

# Winning Space Race with Data Science

Vezikhaya Bomela  
31.10.2025



# Table of Content

---

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

# Executive Summary

---

- Data collection using API and webscraping
- Data Cleaning
- EDA using SQL
- EDA using Data Visualization
- Interactive visual analysis
- Machine learning for Prediction analysis

# Introduction

---

We will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; while other providers may cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore we can determine the cost of the first stage launch if it does land.

Section 1

# Methodology

# Methodology

---

## Executive Summary

- **Data collection methodology:**

Dataset extracted using SpaceX API

Collecting webscraped data from Wikipedia using BeautifulSoup

- **Perform data wrangling**

Performed EDA

Determined labels for training

- **Perform exploratory data analysis (EDA) using visualization and SQL**

Used scatterplots to visualize and analyze data

- **Perform interactive visual analytics using Folium and Plotly Dash**

Built interactive map for geospatial analysis and dashboard for viewing graphs

- **Perform predictive analysis using classification models**

Trained logistic regression model, SVM, decision tree classifier, and KNN to determine the best fit model

# Data Collection

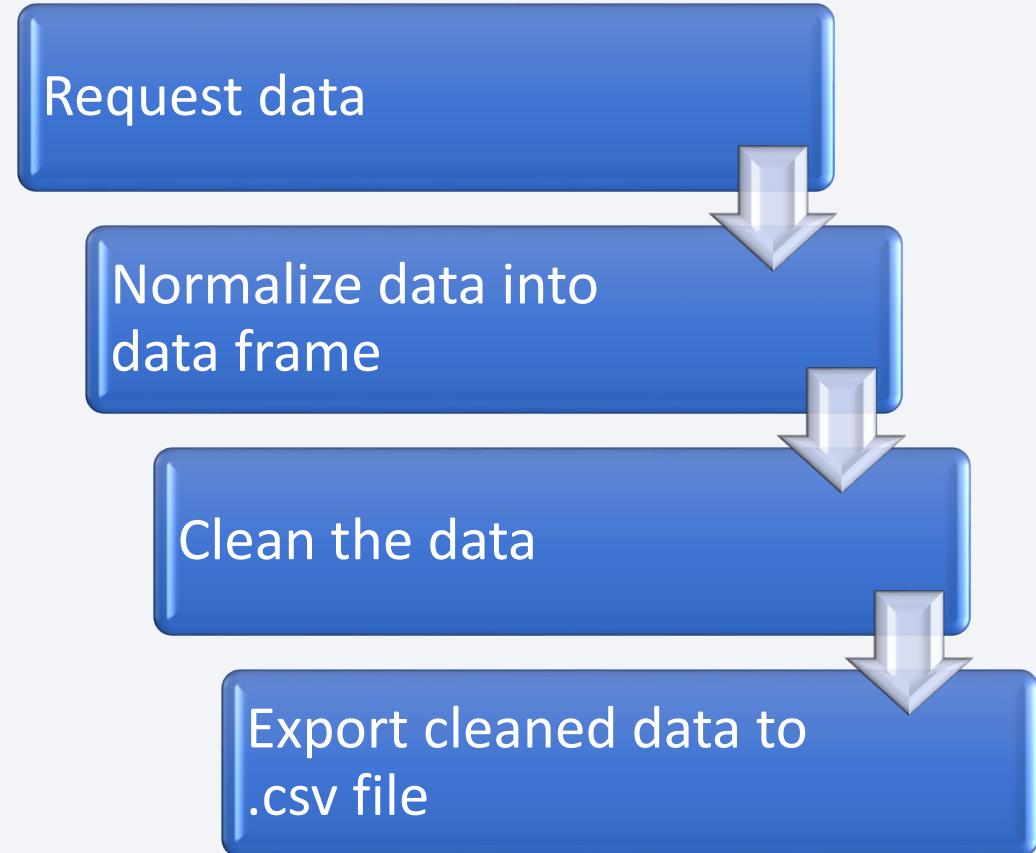
---

- SpaceX API - <https://api.spacexdata.com/v4/>  
API used to collect required data from SpaceX  
Data is cleaned for further analysis
- Webscraping Falcon 9 data:  
[https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)  
Extract HTML table and converting it to a dataframe using BeautifulSoup

# Data Collection – SpaceX API

---

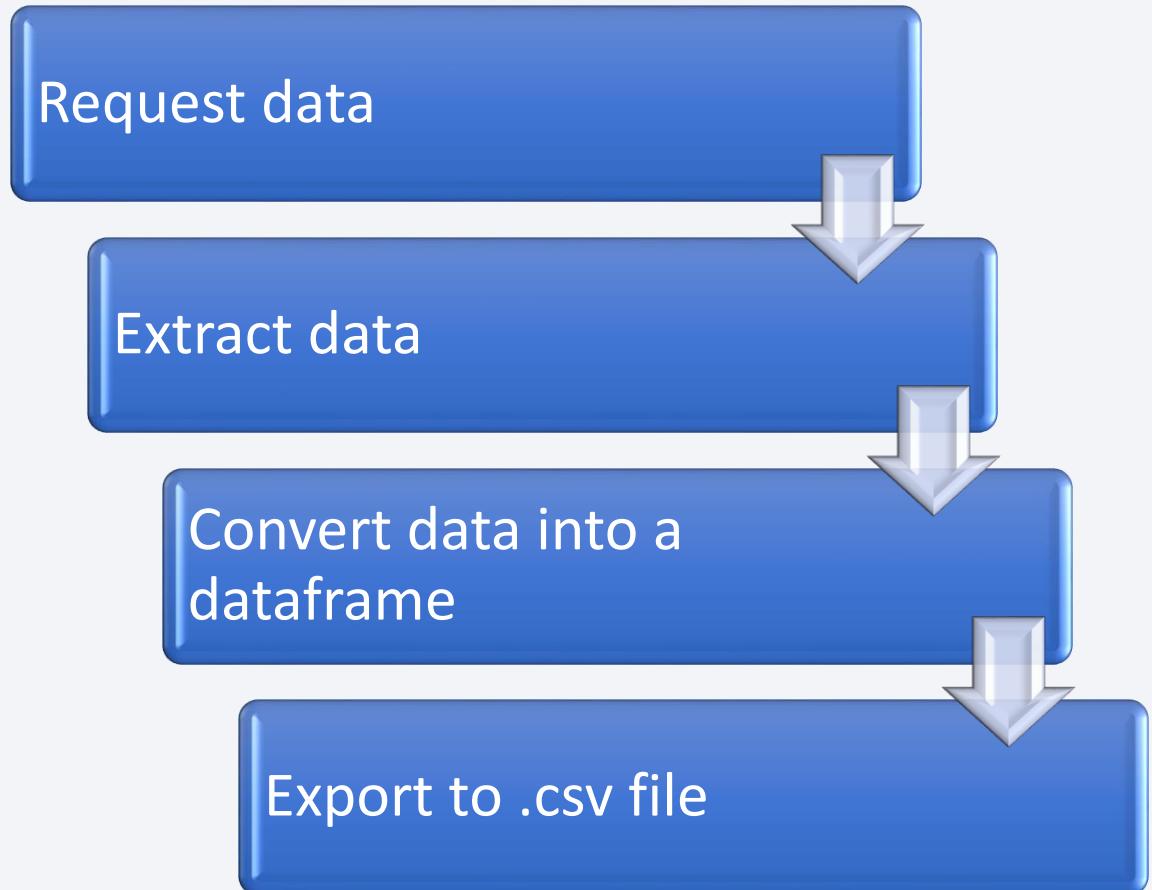
- Request data from SpaceX API
- Normalize .json data into a dataframe
- Dataset is cleaned and filtered required information for further processing
- Cleaned data is exported to a csv file
- GitHub URL SpaceX API:  
<https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/DataCollectionAPI.ipynb>



# Data Collection - Scraping

---

- Request Falcon 9 data from Wikipedia URL
- Extracted HTML table data using BeautifulSoup
- Data is parsed to into a dataframe
- Dataframe is exported to a csv file
- GitHub URL Web Scraping:  
<https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/DataCollectionWebScraping.ipynb>



# Data Wrangling

---

- Identify and filter missing values
- Calculate number of launches on each site
- Calculate number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Data is exported to a csv file
- GitHub URL Data Wrangling: <https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/DataWrangling.ipynb>

# EDA with Data Visualization

---

- The scatter plot was used to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, Flight Number and Orbit type, and Payload and Orbit type as it is best for observing relationships between two variables (categorical data)
- The bar plot was used to visualize relationship of the success rate of each orbit type as it is best for depicting group of variables (categorical data)
- The line plot was used to visualize the launch success yearly trend as it best for showing time series data
- GitHub URL EDA with data visualization: <https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/EDADataVisualision.ipynb>

# EDA with SQL

---

- **SQL queries performed to get insights:**
  - Unique launch sites.
  - Launch sites beginning with CCA.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - Date when the first successful landing outcome in ground pad was achieved.
  - The total no of successful and failed outcomes.
  - Names of the booster versions which have carried the maximum payload mass.
  - Rank of the count of landing outcomes between a specified date.
- GitHub URL EDA with SQL: [https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/EDA%20\\_SQL%20\(3\).ipynb](https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/EDA%20_SQL%20(3).ipynb)

# Build an Interactive Map with Folium

---

- Map objects used: Circles, Markers, Marker clusters, Mouse position and Line;  
**Circles** - For adding highlighted circle area with a label on a site.  
**Markers** - Mark the site  
**Marker Clusters** - To simplify the map as it contained any markers having the same coordinate  
**Mouse Position** - Coordinates for the position the mouse points on the map  
**Line** - Draws a line from a site to the nearest coast, rail and highway.
- GitHub URL Folium: [https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/AnalysisFolium%20\(2\).ipynb](https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/AnalysisFolium%20(2).ipynb)

# Build a Dashboard with Plotly Dash

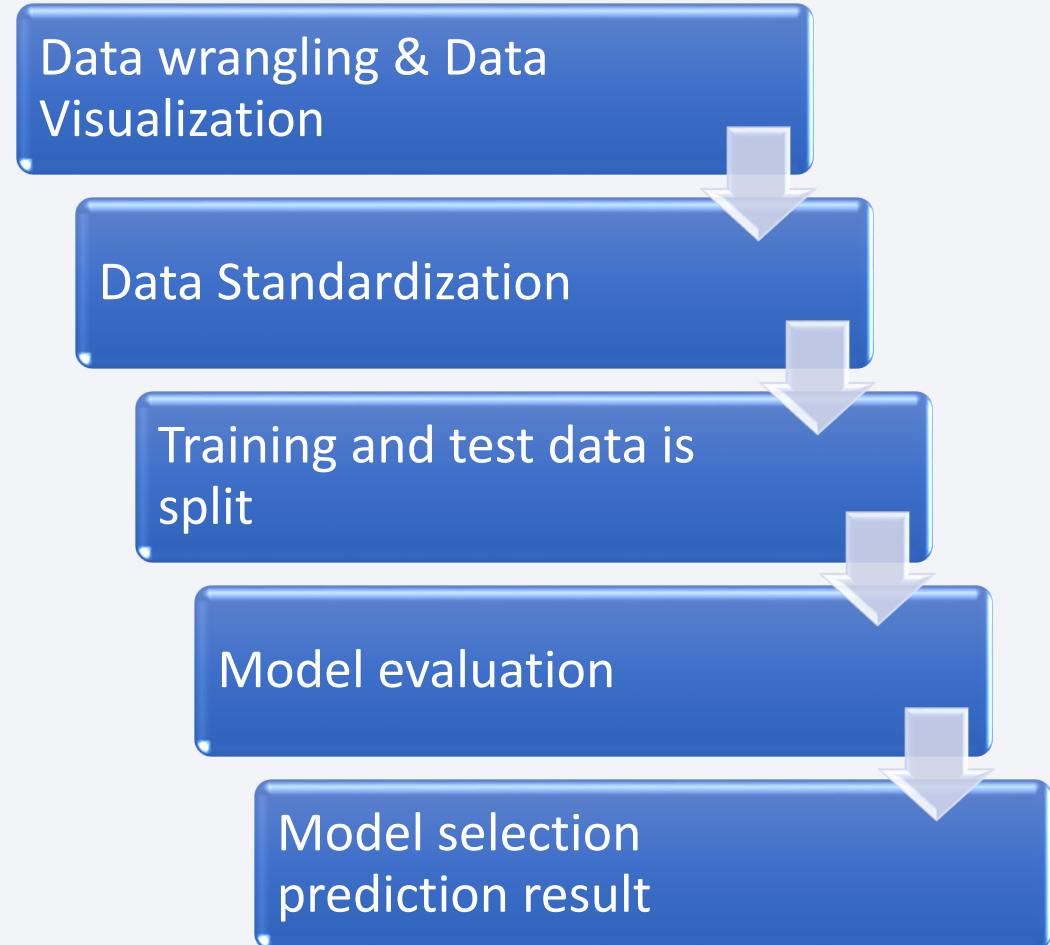
---

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

# Predictive Analysis (Classification)

---

- Data wrangling and Data Visualization datasets were loaded into respective dataframes before standardizing data for prediction.
- After data is standardized it is split into training data and test data
- Classification models logistic regression model, SVM, decision tree, and KNN were used to determine the best fit model
- GitHub URL predictive analysis:  
<https://github.com/Vezikhaya/Data-Science-Capstone-Project/blob/main/PredictionAnalysis.ipynb>



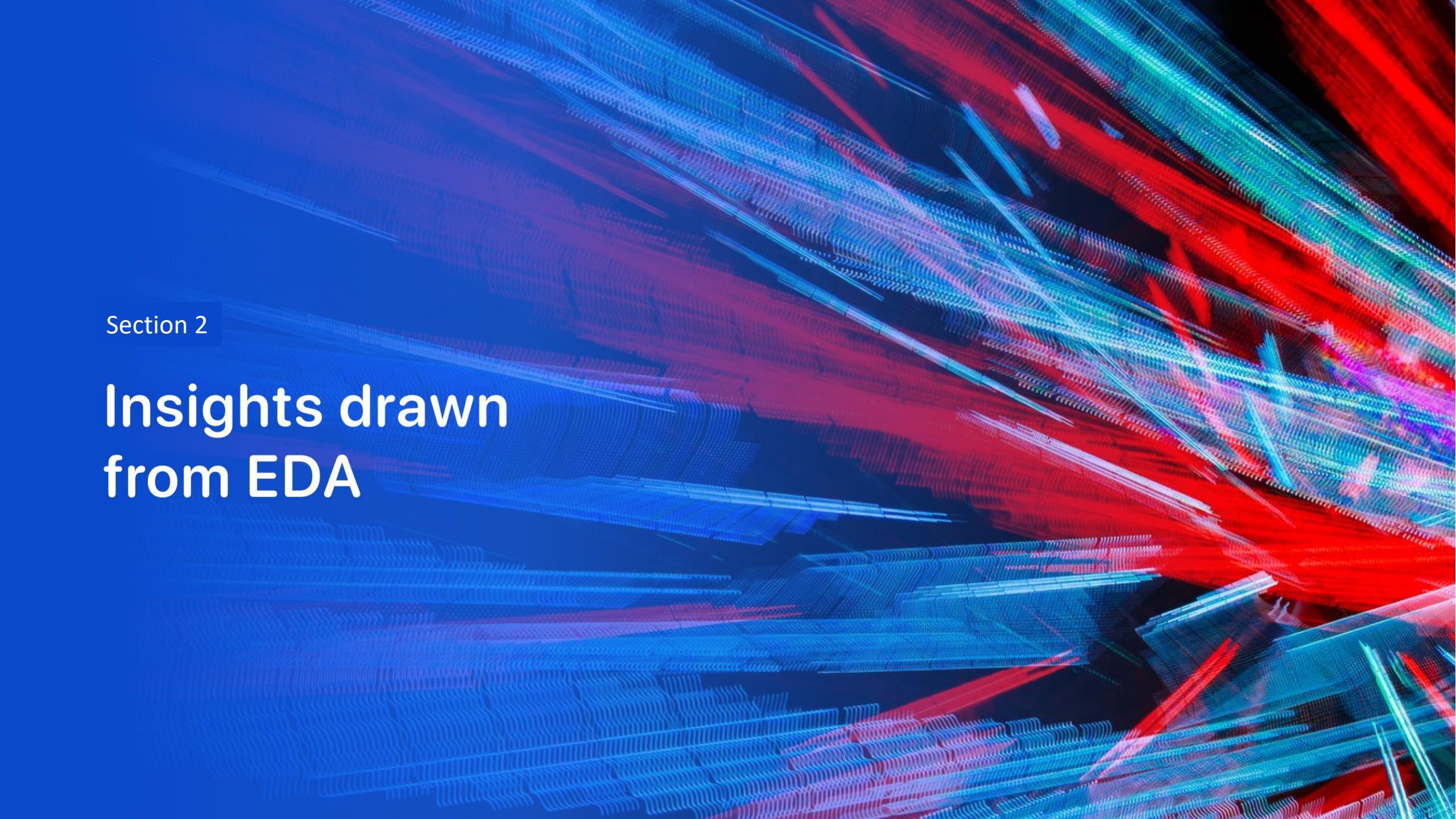
# Results

---

- Exploratory data analysis results
  - Payloads over 8000kg have the highest success rate
  - SpaceX has three launch sites KSC LC-39A , VAFB SLC 4E and CCAFS LC-40
  - The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing
- Interactive analytics demo in screenshots



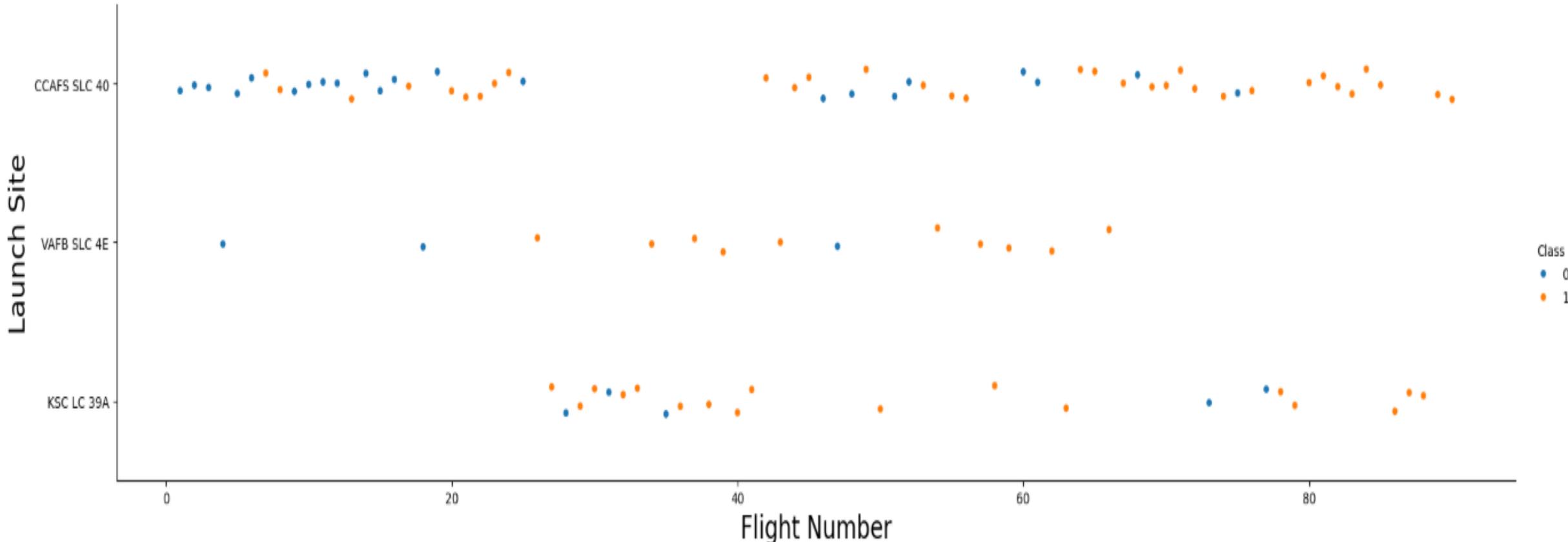
- Predictive analysis results
  - According to the model evaluation results the decision tree is best the classification model for this problem.

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a three-dimensional space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

## Insights drawn from EDA

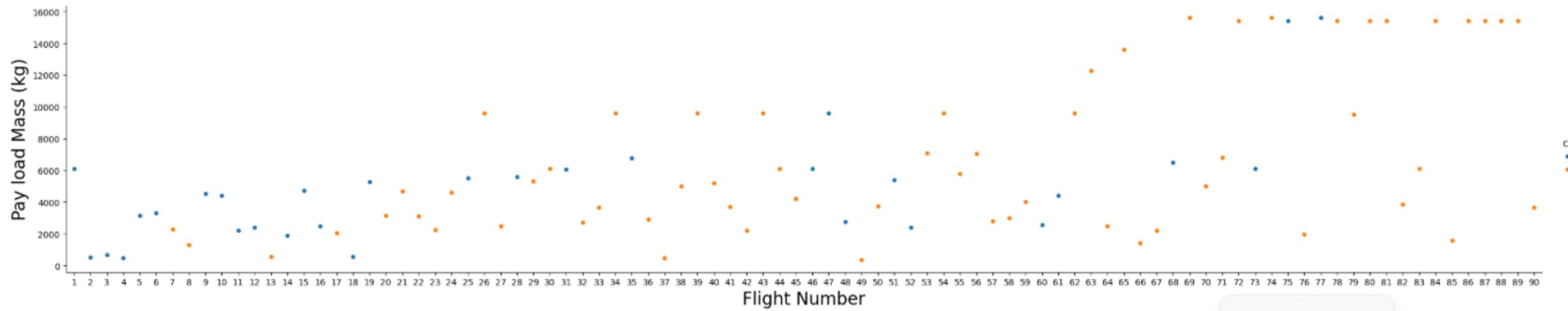
# Flight Number vs. Launch Site



CCAFS LC-40 has low success rate compared to the other two as it failed a lot during initial flights. VAFB SLC 4E and KSC LC-39A have almost same success rate, and they have a relatively higher flight number so failure rate is low

# Payload vs. Launch Site

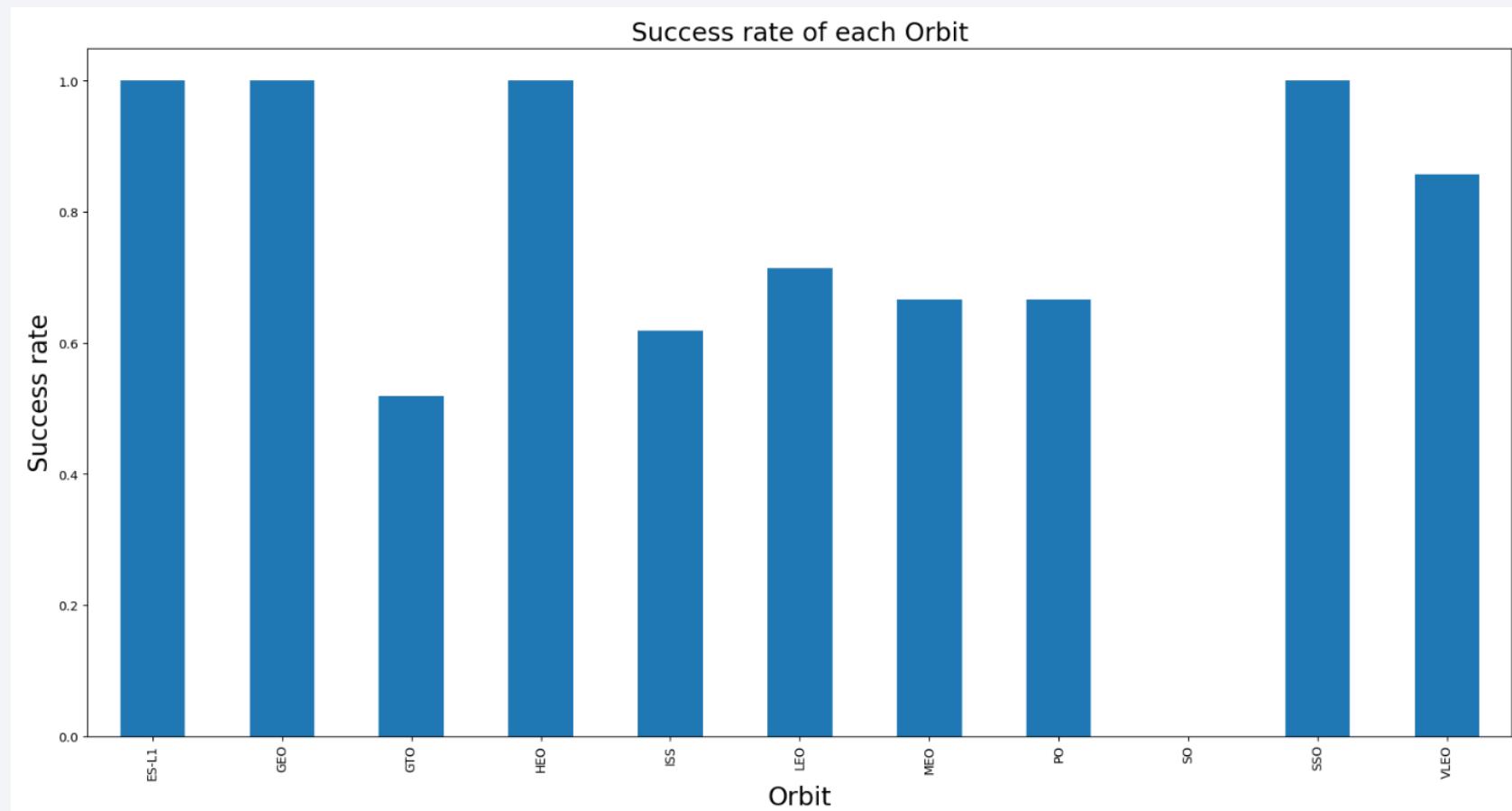
---



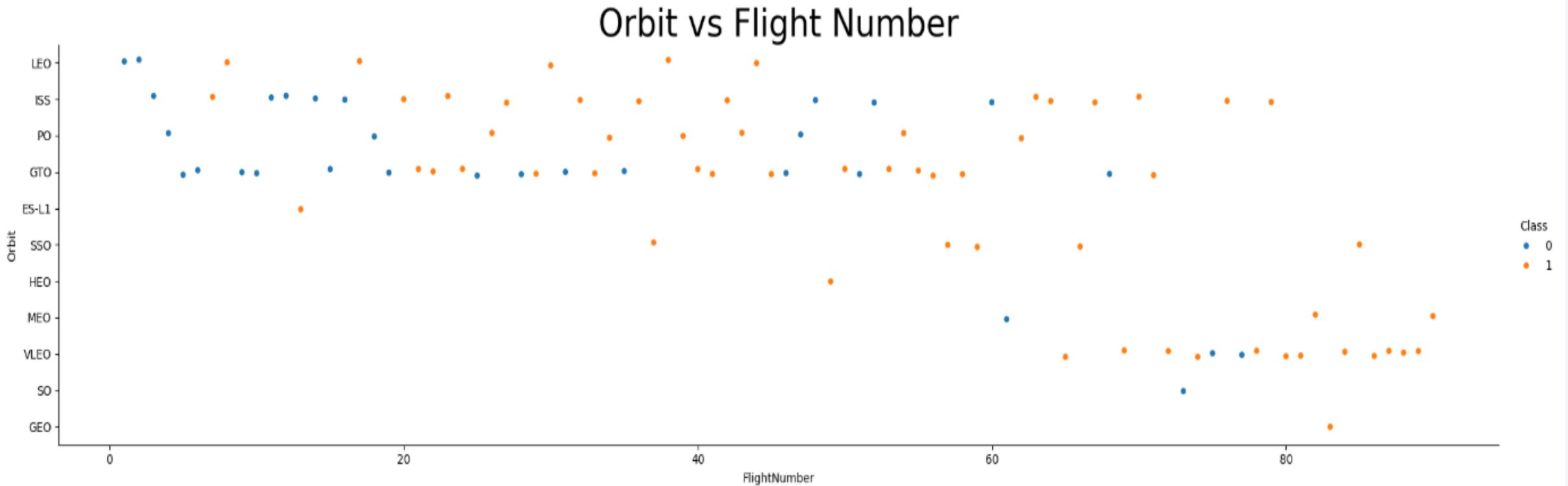
- Show a scatter plot  
of Payload vs. Launch Site

# Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations

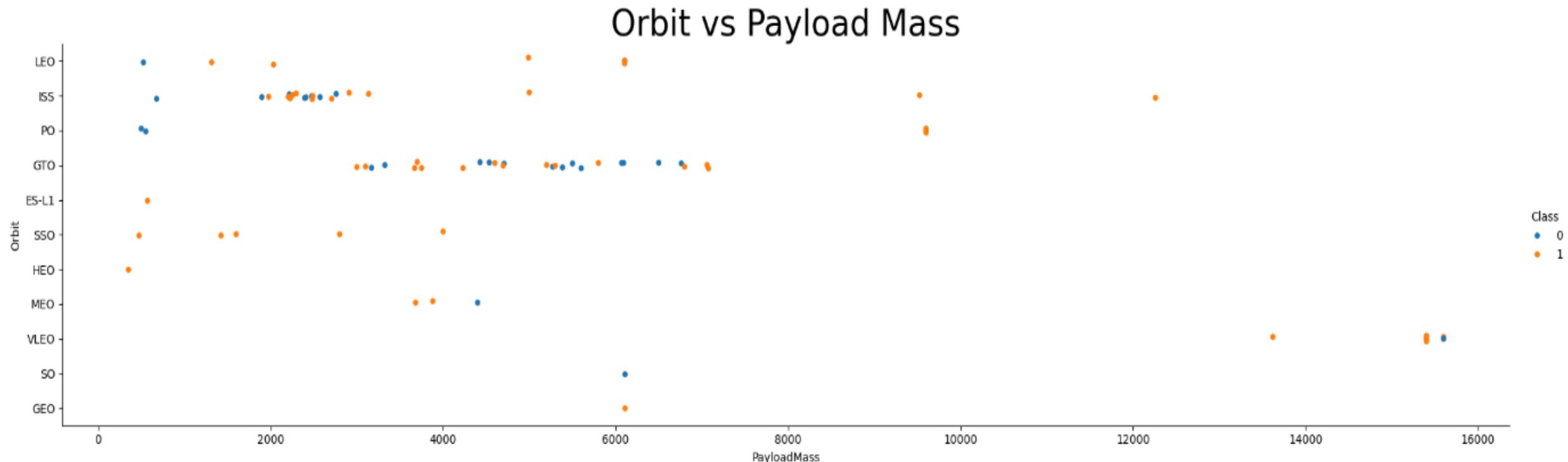


# Flight Number vs. Orbit Type



You should 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.

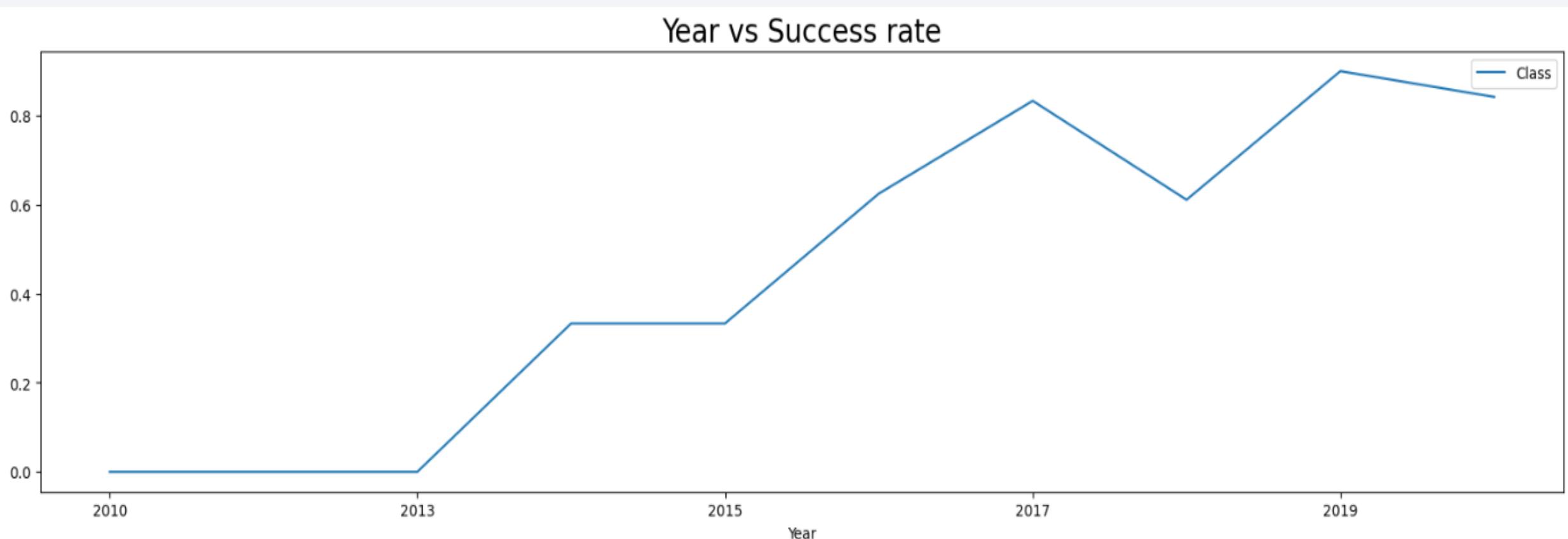
# Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, 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



You 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

---

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

.4] : %sql SELECT DISTINCT Launch\_Site FROM SPACEXTBL;

\* sqlite:///my\_data1.db

Done.

.4] : Launch\_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

# Launch Site Names Begin with 'CCA'

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

```
: %sql SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

```
Done.
```

```
: Launch_Site
```

```
CCAFS LC-40
```

# Total Payload Mass

---

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

```
: %sql SELECT SUM(PAYLOAD_MASS_KG_) AS total_payload FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)';
```

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

Done.

```
: total_payload
```

---

45596

# Average Payload Mass by F9 v1.1

---

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

```
: %sql SELECT avg(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';
```

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

```
Done.
```

```
: Avg_Payload
```

---

```
2928.4
```

# First Successful Ground Landing Date

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

*Hint:Use min function*

```
%sql SELECT min(date) AS Early_Date FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

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

Done.

**Early\_Date**

---

2015-12-22

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

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

```
%sql SELECT DISTINCT Customer, Landing_Outcome, PAYLOAD_MASS_KG_ FROM SPACEXTBL WHERE Landing_Outcome ='Success (drone ship)' AND
```

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

```
Done.
```

Customer	Landing_Outcome	PAYLOAD_MASS_KG_
SKY Perfect JSAT Group	Success (drone ship)	4696
SKY Perfect JSAT Group	Success (drone ship)	4600
SES	Success (drone ship)	5300
SES EchoStar	Success (drone ship)	5200

# Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, Count(*) AS Numbers FROM SPACEXTBL GROUP BY Mission_Outcome;  
* sqlite:///my_data1.db  
Done.
```

Mission_Outcome	Numbers
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

# Boosters Carried Maximum Payload

List all the booster\_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
| : %sql SELECT Booster_Version, Max_Payload FROM (SELECT Booster_Version, MAX(PAYLOAD_MASS__KG_) AS Max_Payload FROM S
| : * sqlite:///my_data1.db
| : Done.
```

Booster_Version	Max_Payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.1	3600

# 2015 Launch Records

---

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: SQLite 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.**

```
, 6, 2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date, 0,
```

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

```
Done.
```

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

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

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.

```
: numbers FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%' AND Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Numbers DE
```

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

```
Done.
```

Landing_Outcome	Numbers
-----------------	---------

Success (drone ship)	5
----------------------	---

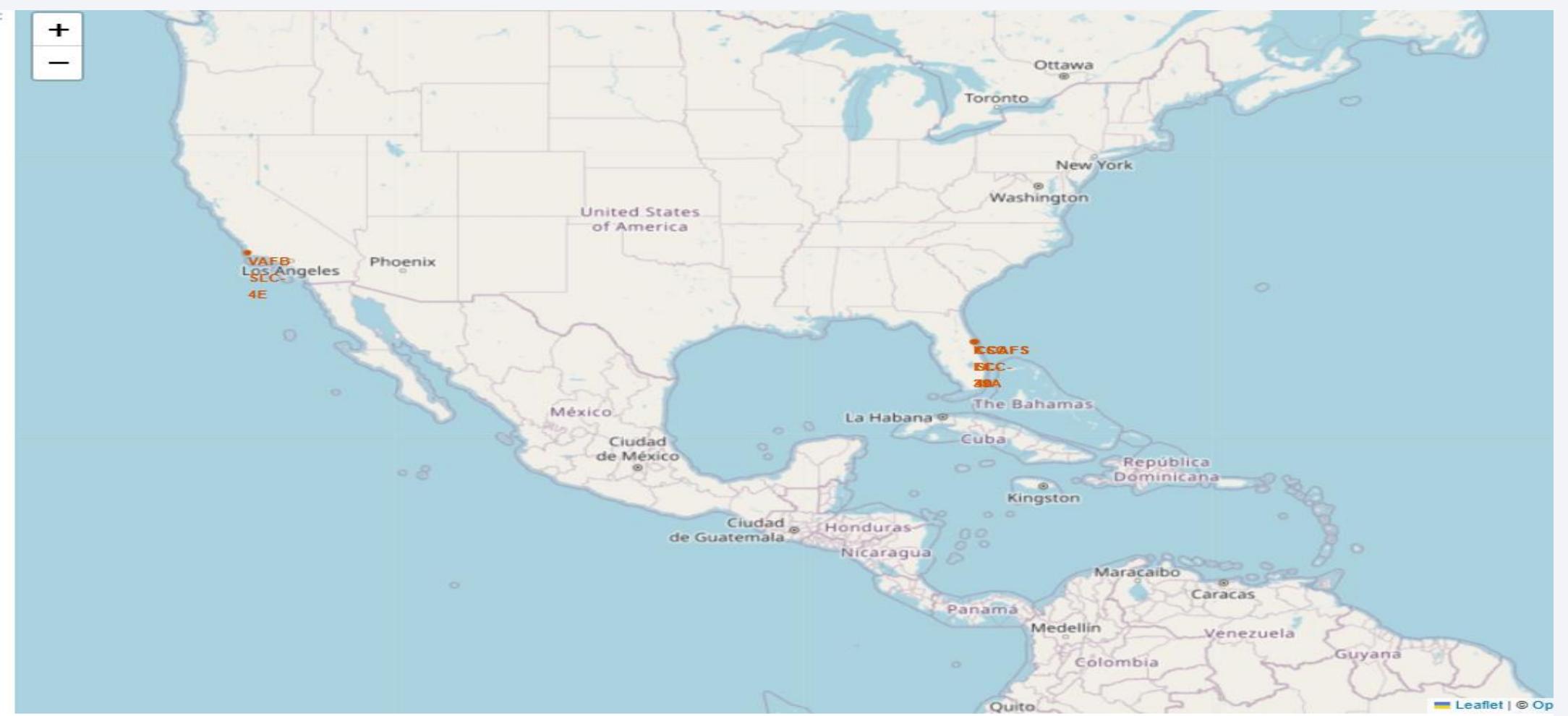
Success (ground pad)	3
----------------------	---

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth's horizon against a dark blue sky. Numerous glowing yellow and white points represent city lights, concentrated in coastal and urban areas. In the upper right quadrant, there are bright green and yellow bands of light, likely the Aurora Borealis or Australis. The overall atmosphere is dark and mysterious.

Section 3

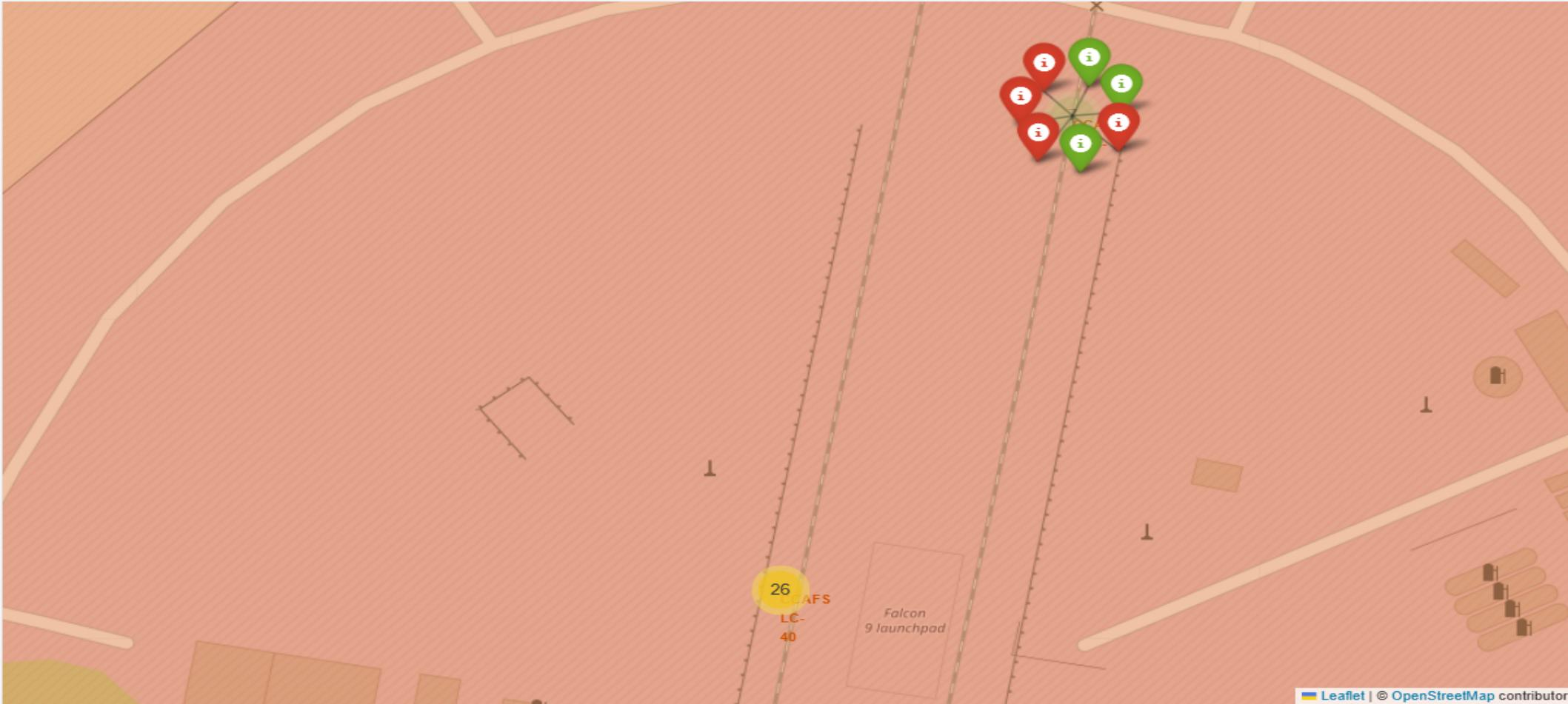
# Launch Sites Proximities Analysis

# <Folium Map Screenshot 1>



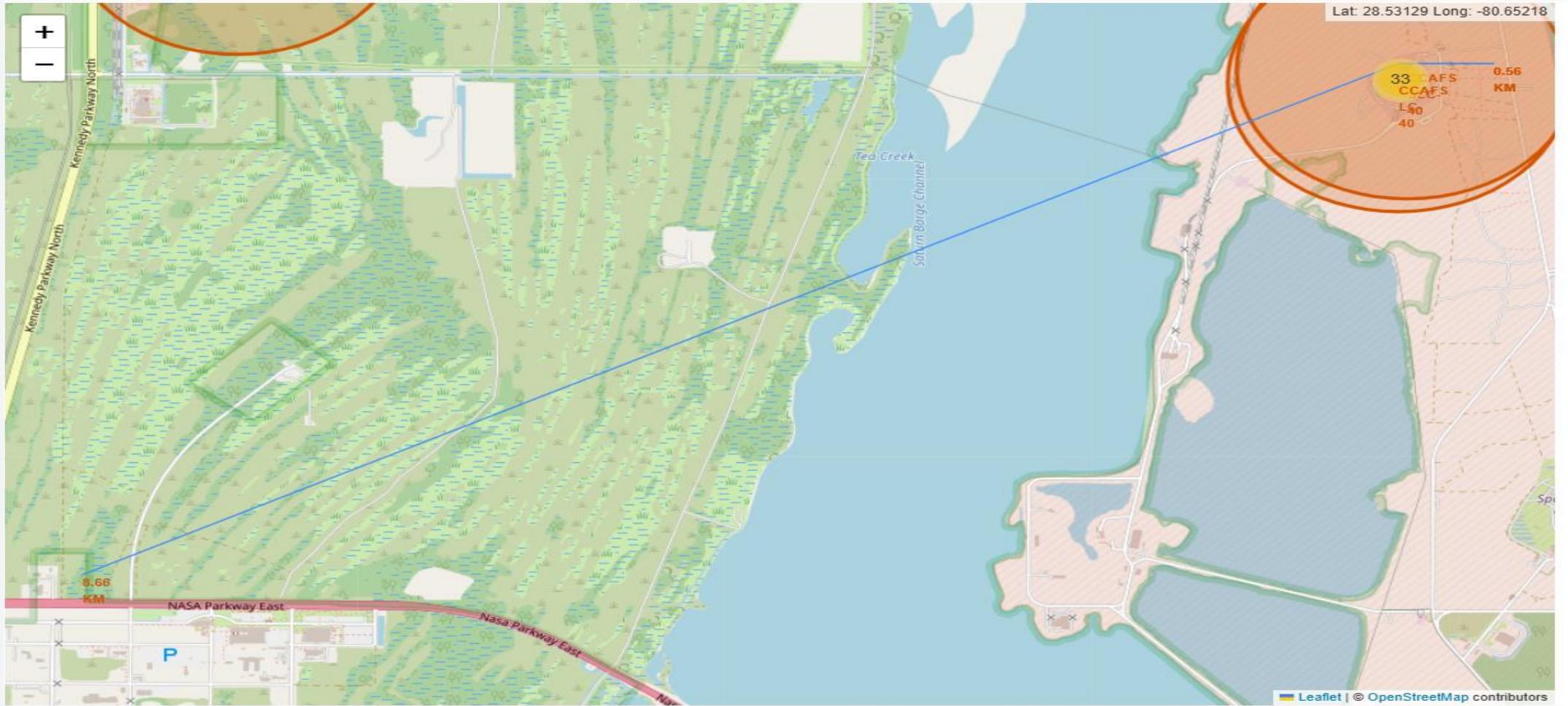
Launch sites in proximity to the Equator line and are in very close proximity to the coast

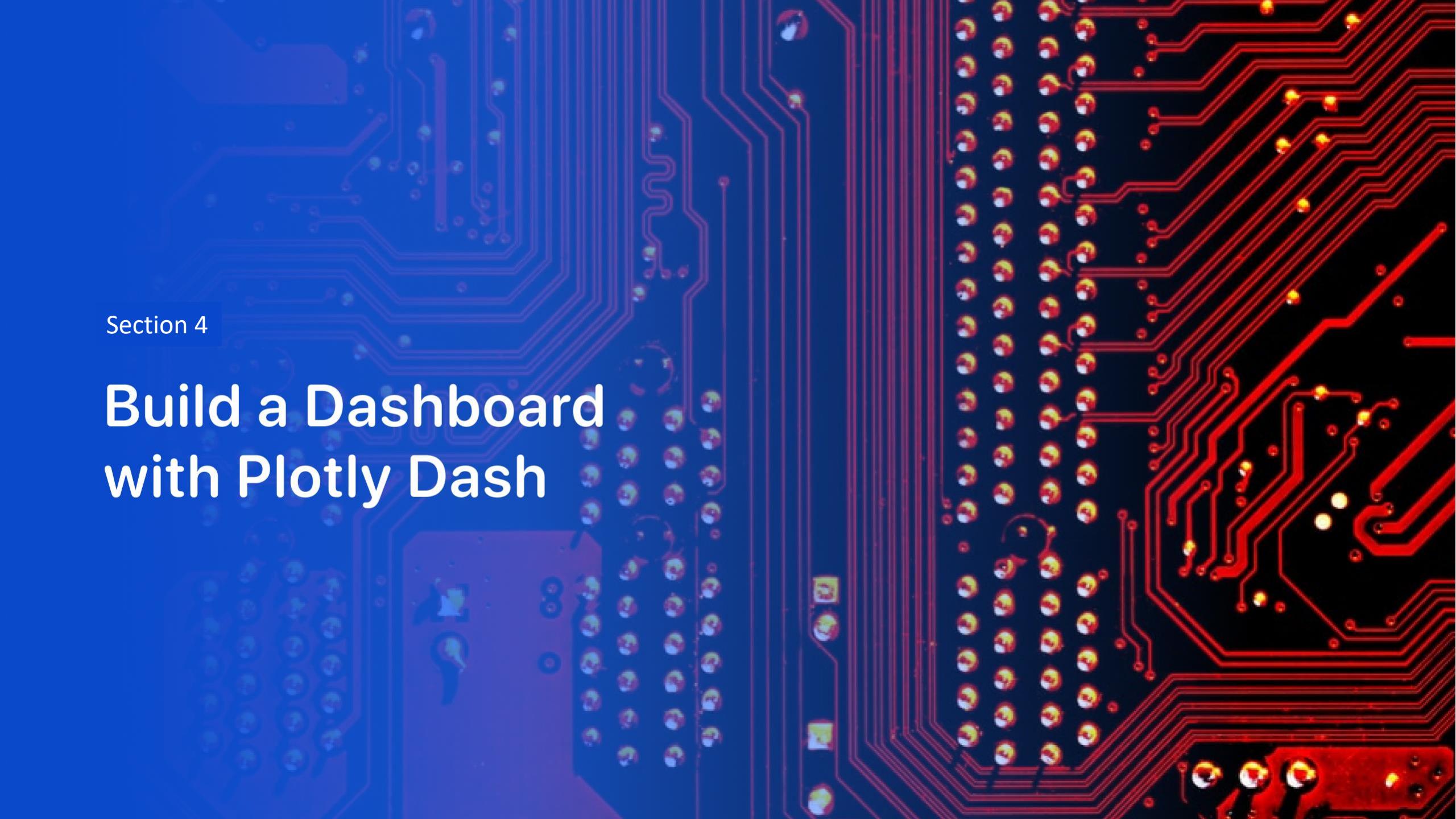
# <Folium Map Screenshot 2>



From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

# <Folium Map Screenshot 3>





Section 4

# Build a Dashboard with Plotly Dash

# <Dashboard Screenshot 1>

---

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

## <Dashboard Screenshot 2>

---

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

## <Dashboard Screenshot 3>

---

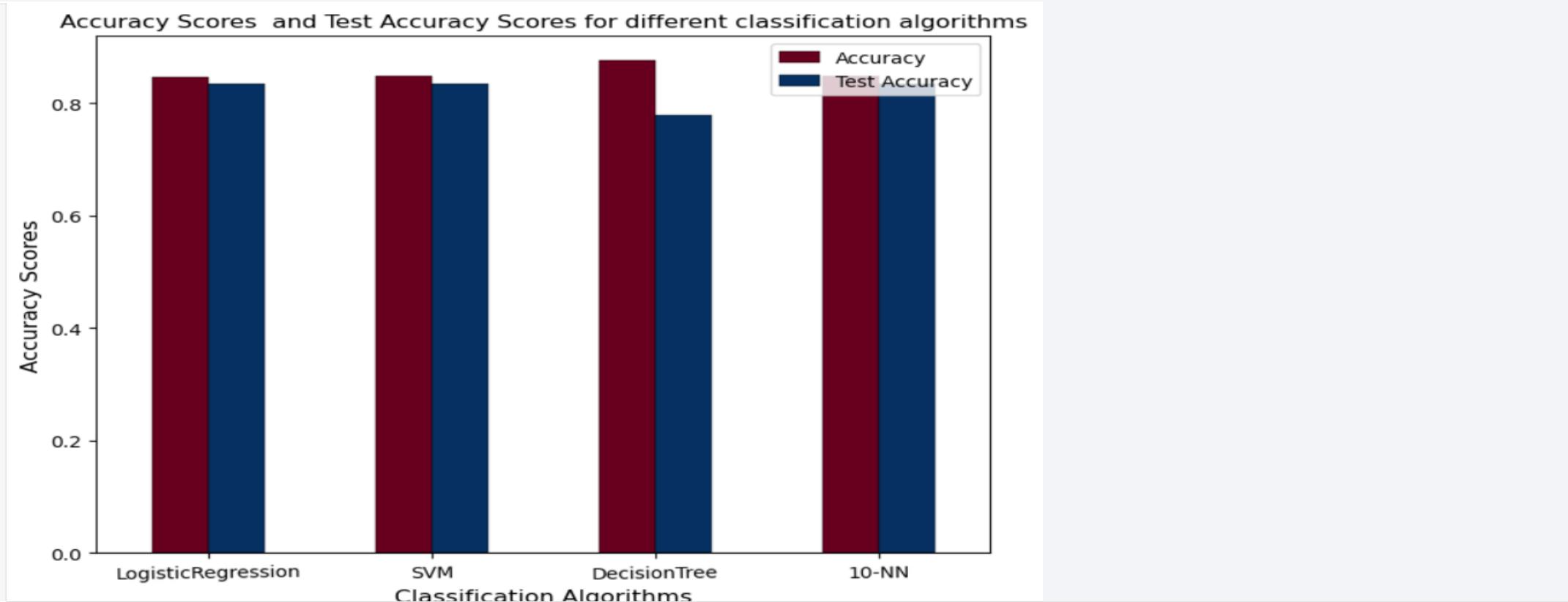
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

The background of the slide features a dynamic, abstract design. It consists of several curved, overlapping bands of color. A prominent band on the left is a bright blue, while another on the right is a warm yellow. These colors transition into lighter shades of blue and yellow towards the edges. The overall effect is one of motion and depth, resembling a tunnel or a stylized landscape.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy



The best classification model is SVM

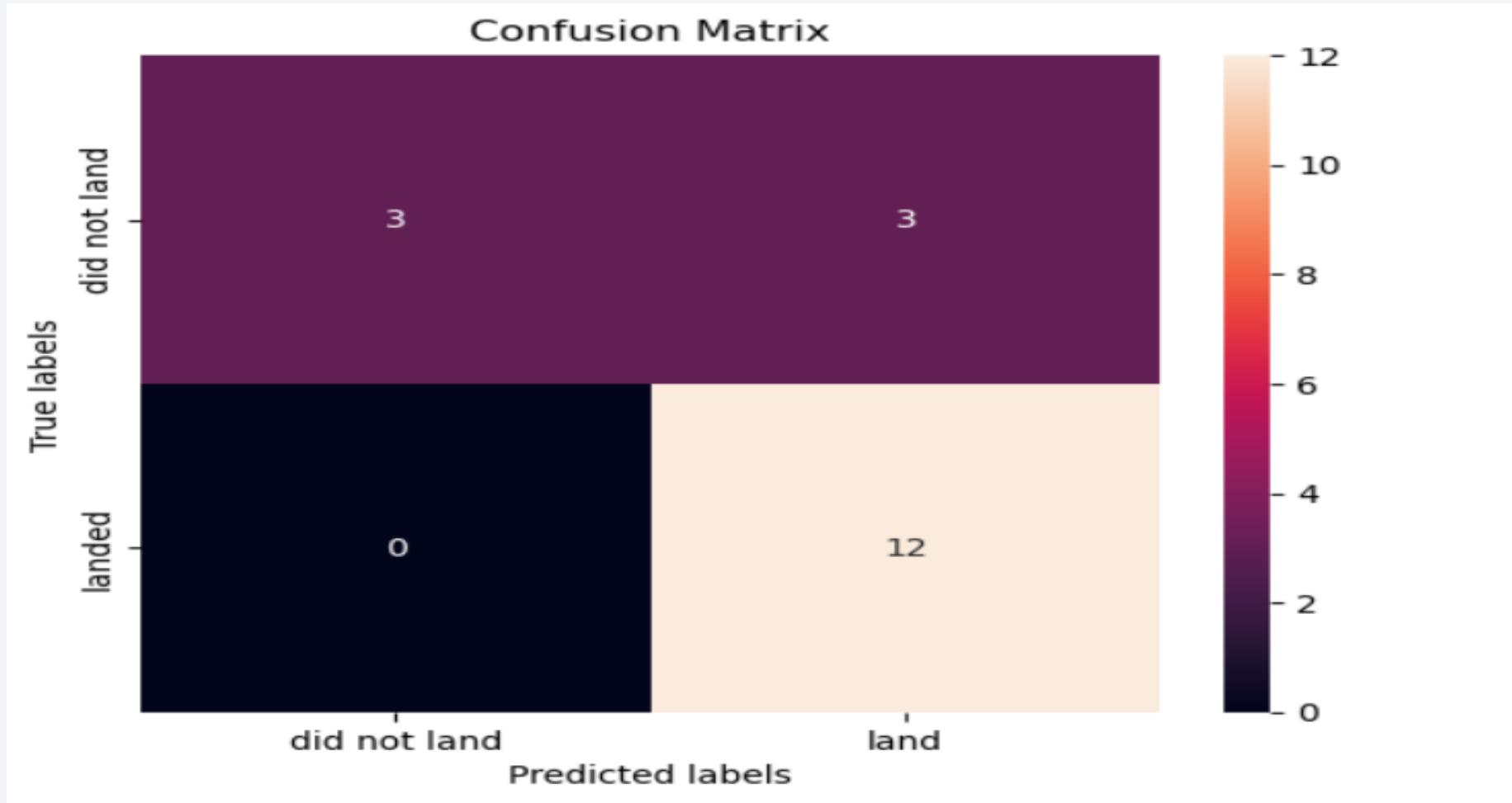
10-NN, with a mean accuracy of 0.8407738095238095

The best parameters for the Decision Tree model are:

```
{'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'} 43
```

# Confusion Matrix

---



# Conclusions

---

- Launches with payloads over 8000kg have a high success rate and a low failure rate in new launches
- SSO, GEO, ES-L1, and HEO have the highest success rate while SO has a very low success rate
- Launch sites in proximity to the Equator line and are in very close proximity to the coast
- Decision Tree Classifier is best model to for the problem with an accuracy score of 84%

# Appendix

---

- <https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/TZRZL/hands-on-lab-complete-the-data-collection-api-lab>
- <https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/sVYES/hands-on-lab-complete-the-data-collection-with-web-scraping-lab>
- <https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/fSWin/hands-on-lab-data-wrangling>
- <https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/XTU2I/hands-on-lab-complete-the-eda-with-sql>
- <https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/BFhq5/hands-on-lab-interactive-visual-analytics-with-folium-lab>
- <https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/EUhln/hands-on-lab-complete-the-machine-learning-prediction-lab>

Thank you!

