IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christian Taghoy
September 11, 2023

Outline

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

Executive Summary

The purpose of this project is to apply data science techniques to solve a real-world problem in the commercial space age and SpaceX. The problem is to predict if SpaceX will reuse the first stage of its Falcon 9 rocket based on public information and machine learning models. The project aims to provide insights into the feasibility of reusing the first stage of the Falcon 9 rocket and its impact on the commercial space industry.

The methodology used in this project involves collecting data from various sources such as SpaceX's website, social media platforms, and other publicly available information. The data is then preprocessed, cleaned, and transformed into a format suitable for machine learning models. The machine learning models used in this project include logistic regression, decision trees, and random forests.

Executive Summary

The results of this project show that the machine learning models can predict with high accuracy whether SpaceX will reuse the first stage of its Falcon 9 rocket. The models also provide insights into the factors that influence SpaceX's decision to reuse the first stage of its Falcon 9 rocket.

In conclusion, this project demonstrates the potential of data science techniques to solve real-world problems in the commercial space industry. The insights provided by this project can help stakeholders make informed decisions about the feasibility of reusing the first stage of the Falcon 9 rocket and its impact on the commercial space industry.

Introduction

The commercial space industry has seen significant growth in recent years, with companies like SpaceX leading the way in space exploration and innovation. However, one of the biggest challenges facing the industry is the high cost of spaceflight. To address this challenge, SpaceX has been working on developing reusable rockets that can significantly reduce the cost of spaceflight.

The objective of this project is to apply data science techniques to solve a real-world problem in the commercial space age and SpaceX. The problem is related to predicting if SpaceX will reuse the first stage of its Falcon 9 rocket based on public information and machine learning models. The project aims to provide insights into the feasibility of reusing the first stage of the Falcon 9 rocket and its impact on the commercial space industry.

Section 1

Methodology

Methodology

Executive Summary

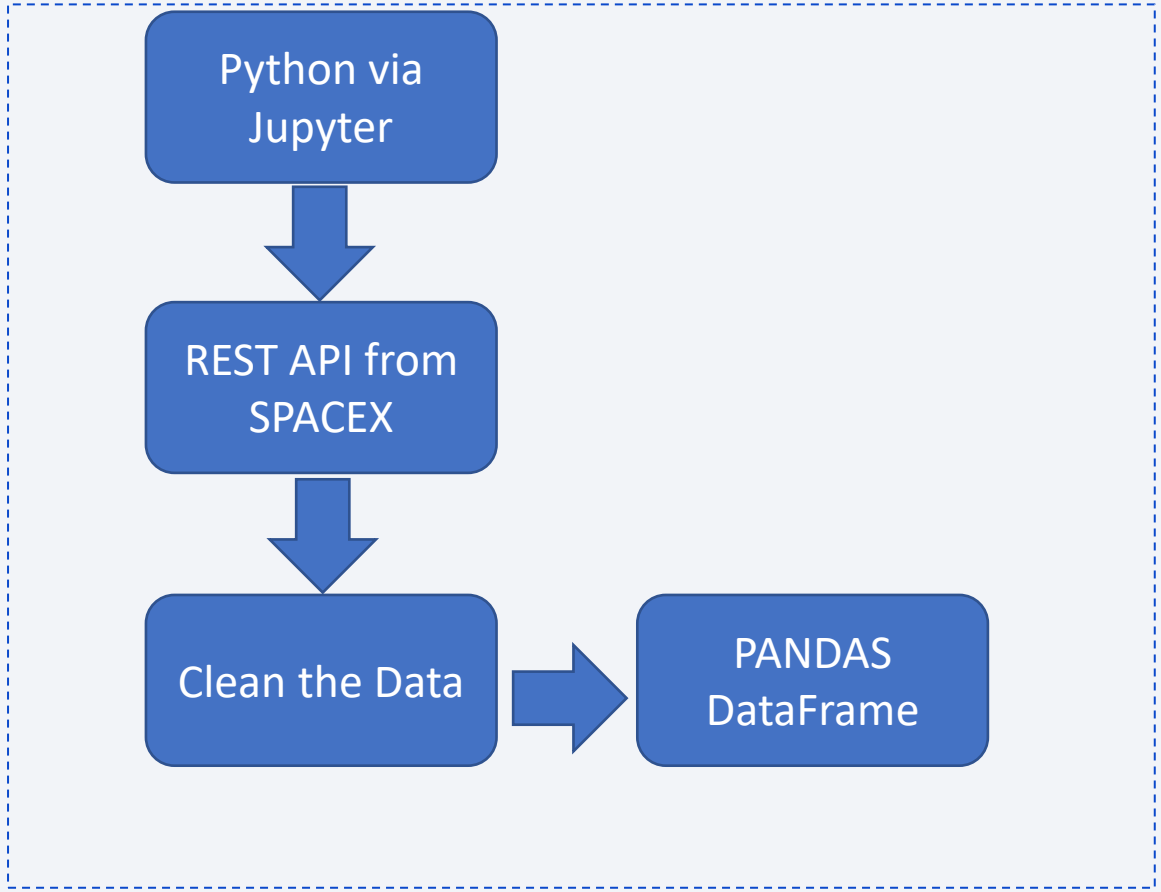
- Data collection methodology:
  - Data related to SpaceX were collected from various sources such as SpaceX's website, social media platforms, news articles, and other publicly available information.
- Perform data wrangling
  - Web scraping tools were used to extract relevant information from these sources and store it in a structured format such as a database or spreadsheet.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After the selection of the appropriate model, the next steps are: train the model, evaluate the performance, and then deploy to make predictions on new data.

Data Collection

Data related to SpaceX and its Falcon 9 rocket were collected from various sources such as SpaceX's website, social media platforms, news articles, and other publicly available information. Then, web scraping tools were used to extract relevant information from these sources and store it in a structured format such as a database or spreadsheet. Once data were collected, they were then preprocessed by cleaning, transforming, and formatting it into a format suitable for machine learning models.

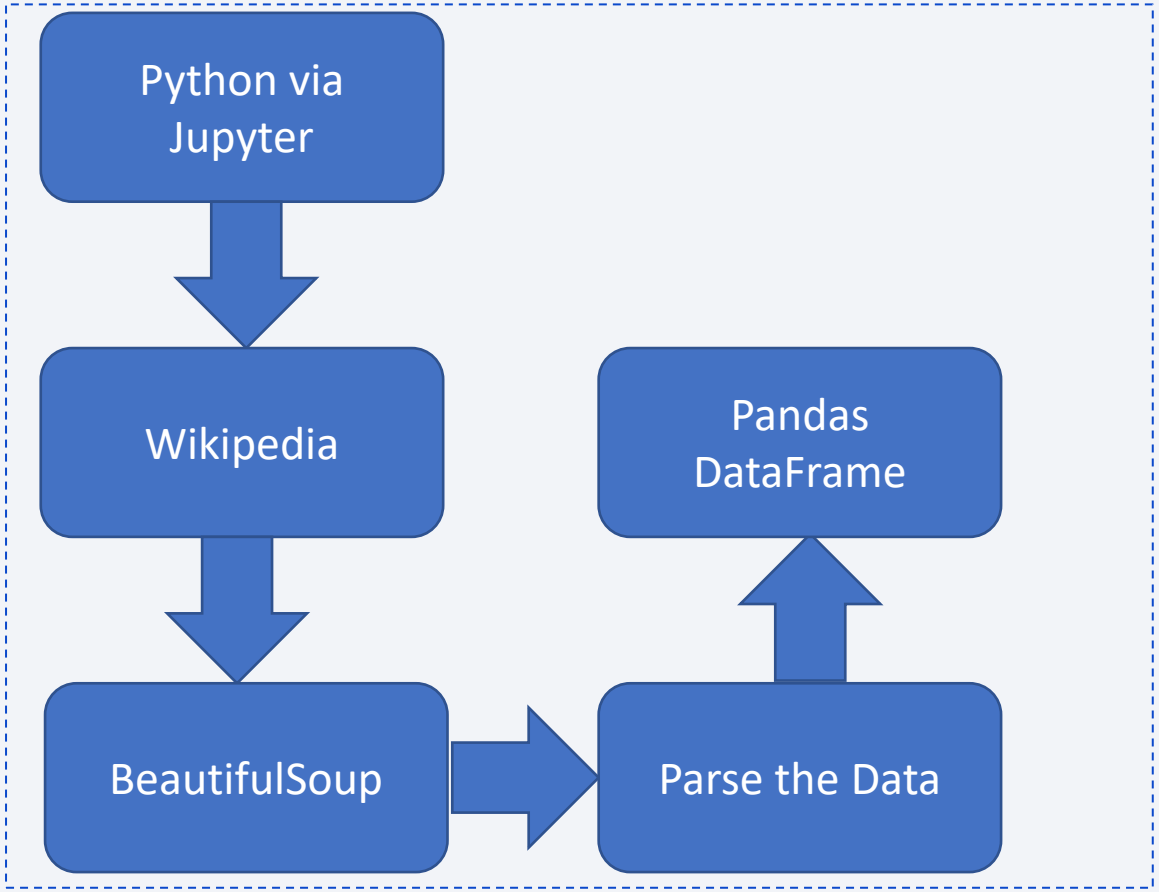
Data Collection – SpaceX API

- There is an open-source REST API for SpaceX launch, rocket, core, launchpad, and landing pad data provided by r-spacex/SpaceX-API on GitHub. Authentication was done by passing the header spacex-key with an API key.
- GitHub URL: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/01_SpaceX-Data-Collection-API.ipynb



Data Collection - Scraping

- Extract relevant information related to SpaceX and its Falcon 9 rocket from Wikipedia. Use BeautifulSoup. Parse the data into a dictionary. Then create Pandas DataFrame.
- GitHub URL: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/02_Webscraping.ipynb

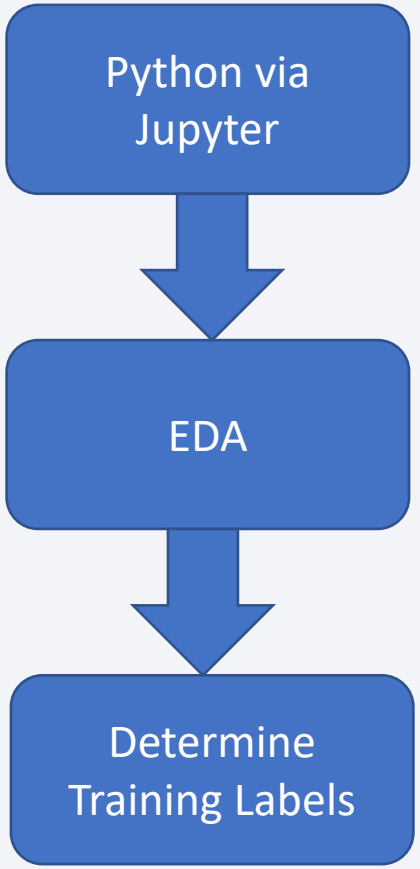


Data Wrangling

- Convert raw data into a usable form. Perform Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

- GitHub URL:

https://github.com/ctaghoy/applied_data_science_capstone/blob/main/03_Data_Wrangling.ipynb



EDA with Data Visualization

- Gain insights from visualizing data using bar charts, line charts, scatter plots. Through these chart types we can visualize and answer some valuable research questions. For example, bar charts are great when we want to track the development of one or two variables over time, while scatter plots are useful for visualizing the relationship between two numerical variables.
- GitHub URL: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/05_EDA-Dataviz.ipynb

EDA with SQL

- 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

EDA with SQL

- List the names of the booster versions which have carried the maximum payload mass. Use 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: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/04_EDA-SQL.ipynb

Build an Interactive Map with Folium

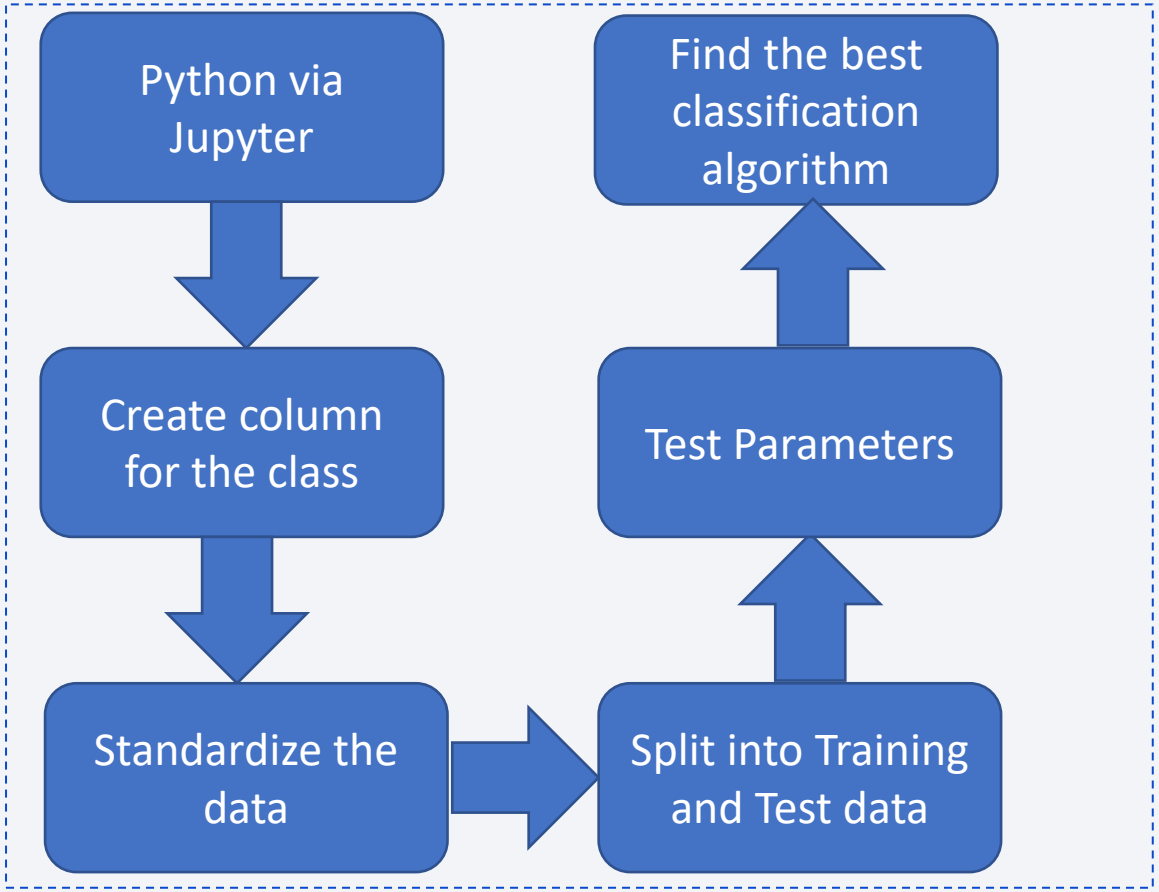
- Perform more interactive visual analytics using Folium
- The launch success may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. We could discover some of the factors by analyzing the existing launch site locations.
- The following tasks were performed:
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
- GitHub URL: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/06_Launch_Site_Location_Folium.ipynb

Build a Dashboard with Plotly Dash

- Interactions in dashboards allow for exploration of the data on a deeper level and make well-informed, data-driven business decisions.
- The dashboard includes the following:
  - Dropdown list to enable Launch Site selection.
  - Pie chart that shows successful launches count for one or all sites. If a specific launch site was selected, it shows the Success vs. Failed counts.
  - A slider to select payload range
  - A scatter chart to show the correlation between payload and launch success.
- GitHub URL: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/07_SpaceX_Plotly_Dash_App.py

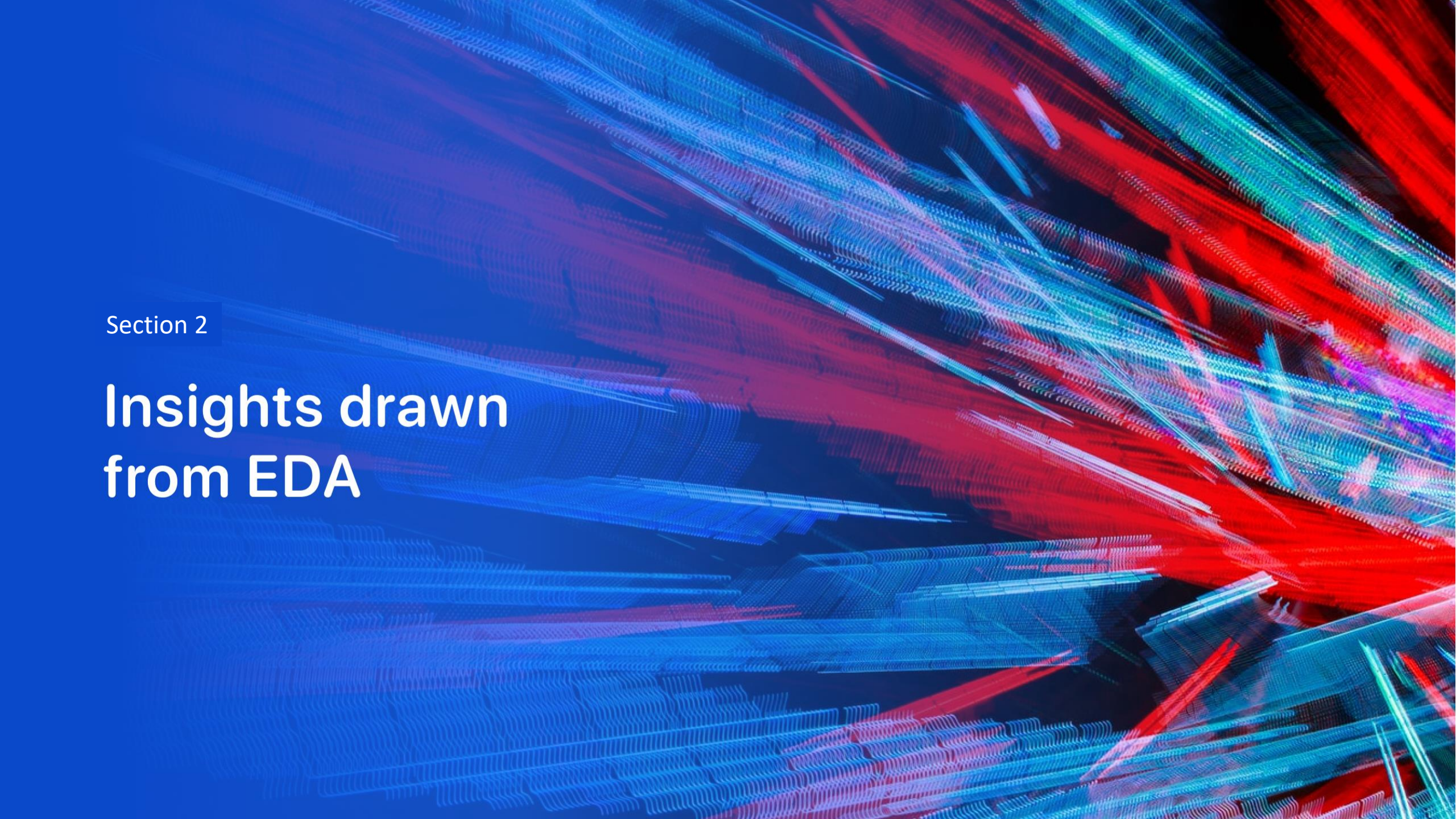
Predictive Analysis (Classification)

- The model development process involves:
  - Preprocessing to standardize the data
  - Split data into training and testing data
  - Test parameters of classification algorithms and find the best one through Sklearn from the following:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K Nearest Neighbor
- GitHub URL: https://github.com/ctaghoy/applied_data_science_capstone/blob/main/08_Machine_Learning_Prediction.ipynb

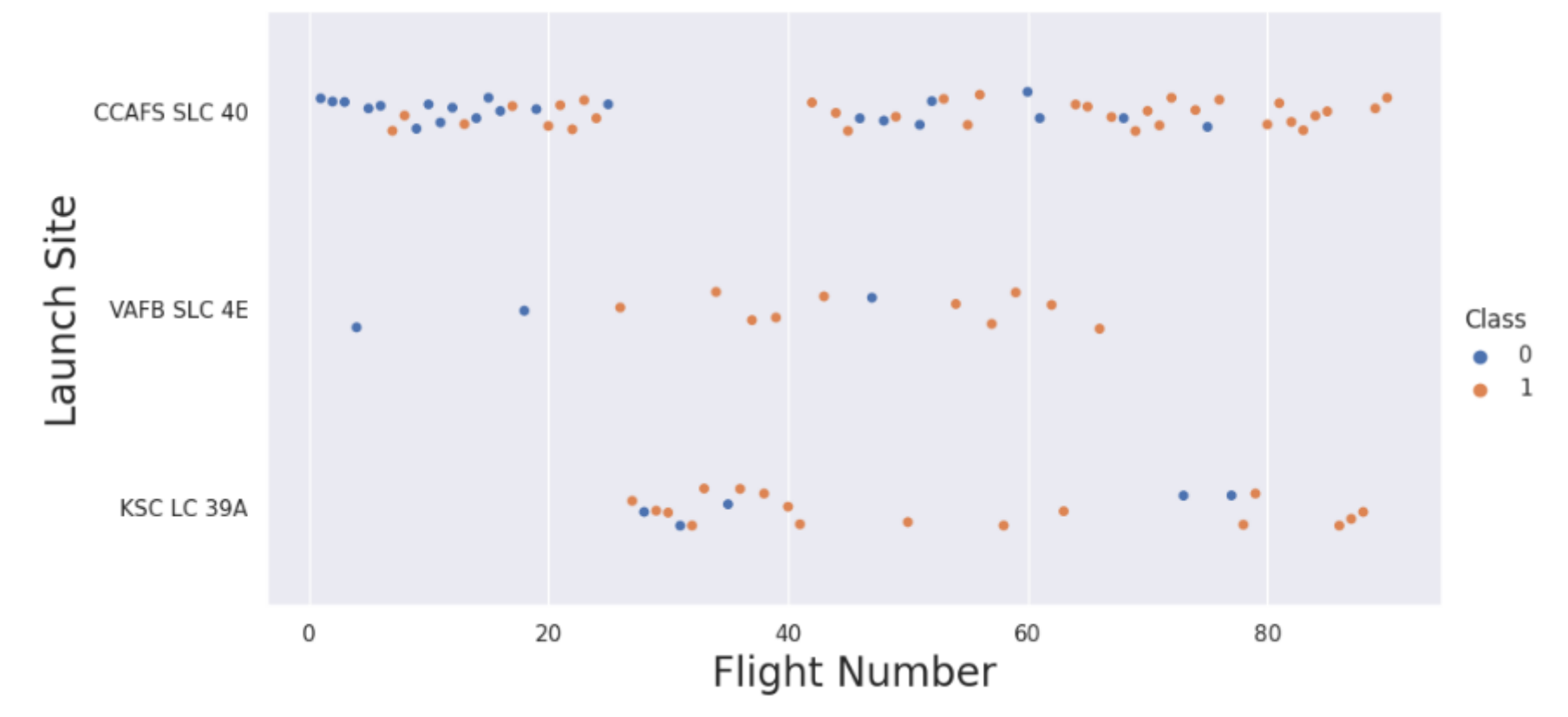


Results

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

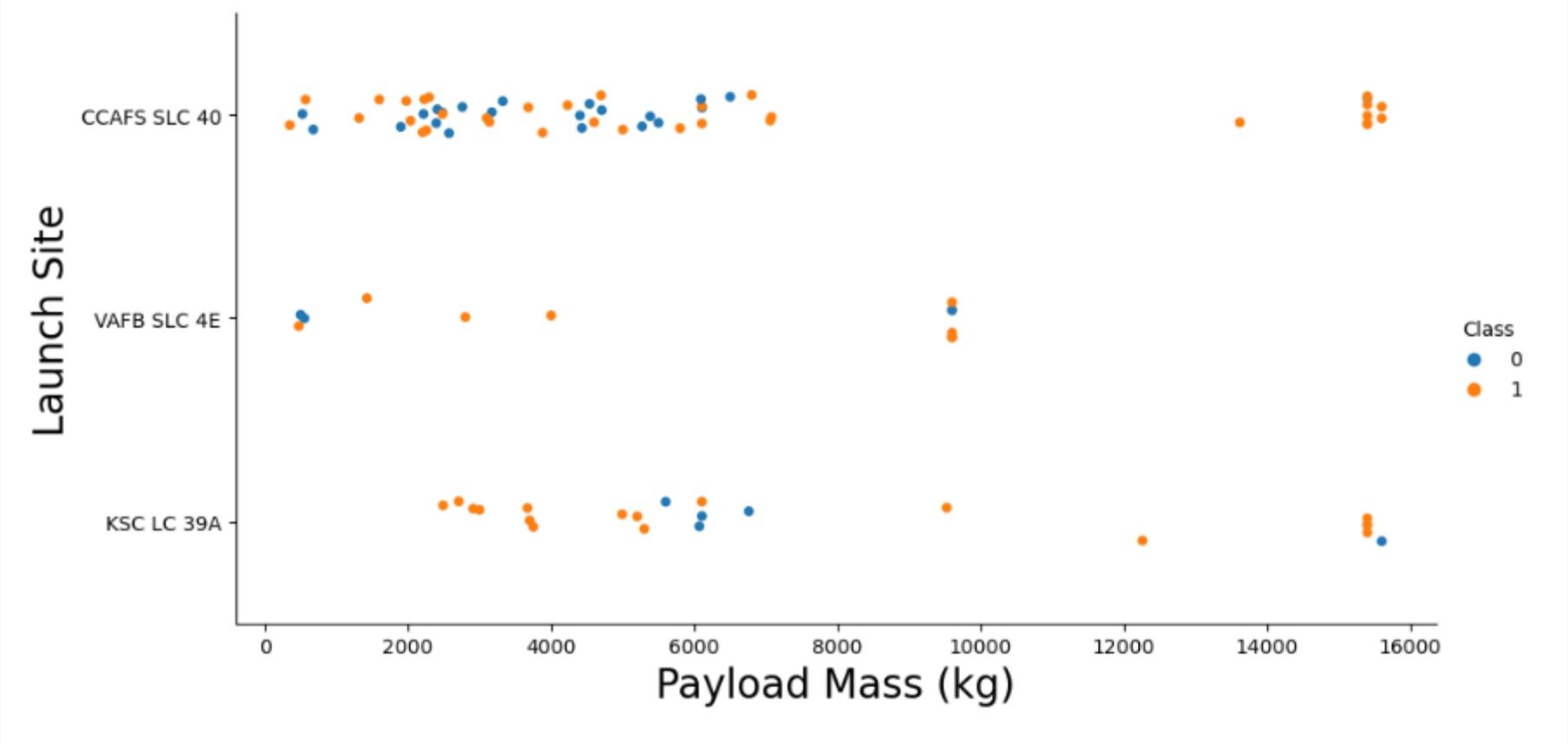

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



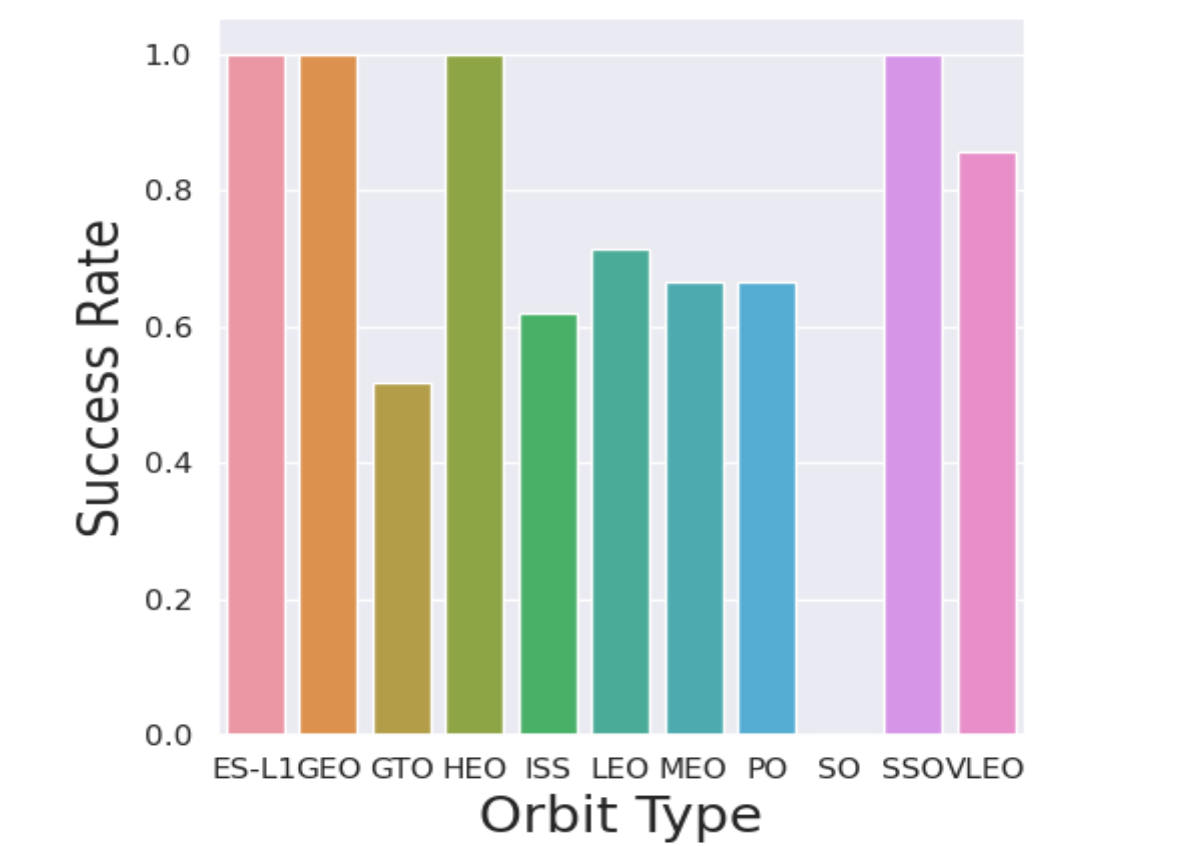
The data shows that as the flight numbers increase, the success rate also increased. The number of flights at Launch Site CCAFS SLC 40 seems to be greater than the others.

Payload vs. Launch Site



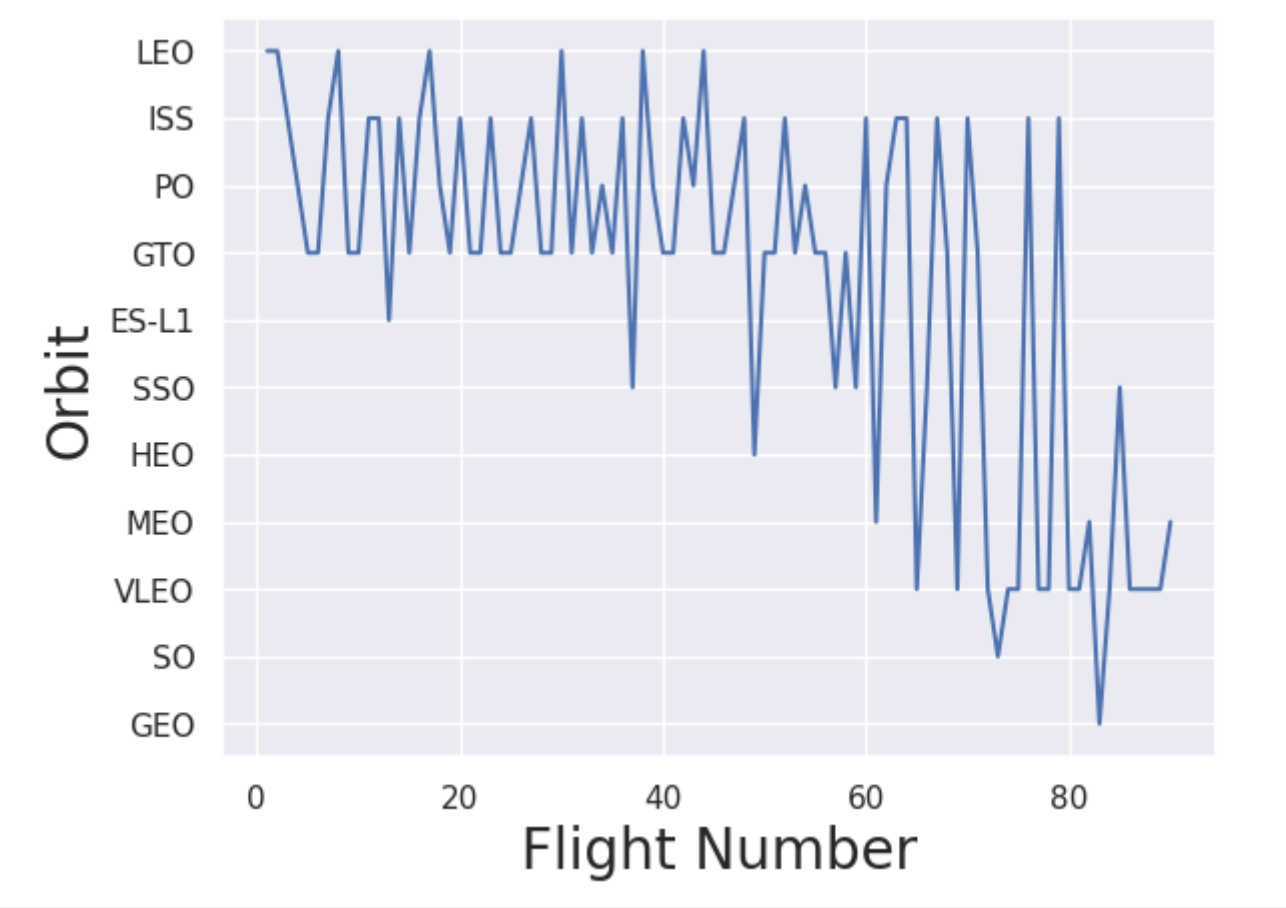
- There were successful launches from CCAFS SLC 40 and KSC LC 39A with higher payloads of about15,000 kg payloads while VAFB SLC 4E did not show any launch from payload mass greater than 10,000 kg.

Success Rate vs. Orbit Type



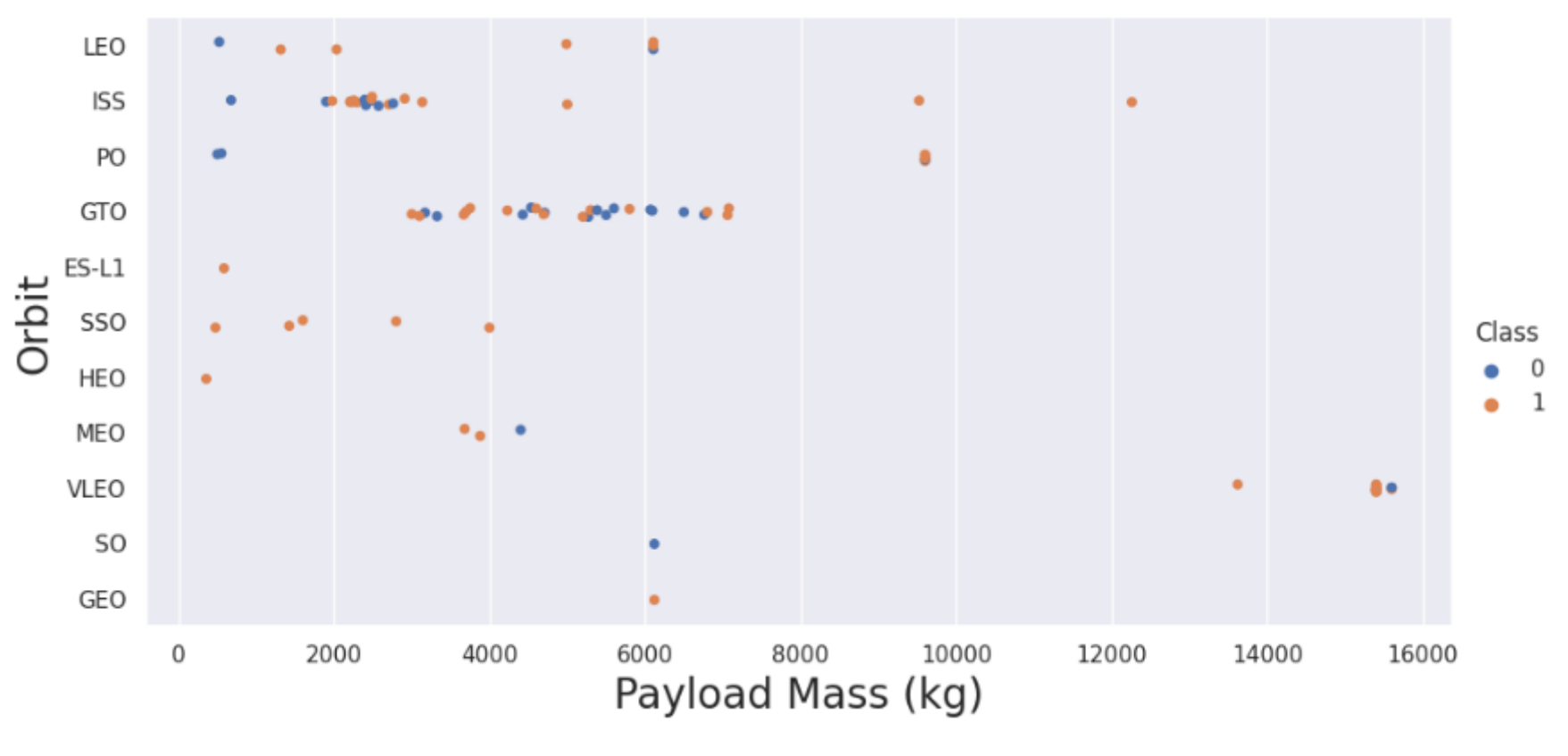
The Orbit Types ES-LI, GEO, HEO, and SSOV have high success rates.

Flight Number vs. Orbit Type



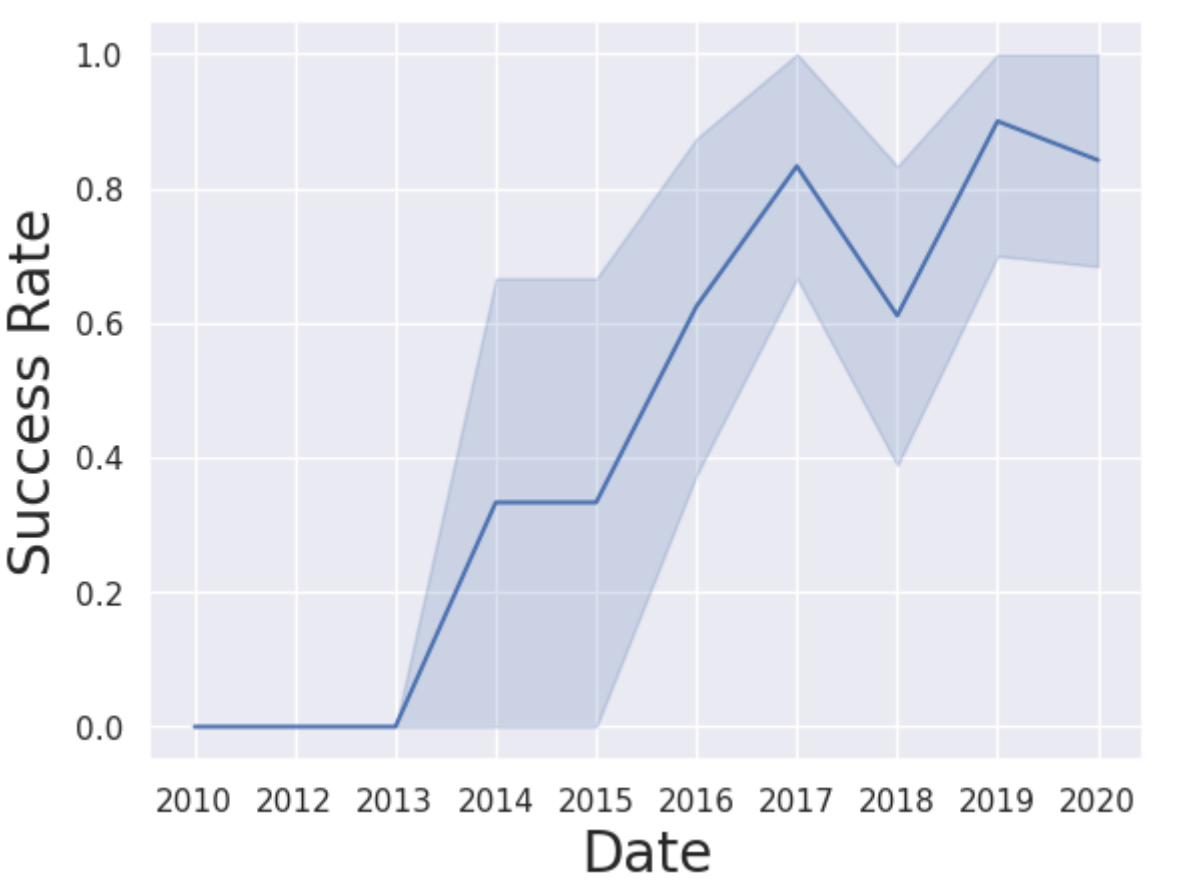
In the VLEO orbit the Success appears related to the number of flights. On the other hand, there seems to be no relationship between flight number when in GTO orbit.

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.

Launch Success Yearly Trend



It can be observed that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

```
In [8]: %sql SELECT DISTINCT launch_site as "Launch_Sites" FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

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

Launch Site Names Begin with 'CCA'

```
In [48]: %sql SELECT * FROM SPACEXTABLE WHERE launch_site LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[48]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

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

Total Payload Mass

```
In [49]: %sql SELECT SUM (PAYLOAD_MASS__Kg_) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[49]: SUM (PAYLOAD_MASS__Kg_)
         45596
```

Calculate the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

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

Calculate the average payload mass carried by booster version F9 v1.

First Successful Ground Landing Date

```
In [98]:  %%sql
          SELECT MIN (DATE) AS "First Successful Landing" FROM SPACEXTABLE
          WHERE landing_outcome = "Success (ground pad)";

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

Out[98]:

First Successful Landing
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

In [51]:
```
%sql SELECT booster_version FROM SPACEXTABLE WHERE \
landing_outcome = 'Success (drone ship)' and payload_mass__Kg_ > 4000 \
AND payload_mass__Kg_ < 6000;
```

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

Out[51]:

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

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

In [97]:
```
%%sql
SELECT COUNT(*) as "Total Missions",
SUM(CASE WHEN mission_outcome LIKE "Success%" THEN 1 ELSE 0 END) as "Successful Missions",
SUM(CASE WHEN mission_outcome LIKE "Failure%" THEN 1 ELSE 0 END) as "Failed Missions"
from SPACEXTABLE;
```

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

Out[97]:

Total Missions	Successful Missions	Failed Missions
101	100	1

Calculate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

In [24]:
```
%sql SELECT booster_version FROM SPACEXTABLE WHERE \
payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTABLE);
```

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

Out[24]:

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

There are 12 booster versions which have carried the maximum payload mass

2015 Launch Records

In [99]:

```
%%sql
SELECT
CASE substr(Date, 6, 2)
    WHEN '01' THEN 'January'
    WHEN '02' THEN 'February'
    WHEN '03' THEN 'March'
    WHEN '04' THEN 'April'
    WHEN '05' THEN 'May'
    WHEN '06' THEN 'June'
    WHEN '07' THEN 'July'
    WHEN '08' THEN 'August'
    WHEN '09' THEN 'September'
    WHEN '10' THEN 'October'
    WHEN '11' THEN 'November'
    WHEN '12' THEN 'December'
END
    AS Month,
date, booster_version, launch_site, [Landing_Outcome] FROM SPACEXTABLE
WHERE [Landing_Outcome] LIKE 'Failure%' AND substr(Date, 1, 4) = '2015';
```

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

Out[99]:

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
October	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
[17]: %%sql
SELECT Landing_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
WHERE DATE BETWEEN '2010-06-04'
AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC;
```

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

[17]:

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

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


Section 3
Launch Sites
Proximities Analysis

Location of Launch Sites

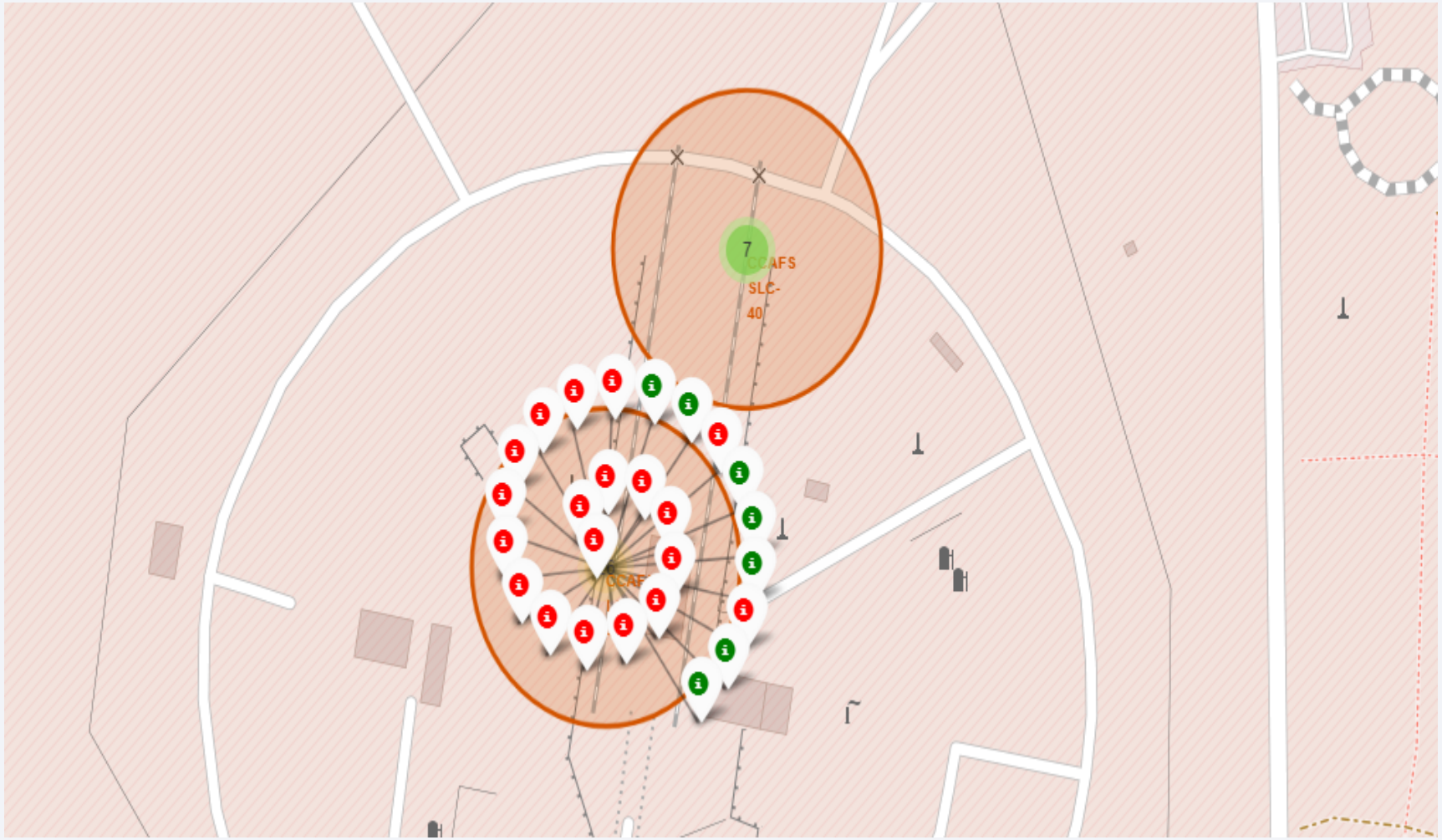
All launch sites' location markers on a global map

Folium Map showing Success and Fail Markers

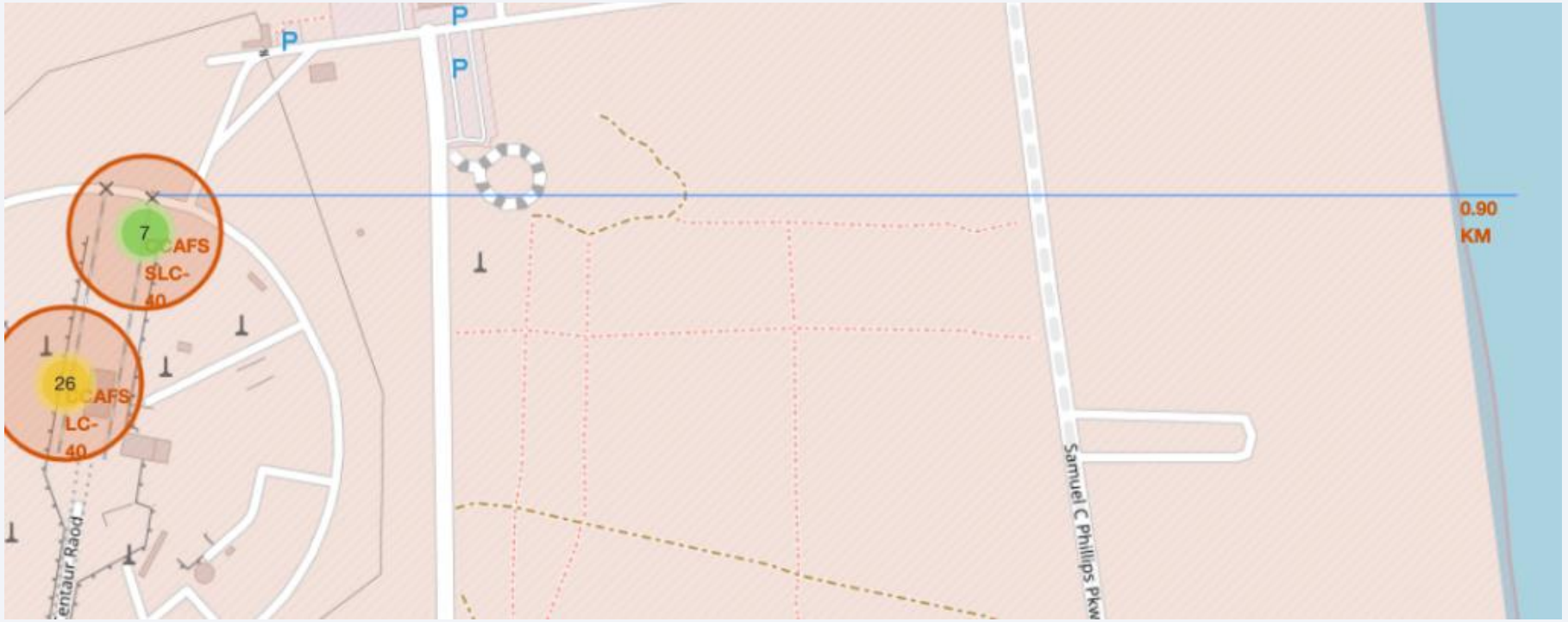
Enhanced map by adding the launch outcomes for each site.

Successful launch (class 1) = green marker

Failed launch (class 0) = red marker

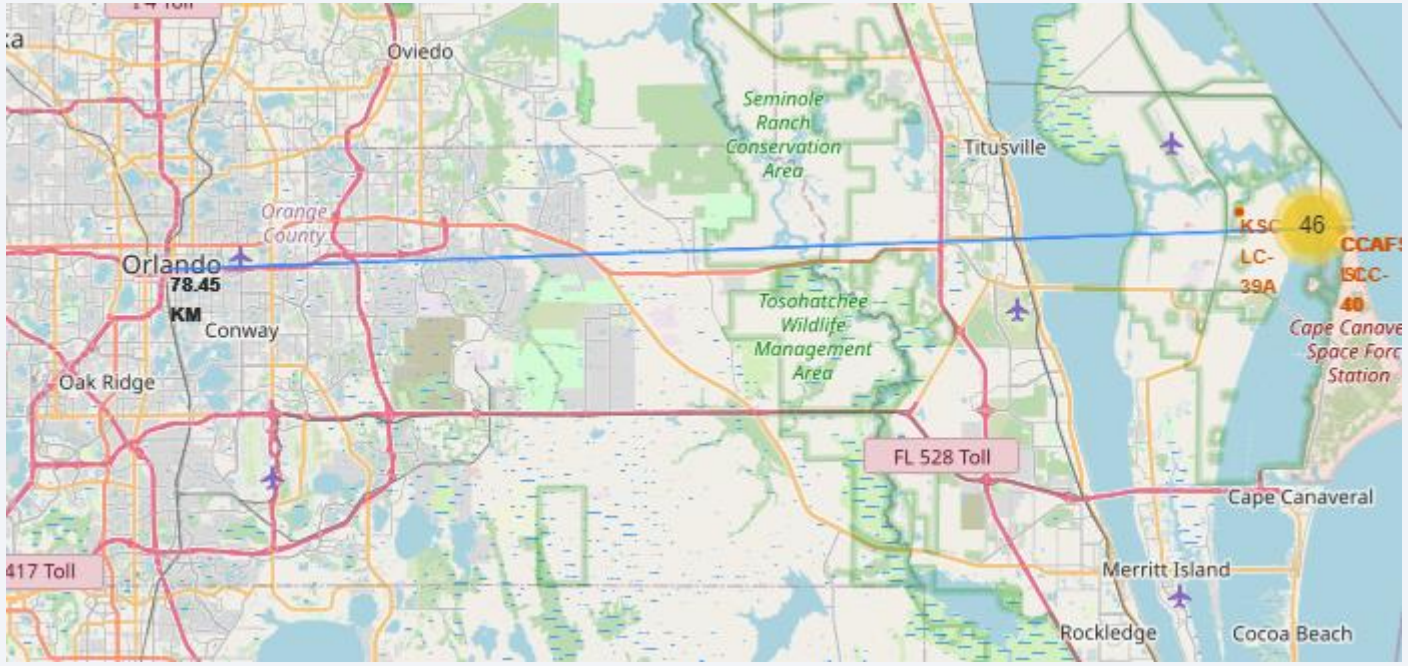


Folium Map showing proximities of launch sites

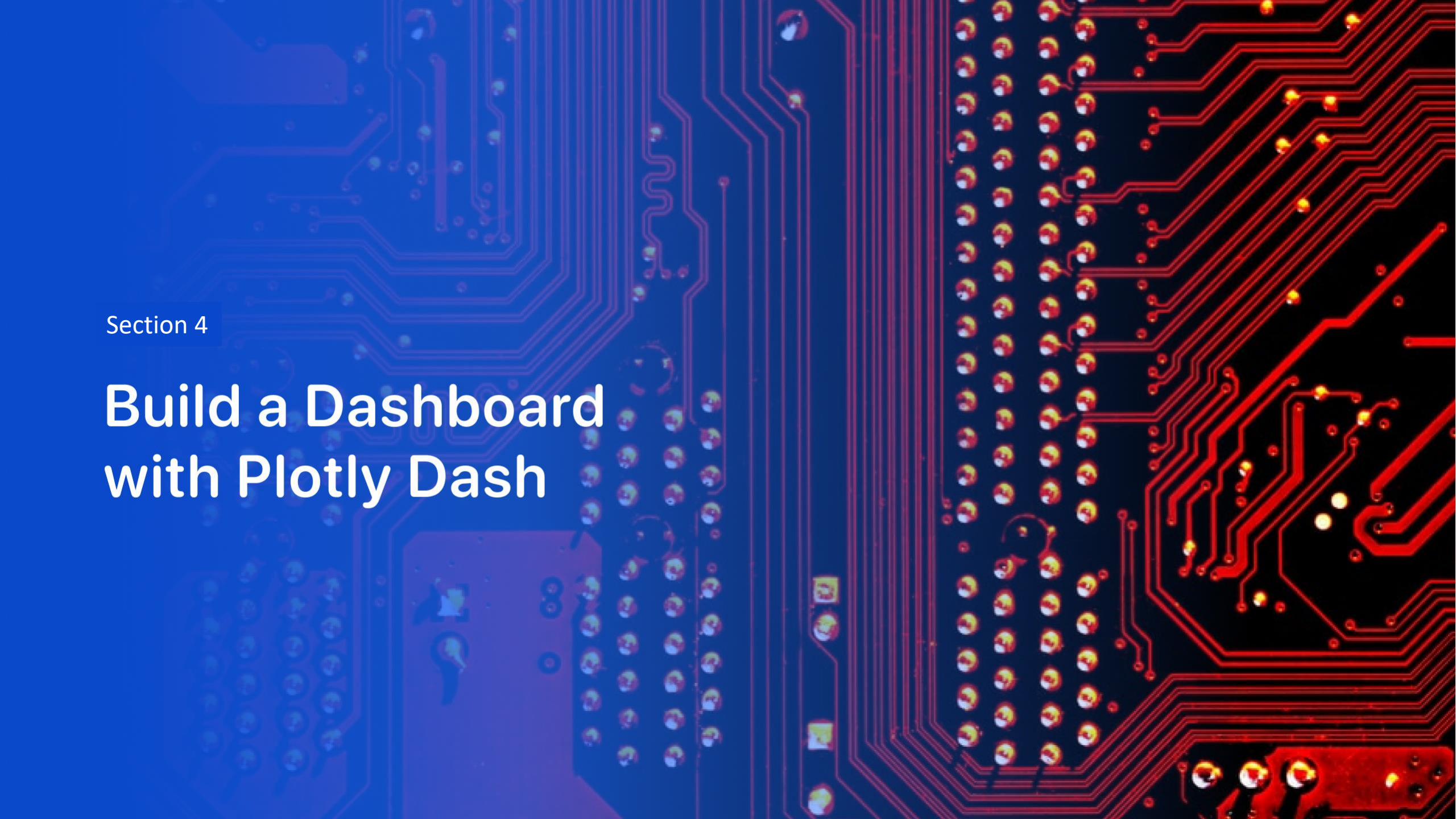


Are launch sites in close proximity to coastline? No.

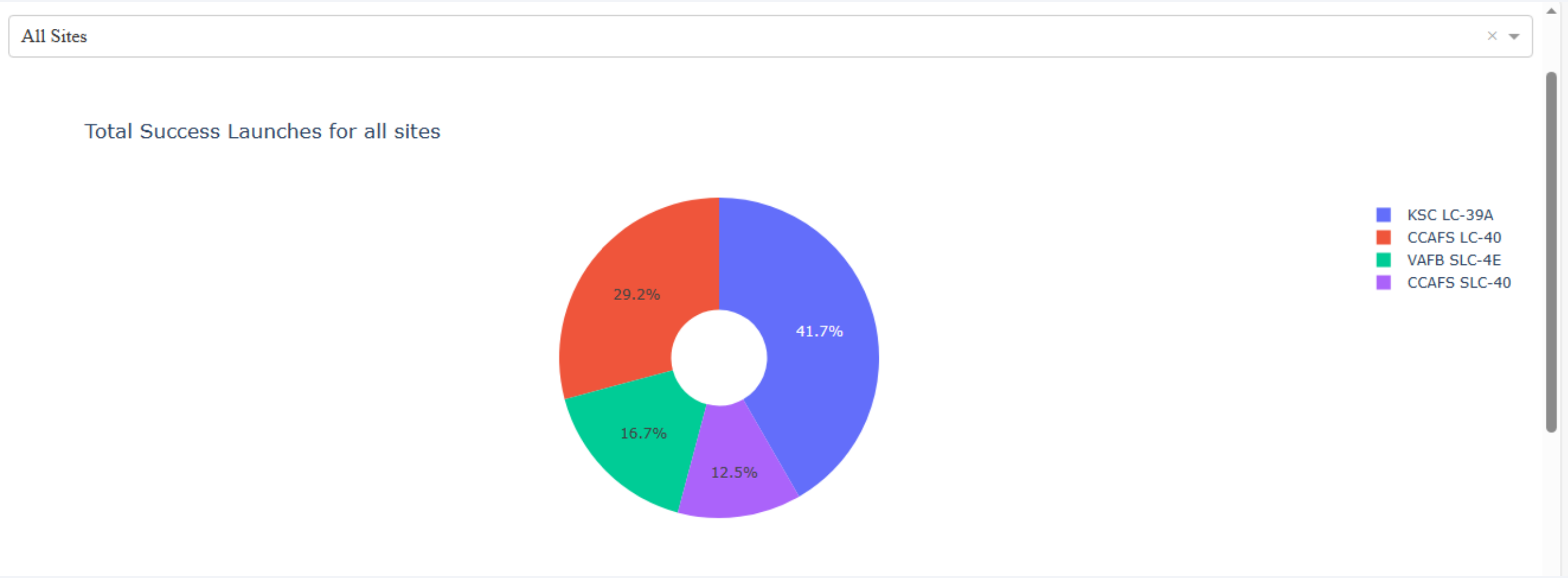
Folium Map showing proximities of launch sites



Are launch sites in close proximity to railways? No.
Are launch sites in close proximity to highways? No.
Do launch sites keep certain distance away from cities? Yes.

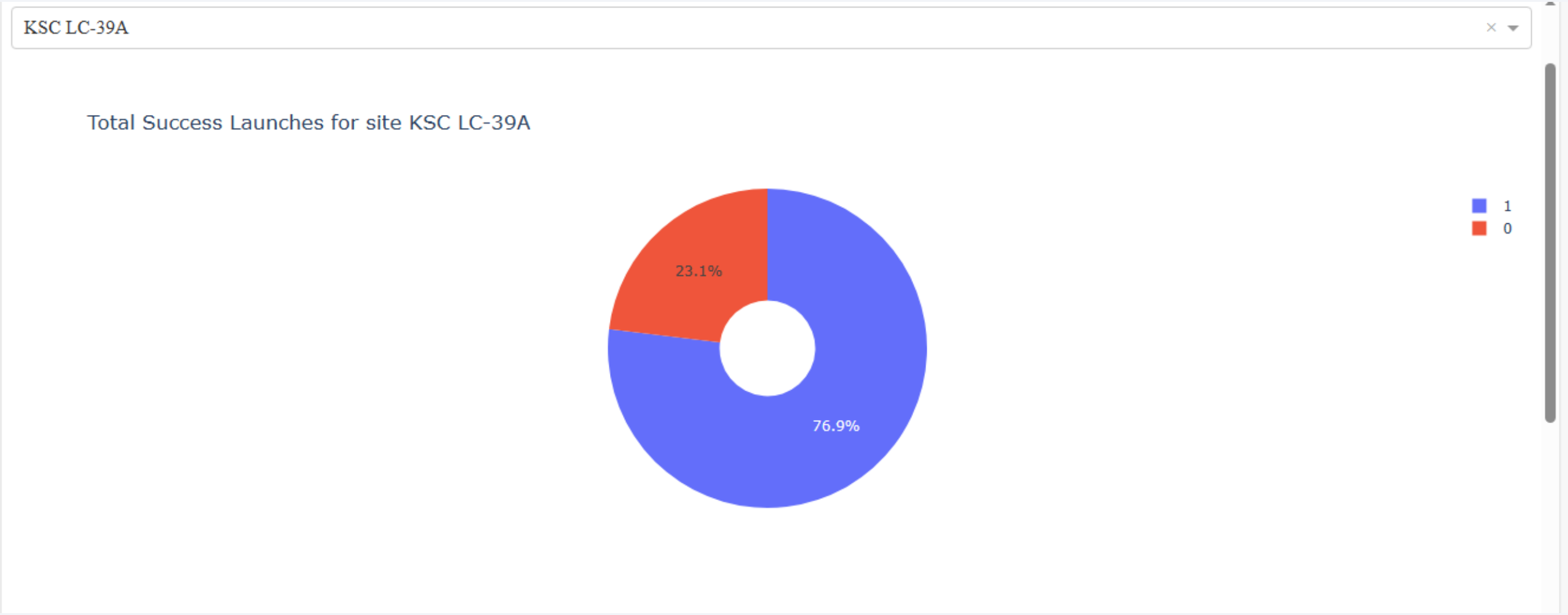

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Plotly Dash – Pie Chart



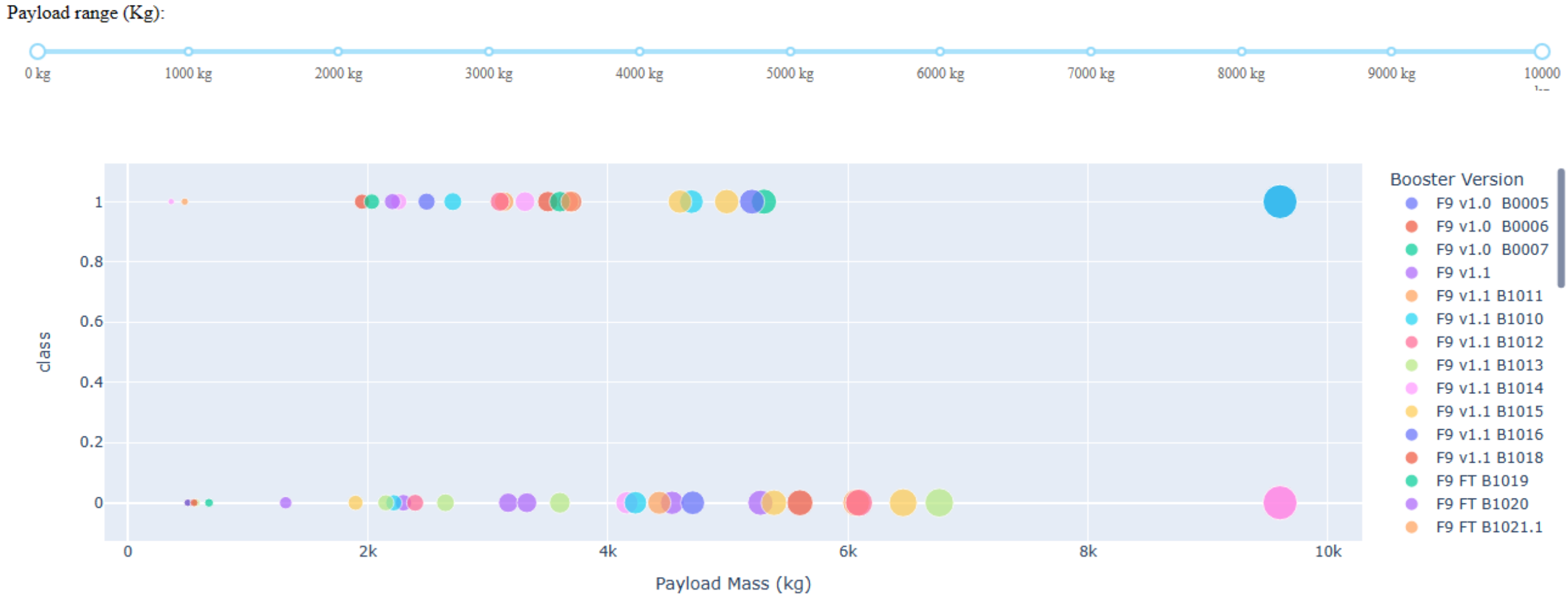
Launch success count for all sites in a pie chart

Dashboard – Plotly Dash – Pie Chart



Pie chart for the launch site with highest launch success ratio

Dashboard - Payload vs. Launch Outcome scatter plot



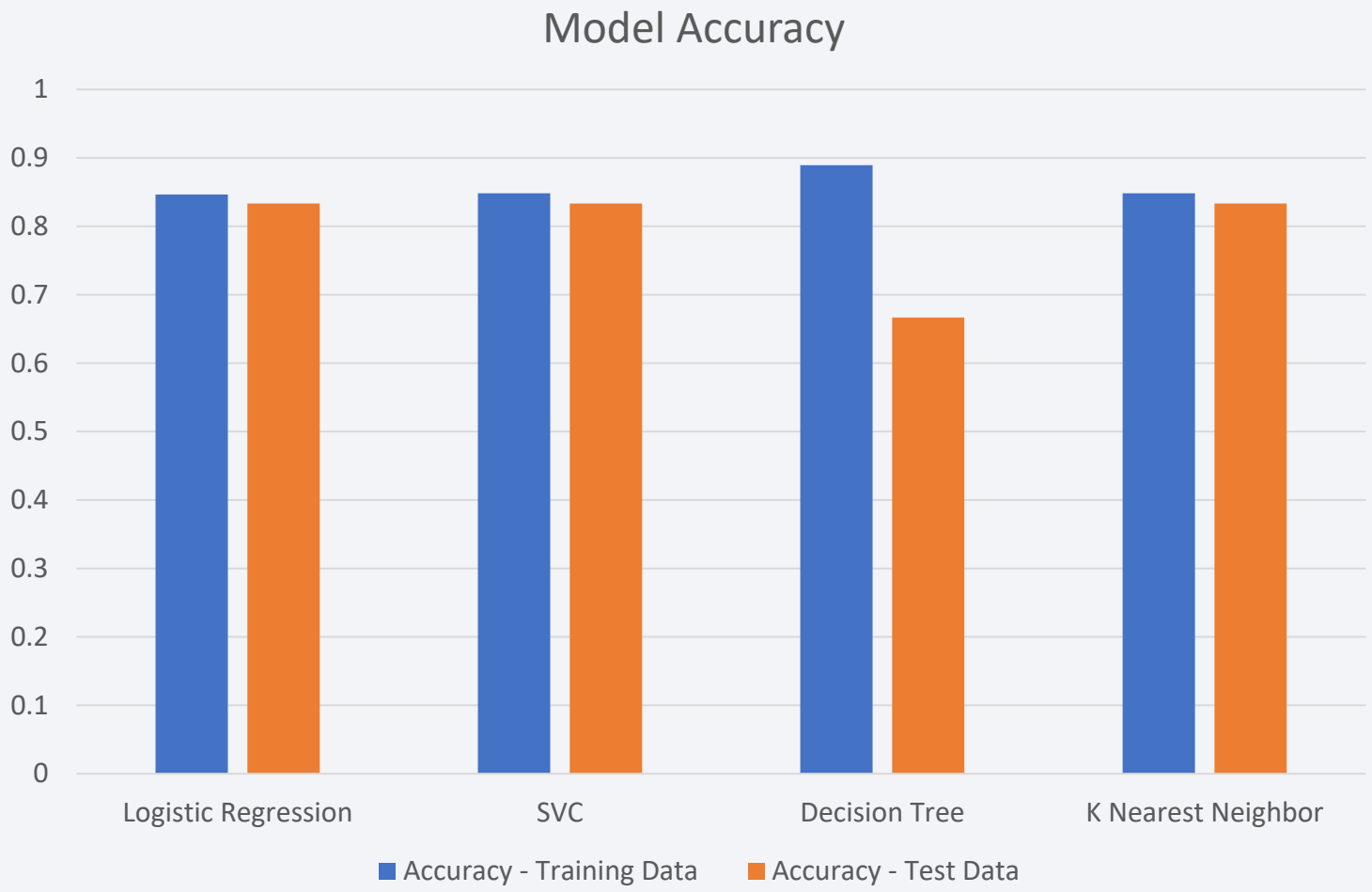
Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Section 5

Predictive Analysis (Classification)

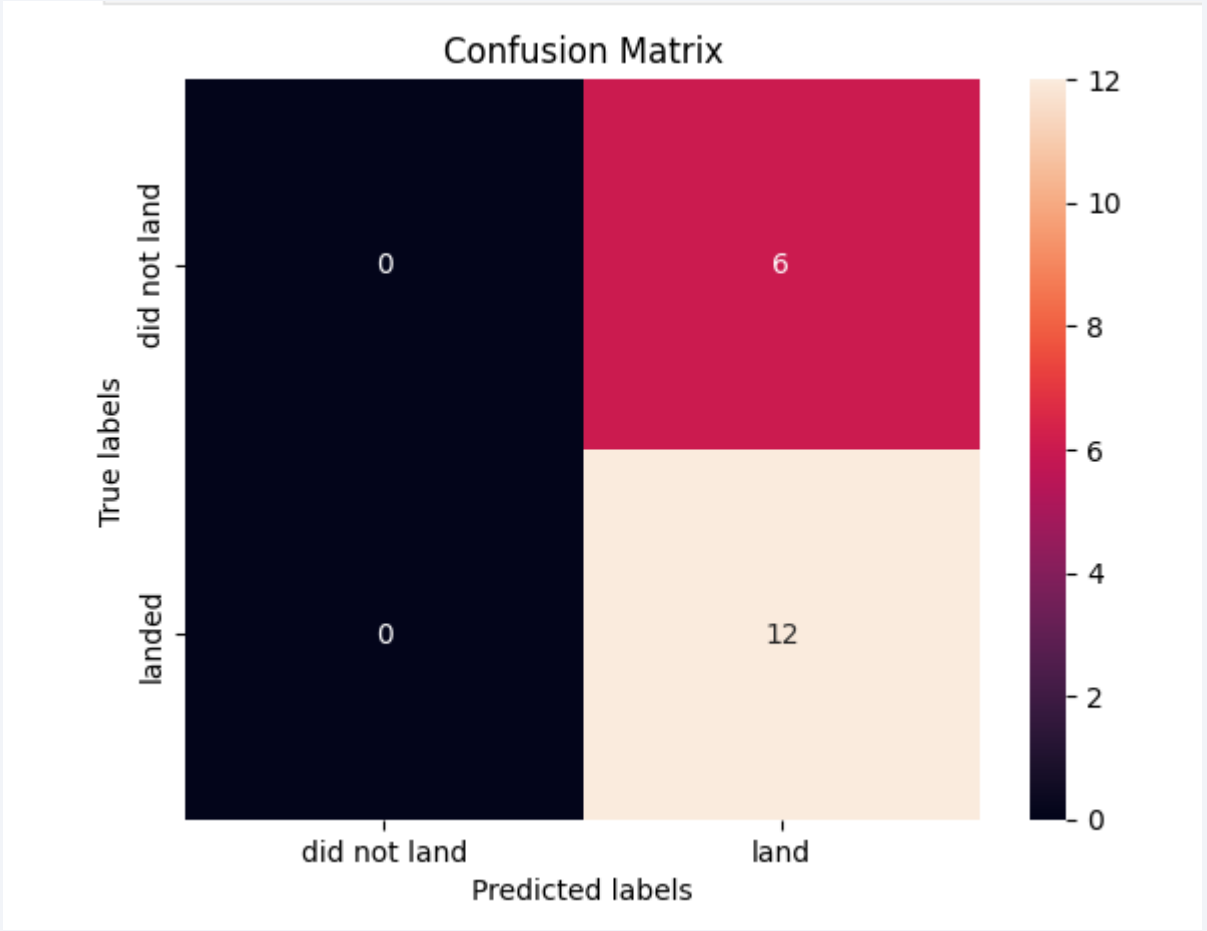
Classification Accuracy

Model that has the highest classification accuracy : Decision Tree



Confusion Matrix – Decision Tree

- A confusion matrix is a table that is used to evaluate how well a machine learning model performs
- The number 12 in the bottom right part of the matrix tells us that the machine learning model predicted that a rocket would land 12 times, and it was correct every time! That's really good!



Conclusions

1. The Decision Tree Classifier Algorithm is the best Machine Learning approach for this dataset.
2. The low weighted payloads (which are defined as 5000kg and below) performed better than the heavy weighted payloads.
3. Starting from 2013, the success rate for SpaceX launches has been increasing every year, and it is expected to continue to improve in the future.
4. KSC LC-39A has the highest success rate of any launch site, with a success rate of 76.9%.
5. Additionally, the Orbit Types ES-LI, GEO, HEO, and SSOV have a success rate of 100% and with more than one occurrence.

Appendix: GitHub Repository

Please refer to the GitHub repository that contains the code and data for this project.

- Repository name: applied_data_science_capstone
- Repository owner: ctaghoy
- Repository URL: https://github.com/ctaghoy/applied_data_science_capstone/tree/main
- Version: main

The code in this repository can be used to reproduce the results of this project. The data used in this project is also available in this repository.

Thank you!