



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Joko Eliyanto  
June 8<sup>th</sup>, 2025



# Outline

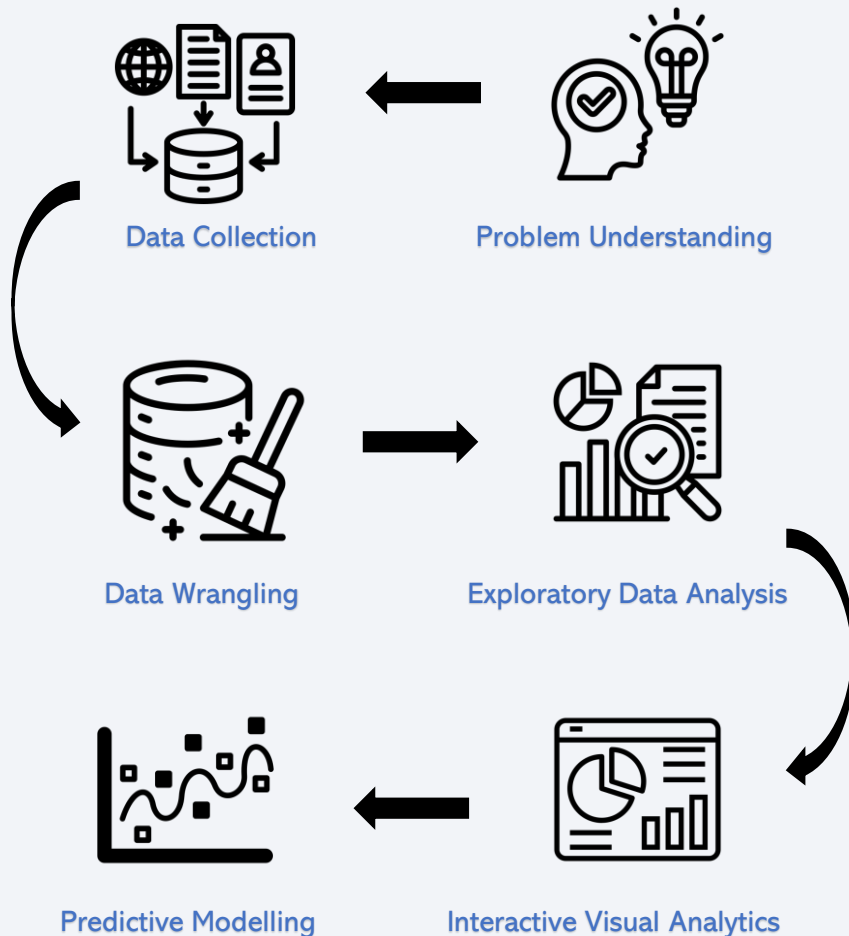
---

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

# Executive Summary

---

## Methodologies Summary




## Result Summary

The analysis reveals that launches to LEO and ISS orbits have the highest success rates, with heavier payloads typically launched from more reliable sites. Certain orbits like GTO and GEO show increased risk for heavier payloads, while launch success has improved over time due to better technology. Machine learning models confirmed these insights, with the Decision Tree model achieving the highest accuracy of 94%. Together, visual exploration and modeling provide a strong foundation for optimizing launch decisions, mission planning, and risk management.

# Introduction

CAPABILITIES & SERVICES	
<small>SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low Earth orbit (LEO) destinations. See additional information at <a href="https://www.spacex.com/falcon9">spacex.com/falcon9</a>.</small>	
PRICE*	FALCON 9
STANDARD PAYMENT PLAN (through 2025)	<b>\$69.85 M</b> Up to 5.5 MT TO GEO
DESTINATION	
LOW-EARTH ORBIT (LEO)	22,000 kg 50,240 lbs
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	8,300 kg 18,300 lbs
PAYLOAD TO MARS	4,020 kg 8,840 lbs

\*Pricing adjustments made in 2025 account for inflation. Missions purchased in 2025 but flown beyond 2025 may be subject to additional adjustments due to inflation. Performance represents best capability with future capabilities subject to change.



## Project background and context

SpaceX offers Falcon 9 rocket launches at a significantly lower cost—\$62 million—thanks to the reusability of the first stage, which alone costs around \$15 million to build. However, depending on mission parameters like payload, orbit, or customer requirements, SpaceX sometimes opts not to recover the first stage.

## Problems to find answers

**Can we accurately predict whether the Falcon 9 first stage will successfully land?** This prediction acts as a proxy to estimate launch cost, which is valuable for alternative providers evaluating competitive bids against SpaceX.



Section 1

# Methodology

# Methodology

---

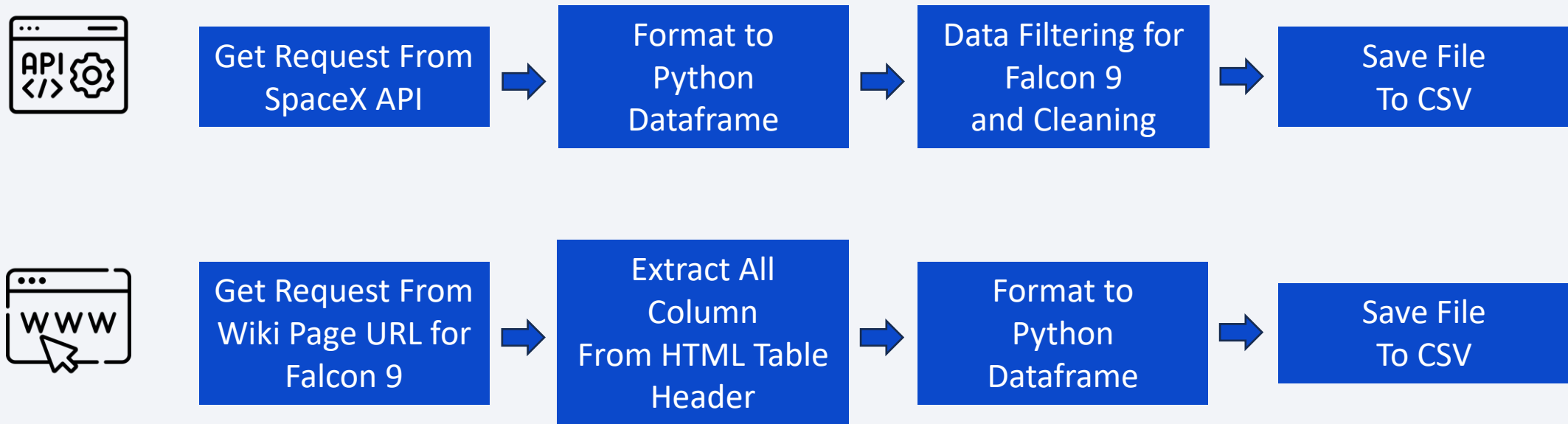
## Executive Summary

- Data collection methodology:
  - Data was collected using both an API call to the SpaceX launch data endpoint and web scraping from the launch success history page. The collected data was then formatted into a DataFrame, cleaned, and saved as a .csv file.
- Perform data wrangling
  - The data was cleaned by removing missing values and duplicates. At this stage, landing success labels were also assigned.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build model using 80% as data training, try 4 different classification model (logistic regression, svm, knn, and decision tree), calculate the accuracy on test data, and plot the confusion matrix for evaluate the model

# Data Collection

---

- Data collected both from spacex api using api request and wiki page url using web scraping technique, the data collected formatted into dataframe, cleaned and save it into csv files



# Data Collection – SpaceX API

## SpaceX Data Collection Pipeline (via REST API) – Key Phrase

### Main API Endpoint:

<https://api.spacexdata.com/v4/launches/past> – for historical launch data

### Data Retrieval & Preprocessing

- Used requests + json\_normalize() to convert JSON into Pandas DataFrame
- Filtered for:
  - Single-core & single-payload launches
  - Launches before Nov 13, 2020
- Extracted key fields: rocket, payloads, launchpad, cores, flight\_number, date\_utc

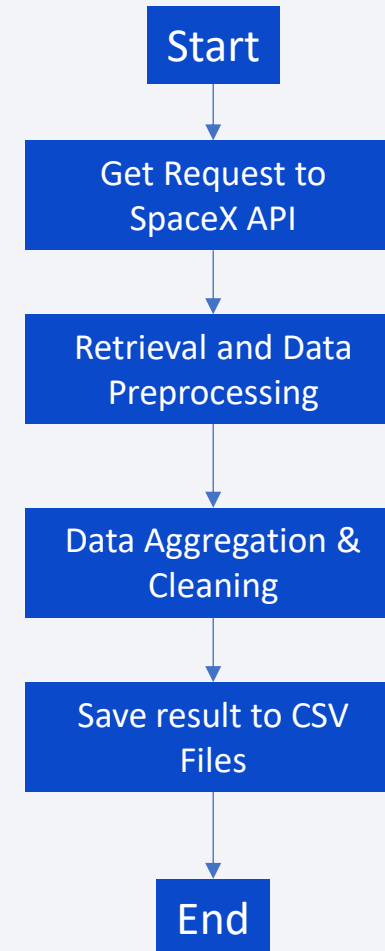
### Secondary API Calls

Used ID references from main data:

- /v4/rockets/ → Booster version
- /v4/launchpads/ → Launch site name, lat, long
- /v4/payloads/ → Payload mass, orbit
- /v4/cores/ → Landing outcome, grid fins, reused, legs, etc.

### Data Aggregation & Cleaning

- Collected results into lists, built final DataFrame
- Dropped Falcon 1 launches
- Reindexed FlightNumber
- Filled missing PayloadMass with mean



[Data Collection – SpaceX API => Notebook](#)



# Data Collection - Scraping

## SpaceX Web Scraping Process – Key Phrases

 **Source:** Wikipedia archive page

[List of Falcon 9 and Falcon Heavy launches](#)

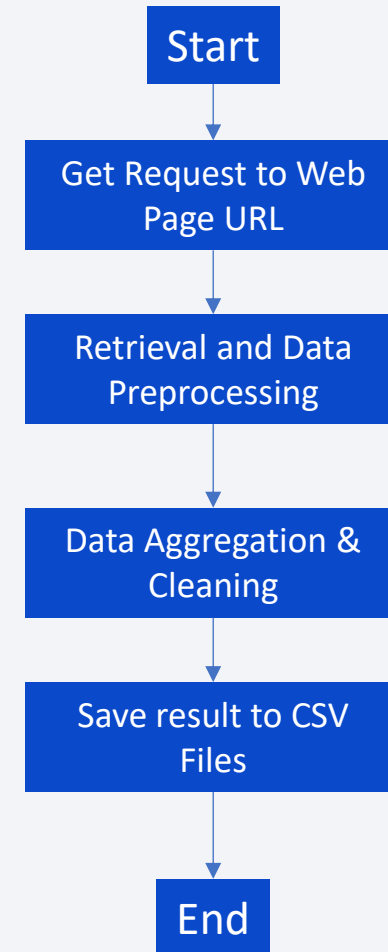
**Target:** Extracted structured launch data from HTML tables with class wikitable plainrowheaders collapsible

### Data Fields Extracted:

- Flight No., Date, Time
- Version Booster, Launch Site, Payload
- Payload mass, Orbit, Customer
- Launch outcome, Booster landing

### Preprocessing:


- Removed irrelevant columns
- Normalized units (e.g. payload mass in kg)
- Cleaned text and stripped tags
- Used helper functions to extract embedded value



[Data Collection – Scraping => Notebook](#)

# Data Wrangling

## Data Wrangling Process for Rocket Landing Prediction– Key Phrase

 **Dataset Source:** `dataset_part_1.csv`

### Data Understanding

- Used `.head()` to preview data
- Check missing value using `.isnull().sum()/len(df)*100`
- Inspect data types using `.dtypes`

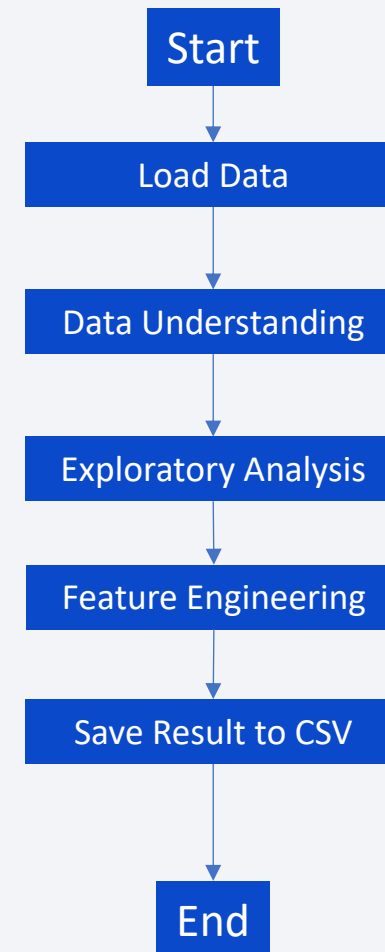
### Exploratory Analysis

Used `.value_counts()` to examine categorical distributions for:

- `LaunchSite`
- `Orbit`
- `Outcome`.

### Feature Engineering

- Make a binary classification using `Outcome` :
  - Labeled as 0 if the outcome was **failure or non-success** (bad outcomes)
  - Labeled as 1 if the outcome was **successful** landing or reuse
- Creating new column `Class` based on binary classification from `Outcome`



[Data Wrangling => Notebook](#)

# EDA with Data Visualization

---

List of plotted chart and their purpose:

**1. Scatter plot between Flight Number and Launch Site**

→ Purpose: To observe the distribution of launches across different sites as flight numbers increase, and to identify success patterns based on launch location.

**2. Scatter plot between Payload Mass and Launch Site**

→ Purpose: To evaluate the relationship between payload mass and launch site, and how it impacts the success of launches.

**3. Bar plot showing the success rate of each orbit type**

→ Purpose: To compare the launch success rate across different orbit types and identify which orbits are more likely to result in successful missions.

**4. Scatter plot between Flight Number and Orbit type**

→ Purpose: To explore how orbit types vary over time (as represented by flight number) and how that relates to mission success.

**5. Scatter plot between Payload Mass and Orbit type**

→ Purpose: To analyze how payload mass is distributed across different orbit types and whether it influences mission outcomes.

**6. Line plot of launch success yearly trend**

→ Purpose: To visualize the trend of launch success rates over the years, indicating whether SpaceX's mission performance has improved over time.

[EDA with Data Visualization => Notebook](#)

# EDA with SQL

---

List of SQL query performed for EDA:

- All Launch Site Names
- Launch Site Name Begin with 'CCA'
- Total Payload Mass
- Average Payload Mass by F9 v.1.1
- First Successful Ground Landing Date
- Successful Drone Ship Landing with Payload between 4000 and 6000
- Total Number of Successful and Failure Mission Outcomes
- Boosters Carried Maximum Payload
- 2015 Launch Records
- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

# Build an Interactive Map with Folium

---

Object list on Folium Map:

- **Markers:** Show all launch site locations with launch counts.
- **Circle Markers (colored):** Indicate success (green) and failure (red) for each launch at SLC-40.
- **Concentric Circles:** Visualize safety zones and distances to key features (city, highway, etc.).
- **Polylines:** Depict proximity to infrastructure (railways, roads, coastlines).

*Purpose:*

To visualize launch site distribution, success rates, and strategic placement based on geography, accessibility, and safety.



# Build a Dashboard with Plotly Dash

---

## User Interactions:

- Dropdown 1: Selects report type – Yearly or Recession Statistics
- Dropdown 2: Selects year (enabled only for Yearly Statistics)

## Visuals by Report Type

### **Recession Period Statistics:**

- Line Chart: Avg. automobile sales over recession years
- Bar Chart: Avg. vehicles sold by type
- Pie Chart: Ad spend share by vehicle type
- Bar Chart: Unemployment rate impact on vehicle sales

### **Yearly Statistics (Selected Year):**

- Line Chart: Sales trend across years
- Line Chart: Monthly sales in the selected year
- Bar Chart: Avg. vehicles sold by type
- Pie Chart: Ad spend by vehicle type

## Purpose:

- These visualizations help analyze sales trends, consumer behavior, and marketing focus across economic conditions and time periods — enabling data-driven insights in a single interactive dashboard.

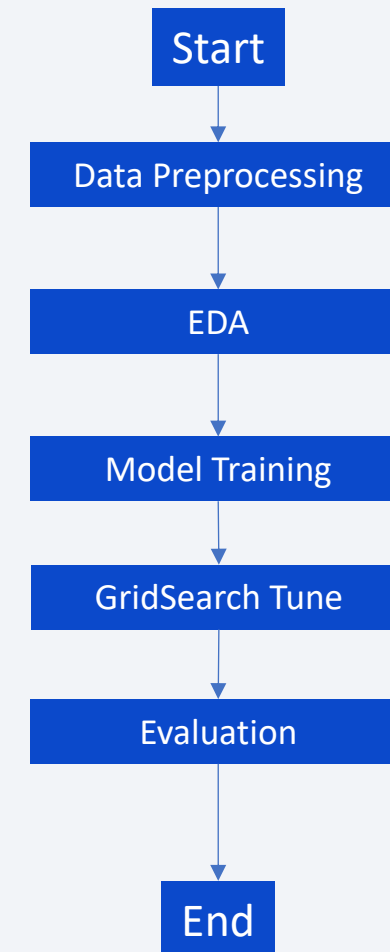
[Dashboard with Plotly Dash => Code](#)

# Predictive Analysis (Classification)

---

## ✓ Model Development Summary (Concise)

- **Preprocessing**
  - Cleaned data, encoded categories, scaled features
- **EDA**
  - Explored feature impact on launch success
- **Model Training**
  - Trained **LogReg, KNN, SVM, Decision Tree**
- **Hyperparameter Tuning**
  - Applied **GridSearchCV** to optimize each model
- **Evaluation**
  - Compared models using **accuracy**
  - **Decision Tree** achieved best accuracy: **0.94**



# Results

---

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



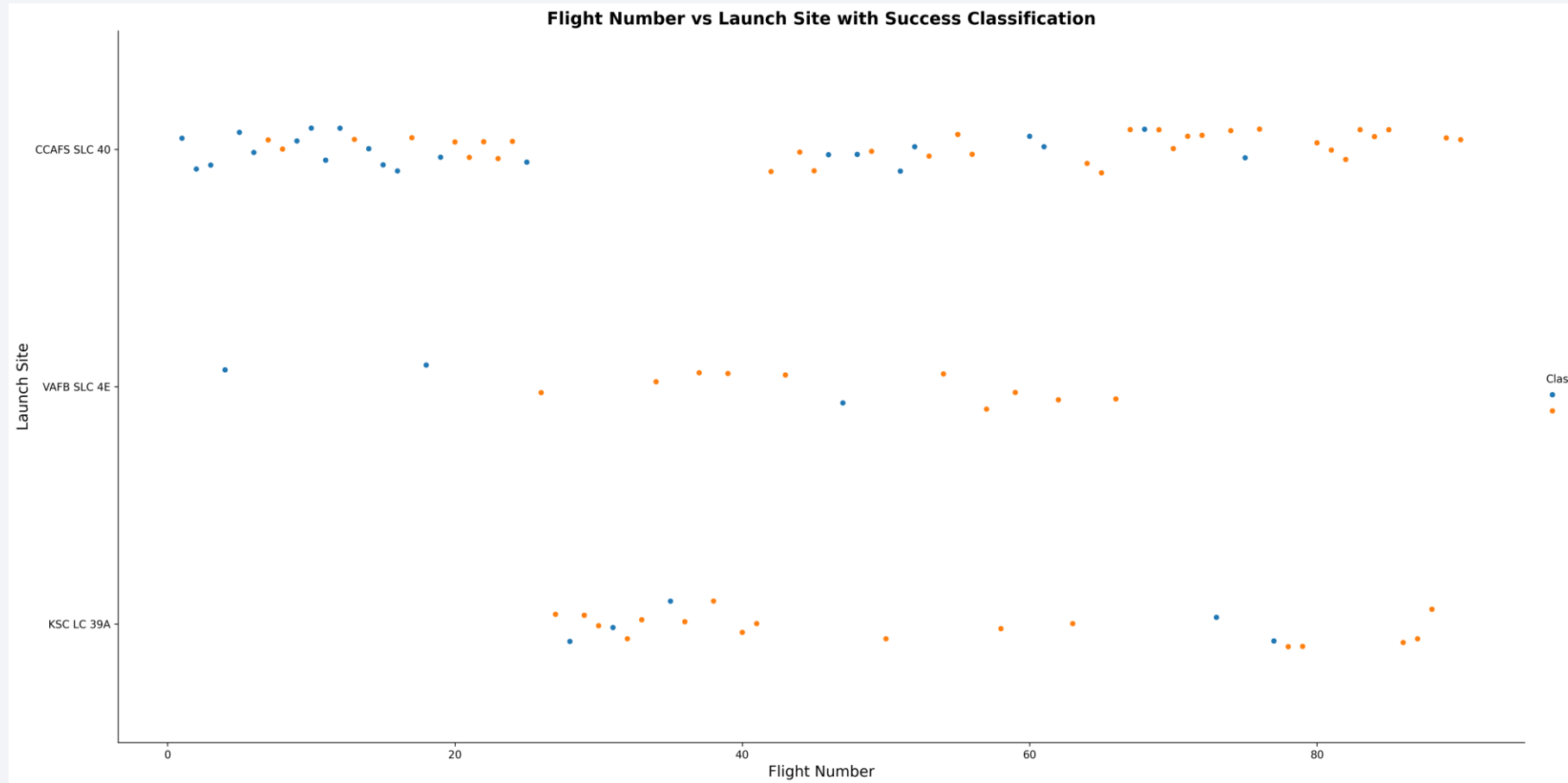
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is dynamic and technological.

Section 2

# Insights drawn from EDA



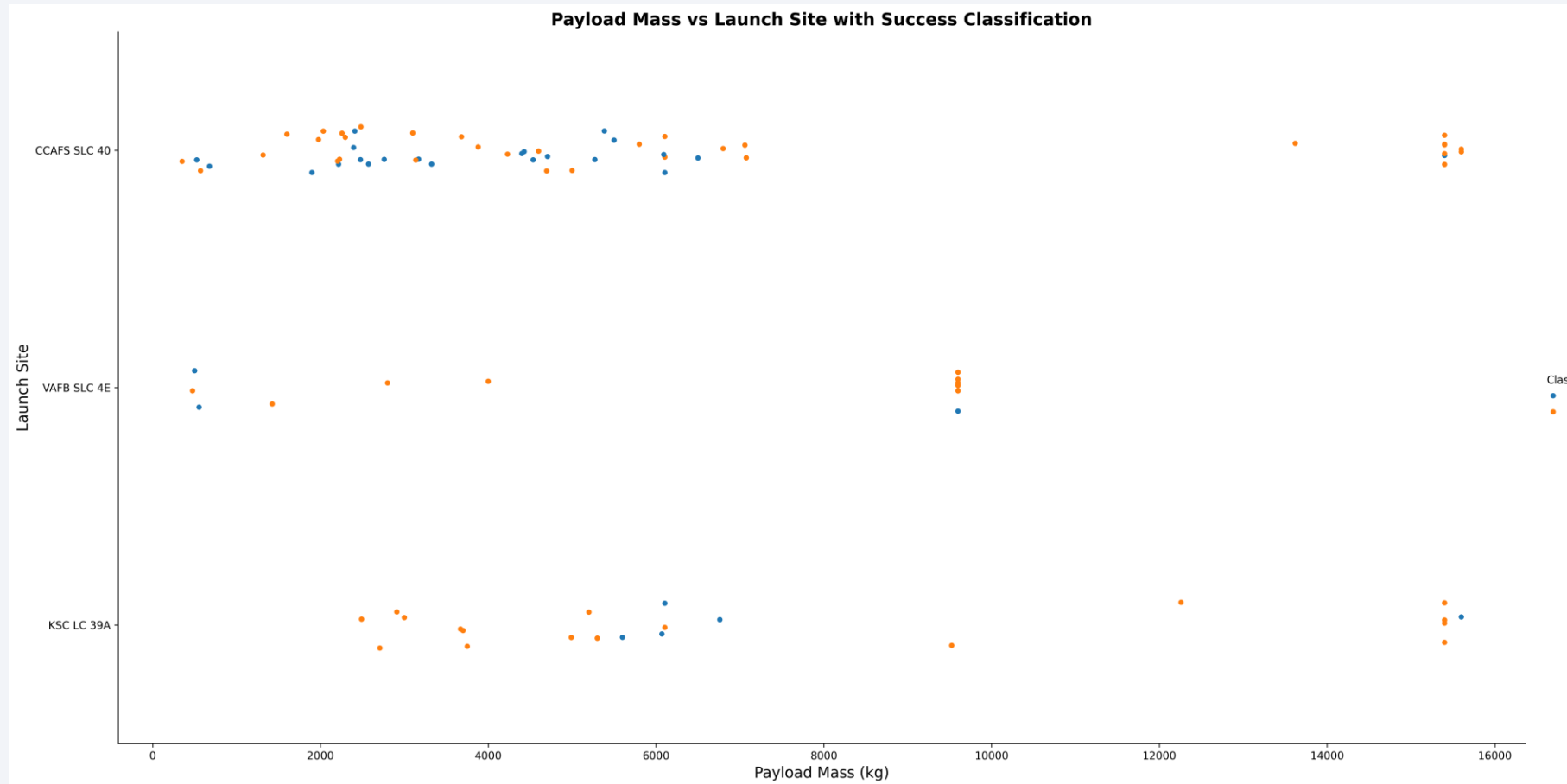
# Flight Number vs. Launch Site



Launches from **CCAFS SLC 40** appear more frequent and show a consistent trend of increasing success (Class = 1) with higher flight numbers compared to other sites.

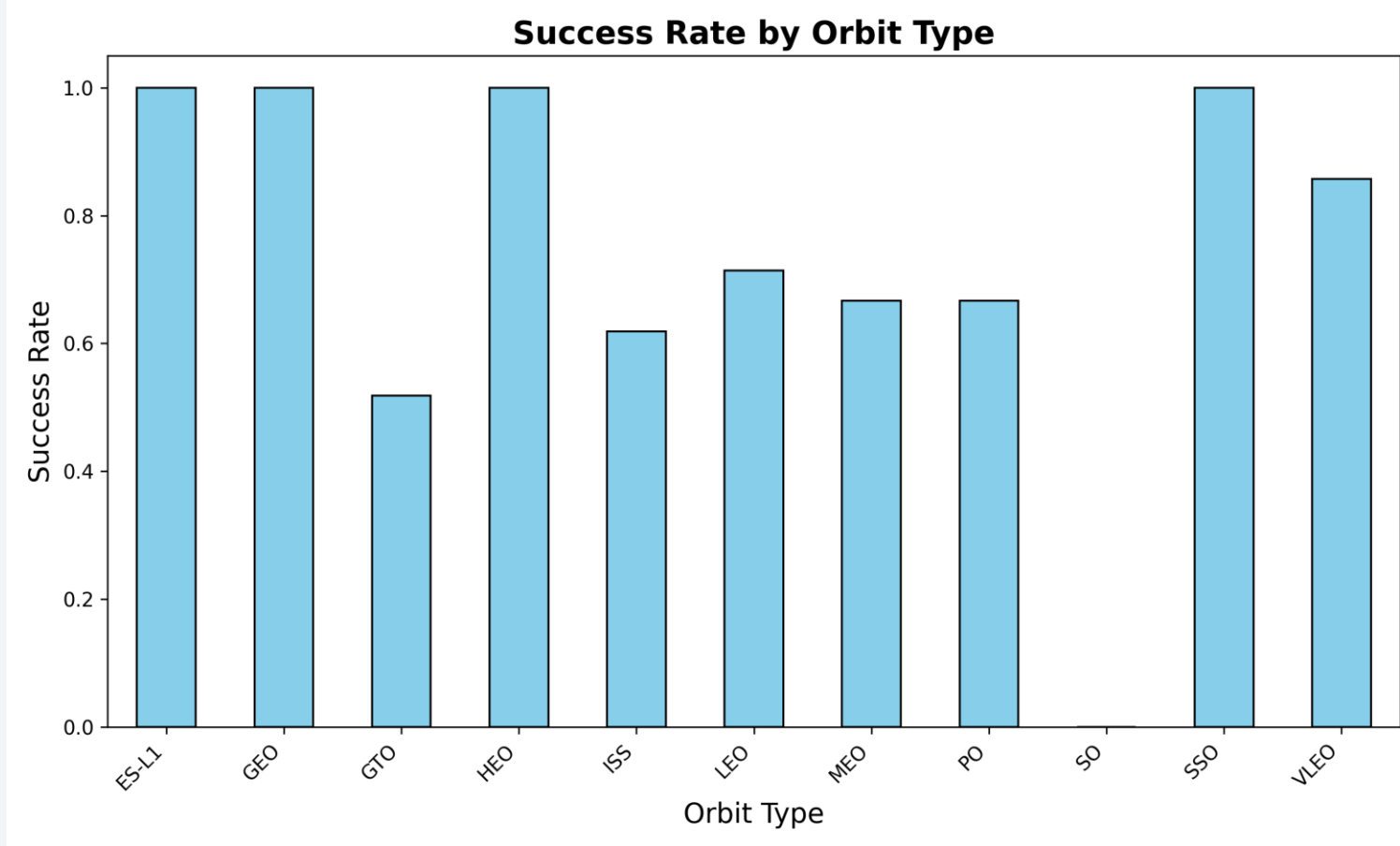


# Payload vs. Launch Site



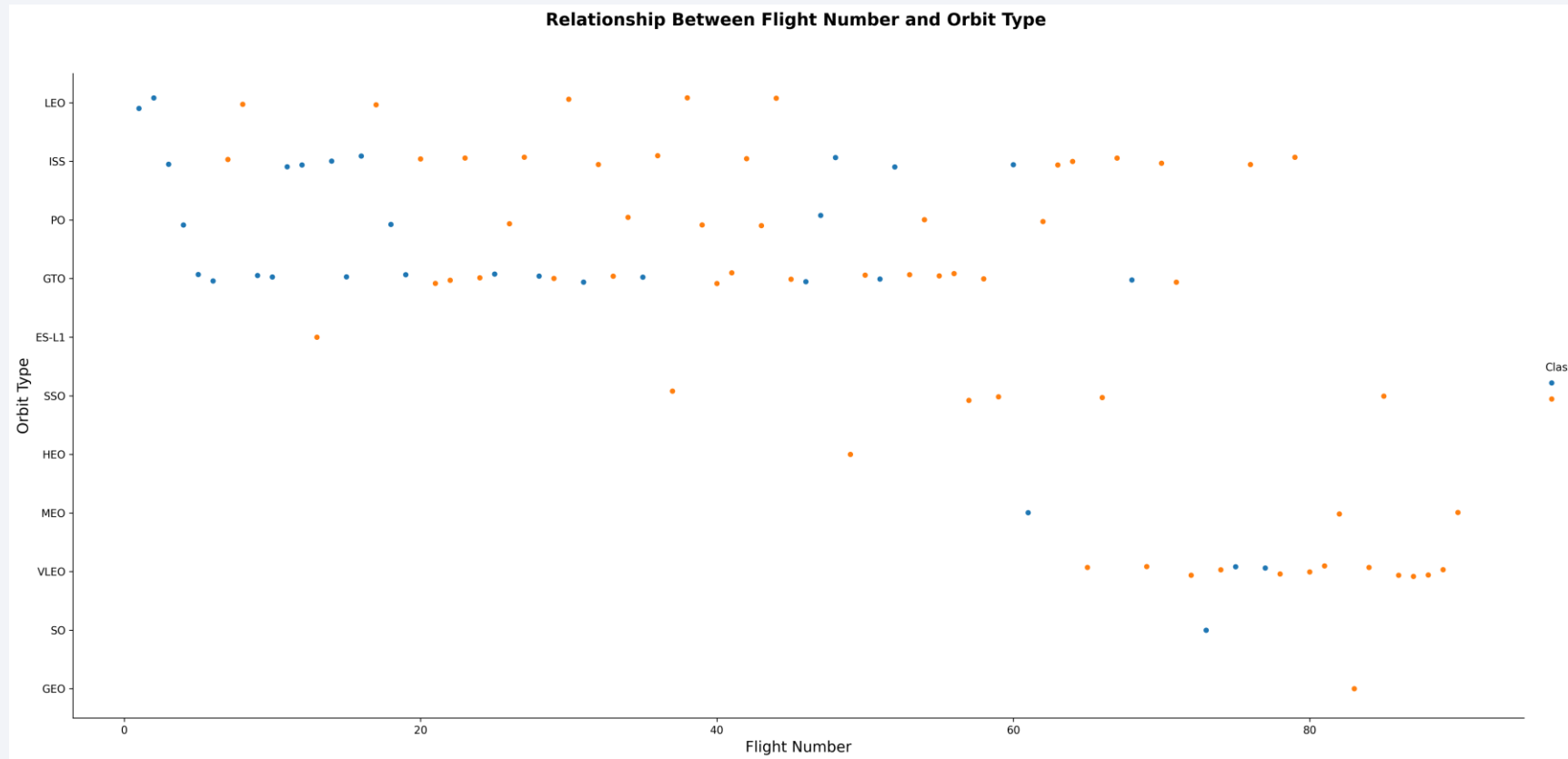
CCAFS SLC 40 handled the widest range of payload masses with a mix of success and failure, while VAFB SLC 4E and KSC LC 39A had fewer launches with mostly successful outcomes.

# Success Rate vs. Orbit Type



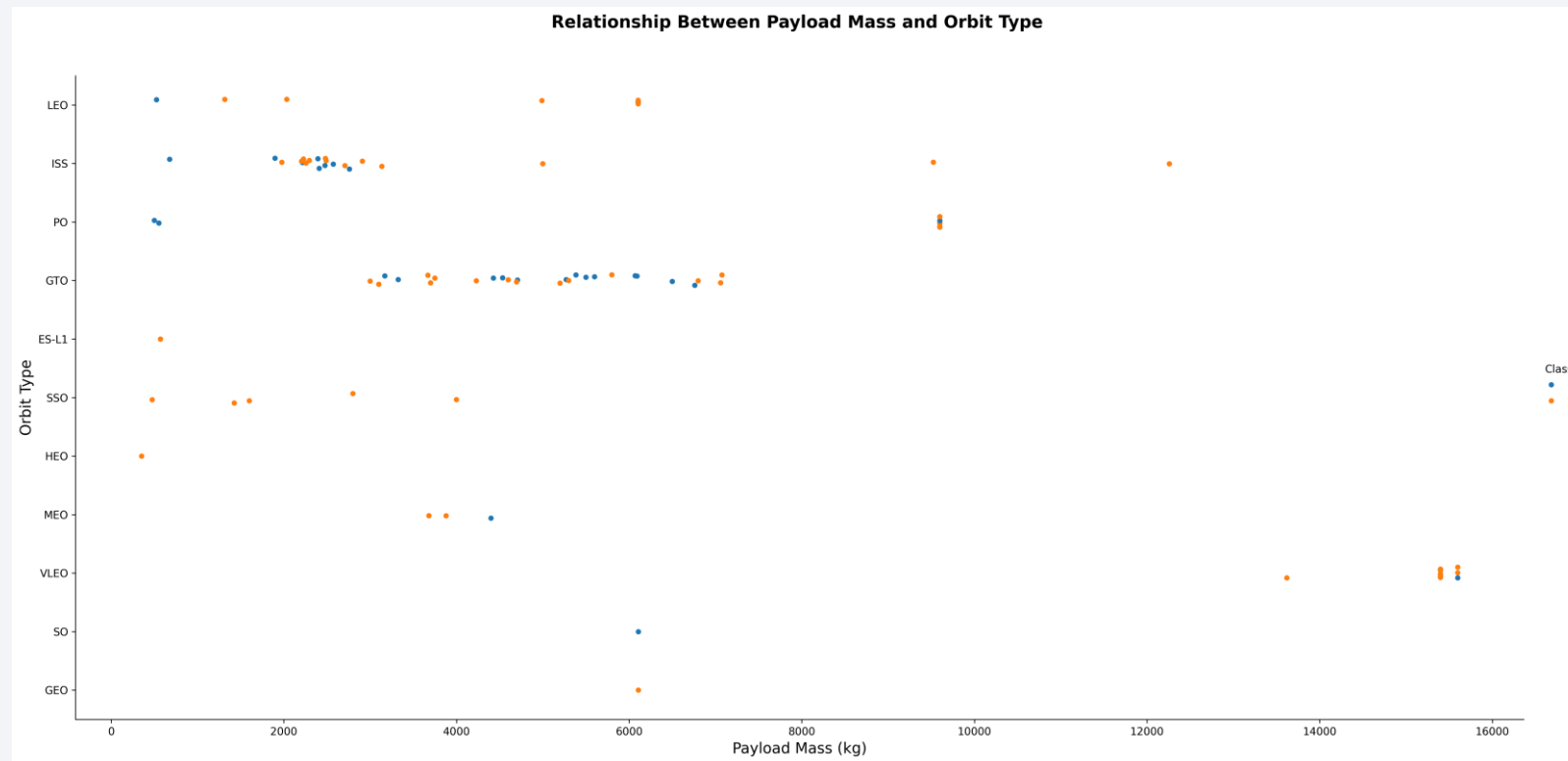
Orbits ES-L1, GEO, HEO, and SSO achieved a perfect 100% success rate, while GTO had the lowest success rate among all orbit types.

# Flight Number vs. Orbit Type



As flight numbers increase, missions diversify into a wider range of orbit types, with Class 1 missions becoming more prominent in later flights.

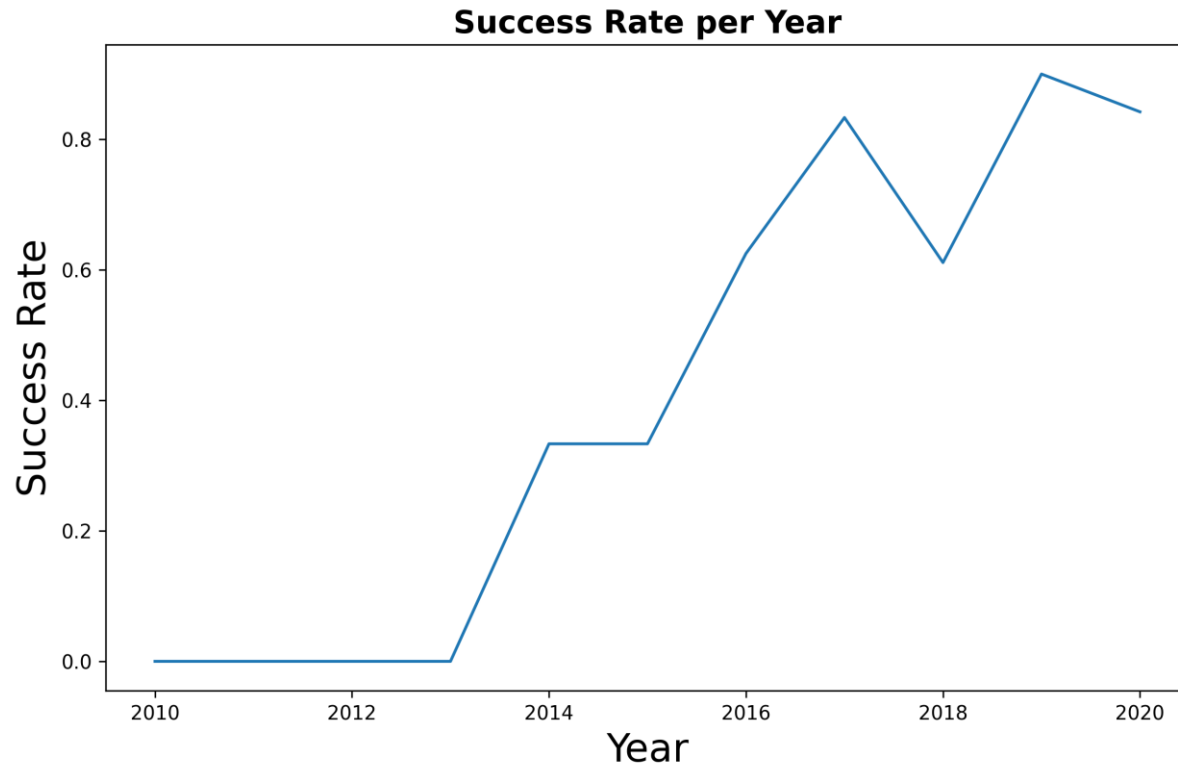
# Payload vs. Orbit Type



LEO & GTO: Wide payload range, mixed success.  
SSO & ES-L1: Light payloads, high success.  
GEO, HEO, SO: Few launches, mostly successful.  
VLEO & ISS: Varied payloads, mixed outcomes.

# Launch Success Yearly Trend

---



Success Rate Trend (2010–2020):  
Gradual improvement with a major rise after 2014,  
peaking in 2019.



# All Launch Site Names

---

## Launch\_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

1. CCAFS LC-40  
Cape Canaveral Air Force Station Launch Complex 40, used primarily by SpaceX for Falcon 9 launches.
2. VAFB SLC-4E  
Vandenberg Air Force Base Space Launch Complex 4E, located in California, used for polar orbit launches.
3. KSC LC-39A  
Kennedy Space Center Launch Complex 39A, a historic pad used for Apollo and Space Shuttle missions, now leased by SpaceX.
4. CCAFS SLC-40  
Duplicate entry or alternate naming for LC-40 at Cape Canaveral, used by SpaceX (same as the first).

# Launch Site Names Begin with 'CCA'

---

## Launch\_Site

CCAFS LC-40

CCAFS LC-40

CCAFS LC-40

CCAFS LC-40

CCAFS LC-40

**This indicates that the query successfully filtered launch sites starting with 'CCA',**  
returning 5 matching entries — all from Cape Canaveral Air Force Station Launch Complex 40.

# Total Payload Mass

---

SUM(PAYLOAD_MASS_KG_)
45596

**Total Payload Mass: 45,596 kg**

This means that SpaceX boosters launched under the NASA (CRS) (Commercial Resupply Services) program carried a combined payload of 45,596 kilograms.

# Average Payload Mass by F9 v1.1

---

AVG(PAYLOAD_MASS_KG_)
2534.6666666666665

**Average Payload Mass (F9 v1.1): 2,534.67 kg** (rounded to two decimal places)

This means that boosters of version Falcon 9 v1.1 carried, on average, approximately 2,535 kilograms of payload per launch.

# First Successful Ground Landing Date

---

MIN(Date)
2015-12-22

## **Date of First Successful Ground Pad Landing: 2015-12-22**

This indicates that on December 22, 2015, SpaceX successfully landed a booster on a ground landing pad for the first time — a major milestone in reusable rocket technology.



# Successful Drone Ship Landing with Payload between 4000 and 6000

---

## Booster\_Version

F9 FT B1021.1

F9 FT B1022

F9 FT B1023.1

F9 FT B1026

F9 FT B1029.1

F9 FT B1021.2

F9 FT B1029.2

F9 FT B1036.1

F9 FT B1038.1

F9 B4 B1041.1

F9 FT B1031.2

F9 B4 B1042.1

F9 B4 B1045.1

F9 B5 B1046.1

These are all Falcon 9 boosters of various versions (FT, B4, B5) that achieved drone ship landings while carrying moderate to heavy payloads in the 4–6 metric ton range. This reflects SpaceX's operational success in reusability and heavy-lift capability.

# Total Number of Successful and Failure Mission Outcomes

---

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- There are 2 separate "Success" entries — likely due to whitespace or formatting differences in the database.
- Merging them gives a total of 99 successful missions.
- There was 1 failed mission and 1 mission where the payload status was unclear, though the launch itself succeeded.

# Boosters Carried Maximum Payload

---

Booster_Version
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

These booster versions — all from the Falcon 9 Block 5 (F9 B5) family — each completed missions that carried the heaviest payloads logged in the SPACEXTBL database. The results confirm the high performance and reuse capability of Block 5 boosters.

# 2015 Launch Records

---

Month_Year	Landing_Outcome	Booster_Version	Launch_Site
January 2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April 2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The query lists all records from 2015 where the booster failed to land on the drone ship. It shows the month and year of the mission, the landing outcome, the booster version, and the launch site. The results reveal two such failures in January and April 2015, both launched from CCAFS LC-40 using the F9 v1.1 booster versions.

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

---

Landing_Outcome	TOTAL_NUMBER
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

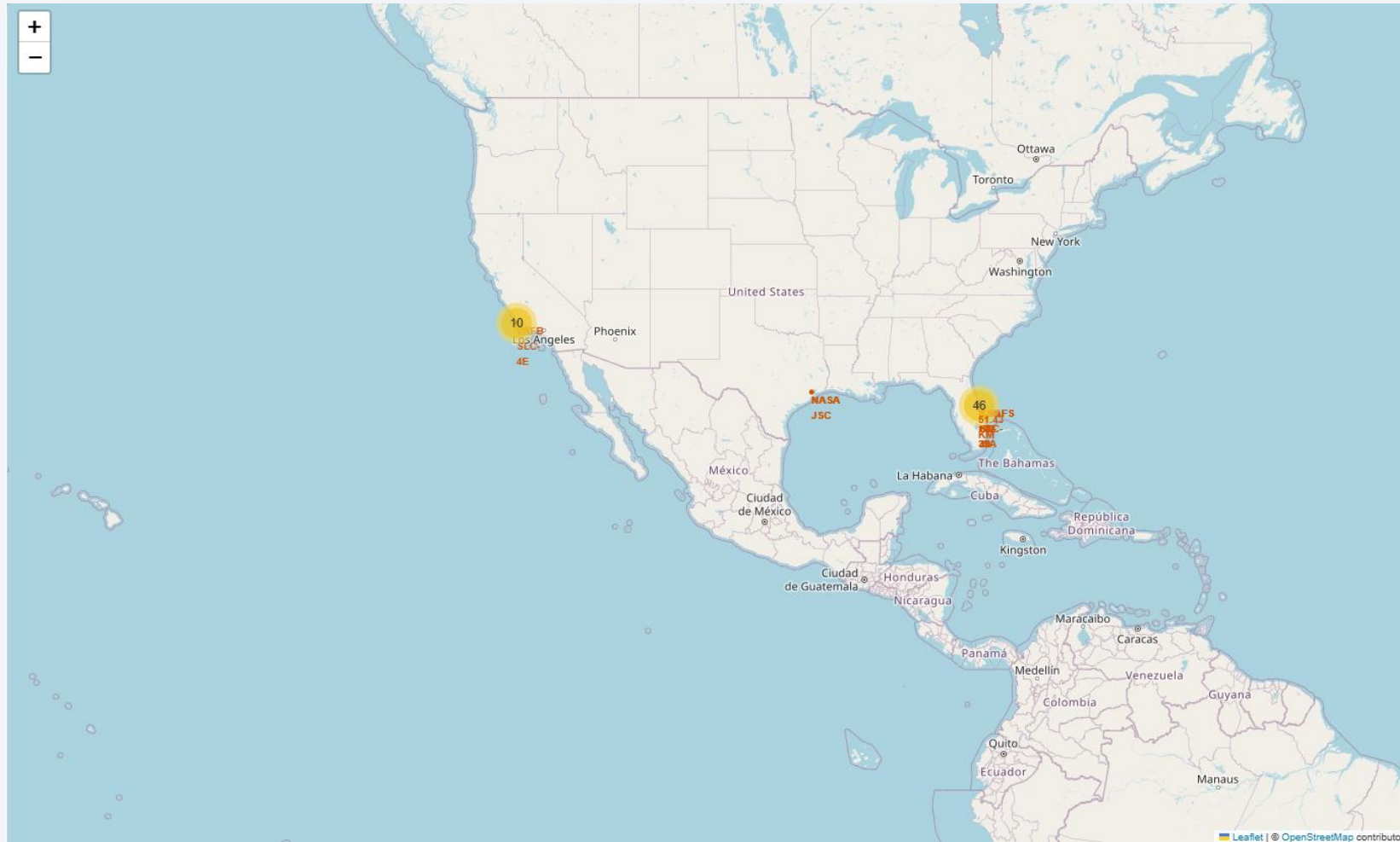
- The most common outcome was "No attempt" with 10 records.
- Successful landings on the drone ship occurred 5 times, matching the number of failures there.
- Ground pad successes and other outcomes had fewer counts.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

# Launch Sites Proximities Analysis

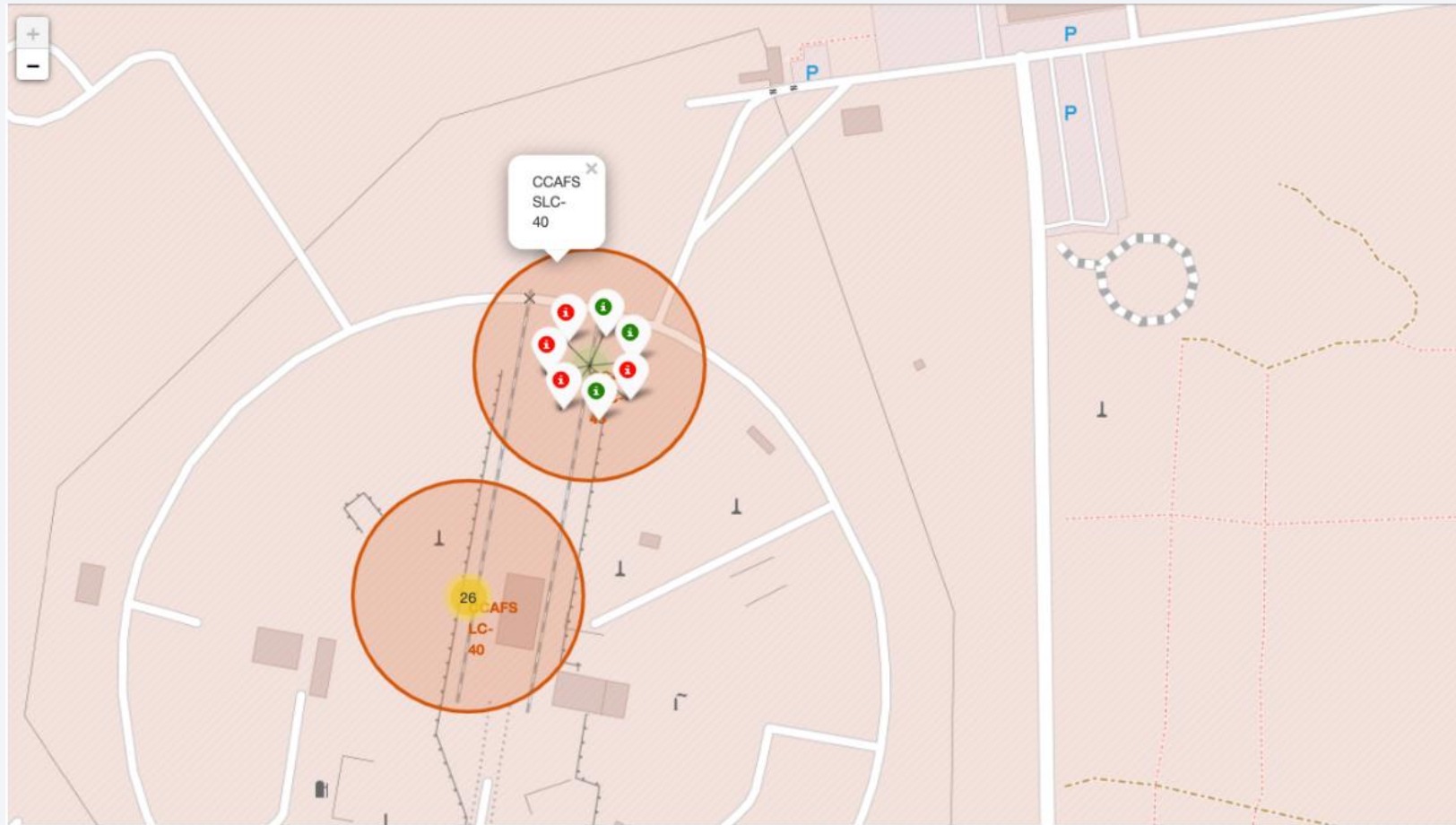
# All Launch Site Locations



All launch sites are located near the equator and along the coast, which is strategic. Launching from the equator requires less fuel thanks to Earth's rotational boost, and coastal locations help minimize risk by allowing rockets to safely drop debris over the ocean.



# Success/failed launches for each site

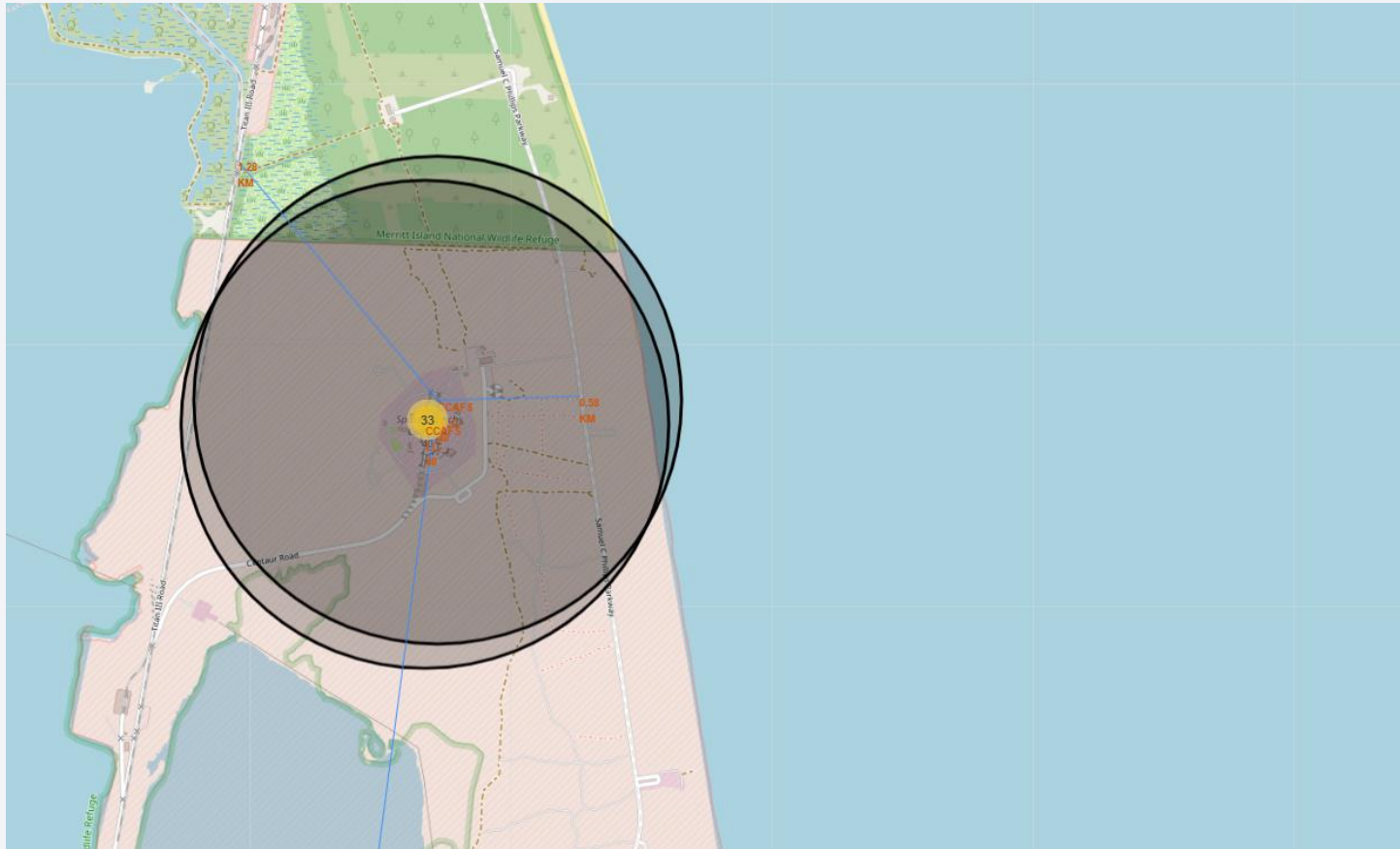


Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rates, namely SLC-40



# Launch Site Location Analysis

---



Visualizing the proximity of railways, highways, coastlines, and cities for each launch site helps us understand their strategic placement. For example, the proximities for **CCAFS SLC-40** are:

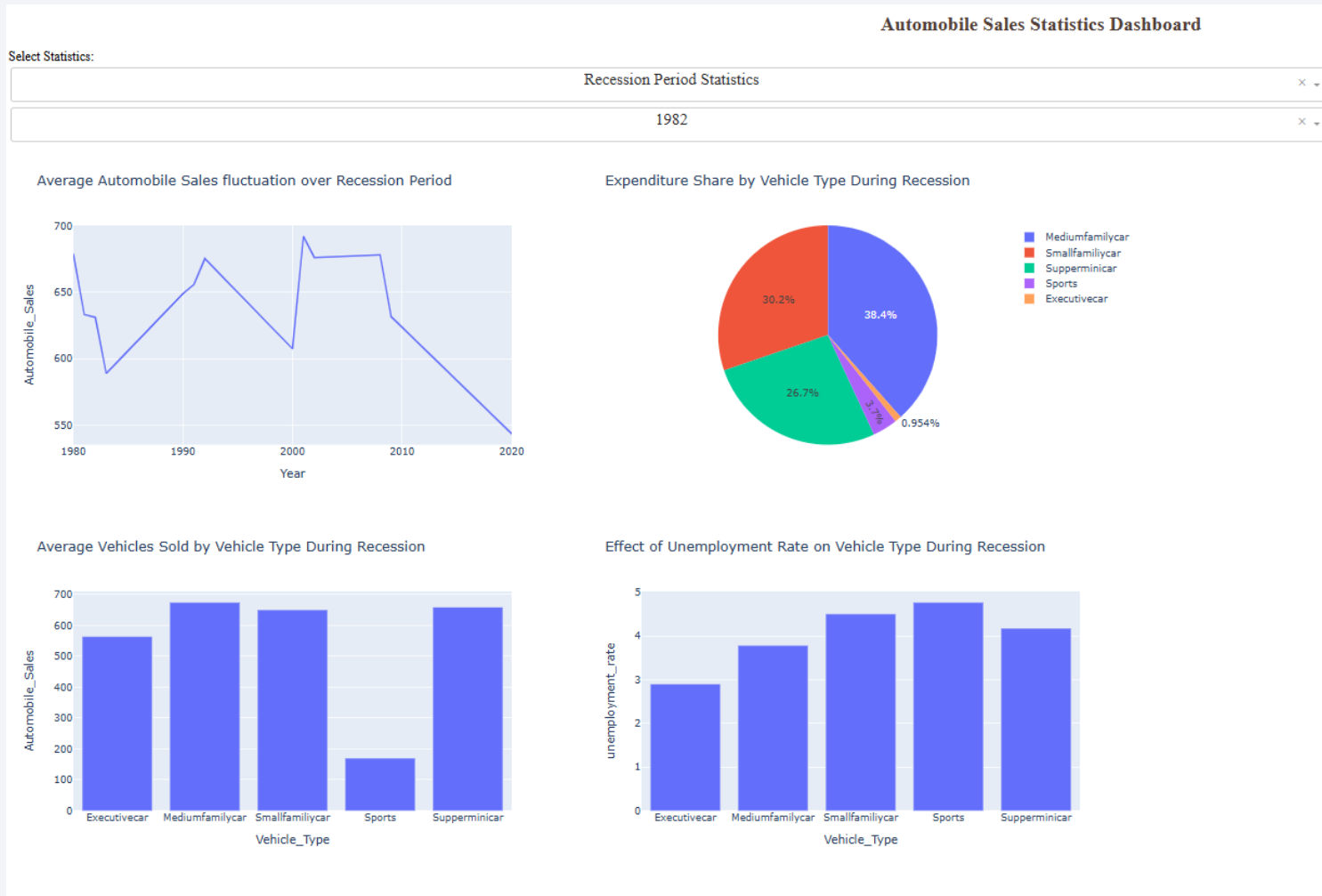
- **Railway: 1.28 km** – enables transport of heavy cargo
- **Highway: 0.58 km** – allows easy movement of personnel and equipment
- **Coastline: 0.86 km** – provides the option to abort launches over water and reduces risk from falling debris
- **City: 51.43 km** – ensures safe distance from densely populated areas to minimize danger during launches



Section 4

# Build a Dashboard with Plotly Dash

# Recession Period Statistics



→ Filter for type of Topic Analysis  
→ Year Filter

Visualization of Automobile Sales Statistics during Recession Period

# Yearly Statistics

## Automobile Sales Statistics Dashboard

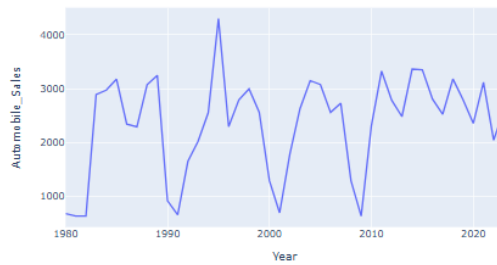
Select Statistics:

Yearly Statistics

1982

- Filter for type of Topic Analysis
- Year Filter

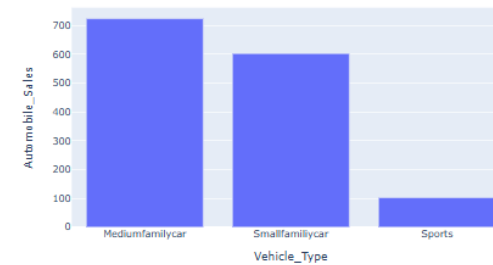
Yearly Automobile Sales Over Time



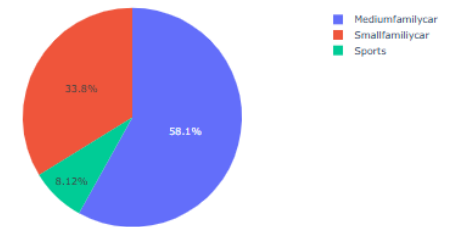
Total Monthly Automobile Sales for the Year 1982



Average Vehicles Sold by Vehicle Type in the year 1982



Advertisement Expenditure by Vehicle Type for the Year 1982



## Visualization of Automobile Sales In a Year

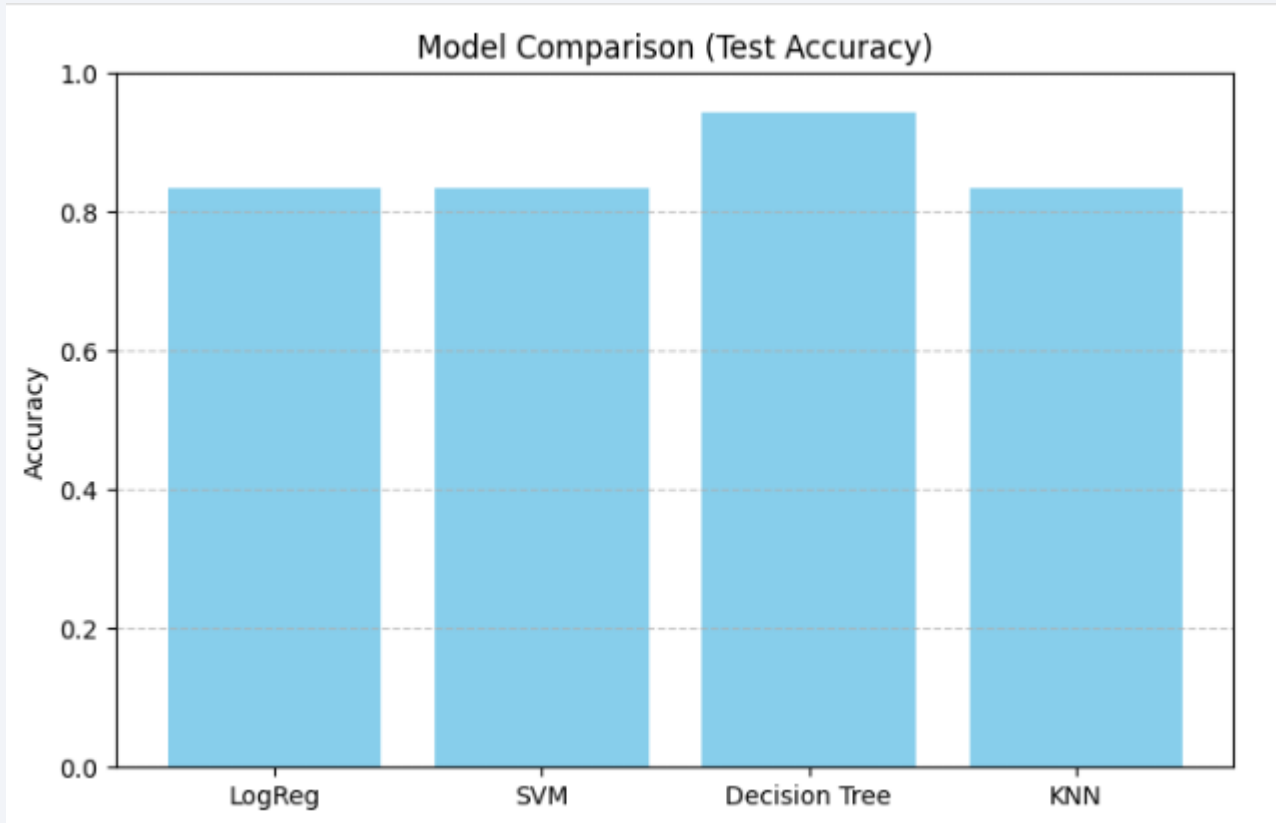


Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

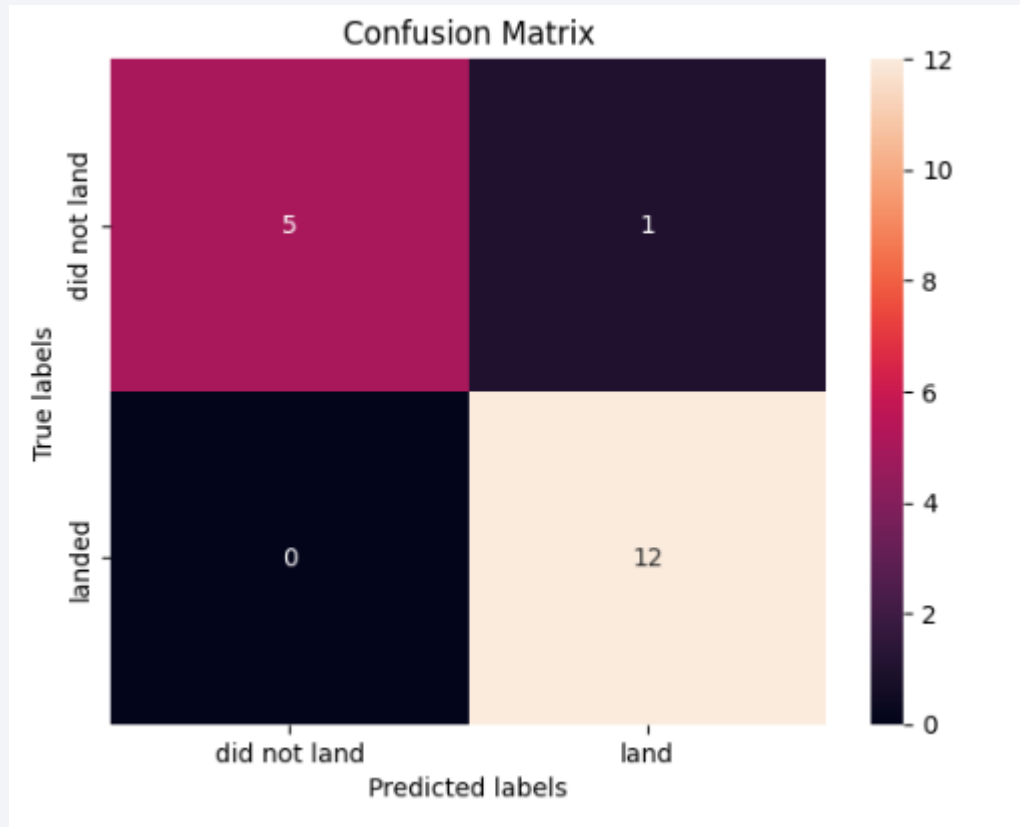
---



This bar chart compares test accuracy across four models. Decision Tree performs best (0.94 accuracy) , followed closely by LogReg, SVM, and KNN, which have similar accuracy levels around 0.83.

# Confusion Matrix

---



This confusion matrix shows strong model performance:

- Correct predictions: 5 (did not land), 12 (landed)
- Only 1 misclassification (predicted "land" but actually "did not land")
- Overall, the model is highly accurate with minimal error.

# Conclusions

---

The combination of visual exploration and machine learning modeling provides strong insights into launch success factors:

- **Visual Insights:**
  - Launches to **LEO** and **ISS** orbits have the **highest success rates**.
  - **Heavier payloads** are typically launched from more capable and reliable launch sites.
  - Certain **orbits** (e.g., GTO, GEO) show **greater risk** associated with heavier payloads.
  - **Success rate by launch site** indicates that some sites have a better reliability record.
  - There is a **clear improvement in launch success rate over time**, reflecting better technology and procedures.
- **Modeling Results:**
  - All models (KNN, SVM, Logistic Regression) achieved an accuracy of **0.83**.
  - The **Decision Tree** model performed best with an accuracy of **0.94**, suggesting it captures complex patterns well.
- **Implications:**
  - Visualizations helped identify key relationships in the data (e.g., payload vs. orbit type).
  - Machine learning models validated these patterns and offer a **predictive advantage**.
  - This combined approach supports **better decision-making** for:
    - Choosing orbits and launch sites,
    - Managing payload size and mission planning,
    - Minimizing mission risk.



# Appendix

---

- Github Repository :  
[https://github.com/jokoeliyanto/coursera\\_capstone\\_project/tree/main](https://github.com/jokoeliyanto/coursera_capstone_project/tree/main)

Thank you!

