IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

INCHEKEL Massinissa Nacim
05 October 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result from Machine Learning Lab

Introduction

SpaceX has transformed space travel with its reusable Falcon 9 rocket, significantly lowering launch costs by recovering and reusing the first stage. While SpaceX offers launches at 62 million dollars, far less than the 165 million dollars of other providers, much of the savings come from this reusability. Accurately predicting whether the first stage will land successfully is critical, as it impacts overall mission cost-effectiveness. For competitors or government agencies looking to bid against SpaceX, such predictions are valuable for estimating launch costs and making informed decisions.

In this capstone project, the following key questions will be addressed:

Can we accurately predict whether the Falcon 9 first stage will land successfully after a launch?

What factors most influence the success of the first stage landing?

How can the prediction of successful landings be used to estimate the overall cost of a Falcon 9 launch?

How does the ability to predict the success of a landing compare to current industry standards for space launches?

By collecting, cleaning, and analyzing the data from SpaceX launches, the project aims to provide valuable insights into the success of the Falcon 9 first stage landings and their potential cost implications.

Section 1

Methodology

Methodology

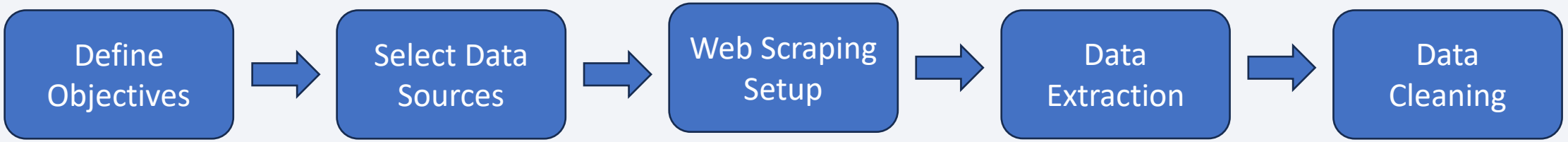
Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping (Wikipedia)
- Perform data wrangling
  - Data Inspection / Data Cleaning / Data Transformation / Data Validation
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

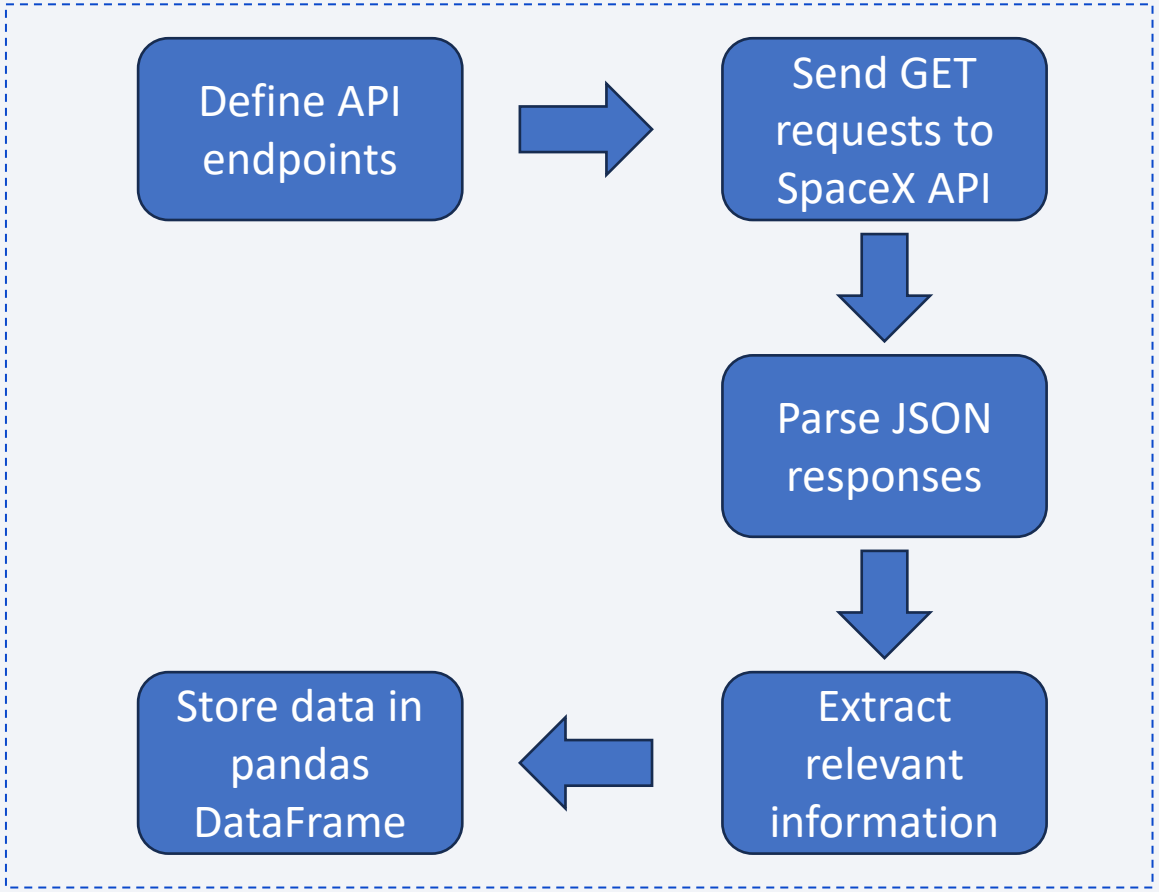
- Datasets are collected from Rest SpaceX API and Webscrapping Wikipedia :
  - Rest SpaceX API : The data sets were collected from public APIs, specifically targeting information related to SpaceX launches.

  - Webscrapping Wikipedia : Choose websites or online databases relevant to the objectives and ensure that the sources are reliable and contain the required information.



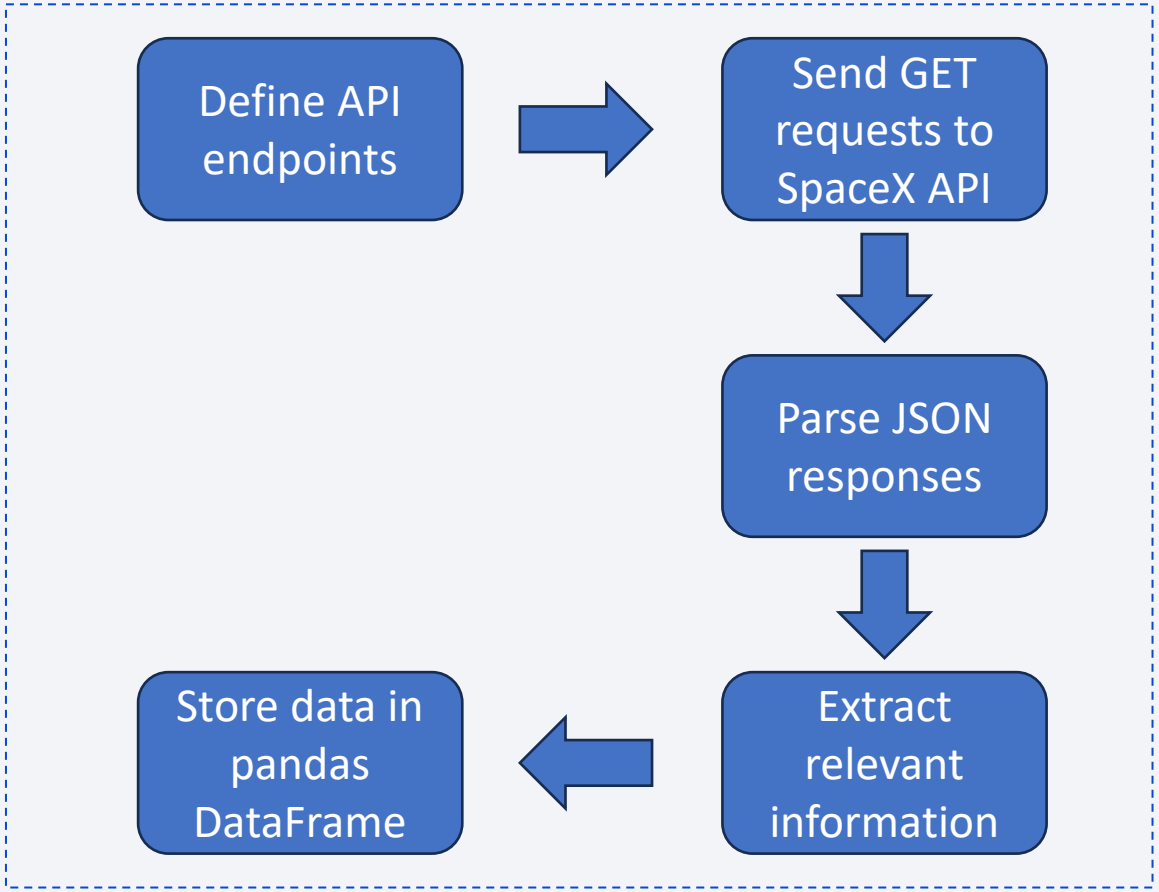
Data Collection – SpaceX API

- The data collection process involves gathering detailed SpaceX launch information through their API. By defining specific endpoints and sending GET requests, the API responses are parsed to extract key data such as launch dates, rocket specifications, payload details, and landing outcomes. This information is then organized and stored in a pandas DataFrame for further analysis.
- The GitHub URL of the completed SpaceX API calls notebook



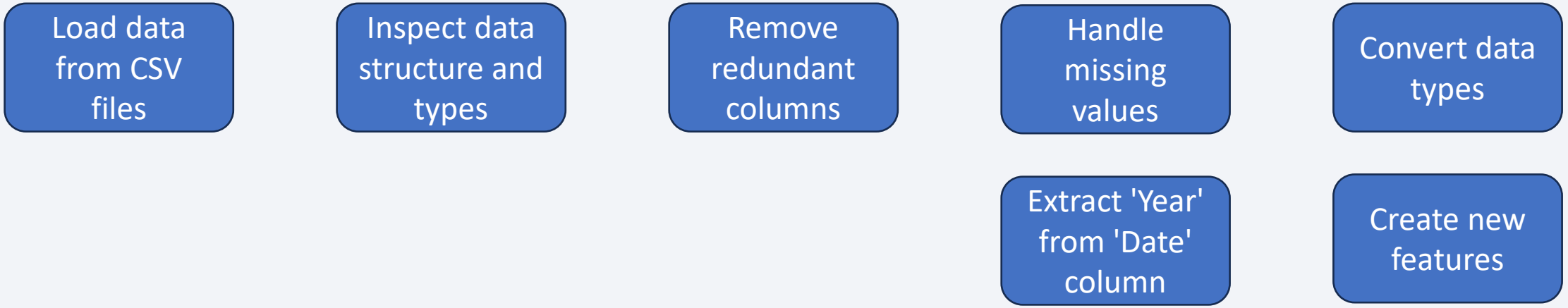
Data Collection - Scraping

- This process involves collecting additional SpaceX launch data from Wikipedia pages using web scraping tools like Beautiful Soup and Requests. After identifying the target URLs, HTTP requests are sent to retrieve the HTML content. Relevant tables and text are then parsed, cleaned, and structured. The data collected includes supplementary launch details, historical context, and additional rocket information for a more comprehensive dataset.
- The GitHub URL of the completed web scraping notebook



Data Wrangling

- The data wrangling process focuses on cleaning and preparing the collected data for analysis. It involves loading data from CSV files, inspecting the structure, removing redundant columns, handling missing values, and converting data types for consistency. New features, such as 'Class' for landing outcomes, are created, and 'Year' is extracted from the 'Date' column. The outcome is a clean, well-structured dataset, ready for exploratory data analysis (EDA) and modeling.
- The GitHub URL of your completed data wrangling related notebooks .



EDA with Data Visualization

- The exploratory data analysis (EDA) phase uses visualization tools like Matplotlib and Seaborn to uncover patterns and relationships within the data. Key visualizations include scatter plots for flight numbers, payloads, launch sites, and orbit types, as well as a bar chart showing success rates by orbit type. A line chart illustrates the yearly trend of launch success. These visualizations help identify trends and insights that guide further analysis.

- The GitHub URL of your completed EDA with data visualization notebook .

EDA with SQL

Summary of SQL Queries Performed:

- Identified unique launch sites.
- Queried launches from specific sites (e.g., CCA).
- Calculated total payload for NASA boosters.
- Found average payload for F9 v1.1 boosters.
- Retrieved the date of the first successful ground landing.
- Queried successful drone ship landings within a specific payload range.
- Generated statistics on mission outcomes.
- These queries provide deeper insights into launch performance and trends.

Build an Interactive Map with Folium

Interactive Map Summary:

- An interactive map was created using Folium to visualize SpaceX launch sites. Markers were added to identify each site, along with popup information for user interaction. Circle markers were used to indicate the number of launches and success rates, with colors representing performance (Green for successful landing and Red for unsuccessful landing).
- Additionally, proximity circles show a 100 km radius around each site to visualize potential impact areas. This map provides a clear view of SpaceX's launch site distribution and highlights key insights about site location strategy.
- The GitHub URL of your completed interactive map with Folium map .

Build a Dashboard with Plotly Dash

Dashboard Summary:

- The interactive dashboard, built with Plotly Dash, allows users to explore SpaceX launch data through several dynamic features. A launch site dropdown filters data by location, while a success rate pie chart updates to show the proportion of successful and failed launches.
- A payload range slider enables users to filter launches based on payload mass, and a scatter plot visualizes the relationship between payload and launch success. These features provide real-time updates and insights, making the dashboard an intuitive tool for analyzing SpaceX's launch performance.
- The GitHub URL of your completed Plotly Dash lab .

Predictive Analysis (Classification)

In this predictive analysis, we developed and evaluated classification models to predict Falcon 9 first stage landing outcomes. The process involved the following key steps:

Data Preprocessing:

- Feature selection using correlation analysis.
- One-hot encoding of categorical variables.
- Standardization of numerical features.
- Split data into training (80%) and testing (20%) sets.

Model Development:

- Four classification models were implemented:
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)

Model Evaluation:

- Evaluated models using the following metrics:
  - Accuracy
  - Jaccard Index
  - F1-Score
  - Log Loss (for Logistic Regression)

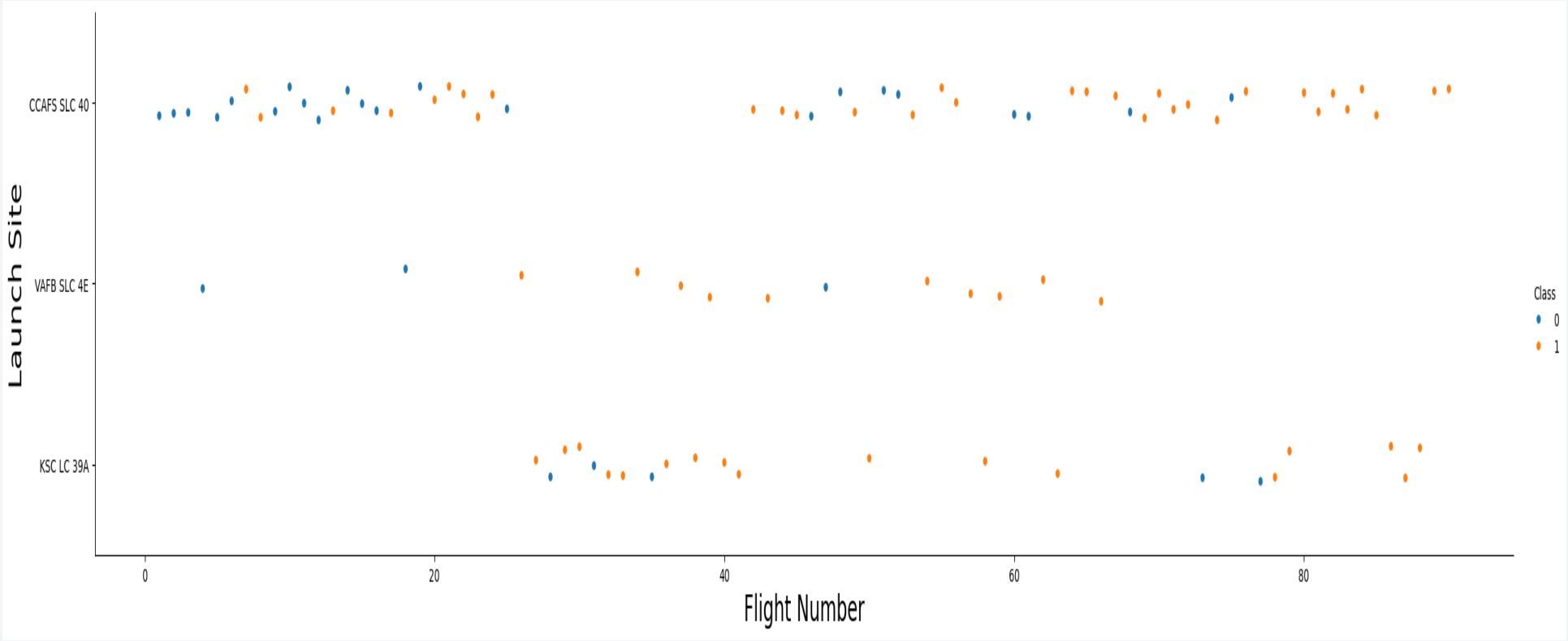
Results

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

Section 2

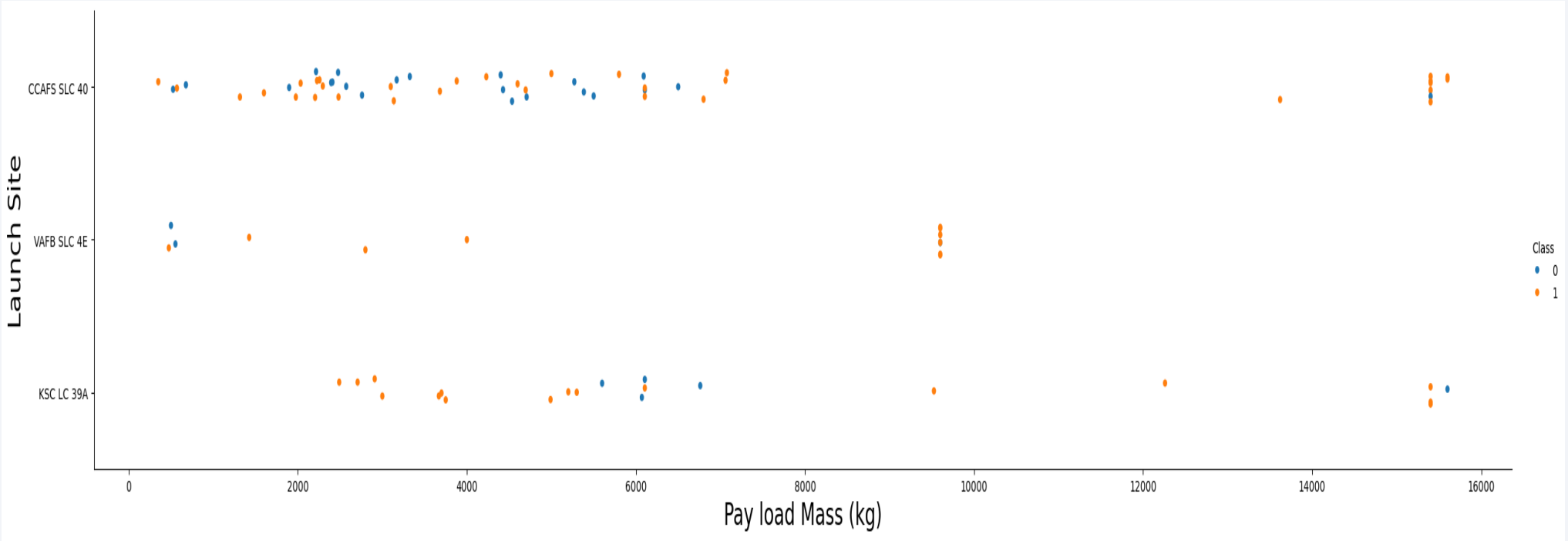
Insights drawn from EDA

Flight Number vs. Launch Site



The data indicates that both CCAFS SLC 40 and KSC LC 39A have higher launch frequencies and show an improvement in success rates over time. VAFB SLC 4E has fewer launches, making it harder to discern clear patterns. Overall, the increasing success rate with higher flight numbers suggests advancements in launch technologies or methodologies.

Payload vs. Launch Site



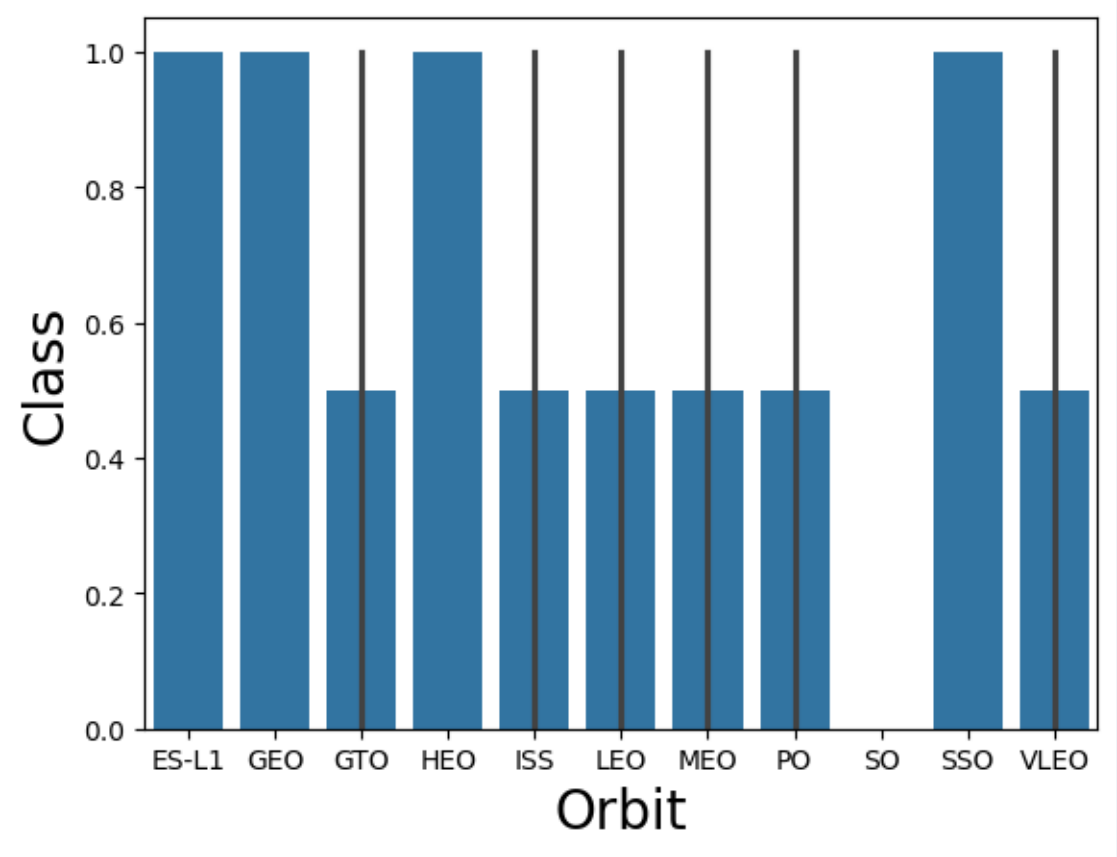
Across all sites, successful launches are observed across a wide range of payload masses. However, larger payloads tend to have higher success rates, indicating improved capabilities in handling heavier missions.

Success Rate vs. Orbit Type

The bar chart shows the success rates for different orbits. The orbits with the highest success rates, indicated by a value of 1.0, are:

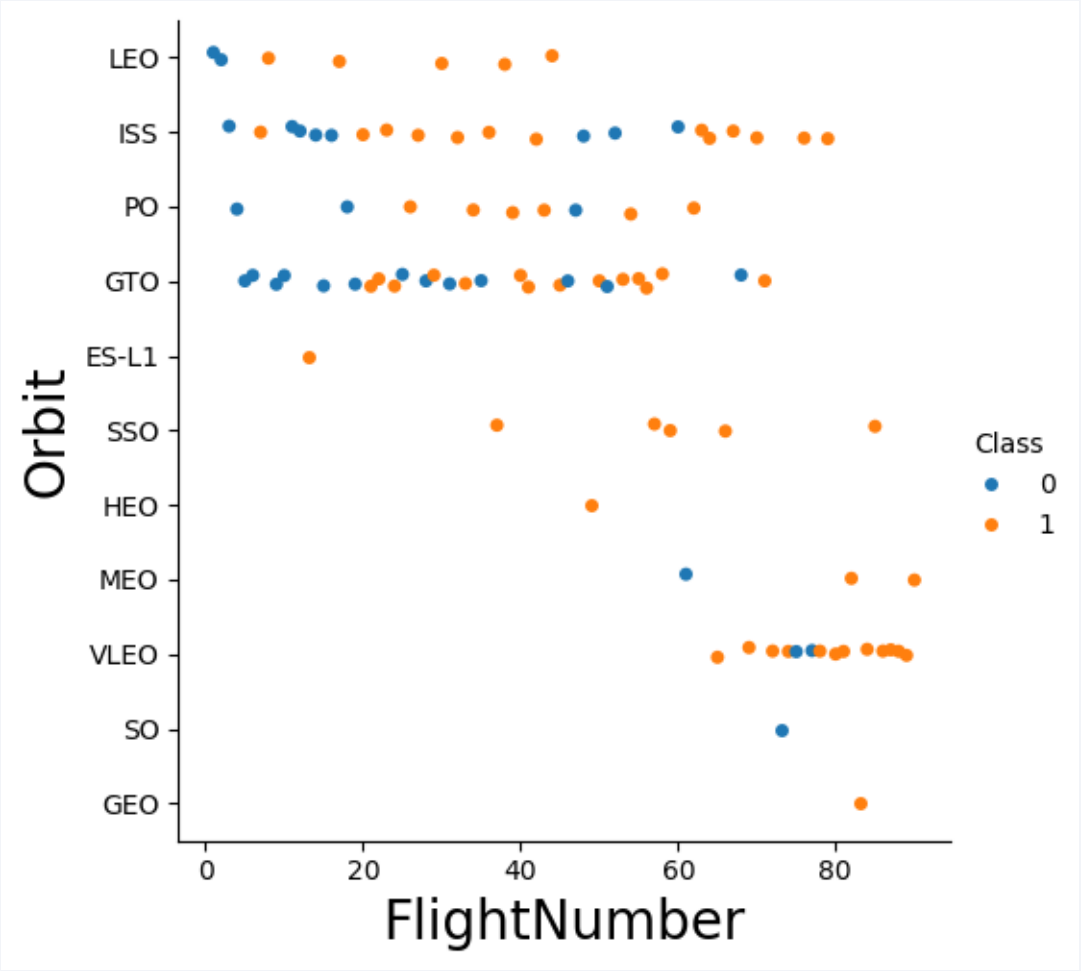
- ES-L1
- GEO
- HEO
- SSO

These orbits have bars reaching the top of the chart, indicating a 100% success rate.



Flight Number vs. Orbit Type

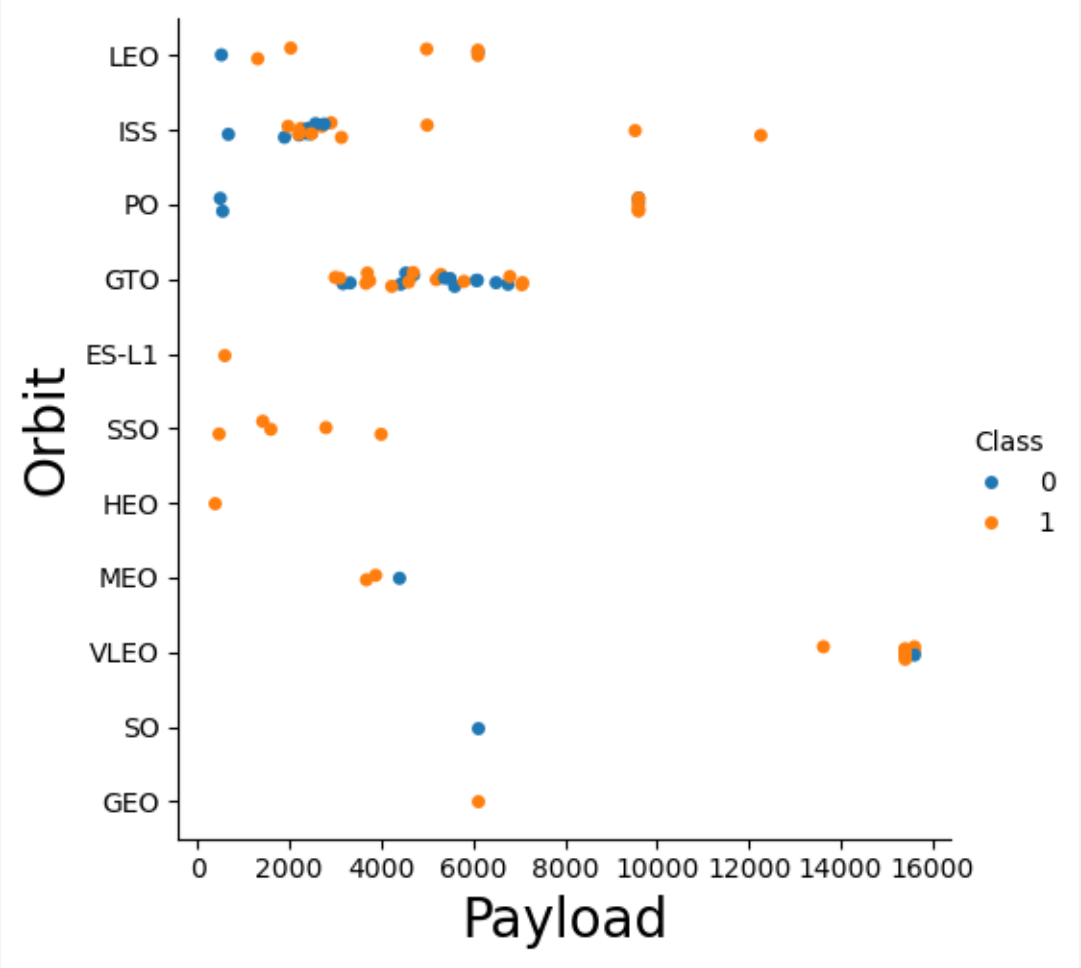
You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

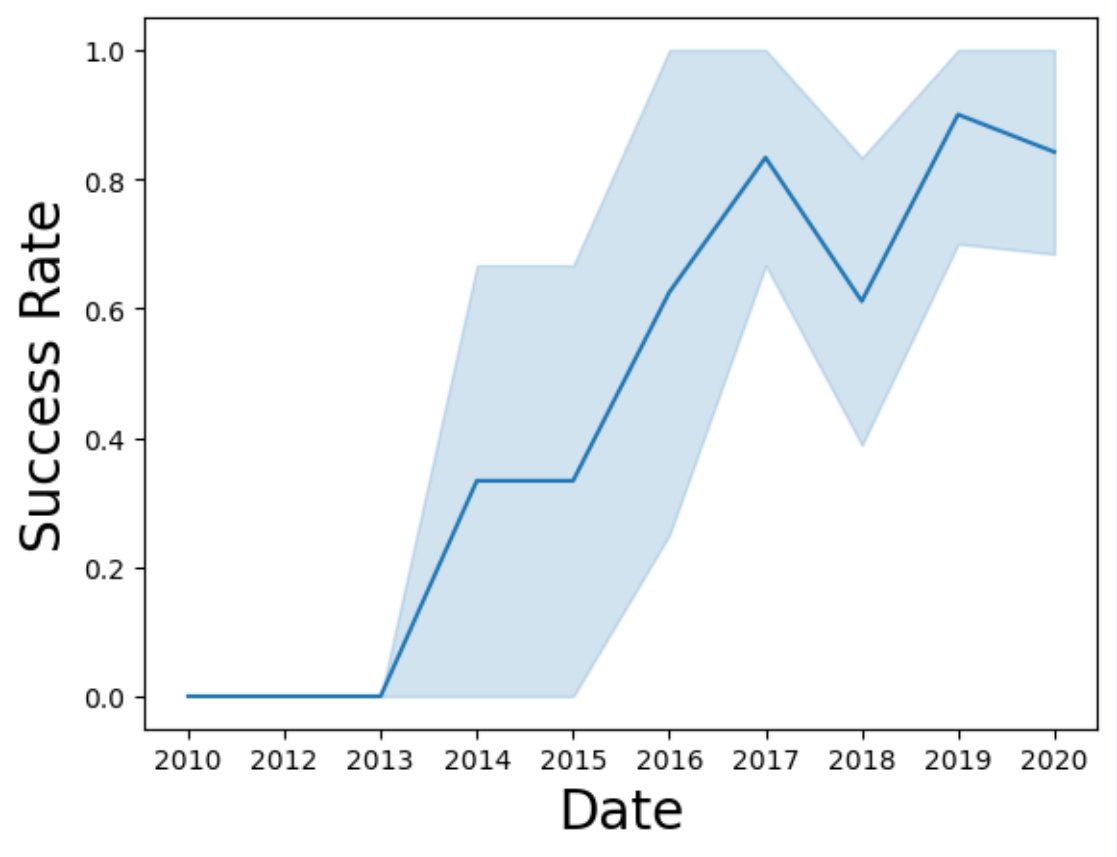
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

you can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

The query in the image retrieves distinct values from the LAUNCH_SITE column in the SPACEXTBL table using SQLite. The result shows a list of unique launch sites, which are:

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

This indicates the different locations where launches have occurred according to the database

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
 * 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'

The query retrieves five records from the SPACEXTBL table where the launch sites start with "CCA". The results are limited to five entries, showing launches from the CCAFS LC-40 site.

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

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

```
* 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

Total Payload Mass

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

This command sums up the PAYLOAD_MASS__KG_ column for all records in the SPACEXTBL table where the CUSTOMER is 'NASA (CRS)'. The result is a total payload mass of 45,596 kg.

Average Payload Mass by F9 v1.1

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'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

The query calculates the average payload mass carried by the booster version "F9 v1.1" from the database table SPACEXTBL.

The result shows that the average payload mass is approximately 2928.4 kg.

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(DATE)
2015-12-22

The query retrieves the earliest date when a successful landing on a ground pad was achieved. It uses the SQL min function to find the minimum date from the SPACEXTBL table where the Landing_Outcome is 'Success (ground pad)'. The result shows that the first successful landing on a ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 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 = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and
```

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

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

The query retrieves the names of SpaceX boosters that successfully landed on a drone ship with a payload mass between 4000 and 6000 kg. The SQL command selects the BOOSTER_VERSION from the SPACEXTBL table, filtering for successful drone ship landings and the specified payload mass range.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

The query in the image counts the total number of missions with outcomes labeled as either "Success" or "Failure (in flight)" from the SPACEXTBL table. The result shows that there are 99 such mission outcomes in total.

Boosters Carried Maximum Payload

- The result lists several booster versions, such as F9 B5 B1048.4 and F9 B5 B1049.4, indicating multiple boosters have achieved this maximum payload capacity.

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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

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, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE substr(DATE, 1,
```

```
* 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

The SQL query in the image retrieves records from a database table named SPACEXTBL. It selects the month, landing outcome, booster version, and launch site for launches that occurred in 2015 with a specific landing outcome ("Failure (drone ship)"). The query uses the substr function to extract the month and year from the DATE

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

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 '20
```

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

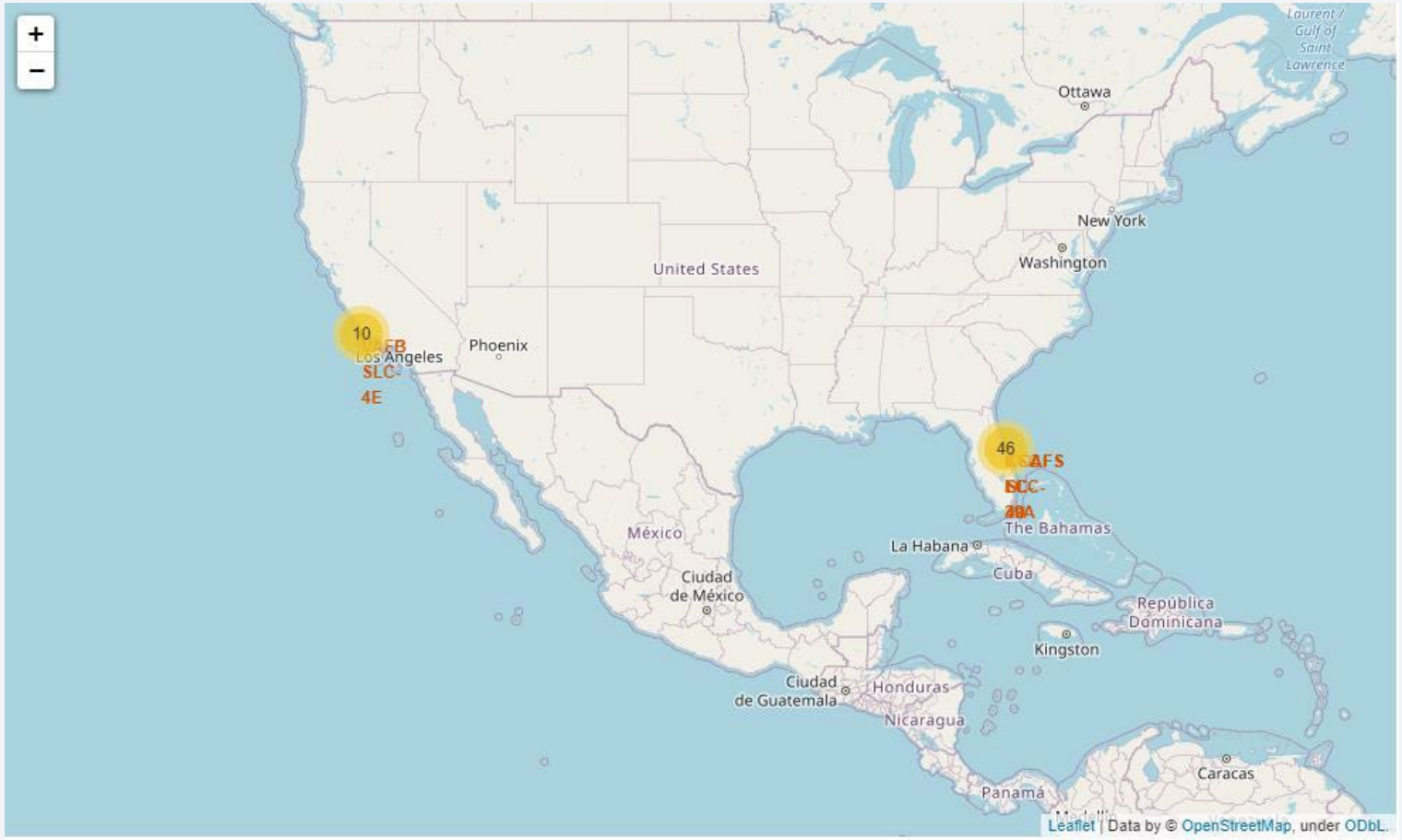
Landing_Outcome	OUTCOME_COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The query result shows the count of different landing outcomes for SpaceX launches between June 4, 2010, and March 20, 2017. The outcomes are ranked in descending order based on their frequency


Section 3
Launch Sites
Proximities Analysis

All launch sites' location markers on a global map

This visualization helps in understanding the concentration and distribution of launch sites across these regions

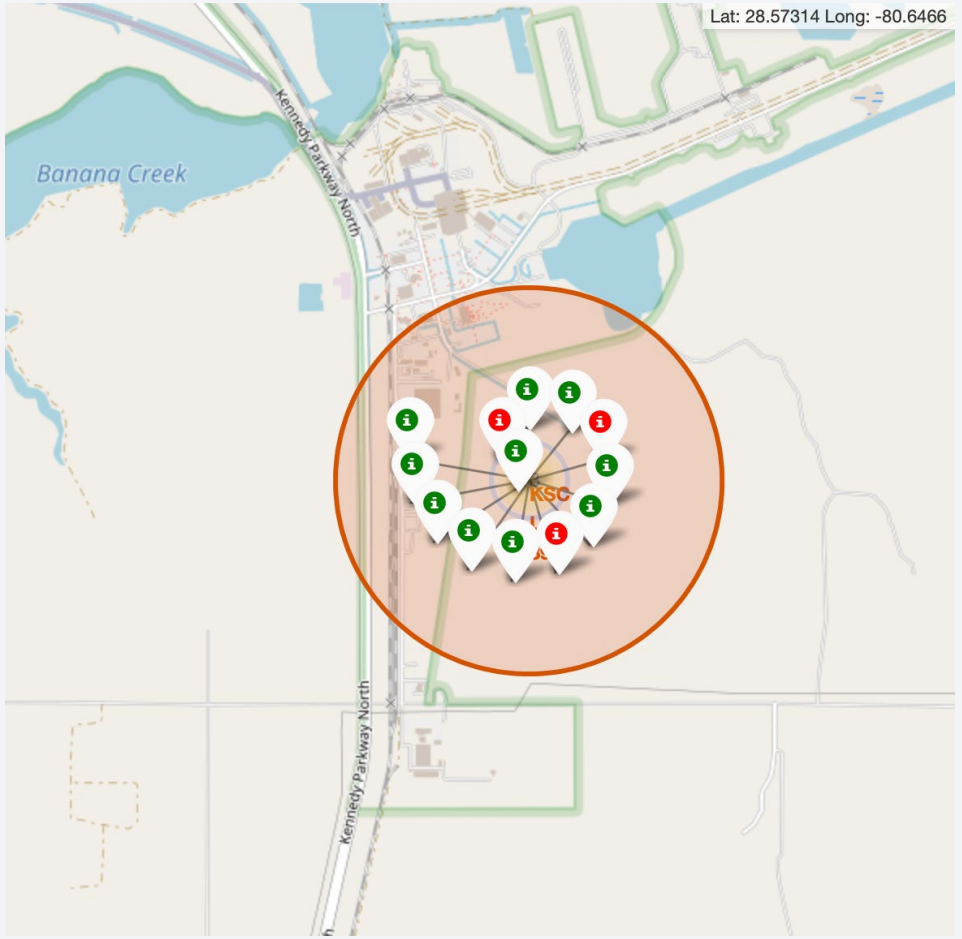


Markers showing launch sites with color labels

The screenshot shows a map centered around the Kennedy Space Center (KSC) with markers indicating launch outcomes. Here are the key elements and findings:

- Green Markers: Indicate successful launch outcomes.
- Red Markers: Indicate unsuccessful launch outcomes.

This visualization helps in quickly assessing the success rate of launches from this site, with more green markers indicating a higher success rate.

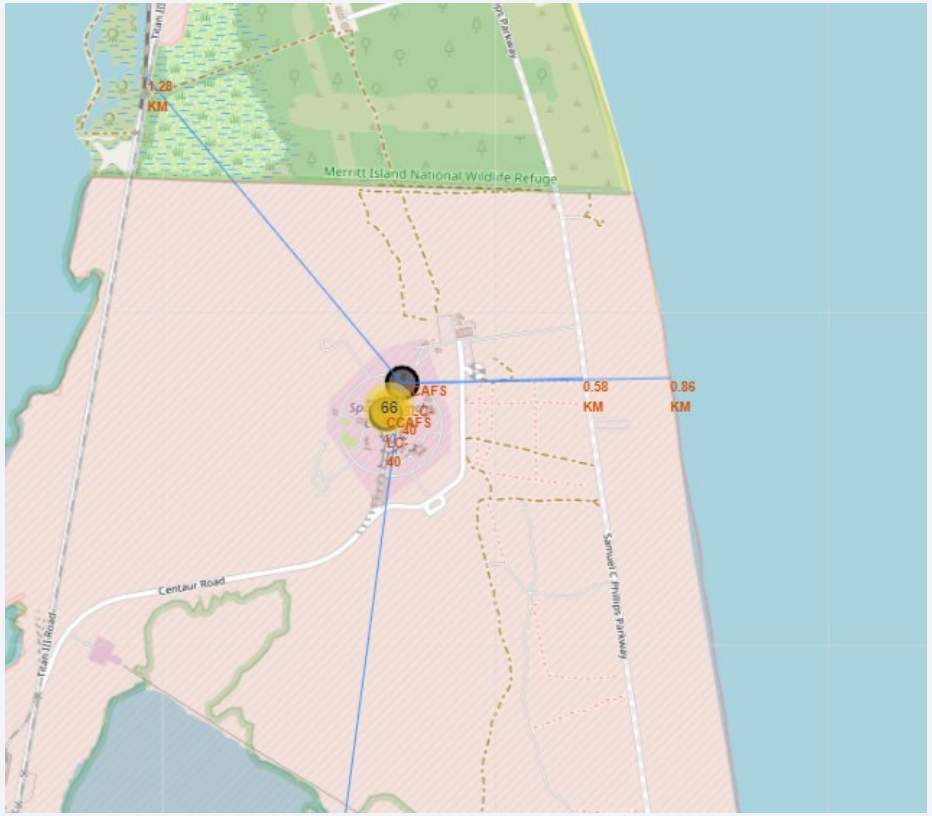


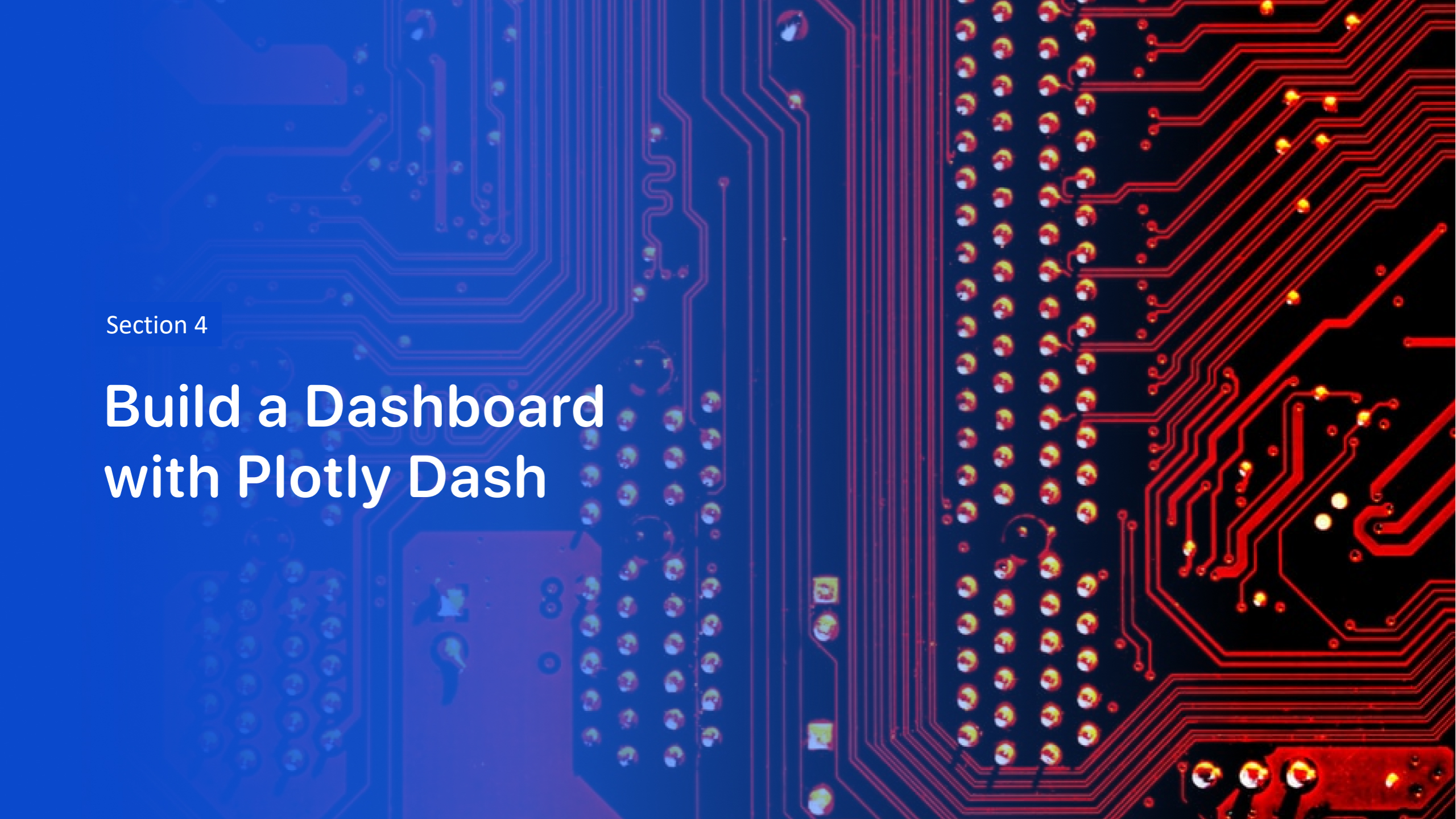
<Folium Map Screenshot 3>

The screenshot of the folium map highlights the following important elements and findings regarding the launch site and its proximities:

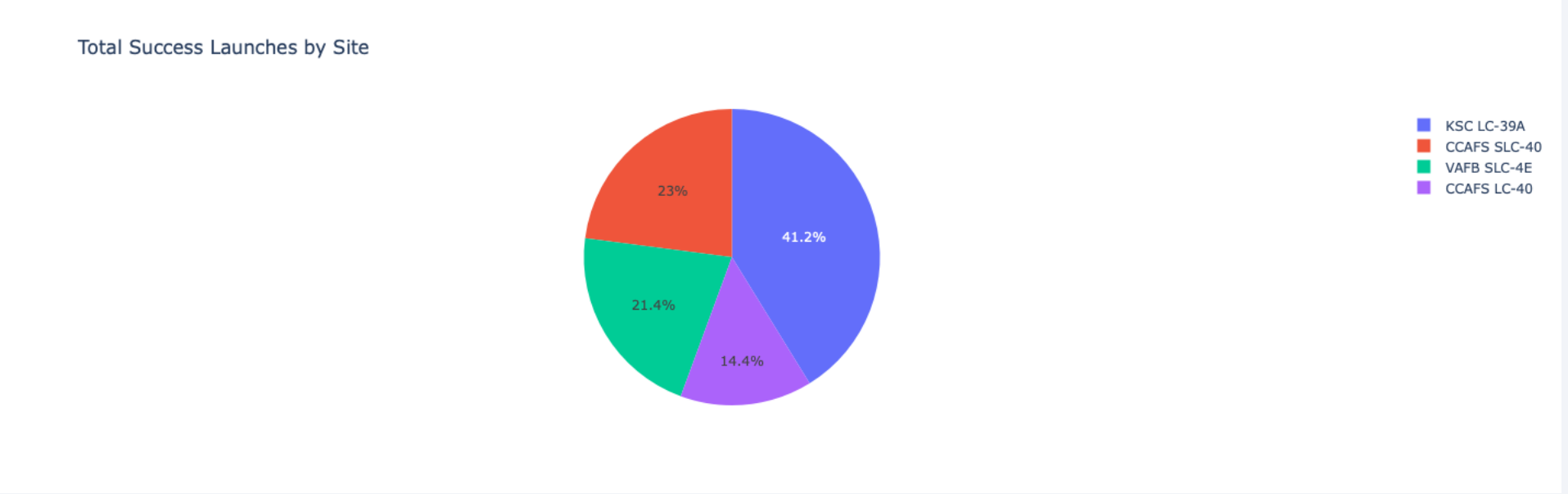
- Launch Site: The central point marked as "CCAFS" (Cape Canaveral Air Force Station) is the location of the launch site.
- Proximities:
    - Highway: The Samuel C. Phillips Parkway is shown with a calculated distance of approximately 0.58 km from the launch site.
    - Railroad: The railway is located about 1.28 km from the launch site, providing logistical support for transportation.
    - Coastline: The map shows the proximity to the coastline, which is important for launches over water.
    - Distance to City: The nearest city is approximately 51.43 km away, indicating a significant buffer zone for safety and noise reduction.

These elements are crucial for understanding the infrastructure and environmental context of the launch site, impacting logistics, safety protocols, and environmental management.




Section 4
Build a Dashboard with Plotly Dash

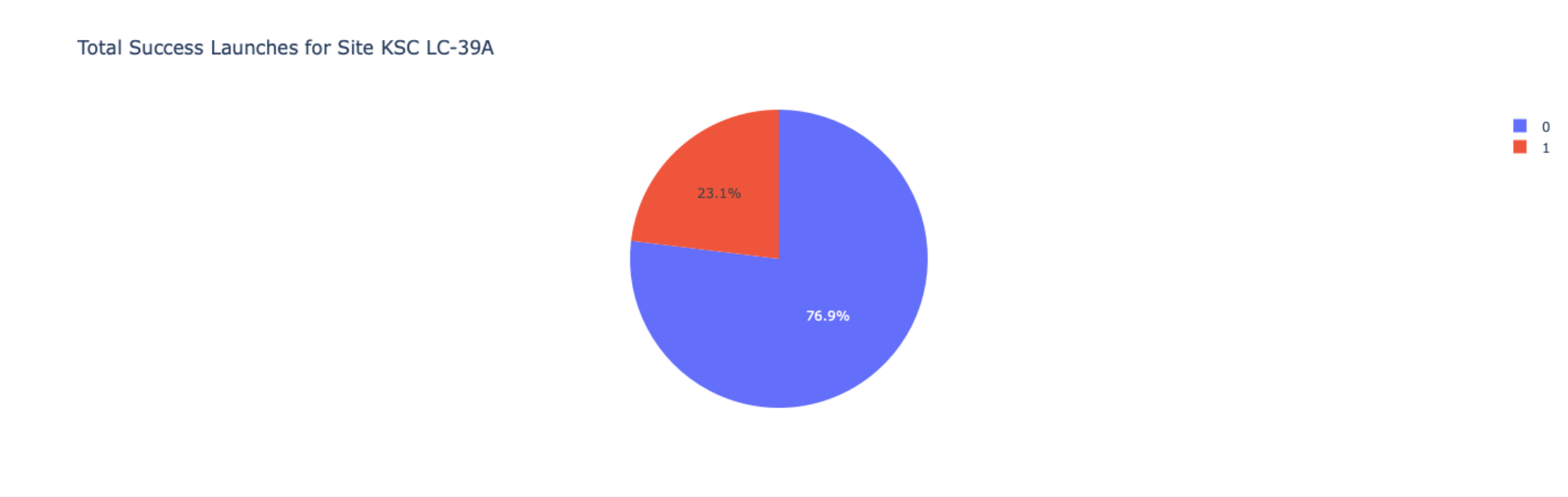
Total Success Launches by Site



- KSC LC-39A is the most frequently used site for successful launches, making up 41.2% of the total.
- The other sites contribute significantly but less than KSC LC-39A, with CCAFS SLC-40 and VAFB SLC-4E having similar contributions.

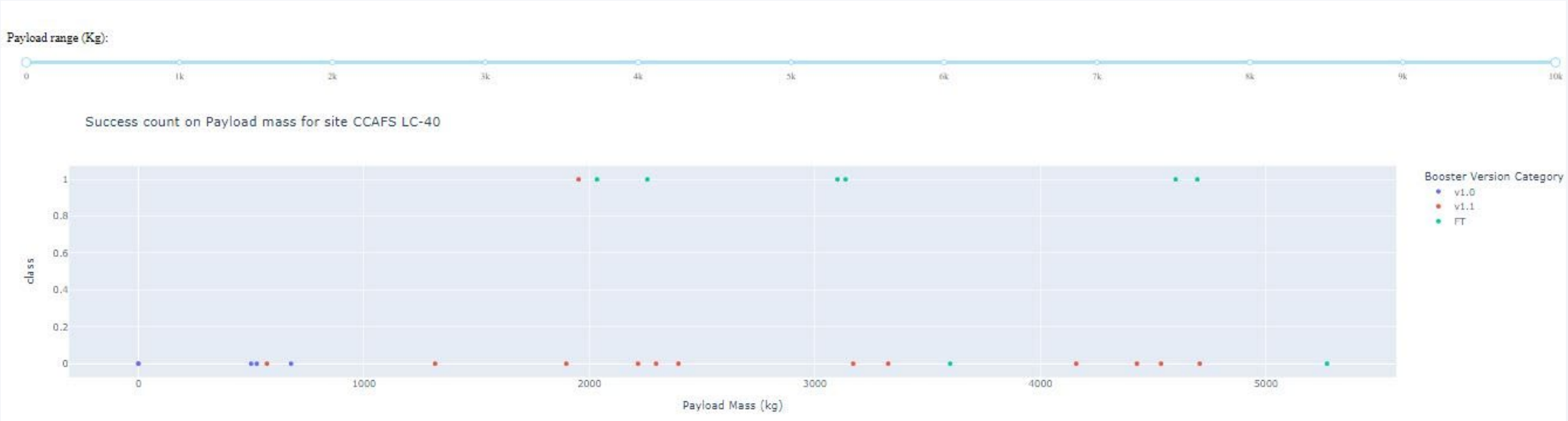
This distribution can help in understanding which sites are more active or preferred for successful launches, potentially guiding future planning and resource allocation.

Launch Success Distribution for KSC LC-39A



This analysis provides a clear overview of the success rates and highlights the efficiency of launches from KSC LC-39A.

Payload vs. Launch Outcome Scatter Plot for All Sites



These visualizations highlight which booster versions and payload ranges are most successful, providing valuable insights for optimizing future launches.

Section 5

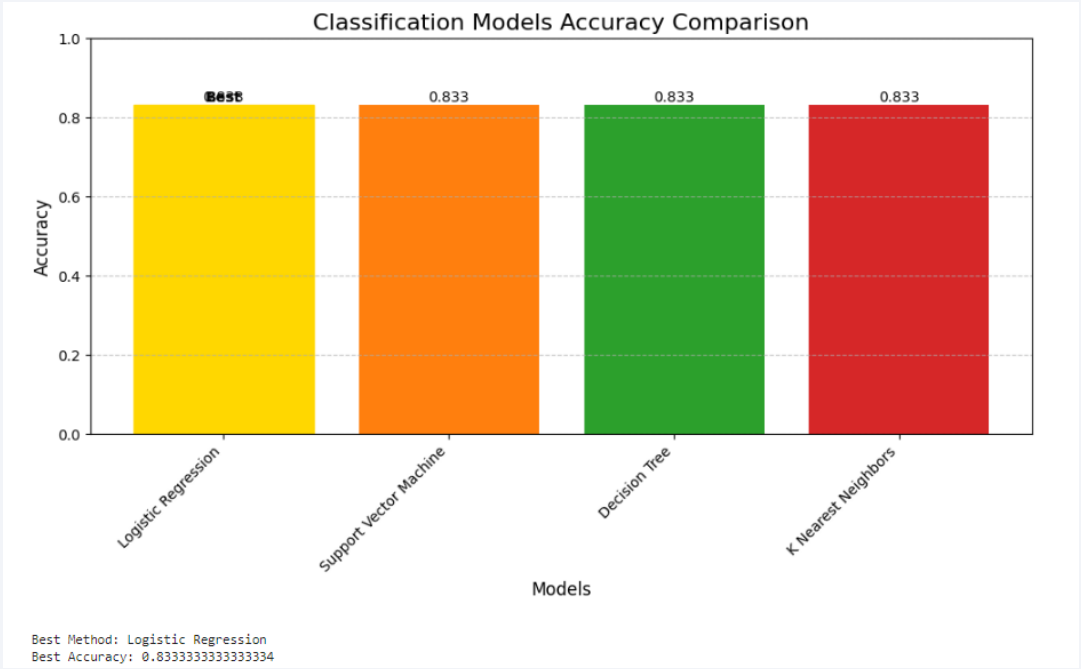
Predictive Analysis (Classification)

Classification Accuracy

The bar chart compares the accuracy of four classification models: Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors. All models have an accuracy of 0.833. However, Logistic Regression is highlighted as the best method, possibly due to other factors like simplicity or interpretability.

- Summary:
- Logistic Regression: 0.833 (Best Method)
- Support Vector Machine: 0.833
- Decision Tree: 0.833
- K Nearest Neighbors: 0.833

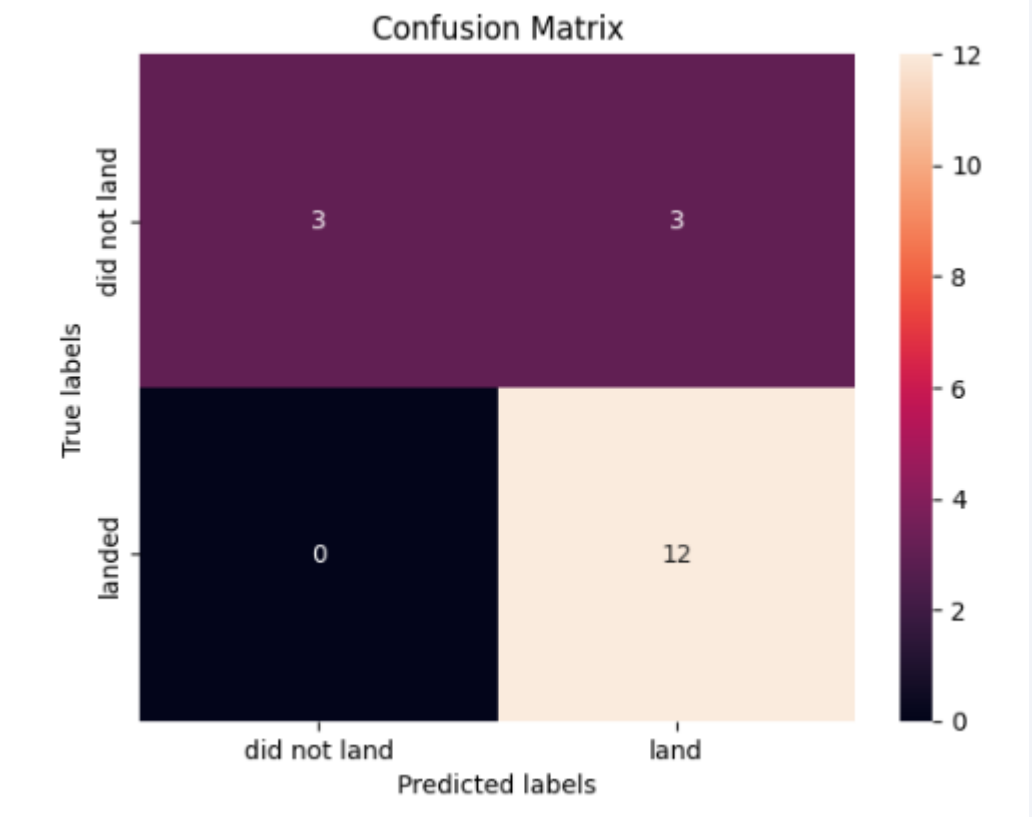
All models have the same accuracy of 0.833, with Logistic Regression noted as the best method in this context.



Confusion Matrix

The confusion matrix shown in the image provides a summary of the performance of a classification model. It compares the predicted labels with the true labels for a binary classification problem,

This matrix indicates that the model performs well, especially in predicting when something will land, with perfect recall but slightly lower precision due to some false positives.



Conclusions

Successfully developed a predictive model for SpaceX Falcon 9 first stage landings

- Key findings:
  - Launch site and orbit type significantly influence landing success
  - Payload mass shows correlation with landing outcomes
  - Success rates have improved over time across all launch sites
- Model performance:
  - All classification models achieved 83.3% accuracy
  - Logistic Regression identified as the best method, balancing accuracy and interpretability

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!