IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amrutha M
23-03-2023

Outline

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

Executive Summary

Summary of methodologies:

- The project collected data from public sources such as the SpaceX API and Wikipedia page, and processed it to prepare it for analysis.
- The data was explored and analyzed through various techniques, including SQL queries, visualizations, and interactive maps generated using Folium.
- An interactive dashboard was developed using Plotly Dash to provide a user-friendly interface for exploring the data.
- The project also trained machine learning models using classification techniques to predict successful landings, based on the data collected and processed.

Summary of results:

- The models developed as part of the project were consistent in their performance, with an accuracy rate of approximately 83.33% when tested with a separate test data set.
- It was observed that the models tended to predict successful landings more often than not, indicating a slight bias towards positive outcomes.
- Further improvements in accuracy could potentially be achieved by incorporating more training data and optimizing the model parameters accordingly.

Introduction

Project background and context:

In recent years, the commercial space industry has been rapidly expanding, with several companies offering affordable space travel options. One of the most successful companies in this space is SpaceX, known for its cost-effective rocket launches. One of the key factors contributing to SpaceX's cost savings is the ability to reuse the first stage of the rocket.

This project aims to predict the likelihood of a successful landing of the Falcon 9 first stage, which would provide valuable information for determining the cost of a launch. If an alternate company is looking to bid against SpaceX for a rocket launch, this information would be particularly useful. By predicting the success of the first stage landing, we can determine the overall cost of the launch, which would be a key factor in deciding whether to pursue a partnership with SpaceX or explore other options.

Problems you want to find answers:

- What factors influence the likelihood of a rocket successfully landing? This would involve identifying the correlations between different variables associated with the rocket and its landing rate.
- What are the optimal conditions required to achieve a successful landing? This would involve determining the conditions that yield the best results and ensure a higher likelihood of success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX REST API and web scraping from SpaceX Wikipedia page.
- Perform data wrangling
  - Irrelevant fields removed, records with null values removed.
  - Transform the results into training labels, classifying each observation as either a successful or unsuccessful landing of the booster.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - SVM, Classification Trees and Logistic Regression

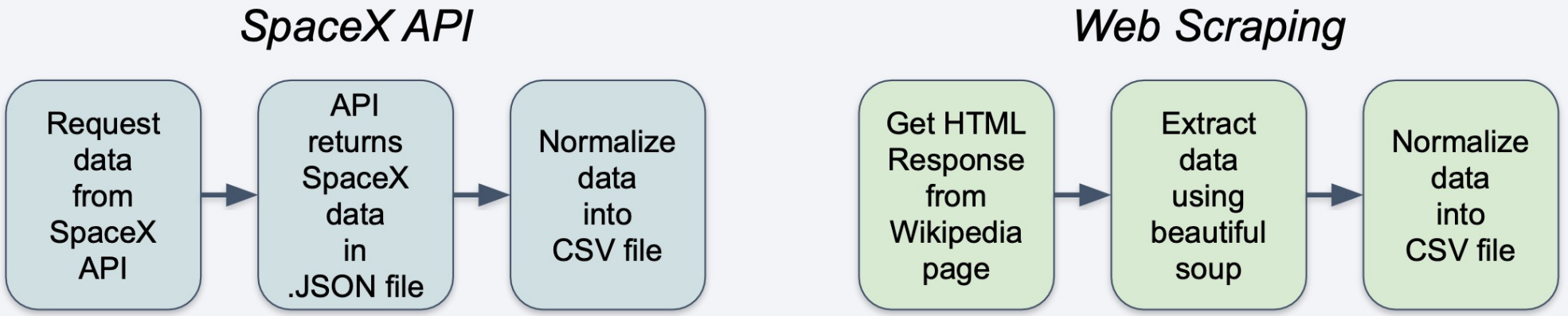
Data Collection

- **SpaceX REST API**

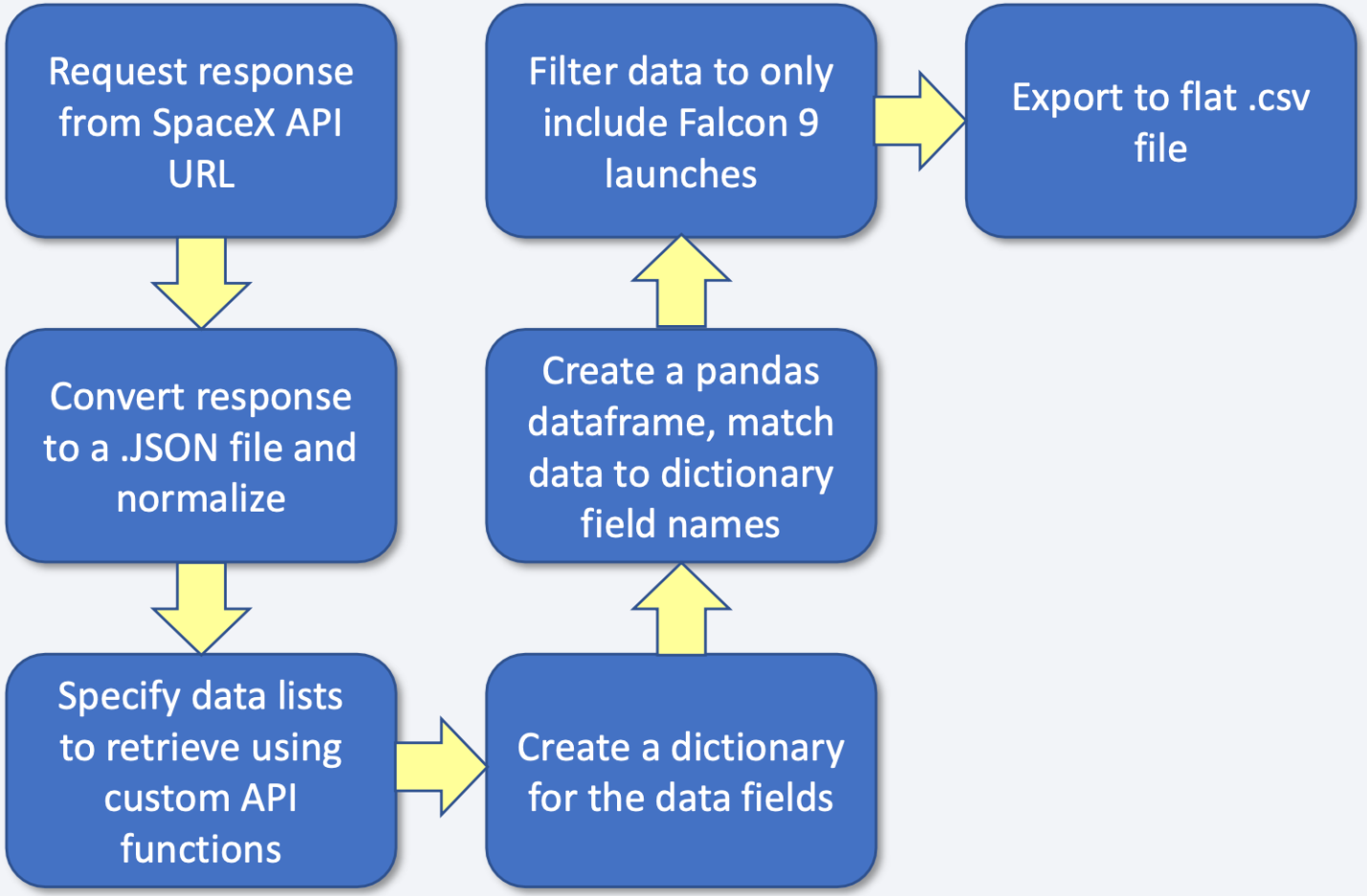
Data was collected from the SpaceX REST API. The data included information about the rocket used, payload, landing outcome, and other launch & landing specifications in the form of a .JSON file. Data is normalized into a flat .csv file.

- **Web Scraping**

Data was also collected by web scraping a table from the SpaceX Wikipedia page using the BeautifulSoup python package. Information can be appended to our dataset by using the rocket/flight id as a key. Data is normalized into a flat .csv file.

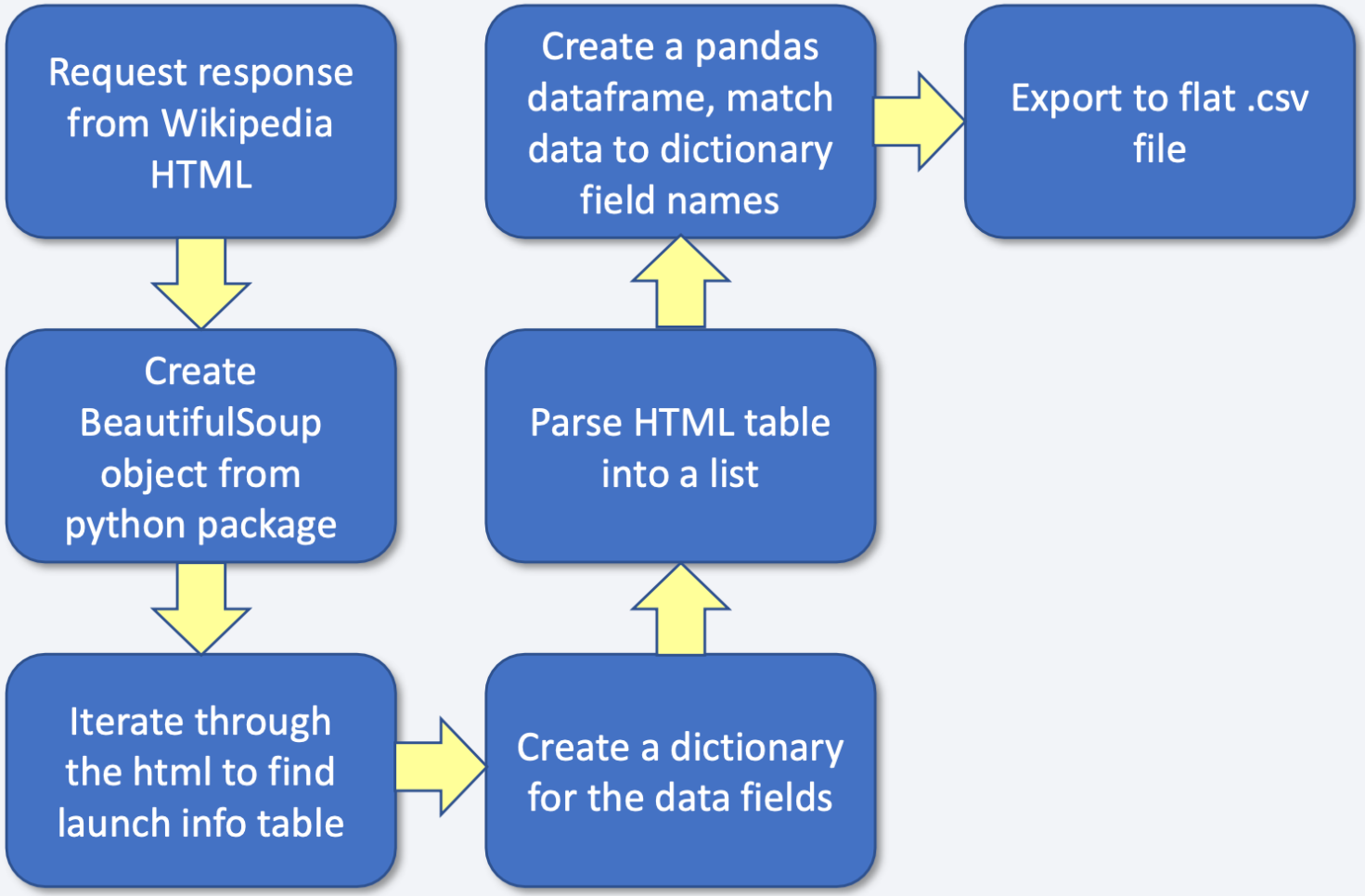


Data Collection – SpaceX API



GitHub URL

Data Collection - Scraping



GitHub Link

Data Wrangling

As part of the data wrangling process in this project, we will conduct exploratory data analysis to identify patterns and determine appropriate labels for training supervised models. The dataset includes various scenarios where the Falcon 9 booster did not land successfully. For example, a landing may have been attempted but failed due to an accident, resulting in outcomes such as True Ocean (successfully landed in the ocean) or False Ocean (unsuccessfully landed in the ocean), True RTLS (successfully landed on a ground pad) or False RTLS (unsuccessfully landed on a ground pad), and True ASDS (successfully landed on a drone ship) or False ASDS (unsuccessfully landed on a drone ship).

Our main objective in this project is to convert these outcomes into training labels, where a value of 1 represents a successful landing and 0 indicates an unsuccessful landing. By doing so, we aim to generate a dataset that can be used to train machine learning models to predict the likelihood of a successful landing for the Falcon 9 first stage.

Data Wrangling

The number and occurence of mission outcome per orbit type

```
landing_outcomes = df.Outcome.value_counts()
landing_outcomes
```

```
True ASDS       41
None None       19
True RTLS       14
False ASDS       6
True Ocean       5
False Ocean      2
None ASDS        2
False RTLS       1
Name: Outcome, dtype: int64
```

Calculating the success rate for every landing in dataset

```
df["Class"].mean()
```

```
0.6666666666666666
```

GitHub URL

EDA with Data Visualization

- Scatter chart:

-Flight Number vs. Launch Site

-Payload vs. Launch Site

-Flight Number vs. Orbit Type

-Payload vs. Orbit Type

GitHub URL

- Bar chart:

-Orbit Type vs. Success Rate

- Line chart:

-Year vs. Success Rate

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CAA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch site for the months in year 2015
- Ranking the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

GitHub URL

Build an Interactive Map with Folium

- We added several objects to a Folium map, including markers that display all launch sites and their corresponding success/failure statuses. We also added lines that show the distances between each launch site and its surrounding areas. By incorporating these objects, we were able to identify various geographical patterns related to launch sites.
- Our analysis revealed that launch sites are generally located in close proximity to railways, highways, and coastlines. We also found that launch sites tend to maintain a certain distance away from nearby cities.

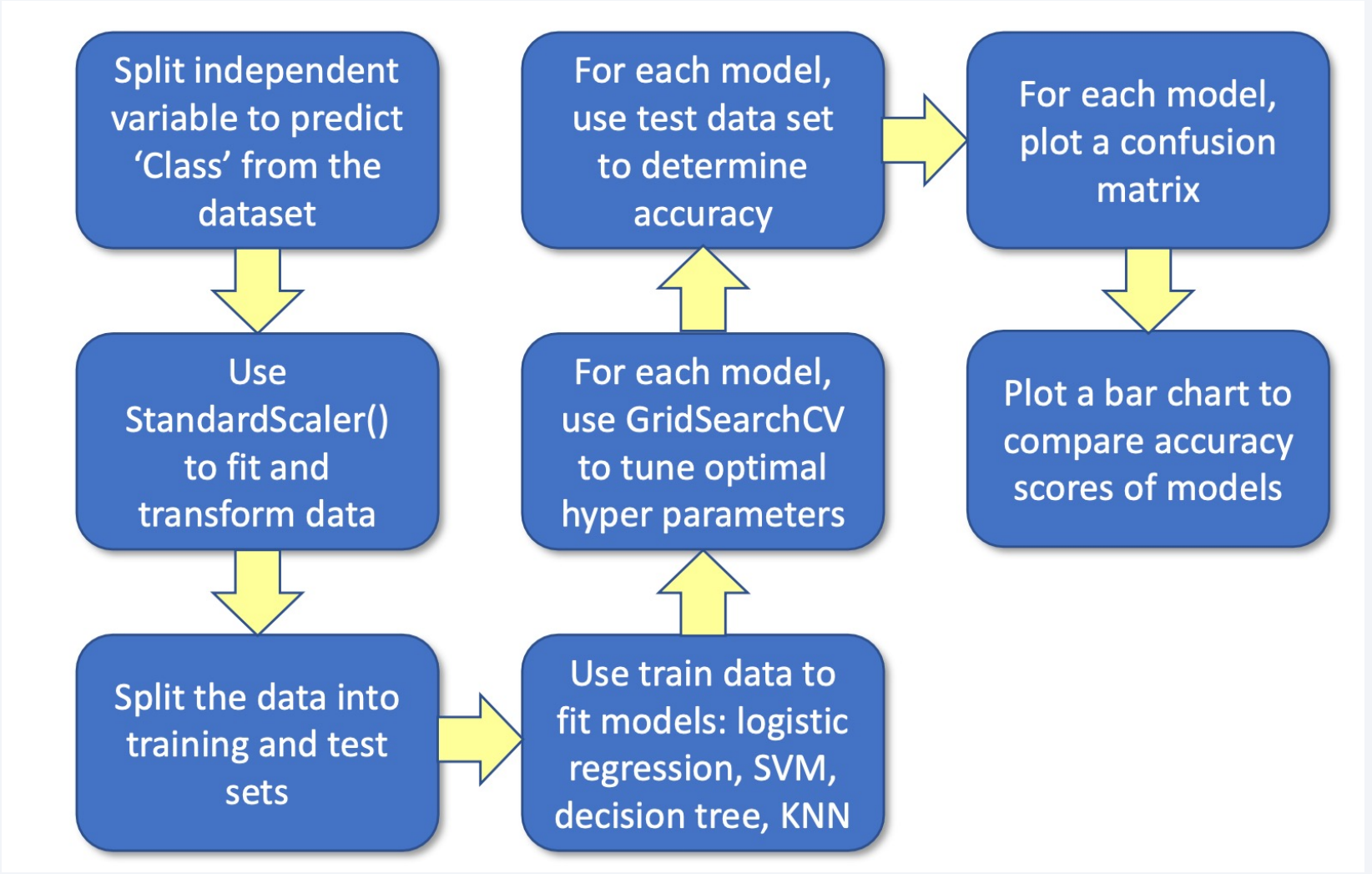
GitHub URL

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Interactive pie chart used to visualise launch site success rate; showing distribution of successful landings across all launch sites or distribution of successful landings for specific individual launch site.
- Scatter plot used to visualise how success varies dependent on payload mass and booster version category.

GitHub URL

Predictive Analysis (Classification)

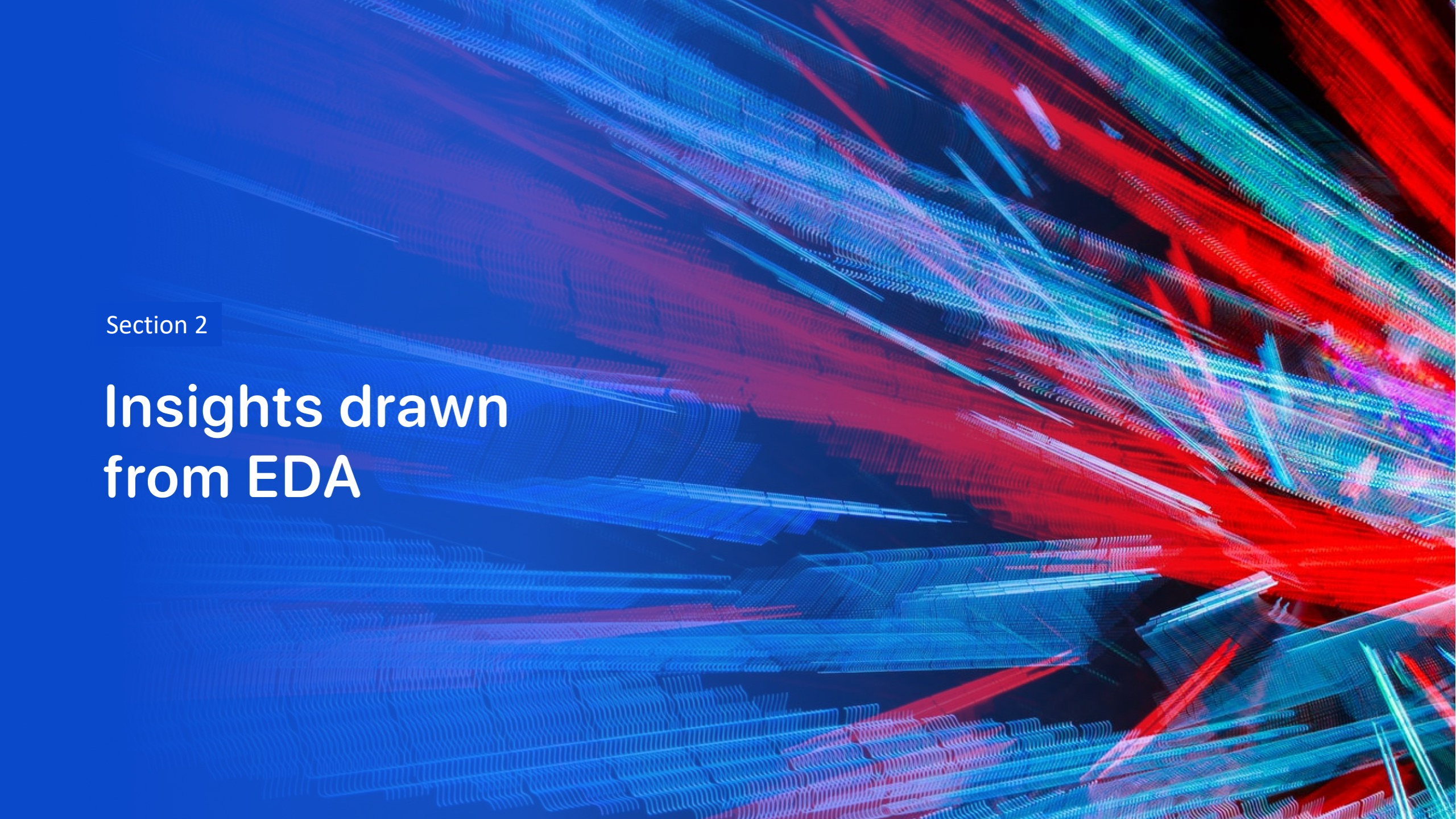


GitHub URL

Results

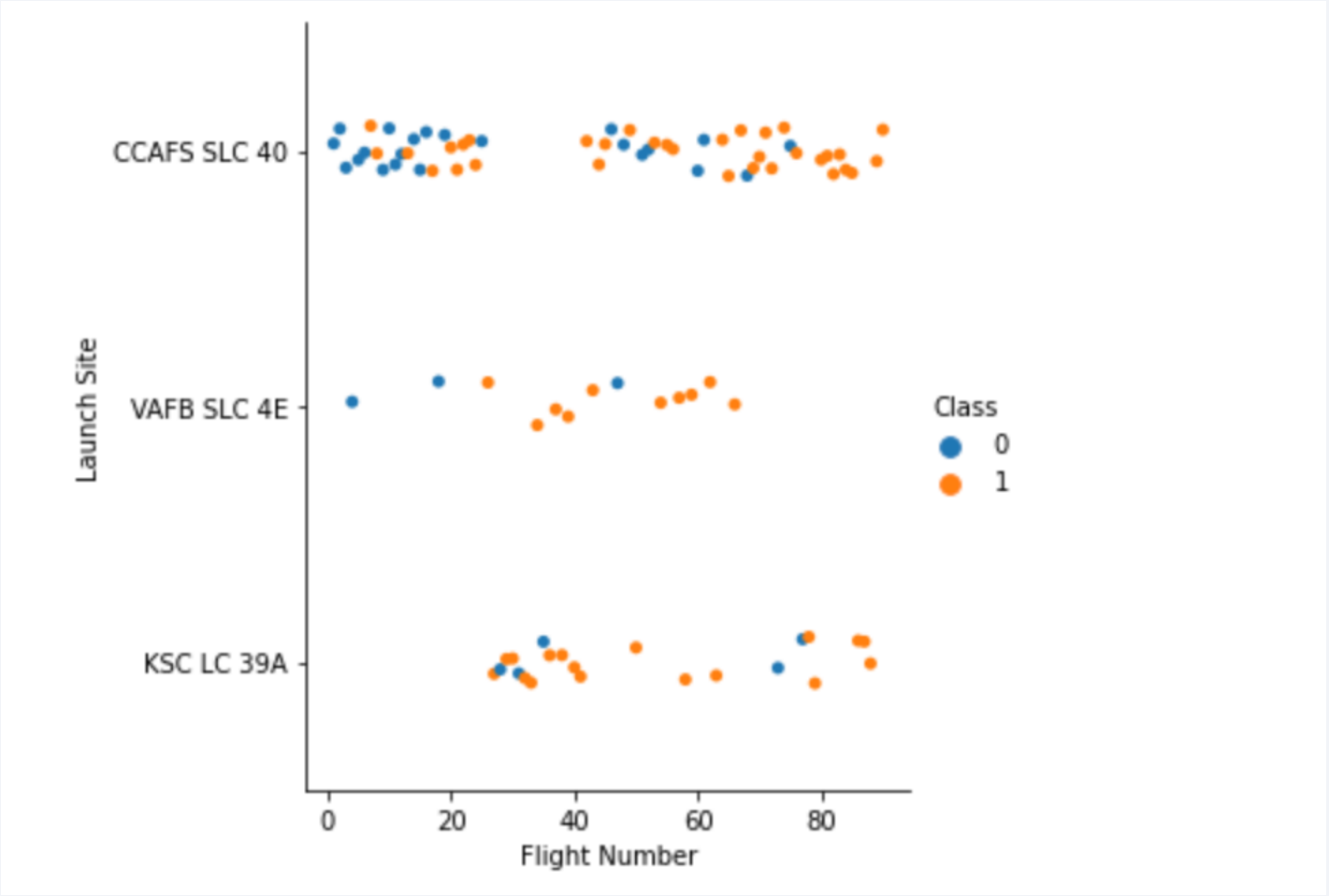
- The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and Interactive Dashboard will be shown in the next slides.

- Comparing the accuracy of the four methods, all return the same accuracy of about 83% for test data.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

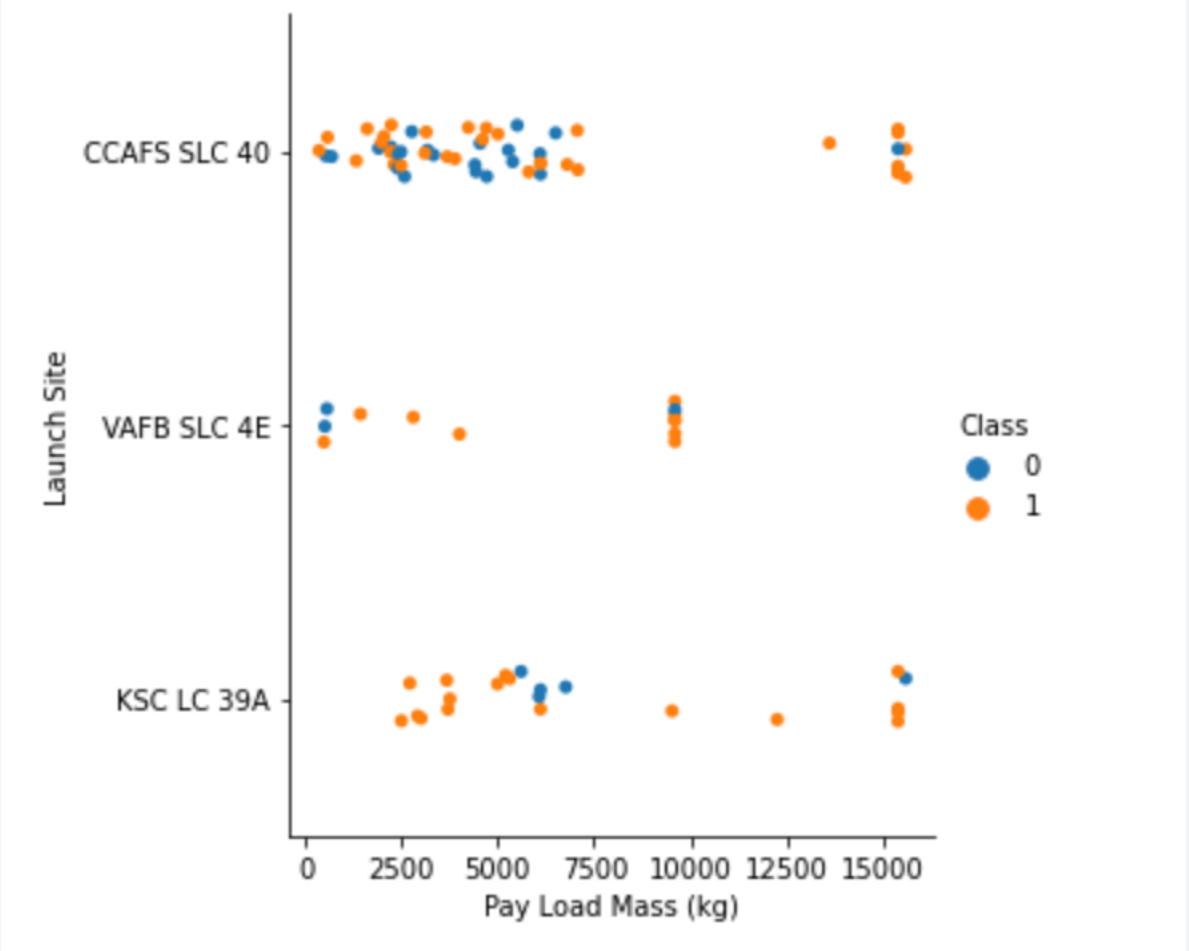
- `Class 0` (blue) represents unsuccessful launch, and `Class 1` (orange) represents successful launch.
- This figure shows that **the success rate increased as the number of flights increased**.
- As the success rate has increased considerably since the *20th* flight, this point seems to be a big breakthrough.



Payload vs. Launch Site

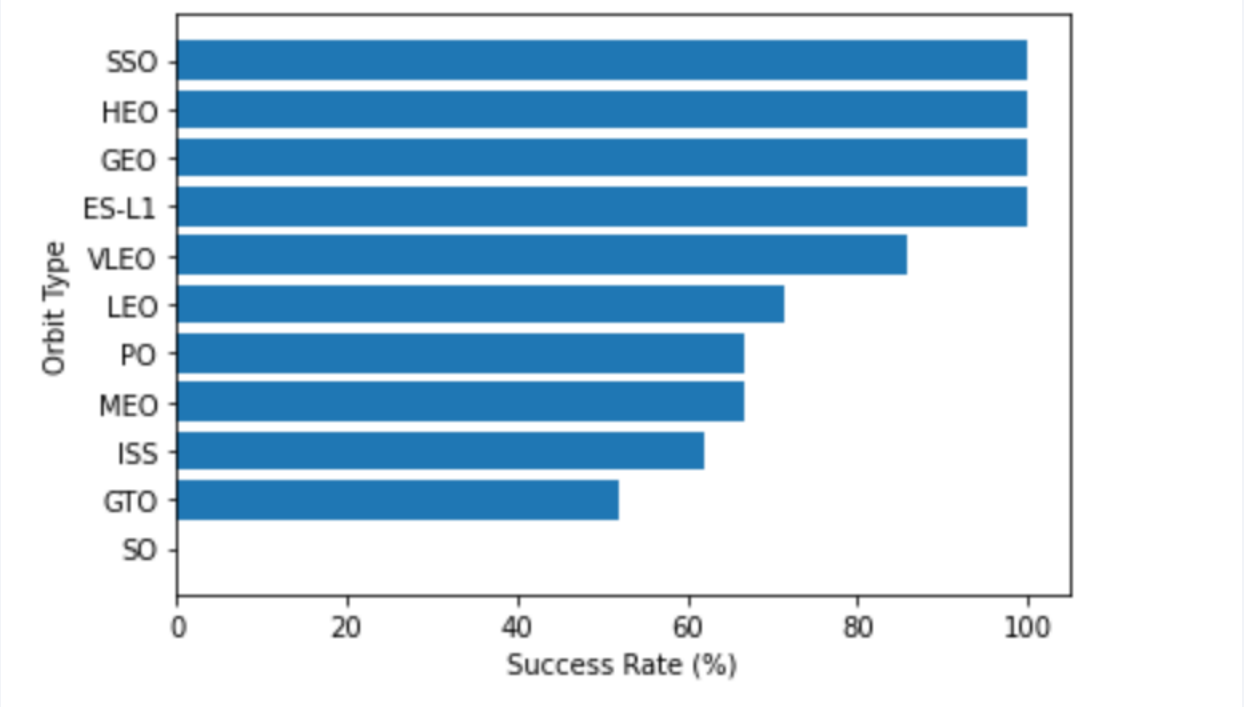
- `Class 0` (blue) represents unsuccessful launch, and `Class 1` (orange) represents successful launch.

- At first glance, the larger pay load mass, the higher the rocket's success rate, but it seems difficult to make decisions based on this figure because **no clear pattern can be found between successful launch and Pay Load Mass**.



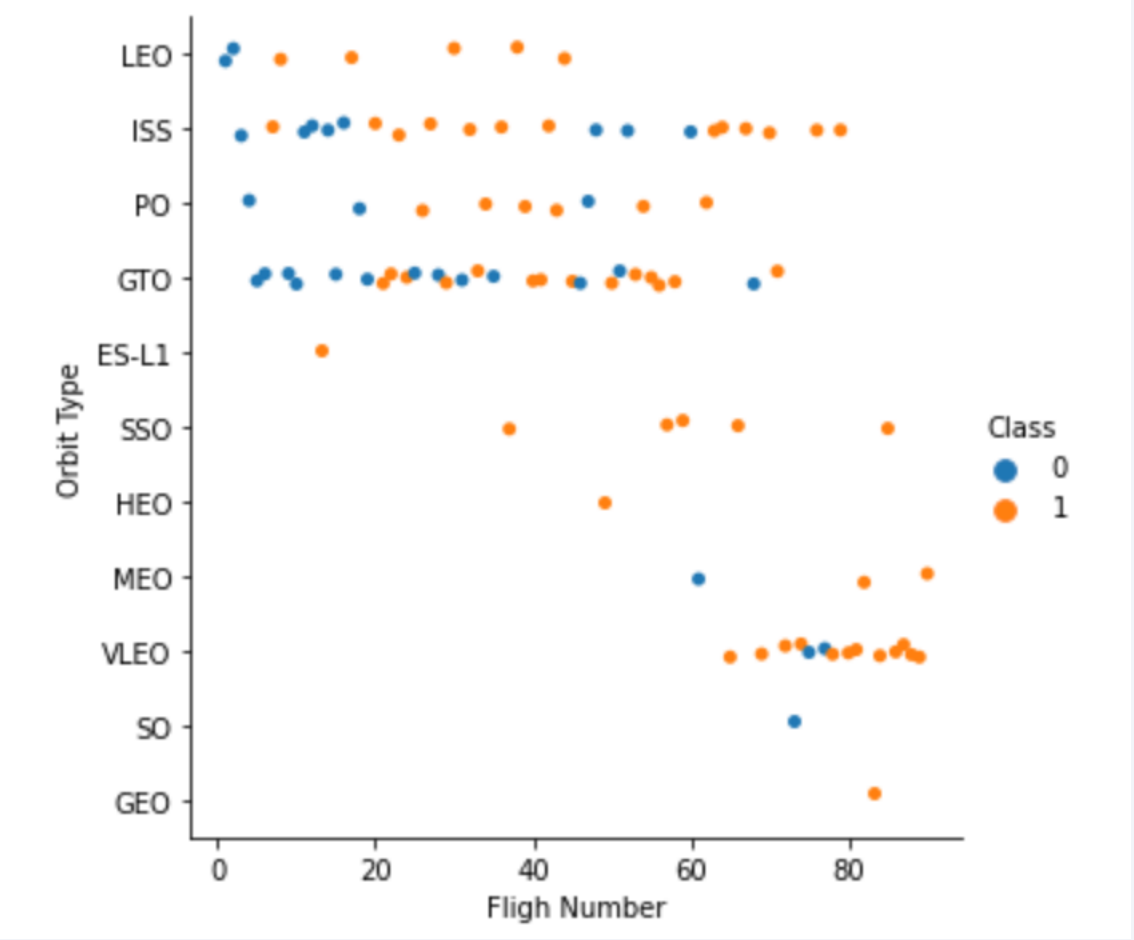
Success Rate vs. Orbit Type

- Orbit types **SSO**, **HEO**, **GEO**, and **ES-L1** have **the highest success rates** (*100%*).
- On the other hand, the success rate of orbit type **GTO** is only *50%*, and it is the **lowest** except for type SO, which recorded failure in a single attempt.



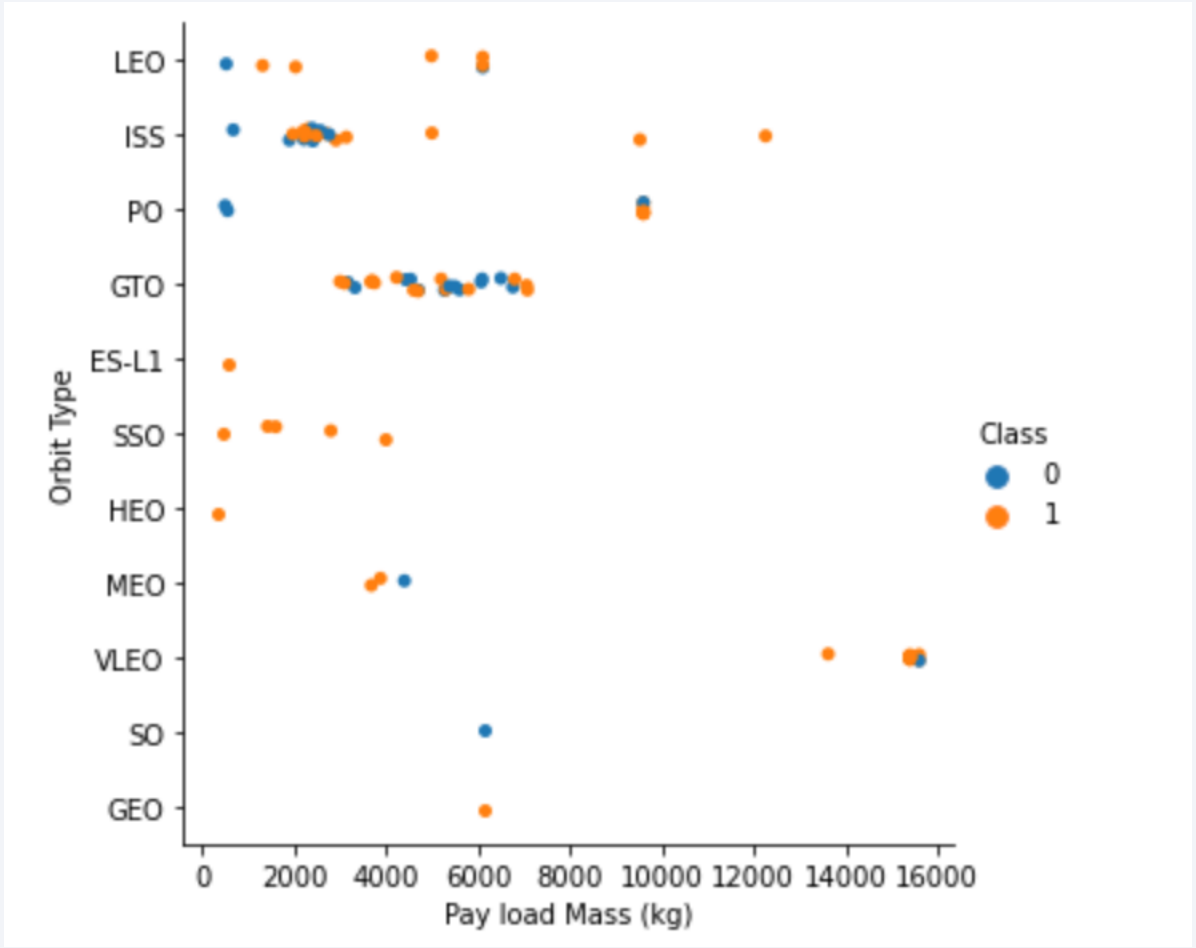
Flight Number vs. Orbit Type

- `Class 0` (blue) represents unsuccessful launch, and `Class 1` (orange) represents successful launch.
- In most cases, the launch outcome seems to be correlated with the flight number.
- On the other hand, in **GTO** orbit, there seems to be **no** relationship between flight numbers and success rate.
- SpaceX starts with LEO with a moderate success rate, and it seems that VLEO, which has a high success rate, is used the most in recent launches.



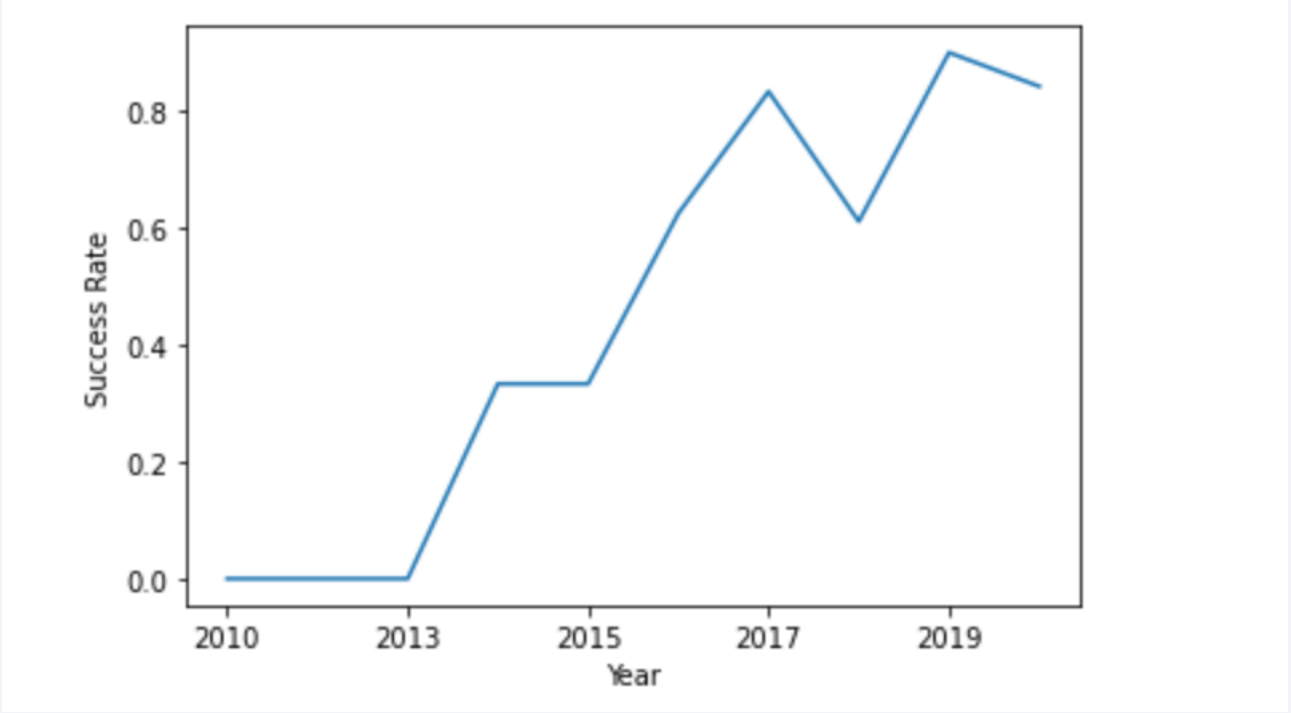
Payload vs. Orbit Type

- `Class 0` (blue) represents unsuccessful launch, and `Class 1` (orange) represents successful launch.
- With heavy payloads the successful landing or positive landing rate are more for LEO and ISS.
- However, in the case of GTO, it is hard to distinguish between the positive landing rate and the negative landing because they are all gathered together.



Launch Success Yearly Trend

- Since 2013, the success rate has continued to **increase** until 2017.
- The rate decreased slightly in 2018.
- Recently, it has shown a success rate of about *80%*.



All Launch Site Names

SQL Query:

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

Output of distinct launch sites:

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

Returns unique launch site names from the database.

Launch Site Names Begin with 'CCA'

SQL Query:

```
%%sql
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

Output:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Returns first five records in the database where launch site name begins with “CCA”.

Total Payload Mass

SQL Query:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

Output:

total_payload_mass_kg
45596

Returns the sum of payload mass (kg) for all records where the customer name is “NASA (CRS)”.

Average Payload Mass by F9 v1.1

SQL Query:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

Output:

avg_payload_mass_kg
2928

Returns the average of payload mass (kg) for records where the booster version name begins with “F9 v1.1”.

First Successful Ground Landing Date

SQL Query:

```
%%sql
SELECT MIN(DATE) AS first_successful_landing_date
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

Output:

first_successful_landing_date
2015-12-22

Returns the date of the first successful launch landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query:

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

Output:

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

Returns the booster versions of records which successfully landed a drone ship landing and where the payload was between 4000kg and 6000kg.

Total Number of Successful and Failure Mission Outcomes

SQL Query:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

Output:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Returns all types of mission outcome, and the number of occurrences. Most missions are
successful, and is not an indicator of whether or not the landing of the first stage was successful.

Boosters Carried Maximum Payload

SQL Query:

```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

Returns the names of the booster versions which have carried the maximum payload mass.

Output:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

SQL Query:

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = '2015'
```

Output:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Returns the month, landing outcome, booster version, payload mass (kg) and launch site of 2015 launches which failed to land a drone ship landing.

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

SQL Query:

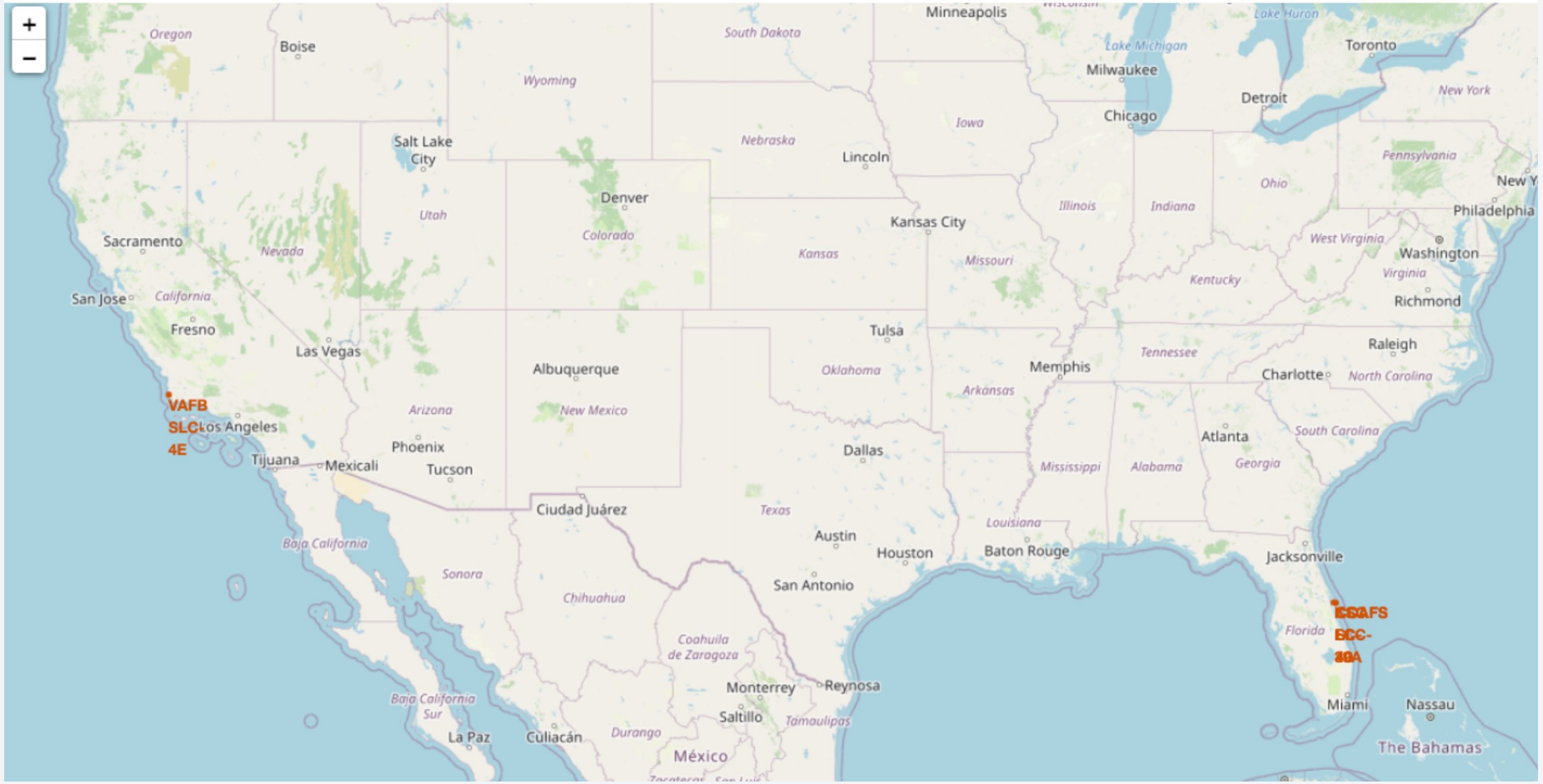
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

Output:

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

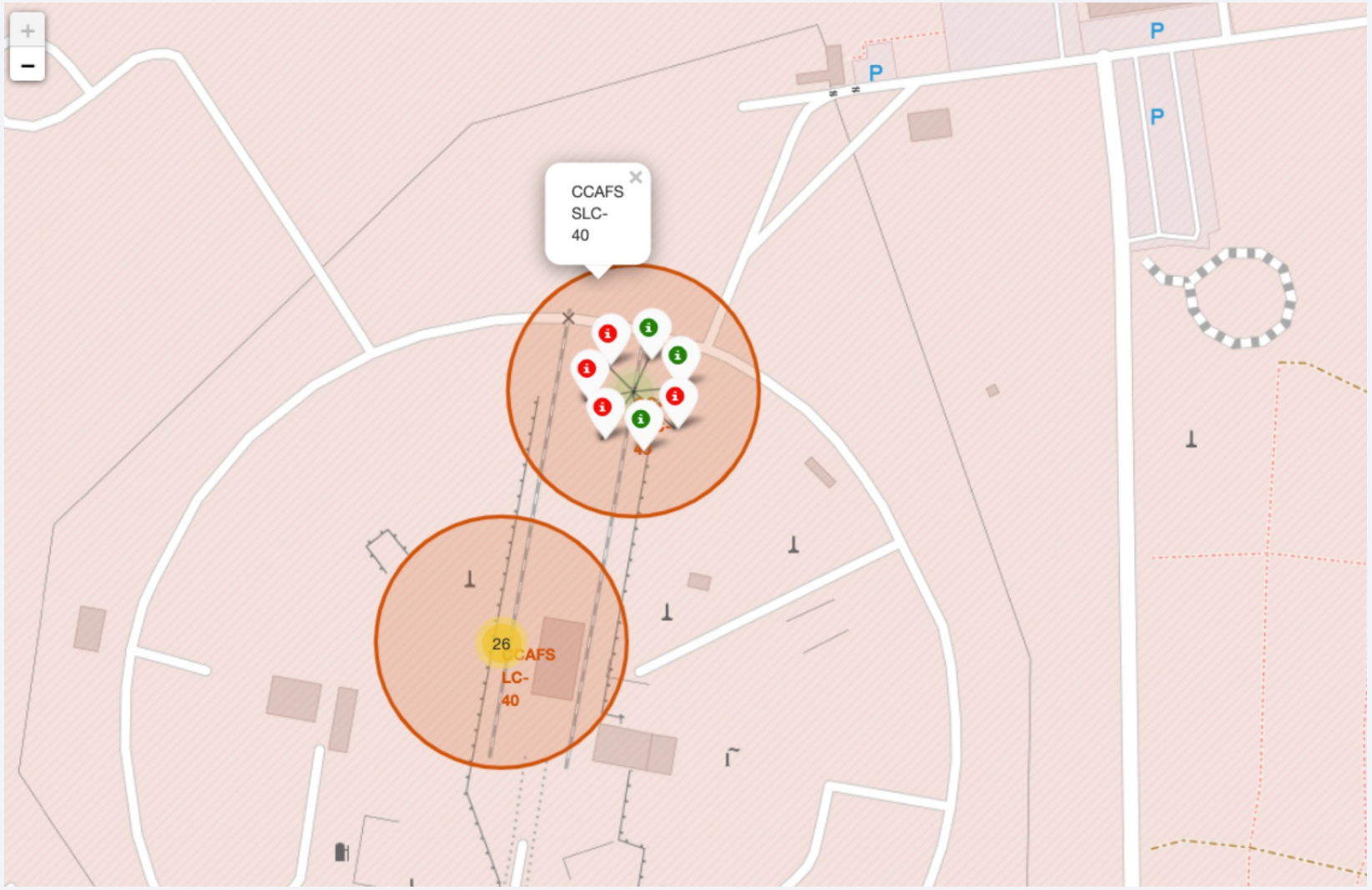

Section 3
Launch Sites
Proximities Analysis

All Launch Sites on a map



We can see that all launch sites are located in North America and that all launch sites are located near to coastlines, specifically the coasts of Florida and California.

Color-labeled Launch Outcomes

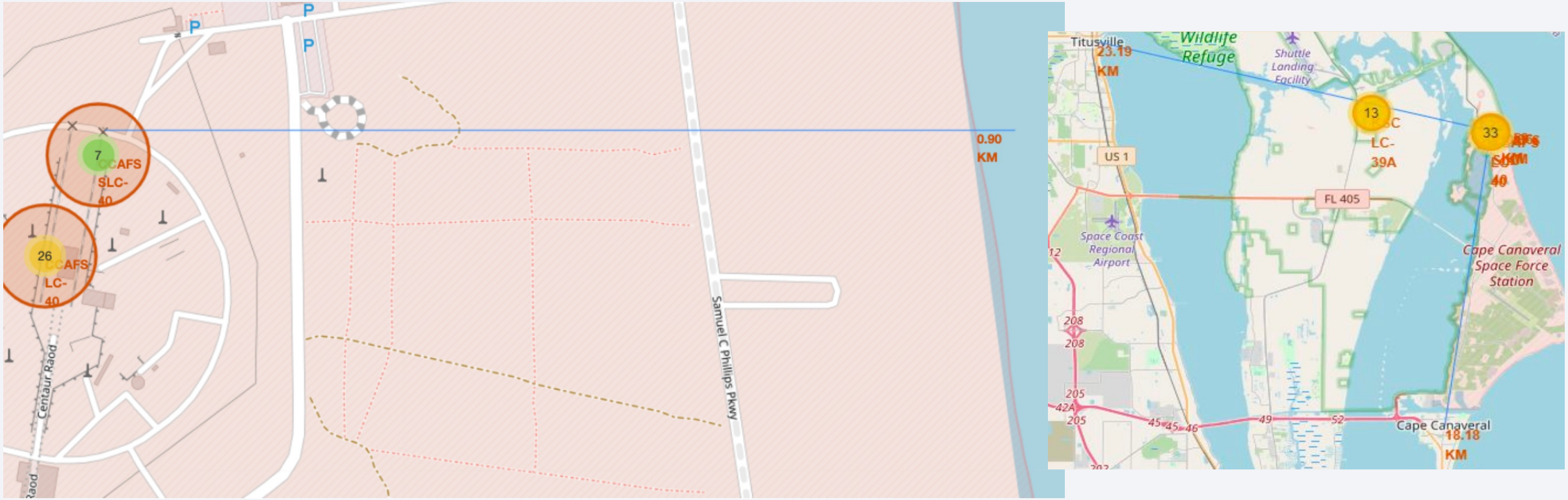


The map indicates the launch outcomes at the location CCAFS SLC-40.

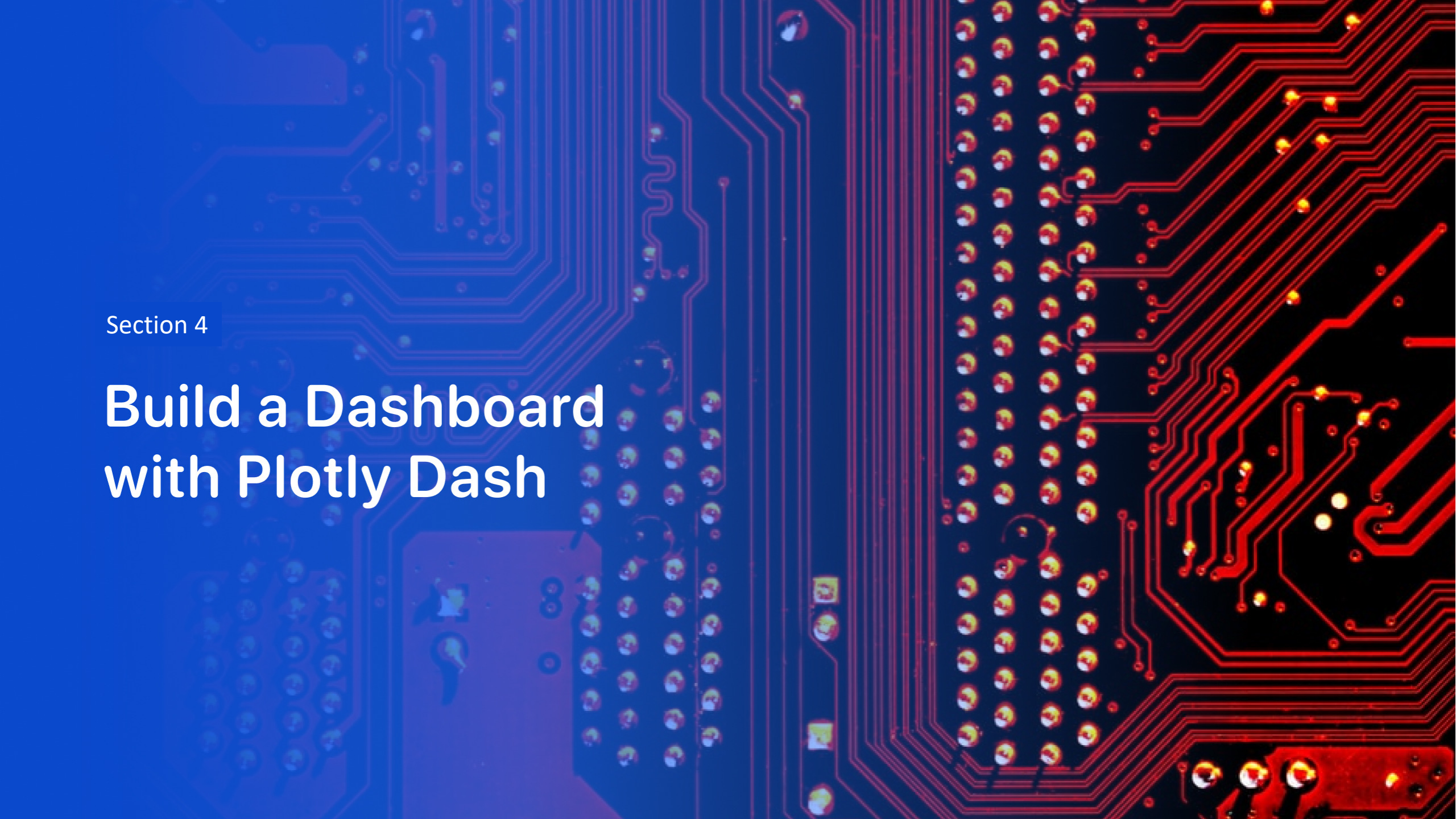
Green- Indicates successful landings
Red- Indicates unsuccessful landings

We can further drill down on the map by clicking on the cluster.

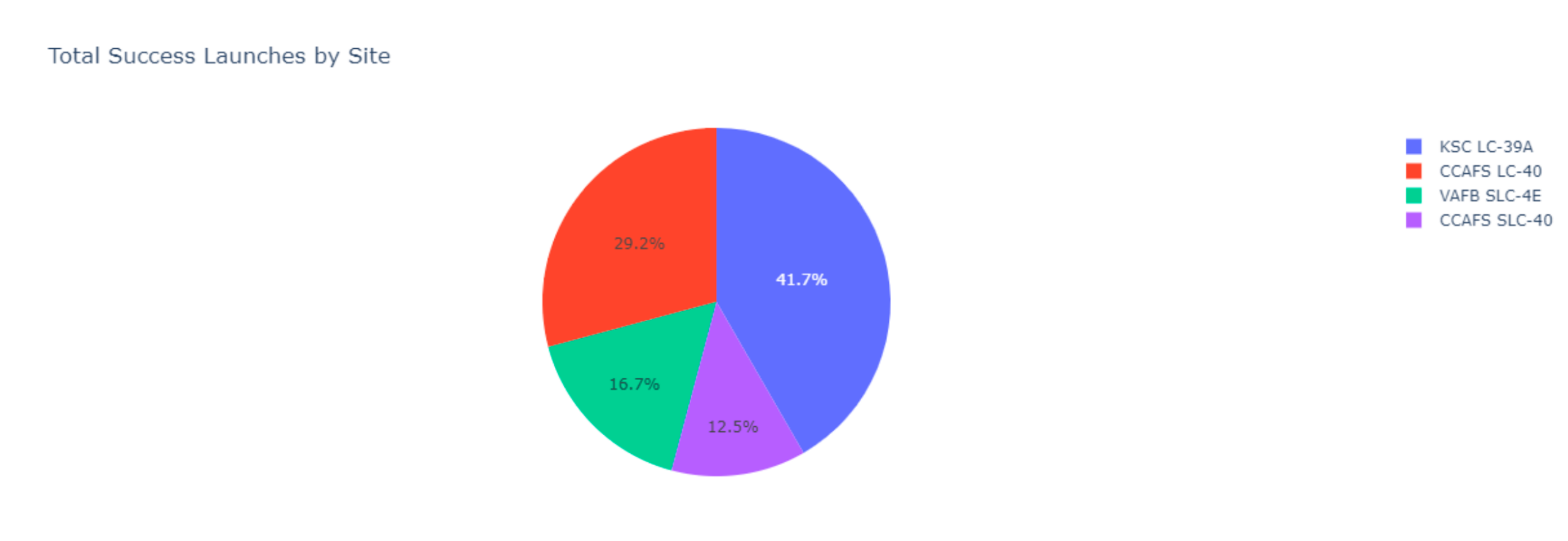
Proximities of Launch Sites



It can be found that the launch site is **close to railways and highways** for transportation of equipment or personnel, and is also **close to coastline** and relatively **far from the cities** so that launch failure does not pose a threat.

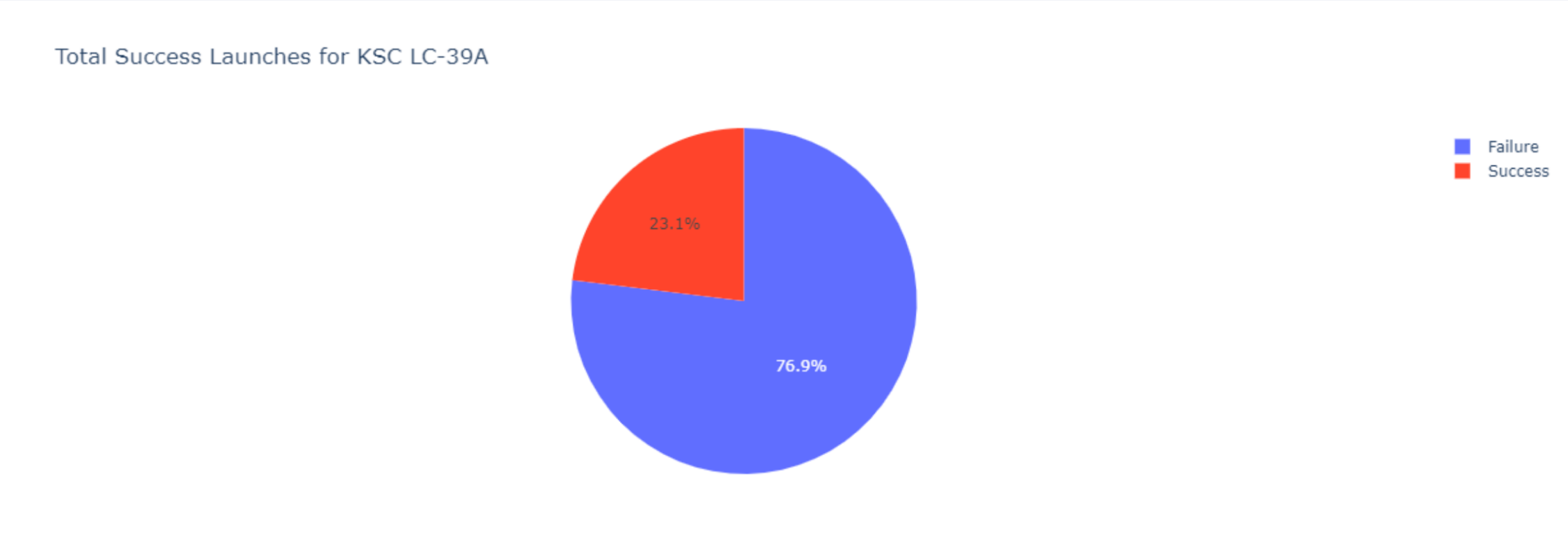

Section 4
Build a Dashboard with Plotly Dash

Successful Stage 1 Landings By Launch Site



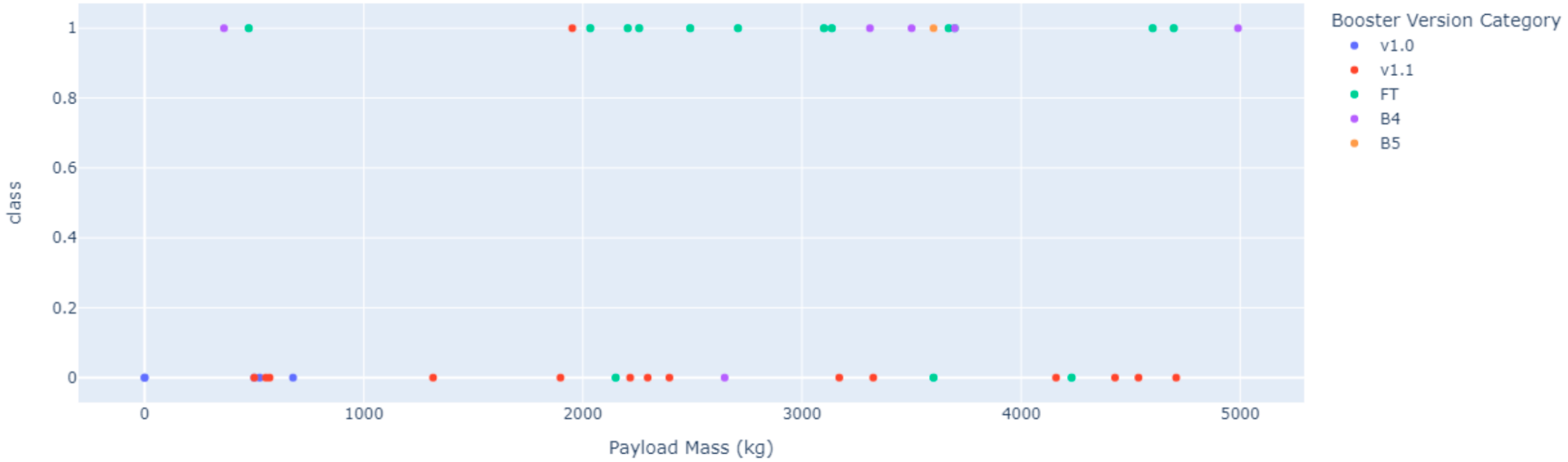
We can see that most successful landings were launches from KSC LC-39A.
The least successful landings were launches from CCAFS SLC-40.

Launch Site with Highest Launch Success Ratio



KSLC-39A has the highest success rate with 10 landing successes *(76.9%)* and 3 landing failures *(23.1%)*.

Payload vs. Launch Outcome Scatter Plot

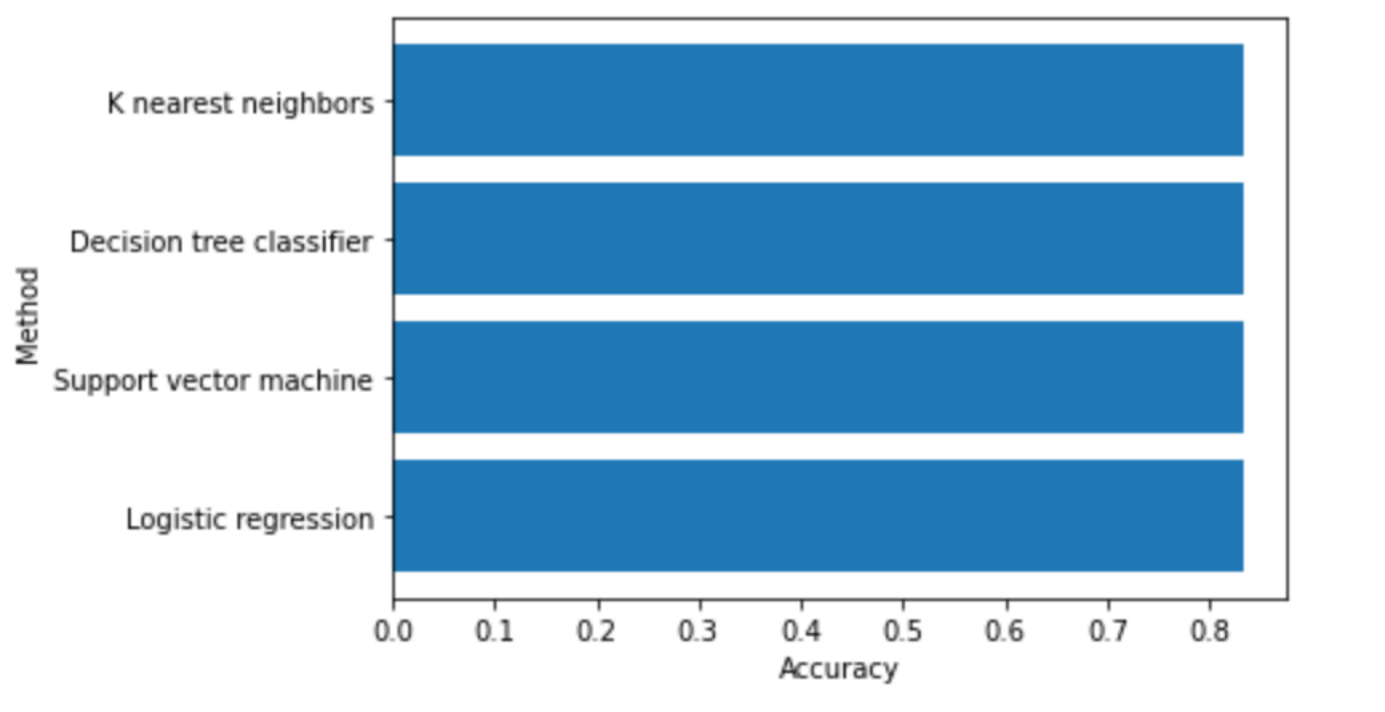


We can see that booster version category FT has many successes and few failures. In contrast, v1.1 has many failures and few successes. There are recorded failures with payload mass of 0kg, this may be a data entry error.

Section 5

Predictive Analysis (Classification)

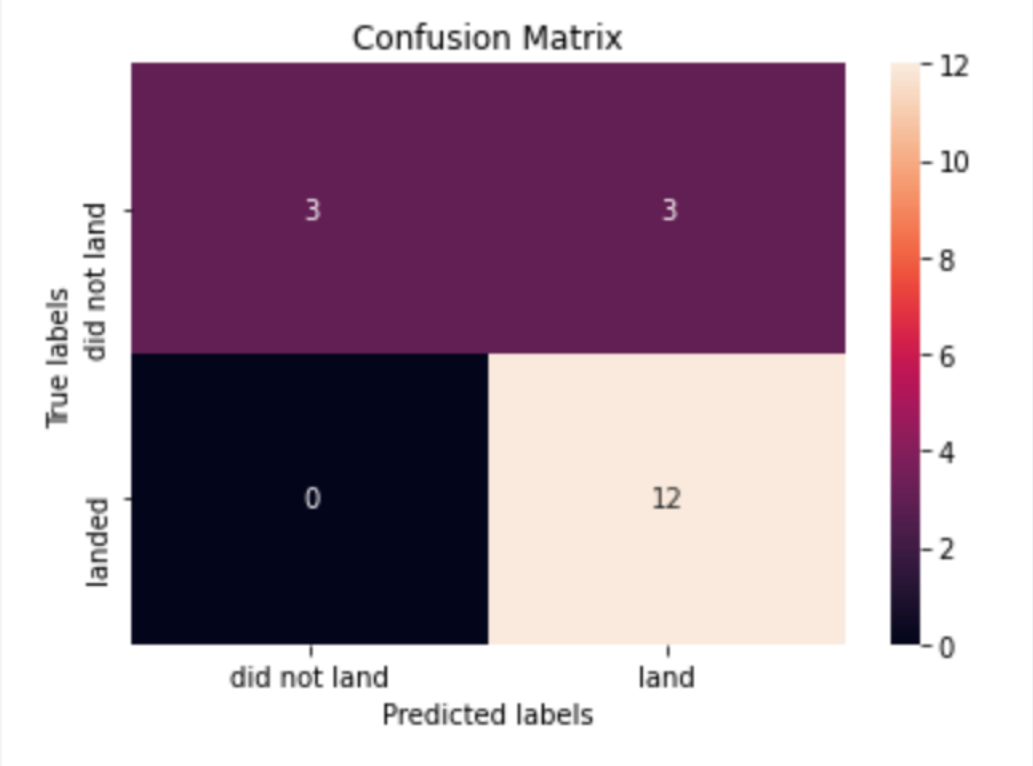
Classification Accuracy



- In the test set, **the accuracy of all models** was virtually the **same** at **83.33%**.
- It should be noted that the test size was small at 18. Therefore, more data is needed to determine the optimal model.

Confusion Matrix

- The confusion matrix is the same for all models because all models performed the same for the test set.
- The models predicted 12 successful landings when the true label was successful and 3 failed landings when the true label was failure. But there were also 3 predictions that said successful landings when the true label was failure *(false positive)*.
- Overall, **these models predict successful landings**.



Conclusions

- Our goal was to develop a machine learning model to predict if stage 1 will successfully land for a given launch.
- We developed four machine learning models, which all predicted successful landings with ~83.33% accuracy for some test data. The models tend to over predict successful landings, the models could be improved by using more data.
- In addition, we found that:

  -Success rate of stage 1 landings has improved over time.

  -ES-L1, GEO, HEO, SSO orbits have the best success rate. •

  -Launch sites are typically located close to coastlines.

Appendix

- GitHub URL
- Coursera IBM Data Science Specialisation URL

Thank you!