IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noppon
29 Nov 2025

Outline

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

Executive Summary

Summary of Methodologies

1. **Data Acquisition and Preparation:** Launch records were collected via web scraping and processed using Data Wrangling techniques to clean, transform, and label key features such as booster version, launch site, payload mass, and orbit type.
2. **Exploratory Data Analysis (EDA):** Visualizations were created to analyze trends in mission outcomes over time and across key variables, providing initial insights into success drivers.
3. **Classification Modeling:** Four Machine Learning algorithms—**Logistic Regression**, **SVM**, **Decision Tree**, and **KNN**—were trained and tested to predict the binary launch outcome (Success/Failure).

Executive Summary

Summary of Key Results

1. **Booster Version Dominance:** The most critical determinant of success is the **Booster Version Category**. Newer booster versions (**FT, B4, B5**) exhibit near-perfect success rates, effectively eliminating the risk associated with payload mass or launch site. Failures are primarily historical, linked to older **v1.0** and **v1.1** boosters.
2. **Operational Maturity:** The Yearly Launch Success Trend shows a clear transition from a developmental phase (0% success from 2010–2013) to a high-reliability operational phase, stabilizing at **over 80%** success by 2019–2020.
3. **Launch Site Performance: KSC LC-39A** accounts for the largest share of successful launches (41.7%), reflecting its use as the high-capacity, mature launch platform. Conversely, the high failure rate at **CCAFS SLC-40** (57.1% Failures) reflects its use during the program's initial, higher-risk learning phase.
4. **High Prediction Accuracy:** All four classification models achieved an identical test accuracy of **0.83**, confirming the high feasibility of accurately predicting first-stage reusability based on mission parameters.

Introduction

Project background and context

This project analyzes SpaceX Falcon 9 and Falcon Heavy launch data to understand factors influencing launch success and cost efficiency. The dataset includes detailed records of each mission, such as launch site, payload, booster version, and landing outcome. The goal is to leverage data science techniques—web scraping, SQL, EDA, and visualization—to extract insights that can inform business decisions and competitive strategies in the commercial space industry.

Problems you want to find answers

1. What factors most influence the success or failure of SpaceX launches?
2. How do launch site, payload mass, and orbit type affect outcomes?
3. What patterns exist in mission outcomes over time?
4. How can data-driven insights help predict future launch success and optimize costs?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data for the SpaceX Falcon 9 launches was collected from publicly available sources
- Perform data wrangling
  - The data was cleaned, transformed, and labeled to ensure consistency and usability for analysis. Key features were engineered and the processed dataset was saved for further visualization and modeling.
- 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 Collection

Data Collection Process for SpaceX Launch Dataset

Key Phrases:

- Gathered from public sources: SpaceX API, Wikipedia, IBM Skills Network
- Automated data download (CSV, API, web scraping)
- Data validation and enrichment (add coordinates, clean records)

Flowchart:

Data Collection – SpaceX API

1. **Identify the SpaceX API Endpoint**
   Find the official SpaceX API URL, such as https://api.spacexdata.com/v4/launches.

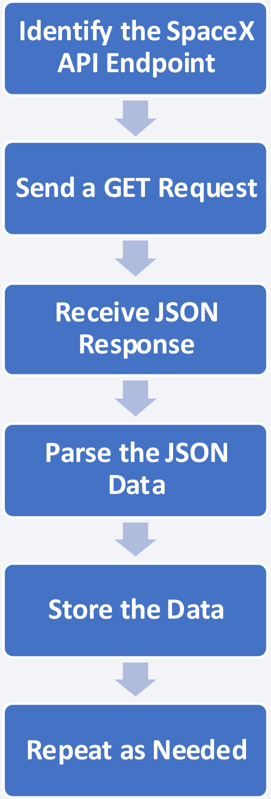
2. **Send a GET Request**
   Use a tool or library (e.g., Python's requests) to send an HTTP GET request to the endpoint.

3. **Receive JSON Response**
   The API returns data in JSON format containing launch details.

4. **Parse the JSON Data**
   Extract relevant fields from the JSON response, such as launch date, site, payload, and outcome.

5. **Store the Data**
   Save the parsed data into a structured format (e.g., Pandas DataFrame or CSV file) for further analysis.

6. **Repeat as Needed**
   If required, make additional API calls to other endpoints (e.g., rockets, payloads) to enrich your dataset.



[Link: Space X API Lab Link]

Data Collection - Scraping

- **Identify target Wikipedia page:**
  Find the URL with Falcon 9 launch records.

- **Send HTTP GET request:**
  Use requests to download the page's HTML.

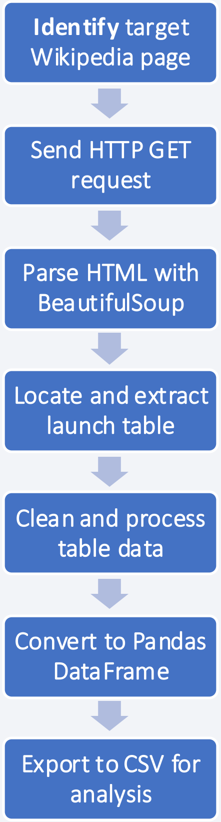
- **Parse HTML with BeautifulSoup**:
  Create a BeautifulSoup object to read the HTML.

- **Locate and extract launch table:**
  Find and select the table with launch data.

- **Clean and process table data:**
  Remove noise and format the data.

- **Convert to Pandas DataFrame:**
  Organize the data for analysis.

- **Export to CSV for analysis:**
  Save the data for future use.



[Link: Web Scraping Lab]()

Data Wrangling

How Data Were Processed (Data Wrangling):

- Cleaned raw scraped data by removing noise and handling missing values.
- Standardized column names and formats.
- Converted relevant columns to appropriate data types.
- Organized data into a Pandas DataFrame for analysis.
- Exported the cleaned data to CSV for further use.

Key Phrases:

- Remove noise and missing values
- Standardize columns and formats
- Convert data types
- Organize into DataFrame
- Export to CSV

Link: Data Wranging Lab

EDA with Data Visualization

Summary of the charts plotted and their purposes:

- **Scatter plots (catplot):**
  Plotted Flight Number vs. Payload Mass, Launch Site, and Orbit to visualize relationships and trends in launch success.

- **Bar chart:**
  Showed success rate by orbit type to compare performance across different orbits.

- **Line chart:**
  Displayed yearly launch success trends to observe improvements over time.

These charts were used to identify patterns, correlations, and trends in the SpaceX launch data, helping to understand factors affecting launch outcomes. You can check chart as below link.

Link: EDA with Data Visualization

EDA with SQL

Summary of the SQL queries

- Selected unique launch site names from the mission table.
- Displayed 5 records where launch sites begin with 'CCA'.
- Calculated total payload mass for boosters launched by NASA (CRS).
- Computed average payload mass for booster version F9 v1.1.
- Found the date of the first successful ground pad landing.
- Listed boosters with successful drone ship landings and payload mass between 4000 and 6000 kg.
- Counted total successful and failed mission outcomes.
- Identified booster versions that carried the maximum payload mass.
- Listed records with month, failed drone ship landings, booster version, and launch site for 2015.
- Ranked landing outcomes by count between specific dates.

[Link: EDA with SQL]

Build an Interactive Map with Folium

Map Objects Added:

- Circle: Placed at each launch site’s coordinates to visually highlight the location.
- Marker (with DivIcon): Added at each launch site to display the site’s name as a label.
- MarkerCluster: Used to group individual launch markers (success/failure) for better visualization when many launches occur at the same site.
- Colored Markers: Green for successful launches, red for failed launches, showing outcomes at each site.
- MousePosition Plugin: Allows users to see coordinates by hovering over the map, useful for identifying proximity points.
- Distance Markers (with DivIcon): Placed at points of interest (coastline, city, railway, highway) to show calculated distances from the launch site.
- Polylines: Drawn between launch sites and proximity points (coastline, city, railway, highway) to visually represent and measure distances.

Link: Build an Interactive Map with Folium

Build an Interactive Map with Folium

Reasons for Adding These Objects:

- Circles and Markers: To clearly identify and label each launch site on the map, making locations easy to spot and interpret.
- MarkerCluster and Colored Markers: To visualize launch outcomes (success/failure) at each site, helping users quickly assess performance and patterns.
- MousePosition: To interactively find coordinates of nearby features, supporting further analysis.
- Distance Markers and Polylines: To measure and display distances between launch sites and important proximities (coastline, city, railway, highway), enabling analysis of geographical factors that may influence launch success.

Link: Build an Interactive Map with Folium

Build a Dashboard with Plotly Dash

Summary of Dashboard Plots/Graphs and Interactions:

- **Dropdown for Launch Site Selection:**
    - Allows users to select either "All Sites" or a specific launch site.
    - Interacts with both the pie chart and scatter plot to filter data accordingly.
- **Pie Chart (Success Counts):**
    - Shows total successful launches by site when "All Sites" is selected.
    - Shows success vs. failure launches for a selected site.
    - Updates dynamically based on dropdown selection.
- **Payload Range Slider:**
    - Lets users filter launches by payload mass range.
    - Interacts with the scatter plot to show only launcheswithin the selected payload range.
- **Scatter Plot (Payload vs. Success):**
    - Displays the correlation between payload mass and launch success.
    - Color-coded by Booster Version Category.
    - Updates based on both site selection and payload range.

Link: SpaceX Dash

Build a Dashboard with Plotly Dash

Explanation for Adding These Plots and Interactions:

- **Dropdown & Pie Chart:**
  These provide a quick overview of launch success rates across all sites or for a specific site, helping users identify which sites have higher success rates.

- **Payload Slider & Scatter Plot:**
  These allow users to explore how payload mass affects launch success, and whether certain booster versions perform better with heavier or lighter payloads. The interactive filtering helps users discover patterns and correlations in the data.

- **Overall:**
  The combination of interactive controls and visualizations enables users to perform exploratory data analysis directly in the dashboard, making it easy to answer questions about launch outcomes, payload effects, and site performance.

Link: SpaceX Dash

Predictive Analysis (Classification)

Model Development Process Summary

1. **Data Preparation:**
   Loaded SpaceX launch data, selected relevant features, and standardized the feature matrix using StandardScaler.
2. **Train/Test Split:**
   Split the data into training and test sets (80/20 split).
3. **Model Building:**
   Built four classification models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
4. **Hyperparameter Tuning:**
   Used GridSearchCV for each model to find the best hyperparameters with cross-validation (cv=10).
5. **Model Evaluation:**
   Evaluated each model using test accuracy and confusion matrix.
6. **Best Model Selection:**
   Compared test accuracies of all models and selected the best performing one.

Link: Classification Prediction

Predictive Analysis (Classification)

Key phrases:

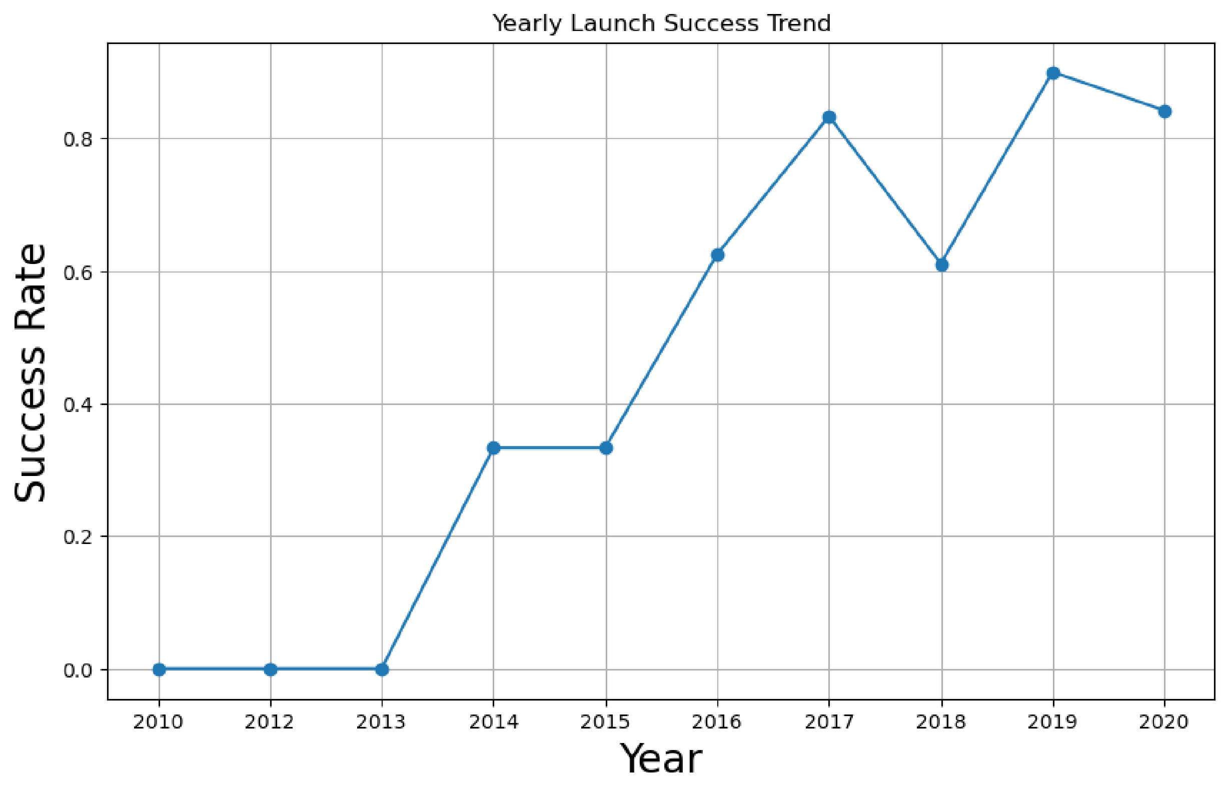
1. Data loading
2. Feature selection
3. Standardization
4. Train/test split
5. Model selection (Logistic Regression, SVM, Decision Tree, KNN)
6. Hyperparameter tuning (GridSearchCV)
7. Model evaluation (accuracy, confusion matrix)
8. Best model selection

Link: Classification Prediction

Results

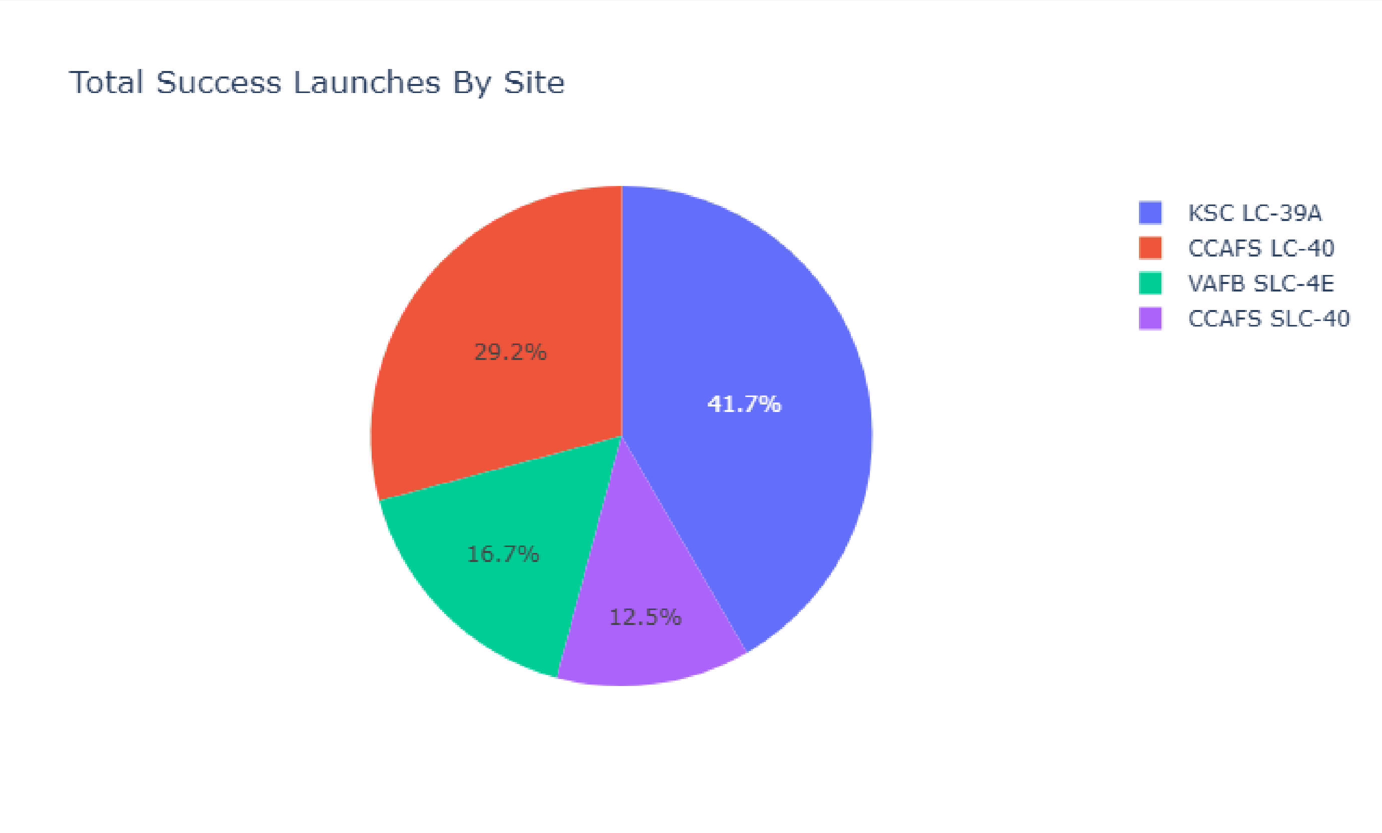
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Results



the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

Results



Total success launches of CCAFS SLC-40 is highest rate success

Results

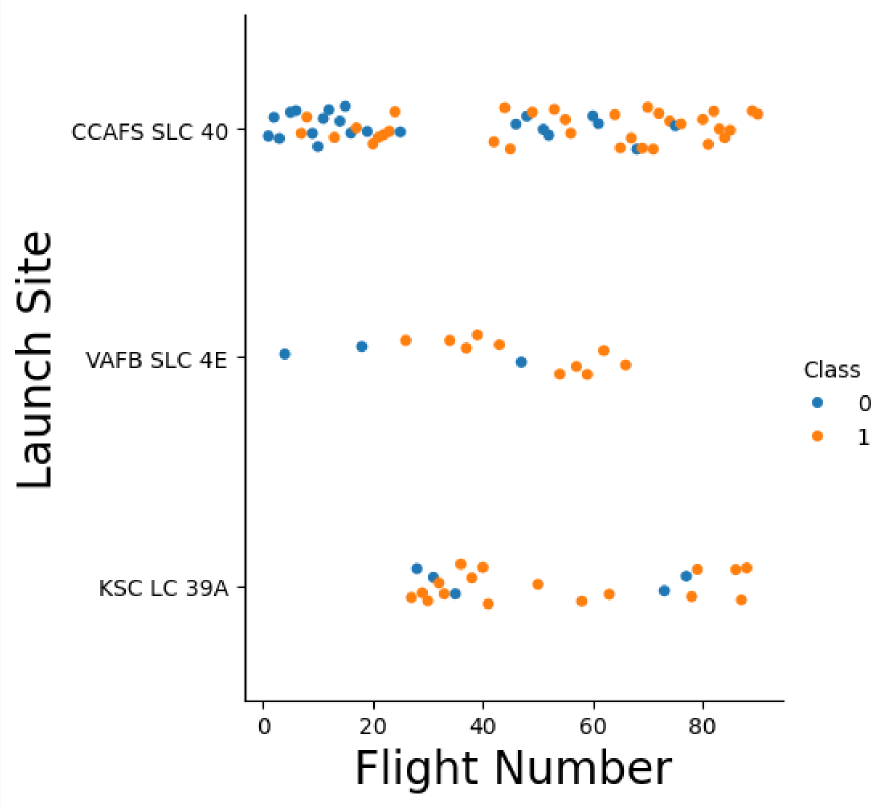
Predictive analysis results

1. Logistic Regression Test Accuracy: 0.8333333333333334
2. SVM Test Accuracy: 0.8333333333333334
3. Decision Tree Test Accuracy: 0.8333333333333334
4. KNN Test Accuracy: 0.8333333333333334

Section 2

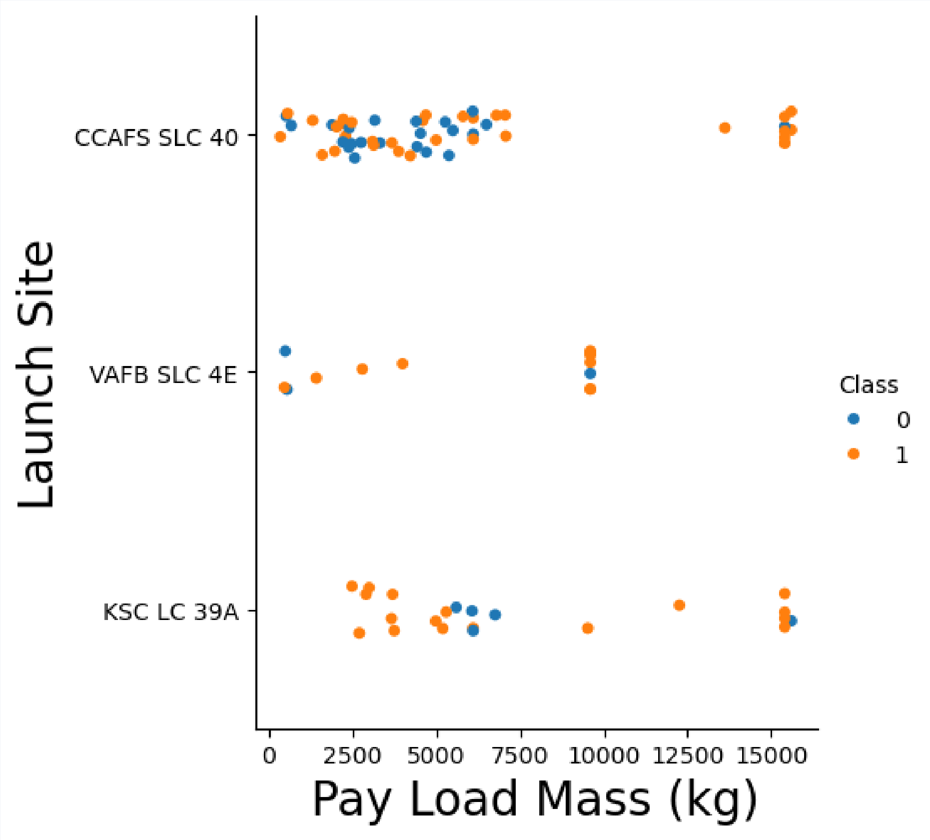
Insights drawn from EDA

Flight Number vs. Launch Site



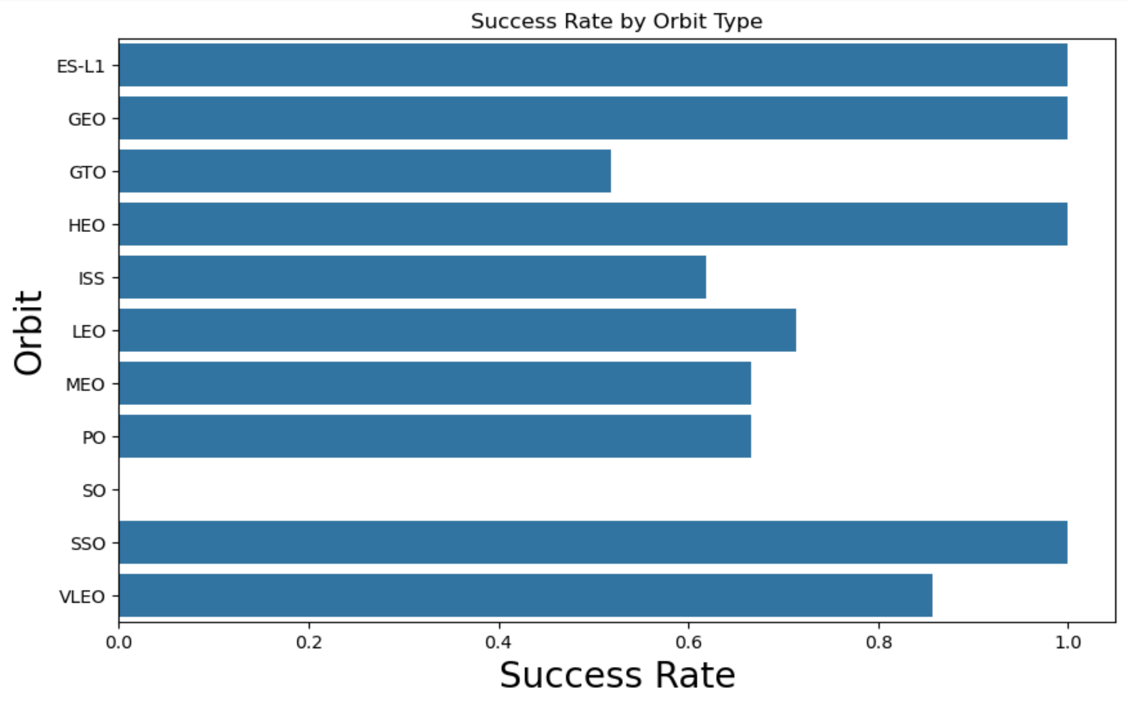
1. CCAFS SLC 40 (The Workhorse):
   - This is the **most frequently used launch site**, spanning the entire range of flight numbers.
   - It clearly demonstrates the **maturity curve** of the launch vehicle: early flights show a mix of failures (Class 0 - blue) and successes (Class 1 - orange), but **later flights (higher flight numbers)** show a **dominant proportion of successes (Class 1)**, reflecting continuous reliability improvement.
2. KSC LC 39A (The Modern Platform):
   - Introduced **later in the flight sequence** (mid-to-late flights).
   - Exhibits an **extremely high initial success rate (Class 1)**. This suggests the site was brought online after the core technology and processes were mature, inheriting a high operational reliability from the start.
3. VAFB SLC 4E (The Specialized Site):
   - Used **less frequently** (typically for polar orbits).
   - Shows a **consistently high success rate (Class 1)** across its flights.

Payload vs. Launch Site



1. CCAFS SLC 40 (The Workhorse):
   - Handles the **widest spectrum of payloads**, from light (near 0 kg) up to the heaviest missions (approx. 15,000 kg).
   - Failures (Class 0) are mostly concentrated in the **mid-range payloads** (3,000 kg – 7,500 kg), while the heaviest payloads show high success.
2. KSC LC 39A (Newer, Reliable Site):
   - Concentrates on **mid-to-heavy payloads** (2,500 kg up to 15,000 kg) with distinct clusters.
   - Demonstrates an **exceptionally high success rate (Class 1)** across all mass categories it launches.
3. VAFB SLC 4E (Specialized Missions):
   - Primarily handles **lighter payloads** (under 2,500 kg) and a separate cluster of heavier missions (9,000 – 10,000 kg).
   - Maintains a **high overall success rate (Class 1)**, indicating high reliability for its specialized mission profiles.

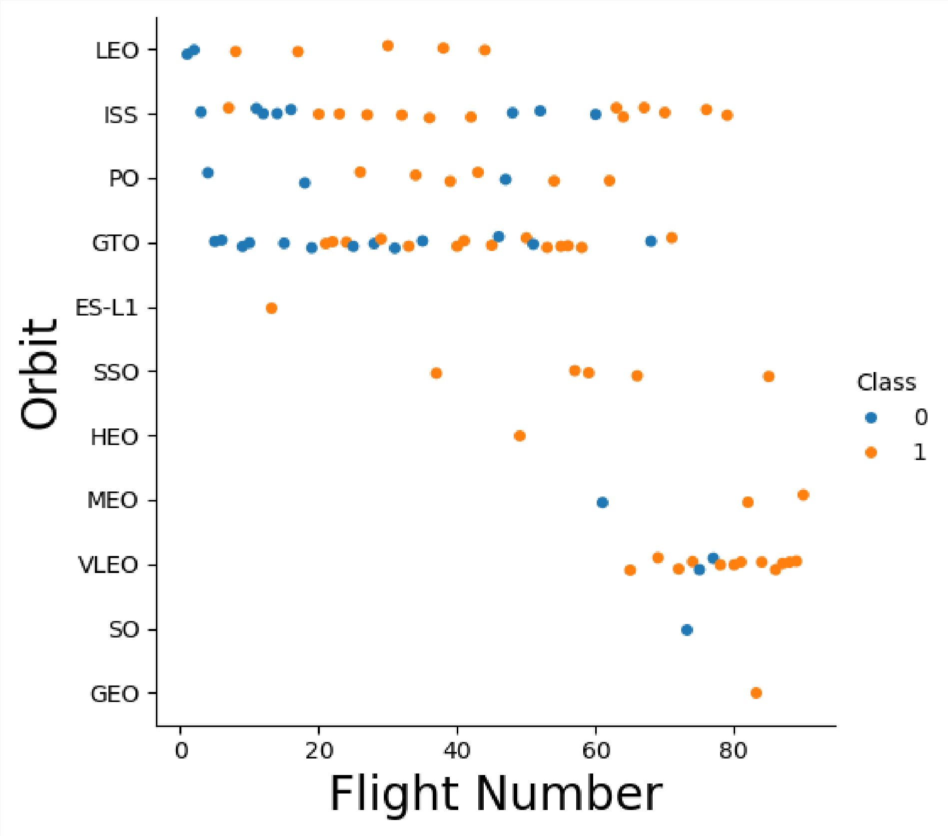
Success Rate vs. Orbit Type



1. Near-Perfect Success (Success Rate approximate 1.0)
   - ES-L1, GEO, HEO, and SSO all achieve near-perfect success rates. This confirms maximum reliability for high-energy (GEO/HEO) and standardized missions (SSO).

2. Mid-to-Low Success (Success Rate 0.58 to 0.85)
   - **LEO, MEO, PO, and ISS** show reliable but lowersuccess rates (approximate **0.65 to $0.75**). **GTO(0.58)** and **VLEO (0.85)** mark the historical high-risk boundaries, with GTO showing the lowestreliability, likely due to early developmentalchallenges for high-energy transfers.

Flight Number vs. Orbit Type



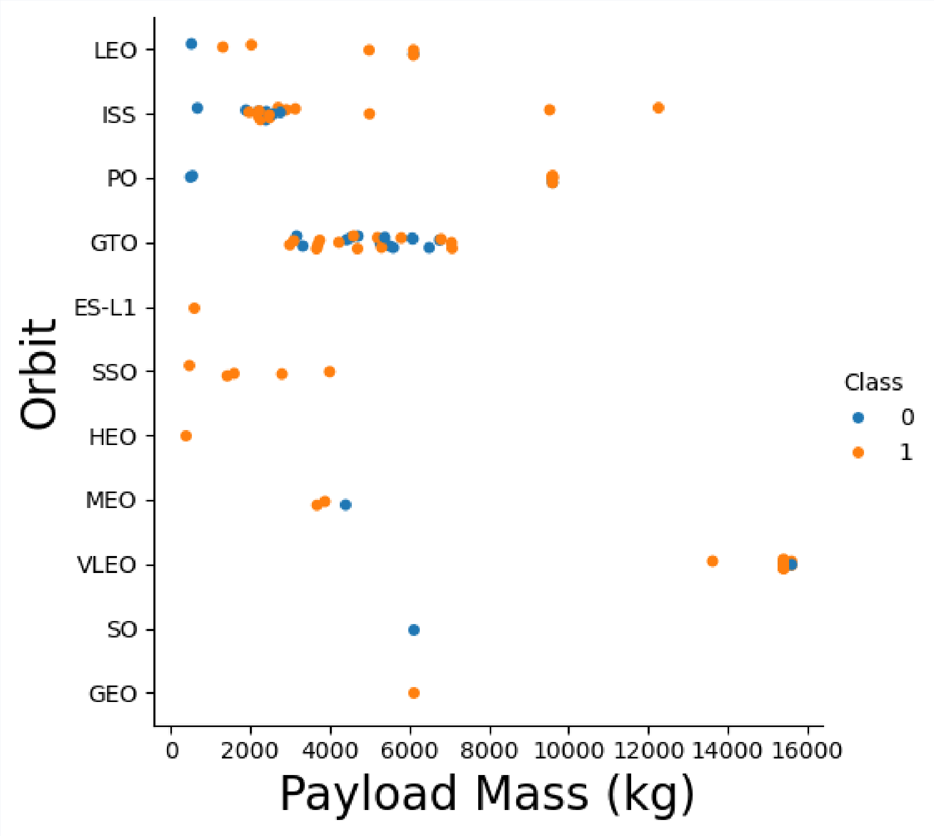
Early Flights (Low Flight Numbers)

- The early phase (Flight Numbers 0-40) was dominated by **LEO, ISS, PO, and GTO** missions.
- **Failures (Class 0 - Blue)** were common in this period, particularly for **LEO, ISS, and GTO**. This indicates the initial R&D phase where success was not guaranteed for these foundational mission types.
- **ES-L1** missions appear very early with high success.

Operational Maturity (Mid-to-High Flight Numbers)

- As the **Flight Number increases (above ~40)**, the number of **failures (Class 0)** across all orbits **dramatically decreases**.
  - **GTO** missions show a near-perfect transition to success.
  - **LEO and ISS** missions become overwhelmingly successful.
- The launch manifest **expands significantly** to include newer, specialized orbits like **SSO, VLEO, MEO, HEO, and GEO**.
  - **VLEO** missions appear later and show a **high success rate** immediately upon introduction.

Payload vs. Orbit Type



Payload Mass Distribution

- **Common Orbits (GTO, ISS, LEO):** Primarily handle **mid-range payloads** (2,000 kg to 7,000 kg).
- **VLEO** stands out, handling an **extremely heavy cluster** of successful missions (around 15,000 kg).

Success/Failure Patterns

- The majority of **failures (Class 0)** occurred with **mid-range payloads** (2,000–7,000 kg), mostly destined for **GTO**, reflecting early developmental challenges.
- **Max-capacity launches (15,000 kg)** show **near-perfect success (Class 1)**, confirming high technological maturity for the heaviest missions.

Launch Success Yearly Trend

Early Phase (2010–2015)

- **Initial Struggle (2010–2013):** Success rate was **0.0 (0%)** before stabilizing at approximately **0.34 (34%)** for 2014-2015.
- This represents the challenging **developmental and testing period**.

Maturity Phase (2016–2020)

- **Rapid Jump:** Success rates rapidly **soared past 0.60** in 2016 and peaked at **over 0.90 (90%) in 2019**.
- This phase demonstrates the achievement of **sustained, near-perfect operational reliability**.

All Launch Site Names

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

The unique launch sites used in the SpaceX missions, based on the provided list, are:

- **CCAFS LC-40** (Cape Canaveral Air Force Station Launch Complex 40)
- **VAFB SLC-4E** (Vandenberg Air Force Base Space Launch Complex 4E)
- **KSC LC-39A** (Kennedy Space Center Launch Complex 39A)
- **CCAFS SLC-40** (Cape Canaveral Air Force Station Space Launch Complex 40)

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

Total Payload Mass

Total payload mass carried by boosters launched by NASA (CRS)

Customer	Total_Payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

Booster_Version	Average_Payload_KG
F9 v1.1 B1003	2534.6666666666665

First Successful Ground Landing Date

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

Landing_Outcome	First_Successful_Landing_Date
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1020	Failure (drone ship)	5271
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

Mission_Outcome	Outcome_Count
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.

Booster_Version	MAX(PAYLOAD_MASS__KG_)
F9 B5 B1060.3	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1056.4	15600
F9 B5 B1051.6	15600

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.

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

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

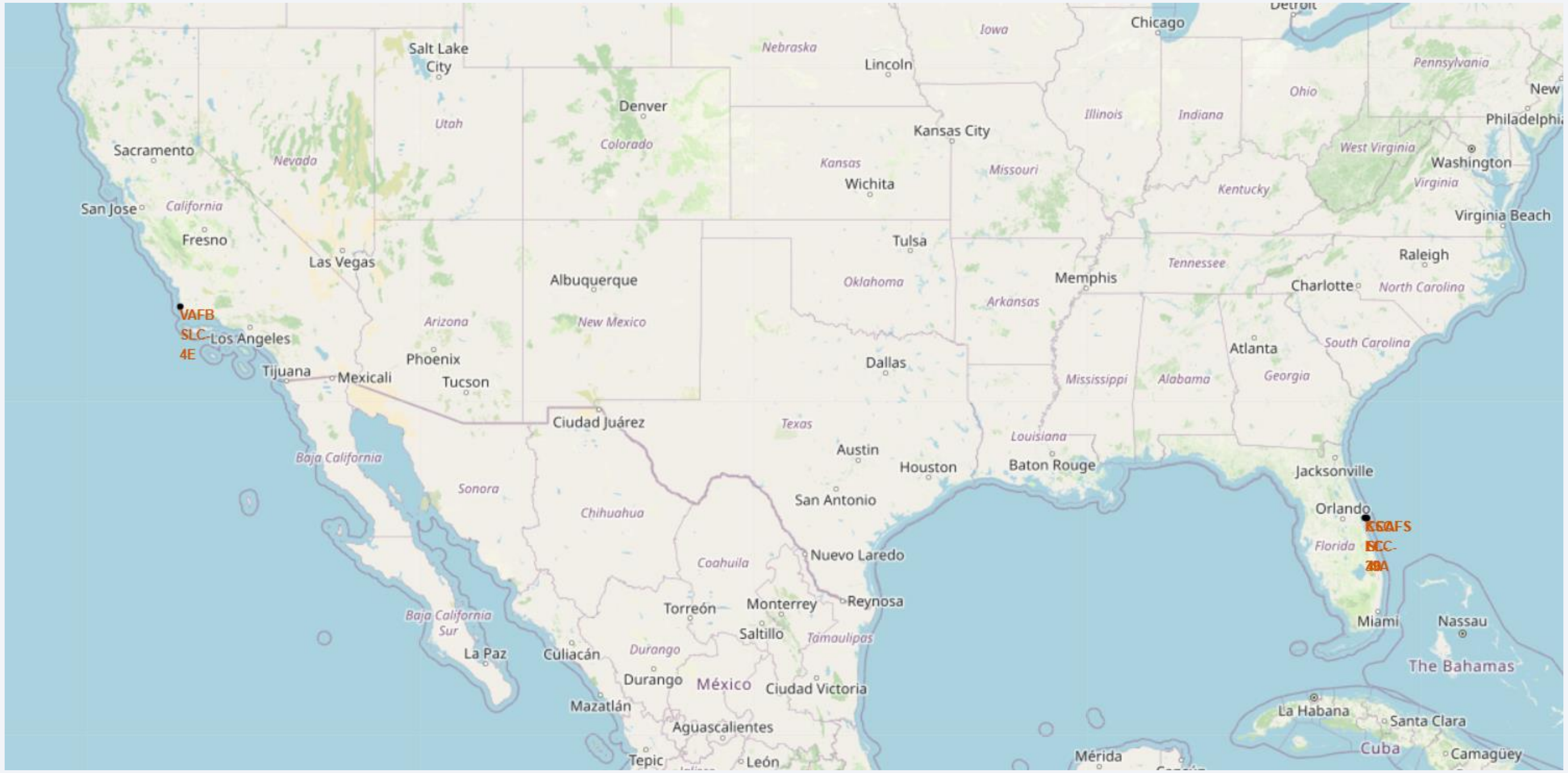
Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20

Landing_Outcome	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

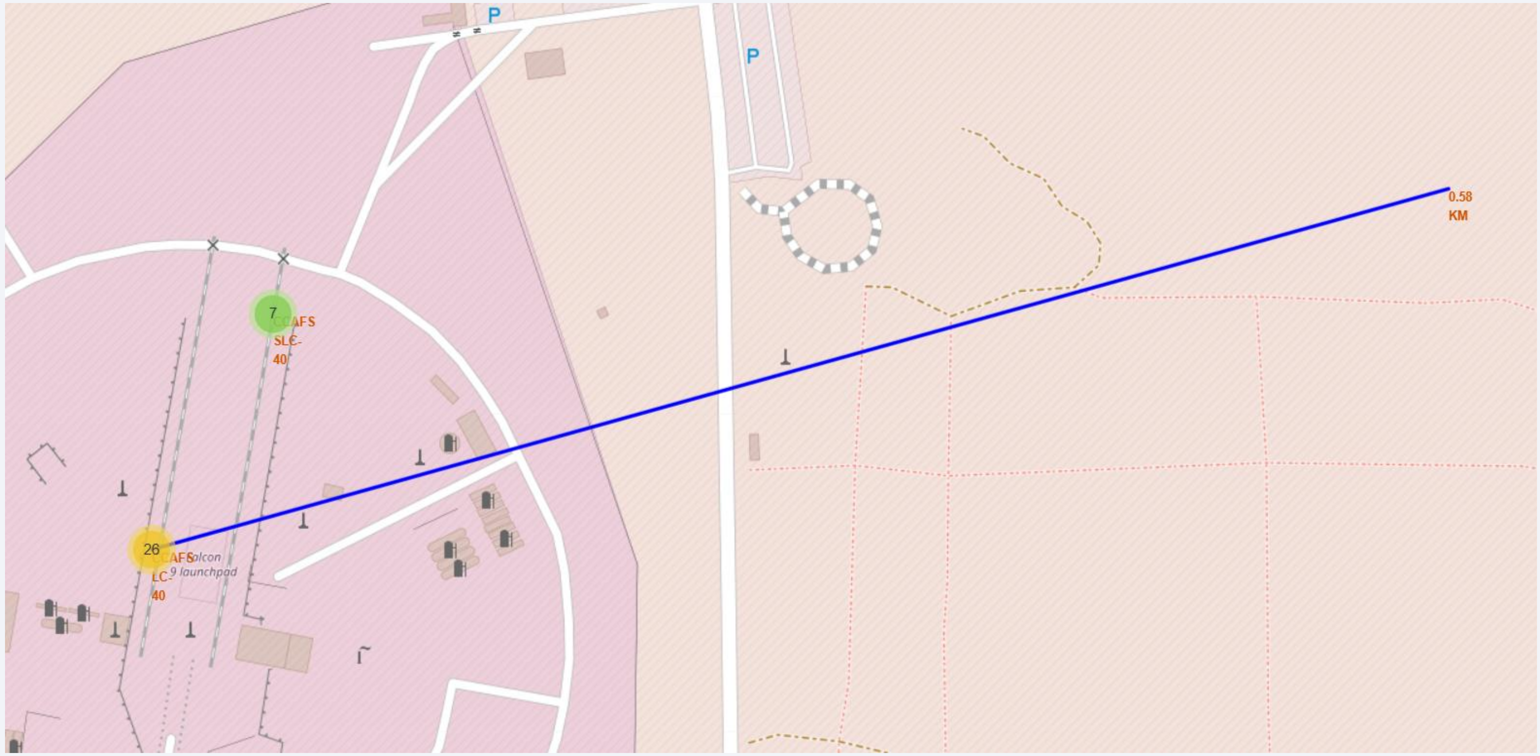
Launch Sites Proximities Analysis

All Launch Sites



Success and failed launches for each site

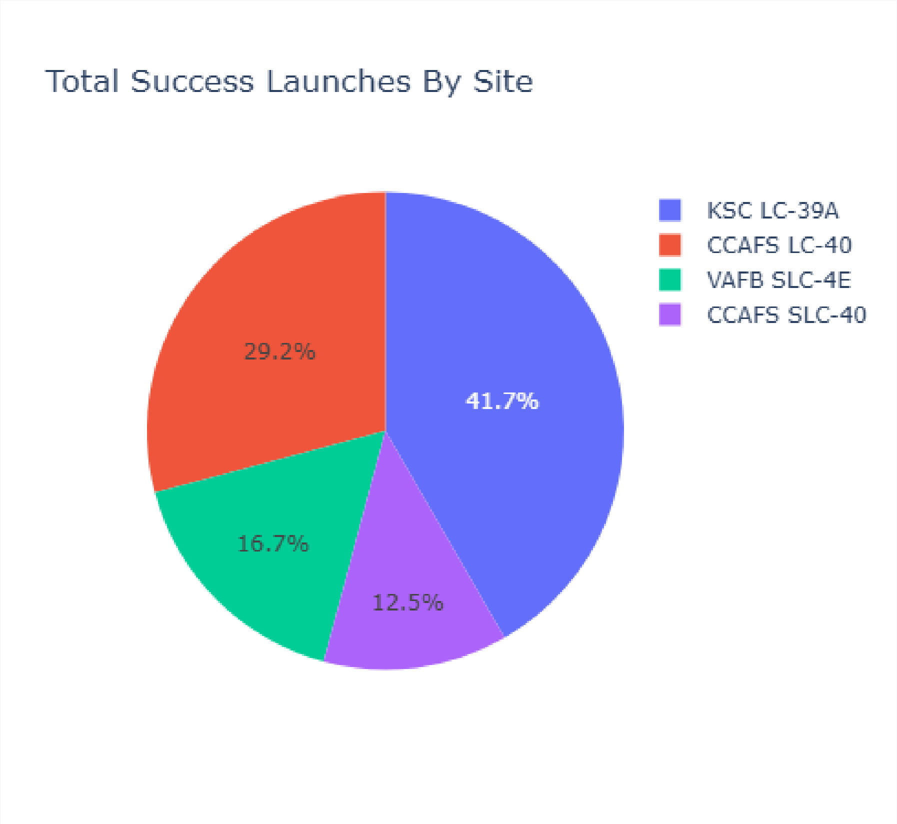
Launch Site and distance



Section 4

Build a Dashboard with Plotly Dash

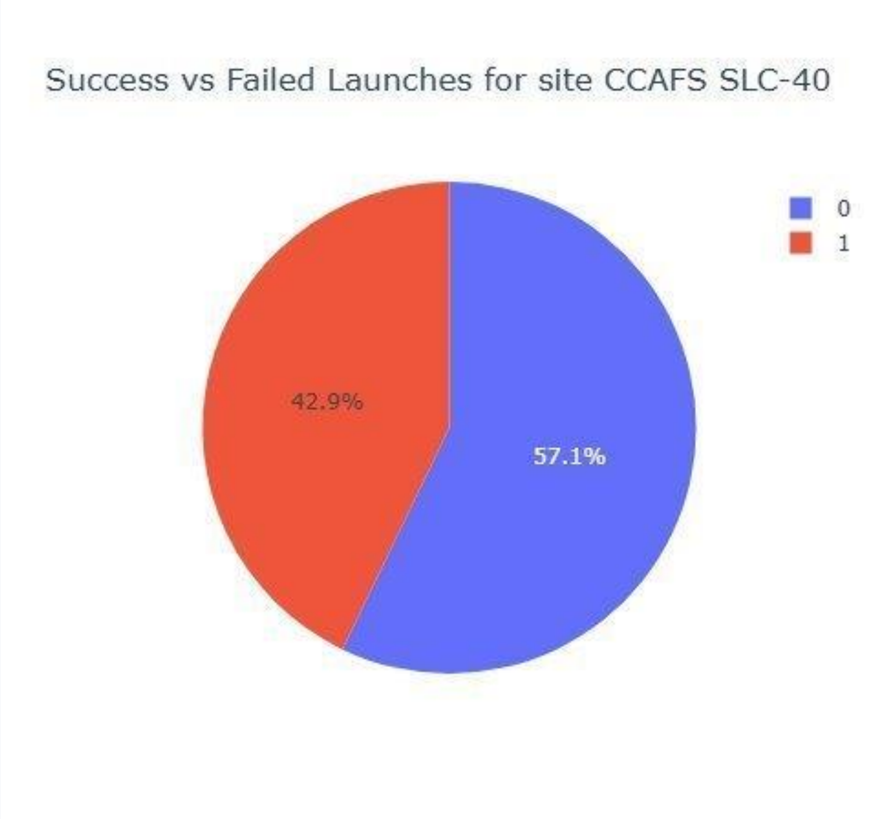
SpaceX Launch Records Dashboard



Summary of Total Successful Launches by Site

- **KSC LC-39A is the dominant site**, accounting for the largest share of successful launches at **41.7%**.
- **CCAFS LC-40** is the next major contributor with **29.2%** of successful missions.
- **VAFB SLC-4E** (16.7%) and **CCAFS SLC-40** (12.5%) make up the remainder.

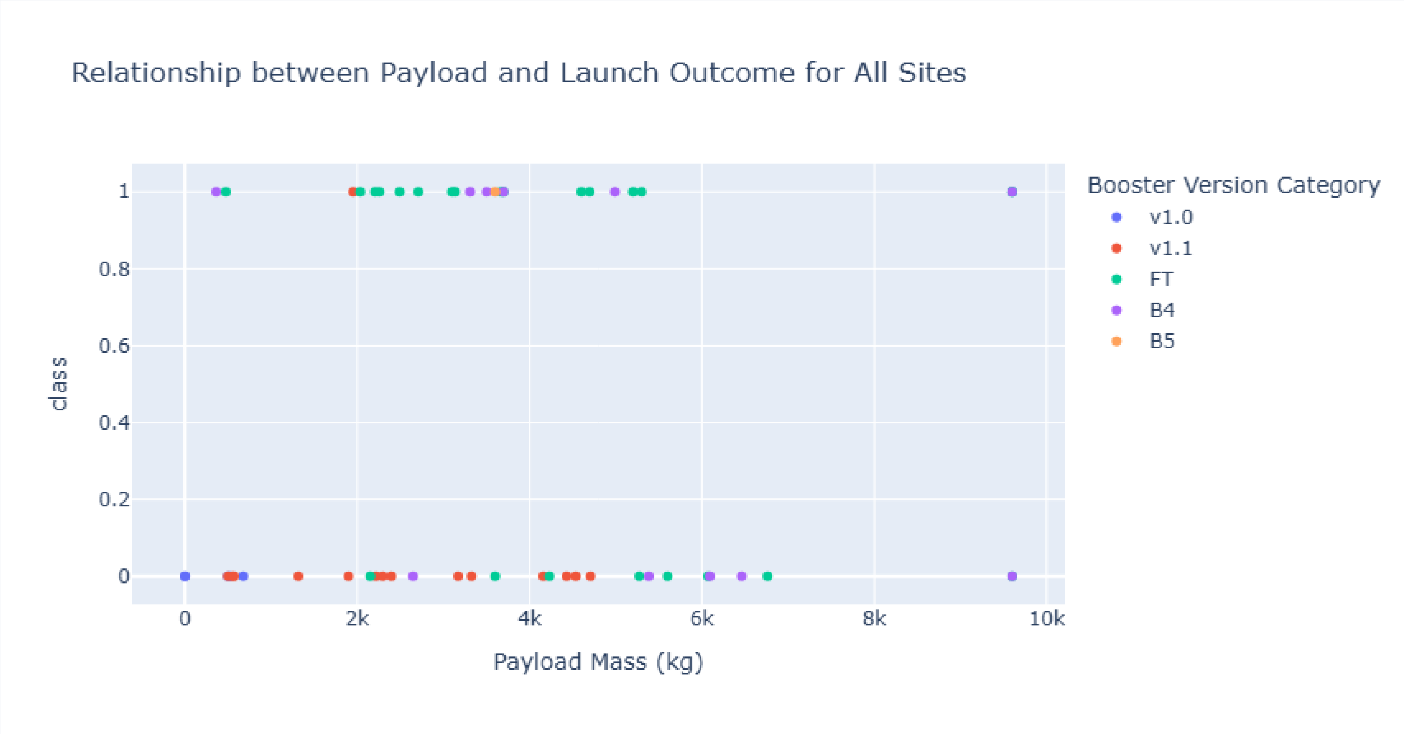
SpaceX Launch Records Dashboard



Success vs. Failed Launches for CCAFS SLC-40

- **Failures (Class 0 - Blue):** Missions launched from CCAFS SLC-40 resulted in **Failure** (Class 0) **57.1%** of the time.
- **Successes (Class 1 - Red):** Missions launched from CCAFS SLC-40 resulted in **Success** (Class 1) **42.9%** of the time.

SpaceX Launch Records Dashboard

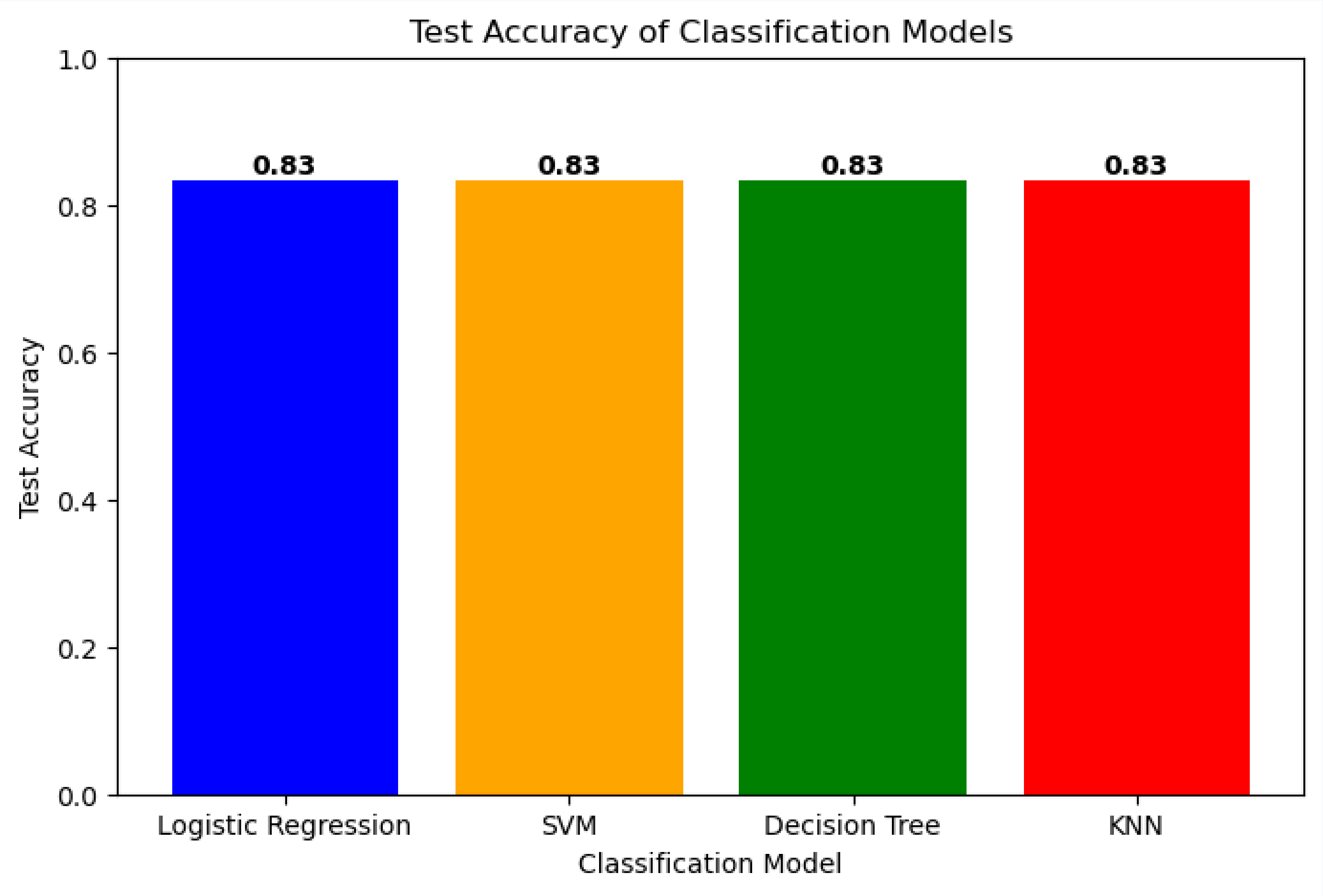


- **Launch Failures (Class 0)** are **exclusively associated with older booster versions (v1.0, v1.1)**, regardless of payload mass.
- **Launch Successes (Class 1)** are **predominantly achieved by newer booster versions (FT, B4, B5)**, spanning all payload masses they handle.

Section 5

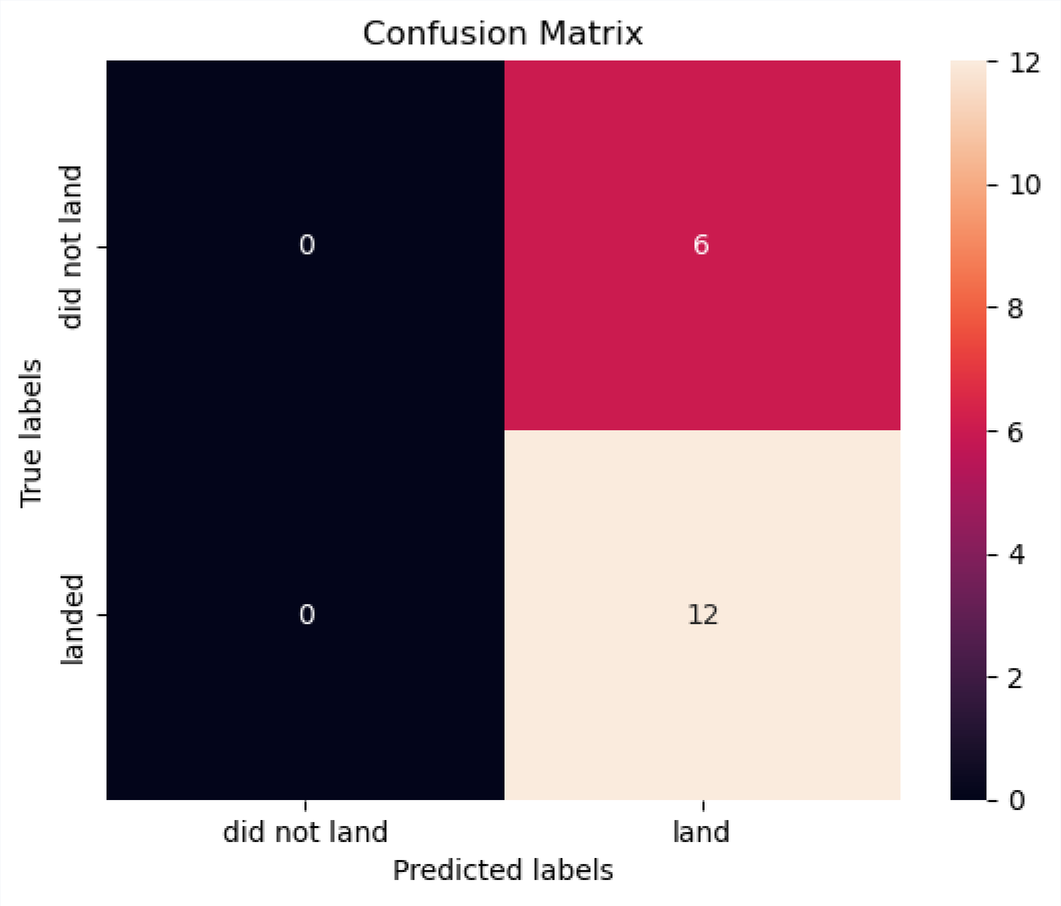
Predictive Analysis (Classification)

Classification Accuracy



- **Identical Performance:** The four distinct classification algorithms—**Logistic Regression**, **SVM** (Support Vector Machine), **Decision Tree**, and **KNN** (K-Nearest Neighbors)—all produced the **same test accuracy of 0.83**.

Confusion Matrix



- **Perfect Recall (1.00):** The model **correctly identified every single successful mission** (TP=12, FN=0). It has no False Negatives (FN).
- **Low Precision (0.67):** However, the model **incorrectly predicted success 6 times** (False Positives - FP=6) when the mission actually failed.

Conclusions

Key Findings & Strategic Implications

1. **Technology is the Success Factor, Not Payload Mass.**
   - Launch success is driven by **Booster Version** (**FT, B4, B5**), achieving near-perfect reliability.
   - Failures are **historical artifacts** tied exclusively to older versions (**v1.0, v1.1**).
2. **High Predictability for Reusability.**
   - Machine Learning models achieved **83% Test Accuracy**, confirming the high feasibility of predicting successful first-stage landing and reuse.
3. **Operational Maturity Achieved.**
   - The yearly success rate rapidly increased, stabilizing **above 80%** by 2019-2020.
   - **KSC LC-39A** is the dominant site for successful missions (41.7%), reflecting the transition to a highly reliable operational hub.
4. **Competitive Edge:**
   - SpaceX maintains a significant competitive advantage due to predictable, high-reliability reusable technology. Competitors must bridge the technological gap to challenge the $62M launch price.

Appendix

- https://github.com/Noppon-Ch/Winning-Space-Race-With-Data-Science-IBM-Lab/tree/main

Thank you!