IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Stuart
Dec 5 2024

# Outline

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

# Executive Summary

In this project, I used many different data cleaning, data wrangling, visualization, and machine learning models to help determine if we can predict whether the first stage would land. After all the analysis and modeling, I came to the conclusion that we can predict whether or not the first stage will land at about 83% accuracy.

# Introduction

- SpaceX advertisies their Falcon 9 rocket launches on their website, with a cost of 62 million dollars, less than half the budget of their competitors. The purpose of this project is to predict if and where the first stage will land, which will help determine the cost of a launch.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Data was collected through the Space X API and via web scrapes of Wikipedia
- Perform data wrangling
  - Data was cleaned, and missing values were replaced.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Accuracy scores were analyzed of many different machine learning models

# Data Collection

- Data Sets were collected by:
  - SpaceX API requests
  - Web Scraping using BeautifulSoup

1. SpaceX Launch Data was requested
2. Data was decoded from JSON and turned into a Pandas Data Frame
3. Data was filtered to only included Falcon 9 Launches, and null values were placed with the mean where applicable.

# Data Collection – SpaceX API

https://github.com/stuart-reoch/capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

# Data Collection - Scraping

- https://github.com/stuart-reoch/capstone/blob/main/jupyter-labs-webscraping.ipynb

Request the Falcon9 Launch Wiki Page from its URL

Extracted all column/variable names from the HTML table header

Parsed the launch HTML tables and created a data frame

# Data Wrangling

- https://github.com/stuart-reoch/capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Calculated the number of launches on each site

Calculated the number and occurrences of each orbit

Calculated the number and occurrences of mission outcomes of each orbit

Created a landing outcome label from the Outcome column

# EDA with Data Visualization

- I used scatter plots to visualize relationships between variables such as Flight Number, Launch Site, and Payload Mass.

- https://github.com/stuart-reoch/capstone/blob/main/edadataviz.ipynb

Visualized relationships by using scatter plots

Visualized relationship between success rate and orbit type by using a bar plot.

# EDA with SQL

- Found the names of the unique launch sites in the space mission
- Found 5 records where the Launch Site began with the string ‘CCA’
- Found the total payload mass from boosters launched by NASA (CRS)
- Found the average payload mass from booster version F9 v1.1
- Found the date of the first successful landing outcome in ground pad
- Listed the names of the boosters which have been successful in drone ship, and had a payload mass of greater than 4000, but less than 6000
- Listed the total number of successful and failed missions
- Listed the names of the booster versions which carried the maximum payload mass by using a subquery
- Listed the failed landing outcomes in drone ship booster versions by launch site and month in the year 2015.
- Ranked the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order
- https://github.com/stuart-reoch/capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# Build an Interactive Map with Folium

- Markers were used to show Launch sites and their nearest mode of transportation, such as railways, highways, cities and the coast line.
- PolyLines were used to connect the launch site to the nearest mode.

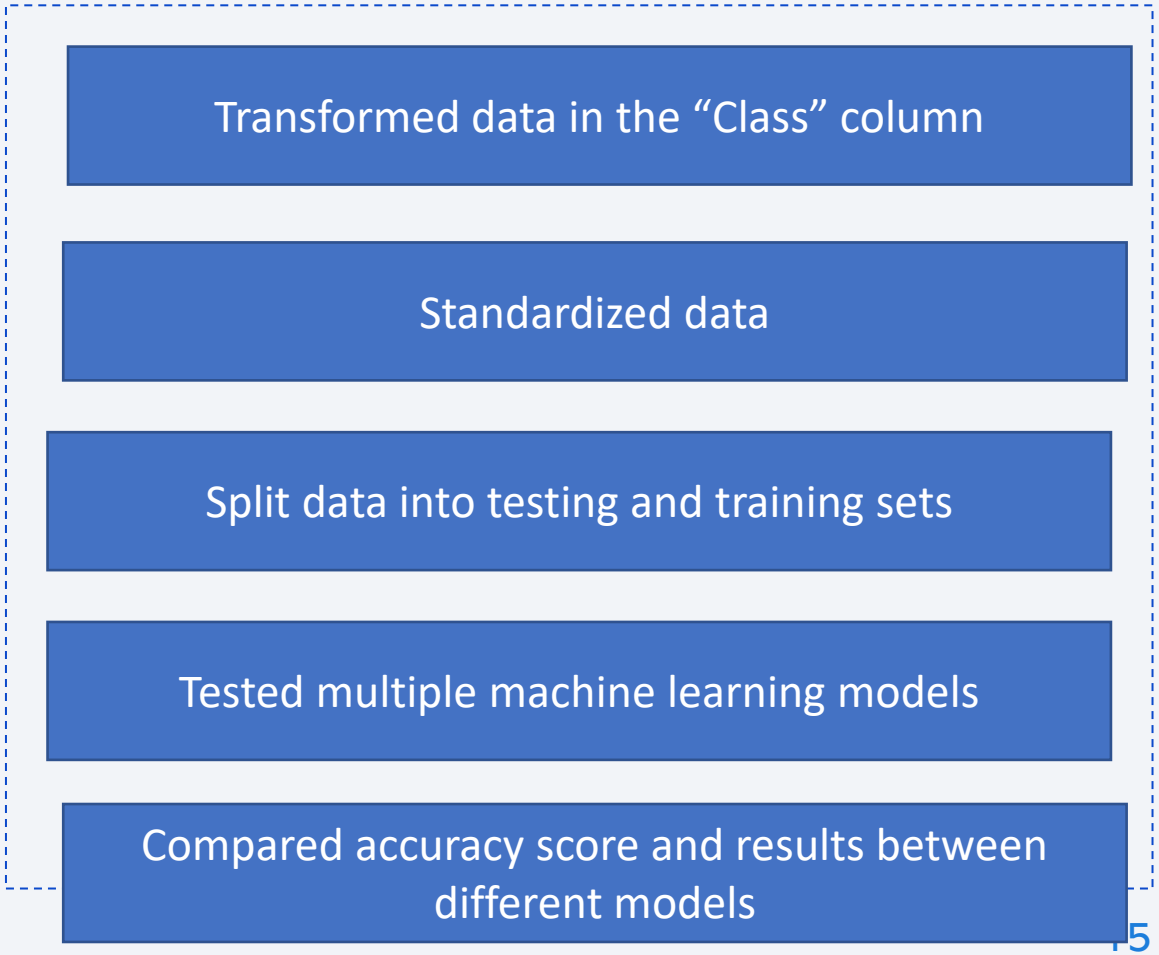
https://github.com/stuart-reoch/capstone/blob/main/lab_jupyter_launch_site_location.ipynb

# Build a Dashboard with Plotly Dash

- Pie graphs and scatter plots were used to visualize rocket launch success rate per launch site. These visualizations showed the relationship between different variables and how they influence the launch success rate.
- https://github.com/stuart-reoch/capstone/blob/main/spacex_dash_app.py

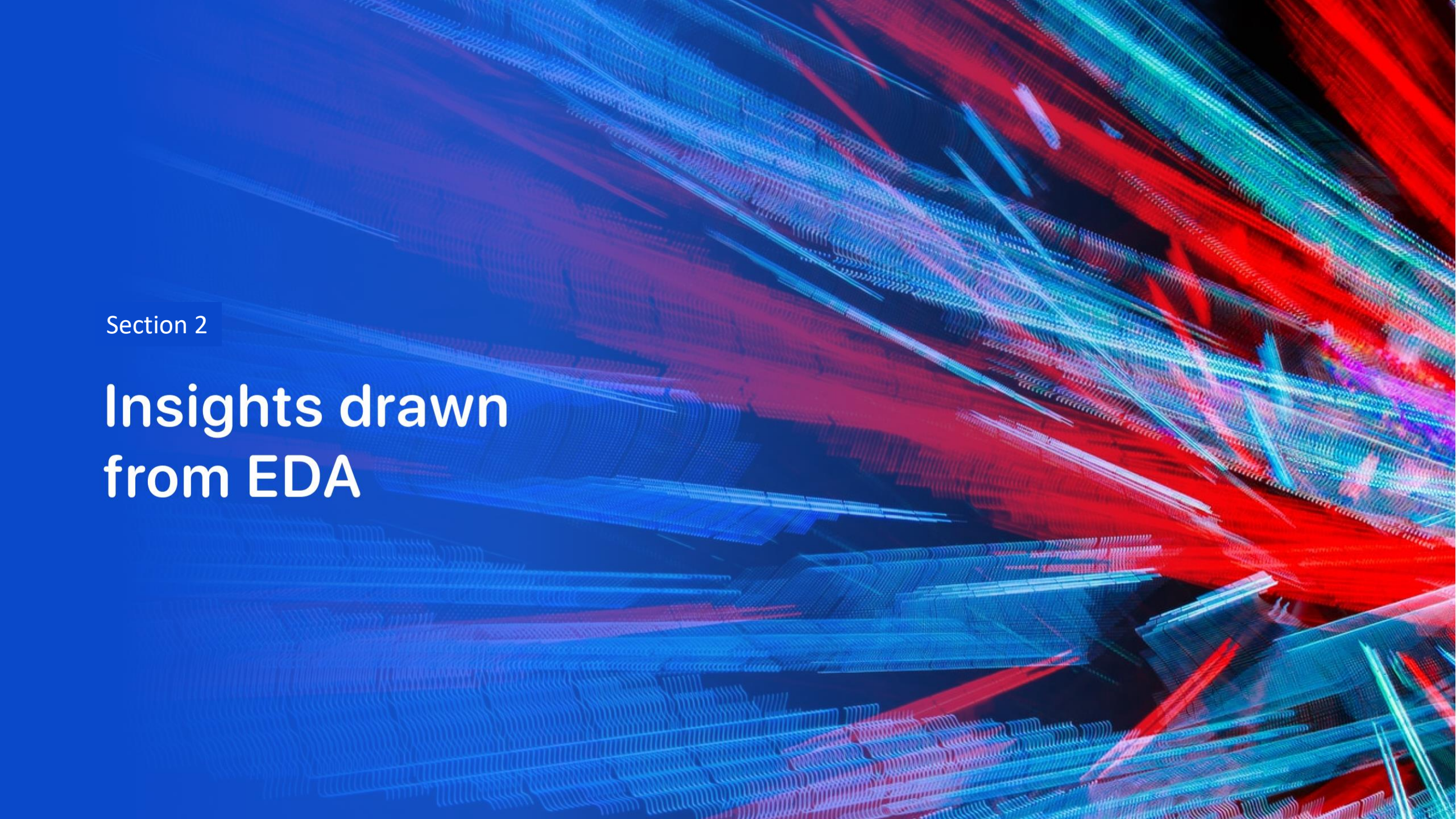
# Predictive Analysis (Classification)

- Many different machine learning models were used with the sci-kit learn package and the accuracy of these models was calculated and visualized using a confusion matrix.
- https://github.com/stuart-reoch/capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



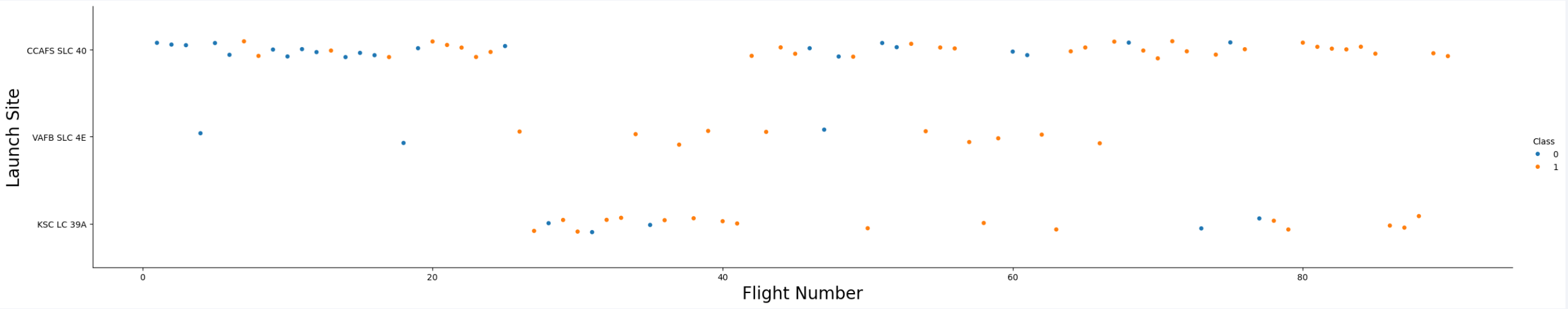
# Results

- The EDA process led me to conclude that successful landing outcomes are correlated with Flight Number. Also, the amount of successful landings has increased since 2015.
- All launch sites are near the coast line. I came to the conclusion this is for the safety of the population.
  - Sites are also near transportation modes such as highways and railways, perhaps to save costs and time on transportation.
- Our Machine Learning Models were quite accurate, with a score of around 83%


Section 2
Insights drawn from EDA

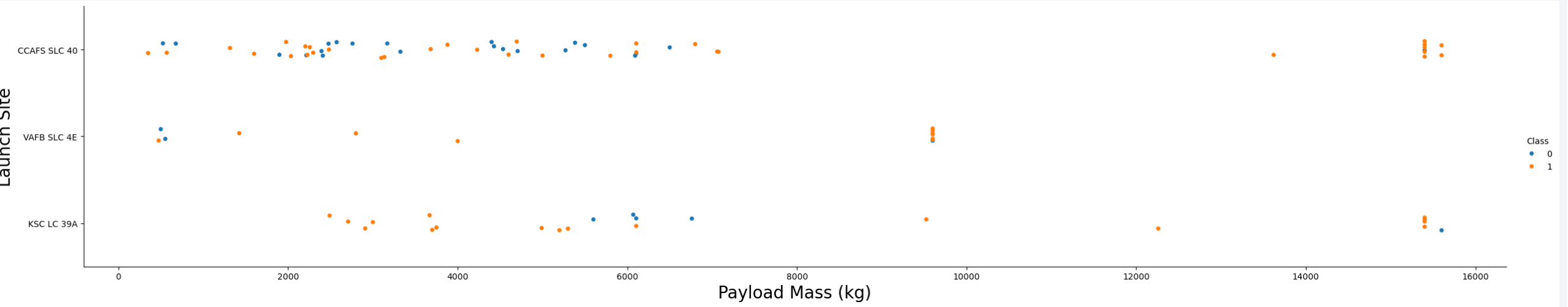
# Flight Number vs. Launch Site

- Flights were more successful as flight numbers increased. The launch site CCAFS SLC 40 had the most attempts, and most successful landings. The VAFB SLC 4E launch site only had 3 failed landings



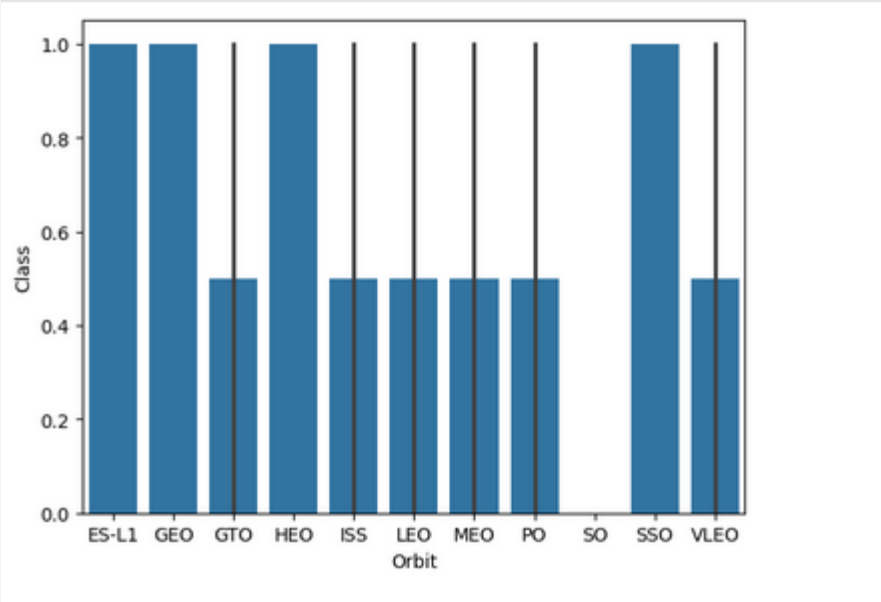
# Payload vs. Launch Site

- There are very few launches of rockets that have a payload mass of between 10000 and 16000 kg.



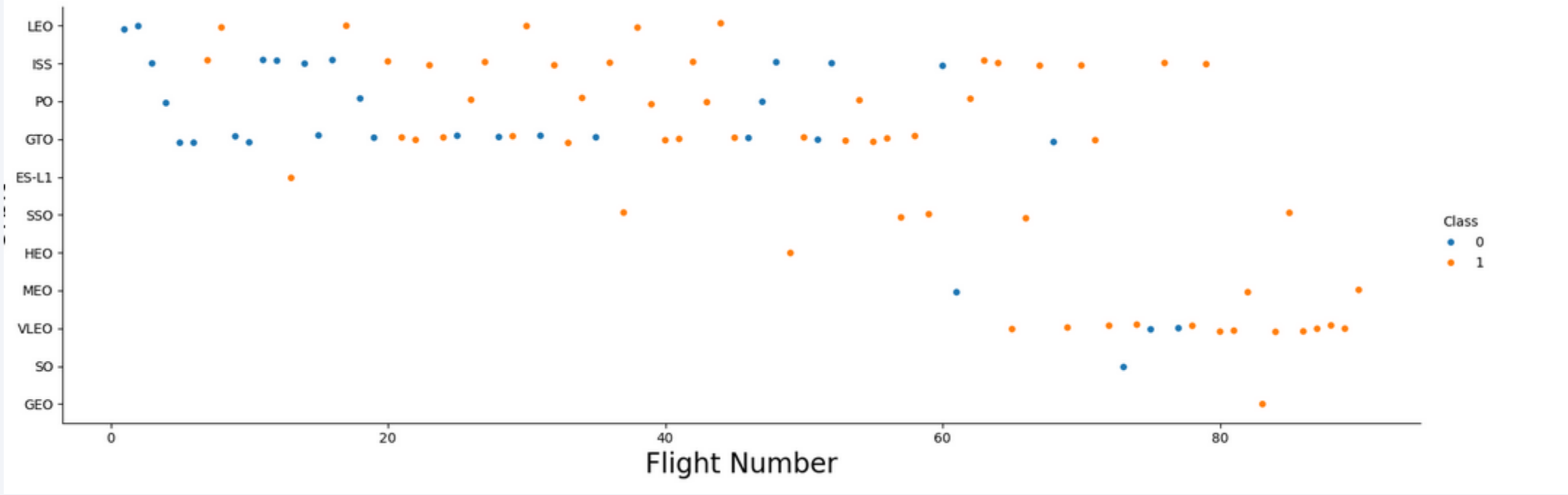
# Success Rate vs. Orbit Type

- The following orbit types were the most successful:
  - ES-L1
  - GEO
  - HEO
  - SSO



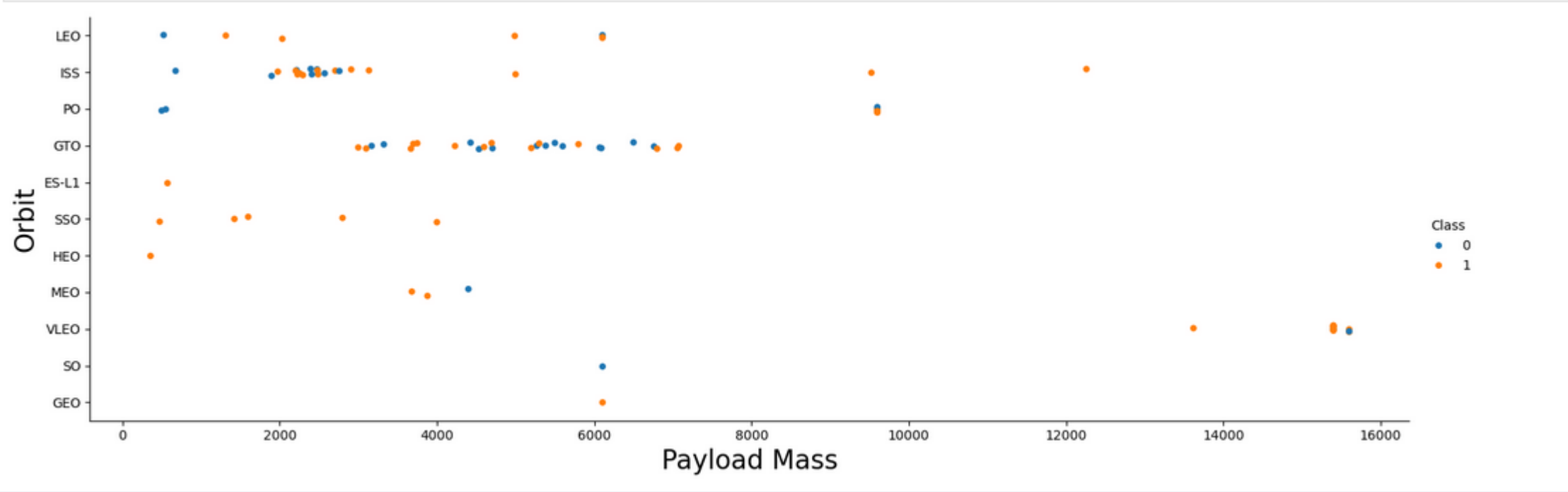
# Flight Number vs. Orbit Type

- The number of flights and flight success are positively correlated.



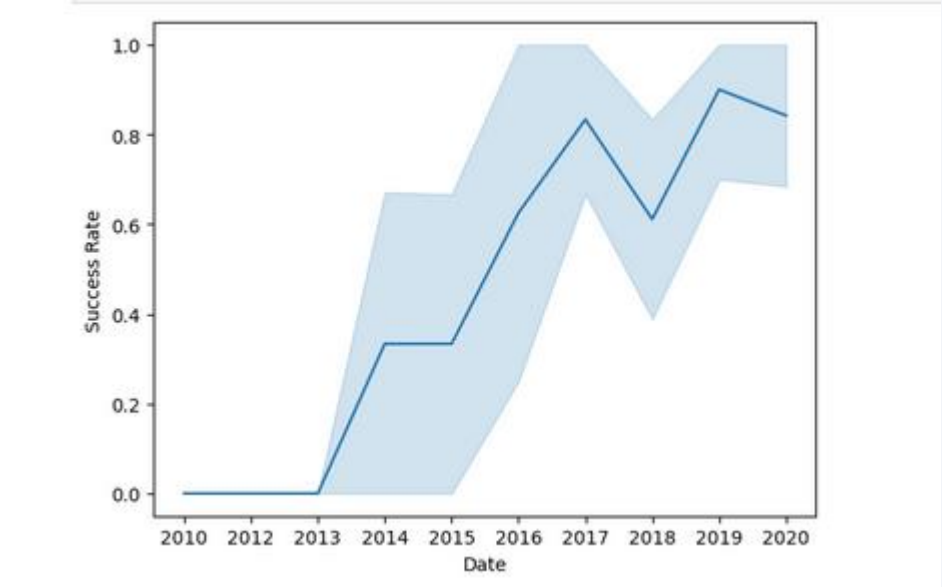
# Payload vs. Orbit Type

- Heavier payloads seem to be more beneficial for ISS.



# Launch Success Yearly Trend

- There is a large increase in success rate post 2015. There was a slight drop in the trend line in 2018, but has since recovered



# All Launch Site Names

- I used SELECT DISTINCT to find only the unique launch sites.

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

In [13]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

Out[13]:

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

# Launch Site Names Begin with 'CCA'

- LIMIT 5 would show only 5 records

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [15]:
```
%sql SELECT Launch_Site FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%'
```

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

Out[15]:

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

# Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [16]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

Out[16]:

| SUM(PAYLOAD_MASS__KG_) |
|---|
| 45596 |

Task 4

Where clause used to filter

# Average Payload Mass by F9 v1.1

- Where clause filters for only booster_version in question

Task 4

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

In [17]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

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

Out[17]:

| AVG(PAYLOAD_MASS__KG_) |
|---|
| 2928.4 |

# First Successful Ground Landing Date

- Found the first date using the MIN function

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

*Hint:Use min function*

```
[17]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

[17]:

| MIN(Date) |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- Used the where clause to find only records that fit with the criteria

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[19]:
```
%sql SELECT booster_version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

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

[19]:

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

# Total Number of Successful and Failure Mission Outcomes

- Used the COUNT function to see how many times each mission outcome appeared in the dataset

Task 7

List the total number of successful and failure mission outcomes

```
[20]: %sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

[20]:

| Mission_Outcome | COUNT(*) |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- This query finds the booster_version for all the records that contain the maximum Payload Mass

Task 8

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

```
[27]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

[27]:

| 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

- This query uses substr functions to parse through the Date column, and the where clause to filter out unwanted records

Task 9

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.

**Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.**

[28]:
```
%sql SELECT substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site  FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome = 'Failure (drone ship)'
```

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

[28]:

| 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

- Using window function to rank records in descending order based on count that is in subquery

Task 10

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, in descending order.

[32]:
```
%sql SELECT *, RANK() OVER (ORDER BY cnt DESC) AS rnk FROM (SELECT Landing_Outcome, COUNT(*) as cnt FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' AND Landing_Outcome IN ('Failure (drone ship)', 'Success (ground pad)') GROUP BY 1) sub
```

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

[32]:

| Landing_Outcome | cnt | rnk |
|---|---|---|
| Failure (drone ship) | 5 | 1 |
| Success (ground pad) | 3 | 2 |

Section 3

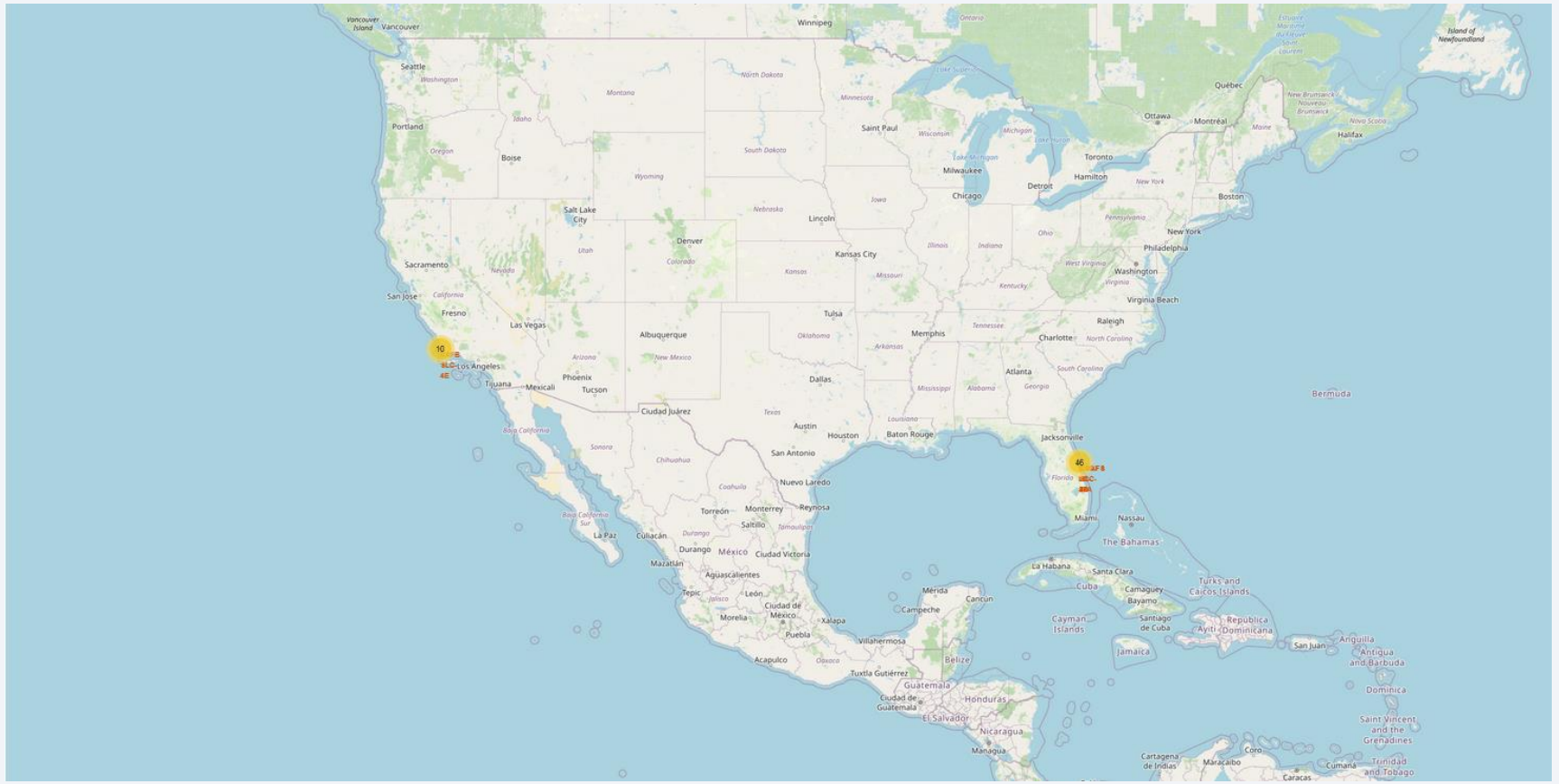
# Launch Sites Proximities Analysis

# North America Launch Sites

- All launch sites are near the coast

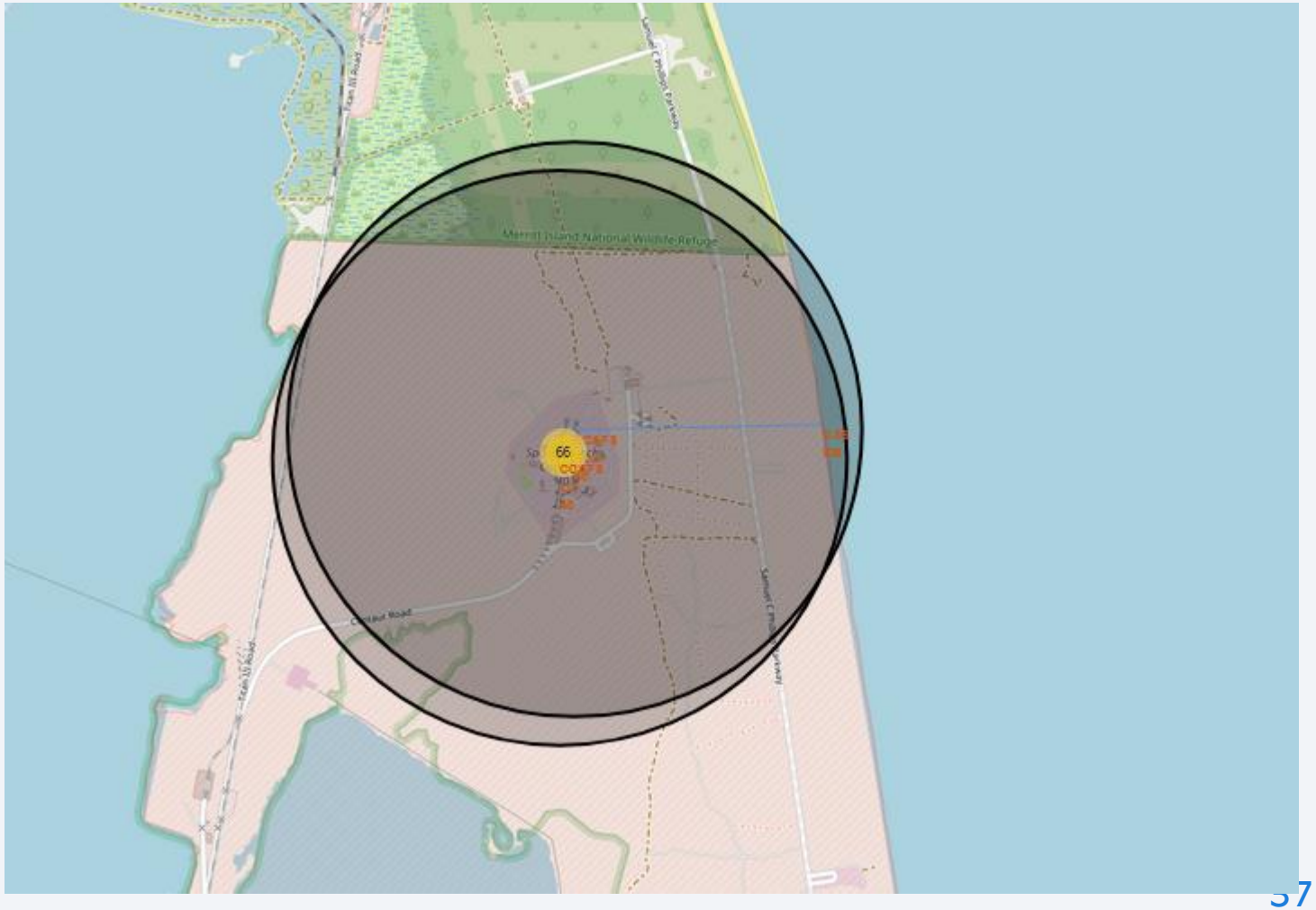
# Launch Sites with Markers

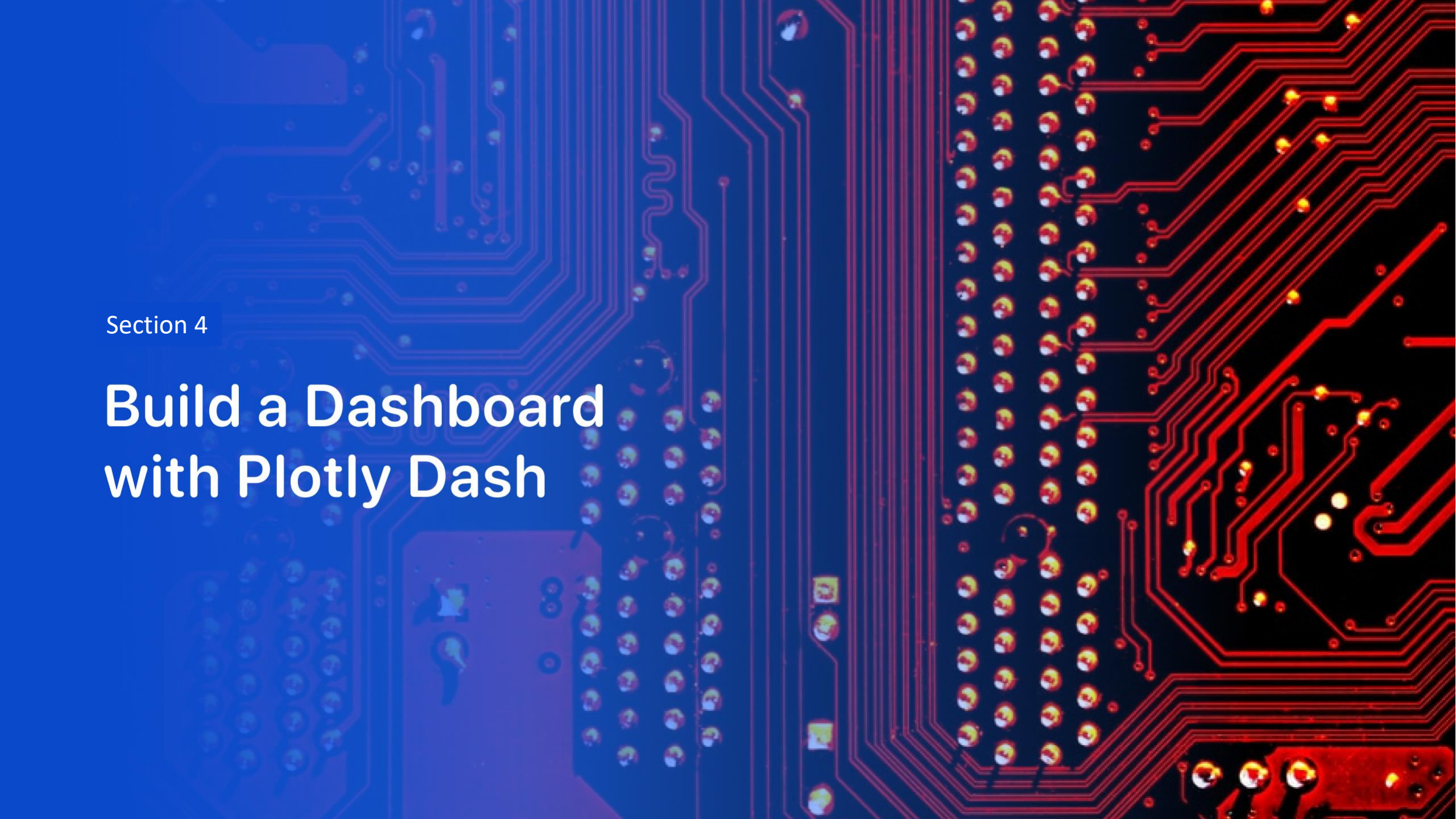
- This map includes the number of successful and failed launches at each launch site



# Launch Site Proximity to Key Transportation

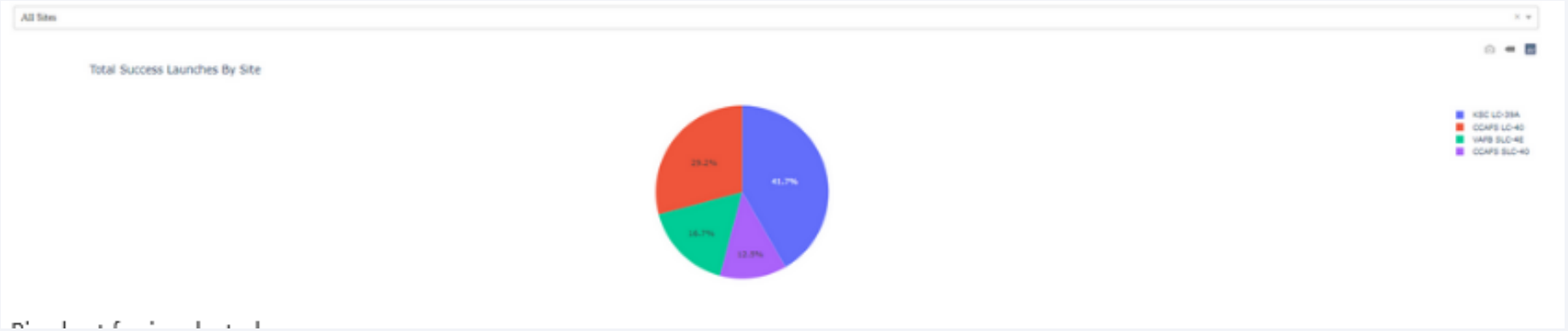
- This map shows the proximity to the coast for this launch site.




Section 4
Build a Dashboard with Plotly Dash

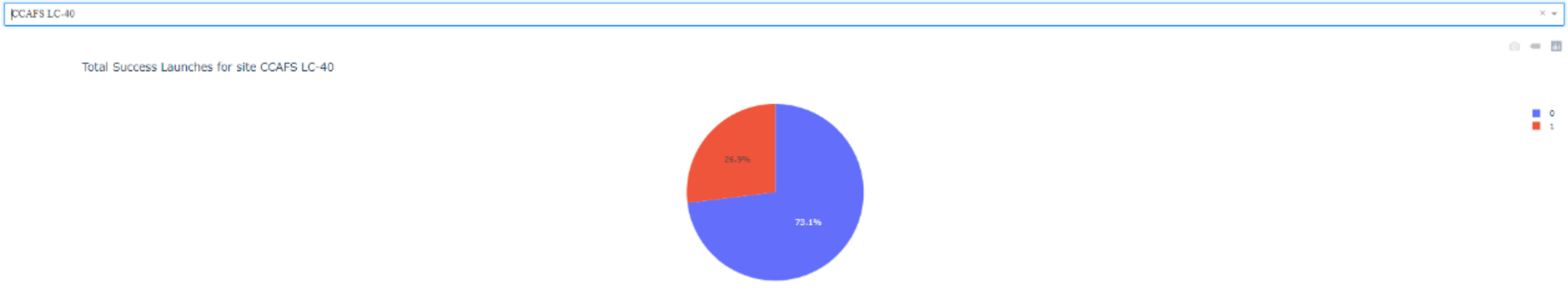
# Pie Chart for all successful launches by site

- The majority of successful launches comes form the launch site KSC LC-39A



# Pie chart for Launch Site CCAFS LC-40

- Launch site CCAFS LC-40 has the highest launch site success ratio.



# Payload Mass and Launch Success

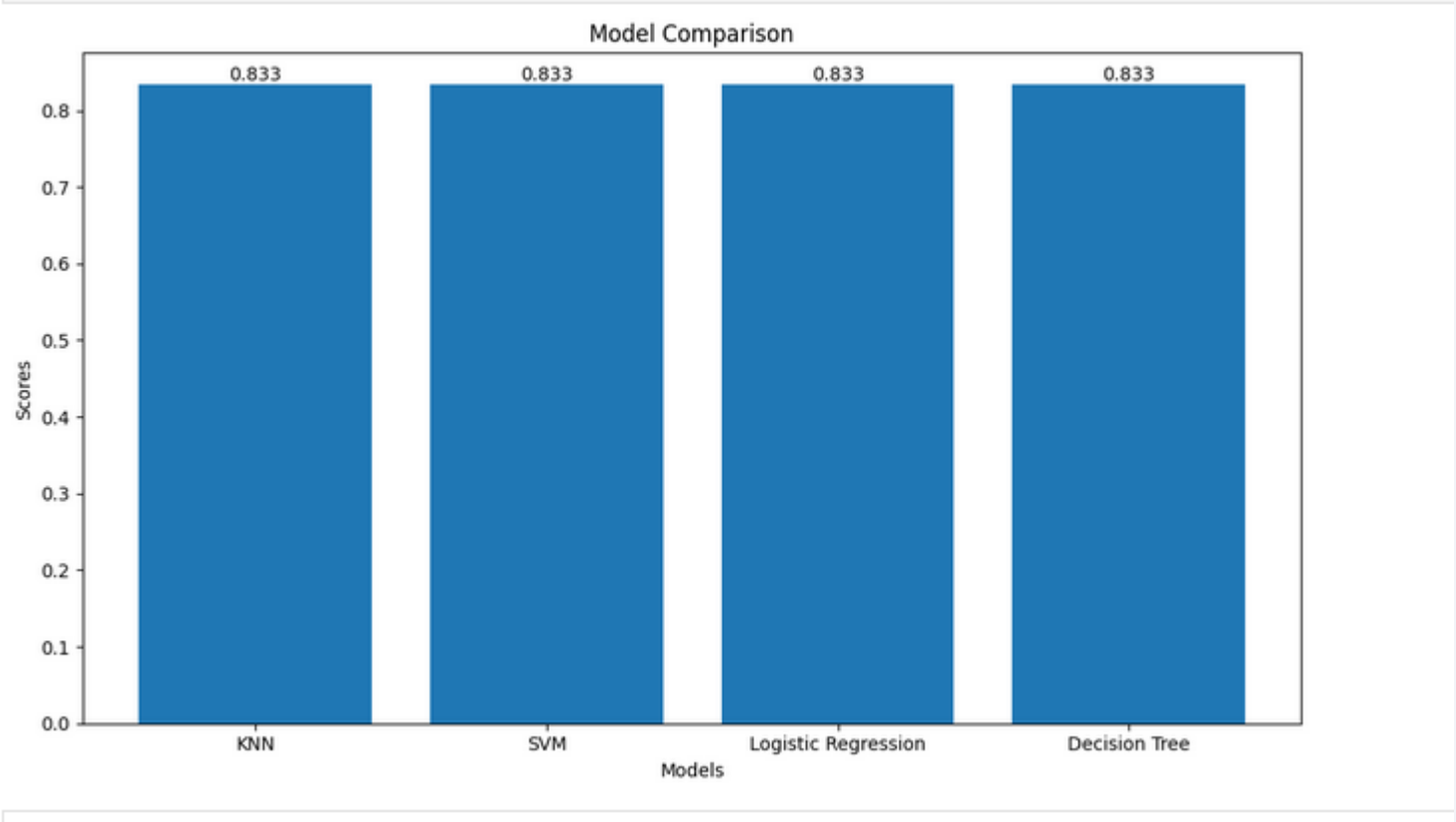
- The launches that were between 2000 and 4000 kg have the best success rate

Section 5

# Predictive Analysis (Classification)

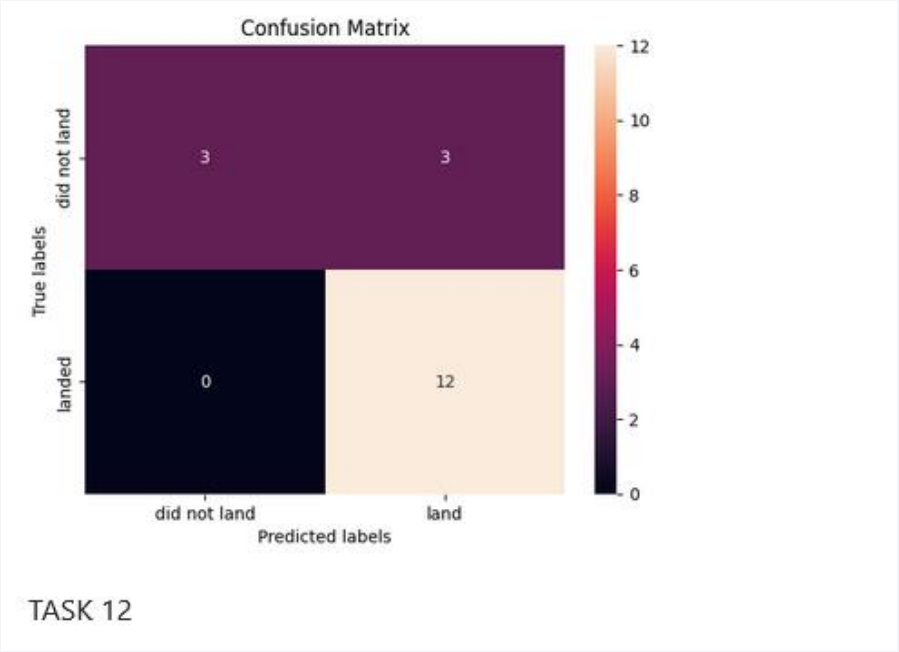
# Classification Accuracy

- All the models performed the same, at 83% accuracy.



# Confusion Matrix

- All of the models performed at the same accuracy, 83%. This confusion matrix shows that the model predicted 12/12 landings correctly, but also though that 3 launches that did not land, would land.



# Conclusions

- We are able to predict launch success rate with 83% accuracy.
- Success rate of landings is increasing in more recent years
- The launch site KSC LC-39A has the highest launch success rate out of all the launch sites that are included in this analysis.
- Launch sites are strategically placed on coasts to protect most people in case of a failed launch.

# Thank you!