IBM Developer
SKILLS NETWORK

DATA SCIENCE CAPSTONE PROJECT

Winning Space Race with Data Science

Prepared by:

Barbara Zasowski

Date: 01/24/2025

Outline

Executive Summary

Summary of methodologies

- **Data Collection**: Obtained SpaceX data using REST APIs and web scraping
- **Data Wrangling**: Cleaned data for accurate analysis (success/fail outcome variables).
- **Exploratory Data Analysis (EDA) with visualization**: Visualization tools to test factors such as payload, launch site, flight numer etc.
- **Data Analysis with SQL:** Used SQL and Python for trend analysis (calculate statistics as total payload, range of successfull lauches and successful/failed outcomes.
- **Interactive Visualization: Map with Folium and Dashboard with Plotly Dash**
- **Predictive Analytics (Classificatio)**: Built and evaluated machine learning models predicting landing outcomes using logistic regression, support vector machines (SVM), decision tree, and K-nearest neighbors (KNN)

Summary of all results

Exploratory Data Analysis

- Launch success rates improved over time.
- Specific payload ranges showed higher success rates.

Predictive Analysis

- Predictive models achieved over 85% accuracy.
- All models performed similarly on the test

Visualization Analysis

- Most launch sites are near the equator and all are close to the coast

Introduction

Project background and context

- SpaceX: Revolutionizing Space Travel
- Key Achievements
- Leading commercial space company
- Dramatically reduced launch costs
- Pioneered reusable rocket technology
- Cost Breakthrough
- Falcon 9 launch cost: $62 million
- Competitor launch costs: $165+ million
- Savings achieved through first-stage rocket reusa

Innovation Strategy

•Machine learning models predict first-stage landing potential

•Reusability = Lower launch expenses

•Making space exploration more accessible and affordable

Impact

•Transformed space travel from government-only to commercial reality

•Consistently successful International Space Station missions

•Setting new standards in aerospace engineering

Questions to ask

- What factors influence mission success?
- How do payload and orbit affect outcomes?
- Can predictive models improve future mission planning?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Performed using SpaceX REST APIs and web scraping techniques
- Perform data wrangling
  - Filtering the data
  - Augmenting with missing values
  - Applying One Hot Encoding to prepare the data for analysis ad modeling
- 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 and evaluation of classification models to ensure the best results

Data Collection – REST APIs

- Request Data- a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX Wikipedia entry. Both data collection methods were necessary to get complete information about the launches of rocket for a more detailed analysis.
- Decode Response – using .json() and converted to pandas dataframes using json_normalize() function
- Request Information - requested additional information using custom functions
- Create Dictionary – from the data collected
- Create DataFrames – from the dictionary
- Filter DataFrames – to ensure that dataframes hold Falcon 9 launches data only
- Replace Missing Values – Payload Mass attributes using .mean() function
- Export the result sets into CSV files – to use in later analysis

Data Collection - WEB Scraping

Data Source – Wikipedia website to collect Falcon 9 launch data	Decode Response – used BeautifulSoup object to parse the HTML respnse data	Column Name Extraction – used HTMLS table header to extract the column names	Data Extraction – collecting the data by parsing HTML tables	Create Data Dictionary- from collected data	Create DataFrames – from the dictionary	Export the result sets into CSV files – to use in later analysis

Data Wrangling

- Perform exploratory Data Analysis and determine Training Labels
- Identified bad outcome for landings
- Augmented new classification feature to the dataset for landing outcome
- Identified Successful and failed landing attempts for different types
- Identified the percentage of mising information or NaN
- Augmented missing data attributes using .mean() functions
- Export the data to CSV

EDA with Data Visualization

Charts to visualize the relationship

- Flight Number and Launch Site
- Payload and Launch Site
- Success rate of each orbit type
- Flight Number and Orbit type
- Pay Load and Orbit type
- Success yearly trend

Analysis:

- Viewed relationship by scatter plots. The variables could be used for machine learning if a relationship exists
- Bar Charts show comparisons among discrete categories. They show the relationship among the categories and a measured value

EDA with SQL

Performed SQL queries:

Display:

- Names of unique launch sites
- 5 records where launch site begins with „CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

List:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total of successful and failed missions
- Names of booster versions which have carried the max payload mass
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Interactive Map with Folium

Markers identifying Launch Sites:

- Circle added at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longtitude coordinates
- Red circles added at all launch sites coordinates with popup label showing its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes:

- Colored markers added of successful (Green) and unsuccessful (Red) launches using Marker Cluster to identify which launch site have high success rates.

Distances between a Launch Site to Proximties:

- Colored lines added to show distance between launch sites and its proximity to the nearest coastline, railway, hightway, and city

Dashboard with Plotly Dash

Developed a dashboard application containing input components such as dropdown list and a range slider to interact a pie chart and a scatter point chart. These features were developed to answer the insights such as:

- Site with the largest successful Launch
- Site with the highest success rate
- Payload ranges with highest launch success rate
- Payload ranges with lowest launch success rate
- F9 Booser version. Payload vs. Success rate

Predictive Analysis (Classification)

There are several steps when creating a high-performing classification model:

- Data Preparation – Clean, preprocess, and split the dataset into training and testing datasets
- Build Model – Train multiple classification models such as Logistic Regression, SVM, Decision Tree, KNN)
- Evaluate – Use accuracy, score, and confusion matrix to evaluate each model
- Improve – Optimize models via hyperparameter tuning and cross-validation
- Best Result selection– Idenitify the model with the highest performance

Results

Exploratory data analysis

Executive Summary Visualization/Analytics

Visualizations Analytics

Predictive analysis

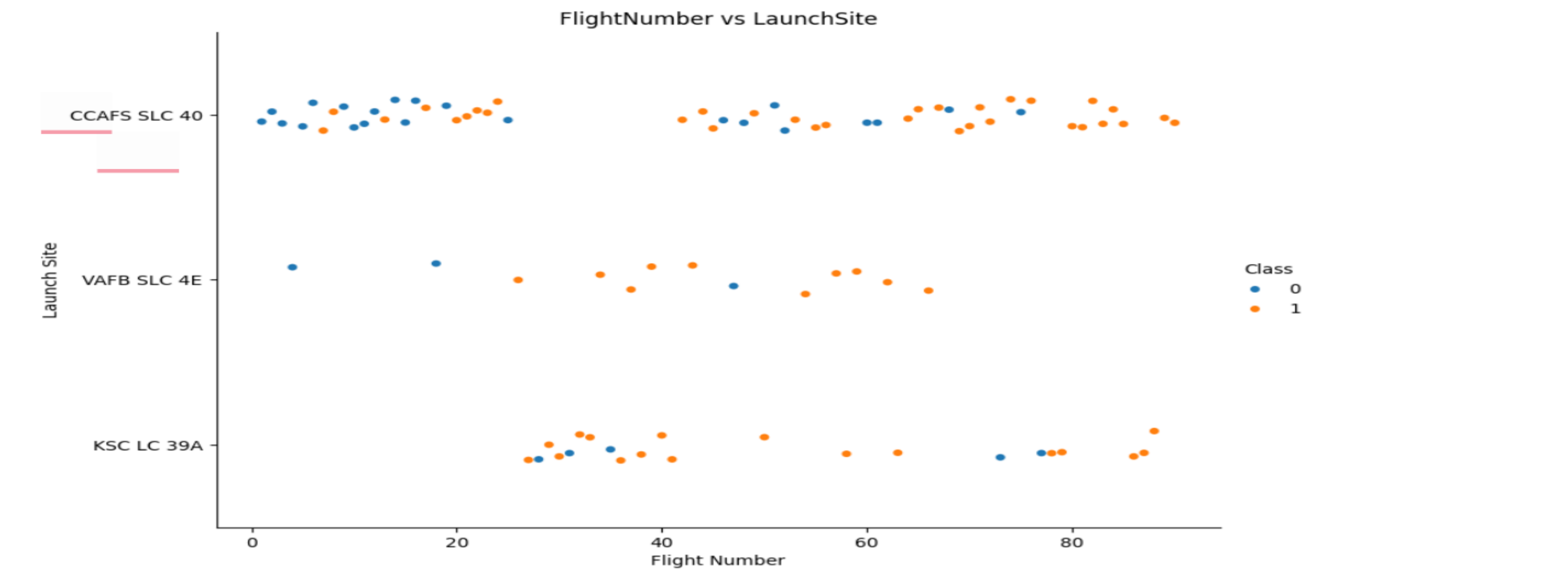
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Data Analysis Explanation:

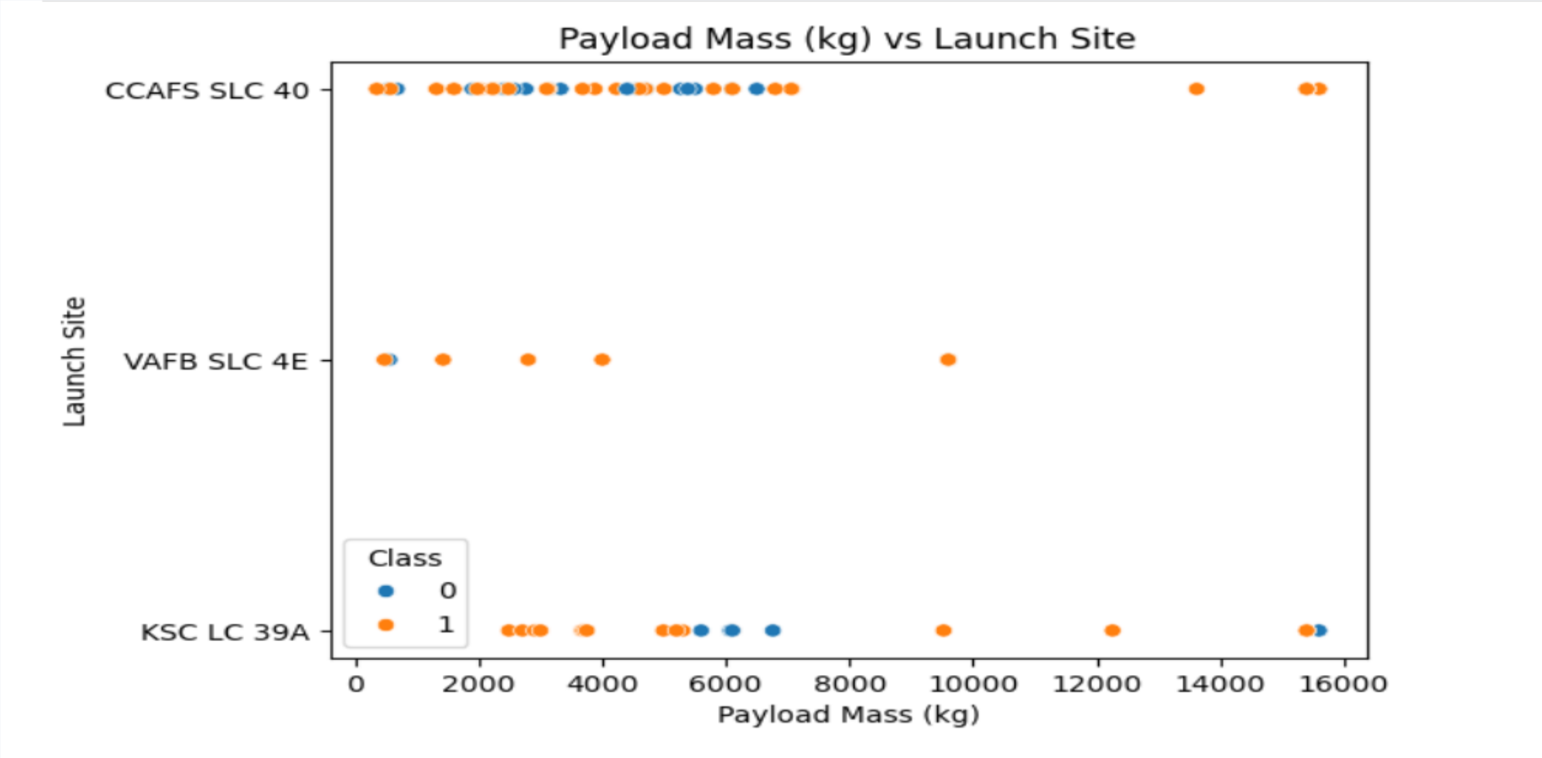
- Initial flights had lower rate of success
- Following flights had higher rate of success
- VAFB SLC 4E and KSC LC 39A have higher rates of success
- Most of the launches were from CCAFS SLC 40
- It is assumed that each new launch has a higher rate of success



Payload vs. Launch Site

Data Analysis Explaination:

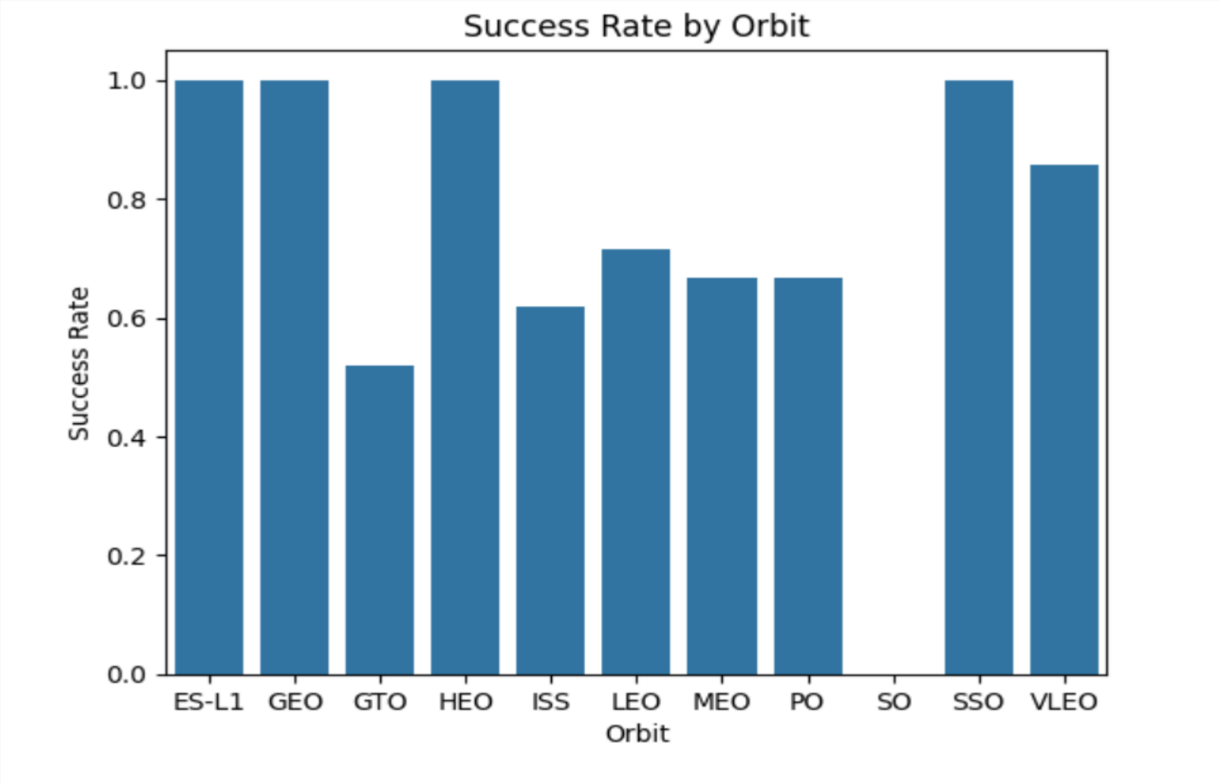
- The Higher the Payload Mass (kg), the higher rate of success
- Most launches with a payload mass of 7000 kg or higher had a successfull outcome
- KSC LC 39A has a 100% rate of success for launches with payload mass under 5500 kg. In addition, a success rate for launches with payload mass between 9,000 – 15,000 kg was higher
- Most of the launches took place from CCAFS SLC 40



Success Rate vs. Orbit Type

Data Analysis Explaination:

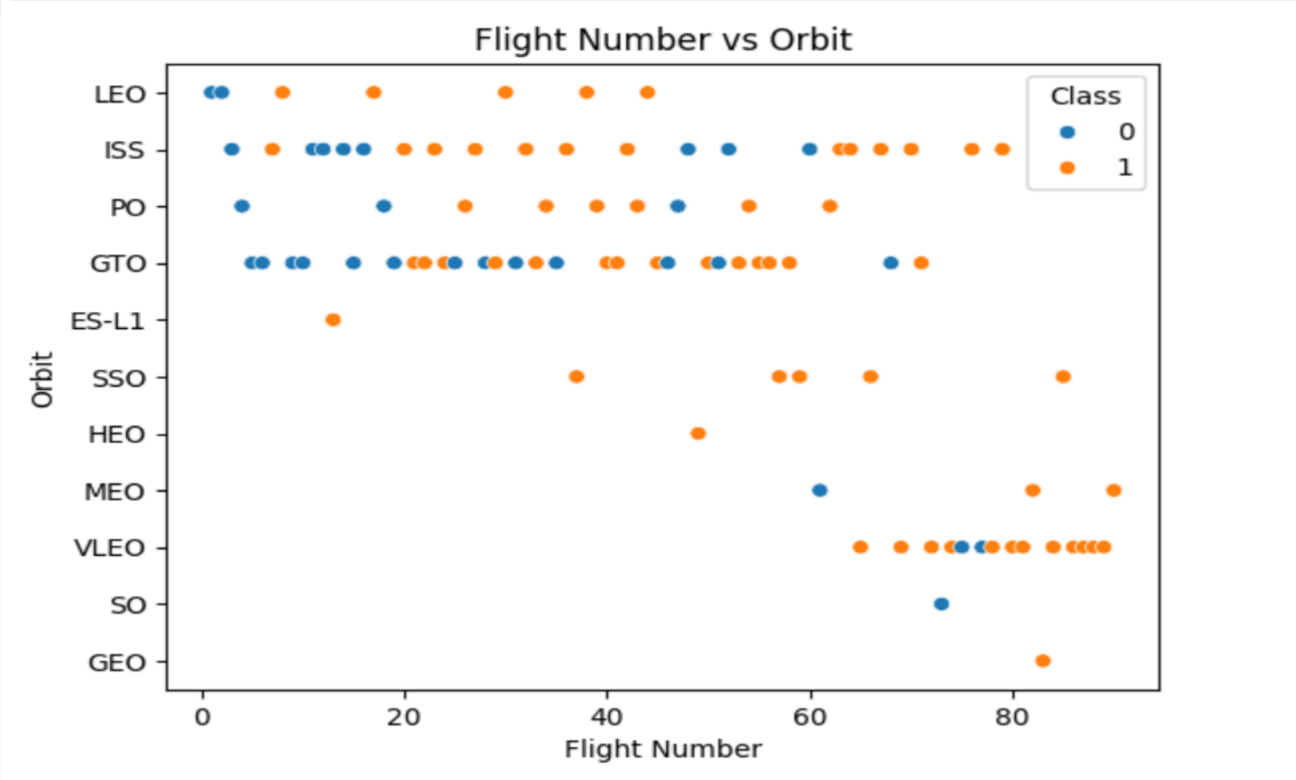
- Orbits with 100% Success Rate : ES-L1, GEO, HEO and SSO
- Orbits with 0% Success Rate : SO
- Orbits with success rate between 50% and 85% : GTO, ISS, LEO, MEO, PO



Flight Number vs. Orbit Type

Data Analysis Explaination:

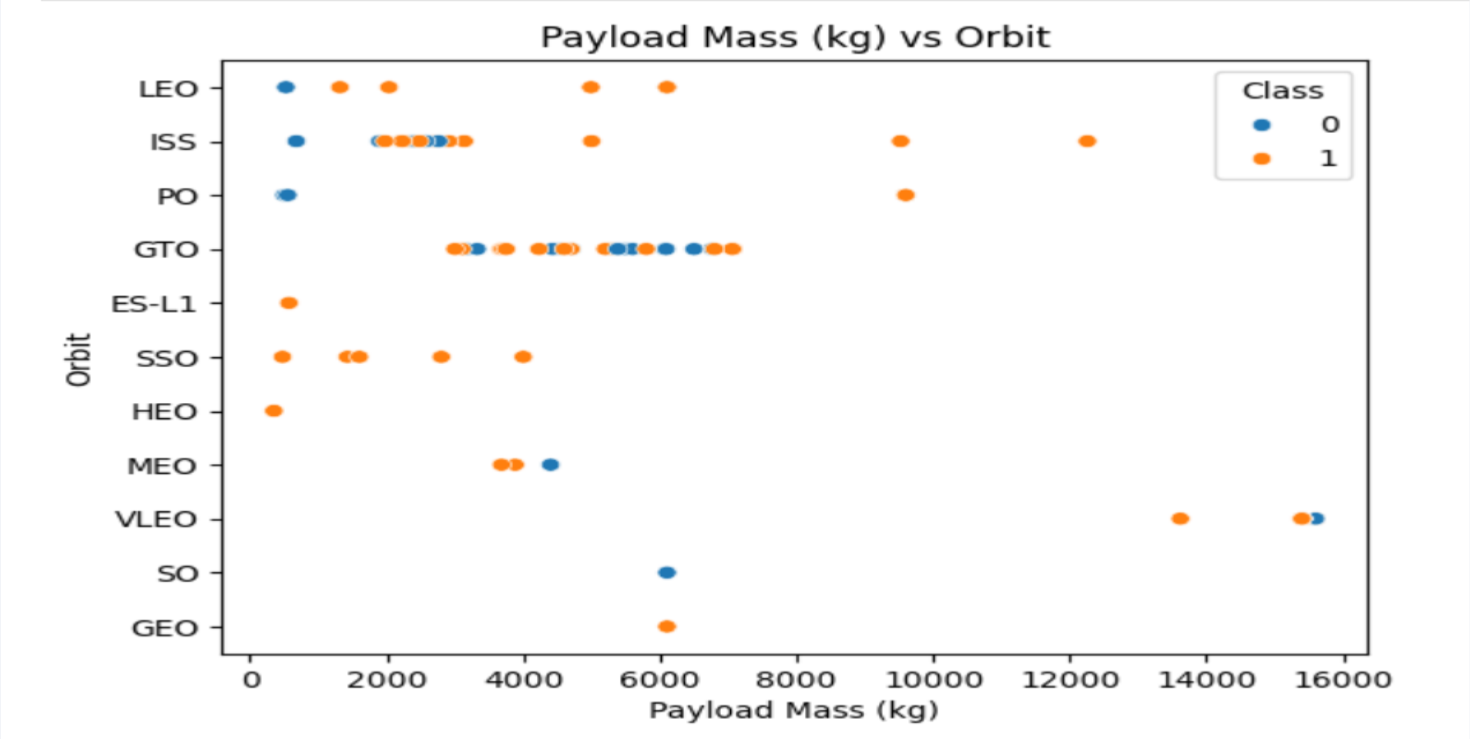
- A Higher Success Rate is achieved with each attempt in the orbit
- A Higher Success Relation appears related for the LEO Orbit
- GTO Orbit does not present any relationship



Payload vs. Orbit Type

Data Analysis Explaination:

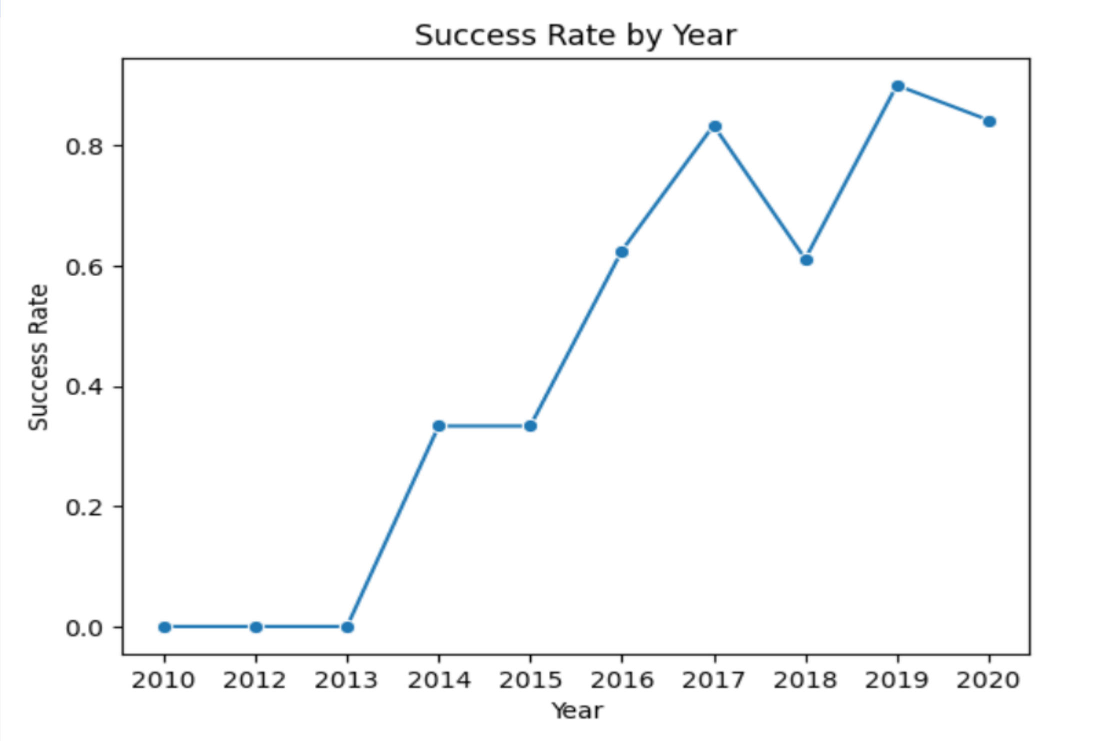
- Heavy payload mass has a higher rate of success for LEO, ISS and PO orbits
- SSO seems to have success rate for up to 4,000 kg payload
- GTO seems to have mixed outcome for heavier payloads



Launch Success Yearly Trend

Data Analysis Explaination:

- Success rate has increased significantly between 2013 and 2017, and between 2018 through 2019
- Success rate has decreased between 2017 and 2018, and between 2019 and 2020
- In general, the rate of success has been increasing since 2013



All Launch Site Names

Data Analysis Explaination:

- Displaying the names of the unique launch sites

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

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

Data Analysis Explaination:

- Displaying 5 Launch Sites that Begin with the string „CCA”

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

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Data Analysis Explaination:

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

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like '%NASA (CRS)%'
```

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

sum(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

Data Analysis Explaination:

- Displaying average payload mass carried by booster version F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Data Analysis Explaination:

- Displaying the date of the first successful landing in ground pad was achieved.

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

Hint:Use min function

```
%sql select Date from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)' order by Date asc limit 1
```

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

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Data Analysis Explaination:

- Displaying the names of the bosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

```
%sql select distinct(Booster_Version) from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and
```

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

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

Total Number of Successful and Failure Mission Outcomes

Data Analysis Explaination:

- Displaying the total numer of successful and failure mission outcomes
  - 1 Failure in Flight
  - 99 Success
  - 1 Success (Payload status unclear)

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome order by Mis
```

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

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

Boosters Carried Maximum Payload

Data Analysis Explaination:

- Displaying the names of the booster versions which have carried the maximum payload mass

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

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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

2015 Launch Records

Data Analysis Explaination:

- Displaying the failed landings in drone ship, their booster versions and launch site names for the months in year 2015

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.

```
%sql SELECT substr(Date, 6,2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHE
```

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

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

Data Analysis Eplaination:

- Ranking the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

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.

```
%sql SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-
```

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

Landing_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

Launch Site Analysis on Global Map

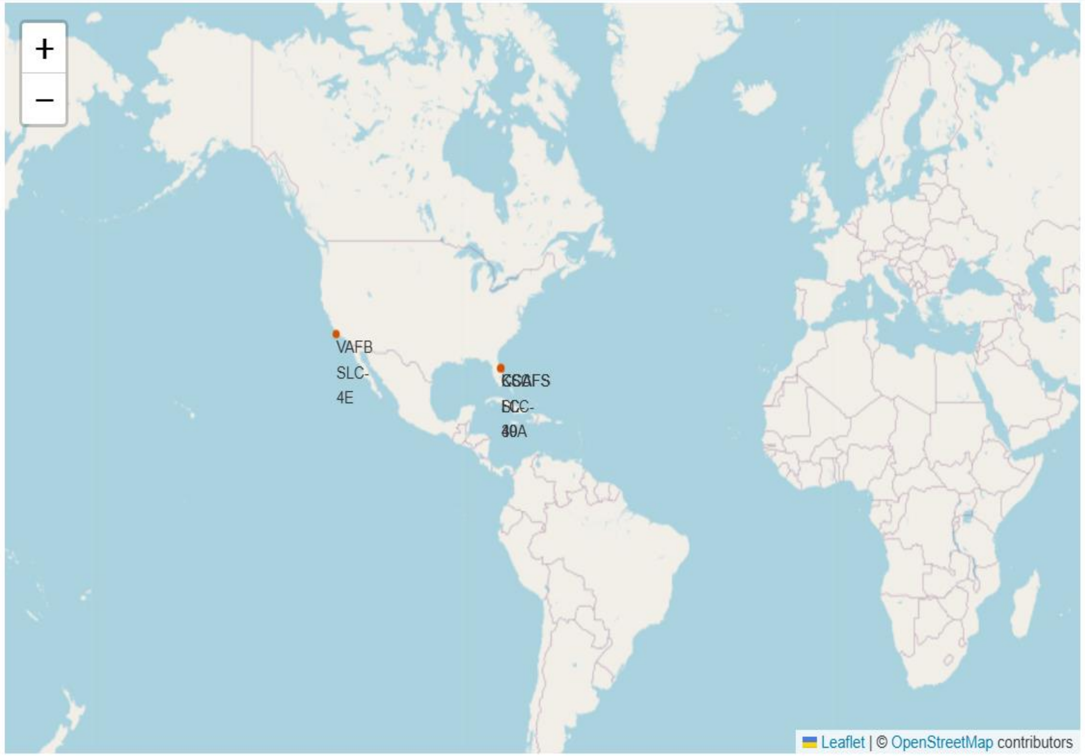
Proximity to the Equator:

- Launch sites near the equator provide a natural advantage due to Earth's rotational speed. Rockets launched from these locations get an extra velocity boost, reducing fuel needs and overall costs.
- The Earth rotates at about **1670 km/h** at the equator, meaning spacecraft launched from there already carry significant momentum, aiding in achieving and maintaining orbit efficiently.

Coastal Proximity:

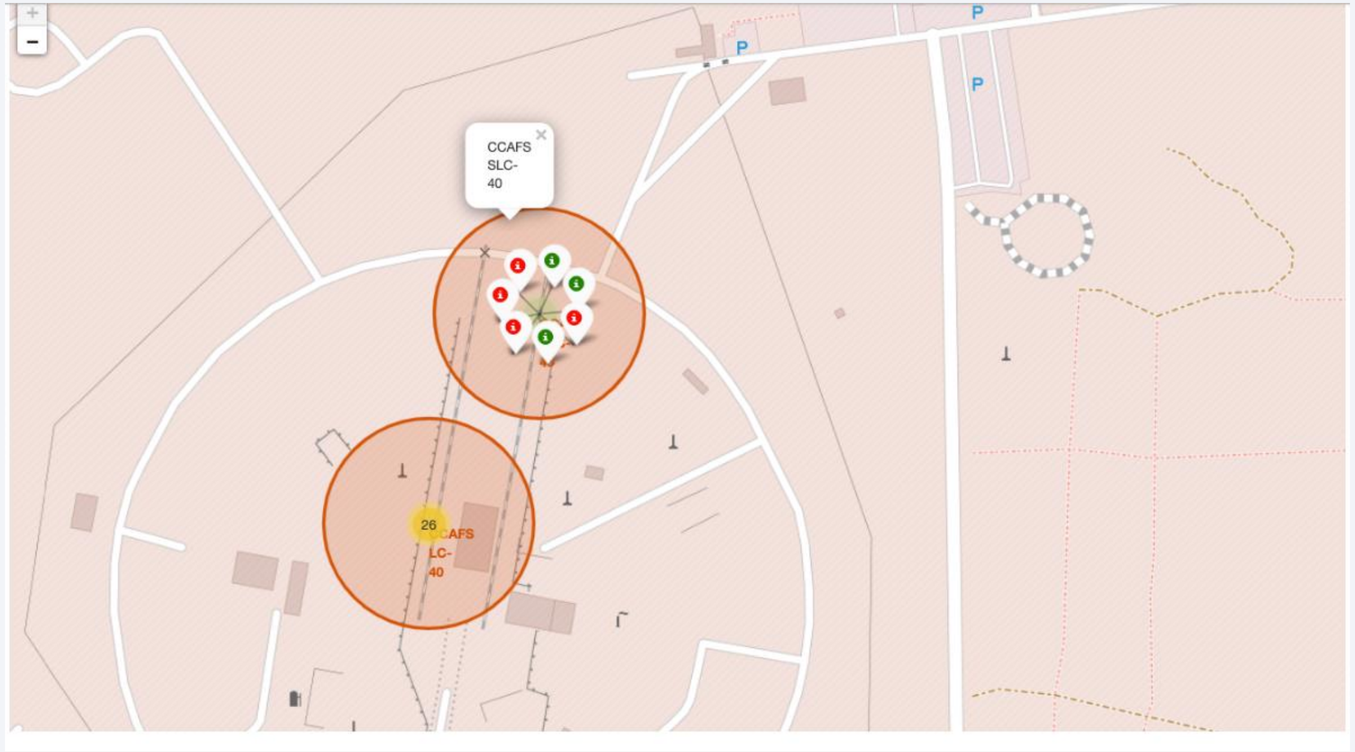
- Most launch sites are located near coastlines. This strategic placement minimizes risks, ensuring that if any debris falls or an explosion occurs, it happens over the ocean rather than populated areas.

Essentially, launching from near the equator and coastal regions makes space travel more efficient, cost-effective, and safer



Launch Site with Outcomes Colored Labels

- GREEN Marker - Successful Launches
- RED Marker- Failed Launches
- Launch Site CCASFS SLC-40 has a success rate of 43%



From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Launch Site Proximity to Key Sites

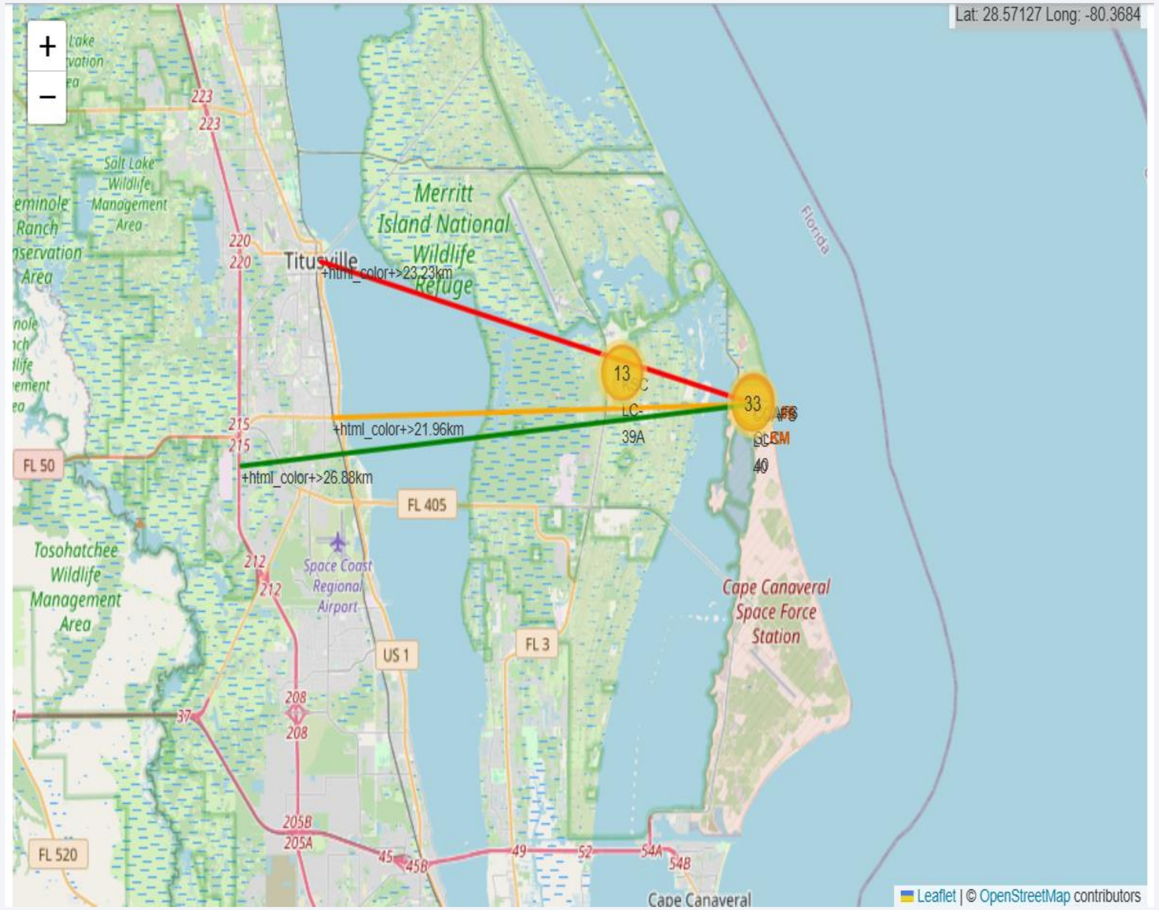
CCAFS SLC-40:

- **0.86 km** from the nearest coastline
- **21.96 km** from the nearest railway
- **23.23 km** from the nearest city
- **26.88 km** from the nearest highway

Why Location Matters

- **Coastal Advantage:** Placing launch sites near the ocean ensures that any discarded rocket stages or failed launches land in the water, minimizing risks to people and property.
- **Safety & Security:** A designated exclusion zone around the launch site keeps unauthorized individuals at a safe distance, reducing potential hazards and ensuring secure operations.
- **Balancing Accessibility & Distance:** While launch sites need to be far enough from cities and infrastructure to prevent damage from potential failures, they must also remain accessible via roads, railways, and ports to support logistics and transport needs efficiently.

This strategic placement ensures both safety and operational efficiency, making launches smoother and more secure.



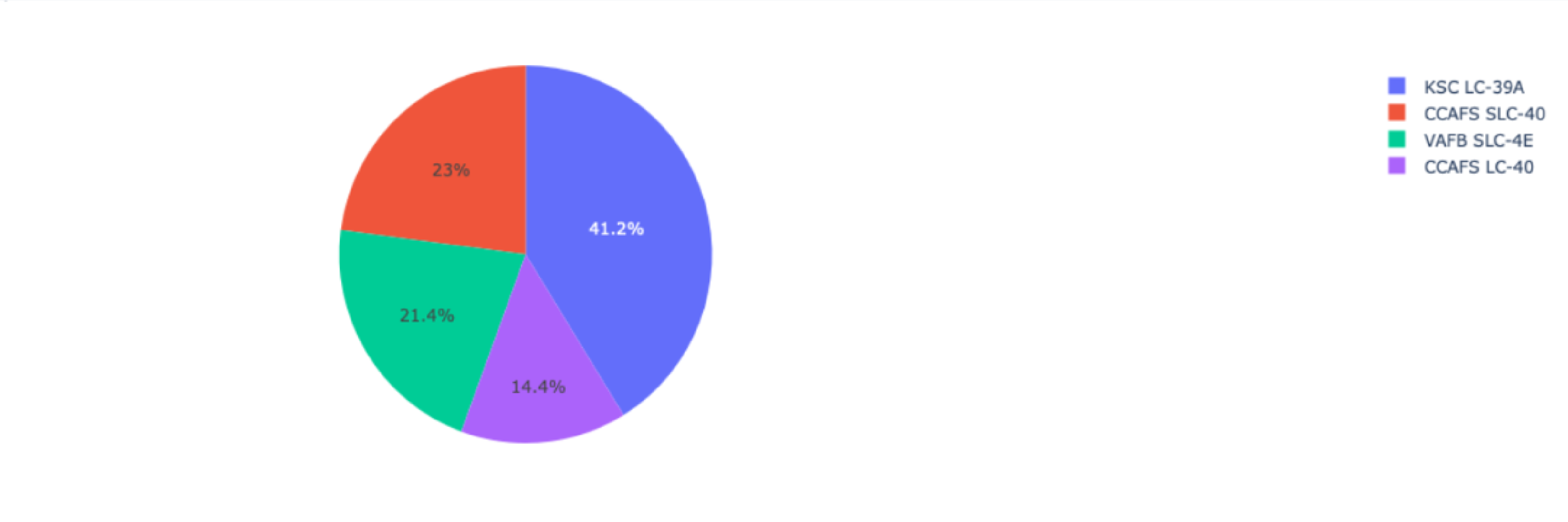
Section 4

Build a Dashboard with Plotly Dash

Dashboard with Plotly success rate count for all launch sites

Data Analysis Eplaination:

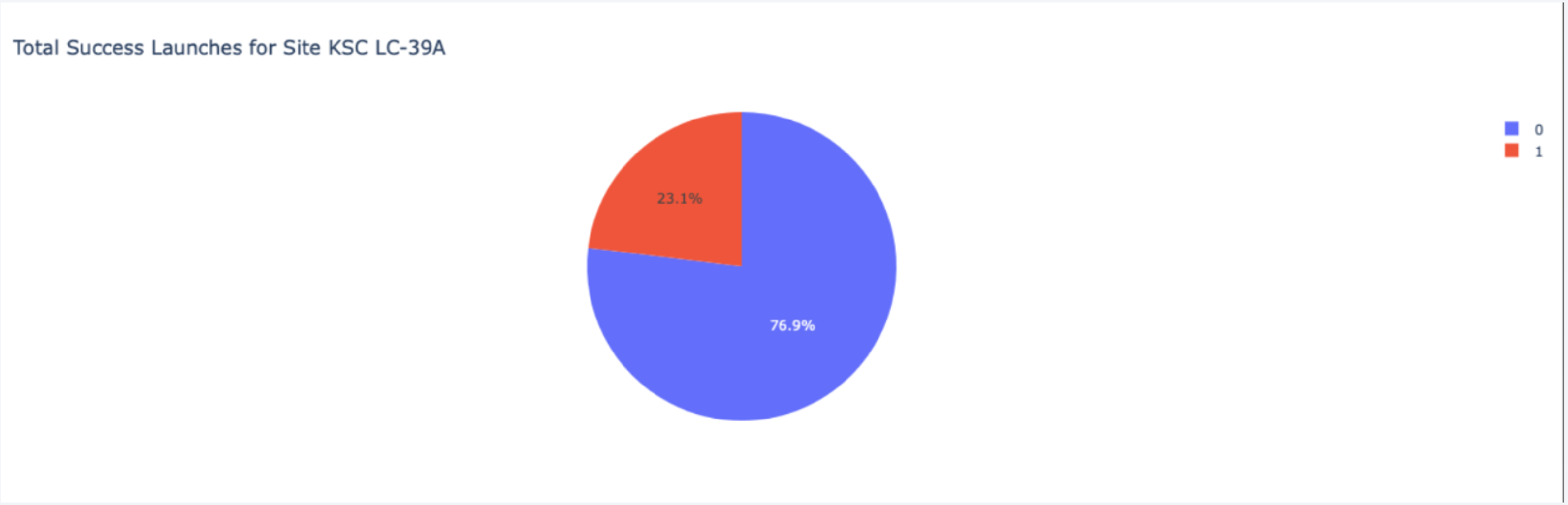
- Dashboard with Plotly built to analyze the succes rate count for all launch sites



Launch Site with Highest success ratio

Data Analysis Explaination:

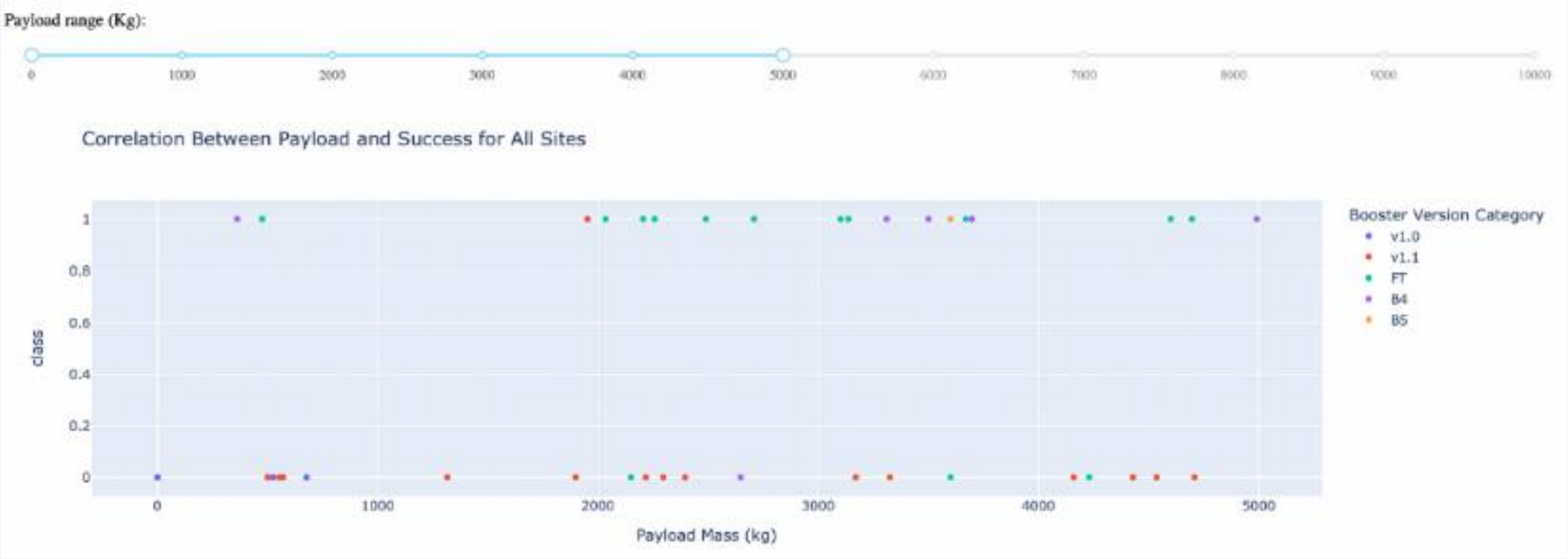
- KSC LC 39A has the highest success rate among the launch sites.



Payload Mass vs Launch Outcome

Data Analysis Explaination:

- Payloads between 2,000 kg and 5,500 kg have the highest success rate



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Data Analysis Explaination:

- All the models performed similarly with comparable scores and accuracy, thus there isn't one method that performs best
- The most probable cause is due to the small dataset

Find the method performs best:

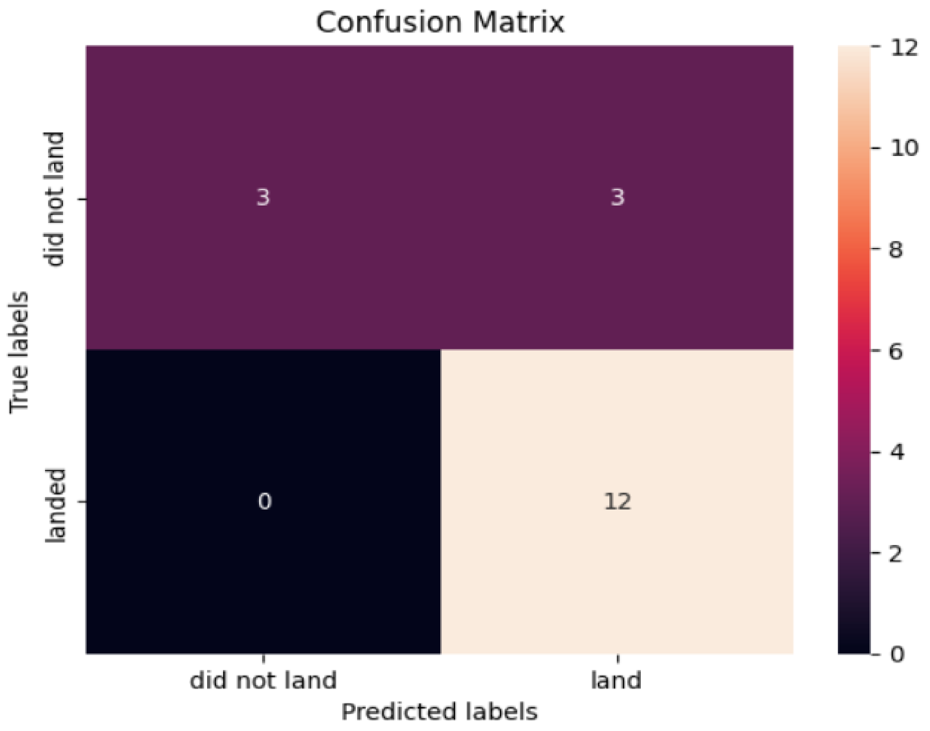
```
print("Test accuracy for Logistic Regression:", test_accuracy_logreg)
print("Test accuracy for SVM:", test_accuracy_svm)
print("Test accuracy for Decision Tree:", test_accuracy_tree)
print("Test accuracy for KNN:", test_accuracy_knn)

# Determine which method performs the best
best_model = max(test_accuracy_logreg, test_accuracy_svm, test_accuracy_tree, test_accuracy_knn)
print("The best performing model is:", best_model)
```

```
Test accuracy for Logistic Regression: 0.8333333333333334
Test accuracy for SVM: 0.8333333333333334
Test accuracy for Decision Tree: 0.7777777777777778
Test accuracy for KNN: 0.8333333333333334
The best performing model is: 0.8333333333333334
```

Confusion Matrix

Data Analysis Explaination:

- The logistic regression can be distinguished between the different classes. The problem is the false positives.



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

True Postive - 12 (True label is landed, Predicted label is also landed)

False Postive - 3 (True label is not landed, Predicted label is landed)

Conclusions

- My analysis found that **most models performed similarly**, with comparable scores and accuracy—no single method stood out as the best
- **Launch success rates have improved over time**, with **KSC LC-39A having the highest success rate**, especially for payloads under 5,500 kg.
- Most launch sites are **near the equator and coast**, optimizing fuel efficiency and safety.
- **Orbits ES-L1, GEO, HEO, and SSO** showed a **100% success rate**, and **heavier payloads correlated with higher success rates**.

Appendix

- SpaceX API-Based Data Collection
- Data Collection Using Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with SQL
- EDA with Data Visualization Tools
- Folium Maps
- SpaceX Dashboard Application
- Machine Learning Prediction

Thank you!