IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anirudh Yadav
March 17, 2025

Outline

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

Executive Summary

- Summary of methodologies
    1. Data Collection using API
    2. Data Collection with Web Scraping
    3. Data Wrangling
    4. Exploratory Data Analysis with SQL
    5. Exploratory Data Analysis with Data Visualization
    6. Interactive Visual Analytics with Folium
    7. Prediction using Machine Learning
- Summary of all results
    1. Exploratory Data Analysis result
    2. Interactive analytics in screenshots
    3. Predictive Analytics result from Machine Learning Lab

Introduction

SpaceX has transformed the commercial space industry by offering Falcon 9 rocket launches at $62 million, significantly lower than competitors' $165 million launches. This cost reduction is due to their innovative reuse of the first stage, which can be landed and reused to further lower costs. As a data scientist at a startup competing with SpaceX, this project aims to build a machine learning pipeline using publicly available data to predict the success of the first stage landing. By forecasting landing outcomes, we can estimate launch costs and determine competitive pricing strategies.

The project answers the following questions:

1. How do variables such as payload mass, launch site, number of flights, and orbits influence the likelihood of a successful landing?
2. Which model provides the most reliable results for estimating the likelihood of a successful landing based on the given dataset?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected in 2 ways: SpaceX API and Web Scraping
- Perform data wrangling
  - Handling missing values
  - Applying One-Hot Encoding to preprocess the categorical data for binary 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 was split into training and test sets
  - Multiple models were built, trained on training data and evaluated on test data

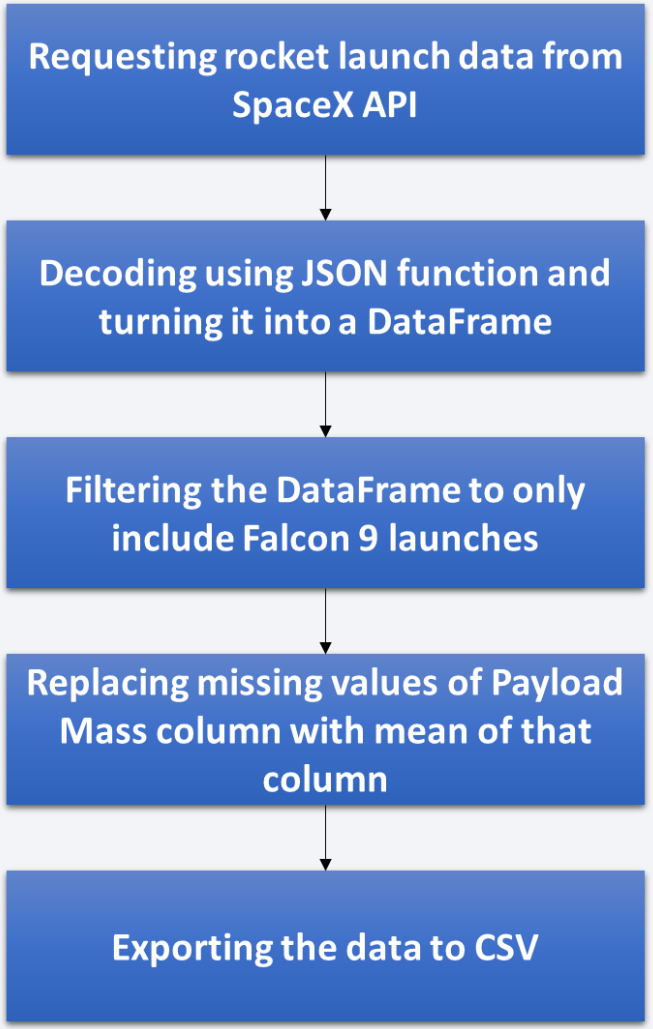
Data Collection

Data collection is the process of gathering information or data from various sources to answer specific questions, solve problems, or evaluate outcomes. It involves selecting relevant variables or factors and recording them in a structured way for analysis. As mentioned earlier, the data in this project was collected using two ways:

1. SpaceX API - Data was collected directly from the SpaceX API, which provides real-time information on Falcon launches and other relevant details.
2. Web Scraping - Data was also gathered through web scraping from Wikipedia, extracting relevant launch details and statistics for analysis,

Data Collection – SpaceX API

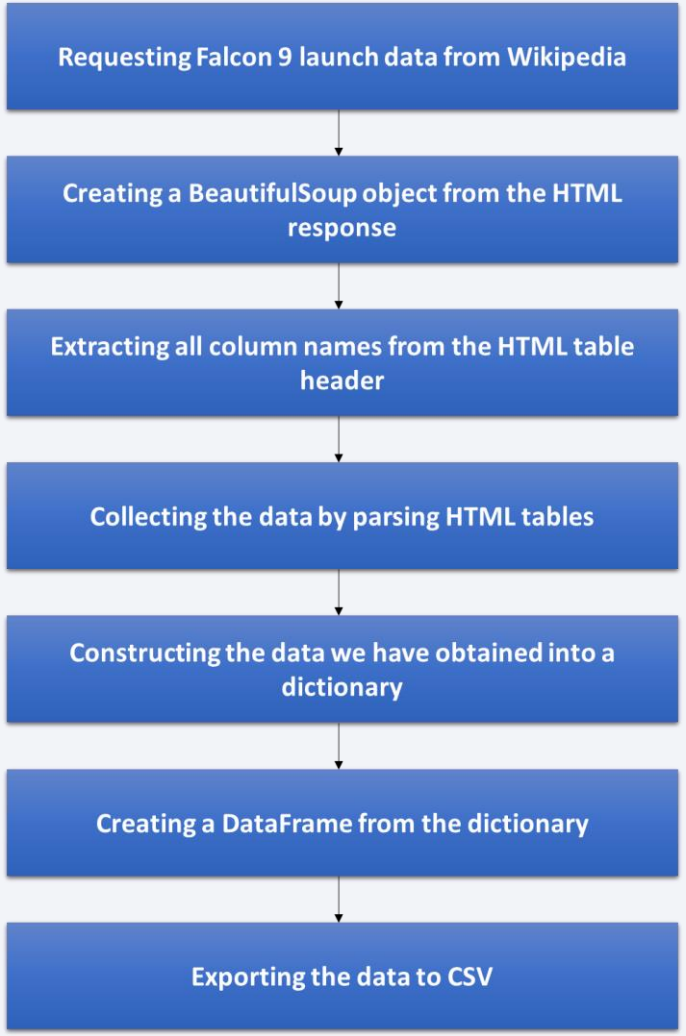
- The data was collected from the SpaceX API : https://api.spacexdata.com/v4/rockets/
- The API provides data on various SpaceX rocket launches, but the data was filtered to include only Falcon 9 launches.
- A GET request was made to fetch the data from the API. The response content was decoded as JSON, which was then converted into a Pandas DataFrame using the json_normalize() method.
- After converting the data, it was cleaned by:
  1. Checking for missing values.
  2. Filling missing values with the mean of the respective columns.
- The final dataset consists of 90 rows and 17 columns.



GitHub: SpaceX API

Data Collection - Scraping

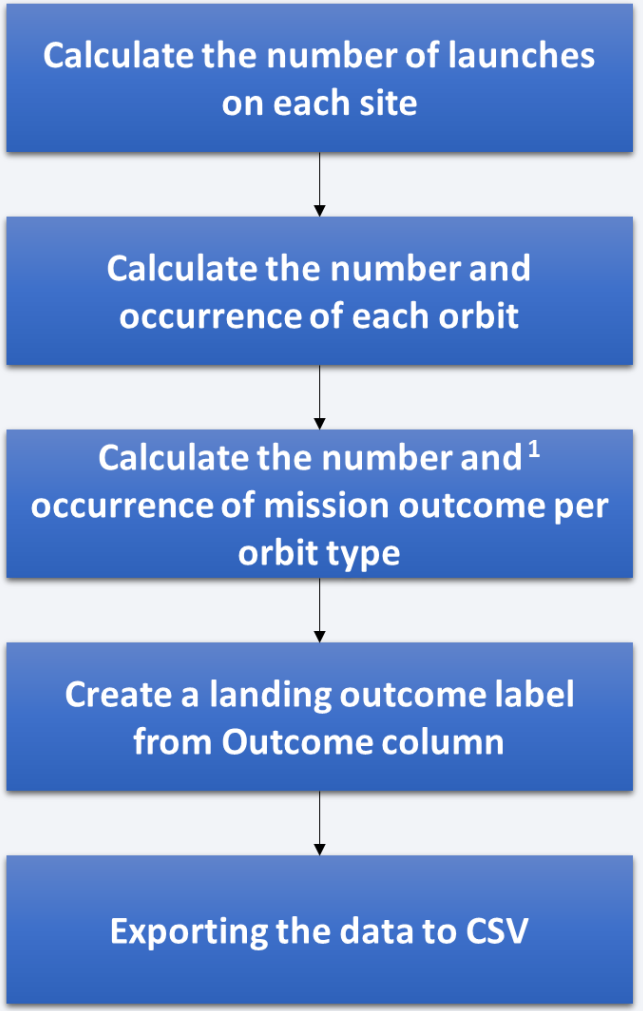
- The data was scraped from the Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- The website contains data only about Falcon 9 launches.
- BeautifulSoup was used to extract the launch records from the HTML table.
- The extracted table data was then parsed and converted into a Pandas DataFrame for further analysis.
- The final dataset consists of 121 rows and 11 columns.



GitHub: Web Scraping

Data Wrangling

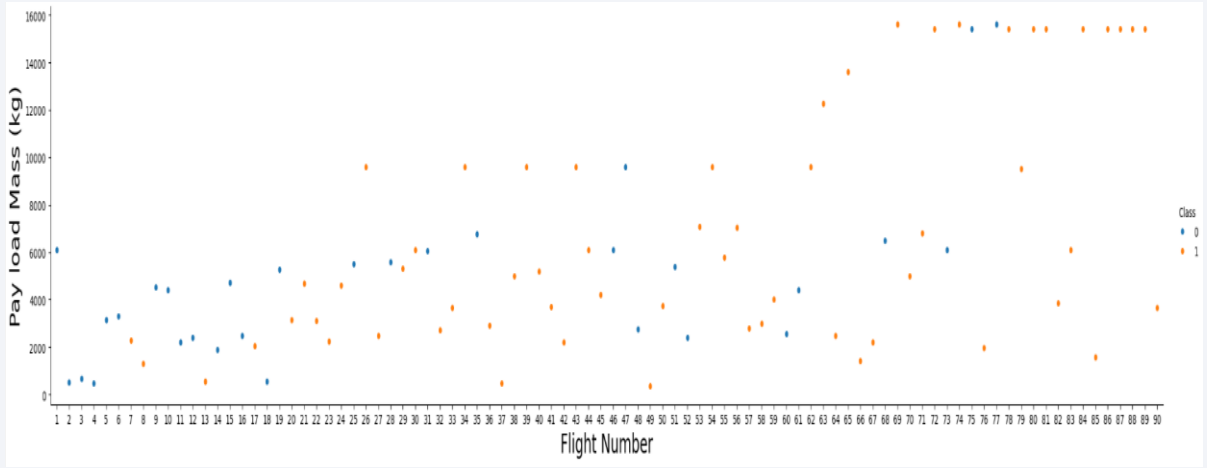
- Data wrangling is the process of cleaning and organizing messy, complex data to make it easier to access and analyze.
- In this project, we first calculate the number of launches for each site and then determine the frequency of mission outcomes for each orbit type.
- Next, we create a landing outcome label based on the outcome column. We convert the mission outcomes into training labels, where “1” indicates a successful booster landing and “0” indicates an unsuccessful landing.
- Finally, the cleaned and processed data is exported to a CSV file.



GitHub: Data Wrangling

EDA with Data Visualization

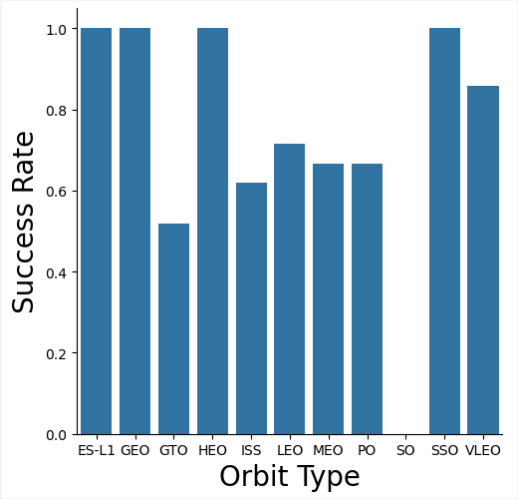
- We started the exploratory data analysis by using scatter plots to examine the relationships between various attributes. These included:
  - Payload and Flight Number
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit Type
  - Payload and Orbit Type
- Scatter plots are useful for identifying dependencies between attributes and once patterns are observed, they help determine which factors most affect the success of landing outcomes.
- If a strong relationship exists, these variables could be valuable in building machine learning models.

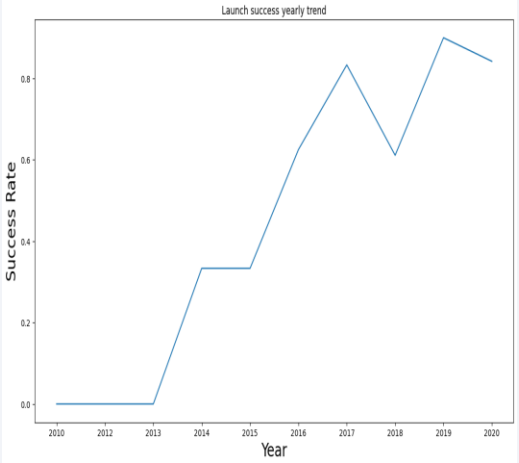


GitHub: Data Visualization

EDA with Data Visualization

- We used bar charts to compare discrete categories and identify orbit types with the highest success rates. These charts help visualize relationships between specific categories and their measured values.
- Line charts were employed to reveal trends in the data over time, specifically focusing on yearly trends in launch success rates. They provide insights into how success rates change over the years.
- After analyzing relationships through scatter and bar charts, we performed feature engineering by creating dummy variables for categorical columns. This prepares the data for future machine learning models to predict landing success.





GitHub: Data Visualization

EDA with SQL

We performed the following SQL queries to get a better understanding of the dataset:

- Displaying the unique names of the launch sites.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch sites names for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub: EDA with SQL

Build an Interactive Map with Folium

- To visualize the launch data, we created an interactive map with circle markers for each launch site using their latitude and longitude coordinates, labeled with the site names.
- We used colored markers (Green for success and Red for failure) to indicate launch outcomes and grouped them in a MarkerCluster() to assess success rates across sites.
- Additionally, we calculated the distances between launch sites and nearby landmarks, such as railways, highways, coastlines and cities, using Haversine's formula to explore the proximity of launch sites to these features.
- After completing the above, we were able to identify some geographical patterns related to the launch sites.

GitHub: Folium

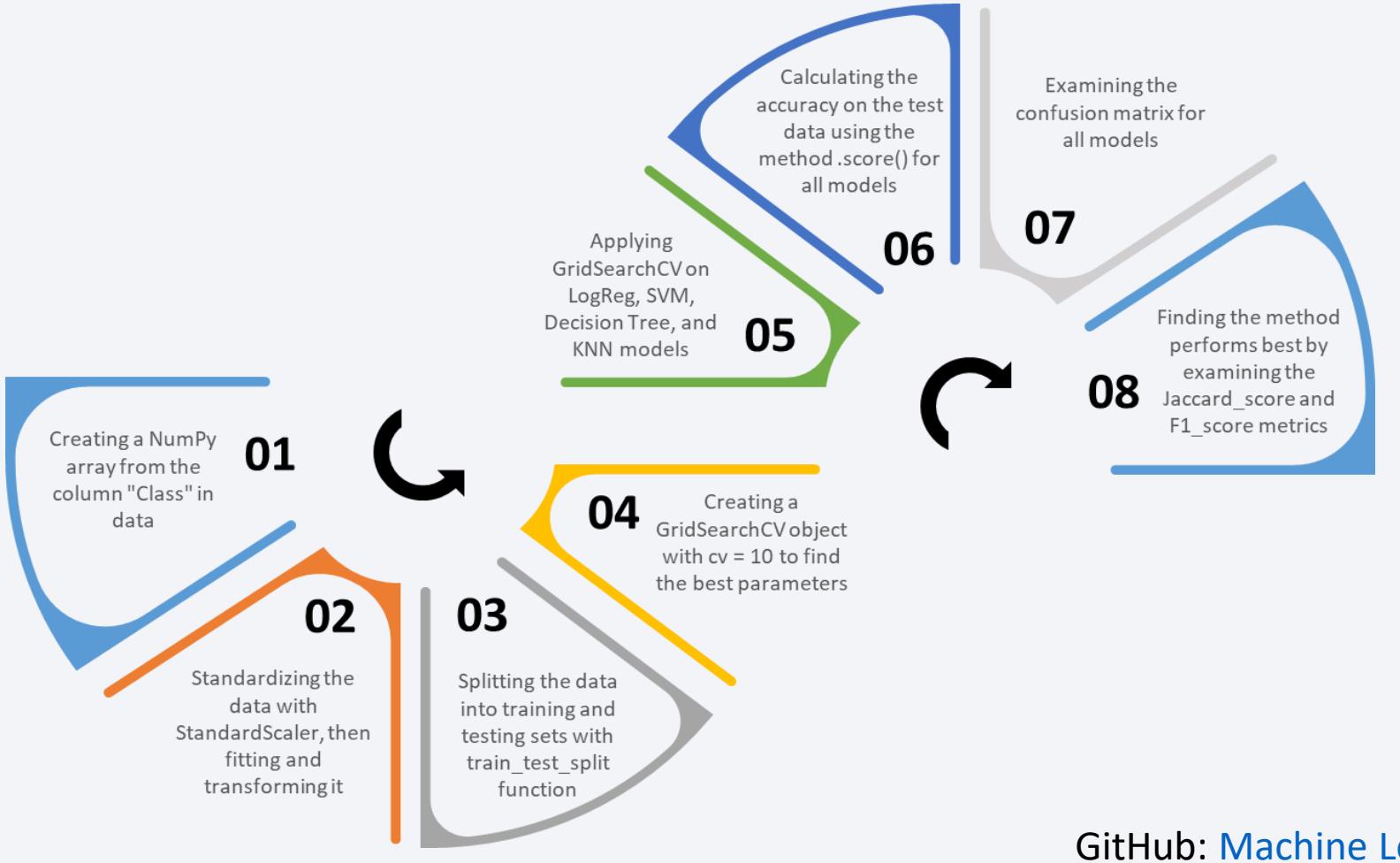
Build a Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly Dash, allowing users to explore the data as needed.
- Added a dropdown list for selecting launch sites and a pie chart to show total successful launches, as well as success vs. failure counts for a specific site.
- Included a slider to select a payload mass range and a scatter chart to visualize the correlation between payload mass and launch success for different booster versions.
- Plotted additional pie charts and scatter graphs to display total launches by site and the relationship between outcome and payload mass for different booster versions.

GitHub: Plotly

Predictive Analysis (Classification)

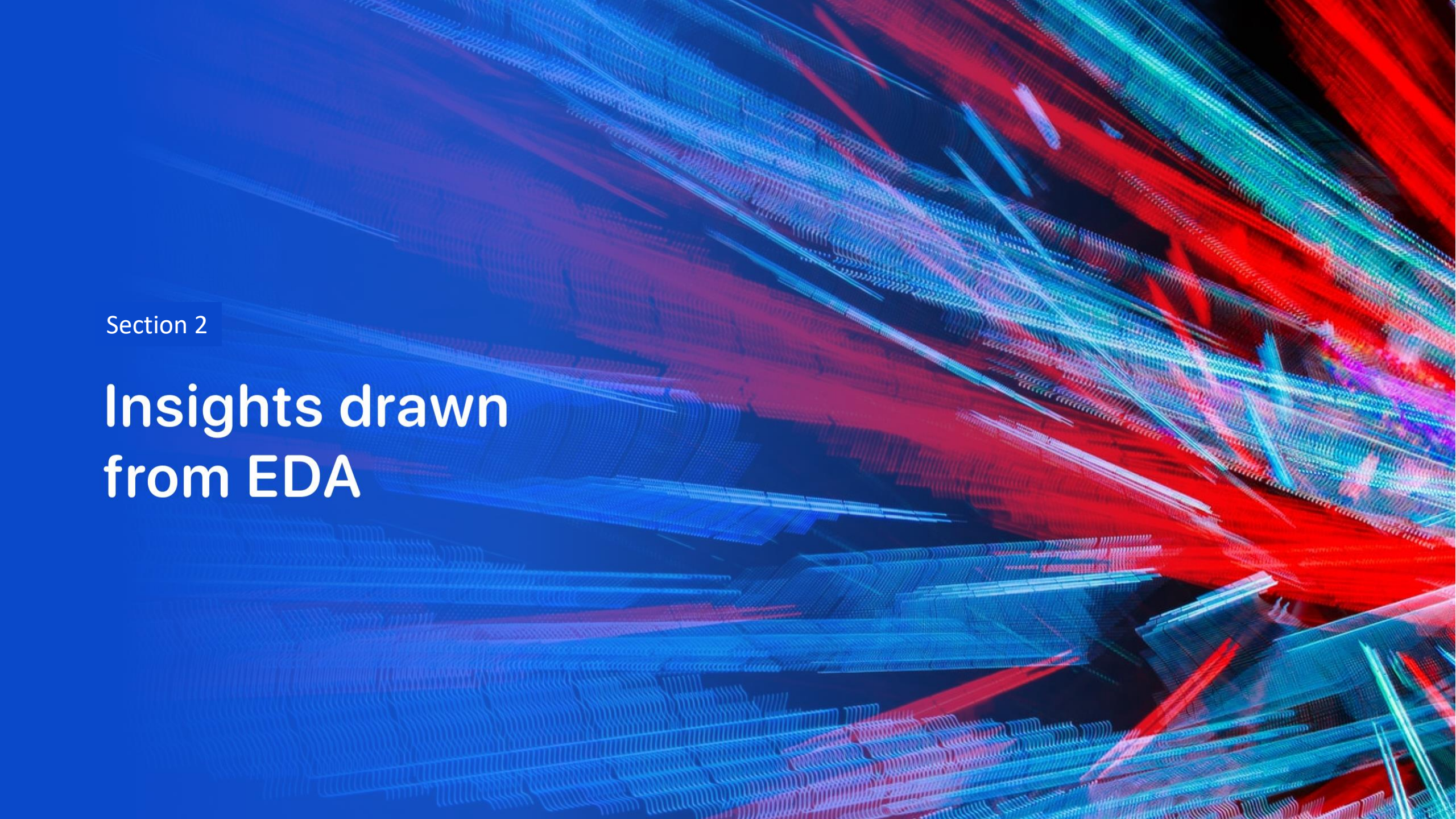
- Functions from the Scikit-learn library were used to create our machine learning models:



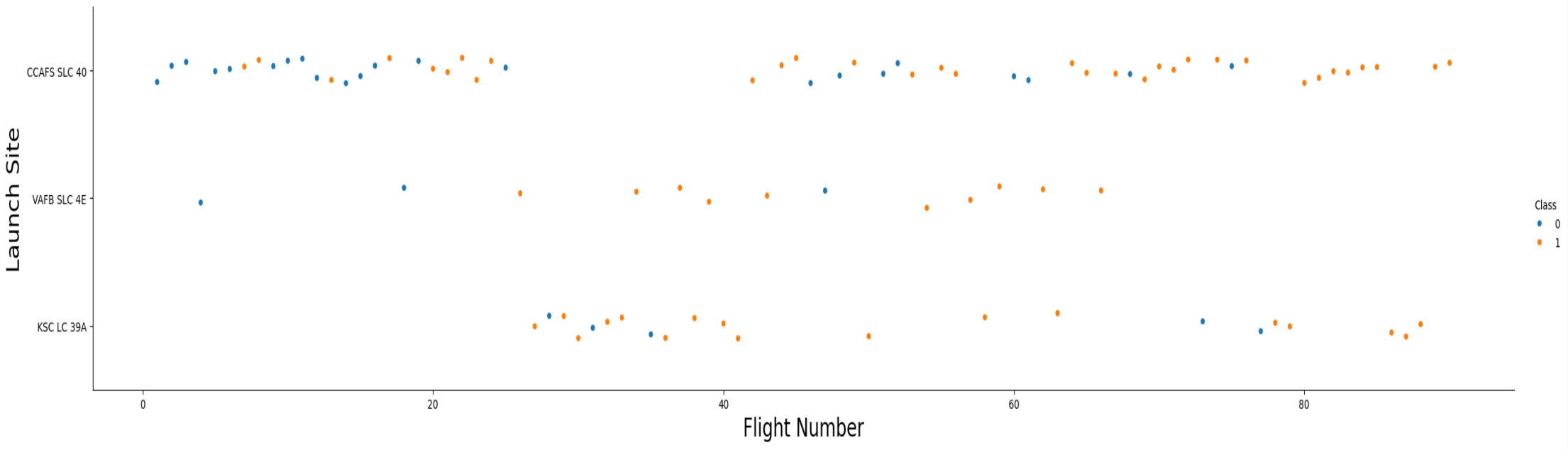
GitHub: [Machine Learning]

Results

- The results are split into 4 sections:
    1. Exploratory Data Analysis
        i. SQL
        ii. Visualization
    2. Folium
    3. Plotly Dash
    4. Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

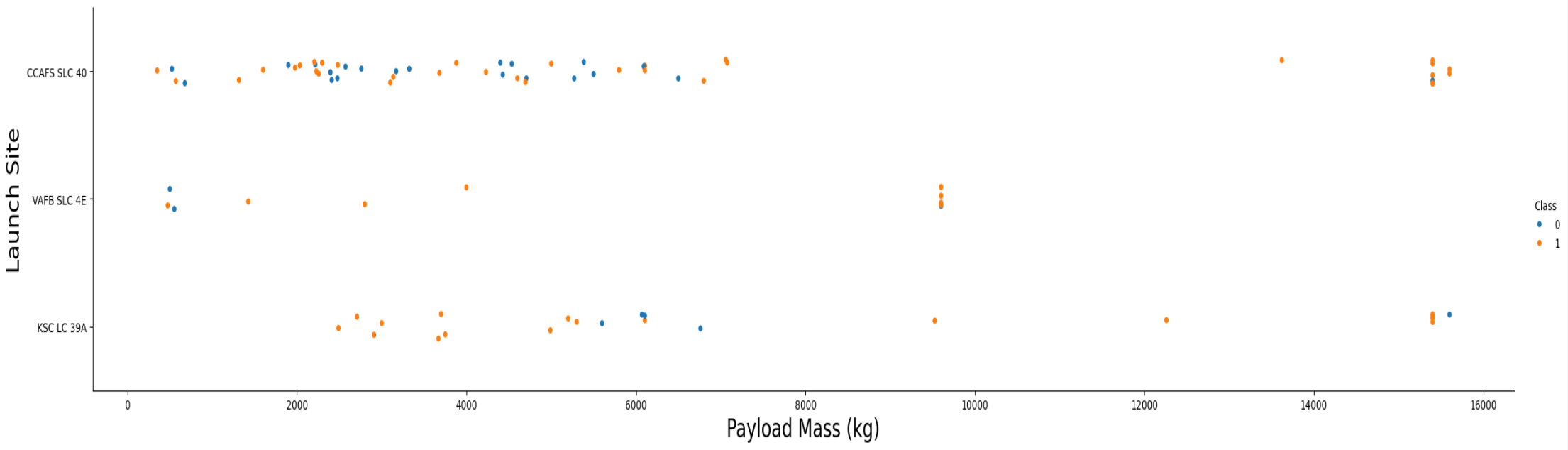

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- 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.

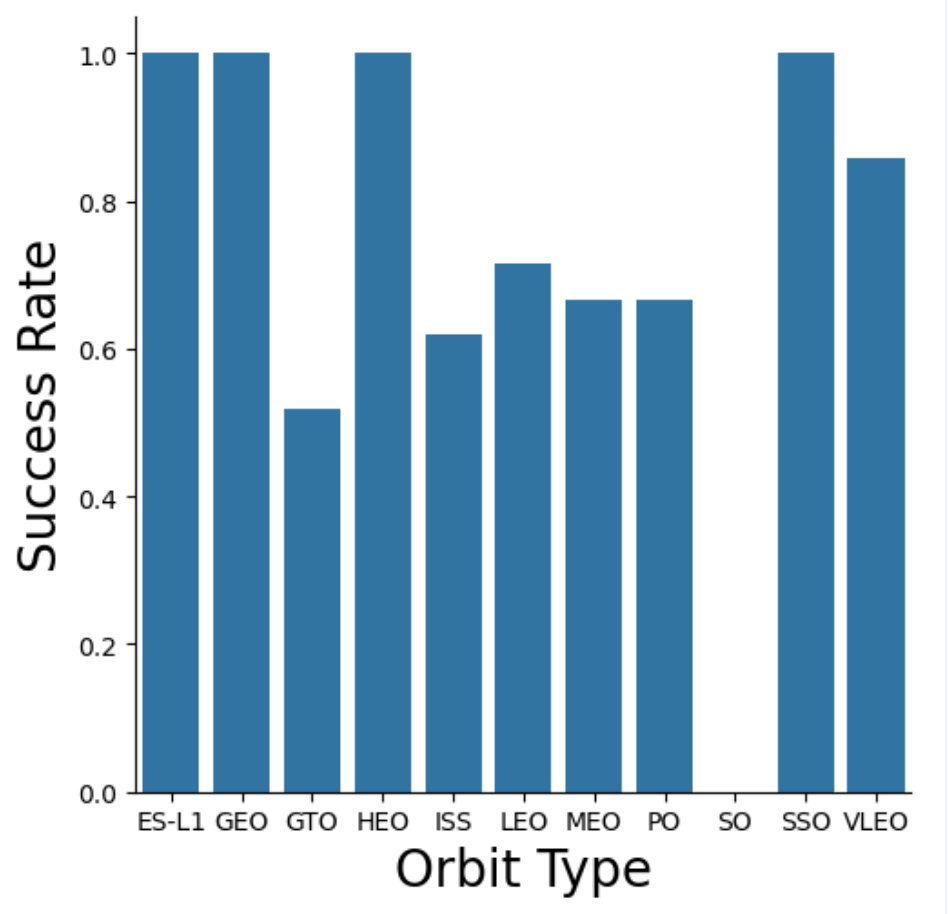
Payload vs. Launch Site



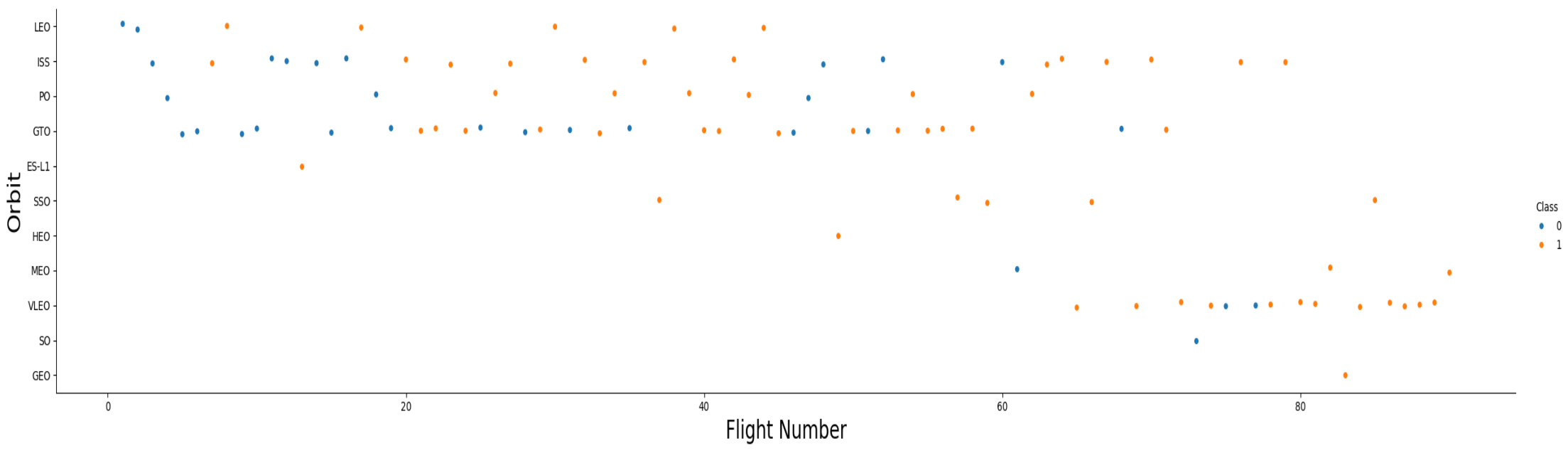
- This scatter plot shows once the pay load mass is greater than 7000kg, the probability of the success rate will be highly increased.
- For every launch site the higher the payload mass, the higher the success rate.

Success Rate vs. Orbit Type

- **Orbits with a 100% success rate:** These orbits, including ES-L1, GEO, HEO and SSO, have consistently shown successful launch outcomes without any failures.
- **Orbits with a 0% success rate:** The SO orbit has had no successful launches, with all attempts resulting in failure.
- **Orbits with a success rate between 50% and 85%:** Orbits such as GTO, ISS, LEO, MEO and PO have a mixed track record, with success rates ranging from 50% to 85%, indicating a relatively higher number of failed launches compared to the successful ones.

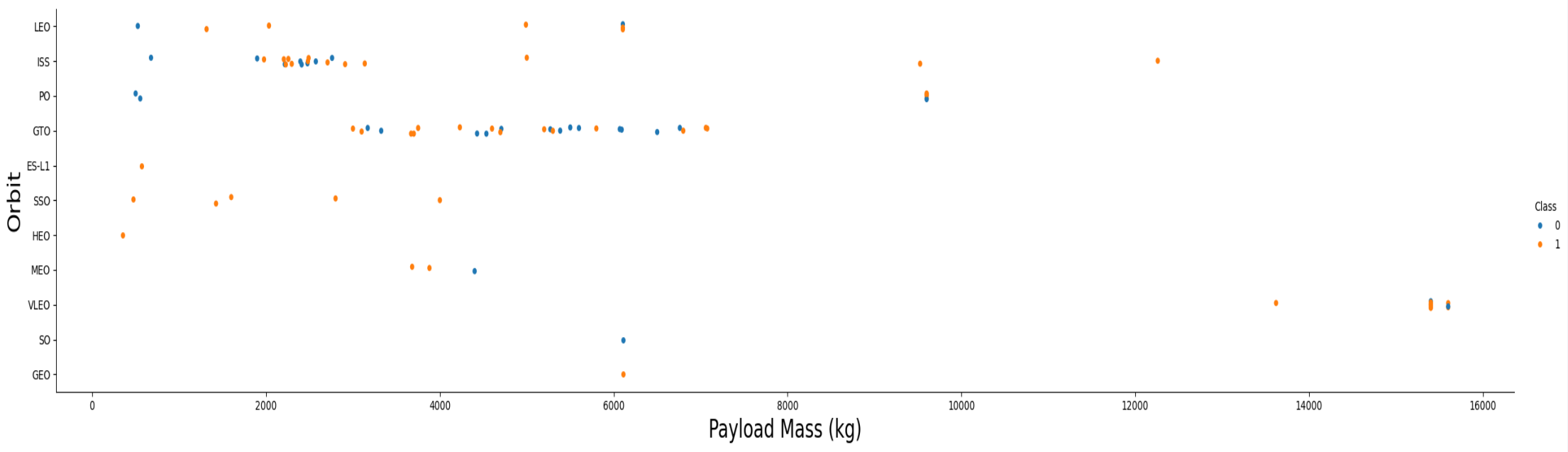


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights.
- On the contrary, there seems to be no relationship between flight number when in GTO orbit.

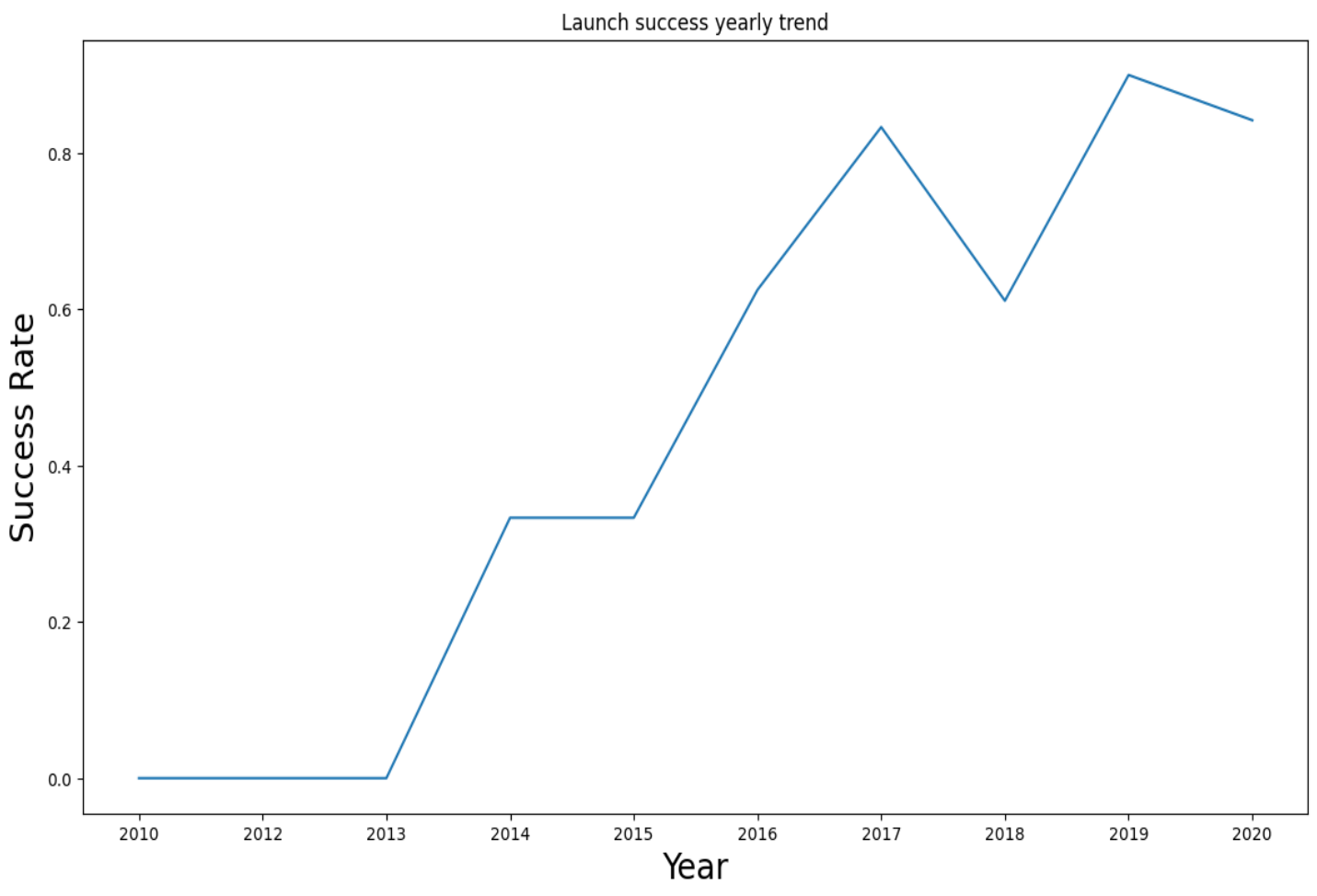
Payload vs. Orbit Type



- Heavy payloads have a positive impact on LEO, ISS and PO orbits, while they have a negative impact on MEO and VLEO orbits.
- GTO orbits show no significant relationship between payload weight and launch success or failure.

Launch Success Yearly Trend

- We observe that over time we see an increase in success rate, with 2017-2018 being an exception.



All Launch Site Names

We used the DISTINCT keyword for displaying the unique launch sites.

```
In [13]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[13]:

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

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with `CCA` using LIKE keyword:

In [14]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[14]:

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

- Calculating the total payload carried by boosters from NASA using the SUM() function.

```
In [15]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

Out[15]:

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1 using AVG() function.

```
In [16]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload" FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Average payload
2928.4

First Successful Ground Landing Date

- Listing the date when the first successful landing outcome in ground pad was achieved.

In [18]:
```
%sql SELECT MIN(DATE) AS "First successful landing outcome in ground pad" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (
```

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

Out[18]:

First successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [21]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000
 * sqlite:///my_data1.db
Done.
```

Out[21]:

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 used wildcard like ‘%’ to filter for WHERE MissionOutcome was a success or a failure

List the total number of successful and failure mission outcomes

In [23]:
```
%sql SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM
```

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

Out[23]:

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

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

```
In [24]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[24]:

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

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

In [30]:
```
%%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[30]:

month	Booster_Version	Launch_Site
01	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

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

In [31]:
```
%sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS Landing_Count FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '20
```

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

Out[31]:

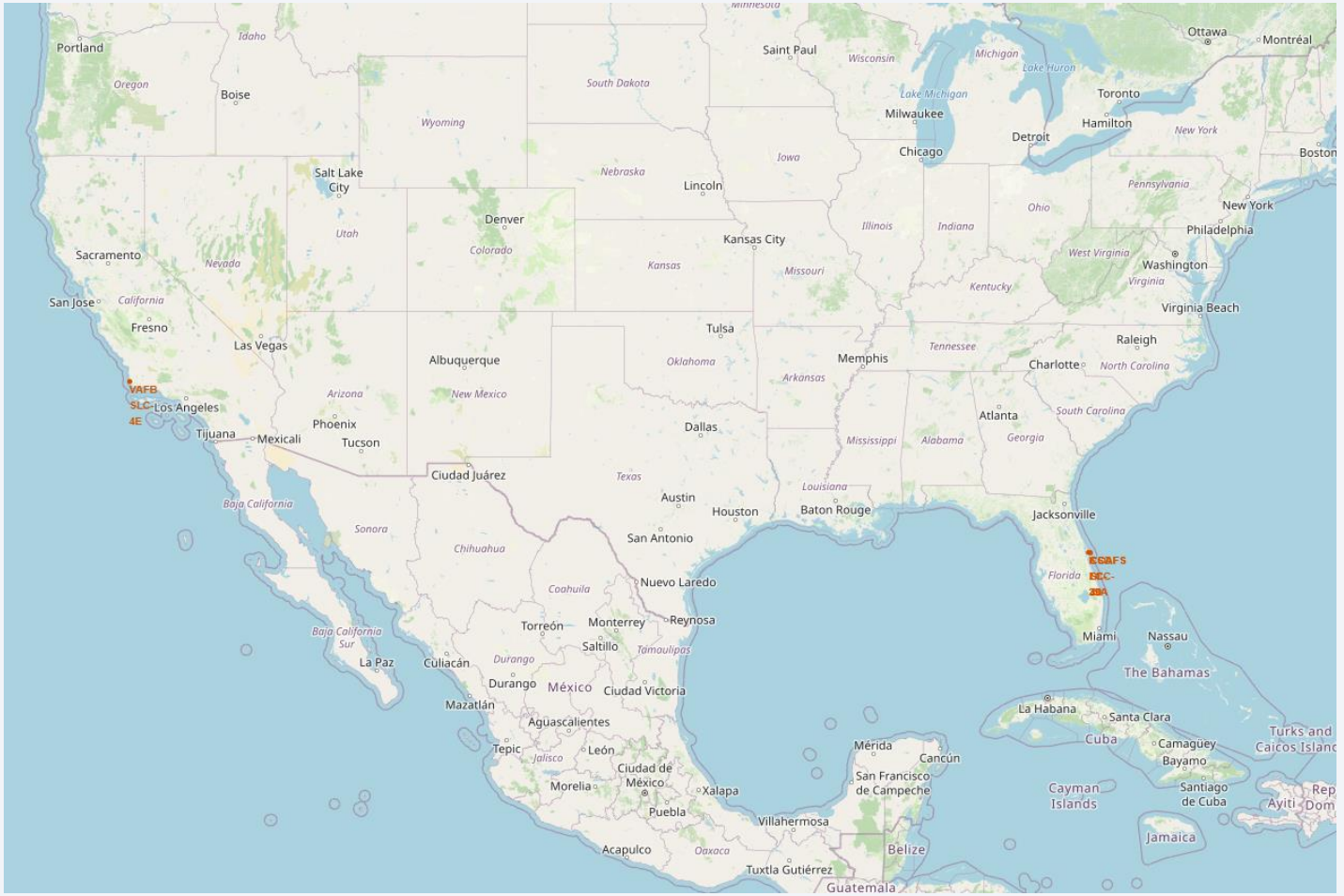
Landing_Outcome	Landing_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

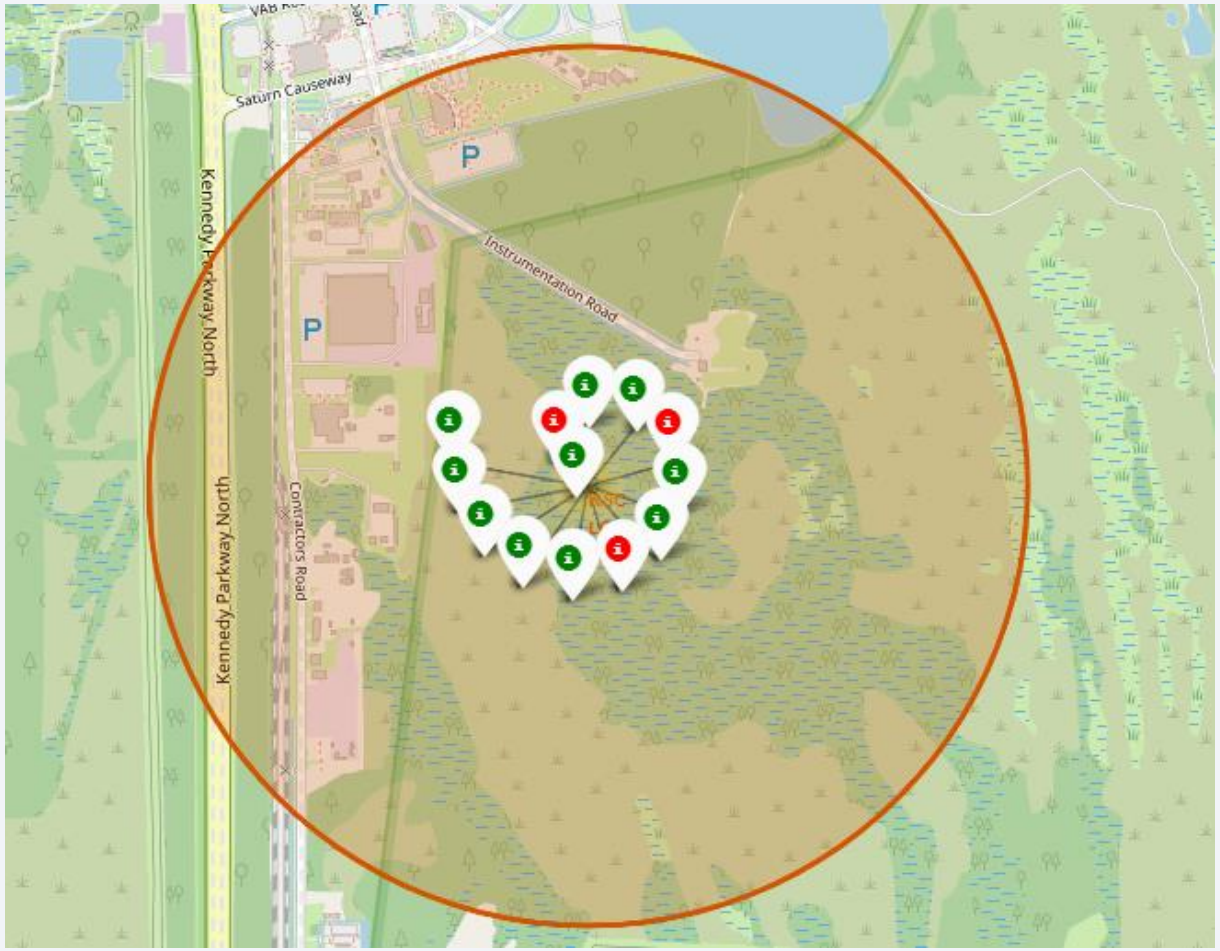
All launch sites' location markers on a global map

- All SpaceX launch sites are located within the United States.
- The launch sites are in close proximity to the coast, reducing the risk of debris falling or exploding near populated areas by launching rockets over the ocean.



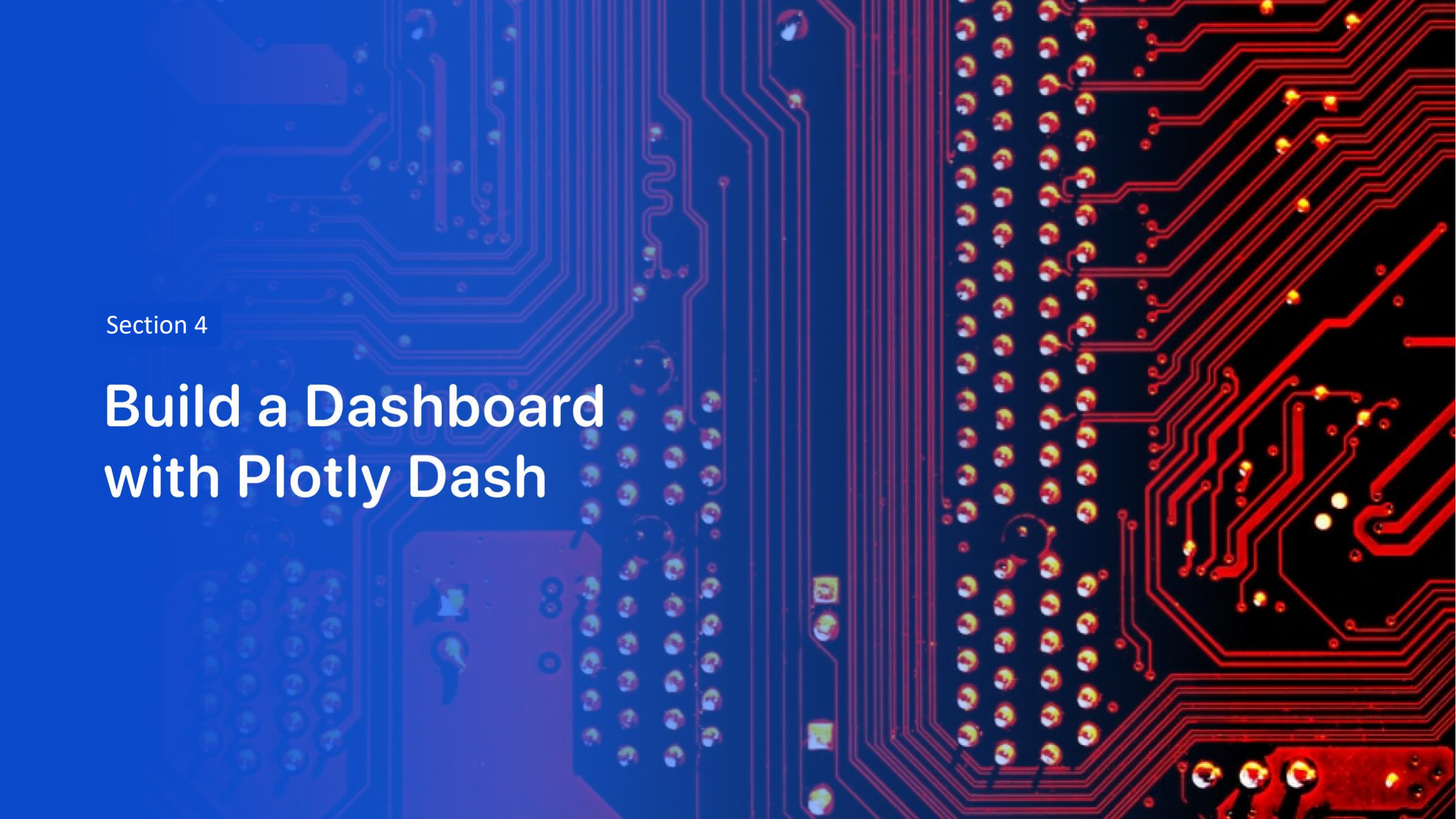
Color-labeled launch records on the map

- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
    1. Green Marker = Successful
    2. Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate

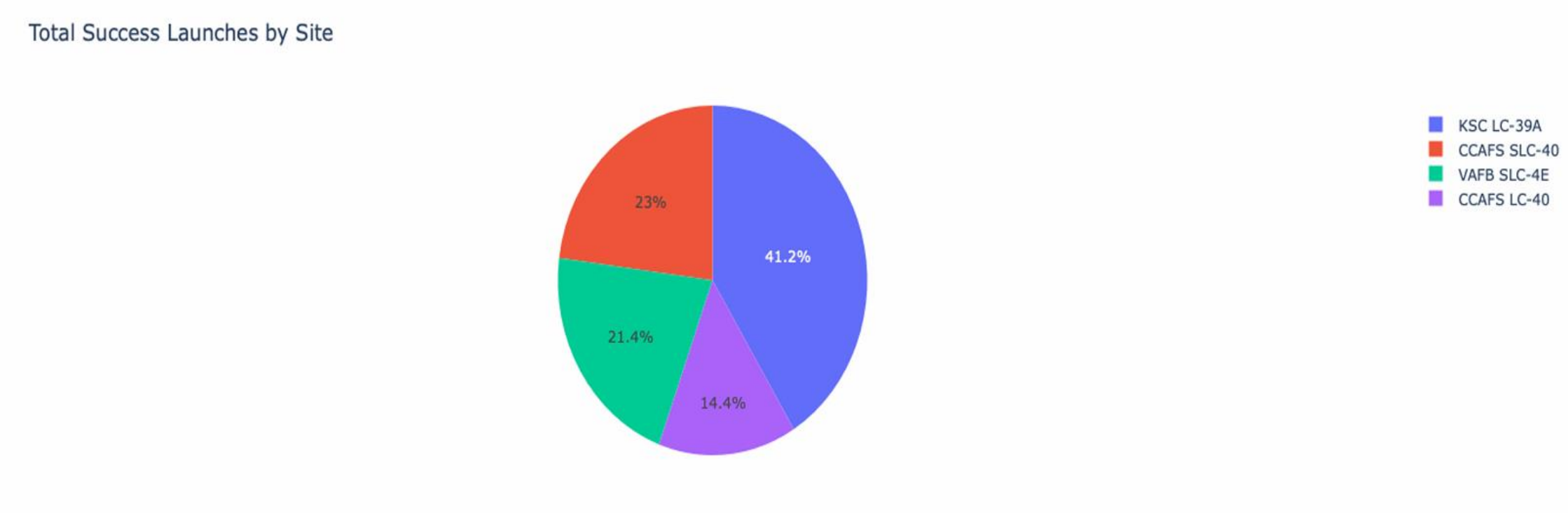


Distance from the Launch Site

- From the map we see that:
    1. Launch sites are not in close proximity to railways. No
    2. Launch sites are sometimes in close proximity to highways.
    3. Launch sites are in close proximity to coastline?
    4. Launch sites keep certain distance away from cities.

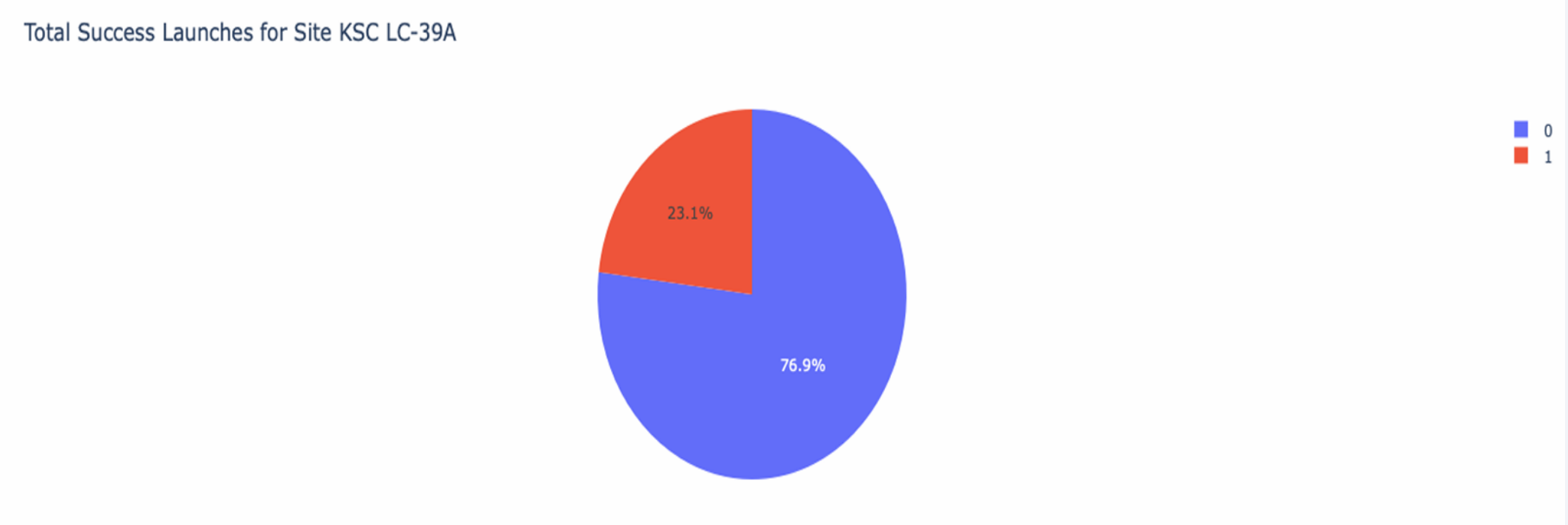

Section 4
Build a Dashboard with Plotly Dash

Site wise distribution of Successful Launch



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

Launch site with highest launch success ratio



- KSC LC-39A has the highest launch success rate (76.9%).

Payload vs Launch Outcome Scatter Plot

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
best_score = {'Logistic regresssion': [logreg_cv.best_score_], 'SVM': [svm_cv.best_score_], 'Decision tree': [tree_cv.best_score_], 'KNN': [knn_cv.best_score_]}
df = pd.DataFrame.from_dict(best_score, orient='index', columns=['Best scores'])
df
```

[32]

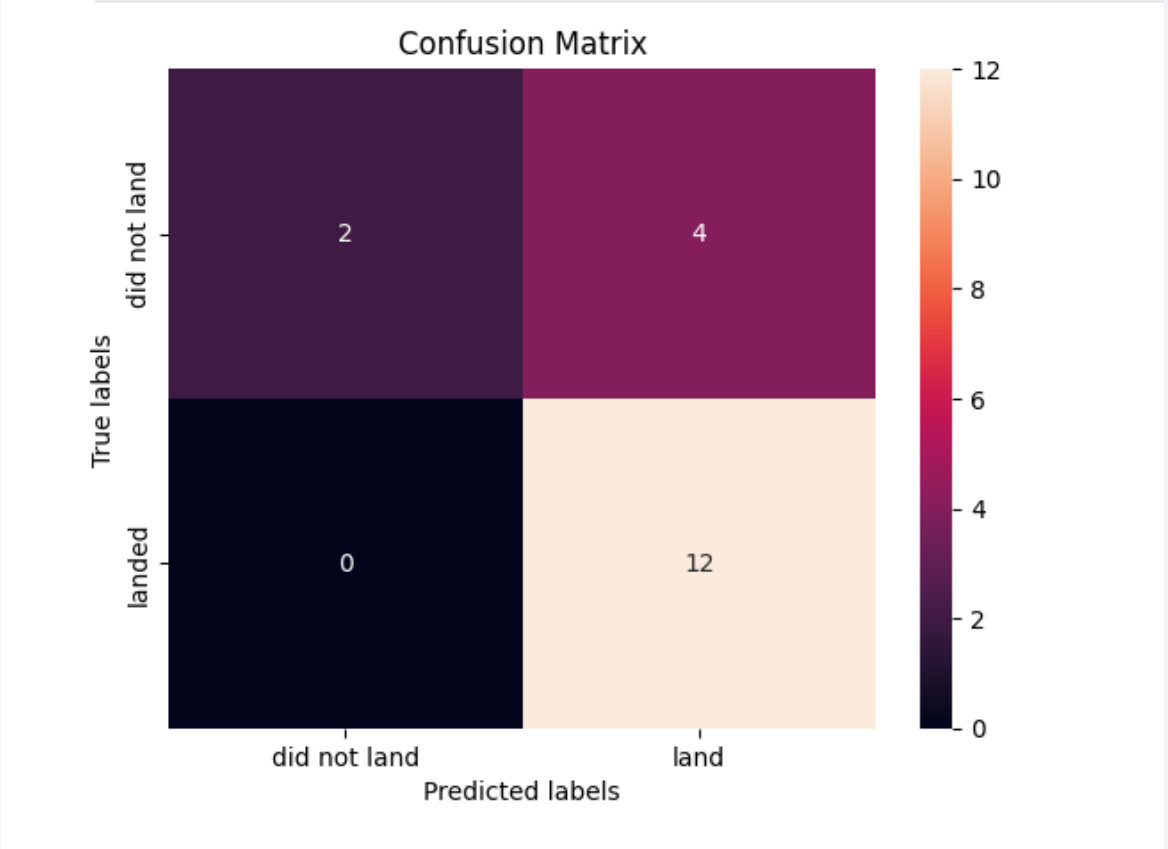
Python

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.873214
KNN	0.848214

- Using the above code snippet, we could identify that the best algorithm to be the Decision Tree Algorithm which has the highest classification accuracy.

Confusion Matrix

- Examining the confusion matrix, we see that Decision Tree can distinguish between the different classes.
- We see that the major problem is false positives.



Conclusions

- The Decision Tree algorithm is the best model for this dataset, providing the most accurate predictions.
- Launches with low payload masses (4000kg and below) tend to have higher success rates compared to those with heavier payloads.
- The success rate of SpaceX launches has increased steadily since 2013, with a clear upward trend continuing toward 2020 and beyond, indicating improved performance over time.
- KSC LC-39A has the highest success rate among all launch sites, with 76.9% successful launches.
- SSO orbit has a 100% success rate, with more than one occurrence, showing the best consistency in successful launches.
- Most launch sites are located near the Equator and in close proximity to the coast, minimizing risks during launches.

Thank you!