IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zair Naqvi

January 15, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis With SQL
  - Building an interactive map with Folium
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Demo in Screenshots
  - Predictive Analysis Results

Introduction

Project background and context

SpaceX is the leading company in the commercial space industry, revolutionizing the affordability of space travel. The company lists Falcon 9 rocket launches on its website at a cost of 62 million dollars, while competitors charge over 165 million dollars per launch. This price difference is mainly due to SpaceX's ability to reuse the first stage of the rocket. By predicting whether the first stage will successfully land, we can estimate the cost of a launch. Using publicly available data and machine learning models, our goal is to predict whether SpaceX will reuse the first stage of a launch.

Problems you want to find answers

- How do factors such as payload mass, launch site, number of flights, and orbital parameters impact the likelihood of a successful first stage landing?
- Has the rate of successful landings improved over the years?
- Which algorithm is most effective for binary classification in this scenario?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering Data
  - Dealing With Missing Values
  - Using one Hot Encoding To prepare the data for Binary Classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, Tuning and evaluation of the classification models to ensure that we get the best results

Data Collection

Data Collection Process

The data collection process utilized a combination of API requests from the SpaceX REST API and web scraping from a table on SpaceX’s Wikipedia page. Both methods were necessary to compile a complete dataset for a more comprehensive analysis of the launches.

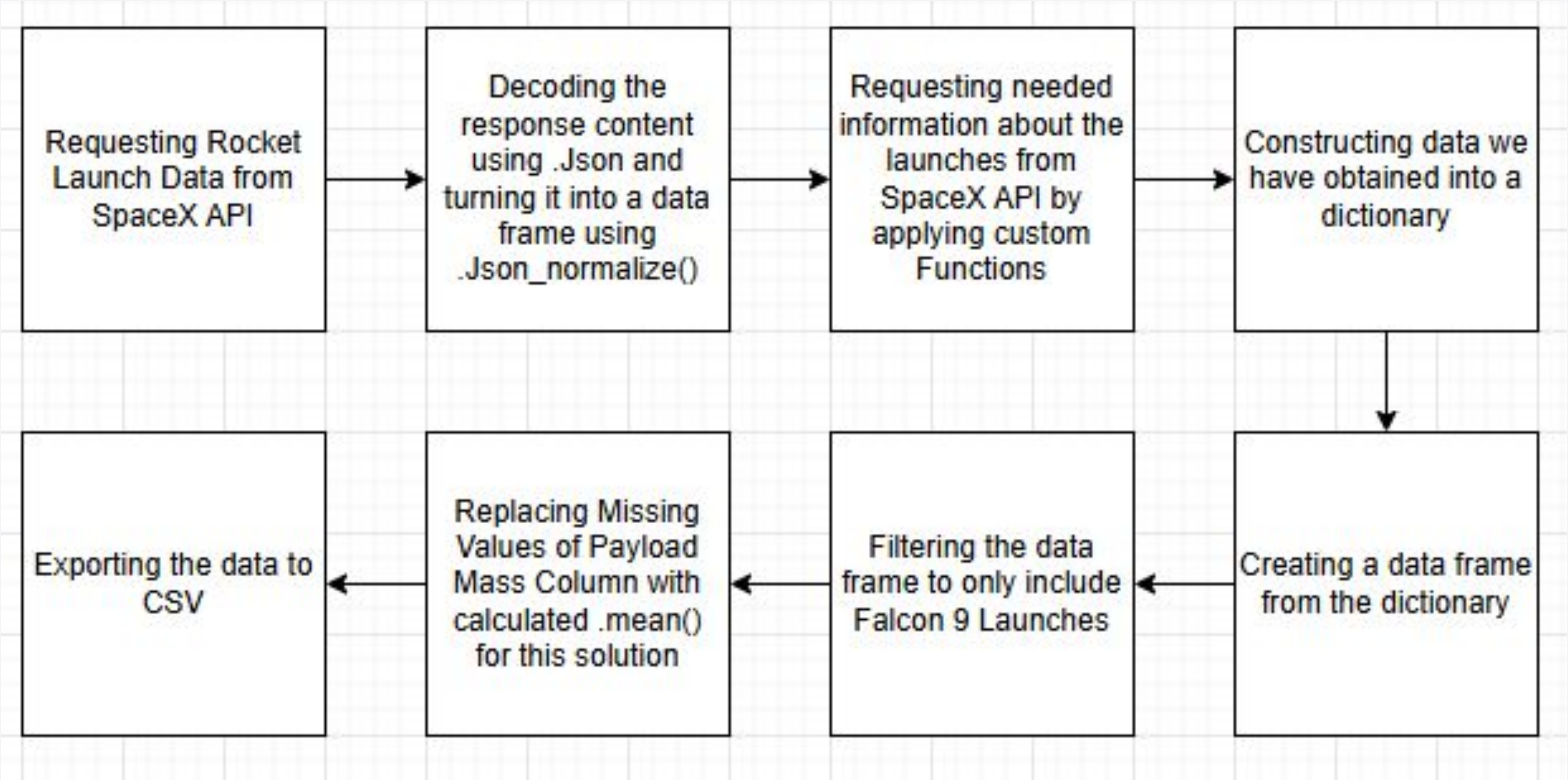
Data Columns Retrieved via SpaceX REST API:

The columns obtained include FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

Data Columns Retrieved via Wikipedia Web Scraping:

The columns obtained include Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, and Time.

Data Collection – SpaceX API



Requesting Rocket Launch Data from SpaceX API

Decoding the response content using .Json and turning it into a data frame using .Json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom Functions

Constructing data we have obtained into a dictionary

Creating a data frame from the dictionary

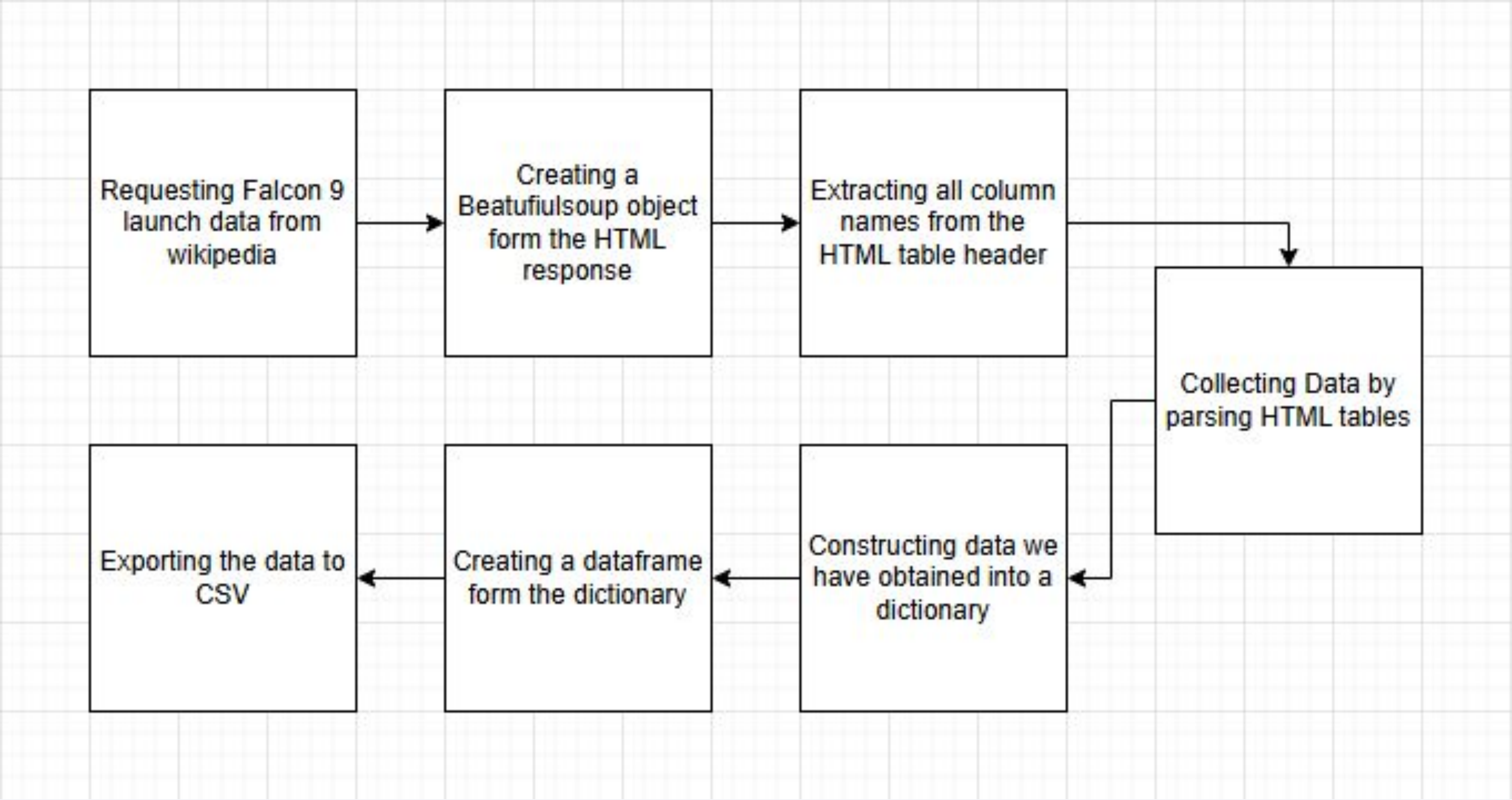
Filtering the data frame to only include Falcon 9 Launches

Replacing Missing Values of Payload Mass Column with calculated .mean() for this solution

Exporting the data to CSV

Data Collection – SpaceX API - Github

Data Collection - Scraping



Requesting Falcon 9 launch data from wikipedia

Creating a Beatufiulsoup object form the HTML response

Extracting all column names from the HTML table header

Collecting Data by parsing HTML tables

Constructing data we have obtained into a dictionary

Creating a dataframe form the dictionary

Exporting the data to CSV

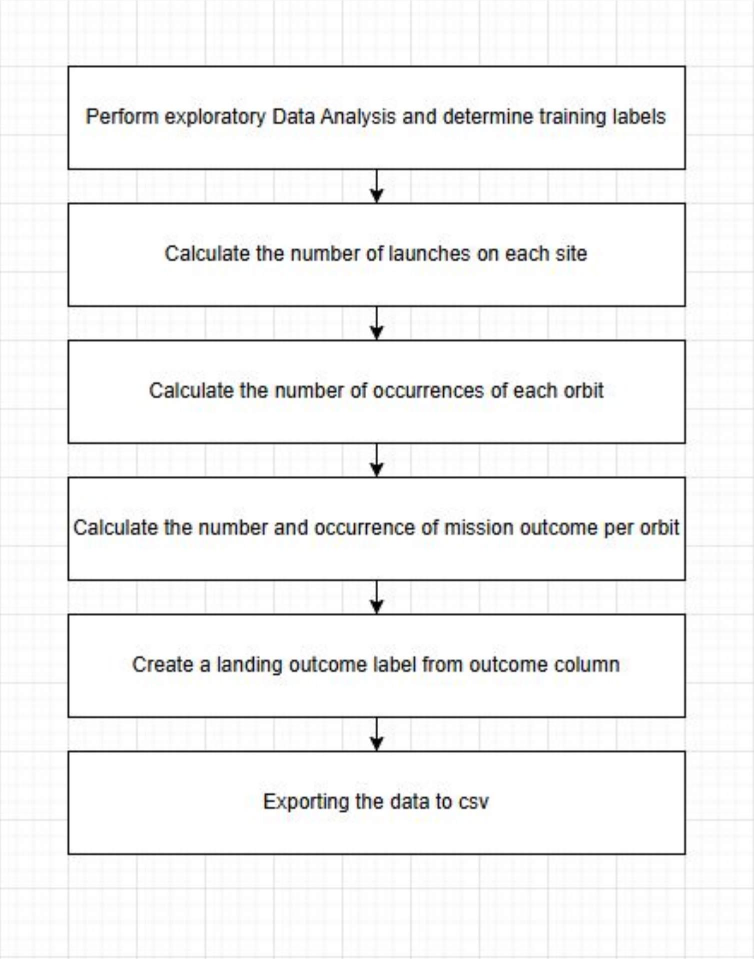
Data Collection - Scraping

Data Wrangling

In the dataset, there are multiple scenarios where the booster did not successfully land. Occasionally, a landing attempt failed due to an accident. For instance, "True Ocean" indicates that the mission outcome was a successful landing in a designated region of the ocean, while "False Ocean" means the landing attempt in the ocean region was unsuccessful. Similarly, "True RTLS" refers to a successful landing on a ground pad, whereas "False RTLS" indicates an unsuccessful attempt on a ground pad. "True ASDS" signifies a successful landing on a drone ship, while "False ASDS" denotes a failed attempt on a drone ship.

To simplify the outcomes, they are converted into training labels: "1" represents a successful booster landing, and "0" represents an unsuccessful attempt.

[Data Wrangling]



Perform exploratory Data Analysis and determine training labels

Calculate the number of launches on each site

Calculate the number of occurrences of each orbit

Calculate the number and occurrence of mission outcome per orbit

Create a landing outcome label from outcome column

Exporting the data to csv

EDA with Data Visualization

The following relationships were analyzed through visualizations:

- Flight Number and Payload Mass
- Flight Number and Launch Site
- Payload Mass and Launch Site
- Orbit Type and Success Rate
- Flight Number and Orbit Type
- Payload Mass and Orbit Type
- Yearly Trend of Success Rate

Purpose of the Chart Types:

- **Scatter Plots:** These reveal relationships between variables, helping to identify patterns that could be useful for machine learning models.
- **Bar Charts:** These are used to compare discrete categories and highlight the connections between categories and specific measured values.
- **Line Charts:** These depict trends over time, focusing on time series data.

EDA with Data Visualization

EDA with SQL

Exploratory Data Analysis (EDA) with SQL

The following SQL queries were performed to analyze the data:

Retrieved the names of unique launch sites used in space missions.

Displayed five records where launch site names start with the string "CCA".

Calculated the total payload mass carried by boosters launched for NASA (CRS missions).

Displayed the average payload mass carried by booster version F9 v1.1.

Identified the date of the first successful landing on a ground pad.

Listed the boosters that successfully landed on a drone ship and carried a payload mass greater than 4000 but less than 6000.

Calculated the total number of successful and failed mission outcomes.

Identified the booster versions that carried the maximum payload mass.

Listed failed drone ship landings, their booster versions, and launch site names for the year 2015.

Ranked landing outcomes (e.g., Failure on a drone ship, Success on a ground pad) by count, within the date range 2010-06-04 to 2017-03-20, in descending order.

EDA with SQL

Build an Interactive Map with Folium

Launch Sites Dropdown List:

A dropdown list was added to enable the selection of specific launch sites.

Pie Chart Displaying Success Launches (All Sites/Specific Site):

A pie chart was created to display the total count of successful launches across all sites. If a specific launch site is selected, the chart shows the breakdown of successful versus failed launches for that site.

Slider for Payload Mass Range:

A slider was implemented to allow users to select a specific range for payload mass.

Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:

A scatter chart was added to visualize the correlation between payload mass and launch success for various booster versions.

Build an Interactive Map with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

A dropdown list was added to enable the selection of specific launch sites.

Pie Chart Displaying Success Launches (All Sites/Specific Site):

A pie chart was created to display the total count of successful launches across all sites. If a specific launch site is selected, the chart shows the breakdown of successful versus failed launches for that site.

Slider for Payload Mass Range:

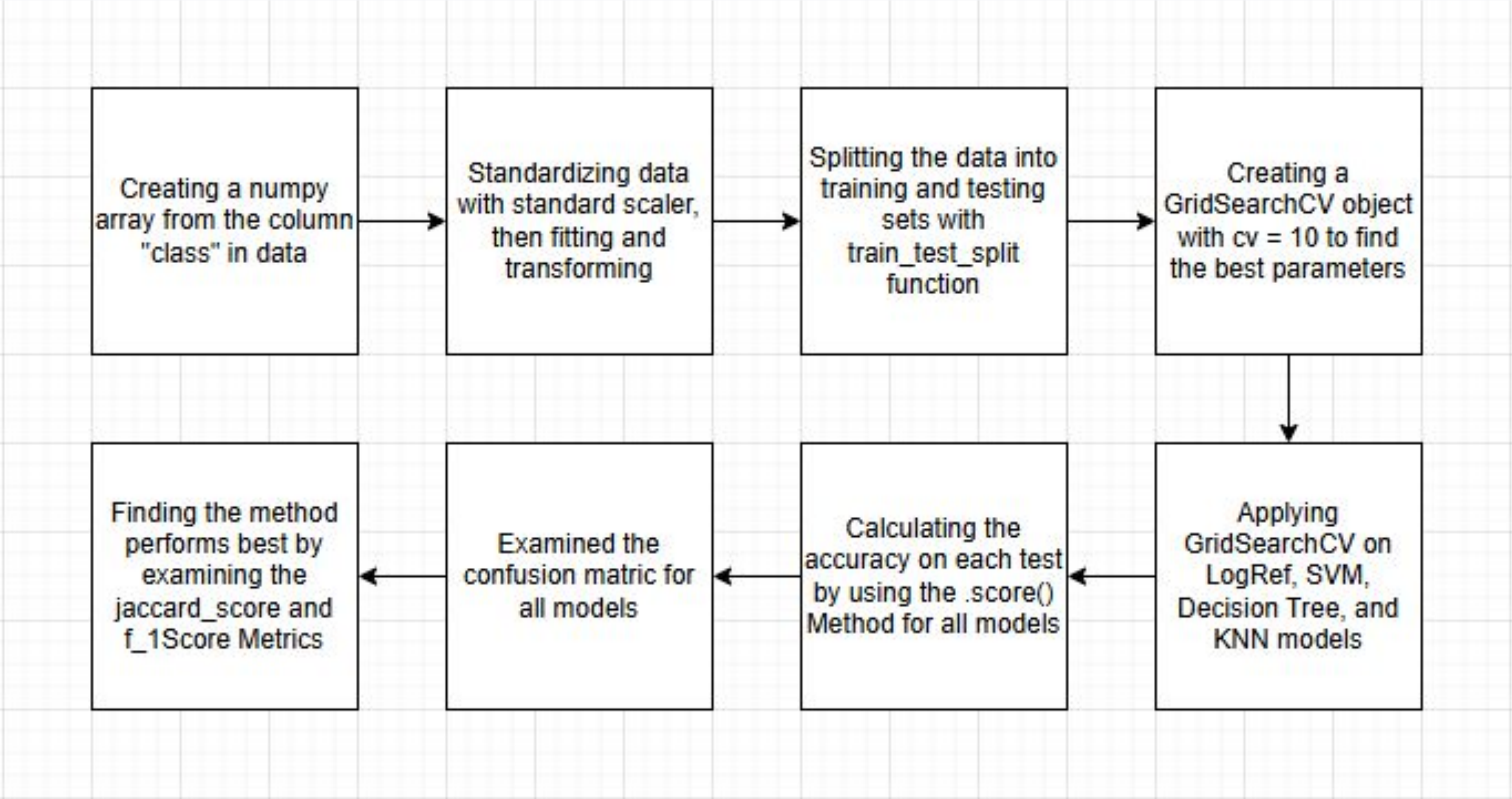
A slider was implemented to allow users to select a specific range for payload mass.

Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:

A scatter chart was added to visualize the correlation between payload mass and launch success for various booster versions.

[Build a Dashboard with Plotly Dash]

Predictive Analysis (Classification)



Creating a numpy array from the column "class" in data

Standardizing data with standard scaler, then fitting and transforming

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogRef, SVM, Decision Tree, and KNN models

Calculating the accuracy on each test by using the .score() Method for all models

Examined the confusion matric for all models

Finding the method performs best by examining the jaccard_score and f_1Score Metrics

Predictive Analysis (Classification)

Results

Exploratory Data Analysis Results:

- Higher payload mass correlates with lower success rates in certain orbits.
- Success rates have improved over time, with specific launch sites and booster versions performing better.

Interactive Analytics Demo:

- Dashboard features include a launch site filter, payload range slider, pie chart for success/failure ratios, and scatter plots showing payload mass vs. success rates.

Predictive Analysis Results:

- The model effectively predicts booster landing success, aiding cost optimization

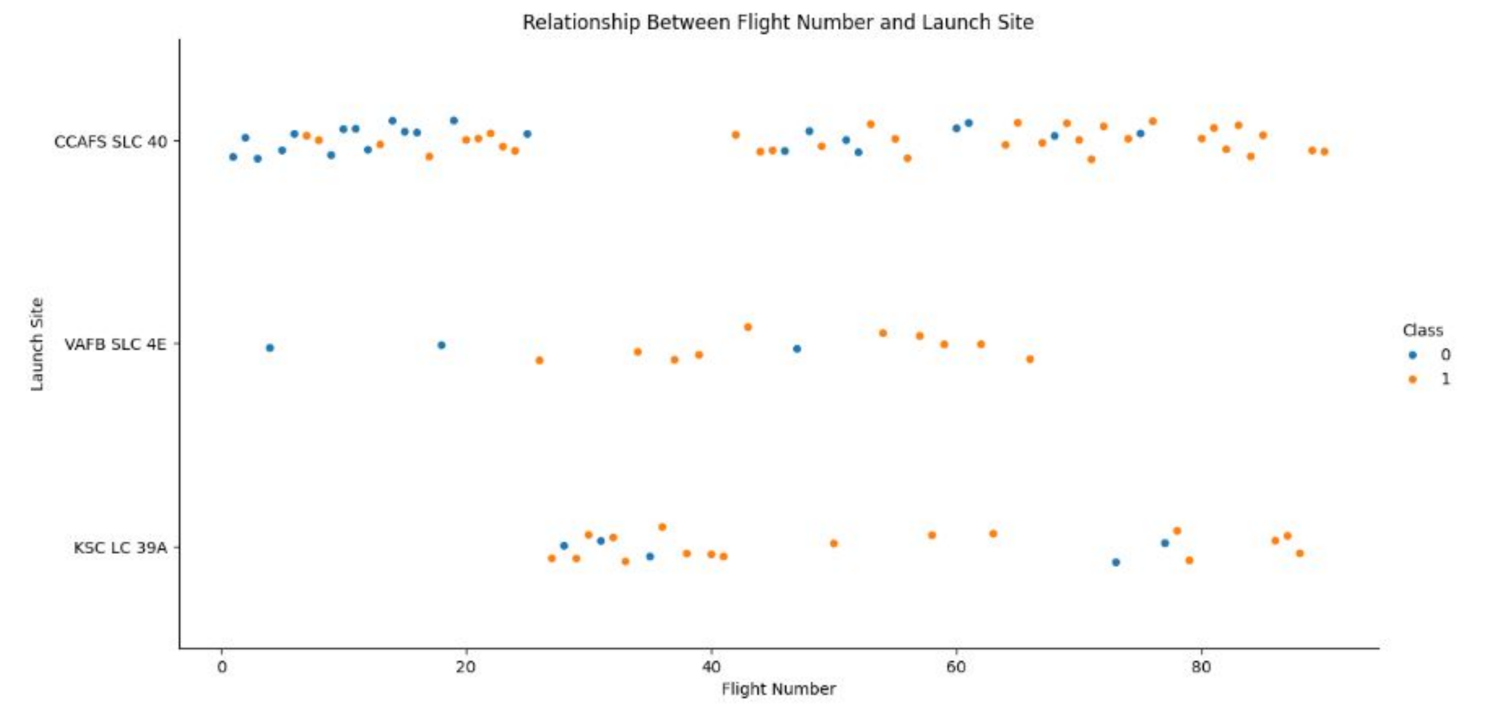
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Explanation:

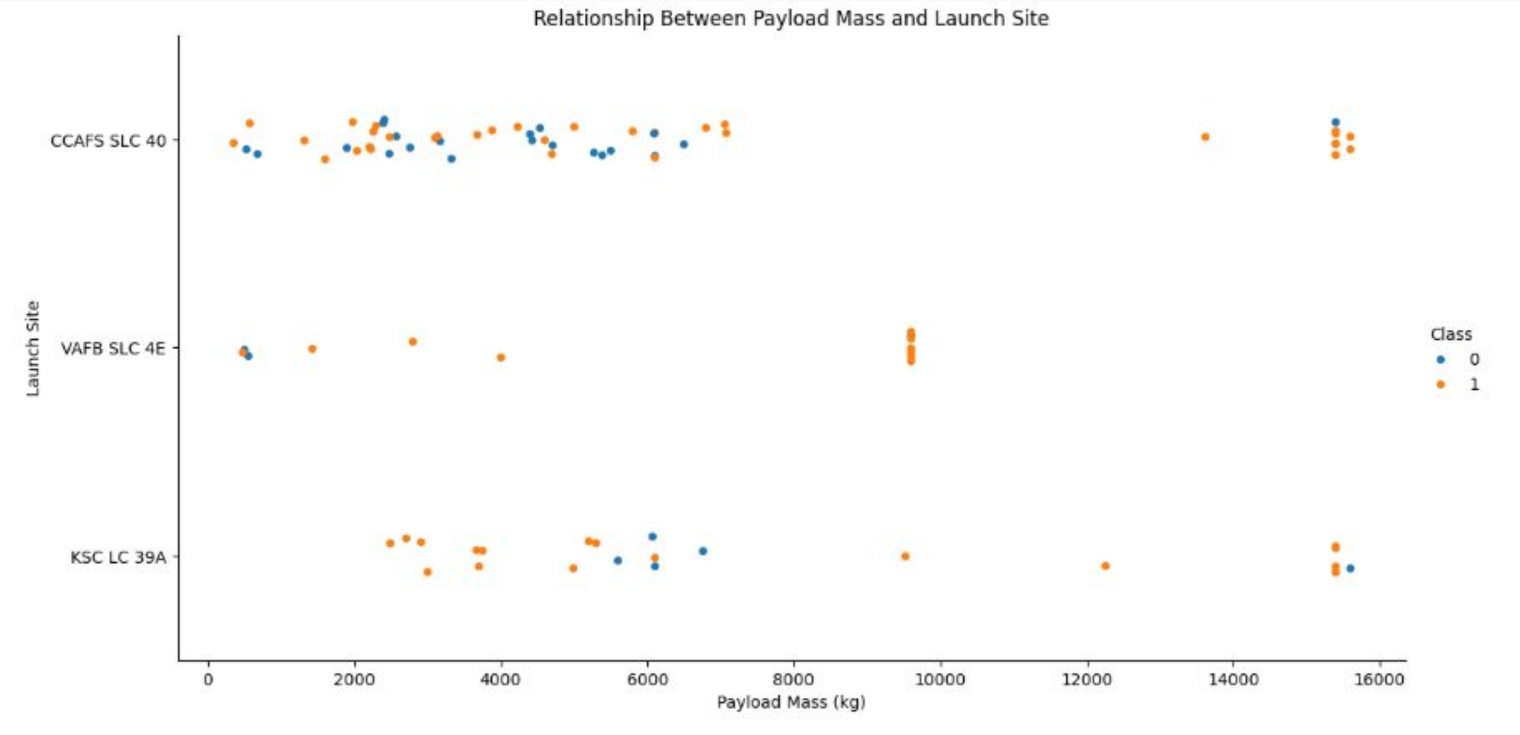
- Early flights failed, while recent ones succeeded.
- CCAFS SLC 40 accounts for nearly half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- Success rates increase with each new launch.



Payload vs. Launch Site

Explanation:

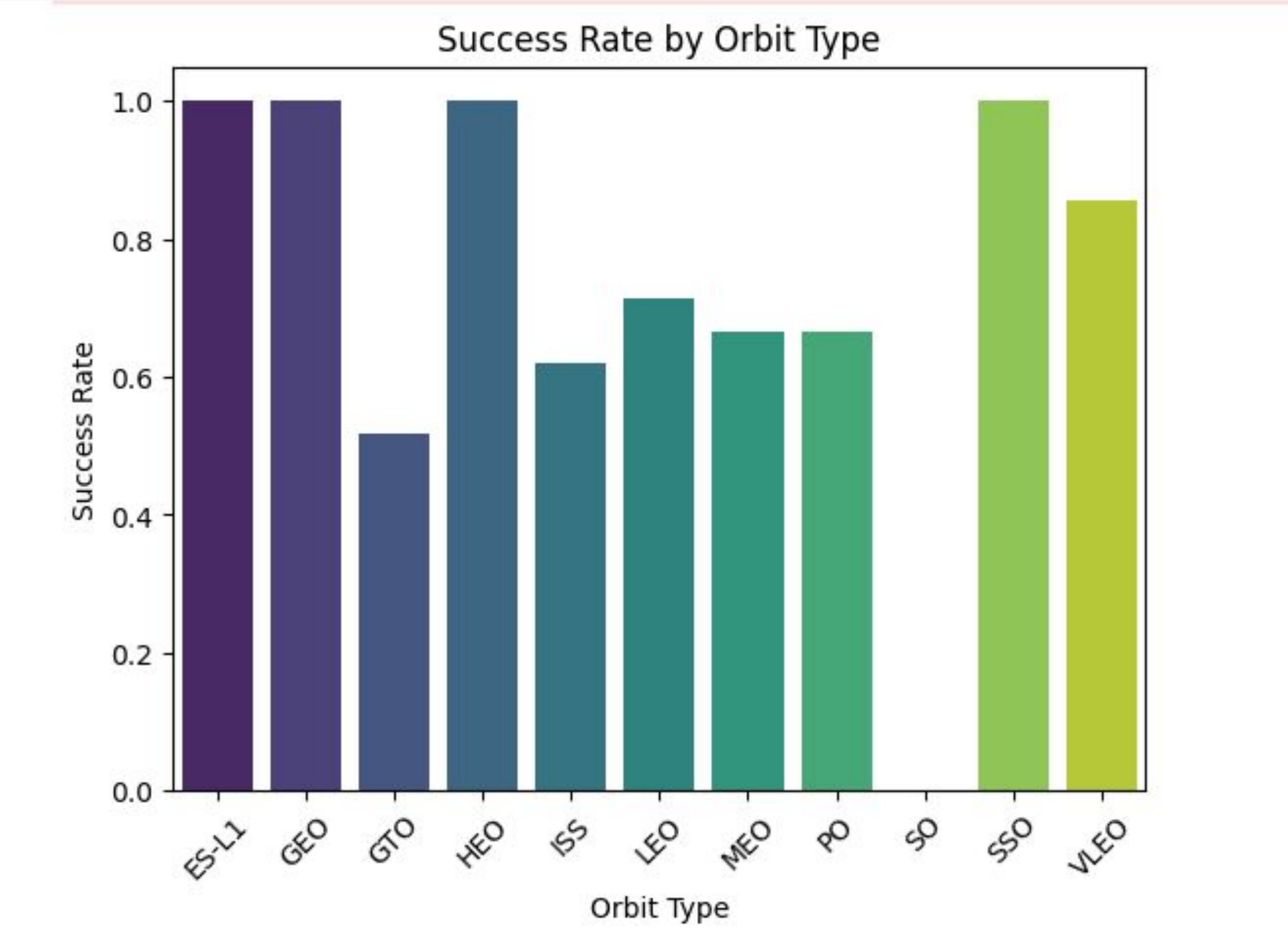
- Higher payload mass generally leads to higher success rates across all launch sites.
- Most launches with payloads over 7000 kg were successful.
- KSC LC 39A achieved a 100% success rate for payloads under 5500 kg.



Success Rate vs. Orbit Type

Explanation:

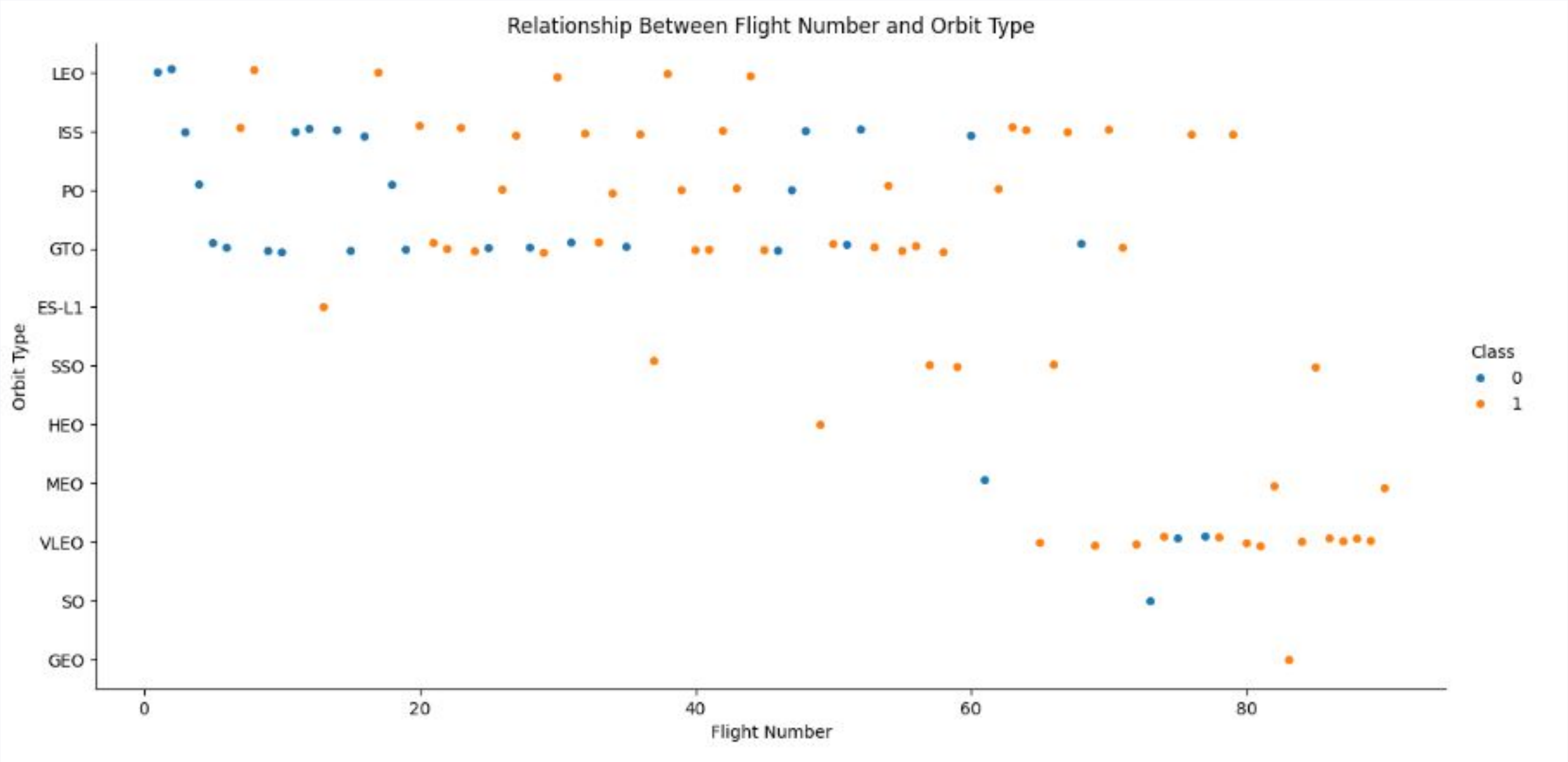
- **100% Success Orbits:** ES-L1, GEO, HEO, SSO
- **0% Success Orbit:** SO
- **50%-85% Success Orbits:** GTO, ISS, LEO, MEO, PO



Flight Number vs. Orbit Type

Explanation:

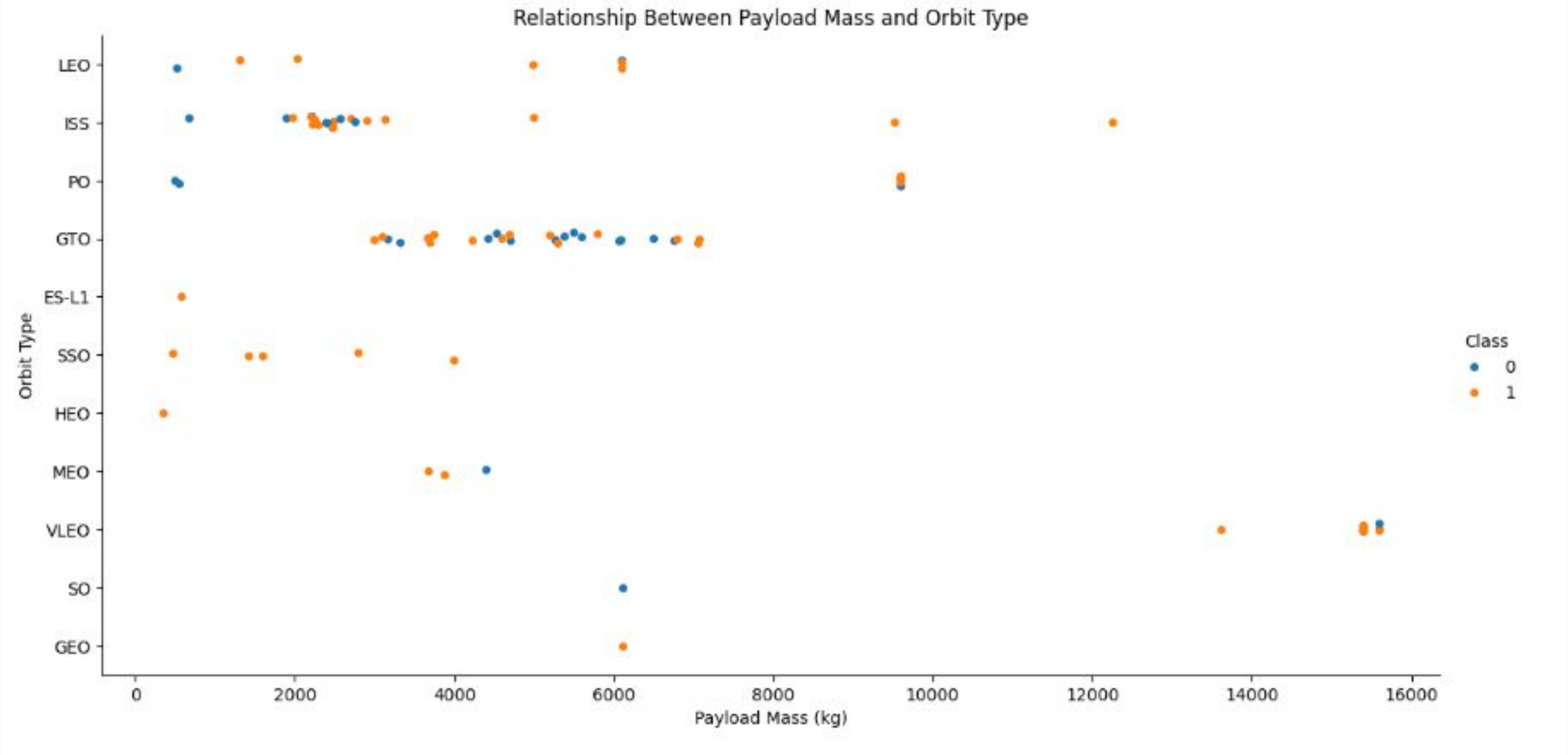
- In the LEO orbit, success seems related to the number of flights.
- In the GTO orbit, no clear relationship exists between success and flight number.



Payload vs. Orbit Type

Explanation:

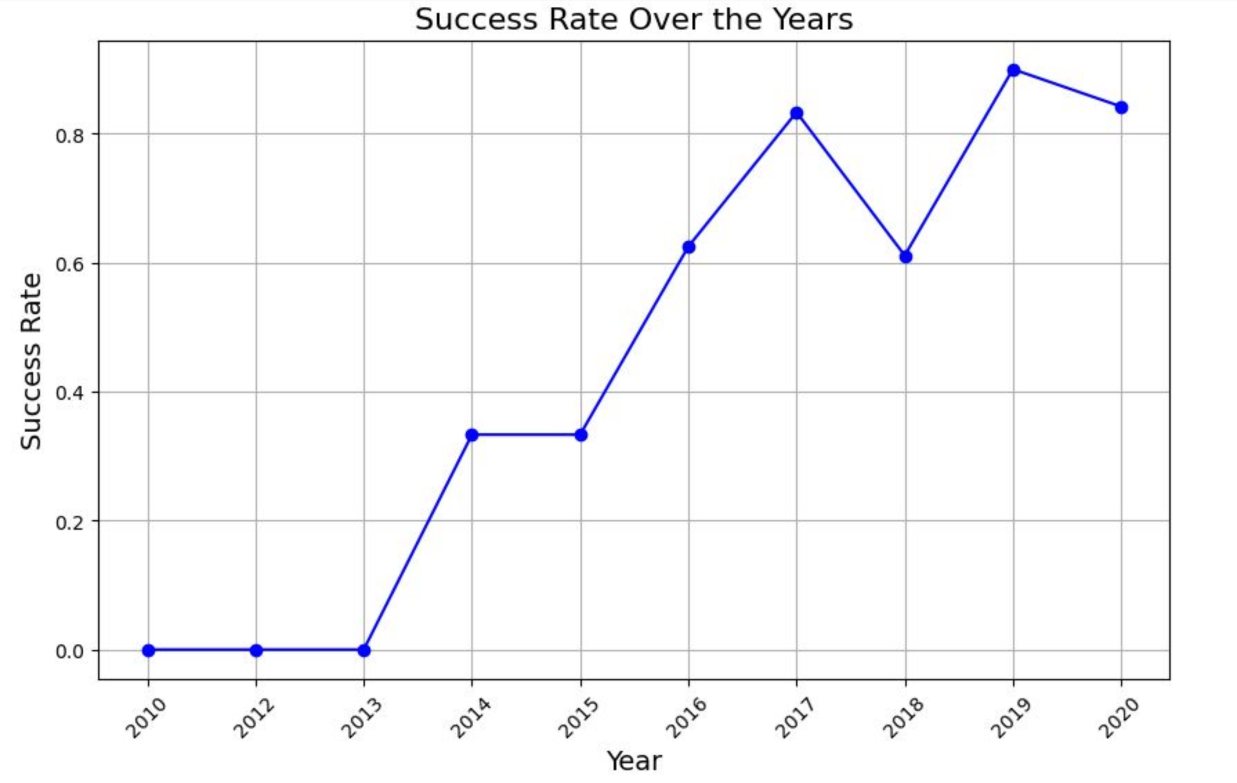
- Heavy payloads negatively impact GTO orbits.
- Heavy payloads positively influence GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

Explanation:

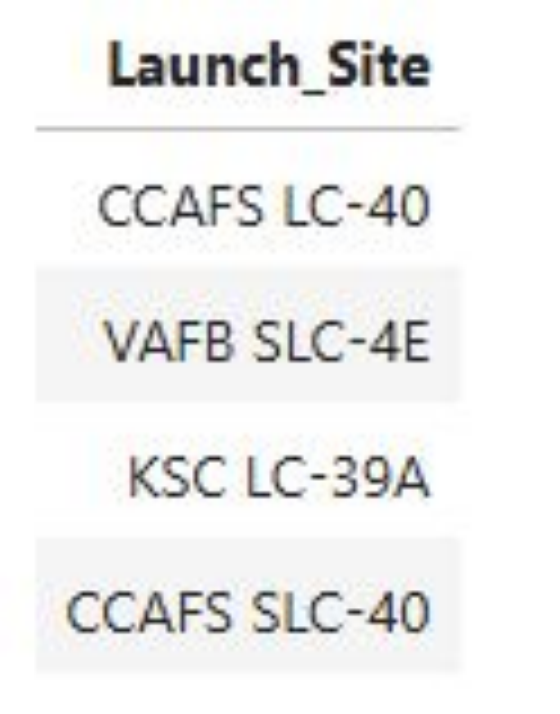
- The success rate steadily increased from 2013 to 2020.



All Launch Site Names

Explanation:

- Displaying all launch site names



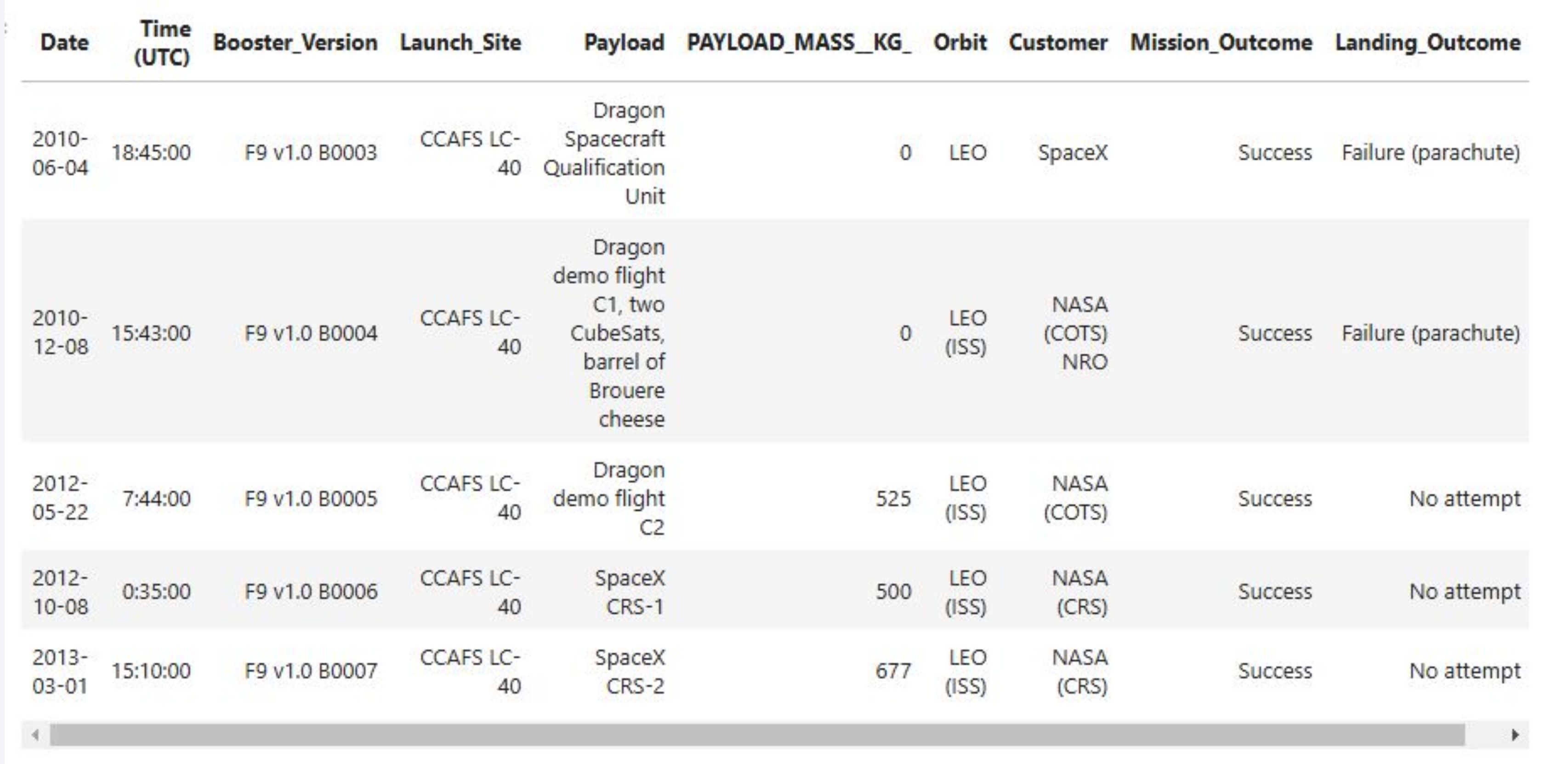
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Explanation:

- Displaying the first 5 records where launch site begins with the string ‘CCA’

-



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

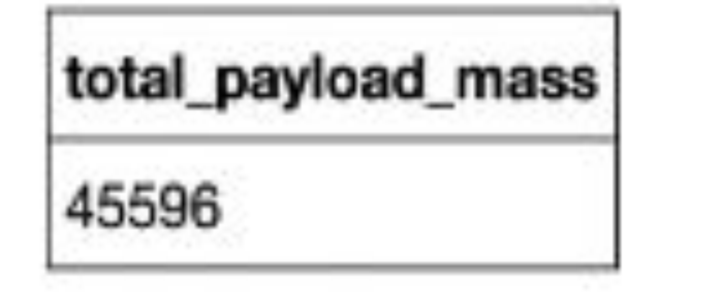
Total Payload Mass

Explanation:

- Calculating the total payload carried by boosters from NASA

Explanation:

- Calculating the total payload carried by boosters from NASA

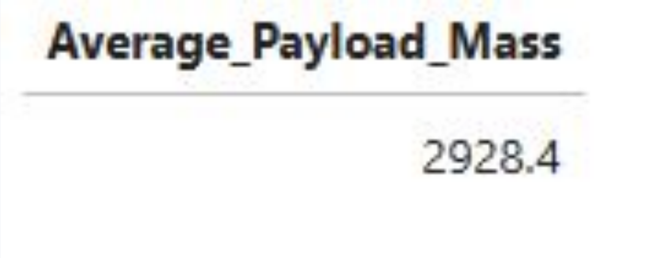


total_payload_mass
45596

Average Payload Mass by F9 v1.1

Explanation:

- Calculating the average payload carried by boosters from F9 V1.1

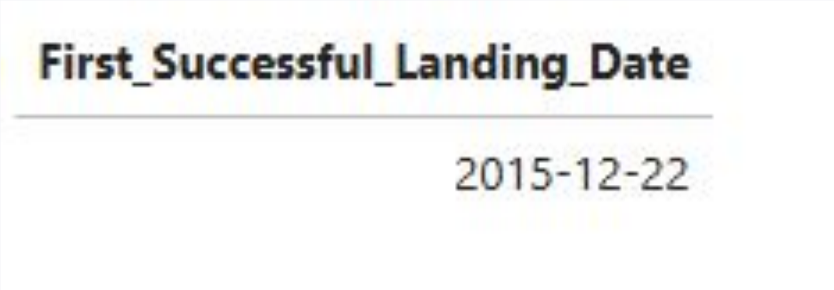


Average_Payload_Mass
2928.4

First Successful Ground Landing Date

Explanation:

- Finding the dates of the first successful landing outcome on ground pad

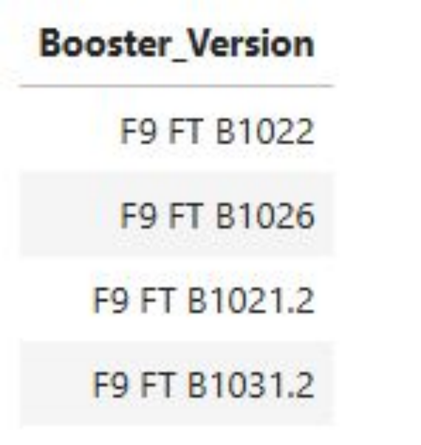


First_Successful_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation:

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

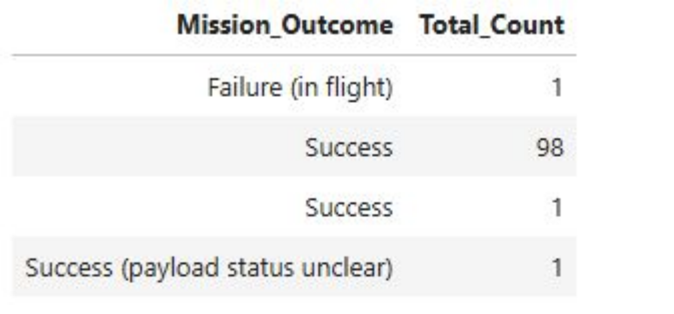


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Explanation:

- Calculate the total number of successful and failure mission outcomes

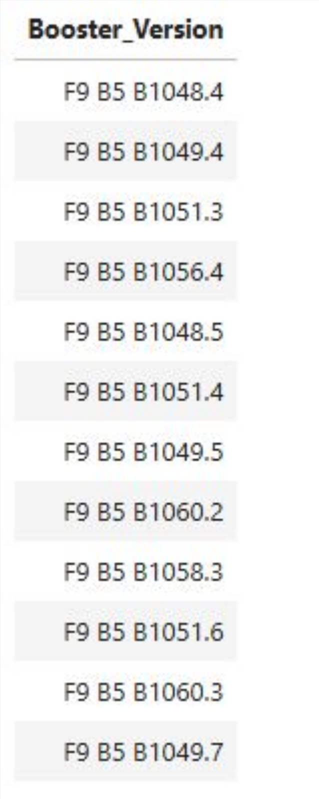


Mission_Outcome	Total_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Explanation:

- List the names of the booster which have carried the maximum payload mass

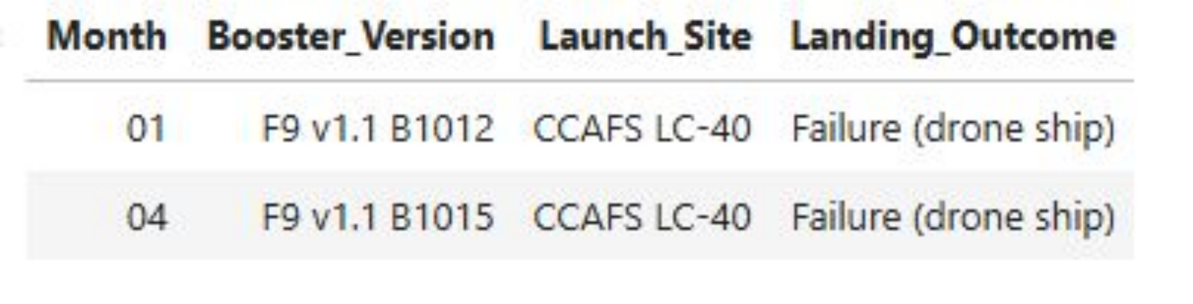


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

Explanation:

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

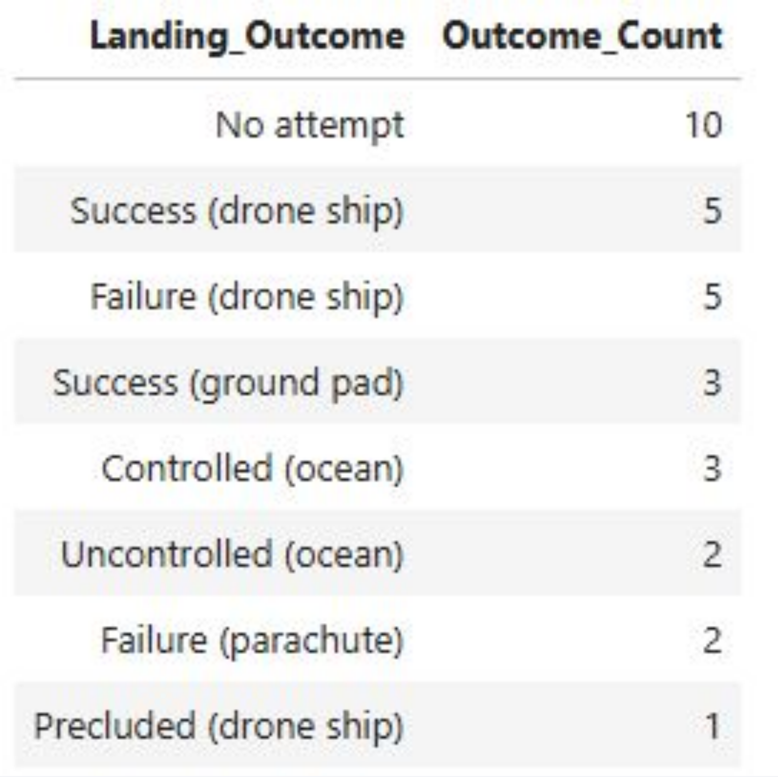


Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Explanation:

- 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

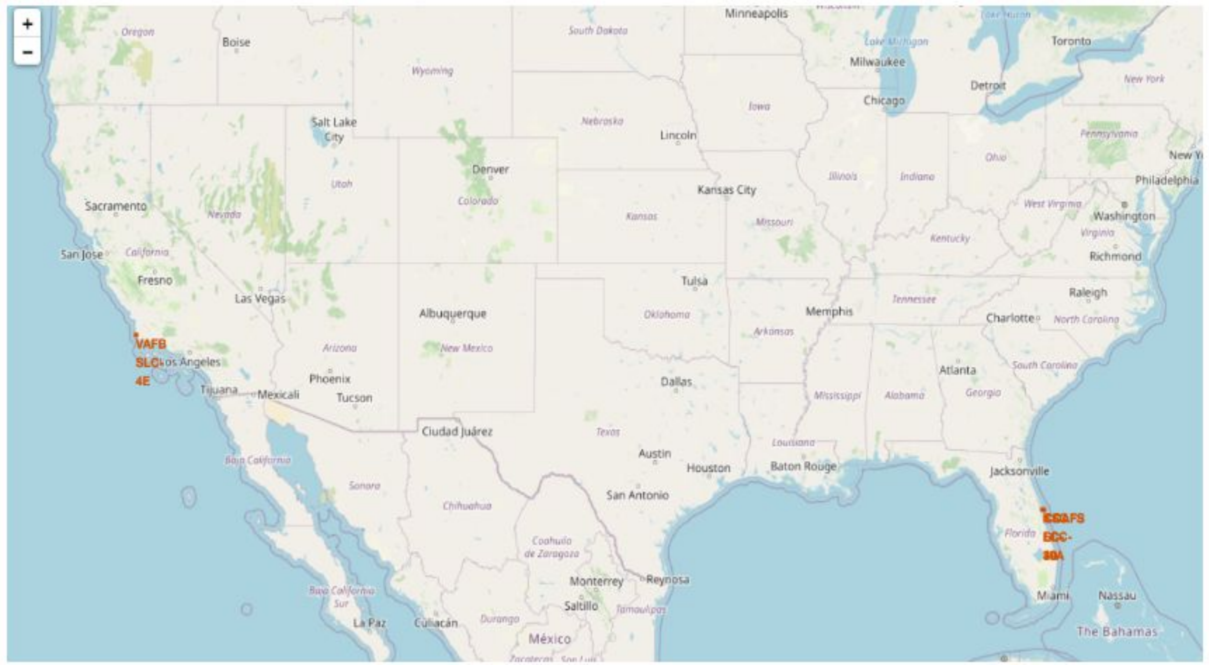


Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

all launch sites’ location markers on a global map



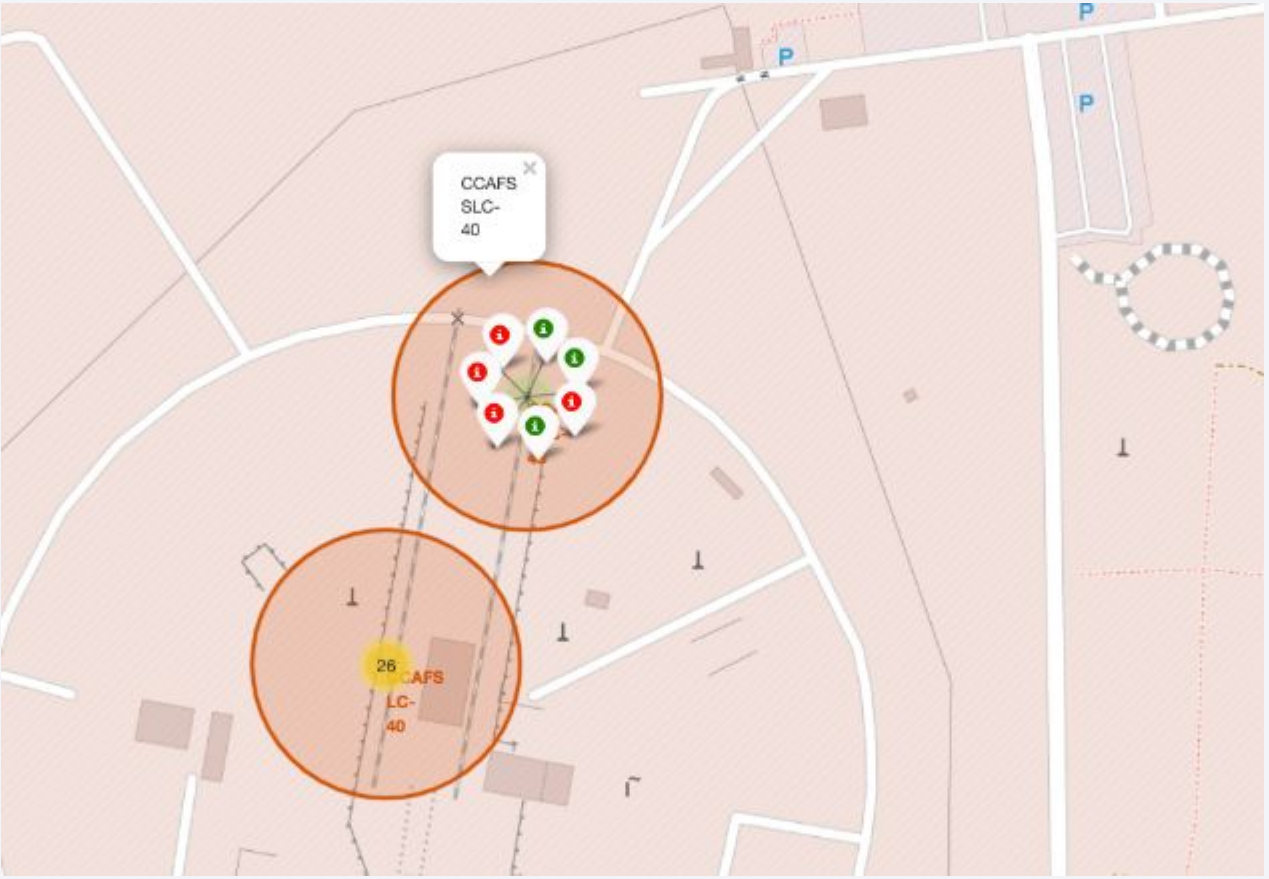
Explanation:

- Most launch sites are near the equator to take advantage of Earth's rotational speed, which provides additional velocity to spacecraft.
- They are also close to the coast to ensure rockets launch over the ocean, reducing risks from debris falling near populated areas.

Color-labeled launch outcomes on the map

Explanation:

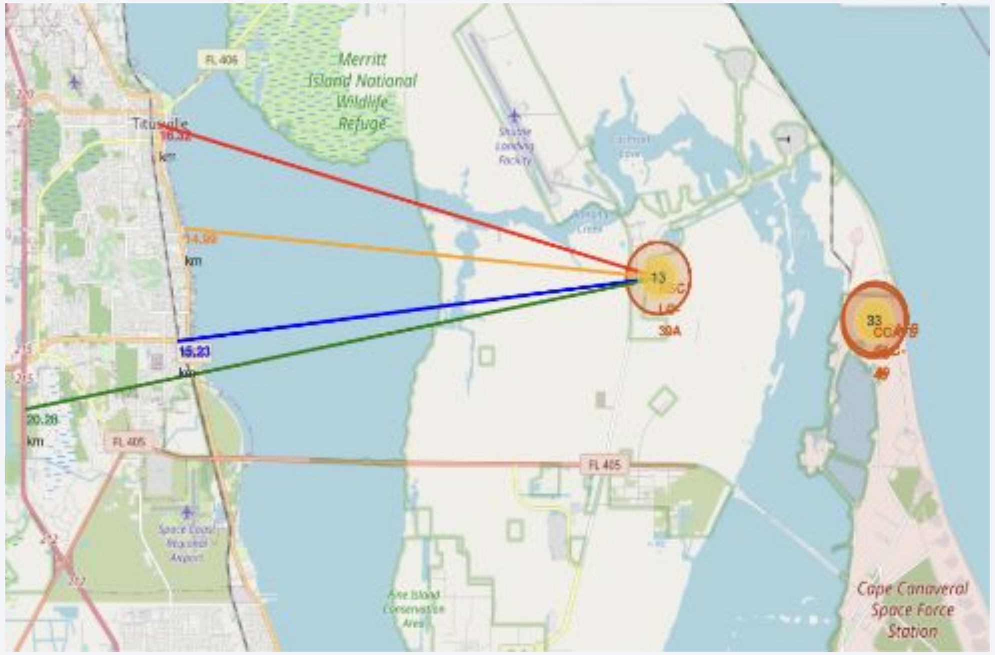
- Color-coded markers (green for success, red for failure) provide a quick visual indication of launch site success rates.
- KSC LC-39A stands out with a very high success rate.



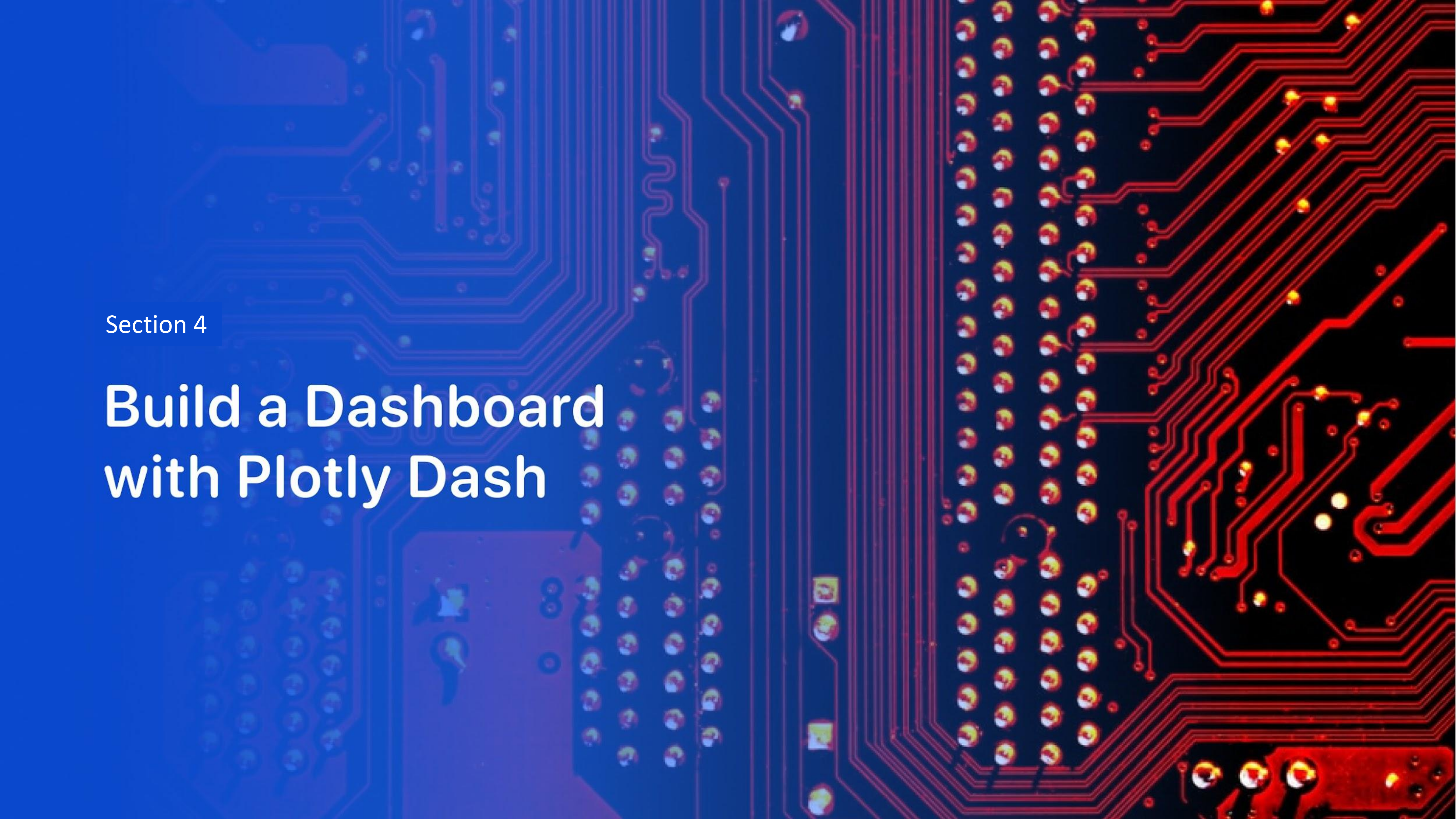
Distance from Launch Site to Proximities

Explanation:

- Launch site KSC LC-39A is:
    - Close to a railway (15.23 km).
    - Near a highway (20.28 km).
    - Close to the coastline (14.99 km).
- Its proximity to the city of Titusville (16.32 km) is notable.
- High-speed rocket failures can pose risks to areas within a 15–20 km range.



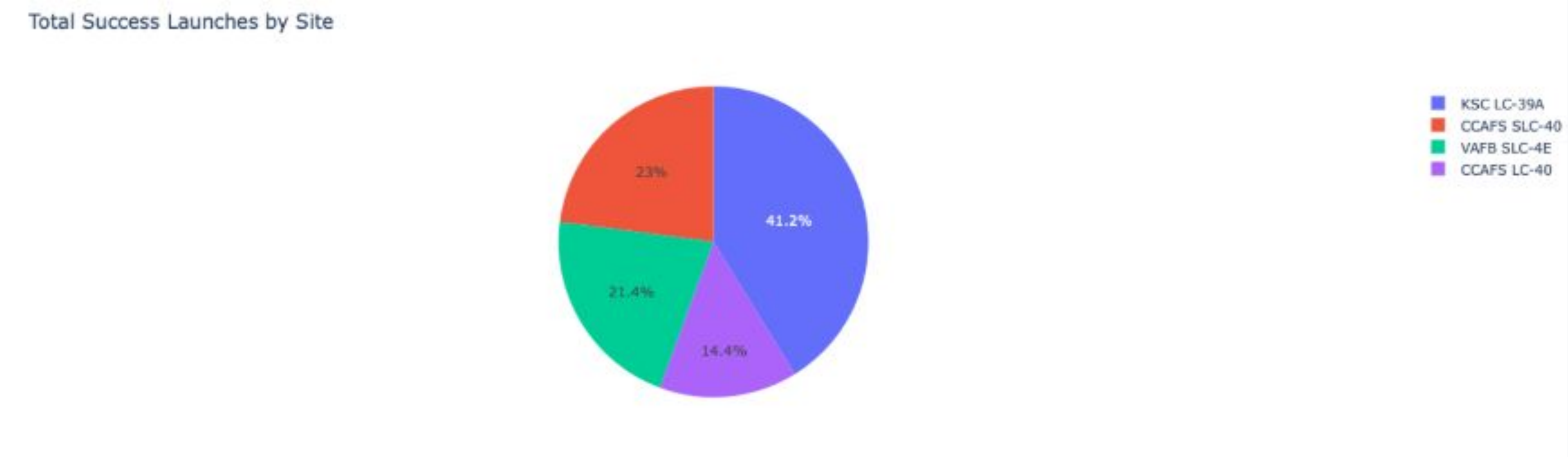
FL 406
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
13
KSC LC-39A
33
CCAFS
FL 405
Cape Canaveral Space Force Station



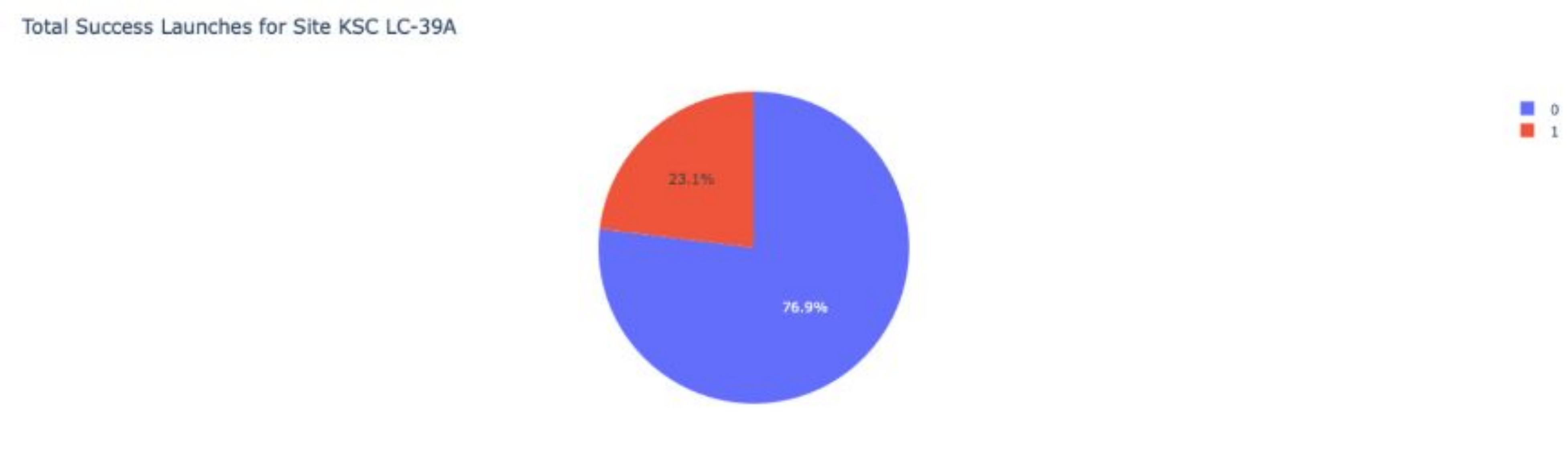
Section 4

Build a Dashboard with Plotly Dash

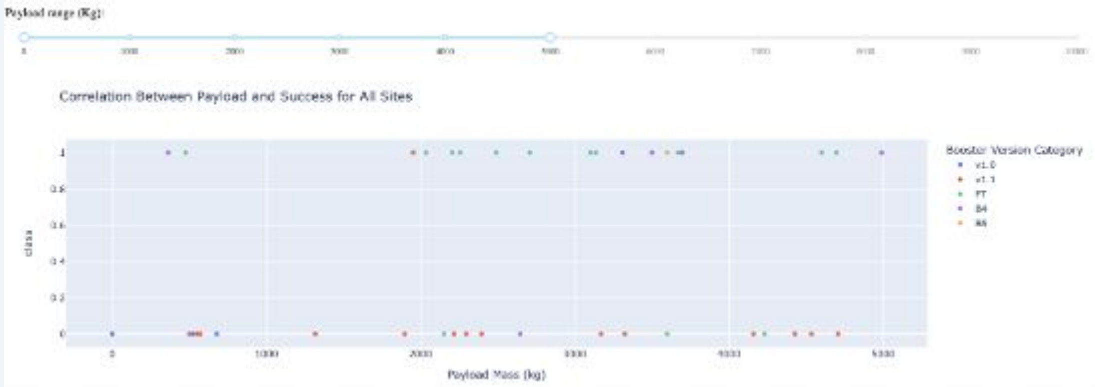
Total Launch Success



Most Successful Launch Site

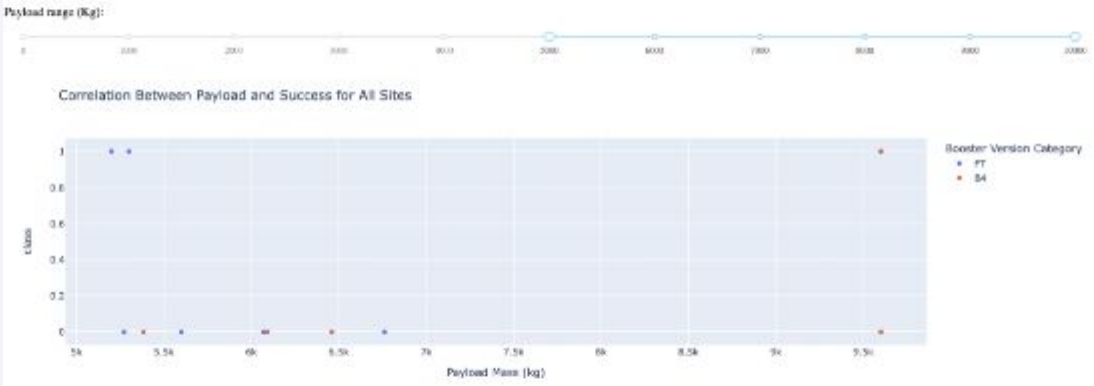


Payload Mass Vs. Launch Outcome from Sites



Payload range (Kg):

Correlation Between Payload and Success for All Sites



Payload range (Kg):

Correlation Between Payload and Success for All Sites

Section 5

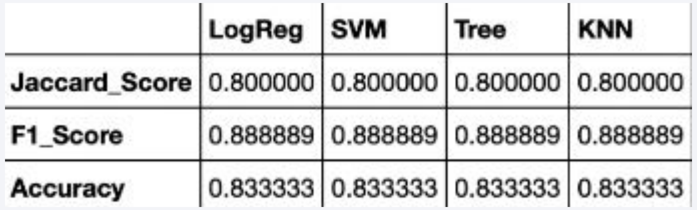
Predictive Analysis (Classification)

Classification Accuracy

Explanation:

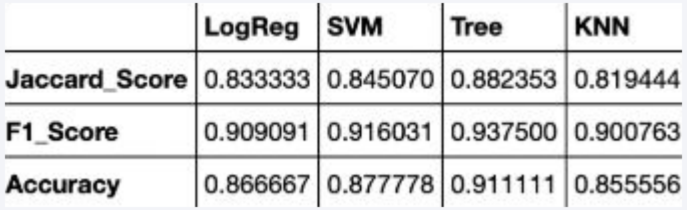
- Test set scores alone cannot confirm the best-performing method due to a small sample size (18 samples).
- Testing on the entire dataset shows the Decision Tree model performs best.
- This model achieves the highest accuracy and scores overall.

Scores & Accuracy of the Test Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores & Accuracy of the Test Set

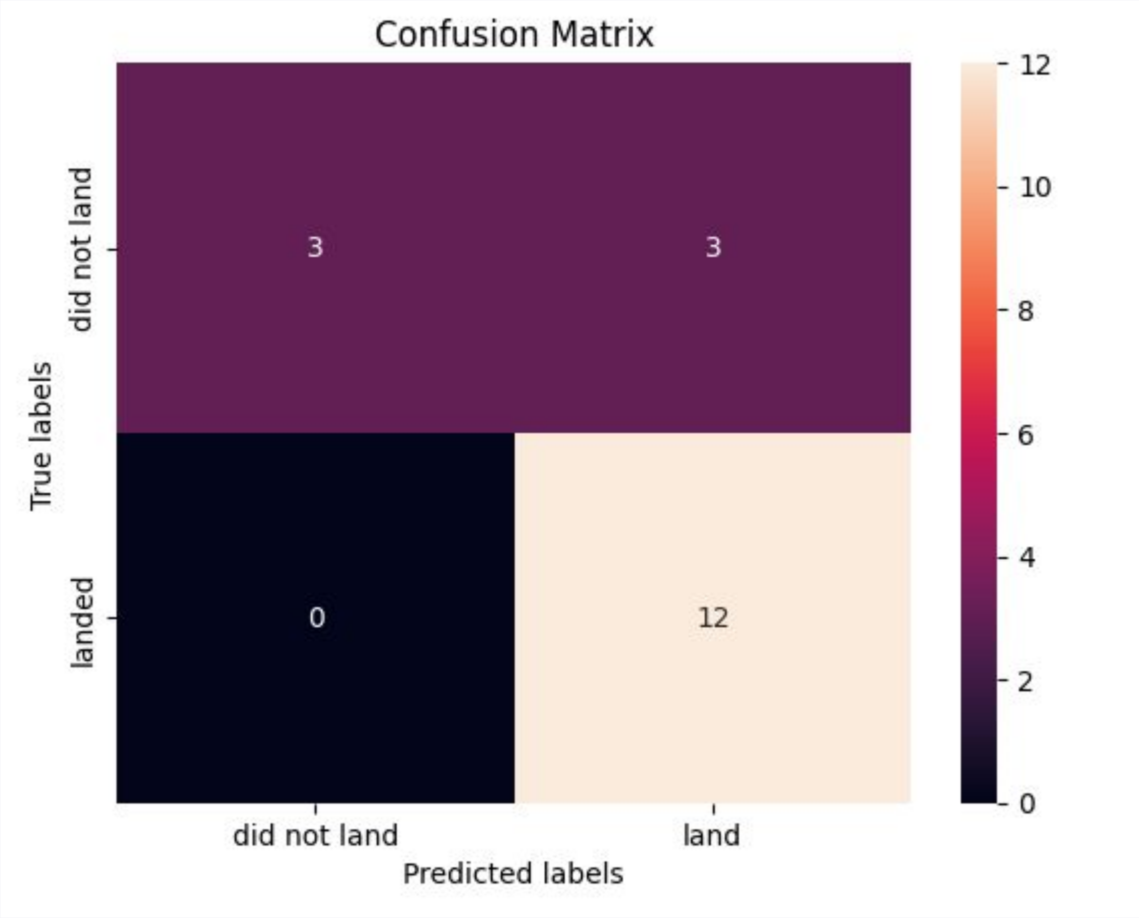


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation:

- The confusion matrix reveals that logistic regression effectively distinguishes between classes.
- However, the primary issue lies in the occurrence of false positives.



Conclusions

In summary, the analysis highlights several key findings:

1. Success rates have improved significantly over time, with newer launches demonstrating better outcomes.
2. Payload mass and orbit type play critical roles in influencing launch success, with certain orbits and payload ranges showing distinct patterns.
3. Launch site proximity to specific geographical features, such as coastlines and populated areas, impacts safety and risk factors.
4. Among predictive models, the Decision Tree Model emerges as the most accurate, but logistic regression highlights challenges with false positives.

These insights provide a solid foundation for optimizing future launches and improving predictive modeling for better decision-making.

Thank you!