IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jones Wong
18th February 2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceXAPI(https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Data was cleaned, analyzed, and enriched by creating a landing outcome label, one-hot encoding categorical features, and finding patterns to train supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - A machine learning pipeline was created to predict first-stage landing using normalized data, dividing it into training and test sets, evaluating with four classification models, and training the best-performing model for accurate predictions.

Data Collection

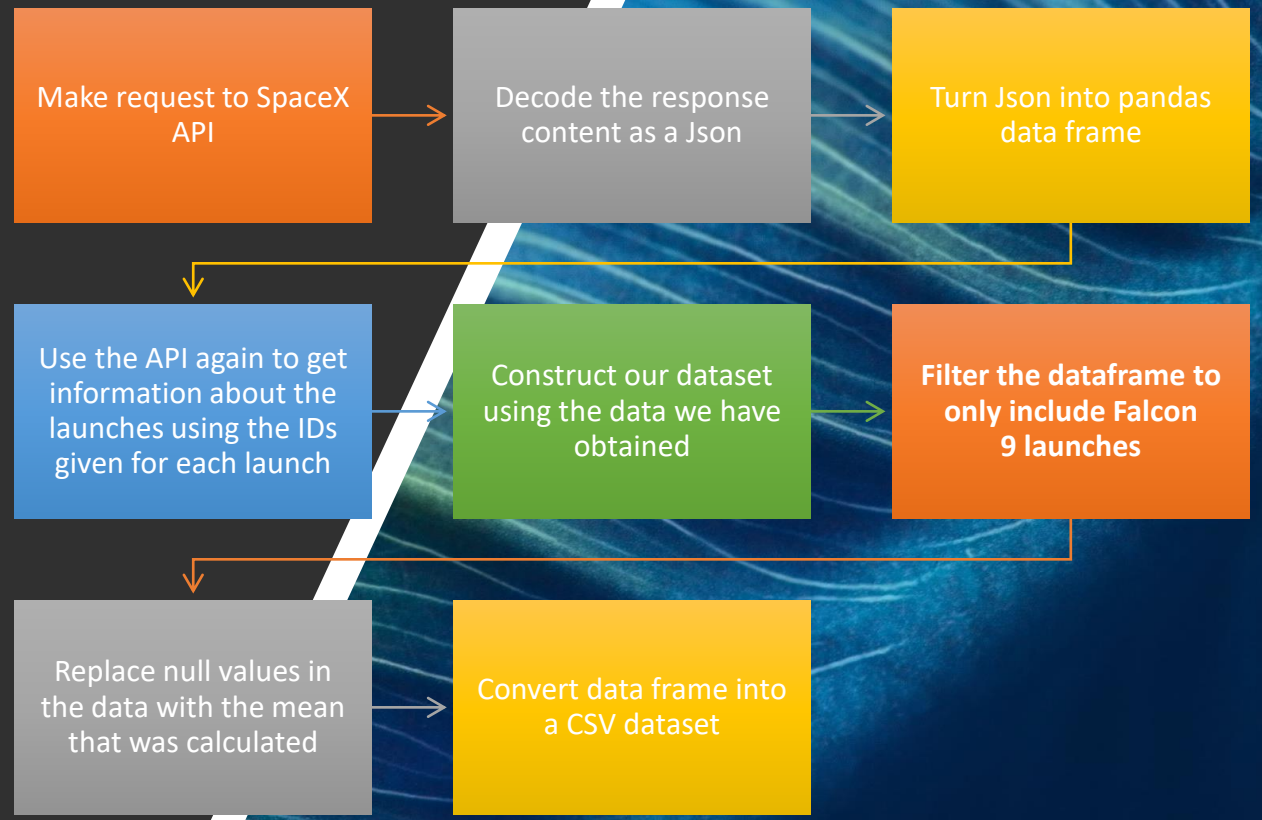
The data collection stage is essential in this project as it provides the input data to train the machine learning models to make accurate predictions. Two methods were used to collect data, REST API request and Web Scraping, which are cost-effective and only require a stable internet connection.

For the REST API request, the GET request was sent to retrieve the data, which was then decoded as Json and converted to a pandas data frame using the json_normalize() function. The data was cleaned and missing values were filled in to ensure completeness and accuracy.

For web scraping, the BeautifulSoup library was used to extract the launch records from Wikipedia as an HTML table. The table was then parsed and converted to pandas dataframe for further analysis and processing. These two methods of data collection were used to gather relevant information and answer relevant questions to evaluate outcomes and improve accuracy.

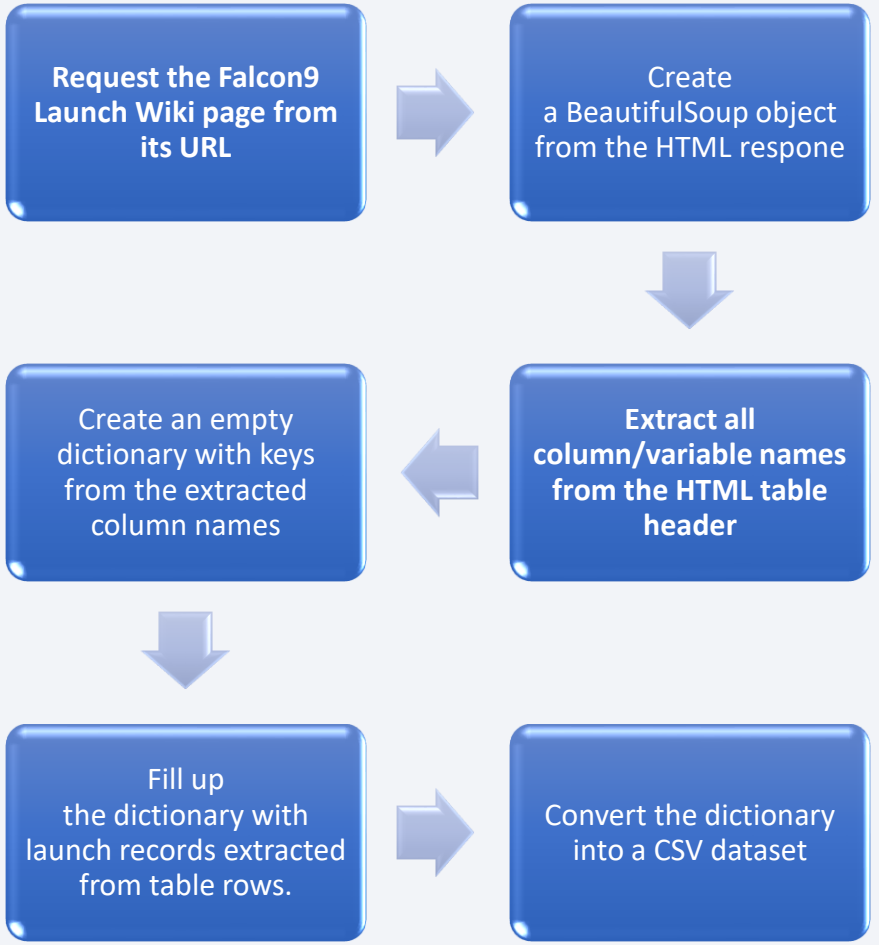
Data Collection – SpaceX API

- The SpaceX API was used to obtain data and persist it, following the flowchart. A request was made to the API to ensure that the data was in the correct format. Basic data wrangling and formatting were performed to clean the requested data, which was then converted into a CSV dataset by converting the data frame.
- [IBM-data-science-capstone/1-spacex-data-collection-api.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)]



Data Collection - Scraping

- Data from SpaceX launches was also obtained from Wikipedia following the flowchart. BeautifulSoup was used to perform web scraping on the Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches." The launch records were stored in an HTML table and parsed, then converted into a CSV dataset.
- [IBM-data-science-capstone/2-webscraping.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)]



Data Wrangling

- Data Wrangling involves cleaning and transforming complex data for analysis and machine learning. Unsuccessful rocket landing cases were transformed into training labels (1 for success, 0 for failure) through initial EDA. The landing outcome label was created from the Outcome column, with the goal of finding patterns in the data and determining the label for training models. String variables were transformed into categorical variables (1 for success, 0 for failure).
- [IBM-data-science-capstone/3-Data wrangling.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)]

Data cleaning - Removing or correcting any inaccurate or inconsistent data in the dataset.

Data transformation - Converting string variables into categorical variables (1 for successful mission, 0 for failure).

Exploratory Data Analysis (EDA) - Performing initial analysis on the dataset to find patterns and relationships.

Summary calculations - Calculating the summary launches per site and occurrences of mission outcome per orbit type.

Landing outcome label creation - Creating a landing outcome label from the Outcome column for easier analysis and machine learning.

Export to CSV - Saving the cleaned and transformed dataset in a CSV format for further analysis and modeling.

EDA with Data Visualization

We first analyzed the relationship between attributes using scatter plots and further visualization tools like bar graphs and line plots.

- Scatter plots showed the dependency of attributes on each other and helped determine the factors affecting the success of landing outcomes.
- Bar graphs were used to determine which orbits had the highest probability of success.
- Line graphs showed a trend or pattern of the attribute over time, in this case, the launch success yearly trend.
- Feature Engineering was performed by creating dummy variables into categorical columns for future success prediction.

IBM-data-science-capstone/5-eda-dataviz.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)

EDA with SQL

- SQL queries were performed:
  - Names of unique launch sites
  - Top 5 launch sites starting with "CCA"
  - Total payload mass carried by NASA (CRS)
  - Average payload mass carried by F9 v1.1 booster
  - Date of first successful landing outcome in ground pad
  - Boosters with successful drone ship landing and payload mass between 4000-6000 kg
  - Total number of successful and failed missions
  - Boosters with maximum payload mass
  - Failed landing outcomes in drone ship for 2015 with booster version and launch site names
  - Rank of landing outcomes (failure/success) between 2010-06-04 and 2017-03-20

IBM-data-science-capstone/4-eda-sql.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)

Build an Interactive Map with Folium

- Latitude and longitude coordinates for each launch site marked with circle markers and labeled with the launch site name
- Launch outcomes (failure/success) marked with Red/Green markers using MarkerCluster()
- Haversine's formula used to calculate the distance to landmarks (railways, highways, cities, coastlines)
- Markers, circles, lines, and marker clusters are used to visualize launch sites and distances to landmarks.

- IBM-data-science-capstone/6-launch_site_location.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)

Build a Dashboard with Plotly Dash

We built an interactive dashboard using Plotly dash that allows the user to interact with the data as needed. The dashboard includes:

- Dropdown component to select a launch site or all sites
- Pie chart displaying the success and failure rate of the selected launch site
- Rangeslider to select a specific payload mass range
- Scatter chart showing the relationship between success and payload mass for different booster versions.

- IBM-data-science-capstone/spacex_dash_app.py at main · joneswong96/IBM-data-science-capstone (github.com)

Predictive Analysis (Classification)

Data Preparation:

- Load dataset
- Normalize data
- Split data into training and test sets

Model Preparation:

- Select machine learning algorithms
- Set parameters using GridSearchCV
- Train models using the training dataset

Model Evaluation:

- Find the best hyperparameters
- Compute accuracy using the test dataset
- Plot Confusion Matrix

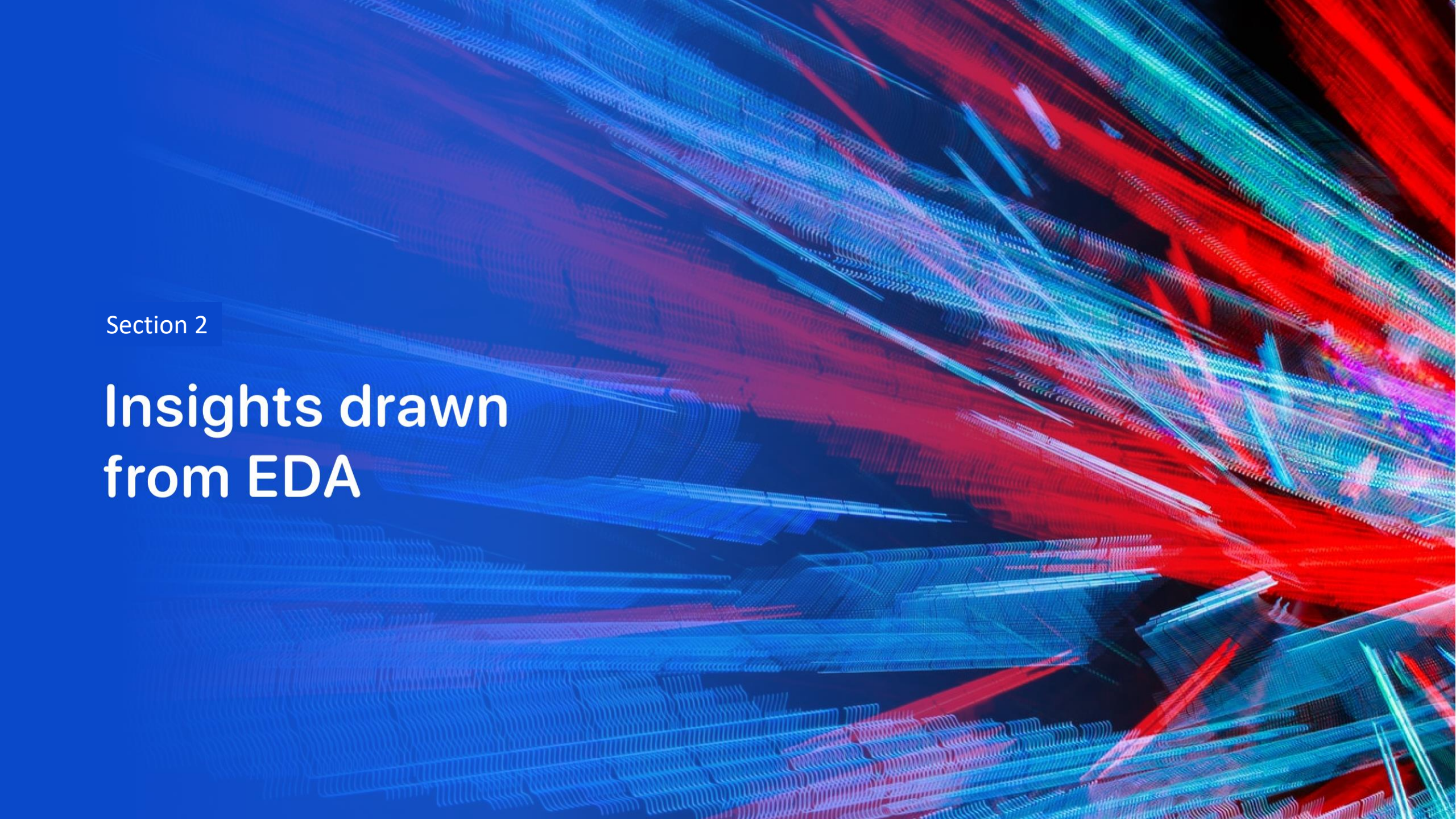
Model Comparison:

- Compare the accuracy of models
- Choose the model with the best accuracy

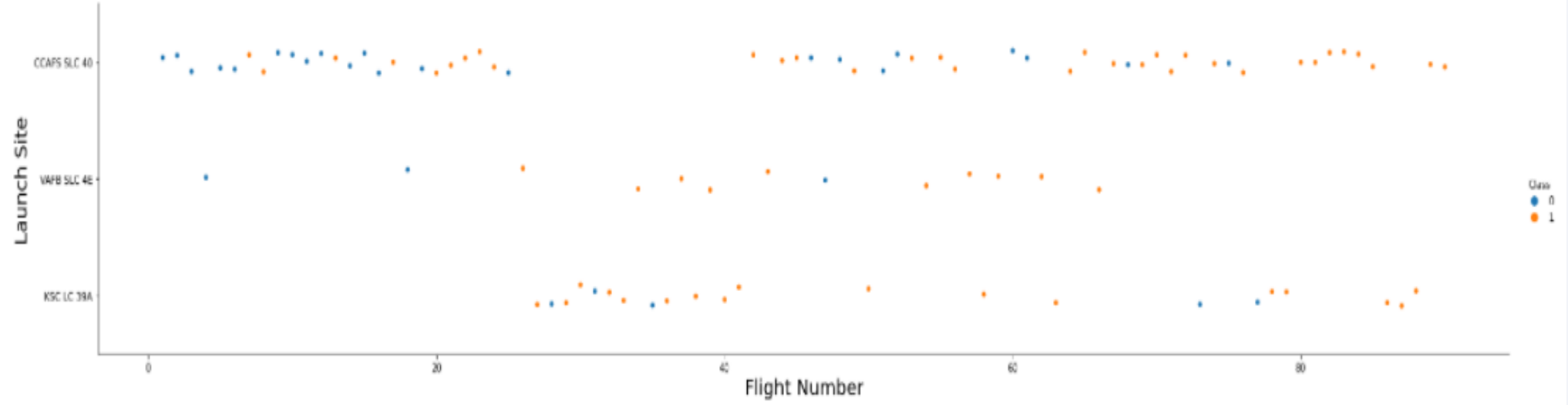
- IBM-data-science-capstone/7-Machine_Learning_Prediction.ipynb at main · joneswong96/IBM-data-science-capstone (github.com)

Results

- Space X has 4 launch sites and primarily launched to Space X and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first successful landing outcome was in 2015, 5 years after the first launch
- Most successful landing outcomes were seen with Falcon 9 booster versions with payload above average
- Majority of mission outcomes were successful
- F9 v1.1 B1012 and F9 v1.1 B1015 failed landing in 2015
- Landing outcomes improved over the years
- Launch sites near coastlines and with easy transportation access
- Machine learning models had an accuracy score of 83.33% for landing prediction

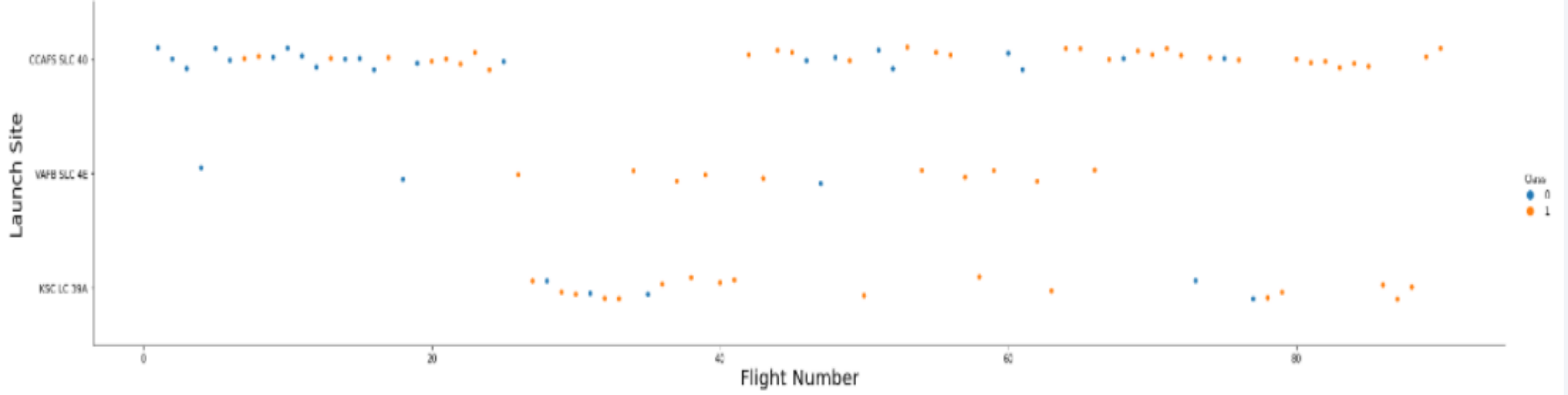

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



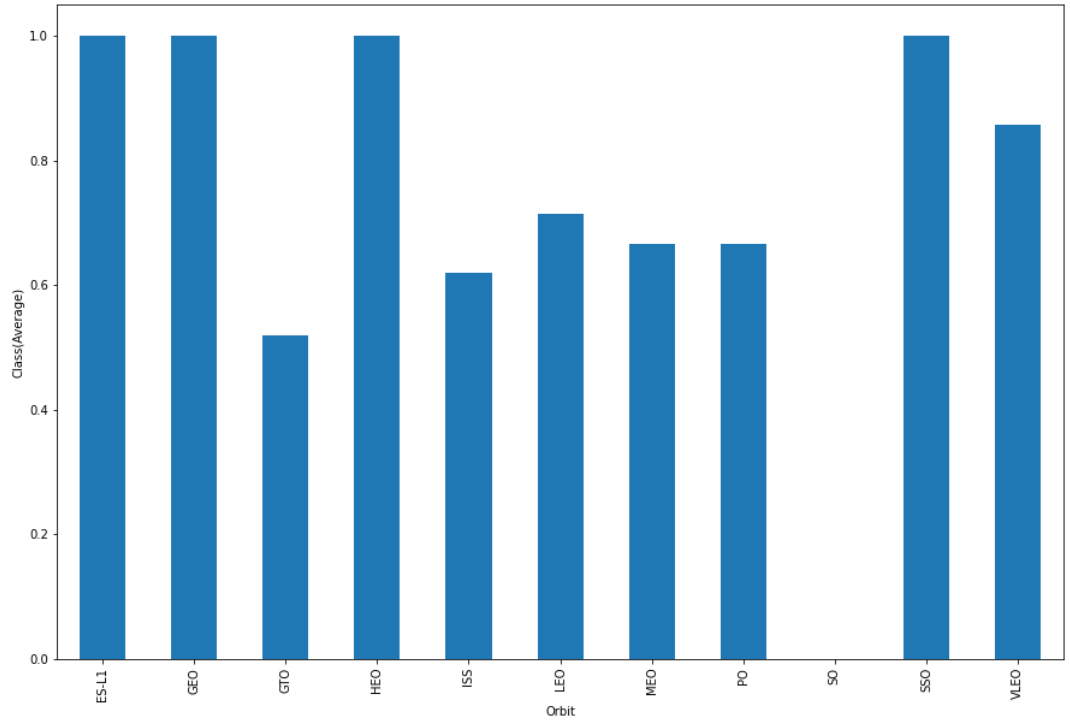
- The scatter plot indicates that as the number of flights at a launch site increases, the success rate also tends to increase, except for CCAFS SLC40.
- The chart also reveals that there were no launches with a heavy payload (greater than 10,000 kg) at the VAFB-SLC launch site.

Payload vs. Launch Site



- The launch site should be considered when determining the appropriate payload for a successful landing. However, if the payload is too heavy, it could result in a failed landing.
- The scatter plot indicates that a higher payload mass leads to a higher probability of success. However, the relationship between the launch site and payload mass for success rate is not clear.

Success Rate vs. Orbit Type

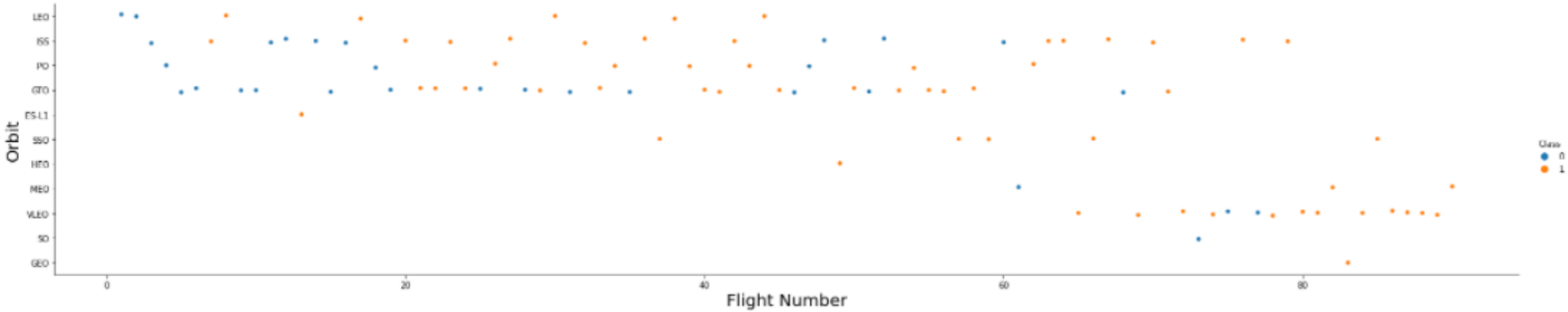


The orbit types **SSO**, **HEO**, **GEO** and **ES-L1** had the higghest success rate.

- The success rate of landing outcomes was influenced by the orbit types. SSO, HEO, GEO, and ES-L1 orbits had the highest success rate while the SO orbit produced 0% rate of success. However, further analysis showed that some of these orbits had only one occurrence, indicating the need for more data to draw conclusions.

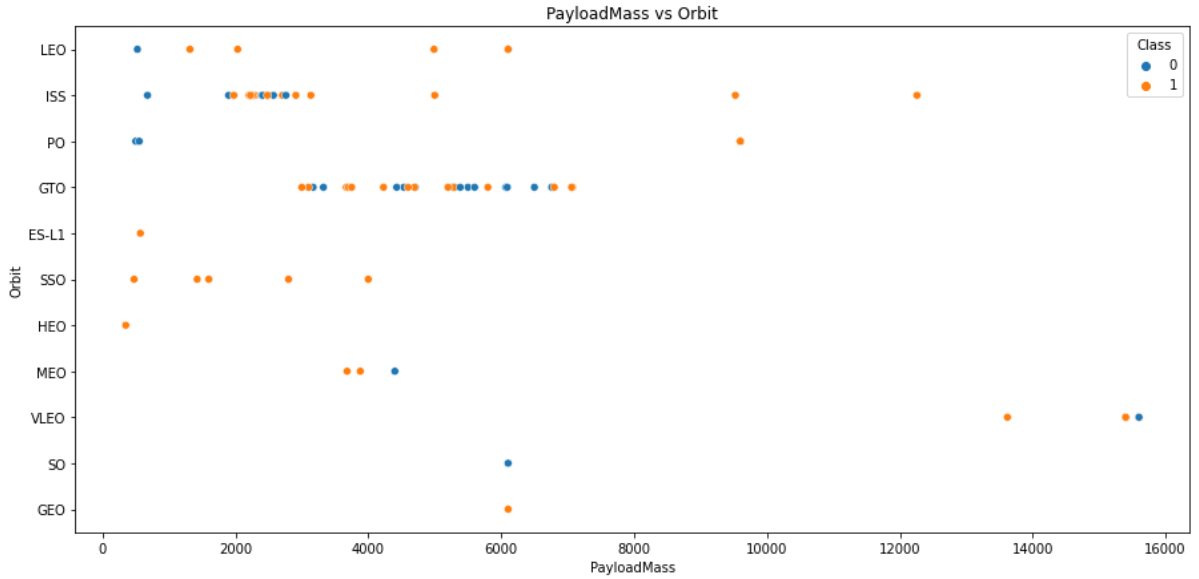
Flight Number vs. Orbit Type

- The plot displays Flight Number vs. Orbit type. It was observed that success in LEO orbit is related to the number of flights. In contrast, there seems to be no correlation between the number of flights and success rate in the GTO orbit. However, some orbits such as SSO or HEO had high success rates due to the knowledge gained from previous launches for other orbits.



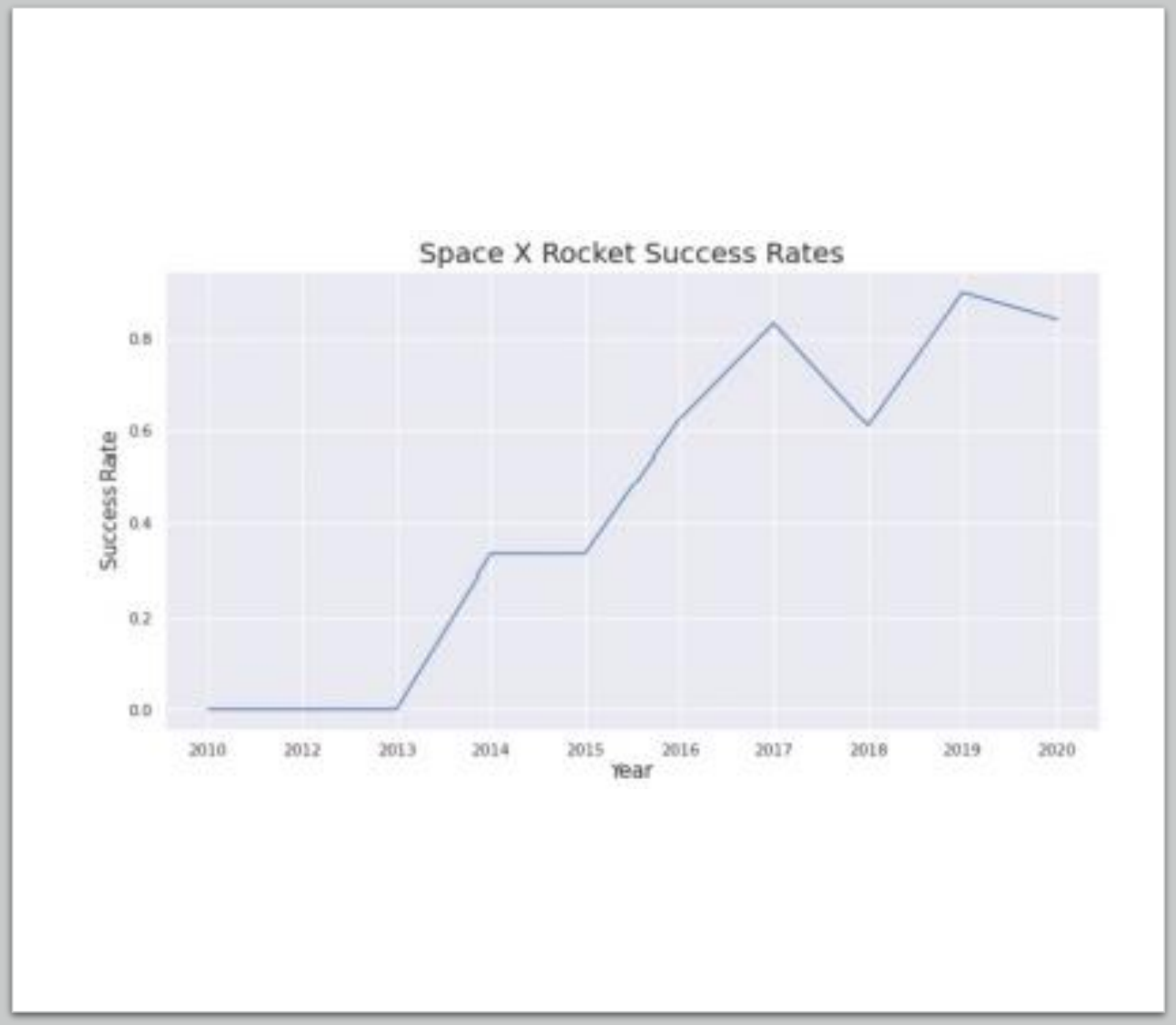
Payload vs. Orbit Type

- The analysis revealed patterns in the data, including correlations between certain launch sites, payload weights, and orbit types with successful landing outcomes. For example, heavier payloads improved success rates for certain orbits such as LEO, while decreasing payload weight for a GTO orbit increased success. The machine learning models were able to predict landing success with 83.33% accuracy, which could be improved with more data.



Launch Success Yearly Trend

- The figures show a clear increasing trend in success rate from 2013 to 2020, indicating a steady improvement. If this trend continues in the future, we may see a success rate of 100% in the coming years



All Launch Site Names

- The SELECT DISTINCT query was used to retrieve unique launch site names from the spacextbl.
- The launch sites where different rocket landings were attempted include
  CCAFS LC-40,
  CCAFS SLC-40,
  KSC LC-39A, and
  VAFB SLC-4E.
- The DISTINCT keyword helped to eliminate duplicate entries in the results.

Launch Site Names Begin with 'CCA'

- By using the WHERE clause followed by the LIKE clause with the substring "CCA", the query filters the launch sites that contain this substring. This limits the results to only five records.

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 attemp

Total Payload Mass

Customer	Total_Payload_Mass
NASA (CRS)	45596

- The payload mass carried by boosters from NASA was calculated to be 45596.
- The query calculates the sum of payload masses where the customer is NASA (CRS).

Average Payload Mass by F9 v1.1

- The average payload mass carried by the booster version F9 v1.1 was calculated to be 2928.4 kg.
- To obtain this value, a query was used to filter the data based on the booster version containing the substring F9 v1.1 and then calculate the average payload mass.

Average_Payload_Mass (kg)	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- The first successful landing on a ground pad took place on December 22, 2015, which was a major achievement for SpaceX and the world. This was observed by filtering data for successful ground pad landings and obtaining the minimum date value.

In [14]:
```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT BoosterVersion FROM SpaceX WHERE LandingOutcome = 'Success (drone ship)' AND PayloadMassKG > 4000 AND PayloadMassKG < 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- The SELECT statement is used to retrieve the booster version from the SpaceX dataset based on two filtering conditions. The WHERE clause filters for successful drone ship landings, and the AND clause further filters for payload mass between 4000 and 6000 kg. The resulting query returns the booster version that satisfies both conditions.

Total Number of Successful and Failure Mission Outcomes

- Using a query with the wildcard symbol '%', we filtered the SpaceX data for missions that were successful or had failed. We found that the majority of missions were successful except for one failure. To accomplish this, we used subqueries and the COUNT function, and applied a filter using the WHERE clause and LIKE clause

List the total number of successful and failure mission outcomes

In [16]:
```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

Out[16]:

The total number of successful mission outcome is:

	successoutcome
0	100

The total number of failed mission outcome is:

	failureoutcome
0	1

Boosters Carried Maximum Payload

- We identified the boosters that carried the maximum payload mass of 15600 kg using a subquery with MAX function. The main query then returns the distinct booster versions that carried this maximum payload.
- This query returns a list of 12 boosters with their respective booster versions that carried the maximum payload mass. Since the booster version names are similar, it is possible that they were manufactured by the same company.

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

In [17]:

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- By combining the WHERE clause, LIKE, AND, and BETWEEN conditions, we filtered the data to show failed landing outcomes on drone ship, booster versions, and launch site names for the year 2015.
- Two boosters failed to land at the beginning of the year, while the first successful landing occurred in December of the same year.

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [18]:
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Using the WHERE clause, we filtered the data to show landing outcomes between two specific dates. Then, we selected the landing outcomes and counted the occurrences.
- By grouping the results using the GROUP BY clause, we were able to see the frequency of each outcome.
- Finally, we ordered the results in descending order using the ORDER BY clause.

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

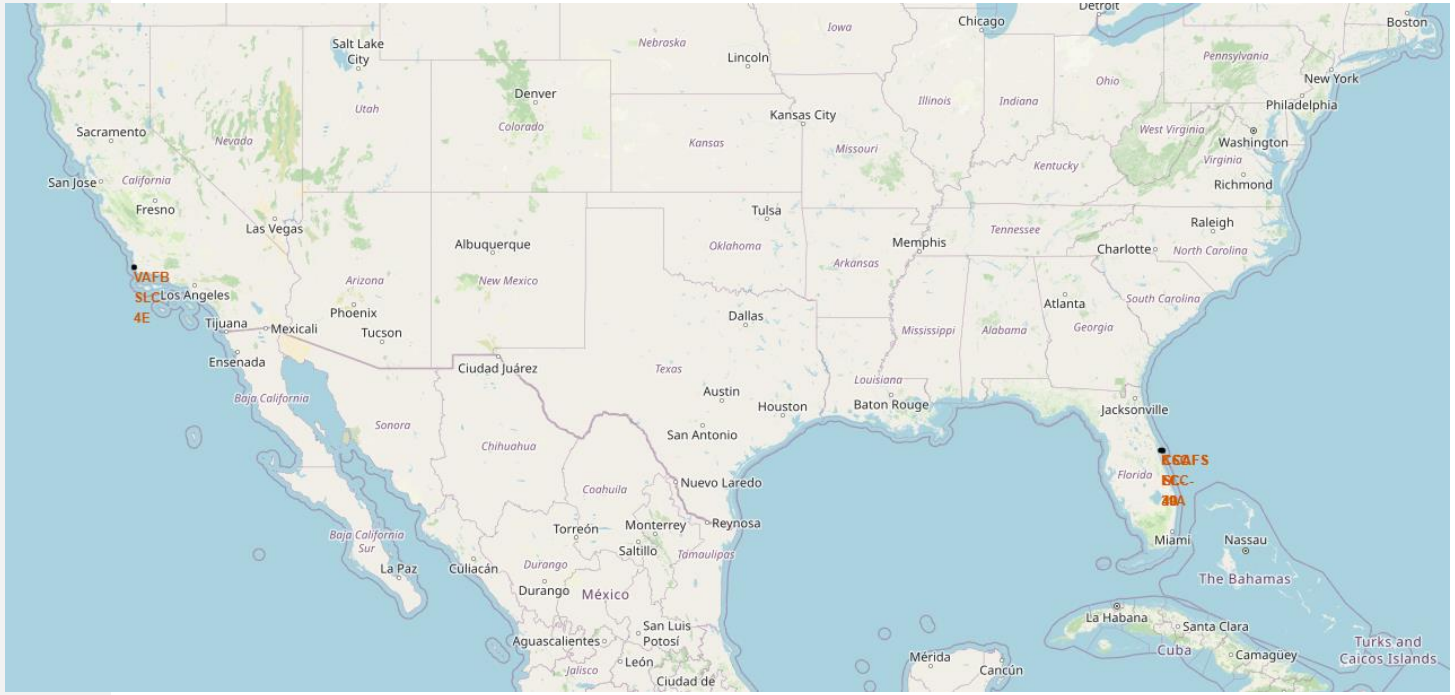
In [19]:
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

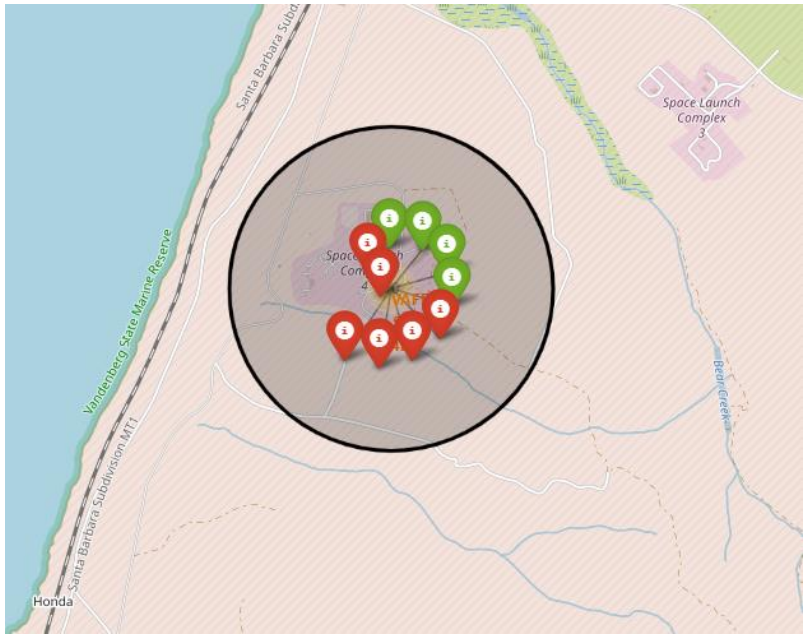

Section 3
Launch Sites
Proximities Analysis

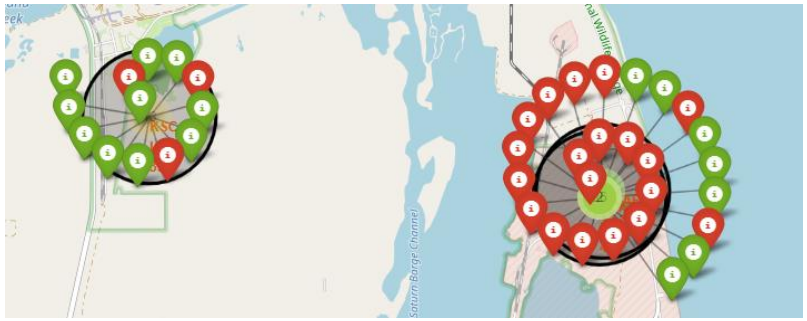
Location of all the Launch Sites



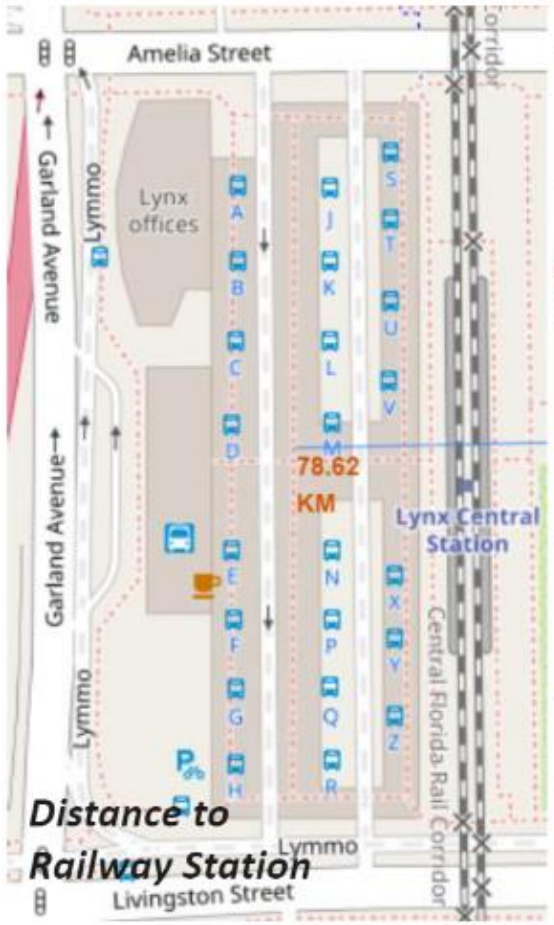
- We can see that all the SpaceX launch sites are located inside the United States
- Not all launch sites are in close proximity to the Equator line, but most of them are. Most launch sites are also close to the coast, but there are a few exceptions. This is because launching from sites closer to the Equator can take advantage of the Earth's rotation to gain additional velocity, and launching from sites near the coast can provide safety and logistical advantages.

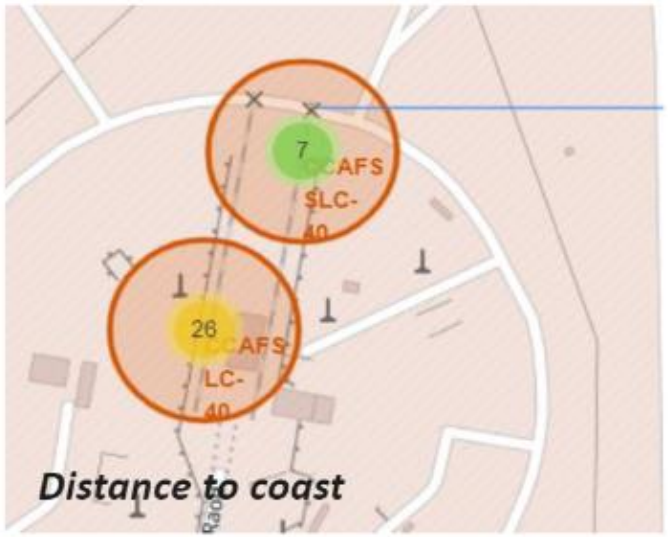
The success/failed launches for each site on the map

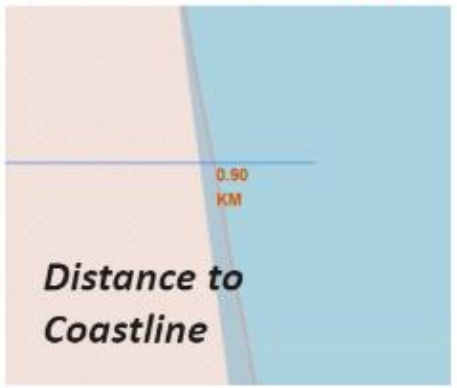


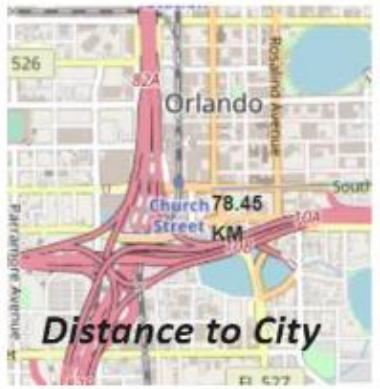


Launch Site distance to landmarks

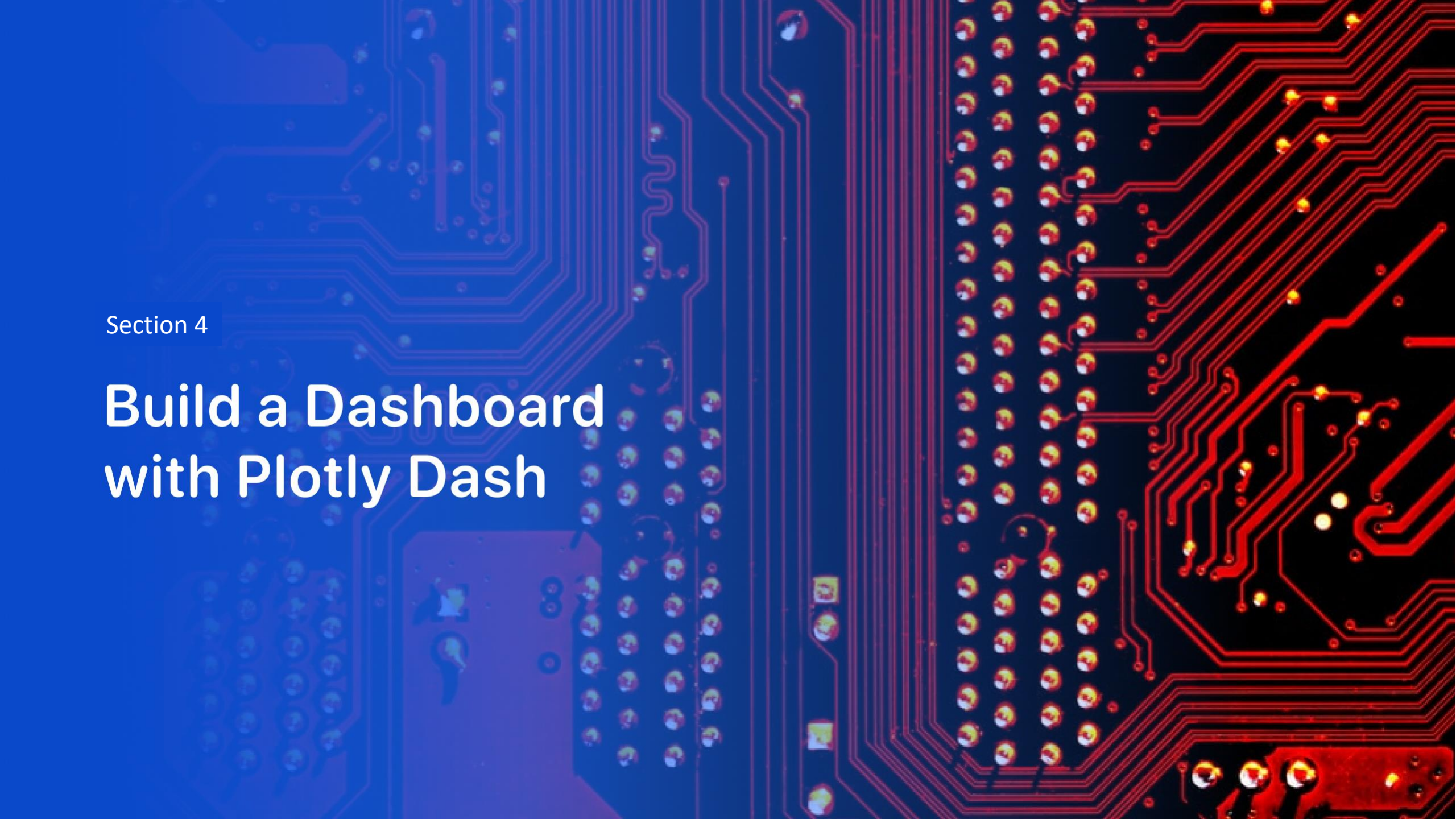






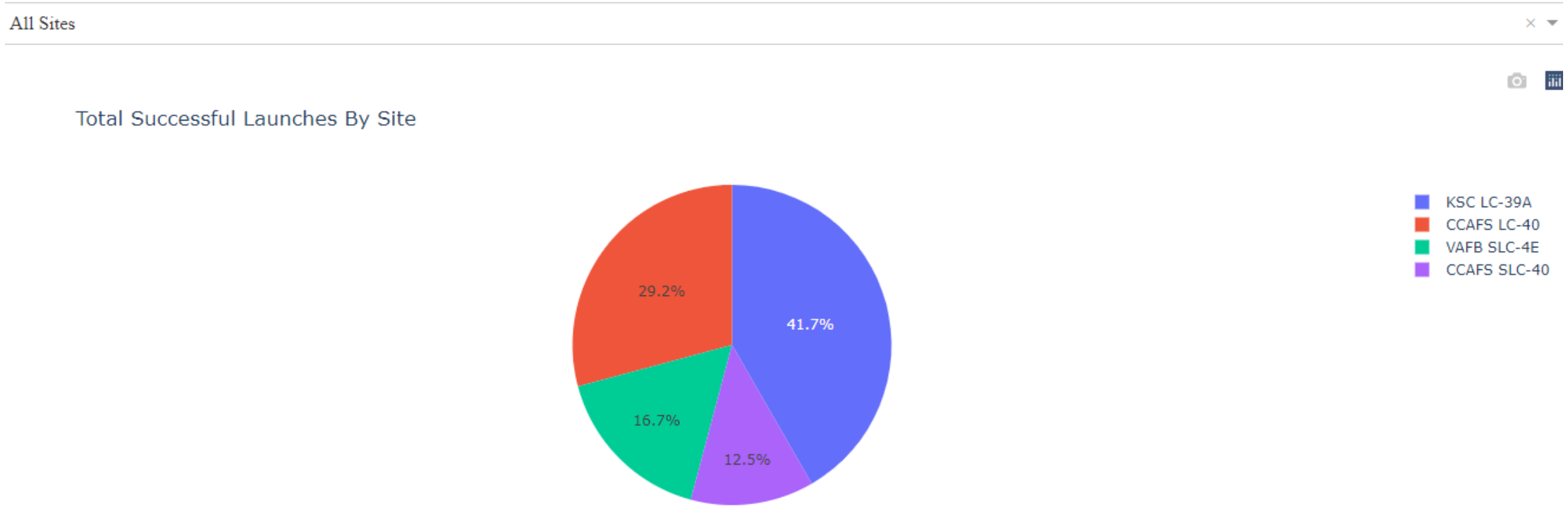


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

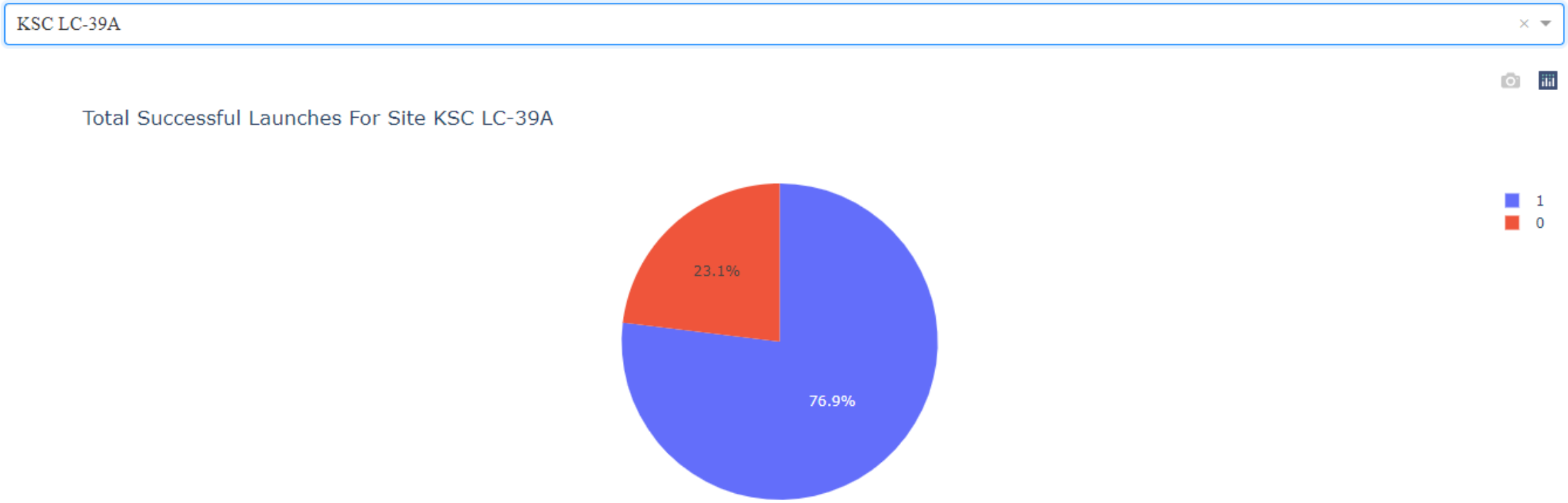
The success percentage by each sites.

- KSC LC-39A has the highest launch success rate and the largest number of successful launches among all the launch sites.
- Further investigation may be needed to determine the reasons why KSC LC-39A is the preferred launch site.



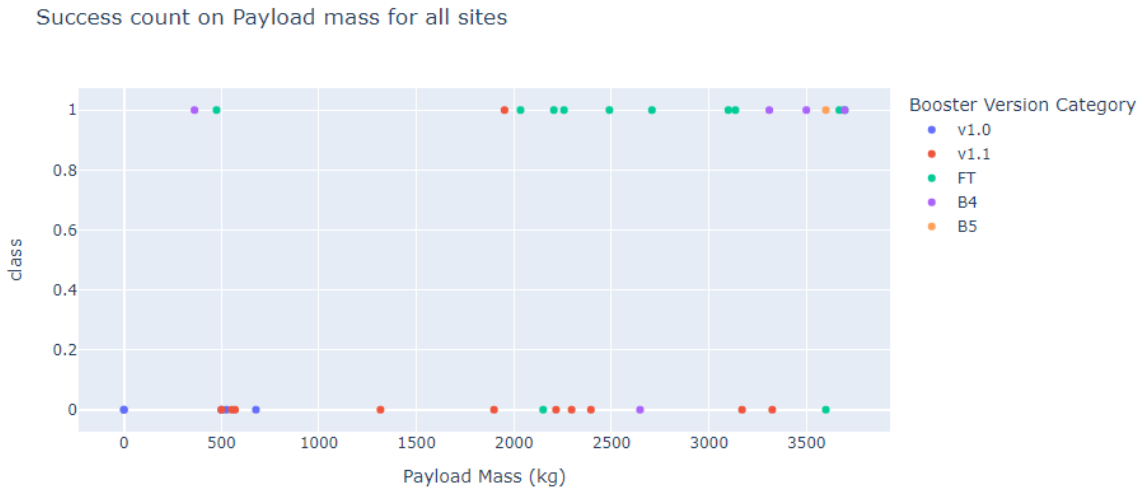
Total Successful Launches for Site KSC LC-39A

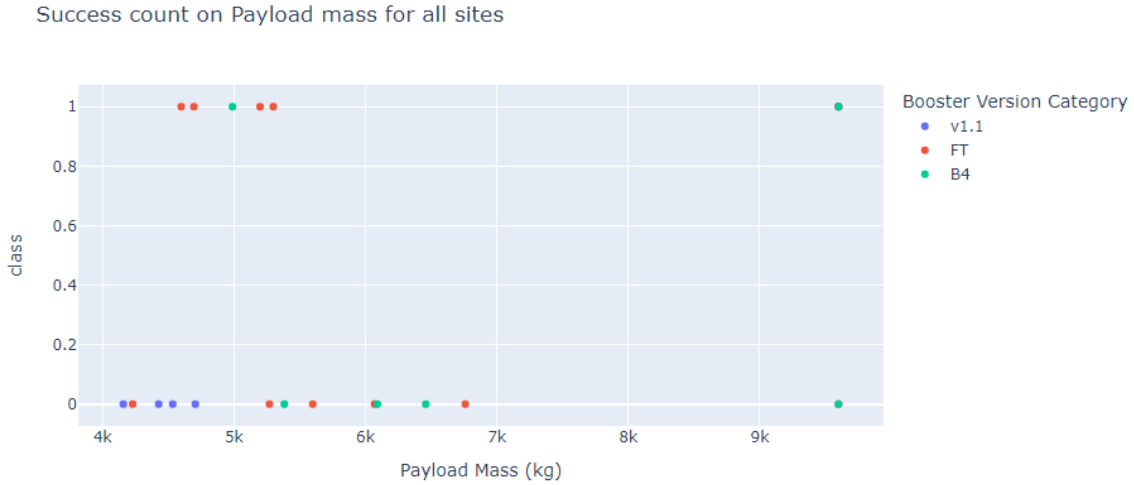
- KSC LC-39A has the highest success rate of 76.9% among all launch sites, which is only about 3% higher than the runner up, CCAFS LC-40."



Low-weighted payloads vs Heavy weighted payloads.

- Low weighted payloads have a better success rate than the heavy weighted payloads.



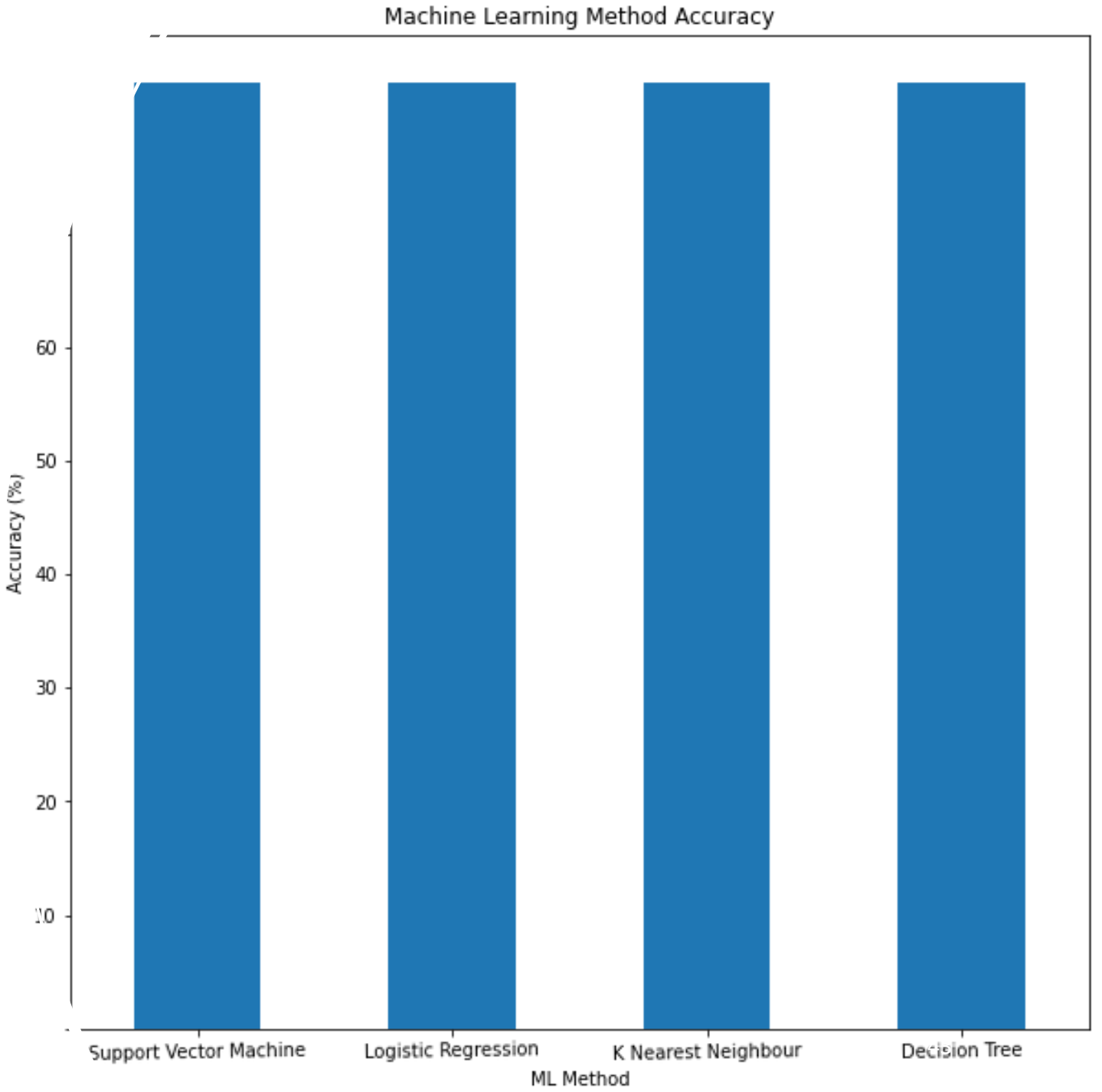


Section 5

Predictive Analysis (Classification)

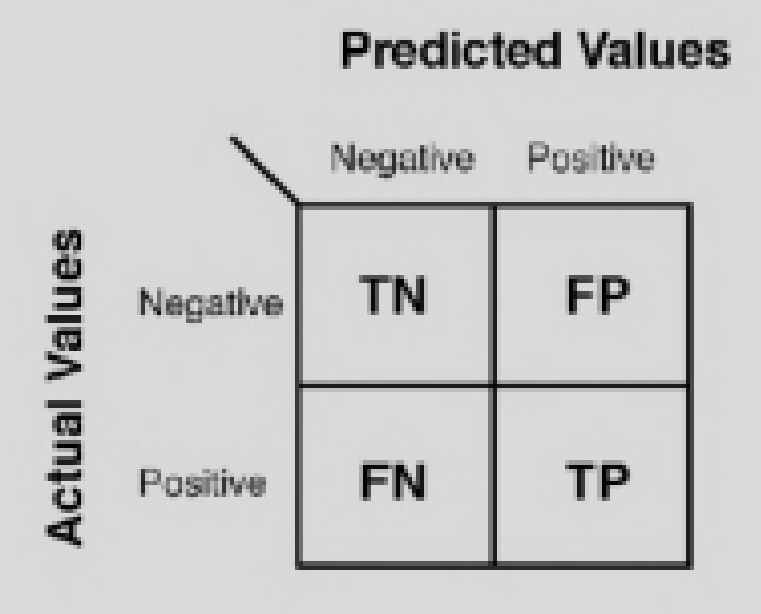
Classification Accuracy

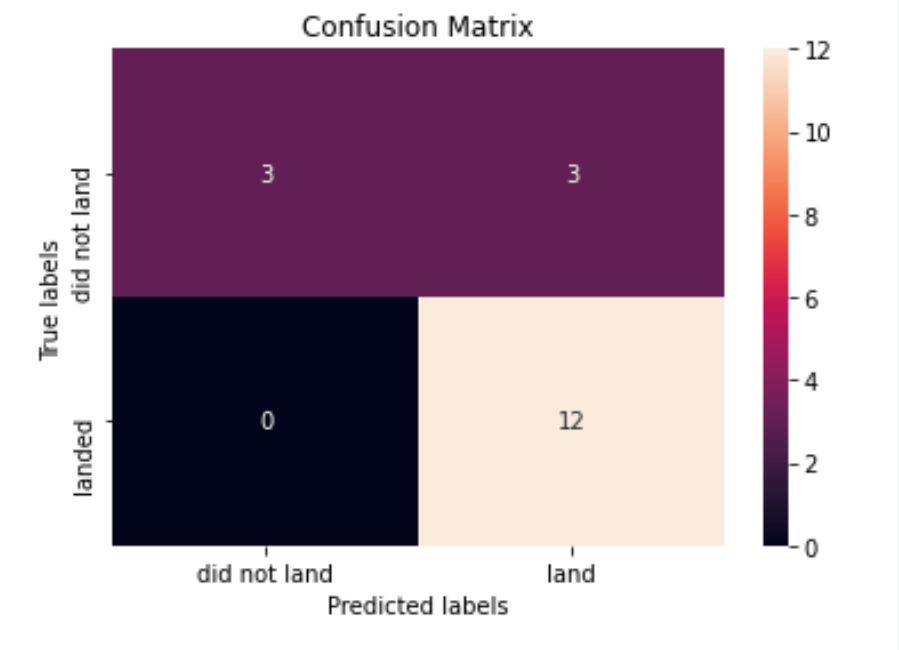
- Due to identical accuracy scores of 83.33% across all methods, Logistic Regression was selected for classification



Confusion Matrix

- The confusion matrix indicates that the decision tree classifier is capable of distinguishing between the different classes, but the major issue lies in false positives where unsuccessful landings are classified as successful.





Conclusions

- Through analysis of SpaceX launch data, key success factors were identified.
- Launch site location near the coast was a common factor for successful launches.
- KSC LC-39A had the highest launch success rate.
- Success rate increased with flight number and over time.
- Machine learning was used to predict landing outcomes with 83.33% accuracy.
- Decision Tree Algorithm was chosen as the best model.
- Orbits GEO, HEO, SSO, ES-L1 had the best success rates.
- Payload mass can impact success based on orbit requirements.
- More data on atmospheric or relevant factors could explain why some launch sites are better than others.

Thank you!