IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Domingo Jesús Sánchez Blanco
16/01/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection and wrangling
  - Exploratory Data Analysis with Data Visualization and SQL
  - Building an interactive map and Dashboard
  - Predictive analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

The commercial space industry is rapidly advancing, with companies like Virgin Galactic offering suborbital flights, Rocket Lab focusing on small satellites, and Blue Origin developing reusable rockets. SpaceX, the industry leader, has achieved significant milestones such as sending spacecraft to the International Space Station, launching Starlink for satellite internet, and conducting manned missions. SpaceX's ability to reduce launch costs—$62 million for a Falcon 9 launch—stems from reusing the rocket's first stage, unlike competitors who charge much more. In this project, you will act as a data scientist for Space Y, a new company founded by billionaire Allon Musk, aiming to compete with SpaceX. Your role involves analyzing SpaceX's operations to determine launch costs and predicting whether SpaceX will reuse the Falcon 9's first stage using machine learning models and publicly available data.

- Problems you want to find answers
  - How do variables like payload mass, launch site, number of flight, and orbits influence the success of the first stage landing?
  - Has the rate of successful landings improved over the years?
  - What is the most effective algorithm for classification in this scenario?

Section 1

Methodology

Methodology

Executive Summary

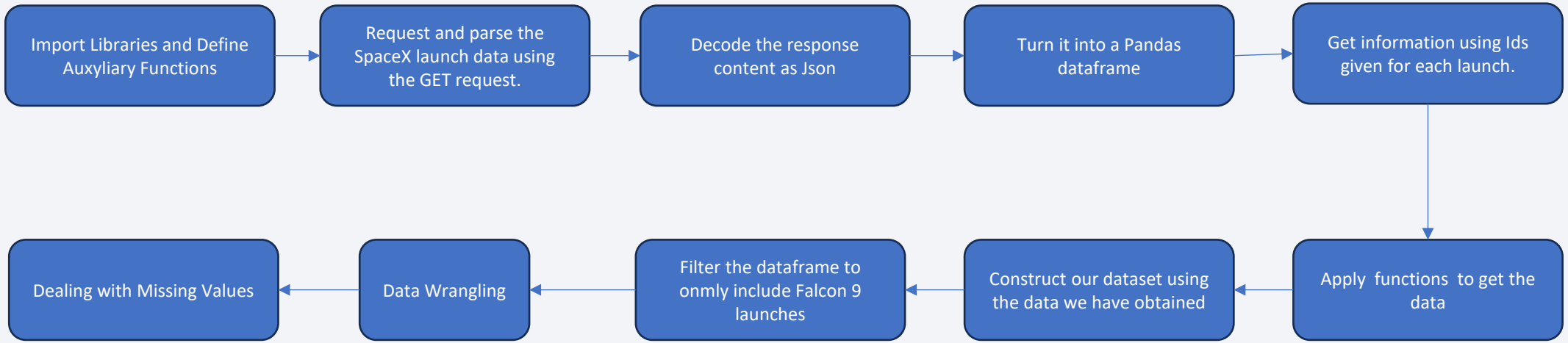
- Data collection methodology:
  - Using SpaceX API Rest and Web Scraping
- Perform data wrangling
  - The process involve filtering the data and handling missing values to prepare the data for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Developing, optimizing, and assessing classification models to achieve the best result

Data Collection

- The data collection process combined API requests from SpaceX API Rest and the use of web scraping. Both methods were necessary to gather comprehensive information about the launches.

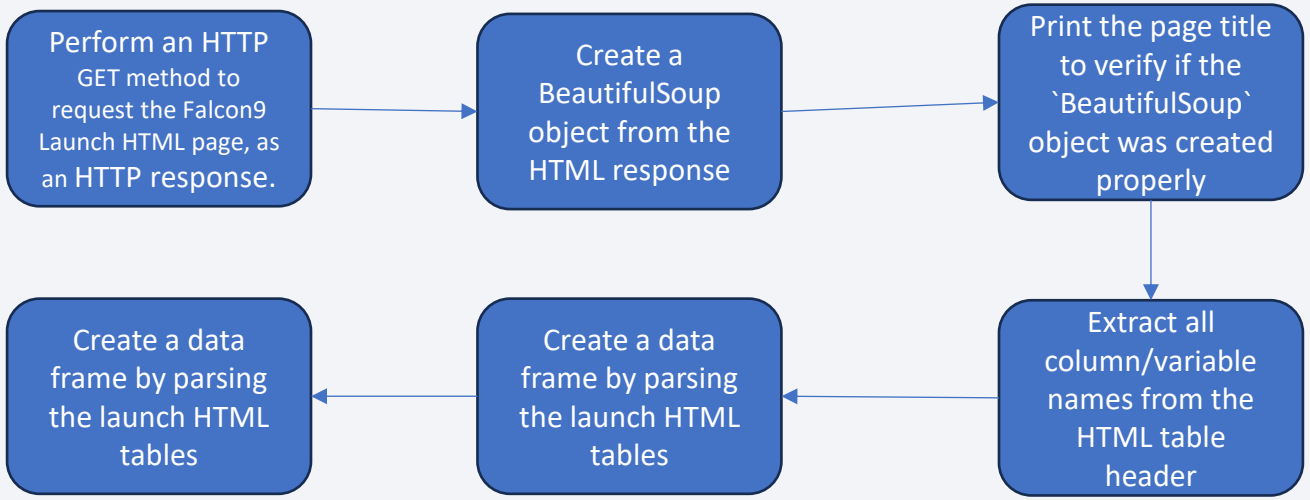
Data Collection – SpaceX API

Flowchart of SpaceX API calls



- Lab 1: Collecting the data

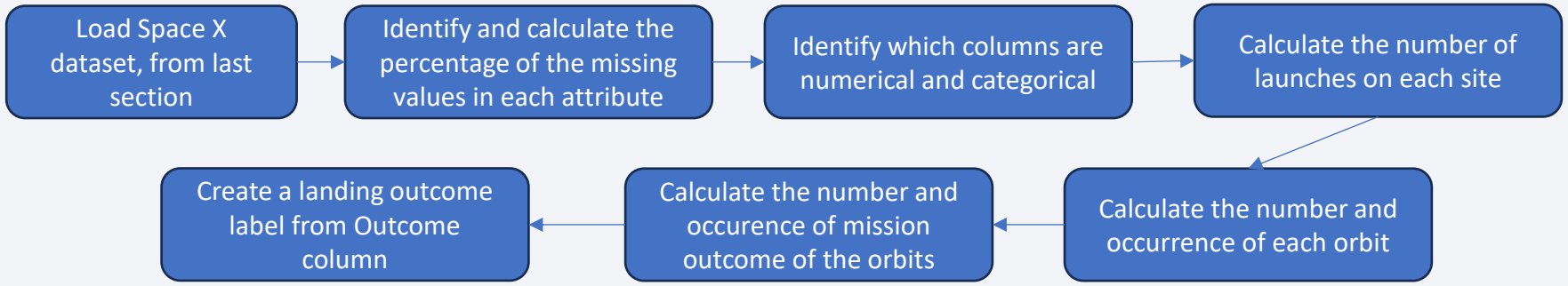
Data Collection - Scraping



Web scraping notebook

Data Wrangling

- In this section, we performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Flow:



- GitHub URL of data wrangling notebooks

EDA with Data Visualization

- Charts were plotted:
  - Visualize the relationship between Flight Number and Launch Site
  - Visualize the relationship between Payload and Launch Site
  - Visualize the relationship between success rate of each orbit type
  - Visualize the relationship between FlightNumber and Orbit type
  - Visualize the relationship between Payload and Orbit type
  - Visualize the launch success yearly trend

GitHub URL of EDA with data visualization notebook.

EDA with SQL

- Performed SQL queries:
  - **Display the names of the unique launch sites in the space mission**
  - **Display 5 records where launch sites begin with the string 'CCA'**
  - **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - **Display average payload mass carried by booster version F9 v1.1**
  - **List the date when the first succesful landing outcome in ground pad was acheived.**
  - **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
  - **List the total number of successful and failure mission outcomes**
  - **List the names of the booster_versions which have carried the maximum payload mass using a subquery**
  - **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, in descending order.**
- GitHub URL of EDA with SQL notebook

Build an Interactive Map with Folium

- Map objects created and added to a folium map:
  - All launch sites on a map
  - The success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
  - The reason to mark this points in the mark was to be able ro find some geographical patterns about launch sites

- GitHub URL of the interactive map with Folium

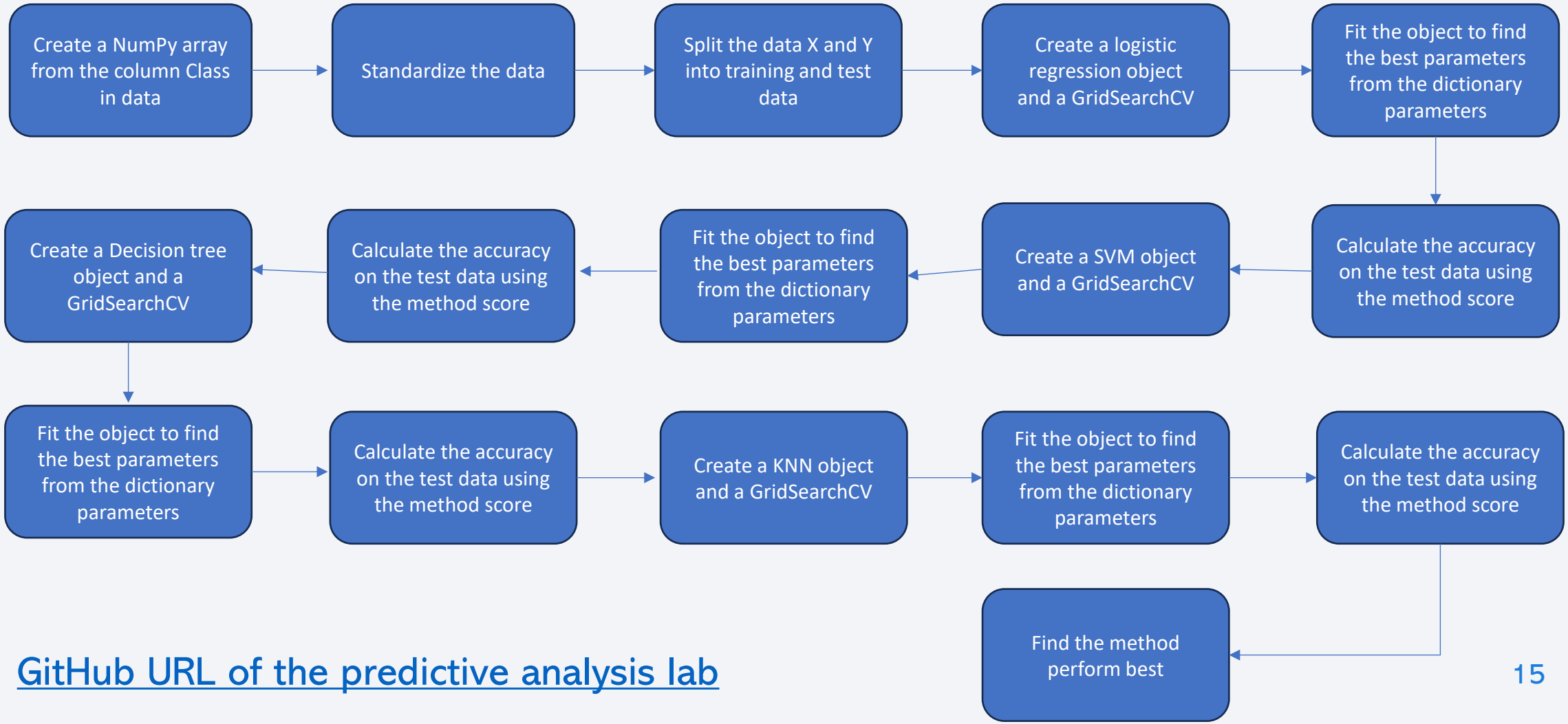
Build a Dashboard with Plotly Dash

- Plots/graphs and interactions you have added to a dashboard
  - A Launch Site Drop-down Input Component
  - A callback function to render success-pi-chart based on selected site dropdown
  - A Range Slider to select Payload
  - A callback function to render success-payload-scatter-chart
- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- After visual analysis using the dashboard, you should be able to obtain some insights to answer the following five questions:

  Which site has the largest successful launches? Which site has the highest launch success rate? Which payload range(s) has the highest launch success rate? Which payload range(s) has the lowest launch success rate? Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

- GitHub URL of the Plotly Dash lab

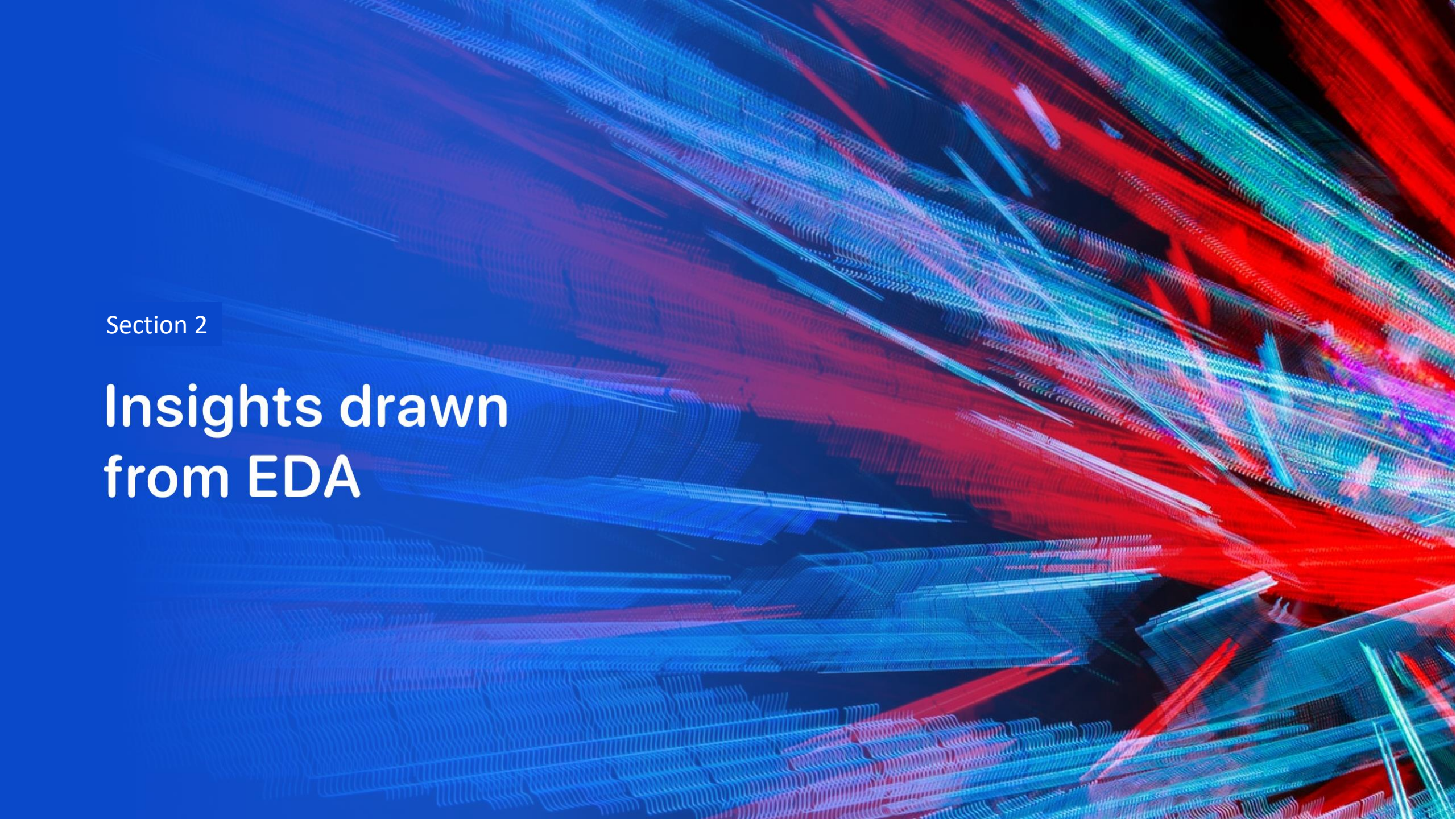
Predictive Analysis (Classification)



- [GitHub URL of the predictive analysis lab]

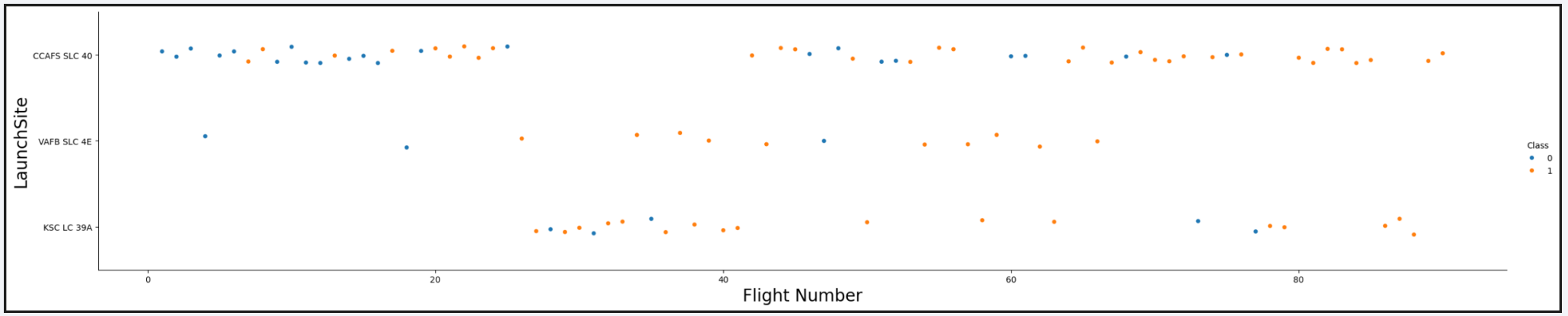
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

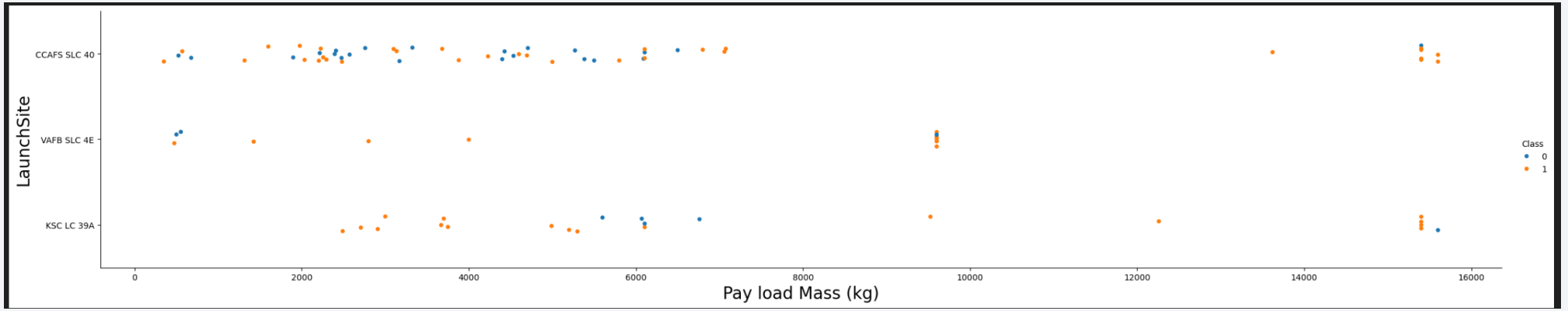
- Scatter plot of Flight Number vs. Launch Site



- Explanations:
  - The CCAFS SLC 40 launch site has about a half of all launches
  - VAFB SLC 4E and KSC LC 39AHave the higher success rates

Payload vs. Launch Site

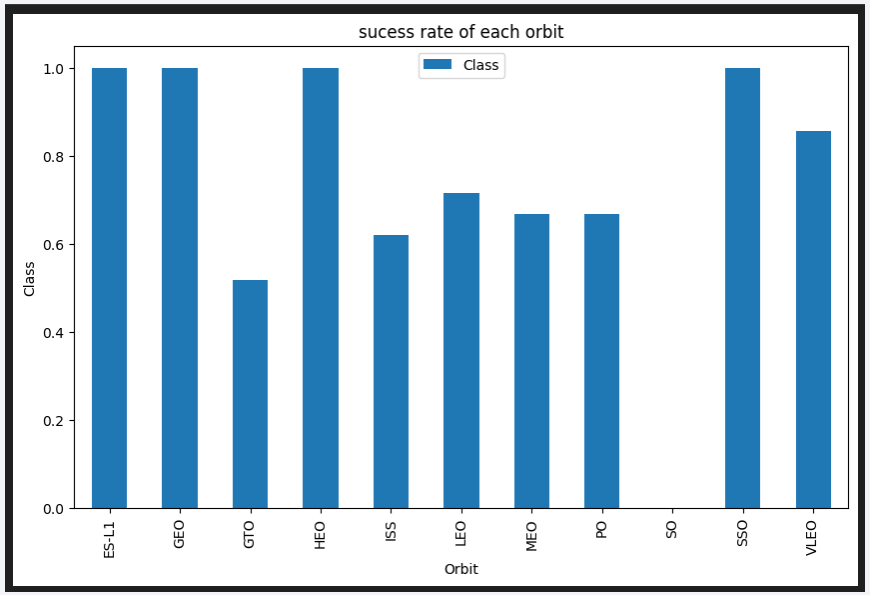
- Scatter plot of Payload vs. Launch Site



- Explanations:
  - Most of the launches with payload mass over 7000 kg were successful
  - KSC LC 39A has a 100% success rate for payload mass under 5000 kg

Success Rate vs. Orbit Type

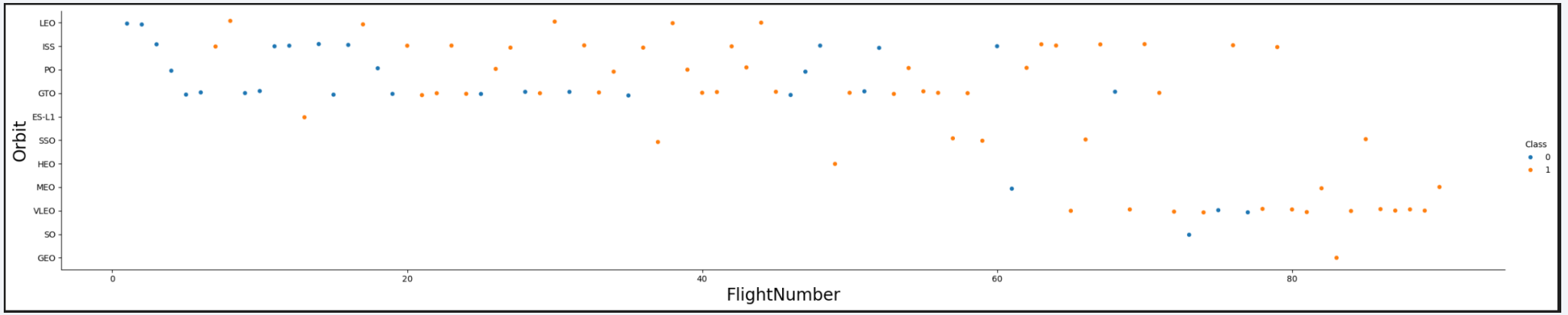
- Bar chart for the success rate of each orbit type



- Explanations:
  - We have 4 orbits with a 100% success rate:
    - ES-L1, GEO, HEO, SSO
  - The orbit SO has a 0% success rate.
  - The rest of the orbits have a success rate between 50-85%

Flight Number vs. Orbit Type

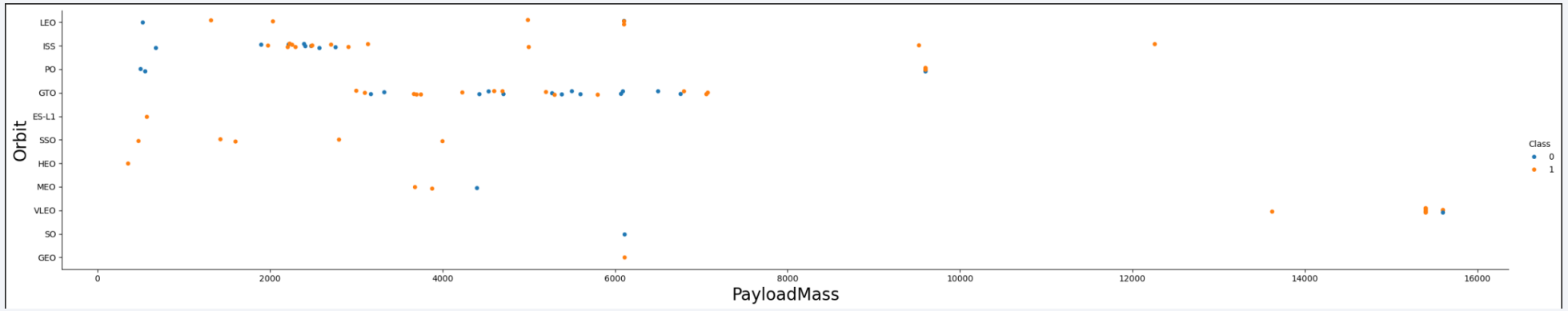
- Show a scatter point of Flight number vs. Orbit type



- Explanations:
  - In the LEO orbit, success appears to be linked to the number of flights, whereas no clear relationship is observed between flight number and success for the rest orbits

Payload vs. Orbit Type

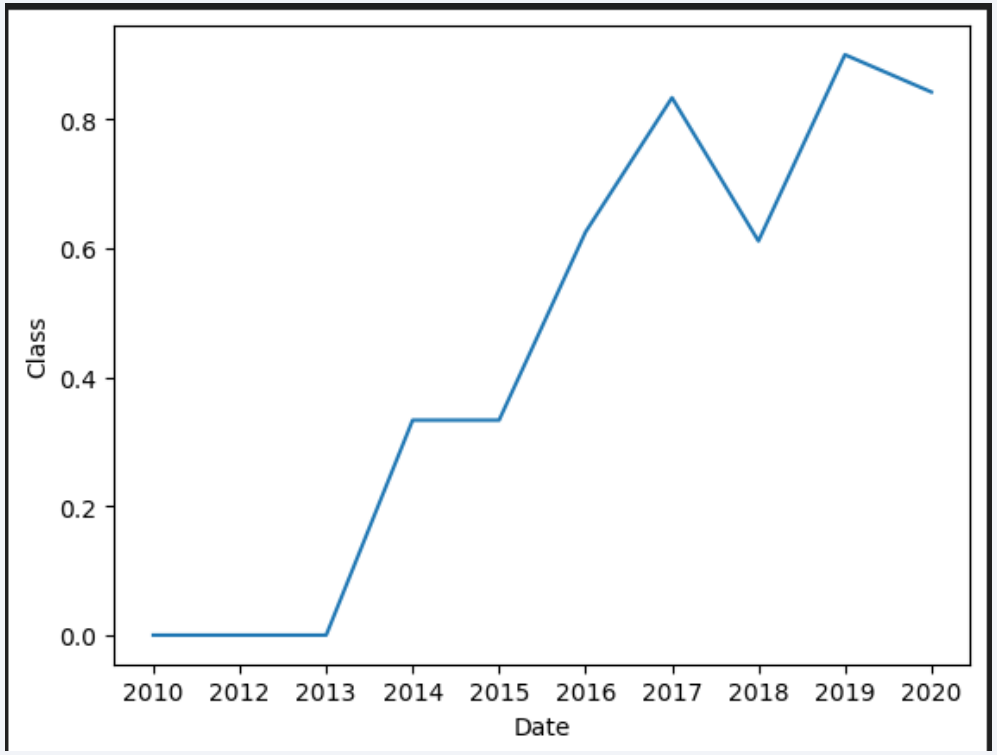
- Scatter point of payload vs. orbit type



- Explanations:
  - With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
  - However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Line chart of yearly average success rate



- Explanations:
  - You can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

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

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

- Explanation:
  - The SELECT DISTINCT statement is used to return only distinct (different) values.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" like "CCA%" LIMIT 5
```

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

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

- Explanation:
  - The LIKE operator is used in a WHERE clause to search for a specified pattern in a column.
  - There are two wildcards often used in conjunction with the LIKE operator:
  - The percent sign % represents zero, one, or multiple characters
  - The underscore sign _ represents one, single character

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

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 = 'NASA (CRS)'
```

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

Total_Payload
45596

- Explanation:
  - The SUM() function returns the total sum of a numeric column.
  - The WHERE clause is used to filter records. It is used to extract only those records that fulfill a specified condition.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version 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 = 'F9 v1.1'
```

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

AVG_Payload
2928.4

- Explanation:
  - The AVG() function returns the average value of a numeric column.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

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

Hint:Use min function

```
%sql SELECT Date FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

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

Date
2015-12-22
2016-07-18
2017-02-19
2017-05-01
2017-06-03
2017-08-14
2017-09-07
2017-12-15
2018-01-08

- Explanation:
  - The answer was achieved listing the dates when the firsts successful landing outcome in ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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 Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4001
```

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

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

- Explanation:
  - The WHERE clause can contain one or many AND operators.
  - The AND operator is used to filter records based on more than one condition

  The BETWEEN operator selects values within a given range. The values can be numbers, text, or dates.

  The BETWEEN operator is inclusive: begin and end values are included.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, count(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

- Explanation:
  - The GROUP BY statement groups rows that have the same values into summary rows, like "find the number of customers in each country".
  - The GROUP BY statement is often used with aggregate functions (COUNT(), MAX(), MIN(), SUM(), AVG()) to group the result-set by one or more columns.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

- Explanation:
  - Subqueries allow you to nest one query inside another, enabling more complex and efficient data retrieval

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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, in descending order.

```
%sql SELECT count(Landing_Outcome), Landing_Outcome FROM SPACEXTBL WHERE (Landing_Outcome = 'Failure (drone ship)' OR Landir
```

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

count(Landing_Outcome)	Landing_Outcome
8	Failure (drone ship)

- Explanation:
  - The COUNT() function returns the number of rows that matches a specified criterion.

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

```
me FROM SPACEXTBL WHERE Date between '2010-06-04' and '2017-03-20' Group by Landing_Outcome ORDER BY count(Landing_Outcome) D
```

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

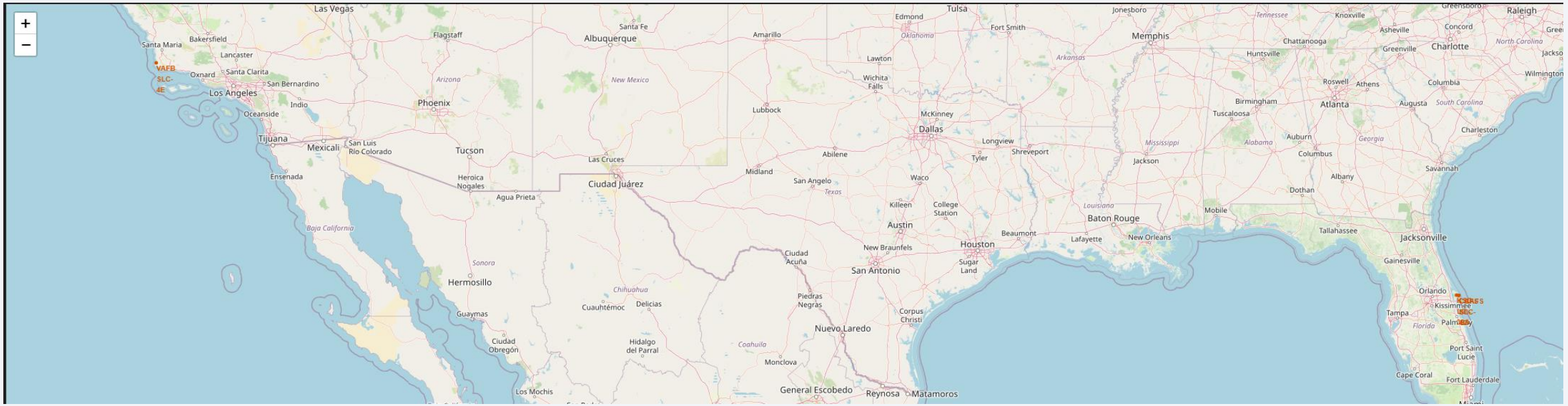
Total	Landing_Outcome
1	Precluded (drone ship)
2	Failure (parachute)
2	Uncontrolled (ocean)
3	Controlled (ocean)
3	Success (ground pad)
5	Failure (drone ship)
5	Success (drone ship)
10	No attempt

- Explanation:
  - The ORDER BY keyword is used to sort the result-set in ascending or descending order.

Section 3

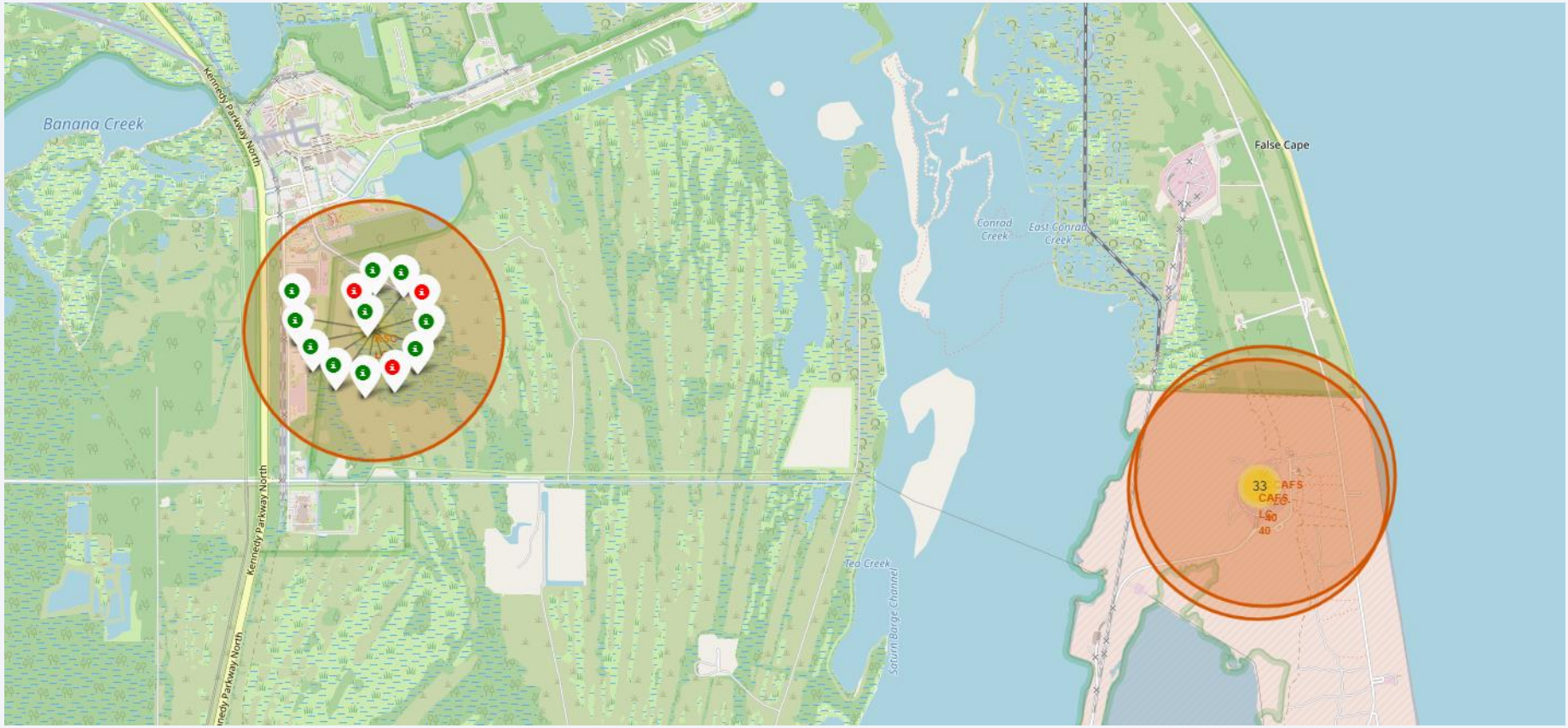
Launch Sites Proximities Analysis

Launch sites



- Launch sites are often near the equator because the Earth's rotation speed is fastest there.
- Launch sites are located near the coast to minimize the risk of debris affecting populated areas.

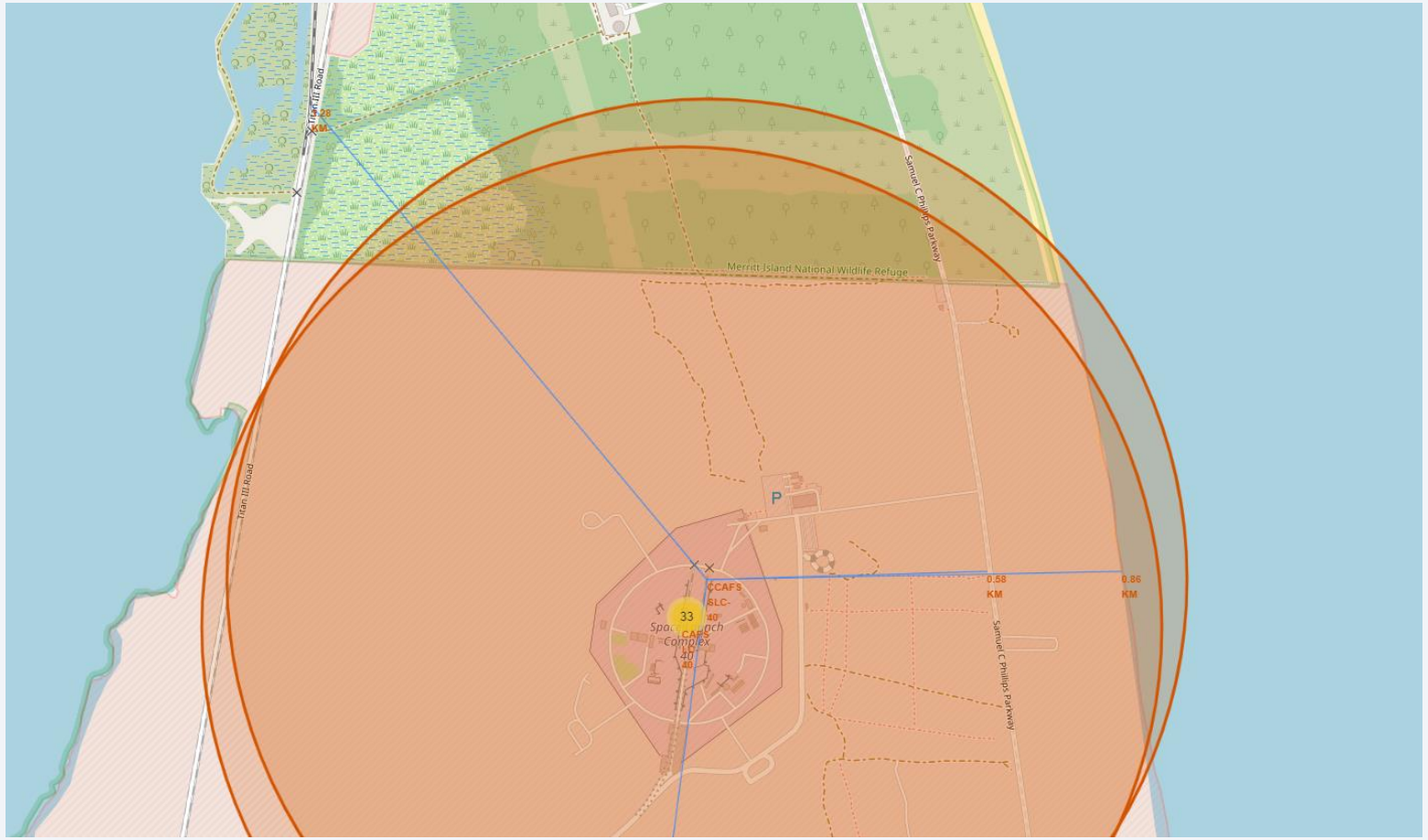
Successful and Failed Launch

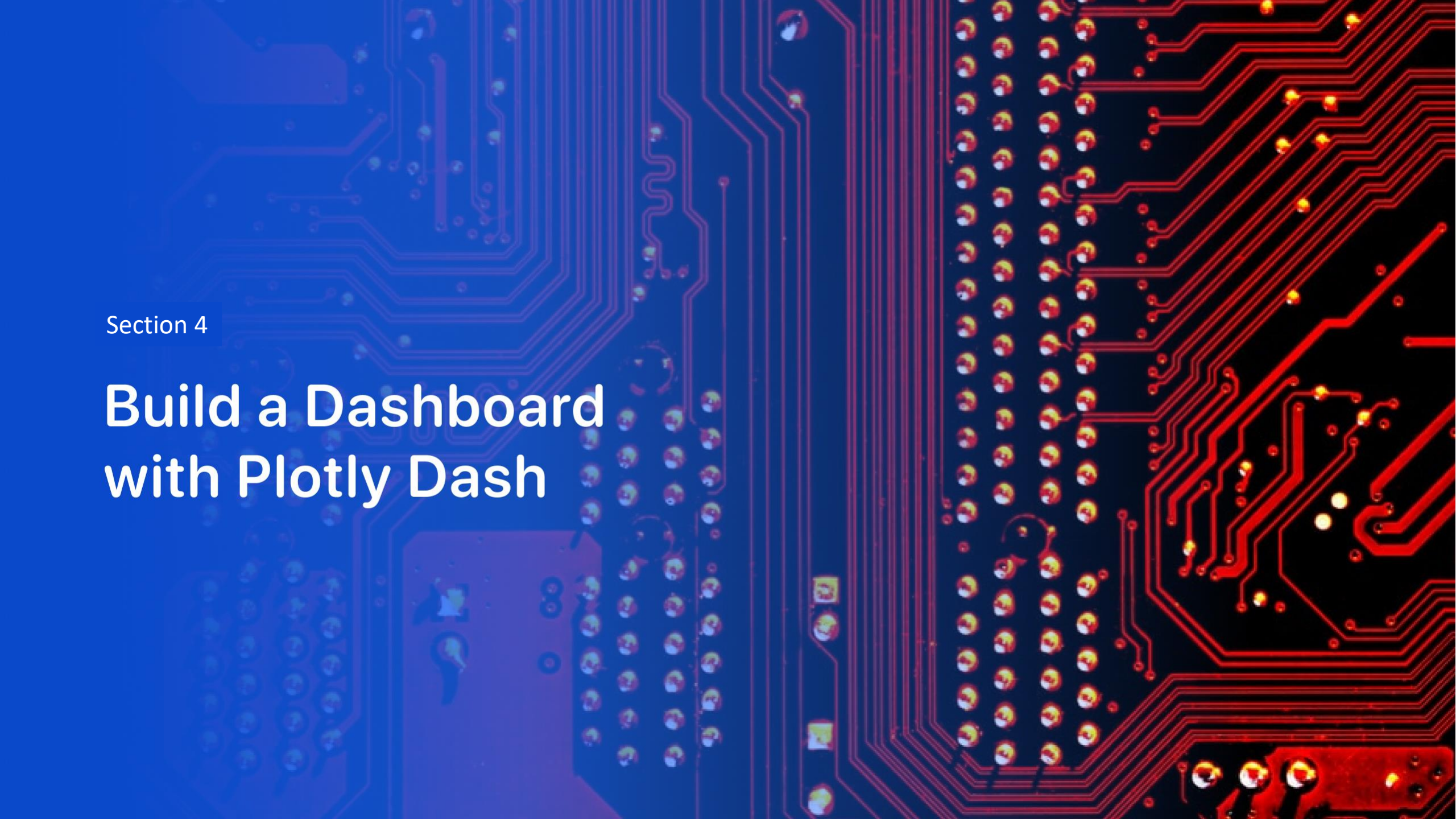


- Green color for successful launch
- Red color for Failes Launch

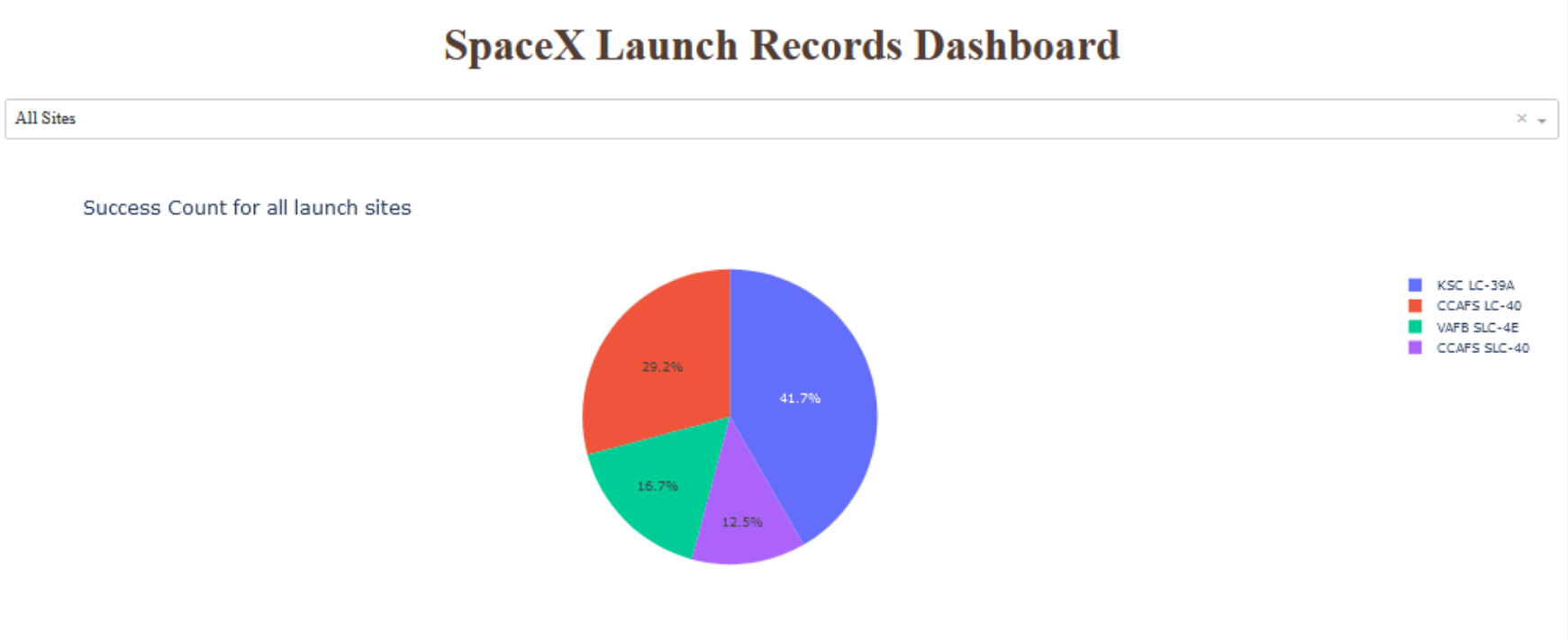
Distances from the launch site

- The launch sites is close to coastline and railway



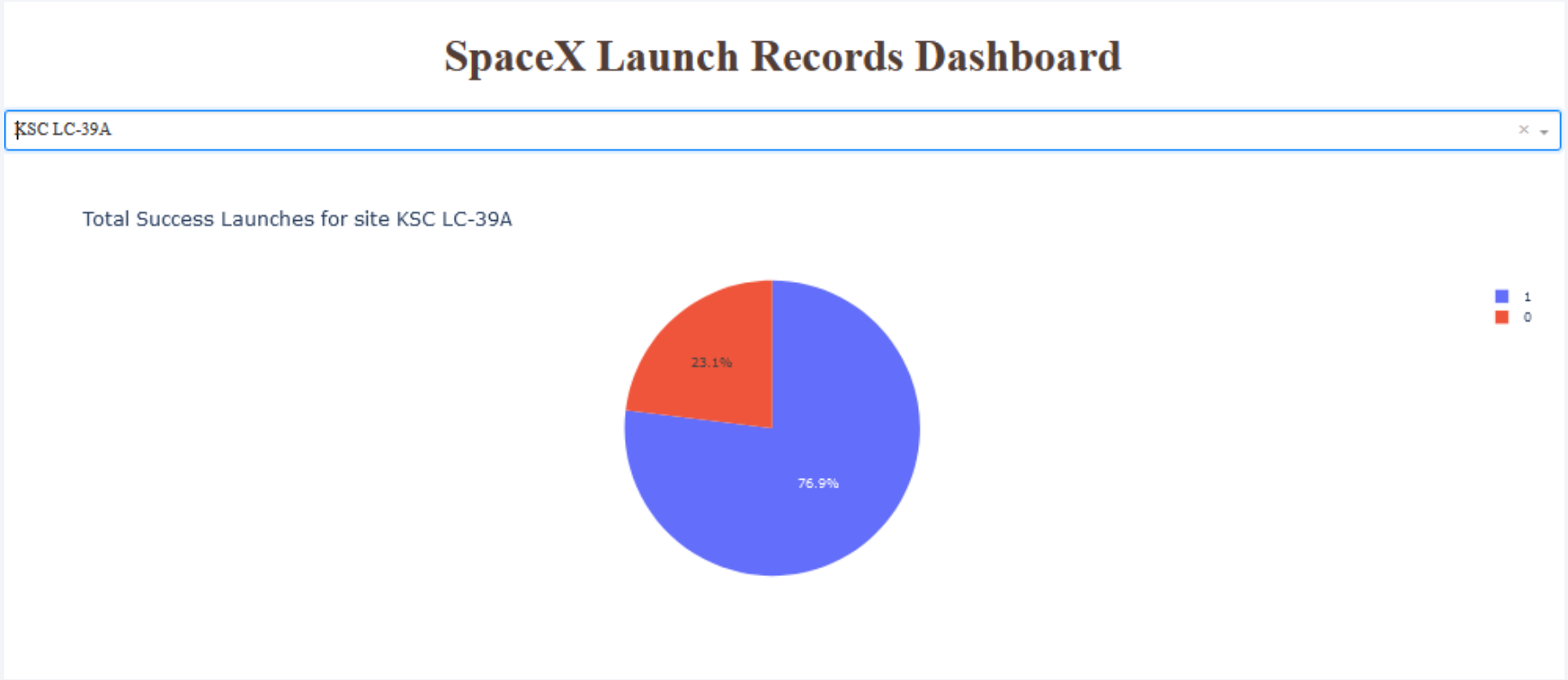

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- Explanation:
  - The KSC LC-39A site has the highest success rate (41.7%), suggesting that it is the most widely used site or has the best track record of successful launches, followed by CCAFS LC-40 with 29.2%.
  - VAFB SLC-4E and CCAFS SLC-40 have lower representation, with 16.7% and 12.5%, respectively.

Total Success Launches for site KSC LC-39A



Explanation:

- This chart highlights that the KSC LC-39A site has a high success rate, with more than three-quarters of launches successfully completed.
- However, the 23.1% failure rate indicates that there is still a significant percentage of launches that were not successful.

Payload vs. Launch Outcome scatter plot for all sites

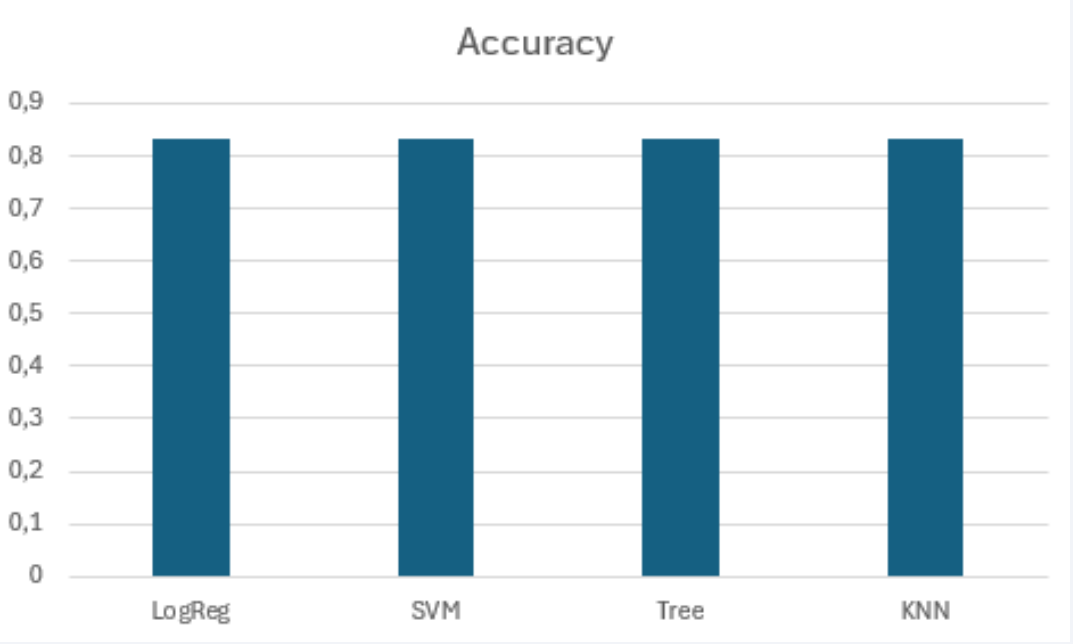
- Explanation:
  - The chart show that payloads between 2K and 5K has the highest success rates

Section 5

Predictive Analysis (Classification)

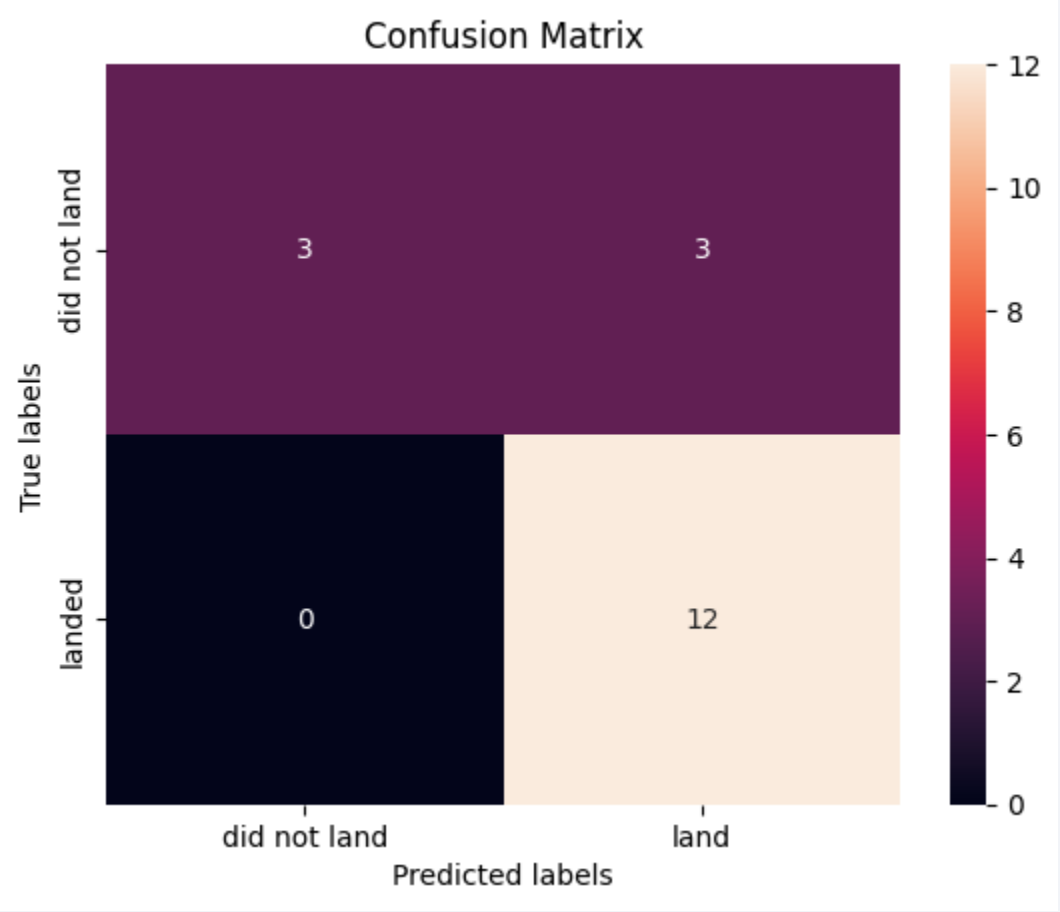
Classification Accuracy

- All the model show the same accuracy: 0.833333



Confusion Matrix

- Explanation:
  - According to the matrix, the model correctly classifies 15 of the cases as true positives or negatives.
  - In the case of erroneous classification, only 3 false positives were detected.



Conclusions

- **Launch Success**: KSC LC-39A launch site has the highest success rate at 41.7%, followed by CCAFS LC-40 at 29.2%. This suggests that KSC LC-39A is either the most used site or has the best track record of successful launches.
- **Relationship between payload and launch success**: Launches with payload mass above 7000 kg were mostly successful. Furthermore, KSC LC-39A has a 100% success rate for payloads below 5000 kg.
- **Annual Launch Success Trend**: The launch success rate has been increasing from 2013 to 2020. This indicates continuous improvement in launch technology and procedures.
- **Predictive Analysis**: The developed classification models showed an accuracy of 83.33%. This suggests that the models are fairly accurate at predicting launch success based on the available data.

Appendix

- In this repository you will find various datasets, notebooks, and python code used during the course

Thank you!