IBM DATA SCIENCE CAPSTONE PROJECT

ARMANDO IBARRA MARQUUEZ

30/10/24

EXECUTIVE SUMMARY

This capstone project aims to apply advanced data science methodologies to support decision-making for a private space launch company. As a data scientist, we will start by gathering relevant data from multiple sources, followed by data wrangling to ensure data quality and consistency for effective analysis.

The project methodology includes several key phases:

- Data Collection to gather comprehensive and relevant information.
- Data Wrangling to refine the raw data for accuracy and usability.
- Exploratory Data Analysis (EDA) using both data visualization and SQL queries to uncover initial insights and relationships between variables. This analysis will further segment data by categorical variables, providing a detailed understanding of the factors influencing space launch contexts.
- Interactive Analytics through tools like Folium to create dynamic maps and Plotly Dash to design an interactive dashboard, enhancing the visualization of complex data relationships.
- Predictive Analysis using classification models to generate actionable insights, anticipating trends that could inform strategic decisions.

INTRODUCTION

In this project, we will explore the commercial space industry with a focus on SpaceX's innovative approach and achievements. With companies like Virgin Galactic, Rocket Lab, and Blue Origin offering suborbital flights, small satellite services, and reusable rockets, SpaceX has emerged as a leader by significantly lowering launch costs through its rocket reusability, particularly with the Falcon 9. This has enabled SpaceX to offer launches at $62 million compared to the industry average of $165 million.

We will examine the stages of the Falcon 9 rocket, focusing on the first stage, which does most of the work during a launch and represents a substantial investment. This stage is often recovered and reused, a unique feature that has helped SpaceX cut costs. However, recovery is not always possible, as some missions require sacrificing the first stage due to specific payload, orbit, or customer requirements.

In this capstone project, we will take on the role of data scientists for a new space company, "Space Y," aiming to compete with SpaceX. Our work will involve estimating the cost of each launch, analyzing factors that affect the success of first-stage landings, and using machine learning to predict whether the first stage will be recovered. Rather than relying solely on rocket science, we will leverage data analysis and public information to build dashboards and predictive models, helping to forecast the reusability of SpaceX's first stage and supporting Space Y's competitive strategy.

METHODOLOGY

In this course, we will start by gathering data through the SpaceX REST API and web scraping from Wikipedia, assembling relevant information for our analysis.
Following data collection, we'll conduct data wrangling, including filtering, handling missing values, and applying One-Hot Encoding to prepare the dataset for binary classification tasks.
Our next step will involve Exploratory Data Analysis (EDA), using both data visualization techniques and SQL queries to uncover initial patterns and insights within the data. To enhance our analysis, we will create interactive visualizations with Folium and build a dashboard with Plotly Dash, providing an engaging way to explore the data.
Finally, we'll perform predictive analysis using classification models. This includes building, tuning, and evaluating models to achieve optimal accuracy and reliable results.

DATA COLLECTION

In this project, we gather data on SpaceX launches primarily through the SpaceX REST API and web scraping from Wikipedia. The SpaceX API provides launch data, including details on rockets, payloads, launch specifics, and landing outcomes, which will support predictions on landing attempts. By targeting specific API endpoints (e.g., /launches/past), we retrieve historical launch data in JSON format, which is then converted to a structured table using json_normalize.
Additionally, we use the BeautifulSoup library to scrape launch data tables from Wikipedia, which we convert into Pandas DataFrames. Data wrangling steps include sampling only Falcon 9 launches, handling missing values—such as calculating the mean for PayloadMass nulls—and resolving data with identification codes by querying further API endpoints for detailed information. The final data is stored in a clean dataset, ready for analysis.

DATA WRANGLING

Data wrangling is a crucial process in data analysis that involves cleaning and transforming raw data into a structured format suitable for analysis. In this project, we focus on wrangling SpaceX launch data to prepare it for further exploration and modeling.

We begin by examining key attributes of the dataset, including Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, and Launch Outcome. Each attribute provides valuable information:

- Launch Site indicates the different locations where launches occur, such as Vandenberg AFB and Kennedy Space Center.
- Orbit describes the paths taken by payloads, including Low Earth Orbit (LEO) and Geosynchronous Transfer Orbit (GTO).
- Outcome reveals the success of the first-stage landings, with various classifications representing different landing scenarios (e.g., successfully landing on a drone ship or failing to land).

To facilitate analysis, we convert the "Outcome" attribute into a binary classification. We label successful landings as 1 and unsuccessful ones as 0, creating a new variable, Y, that serves as our classification label for training supervised models.

During the data wrangling process, we also handle specific cases where landings were attempted but not successful. For example, outcomes such as "True ASDS" indicate a successful landing on a drone ship, while "False ASDS" signifies an unsuccessful attempt. By categorizing these outcomes into training labels, we ensure that our dataset is well-prepared for exploratory data analysis (EDA) and subsequent predictive modeling. This thorough data wrangling lays the foundation for deeper insights and more effective analysis in our project.

EDA WITH DATA VISUALIZATION

Exploratory Data Analysis (EDA) is an essential first step in any data science project, allowing us to investigate and summarize the main characteristics of a dataset. In this project, we will conduct EDA to analyze SpaceX launch data, focusing on identifying patterns and relationships that can inform our predictive modeling efforts.

During our initial analysis, we will assess whether the data can help us automatically determine if the Falcon 9's first stage will successfully land. We will examine various attributes that may influence landing outcomes, including launch number, which provides insight into the historical success rate since 2013.

Our findings indicate that different launch sites exhibit varying success rates; for instance, CCAFS LC-40 has a success rate of 60%, while KSC LC-39A and VAFB SLC 4E achieve around 77%.

Moreover, we will explore how combining multiple features enhances our understanding of the data. For example, while CCAFS LC-40 shows a 60% success rate overall, its success rate rises to 100% for payloads exceeding 10,000 kg. This kind of analysis will allow us to identify attributes that are correlated with successful landings.

To prepare the data for machine learning, we will apply one-hot encoding to categorical variables, transforming them into a format suitable for model training. By doing so, we aim to develop a predictive model that can accurately assess the likelihood of the first stage successfully landing.

SpaceX's achievements, such as being the only private company to return a spacecraft from low Earth orbit and offering competitive launch prices (e.g., $62 million for a Falcon 9 rocket), further emphasize the importance of understanding these dynamics. This information not only enhances our predictive capabilities but also provides valuable insights for any alternative companies looking to compete with SpaceX in the rocket launch market.

Overall, EDA will help us derive meaningful insights from the dataset, setting the stage for effective predictive modeling and a deeper understanding of the factors that influence SpaceX's launch success.

PREDICTIVE ANALYSIS

In this segment, we will focus on Predictive Analysis, where we will develop a machine learning pipeline to determine whether the first stage of the Falcon 9 rocket successfully lands. This analysis is critical, as understanding the likelihood of a successful landing can help inform pricing strategies and competitive bidding for rocket launches.

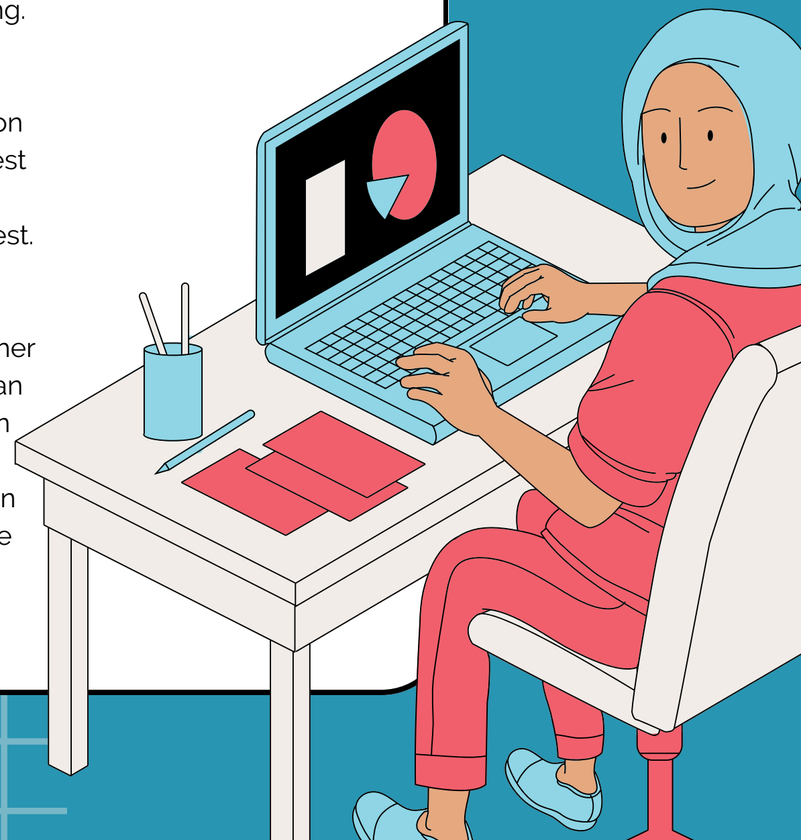
The process begins with data preprocessing, which will standardize our dataset and ensure that it is ready for model training. We will then implement train-test splitting to divide the data into training and testing sets, allowing us to evaluate the performance of our models effectively.

Next, we will train several machine learning models, including Logistic Regression, Support Vector Machines (SVM), Decision Tree Classifier, and K-Nearest Neighbors (KNN). To optimize their performance, we will utilize Grid Search to identify the best hyperparameters for each algorithm. This step is essential for maximizing the accuracy of our predictive models.

Once the models are trained, we will assess their accuracy using the training data and determine which model performs best. The ultimate goal is to produce a confusion matrix that provides a clear representation of each model's performance in predicting landing outcomes.

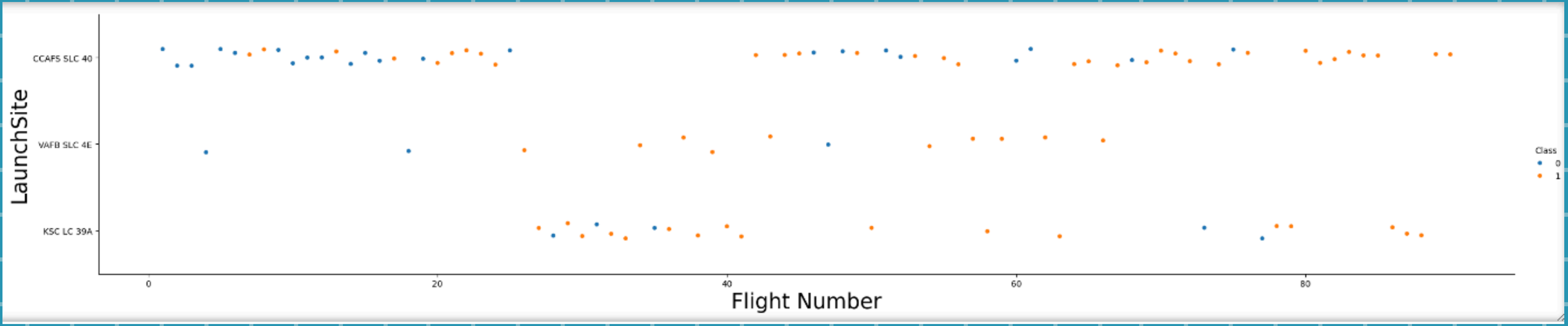
SpaceX's ability to offer competitive launch costs—$62 million for a Falcon 9 rocket compared to over $165 million from other providers—highlights the significance of this analysis. If we can accurately predict the success of first-stage landings, we can gain insights into the overall cost of a launch, which is crucial for any alternative companies looking to bid against SpaceX in the aerospace market.

By the end of this lab, we will have established a robust machine learning pipeline capable of predicting whether the Falcon 9's first stage will land successfully, using the comprehensive data collected and analyzed in previous steps. This predictive capability will not only enhance our understanding of rocket landings but also support strategic decision-making in the competitive landscape of space travel.



EDA WITH VISUALIZATION

FLIGHT NUMBER VS LAUNCH SITE



This scatter plot shows space launch data across three different launch sites:

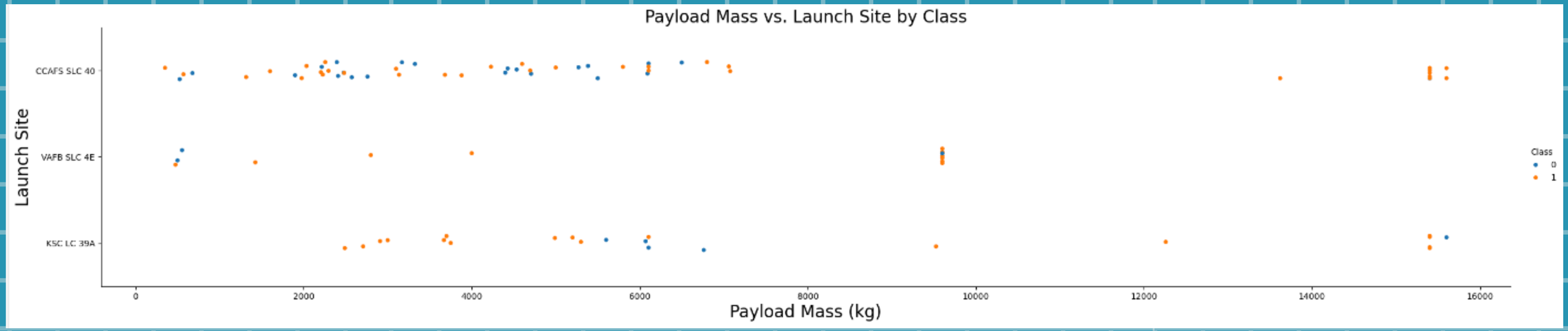
- CCAFS SLC-40
- VAFB SLC-4E
- KSC LC-39A

The x-axis shows flight numbers (0-80), while points are color-coded into two classes (blue and orange). Each point represents a distinct launch, with the launch site location on the y-axis.

The distribution suggests varying patterns of usage across these launch facilities over the sequence of flights.

EDA WITH VISUALIZATION

PAYLOAD VS LAUNCH SITE



This scatter plot shows the relationship between payload mass (x-axis, measured in kg) and launch sites (y-axis), similar to the previous graph:

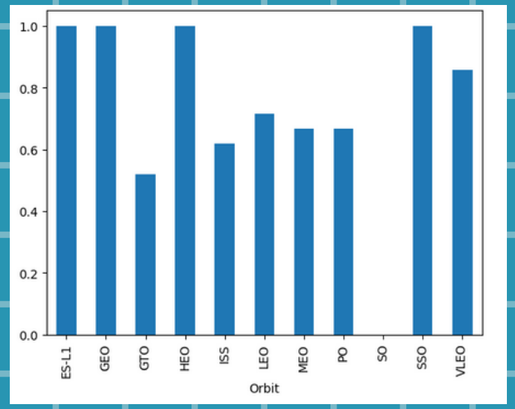
- CCAFS SLC-40
- VAFB SLC-4E
- KSC LC-39A

The data points are classified into two categories (blue and orange), with payload masses ranging from 0 to 16,000 kg. CCAFS SLC-40 appears to have the most launches, while VAFB SLC-4E has fewer launches but handles various payload masses.

The title explicitly states "Payload Mass vs. Launch Site by Class", making clear the variables being compared.

EDA WITH VISUALIZATION

SUCCESS RATE OF EACH ORBIT TYPE



This bar chart shows success rates for different orbital destinations/missions, where 1.0 represents 100% success. From the data:

Highest Success Rates (100%):

- ES-L1 (Earth-Sun Lagrange Point 1)
- GEO (Geostationary Earth Orbit)
- HEO (Highly Elliptical Orbit)
- VELO (likely Velox mission)

Medium Success Rates (60-80%):

- ISS (International Space Station)
- LEO (Low Earth Orbit)
- MEO (Medium Earth Orbit)
- PO (Polar Orbit)
- SO (Sun-synchronous Orbit)

Lowest Success Rate:

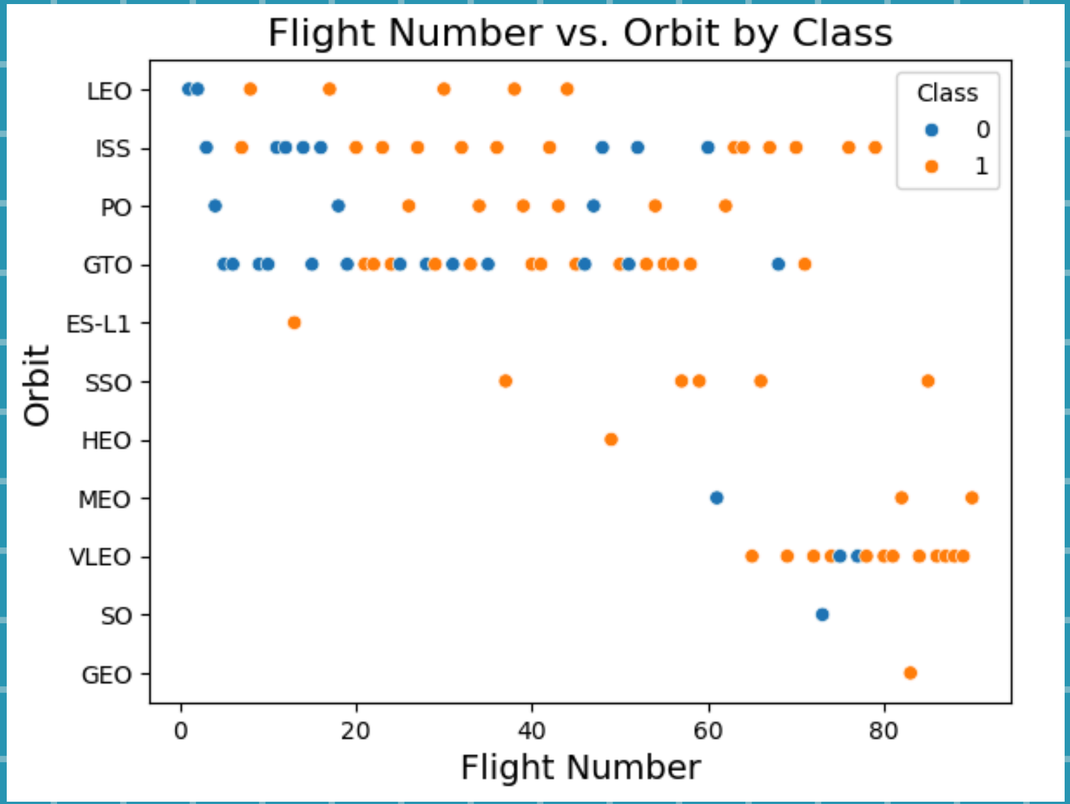
- GTO (Geosynchronous Transfer Orbit) at around 50%

This visualization helps identify which orbital missions have been historically more successful and which ones have faced more challenges.

EDA WITH VISUALIZATION

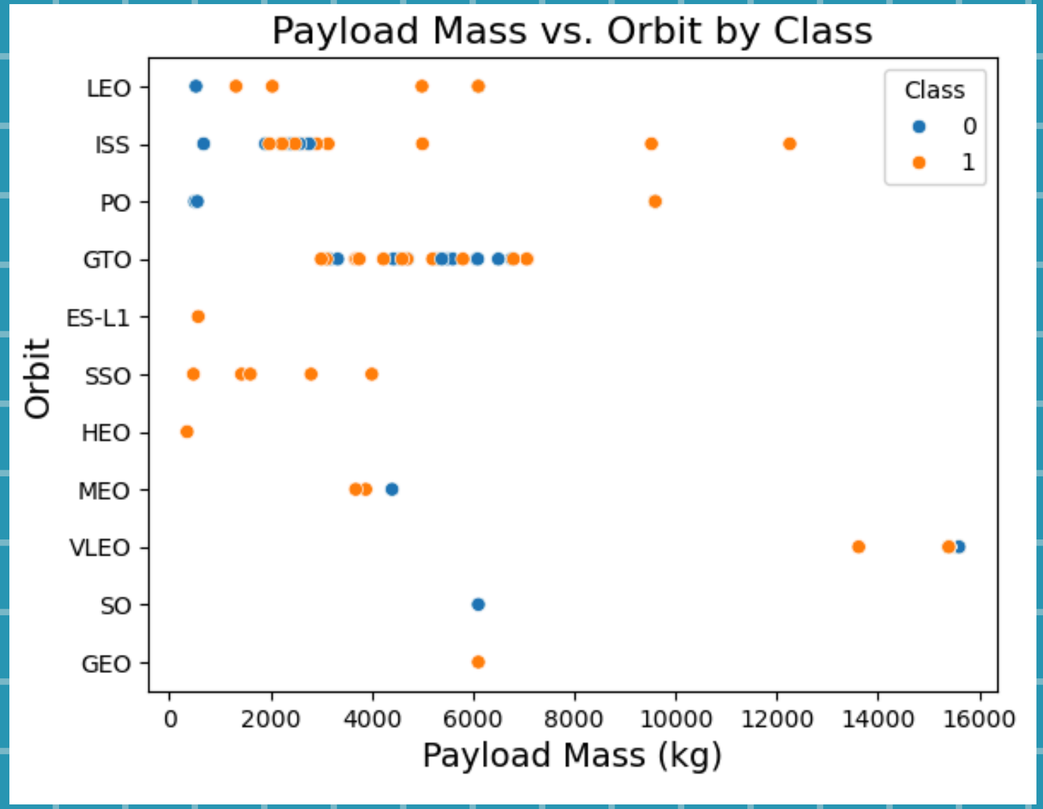
FLIGHT NUMBER VS. ORBIT TYPE

This scatter plot shows "Flight Number vs. Orbit by Class" where:
The y-axis displays different orbit types:
LEO (Low Earth Orbit)
ISS (International Space Station)
PO (Polar Orbit)
GTO (Geosynchronous Transfer Orbit)
ES-L1 (Earth-Sun Lagrange Point 1)
SSO (Sun-Synchronous Orbit)
HEO (Highly Elliptical Orbit)
MEO (Medium Earth Orbit)
VLEO (Very Low Earth Orbit)
SO (Sun-synchronous Orbit)
GEO (Geostationary Earth Orbit)
Points are colored by two classes (blue: 0, orange: 1), and the x-axis shows flight numbers from 0 to about 90.



EDA WITH VISUALIZATION

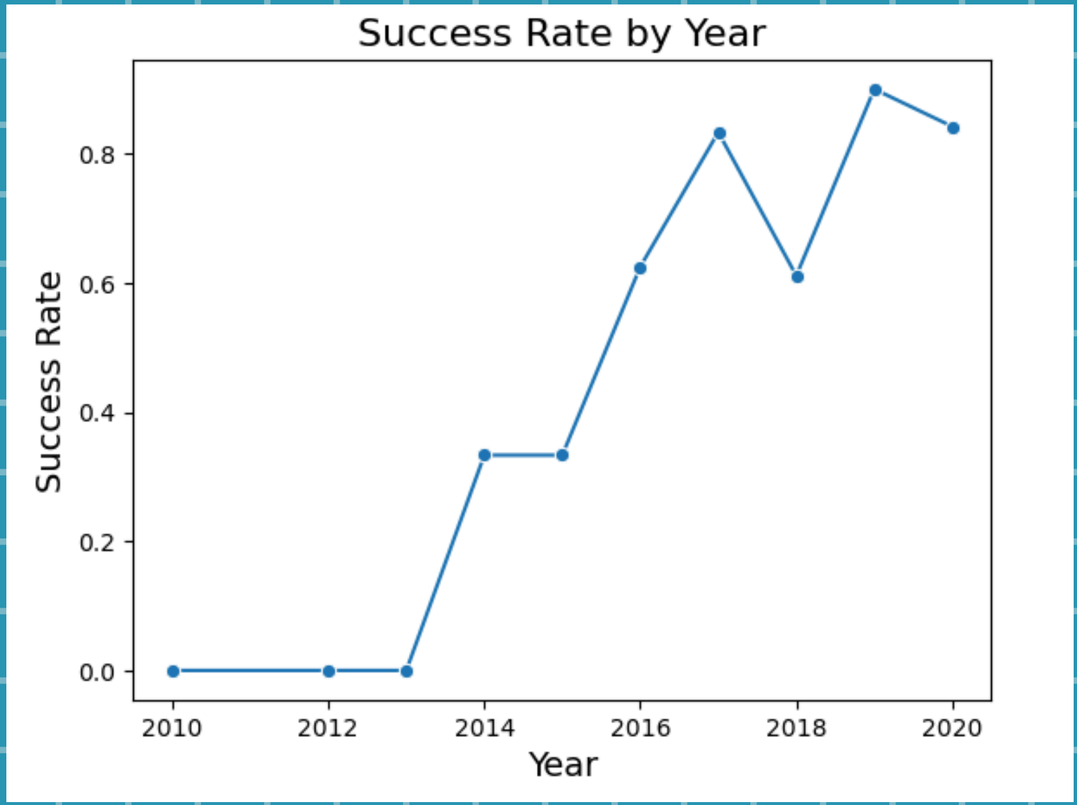
PAYLOAD MASS VS. ORBIT TYPE



This scatter plot shows "Payload Mass vs. Orbit by Class" where:
The y-axis shows the same orbit types as the previous graph (LEO, ISS, PO, GTO, etc.), but now the x-axis displays payload mass in kilograms (0-16,000 kg).

EDA WITH VISUALIZATION

SUCCESS RATE BY YEAR



Overall, the graph shows a general upward trend in success rates over the decade, with the most dramatic improvements occurring between 2014-2017, and maintaining relatively high success rates (>60%) from 2017 onwards.

EDA WITH SQL

ALL LAUNCH SITE NAMES

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Showing the names of the distinct launch sites in the space mission

EDA WITH SQL

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

Showing 5 entries where the launch sites start with the string 'CCA'.

EDA WITH SQL

TOTAL PAYLOAD MASS

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

TOTAL_PAYLOAD
111268

Displaying the total payload mass

EDA WITH SQL

AVERAGE PAYLOAD MASS

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

Displaying average payload mass

EDA WITH SQL

AVERAGE PAYLOAD MASS

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

Displaying average payload mass

FIRST SUCCESSFUL GROUND LANDING DATE

first_successful_landing
2015-12-22

Displaying first succesful landing

EDA WITH SQL

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

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

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

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

DISPLAYING NUMBER OF SUCCESSFULL AND FAILURE OUTCOMES

EDA WITH SQL

NAMES OF THE BOOSTER_VERSIONS WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS

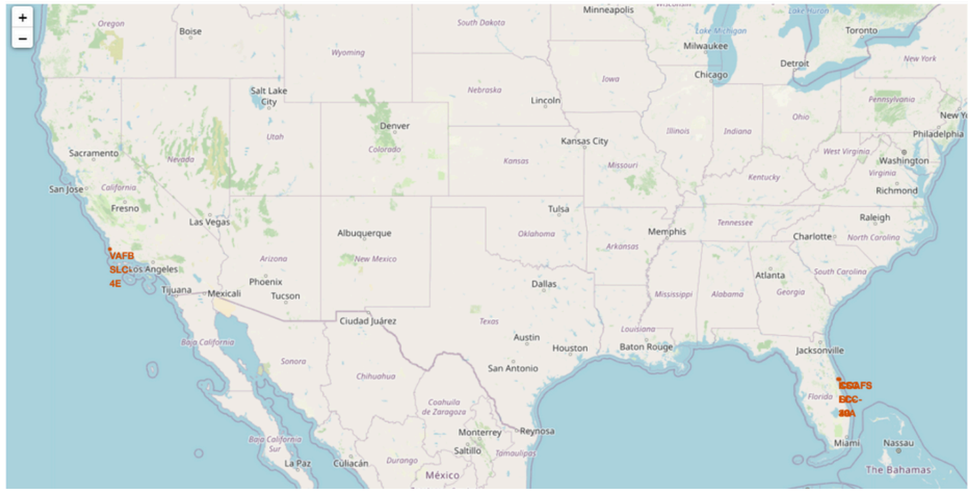
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

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

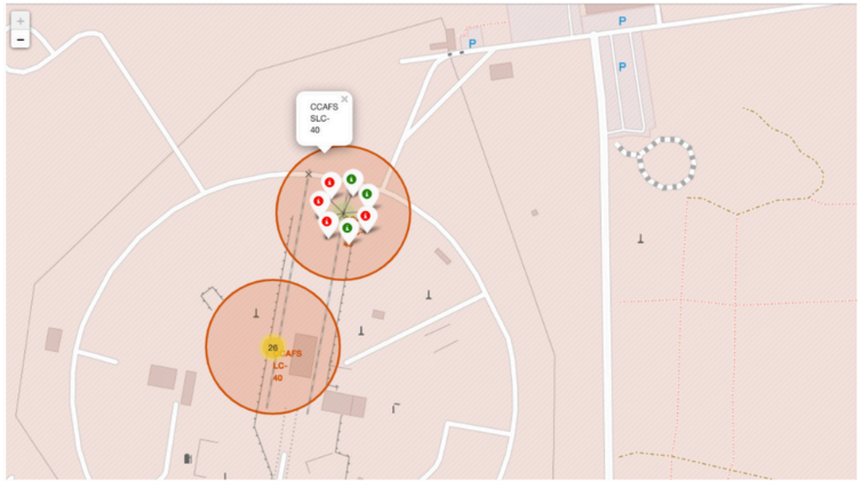
DISPLAYING names of the booster_versions which have carried the maximum payload mass

INTERACTIVE MAP WITH FOLIUM



Most launch sites are close to the equator, where Earth's surface moves fastest at 1670 km/h, giving spacecraft extra speed to help them stay in orbit due to inertia. Additionally, launch sites are near the coast to reduce the risk of debris falling near populated areas when rockets are launched over the ocean.

INTERACTIVE MAP WITH FOLIUM

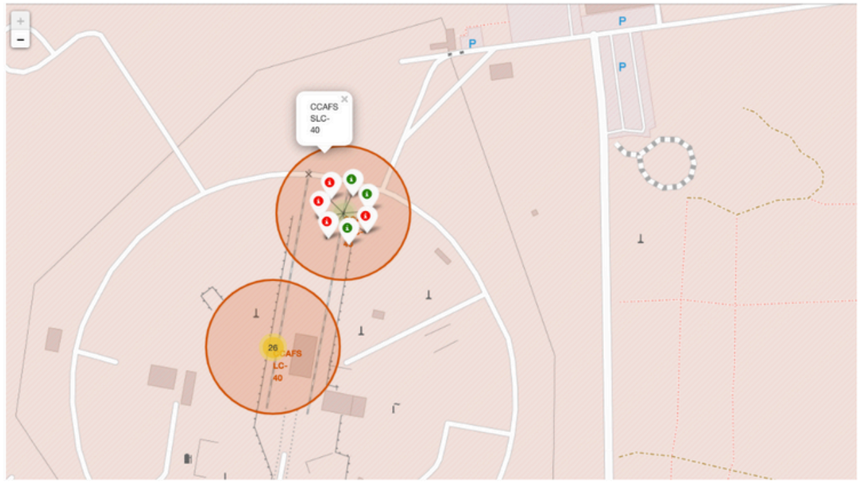


This map shows the CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40) launch site. Key features include:

1. Interactive map controls (+/-) in the top left corner
2. Two orange circles highlighting key areas of the launch complex
3. A cluster of color-coded markers (red and green) in the upper area, indicating launch success rates according to the caption
4. The specific "AFS LC-40" location marked in yellow with number "26"
5. Various support roads and structures around the complex

The caption indicates that the color-coded markers help identify which launch sites have relatively high success rates. Based on the number of green markers, this site appears to have a good track record of successful launches.

INTERACTIVE MAP WITH FOLIUM

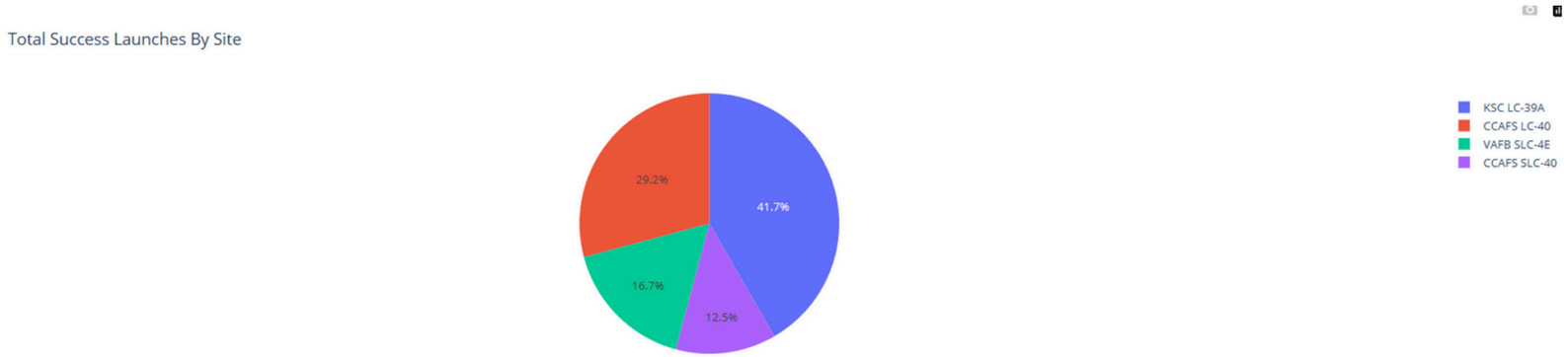


This map shows the CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40) launch site. Key features include:

1. Interactive map controls (+/-) in the top left corner
2. Two orange circles highlighting key areas of the launch complex
3. A cluster of color-coded markers (red and green) in the upper area, indicating launch success rates according to the caption
4. The specific "AFS LC-40" location marked in yellow with number "26"
5. Various support roads and structures around the complex

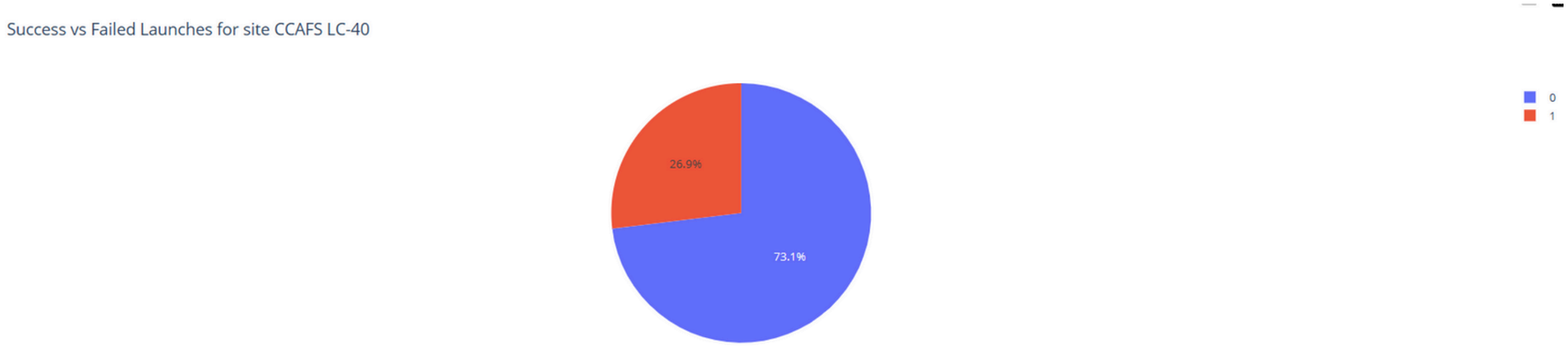
The caption indicates that the color-coded markers help identify which launch sites have relatively high success rates. Based on the number of green markers, this site appears to have a good track record of successful launches.

BUILD A DASHBOARD WITH PLOTLY DASH



The chart clearly indicates that KSC LC-39A has the highest number of successful launches among all the sites.

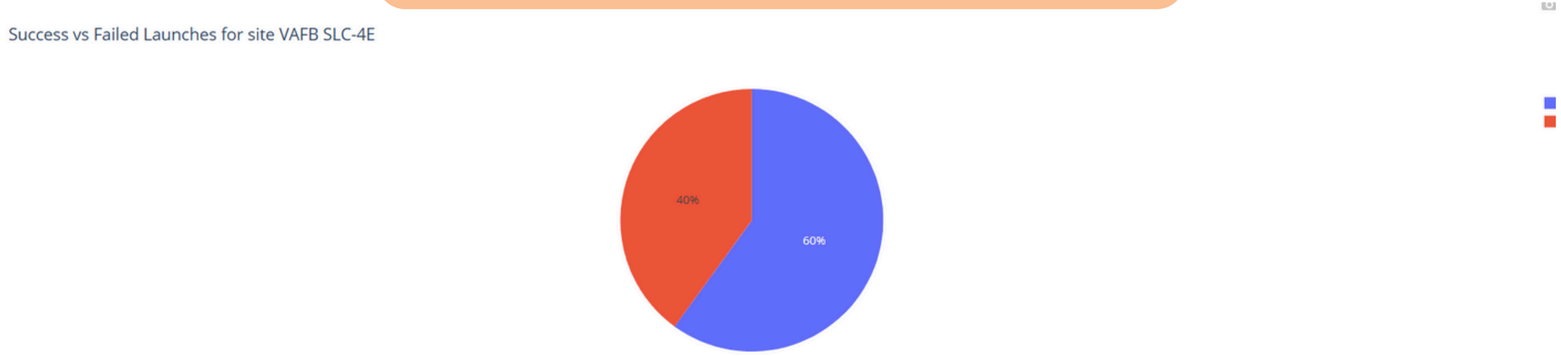
BUILD A DASHBOARD WITH PLOTLY DASH



This image shows a SpaceX Launch Records Dashboard with two main components:

1. A pie chart showing "Success vs Failed Launches for site CCAFS LC-40":
   - 73.1% success rate (shown in blue)
   - 26.9% failure rate (shown in red)
2. A slider at the bottom for "Payload Range (kg)" ranging from 0 to 9000 kg

BUILD A DASHBOARD WITH PLOTLY DASH



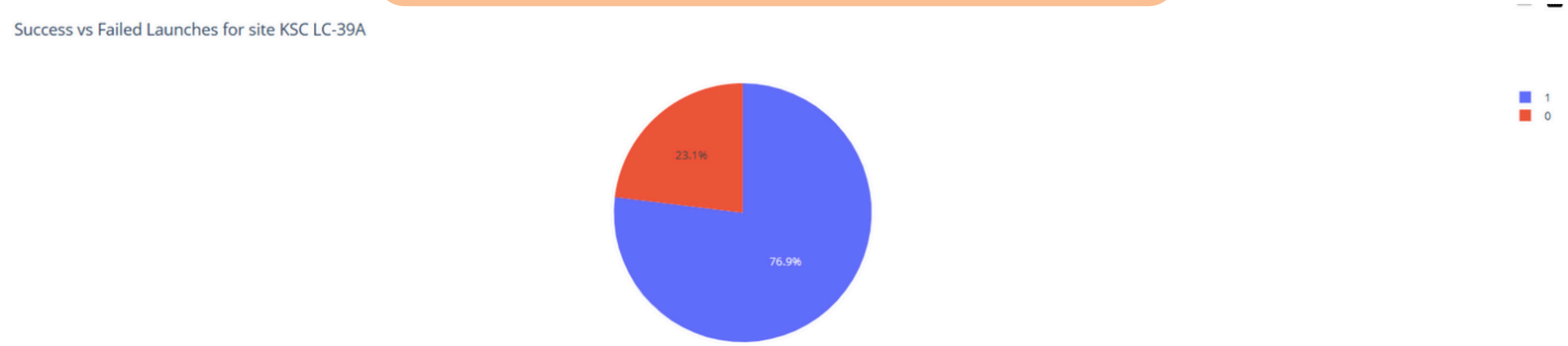
ıd Range (kg):

This pie chart shows "Success vs Failed Launches for site VAFB SLC-4E" (Vandenberg Air Force Base Space Launch Complex 4E):

- 60% success rate (shown in blue)
- 40% failure rate (shown in red)

Compared to the previous CCAFS LC-40 site (which had 73.1% success), VAFB SLC-4E has a lower success rate. A payload range slider appears to be present at the bottom of the dashboard, though it's partially cut off in this image.

BUILD A DASHBOARD WITH PLOTLY DASH



This pie chart shows "Success vs Failed Launches for site KSC LC-39A" (Kennedy Space Center Launch Complex 39A):

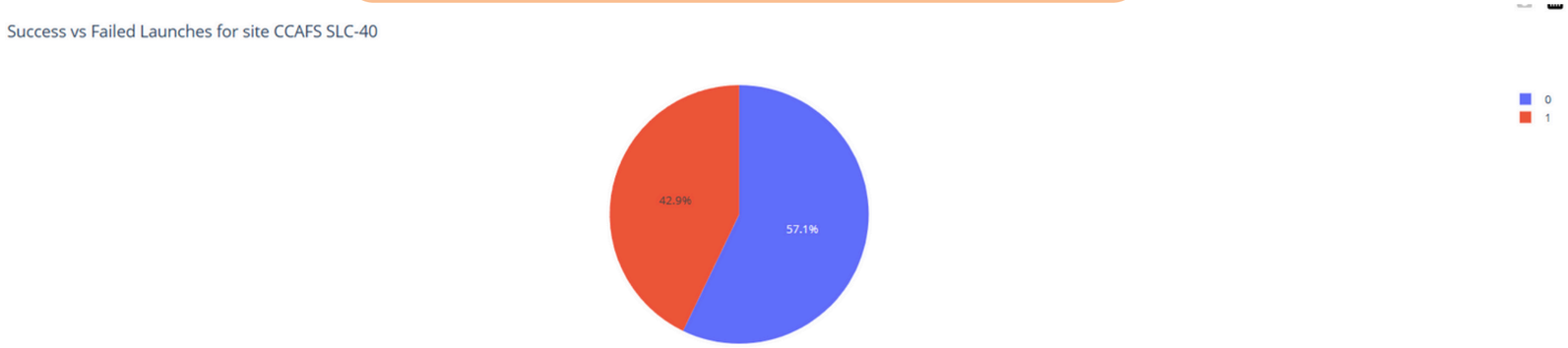
- 76.9% success rate (shown in blue)
- 23.1% failure rate (shown in red)

Comparing all three launch sites:

1. KSC LC-39A: 76.9% success (best performance)
2. CCAFS LC-40: 73.1% success (close second)
3. VAFB SLC-4E: 60% success (lowest success rate)

KSC LC-39A shows the highest success rate of all three launch sites analyzed, with more than three-quarters of launches being successful.

BUILD A DASHBOARD WITH PLOTLY DASH



This is another pie chart for "Success vs Failed Launches for site CCAFS SLC-40", but with different percentages than the earlier version:

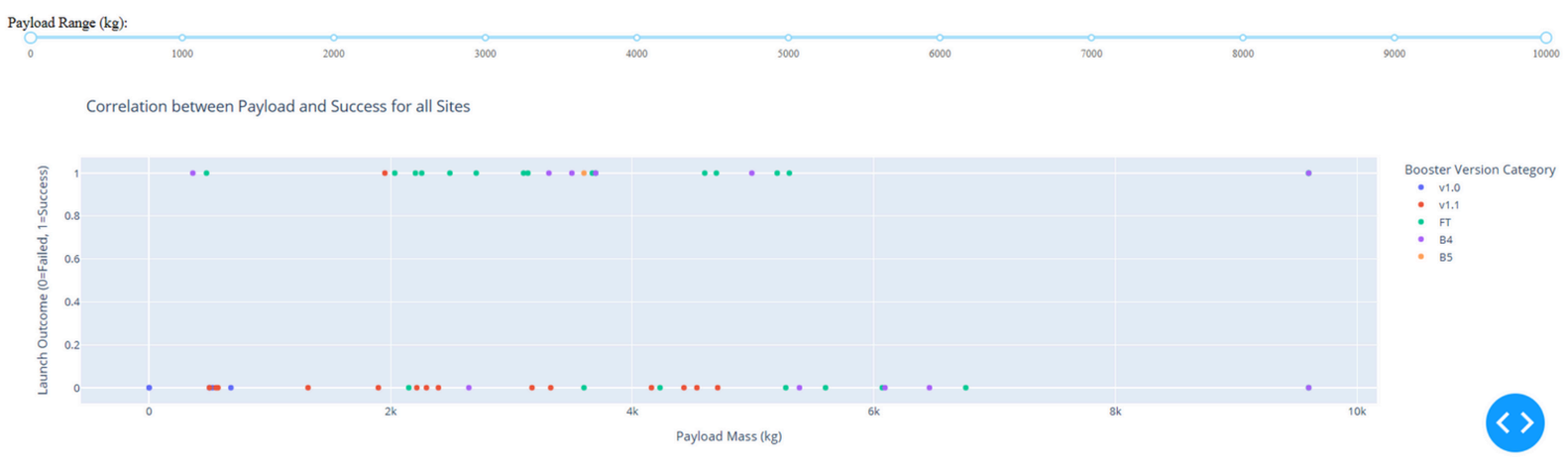
- 57.1% success rate (shown in blue)
- 42.9% failure rate (shown in red)

This appears to show different data than the previous CCAFS SLC-40 chart (which showed 73.1% success). The difference might be due to:

1. Data from a different time period
2. Different payload range selection (using the slider at the bottom)
3. Different filtering criteria

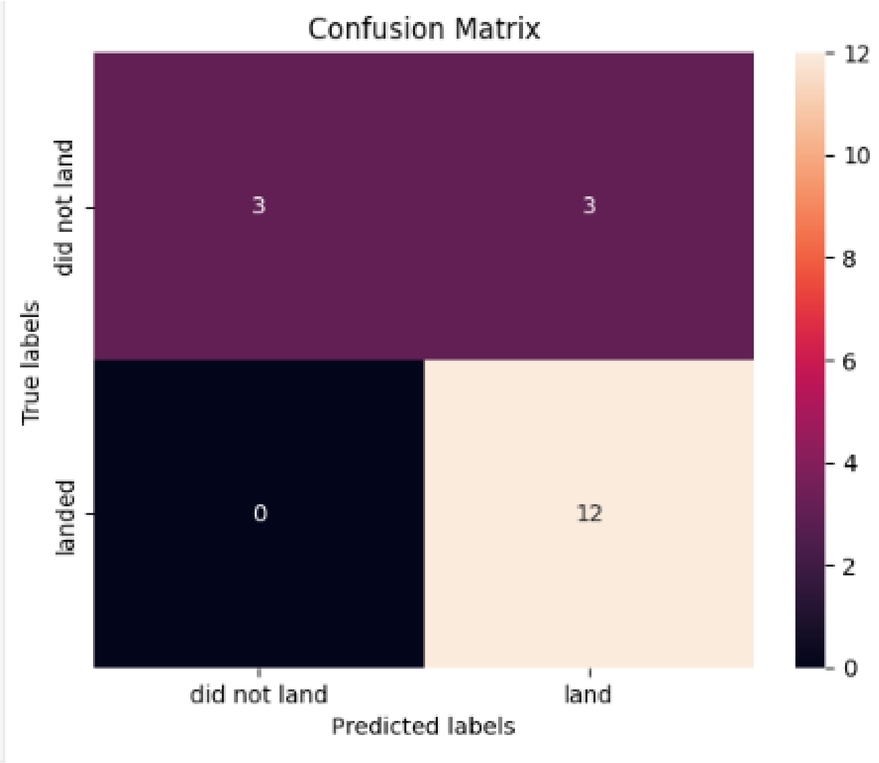
The success rate is notably lower in this view, suggesting these might be launches under specific conditions or during a particular period.

BUILD A DASHBOARD WITH PLOTLY DASH



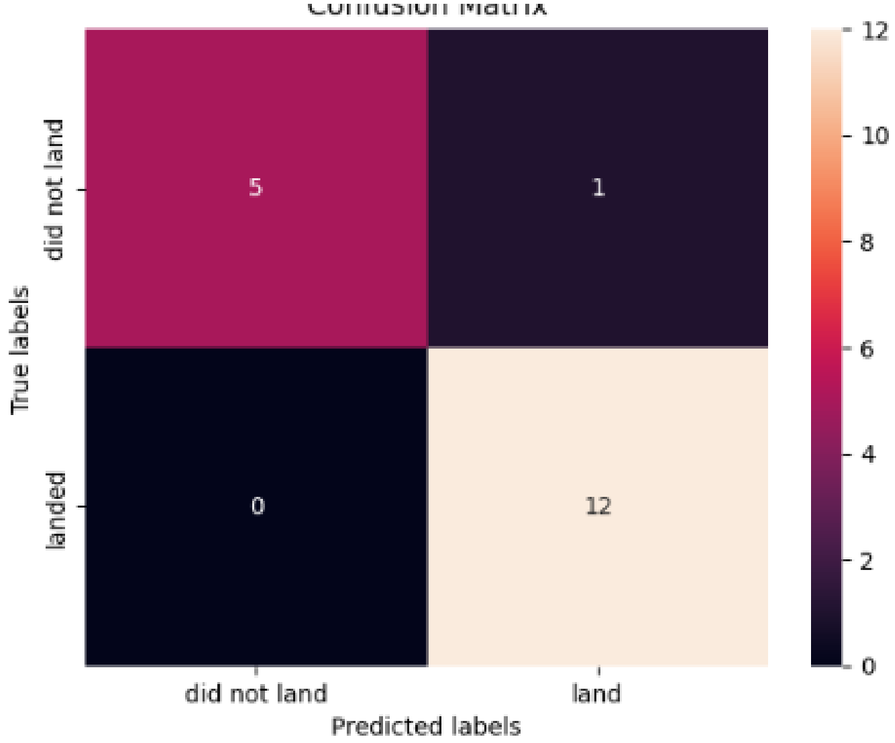
This chart illustrates the correlation between payload mass (in kilograms) and the success of launches across all sites, categorized by booster versions. The x-axis shows the payload mass, while the y-axis represents the launch outcome, with 0 indicating a failed launch and 1 representing a successful one. Each colored dot signifies a different booster version category, with a legend on the right indicating which color corresponds to each version (v1.0, v1.1, FT, B4, and B5).

PREDICTIVE ANALYSIS (CLASSIFICATION)



This task calculates the model's accuracy on test data, which is 83.3%. The confusion matrix shows that out of 18 predictions, the model correctly predicted 12 successful landings and 3 non-landings, with 3 misclassifications.

PREDICTIVE ANALYSIS (CLASSIFICATION)



This confusion matrix shows that the model correctly predicted 12 successful landings and 5 non-landings, with only 1 incorrect prediction.

CONCLUSION

This project provided a comprehensive analysis of SpaceX's Falcon 9 launches to predict first-stage landing success, impacting cost estimations. Starting with data collection via SpaceX's API and web scraping, we gathered essential launch information. Through data wrangling, we cleaned and structured the data for analysis. Exploratory Data Analysis (EDA) helped uncover patterns and relationships impacting launch outcomes. Finally, using predictive analysis, we developed a machine learning pipeline to optimize models and accurately forecast landing success, helping to make competitive cost estimations. This workflow highlights key insights and supports data-driven decisions in the space industry.

SPECIAL THANKS TO:

IBM

UABC