IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gracjan Anioł

April 2023

TABLE OF CONTENTS

EXECUTIVE SUMMARY

- **Summary of methodologies**
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics
  - Machine Learning Prediction

- **Summary of all results**
  - Exploratory Data Analysis Result
  - Interactive Analytics in Screenshots
  - Predictive Analytics Result

INTRODUCTION

Project background and context

SpaceX promotes the launches of its Falcon 9 rockets on their website for a price of 62 million dollars, which is much less than the cost of other providers who charge upwards of 165 million dollars for each launch. One of the reasons for this price difference is that SpaceX is able to reuse the first stage of their rockets.

Therefore, if we can accurately predict whether or not the first stage will successfully land, we can determine the overall cost of a launch. This information can be beneficial for other companies who are looking to bid against SpaceX for a rocket launch.

The objective of this project is to develop a machine learning pipeline that can predict the success rate of the first stage landing.

Problems I want to find answers

- What are the factors that influence the success of a rocket landing?
- How do various features interact with each other to affect the success rate of rocket landing?
- What are the necessary operating conditions for a successful rocket landing program?

“

METHODOLOGY

SECTION 1

EXECUTIVE SUMMARY

- **Data collection methodology:**
  - Data was collected using Space X API
  - Web scraping from Wikipedia

- **Perform data wrangling**
  - One-hot encoding was applied to categorical features

- **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 data that was gathered until this point was processed through a normalization technique to ensure consistency across the data. The dataset was then divided into training and test sets to train and evaluate different classification models. The accuracy of each model was assessed using various combinations of parameters. Overall, four different classification models were evaluated during this process.

DATA COLLECTION

- **How data sets were collected**
  - Data was collected using a GET request to the SpaceX API.
  - The response content was decoded as a JSON using the ".json()" function and then transformed into a Pandas dataframe using ".json_normalize()".
  - Data cleaning was done next, which involved checking for missing values and filling them in where necessary.
  - Web scraping was also performed from Wikipedia for Falcon 9 launch records using BeautifulSoup.
  - The aim was to extract the launch records as an HTML table, parse the table, and convert it to a Pandas dataframe for further analysis.

- I used the get request to the public SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- This is the link to GitHub URL with the completed SpaceX API calls notebook

COLLECTING DATA WITH SPACEX REST CALLS:

- Sending GET request to the SpaceX API to retrieve data.

DECODING RESPONSE CONTENT WITH JSON:

- Decoding response content with ".json()" function call.
- Transforming response content into a Pandas dataframe using ".json_normalize()".

DECODING RESPONSE CONTENT WITH JSON:

- Checking for missing values.
- Filling in missing values where necessary.

DATA COLLECTION – SCRAPING

- **I applied scraping additional data from Wikipedia with BeautifulSoup. Data are downloaded from Wikipedia according to the flowchart and then persisted.**
- **This is the link to GitHub URL with the completed web scraping calls notebook**

ACCESSING THE WIKIPEDIA PAGE AND INSPECTING THE HTML CODE:

- Using a web browser to access the Wikipedia page with Falcon 9 launch records.
- Using the browser's developer tools to inspect the HTML code of the Wikipedia page.
- Identifying the HTML elements that contain the launch record data of interest.

PARSING THE HTML CODE AND CREATING A DATA FRAME:

- Using a web scraping tool like BeautifulSoup to parse the HTML code.
- Extracting the table that contains the Falcon 9 launch records.
- Renaming columns to be more informative.

EXPORTING THE DATA:

- Saving the cleaned data to a CSV file for future use.

- This is the link to GitHub URL with the completed web scraping calls notebook

I conducted exploratory data analysis on the dataset to gain insights and understanding.

I calculated the number of launches at each site and the number of occurrences of each orbit type to identify patterns and trends

Additionally, I created a landing outcome label based on the outcome column to classify the success or failure of the landing.

The resulting data from the exploratory analysis and outcome label creation were exported to a CSV file for further use.

EDA WITH DATA VISUALIZATION

In the exploratory data analysis (EDA), I used various data visualization techniques to understand the relationships between different variables. Here's a summary of the charts I plotted and why I chose them:

- **Catplot:** I visualized the relationship between Flight Number and Launch Site using a categorical plot to compare the number of launches for each site.
- **Scatterplot:** I plotted the relationship between Payload and Launch Site using a scatter plot to identify any trends or patterns in the data.
- **Bar chart:** To understand the success rate of each orbit type, I used a bar chart to compare the number of successful and unsuccessful launches for each orbit type.
- **Scatterplot:** I visualized the relationship between Flight Number and Orbit type using a scatter plot to identify any trends or patterns in the data.
- **Scatterplot:** Similarly, I plotted the relationship between Payload and Orbit type using a scatter plot to understand how the payload varies for each orbit type.
- **Line plot: Finally,** I visualized the launch success yearly trend using a line plot to understand how the launch success has varied over the years.

By using these different charts, I was able to gain insights into the patterns and relationships between different variables in the dataset
This is the link to GitHub URL with the completed EDA with data visualization

EDA WITH SQL

I applied EDA with SQL to get insight from the data. The following SQL queries were performed to:

- 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)
- Display average payload mass carried by booster version F9 v1.1
- 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 names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

This is the link to GitHub URL with the completed EDA with SQL

BUILD AN INTERACTIVE MAP WITH FOLIUM

I created various map objects and added them to a folium map such as markers, circles, lines, and marker clusters.

- I marked all launch sites on a folium map
- Map objects such as markers, circles, lines and marker clusters were added to indicate the success or failure of launches for each site
- Launch outcomes (failure or success) were assigned to class 0 and 1, respectively
- Color-labeled marker clusters were used to identify launch sites with high success rates
- Distances between launch sites and nearby features (e.g. railways, highways, coastlines, cities) were calculated and analyzed to answer specific questions.

•**This is the link to GitHub URL of my completed interactive map with Folium map**

The interactive dashboard I created using Plotly Dash includes two types of visualizations: pie charts and scatter graphs.

- The pie charts display the percentage of launches by site,
- The scatter graph shows the relationship between outcome and payload mass for different booster versions.

I added these plots and interactions to provide a clear and intuitive way to visualize the data.

- The pie charts allow us to easily compare the percentage of launches by site, giving us insights into which launch sites are the most active.
- The scatter graph helps us understand the relationship between outcome and payload mass, which can be useful for identifying trends and making predictions.

This is the link to GitHub URL of my completed Plotly Dash lab

PREDICTIVE ANALYSIS (CLASSIFICATION)

Here is a summary of my model development process:

Building and Preparing Data:

- Loaded the data using NumPy and Pandas
- Transformed the data
- Split data into training and testing sets

Building and Tuning Models:

- Built different machine learning models: logistic regression, support vector machine, decision tree and k nearest neighbors
- Tuned different hyperparameters using GridSearchCV
- Used accuracy as the metric for our model

Improving Model:

- Improved the model using feature engineering and algorithm tuning

Finding the Best Performing Model:

- Compared the four classification models
- Found the best performing classification model

This is the link to GitHub URL of my completed predictive analysis

Here is the summary of exploratory data analysis results:

- Space X uses four different launch sites for their missions.
- The average payload carried by the F9 v1.1 booster is 2,928 kg.
- Many Falcon 9 booster versions have successfully landed on drone ships, carrying payloads above the average.
- The number of successful landing outcomes improved over the years.

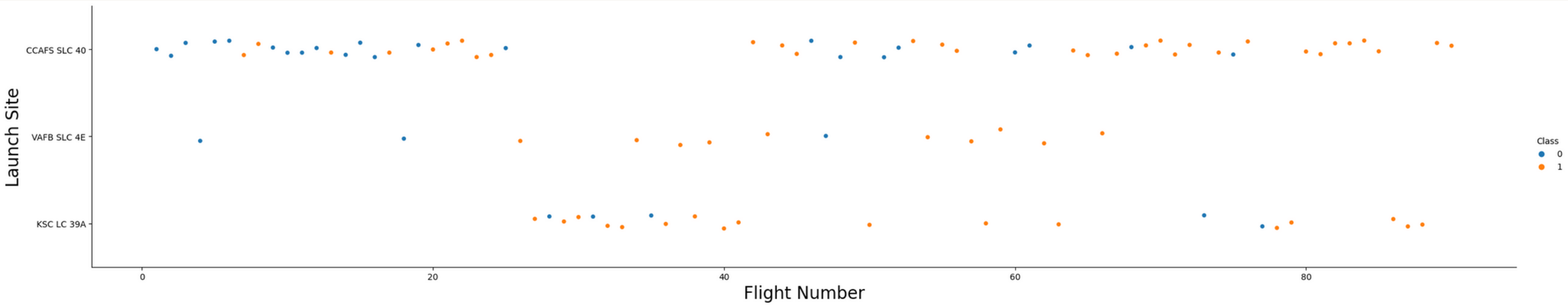
Here is the summary of interactive data analysis results:

- Most launches happens at east cost launch sites.
- Launch sites use to be in safety places, near sea

INSIGHTS DRAWN FROM EDA

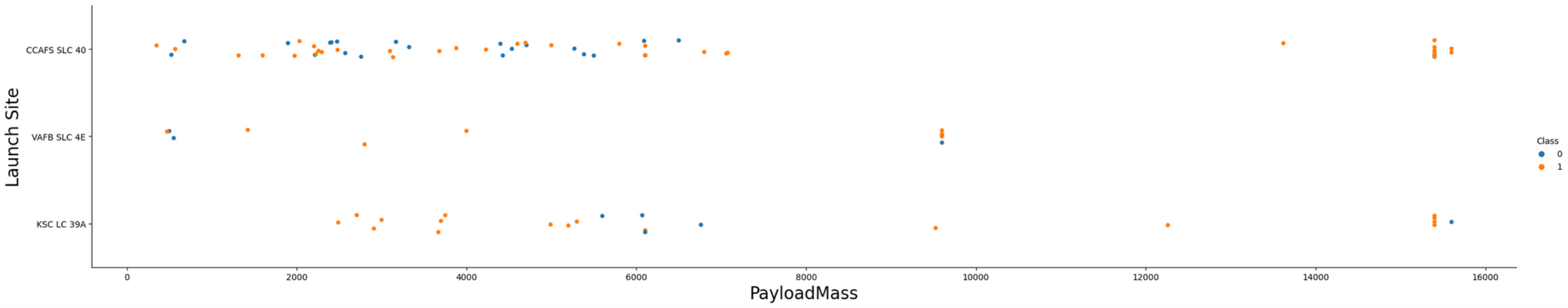
SECTION 2

FLIGHT NUMBER VS. LAUNCH SITE



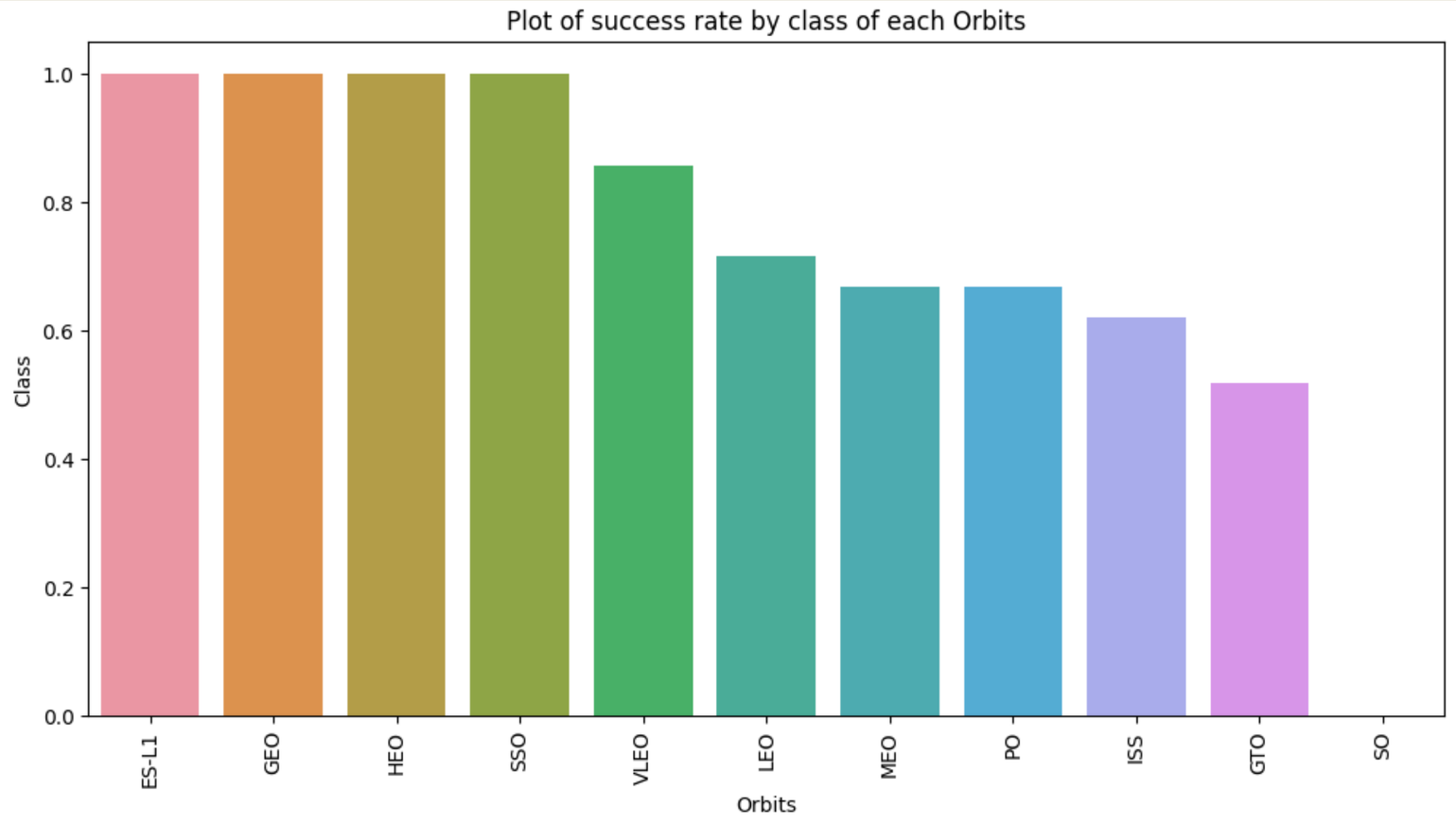
The catplot helped us visualize the relationship between flight amount and success rate at a launch site. We observed that as the flight amount increased, the success rate also increased, and the plot also showed the improvement of the general success rate over time.

PAYLOAD VS. LAUNCH SITE



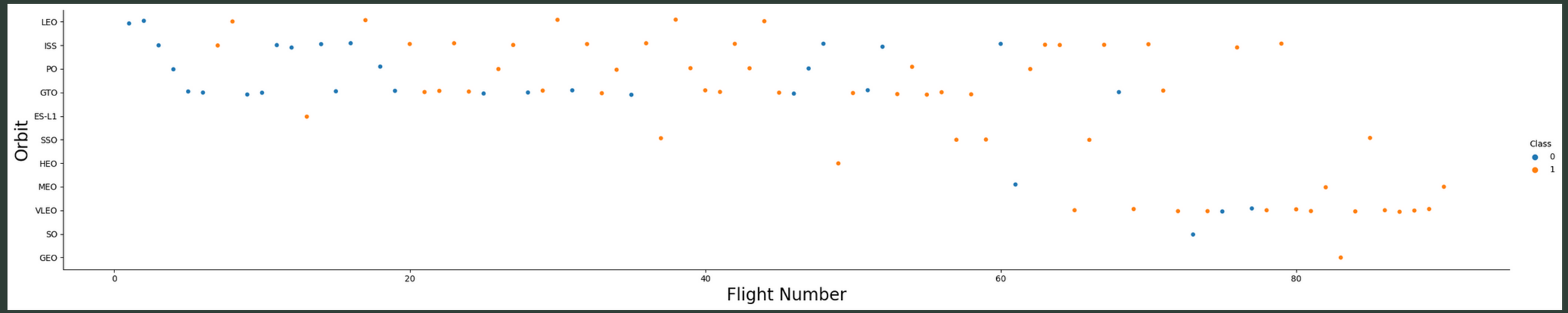
The catplot helped us visualize the relationship between payload and success rate at a launch site. We see the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000) and the greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.

SUCCESS RATE VS. ORBIT TYPE



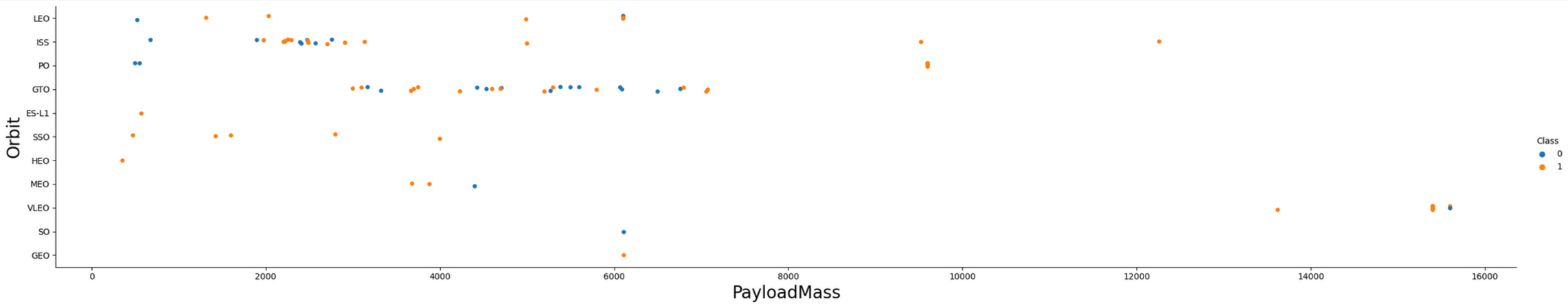
From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

FLIGHT NUMBER VS. ORBIT TYPE



The catplot helped us visualize the relationship between Flight Numer and Orbit Type. We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

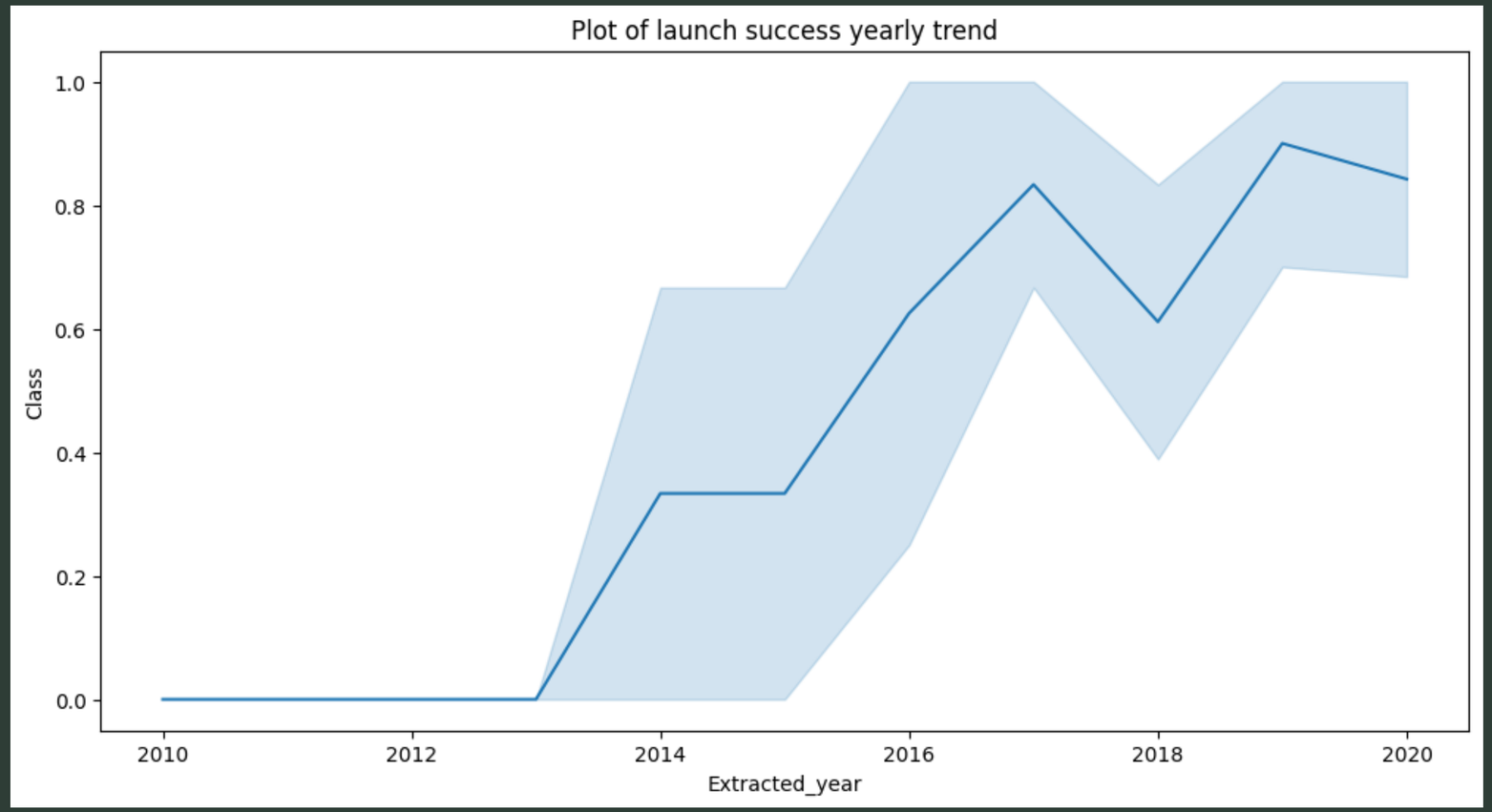
PAYLOAD VS. ORBIT TYPE



The catplot helped us visualize the relationship between Payload Mass and Orbit Type. We see that heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

LAUNCH SUCCESS YEARLY TREND

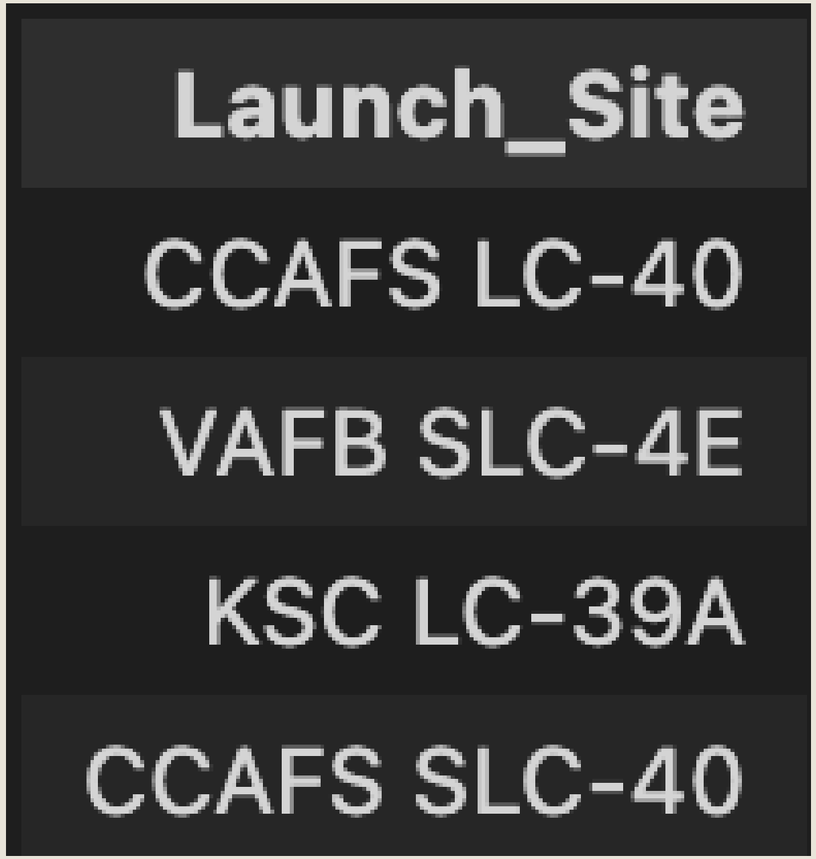


The line chart helped us visualize the relationship between years and success rate. We observe that the success rate since 2013 kept increasing till 2020.

ALL LAUNCH SITE NAMES

I used SQL to extract data from a database and found the unique launch site names.

This query result provides a list of the different sites from which launches have occurred, which can be useful for further analysis and understanding of the data.



LAUNCH SITE NAMES BEGIN WITH 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

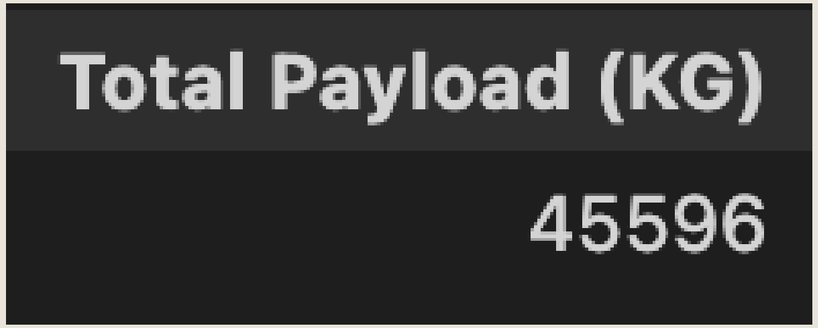
I used SQL to extract data from a database and found 5 records where the launch sites begin with CCA.

This query result provides a subset of the launch data that is specific to launch sites that have names starting with CCA, which may be useful for analyzing launches from these particular sites.

TOTAL PAYLOAD MASS

I used SQL to extract data from a database and calculated the total payload carried by boosters from NASA, which was found to be 2,928 kg.

This query result provides an important metric for understanding the performance of NASA's boosters and can be useful for comparing their payloads to those of other organizations.



AVERAGE PAYLOAD MASS BY F9 V1.1

I used SQL to extract data from a database and calculated the average payload mass carried by booster version F9 v1.1, which was found to be 2928.4 kg.

This query result provides an important metric for understanding the performance of the F9 v1.1 booster and can be useful for comparing its average payload to other booster versions or organizations.

AVG Payload Mass (KG)
2928.4

FIRST SUCCESSFUL GROUND LANDING DATE

I used SQL to extract data from a database and found the date (01-05-2017) of the first successful landing outcome on a ground pad.

This query result is significant as it marks a milestone in SpaceX's history of reusable rocket technology, paving the way for more cost-effective and sustainable space launches.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

I used SQL to extract data from a database and found list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

The result of this query is a list of booster names that meet the specified criteria: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

This information can be useful for analyzing the performance of SpaceX boosters and comparing them to those of other organizations.

Booster_Version	payloadmass
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

Based on the SQL query, the total number of successful and failure mission outcomes was calculated from the database.

The query result shows that there were 98 successful mission outcomes, 1 failure in-flight, and 2 other success outcomes with unclear payload status.

This metric provides important information about the overall success rate of missions conducted by the organization and can be used to monitor and improve the performance of future missions.

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

BOOSTERS CARRIED MAXIMUM PAYLOAD

Using SQL, I extracted data from a database and identified 12 boosters that have carried the maximum payload mass (15600 KG).

This information could be useful for understanding the capabilities of these boosters and for comparison with other boosters or organizations.

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

2015 LAUNCH RECORDS

Using SQL, I extracted data from the database to list the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015.

This query allows us to identify the specific failures that occurred during that year, the booster versions involved, and the launch sites from which the missions were launched.

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

Using SQL, I extracted data from the database to rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20.

This information can provide insights into the success rate of landing outcomes and help in identifying any patterns or trends in the data.

Landing _Outcome	Counting
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

LAUNCH SITES PROXIMITIES ANALYSIS

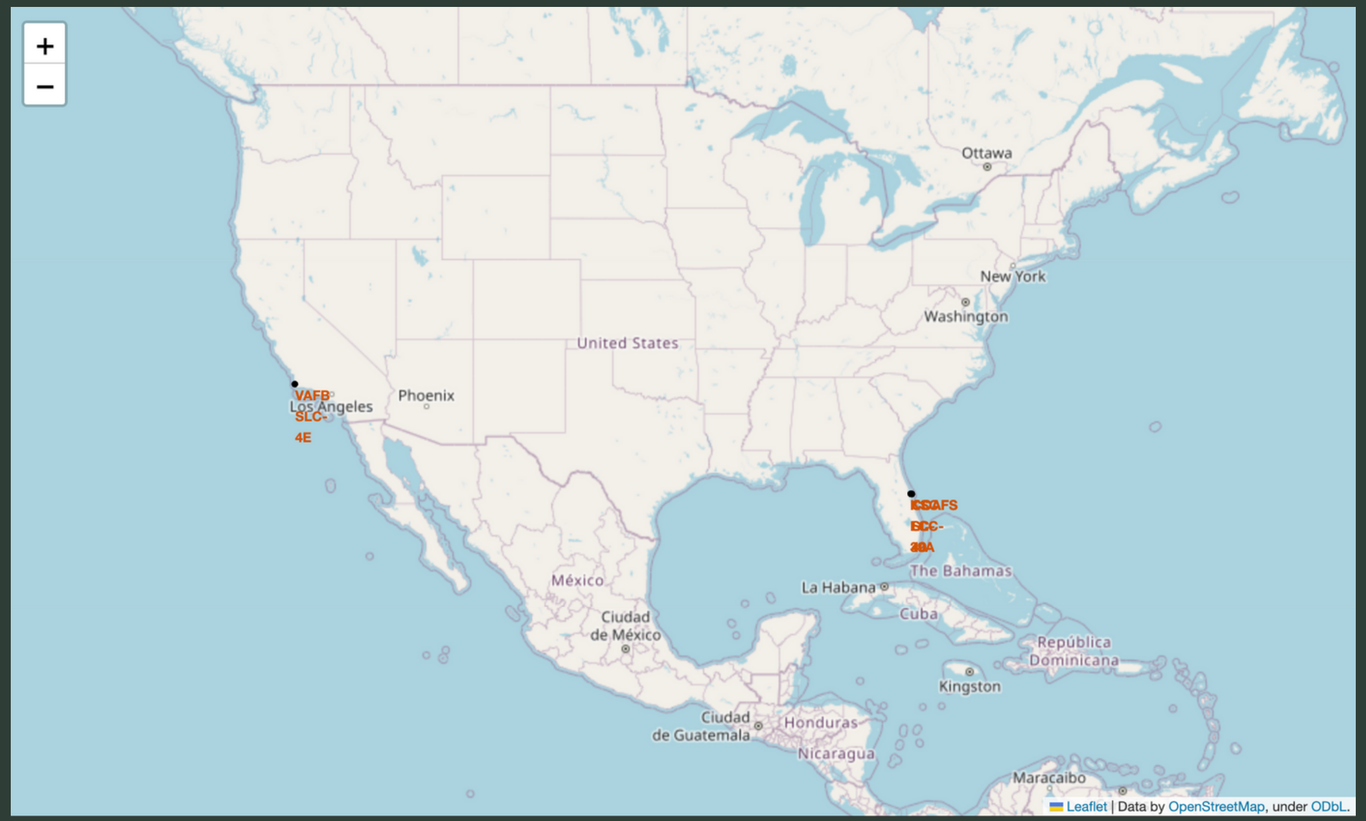
SECTION 3

ALL LAUNCH SITES

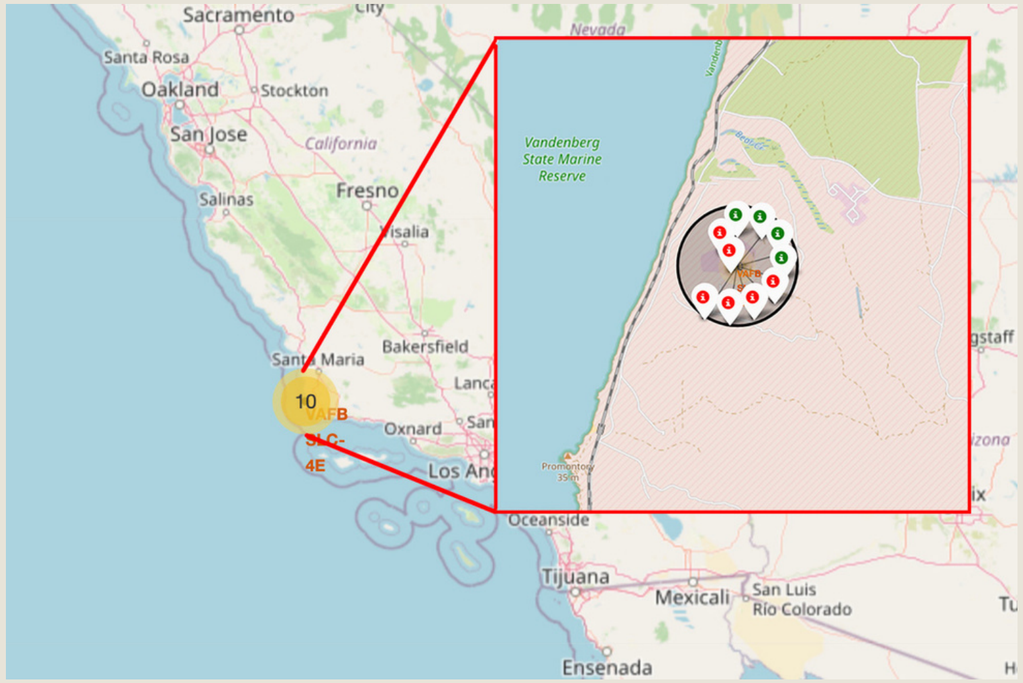
Based on the generated Folium map, it appears that the launch sites are close to the Equator.

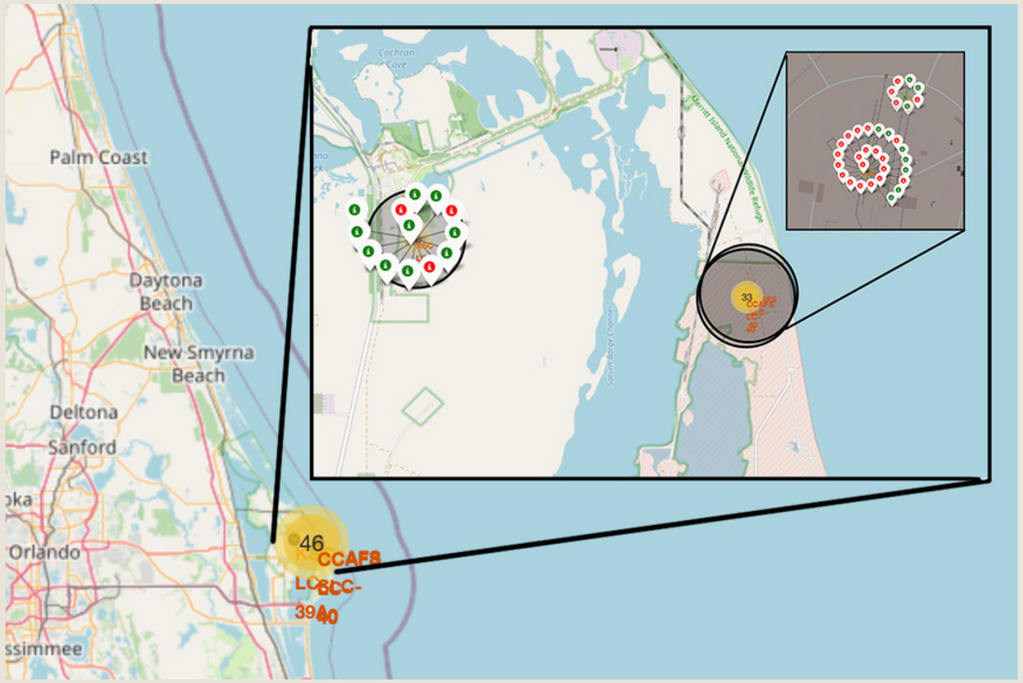
Regarding the proximity to the coast, most launch sites are located near the coastlines, which is likely due to safety and logistical reasons such as the ability to transport rockets and equipment by sea.

Overall, the Folium map provides a useful visual representation of the global distribution of launch sites and highlights the variety of factors that are considered when selecting a launch site.



LAUNCH OUTCOMES BY SITE





The screenshots show a map with markers indicating the locations of the various launch sites. Each marker have a popup label with the name of the launch site. Additionally, the markers are color-coded to indicate the launch outcomes (i.e., success or failure) at each site. Sites with a high success rate would be marked with green, while sites with a high failure rate would be marked with red.

This information could be useful in identifying trends and patterns in the launch data and could help inform future launch decisions. Additionally, the map could be used to identify areas where additional launch sites might be needed to improve access to space.

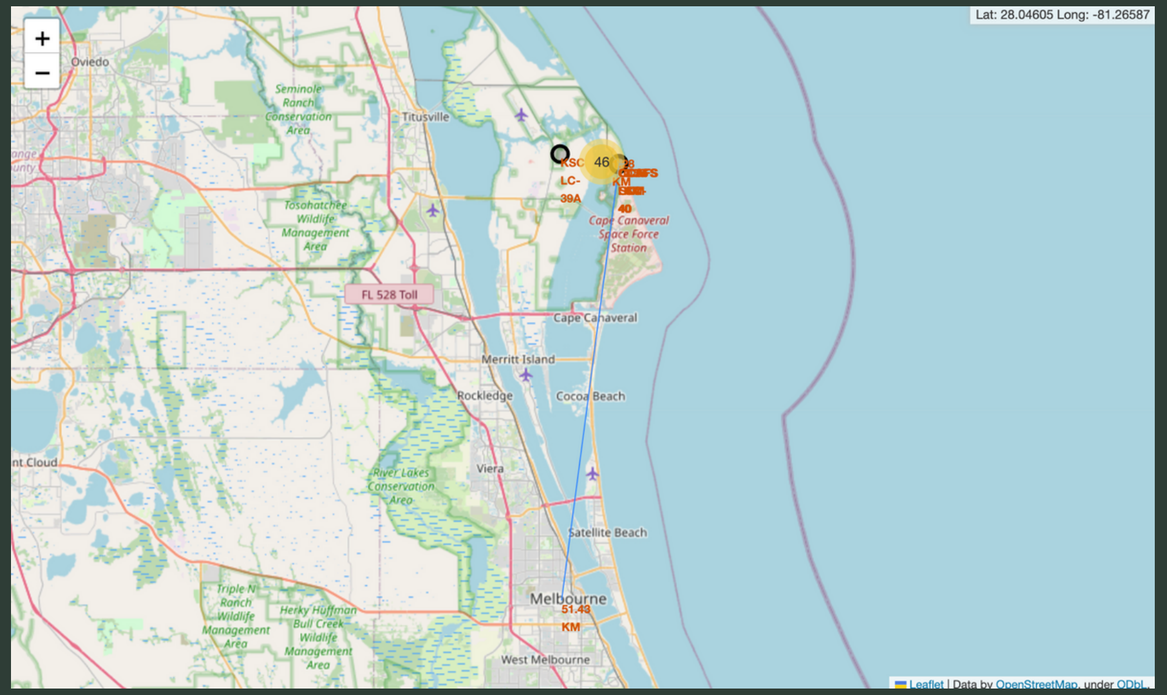
LAUNCH SITE DISTANCE TO LANDMARKS (LOGISTICS AND SAFETY)

Launch sites are typically located near highways to facilitate easy transportation of people and equipment.

They are also located close to railways, which is helpful in transporting heavy cargo.

Additionally, launch sites are usually situated far away from densely populated areas to minimize the risk to human life in the event of an accident or malfunction.

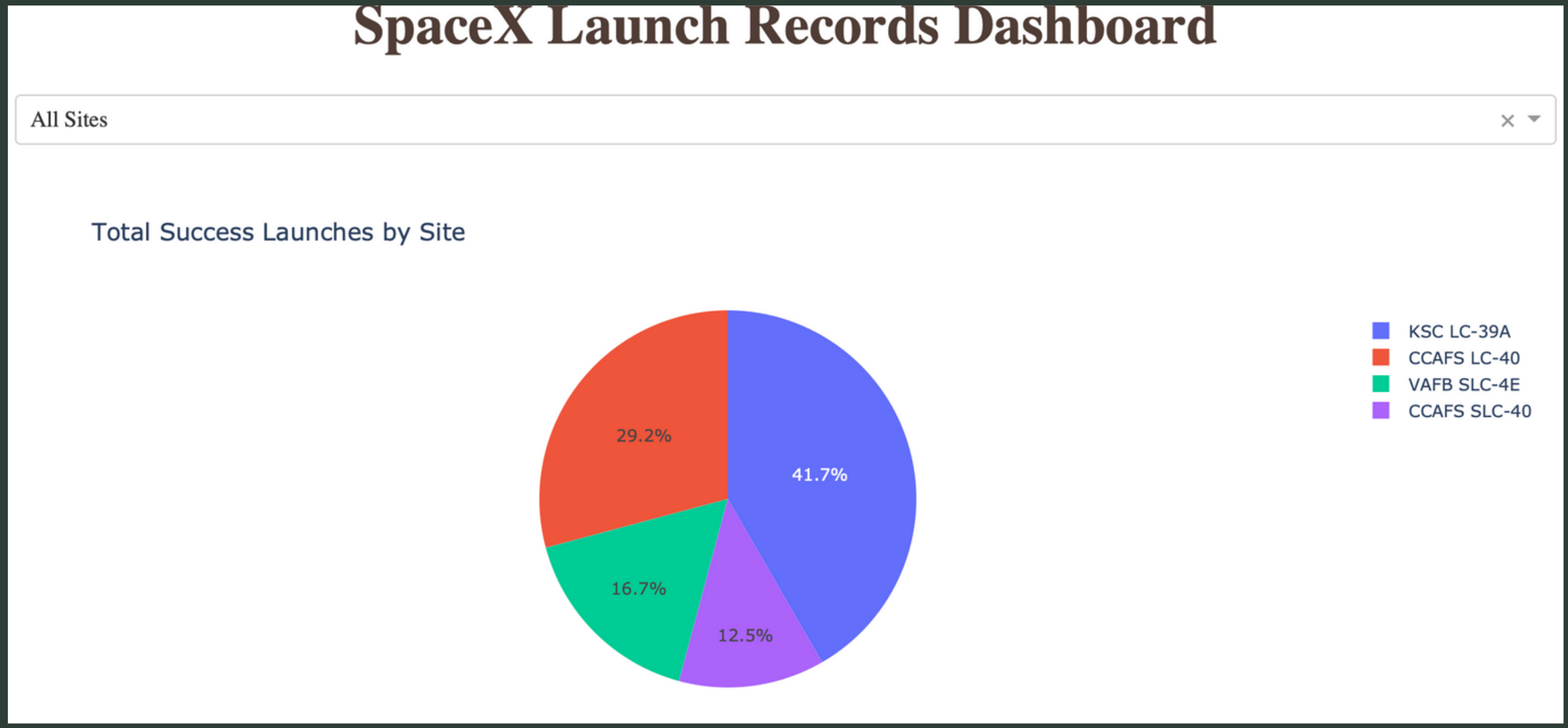
This strategic location selection ensures that the launch process can be carried out with minimal risk to both human life and property.



BUILD A DASHBOARD WITH PLOTLY DASH

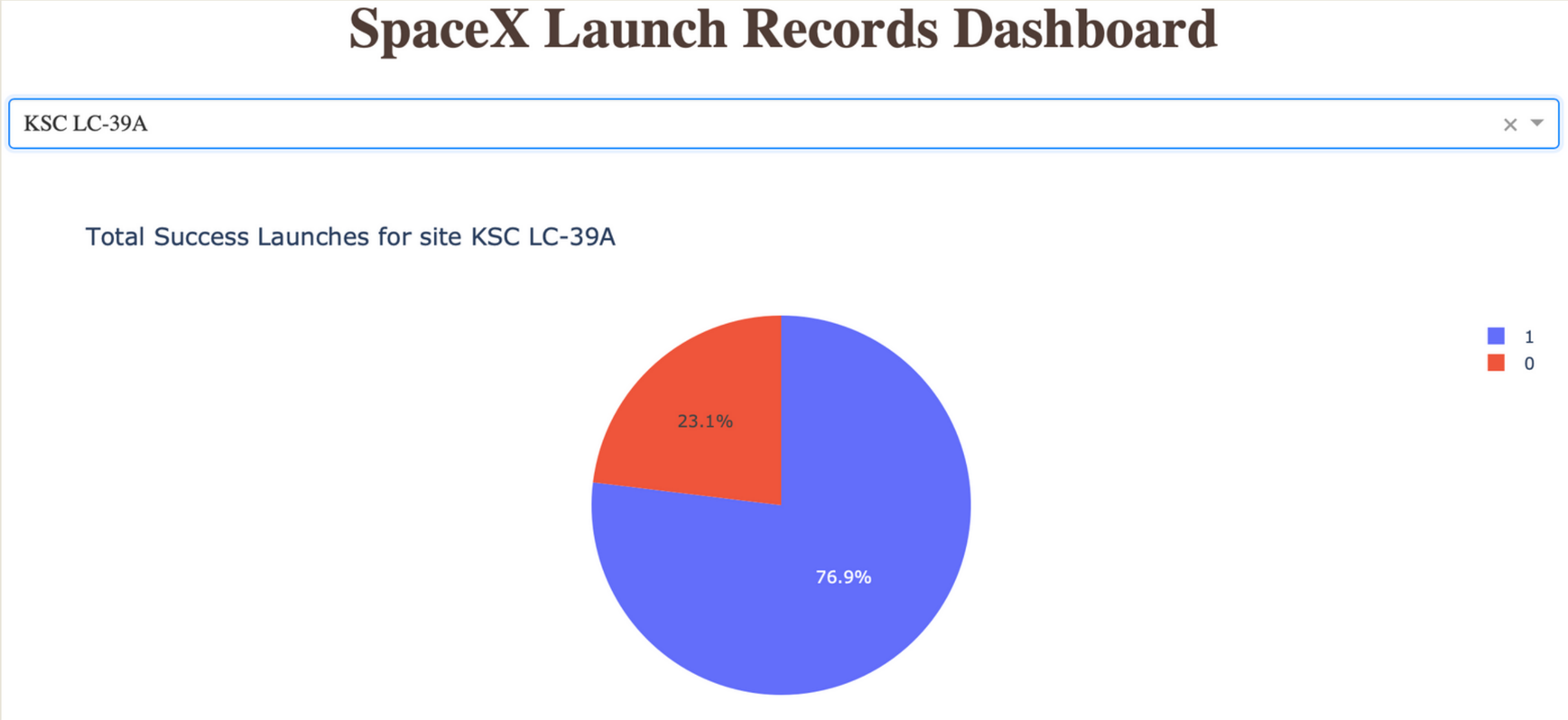
SECTION 4

SUCCESSFUL LAUNCHES BY SITE



The location where rocket launches take place plays a crucial role in determining the success of missions. It has been observed that KSC LC-39A, among all the launch sites, has the highest number of successful launches.

LAUNCH SUCCESS RATIO FOR KSCLC-39A



The piechart shows that KSC LC-39A, a launch site, has achieved a 76.9% success rate in launching rockets. On the other hand, the remaining 23.1% of launches were failures.
This information is a finding that suggests KSC LC-39A is a successful launch site, which has been able to accomplish a high rate of successful missions. The success rate of a launch site is an important element to consider when planning a rocket mission, as it can greatly impact the outcome of the mission.

PAYLOAD VS. LAUNCH OUTCOME

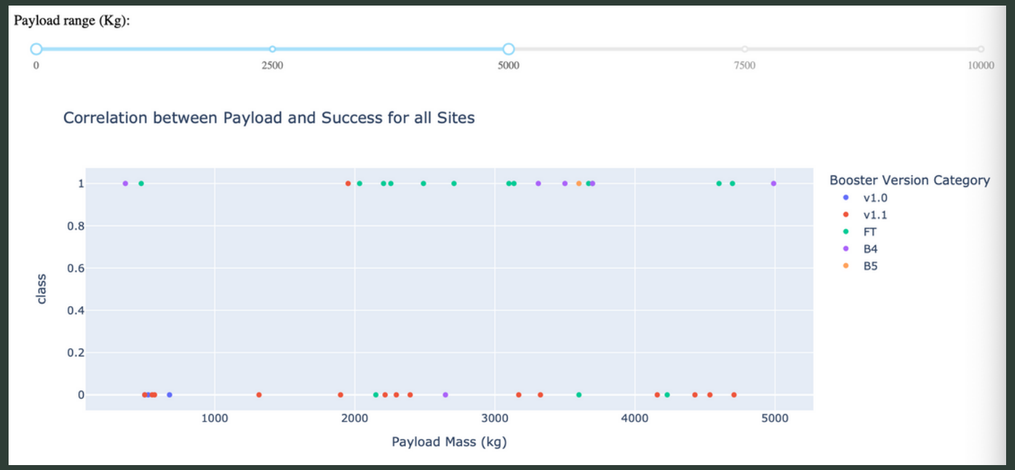
The scatter plot shows the relationship between the payload weight and launch outcome for all launch sites.

- **The findings show that launches with a payload between 5000 and 10000 kg had a relatively low success rate, with only 3 out of 11 launches being successful. Furthermore, this payload range had only two booster versions.**
- **In contrast, the payload range from 0 to 5000 kg had a much higher success rate, with 18 successful launches out of 39 total launches**

Overall, the scatter plot suggests that payload weight plays a significant role in determining the launch success rate.

The findings indicate that launching payloads within a certain weight range and with specific booster versions can increase the likelihood of a successful launch.

This information can be used to inform decision-making in planning and executing future rocket missions.



PREDICTIVE ANALYSIS (CLASSIFICATION)

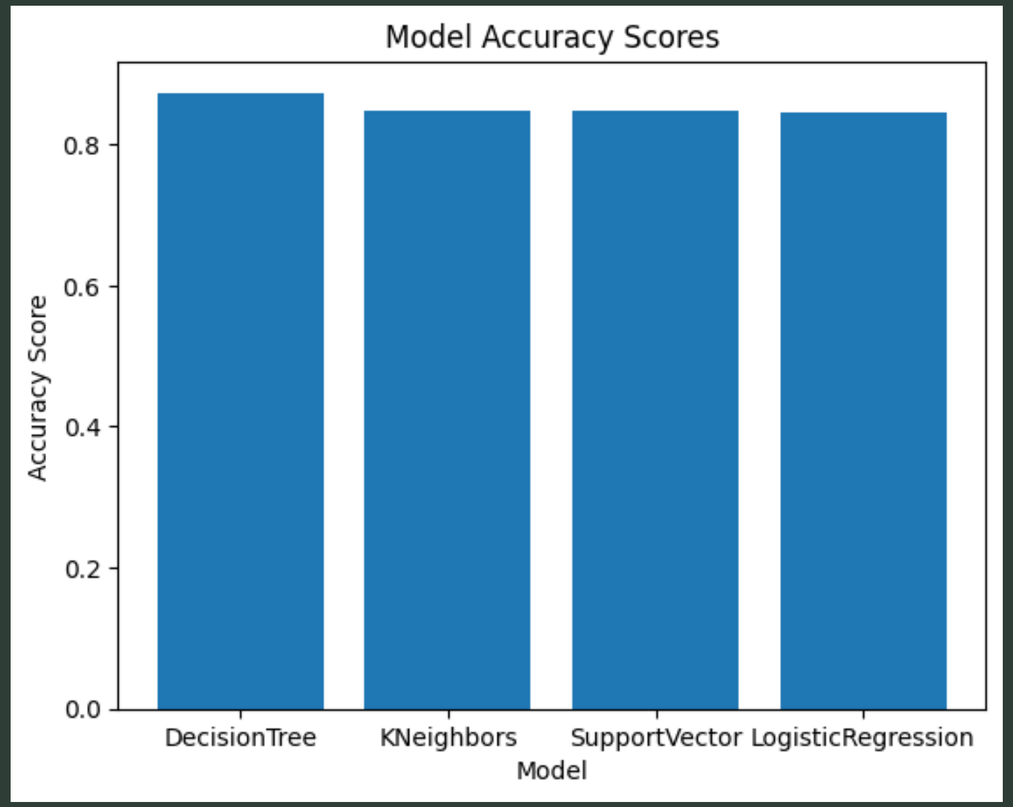
SECTION 5

CLASSIFICATION ACCURACY

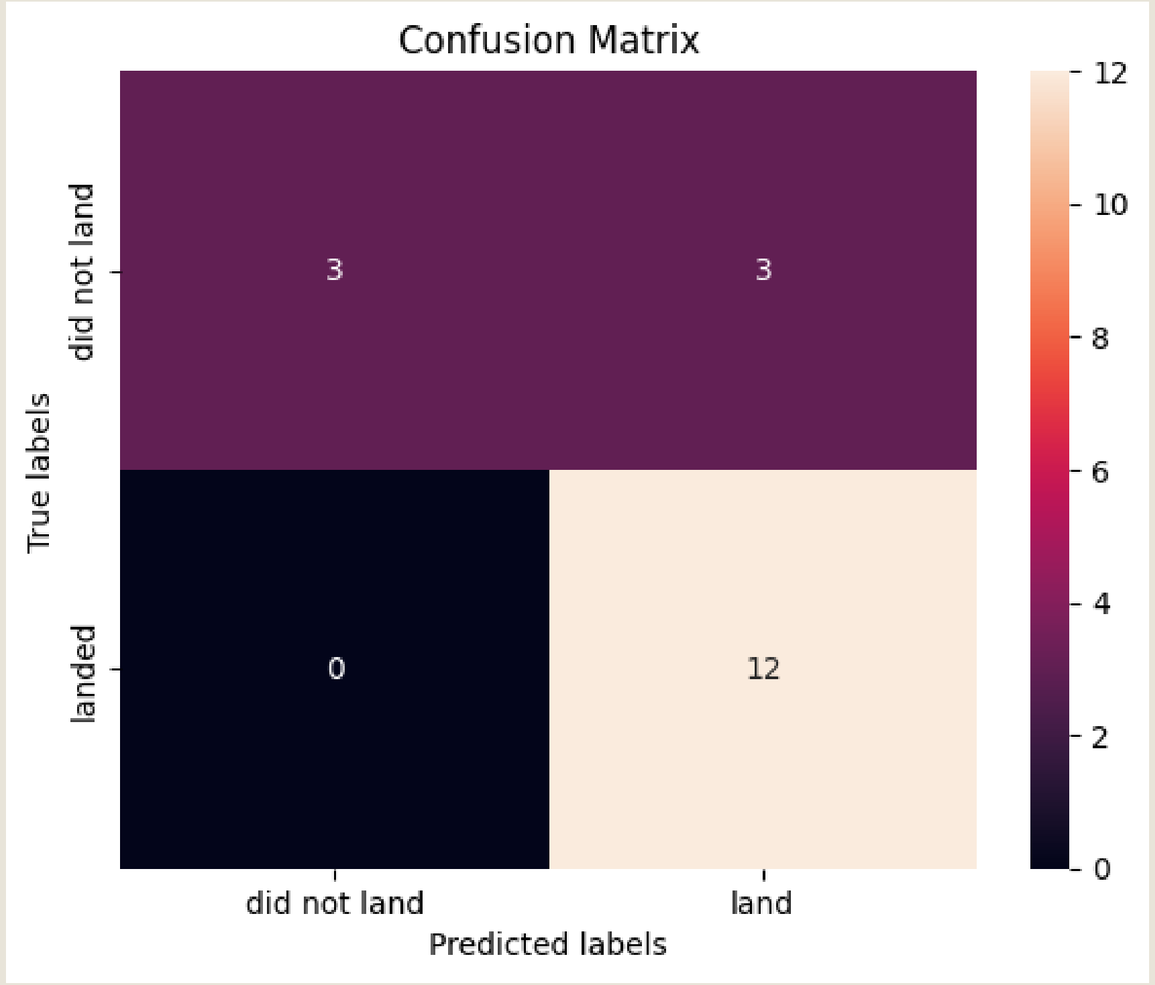
To compare the accuracy of different classification models, a bar chart can be used to visualize their performance. The accuracy scores of each model are plotted on the y-axis and the names of the models are listed on the x-axis.

In this case, four classification models were tested, including K-Nearest Neighbors, Decision Tree, Logistic Regression, and Support Vector Machine.

After comparing their accuracy scores, it was determined that the Decision Tree Classifier had the highest classification accuracy, with scores over 87%. Therefore, the Decision Tree Classifier is the best-performing model among the four tested models.



The confusion matrix (Decision Tree Classifier) is a table that summarizes the performance of a classification model by comparing the actual (true) labels of the data with the predicted labels.



Based on the analysis of the data sources, we can draw the following conclusions:

- **KSC LC-39A is the most successful launch site among all the sites analyzed.**
- **Orbits ES-L1, GEO, HEO, SSO, VLEO had the highest success rates.**
- **Successful landing outcomes have improved over time, thanks to the evolution of processes and rockets. The launch success rate has been increasing since 2013 and peaked in 2020.**

IBM Developer
SKILLS NETWORK

THANK YOU

Gracjan Anioł
April 2023