Winning Space Race with Data Science

Jonathan Hayes
September 27th, 2023

Outline

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

Executive Summary

Summary of methodologies

1. Data collection with API
2. Data collection through web scraping
3. Data Wrangling
4. Exploratory Data Analysis using SQL
5. Exploratory Data Analysis using Visualization
6. Interactive Visual Analytics and Dashboarding with Folium and Plotly Dash
7. Predictive Modeling using Machine Learning

Summary of all results

1. Exploratory Data Analysis Results
2. Visualization Results
3. Predictive Analysis Results

Introduction

Project background and context

SpaceX advertises the Falcon 9 Rocket Launch with a price tag of 62 million dollars; other providers cost upward of 165 million dollars each. These savings are derived from the fact that SpaceX can reuse the first stage of their rocket launch. Therefore, if we can determine if the first stage will land successfully, we can determine the overall cost of a launch, information which can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems you want to find answers

1. What factors determine whether the first stage lands successfully?
2. With these factors taken into consideration, what is the overall success/failure rate of Falcon 9 Rocket Launches?
3. How does this success rate affect the overall cost of Falcon 9 launches?
4. Can we develop a predictive modeling method that accurately predicts if a certain launch will be a success or a failure?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Data was collected through the SpaceX API and via web scraping from Wikipedia.

Perform data wrangling

- The data was prepared for analysis through filtering, the handling of null values, and the application of one-hot encoding to categorical variables.

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

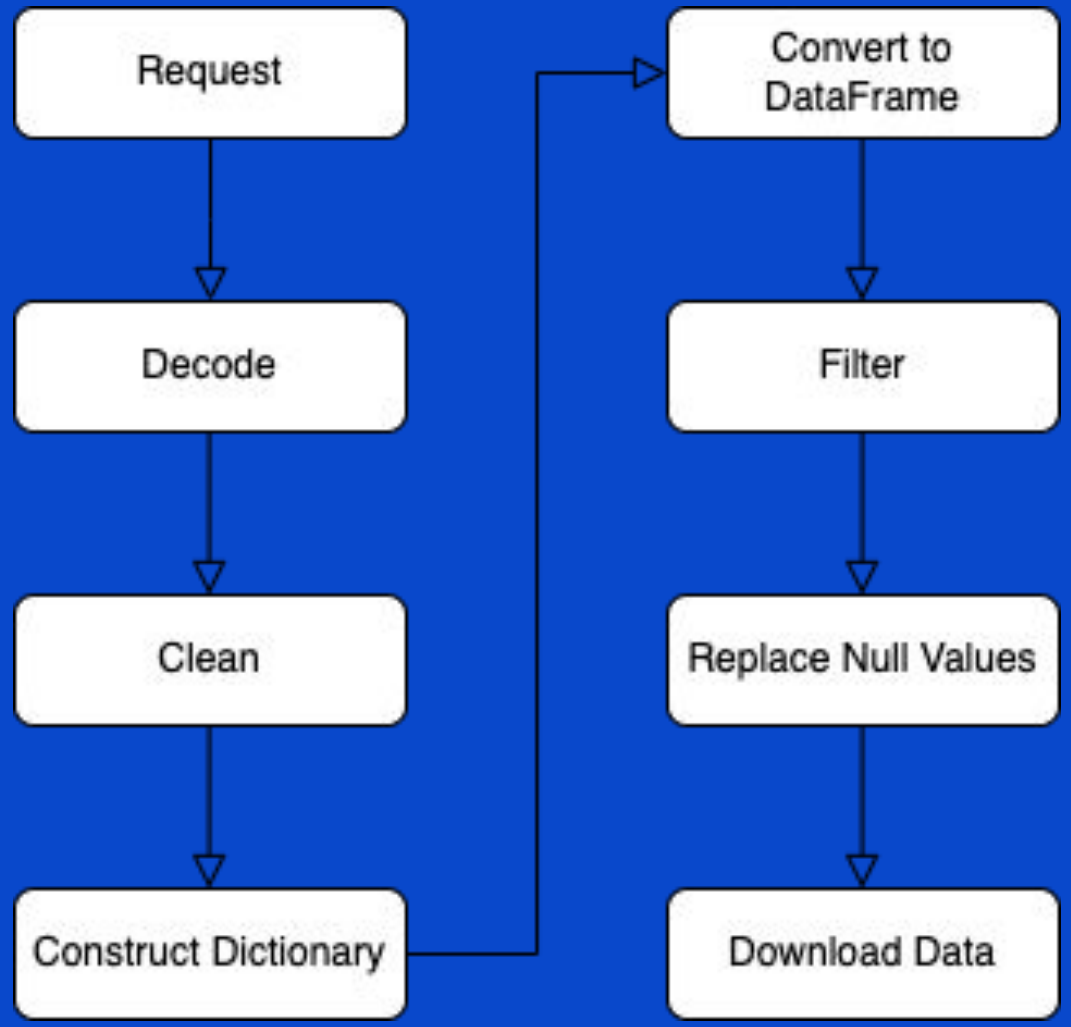
Perform predictive analysis using classification models

- A variety of classification models were used, and their hyper-parameters were tuned using Grid Search cross validation. These models were then evaluated through various accuracy scores.

Data Collection – SpaceX API

Steps:

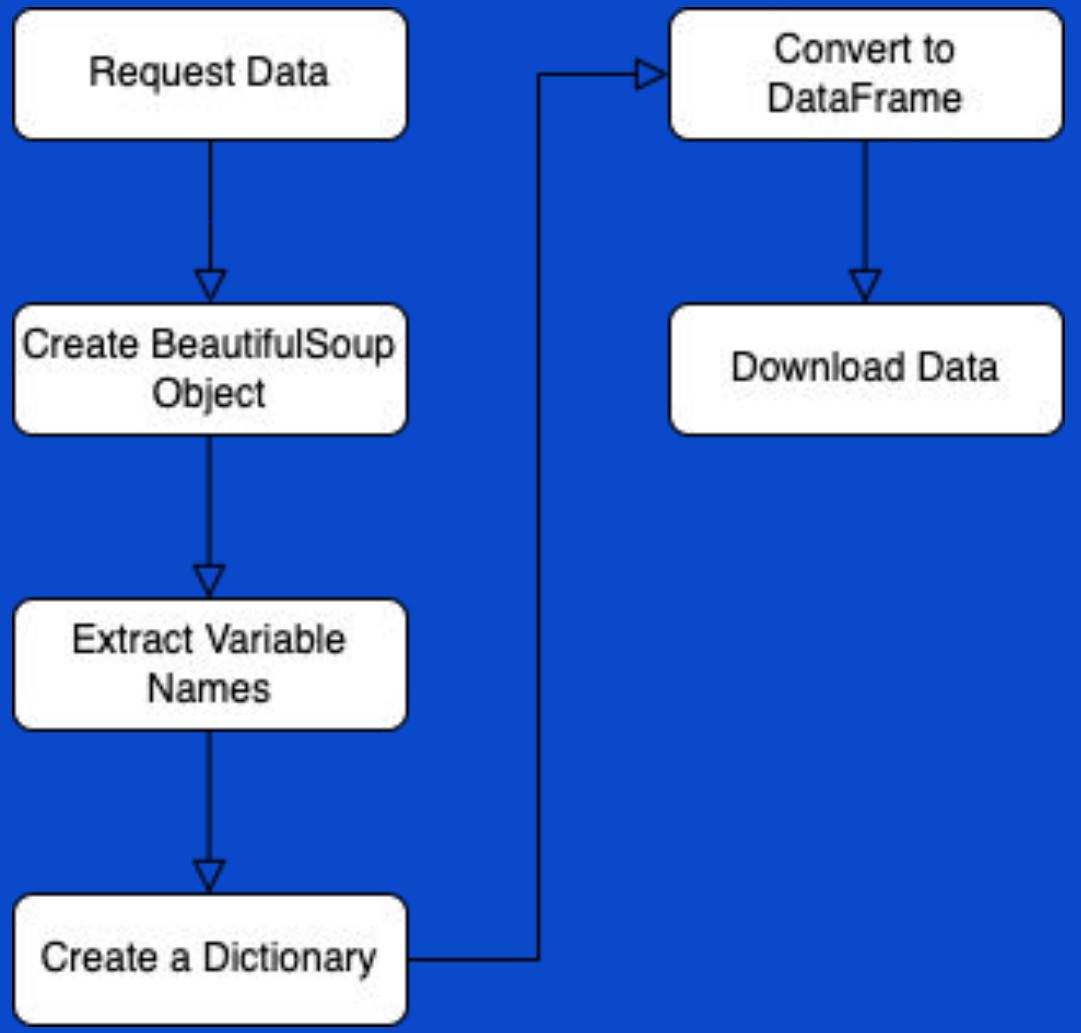
1. **Request** rocket launch data from the SpaceX API
2. **Decode** the response content from json data into a Pandas Dataframe
3. **Clean/Clarify** the dataframe by accessing the API to get more impactful column data
4. **Construct** a dictionary object from this data
5. **Convert** the dictionary object into a new dataframe
6. **Filter** the data to include only Falcon 9 Launches
7. **Replace** null Payload Mass values with the column mean
8. **Download** the data via a CSV file



Data Collection - Scraping

Steps:

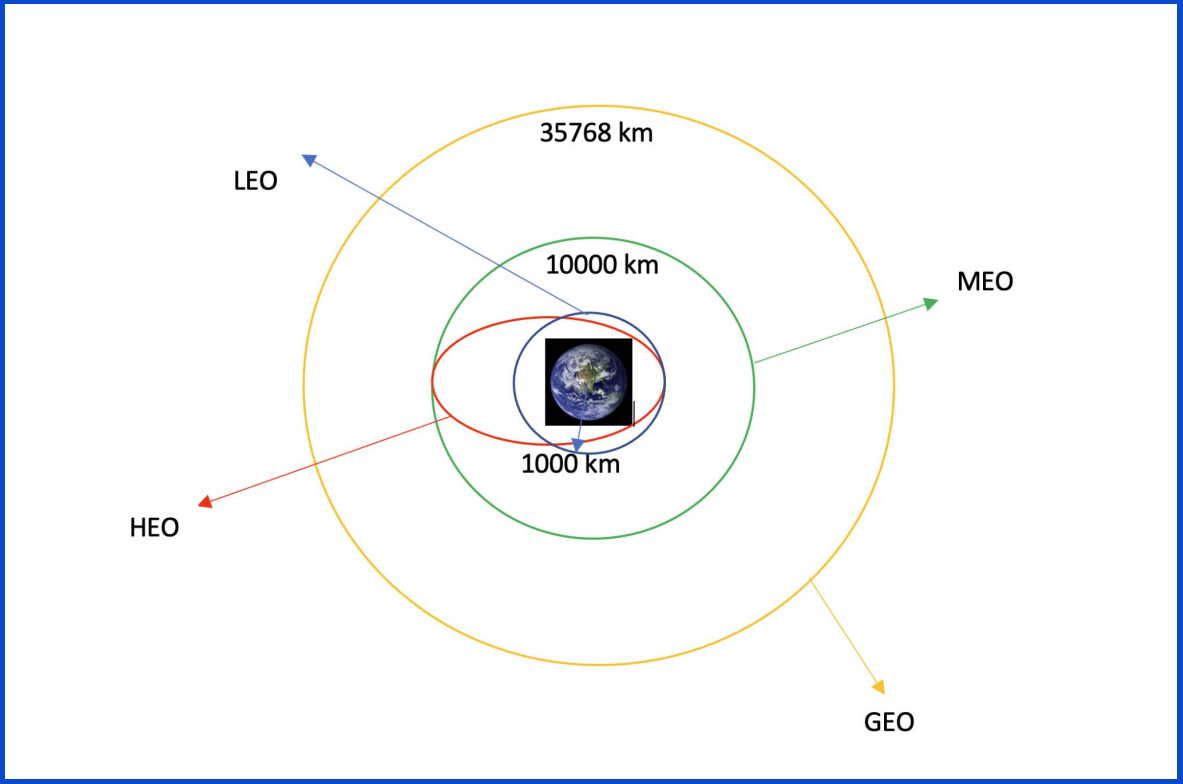
1. **Request** Falcon 9 launch data from Wikipedia
2. **Create** a BeautifulSoup object from the HTML response
3. **Extract** all column/variable names from the HTML table header
4. **Create** a dictionary by parsing the launch HTML tables
5. **Convert** dictionary to a dataframe
6. **Download** the dataframe as a CSV file



Data Wrangling

Key Data Wrangling and EDA Processes:

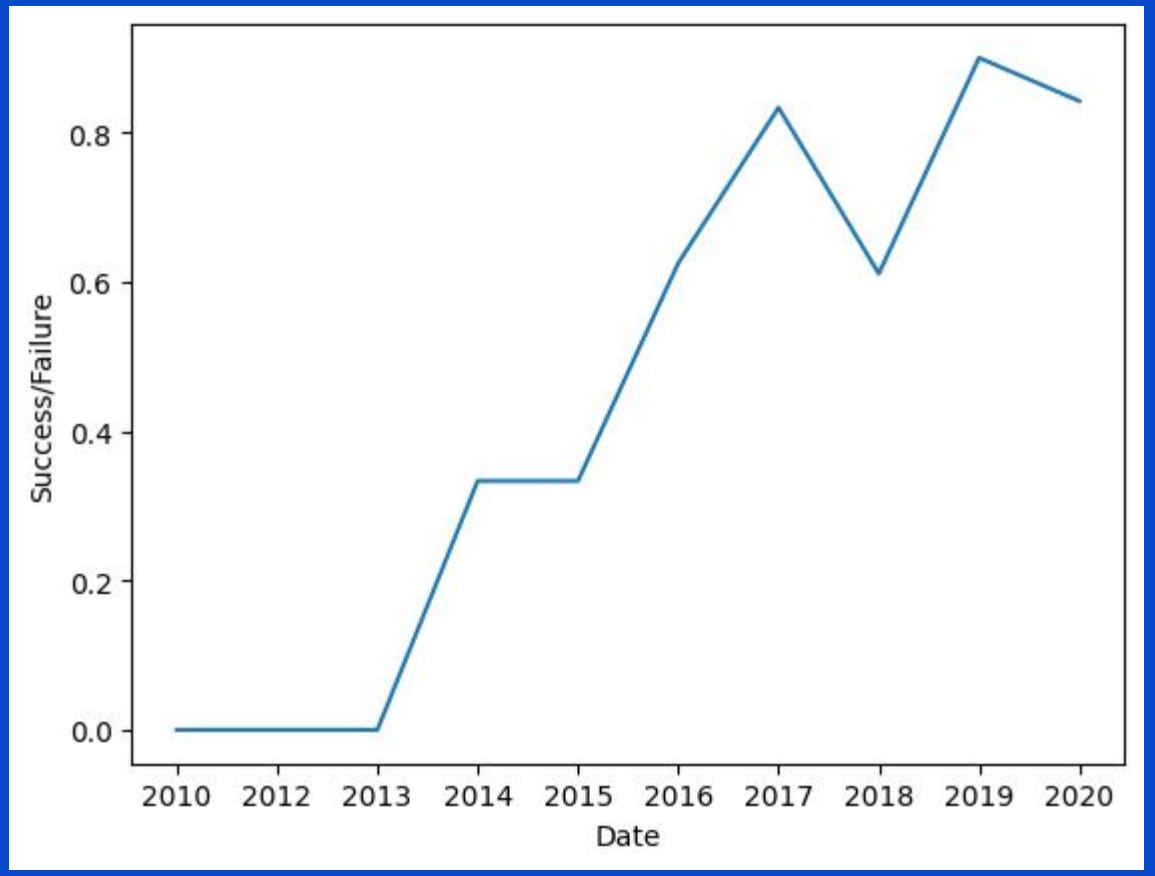
1. Calculate the Number of Launches for each Site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurrence of mission outcome per orbit type
4. Create a landing outcome label from the outcome column
5. Convert these outcomes into binary variables for success and failure



EDA with Data Visualization

Charts that were Plotted:

- **Flight Number vs. Payload Mass (scatter):** does payload mass change over time?
- **Flight Number vs. Launch Site (scatter):** which flights were launched from what locations?
- **Payload Mass (kg) vs. Launch Site (scatter):** does payload mass differ per launch site?
- **Orbit Type Mean Success Rate (bar):** are some orbits more successful than others?
- **Flight Number vs. Orbit Type (scatter):** does orbit type differ among flight numbers?
- **Payload Mass (kg) vs. Orbit Type (scatter):** do different orbit types take different mass levels?
- **Launch Success Yearly Trend (time-series line):** has landing success rate changed over time?



EDA with SQL

SQL Queries Performed:

- The names of the unique launch sites
- 5 records where launch sites begin with the string ‘CCA’
- The total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing on ground pad
- Names of the boosters which had success in drone ship landings and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure outcomes
- The names of the booster versions which have carried the maximum payload mass
- The records which display the months, failure outcomes in drone ships, booster versions, and launch sites for the months in year 2015
- Descending ranked count of landing outcomes between the date 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

Mark all launch sites on the map

- Use longitude and latitude coordinates to plot each launch site on a map with circle markers
- This allows us to easily visualize where each launch is taking place

Mark the success/failed launches for each site

- Create a new variable for launch outcome (green = success, red = failure)
- Using marker_cluster function, plot the successful (green) and unsuccessful (red) landing outcomes at each launch site, and denote the count of each
- This allows us to easily see which sites have a higher/lower success rate

Calculate the distances between launch sites and proximities

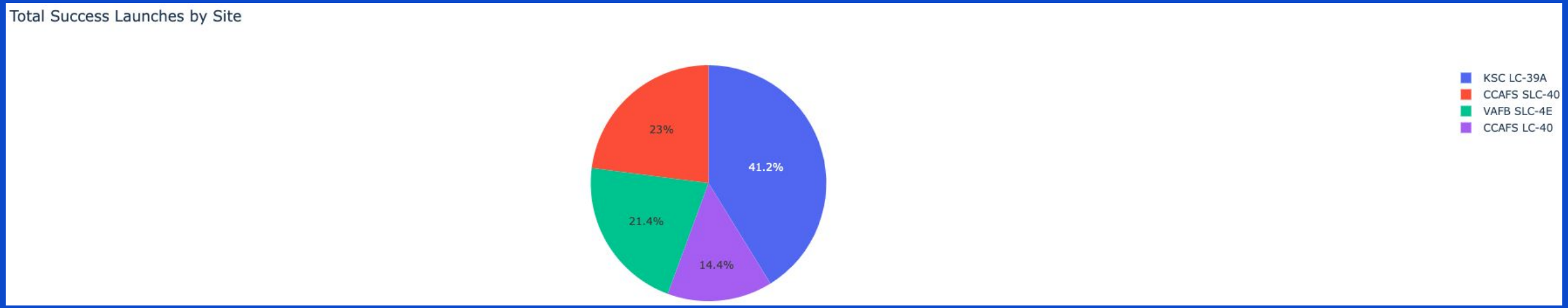
- Find the distance between a launch site and its nearest coastline, and plot a line displaying this distance.
- Find the distance from this launch site to the nearest highway, railroad, and city.
- This allows us to easily visualize how close the launch site is to other notable features

Build a Dashboard with Plotly Dash

Synopsis:

Utilizing Plotly Dash, we created a dashboard displaying the following charts/interactions

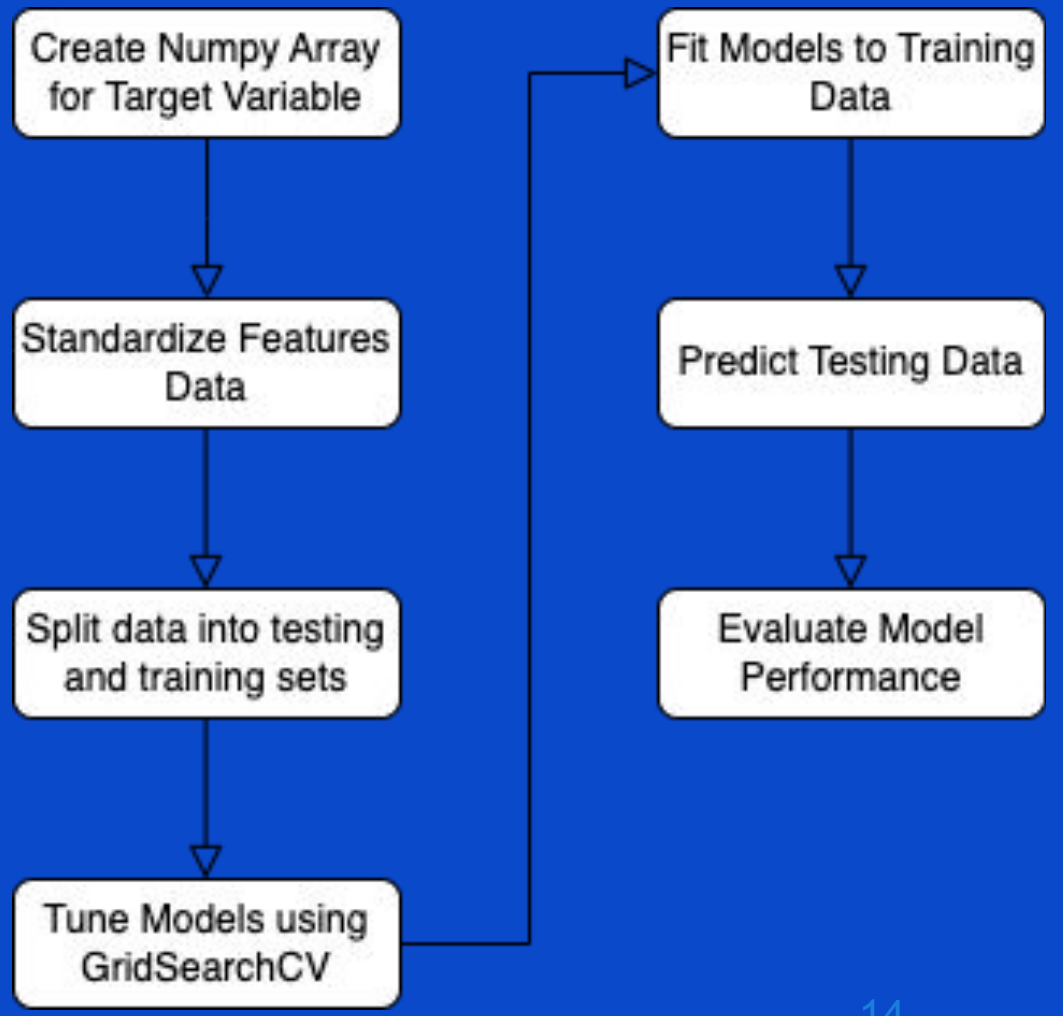
1. An interactive pie chart showing launch success rate per launch site
2. A scatter plot showing the correlation between payload mass and success rate
   a. An interactive mass slider to manipulate the masses considered in the above scatter plot.
   b. This allows us to better judge the relationship between payload mass and success rate.



Predictive Analysis (Classification)

Process:

- **Create** a Numpy Array from the target variable (Y)
- **Standardize** the data in the features dataset using StandardScaler function
- **Split** the data into testing and training sets using train_test_split function
- **Tune** models using GridSearchCV to find the optimal hyperparameters
- **Apply** GridSearchCV parameter selection to a variety of different models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
- **Fit** each model to the training data
- **Predict** the testing data
- **Evaluate** model performance using accuracy scores and confusion matrices.



Results

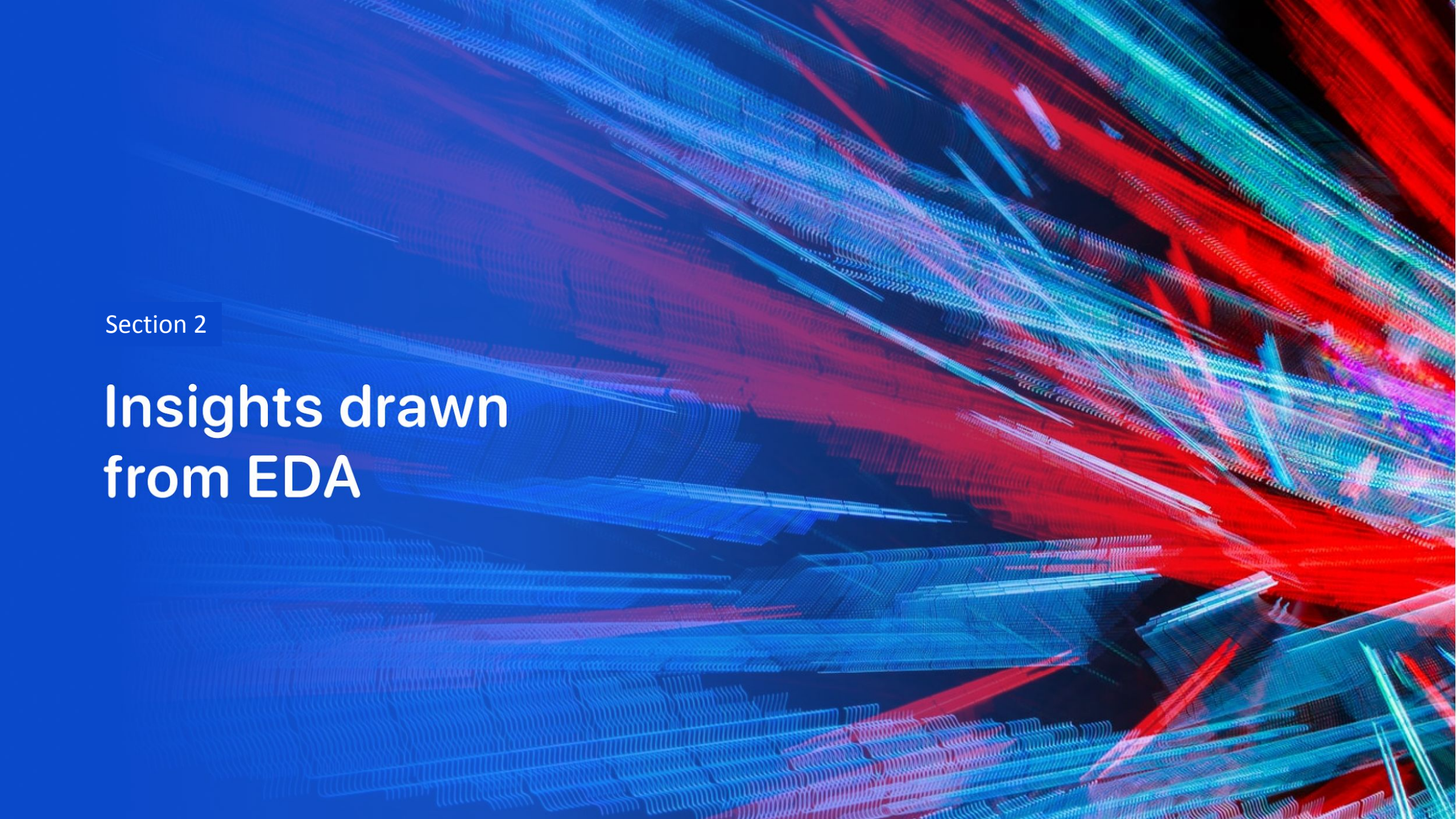
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

Section 2

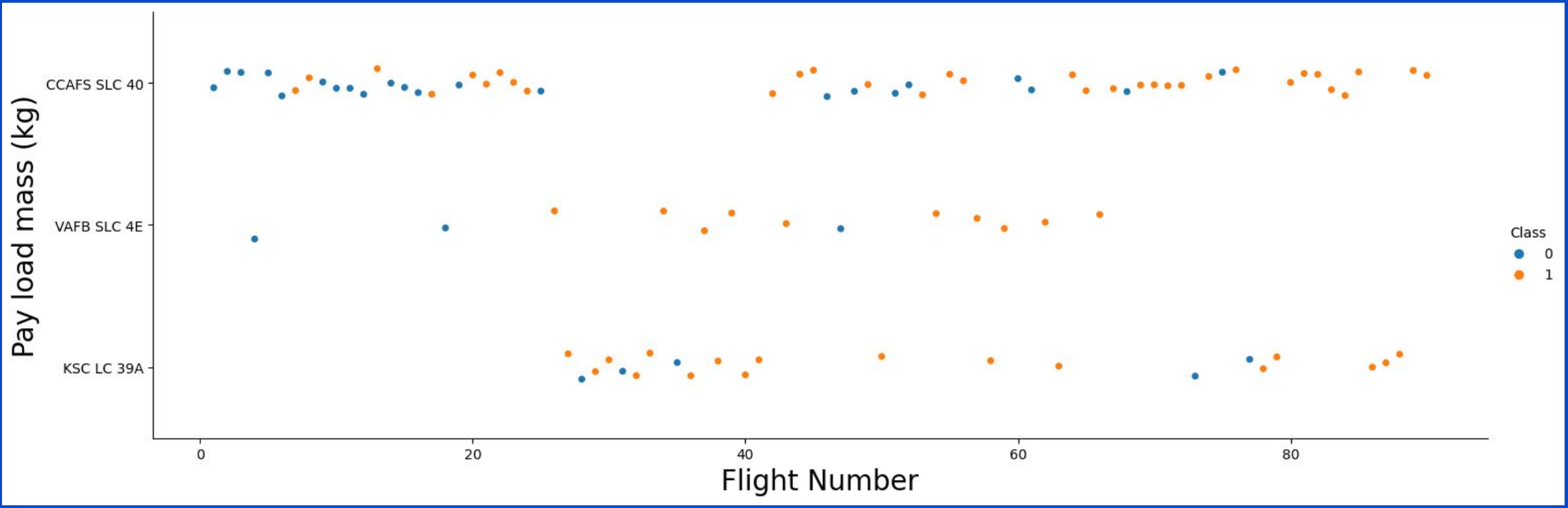
Insights drawn from EDA



Flight Number vs. Launch Site

Insights:

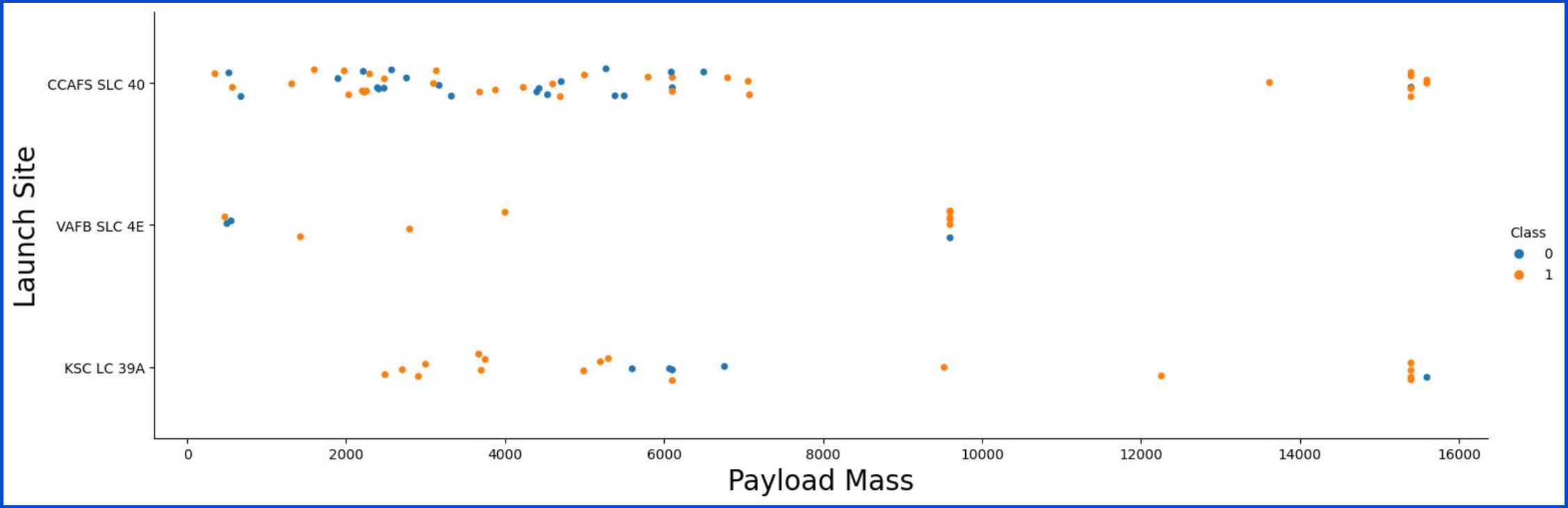
- Later flights had a higher success rate (indicated by orange circles)
- The majority of the launches were from site CCAFS SLC 40



Payload vs. Launch Site

Insights:

- Payloads with higher masses seemed to have a higher success rate
- All of the payloads with a mass greater than ~10,000 were launched from either CCAFS SLC 40 or KSC LC 39A



Success Rate vs. Orbit Type

Insights:

100% Success Rate:

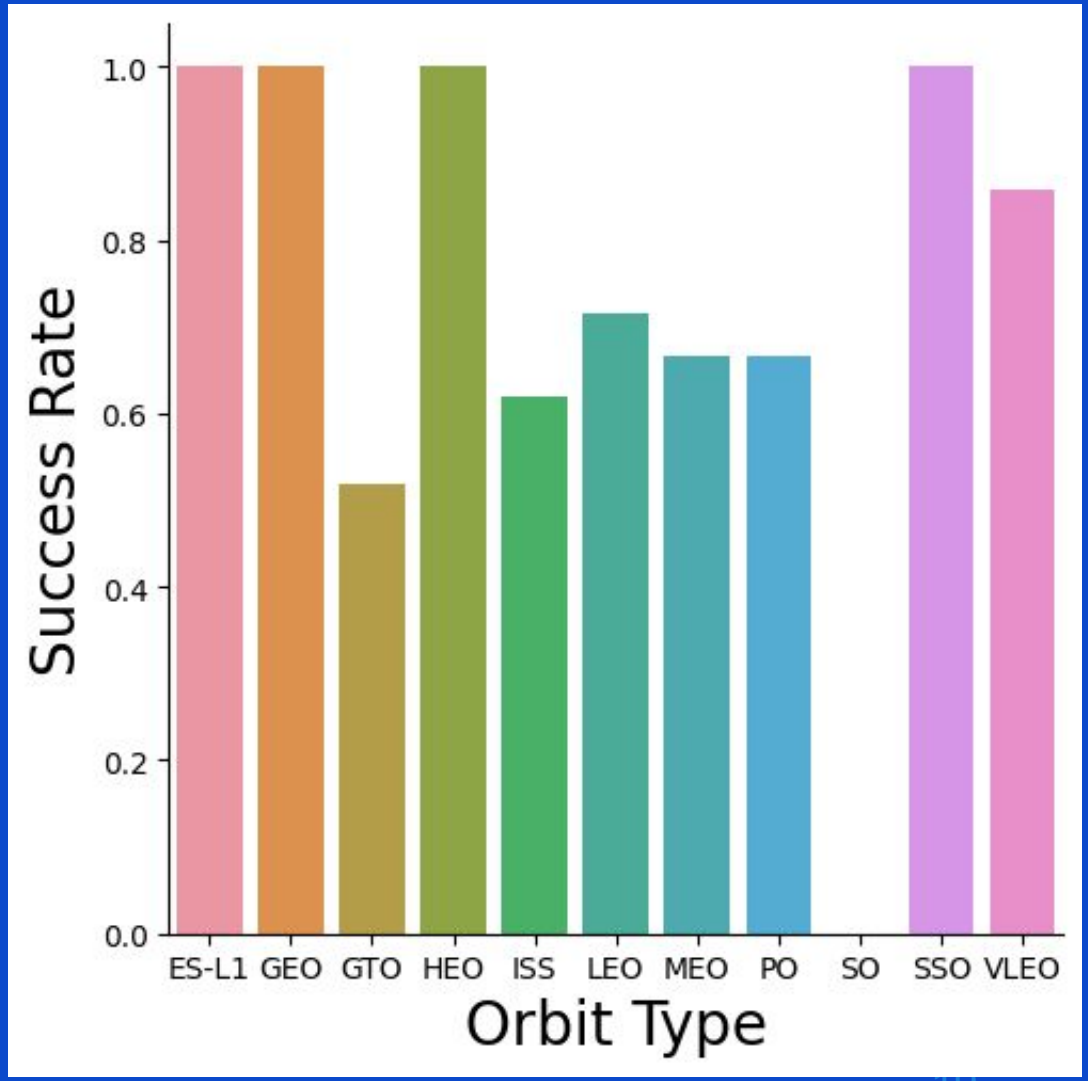
- ES-L1
- GEO
- HEO
- SSO

50 to 80% Success Rate:

- GTO
- ISS
- LEO
- MEO
- PO
- VLEO

0% Success Rate:

- SO



Flight Number vs. Orbit Type

Insights:

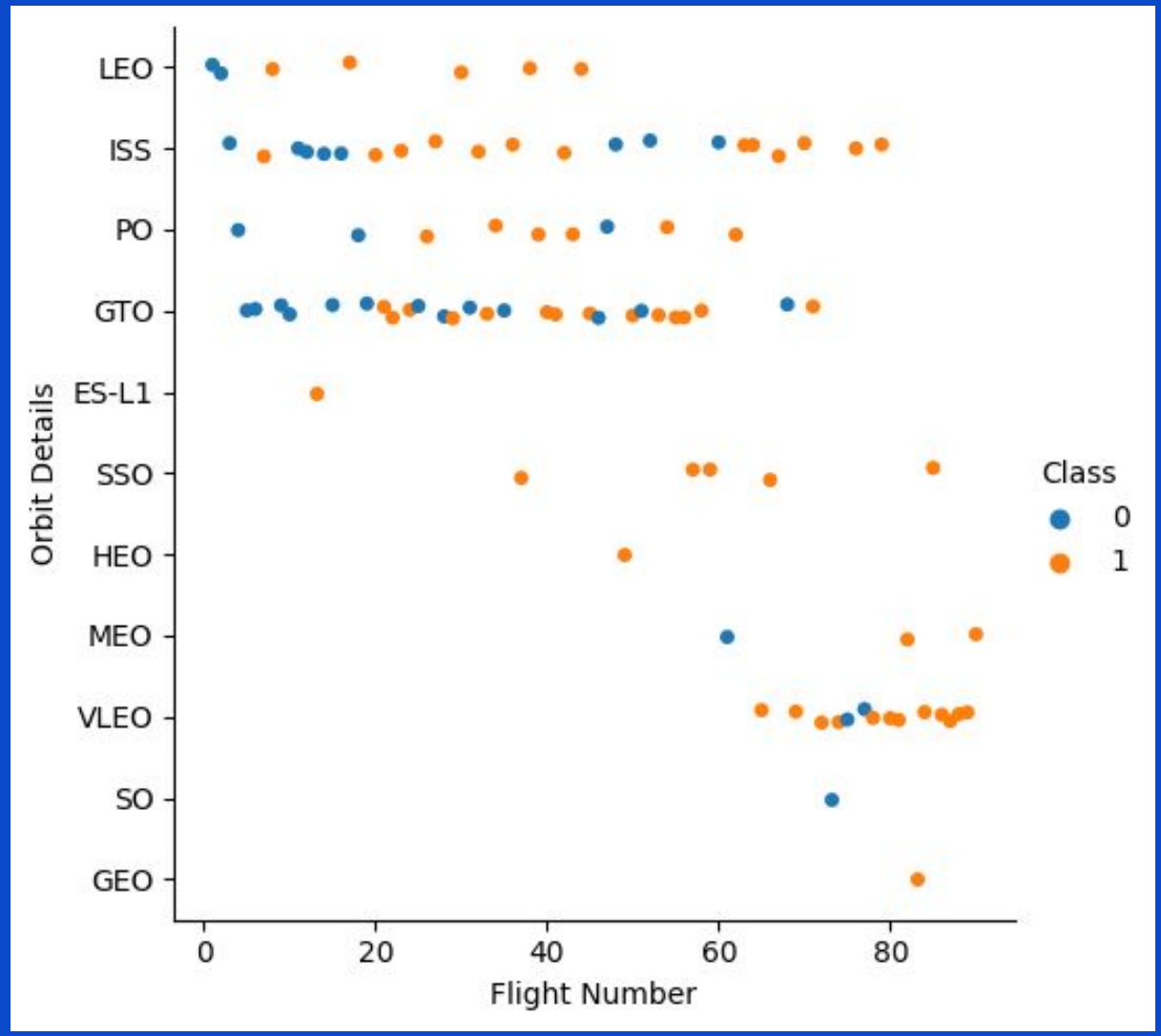
Earlier flights were heavily focused on orbit types LEO, ISS, PO, and GTO.

These orbit types also have the highest number of launches

For these orbit types, the there is a trend towards higher numbers of success as flight number increases

Flights in ES-L1, SSO, HEO, MEO, VLEO, SO, and GEO begin appearing significantly after flight #40

Although more sparsely populated, these sites boast higher rates of success



Payload vs. Orbit Type

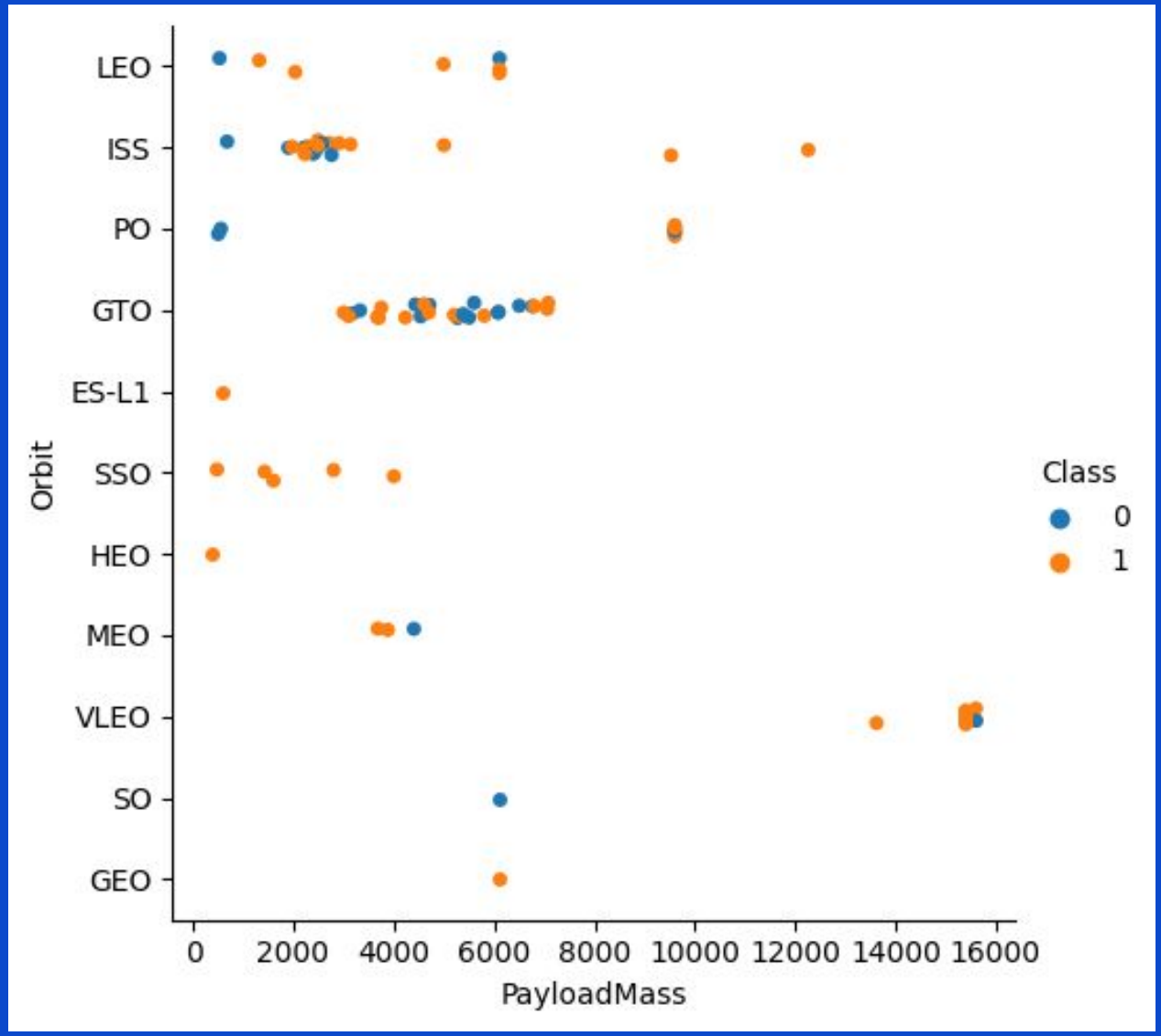
Insights:

Orbit type VLEO generally is matched with taking the highest payload mass

Orbit type GTO has exclusively taken payload masses between ~2500 and ~8000

Orbit type SSO is the only orbit that has taken multiple payloads with a 100% success rate

Multiple orbit types have only taken one payload

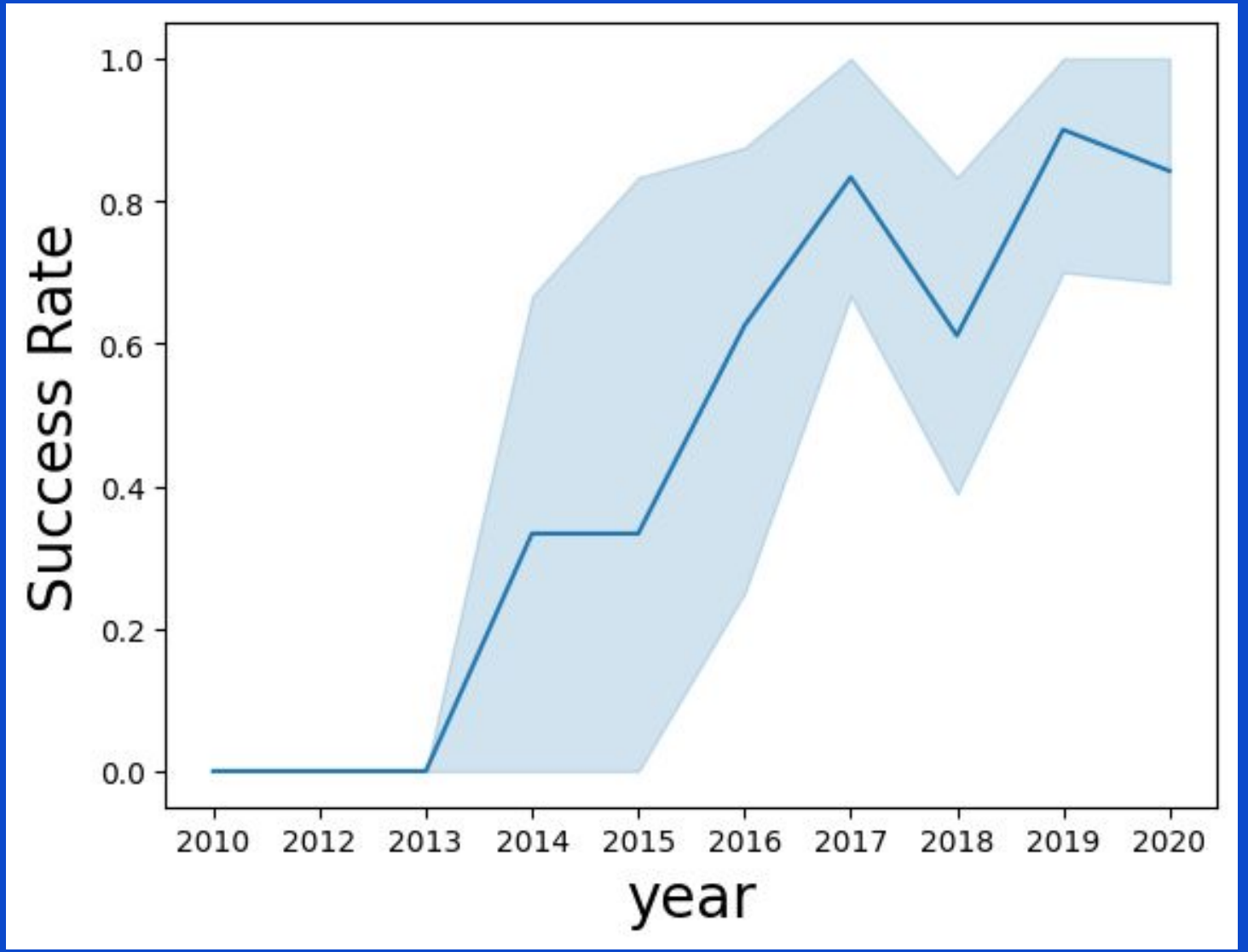


Launch Success Yearly Trend

Insights:

The success rate of launches has improved drastically over time (0% in 2010 to 80% in 2020)

This is despite success rate decreases from 2017 - 2018, and 2019 - 2020.



All Launch Site Names

Synopsis:

The query to the right returns the distinct names of the launch sites within the data set. As we can see, these are:

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

```
%%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

Launch Site Names Begin with 'CCA'

Synopsis:

The query below returns 5 rows of data pertaining to launch sites that begin with the letters ‘CCA’. Although there are two launch sites starting with these characters, we see that all of the launch sites below are CCAFS LC-40.

```
%%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
2010-04-06	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

Total Payload Mass

Synopsis:

By utilizing the SUM function, the query to the right returns the total payload mass that rockets with Nasa as a customer carried on their flights.

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Synopsis:

By utilizing the AVG function, the query to the right returns the average payload mass carried by all launch instances of the F9 v1.1 rocket.

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Synopsis:

By utilizing the MIN function, the query to the rate returns the first successful ground pad landing represented in the data set.

```
%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Synopsis:

By utilizing the BETWEEN function, the query to the right returns the booster version successful landings that were carrying a payload mass between 4000 and 6000.

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * 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

Synopsis:

By utilizing the COUNT function, the query to the right returns the total number of distinct successful and unsuccessful mission outcomes.

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) as total_number
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

Synopsis:

The query to the right returns the name of the boosters which carried a maximum payload during their missions.

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

```
 * 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

Synopsis:

The query below returns the launch records (month, data, landing outcome, booster version, and launch site) of launches that took place in 2015

To extract the month value, the substring function was utilized.

```
%%sql
SELECT substring(date,6,2) as Month, Date, landing_outcome, booster_version, launch_site
FROM spacextbl
WHERE landing_outcome LIKE 'Failure (drone ship)' AND DATE LIKE '%2015%';
```

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

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

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

Synopsis:

The query to the right displays and ranks, in descending order, the landing outcome types between 2010-06-04 and 2017-03-20.

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

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

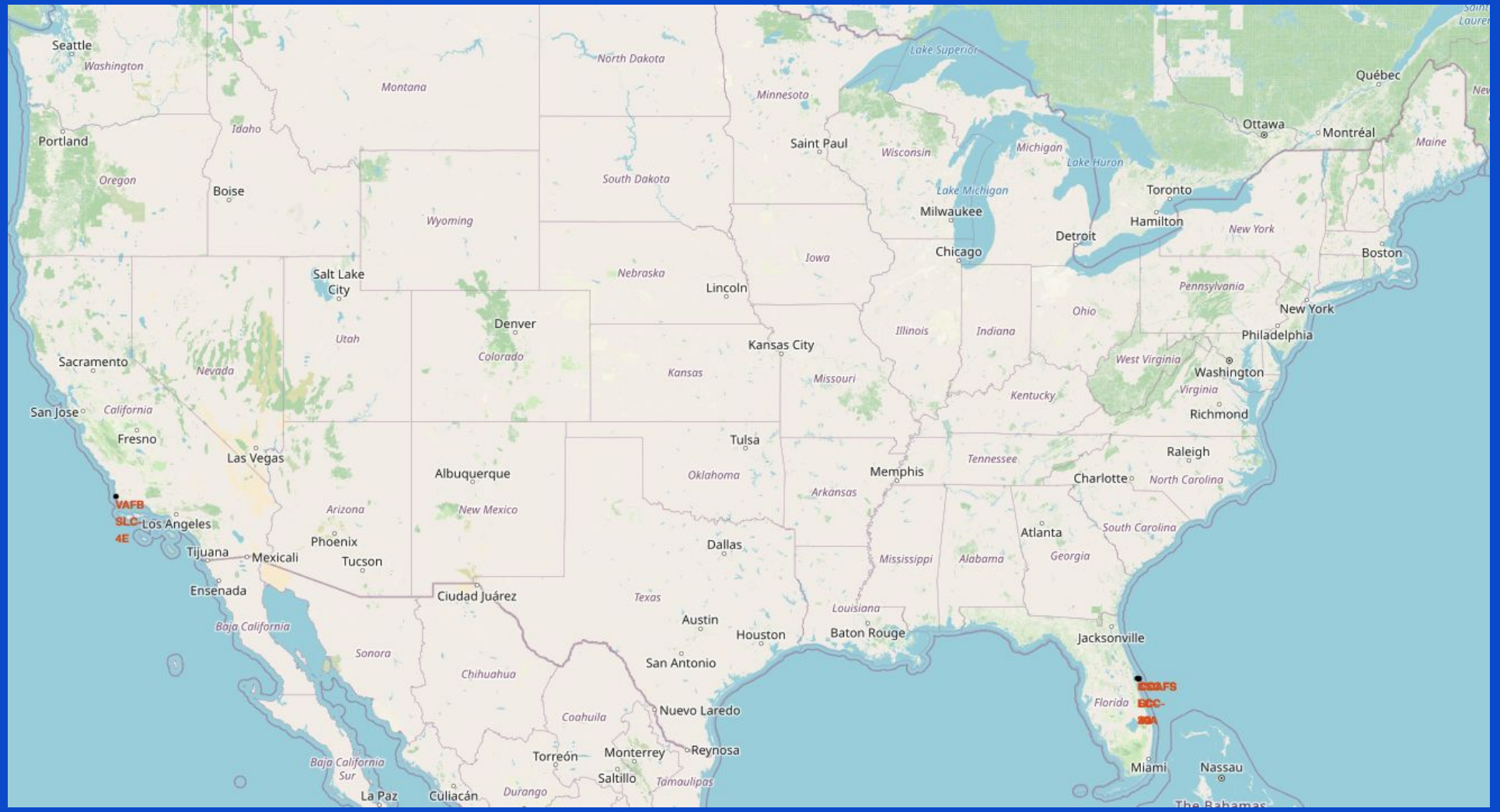
Landing_Outcome	COUNT
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Sites on Map

Insights:

As we can see, the launch sites are in close proximity to both the equator and a coast line.

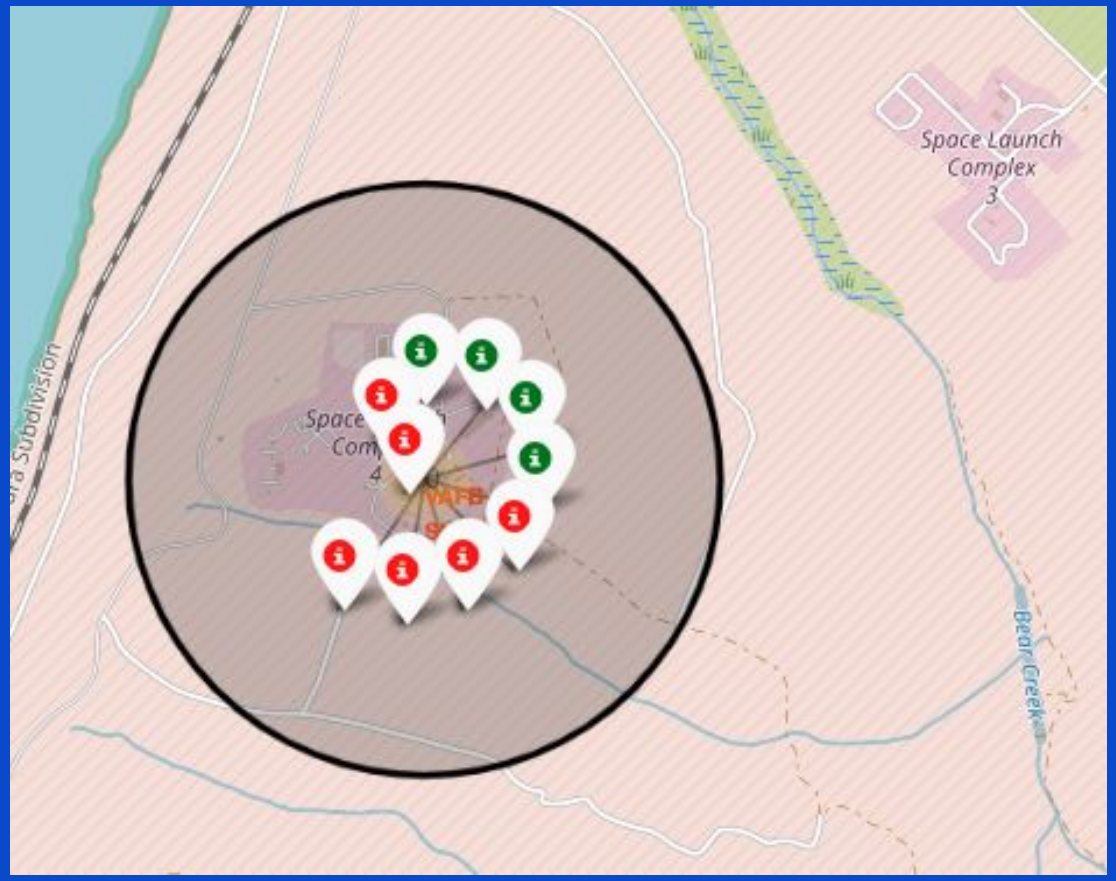


SpaceX Launch Outcomes on Map

Synopsis:

To the right, we see that successful launches are marked with a green marker, whereas unsuccessful attempts are marked with a red marker.

At this particular site, we see that there is a success rate of 40%.

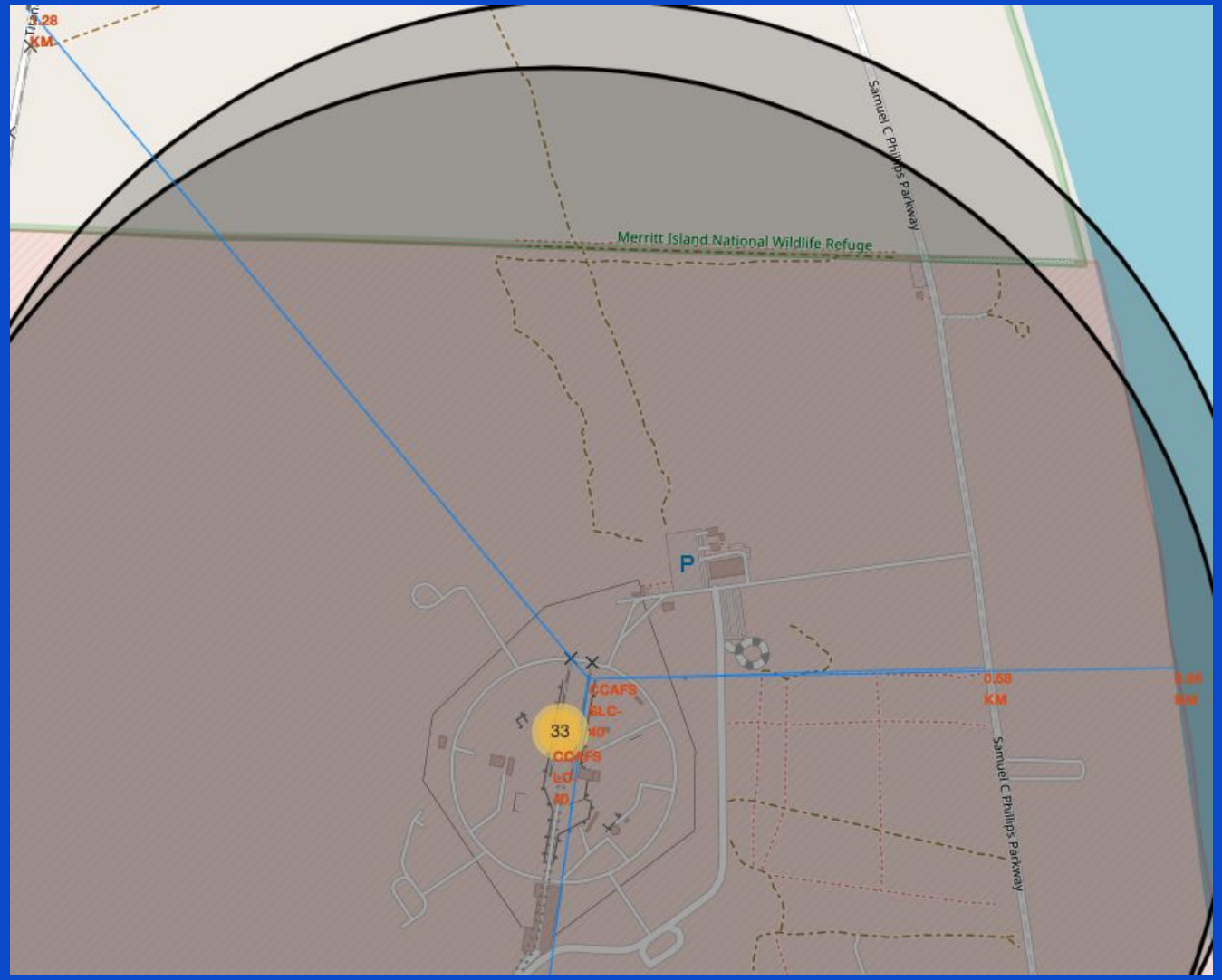


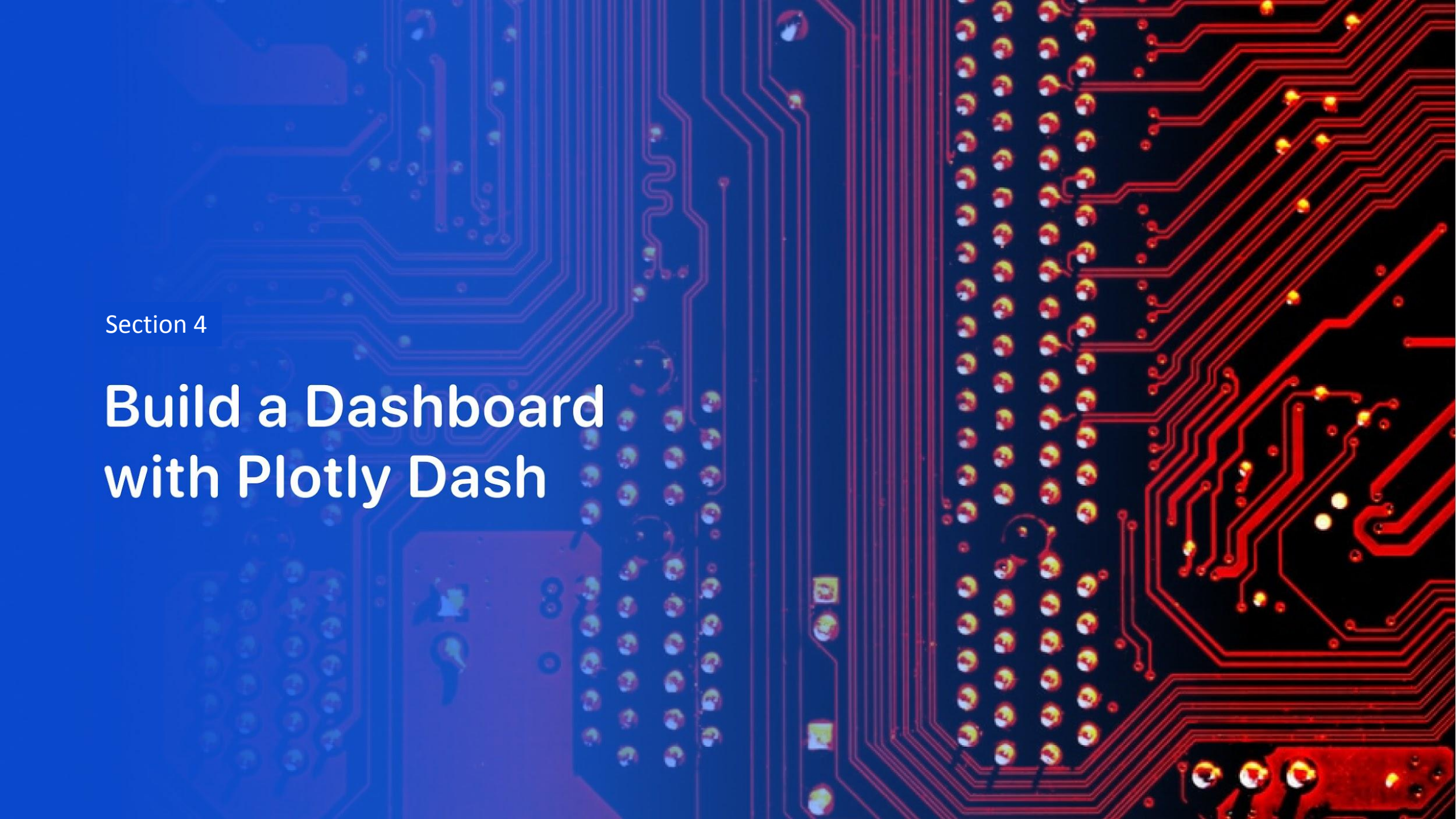
Launch Site Proximities to Local Features

Synopsis:

On the right, we see a launch site and lines marking its proximity to important local features.

- **Coastline**: .88 km
- **Highway**: .58 km
- **Railroad**: 1.28 km
- **City**: 51.43 km




Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site

Synopsis:

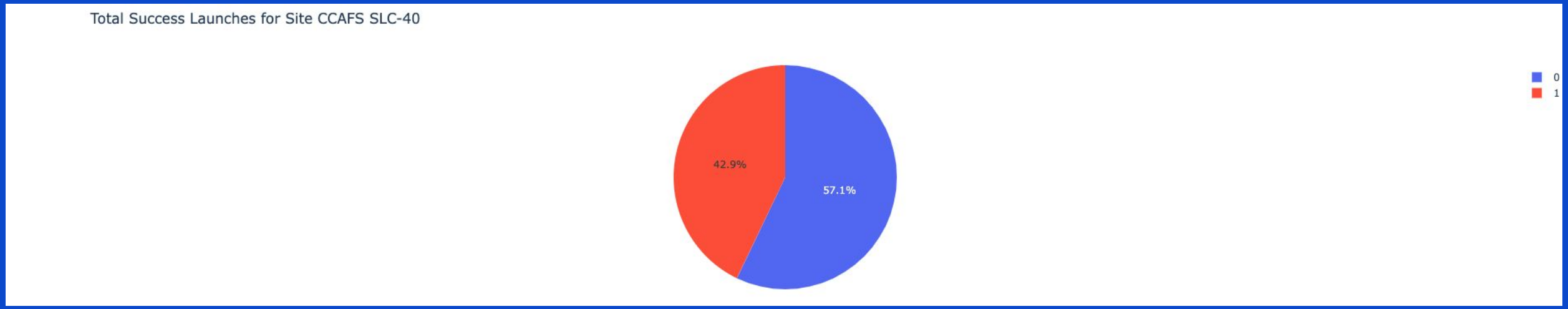
The pie chart below displays each launch sites success rate as percent of total success rate from all launch sites. i.e each sites share of success.

We can see that site KSC LC-391 has had the most successful launches among the all launch sites.

Launch Site with Highest Success Ratio

Synopsis:

As represented by the pie chart below, the launch site with the highest success rate is CCAFS SLC-40, with a success rate of 42.9%.



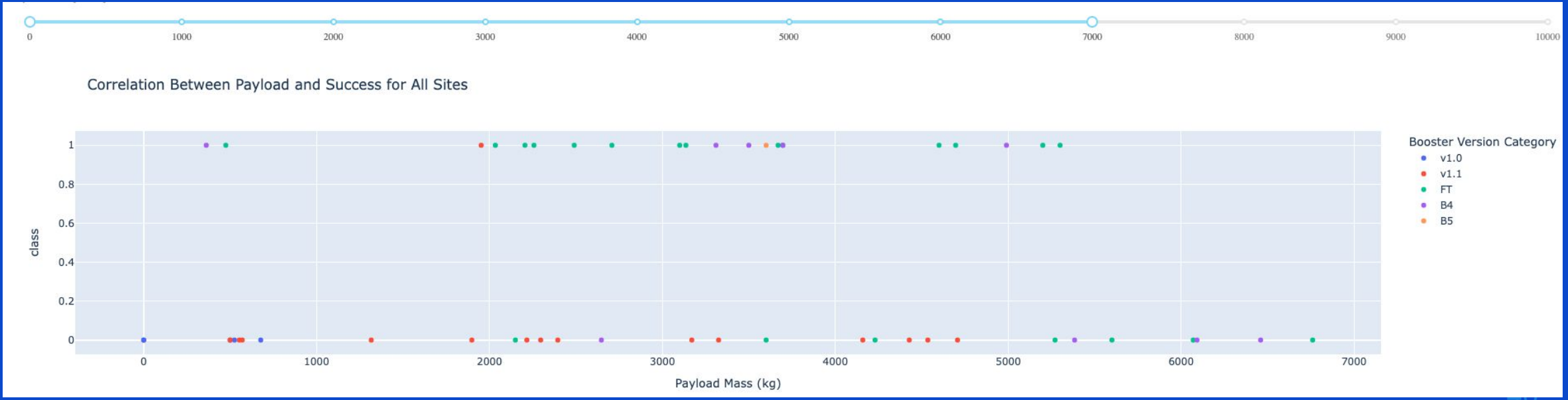
Correlation between Payload Mass and Success Rate

Synopsis:

We can see that lighter payloads have had a higher failure rate

We see very low rates of success in payloads between 4000 and 5000

Payloads between 2000 and 4000 have had many successful launches



Section 5

Predictive Analysis (Classification)

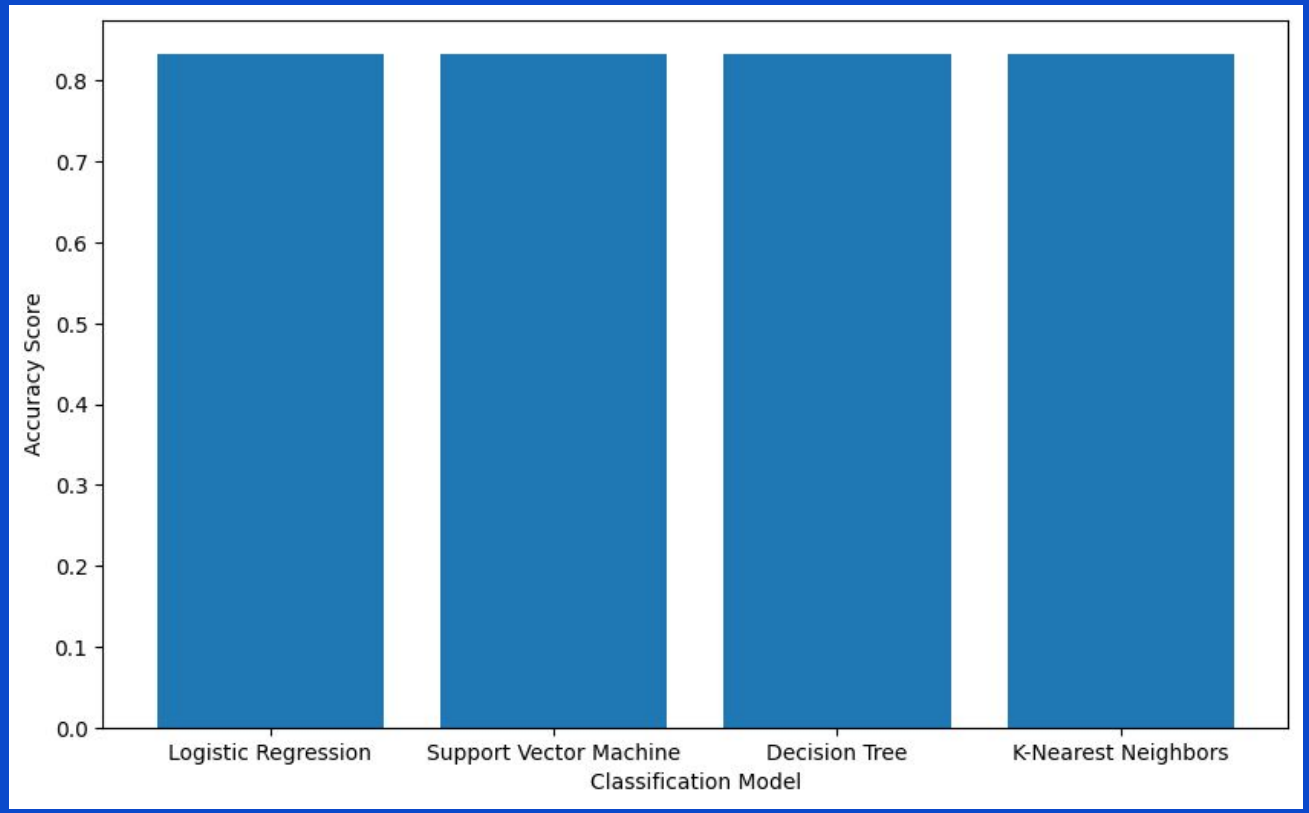
Classification Accuracy

Synopsis:

Upon tuning, training, testing, then evaluating the models, all the models received the same accuracy score.

This was done with a score of approximately .833333, as indicated by the plot to the right.

Although not perfect, this score is fairly high indicating that our models effectively predicted outcomes.

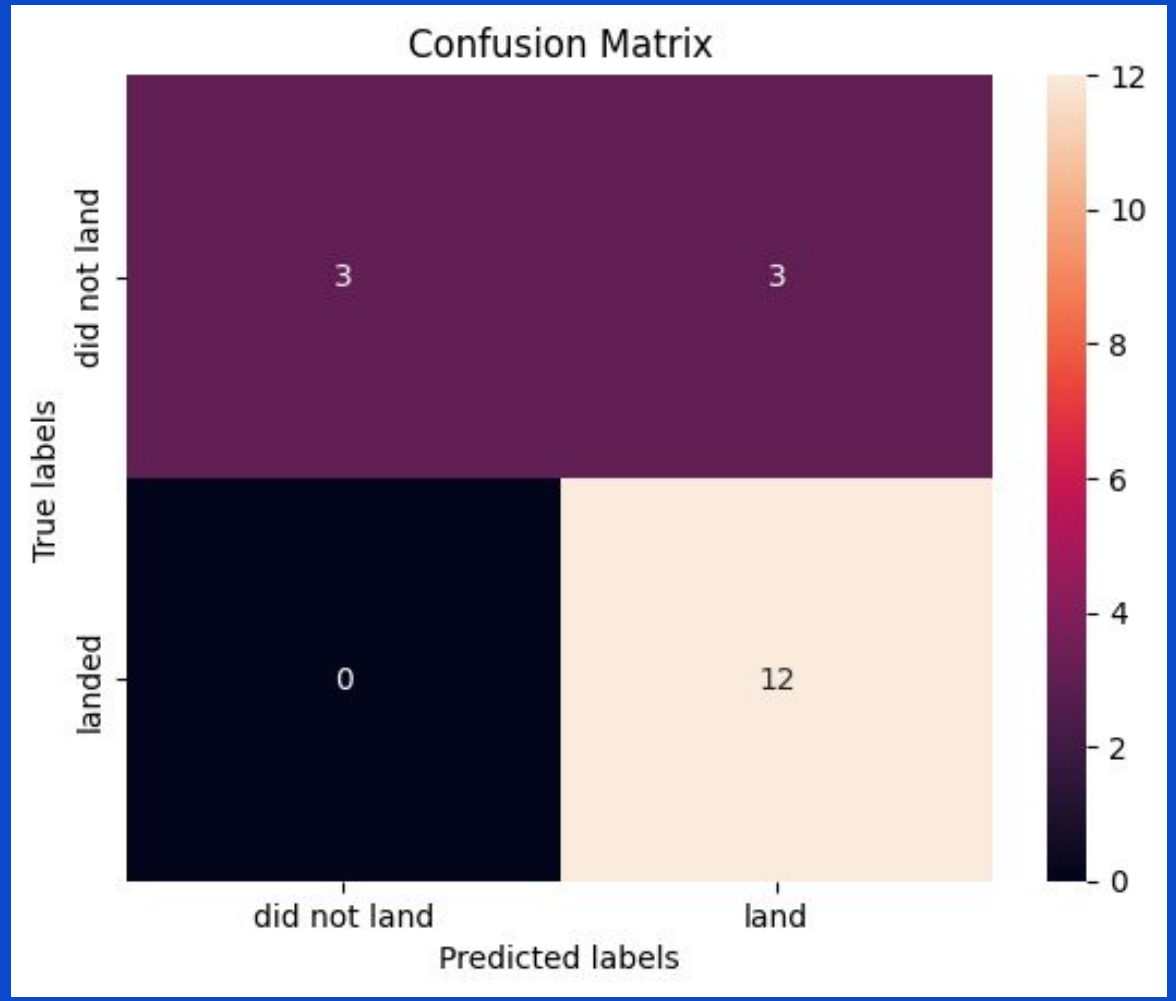


Confusion Matrix

Synopsis:
A confusion matrix is useful when determining the performance of a classification model as it indicates the number of true positives, true negatives, false positives, and false negatives.

In our case, the confusion matrices across the models were identical with:

- 12 true positives
- 3 true negatives
- 3 false positives



Conclusions

From our time series visualization, it is evident that launch success rate has improved drastically over time, with temporary periods of slight decreases over the years.

In terms of launch sites, they are all in close proximity to both the coast and the equator.

Site CCAFS SLC-40 has the highest success rate of launch sites

Orbit types ES-L1, GEO, HEO, and SSO have a 100% success rate

In general, it seems that the higher the payload mass, the higher the success rate

Overall, our predictive models performed well, and all delivered an accuracy score of 83%

Thank you!