IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Priyaranjan Mishra
12th August 2023

Outline

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

Executive Summary

Summary of methodologies

- The research aims to identify the key determinants of a successful rocket landing.
- To achieve this goal, the study employed the following methodologies:

i. Data collection was carried out using both the SpaceX REST API and web scraping techniques.

ii. Data was processed to create a binary outcome variable indicating success or failure of rocket landings.

iii. The data was then explored using data visualization methods, focusing on factors such as payload, launch site, flight number, and yearly trends.

iv. Statistical analysis was conducted using SQL to compute metrics such as total payload, payload range for successful landings, and counts of successful and failed outcomes.

v. The success rates of different launch sites were examined, including their proximity to geographical markers.

vi. Launch sites with the highest success rates and successful payload ranges were visualized

vii. Various machine learning models, including logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN), were developed to predict landing outcomes.

Summary of all results

- Exploring the Data:

i. The analysis revealed a progressive enhancement in launch success over the years.

ii. Among the landing sites, KSC LC-39A exhibited the most remarkable success rate.

iii. Certain orbital destinations like ES-L1, GEO, HEO, and SSO attained a perfect 100% success rate.

- Visualizing and Analyzing:

i. Majority of launch sites are situated near the equator and in proximity to coastal areas.

- Predictive Analysis:

i. All the predictive models displayed comparable performance on the test dataset.

ii. Notably, the decision tree model exhibited a slightly superior performance compared to the others.

INTRODUCTION

Project background and context

- SpaceX, a prominent leader in the space sector, is committed to democratizing the affordability of space travel.
- Their accomplishments encompass tasks like sending spacecraft to the international space station, establishing a satellite network to deliver internet services, and executing crewed space missions.
- The key to SpaceX's cost-effectiveness is their innovative approach: they recycle the initial stage of the Falcon 9 rocket, leading to relatively inexpensive launches at around $62 million each.
- In contrast, other companies that lack reusability capabilities face significantly higher launch costs, often exceeding $165 million per launch.
- The determining factor for launch pricing lies in predicting the successful landing of the first rocket stage.
- By utilizing publicly accessible data and employing machine learning models, it becomes possible to forecast whether either SpaceX or its competitors can achieve successful reuse of the initial rocket stage.

Problems you want to find answers

Examining	Analyzing	Identifying
Examining how the weight of the payload, the chosen launch site, the frequency of flights, and the specific orbital paths influence the likelihood of a successful first-stage landing.	Analyzing the historical trend in successful landing rates of the first stage over a period of time.	Identifying the optimal predictive model for accurately determining the success or failure of first-stage landings through binary classification.

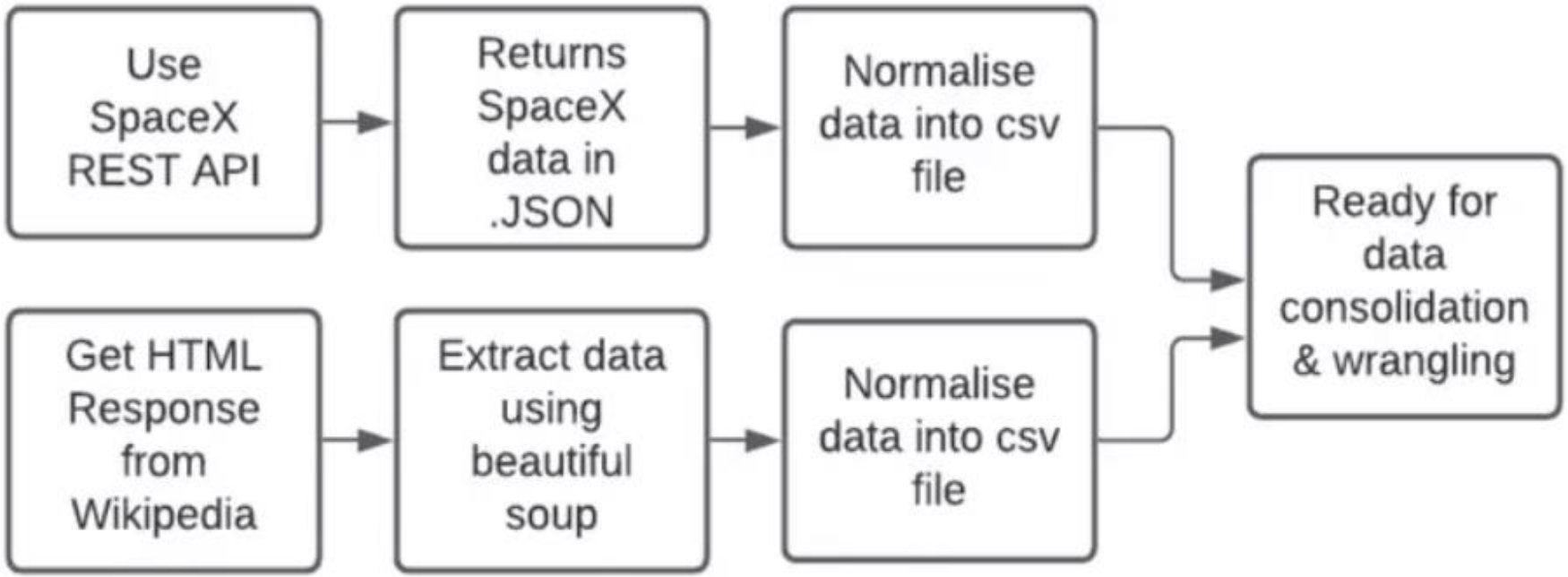
Section 1

Methodology

Methodology

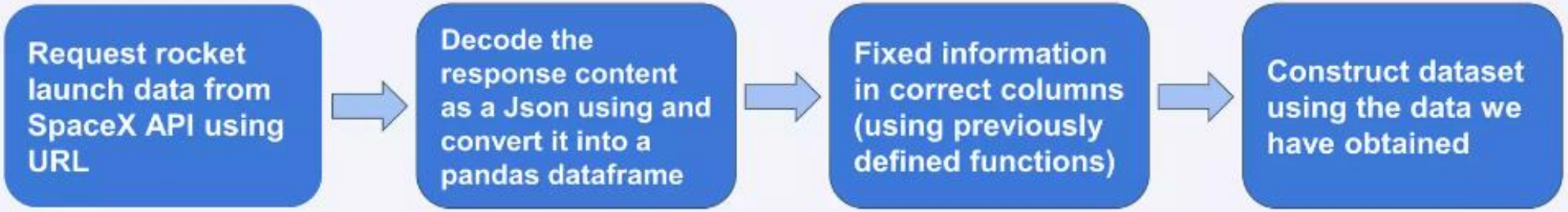
- **Executive Summary**
- **Data collection methodology:**
  - **Collect data using SpaceX REST API and web scraping techniques**
- **Perform data wrangling**
  - **Wrangle data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling**
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - **Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameter**

Data Collection



Data Collection – SpaceX API

- Retrieve rocket launch data by making requests to the SpaceX API.
- Decode the API response using the .json() method and transform it into a dataframe using .json_normalize().
- Utilize custom functions to fetch launch details from the SpaceX API.
- Construct a dictionary using the obtained information.
- Convert the dictionary into a dataframe.
- Apply filters to the dataframe to retain only launches involving Falcon 9 rockets.
- Fill in any missing Payload Mass values with the computed mean.
- Export the data to a CSV file for storage.



Data Collection - SpaceX API Github link

Data Collection – Scraping

- Retrieve Falcon 9 launch data from Wikipedia through data requests.
- Generate a BeautifulSoup object from the HTML response for parsing.
- Extract column names by analyzing the headers of HTML tables.
- Gather relevant data by parsing the HTML tables.
- Compile the collected data into a dictionary format.
- Convert the dictionary into a structured dataframe.
- Save the data to a CSV file for external storage.

Data Collection - Scraping Github link

Data Wrangling

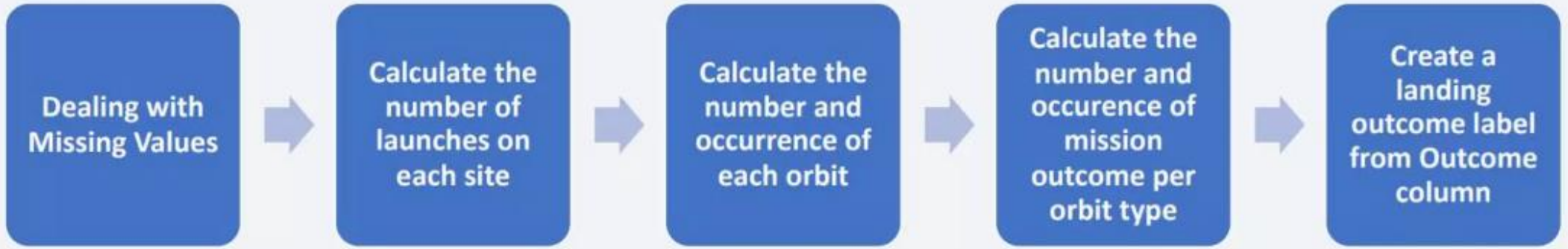
Conduct exploratory data analysis (EDA) to understand the dataset and define data labels.

- Compute the following metrics:

- Number of launches for each launch site.
- Count and frequency of different orbits.
- Count and frequency of mission outcomes based on orbit types.

Generate a new binary column indicating the landing outcome (as the dependent variable).

- Save the updated dataset to a CSV file.



Data Wrangling Github link

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

• View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists

• Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

Data Visualization Github link

EDA with SQL

Queries

Display:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on ground pad
- 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
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

EDA with SQL Github link

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 Map with Folium

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

Interactive map with folium Github link

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

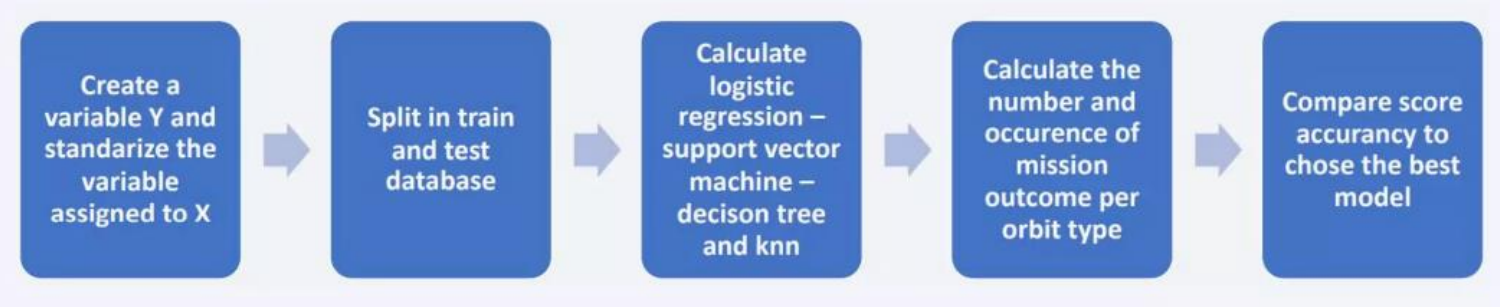
Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

Charts

- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- 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



Predictive Analysis Github link

Results

Exploratory Data Analysis

- 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

Visual Analytics

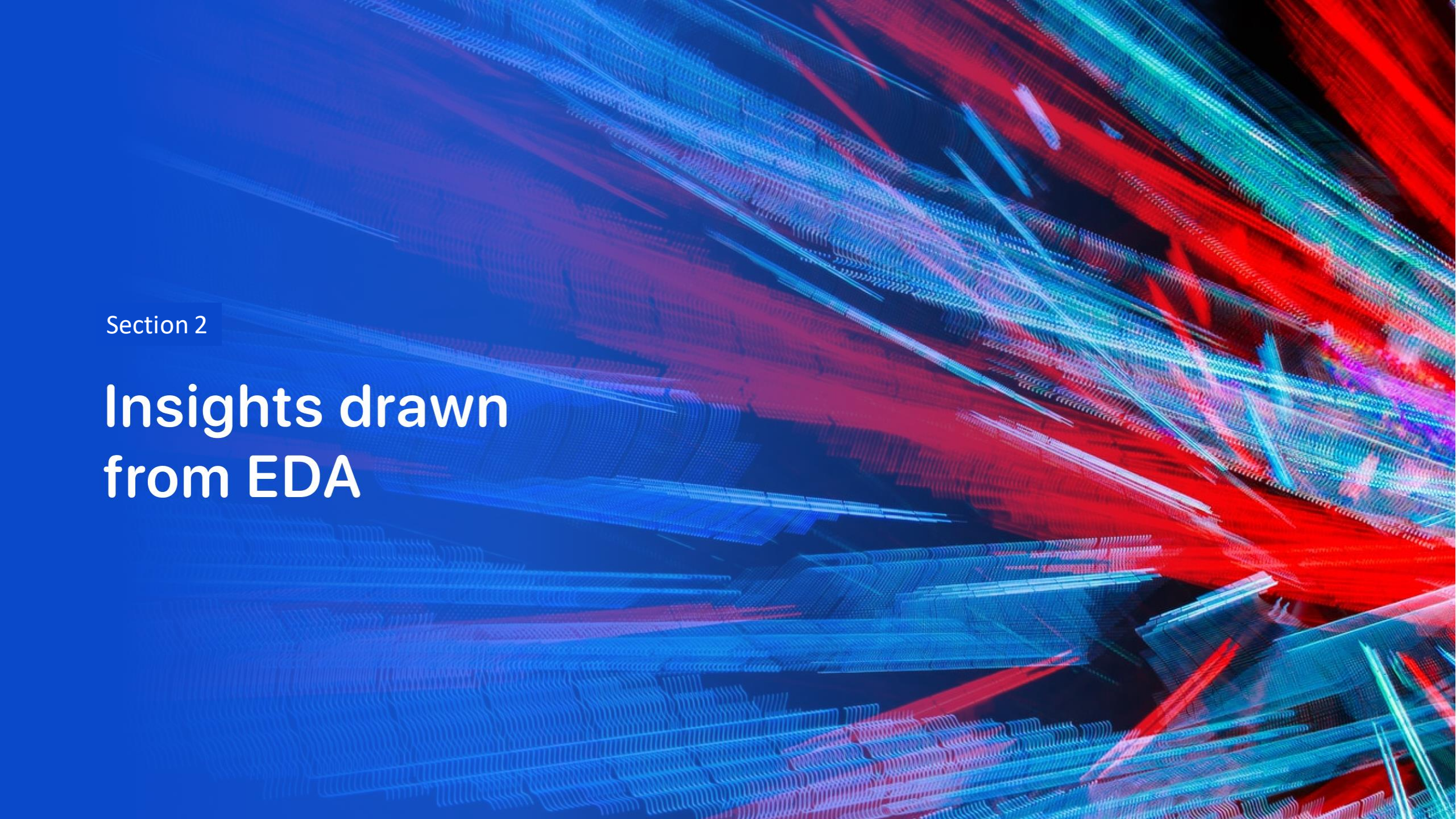
- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

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

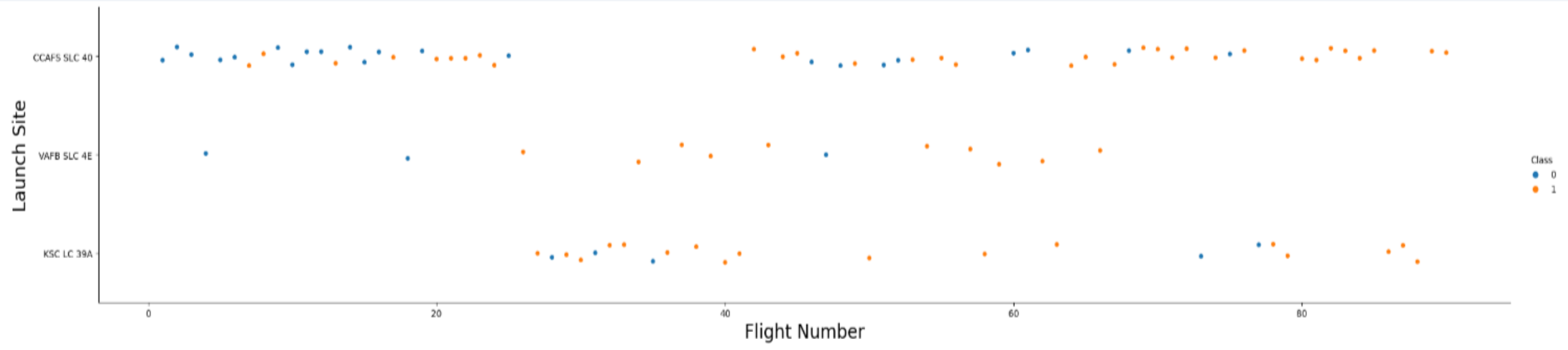
Section 2

Insights drawn from EDA



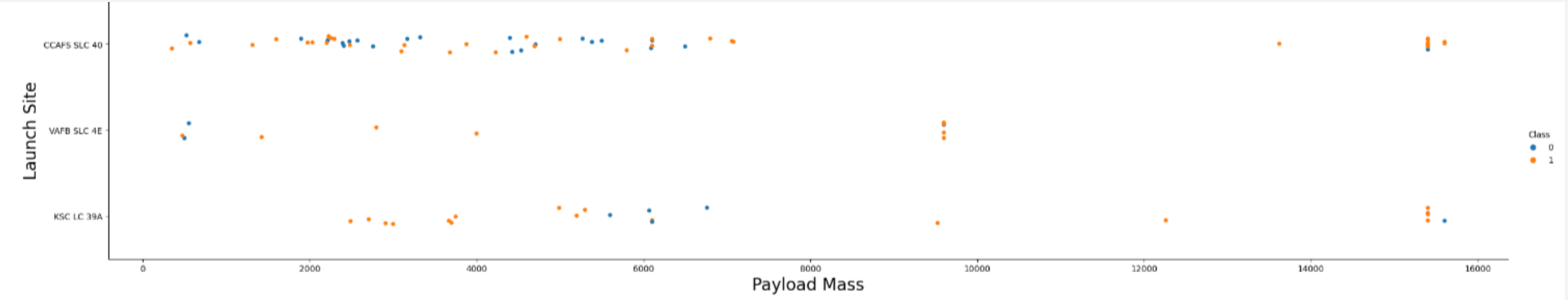
Flight Number vs. Launch Site

- 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



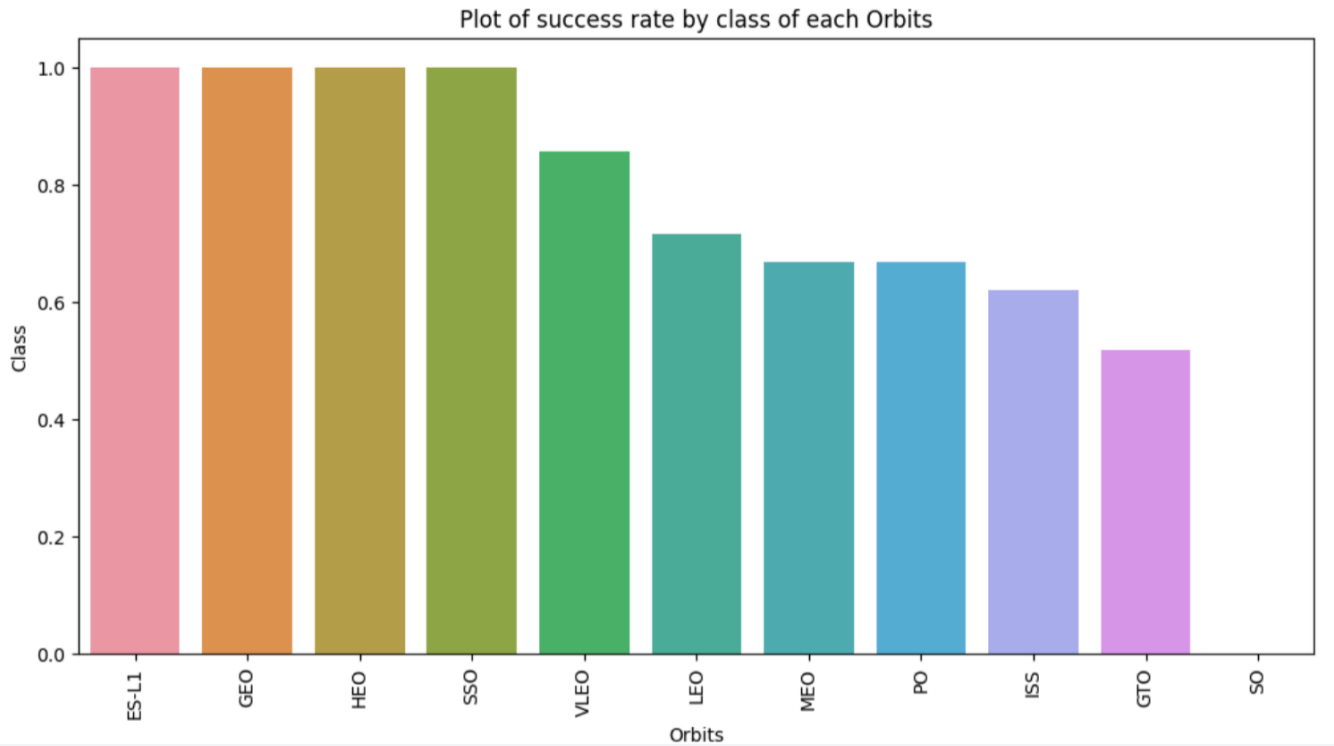
Payload vs. Launch Site

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces 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



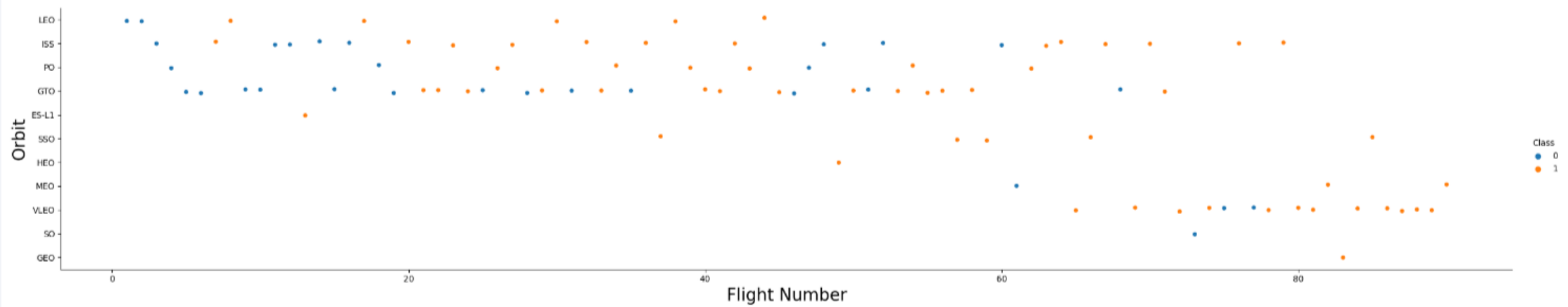
Success Rate vs. Orbit Type

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



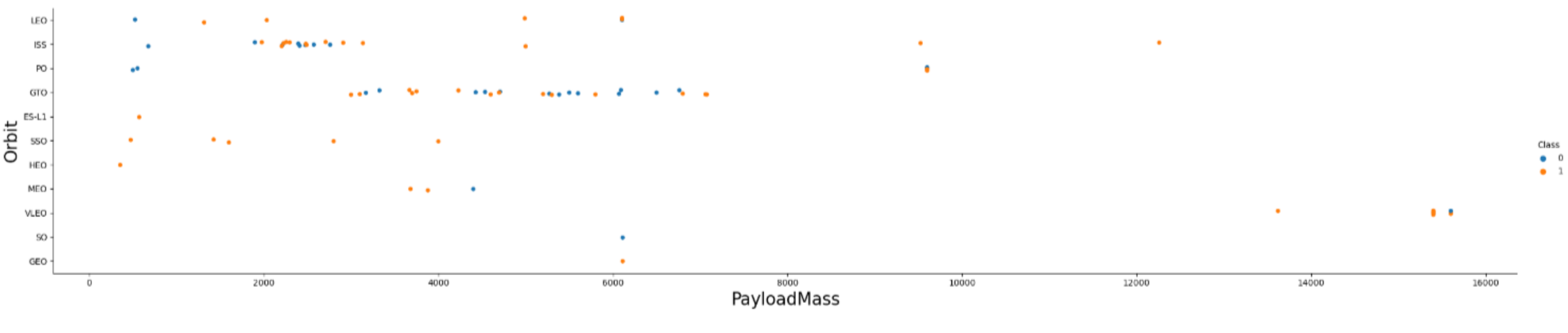
Flight Number vs. Orbit Type

- 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



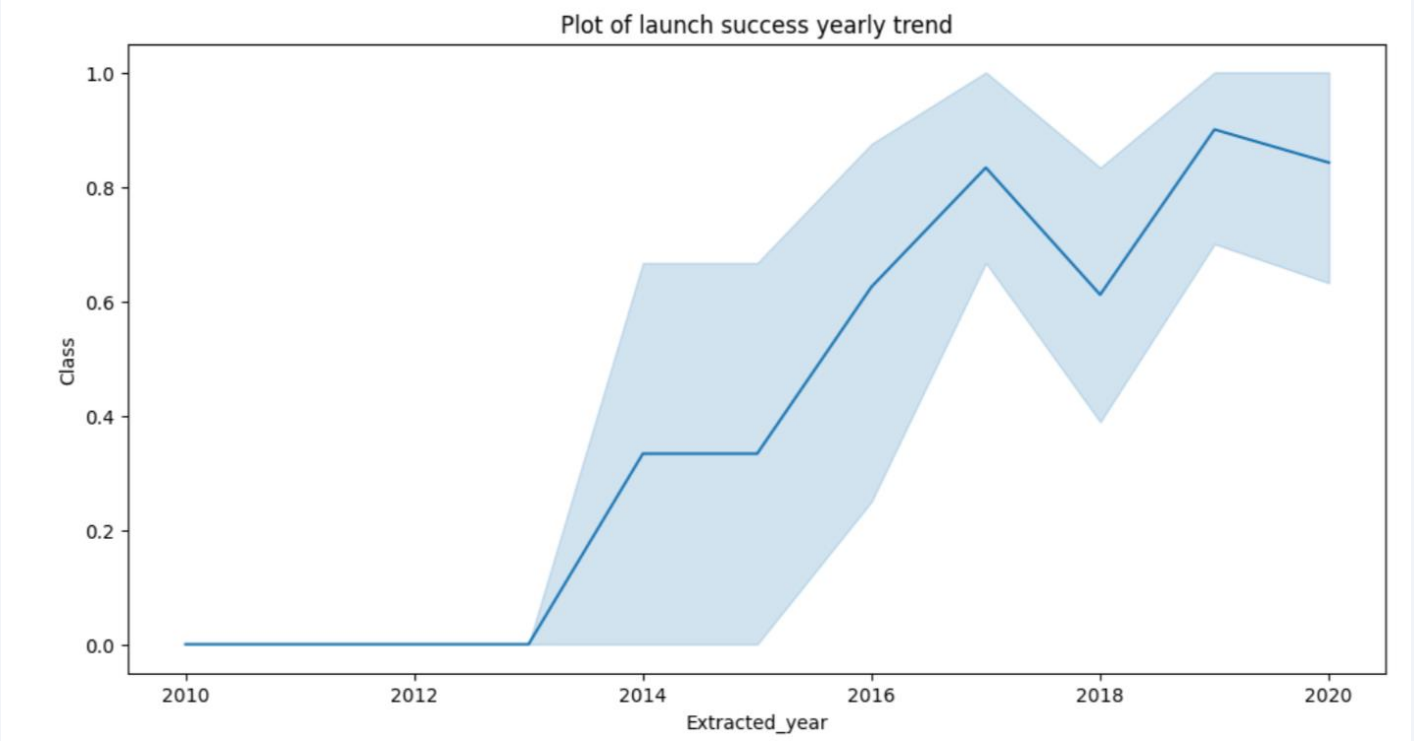
Payload vs. Orbit Type

- 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

Task 1

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

In [15]:
```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

Out[15]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

Task 2

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

In [16]:
```
%%sql
SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Out[16]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
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 att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [21]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer == 'NASA (CRS)';
```

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

Out[21]:

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [26]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Out[26]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [28]:
```
%%sql
SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome == 'Success (ground pad)'
```

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

Out[28]:

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [31]:

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

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

Out[31]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [34]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

Out[34]:

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

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [37]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Out[37]:

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

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [45]:
```
%%sql
SELECT SUBSTR(Date, 4, 2) AS MONTH, Landing_Outcome AS Failure_Landing_Outcomes, Booster_Version, Launch_Site
FROM SPACEXTBL WHERE SUBSTR(Date, 7, 4) = '2015' AND LANDING_OUTCOME LIKE 'Failure%';
```

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

Out[45]:

MONTH	Failure_Landing_Outcomes	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

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.

In [47]:

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

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

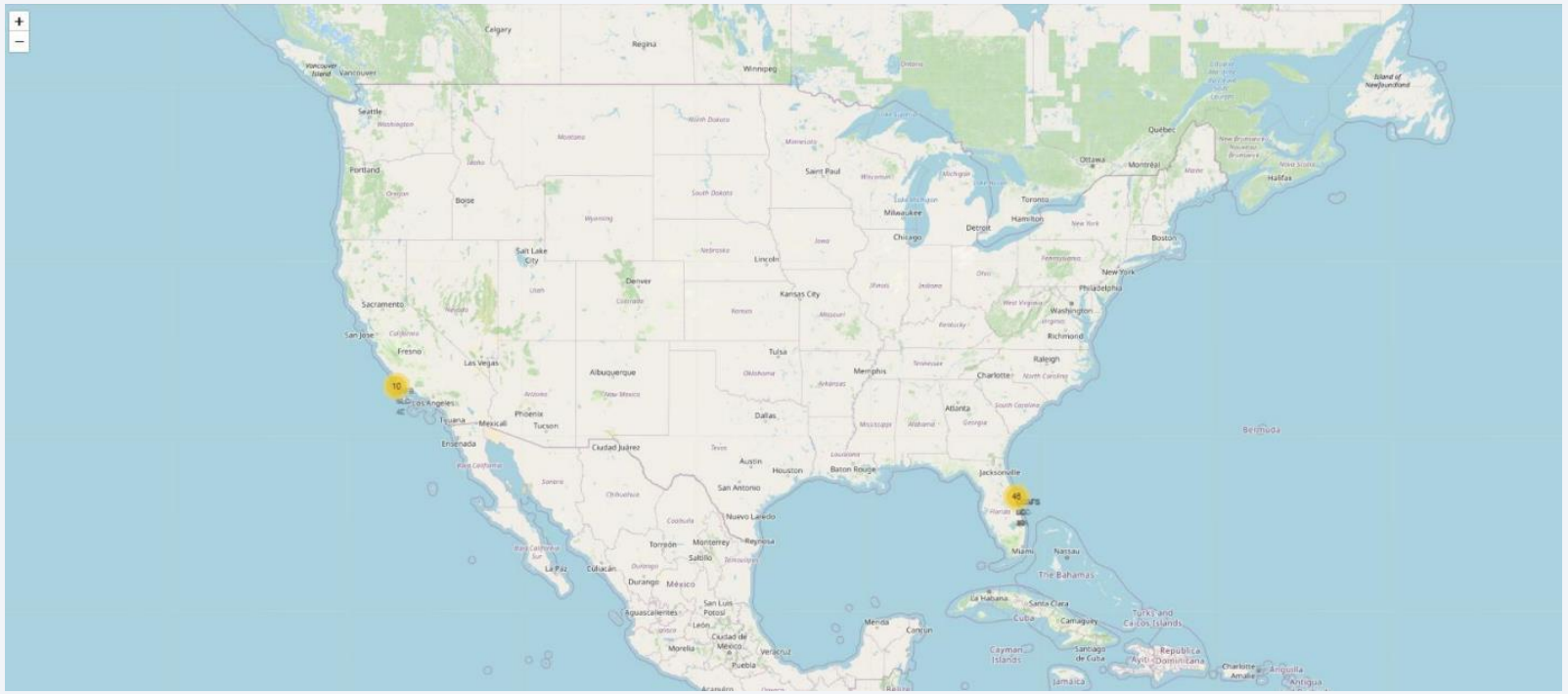
Out[47]:

Landing_Outcome	TOTAL_NUMBER
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

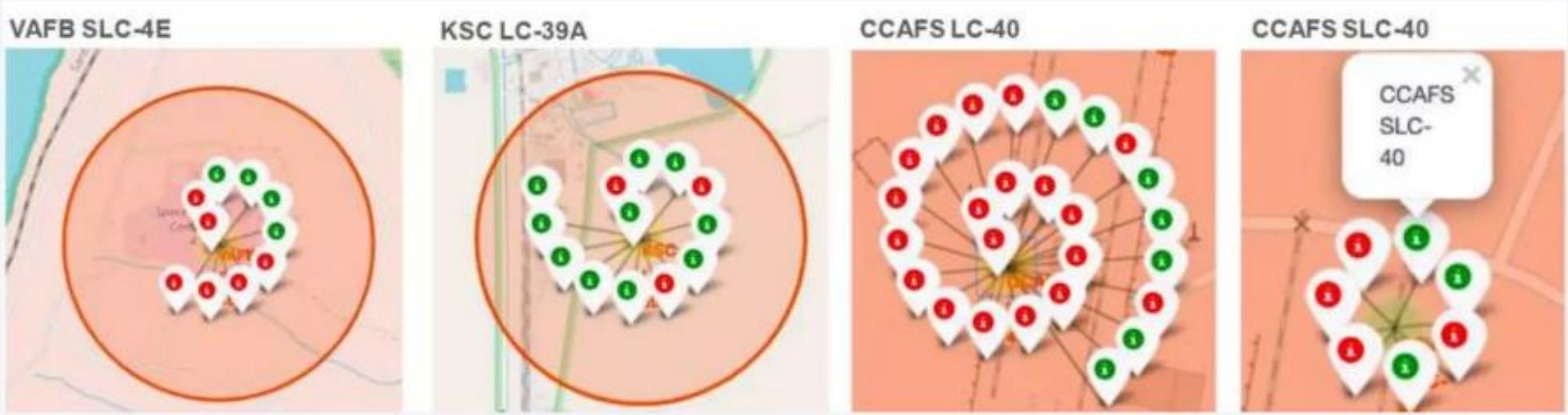
Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps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



Launch Outcomes

Outcomes:

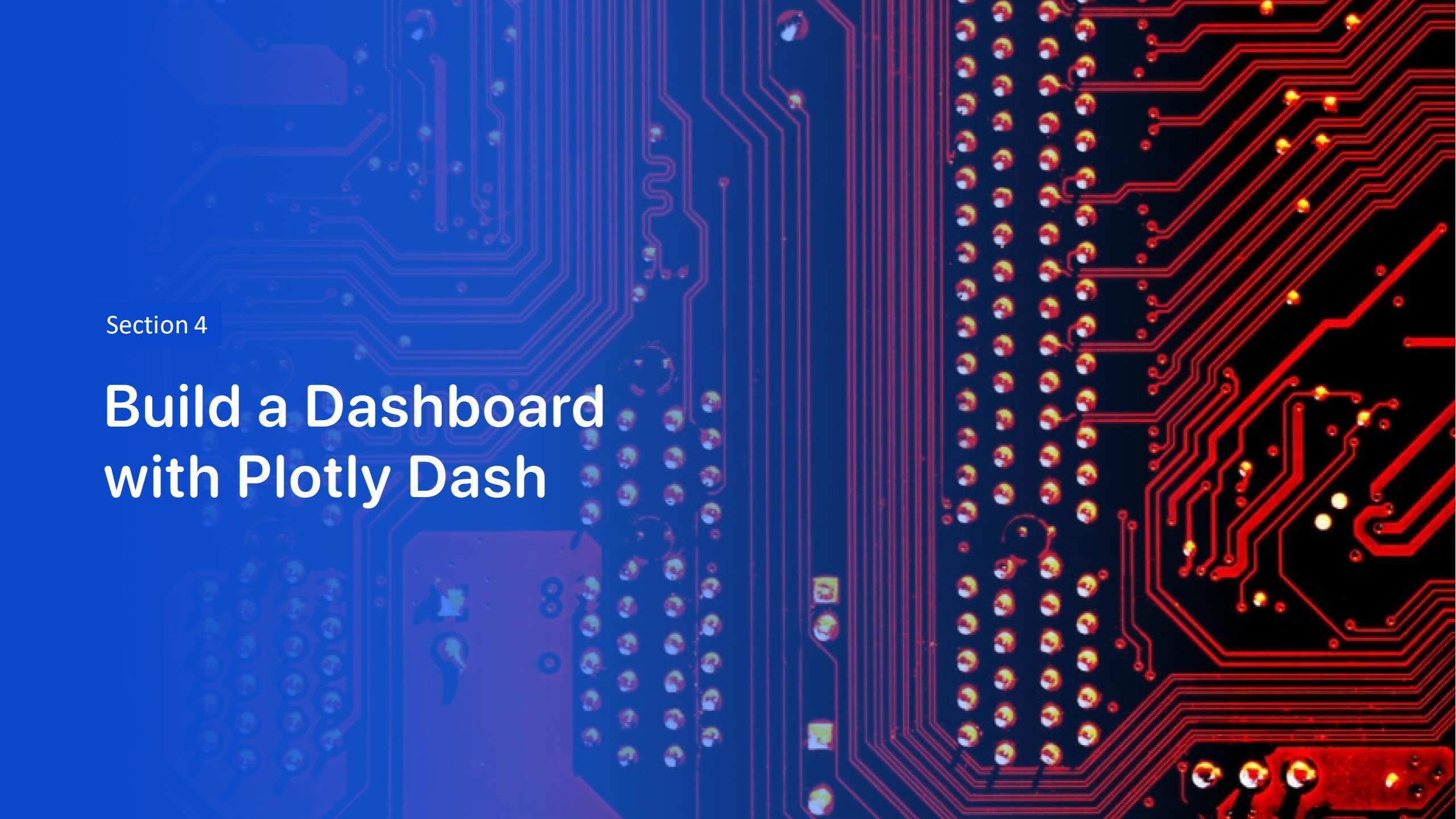
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Distance to Proximities

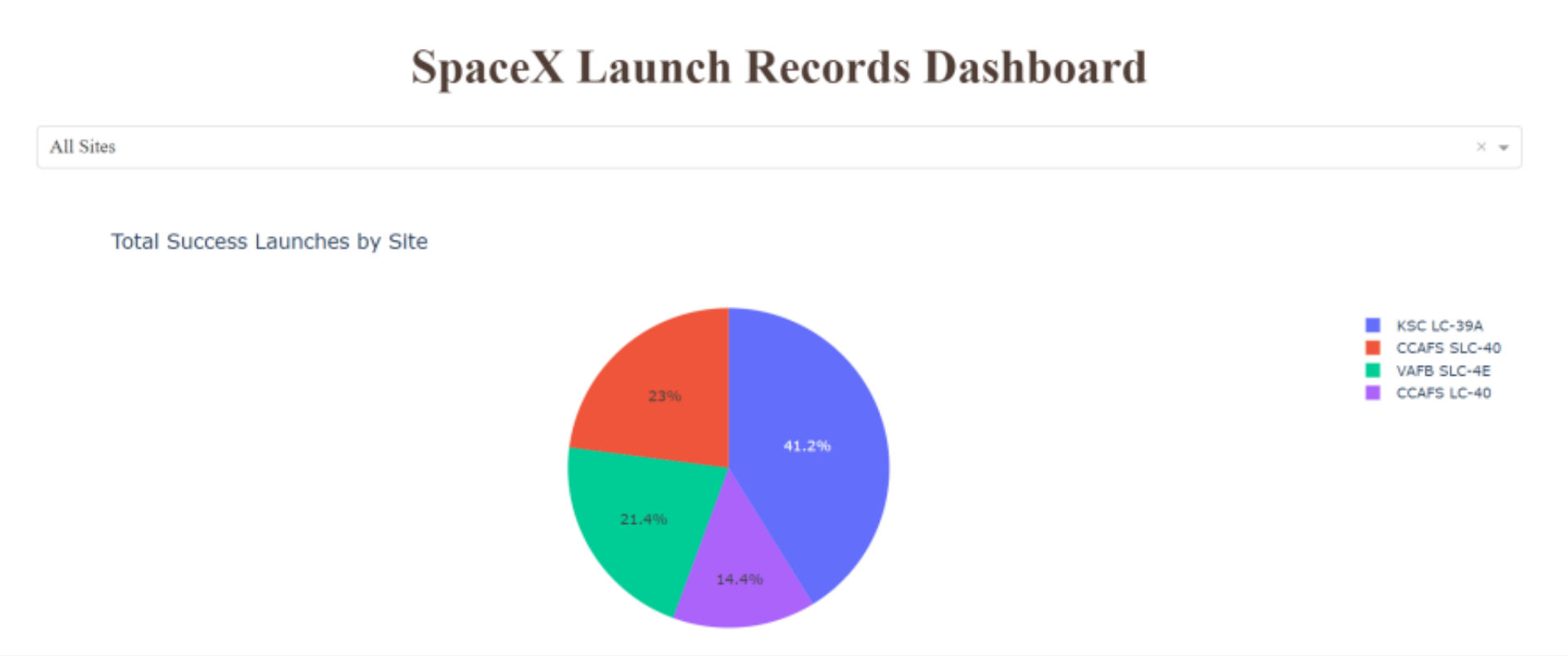
CCAFS SLC-40

- 0.86 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

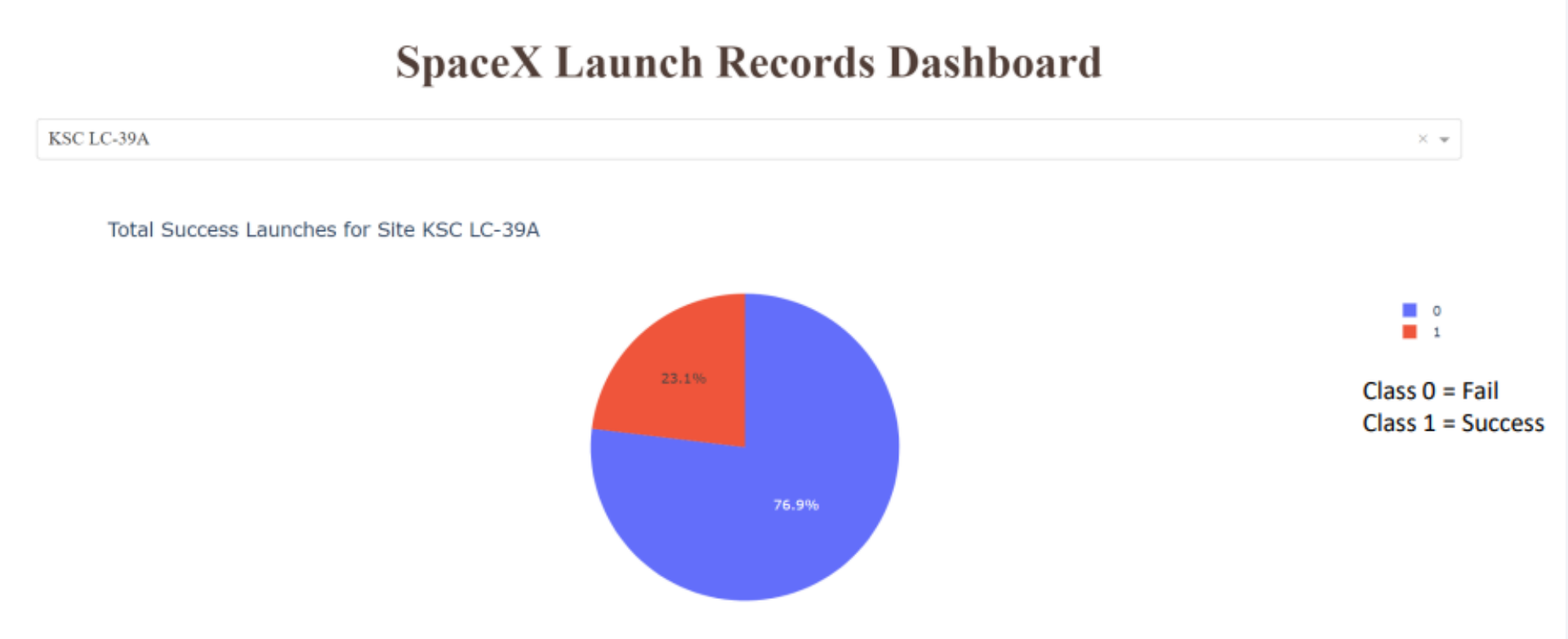
Launch Success by Site

- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



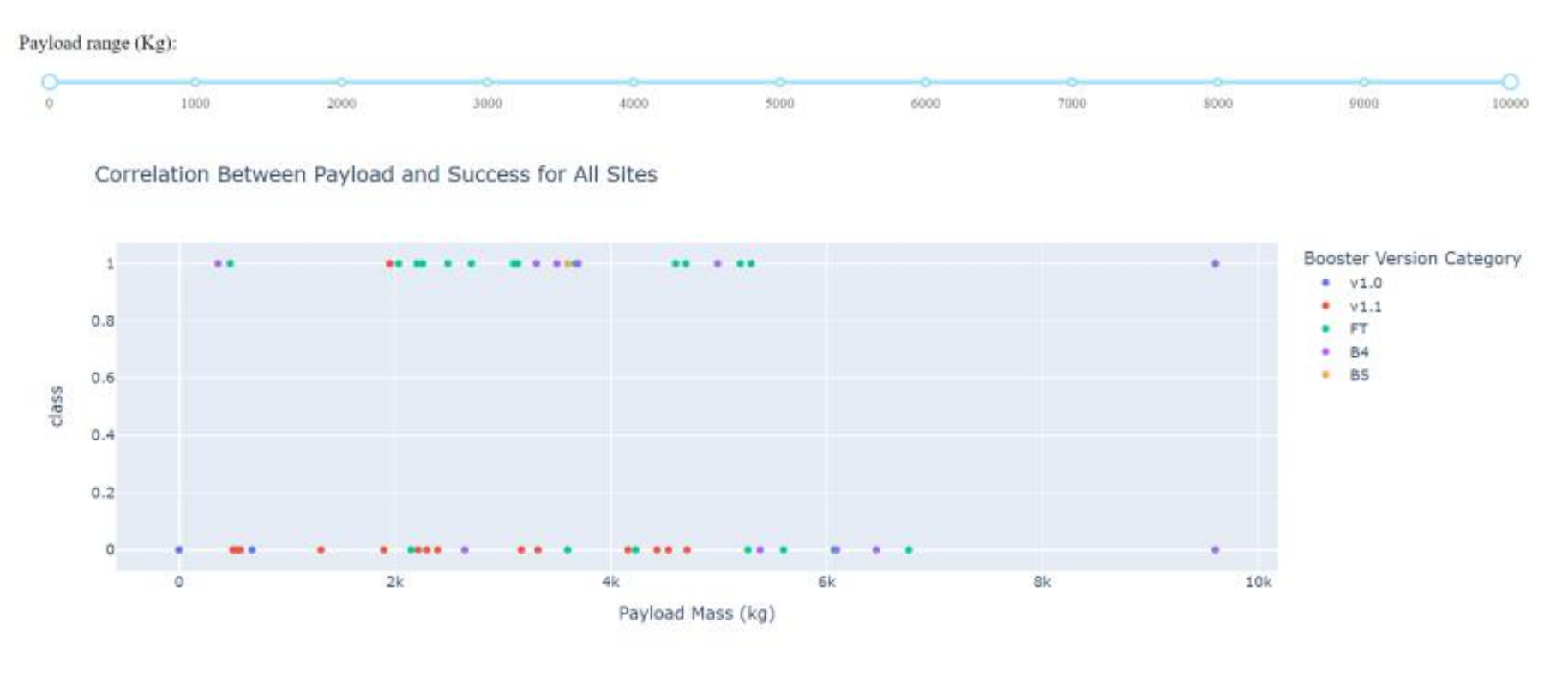
Launch Success (KSC LC-29A)

- 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

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

[46]:

	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

[47]:

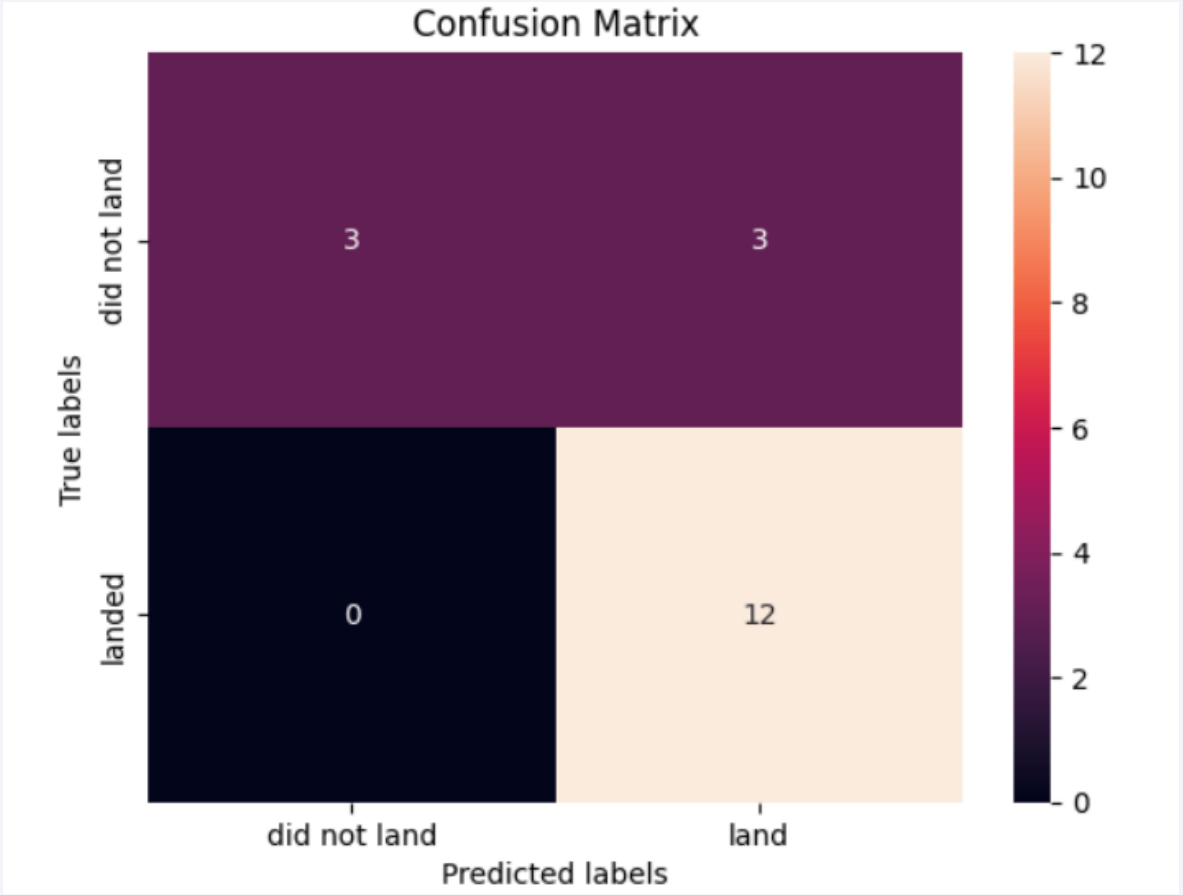
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.875
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'r
andom'}
```

Confusion Matrix

Confusion Matrix Outputs:

- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative



Conclusions

Research Findings:

• Model Performance: The models exhibited comparable performance on the test dataset, with a slight advantage for the decision tree model.

• Equatorial Advantage: Most launch sites are strategically positioned near the equator, leveraging Earth's rotational speed for an inherent boost. This minimizes the need for extra fuel and boosters, leading to cost savings.

• Coastal Proximity: All launch sites are situated close to coastal areas.

• Launch Success Trend: Launch success rates have shown a consistent upward trend over time.

• KSC LC-39A Superiority: Among the launch sites, KSC LC-39A stands out with the highest success rate. Notably, it achieves a perfect 100% success rate for launches with payloads under 5,500 kg.

• Orbit Performance: Orbits such as ES-L1, GEO, HEO, and SSO maintain a flawless 100% success rate.

• Payload Impact: Irrespective of the launch site, a positive correlation exists between higher payload mass (measured in kg) and a heightened success rate.

Things to Consider

• Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set

• Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy

• XGBoost: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models

Appendix

- https://github.com/priyam-op-007/SpaceX_falcon9_landing_prediction/tree/main

Thank you!