IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Umeh johnpaul
14th january,
2026

# Outline

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

# Executive Summary

- **Executive Summary**
- This project leverages historical SpaceX launch data obtained from the SpaceX API to predict whether a Falcon 9 first-stage rocket will successfully land. Accurate landing predictions are critical for mission planning and cost reduction, as reusability is a key factor in SpaceX’s launch strategy.
- After collecting, cleaning, and exploring the data, several machine learning models were evaluated, including Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors. Logistic Regression achieved the best performance, with an accuracy of **83.33%** on the test dataset.
- The results demonstrate that launch-related features such as payload mass, orbit type, and launch site play a significant role in determining landing success. This model can support early risk assessment and decision-making in future launch operations.

# Introduction

SpaceX’s launch strategy depends heavily on recovering and reusing Falcon 9 first-stage boosters. While this approach significantly reduces cost, landing success is not guaranteed for every mission.

Understanding the factors that influence landing outcomes is therefore essential. This creates an opportunity to apply data science and machine learning to historical launch data in order to support better, data-driven decisions before launch.

**Based on this context, the project aims to answer four key questions**.

First, whether historical launch data can be used to reliably predict landing success. Second, which launch characteristics have the greatest impact on the outcome. Third, which machine learning model performs best for this task. And finally, how well the selected model generalizes to new, unseen data.”

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

- Data collected from the **SpaceX REST API**
- Through request API
- Unto web scrapping
- Launch, payload, orbit, and landing outcome data retrieved
- Multiple API endpoints merged into a unified dataset
- Data stored and processed using Python for analysis

# Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

- Data collected from the **SpaceX REST API**
- Through request API
- Launch, payload, orbit, and landing outcome data retrieved
- Multiple API endpoints merged into a unified dataset
- Data stored and processed using Python for analysis

https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/jupyter-labs-spacex-data-collection-api%20(1).ipynb

# Data Collection - Scraping

- Used Beautiful soup
- Imported rt library
- Extracted columns/variables from html header
- Used pandas for the dataframe.

- https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/jupyter-labs-webscraping%20(1).ipynb

# Data Wrangling

- Loaded spaceX dataset from data collection
- Identified and calculated the missing values in attribute
- Identified which columns were numerical and categorical
- Calculated the number of launches on each site
- Calculated the number and occurrence of each orbit
- Calculated the number and occurence of mission outcome of the orbits
- Created a landing outcome label from Outcome column
- https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/labs_jupyter_spacex_Data_wrangling.ipynb

# EDA with Data Visualization

- Few charts were plotted, scatter plot, bar graph, and line chart
- I used scatter plots to the relationship between Flightnumber and LunchSite. The relationship between payload Mass and LunchSite, between Flightnumber and orbit type
- I plotted bar to visualize the relationship between success rate of each orbit type
- Then used line chart to Visualize the launch success yearly trend
- https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/edadataviz%20(1).ipynb

# EDA with SQL

- I Displayed the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

# Build an Interactive Map with Folium

To better understand the geographic patterns of Falcon 9 launches and landings, an interactive map was created using the Folium library.

Markers were used to represent SpaceX launch sites, while circles were added to visualize landing success and failure outcomes. Lines were included to measure distances between launch sites and nearby infrastructure such as coastlines or cities. Interactive popups were added to provide additional contextual information for each location.

https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/lab_jupyter_launch_site_location%20(1).ipynb

# Build a Dashboard with Plotly Dash

- Purpose of the Plots and Interactions
- The **pie chart** was added to quickly summarize launch success rates and compare performance across different launch sites.
- The **scatter plot** helps analyze the correlation between payload mass and launch success, revealing trends that may not be visible in aggregate statistics.
- The **dropdown menu** improves usability by allowing focused analysis on individual launch sites.
- The **payload range slider** enhances interactivity and exploratory analysis by enabling users to isolate specific payload ranges and observe how they affect launch outcomes

# Predictive Analysis (Classification)

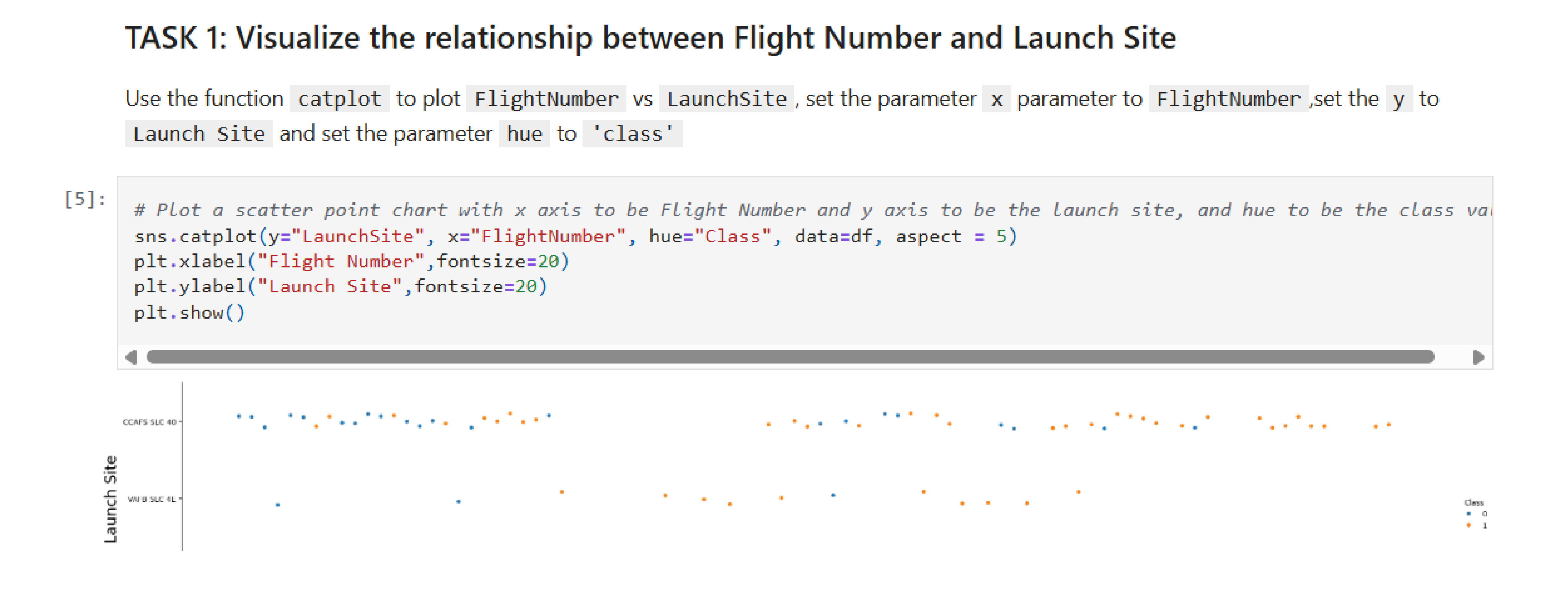
- Problem Definition
- ↓
- Data Cleaning & Preprocessing
- ↓
- Feature Engineering & Scaling
- ↓
- Train-Test Split
- ↓
- Model Training (LR | SVM | DT | KNN)
- ↓
- Model Evaluation (Accuracy | Confusion Matrix | Recall)
- ↓
- Hyperparameter Tuning
- ↓
- Model Comparison
- ↓
- Best Model Selection
- https://github.com/Johnpaul10j/Datascience-Capstone-project/blob/a270627b499ebf43cda007c762bdde6db877a550/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

# Results

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

Section 2

# Insights drawn from EDA



## TASK 1: Visualize the relationship between Flight Number and Launch Site

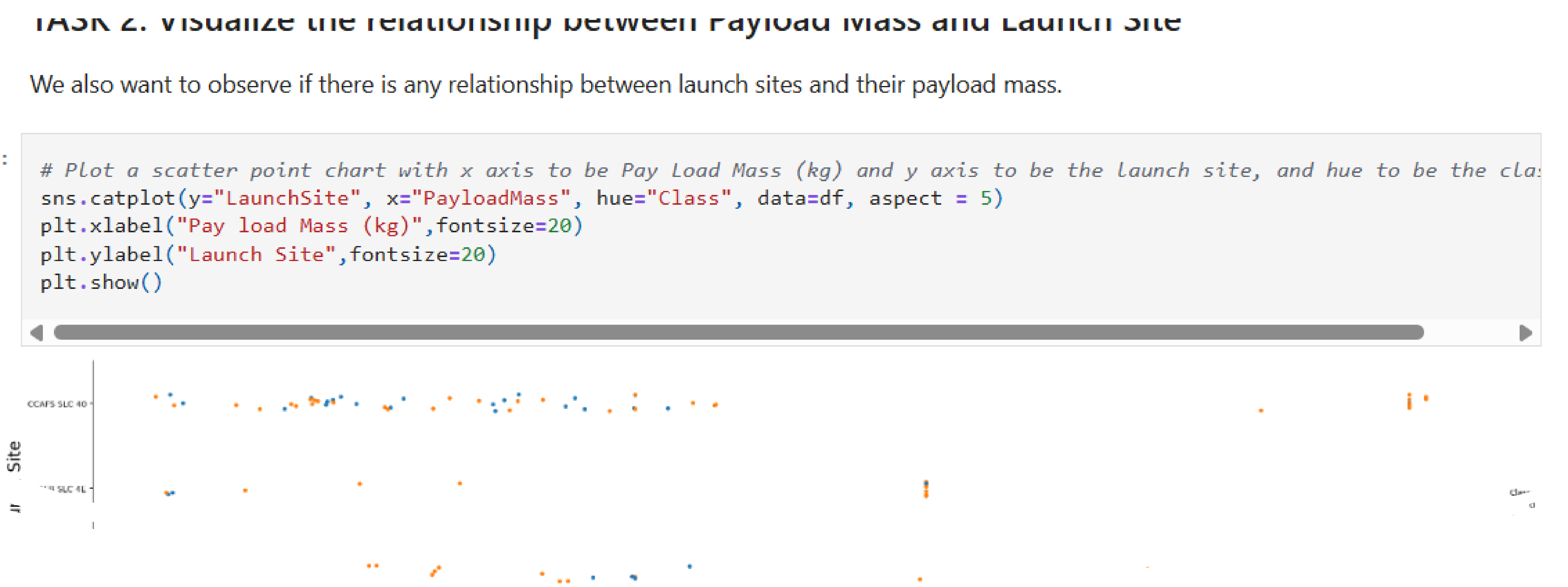
Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber` ,set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

[5]:

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class val
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

# Flight Number vs. Launch Site

The scatter plot shows that as SpaceX gained more launch experience over time, launches expanded across more sites and landing success rates improved significantly.



## TASK 2: Visualize the relationship between Payload Mass and Launch Site

We also want to observe if there is any relationship between launch sites and their payload mass.

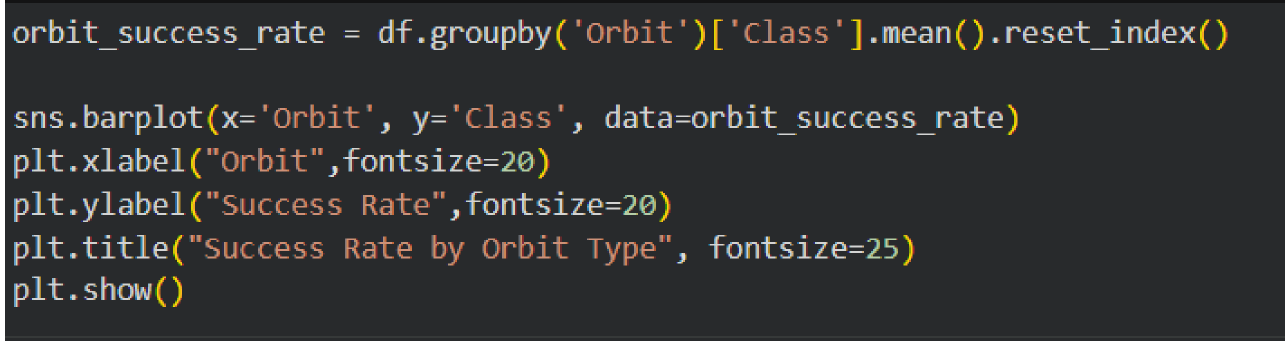
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the cla
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

# Payload vs. Launch Site

I observed that for the VAFB-SLC launch site, there were no rockets launched with heavy payload masses (greater than 10,000 kg). This task helped understand how payload mass varies across different launch sites and its potential impact on launch success.

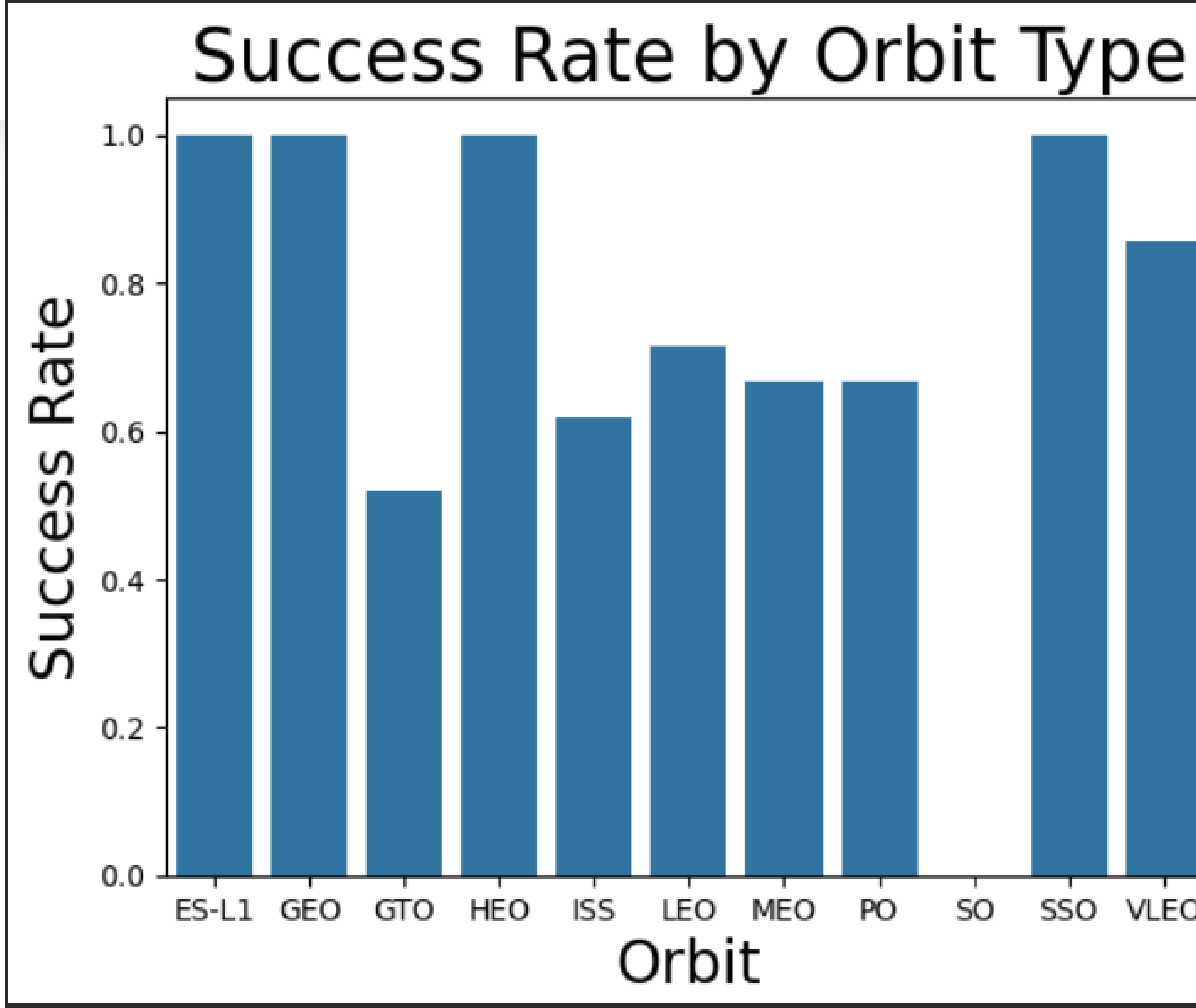
# Success Rate vs. Orbit Type

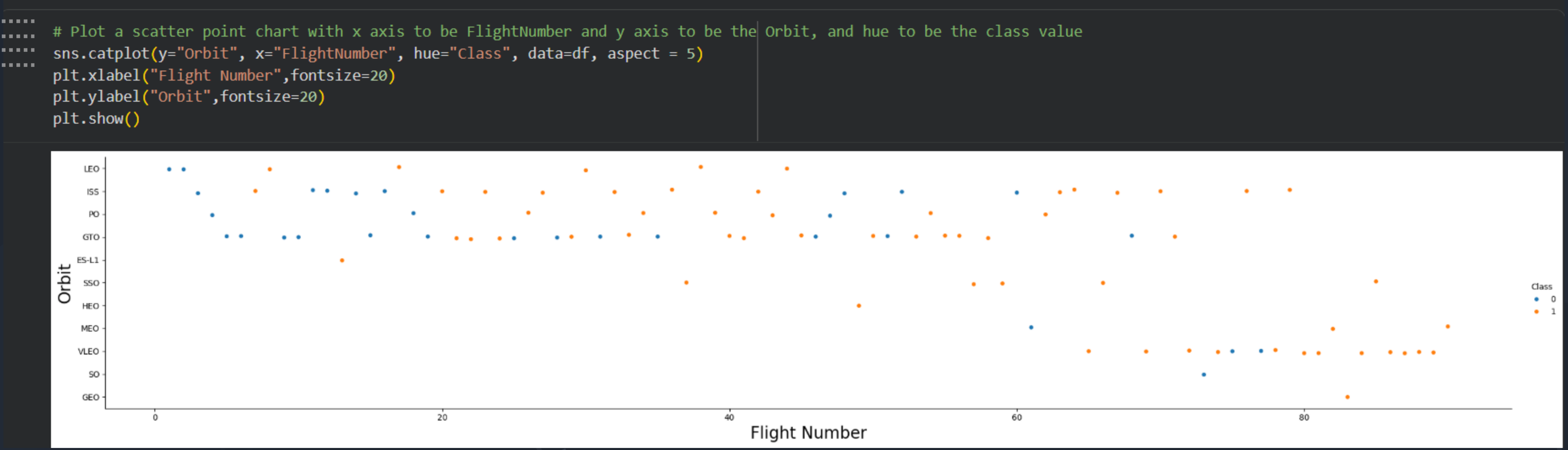
- Landing success rates vary significantly by orbit type, with high-energy missions such as GTO showing lower success compared to other orbits.



```
orbit_success_rate = df.groupby('Orbit')['Class'].mean().reset_index()

sns.barplot(x='Orbit', y='Class', data=orbit_success_rate)
plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.title("Success Rate by Orbit Type", fontsize=25)
plt.show()
```





```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

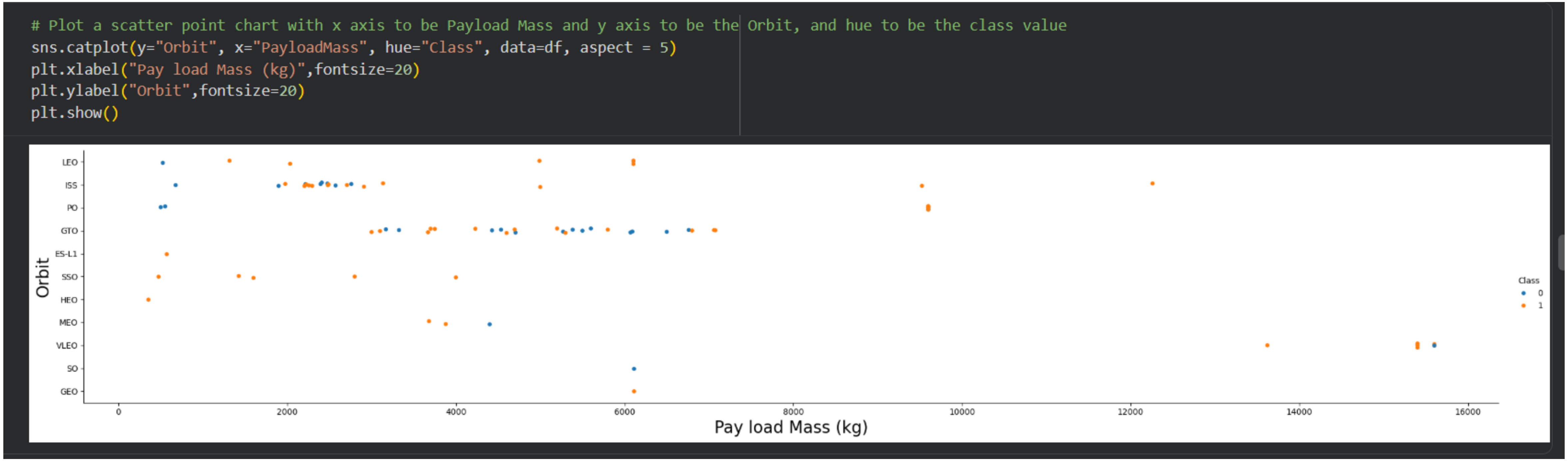
# Flight Number vs. Orbit Type

- As flight experience increased, SpaceX expanded into more complex orbit types while achieving higher landing success rates.”

# Payload vs. Orbit Type

- High Success for Heavy Payloads in Specific Orbits: For Polar, LEO (Low Earth Orbit), and ISS (International Space Station) orbits, successful landings or positive landing rates are more prevalent even with heavy payloads.

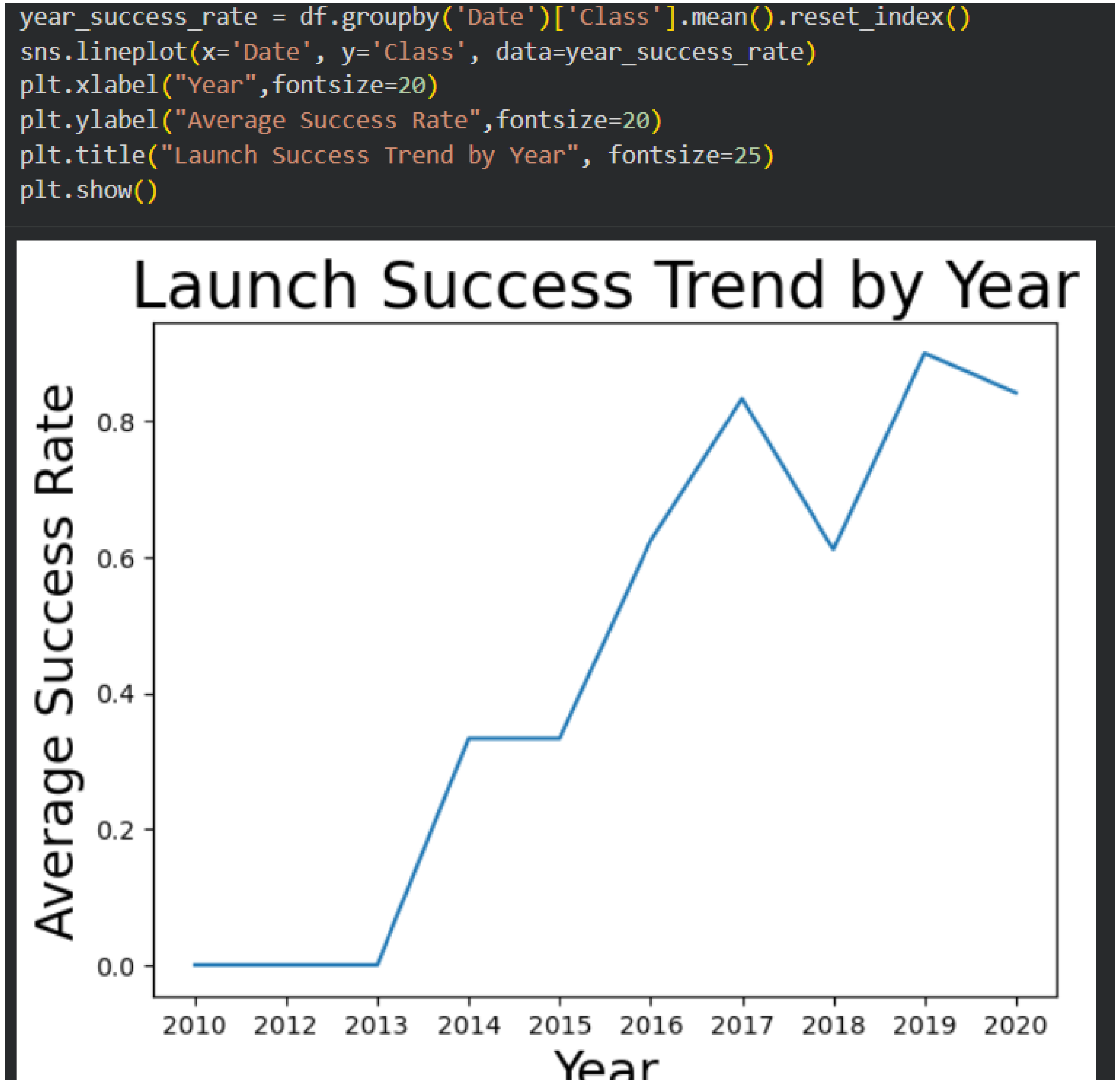
- Difficulty in Distinguishing Success for GTO with Heavy Payloads: For GTO (Geostationary Transfer Orbit), it's more challenging to differentiate between successful and unsuccessful landings when dealing with heavy payloads, as both outcomes are observed. This suggests that while heavy payloads might succeed in GTO, the success is not as consistently high or clearly separable as in other orbits.



```
# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

# Launch Success Yearly Trend

- Increasing Success Rate: The launch success rate for Falcon 9 first stage landings consistently increased since 2013 and continued to rise until 2020. This indicates a significant improvement in SpaceX's launch and landing capabilities over this period.



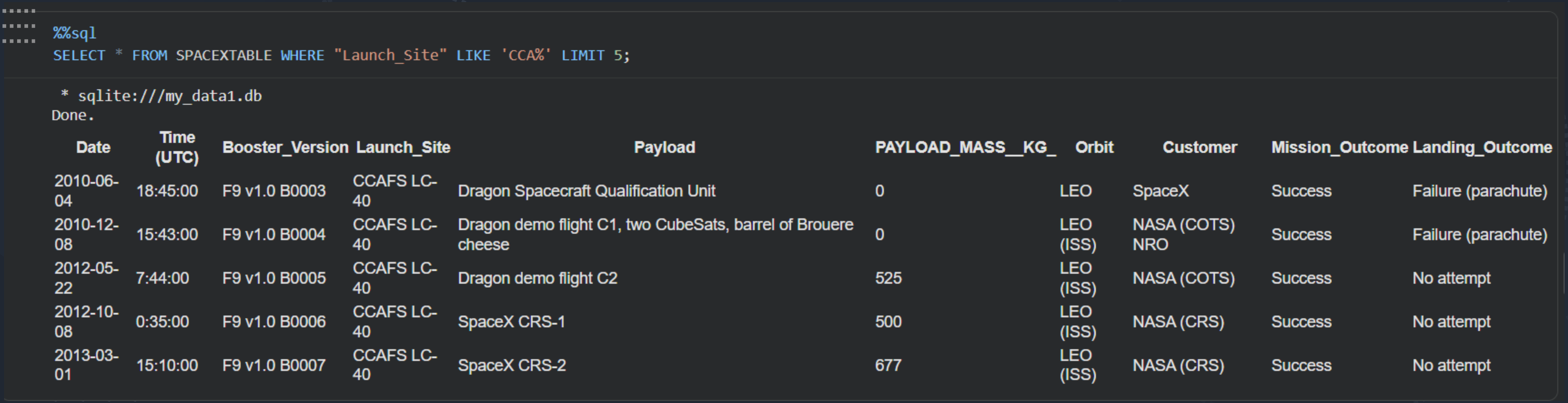
```
year_success_rate = df.groupby('Date')['Class'].mean().reset_index()
sns.lineplot(x='Date', y='Class', data=year_success_rate)
plt.xlabel("Year",fontsize=20)
plt.ylabel("Average Success Rate",fontsize=20)
plt.title("Launch Success Trend by Year", fontsize=25)
plt.show()
```

# All Launch Site Names

- Findings: The query successfully retrieved the unique launch sites from the SPACEXTABLE. The distinct launch sites identified are:
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

# Launch Site Names Begin with 'CCA'

- These records include details about the launch date, time, booster version, payload, orbit, customer, mission outcome, and landing outcome for the earliest missions from these sites.



```
%%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 mass carried by boosters launched for NASA (CRS) is 45596 kg.

# Average Payload Mass by F9 v1.1

The average payload mass carried by the F9 v1.1 booster version is 2928.4 kg

# First Successful Ground Landing Date

- The date when the first successful landing outcome on a ground pad was achieved is 2015-12-22.

# Successful Drone Ship Landing with Payload between 4000 and 6000

The boosters that meet these criteria are:

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

# Total Number of Successful and Failure Mission Outcomes

The findings indicate the following mission outcomes:

- Failure (in flight): 1 mission
- Success: 98 missions
- Success : 1 mission (Note: This appears to be a success outcome with a trailing space, indicating a potential data cleaning opportunity)
- Success (payload status unclear): 1 mission

# Boosters Carried Maximum Payload

The booster versions that have carried the maximum payload mass are:

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

The results show:

In January 2015, booster F9 v1.1 B1012 from CCAFS LC-40 had a Failure (drone ship) landing outcome.

In April 2015, booster F9 v1.1 B1015 from CCAFS LC-40 also had a Failure (drone ship) landing outcome.

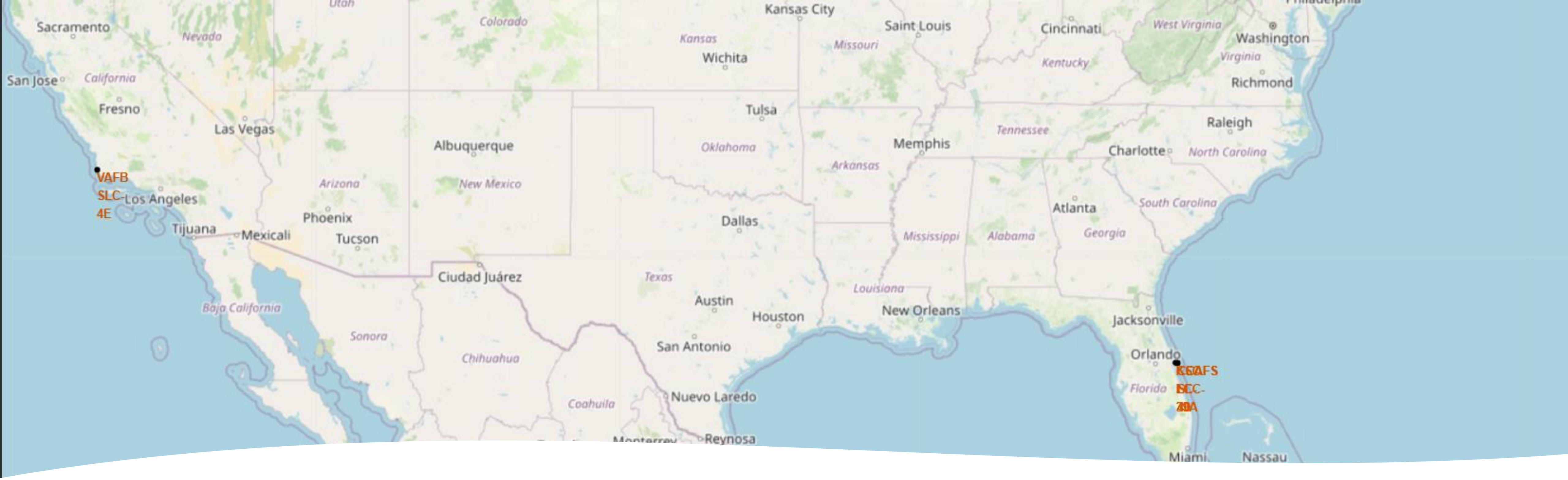
# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Here are the landing outcomes and their counts, in descending order:

- '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

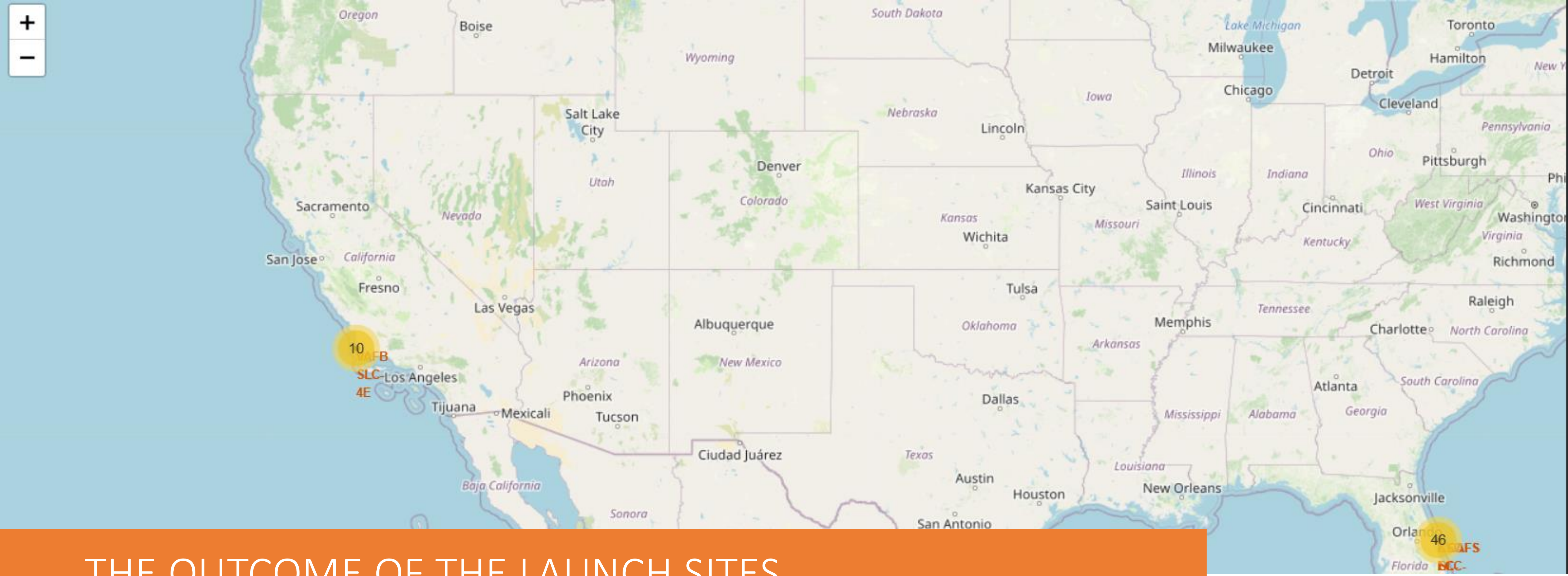
Section 3

# Launch Sites Proximities Analysis



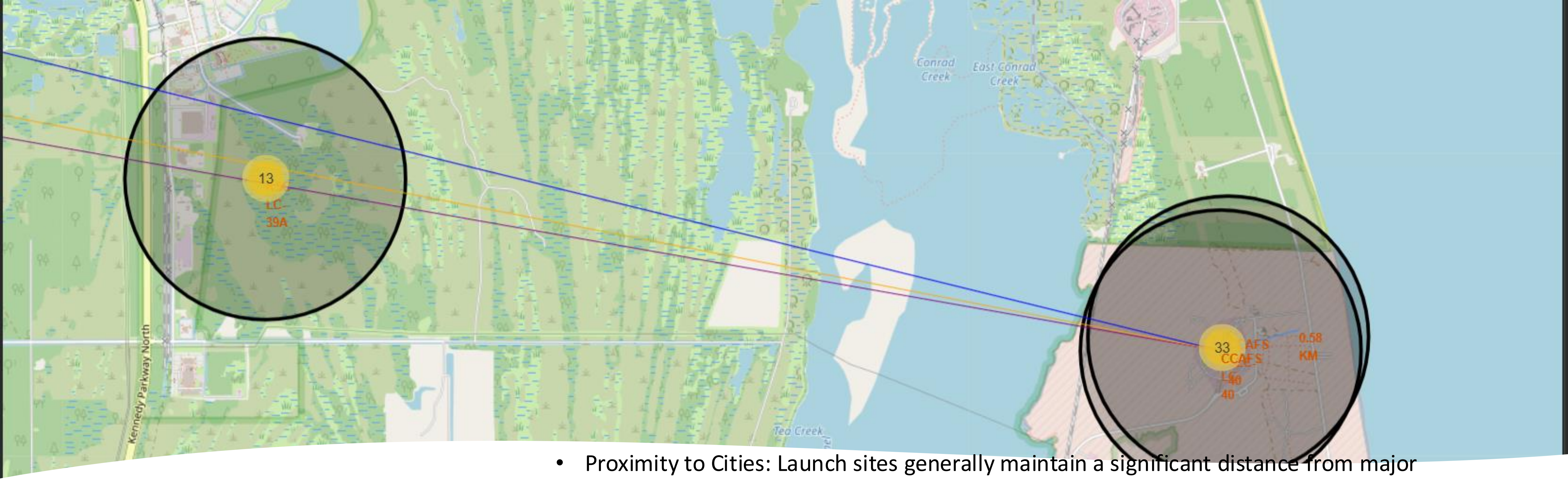
# LOCATIONS OF ALL LAUNCH SITES

- This is evident from their geographical locations on the map (Florida for the CCAFS and KSC sites, and California for VAFB SLC-4E), which are typical locations for launch facilities due to safety and logistical reasons (e.g., eastward launches over the ocean)



# THE OUTCOME OF THE LAUNCH SITES

- KSC LC-39A (Kennedy Space Center Launch Complex 39A) appears to have a high concentration of green markers, indicating a good success rate.
- VAFB SLC-4E (Vandenberg Space Force Base Space Launch Complex 4E) shows a mix of green and red markers, suggesting a moderate success rate with some failures.
- CCAFS LC-40 (Cape Canaveral Space Force Station Space Launch Complex 40) and CCAFS SLC-40 (Cape Canaveral Space Force Station Space Launch Complex 40) also show varying degrees of success and failure, which can be further analyzed by counting the green and red markers within their clusters.



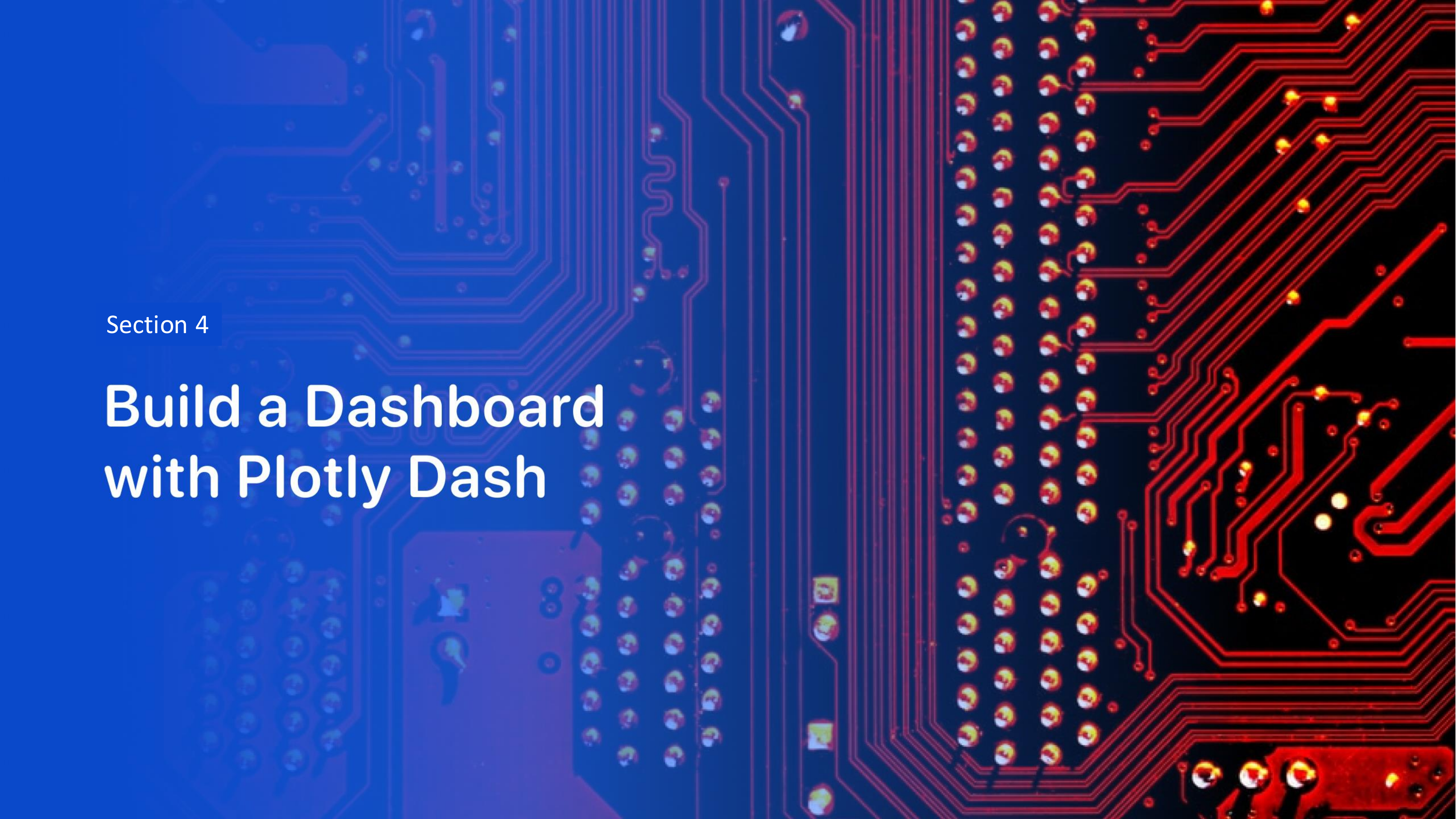
# The Proximities of Launch Sites

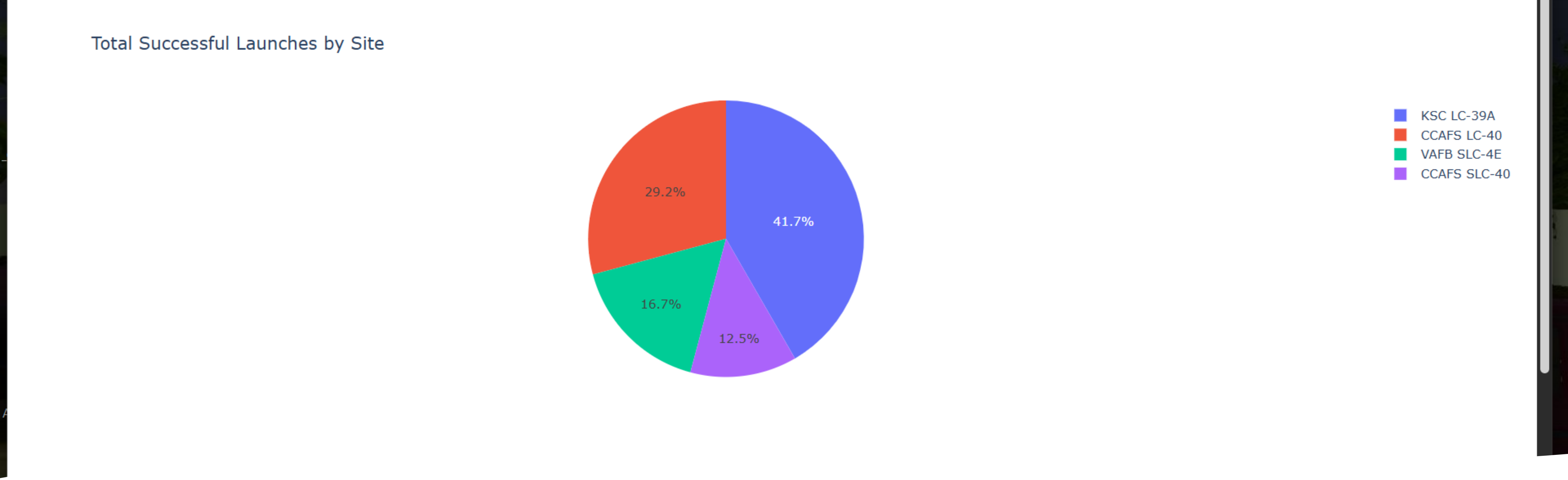
- Proximity to Cities: Launch sites generally maintain a significant distance from major population centers like cities. For example, the CCAFS LC-40 launch site is approximately 23.25 KM from Titusville. This distance is crucial for safety reasons, minimizing risks to urban populations in case of launch failures or falling debris.

- Proximity to Highways: Launch sites are located relatively close to highways. For the CCAFS LC-40 site, the distance to a nearby highway (US-1) was around 22.74 KM. This proximity to highways is essential for logistics, allowing for the efficient transportation of rocket components, fuel, equipment, and personnel to and from the launch complex.

Section 4

# Build a Dashboard with Plotly Dash



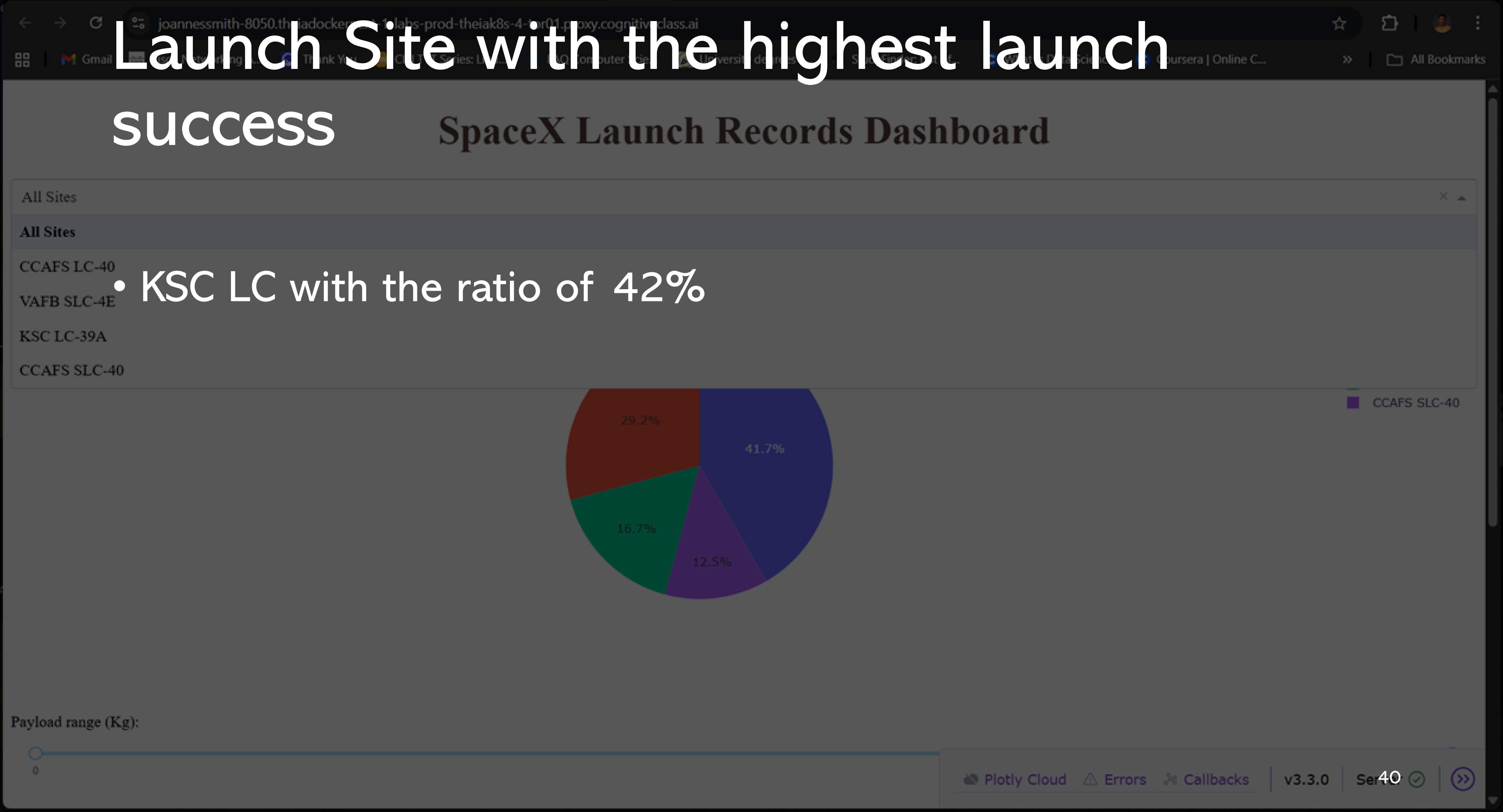


# Launch Success Count for all sites

- **Pie Chart – Launch Success Distribution**
- Displays the total number of successful launches for all SpaceX launch sites.
- When a specific launch site is selected, the pie chart shows the distribution of **successful vs failed launches** for that site.

# Launch Site with the highest launch success

- KSC LC with the ratio of 42%



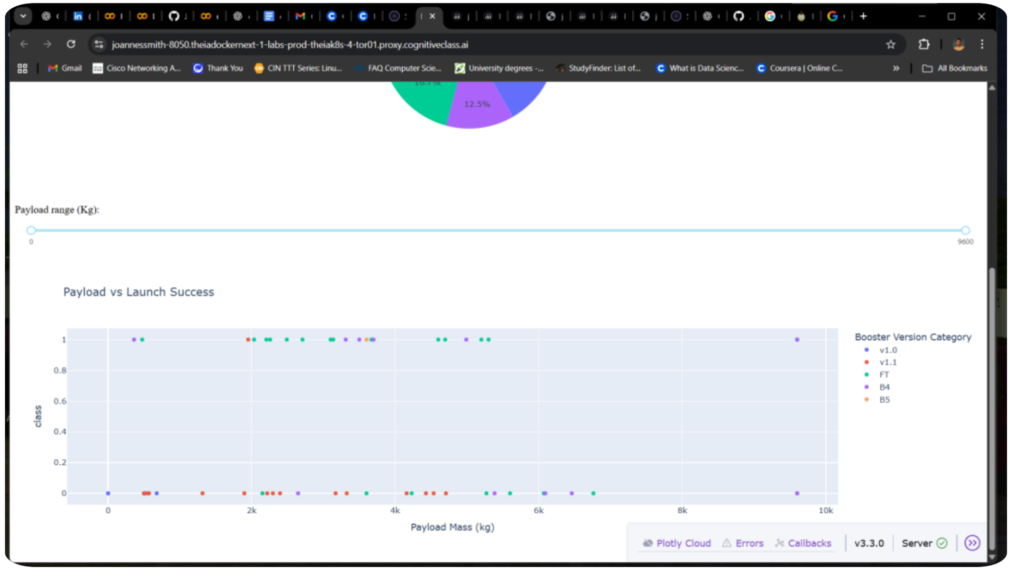
SpaceX Launch Records Dashboard

All Sites

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

Payload range (Kg):

# The Relationship between Payload vs Launch Outcome



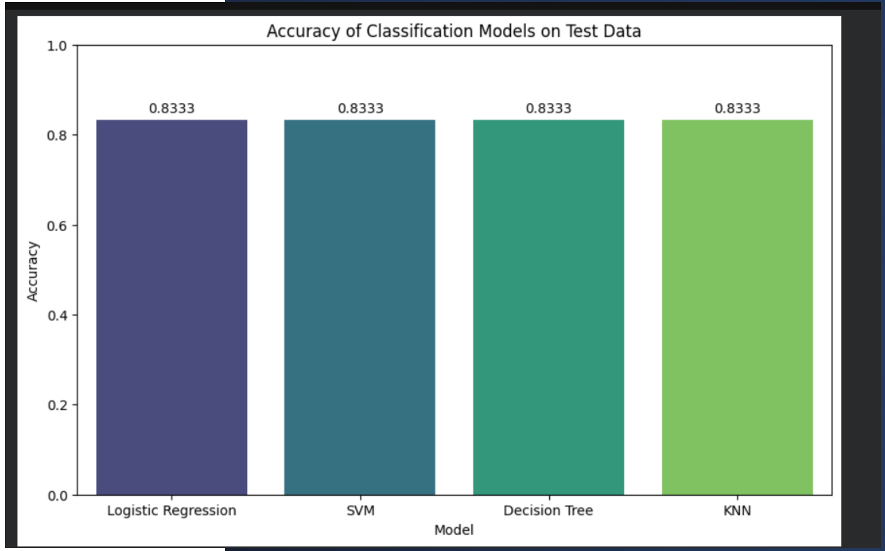
- **Scatter Plot – Payload vs Launch Success**
- Shows the relationship between **payload mass (kg)** and **launch success (0 = failure, 1 = success)**.
- Data points are color-coded by **booster version category** to provide additional insights into performance variations.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- Data Analysis Key Findings
- All four classification models—Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors—achieved an identical test accuracy of 0.8333.
- This equal performance indicates that, based on the X_test and Y_test split, no single model outperformed the others in terms of accuracy.
- The performance was visualized using a bar chart, clearly showing the uniform accuracy across all evaluated models.

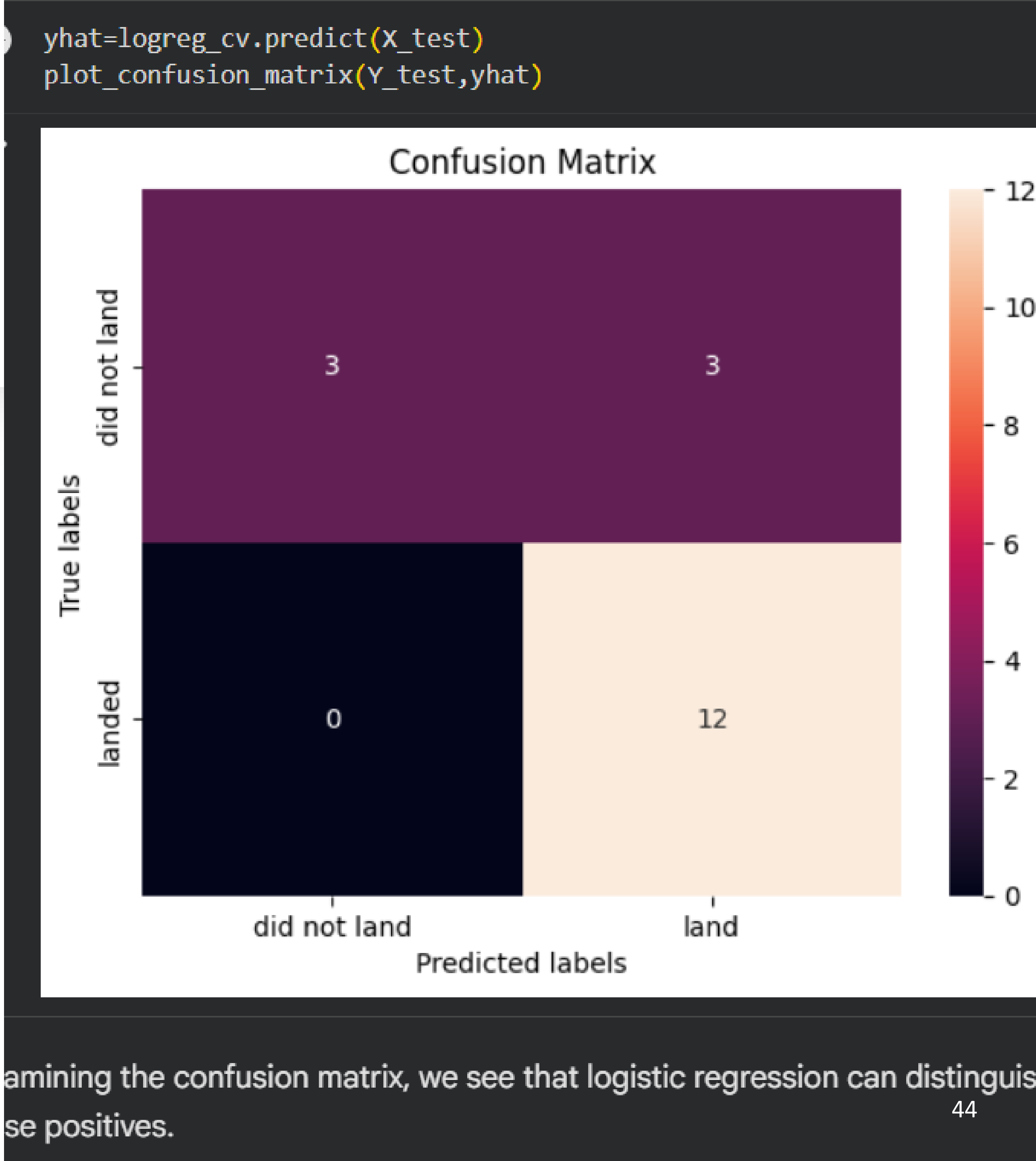


# Confusion Matrix

- Which model(s) performed best on the test data? All four models—Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors—performed equally best, each achieving a test accuracy of 0.8333.

- They are considered equally best because all models achieved the exact same highest accuracy score on the test data, indicating that for this specific dataset and split, their predictive power is identical based on the accuracy metric.

- But here is a Logistic Regression confusion matrix



```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

amining the confusion matrix, we see that logistic regression can distinguis
se positives.

# Conclusions

Certainly! Here are 5 key insights and findings from this machine learning prediction lab:

- Objective Achievement: The primary goal of building a machine learning pipeline to predict Falcon 9 first-stage landings was successfully addressed, highlighting the economic importance of reusability for Space X.
- Comprehensive Data Preparation: The dataset underwent crucial preprocessing steps, including the creation of a Class column for the target variable Y, standardization of features in X, and splitting into training and test sets (X_train, X_test, Y_train, Y_test).
- Multiple Model Evaluation: Four different classification algorithms—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN)—were implemented and thoroughly tuned using GridSearchCV with cross-validation (cv=10) to find optimal hyperparameters.
- Consistent Test Accuracy: Despite varying best validation scores during hyperparameter tuning, all four models ultimately achieved an identical test accuracy of approximately 83.33%. This suggests that, based solely on this metric, no single model definitively outperforms the others on this particular test set.
- Further Refinement Opportunities: While the models show good predictive power, a deeper dive into other evaluation metrics (e.g., precision, recall, F1-score) and their respective confusion matrices could reveal subtle differences in performance, especially concerning false positives or false negatives, which might be critical depending on the cost of misclassification in a real-world scenario. Addressing warnings, such as the one encountered during GridSearchCV for the Decision Tree, could also lead to more robust models

# Appendix

- The calculations for precision, recall, and F1-score have been successfully executed. Here are the results:
- Model Performance Metrics on Test Data:
- Logistic Regression:
- Precision: 0.8000
- Recall: 1.0000
- F1-Score: 0.8889
- SVM:
- Precision: 0.8000
- Recall: 1.0000
- F1-Score: 0.8889
- Decision Tree:
- Precision: 0.8000
- Recall: 1.0000
- F1-Score: 0.8889
- KNN:
- Precision: 0.8000
- Recall: 1.0000
- F1-Score: 0.8889
- Key Observations:
- Identical Performance: Similar to the accuracy scores, all four models exhibit the exact same precision, recall, and F1-score on this test dataset. This reinforces the idea that, based on these metrics, there's no single 'best' model among them for this specific test set.
- High Recall (1.0000): A recall of 1.0000 means that all actual positive cases (rockets that did land successfully) were correctly identified by every model. There were no false negatives.
- Moderate Precision (0.8000): A precision of 0.8000 indicates that when a model predicted a rocket would land, it was correct 80% of the time. This implies there were some false positives (cases where the model predicted a landing, but it did not occur).
- Strong F1-Score (0.8889): The F1-score is the harmonic mean of precision and recall. A score close to 1 suggests a good balance between identifying all positive cases and not incorrectly classifying too many negative cases as positive. Given the perfect recall, the F1-score is largely influenced by the precision.

# Thank you!