IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Isaac D
29 July 2025

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

Executive Summary

Overview of Methodologies

- Data Acquisition
- Data Cleaning and Preparation
- Exploratory Data Analysis using Data Visualization Techniques
- Exploratory Data Analysis through SQL Queries
- Creation of Interactive Maps with Folium
- Development of Dashboards using Plotly Dash
- Predictive Modeling (Classification)

Summary of Findings

- Comprehensive EDA and Interactive Analytical Visualizations
- Results from Predictive Modeling

Introduction

Project Introduction: Predicting SpaceX Launch Success

- While SpaceX has increased launch frequency, each mission remains a high-stakes endeavor with significant financial and scientific risks. The ability to accurately predict a launch's outcome is therefore invaluable for risk management.

Questions to be Addressed:

- This project aims to identify the best method and model for predicting Falcon 9 launch success, defined by the payload's deployment, using SpaceX REST API to find key success factors.

Section 1

Methodology

Methodology

Executive Summary

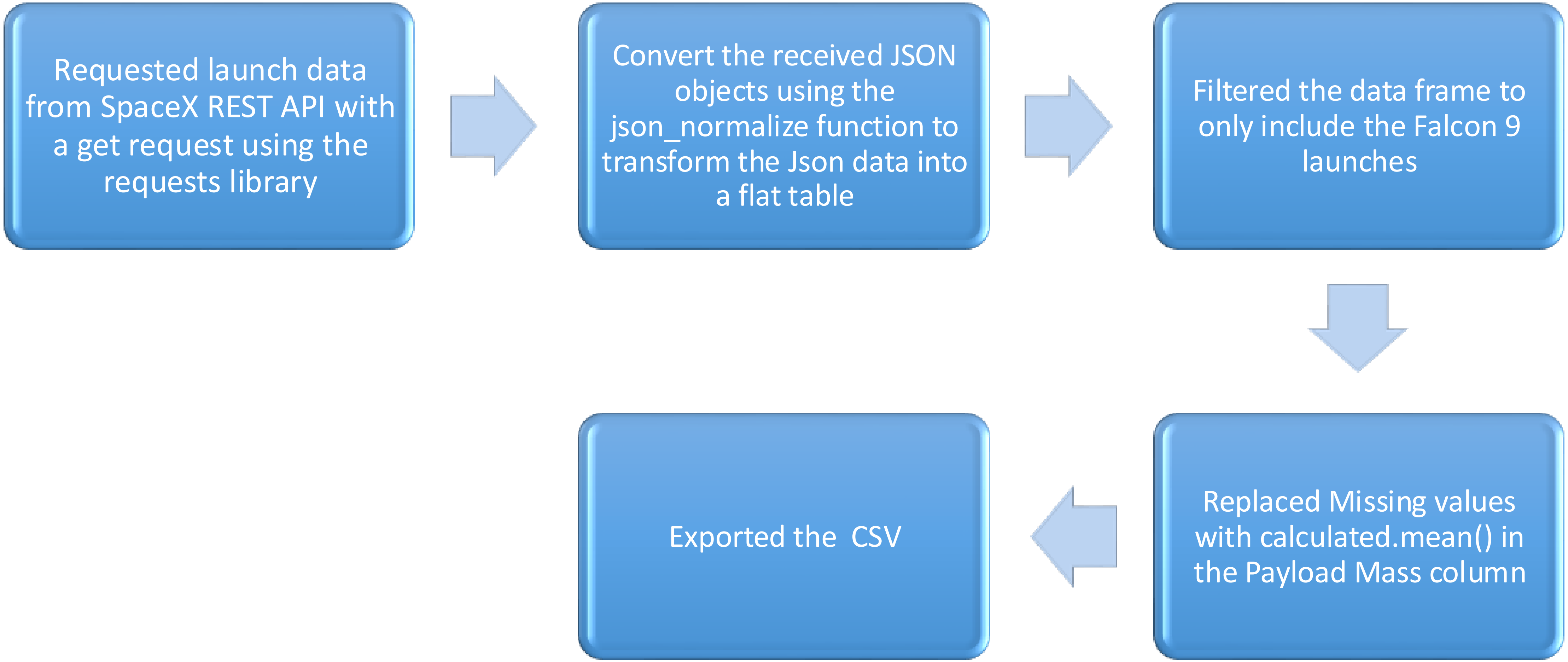
- **Data collection methodology:**
  - Data collection (API) and Web Scraping
- **Perform data wrangling**
  - Describe how data was processed
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- Perform interactive visual analytics using Folium and Plotly Dash
- **Perform predictive analysis using classification models**
- **Build:** Prepare data and train a baseline model on a training set.
- **Tune:** Use cross-validation with Grid/Random Search to find optimal hyperparameters.
- **Evaluate:** Test the final model on an unseen test set using metrics like Accuracy and F1-Score.

Data Collection

- The Data Gathered from the SpaceX REST API : Flight number, Date, Booster Version, Payload mass, orbit, Launch Site, Outcome, Flights, Grid Fins, Legs, Block, Landing Pad, Reused count, Longitude, Latitude

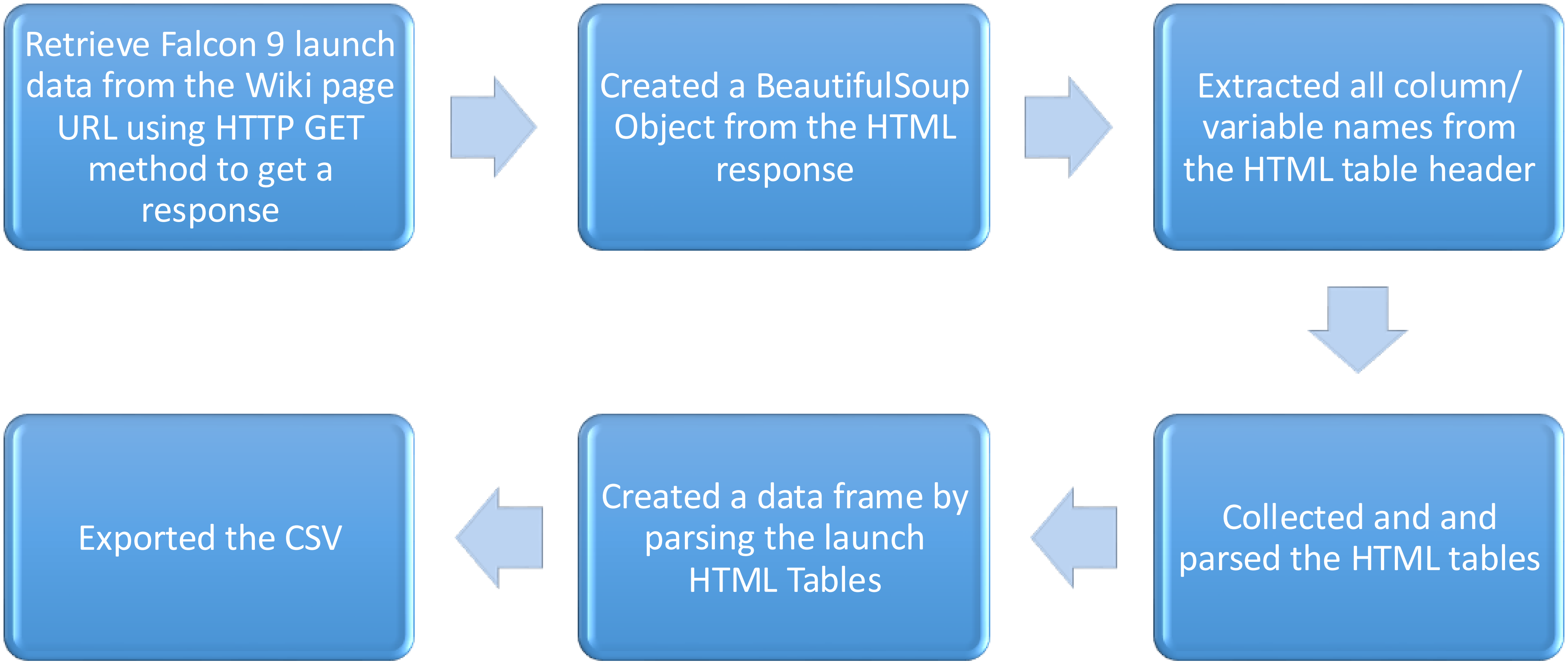
- Web Scraped HTML tables in the SpaceX Wikipedia page using the BeautifulSoup package: Flight Number, Launch site payload mass, payload, launch outcome, orbit, Booster Version, Booster landing, Date, time

Data Collection – SpaceX API



[Data Collection API Git Hub]

Data Collection - Scraping



[Web scraping Falcon 9 Git Hub]

Data Wrangling

Load the SpaceX dataset, determine the percentage of missing values in each attribute, and identify the numerical and categorical columns.

Determine the number of launches per site using value_counts() on the launch site column.

Utilize the .value_counts() method to ascertain the count of landing outcomes and the total number of mission outcomes for the specified orbits.

Develop a landing outcome label from the Outcome column, numerically classifying the Outcome to ascertain the success rate.

Export the CSV

[Data Wrangling Git Hub]()

EDA with Data Visualization

I employed various plots and charts to effectively visualize the data:

- Scatterplot – to examine the relationships between payload mass and flight number, as well as payload mass and launch sites.
- Bar chart – to depict correlations between success rates and orbit types.
- Line chart – to illustrate trends over time by plotting the year against the average success rate.
- Additionally, we reviewed summaries of the pandas data frame to determine the appropriate columns for the x and y axes in our visualizations.

EDA with Data Visualization Git Hub

EDA with SQL

10 SQL queries to obtain data from the SpaceX Data Set were used to display the following:

1. display the names of the unique launch sites in the space mission
2. display the 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 all the booster versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
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.

Build an Interactive Map with Folium

map marker and objects used in building various interactive maps:

- folium.circle - add a highlighted circle area with a text label on a specific coordinate
- folium.marker - pin each launch site location
- Folium marker - mark locations of launch success and failure
- Mouse_position - to get the coordinate (Lat, Long) for a mouse over on the map
- PolyLine – draw a line between launch site and coastline point

Interactive Visual Analytics with Folium Git Hub

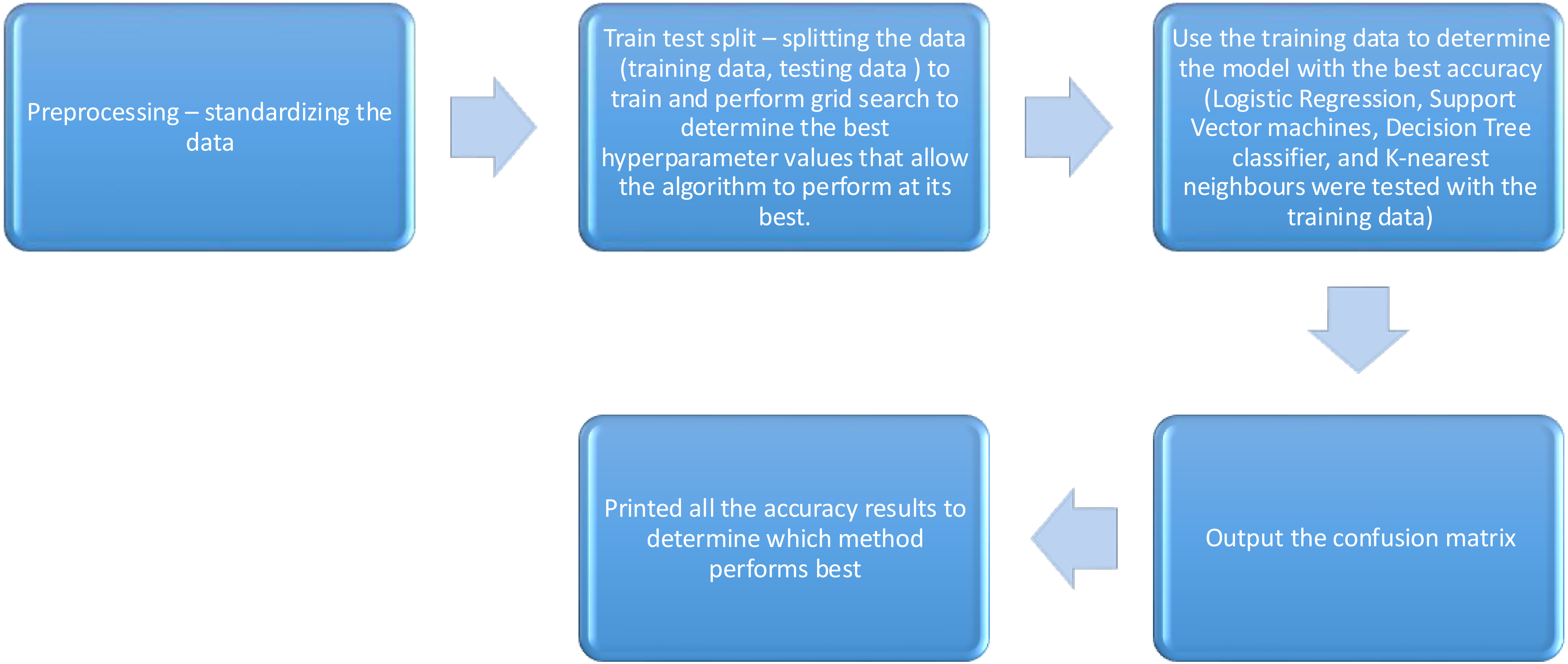
Build a Dashboard with Plotly Dash

I developed a dashboard application to provide an interactive visualization of the data, featuring:

- A success pie chart accompanied by a dropdown menu that allows users to select and view the total number of successful launches at specific sites.
- A success versus payload scatterplot with a payload slider, enabling users to explore the relationship between successful launches and payload weight at chosen launch locations.

Interactive dashboard Git Hub

Predictive Analysis (Classification)



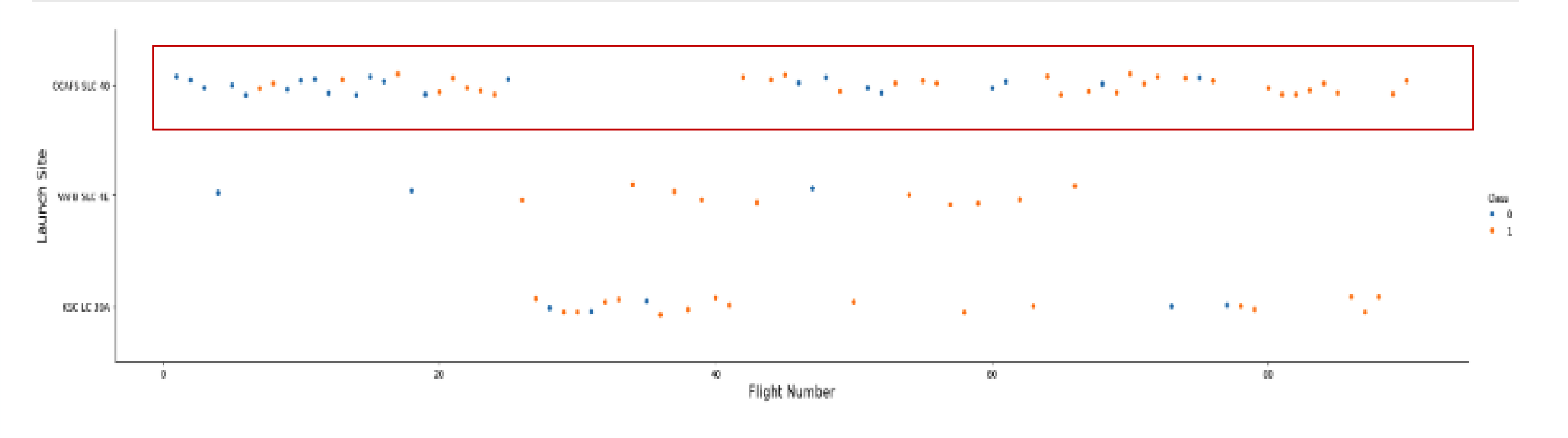
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

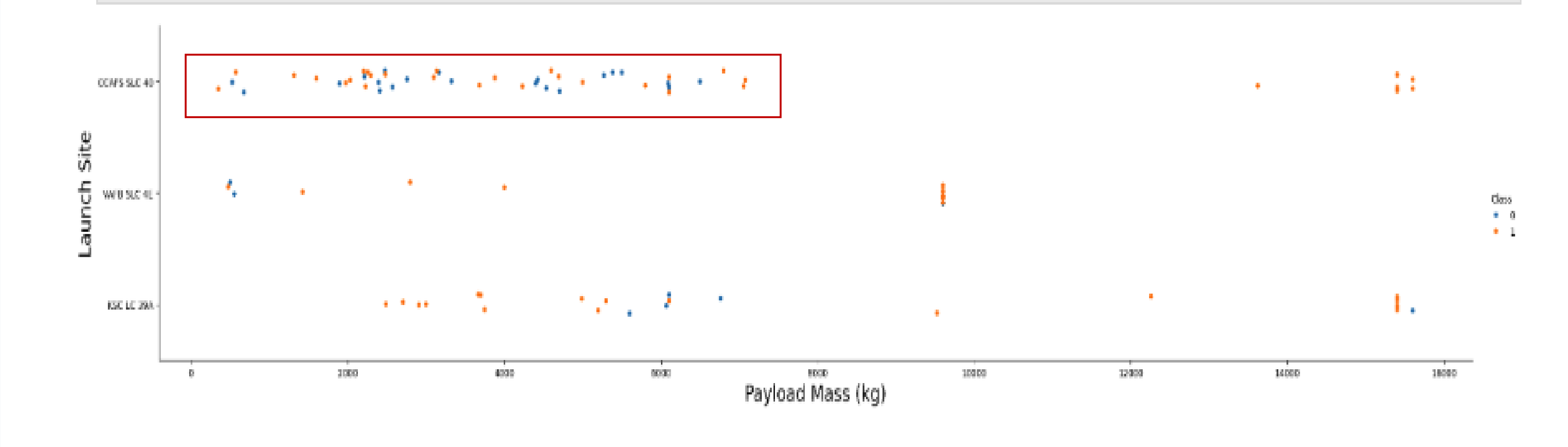
Insights drawn from EDA

Flight Number vs. Launch Site



- This scatterplot displays the relationship between flight counts and launch locations, highlighting a clear pattern in which the CCAFS SLC 40 site hosts a higher volume of multi-class rocket launches. The various rocket classes are distinguished by color coding.

Payload vs. Launch Site



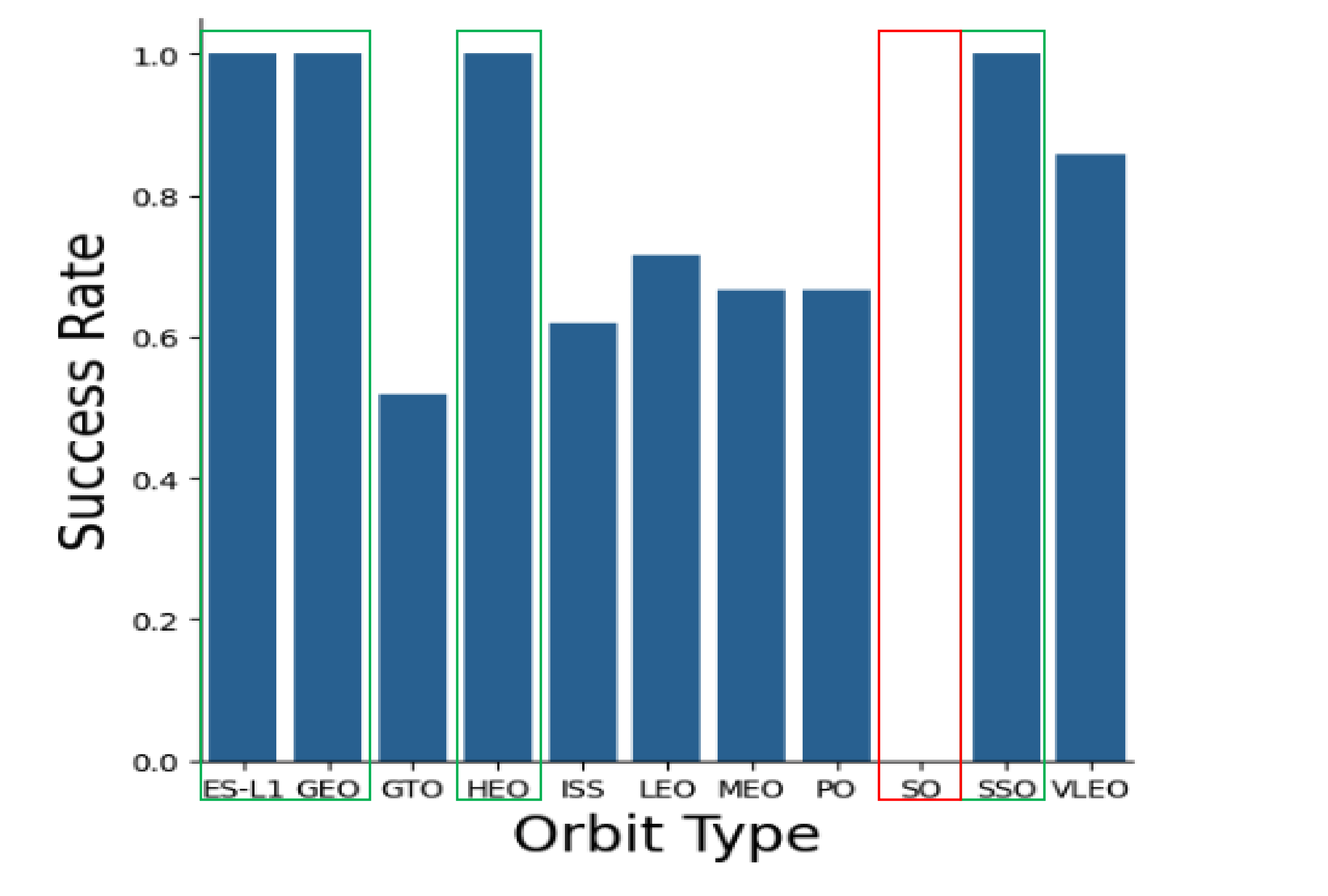
The scatter plot illustrates the relationship between Payload and Launchsite, with the two classes differentiated by color.

- The plot reveals a trend in the data, indicating that most class types fall within the 0 kg to 7500 kg range.

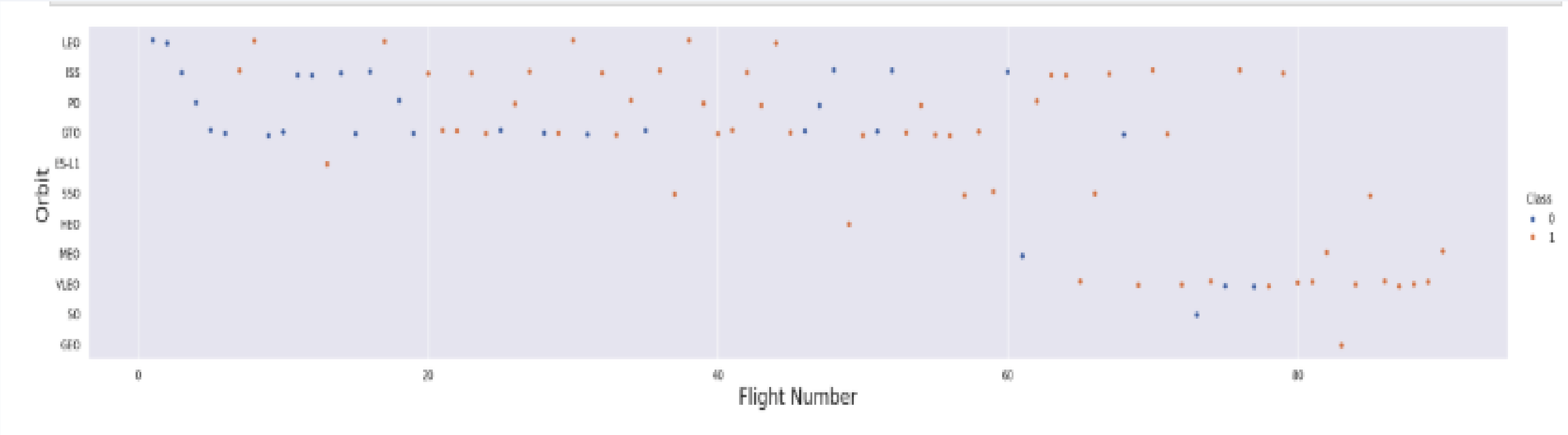
Success Rate vs. Orbit Type

The bar chart depicts the correlation between different orbit types and their corresponding success rates.

- Among the orbits, ES-L1, GEO, HEO, and SSO exhibit the highest success rates (highlighted in green), while the SO orbit category shows the lowest success rate (highlighted in red).

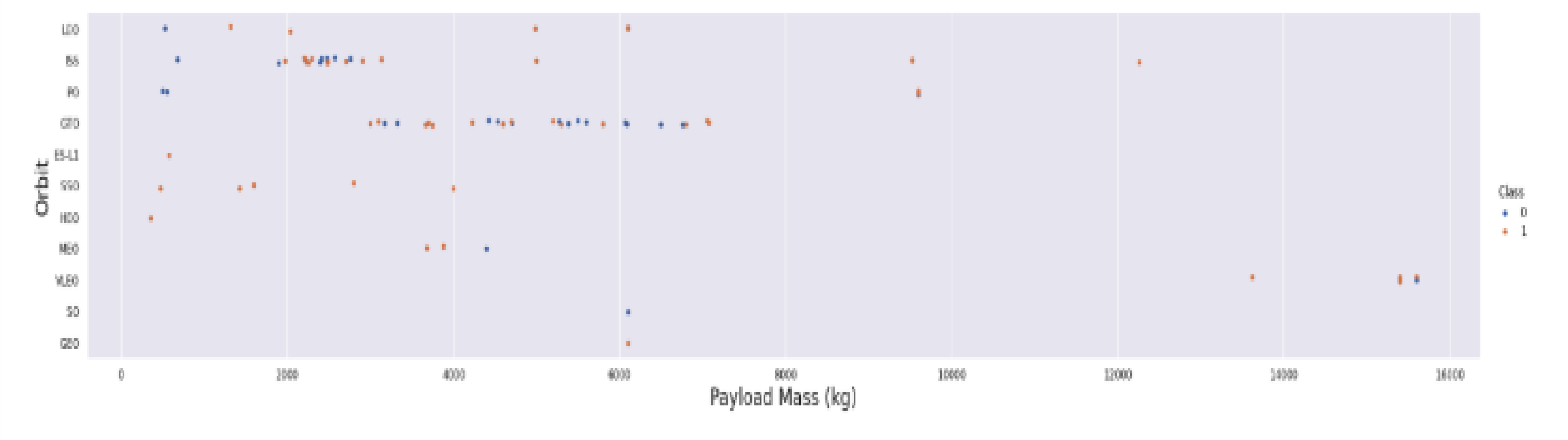


Flight Number vs. Orbit Type



The scatterplot illustrates the correlations between orbit type and flight number, suggesting that the success of the LEO orbit type is associated with the number of flights. In contrast, there appears to be no correlation between flight number and the GTO orbit, as observed in both the preceding bar graph and this scatterplot.

Payload vs. Orbit Type

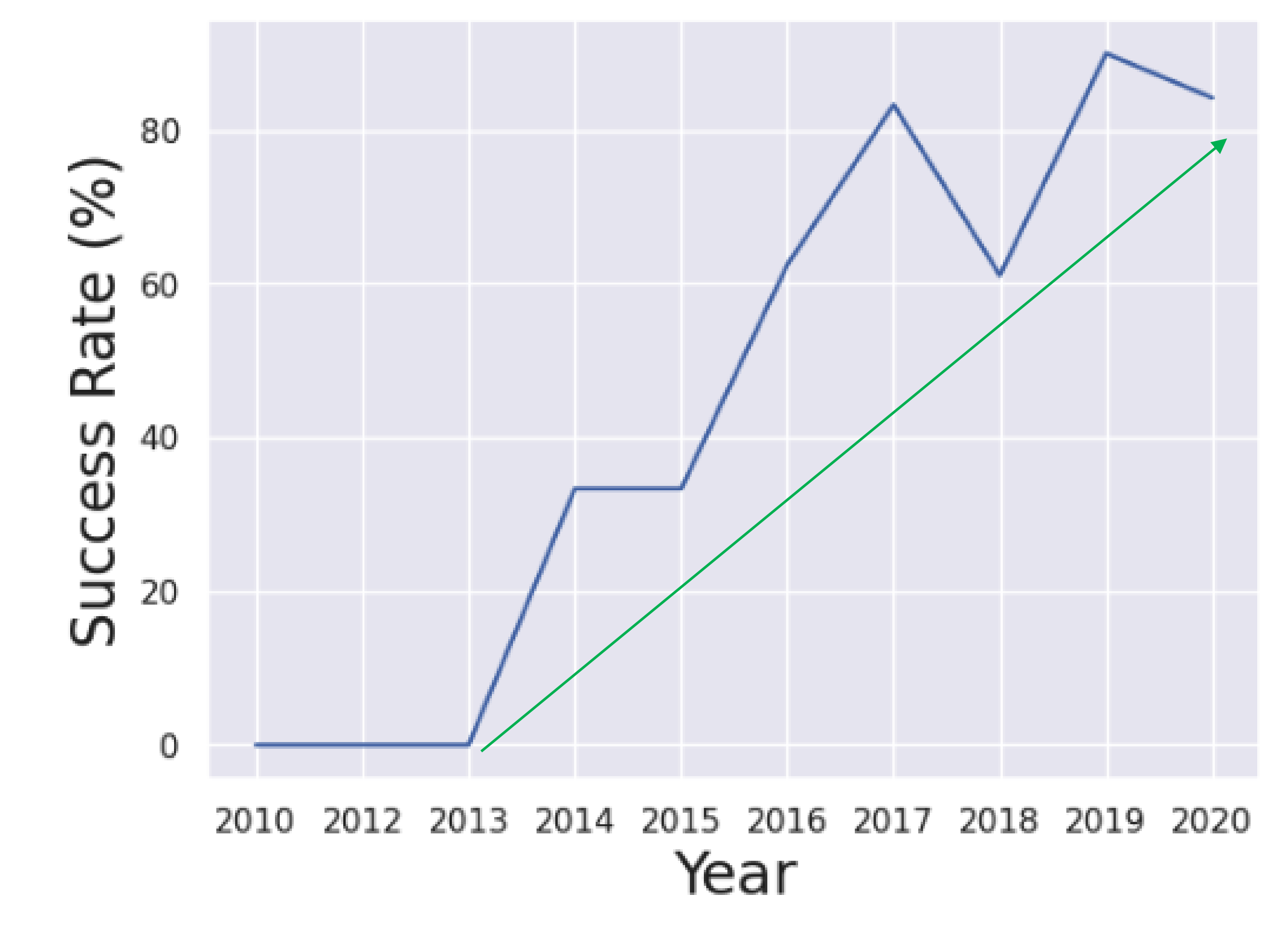


The scatterplot illustrates the relationship between Payload Mass (Kg) and Orbit Type. It reveals a small number of rockets with higher payload masses in the VLEO orbit category, while a greater proportion of rockets carrying lower payload masses are distributed across the other orbit types.

Launch Success Yearly Trend

This line chart illustrates the trend in the average success rate from 2010 to 2020.

- The data clearly demonstrates a rise in the average success rate beginning in 2013 and continuing through to 2020.



All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" from SPACEXTBL;
```

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

I performed an SQL query to identify the distinct launch sites in the given SpaceX dataset. The query utilized the SELECT and DISTINCT statements, specifying the column Launch site and retrieving data from the SPACEXTBL table.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_(
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N(
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N(
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N(

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_site" LIKE 'CCA%' LIMIT 5;
```

I executed a query to retrieve five records where the launch site name starts with the string ‘CAA’.

Total Payload Mass

Total_Payload_Mass
None

%sql SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass FROM SPACEXTBL WHERE "Launch_Service_Provider" = 'NASA (CRS)';

I executed this SQL query to calculate the total payload mass by boosters launched under NASA (CRS).

Average Payload Mass by F9 v1.1

Average_Payload_mass
2928.4

```
%sql SELECT AVG("Payload_Mass__kg_") AS Average_Payload_mass FROM SPACEXTBL WHERE "Booster_Version" = 'F9 v1.1';
```

The mean payload mass transported by the F9 v1.1 booster version was 2,928.4 kg, calculated by applying the AVG("Payload_Mass_kg") statement.

First Successful Ground Landing Date

I executed the query to identify the earliest successful landing date by using the MIN("Date") statement. The first successful landing occurred on December 12, 2015.

First_Successful_Landing_Date
2015-12-22

```
%sql SELECT MIN("Date") AS First_Successful_Landing_Date FROM SPACEXTBL WHERE
"Landing_Outcome" = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

I executed an SQL query to retrieve successful drone ship landings with payloads ranging from 4,000 to 6,000. The query selected the booster version and successful landing status from the SpaceX data frame, specifying the relevant column names and using > , < as well as the AND statement to filter the payload mass within the defined range.

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

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)'
AND "Payload_Mass__kg_" > 4000 AND "Payload_Mass__kg_" < 6000;
```

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	total_count
Success	98

```
%sql SELECT "Mission_Outcome", COUNT(*) AS total_count FROM SPACEXTBL WHERE "Mission_Outcome" IN
('Success', 'Failure') GROUP BY "Mission_Outcome";
```

The total number of successful mission Outcome was 98, I did not manage to retrieve the total number of total failures

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE
Payload_Mass__kg_ = (SELECT MAX(Payload_Mass__kg_)FROM SPACEXTBL);
```

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

I executed a query with a sub-query to identify the Booster versions that have transported the maximum payload mass; there are 12 distinct booster versions that have achieved this maximum payload capacity.

2015 Launch Records

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

```
%%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" = 'Failure (drone ship)';
```

I executed a SQL query to extract the month name, landing outcome, booster version, and launch site. The data indicates that landing failures occurred in January and April of 2015.

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

```
%sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count
FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND
'2017-03-20' GROUP BY "Landing_Outcome" ORDER BY
outcome_count DESC;
```

This SQL query ranked landing outcomes in descending order for the period between June 4, 2010, and March 20, 2017. The outcome count provides detailed information on successes and failures, including the conditions for success and types of failure.

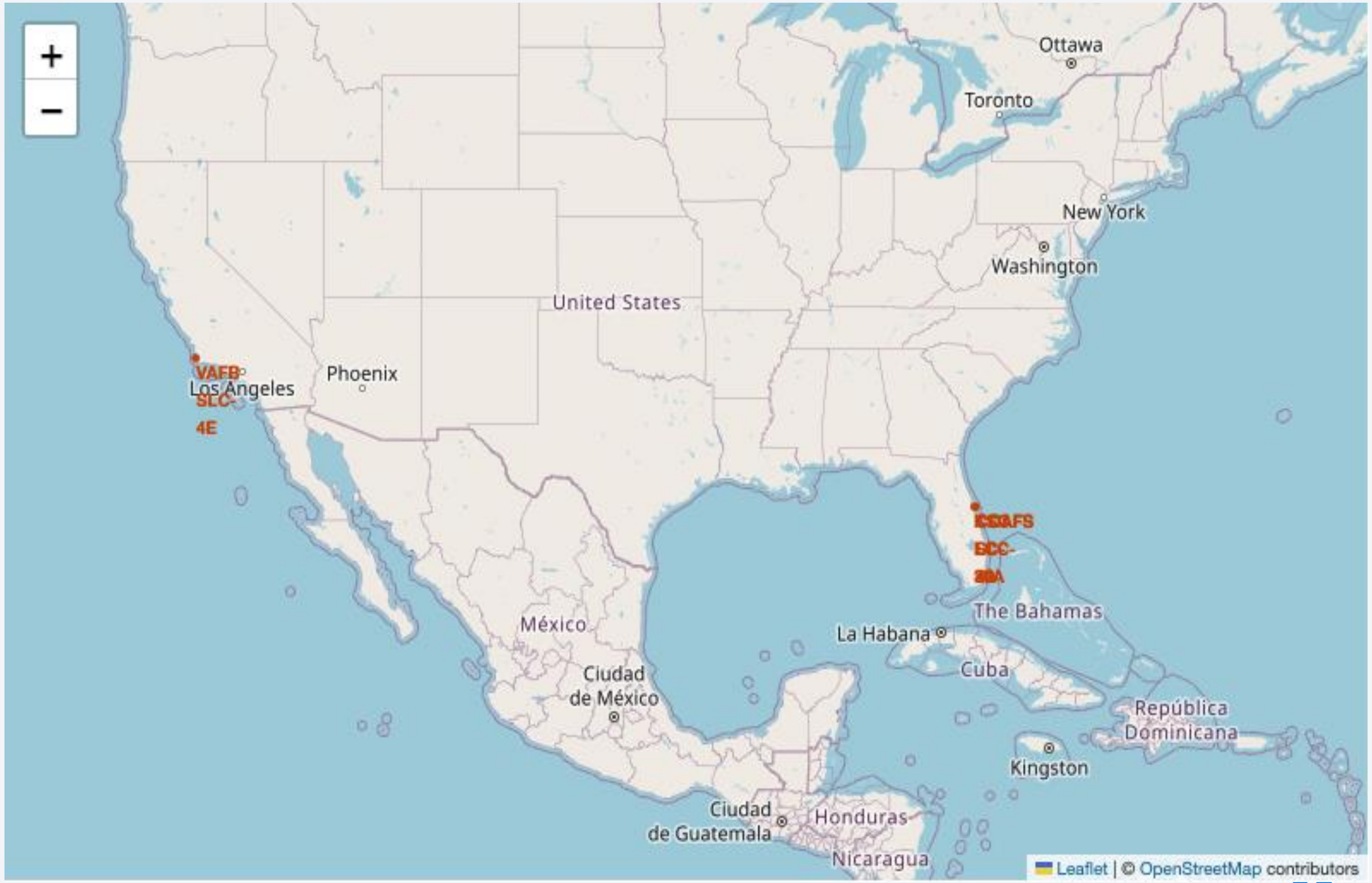
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

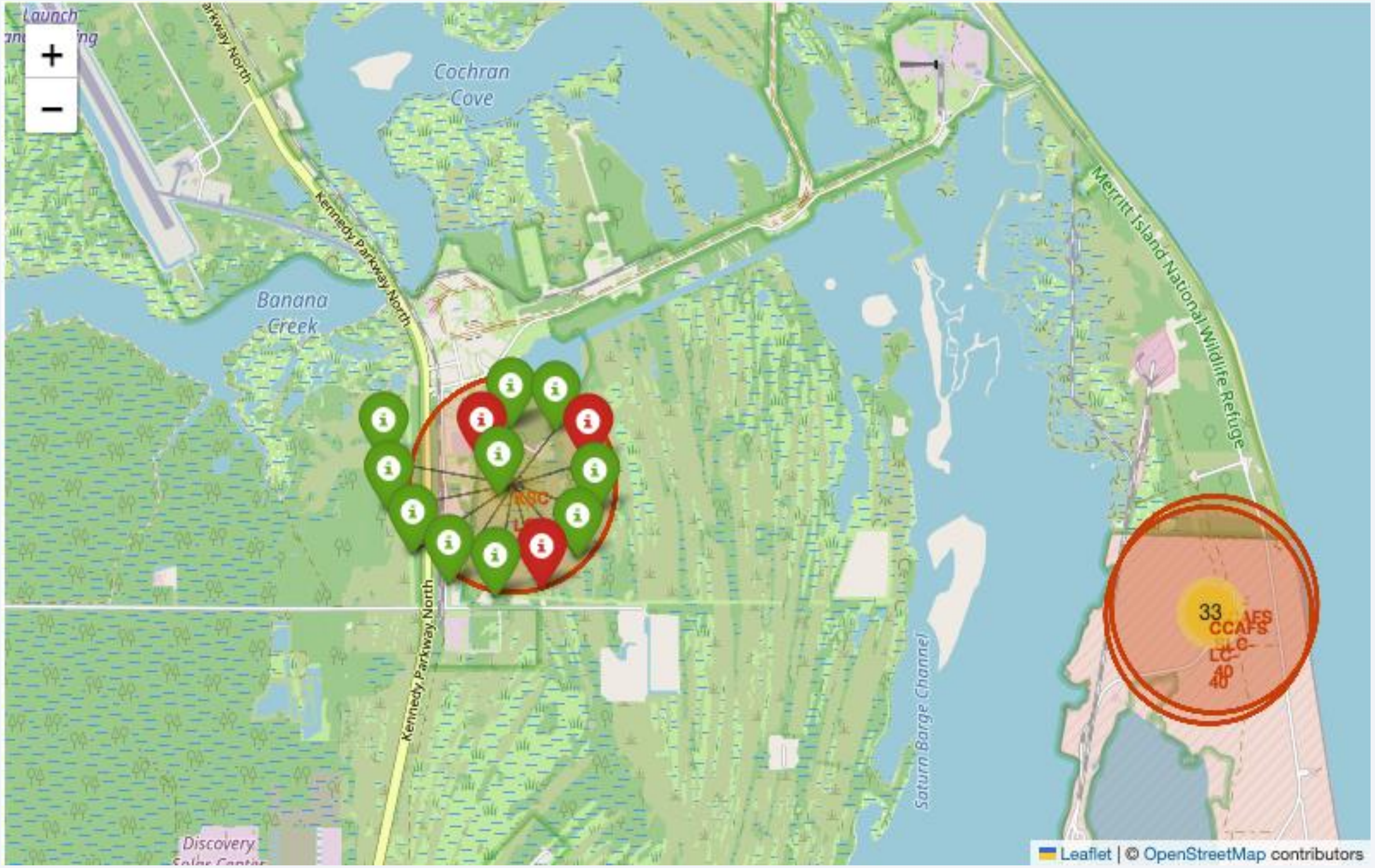
All SpaceX Launch Sites in the US (Folium)

All launch site locations are indicated and labelled in red. The majority of these sites are based in Florida, with an additional launch site located in California.



Color Labeled Launch Outcomes with Folium

This folium-generated labeled map highlights launch outcomes at designated sites, marking successful launches in green and unsuccessful ones in red. The launch site shown here is situated in Florida.



Calculating distances In Folium

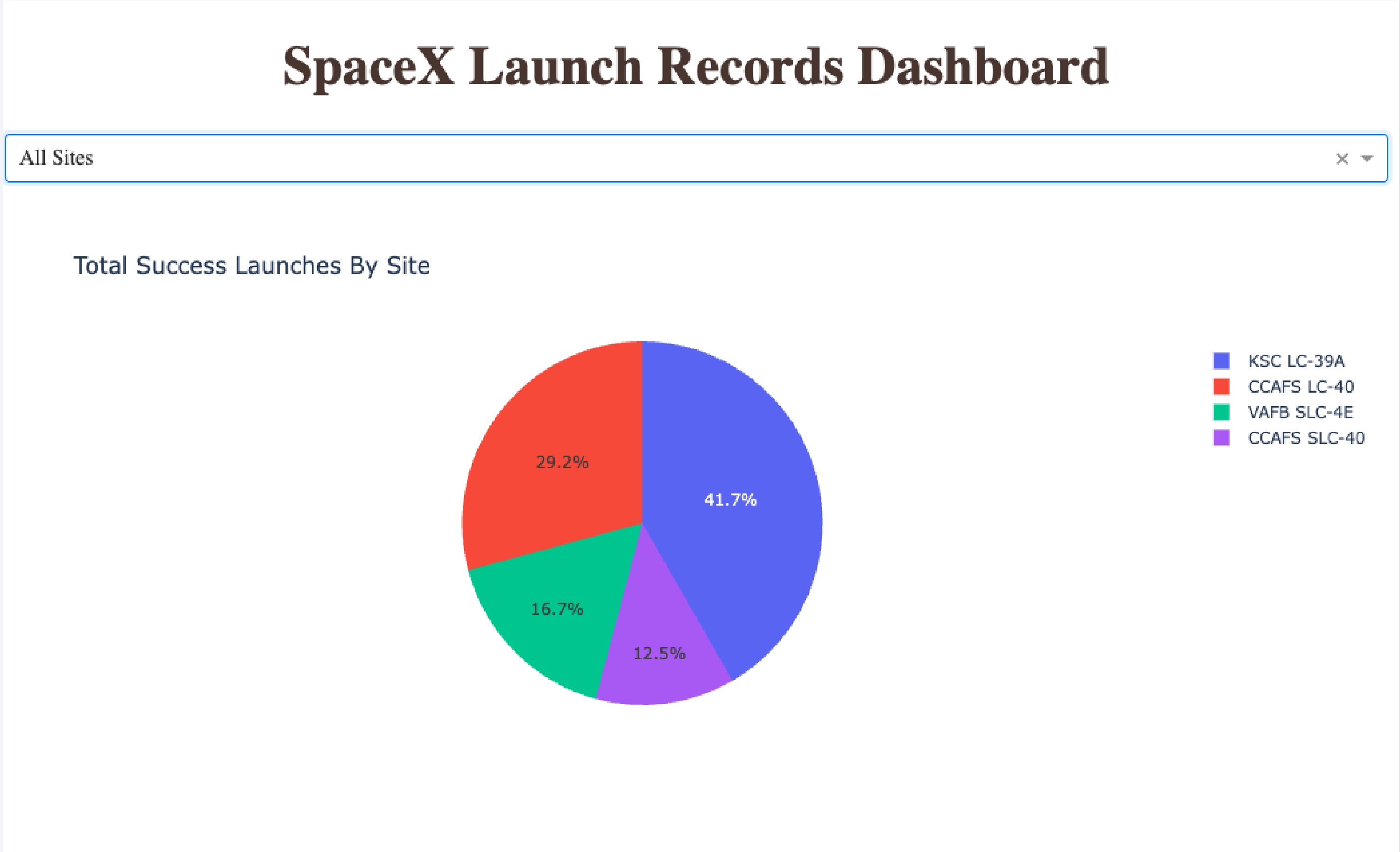
- The map displays the distance from a launch site to the nearest coastline. By using MousePosition to determine the coordinates of the closest coastline, the distance was initially calculated. A polyline was then employed to illustrate this distance, accompanied by a distance_marker to indicate the measurement.

Section 4

Build a Dashboard with Plotly Dash

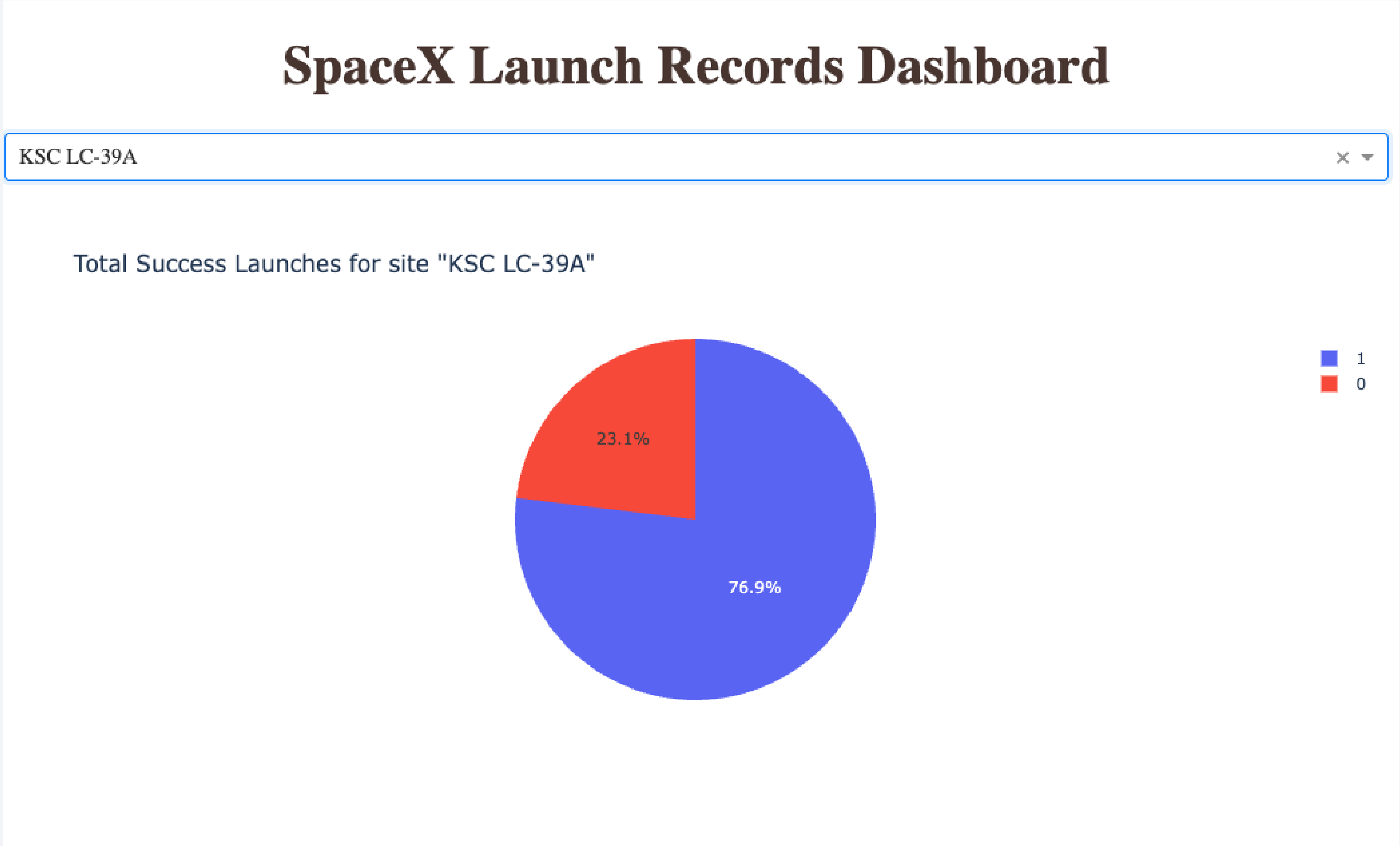
SpaceX launch Records Dashboard Piechart

This pie chart illustrates the successful launches across various sites, with each segment color-coded to represent a specific launch location. Notably, KCS LC-39A leads with the highest proportion of successful launches, comprising 41.7% of the total.



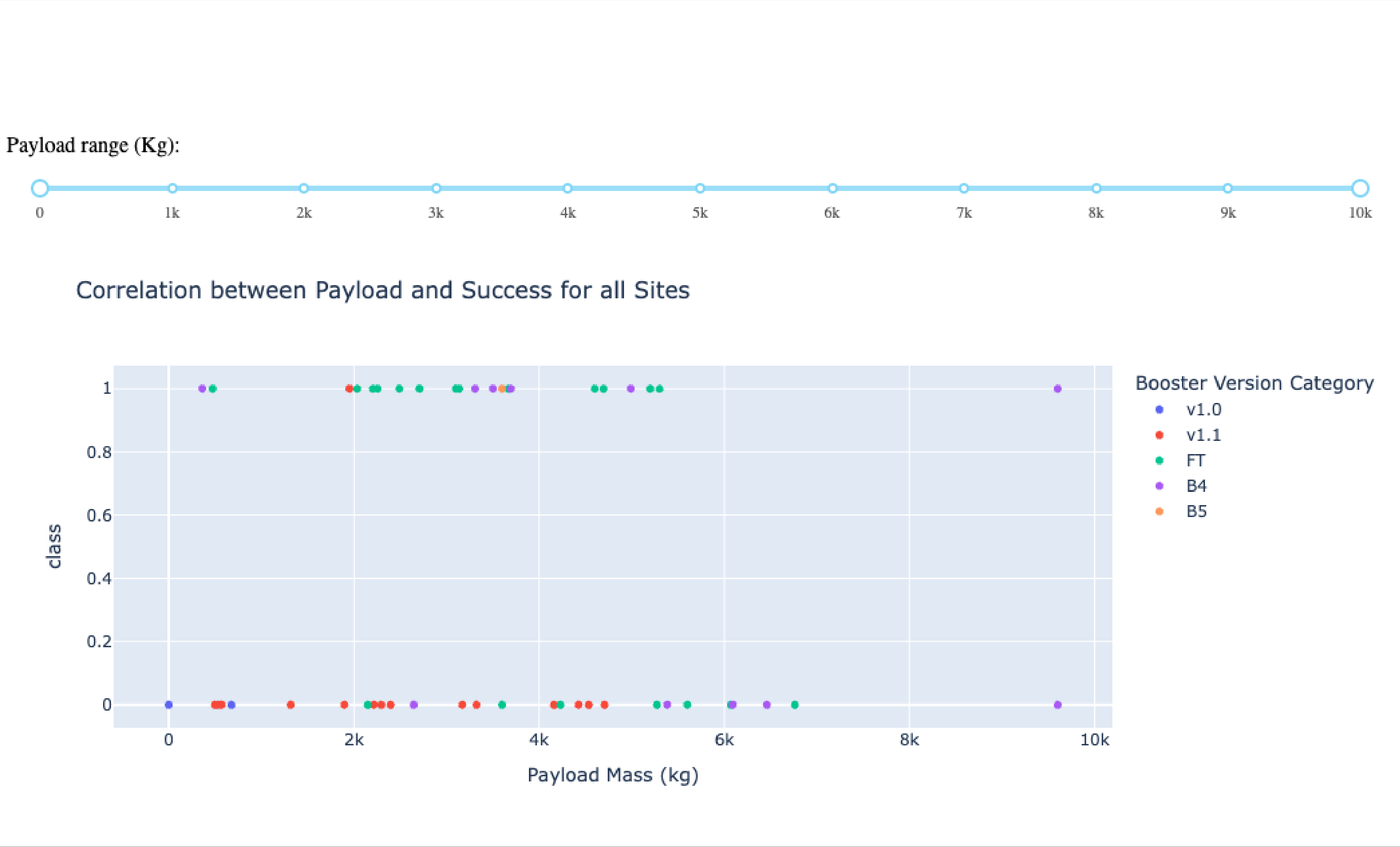
SpaceX Launch Records Dashboard

This pie chart illustrates the highest launch success rate, with successful launches accounting for 76.9% of the total.



Interactive Analytics Dashboard with plotly

This section of the dashboard application features a slider for choosing various payload masses, with colours representing different booster version categories. The scatter plot illustrates the relationship between payload mass and mission success across all launch sites.

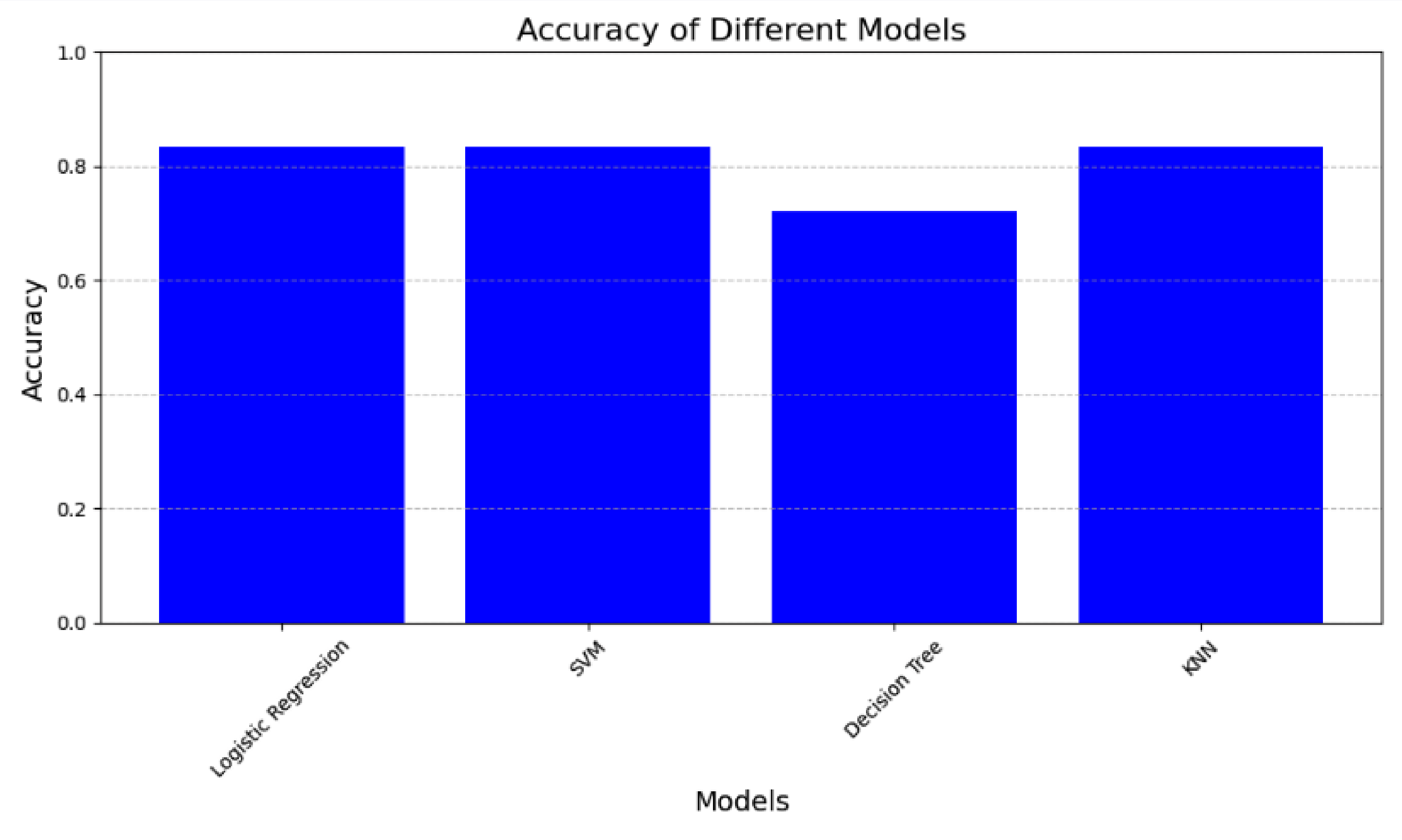


Section 5

Predictive Analysis (Classification)

Classification Accuracy

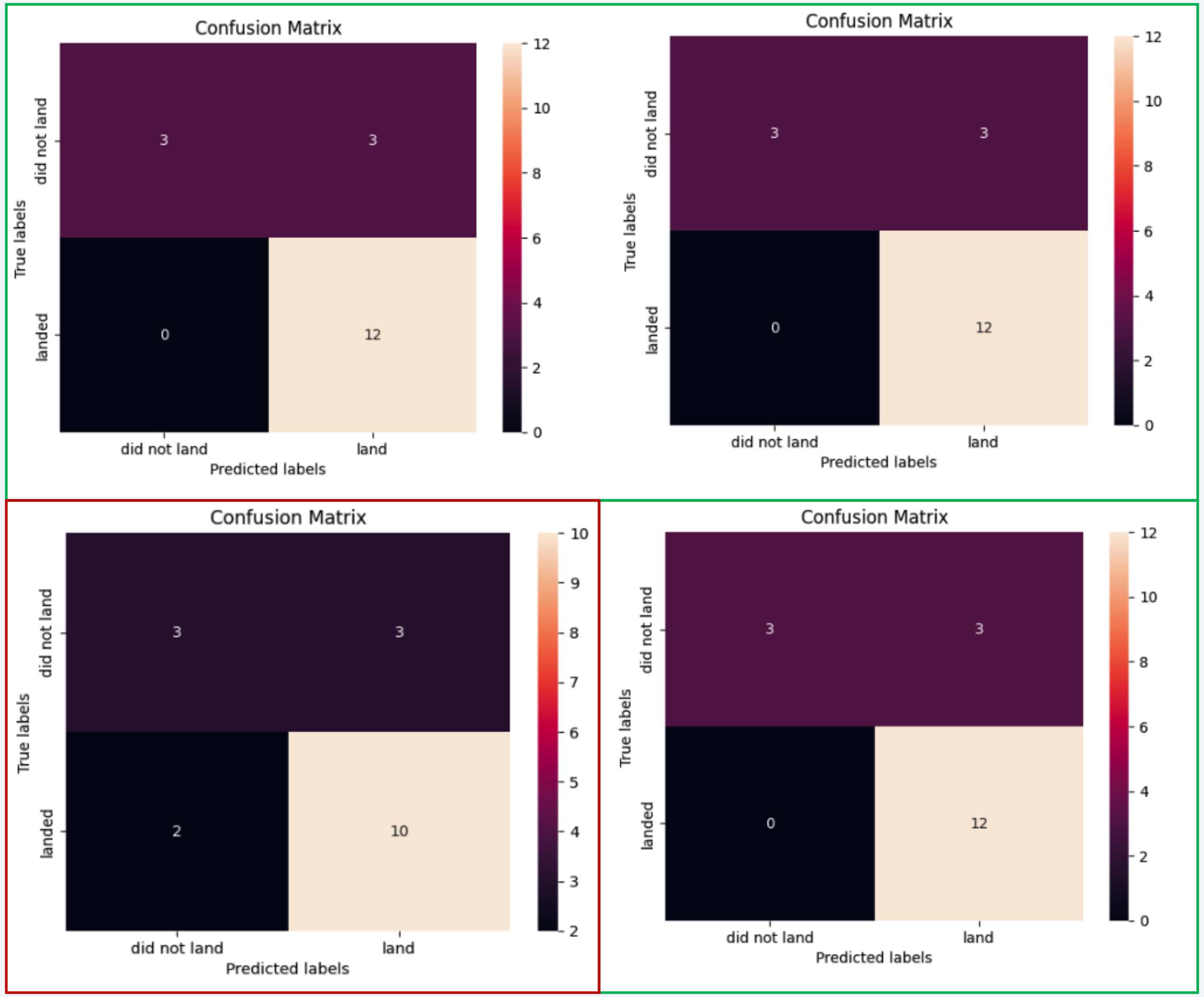
After assessing four different models with the test dataset and illustrating the outcomes in a bar chart, it is clear that the KNN, SVM, and Logistic Regression classifiers exhibit the strongest performance.



Confusion Matrix

The test accuracies of LR, SVM, and KNN were identical, demonstrating that all three models outperformed the DT model.

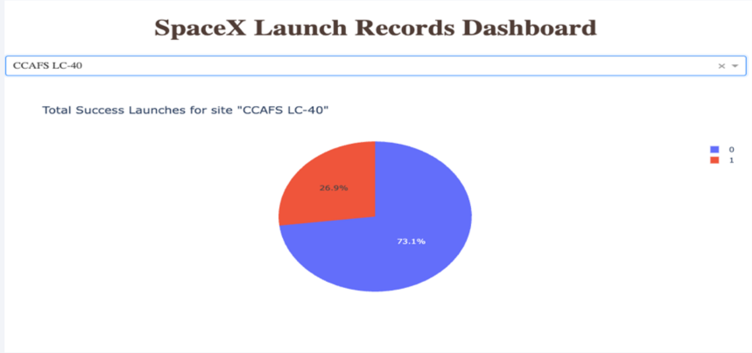
```
LR Test Accuracy: 0.8333333333333334
SVM Test Accuracy: 0.8333333333333334
KNN Test Accuracy: 0.8333333333333334
DT Test Accuracy: 0.7222222222222222
```

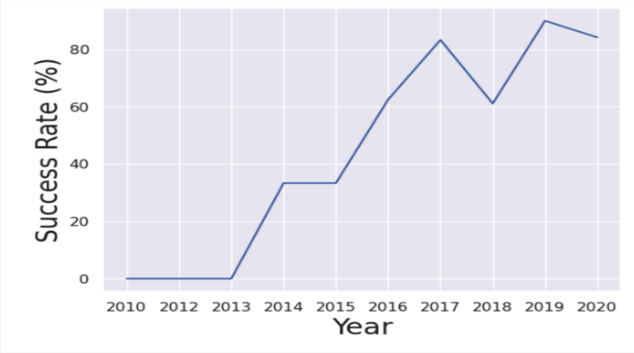


Conclusions

- A significant correlation exists between booster versions carrying lighter payloads and higher launch success rates compared to those with heavier payloads.
- Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) have proven to be the most effective models for predicting outcomes using SpaceX data.
- Among all launch sites, Kennedy Space Center's LC-39A boasts the highest success rate for launches.
- Based on lessons learned from previous launches and ongoing development, SpaceX's launch success rate is expected to increase substantially, indicating a positive trend for future missions.

```
LR Test Accuracy: 0.8333333333333334
SVM Test Accuracy: 0.8333333333333334
KNN Test Accuracy: 0.8333333333333334
DT Test Accuracy: 0.7222222222222222
```





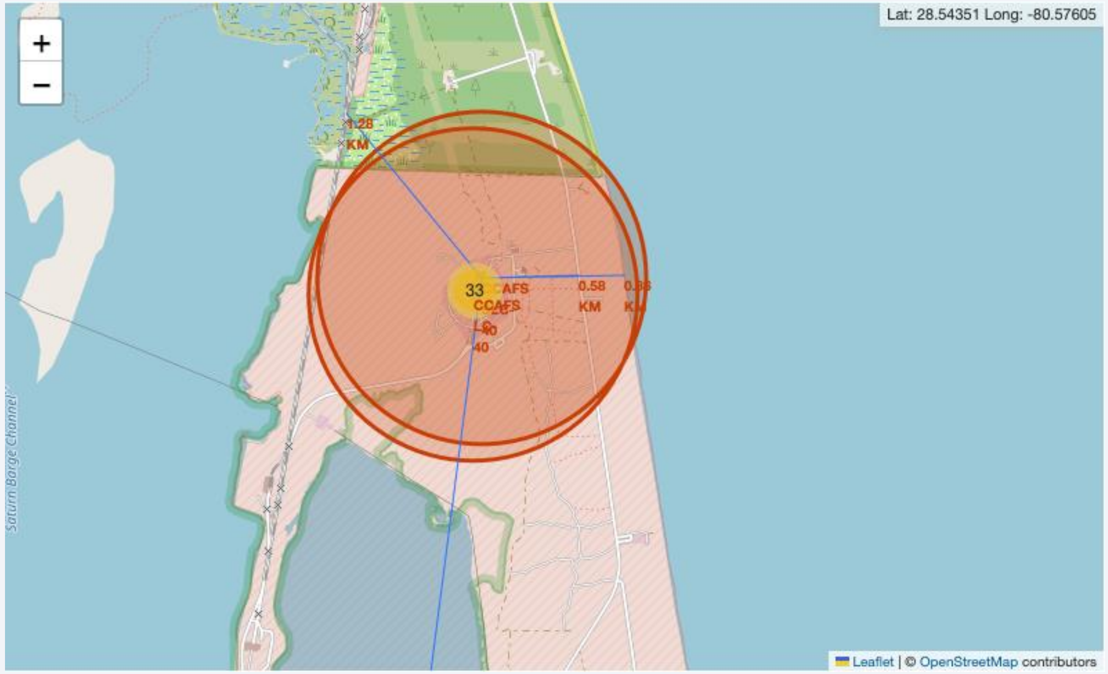
Appendix

```
distance_highway = 0.5834695366934144  km
distance_railroad = 1.2845344718142522  km
distance_city = 51.434169995172326  km
```

Data frame generated from web scraping results.

Distances to nearest highway, railroad and city with folium

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX Capella Space and Tyvak	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA (CRS)	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26



Thank you!