IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

DAVID EDNA BUNES
16/07/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

This project utilized SpaceX REST API and web scraping to collect data on rocket launches, enabling the creation of a success/fail outcome variable. Data visualization and SQL techniques were employed to analyze various factors such as payload, launch site, flight number, and yearly trends.

Predictive models including logistic regression, SVM, decision tree, and K-nearest neighbor were developed to forecast rocket landing outcomes.

The findings suggest an improvement in launch success over time, with KSC LC-39A identified as the landing site with the highest success rate. The tested models performed similarly on the test set, with a slight advantage observed for the decision tree model in predictive analytics.

These results offer valuable insights into the factors influencing rocket landing outcomes and contribute to understanding SpaceX's achievements in the space industry.

Introduction

Project background and context

- SpaceX, a prominent player in the space industry, aims to democratize space travel by making it affordable for everyone. They have achieved significant milestones such as delivering payloads to the international space station, deploying an internet satellite network, and conducting manned missions. SpaceX's ability to offer relatively inexpensive rocket launches, costing $62 million per launch compared to competitors' prices of $165 million, is attributed to their groundbreaking practice of reusing the first stage of their Falcon 9 rocket. By utilizing public data and machine learning models, it becomes possible to predict the viability of first stage reuse, which directly impacts the launch cost for SpaceX and other companies in the market.

Introduction

Objectives

- What is the trend of successful landings over time?
- Which predictive model performs best in determining the success or failure of first-stage landings?
- How does payload mass, launch site, number of flights, and orbits influence the success of first-stage landings?

Section 1

Methodology

Methodology

Here are the steps involved in the analysis and modeling process for predicting landing outcomes using SpaceX data:

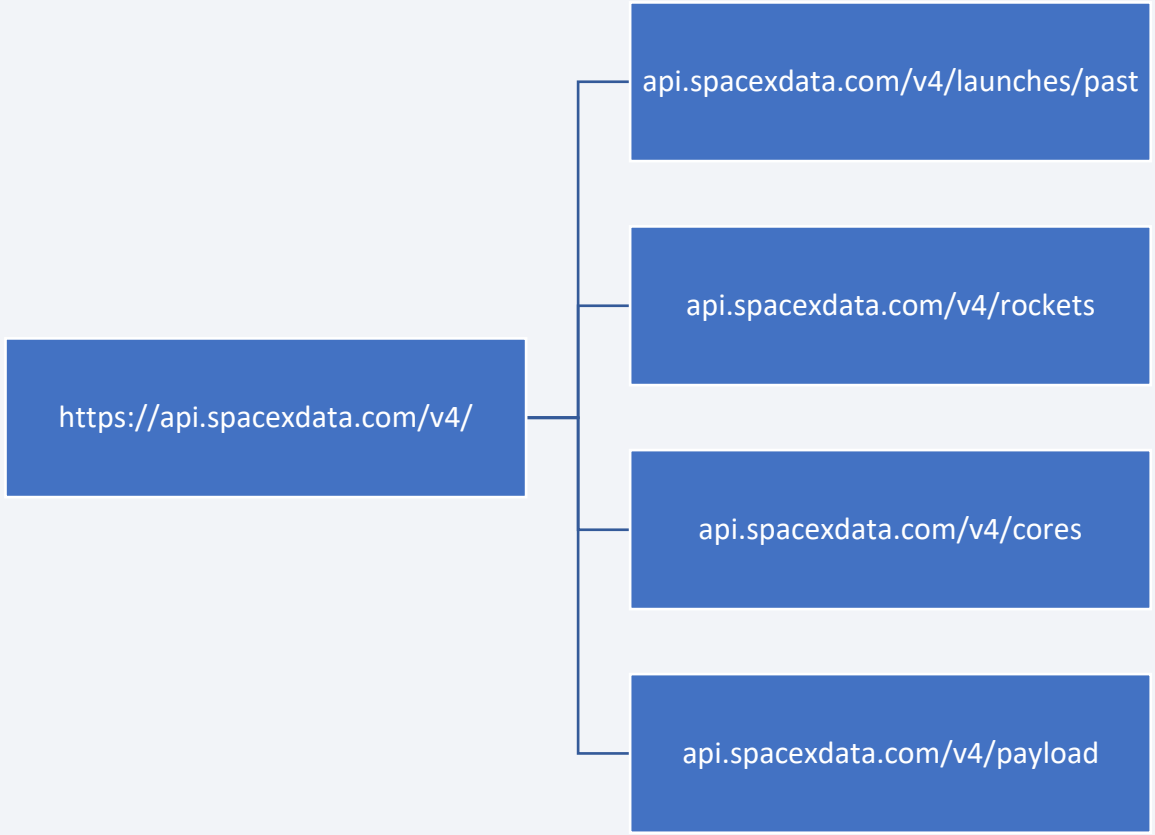
- Data Collection: Gather relevant data using SpaceX REST API and web scraping techniques.
- Data Wrangling: Prepare the data for analysis by filtering, handling missing values, and applying one-hot encoding.
- Exploratory Data Analysis (EDA): Explore the data using SQL queries and data visualization techniques.
- Data Visualization: Utilize Folium and Plotly Dash to create visual representations of the data.
- Model Building: Construct classification models to predict landing outcomes. Fine-tune and evaluate the models to identify the best model and optimal parameters for accurate predictions.

Data Collection – SpaceX API

Here are the steps involved in the data retrieval and preparation process:

- Retrieve rocket launch data from SpaceX API by making a request.
- Utilize custom functions to request specific launch information from the SpaceX API.
- Transform the response into a DataFrame.
- Filter the DataFrame to include only launches related to Falcon 9 rockets.
- Export the prepared data to a CSV file.

https://github.com/ednadavid/IBM_Data_Science_Capstone_Project

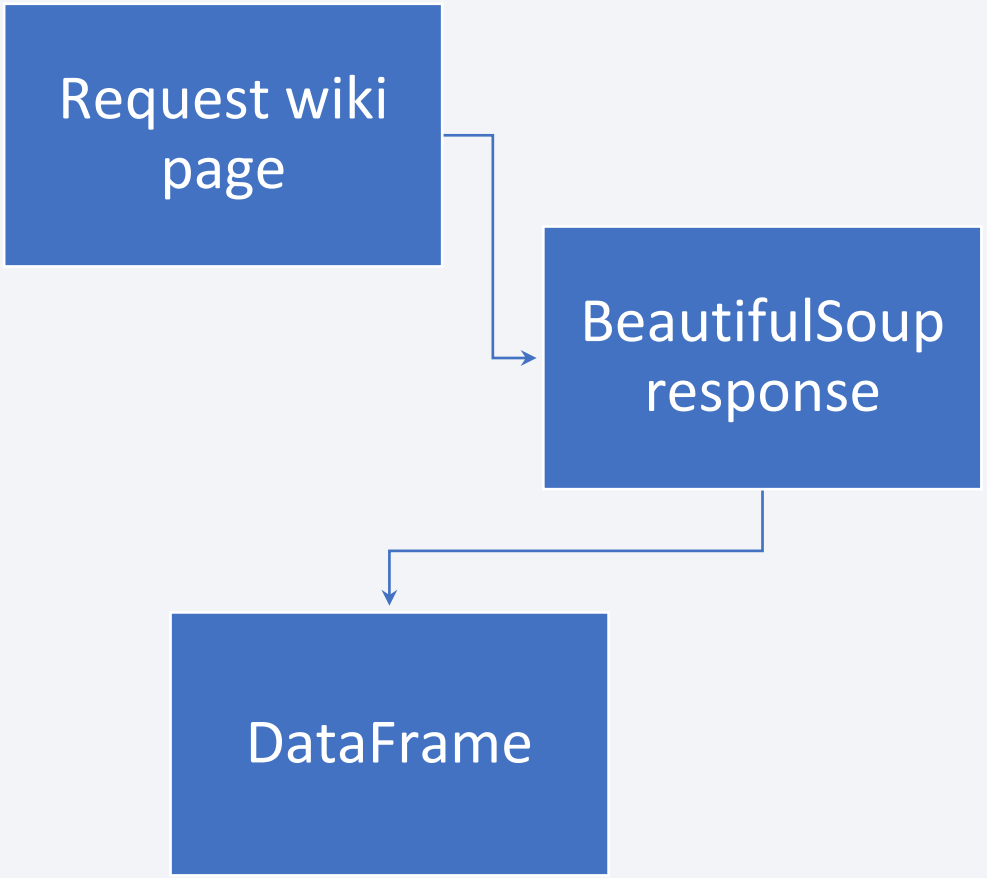


Data Collection - Scraping

Steps taken to retrieve and process Falcon 9 launch data from Wikipedia:

- Request the data from Wikipedia's Falcon 9 launch page.
- Utilize Beautiful Soup to create a structured object from the HTML response.
- Parse the HTML tables to collect the relevant data.
- Move data into a DataFrame.
- Export the processed data to a CSV file.

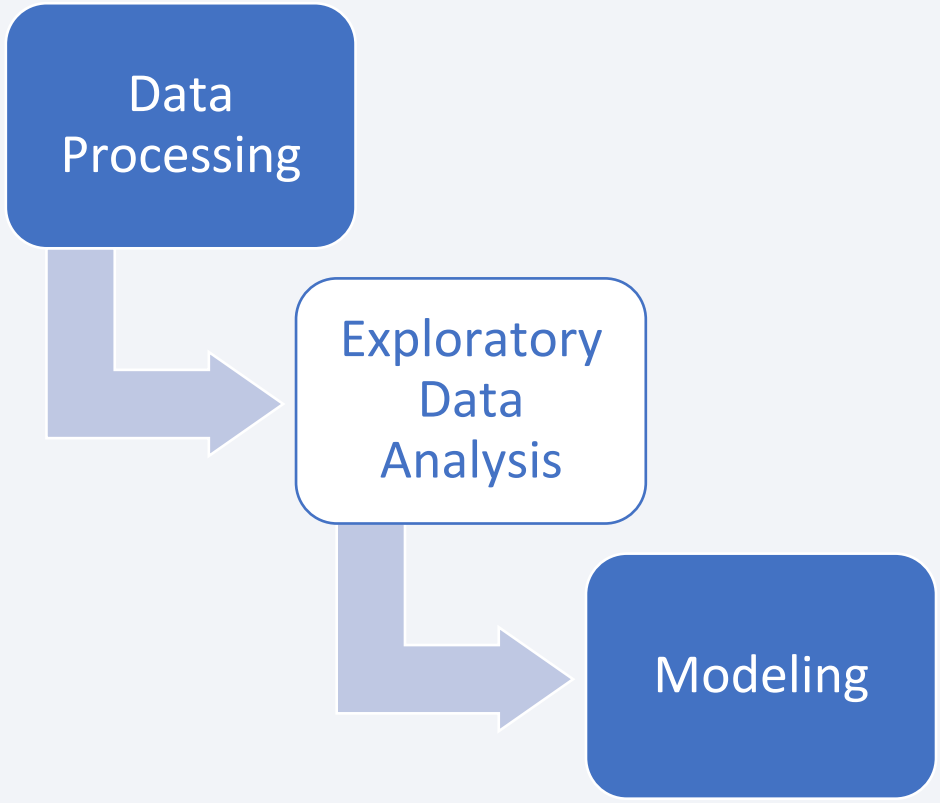
https://github.com/ednadavid/IBM_Data_Science_Capstone_Project



Data Wrangling

- Conduct exploratory data analysis (EDA) and identify data labels.
- Calculate the number of launches for each launch site.
- Determine the count and occurrence of different orbit types.
- Analyze the count and occurrence of mission outcomes for each orbit type.
- Create a binary column to represent the landing outcome (dependent variable).
- Export the data to a CSV file for further analysis.

https://github.com/ednadavid/IBM_Data_Science_Capstone_Project



EDA with Data Visualization

Charts:

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis:

- Utilized scatter plots to examine the relationships between variables and assess their potential usefulness for machine learning applications.
- Employed bar charts to compare and visualize relationships between discrete categories and their corresponding measured values.

EDA with SQL

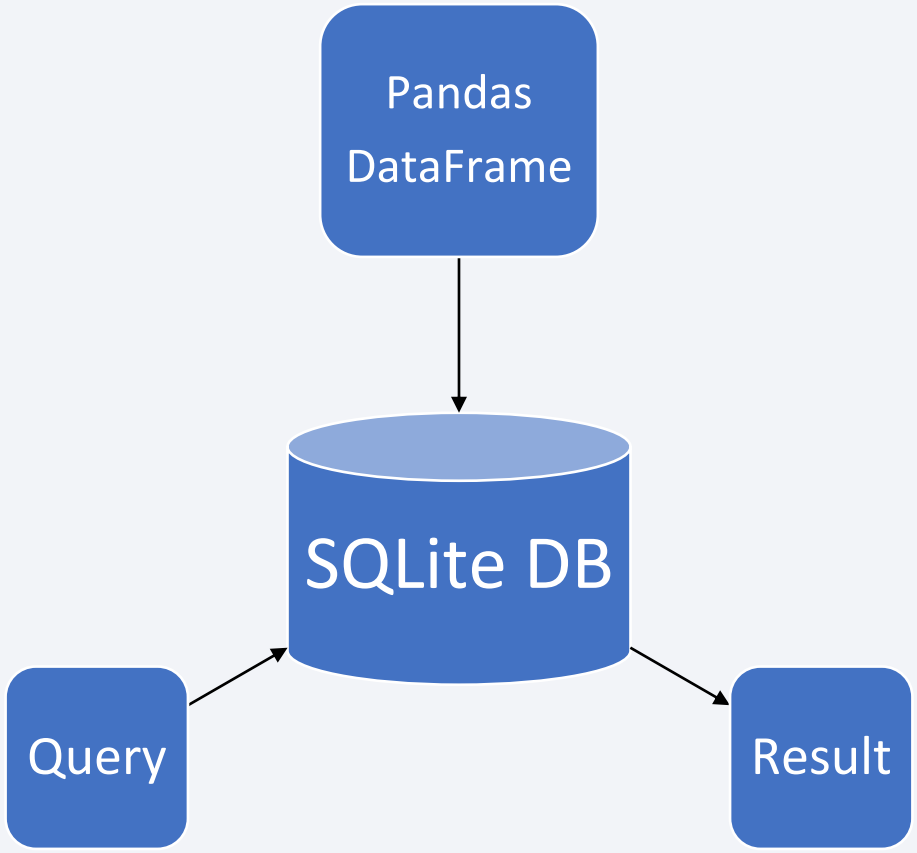
Display:

- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions

https://github.com/ednadavid/IBM_Data_Science_Capstone_Project



Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates

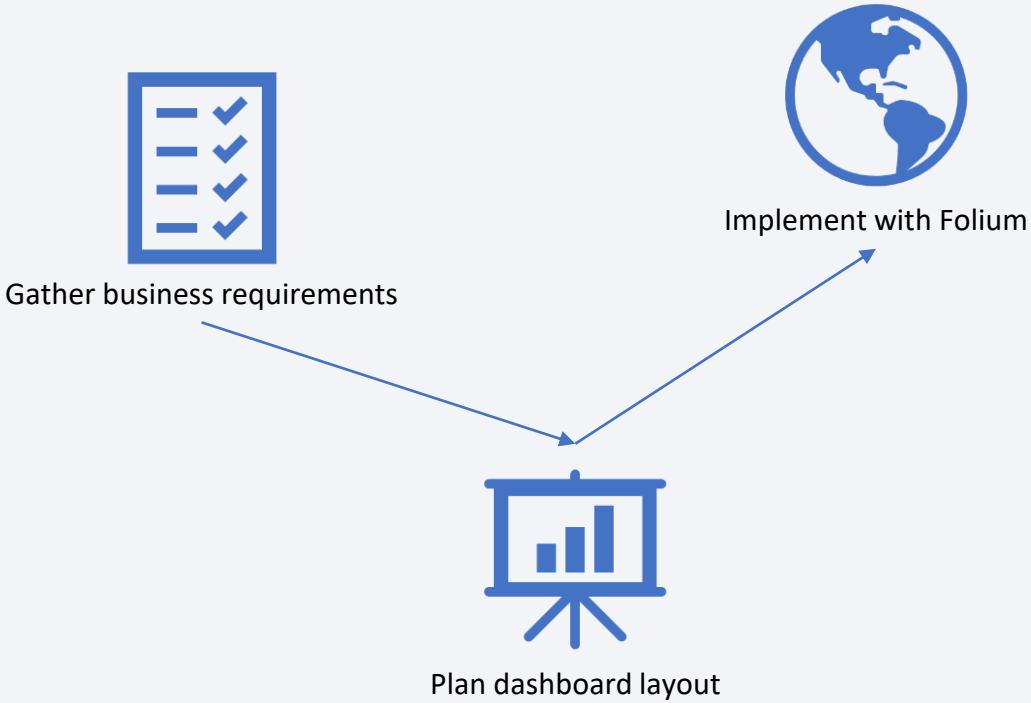
Colored Markers of Launch Outcomes

- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added colored lines to show distance between launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city

https://github.com/ednadavid/IBM_Data_Science_Capstone_Project



Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites
  - Allow user to select all launch sites or a certain launch site
- Slider of Payload Mass Range
  - Allow user to select payload mass range
- Pie Chart Showing Successful Launches
  - Allow user to see successful and unsuccessful launches as a percent of the total
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
  - Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_Score and Accuracy

Results

Exploratory data analysis results

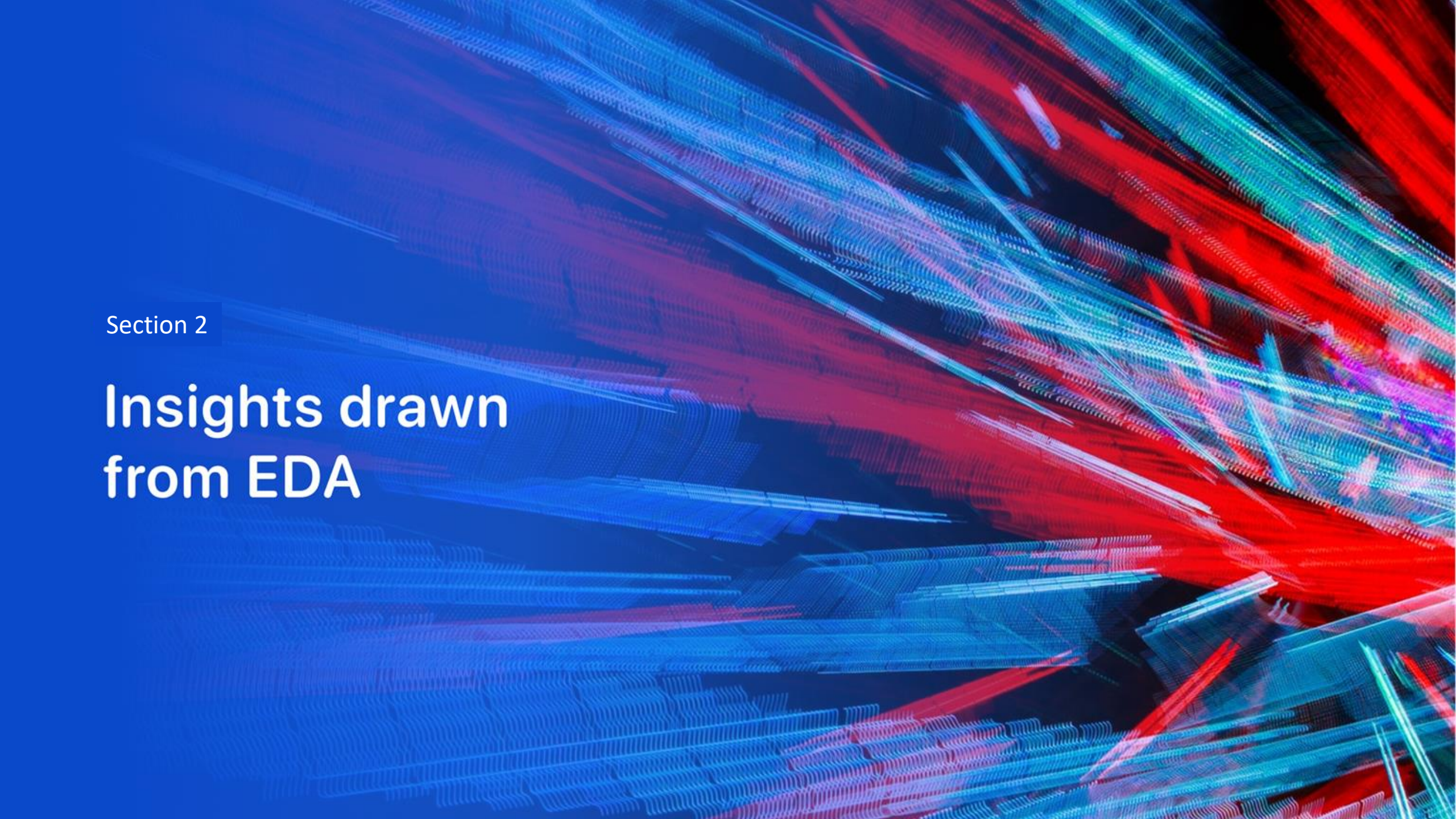
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Predictive analysis results

- Decision Tree model is the best predictive model for the dataset

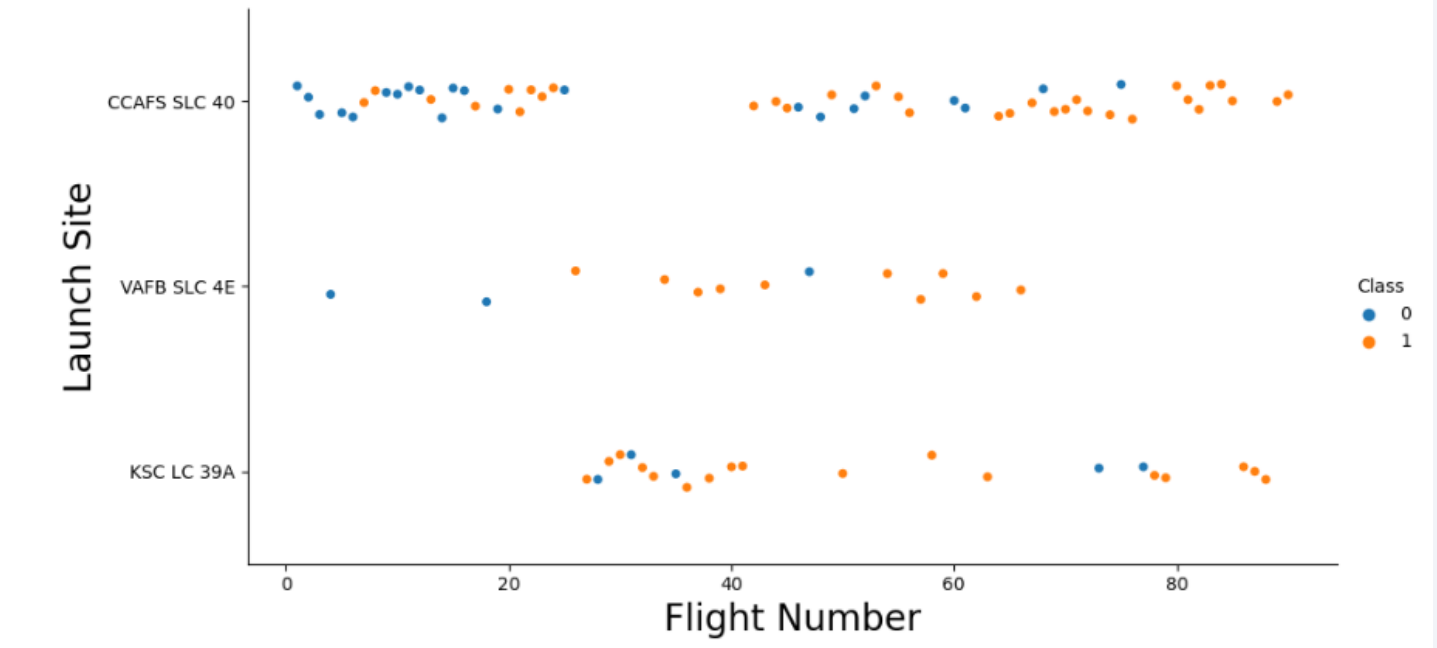
Interactive analytics demo in screenshots

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

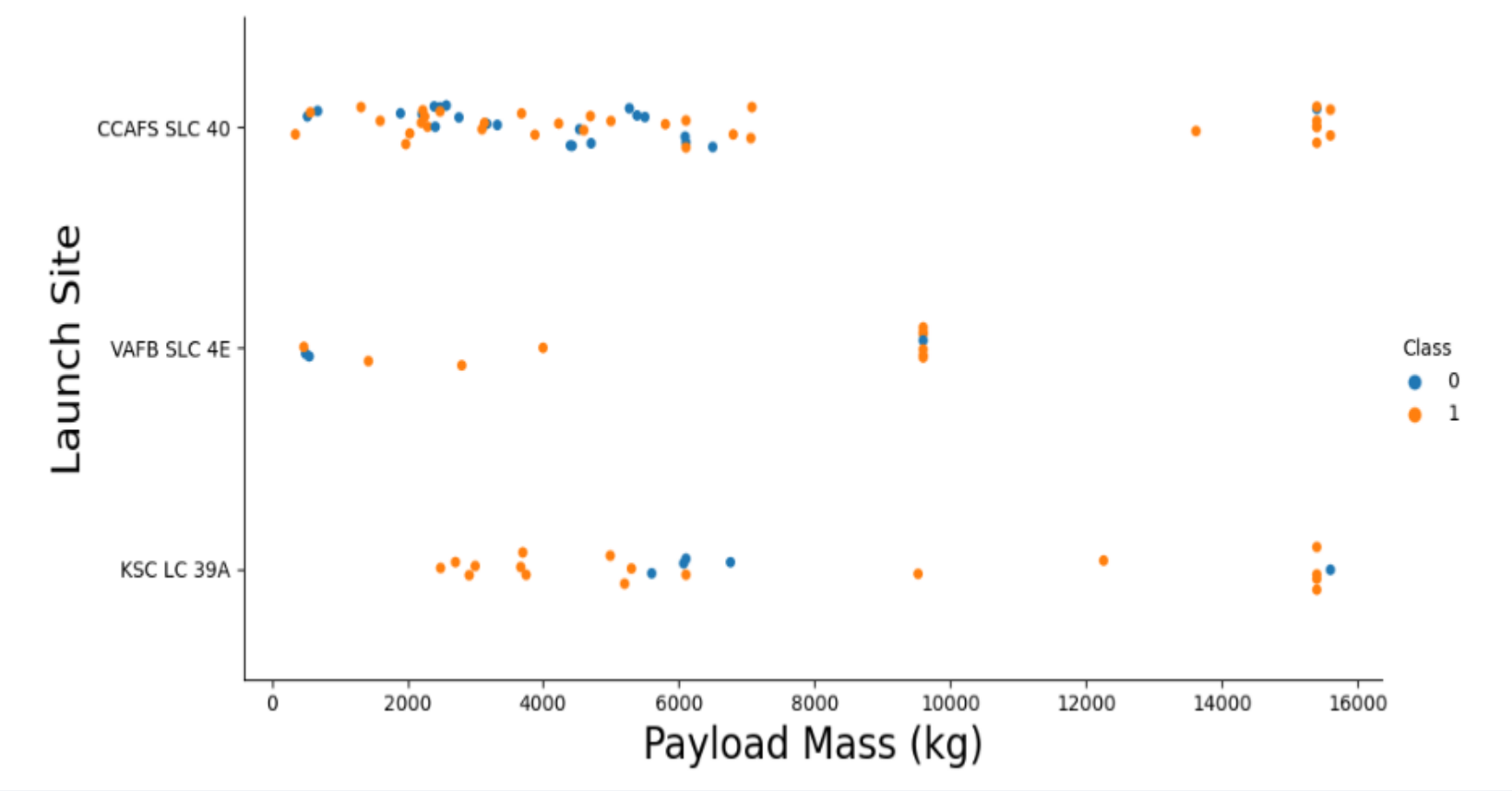

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



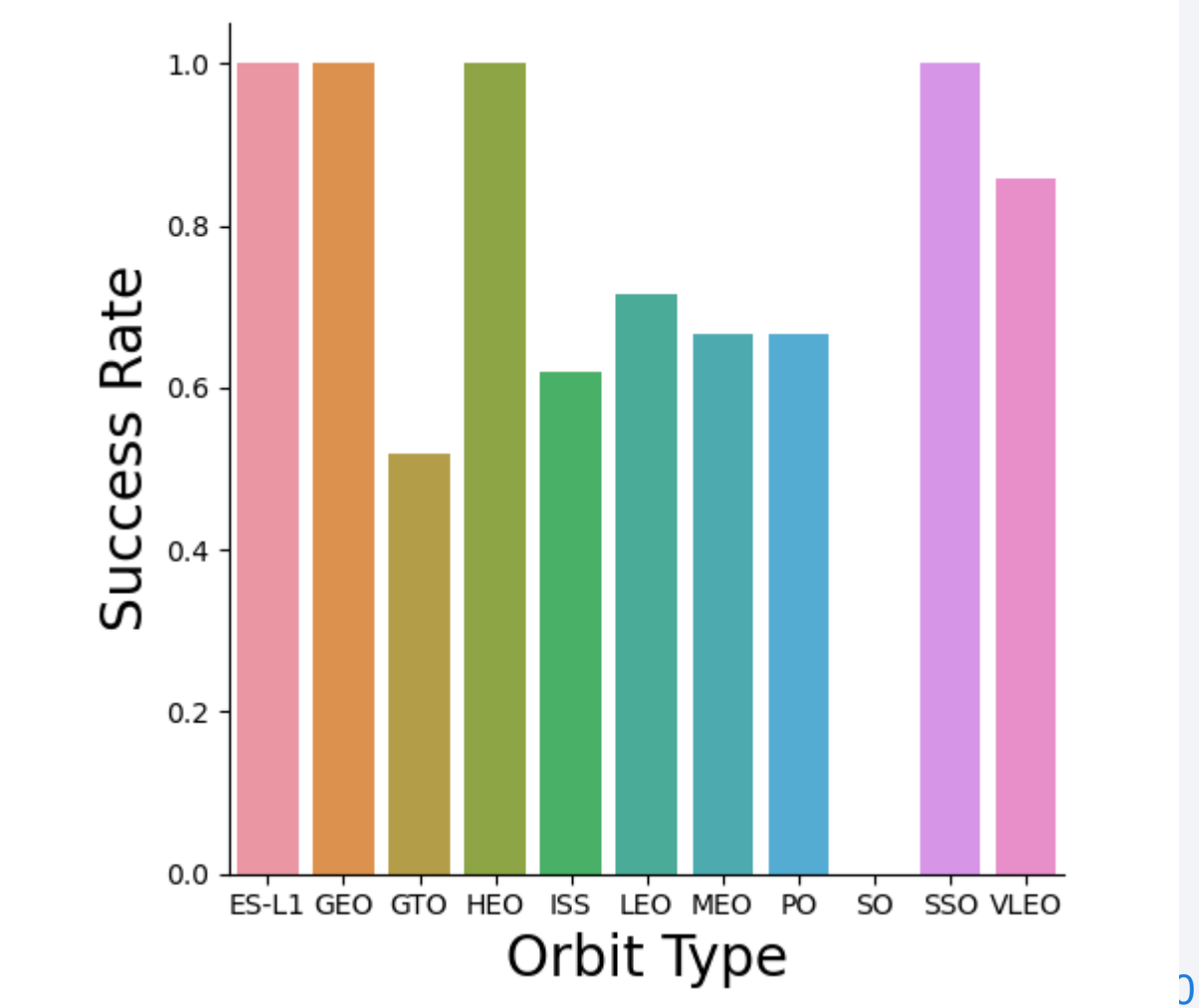
Payload vs. Launch Site



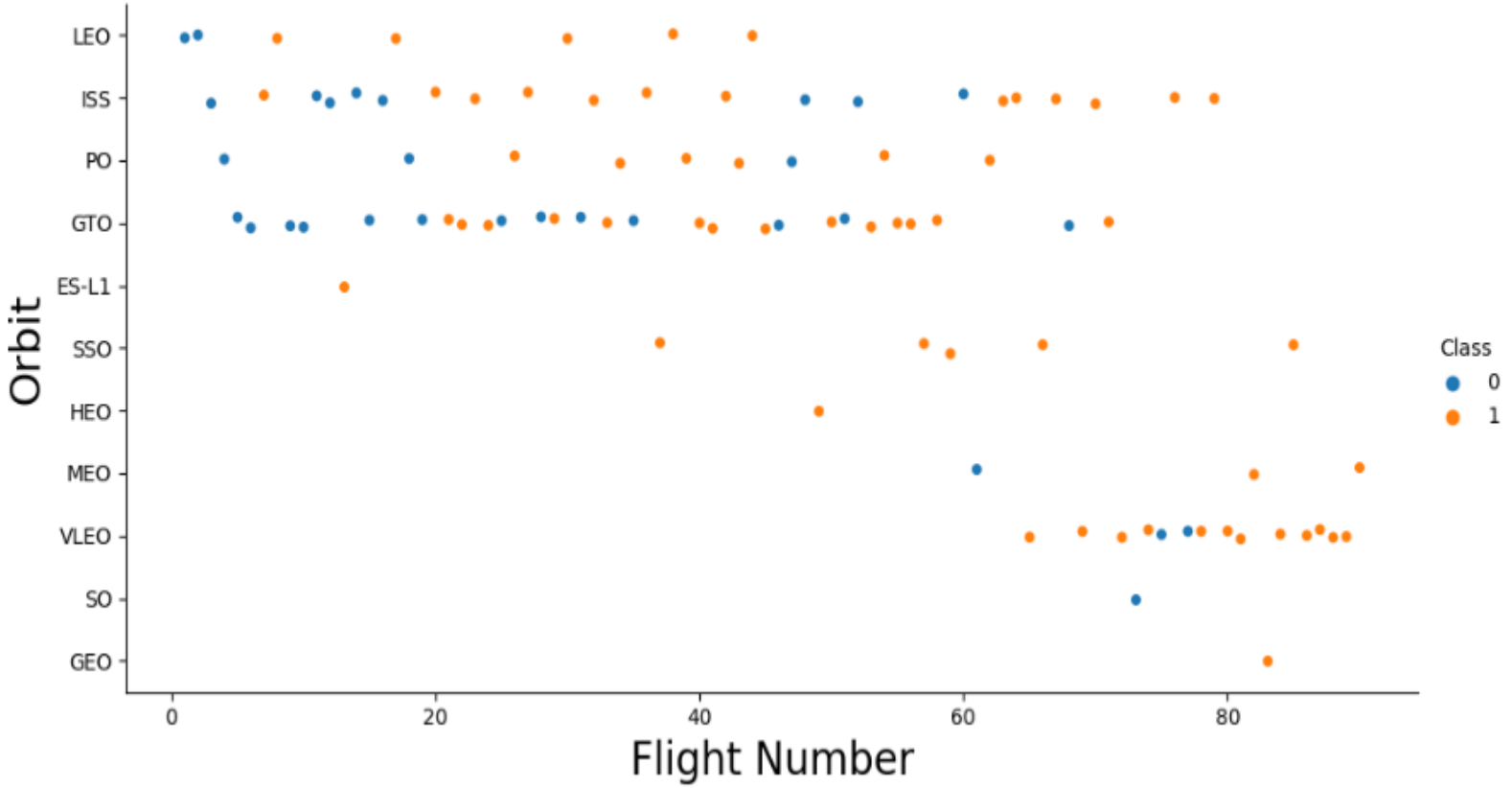
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg

Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



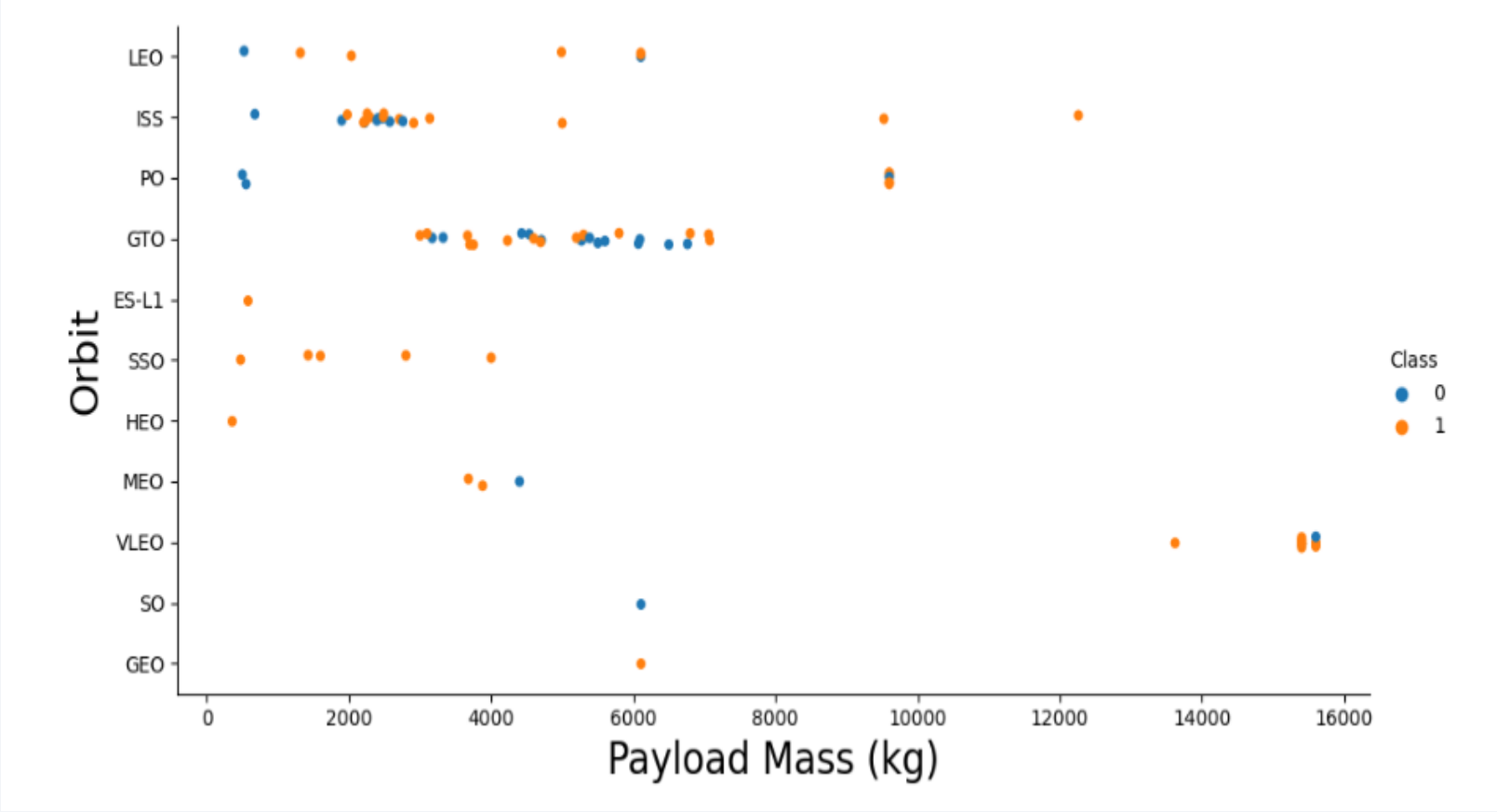
Flight Number vs. Orbit Type



- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend

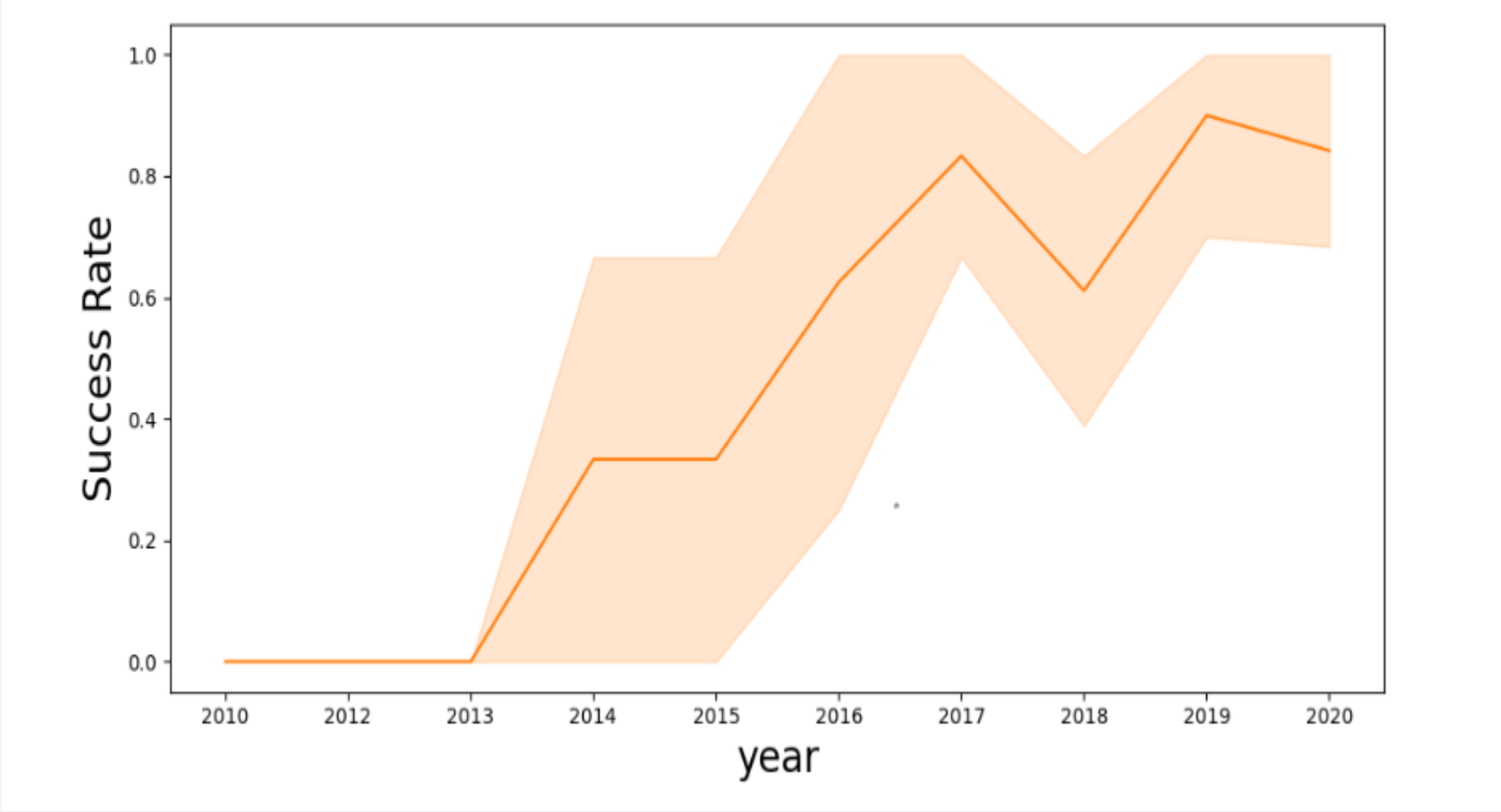
Payload vs. Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

Launch Site Names

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

```
[16]: %%sql

SELECT DISTINCT "Launch_Site"
FROM SPACEXTBL;
```

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

[16]:

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

Launch Site Names Begin with 'CCA'

Displaying 5 records below

```
SELECT *
FROM SPACEXTBL
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

[18]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
[20]: %%sql

      SELECT SUM(PAYLOAD_MASS__KG_)
      FROM SPACEXTBL
      WHERE Customer LIKE 'NASA (CRS)';

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

[20]:

SUM(PAYLOAD_MASS__KG_)
45596.0

- 45,596 kg (total) carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
[22]: %%sql

      SELECT AVG(PAYLOAD_MASS__KG_)
      FROM SPACEXTBL
      WHERE Booster_Version = 'F9 v1.1';

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

[22]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- 2,928 kg (average) carried by booster version F9 v1.1

First Successful Ground Landing Date

```
[23]: %%sql

SELECT MIN(DATE)
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

[23]:

MIN(DATE)
01/08/2018

- 1st Successful Landing in Ground Pad 01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

[25]:
```
%%sql

SELECT Booster_Version
FROM SPACEXTBL
WHERE (LANDING_OUTCOME = 'Success (drone ship)')
& (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

[25]:

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

- F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
[18]: %%sql

SELECT COUNT(Mission_Outcome), Mission_Outcome
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

[18]:

COUNT(Mission_Outcome)	Mission_Outcome
0	None
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

- Overall, the total number of successful missions is 99 while, that of failed missions is 1.

Boosters Carried Maximum Payload

```
[15]: %%sql

SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ =
    (SELECT MAX(PAYLOAD_MASS__KG_)
     FROM SPACEXTBL);
```

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

[15]:

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

- 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

Showing month, date, booster version, launch site and landing outcome

```
[16]: %%sql

SELECT substr(Date,4,2) as month, DATE, BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome
FROM SPACEXTBL
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

[16]:

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

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

Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
SELECT Landing_Outcome, count(*) as count_outcomes
FROM SPACEXTBL
WHERE DATE between '04-06-2010' and '20-03-2017' group by Landing_Outcome order by count_outcomes DESC;
```

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

[17]:

Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

Launch Sites

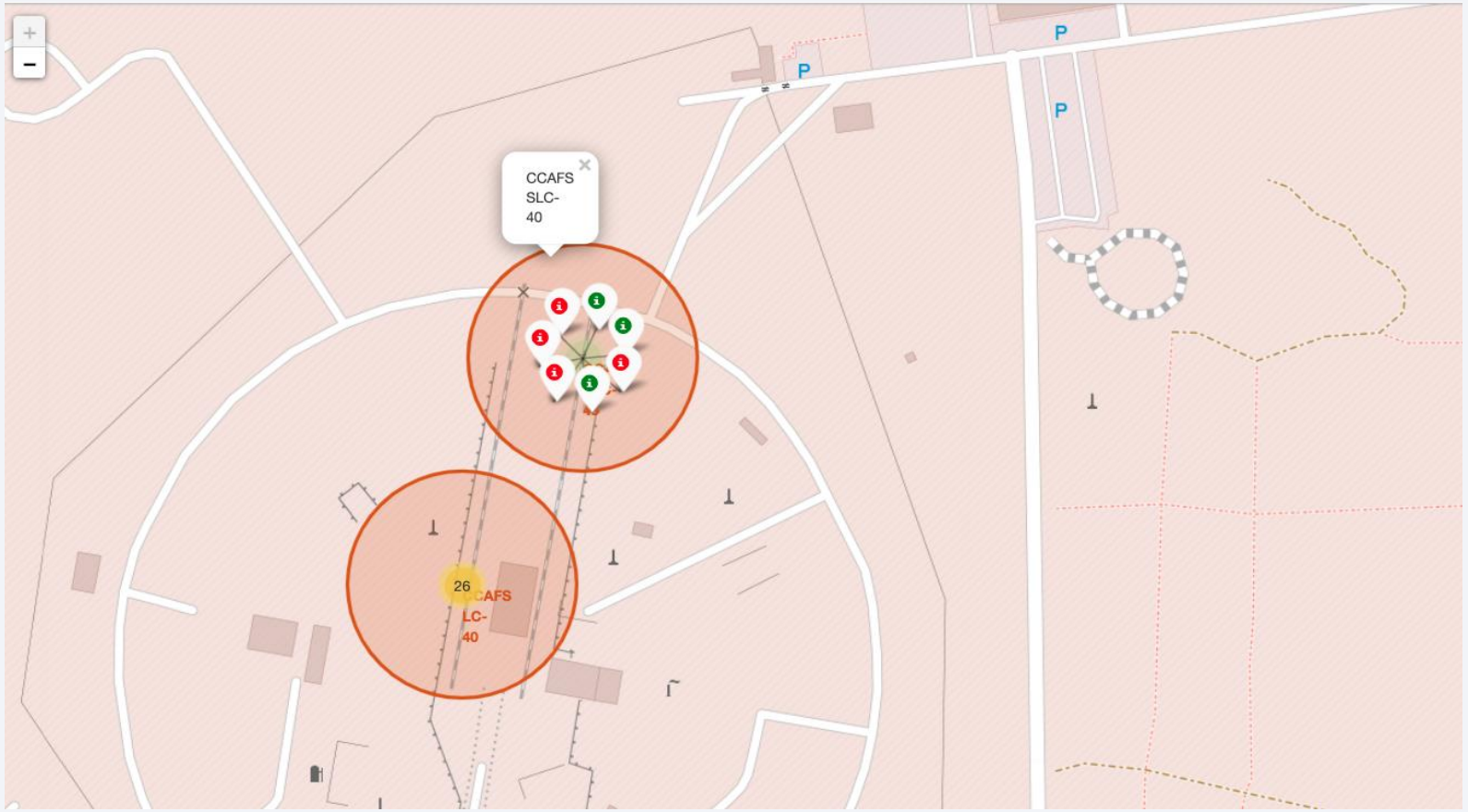
Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.

Launch Outcomes

At Each Launch Site

Outcomes:

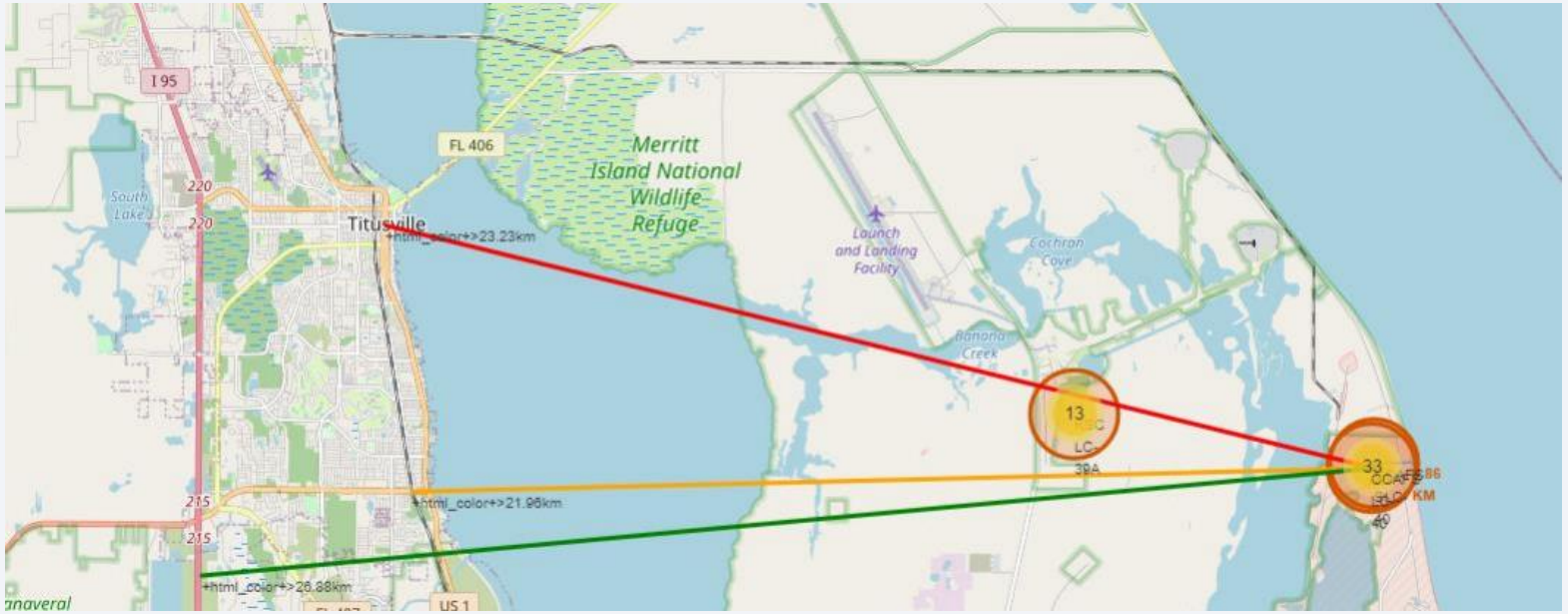
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

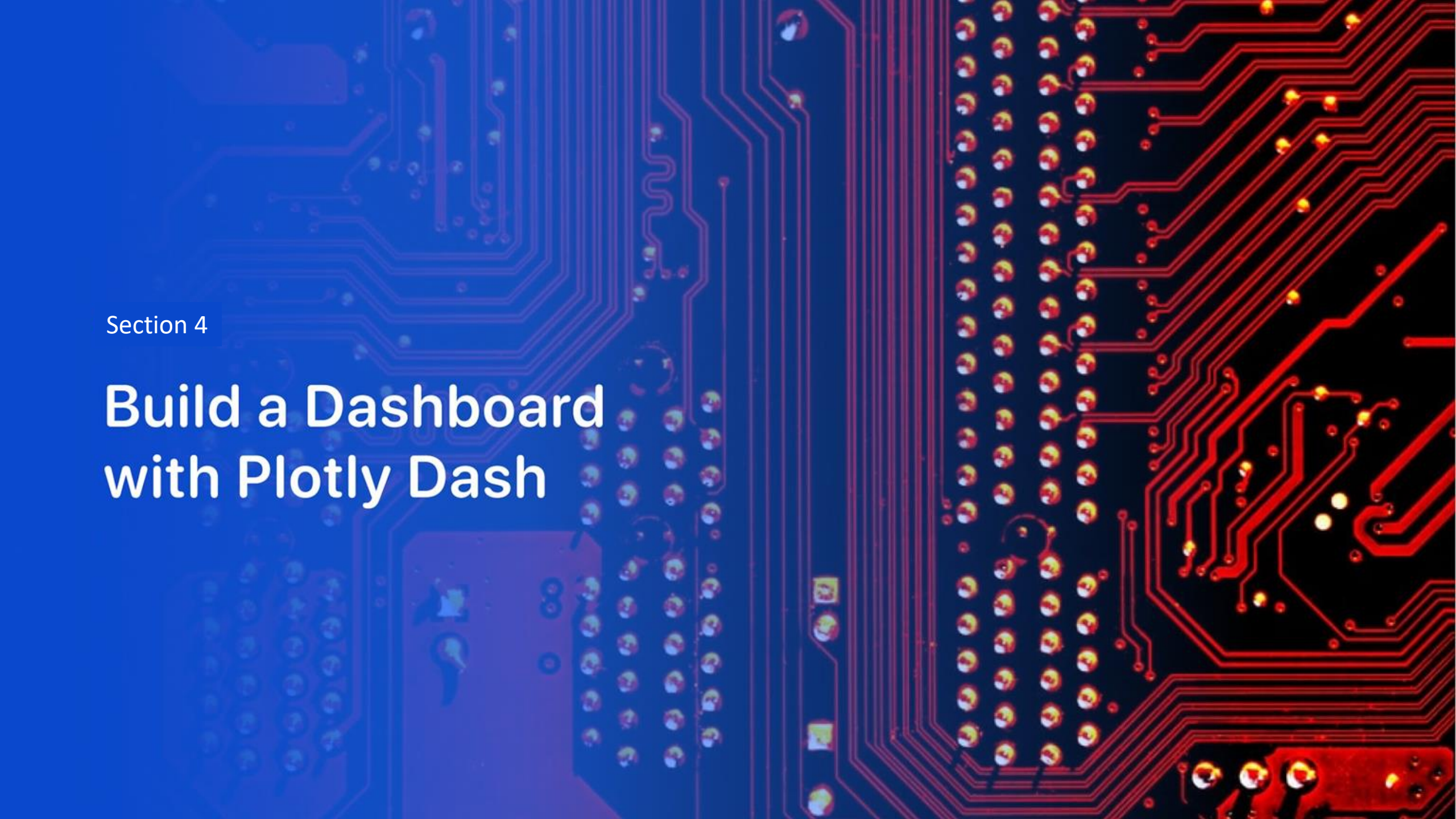


Distance to Proximities

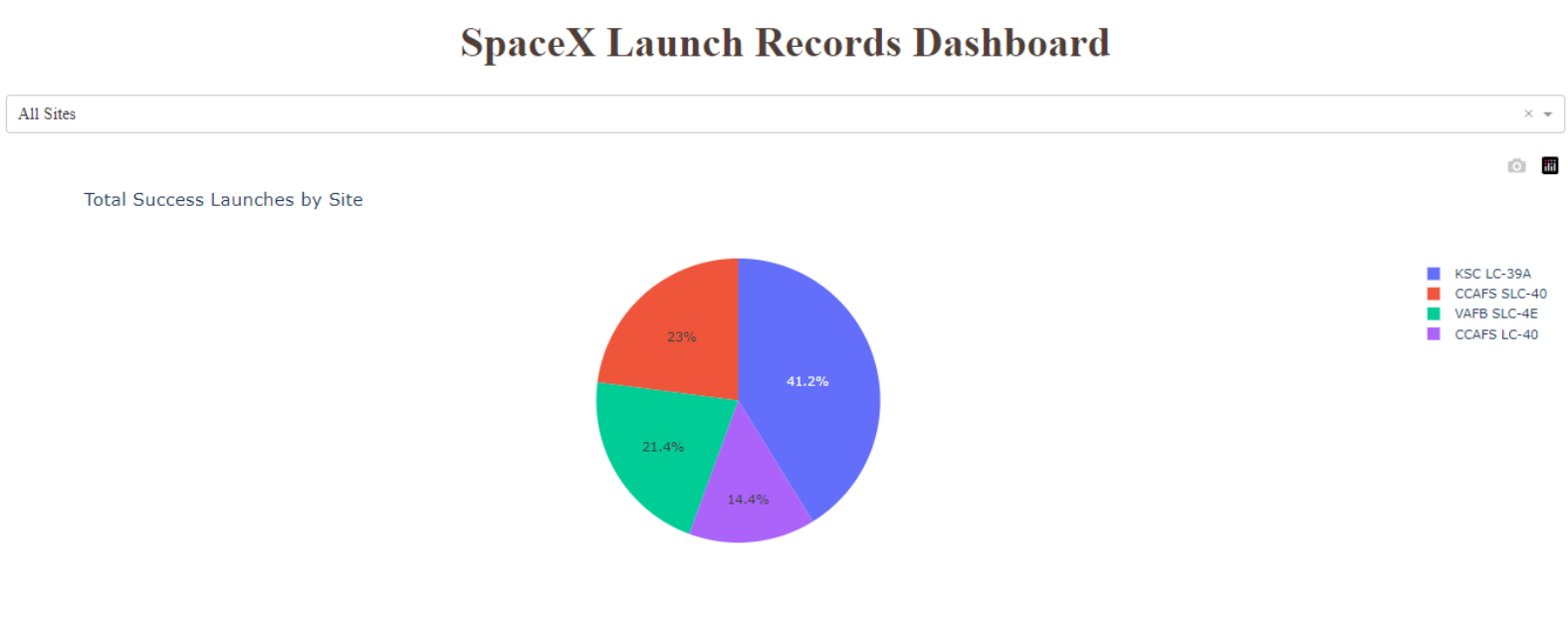
CCAFS SLC-40

- .86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway



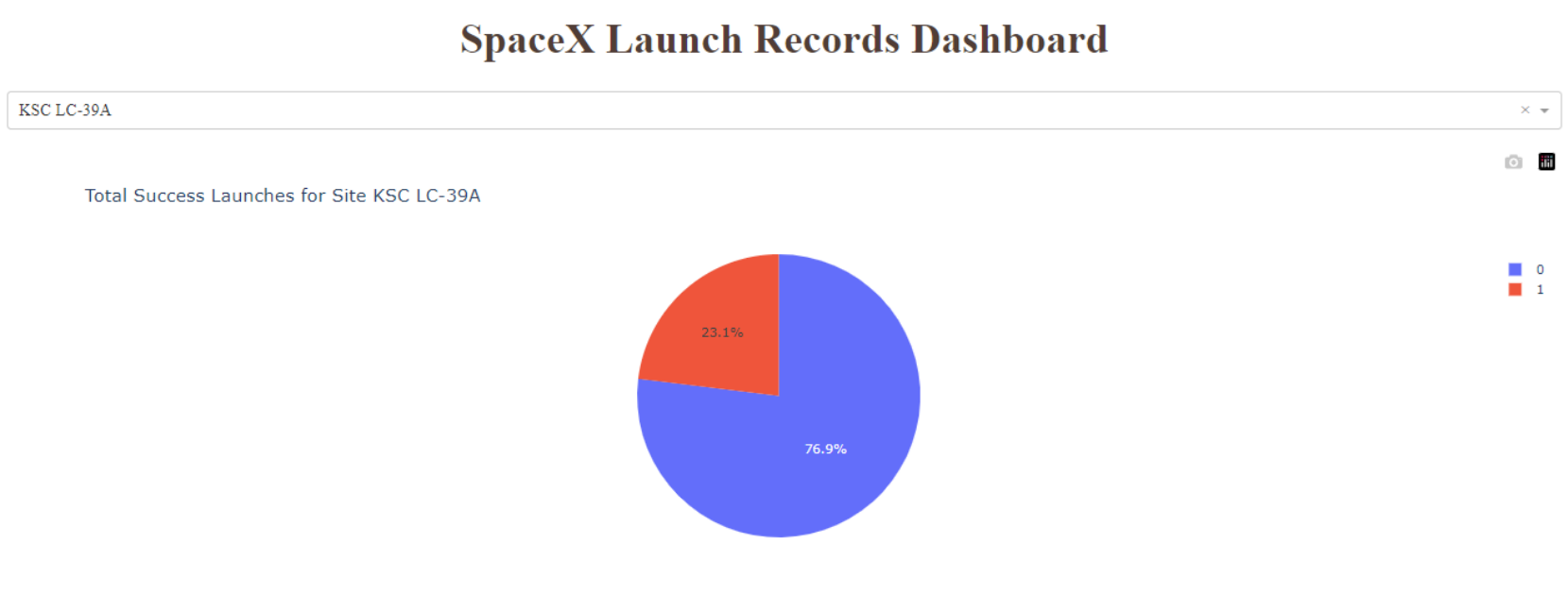

Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site



Success as Percent of Total:
KSC LC-39A has the most successful launches amongst launch sites (41.2%)

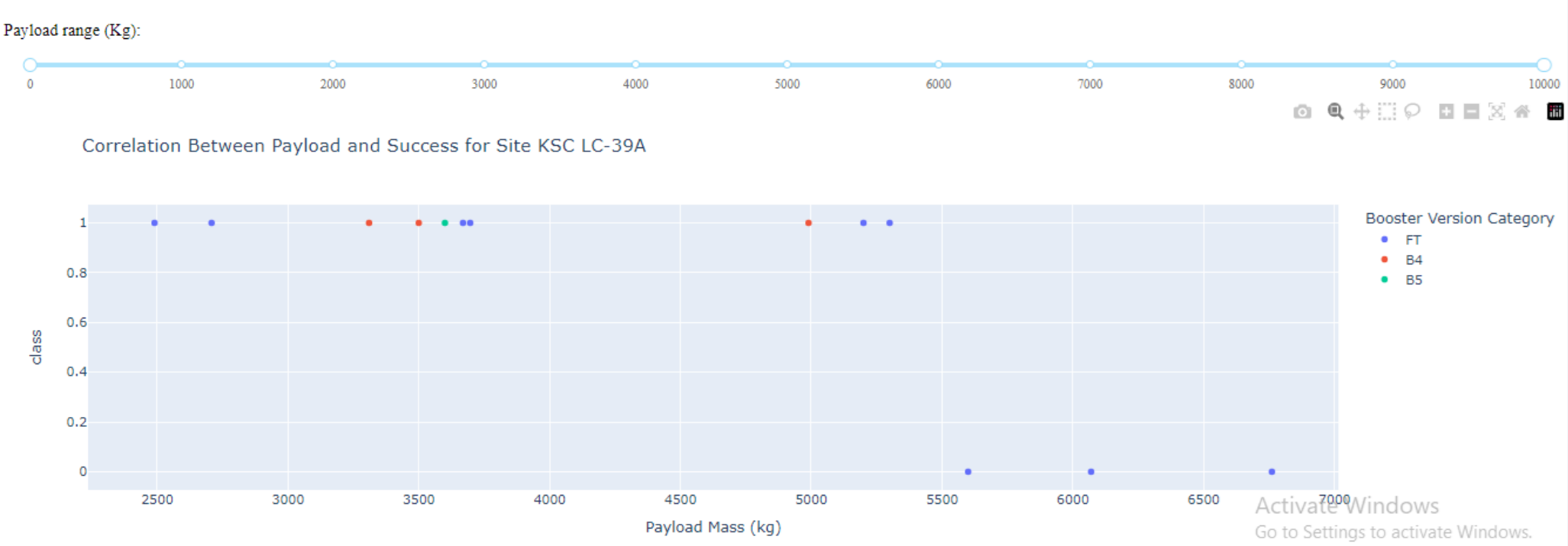
Launch Success (KSC LC-39A)



Success as Percent of Total:

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches

Payload Mass and Success



- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

Section 5

Predictive Analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

.best_score_ is the average of all cv folds for a single combination of the parameters

Accuracy

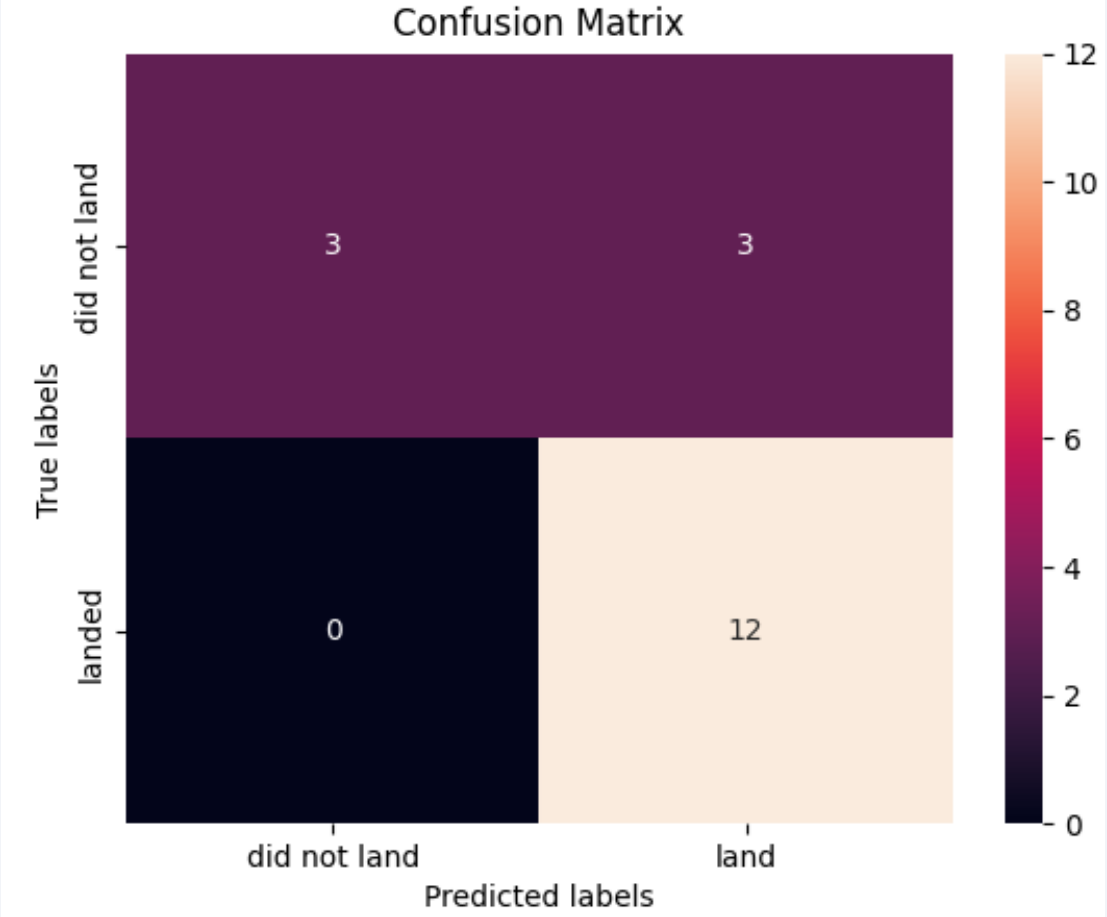
All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset.

The Decision Tree model slightly outperformed the rest when looking at .best_score_

Confusion Matrix

A confusion matrix summarizes the performance of a classification algorithm

- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative



Conclusions

- Model Performance: The decision tree model slightly outperformed other models in the test set.
- Launch Site Factors: Launch sites are strategically located near the equator to leverage the Earth's rotational speed, minimizing fuel and booster requirements. Additionally, all launch sites are situated close to the coast.
- Increasing Launch Success: Over time, the overall success rate of launches has shown an upward trend.
- KSC LC-39A: Among all launch sites, KSC LC-39A exhibits the highest success rate. Specifically, it achieves a perfect 100% success rate for launches with payloads less than 5,500 kg.
- Payload Mass Impact: Irrespective of launch site, higher payload masses (kg) are associated with higher success rates.

Thank you!