Winning Space Race with Data Science

Shahzaib
29-Sep-2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- SpaceX stands out as the most triumphant enterprise in the era of commercial space exploration, effectively democratizing space travel. On its official website, the company prominently showcases its Falcon 9 rocket launches, priced at a reasonable $62 million per launch. In stark contrast, competing space launch providers charge upwards of $165 million for similar services. A substantial portion of this cost advantage stems from SpaceX's groundbreaking ability to recycle and reuse the initial stage of its rockets. Consequently, by gauging the likelihood of a successful first stage recovery, we can reliably estimate the overall cost of a launch. Leveraging publicly available data and cutting-edge machine learning models, we are poised to forecast whether SpaceX will indeed recover and reuse the first stage in upcoming missions.
- Questions to be answered?
- **Comprehensive Factor Identification:** This entailed the challenge of identifying and cataloging all the variables and elements that exert an influence on the ultimate outcome of the landing.
- **Understanding Variable Interrelationships:** There was a need to gain a deeper understanding of the interconnections between these variables and how each one impacted the overall outcome.
- **Optimal Conditions for Success:** Another challenge involved determining the most favorable and conducive conditions that would enhance the likelihood of achieving a successful landing.

Section 1

Methodology

Methodology

Executive Summary

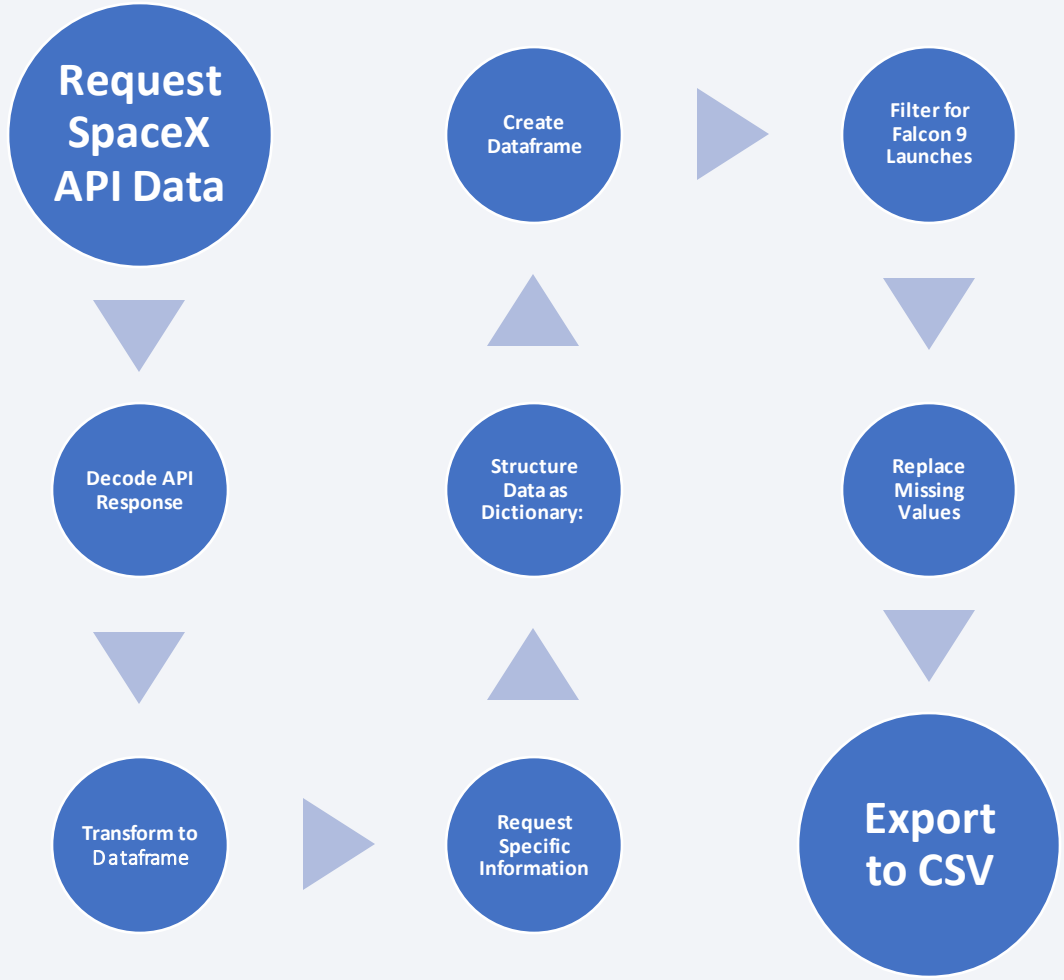
- Data collection methodology:
  - Data was collected through Web Scrapping from Wikipedia and from SpaceX Rest API
- Perform data wrangling
  - The data underwent one-hot encoding for categorical attributes.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data collection involved two main methods: REST API retrieval and web scraping from Wikipedia.
- For the REST API, we initiated a GET request, decoded the JSON response, and converted it into a pandas dataframe using json_normalize(). Afterward, data was cleaned, and missing values were handled.
- Web scraping utilized Beautiful Soup to extract launch records from an HTML table on a webpage. The scraped data was parsed and transformed into a pandas dataframe for analysis.

Data Collection – SpaceX API

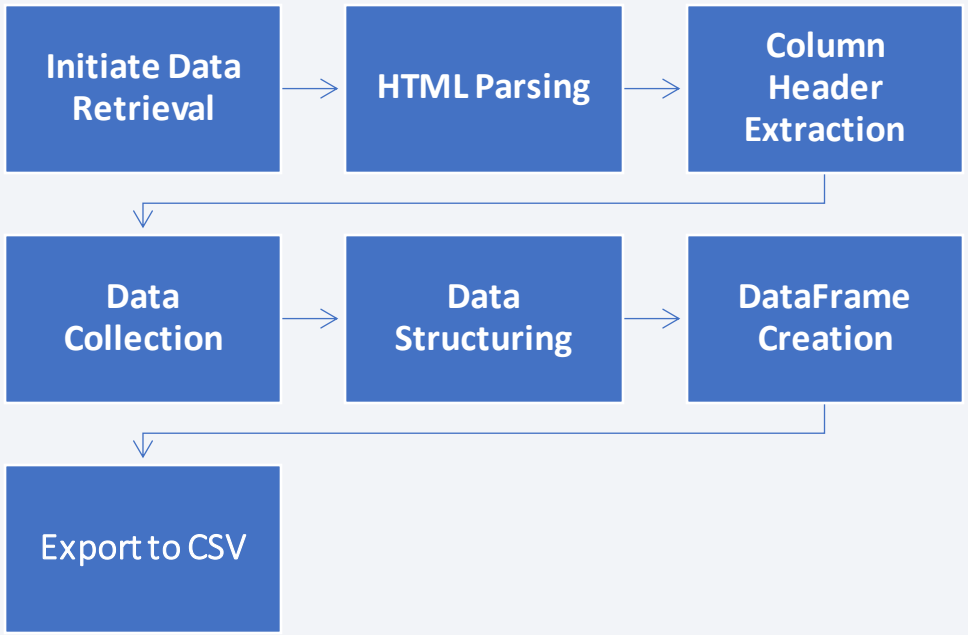
1. Requested SpaceX rocket launch data via the SpaceX API.
2. Decoded the API response content using .json() and transformed it into a dataframe with .json_normalize().
3. Extracted specific launch details from SpaceX API using custom functions.
4. Structured the acquired data into a dictionary.
5. Constructed a dataframe from the dictionary.
6. Filtered the dataframe to retain only Falcon 9 launches.
7. Replaced missing values in the Payload Mass column with the calculated mean.
8. Exported the processed data to a CSV file.



- Ctrl+Click -> Github Data Collection

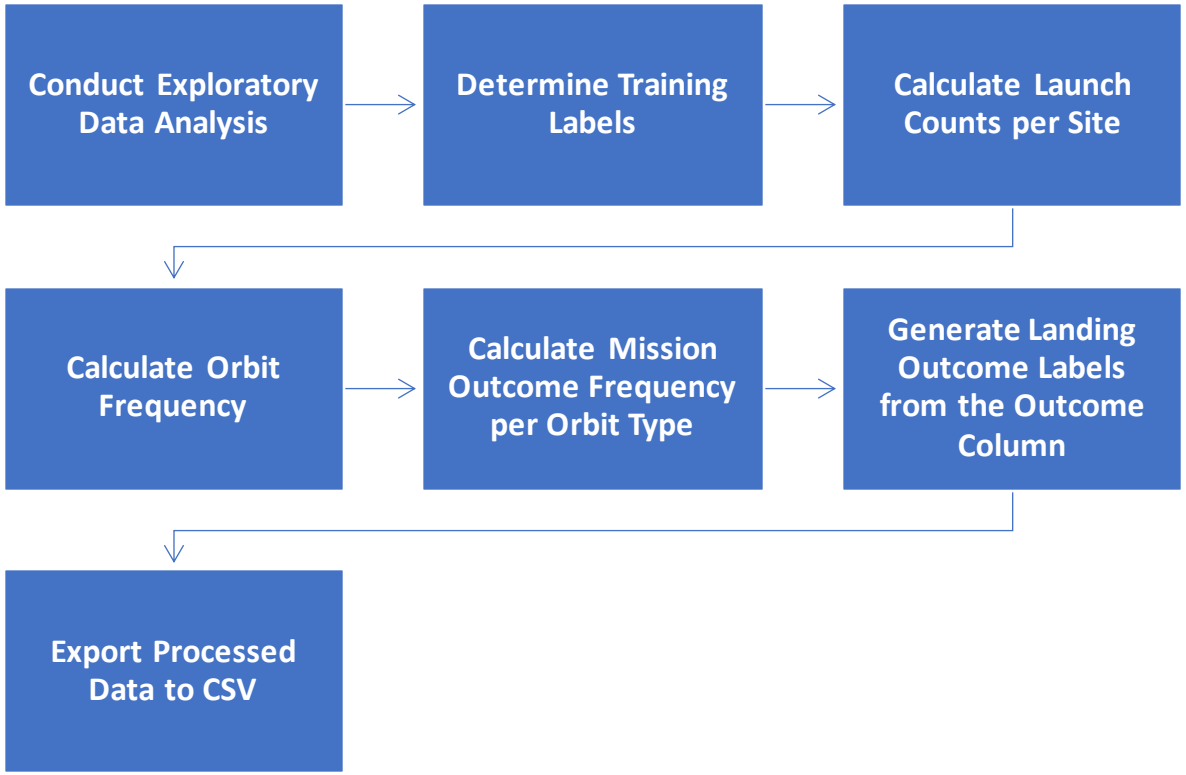
Data Collection - Scraping

- 1. Requesting Falcon 9 Launch Data from Wikipedia
- 2. Generating a BeautifulSoup Object from the HTML Response
- 3. Retrieving All Column Names from the HTML Table Header
- 4. Gathering Data by Parsing HTML Tables
- 5. Structuring Acquired Data into a Dictionary
- 6. Creating a Dataframe from the Dictionary
- 7. Exporting the Data to CSV
- Ctrl + Click -> GitHub WebScrapping



Data Wrangling

- Within the dataset, there are various scenarios where the booster did not achieve a successful landing. At times, landing attempts were made but resulted in failure due to accidents. For instance, "True Ocean" signifies a successful landing in a specific ocean region, while "False Ocean" indicates an unsuccessful ocean landing. Similarly, "True RTLS" represents a successful ground pad landing, whereas "False RTLS" denotes an unsuccessful ground pad landing. "True ASDS" signifies a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful drone ship landing.
- To simplify these outcomes, we have converted them into training labels. A label of "1" signifies a successful booster landing, while "0" denotes an unsuccessful landing.



Ctrl + Click -> GitHub Data Wrangling

EDA with Data Visualization

- Here's a summary of the charts that were plotted and the reasons for using them:
- **Flight Number vs. Payload Mass:** This chart was created to examine if there is any correlation between the flight number and the payload mass, helping us understand if there's a trend in payload weight over time.
- **Flight Number vs. Launch Site:** It was generated to explore how launch sites might change over time and if there's any pattern in the selection of launch sites.
- **Payload Mass vs. Launch Site:** This chart allowed us to compare payload masses at different launch sites, helping identify variations in payload mass based on launch location.
- **Orbit Type vs. Success Rate:** By plotting this chart, we aimed to understand how different orbit types correlate with the success rate of launches.
- **Flight Number vs. Orbit Type:** We used this chart to investigate whether there's a preference for specific orbit types over time.
- **Payload Mass vs. Orbit Type:** This chart provided insights into how payload masses are distributed across different orbit types.
- **Yearly Trend in Success Rate:** This line chart was employed to visualize the success rate trend over the years, allowing us to identify long-term patterns.
- The purpose of these charts was to gain a deeper understanding of the data, identify potential relationships between variables, and reveal trends or patterns that might be useful for further analysis or modeling.
- Ctrl+Click -> Github EDA Visualization

EDA with SQL

Here's a summary of the SQL queries in different wording:

For Displaying:

We retrieved the unique names of launch sites.

We identified and displayed 5 records where the launch site's name starts with 'CCA'.

We calculated the total payload mass carried by boosters launched by NASA under the CRS program.

We calculated the average payload mass carried by boosters of the specific version, F9 v1.1.

For Listing:

We found and displayed the date of the first successful landing on a ground pad.

We listed the names of boosters that successfully landed on drone ships with payload masses ranging from 4,000 to 6,000.

We determined and displayed the total counts of both successful and failed missions.

We identified the booster versions that carried the maximum payload.

We gathered data on failed landing outcomes on drone ships, including booster versions and launch sites, specifically for the months in the year 2015.

We counted and arranged landing outcomes between the dates 2010-06-04 and 2017-03-20 in descending order.

Ctrl+Click -> GitHub EDA With SQL

Build an Interactive Map with Folium

- **Markers:**
- Blue circle marker at NASA Johnson Space Center.
- Red circle markers at all launch sites.
- **Colored Markers:**
- Green for successful launches.
- Red for unsuccessful launches.
- **Distance Lines:**
- Colored lines show distances between launch site CCAFS SLC-40 and its nearest coastline, railway, highway, and city.
- These elements were added to enhance the map's visual clarity, display launch outcomes, and depict distances for better spatial understanding.
- Ctrl+Click -> GitHub Interactive Folium

Build a Dashboard with Plotly Dash

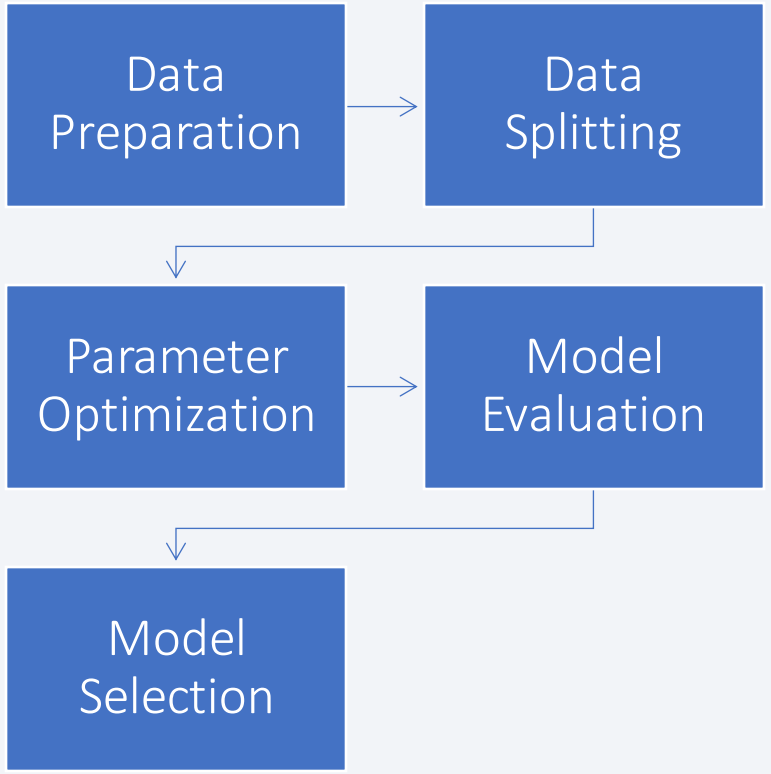
- **Dropdown List for Launch Sites:**
  - Users can select specific launch sites or view data for all launch sites.
  - Added to provide users with flexibility in exploring launch site data.
- **Slider for Payload Mass Range:**
  - Users can choose a payload mass range for data filtering.
  - Included to enable users to focus on specific payload ranges of interest.
- **Pie Chart for Successful Launches:**
  - Users can visualize the percentage of successful and unsuccessful launches relative to the total.
  - Added to offer a clear and concise overview of launch success rates.
- **Scatter Chart for Payload Mass vs. Success Rate by Booster Version:**
  - Users can examine the correlation between payload mass and launch success.
  - Included to provide insights into how payload mass relates to launch outcomes and booster versions.
  - These plots and interactive elements were incorporated to enhance user engagement, facilitate data exploration, and convey meaningful insights related to launch sites, payload masses, and launch success rates within the dashboard.

Ctrl+Click -> GitHub Dash_Plotly

Predictive Analysis (Classification)

- **Data Preparation:**
  - Create a NumPy array from the Class column.
  - Standardize the data using StandardScaler.
- **Data Splitting:**
  - Split the data into training and testing sets using train_test_split.
- **Parameter Optimization:**
  - Create a GridSearchCV object with cross-validation (cv=10) for hyperparameter tuning.
- **Model Evaluation:**
  - Apply GridSearchCV on different algorithms: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbor (KNN).
  - Calculate accuracy on the test data using .score() for all models.
  - Assess the confusion matrix for all models.
- **Model Selection:**
  - Identify the best model using evaluation metrics such as Jaccard Score, F1 Score, and Accuracy.

- Ctrl+Click -> GitHub Predictive_Analysis

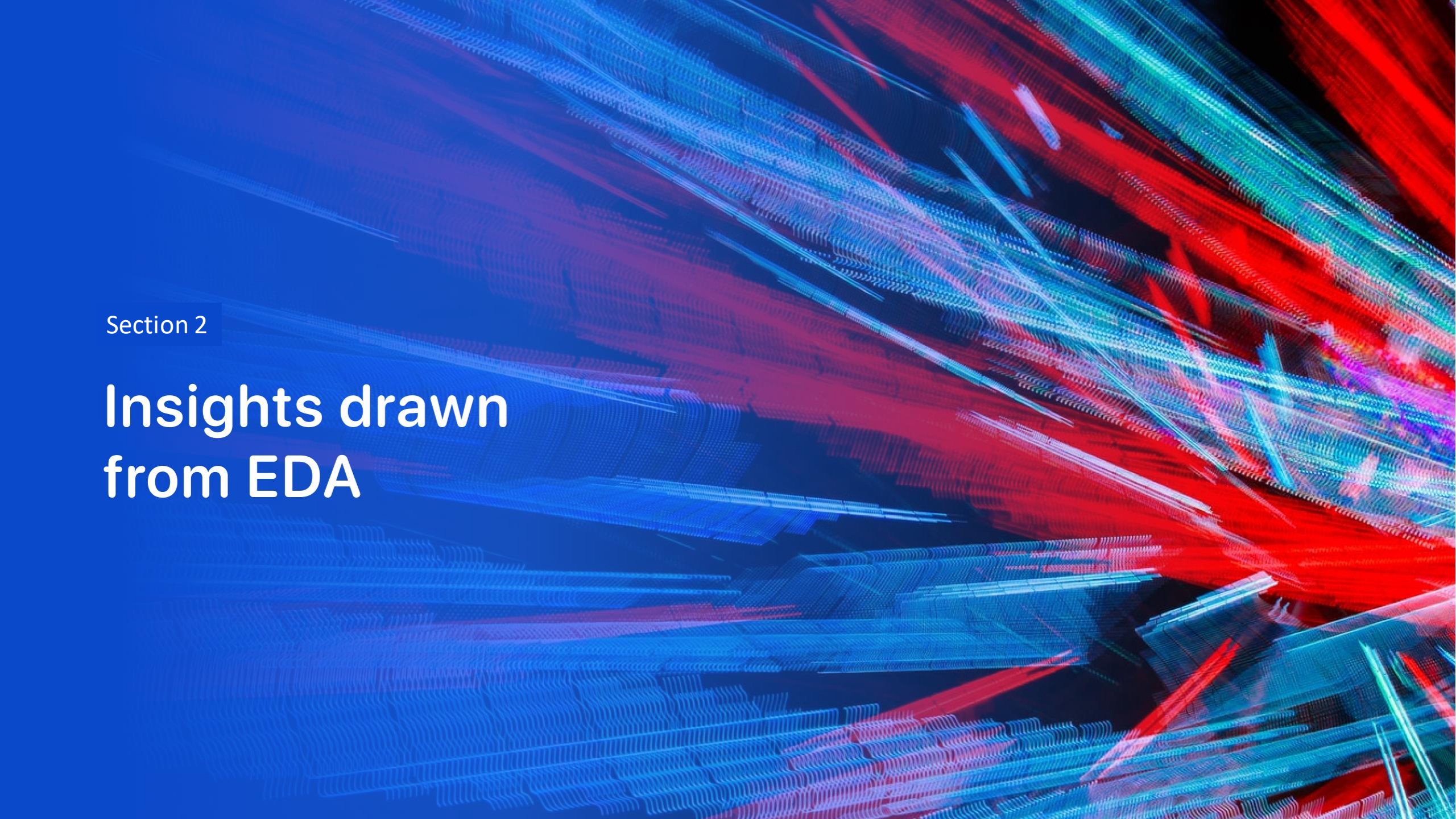


Results

- **Exploratory Data Insights:**
- A notable improvement in launch success rates has been observed over time.
- Among the landing sites, KSC LC-39A stands out with the highest success rate.
- Exceptional success rates of 100% have been achieved in orbits ES-L1, GEO, HEO, and SSO.
- **Visual Data Analysis:**
- Most launch sites are positioned in close proximity to the equator, capitalizing on the Earth's rotational speed to reduce launch costs.
- Launch sites are strategically located to ensure safety, with sufficient distance from potential impact areas such as cities, highways, and railways while maintaining accessibility for launch support.
- **Predictive Analytics Outcome:**
- Based on predictive modeling, the Decision Tree algorithm emerged as the most effective for this dataset.

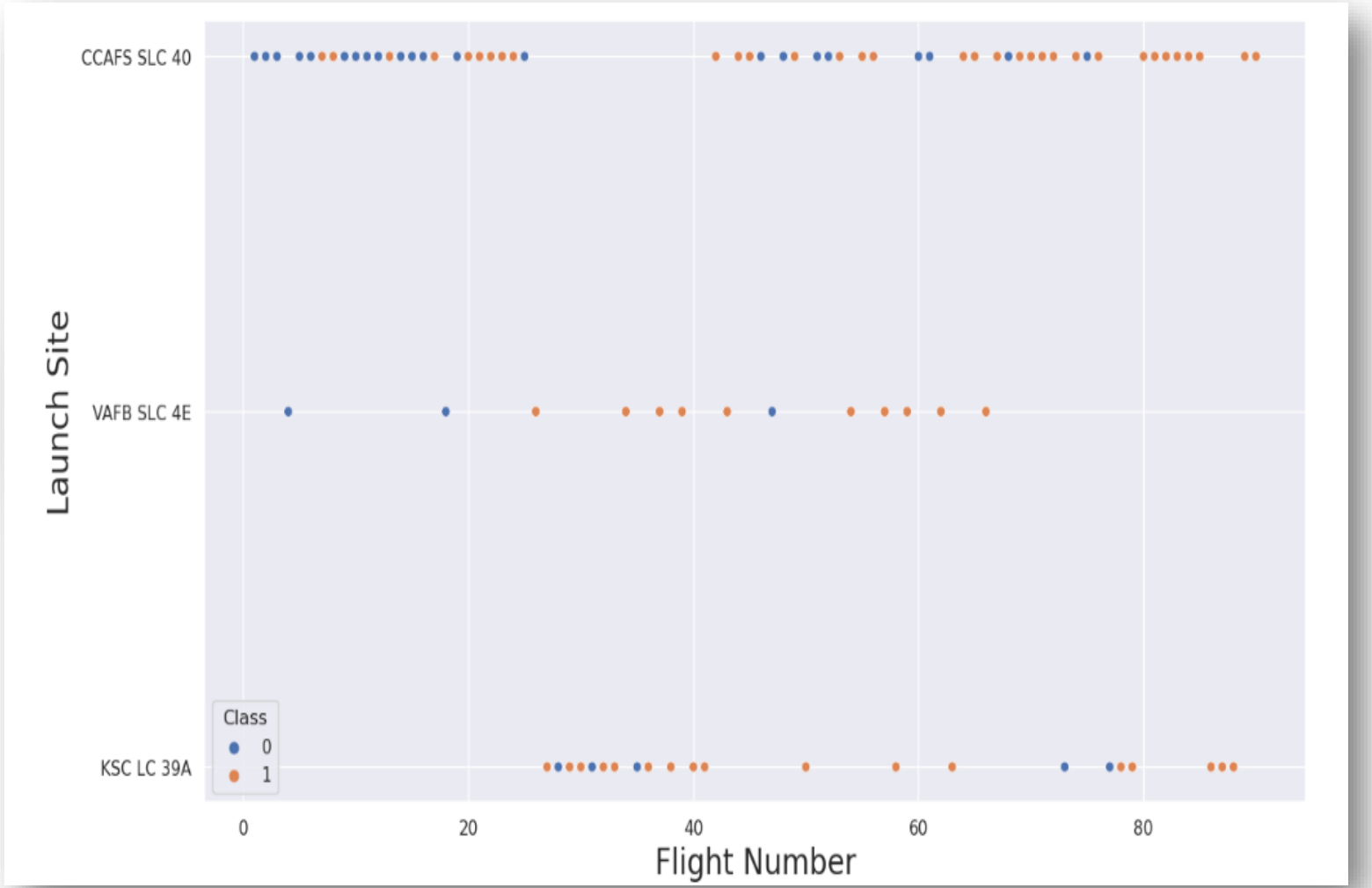
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

- The scatter plot illustrates that as the number of flights from the launch site increases, the success rate tends to rise.
- However, it's worth noting that site CCAFS SLC40 exhibits the least correlation with this pattern.

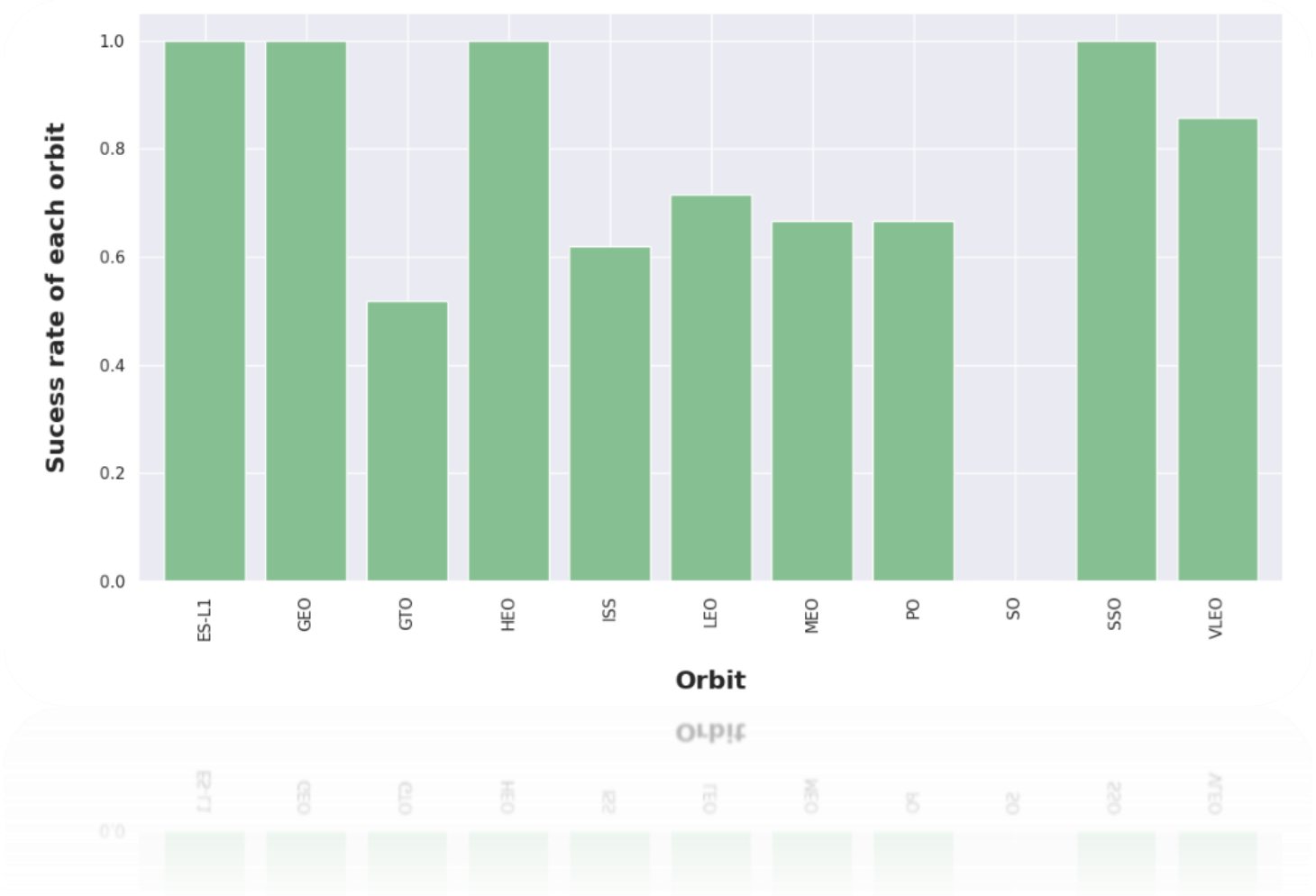


Payload vs. Launch Site

- The scatter plot reveals that once the payload mass exceeds 7000kg, the probability of a successful launch significantly increases.
- Nevertheless, there is no clear pattern to suggest that the launch site is dependent on payload mass for the success rate.

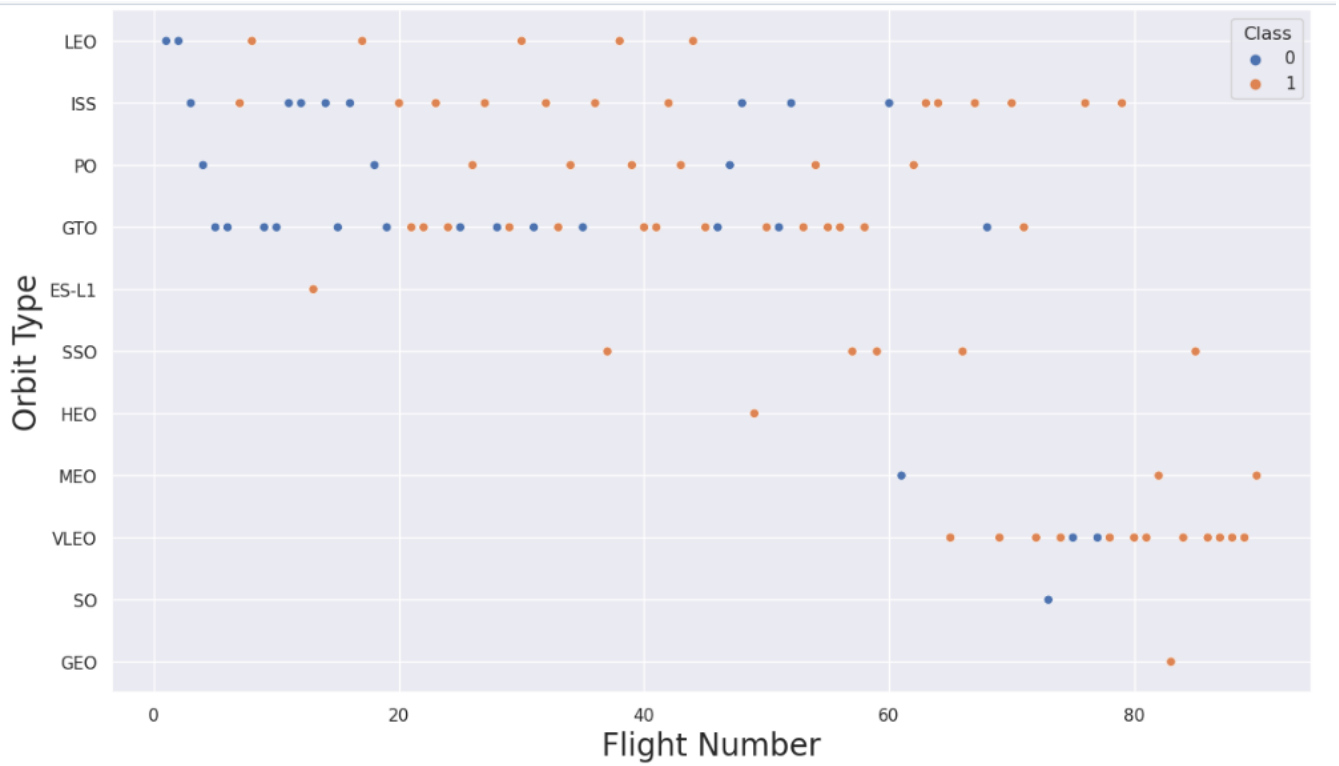
Success Rate vs. Orbit Type

- Orbits with a flawless 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with an unfortunate 0% success rate: SO
- Orbits with success rates ranging from 50% to 85%: GTO, ISS, LEO, MEO, PO



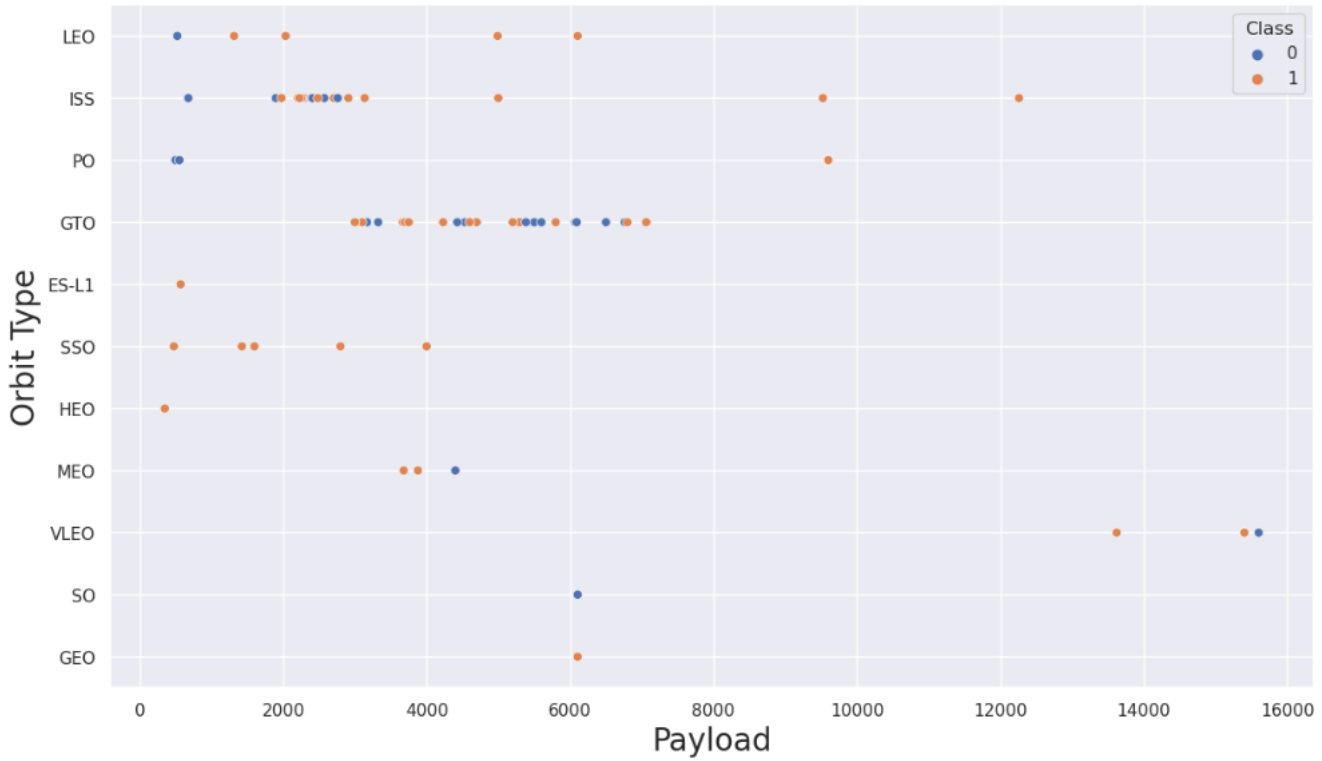
Flight Number vs. Orbit Type

- In LEO orbit, success appears correlated with the number of flights.
- In contrast, there seems to be no connection between flight count and success in GTO orbit.



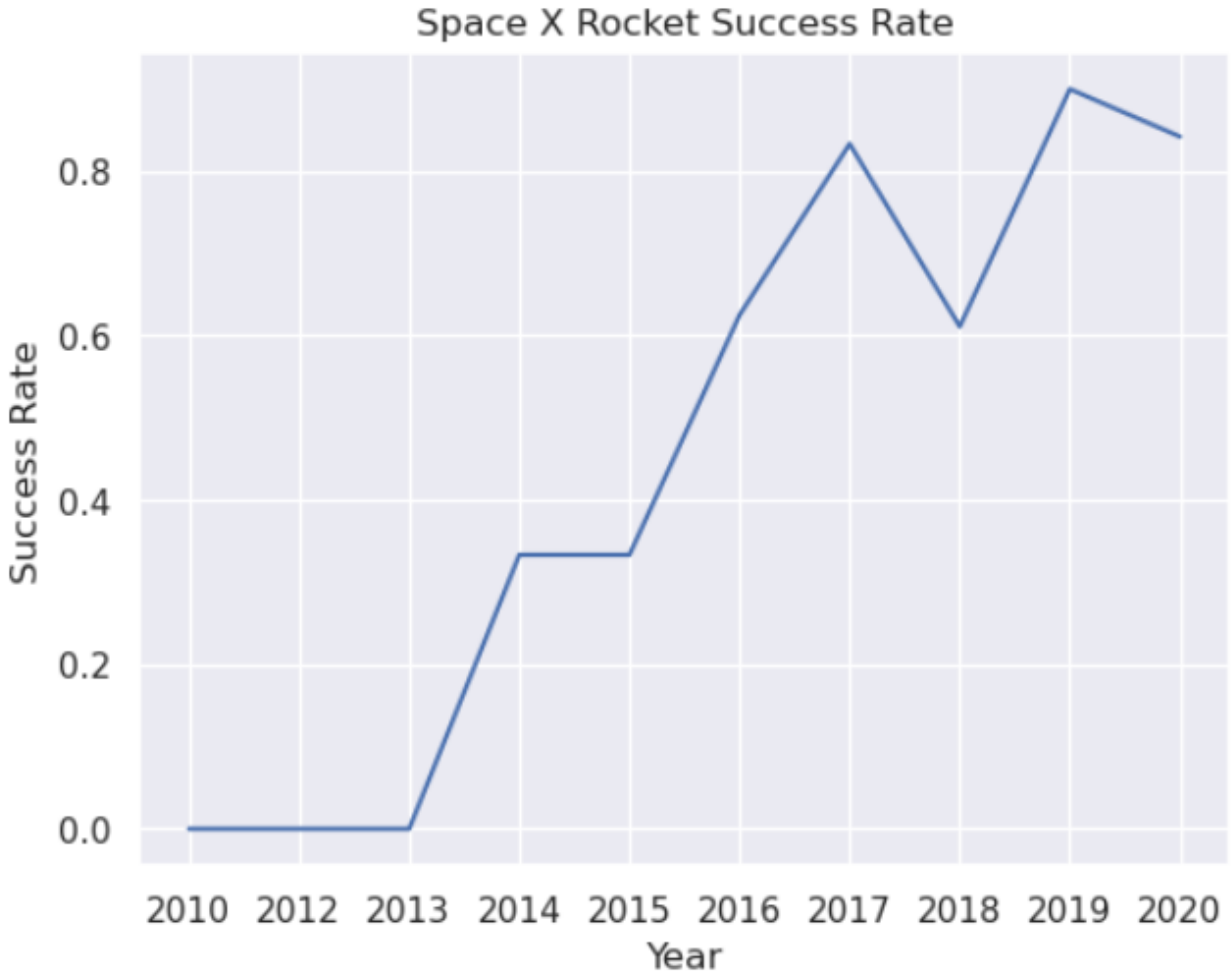
Payload vs. Orbit Type

- Heavy payloads exert a negative impact on GTO orbits but have a positive effect on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- Success rate exhibited improvement during 2013-2017 and 2018-2019.
- Success rate declined between 2017-2018 and from 2019-2020.
- Overall, there has been an improvement in the success rate since 2013.



All Launch Site Names

- We employed DISTINCT to exclusively display unique SpaceX launch sites.

In [9]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

Out[9]:

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

Launch Site Names Begin with 'CCA'

- We utilized a query to exhibit the first 5 records where launch sites commence with 'CCA.'

```
In [10]: %sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[10]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Total payload transported by NASA boosters amounted to 45596, as determined via the following query.

```
In [11]: %sql SELECT SUM (PAYLOAD_MASS__kg_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)' ;
 * sqlite:///my_data1.db
Done.
Out[11]: SUM (PAYLOAD_MASS__kg_)
         45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was computed as 2928.4.

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

In [12]:
```
%sql SELECT AVERAGE (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

First Successful Ground Landing Date

- We employed the min() function to ascertain that the initial successful ground pad landing took place on December 22, 2015.

In [13]:
```
%sql SELECT MIN (DATE) AS "First Successful Landing" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

Out[13]:

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- By using the WHERE clause and AND condition, we filtered for boosters with successful drone ship landings and payload masses between 4000 and 6000.

In [14]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND P
```

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

Out[14]:

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

Total Number of Successful and Failure Mission Outcomes

- We harnessed the '%' wildcard to filter WHERE MissionOutcome was either a success or failure.

List the total number of successful and failure mission outcomes

In [15]:
```
%sql SELECT sum(case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end) AS "Successful Mission", \
sum(case when MISSION_OUTCOME LIKE '%Failure%' then 1 else 0 end) AS "Failure Mission" \
FROM SPACEXTBL;
```

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

Out[15]:

Successful Mission	Failure Mission
100	1

Boosters Carried Maximum Payload

- The booster that carried the maximum payload was identified through a subquery within the WHERE clause and the MAX() function.

In [16]:
```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" \
FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[16]:

Booster Versions which carried the Maximum Payload Mass
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

- To filter for drone ship landing failures, their booster versions, and launch site names in 2015, we amalgamated the WHERE clause, and AND conditions.

In [23]:

```
%sql SELECT substr(DATE, 6, 2) AS Month, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
WHERE substr(DATE, 1, 4) = '2015' AND LANDING_OUTCOME = 'Failure (drone ship)';
```

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

Out[23]:

Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

In [24]:
```
%sql SELECT LANDING_OUTCOME as "Landing Outcome", COUNT(LANDING_OUTCO
AS "Total Count" FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '
GROUP BY  LANDING_OUTCOME ORDER BY COUNT(LANDING_OUTCOME) DESC ;
```

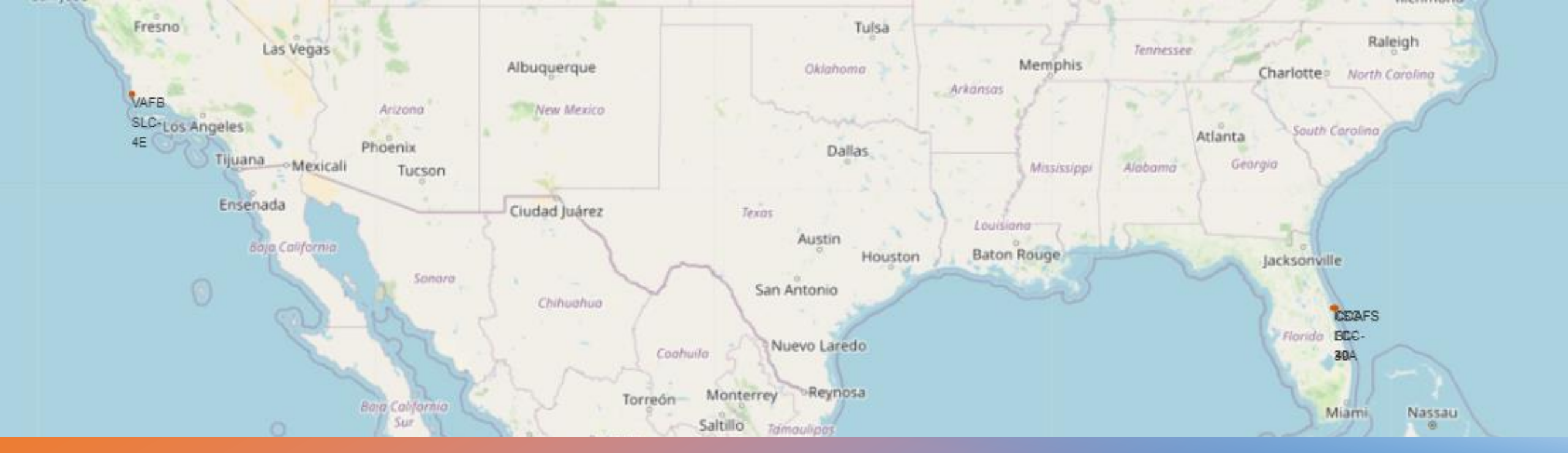
```
 * sqlite:///my_data1.db
Done.
```

Out[24]:

Landing Outcome	Total Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- Landing outcomes and the COUNT of landing outcomes were selected from the data. The WHERE clause was used to filter for landing outcomes between June 4, 2010, and March 20, 2010.
- We applied GROUP BY to group landing outcomes and ORDER BY to sort them in descending order.


Section 3
Launch Sites
Proximities Analysis

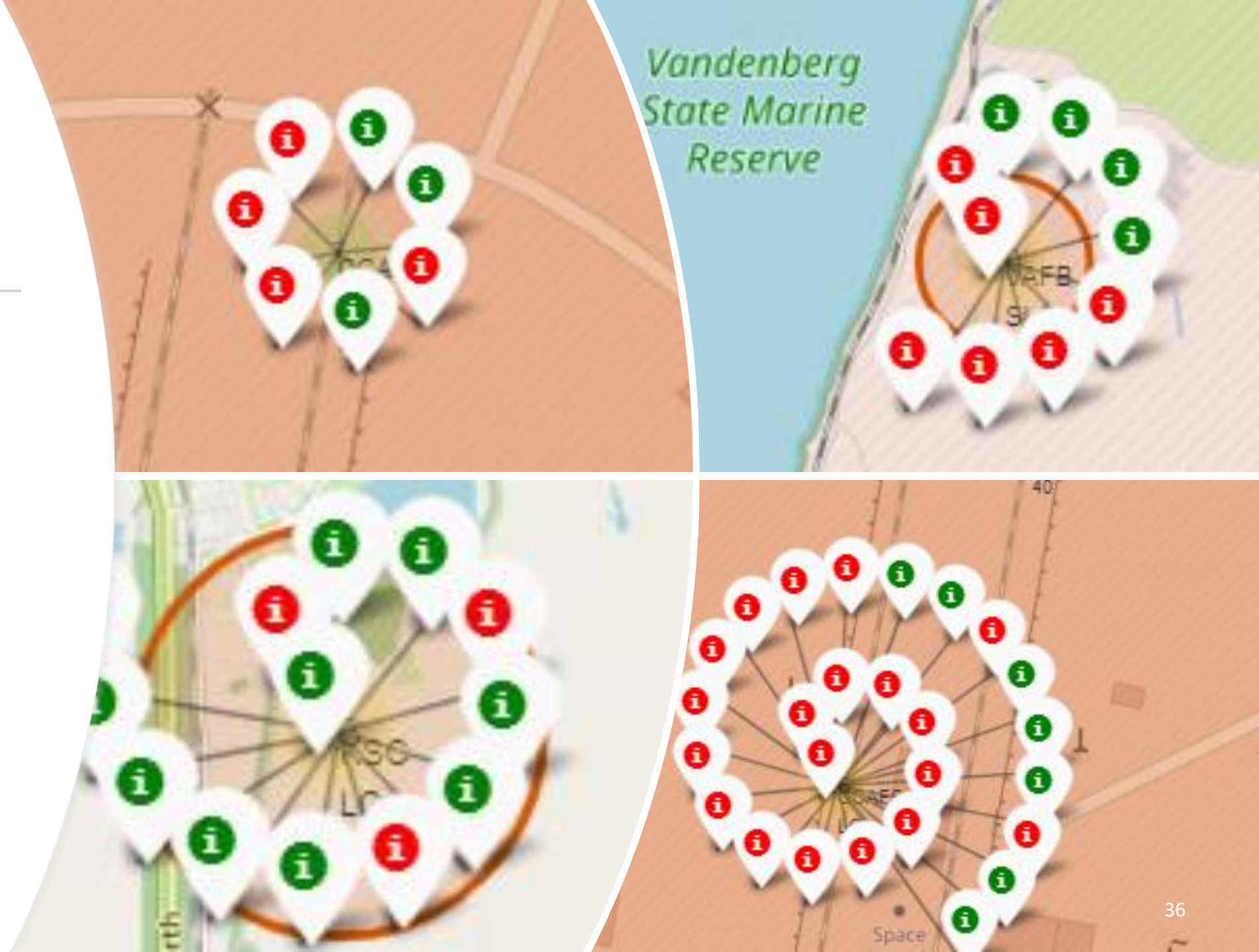


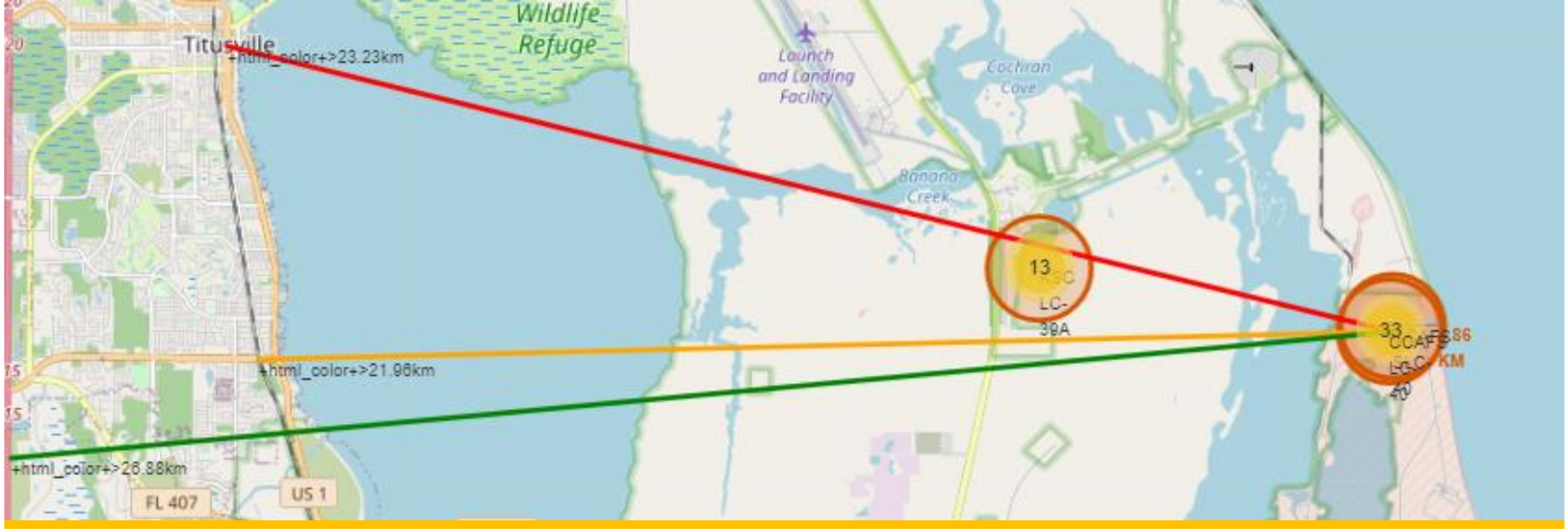
Lauch Sites Location on a Global Map

- Launch sites are near the equator to benefit from Earth's rotation, and near the coast to minimize debris risk.

Markers to show colored labeled Launch Sites

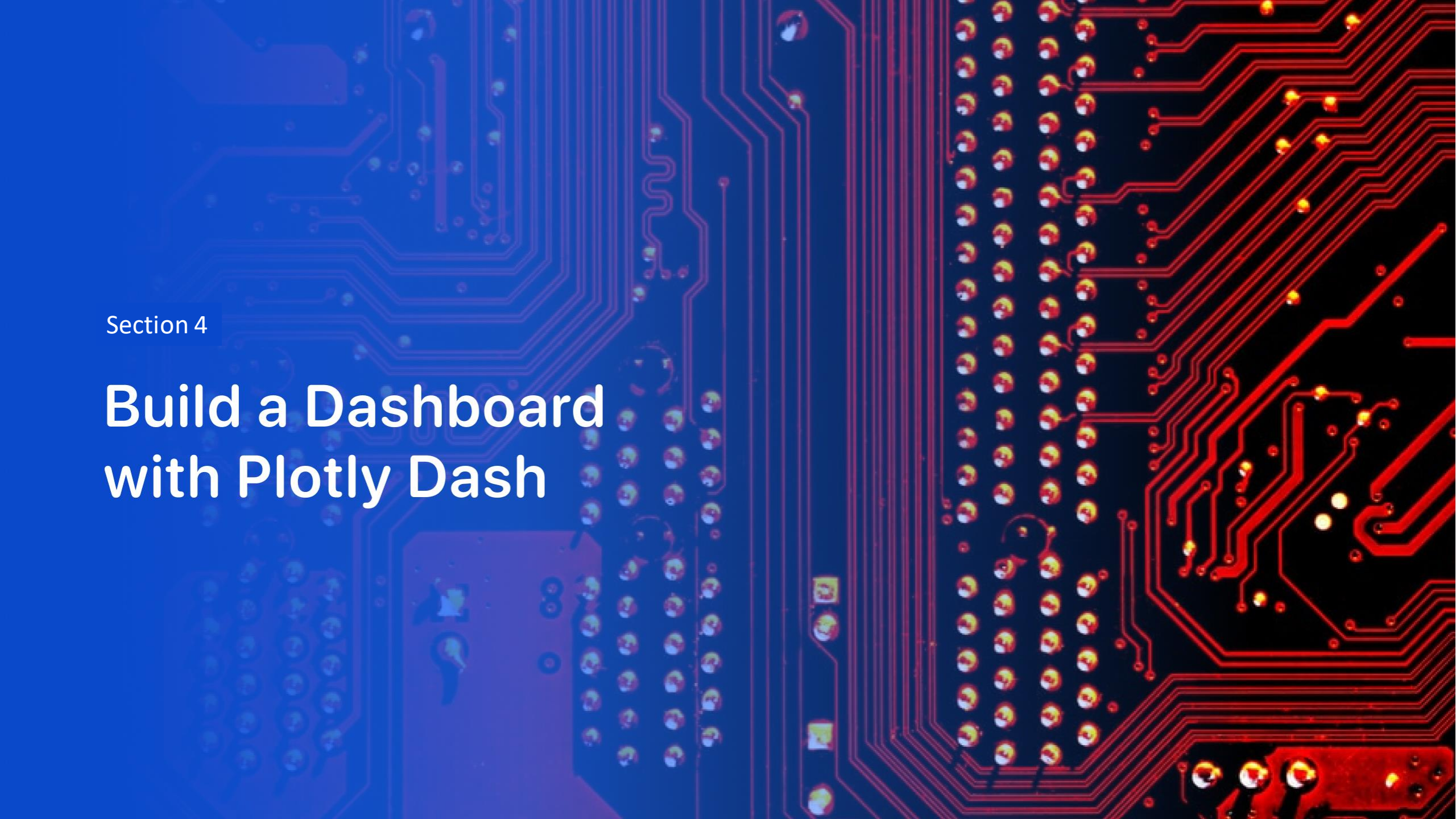
- Color-coded markers identify launch sites with high success rates
- Green: success, red: failure.
- KSC LC-39A has a very high success rate.





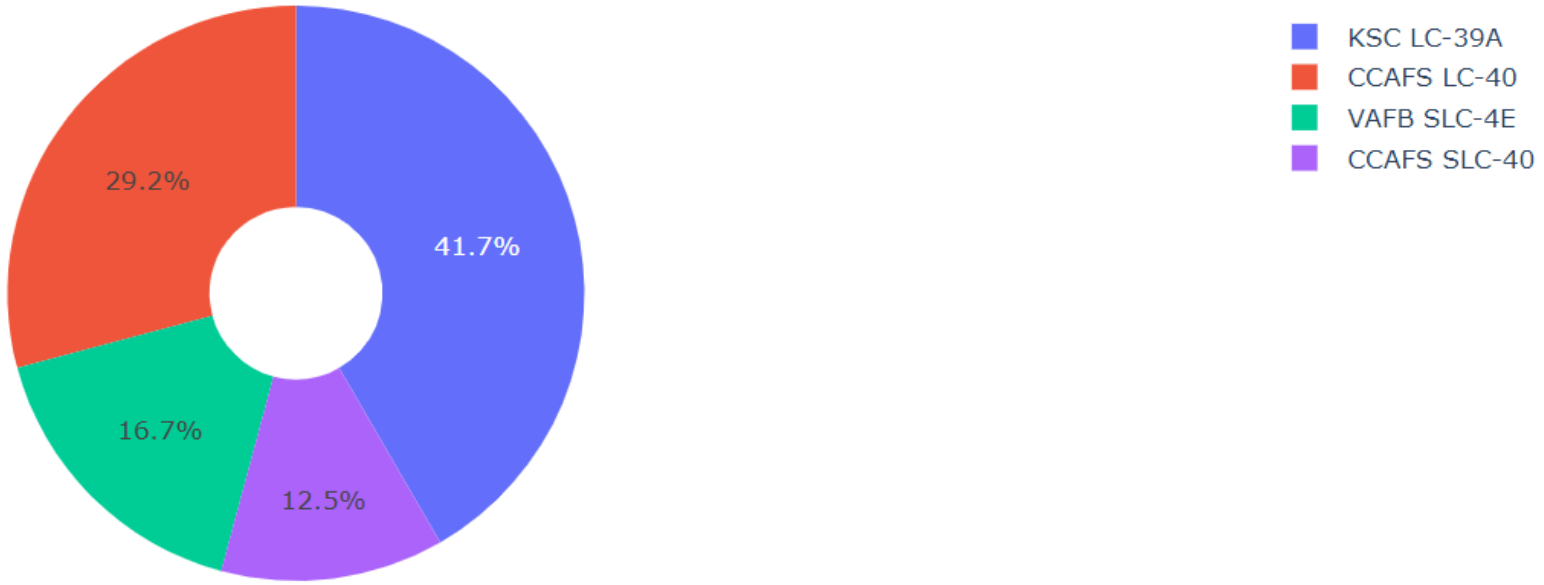
Launch Sites Distance Landmarks:

- KSC LC-39A is close to major transportation routes and populated areas as compared to CCAFS SLC-40. Failed rockets can travel at high speeds and reach these areas in seconds, posing a potential safety hazard.


Section 4
Build a Dashboard with Plotly Dash

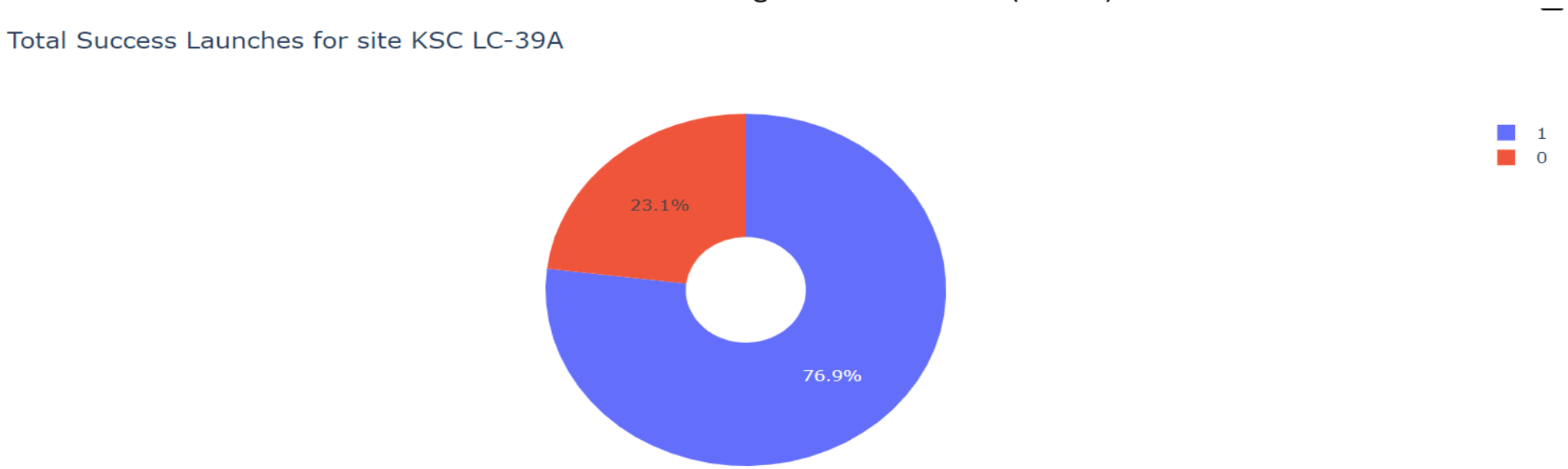
Pie Chart To Display Launch Success Count For All Sites

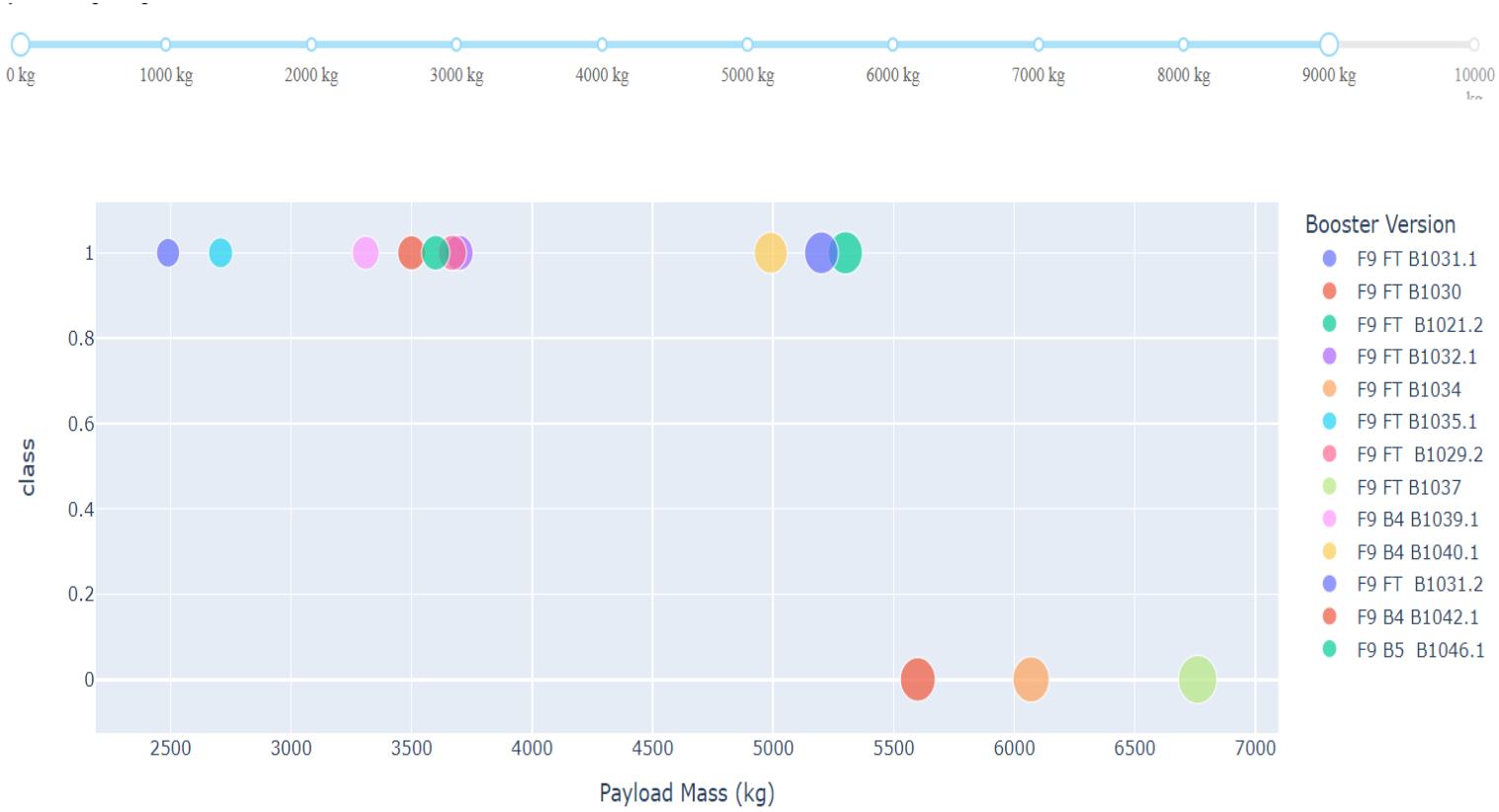
- KSC LC-39A has the highest launch success rate of any launch site (41.2%).



Highest Launch Success Ratio

- KSC LC-39A is the launch site with the highest success rate (76.9%).





Payload vs. Launch Outcome

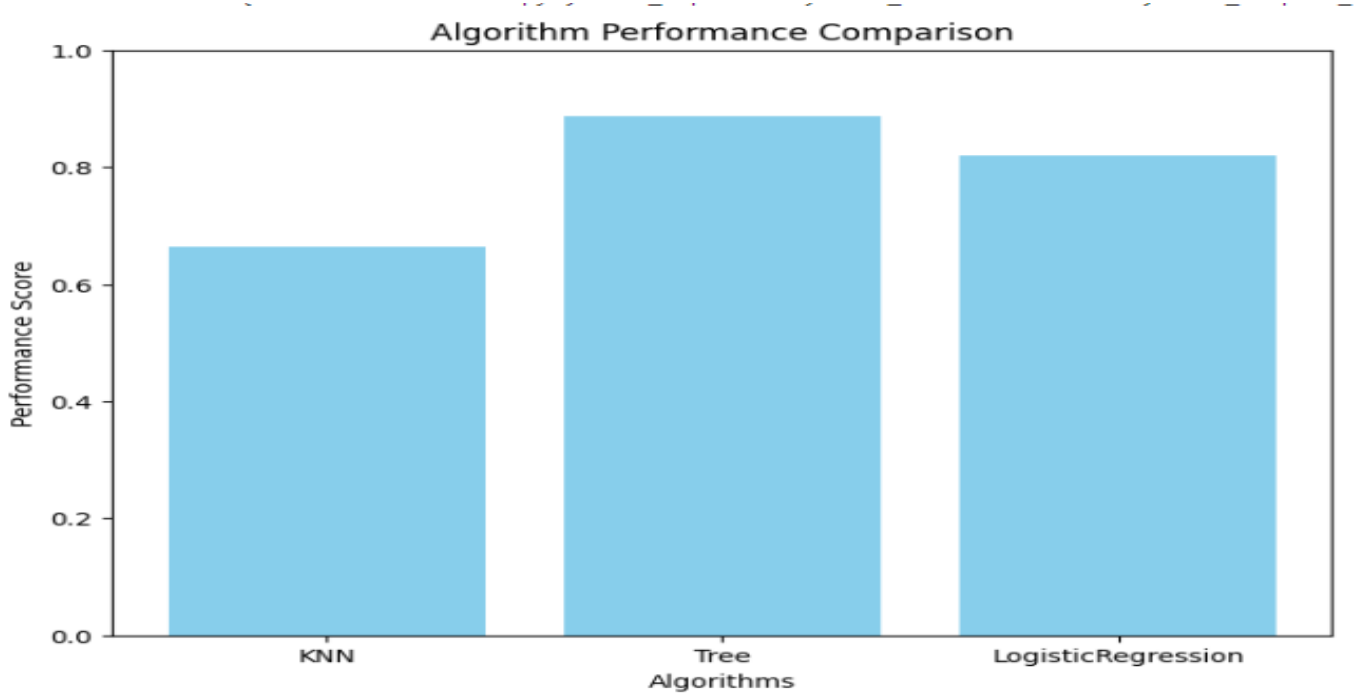
- Payloads in the 2-5 ton range have the best chance of success.
- 1 = success, 0 = failure.

Section 5

Predictive Analysis (Classification)

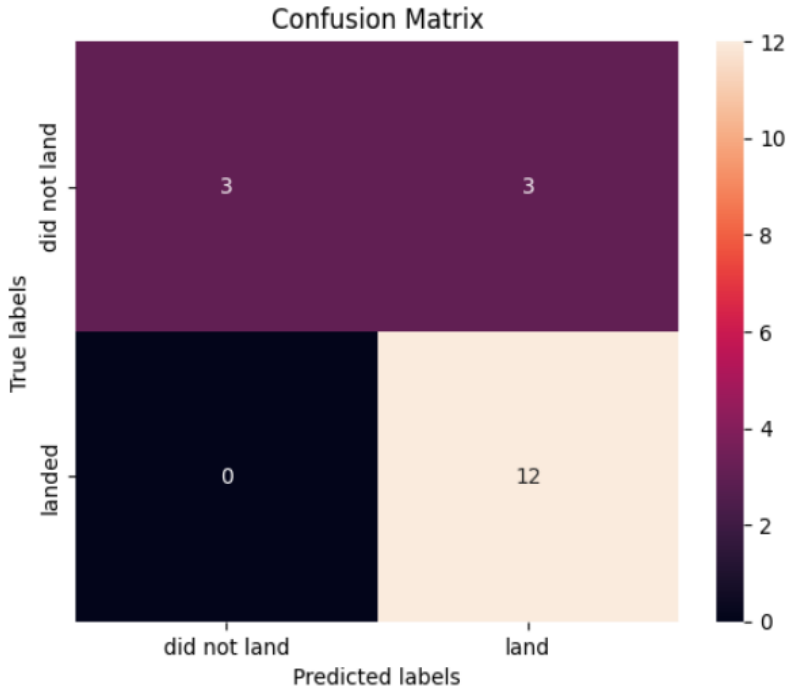
Classification Accuracy

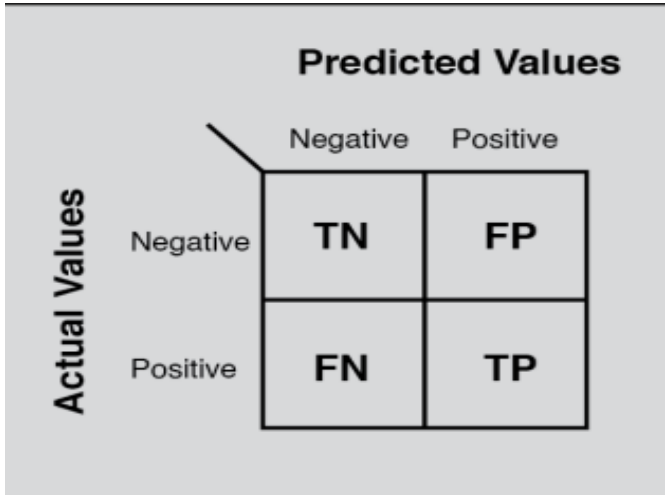
- Best Algorithm is Tree with a score of 0.8875



Confusion Matrix

- Logistic regression can identify different classes, but it produces too many false positives.





Conclusions

Model Performance: The models exhibited comparable performance on the test dataset, with the decision tree model showing a slight edge.

Proximity to Equator: Most launch sites are situated near the equator, leveraging the Earth's rotational speed to reduce the need for extra fuel and boosters, thereby saving costs.

Coastal Locations: All launch sites are strategically located near coastlines.

Positive Launch Success Trend: There is an observable trend of increasing launch success rates over time.

KSC LC-39A: This launch site stands out with the highest success rate among all launch sites. Notably, it achieves a 100% success rate for launches with a payload mass less than 5,500 kg.

Exceptional Orbits: Orbits ES-L1, GEO, HEO, and SSO have maintained a remarkable 100% success rate.

Payload Mass Influence: Across all launch sites, there is a correlation where higher payload masses (measured in kilograms) correspond to higher success rates.

Appendix

Appendix A: Data Dictionary

This appendix provides a detailed explanation of the columns in the dataset, including their names, data types, and descriptions.

Column Name	Data Type	Description
FlightNumber	Integer	Numeric identifier for each flight
LaunchSite	String	The name of the launch site
PayloadMass	Float	The mass of the payload in kilograms
Orbit	String	The orbit type of the mission
LaunchOutcome	String	The outcome of the launch (success/failure)
BoosterVersion	String	The version of the booster used
FlightDate	Date	The date of the flight
BoosterSerial	String	The serial number of the booster
PayloadType	String	The type of payload

Thank you!