IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nimra Ghazal

11-03-2-24

Table of Content

01 [Executive Summary]
A brief overview of key findings and conclusions.

02 [Introduction]
The background and outlines of problems addressed.

03 [Methodology]
The summary from data collection to model evaluation.

04 [Results]
The key findings and outcomes of the project.

05 [Conclusion]
The summary of key insights drawn

06 [Appendix]
Supplementary information

Executive Summary (1/2)

Objective:

- **SpaceX has revolutionized the commercial space industry with cost-effective launches**, particularly with the **Falcon 9 rocket**. In response, **Space Y aims to compete with SpaceX by predicting first-stage reusability** using machine learning and public data.
- Our aim was to determine **the factors influencing the success of rocket landings,** assess the **contribution of various features to the success rate**, and identify the most effective algorithm for predicting success.

Methodology:

- We utilized **SpaceX REST API and web scraping from Wikipedia,** performed **standardized, filtered, and encoded data** for analysis and machine learning. Employed various **Exploratory Data Analysis (EDA)** techniques such as scatter plots, bar graphs, and **SQL queries** to analyze relationships and gain insights. **Utilized Folium and Plotly Dash** for interactive visualizations and dashboards. Implemented **classification models and grid search for algorithm** selection and **hyperparameter tuning**.

Executive Summary (2/2)

Key Outcomes:

- Identified various factors contributing to rocket landing success, including booster version, orbit type, payload weight, launch site, and landing type. Orbits like **ES-L1, GEO, HEO, SSO, and VLEO exhibit high success rates**, while **lower-weighted payloads tend to perform better**. Drone Ship considered is observed to be the most used and successful landing type landed among all.
- Launch sites with **larger flight volumes show higher success rates.** Booster versions like **FT, B4, and B5 dominate successful launches**. Kennedy Space Center Launch Complex has the highest success rate of launches. Determined **Decision Tree Classifier as the most effective for predicting** rocket landing success.

Recommendation:

- These findings provide valuable insights for optimizing rocket launch operations and improving success rates in the commercial space industry. **Space Y should prioritize launching payloads into orbits with higher success rates, strategically select launch sites based on safety and operational efficiency, carefully plan payload distribution, and optimize booster version selection.** Leveraging machine learning models like **Decision Trees can aid in predictive analysis and informed decision-making** for maximizing success rates.

Introduction

- The commercial space age has arrived, with companies making space travel more affordable.
- SpaceX launched Falcon 9 **for $62 million, 62.42% cheaper** than other providers costing upwards of $165 million each.
- SpaceX's cost-effectiveness is attributed to **reusability**, particularly with the **recovery of the first stage of Falcon 9 rockets.**
- Despite successful recoveries, there are instances where the first stage does not land, crashes, or is sacrificed due to mission parameters.
- As part of a new rocket company, Space Y, competing with SpaceX, the task is to determine launch prices and predict first stage reusability using machine learning and public data.

Questions We Want To Answer

1. **What factors determine the success of a rocket landing?**
2. **How do various features contribute to the success rate of rocket landings?**
3. **Which algorithm is most effective for predicting the success of rocket landings?**

Section 1

Methodology

Methodology (1/2)

- **Data Collection Methodology**:
  - Utilized **SpaceX REST API** and performed **web scraping from Wikipedia**.
- **Data Wrangling**:
  - Conducted **data standardization, filtering, and one-hot encoding** to prepare data for EDA and machine learning.
- **Exploratory Data Analysis (EDA)**:
  - Utilized **scatter graphs, bar graphs, and line graphs** to analyze relationships between **payloads, launch sites, number of flights, and orbit,** gaining insights into data patterns.
  - Employed **SQL** queries **for statistical insights and summarization** of dataset aspects.

Methodology (2/2)

- **Interactive Visual Analytics**:
  - **Utilized Folium to perform interactive visual** analytics, visualizing launch sites, their success/failure rates, and distances from proximity.
  - **Implemented a Plotly Dash dashboard** to display launch site success/failure rates and utilize scatter plots to visualize success/failure rates of different booster categories with payload mass variation.
- **Predictive Analysis**:
  - Employed **classification models and grid search** to select optimal **algorithms and hyperparameters**.
  - Conducted **comparative analysis of accuracies** to assess model performance.

Data Collection

Data sets were collected by using two ways:

Web API

This method allowed us to receive **structured data** directly from **predefined endpoints**, including **JSON or XML data formats** commonly used for transmitting information over the internet.

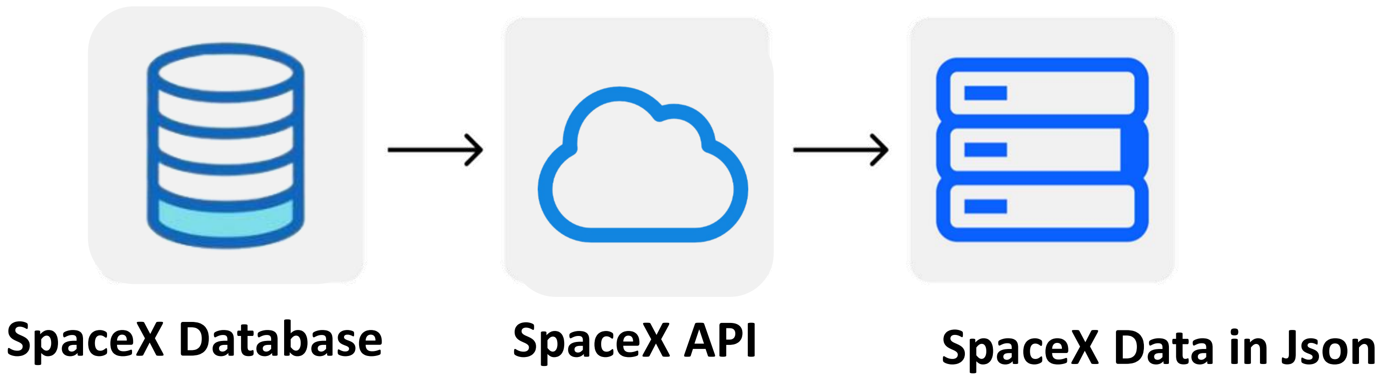
Web Scrapping

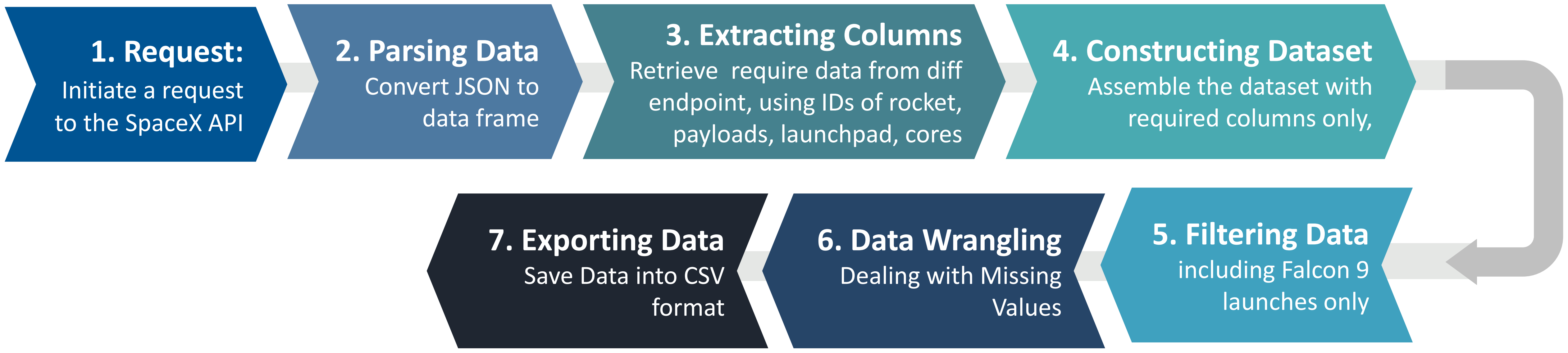
Through this technique, we extracted **unstructured or semi-structured data** from **web pages**, including **text, images, tables**, and other content by parsing **HTML and CSS** elements.

Data Collection – SpaceX API

SpaceX API

The [SpaceX API] is a public API that provides a wealth of information about SpaceX's space missions, that is **launch, rocket, core, capsule, starlink, launchpad, and landing pad data.**





[Notebook - Data collection using api]

Data Collection – Web Scraping

SpaceX Wikipedia Page

The **Historical Falcon 9 launch records from a Wikipedia**, provides extensive information on the SpaceX's **space missions**, encompassing **lists of Falcon 9 and Falcon Heavy launches, launch statistics, historical records of past launches, and insights into upcoming missions.**

SpaceX Wikipedia Page → **Web Scrapping** → **SpaceX Data in HTML elements**

1. REQUEST:
Request the Falcon9 Launch Wiki page from its URL

2. PARSING DATA
Create **BeautifulSoup** Object for html data extraction

3. EXTRACTING COLUMNS
Identify and extract column names from the HTML table headers.

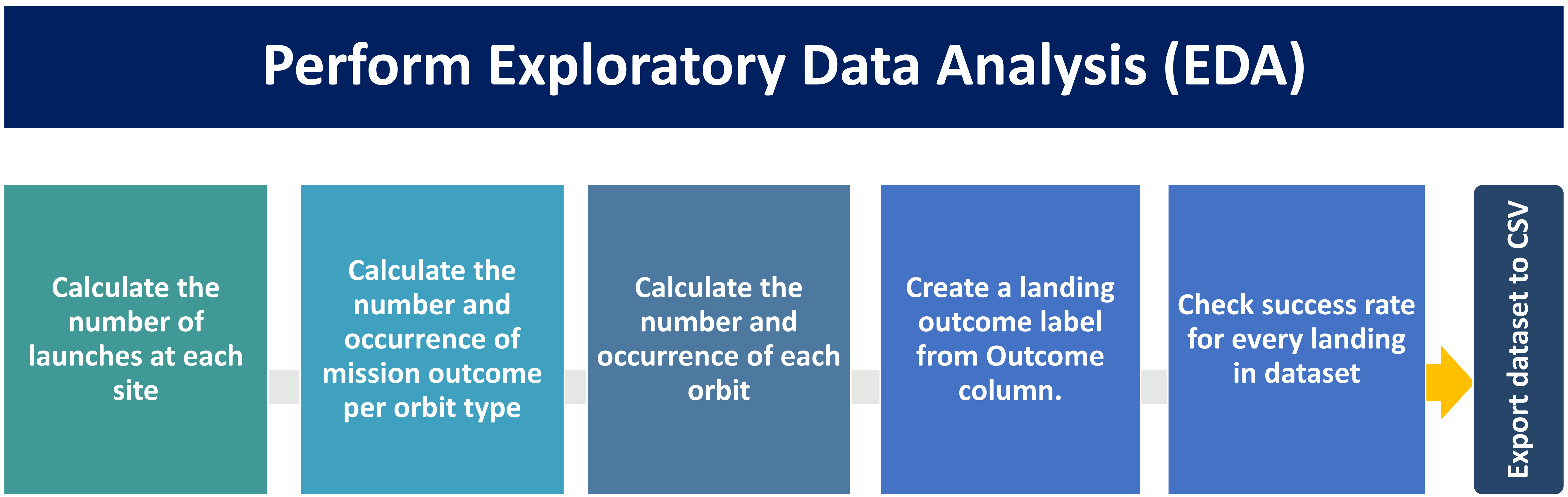
4. CONSTRUCTING DATASET
Populate a dataframe with launch records by iterating through table rows.

7. EXPORTING DATA
Save Data into CSV format

[Notebook - Data collection using web scraping]

Data Wrangling

Data wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze. In this project, we use **One-Hot Encoding** to convert training labels with 1 as landing success and 0as landing failure



During the data wrangling process, several insights and observations were made:

- Data Wrangling Observations- Landing Types and Launch Sites
- Data Wrangling Visualization - Number of launch occurrences in each orbit.
- Data Wrangling Visualization - Number of occurrences in each landing type.

Notebook - Wrangling

EDA with Data Visualization

Seaborn plots were employed for exploratory data analysis, with specific graphs serving distinct purposes:

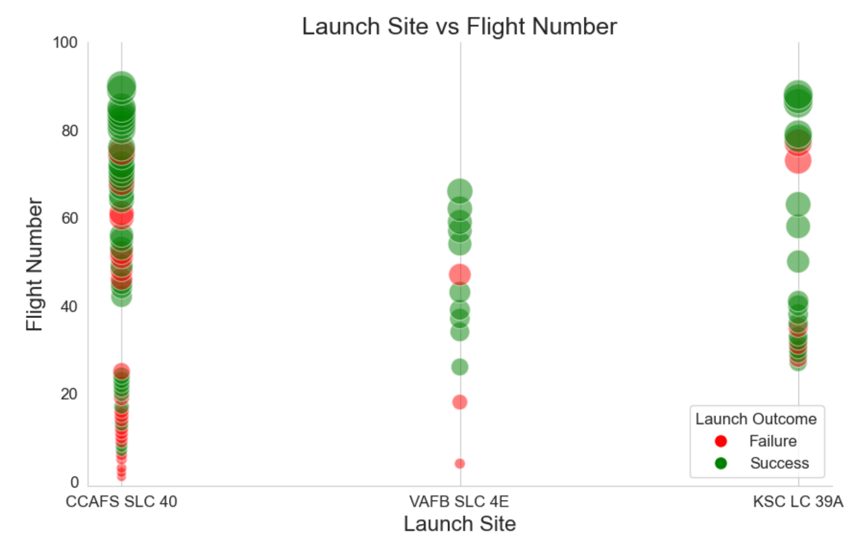
Scatter Plots: Employed to determine correlations between variables, including:

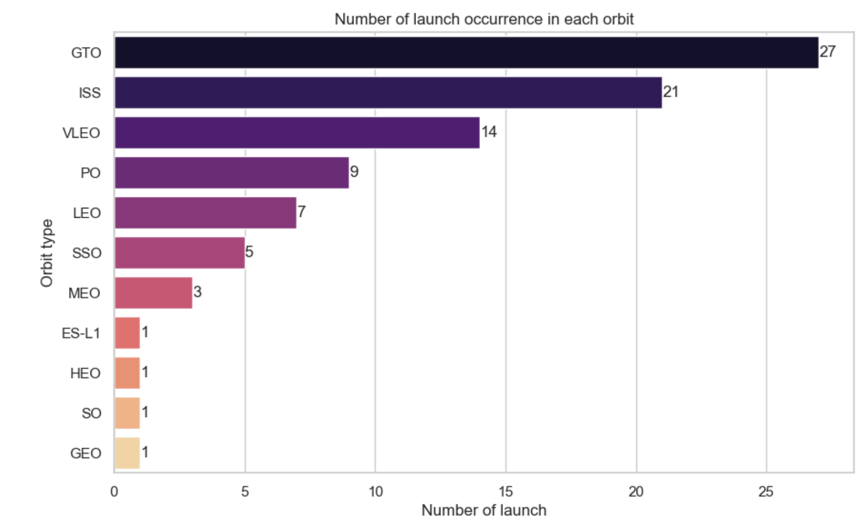
- [Flight number vs. launch site.]
- [Payload vs. launch site.]
- [Flight number vs. orbit type.]
- [Payload vs. orbit type.]

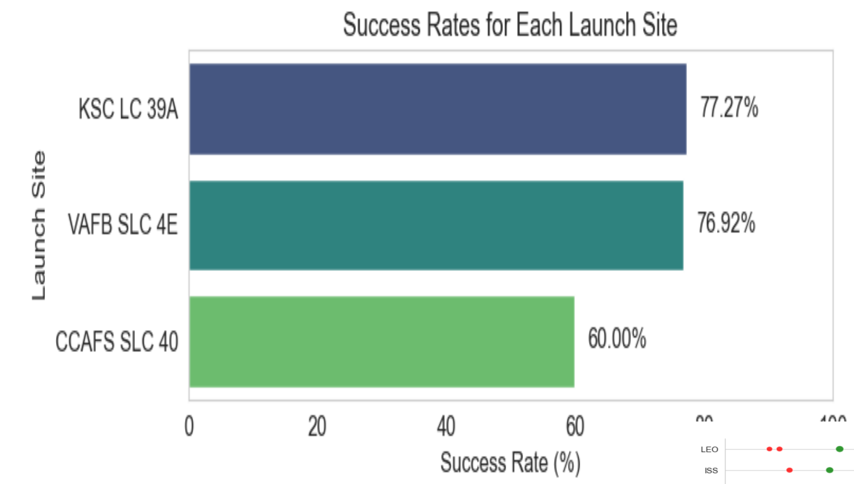
Bar Graphs: Utilized to represent and compare discrete data across different categories or groups. include:

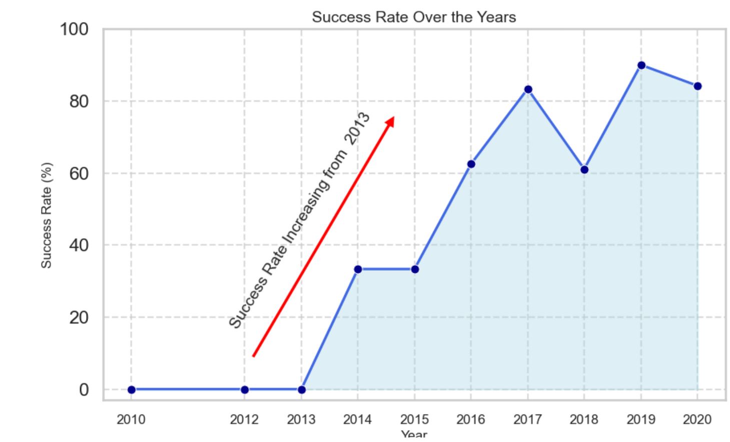
- [Number of launch occurrences in each orbit.]
- [Number of occurrences in each landing type.]
- [Success rate vs. orbit type.]
- [Success rate for each launch site.]
- [Success rate of each landing type.]

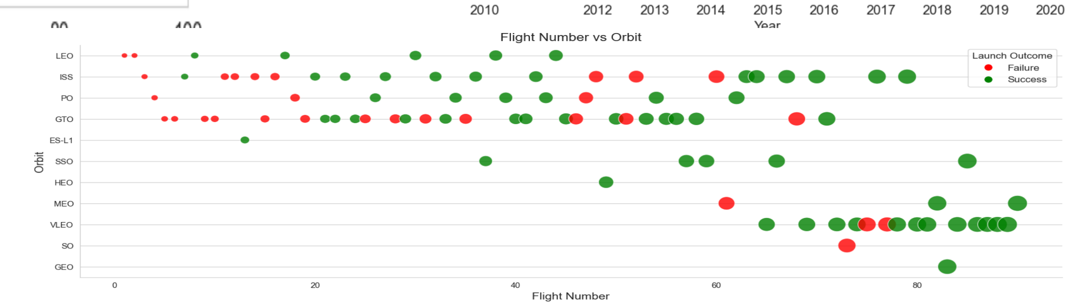
[Notebook - EDA with Data Visualization]











Line Charts: Utilized to visualize trends or changes over time. Specifically used for:

[Launch success yearly trend.]

These visualization techniques were chosen to effectively analyze and interpret the dataset's characteristics and trends.

EDA with SQL

SQL is an essential tool for querying relational databases like IBM's DB2 for Cloud, aiding data scientists in understanding and analyzing datasets. **Queries were conducted to extract information about launch site names, mission outcomes, payload sizes, booster versions, and landing outcomes, providing valuable insights for the project**.

The following SQL queries were performed:

1. Names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7. List the total number of successful and failure mission outcomes.
8. List the names of the booster versions which have carried the maximum payload mass. Use a subquery.
9. 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.
10. 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.

Notebook - EDA with SQL

Build an Interactive Map with Folium

Launch Sites Locations Analysis with Folium

Mark all launch sites on a map	Mark the success/failed launches for each site on the map	Calculated the distances between a launch site to its proximities

We utilized various visualization elements on **Folium Maps, including markers, circles, lines, and marker clusters.** These elements served specific purposes:

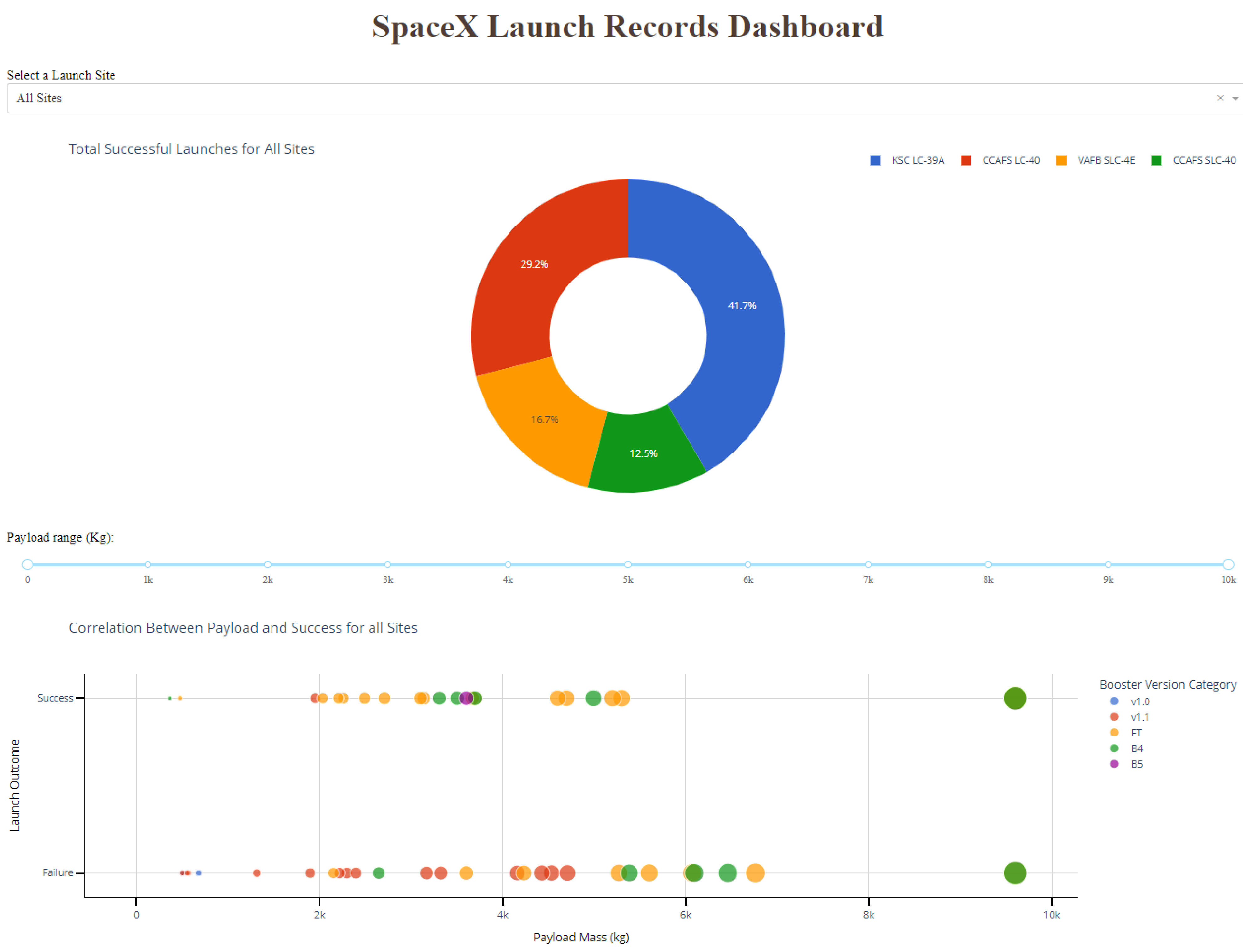
- Markers pinpointed the locations of launch sites.
- Circles highlighted areas around specific coordinates, such as the NASA Johnson Space Center and all three launch sites.
- Marker clusters grouped successful/failure launch occurrences at particular sites.
- Poly Lines illustrated distances between two coordinates, aiding in spatial analysis.
- Haversine's formula was employed to calculate distances from the Launch Site to landmarks like Railway, Highway, Coastline, and Nearest City.
- This facilitated trend identification and pattern analysis in the vicinity.
- The lines drawn on the map visually represented these distances, further aiding in spatial analysis.

Github - Launch Sites Locations Analysis with Folium

jupyter nbviewer Notebook Viewer (for rendered map) - Launch Sites Locations Analysis

Dashboard with Plotly Dash

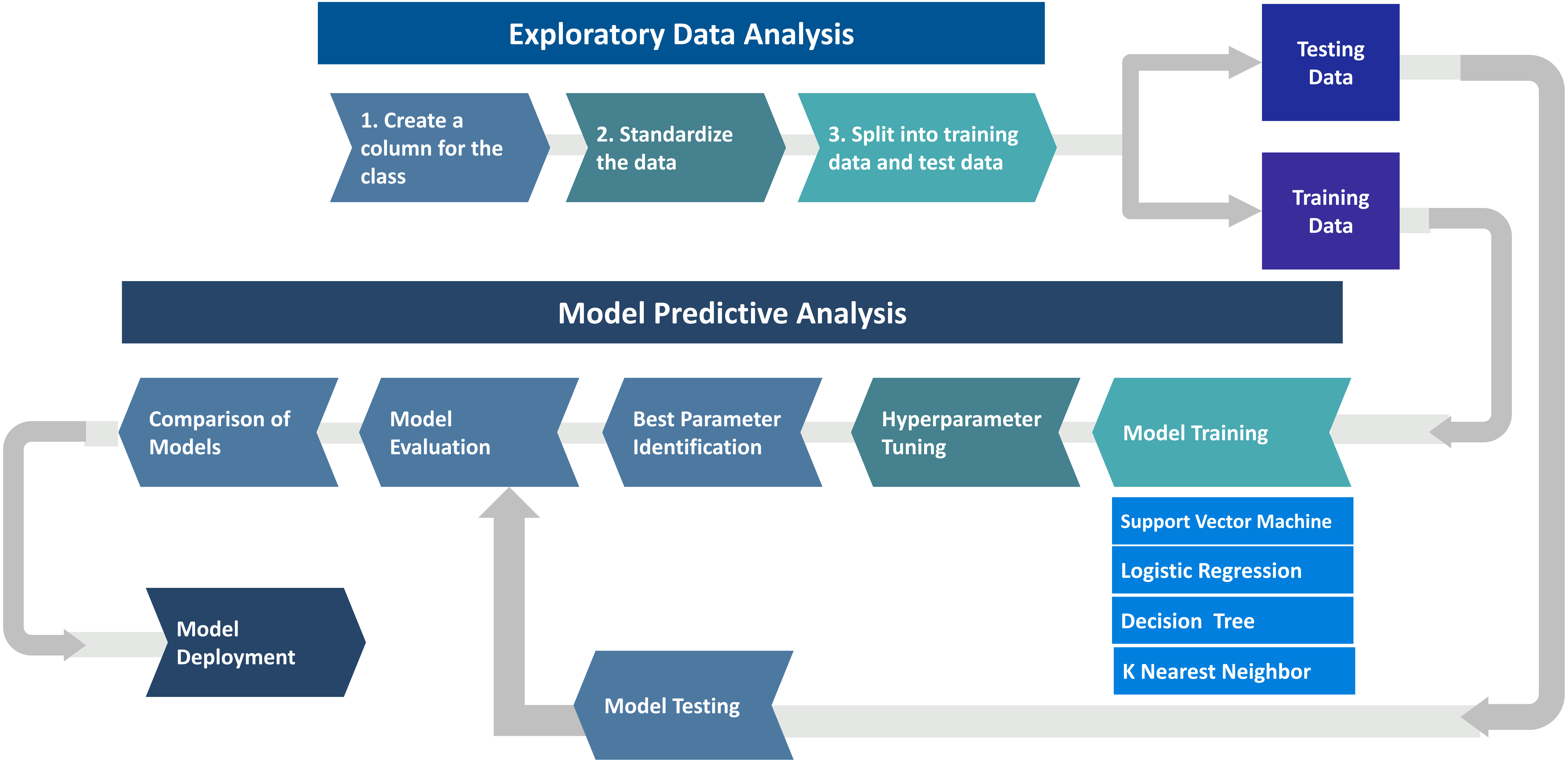
- The dashboard is built using **Flask and Dash web frameworks.**
- It includes **a pie chart and a scatter plot**.
- The pie chart can be toggled to display the distribution of **successful landings across all launch sites or individual launch site success rates.**
- The scatter plot allows selection between viewing data for all sites or a specific site.
- Payload mass can be adjusted using a **slider ranging from 0 to 10000 kg.**
- The pie chart visualizes launch site success rates.
- The scatter plot illustrates the relationship between outcome and payload mass (in kg) for different booster versions.



[Notebook - Dashboard with Plotly Dash]

[Live - SpaceX Launch Records Dashboard]

Predictive Analysis (Classification)



[Notebook - Predictive Analysis (Classification)]

Results – EDA with Visualization

Exploratory data analysis with Visualization reveals several key findings:

- Various factors such as **booster version, orbits, payload weight, launch site, flight volume, and landing type** impact rocket landing success.
- Drone Ship – Most used and successful landing type landed among all.
- Orbits like **ES-L1, GEO, HEO, SSO, and VLEO show higher success rates.**
- **Launch sites with higher flight volumes** tend to have **higher success rates**.
- **The number of successful landing outcomes improved over time.**
- **Overal success rate of landing is 66.67%**

Results – EDA with SQL

Exploratory data analysis with SQL reveals several key findings:

- SpaceX operates from **4 distinct launch sites**.
- Initial launches targeted both SpaceX's own facilities and NASA.
- The average payload of the **F9 v1.1 booster is 2,928 kg**.
- F9 Booster version with the **highest launch success rate: F9 B4 B1039.1**
- The **first successful landing occurred in 2015**, five years after the initial launch.
- Many **Falcon 9 booster versions achieved successful landings on drone ships**, particularly with payloads exceeding the average.
- **Nearly 100% of mission outcomes were successful**.
- Two booster versions, **F9 v1.1 B1012 and F9 v1.1 B1015, failed at landing on drone ships in 2015.**

Results – Interactive Analytics Dashboard

Interactive analytics using Folium and Plotly Dash Key findings:

- **KSC LC-39A** stands out with the **highest number of successful launches among all sites**.
- **Lower payload weights tend to have higher success rates**, with certain mass ranges showing positive correlations.
- A **failed landing occurred even with a payload of 0**, suggesting factors beyond payload mass may influence outcomes.
- Payloads in the range of 1900-1999 Kg exhibit the highest launch success rate.
- **Booster versions FT, B4, and B5 dominate successful launches**, with **B4 and FT** capable of carrying **payloads exceeding 5,000 kg.**
- Launch sites like **CCAFS LC-40** are strategically located near transportation networks and coastlines for efficient operations and minimized risks

Results – Predictive Analysis

Predictive analysis results:

- The **Decision Tree Classifier is the most effective algorithm** for predicting rocket landing success.
- It achieves the highest accuracy among all models**, scoring 89%.**
- Unlike other algorithms, **the Decision Tree model shows improved test accuracy c**ompared to its train accuracy.

Section 2

Insights drawn from EDA

EDA with Visualization

Data Wrangling Observations and Exploratory Data Analysis

Data Wrangling - Observations

There are three landing types defined in the data

Ocean Region

Ground Pad

Drone ship

The data contains following Space X launch facilities:

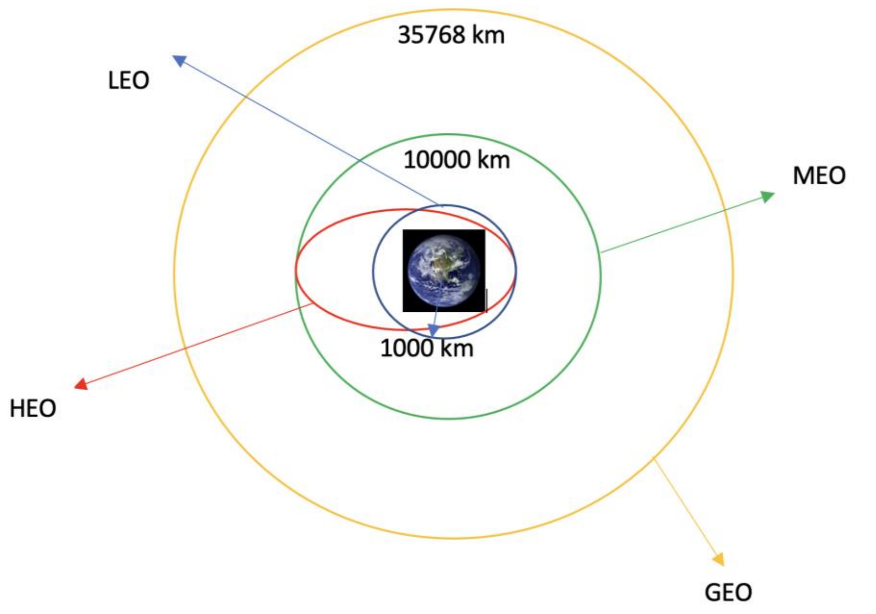
Number of launches for each site

55
Cape Canaveral Space
Launch Complex
40 VAFB SLC 4E

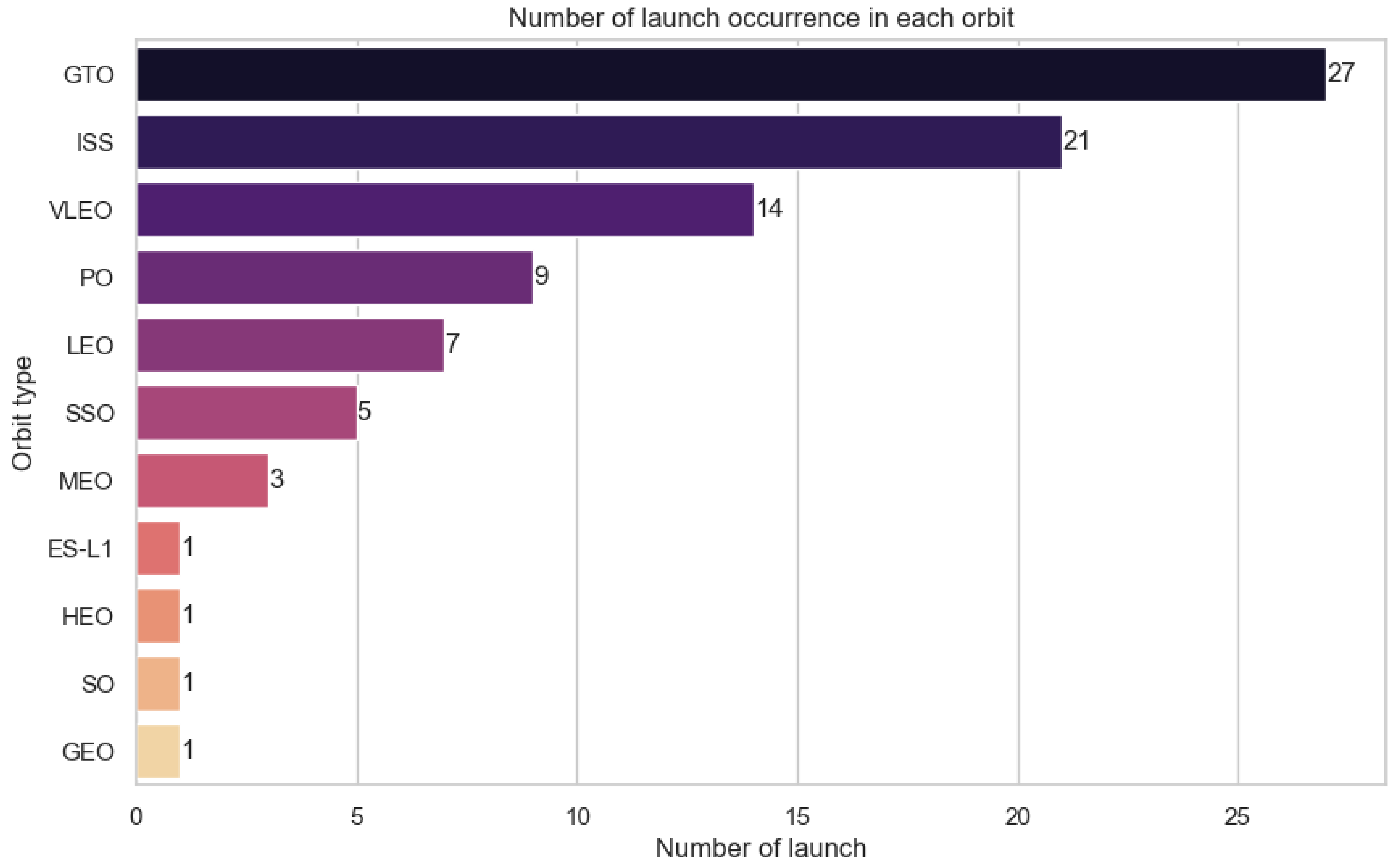
22
Vandenberg Air Force Base
Space Launch Complex
4E (SLC-4E)

13
Kennedy Space Center
Launch Complex 39A
KSC LC 39A

Data Wrangling - Number of Launch Occurrence in Each Orbit.



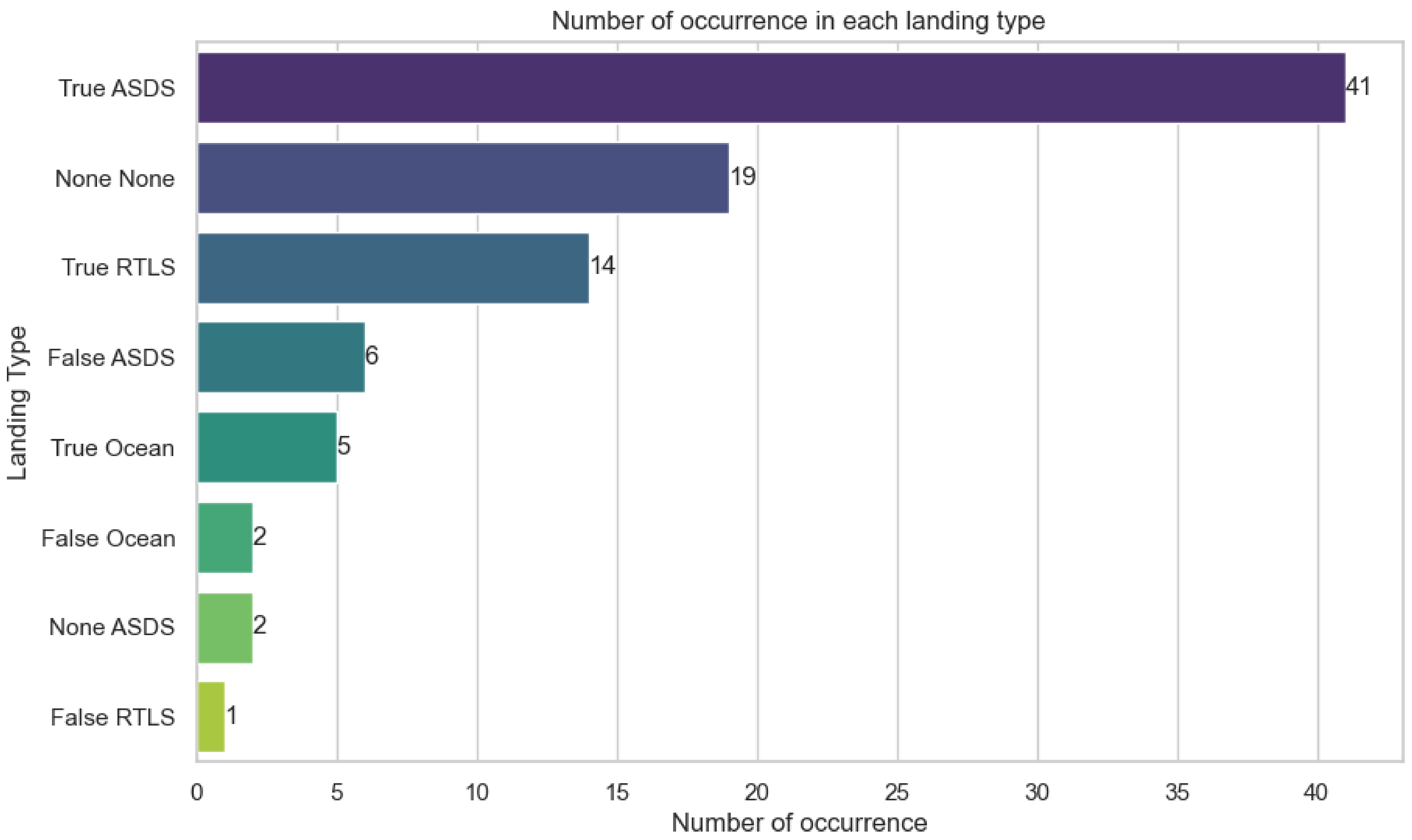
Each launch aims to dedicated orbit, and here are Number of launch occurrence in each orbit.



Geosynchronous orbit (GTO) has the highest number of launch. As it is a **high Earth orbit that allows satellites to match Earth's rotation**. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot **for monitoring weather, communications and surveillance**.
While **HEO, SO, and GEO** has the **lowest occurrences** and all of three as all of three are at **high altitude**

Data Wrangling - Number of Occurrence in each Landing type.

Each launch is specified to land on a particular landing region, here are Number of occurrence in each landing type.

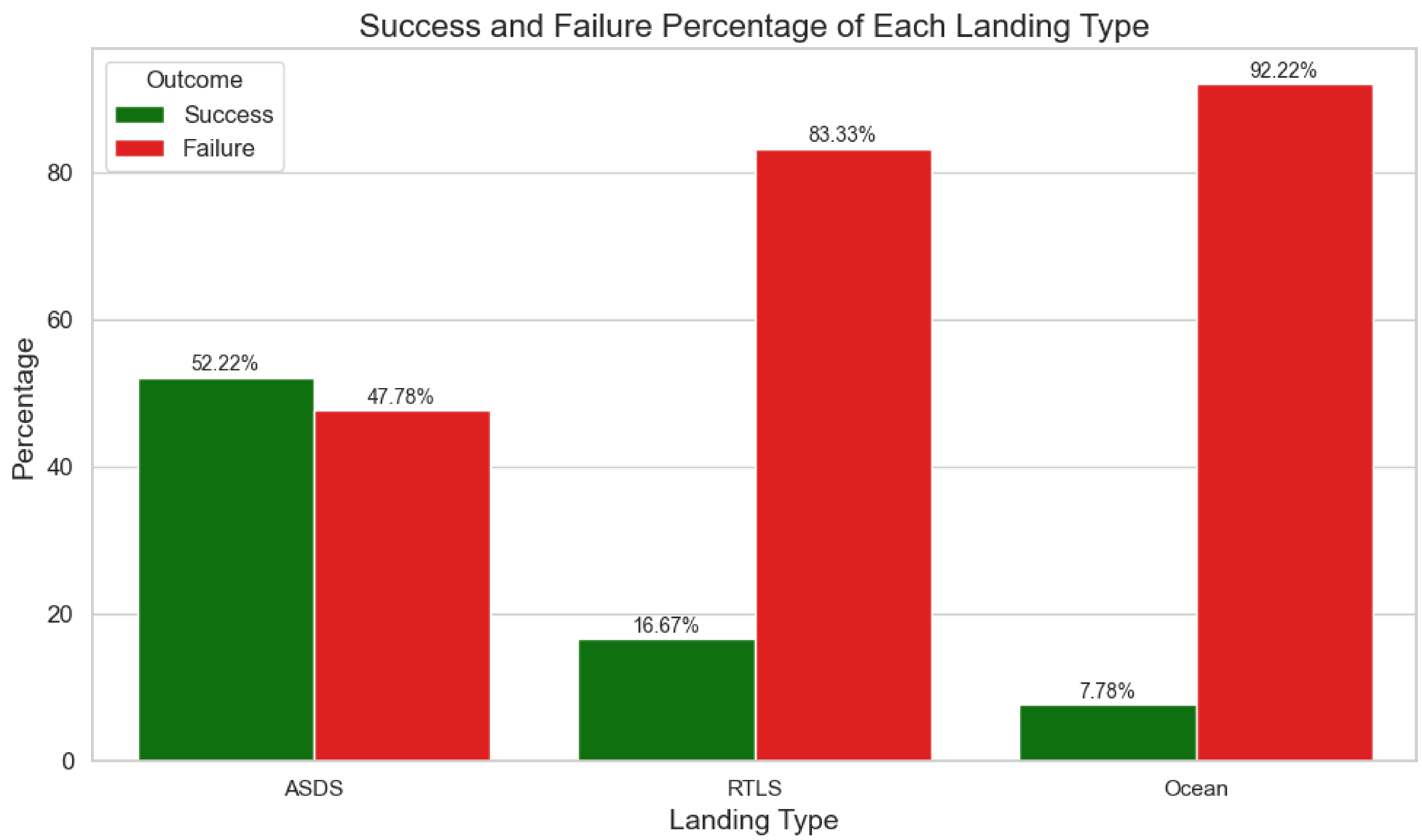


Drone Ship – (True ASDS) Most used and **successful landing type** landed among all.

While Ocean – (False Ocean) has the maximum rate of failure.

Data Wrangling – Success rate of Each Landing Type

The success and failure rates of each landing type show that ASDS has the highest success rate, while ocean landings have the highest failure rate. Overall, the success rate of the attempts follows.



OVERALL, SUCCESS RATE

66.67%

Data Wrangling - Convert outcomes into Training Labels

Created a landing outcome label from Outcome column, by separating where the second stage did not land successfully.

Convert those outcomes into Training Labels with `1` means the **booster successfully landed** `0` means it was **unsuccessful.**

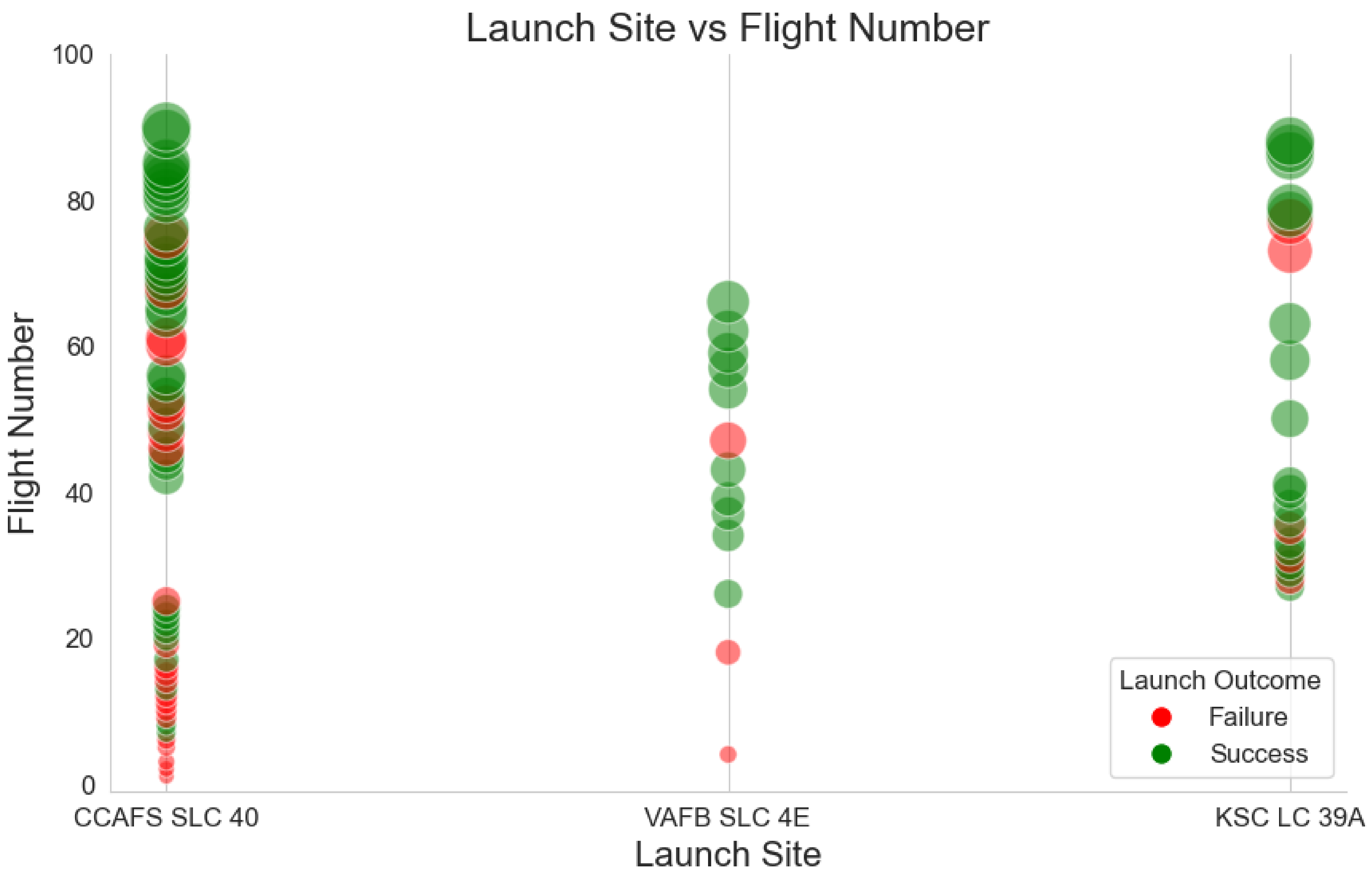
Outcome

- **1 — Successful**: The booster successfully landed
  - **True Ocean**: landed to a specific region of the ocean.
  - **True RTLS**: successfully landed to a ground pad.
  - **True ASDS**: successfully landed on a drone ship.
- **0 — Failed**: The booster Failed the landing
  - **False Ocean**: unsuccessfully landed to a specific region of the ocean.
  - **False RTLS**: unsuccessfully landed to a ground pad.
  - **False ASDS**: unsuccessfully landed on a drone ship.

Flight Number vs. Launch Site

Positive relationship between flight frequency and success rates.

- A **higher number of flights at a launch site correlates with an increased success rate at that site**, suggesting a **positive relationship** between flight frequency and success rates.
- It is evident that the **CCAFS LC-40 launch has more launch attemptsites compared to KSC LC-39A and VAFB SLC 4E**.
- This suggests that **CCAFS** is the **primary launch site due to its higher volume** of launches.

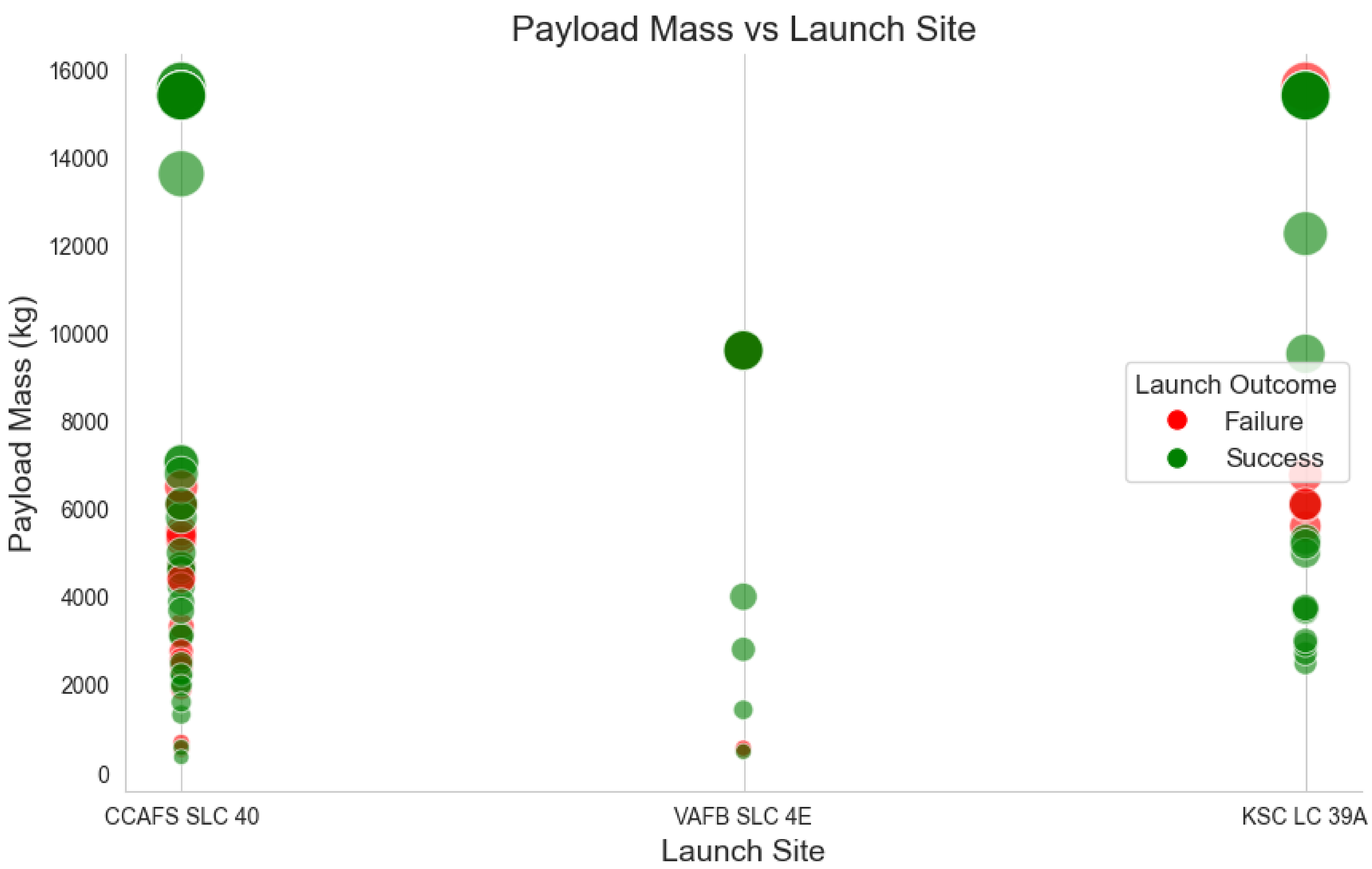


Payload vs. Launch Site

- The **relationship between payload mass and success rate is notably clear at Launch Site CCAFS SLC 40,** where **a higher payload mass tends to correspond with a higher success rate** for the rocket launches.

- However, **there isn't a distinct pattern observed for the other launch sites.**

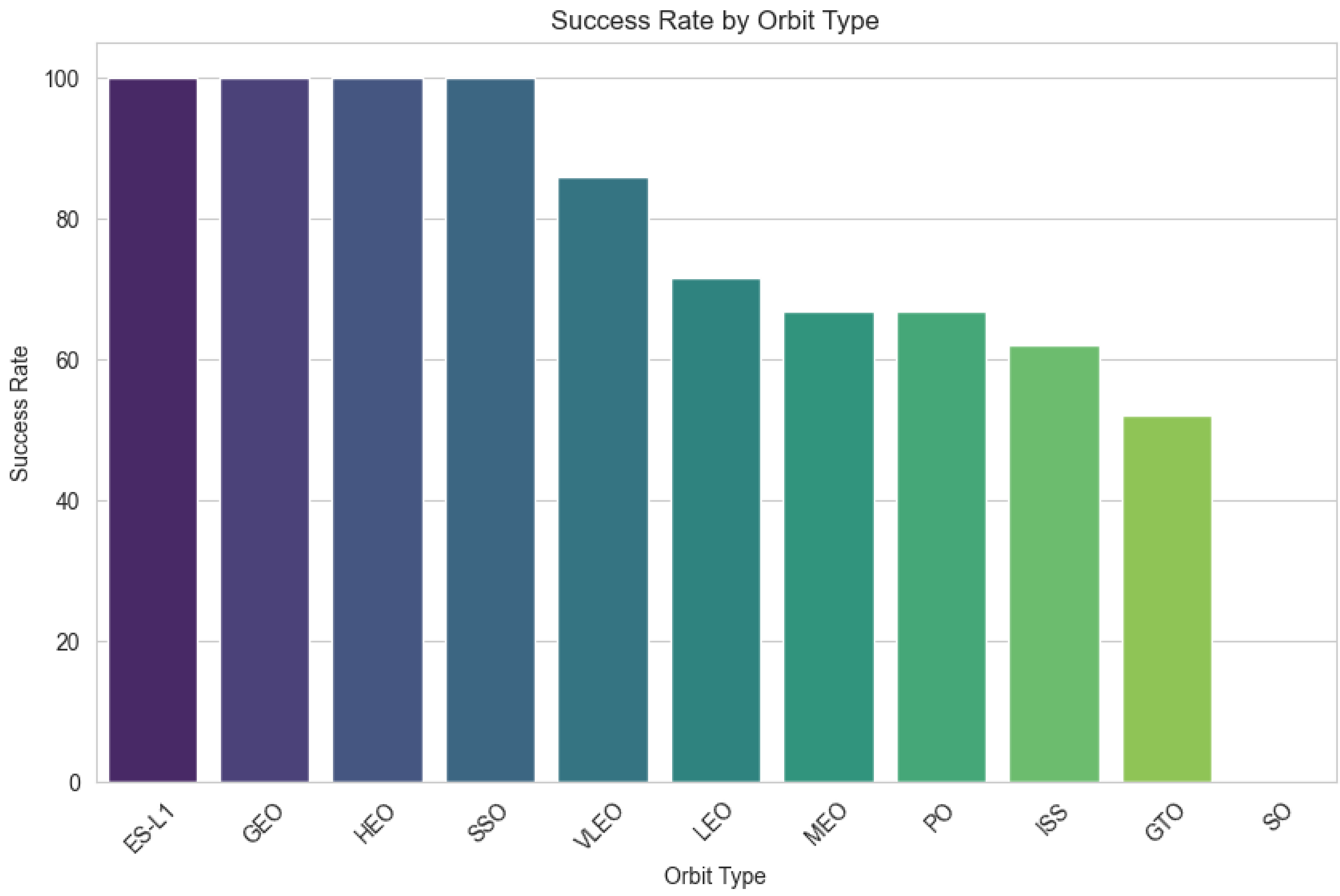
- Payloads exceeding **9,000 kg exhibit an excellent success rate, approximately equivalent to the weight of a school bus.**



- Additionally, payloads **exceeding 12,000 kg seem feasible only at Launch Sites CCAFS SLC 40 and KSC LC39A**. Notably, at the **VAFB-SLC launch site, there are no rockets launched with heavy payload masses** exceeding 10,000 kg.

Success Rate vs. Orbit Type

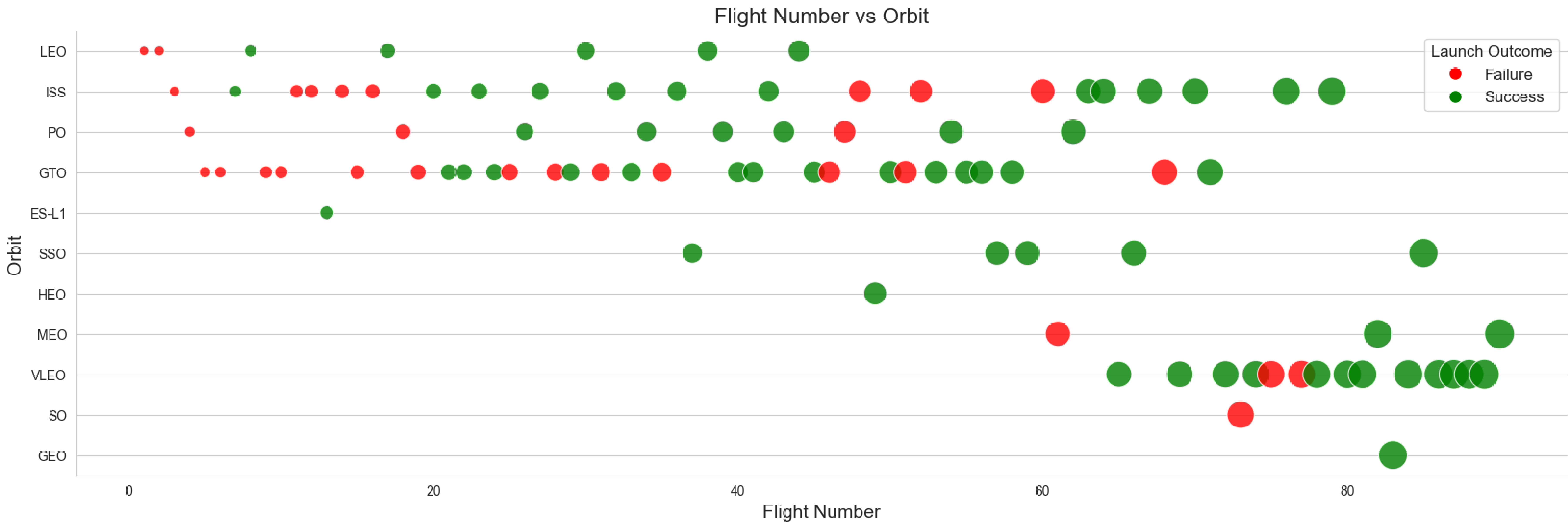
- The best success rates are observed for orbits **ES-L1, GEO, HEO, and SSO**.
- Despite the majority of attempts**, GTO (27 attempts**) exhibits approximately **a 50% success rate.**
- **Orbit SO** has **no recorded success rate.**



Flight Number vs. Orbit Type

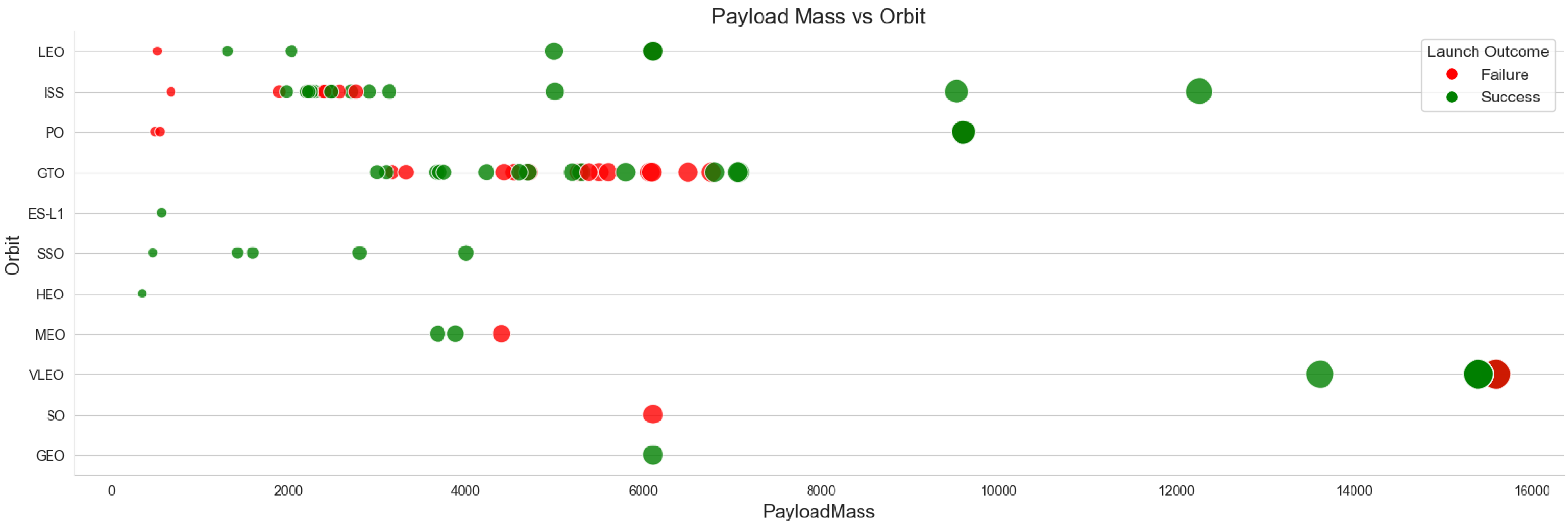
Overall, success rates have exhibited improvement over time with the increase in flight numbers across all orbits.

- Success in **LEO and MEO orbits appears to correlate with the number of flights.**
- Conversely, there is **no apparent relationship between flight number and success in GTO, ISS, and PO orbits.**



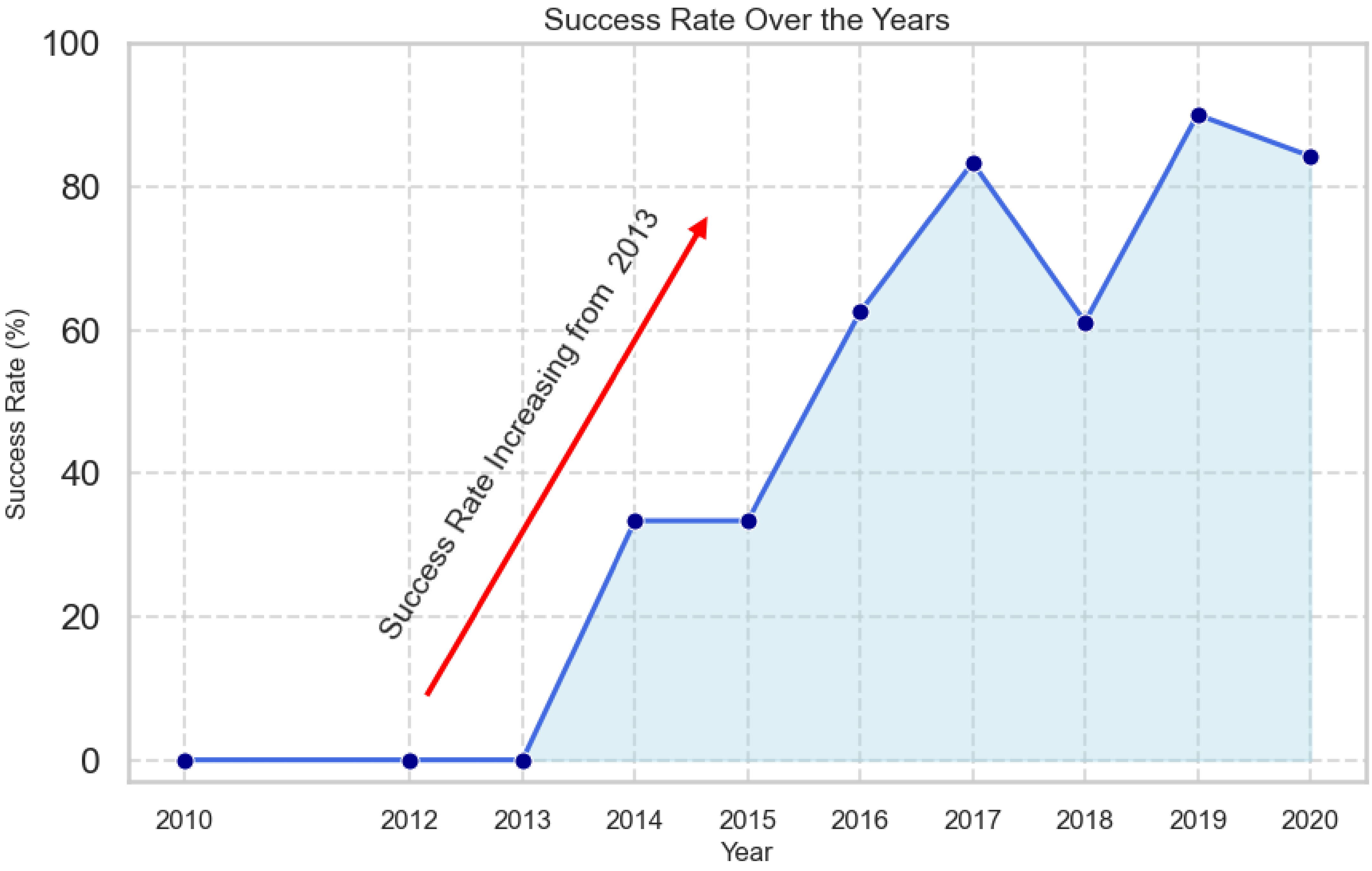
Payload vs. Orbit Type

- **Successful landing rates for heavy payloads are higher for Polar, LEO, and ISS orbits**.
- There appears to be no clear relationship between payload and success rate for GTO and VLEO orbit.
- ISS orbit demonstrates the widest range of payloads and maintains a good success rate.
- There are few launches to the orbits SO and GEO.



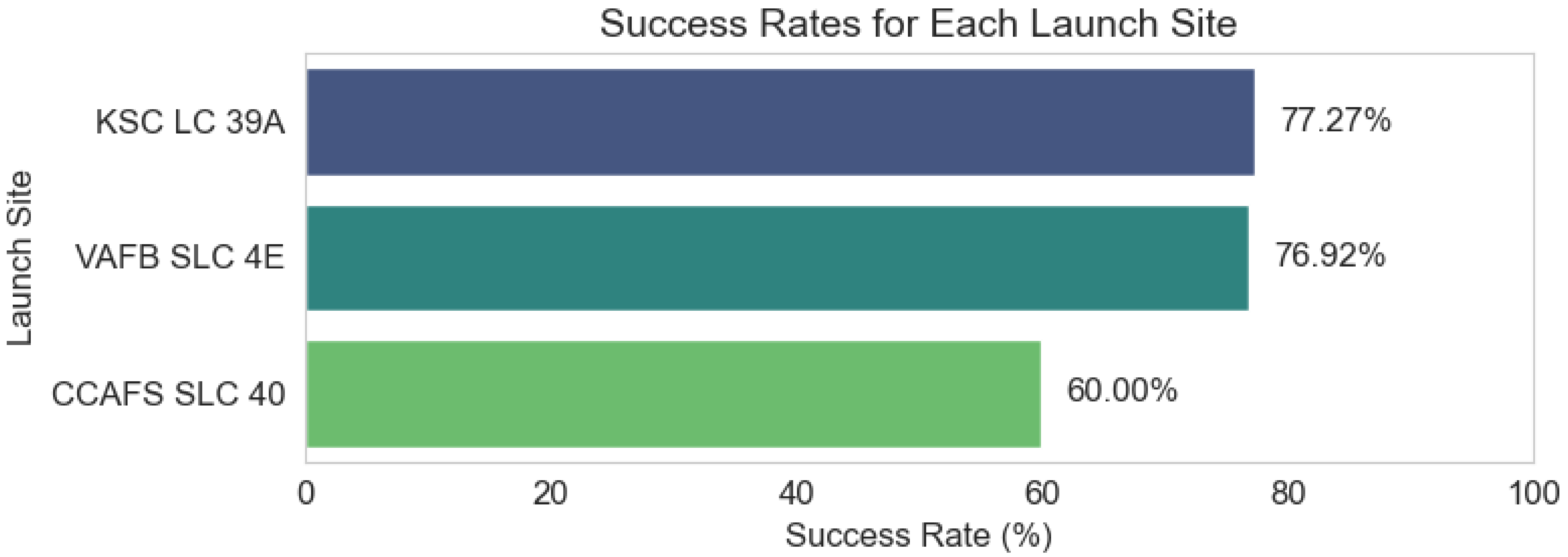
Launch Success Yearly Trend

- The success rate has been steadily **increasing since 2013**, with stability observed in 2014 and a continuous rise thereafter.
- The **initial three years (2010 -2013)** may have represented a **period of technological improvement,** as evidenced by the **absence of success rates during that time**.
- **Success rates in recent years have reached around 80%.**
- These **trends indicate a positive trajectory in launch success rates**, reflecting advancements in technology and operational practices over time.



Success Rate for Each Launch Site

KSC LC 39A exhibits the highest success rate among all launch sites.



EDA with SQL

Exploratory data analysis with Structured Query Language (SQL)

All Launch Site Names

Query and output:

Names of the unique launch sites in the space mission.

```
%sql select DISTINCT "Launch_Site" from SPACEXTBL
```

✓ 0.0s

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

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

Note: The dataset provided for the data wrangling process [1] only contains three unique launch site values: CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E. However, here we also have CCAFS LC-40. It's possible that there is missing data in the previous dataset provided for this specific purpose, Upon further investigation, it was found that CCAFS LC-40 is the old name of CCAFS SLC-4 [2] . Therefore, we observe the same four launch sites for further processing.

Query Explanation

Using the keyword **DISTINCT** in the query indicates that it will only display **Unique values** in the **Launch_Site** column from **SPACEXTABL**

Launch Site Names Begin with 'CCA'

Query and output:

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

```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5;
```

✓ 0.0s Python

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

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

Query Explanation

Using the keyword **TOP 5** in the query indicates that only 5 records will be displayed from the **SPACEXTABL.** The "**LIKE**" keyword with the phrase **'CCA%**' includes a wildcard **(%)** at the end, suggesting that the **Launch_Site** name must start with "**CCA"**. The wildcard (%) allows for any characters to follow '**CCA**', meaning that it matches any **Launch_Site** name starting with **'CCA**'.

Total Payload Mass

Query and output:

Display 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_") as Total_Payload_Mass from SPACEXTBL where Customer like 'NASA (CRS)';
```

✓ 0.0s

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

Total_Payload_Mass
45596

Query Explanation

The function **SUM** calculates the total sum of values in the **column PAYLOAD_MASS_KG_.** The **WHERE** clause filters the dataset to only include records where the **Customer** is identified as **NASA (CRS**), ensuring that calculations are performed solely on payloads associated with this customer

Average Payload Mass by F9 v1.1

Query and output:

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

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

✓ 0.0s

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

avg(PAYLOAD_MASS__KG_)
2928.4

Query Explanation

The function **AVG** computes the average value in the column **PAYLOAD_MASS__KG_.** The **WHERE** clause filters the dataset to only include records where the **Booster_Version** is identified as **'F9 v1.1**'

First Successful Ground Landing Date

Query and output:

List the date when the first successful landing outcome in ground pad was achieved

```
%sql select min("Date") as First_Sucessful_groundpad_landing_Date from SPACEXTBL where "Landing_Outcome" like 'Success (ground pad)';

✓ 0.0s

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

First_Sucessful_groundpad_landing_Date
2015-12-22

Query Explanation

The query selects the earliest date of successful ground pad landing from the **SPACEXTBL** table and labels it as **First_Successful_groundpad_landing_Date**. It filters the data to include only records where the **Landing_Outcome** is identified as **'Success (ground pad)'.**

Successful Drone Ship Landing with Payload between 4000 and 6000

Query and output:

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

```
%%sql
select Booster_Version
from SPACEXTBL
where "Landing_Outcome" like 'Success (drone ship)'
and "PAYLOAD_MASS__KG_" > 4000
and  "PAYLOAD_MASS__KG_" < 6000
```

0.0s

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

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

Query Explanation

It filters the dataset using the "**AND**" keyword, ensuring that only where the **Landing_Outcome** is identified as **'Success (drone ship)'** and the **PAYLOAD_MASS__KG** is greater than 4000 kg and less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Query and output:

List the total number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

```
%%sql
SELECT 'Failure' AS Outcome_Type, COUNT("Mission_Outcome") AS Outcome_Count
FROM SPACEXTBL
WHERE "Mission_Outcome" LIKE 'Failure%'

UNION ALL

SELECT 'Success' AS Outcome_Type, COUNT("Mission_Outcome") AS Outcome_Count
FROM SPACEXTBL
WHERE "Mission_Outcome" LIKE 'Success%';
```

✓ 0.0s

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

Outcome_Type	Outcome_Count
Failure	1
Success	100

Query Explanation

This query combines two SELECT statements using the **UNION ALL** operator to retrieve the **count** of mission outcomes categorized as 'Success' and 'Failure' from the **SPACEXTBL** table. In the first **SELECT** statement, records with a **Mission_Outcome** starting with **'Failure'** are counted and labeled as **'Failure'** in **the Outcome_Type** column. Similarly, the second **SELECT** statement counts records with a **Mission_Outcome** starting with **'Success'** and labels them as **'Success'** in the **Outcome_Type** column. This query provides a concise summary of mission outcomes, distinguishing between successes and failures in the space missions recorded in the dataset.

Boosters Carried Maximum Payload

Query and output:

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

```
%%sql
select
"Booster_Version", "PAYLOAD_MASS__KG_"
from SPACEXTBL
where
"PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTBL )
```

✓ 0.0s

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

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

This query retrieves the **booster_version** and payload mass from the **SPACEXTBL** table. It filters the records based on the condition where the payload mass matches the maximum value in the dataset. The maximum value is determined using the **MAX()** function, implemented within a subquery in the **WHERE** clause.

2015 Launch Records

Query and output:

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.

```
%%sql
select
substr("Date", 6, 2) as month,
"Landing_Outcome",
"Booster_Version",
"Launch_Site"
from
SPACEXTBL
where
substr("Date", 0, 5) = '2015'
and "Landing_Outcome" like 'Failure (drone ship)';
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

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

Query Explanation

It filters the dataset to include records where the year portion of the Date column is '2015' using the **SUBSTR** function to extract the year. Additionally, it filters the dataset to include records where the **Landing_Outcome** column starts with '**Failure (drone ship)**'. The **SUBSTR** function is used to extract the month portion of the **Date** column.

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

Query and output:

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("Landing_Outcome") as outcome_count
from SPACEXTBL
WHERE "Date" between '2010-06-04' and '2017-03-20'
group by "Landing_Outcome"
order by "outcome_count" DESC;
```

✓ 0.0s

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

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

Query Explanation

This query selects the **Landing_Outcome** column and **counts** the occurrences of each landing outcome in the **SPACEXTBL** table. It filters the dataset to include records using **where** clause for the **Date** column falls between '2010-06-04' and '2017-03-20'. The results are grouped using **GROUP BY** the **Landing_Outcome** column to aggregate the counts for each unique landing outcome. The outcome counts are then ordered in descending order using the **ORDER BY** clause.

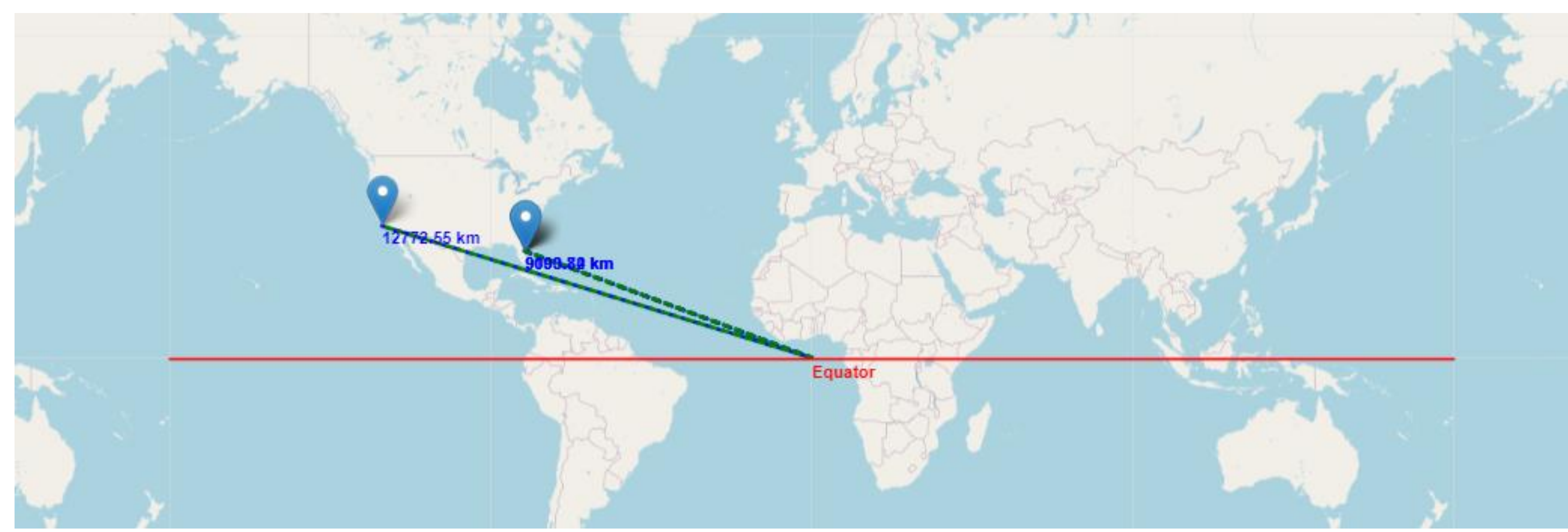
Section 4

Launch Sites Proximities Analysis

Mark All Launch Sites on a Map

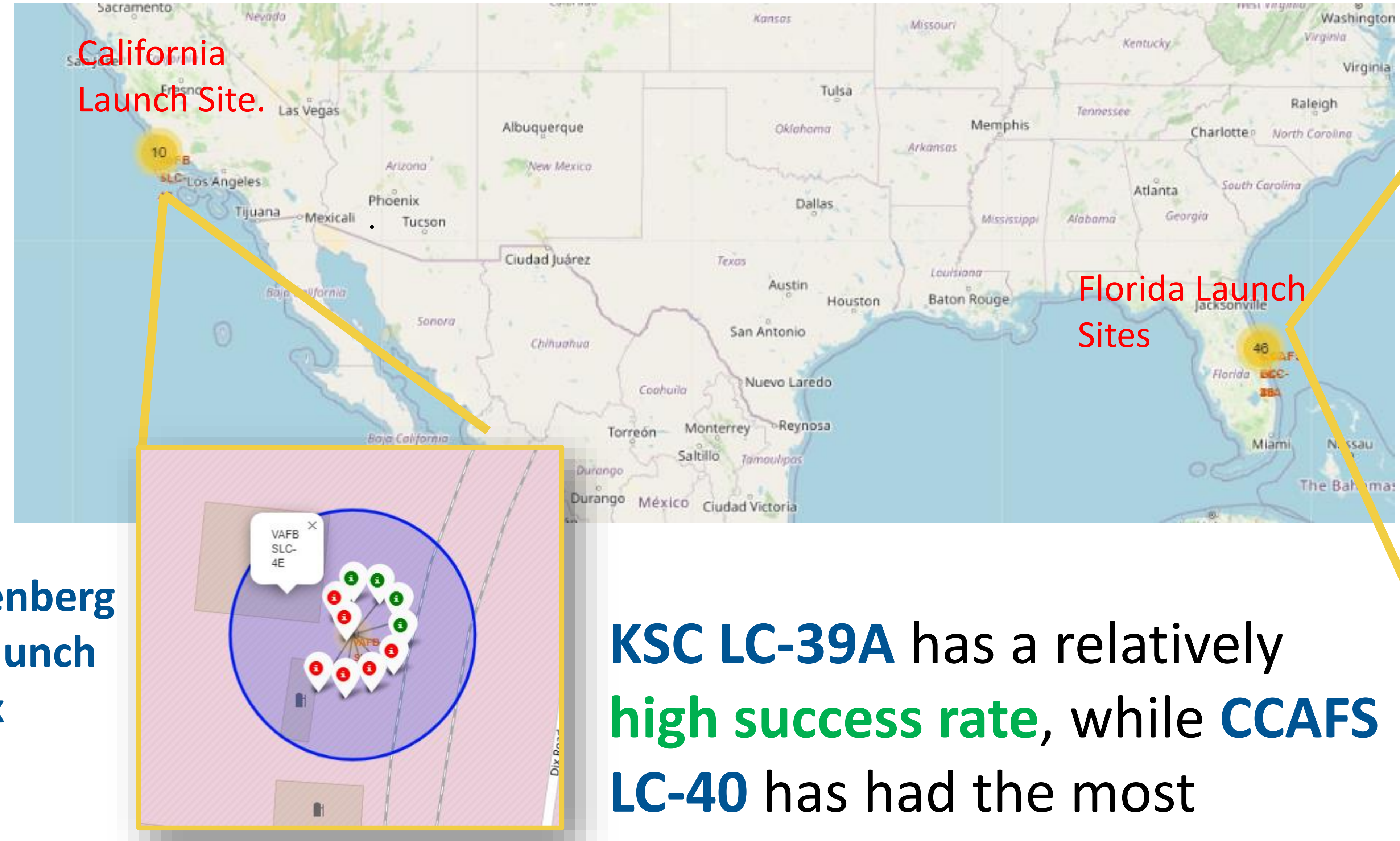
Three sites are close to each other near the **Cost of Florida.**

We can see that **all four SpaceX launch sites** are located in the **United States of America**, **near the coasts of California and Florida.**



All launch sites are not in the proximity of the equator line as we can see here.

Success/Failed Launches for Each Site



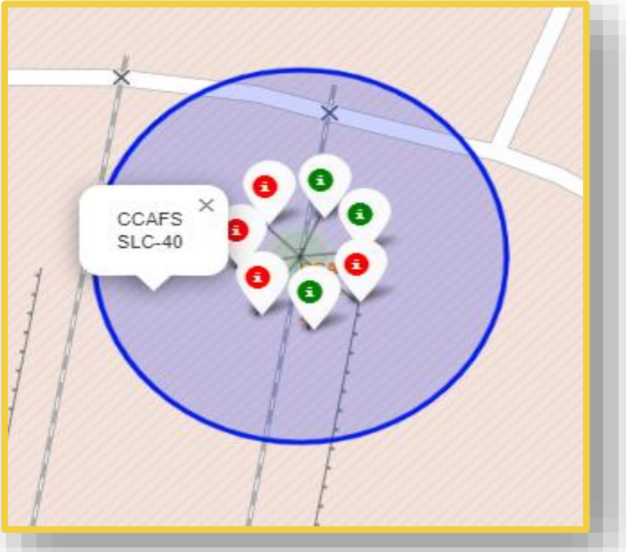
1. Vandenberg Space Launch Complex

VAFB SLC-4E
Total Successful Launch: 4
Total Failed Launch: 6

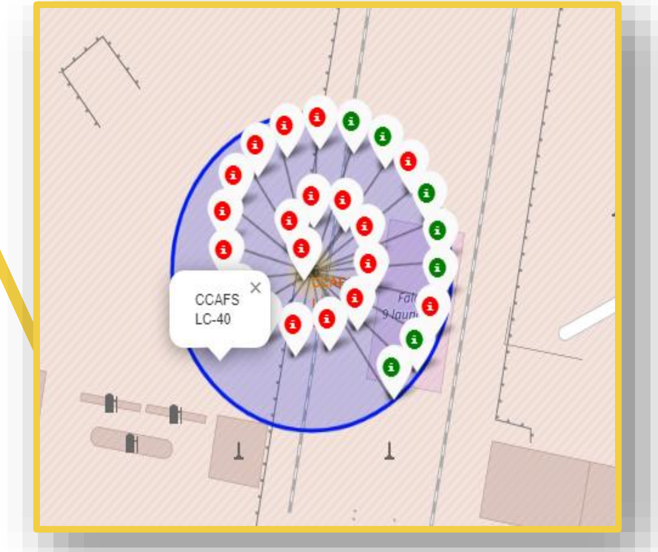
2. Kennedy Space Center Launch Complex

KSC LC-39A
Total Successful Launch: 10
Total Failed Launch: 3

3. Cape Canaveral Space Launch Complex



CCAFS SLC-40
Total Successful Launch: 3
Total Failed Launch: 4

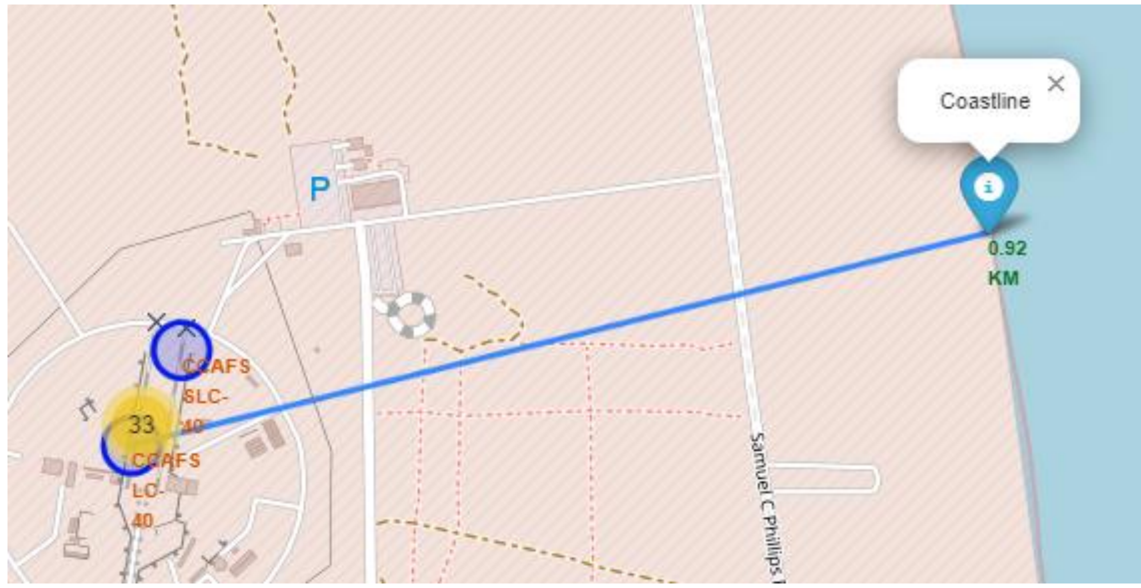


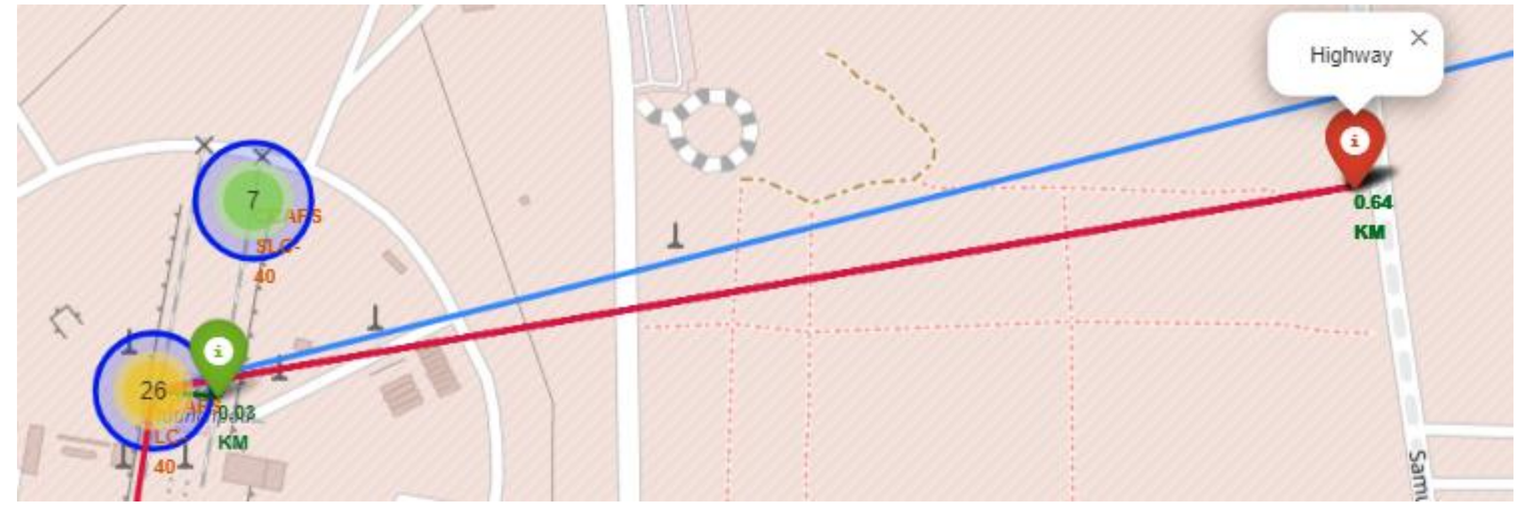
CCAFS LC-40
Total Successful Launch: 7
Total Failed Launch: 19

KSC LC-39A has a relatively **high success rate**, while **CCAFS LC-40** has had the most attempts and **the highest failure rate.**

Total launches of each site are represented in a cluster, while **green** marking successful launches and **red** indicating unsuccessful ones.

Distances between a launch site to its proximities





Launch sites is **very close proximity to Railway line is around 0.03 Km**

Launch sites is **close proximity to Highway around 0.63 Km.**

Taking **CCAFS LC-40** as reference, it can be observed that
Launch sites in **close proximity to coastline around 0.92 Km**

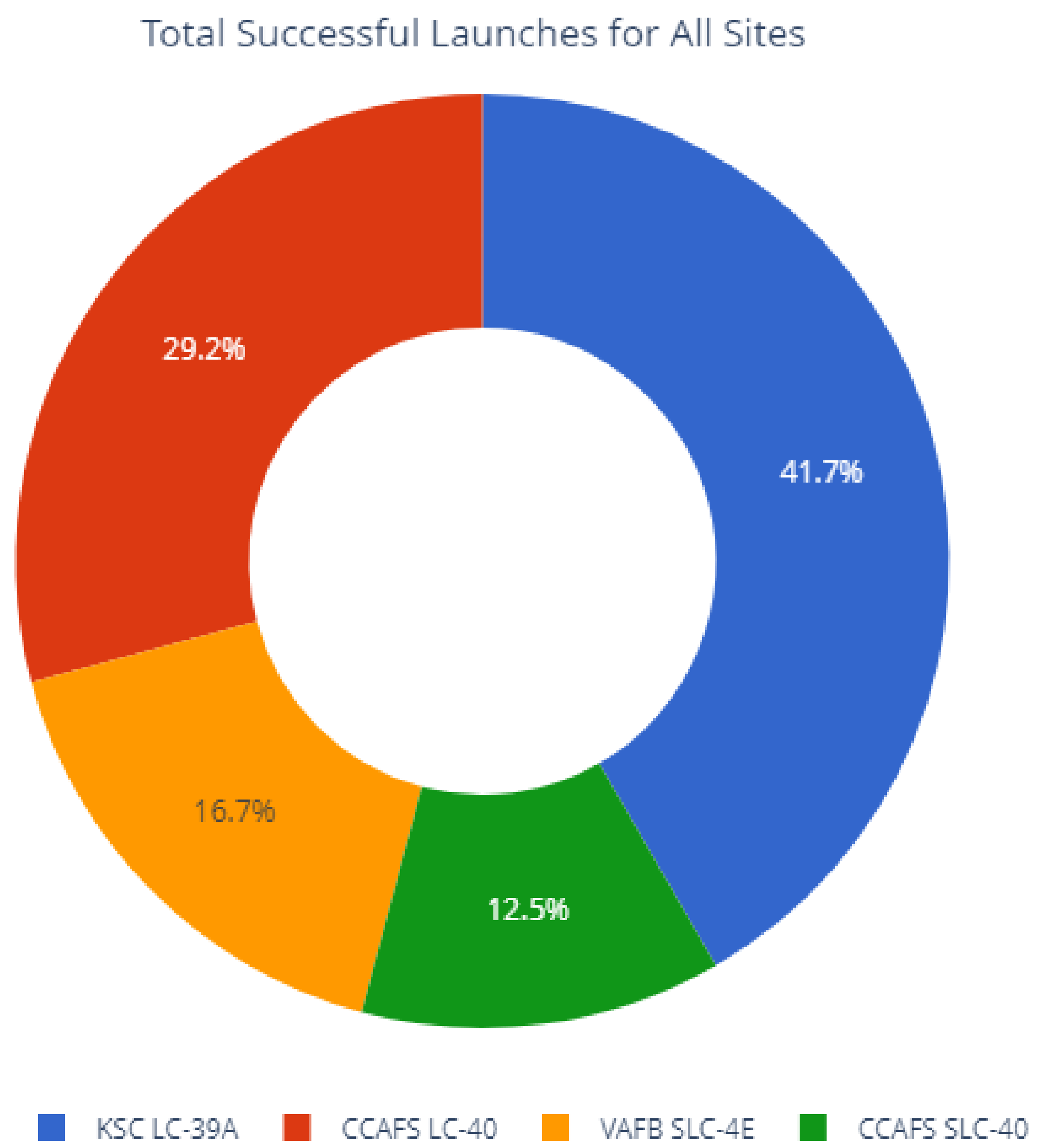
Launch sites is **quite far from the nearest city i.e., Cape Canaveral around 17.27 Km**.

- Observations suggest that **launch sites like CCAFS LC-40 are strategically located near railways and highways for efficient transportation of goods.**
- Moreover, **their proximity to coastlines minimizes risks** by directing launch failures towards the sea, away from densely populated areas.
- These strategic choices prioritize operational efficiency and safety measures in space launch activities.

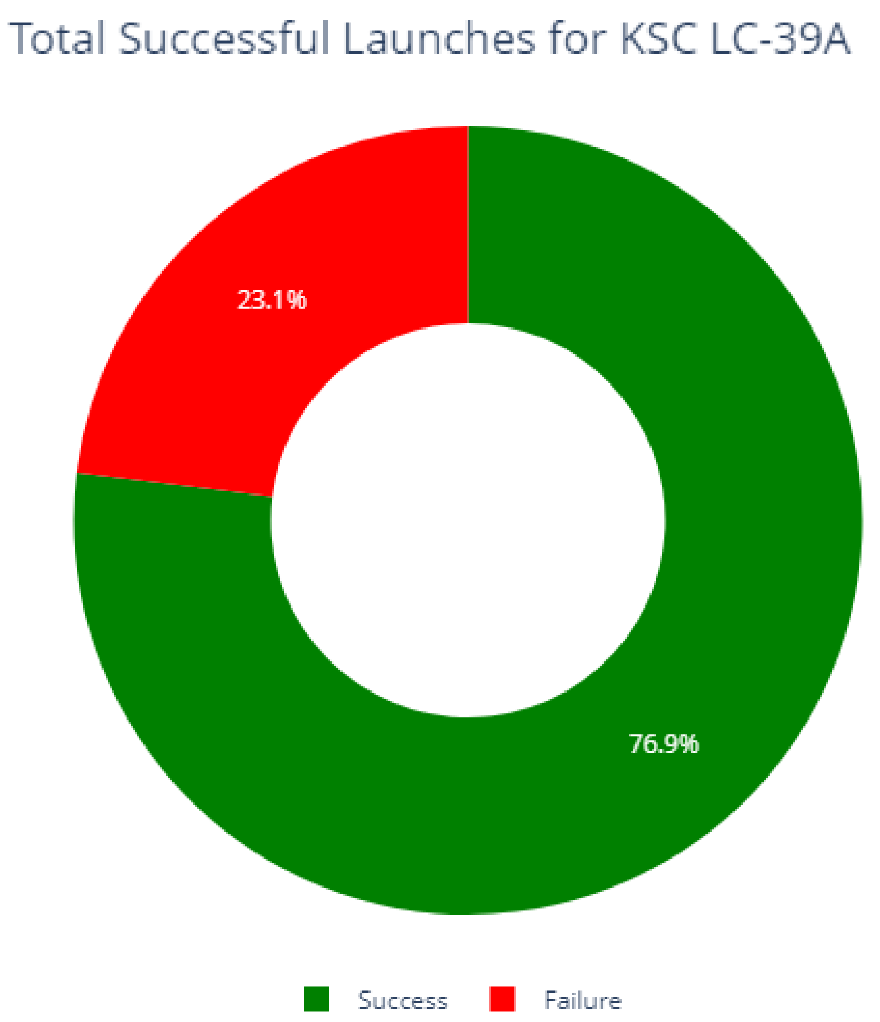
Section 5

Build a Dashboard with Plotly Dash

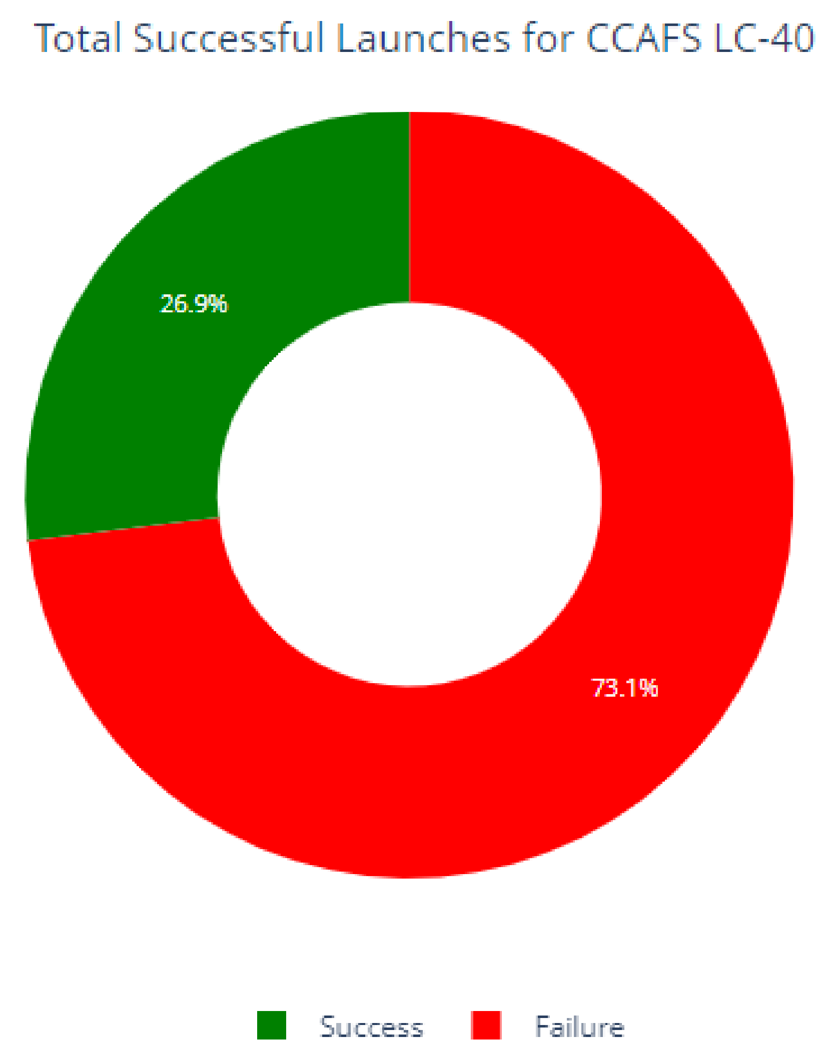
Successful Launches Across Launch Sites



We can observe that **KSC LC-39A** had **the most successful** launches among all the sites, while **CCAFS LC-40 had fewer successful launches** in comparison with other sites.



KSC LC-39A achieved a **success rate of 76.9%,** the highest among all sites, with a **failure rate of 23.1%.**

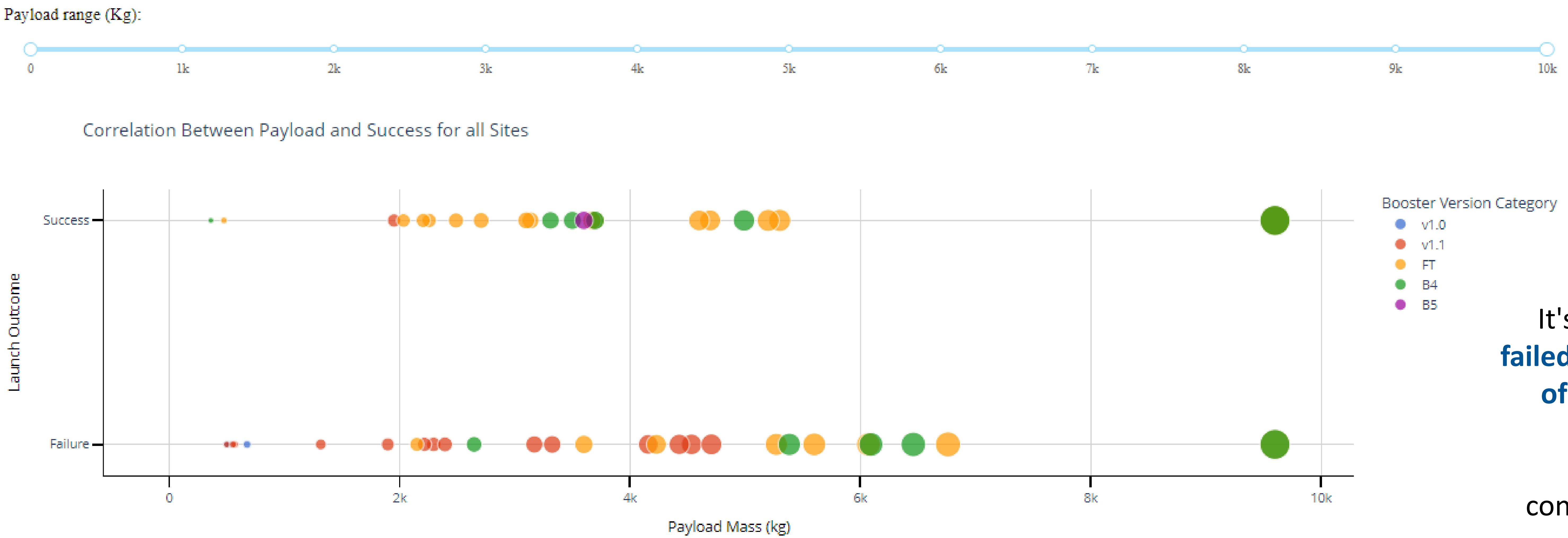


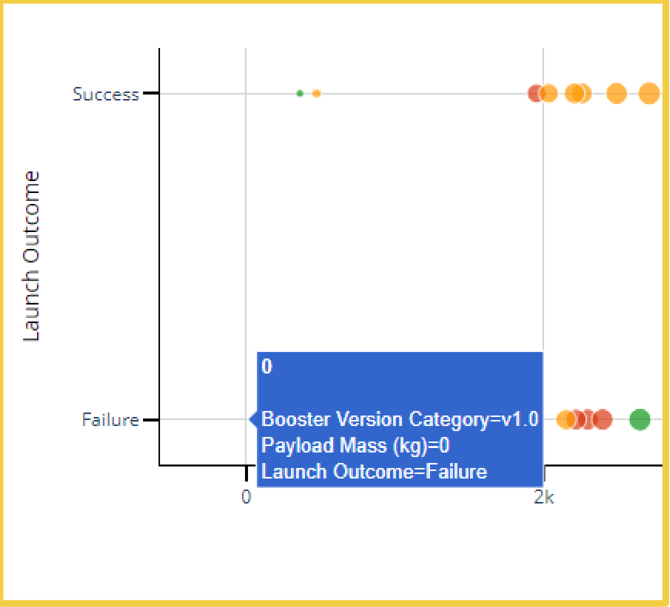
We observe that **CCAFS LC-40** experienced the **highest number of launch failures** among all the sites.

Payload Mass vs. Launch Outcome vs. Booster Version Category

Negative correlation between payload weight and success rate

We observe that the **success rates for low-weighted payloads are higher than those for heavy-weighted payloads**, indicating that as **payload weight increases**, **success rates tend to Decrease**.





It's noteworthy that there was a **failed landing even with a payload of 0**, indicating potential factors beyond payload mass such as technical issues, weather conditions, or operational errors.

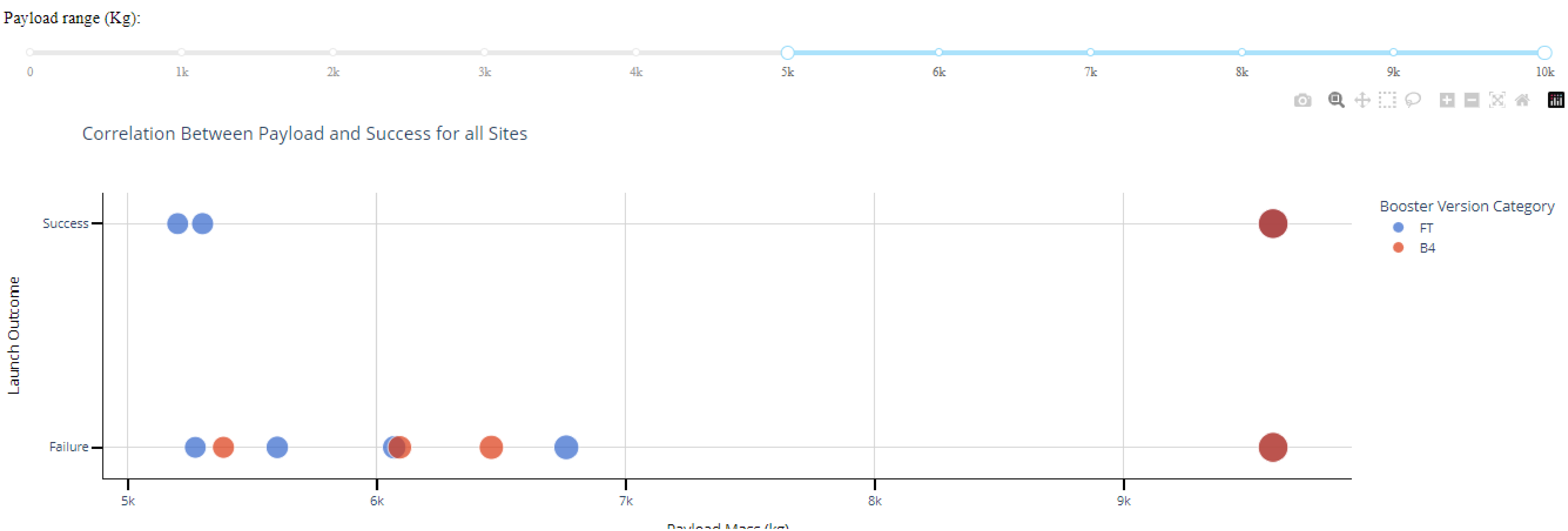
The **B4 booster version** stands out as the **only one** capable of **achieving success with payloads as low as 362 Kg**, while also **boasting the highest capacity of 9600 Kg**, distinguishing it from other variants.

It has been observed that the **FT booster** demonstrates the **best performance**, characterized by the maximum number of **successful attempts**. Conversely, **the v1.1 booster is identified as the worst performing** among all variants.

Payload vs. Launch Outcome – Range (5K – 10K)

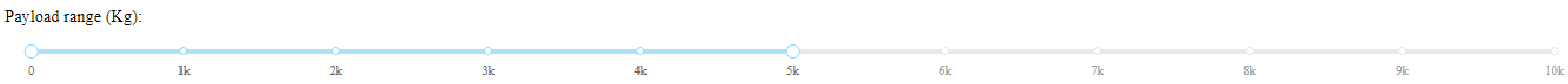
Positive correlation between the payload mass within this range and the launch outcome.

The **B4 and FT** versions are the only ones capable **of carrying payloads exceeding 5,000Kg**, despite the majority of attempts ending unsuccessfully.



Payload vs. Launch Outcome – Range (0K – 5K)

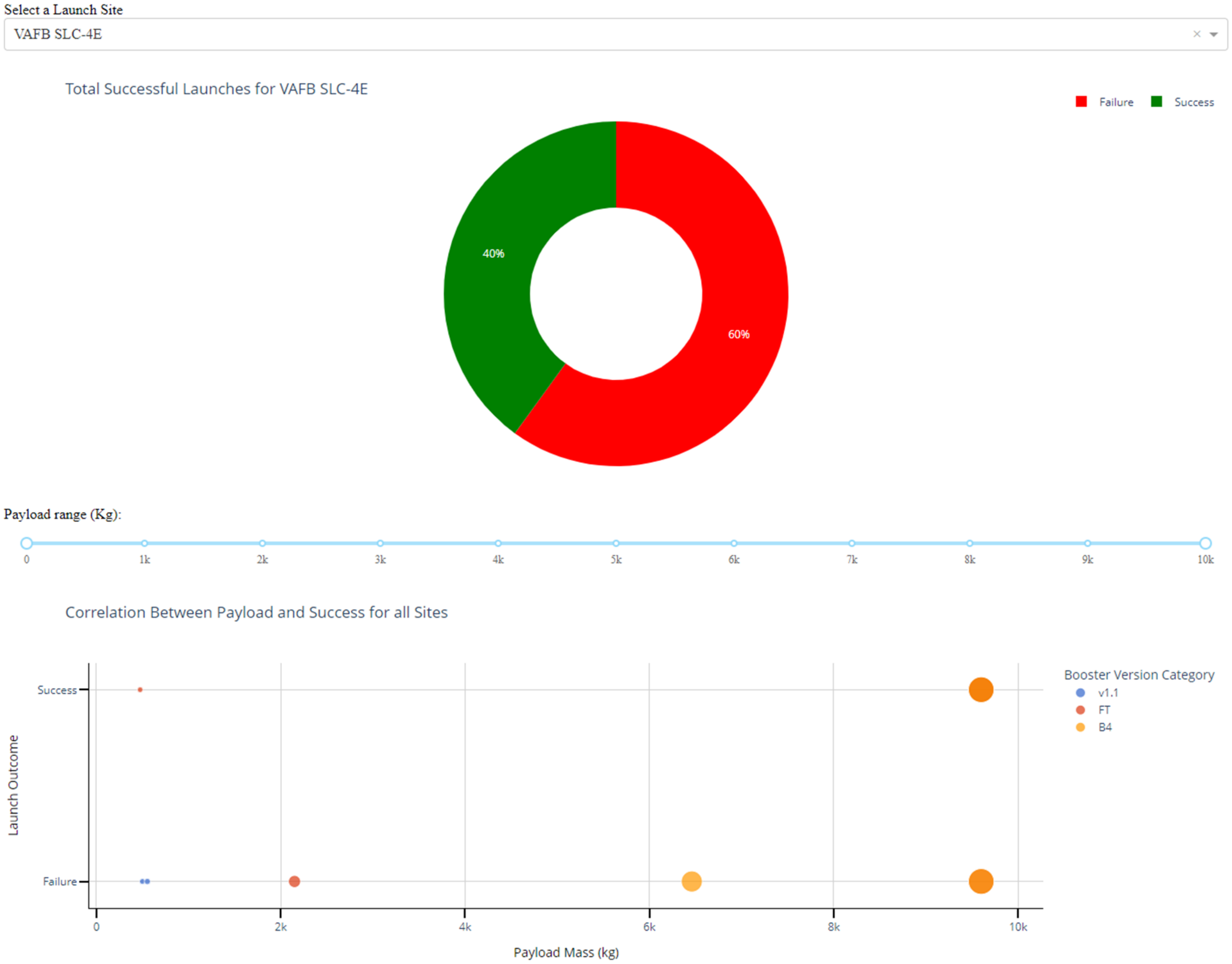
- Overall, it is evident that **most successful attempts involve payloads under 5000 kg**.
- Additionally, **B5 appears to have only one attempt**, which was successful.



Most Successful Launch Site - Payload vs. Launch Outcome

As observed, the most successful site, **KSC-LC 39A**, primarily witnessed **successful launches with the FT booster version**, **followed** by the **B4 and B5 variants.**

Most Unsuccessful Launch Site - Payload vs. Launch Outcome



As observed, the **most unsuccessful site, CCAFS LC-40**, had unsuccessful launches with **booster versions v1.1, B4, and FT.**

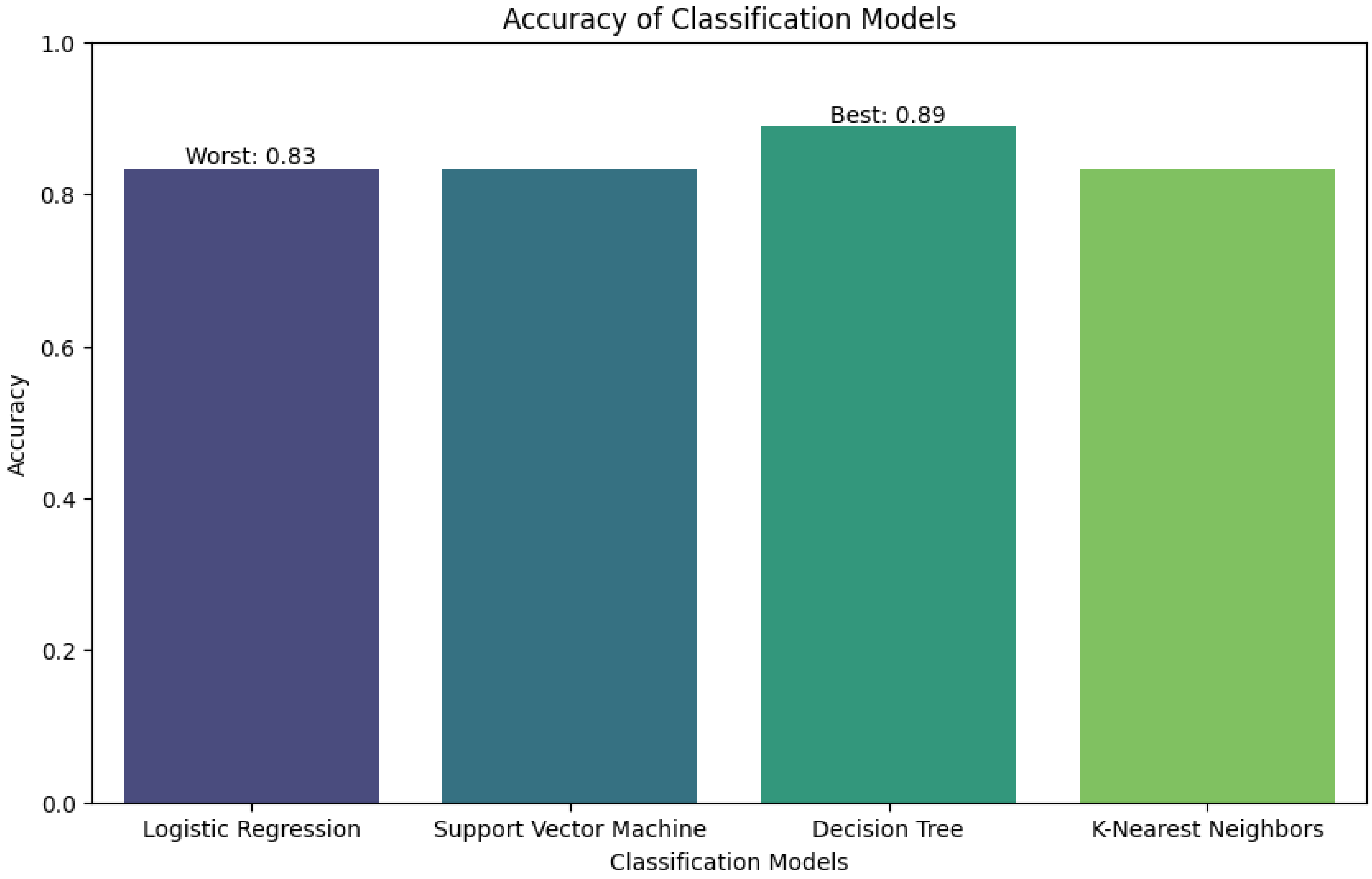
Section 6

Section 5

Predictive Analysis (Classification)

Accuracy For All Built Classification Models

The Decision Tree Model is considered the best classification model for this task.



The **Decision Tree Classifier** achieved the **highest accuracy among all models,** with a **score of 89%.**
After tuning its **hyperparameters**, the model achieved even better performance.
The best parameters for the decision tree model are as follows:
criterion='entropy’,
max_depth = 8,
max_features='sqrt',
min_samples_leaf=4,
min_samples_split=5, and
splitter='random'.

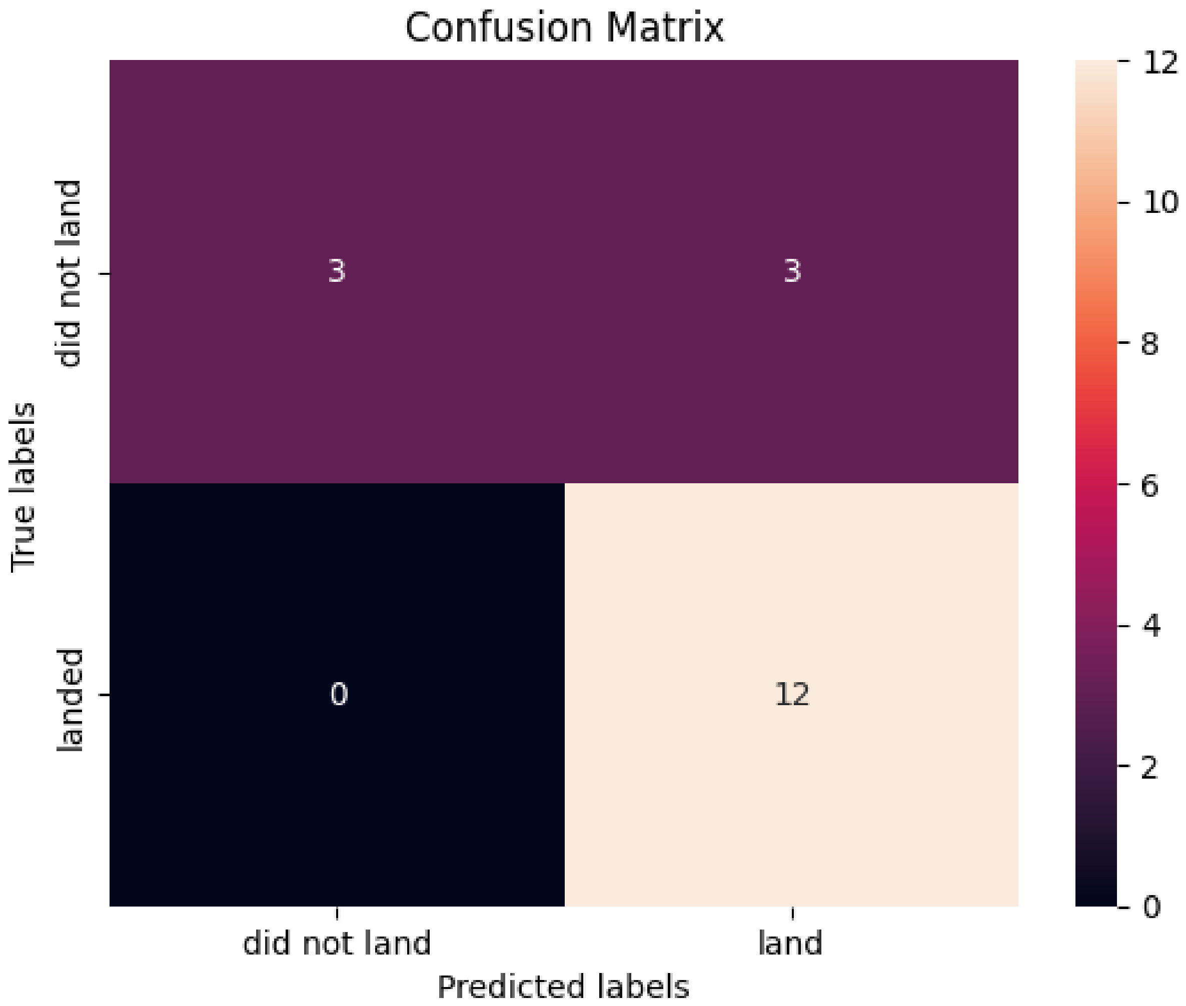
Comparison of Test and Train Accuracies

It's notable that only the **Decision Tree model exhibits an improvement in test accuracy compared to its train accuracy,** while all other algorithms show the opposite trend.

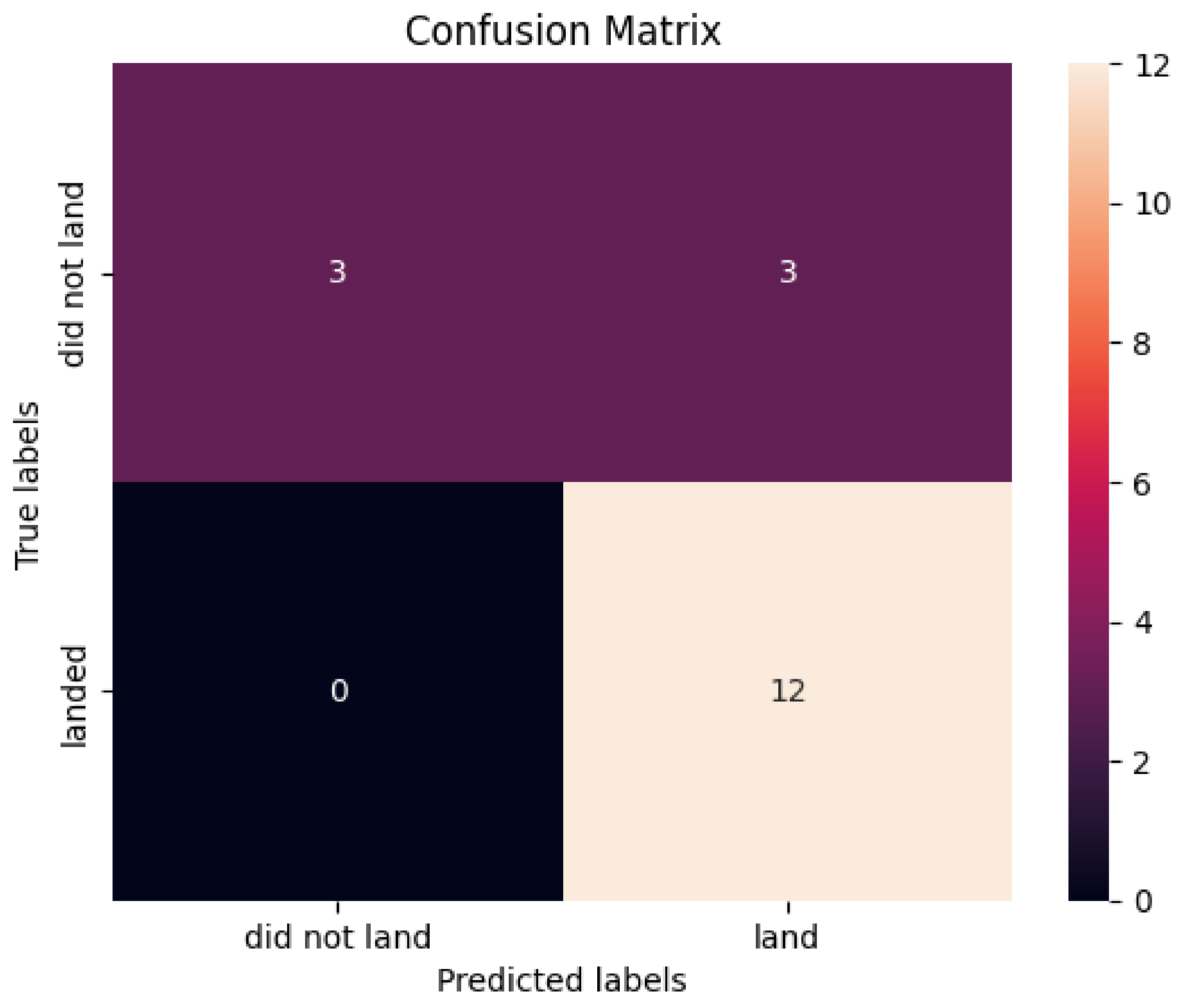
Confusion Matrix

The logistic **Regression Model And Support Vector Machine (SVM)** both achieved **identical results** on the confusion matrix, indicating consistent performance in accurately identifying both positive and negative cases. However, the presence of a few false positive predictions suggests a potential area for enhancement in precision.

Logistic Regression
Train Accuracy: 0.8464285
Test Accuracy:0.833333S



SVM
Train Accuracy: 0.8482142•
Test Accuracy:0.833333



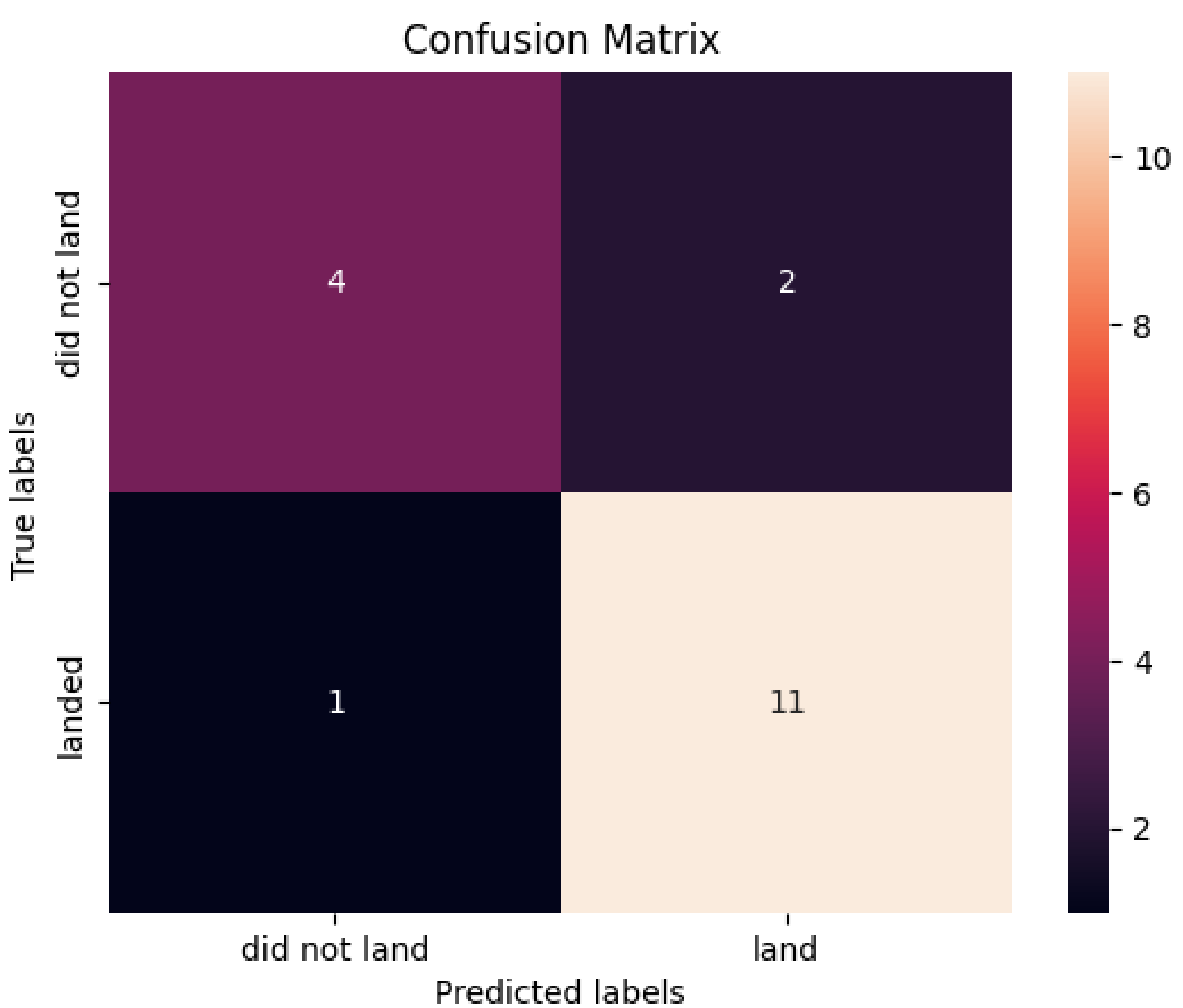
Confusion Matrix

The **decision tree classifier showed a distinct confusion matrix**, featuring fewer true positives and a slightly elevated false negative count. Despite these deviations, the decision tree model maintained acceptable performance levels.

Decision Tree

Train Accuracy: 0.8767857142857143

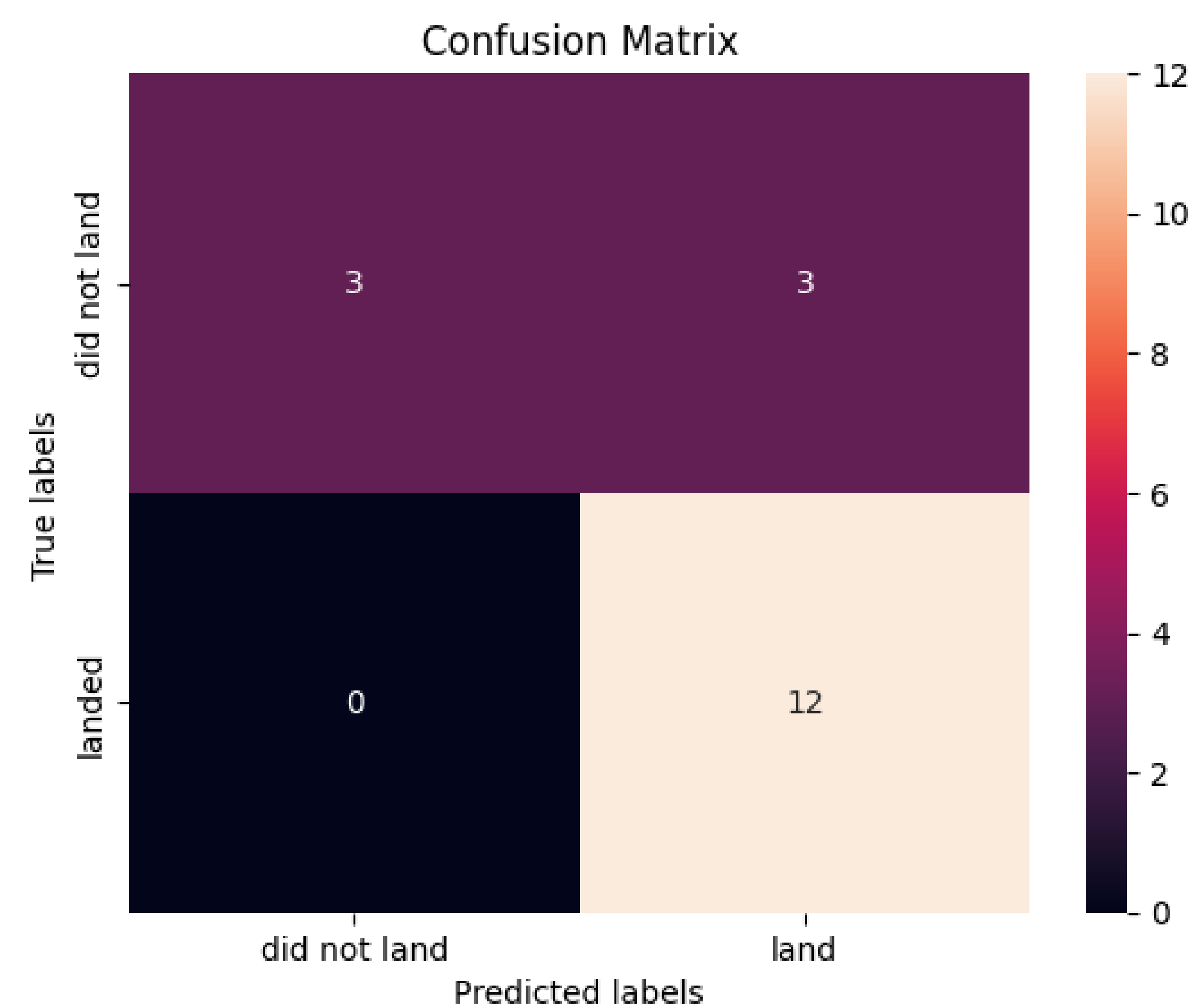
Test Accuracy: 0.8333333333333334



KNN

Train Accuracy: 0.8482142

Test Accuracy:0.833333



Conclusion

In conclusion, addressing the problem statement of our study:

1. **Factors determining rocket landing success:**
   - Various factors such as **Booster Version, Orbits, Payload Weight, Launch Site, Flight Volume, and Landing Type** contribute to the success of a rocket landing.
2. **Contribution of different features to success rate:**
   - **Orbits Type:** Orbits ES-L1, GEO, HEO, SSO, and VLEO exhibit the highest success rates.
   - **Payload Weight:** Lower-weighted payloads perform better than heavier ones, with a negative correlation observed.
   - **Launch Site Sites:** Launch Site Sites with larger flight volumes tend to have higher success rates.
   - **Flight Number:** Higher flight volumes correlate with increased success rates at launch sites.
   - **Payload Mass:** Positive correlation noted within certain mass ranges and launch outcomes.
   - **Landing Type:** Drone Ship (ASDS) stands out as the most used and successful landing method.
3. **Most effective algorithm for success prediction:**
   - The Decision Tree Classifier emerges as the most effective machine learning algorithm for predicting rocket landing success.

Additional Insights and Observations:

- **Launch Site Success:** KSC LC-39A stands out with the highest number of successful launches, indicating its reliability and operational excellence.
- **Payload and Orbit Types:** Successful landings are more common for heavy payloads in Polar, LEO, and ISS orbits, suggesting optimized mission planning and payload distribution.
- **Booster Version Impact:** The choice of booster version significantly affects launch success, with FT, B4, and B5 variants dominating successful launches, particularly for payloads exceeding 5,000 kg.
- **Strategic Launch Site Location: Sites** like CCAFS LC-40 are strategically located near transportation networks and coastlines, prioritizing operational efficiency and safety by minimizing risks to populated areas.
- **SpaceX Mission Success:** SpaceX demonstrates remarkable mission success rates, with nearly 99% success in achieving mission objectives, highlighting their expertise and reliability in space launch operations
- **Flight Number and technology Advancement:** Monitoring trends across multiple flights provides insights into the learning curve and landing success rates' evolution with technological progress.

Recommendation for SpaceY (1/2)

Considering the objectives of the assignment and aiming to aid Space Y in its competition with SpaceX, several key considerations should be taken into account:

- **Orbit Selection:** Space Y should prioritize launching payloads into orbits with higher success rates, such as Polar, LEO, and ISS orbits. These orbits have demonstrated better success rates for heavy payloads, potentially increasing mission success for Space Y.
- **Launch Site Strategy:** Space Y should carefully select launch sites based on factors like proximity to transportation networks, coastlines, and operational efficiency. Sites similar to KSC LC-39A or CCAFS LC-40 offer strategic advantages in terms of safety, accessibility, and operational excellence.
- **Payload Planning:** Space Y should analyze payload distribution and weight carefully, considering the impact on launch success. By prioritizing payloads within a certain weight range and suitable for specific orbits, Space Y can optimize success rates and operational efficiency.

Recommendation for SpaceY (2/2)

- **Booster Version Selection:** Space Y should assess the performance of different booster versions and prioritize those with higher success rates, similar to successful variants used by SpaceX like FT, B4, and B5. This ensures reliability and consistency in launch operations.
- **Machine Learning Implementation:** Implementing machine learning algorithms like Decision Trees, as demonstrated by SpaceX's success, can significantly aid Space Y in predicting launch outcomes and optimizing operational decisions. By leveraging historical data and predictive modeling, Space Y can make informed decisions to minimize risks and maximize success rates

By considering these factors and leveraging insights gained from SpaceX's experiences, Space Y can strategically position itself for success in the competitive space industry, ultimately achieving its goals of efficient operations and competitive positioning.

Appendix

- **Link to GitHub Repo including completed courses:**
    - Link to Github Repo
- **Link to the deployed Dash application on PythonAnywhere:**
    - Link to - SpaceX Launch Records Dashboard
- **The Folium maps may not render directly on GitHub, so I have provided a link to** GitHub Notebook Viewer **for proper visualization**
    - Link to Launch Site Visualization - Notebook

Thank you!