DATA SCIENCE PROJECT

LOGAN PEARSON

JULY 3, 2024

OUTLINE

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion

EXECUTIVE SUMMARY

- **Summary of methodologies:**
  - Data collection
  - Data wrangling with Python
  - Exploratory data analysis with visualization
  - Exploratory data analysis with SQL
  - Interactive map with Folium
  - Interactive dashboard with Plotly Dash
  - Predictive analysis (Classification)
- **Summary of all results:**
  - Exploratory data analysis results
  - Interactive map and dashboard
  - Predictive analysis results

INTRODUCTION

➢ Project background and context

- SpaceX leads the commercial space industry by making space travel more affordable, thanks to its reusable Falcon 9 rockets. Priced at $62 million per launch, significantly less than competitors, SpaceX’s cost savings come from reusing the first stage. This project involves predicting the likelihood of SpaceX reusing the first stage, using machine learning and public data, simulating the role of a data scientist.

➢ Problems we want to answer

- Do payload specifications, launch site conditions, flight history, and mission requirements influence the likelihood of successful first-stage landings?
- Does the rate of first-stage reuse exhibit temporal trends or vary based on mission specifics?
- Which machine learning algorithm is most effective in predicting SpaceX's first-stage reuse under different operational scenarios?

METHODOLOGY

SECTION 1

METHODOLOGY

SUMMARY

- Data collection methodology:
  - SpaceX Rest API
  - Web scraping
- Perform data wrangling
  - Replacing and removing null data values
  - Filtering data
  - Implementing binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluating classification models

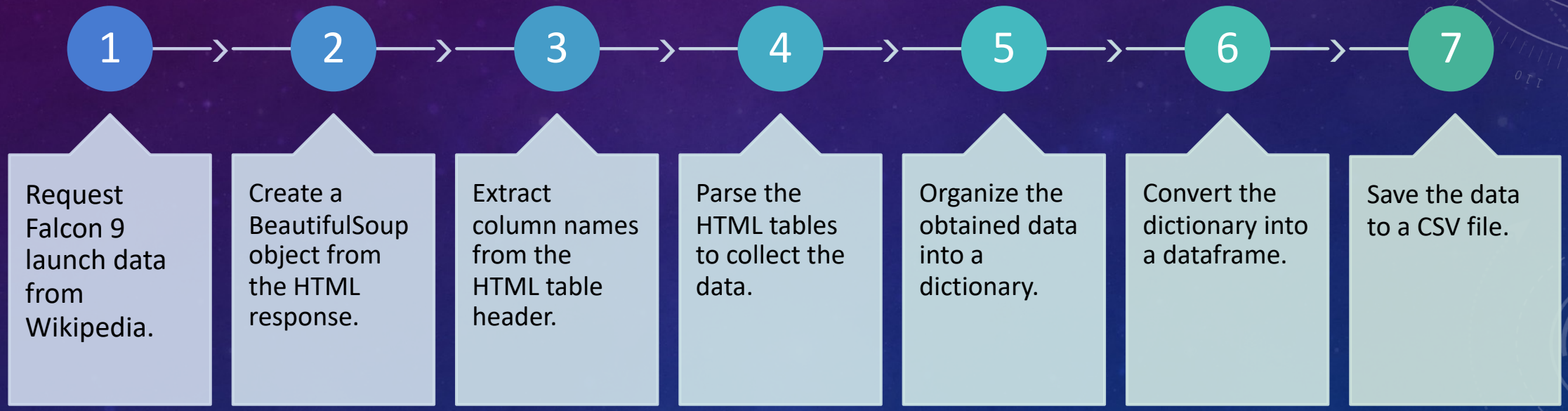
DATA COLLECTION – SPACEX API

Utilized SpaceX REST API to gather comprehensive data on rocket launches

1. Request rocket launch data from the SpaceX API.
2. Decode the response content using `json()` and convert it into a dataframe using `json_normalize()`.
3. Request necessary launch information from the SpaceX API using custom functions.
4. Construct the obtained data into a dictionary.
5. Create a dataframe from the constructed dictionary.
6. Filter the dataframe to include only Falcon 9 launches.
7. Replace missing values in the Payload Mass column with the calculated mean for this column.
8. Export the data to a CSV file.

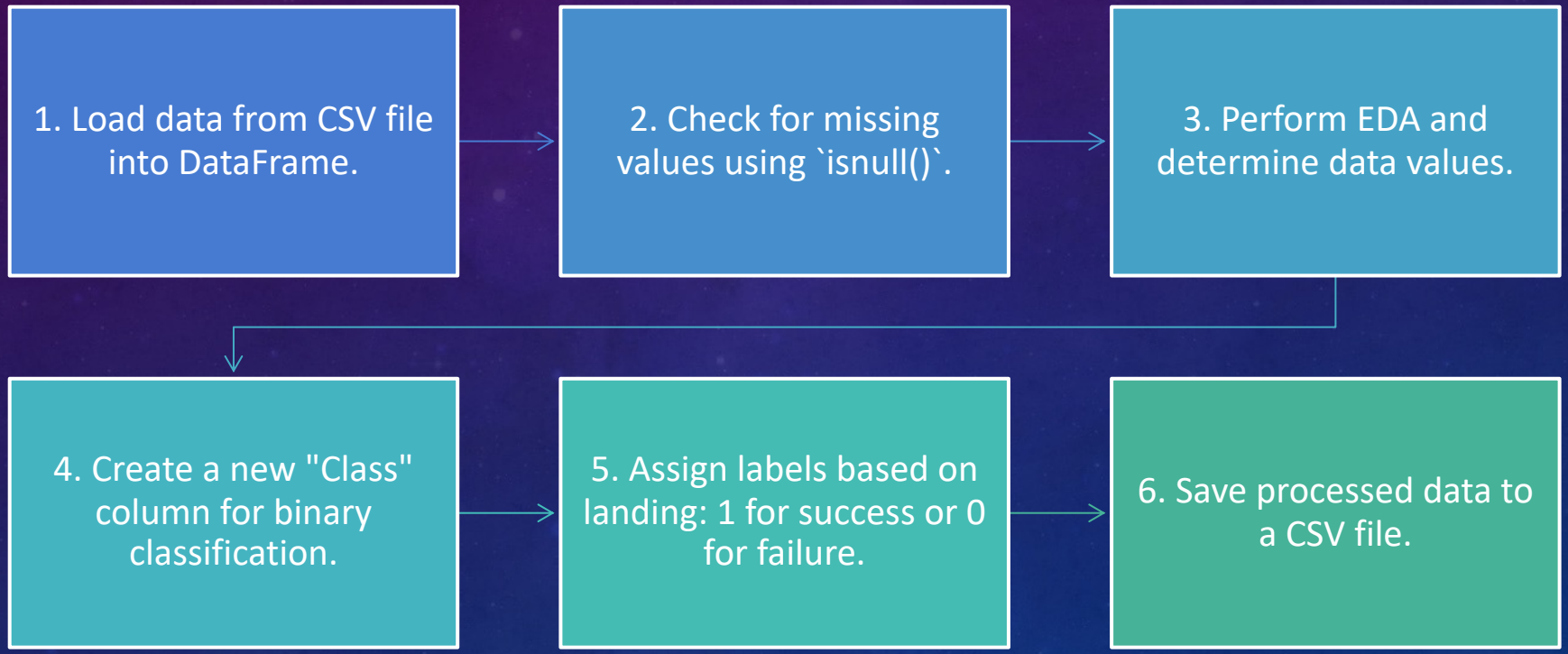
DATA COLLECTION – WEB SCRAPING

Extracted additional details from SpaceX's Wikipedia page to ensure complete information



DATA WRANGLING

Processed the collected data using Python in Jupyter Notebooks



EDA WITH DATA VISUALIZATIONS

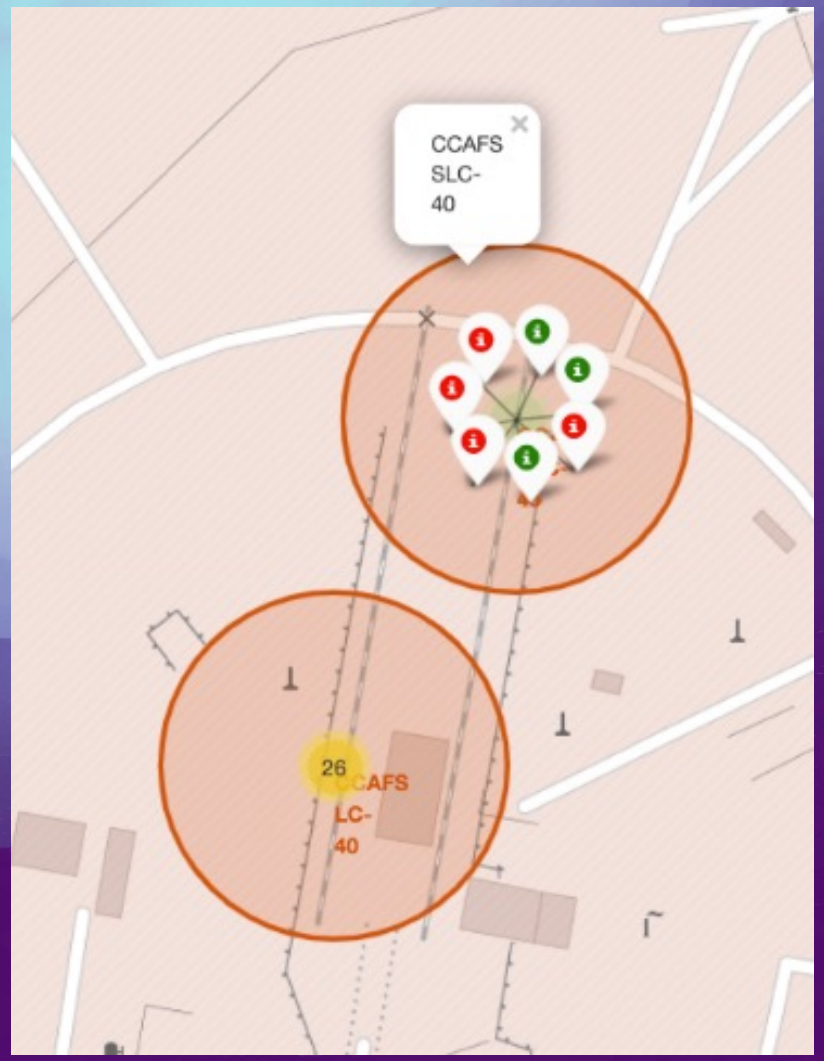
SUMMARY OF VISUALIZATIONS

- **FlightNumber vs. PayloadMass (Catplot):** Purpose: To observe if higher FlightNumber or PayloadMass affects landing outcome.
- **FlightNumber vs. LaunchSite (Catplot):** Purpose: To analyze if LaunchSite impacts the success of the landing.
- **PayloadMass vs. LaunchSite (Catplot):** Purpose: To understand how PayloadMass varies with LaunchSite and its impact on landing success.
- **Success Rate by Orbit Type (Bar Chart):** Purpose: To visualize the success rates across different orbital paths.
- **FlightNumber vs. Orbit Type (Scatter Plot):** Purpose: To explore how FlightNumber correlates with different types of orbits in terms of landing success.
- **PayloadMass vs. Orbit Type (Scatter Plot):** Purpose: To examine the relationship between PayloadMass and different orbital types in terms of landing outcomes.
- **Yearly Launch Success Rate Trend (Line Chart):** Purpose: To track the trend of launch success rates over the years and identify any patterns or improvements.

SUMMARY OF SQL QUERIES

WITH SQLITE

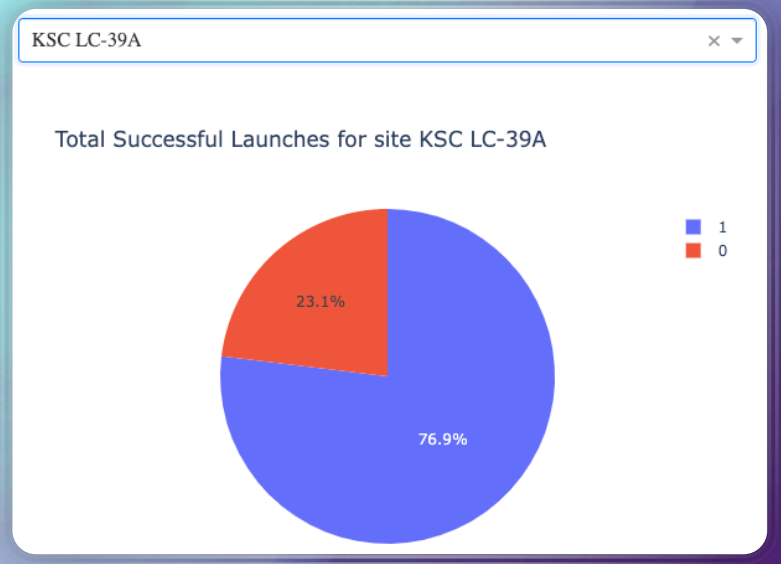
- Listed unique launch sites.
- Displayed 5 records of launch sites starting with 'CCA'.
- Total payload mass of boosters launched by NASA (CRS).
- Average payload mass of booster version F9 v1.1.
- Date of first successful ground pad landing.
- Boosters successful on drone ship with payload 4000-6000 kg.
- Count of successful and failed mission outcomes.
- Booster versions with maximum payload mass using a subquery.
- Records showing month, failure on drone ship, booster versions, launch site in 2015.
- Ranked landing outcomes between 2010-06-04 and 2017-03-20.

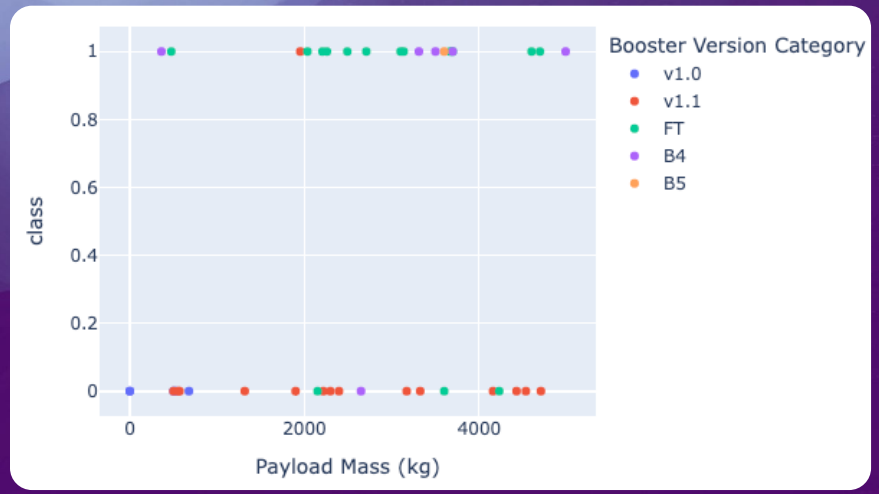


BUILDING AN INTERACTIVE MAP

WITH FOLIUM

- **Markers of All Launch Sites:**
  - Added markers with circles, popup labels, and text labels for each launch site. For example, NASA Johnson Space Center was marked using its latitude and longitude coordinates as the start location. This setup visually represents the geographical locations and their proximity to the Equator and coasts.
- **Colored Markers of Launch Outcomes:**
  - Integrated colored markers to distinguish between success (Green) and failed (Red) launches. Marker clustering was used to identify launch sites with higher success rates, providing a clear visual indication.
- **Distances Between Launch Site and Proximities:**
  - Included colored lines to illustrate distances from launch sites (e.g., KSC LC-39A) to nearby features such as railways, highways, coastlines, and closest cities. This feature enhances understanding of spatial relationships and logistical considerations for each launch site.





BUILDING A DASHBOARD

WITH PLOTLY DASH

- **Dropdown List for Launch Site Selection:**
  - Allows users to filter launches by site for detailed analysis.
- **Pie Chart Showing Total Successful Launches:**
  - Provides an overview of success distribution across sites or for all launches.
- **Slider to Select Payload Range (kg):**
  - Enables users to filter data by payload range, facilitating focused analysis.
- **Scatter Chart Depicting Payload Mass and Launch Success:**
  - Visualizes the correlation between payload mass and launch outcome, with site-specific or overall insights.

PREDICTIVE ANALYSIS (CLASSIFICATION)

OVERVIEW

Creating	Creating a NumPy array from the "Class" column in the dataset.
Standardizing	Standardizing the data using StandardScaler, then fitting and transforming it.
Splitting	Splitting the data into training and testing sets using train_test_split function.
Creating	Creating a GridSearchCV object with cv = 10 to find the best parameters.
Applying	Applying GridSearchCV on Logistic Regression, SVM, Decision Tree, and KNN models.
Calculating	Calculating the accuracy on the test data using the score method for all models.
Examining	Examining the confusion matrix for all models.
Determining	Determining the best-performing method by evaluating Jaccard score and F1 score metrics.

TYPES OF RESULTS

Exploratory data analysis results

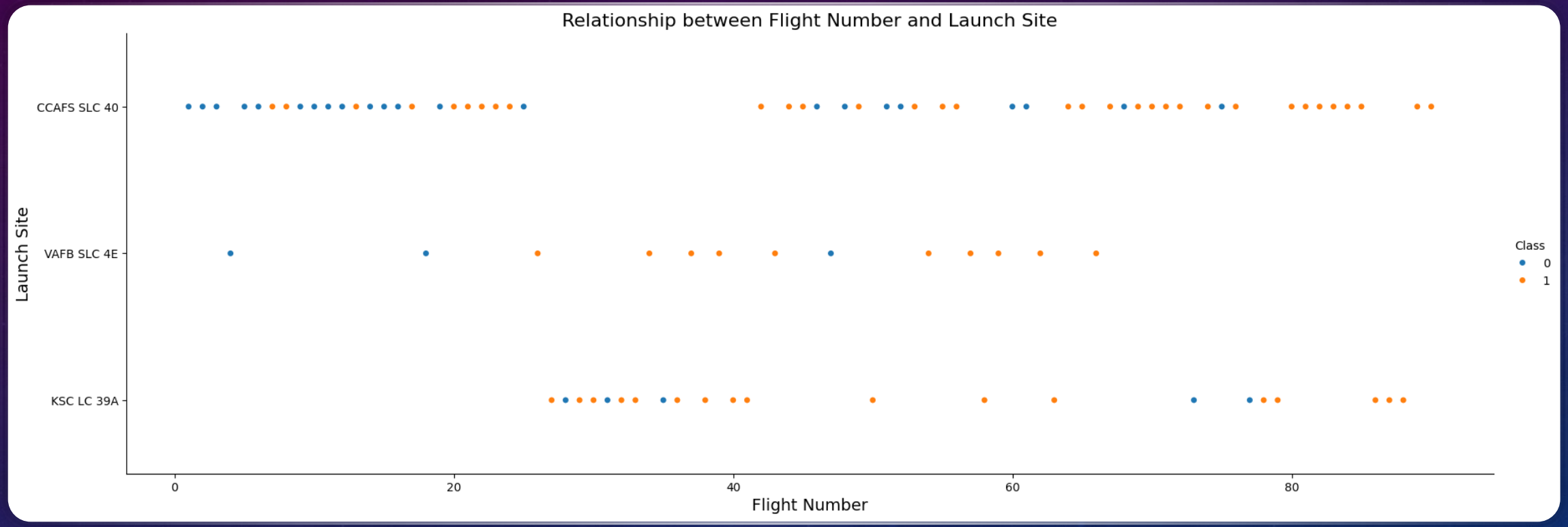
Interactive analytics demo in screenshots

Predictive analysis results

INSIGHTS FROM EDA

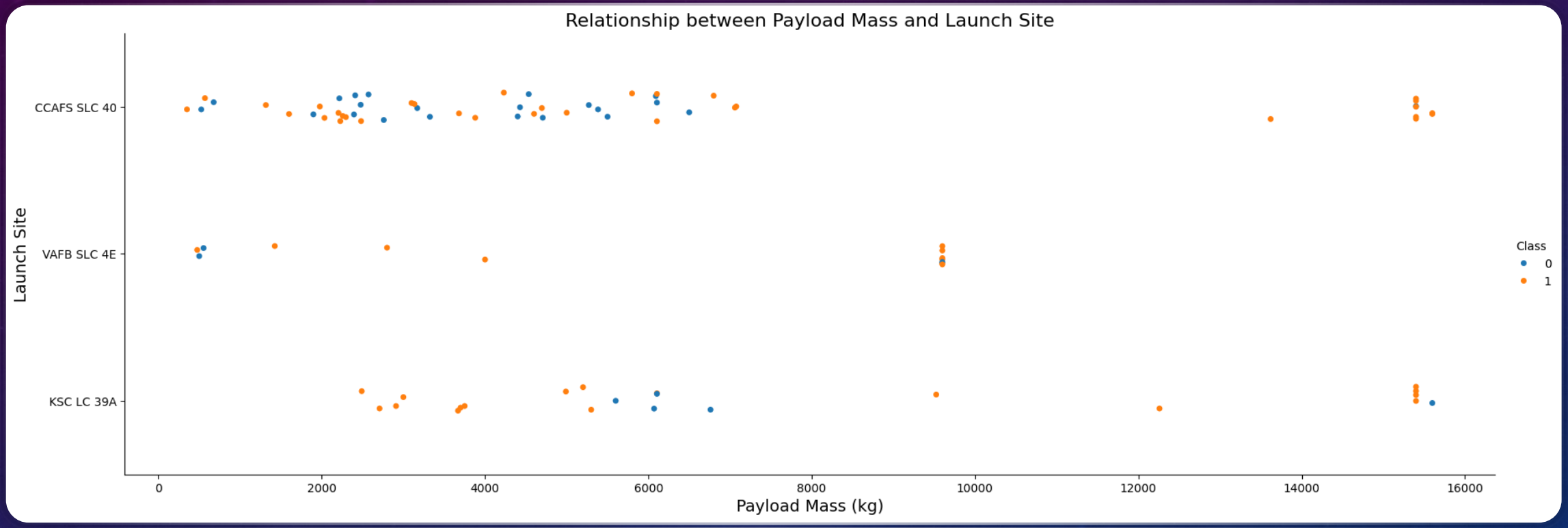
SECTION 2

FLIGHT NUMBER VS. LAUNCH SITE

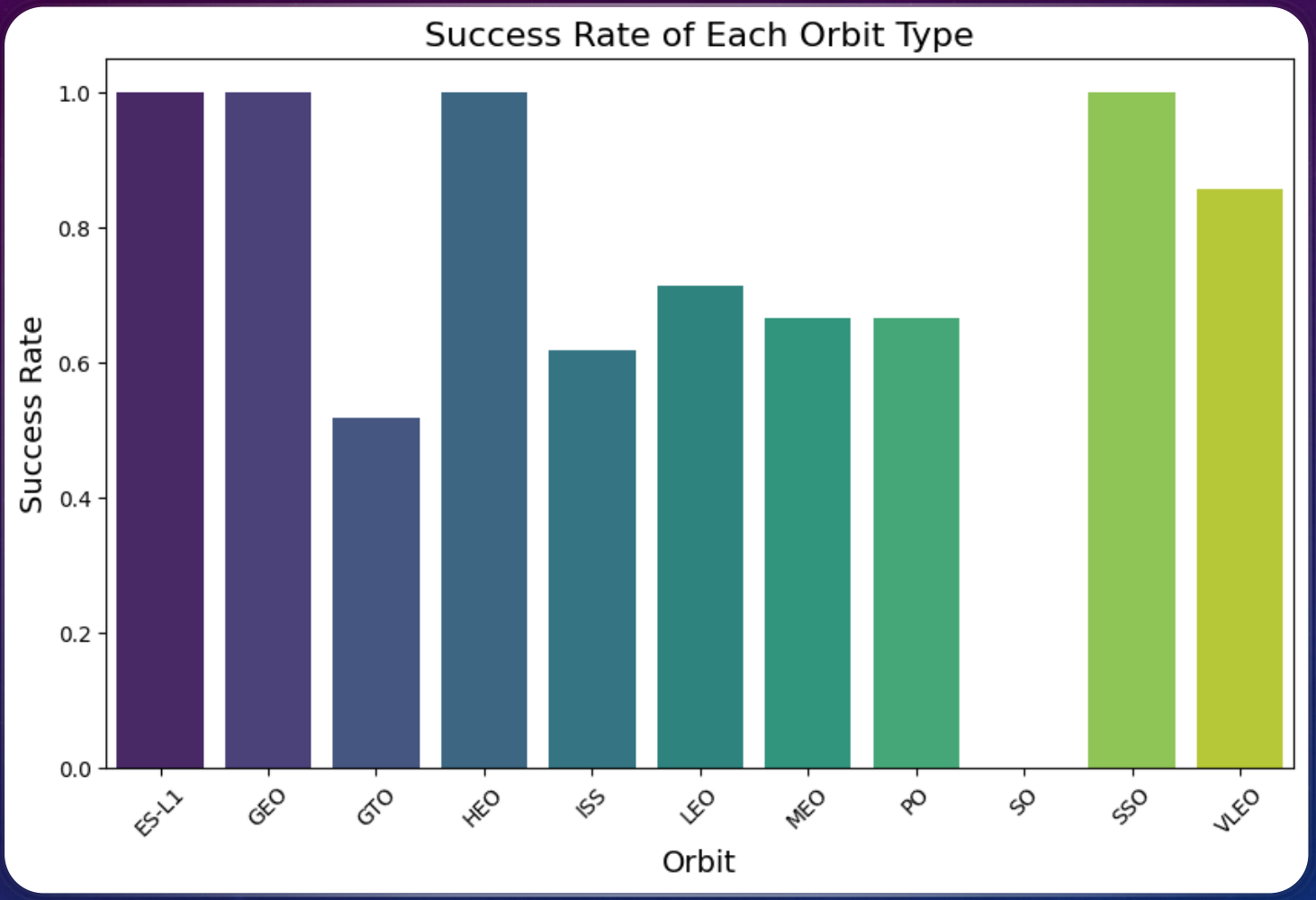


- The earliest 5 flights all failed while the last 13 all succeeded.
- As flight numbers increase, we see success rate increase as well.

PAYLOAD VS. LAUNCH SITE

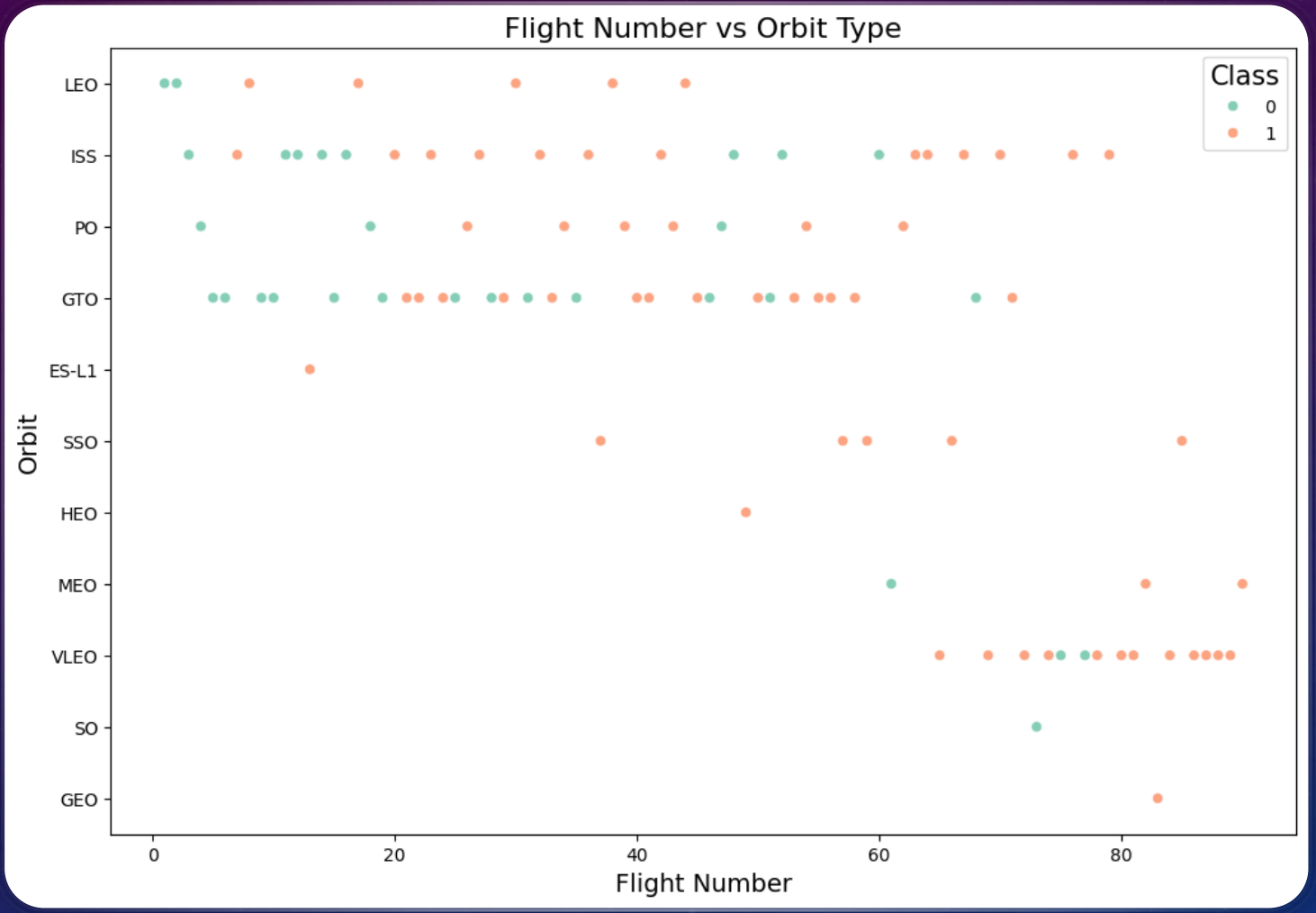


- As payload mass increases, success rate increases as well.
- Most of the launches after 8,000 kg were successful.



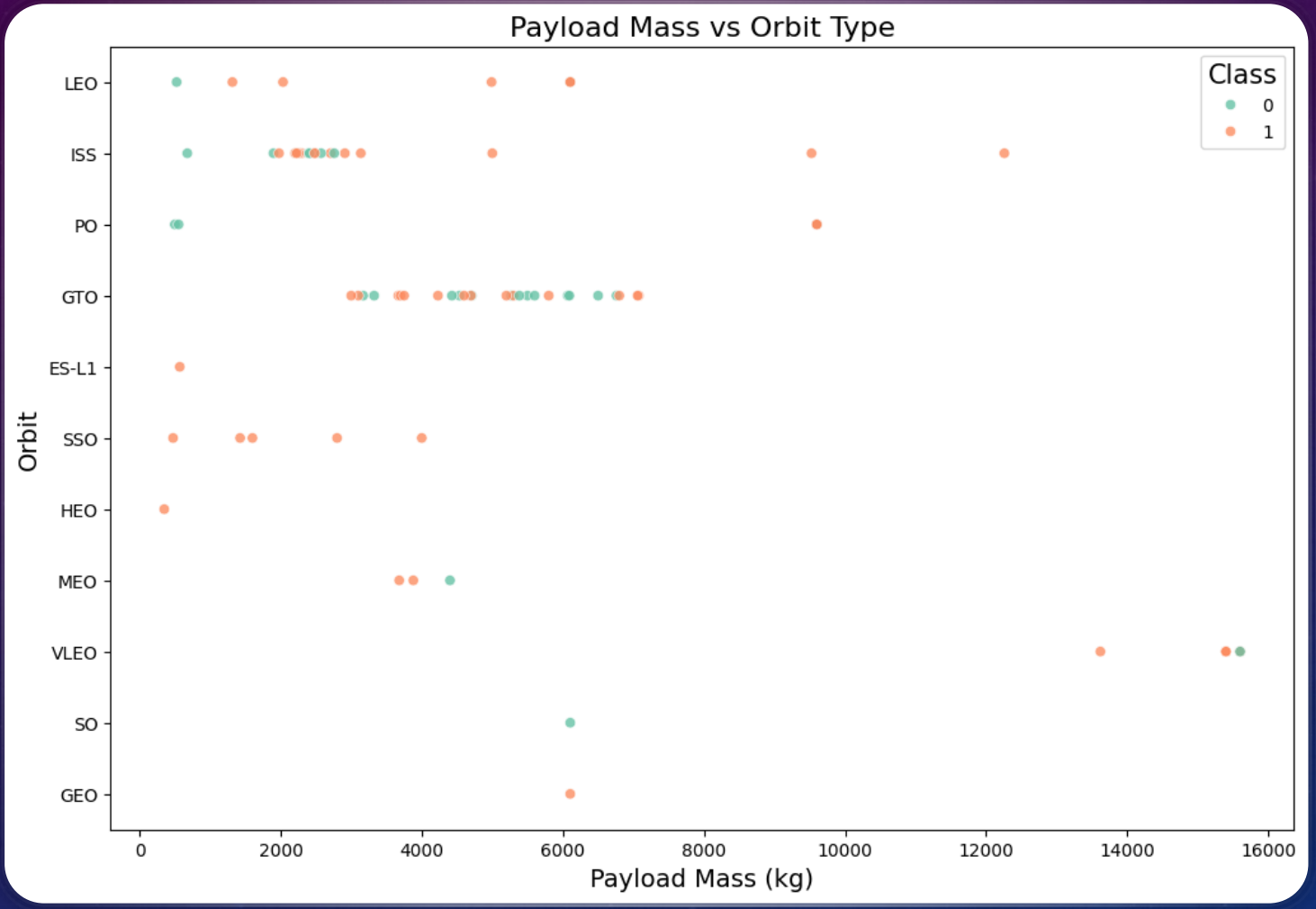
SUCCESS RATE VS. ORBIT TYPE

- There are 4 orbit types with 100% success rates.
- There are 5 orbit types between 50% and 70% success rates.
- There is 1 orbit type, SO, with a 0% success rate.



FLIGHT NUMBER VS. ORBIT TYPE

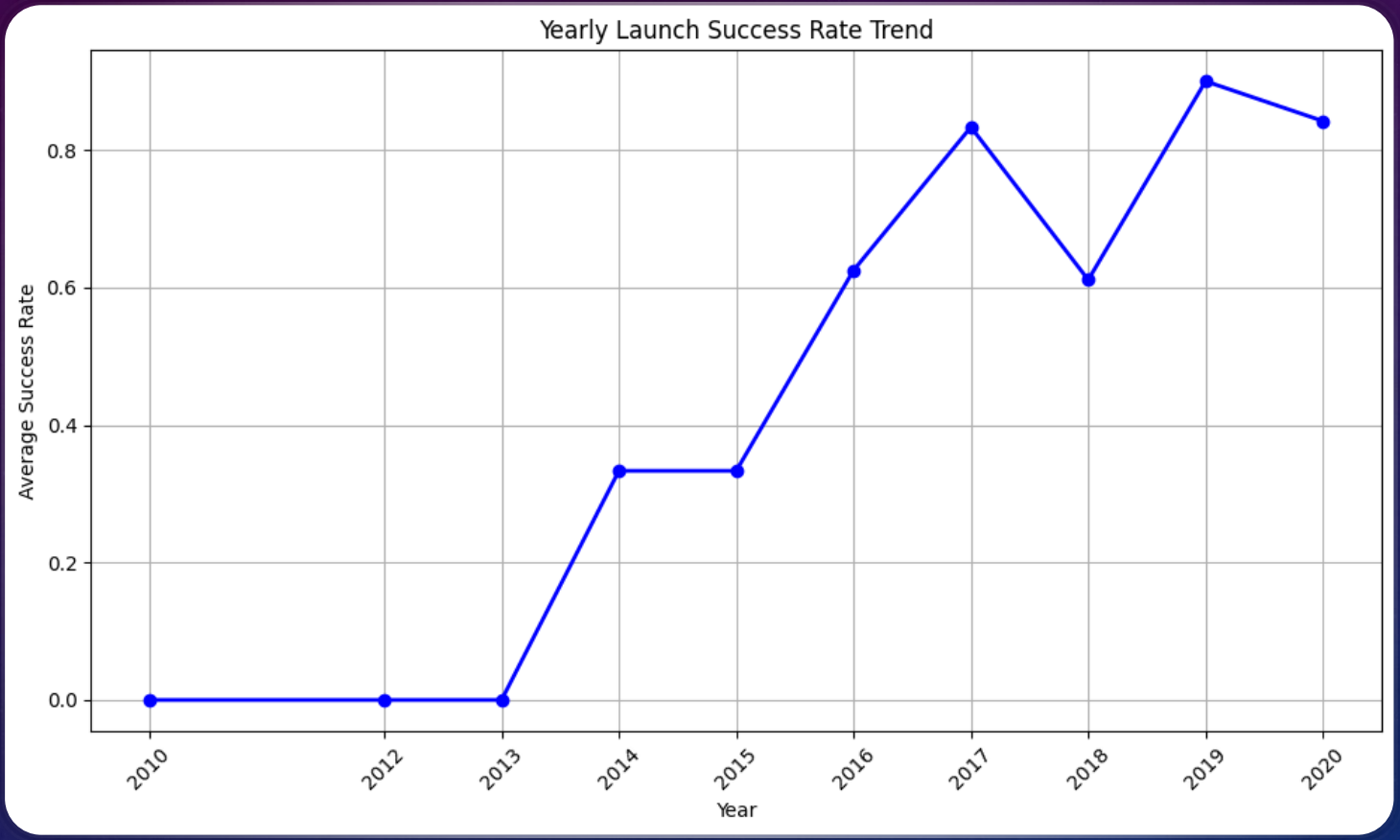
- In some cases, for example, LEO, it appears that when the flight number increases, the success rate increases.
- Overall, there does not seem to be a significant relationship between flight number and GTO orbit.



PAYLOAD VS. ORBIT TYPE

- Greater payloads seem to have a positive success rate effect on LEO, and ISS.
- For the rest, the payload mass appears insignificant or unclear whether if affects the the orbit's success rate.

LAUNCH SUCCESS YEARLY TREND



- Overall, the average success rate increases as years increase.

EDA WITH SQL

SECTION 3

```
[20]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

[20]:

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

LAUNCH SITE NAMES

FINDS THE NAME OF EACH UNIQUE LAUNCH SITE

```
[19]: %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

[19]:

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

LAUNCH SITE RECORDS THAT BEGIN WITH 'CCA'

FINDS 5 RECORDS WHERE LAUNCH SITES BEGIN WITH `CCA`

```
[36]: %%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS "TotalPayloadMass"
FROM SPACEXTABLE
WHERE "Booster_Version" LIKE 'F9%'
  AND "PAYLOAD_MASS__KG_" IS NOT NULL;

 * sqlite:///my_data1.db
Done.

[36]:
```

TotalPayloadMass
619967

TOTAL PAYLOAD MASS

CALCULATES THE TOTAL PAYLOAD CARRIED BY BOOSTERS FROM NASA

```
[37]: %%sql
SELECT AVG("PAYLOAD_MASS__KG_") AS AveragePayloadMass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1'
  AND "PAYLOAD_MASS__KG_" IS NOT NULL;

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

[37]:

AveragePayloadMass
2928.4

AVERAGE PAYLOAD MASS BY F9 V1.1

CALCULATES THE AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1

[38]:
```
%%sql
SELECT MIN("Date") AS FirstSuccessfulLanding
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Success%Ground%Pad%';
```

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

[38]:

FirstSuccessfulLanding
2015-12-22

FIRST SUCCESSFUL GROUND LANDING DATE

FINDS THE DATE OF THE FIRST SUCCESSFUL LANDING OUTCOME ON GROUND PAD

[39]:
```
%%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Success%Drone%Ship%'
  AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

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

[39]:

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

SUCCESSFUL DRONE SHIP LANDING

FINDS THE BOOSTER VERSIONS OF SUCCESSFUL LANDINGS WITH PAYLOADS BETWEEN 4000 AND 6000

```
[40]: %%sql
SELECT COUNT(*) AS TotalSuccessfulOutcomes
FROM SPACEXTABLE
WHERE "Mission_Outcome" LIKE 'Success%'
UNION ALL
SELECT COUNT(*) AS TotalFailureOutcomes
FROM SPACEXTABLE
WHERE "Mission_Outcome" LIKE 'Failure%';

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

[40]:

TotalSuccessfulOutcomes
100
1

TOTAL SUCCESSFUL & FAILURE MISSION OUTCOMES

CALCULATES THE TOTAL NUMBER OF SUCCESSFUL OUTCOMES IN THE FIRST ROW, THEN THE NUMBER OF FAILURE MISSION OUTCOMES IN THE SECOND ROW

```sql
%%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

BOOSTERS THAT CARRIED THE MAXIMUM PAYLOAD

FINDS THE NAMES OF THE BOOSTERS THAT 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

[42]:
```
%%sql
SELECT SUBSTR("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTABLE
WHERE SUBSTR("Date", 1, 4) = '2015'
  AND "Landing_Outcome" LIKE 'Failure%Drone%Ship%';
```

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

[42]:

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

2015 LAUNCH FAILURE RECORDS

FINDS THE FAILED LANDING OUTCOMES IN DRONE SHIPS, THEIR BOOSTER VERSIONS, AND LAUNCH SITE NAMES FOR THE YEAR 2015

```sql
%%sql
SELECT "Landing_Outcome", COUNT(*) AS CountLandingOutcomes
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY CountLandingOutcomes DESC;
```

Landing_Outcome	CountLandingOutcomes
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

RANKED LANDING OUTCOMES

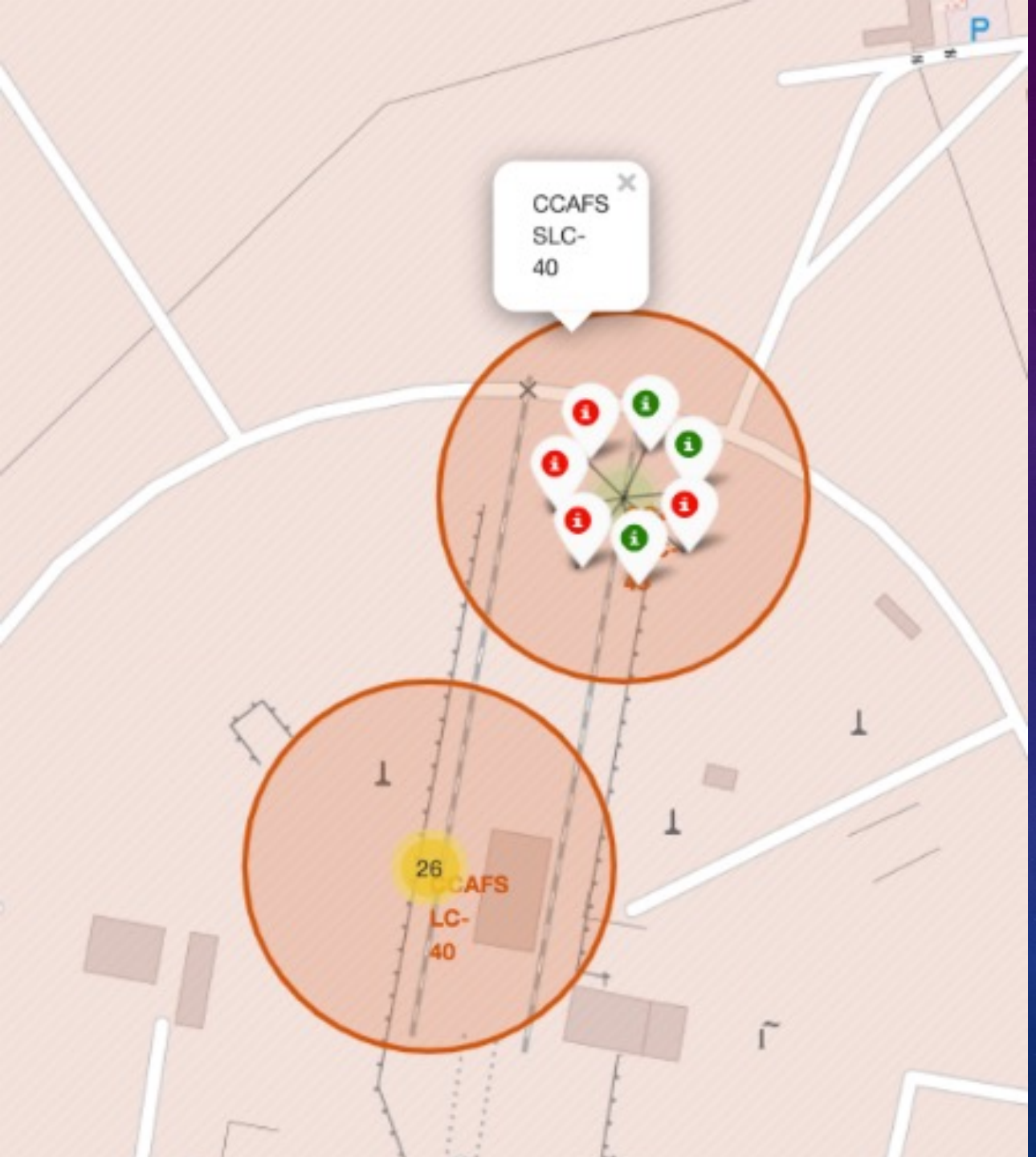
RANKS THE COUNT OF LANDING OUTCOMES (SUCH AS FAILURE (DRONE SHIP) OR SUCCESS (GROUND PAD)) BETWEEN THE DATE 2010-06-04 AND 2017-03-20, IN DESCENDING ORDER

LAUNCH SITE PROXIMITIES ANALYSIS

SECTION 4

LAUNCH SITES WORLD MAP

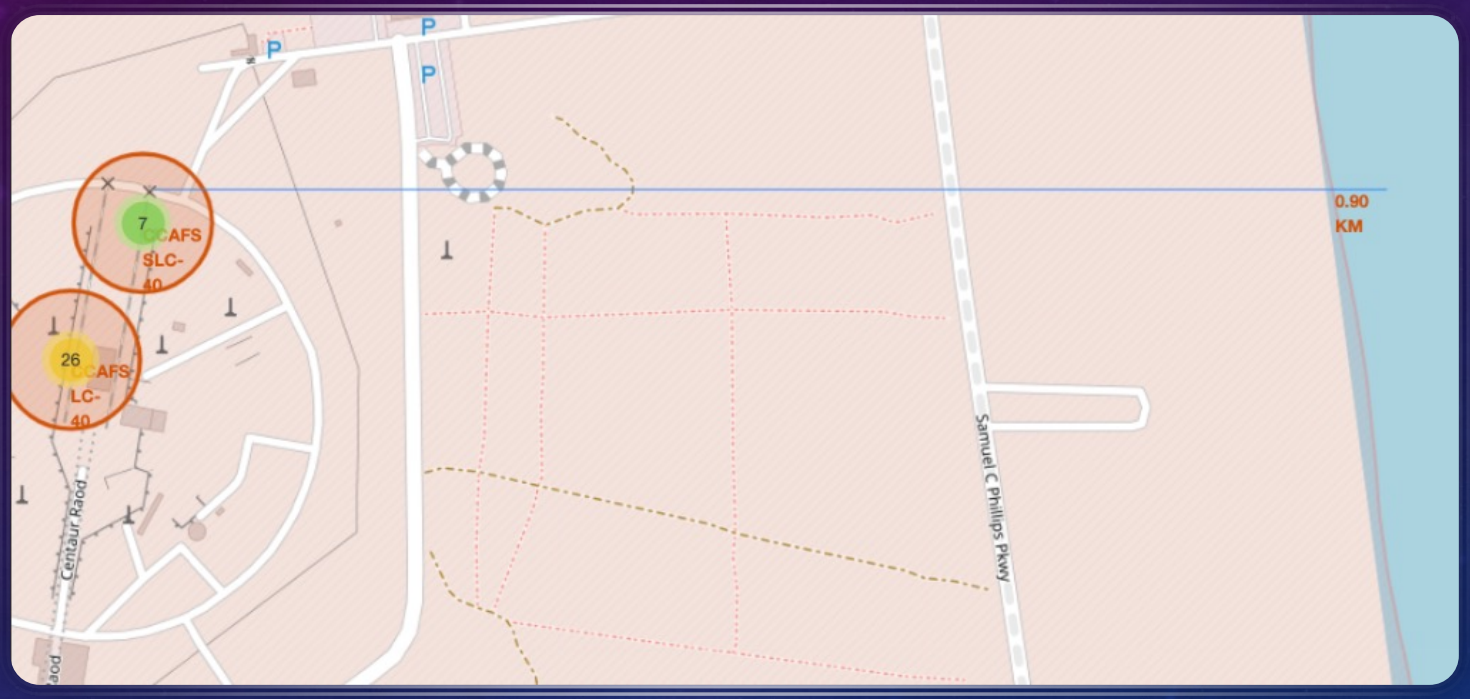
This map shows us that SpaceX launch sites are only in the United States. They are all located on the coasts of California and Florida.



COLOR-LABELED LAUNCH SITE MARKERS

- **Green Markers:**
  - ✓ Show successful launches
- **Red Markers:**
  - ❖ Show failed launches

LAUNCH SITE'S DISTANCE TO COASTLINE

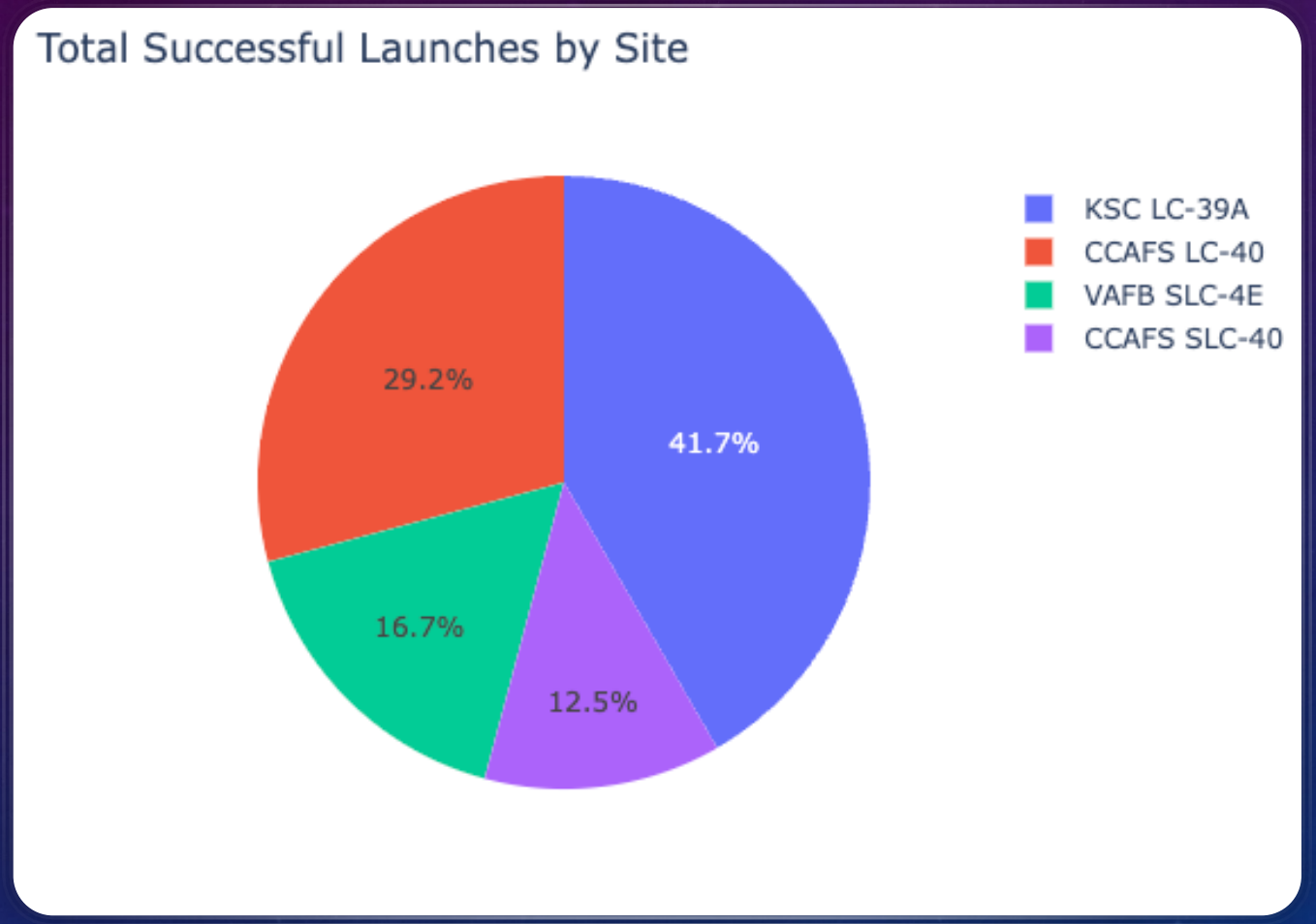


This visual shows a line from a launch site to the coastline that is 0.9 KM away.

BUILDING A DASHBOARD WITH PLOTLY DASH

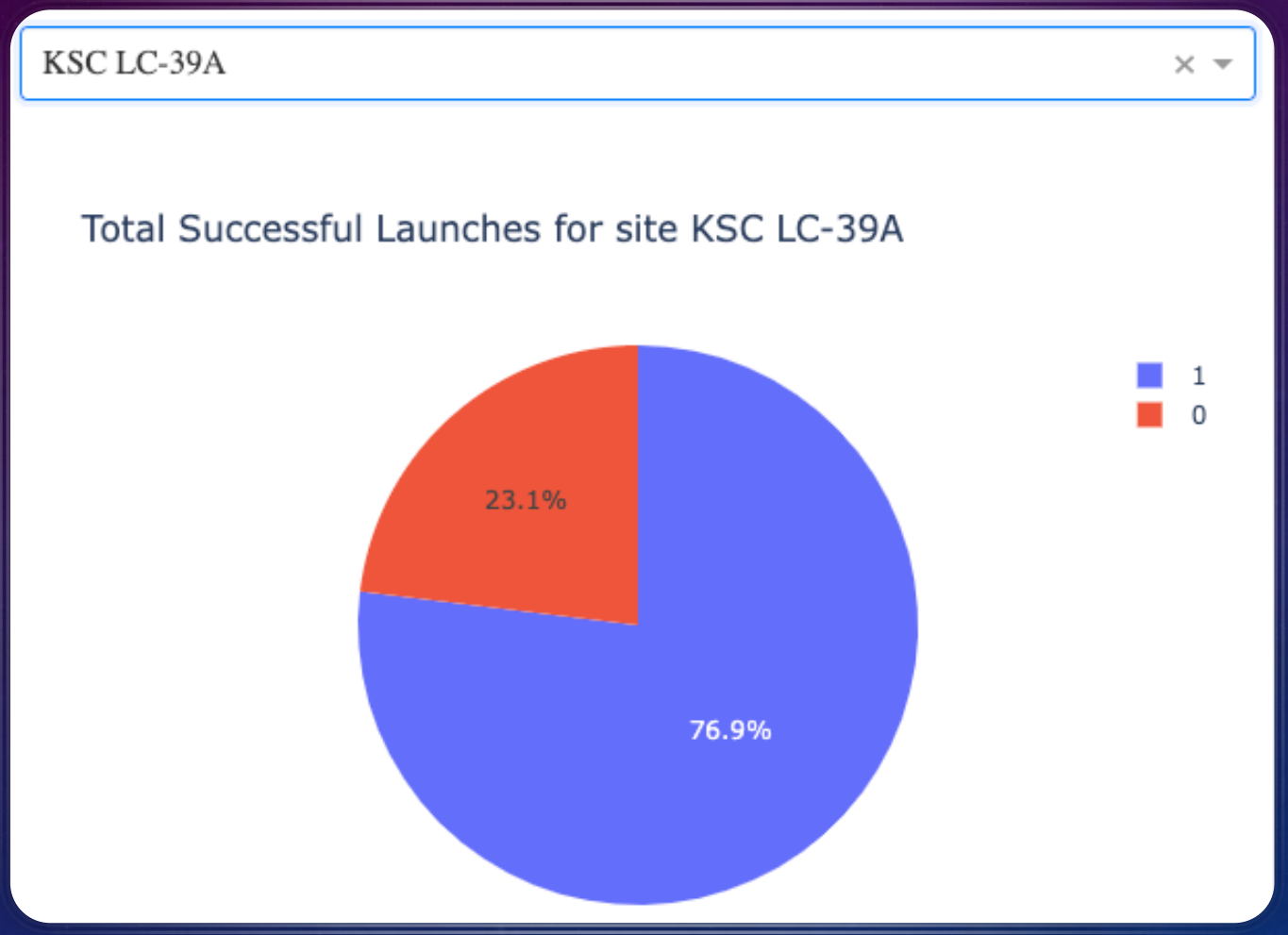
SECTION 5

TOTAL SUCCESS CONTRIBUTION OF EACH LAUNCH SITE

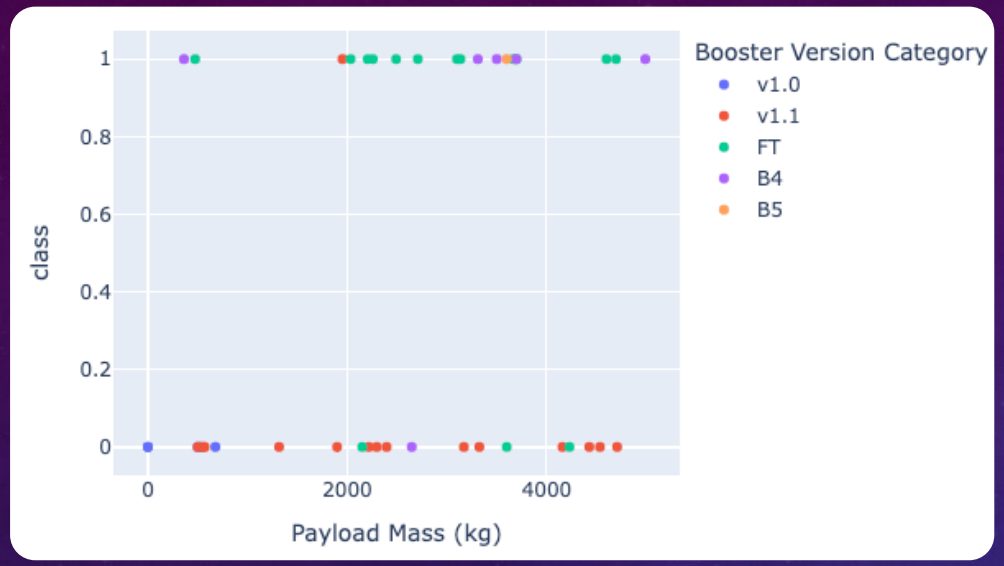


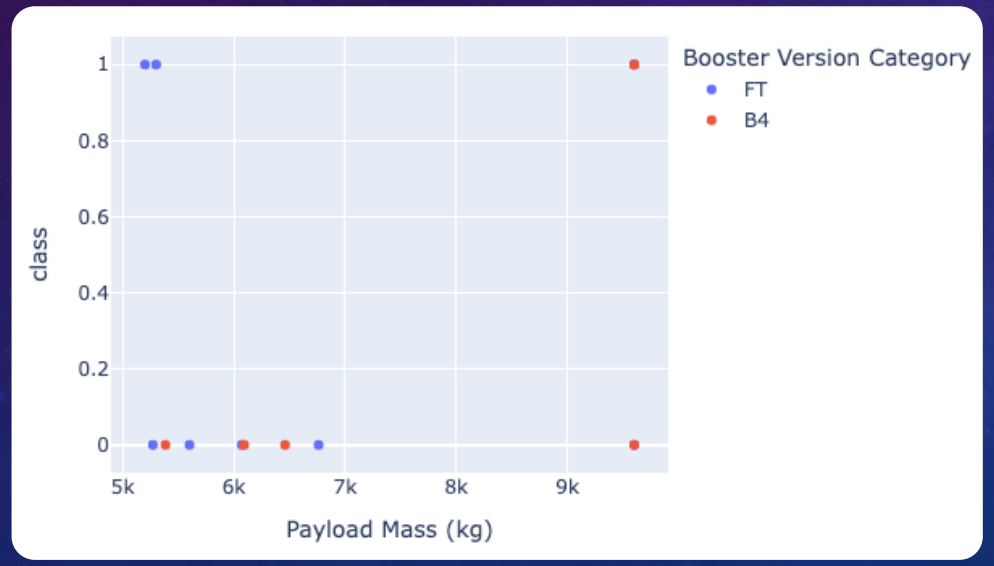
KSC LC-39A had the most successful launches making up 41.7% of the total successful launches.

LAUNCH SITE WITH HIGHEST SUCCESS RATE



The launch site KSC LC-39A has the highest success rate being successful 76.9% of the time.



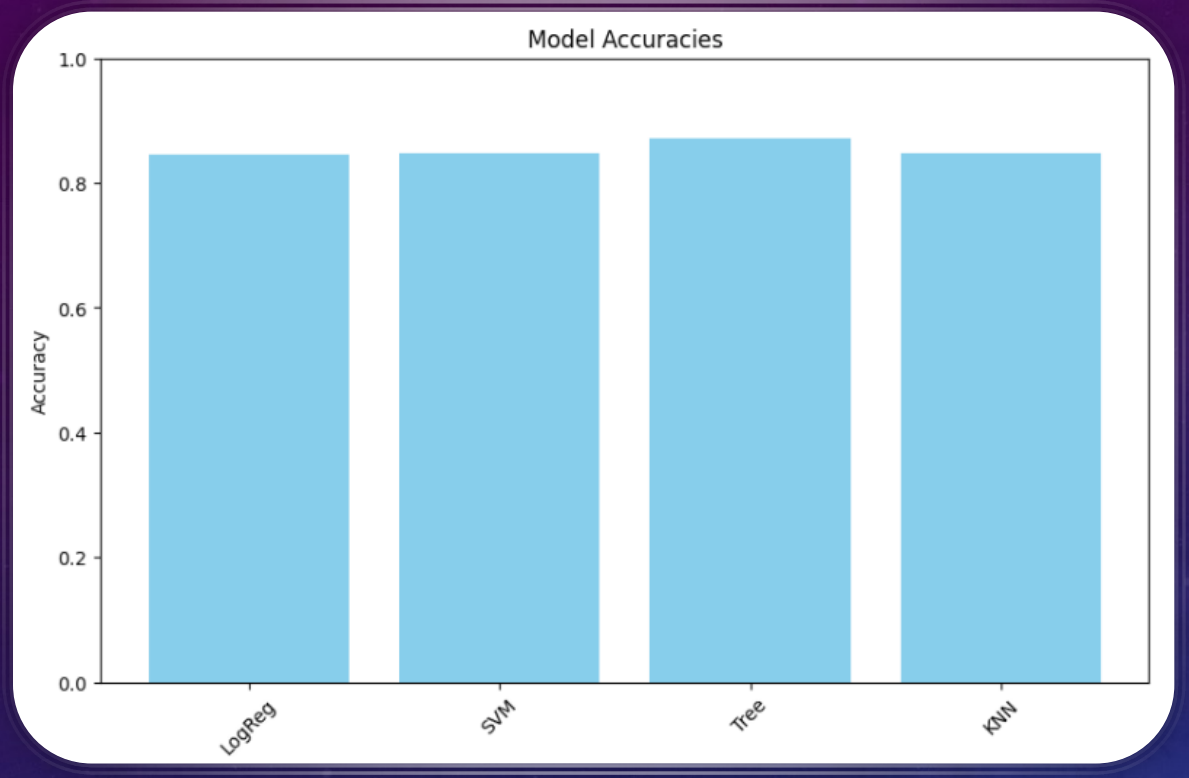


PAYLOAD MASS VS LAUNCH OUTCOME

- This app uses an interactive range slider to select the range of payloads visible in the plot.
- The top plot shows the successes and failures for low payload masses ranging from 0 kg to 5,000 kg.
- The bottom plot shows the successes and failures for high payload masses ranging from 5,000 kg to 10,000 kg.

PREDICTIVE ANALYSIS (CLASSIFICATION)

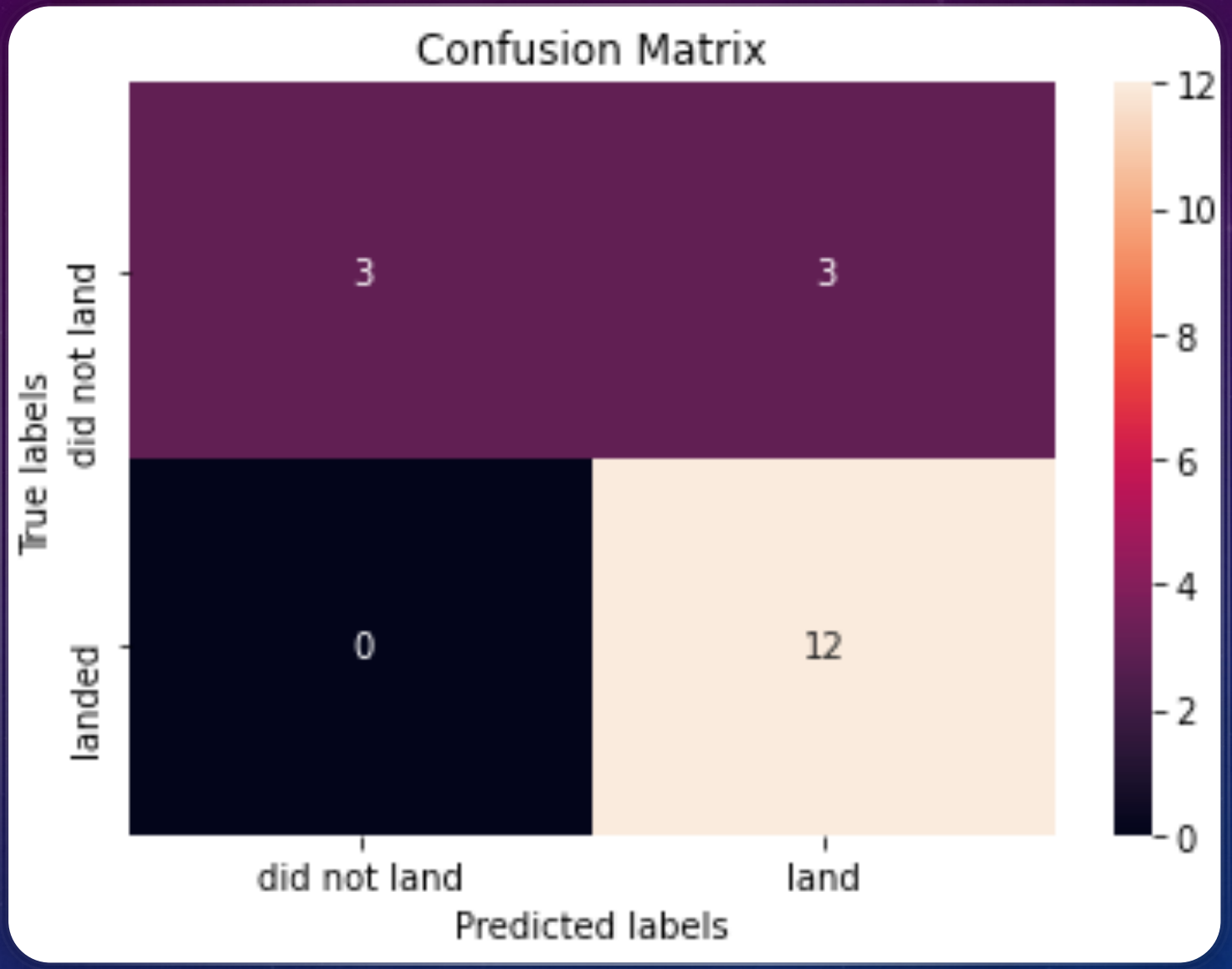
SECTION 6



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.733333	0.800000
F1_Score	0.888889	0.888889	0.846154	0.888889
Accuracy	0.846429	0.848214	0.871429	0.848214

CLASSIFICATION ACCURACY

- Evaluated four machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors.
- This bar chart shows the built model accuracy for all built classification models.
- The model with the highest accuracy is the decision tree with an accuracy of 0.871.



CONFUSION MATRIX

- This confusion matrix for the decision tree classifier demonstrates its ability to differentiate between various classes.
- One significant issue arises with false positives, where the classifier incorrectly marks unsuccessful landings as successful ones which happened 3 times.

CONCLUSIONS

- The more flights that happen at a launch site, the greater the success rate will be at that launch site.
- As payload mass increases, the success rate increases as well.
- KSC LC-39A had the most successful launches making up 41.7% of the total successful launches.
- The decision tree is the most effective algorithm in predicting SpaceX's first-stage reuse across various operational scenarios achieving an accuracy of 0.871.

THANK YOU