IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ROSTYSLAV LEVCHENKO

08/2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

Project Background and Context: This project focuses on SpaceX, a prominent aerospace company founded by Elon Musk. We aim to analyze data related to SpaceX launches, exploring their success rates, payload trends, rocket performance, landing outcomes, launch site choices, and temporal patterns.

Problems :

Launch Success: What factors contribute to the success of SpaceX launches?
Payload Analysis: How has payload mass and types evolved over time, and do they correlate with launch outcomes?
Rocket Performance: What can we learn about the reliability and reusability of SpaceX's rockets?
Landing Outcomes: What patterns exist in the outcomes of first stage landings?
Launch Site Assessment: How do launch site choices affect mission success?
Temporal Trends: Are there any notable trends or seasonality in SpaceX's launch history?

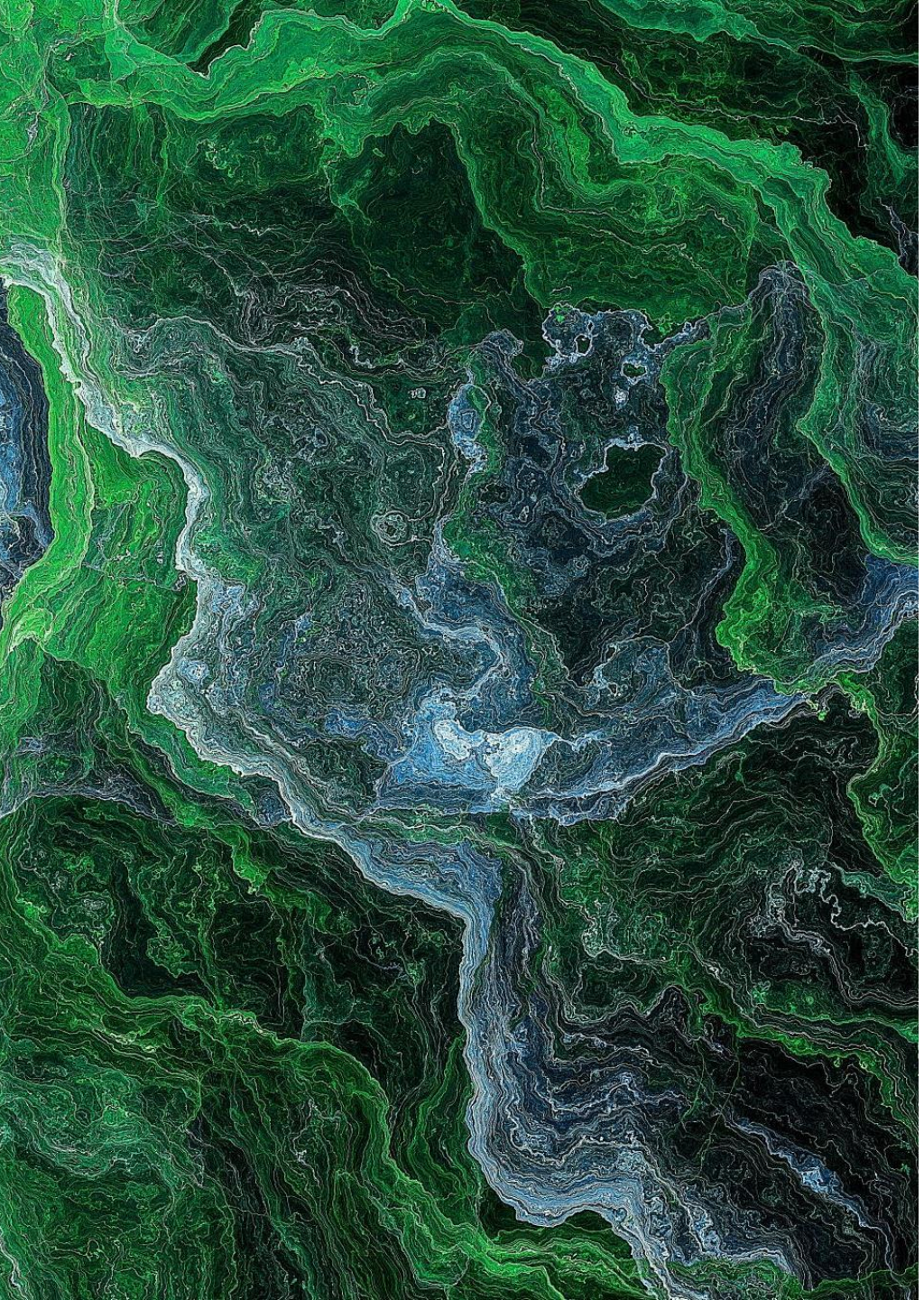
Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- 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
  - How to build, tune, evaluate classification models

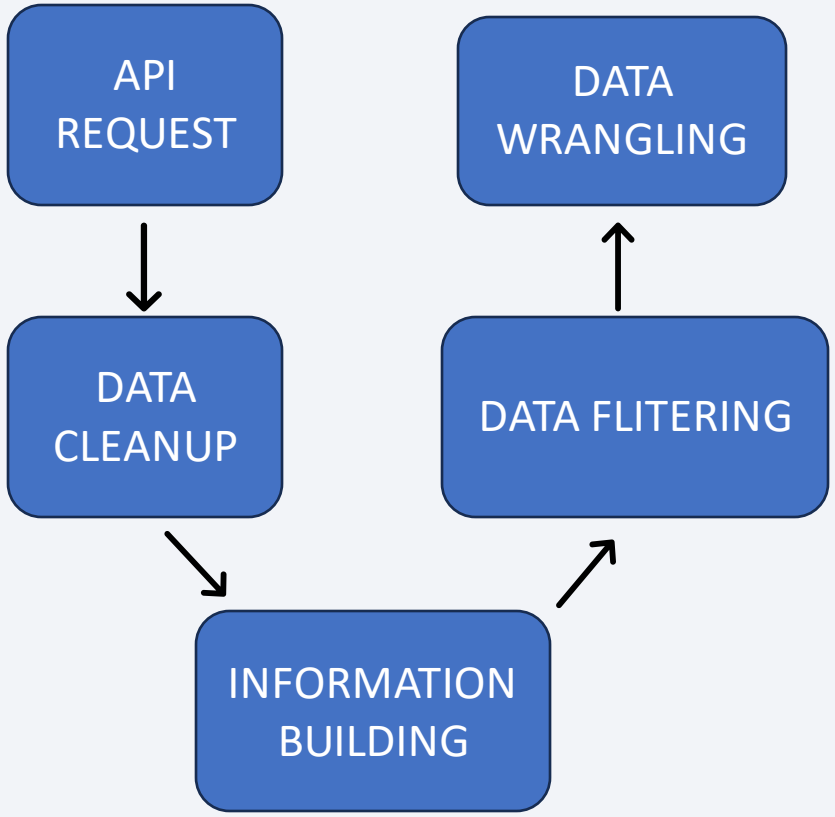


Data Collection

Initial Setup:	• Identify data source: SpaceX official website. • Set up web scraping tools and environment.
Website Access:	• Use web scraping libraries to access SpaceX's website.
Page Navigation:	• Navigate through web pages to locate relevant data. • Access SpaceX launch history page.
Data Extraction:	• Extract data from web pages.
Data Cleaning:	• Clean extracted data to remove inconsistencies and errors. • Check for missing values.
Data Storage:	• Export processed data into a structured format like CSV.
Analysis:	• Utilize collected data for various analyses, e.g., predicting Falcon 9 first stage landings.

Data Collection – SpaceX API

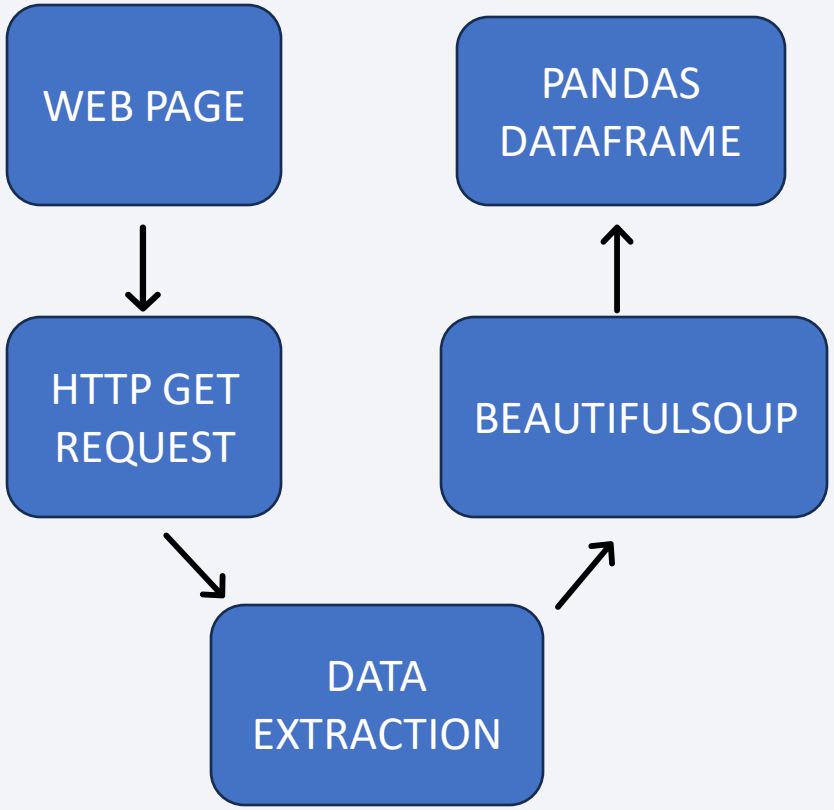
- **API Request:** Utilized SpaceX API to gather launch data.
- **Data Cleanup:** Formatted and cleaned the obtained data.
- **Information Extracted:**
  - Booster Version
  - Payload Mass & Orbit
  - Launch Site (Longitude & Latitude)
  - Landing Outcomes & Types
  - Flight Counts
  - Gridfins, Reuse, and Legs
  - Landing Pads
  - Core Blocks, Reuse Counts, and Serials
- **Data Filtering:** Retained only Falcon 9 launches.
- **Data Wrangling:** Imputed missing Payload Mass values with the mean.



Github - Link

Data Collection - Scraping

- **Objective:** Web scrape Falcon 9 launch records from Wikipedia.
- **Tools Used:** BeautifulSoup, Requests, Pandas.
- **Step 1:** Request the Falcon 9 Launch Wiki page.
- **Step 2:** Create a BeautifulSoup object.
- **Step 3:** Extract column names from the HTML table header.
- **Step 4:** Create a dictionary with column names.
- **Step 5:** Parse and populate the dictionary with launch records.
- **Step 6:** Convert the dictionary into a Pandas dataframe.
- **Step 7:** Export the dataframe to a CSV file.



Github - Link

Data Wrangling

- **Data Import:**
  - Imported data from a CSV file containing SpaceX Falcon 9 launch data.
- **Exploratory Data Analysis (EDA):**
  - Determined patterns and missing values.
  - Identified various attributes including launch site, orbit type, and landing outcomes.
- **Identifying Launch Site Data:**
  - Calculated the number of launches for each launch site.
  - Analyzed the launch site data to determine site-specific trends.
- **Analyzing Orbit Types:**
  - Examined the number and occurrence of each orbit type.
- **Determining Mission Outcomes:**
  - Identified mission outcomes (landing results) and their frequencies.
  - Created a set of unsuccessful outcomes (bad_outcomes).
- **Creating Classification Labels:**
  - Generated a classification variable (landing_class) where 0 represents unsuccessful landings and 1 represents successful landings.
- **Data Export:**
  - Exported the processed data to a CSV file (dataset_part_2.csv) for further analysis.
  - **Flowchart:**
- **Data Import ➞ EDA ➞ Identify Launch Site Data ➞ Analyze Orbit Types ➞ Determine Mission Outcomes ➞ Create Classification Labels ➞ Data Export**

Github - Link

EDA with Data Visualization

- **FlightNumber vs. PayloadMass Scatter Plot:**
  - Purpose: Explore the relationship between flight number and payload mass.
- **FlightNumber vs. Launch Site Scatter Plot:**
  - Purpose: Explore the relationship between flight number and launch site.
- **PayloadMass vs. Launch Site Scatter Plot:**
  - Purpose: Explore the relationship between payload mass and launch site.
- **Success Rate vs. Orbit Bar Chart:**
  - Purpose: Visualize the success rate of different orbit types.
- **FlightNumber vs. Orbit Scatter Plot:**
  - Purpose: Explore the relationship between flight number and orbit type.
- **PayloadMass vs. Orbit Scatter Plot:**
  - Purpose: Explore the relationship between payload mass and orbit type.
- **Launch Success Yearly Trend Line Chart:**
  - Purpose: Analyze the trend in average launch success rate over the years.

These charts were used for exploratory data analysis (EDA) and feature engineering to understand the SpaceX dataset better. They provide insights into factors influencing launch success, such as flight number, payload mass, launch site, orbit type, and launch success trends.

Github - Link

EDA with SQL

- **Unique Launch Sites:** Displayed unique launch site names.
- **Launch Sites Starting with 'CCA':** Displayed 5 records with launch sites starting with 'CCA'.
- **Total Payload Mass for NASA (CRS):** Calculated the total payload mass for NASA (CRS) missions.
- **Average Payload Mass for F9 v1.1:** Calculated the average payload mass for booster version F9 v1.1.
- **First Successful Ground Pad Landing Date:** Found the date of the first successful ground pad landing.
- **Boosters with Successful Drone Ship Landings:** Listed boosters that successfully landed on drone ships with a payload mass between 4000 and 6000.
- **Mission Outcomes:** Counted and listed the total number of successful and failed mission outcomes.
- **Max Payload Mass Booster Versions:** Identified booster versions with the maximum payload mass (using a subquery).
- **Failed Drone Ship Landings in 2015:** Listed records for failed drone ship landings in 2015, including month names.
- **Ranking Landing Outcomes:** Ranked landing outcomes between specific dates in descending order.

Github - Link

Build an Interactive Map with Folium

In the Folium map, I added the following map objects:

- **Markers:** These represent specific locations or points of interest on the map. Markers display information when clicked and are useful for highlighting important locations.
- **Polyline:** I used polylines to connect multiple points on the map, creating a visual path or route. This helps users understand connections between locations.
- **Marker Cluster:** To prevent map clutter, I grouped nearby markers into clusters. This enhances map readability, especially when multiple markers are close together.
- I added these objects for better visualization, connectivity representation, and improved user experience.

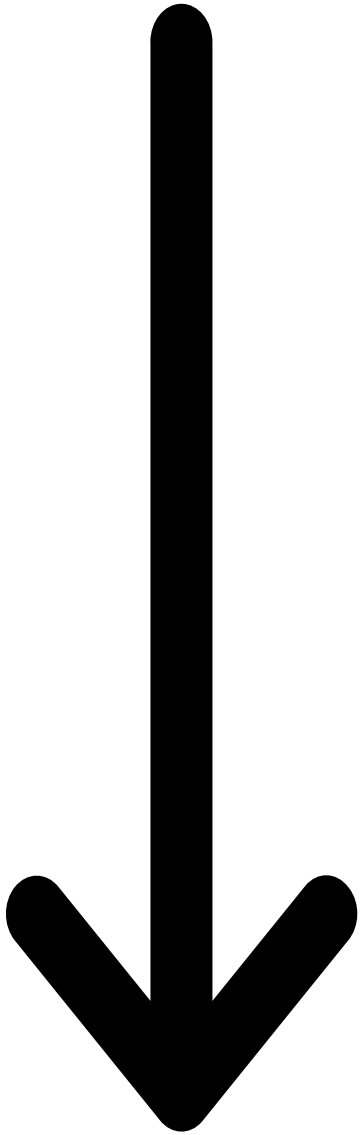
Github - Link

Build a Dashboard with Plotly Dash

- **Dropdown Selection for Statistics:**
  - Users can choose "Yearly Statistics" or "Recession Period Statistics."
- **Dropdown Selection for Year:**
  - Users can select a specific year for "Yearly Statistics."
- **Disabled Year Dropdown:**
  - The year dropdown is disabled for "Recession Period Statistics."
- **Graph Output Container:**
  - Dynamic space for displaying interactive graphs.
- **Yearly Statistics Graph (Example):**
  - Bar chart showing monthly automobile sales for a selected year.
- **Recession Period Statistics Graph (Example):**
  - Line chart illustrating automobile sales during recession periods.
  - **Purpose:**
    - Explore historical automobile sales data.
    - Analyze annual sales trends.
    - Understand the impact of recessions on sales.

Github - Link

Predictive Analysis (Classification)



- **Data Exploration:**
  - Load dataset
  - Explore data
- **Data Preprocessing:**
  - Create binary target variable
  - Standardize data
  - Split data (80% train, 20% test)
- **Model Selection:**
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- **Hyperparameter Tuning:**
  - GridSearchCV (cv=10)
  - Tune hyperparameters
- **Model Training & Evaluation:**
  - Train with best parameters
  - Test accuracy on test data
- **Confusion Matrix:**
  - Analyze false positives/negatives
- **Best Model:**
  - Decision Tree (66.67% accuracy)
- **Conclusion:**
  - Model predicts rocket landings
  - Room for improvement

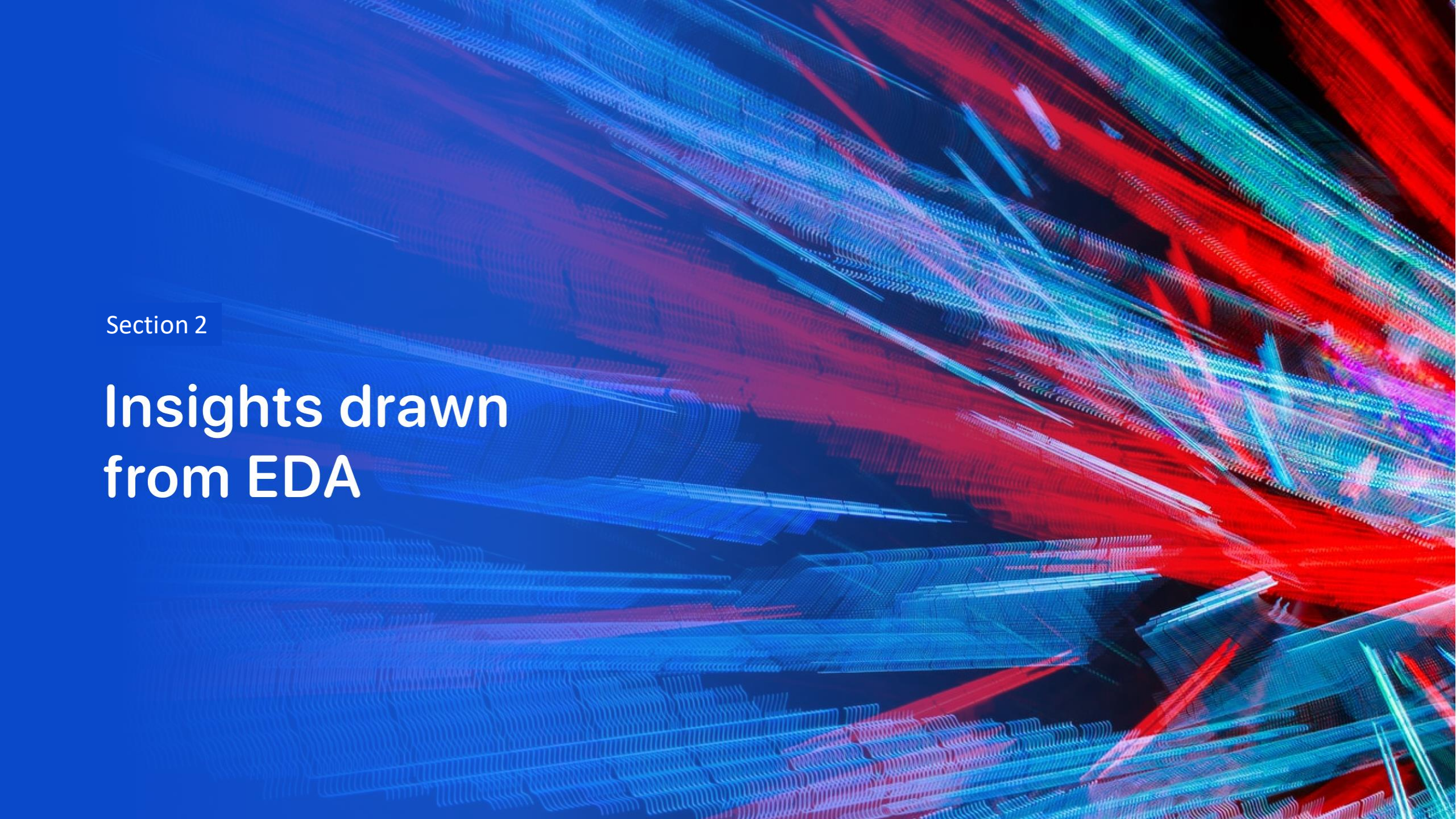
Github - Link

Results

- **Exploratory Data Analysis Results:**
- Summary stats, data visualization, correlations.
- Data anomalies and feature engineering.
- **Interactive Analytics Demo:**
- Screenshots of interactive dashboards.
- Filter, selection, drill-down options.
- Real-time data updates (if applicable).
- **Predictive Analysis Results:**
- Model selection (Logistic Regression, SVM, Decision Trees, K-NN).
- Hyperparameter tuning.
- Training, testing, and model evaluation.
- Best-performing model identified.

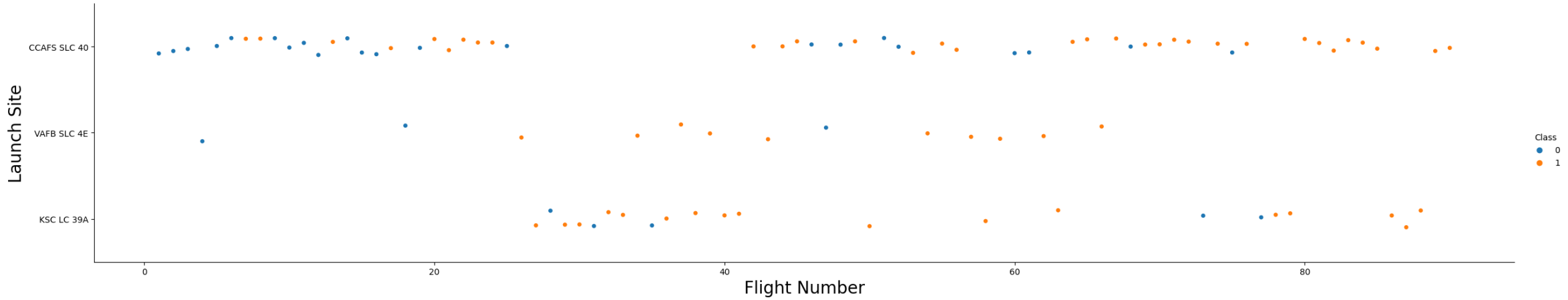
Section 2

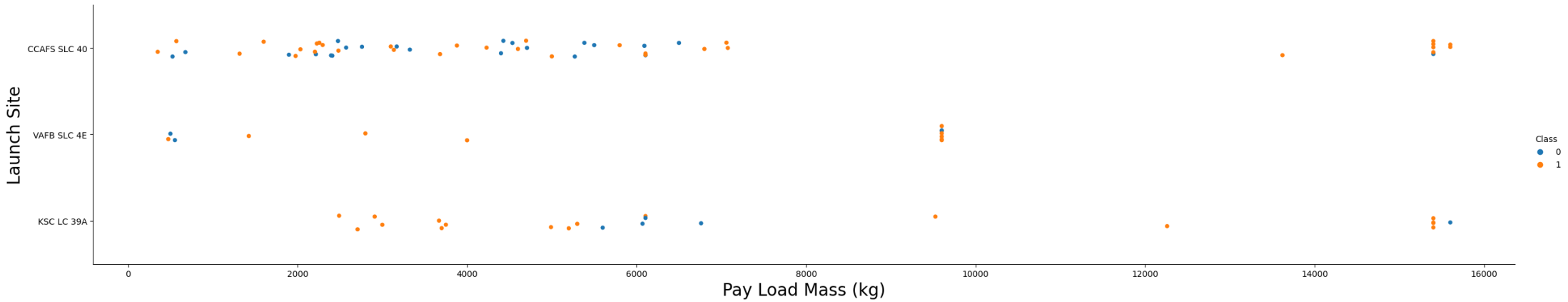
Insights drawn from EDA



Flight Number vs. Launch Site

- We can observe that Cape Canaveral Air Force Station Space Launch Complex 40 (CCAFS SLC 40) had more flights than the other launch sites



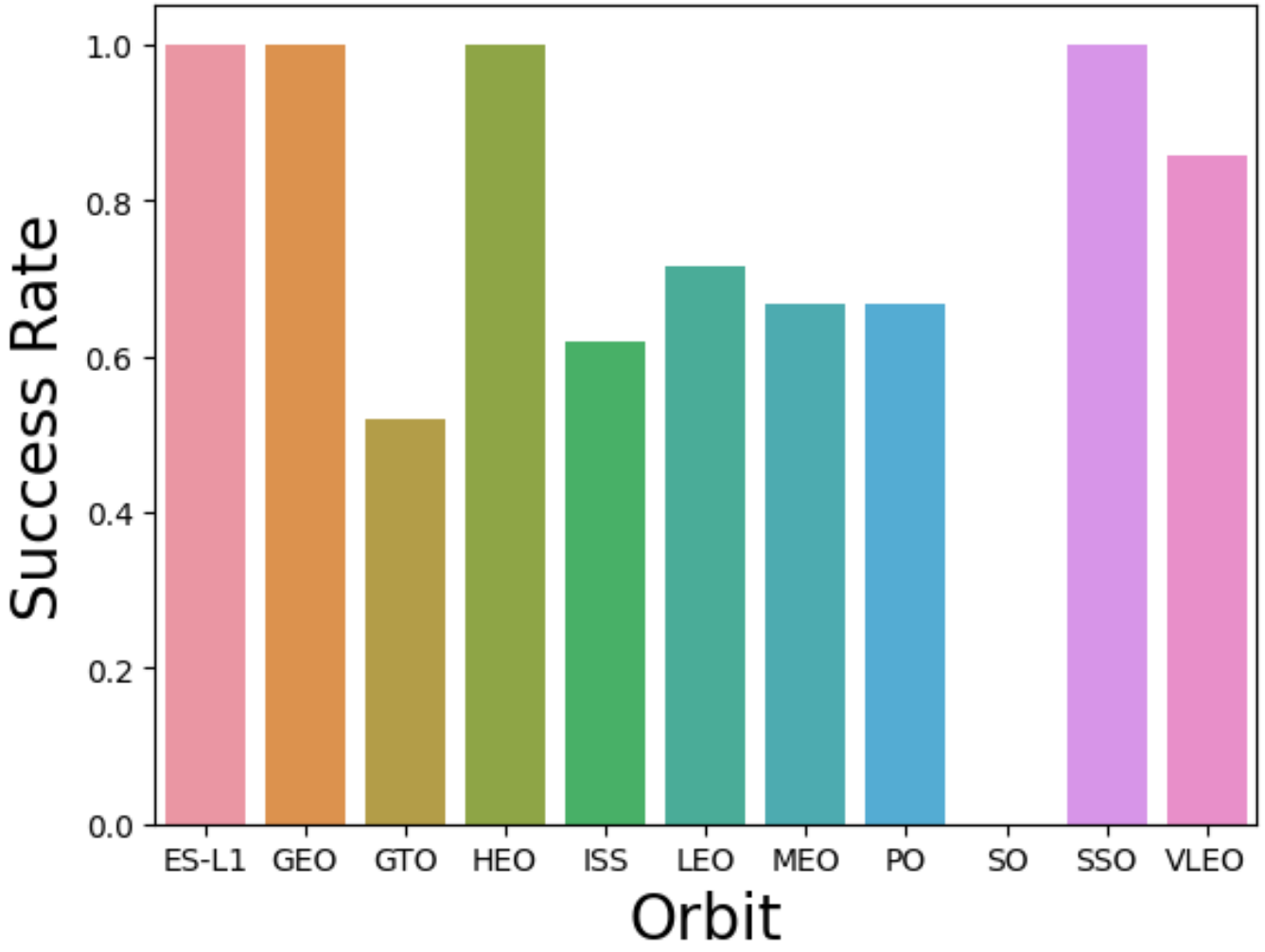


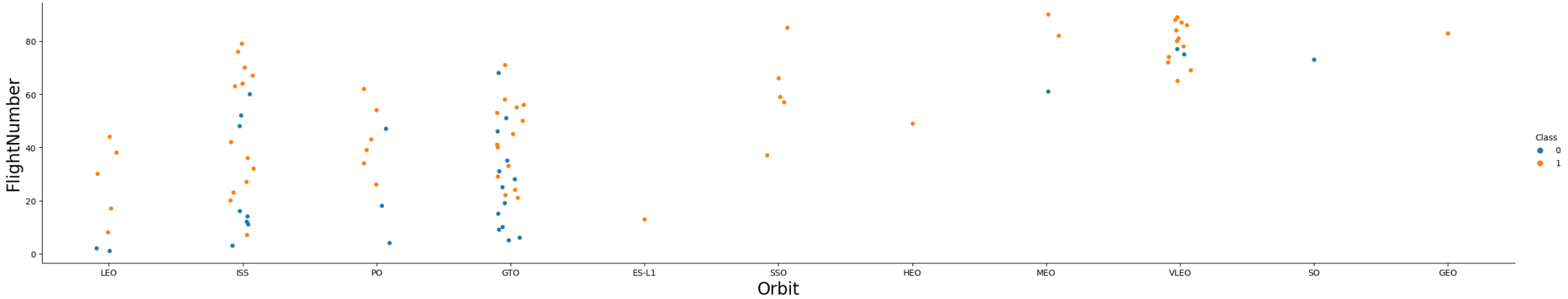
Payload vs. Launch Site

- Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type

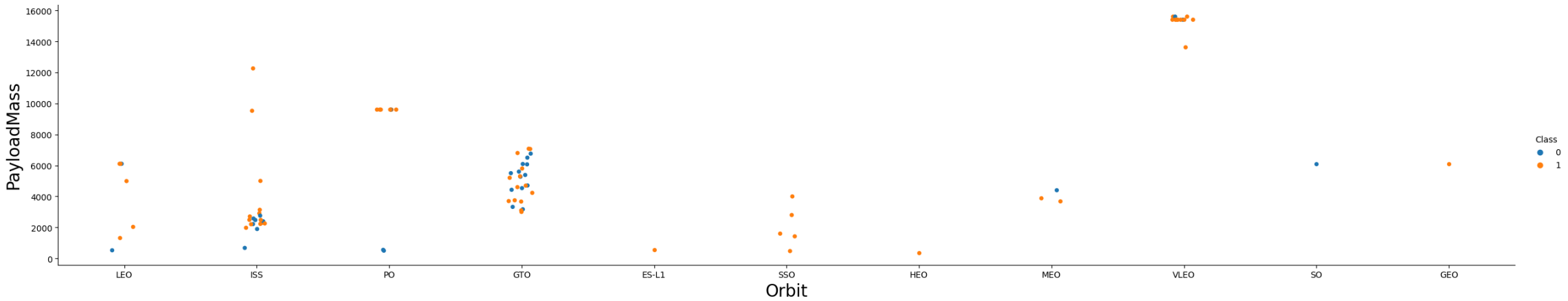
- We observe that missions to Earth-Sun L1 (ES-L1), Geostationary Orbit (GEO), Sun-Synchronous Orbit (SSO), and Highly Elliptical Orbit (HEO) had the highest success rates





Flight Number vs. Orbit Type

- 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.

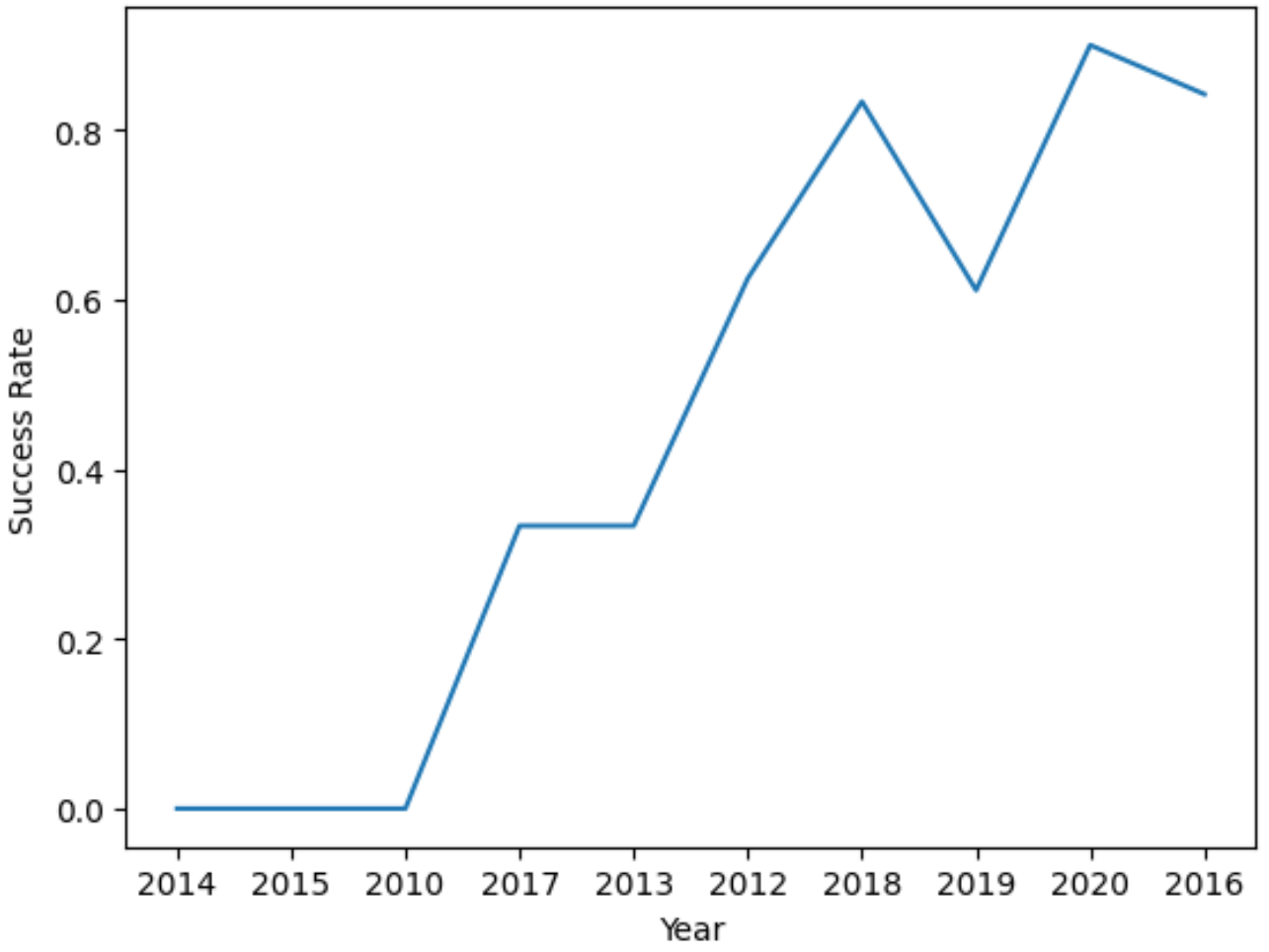


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- the success rate since 2013 kept increasing till 2020



All Launch Site Names

- SQL query using the **SELECT DISTINCT** statement to retrieve unique values from the "Launch_Site" column in a table called "SPACEXTABLE."

```
cur.execute('select distinct Launch_Site from SPACEXTABLE')
cur.fetchall()
```

```
[('CCAFS LC-40',), ('VAFB SLC-4E',), ('KSC LC-39A',), ('CCAFS SLC-40',)]
```

Launch Site Names Begin with 'CCA'

- receive the first 5 rows from the "SPACEXTABLE" table where the "Launch_Site" column starts with "CCA"

```
cur.execute('select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5')
cur.fetchall()
```

Total Payload Mass

- the sum of the "PAYLOAD_MASS__KG_" column for rows where the "Customer" column is equal to "NASA (CRS).

```
cur.execute('select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"')
cur.fetchall()
```

```
[(45596,)]
```

Average Payload Mass by F9 v1.1

- the average payload mass for rows in the "SPACEXTABLE" where the "Booster_Version" is equal to "F9 v1.1."

```
cur.execute('select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"')
cur.fetchall()
```

```
[(2928.4,)]
```

First Successful Ground Landing Date

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

```
cur.execute('select min(Date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"')
cur.fetchall()
```

```
[('2015-12-22',)]
```

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

```
cur.execute('select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ be
cur.fetchall()
```

```
[('F9 FT B1022',), ('F9 FT B1026',), ('F9 FT  B1021.2',), ('F9 FT  B1031.2',)]
```

Total Number of Successful and Failure Mission Outcomes

- List the total number of successful and failure mission outcomes

```
cur.execute('select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome')
cur.fetchall()
```

```
[('Failure (in flight)', 1),
 ('Success', 98),
 ('Success ', 1),
 ('Success (payload status unclear)', 1)]
```

Boosters Carried Maximum Payload

- List the names of the booster_versions which have carried the maximum payload mass

```
cur.execute('select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max (PAYLOAD_MASS__KG_) from SPACEXTA
cur.fetchall()
```

```
[('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.

```
import calendar
cur.execute('select substr(Date, 6, 2) as month_name, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome =
res = cur.fetchall()
res = [(calendar.month_name[int(t[0])], t[1], t[2]) for t in res]
res
```

```
[('October', 'F9 v1.1 B1012', 'CCAFS LC-40'),
 ('April', 'F9 v1.1 B1015', 'CCAFS LC-40')]
```

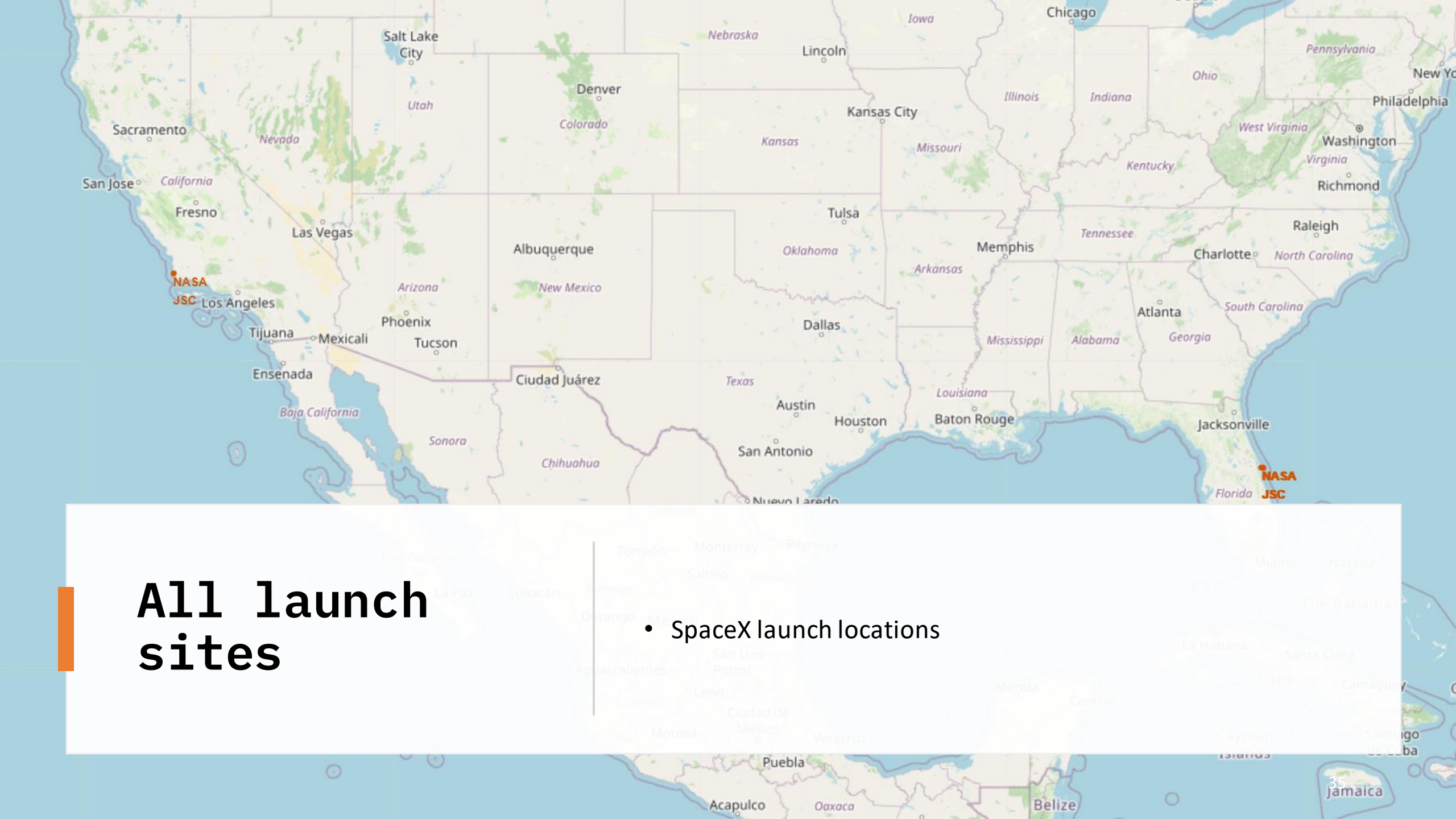
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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.

```
cur.execute('select count(*) as slurp , Landing_Outcome from SPACEXTABLE where Date between "2010-06-04" and "2017-03-20" gr
cur.fetchall()
```

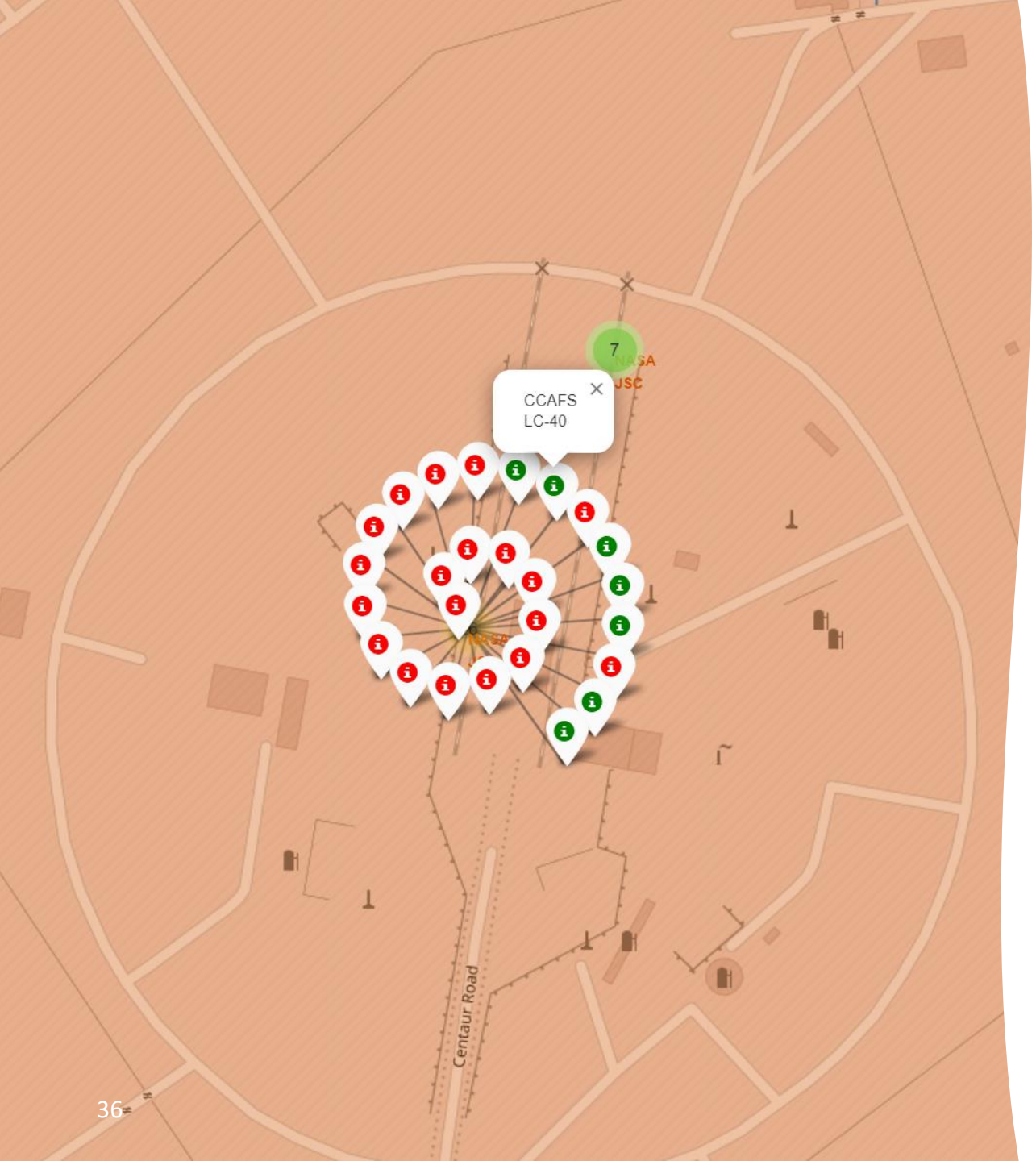
```
[(10, 'No attempt'),
 (5, 'Success (ground pad)'),
 (5, 'Success (drone ship)'),
 (5, 'Failure (drone ship)'),
 (3, 'Controlled (ocean)'),
 (2, 'Uncontrolled (ocean)'),
 (1, 'Precluded (drone ship)'),
 (1, 'Failure (parachute)')]
```


Section 3
Launch Sites
Proximities Analysis



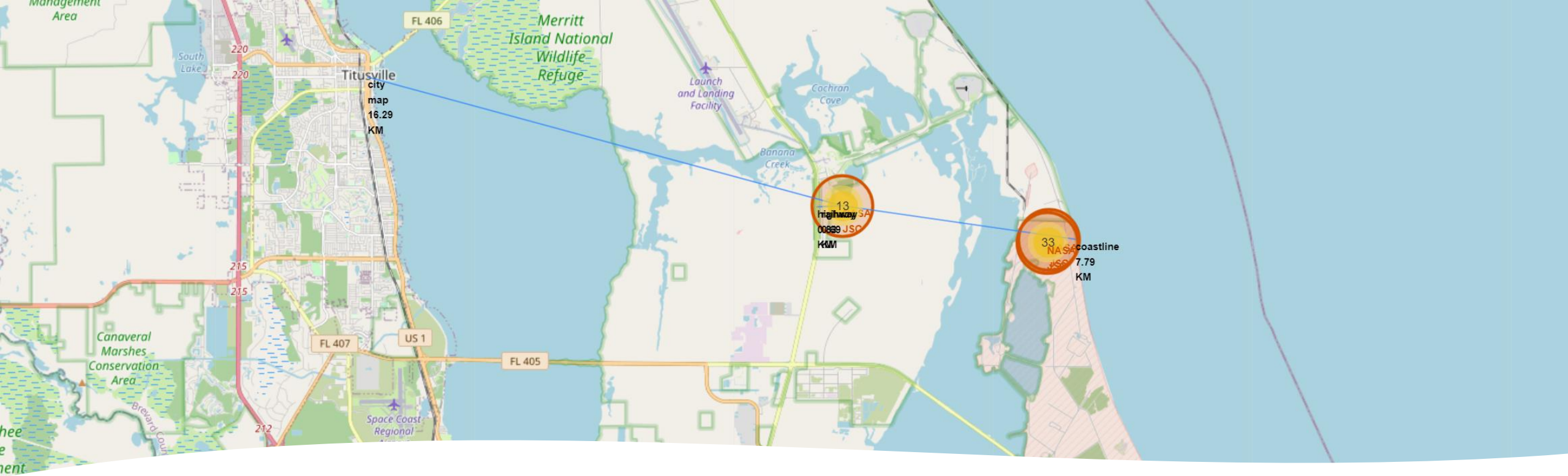
All launch sites

- SpaceX launch locations



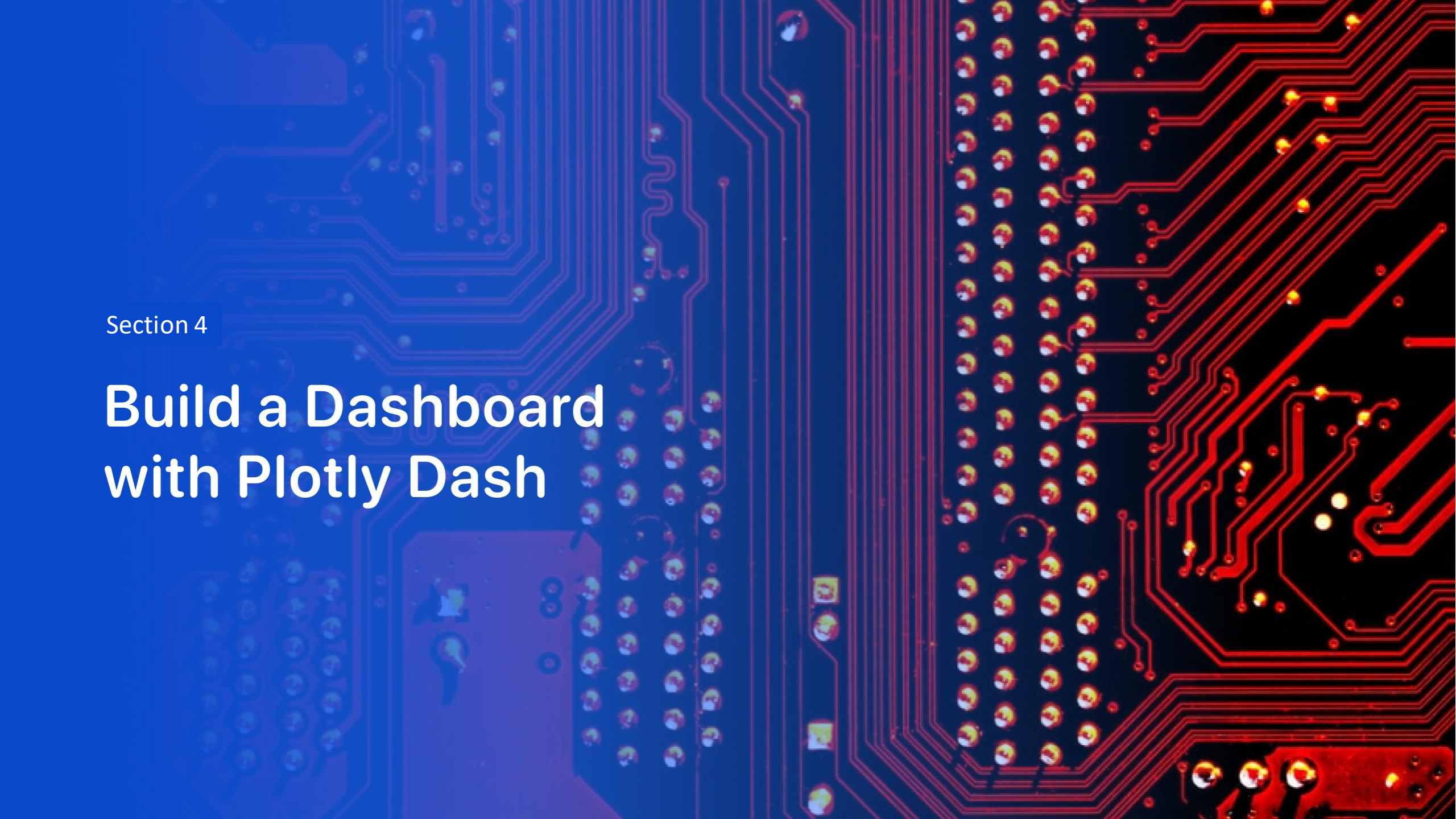
Success/failed launches

- The launch outcomes for each site with a color based on their success

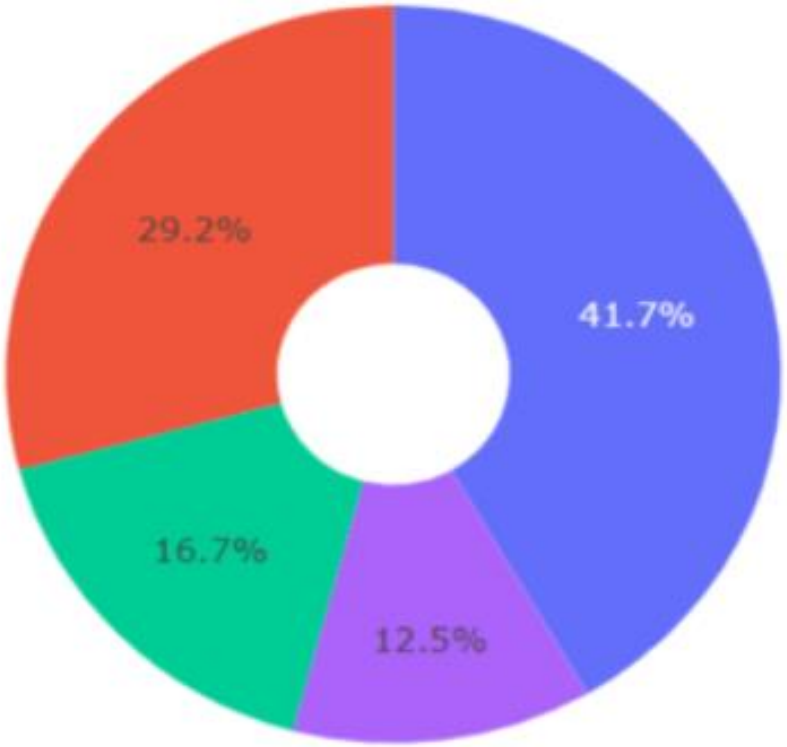


Distances between a launch site to its proximities

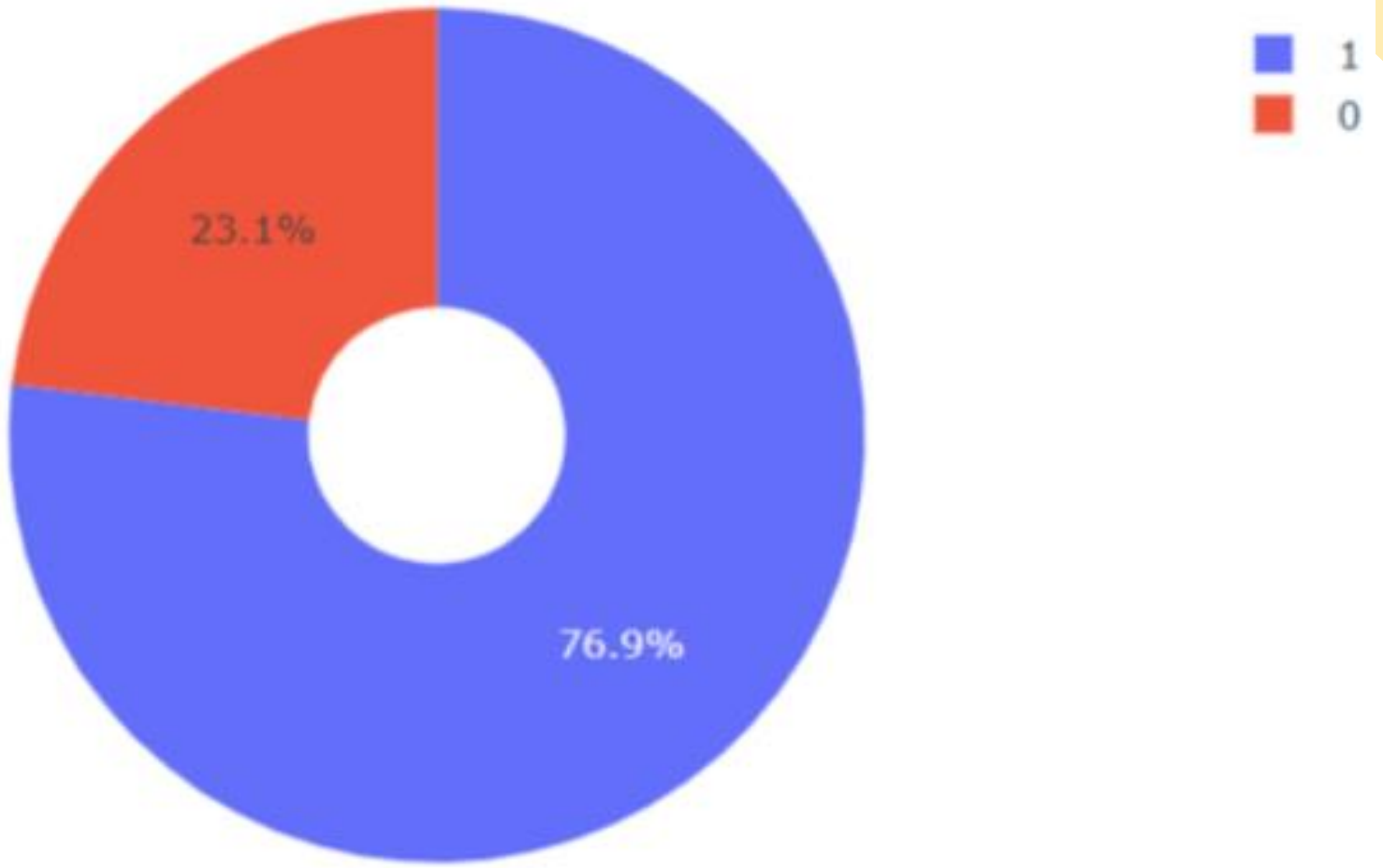
- Distance by KM


Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By all sites

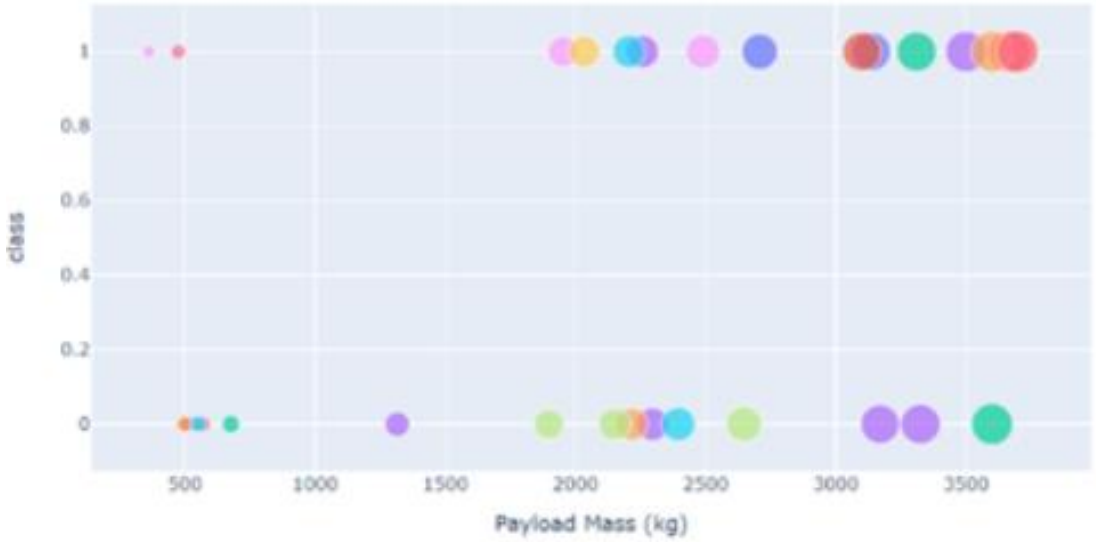


We can see that KSC LC-39A had the most successful launches from all the sites

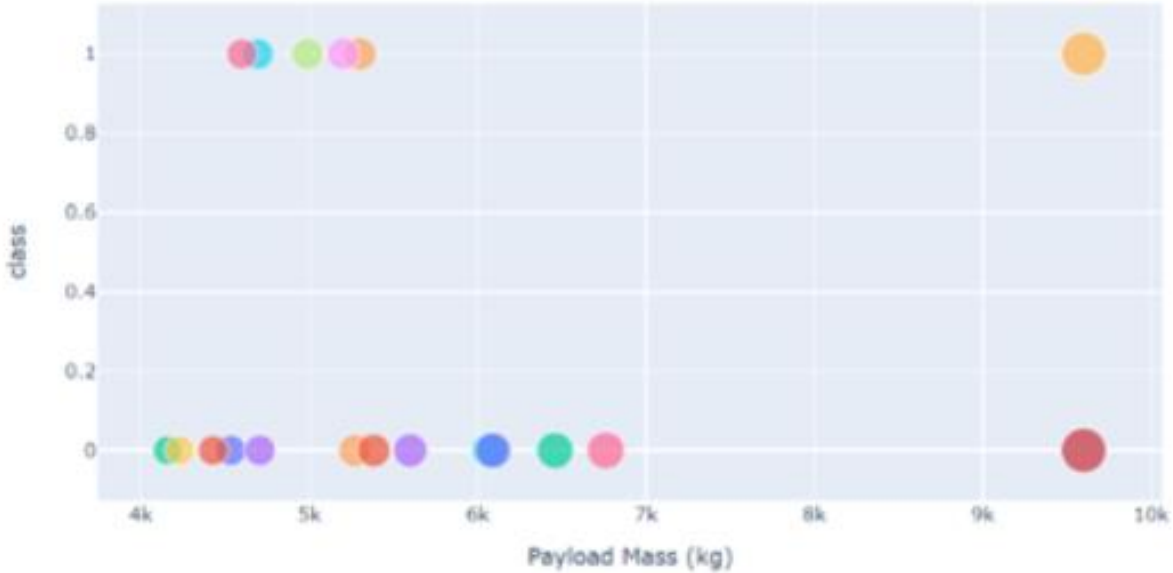


KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Low Weighted Payload 0kg – 4000kg



Heavy Weighted Payload 4000kg – 10000kg

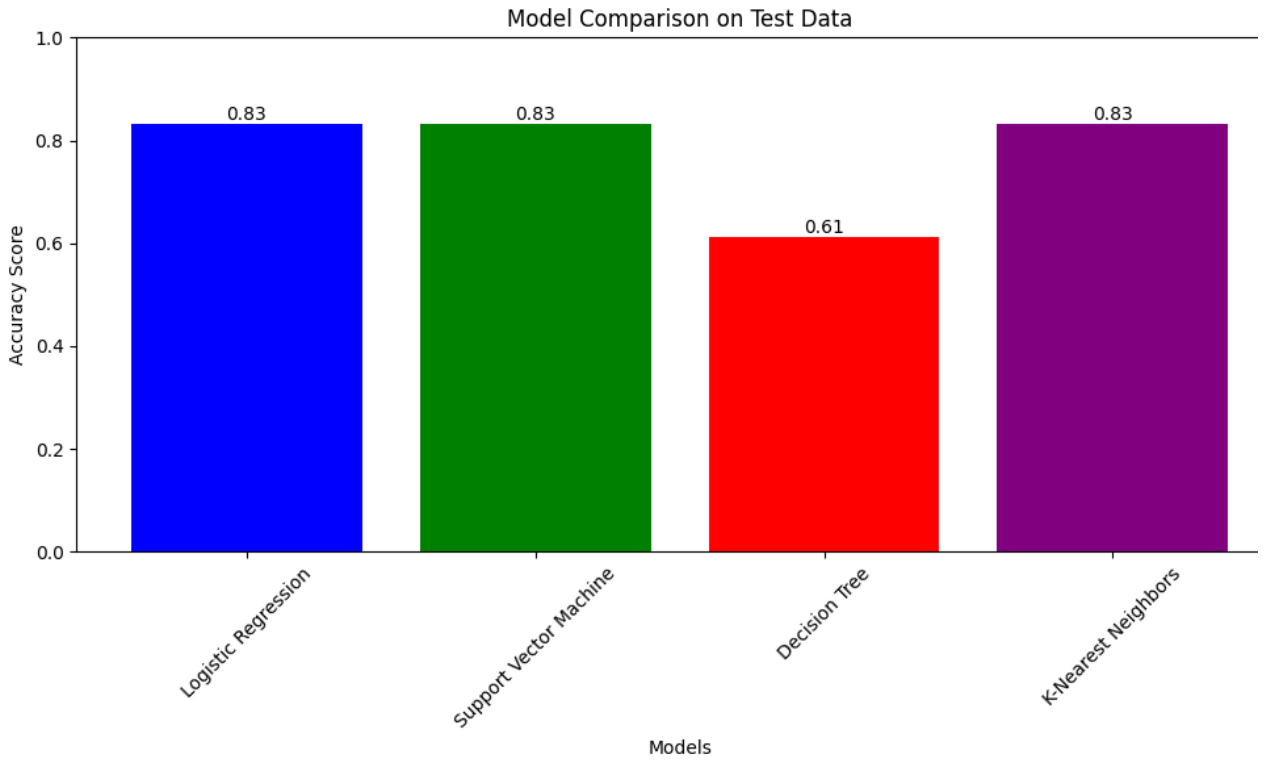


We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

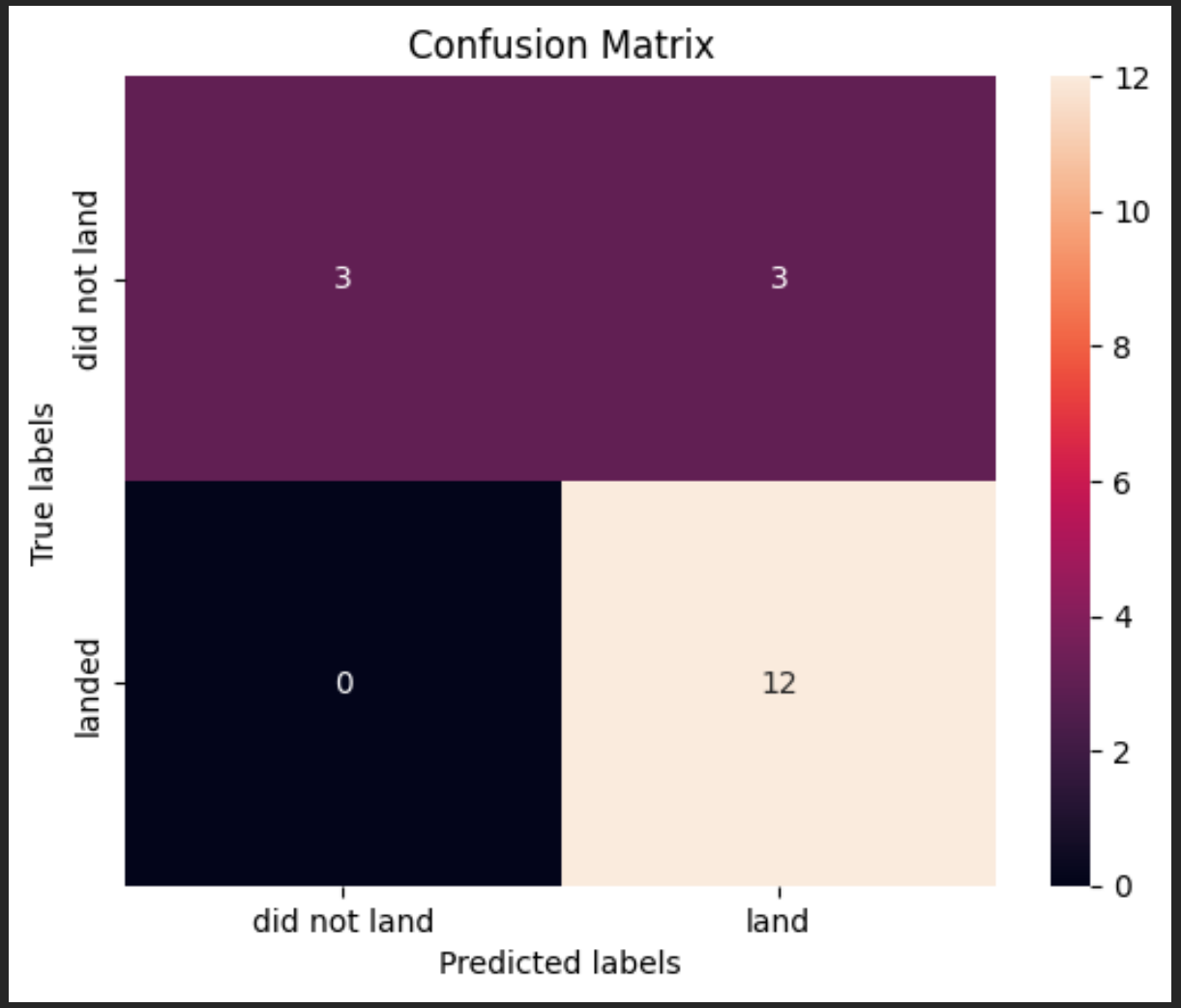
Section 5

Predictive Analysis (Classification)

Classification Accuracy



- All of LR, SVM and KNN gave a good high accuracy



Confusion Matrix

- We can notice that the accuracy is pretty high by matching 12 land - land

Conclusions

- In this project, we embarked on a comprehensive analysis of SpaceX's launch data, aiming to gain insights into the company's launch history, success rates, and payload trends. Our exploration led to several key findings:
- **Launch Site Analysis**: We identified that Cape Canaveral Air Force Station Space Launch Complex 40 (CCAFS SLC 40) had the highest number of launches among the sites considered.
- **Success Rate**: Launches to Geostationary Transfer Orbit (GTO), Highly Elliptical Orbit (HEO), and Sun-Synchronous Orbit (SSO) exhibited the highest success rates, underlining the reliability of these missions.
- **Payload Mass Trends**: An upward trend in payload mass over the years was observed, indicative of SpaceX's evolving capabilities and growing ambitions.
- **Customer Relations**: NASA (CRS) emerged as the primary customer for SpaceX, contributing significantly to the payload mass launched.

Appendix

- **Python Code Snippets**: Various Python code snippets were employed for data collection, preprocessing, and analysis. These code snippets are available in the project's source code.
- **SQL Queries**: SQL queries were used to retrieve and analyze data from the database. The specific queries can be found in the project's SQL script.
- **Charts and Visualizations**: Visual aids, including bar charts and line charts, were created to illustrate key findings and trends. These charts are available in the project's report.
- **Jupyter Notebook Outputs**: A Jupyter Notebook was utilized for in-depth data analysis and exploration. The complete notebook and its outputs can be accessed in the project's repository.
- **Data Sets**: Links to both the raw and cleaned data sets used for the project are provided for reference. These data sets can be accessed in the project's data directory.

Thank you!