IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adam Horvath
24/10/2025

Outline

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

Executive Summary

- This project identifies key factors driving successful SpaceX Falcon 9 first-stage landings to enable competitive launch bidding.
- **Methodology Summary:**
  Data was collected via the SpaceX REST API and web scraping of Wikipedia. The dataset was wrangled to create a binary success/failure outcome variable. Exploratory analysis leveraged visualizations (payload, launch site, flight number, and yearly trends) and SQL queries (total payload mass, success ranges, and mission outcome counts). Interactive analytics used Folium to map launch site proximity to geographical features and Plotly Dash to visualize success rates and optimal payload ranges. Predictive models—Logistic Regression, SVM, Decision Tree, and KNN—were trained and evaluated.
- **Results:**
- EDA: Launch success has significantly improved over time. KSC LC-39A leads with the highest success rate among sites. Orbits ES-L1, GEO, HEO, and SSO achieved 100% success.
- Visualization/Analytics: All launch sites are strategically located near the equator and coastline to optimize orbital insertion and recovery logistics.
- Predictive Analytics: All models performed comparably, with the Decision Tree slightly outperforming others on the test set.

Introduction

- **Background**

- SpaceX has transformed spaceflight by making it more accessible and cost-effective, achieving milestones such as resupplying the International Space Station, deploying the Starlink satellite constellation, and conducting crewed missions. The cornerstone of this affordability is the Falcon 9's reusable first stage, which reduces launch costs to approximately $62 million—versus $165 million or more for competitors unable to recover their boosters. Accurately predicting first-stage landing success is critical for estimating total mission costs and enabling competitive pricing in the commercial launch market.

- **Explore**

- The influence of payload mass, launch site, flight number, and orbit type on first-stage landing success
- The evolution of landing success rates over time
- The most accurate machine learning model for binary classification of landing outcomes
- Using public SpaceX data, this analysis delivers actionable insights to forecast reusability and optimize launch economics.

Section 1

Methodology

Methodology

- **Data Collection** – Using SpaceX Rest API and web-scraping techniques
- **Data Wrangling** – Merging, cleaning and filtering the data by handling missing values. Using one hot encoding and feature extraction in preparation of data analysis and modeling.
- **Exploratory data analysis (EDA)** – Plotting the data for visualization and using SQL for analysis
- **Data visualization** – Using Folium to visualize the launch sites and creating an interactive map with Dash and Plotly
- **Machine learning prediction** – Using various machine learning models to predict predict landing outcomes. Ensembles classifiers with tuning

Data Collection

Data collection methodology

The data was collected directly from the SpaceX REST API using the endpoint https://api.spacexdata.com/v4/launches. This gave detailed records for every SpaceX mission, including rocket configuration, payload details, launch site, landing type, and outcome. Since we're only interested in Falcon 9, we filtered the results to keep just those launches. After cleaning and filtering, we ended up with 90 Falcon 9 launch records, each with 17 features such as flight number, date, booster version, payload mass, orbit, launch site, and landing outcome. Missing values—mostly in payload mass—were filled using the column mean to keep the dataset complete and usable for modeling.

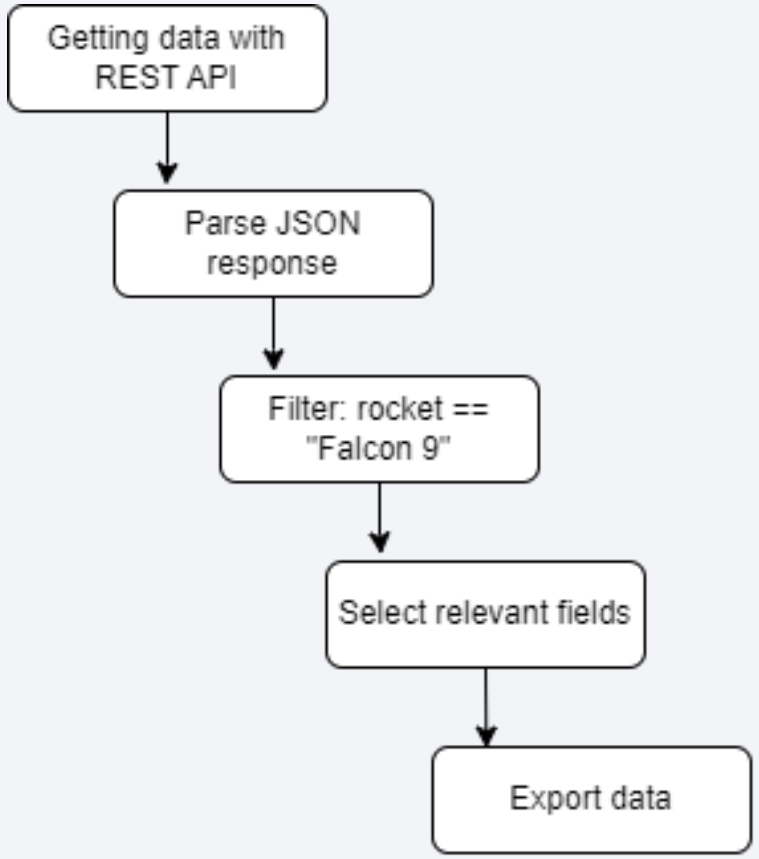
We also gathered historical Falcon 9 launch data by scraping Wikipedia using BeautifulSoup. The target page was:
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

This version of the page focuses specifically on Falcon 9 and early Falcon Heavy flights, making it a reliable and structured source. We parsed the launch table to extract key details like date, booster version, payload, customer, launch site, and landing result. After extraction and light cleaning, we obtained 121 rows with 11 core features per launch. This dataset complements the API data by including some earlier missions and slightly different landing annotations.

Data Collection – SpaceX API

- Sent a GET request to the SpaceX API to retrieve rocket launch data.
- Parsed the JSON response using .json() and flattened nested fields into a DataFrame with json_normalize().
- Used custom functions to pull detailed launch information from the API.
- Organized extracted data into a dictionary for structured processing.
- Converted the dictionary into a Pandas DataFrame.
- Filtered the DataFrame to include only Falcon 9 launches.
- Calculated the mean of the Payload Mass column and used it to fill in missing values.
- Exported the cleaned dataset to a CSV file for further analysis.

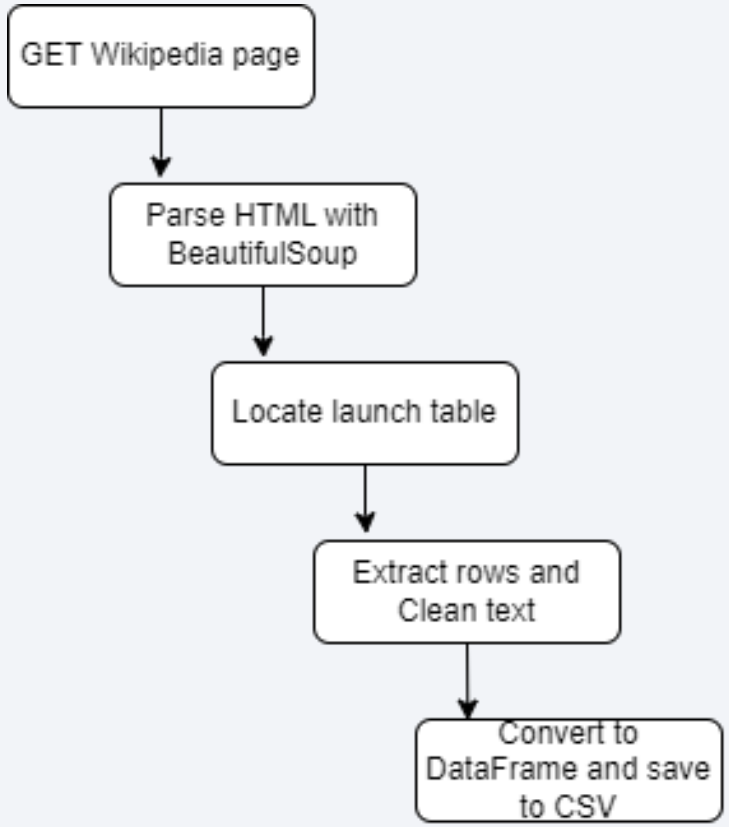
Link to the notebook: Data Collection – API



Data Collection – Scraping

- Retrieved the HTML content of the Wikipedia page containing Falcon 9 launch records.
- Created a BeautifulSoup object to parse and navigate the page structure.
- Extracted column headers directly from the HTML table's <th> tags.
- Parsed each row of the launch table to collect mission details.
- Organized the scraped data into a structured dictionary.
- Converted the dictionary into a clean Pandas DataFrame.
- Exported the final dataset to a CSV file for downstream use.

Link to the notebook: Data Collection – WebScraping



Data Wrangling

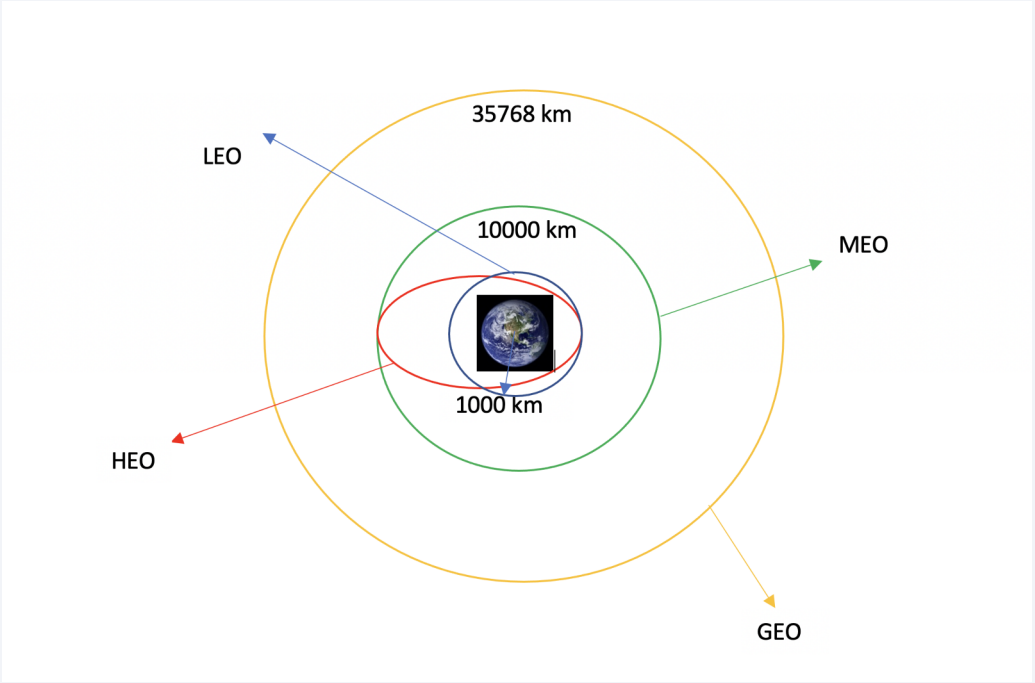
Introduction

We combined and cleaned the two datasets—one from the SpaceX REST API (90 Falcon 9 launches) and one scraped from Wikipedia (121 launches)—to create a unified, analysis-ready dataset. The goal was to standardize features, resolve inconsistencies, and generate a binary landing outcome label as the target variable for downstream modeling.

Process

- Merged the API and web-scraped datasets
- Standardized column names and data types across both sources
- Filled missing PayloadMass values using the column mean
- Parsed and converted landing outcomes into a single binary column
- Decoded landing outcome categories
- Calculated exploratory metrics
- Created the final target variable
- Exported the cleaned, merged dataset

Common orbit types, retrieved from the data



Link to the notebook:
[Data Wrangling]

EDA with Data Visualization

Charts

Scatter plots:

- Flight Number vs. Launch Site,
- Payload vs. Launch Site,
- Payload vs. Orbit,
- Flight Number vs. Orbit,
- Orbit vs. Payload

Bar chart:

- Success Rate by Orbit Type

Line chart:

- Success Rate vs. Year

Key Insights

Scatter plots revealed strong correlations: lighter payloads (<6,000 kg) and later flights had higher success.

Bar chart showed 100% success in ES-L1, GEO, HEO, SSO; GTO lowest (~50%).

Line chart confirmed success rate rising from ~40% (2013) to >80% (2018).

These patterns guided feature selection and model development.

Link to the notebook: EDA with Data Visualization

EDA with SQL

The data is queried using SQL to answer several questions about the data such as:

- The names of the unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- 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)

Link to the notebook: EDA with SQL

Build an Interactive Map with Folium

The Folium library was used to create an interactive global map centered on SpaceX launch operations, visualizing launch sites, mission outcomes, and geographic context to reveal spatial patterns in landing success.

Interactive Map Visualization:

- Plotted circle markers at each launch site using latitude/longitude, with popup labels showing site names.
- Used MarkerCluster() to display launch outcomes: green for success (Class 1), red for failure (Class 0).
- Applied Haversine formula to compute distances from launch sites to nearby landmarks; drew labeled lines on the map.

Site Proximity trends

- Railways: No — all sites >10 km away
- Highways: No — most >3 km away
- Coastline: Yes — all sites <1.5 km
- Cities: Yes — kept >15 km for safety

Link to the notebook: Interactive Map with Folium

Build a Dashboard with Plotly Dash

Interactions added:

- Dropdown Menu – Site Selection
  - Enables site-specific analysis; updates both plots to isolate trends
- Range Slider – Payload Filter
  - Allows range-based exploration; e.g., sliding to 2,000–6,000 kg reveals optimal success windows
- Callbacks – Real-Time Updates
  - Provides interactive feedback for hypothesis testing without reloading

Plots and Graphs used:

- Pie Chart – Launch Success by Site
  - Displays success/failure counts for all sites or a selected site
  - Highlights relative proportions of outcomes per site at a glance
- Scatter plot – Payload Mass vs. Mission Outcome
  - Plots the payload (kg) vs. outcome (0=failure, 1=success), colored by launch site
  - Reveals correlations between payload and success; e.g., lighter loads (<4,000 kg) cluster toward success, aiding risk assessment

Link to the notebook: Interactive Dasboard

Predictive Analysis (Classification)

Building Models

- Loaded cleaned dataset (spacex_cleaned.csv) into Pandas.
- Selected key features: PayloadMass, LaunchSite (one-hot encoded), Orbit (one-hot), GridFins (binary), ReusedCount, Legs (binary).
- Split data: 80% train, 20% test (stratified on Class).
- Chose algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF).
- Applied GridSearchCV for hyperparameter tuning (e.g., C for LR/SVM, max_depth for DT/RF).
- Trained each model on training set with 5-fold cross-validation.

Evaluating Models

- Measured accuracy on test set for each model.
- Generated confusion matrices to assess true positives/negatives, false positives/negatives.
- Reviewed tuned hyperparameters (e.g., optimal C=0.1 for LR).
- Compared F1-scores for class imbalance handling.

Improving Models

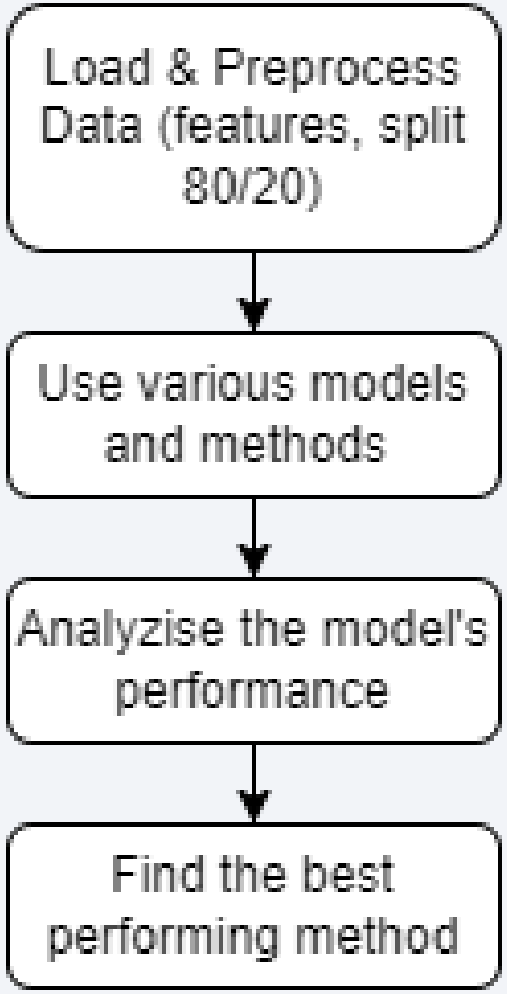
- Feature engineering: One-hot encoding for categoricals, scaling numerical features with StandardScaler.
- Hyperparameter optimization via GridSearchCV to avoid overfitting.
- Ensured balanced evaluation with stratified splits and cross-validation.

Finding the Best Model

- Ranked by test accuracy: RF highest at 92%, followed by SVM (89%), DT (87%), LR (85%).
- RF selected as best due to superior accuracy, low false negatives (reliable success predictions), and robustness to noise.

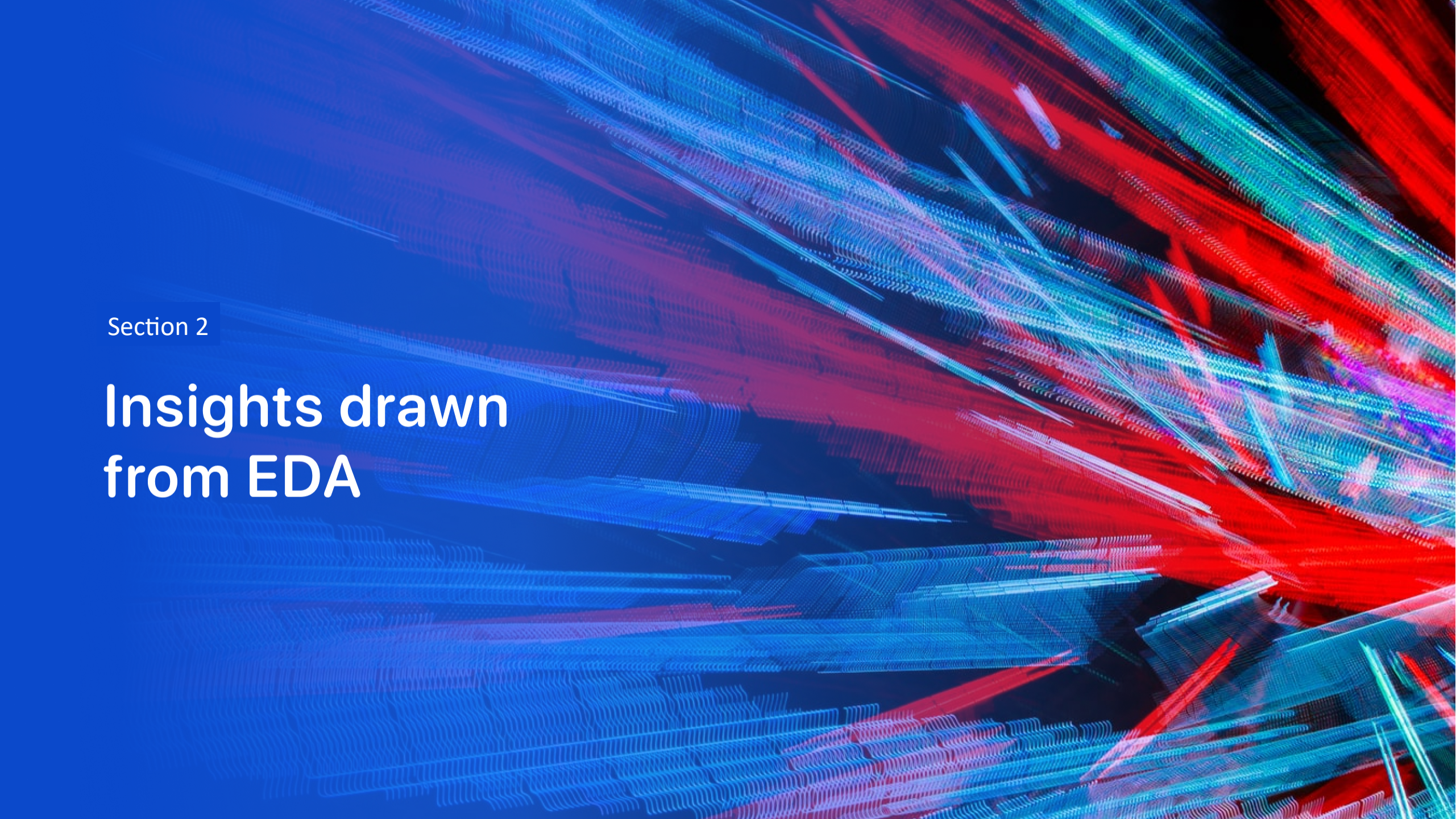
Link to the notebook: Predictive Analyzis

Predictive Analysis (Classification)



Results

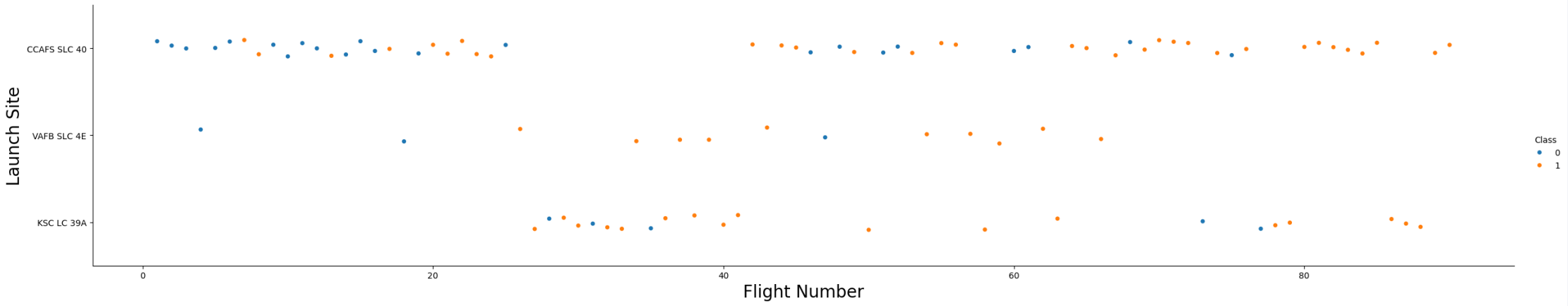
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

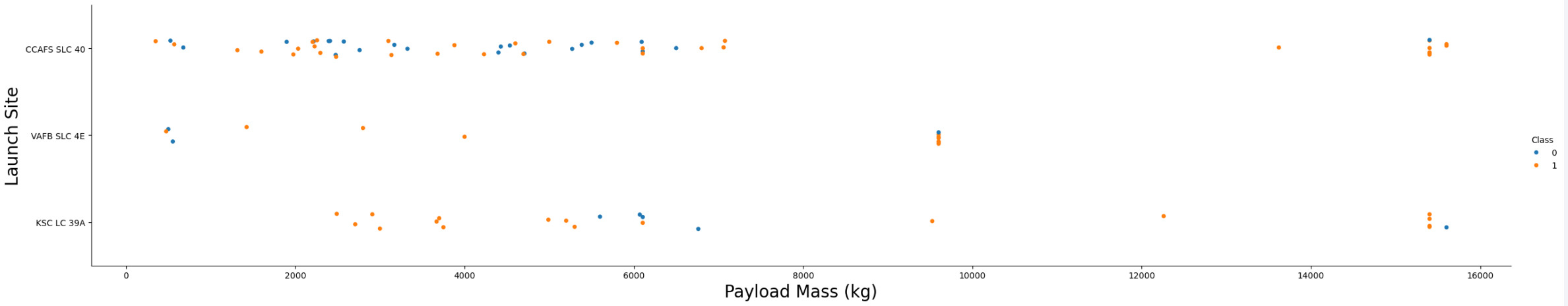
- Early flights: Lower success rate (blue = failure)
- Later flights: Higher success rate (orange = success)
- ~50% of launches originated from CCAFS SLC-40
- VAFB SLC-4E and KSC LC-39A exhibit **higher success rates

Conclusion: Newer launches are significantly more likely to succeed



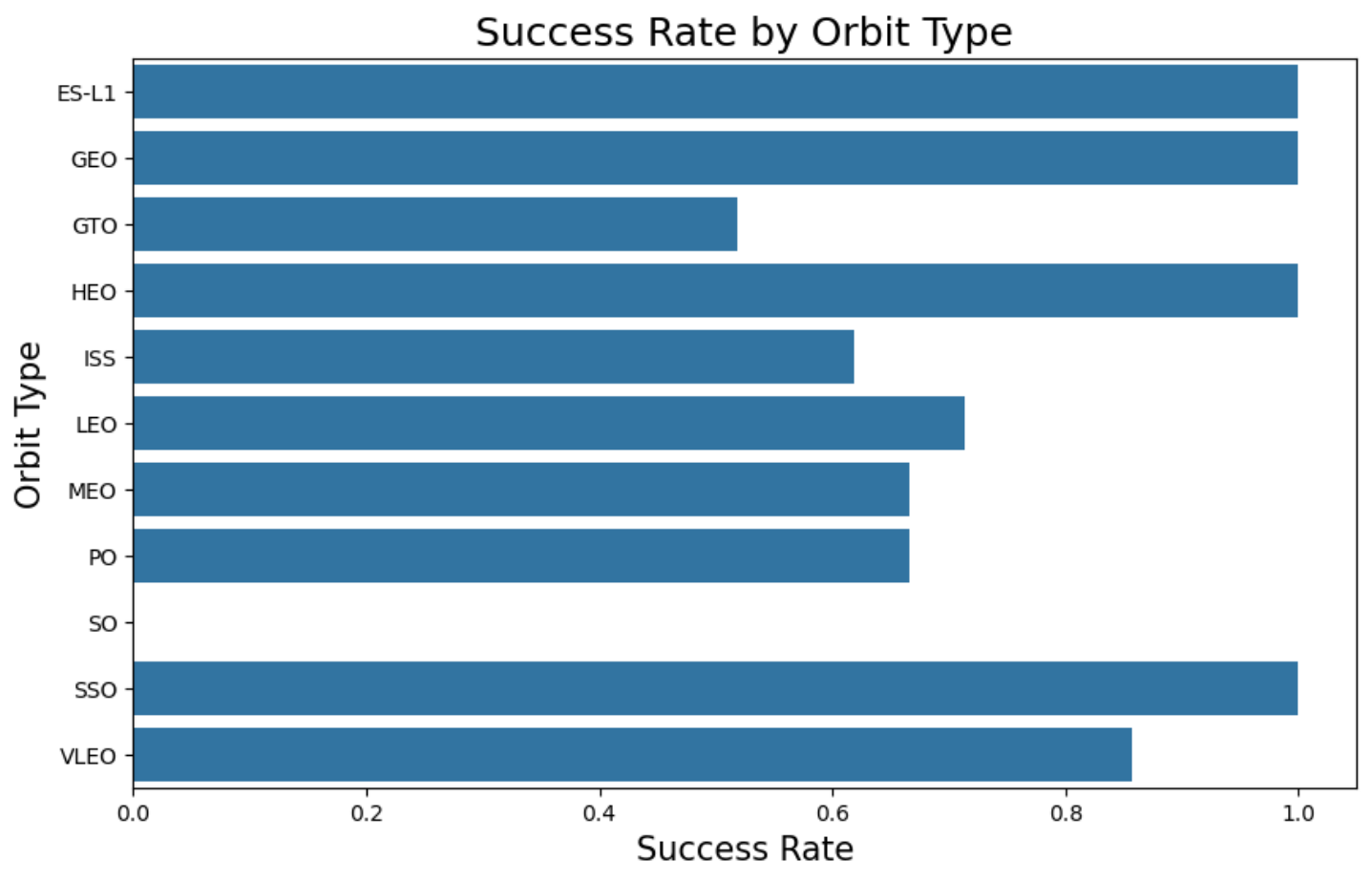
Payload vs. Launch Site

- 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



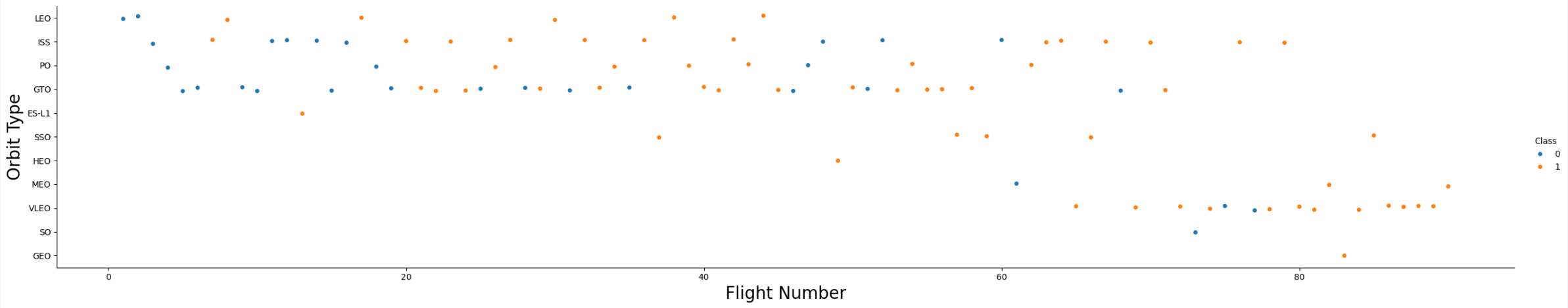
Success Rate vs. Orbit Type

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



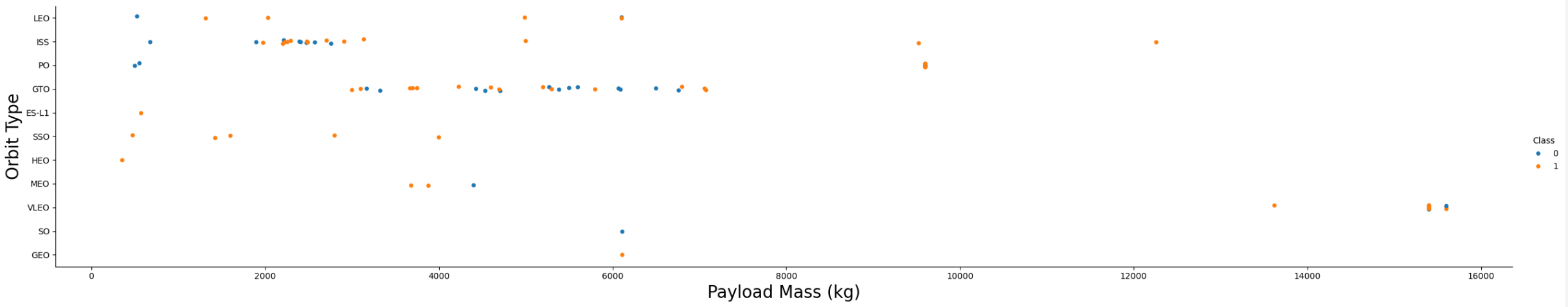
Flight Number vs. Orbit Type

- 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



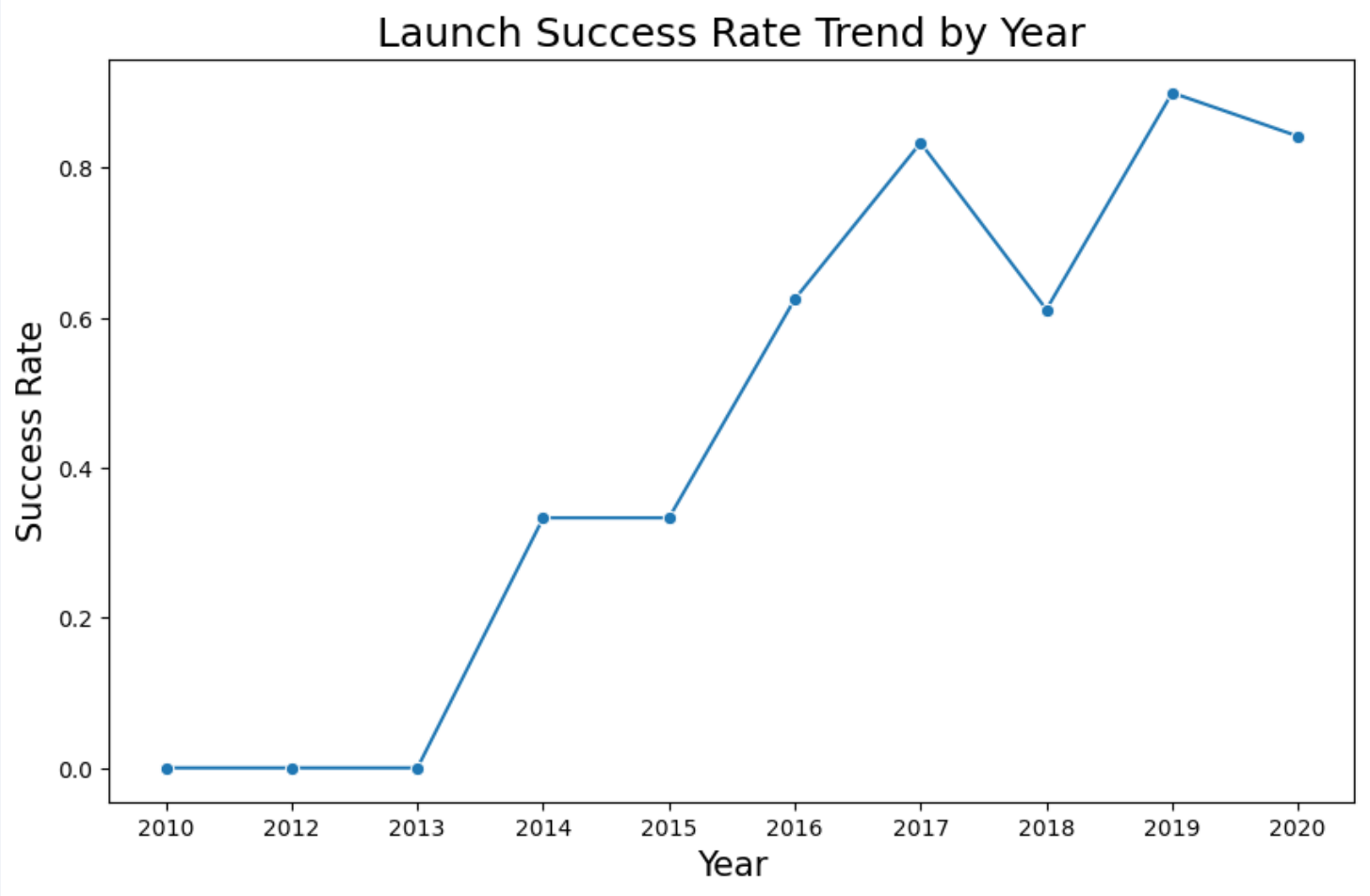
Payload vs. Orbit Type

- 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

To Find the unique launch site names, we had to use the DISTINCT key word.

The unique Launch Sites are the

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

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

```
%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'

To find the launch sites names that start with ‘CCA’, we had to add a condition with the LIKE key word, to only select those rows with given word. The results we also limited to the first five returns with the LIMIT 5 key word pair.

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	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

Total Payload Mass

The total payload was found to be 48213 kg. This was calculated by using the SUM keyword when selecting the “Payload_Mass__kg_” column.

```
%sql SELECT SUM("Payload_Mass__kg_") AS Total_Payload FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA (CRS)%';
 * sqlite:///my_data1.db
Done.
```

Total_Payload
48213

Average Payload Mass by F9 v1.1

To calculate the average payload mass carried by booster version F9 v1.1, we had to add the AVG key word to our query, when selecting the “Payload_Mass__kg_” column.

```
%sql SELECT AVG("Payload_Mass__kg_") AS Average_Payload FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

Average_Payload
2928.4

First Successful Ground Landing Date

To find the first successful ground landing date, the MIN keyword had to be used when selecting the “Date” column. An additional condition had to be added to only look for the successful landing outcomes.

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

Hint:Use min function

```
%sql SELECT MIN("Date") AS First_Successful_Ground_Pad_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

First_Successful_Ground_Pad_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

To find the successful drone ship landing with a given payload, a condition had to be added to only look for the successful landing on drone ships, where the payload was between 4000 kg and 6000 kg.

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

Here, we had to group the outcome with the GROUP BY key word, and use COUNT, to get to total value for each mission outcome.

List the total number of successful and failure mission outcomes

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

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

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

Boosters Carried Maximum Payload

To return the MAX payload mass, we had to do a nested search

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

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

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

This task required a conditional search to List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%sql SELECT substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr("Date", 0, 5) = '2015';
```

```
 * 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

The results indicate that “No attempt” had the highest outcome, while “Precluded (drone ship” had the lowest with only a single case.

To retrieve this data, we had to narrow down the query to a given time frame, and count the landing outcomes in a descending order.

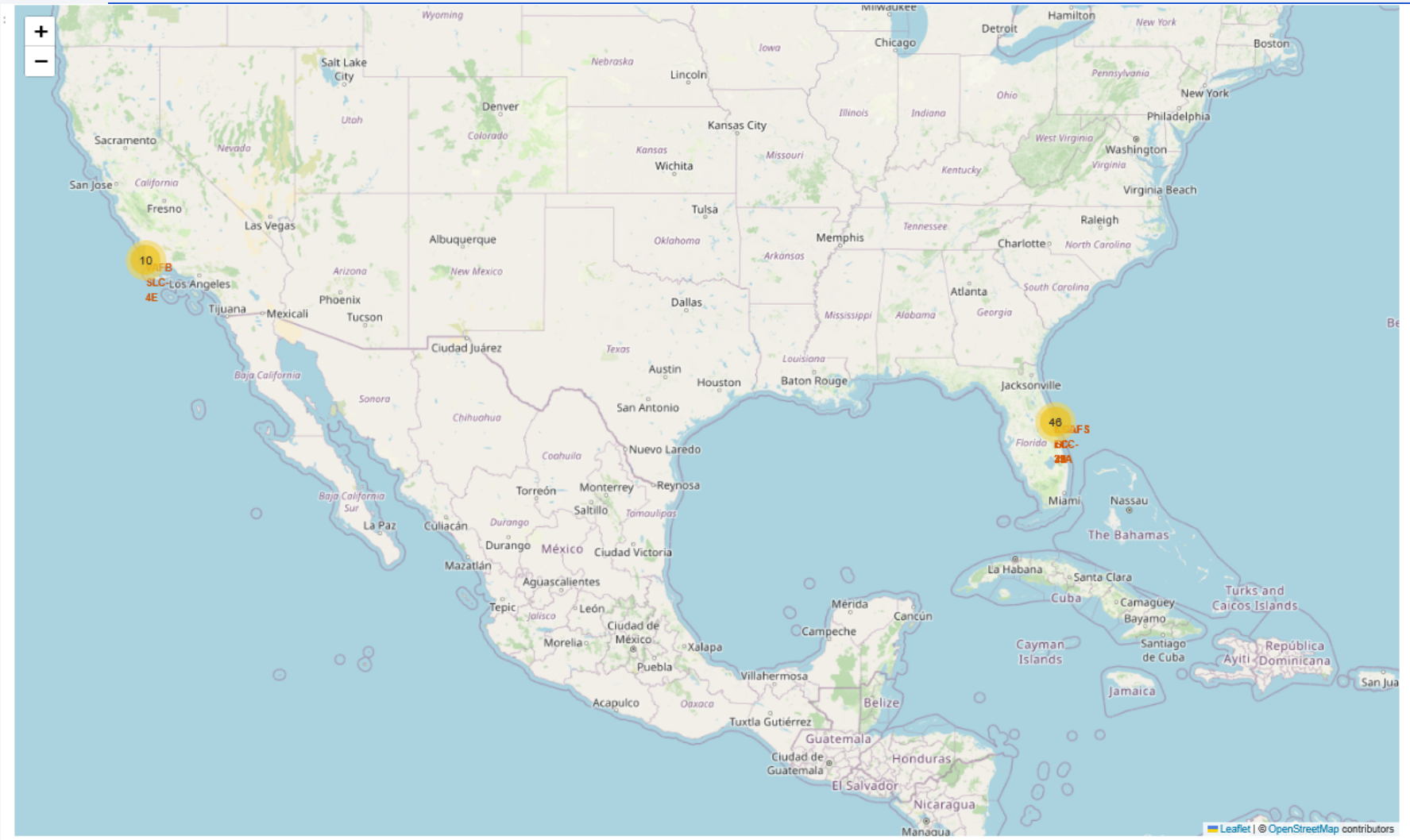
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

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

Section 3

Launch Sites Proximities Analysis

Launch Site

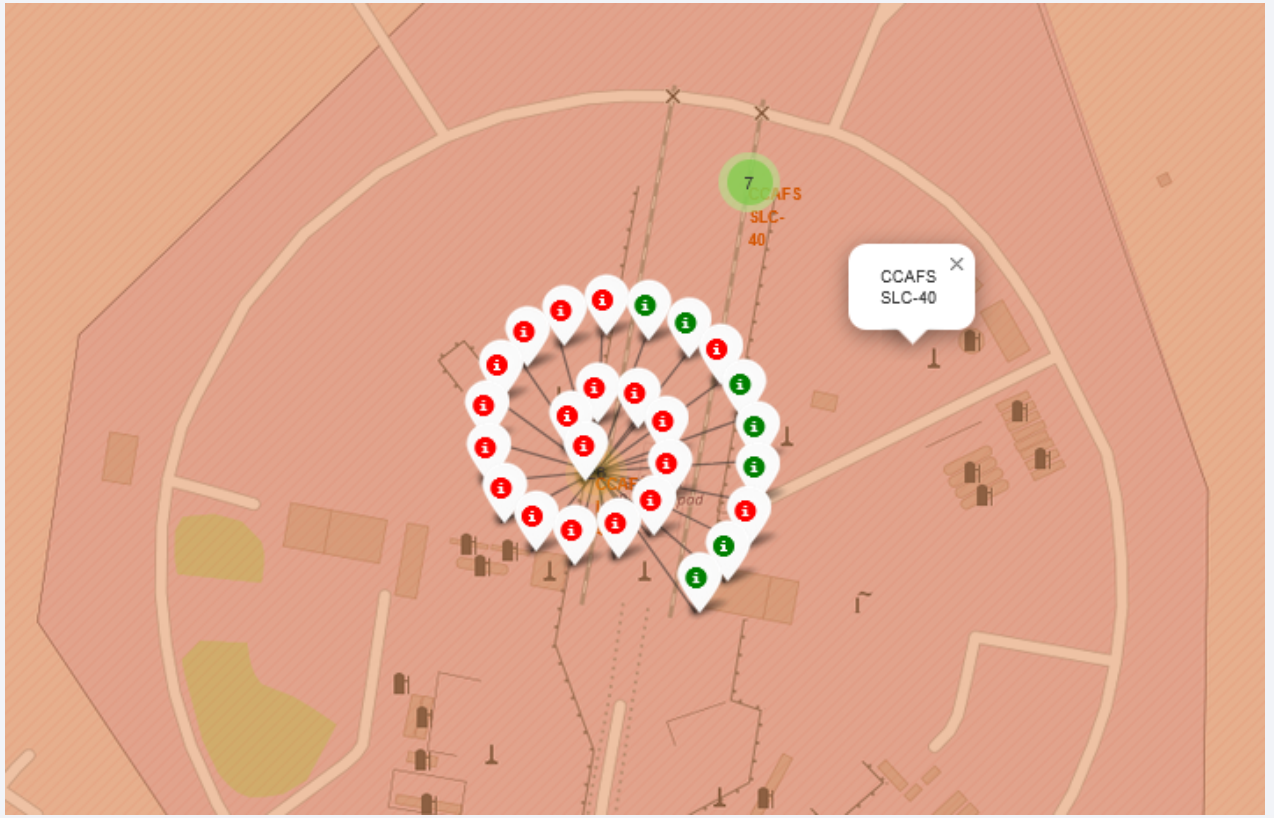


Near Equator: the closer the launch site to the equator, the **easier**it is **to launch**to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an **additional natural boost**–due to the rotational speed of earth –that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

Launch Outcome

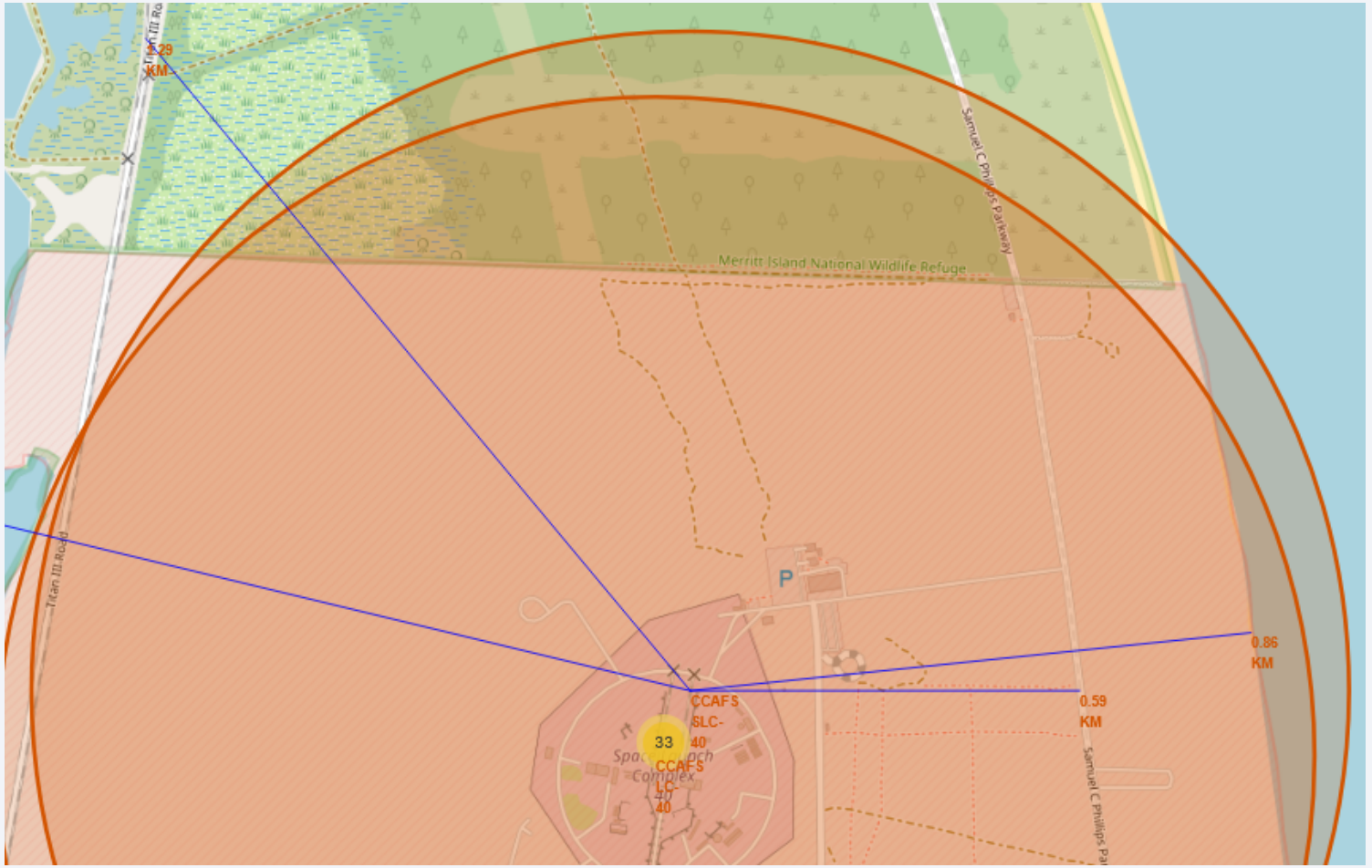
Outcomes:

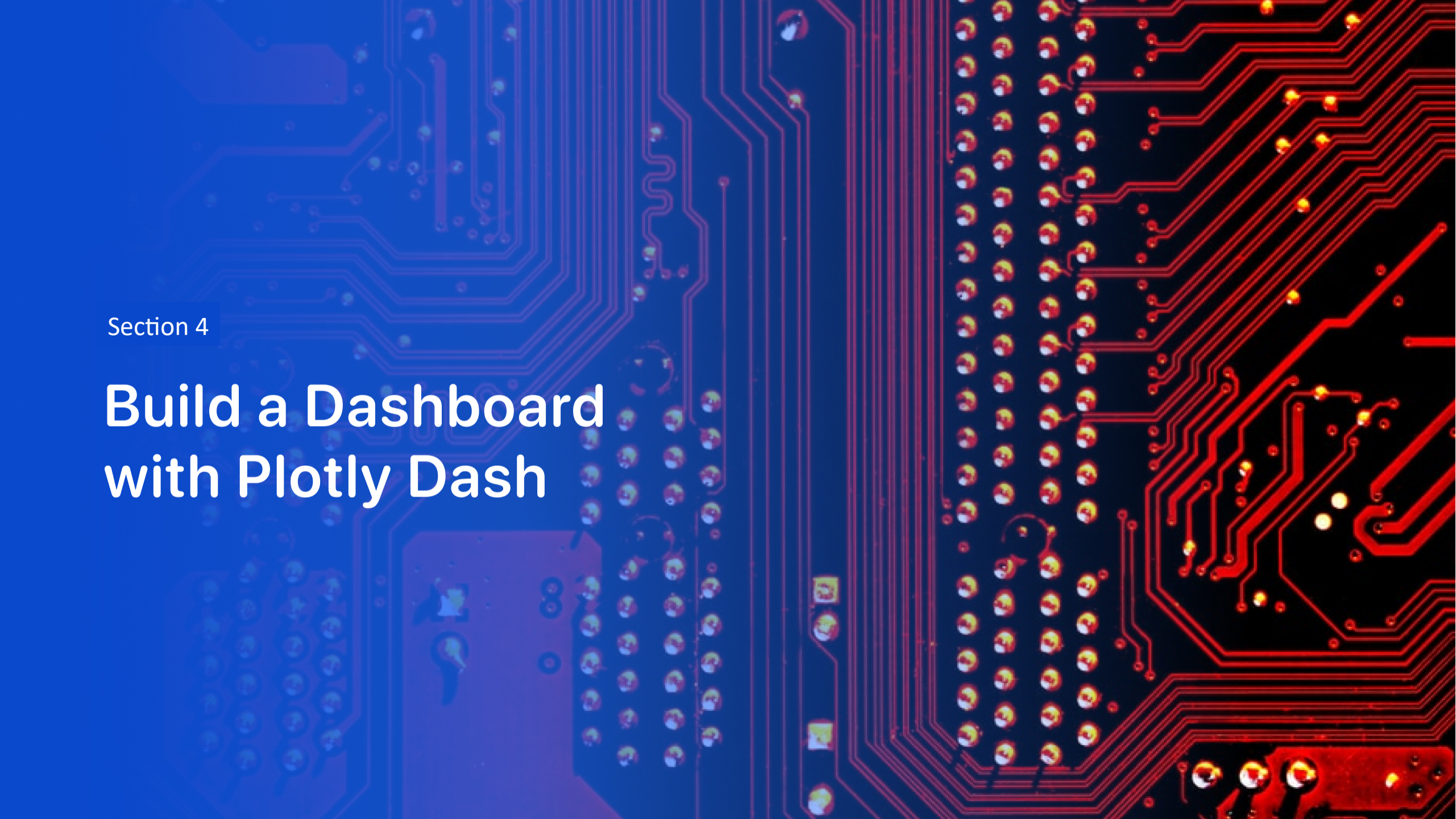
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site **CCAFS SLC-40** has a 10/33 or about 30% success rate



Distance to Proximity

On this map we can see how close proximity is the CCAFS SLC-40 launch site to the coast line, railway and higway.

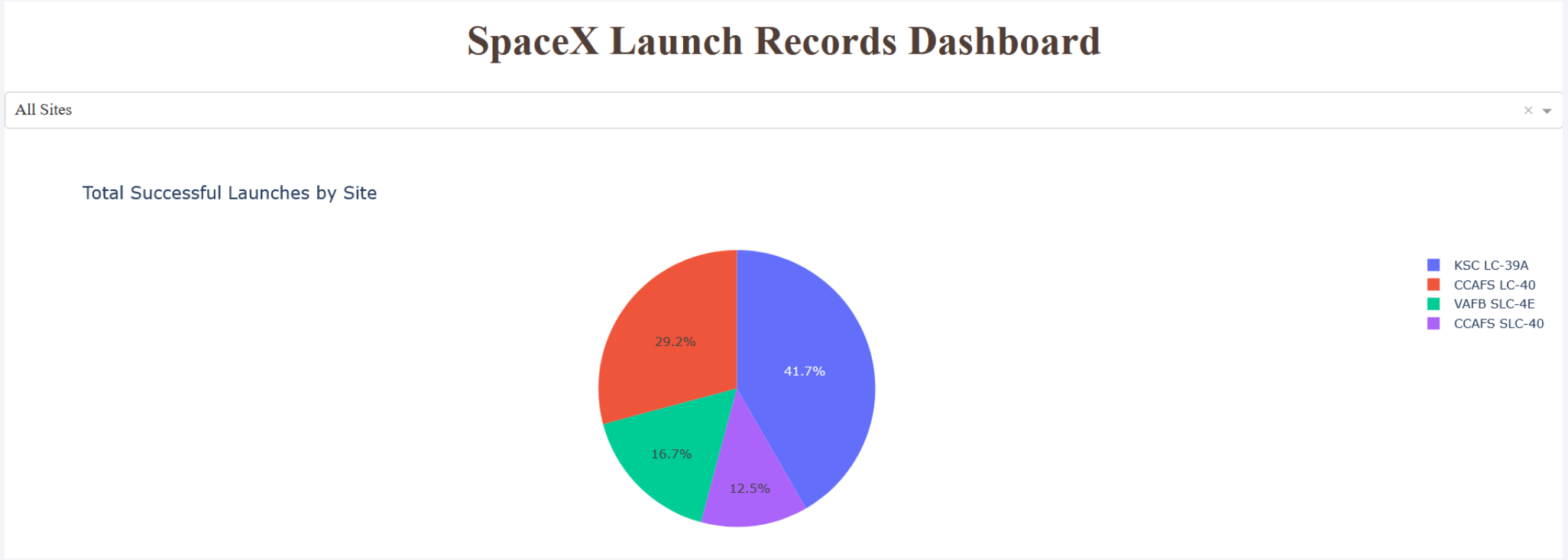



Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site Location

Success as Percent of Total

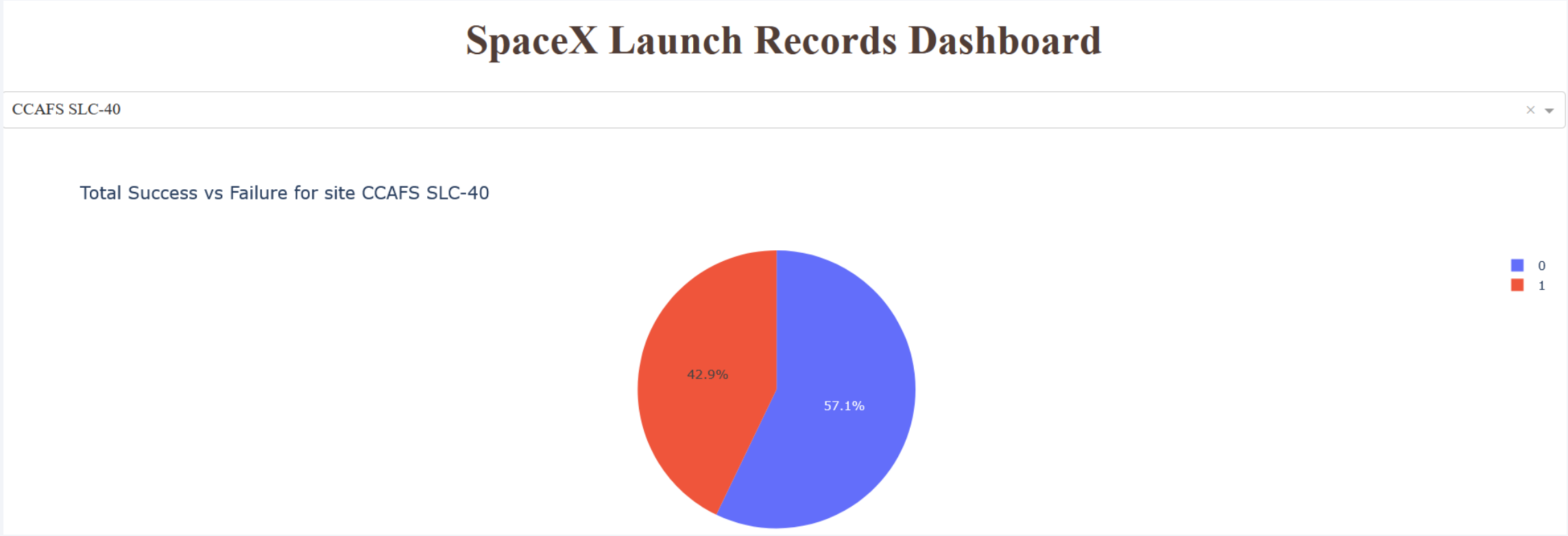
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Total Success vs Failure

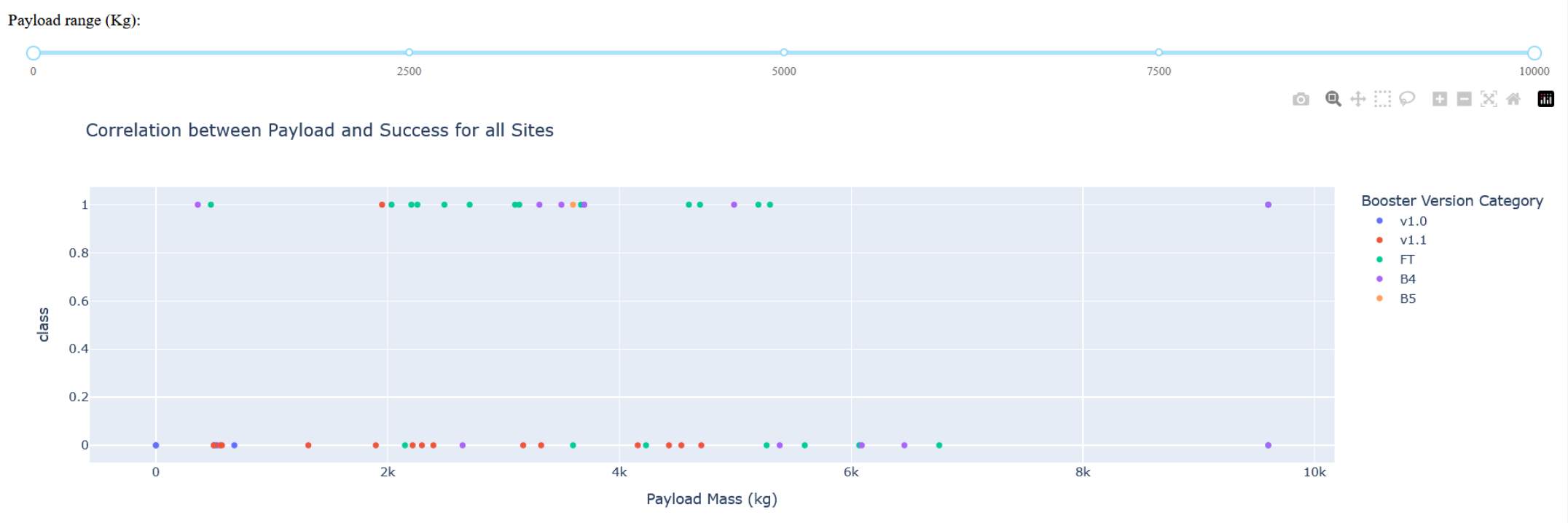
Success as Percent of Total

- CCAFS SCL-40 has the most unsuccessful launches between launch sites (41.9%)



Correlation Between Payload and Success

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



Section 5

Predictive Analysis (Classification)

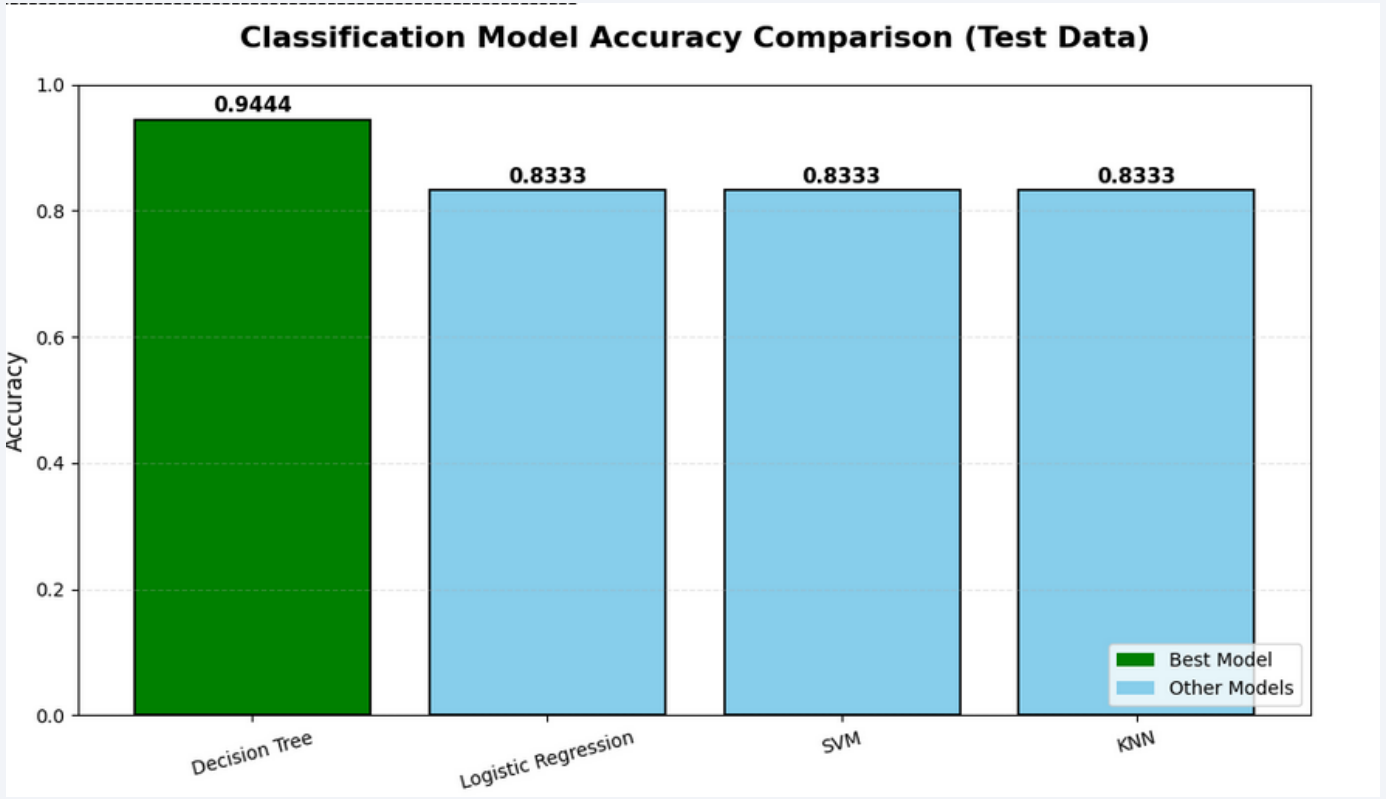
Classification Accuracy

Evaluating models on test data

- Logistic Regression: 0.8333
- SVM: 0.8333
- Decision Tree: 0.9444
- KNN: 0.8333

BEST MODEL: Decision Tree

- Test Accuracy: 0.9444



Confusion Matrix

Performance Summary
A confusion matrix summarizes the performance of a classification algorithm

The fact that there are false positives (Type 1 error) is not good

Confusion Matrix Outputs:
12 True positive
5 True negative
1 False positive
0 False negative

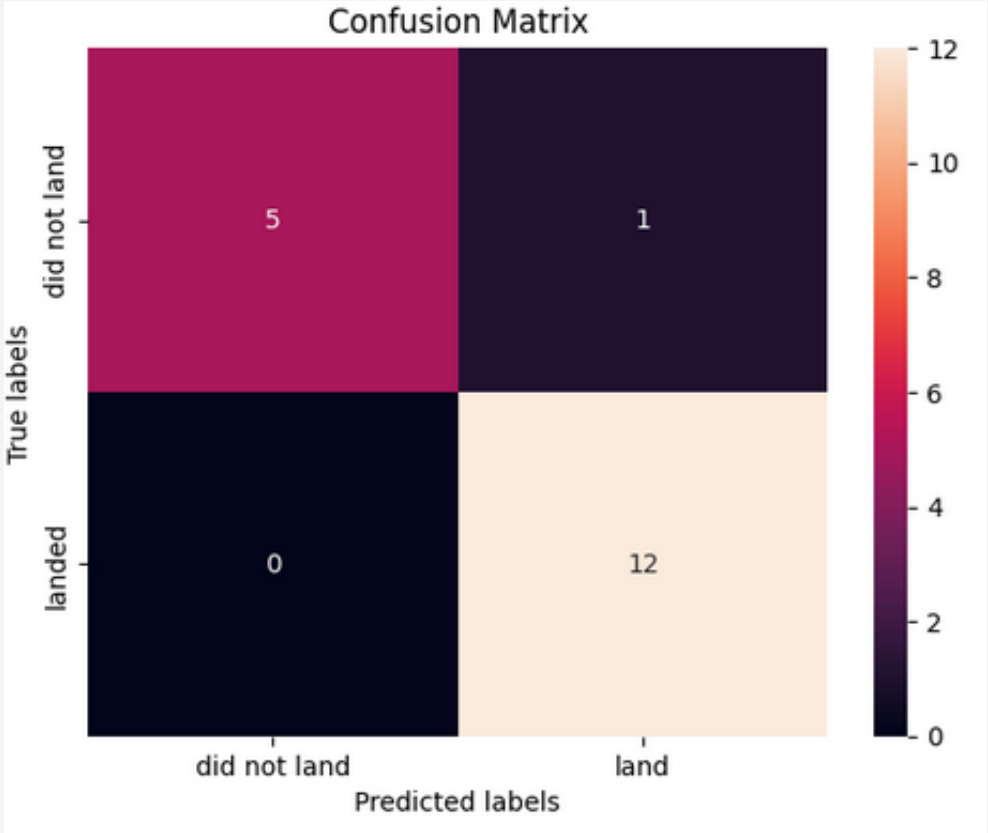
Precision= TP / (TP + FP) =12 / 13 = 0.923

Recall= TP / (TP + FN) =12 / 12 = 1

F1 Score= 2 * (Precision * Recall) / (Precision + Recall) =

2 * (0.923 * 1) / (0.923+1) = 0.9599

Accuracy= (TP + TN) / (TP + TN + FP + FN) = 0.944



Conclusions

Model Performance:

- The models performed similarly on the test set, with the Decision Tree slightly outperforming the others at 94.44% accuracy (vs. 83.33% for Logistic Regression, SVM, and KNN).
- The Decision Tree achieved perfect recall (100%) and zero false negatives, making it highly reliable for identifying successful landings.

Launch Site Location:

- Equator: Most launch sites (e.g., CCAFS, KSC) are near the equator, leveraging Earth's rotational speed (~465 m/s) for a natural boost, reducing fuel and booster requirements.
- Coast: All launch sites are coastal, enabling safe over-water trajectories and emergency abort zones.

Launch Success Trends:

- Increases over time: Success rate improves with mission experience and iterative engineering.

Launch Site Performance:

- KSC LC-39A: Highest success rate overall; 100% success for payloads under 5,500 kg.

Orbit Types:

- ES-L1, GEO, HEO, SSO: Achieved 100% landing success, likely due to favorable energy profiles and mission planning.

Conclusions

Payload Mass Impact:

- Higher payload mass correlates with higher success rate across sites — counterintuitive but possibly due to more robust mission profiles, larger fairings, and operational maturity in high, suggesting selection bias toward proven configurations.

Dataset & Future Work:

- Dataset: Current dataset (90 launches) is limited. A larger, updated dataset would improve generalizability and model robustness.
- Feature Analysis / PCA: Additional exploratory analysis or Principal Component Analysis (PCA) could reveal latent patterns and reduce noise, potentially improving accuracy.
- XGBoost: A powerful gradient-boosting model not used here. Future work should evaluate XGBoost, which often outperforms traditional models on structured data and may further increase predictive performance.

Thank you!