Data Science capstone

Collin Youngerman

June 05, 2025

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

Executive Summary

Methodologies

- Applied CRISP-DM methodology to structure the predictive analysis lifecycle
- Utilized data wrangling and feature engineering on Falcon 9 launch records
- Conducted exploratory and interactive visual analysis to identify influential variables
- Trained and validated multiple classification models (SVM, Logistic Regression, Decision Tree)
- Evaluated models based on validation performance and test accuracy

Results

- Strong correlation found between landing success and features like orbit type, payload mass, and launch site
- Decision Tree classifier outperformed all others
- Model output supports future decision-making on launch strategy and risk assessment
- Sigmoid kernel yielded the best result for support vector machines during validation

Introduction

- This capstone project focuses on predicting whether the first stage of SpaceX's Falcon 9 rocket will land successfully using machine learning classification models.
- Falcon 9 launches cost approximately $62 million, significantly less than competitors, due to the reusability of its first stage. Accurately predicting landing success helps estimate launch costs and supports competitive bidding for launch contracts.
- While some landings are planned to fail (e.g., controlled ocean landings), most are intended to succeed.
- Central Question
  - Given features like payload mass, orbit type, launch site, and rocket version, can we predict whether the first stage will land successfully?

Section 1

Methodology

Methodology

Overall methodology includes

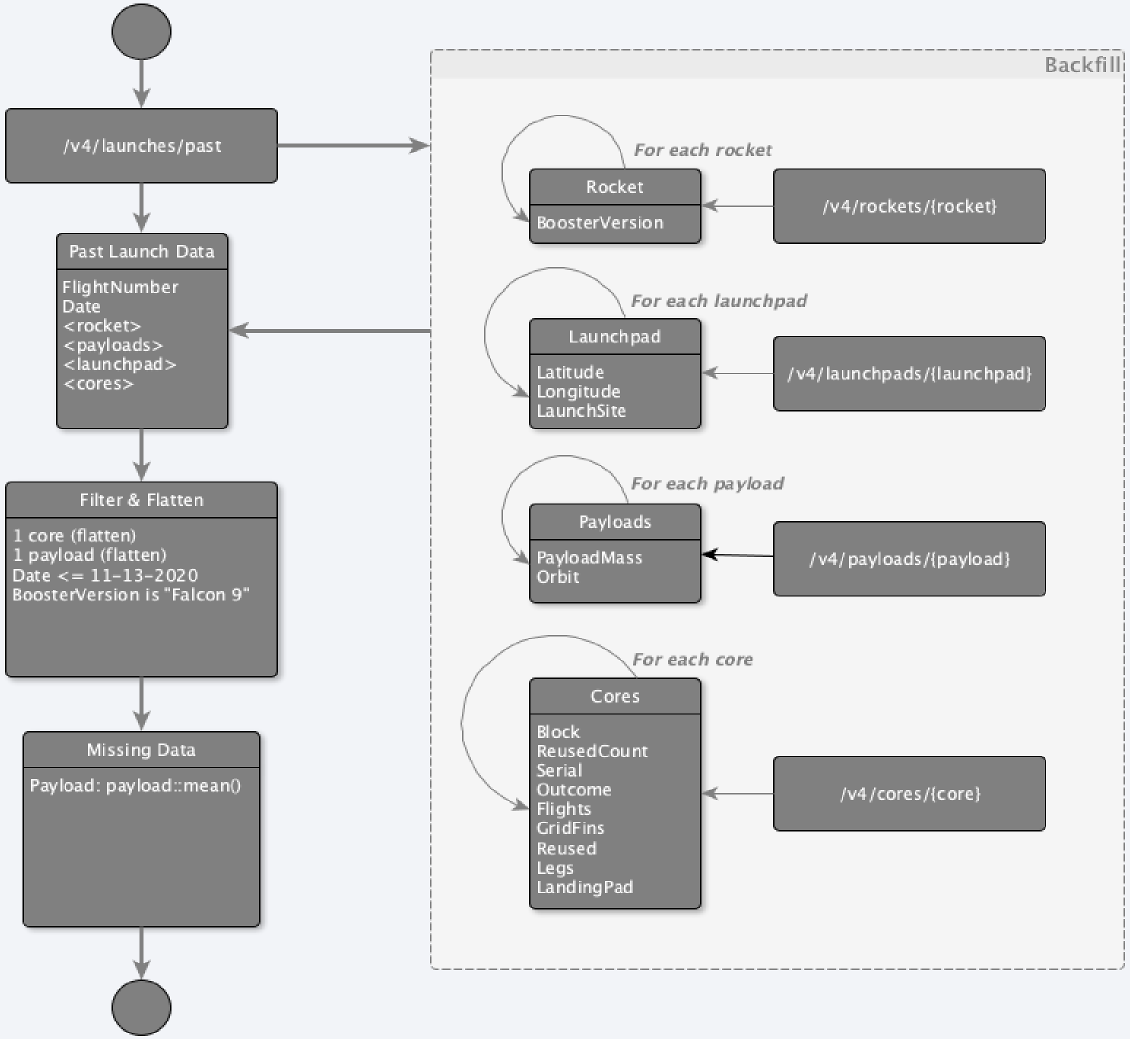
- Data collection:
  - Extracted Falcon 9 launch data via the official SpaceX API and web scraping from SpaceX's site
- Data wrangling:
  - Data was process and cleaned using Pandas and NumPy, ensuring structure and consistency for modeling
- Exploratory data analysis (EDA), using:
  - Using python and SQL I investigated launch outcomes with respect to orbit type, payload mass and launch site.
- Predictive Modeling
  - Built and evaluated classification models:
    - Logistic regression, svm, decision tree, and KNN
  - Tuned models on validation data and tested accuracy

Data Collection

- Data was collected using a combination of retrieval techniques:
  - HTTP requests used against various SpaceX API endpoints:
    - Primary launch records retrieved from: /v4/launches/past
    - Supplementary data backfilled from:
      - /v4/rockets
      - /v4/launchpads
      - /v4/payloads
      - /v4/cores
    - Wikipedia
      - Tabular data referenced from the list of Falcon 9 and Falcon Heavy launches

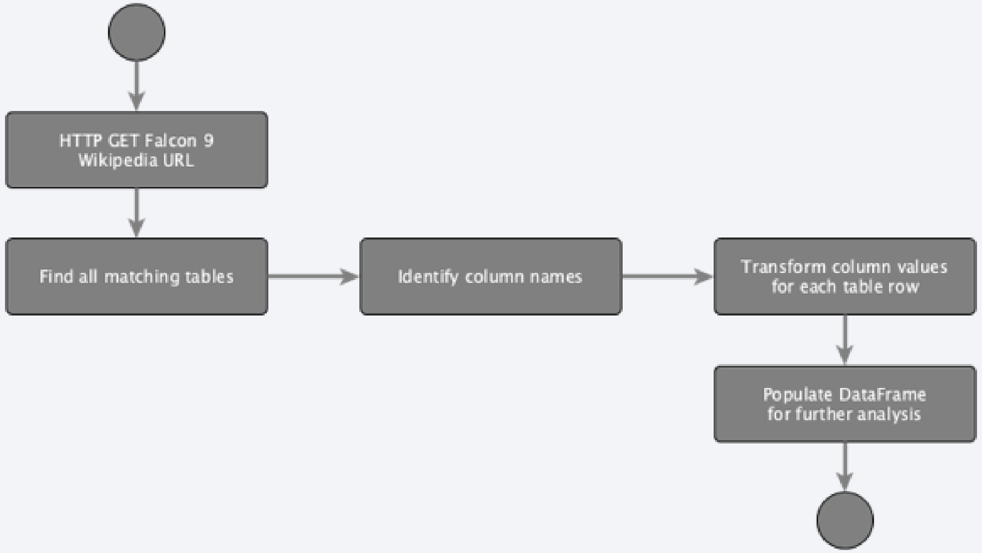
Data Collection – SpaceX API

- Primary launch data retrieved from the /v4/launces/past endpoint
- Additional data was backfilled using corresponding IDs from :
  - /v4/rockets
  - /v4/launchpads
  - /v4/payloads
  - /v4/cores
- Objective:
  - To enrich the core dataset by merging related records and building a complete view of each Falcon 9 launch.
- Reference
  - Full implementation notebook: [Notebook (GitHub)]()



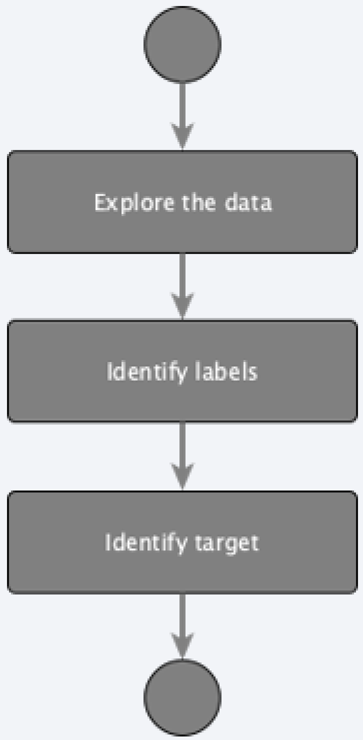
Data Collection - Scraping

- Objective
  - To extract structured launch data from the Falcon 9 Wikipedia page when unavailable via API
- Overview of Process
  - Sent HTTP request to retrieve HTML content from the official Falcon 9 launch table
  - Parsed HTML using BeautifulSoup to locate relevant tables
  - Extracted launch details and transformed into a structured Pandas DataFrame for analysis
- Reference
  - Full implementation: [Notebook (GitHub)]



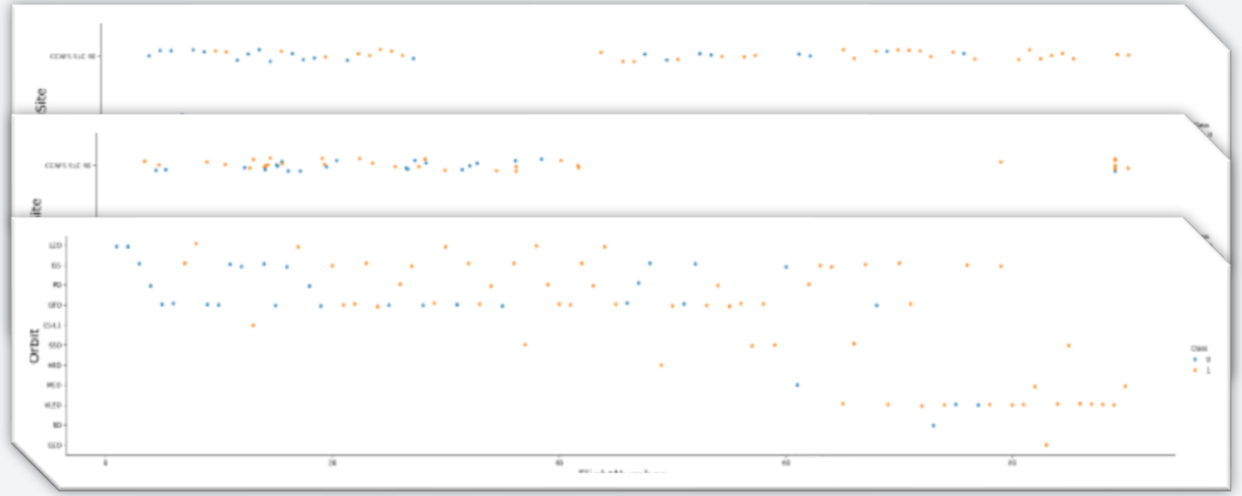
Data Wrangling

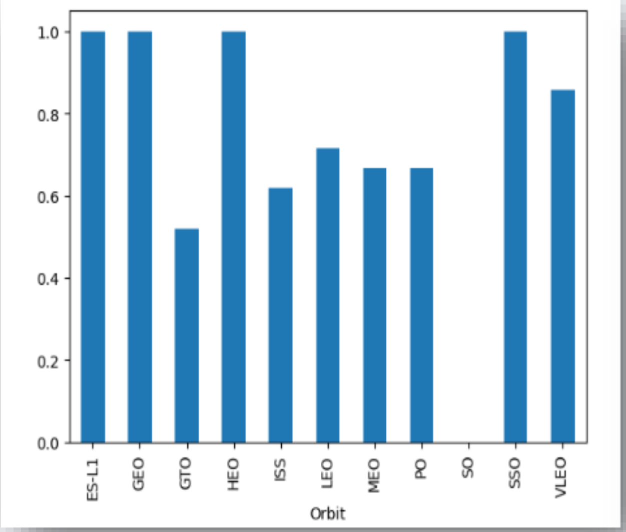
- Purpose
  - Analyze and clean the dataset to support supervised learning
  - Identify meaningful patterns and assign binary class labels for model training
- Key Actions
  - Assessed missing values across all attributes
  - Classified columns as numerical or categorical
  - Show launch distribution by site and orbit type
  - Transformed outcome values into a binary “Class” label (Success vs. Failure) for modeling
- Full code available in [Notebook (Github)]

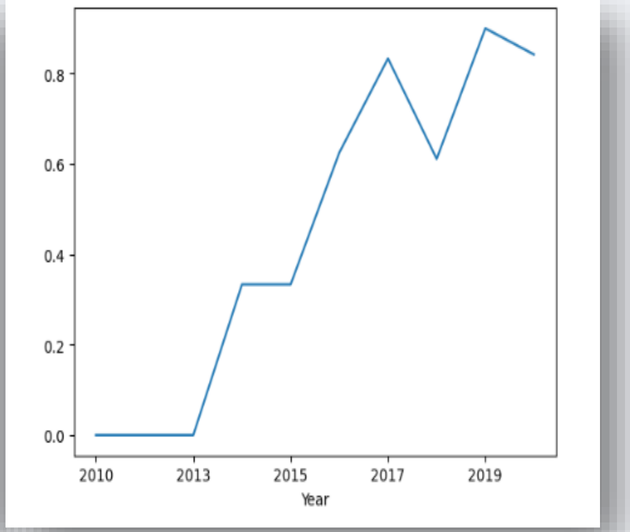


EDA with Data Visualization

- Objective:
  - Identify feature relevance and patterns driving launch outcomes through visual analysis
- Key Insights Explored:
  - Correlation between:
    - Flight Number ⟷ Outcome
    - Launch site ⟷ Outcome
    - Payload ⟷ Launch Site & Orbit
    - Orbit ⟷ Outcome
    - Yearly success rate
  - Data Transformation
    - Converted categorical features to dummy variables for more input
    - Cast numerical columns to float64 to ensure compatibility with machine learning algorithms
- Reference:
  - [Notebook (Github)]







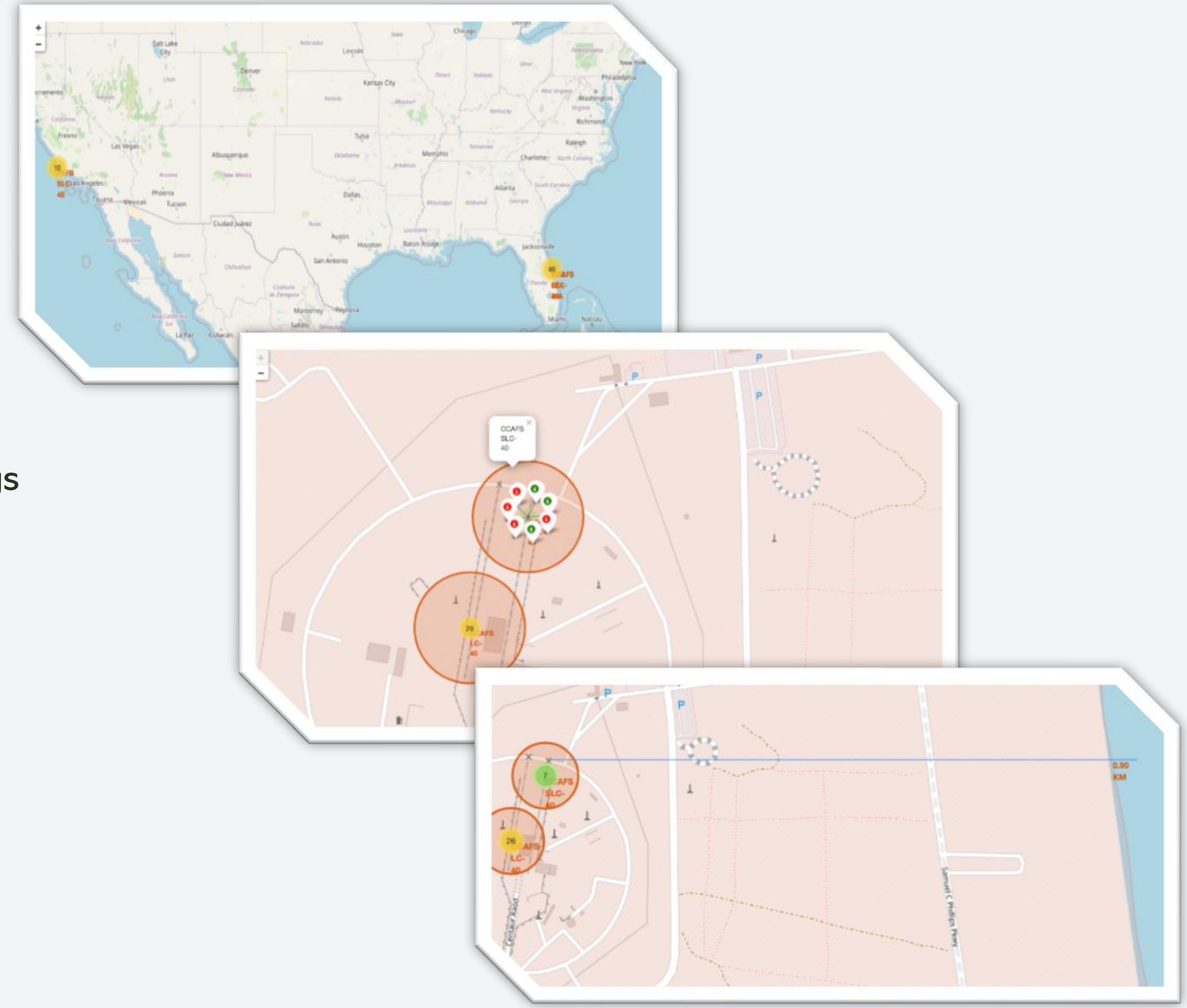
EDA with SQL

- Purpose
  - Utilized SQL queries to explore mission-level insights, identify patterns, and extract key statistics from the launch dataset
- Representative Queries Executed
  - Retrieve distinct launch sites and sample missions from each (LIKE 'CCA%')
  - Aggregate total and average payload mass by booster version and customer (e.g., 'NASA (CRS), v1.1)
  - Identify the first successful ground landing
  - Filter booster versions with successful drone ship landings carrying payloads between 4,000 and 6,000kg
  - Summarize total successful vs failed outcomes
  - Join and filter on multiple dimensions to:
    - Identify boosters with max payloads and failure outcomes
    - Analyze failure trends on drone ship landings (2015)
    - Visualize mission outcome trends between June 4th, 2010 and March 20th, 2017

- [Notebook (GitHub)]

Build an Interactive Map with Folium

- Objective:
  - Leverage geospatial visualizations to uncover spatial trends in launch performance and site distribution
- Mapped Insights
  - Launch Site Locations
    - Displayed all Falcon 9 launch sites on an interactive map
  - Mission Outcomes
    - Differentiated between successful and failed landings with color-coded markers
  - Geographic Context
    - Calculated distances from each site to nearby infrastructure or landmarks
- Outcome
  - Enhanced spatial understanding of mission patterns, enabling visual clustering of high-success launch infrastructure or landmarks
- [Notebook (GitHub)]

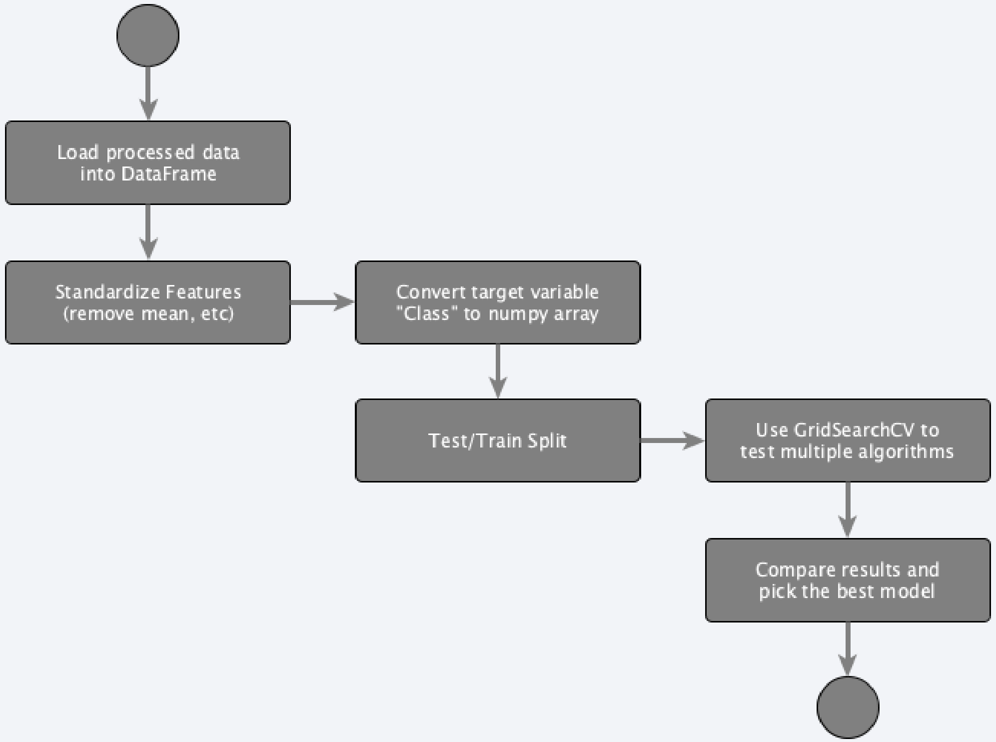


Build a Dashboard with Plotly Dash

- Objective
  - Enable interactive data exploration through a dynamic dashboard that allows users to uncover launch performance patterns by site and payload characteristics
- Key Features
  - Launch Site Selector (Dropdown)
    - Filters visuals based on selected site
  - Pie Chart Visualization
    - All Sites: Displays success distribution across all launch sites
    - Selected Site: Compares success vs failure for the chosen site
  - Scatter Plot Visualization
    - All Sites: Plots mission outcome by payload mass and booster version
    - Selected Site: Focuses on missions from that specific location
  - Payload Mass Range Filter
    - Interactively filters data points on the scatter plot to study trends by payload size
- Notebook (Github)

Predictive Analysis (Classification)

- Objective
  - Train and evaluate machine learning models to predict the success of Falcon 9 first-stage landings
- Process Overview
  - Data Preparation
    - Load processed data into a Pandas DataFrame
    - Standardize feature set X using StandardScaler
    - Convert target labels Y to NumPy array
  - Train/Test Split
    - Partition dataset into training and testing subsets
  - Model Tuning with GridSearchCV
    - Perform hyperparameter optimization across multiple classification algorithms:
      - Logistic Regression
      - Support Vector Machine (SVM)
      - Decision Tree Classifier
      - K-Nearest Neighbors (KNN)
- Outcome
  - The best-performing model is selected based on test accuracy and validation performance after tuning (Decision Tree)
- [Notebook (GitHub)]



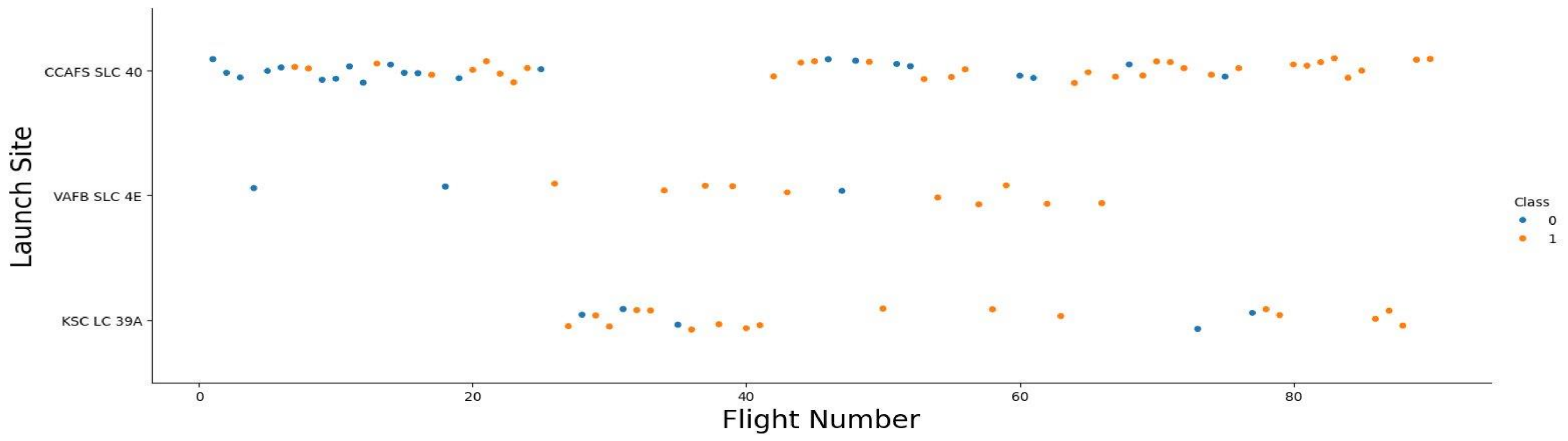
Results Slides to Follow

- Exploratory data analysis results
- Launch sites proximities analysis
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

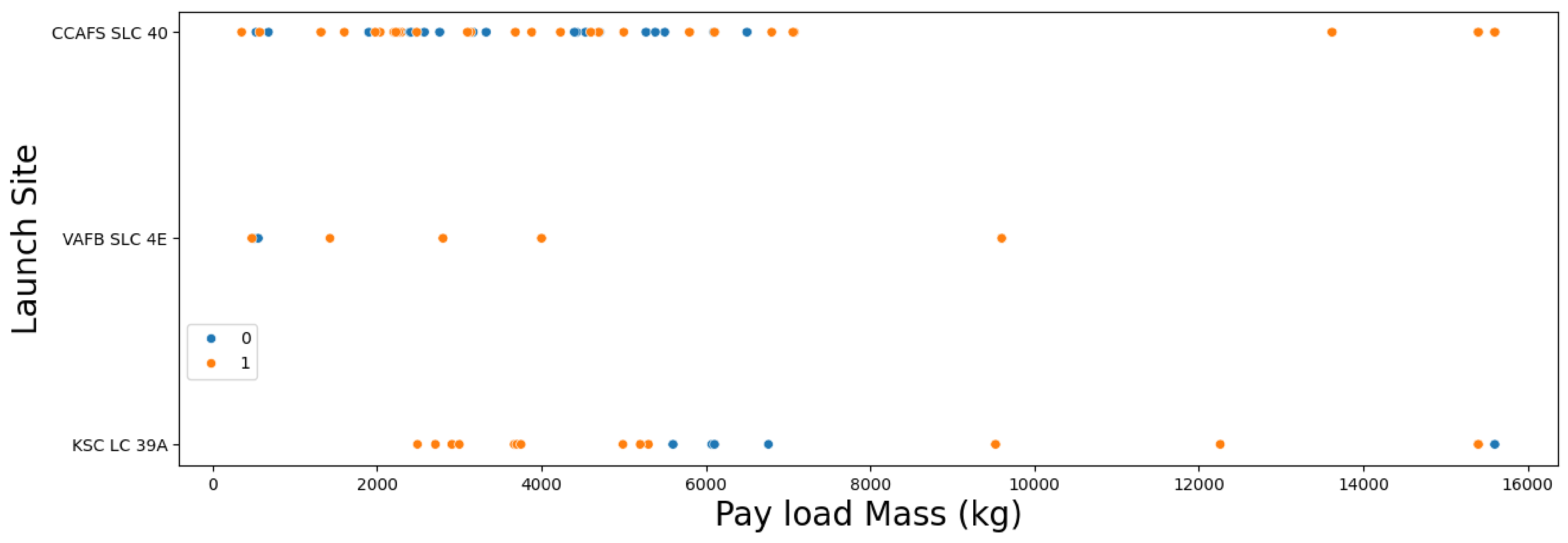
Insights drawn from EDA

Flight Number vs. Launch Site



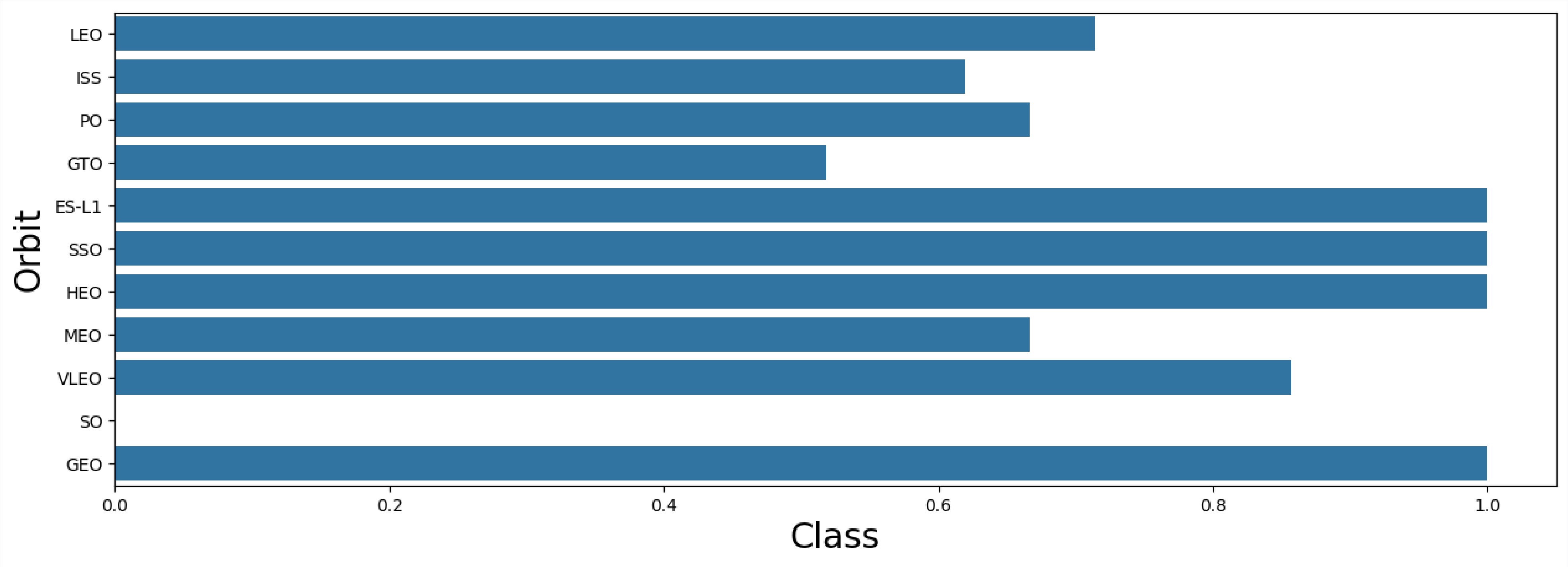
- Insight Summary
  - All launch sites exhibit a mix of landing successes and failures across flights
  - Earlier missions show a higher failure rate, highlighting a trend of technological and operational improvement over time.
  - CCAFS SLC 40 handled the largest number of launches, but:
    - VAFB SLC 4E shows a higher success rate relative to its flight count.
  - Success Outcomes (orange) become more frequent with higher flight numbers, suggesting maturation of Falcon 9's recovery system

Payload vs. Launch Site



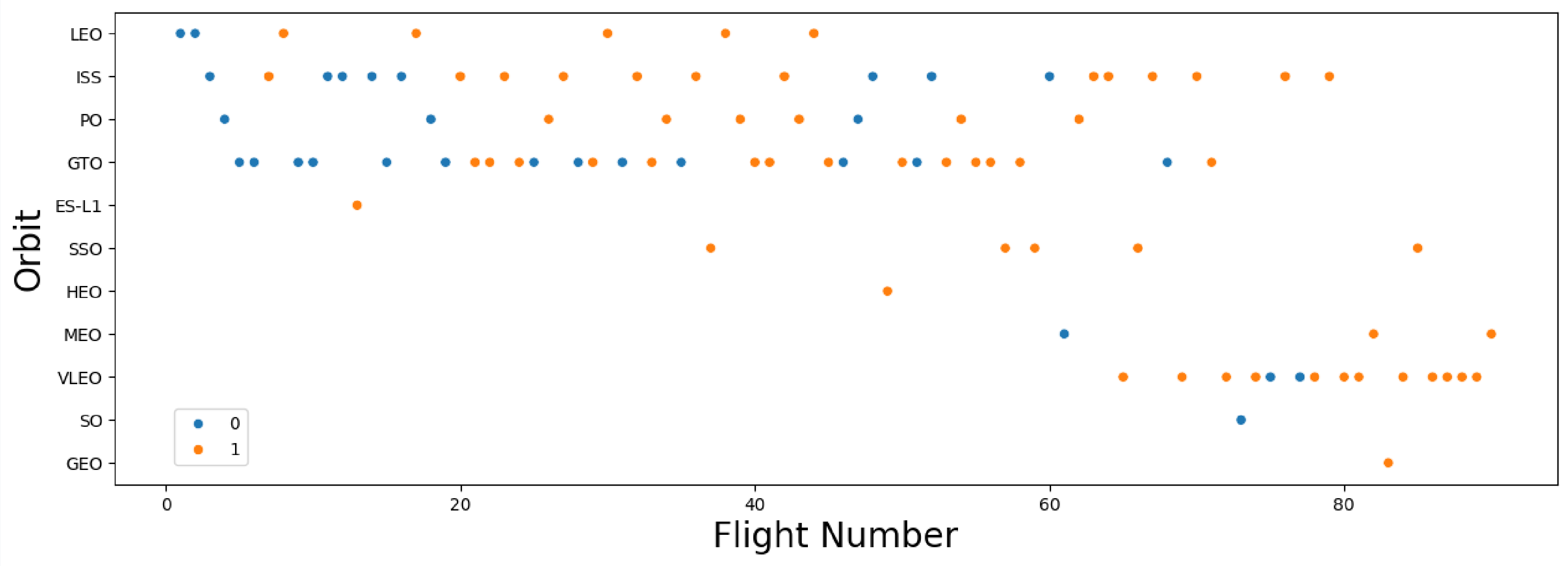
- Key Insights
  - All sites handled launches across a wide range of payload weights, from light missions to heavy lift operations.
  - Failures are more frequent at lower payload masses, which are typically associated with earlier launches.
  - As payload weight increases, so does the rate of successful landings, reflecting maturing technology and mission confidence.
  - CCAFS SLC 40 demonstrates the most frequent heavy payload launches, while VAFB SLC 4E shows consistent performance across all weights

Success Rate vs. Orbit Type



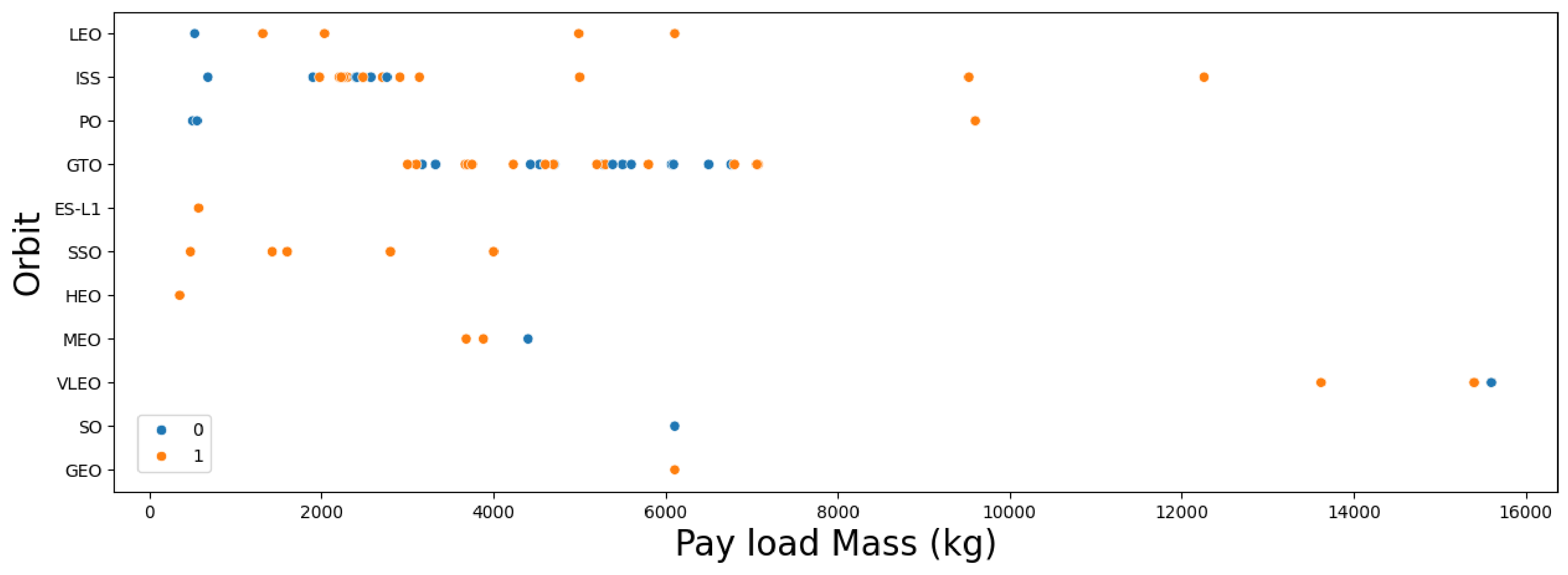
- Key Insights
  - **High Success Orbits**: Missions to ES-L1, SSO, HEO, and GEO consistently achieve 100% success, indicating strong operational reliability for these orbits
  - **Mixed Results**: GTO missions show greater variability, hinting at added mission complexity or technical risk
  - **Emerging Orbits**: Some orbits like SO have only one data point, making statistical conclusions unreliable

Flight Number vs. Orbit Type



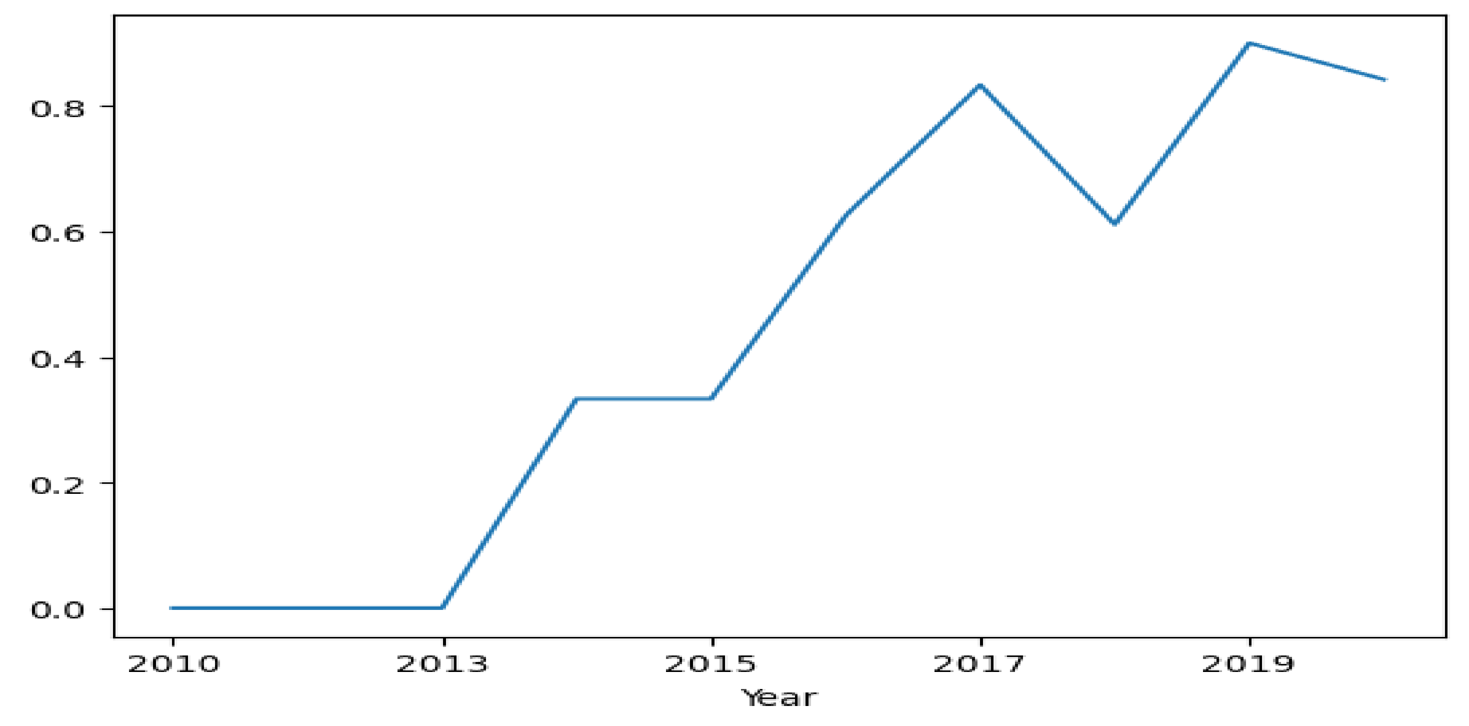
- **Orbit Diversity Grows Over Time**: Early missions are concentrated in fewer orbits. Over time, SpaceX expands into a broader range of orbits as confidence and capability increase
- **Improved Success Rate with Experience**: Later flights (higher flight numbers) show fewer failures, reflecting technical maturation, process refinement, and operational learning
- GTO & SSO missions appear later in the timeline, aligning with more complex mission readiness

Payload vs. Orbit Type



- **Wide Payload Range Across Orbits**: Most orbit types show broad payload mass variation, while others like SSO, MEO, HEO, and GEO have more constrained ranges
- **Orbit-Specific Payload Trends**: Missions to SSO and GEO orbits often involve heavier payloads, requiring more advanced technology and precise trajectory planning
- **Success Distribution**: Orbits with narrower payload profiles (SSO, MEO) tend to show higher consistency in landing outcomes, potentially reflecting better mission planning or tighter constraints

Launch Success Yearly Trend



- Key Insights
  - SpaceX demonstrates a clear upward trajectory in first stage landing success from 2013 onward, reflecting rapid technology maturity
  - Significant improvement begins in 2015, with success rates rising above 50%
  - From 2016-2020, success became increasingly reliable, peaking above 90%. This highlights consistent gains in precision landing systems
  - The minor dip in 2018 reflects a temporary challenge, but was quickly corrected in subsequent years

All Launch Site Names

There are four unique Launch Sites

- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

```
SELECT DISTINCT Launch_Site from SPACEXTABLE
```

Launch Site Names Begin with 'CCA'

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

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

Total Payload Mass

The total payload carried by boosters from NASA (CRS) is **45,596kg**.

```
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD
  FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is **2928.4kg**

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS
  FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1%'
```

First Successful Ground Landing Date

- The first successful landing outcome on ground pad occurred on **December 22nd, 2015**.

```
SELECT MIN(Date) as LaunchDate
  FROM SPACEXTABLE
  WHERE Landing_Outcome = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

Booster	Payload Mass
F9 FT B1022	4,696kg
F9 FT B1026	4,600kg
F9 FT B1021.2	5,300kg
F9 FT B1031.2	5,200kg

- SELECT Booster_Version, PAYLOAD_MASS__KG_
- FROM SPACEXTABLE
- WHERE Landing_Outcome = 'Success (drone ship)'
- AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Mission Status	Count
Failure	1
Success	100

```
SELECT CASE
    WHEN Mission_Outcome LIKE 'Success%' THEN
 'Success’
    WHEN Mission_Outcome LIKE 'Failure%' THEN
 'Failure’
END as Mission_Status, COUNT(*)
FROM SPACEXTABLE
GROUP BY Mission_Status
```

Boosters Carried Maximum Payload

- The maximum payload sent was **15,600kg**.
- The boosters that carried the maximum payload are:

```
SELECT
  DISTINCT Booster_Version,
  PAYLOAD_MASS__KG_
FROM SPACEXTABLE
  WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
  )
ORDER BY Booster_Version
```

Booster Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Outcome	Booster	Launch Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
SELECT
    CASE strftime('%m', Date)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as Month,
    Landing_Outcome, Booster_Version, Launch_Site, Date
FROM SPACEXTABLE
WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = '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:

```
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
```

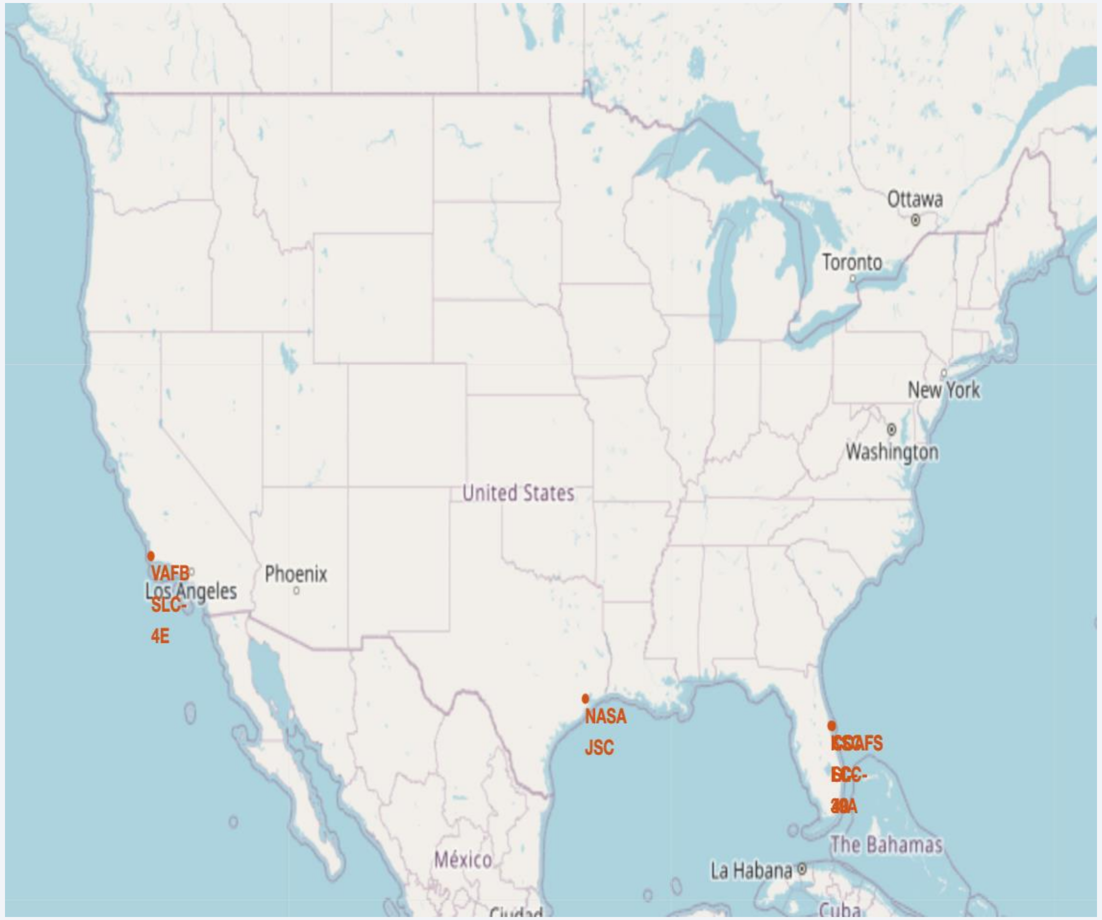
Landing Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

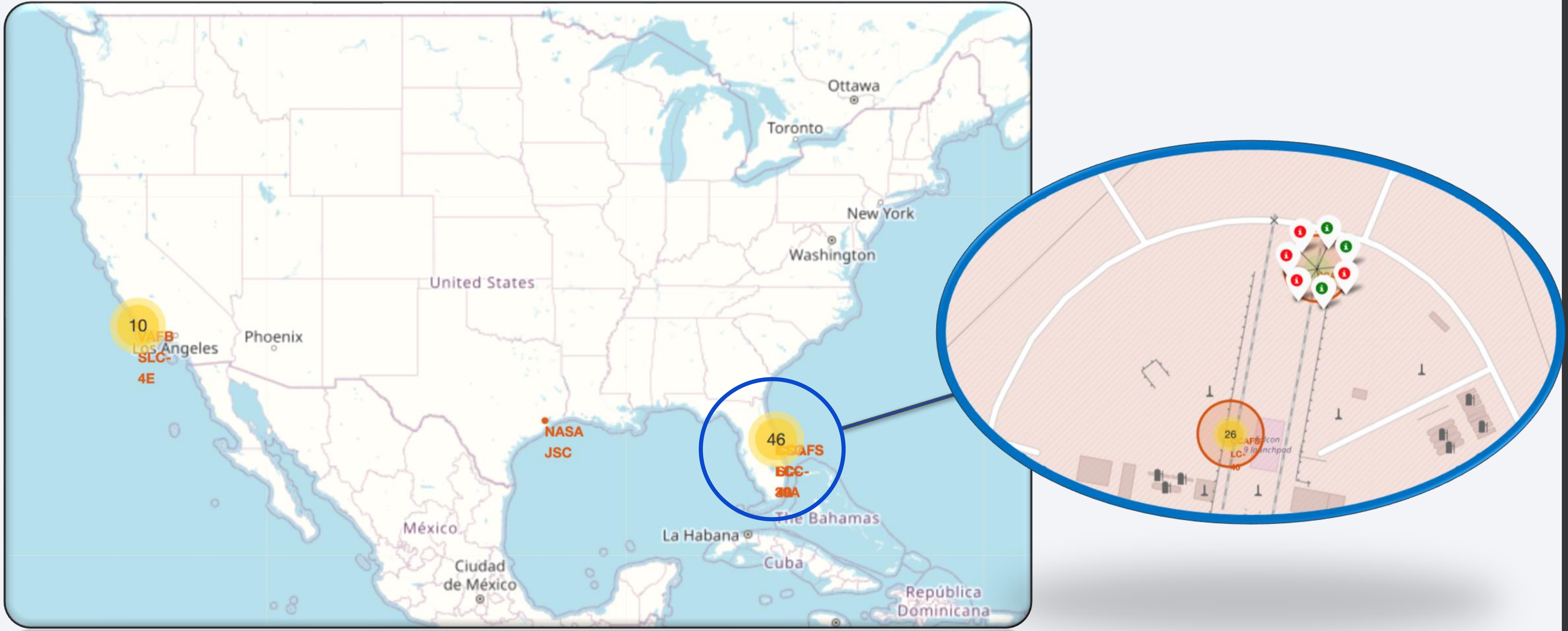
Launch Sites Proximities Analysis

Launch Site Locations

- Launch sites are positioned near coastal regions in Florida and Carolina to minimize risk to populated areas in the event of launch failure

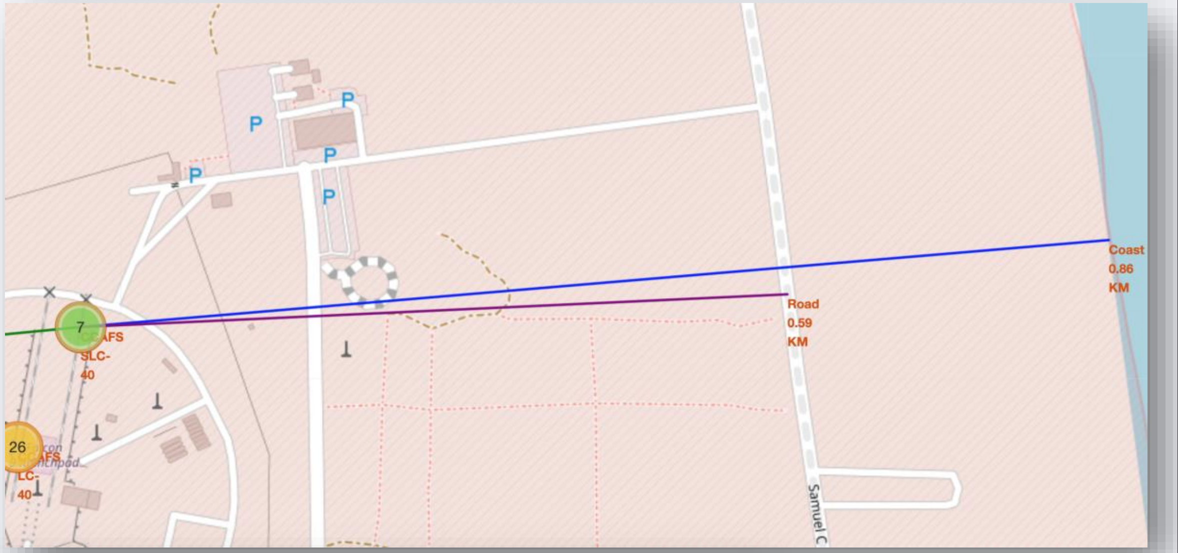


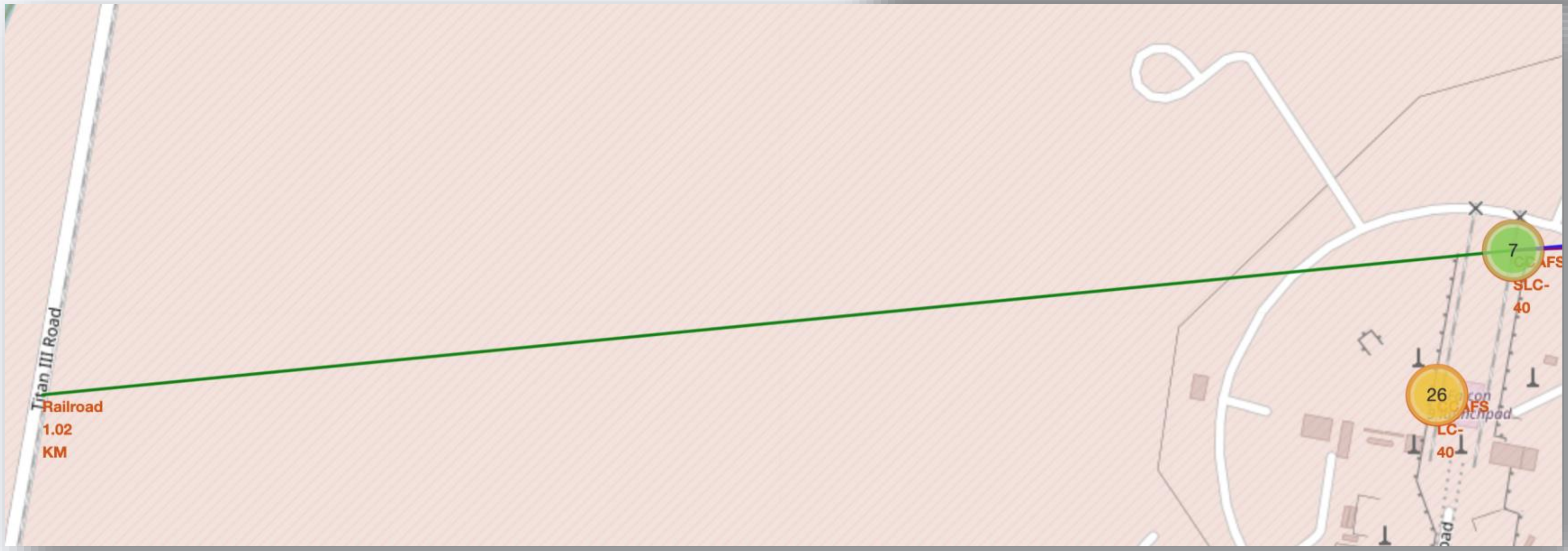
Launch Outcomes



Notable Nearby Locations

- Notable Locations
  - Railway
  - Road
  - Coast



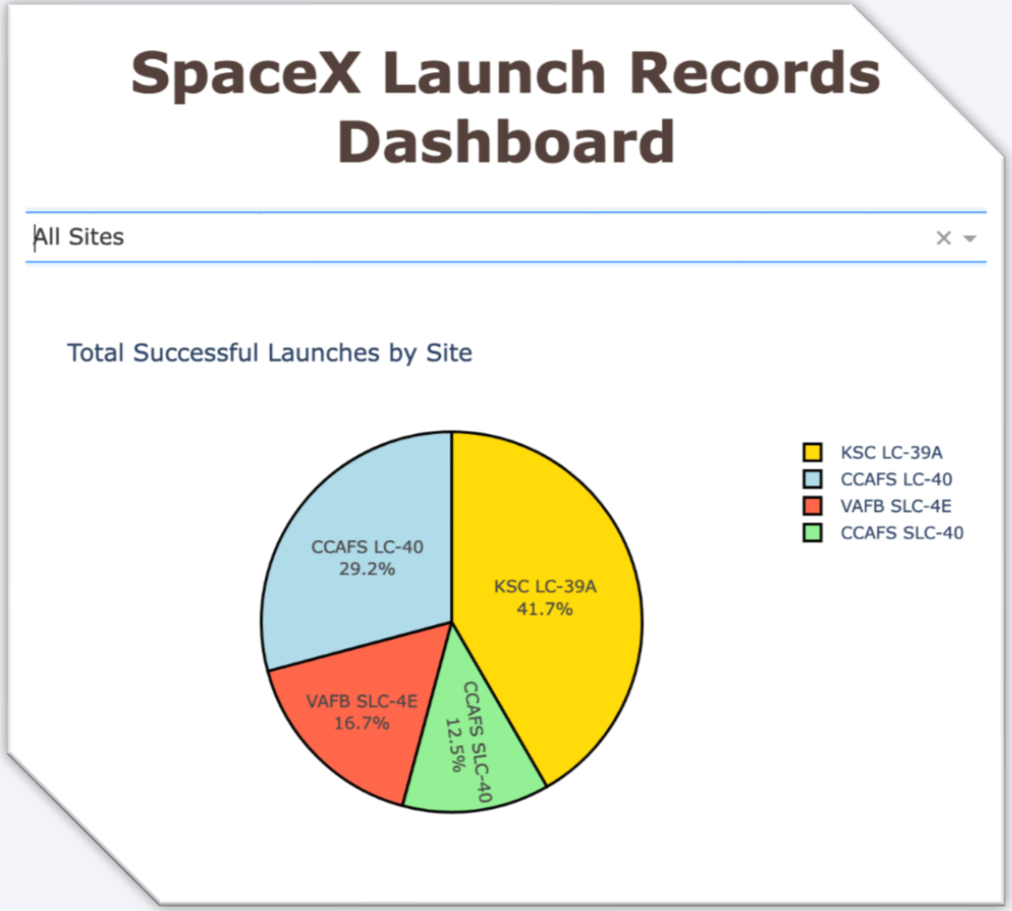


Section 4

Build a Dashboard with Plotly Dash

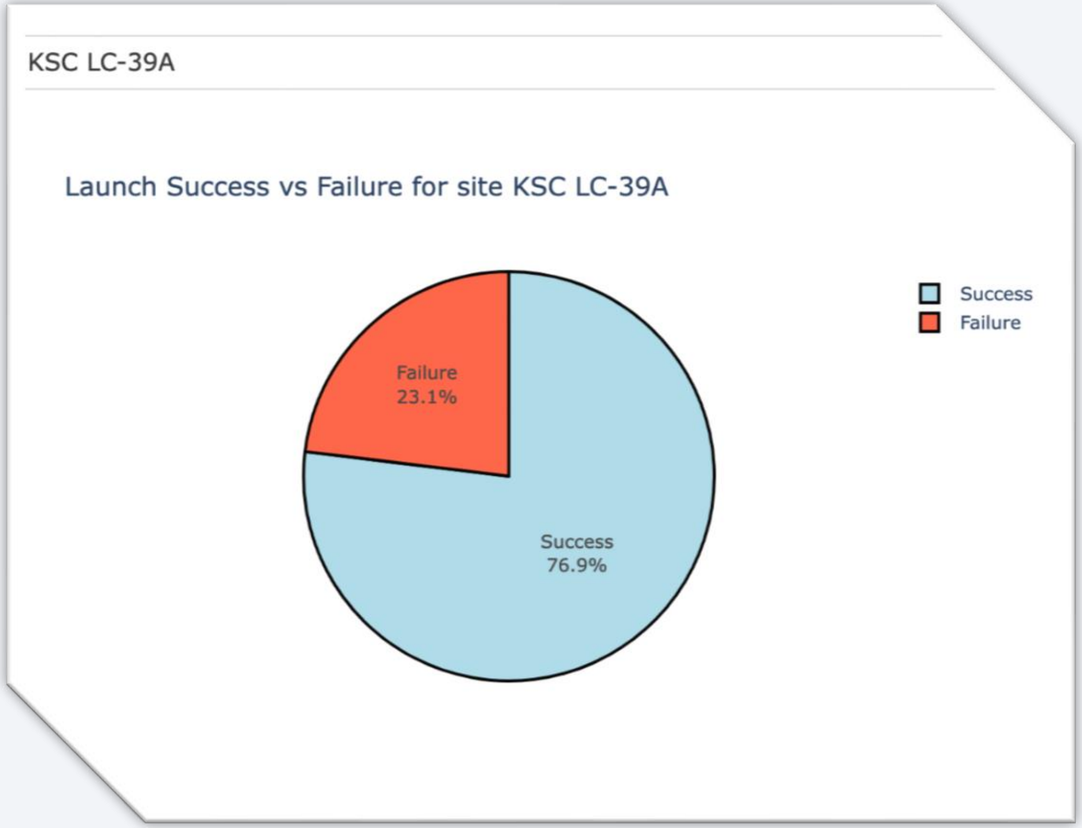
Launch Site Performance: Success Rate Comparisons

- KSC LC-39A achieved the highest success rate in first-stage landings
- CCAFS LC-40 followed with a strong volume of launches and a solid success ratio
- VAFB SLC-4E recorded the lowest success rate, highlighting variability by location



Launch Success Rate

- KSC LC-39A recorded the highest landing success rate among all launch sites, with over 75% of missions landing successfully



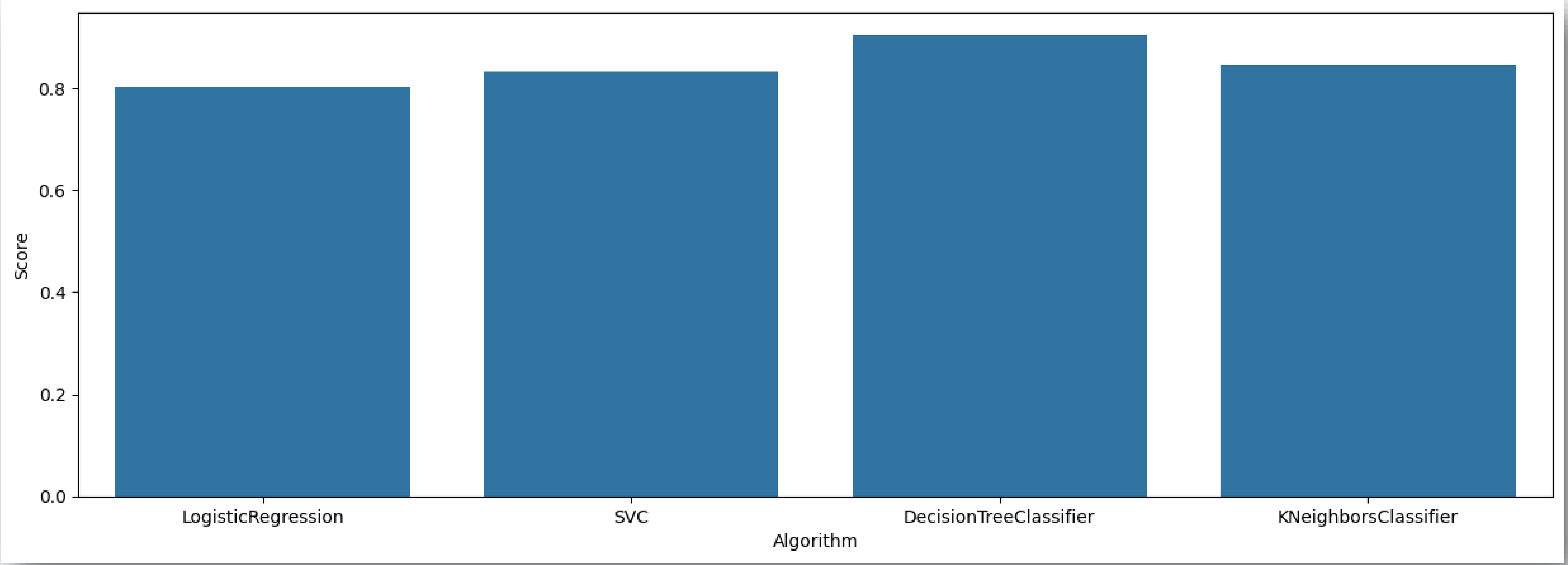
Booster Performance by Payload Range

- Between 3,000 and 5,000 kg, v1.1 boosters showed the lowest success rate
- In the same range, B4 and B5 boosters achieved the highest success rates, followed closely by FT
- Performance appears to be influenced more by booster version than payload weight alone

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree Classifier was the most accurate model

Confusion Matrix

- **14 true positives** – correctly predicted successful landings
- **3 true negatives** – correctly predicted failed landings
- **1 false positive** – incorrectly predicted failure as a success
- **0 false negatives** – No missed failed landings
- **Interpretation :**
  - The decision tree model shows strong performance, with only one misclassification and zero false negatives, making it highly reliable for predicting failed landings.

Conclusions

- Landing success rates improved over time, reflecting steady operational and technical advancement
- Orbit type impacts success rates
  - ES-L1, SSO, HEO, and GEO yielded consistently strong outcomes
- Launch site was a strong predictor
  - KSC LC-39A led in success rate, followed closely by CCAFS LC-40
- Among all models, the decision tree classifier performed best, achieving high accuracy, precision, and recall in predicting landing outcomes

Appendix

- Data Sources
  - SpaceX API
    - Data: API Dataset
    - Wikipedia
      - Tabular data referenced from the list of Falcon 9 and Falcon Heavy launches
    - Data post wrangling: spacex_web_scraped_tpf
    - Geographical data: spacex_launch_geo

Thank you!