IBM Data Science Exam

Jordan Moles

22/05/2024

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results:
  - Data collection and wrangling
  - EDA
  - Predictive analysis
- Conclusion
- Appendix

EXECUTIVE SUMMARY

In this capstone project, we predict the successful landing of SpaceX's Falcon 9 first stage rocket. SpaceX offers launches at 62 million dollars each, significantly less than competitors who charge upwards of 165 million dollars, due to the reusability of the first stage. Accurate predictions of landing success help estimate launch costs, providing critical insights for companies competing against SpaceX.

We collected and formatted data from the SpaceX API and web scraped Falcon 9 launch records from Wikipedia. Exploratory Data Analysis (EDA) revealed patterns, and we engineered features to train supervised models.

We evaluated various models (Logistic Regression, SVM, Decision Tree, KNN) and identified the best one using GridSearchCV, which optimizes hyperparameters by systematically testing multiple combinations. The best model is Decision Tree and demonstrated superior performance across metrics like Jaccard Index, F1-Score, and LogLoss.

Our approach ensures the selected model and its hyperparameters are optimized, leading to reliable predictions of Falcon 9 first stage landings and valuable insights for cost estimation and competitive bidding.

INTRODUCTION

The commercial space industry has seen significant advancements, with SpaceX leading the charge in reducing launch costs through innovative technologies. One of the key factors contributing to SpaceX's cost efficiency is the reusability of the Falcon 9 rocket's first stage. SpaceX advertises the cost of a Falcon 9 launch at 62 million dollars, which is substantially lower than the 165 million dollars charged by other providers. This price reduction is largely due to the ability to successfully land and reuse the first stage of the rocket.

Predicting the successful landing of the Falcon 9 first stage is crucial for estimating launch costs accurately. This prediction can provide critical insights for companies looking to compete against SpaceX for rocket launch contracts. Understanding the factors that influence the success of these landings can also inform improvements in launch technology and operations.

INTRODUCTION

In this capstone project, we aim to predict the landing success of the Falcon 9 first stage using various classification algorithms. We will collect data from the SpaceX API and supplement it with additional information obtained through web scraping Falcon 9 launch records from Wikipedia. The collected data will undergo thorough Exploratory Data Analysis (EDA) to identify patterns and trends. We will then engineer features to train our predictive models.

The project will involve:

- Data Collection: Gathering data from the SpaceX API and web scraping launch records.
- Data Wrangling: Cleaning and formatting the data for analysis.
- Exploratory Data Analysis: Using data visualization techniques to uncover patterns.
- Feature Engineering: Creating features that will be used in predictive models.
- Model Training and Evaluation: Training various classification algorithms and evaluating their performance using metrics such as accuracy, Jaccard Index, F1-Score, and LogLoss.

Our goal is to build a model that can reliably predict the success of Falcon 9 first stage landings. By using GridSearchCV, we will optimize the hyperparameters of our models to ensure the best performance. The results of this project will not only aid in cost estimation but also provide valuable insights for companies in the competitive space launch market.

METHODOLOGY

- Data collection
  - SpaceX Rest API
  - Web scrapping from Wikipedia
- Data wrangling
  - Filtering the data
  - Dealing with missing values
  - Encoding the data
- Exploratory data analysis
  - With visualization
  - With SQL
- Interactive visual analytics
  - With Folium
  - With Plotly Dash
- Predictive data analysis
  - Modeling, optimizing and evaluating classification models

DATA COLLECTION

Objective: To gather comprehensive data on Falcon 9 rocket launches to predict the success of first stage landings

- SpaceX API Data Collection:
  - Source: SpaceX REST API
  - Data Collected: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Methods: Sending HTTP requests to SpaceX API endpoints Parsing JSON responses to extract relevant data Storing data in a structured format (e.g., Pandas DataFrame)
- Web Scraping Wikipedia:
  - Source: Wikipedia page on Falcon 9 launch records
  - Data Collected: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
  - Methods: Using BeautifulSoup for HTML parsing Identifying and extracting the HTML table containing launch data Converting the HTML table into a Pandas DataFrame for analysis

DATA WRANGLING

Objective: To prepare the collected dataset for analysis and modeling by addressing various data quality issues and enhancing its suitability for machine learning tasks.

- Cleaning and Formatting:
  - Handling missing values
  - Correcting data types
- Feature Engineering: Creating binary labels for landing outcomes (1 for success, 0 for failure) Generating new features from existing data to improve model performance
  - True ASDS, True RTLS, & True Ocean gives 1
  - None None, False ASDS, None ASDS, False Ocean, False RTLS gives 0

EDA WITH SQL

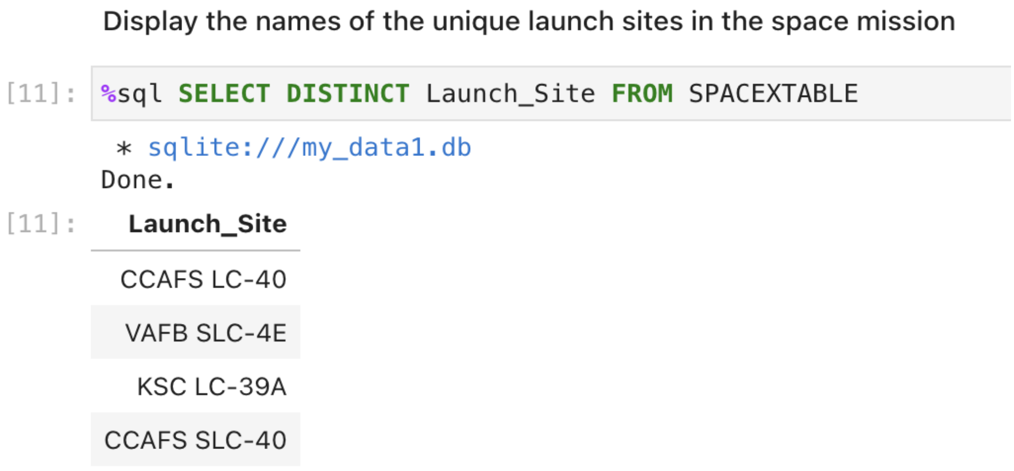
Objective: To Utilize SQL queries to generate insightful visualizations for data analysis and exploration.

List of requests:

- 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 succesful landing outcome in ground pad was acheived
- 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.
- 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

EDA WITH SQL

Objective: To Utilize SQL queries to generate insightful visualizations for data analysis and exploration.



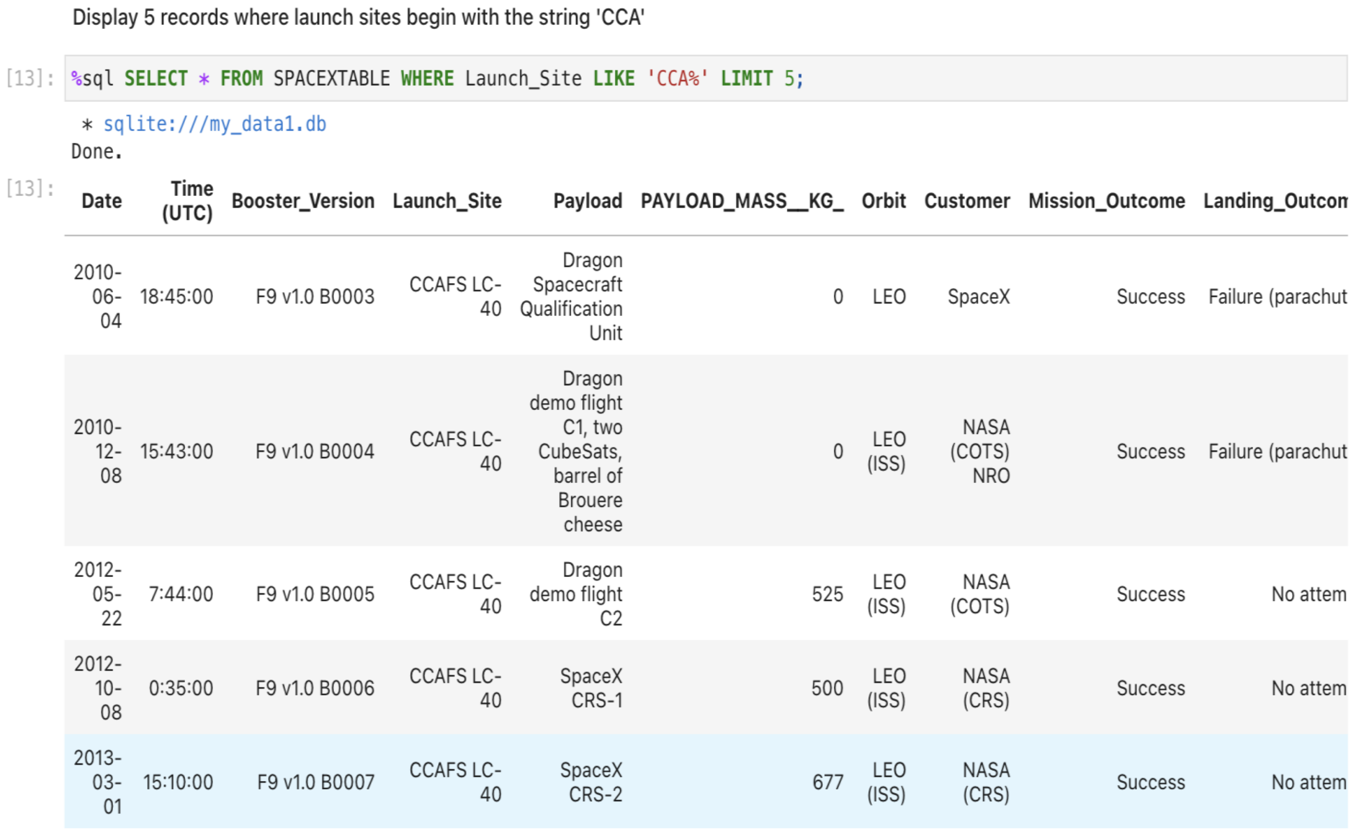
Display the names of the unique launch sites in the space mission

```
[11]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE

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

[11]:

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



Display 5 records where launch sites begin with the string 'CCA'

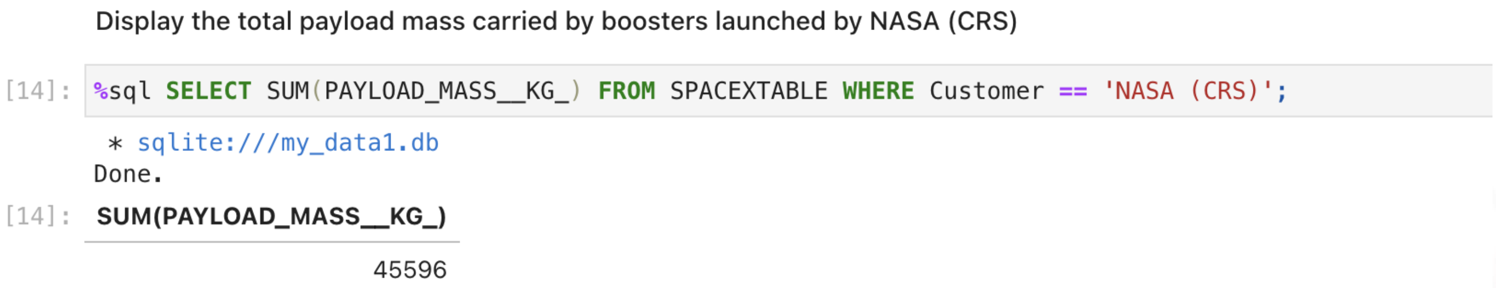
```
[13]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

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

[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcon
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
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 (parachut
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

EDA WITH SQL

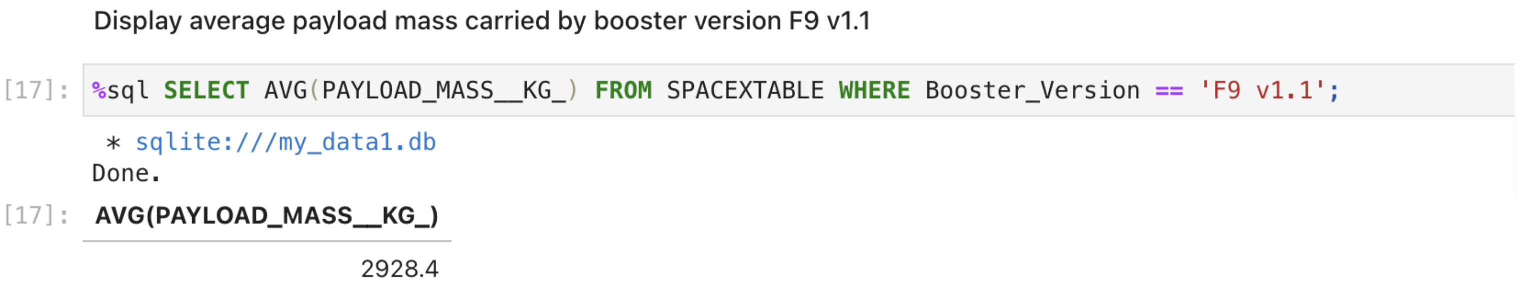


Display the total payload mass carried by boosters launched by NASA (CRS)

```
[14]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.

[14]: SUM(PAYLOAD_MASS__KG_)
                      45596
```

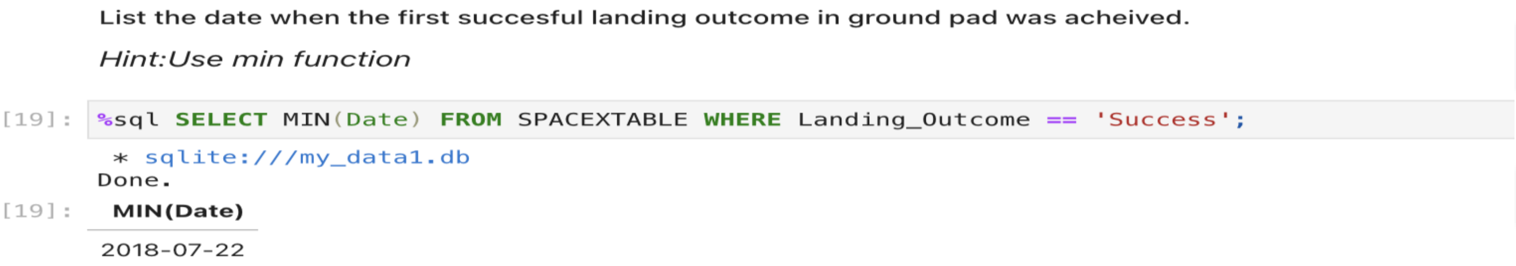


Display average payload mass carried by booster version F9 v1.1

```
[17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version == 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

[17]: AVG(PAYLOAD_MASS__KG_)
                     2928.4
```



List the date when the first succesful landing outcome in ground pad was acheived.

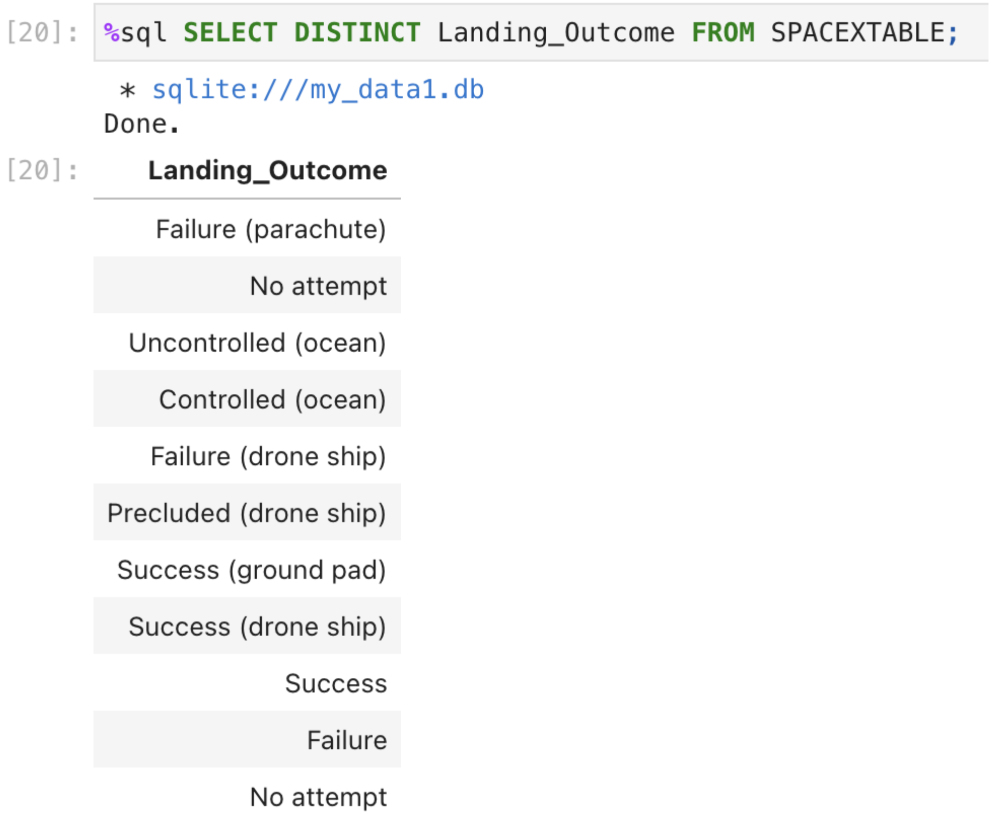
Hint:Use min function

```
[19]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome == 'Success';

 * sqlite:///my_data1.db
Done.

[19]:  MIN(Date)
      2018-07-22
```

EDA WITH SQL

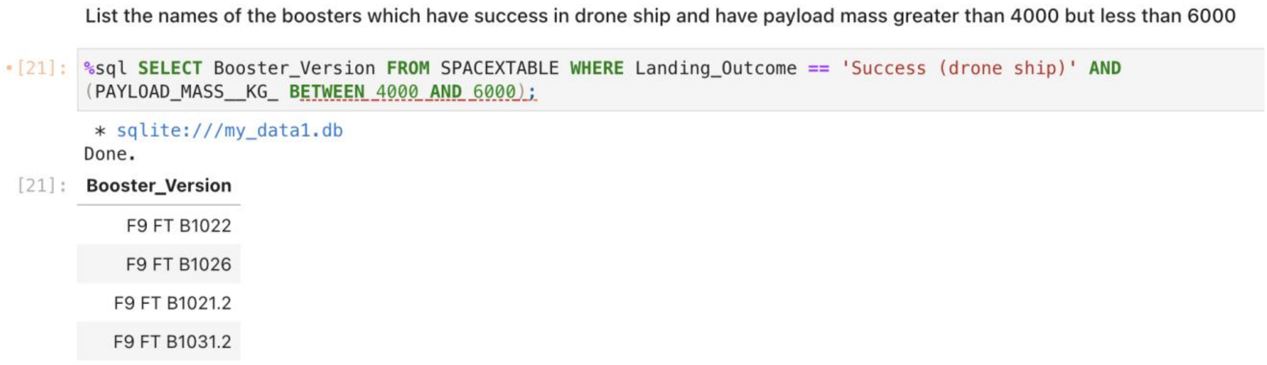


```
[20]: %sql SELECT DISTINCT Landing_Outcome FROM SPACEXTABLE;
```

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

[20]:

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt



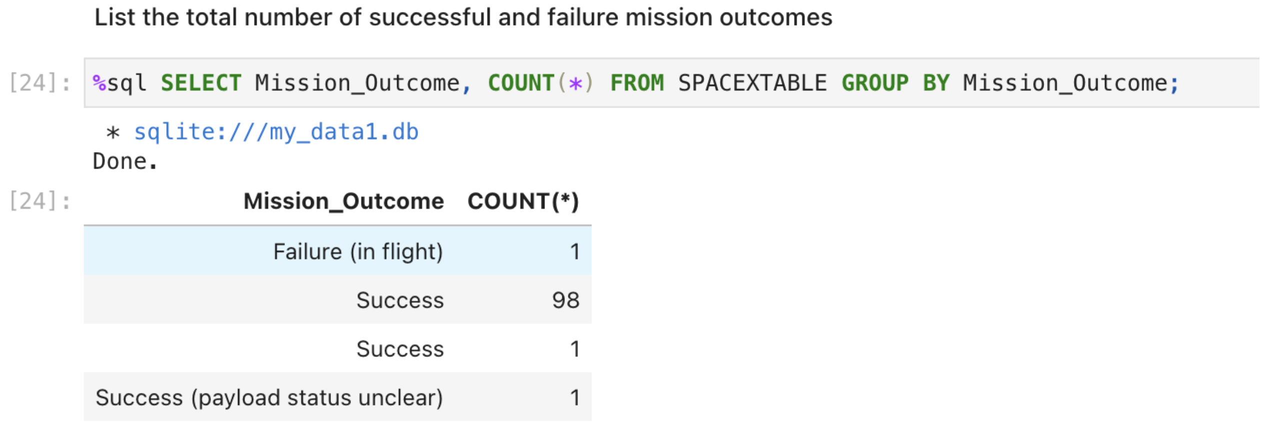
List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[21]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (drone ship)' AND
      (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

[21]:

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



List the total number of successful and failure mission outcomes

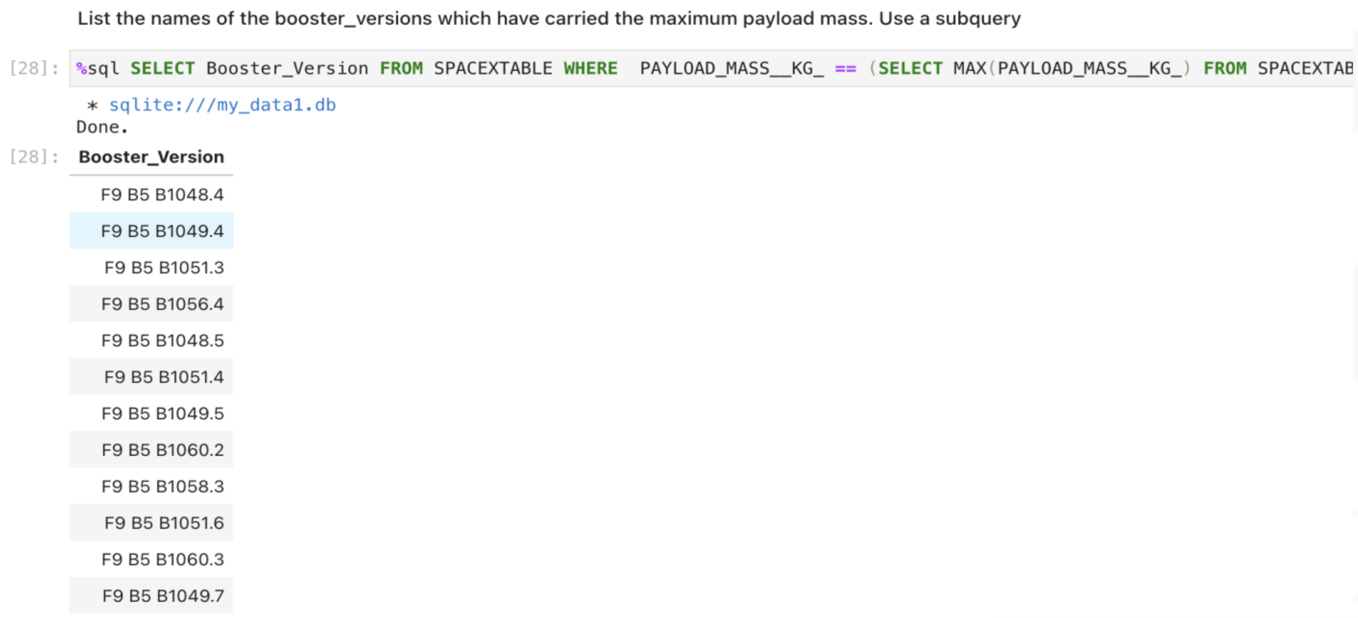
```
[24]: %sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

[24]:

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

EDA WITH SQL



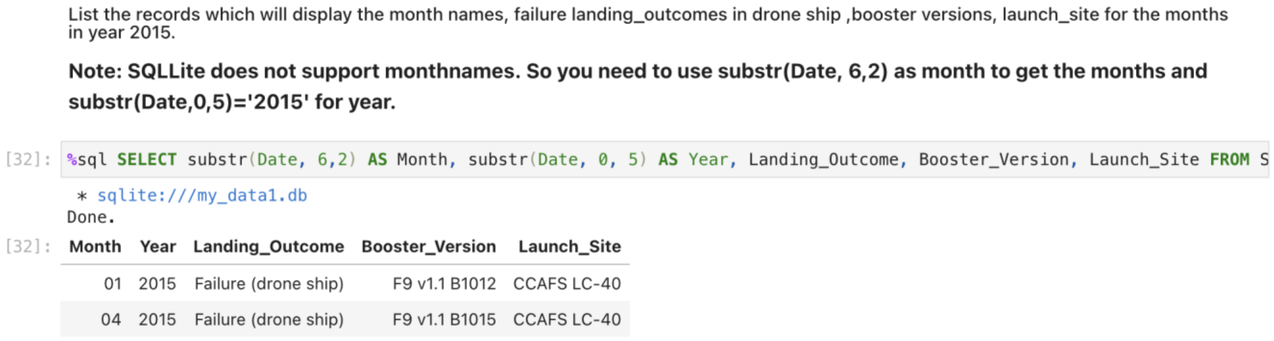
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[28]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE  PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTAB
```

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

[28]:

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



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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
[32]: %sql SELECT substr(Date, 6,2) AS Month, substr(Date, 0, 5) AS Year, Landing_Outcome, Booster_Version, Launch_Site FROM S
```

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

[32]:

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

EDA WITH PANDAS AND MATPLOTLIB

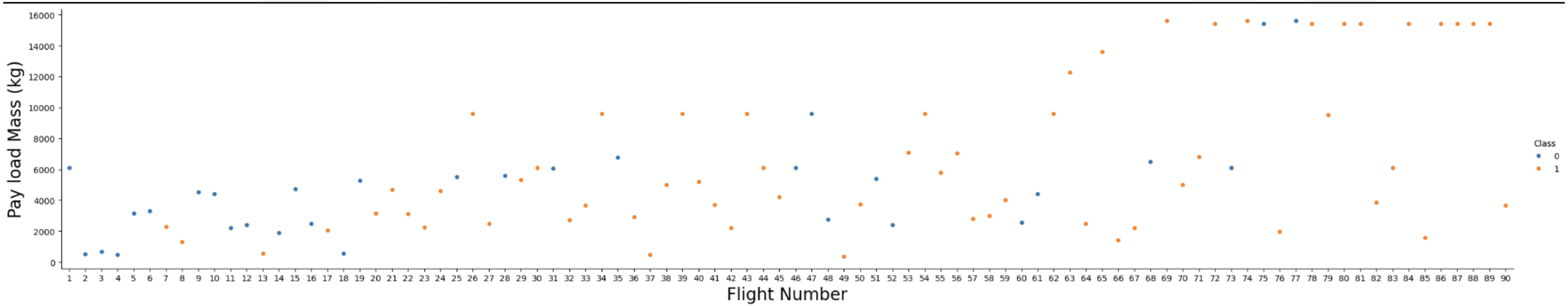
Objectives: To Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib

Plots we visualize:

- FlightNumber vs PayloadMass
- PayloadMass vs Launch Site
- Success Rate vs Orbit
- FlightNumber vs Orbit
- PayloadMass vs Orbit
- Launch Success Yearly Trend

EDA WITH PANDAS AND MATPLOTLIB

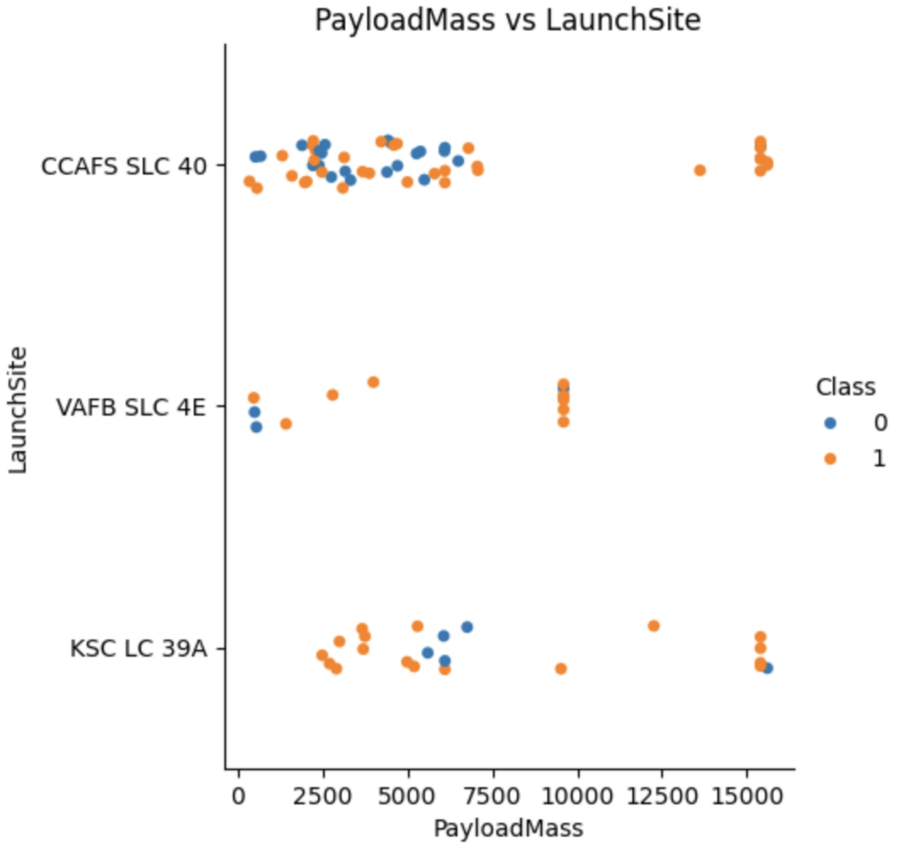
Flight Number vs Payload: Visualize the relationship between FlightNumber and PayloadMass and observe that as FlightNumber increases, PayloadMass tends to decrease.



EDA WITH PANDAS AND MATPLOTLIB

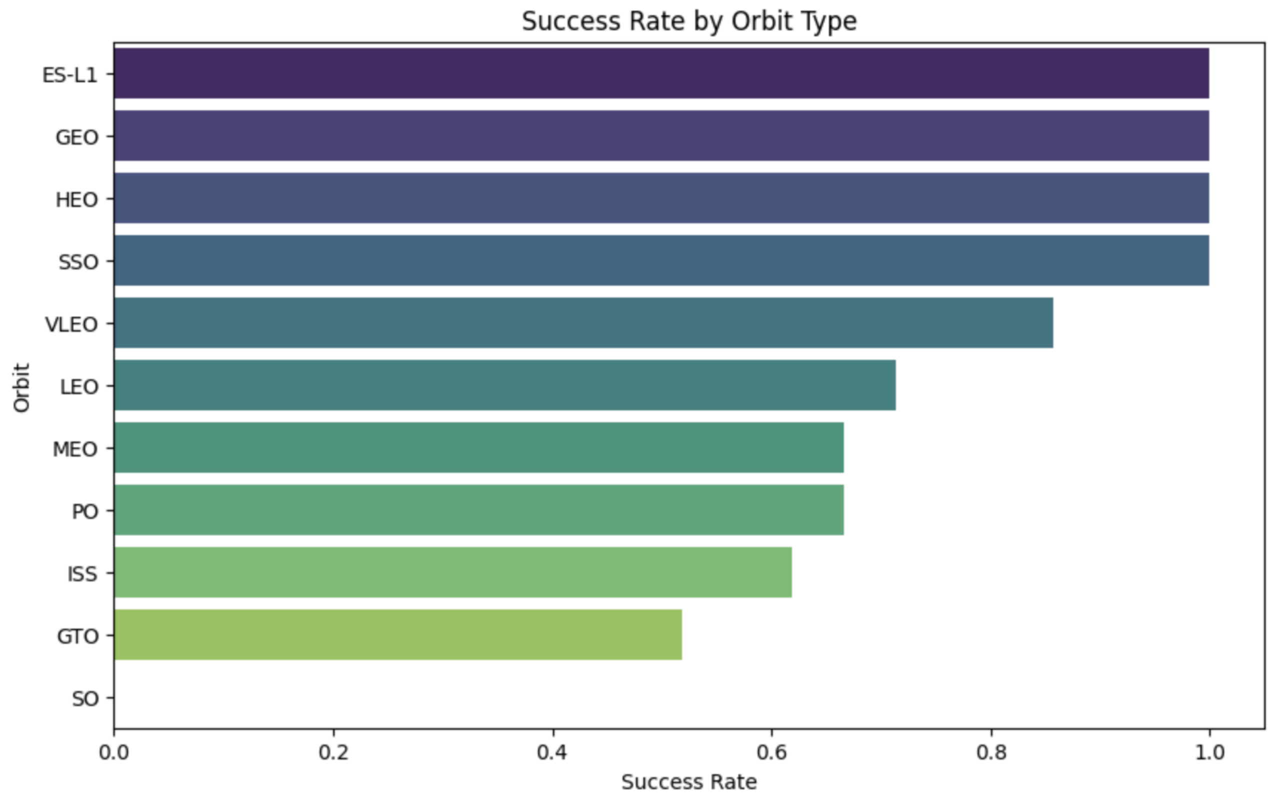
PayloadMass vs LaunchSite: Explored the relationship between PayloadMass and LaunchSite using a scatter plot and identified patterns in payload mass for different orbit types.

Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



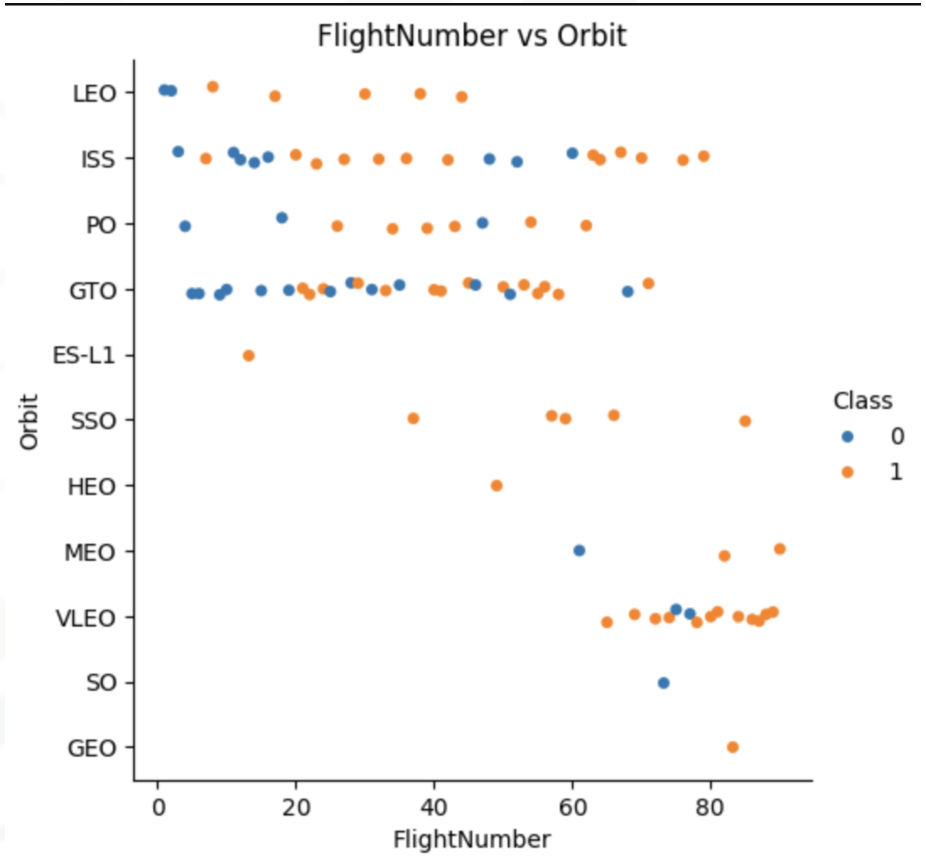
EDA WITH PANDAS AND MATPLOTLIB

SuccessRate vs Orbit: Created a bar chart to display the success rate for each orbit type. Identified orbits with high success rates.



EDA WITH PANDAS AND MATPLOTLIB

FlightNumber vs Orbit: Plot a scatter plot to analyze the relationship between FlightNumber and Orbit and observe variations in the relationship based on orbit type.

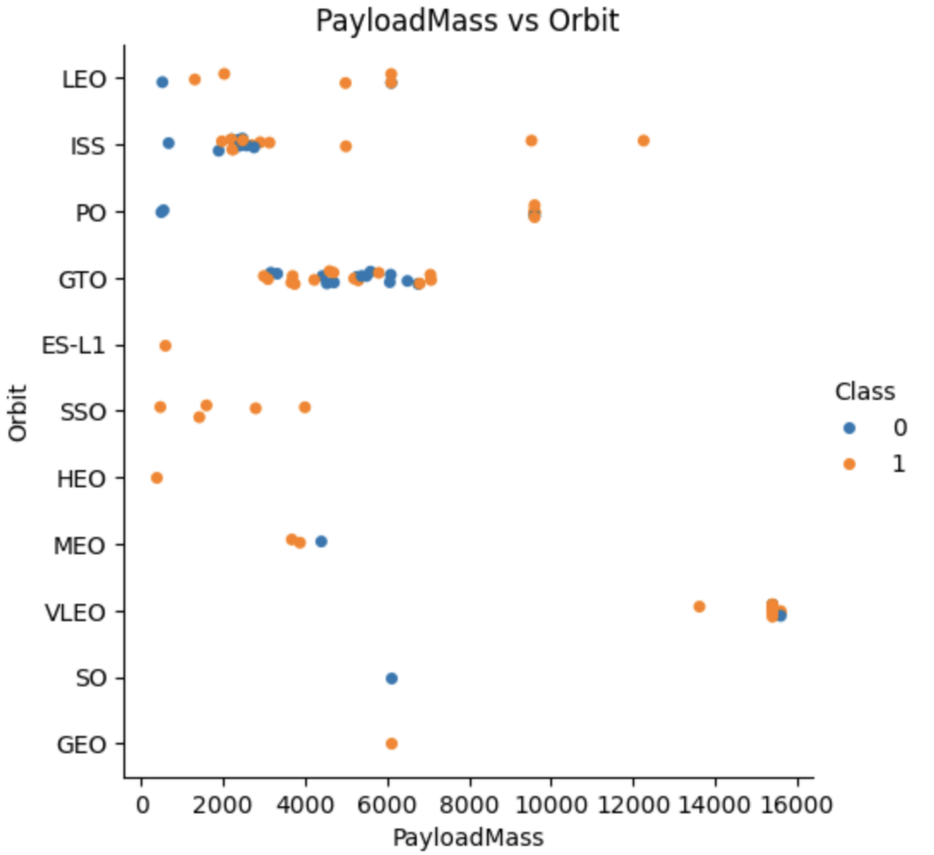


We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

EDA WITH PANDAS AND MATPLOTLIB

PayloadMass vs Orbit: Explore the relationship between PayloadMass and Orbit using a scatter plot and identify patterns in payload mass for different orbit types.

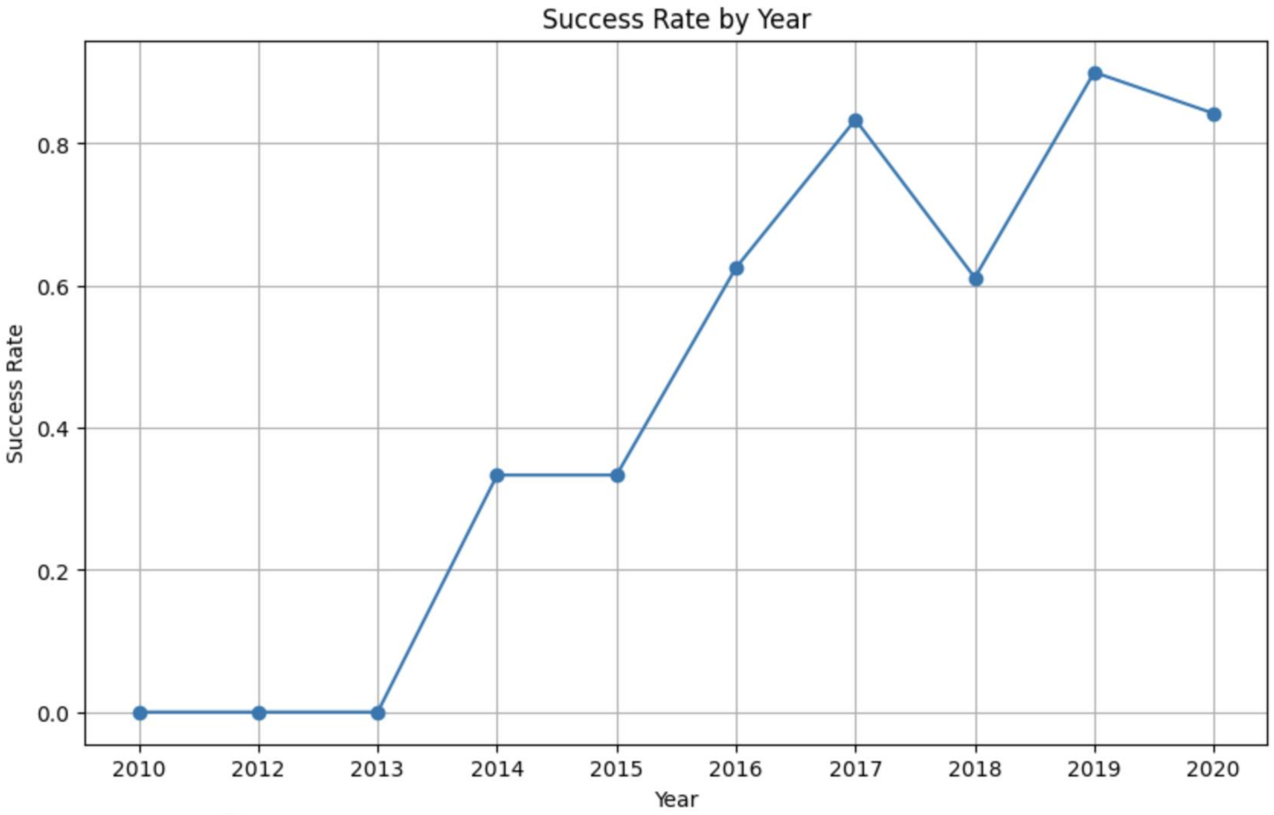
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.



EDA WITH PANDAS AND MATPLOTLIB

Plotted the average success rate trend over the years: Create a line chart to visualize the average success rate trend over the years and observe variations in success rates over time.

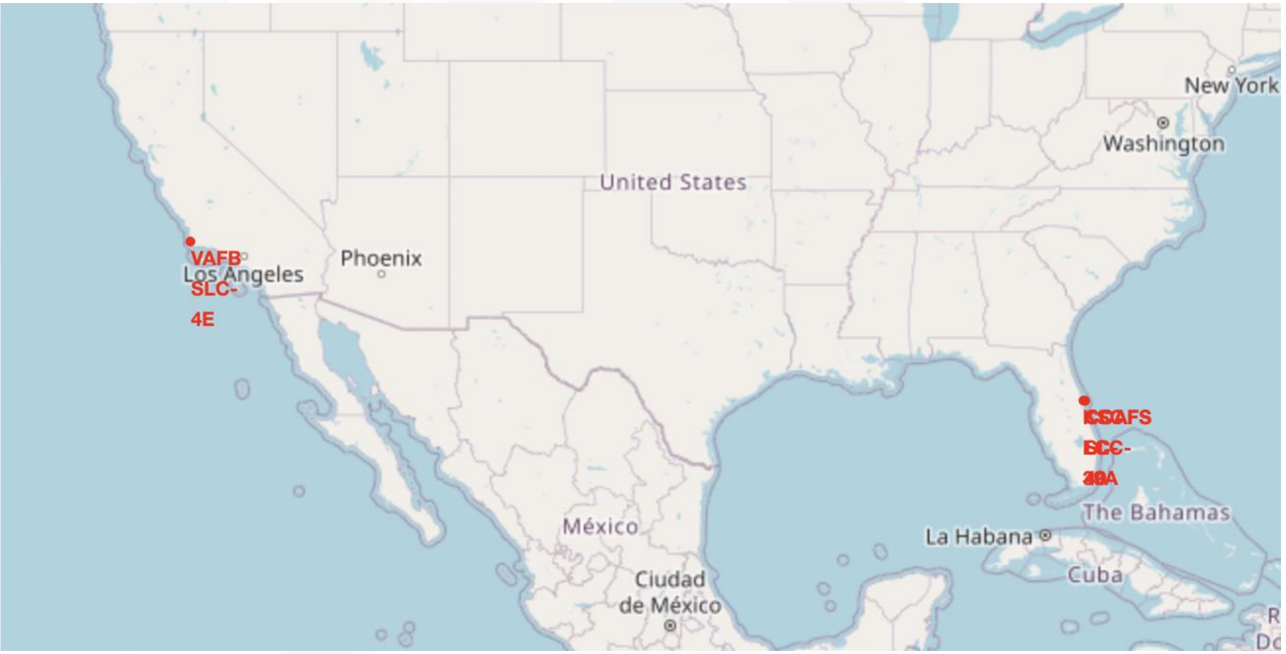
We can observe that the sucess rate since 2013 kept increasing till 2020



INTERACTIVE VISUALS WITH FOLIUM

Objectives: To visually analyze SpaceX launch sites using Folium. This involves marking sites on a map, indicating success/failure of launches, and calculating distances to nearby features like railways and highways.

- Observations: The launch sites are in proximity to the Equator line and very close to the coast.

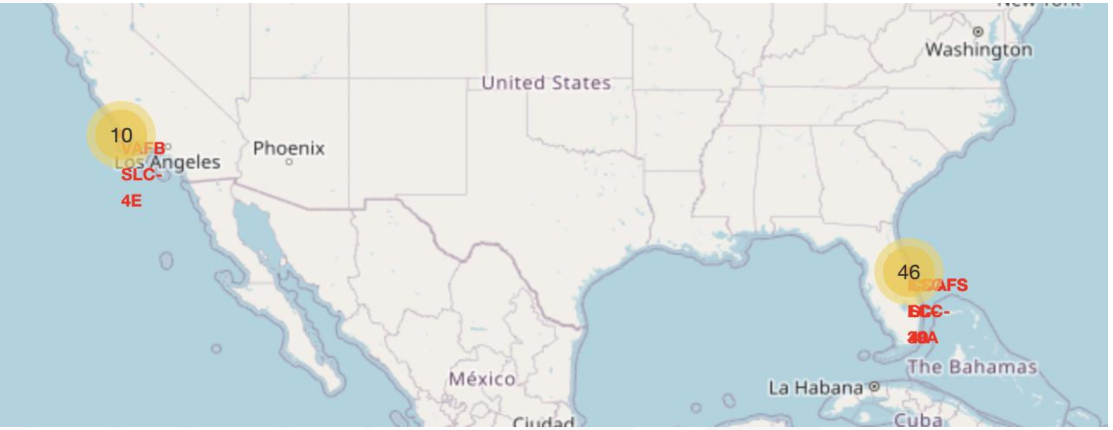


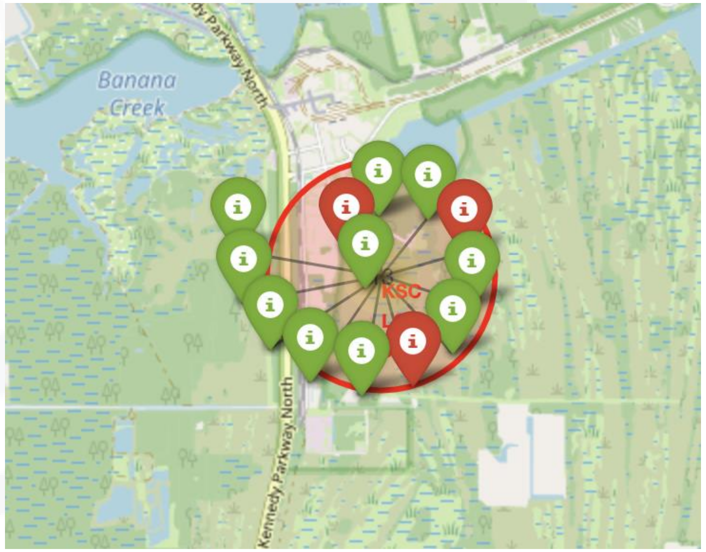
INTERACTIVE VISUALS WITH FOLIUM

Observations:

- Many launch records will have the exact same coordinate. Hence, we create Marker clusters to simplify a map containing many markers having the same coordinate.

- From the color-labeled markers in marker clusters, we are able to easily identify which launch sites have relatively high success rates: KSC LC-39A.





INTERACTIVE VISUALS WITH PLOTLY

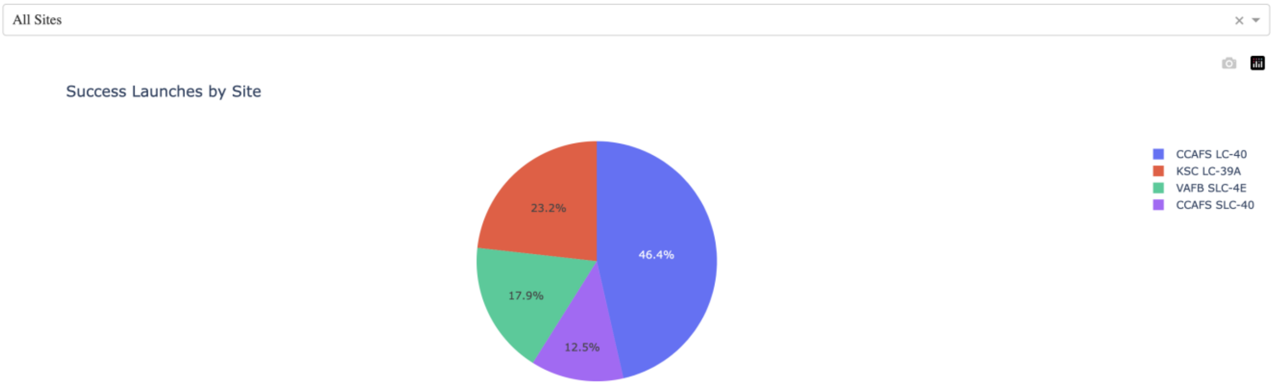
Objectives: to explore and analyze geographical patterns related to launch sites using interactive Plotly visualizations and answer the following questions:

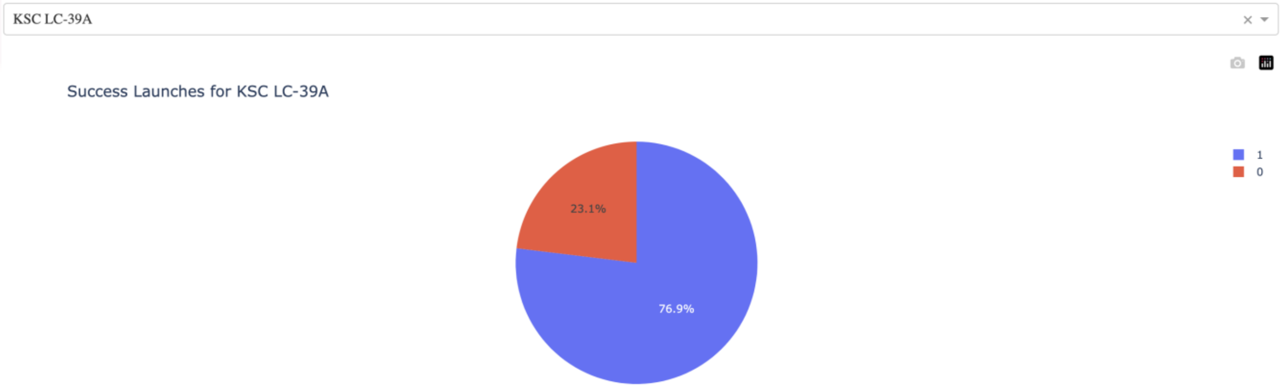
- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

INTERACTIVE VISUALS WITH PLOTLY

Observations:

- KSC LC-39A has the most successful launches

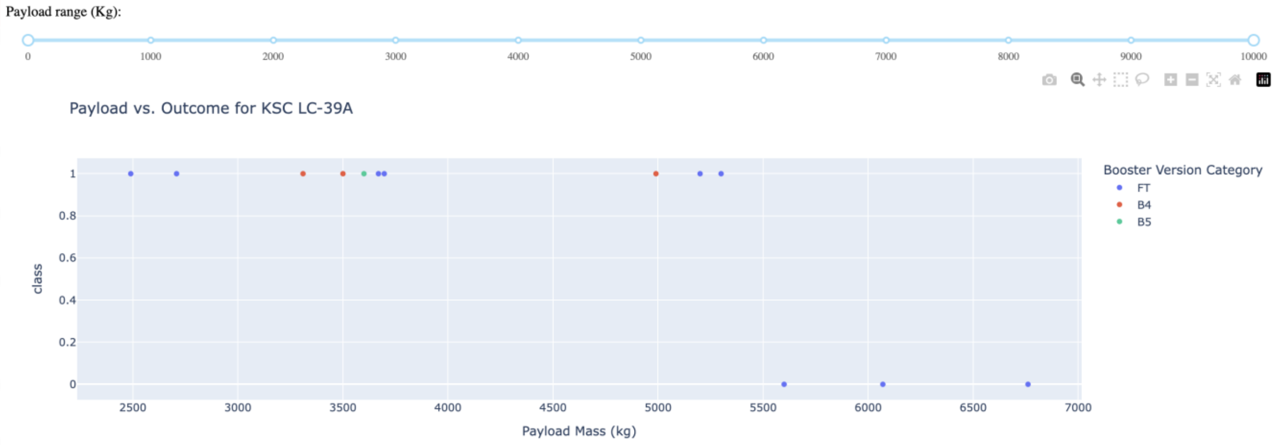


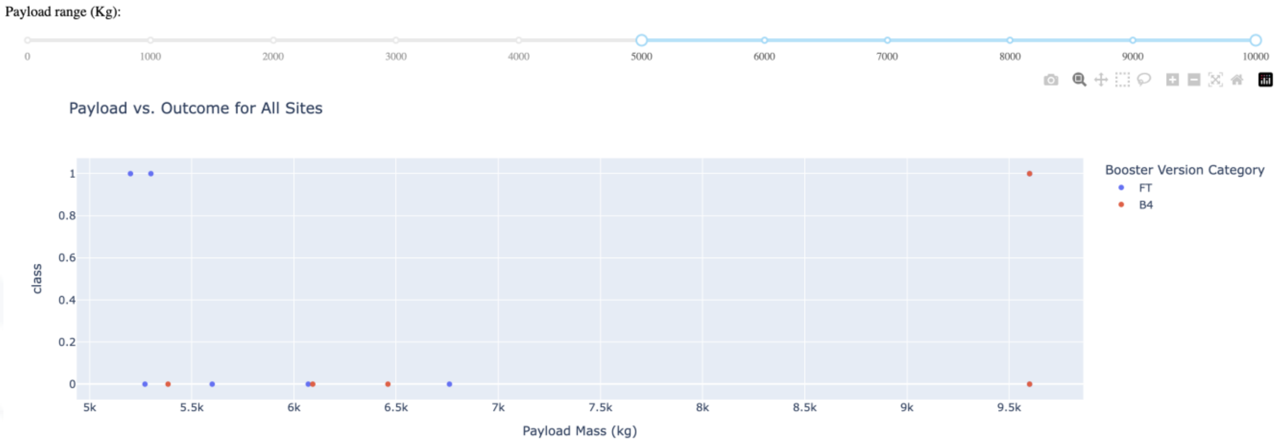


- KSC LC-39A has the highest success rate with 10 successful and 3 failed landing.

INTERACTIVE VISUALS WITH PLOTLY

Observations:





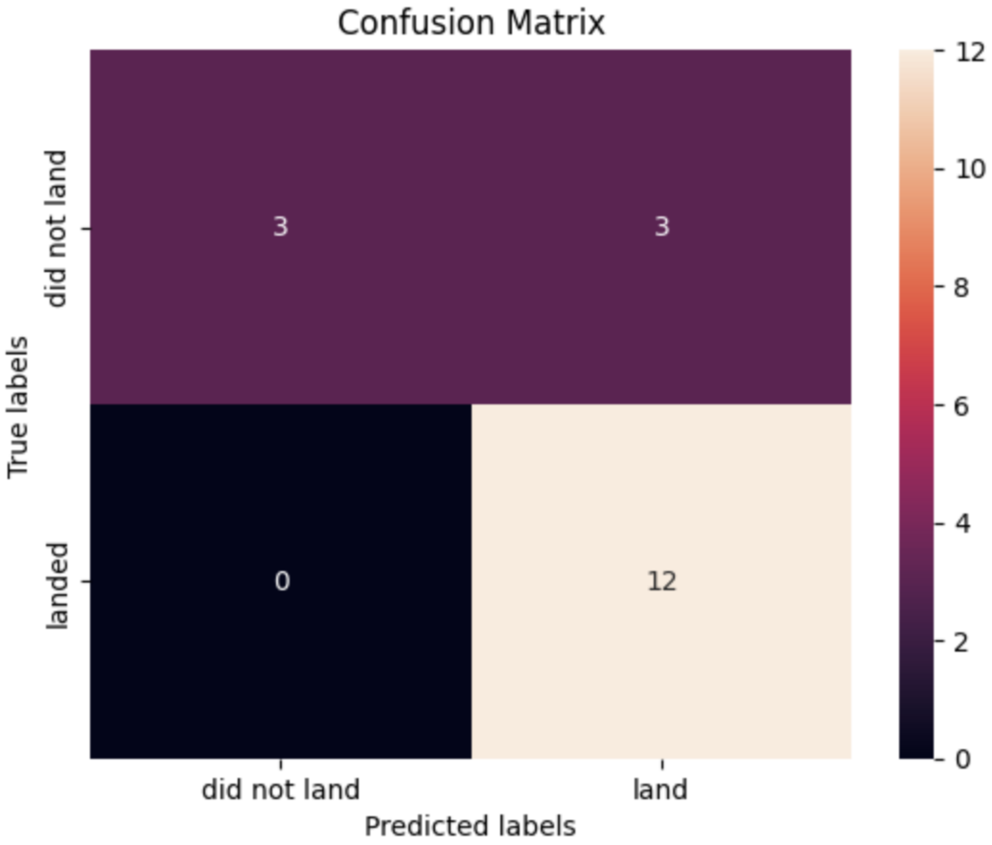
PREDICTIVE DATA ANALYSIS

Objectives:

- To Perform exploratory Data Analysis and determine Training Labels create a column for the class
- Standardize the data Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data

PREDICTIVE DATA ANALYSIS

- Classification Accuracy:
  - Test Set scores inconclusive due to small sample size (18 samples)
  - Evaluation conducted on entire dataset
  - Decision Tree Model identified as the best, demonstrating higher scores and highest accuracy (83%)
- Confusion Matrix Insights:
  - Decision Tree effectively distinguishes between classes
  - Occurrence of false positives with Decision Tree



CONCLUSION

- Objective Clarification: Our primary objective was to develop a machine learning model tailored for Space Y, enabling them to compete effectively with SpaceX. Specifically, our focus was on predicting successful Stage 1 landings, potentially saving up to $100 million USD per launch.
- Data Acquisition: We meticulously gathered data from various sources, including a public SpaceX API and by employing web scraping techniques on SpaceX's Wikipedia page. This extensive dataset served as the foundation for our model development.
- Data Labeling and Storage: With the collected data, we embarked on a labeling process to ensure the accuracy and relevance of each data point.
- Dashboard Creation: Concurrently, we developed an interactive dashboard tailored for visualizing critical insights derived from the dataset.
- Model Development and Performance: Leveraging state-of-the-art machine learning techniques, we engineered a predictive model with an impressive accuracy rate of 83%. This model demonstrated remarkable proficiency in forecasting successful Stage 1 landings, empowering SpaceY with valuable foresight.

ACKNOWLEDGMENT

- I extend my sincere gratitude to the instructors whose dedication and expertise have been invaluable throughout this learning journey. Their guidance and support have played a pivotal role in shaping my understanding of predictive data analysis.
- I am also thankful to Coursera for providing a comprehensive platform that fosters learning and growth.
- Additionally, I appreciate IBM for its contributions to the field of data science and for providing access to cutting-edge tools and resources. Their commitment to education and innovation has been instrumental in enhancing my skills and knowledge in this domain.