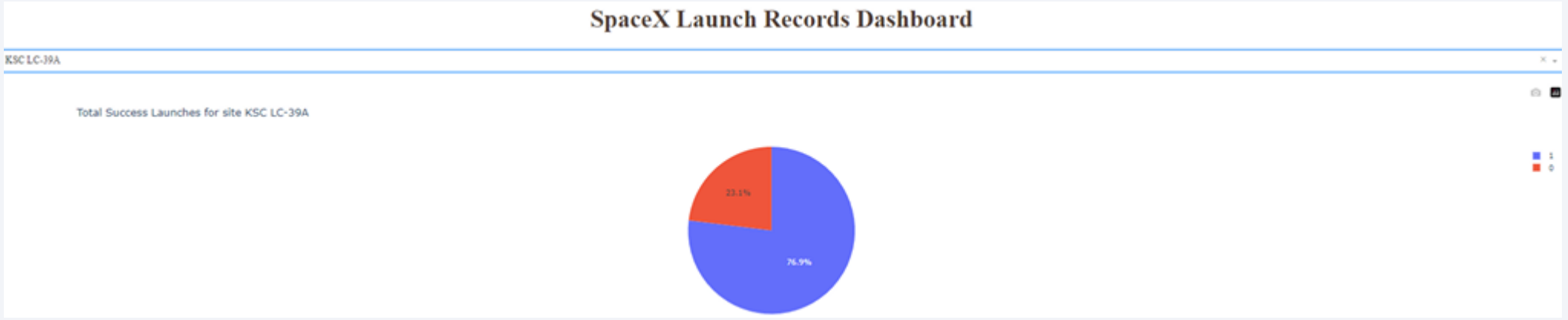

Section 4
Build a Dashboard with Plotly Dash

Dashboard: Launch success count for all sites



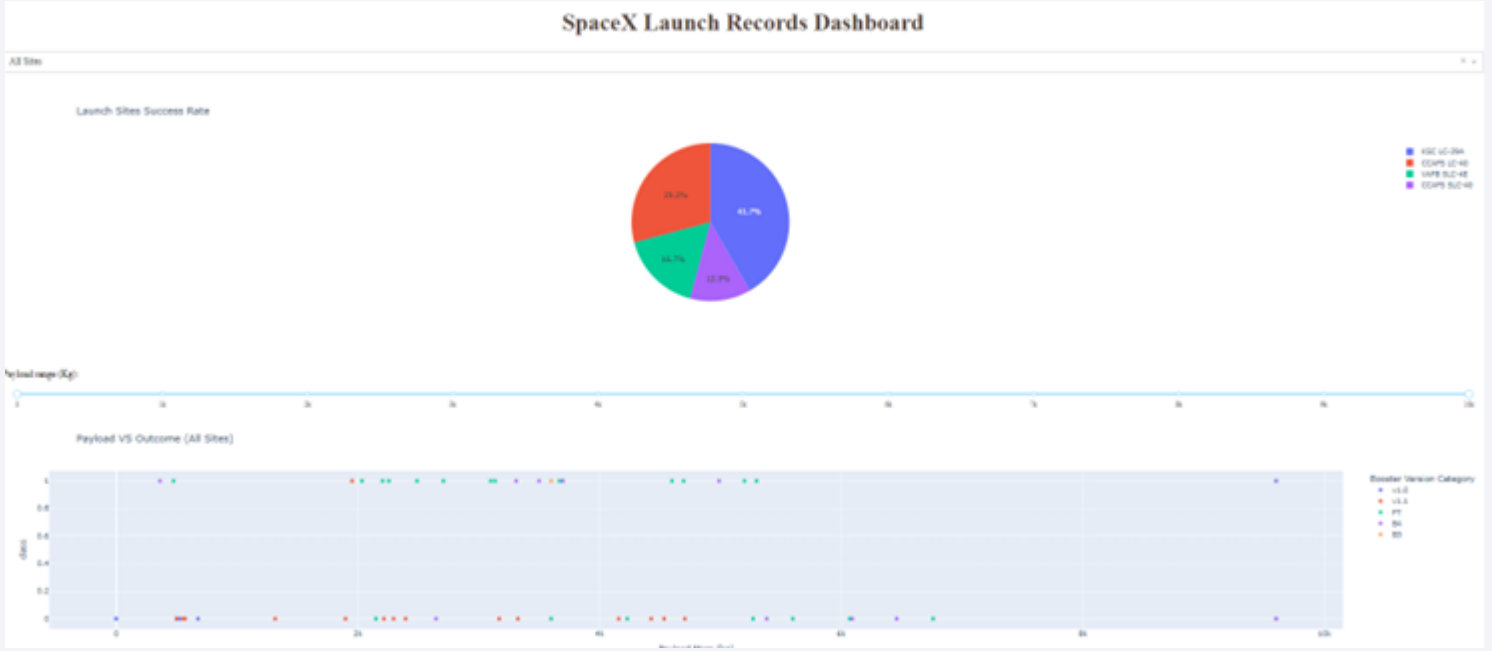
The graph shows the success rates for first-stage returns across launch sites. KSC LC-39A leads with a 41.7% success rate, while CCAFS SLC-40 has the lowest at 12.5%, highlighting differences in recovery performance by location.

Dashboard: Launch success for KSC LC-39A



KSC LC-39A has a success rate of 76.9% and a failure rate of 23.1%, demonstrating its reliability and the effectiveness of SpaceX's operations and technology at this historic launch site.

Dashboard Payload vs. Launch Outcome scatter plot for all sites

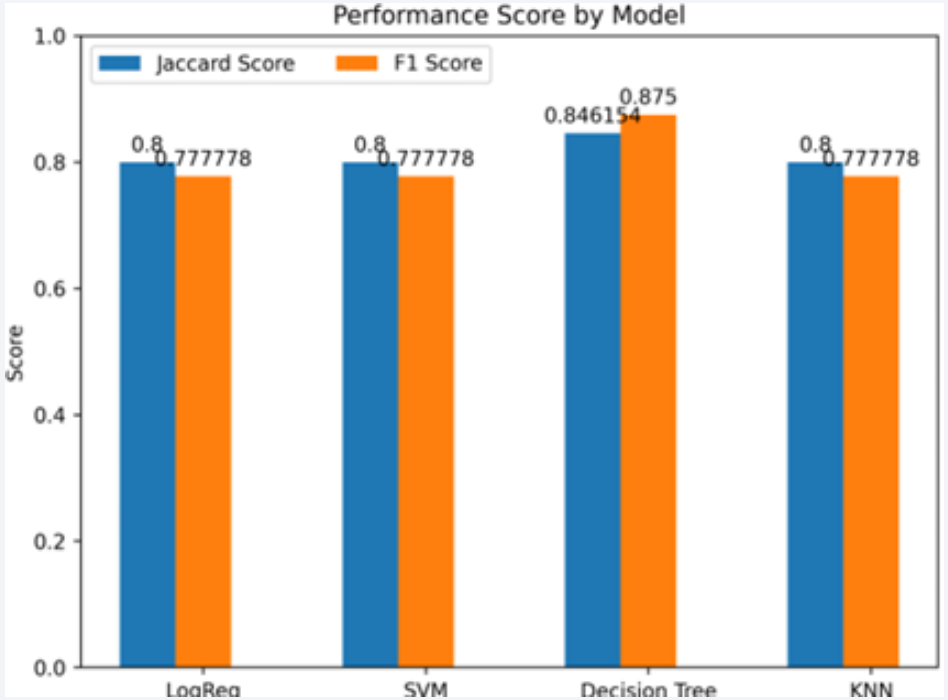


This interactive scatter plot highlights that payloads under 4,000 kg have a higher success rate. It also provides insights into performance trends across booster versions, with adjustable payload mass ranges for detailed analysis.

Section 5

Predictive Analysis (Classification)

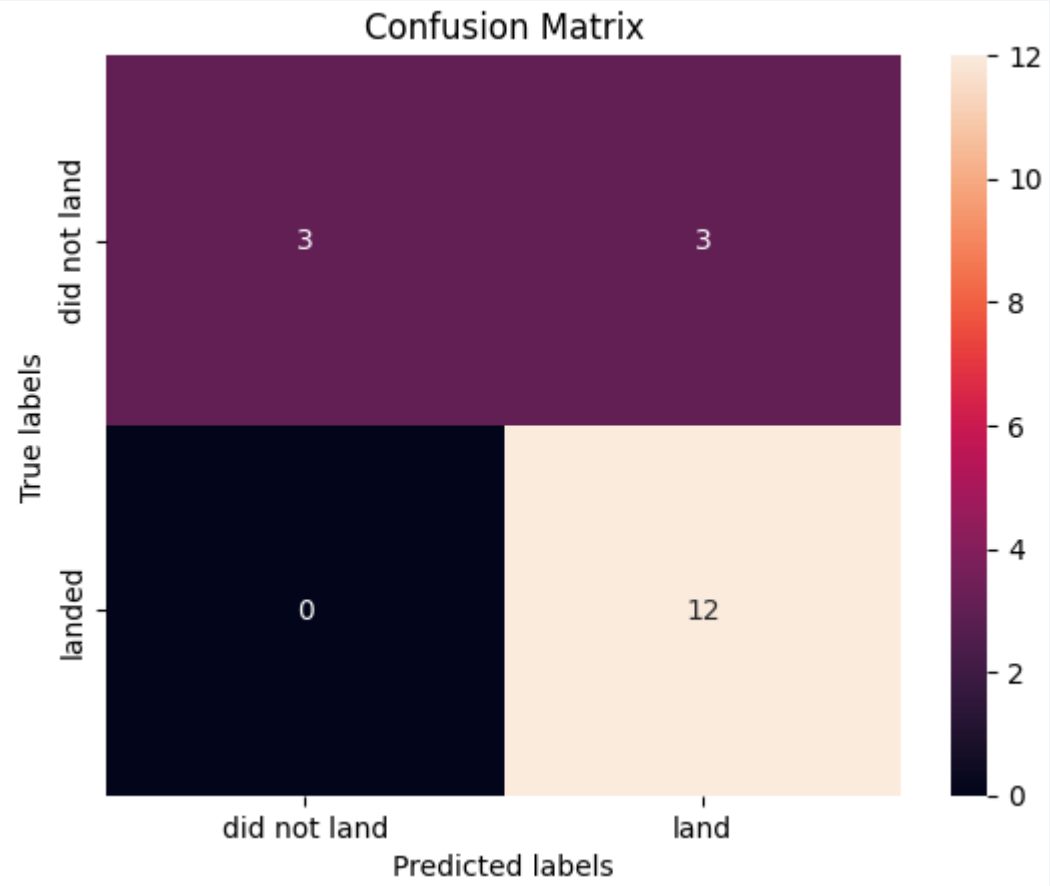
Classification Accuracy



The Decision Tree model demonstrated superior performance, achieving the highest Jaccard Score (0.85) and F1 Score (0.88), reflecting its strong precision and recall in classification tasks. In comparison, Logistic Regression, SVM, and KNN exhibited similar performance, each with a Jaccard Score of 0.8 and F1 Scores of approximately 0.78, indicating comparable effectiveness in their classifications.

Confusion Matrix

The confusion matrix for the Decision Tree classifier demonstrates its ability to differentiate between classes. However, a notable issue is the occurrence of false positives, where unsuccessful landings are incorrectly classified as successful.



Conclusions

SpaceX's focus on reusable first-stage rockets has significantly reduced costs, giving the company a competitive edge. A similar emphasis on reusability could benefit Space Y. Key factors influencing mission success include orbit type and payload mass, with LEO orbits showing higher success rates compared to GTO orbits. Launch sites like KSC LC-39A demonstrate exceptional reliability, emphasizing the importance of strategic location planning. Success rates have steadily improved from 2013 to 2020, with orbits such as ES-L1, GEO, HEO, SSO, and VLEO achieving the highest success rates. Additionally, advanced data analytics and machine learning, particularly Decision Tree classifiers, have proven effective for optimizing success and reducing risks. Continuous evaluation of historical performance and adoption of emerging technologies remain critical for sustaining competitiveness in this dynamic sector.

Appendix

- Github’s webpage of the overall project
- SpaceX Static Wikipedia
- SpaceX data used in ML training

Thank you!