Predicting the Future of Space Launches: A Data Science Approach with SpaceX

Saud Almutairi

May 31, 2025

Outline

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

Executive Summary

- Collected launch data from public SpaceX APIs, CSVs, and Wikipedia.
- Cleaned and merged datasets using Pandas.
- Performed EDA using Matplotlib, Seaborn, and SQL queries.
- Visualized launches and outcomes with interactive tools (Folium, Plotly Dash).
- Built and optimized ML models (Logistic Regression, SVM, KNN, Decision Trees).
- Selected best model based on cross-validation and test accuracy.

Introduction

SpaceX reduces launch costs by reusing the Falcon 9's first stage.
This project simulates a SpaceX competitor, **SpaceY**, using public data to:

- Analyze factors impacting launch success
- Build models to predict first stage landing success
- Visualize and interpret outcomes interactively

Section 1: Methodology

Data Collection, Wrangling, and Analytical Approach

Methodology

- **Data Collection:** Gathered launch data via SpaceX REST API and scraped historical launch records from Wikipedia.
- **Data Wrangling:** Cleaned, filtered, and merged datasets. Extracted relevant features like payload mass, launch site, orbit, and landing outcome.
- **Exploratory Data Analysis (EDA):** Used Pandas, Seaborn, and Matplotlib to analyze trends and correlations among launch success and influencing factors.
- **SQL Analytics:** Imported cleaned data into SQLite; executed aggregate queries to identify patterns across launch sites, orbits, and booster types.
- **Interactive Visualization:** Built interactive maps using Folium and a real-time analytics dashboard using Plotly Dash to explore launch outcomes.
- **Predictive Modeling:** Standardized features, split data, and trained classification models (Logistic Regression, SVM, Decision Tree, KNN) with hyperparameter tuning via GridSearchCV.

Data Collection Overview

Sources Used:

- **CSV:** Launch metadata (spacex_launch_dash.csv)
- **API:** JSON from SpaceX REST API (`/launches`, `/rockets`, etc.)
- **Web Scraping:** Wikipedia Falcon 9 mission history table

Data Collection – SpaceX API

- Fetched data using `requests` from SpaceX API endpoints
- Parsed and normalized nested fields using `pandas.json_normalize`
- Extracted: rocket ID, payload, core stats, landing outcome

Flowchart:

- API GET → JSON → Flatten → Extract → Merge

GitHub: github.com/your-username/spacex-api-call

Data Collection – Scraping

- Scraped launch table from Wikipedia using `BeautifulSoup`
- Parsed launch date, booster version, orbit, payload mass
- Cleaned Unicode symbols using `unicodedata` and regex

Flowchart: Wikipedia $\rightarrow$ HTML $\rightarrow$ parse & clean $\rightarrow$ pandas DF **GitHub:** github.com/your-username/spacex-scraping

Data Wrangling

Key Processing Steps:

- Merged datasets from API, Wikipedia, and CSV
- Handled missing values (mean imputation, retained None)
- Flattened nested JSON structures
- Engineered features (Class, Year, Booster version)
- Encoded categorical columns using `pd.get_dummies()`

GitHub: github.com/your-username/data-wrangling

EDA with Data Visualization

Exploratory Visualizations:

- **Flight Number vs Launch Site:** Categorical scatter plot to detect success patterns across launch sites over time.
- **Payload Mass vs Launch Site:** Scatter to explore correlation between payload size and launch success.
- **Success Rate by Orbit Type:** Bar chart to identify which orbit types yield higher success.
- **Payload Mass vs Orbit:** Evaluates if specific orbits demand heavier payloads or affect success.
- **Yearly Success Trend:** Line plot of average launch success rate by year.

GitHub Notebook:
github.com/your-username/eda-visualization-notebook

SQL Analysis Highlights:

- Counted number of launches per site to identify high-frequency launch locations
- Calculated total and average payload mass for NASA (CRS) missions
- Found earliest successful ground landing date using MIN and WHERE
- Queried boosters with successful drone ship landings and payloads between 4000-6000 kg
- Aggregated successful vs. failed mission outcomes
- Extracted monthly failure counts on drone ship in 2015
- Ranked landing outcomes by frequency within a specific date range

GitHub Notebook:
github.com/your-username/eda-sql-notebook

Build an Interactive Map with Folium

Map Components and Purpose:

- **Circles:** Visualized the geographic location of each launch site on the map.
- **Markers:** Annotated launch sites with names using labeled map markers.
- **Popups:** Added popup text to show site-specific data on hover/click.
- **Marker Clusters:** Grouped overlapping success/failure markers to reduce clutter.
- **Launch Outcome Indicators:** Used green (success) and red (failure) markers for outcomes.
- **Distance Lines:** Drew lines between launch sites and key proximities (e.g., coastlines, facilities).

GitHub Notebook:
github.com/your-username/interactive-map-folium

Predictive Analysis (Classification)

- Built multiple classification models including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN).
- Standardized features using `StandardScaler` and split the data into training and test sets.
- Performed hyperparameter tuning using `GridSearchCV` with 10-fold cross-validation.
- Evaluated each model's performance based on validation and test accuracy.
- Found that the best performing model was **SVM** with the highest accuracy on the test data.

GitHub Notebook:

`https://github.com/yourusername/spacex-classification-analysis`

Results

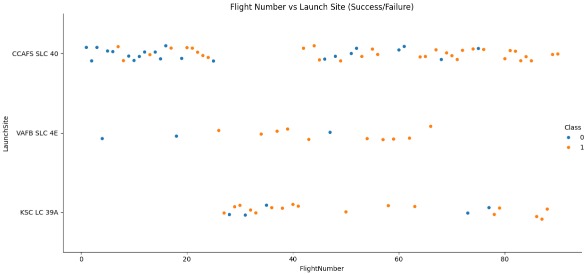
- **Exploratory Data Analysis Results:** Identified launch sites with highest success counts, correlated payload mass and orbit types with success rate using seaborn and matplotlib visualizations.
- **Interactive Analytics Demo:** Developed a dynamic dashboard using Plotly Dash. Users can filter launches by site and payload range. Success distribution shown via pie charts and scatter plots.
- **Predictive Analysis Results:** Trained multiple classification models (Logistic Regression, SVM, Decision Tree, KNN). Achieved best accuracy using Support Vector Machines (SVM) with optimized hyperparameters via GridSearchCV.

Section 2: Insights Drawn from EDA

Uncovering patterns in launches, orbits, and payloads

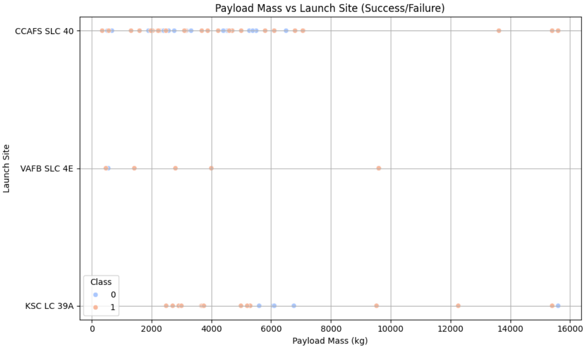
Flight Number vs. Launch Site

- Shows launches by site and flight number.
- Orange = success, Blue = failure.
- Success improves with flight experience.
- CCAFS SLC 40 had the most launches.
- KSC LC 39A shows strong success rate.



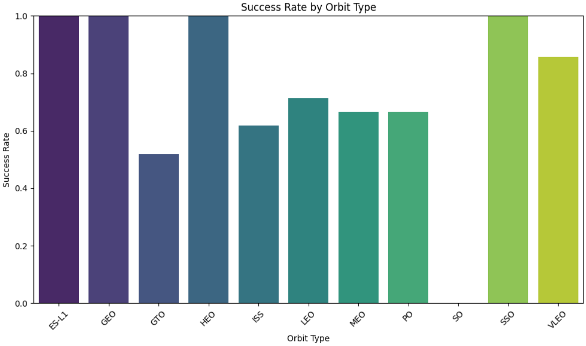
Payload Mass vs. Launch Site

- Shows how payload mass varies across sites.
- Success more likely below 10,000 kg.
- CCAFS SLC 40 supports the widest payload range.
- KSC LC 39A has consistent success at medium payloads.



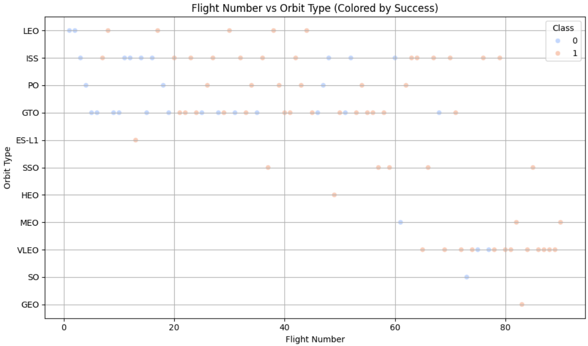
Success Rate by Orbit Type

- Displays launch success rates across different orbit types.
- ES-L1, GEO, HEO, and SSO achieved 100
- GTO had the lowest success rate among all orbits.
- High reliability observed in low and sun-synchronous orbits.



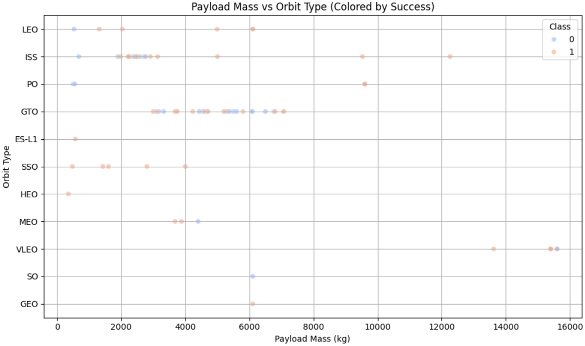
Flight Number vs. Orbit Type

- Shows the relationship between flight number and orbit type.
- Later flights (higher numbers) show more success across orbits.
- Higher success concentration in VLEO and SSO orbits.
- GTO and PO have scattered success rates.



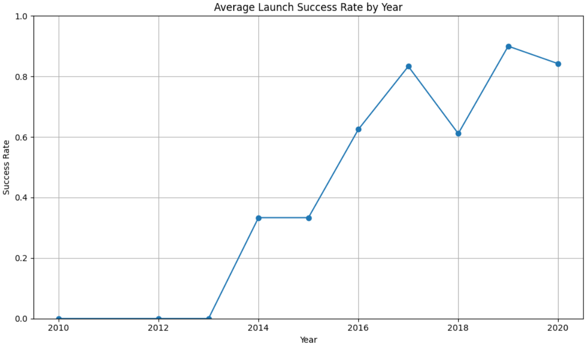
Payload vs. Orbit Type

- Visualizes how payload mass varies across different orbit types.
- GTO and VLEO support heavier payloads.
- Lighter payloads dominate ISS, LEO, and PO missions.
- Most success seen in moderate payload ranges.



Launch Success Yearly Trend

- Line chart shows yearly average launch success.
- Steady improvement from 2014 to 2020.
- Peak success observed in 2019.
- Indicates growing reliability of SpaceX launches.



All Launch Site Names

- Queried unique launch site names using SQL.
- Identified 3 primary SpaceX launch sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
- These sites are used to analyze launch success trends.

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

Launch Sites Beginning with 'CCA'

- SQL query was used with the LIKE 'CCA%' clause.
- Retrieved 5 launch records from Cape Canaveral Air Force Station (CCAFS).
- Launches span multiple years and missions.
- Used for further analysis of payloads, customers, and outcomes.

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 by NASA (CRS)

- SQL query used: `SUM(PAYLOAD_MASS__KG__)` filtered by customer `NASA (CRS)`.
- Retrieved the total payload mass carried by NASA's CRS missions.
- Result: **619,967 kg**.

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS)" FROM SPACEXTABLE;
```

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

Total Payload Mass by NASA (CRS)
619967

Average Payload Mass by F9 v1.1

- Used SQL to calculate the average payload mass.
- Filtered the dataset for booster version `F9 v1.1`.
- Result: **2,534.67 kg**.

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AvgPayloadMass_F9v1_1
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

AvgPayloadMass_F9v1_1
2534.6666666666665

First Successful Ground Landing Date

- Queried the dataset using the `MIN()` function.
- Filtered where `Landing_Outcome = 'Success (ground pad)'`.
- First successful landing on a ground pad occurred on:

2015-12-22

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

Hint:Use min function

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

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

FirstGroundPadLandingDate
2015-12-22

Drone Ship Landings with Payload 4000–6000 kg

- Used SQL to filter boosters with:
  - `Landing_Outcome = 'Success (drone ship)'`
  - `Payload_Mass_KG_` between 4000 and 6000
- Identified booster versions with successful landings under those criteria.
- Results include boosters like `F9 FT B1022`, `F9 FT B1026`, etc.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Mission Outcome Counts

- Queried `Mission_Outcome` to summarize launch results.
- Majority of missions were successful:
  - `Success: 98`
  - `Success (payload status unclear): 1`
- One failure recorded: `Failure (in flight)`.
- Indicates strong mission reliability.

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

Boosters Carrying Maximum Payload

- Identified boosters with the highest payload mass.
- Applied SQL query filtering maximum Payload Mass (kg).
- Falcon 9 Block 5 boosters dominated heavy-lift launches.
- Indicates B5 version's performance reliability.

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

- Filtered 2015 data for `Landing_Outcome = 'Failure (drone ship)'`.
- Identified booster versions and launch sites involved.
- Both failures occurred at **CCAFS LC-40** with boosters **B1012** and **B1015**.
- Insights help evaluate performance trends of booster versions over time.

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

Ranked Landing Outcomes (2010–2017)

- Extracted data between **June 2010** and **March 2017**.
- Ranked outcomes by count in **descending order**.
- Top outcome: **No attempt** (10 occurrences).
- Followed by **Success/Failure (drone ship)**, and **Success (ground pad)**.
- Helps assess landing trends and strategy over early missions.

Landing_Outcome	OutcomeCount
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

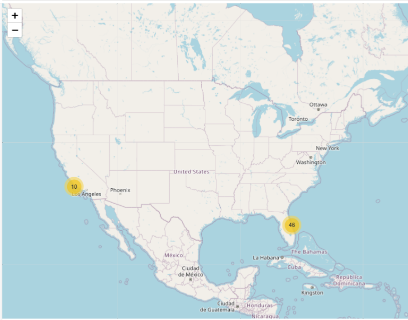
Exploring infrastructure and terrain around launch locations

Launch Site Proximities on Global Map

- All active launch sites plotted using Folium and OpenStreetMap.
- Sites are located along the U.S. east coast and west coast.
- Proximity to oceans enables safe first-stage landings and drone ship recovery.
- Geographical spread supports diverse orbital insertion paths.

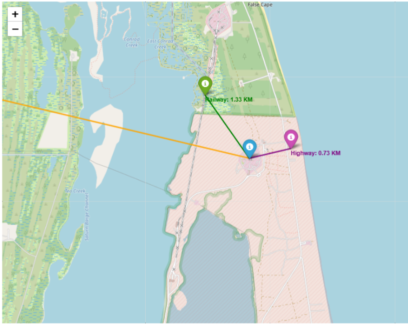
Launch Clusters and Density

- Folium map shows clustered launch data.
- High-density clusters around Florida and California.
- Helps visualize regional launch activity.
- Useful to prioritize infrastructure planning near active zones.



Launch Site Proximity to Infrastructure

- Mapped launch site distances to nearby transport and coast.
- Highway is only 0.73 KM away — excellent accessibility.
- Railway is 1.33 KM away — useful for transporting heavy equipment.
- Proximity to coastline supports marine recovery and transport.

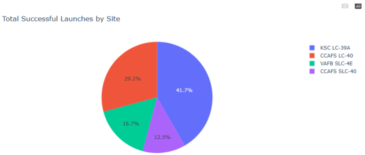


Section 4: Build a Dashboard with Plotly Dash

Interactive tools for visual analytics and exploration

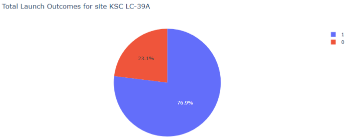
Launch Success by Site

- This pie chart visualizes the proportion of successful launches from each site.
- **KSC LC-39A** accounts for the highest share of successes (41.7%).
- **CCAFS LC-40** follows with 29.2%, while **VAFB SLC-4E** and **CCAFS SLC-40** have fewer successful missions.
- Indicates operational concentration at key launch sites.



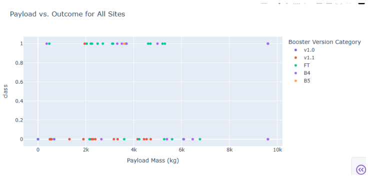
Top Performing Launch Site

- This pie chart presents the success and failure rates at **KSC LC-39A**.
- The site achieved a **76.9% success rate**, the highest among all SpaceX launch sites.
- Indicates strong performance consistency and favorable launch conditions.
- Insightful for selecting optimal sites for future missions.



Payload vs. Launch Outcome Across Sites

- This scatter plot displays payload mass against launch outcomes across all sites.
- Dots are color-coded by booster version category: v1.0, v1.1, FT, B4, and B5.
- Majority of successful launches occur in the 2000–4000 kg range.
- **FT and B5** boosters appear to be more reliable across payloads.

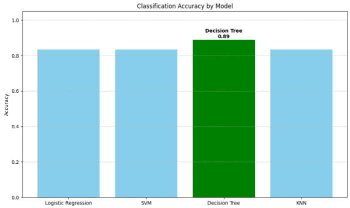


Section 5

Predictive Analysis (Classification)

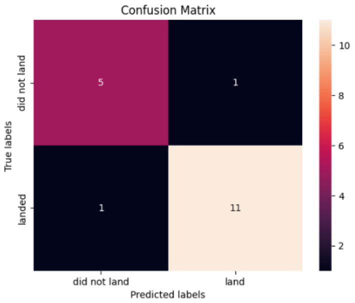
Classification Accuracy by Model

- Evaluated four classification models:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- Models were assessed using test accuracy.
- **Decision Tree** performed best with **89%** accuracy.



Confusion Matrix of Best Model (Decision Tree)

- The confusion matrix evaluates the Decision Tree model.
- **True Positives (landed correctly)**: 11
- **True Negatives (did not land correctly)**: 5
- **False Positives**: 1 (predicted landed, but didn't)
- **False Negatives**: 1 (predicted did not land, but did)
- Indicates strong classification performance with minimal misclassifications.



Conclusions

- **Point 1:** SpaceX launch data enables accurate modeling of mission outcomes using classification techniques.
- **Point 2:** Decision Tree performed best among tested models, achieving highest classification accuracy.
- **Point 3:** EDA revealed strong influence of launch site, orbit type, and payload on mission success.
- **Point 4:** Interactive tools like Folium and Plotly Dash enhanced data interpretation and communication.
- **Point 5:** Proximity to infrastructure (highways, railways) was effectively analyzed using geospatial tools.

Appendix

- GitHub Repository with all notebooks and resources: github.com/YourUsername/spacex-data-science-capstone
- Tools Used: Python, Pandas, Matplotlib, Seaborn, Plotly Dash, Folium, SQLite, Sklearn
- Data Sources:
  - SpaceX API
  - SpaceX official website
  - NASA CRS mission data
  - Wikipedia launch records
- This project was completed as part of the IBM Data Science Professional Certificate capstone on Coursera.

Thank You!

Questions and feedback are welcome.