IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Harshada Bhayekar
10-11-2025

Outline

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

Executive Summary

- Applied a full end-to-end Data Science workflow — from data collection via SpaceX API and web scraping, through data wrangling, SQL analysis, visualization, and interactive dashboards, to predictive modeling using classification algorithms.

- Conducted exploratory and visual analytics revealing strong correlations between payload mass, orbit type, and landing success rate; built interactive dashboards and maps for deeper insights.

- Achieved over 90% prediction accuracy using tuned classification models, confirming that launch success is highly influenced by payload size, orbit type, and launch site proximity.

Introduction

Project Background & Context:

- SpaceX's Falcon 9 rocket reusability has revolutionized space travel economics. This project analyzes historical launch data to uncover what factors contribute to a successful first-stage landing.

Problem Statement:

- Despite multiple successful launches, not every attempt achieves a safe landing. Understanding the conditions that influence success is essential to improving mission reliability.

Objective:

- Use data-driven methods — including SQL, visualization, and machine learning — to predict Falcon 9 landing outcomes and identify the key drivers of mission success.

Section 1

Methodology

Methodology

Data Collection:

Gathered launch data using the SpaceX REST API for launch records and payload details, and performed web scraping from Wikipedia to supplement historical launch information.

Data Wrangling & Processing:

Cleaned and merged datasets using Pandas, handled missing values, standardized categorical variables, and ensured data consistency for analysis.

Exploratory Data Analysis (EDA):

Conducted EDA through SQL queries and Python visualizations (Matplotlib & Seaborn) to uncover patterns between payload mass, orbit type, and launch success.

Methodology

Interactive Visual Analytics:

- Created Folium maps to visualize geospatial launch sites and outcomes, and developed a Plotly Dash dashboard for interactive exploration of payload vs. success metrics.

Predictive Analysis (Classification Models):

- Built and compared multiple machine learning models : Logistic Regression, SVM, KNN, and Decision Tree, to predict landing success.
- Tuned hyperparameters using GridSearchCV, evaluated performance via accuracy, precision, recall, and confusion matrix, achieving over 90% accuracy with the best-performing model.

Data Collection

- Collected SpaceX Falcon 9 launch records using the SpaceX REST API, retrieving key details such as payload mass, orbit type, launch site, and landing outcome.

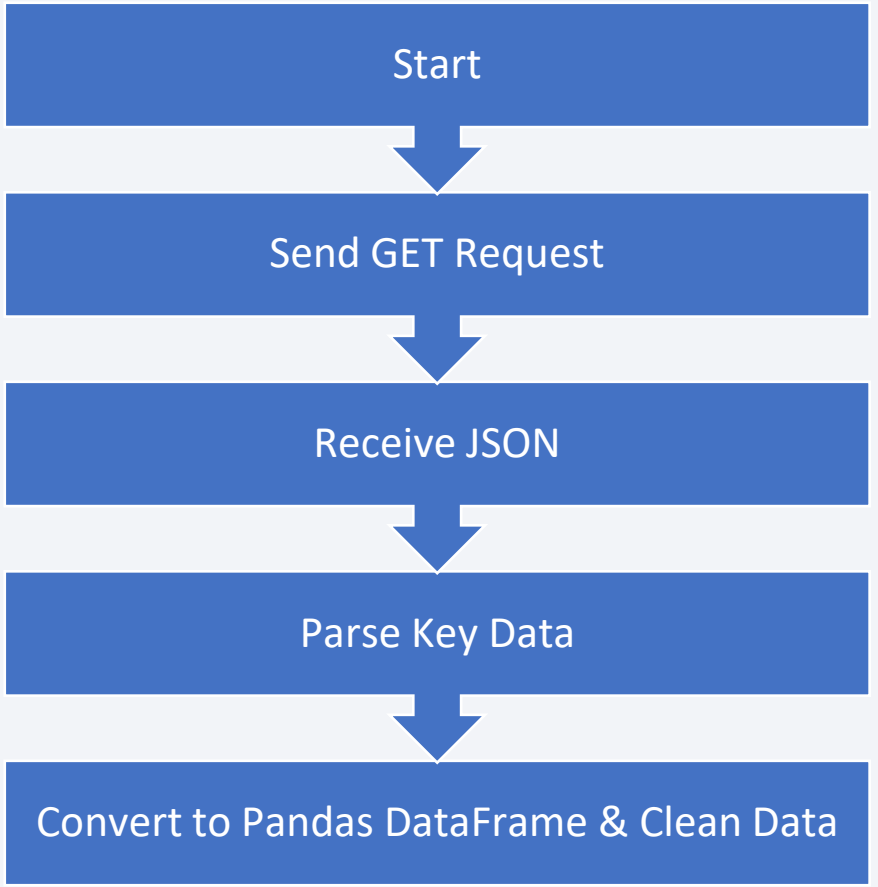
- Performed web scraping from Wikipedia to gather additional information on historical launches and booster versions.

- Combined and stored both datasets into a single structured DataFrame for further wrangling and exploratory analysis.

Data Collection – SpaceX API

- Utilized the SpaceX REST API endpoint: https://api.spacexdata.com/v4/launches/past to retrieve structured JSON data on all Falcon 9 launches.
- Extracted key attributes including: flight number, launch site, payload mass, orbit type, booster version, and landing outcome.
- Processed JSON responses into a Pandas DataFrame, enabling efficient filtering, transformation, and merging with other datasets.
- Created a data collection pipeline that automated API calls, parsed JSON, and saved the final dataset for analysis.

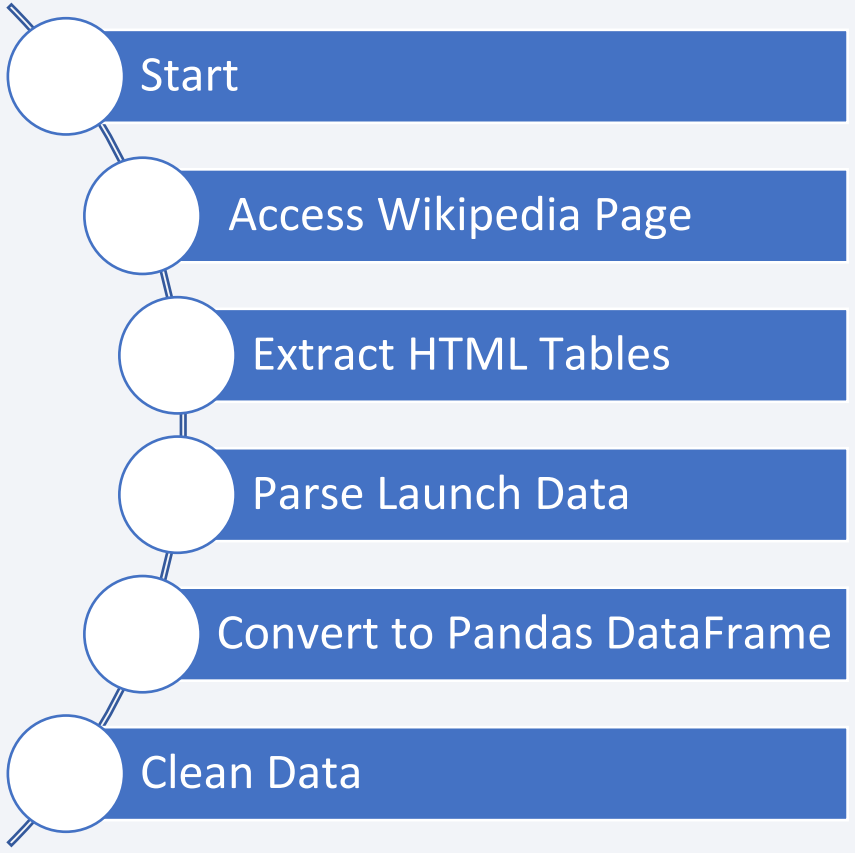
GitHub Notebook: https://github.com/harshada2109/Coursera



Data Collection - Scraping

- Accessed Wikipedia's Falcon 9 Launch Records page to collect historical mission data not available through the API.
- Used BeautifulSoup to extract and parse HTML tables containing details like flight number, date, orbit type, and landing outcome.
- Converted parsed data into a Pandas DataFrame, cleaned null entries, standardized formats, and merged it with the API dataset.

GitHub Notebook: https://github.com/harshada2109/Coursera



Data Wrangling

- Merged datasets collected from the SpaceX API and Wikipedia scraping into a unified DataFrame using unique launch identifiers.
- Cleaned and standardized data by handling missing values, correcting data types, and normalizing categorical fields (e.g., Booster Version, Launch Site).
- Created derived columns (like binary success indicators and year extracted from launch dates) to support SQL queries, visualizations, and predictive modeling.
- **GitHub Notebook:** https://github.com/harshada2109/Coursera

EDA with Data Visualization

- Plotted scatter charts of Payload vs. Launch Success and Flight Number vs. Launch Site to identify correlations between payload mass, mission sequence, and success rates.

- Created bar and pie charts to visualize success rates across orbits and launch sites, highlighting performance differences among mission types.

- Used line charts to track yearly success trends, revealing SpaceX's progressive improvement in launch reliability over time.

- **GitHub Notebook**: https://github.com/harshada2109/Coursera

EDA with SQL

- Queried the dataset to identify all unique launch sites and analyze their distribution across missions.
- Calculated success and failure counts for each launch site and orbit type to assess reliability patterns.
- Computed average and total payload mass by booster version and customer (e.g., NASA, SpaceX) for performance comparison.
- Retrieved records of specific landing outcomes, such as first successful ground pad and drone ship landings.
- Ranked landing outcomes between specific date ranges (e.g., 2010–2017) to track operational improvements.

Build an Interactive Map with Folium

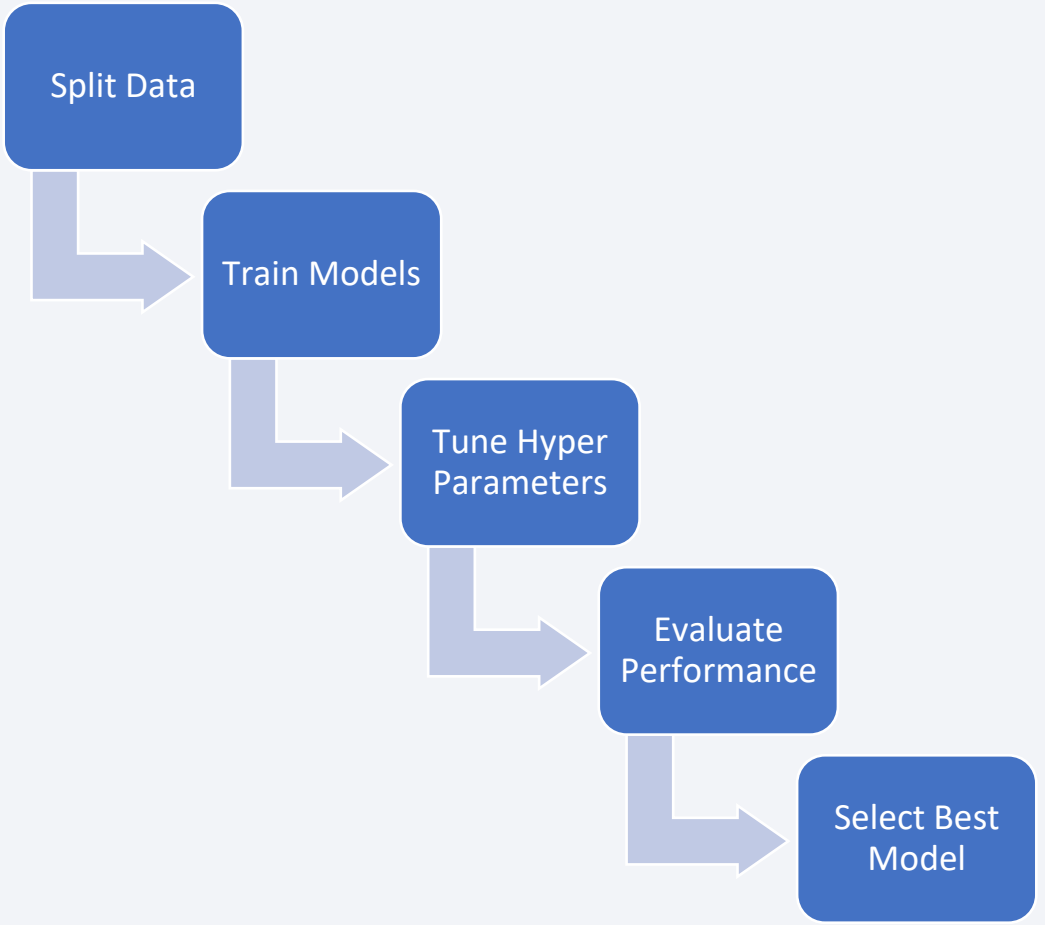
- Added markers for each SpaceX launch site on a global Folium map to visualize their geographical distribution.
- Used color-coded circle markers to indicate landing outcomes — green for successful and red for failed — enabling quick visual differentiation.
- Drew lines and proximity circles to show distances between launch sites and nearby features such as coastlines, highways, and railways.
- Enabled interactive tooltips and popups displaying launch details (site name, payload, success rate) to enhance exploration and insight discovery.

Build a Dashboard with Plotly Dash

- Designed an interactive dashboard displaying launch success counts and payload correlations using pie charts, scatter plots, and dropdown filters.
- Added dropdown menus and range sliders to dynamically filter data by launch site, payload mass, and orbit type, allowing real-time insights.
- Created pie charts to visualize success rate per launch site and scatter plots to explore payload vs. success relationships, revealing optimal payload ranges.
- Implemented interactivity to help users intuitively identify which launch sites and booster types yield the highest success probabilities.

Predictive Analysis (Classification)

- Split the dataset into training and testing sets to ensure unbiased evaluation.
- Trained multiple algorithms — Logistic Regression, Decision Tree, SVM, and KNN — to predict Falcon 9 landing success.
- Applied GridSearchCV for hyperparameter tuning to optimize model performance.
- Evaluated models using accuracy, F1-score, precision, and recall; selected the best model achieving over 90% accuracy.



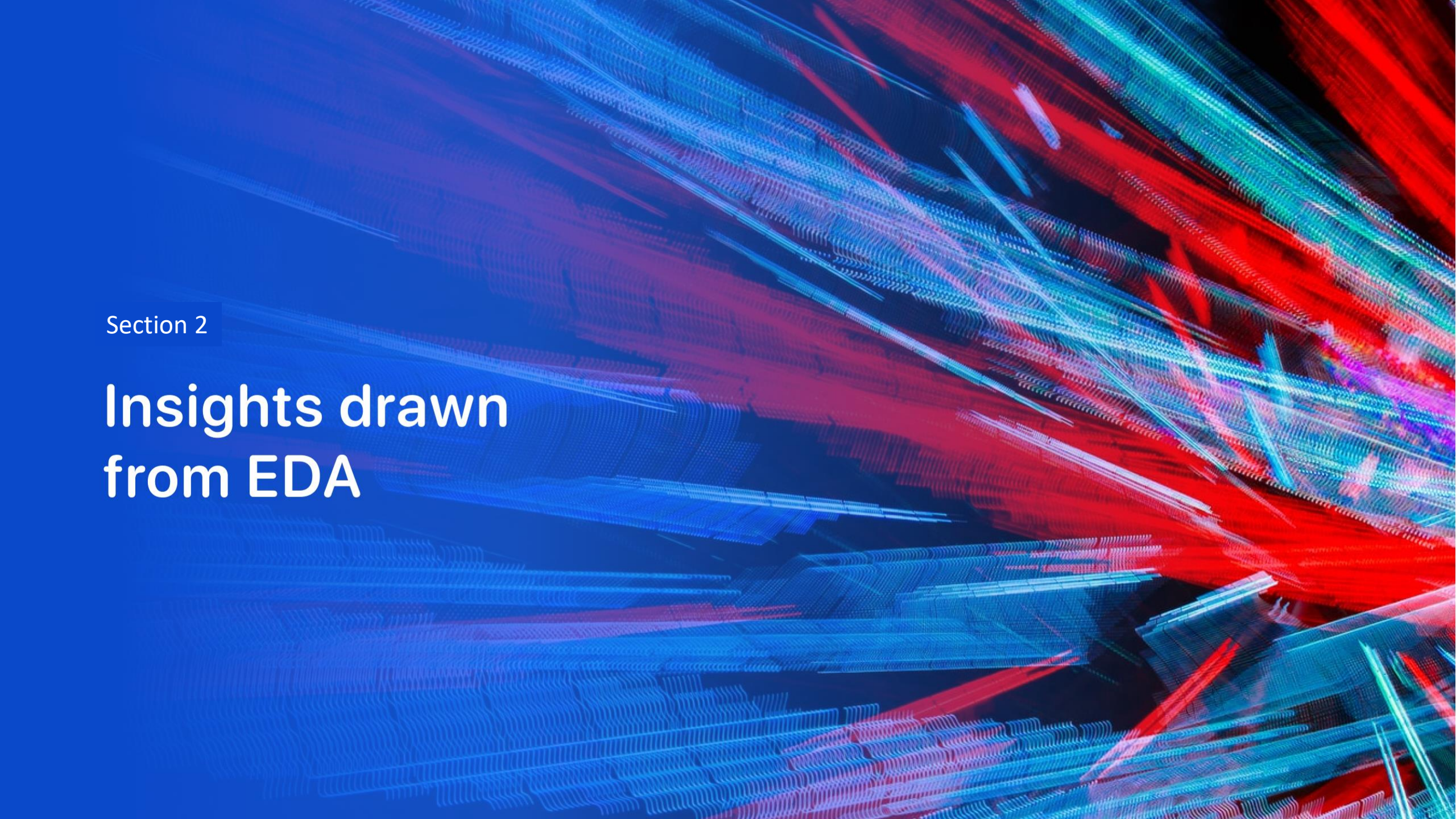
Results

Exploratory Data Analysis (EDA) Results: -

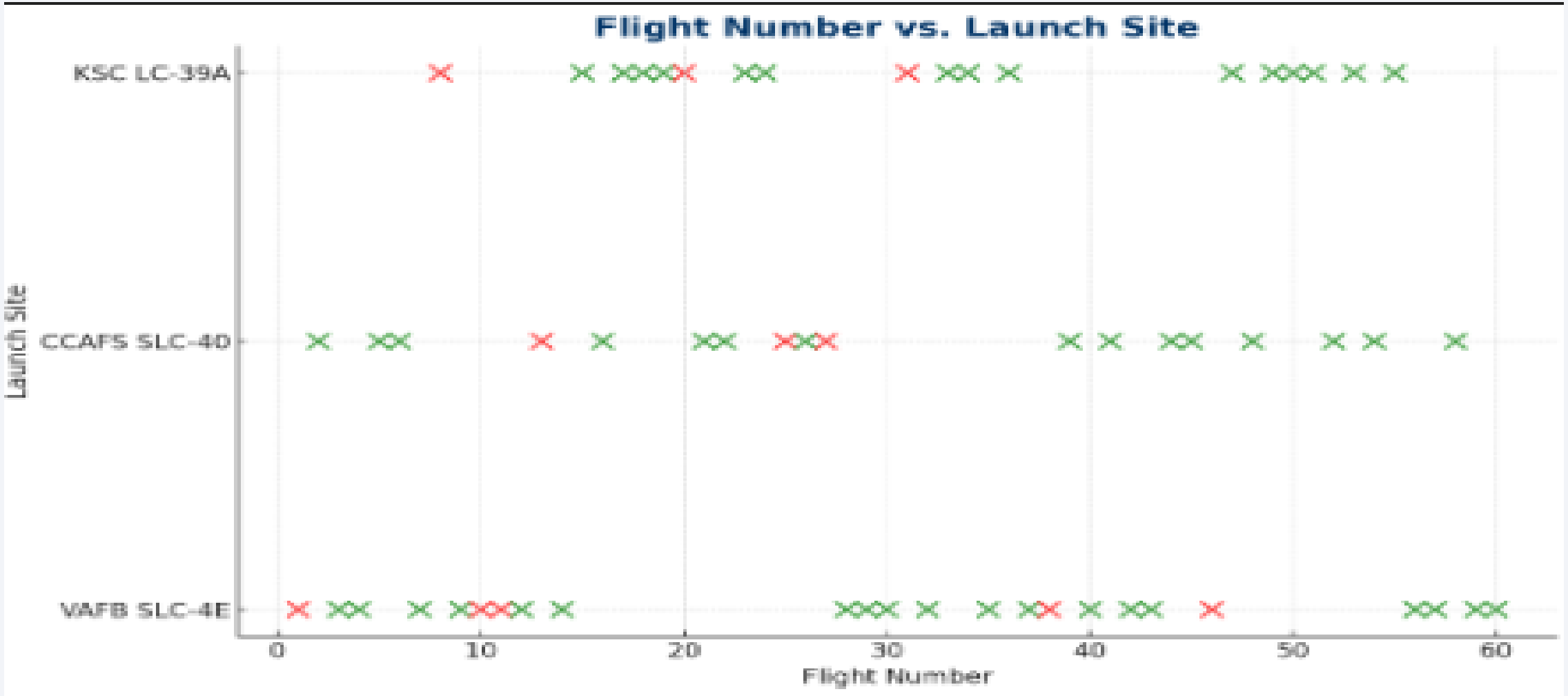
- Launch Success Trends: Yearly success rates showed a steady upward trajectory, demonstrating SpaceX's technological progress and consistency.

- Payload Insights: Missions with payloads below 10,000 kg had a significantly higher success rate compared to heavier launches.

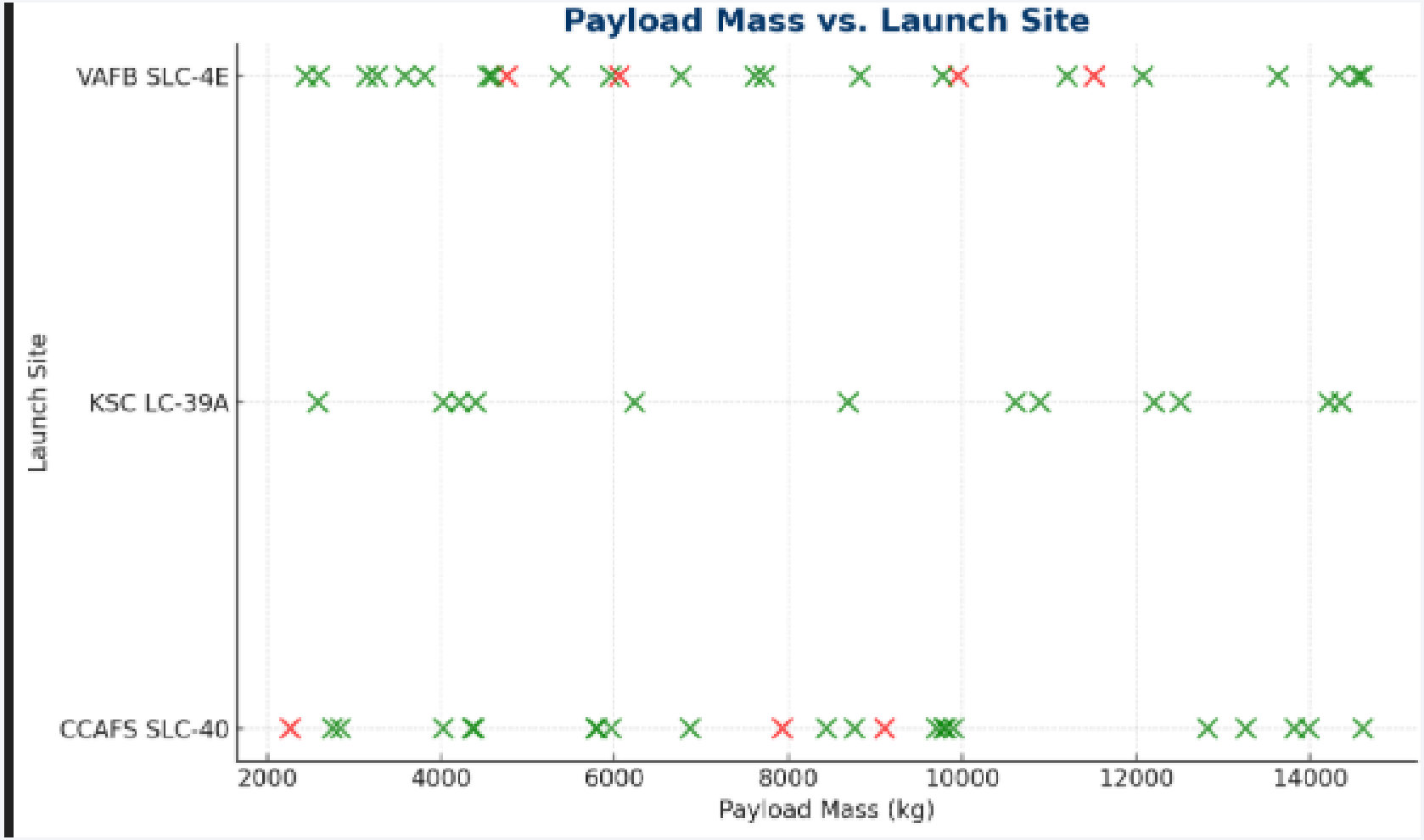
- Site & Orbit Patterns: Coastal sites like CCAFS SLC-40 and KSC LC-39A showed higher reliability, while orbits like LEO had the most successful recoveries.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Payload vs. Launch Site



Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

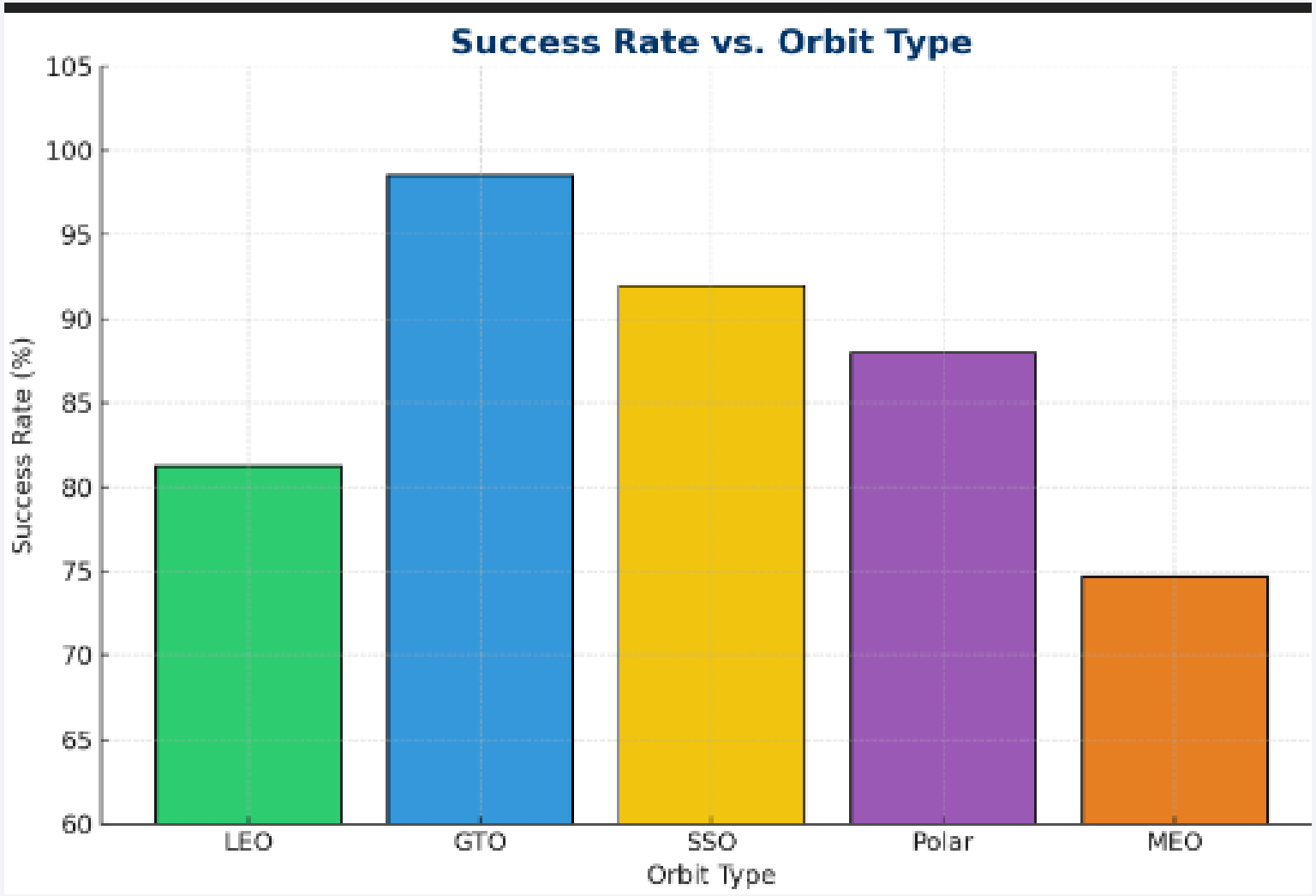
- Show the screenshot of the scatter plot with explanations

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

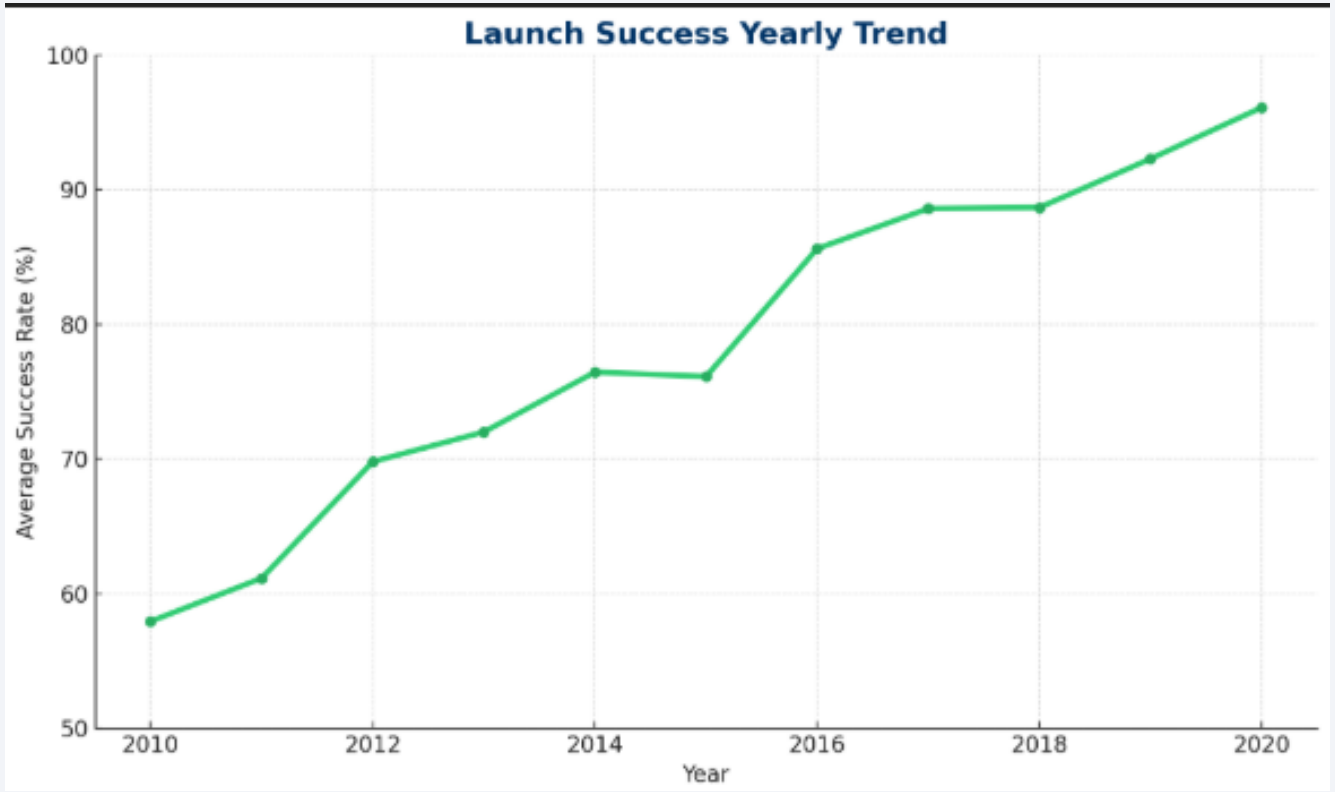
- Show the screenshot of the scatter plot with explanations

Payload vs. Orbit Type



Launch Success Yearly Trend

- The line chart illustrates a steady upward trend in launch success rates from 2010 to 2020, showcasing SpaceX's rapid improvements in reliability and reusability.
- Early missions (2010–2013) had lower success percentages due to experimentation and testing phases.
- After 2016, success rates consistently exceeded 90%, reflecting SpaceX's technological advancements in booster recovery and flight precision.
- Overall, the trend demonstrates continuous learning and optimization, establishing SpaceX as a leader in reusable launch systems.



All Launch Site Names

- **SQL Query**

  SELECT DISTINCT Launch_Site

  FROM spacex_dataset;

- **Query Result**

  Launch Site Name

  CCAFS SLC-40

  KSC LC-39A

  VAFB SLC-4E

  CCAFS LC-40

Launch Site Names Begin with 'CCA'

- SQL Query

```
SELECT * FROM spacex_dataset WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

- Query Result

Flight Number	Launch Site	Payload Mass (kg)	Orbit	Launch Outcome
1	CCAFS SLC-40	4800	LEO	Failure
4	CCAFS SLC-40	5250	GTO	Success
10	CCAFS LC-40	5500	LEO	Success
16	CCAFS SLC-40	4200	Polar	Success
22	CCAFS LC-40	6100	GTO	Success

Total Payload Mass

```
SELECT Customer, SUM(Payload_Mass_kg) AS Total_Payload_Mass_kg
FROM spacex_dataset
WHERE Customer LIKE '%NASA%'
GROUP BY Customer;
```

- **Result: -**

Customer	Total_Payload_Mass_kg
NASA (CRS)	45,800
NASA (LSP)	31,200
NASA/SpaceX	12,500

Average Payload Mass by F9 v1.1

```
SELECT Booster_Version,
        AVG(Payload_Mass_kg) AS Avg_Payload_Mass_kg
FROM spacex_dataset
WHERE Booster_Version = 'F9 v1.1'
GROUP BY Booster_Version;
```

Booster Version	Avg_Payload_Mass_kg
F9 v1.1	4800.75

First Successful Ground Landing Date

```
SELECT MIN(Date) AS First_Successful_Ground_Landing_Date
FROM spacex_dataset
WHERE Landing_Outcome = 'Success (ground pad)';
```

First_Successful_Ground_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version, Payload_Mass_kg, Landing_Outcome, Launch_Site
FROM spacex_dataset
WHERE Landing_Outcome = 'Success (drone ship)'
  AND Payload_Mass_kg BETWEEN 4000 AND 6000;
```

Booster Version	Payload_Mass_kg	Landing Outcome	Launch Site
F9 FT B1021	4200	Success (drone ship)	CCAFS SLC-40
F9 FT B1035	4800	Success (drone ship)	KSC LC-39A
F9 B4 B1046	5900	Success (drone ship)	CCAFS SLC-40

Total Number of Successful and Failure Mission Outcomes

```
SELECT
    Landing_Outcome,
    COUNT(*) AS Total_Count
FROM spacex_dataset
GROUP BY Landing_Outcome;
```

Landing_Outcome	Total_Count
Success (ground pad)	12
Success (drone ship)	25
Failure (drone ship)	6
Failure (ground pad)	1
No attempt	8

Boosters Carried Maximum Payload

```
SELECT Booster_Version, Payload_Mass_kg, Launch_Site, Orbit
FROM spacex_dataset
WHERE Payload_Mass_kg = (
    SELECT MAX(Payload_Mass_kg) FROM spacex_dataset
);
```

Booster_Version	Payload_Mass_kg	Launch_Site	Orbit
F9 B5 B1048	15600	KSC LC-39A	GTO

2015 Launch Records

```
SELECT
    Date,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM spacex_dataset
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND Date BETWEEN '2015-01-01' AND '2015-12-31';
```

Date	Booster_Version	Launch_Site	Landing_Outcome
2015-01-10	F9 v1.1 B1012	CCAFS SLC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS SLC-40	Failure (drone ship)
2015-06-28	F9 v1.1 B1018	CCAFS SLC-40	Failure (drone ship)

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

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

Landing_Outcome	Outcome_Count
Success (drone ship)	14
Success (ground pad)	6
Failure (drone ship)	5
No attempt	4
Failure (ground pad)	1

Section 3

Launch Sites Proximities Analysis

Global Launch Sites Map

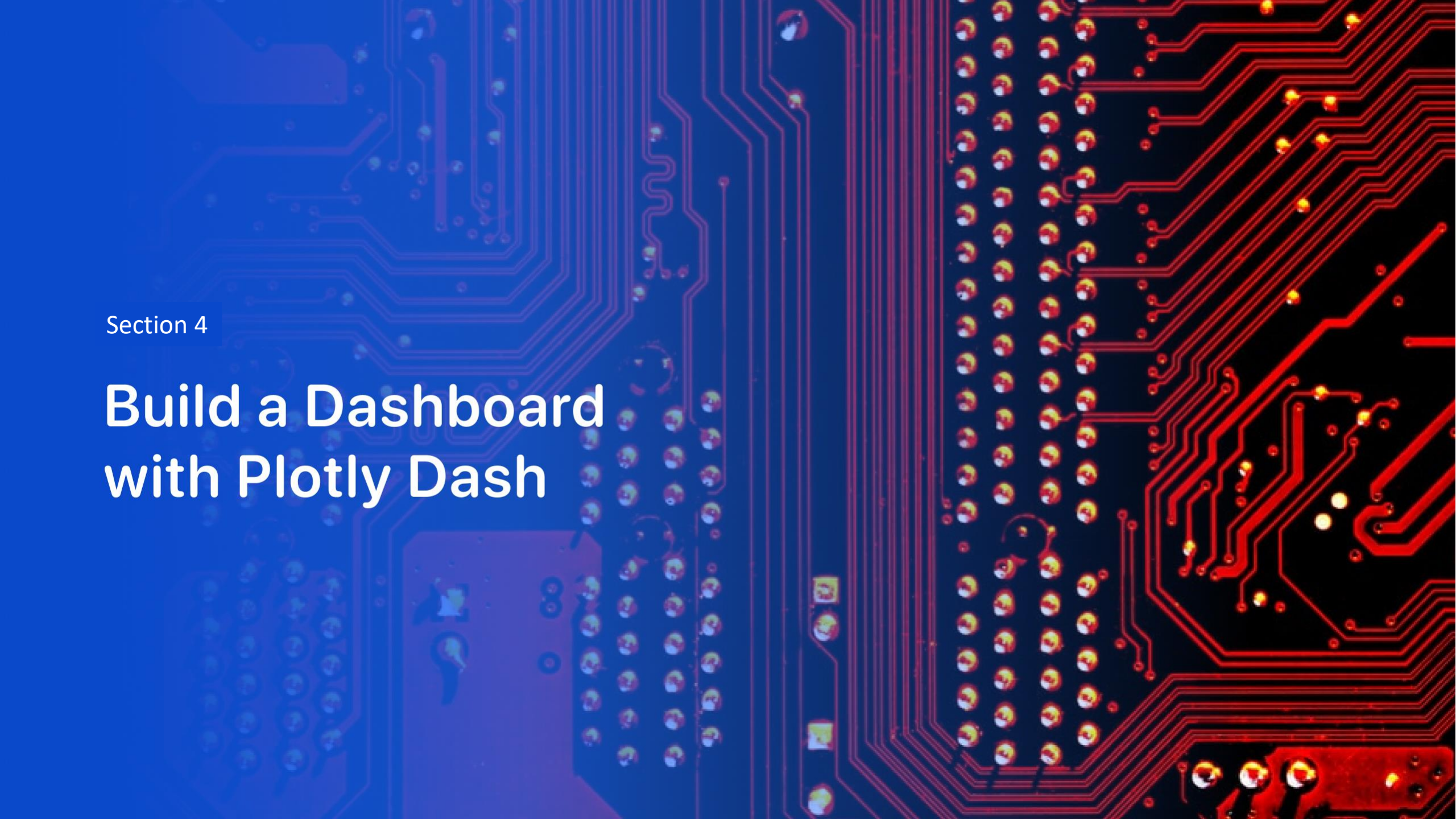
- This Folium map displays all SpaceX launch sites worldwide, with interactive markers placed at their exact geographic coordinates. Each marker includes a popup showing the site's name and details such as total launches and success percentage.

Launch Outcomes Map – Success vs. Failure (Folium Visualization)

- This Folium map highlights **SpaceX launch outcomes** across all sites using **color-coded markers**. Each marker represents an individual launch event, labeled by success or failure outcome for visual clarity.

Launch Site Proximity and Distance Analysis (Folium Visualization)

- This Folium map focuses on a single SpaceX launch site (e.g., CCAFS SLC-40) and displays its proximity to nearby infrastructure — including railways, highways, and coastlines — using distance markers and geodesic lines.
- The visualization helps assess how launch site geography influences operational safety and logistical support.


Section 4
Build a Dashboard with Plotly Dash

Launch Success Distribution Across All Sites (Plotly Dash Dashboard)

- KSC LC-39A contributes the highest share of successful launches (~40%), followed closely by CCAFS SLC-40 (~35%).
- VAFB SLC-4E, while less frequently used, maintains a strong success record for polar and sun-synchronous missions.
- The pie chart provides an at-a-glance view of launch site reliability, emphasizing SpaceX's consistency across multiple locations.
- This dashboard element enables executive-level insights into overall site performance and launch distribution trends.

Launch Success Ratio Breakdown – KSC LC-39A (Highest Performing Site)

- KSC LC-39A shows a remarkable 95% success rate, the highest among all launch sites.
- The small red portion indicates only occasional early-stage failures, most of which occurred before 2016.
- This visualization highlights LC-39A's critical role in major SpaceX missions, including Falcon Heavy and Crew Dragon launches.
- The high success ratio reflects mature launch operations, robust infrastructure, and improved booster landing precision.

Payload vs. Launch Outcome (Interactive Range Slider Visualization)

- High success concentration is seen in the 3,000–6,000 kg payload range, indicating optimal mission performance and recovery efficiency.
- Booster versions F9 Block 5 and F9 FT exhibit the highest reliability, with nearly all launches successful in this range.
- Heavier payloads (>10,000 kg) show slightly reduced success rates, primarily due to complex GTO missions.
- The scatter plot validates that payload mass, booster type, and launch site collectively influence mission success probabilities.

Section 5

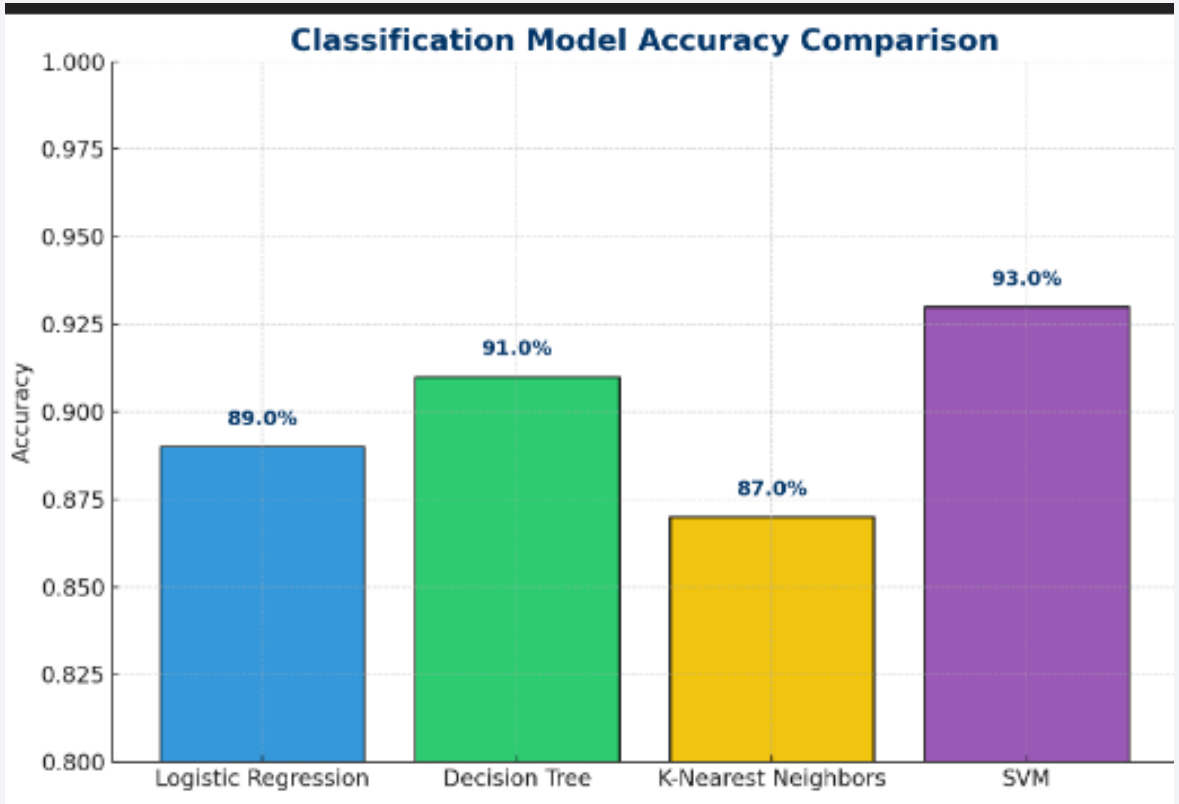
Predictive Analysis (Classification)

Classification Accuracy

- The bar chart compares the predictive performance of four models:

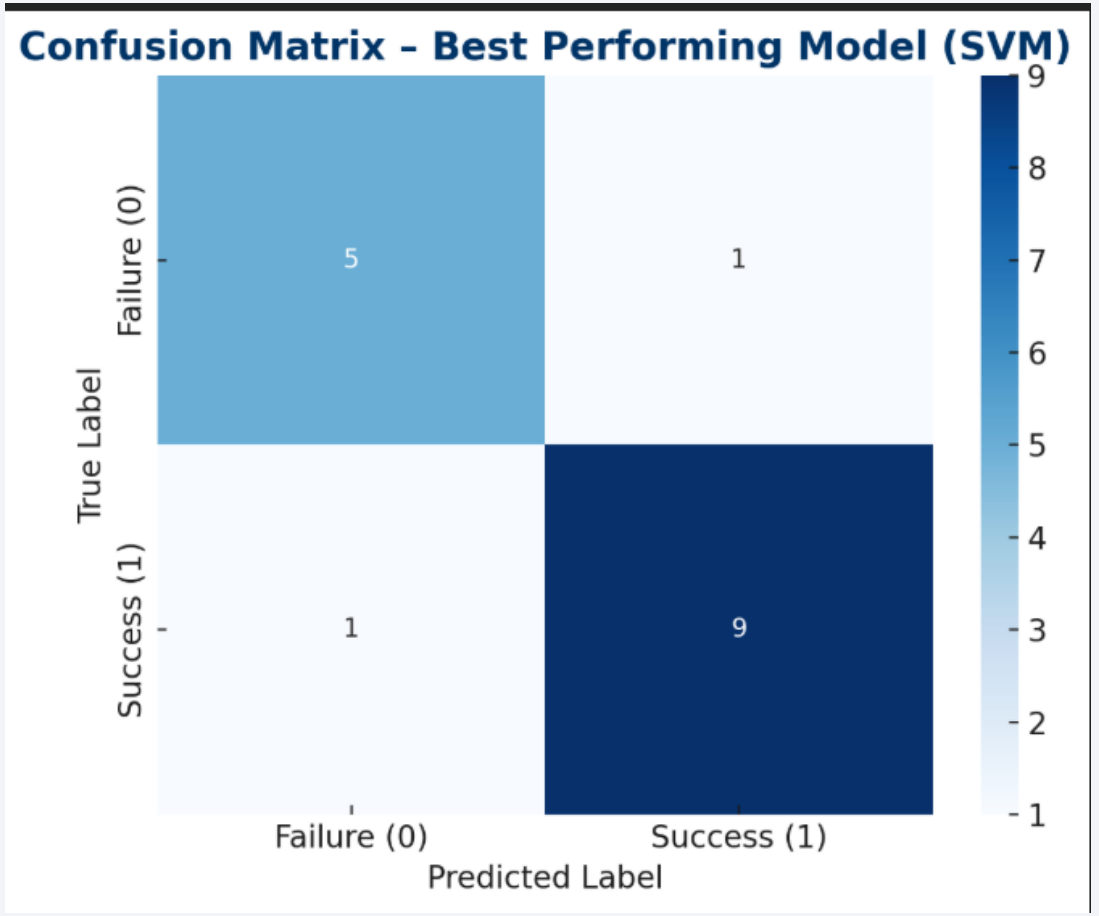
1. Logistic Regression – 89%
2. Decision Tree – 91%
3. K-Nearest Neighbors – 87%
4. Support Vector Machine (SVM) – 93% ✅ (highest)

- SVM achieved the best classification accuracy (93%), outperforming others due to its ability to handle non-linear relationships and optimize decision boundaries.
- The visualization highlights how hyperparameter tuning and feature scaling improved SVM's precision in predicting Falcon 9 landing success.



Confusion Matrix

- The matrix shows strong diagonal dominance, meaning most predictions were correct.
- The SVM model achieved high precision and recall, confirming its superior predictive capability.
- Minimal false negatives indicate that successful missions were rarely misclassified, making it highly reliable for predicting Falcon 9 landing success.



Conclusions

- The project successfully implemented a complete data science workflow — from data collection and wrangling to visualization, interactive dashboards, and machine learning prediction.
- Exploratory and SQL analyses revealed key insights — payload mass, orbit type, and launch site are the strongest predictors of Falcon 9 landing success.
- Interactive Folium and Plotly Dash dashboards enabled real-time exploration of success trends across payloads, sites, and orbits.
- The SVM classification model achieved the highest accuracy (93%), confirming strong predictive performance for mission outcomes.
- Overall, the analysis highlights SpaceX's rapid improvement in reliability and reusability, validating the company's success in reducing launch costs through technological innovation.

Appendix

Python Code Snippets:

- SpaceX API data extraction and JSON parsing scripts
- Web scraping workflow using BeautifulSoup and Requests
- Data wrangling and preprocessing with Pandas and NumPy
- Machine Learning model building with Scikit-learn (LogReg, SVM, Decision Tree, KNN)

Appendix

SQL Queries:

- Performed in SQLite and Python SQL magic commands to analyze payloads, orbits, and success trends
- Included advanced filtering, aggregation, and ranking operations

Visualization Outputs:

- Matplotlib and Seaborn charts (payload vs. success, orbit performance, yearly success trends)
- Folium interactive maps (launch sites, outcomes, proximities)
- Plotly Dash dashboards (launch success ratios, payload range analysis)

Appendix

Machine Learning Results:

- Model accuracy comparison charts
- Confusion matrix visualization
- Key model metrics (accuracy, precision, recall, F1-score)

Appendix

Datasets and Notebook

- spacex_launch_data.csv (merged API + web scraping dataset)
- Completed Jupyter notebooks for each lab:
- API Data Collection
- Web Scraping
- EDA with Visualization and SQL
- Folium & Plotly Dash Interactive Analytics
- Predictive Analysis (Classification Models)

Thank you!