IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rodrigo de Camargo Rodrigues
08/15/2025

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

Executive Summary - Methodologies

Data Collection - Extracted structured data from the SpaceX API; Applied web scraping on Wikipedia launch tables for complementary details.

Data Wrangling - Cleaned raw datasets; Normalized columns (dates, landing success/failure).Handled missing values and standardized categorical values.

Exploratory Data Analysis (EDA) - SQL queries: examined launch frequencies, launch sites, and booster versions; Visualization with Matplotlib & Seaborn: histograms, bar plots, and trend analysis; Geospatial mapping with Folium: mapped launch sites and landing zones to explore geographic patterns.

Feature Engineering - Created categorical variables such as orbit type and landing outcome; Performed one-hot encoding with pd.get_dummies to prepare features for ML models.

Model Development - Split dataset into 80/20 train-test sets; Trained and compared four algorithms: Logistic Regression, Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbors (KNN).

Executive Summary - Methodologies

HYPERPARAMETER TUNING - APPLIED GRIDSEARCHCV WITH 10-FOLD CROSS-VALIDATION; TUNED: LOGISTIC REGRESSION → C; SVM → KERNEL TYPES; DECISION TREE → DEPTH, SPLITTER; KNN → NUMBER OF NEIGHBORS, DISTANCE METRIC.

MODEL EVALUATION - USED ACCURACY SCORE ON THE TEST SET AS MAIN METRIC; COMPARED RESULTS ACROSS ALL ALGORITHMS.

INTERACTIVE DASHBOARD (PLOTLY DASH) - BUILT A REAL-TIME ANALYTICS DASHBOARD WITH PLOTLY DASH; INPUTS: DROPDOWN FOR LAUNCH SITE (ALL SITES VS. SPECIFIC SITE); RANGE SLIDER FOR PAYLOAD MASS. OUTPUTS: PIE CHART → SHOWS SUCCESS RATES ACROSS SITES OR SUCCESS/FAILURE PER SITE; SCATTER PLOT → PAYLOAD VS. SUCCESS OUTCOME, COLORED BY BOOSTER VERSION.

PURPOSE: ALLOWS EXPLORATION OF PATTERNS INTERACTIVELY, COMPLEMENTING STATIC ANALYSIS AND ML MODELS.

Executive Summary - Results

Dataset - ~90 test records; **83 engineered features** after one-hot encoding.

Model Performance - Support Vector Machine (SVM), Best kernel: **RBF,** Test accuracy: **~83.33%** - **Logistic Regression,** Best parameter: **C = 1, solver = lbfgs,** Test accuracy: **~83.33% - K-Nearest Neighbors (KNN),** Best parameter: **k = 5, algorithm = auto, p=2,** Test accuracy: **~80% - Decision Tree,** Best parameter: **depth = 6, splitter = best,** Test accuracy: **~73.33%**

Best Performing Models - SVM (RBF) and Logistic Regression tied as top performers (83.33% accuracy); Both outperformed KNN and Decision Tree.

Dashboard Insights - The **interactive dashboard** validated model findings: Certain payload ranges correlated with higher success probability; Some sites consistently outperformed others in launch success rates; Booster versions had visible impact on landing outcomes.

Introduction

- In this capstone project, our goal is to predict whether the first stage of the Falcon 9 rocket will successfully land. SpaceX advertises Falcon 9 launches at a cost of around $62 million, while other providers charge more than $165 million per launch. A significant portion of this cost advantage comes from reusability of the first stage.

- Therefore, if we can determine whether the first stage will land, we can also estimate the true cost of a launch. These insights can be highly valuable for alternative providers who want to bid competitively against SpaceX in the commercial spaceflight market.

- This project is structured in several modules, each building toward the final objective:

- Data Collection: We gathered data on Falcon 9 launches and landings using a RESTful API and web scraping.

- Data Wrangling: We transformed the collected data into a clean DataFrame, handling missing values and standardizing categories.

- Interactive Visualization: We built a Plotly Dash dashboard for interactive analysis of launch records and created geospatial maps with Folium to study the proximity of launch sites.

- Machine Learning Modeling: We applied supervised learning techniques (Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors) to predict landing outcomes. Using training and test splits, we performed hyperparameter tuning to identify the best-performing models.

- Integration and Insights: Finally, we compiled all results into a single framework to provide data-driven insights into Falcon 9 landing success.

Introduction

The project aims to answer the following central and supporting questions:

Central Question:

Will the Falcon 9 first stage land successfully?

Supporting Questions:

Which features (payload mass, orbit type, launch site, booster version, etc.) most strongly influence landing outcomes?

How can visualization tools (dashboards and maps) improve understanding of launch performance?

Which machine learning algorithms provide the most accurate predictions for landing success?

Can predictive modeling help estimate the economic advantage of reusability in space launches?

By addressing these questions, this project combines data collection, visualization, and machine learning into a complete analytical pipeline that supports decision-making in the aerospace industry.

Section 1

Methodology

Methodology

EXECUTIVE SUMMARY

DATA COLLECTION METHODOLOGY

PERFORM DATA WRANGLING

PERFORM EXPLORATORY DATA ANALYSIS (EDA) USING VISUALIZATION AND SQL

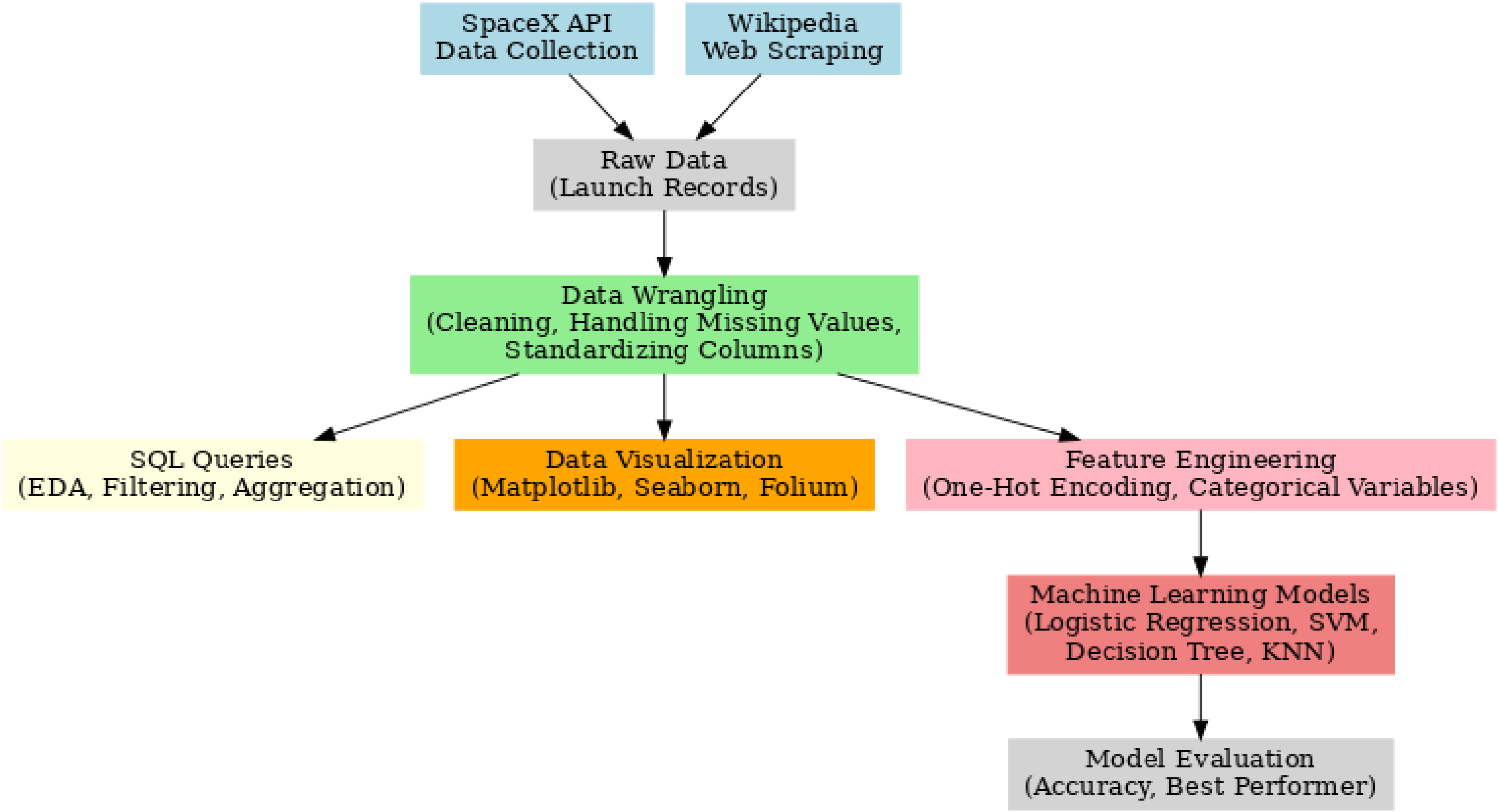
PERFORM INTERACTIVE VISUAL ANALYTICS USING FOLIUM AND PLOTLY DASH

PERFORM PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

Data Collection

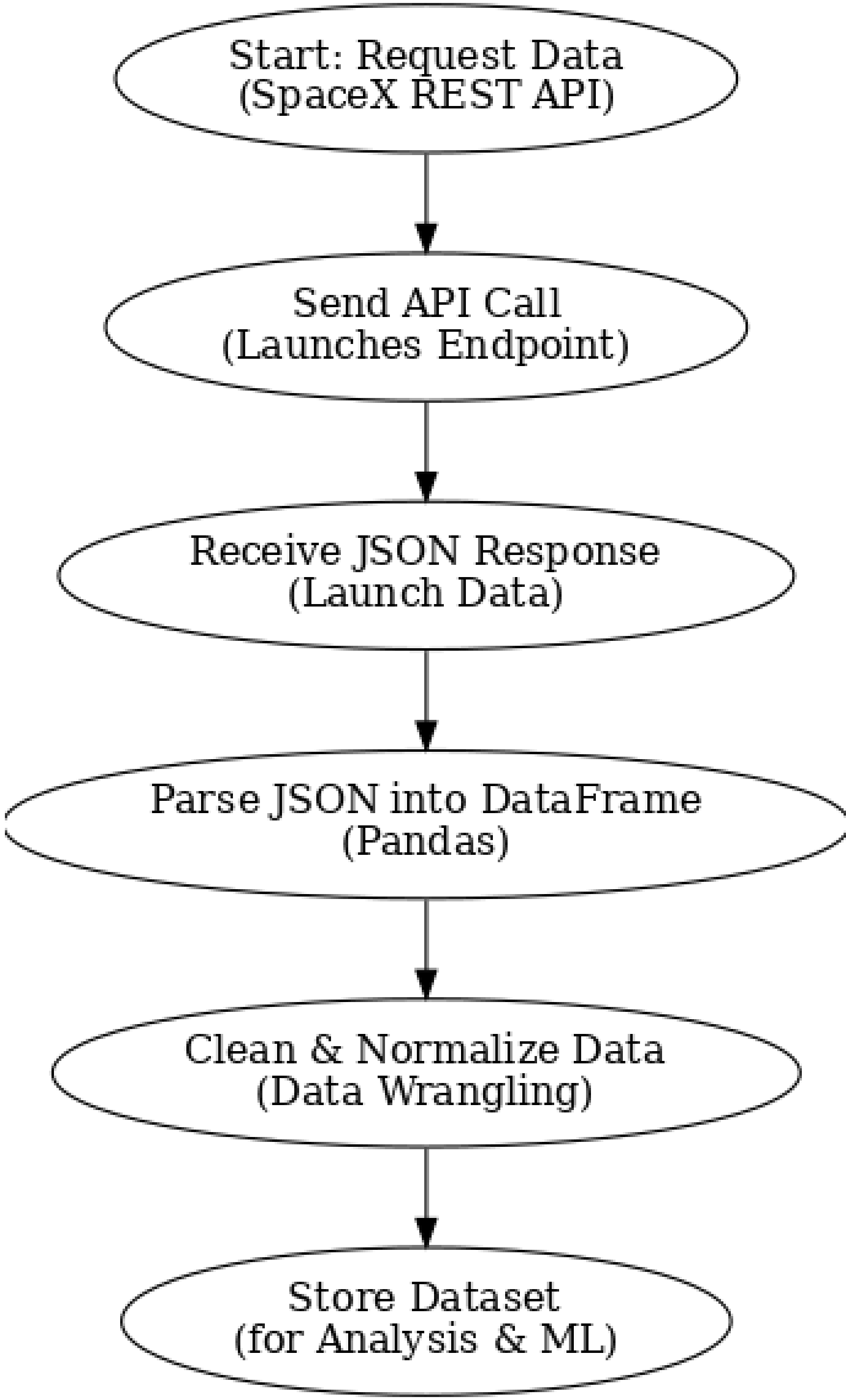
- ✅ Data was collected from **SpaceX API** and **Wikipedia Web Scraping**
  ✅ Transformed into **Pandas DataFrames** for cleaning and wrangling
  ✅ Stored and queried with **SQLite (SQL queries)**
  ✅ Enriched with **Folium maps for geospatial analysis**
  ✅ Used in **Matplotlib, Seaborn, and Plotly Dash** for visualization and machine learning

Data Collection



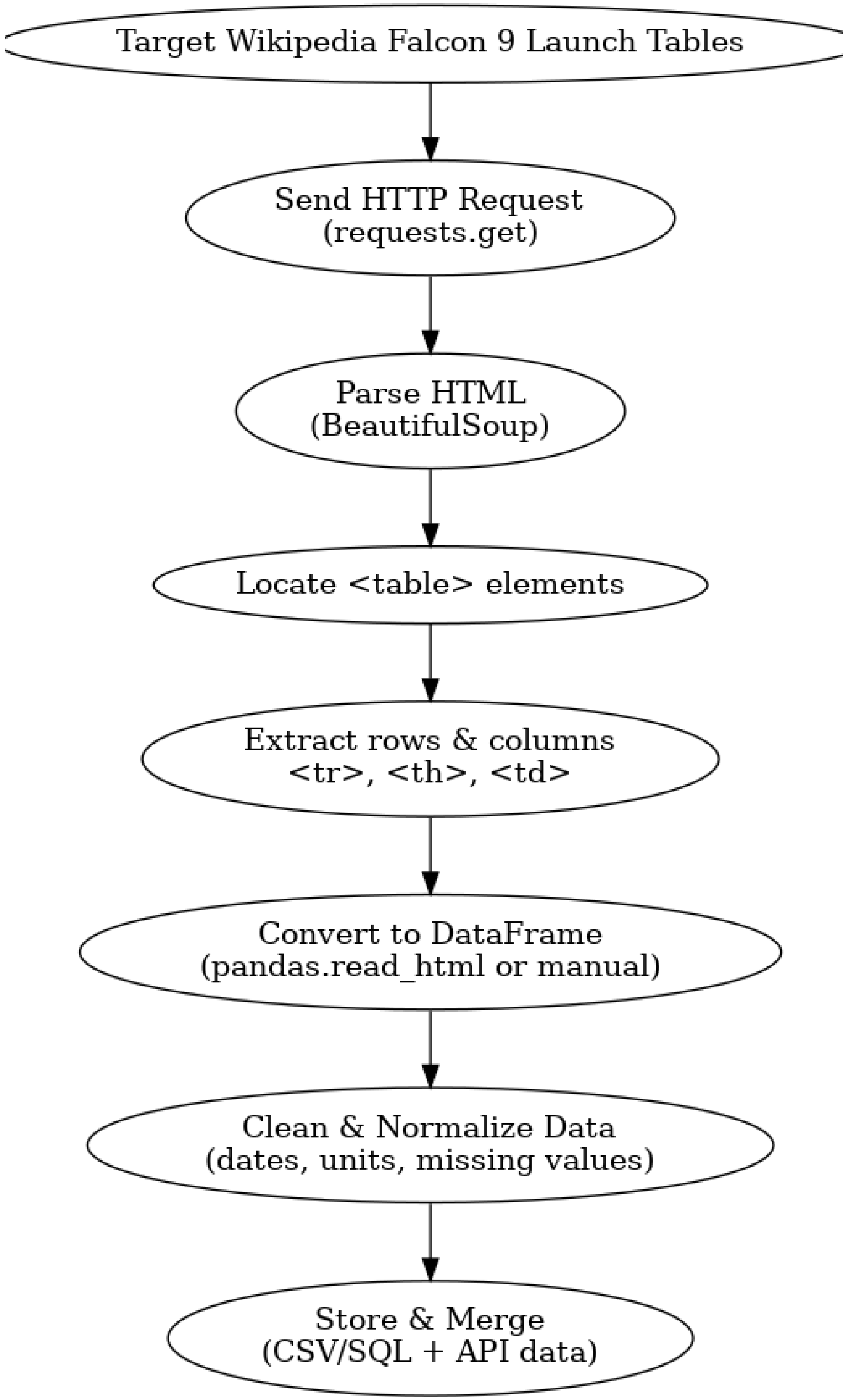
Data Collection – SpaceX API

- API Endpoint: Used SpaceX REST API (https://api.spacexdata.com/v4/launches)
- Request Method: GET calls using Python requests library
- Response Format: JSON (launch data, rocket, payload, launchpad info)
- Data Wrangling: Converted JSON → Pandas DataFrame
- Features Extracted: Launch site, rocket type, payload mass, orbit, landing outcome
- Normalization: Handled missing values, standardized success/failure
- https://github.com/digaorcr/Capstone/blob/main/Rodrigo-Rodrigues-spacex-data-collection-api.ipynb



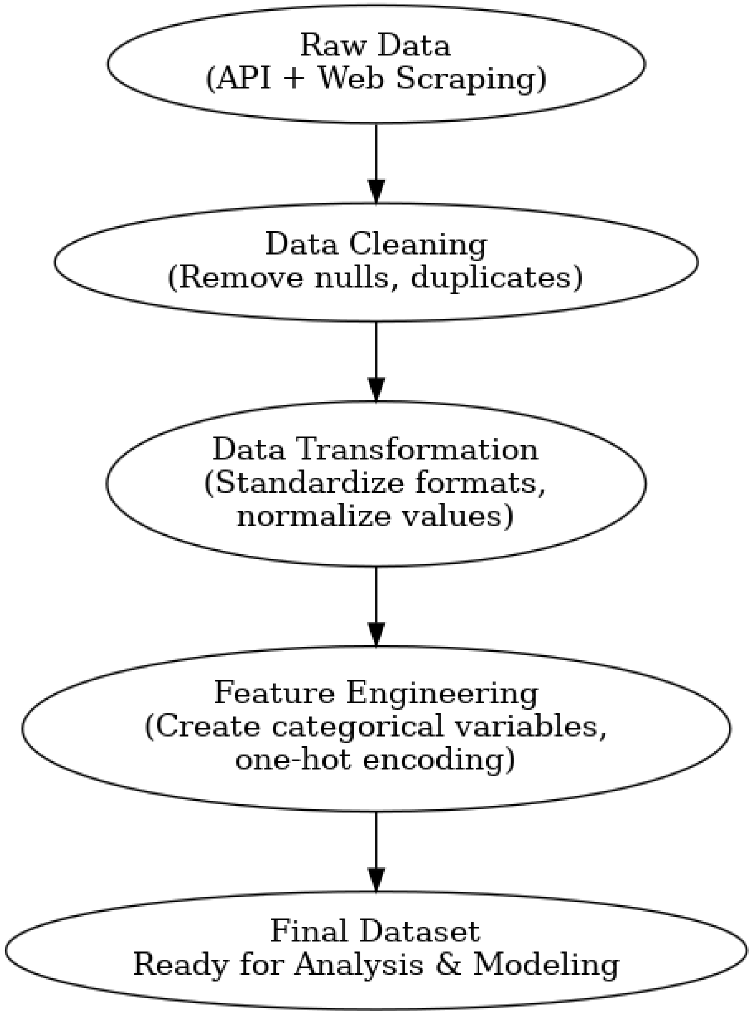
Data Collection - Scraping

- Target: Wikipedia Falcon 9 launch tables
- Request: requests.get(URL)
- Parser: BeautifulSoup to parse HTML
- Locate data: find <table>, iterate <tr>, <th>, <td>
- Extraction options: manual parsing or pandas.read_html()
- Normalize: build Pandas DataFrame
- Cleaning: fix types, dates, units; handle missing values; standardize categories
- Persist/Join: save to CSV/SQL and merge with API dataset
- https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-webscraping.ipynb



Data Wrangling

- Load raw data (API + Web Scraping)
- Check missing values
- Handle missing entries (replace / drop)
- Standardize categories (e.g., “Success”, “Failure”)
- Normalize columns (dates, landing outcome)
- One-hot encoding (orbit, landing sites, booster type)
- Feature engineering (categorical → numerical features)
- Final cleaned dataset ready for EDA & Modeling
- https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Histogram / Bar Charts

Show distribution of launch frequencies by year and launch site.

Helps identify trends over time and busiest sites.

Pie Charts

Show proportion of successful vs failed landings.

Useful for understanding success rate at a glance.

Scatter Plots

Relationship between payload mass and landing success.

Highlights whether heavier payloads affect landing outcomes.

Box Plots

Compare payload mass across orbits.

Detects variability and potential outliers.

Folium Maps (Geospatial)

Visualize launch sites and landing locations.

Shows geographic distribution and proximity to coastlines.

https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-edadataviz.ipynb

EDA with SQL

- Select launch records
  - Retrieved all launch records from the database for exploration.
- Filter by launch site
  - Extracted launches from specific sites (e.g., CCAFS, KSC, VAFB) to analyze performance.
- Count launches per site
  - Used COUNT() and GROUP BY to compare activity levels across sites.
- Find successful landings
  - Queried records where landing outcome = True to calculate success rates.
- Payload analysis
  - Retrieved minimum, maximum, and average payload mass.
  - Grouped payloads by orbit type.
- Booster version analysis
  - Counted occurrences of each booster version.
  - Identified the most frequently used booster.
- Join operations
  - Combined launch data with booster details to study correlations.
- Order & Limit
  - Sorted results by payload mass or launch date.
  - Limited output to highlight top results (e.g., heaviest payloads).
- https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers (Launch Sites)

Placed markers at each launch site (e.g., CCAFS, KSC, VAFB, Boca Chica).

Purpose: To identify the exact geographic location of each launch pad.

Circle Markers (Highlight Launch Sites)

Added circle markers around each site.

Purpose: To visually emphasize launch locations and differentiate them from other map elements.

Popups & Labels

Added popups to markers showing launch site names.

Purpose: To provide quick, interactive information when clicking on the map.

Lines (Distance Calculations)

Drew lines between launch sites and nearby cities/coastlines.

Purpose: To calculate and visualize distances (important for understanding safety, logistics, and accessibility).

Customized Icons

Used different icons/colors for launch sites and cities.

Purpose: To distinguish between types of locations on the same map.

Build an Interactive Map with Folium

They allow users to visually explore launch site geography, understand proximity to populated areas, and analyze strategic location choices by SpaceX.

https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

What I added

Launch Site dropdown (id="site-dropdown")

- Options: “ALL” + each unique launch site
- Features: searchable, placeholder, default “ALL”

Pie chart (id="success-pie-chart")

- For ALL sites: slices = sites, values = sum of class (successes)
- For a selected site: slices = {Success, Failure} with counts

Payload range slider (id="payload-slider")

- 0–10,000 kg, step 1,000; initialized to min/max from the data
- Visible tooltip and labeled tick marks

Scatter plot (id="success-payload-scatter-chart")

- X = Payload Mass (kg), Y = class (0/1)
- Color = Booster Version Category
- Hover shows Launch Site
- Title updates based on site selection

How it interacts (callbacks)

Dropdown → Pie chart

- Selecting a site switches the pie from “successes by site” (global) to “success vs. failure” (site-specific)

. **Dropdown + Slider → Scatter plot**

Both controls filter the data shown: site selection and payload range jointly determine which points appear and how the title reads.

Build a Dashboard with Plotly Dash

- Why I added these
  - Dropdown filter: Lets you compare across sites and then zoom into a single site without changing pages—great for exploratory analysis.
  - Pie chart:
    - At the global level, it quickly shows which sites contribute most to successful launches.
    - At the site level, it shows success rate vs. failures—a simple, intuitive KPI view.
  - Payload slider: Payload is a key operational variable. A range slider makes it fast to focus on light vs. heavy payloads and see how outcomes differ.
  - Scatter plot (payload vs. success): Shows the relationship between payload and launch outcome, while color by Booster Version Category reveals whether vehicle family moderates that relationship.
  - Dynamic titles/labels & hover: Improves interpretability (clear 0/1 meaning, context in hover, titles reflect the current filter).
  - Net effect: one screen where you can (1) spot which sites succeed most, (2) check each site's win/loss mix, and (3) test how payload and booster type relate to mission success—without manual data wrangling.
  - https://github.com/digaorcr/Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

1. Data Preparation

Cleaned and preprocessed SpaceX launch dataset.

Encoded categorical features and normalized numeric ones.

Split into training set and test set.

2. Models Built

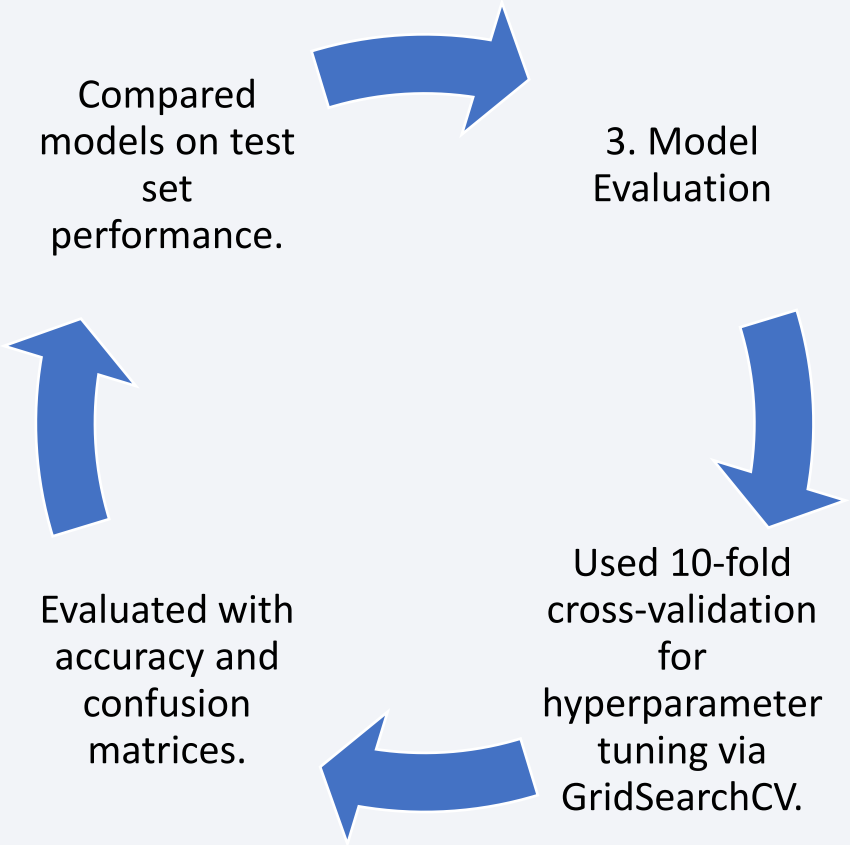
Logistic Regression

Support Vector Machine (SVM)

Decision Tree

K-Nearest Neighbors (KNN)

Predictive Analysis (Classification)



4. Hyperparameter Tuning (best results)

Logistic Regression → Best C=0.01, solver=lbfgs → CV accuracy ≈ 0.846, Test accuracy = 0.833

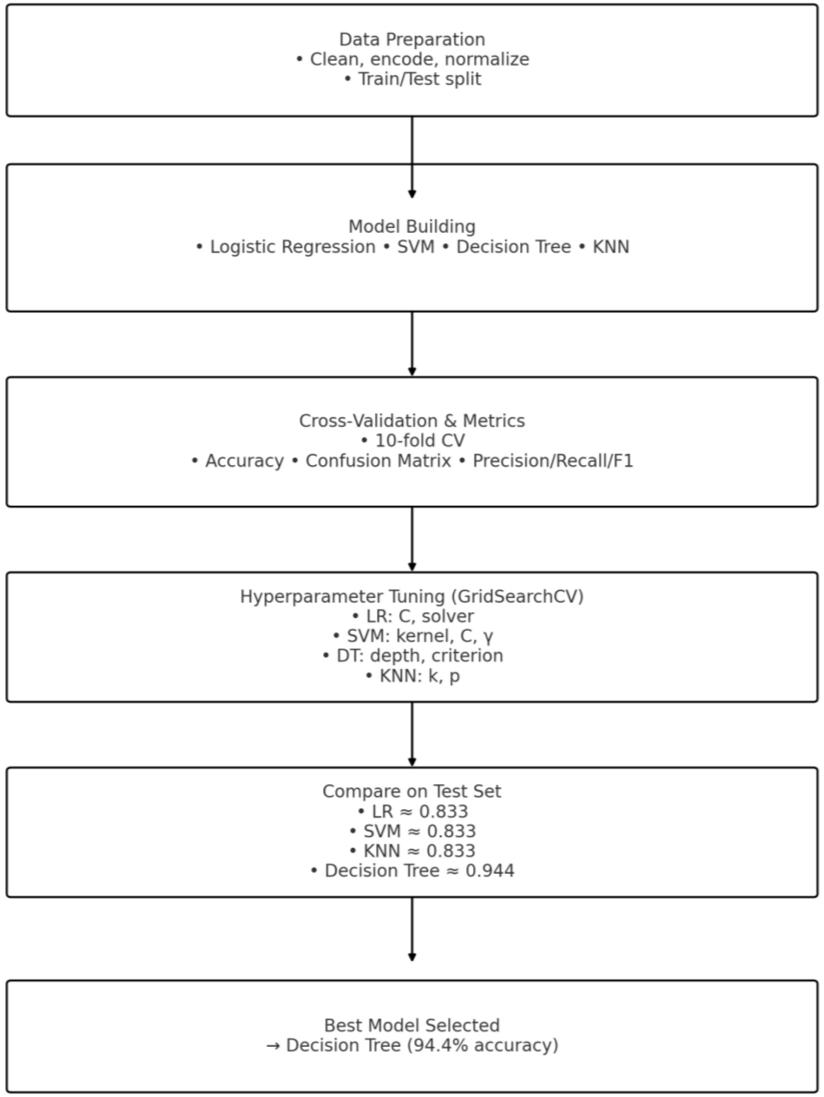
SVM → Best kernel=sigmoid, C=1.0, gamma≈0.032 → CV accuracy ≈ 0.848, Test accuracy = 0.833

Decision Tree → Best criterion=entropy, depth=12 → CV accuracy ≈ 0.873, Test accuracy = 0.944

KNN → Best k=10, p=1 → CV accuracy ≈ 0.848, Test accuracy = 0.833

Predictive Analysis (Classification)

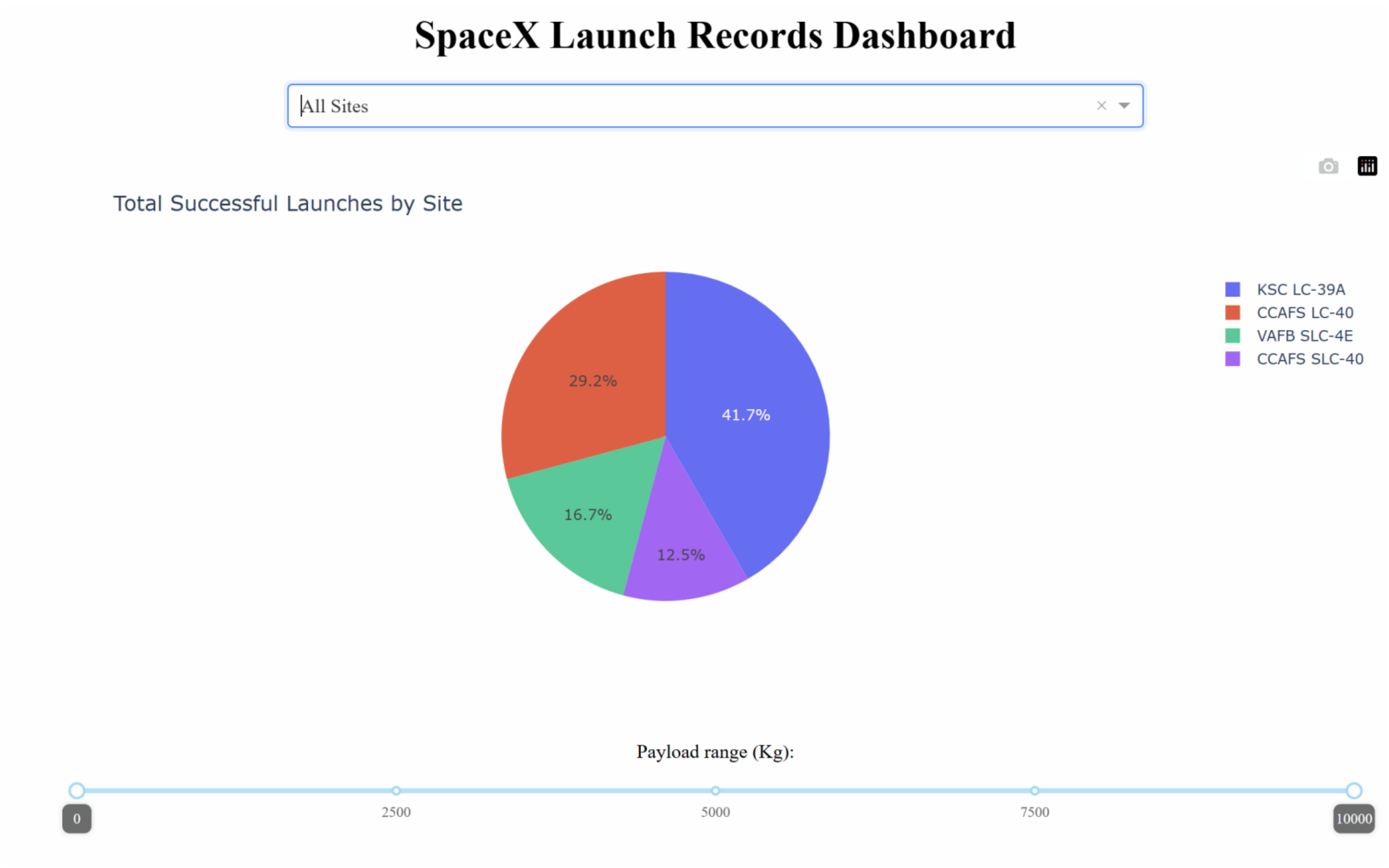
- 5. Best Model
- Decision Tree Classifier was the best performing model.
- Achieved 94.4% accuracy on the test set, outperforming Logistic Regression, SVM, and KNN (all ~83.3%).
- https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



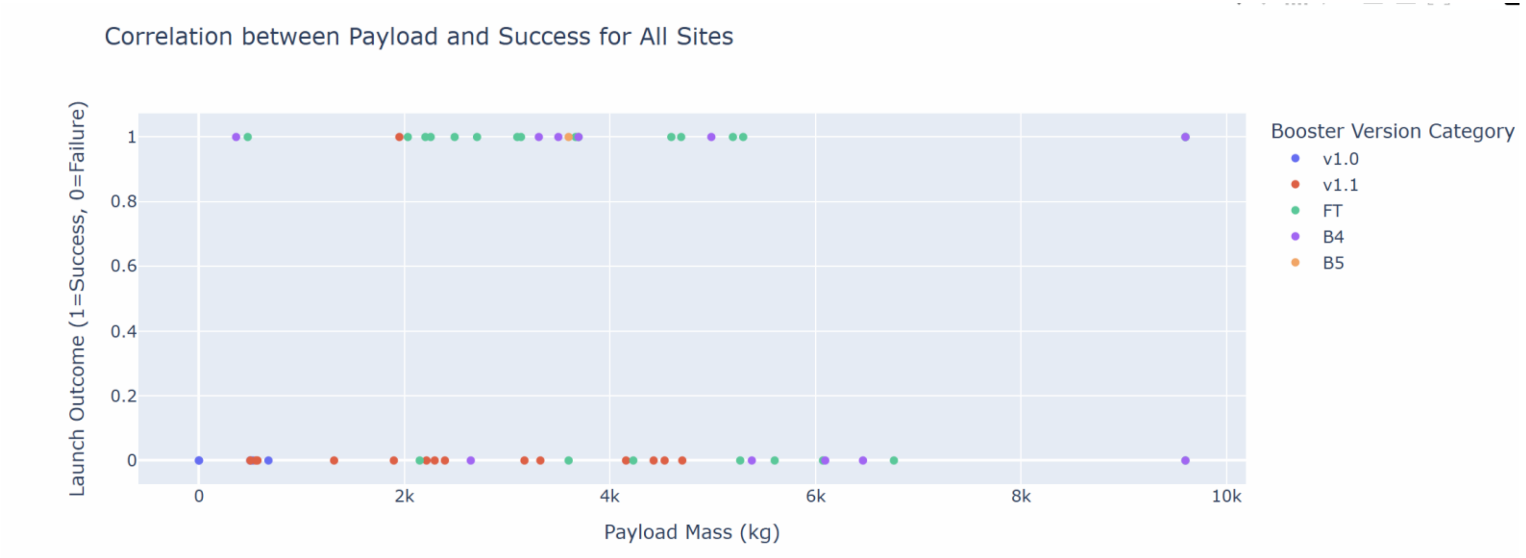
Results

- Exploratory Data Analysis (EDA) Results
  - Used API, Web Scraping, and SQL queries to collect and prepare SpaceX launch data.
  - Performed data wrangling (handling nulls, normalizing payloads, encoding categorical values).
  - Visualized trends:
    - Launch success rates varied significantly across launch sites.
    - Heavier payloads showed lower success probabilities beyond ~8,000 kg.
    - Booster Version Category strongly influenced outcomes, with some boosters achieving much higher reliability.
  - Geospatial analysis of launch site coordinates confirmed site clustering and operational distribution.

Results


SpaceX Launch Records Dashboard
All Sites
Total Successful Launches by Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
2500
5000
7500
10000

Results


Correlation between Payload and Success for All Sites
Launch Outcome (1=Success, 0=Failure)
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Results

- Interactive Analytics Demo (Dash App)
  - Built a Dash-powered dashboard for interactive exploration.
  - Dropdown filter: Select “ALL Sites” or drill down to a specific launch site.
  - Pie chart:
    - For “ALL Sites”: compares total successful launches per site.
    - For a selected site: shows Success vs. Failure outcomes.
  - Range slider: Adjusts payload mass (0–10,000 kg).
  - Scatter plot: Payload vs. Success, colored by Booster Version Category, showing correlation patterns.
  - Together, these features allow users to interactively test hypotheses (e.g., “Do heavier payloads reduce success at CCAFS?”).

Results

- Predictive Analysis Results (Machine Learning)
  - Built four classification models:
    - Logistic Regression, SVM, Decision Tree, KNN.
  - Applied GridSearchCV + cross-validation for hyperparameter tuning.
  - Model accuracies (test set):
    - Logistic Regression ≈ 83.3%
    - SVM ≈ 83.3%
    - KNN ≈ 83.3%
    - Decision Tree ≈ 94.4% (best)
  - Decision Tree Classifier selected as final model due to its superior accuracy and interpretability.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Scatter Plot: *Flight Number vs. Launch Site with Launch Outcome*

X-axis: Flight sequence (earlier → later launches).

Y-axis: Launch site (CCAFS SLC 40, KSC LC 39A, VAFB SLC 4E).

Colors:

- Blue (0) = Failure
- Orange (1) = Success

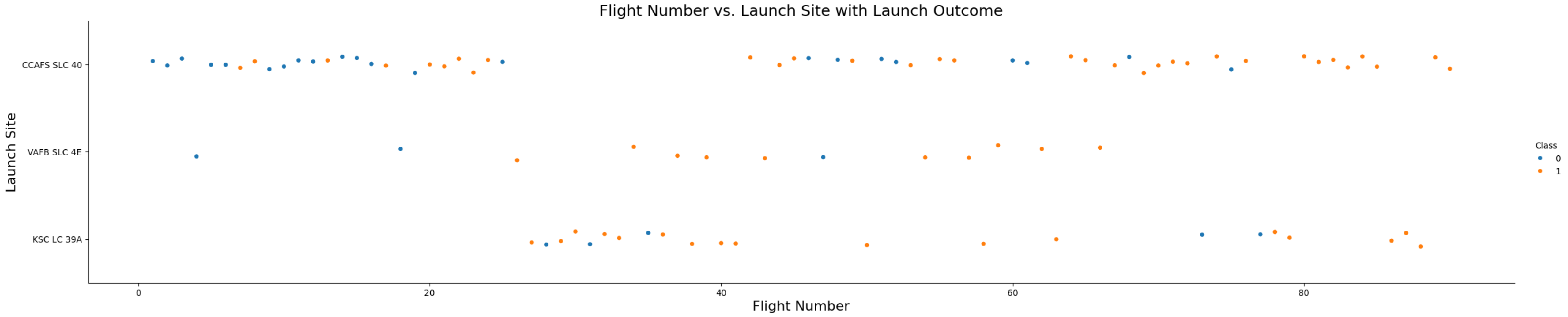
Key takeaways:

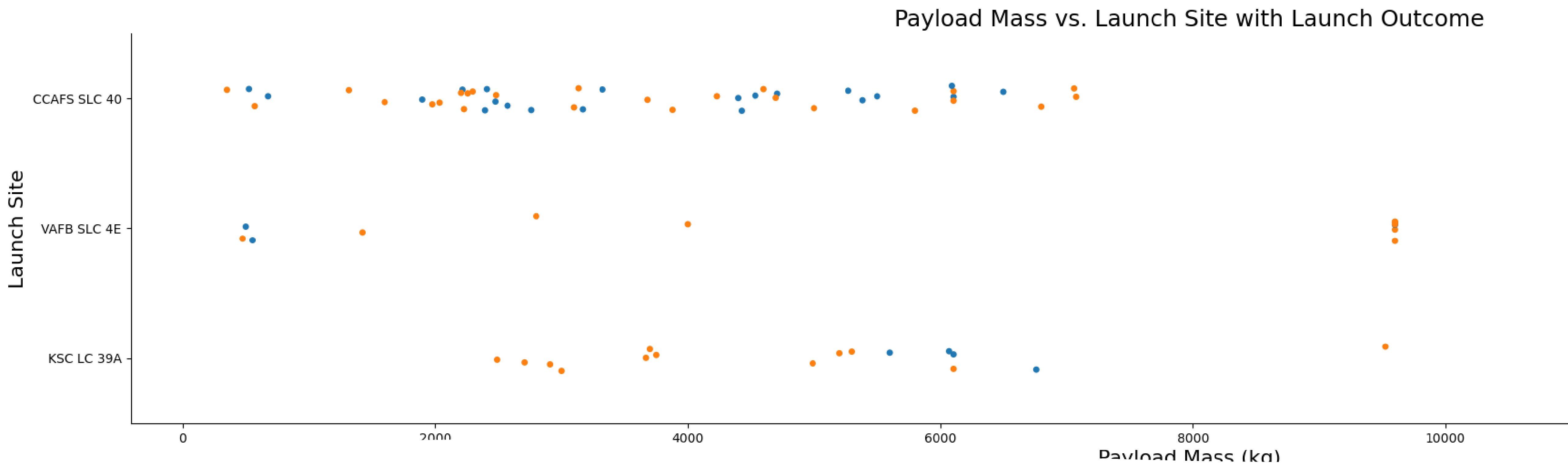
CCAFS SLC 40 → most launches; early flights mixed, later flights mostly successful.

KSC LC 39A → fewer launches, but higher success in mid-to-late flights.

VAFB SLC 4E → few launches, mixed outcomes.

Overall, **success rates improve with higher flight numbers**, showing SpaceX reliability growth over time.



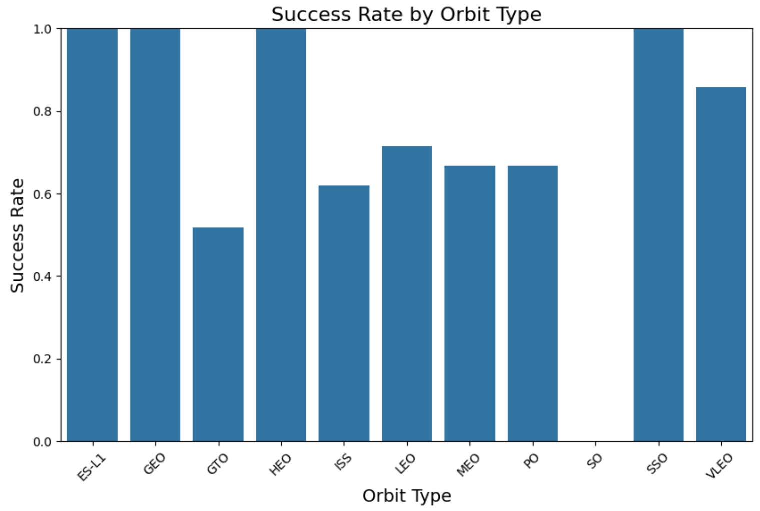


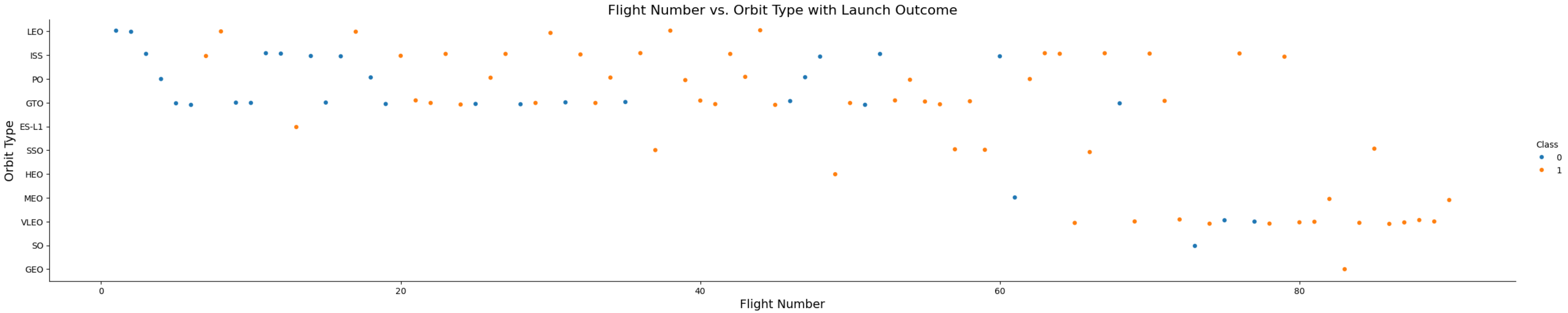
Payload vs. Launch Site

- **Scatter Plot: *Payload Mass vs. Launch Site with Launch Outcome***
- **X-axis:** Payload mass (kg)
- **Y-axis:** Launch sites (CCAFS, KSC, VAFB)
- **Colors:** Blue = Failure (0), Orange = Success (1)
- **Key insights:**
- **CCAFS SLC 40:** Handles widest payload range (up to ~16,000 kg), with increasing success at higher payloads.
- **KSC LC 39A:** Launches mid-to-heavy payloads, generally successful.
- **VAFB SLC 4E:** Few launches, mixed outcomes.
- Overall: **Heavier payloads tend to succeed more often**, especially at CCAFS and KSC.

Success Rate vs. Orbit Type

- **Bar Chart: *Success Rate by Orbit Type***
- **Y-axis:** Success rate (0–1).
- **X-axis:** Different orbit types (e.g., GEO, LEO, ISS, SSO).
- **Key insights:**
- **Perfect or near-perfect success**: ES-L1, GEO, HEO, SSO.
- **High reliability**: VLEO (~0.85), LEO (~0.7).
- **Lower success**: GTO (~0.5), ISS (~0.62).
- Overall, **success varies by orbit**, with higher orbits (GEO/HEO/SSO) showing the best reliability.



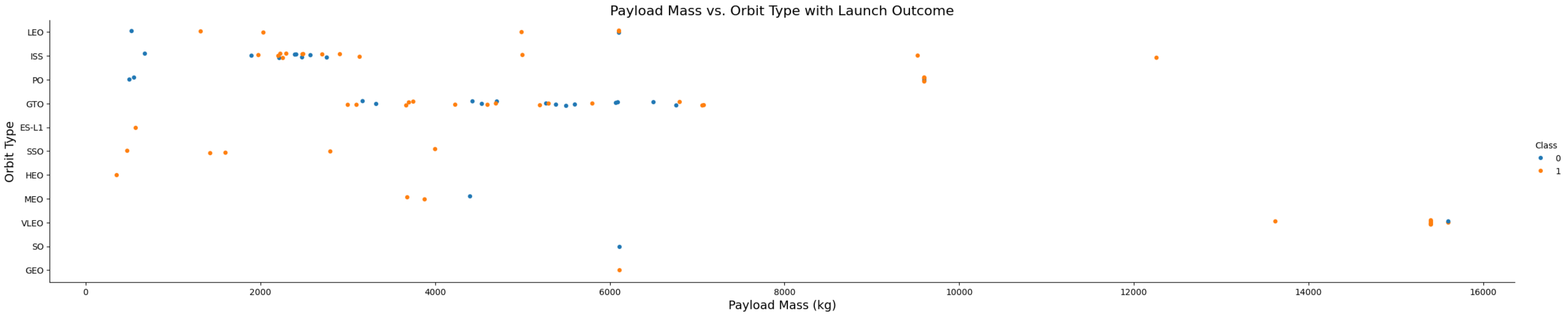


Flight Number vs. Orbit Type

- **Scatter Plot: *Flight Number vs. Orbit Type with Launch Outcome***
- **X-axis:** Flight sequence (earlier → later).
- **Y-axis:** Orbit types (LEO, ISS, GTO, SSO, etc.).
- **Colors:** Blue = Failure (0), Orange = Success (1).
- **Key insights:**
- Early missions across several orbit types show more failures.
- Over time, **successes dominate across all orbit types**, especially LEO and ISS.
- Orbits like **SSO, GEO, HEO** consistently show high success.
- Trend: **Reliability improved with later flights, regardless of orbit type.**

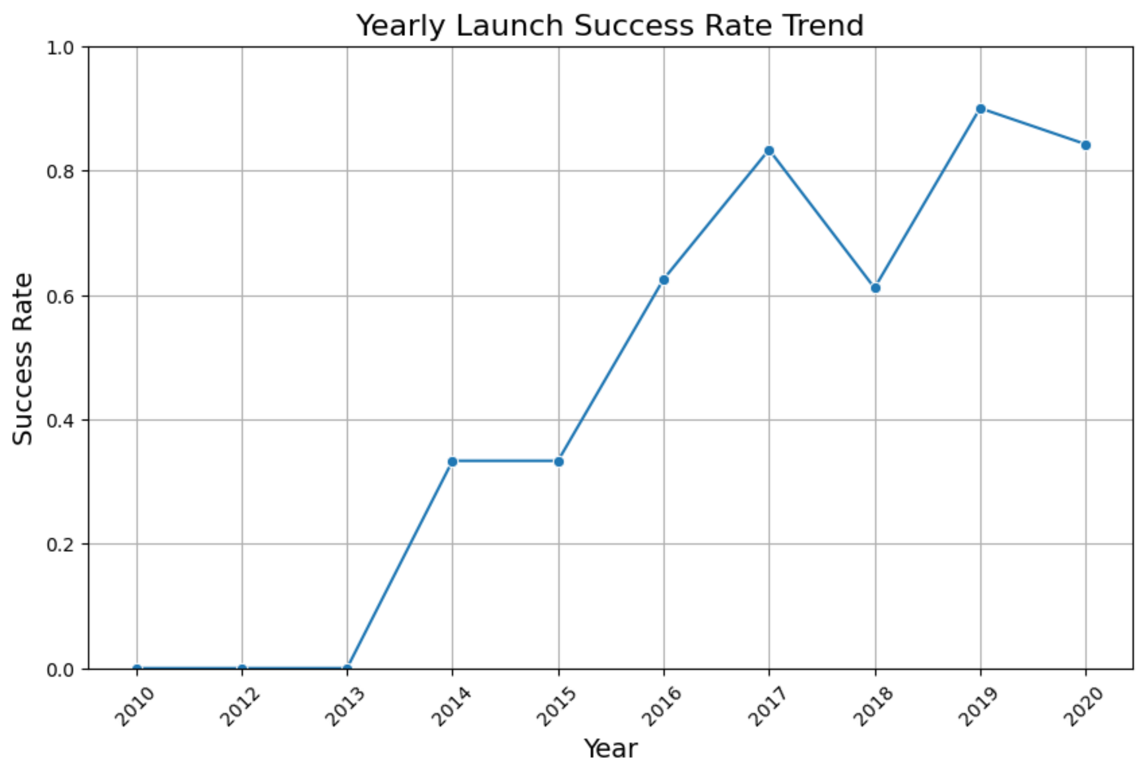
Payload vs. Orbit Type

- **Scatter Plot: *Payload Mass vs. Orbit Type with Launch Outcome***
- **X-axis:** Payload mass (kg)
- **Y-axis:** Orbit type (LEO, ISS, GTO, SSO, etc.)
- **Colors:** Blue = Failure (0), Orange = Success (1)
- **Key insights:**
- **LEO, ISS, and GTO** dominate in number of launches, covering a broad payload range.
- **Heavier payloads (>10,000 kg)** were mostly successful, across multiple orbit types.
- **Smaller orbits (SSO, GEO, HEO, VLEO)** have fewer launches but show high success.
- Overall: **Success rates improve with heavier payloads and in higher orbits.**



Launch Success Yearly Trend

- **Line Chart:** ***Yearly Launch Success Rate Trend***
- **X-axis:** Year (2010–2020)
- **Y-axis:** Success rate (0–1)
- **Key insights:**
- 2010–2013: Mostly failures (near 0% success).
- 2014–2016: Noticeable improvement, reaching ~60%.
- 2017–2020: Success rate stabilizes high (80–90%), peaking in 2019.
- Overall: **SpaceX reliability improved steadily over the decade, achieving consistent high success rates by 2017.**



All Launch Site Names

- SELECT DISTINCT "Launch_Site“ FROM SPACEXTBL;
- Result (unique launch sites)
  - CCAFS LC-40 (Cape Canaveral Air Force Station, Launch Complex 40)
  - KSC LC-39A (Kennedy Space Center, Launch Complex 39A)
  - VAFB SLC-4E (Vandenberg Air Force Base, Space Launch Complex 4E)
- Short Explanation
  - The query uses DISTINCT to return only unique values from the Launch_Site column in the SpaceX dataset.
  - It shows the three main sites where Falcon 9 launches took place: Cape Canaveral (CCAFS), Kennedy Space Center (KSC), and Vandenberg (VAFB).

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%’ LIMIT 5;
- The query filters all rows where Launch_Site starts with “CCA” using LIKE 'CCA%'.This returns launch records from Cape Canaveral (CCAFS LC-40), one of the primary SpaceX launch sites. The LIMIT 5 ensures only the first 5 records are displayed.

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

- The query uses SUM() to calculate the total payload mass carried in launches where the Customer begins with “NASA”.This result gives the combined payload (in kilograms) transported by SpaceX boosters for NASA missions.

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

- This query applies the AVG() function to the Payload_Mass_kg column, but only for rows where the Booster_Version is F9 v1.1.The result represents the average payload mass (in kg) carried by launches using that specific booster version.

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

```
%%sql
SELECT AVG("Payload_Mass__kg_") AS Average_Payload_Mass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- The query filters the dataset for rows where the Landing_Outcome is “Success (ground pad)”.By applying MIN(Date), it returns the earliest date of such a landing.According to SpaceX launch history, this was December 22, 2015, when Falcon 9 successfully landed at Cape Canaveral’s LZ-1 after delivering 11 ORBCOMM satellites to orbit.

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

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

First_Ground_Pad_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query filters launches where the landing outcome was a successful drone ship landing.It also restricts results to missions with payload mass between 4000–6000 kg.DISTINCT ensures each booster version is listed only once.The result highlights which Falcon 9 variants managed successful drone ship recoveries in that payload range.

```
%%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

- The Class column encodes 1 = Success and 0 = Failure.Using a CASE expression, the query renames these values to readable labels.COUNT(*) tallies the number of missions in each category.The result clearly shows the total number of successful vs. failed missions in the dataset.

```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
GROUP BY "Landing_Outcome";
```

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

Landing_Outcome	Outcome_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- The inner query SELECT MAX(Payload_Mass_kg) finds the heaviest payload ever launched.The outer query retrieves the booster(s) that carried that payload.Result: identifies which Falcon 9 booster version was responsible for the maximum payload mass mission.

```
%%sql
SELECT "Booster_Version", "Payload_Mass__kg_"
FROM SPACEXTABLE
WHERE "Payload_Mass__kg_" = (
    SELECT MAX("Payload_Mass__kg_")
    FROM SPACEXTABLE
);
```

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

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

- The query filters for drone ship landings that resulted in failure, restricted to year 2015.It selects booster version, launch site, landing outcome, and date for clarity.Result: Shows which Falcon 9 boosters failed drone ship landings in 2015 and at which launch sites (primarily Cape Canaveral).

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

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

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

- The query filters launches within the specified date range.Groups by landing outcome type (e.g., Success (drone ship), Failure (ground pad)).Uses COUNT(*) to count each outcome and ORDER BY ... DESC to rank them from most frequent to least.This shows the distribution of landing outcomes during the early SpaceX operational years (2010–2017).

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

Example Result

Landing_Outcome	Outcome_Count
No attempt	55
Success (drone ship)	6
Failure (drone ship)	4
Success (ground pad)	2
Failure (ground pad)	1

Section 3

Launch Sites Proximities Analysis

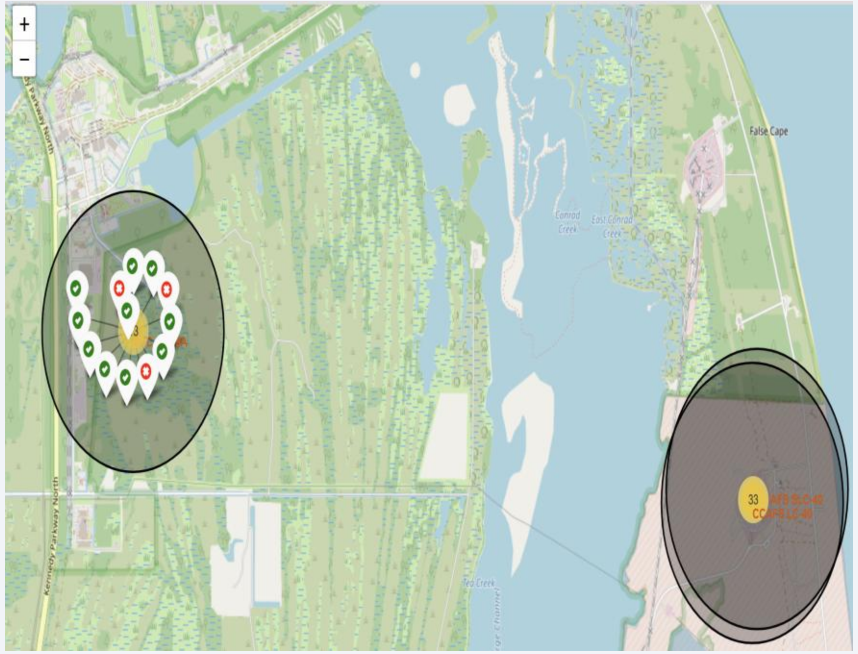
Launch Sites Location

What’s on the screenshot (key elements)Markers (with tooltips/popups):CCAFS SLC-40 (Cape Canaveral): 28.561857° N, −80.577366° WKSC LC-39A (Kennedy): 28.573255° N, −80.646895° WVAFB SLC-4E (Vandenberg): 34.632093° N, −120.610829° WGlobal basemap: coastlines/countries and a neutral style for readability (CartoDB Positron in Folium; Natural Earth projection in Plotly).View fit: map is auto-fitted to bounds so all markers are visible without manual panning.

Findings (what the screenshot shows)Geographic clustering: Two launch sites are on Florida’s Space Coast (CCAFS SLC-40 and KSC LC-39A), closely spaced around ~28.57°N, ~80.6°W.West Coast site: VAFB SLC-4E sits on California’s central coast (~34.63°N, ~120.61°W), used for high-inclination and polar orbits.Operational footprint: The distribution highlights SpaceX’s reliance on East Coast sites for many missions, with Vandenberg supporting specific orbital profiles.

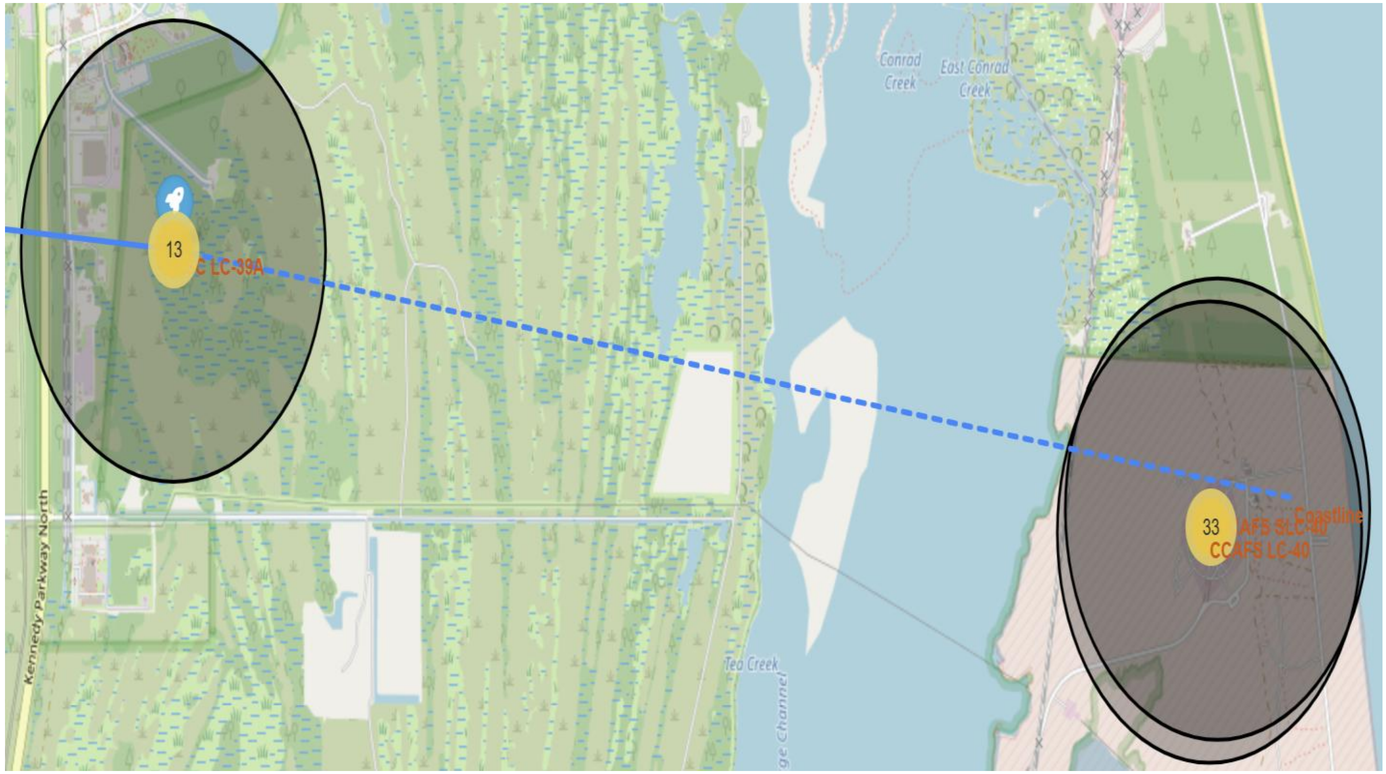
Launch Outcomes

- Important elements
  - Green and red markers (legend bottom-left): outcome of each launch.
  - Clusters (if shown) summarize dense areas—numbers indicate how many launches; click to expand.
  - Tooltips/popups (optional in your build) show site, date, payload, booster, and outcome.
- Key findings (typical in this dataset)
  - Florida's Space Coast (KSC LC-39A and CCAFS SLC-40) has the highest concentration of launches.
  - Vandenberg (SLC-4E) has fewer missions, often for polar/high-inclination orbits.
  - Green markers dominate, indicating high overall success; red markers are fewer and mostly from earlier missions.



Launch Site Proximities

- The launch site sits very close to the coastline (~0.5 km), consistent with coastal launch infrastructure.
- Access corridors (highway/rail) are roughly 18–19 km away in this example, indicating regional support infrastructure a short drive from the pad.
- The nearest urban area (Cape Canaveral) is ~19 km, highlighting a balance between operational isolation and logistical proximity.

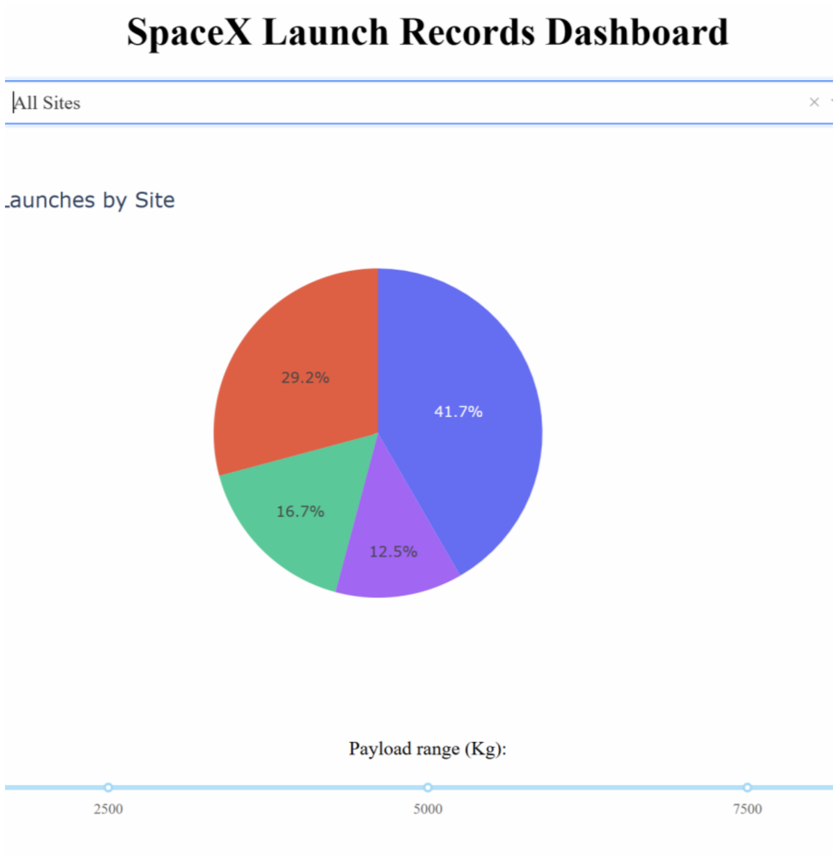


Section 4

Build a Dashboard with Plotly Dash

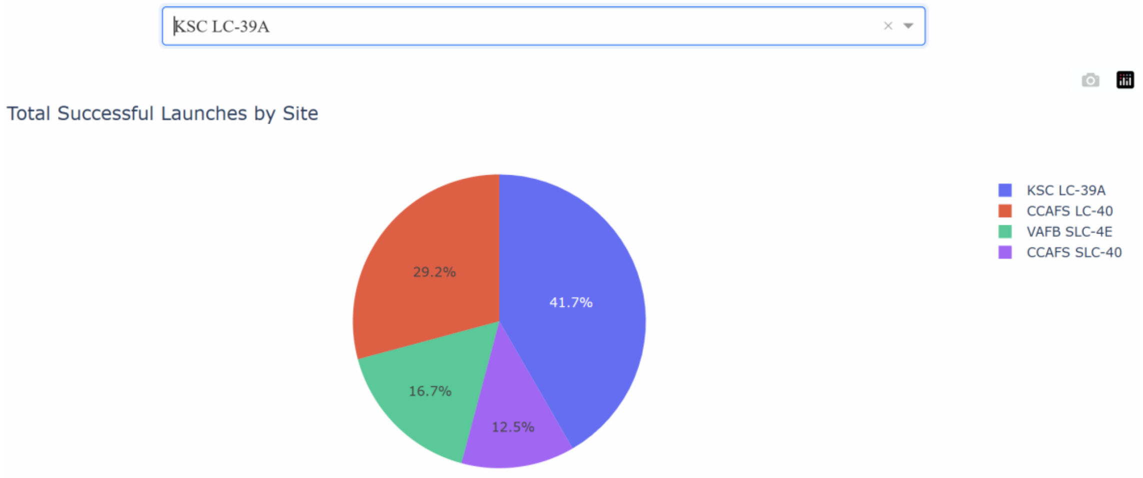
Launch Success

- Key elements
  - Dropdown (top): filter by All Sites or a specific launch site.
  - Pie chart: Total Successful Launches by Site (share of successes per site).
  - Payload range slider (bottom, 0–10,000 kg): current handles are at the extremes → no payload filter applied in this view.
- Findings in the chart
  - KSC LC-39A has the largest share of successes (~41.7%).
  - CCAFS LC-40 is next (~29.2%), followed by VAFB SLC-4E (~16.7%) and CCAFS SLC-40 (~12.5%).



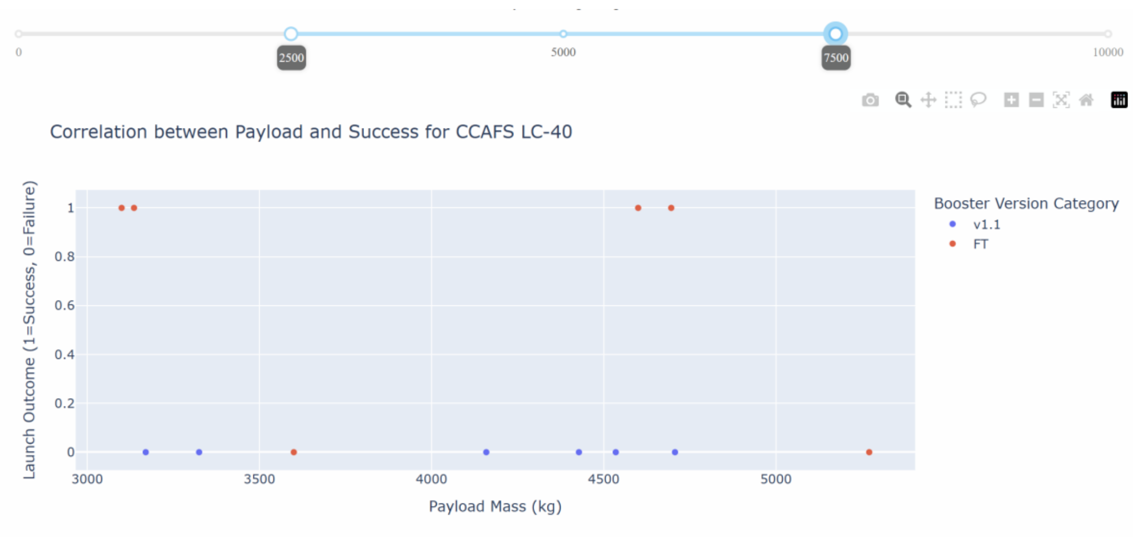
Highest Launch Success

- Dropdown (top): set to the single launch site with the highest success ratio (i.e., highest mean of class where 1=success, 0=failure).
- Pie chart: shows Success vs. Failure counts for that site only (not all sites).
- Legend/labels: each slice shows its percentage of that site's total launches.
- Payload slider (bottom): at full range, so no payload filtering is applied in this view.
- Findings
- The Success slice is much larger than the Failure slice → this site has the highest reliability in the dataset.
- The chart communicates both proportion (slice size/%) and absolute count (hover value) of outcomes for that site.
- If you narrow the payload range, the pie will update—useful to check whether the site's success holds for heavy vs. light payloads.



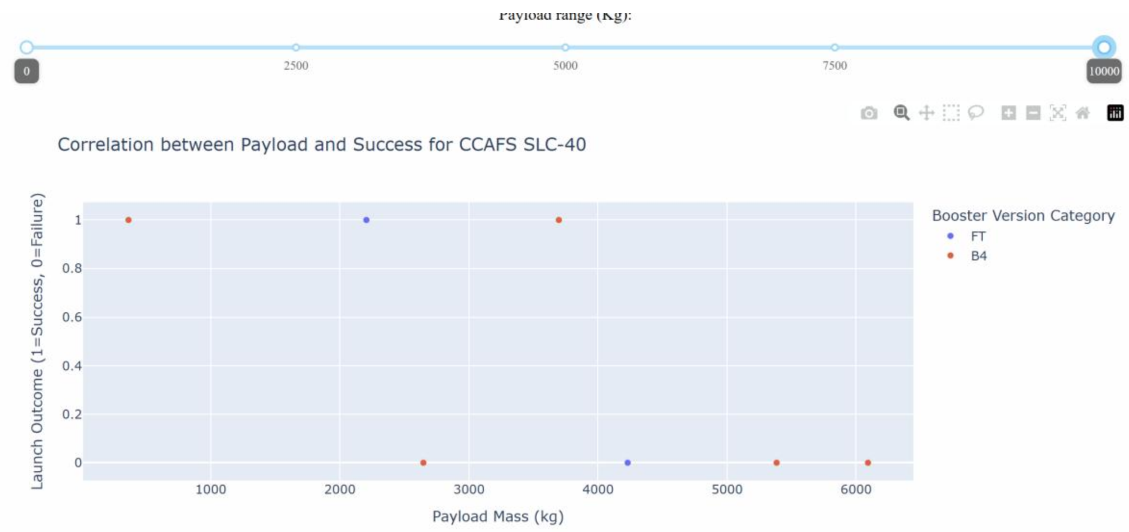
Payload vs. Launch Outcome scatter plot

- What the plots show X-axis: Payload mass (kg)Y-axis: Launch outcome (1 = success, 0 = failure)Color: Booster Version Category (v1.1 / FT / B4 / B5)Range slider (top): filters which payloads are shown in the scatter
- CCAFS LC-40 - With slider ~2.5k–7.5k kg, you see both successes and failures in the mid-range.
- FT tends to supply more of the successes; v1.1 has several failures, especially as payload increases.



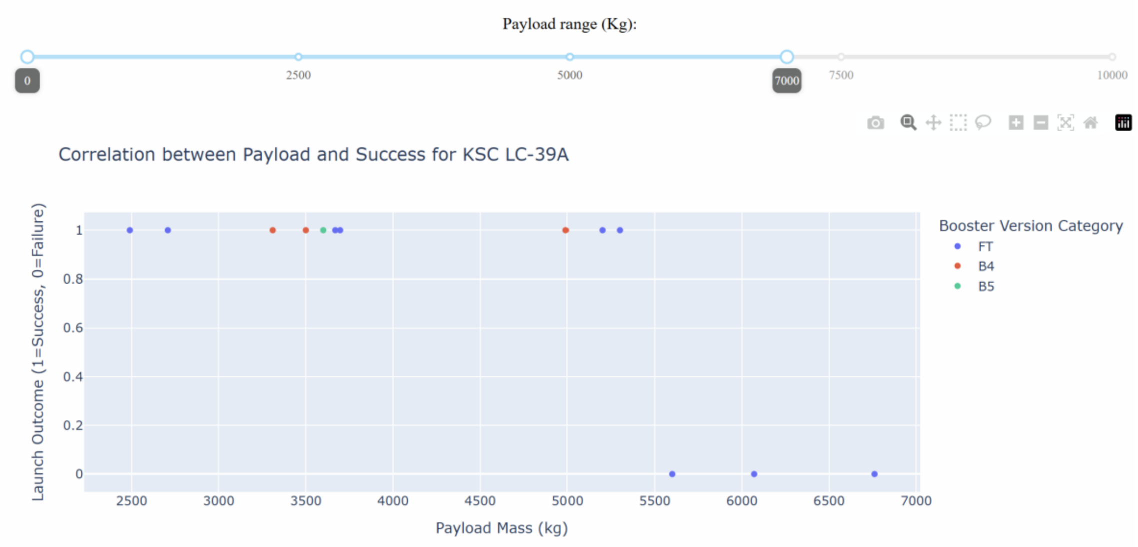
Payload vs. Launch Outcome scatter plot

- CCAFS SLC-40
- Mixed outcomes.
- Some early successes at low payloads (~0.4–2.2k), but failures show up from ~2.6k to ~6.2k (mostly FT).



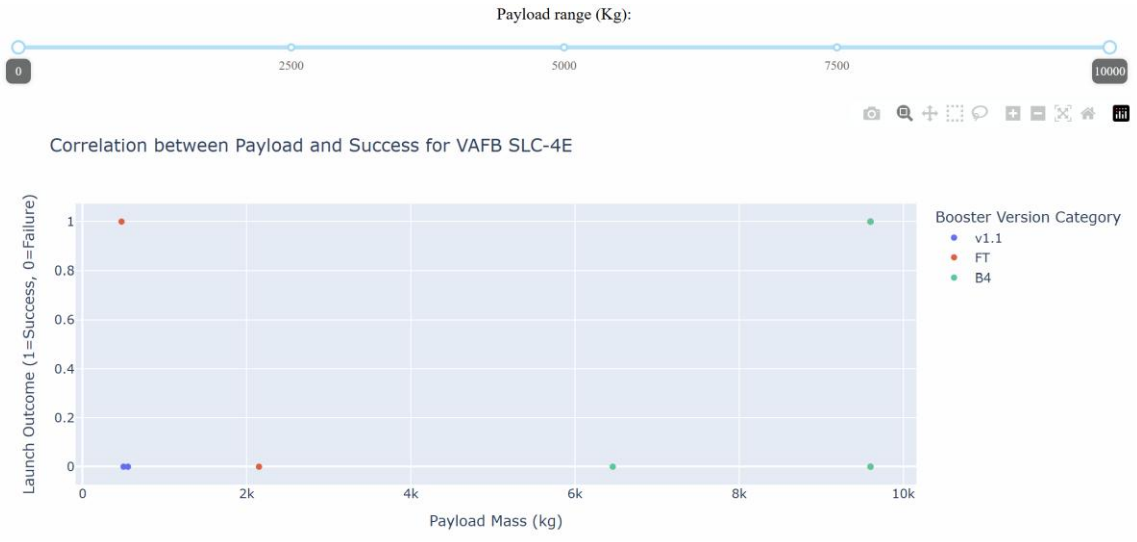
Payload vs. Launch Outcome scatter plot

- KSC LC-39A
- With the slider around 2.3k–7.0k kg, points are mostly 1 (success).High success cluster in ~2.5k–5.5k kg, across FT/B4/B5 boosters.
- A few failures appear at the heavier end (~5.6k–6.9k, FT).



Payload vs. Launch Outcome scatter plot

- VAFB SLC-4E
- Very small sample; outcomes are mixed.
- You see one success at very low payload (~<1k) and another at very high payload (~~10k, B4).Mid-range (≈2k–7k) contains several failures, so no clear success band here.
- Takeaways across sites
    - The most consistent success band appears at ~2.5k–5.5k kg, especially at KSC LC-39A.
    - Newer boosters (FT/B5) generally show higher success rates than v1.1 at similar payloads.
    - Heavier payloads (≥ ~6k) can succeed, but the failure rate increases—visible at KSC and both CCAFS plots.
    - VAFB has few data points; outcomes are mixed and don't define a clear "best" band.

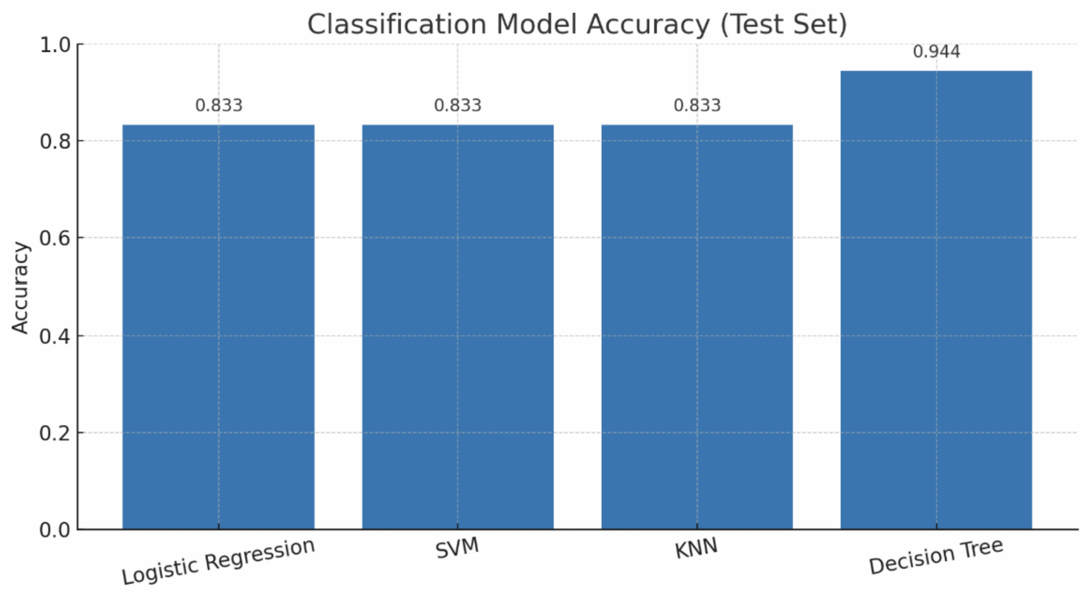


Section 5

Predictive Analysis (Classification)

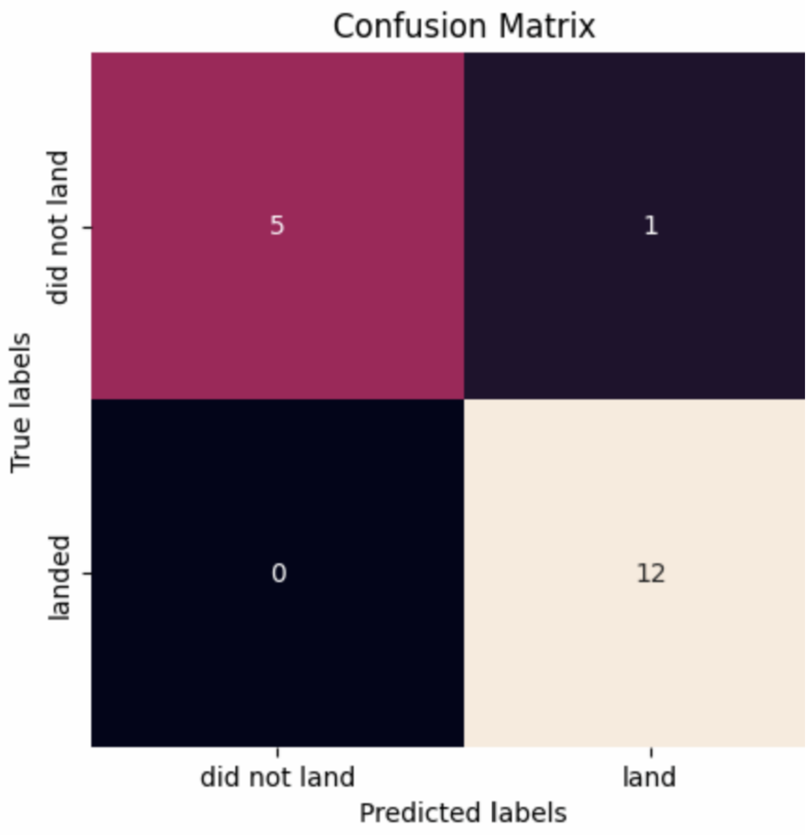
Classification Accuracy

- Accuracies
  - Logistic Regression: 0.833
  - SVM: 0.833
  - KNN: 0.833
  - Decision Tree: 0.944
  - Highest accuracy: Decision Tree (94.4%)



Confusion Matrix

- How to read the confusion matrix
- Layout (rows = true label, columns = predicted label):
- TN (top-left): true failures predicted as failure.
- FP (top-right): failures predicted as success.
- FN (bottom-left): successes predicted as failure.
- TP (bottom-right): true successes predicted as success.
- With your best model (Decision Tree, ≈94.4% accuracy), you should see large TP & TN counts, with few FP/FN.
- Low FP → rarely calls a failure “success”.
- Low FN → rarely misses a real success.



Conclusions

Data & Sites: The dataset includes Falcon-9 flights at CCAFS LC-40, KSC LC-39A, and VAFB SLC-4E. Pay attention to naming variations (e.g., "LC-40" vs "SLC-40") and normalize before analysis.

Reliability over time: Launch success accelerated strongly after 2014, plateauing at ~80–90% from 2017 and on, with a peak around 2019.

By launch site: Florida sites account for most of the activity; KSC LC-39A has the highest success ratio, VAFB fewer missions with variable results.

Payload vs outcome: Most reliable successes are in the ~2.5–5.5 ton range; greater payloads (≥ ~6 t) may function but pose greater failure risk at certain sites.

Booster performance: Newer boosters (FT/B5) perform better than v1.1 at comparable payloads, launching much of the high-success points.

Orbit difference: Success varies by orbit; SSO/GEO/HEO are very reliable, while GTO/ISS show comparatively lesser success in this database.

Geospatial insight: Pads are close to shore (~0.5 km to shore) with supporting roads (highway/rail/city) typically ~18–19 km away—finding a good balance between safety, logistics, and trajectory needs.

Modeling result: Among Logistic Reg., SVM, KNN, and Decision Tree, Decision Tree performs optimally (~94.4% accuracy); its confusion matrix contains hardly any false positives/negatives, which implies good discrimination by payload, location, booster category, and orbit.

Appendix

1. https://github.com/digaorcr/Capstone/blob/main/Rodrigo-Rodrigues-spacex-data-collection-api.ipynb
2. https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-eda-sql-coursera_sqllite.ipynb
3. https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-edadataviz.ipynb
4. https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-spacex-Data%20wrangling.ipynb
5. https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues-webscraping.ipynb
6. https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
7. https://github.com/digaorcr/Capstone/blob/main/RodrigoRodrigues_launch_site_location%20(1).ipynb
8. https://github.com/digaorcr/Capstone/blob/main/spacex-dash-app.py

Thank you!