Winning Space Race with Data Science

Jean Paul Maidana
December 19, 2024

Outline

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

Executive Summary

- SpaceX stands as a leader in the aerospace industry, redefining the economics of space exploration with innovative technologies. Traditional rocket launches cost $165 million, whereas SpaceX achieves similar objectives at $60 million due to its reusable rocket systems.
- This report explores the feasibility of competing with SpaceX by analyzing historical launch data. Utilizing SpaceX's REST API and additional web-scraped data, we conducted data preprocessing, exploratory data analysis (EDA), and predictive modeling to identify key factors behind SpaceX's success.
- Interactive visual analytics were developed using Folium and Plotly Dash, revealing insights into launch site performance, payload dynamics, and success trends. Additionally, machine learning models, including Decision Trees and Logistic Regression, were applied to predict the success of first-stage landings. Our findings offer actionable insights for new entrants, like Space Y, to optimize their operations and enhance 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 long been a barrier, with traditional rocket launches averaging $165 million per launch. SpaceX disrupted this paradigm through groundbreaking advancements, such as reusable first-stage rockets, reducing the cost to approximately $60 million. This revolutionary approach not only made space exploration more feasible but also set a new benchmark for efficiency in the industry.
- In this project, we delve into the critical factors contributing to successful first-stage rocket landings. By employing data science methodologies, we analyze variables such as payload mass, launch site, orbit type, and launch trends. Our insights aim to inform future strategies for competing in the burgeoning private aerospace sector.

Section 1

Methodology

Methodology

Executive Summary

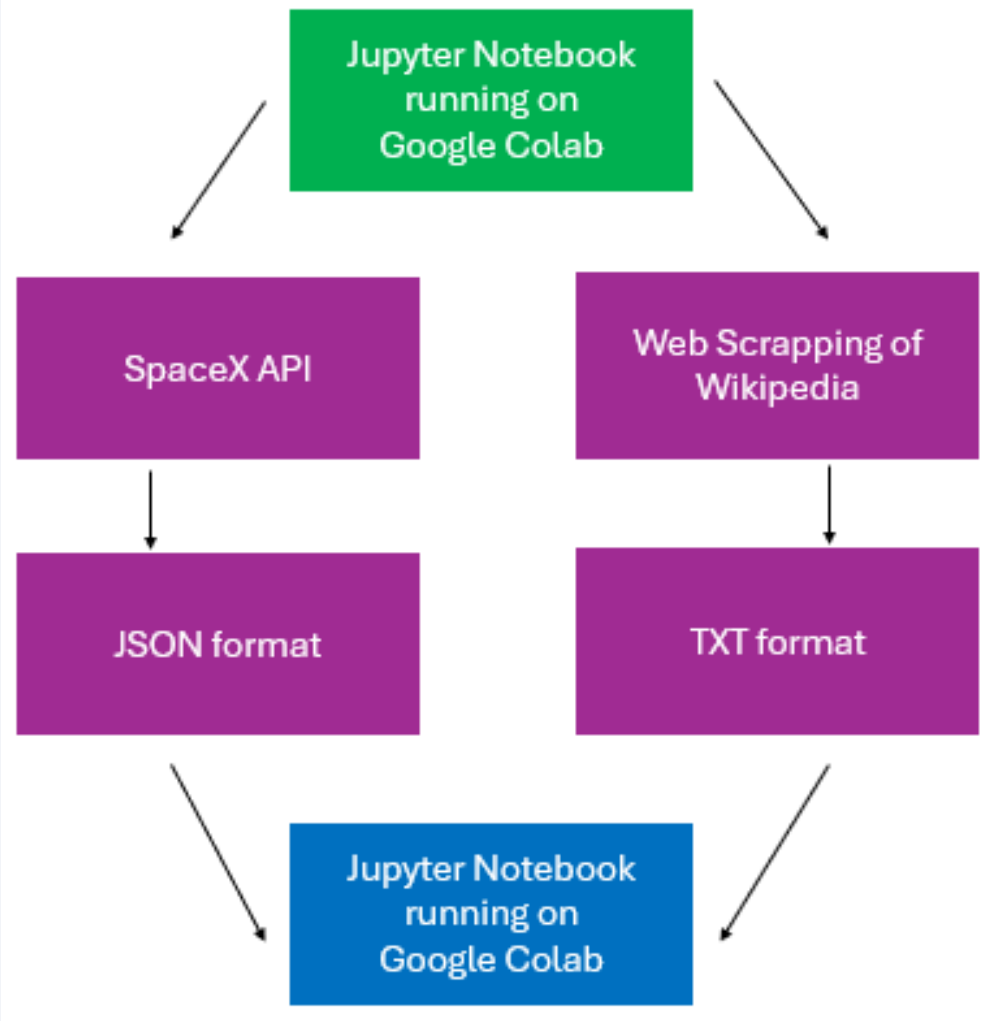
- Data collection methodology: The data was gathered using the SpaceX REST API, supplemented by web scraping techniques to extract additional information from relevant Wikipedia pages.
- Perform data wrangling: The data underwent preprocessing using Pandas and NumPy. Key techniques includes like One-hot encoding, Removal of unnecessary columns and Data normalization and standardization
- Perform exploratory data analysis (EDA): Visualization and statistical exploration were conducted using libraries such as Seaborn and Matplotlib. SQL queries were also employed for in-depth data analysis.
- Perform interactive visual analytics using Folium and Plotly Dash: Dynamic visualizations were created using tools like Folium and Plotly Dash to enhance insights and facilitate interactive data exploration.
- Perform predictive analysis using classification models: The dataset was split into training and testing sets. Optimal algorithms and hyperparameters were identified through Grid Search. The selected model and parameters were adopted for deployment to achieve effective predictions.

Data Collection

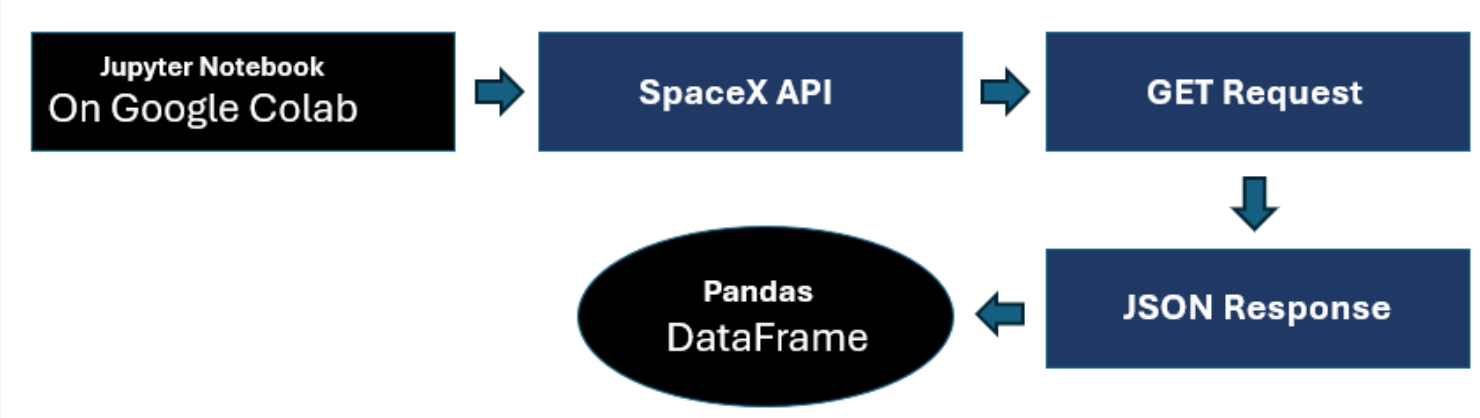
Sources of data collection

- SpaceX API: A publicly available REST API providing comprehensive data on launches, rockets, cores, capsules, Starlink, launchpads, and landing pads.
- Wikipedia: A free, volunteer-driven online encyclopedia that serves as a rich source of information, maintained and hosted by the Wikimedia Foundation.

Data collection diagram



Data Collection – SpaceX API



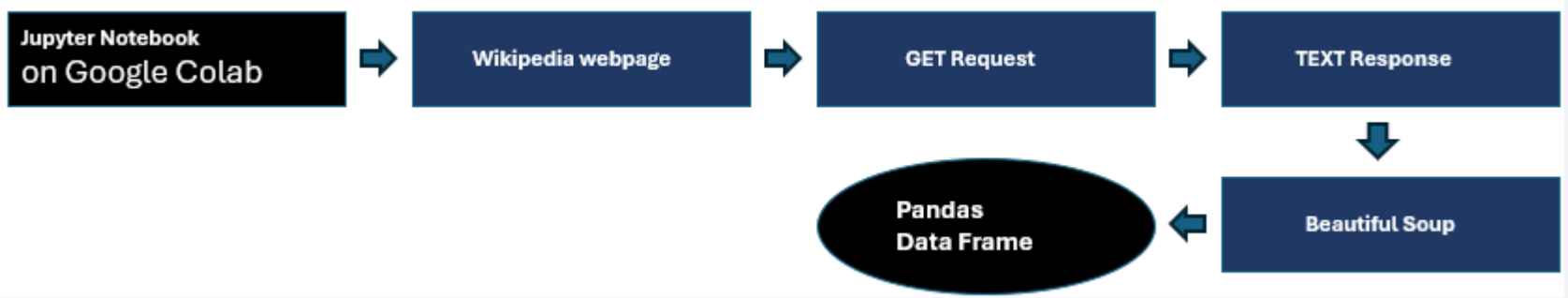
Process Overview

Initiation: Utilized Jupyter Notebook on IBM Watson to import essential libraries such as pandas, NumPy, and requests.
API Request: Established a URL GET request to the SpaceX API.. Extracted relevant data fields, including geospatial information, rocket type, orbit details, flight numbers, and more.
Data Conversion: Received responses in JSON format, which were then transformed into a Pandas DataFrame for further analysis and visualization

Additional Resources

Explore the complete SpaceX API implementation on GitHub: Link

Data Collection - Scraping



Process Overview

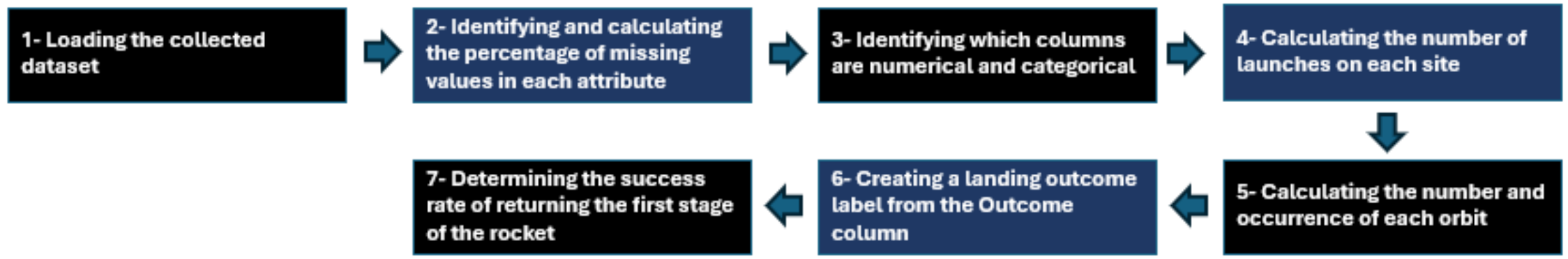
To perform this task, we utilized Python libraries such as BeautifulSoup and requests. The data source was a Wikipedia webpage titled “Space X Falcon 9 First Stage Landing Prediction.”

We initiated an HTTP GET request, receiving the response in text format. Using BeautifulSoup, we effectively extracted tables and columns from the text data.

The extracted information was then converted into a structured format and stored in a Pandas DataFrame for further analysis.

Github repository: The complete web scraping task notebook can be accessed from This Link

Data Wrangling



Process Overview

During this stage, we imported essential Python libraries such as pandas and NumPy. We loaded the collected dataset from the previous step to perform exploratory data analysis (EDA). The primary goal was to clean the dataset and identify key features to train a machine learning model effectively.

Github repository: The completed Data Wrangling notebook can be accessed here Link

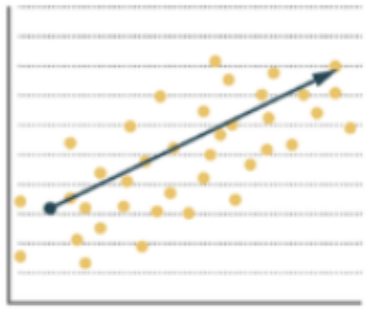
EDA with Data Visualization

Process Overview

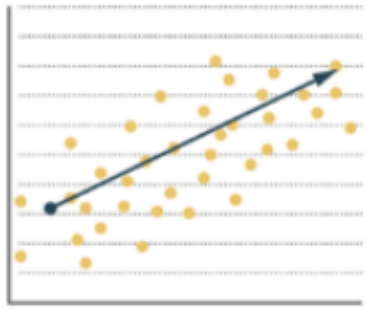
At this stage, we completed the Exploratory Data Analysis (EDA) process by identifying correlations between features and the target variable.
We utilized visualization tools such as Seaborn and Matplotlib to gain insights into the data.
Additionally, feature engineering was performed by converting categorical variables into dummy variables to prepare the dataset for machine learning models.

Github repository: he completed EDA with data visualization notebook can be accessed here Link

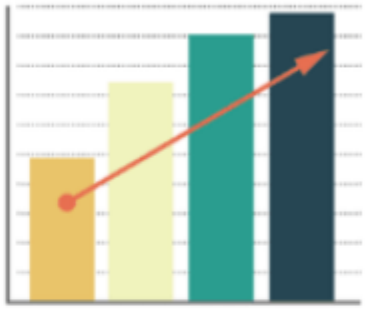
1- Visualization of the relationship between Flight Number and Launch site



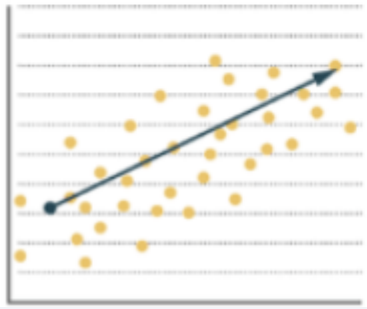
2- Visualization of the relationship between Payload and Launch Site



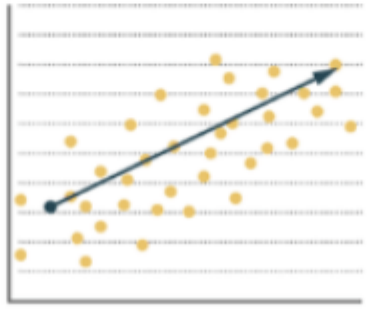
3- Visualization of the relationship between Success Rate for each Orbit Type



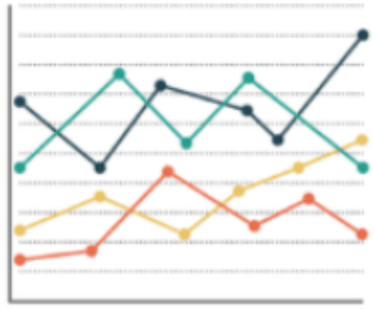
4- Visualization of the relationship between Flight Number and Orbit Type



5- Visualization of the relationship between Payload and Orbit type



6- Visualization of the Launch Success yearly trend



EDA with SQL

SQL Queries

- Display the names of the unique launch sites in the space mission.
- **Display 5 records where launch sites begin with the string 'CCA'.**
- Display the total payload mass carried by boosters launched by NASA (CRS
- **List the date when the first successful landing outcome in ground pad was achieved.**
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- **List the total number of successful and failure mission outcomes**
- List the names of the booster versions which have carried the maximum payload mass . Use a subquery
- **List the failed landing outcomes in drone ship, their booster versions , and launch site names for the in year 2015.**
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010 06 04 and 2017 03 20, in descending order.

Github repository: The completed EDA with SQL notebook can be accessed here

Build an Interactive Map with Folium

Process Overview

We used Folium library to represent our analysis as geospatial data, we do this by drawing markers like circles and lines on an interactive map.

Launch Site Markers

We started by drawing circles on the map for the four Falcon 9 rocket launch sites. The coordinates of 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 added to these locations to indicate whether the first stage of the rocket successfully returned or failed.

Distance Calculation

We calculated the distances from CCAFS LC-40 to three key locations:

- The closest city.
- The coastline.
- A nearby highway.

Using this information, we drew polylines on the map to visually represent these distances.

Github repository: The completed interactive map with Folium map notebook can be accessed here

Build a Dashboard with Plotly Dash

Process Overview

We used Plotly library to represent our analysis in an interactive dashboard. These are the tasks we perform

- We added a dropdown list to enable the Launch site selection including the following options: “All Sites”, “CCAFS LC 40”. “CCAFS SLC-40”. “VAFB SLC-4E” and “KSC LC-39A
- We added a pie chart to show the total successful launches count for all sites
- We added a slider to select payload which ranges from 0-10000
- Finally we added a scatter chart to show the correlation between payload and launch success

Github repository: The completed notebook with interactive dashboard rutines are in the following link

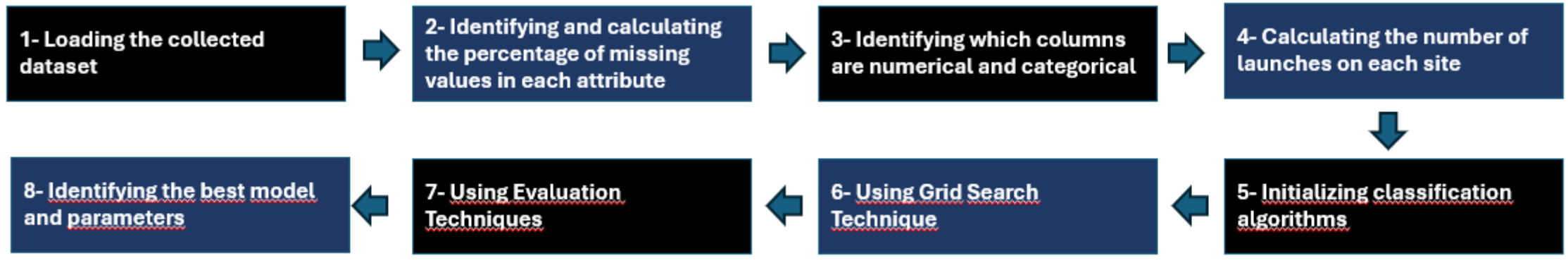
Predictive Analysis (Classification)

Machine Learning Stages

Import Required Libraries: Begin by importing the essential libraries needed for data manipulation and model building.
Load Cleaned Data: Load the pre-processed dataset to ensure that the data is ready for analysis.
Standardize the Data: Apply standardization techniques to normalize the data and minimize bias in model training.
Split the Data: Divide the dataset into training (80%) and testing (20%) subsets to evaluate model performance effectively.
Initialize Classification Algorithms: Set up four distinct classification algorithms for comparison: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT) and K-Nearest Neighbors (KNN)
Optimize Hyperparameters with Grid Search: Utilize Grid Search to systematically explore different parameter combinations and identify the best-performing model.
Evaluate Model Performance: Assess the models using various evaluation metrics, including: Confusion Matrix, F1 Score and Jaccard Score

Additional Resources

Explore the complete Predictive Analysis Lab implementation on GitHub: Link



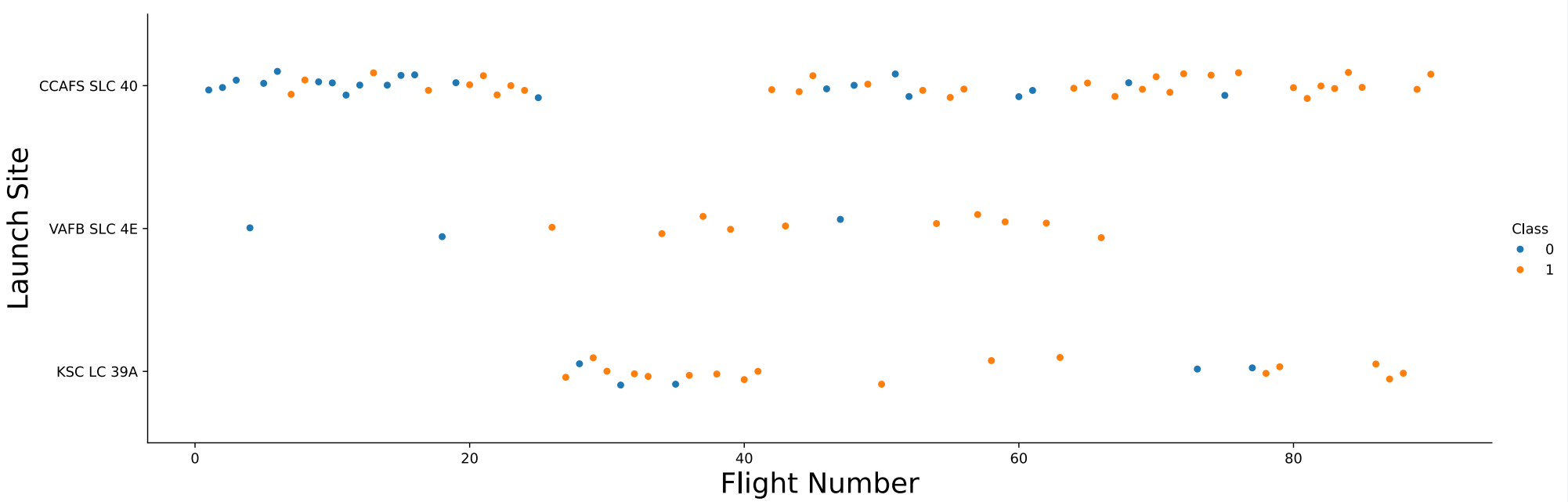
Results

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

Section 2

Insights drawn from EDA

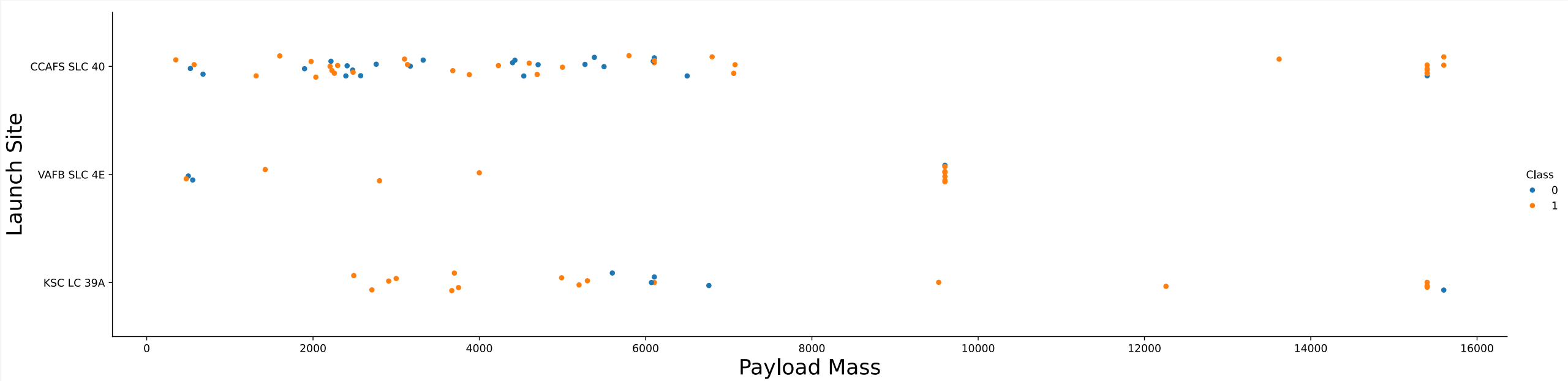
Flight Number vs. Launch Site



Comments

CCAFS SLC 40: This is the most frequently used site for launching SpaceX rockets, having conducted 55 trials. Out of these, 33 were successful, while 22 resulted in failure, yielding a 60% success rate.
VAFB SLC 4E: This site is the least utilized for SpaceX rocket launches, with a total of 13 trials. Of these, 10 were successful and 3 failed, resulting in a 77% success rate.
KSC LC 39A: This site has a moderate performance in launching SpaceX rockets, with 22 trials conducted. Among these, 17 were successful and 5 failed, also achieving a 77% success rate.

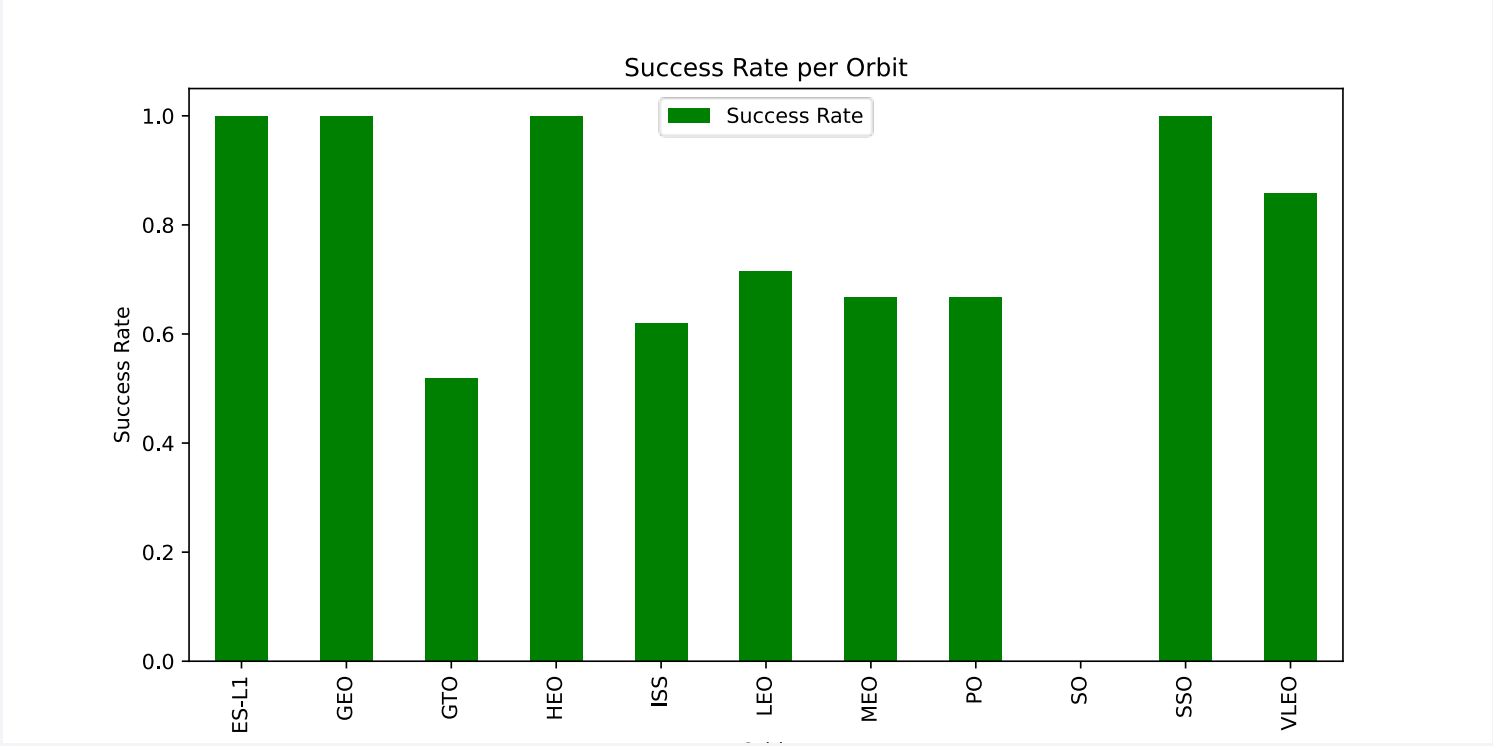
Payload vs. Launch Site



Comments

According to the plot above, there is no significant relationship between payload mass and the success of the first stage return. This is evidenced by the approximately equal number of failed and successful trials.

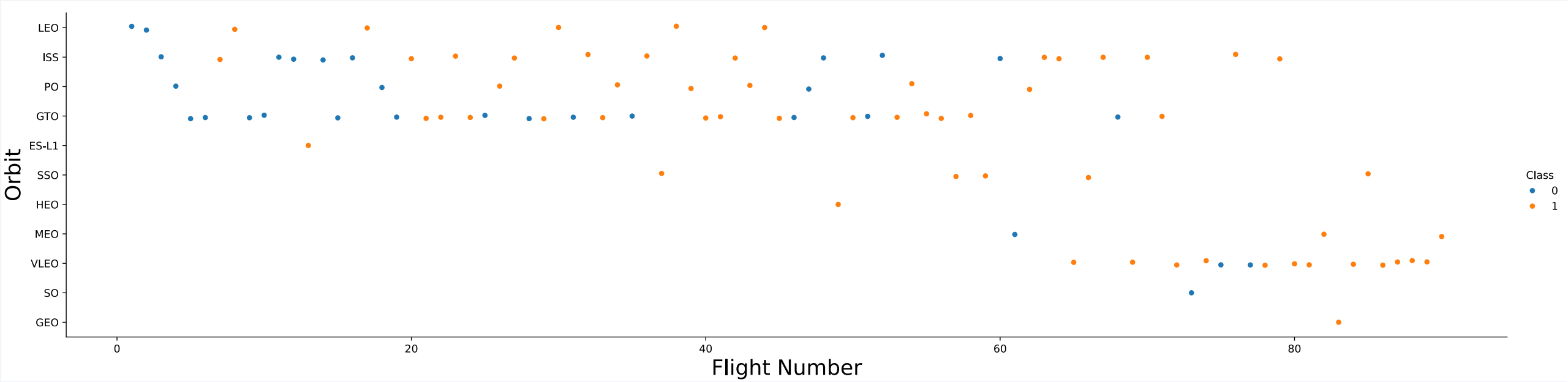
Success Rate vs. Orbit Type



Comments

According to the bar plot, the most successful orbits for first stage returns are ES L1, GEO, HEO, and SSO. In contrast, the worst-performing orbit is GTO. It is important to analyze the reasons behind GTO's poor performance in order to mitigate failures in first stage returns.

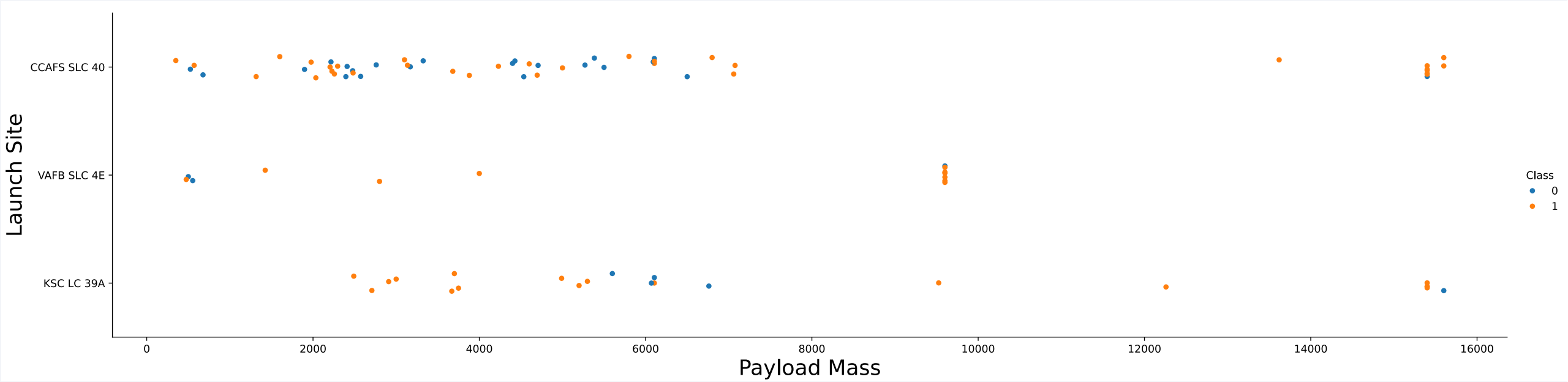
Flight Number vs. Orbit Type



Comments

We observe that in the LEO orbit, success appears to be correlated with the number of flights. In contrast, there seems to be no discernible relationship between the number of flights and success in the GTO orbit.

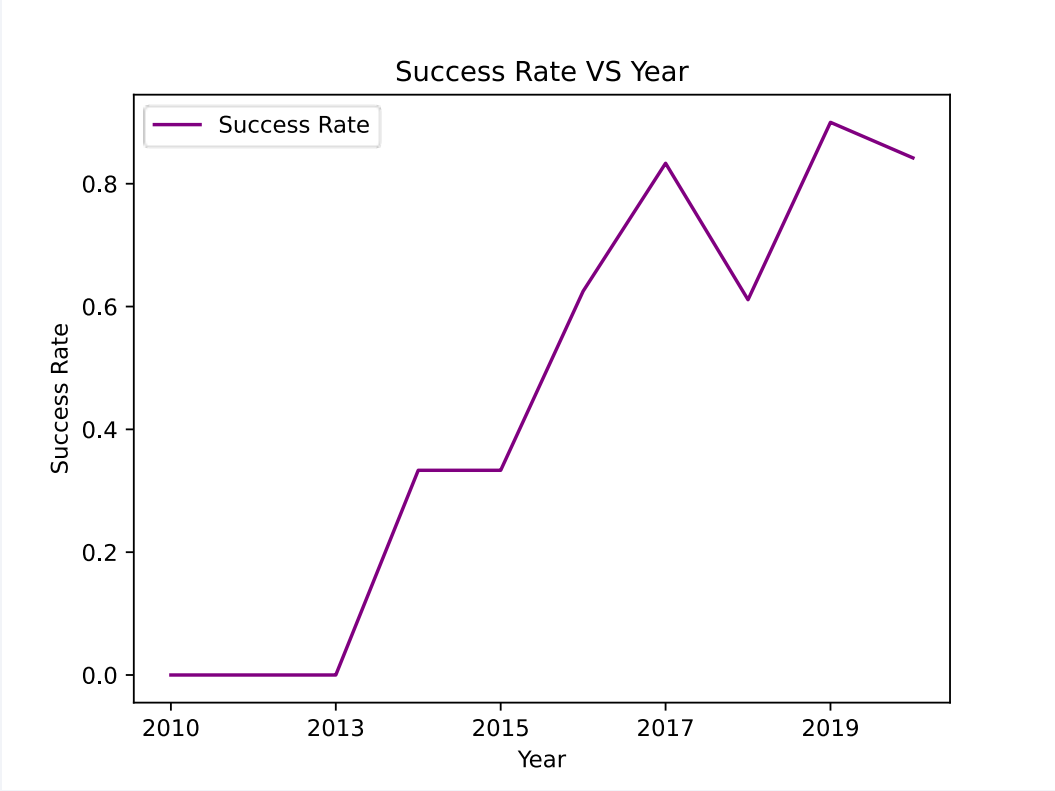
Payload vs. Orbit Type



Comments

We observe that heavy payloads have a negative influence on GTO orbits but a positive impact on Polar, LEO, and ISS orbits. In these latter orbits, the successful landing rate is higher with heavy payloads. However, in GTO orbits, it is difficult to distinguish between successful and unsuccessful landings, as both outcomes occur.

Launch Success Yearly Trend



Comments

We can observe that the success rate has consistently increased from 2013 until 2020.

All Launch Site Names

To find the names of the unique launch sites we use the following query

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

Giving the output:

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

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

As we can see, there are 4 sites for rockets launches.

Launch Site Names Begin with 'CCA'

To find the 5 records where launch sites begin with `CCA` we used the query:

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

With 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

We can see that only one of the five launches sites have customer SpaceX

Total Payload Mass

To calculate the total payload carried by boosters from NASA, we use the query:

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

Which give us the following result:

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

The total amount of payload transported to outer space by NASA using SpaceX rockets is 45,596 kg, which is equivalent to approximately 50.261 US tons.

Average Payload Mass by F9 v1.1

To calculate the average payload mass carried by booster version F9 v1.1, we used the query:

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

With result:

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

The average payload mass carried by the Falcon 9 booster version v1.1 is 2928,4 kg

First Successful Ground Landing Date

To find the dates of the first successful landing outcome on ground pad we used the following query:

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

The result of the query is:

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

The first successful landing of a Falcon 9 rocket on a ground pad occurred on December 22, 2015. This milestone was achieved during the Falcon 9 flight 20 mission, marking a significant advancement in reusable rocket technology for SpaceX. The rocket's first stage successfully returned to Earth and touched down at Cape Canaveral, demonstrating the feasibility of controlled landings after orbital launches

Successful Drone Ship Landing with Payload between 4000 and 6000

To find the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, we use the following query

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

Which give us:

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

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

This shows that only four boosters have success in drone ship landing with payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

To calculate the total number of successful and failure mission outcomes, we use the following query

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

Which give us the following result:

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

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

It is clear that the success rate of mission outcomes is overwhelmingly positive, with only 1 failed mission compared to 99 successful ones. This demonstrates a remarkable track record of success.

Boosters Carried Maximum Payload

To show the list of names for which the booster have carried the maximum payload mass, we use the following query:

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

Giving us the following result:

```
 * 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 booster versions that carry the maximum payload start with F9 B5 and range from B1048 to B1060. These boosters are part of the Falcon 9 Block 5 series, which has been designed for enhanced performance and reusability, allowing for significant payload capacities to Low Earth Orbit (LEO) and beyond.

2015 Launch Records

Now we give the list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, with the following query:

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

Giving the following list

```
 * 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

There were two failed landings in 2015 on a drone ship, both occurring at the same site, CCAFS LC 40, and involving the same booster version, F9 v1.1. These attempts highlight the challenges SpaceX faced in perfecting their landing techniques during that period.

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

Finally, we rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order, with the following query:

```
%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
```

Giving us the list:

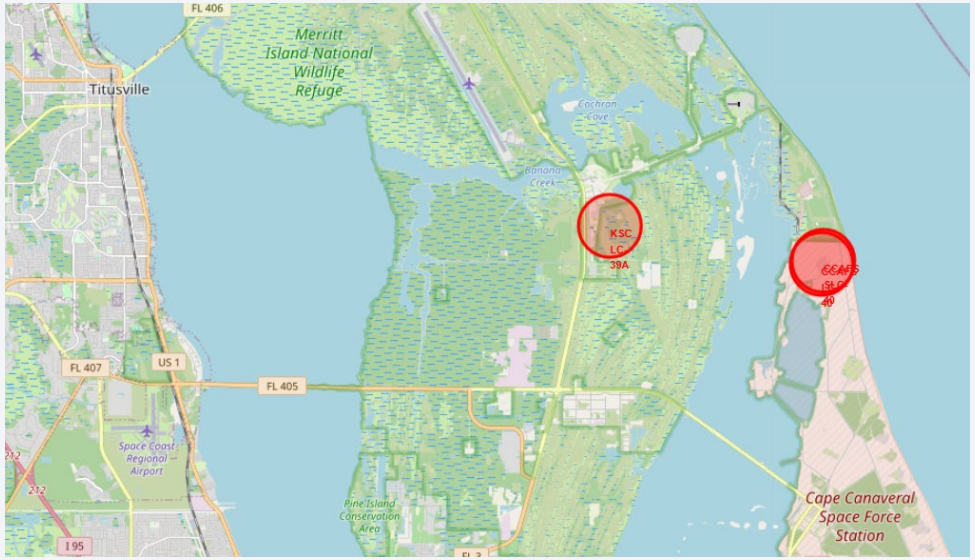
```
 * sqlite:///my_data1.db
Done.
```

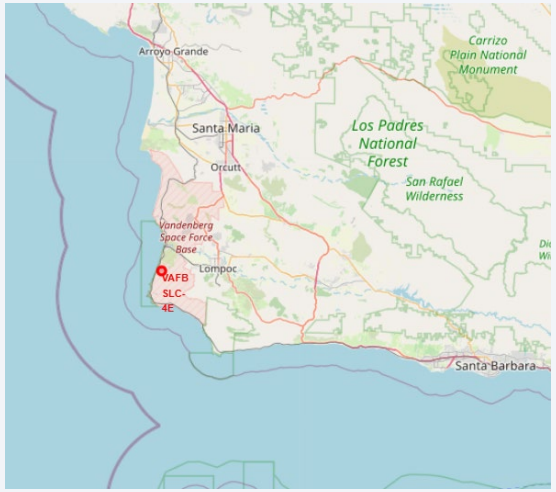
Landing_Outcome	counts_of_landing_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The landing outcomes from June 4, 2010, to March 20, 2017, show significant progress in SpaceX's recovery technology, with equal successes and failures on drone ships, three successful ground pad landings, and numerous missions where no landings were attempted.


Section 3
Launch Sites
Proximities Analysis

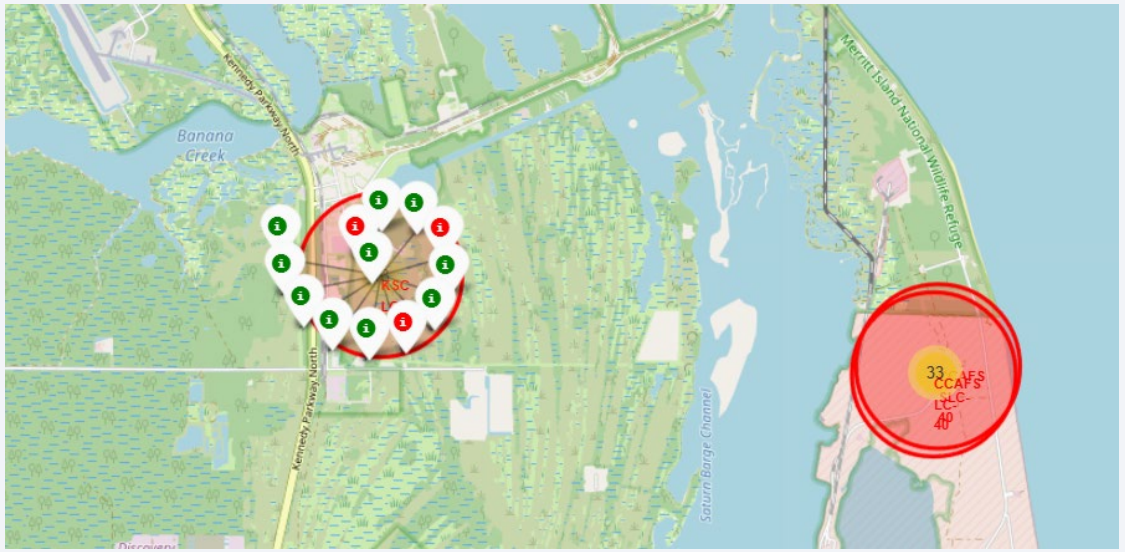
Folium Map: Launch sites overview



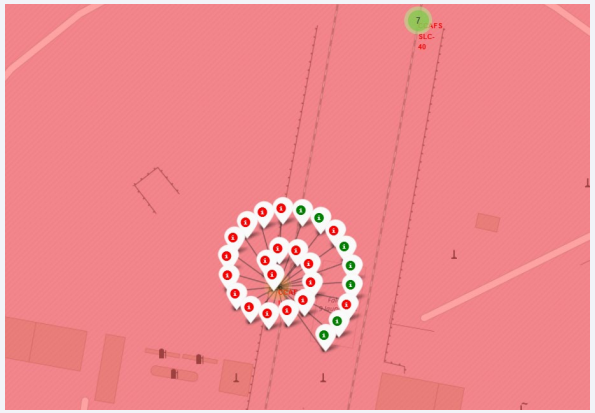


All SpaceX launch sites are strategically located near the coast and the equator, prioritizing proximity to water and the zero latitude line to minimize risks during launches. The launch facilities are distributed across two states: California and Florida.

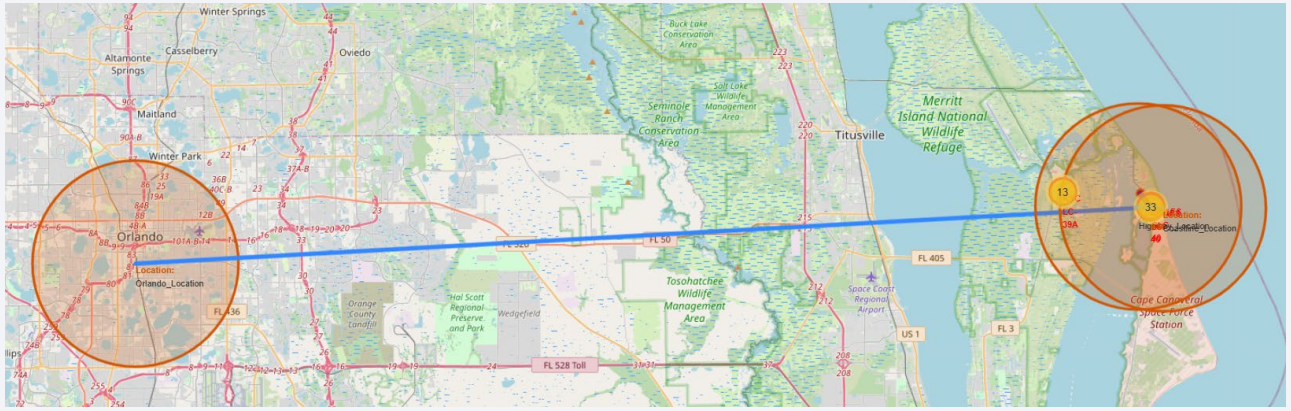
Folium Map: Success rate for each launch location

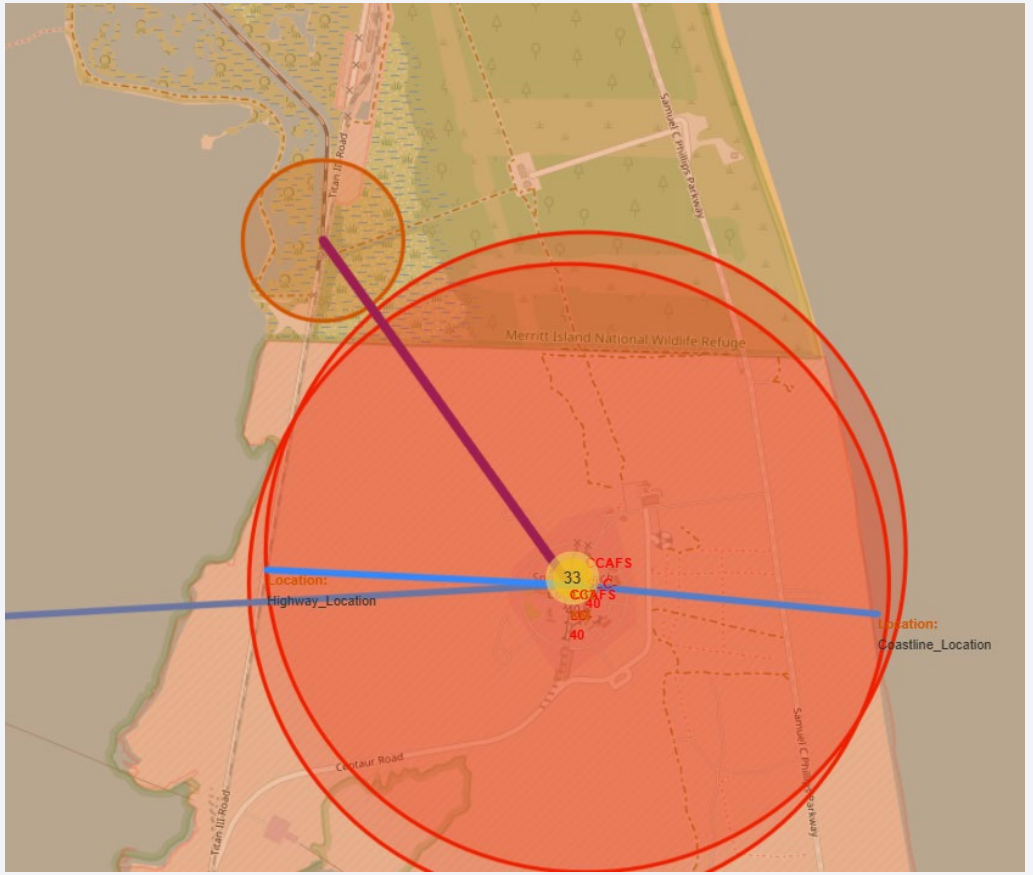


From the color-coded markers in the marker clusters, we can easily identify which launch sites have relatively high success rates. A green marker indicates a successful return, while a red marker signifies a failed return.



Folium Map: Closest proximities to CCAFS LC-40

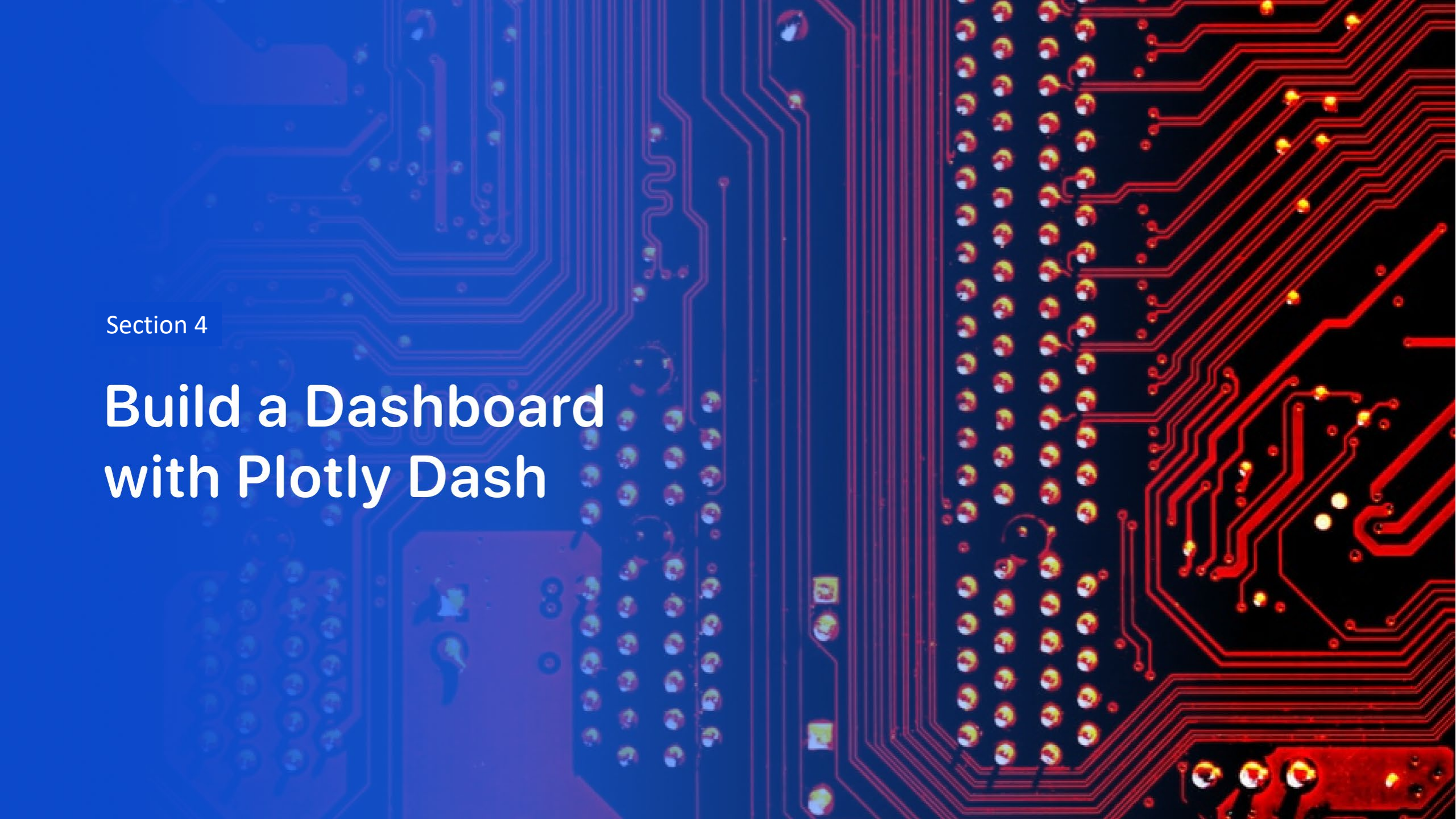




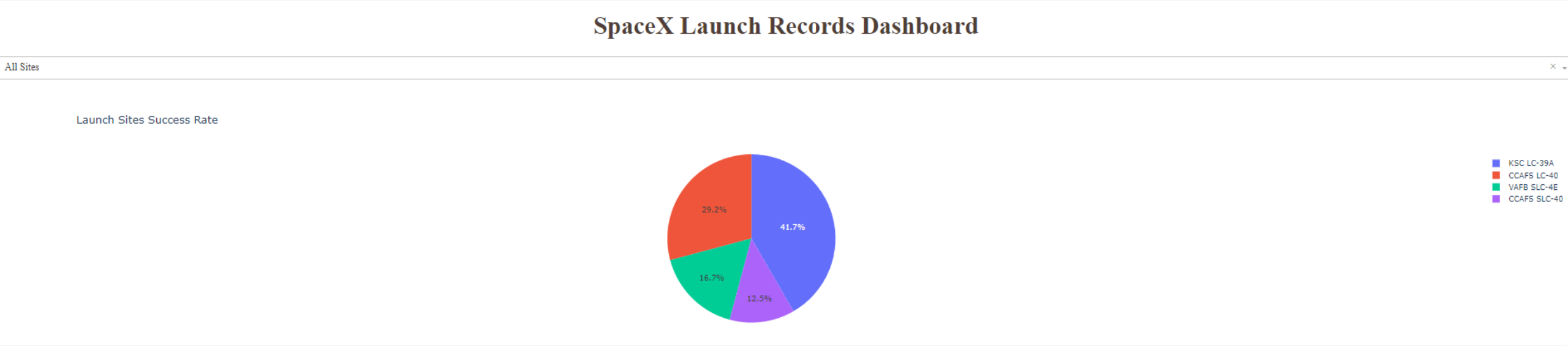
Proximities Coordinates:

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

According to this, we can say that launch sites keep distance between cities, around 78.8km to Orlando. But from coastline and highway is not really far (less than one kilometer).

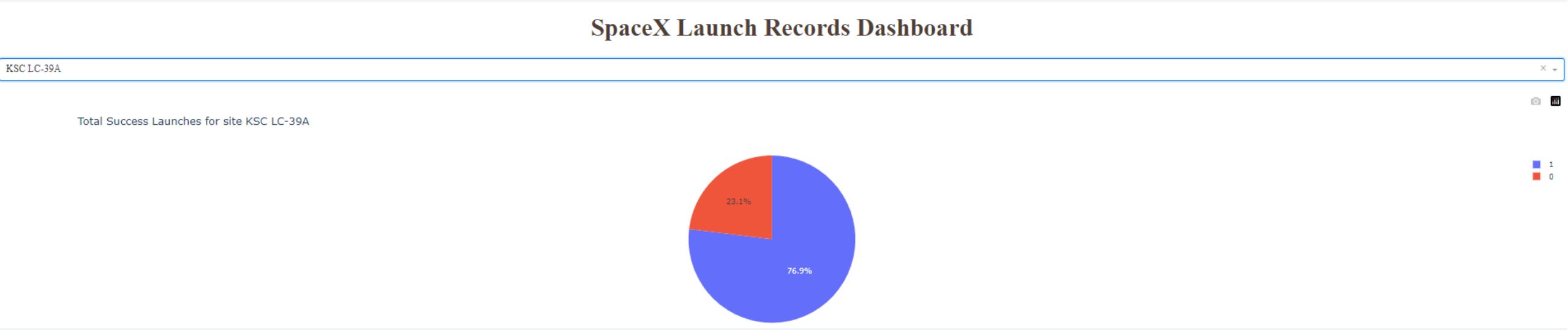

Section 4
Build a Dashboard with Plotly Dash

Dashboard: Launch success count for all sites



The graph indicates the success percentage for first stage returns at various launch sites. The top-performing site is KSC LC 39A, achieving a 41.7% success rate, while CCAFS SLC 40 has the lowest success rate at 12.5%. This disparity highlights the varying effectiveness of different launch locations in facilitating successful rocket recoveries.

Dashboard: Launch success for KSC LC-39A

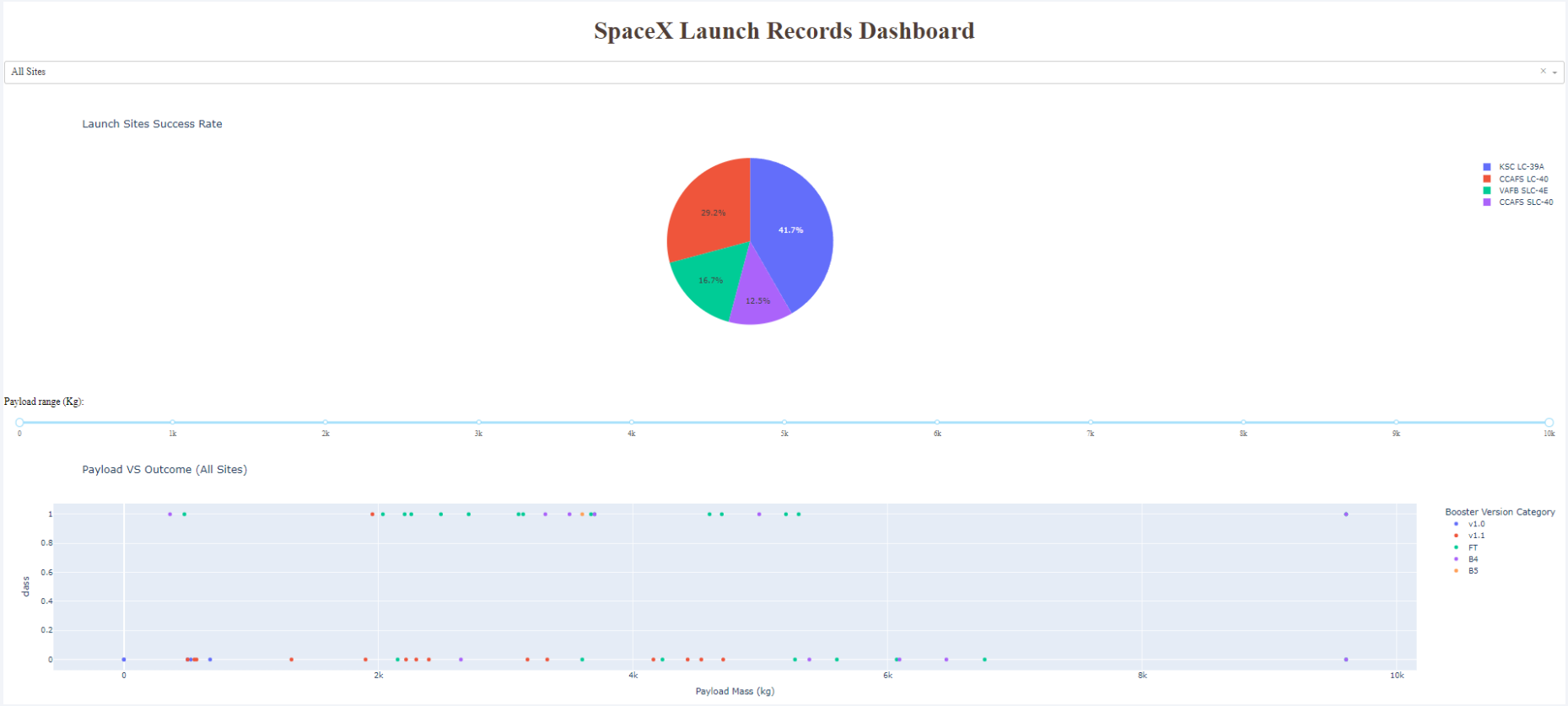


The total success launches for KSC LC 39A are as follows:

- Successful Missions: 76.9%
- Failed Missions: 23.1%

This indicates a strong performance for this launch site, showcasing its reliability in conducting successful missions. The high success rate reflects effective operational practices and technological advancements achieved by SpaceX at this historic launch complex.

Dashboard: Payload vs. Launch Outcome scatter plot for all sites

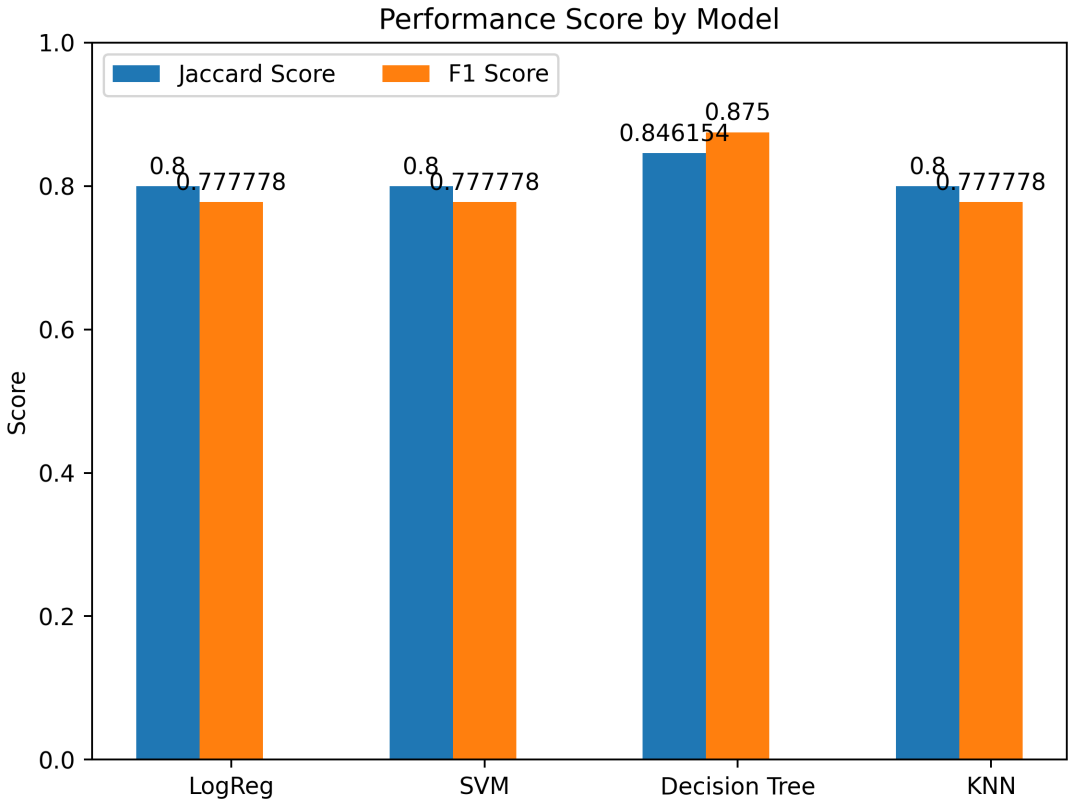


This interactive scatter plot illustrates the launch outcomes in relation to payload mass. It indicates that for payload masses under 4,000 kg, the likelihood of a successful launch is significantly higher. Additionally, the plot allows for analysis of success rates across different booster versions, providing valuable insights into performance trends based on payload capacity, as we can change the range of payload mass interval as we want.

Section 5

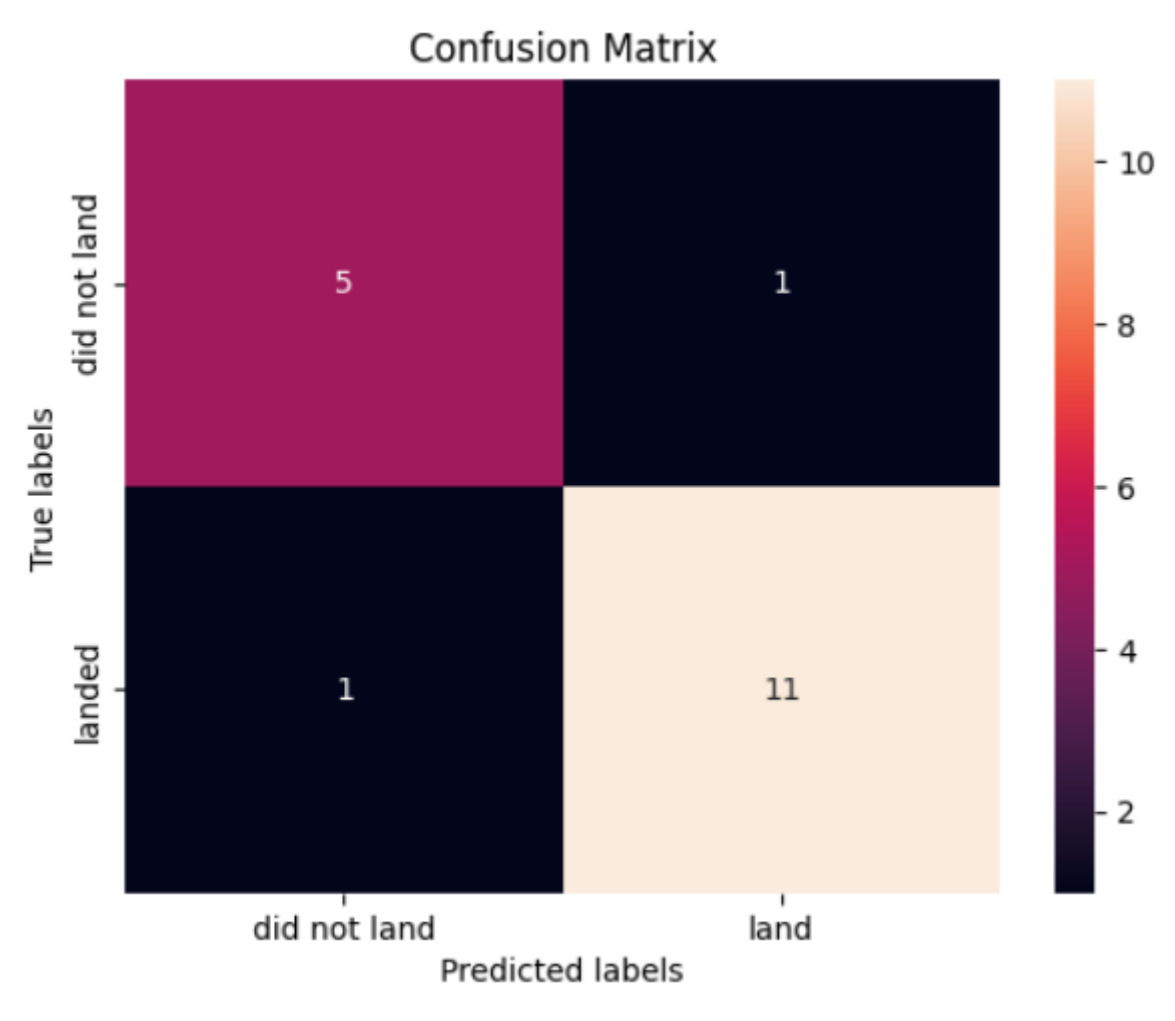
Predictive Analysis (Classification)

Classification Accuracy



The Decision Tree model outperformed the others with the highest Jaccard Score (0.85) and F1 Score (0.88), indicating better precision and recall in classification. Logistic Regression, SVM, and KNN all shared similar performance metrics, with Jaccard Scores of 0.8 and F1 Scores around 0.78, suggesting comparable effectiveness in their classifications.

Confusion Matrix



The Decision Tree model demonstrates strong performance with a total of 11 true positives and 5 true negatives, leading to high accuracy in predicting both successful and failed landings. The low counts of false positives (1) and false negatives (1) suggest that the model is reliable, with only minor misclassifications. Overall, this confusion matrix indicates that the Decision Tree model is effective in distinguishing between successful and unsuccessful rocket landings.

Conclusions

- Reusable Technology's Role: SpaceX's emphasis on reusable first-stage rockets drastically reduces costs, providing a competitive advantage. A similar focus is recommended for Space Y.
- Key Predictors of Success: Variables such as orbit type and payload mass significantly influence landing outcomes. For instance, LEO orbits show higher success rates compared to GTO orbits.
- Site-Specific Insights: Launch sites like KSC LC-39A demonstrate higher reliability, suggesting the importance of strategic location planning for future missions.
- Data-Driven Decision Making: The integration of advanced data analytics and machine learning can optimize mission success rates and reduce operational risks.
- Continuous Improvement: Regular analysis of historical performance and adoption of emerging technologies will ensure sustained competitiveness in this dynamic sector.

Appendix

- Github's webpage of the overall project: Link
- SpaceX Static Wikipedia URL: Link
- SpaceX data used in ML training: Link

Thank you!