

Winning Space Race with Data Science

Lai Jien Weng
12 October 2024



Outline

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

Executive Summary

Summary of Methodologies

- **Collect** data with SpaceX REST API and web scraping
- **Exploratory Data Analysis (EDA)** focusing on: payload mass, launch site, flight number, and year trend
- **Analyze** data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Statistical Modelling** to predict landing outcomes using logistic regression, SVM, decision tree and KNN

Summary of Results:

Exploratory Data Analysis:

- Launch success rates have steadily increased over time.
- KSC LC-39A is the most successful landing site.
- Orbits such as ES-L1, GEO, HEO, and SSO have a perfect success rate.
- Majority of launch sites are located near the equator and close to coastal areas.

Predictive Analytics:

- All models showed comparable performance on the test set, with the decision tree model achieving slightly better results.

Introduction

Background

SpaceX, a pioneer in the space industry, aims to make space travel accessible to all. Their achievements include delivering spacecraft to the International Space Station, launching a satellite constellation for global internet access, and conducting manned space missions. A key factor in their success is the cost-efficiency of their rocket launches, priced at \$62 million per launch, made possible by reusing the first stage of the Falcon 9 rocket.

In contrast, other providers, unable to reuse the first stage, charge over \$165 million per launch. By predicting the likelihood of a successful first-stage landing, we can estimate the launch cost. This can be achieved using publicly available data and machine learning models to forecast whether SpaceX or its competitors can reuse the first stage.



Section 1

Methodology

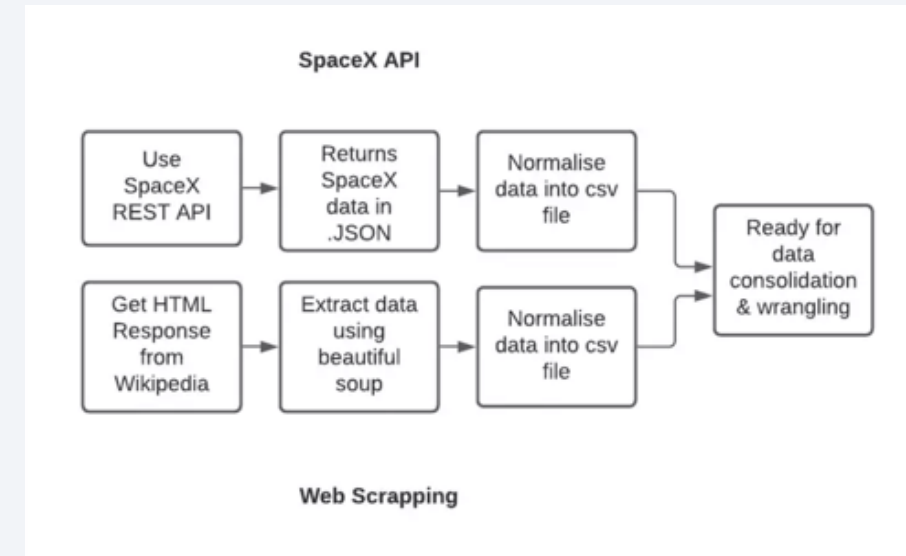
Methodology

Executive Summary

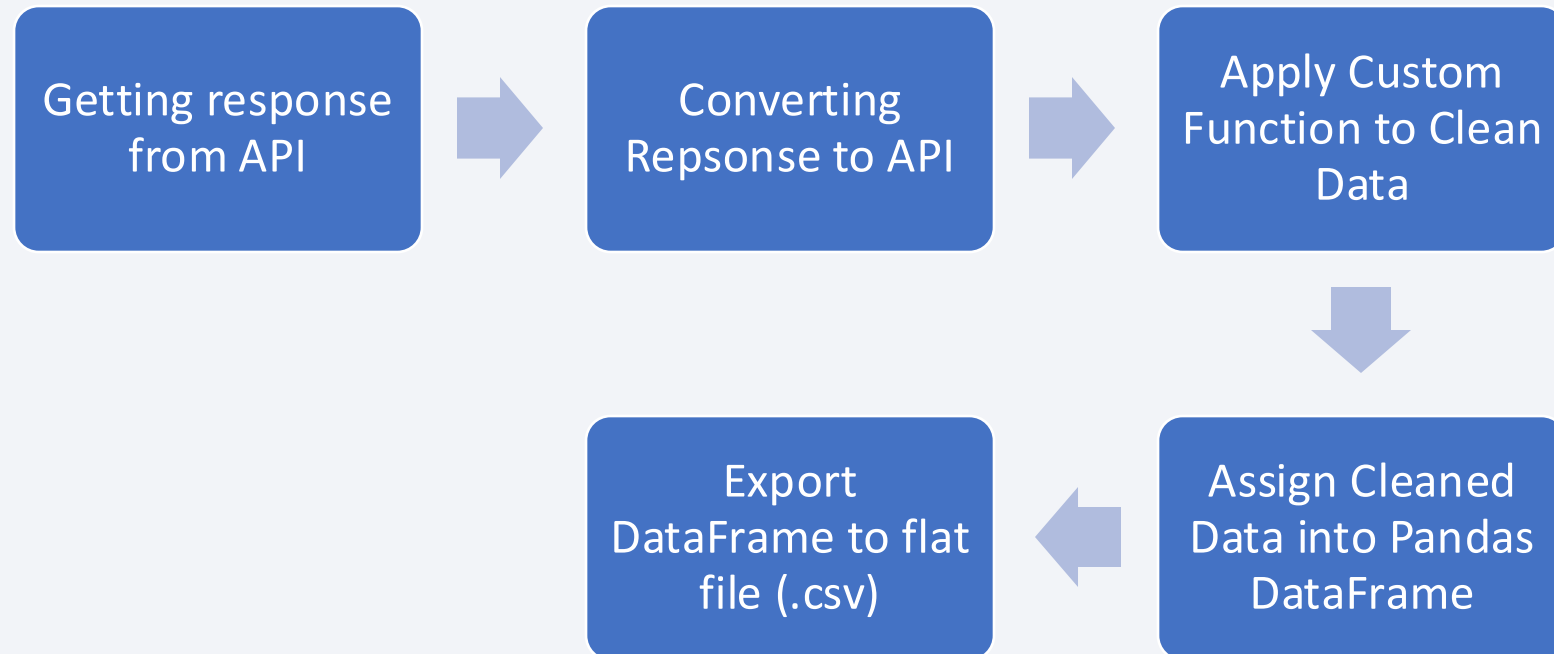
- **Data collection methodology:**
 - Data collected using SpaceX REST API and web scraping technique
- **Perform data wrangling**
 - Address missing values and map each data into binary outcomes
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
 - Build and tune KNN, logistic regression, decision tree and SVM with GridSearchCV

Data Collection

- The dataset was collected by:
 - SpaceX launch data was collected from the SpaceX REST API
 - It provides launches information, including:
 - Payload Mass
 - Launch Specifications
 - Landing Specifications
 - Landing Outcomes
 - The URL is: api.spacexdata.com/v4/
 - Data of Falcon 9 Launch is obtained through web scraping with BeautifulSoup on Wikipedia

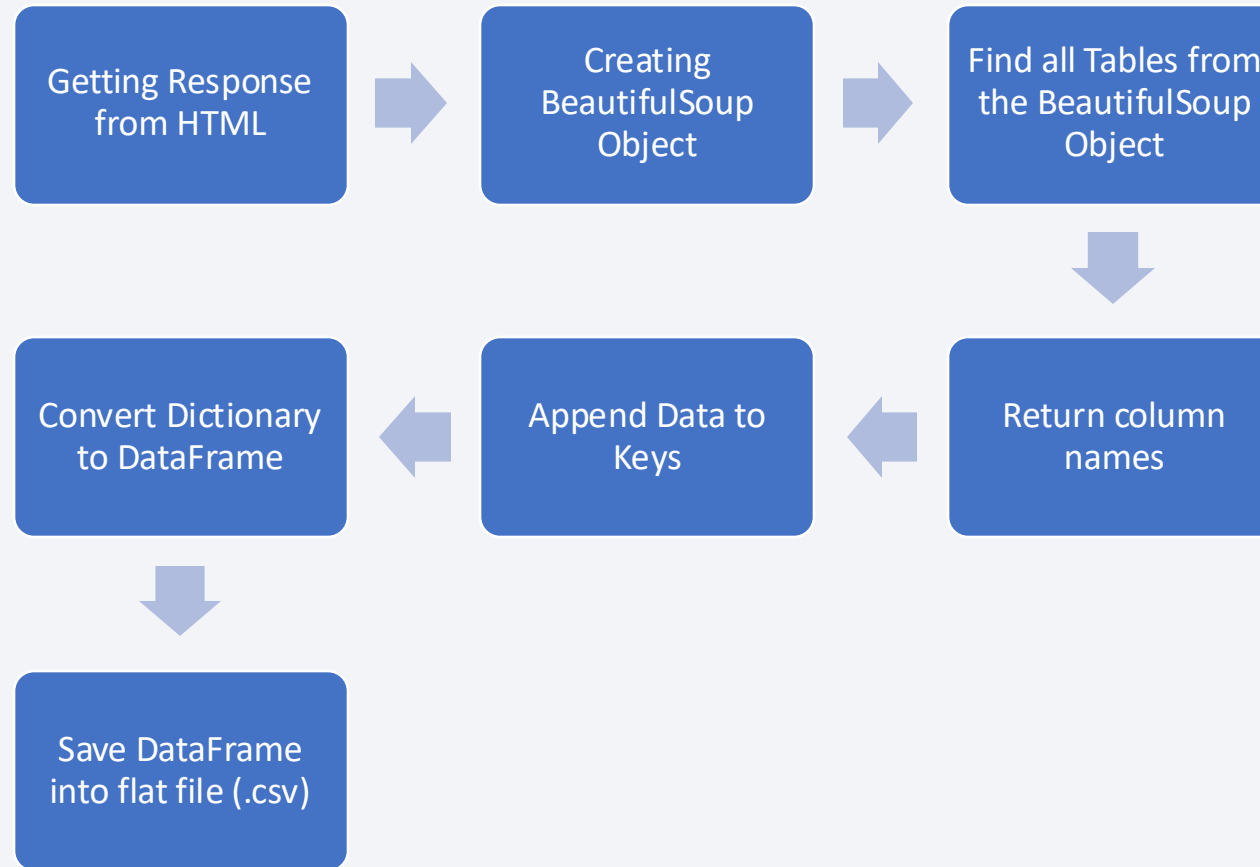


Data Collection – SpaceX API



The source file of Data Collection can be accessed through [GitHub](#)

Data Collection - Scraping



The source file of Data Collection can be accessed through [GitHub](#)

Data Wrangling

- **In this section, we performed:**
 - Exploratory Data Analysis
 - Determine Training Labels
- **The targeted variable is mapped into 0 & 1, which:**
 - 0: Unsuccessful landing
 - 1: Successful landing
- **We found that the success rate of landing is 66.67% from EDA**

The source file of Data Collection can be accessed through [GitHub](#)

EDA with Data Visualization

- Data were visualized through several visualization techniques:
 - Scatterplot
 - Bar Plot
 - Count Plot
 - Line Plot
- Matplotlib, Seaborn and NumPy libraries were utilized to visualize the data.

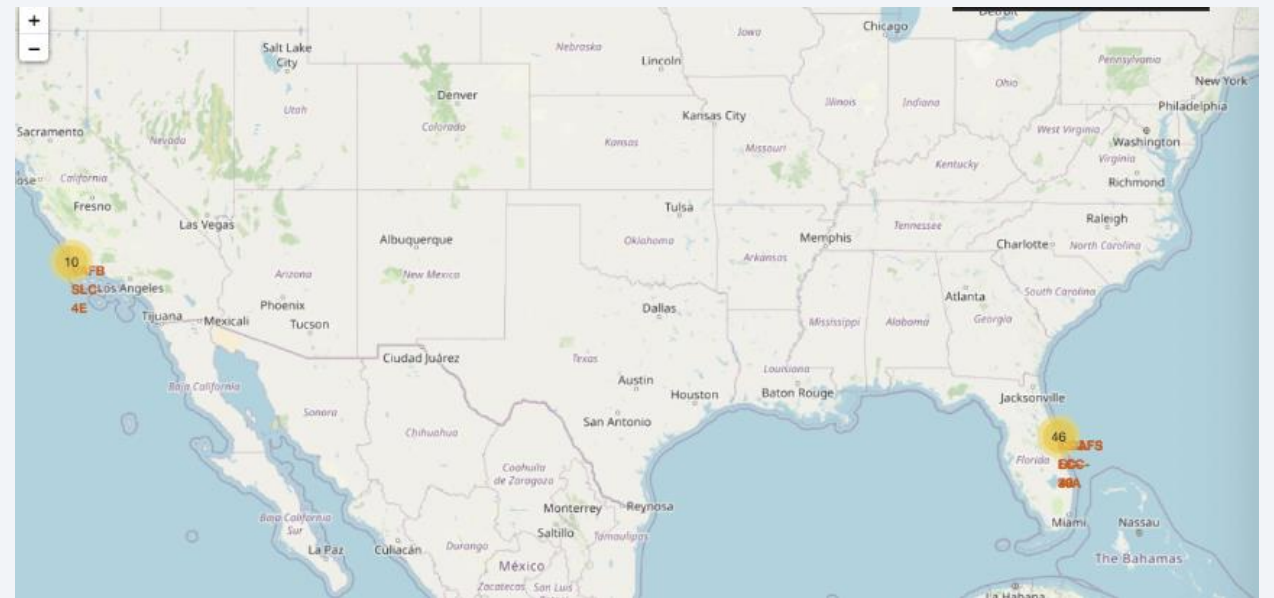
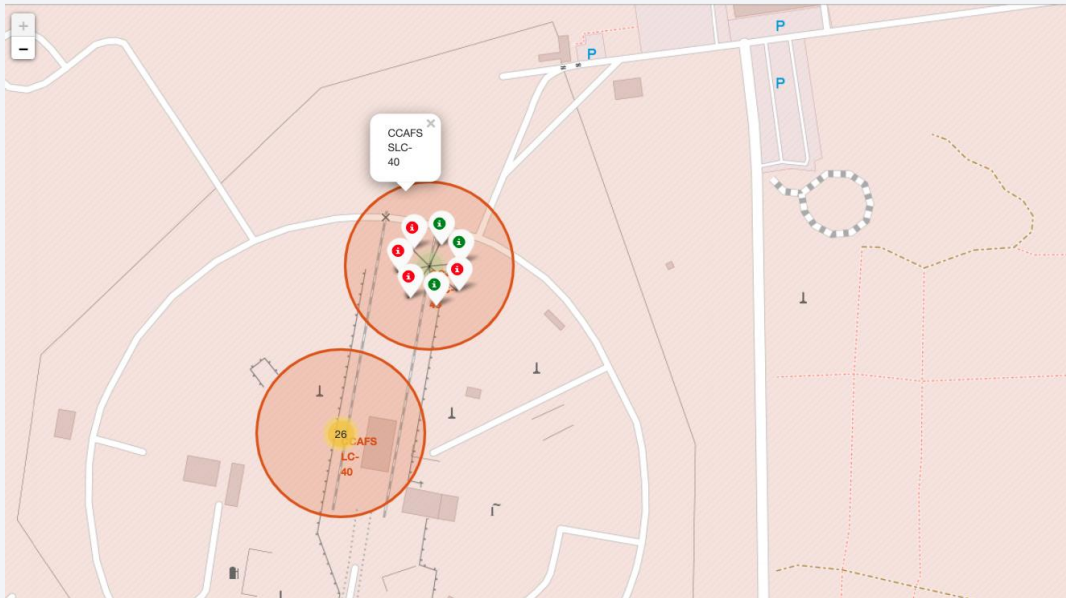
The source file of Data Collection can be accessed through [GitHub](#)

EDA with SQL

- **The following analysis is performed in SQL queries:**
 - Names of unique launch sites in space mission
 - First 5 records where the launch sites begin with "CCA"
 - Total payload mass carried by boosters launched by NASA (CRS)
 - Average payload mass carried by booster version F9 v1.1
 - The day when the first successful landing in ground pad was achieved
 - The boosters which have success in drone ship and have payload mass between 4,000 to 6,000
 - Total number of successful and failed mission outcomes
 - Boosters versions which have carried the maximum payload mass
 - Records of failed landing in drone ship with respective information in 2015
 - Rank the count of landing outcomes between 2010-06-04 to 2017-03-20

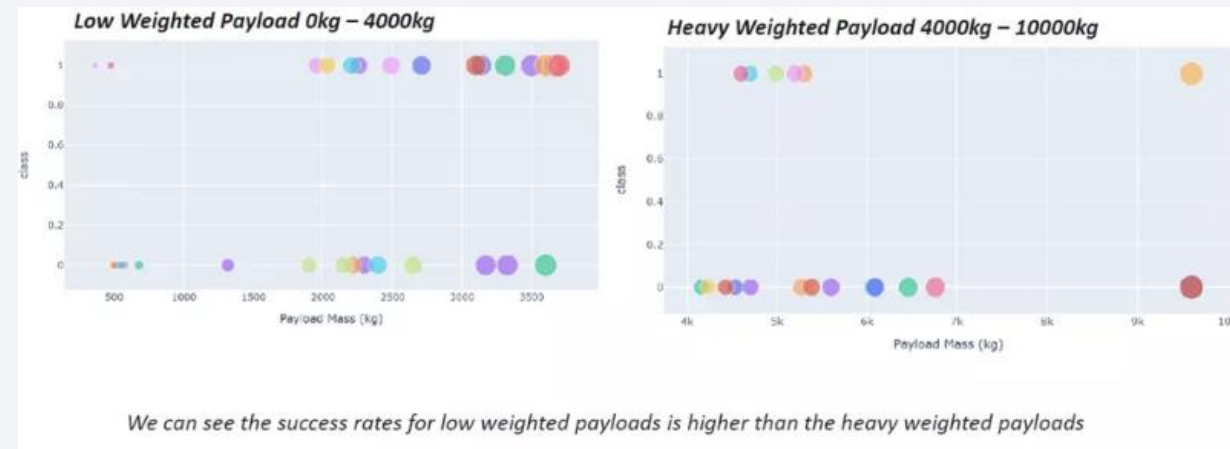
The source file of Data Collection can be accessed through [GitHub](#)

Build an Interactive Map with Folium



The source file of Data Collection can be accessed through [GitHub](#)

Build a Dashboard with Plotly Dash



The source file of Data Collection can be accessed through [GitHub](#)

Predictive Analysis (Classification)

- **Process of Modelling:**
 - Standardized data
 - Train test splitting into 80% for training set & 20% for testing set
 - The following model are built:
 - K-Nearest Neighbors (KNN)
 - Logistic Regression
 - Decision Tree
 - Support Vector Machine (SVM)
 - Each model was hyperparameter-tuned and the best model is selected
 - Each model is evaluated with accuracy and confusion matrix
 - The model with highest "out-of-training accuracy" is selected

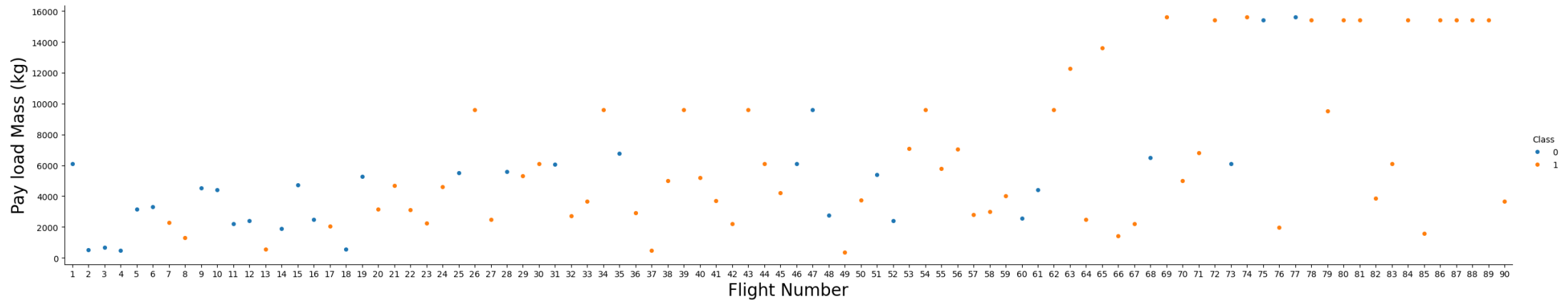
The source file of Data Collection can be accessed through [GitHub](#)

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

Section 2

Insights drawn from EDA

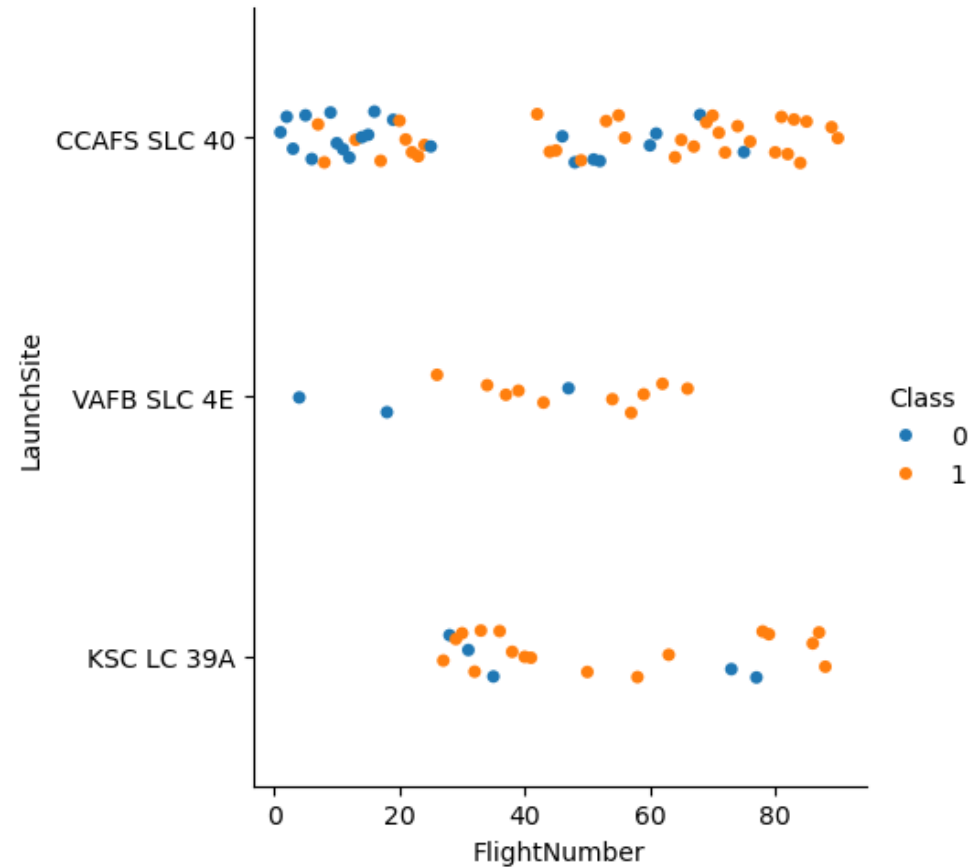
Flight Number vs. Payload Mass



1. As flight number increase, first stage is more likely to land successfully.
2. As payload mass increase, the less likely the first stage will return

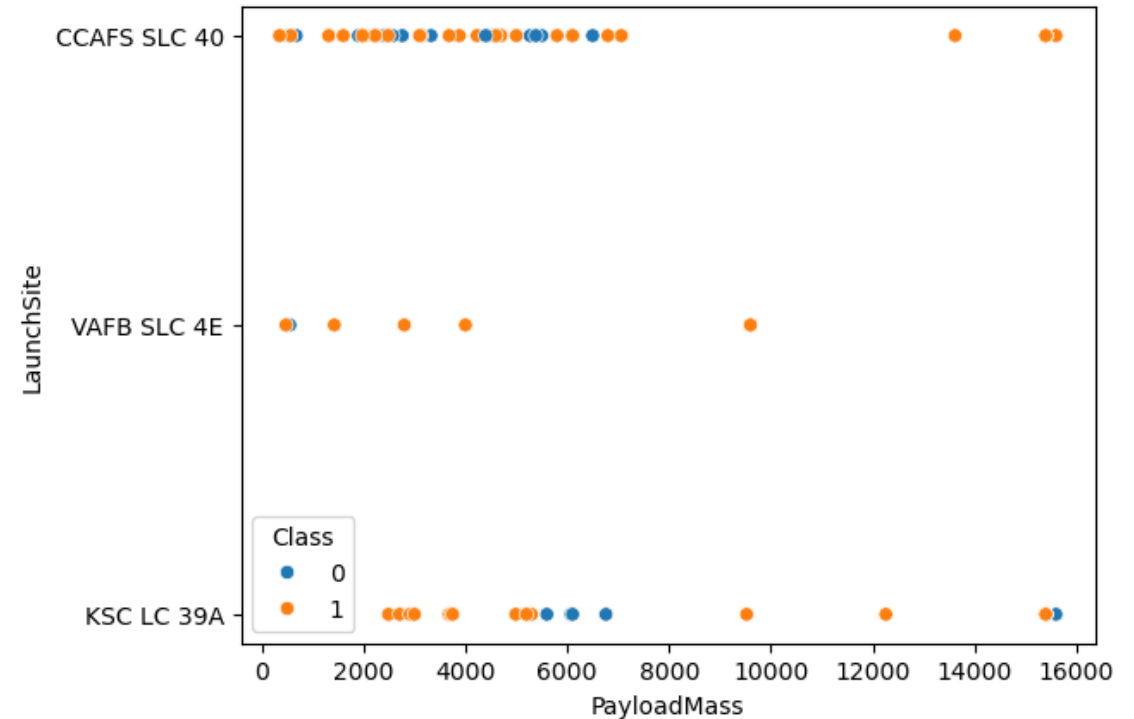
Flight Number vs. Launch Site

1. High proportion of KSC LC 39A land successfully.
2. As Flight Number increases, it more likely to land successfully



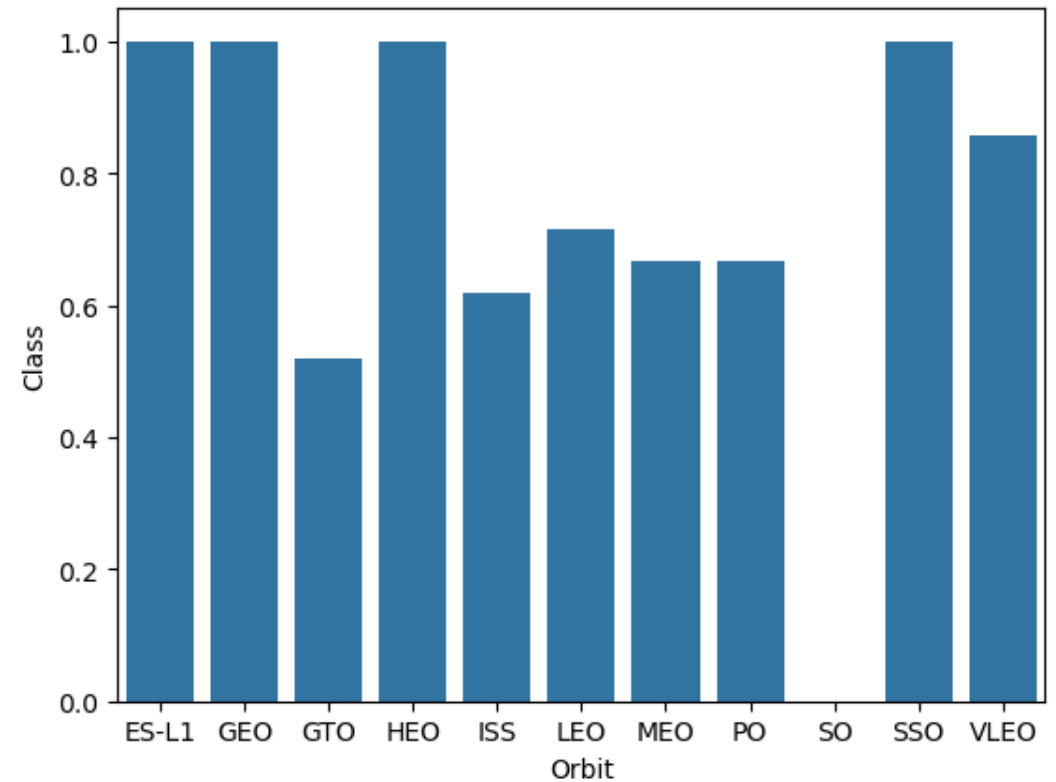
Payload Mass vs. Launch Site

1. There are no rockets launch for heavy payload mass at VAFB-SLC 4E
2. Note that heavy payload flights is consider greater than 10,000kg



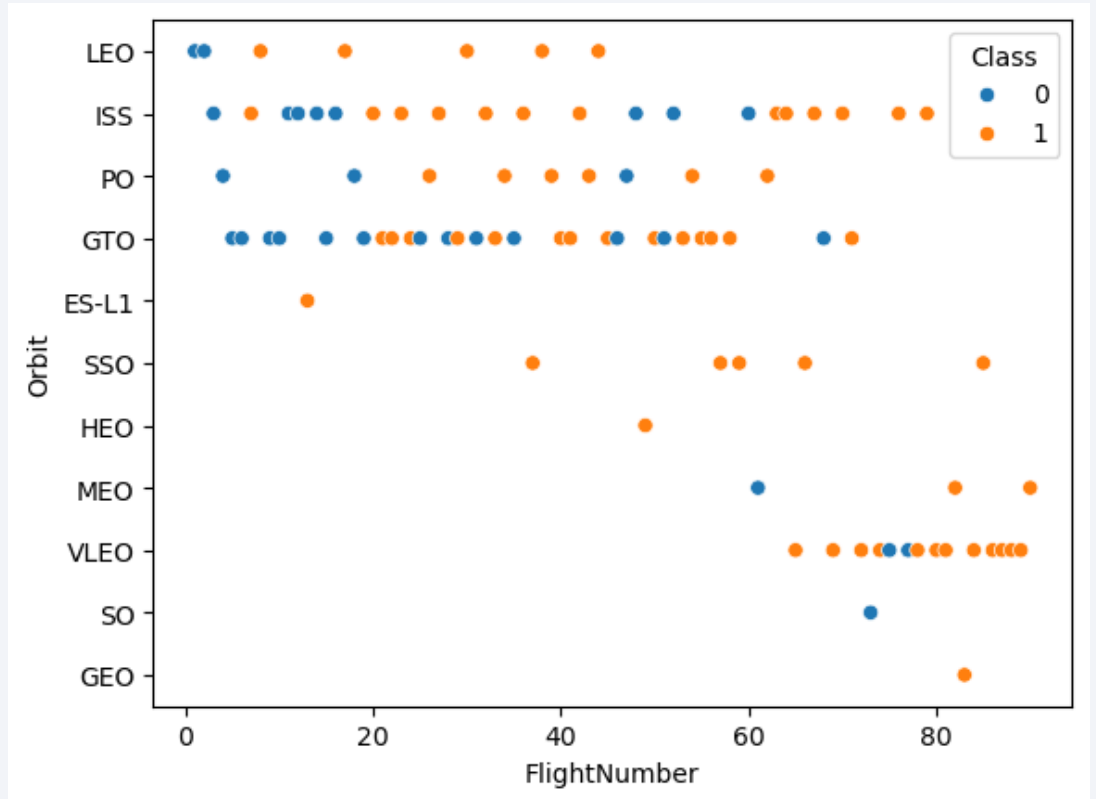
Success rate by Orbit Type

1. These Orbits have 100% success rate for landing:
 1. ES-L1
 2. GEO
 3. HEO
 4. SSO
2. All landing from SO Orbit failed.



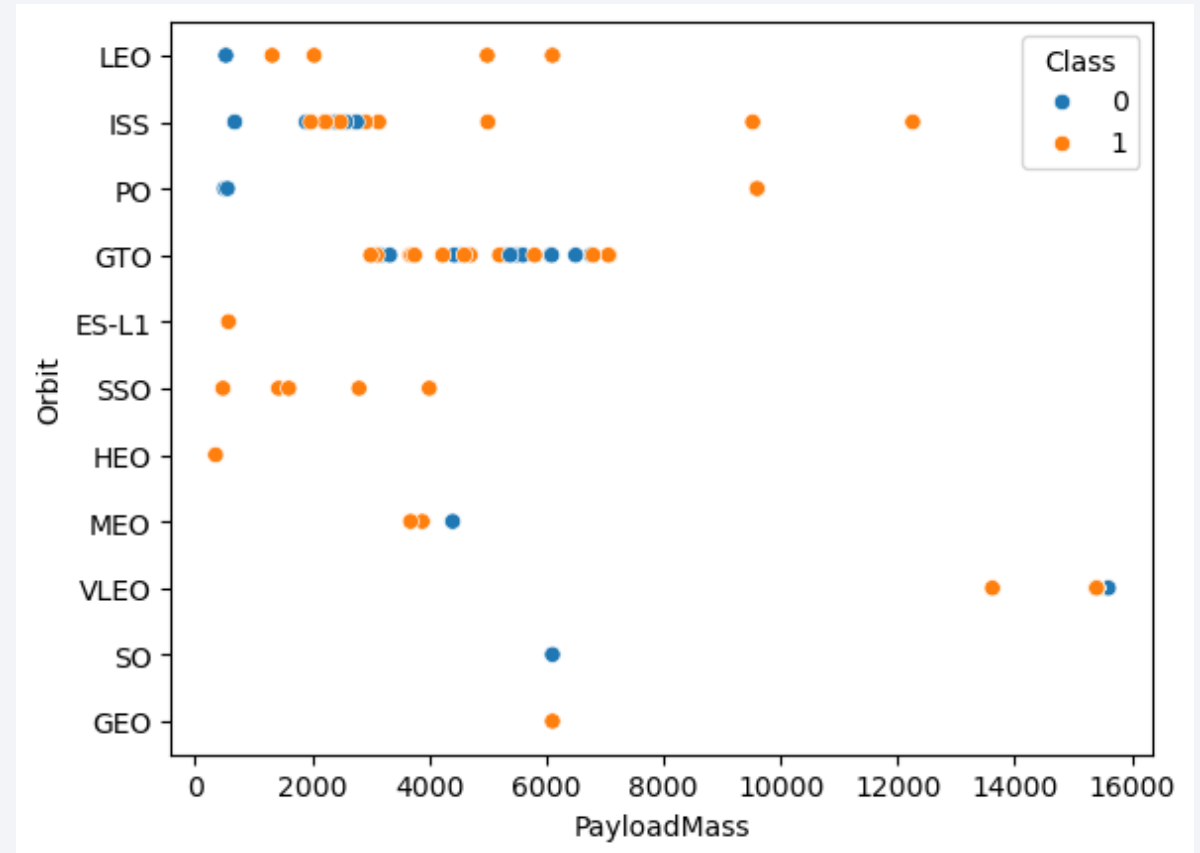
Flight Number vs. Orbit Type

- LEO Orbit appears to be related to the number of flights of having successful landing
- There's seems to have no relationship between flight number when in GTO orbit.



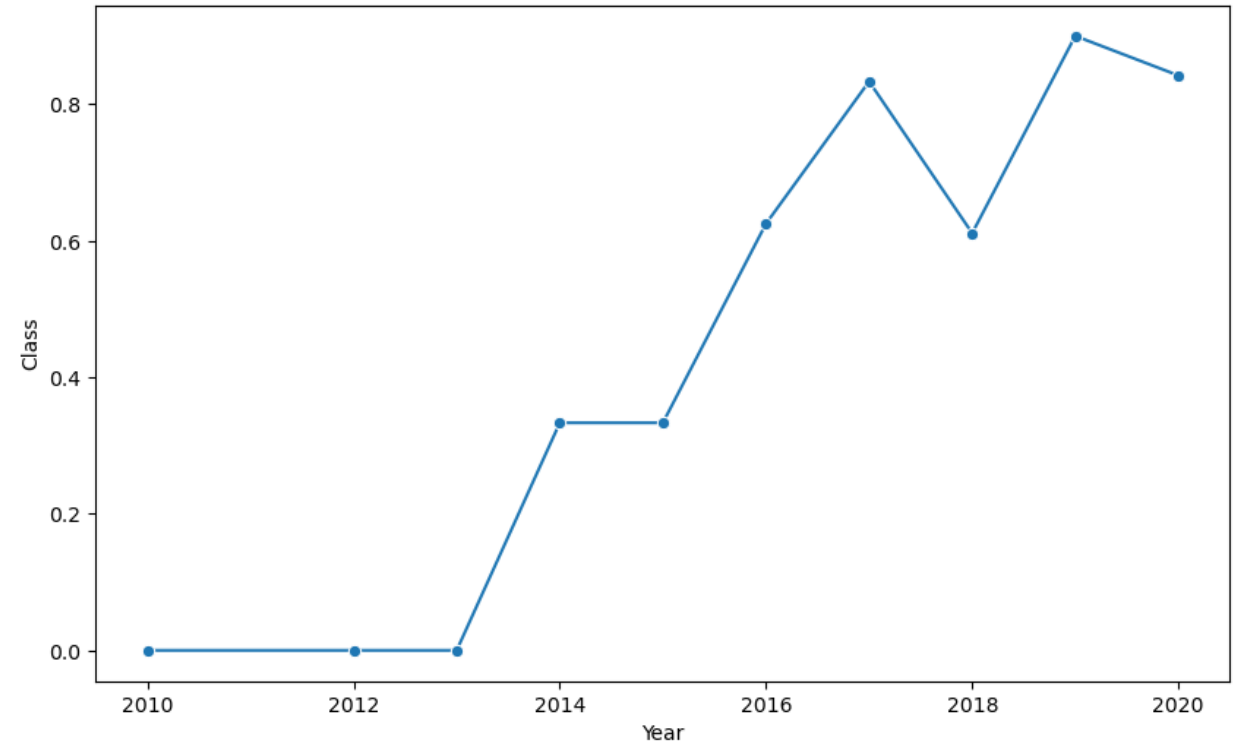
Payload vs. Orbit Type

- With heavy payloads, the successful landing are higher for Polar, LEO and ISS Orbit.
- GTO cannot distinguish as the successful and unsuccessful landing are both there.

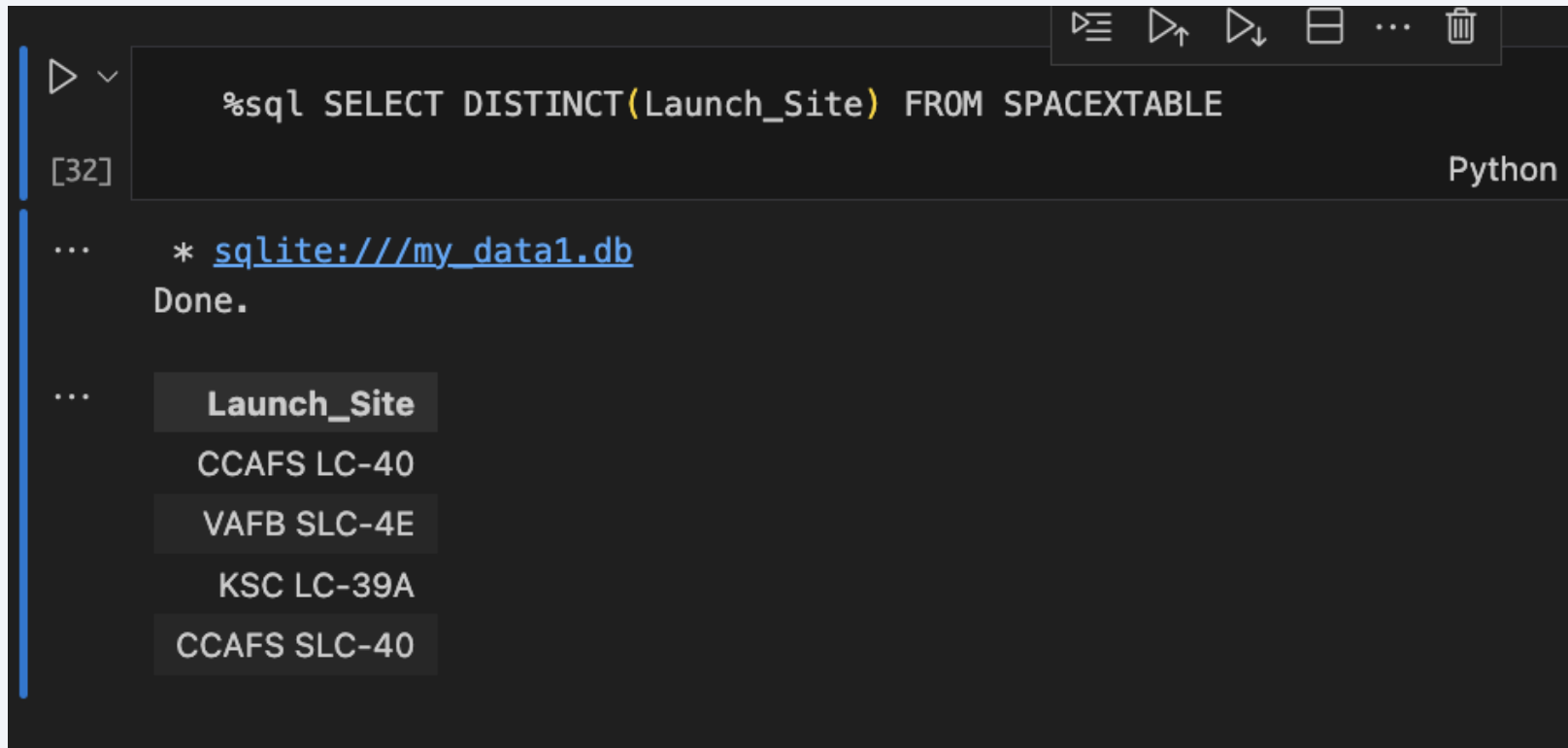


Yearly Trend of Success Launch

1. Success rate of landing kept increasing from since 2013 until 2017.
2. The success rate dropped in 2018 but bounced higher than 2017 in 2019.



All Launch Site Names



A screenshot of a Jupyter Notebook interface. The top toolbar contains icons for running, stepping through, and other code execution functions. The main area shows a code cell with the following content:

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

Below the code cell, the output is displayed. It starts with a prompt character followed by the connection string and the word "Done.", then a table of results.

```
[32] Python
```

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

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

Launch Site Names Begin with 'CCA'

```

%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
[33]
Python

... * 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 |

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS_KG_) AS Total_Payload_Mass FROM SPACEXTABL
```

Python

* [sqlite:///my_data1.db](#)
Done.

| Total_Payload_Mass |
|--------------------|
| 45596 |

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPAC
```

Python

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

Done.

| Average_Payload_Mass |
|----------------------|
|----------------------|

| |
|--------|
| 2928.4 |
|--------|

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) AS First_Successful_Landing FROM SPACEXTABLE W
```

9]

Python

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

Done.

| First_Successful_Landing |
|--------------------------|
|--------------------------|

| |
|------------|
| 2015-12-22 |
|------------|

Successful Drone Ship Landing with Payload between 4000 and 6000

```
▶ %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome =  
[41] Python  
... * sqlite:///my\_data1.db  
Done.  
... 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 SELECT Landing_Outcome, COUNT(*) AS Total_Count FROM SPACEXTABL
```

Python

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

| Landing_Outcome | Total_Count |
|------------------------|-------------|
| Controlled (ocean) | 5 |
| Failure | 3 |
| Failure (drone ship) | 5 |
| Failure (parachute) | 2 |
| No attempt | 21 |
| No attempt | 1 |
| Precluded (drone ship) | 1 |
| Success | 38 |
| Success (drone ship) | 14 |
| Success (ground pad) | 9 |
| Uncontrolled (ocean) | 2 |

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS_KG_
Python

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

| 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 SELECT substr(Date ,6, 2) AS Month, Landing_Outcome, Booster_Ve
```

Python

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

Done.

| 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

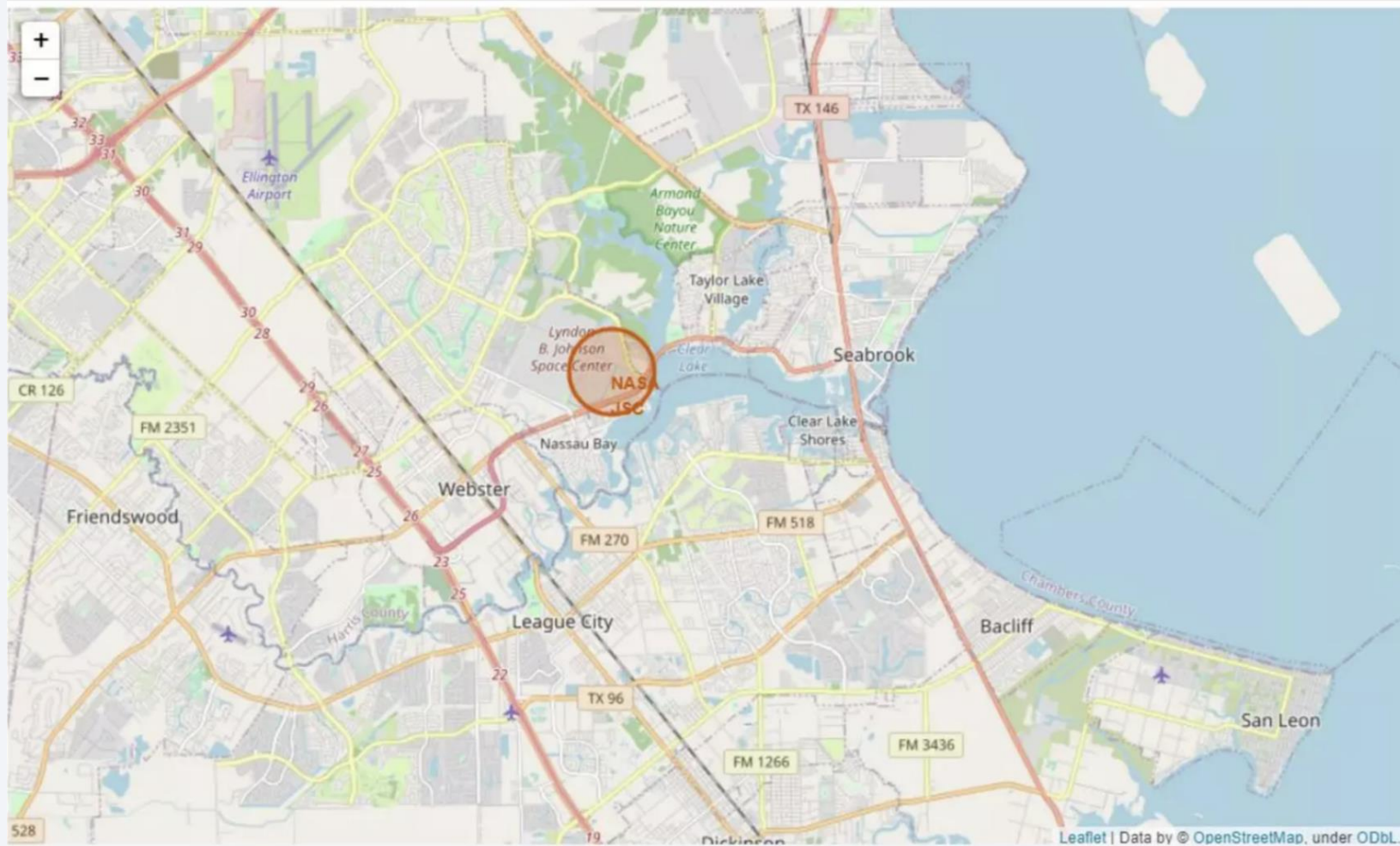
| Landing_Outcome | Total_Outcome |
|------------------------|---------------|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark, with numerous bright yellow and orange lights representing cities and urban areas. The horizon of the Earth is visible as a thin, curved line separating the dark surface from the deep blue of space.

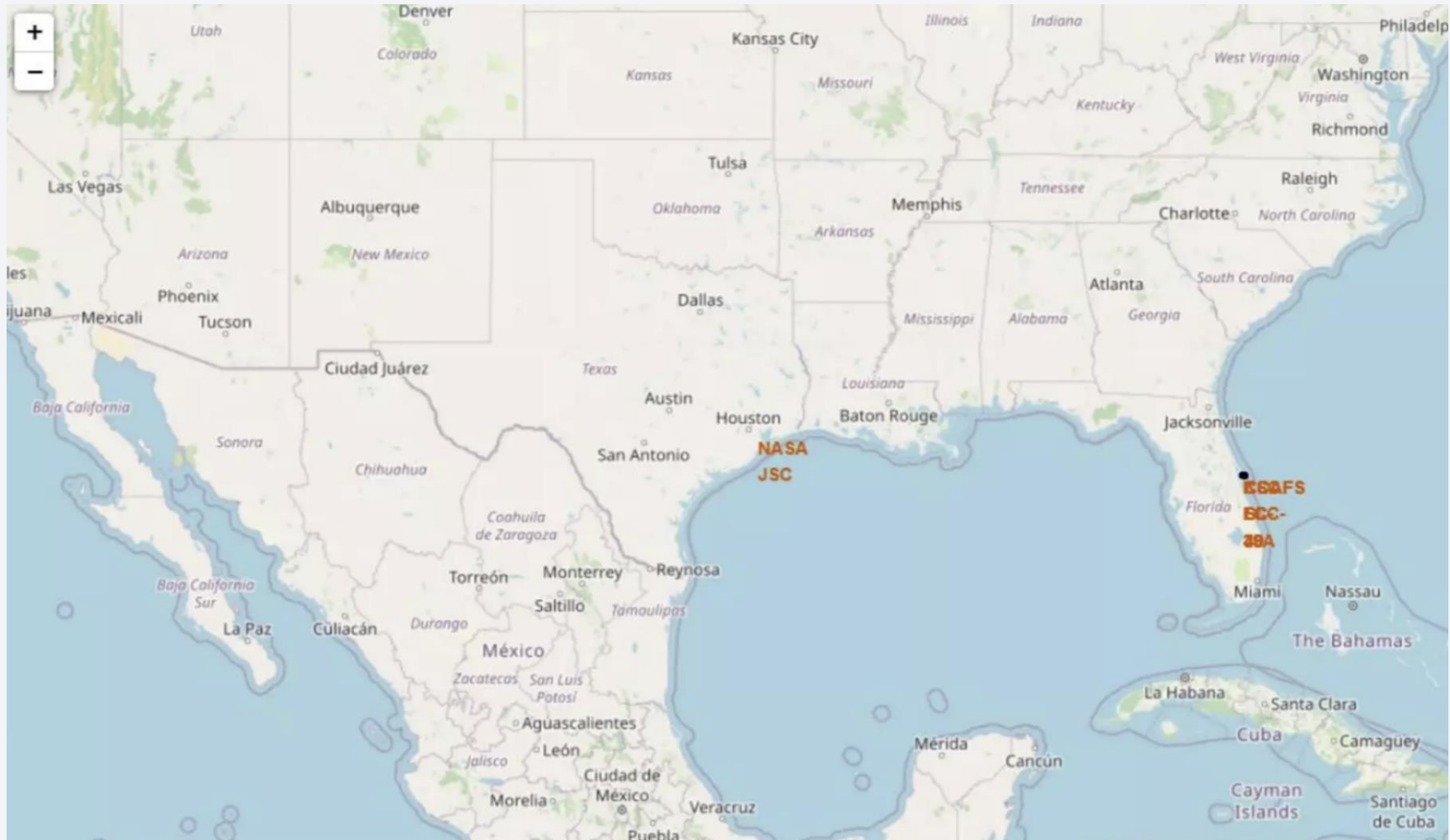
Section 3

Launch Sites Proximities Analysis

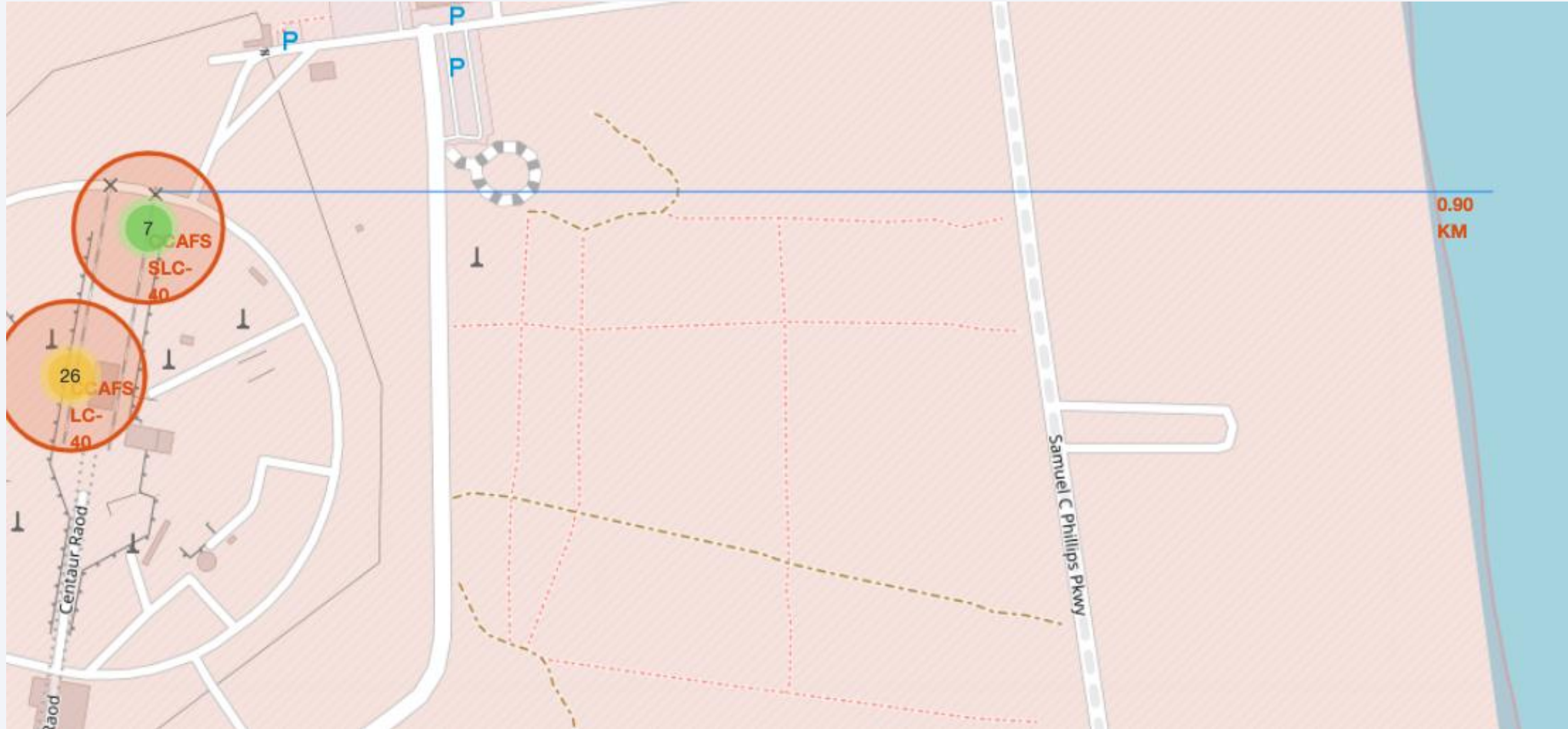
All launch sites marked on a map



Successful / Failed Launches marked on map



Launches site distance from nearest coastline

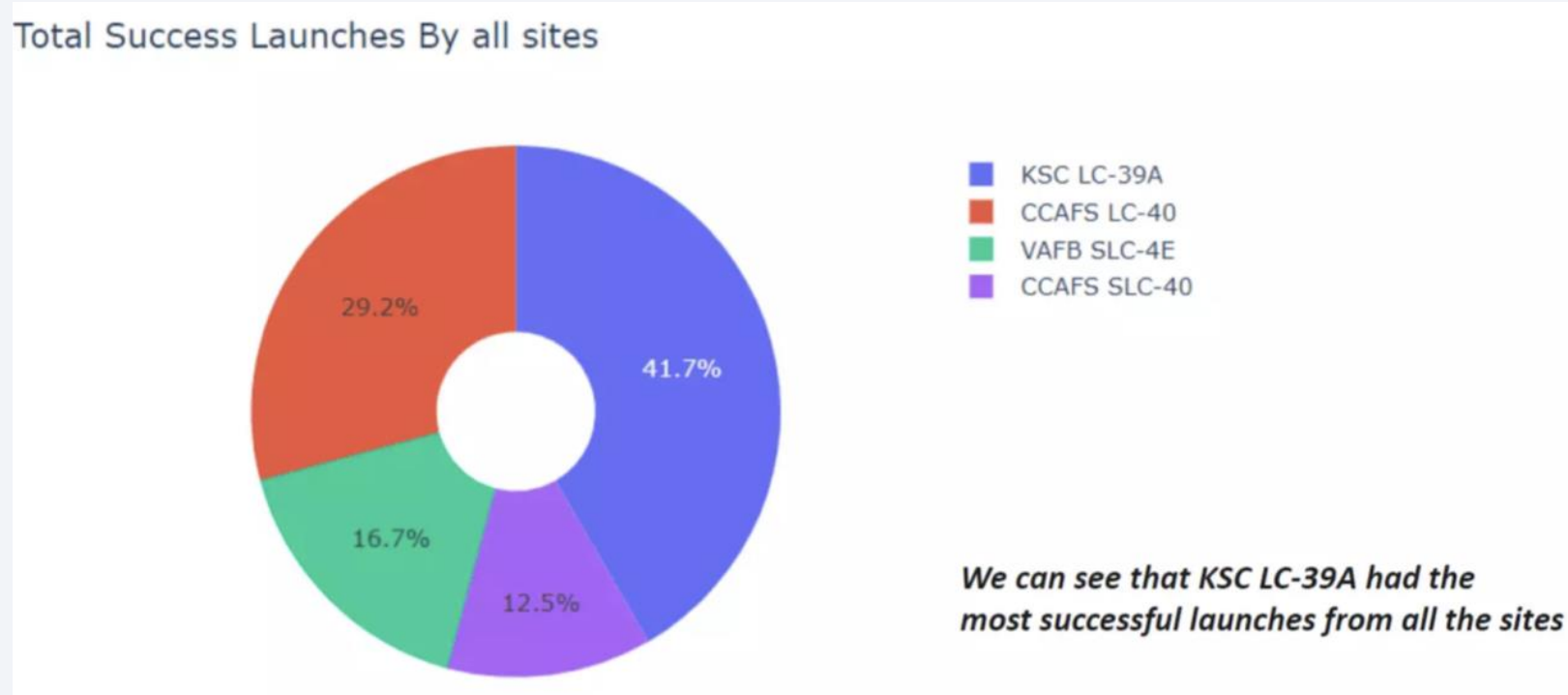




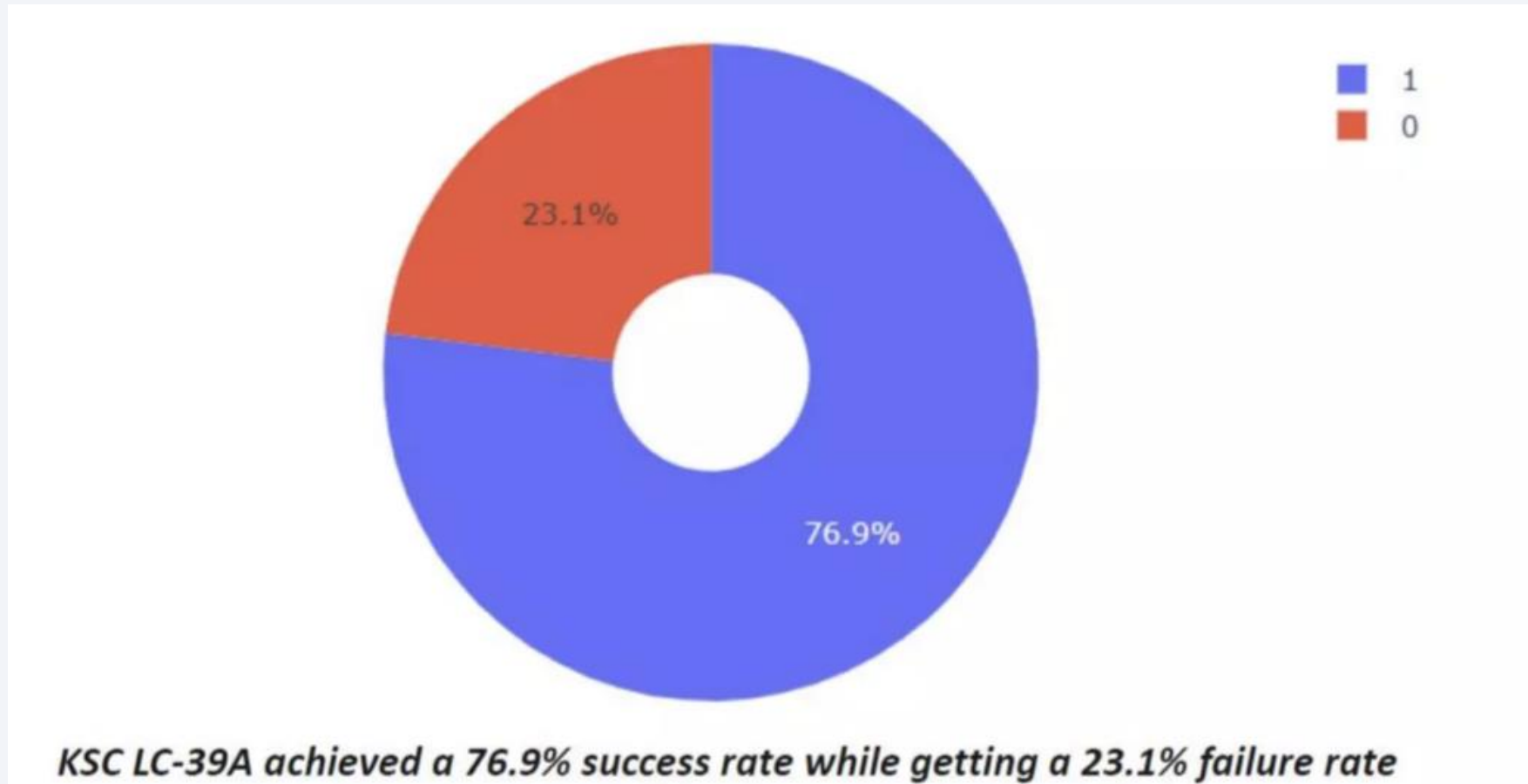
Section 4

Build a Dashboard with Plotly Dash

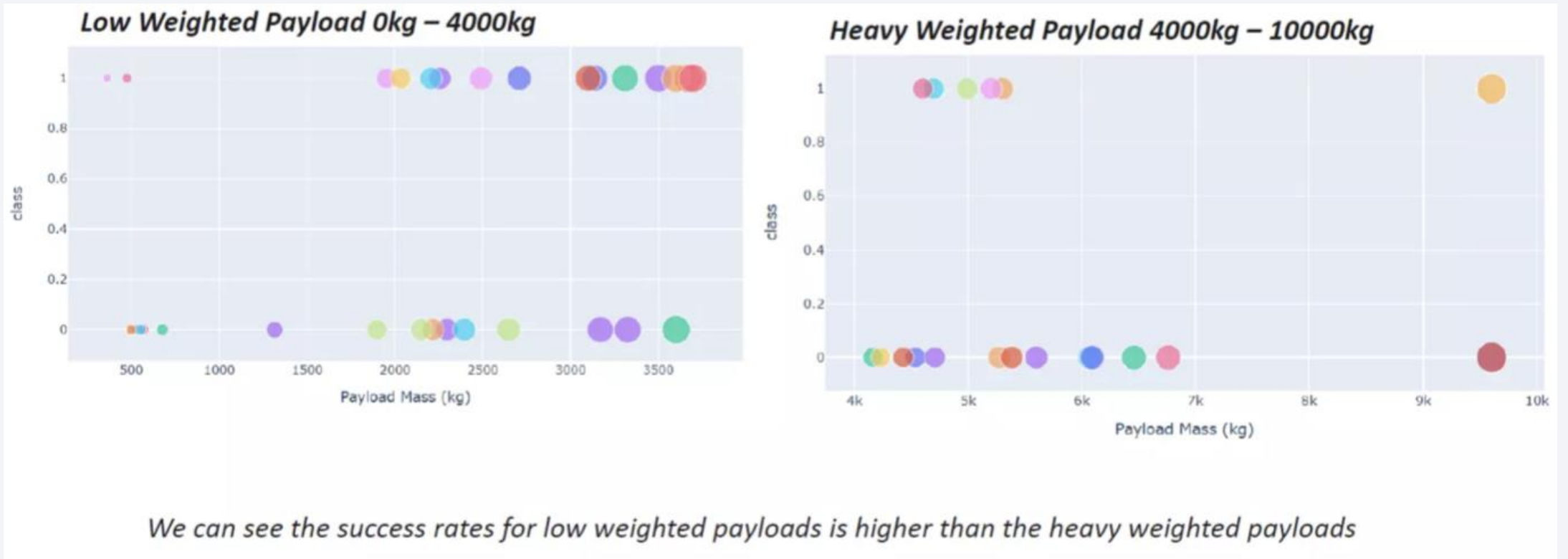
Total Success Launches by All Sites



Success rate by KSC LC-39A



Payload Mass vs. Launch Outcomes

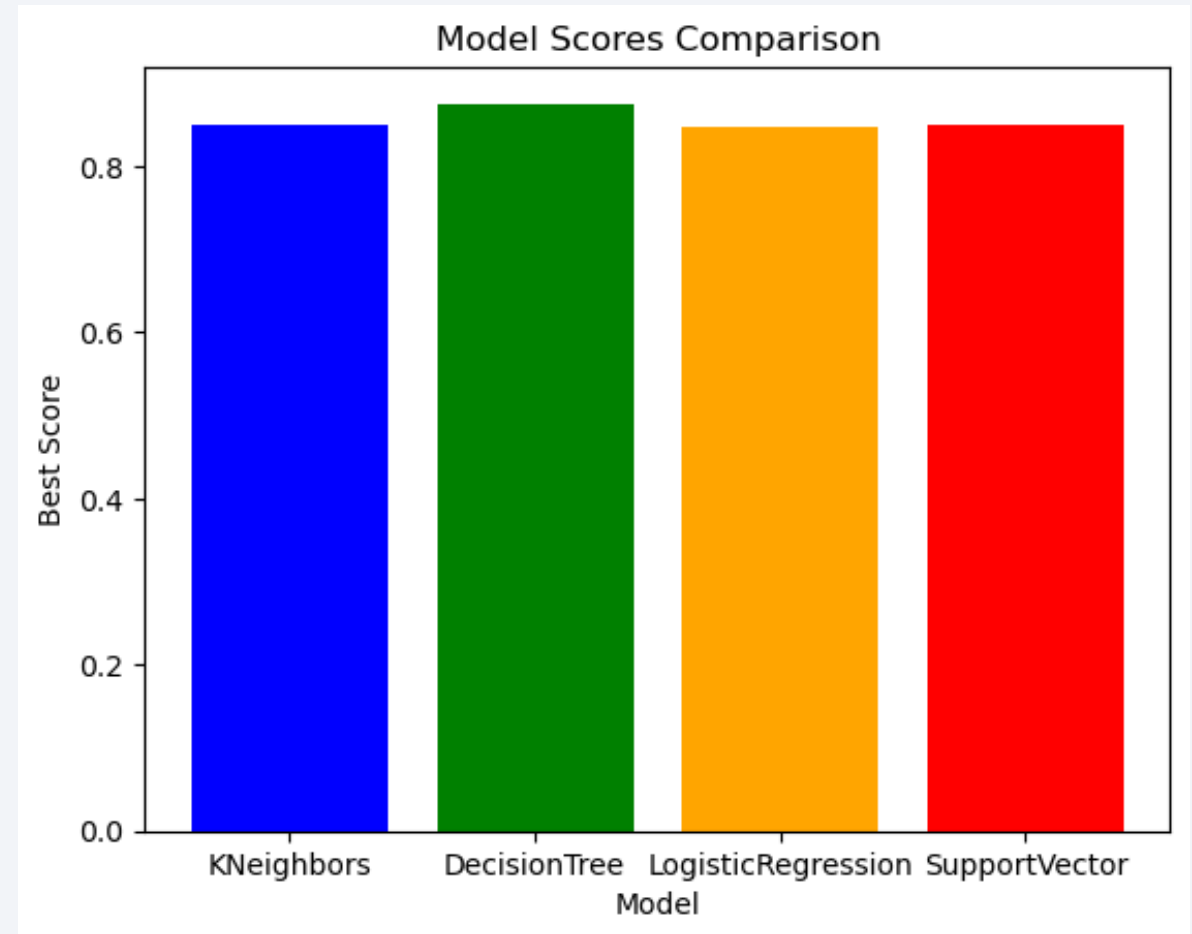


Section 5

Predictive Analysis (Classification)

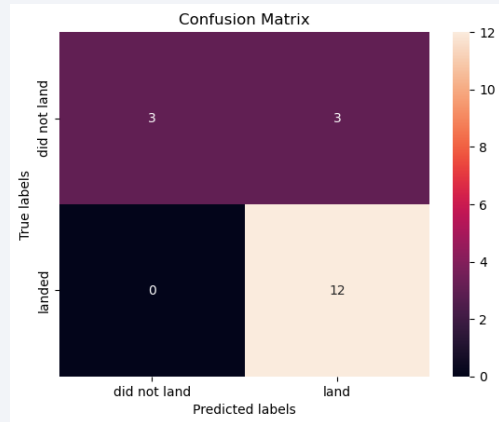
Classification Accuracy

- Decision Tree have the highest Classification Accuracy among KNN, Decision Tree, Logistic Regression and SVM.

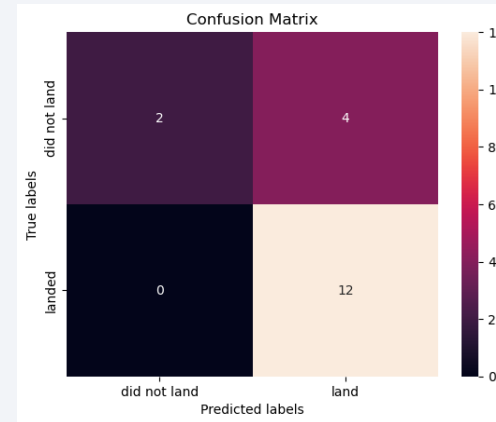


Confusion Matrix

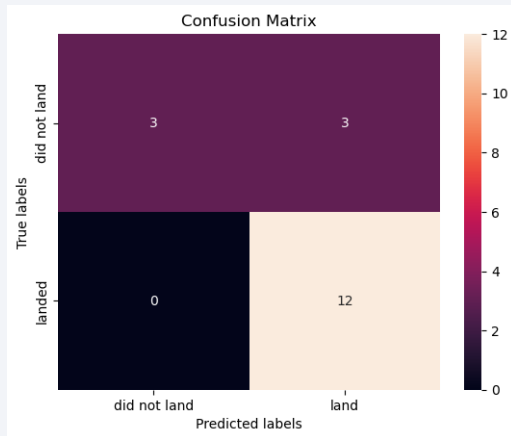
Logistic Regression



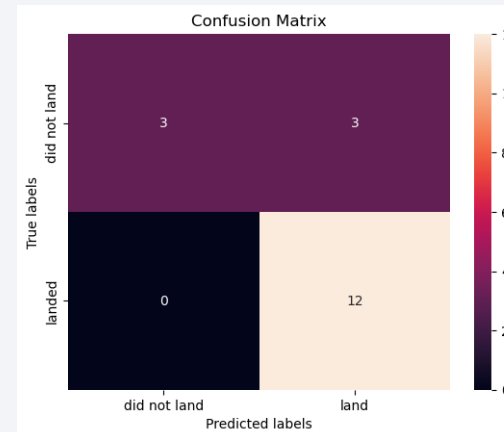
Decision Tree



SVM



KNN



Conclusions

- Model Performance: Decision Tree slightly outperform other models.
- All launch sites are close to the coast
- KSC LC-39A has the highest success rate among launch sites
- ES-L!, GEO, HEO & SSO Orbits has 100% success rate
- Most of the launch sites are near to the equator. It helps launches with additional natural boost, which save cost of fuel and boosters.

Thank you!

