IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adrian Graumnitz
06.03.2025

Outline

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

Executive Summary

- Data Collection
  - with the SpaceX API
  - Web scraping from Wikipedia
    - Summary: Achieve 87% accuracy in data recovery
- Data Wrangling
  - Prepare the Data for Analysis and Prediction

Executive Summary 2

- Analysis through SQL and visualize
  - Store the Data in a Database
  - Create visualization for the Data
- Build interactive Dashboards
  - Build Dashboards for better insights
- ML Prediction on the data
  - Train a model to predict future outcomes based on the data

Introduction

- The goal is to develop reusable rockets by recovering the first stage.
- Compare different types of data, such as launch site and payload, to identify the best conditions for success.
- To achieve this goal python as programming language was used

Section 1

Methodology

Methodology

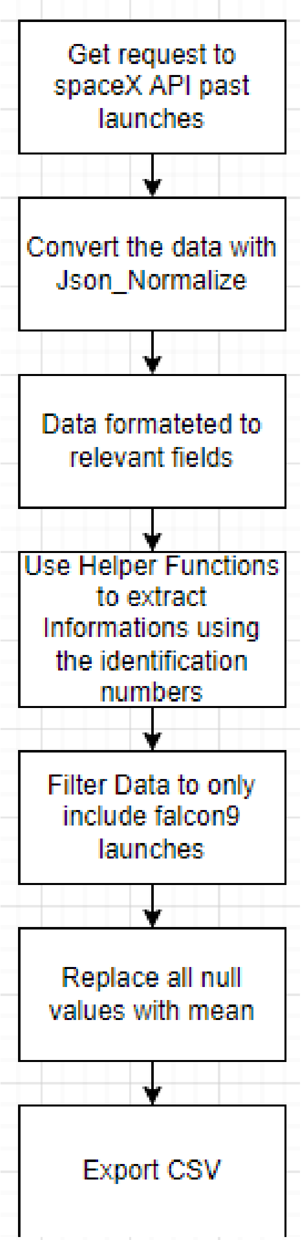
- Executive Summary
- Data collection methodology:
  - Create two Data Sets through SpacX API and the Web Scraper Beautyful Soup
- Perform data wrangling
  - Determine Training Labels and store them in the Data
  - Preparing the Data for the ML Model
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform EDA using visualizations and SQL queries to uncover trends and insights.
  - Visualize through Matplotlib and Seaborn to analyze the Data

Methodology

- Executive Summary
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize the Data
  - Use various approaches to identify the best solution, including Grid Search and Support Vector Machine classification, among others.

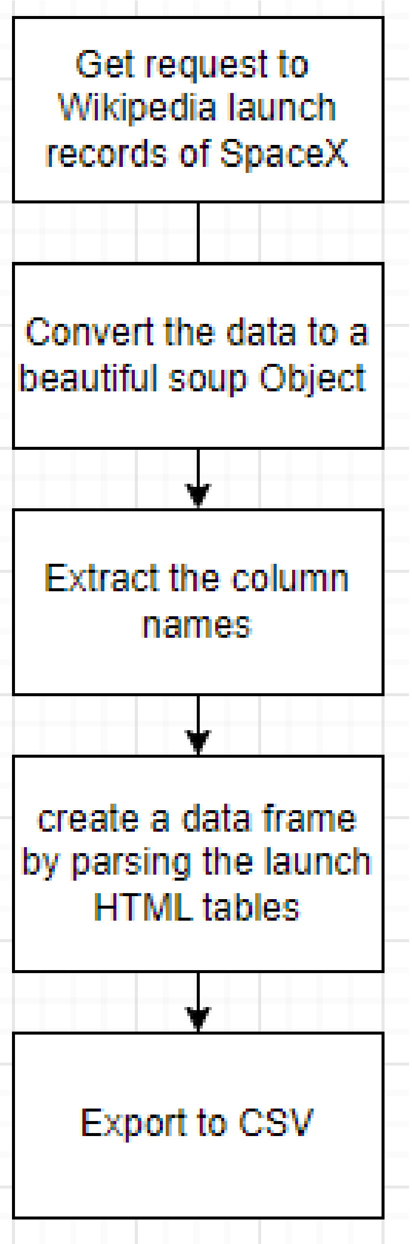
Data Collection – SpaceX API

- Collect data from the SpaceXAPI
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/1.%20Collecting%20the%20Data/1.%20spacex-data-collection-api-v2.ipynb



Data Collection - Scraping

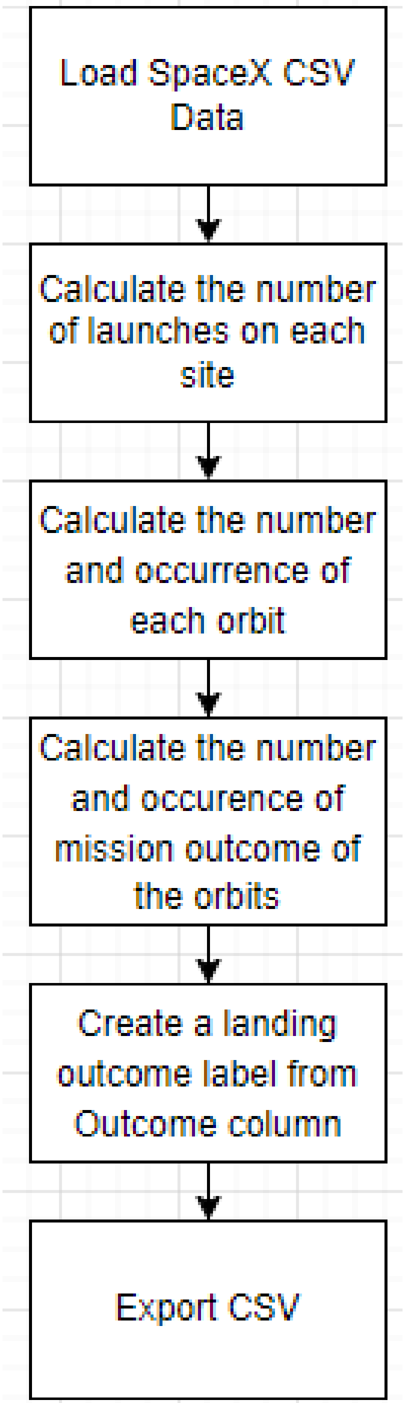
- To collect historical launch data, I used web scraping techniques to gather information from Wikipedia.
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/1.%20Collecting%20the%20Data/2.%20webscraping.ipynb



Data Wrangling 1

- Through the landing outcomes, a 'class' column was created with labels
  - 1 = landed
  - 0 = not landed
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/2.%20Data%20Wrangling/spacex-Data%20wrangling-v2.ipynb

Data Wrangling 2



EDA with Data Visualization

- To predict the success of the Falcon 9 first stage landing, I utilized scatterplots, bar plots, and line charts to explore the relationships between various variables and the landing success.
  - Scatterplots for numerical variables
  - Bar plots for categorical data
  - Line plots for visualizing the success rate trend
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/3.%20Exploratory%20Analysis%20Using%20SQL/eda-datavisualization-with-Pandas-Matplotlib-v2.ipynb

EDA with SQL

- Create a Data Base for the SpaceX Data
- Show the unique launch sites
- Show first successful landing
- Compare Booster Version and payload mass
- List the total number of successful and failure mission outcomes
- Rank the count landing outcomes
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/3.%20Exploratory%20Analysis%20Using%20SQL/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- I create markers for each launch site, to get a better understanding where the launch sites are (near the coast etc.)
- I create circles around the markers to enhance the visualization
- To show the approximates to cities, railroads and coastlines we also use markers and define a line between them and approximated launch site
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/4.%20Interactive%20Dashboards/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

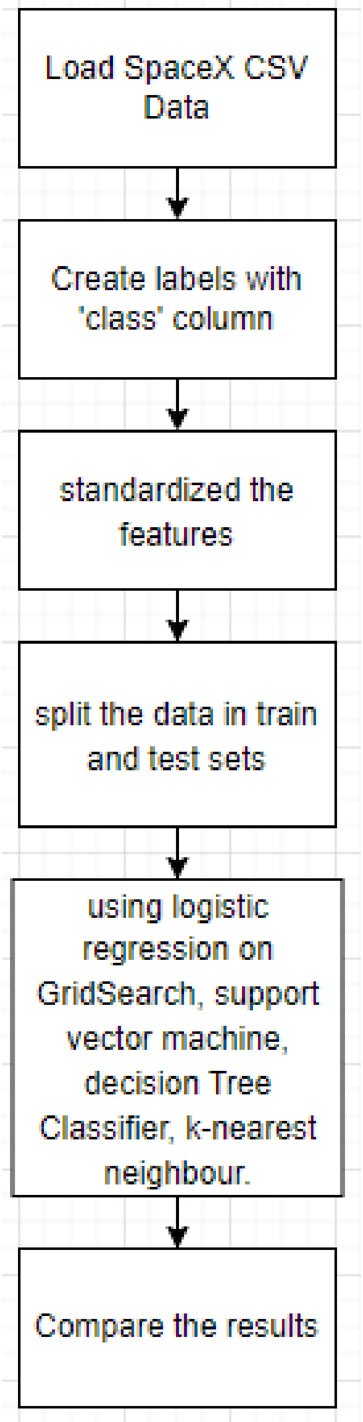
Build a Dashboard with Plotly Dash

- Creating a pie chart to visualize the total success rate by all launching sites and success rate for every launching site
- Creating a range slider with a scatterplot to show the relation between the success rate, the payload mass and the booster version
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/4.%20Interactive%20Dashboards/Interactive%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification) 1

- We create a test and training set and use different concepts to find the best working hyperparameters
- In the end the results are almost the same, the true positive results are perfect, but it also have false positive results (3 by every test)
- The best test accuracy is 83.33%
- https://github.com/AdrianGraumnitz/IBM-Professional-Certificate/blob/main/_Capstone%20Project/5.%20Predicitive%20Analysis/SpaceX-Machine-Learning-Prediction.ipynb

Predictive Analysis (Classification) 2

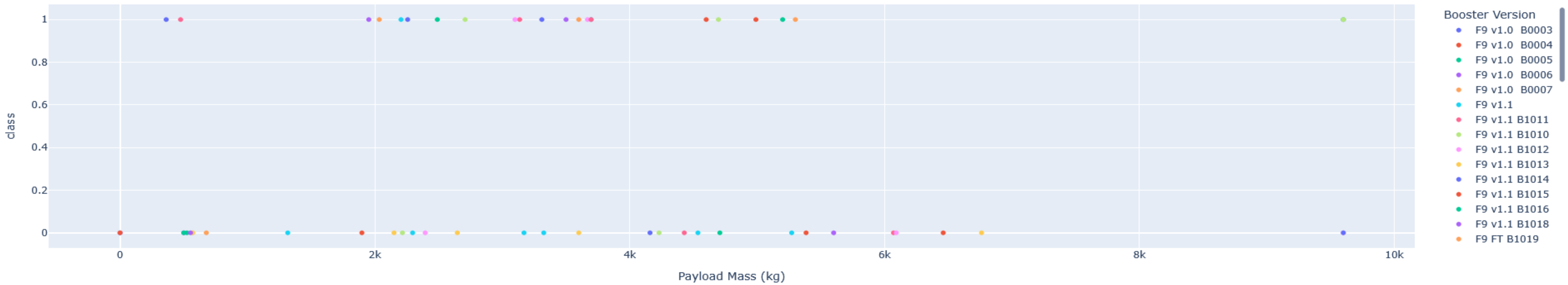


Results 1

- A model was created using SpaceX data to improve the prediction accuracy of the first-stage landing success rate. By analyzing historical data, the model identifies key factors influencing landing outcomes and provides insights that could contribute to higher success rates in future launches.

Select Payload Mass Range (kg)

0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000

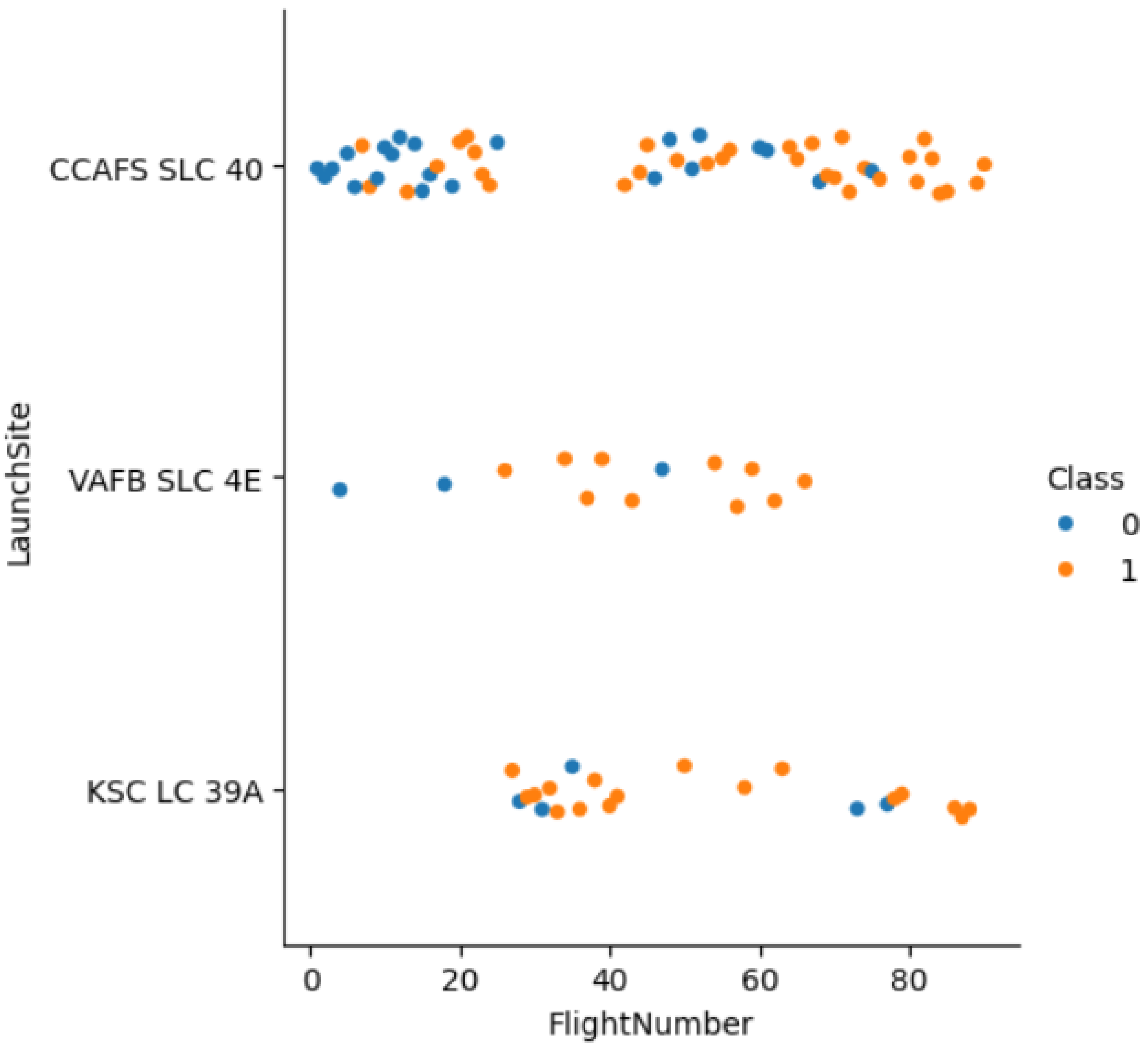


Results 2

Section 2

Insights drawn from EDA

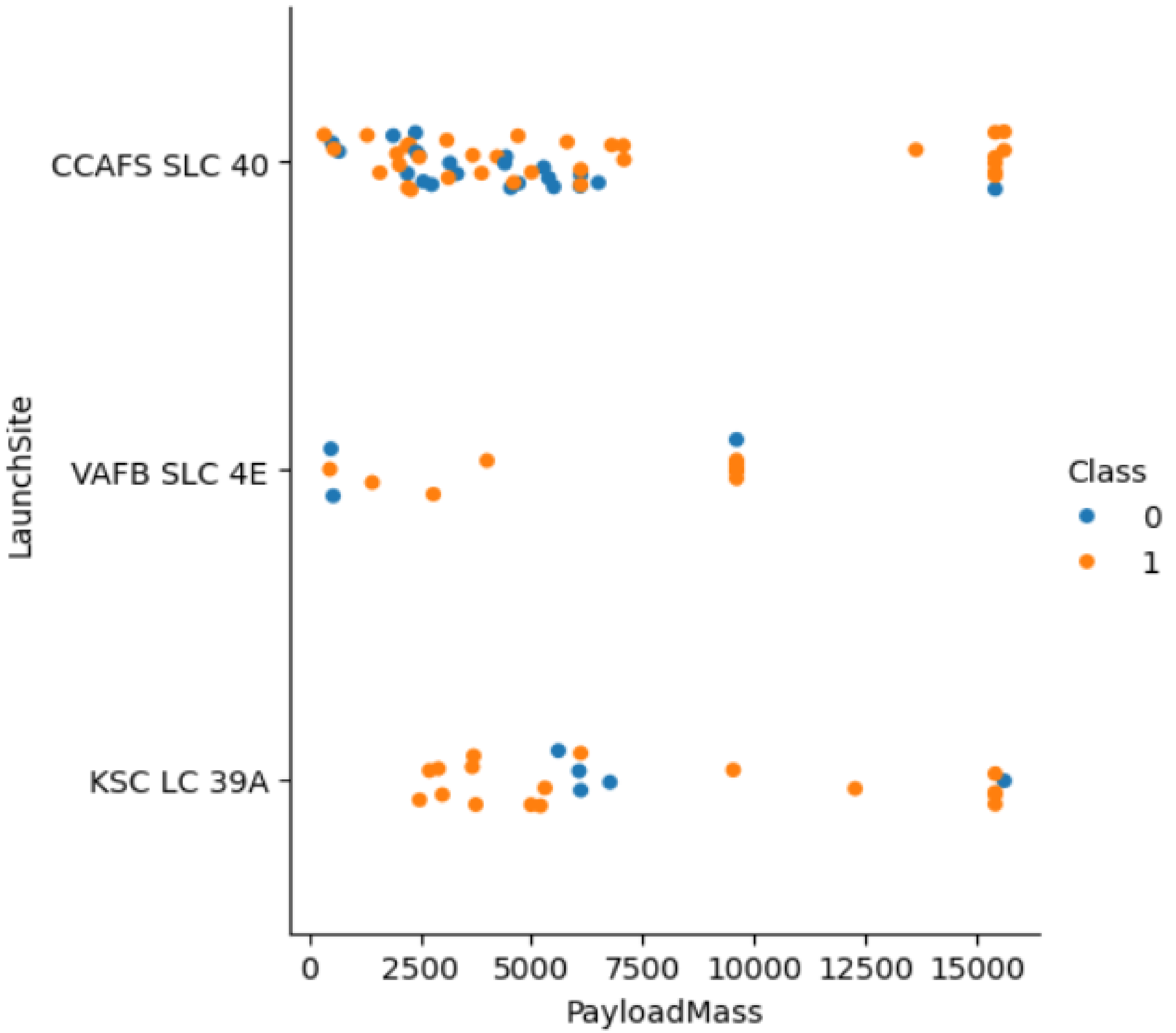
Flight Number vs. Launch Site 1



Flight Number vs. Launch Site 2

- On the foil, you can see that by increasing the flight number, more rockets are landing.

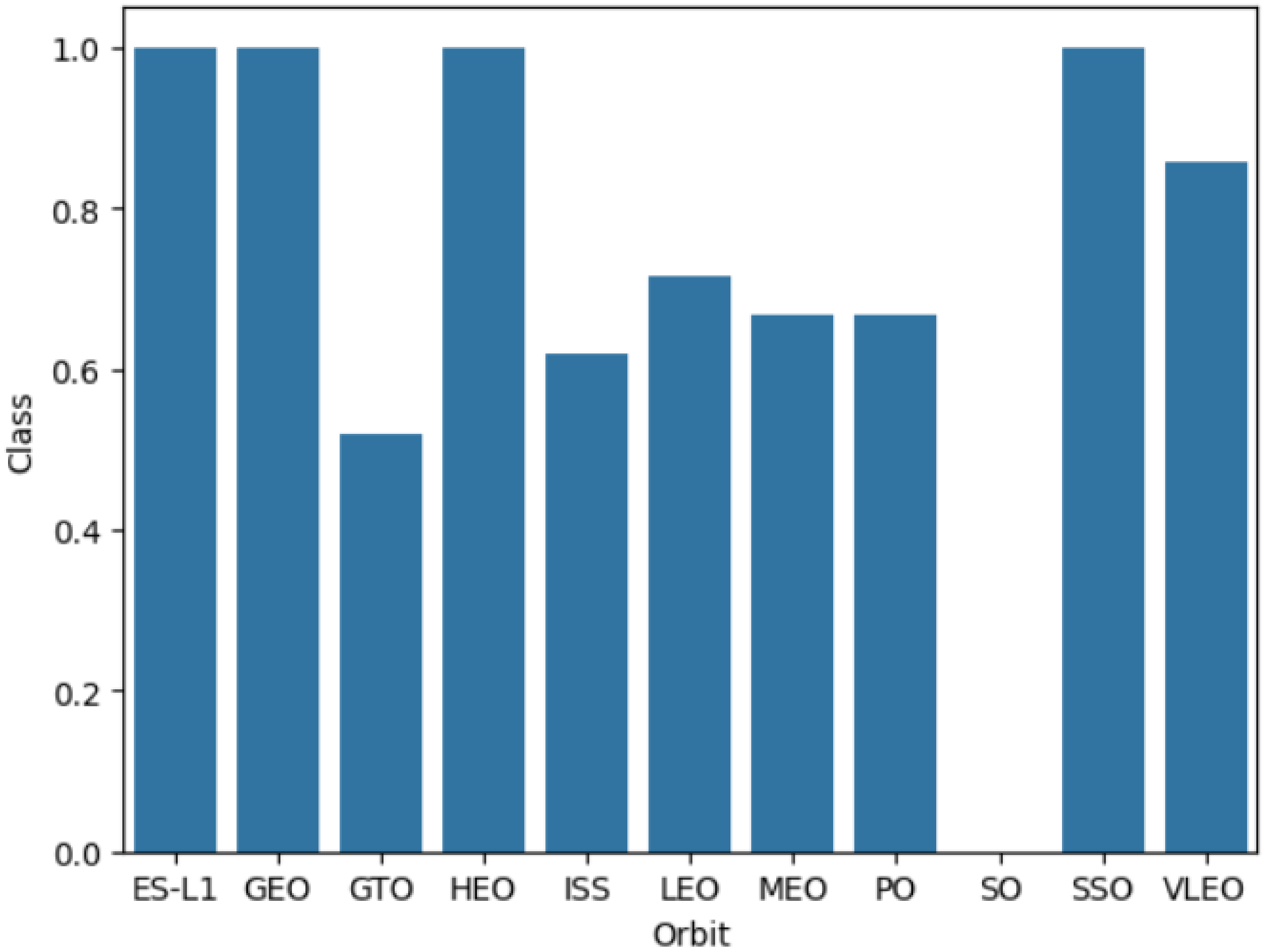
Payload Mass vs. Launch Site 1



Payload Mass vs. Launch Site 2

On the foil, you can see that by increasing the Payload Mass, more rockets are landing.

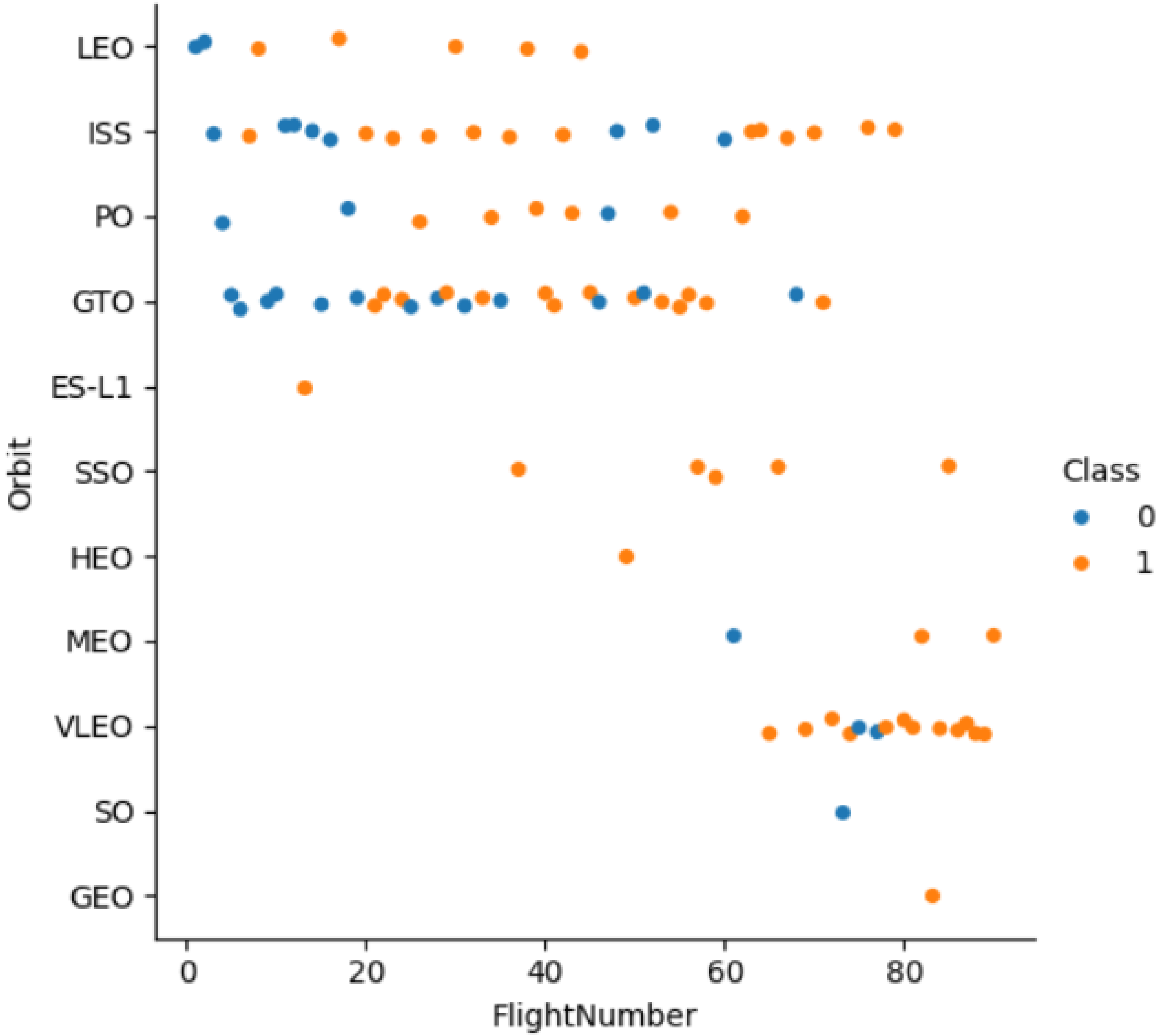
Success rate vs orbit type 1



Success rate vs orbit type 2

- On the foil, you can see that the success rate is the highest by the orbit types:
  - ES-L1
  - GEO
  - HEO
  - SSO

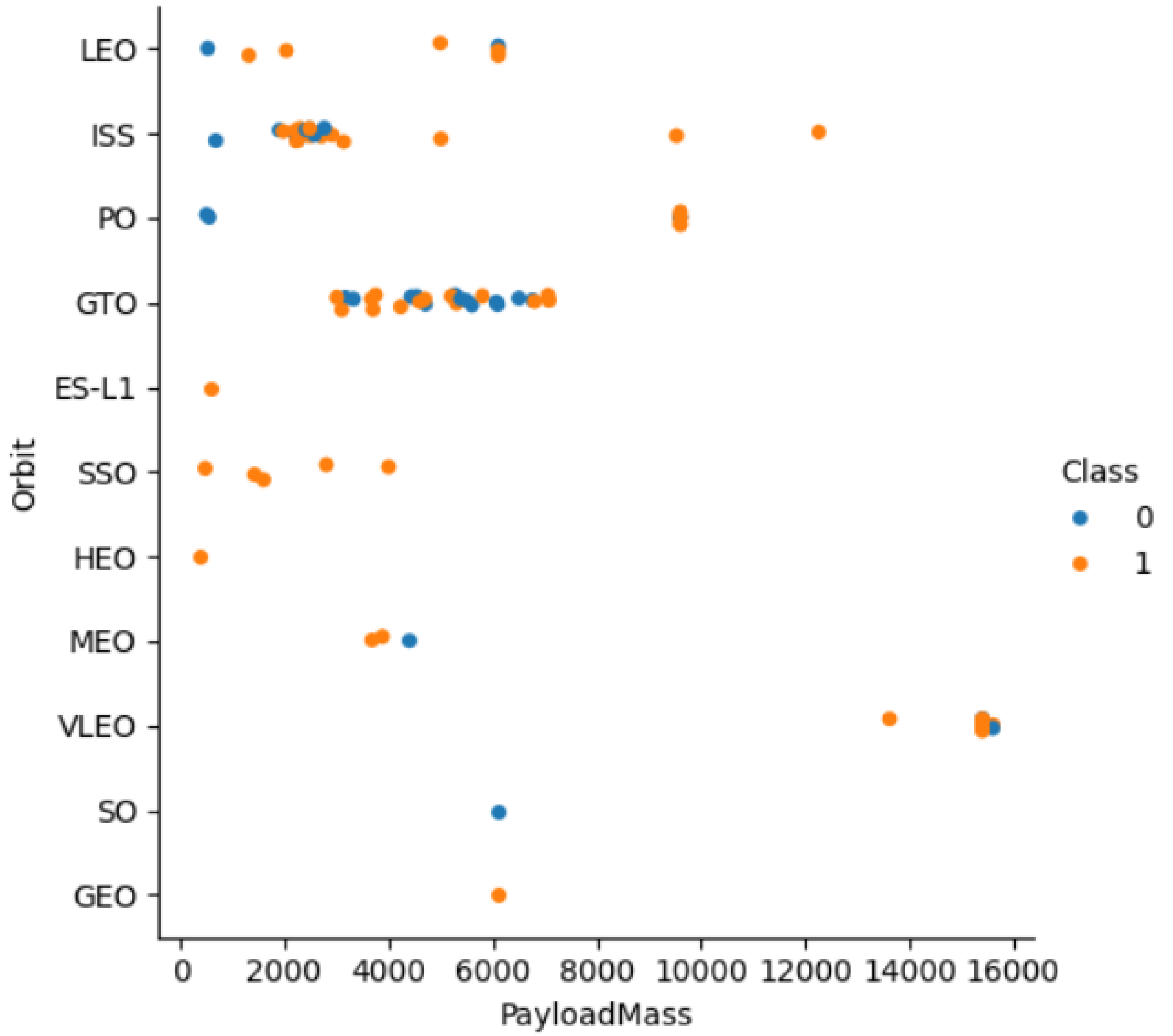
Orbit type vs Flight-number 1



Orbit type vs Flight-number 2

- On the foil, you can see that most flights go to:
  - ISS
  - GTO
- The following orbit stations have a small percentage of failed landings:
  - LEO
  - ISS
  - PO
  - VLEO

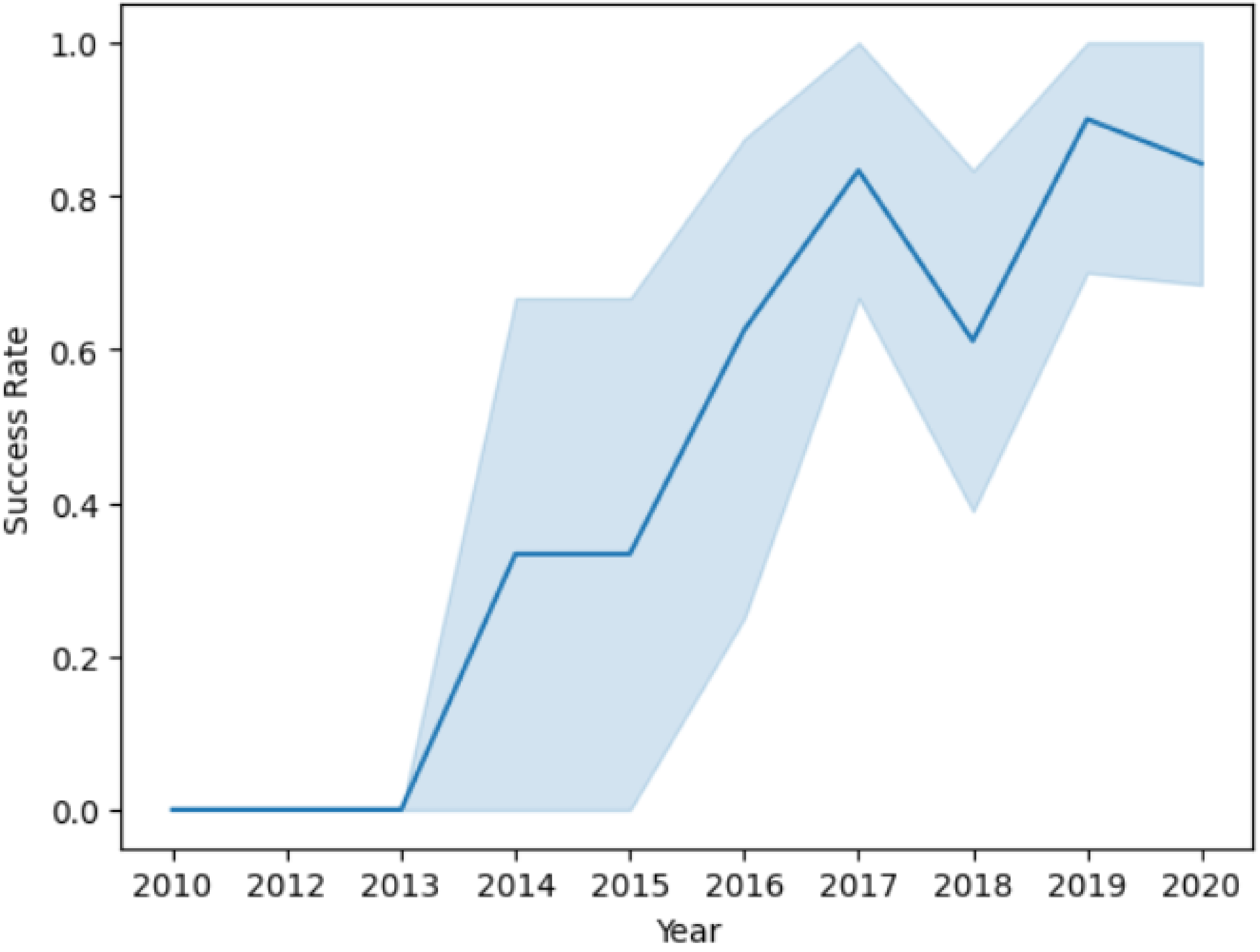
Payload vs orbit type 1



Payload vs orbit type 2

- On the foil, you can see that payload mass tends to increase with certain orbits like GTO and GEO, while successful landings (Class 1) are more frequent in lower payload masses across various orbits

Launch success yearly trend



Launch success yearly trend

- On the foil, you can see that the success rate increases over the years, with break in 2018 and 2020.

- The launch sites are:
  - Cape Canaveral Space Force Station (CCAFS LC-40 and CCAFS SLC-40)
  - Vandenberg Space Force Base (VAFB SLC-4E)
  - Kennedy Space Center (KSC LC-39A)

```
%sql select distinct Launch_Site from spacextable
```

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

All Launch Site Names

```
%sql select * from spacextable where Launch_Site like 'CCA%' Limit 5
```

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

- You can see that none of them have information about a positive landing outcome. They were all going to the LEO orbit station.

Launch Site Names Begin with 'CCA'

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as 'Total Payload Mass' from spacextable where customer like '%NASA%'
```

- The total payload mass carried by boosters launched by NASA

Total Payload Mass
107010

```
%sql select avg(PAYLOAD_MASS__KG_) as 'Average Payload Mass from F9 v1.1' from spacextable where Booster_Version = 'F9 v1.1'
```

- Average payload mass carried by booster version F9 v1.1

Average Payload Mass from F9 v1.1
2928.4

Average Payload Mass by F9 v1.1

```
%sql select min(date) as 'First Successfull Landing' from spacextable where Landing_Outcome like '%Success'
```

- The date when the first successful landing outcome in ground pad was achieved

First Successfull Landing
2018-07-22

First Successful Ground Landing Date

```
%sql select Booster_Version from spacextable where Payload_Mass__Kg_ between 4000 and 6000 and Landing_Outcome = 'Success (drone ship)'
```

- Names of the boosters 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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select count(*) as 'Total Number of successful and failure mission outcomes' from spacextable where Mission_Outcome like '%Success' or Mission_Outcome like '%Failure'
```

Total Number of successful and failure mission outcomes
98

Total Number of Successful and Failure Mission Outcomes

```
%sql select Booster_Version from spacextable where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextable)
```

- Names of the booster versions which have carried the maximum payload mass

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

Boosters Carried Maximum Payload

```
%sql select Booster_Version, Launch_Site, Landing_Outcome,Date, substr(Date,6,2) from spacextable where Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0, 5) ='2015'
```

- Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015

Booster_Version	Launch_Site	Landing_Outcome	Date	substr(Date,6,2)
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10	01
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14	04

2015 Launch Records

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

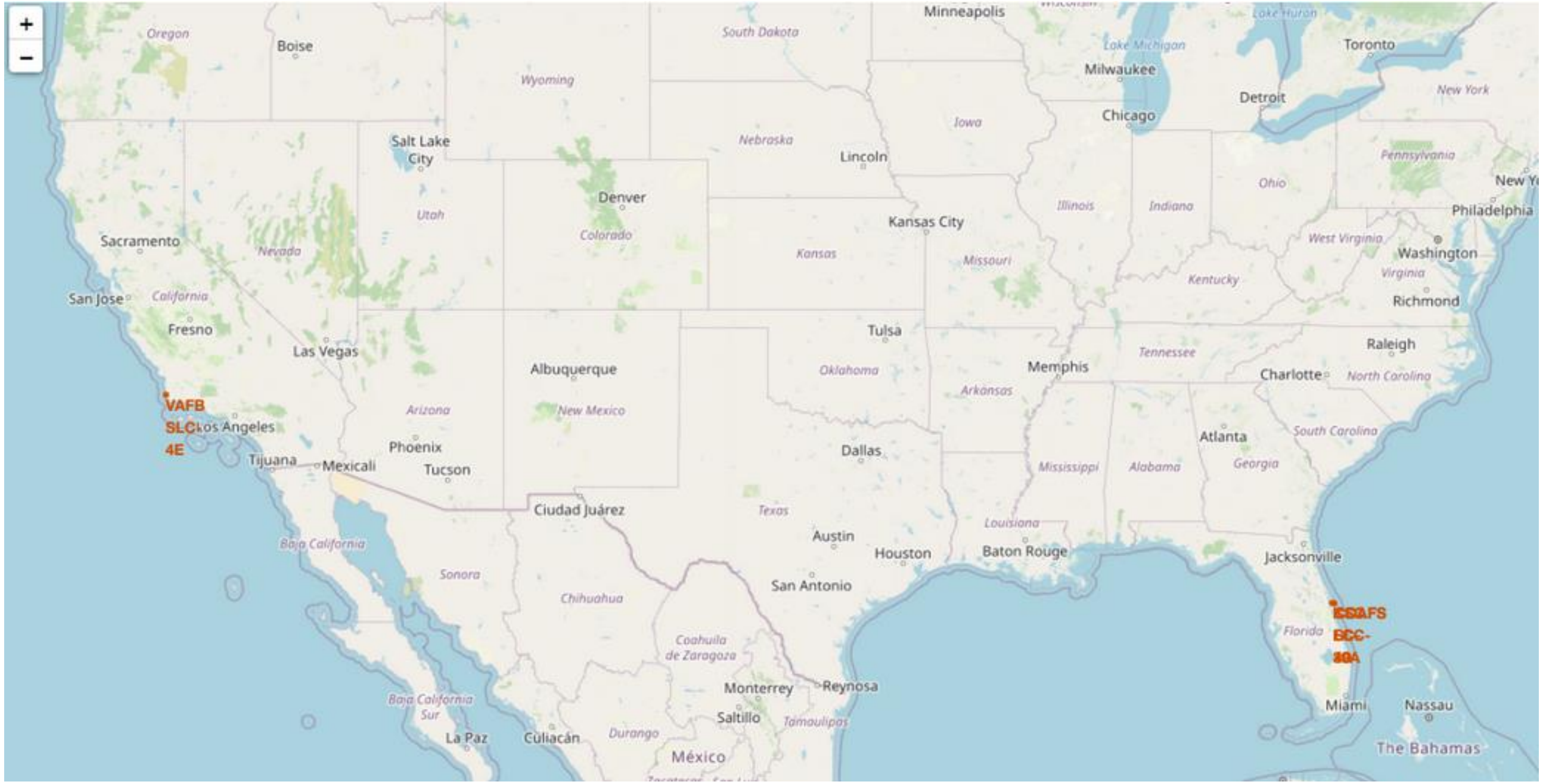
- Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015

```sql
%%sql
SELECT Landing_Outcome, COUNT(*) AS 'Count of Landing Outcomes'
FROM spacextable
WHERE (Landing_Outcome = 'Failure (drone ship)' OR Landing_Outcome = 'Success (ground pad)')
 AND date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY 'Count of Landing Outcomes' DESC;
```

Landing_Outcome	Count of Landing Outcomes
Success (ground pad)	3
Failure (drone ship)	5

Section 3

Launch Sites Proximities Analysis

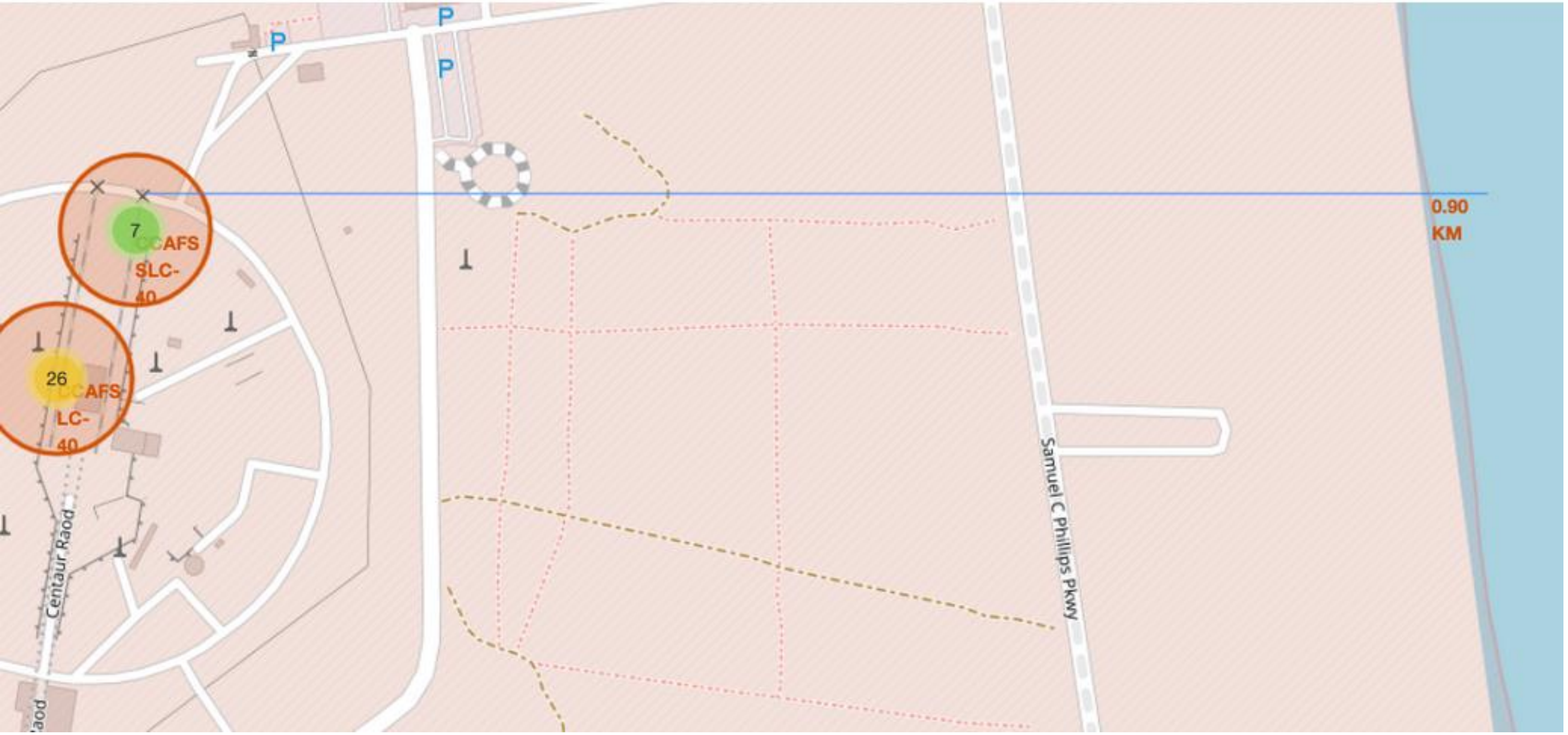


- The launch sites are marked in orange, with their names displayed on the map

Launch sites

- The green launch outcomes indicate that the landing was successful, while the red ones show that it was not

Launch outcomes



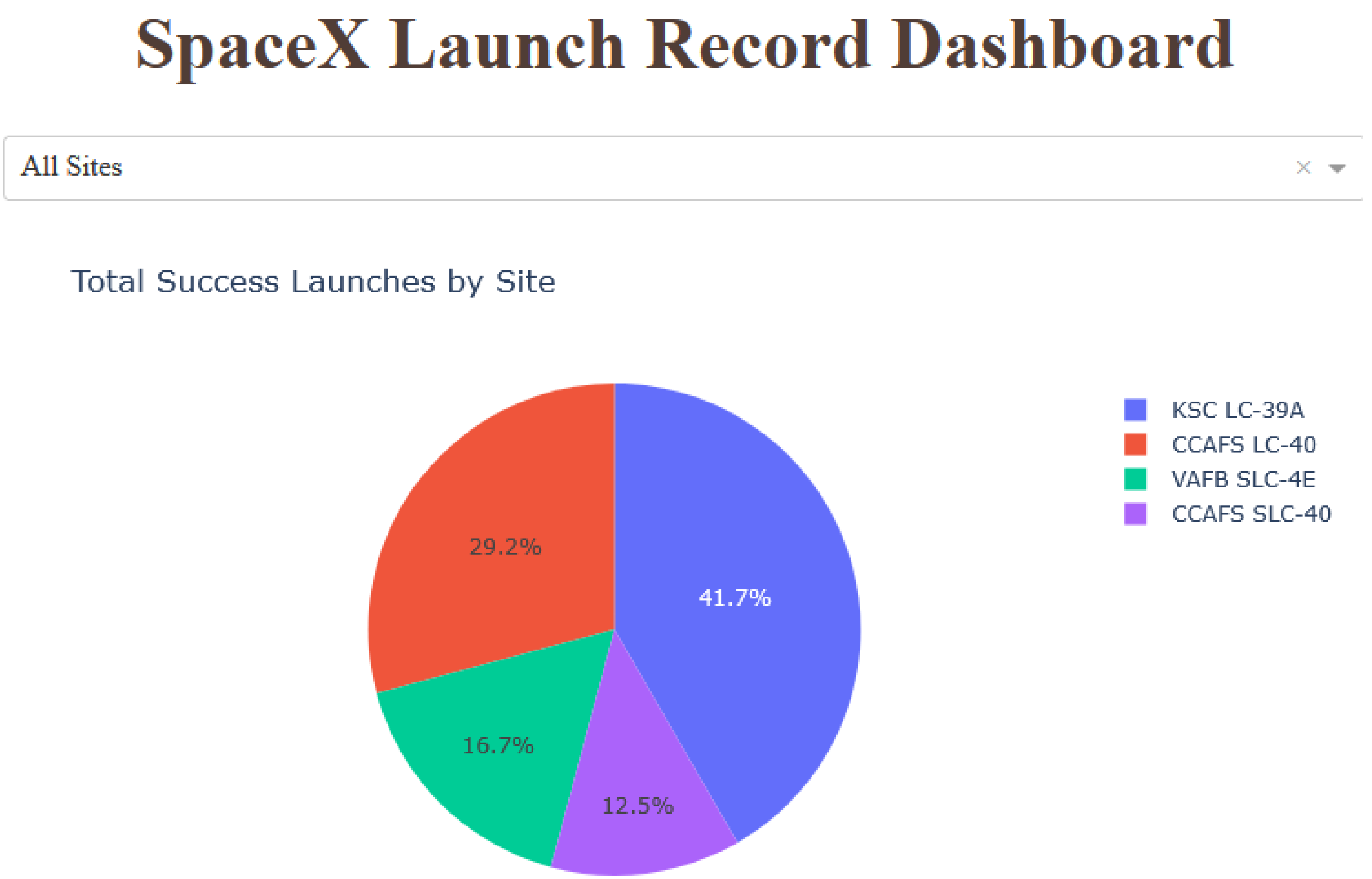
- There are 900 meters to the nearest coastline from Cape Canaveral

Distance to coastline

Section 4

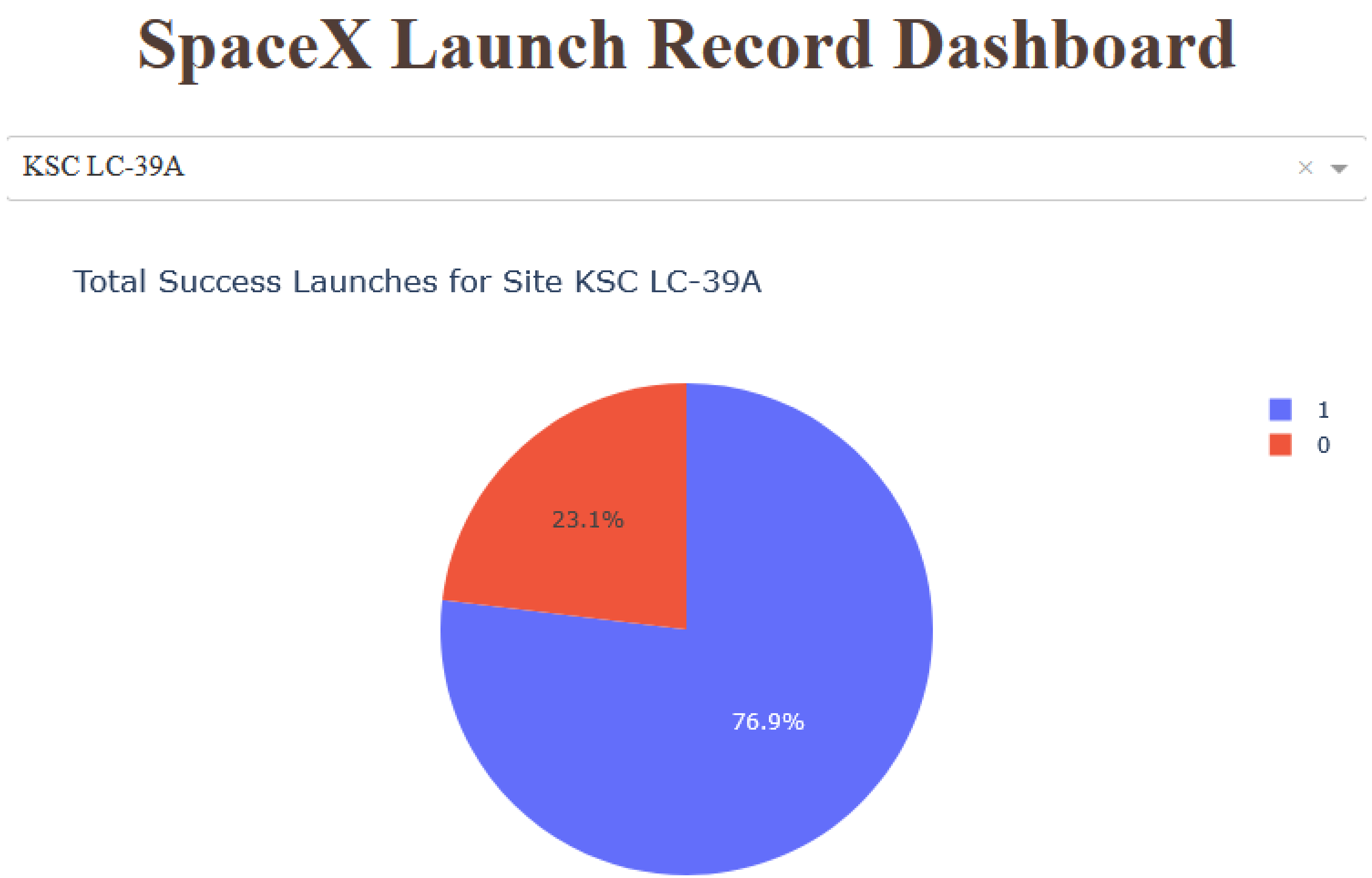
Build a Dashboard with Plotly Dash

Dashboard pie chart 1



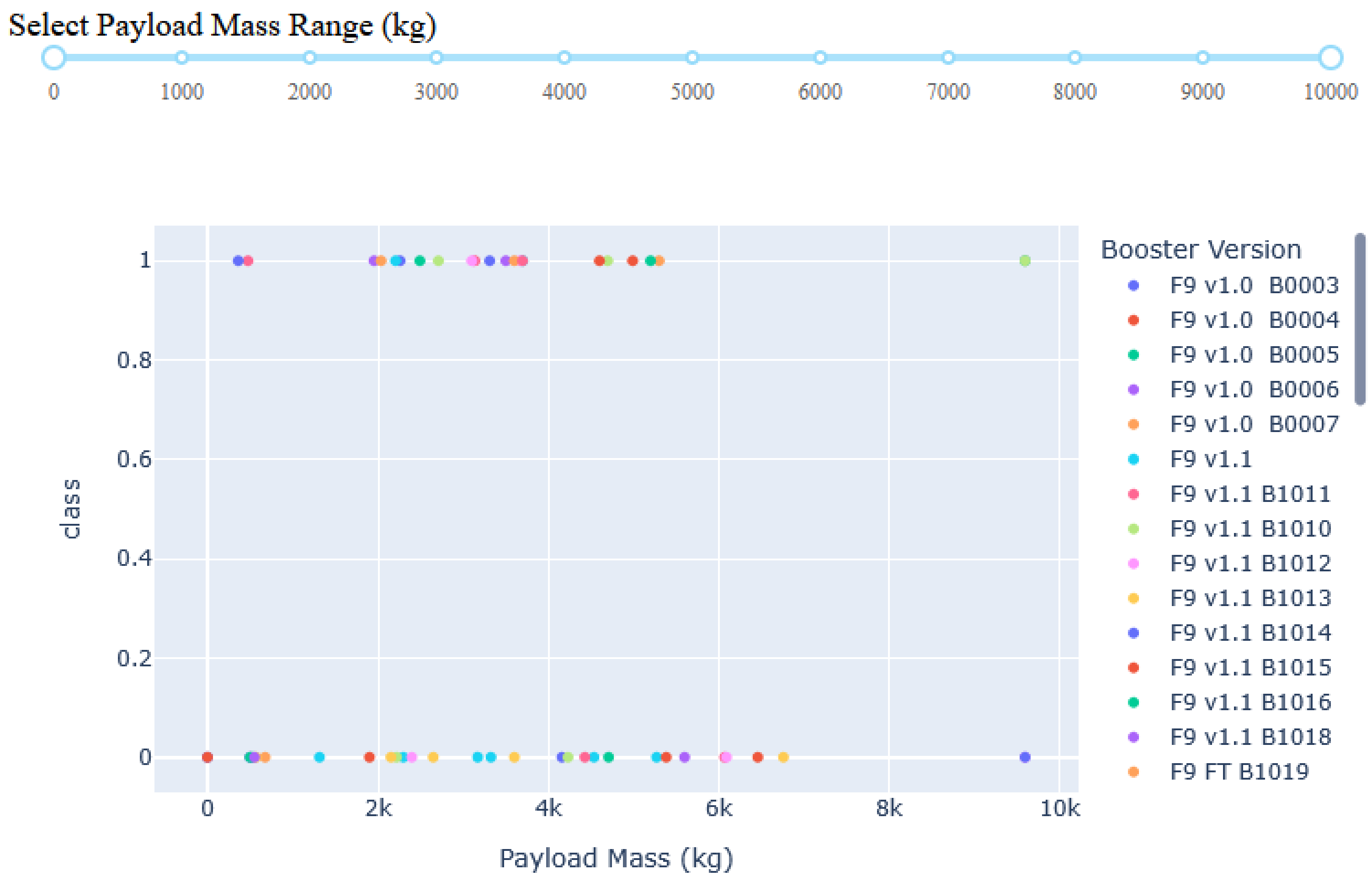
- Kennedy Space Center, followed by Cape Canaveral, has the most successful launches

Dashboard pie chart 2



- Kennedy Space center has the highest success rate for launches

Dashboard range slider, scatter plot

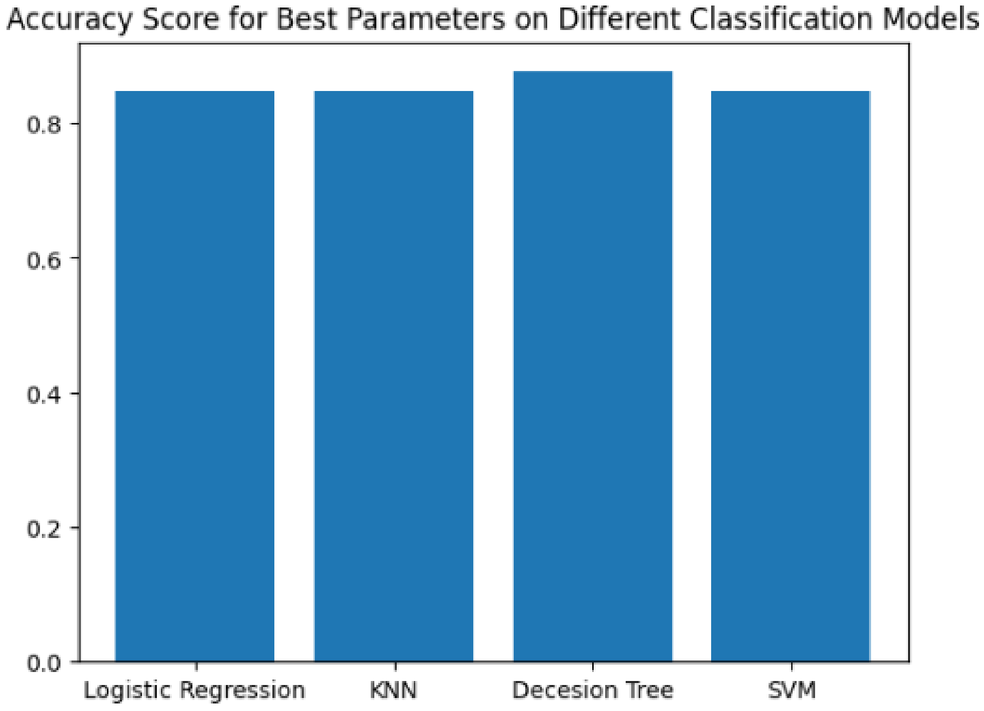


- The highest success rate for landing the first stage is between 2000 kg and 6000 kg

Section 5

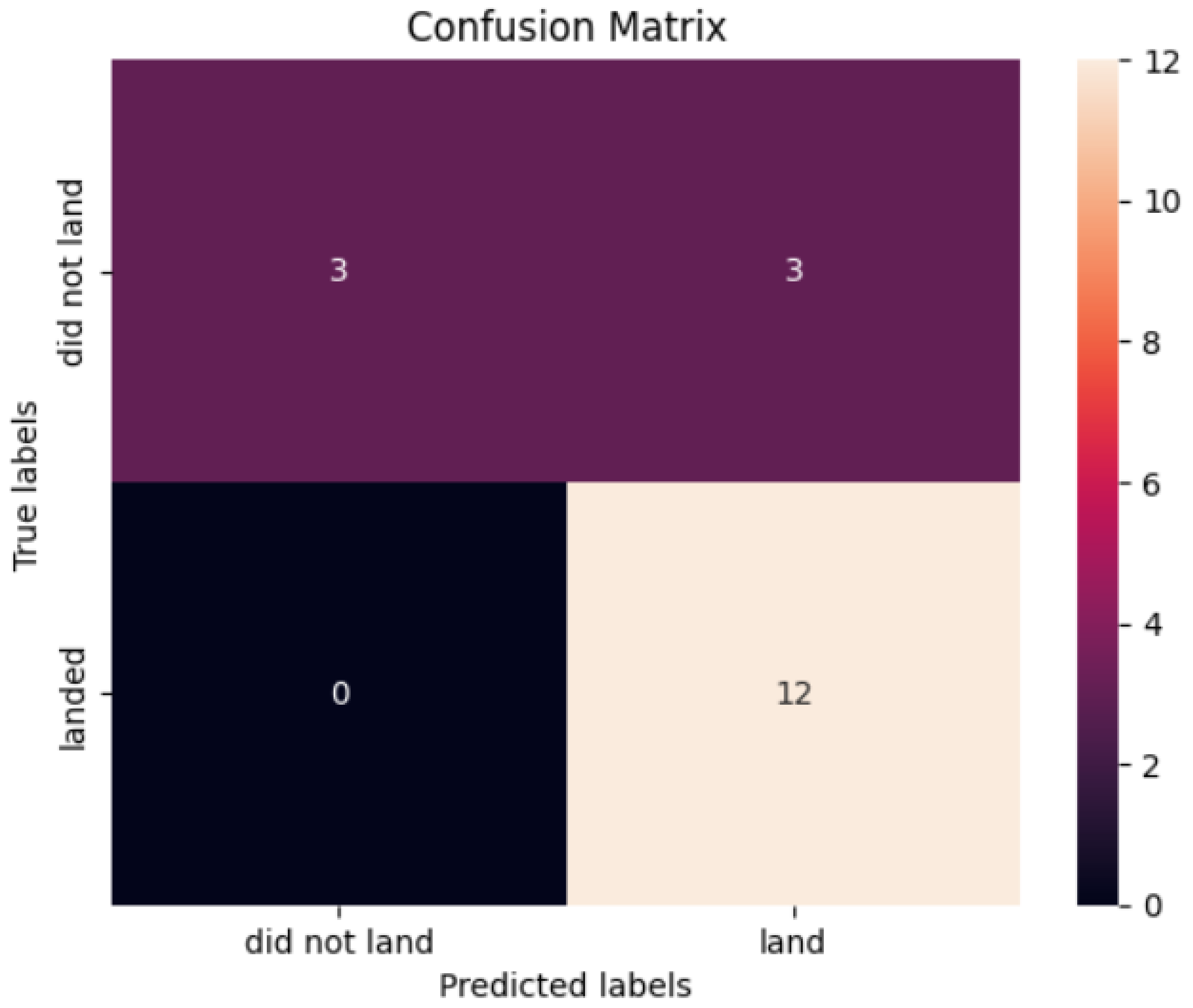
Predictive Analysis (Classification)

Classification Accuracy



- The highest accuracy has the decision Tree with 87.67 %

Confusion Matrix



- All the model have the same result

Conclusion

I formatted the data, created visualizations to compare different categories, and developed an interactive dashboard to explore the data. Additionally, I built a model that helps us better understand which categories are important and how they relate to each other, ultimately increasing the probability of a successful first-stage landing

Thank you!