IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sreeharsha Sadhu
12/02/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Building an Interactive Map using Folium
  - Building a Dashboard using Plotly DASH
  - Predictive Analysis - Classification
- Summary of all results
  - EDA Results
  - Interactive Environment Sreenshots
  - Classification Model Results

Introduction

- Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the irst stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems you want to find answers
  - How is the success of the first stage landing affected by various factors such as payload mass, launch site, no. of flights, orbits, etc?
  - How has success rate varied over time?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Data Filtering
  - Handling Missing Values
  - Using Encoding to restructure data for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

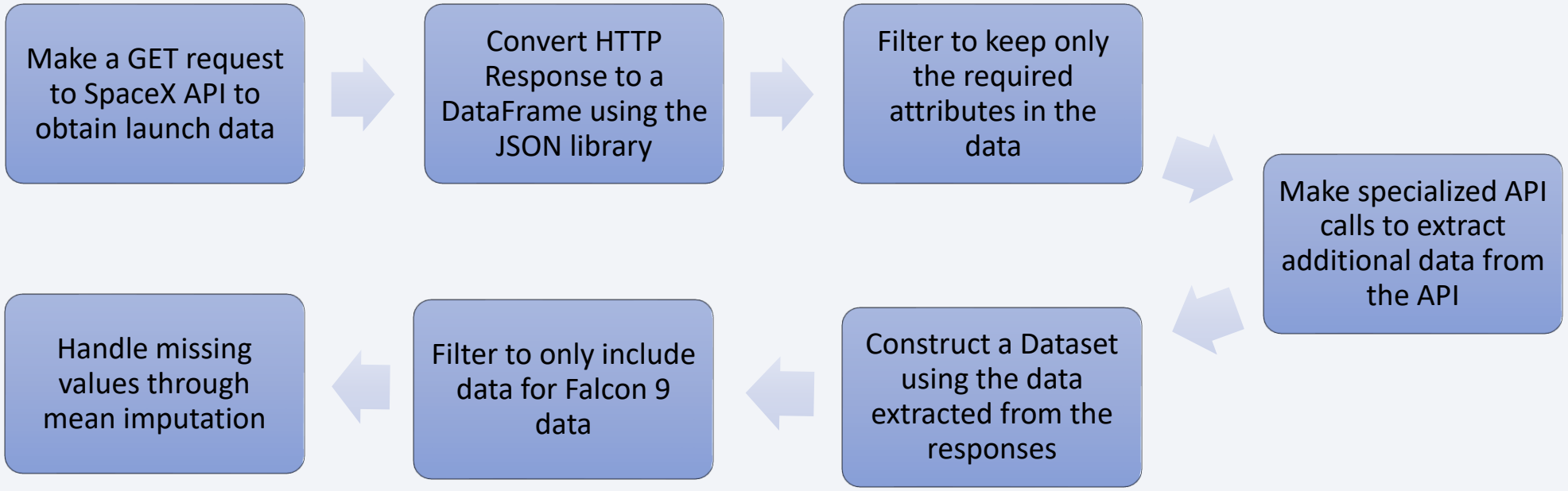
The data collection process involved utilizing a combination of two methods:
API requests from the SpaceX REST API and web scraping data from a table in SpaceX's Wikipedia entry.
These methods were employed together to ensure the collection of complete information on the launches, enabling a more detailed analysis.

From the SpaceX REST API, the following data columns were retrieved: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, and Latitude. These columns provided detailed technical data about each launch, including information about the rocket components, landing details, and geographical information.

In addition, data was obtained through web scraping from SpaceX's Wikipedia page. The columns collected via this method included Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, and Time. This data complemented the information from the API, offering insights such as customer details and the specific timing of the launches.

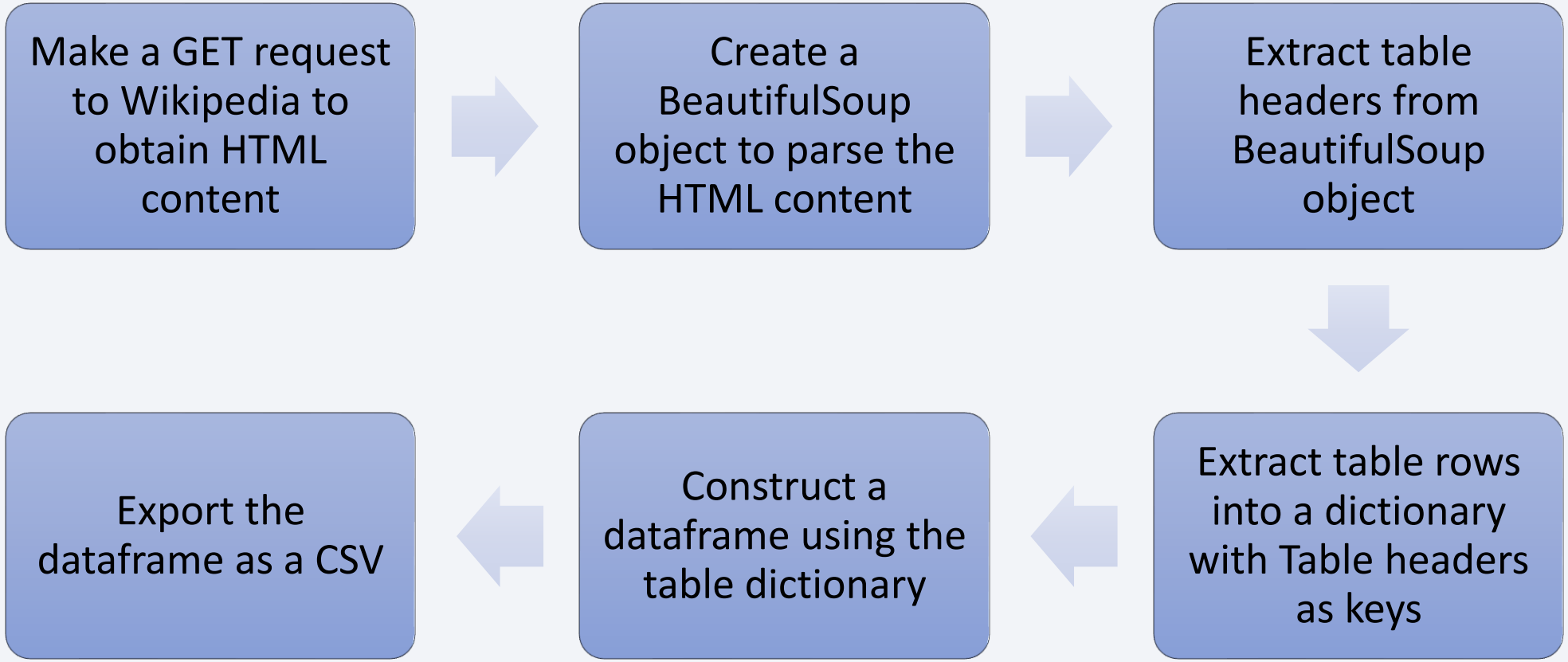
By combining data from both the SpaceX REST API and the Wikipedia web scraping, a comprehensive dataset was created for a more thorough analysis of SpaceX's launch history.

Data Collection – SpaceX API



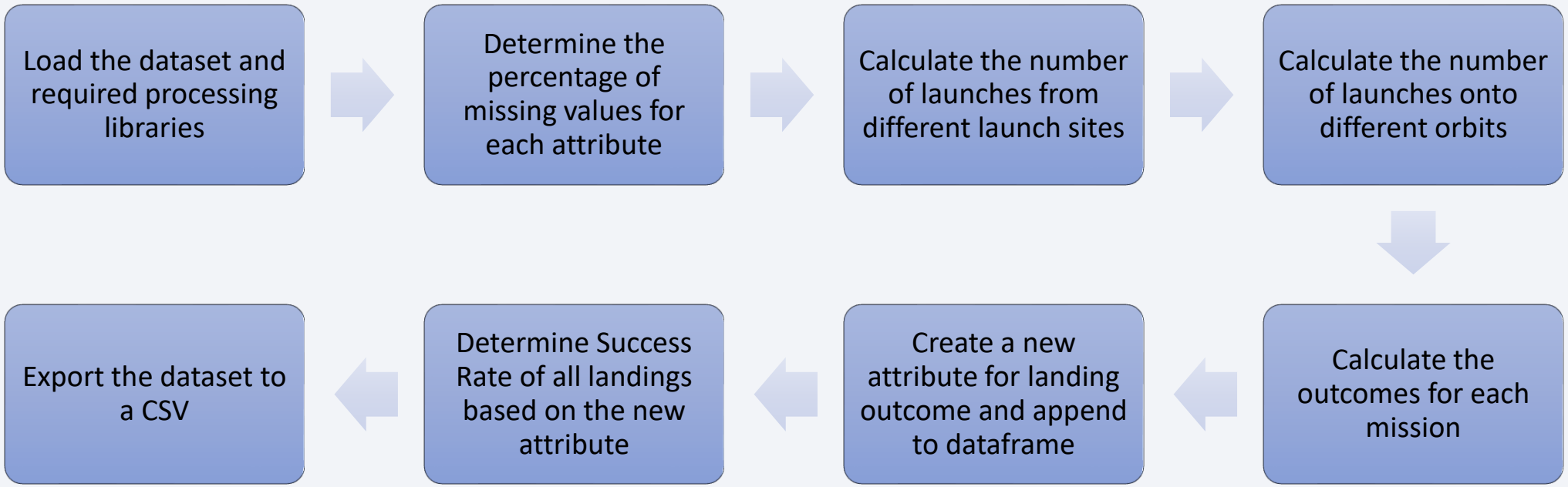
Github: [Data Collection using SpaceX API]

Data Collection - Scraping



Github: [Data Collection using Webscrapping]

Data Wrangling



Github: Data Wrangling

EDA with Data Visualization

1. Scatter Plots:

   Importance: Scatter plots are used to observe relationships between two continuous variables. They help in identifying patterns, trends, and potential correlations.

   Graphs:

   - FlightNumber vs. PayloadMass with Class as hue.
   - FlightNumber vs. LaunchSite with Class as hue.
   - PayloadMass vs. LaunchSite with Class as hue.
   - FlightNumber vs. Orbit with Class as hue.
   - PayloadMass vs. Orbit with Class as hue.

Github: EDA with Data Visualization

EDA with Data Visualization

2. Bar Charts:
    - Importance: Bar charts are used to compare categorical data. They are useful for showing the distribution of data across different categories.
    - Graphs:
        - Success rate of each orbit type.
3. Line Charts:
    - Importance: Line charts are used to display trends over time. They are useful for showing how a variable changes at regular intervals.
    - Graphs:
        - Yearly trend of launch success rates.

Github: EDA with Data Visualization

EDA with SQL

SQL Queries Performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- Display the date when the first successful landing outcome in ground pad was achieved
- 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 total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Github: EDA with SQL

Build an Interactive Map with Folium

1. Markers:
   - Launch Sites: Added markers for each launch site to indicate their locations on the map.
   - Launch Outcomes: Added markers for each launch with different colors (green for success and red for failure) to indicate the outcome of each launch.
2. Circles:
   - Launch Sites: Added Circles around each launch site to highlight their locations.
3. Lines:
   - Proximity Lines: Added lines between launch sites and their proximities (e.g., railway, highway, coastline, city) to visualize the distances.

Github: Interactive Map with Folium

Build a Dashboard with Plotly Dash

1. Dropdown List:
   Launch Site Selection: A dropdown list that allows users to select a specific launch site or view data for all sites. This interaction helps users filter the data based on the launch site of interest.

2. Pie Chart:
   Success Counts: A pie chart that displays the total successful launches count for all sites or the success vs. failed counts for a selected site. This plot provides a visual summary of launch outcomes, making it easy to compare the success rates across different sites.

3. Range Slider:
   Payload Range Selection: A range slider that allows users to select a range of payload masses. This interaction helps users filter the data based on payload mass, enabling analysis of how payload mass affects launch success.

4.Scatter Chart:
   Payload vs. Success Correlation: A scatter chart that shows the correlation between payload mass and launch success. This plot helps users visualize the relationship between payload mass and the likelihood of a successful launch, providing insights into the impact of payload mass on launch outcomes.

Github: Plotly Dash App

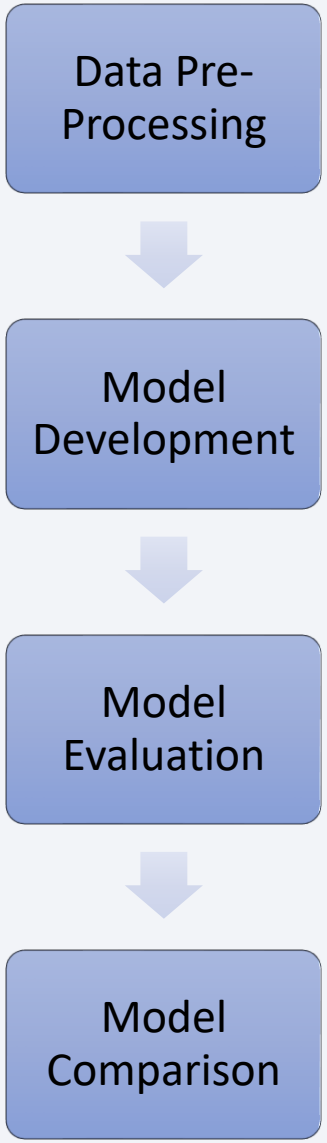
Predictive Analysis (Classification)

Data Pre-Processing:

- Load Data: Imported the dataset
- Handle Missing Values: Imputed null values with mean of the attribute.
- Standardize Data: Used StandardScaler to standardize the features.
- Data Split: Split the total data into Train and Test sets for model training and testing

Model Building:

- Model Selection: Trained different ML models that include Logistic Regression, Support Vector Machine, Decision Tree, and k-Nearest Neighbors
- Hyperparameter Tuning: Utilizied GridSearchCV to find the best hyperparameters for each of the models used



Github: Predictive Analysis

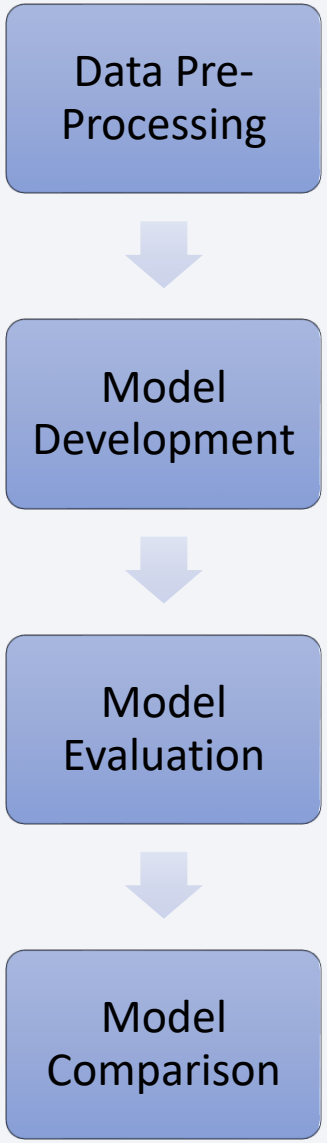
Predictive Analysis (Classification)

Model Evaluation:

- Accuracy: Evaluated the accuracy of each model on the test data.
- Confusion Matrix: Plotted confusion matrices to visualize the performance of each model.
- Jaccard Score and F1 Score: Calculated Jaccard and F1 scores for each model.

Model Comparison:

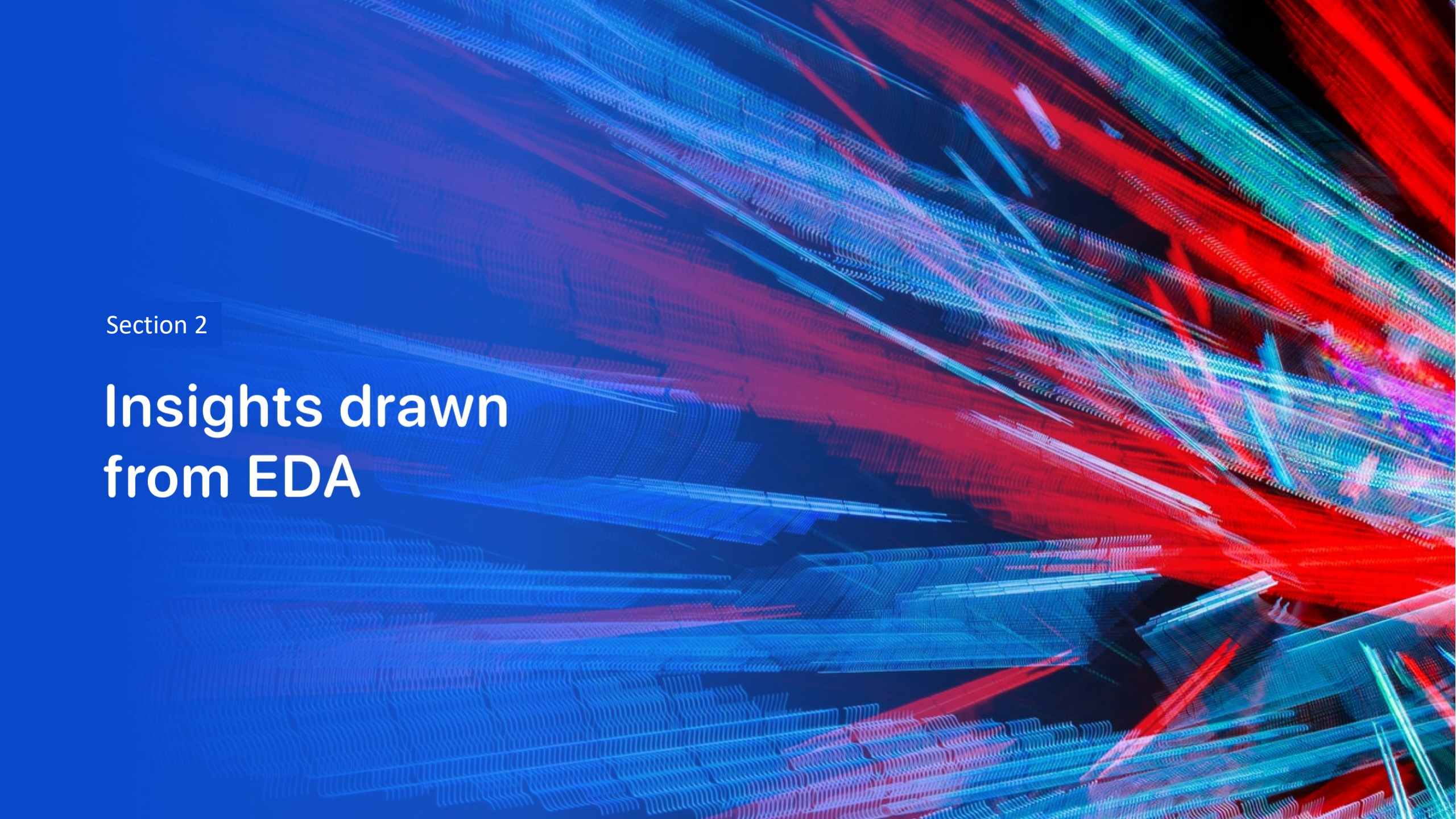
- Compare Scores: Compared the accuracy, Jaccard score, and F1 score of all models.
- Best Model Selection: Selected the model with the highest scores as the best performing model.



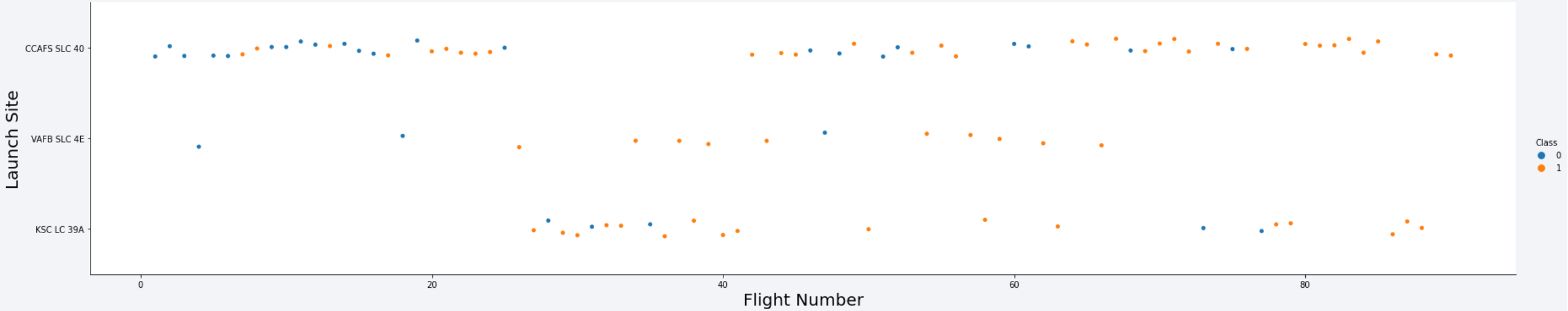
Github: [Predictive Analysis]

Results

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


Section 2
Insights drawn from EDA

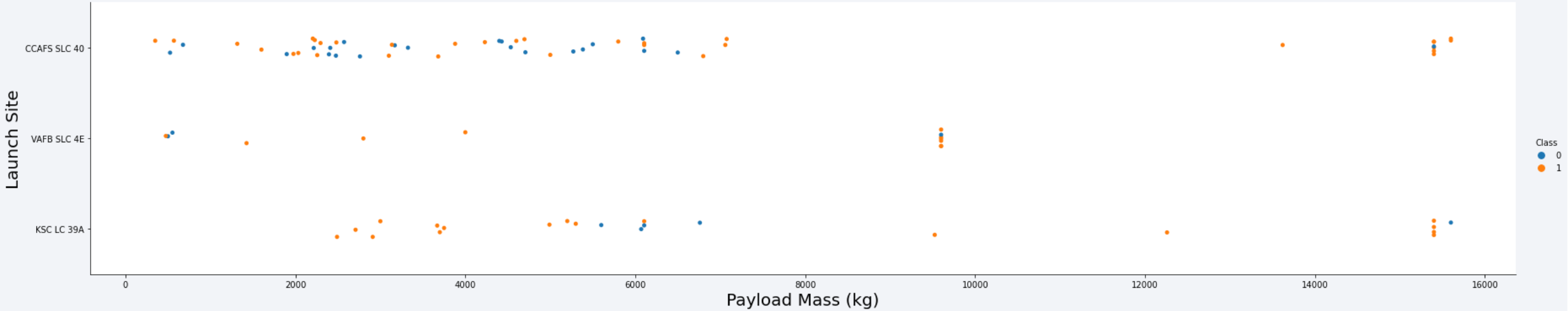
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successfull.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

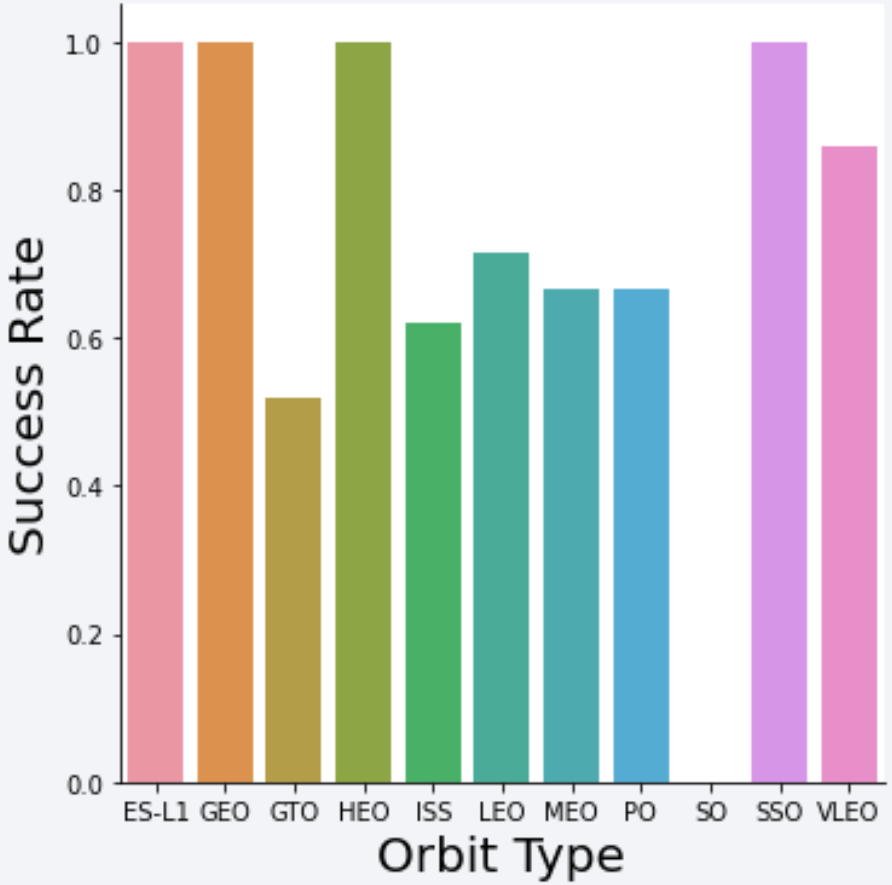
Success Rate vs. Orbit Type

Explanation:

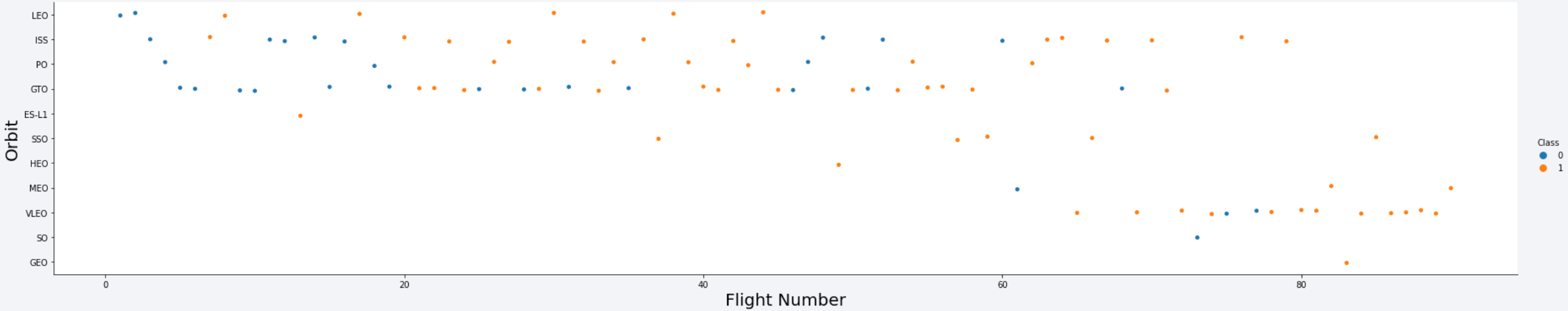
- Orbits with 100% success rate are:
  - ES-L1
  - GEO
  - HEO
  - SSO

- Orbits with 0% success rate are:
  - SO

- Orbits with success rate between 50% and 85%:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO

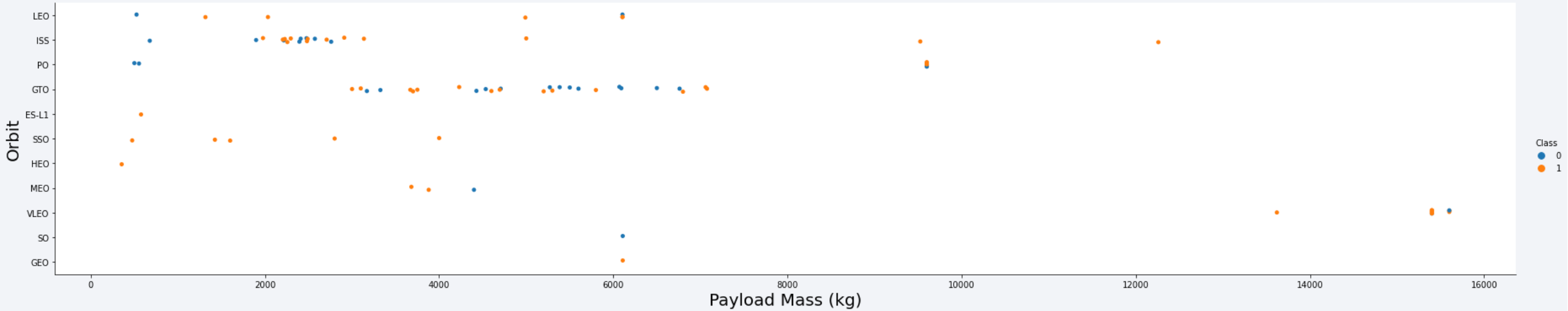


Flight Number vs. Orbit Type



Explanation:
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

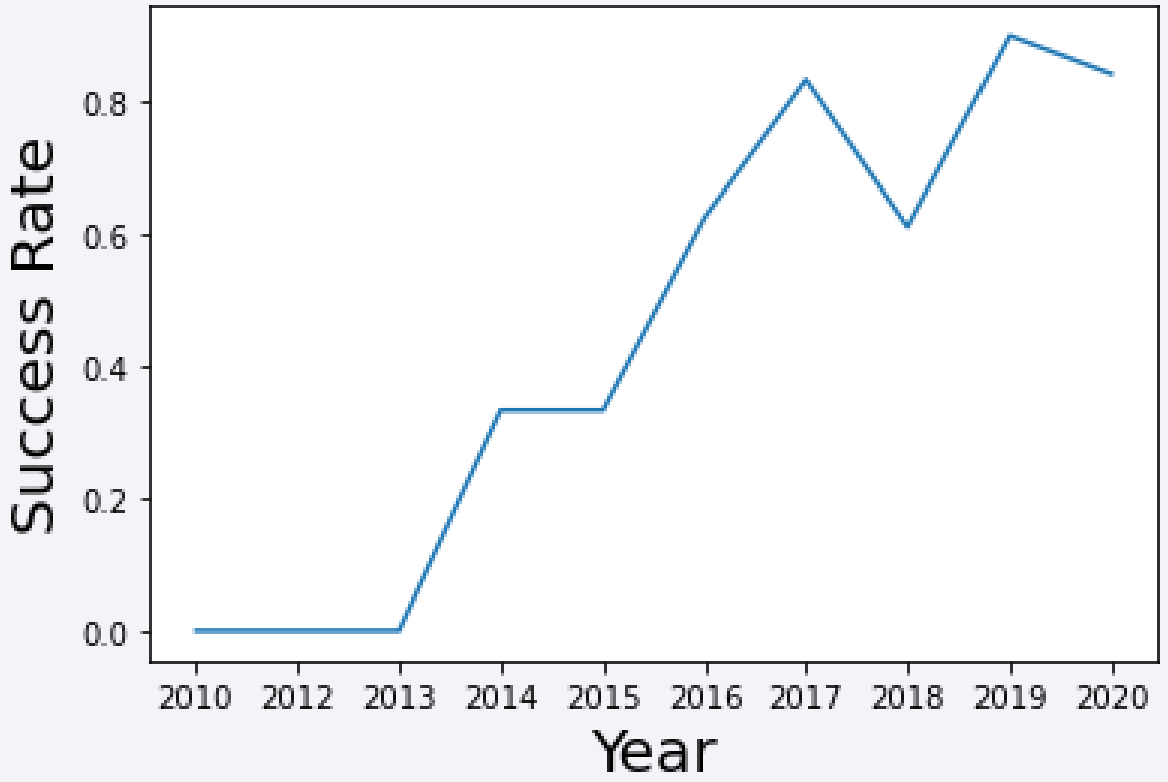
Payload vs. Orbit Type



Explanation:
Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

Explanation:
It can be observed that the success rate kept increasing from 2013 to 2020. There was a small drop in the success rate during 2017-2018



All Launch Site Names

SELECT DISTINCT launch_site FROM SPACEXDATASET

Explanation:
Displays the names of the unique launch sites in the space mission.

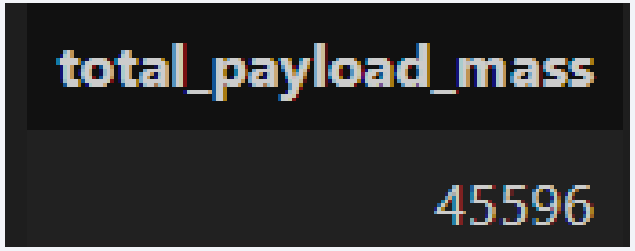
Launch Site Names Begin with 'CCA'

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' LIMIT 5;

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

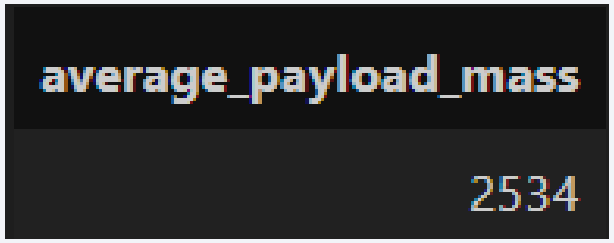
Total Payload Mass




```
SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEXDATASET WHERE customer = 'NASA (CRS)';
```

Explanation:
Displays the total payload mass carried by boosters launched by NASA (CRS).

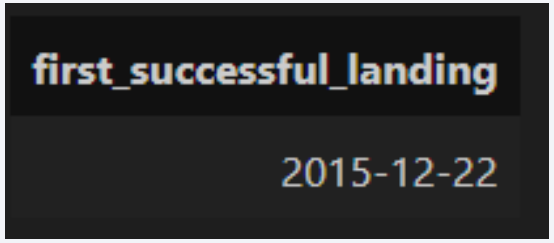
Average Payload Mass by F9 v1.1



SELECT AVG(payload_mass__kg_) AS average_payload_mass FROM SPACEXDATASET WHERE booster_version LIKE '%F9 v1.1%';

Explanation:
Displays average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date




```
SELECT MIN(date) AS first_successful_landing FROM SPACEXDATASET WHERE landing__outcome = 'Success (ground pad)';
```

Explanation:
Lists the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT booster_version FROM SPACEXDATASET WHERE landing__outcome =
'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

SELECT mission_outcome, COUNT(*) AS total_number FROM SPACEXDATASET GROUP BY mission_outcome;

Explanation:
Lists the total number of successful and failure mission outcomes.

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
SELECT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_ =
(SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET);
```

Explanation:
Lists the names of the booster_versions which have carried the maximum payload mass. Use a subquery.

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

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

```
SELECT monthname(date) AS month, date, booster_version, launch_site, landing__outcome FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND year(date) = 2015;
```

Explanation:
Lists the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015.

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

```
SELECT landing__outcome, COUNT(*) AS count_outcomes FROM
SPACEXDATASET
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY count_outcomes DESC;
```

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

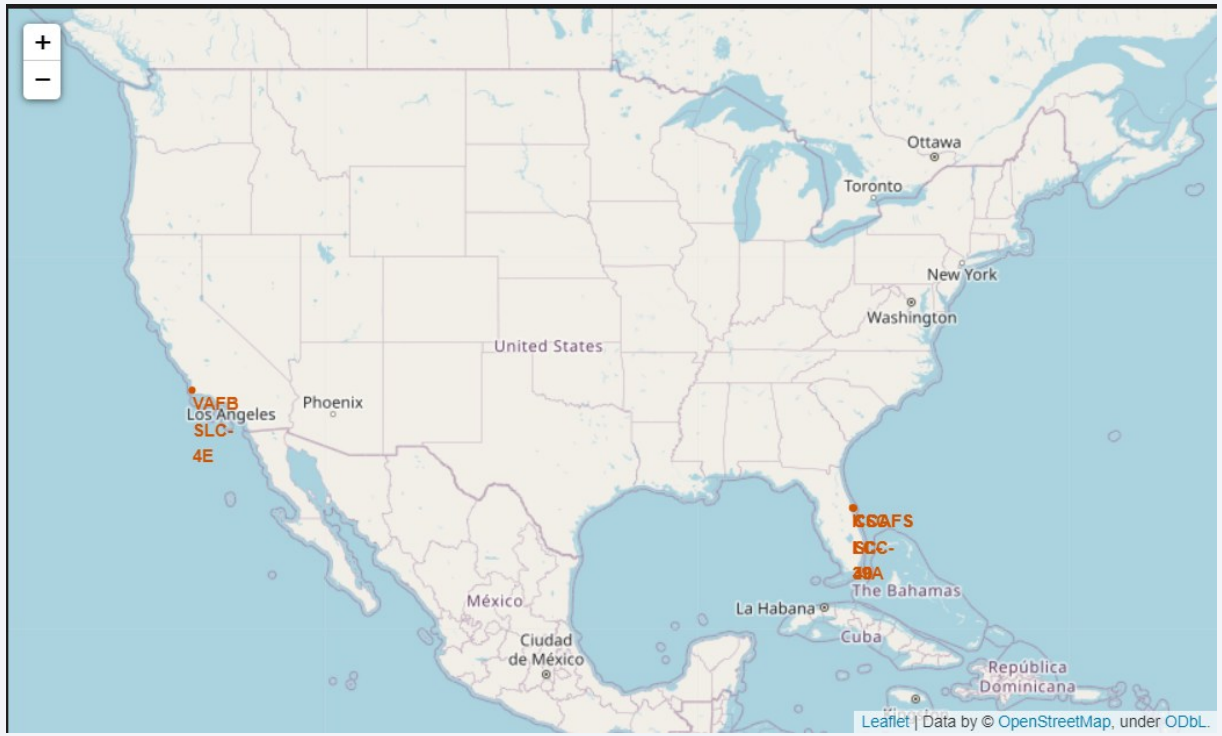
Explanation:
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.


Section 3
Launch Sites
Proximities Analysis

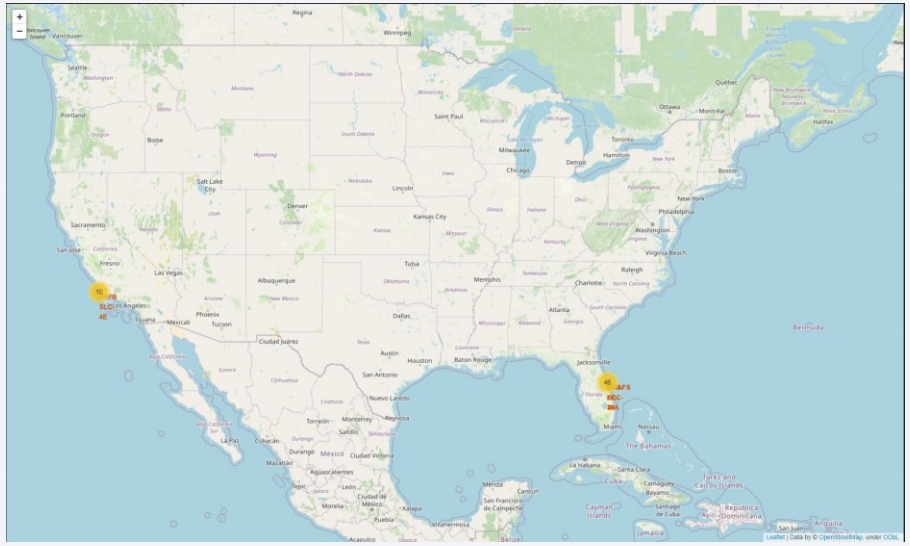
Complete Launch Site Markers

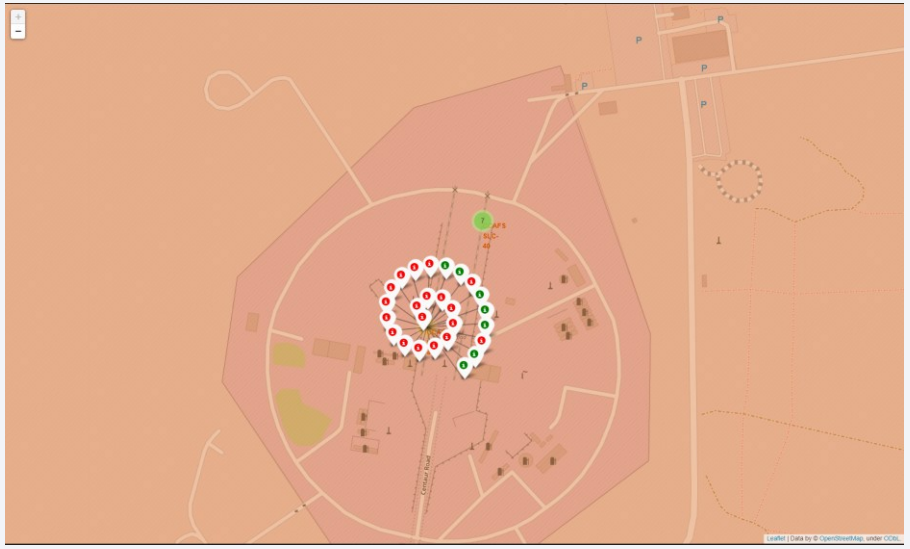
Description: This map includes markers for all SpaceX launch sites using their latitude and longitude coordinates.

Purpose: To visualize the locations of all SpaceX launch sites on a map.



Launch Outcome Markers





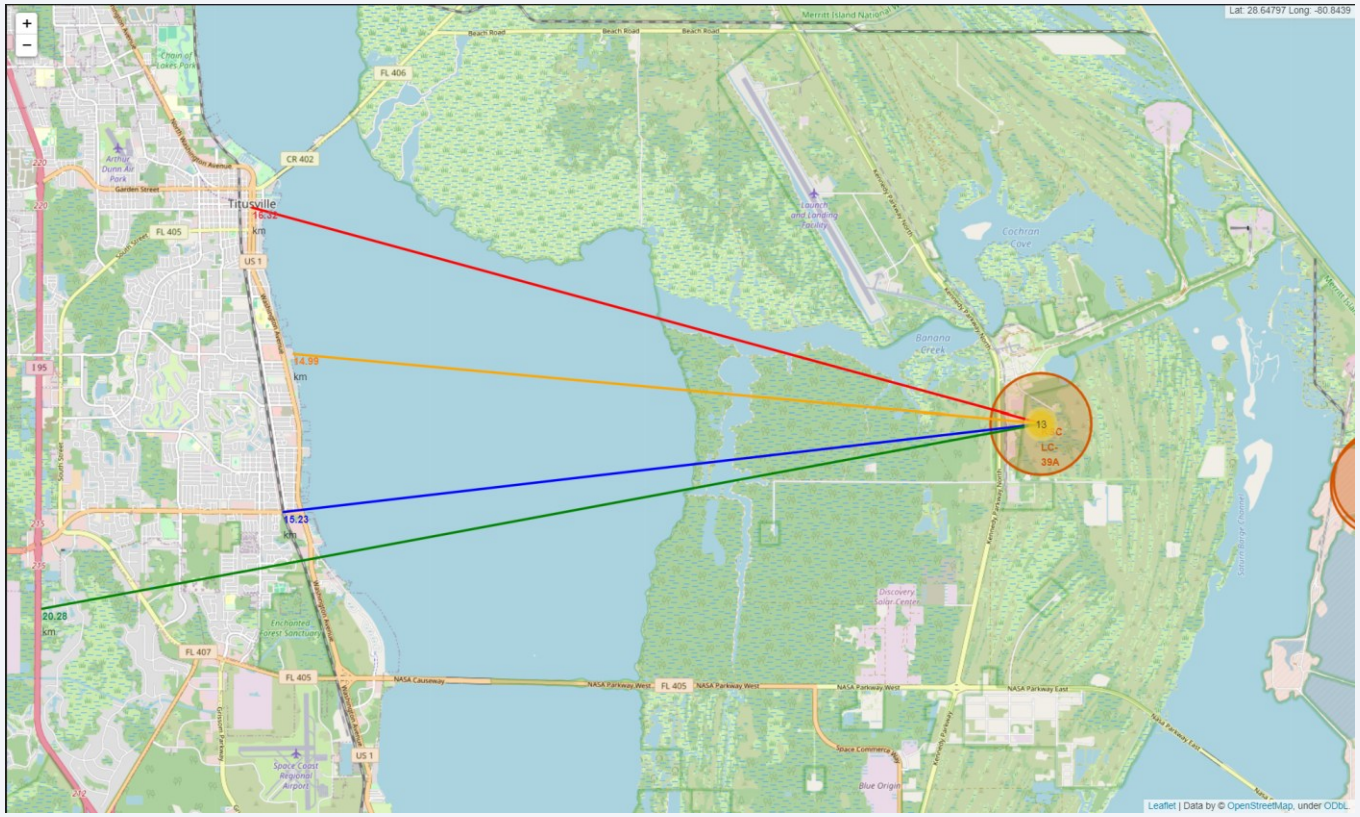
Description: This map enhances the previous map by adding markers for each launch outcome. Green markers indicate successful launches, and red markers indicate failed launches.

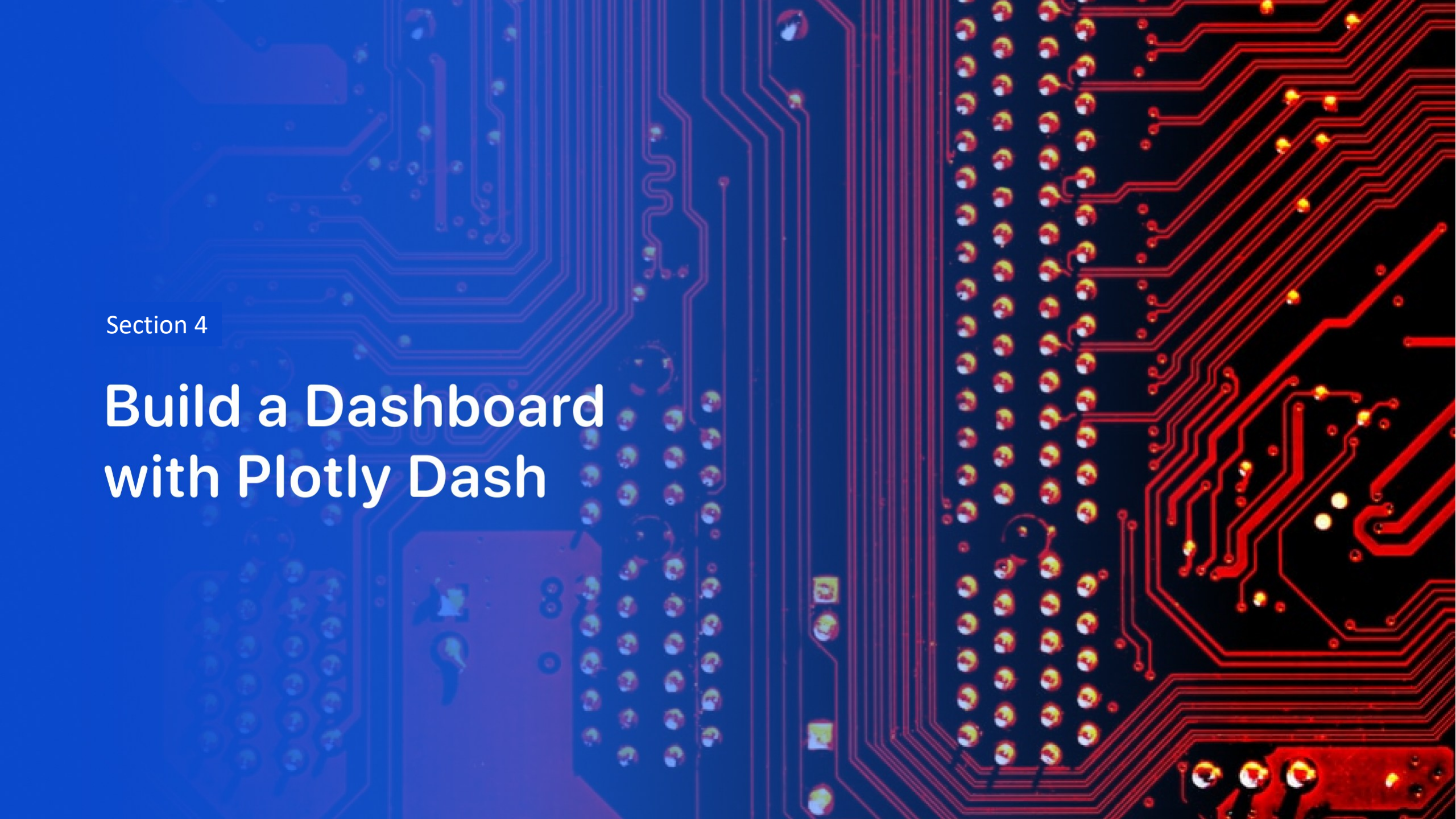
Purpose: To visualize the success and failure rates of launches at each site.

Distance to Nearest Features

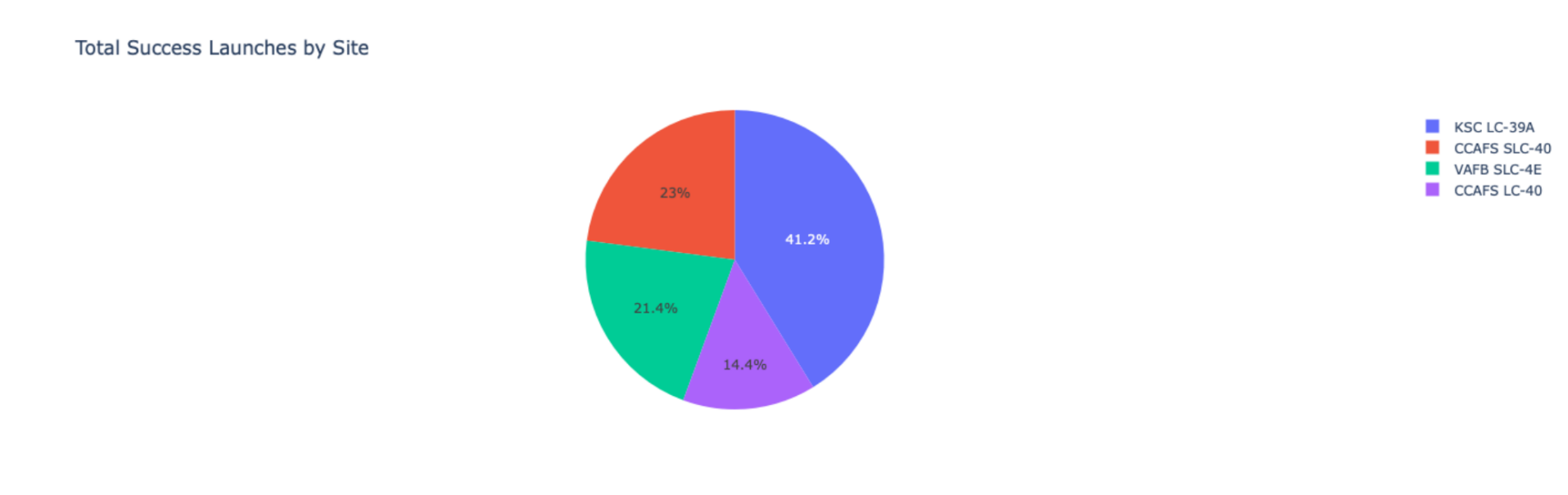
Description: This map includes markers and lines showing the distances from a launch site to the nearest city, coastline, and highway.

Purpose: To calculate and visualize the proximity of launch sites to cities, coastlines, and highways.



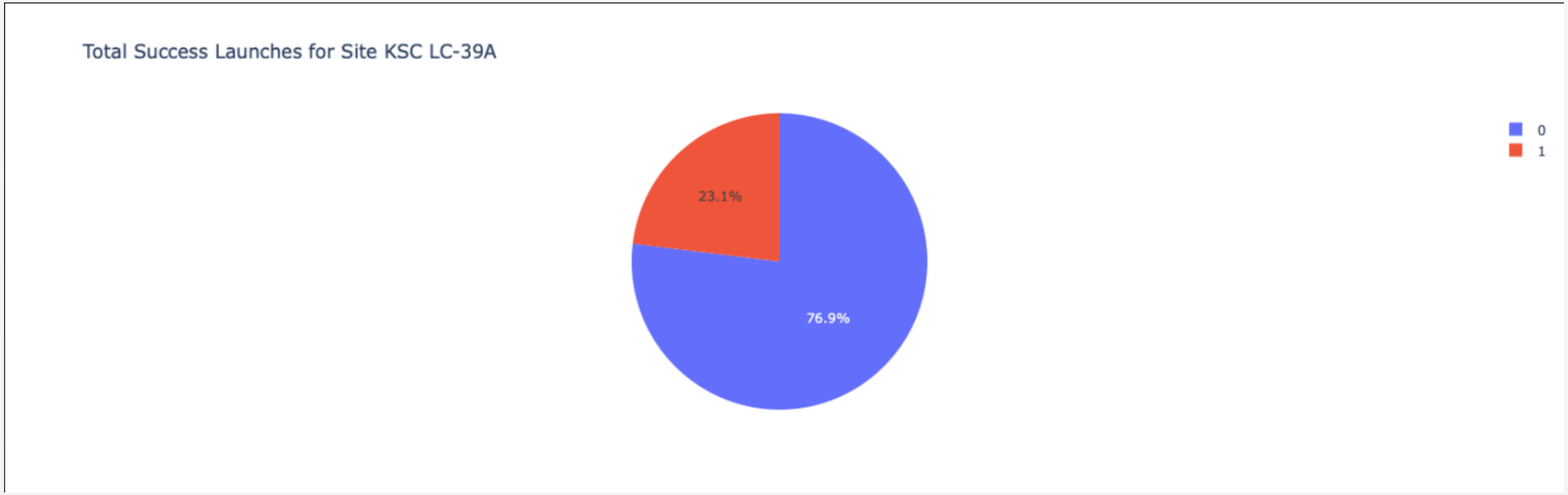

Section 4
Build a Dashboard with Plotly Dash

Launch Success Rate of All Sites



Explanation:
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch Site with Highest Launch Success Rate



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

<Dashboard Screenshot 3>

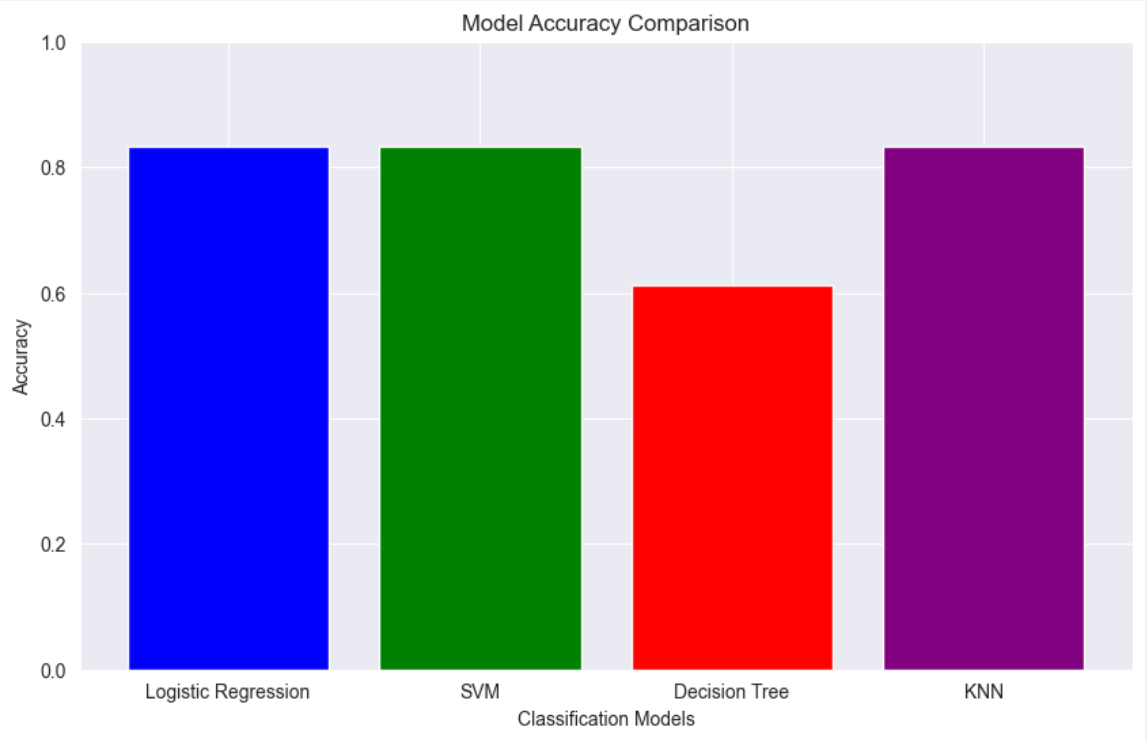
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

The model with the highest accuracy is Logistic Regression with an accuracy of 0.83

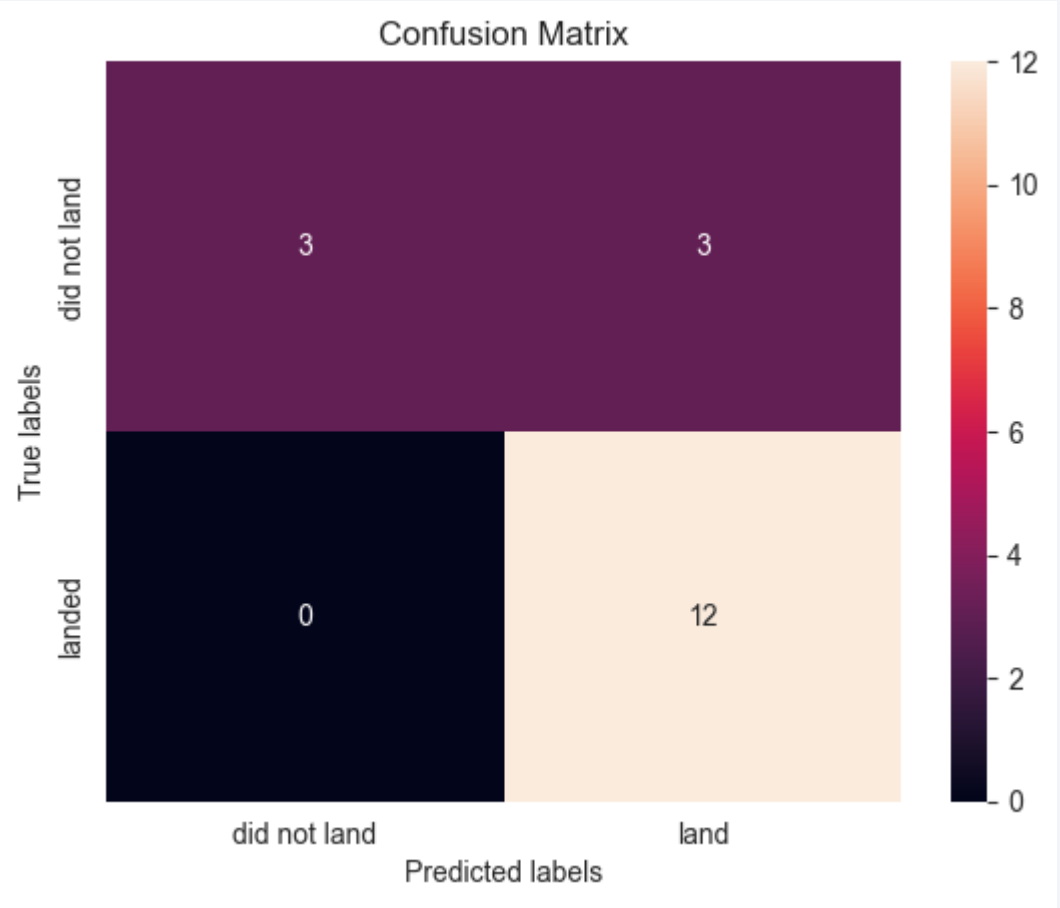


Confusion Matrix

True Labels: The actual labels from the test dataset.

Predicted Labels: The labels predicted by the Decision Tree model.

Confusion Matrix: A table used to describe the performance of a classification model. It shows the number of correct and incorrect predictions made by the model compared to the actual outcomes (true labels).



Conclusions

- Logistic Regression Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!