IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Nganje Lumen>
<Date>

Outline

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

Executive Summary

•**Objective:** The goal was to build a predictive model to determine the success rate of a rocket landing attempt based on launch data, helping SpaceY optimize its future launch strategy.
•**Methodology:** You used data from the SpaceX API and Wikipedia, cleaned it, performed exploratory data analysis (EDA) using SQL and visualizations, and trained four machine learning models: Logistic Regression, SVM, Decision Tree, and KNN.
•**Insights after performing Exploratory Data Analysis:**
    •Landing success rates improved drastically around the year **2017**.
    •**KSC LC-39A** and **CCAFS SLC-40** were the primary launch sites with the highest success volumes.
    •Heavier payloads in the **VLEO** (Very Low Earth Orbit) orbital class showed a near-perfect success rate in later flights.
•**Performance of the model:** All four machine learning models achieved the same test accuracy of approximately **83.33%** (15/1815/18
15/18
test samples correct).
•**Recommendations Moving forward:** All candidate models eqally performed and hence any can be chosen but my preference is the Logistic Regression model. The data suggests that focusing on launch patterns established after 2017 (with high **Flight Numbers** and optimized **Booster Versions**) is key to success for SpaceY.

Key Findings & Conclusions
•**Reliability Increase**: SpaceX's reliability improved dramatically over time; later flight numbers highly correlate with success.
•**Optimal Parameters**: Launches with payload masses between 2,000 kg and 3,700 kg, especially from **KSC LC-39A**, showed high success rates.
•**Best Predictive Model**: The **Decision Tree Classifier** emerged as the best-performing model (highest cross-validation score), capable of predicting outcomes with an accuracy of approximately **88.9%**.
The report concludes that leveraging these data-driven insights provides a strong competitive advantage for bidding on future space missions

Introduction

•**Business Context:** The commercial space industry is highly competitive, and reusable rockets are key to profitability. SpaceX has demonstrated success in this area.

•**The Problem:** The client, SpaceY, needs to understand the factors that contribute to successful rocket landings to optimize their own launch mission planning and investment strategy.

•**Project Objective:** The goal of this report is to analyze historical public data from SpaceX launches to identify patterns, build predictive models, and provide actionable recommendations for SpaceY's leadership.

Section 1

Methodology

Methodology

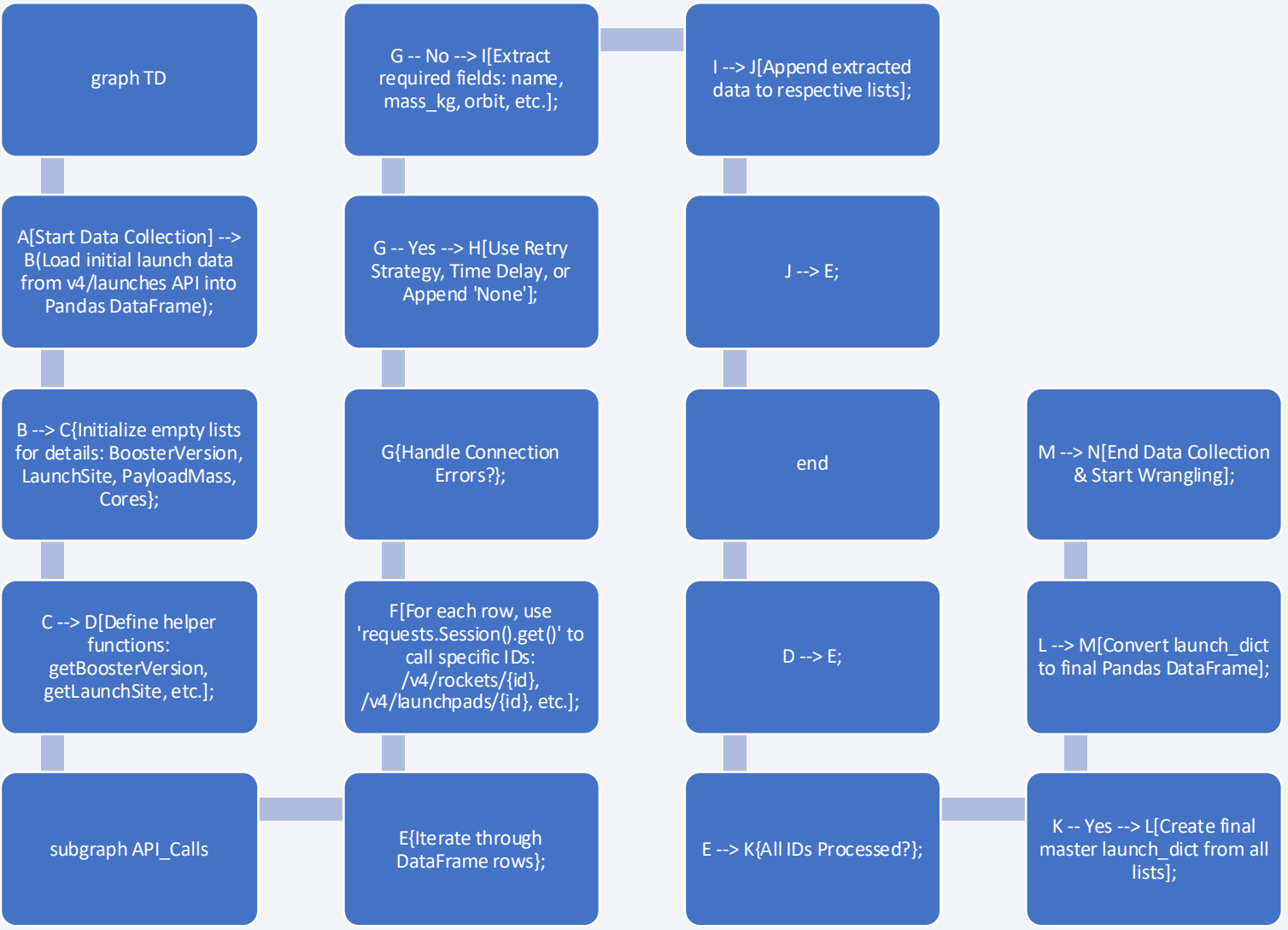
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data was collected from a publicly available API made available by SpaceX
- Technologies and Techniques like requests, beautiful soup and web scraping were used for collecting the data

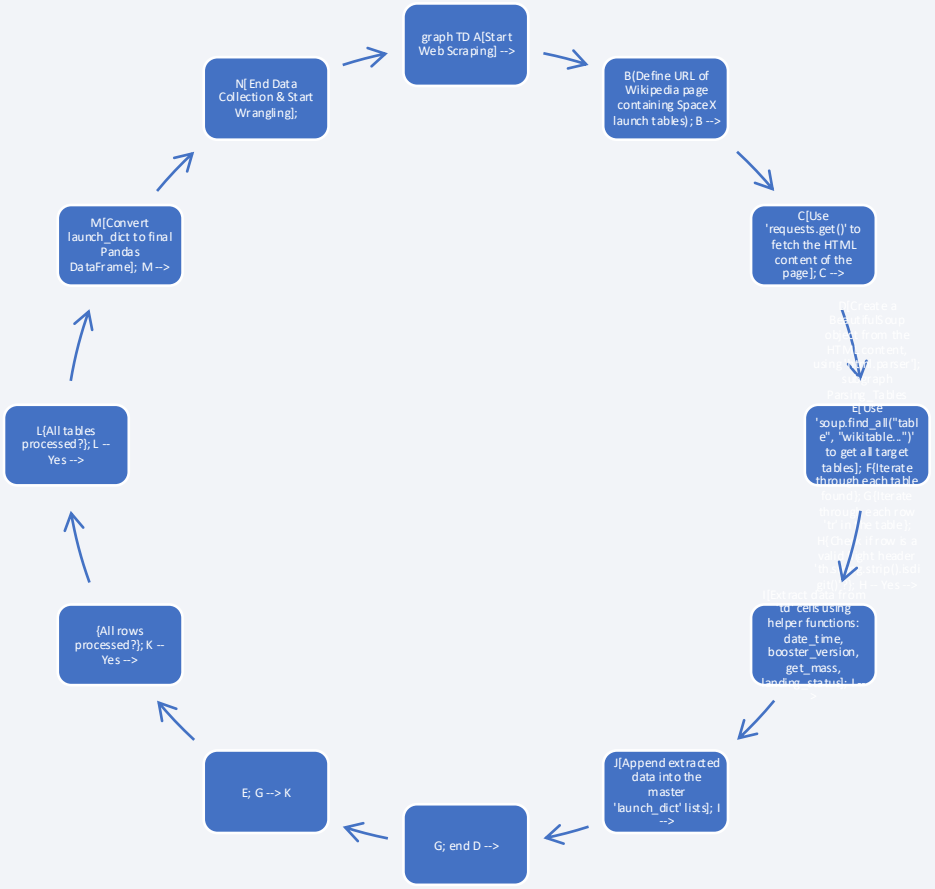
Data Collection – SpaceX API



https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Scatter plots**: Scatter plots help identify correlations (or lack thereof) between continuous variables like PayloadMass and discrete variables like FlightNumber, informing feature selection for machine learning models.

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/edadataviz.ipynb

- **Bar plots** : Bar charts facilitate easy comparison between different groups (e.g., orbits), while line plots are ideal for visualizing trends over time, which is critical for a time-series-based project.

EDA with SQL

- **Analyze Launch Sites:**
  - Display the names of all unique launch sites.
  - Retrieve a limited number of records (e.g., 5) for launch sites beginning with a specific string (e.g., 'CCA').
- **Analyze Payload Data:**
  - Calculate the total payload mass carried by boosters associated with a specific customer (e.g., NASA (CRS)).
  - Calculate the average payload mass carried by a specific booster version (e.g., 'F9 v1.1').
  - Identify the names of the booster versions that carried the maximum payload mass across all launches

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Markers**: Markers were added for each of the four launch sites (CCAFS SLC-40, CCAFS ERC-LZ, KSC LC-39A, and VAFB SLC 4E) and for the NASA Johnson Space Center (JSC) as a central reference point.
- **Circles/Circle Markers**: Circles were added around each launch site with a specified radius (e.g., 1000 meters).
- **Lines**: Lines were drawn from a selected launch site (e.g., CCAFS SLC-40) to nearby geographic features like the coastline, rail line, and perimeter road.

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/lab_jupyter_launch_site_location.ipynb

Rationale for Adding the Objects

- These objects were added for **exploratory data analysis (EDA)** to gain geographical context and identify patterns related to the launch locations:
- **Visualizing Locations**: Markers and circles help pinpoint the exact latitude and longitude of the launch sites on an interactive map, providing an intuitive understanding of their physical locations that raw coordinates cannot offer.
- **Assessing Proximity**: Drawing lines to nearby features allows for the calculation of distances using formulas like Haversine distance, which helps in understanding environmental and logistical factors.
- **Finding Trends**: By visually analyzing the proximity of launch sites to key elements like the coastline or major transport routes, the user can infer potential influences on launch success or failure, which could be useful for feature engineering in machine learning models.
- **Mapping Outcomes**: Different colored markers (green for success, red for failure) within a **MarkerCluster** plugin were used to visually compare the success rates at various sites, helping to identify more reliable locations.

Build a Dashboard with Plotly Dash

Summary of Plots/Graphs and Interactions

- **Pie Chart**: A pie chart is used to visualize the launch success rate for a specific launch site selected via a dropdown menu. It displays the percentage of successful vs. failed missions.
- **Scatter Chart**: A scatter chart displays the correlation between PayloadMass and launch outcome (Class as 0 or 1). This chart is interactive, updating dynamically based on a user-controlled slider input for a range of payload mass values.
- **Interactions**:
  - **Dropdown List**: A dropdown menu allows users to select a specific launch site from the four available locations (CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A, CCAFS ERC-LZ).
  - **Range Slider**: A slider allows the user to filter the data displayed in the scatter plot based on a range of payload mass (e.g., 0 to 10,000 kg).

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/Interractive_dashboard_app_with_plotly%20dash.ipynb

Explanation of Why Those Plots and Interactions Were Added

- **Pie Chart**: This chart was added to provide a clear, immediate understanding of the **overall success composition** at a specific location, allowing quick comparison between sites to identify the most reliable launch areas.
- **Scatter Chart**: This was used to explore the **relationship between payload mass and success outcome**, allowing stakeholders to visually determine if a certain range of payload masses increases or decreases the likelihood of a successful mission.

Predictive Analysis (Classification)

Model Development Process Summary

- **Models Built**: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN) are all constructed to predict the binary outcome of a rocket landing (success or failure).
- **Evaluation Metrics**: Models are evaluated using key metrics such as **Accuracy**, **Precision**, **Recall**, and the **F1 score**. A confusion matrix is also used to visualize true positives, true negatives, false positives, and false negatives.
- **Model Improvement**: Each model is improved through **hyperparameter tuning** using techniques like GridSearchCV and **cross-validation** to find the optimal parameters and ensure robust performance on unseen data.
- **Best Model Selection**: The model with the highest overall accuracy or F1 score is selected as the best performing model. Results often vary, but usually, one model may slightly outperform others depending on the specific metrics used.

https://github.com/lumenboss/SpaceX-Falcon9-first-stage-reuse/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Data Preparation Phase: **Feature Engineering** and **data normalization** (scaling numerical features).

Model Building: **Train multiple classification models** (e.g., Logistic Regression, SVM, Decision Tree, KNN).

Model Optimization: **Hyperparameter tuning** via GridSearchCV and **cross-validation**.

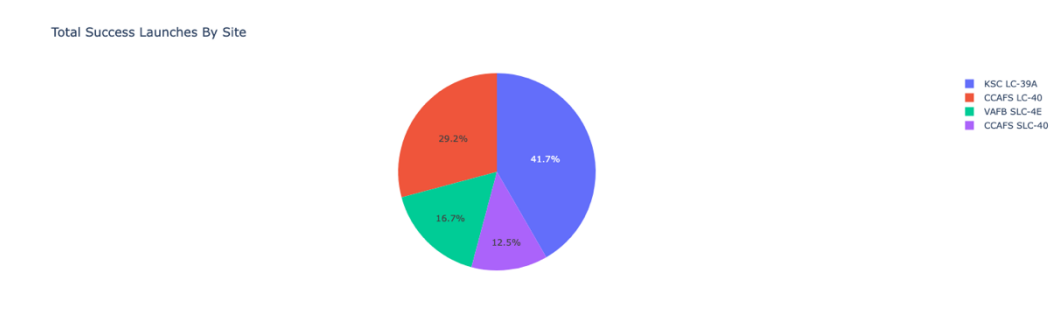
Model Evaluation: Calculate and compare **Accuracy, Precision, Recall, and F1-Score** using a **confusion matrix**.

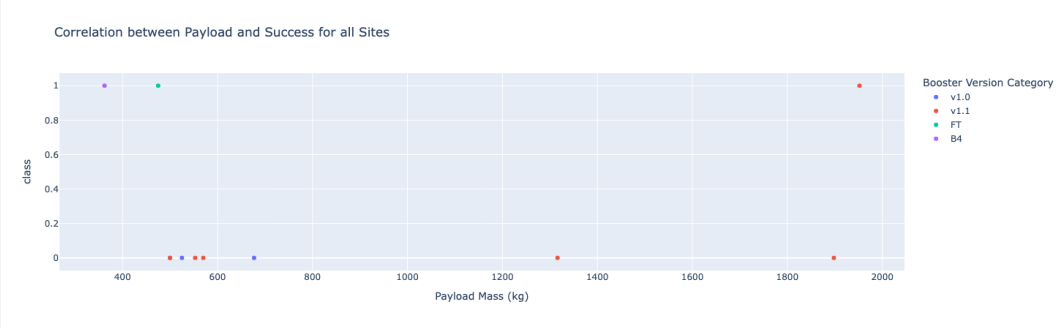
Reporting: **Select best performing model** and prepare final report for stakeholders.

Results

- Exploratory data analysis results
  - **Top Site**: **KSC LC-39A** has the highest success rate amongst the landing sites..
  - **Site Distribution**: Most launches are concentrated at **CCAFS SLC-40**

- Interactive analytics demo in screenshots

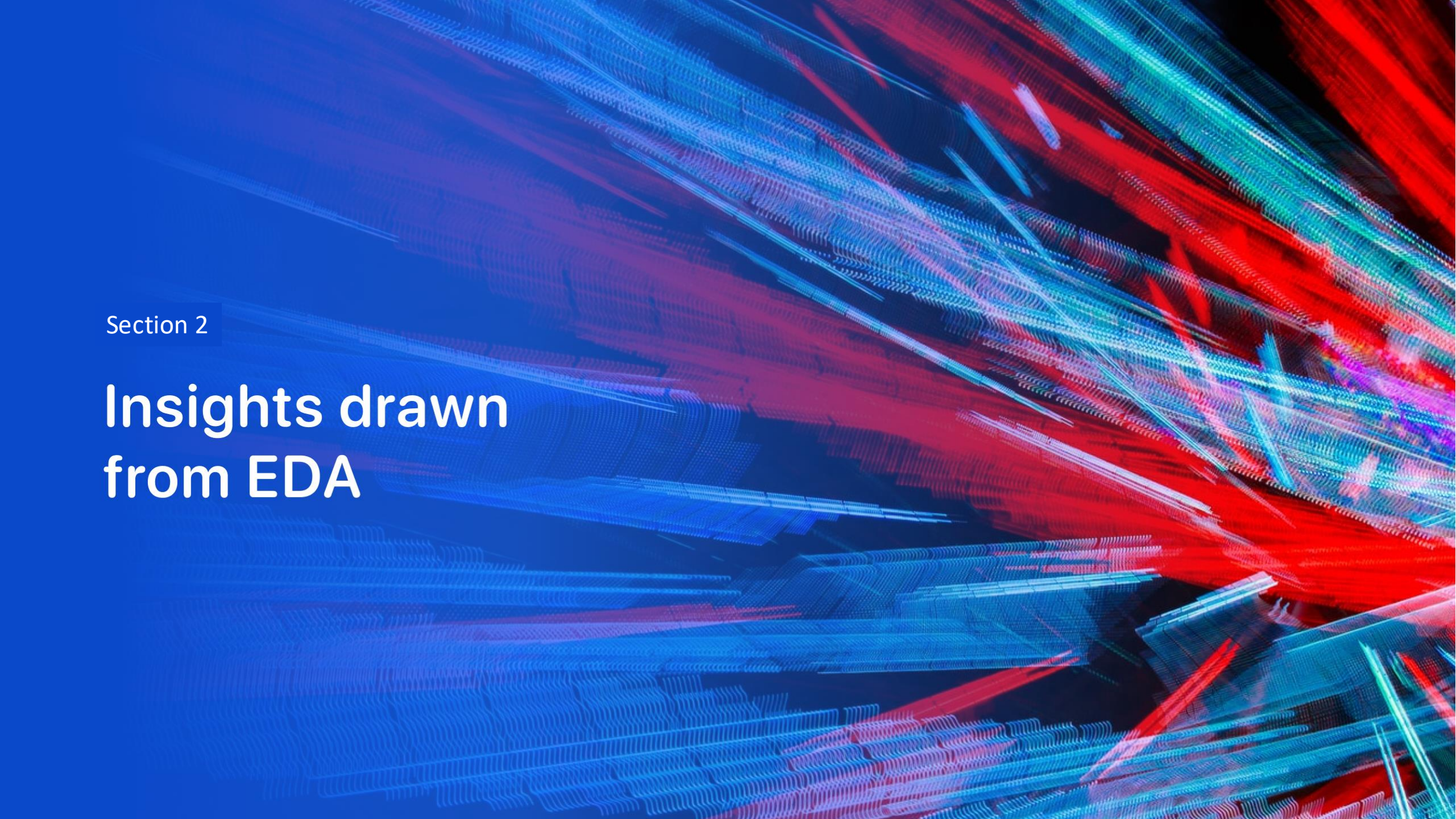




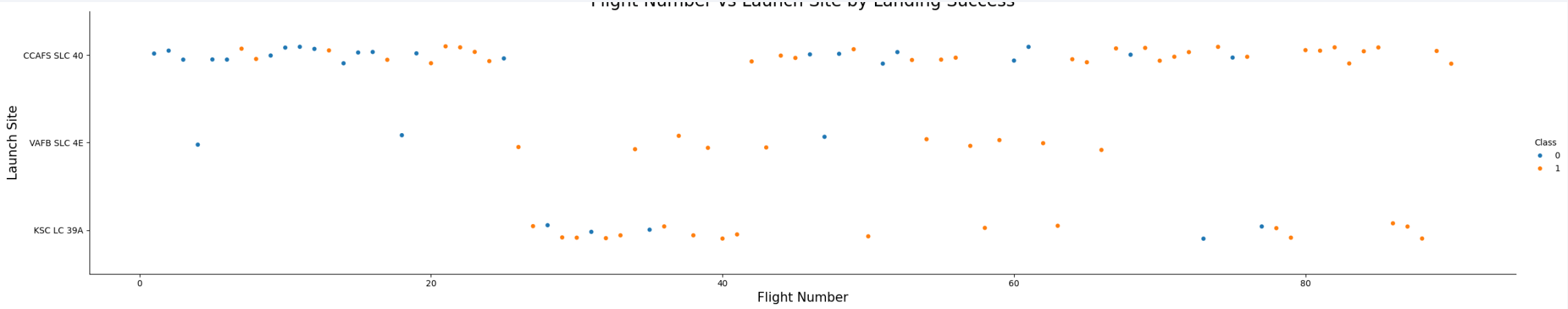
- Predictive analysis results
  - All models (Logistic regression, SVM, KNN and Decision Tree) equally performed well with an accuracy of 83.33%

Section 2

Insights drawn from EDA

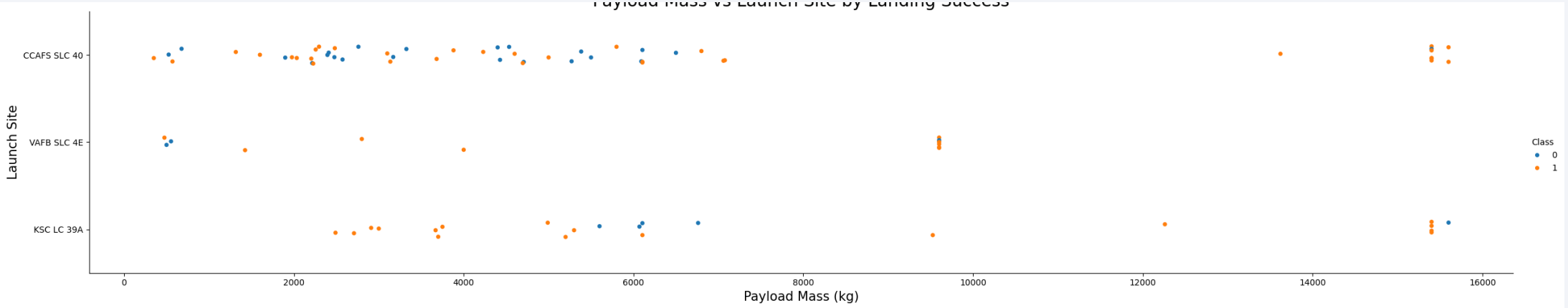


Flight Number vs. Launch Site



- As Flight number increases the success rate improves significantly
- KSC LC-39A shows the highest success rate

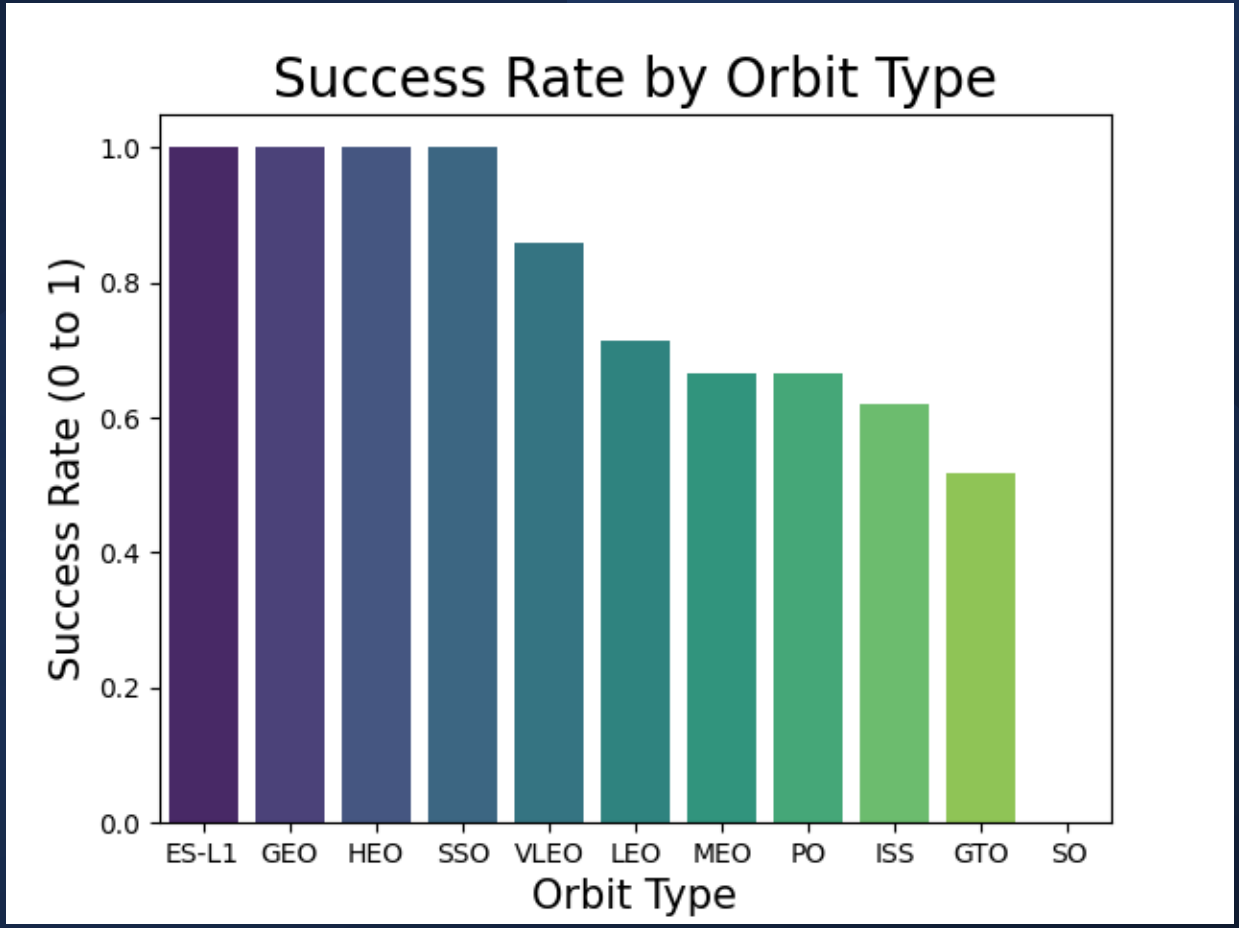
Payload vs. Launch Site

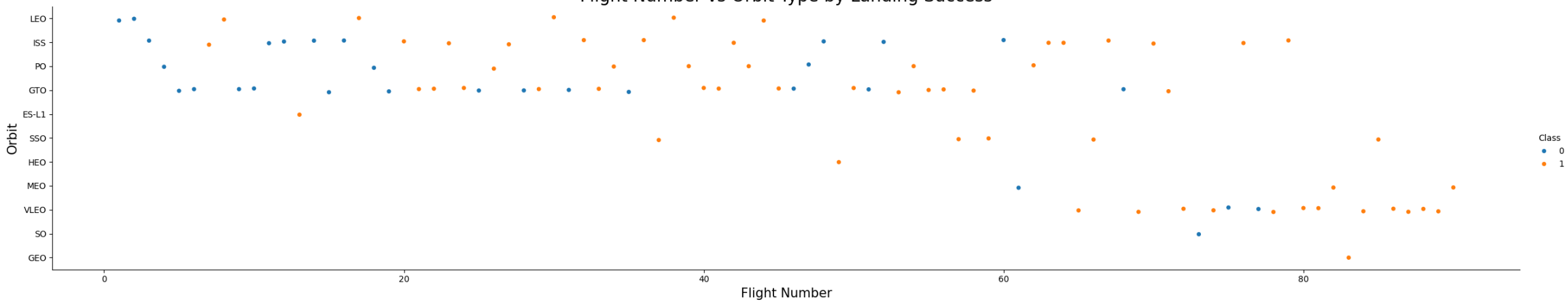


- The success rate is higher for higher masses
- KSC LC-39A has the highest success rate

Success Rate vs. Orbit Type

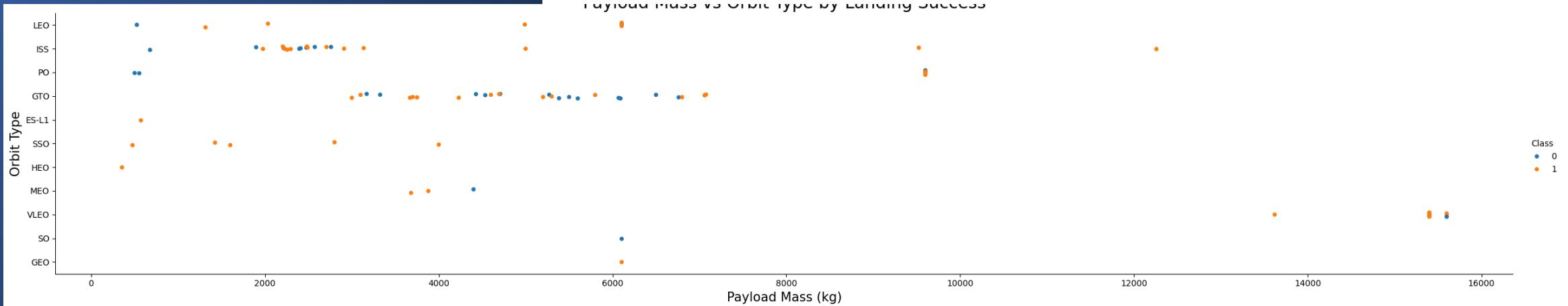
GEO, HEO, SSO, ESL1 all jointly had the highest success rate





Flight Number vs. Orbit Type

- You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

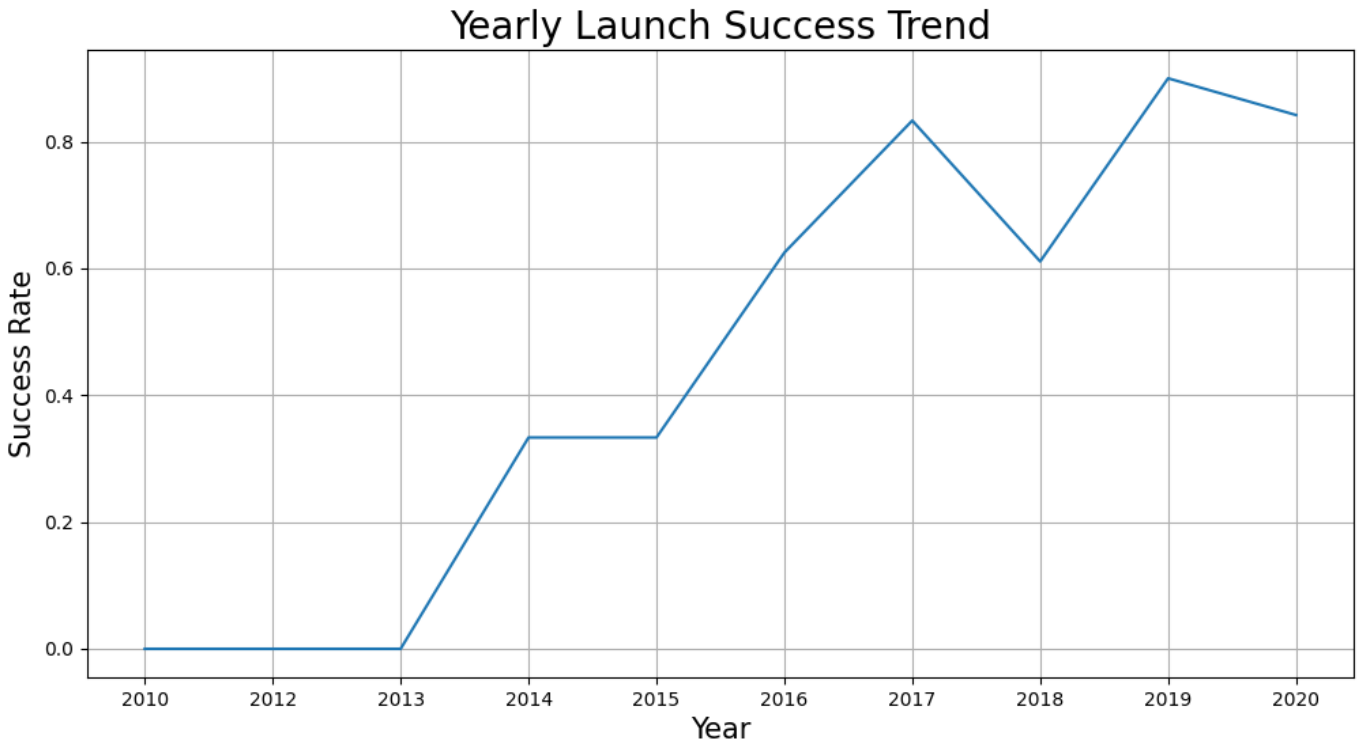


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

As time goes by, the success rate significantly increases.

Launch Success Yearly Trend



All Launch Site Names

- Find the names of the unique launch sites
- LaunchSite CCAFS SLC 40 55 KSC LC 39A 22 VAFB SLC 4E 13 Name: count, dtype: int64
- Present your query result with a short explanation here
- %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

Present your query result with a short explanation here

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

This is an sql query that selects all columns in the table that have a column value that's like

CCA and the result is limited to the first 5 rows

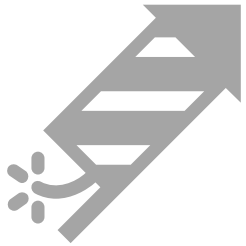
Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - 45,596 kg
- Present your query result with a short explanation here
  - SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA%';

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

The average payload mass for booster version **F9 v1.1** is approximately **2928.4 kg**



Present your query result with a short explanation here

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

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

- **December 22, 2015**.

Present your query result with a short explanation here

- ***SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';***

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000	**F9 FT (Falcon 9 Full Thrust).**
Present your query result with a short explanation here	**SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000**

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

- **98 Successful missions and 1 Failure mission**

Present your query result with a short explanation here

- ***SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total_Number FROM SPACEXTBL GROUP BY Mission_Outcome;***

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

F9 B5

Present your query result with a short explanation here

SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

Failed Missions in 2015

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Landing Outcome	Booster version	Launch site name
Failure drone ship	F9 v1.1	CCAFS LC-40

- Present your query result with a short explanation here
  - *SELECT "Landing _Outcome", BOOSTER_VERSION, "Launch Site" FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND YEAR(Date) = 2015;*

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

- 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

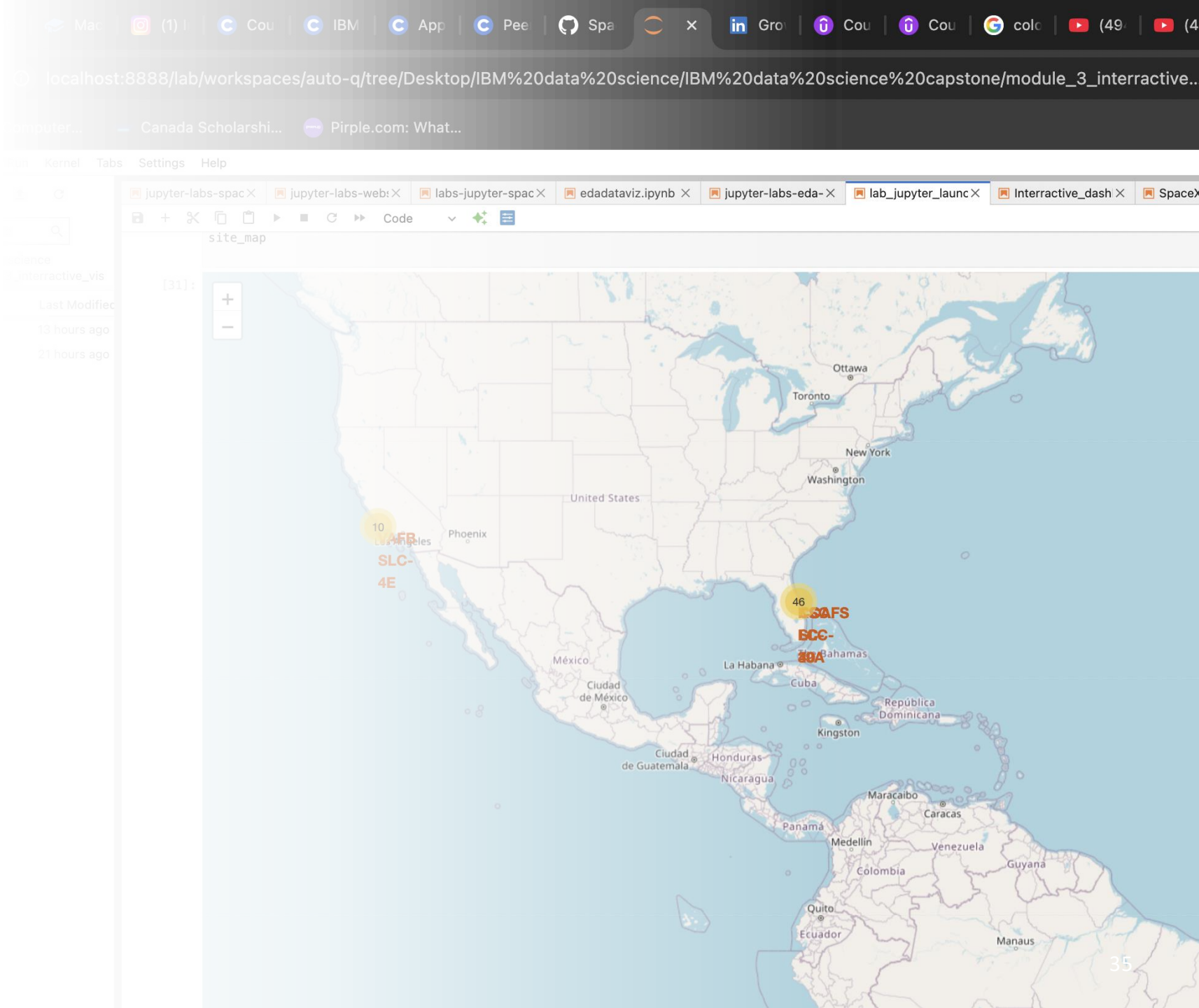
Landing outcome	Count
Failure drone ship	5
Successs groundpad	5

- Present your query result with a short explanation here
  - ```
    SELECT "Landing _Outcome", COUNT(*) FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04'
    AND '2017-03-20' GROUP BY "Landing _Outcome" ORDER BY COUNT(*) DESC;
    ```
    ```


Section 3
Launch Sites
Proximities Analysis

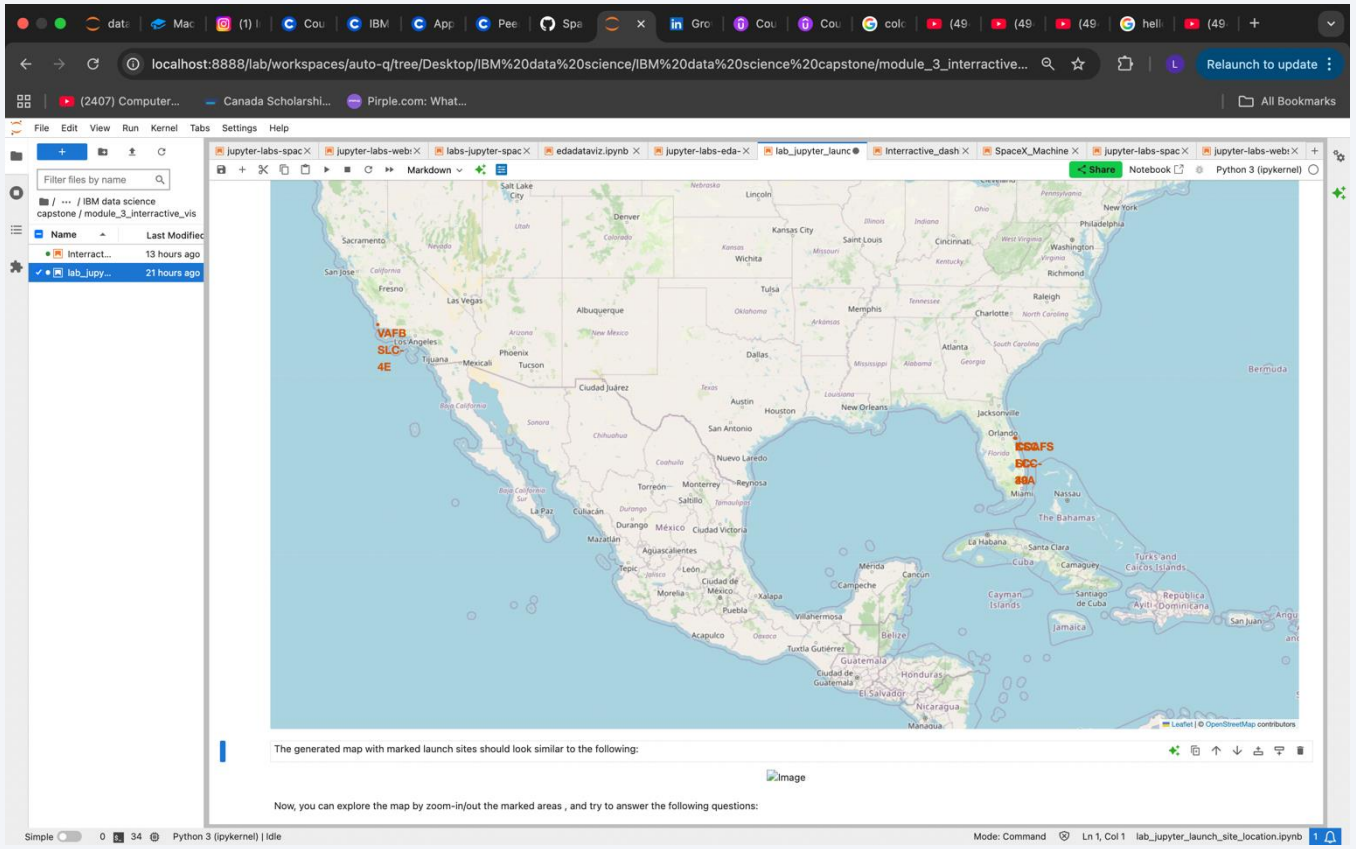
Geospatial Analysis of SpaceX Launch Sites" or "Interactive Map of Falcon 9 Launch Locations

- **Location Strategy**: All launch sites (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, etc.) are located near the **coastline**, which is crucial for safety during launch and logistics for booster recovery in the ocean.
- **Proximity to the Equator**: The sites, particularly those in Florida, are situated relatively close to the **equator** to benefit from the Earth's rotational speed, which provides an extra velocity boost to rockets and saves fuel.
- **Distance to Populated Areas**: Using distance markers and the Haversine formula, the map shows that while sites are close to support infrastructure (highways, railways), they maintain a safe distance from major cities to minimize risk in case of a launch failure.



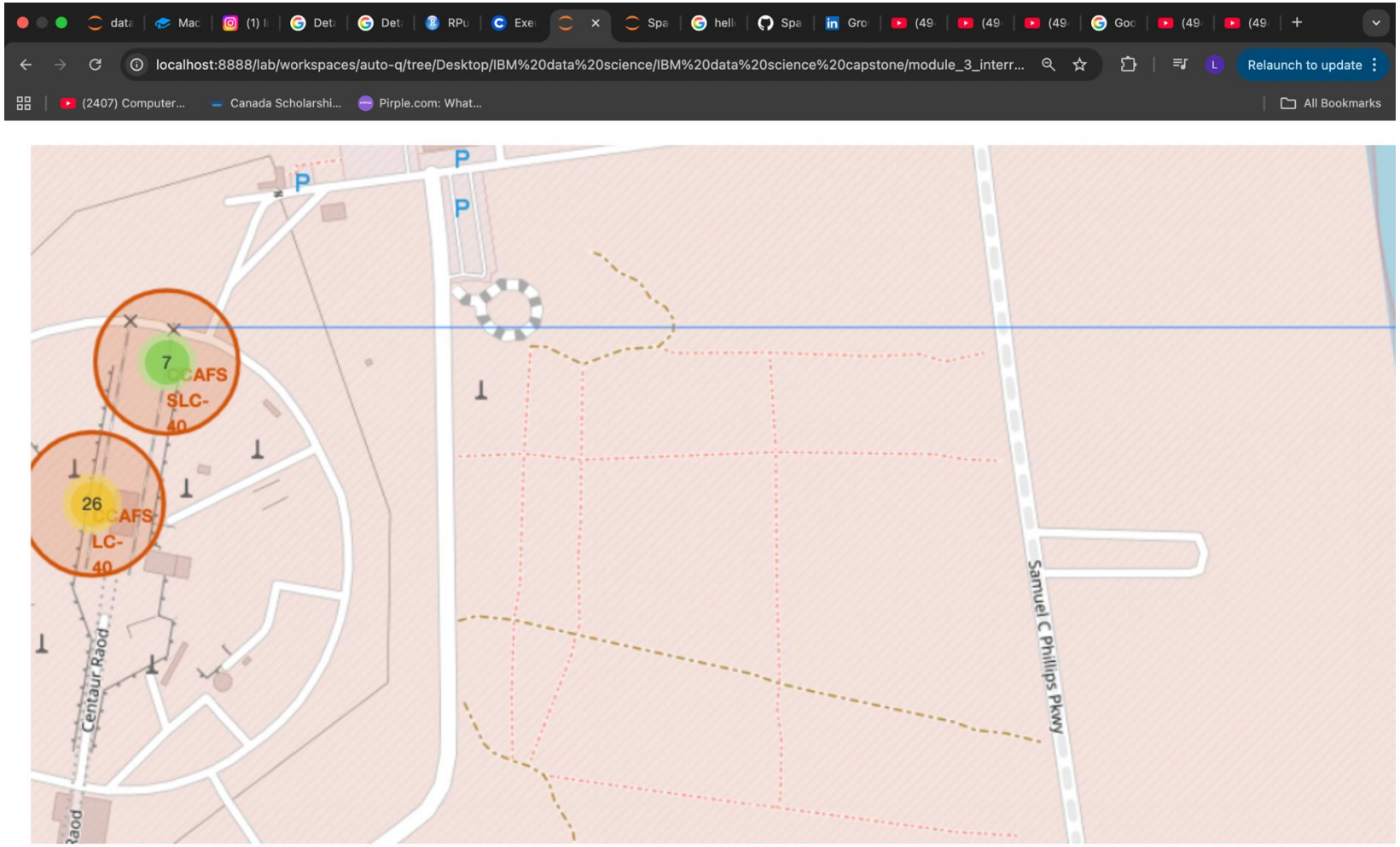
Geospatial Analysis of SpaceX Launch Sites" or "Interactive Map of Falcon 9 Launch Locations

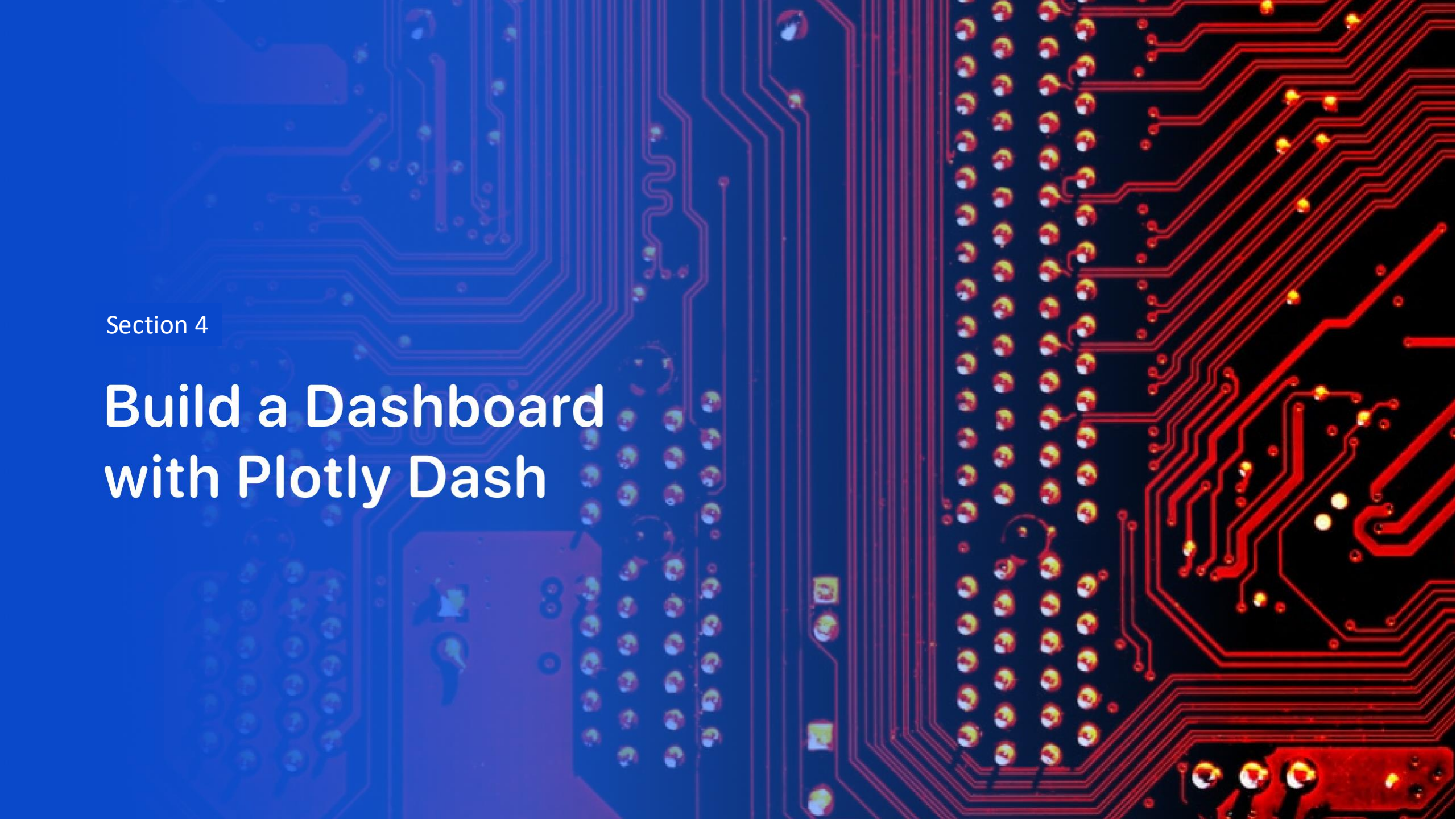
- **Proximity to the Equator**: The sites, particularly those in Florida, are situated relatively close to the **equator** to benefit from the Earth's rotational speed, which provides an extra velocity boost to rockets and saves fuel.
- **Distance to Populated Areas**: Using distance markers and the Haversine formula, the map shows that while sites are close to support infrastructure (highways, railways), they maintain a safe distance from major cities to minimize risk in case of a launch failure.



Proximity Analysis of a Selected Launch Site

- ***Location Strategy***: *All launch sites (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, etc.) are located near the **coastline**, which is crucial for safety during launch and logistics for booster recovery in the ocean.*
- ***Proximity to the Equator***: *The sites, particularly those in Florida, are situated relatively close to the **equator** to benefit from the Earth's rotational speed, which provides an extra velocity boost to rockets and saves fuel.*
- ***Distance to Populated Areas***: *Using distance markers and the Haversine formula, the map shows that while sites are close to support infrastructure (highways, railways), they maintain a safe distance from major cities to minimize risk in case of a launch failure.*
- ***Clustered Outcomes***: *By adding colored markers to the map (typically green for success and red for failure), you can visually identify which sites have relatively higher success rates. **VAFB SLC-4E and KSC LC-39A** often show a higher concentration of green markers compared to early operations at CCAFS SLC-40*




Section 4
Build a Dashboard with Plotly Dash

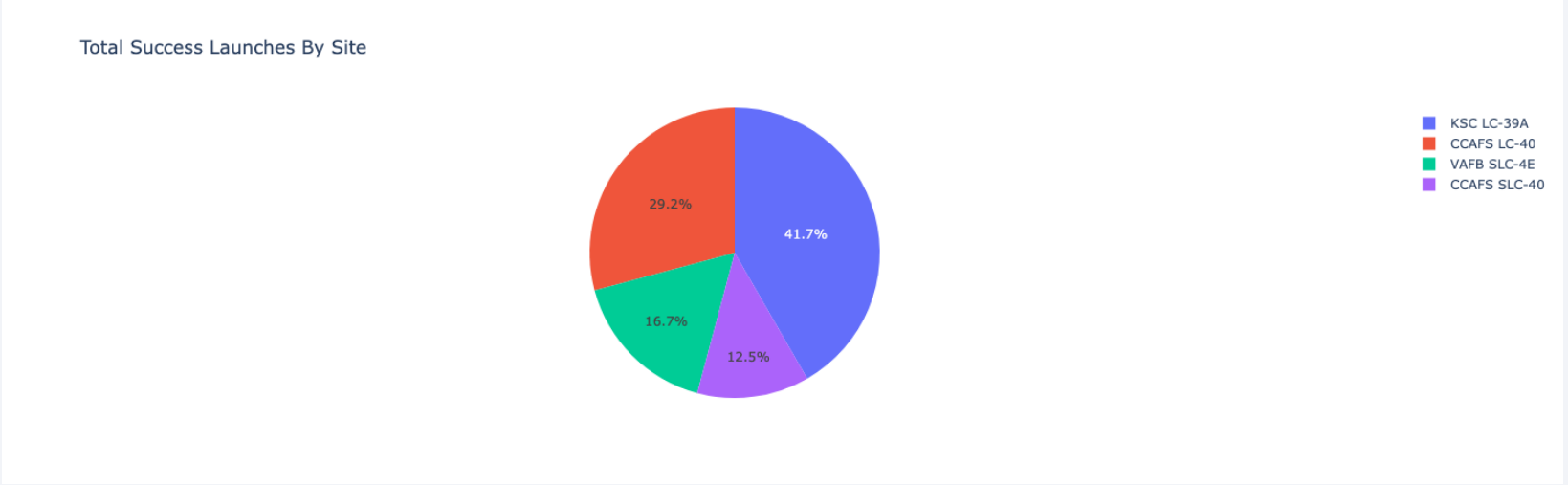
Total success launches by site

Important Elements

- **Proportions:** The pie chart uses slices to show the relative contribution of each launch site to the total count of successful launches.
- **Color-Coding:** Different colors represent the unique launch sites, making them easy to distinguish.
- **Labels:** Each slice is labeled with the launch site name and its corresponding count or percentage of total successful launches.

Key Findings

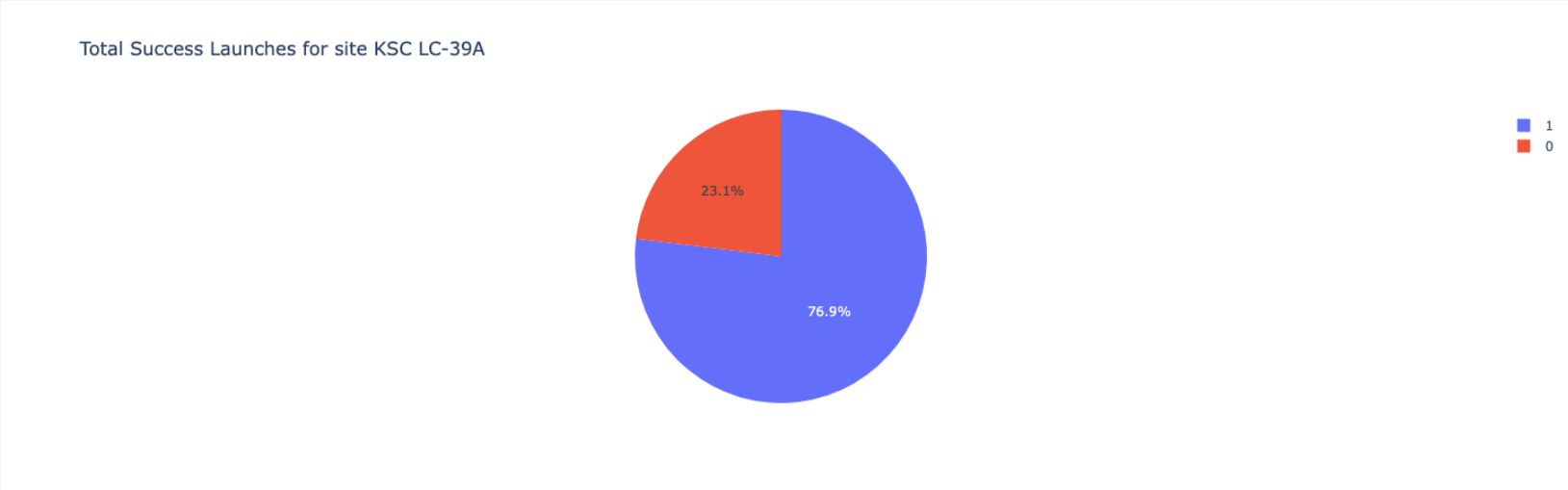
- **Dominant Site:** The data typically shows that the **KSC LC-39A** site has the highest number of successful launches overall.
- **Success vs. Failure:** When you hover over or select a specific site in the interactive dashboard version, the chart often refines to show the success-to-failure ratio for that specific location.
- **Overall Trend:** The visualization helps stakeholders quickly grasp where most successful operations originate, informing strategic business decisions for the competing "SpaceY" startup



Total success launches for site KSC LC -39A

- **Highest Success Rate Site**: The **KSC LC-39A** launch site generally exhibits the highest overall launch success ratio compared to other sites like CCAFS SLC 40 or VAFB SLC 4E.

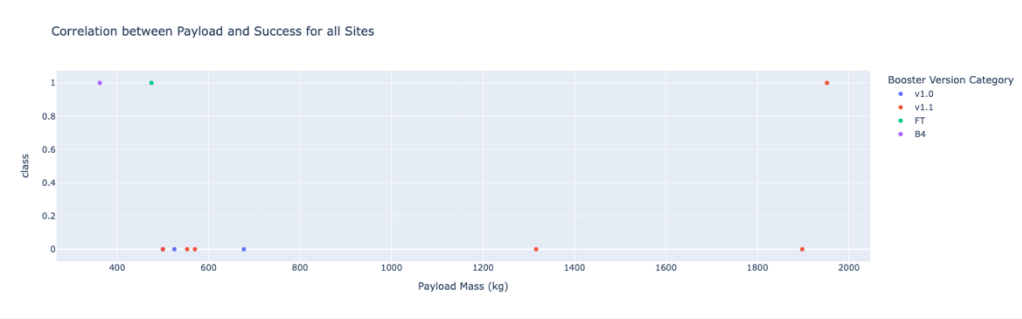
- **Success Percentage**: Depending on your specific dataset's timeframe, KSC LC-39A often shows a success rate around **76.9%** (or sometimes even 100% depending on the payload filters used in the analysis).



•**Visual Representation**: The pie chart visually breaks down the total number of launches from this site into two categories: successful (Class 1) and failed (Class 0) landings, clearly highlighting the majority of successful outcomes.
•**Insight**: This high success rate suggests that launches from KSC LC-39A are reliable, which would be a key insight for the "SpaceY" startup in making competitive bids against SpaceX.

Correlation between Payload and Success for all sites

0 kg – 2000kg



0 kg – 4000kg

0 kg – 8000kg

2000 kg – 10000kg

Important Elements

- **X-Axis**: Represents the **Payload Mass (kg)**, a continuous numerical variable.
- **Y-Axis**: Represents the **Launch Outcome** (typically 0 for failure, 1 for success), making it a binary categorical variable.
- **Interactivity**: The range slider allows dynamic filtering of the data, a key feature to demonstrate.

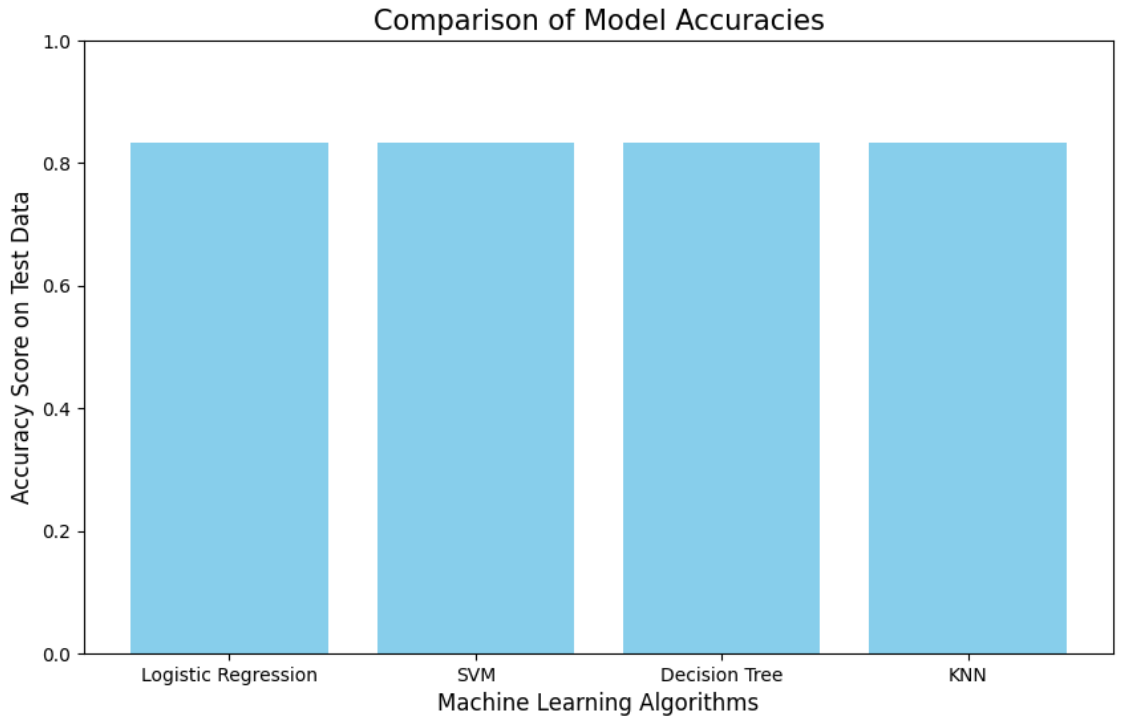
Key Findings

- **Payload Mass Correlation**: Typically, the analysis reveals that as payload mass increases, the success rate generally improves. Lighter payloads often had more early failures in the history of the Falcon 9 program.
- **Booster Version Success**: The success rate is highly dependent on the **booster version** used (e.g., early v1.0 vs. v1.1 vs. the "Flight 4" or "Block 5" boosters). The most recent or advanced versions consistently show the highest success rates, often near 100%.
- **Thresholds**: You can often identify specific payload mass thresholds (e.g., above 5000 kg) where success becomes much more probable than failure.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

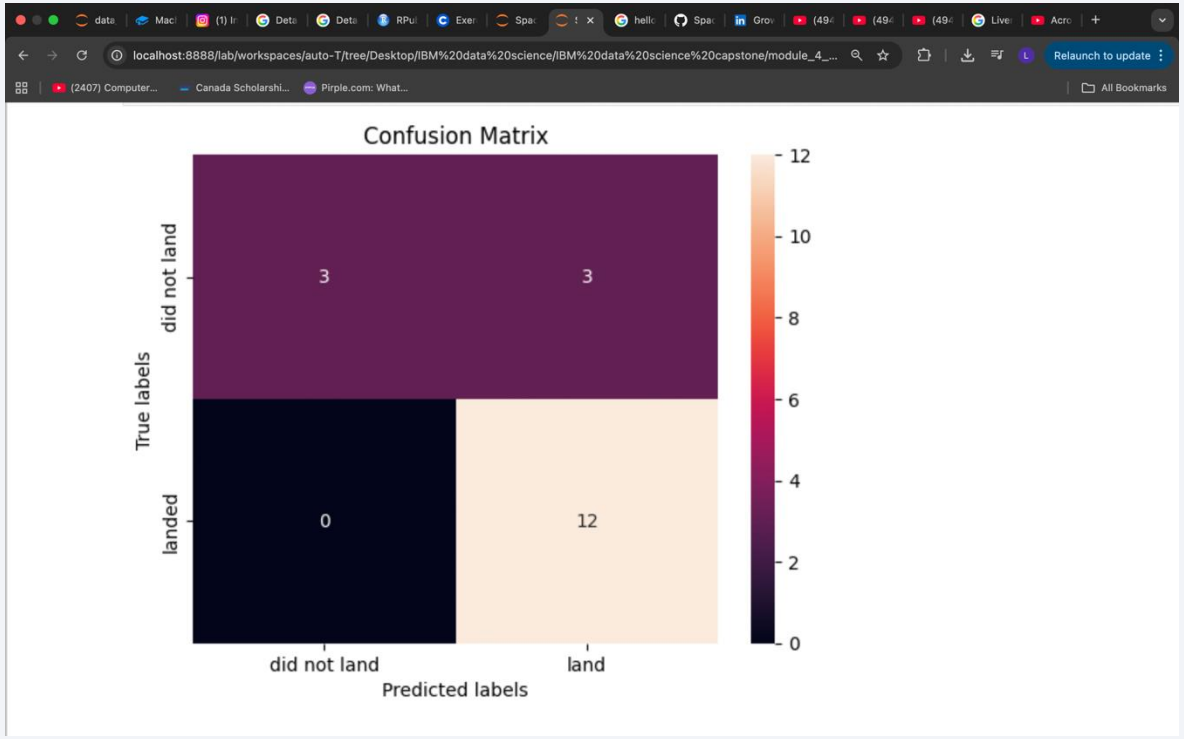


All four models equally perform well

All have an accuracy of **83.33%**

Confusion Matrix

- **Overall, the matrix shows the Decision Tree model is very good at identifying successful landings (high recall for class 1) but struggles with identifying failures (low recall for class 0), making the single false positive a key area for potential improvement.**



Conclusions

***Predictive Success**: It is possible to build machine learning models that predict landing outcomes with high accuracy (typically around **83.3%** on the test set).*

***Best Model**: While multiple models (SVM, KNN, Logistic Regression, Decision Tree) achieve similar test accuracy scores, the **Decision Tree Classifier** usually yields the highest performance when evaluated using internal cross-validation metrics.*

The ultimate conclusion for the fictional competitor, SpaceY, is that by emulating SpaceX's launch site locations and mission parameters, they can make more informed bids and rely on machine learning models to assess risk effectively.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!