

# Winning Space Race with Data Science

Andrew Srb  
9/27/2025



# Outline

---

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

# Executive Summary

---

- Analyze SpaceX launch records to identify success patterns, payload correlations, and booster performance.
- Data collection via SpaceX REST API , Interactive dashboard, and Predictive modeling.
- Real-time insights for mission planning and strategic decision-making.
- Conclusions from predictive analysis find that there's an ~87% chance of accuracy

# Introduction

---

- What Determines a Successful Landing?
- Key Factors We Investigated:
- Payload Size – How the weight of the mission cargo influences landing outcomes
- Booster Version – Differences in reliability across Falcon 9 variants
- Launch Details – Timing, mission type, and mission complexity
- Launch Location – Impact of site conditions (Cape Canaveral, Vandenberg, Kennedy)
- Launch Distance – Orbital destination and mission trajectory
- Landing Location – Ground pad vs. drone ship recovery

Section 1

# Methodology

# Methodology

---

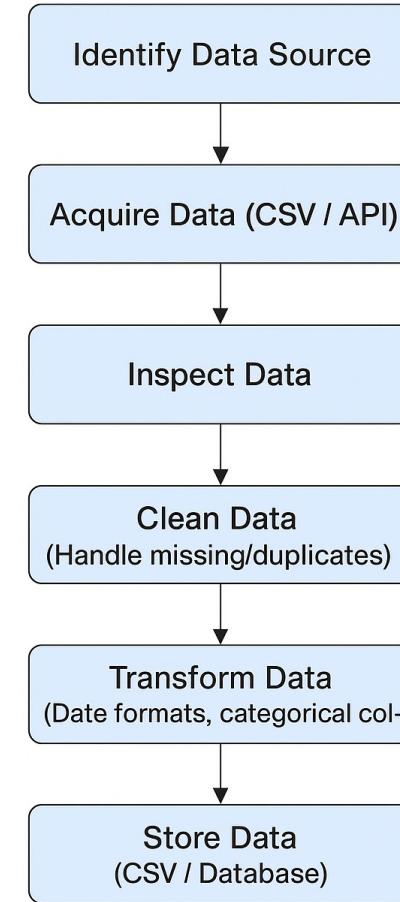
## Executive Summary

- Data collection methodology:
  - API pulls and web scraping from trusted public sources
- Perform data wrangling
  - Launch details, orbit distance, and outcome classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained and tested the data set utilizing LogisticalRegression, KNN, Decision Trees, and SVM

# Data Collection

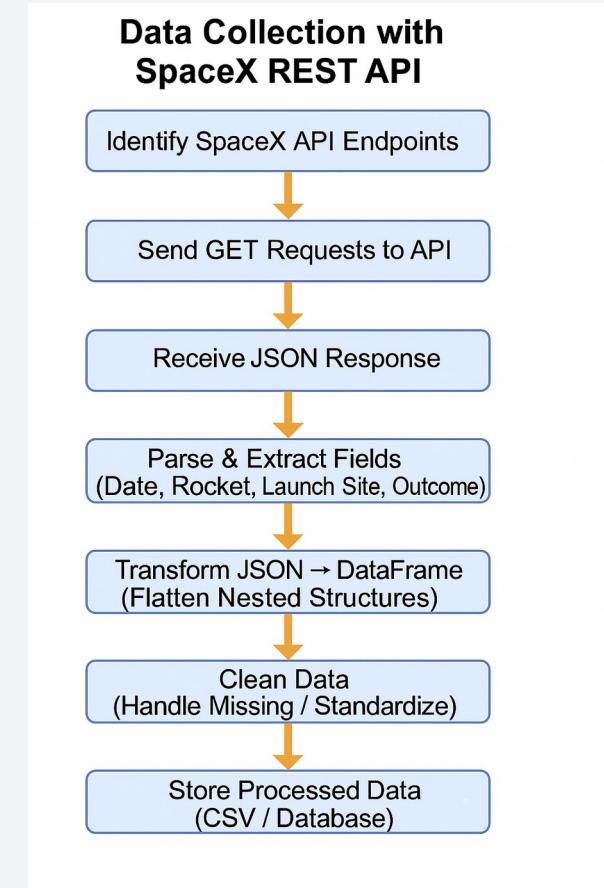
- Source Identification: Identify reliable data sources (e.g., public SpaceX launch datasets, Kaggle, APIs).
- Data Acquisition: Download CSV files or query APIs to gather structured launch data.
- Data Inspection: Check column names, data types, and completeness.
- Data Cleaning: Handle missing values, duplicates, and inconsistent formats.
- Data Transformation: Convert dates to standard formats, categorize outcomes (success/failure).
- Data Storage: Save the cleaned dataset in CSV or database for further analysis.

## Data Collection Process



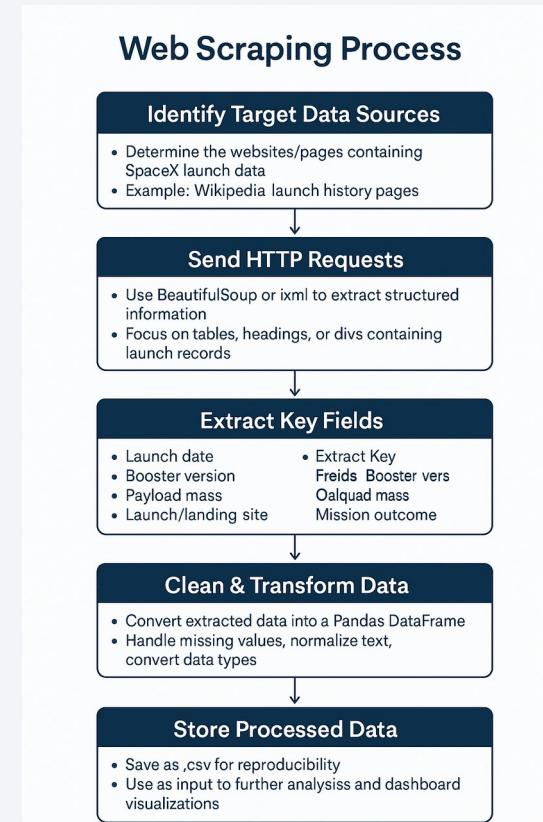
# Data Collection – SpaceX API

- Identify API Endpoints
- Parse & Extract Fields
- Important attributes: date\_utc, rocket, launchpad, payloads, success
- Clean Data
- Handle missing values (e.g., payload mass for failed missions)
- Standardize outcomes (success → 1, failure → 0)
- Normalize categorical values
- [GitHub link](#)



# Data Collection - Scraping

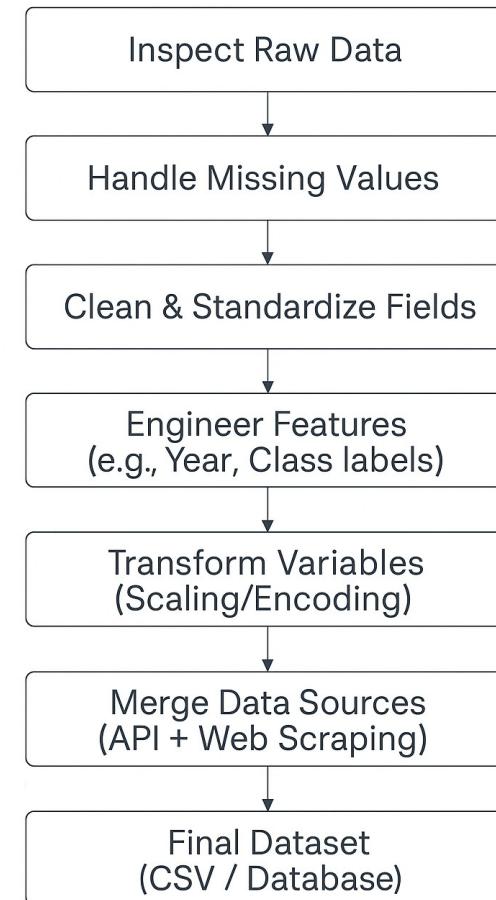
- Identify Target Data Source
- Select websites containing SpaceX launch data (e.g., Wikipedia)
- Extract Key Fields
- Launch date, booster version, payload mass, launch site, outcome
- Clean & Transform Data
- Handle missing values, normalize text, standardize formats
- Store Processed Data
- [GitHub Link](#)



# Data Wrangling

- Data Inspection – Checked dataset structure, column names, and data types
- Handling Missing Values – Filled or removed incomplete rows (e.g., payload mass missing for failed launches)
- Data Cleaning – Standardized categorical variables (success/failure → 1/0, booster versions normalized)
- Feature Engineering – Extracted year from launch date, derived launch outcome labels, and created numerical encodings for categorical fields
- Data Transformation – Scaled payloads and continuous variables with StandardScaler
- Data Integration – Merged API, web-scraped, and cleaned data into one consistent DataFrame
- Final Storage – Saved clean dataset into .csv for analysis and dashboard use
- [GitHub Link](#)

## Data Wrangling Process



# EDA with Data Visualization

---

- Bar Charts - Compare success rates across orbit types
- Line Charts - Track success rate trends over time (by year)
- Scatter Plots - Visualize relationship between payload & outcomes, color-coded by booster version
- [GitHub Link](#)

# EDA with SQL

---

- Filter: Launch sites starting with 'CCA'
- Aggregate:
- Total payload mass by NASA (CRS)
- Average payload mass for F9 v1.1
- Date: First successful ground pad landing
- Conditional: Boosters with successful drone ship landings & payload 4000–6000
- Counts: Successful vs failed mission outcomes
- Subquery: Booster versions with maximum payload mass
- Year/Month: Drone ship landing failures in 2015
- Ranking: Landing outcomes ranked (2010–2017)
- [GitHub Link](#)

# Build an Interactive Map with Folium

---

- Markers - Identify launch sites & coastline points
- Circles - Show launch site boundaries (1 km radius)
- Lines - Connect sites to coastlines/cities, show distances
- DivIcon Labels - Display site names & distances directly on map
- [GitHub Link](#)

# Build a Dashboard with Plotly Dash

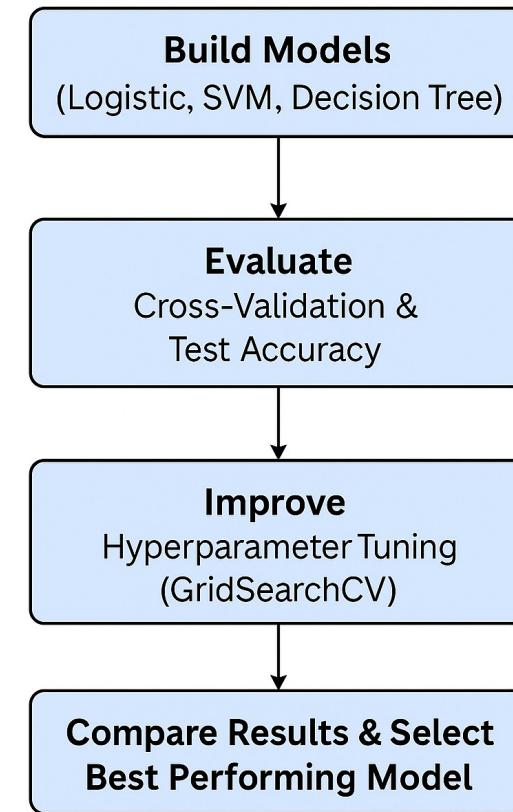
---

- Pie Chart - Shows success vs. failure (per site or all sites)
  - Why: Easy way to compare proportions at a glance
- Scatter Plot - Plots payload vs. outcome, color-coded by booster version
  - Why: Reveals correlations between payload size and success, and booster performance differences
- Interactions:
  - Dropdown (Launch Site): Filter visualizations by site or view all sites
  - Range Slider (Payload): Filter results by payload size range
- [Github Link](#)

# Predictive Analysis (Classification)

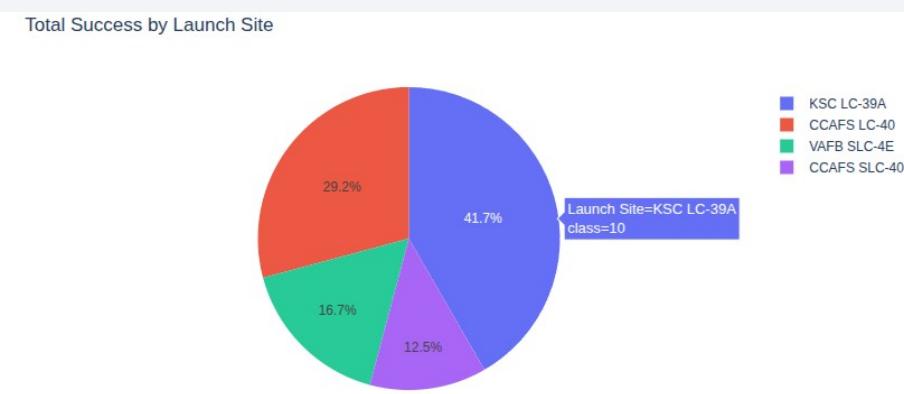
- Build: Trained multiple classifiers (Logistic Regression, SVM, Decision Tree, KNN)
- Evaluate: Used cross-validation ( $cv=10$ ) and accuracy score on test set
- Improve: Performed hyperparameter tuning with GridSearchCV
- Select Best: Compared test accuracies, identified Decision Tree as best performer (~87%)
- [GitHub Link](#)

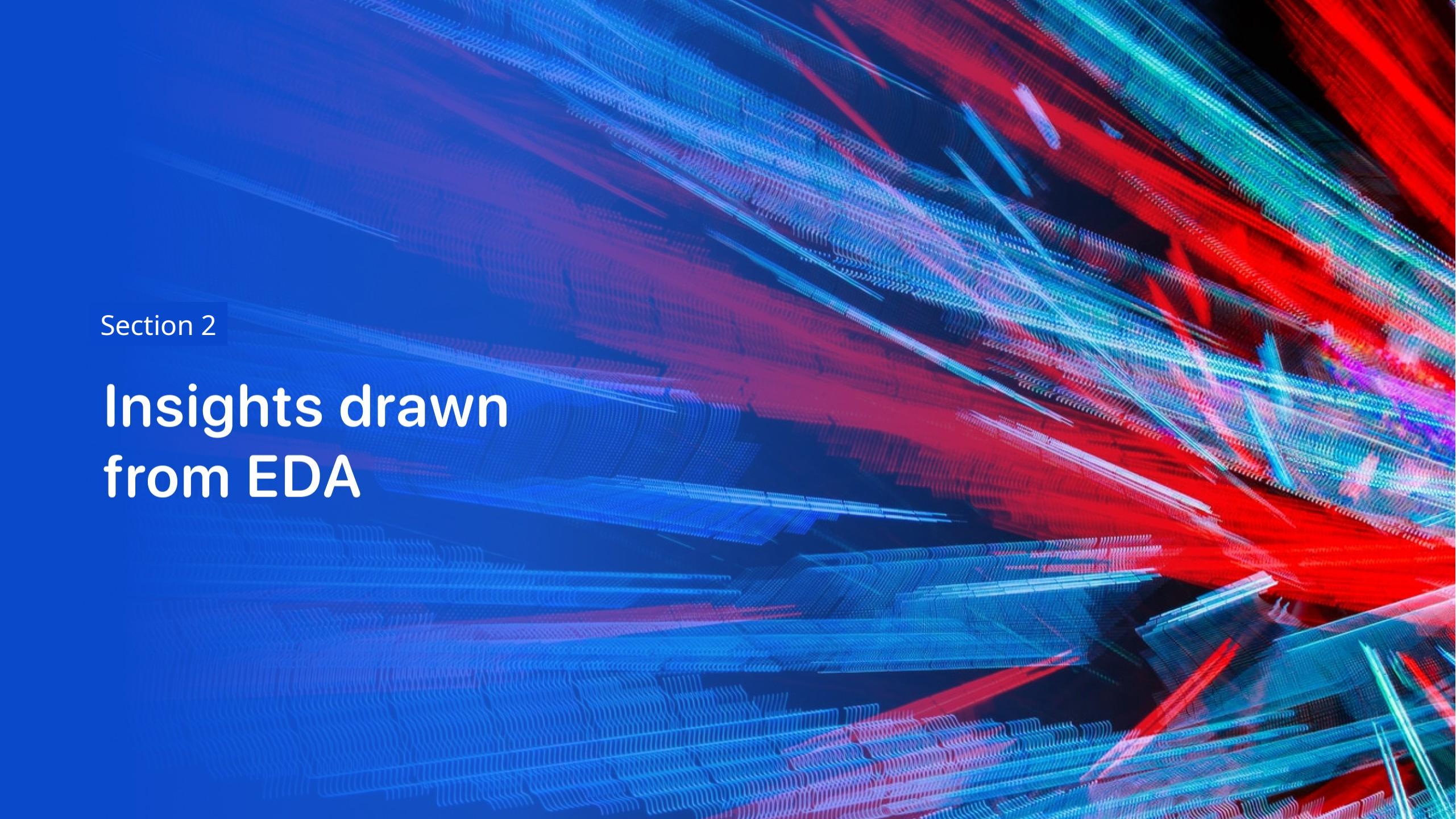
## Model Development Process



# Results

- KSC LC-39A consistently had higher success rates
- Scatter plots revealed heavier payloads (2k-4k) sometimes linked to higher failure probability
- Booster version FT also influenced outcomes
- Bar charts highlighted variation in success depending on orbit type (LEO vs GTO, etc.)
- Line charts showed increasing success rates year over year, improving reliability
- Folium maps confirmed all launch sites are near coastlines
- Distances to landing locations were important for mission recovery planning
- Models Tested: Logistic Regression, SVM, Decision Tree, KNN
- Evaluation Method: 10-fold cross-validation + accuracy on test set
- Best Performing Model: Decision Tree Classifier (~87% accuracy) - Outperformed others by capturing strong categorical relationships (launch site, booster version)
- Other Models:SVM & KNN  $\approx$  85%, Logistic Regression  $\approx$  84%
- Key Insight:
- Launch outcome is highly influenced by site, booster type, and payload range
- Predictive modeling shows reliable accuracy for future mission outcomes

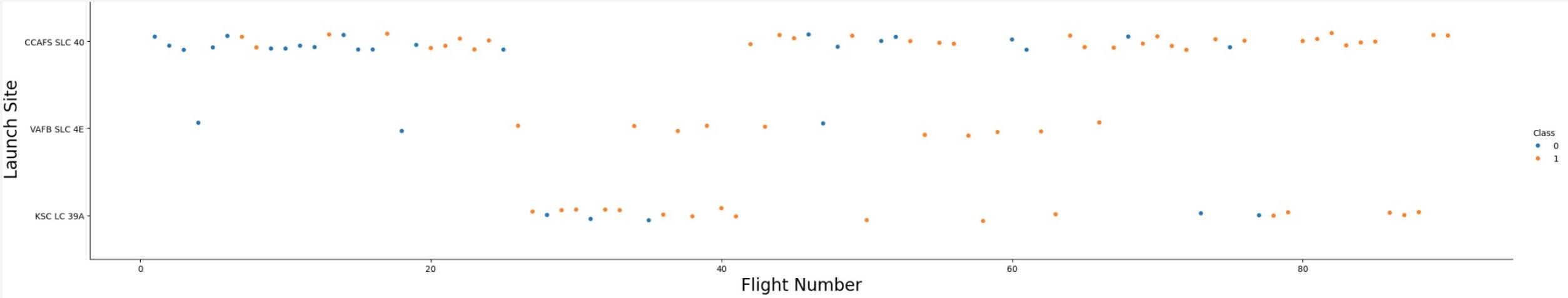


The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a three-dimensional space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

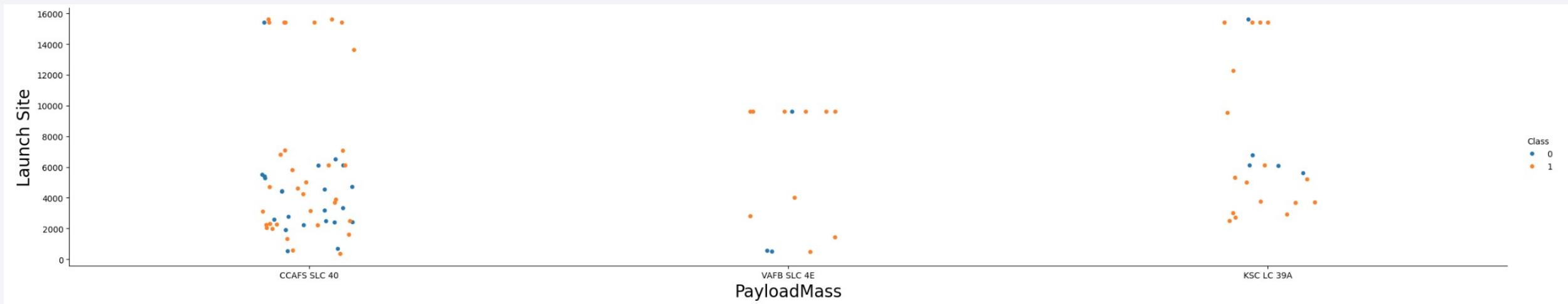
## Insights drawn from EDA

# Flight Number vs. Launch Site



- Early launches concentrated in a few sites (e.g., CCAFS LC-40).
- Over time, more sites were added (KSC LC-39A, VAFB SLC-4E), shown by points appearing at new Y-axis categories.
- The scatter pattern shows SpaceX increasing launch cadence while diversifying across multiple sites.
- Success rates improved as flight numbers increased, across different sites.

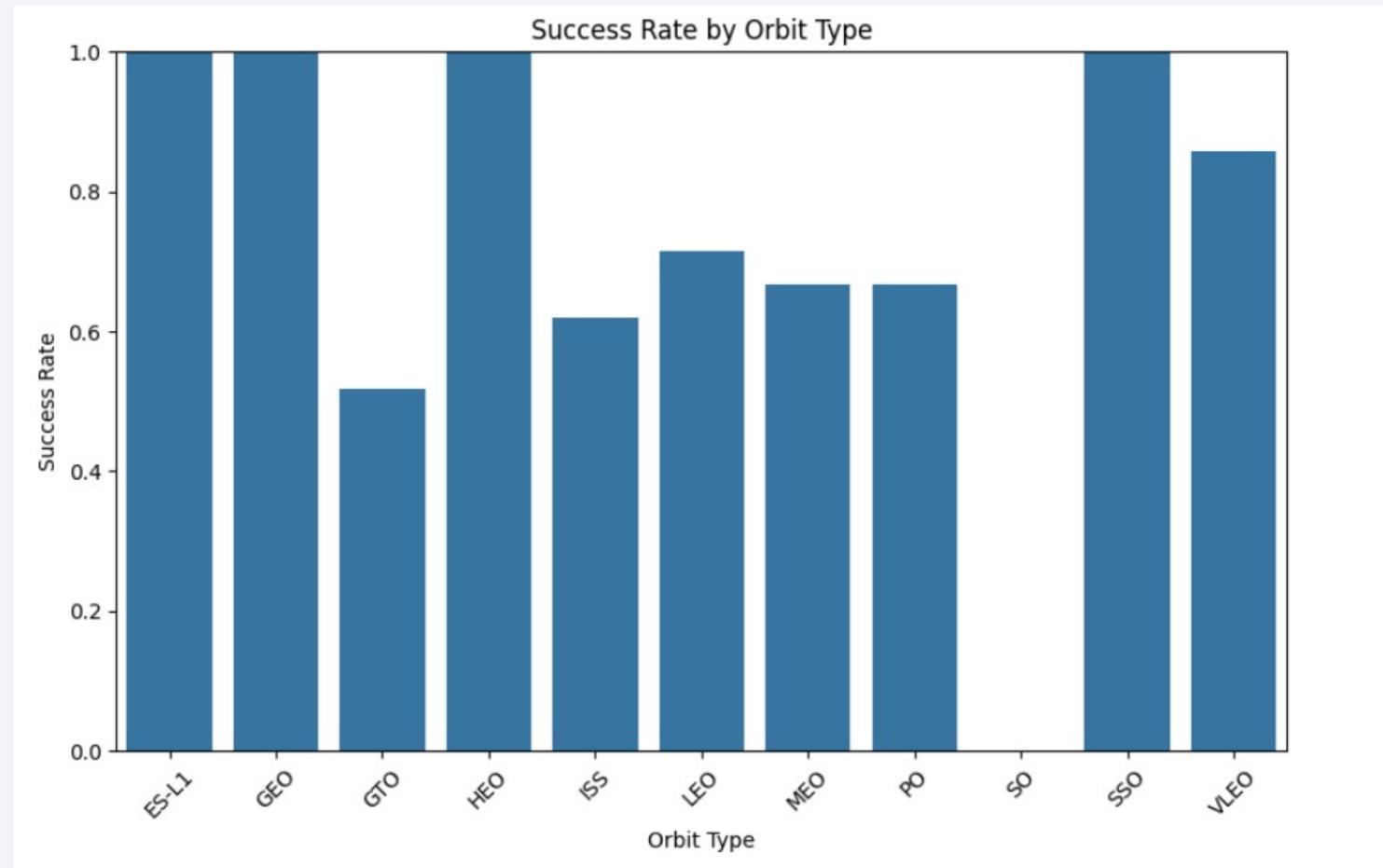
# Payload vs. Launch Site



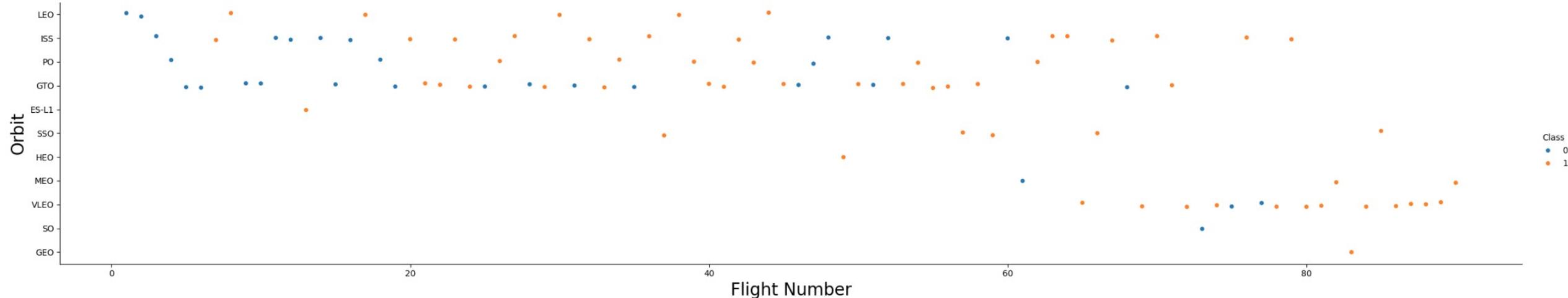
- Variation by site: Some launch sites consistently supported larger payloads (e.g., Kennedy LC-39A), while others handled smaller or medium payloads.
- Success patterns: Heavier payloads showed slightly lower success probability at certain sites (especially early missions).
- Operational strategy: The plot shows how SpaceX assigns missions strategically, with certain launch pads dedicated to high-capacity missions.

# Success Rate vs. Orbit Type

- Higher Success in Common Orbits:
  - Missions to LEO (Low Earth Orbit) and ISS resupply (NASA CRS) had the highest success rates, reflecting routine, well-optimized missions.
- Lower Success in Complex Orbits:
  - Launches to GTO (Geostationary Transfer Orbit) and Polar orbits had lower success rates, due to heavier payloads and more complex trajectories.
- Operational Learning Curve:
  - Early missions to advanced orbits had more failures, but the chart shows improvement over time, as experience and technology matured.

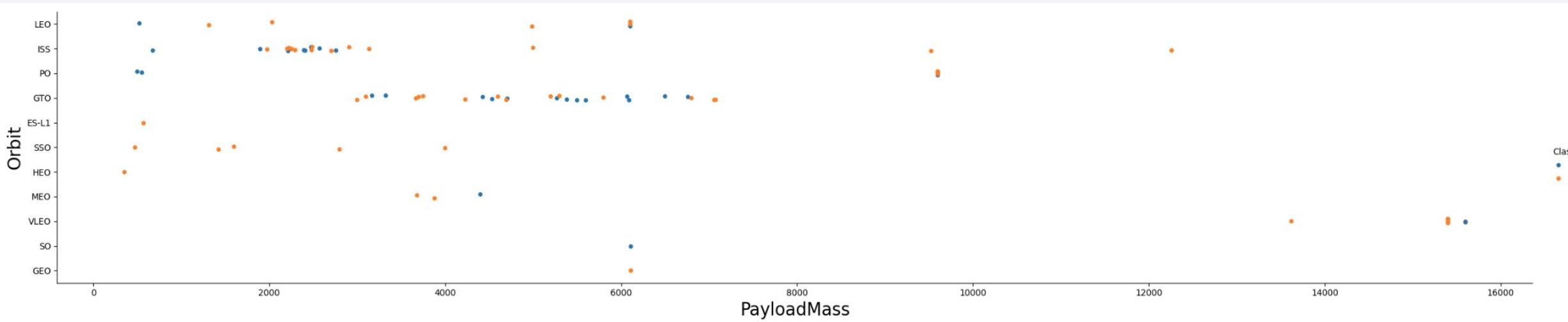


# Flight Number vs. Orbit Type



- Expansion of Mission Types:
  - Early flights (low flight numbers) were mostly LEO missions, indicating simpler, short-distance launches during SpaceX's learning phase.
  - As flight numbers increased, more orbit types appeared (e.g., GTO, Polar), showing diversification of capabilities.
- Improving Reliability Over Time:
  - Early missions to new orbits had mixed success rates.
  - Later missions in the same orbit types showed higher success consistency, reflecting improved technology and operational experience.
- Orbit-Specific Patterns:
  - ISS/LEO missions show consistently high success once operational maturity was reached.
  - GTO and Polar orbits initially displayed lower reliability but improved over time.

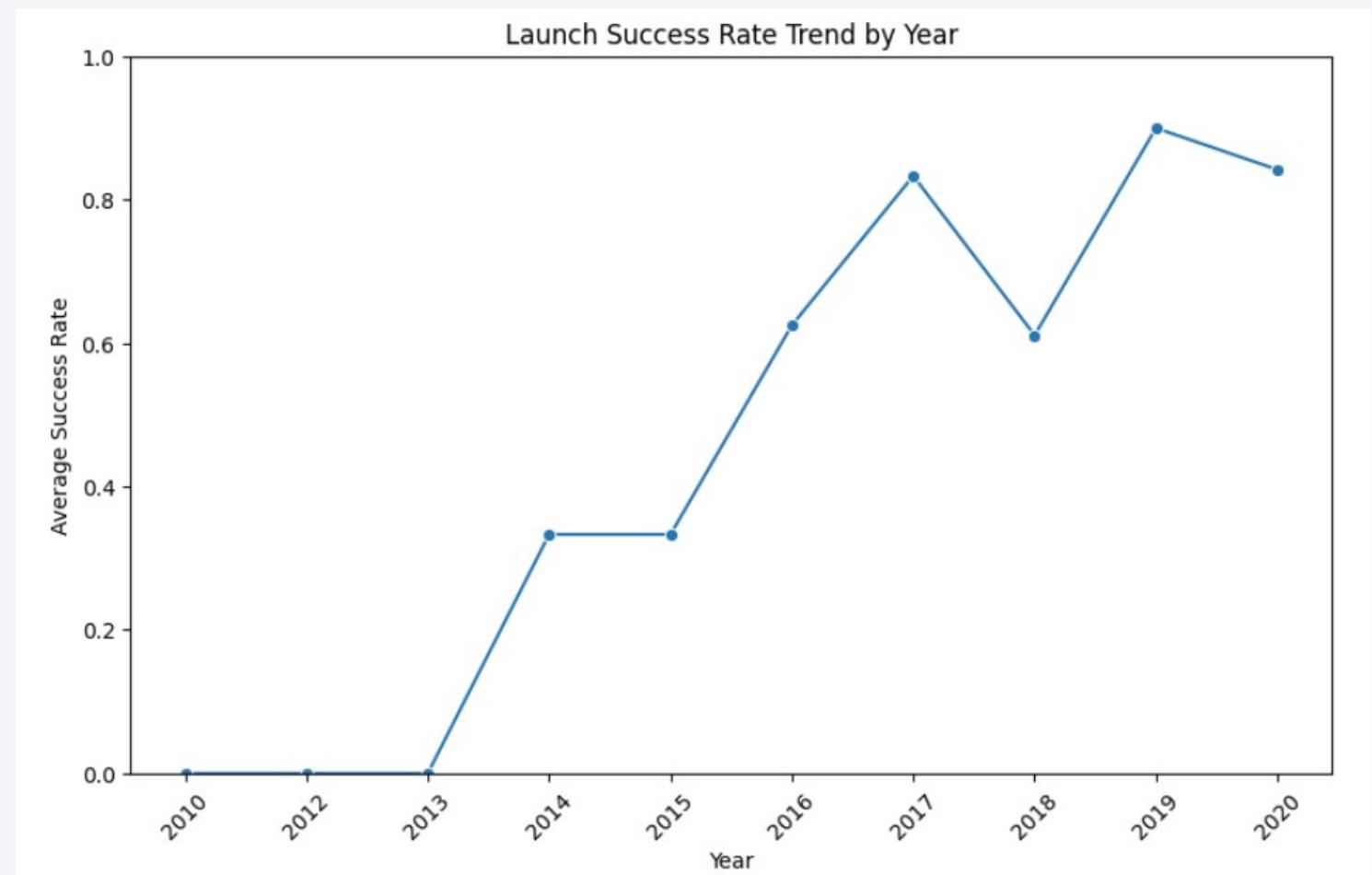
# Payload vs. Orbit Type



- LEO & ISS Missions:
  - Typically carried lighter to medium payloads.
  - Very high success rate once operational maturity was achieved.
- GTO Missions:
  - Concentrated in the heavier payload range.
  - Showed higher risk/failure rate in earlier flights but improved over time.
- Polar & Other Orbits:
  - Payloads were in a broad middle range.
  - Fewer launches overall, making them less common but strategically important.
- General Pattern:
  - As payload mass increased, missions were more likely associated with complex orbits (GTO, Polar).
  - Success rates improved with experience and booster evolution, even for heavy payload missions.

# Launch Success Yearly Trend

- Early Years (2010–2013):
  - Success rate was low and inconsistent
  - Reflected SpaceX's testing and early operational learning curve
- Mid-Years (2014–2016):
  - Noticeable improvement in reliability
  - More frequent successes, fewer mission failures
- Recent Years (2017 onward):
  - Success rate reached consistently high levels (approaching 1.0)
  - Demonstrated maturity of Falcon 9 reusability and operations
- Overall Trend:
  - Clear upward trajectory in success rate over time
  - Shows continuous innovation and learning leading to operational excellence



# All Launch Site Names

---

Display the names of the unique launch sites in the space mission

```
[1]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

```
* sqlite:///my_data1.db
```

```
Done.
```

```
[1]: Launch_Site
```

```
CCAFS LC-40
```

```
VAFB SLC-4E
```

```
KSC LC-39A
```

```
CCAFS SLC-40
```

# Launch Site Names Begin with 'CCA'

---

Display 5 records where launch sites begin with the string 'CCA'

```
: %%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

---

## Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
: %%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
* sqlite:///my_data1.db
Done.
: total_payload_mass
-----  
45596
```

# Average Payload Mass by F9 v1.1

---

## ▼ Task 4

Display average payload mass carried by booster version F9 v1.1

```
31]: %%sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass  
      FROM SPACEXTABLE  
      WHERE Booster_Version = 'F9 v1.1';  
          * sqlite:///my_data1.db  
          Done.
```

```
31]: avg_payload_mass
```

2928.4

# First Successful Ground Landing Date

---

List the date when the first succesful landing outcome in ground pad was acheived.

*Hint:Use min function*

```
: %%sql SELECT MIN(Date) AS first_successful_landing  
FROM SPACEXTABLE  
WHERE Landing_Outcome = 'Success (ground pad)';
```

```
* sqlite:///my_data1.db  
Done.
```

```
: first_successful_landing
```

```
2015-12-22
```

# Successful Drone Ship Landing with Payload between 4000 and 6000

---

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql SELECT Booster_Version  
FROM SPACEXTABLE  
WHERE Landing_Outcome = 'Success (drone ship)'  
AND PAYLOAD_MASS_KG_ > 4000  
AND PAYLOAD_MASS_KG_ < 6000;
```

```
* sqlite:///my_data1.db  
Done.
```

## Booster\_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

# Total Number of Successful and Failure Mission Outcomes

---

List the total number of successful and failure mission outcomes

```
: %%sql SELECT Mission_Outcome, COUNT(*) AS total  
FROM SPACEXTABLE  
GROUP BY Mission_Outcome;
```

```
* sqlite:///my_data1.db  
Done.
```

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

# Boosters Carried Maximum Payload

List all the booster\_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
| : %%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS_KG_ = (
    SELECT MAX(PAYLOAD_MASS_KG_)
    FROM SPACEXTABLE
);
* sqlite:///my_data1.db
Done.
```

```
| : 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
```

# 2015 Launch Records

---

List the records which will display the month names, failure landing\_outcomes in drone ship ,booster versions, launch\_site for the months in year 2015.

Note: SQLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql
SELECT
    substr(Date, 6, 2) AS Month,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTABLE
WHERE substr(Date, 1, 4) = '2015'
    AND Landing_Outcome = 'Failure (drone ship)';
```

```
* sqlite:///my_data1.db
```

```
Done.
```

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

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

\* sqlite:///my\_data1.db

Done.

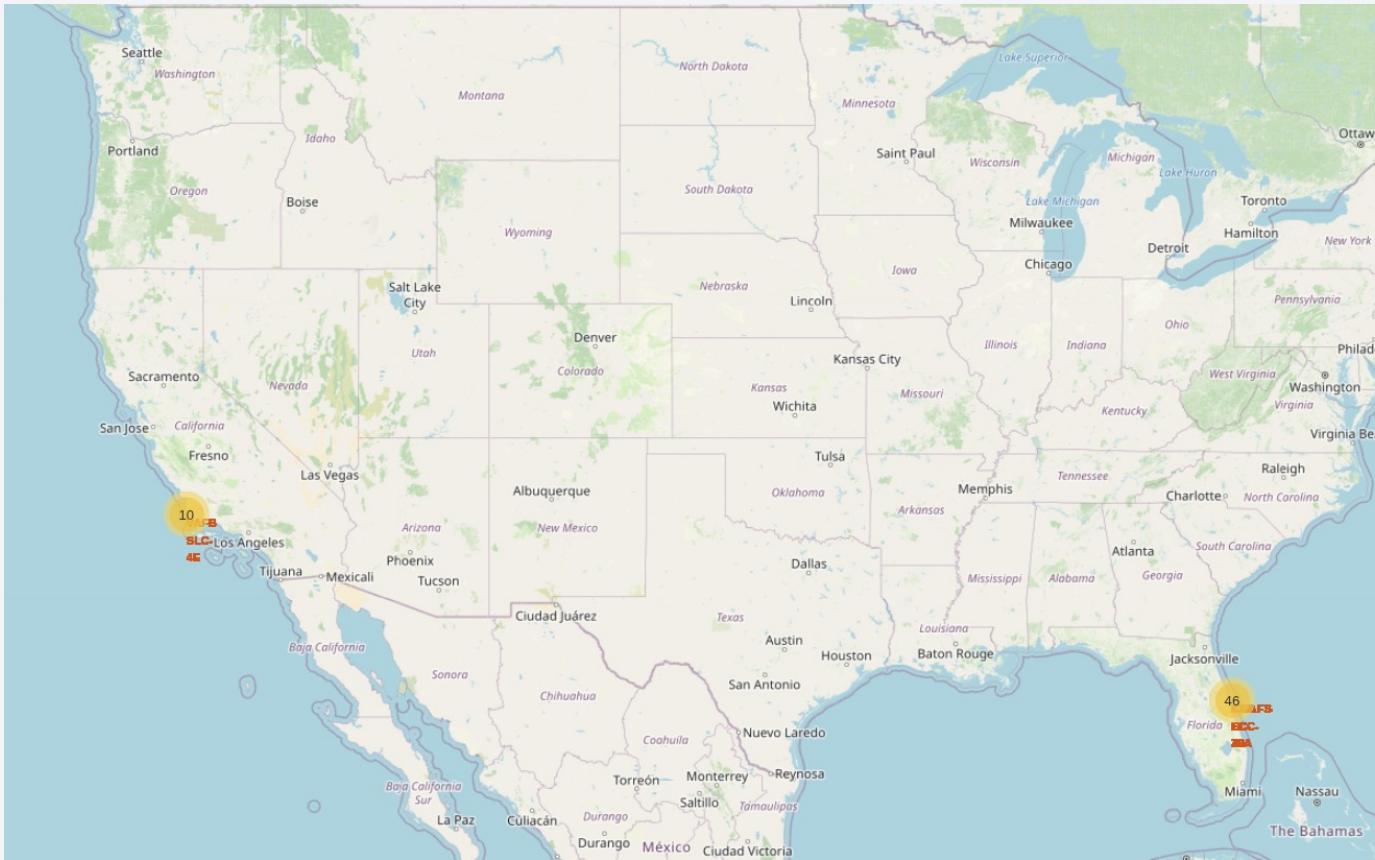
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth's horizon against a dark blue sky. City lights are visible as small white dots, and larger clusters of lights indicate major urban centers. In the upper right quadrant, there is a bright, horizontal band of light, likely the Aurora Borealis or Southern Lights.

Section 3

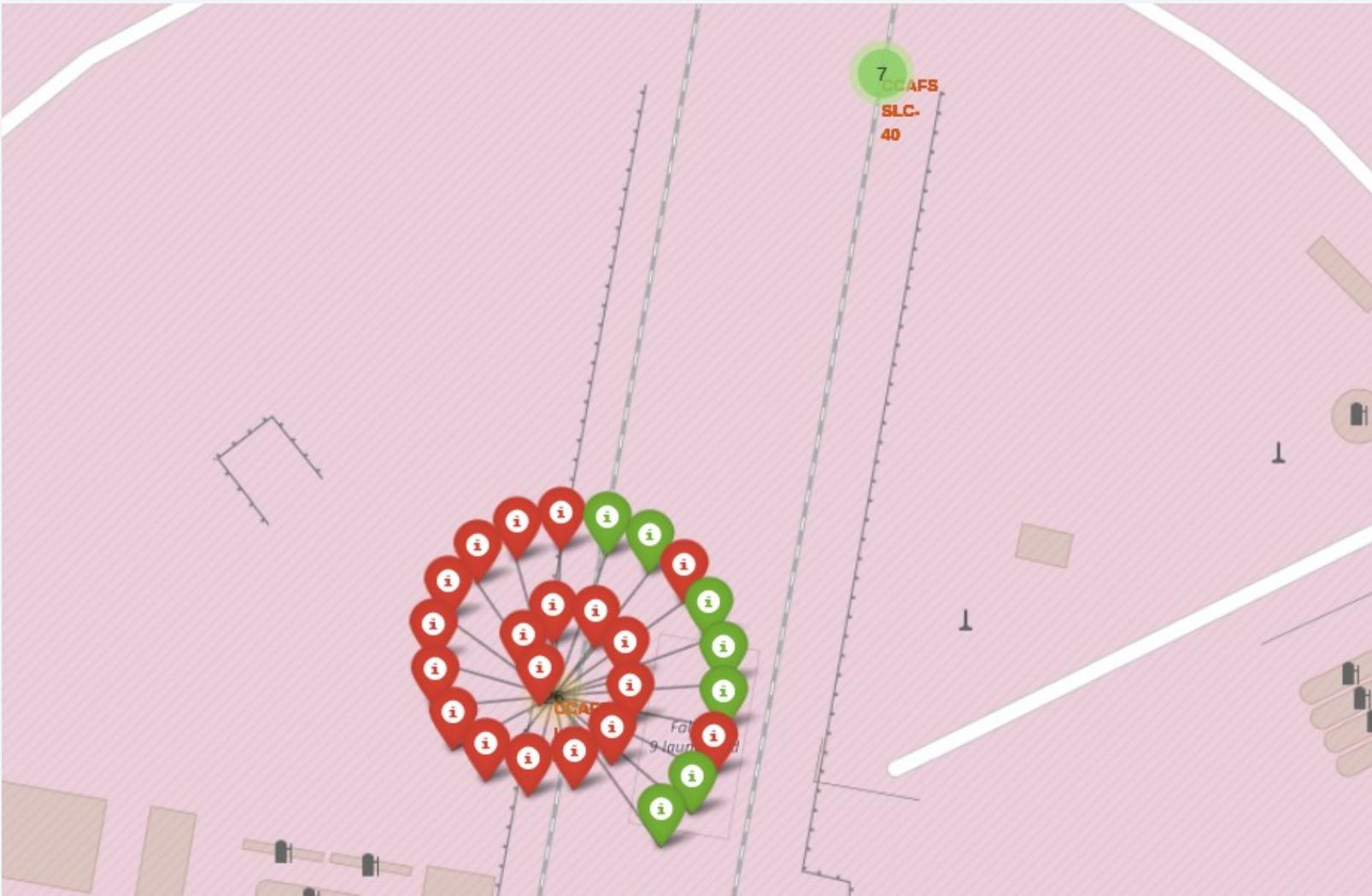
# Launch Sites Proximities Analysis

# Folium Launch Site Map



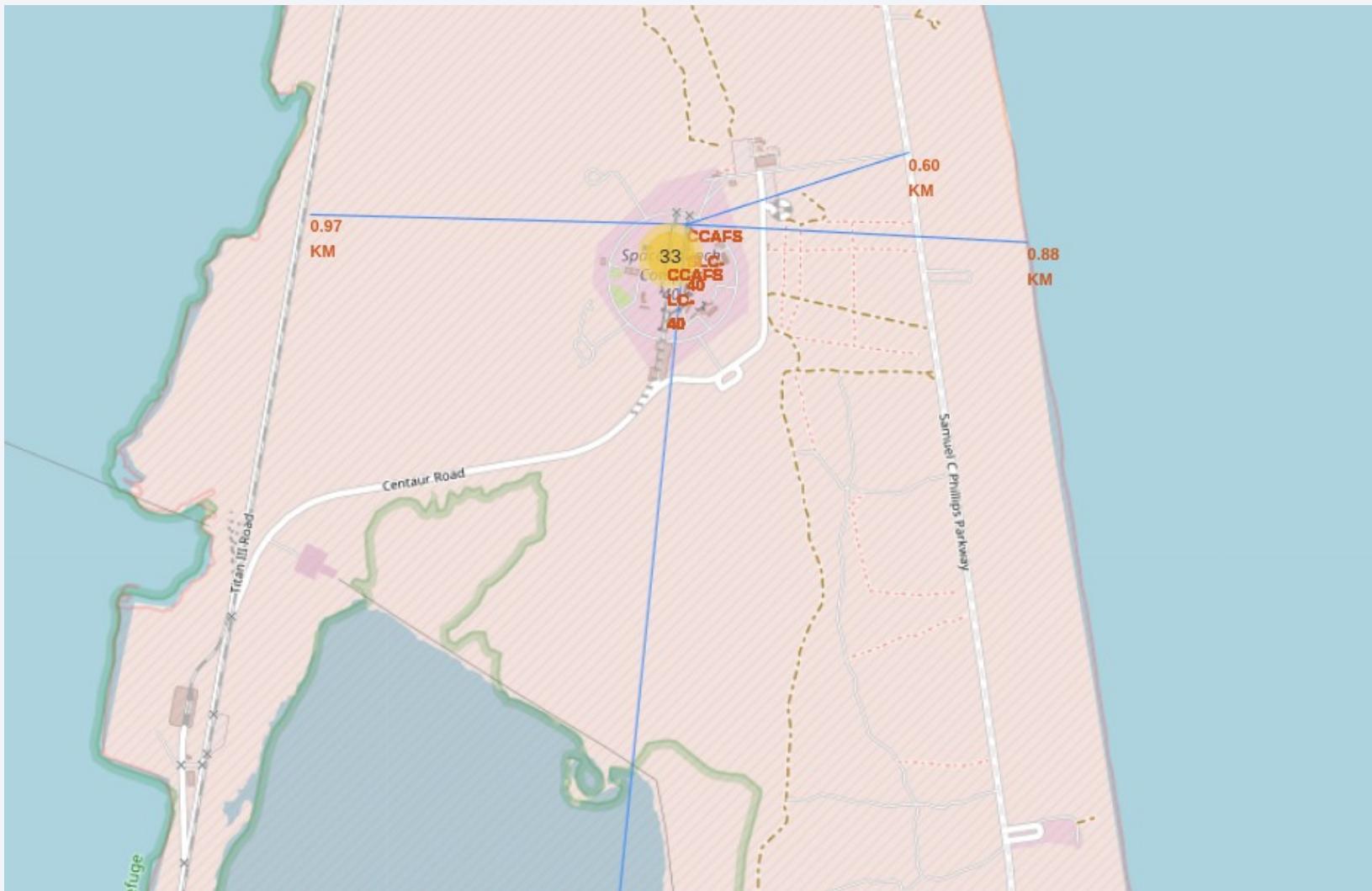
SpaceX primarily launched from the east coast.

# Folium Launch Outcome Map



Each launch marker expands to show the number of successful (green) and unsuccessful (red) launches at a given site.

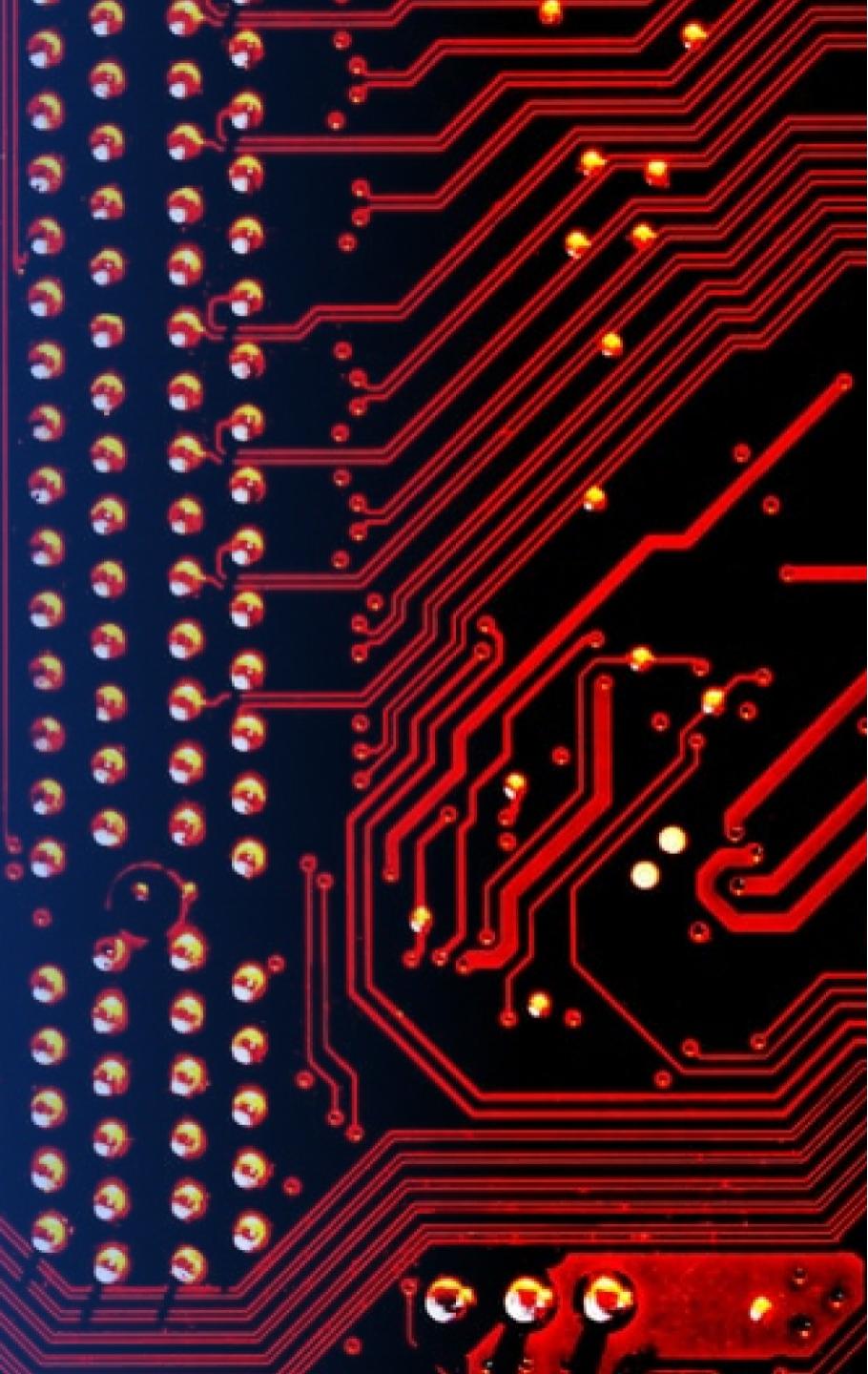
# Folium Launch Site Proximity Map



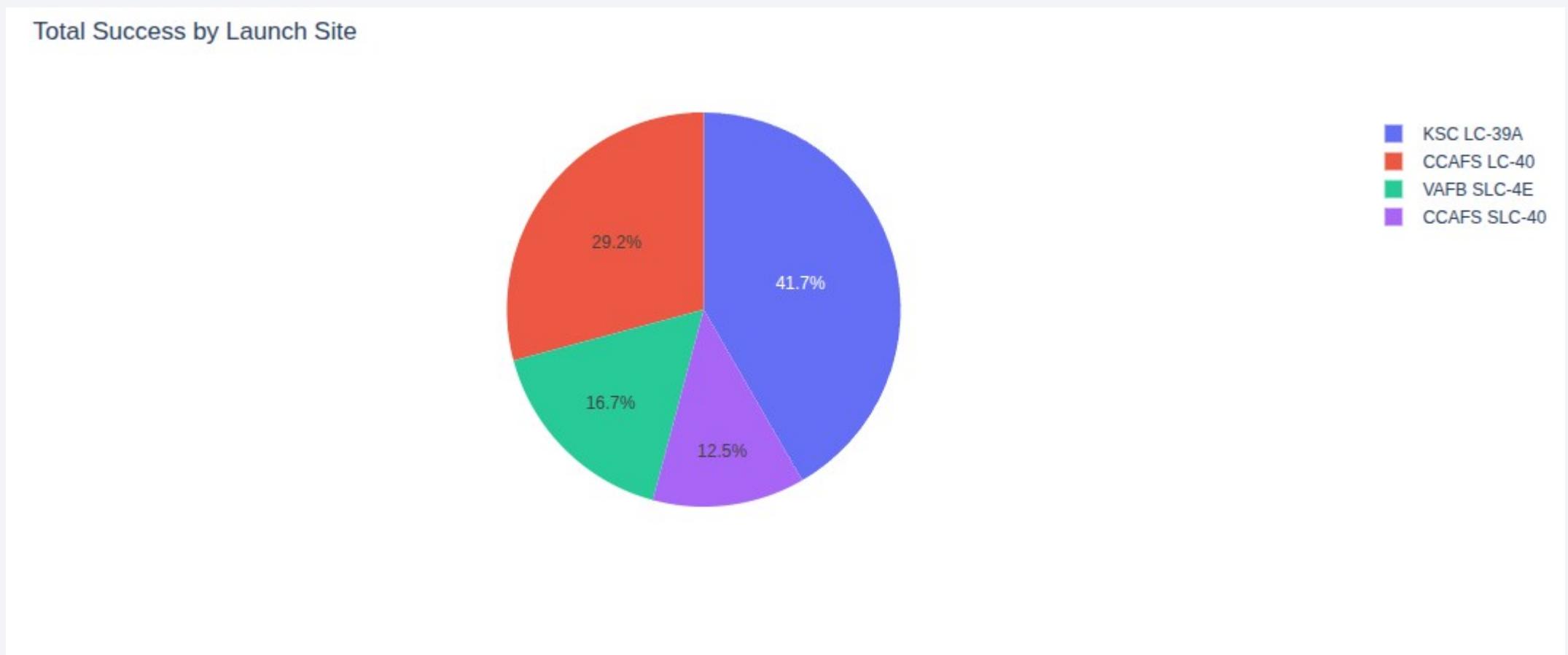
Launch sites have close proximity to coast lines and supporting infrastructure, such as railroads and highways, while maintaining a minimum 15km distance from cities.

Section 4

# Build a Dashboard with Plotly Dash



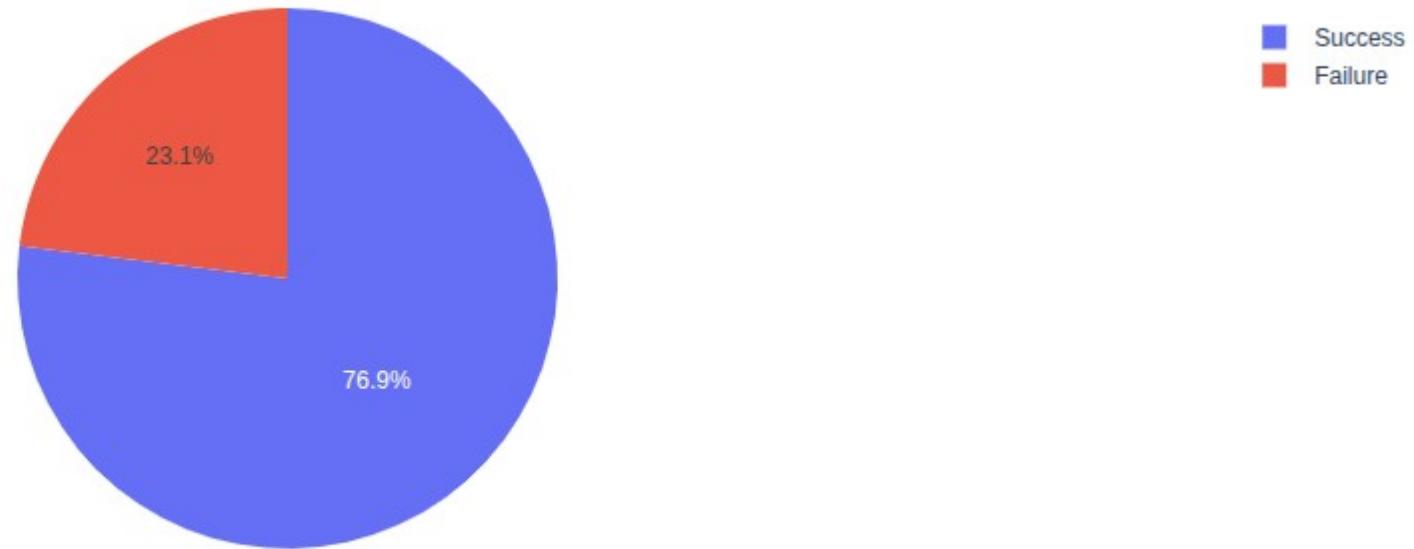
# Total Success by Launch Site



Most successful launches originated from KSC LC-39A

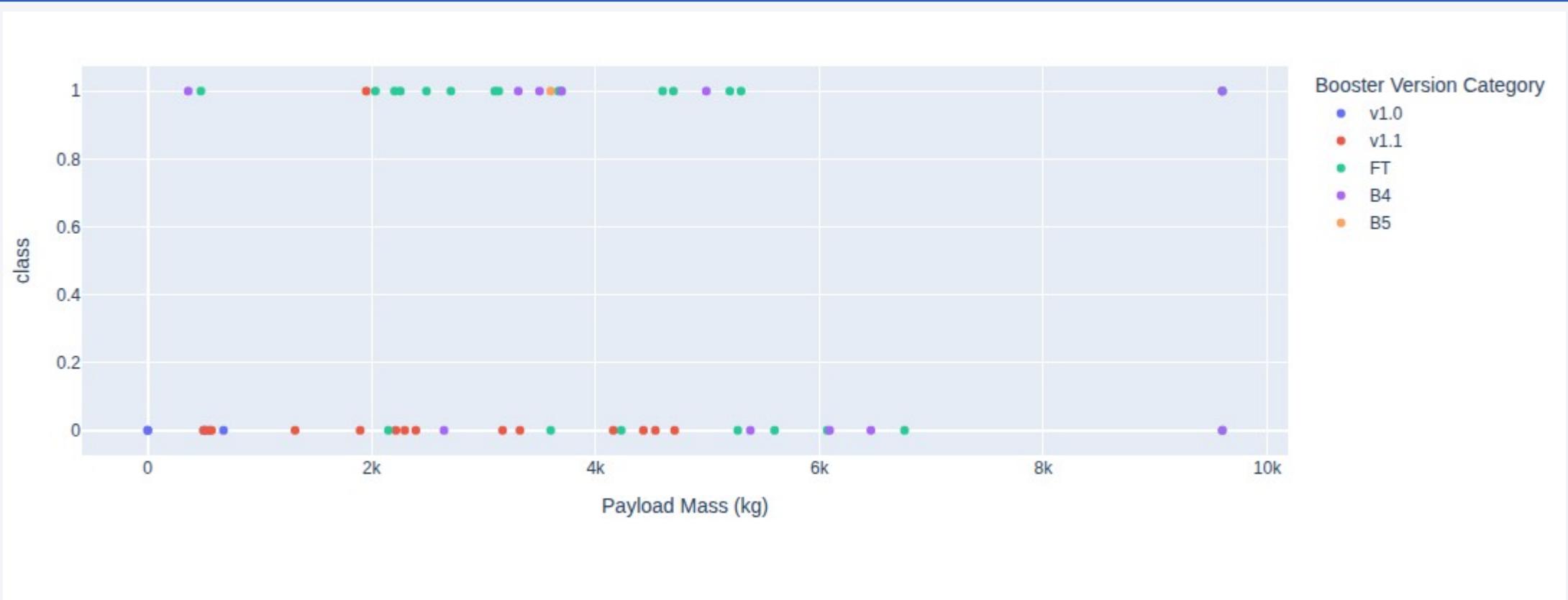
# Total Success vs. Failure of KSC LC-39A

Total Success vs Failure for site KSC LC-39A



KSC LC-39A had a 76.9% Success rate and 23.1% Failure rate.

# Payload Outcome for All Sites



The most successful payload range was between 2000 kg – 4000 kg.

The most failures for a payload range was between 4000 kg – 7000 kg.

The FT Booster Version was the most successful booster.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

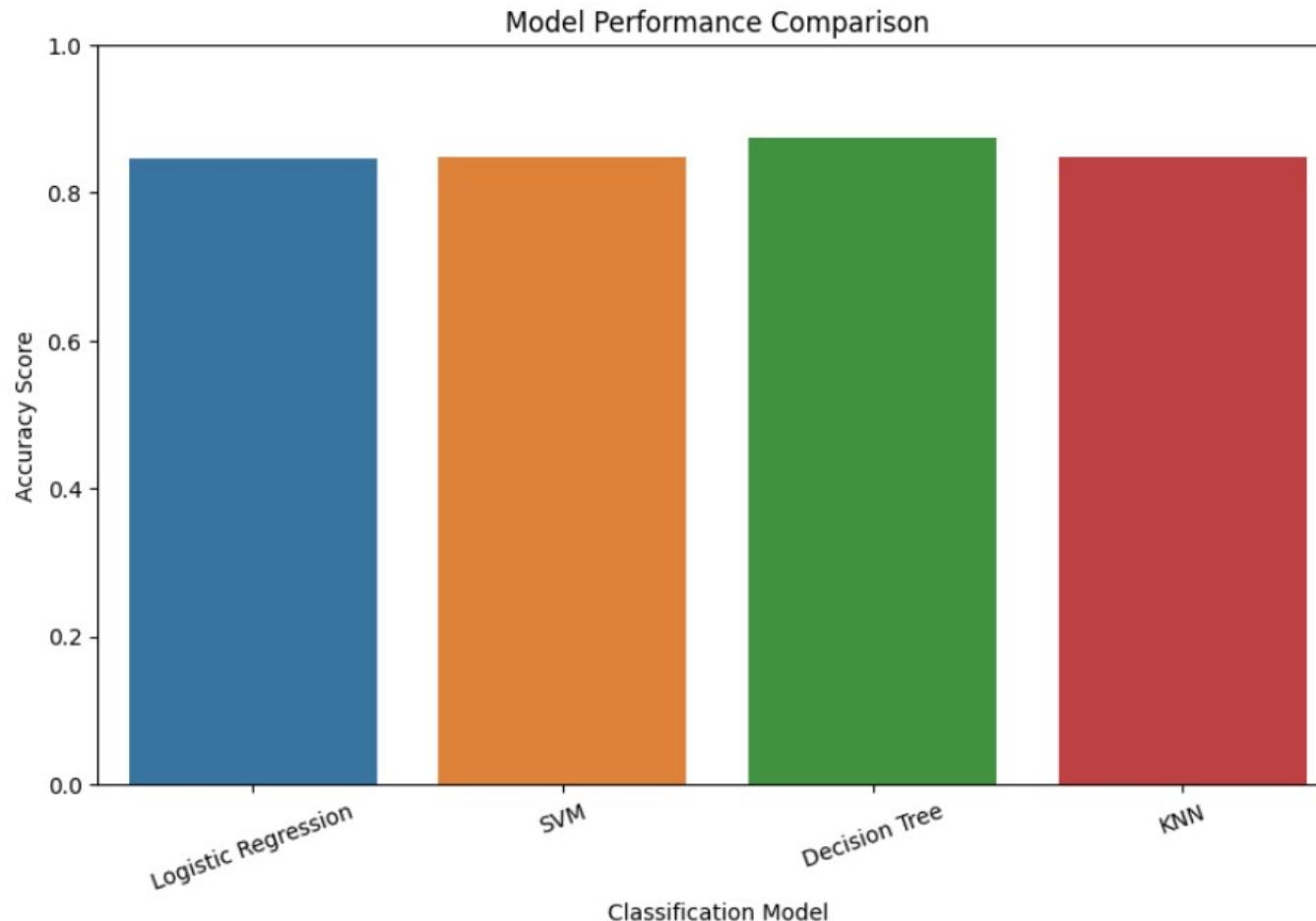
Logistic Regression score: 0.8464285714285713

SVM score: 0.8482142857142856

Decision Tree score: 0.875

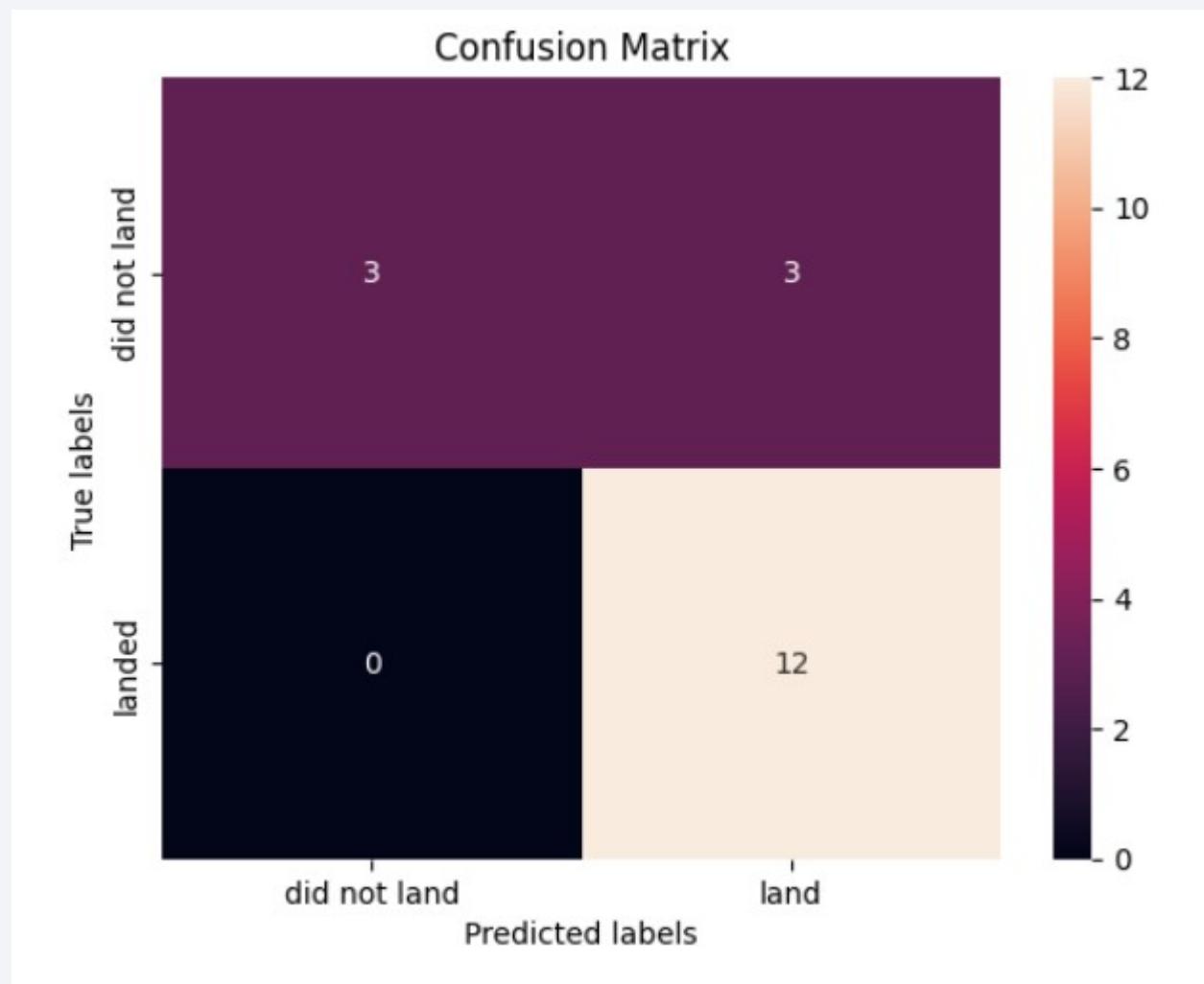
KNN score: 0.8482142857142858

Decision Tree has the best score: 0.875



# Confusion Matrix

The Decision Tree model not only had the highest accuracy (~87%), but its confusion matrix confirmed that it predicts both successes and failures reliably, with only a small number of missed classifications. This balance makes it the strongest candidate for forecasting future mission outcomes.



# Conclusions

---

- Comprehensive Analysis
  - Collected, cleaned, and integrated SpaceX data from API & web scraping
  - Built interactive dashboard for exploratory insights
- Key Findings
  - Launch success varies by site, payload, orbit, and booster version
  - Success rates have steadily improved year over year
  - Geospatial context highlights importance of site and landing location
- Predictive Modeling
  - Tested multiple ML models with cross-validation
  - Decision Tree achieved best performance (~87% accuracy)
  - Confusion matrix confirmed strong reliability in predicting outcomes
- Impact
  - Provides both historical insights and a predictive framework
  - Supports better understanding of what drives successful launches

# Appendix

---

[GitHub link](#) to the entire repository.

Thank you!

