IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

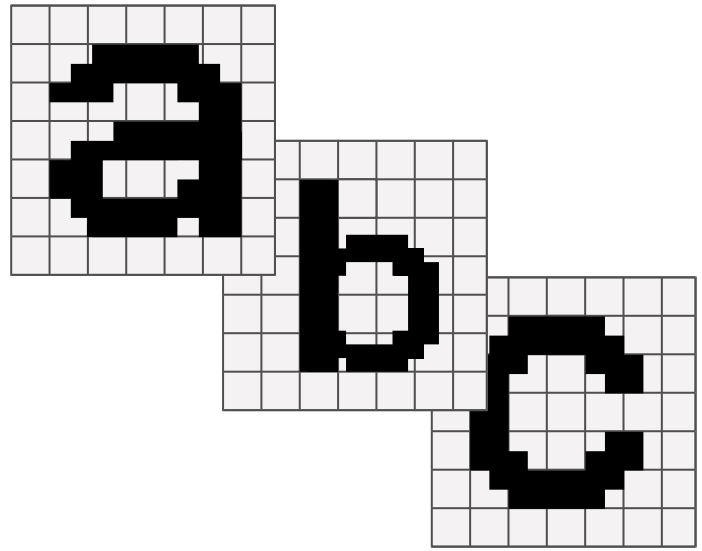
<Luís Martins>
<28-03-2025>

Name Department Course Number Section Time Location DATE

SpaceX Capstone

Outline

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



Executive Summary

Methodology

- Data Collection: Utilized web scraping techniques and the SpaceX REST API.
- Data Processing: Performed data wrangling to define a binary outcome variable (successful/failed launch).
- Analysis: Conducted exploratory data analysis (EDA) using SQL and data visualization tools.
- Dashboard: Developed a dashboard to visualize key metrics, including successful launch sites and payload ranges.
- Predictive Modeling: Built classification models (Logistic Regression, SVM, K-NN, Decision Tree) to predict landing outcomes.

Results

- Trend: Overall launch success rates have shown improvement over time.
- Landing Sites: KSC LC-39A demonstrated the highest success rate among landing locations.
- Success Factors: Payload mass was identified as a major contributor to launch success, especially when considering orbit type and booster version.
- Orbit Success: Specific orbits (ES-L1, GEO, HEO, SSO) achieved a 100% success rate in the dataset.
- Geography: Most launch sites are situated near coastal areas.
- Model Performance: All classification models yielded similar performance on the test data, with the Decision Tree model showing a slight edge.

Introduction: The SpaceX Advantage

Topic

Mission: SpaceX aims to make space travel affordable, driven by reusable rocket technology.

Falcon 9 Reuse: Reusing the first stage significantly reduces launch costs to ~$62 million, compared to ~$165 million+ for expendable rockets.

Cost Prediction: Determining if the first stage will land successfully is key to estimating the true cost of a launch.

Application: This cost insight is valuable for competitors bidding against SpaceX.

Key Research Questions

- Factors Influencing Landing Success:
  - How do payload mass, launch site, flight number, and orbit type impact the success of first-stage landings?
- Geographic Impact:
  - Does the specific location and proximity features of a launch site affect its landing success rate?
- Predictive Modeling:
  - What is the most accurate predictive model for determining if a Falcon 9 first stage will land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - **SpaceX API and Web Scraping from Wikipedia**
- Perform data wrangling
  - **Converting outcomes into data**
  - **‘1’ if the booster successfully landed, otherwise ‘0’.**
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - **Logistic Regression, SVM, Decision Tree and KNN.**

Data Collection

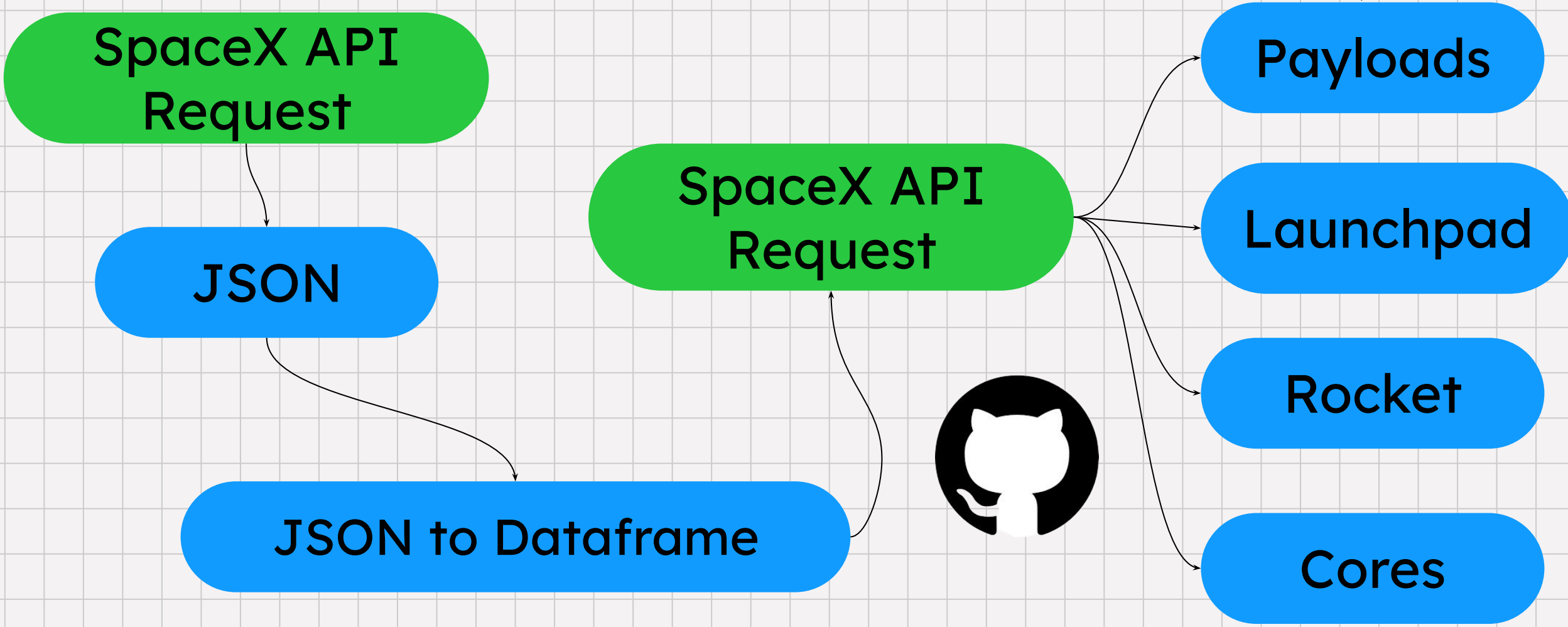
SpaceX REST API

- Source: Directly queried the official SpaceX REST API.
- Data Collected:
  - Rocket types
  - Payload details (mass, destination)
  - Launch specifications
  - Landing specifications
  - Landing outcome (success/failure)[1, 0]

Web Scraping

- Source:
  - Tables containing Falcon 9 launch records on Wikipedia pages.
- Tool:
  - Utilized the Python BeautifulSoup library for extracting data from the web pages.

Data Collection – SpaceX API



Data Collection – SpaceX API

Wikipedia

Request Data

BeautifulSoup

Tables and Features

Dictionary

Dataframe

CSV

Data Wrangling

Tools & Initial Setup

- Libraries Used: Performed analysis primarily using pandas and numpy in Python.
- Dataset Loading: Loaded the collected dataset into a suitable structure (e.g., pandas DataFrame).
- Variable Identification: Identified numerical and categorical columns within the dataset.

Target Variable Creation

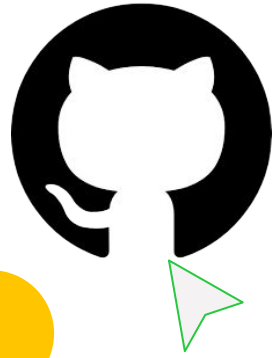
- Objective: Determine training labels for predictive modeling.
- Dependent Variable: Created a binary landing outcome column (named 'Class').
  - **0 = First stage did not land successfully.**
  - **1 = First stage landed successfully.**
- Derivation: Generated the 'Class' column based on an existing 'Outcome' column in the data.

EDA with Data Visualization

Visualization Goals
To deduce potential relationships between variables (useful for feature selection/engineering in machine learning).
To compare performance metrics across different categories (e.g., orbits).
To understand trends over time.

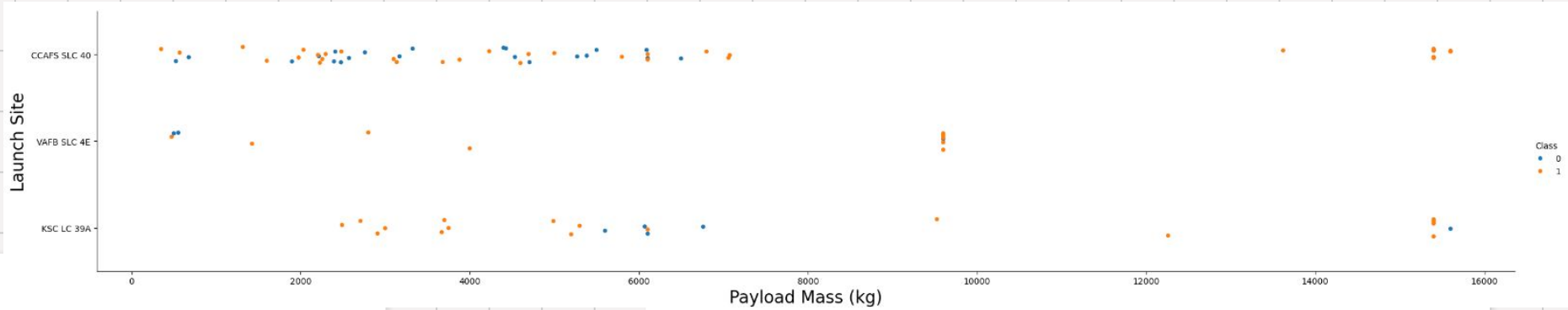
- Scatter Plots: Used to identify potential correlations.
  - "Flight Number" vs. "Launch Site"
  - "Payload Mass (kg)" vs. "Launch Site"
  - "Flight Number" vs. "Orbit type"
  - "Payload Mass (kg)" vs. "Orbit type"
- Bar Chart: Used for comparing categorical success rates.
  - "Success rate" per "Orbit type"
- Line Chart: Used to visualize trends over time.
  - Launch "Success Rate" trend yearly

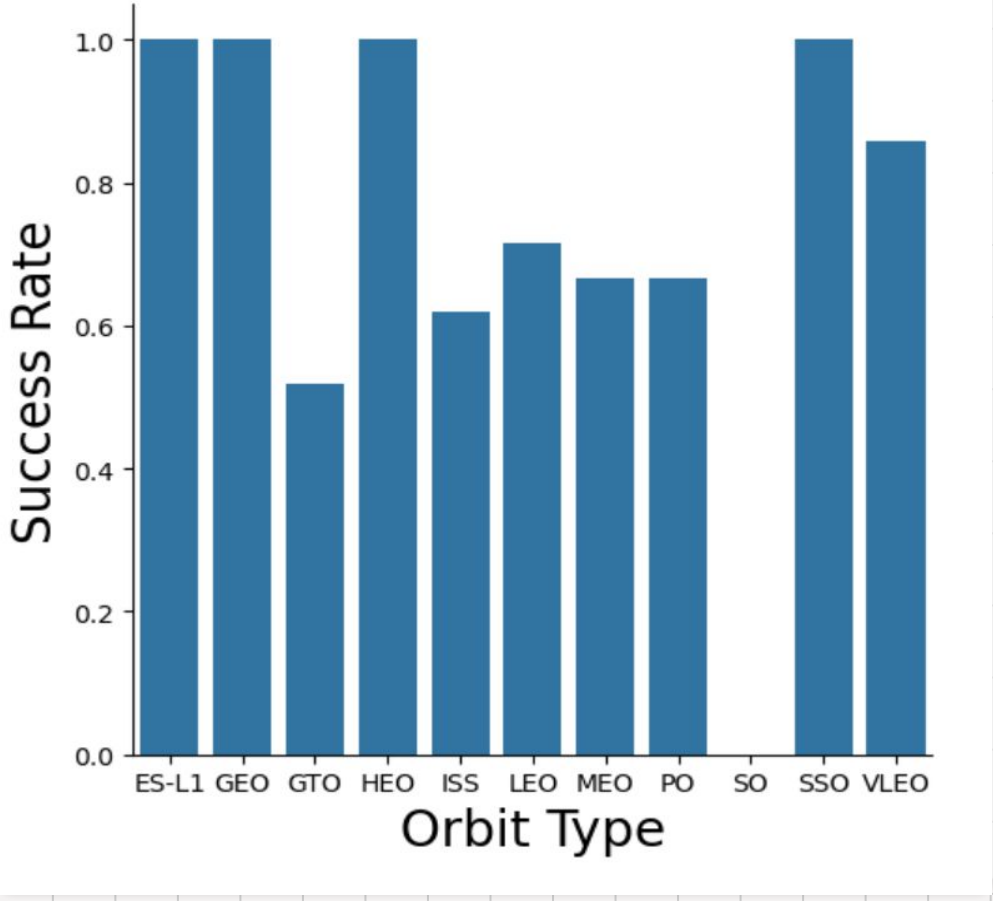
Click in the image for access the github for me info; **Next slide the charts**

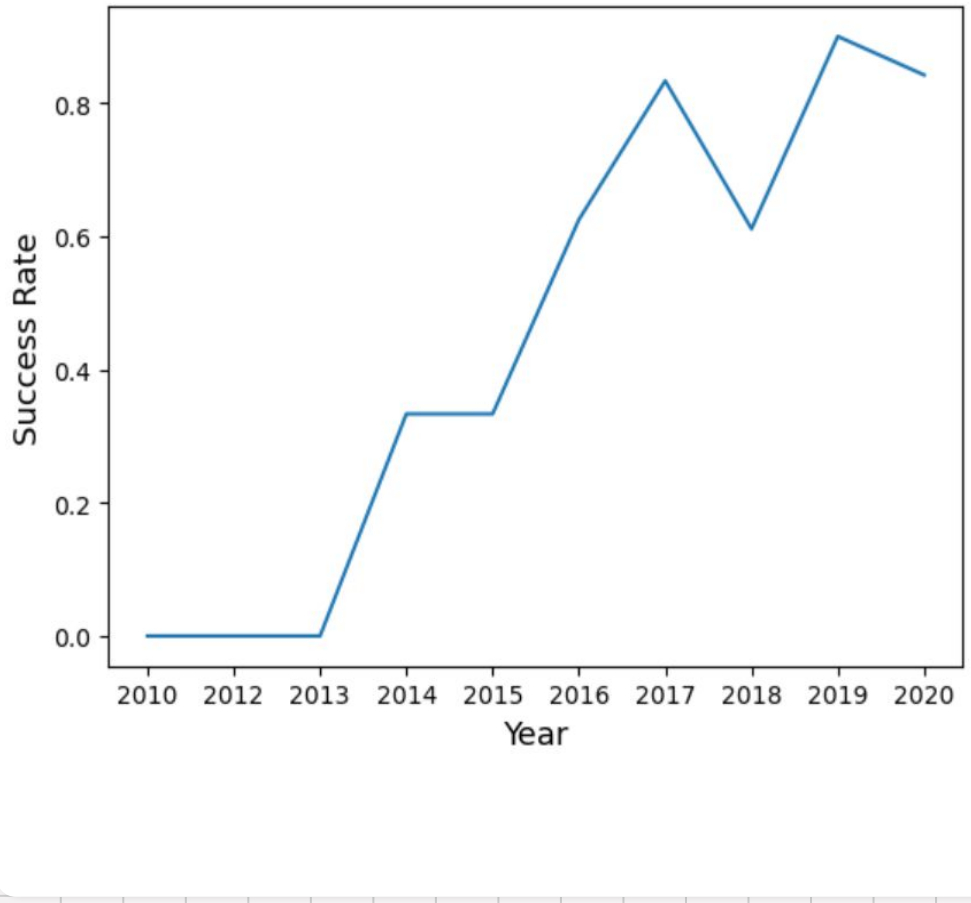


Tips

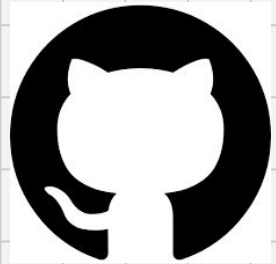
Charts


CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Launch Site
0
2000
4000
6000
8000
10000
12000
14000
16000
Payload Mass (kg)
Class
0
1


Success Rate
1.0
0.8
0.6
0.4
0.2
0.0
ES-L1 GEO GTO HEO ISS LEO MEO PO SO SSO VLEO
Orbit Type


Success Rate
0.8
0.6
0.4
0.2
0.0
2010 2012 2013 2014 2015 2016 2017 2018 2019 2020
Year

EDA with SQL



General SQL Findings

Identified unique launch site names used by SpaceX.

Determined the frequency of different landing outcomes (success/failure, ground/drone ship).

Calculated total and average payload mass for specific booster versions.

Isolated boosters associated with successful drone ship landings within defined payload mass ranges.

Specific SQL Queries

Identified boosters responsible for carrying the maximum payload mass observed.

Displayed unique launch site names.
Filtered and displayed records for specific sites (e.g., launch sites starting with 'CCA').

Calculated aggregate payload mass for specific customers/missions (e.g., NASA CRS).
Computed average payload mass for specific booster versions (e.g., F9 v1.1).

Pinpointed the date of the first successful landing on a ground pad.

Tasks Performed

- Queried for boosters meeting specific criteria (e.g., successful drone ship landing with payload between 4,000 kg and 6,000 kg).
- Counted the total number of successful vs. failed missions overall.
- Identified booster versions associated with carrying the maximum recorded payload.
- Investigated specific failure scenarios (e.g., failed drone ship landings in 2015, showing booster, site, and outcome).

Build an Interactive Map with Folium

Objective

To visually explore the geographic distribution of launch sites and analyze potential correlations between location, proximities, and launch success rates.
To understand if factors like proximity to coastlines, railways, cities, or highways influence landing outcomes.

Rationale: Investigating if proximity to these features correlates with launch success, potentially informing optimal site selection strategies.

- **Map Creation**: Built an interactive map using the Python Folium library.
- **Site Marking**:
  - Placed Circle and Marker objects on the map for each launch site using its geographical coordinates.
  - Employed MarkerCluster objects to handle multiple launch records originating from the exact same coordinates, preventing clutter.
- **Success/Failure Visualization**:
  - Color-coded markers based on the landing outcome ('class' column):
    - Green: Successful landing (class=1)
    - Red: Failed landing (class=0)

Distance Calculation:

Nearest Coastline
Nearest Railway
Nearest City
Nearest Highway

Build a Dashboard with Plotly Dash

Goal: To facilitate quick identification of launch sites and payload ranges that exhibit the highest success rates through interactive exploration.

Objective

To provide an interactive platform for exploring launch success factors, specifically focusing on launch sites, payload mass, and booster versions. To allow users to easily filter and visualize data to gain insights into success patterns.

- **Dashboard Components & Features**
  - **Framework:** Built using the Python Plotly Dash library.
  - **Interactive Controls:**
    - **Launch Site Selection**: A dropdown list enables users to select data for a specific launch site or view data for all sites combined.
    - **Payload Mass Filter**: A range slider allows users to filter the data based on a selected payload mass range.

Visualizations:

Pie Chart: Displays the total successful launch counts;
Scatter Plot: Illustrates the relationship between Payload Mass and Launch Success Rate across different Booster Versions

Predictive Analysis (Classification)

Data Preparation

Target Variable: Created a NumPy array directly from the 'Class' column (landing outcome: 0 or 1).

Feature Scaling: Standardized the feature data using StandardScaler

Data Splitting: Divided the dataset into training and testing sets using train_test_split

Model Training & Hyperparameter Tuning

Hyperparameter Optimization: Utilized GridSearchCV with 10-fold cross-validation (cv=10) to systematically find the optimal hyperparameters for each machine learning model being tested.

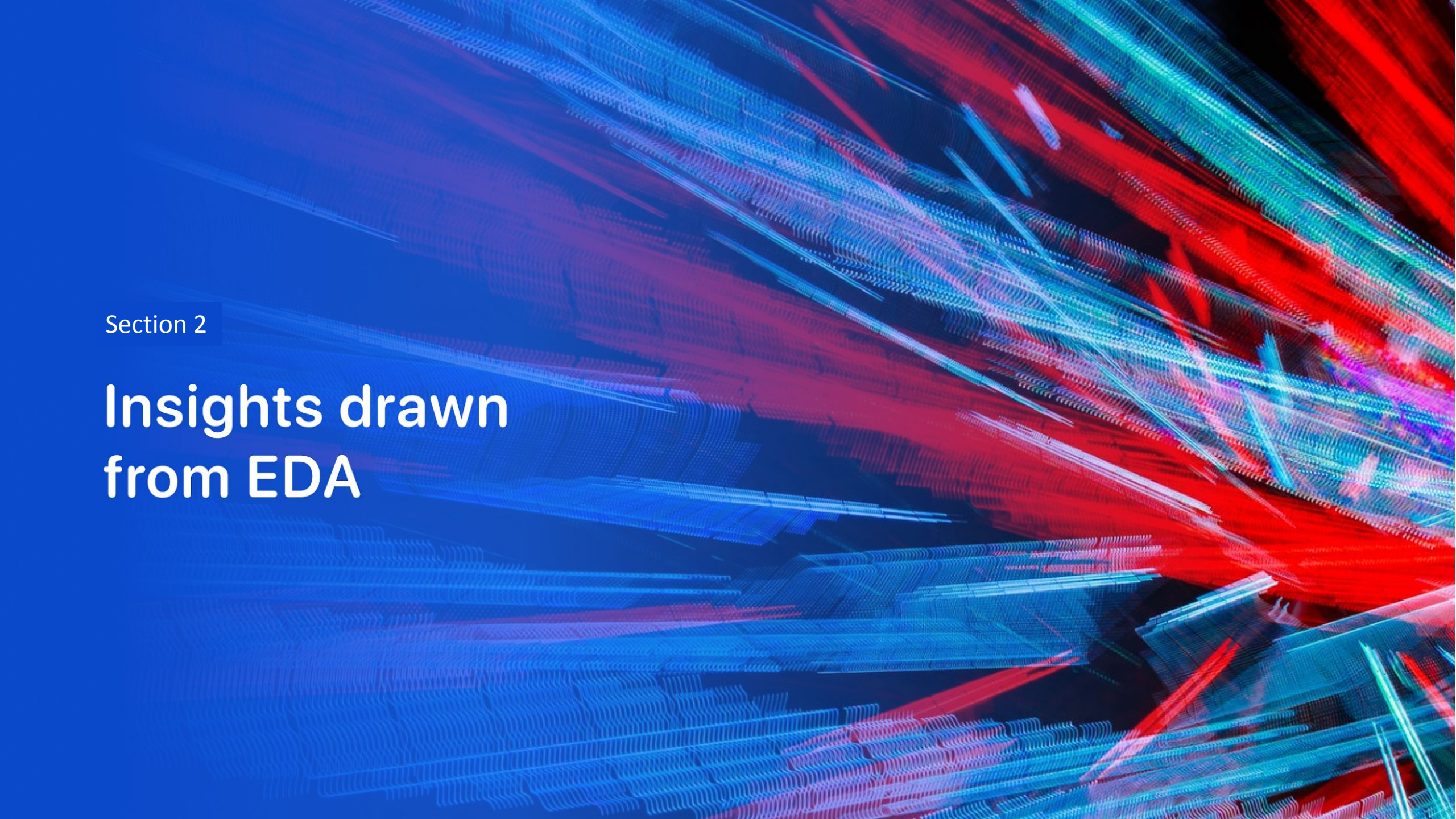
Model Application: Applied GridSearchCV across the different classification algorithms selected for predicting landing success (e.g., Logistic Regression, SVM, K-NN, Decision Tree).

Model Evaluation

- Accuracy Assessment: Calculated the accuracy of each trained model on the held-out test data using the .score() method.
- Confusion Matrix: Generated a confusion matrix for every model
- Metric Comparison: Compared the performance of the different models using multiple evaluation metrics:
  - F1-Score
  - Jaccard Score
  - Accuracy

Section 2

Insights drawn from EDA



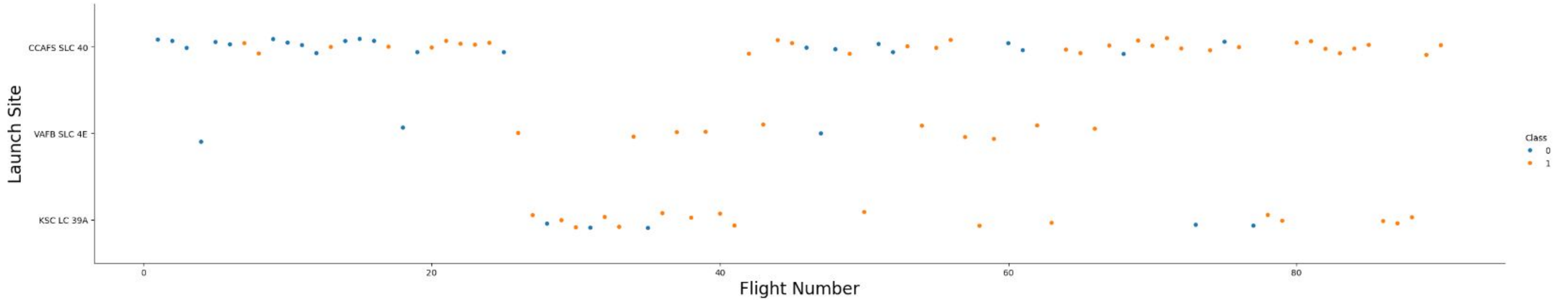
Flight Number vs. Launch Site

Success Rate Over Time

- **Observation:** Earlier launches generally exhibit lower success rates (represented visually, e.g., as blue points for class 0 failures).
- **Trend:** The most recent launches demonstrate significantly higher success rates.
- **Supporting Evidence (Scatter Plot):** Analysis of flight number versus landing outcome shows that as the flight number increases (indicating later launches), the first stage is more likely to land successfully.
    - **Inference:** Experience and technological iteration likely contribute to improved success over time.

Launch Site Distribution

- **Observation:** A significant portion of the analyzed launches originated from one primary site.
- **Dominant Site:** More than half of the total launches occurred from the CCAFS SLC-40 launch site.



Payload vs. Launch Site

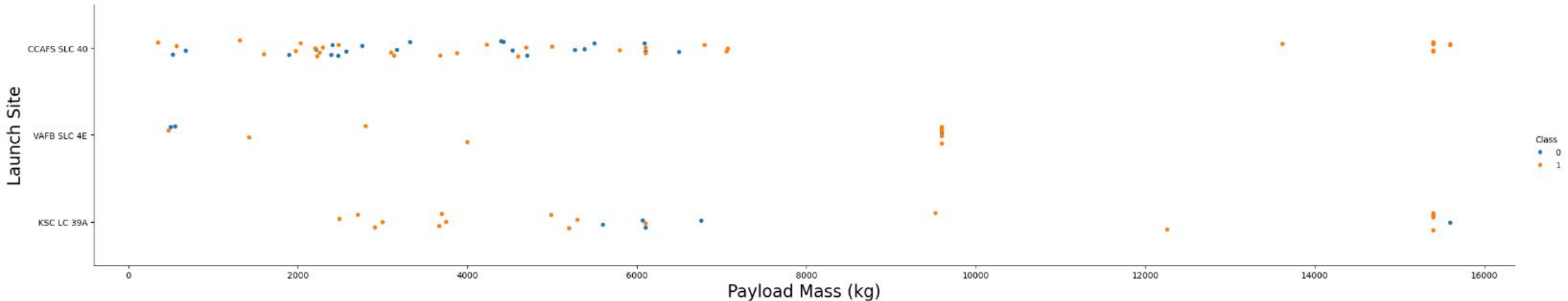
Observations: Payload Mass vs. Success Rate

General Trend (from Scatter Plot)

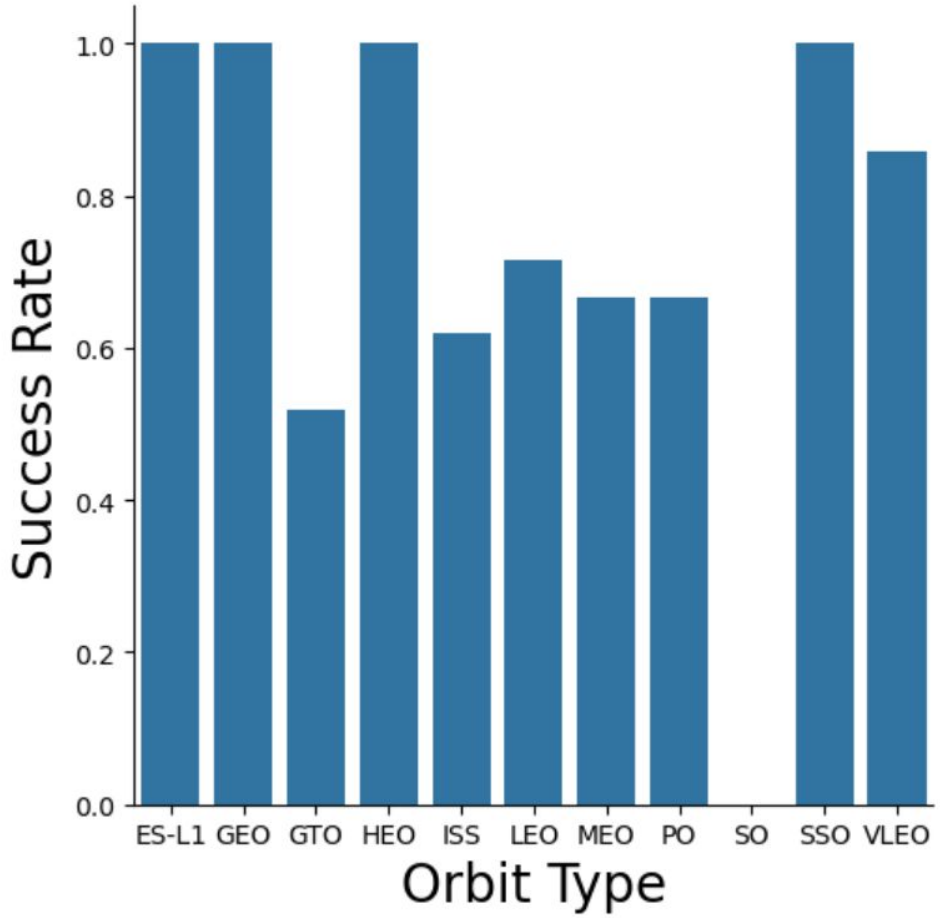
- **Observation:** A positive correlation exists between payload mass (kg) and the first-stage landing success rate.
- **Interpretation:** Launches with higher payload masses tended to have a higher likelihood of successful first-stage landings.

Specific Findings

- **High Payload Success:** The majority of launches carrying payloads greater than 7,000 kg resulted in successful first-stage landings.
- **Site-Specific Observation (CCAFS SLC-40):** For launches originating from CCAFS SLC-40, there appears to be a trend where lower payload masses are associated with a lower probability of successful first-stage landings.



Success Rate vs. Orbit Type



Observations: Orbit Type vs. Success Rate (from Bar Chart)

Highest Success Rate Orbits

- **Observation:** Several orbit types demonstrated a perfect success record in the analyzed data.
- **100% Success Rate:** ES-L1, GEO, HEO, SSO.

Moderate Success Rate Orbits

- **Observation:** A group of common orbits showed success rates above 50%.
- **>50% Success Rate:** LEO, MEO, PO, VLEO, GTO, ISS.

Lowest Success Rate Orbits

- **Observation:** At least one orbit type had no successful first-stage landings in the dataset.
- **0% Success Rate:** SO (Sub-orbital).

Flight Number vs. Orbit Type

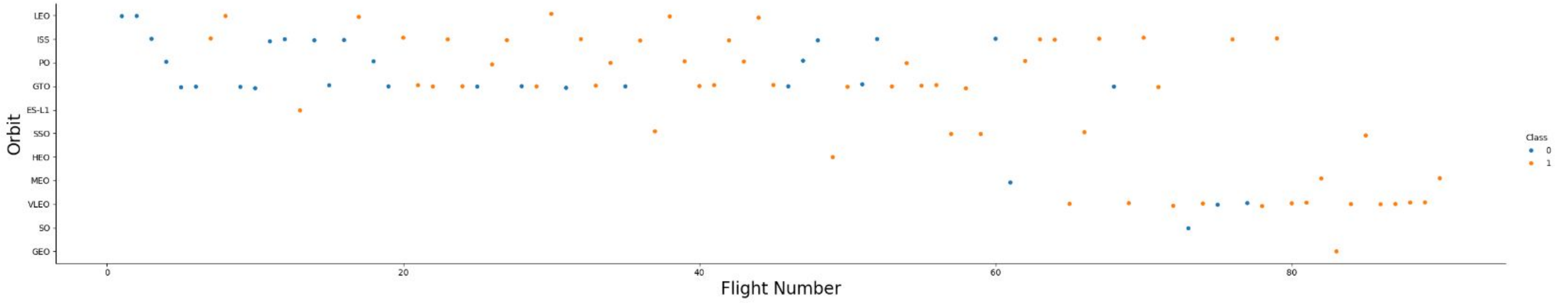
Observations: Flight Number, Orbit & Success

General Trends

- **Reiteration:** Recent launches consistently show better success rates compared to earlier ones.
- **Recent Orbit Focus:** The most recent launches observed in the dataset primarily targeted the VLEO orbit.

Flight Number vs. Success Rate (Orbit Specific)

- **Overall View (Flight # vs. Orbit Type Scatter Plot):** This specific visualization does *not* directly reveal a simple, universal relationship between flight number and successful landings when viewed across all orbit types simultaneously.
- **LEO Orbit Trend:** For launches targeting LEO, a positive trend is visible – successful landings become more likely as the flight number increases.
- **GTO Orbit Trend:** For launches targeting GTO, there appears to be no discernible relationship between the flight number and the likelihood of a successful landing.



Payload vs. Orbit Type

Observations: Payload Mass, Orbit & Success Interaction

Payload Mass vs. Success Varies by Orbit

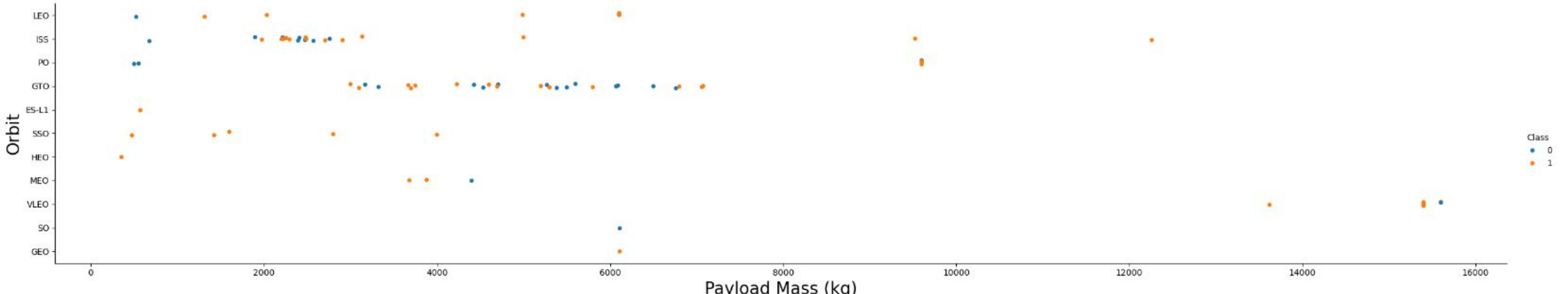
- **Observation:** The relationship between payload mass and the likelihood of a successful first-stage landing is not uniform across all orbit types.

Orbit Types Favoring Higher Payload Mass

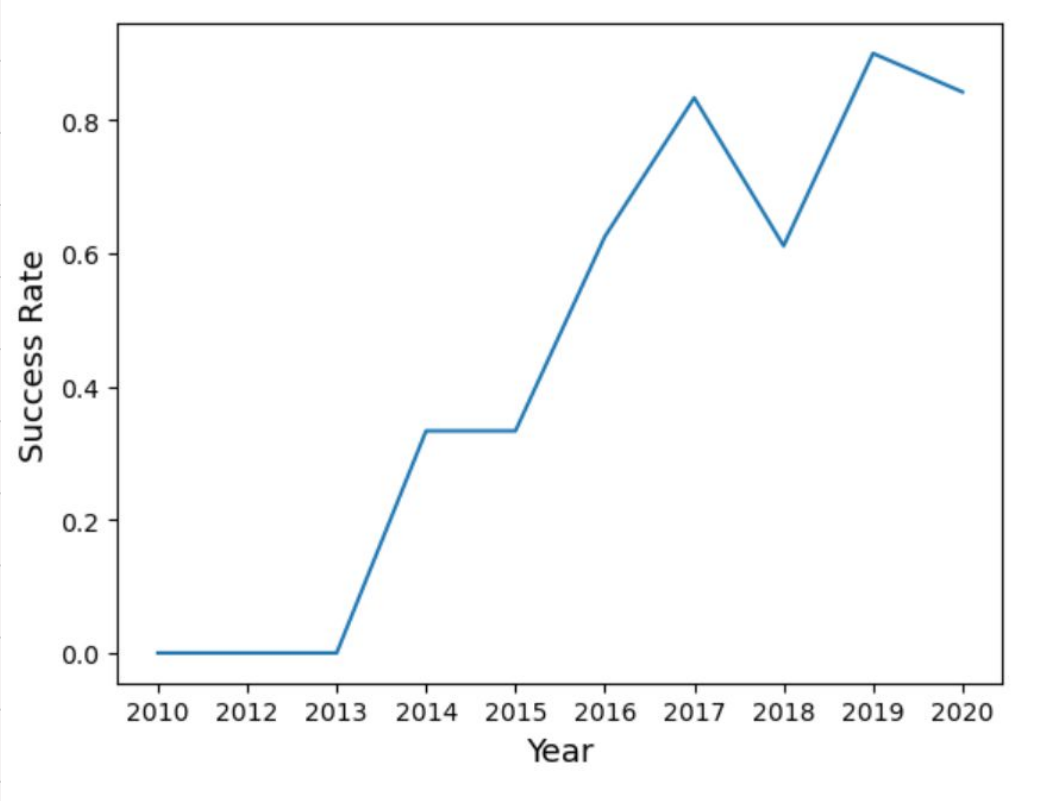
- **Finding:** For certain orbits, successful landings are more likely with *greater* payload masses.
- **Orbits:** LEO, ISS, PO.

Orbit Types Associated with Lower Payload Mass Success

- **Finding:** For other orbits, successful landings were observed primarily with *lower* payload masses.
- **Orbits:** SSO, MEO, HEO.



Launch Success Yearly Trend



Observations: Yearly Launch Success Rate Trend

Overall Trend

- **Observation:** The success rate of first-stage landings has demonstrated a clear upward trend, particularly starting after the year 2013.

Specific Fluctuation

- **Observation:** Within the general upward trend, a slight decrease in the success rate was observed between the years 2017 and 2018.

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

```
 * 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'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE 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

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 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

```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

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

```
 * 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

```
%sql SELECT SUBSTR(Date, 6, 2)AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5) = '2015' AND Landing_Ou
```

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

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

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS COUNT from SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUN
```

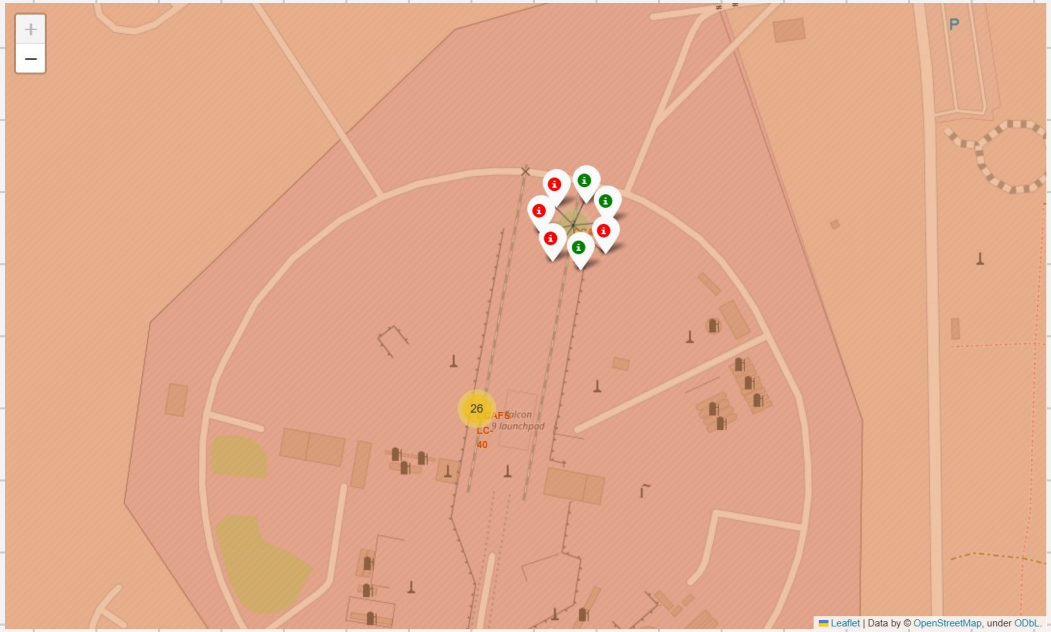
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	COUNT
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


Section 3
Launch Sites
Proximities Analysis

Launch Sites Location

Launch Sites Success Rate



Observations from Interactive Map: Site Success Rates

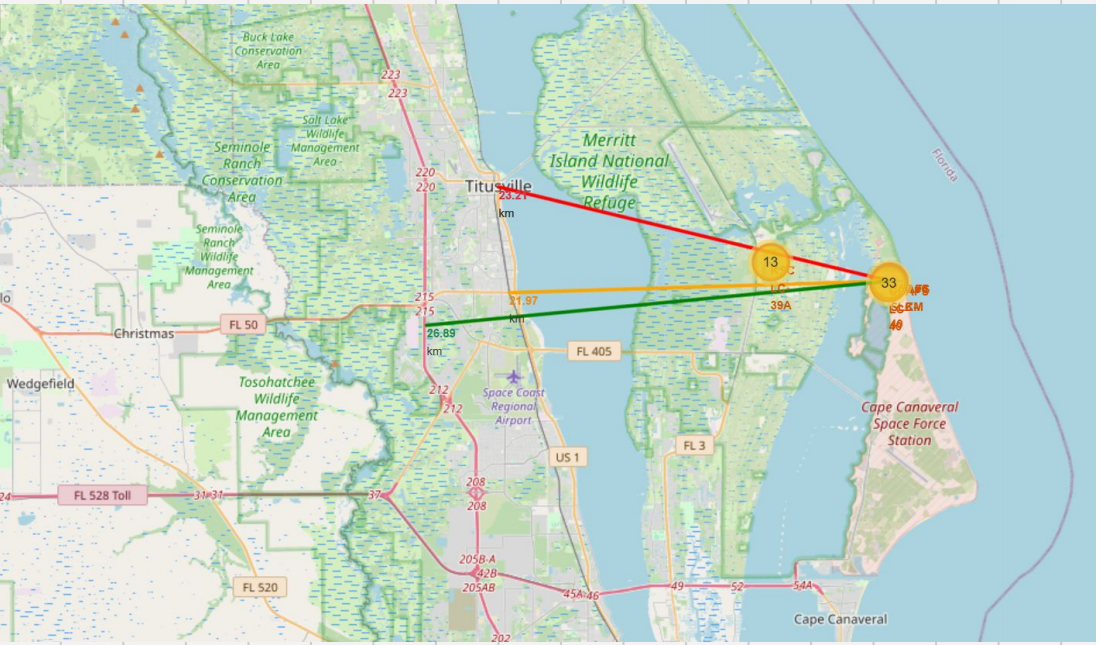
Visual Representation

- **Color Coding:** On the interactive map (created with Folium), launch outcomes are visually marked:
  - **Green Markers:** Indicate successful first-stage landings.
  - **Red Markers:** Indicate unsuccessful first-stage landings.

Launch Site Performance Comparison

- **Highest Success Rate:** Visual inspection of the markers confirms that launch site **KSC LC-39A** exhibits the highest proportion of green markers, indicating the highest success rate.
- **Lowest Success Rate:** Conversely, launch site **CCAFS LC-40** shows a higher proportion of red markers, indicating the lowest success rate among the analyzed sites.

CCAFS SLC-40 and proximities



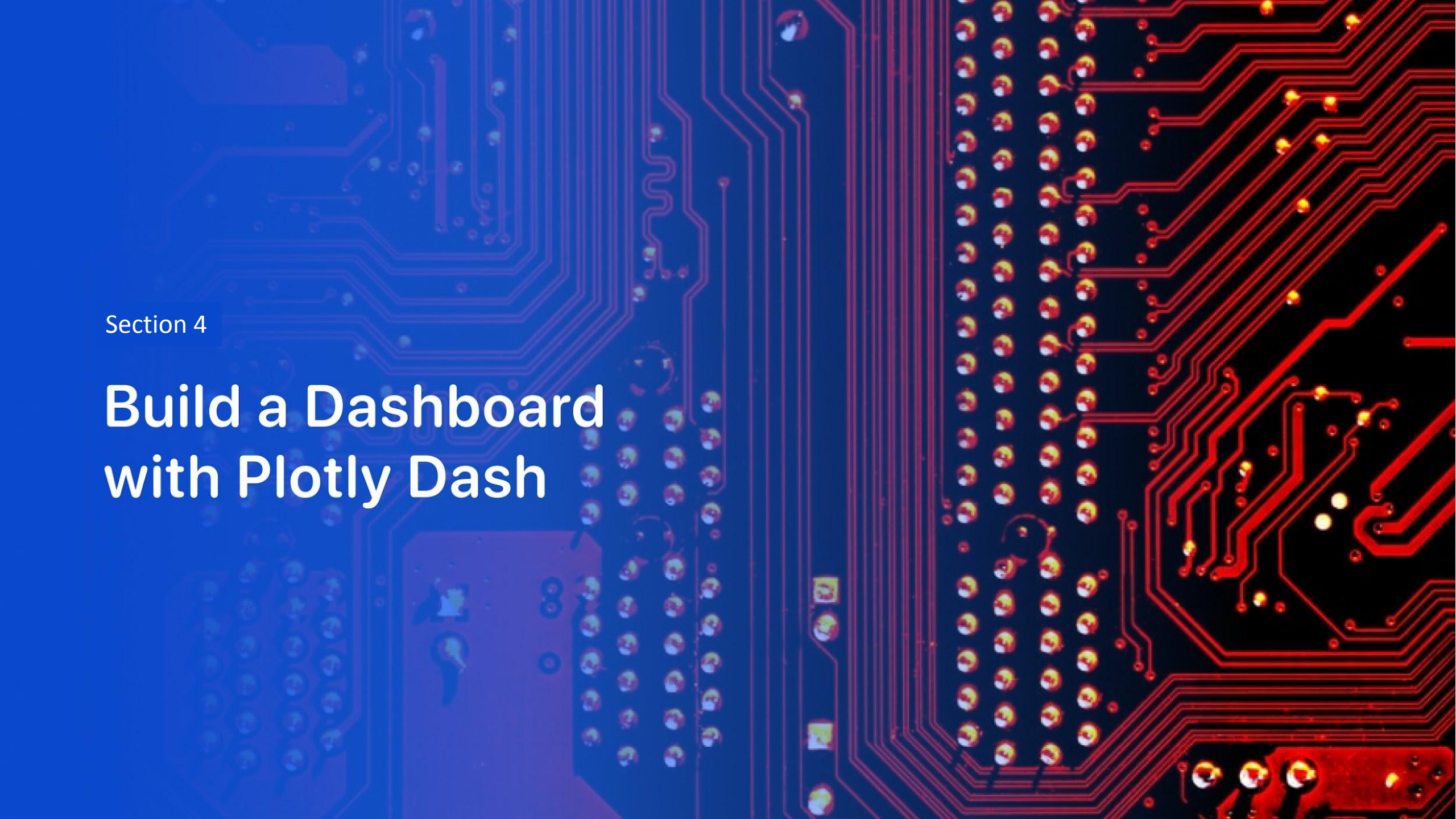
Proximity Analysis Example: CCAFS SLC-40

Methodology

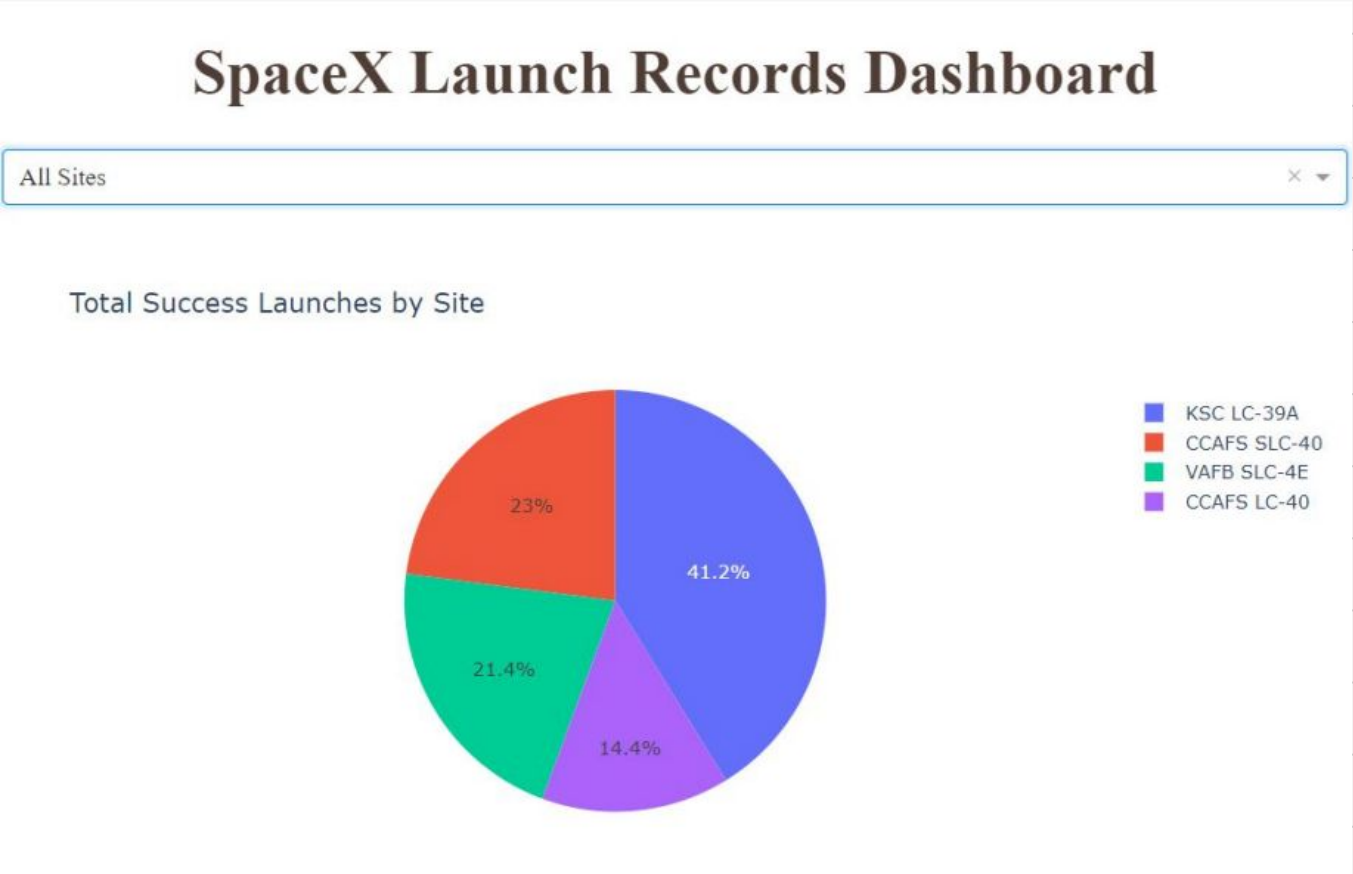
- Distances between the launch site and nearby geographical features were calculated using capabilities within the `Folium` map visualization.

Calculated Distances for CCAFS SLC-40

- **To Nearest Railway:** 21.97 km
- **To Nearest Highway:** 26.89 km
- **To Nearest Coastline:** 0.9 km
- **To Nearest City:** 23.21 km


Section 4
Build a Dashboard with Plotly Dash

Success Launches by Site



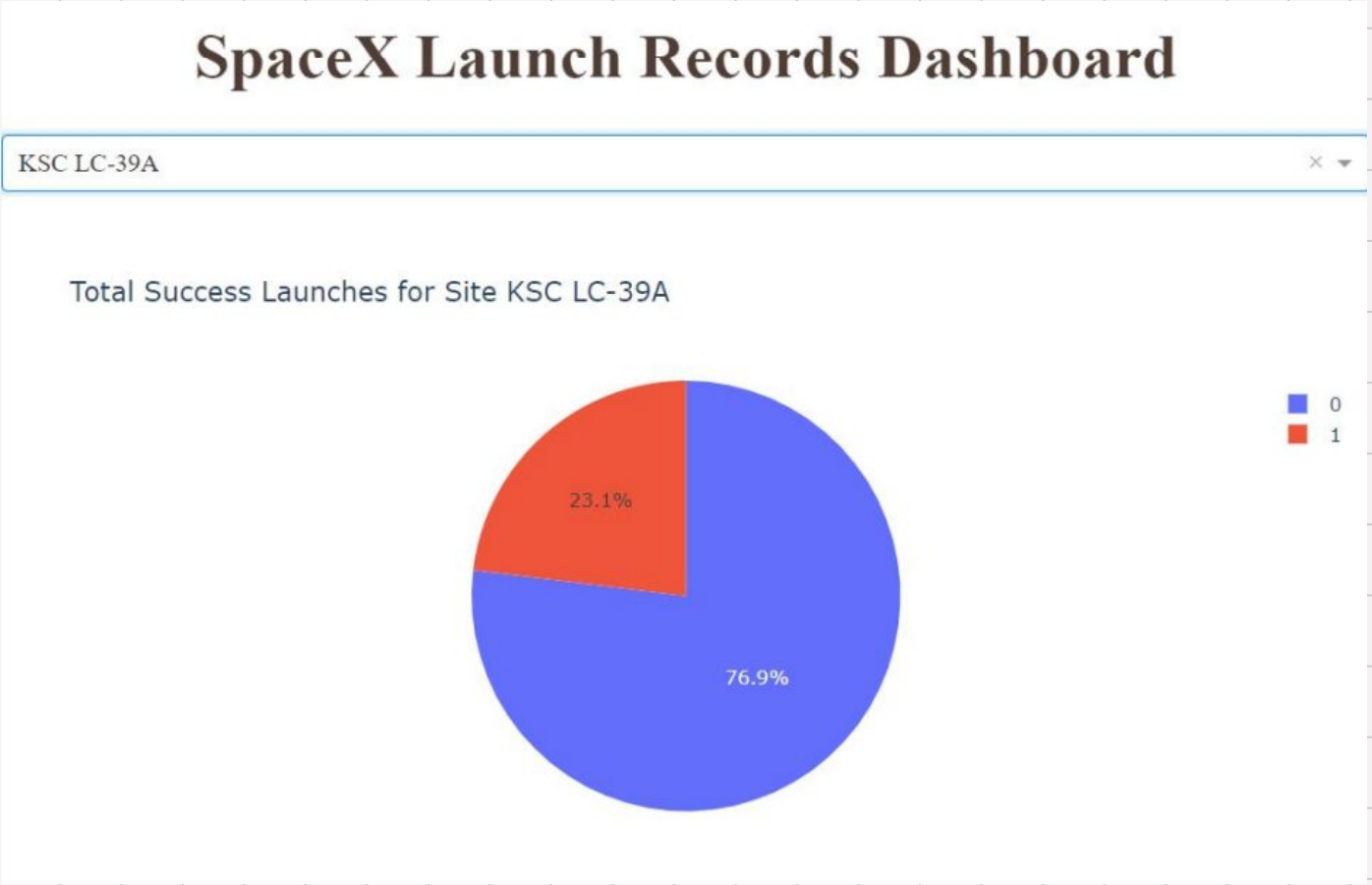
Context & Purpose

- **Source:** This pie chart is taken from the "SpaceX Launch Records Dashboard".
- **Objective:** The chart visualizes the proportion of *total successful first-stage landings* attributed to each launch site within the dataset.

Breakdown of Successful Launches

- **KSC LC-39A (Blue):** Accounts for the largest share, representing **41.2%** of all successful launches.
- **VAFB SLC-4E (Green):** Responsible for **21.4%** of successful launches.
- **CCAFS (Orange & Purple):** Represented by two separate slices:
  - Orange Slice: **23%**
  - Purple Slice: **14.4%**

Launch Site Highest Success Rate

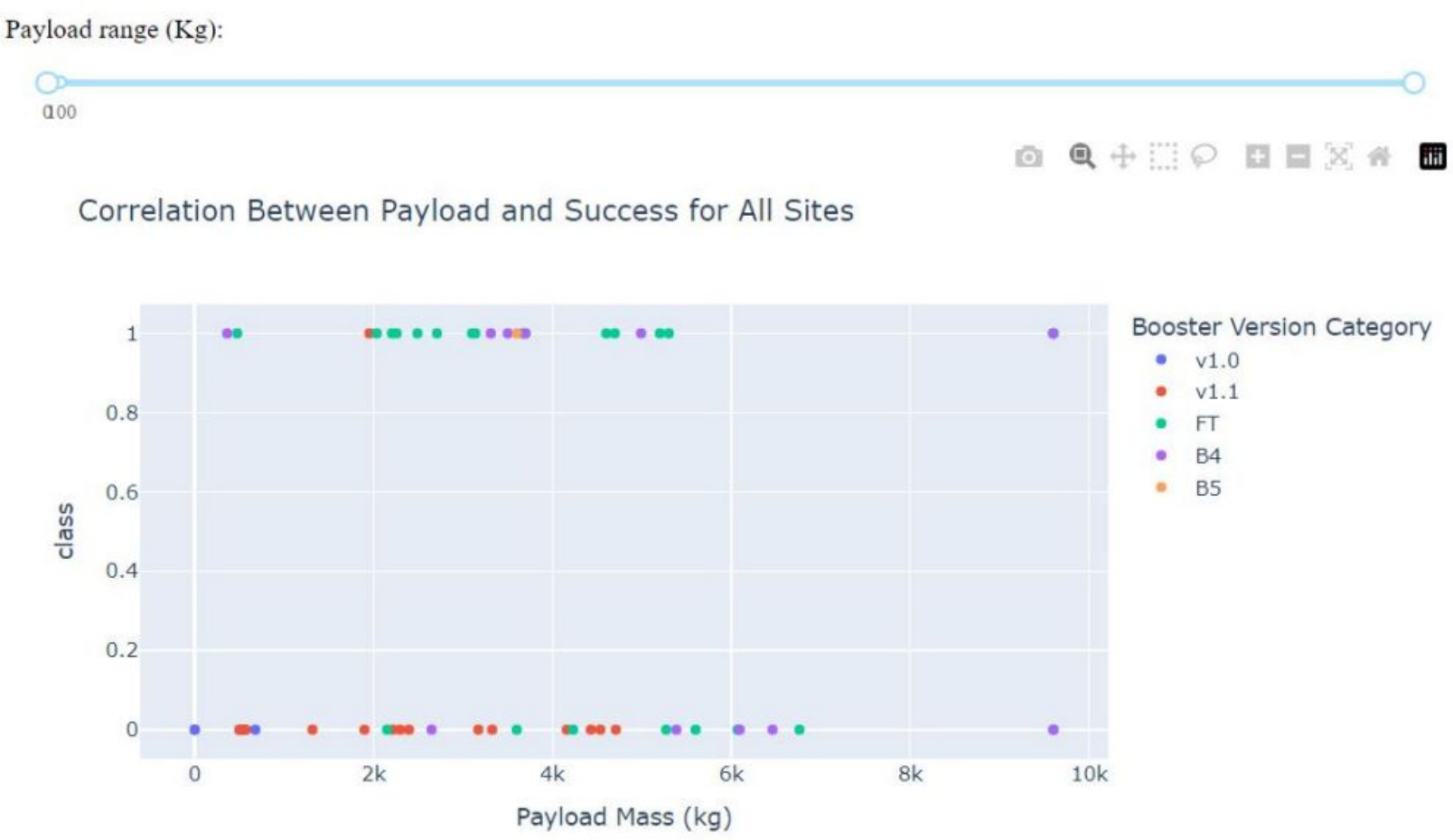


- **Source:** This pie chart is from the "SpaceX Launch Records Dashboard".
- **Objective:** The chart illustrates the breakdown of launch outcomes (successful vs. failed first-stage landings) *exclusively for launches from KSC LC-39A*.

Breakdown of Launch Outcomes for KSC LC-39A

- **Successful Landings (Blue Slice):** Represents **76.9%** of the launches from this site, indicating successful first-stage landings (Class=1).
- **Failed Landings (Red Slice):** Represents **23.1%** of the launches from this site, indicating unsuccessful first-stage landings (Class=0).
- **Note on Legend:** While the chart's legend labels might seem inverted ('0' next to the larger blue slice, '1' next to the red), the accompanying text explicitly states a **76.9% success rate** for KSC LC-39A. Therefore, the blue slice corresponds to successful outcomes (Class=1) and the red slice to failures (Class=0).

Payload Mass vs Success



Key Observation

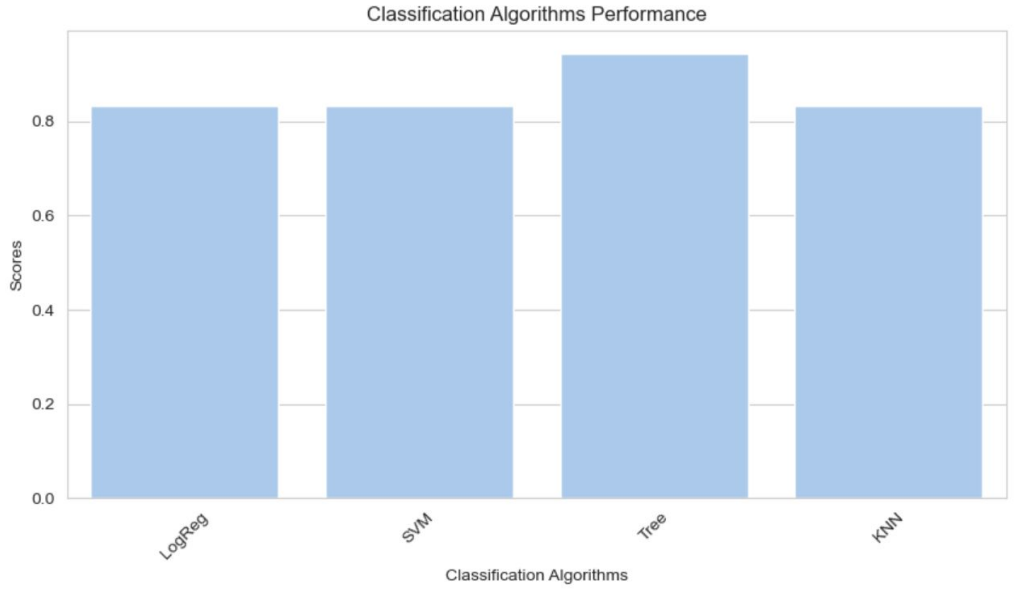
While successful launches occur across various payload masses, there appears to be a noticeable concentration of successful outcomes (dots at y=1) for launches with payload masses roughly between **2,000 kg and 5,500 kg**. This suggests this payload range might have a higher success rate overall within the dataset shown, although performance can vary significantly based on the specific booster version (color) and other factors.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
F1_Score	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333



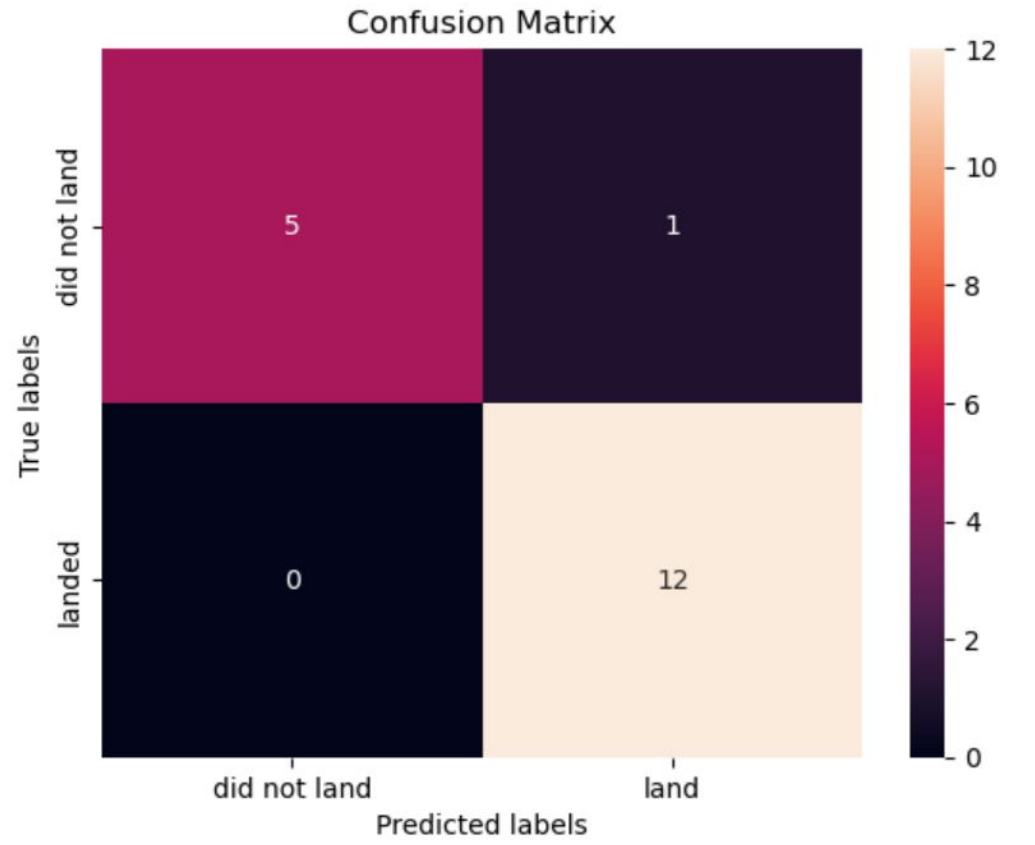
Preferred Model Selection: Decision Tree

Best Performing Model

- **Selection:** Based on a detailed evaluation, the **Decision Tree model** was identified as the best performer overall.
- **Achieved Scores:** This model achieved top scores during the evaluation process:
  - **Accuracy:** 0.94 (or 94 %)
  - **F1-Score:** 0.96 (or 96%)
-

Confusion Matrix

```
yhat_tree = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat_tree)
```



Understanding the Matrix:

- **True Labels (Rows):** Show the actual outcomes.
  - `did not land`: The rocket stage actually failed to land (Actual Negative).
  - `landed`: The rocket stage actually landed successfully (Actual Positive).
- **Predicted Labels (Columns):** Show what the model predicted.
  - `did not land`: The model predicted a failed landing (Predicted Negative).
  - `land`: The model predicted a successful landing (Predicted Positive).

Breaking Down the Numbers:

- **True Negatives (TN): 5**
  - The model correctly predicted "did not land" when the stage actually did not land.
- **False Positives (FP): 1**
  - The model incorrectly predicted "land" when the stage actually did not land (Type I Error).
- **False Negatives (FN): 0**
  - The model incorrectly predicted "did not land" when the stage actually landed successfully (Type II Error). **Crucially, there are zero instances of this error.**
- **True Positives (TP): 12**
  - The model correctly predicted "land" when the stage actually landed successfully.

Conclusions

Trends & Location Factors

- **Success Over Time:** Launch success rates demonstrated a clear increase over time, particularly after 2013.
- **Launch Site Geography:**
  - All analyzed launch sites are situated near coastlines.
  - Proximity to the equator may offer cost-saving advantages for launch sites.

Operational Factors Influencing Success

- **Payload Mass:** A general trend observed suggests that launches with greater payload mass had a higher chance of success *(note: earlier analysis showed this varies significantly by orbit/booster)*.
- **High-Performing Payload Range:** Payloads between 2,000 kg and 5,500 kg showed high success rates, particularly with 'FT' and 'B4' booster versions.
- **Orbit Performance:** Specific orbits achieved perfect success records: ES-L1, GEO, HEO, and SSO demonstrated a 100% success rate in the dataset.

Site & Model Performance

- **Launch Site Comparison:**
  - **KSC LC-39A:** Showed the highest overall success rate, accounted for the most successful launches, and achieved 100% success for payloads between 2k-5.5k kg.
  - **CCAFS LC-40:** Exhibited the lowest success rate among the sites analyzed.
- **Predictive Modeling:**
  - All tested models (KNN, Logistic Regression, SVM, Decision Tree) performed similarly on the final test set.
  - The **Decision Tree model** showed a slight advantage based on cross-validation metrics (`.best_score_`) and could be considered the preferred algorithm.

Thank you!