IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Seiha Vat
10 Dec 2025

Outline

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

Executive Summary

Goal: To build and tune multiple Machine Learning models to accurately predict the success (Class = 1) or failure (Class = 0) of SpaceX rocket launches based on pre-launch parameters.

Key Finding: All highly-tuned classifiers achieved a very high, consistent accuracy (around 83% on the unseen test data), demonstrating the strong predictive power of the mission telemetry and characteristics.

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using API and web scraping
- Perform data wrangling
  - Imputing missing data, transforming and feature engineering data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data preprocessing, Model selection and Tuning

Data Collection

The raw data for this analysis was not sourced from a single file, but was systematically collected from two distinct public online sources to ensure comprehensive feature coverage:

flowchart of Data Collection:

Stream 1: SpaceX API

API Request - JSON Data

Features Collected**:** Payload Mass, Orbit, Launch Site, Target Class

Stream 2: Wikipedia

Web Scraping - using Beautiful Soup, HTML Tables

Features Collected: Historical Booster Version Details, External Context

Data Collection – SpaceX API

Key Phrases:

- Public API, Version 4, Launch Endpoint, Payload Endpoint
- Programmatic Access, GET Request, JSON Response
- JSON Parsing, Flattening Nested Data, Schema Transformation
- Target Variable Isolation, Mission Status, Primary Key

flowchart of SpaceX API calls:

- **API Call (GET Request)-** Target Endpoint
- **Receive JSON Response -** (Hierarchical Data)
- **JSON Parsing and Flattening** (Extract Payload Mass, Orbit Launch Site)
- **Create Pandas Data Frame**

GitHub URL : https://github.com/vatsun07-web/Data-Science-Capstone-Project

Data Collection - Scraping

Key Phrases:

Stable URL, HTML Structure, Structured Tables

HTTP GET Request, Python requests Library

Beautiful Soup, DOM Traversal, Tag Identification

Table Conversion, Feature Enrichment, Data Standardization

GitHub URL : https://github.com/vatsun07-web/Data-Science-Capstone-Project

flowchart of web scraping:

1. **Target URL - HTTP Request**
2. **Receive HTML Content**
3. **Parse HTML (Beautiful Soup)** -Identify Target Tables
4. **Extract Data** (Cell by Cell)
5. **Convert to Pandas DataFrame**

Data Wrangling

After acquiring and merging the multi-source data, a rigorous wrangling process was applied to clean, transform, and encode the features. This step is mandatory to ensure the data is in a mathematical format suitable for training the classification models.

flowchart of Data Wrangling:

1. **Merged Raw Data** (API + Wikipedia)
2. **Missing Data Handling** -Imputation / Removal
3. **Categorical Encoding** - One-Hot Encoding
4. **Numerical Scaling** - Standardization
5. **Final Split** - *X*(Features) and *Y*(Target)
6. **Model Training Ready**

EDA with Data Visualization

Summarize what charts were plotted:

1. **Categorical chart** : To visualize the density and distribution of successful and failed launches across continuous numerical variables (Payload) and the sequence of launches (Flight Number).
2. **Line chart** :To track temporal trends and identify changes in performance over time.
3. **Bar chart**: To compare the aggregated average performance (success rate) across distinct categorical groups (Launch Sites). A bar plot is ideal for comparing simple metrics across categories

GitHub URL : https://github.com/vatsun07-web/Data-Science-Capstone-Project

EDA with SQL

- select distinct Launch Site from SPACEXTABLE;
- select * from SPACEXTABLE where Launch Site like 'CCA%' limit 5;
- select sum(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
- select AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTABLE WHERE Booster Version = 'F9 v1.1';
- select min(Date) as Date from SPACEXTABLE where Landing Outcome = 'Success (ground pad)';
- select Booster Version, Payload from SPACEXTABLE where Landing Outcome = 'Success (drone ship)' and Payload_Mass__KG_ between 4000 and 6000;
- select Mission_Outcome, count(*) as Total_Number from SPACEXTABLE group by Mission_Outcome;
- select DISTINCT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

EDA with SQL

- select case substr(Date,6,2) 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 name, Landing Outcome, Booster Version, Launch Site from SPACEXTABLE where Landing Outcome = 'Failure (drone ship)' and substr(Date,1,4) ='2015';
- select Landing Outcome, count(*) AS TOTAL_N from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' Group by Landing Outcome order by TOTAL_N DESC;
- GitHub URL : https://github.com/vatsun07-web/Data-Science-Capstone-Project

Build an Interactive Map with Folium

Marker : Displays the Name of the Launch Site directly on the map

Circle : Illustrates the Launch Site boundaries or safety zones

MarkerCluster : Displays individual launch outcomes (Success: Green, Failure: Red) that aggregate into a single cluster icon when zoomed out

PolyLine (Line): Connects the launch site to the coastline, nearby streets, and railroads.

GitHub URL : https://github.com/vatsun07-web/Data-Science-Capstone-Project

Build a Dashboard with Plotly Dash

1.Success vs. Failure Pie Chart:

Performance Aggregation: 'ALL' Selected: Illustrates the total success count aggregated by each Launch Site, showing which sites contribute most to overall success.

Specific Site Selected: Swaps to show the direct Success vs. Failure percentage for only that site, enabling fast performance evaluation.

2. Payload Mass vs. Success Scatter Chart:

Correlation and Feature Analysis: Relationship: Reveals the correlation (or lack thereof) between the mass carried and the launch outcome.

Multivariate Analysis: Color-coding by Booster Version allows us to see if a specific booster type performs better or worse within a certain payload range, which is critical for model feature selection.

GitHub URL : https://github.com/vatsun07-web/Data-Science-Capstone-Project

Predictive Analysis (Classification)

Key phrases:

1. Baseline Models, Classifier Objects, Diverse Algorithms
2. GridSearchCV, Parameter Grid, 10-Fold Cross-Validation (cv=10)
3. Unseen Test Set, Accuracy Score, Generalization
4. Performance Comparison, Simplicity, Interpretability

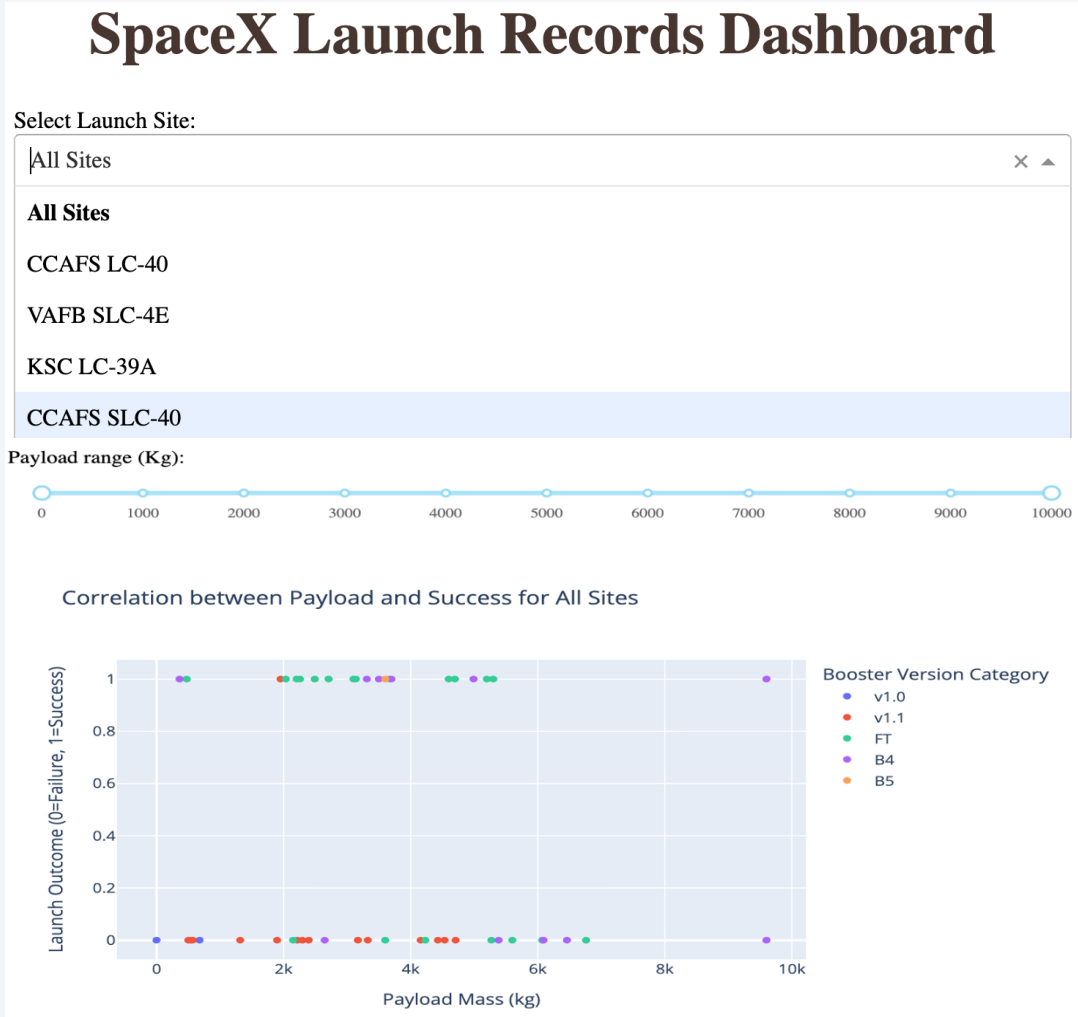
flowchart of web scraping:

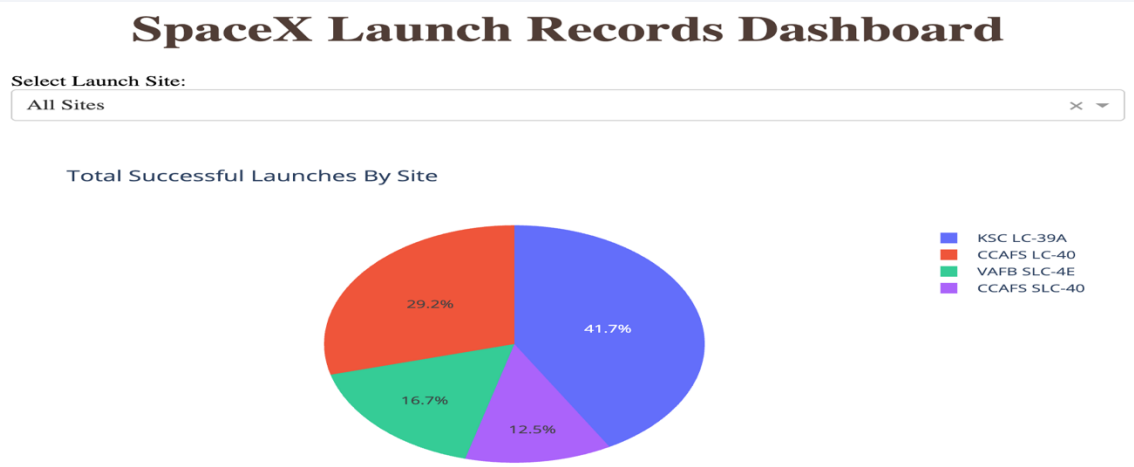
1. **Define Models** (LR, SVM, DT, KNN)
2. **Hyperparameter Tuning (GridSearchCV) -Best Parameters Found**
3. **Final Model Training** (Using Best Parameters)
4. **Evaluate on Test Set - Accuracy Score & Confusion Matrix**
5. **Compare All Models**
6. **Select Best Model** (Decision Tree)

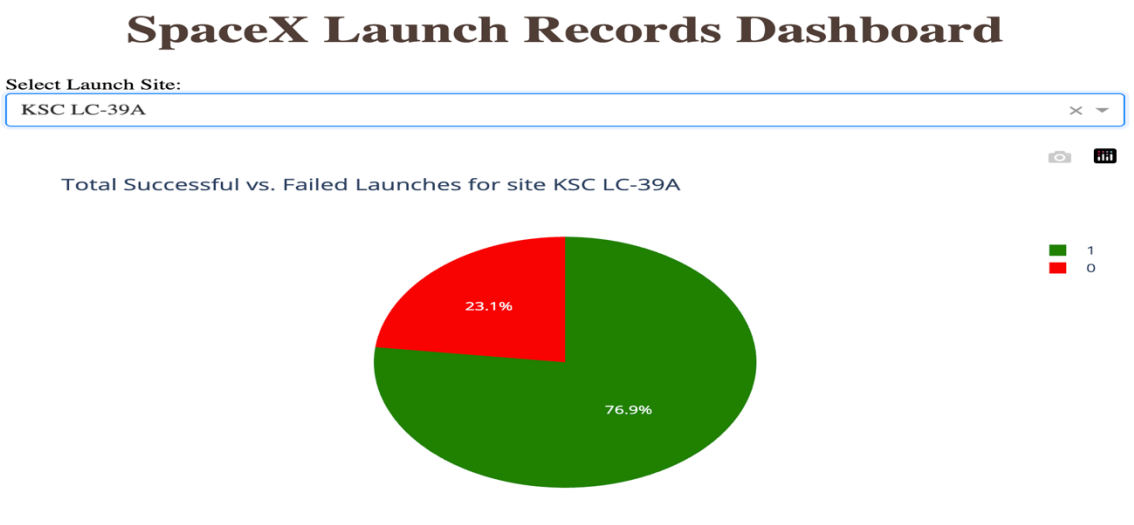
Results - Exploratory data analysis

Chart Type	Key Feature(s) Visualized	Primary Insight Derived
Line plot	Year vs. Success Rate	Confirmed a strong **upward trend in Success Rate over time**, validating the learning curve hypothesis.
Bar plot	Launch Site vs. Success Rate	Showed **significant variability in success rates by Launch Site**, proving location is a key predictor.
Category Plot	Flight Number vs. Payload Mass	Revealed that high failure rates were concentrated in **early flights**, with success stabilizing as payload mass increased up to a certain threshold.

Results- Interactive analytics demo in screenshots







Seiha Vat DS

Results – Predictive Analysis Results

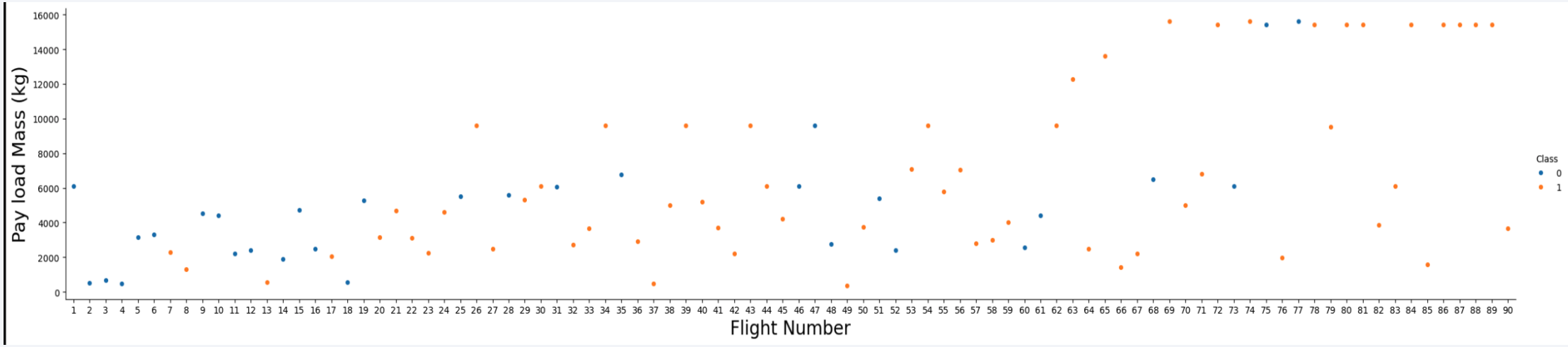
Model	Cross-Validation Score	Test Set Accuracy	Best Model Justification
Decision Tree	**0.877**	0.83	**Highest CV Score & Superior Interpretability**
SVM	0.848	0.83	Robust Non-linear Model
KNN	0.848	0.83	Simple Distance-Based Model
Logistic Regression	0.846	0.83	Simple, Highly Maintainable

Section 2

Insights drawn from EDA

Seiha Vat DS

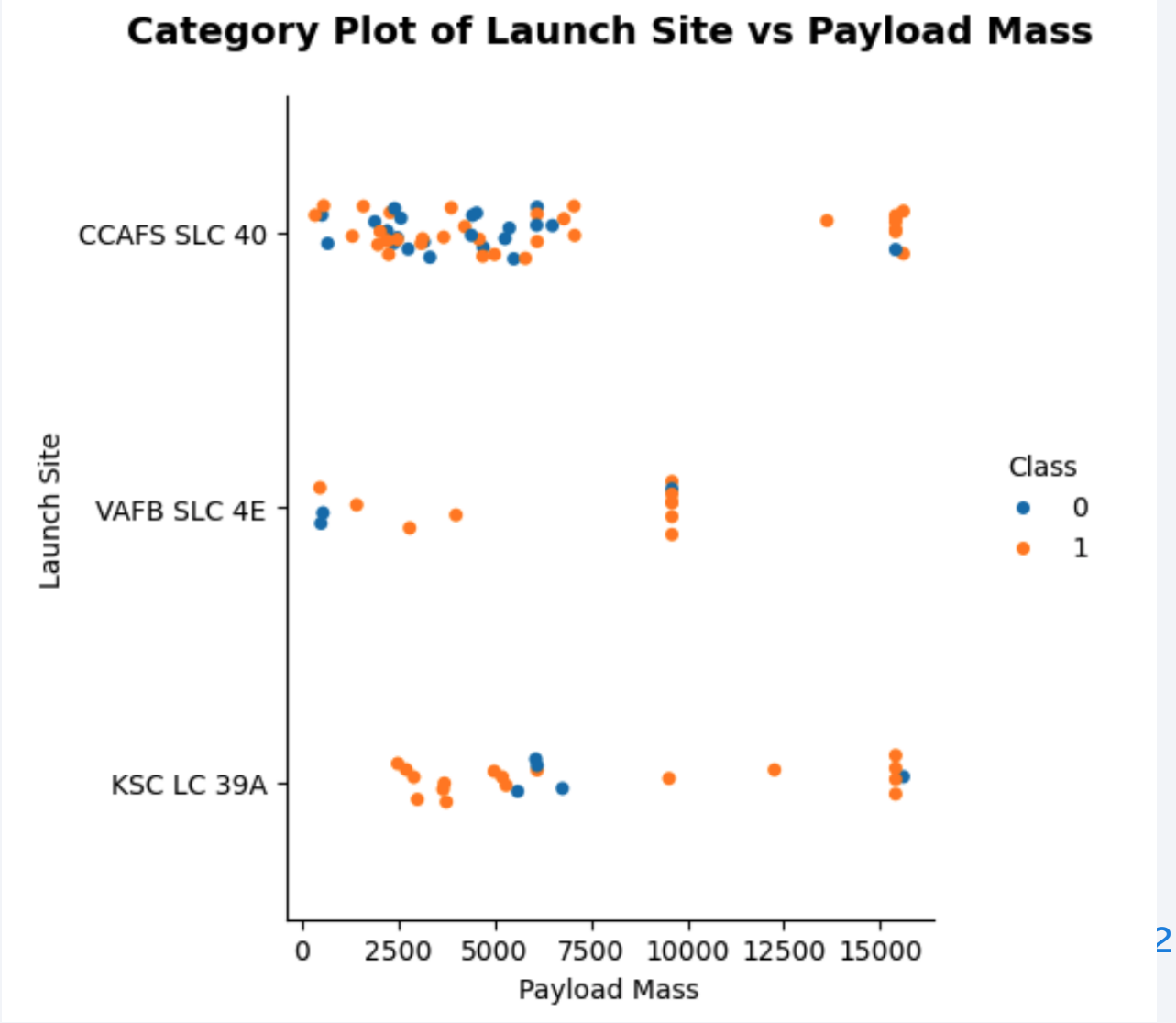
Flight Number vs. Launch Site



- We see that as the flight number increases; the first stage is more likely to land successfully. The payload mass also appears to be a factor; even with more massive payloads, the first stage often returns successfully.

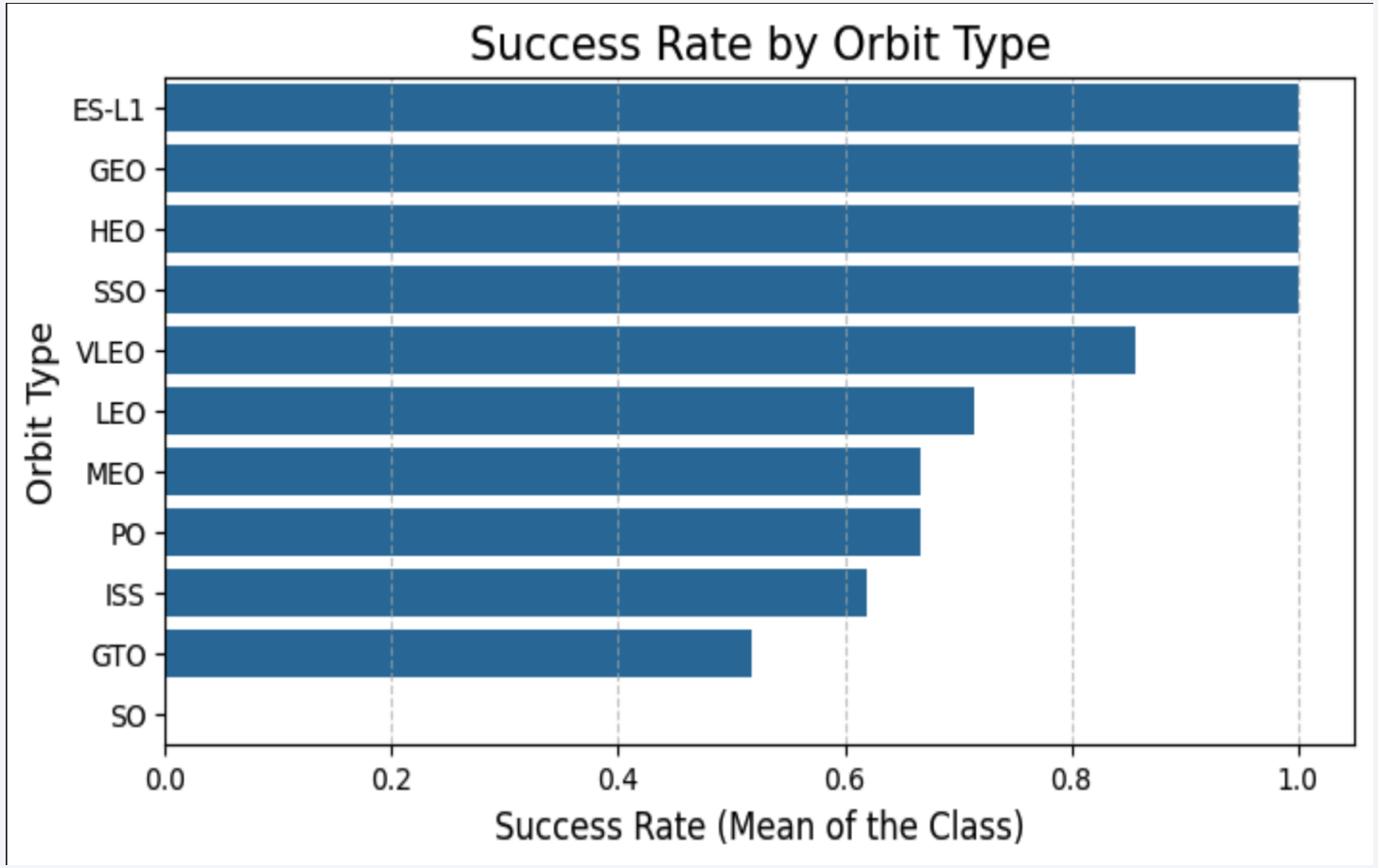
Payload vs. Launch Site

- The Payload Mass Vs. Launch Site scatter point chart showed for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

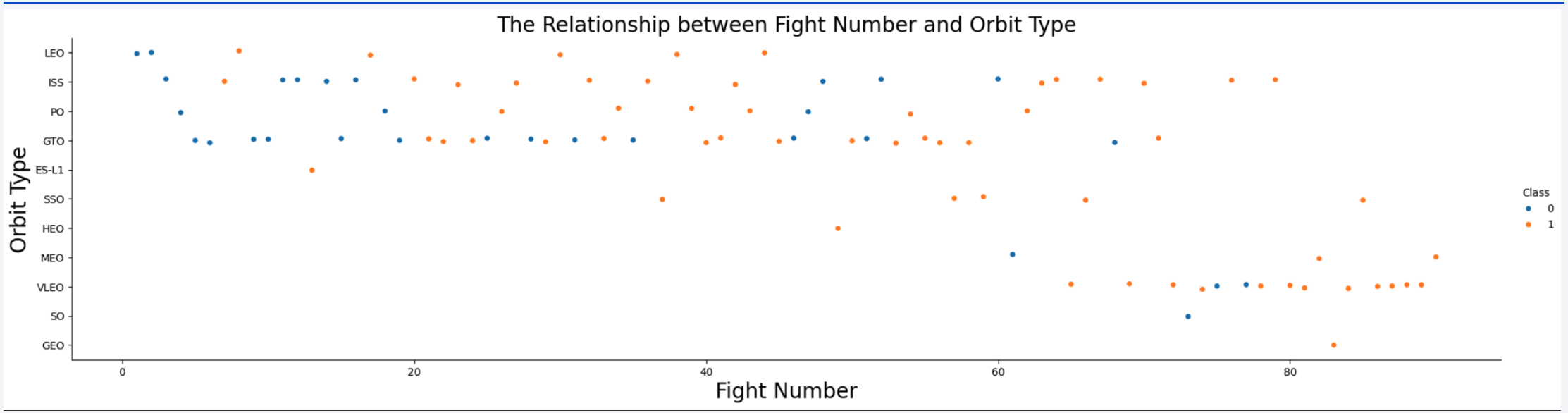


Success Rate vs. Orbit Type

- The successful rate are varied by each Orbit Type, we noticed the most successful rate Orbit are ES-L1, GEO, HEO, SSO

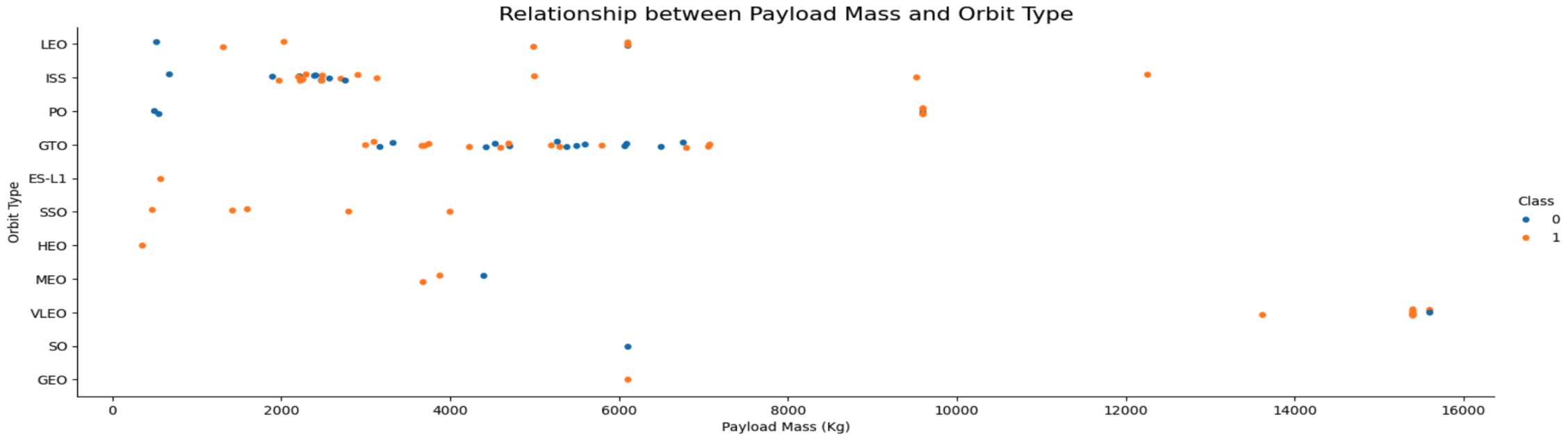


Flight Number vs. Orbit Type



- We observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

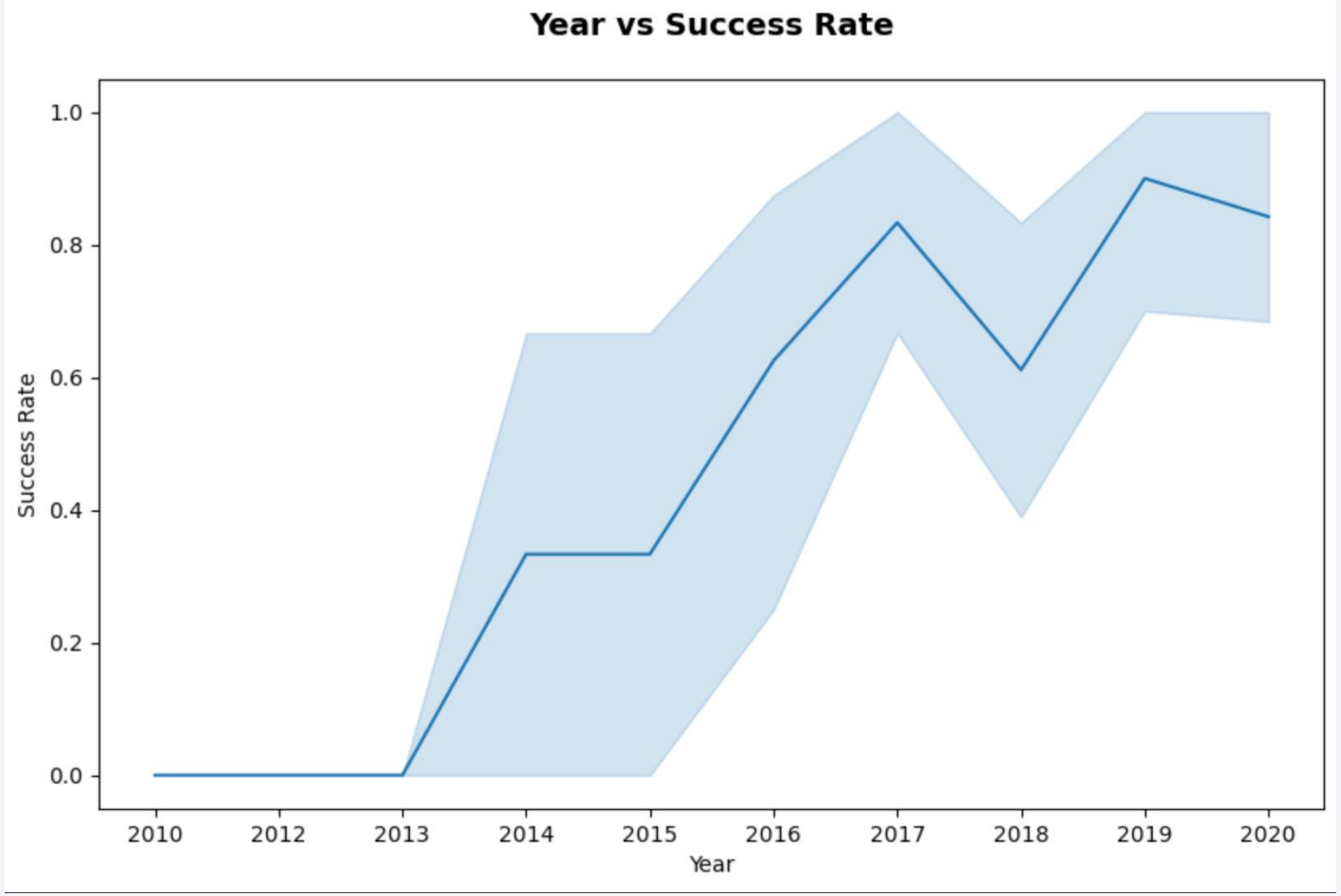
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

The query identified **four distinct, unique launch sites** used by SpaceX in the historical dataset:
CCAFS SLC-40
VAFB SLC-4E
KSC LC-39A
CCAFS ESC-40

- The launch location is a crucial **categorical feature** that heavily influences mission success due to varying factors like orbital inclination limits, range safety constraints, and local weather patterns (as confirmed by the earlier bar plot analysis).

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

The launch site is CCAFS LC-40, however the result of Landing Outcome are unpleasant with 2 Failure (parachute) and 3 No attempt.

The "Failure" designation confirms that the system did not function as intended, resulting in the loss of the booster stage upon ocean impact.

Total Payload Mass

sum(PAYLOAD_MASS__KG_)
45596

- Total Payload Mass (Kg) by boosters from NASA: **45,596 kg**

- **Result Interpretation:** The figure of **45,596 kg** represents a significant portion of the total mass launched. Because NASA missions often involve complex, high-value satellites or resupply capsules (like Dragon missions to the ISS), this metric provides context on the **quality and complexity** of the missions flown for this specific government partner.

Average Payload Mass by F9 v1.1

AVG_PAYLOAD
2928.4

- Average Payload Mass (Kg) by F9 v1.1: **2,928.40 kg**

- The figure 2,928.40 kg represents the **typical payload capacity profile** for the F9 v1.1 generation of the booster. It establishes the baseline expectation for missions assigned to this hardware version.

First Successful Ground Landing Date

Date
2015-12-22

- This result confirms a major milestone: the mission on **December 22, 2015**, marked SpaceX's **first successful recovery of an orbital-class booster stage on solid ground.**
- This accomplishment immediately shifted the paradigm from mere ocean splashdown tests and provided the crucial proof-of-concept for the company's entire reusability vision.

Successful Drone Ship Landing with Payload between 4000 and 6000

Key Insight:

Medium payload missions (4,000–6,000 kg) consistently achieve successful drone ship landings.

Indicates strong booster reliability for mid-range payloads, supporting SpaceX's reusability strategy.

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

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

- The dataset is extremely imbalanced, with a **100:1 ratio** of successful missions to failed missions. The failure class (the minority class) is the event are often most interested in predicting.

Boosters Carried Maximum Payload

- The resulting booster **F9 B5** represents the maximum payload capacity achieved by SpaceX for its missions. This reflects the culmination of years of engine and structural upgrades from the early versions.
- These maximum-payload missions are typically assigned to the most demanding commercial contracts

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

- The multiple failures highlight the extreme **engineering difficulty** involved in slowing down the rocket for a soft landing.
- The repeated use of **CCAFS SLC-40** shows that this site was the primary location for these complex recovery tests.

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

Landing_Outcome	TOTAL_N
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 data shows a balance between operational needs (expending the booster 10 times) and intense R&D efforts (9 definite failures vs. 8 successful recoveries), confirming the steep engineering learning curve required to achieve reusability.

Section 3

Launch Sites Proximities Analysis

Seiha Vat DS

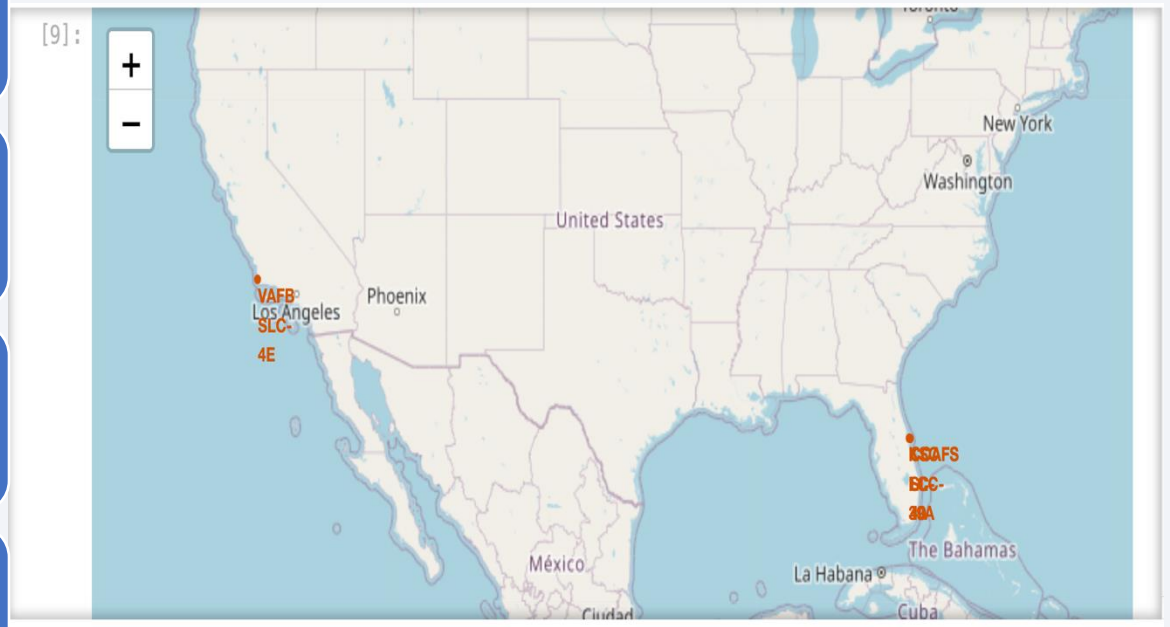
Launch Site Analysis

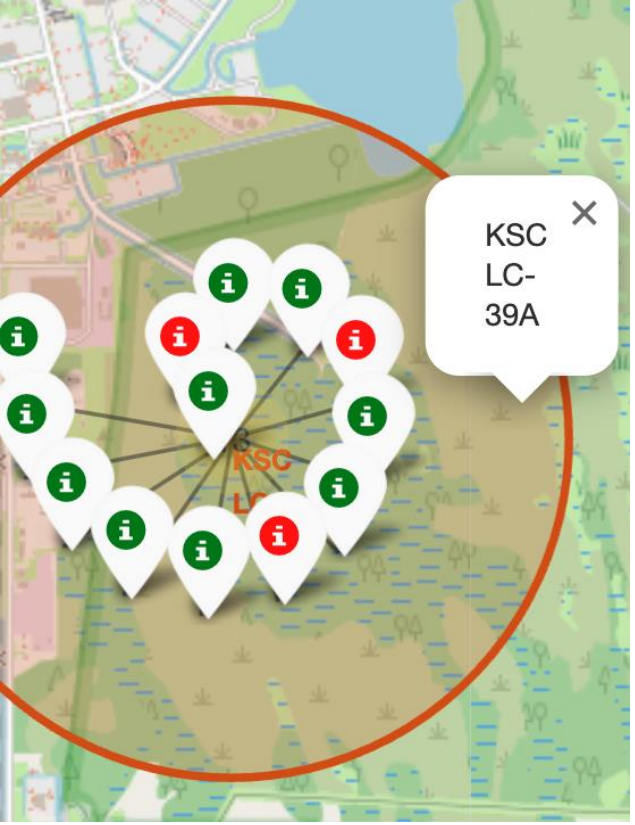
All SpaceX launch sites are **near the coast and equator** for:

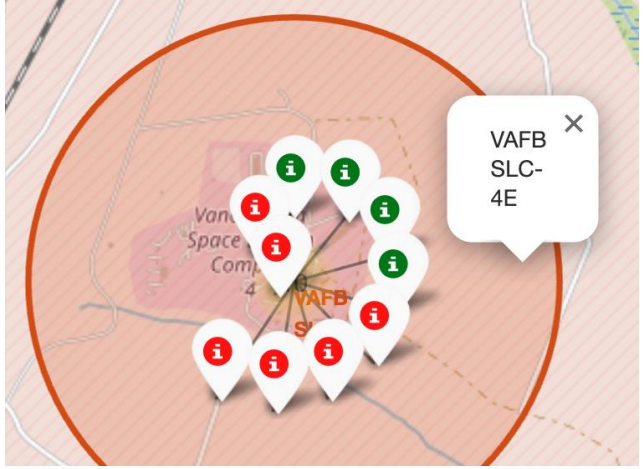
Performance: Earth's rotation gives extra boost for east-bound launches.

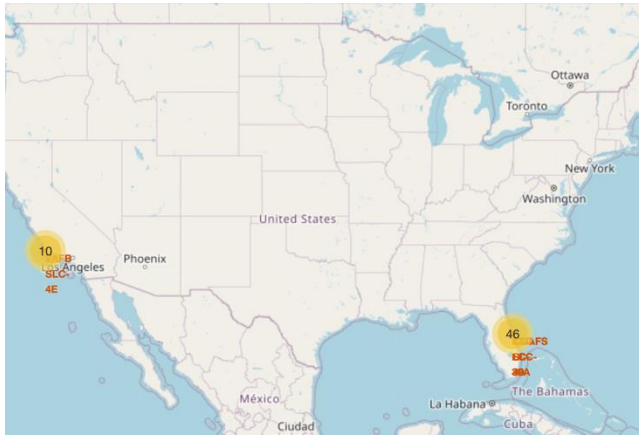
Safety: Ocean trajectories minimize risk to populated areas.

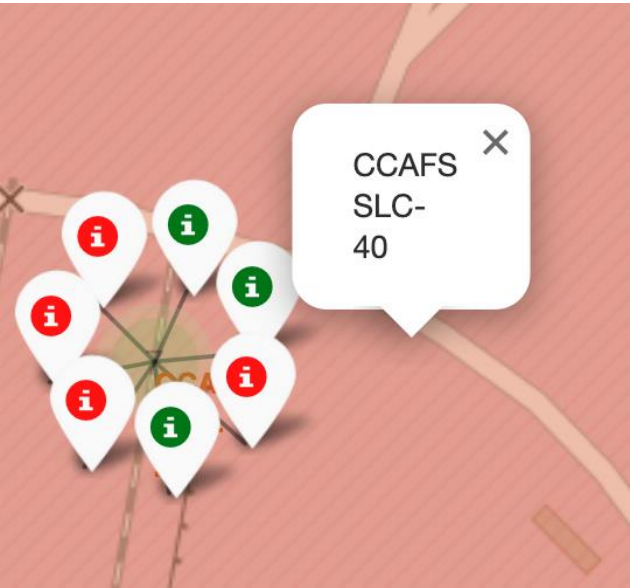
Recovery: Easier booster retrieval using drone ships.

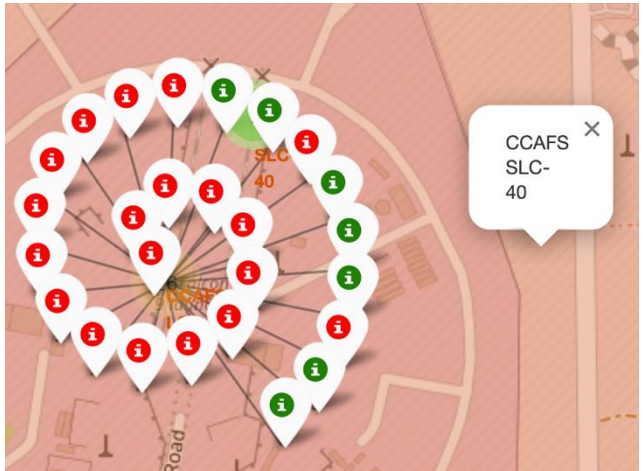








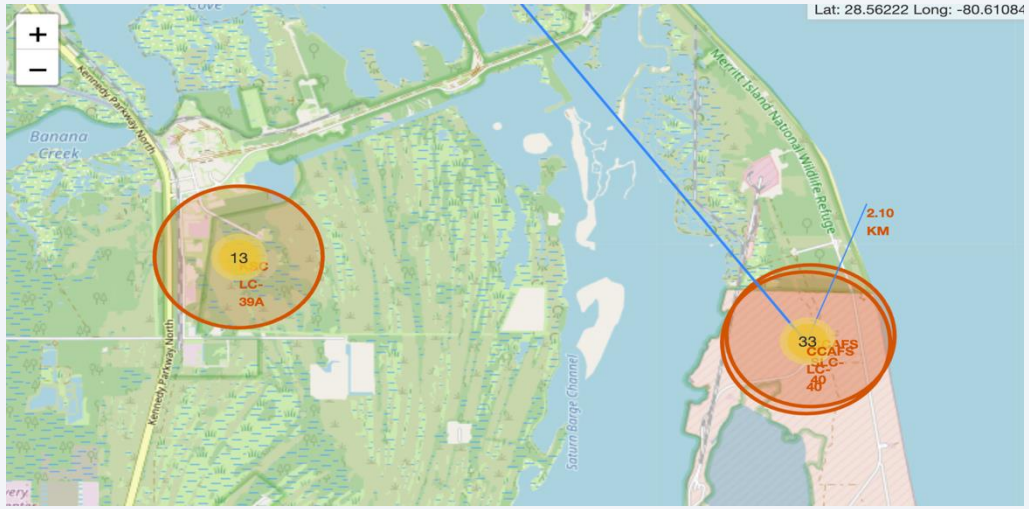




Launch Site Success Analysis

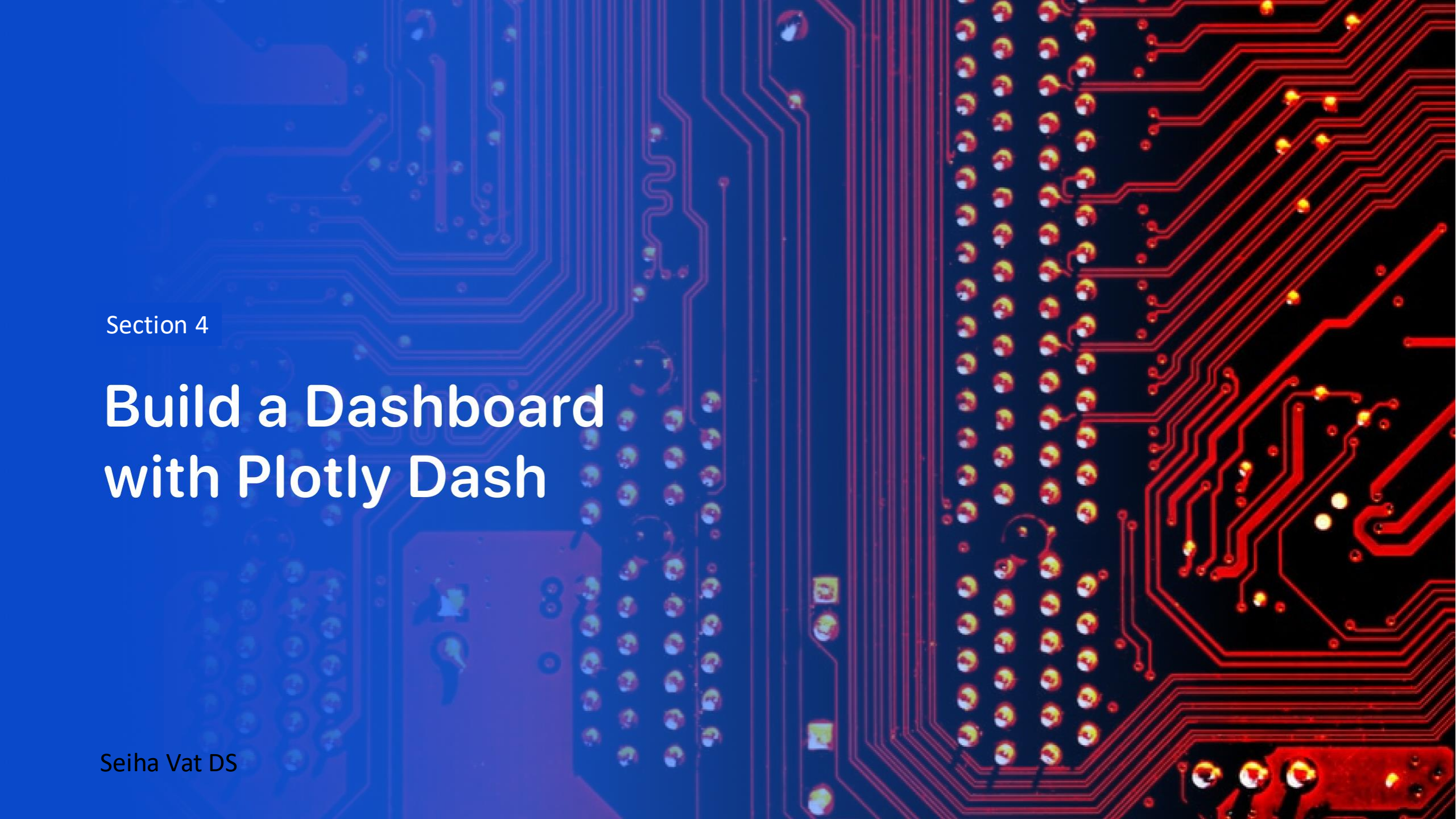
- **KSC LC-39AL :** Shows a very high proportion of **Green markers** with few to no Red markers. It was the highest success rate compared to other 3 launch site.

Why Launch Sites Are Coastal



- **All launch sites are in very close proximity to the coast.** This is a non-negotiable requirement driven by **safety regulations**, ensuring that launch trajectories fly over the ocean so that any debris or potential rocket malfunctions occur over unpopulated areas. It also facilitates easier **booster recovery** onto drone ships.

Seiha Vat DS

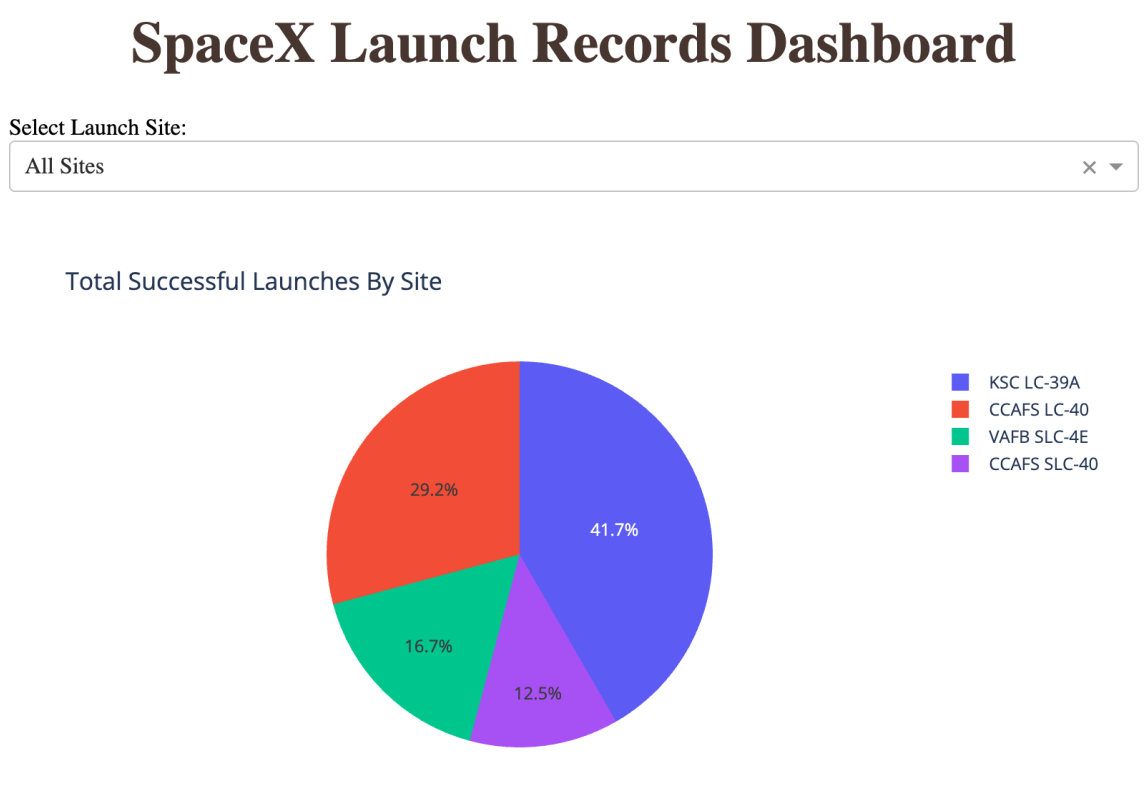


Section 4

Build a Dashboard with Plotly Dash

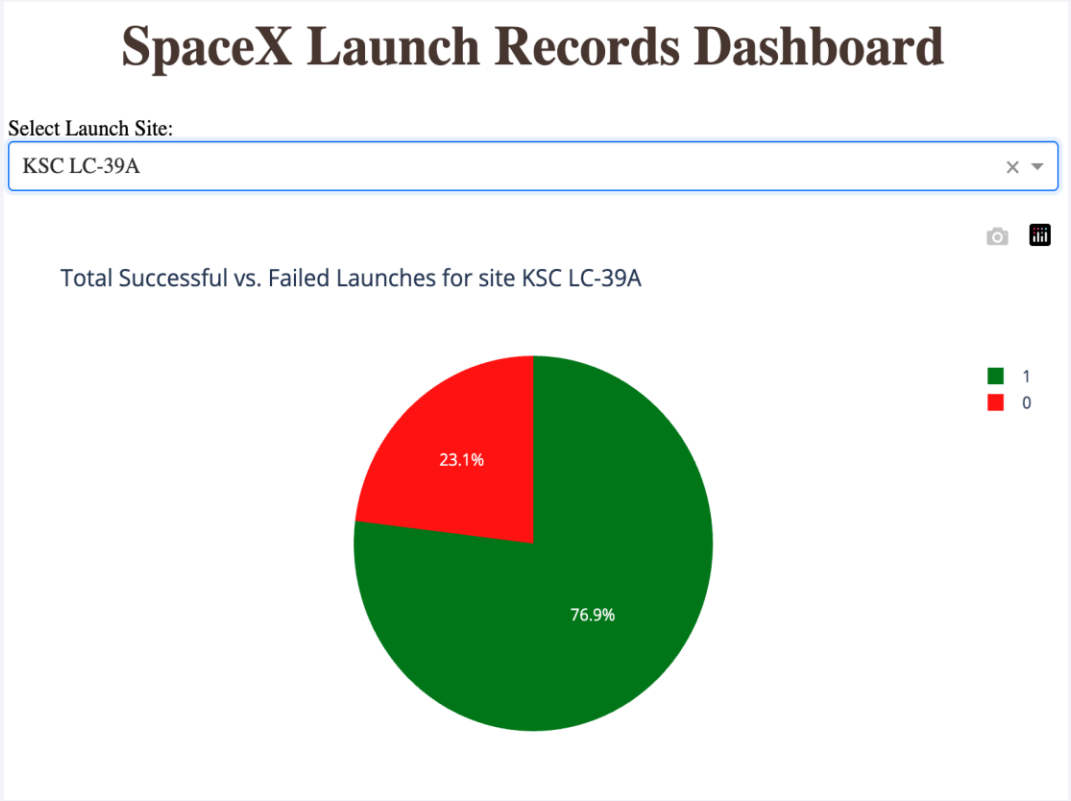
Seiha Vat DS

Launch Site Success Rate Comparation



- **KSC LC-39A:** This launch site recorded the highest success rate, accounting for **41.7%** of all successful missions among the four sites.
- **CCAFS LC-40:** Also demonstrated strong performance, contributing **29.2%** to the overall success rate.

Top Performance Launch Site



- The visual evidence from the KSC LC-39A pie chart confirms that the high success rate of **76.9%.**
- As a key finding, the site's performance validates that the Launch Site feature is a strong predictor of success, demonstrate a significantly lower rate of failure than missions from other launch site

Optimal Payload and Booster

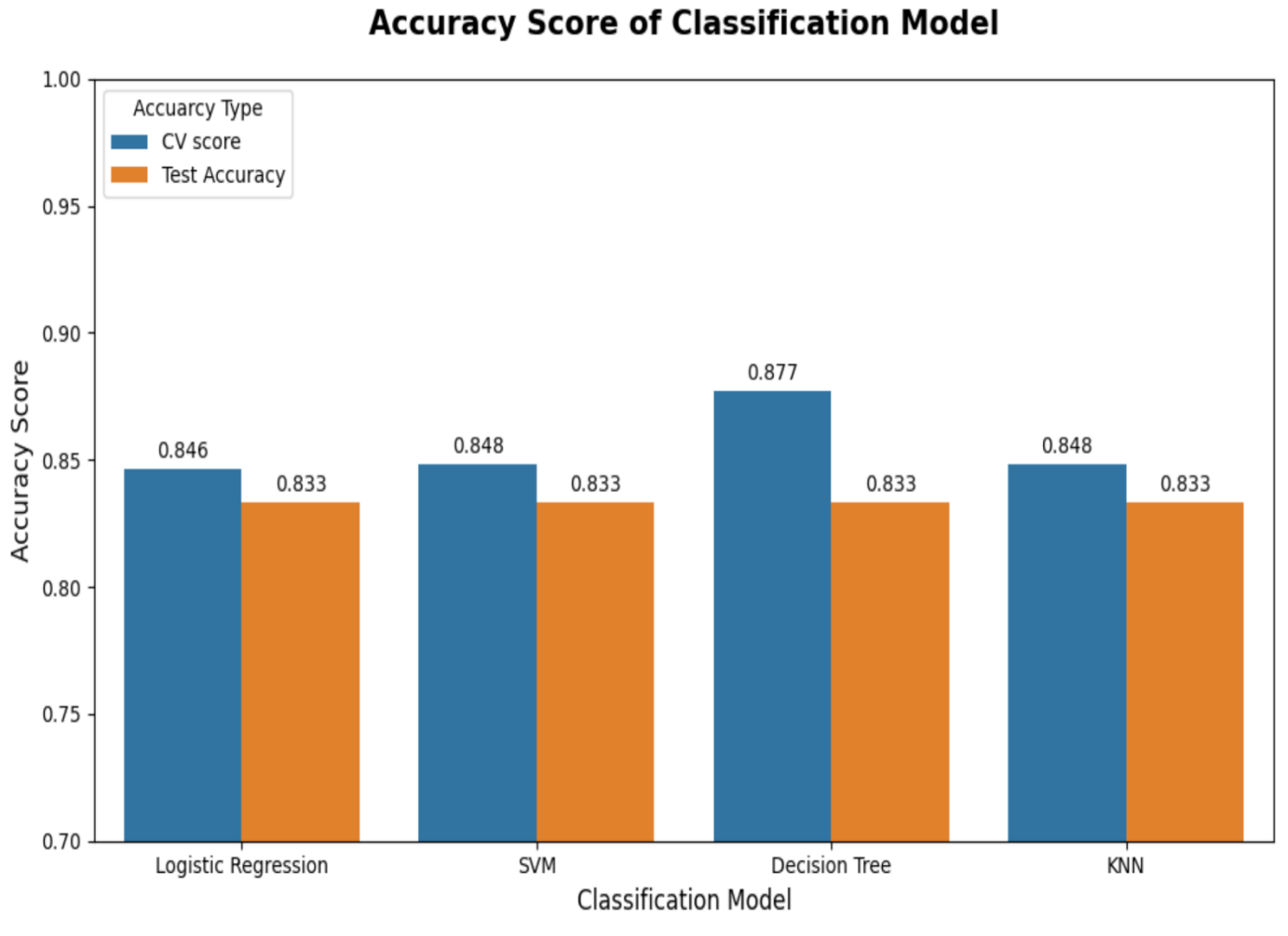
- The screenshot visually validates that the 2,000 kg - 6,000 kg payload mass, when combined with the most reliable booster versions FT, B4, B5), represents the **lowest risk profile** for a SpaceX launch.

Section 5

Predictive Analysis (Classification)

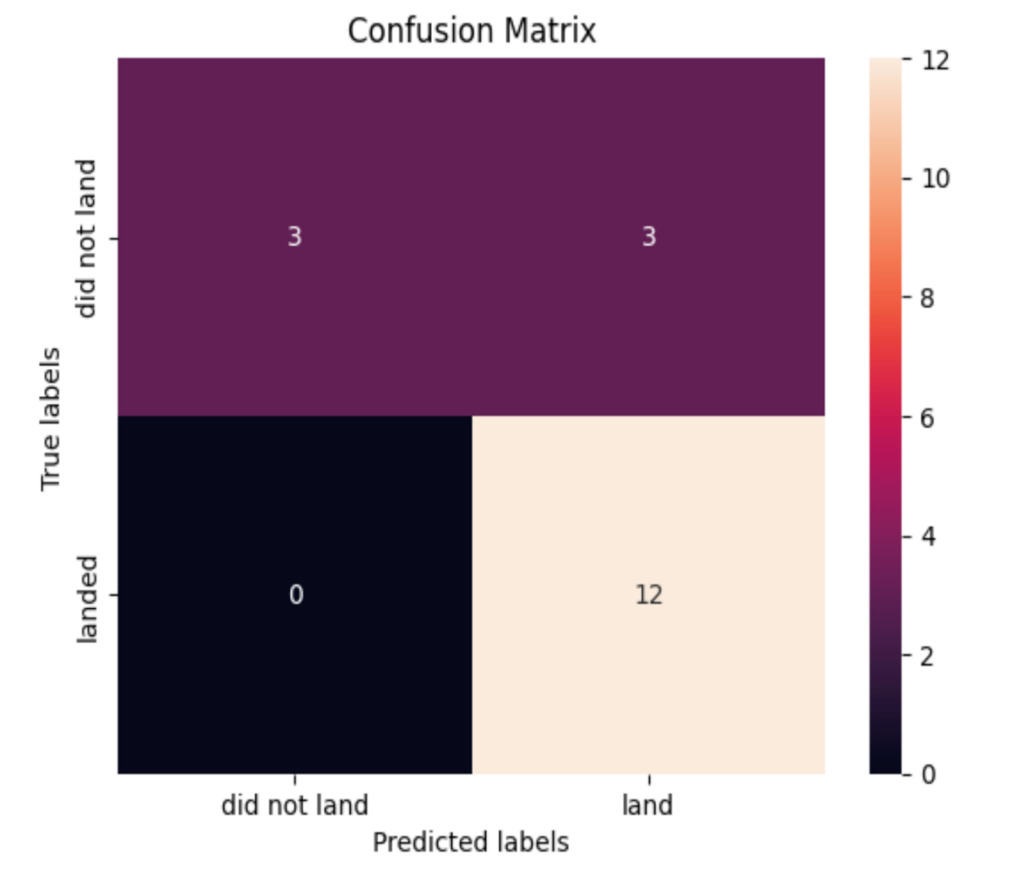
Seiha Vat DS

Classification Accuracy



- All models performed the incredibly well, however, the **Decision Tree** model received the highest score with highest CV score 0.887 and Test Accuracy score 0.83.

Confusion Matrix



Failure Prediction (Risk Area)

•3 failures misclassified as successes (False Positives) → High-risk for space launch planning

•12 successes correctly predicted

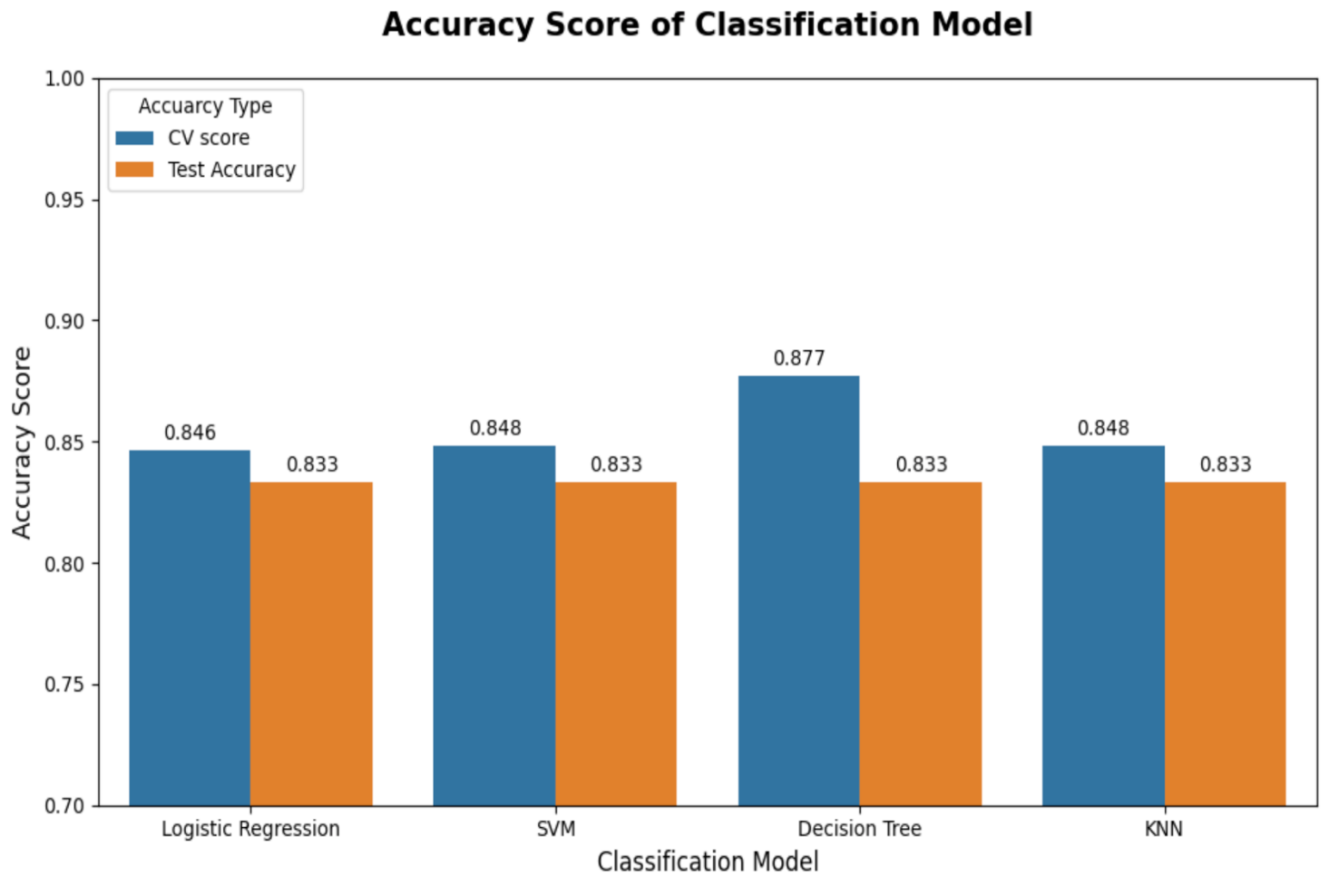
Precision Vulnerability

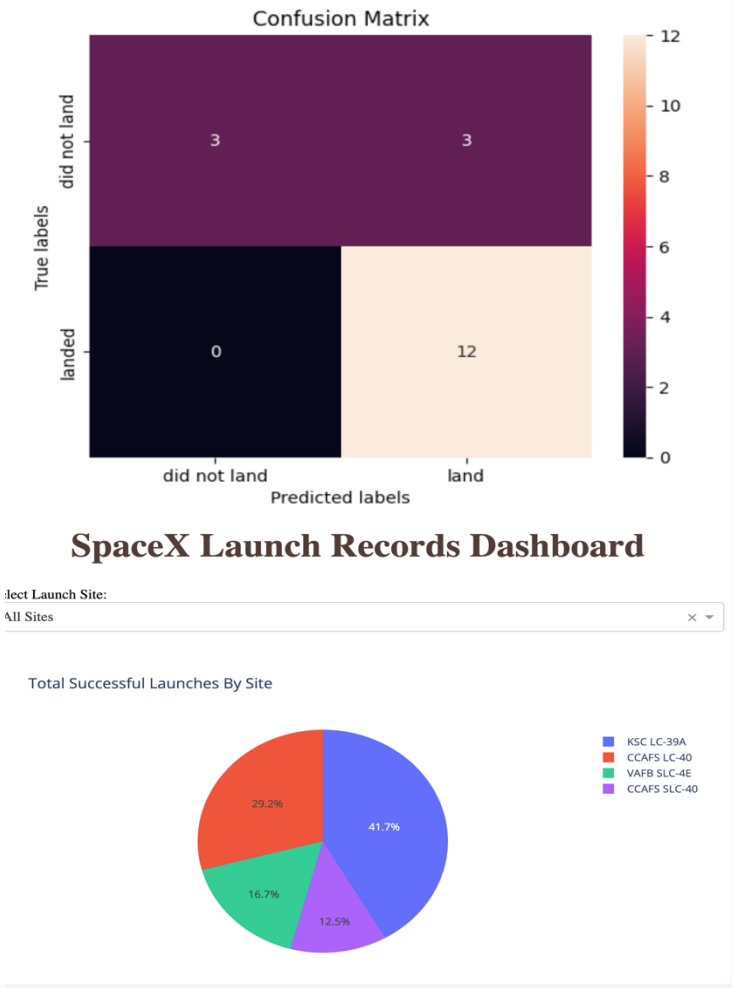
•When predicting "Success," accuracy is **80%** **(Precision = 0.80)**

Conclusions

- **Success Drivers**
  - Mission success depends more on **engineering evolution** than initial launch conditions.
  - **Technological Maturity is Key**: Later Booster Versions (FT, B4, B5) + optimized site **KSC LC-39A → 76.9% success**.
  - Rapid iteration and learning from failures improve reliability.
- **Optimized Payload Window**
  - Best performance: **2,000–6,000 kg** range → minimizes risk and structural stress.
- **Model Performance & Risk**
  - **Decision Tree Classifier**: CV score **0.889** (strongest model).
  - Risk: **3 False Positives** (predicted success for actual failure) → requires threshold tuning for safer mission planning.

Appendix


Accuracy Score of Classification Model
Accuarcy Type
CV score
Test Accuracy
Accuracy Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.846
0.833
0.848
0.833
0.877
0.833
0.848
0.833
Logistic Regression
SVM
Decision Tree
KNN
Classification Model


Confusion Matrix
True labels
did not land
landed
3
3
0
12
did not land
land
Predicted labels
12
10
8
6
4
2
0
SpaceX Launch Records Dashboard
lect Launch Site:
All Sites
Total Successful Launches By Site
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
29.2%
41.7%
16.7%
12.5%

Thank you!