IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yohan Mohan Markose
June 6th 2024

Outline

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

Executive Summary

- Project Overview: This data science project aimed to analyze and predict the success of SpaceX rocket launches using data collected from multiple sources. The ultimate goal was to identify key factors influencing successful landings and to develop a reliable predictive model.
- The analysis indicates that specific parameters, such as payload mass, launch site, flight number, and booster version, play crucial roles in the success of SpaceX rocket launches. By leveraging the Decision Tree classifier, SpaceX can better predict the success of future launches, potentially optimizing costs and improving reliability. These insights are valuable for operational planning and strategic decision-making in future missions.
- This comprehensive approach provides a robust framework for understanding and predicting SpaceX launch outcomes, contributing to more efficient and cost-effective space missions.

Introduction

- Objective: To predict if the Falcon 9 first stage will land successfully and in turn determine the cost of a launch.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- There are several factors that could influence the successful landing of a rocket. The booster version used, payload mass, launch site, distance of the launch site to its proximities.
- We will collect the data needed from Api and websites, examine the factors that influence the success and create a predictive analysis that can predict the outcome of a launch and thereby helping us to estimate the cost.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - ❖ Data was collected using two methods:
    - ➢ Space X API from spacexdata.com
    - ➢ Web scraping from Spacex_Wikipedia
- Perform data wrangling
  - ➢ Combined the API data and web scraped data
  - ➢ Identified the numerical and categorical columns
  - ➢ Created a landing outcome label ('Class') In the data where successful and Failed landings where categorized by values 1 and 0 respectively.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize the data
  - Split it into training and testing data
  - Find best hyperparameters for different classification models
  - Choose the most accurate model

Data Collection

- Data was collected from two main sources: the Space X Wikipedia page and through the Space X API.
  - Space X API :spacexdata.com
  - Spacex Wikipedia
- The collected data was combined and missing values were dropped.
- A new categorical column was created that specifies whether the landing was a success or a failure, corresponding to the 4 different Launch sites

Data Collection – SpaceX API

After receiving the general API response, a function was written to get further details from specific columns using specific API calls.

- **Columns specific API calls**:
  - 'rockets': `"https://api.spacexdata.com/v4/rockets/"`
  - 'launchpad':`"https://api.spacexdata.com/v4/launchpads/"`
  - 'payload':`"https://api.spacexdata.com/v4/payloads/"`
  - 'cores':`"https://api.spacexdata.com/v4/cores/"`

- This was then collected and made into a data frame
- GitHub Link

API link:

```
"https://api.spacexdata.com/v4/launches/past"
        ↓
response = requests.get(spacex_url)
        ↓
data = pd.json_normalize(response.json())
        ↓
    (API calls)
    "https://api.spacexdata.com/v4/rockets/"
    "https://api.spacexdata.com/v4/launchpads/"
    "https://api.spacexdata.com/v4/payloads/"
    "https://api.spacexdata.com/v4/cores/"
        ↓
df = pd.DataFrame(launch_dict)
```

Data Collection - Scraping

- Scraped the space x table from Wikipedia
- Identified important columns
- Formatted the column values
- Created a data frame by parsing the html table values

```
response = requests.get(static_url)
```

↓

```
soup = BeautifulSoup(response.text, "html.parser")
```

↓

```
html_tables = soup.find_all(name="table")
```

↓

Created Launch Dictionary variable form table values

↓

```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

- Github Link

Data Wrangling

- Exploratory Data Analysis was done using pandas and NumPy libraries in python, to find some patterns in the data and determine what would be the label for training supervised models.
- We Identified the numerical and categorical columns and created a categorical column named ‘Class’ that shows the success of each launch corresponding the launching site.
- The ‘Class’ column was created by turning outcomes into Training Labels with `1` implying the booster successfully landed `0` implying, it was unsuccessful.
- GitHub Link

- Combined the API data and web scraped data
- Identified the numerical and categorical columns
- Created a landing outcome label (‘Class’) In the data where successful and Failed landings where categorized by values 1 and 0 respectively.

EDA with Data Visualization

Charts were used to understand the relationship between column data:

Scatter Plot: To find the relationship between Flight number, payload mass, Orbit type and Launch sites with respect to the success rate of the launch ('Class'). By combining the relationship between these data 2 at a time we found that:

- As Flight number increases -> more likely the first stage is successful
- More the payload mass -> less likely first stage successfully lands
- Some launch sites do not have rockets for heavy payloads

Bar Chart: To find relationship between success rate and orbit type. Bar chart is ideal to visualize categorical insights and the same was the reason it was used to understand the relationship between Orbit type and Success of launch of a rocket.

Line Chart: To find average launch success rate by year

GitHub Link

EDA with SQL

- Displayed the names of unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successfully landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20
- GitHub Link

Build an Interactive Map with Folium

- A folium Map object was created, initially with NASA Johnson space center at its center. All Launch sites were marked on the map using `folium.Circle()` to add a highlighted circle area with a text label on each of the coordinates. 'folium.Marker()', was used to add a pop up marker on all launch sites.
- A `MarkerCluster()` object was used to add the outcomes markers with green implying a successful landing and red, a failure. This was to differentiate between outcomes for different launches on each launch site.
- Additionally, a `MousePosition()` object was used on the map to get coordinates for a mouse over a point on the map. This helps in calculating the distance of the launch sites to its proximities such as railway, highway, cities etc.
- Finally, a 'PolyLine()' object was used to draw a line between the launch sites and its proximities, whose coordinates were found using the 'MousePosition()' object

Build an Interactive Map with Folium

Finding an optimal location for building a launch site certainly involves many factors such as location and proximities of a launch site, i.e., the initial position of rocket trajectories.

The Folium objects and markers can be used to identify why each launch sites where used and how effective they were. This will help in choosing an ideal launch site in the future.

GitHub Link

Build a Dashboard with Plotly Dash

- The Space X launch records dashboard contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart
- The dropdown list allows us to select the data of either any one of the four launch sites, or the data of all sites together.
- The pie chart that forms when the selection is made, shows the success rate ("Class") data of either individual sites or all sites when compared together
- A scatter plot is also produced to observe how payload may be correlated with mission outcomes for selected site(s)
- A range slider is also given to enable adjustment of payloads to find if variable payload is correlated to mission outcome
- GitHub Link

Predictive Analysis (Classification)

Standardize the data: Normalize features to have a mean of 0 and a standard deviation of 1.

Split it into training and testing data: Divide the dataset into two parts, one for training the model and one for evaluating its performance.

Find best hyperparameters for different classification models: Use techniques like grid search or random search to optimize model parameters.

Choose the most accurate model: Select the model with the highest accuracy based on testing data performance.

GitHub Link

Library Used : sklearn

↓

Standardize the data : Using `StandardScaler()`

↓

Training and testing data: Test size 20% and random state= 2

↓

Best hyperparameters for different classification models: Using `GridSearchCV()`, `best_params_`

- & `best_score_`

↓

KNN, Logistics Regression, Decision Tree, SVM

↓

Choose the most accurate model by comparing accuracy of all models used

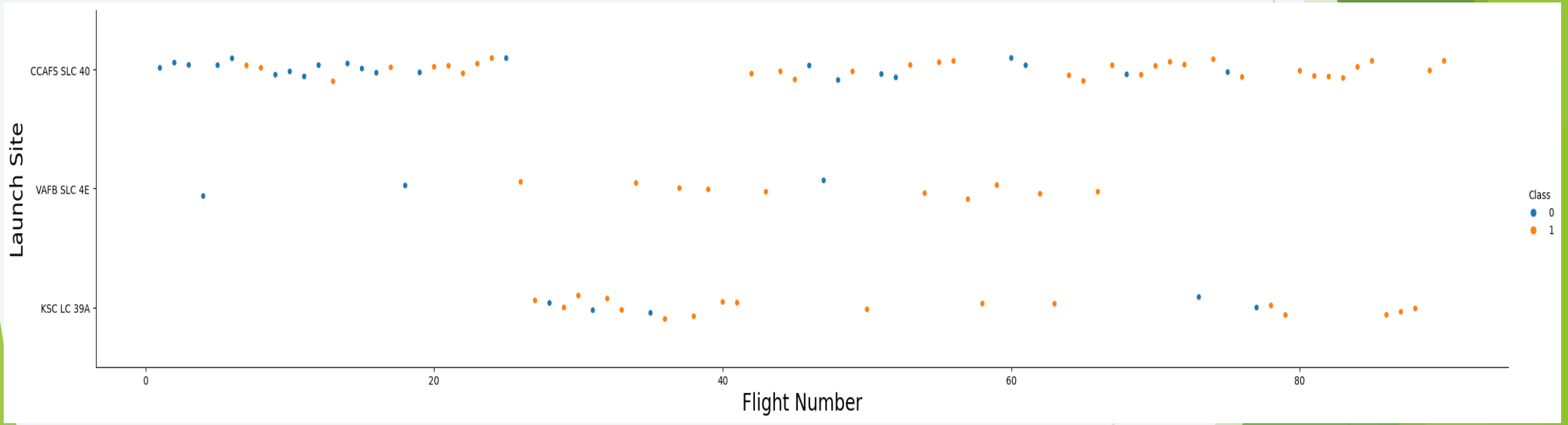
Results

Section 2

Insights drawn from EDA

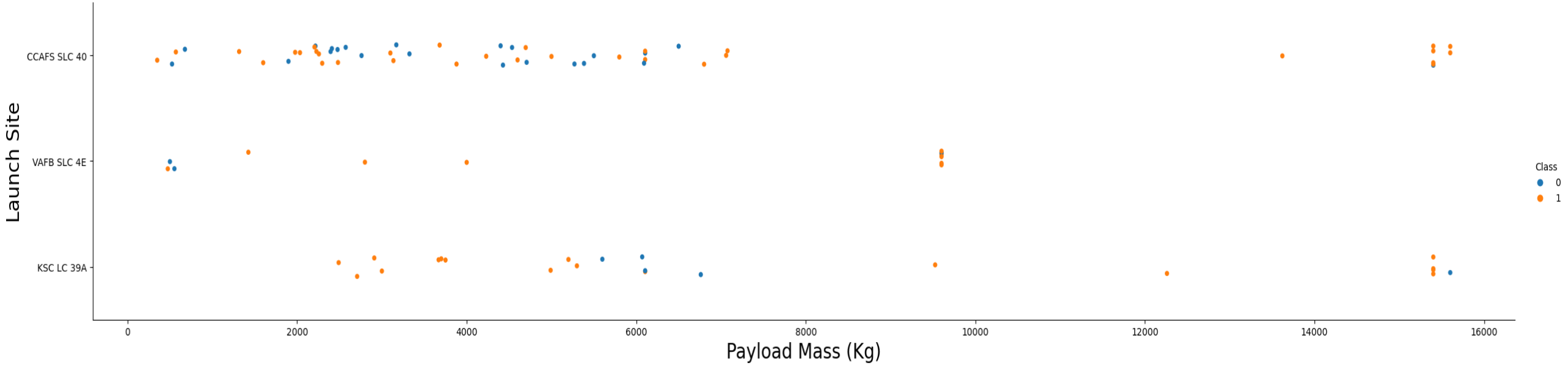
Flight Number vs. Launch Site

The Scatter plot below shows that for each launch site, as the flight number increases the likelihood that the landing is successful increases.



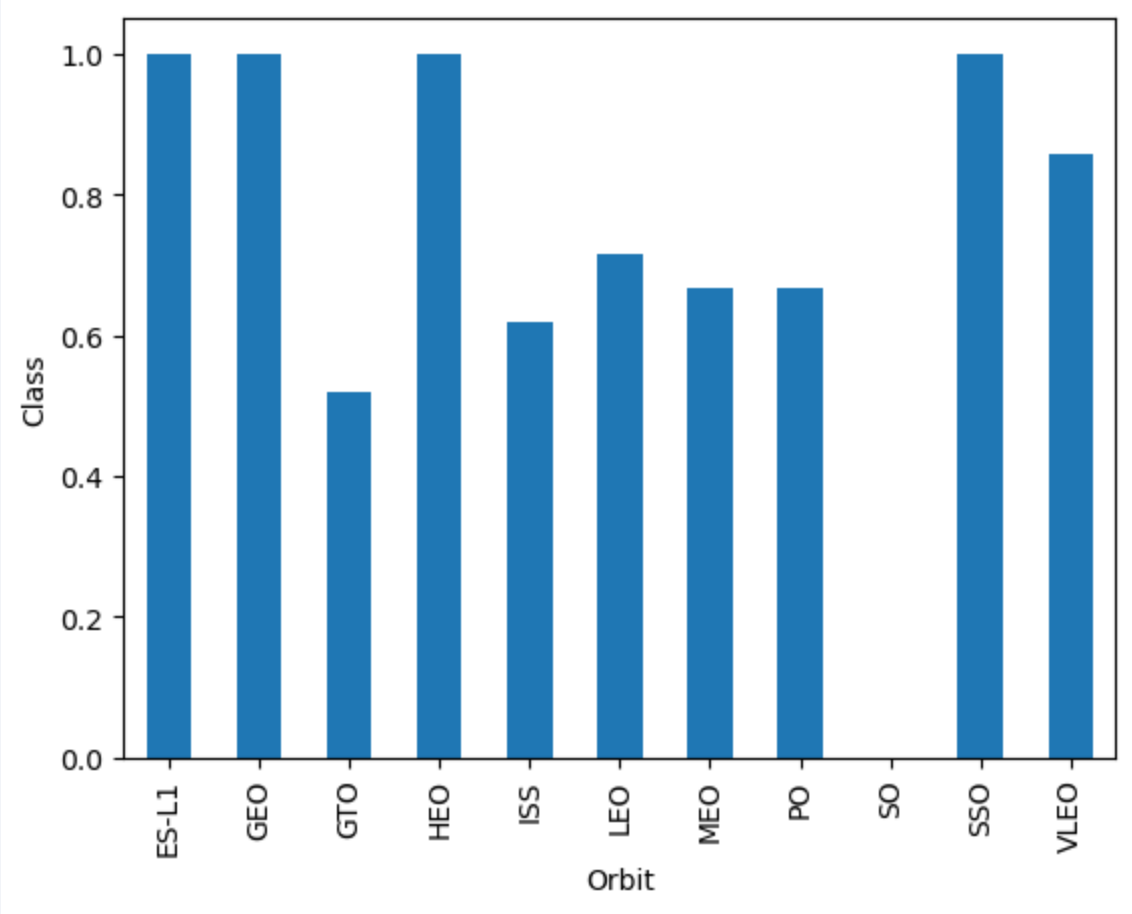
Payload vs. Launch Site

If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).



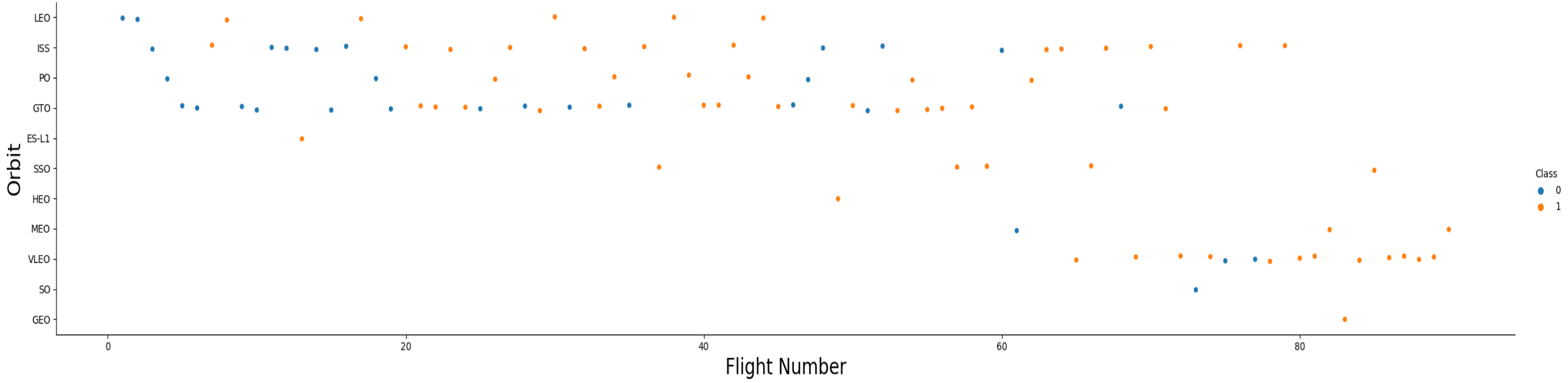
Success Rate vs. Orbit Type

It is observed through the bar chart that the success rate for orbit types ES-L1, GEO, HEO, SSO are more compared to the others



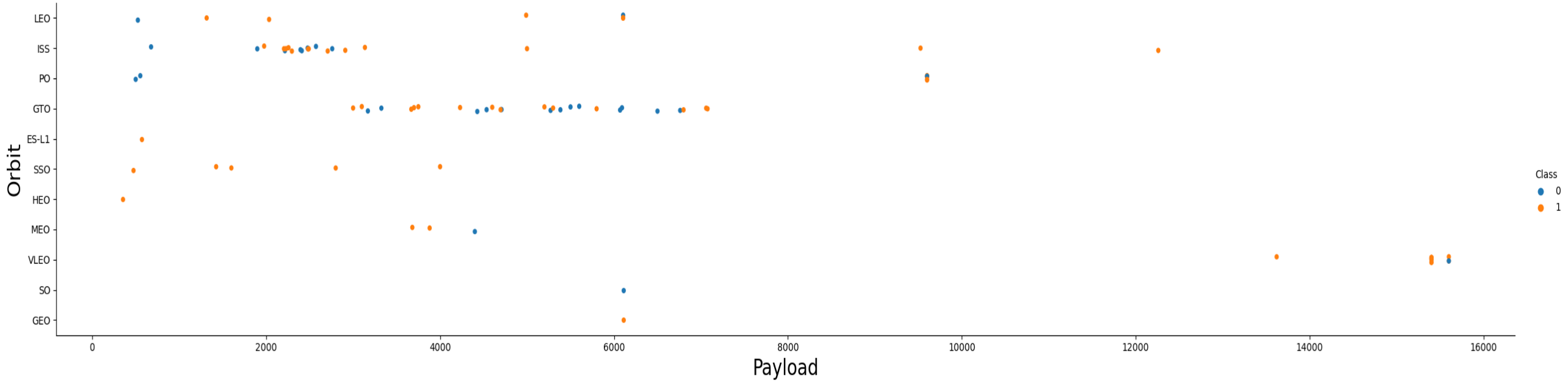
Flight Number vs. Orbit Type

We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



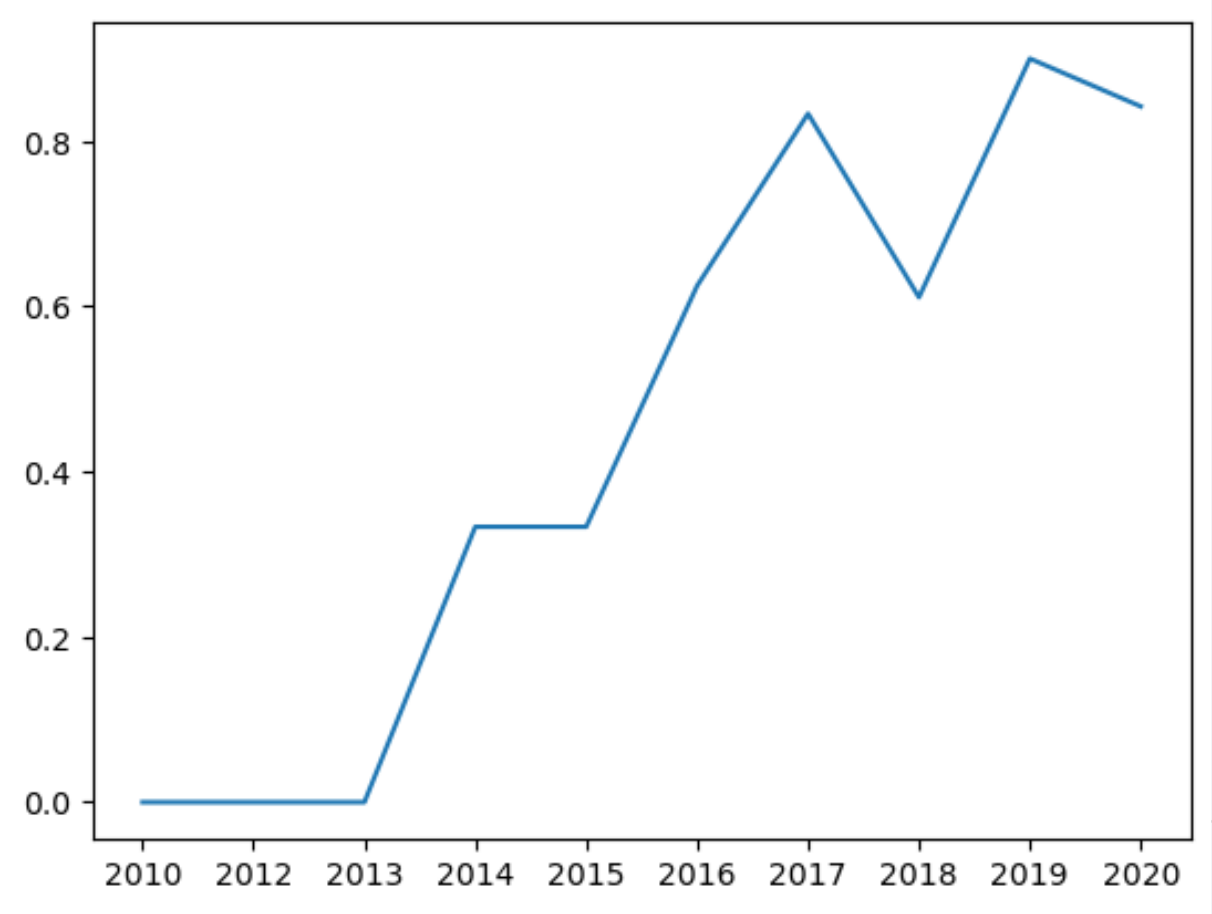
Payload vs. Orbit Type

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



Launch Success Yearly Trend

We can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing



All Launch Site Names

SQL Query

```
SELECT distinct("Launch_Site") FROM SPACEXTBL
```

Finding

There are a total of 4 unique launch sites in our space x data

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

Launch Site Names Begin with 'CCA'

SQL Query

- `SELECT *`
- `FROM SPACEXTBL`
- `WHERE Launch_Site LIKE 'CCA%'`
- `LIMIT 5`

Finding

All records shown below are from CCAFS LC-40 launch site

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

SQL Query

```
SELECT SUM("PAYLOAD_MASS__KG_") AS 'Total Payload Mass'
FROM  SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

Finding

The total payload mass carried by booster launches by NASA (CRS) is 45596 Kg

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

SQL Query

```
SELECT AVG("PAYLOAD_MASS__KG_") AS 'Average Payload Mass'
FROM  SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%'
```

Finding

The average payload mass carried by booster version F9 v1.1 is 2534.66 Kg

Average Payload Mass
2534.6666666666665

First Successful Ground Landing Date

SQL Query

```
SELECT MIN("Date")
FROM  SPACEXTBL
WHERE Landing_Outcome LIKE 'Success%'
```

Finding

The first successful landing outcome in ground pad was achieved on 2015-12-22

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
SELECT "Booster_Version"
FROM  SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

Finding

Shown below are the names of the booster versions which have success in drone ship and have payload mass greater than 4000 but less than 6000:

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

Total Number of Successful and Failure Mission Outcomes

SQL Query

```
SELECT
    COUNT(CASE WHEN Landing_Outcome LIKE 'Success%' THEN 1 END) AS 'Total Success',
    COUNT(CASE WHEN Landing_Outcome LIKE 'Failure%' THEN 1 END) AS 'Total Failure'
FROM SPACEXTBL
```

Finding

There was a total of 61 successful landings and 10 failures

Total Success	Total Failure
61	10

Boosters Carried Maximum Payload

SQL Query

```
SELECT "Booster_Version"
FROM  SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Finding

The names of the booster versions which have carried the maximum payload mass are shown below:

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

SQL Query

```
SELECT
    substr(Date, 6,2) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTBL
WHERE substr(Date,0,5)='2015'
    AND Landing_Outcome = 'Failure (drone ship)'
```

Finding

Below are the records with failed landing outcomes in drone ship, for the months in year 2015

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

SQL Query

```
SELECT Landing_Outcome, COUNT(*) AS Count, RANK() OVER (ORDER BY COUNT(*) DESC) AS Rank
FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome
ORDER BY Rank
```

Finding

We can see that successful drone ship landings comes at rank 2 just below the number of no attempts.

Landing_Outcome	Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

Section 3

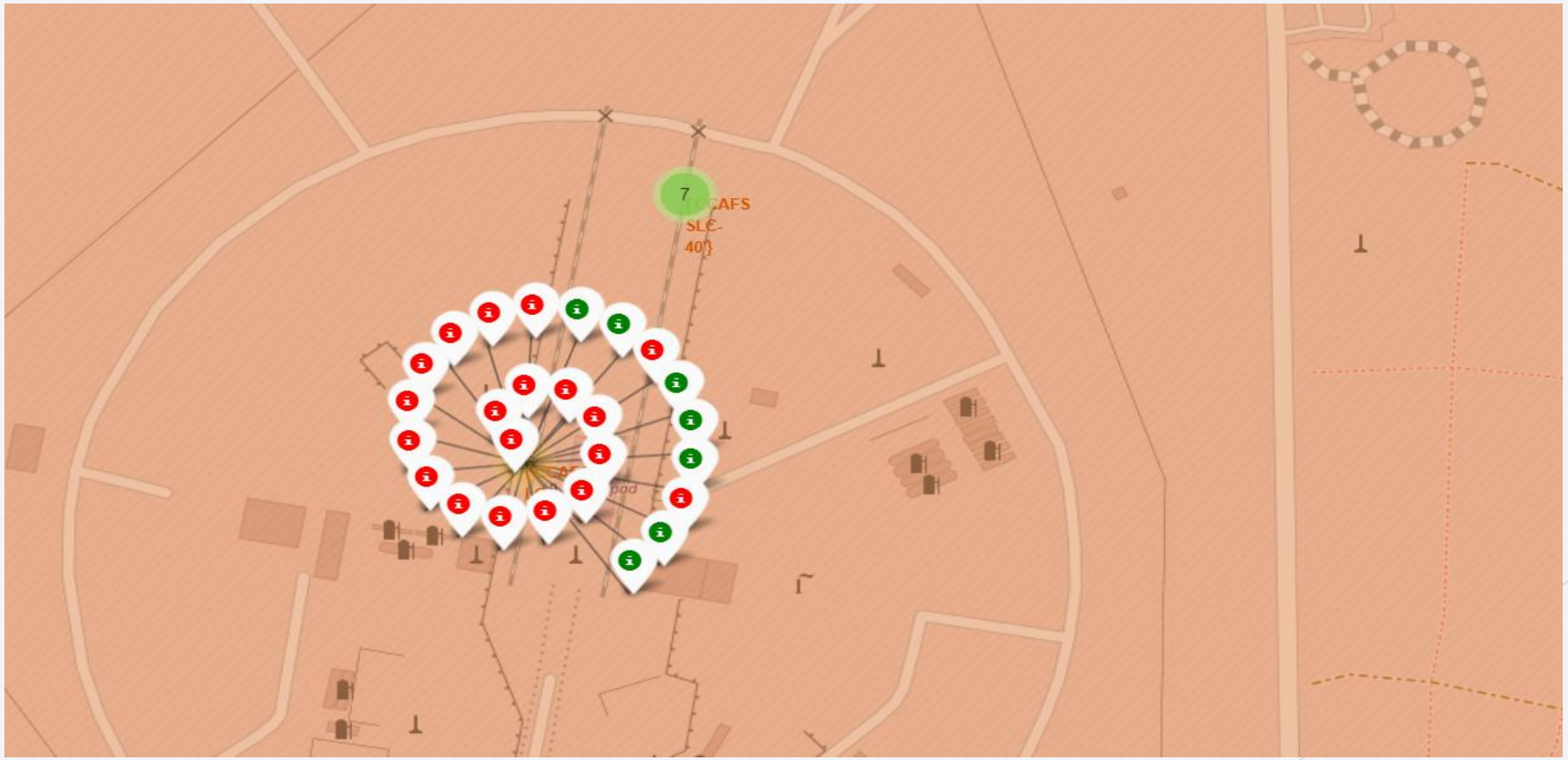
Launch Sites Proximities Analysis

Space X Launch Sites

The WAFB SLC 4E launch site is at completely different location compared to other launch sites

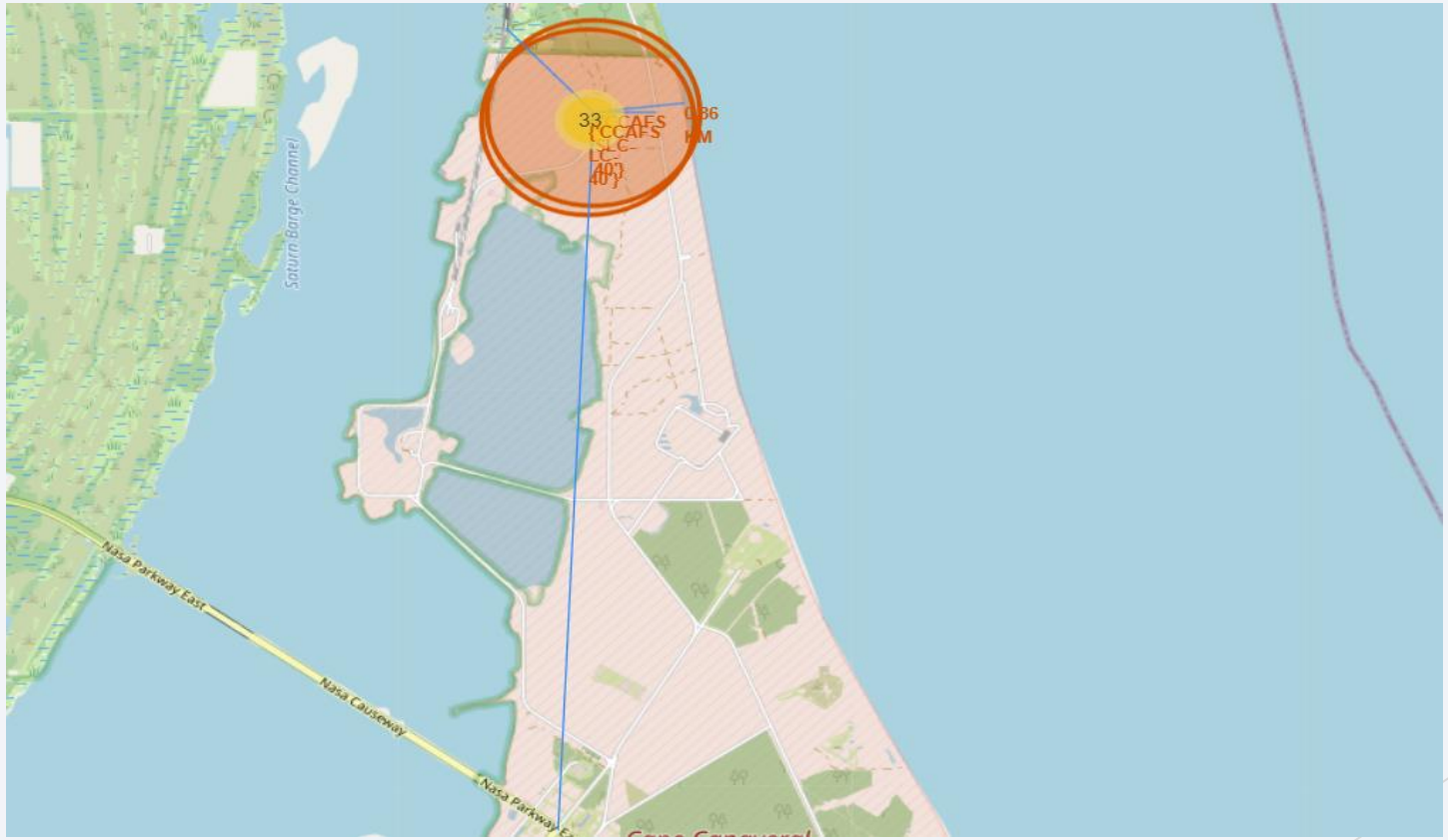
Successful/Failed Outcomes on Launch sites

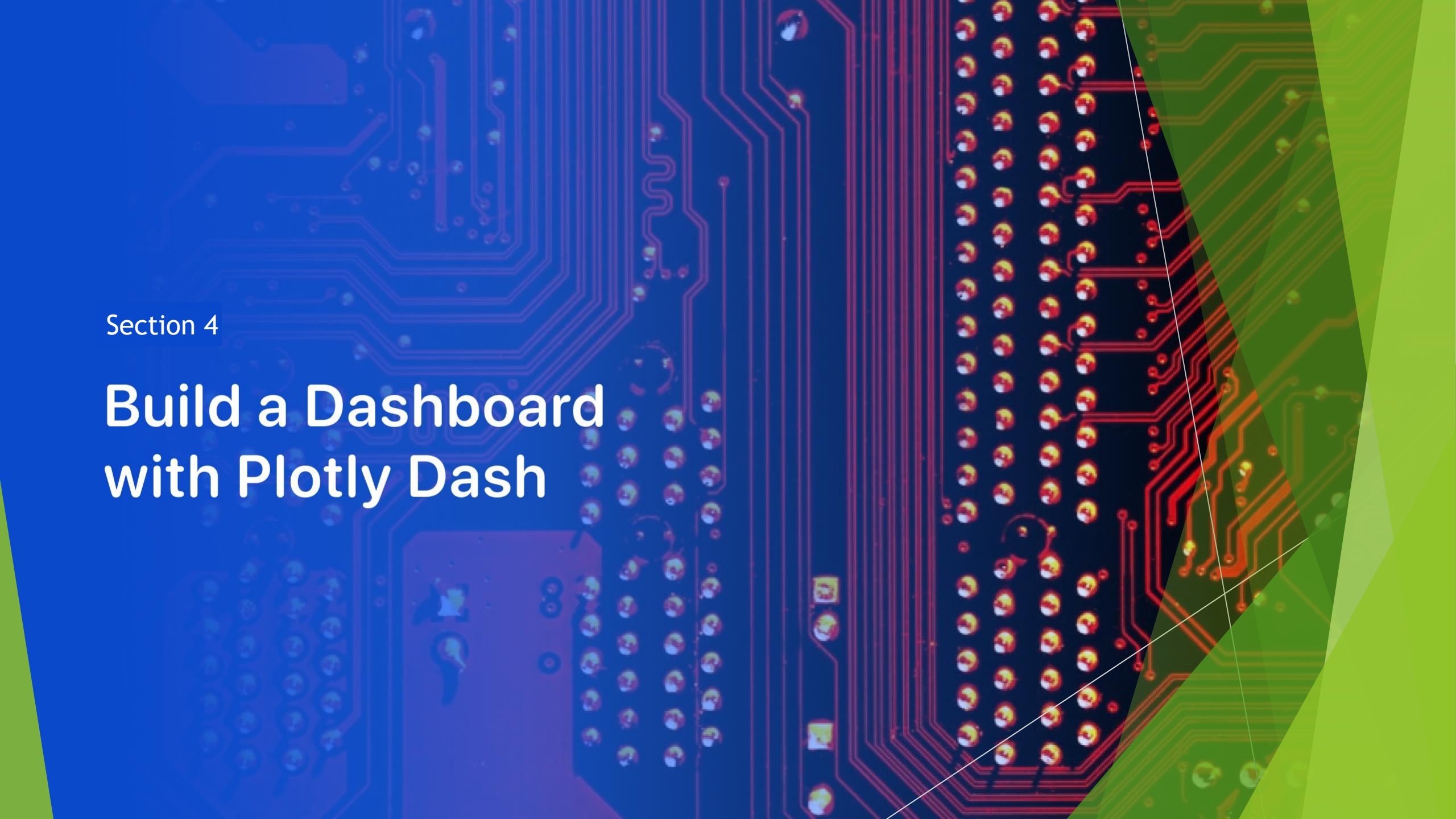
- The outcomes markers shown below shows successful and failed outcomes, with green implying a successful landing and red, a failure



Launch Site Distance to its Proximities

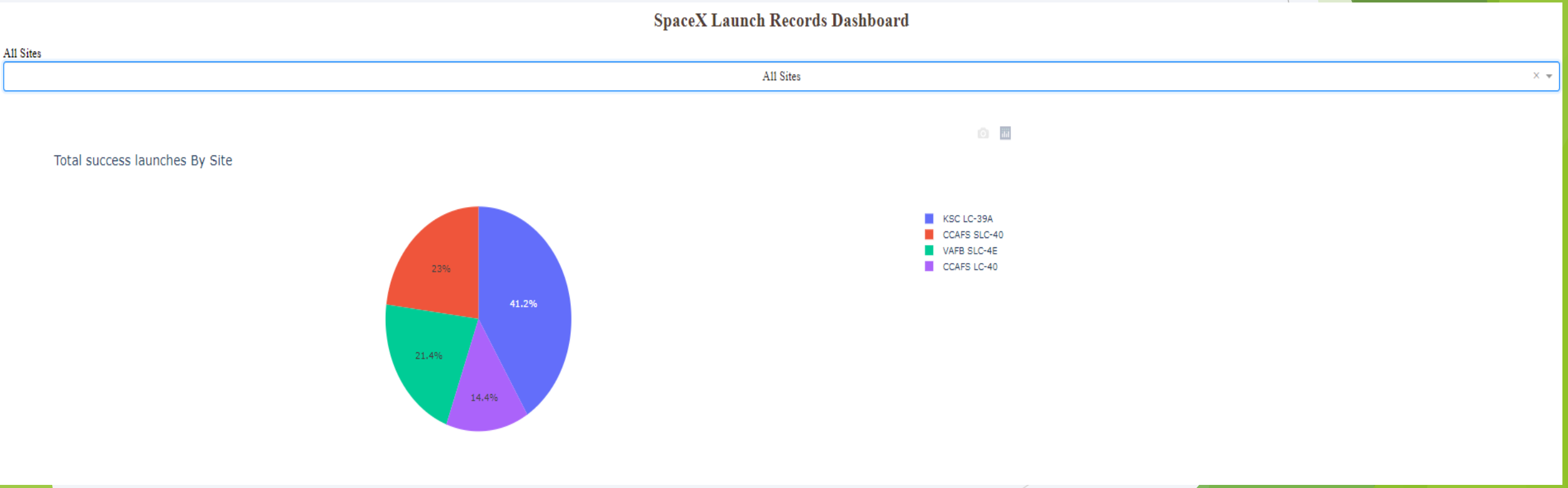
- The distance of each launch site to railways, highways, coastline and cities as shown in the map below, are essential factors to consider when choosing a launch site that achieves the most successful launches.




Section 4
Build a Dashboard with Plotly Dash

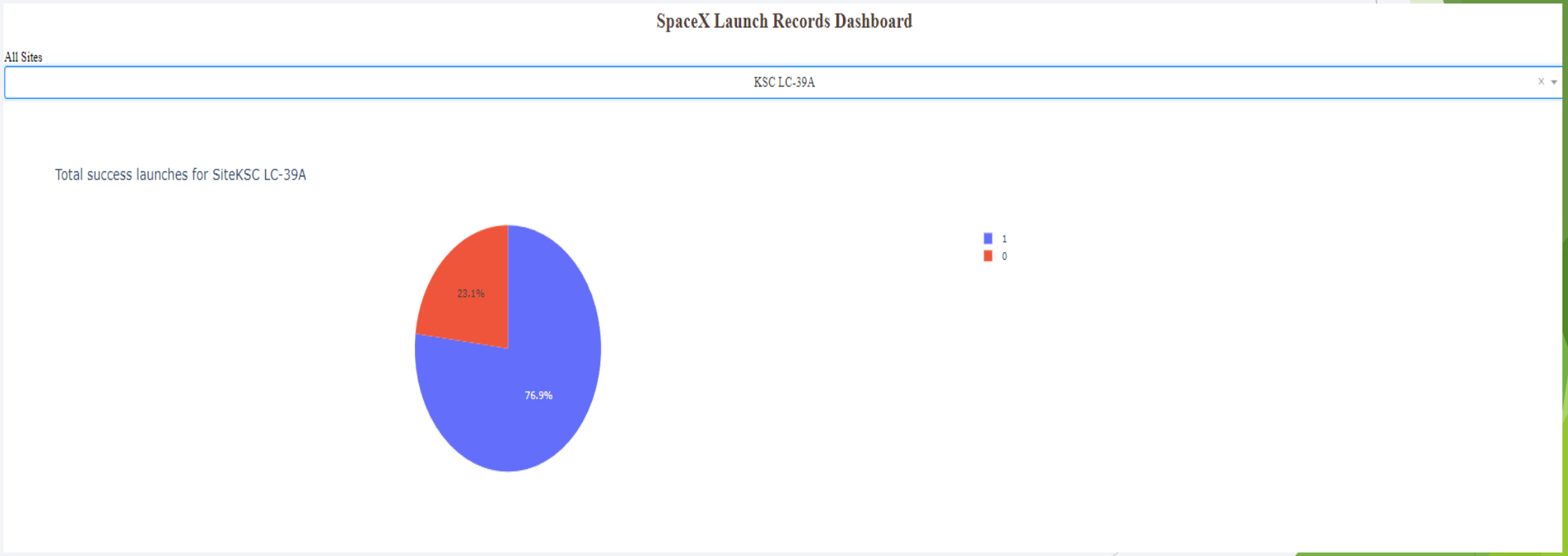
Success Rate of All Launch Sites (Pie chart)

- The pie chart shows the total successful launches data of all the sites with respect to one another. It is seen that the launch site KSC LC-39A has the highest success rate when compared to other launch sites



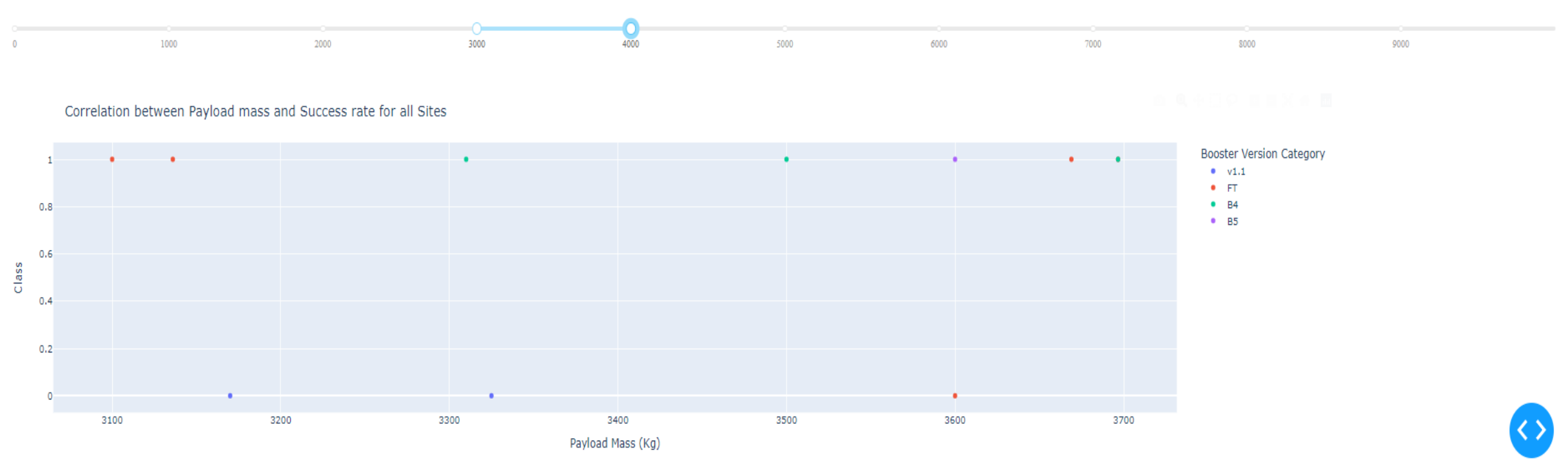
Launch site With Highest Success Rate

- The pie chart shows the Success to failure ratio for the launch site with the highest success rate (KSC LC 39A – 76.9%)



Payload Vs Launch Outcome – All Sites

- When selecting the full range of the payload mass (0 – 10000) it is observed that the booster version category FT has the most number of successful launches. In fact in most payload ranges FT has the most number of successful launches.
- Considering all sites, the success rate seems to be the highest in the payload range of 3000 to 4000

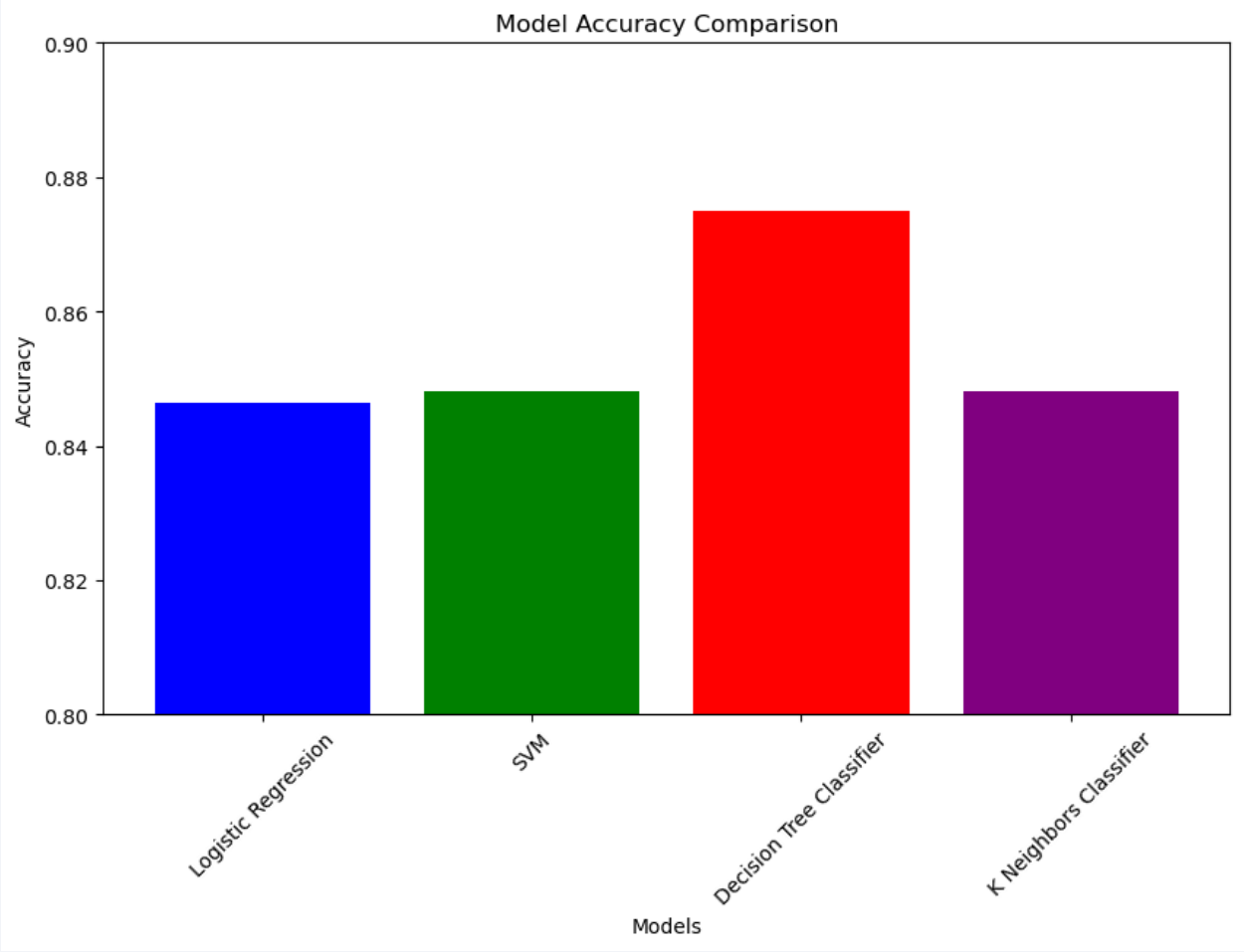


Section 5

Predictive Analysis (Classification)

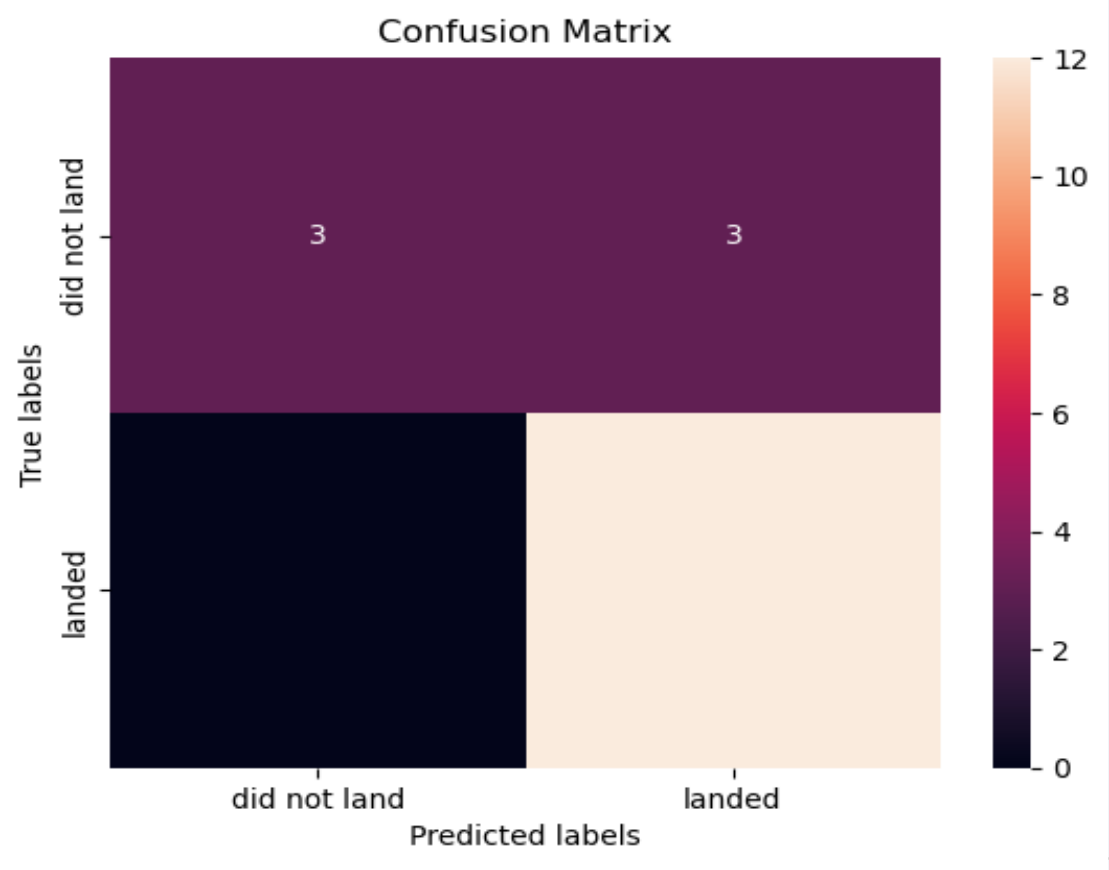
Classification Accuracy

- The Decision Tree classifier has the highest accuracy



Confusion Matrix

- The Confusion Matrix for Decision Tree Classifier is shown below.
- Examining the confusion matrix, we see that the model can distinguish between the different classes. We see that the major problem is false positives



Conclusions

- Space X data was collected from different sources and the launch outcomes have been examined with respect to different parameters. The most important parameters considered were the payload mass, launch site, flight number and booster versions.
- Through EDA and visualizations, we could observe that the ideal payload range for a success rate is between 3000 to 4000 kg
- We have also determined that the Decision Tree classifier model is the most accurate classification model to determine the success of a rocket launch. By keeping a few false positive cases in mind the model can be a reliable one.
- It can also be noted that the booster version category FT has given the most number of successful launches
- All these factors can be considered when calculating if a launch is successful or not. This will in turn help us understand if the cost of Space X rockets would be as low as they are expected to be

Appendix

- Model Performance Data frame to find the most accurate model was created with the code:

```
model_performance = {
    'Models': ['Logistic Regression', 'SVM', 'Decision Tree Classifier', 'K Neighbors Classifier'],
    'Accuracy': [
        logreg_cv.best_score_,
        svm_cv.best_score_,
        tree_cv.best_score_,
        knn_cv.best_score_
    ]
}
performance_df = pd.DataFrame(model_performance)
```

Thank you!