IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Soputhik SENG
05/23/2023

Outline

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

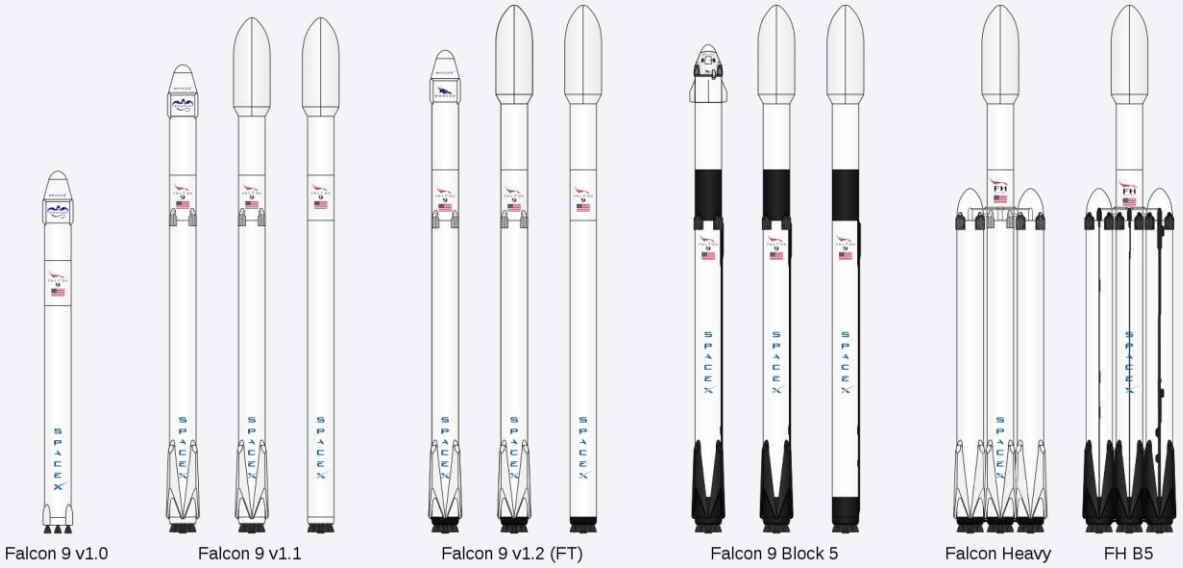
Executive Summary

- Summary of methodologies
  - Data Collection via Rest API
  - Data Collection via Web Scraping (Wikipedia)
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Map with Folium
  - Interactive Dashboard with Dash
  - Predictive Data Analysis
- Summary of all results
  - Results produced from EDA shows conclusive insight on the Falcon9 launches
  - Results produced from Interactive dashboard
  - Predictive Analysis Models that are able to predict the Success Rate of each launch with 83.33% accuracy

Introduction

- Background and Context:
  - SpaceX was able to develop a major breakthrough in reducing the cost of space launches through reusable rockets. Although other companies are also experimenting with this technology, SpaceX has by far the most advanced and well-known solution towards the issue. Since this part of the launch will heavily reduce the costs involved, we would like to be able predict whether the first stage Falcon rockets would be able to land or not.

- Problem we want to find answers:
  - We want to predict whether a launched Falcon Rocket would be able to successfully land.



Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX API in addition to Web Scraping from Wikipedia
- Perform data wrangling
  - The Data was first checked for Null Values, which was dealt with accordingly, we then reduce the data features on landing outcomes to better visualize the success rate
- Perform exploratory data analysis (EDA) using visualization tools such as Seaborn, Matplotlib and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We used the Wrangled data and split into train and test set before training it using multiple models to find out that all models work similarly on the data.

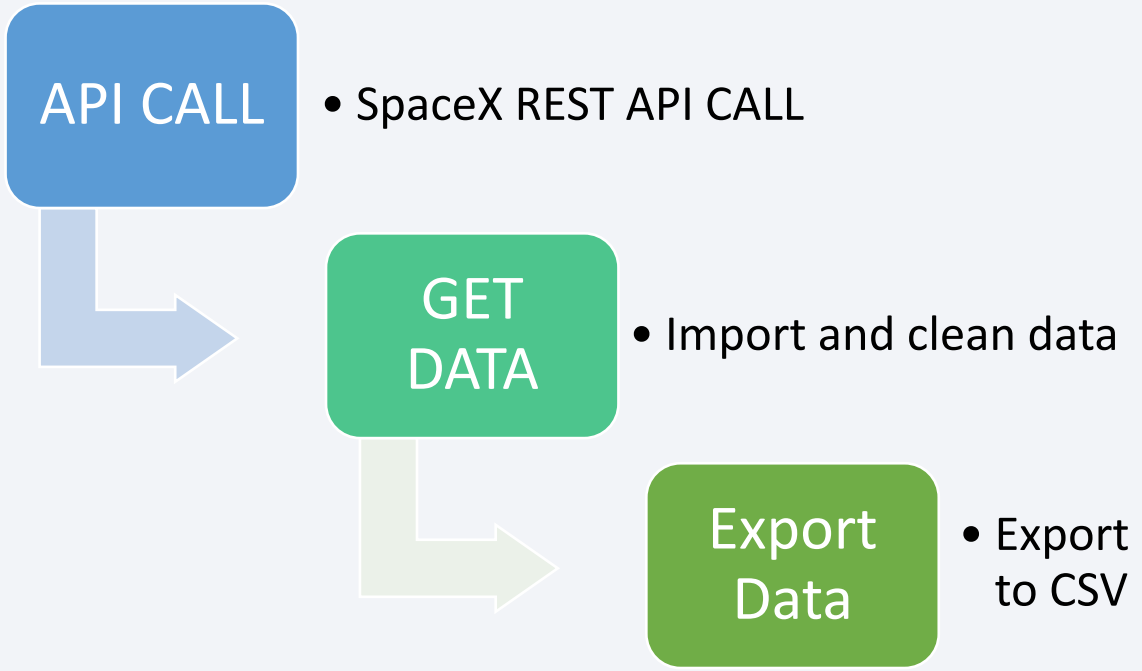
Data Collection

- There were two main sources of the data that was used for our models
  - One of the data was achieved through the **SpaceX API** where we extracted it by calling specifically to the Rest API using a series of helper functions to be able to extract every single data that is useful for our model.
  - The other data set was done through Web Scraping on the **Wikipedia** site to get additional data that were not included in the SpaceX API such as the booster landing outcome, payload mass, and more.

Data Collection – SpaceX API

- We created a GET request towards for the SpaceX API then with the help of a few helper functions we imported the necessary data into a Pandas DataFrame before exporting it into a CSV file.

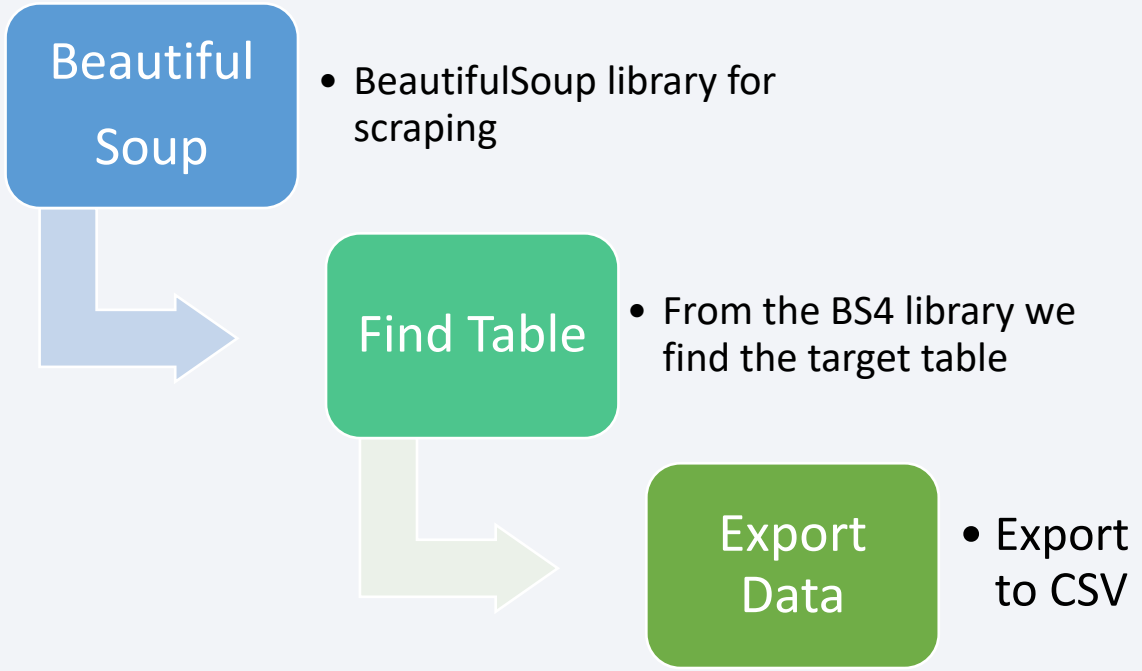
- Here is the link to the GitHub Data Collection Page



Data Collection - Scraping

- Using the Beautiful Soup library, we were able to scrape the Wikipedia web page for the target table that we were looking which contains the data that we need. We then put it into a single Panda DataFrame before exporting it to a CSV file

- Here is the link to the GitHub Web Scraping Page

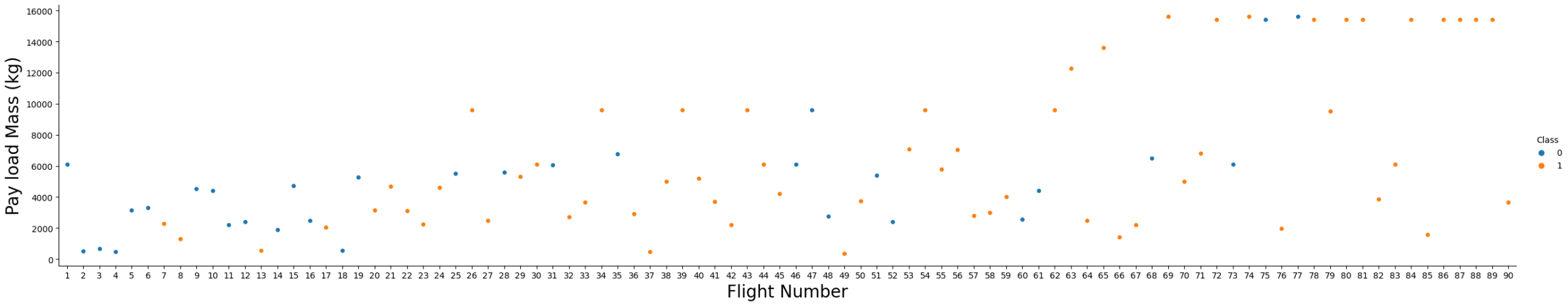


Data Wrangling

- To ensure that the data we have will provide the maximum information possible we need to clean the data in a way that the models could easily find the hidden patterns within.
- We did this by
    1. Exploring the data set looking for duplicates and null values. In our case we found that the LandingPad column has many null values.
    2. We did EDA (Exploratory Data Analysis) on the data acquired we found that there were multiple categorical columns that needed to be converted. We did so by creating **dummy variables** for columns (Orbits, LaunchSite, LandingPad, Serial, and the most important feature which is 'Class' which is used for determining the success/failure of LandingOutcome.
- Here are the links to the two pages where we completed Data Wrangling and EDA

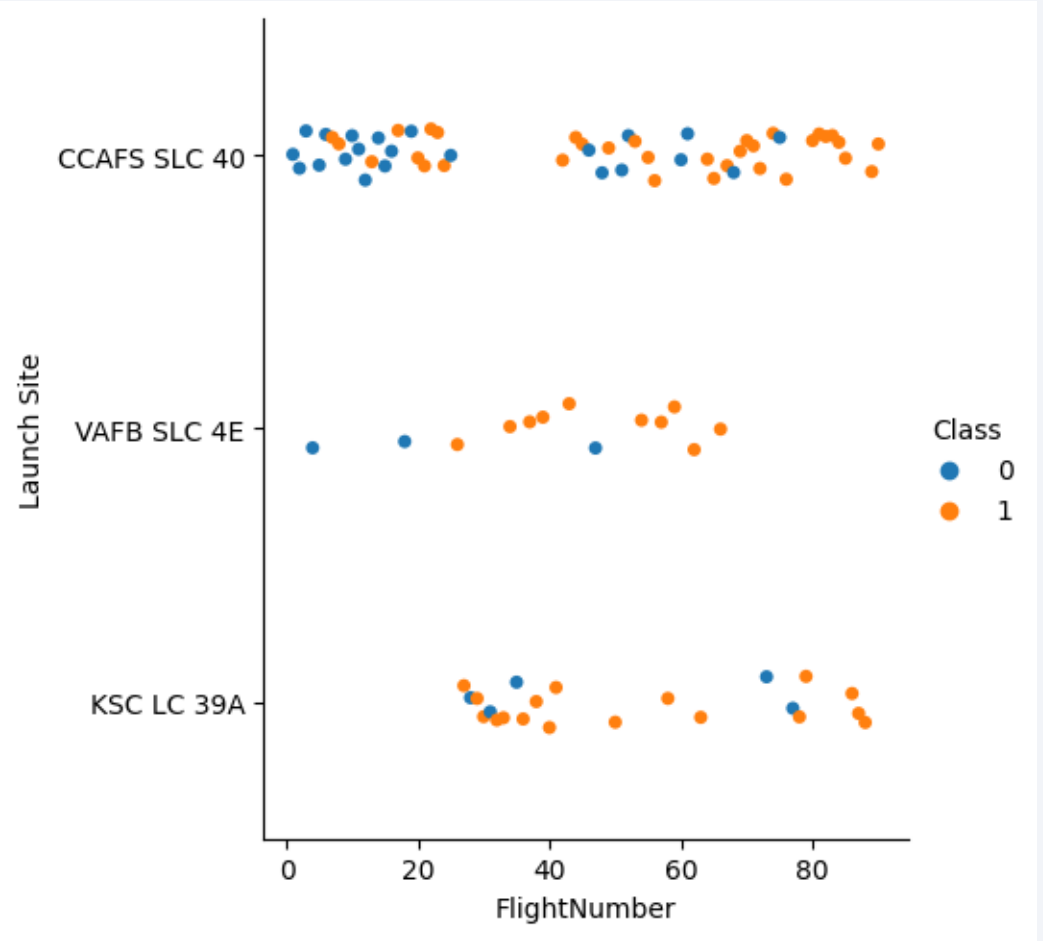
EDA with Data Visualization

- There were a variety of plots that was used for visualization with EDA, each of it has its own uses and meaning, to help us better understand the data we have at hand, and to allow us to identify if there were any issues with our data.
- The first chart that we plotted was a Scatter Plot showing all flights in relation to the Payload Mass that it carries as well as the Class of that flight of whether it had a positive outcome or not.



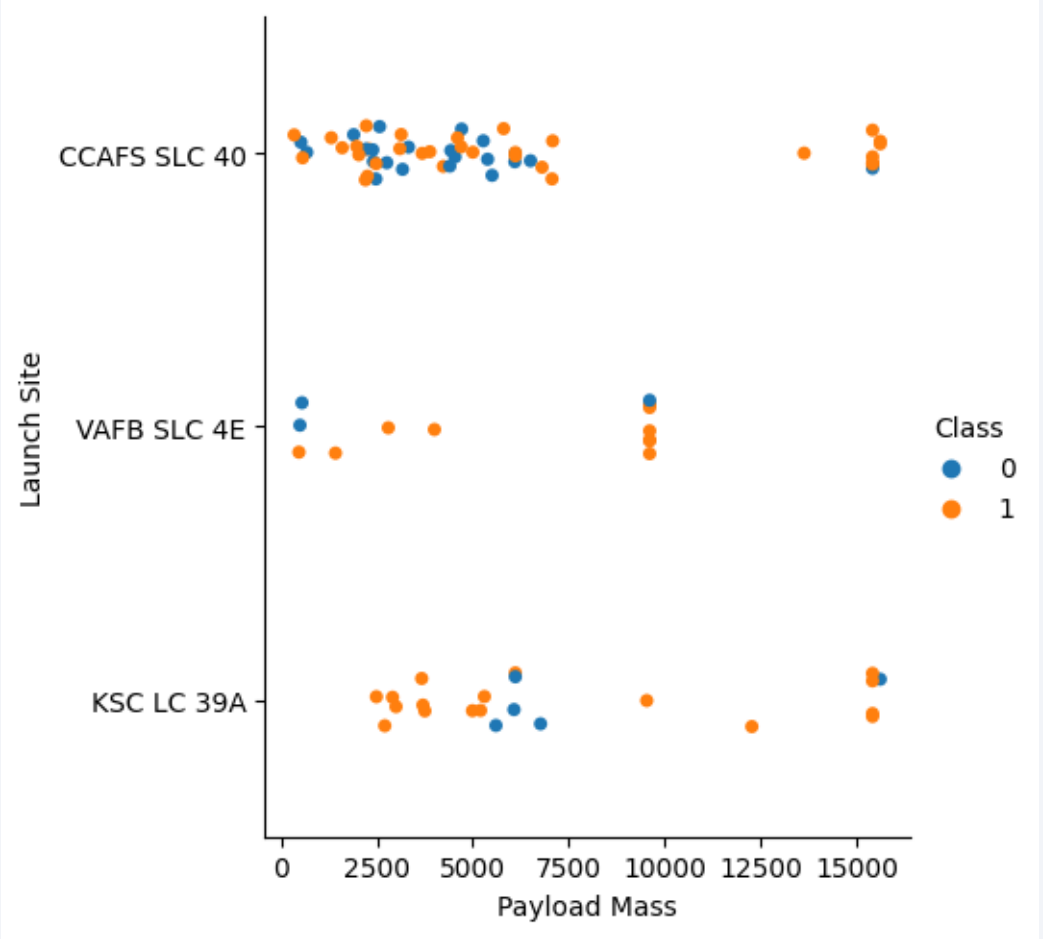
EDA with Data Visualization

- The second chart that we plotted was another Scatter Plot that helps us identify the relationship between the flight number and each specific launch site on top of which also showing us the Class as the hue.



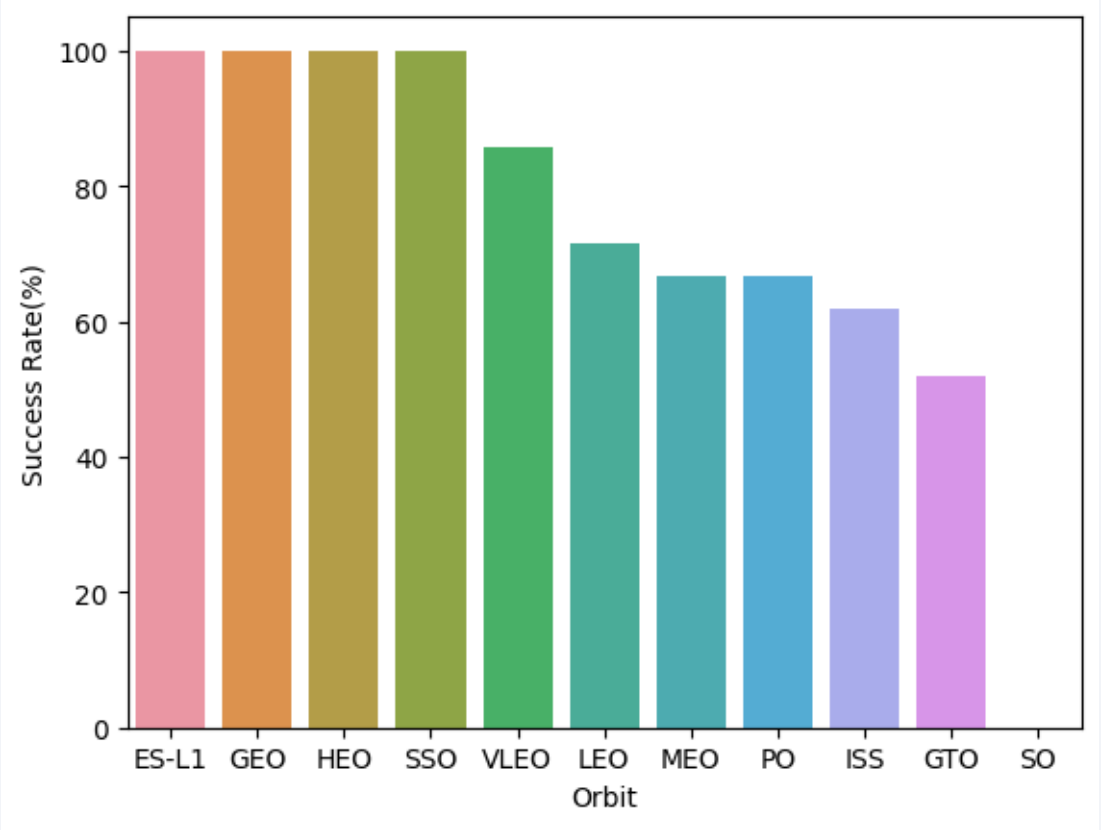
EDA with Data Visualization

- The third chart that we plotted was another Scatter Plot that helps us identify the relationship between the Payload Mass and each specific launch site on top of which also showing us the Class as the hue.



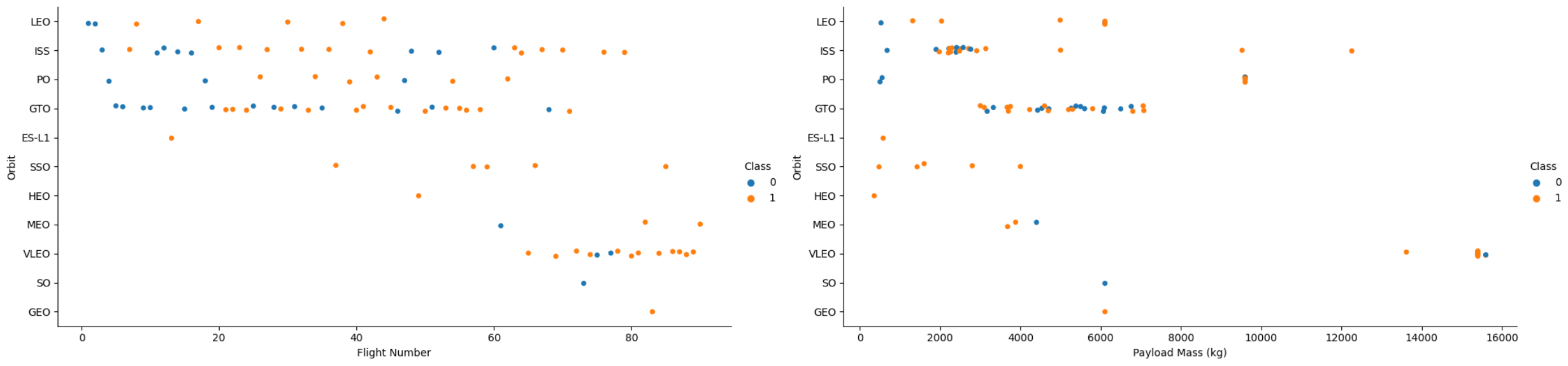
EDA with Data Visualization

- The fourth chart that we plotted was a bar chart which compared the Orbit type towards its Success Rate



EDA with Data Visualization

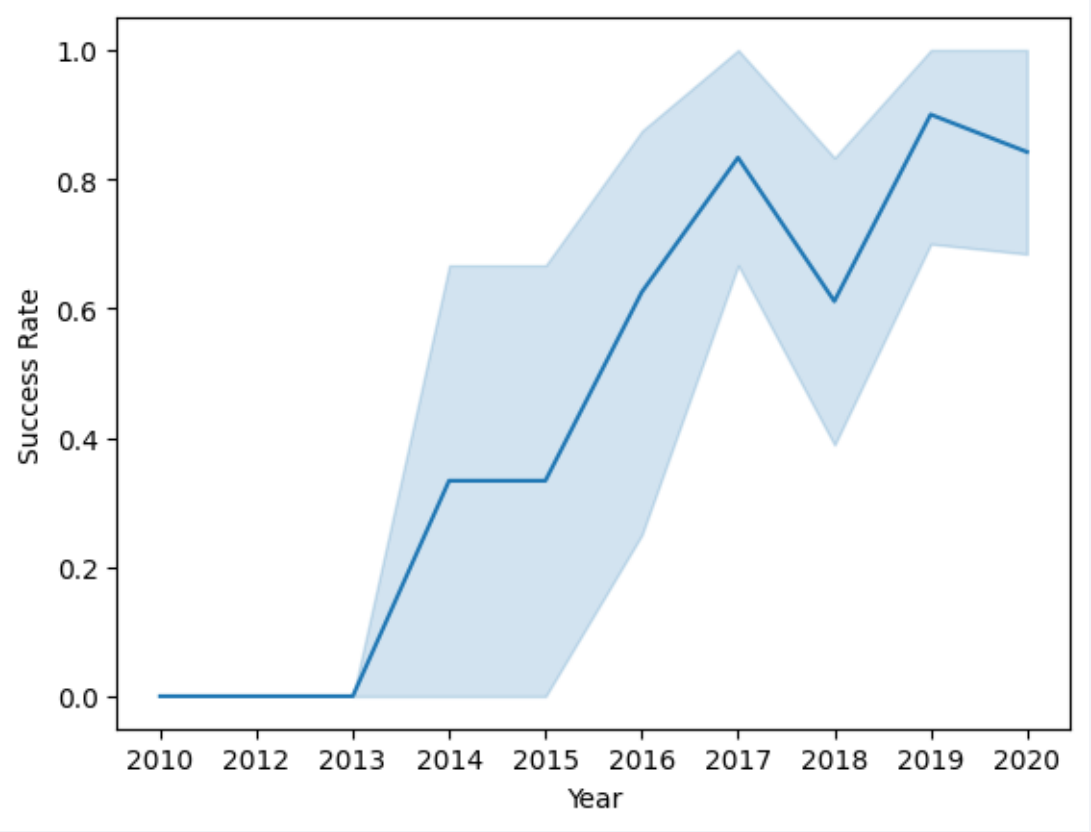
- The fifth and sixth plot represents the relationship of the Orbit type in relations to the flight number and payload mass with the hue as Class.



EDA with Data Visualization

- The last chart that were plotted represented the success rate of all the launches throughout the years.

- Here are the links to the page where we looked into and did EDA with Data Visualization on the data.



EDA with SQL

- There were 10 total queries that we performed:
  - We first looked into the different launch sites that SpaceX had used to launch the Falcon 9
  - We then looked into the first five data of launches made from sites that started with 'CCA'
  - Looked into the total Payload Mass that has been transport so far
  - The average Payload Mass carried by a specific booster version called "F9 v1.1"
  - The first date where a successful landing was carried out on a Ground Pad
  - List booster names that were able to carry Payload with a mass in between 4000 and 6000kg and successfully landed back on a drone ship
  - Summarize the number of mission that were considered as successful or failure
  - List all booster versions that were able to carry the maximum Payload mass
  - List failed Landing Outcome in the year 2015
  - Rank the successful landing outcomes within the time period of 04-06-2010 and 20-03-2017
- Link to GitHub SQL EDA

Build an Interactive Map with Folium

- We added a Folium Circle and Marker object on each launch site before then adding a marker cluster of each launches in each launch site, a Mouse Position is then added so that we are able to get the Latitude and Longitude the mouse is currently on, and last but not least we used the Mouse Position object to add lines onto the map displaying the distance between the launch site and the coast as well as some key infrastructures.
- All of these object's aids in our ability to look at the key details on a map. It allows to easily understand and identify the positions of each launch site, its success rate, and how each launch site is chosen, especially with regards to the infrastructures within its vicinity and location.
- Link to GitHub Folium Map Python Notebook

Build a Dashboard with Plotly Dash

- There were two main graphs that were added to the dashboard, a pie chart and scatter plot. Each of these graphs also have an Input function in terms of a Dropdown filter and a slider to set filter data to what the viewer wants.
- The plots helps viewers identify the launch site and its success rates while the scatter chart helps viewers identify the relationship between the Class with its Payload Mass, the drop down and the sliders aids in getting a better view of a more specific data set that we would like to retrieve, for example a data set of only the launch site KSC LC-39A of Payload between 4000 and 6000 kg.
- Link to GitHub Python Plotly Dashboard

Predictive Analysis (Classification)

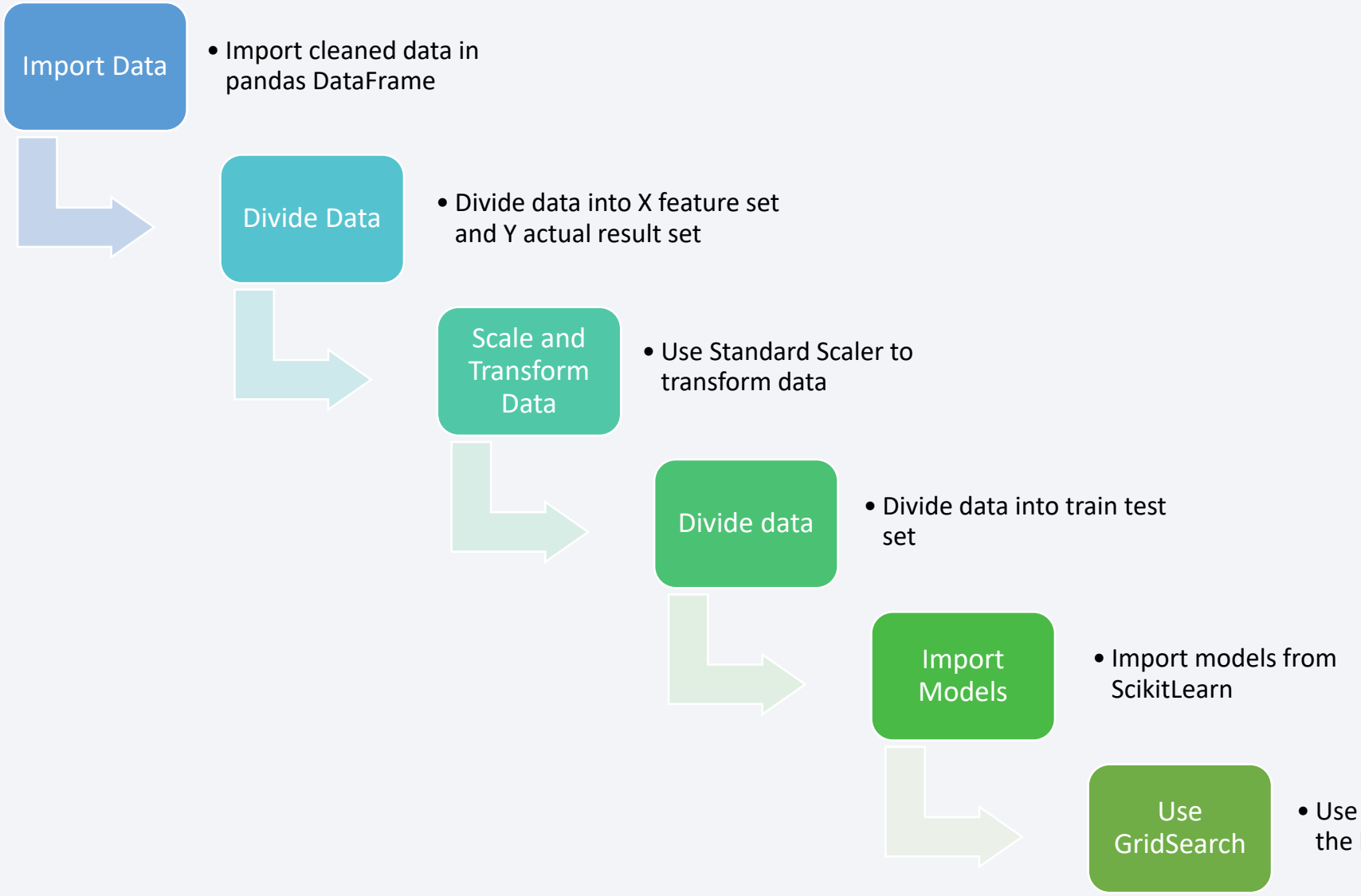
- To create a model to predict the success of a Falcon 9 Launch, we first need import the data into a pandas DataFrame
- Before we could do anything else, we need to divide the data into features set determined by X and the actual result which is determined by Y
- We then scale and transform the features dataset X to ensure that the data is in a standardize form.
- We then divide the into train test set using the train_test_split method
  - We set the test_size to 0.2 and the random_state to 2 to ensure every run will output the same result

Predictive Analysis (Classification)

- In total there were four different models that we had trained our data on. They are Classification Trees, SVM, Logistic Regression, and KNN (K Nearest Neighbor)
- We used GridSearch with specific parameters for each models in order for it to be able to produce the best possible result using the cv=10 parameter.
- Link to GitHub Predictive Analysis Notebook

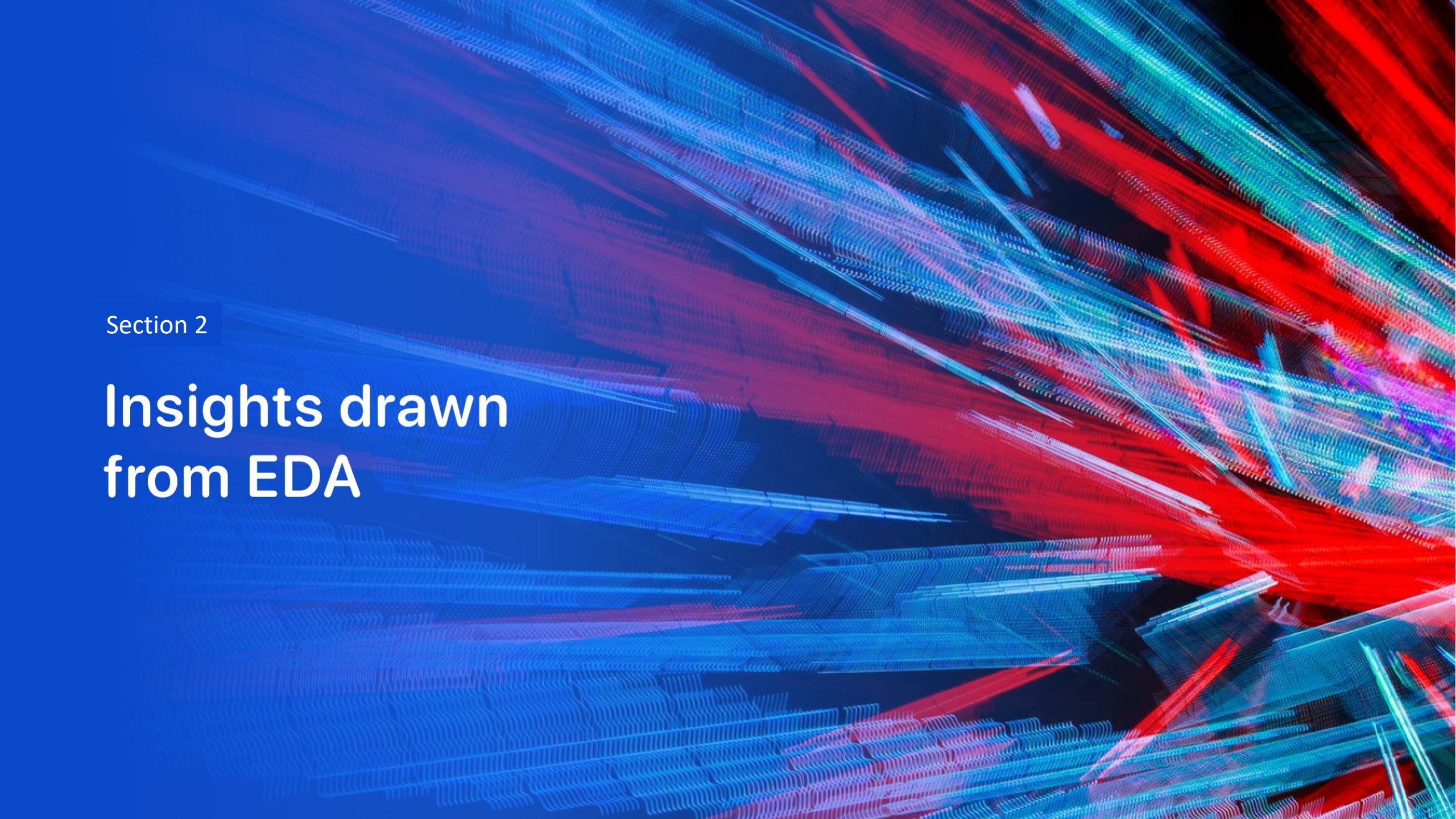
Model	Test Scores	Training Score	Confusion Matrix
knn	0.833333	0.848214	[[3, 3], [0, 12]]
decision_tree	0.833333	0.862500	[[3, 3], [0, 12]]
logistic_regression	0.833333	0.846429	[[3, 3], [0, 12]]
svm	0.833333	0.848214	[[3, 3], [0, 12]]

Predictive Analysis (Classification) Flowchart



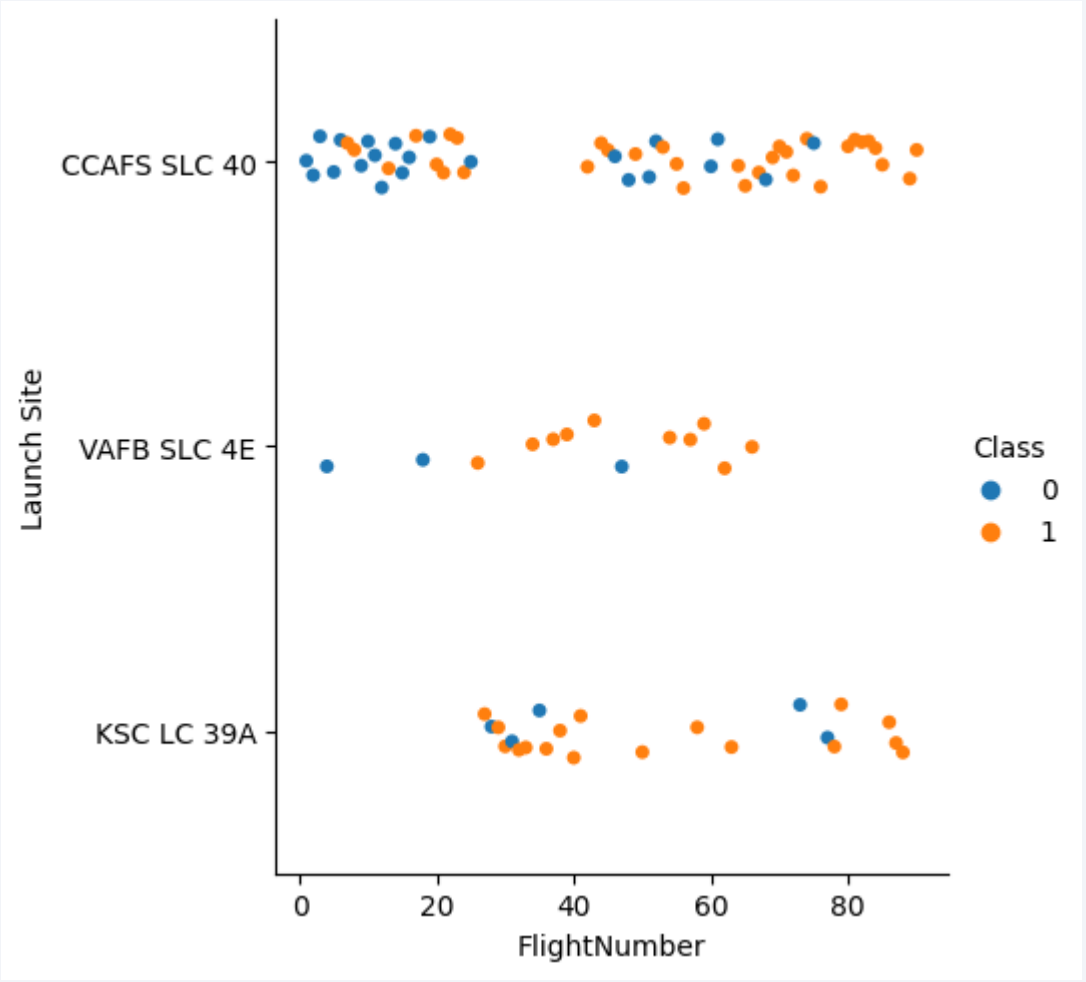
Results

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


Section 2
Insights drawn from EDA

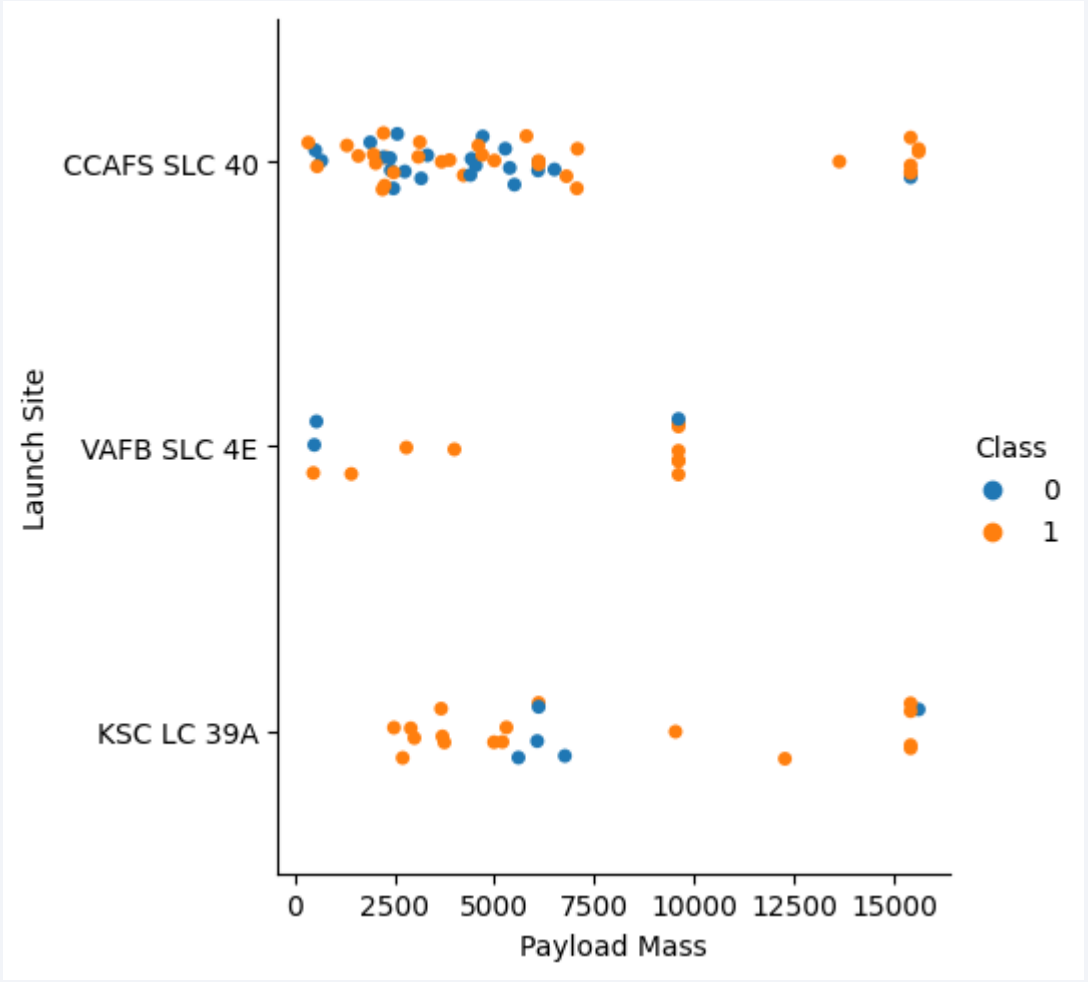
Flight Number vs. Launch Site

- As you can see from the scatter chart you can see that the majority launches are made from Launch Site "**CCAFS SLC 40**"
- The majority of the first 20 flights has a very small success rate, but as the flight number increases, the success rate starts to increase.



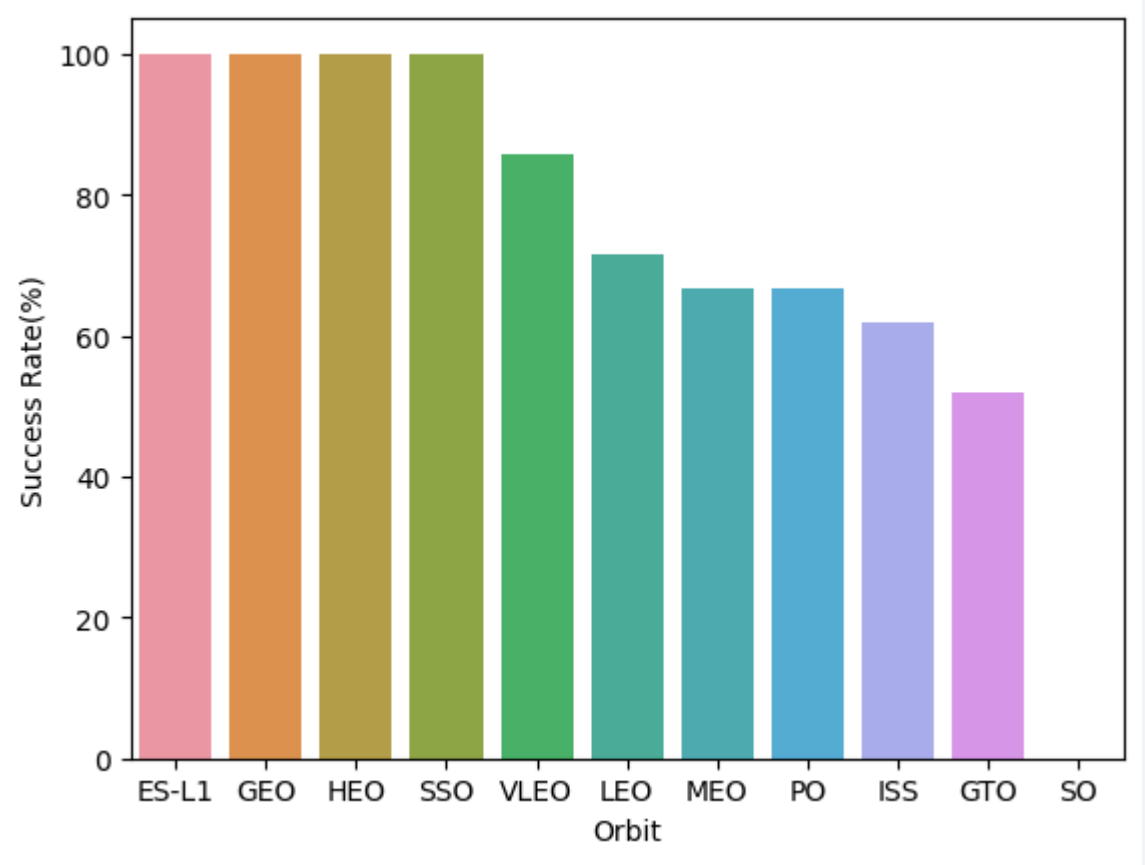
Payload vs. Launch Site

- Most of the launches are made with a payload of around **7500** kg or less.
- However, as the payload gets over 7500 kg, we see a higher success rate.



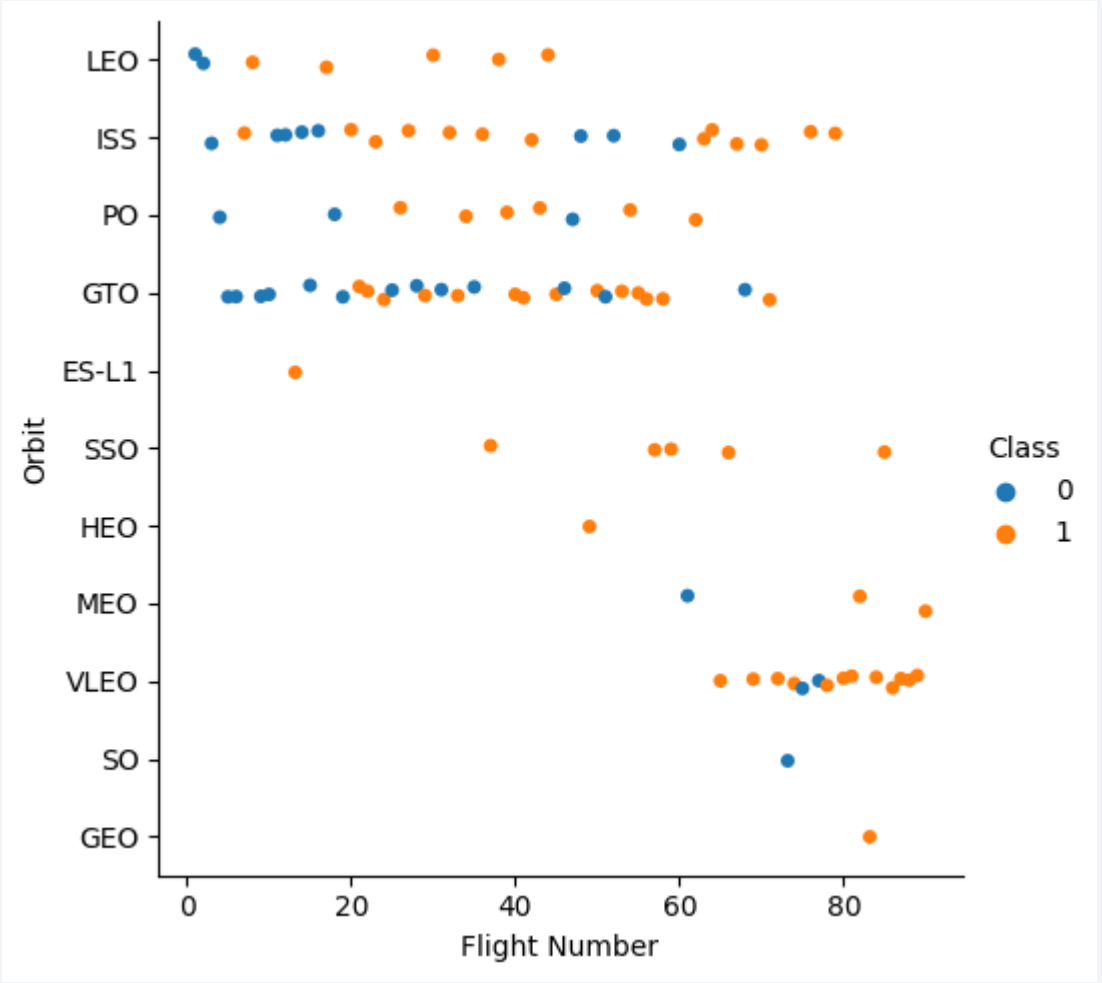
Success Rate vs. Orbit Type

- Launches with payload going to the ES-L1, GEO, HEO, and SSO orbit all have 100% success rate.



Flight Number vs. Orbit Type

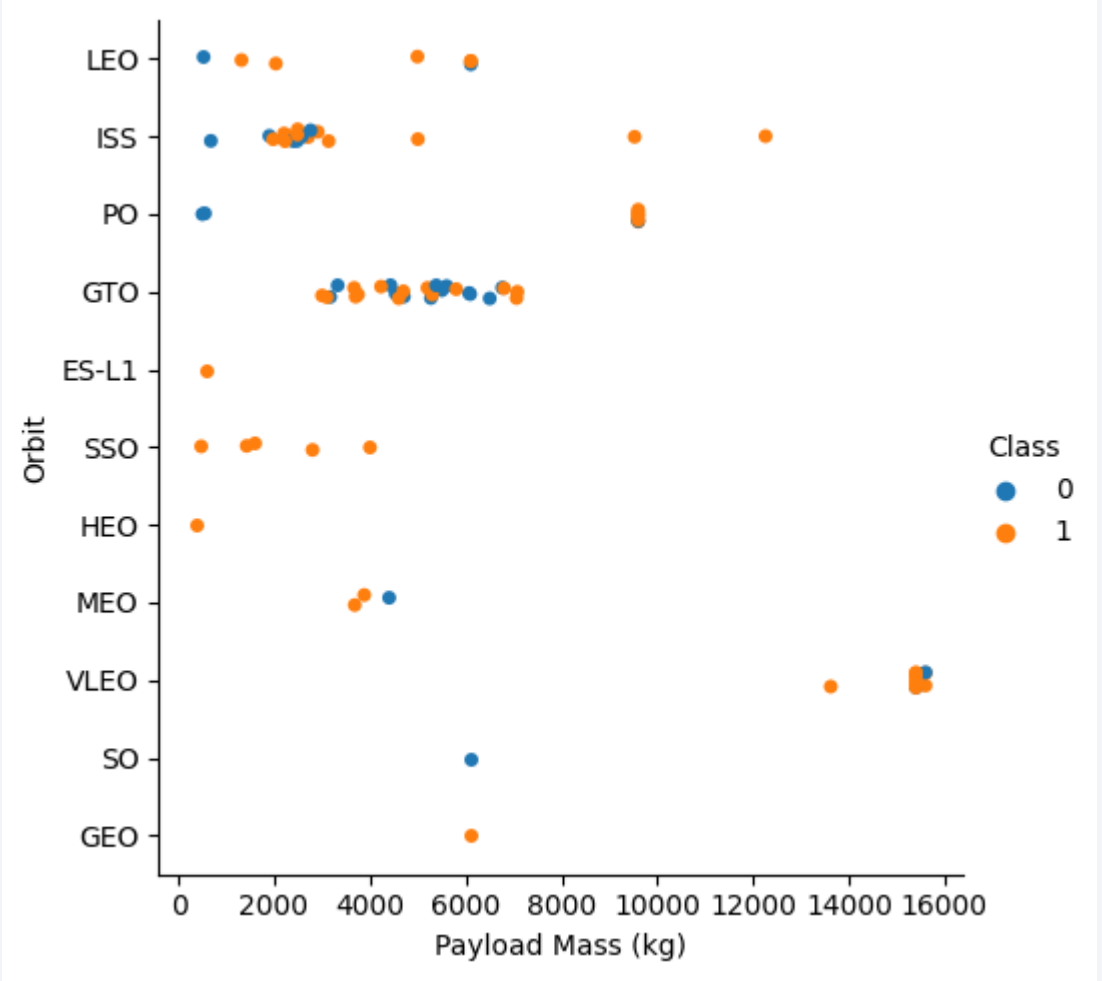
- We can see a lot of variety of the orbit types that SpaceX launches to, particularly only 4 of 11 types of orbit were only launched once.
- SpaceX launched most of its payload to LEO, ISS, PO, GTO, and VLEO type orbit.



Payload vs. Orbit Type

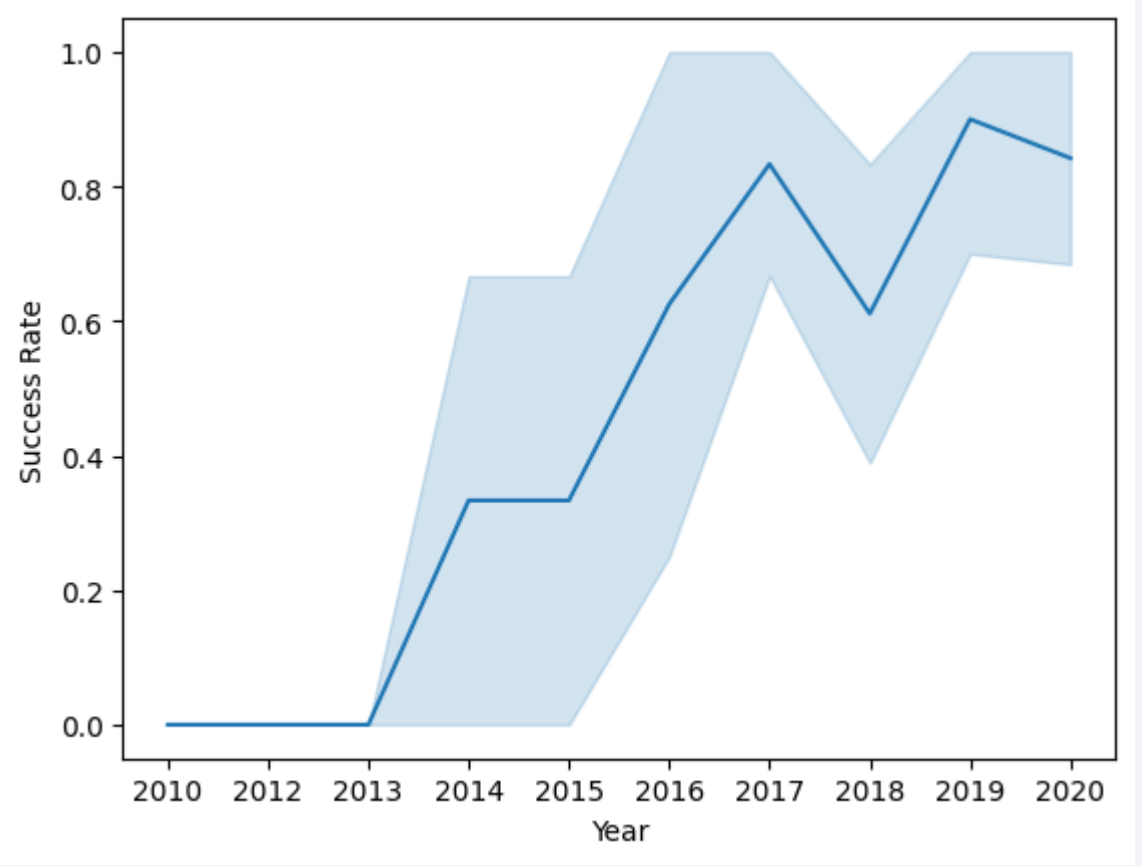
- Most of the payload being sent to orbit have a mass of less 8000 kg

- Only the ISS, PO, and VLEO orbit type have send payload with masses of over 8000 kilograms to orbit.



Launch Success Yearly Trend

- At the beginning launches were very unsuccessful with a 0% rate
- It started to increase as more flights and tests were made as it reached 80% in 2017.
- There were a small dip in the success rate during 2018 but recovered to over 80 percent in 2019.



All Launch Site Names

- There are 4 unique Launch Sites that SpaceX mainly uses for Falcon 9 launches mainly they are: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40

Task 1

Display the names of the unique launch sites in the space mission

[8]:
```
%%sql
SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

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

[8]:

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

None

Launch Site Names Begin with 'CCA'

- The first five launches made from the site that begins with CCA are as follows:
  - 4 of which have NASA as the Customer, while the other being SpaceX

[9]:
```
%%sql
SELECT * FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

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

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The Falcon 9 has totally transported **45596 kilograms** worth of payload during every single one of its missions.

```
[10]: %%sql
      SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_NASA_PAYLOAD FROM SPACEXTBL
      WHERE Customer = 'NASA (CRS)';
       * sqlite:///my_data1.db
      Done.
[10]: TOTAL_NASA_PAYLOAD
      ---------------------
                    45596.0
```

Average Payload Mass by F9 v1.1

- Among the total 45596 kilograms worth of payload that was transported, **2928.4 kilograms** was done so by the Falcon 9 v1.1.

```
[11]: %%sql
      SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_F9v11
      FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

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

[11]:

AVG_PAYLOAD_F9v11
2928.4

First Successful Ground Landing Date

- The first successful landing on a ground pad was made on the **1st of August, 2018**.

```
[12]: %%sql
      SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

[12]:

MIN(DATE)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- There were four different Falcon 9 versions that were able to successfully carry a payload of 4000 to 6000 kilograms and were then able to successfully land on a drone ship. They are **F9 FT B1022**, **F9 FT B1026**, **F9 FT B1021.2**, and **F9 FT B1031.2**.

[21]:
```
%%sql
SELECT BOOSTER_VERSION FROM SPACEXTBL
WHERE (LANDING_OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

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

[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

- Out of the total 101 launches, 100 were considered a success by SpaceX and only 1 were considered a Failure.

[15]:
```
%%sql
SELECT DISTINCT(MISSION_OUTCOME), COUNT(MISSION_OUTCOME) FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

[15]:

Mission_Outcome	COUNT(MISSION_OUTCOME)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are totally 12 different Booster Versions that were able to carry the max payload, they are shown in the image to the right.

```
[16]: %%sql
      SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL
      WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

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

[16]:

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

- There were two different landing failures when trying to land on the drone ship in the year 2015. Both launches were made from the Launch Site CCAFS LC-40.

[17]:
```
%%sql
SELECT substr(Date, 4, 2) AS MONTH, substr(Date, 7, 4) AS YEAR, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME
FROM SPACEXTBL WHERE substr(Date, 7, 4)='2015' AND LANDING_OUTCOME = 'Failure (drone ship)';
```

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

[17]:

MONTH	YEAR	Booster_Version	Launch_Site	Landing_Outcome
10	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- During the period, there were totally 35 successful launch outcomes, 7 on the ground pad, 8 on the drone ship, and 20 other success outcomes.

[24]:
```
%%sql
SELECT LANDING_OUTCOME, COUNT(*) FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Success%' AND (DATE BETWEEN '04-06-2010' AND '20-03-2017')
GROUP BY LANDING_OUTCOME
ORDER BY LANDING_OUTCOME DESC;
```

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

[24]:

Landing_Outcome	COUNT(*)
Success (ground pad)	7
Success (drone ship)	8
Success	20

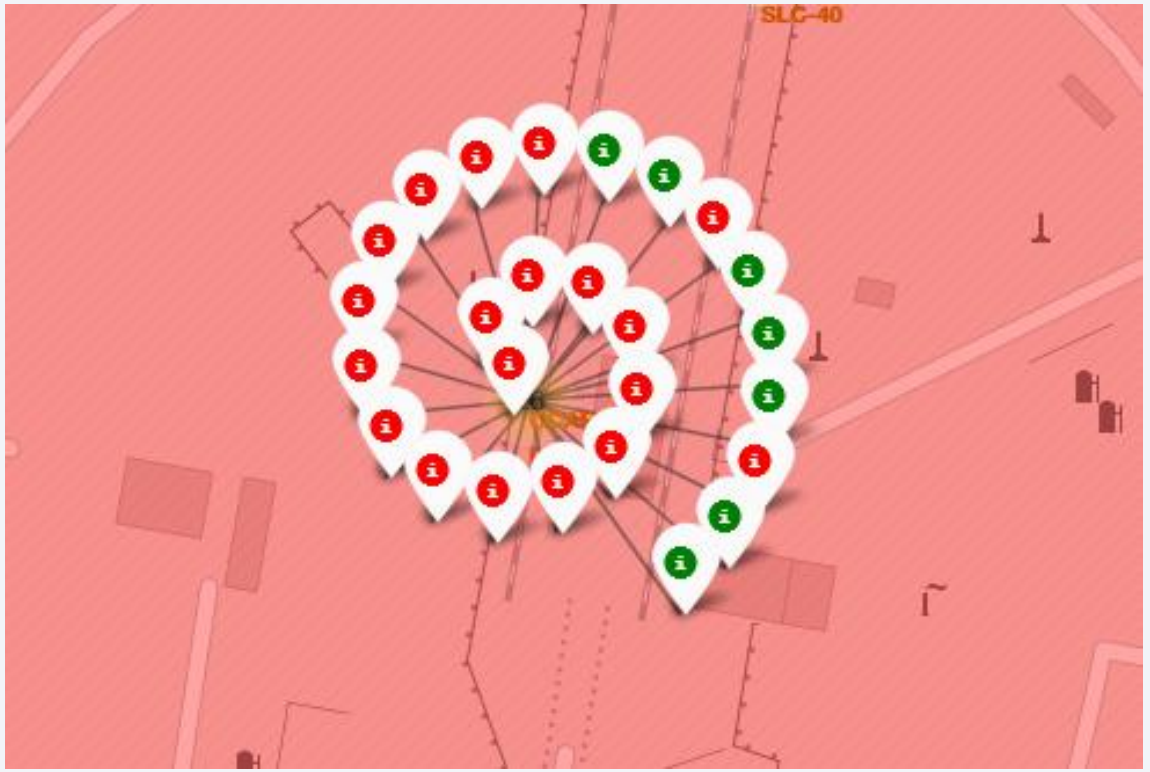

Section 3
Launch Sites
Proximities Analysis

Launch Sites Location

- There are four main launch sites, three were located on the east coast of Florida and one on the west coast of California. All sites also have proximity towards the equator.

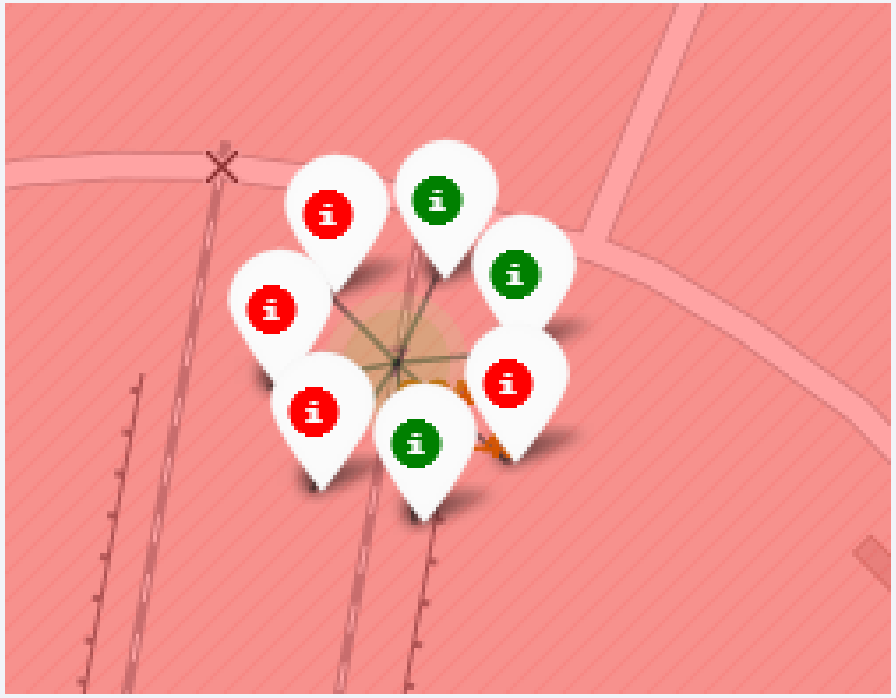
Success/Failure Launches on site CCAFS LC-40

- Many of the launches made on the launch site CCAFS LC-40 fails. The site is located on the East Coast in Florida. There were a total of 26 launches.



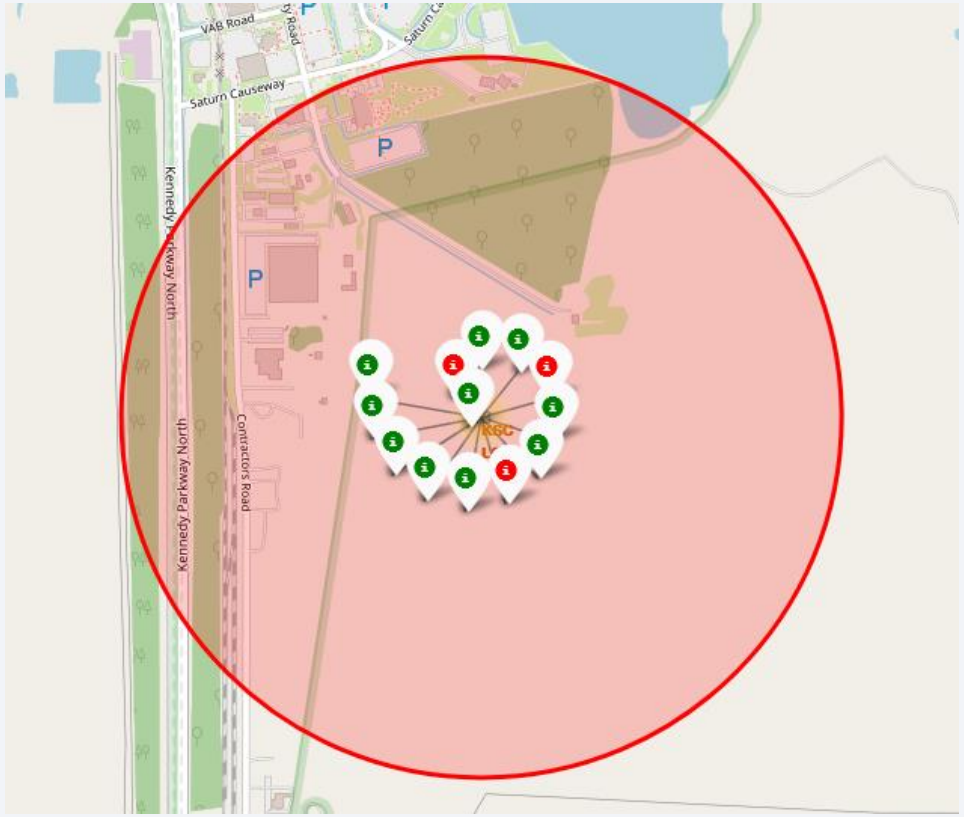
Success/Failure Launches on site CCAFS SLC-40

- Many of the launches made on the launch site CCAFS SLC-40 succeeded, but there were a lot less launches with only 7 that were made here, although it is right next to CCAFS LC-40.



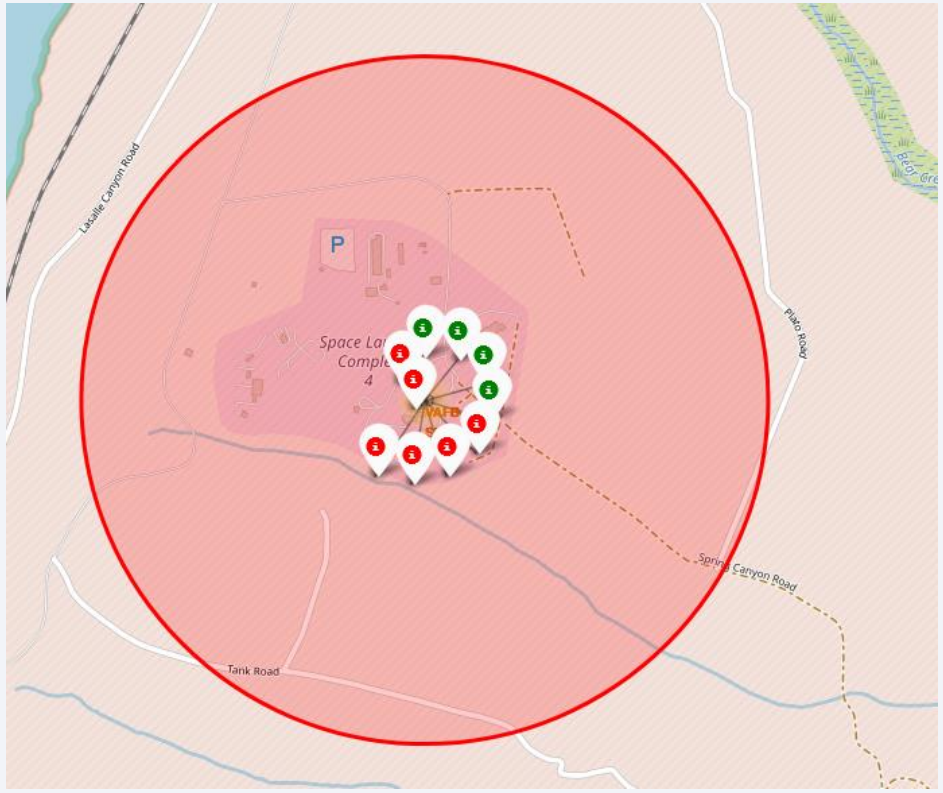
Success/Failure Launches on site KSC LC-39A

- Another launch site in the vicinity of Florida, they were a total of 13 launches here the majority of which succeeded with only 3 failures.



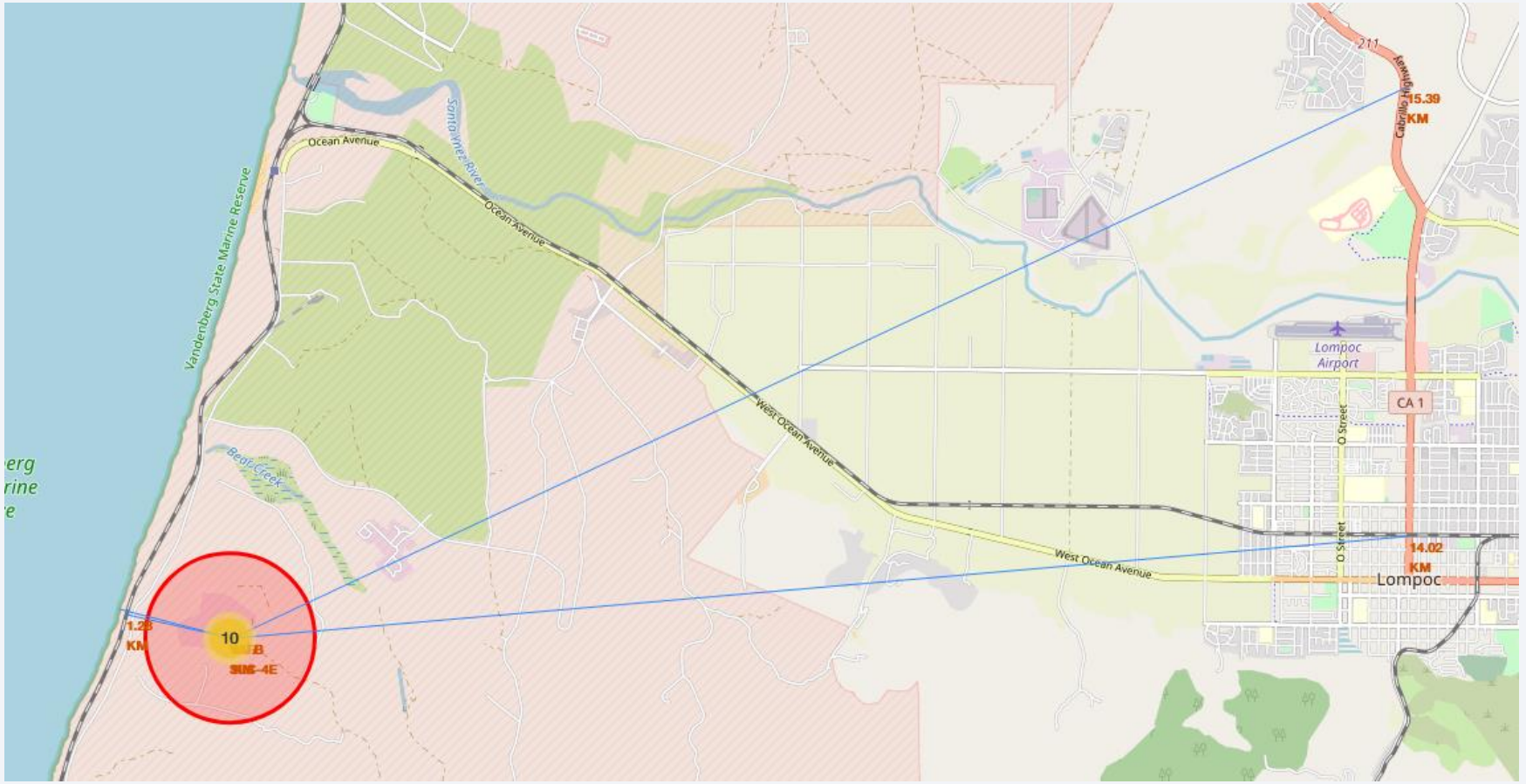
Success/Failure Launches on site VAFB SLC-4E

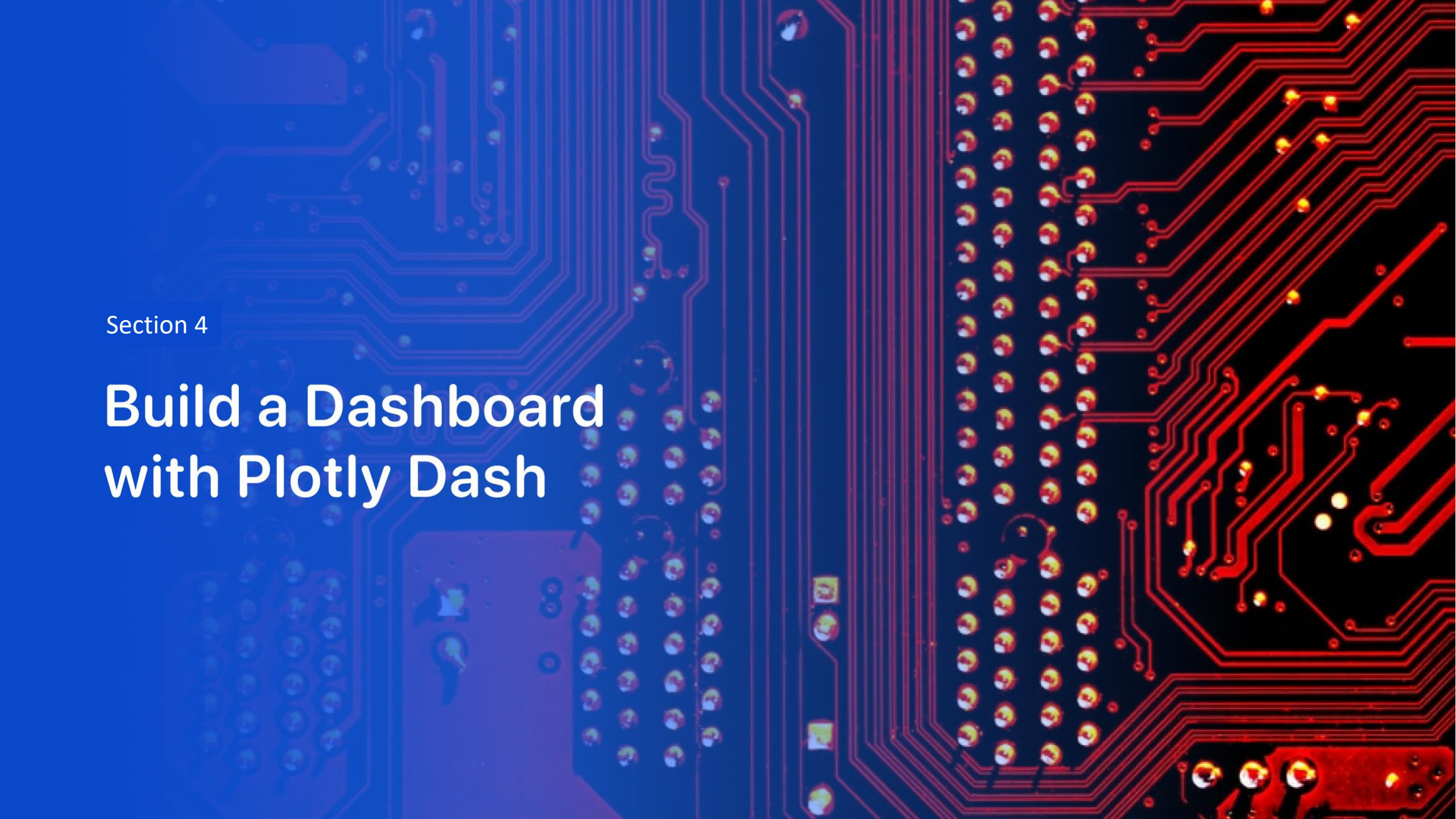
- Contrary to the rest, this site is located on the West Coast, in California. There were totally 10 launches here with a 40% success rate.



VAFB SLC-4E Proximity to Civilian Infrastructure

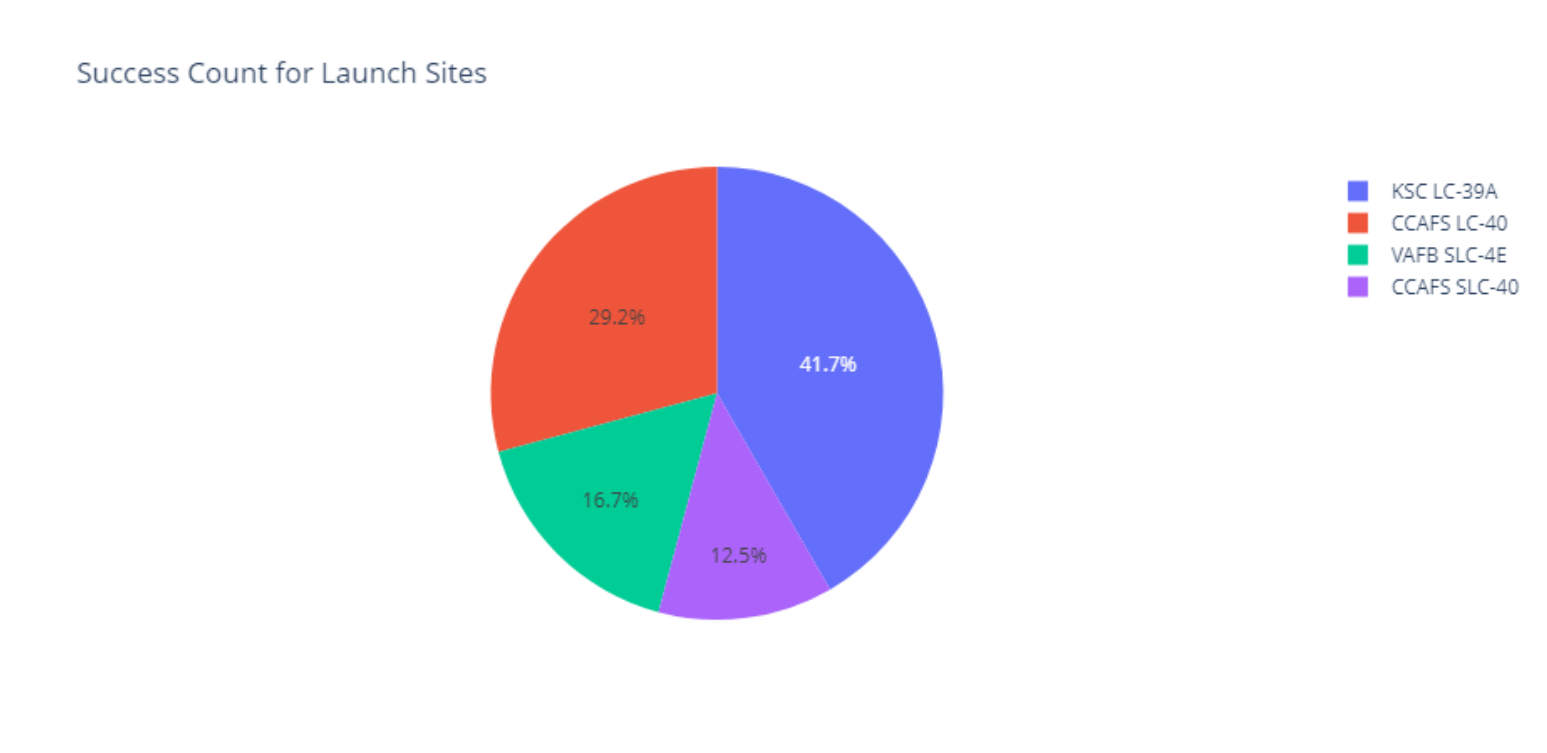
- As shown the launch site is quite close to the coast the distance of only a bit over a kilometer to a coast and a railway, however, it is quite distant from the closest city and highway.




Section 4
Build a Dashboard with Plotly Dash

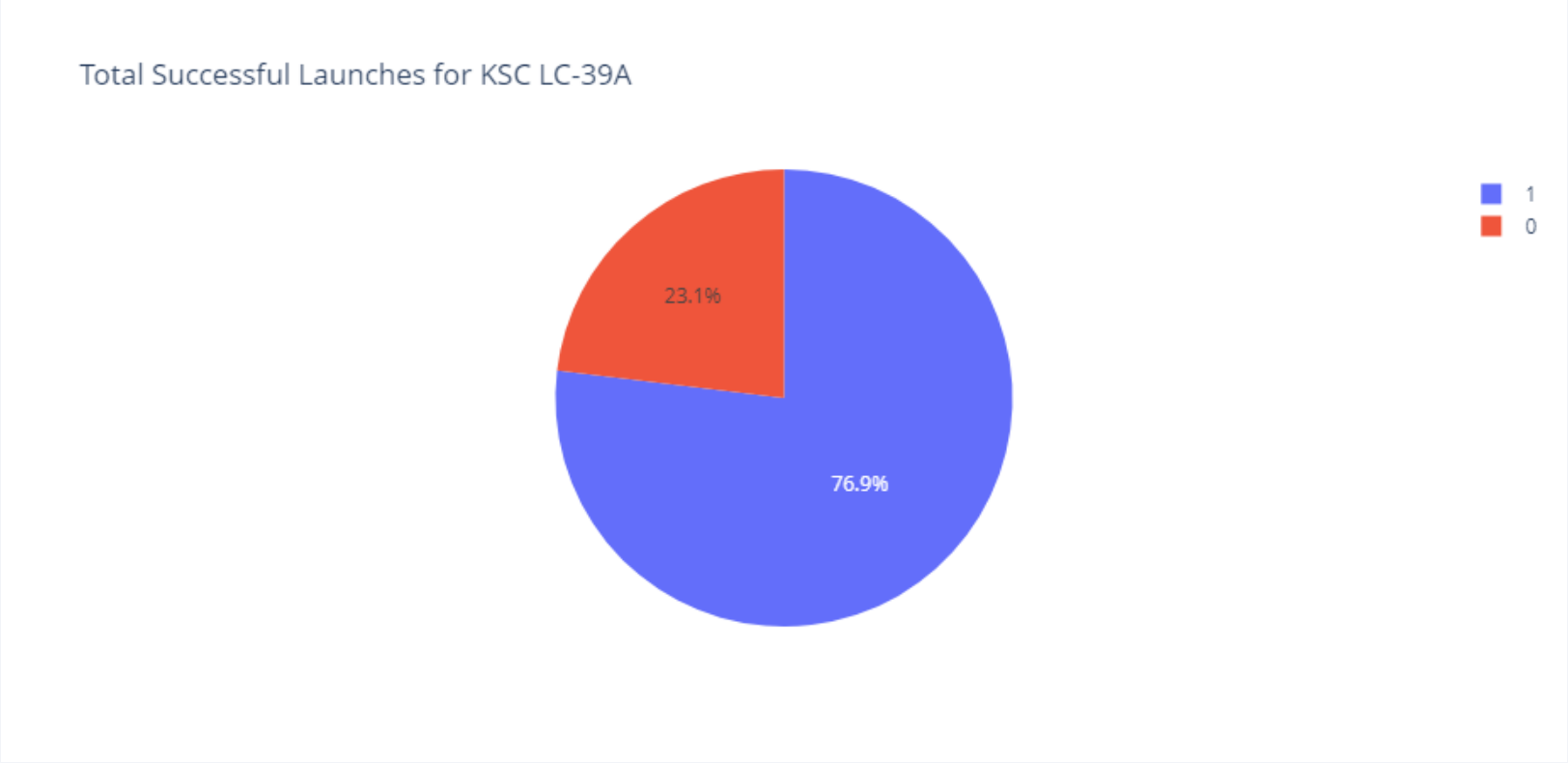
Launch Sites Success Rate

- The highest rates among all four launch sites is site KSC LC-39A, it is located on the Florida coast, west ward of the two other coasts located there as well.
- The least successful site is CCAFS SLC-40.



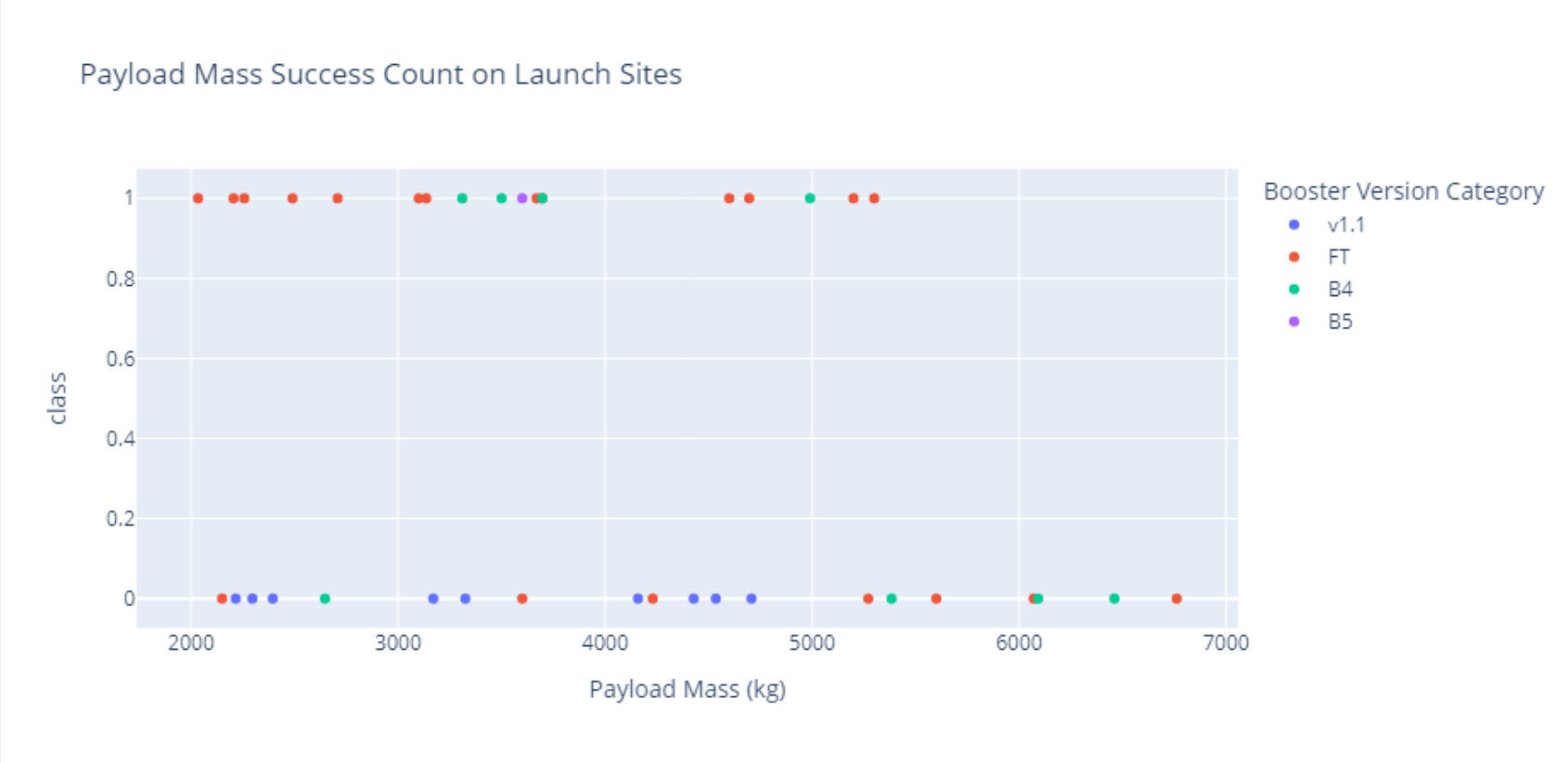
Highest Success Rate Launch Site: KSC LC-39A

- The launch site KSC LC-39A has the highest success to failure ratio standing at **79.6%.**



Payload Mass Success Count

- Based on its payload we could see that the booster with payload less than 4000 is that has the version of FT, B4, and B5 has high success rates, while booster v1.1 has not succeeded at all.

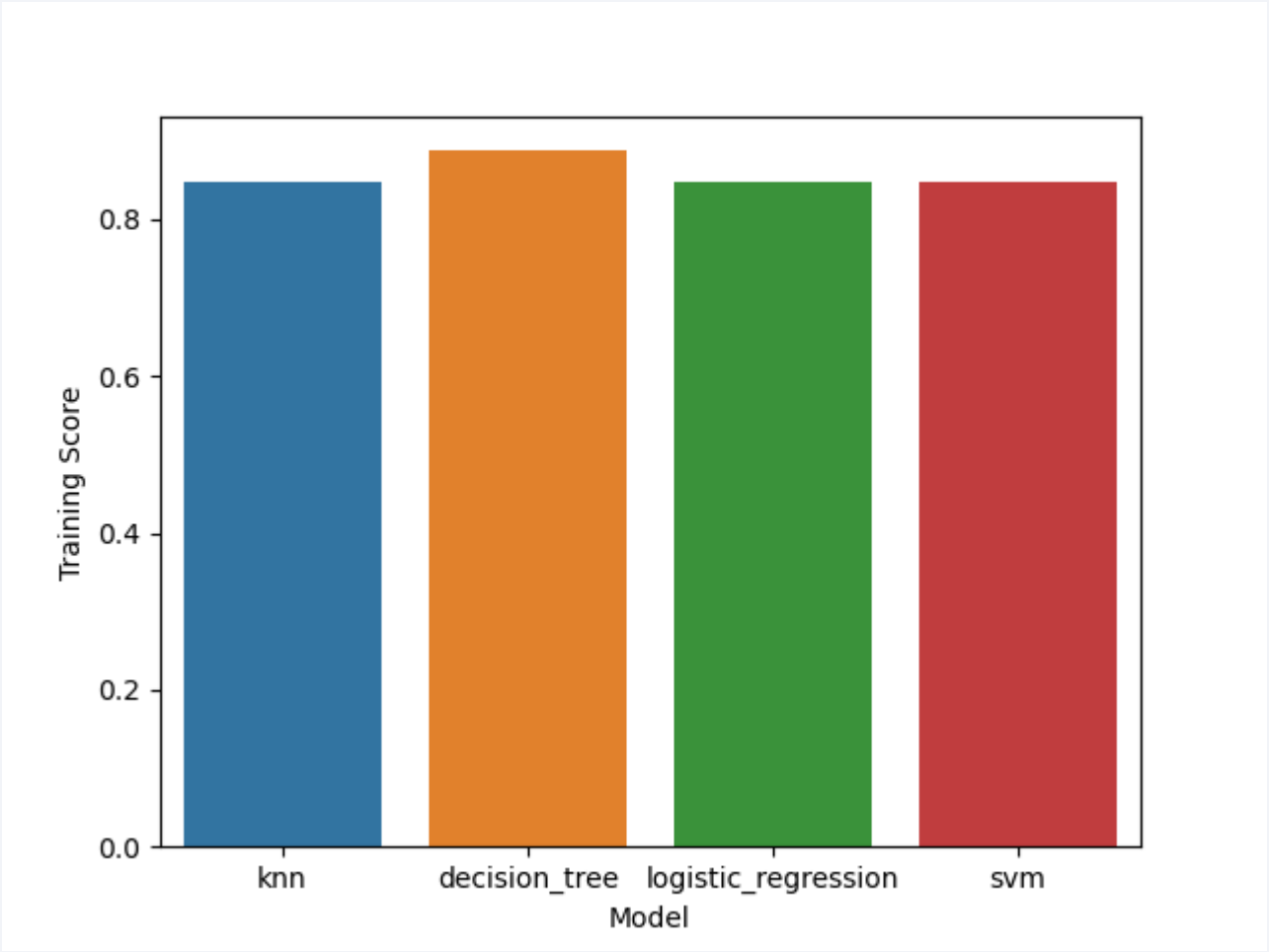


Section 5

Predictive Analysis (Classification)

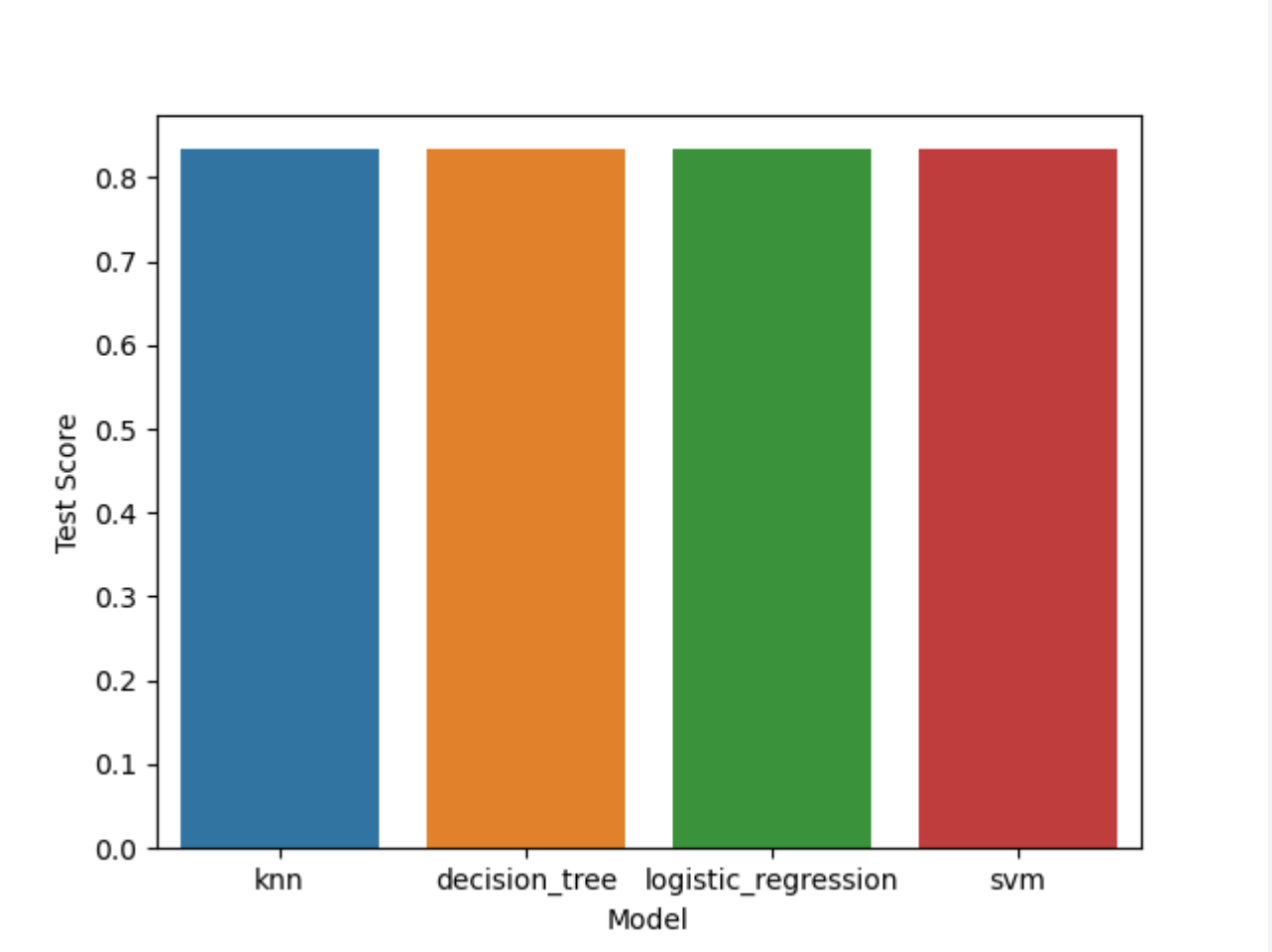
Classification Accuracy: Training Scores

- Among all models that were created, the Decision Tree Model created the model that works best with the Training Data.



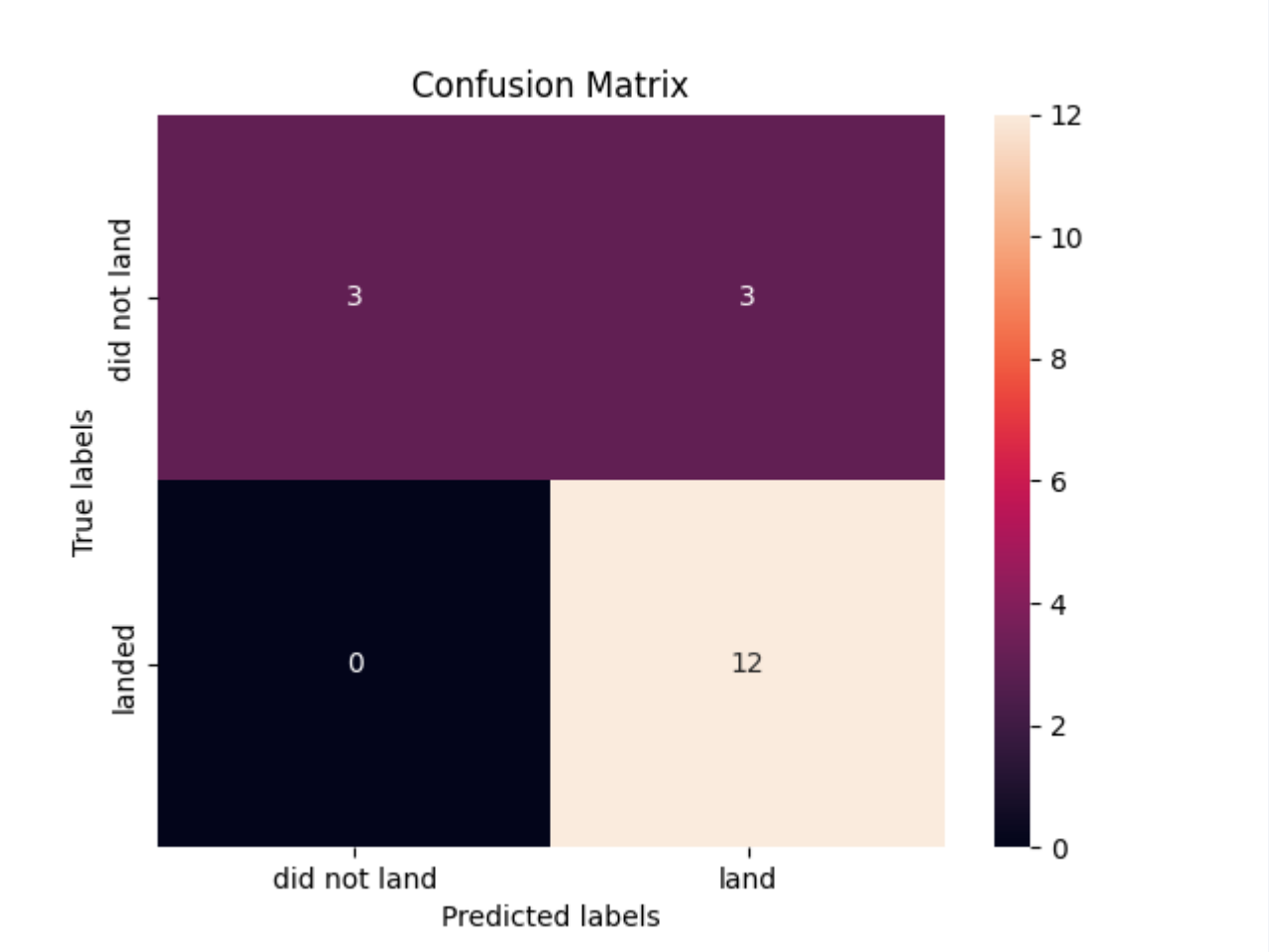
Classification Accuracy: Testing Scores

- During testing, we found out that all models work the same with all receiving test scores of **0.8333**.



Confusion Matrix

- When we look deeper into the confusion matrix, we see that all four models produced the exact same matrix.
- They received 100% accuracy on True Positives, but only 50% accuracy with True Negatives when compared to False Positives.



Conclusions

- We found out that all launch sites are in extremely close to the equator and the coasts.
- SpaceX started the Falcon9 with a rocky success rate, but it started to stabilize and became more successful as launches increase.
- The most successful launch site is the KSC LC-39A launch site.
- The Falcon 9 has transported **45596 kilograms** worth of payload among the data provided in the data set.
- Most of the launches are made with a payload of around **7500** kg or less.
- SpaceX launched most of its payload to LEO, ISS, PO, GTO, and VLEO type orbit.
- With the model we have created, we are only able to accurately predict the outcome only **83.33%** of the time.

Appendix

- The GTO orbit type, has the highest number of payload sent to.

```
[6]: # Apply value_counts on Orbit column
     df['Orbit'].value_counts()
```

```
[6]: GTO      27
     ISS      21
     VLEO     14
     PO        9
     LEO       7
     SSO       5
     MEO       3
     ES-L1     1
     HEO       1
     SO        1
     GEO       1
     Name: Orbit, dtype: int64
```

Thank you!