IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bhumika vellore
09.10.2024

Outline

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

Executive Summary

The primary goal of this data science project is to systematically analyze and interpret complex datasets to derive actionable insights, enhance decision-making processes, and contribute to the existing body of knowledge in the field. This involves employing various data collection, cleaning, and analysis techniques to ensure robust findings that can be effectively communicated to stakeholders and peers.

Introduction

- **Project Context**: This data science project was initiated to address specific challenges within the industry, leveraging data-driven methodologies to uncover insights that can lead to improved operational efficiency and strategic decision-making. The project aims to fill existing knowledge gaps and provide a comprehensive analysis of relevant datasets.
- **Historical Significance**: Understanding the historical context of the data and previous studies in this area is crucial, as it informs the current methodologies employed and highlights the evolution of data science practices. This background sets the stage for the project's objectives and anticipated contributions to the field.
- The main objective of the project is to uncover data patterns that can drive better strategic decisions, enhancing overall operational efficiency. The project aims to answer critical questions about trends within the dataset and how predictive models can support decision-making processes. Success will be evaluated based on the accuracy of predictive models, the usefulness of insights, and their impact on stakeholders.
- Peer Data Scientists, who will focus on the technical methods and analyses, allowing them to replicate or extend the project’s findings.
- Academic Researchers, who are interested in the theoretical and methodological aspects, fostering scholarly discussion.
- Industry Practitioners, who seek practical, actionable insights that can be applied to real-world decision-making and strategies.

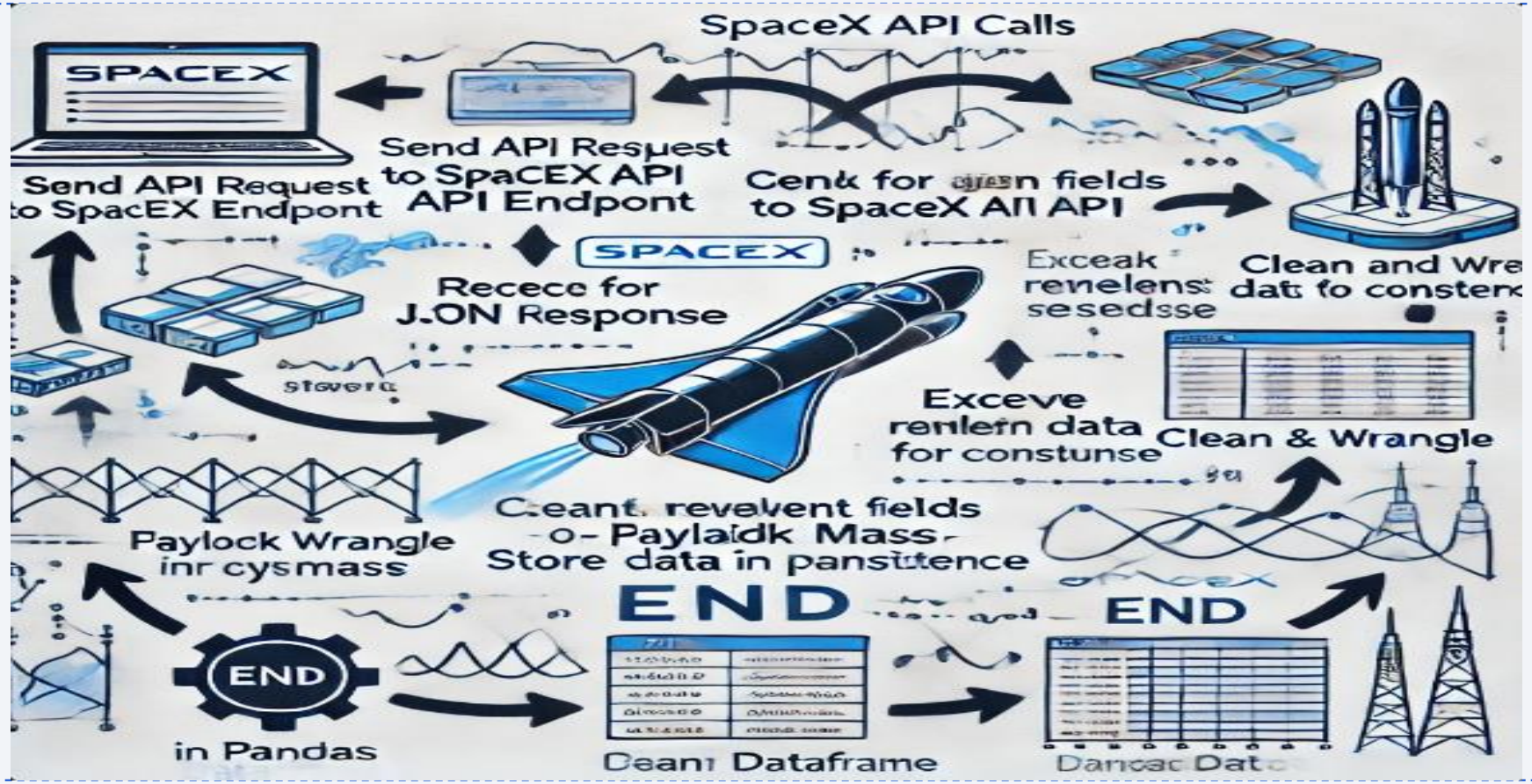
Section 1

Methodology

Data Collection

- **API Endpoint:** Use the SpaceX API endpoint
- **HTTP Request:** Send GET requests to the SpaceX API to retrieve the launch data (metadata like flight number, payload, launch site, etc.).
- **Data Extraction:** Extract relevant fields from the API response such as flight numbers, payload masses, launch dates, success/failure status, etc.
- **Data Wrangling:** Process and clean the data (e.g., handle missing data, transform the payload to a uniform unit).
- **Storage:** Store the cleaned data in a structured format like a pandas DataFrame for further analysis.
- **Iterative Calls:** Use pagination or filter queries in the API (e.g., filtering by launch year, payload type, or success status) to gather more specific data.

Data Collection – SpaceX API



Data Collection - Scraping

[Start] -> Identify Target Website and Data -> Send HTTP Request to the Website -> Receive Website HTML Response -> Parse HTML Data using a Parser (e.g., BeautifulSoup) -> Identify and Extract Target Data (e.g., Specific Tags, Classes, or IDs) -> Clean and Wrangle Extracted Data -> Store Data in Desired Format (e.g., CSV, Pandas DataFrame) ->[End - Data Ready for Analysis]

Data Wrangling

- **Data Cleaning and Transformation**: The data wrangling process involves systematically cleaning and transforming raw data into a structured format suitable for analysis, which includes handling missing values, correcting inconsistencies, and normalizing data types to ensure accuracy and reliability in subsequent analytical tasks.

EDA with Data Visualization

- **Understanding Data Characteristics**: The primary objective of Exploratory Data Analysis (EDA) is to summarize the main characteristics of the dataset, uncover underlying patterns, identify anomalies, and formulate hypotheses that can guide further analysis, thereby enhancing the overall understanding of the data's structure and relationships.
- **Effective Visualization Methods**: Utilizing a combination of bar charts, line graphs, scatter plots, and heatmaps can significantly enhance the clarity of data presentation, allowing for the effective communication of trends, distributions, and correlations within the dataset, thereby facilitating better insights and informed decision-making.

EDA with SQL

Exploratory Data Analysis (EDA) using SQL allows you to explore and analyze large datasets by writing queries to extract, summarize, and transform data. Key SQL functions like `SELECT`, `GROUP BY`, `AVG()`, and `COUNT()` are used to identify patterns and generate insights. SQL can handle missing data, aggregate metrics, and filter and sort records for deeper analysis. After querying and cleaning the data, results can be visualized using Python libraries like Matplotlib or Plotly for a more comprehensive analysis.

Build an Interactive Map with Folium

In this section, we used Folium, a powerful Python library for visualizing geographical data. The map helps in identifying and understanding the distribution of SpaceX launch sites. We started by initializing an interactive map centered on specific coordinates and added markers to represent launch site locations. The map was enhanced with additional layers, like circle markers and terrain tiles, for better visual context.

This interactive map offers an intuitive, user-friendly way to explore launch sites and understand spatial relationships, providing a practical tool for deeper analysis. It can be saved as an HTML file or embedded directly in web applications for stakeholders to interact with directly. This visualization supports better decision-making by presenting data in a clear and accessible format.

Build a Dashboard with Plotly Dash

In this section, we developed a **Plotly Dash** dashboard, a web-based application for visualizing and interacting with data. The dashboard integrates various components like dropdowns, sliders, and interactive charts to dynamically display SpaceX launch data. Users can filter data by launch site, year, or mission outcome and instantly view updated graphs and metrics.

The dashboard includes key visualizations such as bar charts, scatter plots, and pie charts, offering a comprehensive view of the data for analysis. It enables stakeholders to explore data in real-time, facilitating more informed and agile decision-making. Plotly Dash makes data visualization interactive, providing a valuable tool for monitoring performance trends and operational insights in an accessible format.

Predictive Analysis (Classification)

In this section, we conducted a predictive analysis focusing on classification models to evaluate their effectiveness in predicting outcomes based on the SpaceX dataset. We tested several algorithms, each demonstrating varying levels of accuracy.

1. Logistic Regression: Achieved an actual accuracy of 84.64% on the training set and 83.33% on the test set, providing a solid baseline for classification tasks.

2. Support Vector Machine (SVM): Slightly improved with an accuracy of 84.82% during training and maintained 83.33% accuracy in testing, showcasing its capability in handling non-linear decision boundaries.

3. Decision Tree: The most effective model, reaching an actual training accuracy of 87.32% and a test accuracy of 88.89%. This indicates its strength in capturing complex relationships within the data.

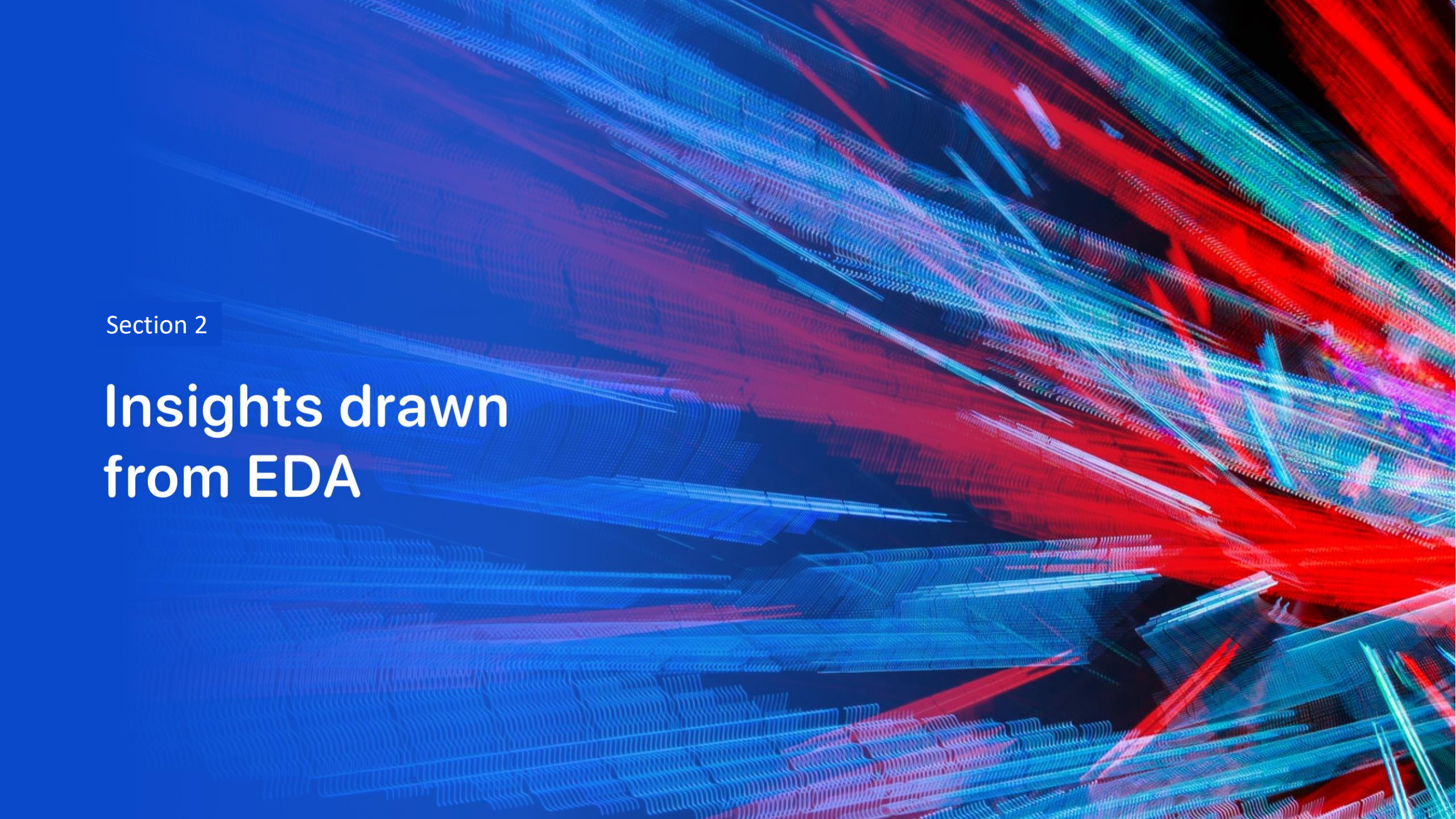
4. K Nearest Neighbor (KNN): Matched the SVM's training accuracy at 84.82%, but performed similarly on the test set with 83.33%, suggesting a potential need for hyperparameter tuning.

Overall, the Decision Tree classifier provided the best performance, making it a strong candidate for predictive tasks in this analysis. The results highlight the importance of model selection and tuning in achieving optimal predictive accuracy.

Results

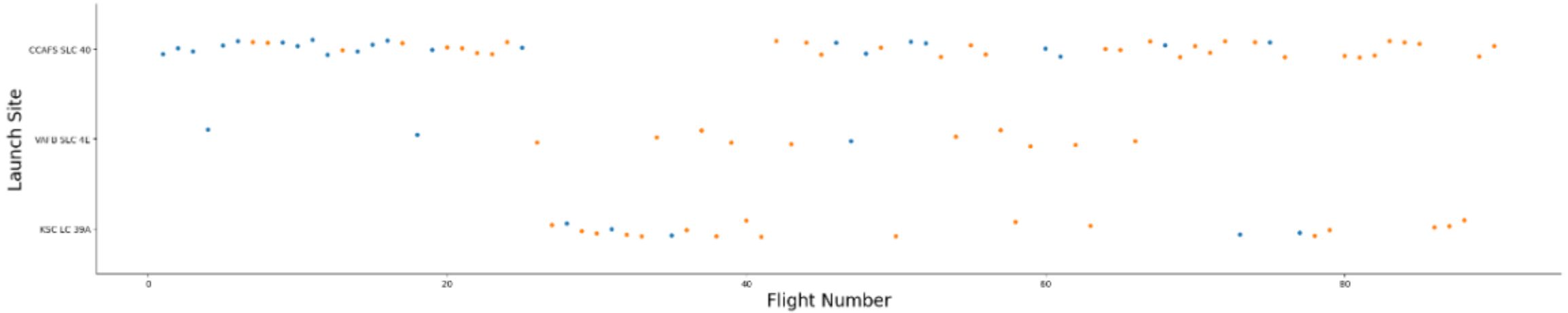
Out[65]:

	Model	Accuracy Model	Accuracy Test
0	Logistic Regression	0.846429	0.833333
1	Support Vector Machine	0.848214	0.833333
2	Decision Tree	0.873214	0.888889
3	K Nearest Neighbor	0.848214	0.833333


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

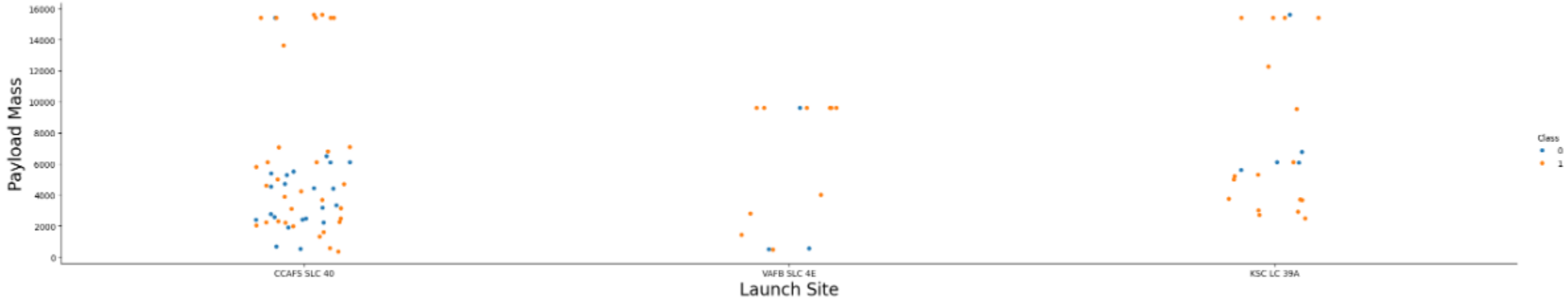
```
Out[92]: Text(32.350641666666675, 0.5, 'Launch Site')
```



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

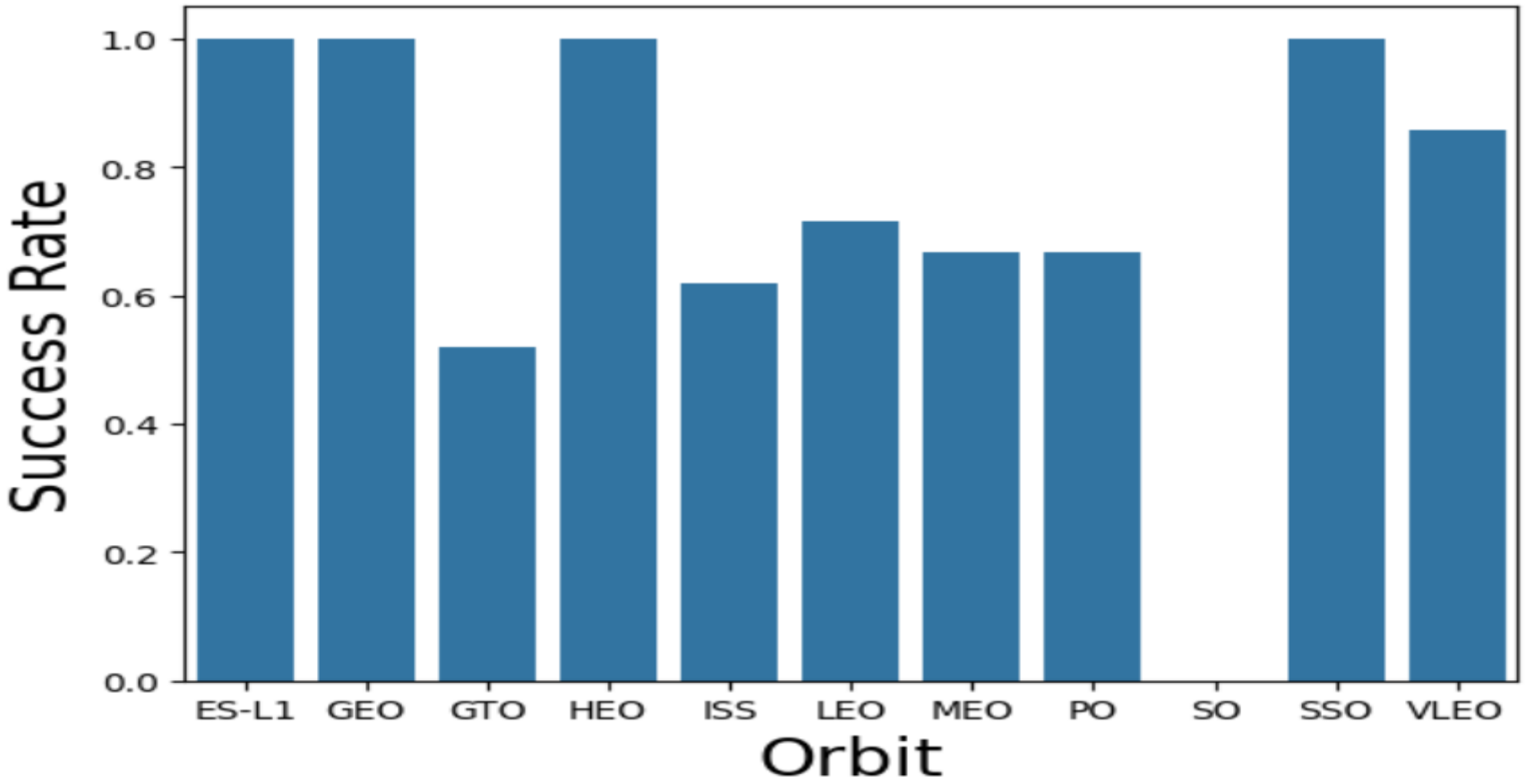
Payload vs. Launch Site

Out[93]: Text(31.07894166666665, 0.5, 'Payload Mass')



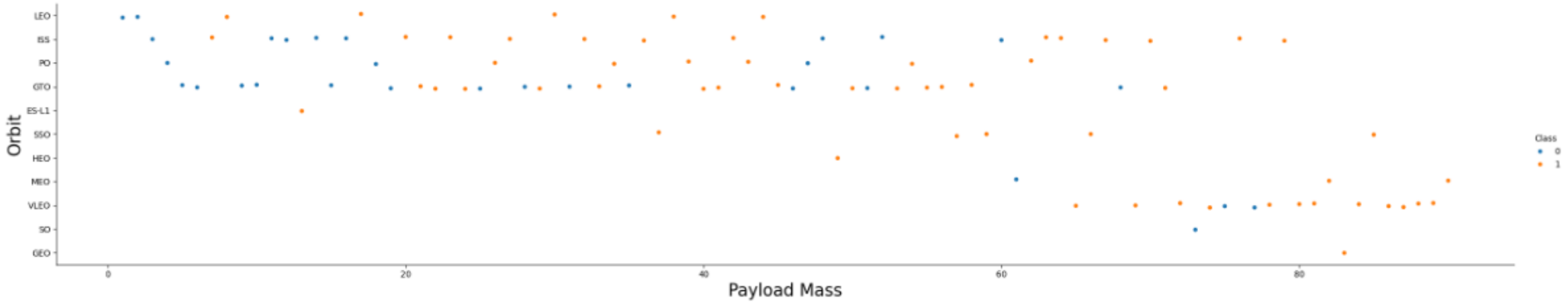
Success Rate vs. Orbit Type

Out[94]: Text(0, 0.5, 'Success Rate')



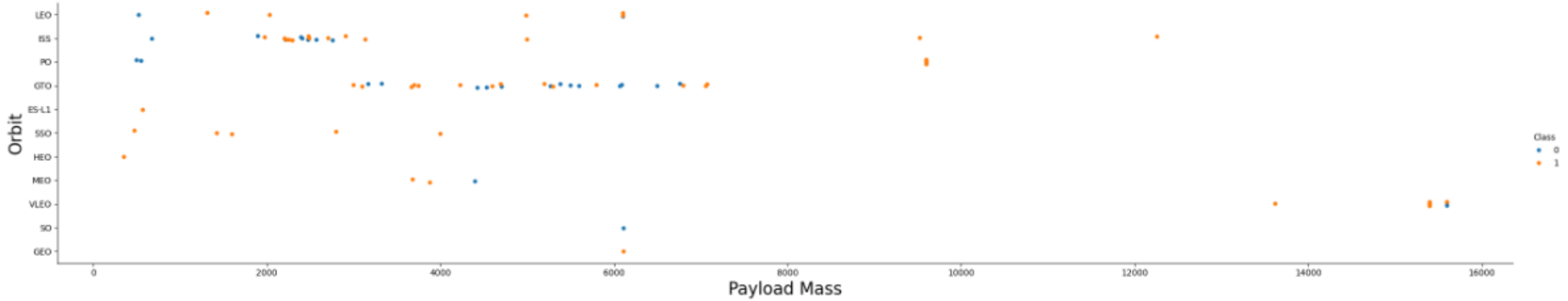
Flight Number vs. Orbit Type

```
Out[95]: Text(30.961191666666664, 0.5, 'Orbit')
```



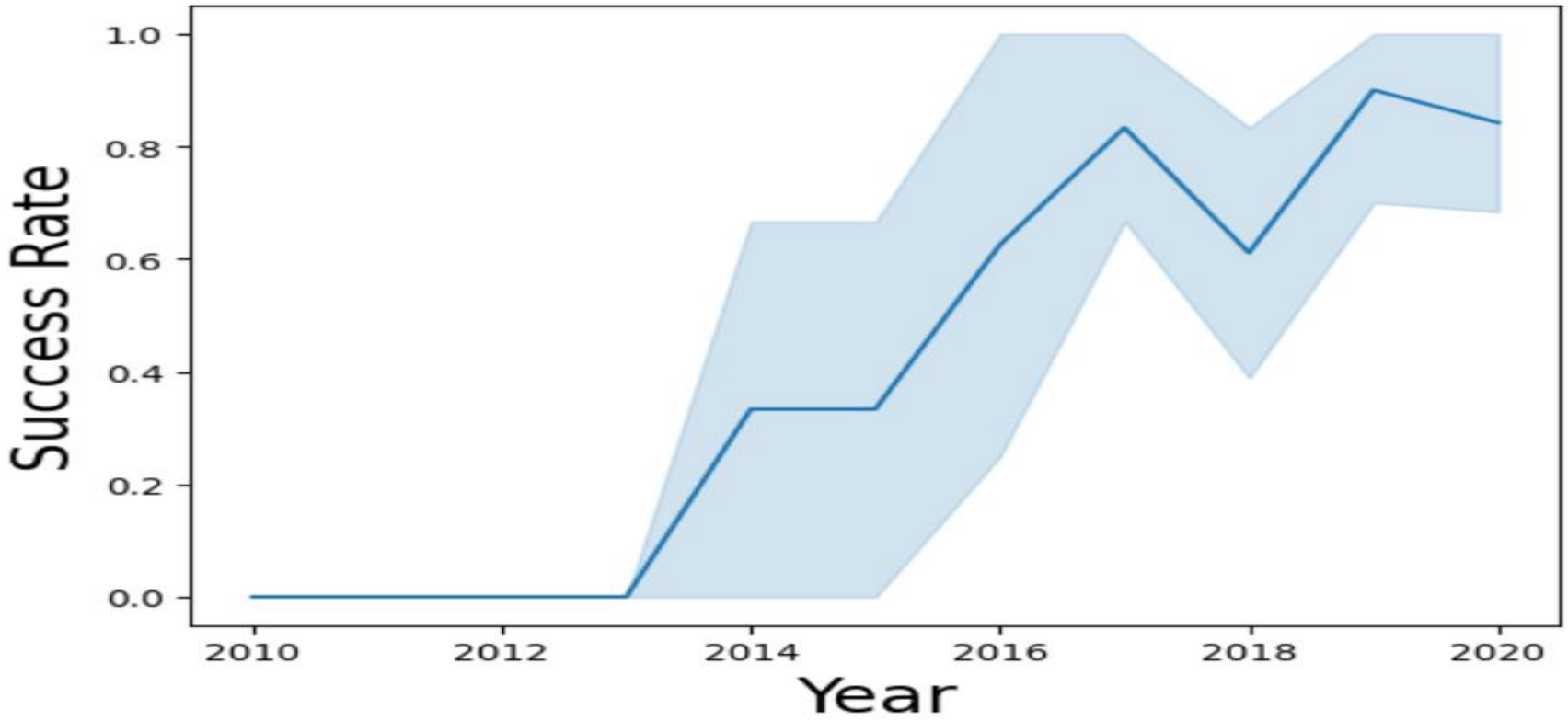
Payload vs. Orbit Type

```
Out[96]: Text(30.961191666666664, 0.5, 'Orbit')
```



Launch Success Yearly Trend

Out[98]: Text(0, 0.5, 'Success Rate')



All Launch Site Names

The SpaceX launch sites are notable locations used for launching various missions. Here are the primary launch sites associated with SpaceX:

1. Cape Canaveral Space Force Station (CCSFS) - Located in Florida, it is one of the most frequently used launch sites by SpaceX.

- -Launch Complex 40 (LC-40) - A site within CCSFS for Falcon 9 launches.
- - Launch Complex 39A (LC-39A) - Originally built for Apollo missions, it has been modified for Falcon 9 and Falcon Heavy launches.

2. Kennedy Space Center (KSC) - Also in Florida, it includes LC-39A, used for crewed missions.

3. Vandenberg Space Force Base (VSFB) - Located in California, this site is mainly used for polar orbit launches.

- - Space Launch Complex 4E (SLC-4E) - A key site at Vandenberg for Falcon 9 launches.

4. Boca Chica - In Texas, this site is primarily used for SpaceX's Starship development and testing.

Launch Site Names Begin with 'CCA'

Out[69]:

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

First Successful Ground Landing Date

Out[73]:

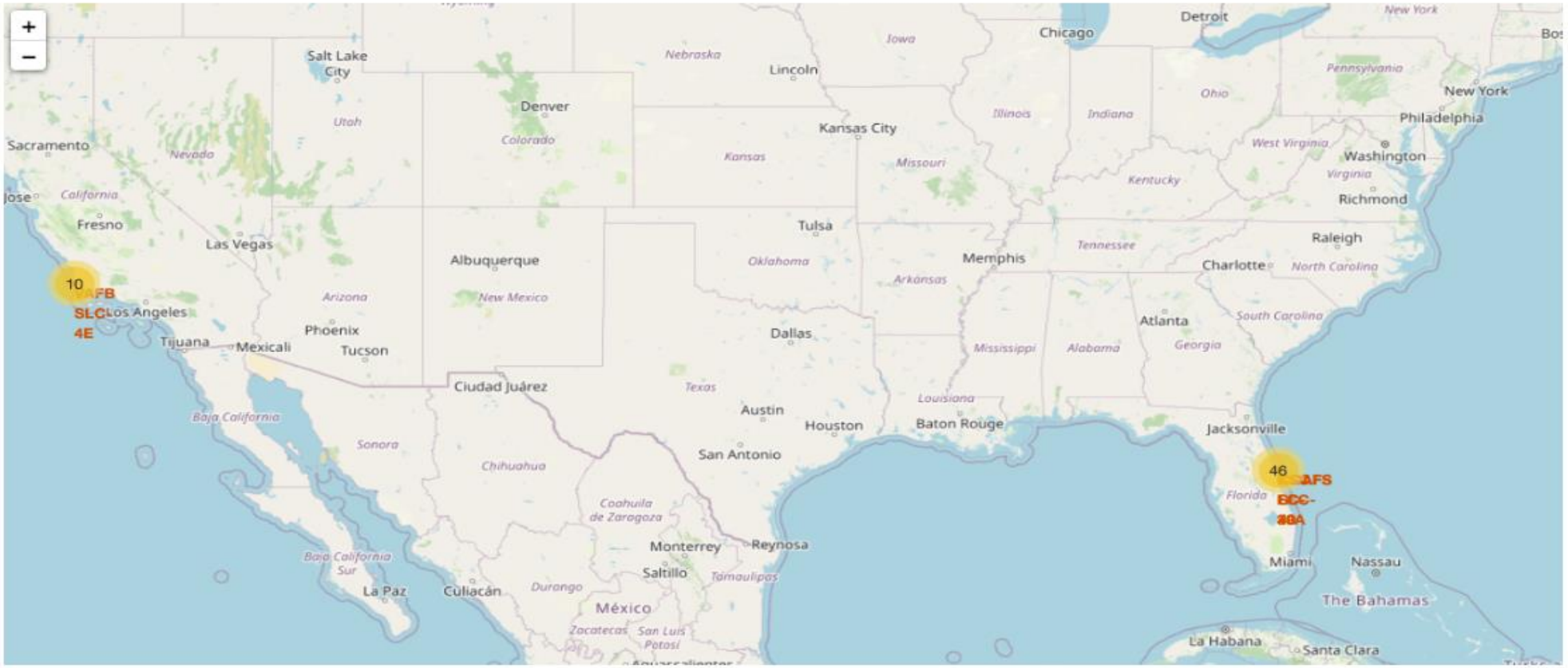
	Launch Site	Lat	Long	class
46	KSC LC-39A	28.573255	-80.646895	1
47	KSC LC-39A	28.573255	-80.646895	1
48	KSC LC-39A	28.573255	-80.646895	1
49	CCAFS SLC-40	28.563197	-80.576820	1
50	CCAFS SLC-40	28.563197	-80.576820	1
51	CCAFS SLC-40	28.563197	-80.576820	0
52	CCAFS SLC-40	28.563197	-80.576820	0
53	CCAFS SLC-40	28.563197	-80.576820	0
54	CCAFS SLC-40	28.563197	-80.576820	1
55	CCAFS SLC-40	28.563197	-80.576820	0

Section 3

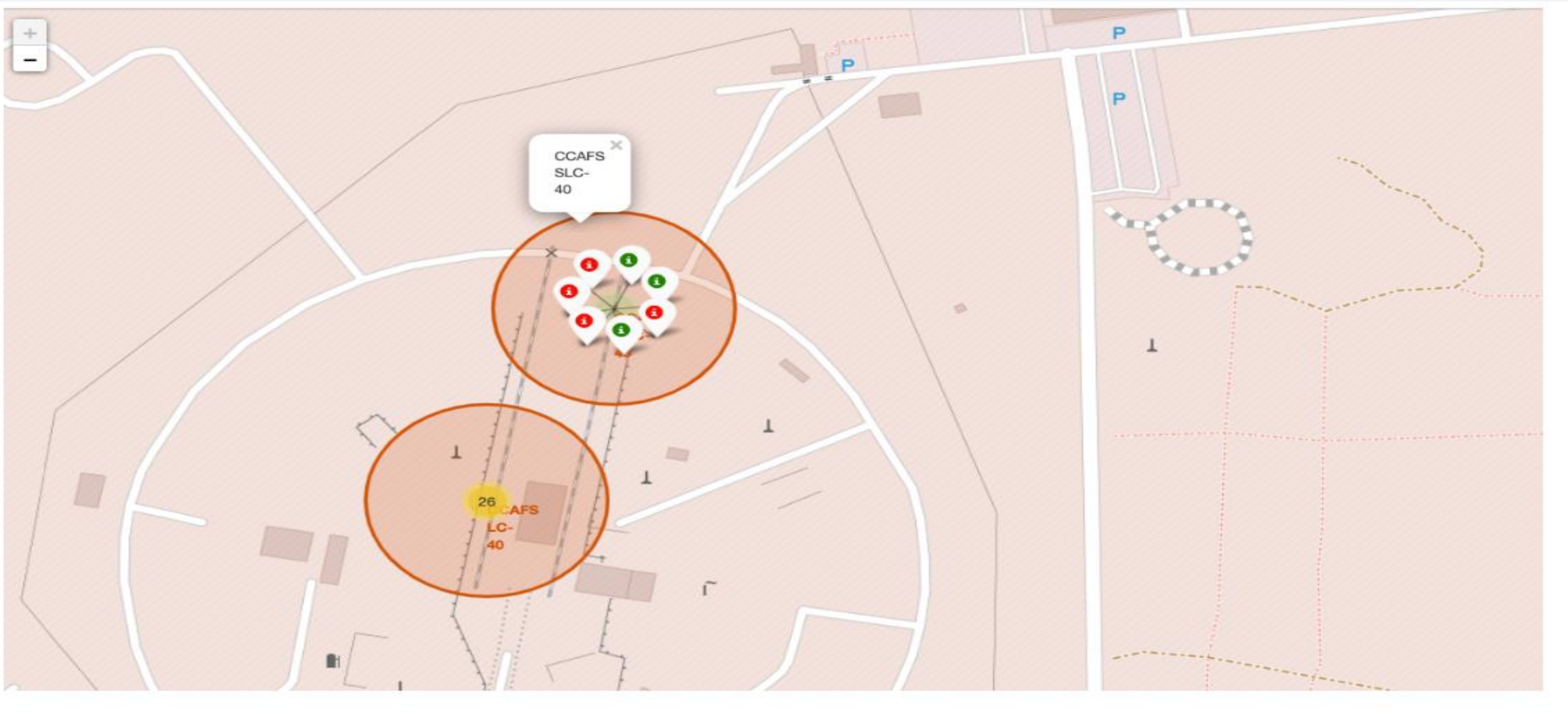
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

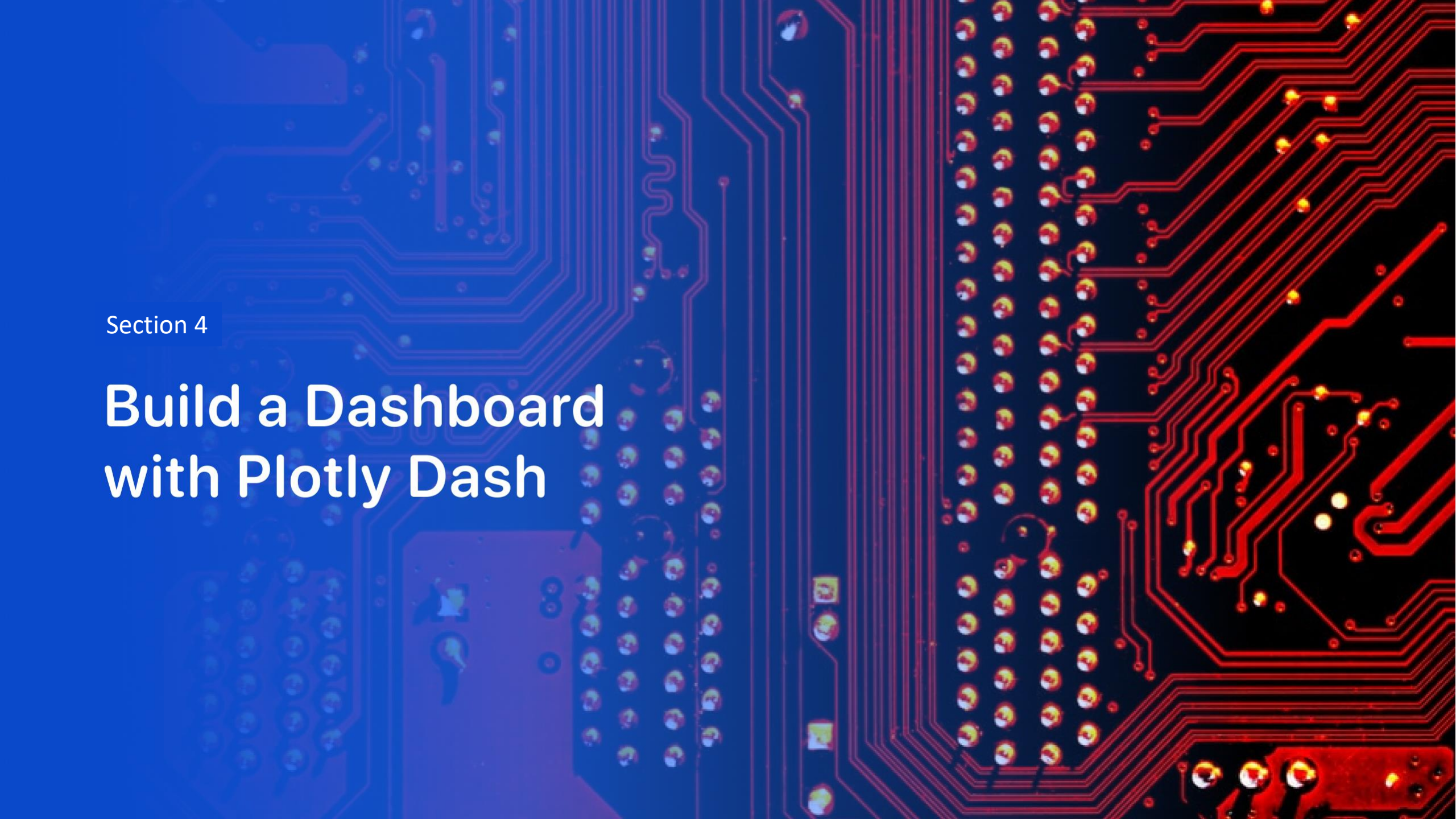
Your updated map may look like the following screenshots:



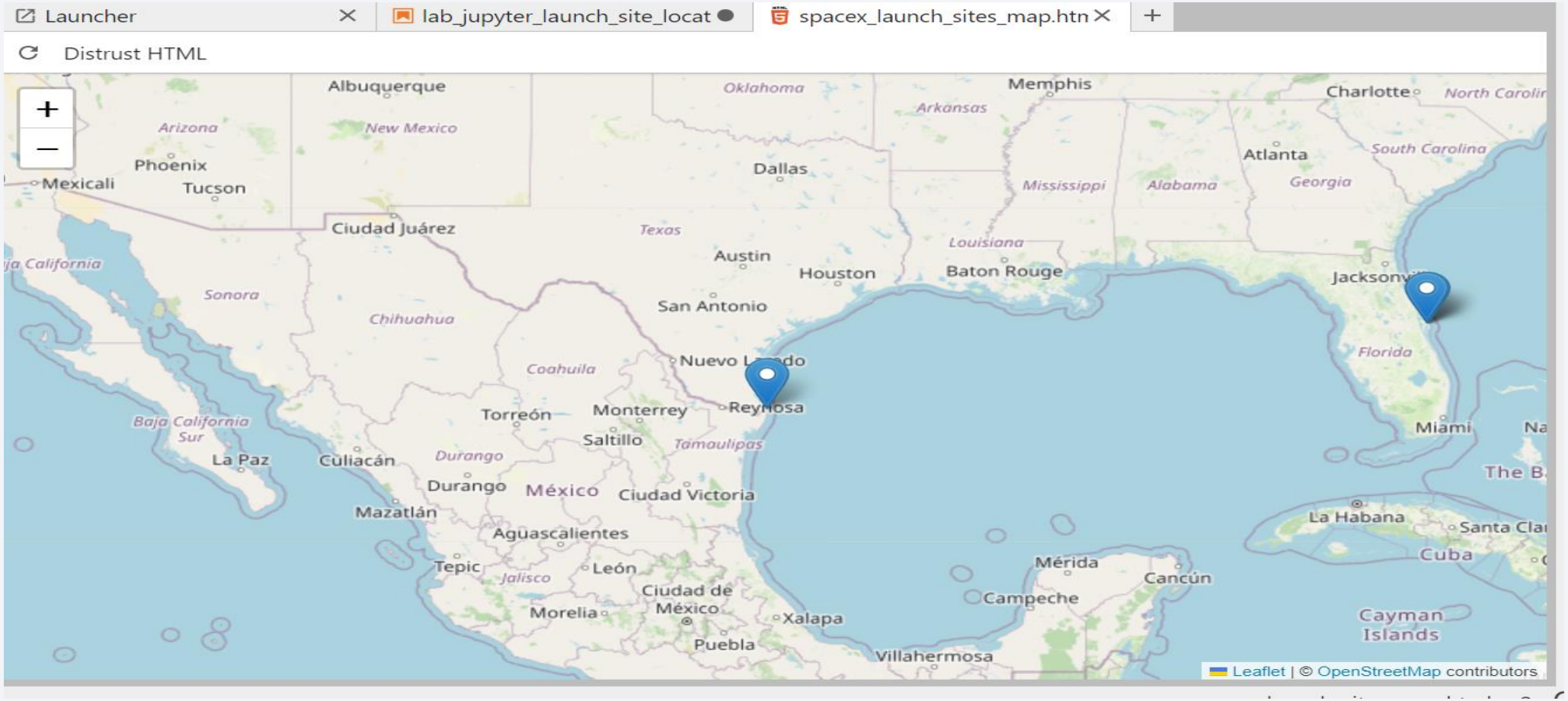
<Folium Map Screenshot 2>



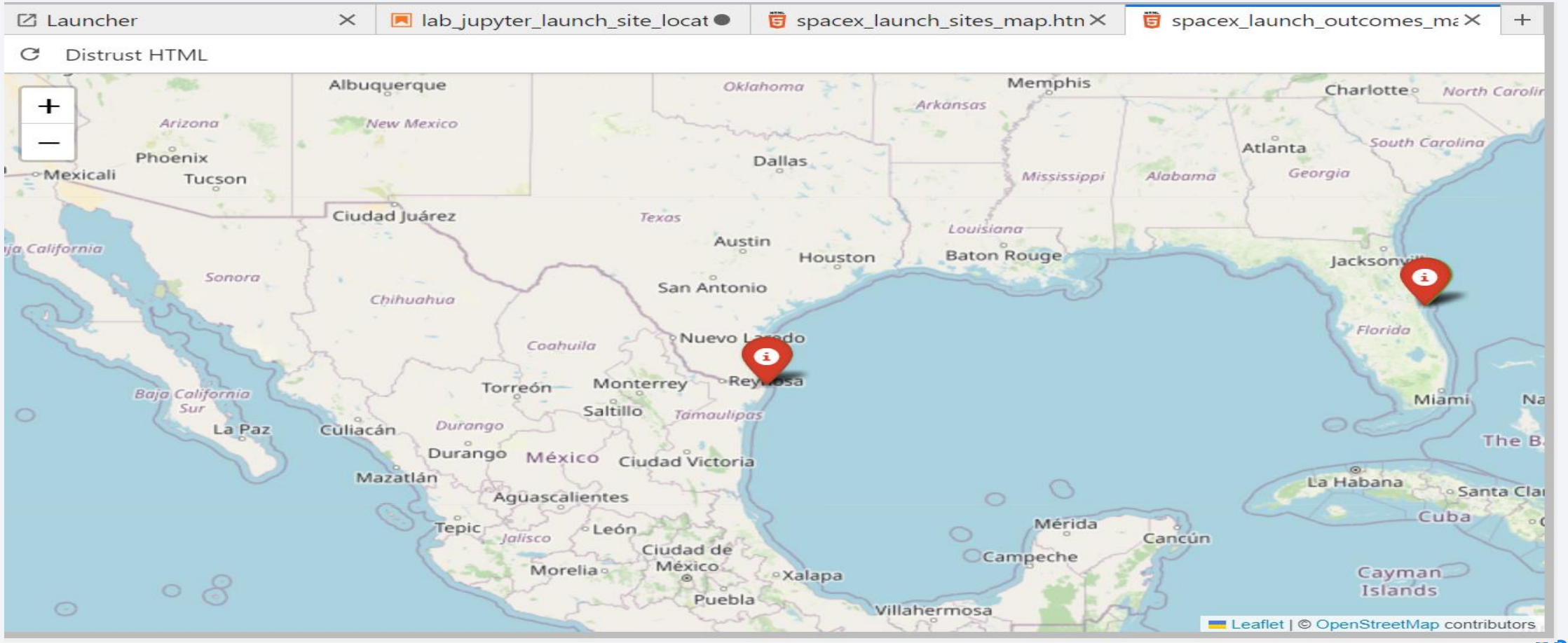
<Folium Map Screenshot 3>


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



<Dashboard Screenshot 3>

Section 5

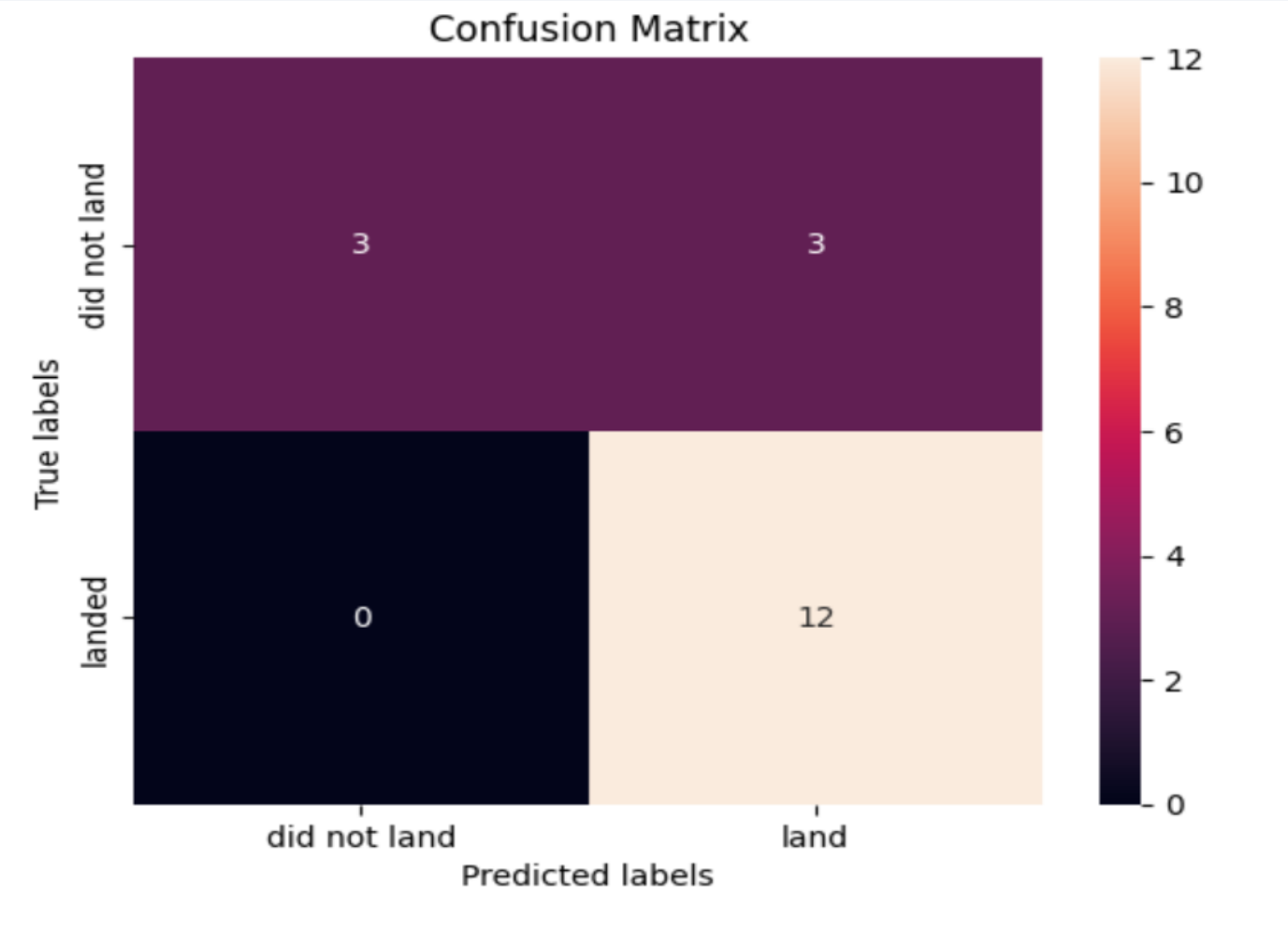
Predictive Analysis (Classification)

Classification Accuracy

Out[63]:

	Model	Accuracy Model	Accuracy Test
0	Logistic Regression	0.846429	0.833333
1	Support Vector Machine	0.848214	0.833333
2	Decision Tree	0.873214	0.888889
3	K Nearest Neighbor	0.848214	0.833333

Confusion Matrix



Conclusions

In this project, we navigated the complete data science pipeline using the SpaceX dataset.

1. Data Collection: We gathered data through APIs and web scraping, ensuring a rich dataset for analysis.

2. Exploratory Data Analysis (EDA): Key trends and patterns were identified, enhancing our understanding of launch success rates and influential variables.

3. Predictive Analysis: We implemented various classification algorithms, with Decision Trees yielding the highest accuracy. This highlights the importance of model selection based on data characteristics.

4. Visualization: Tools like Folium and Plotly Dash helped present our findings effectively, making insights more accessible for stakeholders.

Thank you!