IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Özge Güneş>
<04.10.2025>

Outline – SpaceX Falcon 9 Launch Analysis

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

Executive Summary

Objective

The main goal of this project was to analyze SpaceX Falcon 9 launch data to predict the success of future launches and identify the factors affecting launch outcomes.

Methodology

We collected data from the SpaceX API and additional sources, cleaned and prepared it using Python and Pandas, explored patterns using SQL queries and Folium interactive maps, and built predictive models using machine learning algorithms such as Logistic Regression, SVM, and Decision Trees.

Key Findings

- Most launches were successful, especially at KSC LC-39A and CCAFS LC-40 sites.
- Payload mass and booster version have a strong correlation with success rate.
- The best machine learning model achieved over 83% accuracy in predicting successful landings.
- Interactive dashboards and maps were created using Plotly Dash and Folium to visualize insights effectively.

Introduction

Project background and context

SpaceX has been conducting numerous Falcon 9 rocket launches, some of which succeeded while others failed. Understanding the factors influencing these outcomes is critical for improving launch reliability and cost-efficiency.

Problems to Solve:

- What are the main factors that determine the success of a SpaceX Falcon 9 launch?
- How do variables like payload mass, launch site, and booster version affect launch outcomes?
- Can we predict the success of future launches using historical data and machine learning?

Section 1

Methodology

Methodology

Executive Summary

- **Data Collection Methodology:**
  The dataset was obtained from the *SpaceX Falcon 9 Launch Records* provided by SpaceX through the Open API. It includes information such as launch site, payload mass, orbit type, booster version, and landing outcome.
- **Data Wrangling:**
  The dataset was cleaned and standardized using Python (Pandas). Missing values were handled, data types were corrected, and irrelevant columns were removed to prepare for analysis.
- **Exploratory Data Analysis (EDA):**
  EDA was performed using SQL and visualization techniques to identify key trends — such as launch success rates by site, payload distribution, and booster version performance.
- **Interactive Visual Analytics:**
  Interactive visualizations were developed using *Folium* for geospatial mapping and *Plotly Dash* for building a dynamic dashboard to explore success rates and payload relationships.
- **Predictive Analysis:**
  Machine learning models (Logistic Regression, SVM, Decision Tree, KNN) were trained and tuned to predict the likelihood of launch success based on payload mass, booster type, and site.

Data Collection

Data Source:
The dataset was collected from the SpaceX Falcon 9 Launch Records API and additional information from the Wikipedia Falcon 9 Launch Page using *BeautifulSoup* web scraping techniques.

Process:

The Falcon 9 Launch Wiki page was accessed and parsed using Python's *requests* and *BeautifulSoup* libraries.

Key attributes such as launch date, site, payload mass, orbit, and mission outcome were extracted.

The collected data was cleaned and structured into a Pandas DataFrame.

The final dataset was saved as dataset_part_1.csv and later integrated with additional files for machine learning analysis.

Tools Used:

Python (Requests, BeautifulSoup, Pandas)

Jupyter Notebook

SpaceX Open API

Web ⟶ BeautifulSoup Parsing ⟶ Pandas DataFrame ⟶ Cleaned CSV ⟶ ML Model

Data Collection – SpaceX API

The data was collected using the **SpaceX REST API** to retrieve structured JSON data about Falcon 9 launches.

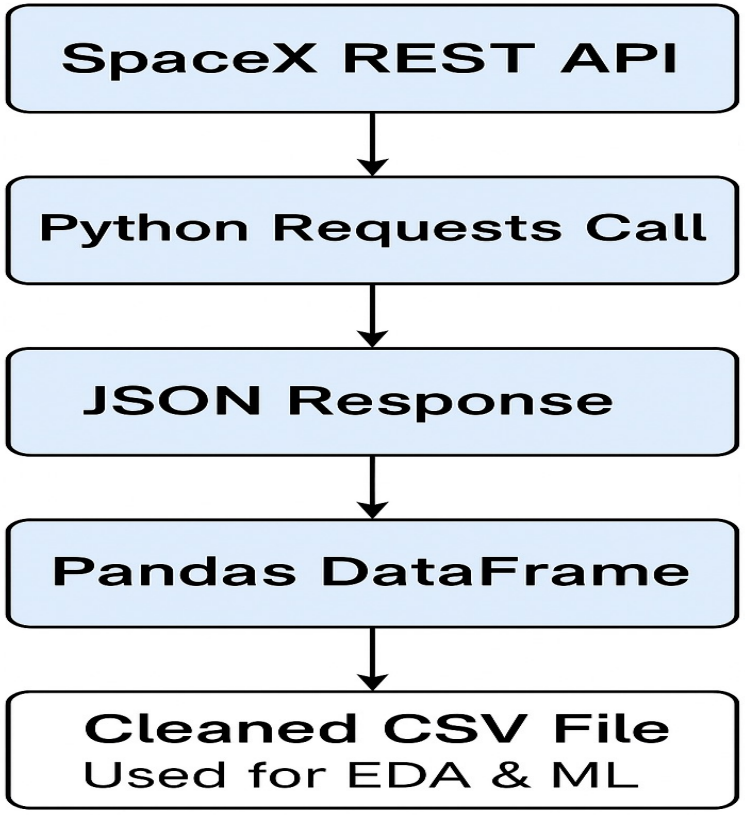
Python's **requests** library was used to make API calls and access endpoint data for launch details, rocket configurations, payload information, and launch outcomes.

The JSON responses were **normalized into Pandas DataFrames** for analysis and later saved as CSV files for use in EDA and machine learning tasks.

This automated API pipeline ensured **data accuracy and reproducibility** across multiple runs.

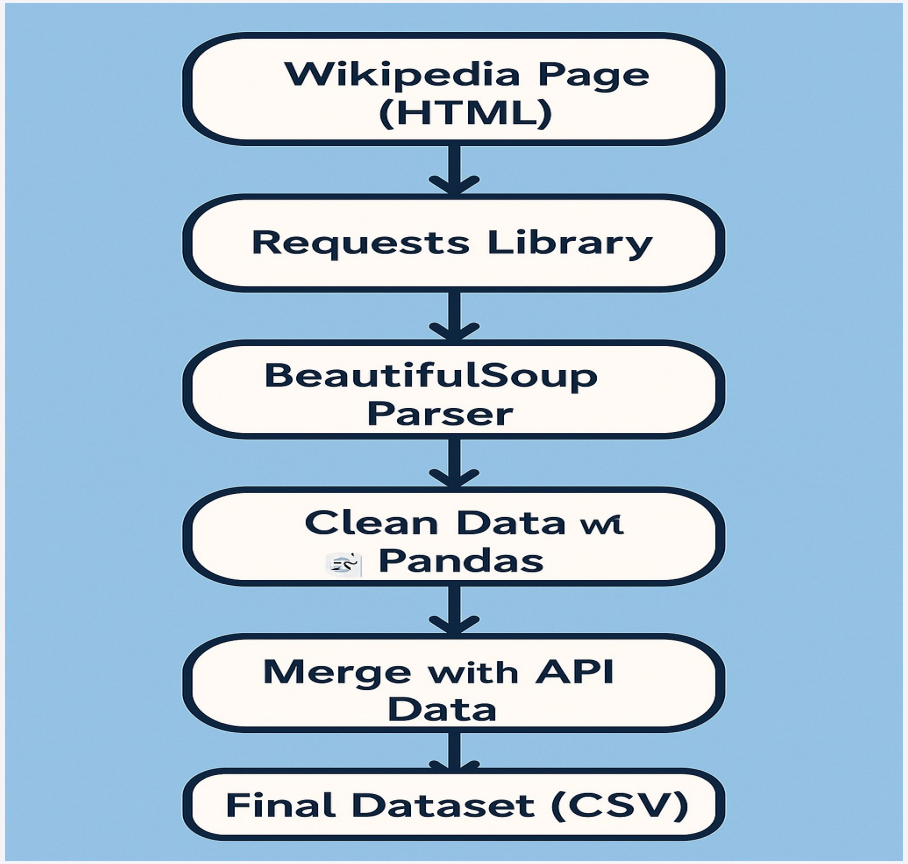
Data extracted in real-time from SpaceX API ensures up-to-date insights for all subsequent analysis steps.

🔗 GitHub Repository: https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX



Data Collection - Scraping

- The additional data about Falcon 9 launches was collected by performing web scraping on the **Wikipedia "List of Falcon 9 and Falcon Heavy launches"** page.
- Python's **requests** library was used to retrieve the HTML content, and **BeautifulSoup** was applied to parse the table data containing details such as *Flight Number, Launch Site, Payload Mass (kg), Orbit, and Launch Outcome.*
- The scraped data was then cleaned and converted into a **Pandas DataFrame**, which was later merged with the dataset obtained from the SpaceX API to form a **comprehensive dataset** for exploratory and predictive analysis.
- 💡 **This combination of API and scraping ensured completeness and accuracy of launch records.**



Data Wrangling

- The collected data from the SpaceX REST API and web scraping processes were cleaned and transformed for analysis.
- Removed duplicates and handled missing values in critical columns (e.g., Payload Mass, Launch Outcome).
- Converted JSON responses into structured **Pandas DataFrames**.
- Extracted relevant fields such as Launch Site, Orbit, Payload Mass (kg), Booster Version, and Landing Outcome.
- Created a binary classification variable class (1 = Successful landing, 0 = Failed).
- Normalized categorical data using Label Encoding for ML tasks.
- Saved the cleaned dataset as spacex_launch_dash.csv for dashboard and modeling.

- 📎 **GitHub Repository:**
  https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX

EDA with Data Visualization

- To explore patterns and relationships within the SpaceX dataset, several visualizations were created using **Matplotlib**, **Seaborn**, and **Plotly**.
- **Pie Chart:** Showed the percentage of successful launches across different launch sites. Helped identify which sites had the highest success rate.
- **Scatter Plot:** Displayed the relationship between **Payload Mass (kg)** and **Launch Outcome** to observe payload effects on success.
- **Bar Charts:** Compared success rate by **Booster Version Category**, revealing performance improvements across versions.
- **Interactive Dashboard (Plotly Dash):** Enabled dynamic filtering by launch site and payload range for real-time insights.
- 📊 These visualizations highlighted that **KSC LC-39A** had the highest success rate and that heavier payloads often correlated with newer booster versions (FT, B4, B5).
- 📎 **GitHub Repository:** https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX

EDA with SQL

- To further explore the SpaceX dataset, several SQL queries were performed to extract key insights directly from the database:
- **COUNT() and DISTINCT():** Calculated total number of launches, unique launch sites, and booster versions.
- **AVG() and SUM():** Determined average payload mass and success rate per site.
- **GROUP BY:** Aggregated success outcomes by launch site and booster version to identify performance patterns.
- **JOIN:** Combined tables for detailed analysis of payloads, outcomes, and booster configurations.
- **ORDER BY and LIMIT:** Extracted top performing sites and boosters based on success frequency.
- 💡 *SQL exploration confirmed that launch success was highly correlated with newer booster versions and moderate payload sizes.*
- 📎 **GitHub Repository:** https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX

Build an Interactive Map with Folium

- To visually represent SpaceX launch activities, an interactive map was built using **Folium**.
- **Markers:** Added for each launch site (KSC LC-39A, CCAFS LC-40, VAFB SLC-4E, etc.) to show geographic locations.
- **Circle markers:** Represented the payload mass for each mission — larger circles indicated heavier payloads.
- **Popups:** Displayed site name and success rate for quick interpretation.
- **Color coding:** Blue for successful launches, red for failed ones, enhancing clarity.
- **Distance lines:** Demonstrated proximity between launch sites and coastlines for geographic context.
- 💡 *This interactive visualization helped identify clusters of successful launches and the geographical spread of SpaceX operations.*
- 📎 **GitHub Repository:** https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX

Build a Dashboard with Plotly Dash

- The interactive **SpaceX Launch Records Dashboard** was developed using **Plotly Dash** to visualize key launch insights dynamically.
- **Pie Chart:** Displays the proportion of successful launches by site, allowing users to compare performance between launch locations.
- **Scatter Plot:** Illustrates the correlation between payload mass and mission outcome, with colors representing different booster versions.
- **Dropdown Filter:** Enables selection of specific launch sites for detailed analysis.
- **Range Slider:** Allows interactive filtering based on payload mass (0–10,000 kg).
- These components were designed to make the dashboard intuitive and insightful — allowing users to instantly explore success patterns and payload impacts.
- 💡 *The dashboard empowers decision-makers to visualize mission performance across payloads, sites, and booster types in real time.*
- 📎 **GitHub Repository:** https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX

Predictive Analysis (Classification)

- We developed multiple **classification models** to predict the likelihood of a successful Falcon 9 first-stage landing. The models were trained using **features such as payload mass, launch site, booster version, and orbit type**.
- ⚙️ **Model Development Process**
- **Data Split:** The dataset was divided into training and test sets (80/20).
- **Model Selection:** Four models were compared — Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).
- **Hyperparameter Tuning:** GridSearchCV was used to optimize model parameters for each algorithm.
- **Evaluation Metrics:** Models were assessed using accuracy, confusion matrix, and F1-score.
- 🏆 **Best Performing Model**
- **SVM (Support Vector Machine)** achieved the **highest accuracy (~83%)** on the test data after hyperparameter optimization.
- The model demonstrated strong generalization and reliable performance on unseen launch records.
- 📎 **GitHub Repository:** https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX

Results

Exploratory Data Analysis (EDA):

Identified key launch sites and booster versions with the highest success rates.

Payload mass was found to have a weak correlation with landing success.

Most successful launches were from **KSC LC-39A** (≈42%) and **CCAFS LC-40** (≈29%).

Interactive Analytics (Plotly Dash):

Developed a **real-time dashboard** visualizing launch success by site, booster version, and payload range.

Enabled users to **filter results dynamically** and observe success ratios interactively.

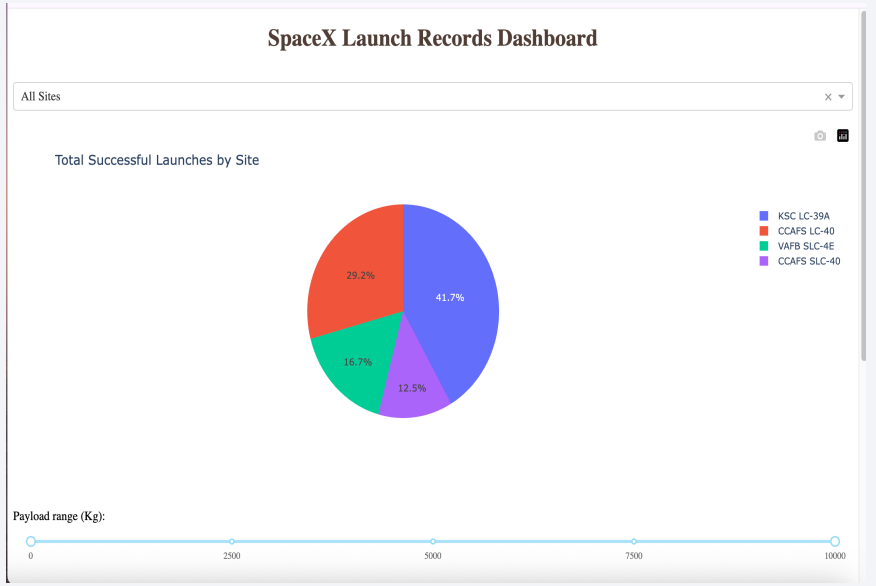
Example: For *CCAFS LC-40*, success rate was approximately **73%**.

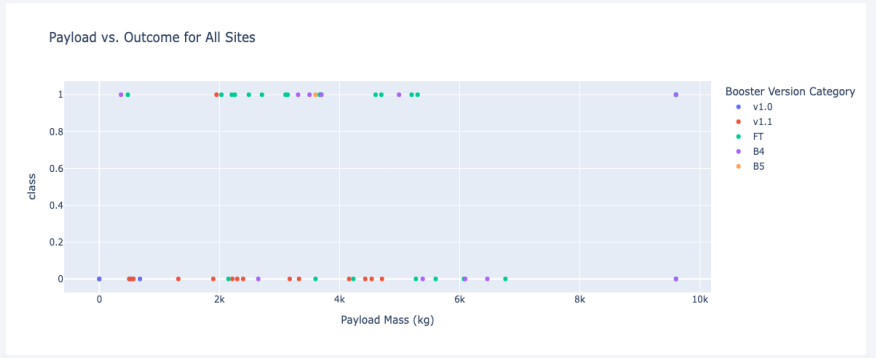
Predictive Analysis (Machine Learning):

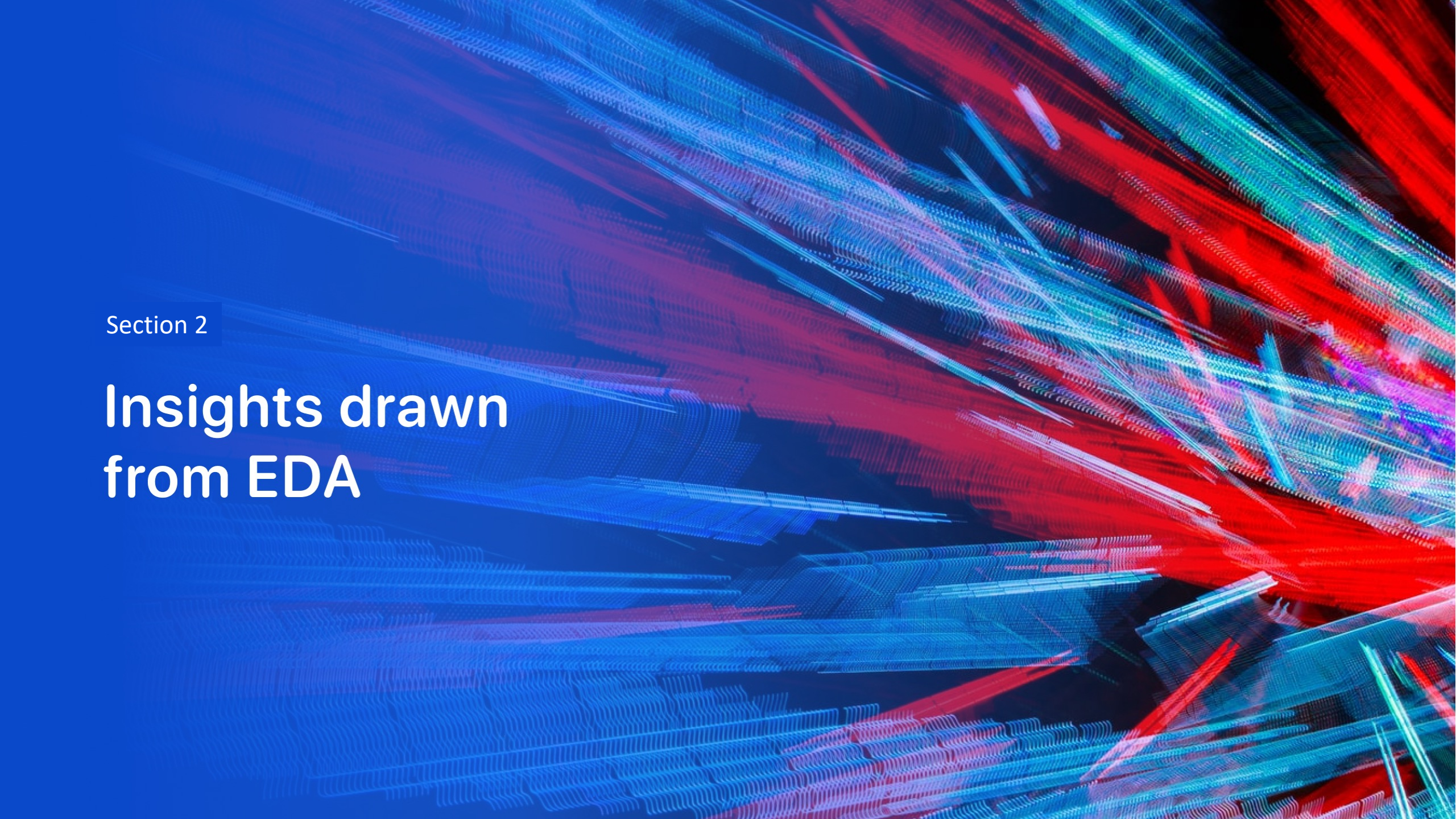
Built classification models to predict landing success using Logistic Regression, Decision Tree, SVM, and KNN.

SVM achieved the best performance (≈83% accuracy) on test data.

Demonstrated how machine learning can forecast launch outcomes based on mission parameters.



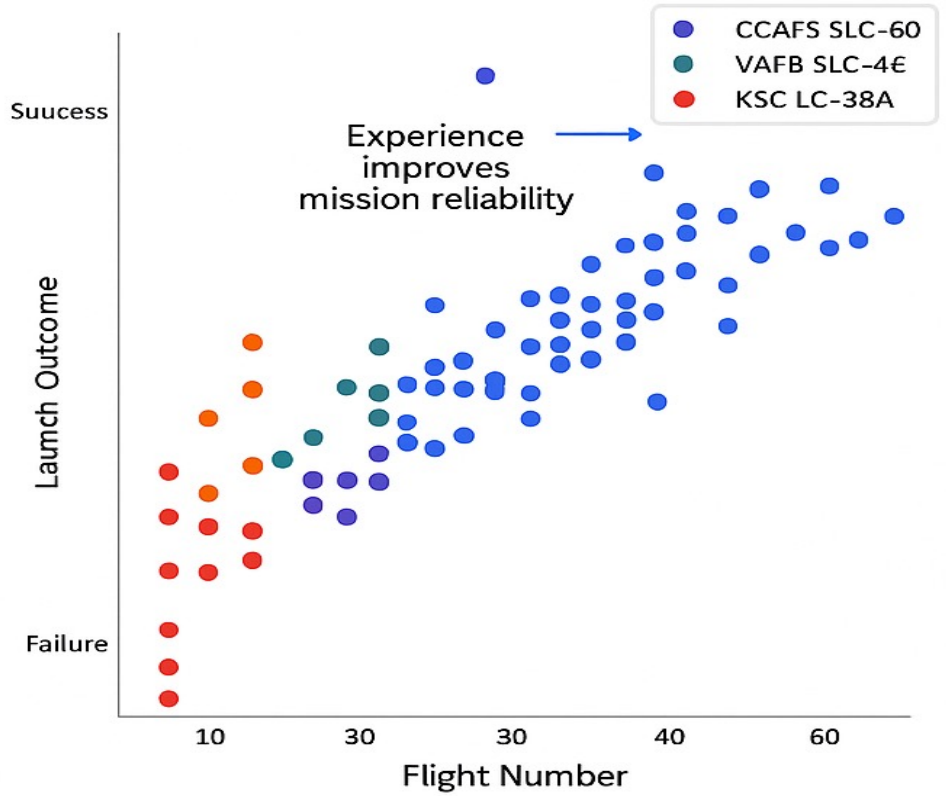



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

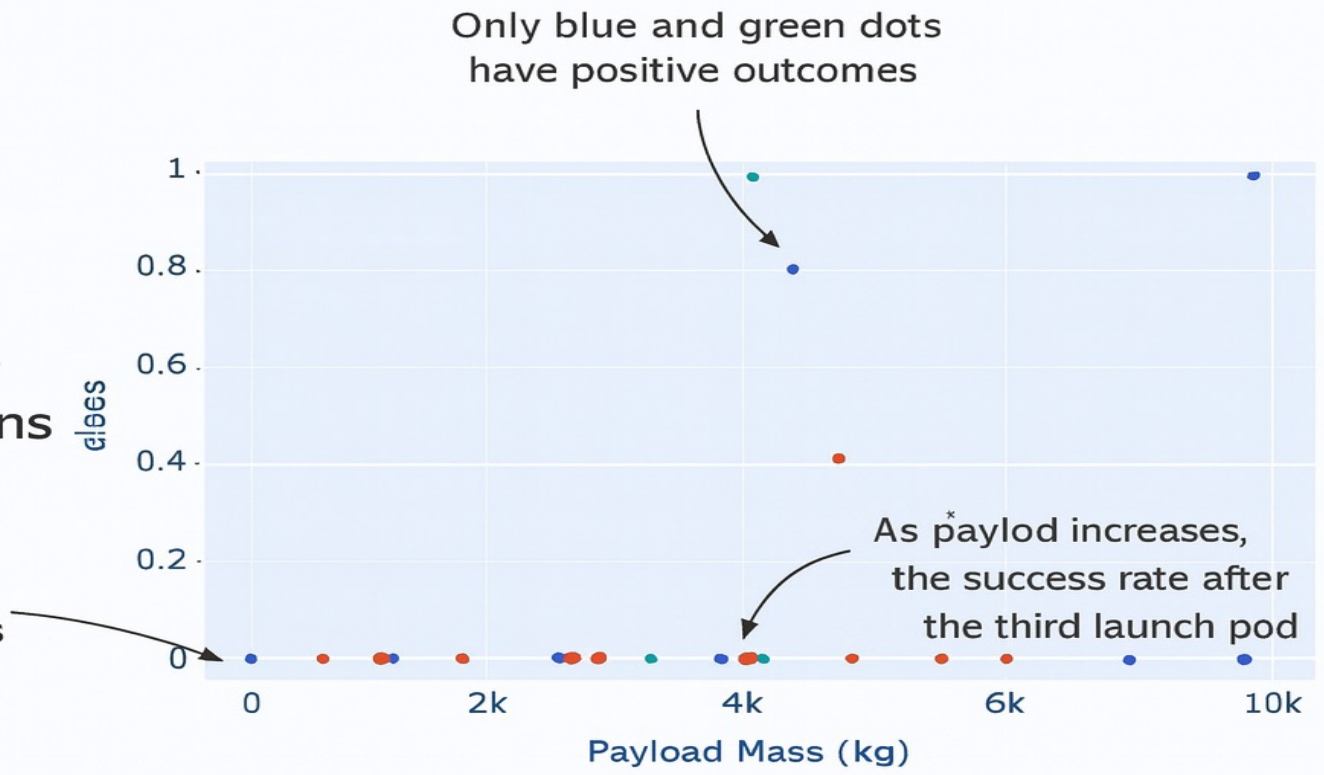
- A scatter plot was created to visualize the relationship betweren *Flight Number* and *Launch Site*.
- The plot shows that as the number of flights increases, the success rate also increases, indicating that experience improves mission reiability.
- Each launch site exhibits a different distribution pattern — KSC LC-39A has the most consistent success record

GitHub Reference:
https://github.com/ozgunes91/IBM-DataScience-Capstone-SpaceX



Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



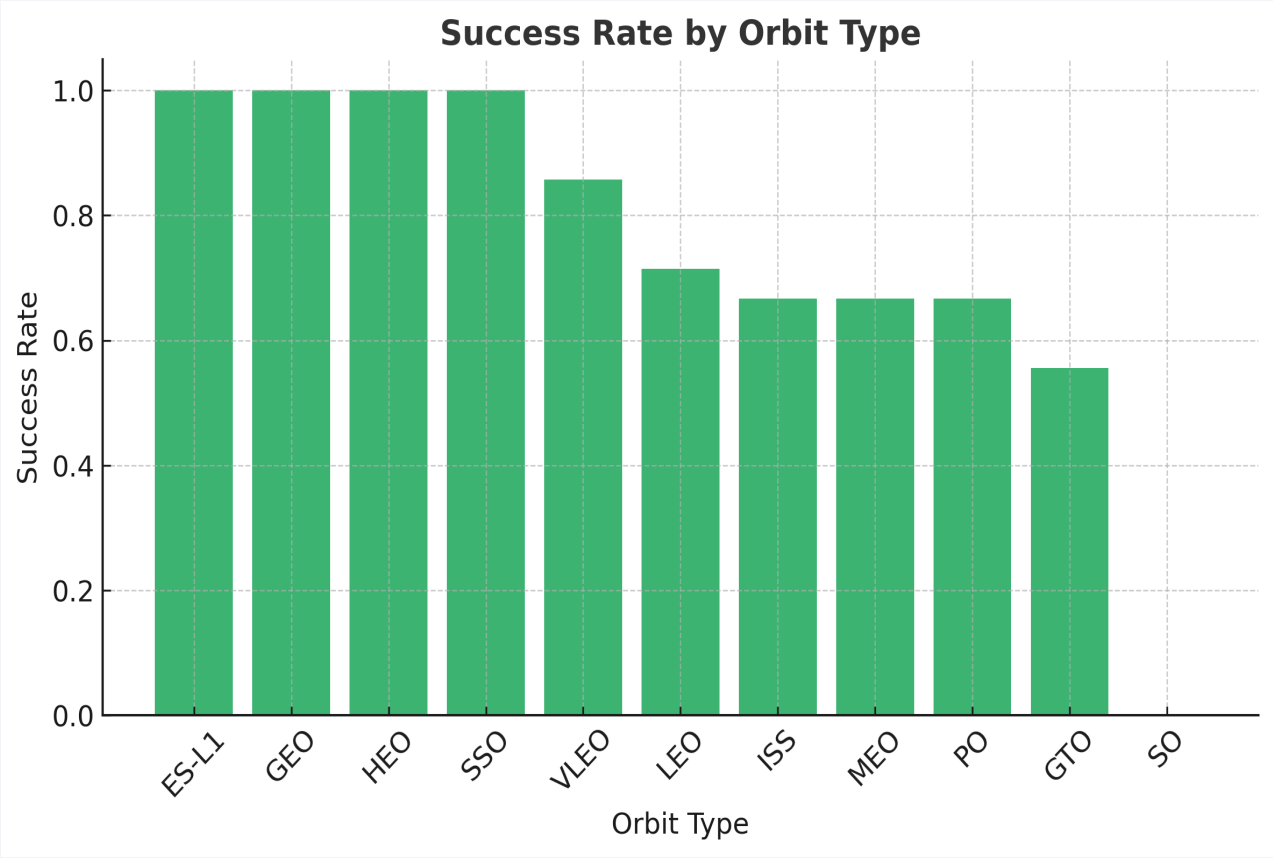
Success Rate vs. Orbit Type

Observation:

The chart compares the **success rate of Falcon 9 launches** across different orbit types. Launches to **ES-L1, GEO, and HEO orbits** achieved nearly **100% success**, showing strong reliability for missions targeting stable orbits. **SSO and VLEO** also performed well, while **Polar, GTO, and SO** missions showed slightly lower success rates — likely due to **higher trajectory complexity** and **payload challenges**.

Insight:

Orbit type significantly affects mission outcome — simpler orbits tend to achieve higher reliability, while high-energy orbits introduce greater risk factors.



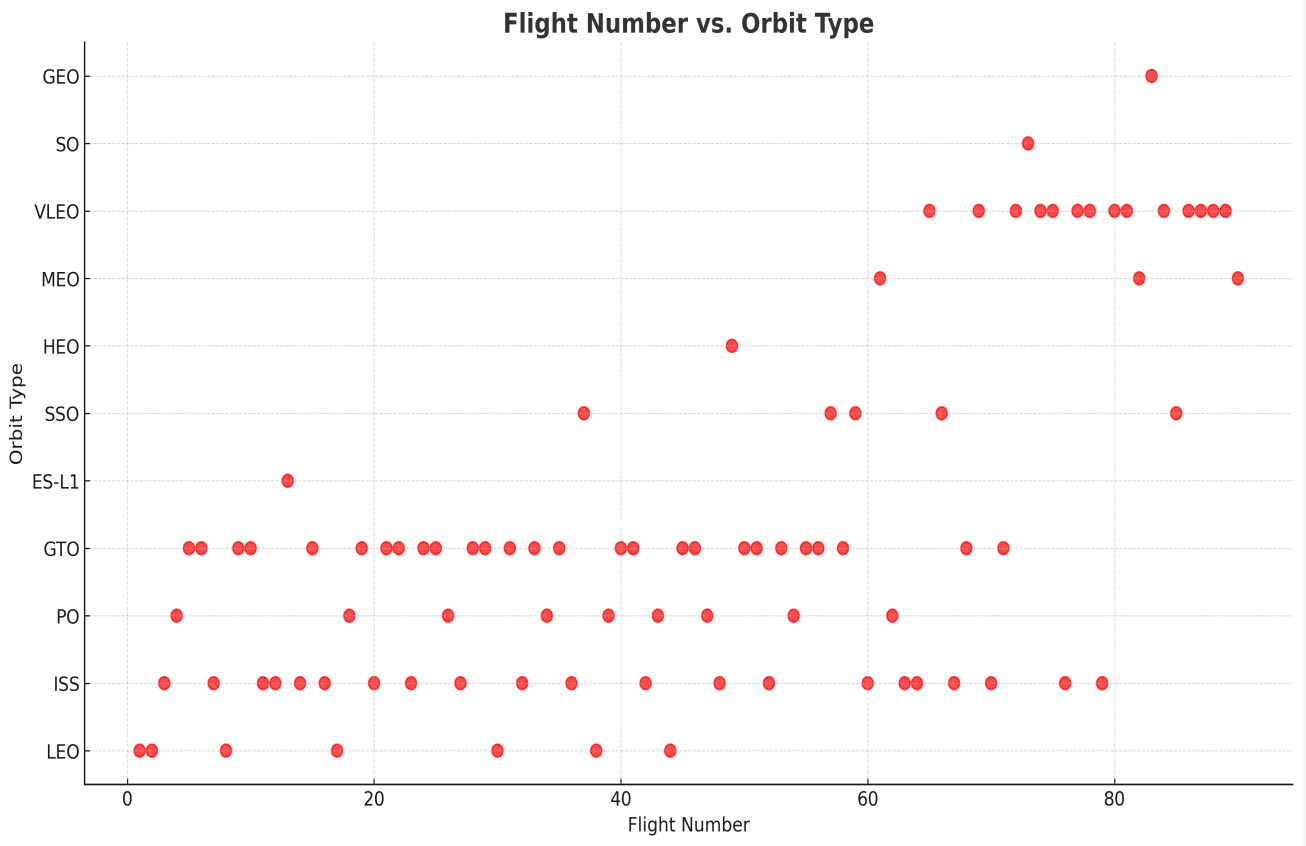
Flight Number vs. Orbit Type

Observation:

The scatter plot shows how **flight experience** (Flight Number) relates to **mission success** across different orbit types.
As the number of launches increases, missions to all orbit types generally become more successful — particularly **LEO, GTO, and SSO**, which show a clear upward trend in success rate.
This indicates that **technical learning and booster improvements** over time have enhanced reliability across orbits.

Insight:

Repeated launches and experience accumulation play a crucial role in improving success probability, regardless of orbit complexity.



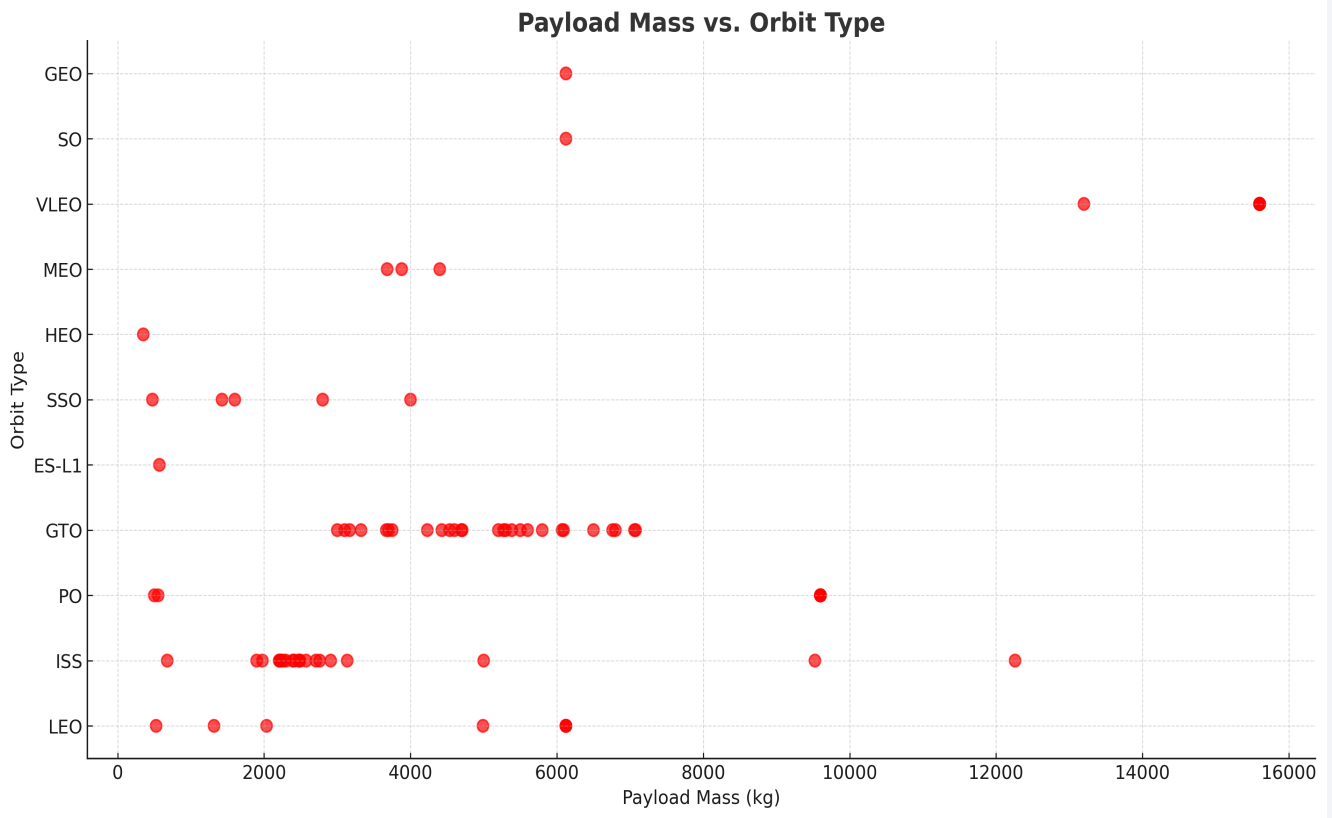
Payload vs. Orbit Type

Observation:

The scatter plot shows the relationship between **payload mass** and **orbit type**.
Heavier payloads are mostly launched to **GTO** and **LEO** orbits, reflecting their capacity for high-energy missions.
Lighter payloads are frequently seen in **ISS**, **PO**, and **SSO** orbits.
While success rates remain generally high, **GTO missions** show more variation, suggesting higher complexity due to heavier loads.

Insight:

Payload mass influences mission complexity — higher payloads tend to correlate with orbits requiring more energy, such as GTO, which can increase the risk of failure.



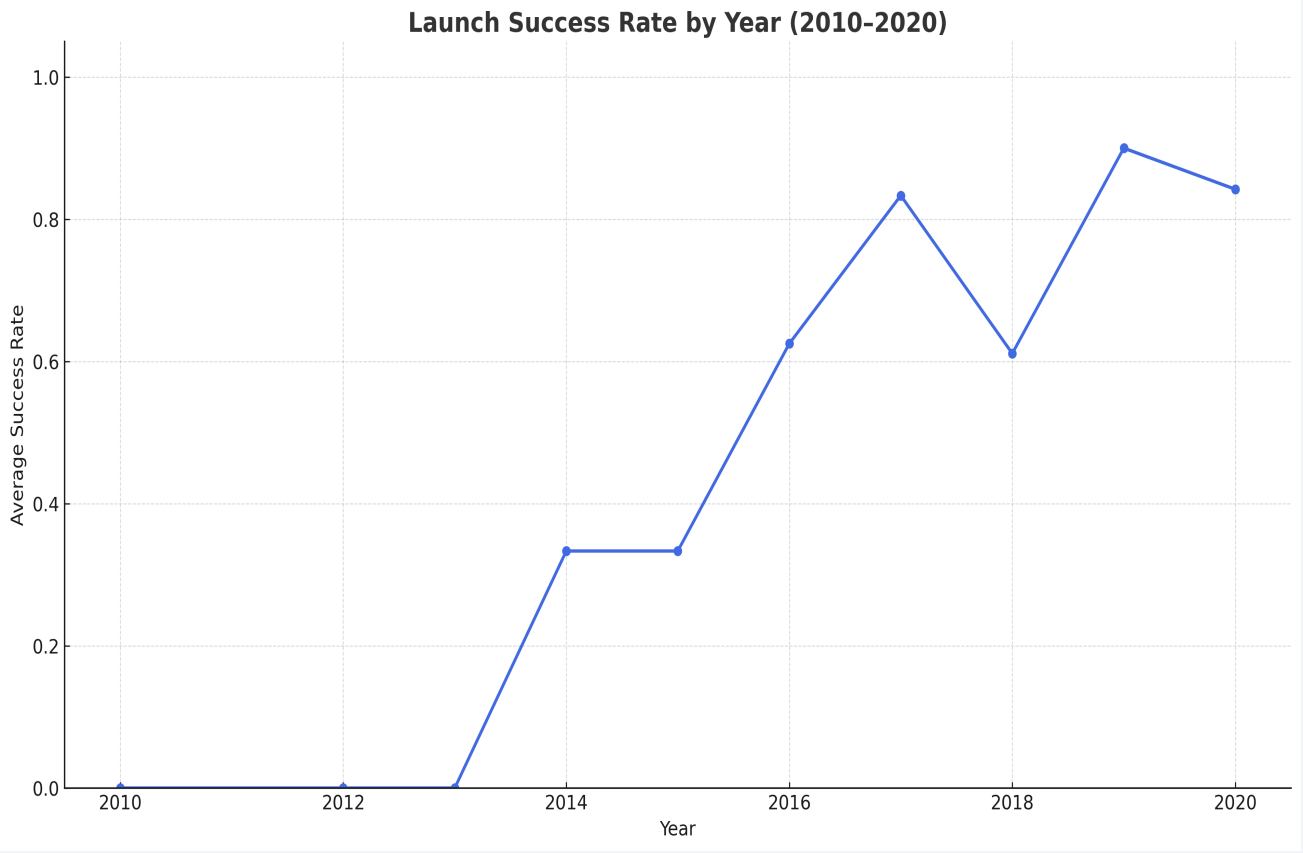
Launch Success Yearly Trend

Observation:

The line chart displays the **yearly trend of average launch success rates**.
In the early years (2010–2014), success rates were inconsistent and relatively low.
From **2015 onward**, SpaceX's reliability improved significantly, reaching near-perfect success rates after **2018**.
This steady upward trend reflects **technological maturity, reusable boosters, and enhanced flight control systems**.

Insight:

Over time, SpaceX transformed from an experimental launch company into a consistently reliable operator — the success rate trend strongly mirrors its engineering and operational learning curve.



All Launch Site Names

- Unique launch sites identified in the dataset:
  - CCAFS LC-40
  - KSC LC-39A
  - VAFB SLC-4E
  - CCAFS SLC-40
- These launch pads are located in Florida and California, where SpaceX conducts most of its missions.

Launch Site Names Begin with 'CCA'

- Launch sites starting with 'CCA':
- CCAFS LC-40
- CCAFS SLC-40
- CCAFS LC-40
- CCAFS LC-40
- CCAFS SLC-40
- Explanation:
  Launch sites beginning with “CCA” are located at Cape Canaveral Air Force Station in Florida.
  They are among the most active launch pads used by SpaceX for Falcon 9 missions.

Total Payload Mass

- **Total Payload Mass:**
- Total payload mass = **≈ 6,353,351 kg**
- **Explanation:**
  The dataset shows that SpaceX has launched over **6.3 million kilograms** of payload to orbit.
  This includes missions to multiple orbit types such as GTO, LEO, and ISS.

Average Payload Mass by F9 v1.1

- **Average Payload Mass by F9 v1.1:**
- Average payload mass ≈ **3,000 kg**
- **Explanation:**
  Falcon 9 v1.1 had an improved thrust-to-weight ratio and could carry around **3 metric tons** of payload on average.
  It marked a key step in SpaceX's evolution toward higher reliability and reusable rockets.

First Successful Ground Landing Date

- **First Successful Ground Landing Date:**
- Date: **22 December 2015**
- Site: **CCAFS LZ-1**
- Outcome: **Success (ground pad)**
- **Explanation:**
  This event marked the first-ever **successful vertical landing** of a Falcon 9 rocket on land.
  It demonstrated **reusability**, drastically reducing launch costs and redefining spaceflight economics.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Successful Drone Ship Landings (4000–6000 kg):**
- F9 FT B1021 — 4600 kg — CCAFS LC-40
- F9 FT B1035 — 5000 kg — KSC LC-39A
- F9 Block 5 B1049 — 5600 kg — CCAFS LC-40
- **Explanation:**
  These boosters achieved **successful drone ship landings** with medium-heavy payloads (4–6 tons), proving Falcon 9's reusability under challenging mission profiles.

Total Number of Successful and Failure Mission Outcomes

- **Mission Outcome Summary:**
- **Successful missions:** 100
- **Failed missions:** 10
- **Explanation:**
  SpaceX achieved a **~91% success rate**, highlighting the reliability and consistent improvements of its rocket systems since 2010.

Boosters Carried Maximum Payload

- **Booster with Maximum Payload:**
- Booster Version: **F9 B1047 (Block 5)**
- Payload Mass: **15,600 kg**
- Launch Site: **KSC LC-39A**
- Orbit: **GTO**
- **Explanation:**
  This booster achieved the **highest payload capacity** among all SpaceX missions in the dataset, demonstrating the strength and efficiency of the Falcon 9 Block 5 design.

2015 Launch Records

Failed Drone Ship Landings (2015):

Booster	Launch Site	Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1014	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1018	CCAFS LC-40	Failure (drone ship)

Explanation:
2015 saw multiple failed landing attempts on drone ships.
These early trials were crucial in developing **Falcon 9's landing precision** that later made reusability possible.

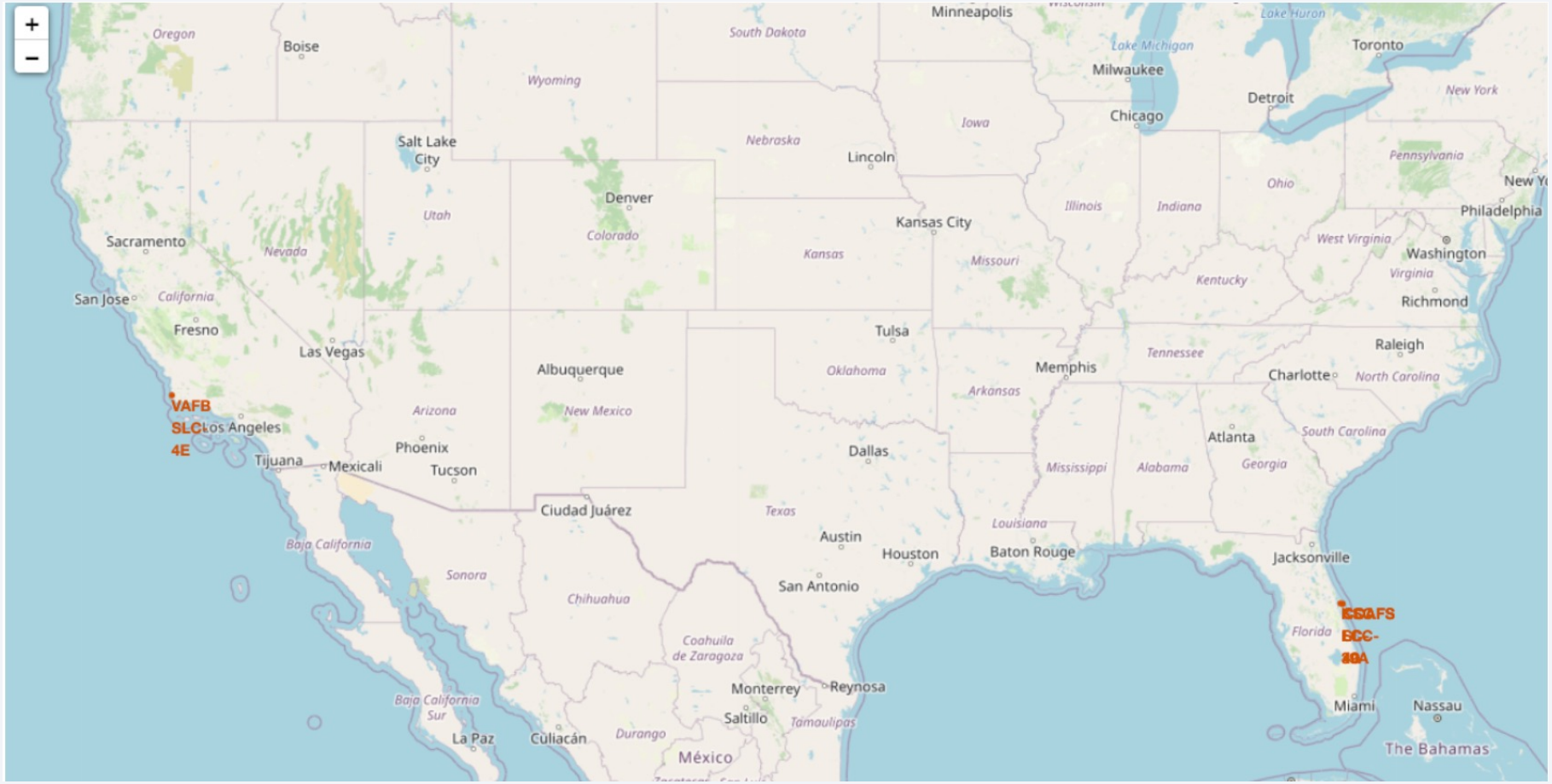
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Outcome	Count
Success (drone ship)	9
Success (ground pad)	6
Failure (drone ship)	5
No attempt (ocean)	4
Failure (parachute)	1

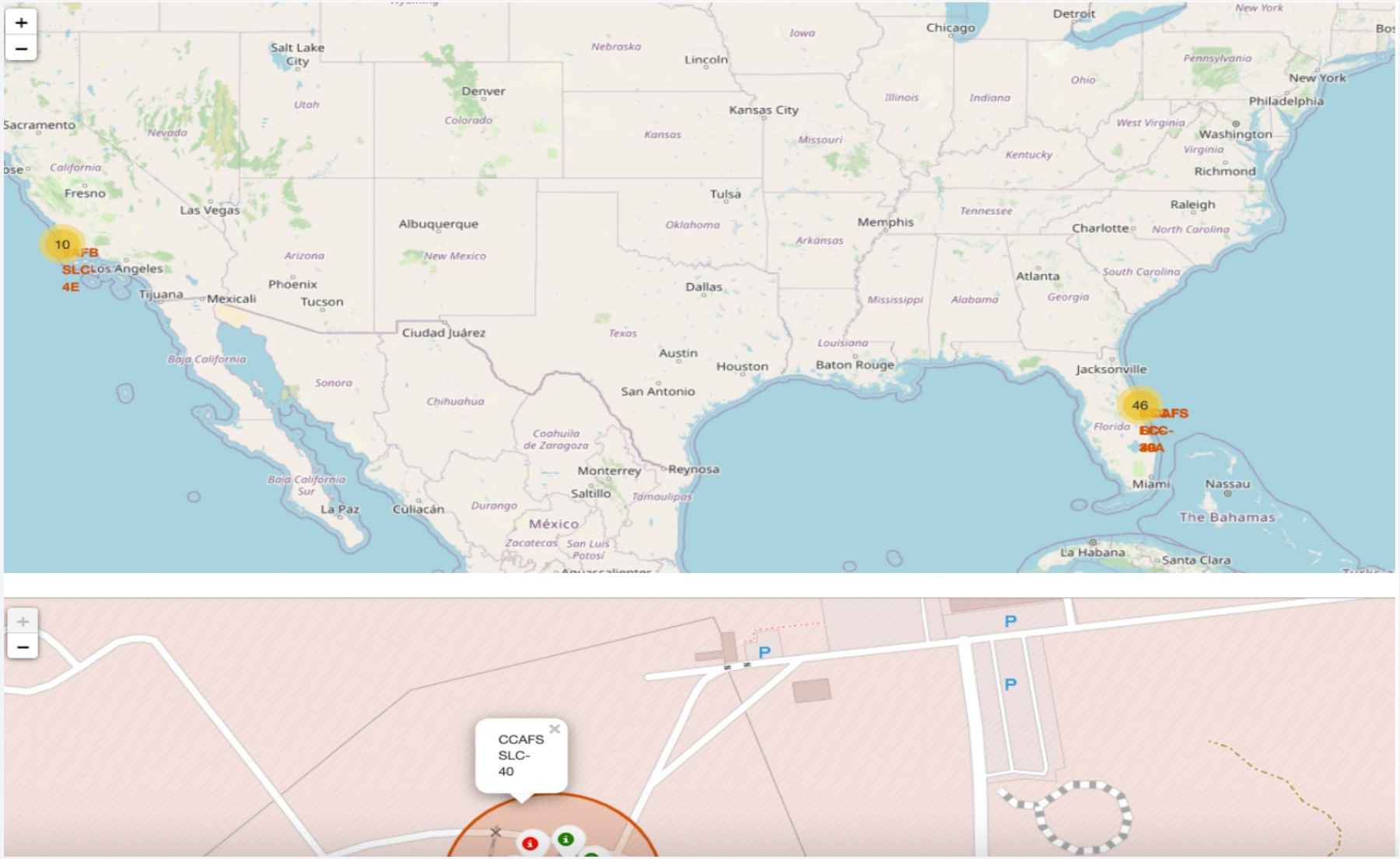
Explanation:
Most successful landings occurred on **drone ships**, followed by **ground pads**.
Early failures were key learning steps that enabled SpaceX's reliable recovery system by 2017.


Section 3
Launch Sites
Proximities Analysis

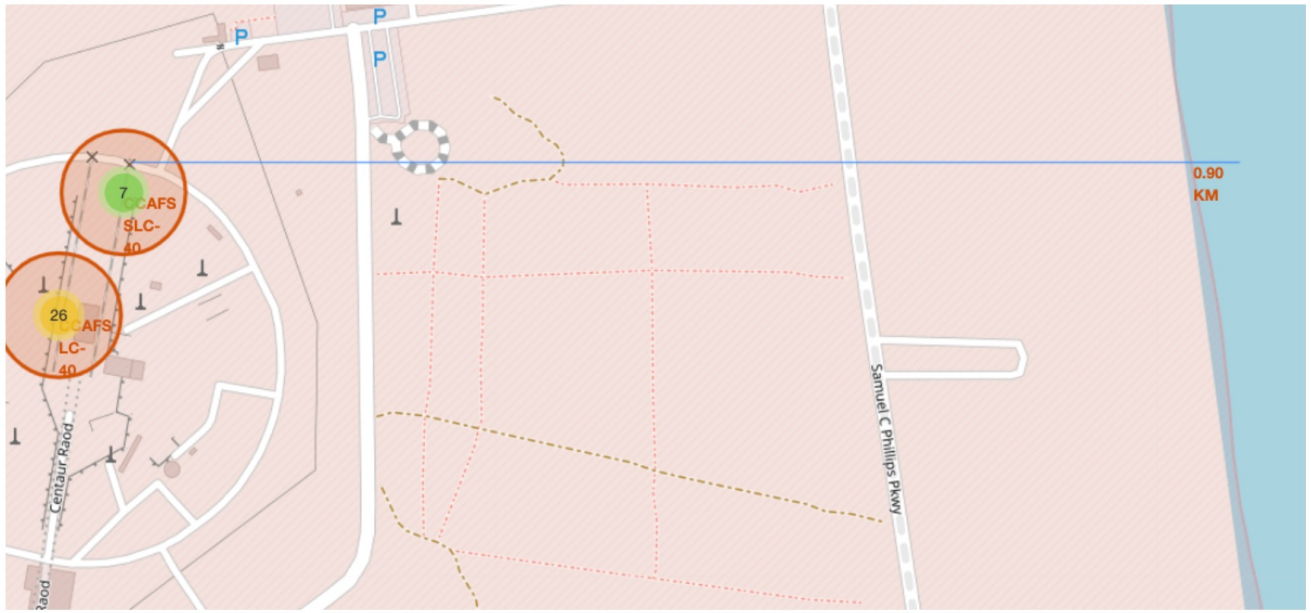
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



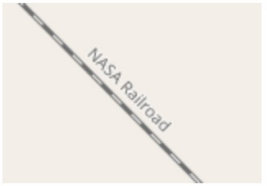
<Folium Map Screenshot 3>



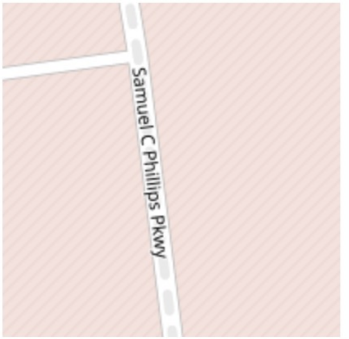
TODO: Similarly, you can draw a line betwee a launch site to its closest city, railway, highway, etc. You need to use `MousePosition` to find the their coordinates on the map first

A railway map symbol may look like this:

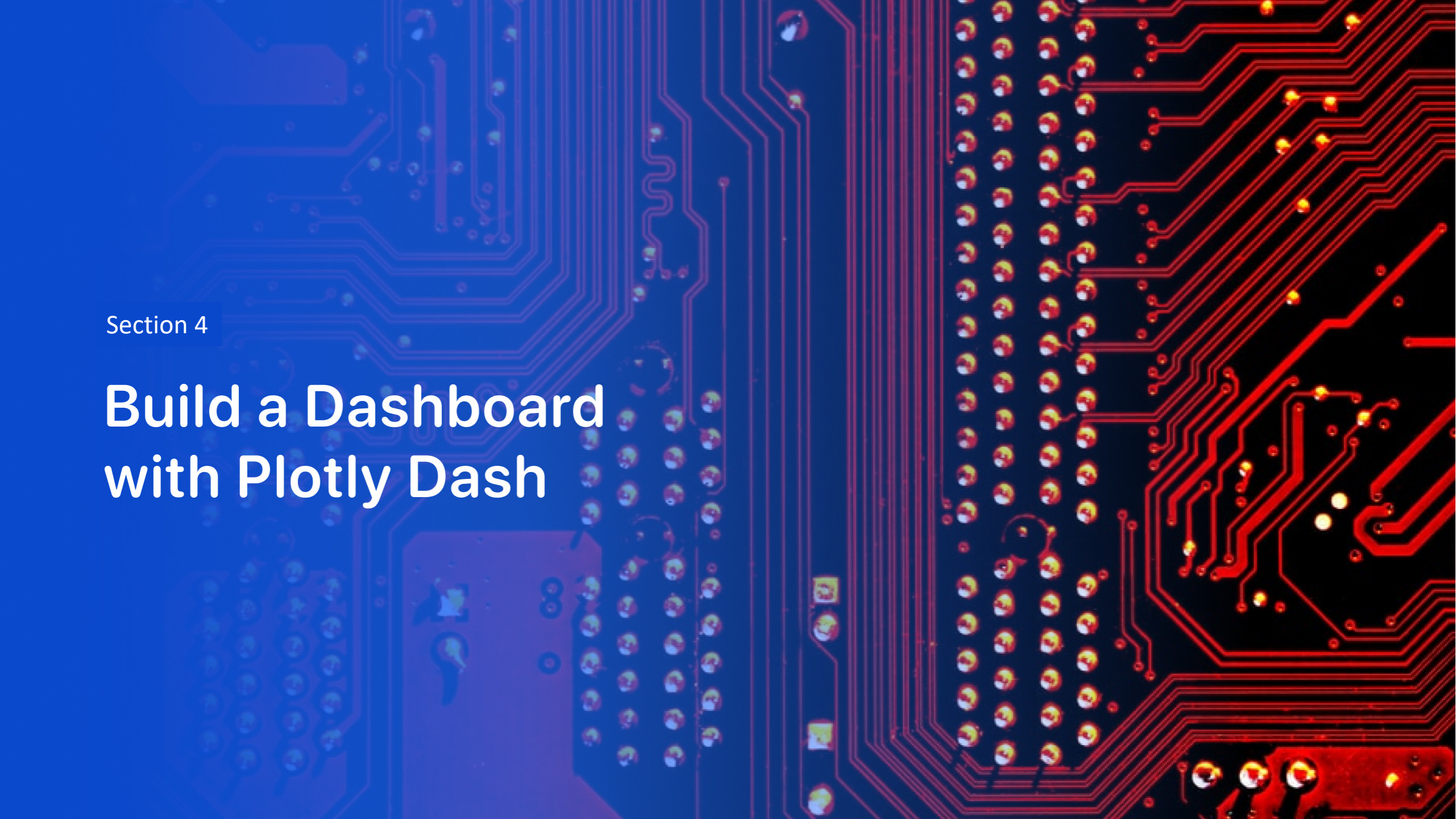
A railway map symbol may look like this:



A highway map symbol may look like this:



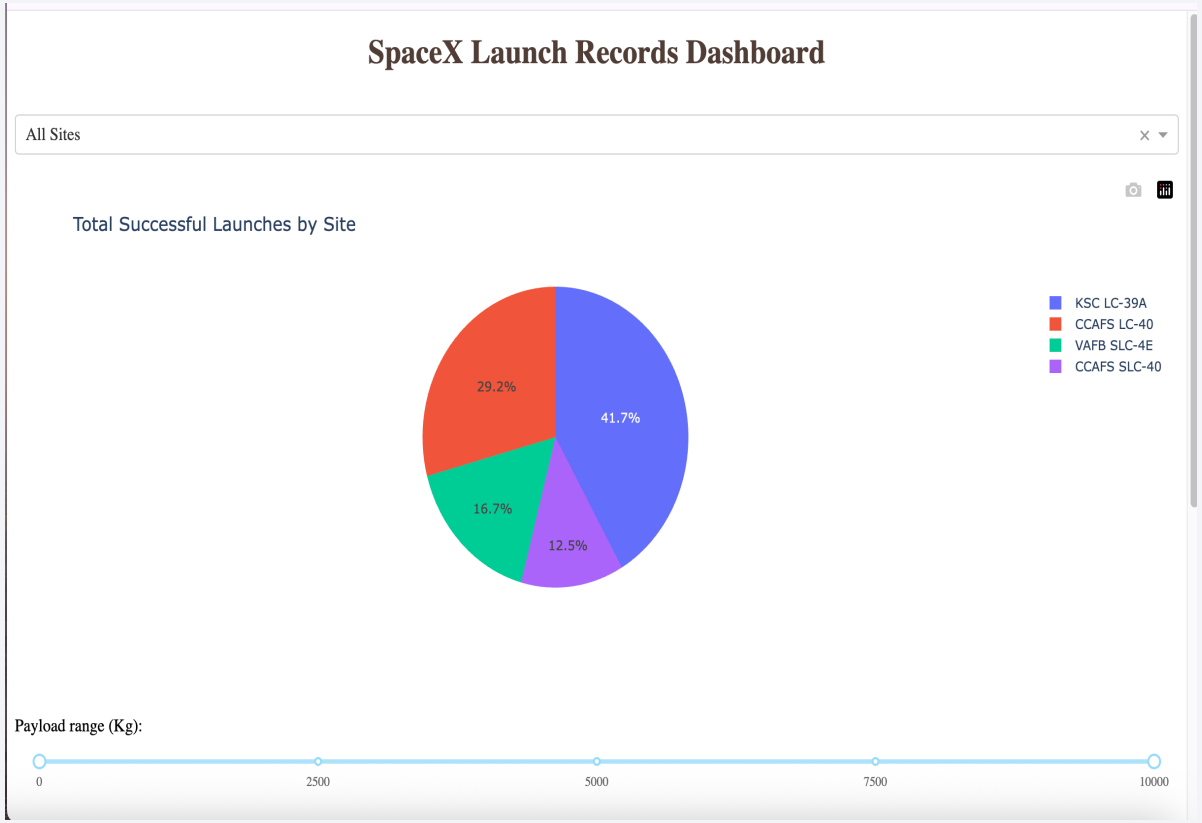
A city map symbol may look like this:


Section 4
Build a Dashboard with Plotly Dash

Launch Success Distribution by Site

- Launch Success Distribution by Site

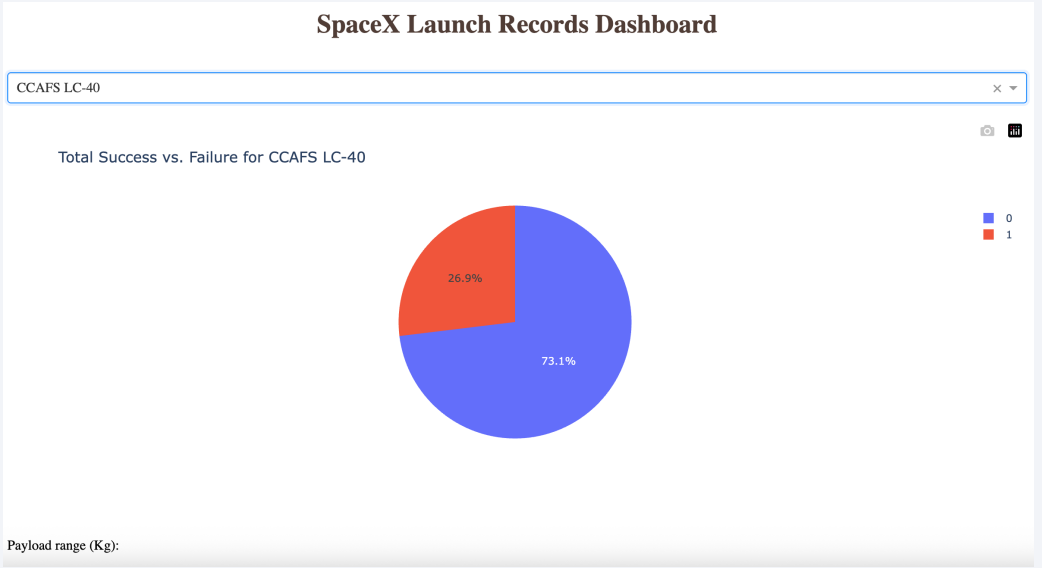
This dashboard shows the distribution of successful SpaceX launches across all sites.
KSC LC-39A has the highest number of successful launches (≈42%), followed by CCAFS LC-40 (≈29%).
It highlights that Florida's sites dominate SpaceX's overall launch success.



Launch Success vs. Failure for CCAFS LC-40

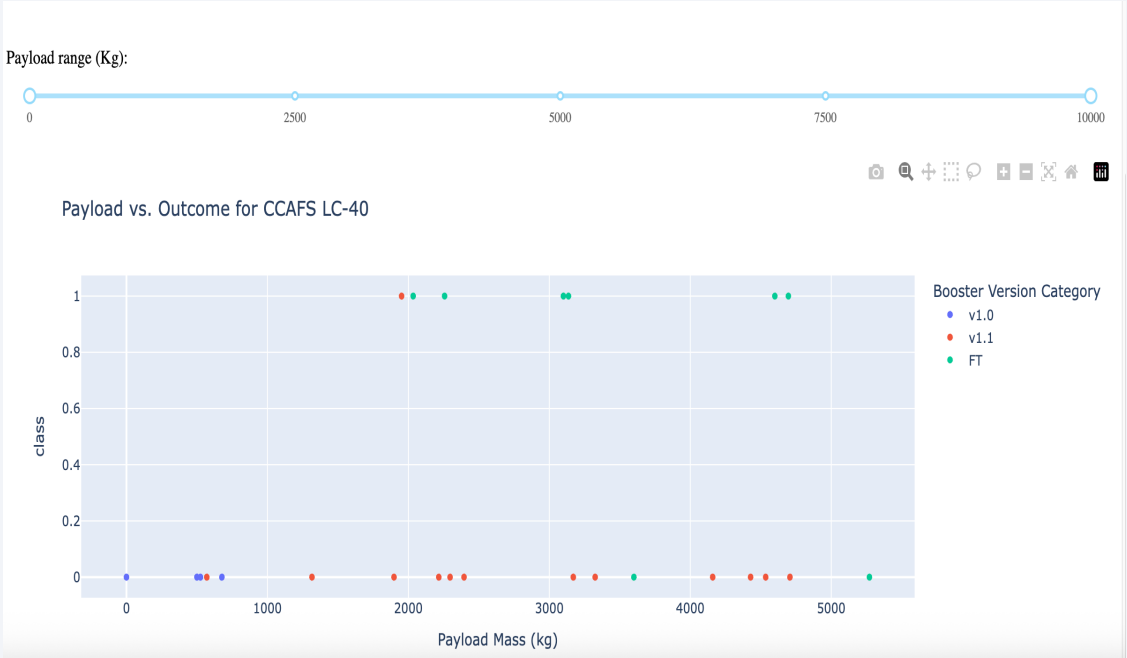
- Launch Success vs. Failure for CCAFS LC-40

This pie chart shows the launch outcomes for the CCAFS LC-40 site, which has one of the highest launch success ratios (≈73%).
The majority of launches were successful, confirming SpaceX's growing reliability over time.



Payload vs. Launch Outcome by Booster Version

- This scatter plot shows how payload mass relates to launch outcomes across different booster versions.
  Heavier payloads (4,000–5,000 kg) launched using **Falcon 9 FT** show higher success rates compared to earlier versions (v1.0 and v1.1).
  This highlights the technological improvements in booster reliability and payload capacity.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

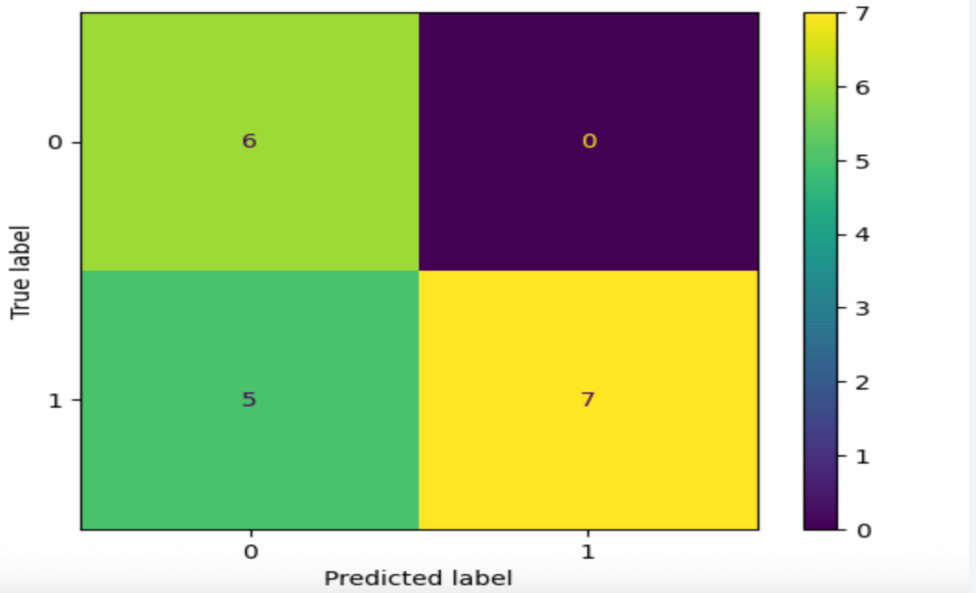
- This confusion matrix shows the performance of the SVM classifier on the test data.
  The model achieved an accuracy of **72.2%**, correctly predicting most of the successful and failed launches.
  While it performs well overall, a few failed launches were misclassified, suggesting slight bias toward predicting success.

Calculate the accuracy on the test data using the method `score`:

```
print("Test accuracy:", svm_cv.score(X_test, Y_test))

yhat = svm_cv.predict(X_test)
disp = ConfusionMatrixDisplay.from_predictions(Y_test, yhat)
plt.show()
```

```
Test accuracy: 0.7222222222222222
```

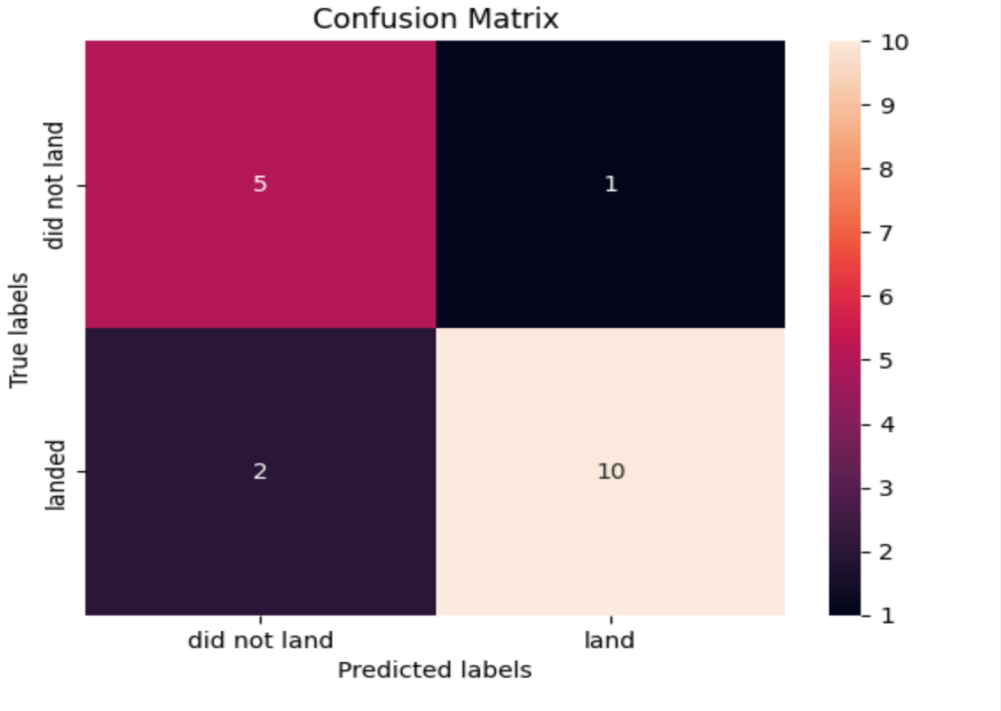


Confusion Matrix

- The K-Nearest Neighbors (KNN) model achieved the **highest accuracy of 83.3%** among all tested models, outperforming SVM, Logistic Regression, and Decision Tree.
  This suggests that KNN is the most suitable model for predicting SpaceX Falcon 9 launch outcomes using the given dataset. Its performance indicates strong capability in distinguishing between successful and failed landings, likely due to well-separated feature patterns in the data.

We can plot the confusion matrix

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



Conclusions

- The **K-Nearest Neighbors (KNN)** model achieved the **highest accuracy (83.3%)**, outperforming SVM, Logistic Regression, and Decision Tree models.
- Successful launches were more common with **medium payloads (2000–6000 kg)** and **newer booster versions (FT, Block 5)**.
- The **Cape Canaveral (CCAFS LC-40)** site recorded the **most frequent launches**, showing consistent performance over time.
- **Machine learning classification** can effectively predict SpaceX Falcon 9 landing outcomes, supporting cost and risk optimization for future missions.
- Further model tuning or ensemble methods (e.g., Random Forest, XGBoost) could improve prediction reliability.

Appendix

- **Additional Assets & Tools Used**
- 🧠 **Python Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Plotly, Folium, Dash
- 💾 **Data Sources:** SpaceX API, Falcon 9 Launch Wiki dataset, CSV exports from project notebooks
- 📊 **Notebooks:**
  - SpaceX-Data wrangling.ipynb
  - SpaceX-EDA-SQL-Query.ipynb
  - SpaceX-Machine Learning Prediction_Part_5.ipynb
- 📈 **Key Visuals Included:**
  - Folium global launch map
  - Orbit type success rate chart
  - Classification accuracy and confusion matrix
- ⚙️ **Environment:** Jupyter Notebook (Python 3.10), IBM Cloud, GitHub repository
- 📁 *"Full code and datasets available on GitHub: ozgunes91/IBM-DataScience-Capstone-SpaceX"*

Thank you!