IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Celia Sánchez Gaitán
Jan 2025

Outline

Executive Summary

Summary of methodologies

- Data collection from SpaceX API
- Data collection with Web Scraping
- Data wrangling
- EDA with Data Visualizations
- EDA with SQL
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Summary of all results

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

Introduction

- **Project background and context**
  - This capstone project focuses on predicting whether the first stage of SpaceX's Falcon 9 rocket will land successfully. SpaceX advertises its Falcon 9 launches at $62 million per launch, a significantly lower cost compared to other providers, who charge upwards of $165 million. The main reason behind this cost difference is SpaceX's ability to reuse the first stage of the rocket. By being able to predict if the first stage will successfully land, we can better understand the launch cost and provide insights for other companies looking to compete with SpaceX in the commercial space industry.
- **Problems you want to find answers**
  - What factors influence the success of a rocket's first-stage landing?
  - How do various features interact to impact the success rate of a landing?
  - What operational conditions are required to establish a reliable and successful landing program?

Section 1
Section 1

Methodology

Methodology

Executive Summary

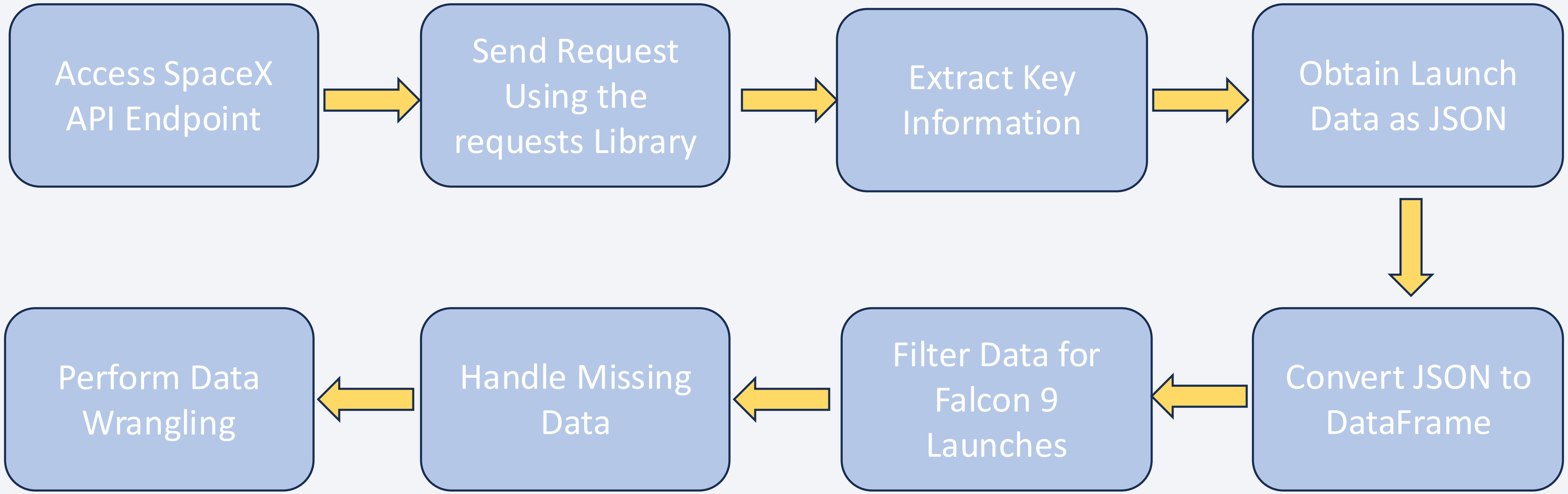
- Data collection methodology:
  - Data was collected using SpaceX API and web scraping form Wikipedia
- Perform data wrangling
  - Data convert form Json and HTML to DF
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use of machine learning to Know if the first stage of Falcon 9 will land successfully.

Data Collection

- **Describe how data sets were collected.**
  - Use the SpaceX REST API endpoint to access available information.
  - Perform a data request using Python's requests library.
  - Retrieve historical launch data in JSON format.
  - Convert the JSON content into a DataFrame for easier analysis.

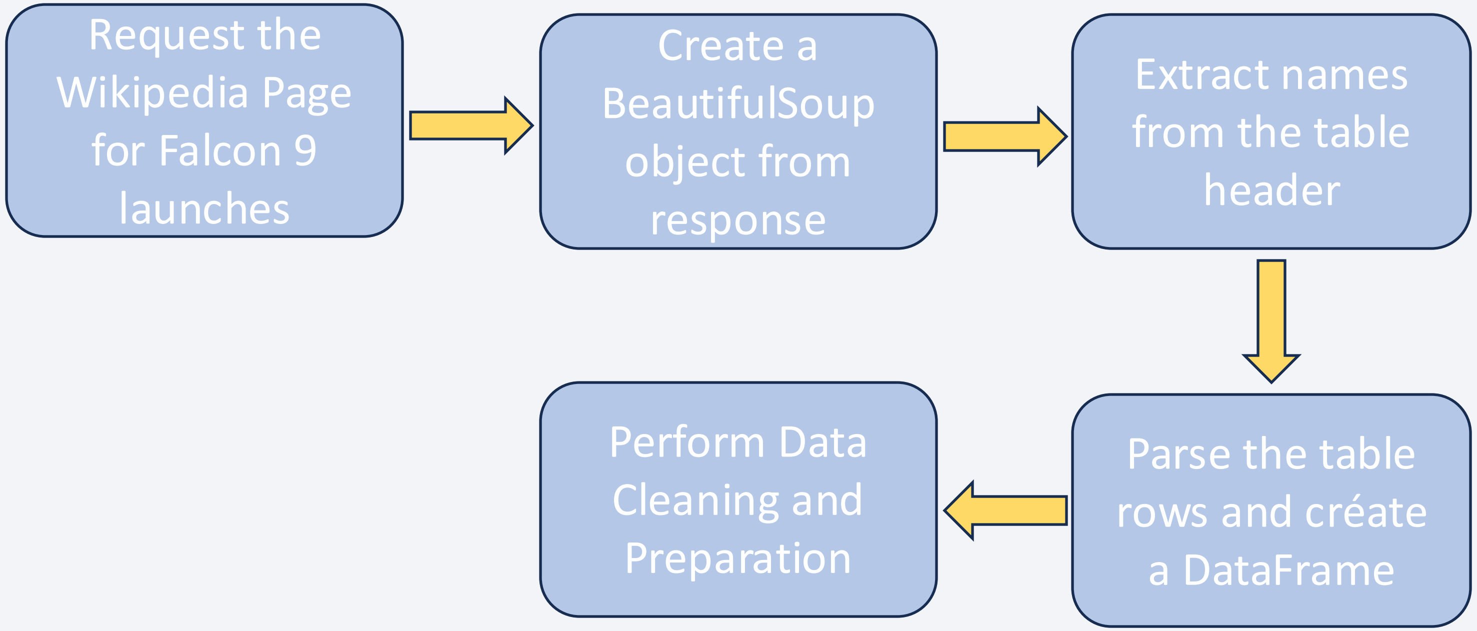
- **You need to present your data collection process use key phrases and flowcharts**
  - Collect Falcon 9 launch records using web scraping techniques.
  - Use the BeautifulSoup library to scrape HTML tables from web pages.
  - Parse and organize data from the extracted tables.
  - Convert the extracted tables into a DataFrame to integrate them with other data sources.

Data Collection – SpaceX API



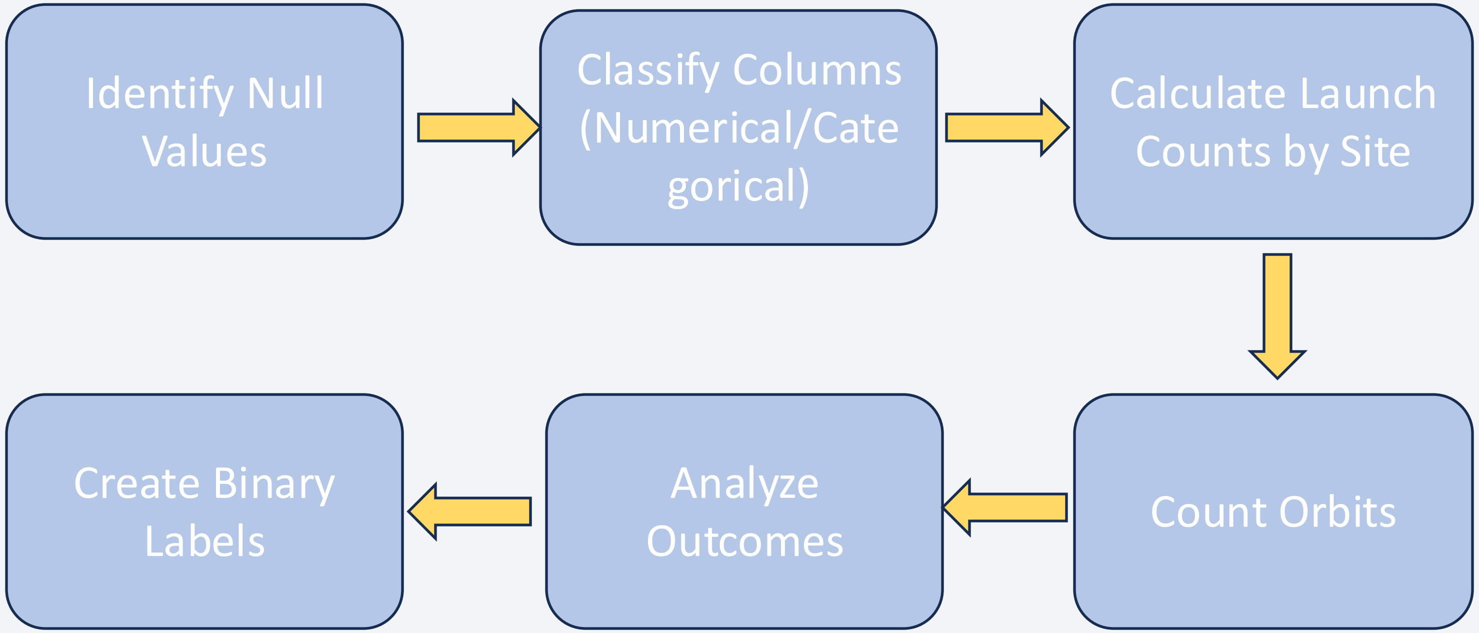
https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/2-jupyter-labs-webscraping.ipynb

Data Wrangling



https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/3-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

We performed Exploratory Data Analysis (EDA) to explore the connections between flight numbers and launch locations, payloads and launch sites, success rates across various orbit types, the relationship between flight numbers and orbit types, and the yearly trends in launch success rates.

https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/4-edadataviz.ipynb

EDA with SQL

- **We utilized SQL for exploratory data analysis to extract meaningful insights. Through various queries, we identified:**
  - The distinct launch sites used for space missions.
  - The total payload mass transported by boosters launched under NASA's missions.
  - The average payload mass handled by the booster version F9 v1.1.
  - The count of successful and unsuccessful mission outcomes.
  - Details about failed landings on drone ships, including their booster versions and launch site names.

https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/5-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **1 - Map Markers**

  Markers were added to the map to pinpoint the locations of the launch sites. Each marker includes a description of the site.

- **2 - Launch Outcomes**

  Launch outcomes were categorized into two classes (0 for failures and 1 for successes), and different colors were used on the markers to represent these results on the map visually.

- **3 - Proximity Circles**

  Circles were drawn around the launch sites to illustrate proximity zones. This helps to visually identify areas of impact or interest.

- **4 - Distance Calculations**

  Distances were calculated from each launch site to key nearby points. These distances were added as pop-up information to the interactive map.

https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/6-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Interactive Dashboard:** Created a dashboard using Plotly Dash to visualize SpaceX data interactively.
- **Pie Charts:** Included pie charts to display the total number of launches at specific launch sites.
- **Scatter Plot:** Added scatter plots to showcase the relationship between payload mass and landing outcomes, categorized by booster versions.
- **Interactivity:** Implemented features for filtering and exploring data dynamically.

Predictive Analysis (Classification)

- **Data Preparation:**
  - Loaded the dataset using NumPy and Pandas libraries.
  - Cleaned and transformed the data to prepare it for machine learning models.
  - Split the dataset into training and testing subsets for evaluation.
- **Model Building:**
  - Experimented with various classification models like Logistic Regression, SVM, Decision Trees, and others.
  - Fine-tuned model hyperparameters using GridSearchCV for optimal performance.
- **Evaluation Metrics:**
  - Evaluated the models using accuracy and other relevant metrics.
  - Conducted feature engineering to improve predictive power and model interpretability.
- **Outcome:**
  - Successfully identified the best-performing model for classifying SpaceX landing outcomes based on key features.

https://github.com/celiapru/Winning-Space-Race-With-Data-Science/blob/main/7-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory data analysis results

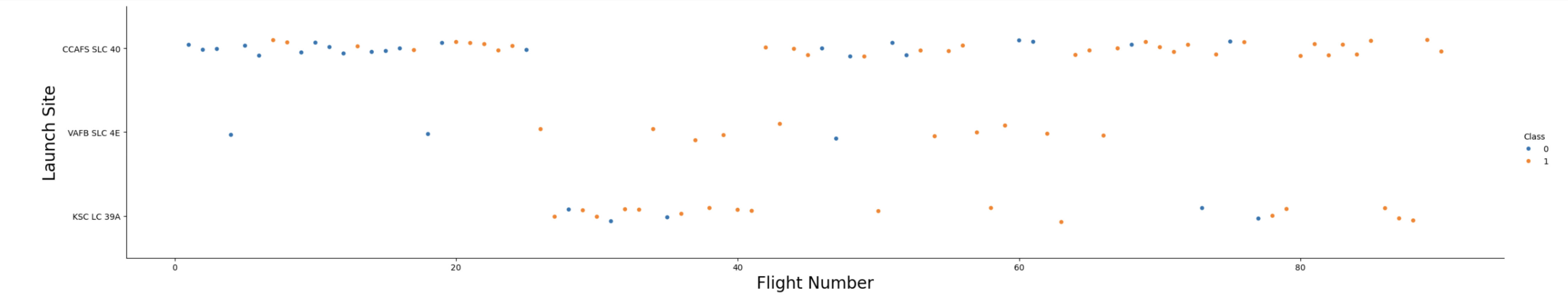
Interactive analytics demo in screenshots

Predictive analysis results

Section 2

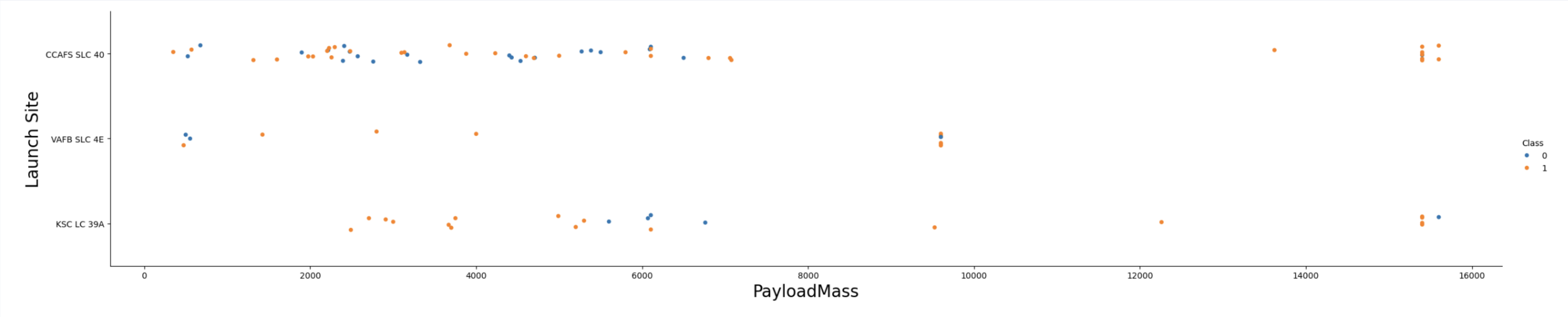
Insights drawn from EDA

Flight Number vs. Launch Site



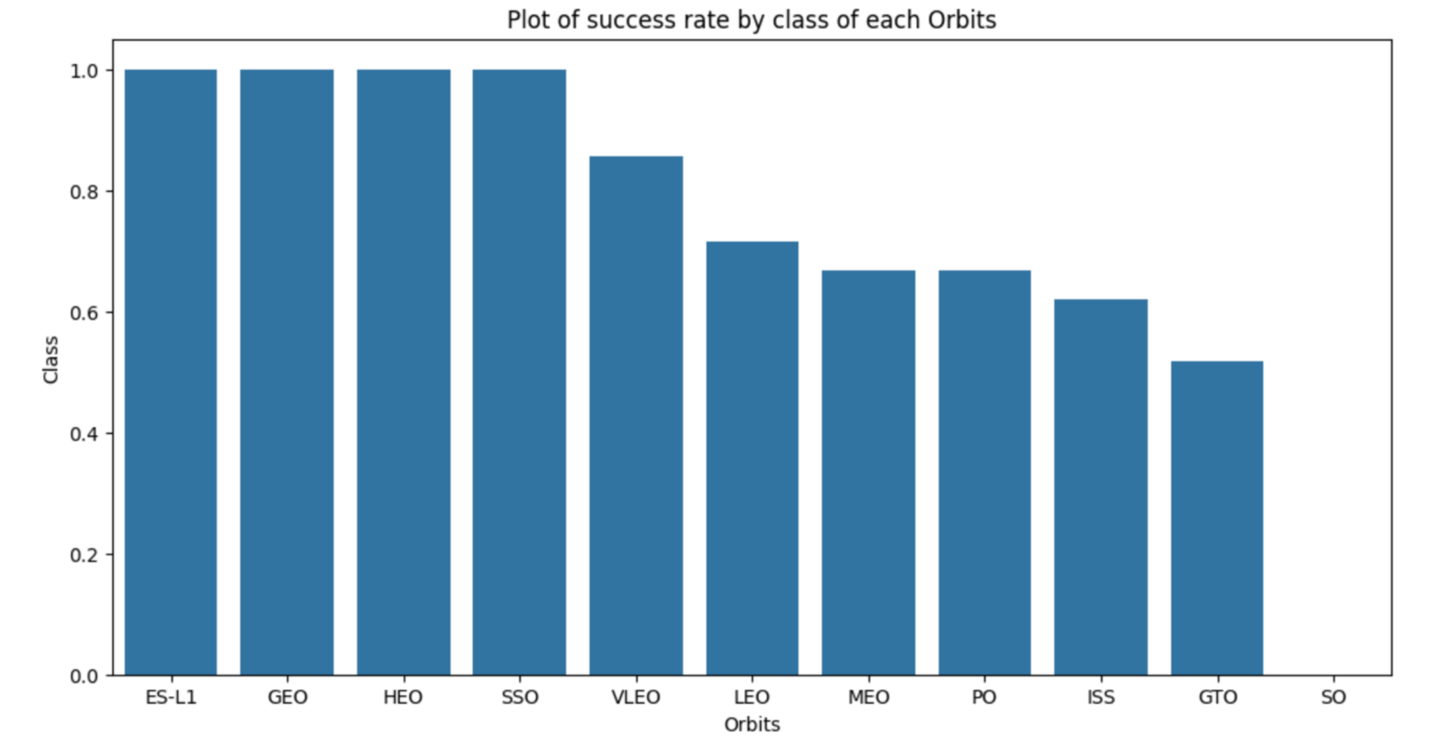
The plot analysis revealed that as the number of flights from a specific launch site increases, the likelihood of success for launches at that site also improves.

Payload vs. Launch Site



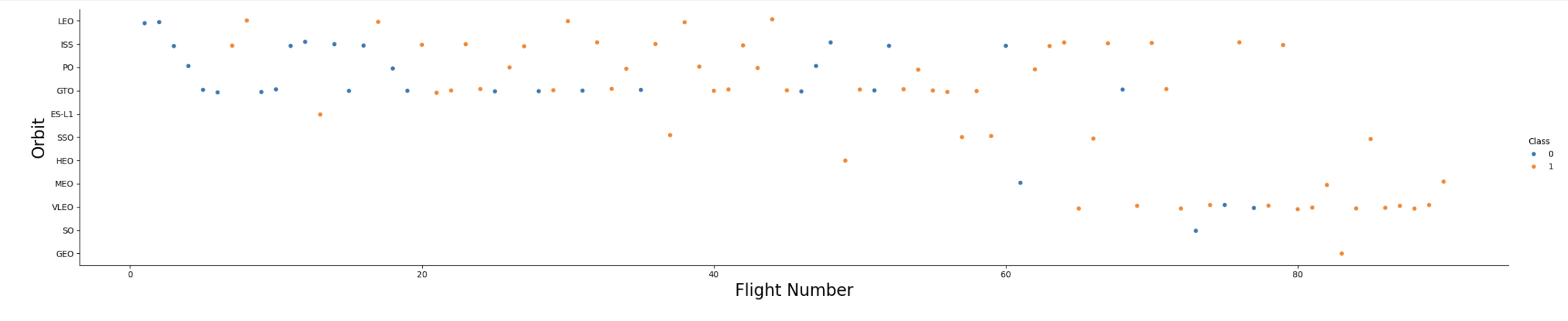
For the CCAFS SLC 40 launch site, an increase in payload mass is associated with a higher success rate for the rocket launches.

Success Rate vs. Orbit Type



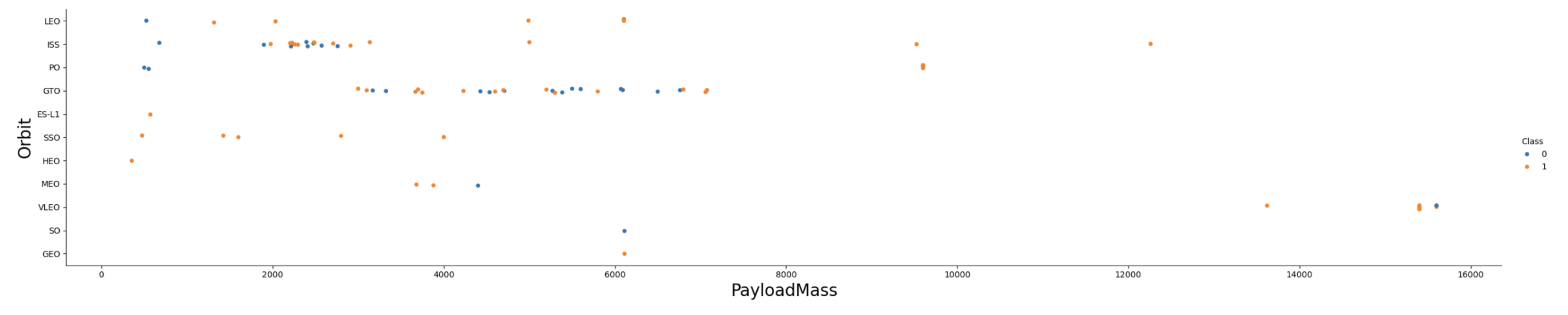
It is evident that ES-L1, GEO, HEO, and SSO achieved the highest success rates.

Flight Number vs. Orbit Type



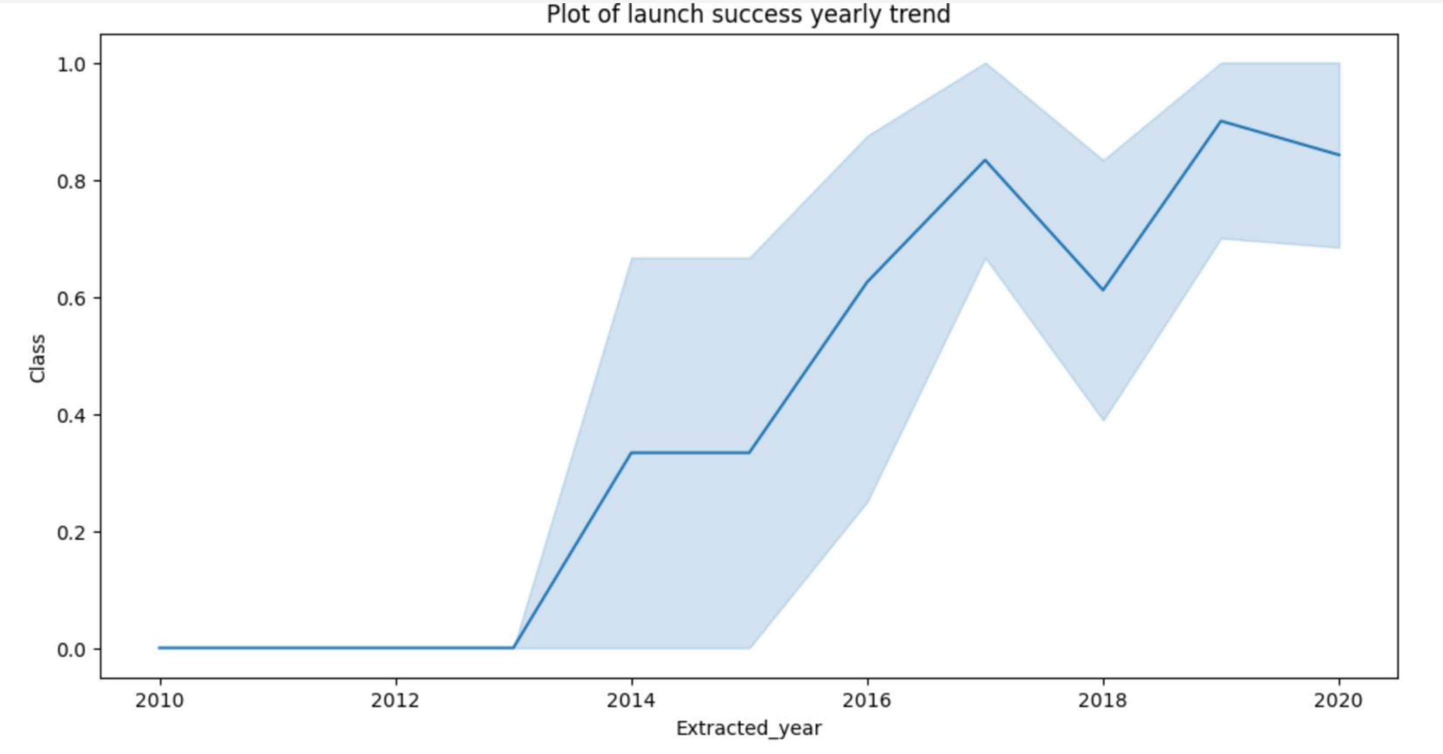
The following graph highlights the relationship between flight number and orbit type. It shows that in the LEO orbit, success correlates with the number of flights, while in the GTO orbit, there is no evident connection between flight number and the orbit.

Payload vs. Orbit Type



With heavier payloads, successful landings are more frequently observed in the PO, LEO, and ISS orbits.

Launch Success Yearly Trend



The plot indicates that the success rate has consistently improved from 2013 through 2020.

All Launch Site Names

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

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

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

We employed the keyword DISTINCT to retrieve and display only the unique launch sites present in the SpaceX dataset.

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_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

The query filters the data from the SpaceX table to display the first five entries where the launch site contains "CCA" in its name.

Total Payload Mass

```
%sql SELECT sum("PAYLOAD_MASS__KG_") FROM SPACEXTBL where "Customer" like "NASA (CRS)"
```

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

sum("PAYLOAD_MASS__KG_")
45596

This query calculates the total payload mass (in kilograms) carried by missions where NASA (CRS) was the customer, resulting in a total of 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT avg("PAYLOAD_MASS__KG_") FROM SPACEXTBL where "Booster_Version" like "F9 v1.1"

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

avg("PAYLOAD_MASS__KG_")
2928.4

This query calculates the average payload mass (in kilograms) for missions using the "F9 v1.1" booster version, resulting in an average of 2,928.4 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTBL where "Landing_Outcome" LIKE 'Success (ground pad)';
```

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

MIN(Date)
2015-12-22

This query retrieves the earliest date (December 22, 2015) when a successful landing occurred on a ground pad, as recorded in the "Landing_Outcome" column.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTBL where ("Landing_Outcome" like "Success (drone ship)" AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000)
```

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

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

This query identifies the "Booster_Version" values from the "SPACEXTBL" table where the "Landing_Outcome" is a "Success (drone ship)" and the "PAYLOAD_MASS__KG_" falls between 4000 and 6000 kilograms. The results list the booster versions that meet these criteria.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT count("Mission_Outcome") FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Success%'
```

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

count("Mission_Outcome")
100

```
%sql SELECT count("Mission_Outcome") FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%'
```

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

count("Mission_Outcome")
1

The first query counts 100 successful missions, while the second counts 1 failure.

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL) ORDER BY "Booster_Version"
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

The query retrieves booster versions with the maximum payload mass (15600 kg), sorted by booster version.

2015 Launch Records

```
%sql SELECT substr(Date, 6,2), "Booster_Version", "Launch_Site", "Landing_Outcome" FROM SPACEXTBL WHERE Landing_Outcome like 'Failure (drone ship)' AND substr(Date,0,5)='2015'
```

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

substr(Date, 6,2)	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

The query extracts the year and retrieves the booster version, launch site, and landing outcome for drone ship landing failures that occurred in 2015.

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

```
%sql SELECT "Landing_Outcome", count("Landing_Outcome") 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.
```

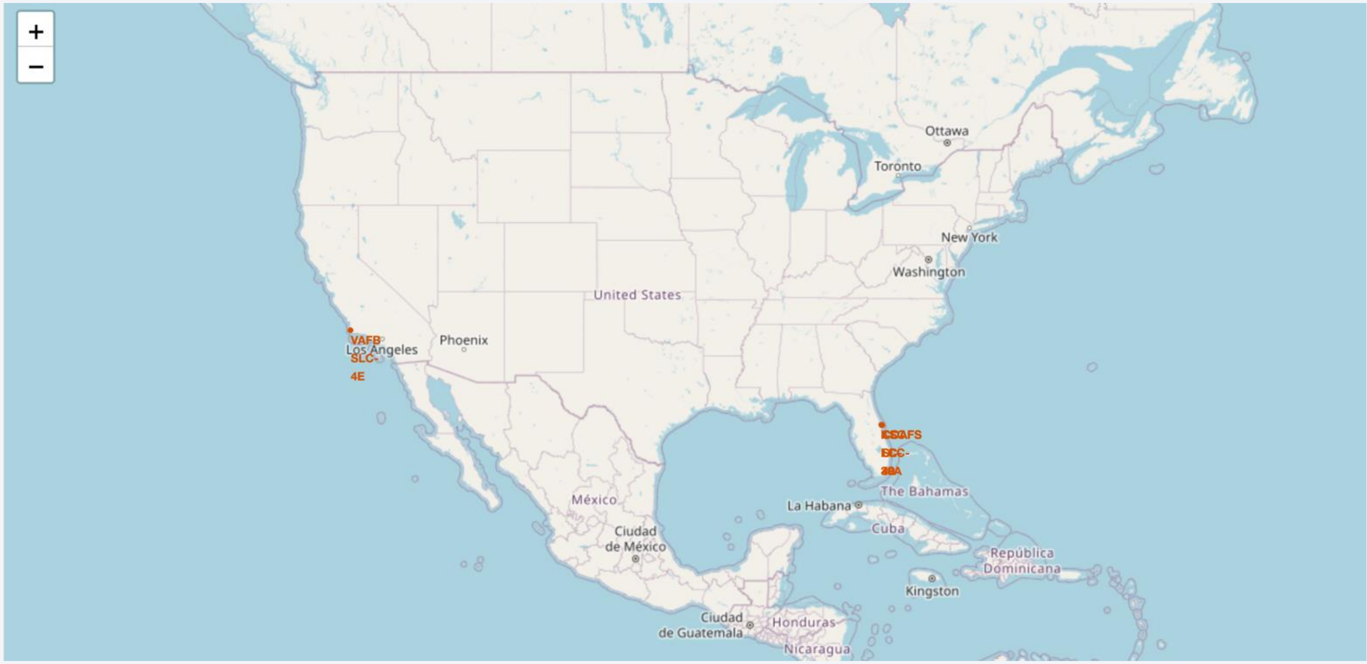
Landing_Outcome	count("Landing_Outcome")
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

This query retrieves the count of each unique landing outcome for SpaceX launches between the dates "2010-06-04" and "2017-03-20." It groups the results by the landing outcome and orders them in descending order of count. The most frequent outcome in this period is "No attempt."

Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map

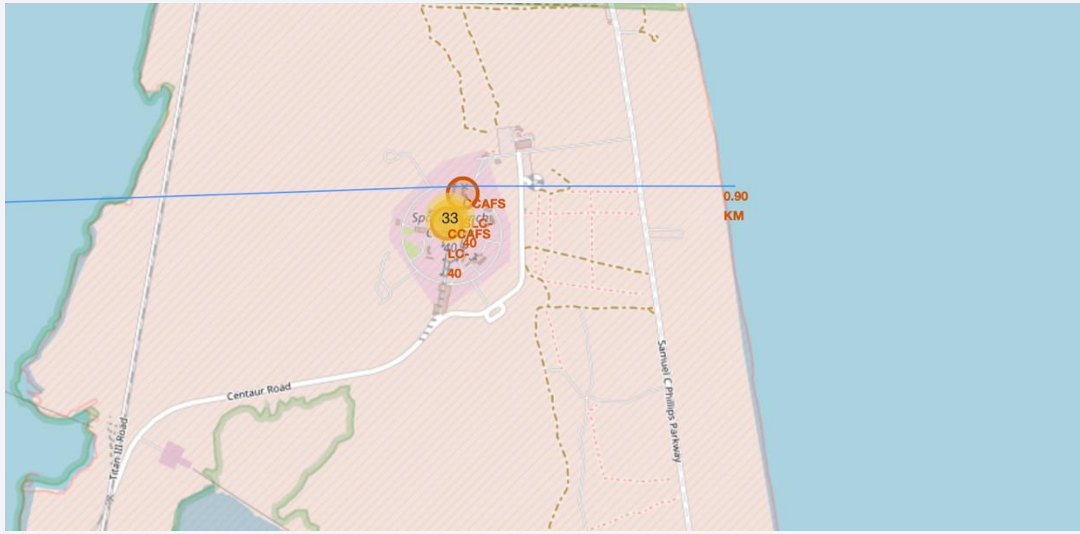


These sites are strategically positioned near the coast to ensure safe rocket launches and landings, minimizing risks to populated areas. The map provides a clear visualization of the geographic distribution of SpaceX's key launch facilities.

Success and Failure Visualized by Colors

On the map, green markers represent successful launches, while red markers indicate failed launches.

Distance

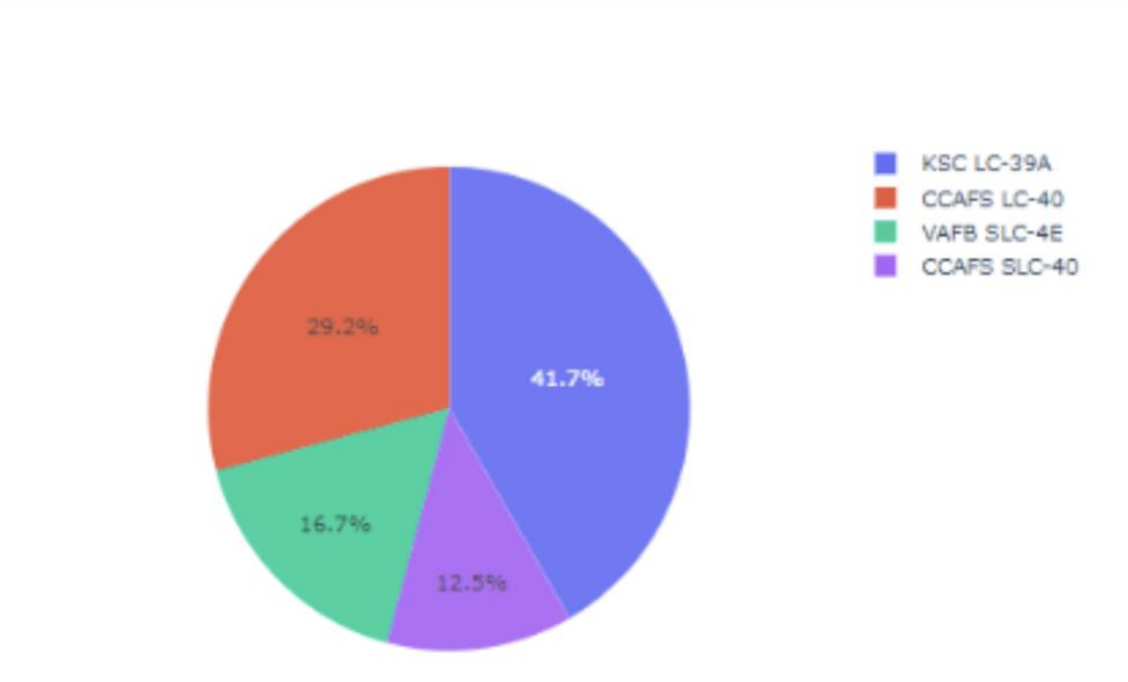


The map highlights the CCAFS launch site and its proximity to key geographical features, such as the coastline and infrastructure like roads. The blue line represents the calculated distance to the coastline, which is approximately 0.90 km. This analysis provides insights into the location's accessibility and its proximity to critical facilities.

Section 4

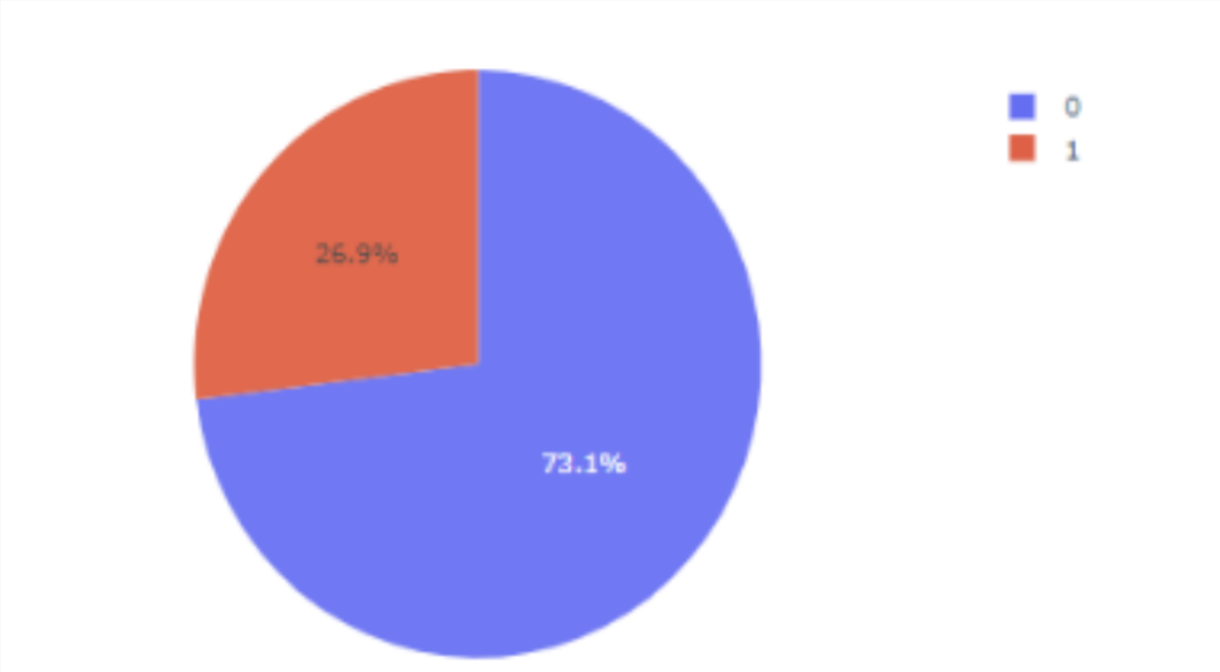
Build a Dashboard with Plotly Dash

Launch Success Count by Sites



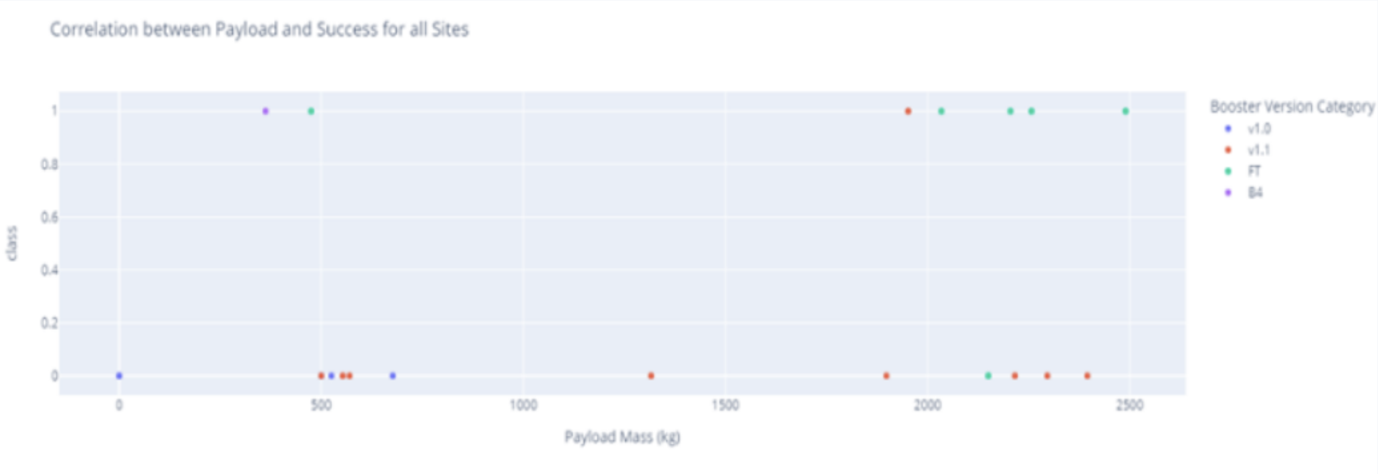
The pie chart displays the percentage of launches from each site, highlighting that KSC LC-39A has the largest share.

Launch Success vs Failure Distribution



The pie chart categorizes launches as successful or failed, showing that 73.1% were successful.

Correlation Between Payload Mass and Launch Success

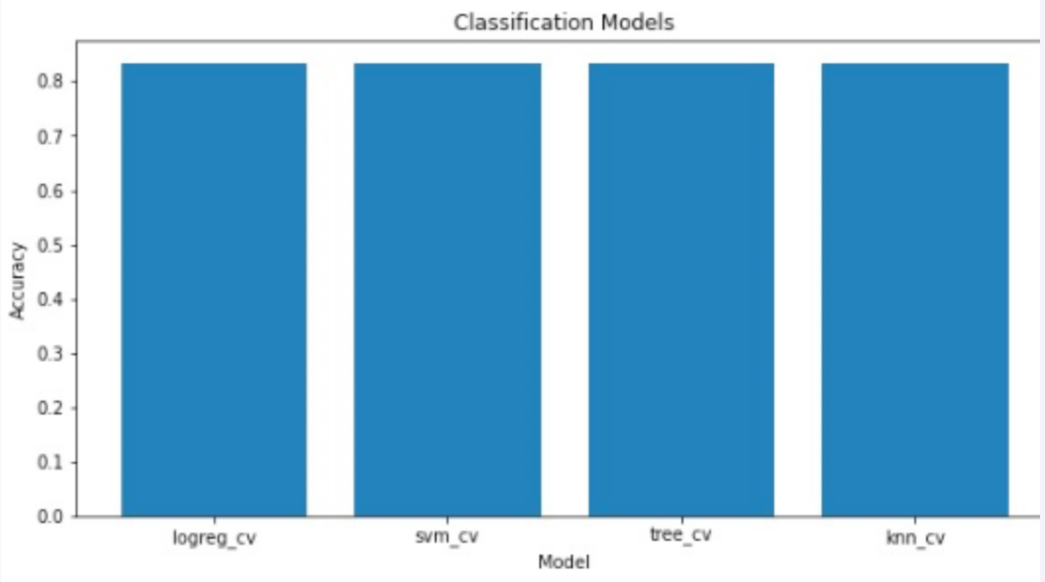


The scatter plots illustrate the correlation between payload mass, booster versions, and launch success rates.

Section 5

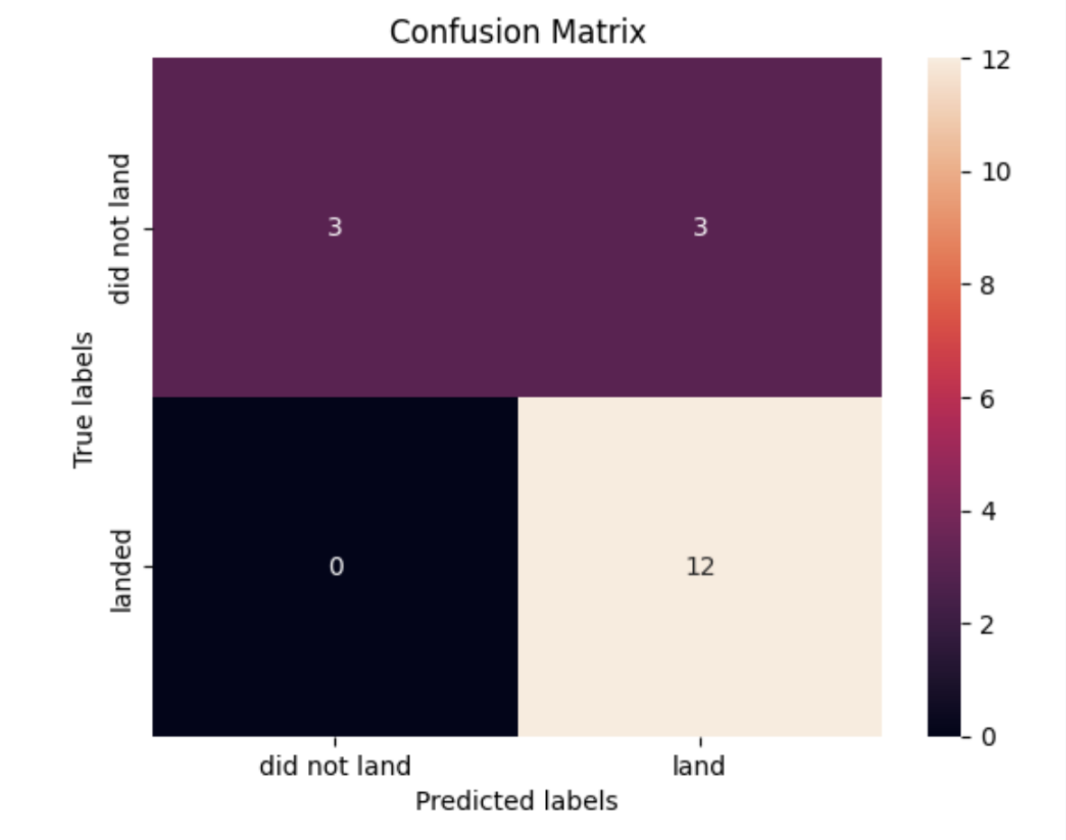
Predictive Analysis (Classification)

Classification Accuracy



The bar chart compares model accuracies, with all models achieving around **80% accuracy** after cross-validation.

Confusion Matrix



The matrix highlights predictions, where **12 landings were correctly predicted**, and **3 landings were missed**.

Conclusions

- KSC LC-39A achieved the highest number of successful launches compared to other sites, indicating its reliability.
- The decision tree classifier outperformed other models, achieving the best accuracy for predicting launch outcomes (~80%).
- Orbits such as ES-L1, GEO, HEO, SSO, and VLEO exhibited the highest success rates, showcasing their importance for mission success.
- Launch success rates steadily increased after 2013, reaching their peak by 2020, reflecting advancements in technology and operational efficiency.
- Higher payload masses were positively correlated with launch success, particularly for CCAFS SLC-40.
- Missions with droneship landings had a slightly lower success rate compared to ground pad landings, showing room for improvement in drone ship recovery systems.

Thank you!