IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alif Adwitiya Pratama
June 28, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) : SQL, Static, and Dashboard
  - Predictive analysis (Classification)
- Summary of all results
  - Valuable insights were garnered from publicly available sources.
  - Exploratory data analysis (EDA) determined the most effective features for predicting the success of launchings
  - The resulting model accurately predicted the likelihood of successful launchings based on key characteristics.

Introduction

- Project background and context

In the pursuit of disrupting the space industry, SpaceX has introduced a game-changing approach to rocket launches, offering Falcon 9 launches at an unprecedented $62 million, significantly undercutting competitors. The secret to this cost advantage lies in the company's innovative first stage reusability feature. As a data scientist for a rival startup, this project aims to analyze and develop a machine learning capable of predicting the likelihood of successful first stage landings, enabling informed bidding strategies against SpaceX.

- Problems you want to find answers
    - What are the crucial factors governing the outcome of first stage landings?
    - How do these factors interact and influence the probability of success?
    - What conditions are necessary to maximize the chances of a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data used in this study was gathered by querying the SpaceX REST API and web scraping techniques to extract information from Wikipedia.
- Perform data wrangling
  - filtering, handling missing values, and applying One Hot Encoding 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
  - Developed and fine-tuned classification models utilizing GridSearchCV to optimize hyperparameters and ensure optimal performance

Data Collection

The data collection process employed a two approach, combining API requests from the SpaceX REST API and web scraping from the Wikipedia, ensuring comprehensive information for in-depth analysis. From this collection process, the following data was obtained:

- Using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

Data Colletion Process

- Requesting data from SpaceX API
- Decoding json response
- Constructing and applying custom functions to get only relevant information
- Normalizing and create dataframe
- Filtering for Falcon 9 launches
- Handling missing values with mean imputation
- Export to CSV

Notebook

Data Collection - Scraping

Web Scraping Process

- Requesting data from Wikipedia
- Creating BeautifulSoup object
- Extracting column names from HTML table header
- Parsing HTML tables for data collection
- Constructing and applying custom functions to get only relevant information
- Create dataframe
- Export to CSV

Notebook

Data Wrangling

- Perform exploratory Data Analysis to determine:
  - Number of launches per site
  - Number and occurrence of each orbit
  - Number and occurrence of mission outcome
- Converting outcomes to Training Labels
  - 1 for successful outcome
  - 0 for fail outcome
- Export data to CSV format
- Notebook

EDA with Data Visualization

- Charts Plotted
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Orbit Type vs. Success Rate
  - Flight Number vs. Orbit Type
  - Payload Mass vs Orbit Type and Success Rate
  - Success Rate Yearly Trend
- Types of Charts
  - Scatter plots: show relationships between variables (potential machine learning model inputs)
  - Bar charts: compare discrete categories (e.g. launch sites, orbit types)
  - Line charts: show trends in data over time (time series)
- Notebook

EDA with SQL

Developed and executed SQL queries

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

Notebook

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Created interactive markers with circular symbols, popup labels, and text labels for each launch site, precisely plotted on the map using their latitude and longitude coordinates, to visually represent their geographical locations and proximity to the Equator and coastal regions.

Coloured Markers of the launch outcomes for each Launch Site:

- Implemented coloured Markers denoting successful (Green) and failed (Red) launches using Marker Cluster to pinpoint which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to illustrate distances between the Launch Site KSC LC-39A (as an example) and its proximities, including Railway, Highway, Coastline, and Closest City.

Notebook

Build a Dashboard with Plotly Dash

Launch Site Selection:

- Implemented a dropdown list to facilitate the selection of a specific launch site.

Launch Success Analysis:

- Created a pie chart to display the total number of successful launches across all sites, as well as the success versus failure rates for a selected site.

Payload Mass Filtering:

- Added a slider to enable the selection of a specific payload mass range.

Payload Mass vs. Success Rate Correlation:

- Developed a scatter chart to illustrate the relationship between payload mass and launch success rates for different booster versions.

Notebook

Predictive Analysis (Classification)

Step-by-Step Process for Model Selection and Evaluation

- Create a NumPy array from the "Class" column in the data.
- Standardize the data using StandardScaler, and fit and transform it.
- Split the data into training and testing sets using the train_test_split function.
- Create a GridSearchCV object with cv = 10 to find the best parameters.
- Apply GridSearchCV to LogReg, SVM, Decision Tree, and KNN models.
- Calculate the accuracy on the test data using the .score() method for all models.
- Examine the confusion matrix for all models.
- Find the method that performs best by examining best validation score.

Notebook

Results

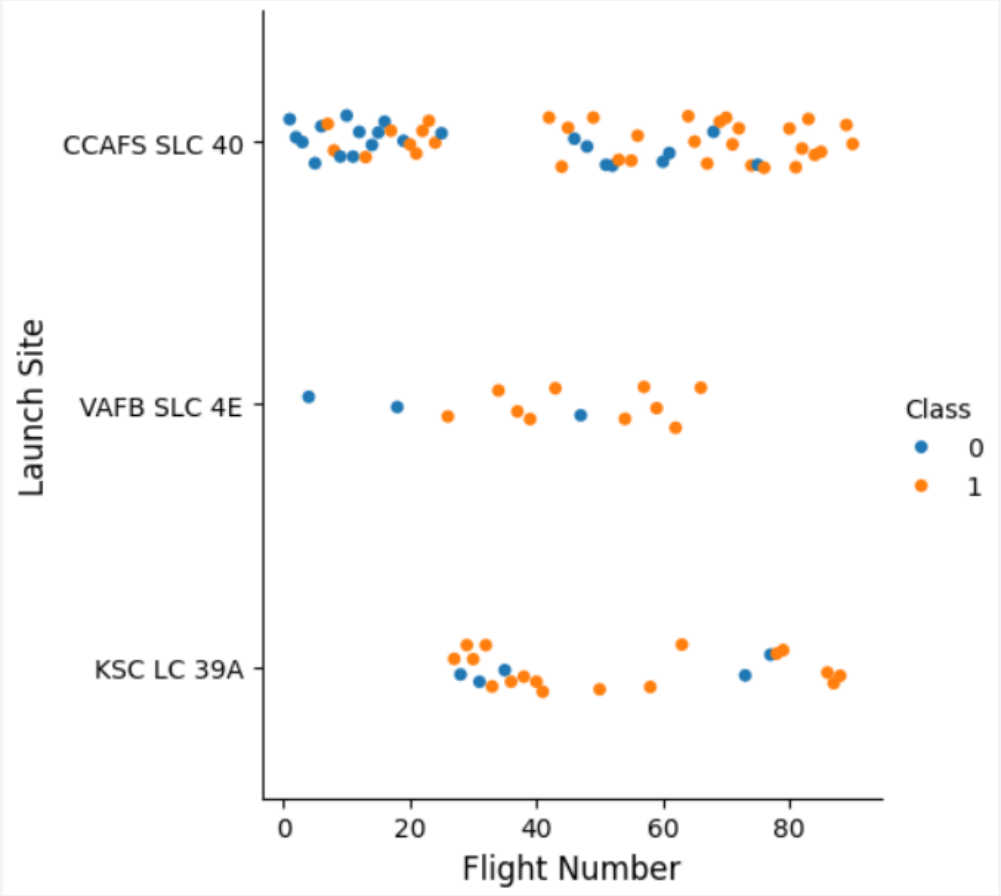
- Insights from Exploratory Data Analysis
- Interactive Analytics Demonstration (as shown in screenshots)
- Predictive Analysis Outcomes

Section 2

Insights drawn from EDA

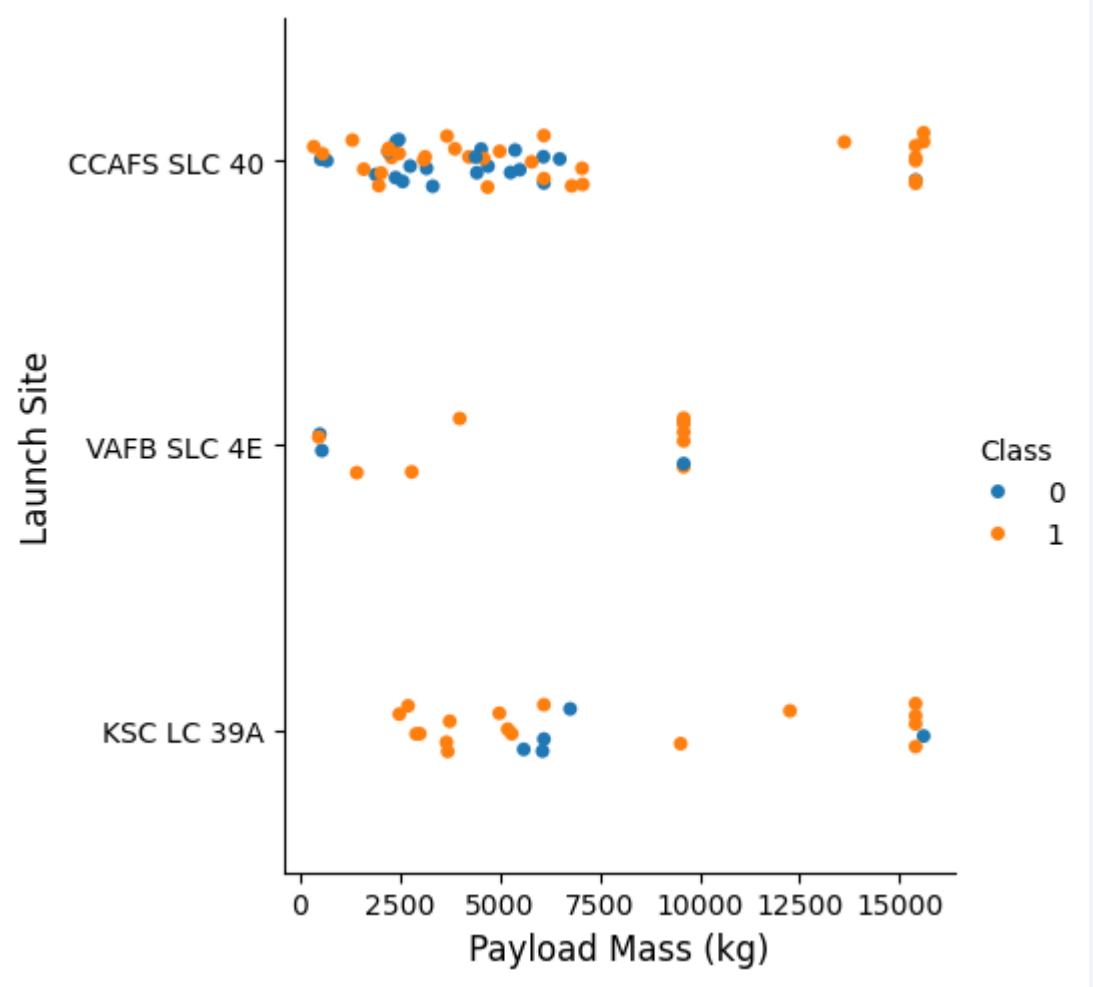
Flight Number vs. Launch Site

- The plot beside reveals that CCAFS SLC 40 is currently the top-performing launch site. KSC LC 39A and VAFB SLC 4E follow closely, ranking second and third, respectively.
- Notably, the data also indicates a steadily improving overall success rate over time, underscoring the progress made in the industry.



Payload vs. Launch Site

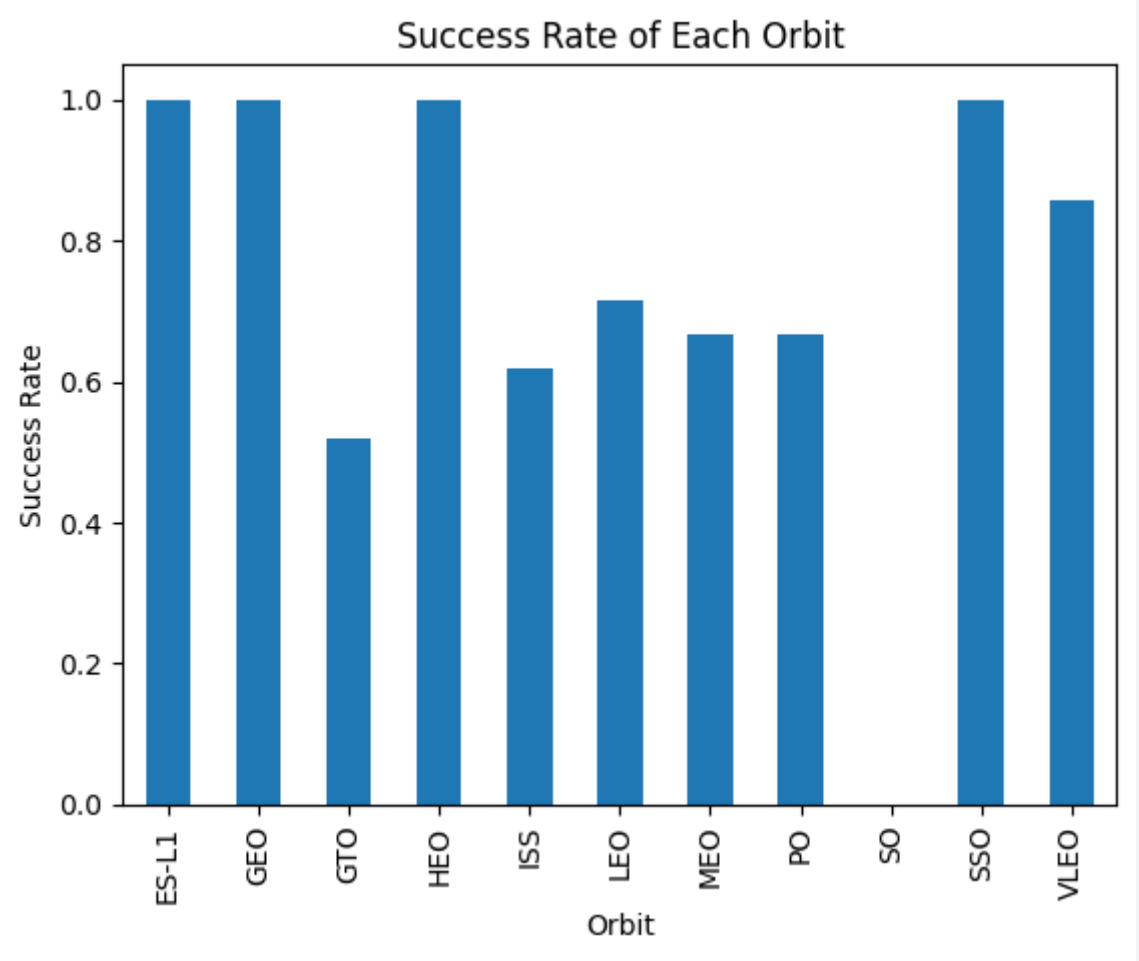
- KSC LC 39A has a 100% success rate for payloads under 5500 kg.
- Payloads over 12,000 kg can only be launched from CCAFS SLC 40 and KSC LC 39A.



Success Rate vs. Orbit Type

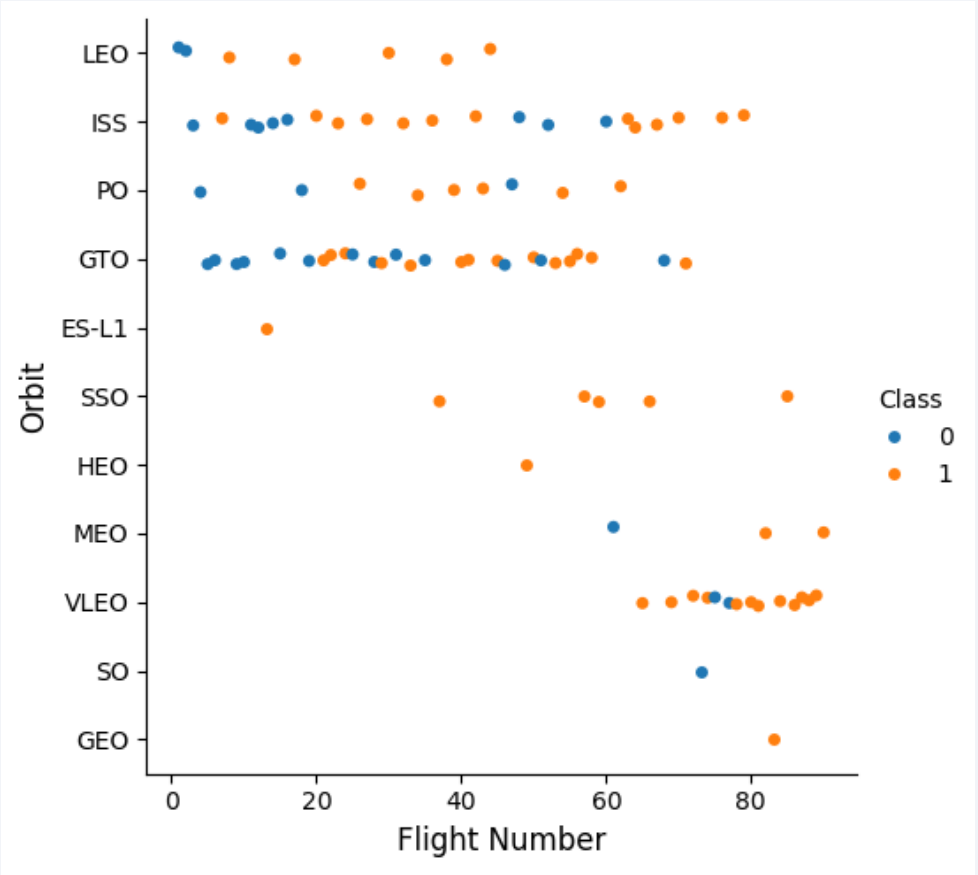
Orbit Recommendation:

- Recommended Orbits: ES-L1, GEO, HEO, SSO (100% success rate)
- Orbits to Avoid: SO (0% success rate)
- Consider with Caution: GTO, ISS, LEO, MEO, PO, VLEO (success rate between 50% and 85%)



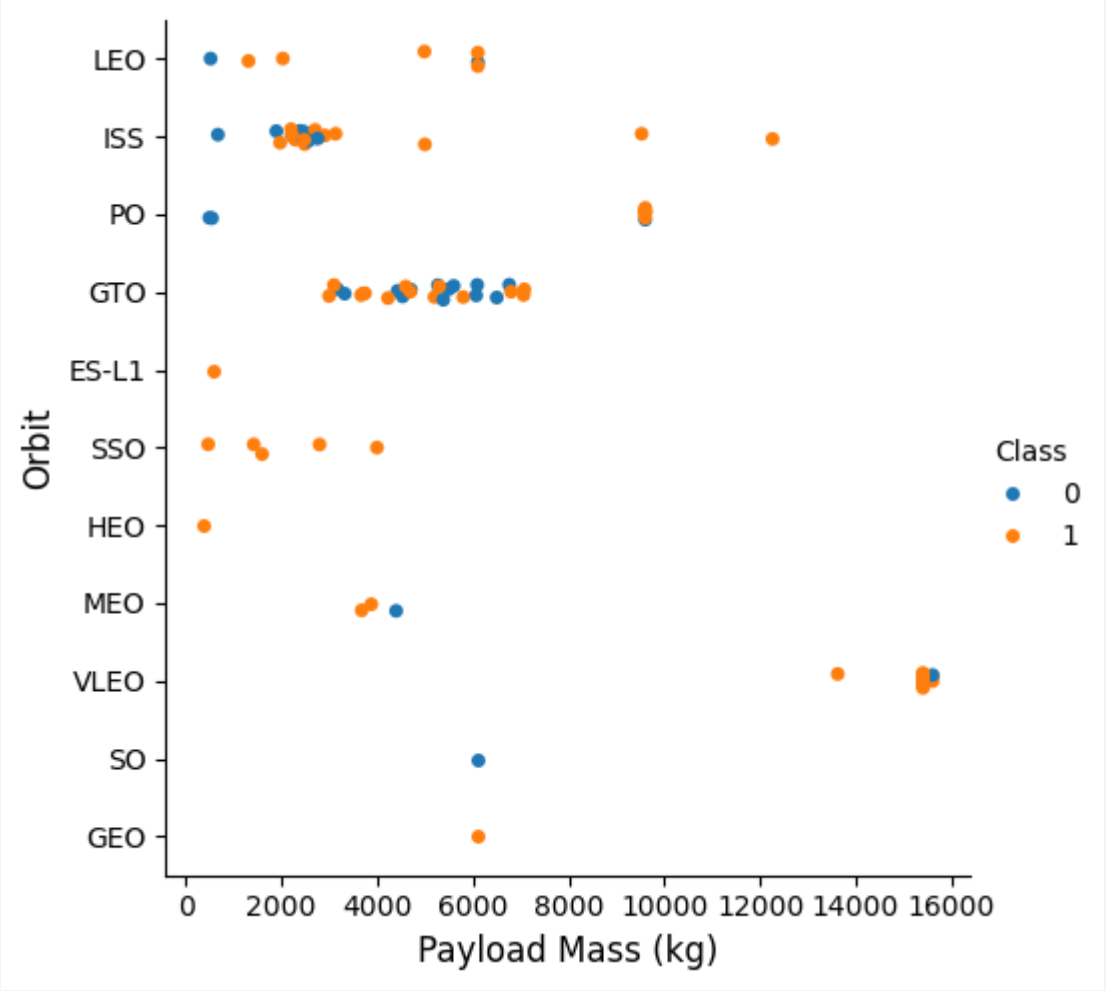
Flight Number vs. Orbit Type

- The success rate of launches has shown a marked improvement over time across all orbits.
- There are few launches to the orbits SO and GEO.



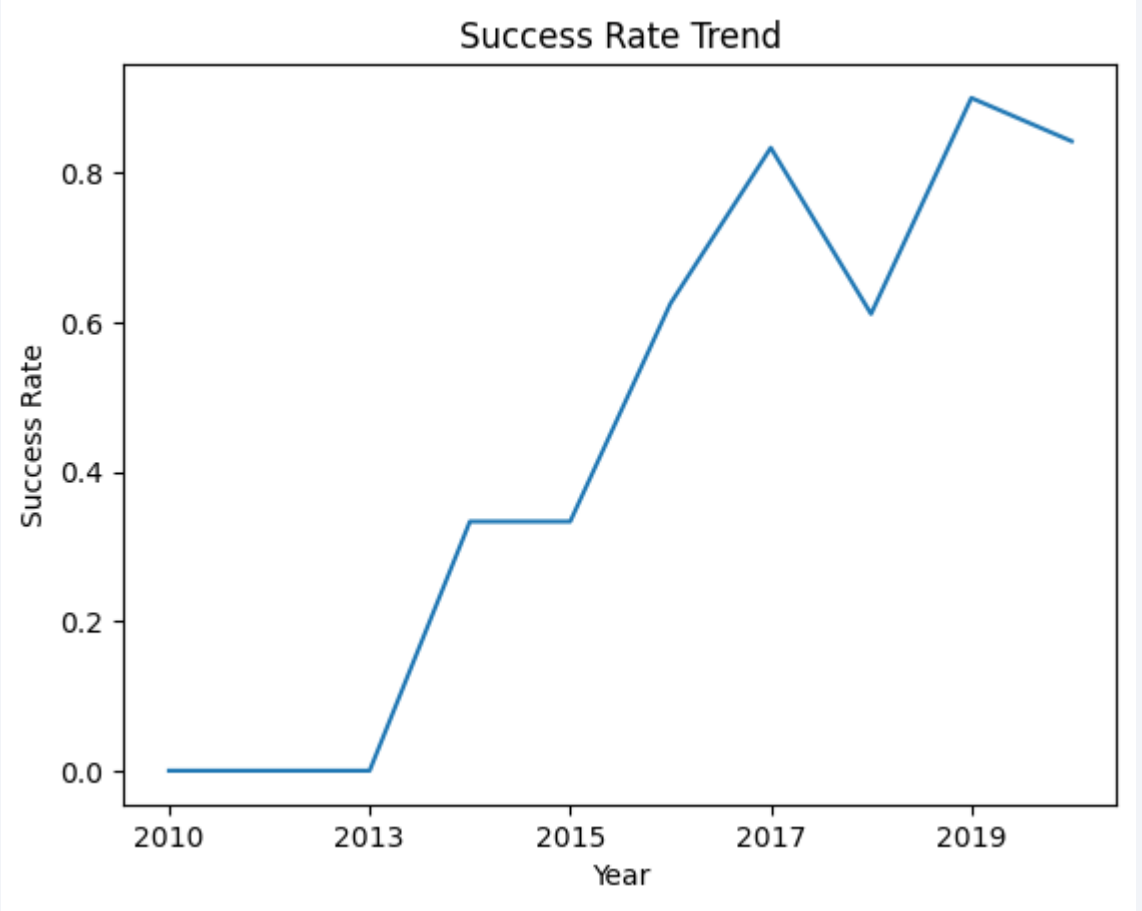
Payload vs. Orbit Type

- The ISS orbit stands out for having the largest range of payloads and a notably high success rate.
- SSO have 100% success rate for payload mass <5000 KG
- VLEO has the highest success rate for payload mass >12000 KG



Launch Success Yearly Trend

- As illustrated by the plot, the success rate has consistently trended upward since 2013, with a steady increase.



All Launch Site Names

- These results were obtained by extracting distinct values of the "launch_site" feature from the dataset.

```
[8]: %sql SELECT DISTINCT launch_site FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

[8]:

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`:

```
[9]: %sql SELECT * FROM SPACEXTABLE WHERE launch_site LIKE 'CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

[9]:

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

- Total payload calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

```
[11]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTABLE WHERE customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
[11]: total_payload_mass
45596
```

Average Payload Mass by F9 v1.1

- Average payload calculated by Filtering data by the booster version and calculating the average payload mass.

```
[12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass FROM SPACEXTABLE WHERE booster_version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

[12]:

average_payload_mass
2928.4

First Successful Ground Landing Date

- This result calculated by filtering data by successful landing outcome on ground pad and getting the minimum value for date.

```
[17]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
[17]: MIN(Date)
      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass betweem 4000 and 6000

```
[19]: %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.
```

[19]:

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

Total Number of Successful and Failure Mission Outcomes

- This query Groups similar "Success"-like mission outcomes together (e.g., "Success", "Success (payload status unclear)", etc.), after that counts the total number of occurrences for each distinct mission outcome and returns a result set with two columns: mission_outcome and total_count

```
[27]: %sql SELECT CASE  WHEN mission_outcome LIKE 'Success%' THEN 'Success'  ELSE mission_outcome  END AS mission_outcome,  COUNT(*) AS total_count  FROM  SPA
 * sqlite:///my_data1.db
Done.
```

[27]:

mission_outcome	total_count
Failure (in flight)	1
Success	100

Boosters Carried Maximum Payload

- This query selects the booster_version from the SPACEXTABLE where the PAYLOAD_MASS__KG_ is equal to the maximum PAYLOAD_MASS__KG_ in the entire SPACEXTABLE

```
[23]: %sql SELECT booster_version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
```

[23]:

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

- This query filtering for rows where the year is 2015 and the landing_outcome contains the phrase "Failure (drone ship)". The query also extracts the month from the Date column, but instead of returning the month as a number, it uses a CASE statement to convert it to a full month name (e.g. "January", "February", etc.). The resulting columns are month, landing_outcome, booster_version, and launch_site.

```
[38]: %sql SELECT  CASE substr(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THEN 'April' WHEN '05' THEN 'May'
 * sqlite:///my_data1.db
Done.
```

[38]:

month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query counts the number of occurrences of each unique landing_outcome in the SPACEXTABLE table, but only for rows where the Date falls between '2010-06-04' and '2017-03-20'. The results are grouped by landing_outcome and sorted in descending order by the count, so the most common landing_outcome appears first.

```
[32]: %sql SELECT landing_outcome, COUNT(*) AS count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY count D
 * sqlite:///my_data1.db
Done.
```

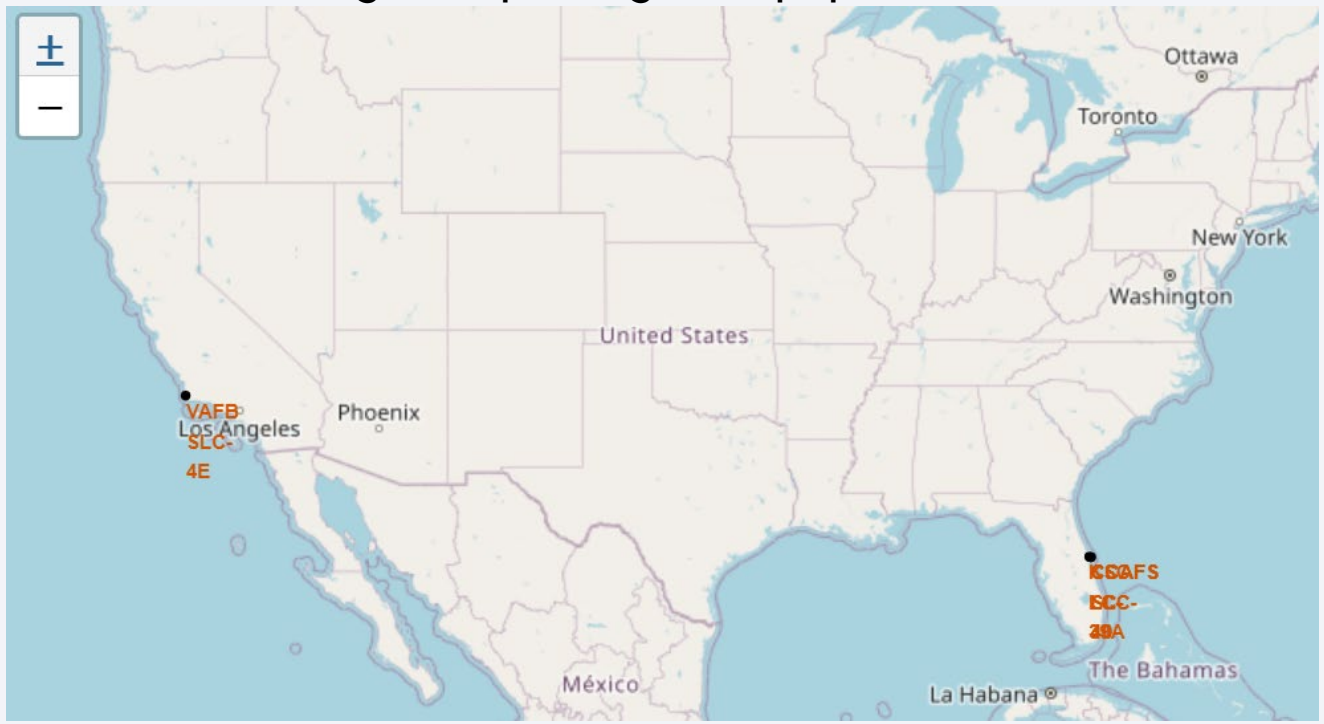
[32]:

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 Sites

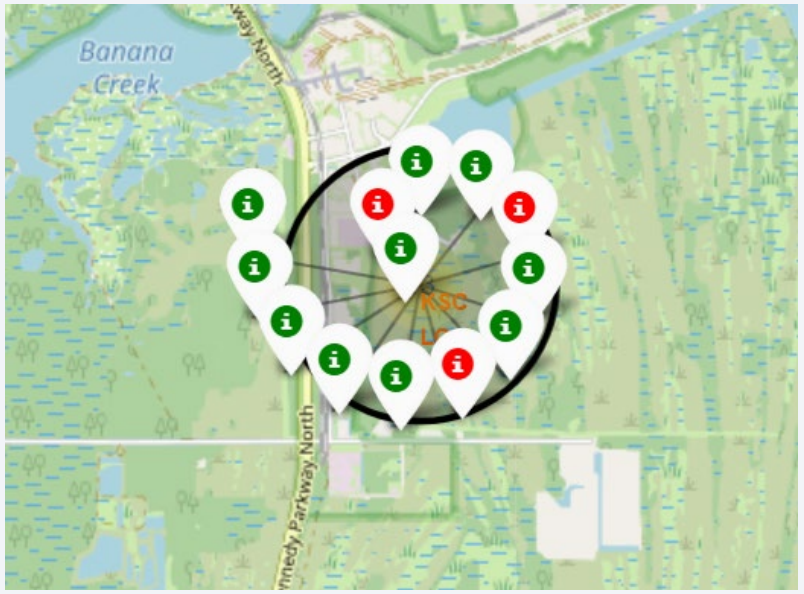
- Most launch sites are located near the Equator, where the Earth's surface moves fastest. This initial velocity helps spacecraft stay in orbit. Additionally, launch sites are situated near coastlines, allowing rockets to launch over the ocean, minimizing the risk of debris falling or exploding near populated areas.



Launch Records by Colour Code on the Map

Easily Identify Launch Site Success Rates with Color-Coded Markers:

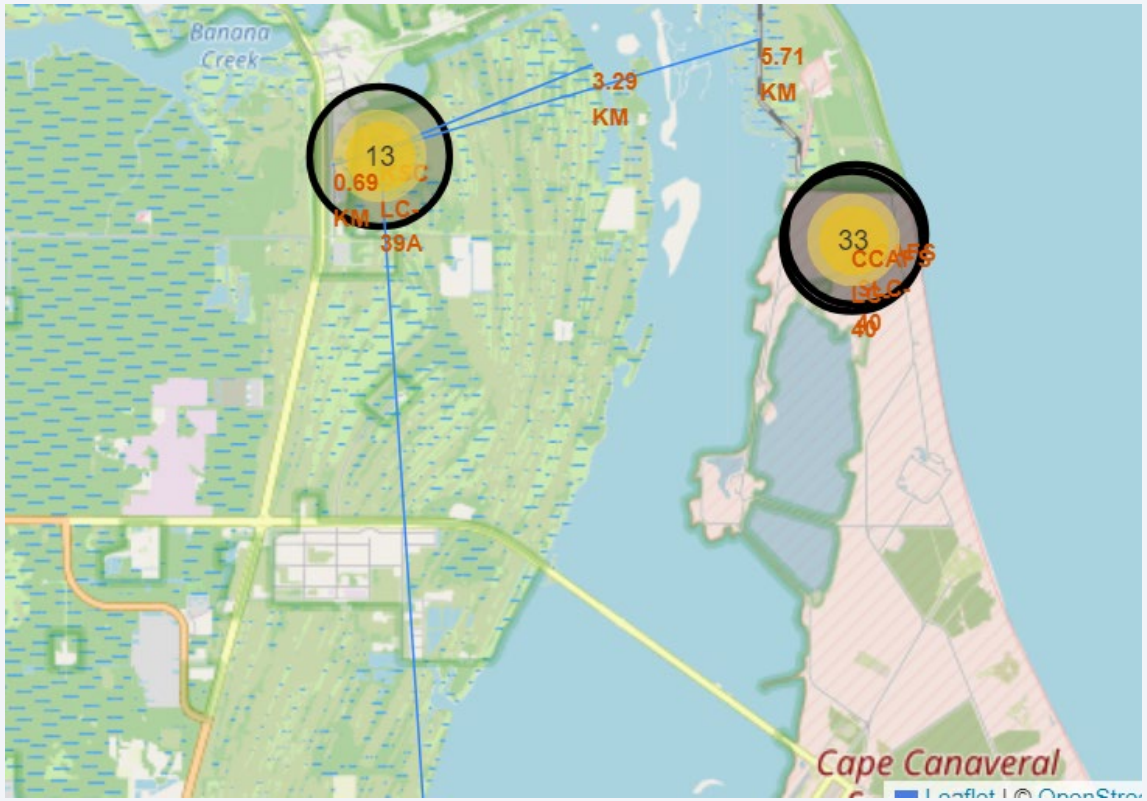
- Green: Successful Launch
- Red: Failed Launch

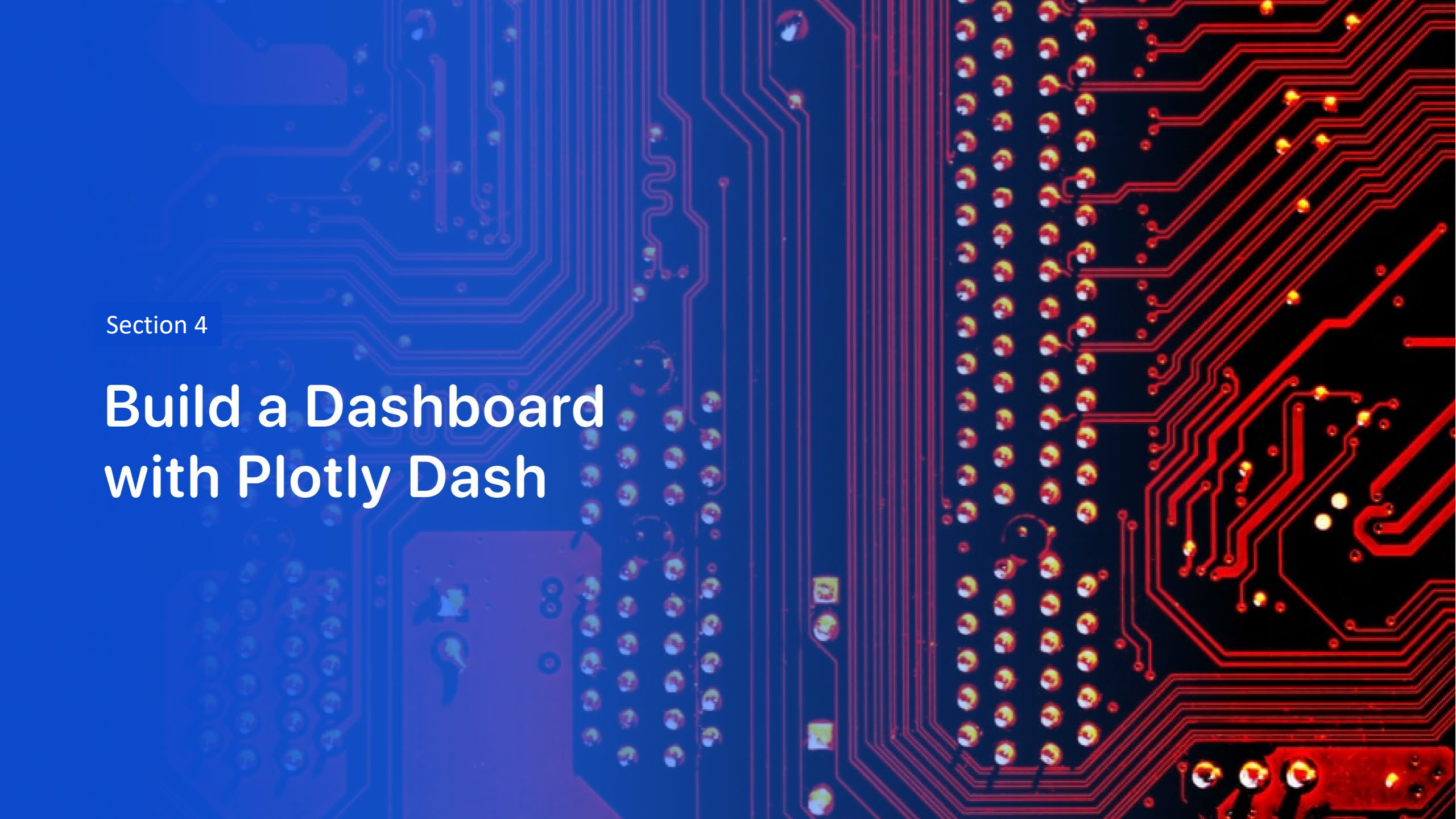


Distances to nearby locations from KSC LC-39A launch site

From the visual analysis of the launch site KSC LC-39A, we can clearly see that:

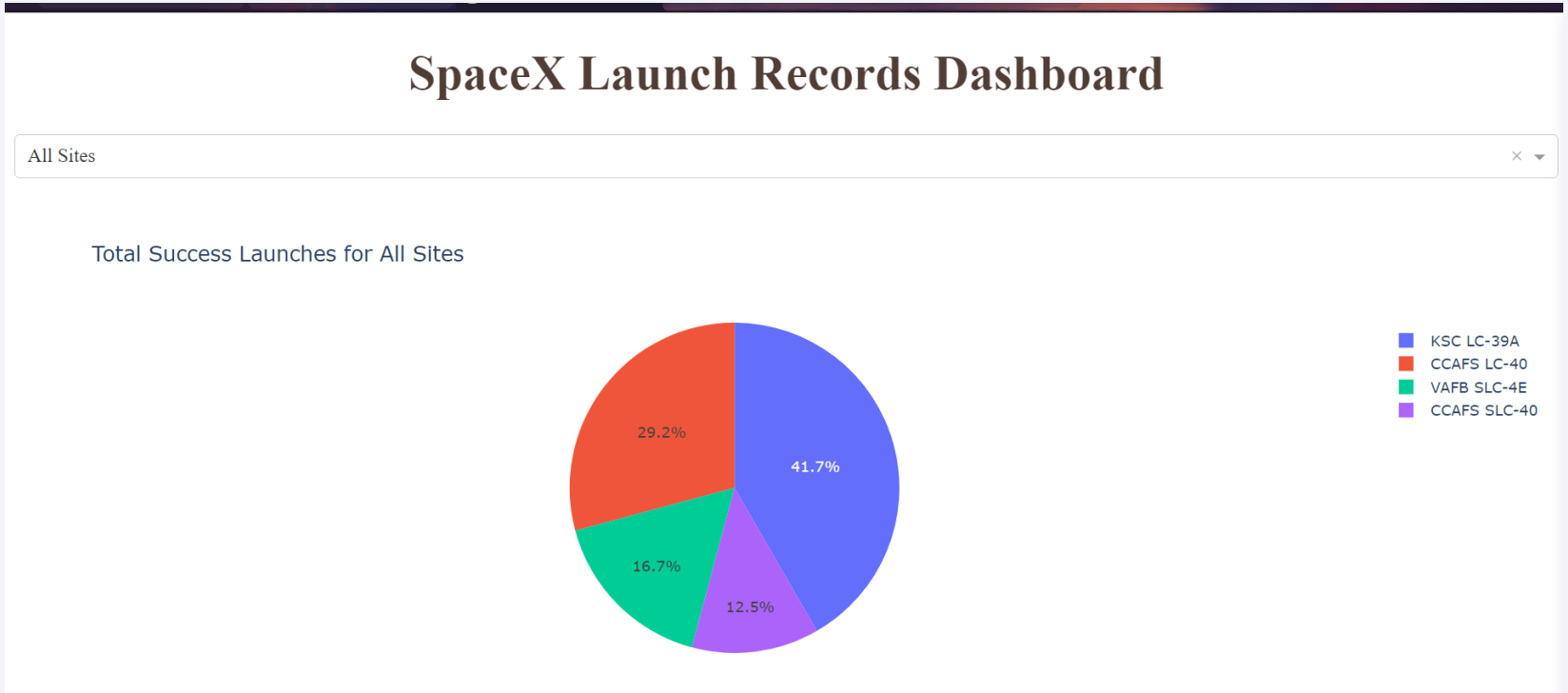
- Railway: 5.71 km - The launch site is situated relatively close to a railway, allowing for easy transportation of crew, equipment, and resources.
- Highway: 0.89 km - The proximity to a highway enables quick and convenient access to the launch site for personnel, equipment, and emergency services.
- Coastline: 3.29 km - The launch site's coastal location allows for over-water launches and provides a natural buffer zone for safety purposes.
- Closest city (Melbourne): 55.26 km - The launch site is intentionally located far from the closest city, Melbourne, for safety reasons, ensuring that the risk of accidental damage or injury to populated areas is minimized.




Section 4
Build a Dashboard with Plotly Dash

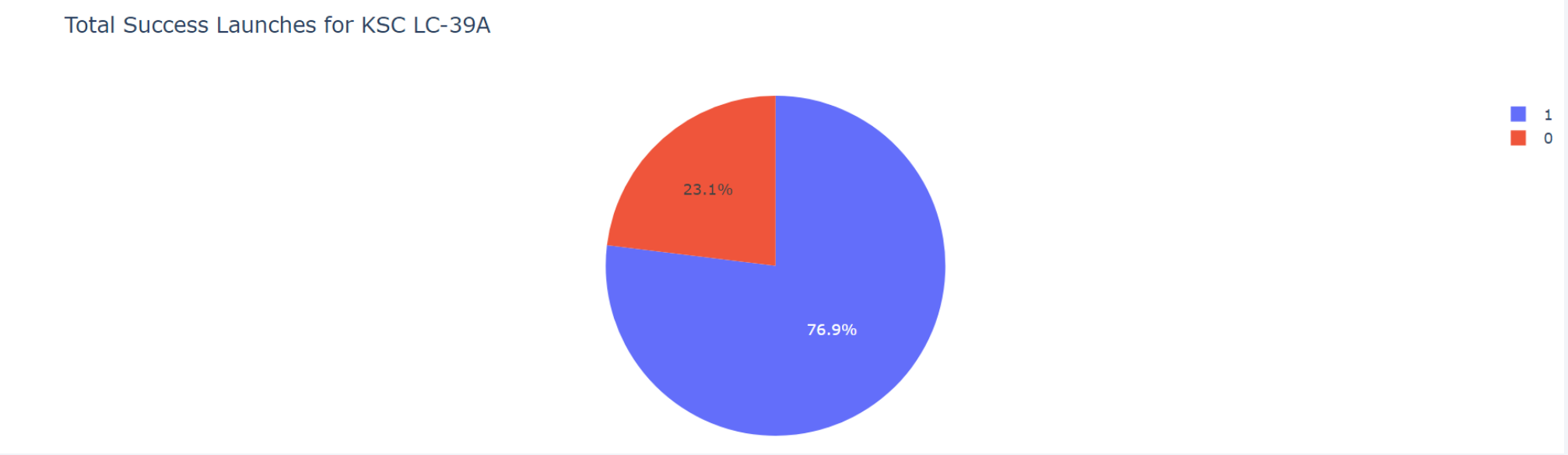
Successful Launches by Site

According to the chart, KSC LC-39A boasts the highest number of successful launches among all sites.



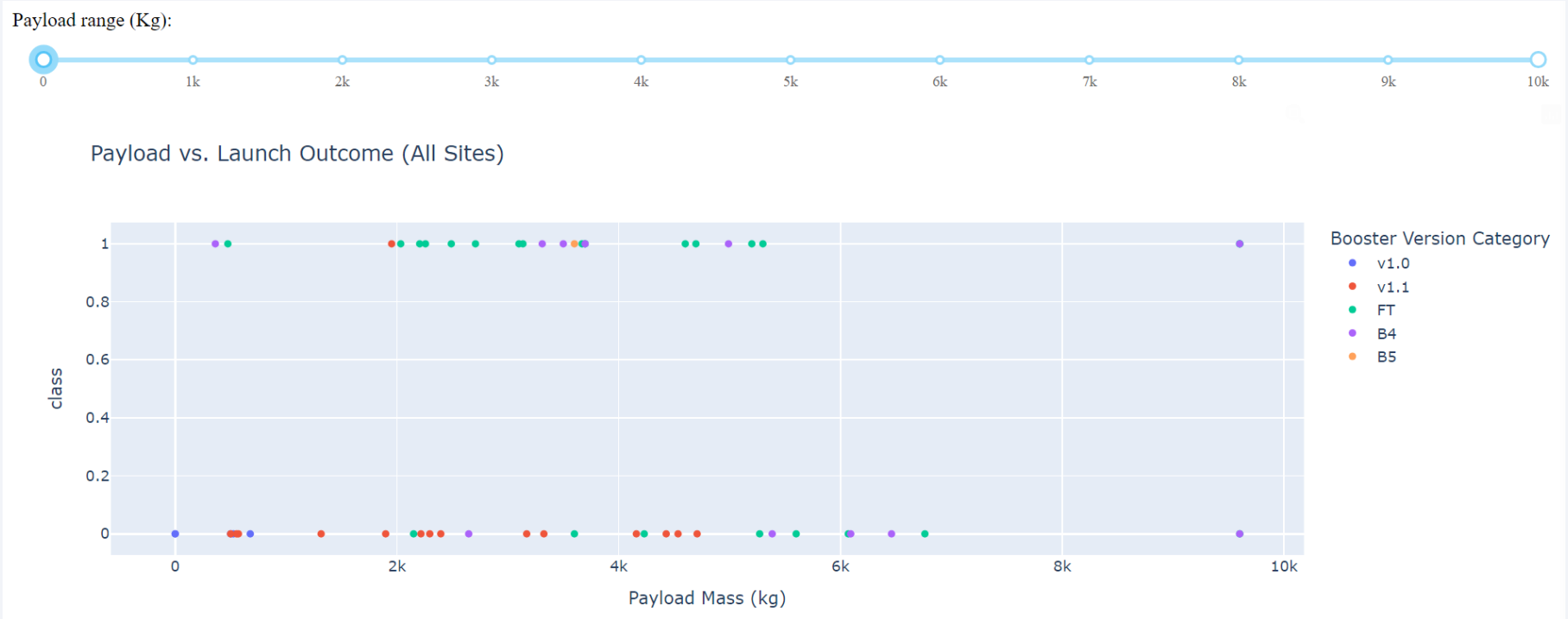
KSC LC 39A Launch Success Ratio

- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



Payload vs. Launch Outcome

- The most successful launch configuration is achieved when pairing payloads weighing under 6,000kg with FT boosters, resulting in the highest success rate.

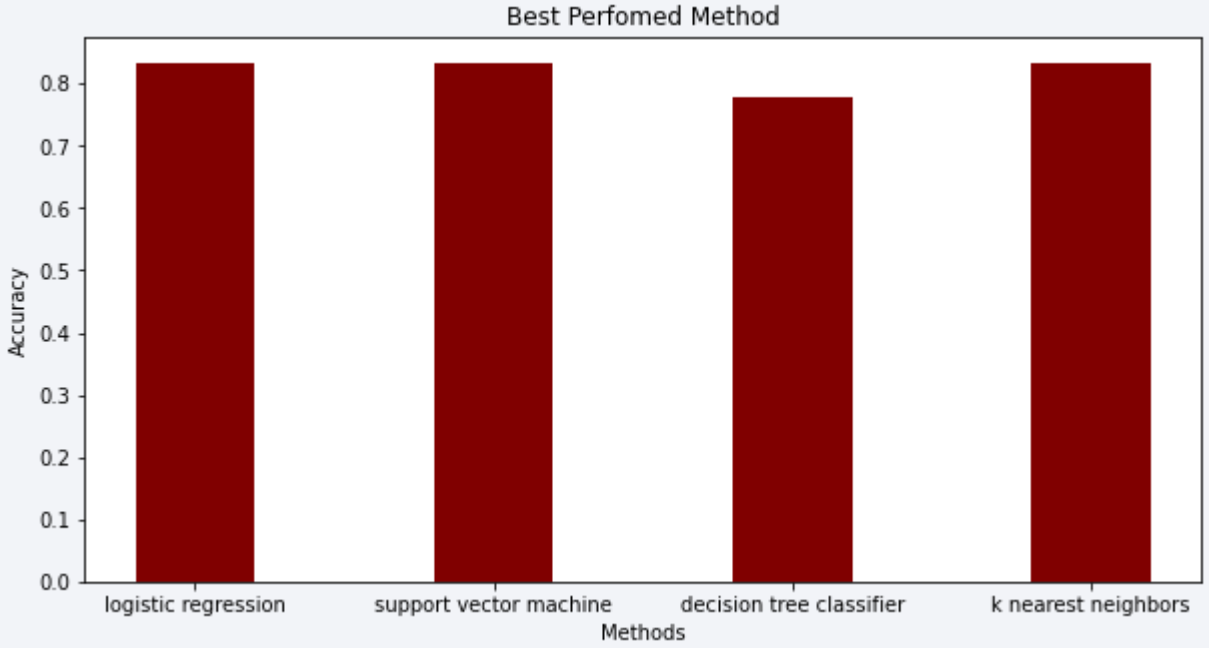


Section 5

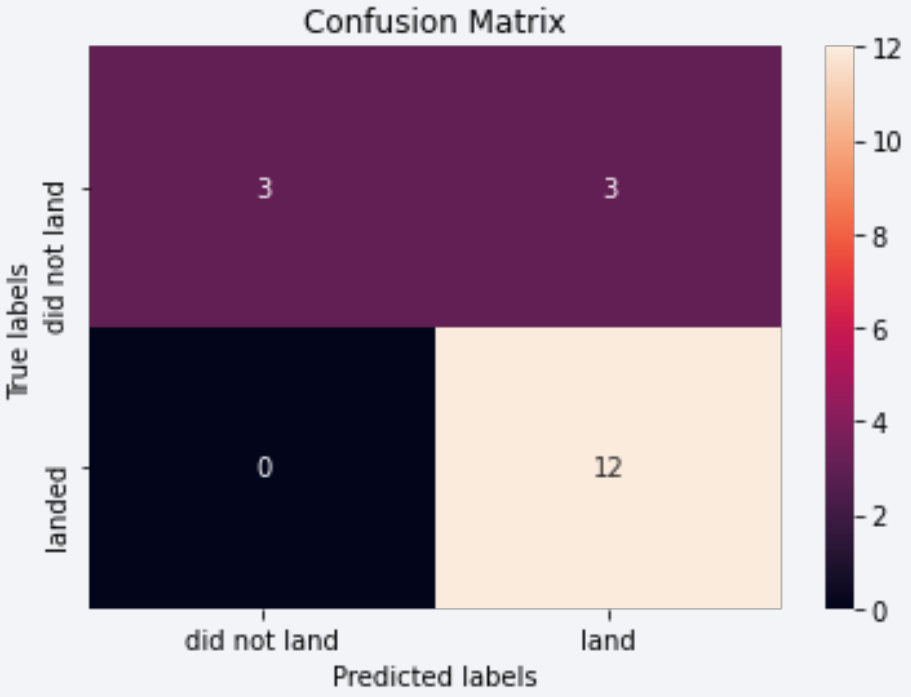
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted.
- In my case, I found that there is no single best model, as three models achieved the same validation accuracy. Therefore, it is reasonable to choose the simplest model, which is logistic regression, due to its lightweight nature.



Confusion Matrix



- The confusion matrix for the logistic regression classifier reveals that the classifier is capable of differentiating between the various classes. However, a major issue arises from the false positive errors, where the classifier mistakenly identifies unsuccessful landings as successful landings.

Conclusions

- It was found that launches with lower payload masses tend to have better outcomes compared to those with larger payload masses.
- The majority of launch sites are located in close proximity to the equator, and all of them are situated near the coastline
- Among all the launch sites, KSC LC-39A boasts the highest success rate for launches.
- The ideal orbit recommendations include ES-L1, GEO, HEO, and SSO, with a remarkable 100% success rate.
- The success rate of launches increases over the years
- logistic regression model is the best choice to predict the outcome of this dataset.

Thank you!