IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kevin McCarthy
1/7/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API & Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Ploty Dash
  - Machine Learning Prediction
- Summary of all results
  - Valuable data was successfully gathered from publicly available sources.
  - Exploratory Data Analysis (EDA) helped pinpoint the most influential features for predicting the success of launches.
  - Machine Learning Prediction identified the optimal model for predicting key characteristics that significantly impact the success of launches, utilizing the entirety of the collected data.

Introduction

- Project background and context
  - In this project, we assess the potential of Space Y, a startup competing with SpaceX, a disruptor in the space industry with its cost-efficient Falcon 9 rocket launches. The objectives include estimating total launch costs, identifying optimal launch locations, understanding factors influencing landing outcomes, exploring variable relationships, and determining conditions for enhancing landing success probability.

- Key Objectives:
  - Identifying the optimal launch locations for maximizing operational efficiency.
  - Comprehensively identifying and understanding all factors influencing the landing outcome.
  - Investigating the relationships between various variables and their impact on the landing outcome.
  - Determining the optimal conditions necessary to enhance the probability of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API (https://api.spacexdata.com/v4/rockets/)
  - Web Scrapping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Augmented the collected data by generating a landing outcome label derived from summarized and analyzed features alongside outcome data.
  - Dropped redundant columns.
  - Applied One Hot Encoding to facilitate classification models.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data has been normalized, segmented into training and test datasets, and subjected to evaluation through four different classification models. The accuracy of each model was assessed using various combinations of parameters.

Data Collection

- Datasets were gathered through distinct techniques, utilizing SpaceX REST API and web scraping from Wikipedia.
  - The API provides details on rockets, launches, and payload information.
    - SpaceX REST API: https://api.spacexdata.com/v4/rockets/

Initiate SpaceX REST API request → Retrieve JSON file from the API → Create a DataFrame using the JSON data → Cleanse the data and export it

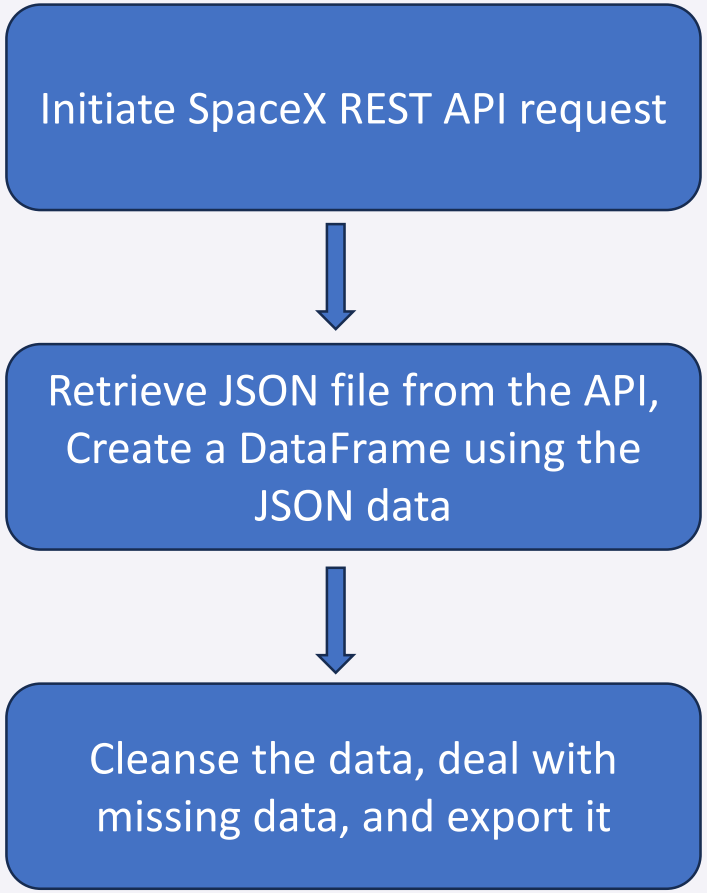
  - The data extracted through Wikipedia includes details about launches, landings, and payload.
    - URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Obtain the HTML response from Wikipedia → Extract data using BeautifulSoup → Transform the data into a DataFrame → Export the processed data

Data Collection – SpaceX API

- SpaceX provides a public API that allows users to access and utilize data. Following the outlined flowchart, I leveraged this API to retrieve information, and subsequently, the acquired data was stored for further use.

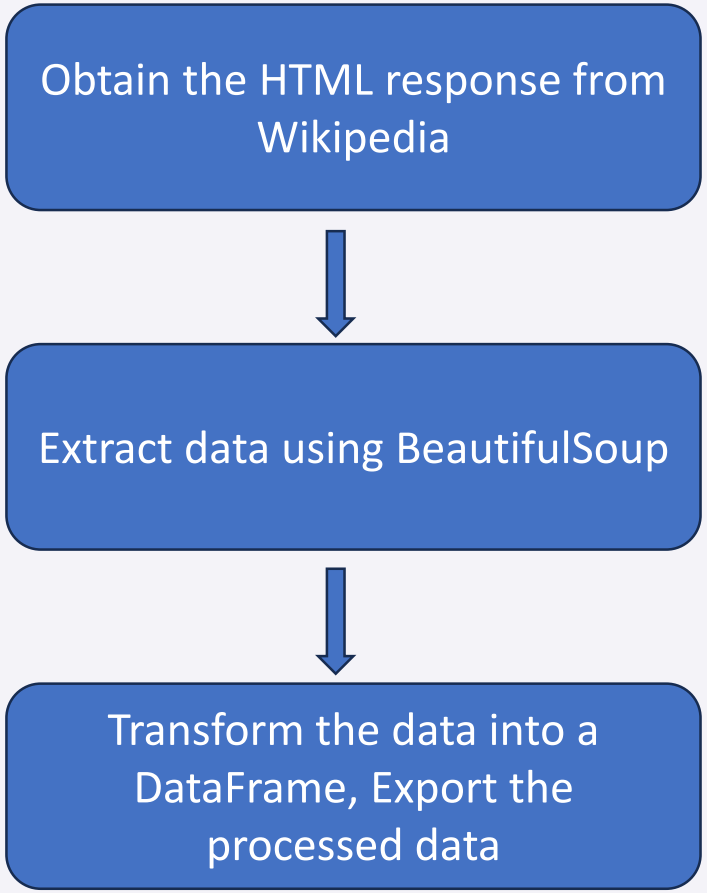
- GitHub URL https://github.com/coder299/space-y/blob/main/Data%20Collection%20Api.ipynb



Data Collection - Scraping

- Obtaining SpaceX launch data from Wikipedia involves web scraping with BeautifulSoup, specifically targeting Falcon 9 and Falcon Heavy Launch Records. The process includes extracting, parsing, and converting relevant information into a Pandas data frame. The downloaded data adheres to a predefined flowchart, ensuring proper persistence for future use.

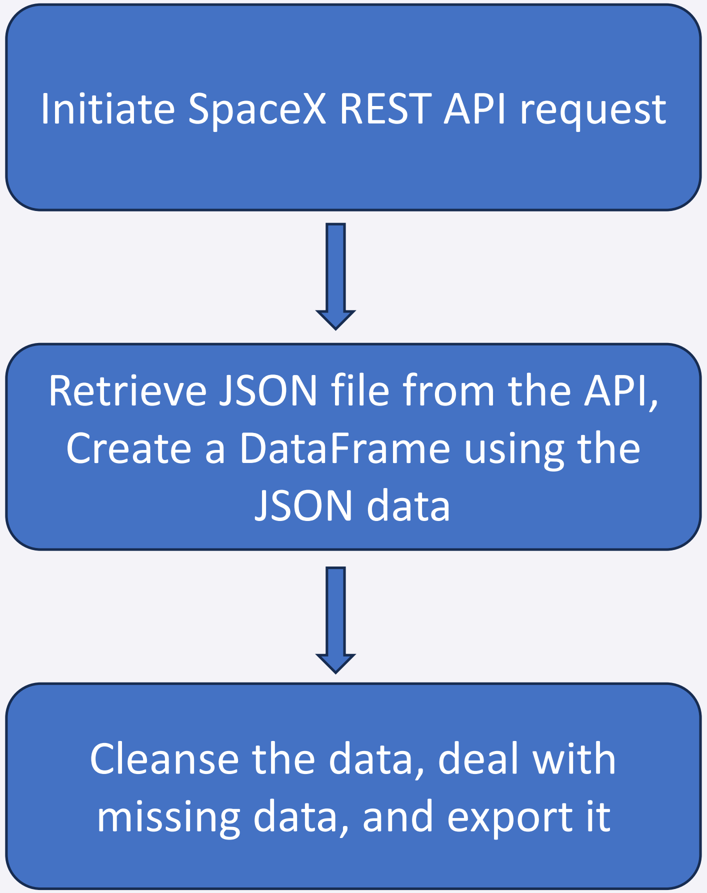
- GitHub URL https://github.com/coder299/space-y/blob/main/Web%20Scraping.ipynb



Data Wrangling

- SpaceX provides a public API that allows users to access and utilize data. Following the outlined flowchart, I leveraged this API to retrieve information, and subsequently, the acquired data was stored for further use.

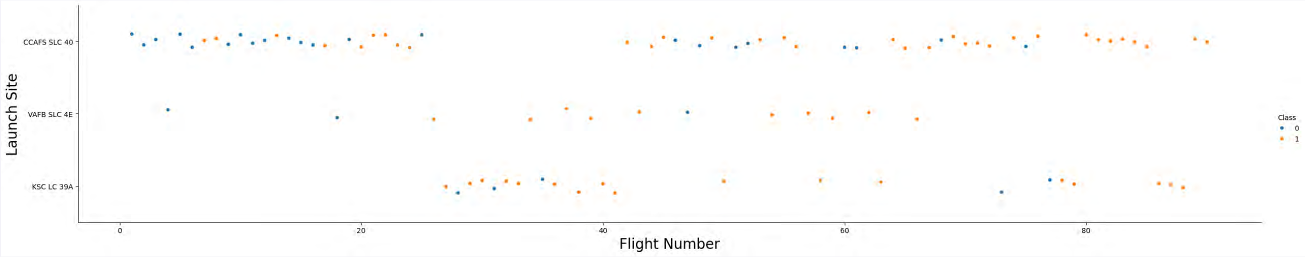
- GitHub URL
  https://github.com/coder299/space-y/blob/main/Data%20Wrangling.ipynb

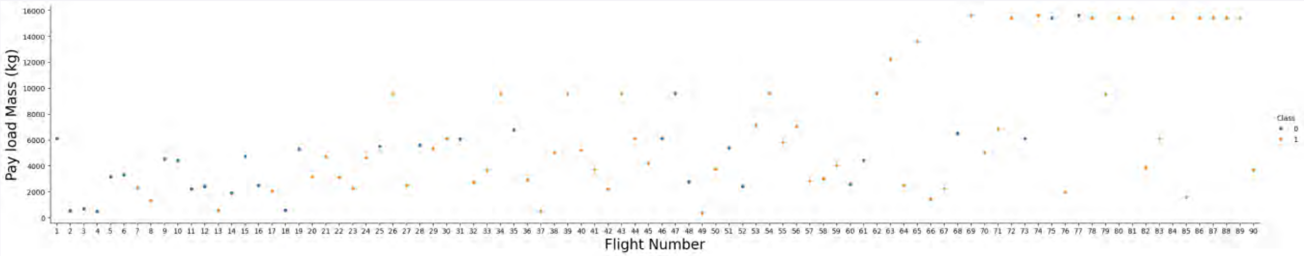


EDA with Data Visualization

Our initial approach involved employing scatter plots to explore relationships between key attributes:

- Payload and Flight Number.
- Flight Number and Launch Site.
- Payload and Launch Site.
- Flight Number and Orbit Type.
- Payload and Orbit Type.





These scatter plots vividly depict the interdependence of attributes. By discerning patterns from the graphs, it becomes evident which factors have the most significant impact on the success of landing outcomes.

GitHub URL https://github.com/coder299/space-y/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

Conducting SQL Queries for Dataset Analysis:

- Display the names of unique launch sites in the space mission.
- Display 5 records where launch sites start with the string 'CCA.'
- Display the total payload mass carried by NASA (CRS) boosters.
- Display the average payload mass carried by booster version F9 v1.1.
- List the date of the first successful landing outcome on the ground pad.
- List the names of boosters with successful drone ship landings and payload mass between 4000 and 6000.
- List the total number of successful and failed mission outcomes.
- List the booster versions that have carried the maximum payload mass.
- Display records showing month names, failure landing outcomes on the drone ship, booster versions, and launch sites for the months in the year 2015.
- Rank the count of successful landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

GitHub URL https://github.com/coder299/space-y/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- The Folium map is centered on NASA Johnson Space Center in Houston, Texas.
    - Place a red circle at NASA Johnson Space Center's coordinates, labeled with its name (using folium.Circle and folium.map.Marker).
    - Add red circles at each launch site's coordinates, with labels displaying the launch site names (using folium.Circle, folium.map.Marker, folium.features.DivIcon).
    - Utilize clustering of points to present multiple types of information for the same coordinates (using folium.plugins.MarkerCluster).
    - Employ markers to signify successful and unsuccessful landings, using green for success and red for failure (using folium.map.Marker and folium.Icon).
    - Include markers to indicate distances between launch sites and key locations (railway, highway, coastline, city) and draw lines connecting them (using folium.map.Marker, folium.PolyLine, folium.features.DivIcon).
- These elements are crafted to enhance comprehension of the problem and the data. They provide a visual representation of all launch sites, their surroundings, and the count of successful and unsuccessful landings.
- **GitHub URL** https://github.com/coder299/space-y/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard features several components including a dropdown, pie chart, rangeslider, and scatter plot.
  - The dropdown, powered by dash_core_components, empowers users to select either a specific launch site or view data for all launch sites.
  - Utilizing plotly.express, the pie chart dynamically displays the total success and failure rates corresponding to the launch site selected via the dropdown.
  - The rangeslider, a dash_core_component.RangeSlider, enables users to precisely choose a payload mass within a specified range.
  - The scatter chart, crafted with plotly.express.scatter, visually represents the relationship between two variables, focusing on the Success vs Payload Mass aspect.
- GitHub URL https://github.com/coder299/space-y/blob/main/SpaceX%20Dash%20App.py

Predictive Analysis (Classification)

- Data Preparation:
  - Load the dataset.
  - Normalize the data.
  - Split the data into training and test sets.
- Model Preparation:
  - Choose machine learning algorithms.
  - Configure parameters for each algorithm using GridSearchCV.
  - Train GridSearchModel models with the training dataset.
- Model Evaluation:
  - Identify the best hyperparameters for each model type.
  - Calculate accuracy for each model using the test dataset.
  - Visualize the Confusion Matrix.
- Model Comparison:
  - Evaluate models based on their accuracy.
  - Select the model with the highest accuracy (Please refer to the Notebook for detailed results).
- GitHub URL https://github.com/coder299/space-y/blob/main/Machine%20Learning%20Prediction.ipynb

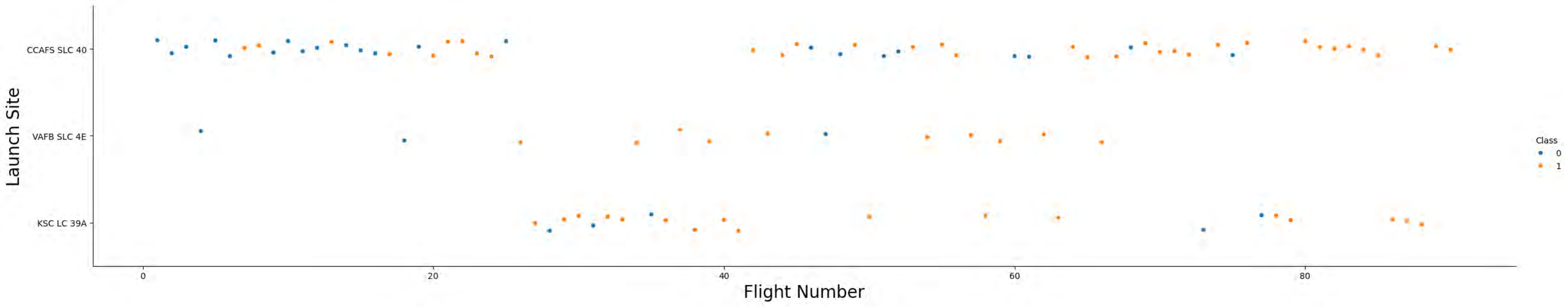
Results

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

Section 2

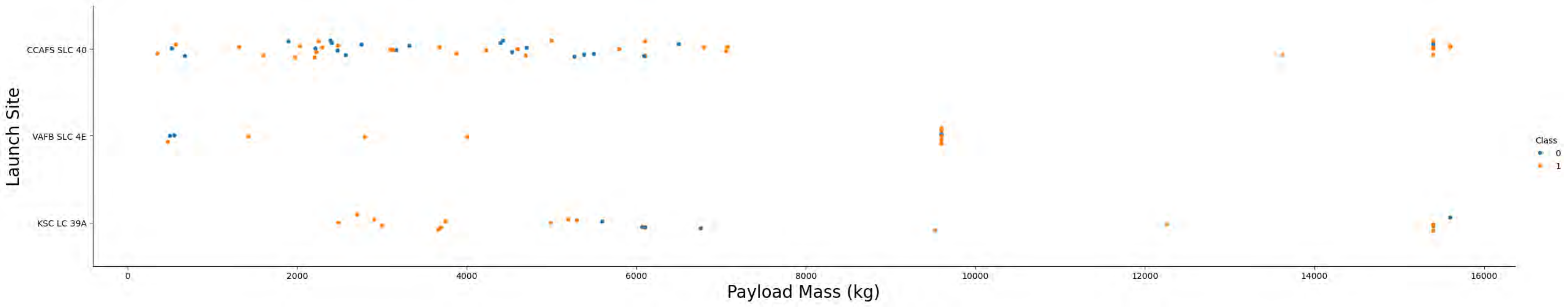
Insights drawn from EDA

Flight Number vs. Launch Site



- Looking at the chart, you can see that the top-performing launch site currently is CCAF5 SLC 40, with a high success rate in recent launches.
- Following closely is VAFB SLC 4E in second place, and KSC LC 39A in third place.
- Additionally, there's a noticeable upward trend in the overall success rate over time.

Payload vs. Launch Site

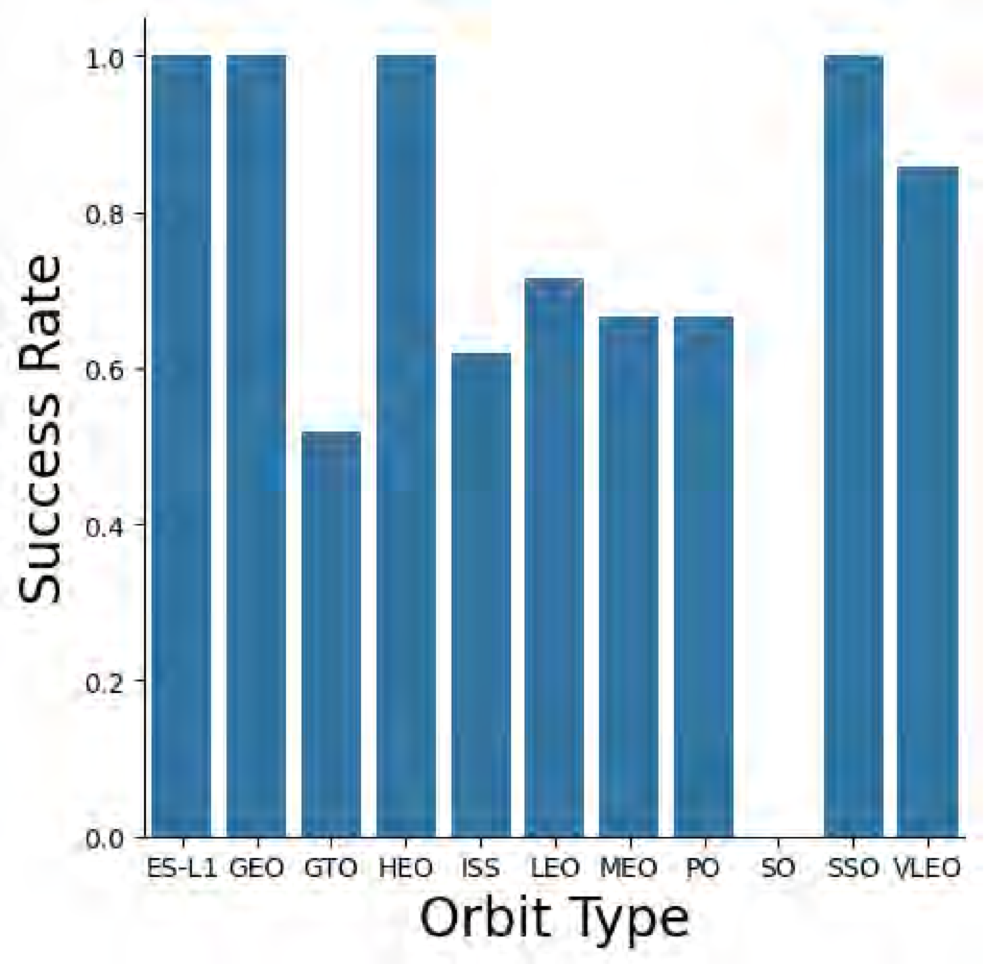


- Payloads exceeding 9,000 kilograms demonstrate an outstanding success rate.
- Payloads surpassing 12,000 kilograms appear feasible exclusively on CCAFS SLC 40 and KSC LC 39A launch sites.

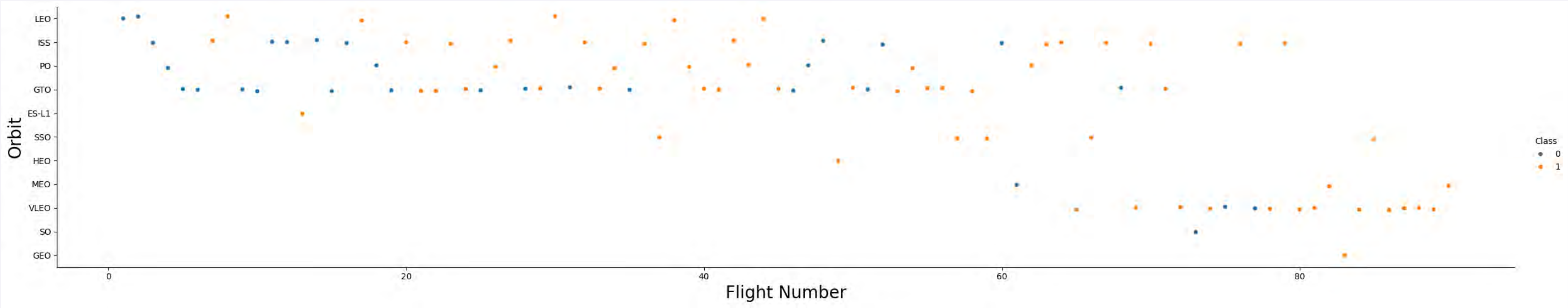
Success Rate vs. Orbit Type

This illustration shows how certain orbits may impact landing outcomes, with 100% success rates observed for specific orbits like SSO, HEO, GEO, and ES-L1, while the SO orbit resulted in a 0% success rate.

However, a more thorough analysis reveals that some orbits, such as GEO, SO, HEO, and ES-L1, have only occurred once, indicating the need for additional datasets to identify patterns or trends before drawing any conclusions.

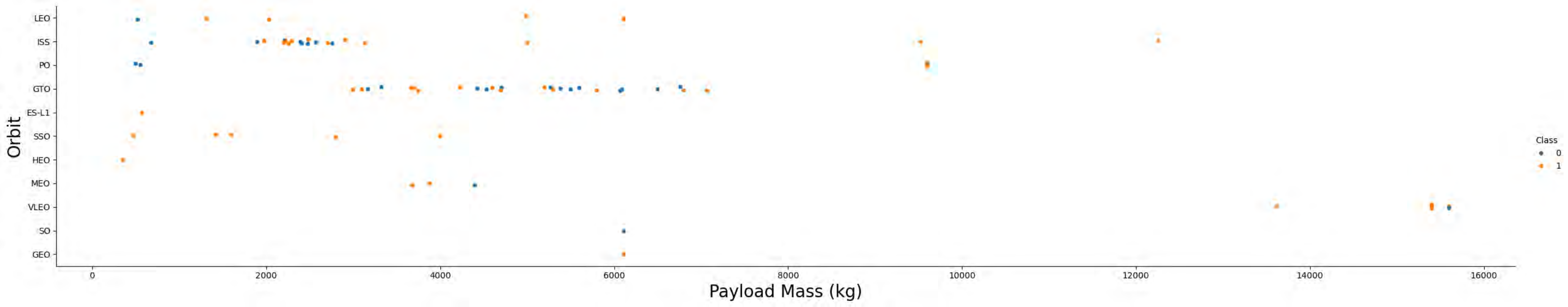


Flight Number vs. Orbit Type



- It appears that the success rate has improved over time for all orbits, with the VLEO orbit emerging as a new business opportunity due to its recent increase in frequency.

Payload vs. Orbit Type

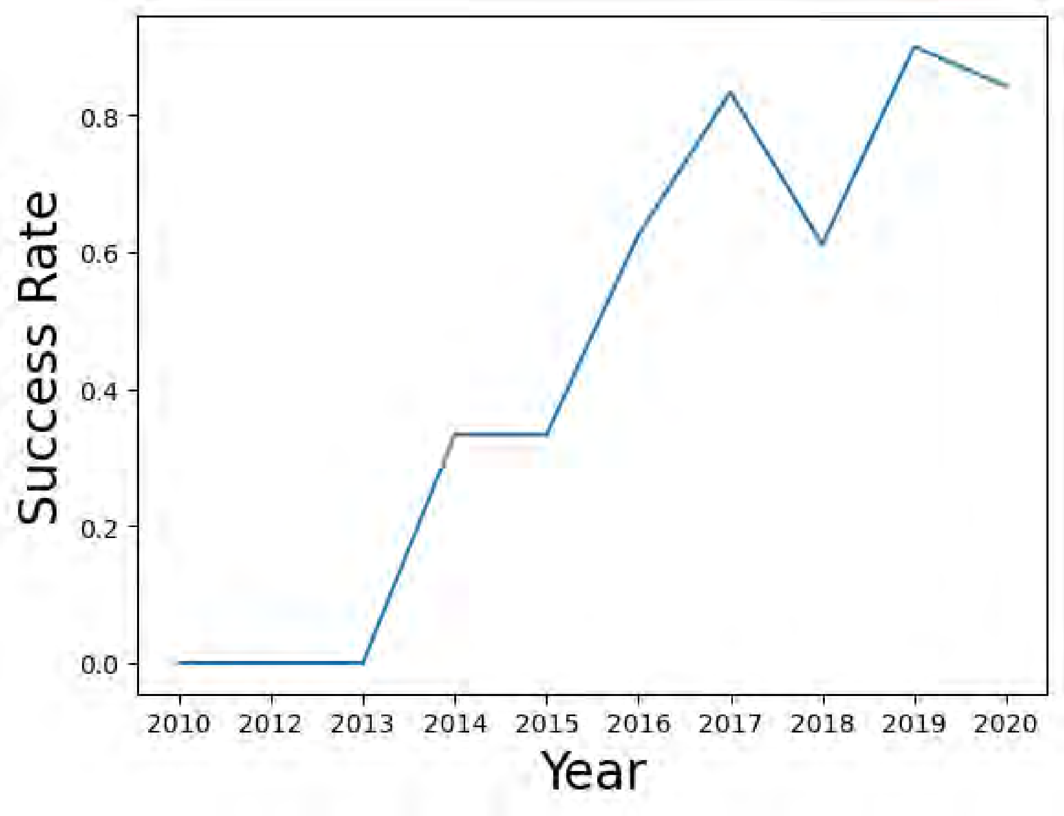


- The payload weight significantly impacts launch success rates in specific orbits. For instance, heavier payloads enhance success rates in Low Earth Orbit (LEO). Conversely, reducing payload weight for a Geostationary Transfer Orbit (GTO) improves launch success.

Launch Success Yearly Trend

The trajectory of success rates for SpaceX launches showed a notable increase from 2013 onwards, steadily climbing until the year 2020.

This upward trend suggests that the initial three years of this period were characterized by a phase of adjustments and technological improvements. During these early years, it is evident that SpaceX was actively refining its processes and enhancing its technological capabilities to achieve the subsequent successes observed in the following years.



All Launch Site Names

```
In [8]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;

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

Out[8]:

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

- In our query, we utilized the keyword DISTINCT to display only unique launch sites from the SpaceX data.

Launch Site Names Begin with 'CCA'

In [9]:

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

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

Out[9]:

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

- Using the WHERE clause in conjunction with the LIKE clause filters launch sites containing the substring "CCA," and applying LIMIT 5 displays the first 5 records resulting from this filter.

Total Payload Mass

```
In [10]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)';

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

Out[10]:

TOTAL_PAYLOAD
45596

- We computed the overall payload transported by NASA boosters to be 45,596 using the query provided.

Average Payload Mass by F9 v1.1

```
In [11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

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

Out[11]:

AVG_PAYLOAD
2928.4

- By filtering the data based on the booster version and computing the average payload mass, we arrived at a value of 2,928.4 kg.

First Successful Ground Landing Date

```
In [12]: %sql SELECT MIN (DATE) AS "First Successful Landing" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)

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

Out[12]:

First Successful Landing
2015-12-22

- Filtering the data for successful ground pad landings and finding the minimum date value reveals the initial occurrence, which occurred on 2015/12/22.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [13]:
```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Out
```

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

Out[13]:

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

- We utilized the WHERE clause to filter for boosters that successfully landed on the drone ship. Additionally, we applied the AND condition to ascertain successful landings with a payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

In [14]:

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Out[14]:

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

- By grouping mission outcomes and tallying records for each category, we arrived at the summarized results provided above.

Boosters Carried Maximum Payload

```
In [15]:  %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM

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

Out[15]:

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

- We employed a subquery to filter the data, focusing on the maximum payload mass using the MAX function. The main query then utilizes the results from the subquery, returning unique booster versions through the SELECT DISTINCT operation along with their corresponding heaviest payload masses.

2015 Launch Records

```
In [16]:  %sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] FROM SPACEXTBL where

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

Out[16]:

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

- In this query, it fetches the month, booster version, launch site with unsuccessful landing, and landing date in the year 2015. The Substr function is used to extract either the month or the year from the date. Substr(DATE, 6, 2) displays the month, while Substr(DATE, 0, 5) displays the year.

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

In [18]:
```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTBL
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

Out[18]:

Landing_Outcome	count_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

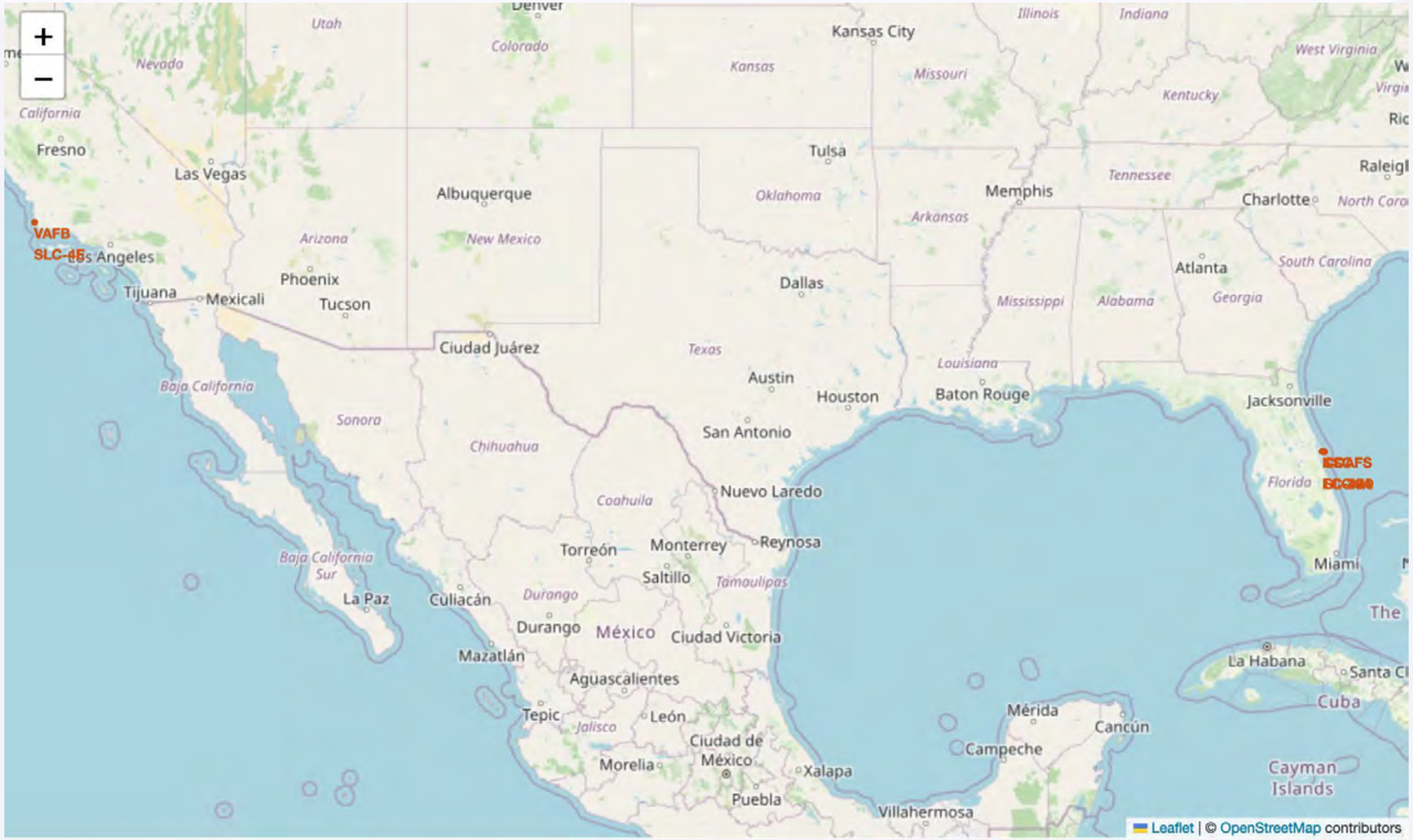
- This query fetches landing outcomes and their respective counts for successful missions that occurred between June 4, 2010, and March 20, 2017. The GROUP BY clause organizes the results based on landing outcomes, and the ORDER BY COUNT DESC arranges the results in descending order.

Section 3

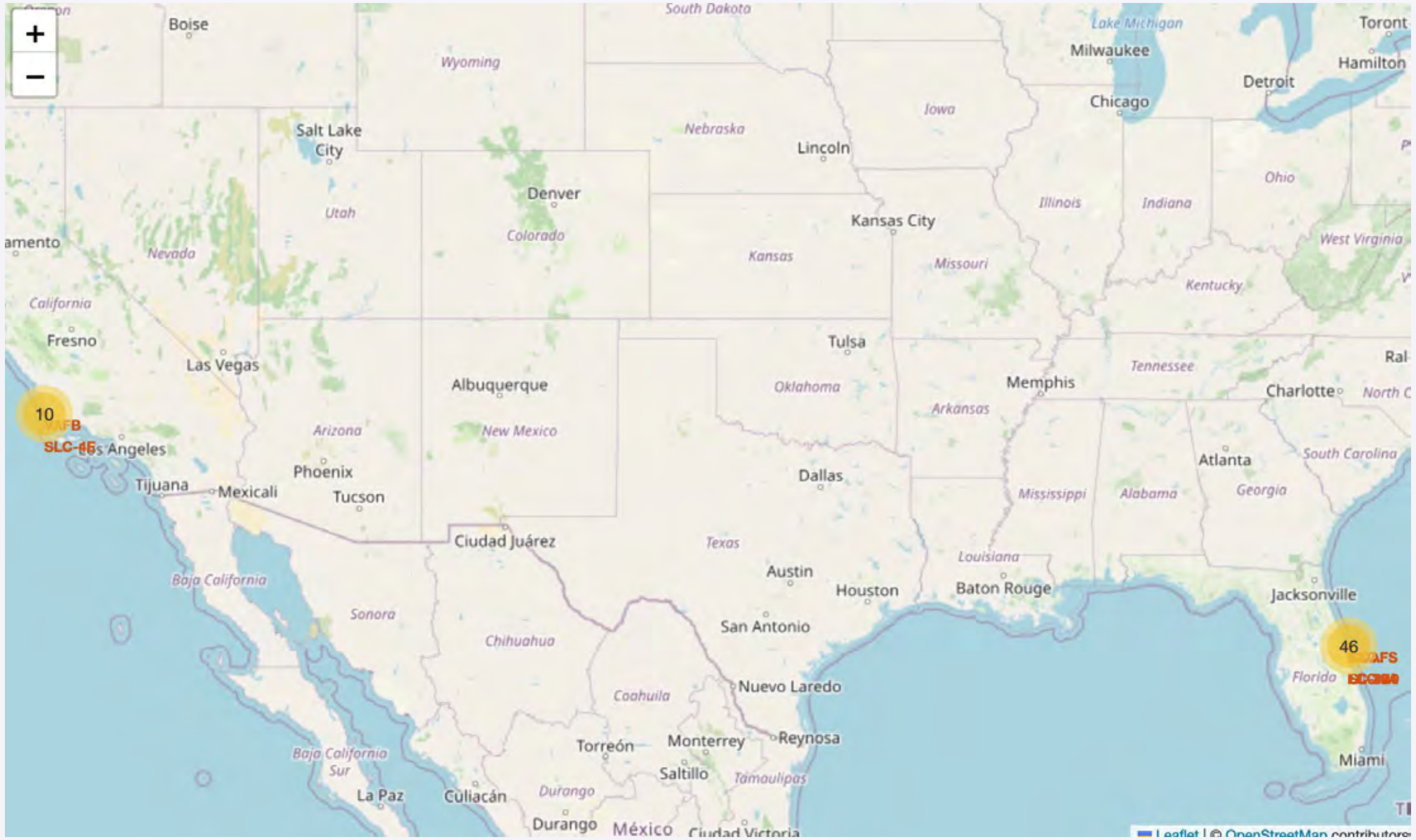
Launch Sites Proximities Analysis

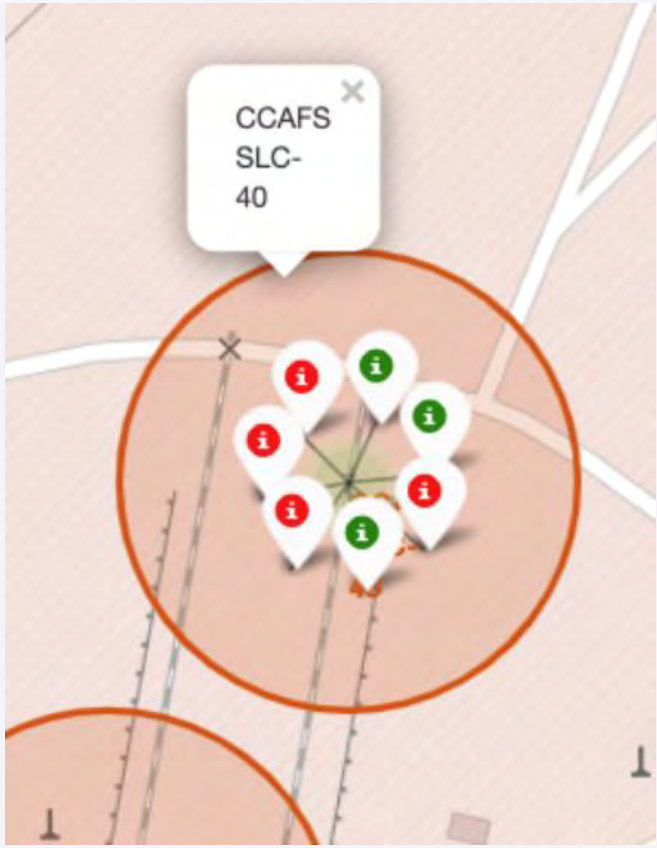
All the Launch Site Locations

SpaceX strategically positions its launch sites along the picturesque coastlines of the United States, a deliberate choice that leverages the expansive and accessible coastal regions for their rocket launches.



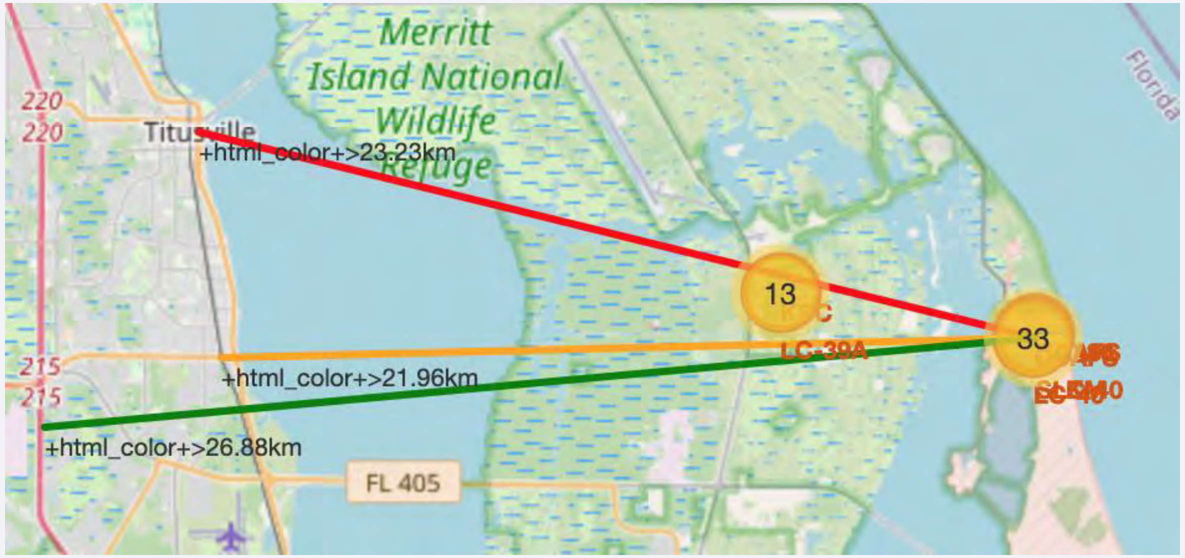
Map with Color-Coded Markers in Folium

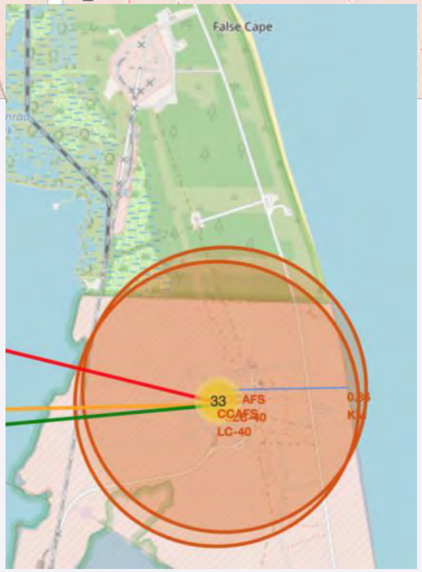




- CCAFS SLC-40: The green marker signifies successful launches, while the red marker indicates unsuccessful ones.

Logistics and Safety





Are launch sites in close proximity to railways? Yes

Are launch sites in close proximity to highways? Yes

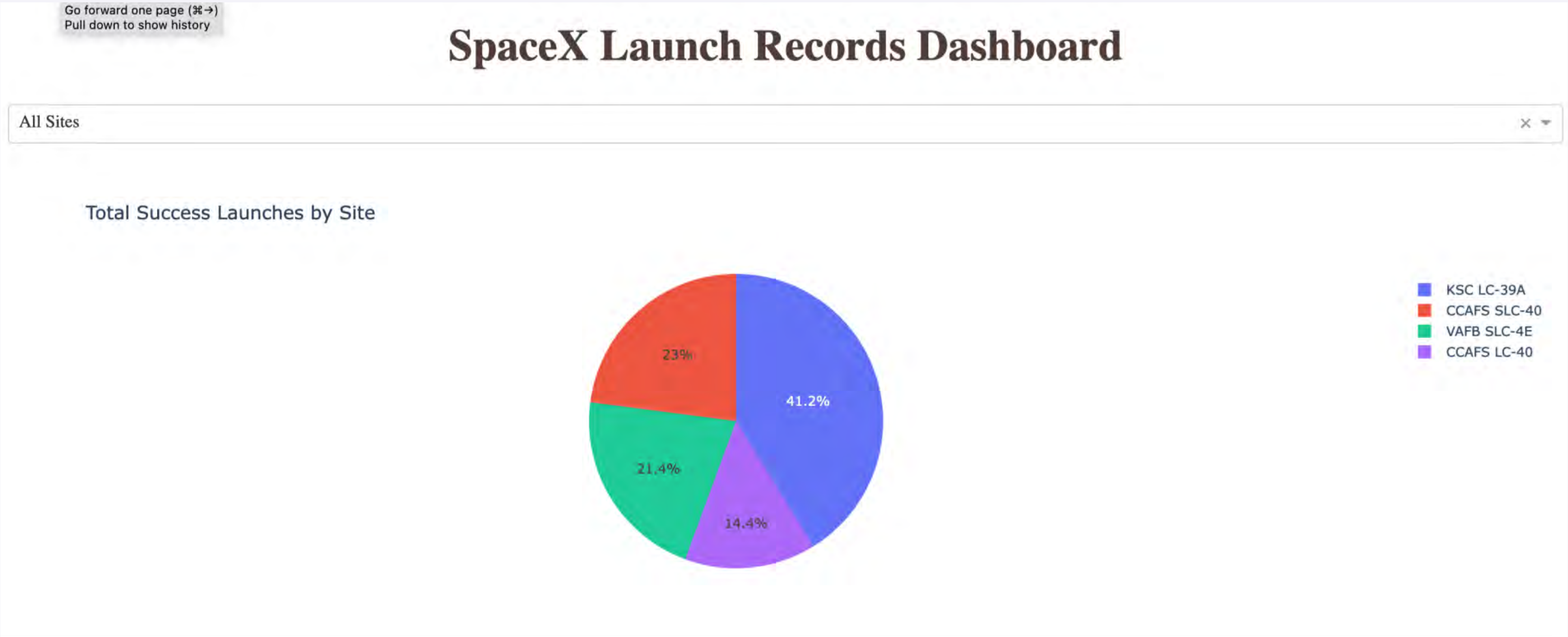
Are launch sites in close proximity to coastline? Yes

Do launch sites keep certain distance away from cities? Yes

Section 4

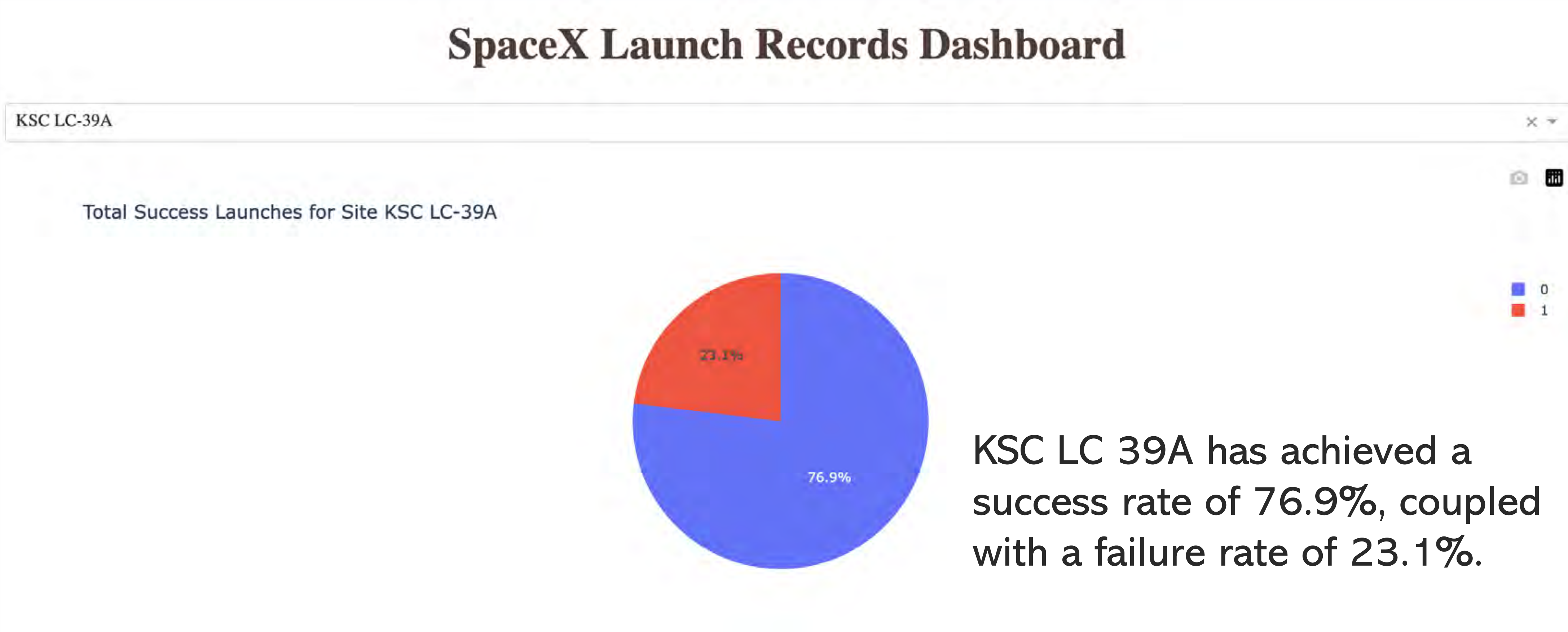
Build a Dashboard with Plotly Dash

Dashboard: Successful Launches Based on Launch Site



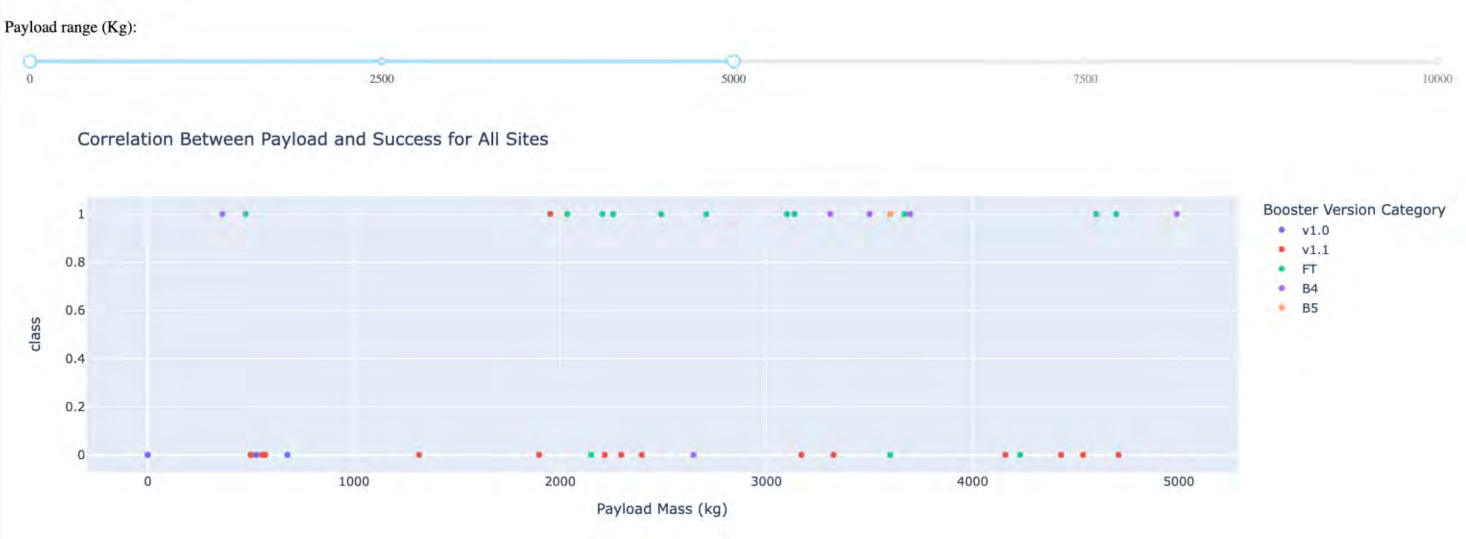
- The location of launch sites appears to play a crucial role in the success of missions. It's evident that KSC LC-39A boasts the highest success rate among launch sites.

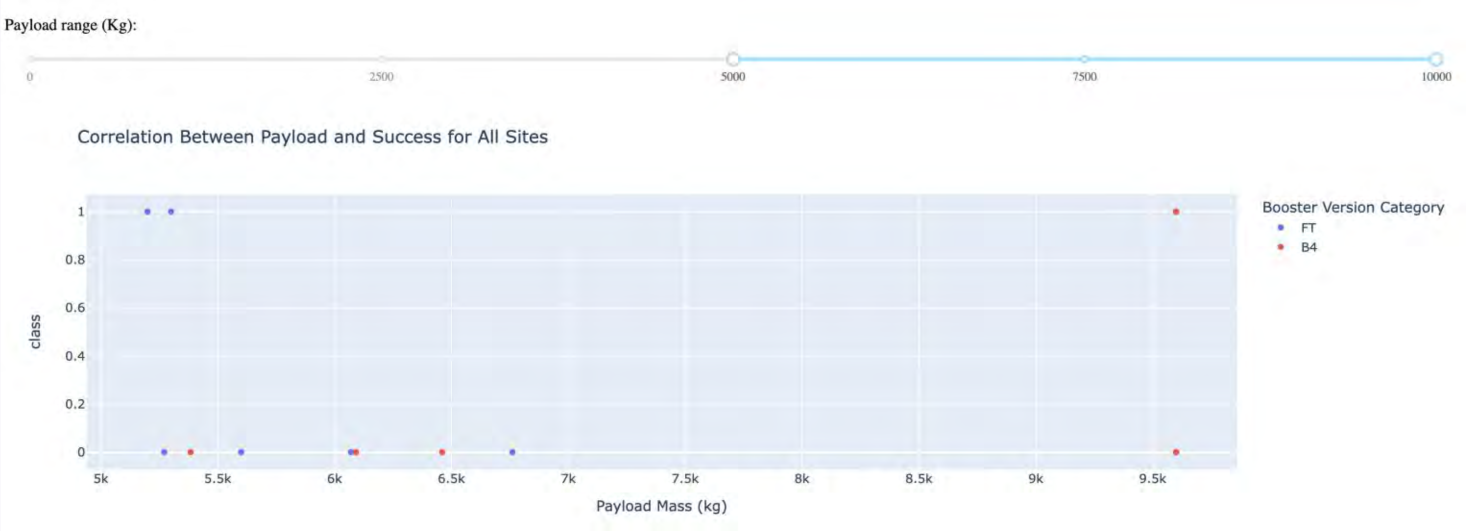
The Highest launch-success Ratio: KSC LC-39A



KSC LC 39A has achieved a success rate of 76.9%, coupled with a failure rate of 23.1%.

Payload vs. Launch Outcome





Payloads with lower weight demonstrate a higher success rate compared to their heavier counterparts.

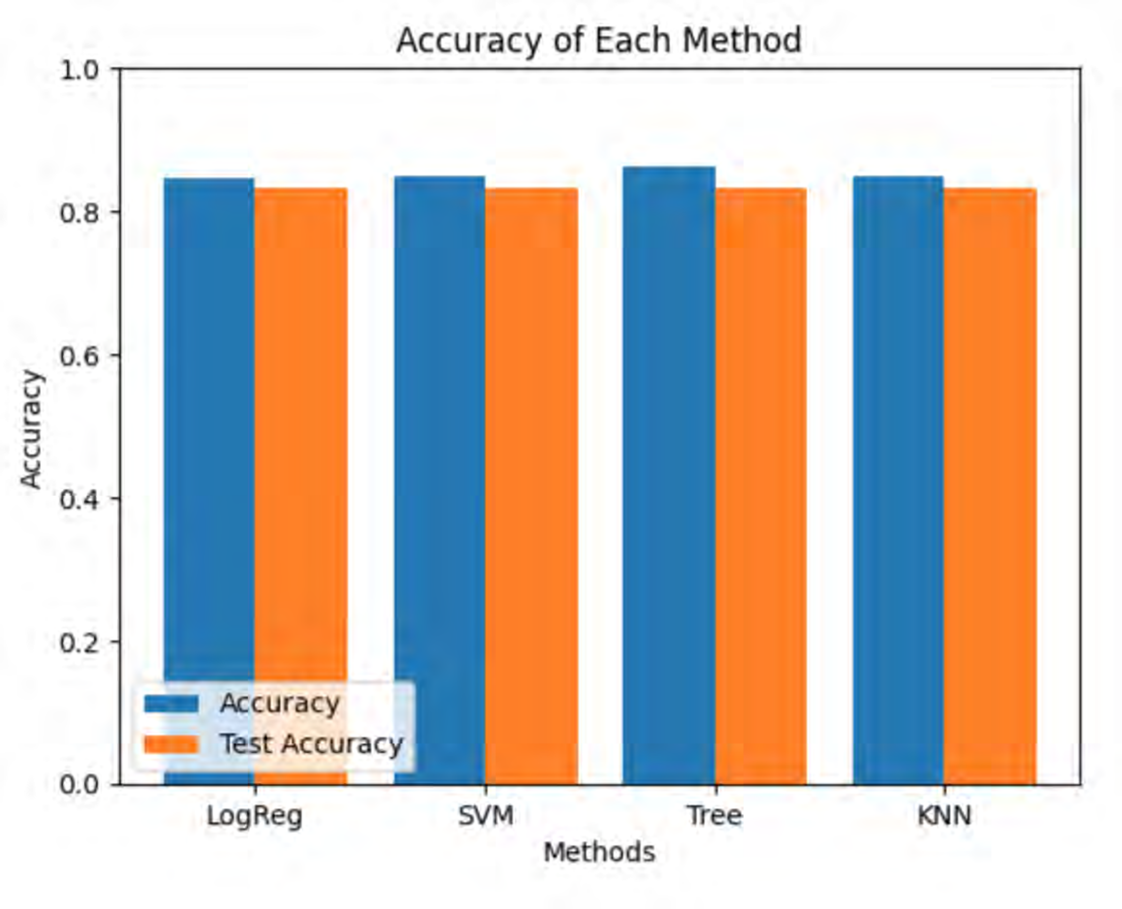
Section 5

Predictive Analysis (Classification)

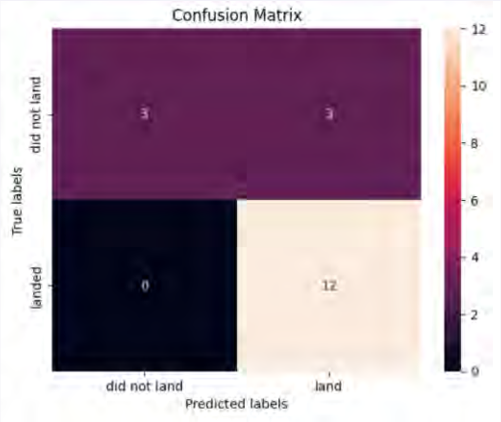
Classification Accuracy

- Best model is DecisionTree with a score of 0.8625

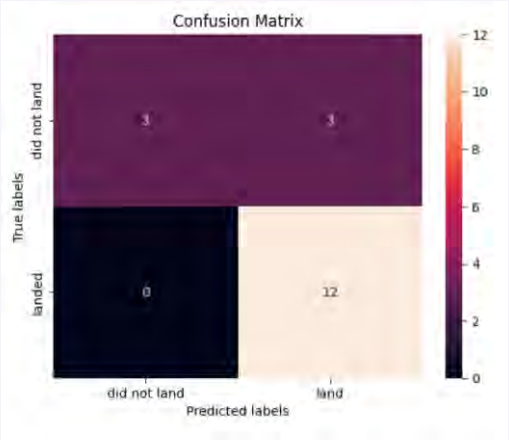
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.8625	0.83333
KNN	0.84821	0.83333



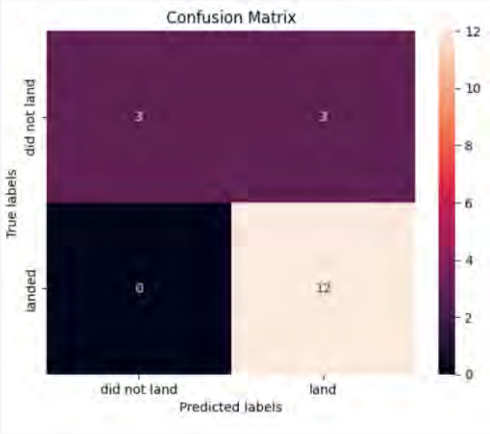
Confusion Matrix



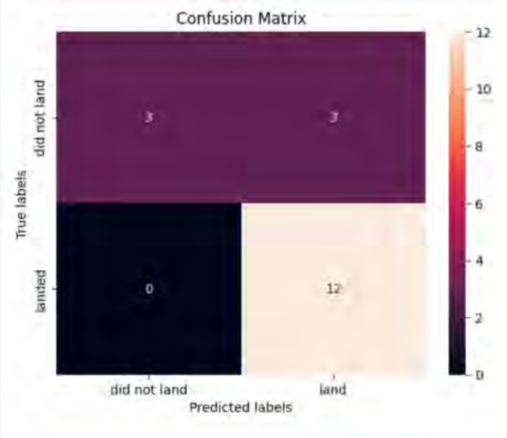
Decision Tree



kNN



Logistic regression



SVM

- The decision tree classifier's confusion matrix indicates its ability to differentiate between various classes. The primary issue lies in false positives, where the classifier wrongly identifies an unsuccessful landing as a successful one.

Conclusions

- Mission success hinges on the launch site, orbit, and prior launches, revealing a learning curve.
- GEO, HEO, SSO, and ES-L1 orbits exhibit the highest success rates.
- Better mission success is generally associated with lighter payloads, especially in specific orbits.
- Payloads weighing ≤4000kg consistently outperform heavier ones, while those exceeding 7,000kg are deemed less risky.
- The preferred model is the Decision Tree Algorithm, chosen for its superior train accuracy, making the Tree Classifier Algorithm optimal for this dataset.
- SpaceX's success rate has continually improved since 2013, showcasing ongoing advancements.
- KSC LC-39A claims the highest success rate among launch sites at 76.9%.
- The SSO orbit demonstrates a flawless 100% success rate in multiple occurrences.

Appendix

- Please check the Folium images and other images on GitHub folder.

Thank you!