Data Science Project on SpaceX Launches

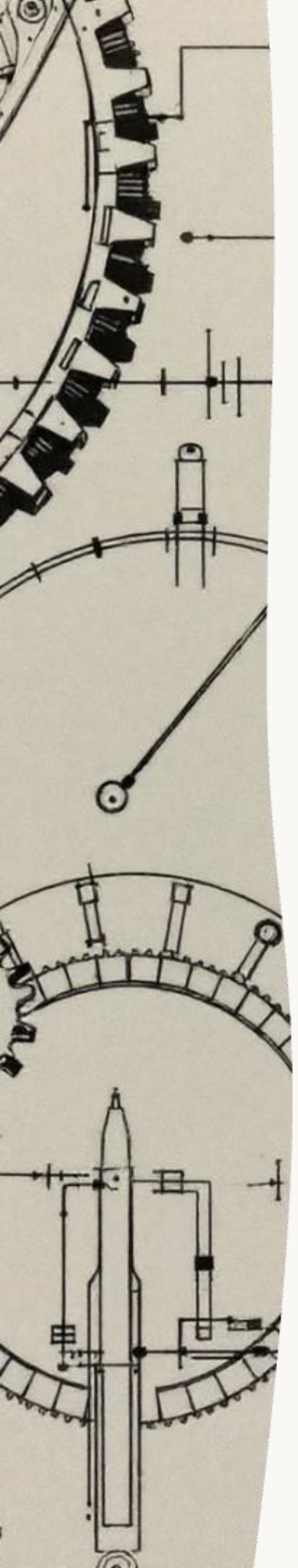
Analyzing Success Factors and Predictions

Natalia Tkachenko

GitHub

December 2024

IBM Developer SKILLS NETWORK



EXECUTIVE SUMMARY

1 Why This Project Matters:

SpaceX has revolutionized space travel, slashing launch costs with its reusable *Falcon 9* rockets. While competitors charge up to $165 million per launch, SpaceX delivers missions at just $62 million, thanks to their groundbreaking ability to recover and reuse the first-stage booster.

2 The Catch:

Not every landing is successful. This project dives deep into the data to uncover the secrets behind successful landings.

3 Unlocking Potential:

By predicting when and why a Falcon 9’s first stage will stick the landing, we unlock the potential to optimize costs, enhance efficiency, and gain a competitive edge in the space industry.

Introduction

Unlocking the Secrets of SpaceX's Falcon 9 Landings

1. ### Revolutionizing Space Travel

   SpaceX has transformed the industry with its reusable **Falcon 9** rockets, slashing launch costs by over 60% compared to competitors.

2. ### The Data-Driven Advantage

   By analyzing the wealth of data generated with each launch, we can uncover the patterns and insights that drive successful **Falcon 9** landings.

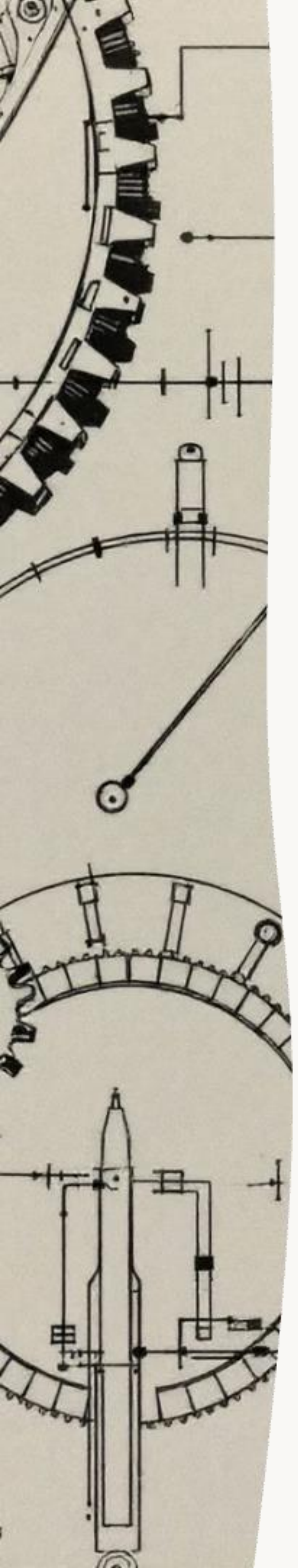
3. ### Predicting the Unpredictable

   Not every landing is perfect - but with advanced data science, we can forecast when and why the first stage will stick the landing.

This project explores key questions:

1. ### How do payload and launch conditions impact success?
2. ### Can we identify trends in landing improvements over time?
3. ### Which machine learning models are best for prediction?

By harnessing the power of data, we unlock the potential to optimize operations, enhance efficiency, and solidify SpaceX's competitive edge in the new space age.



Data Collection and Processing Methodology

Data Collection Methods

To uncover the insights that drive SpaceX's success, we leveraged a robust data collection and processing methodology:

SpaceX REST API

We tapped into the wealth of launch data available through the SpaceX API, extracting details on rockets, payloads, launch sites, and landing outcomes.

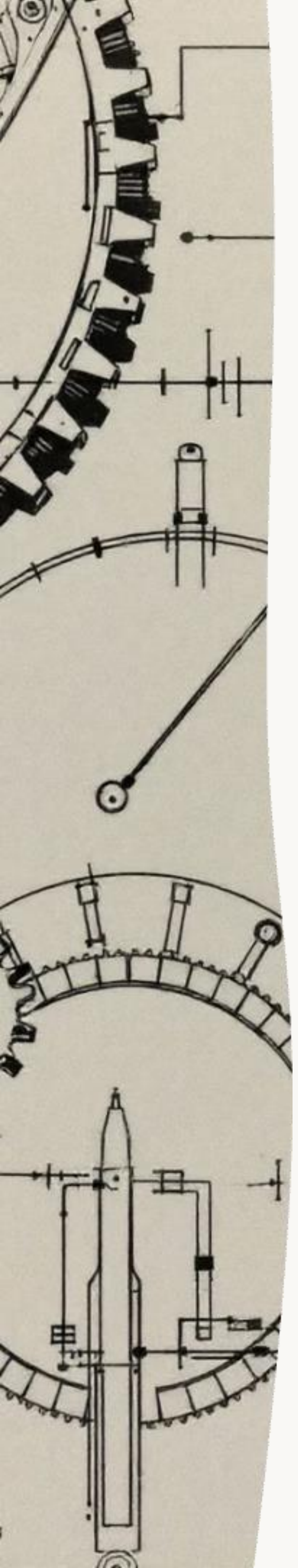
By transforming the raw JSON responses into structured data frames, we set the stage for in-depth analysis.

Web Scraping

To supplement the API data, we turned to historical Falcon 9 and Falcon Heavy launch information on Wikipedia.

Using BeautifulSoup, we parsed HTML tables to collect additional context on payloads, orbits, and customers.

With this comprehensive data foundation in place, we were able to dive deep into the factors driving successful rocket landings.



Data Collection and Processing Methodology

Data Wrangling and Formatting:

Filtering and Cleaning:

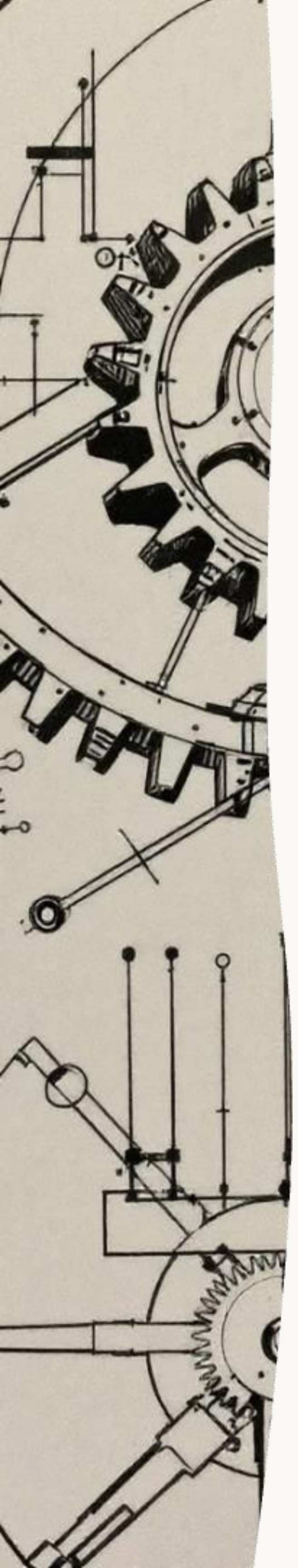
•Removed irrelevant rows and entries with multiple payloads or cores to maintain a uniform structure.

•Handled missing values by replacing them with statistical measures like the mean (e.g., for Payload Mass).

Normalization:

•Standardized categorical features using One-Hot Encoding.

•Reformatted data into 90 rows with 17 consistent features.



Data Collection and Processing Methodology

Tools and Techniques:

Libraries Used

•**Pandas** for data manipulation.

•**NumPy** for mathematical operations.

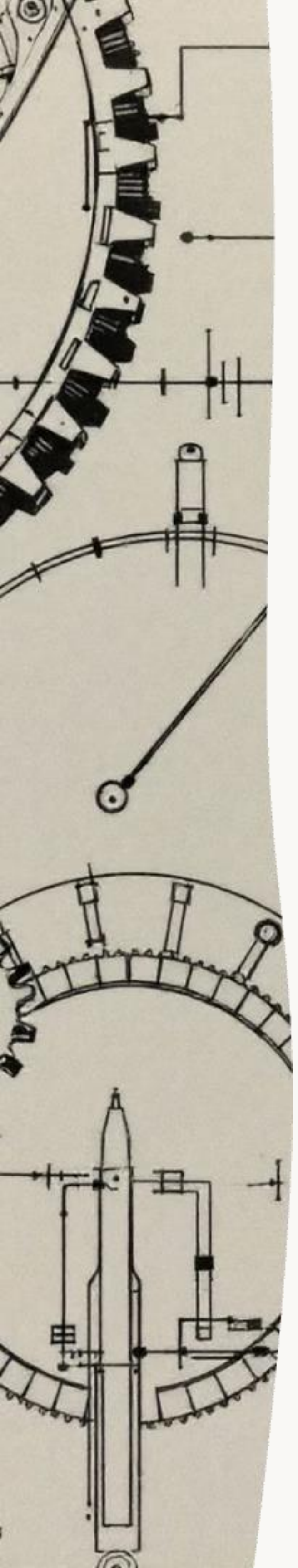
•**BeautifulSoup** and **Requests** for web scraping and API integration.

API Integration

•Automated data extraction with Python functions to retrieve rocket, payload, and core details via SpaceX’s API.

Output

•Cleaned, consolidated dataset ready for exploratory data analysis and predictive modeling.



Methodology: Exploratory Data Analysis (EDA) and Interactive Data Visualization

Exploratory Data Analysis (EDA)

•**Purpose:** EDA helps uncover patterns, relationships, and insights from raw data, preparing it for further modeling and prediction.

Tools Used

Pandas

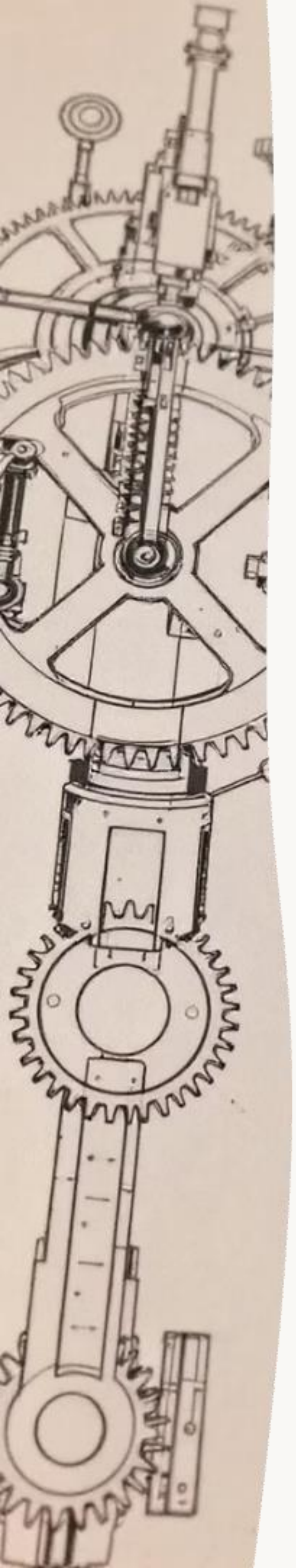
Pandas is a powerful library for data manipulation and aggregation. It provides a wide range of functionalities for cleaning, transforming, and analyzing data.

NumPy

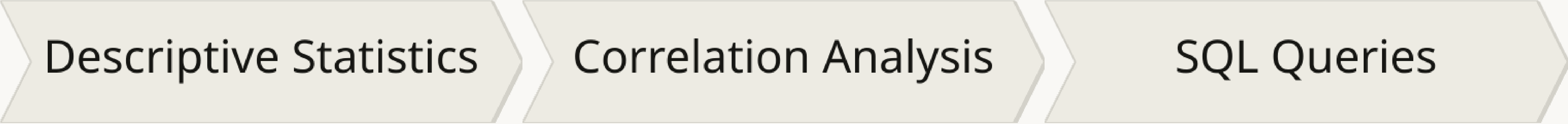
NumPy is used to perform mathematical computations on data. It provides efficient arrays and matrices for numerical operations.

SQL

SQL is used for complex queries and structured data analysis. It is particularly useful for working with large datasets and relational databases.



Methodology: Exploratory Data Analysis (EDA) and Interactive Data Visualization



Descriptive Statistics

Analyzed key metrics such as payload mass, success rates, and flight numbers.

Correlation Analysis

Explored relationships between variables like payload mass and launch success

SQL Queries

Examples:

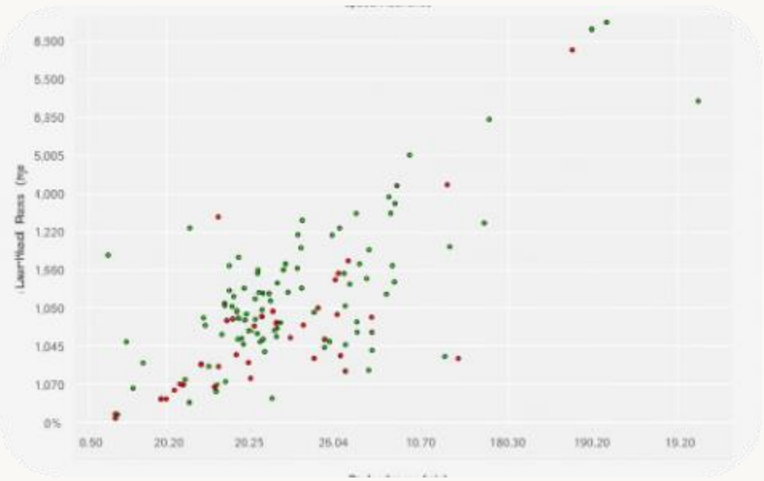
•Count of unique launch sites.

•Payload mass carried by NASA (CRS) missions.

•Success rate by orbit type.

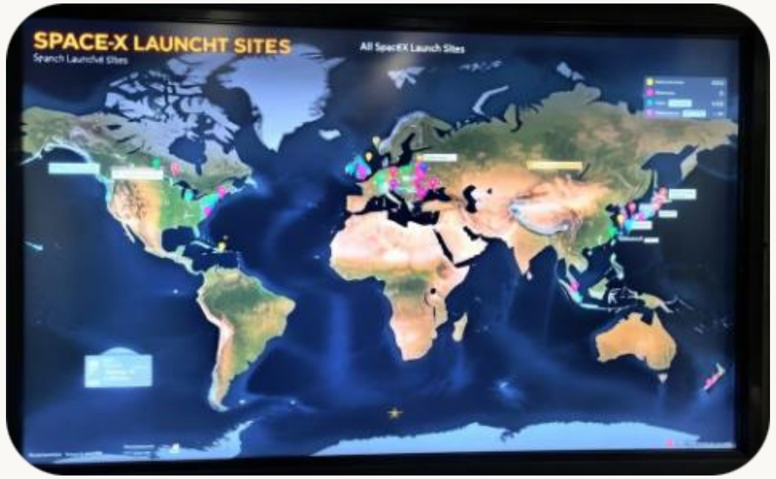
Methodology: Exploratory Data Analysis (EDA) and Interactive Data Visualization

Interactive visuals make it easier to identify trends, outliers, and patterns, facilitating decision-making.



Scatter Plots with Matplotlib and Seaborn

Visualize relationships between variables, like payload mass and launch success, for deeper analysis.

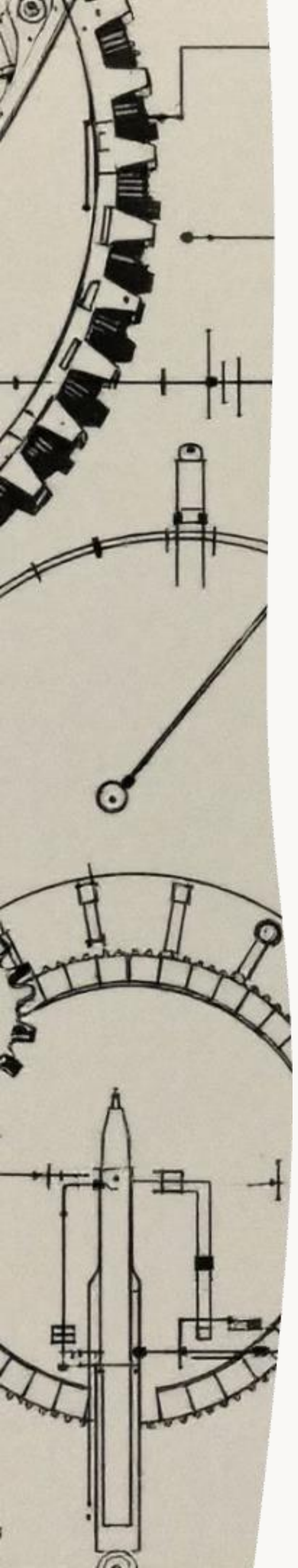


Interactive Maps with Folium

Geospatial data visualization to map launch locations and identify trends in launch site usage.

Interactive Dashboards with Plotly Dash

Combine multiple visualizations to create dynamic dashboards for comprehensive data exploration.



Visualization Examples:

Visualization Examples

Scatter Plots

- Payload Mass vs. Success Rate.
- Flight Number vs. Launch Site.

Bar Charts

- Success rates by orbit type.
- Payload mass distributions

Time Series Analysis

- Success rate trends over the years

Interactive Dashboard

Created using Plotly Dash, allowing users to:

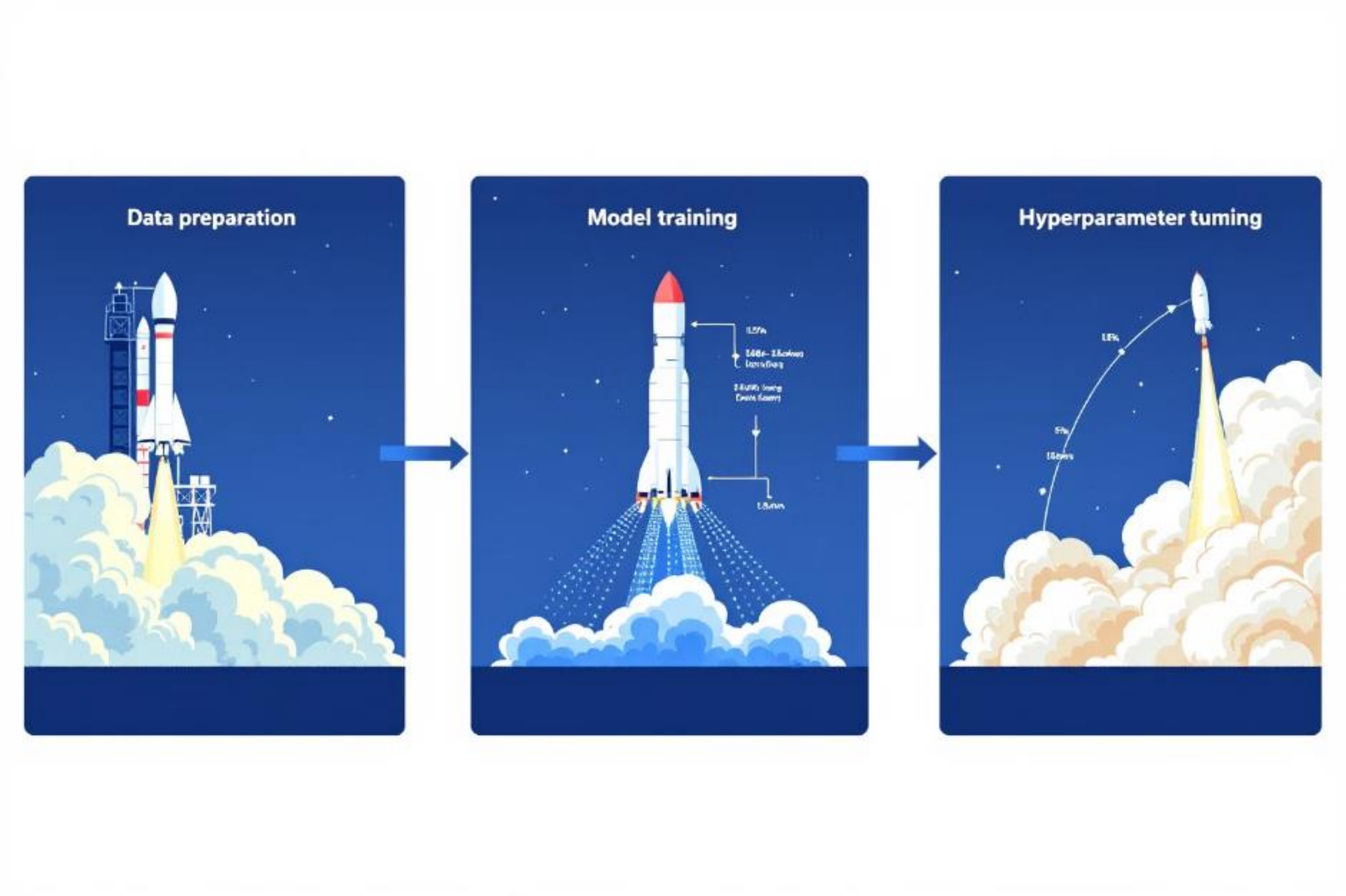
- Filter data by payload range.
- View success rate

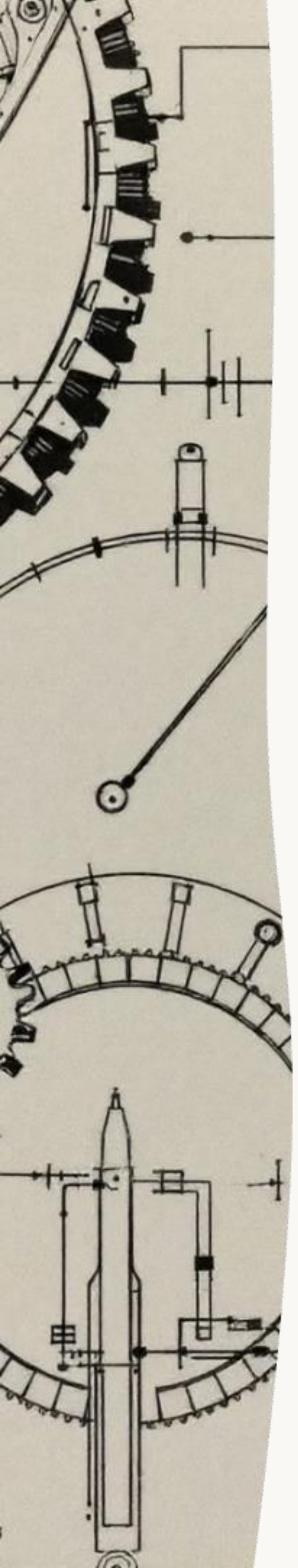
Outcome: Through EDA and visualizations, patterns such as payload mass affecting success rates and launch site performance variations were uncovered. These insights informed the selection of features for predictive modeling and enabled intuitive storytelling through interactive visuals.

Predictive Analysis Methodology

Machine Learning for Falcon 9 First-Stage Landing Prediction

Predicting whether the Falcon 9 first stage will land successfully is critical for optimizing launch costs and improving competitive advantage. This methodology leverages machine learning models to achieve high predictive accuracy through a systematic process of data preparation, model training, and hyperparameter tuning.





Data Preprocessing

- Converted 'Class' column to a NumPy array for labels.
- Standardized features using StandardScaler.
- Dataset split: 80% training, 20% testing.

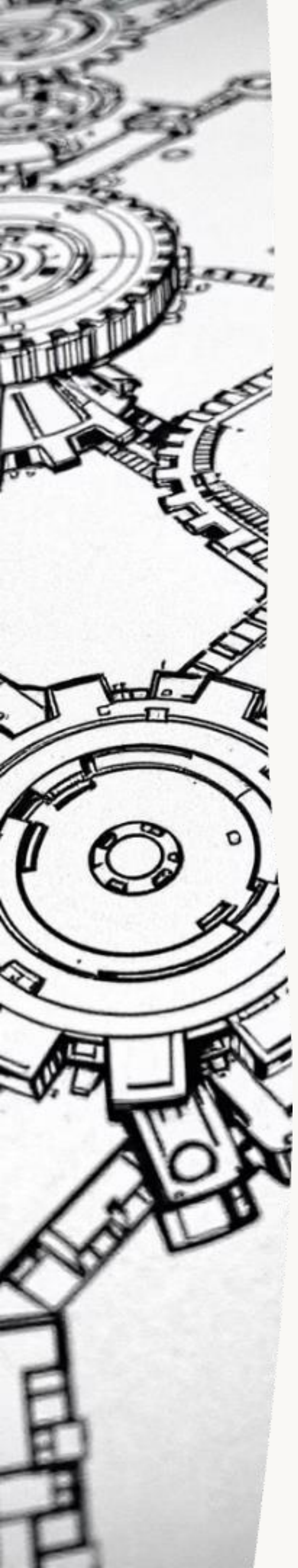
Model Training & Tuning

GridSearchCV optimized hyperparameters for four models:

- Logistic Regression (C, penalty, solver)
- SVM (kernel, C, gamma)
- Decision Tree (depth, criteria, min samples)
- KNN (neighbors, algorithm, metric)

Evaluation

- Metrics: Accuracy, F1-score, confusion matrix.
- Model comparison to identify the best performer.

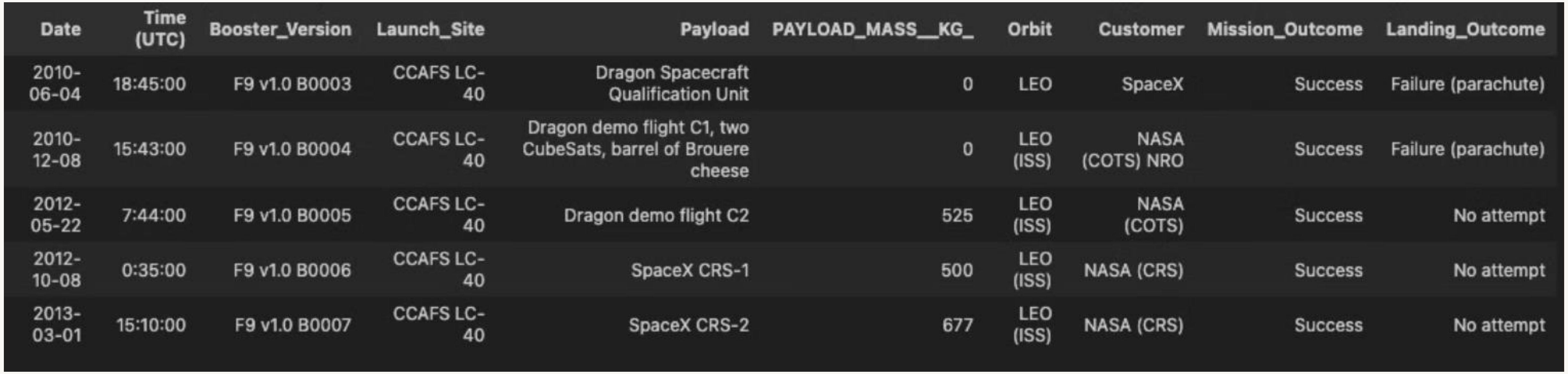


RESULTS

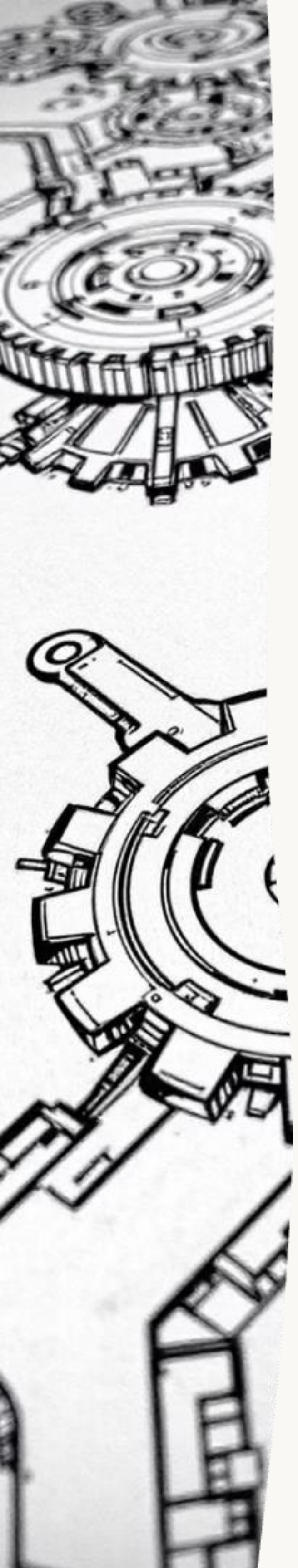
EDA with SQL results

This section presents the tasks of analyzing *SpaceX* data using **Python** and **SQL**. The main goal was to explore the dataset, load it into a Db2 database table, and execute SQL queries to answer specific questions.

Understanding the SpaceX Dataset: Records of the initial launches are presented, including details such as date, time, launch sites, and landing outcomes.

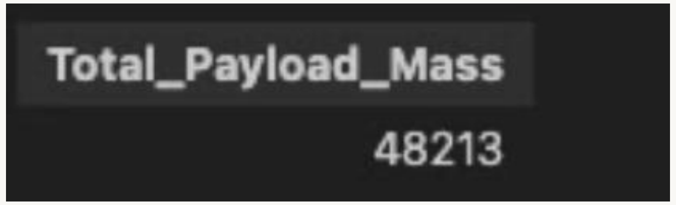


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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



RESULTS

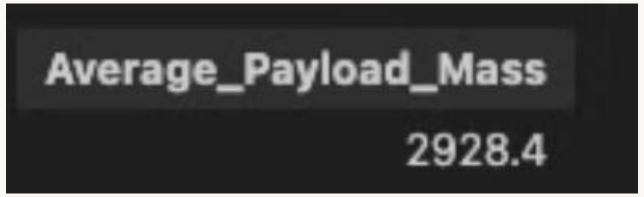
Total Payload Mass



Total_Payload_Mass
48213

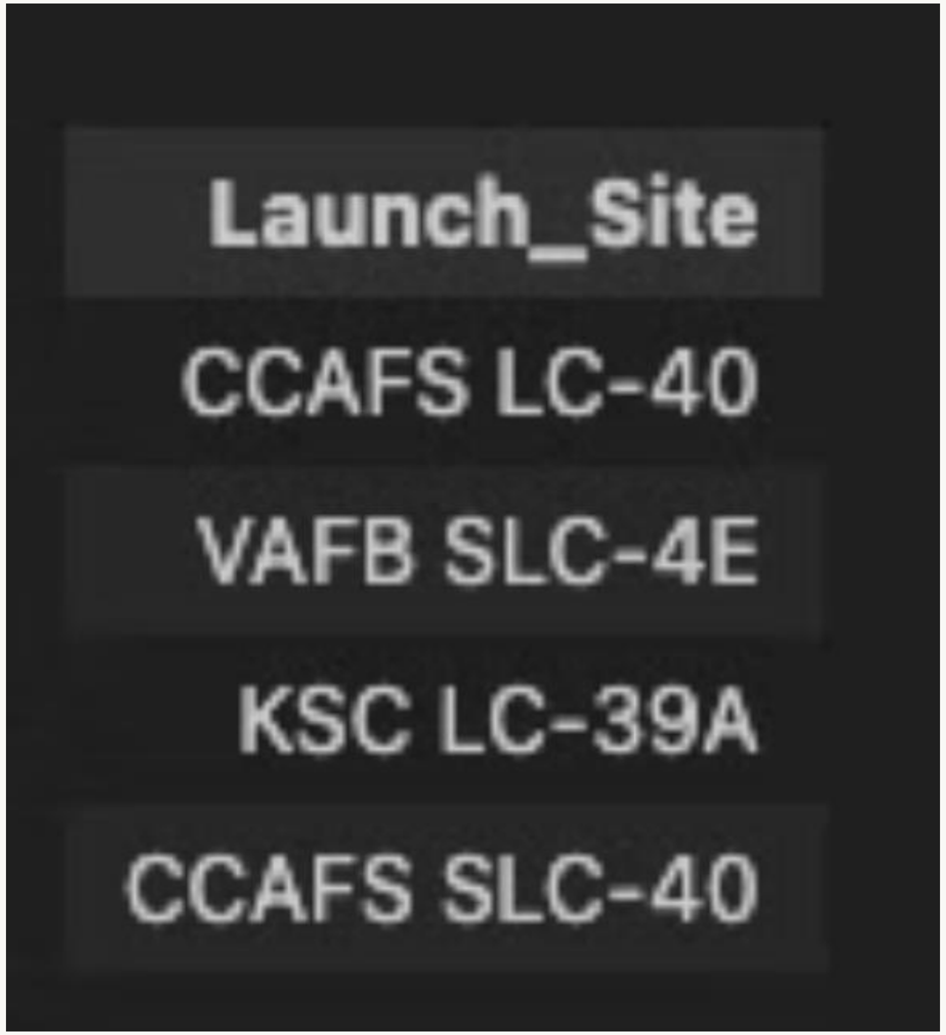
the total payload mass across all missions is 48,213 kg

Average Mass



Average_Payload_Mass
2928.4

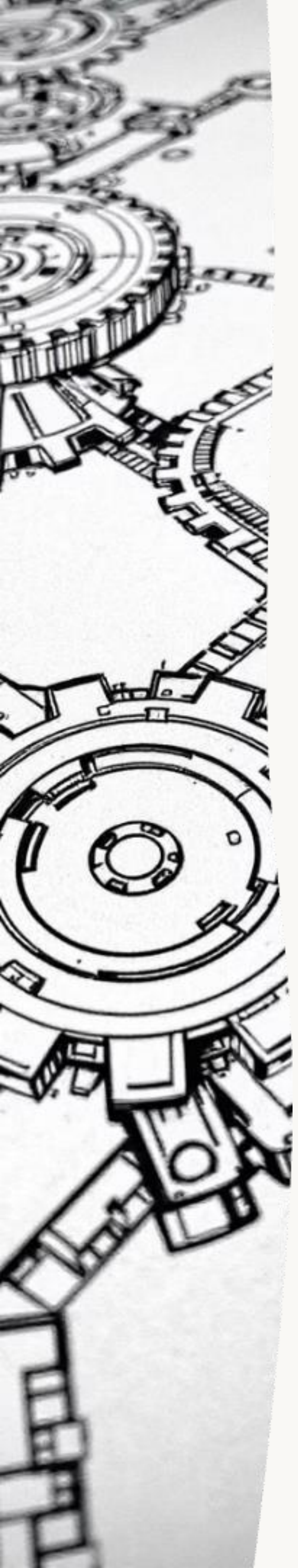
The average payload mass is calculated to be 2928.4 kg.



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

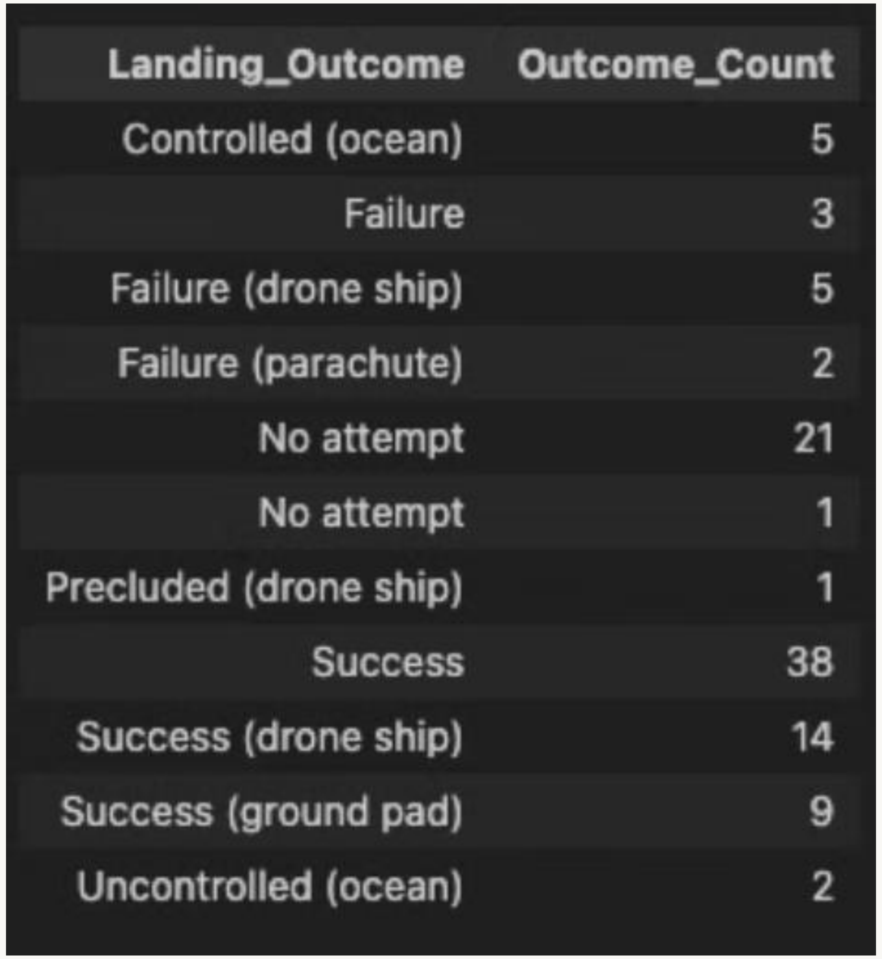
Unique Launch Sites

All launch sites are shown, highlighting the distribution of missions across different locations.



RESULTS

EDA with SQL results

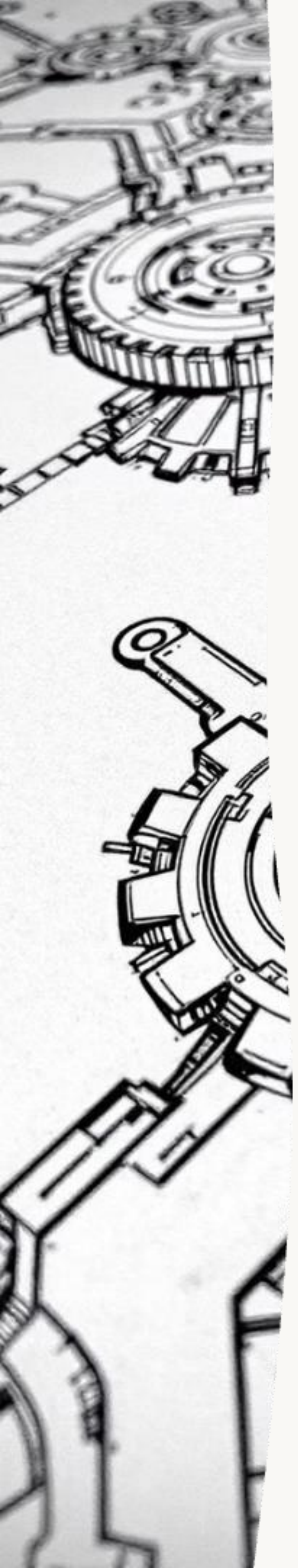


Landing_Outcome	Outcome_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Landing Outcomes

A detailed analysis of landing outcomes, including successful, failed, and missed attempts, with distribution by count.

This approach not only provides an in-depth exploration of SpaceX mission data but also answers specific questions using SQL queries, ensuring a precise and structured analysis.



RESULTS

EDA with SQL results

This section provides insights into *SpaceX's* rocket launch data, focusing on critical factors influencing mission outcomes.

The visualizations explore various relationships, including launch sites, payload mass, orbit types, and flight numbers, relative to success rates.

Observations reveal how different features, such as orbit type or payload mass, correlate with successful landings, emphasizing patterns and trends over the years.

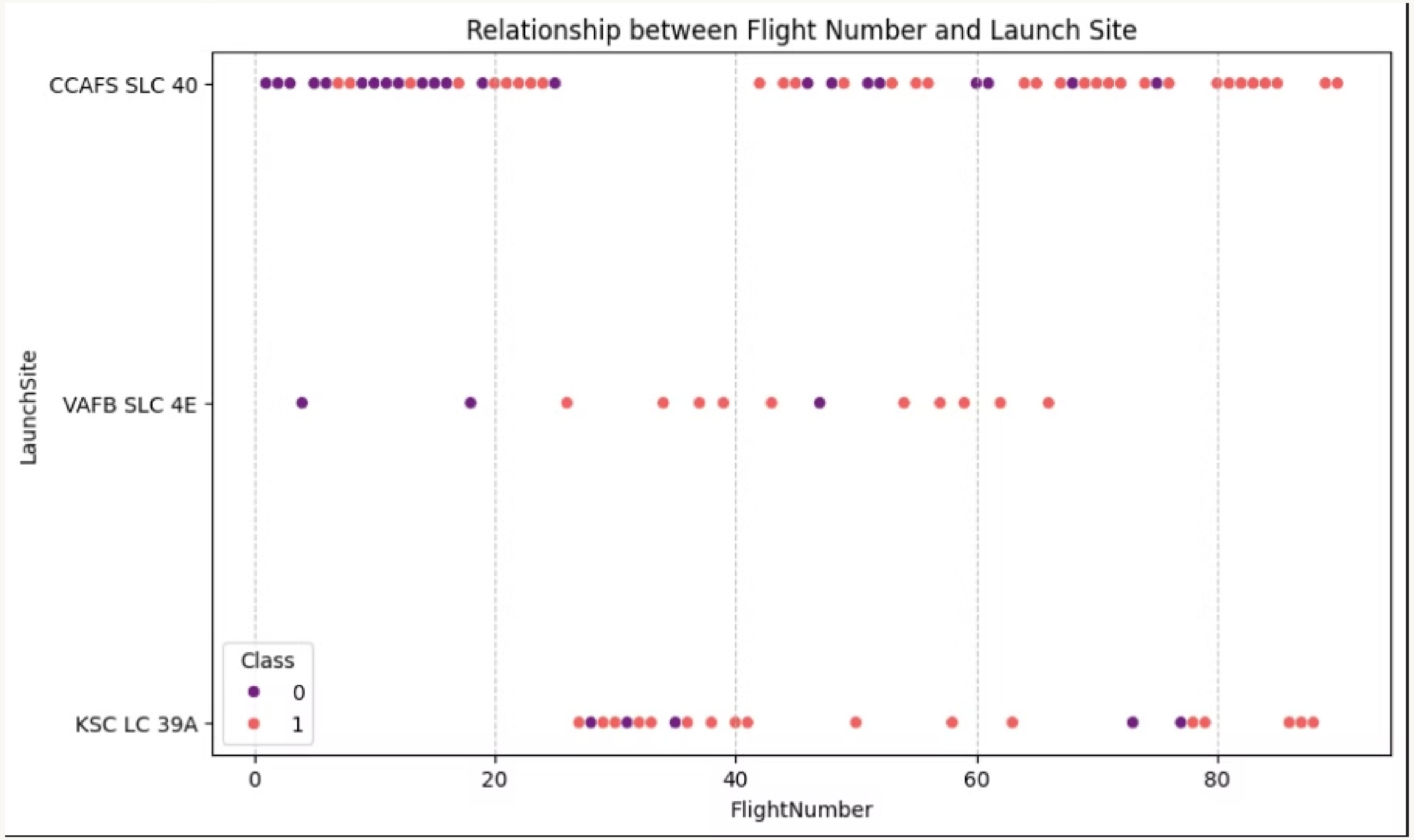
Key findings

- The progression of success rates by year, showcasing SpaceX's improvements.
- The impact of payload mass and flight numbers on launch success across different orbits and launch sites.
- Comparative success rates by orbit types, identifying high-performing orbits like LEO and SSO.

RESULT

EDA with visualization

This scatterplot illustrates the relationship between flight numbers and launch sites (CCAFS SLC-40, VAFB SLC-4E, and KSC LC-39A), with the class variable indicating success (1) or failure (0)

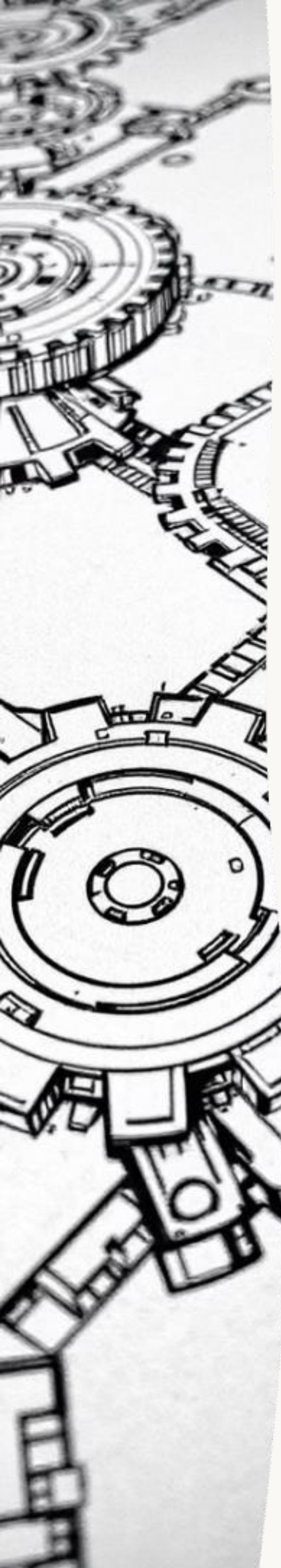


CCAFS SLC-40: This site has the most launches, showing both successes and failures. Success rates seem to improve with higher flight numbers.

VAFB SLC-4E: This site has fewer launches, with mixed outcomes, suggesting it is less utilized than

CCAFS SLC-40.KSC LC-39A: This site has consistently successful outcomes for most launches, especially at higher flight numbers.

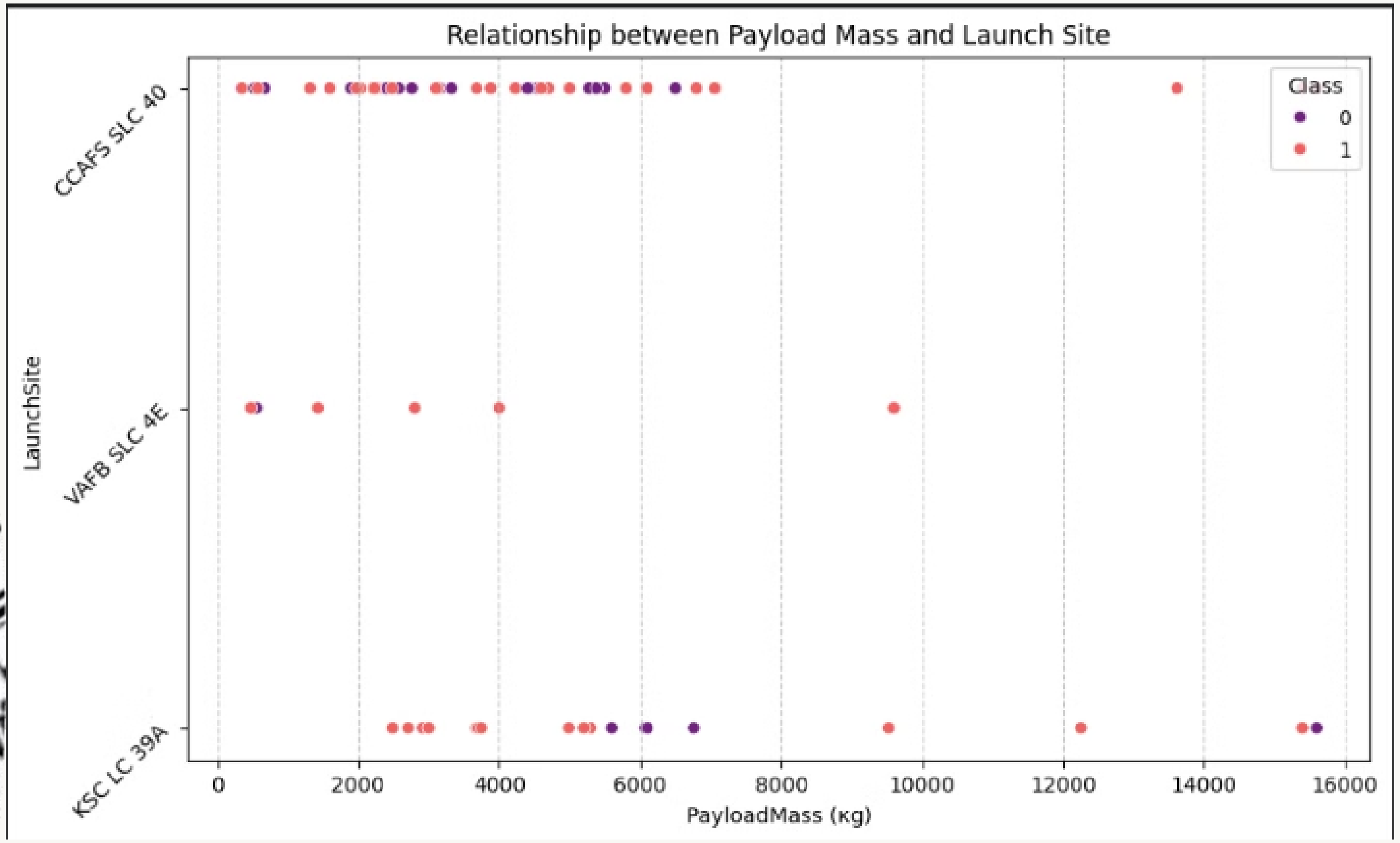
Overall, the trend suggests that higher flight numbers correlate with more successes across all launch sites, demonstrating improved operational reliability over time.



RESULT

EDA with visualization

This scatterplot illustrates the relationship between payload mass (in kilograms) and launch sites (CCAFS SLC-40, VAFB SLC-4E, and KSC LC-39A), with the class variable indicating success (1) or failure (0) of the mission



CCAFS SLC-40: The most launches occur at this site, spanning a wide range of payload masses. Success (1) is more prevalent across payloads up to approximately 10,000 kg.

VAFB SLC-4E: This site handles fewer launches and payloads are typically smaller (below 6,000 kg), with mixed outcomes.

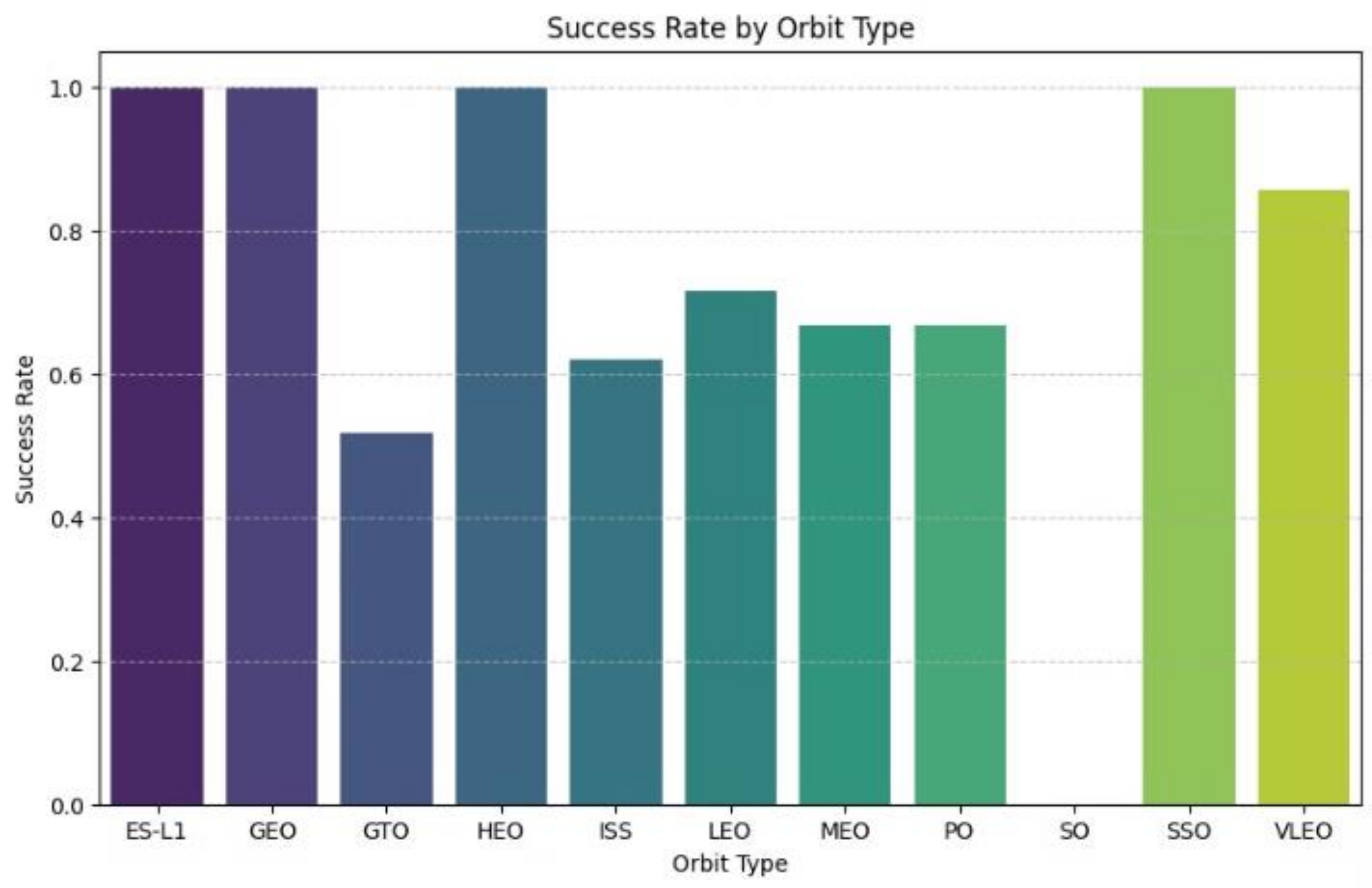
KSC LC-39A: This site is notable for handling larger payloads (up to 16,000 kg). Successful outcomes (1) dominate, especially for higher payload masses.

The overall trend suggests that payload mass does not strongly determine success, as successful launches occur across a range of masses at all sites. However, certain sites like KSC LC-39A tend to handle larger payloads with a higher success rate.

RESULT

EDA with visualization

This bar chart visualizes the success rate of SpaceX launches for different orbit types. Key observations include:



Highest Success Rates (100%): Orbits such as ES-L1, GEO, HEO, and SSO have a perfect success rate, indicating consistent reliability for these missions.

Moderate Success Rates (60-80%): Orbits like LEO, MEO, and PO show moderate success rates, with successful outcomes being more frequent than failures.

Lowest Success Rate (below 50%): The GTO orbit stands out with the lowest success rate, highlighting challenges or complexities in achieving consistent success for missions targeting this orbit.

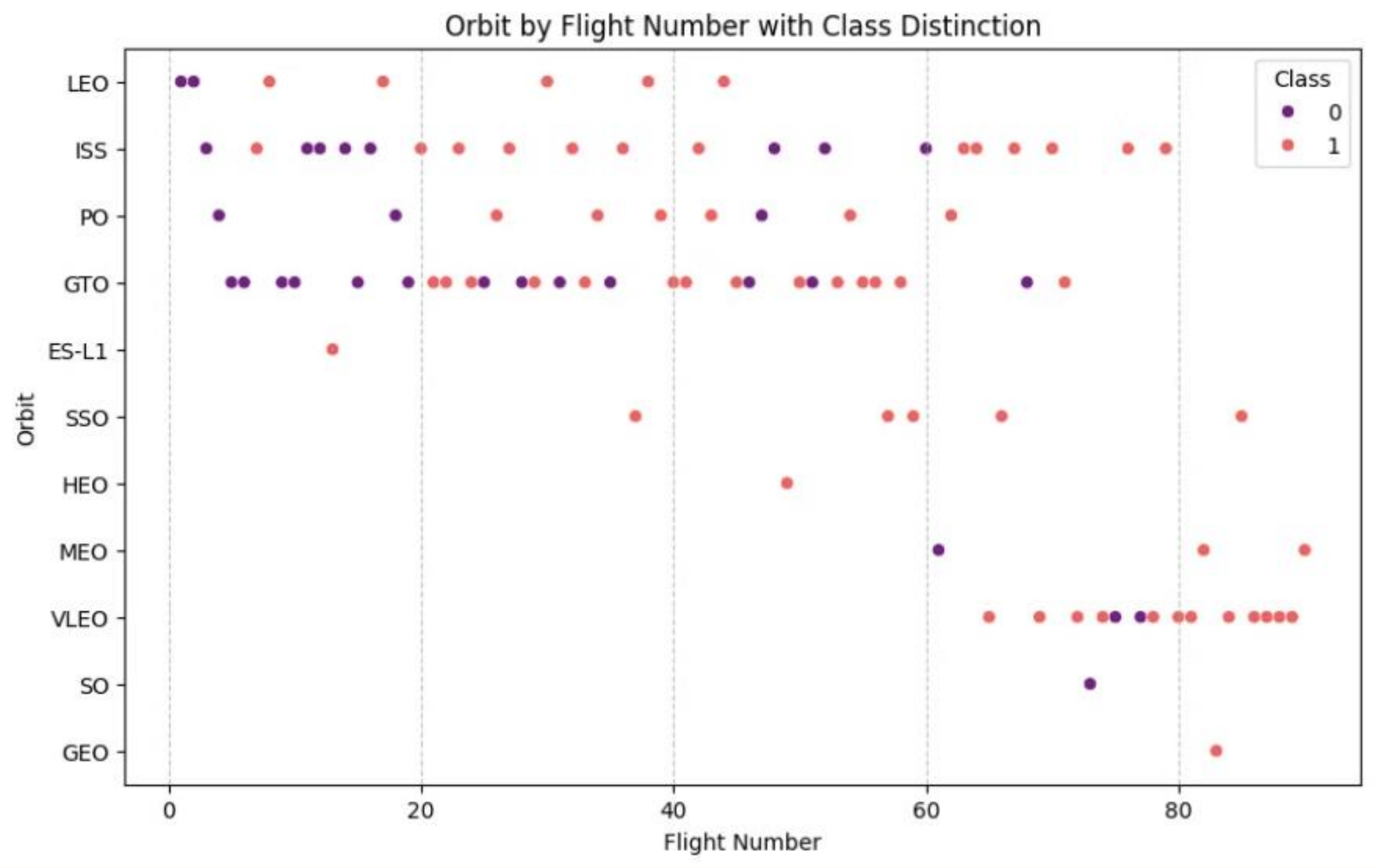
ISS and VLEO Orbits: These orbits exhibit a slightly better-than-average success rate but do not reach the reliability of orbits like GEO or SSO.

This chart highlights that the choice of orbit type can significantly influence mission outcomes, with some orbits proving more challenging than others.

RESULT

EDA with visualization

This scatterplot shows the relationship between flight number and orbit type, highlighting success (red, Class 1) or failure (purple, Class 0):



Successful launches (Class 1): Increase with later flights, indicating improved performance over time.

Failed launches (Class 0): More common in early missions, reflecting initial challenges.

Orbit trends:

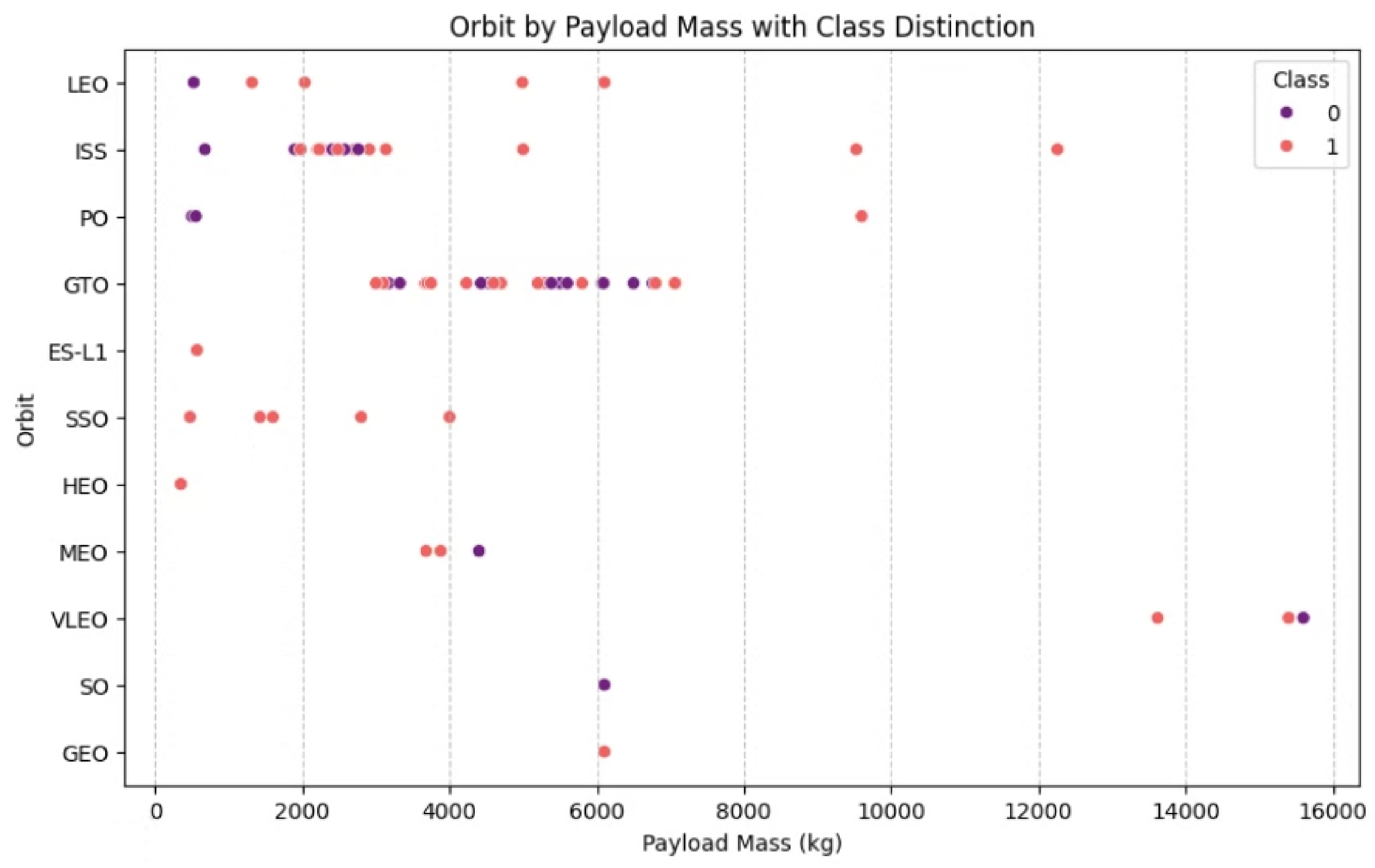
- **LEO and ISS**: Mixed results with increasing success in later flights.
- **GTO**: Balanced mix of successes and failures, showing challenges in this orbit.
- **Rare orbits (GEO, ES-L1, SSO)**: Mostly successful, reflecting specialized missions.

This chart highlights that the choice of orbit type can significantly influence mission outcomes, with some orbits proving more challenging than others.

RESULT

EDA with visualization

This scatterplot shows the relationship between **payload mass (x-axis)** and **orbit type (y-axis)**, with **class distinction** indicating the success (1) or failure (0) of the launches:



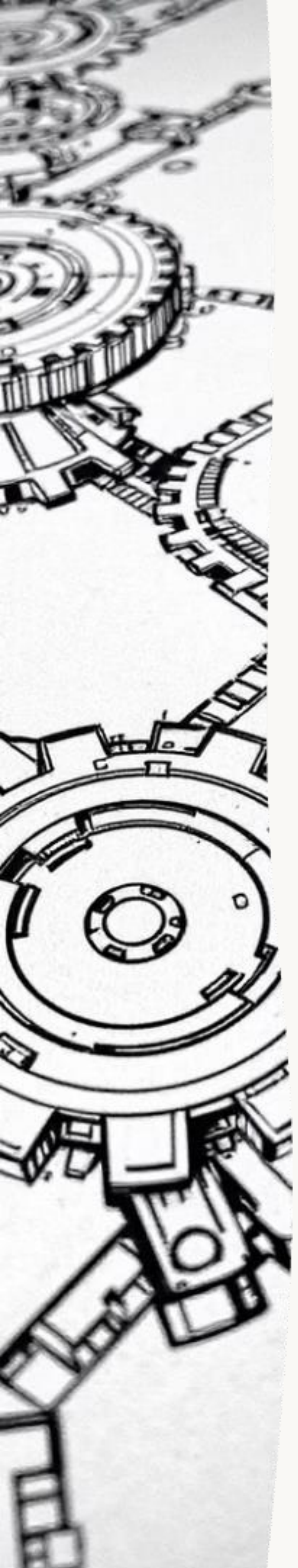
LEO and ISS: Wide payload range with increasing success for higher payloads.

GTO: Mixed outcomes, reflecting challenges across payloads.

High-payload missions (>10,000 kg): Mostly successful, especially in GEO and some LEO missions.

Specialized orbits (SSO, ES-L1, VLEO): Predominantly successful, likely due to specific mission designs.

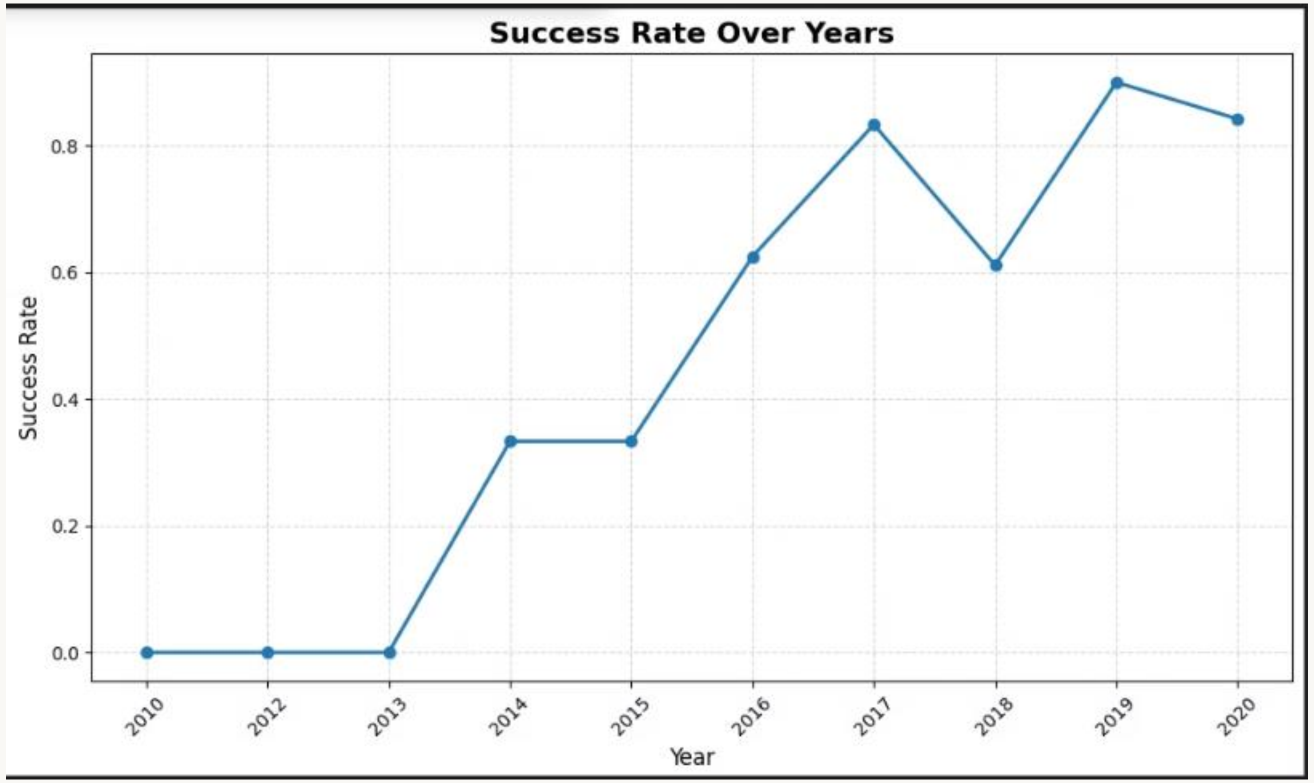
The chart highlights how payload mass influences mission outcomes across different orbits, showcasing SpaceX's growing expertise.



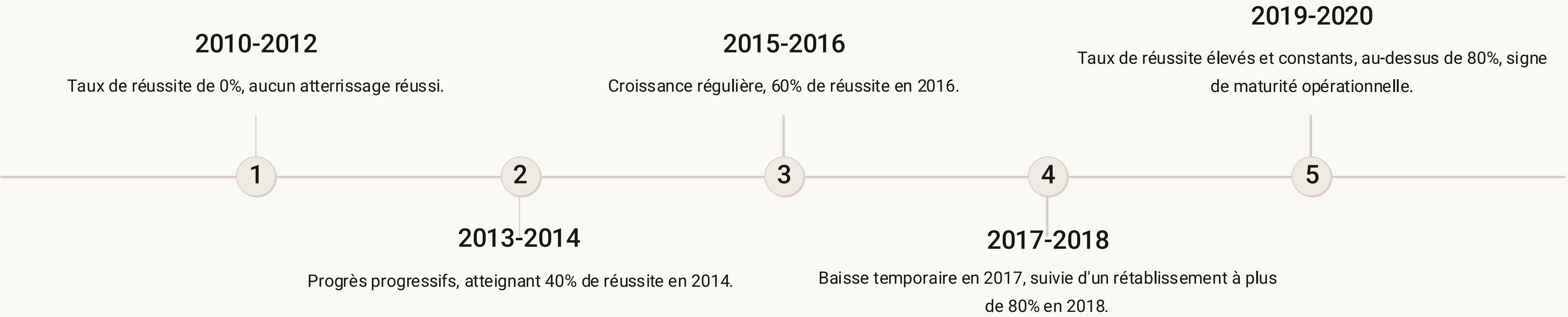
RESULT

EDA with visualization

This graph shows SpaceX's success rate from 2010 to 2020, highlighting key milestones:



The graph illustrates SpaceX's technological advancement and operational learning curve, showcasing how the company has achieved consistent and reliable rocket landings over a decade. This upward trend reflects their success in reusability and cost-efficiency in space exploration.



2010-2012

Taux de réussite de 0%, aucun atterrissage réussi.

2013-2014

Progrès progressifs, atteignant 40% de réussite en 2014.

2015-2016

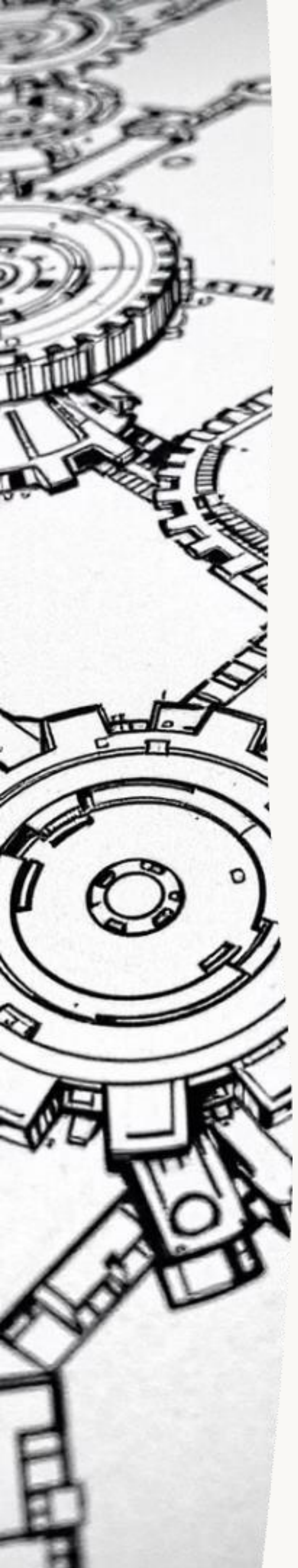
Croissance régulière, 60% de réussite en 2016.

2017-2018

Baisse temporaire en 2017, suivie d'un rétablissement à plus de 80% en 2018.

2019-2020

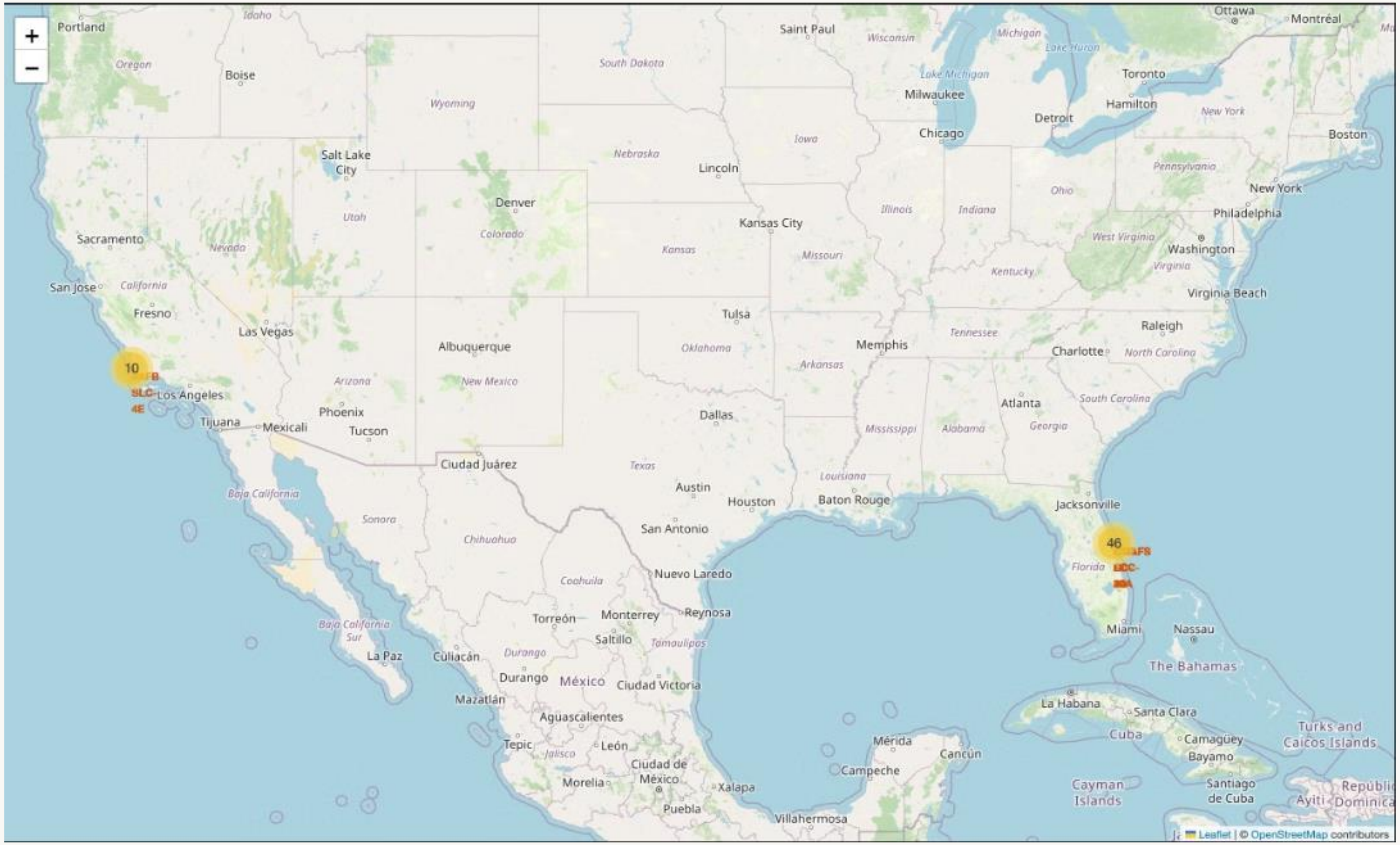
Taux de réussite élevés et constants, au-dessus de 80%, signe de maturité opérationnelle.



RESULT

Interactive map with *Folium*

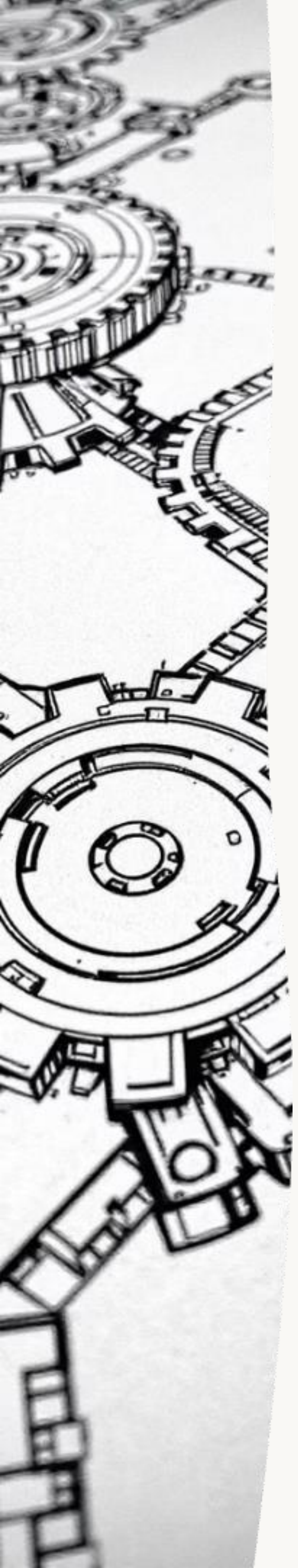
Launch Site Locations



All SpaceX launch sites are represented on the map with blue circles and labeled markers, offering a clear overview of their geographic distribution.

The map reflects the functional differentiation of sites, such as VAFB for polar orbits and Florida sites for equatorial orbits, showcasting SpaseX’s strategic considerations for optimal operations.

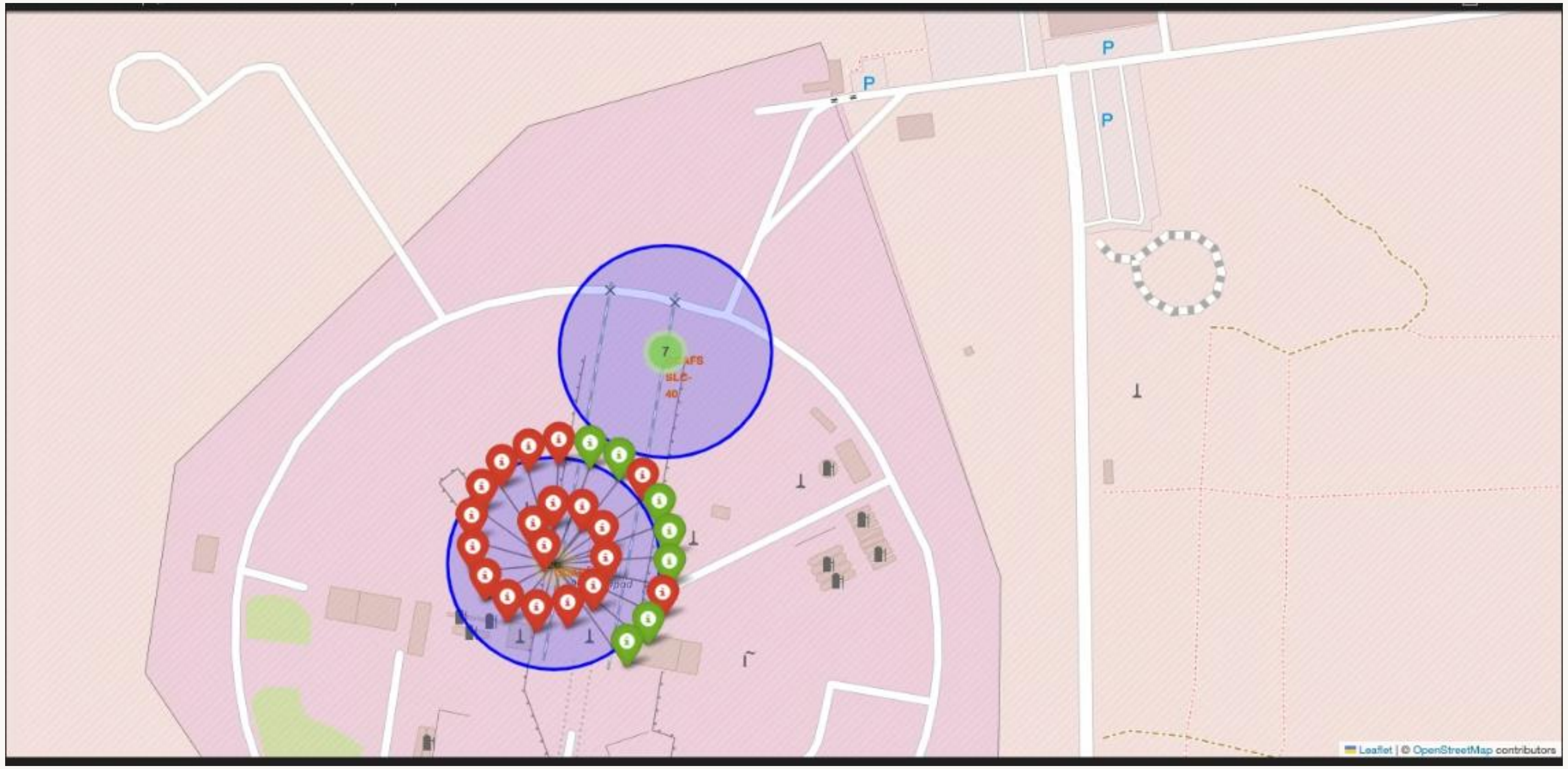
Sites are strategically positioned near coastlines to enhance safety during launches.



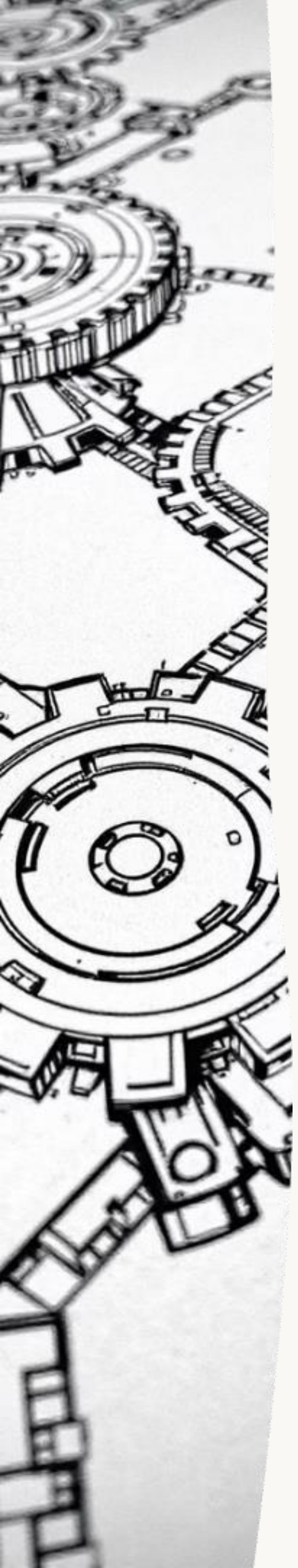
RESULT

Interactive map with *Folium*

Success and Failure Visualization



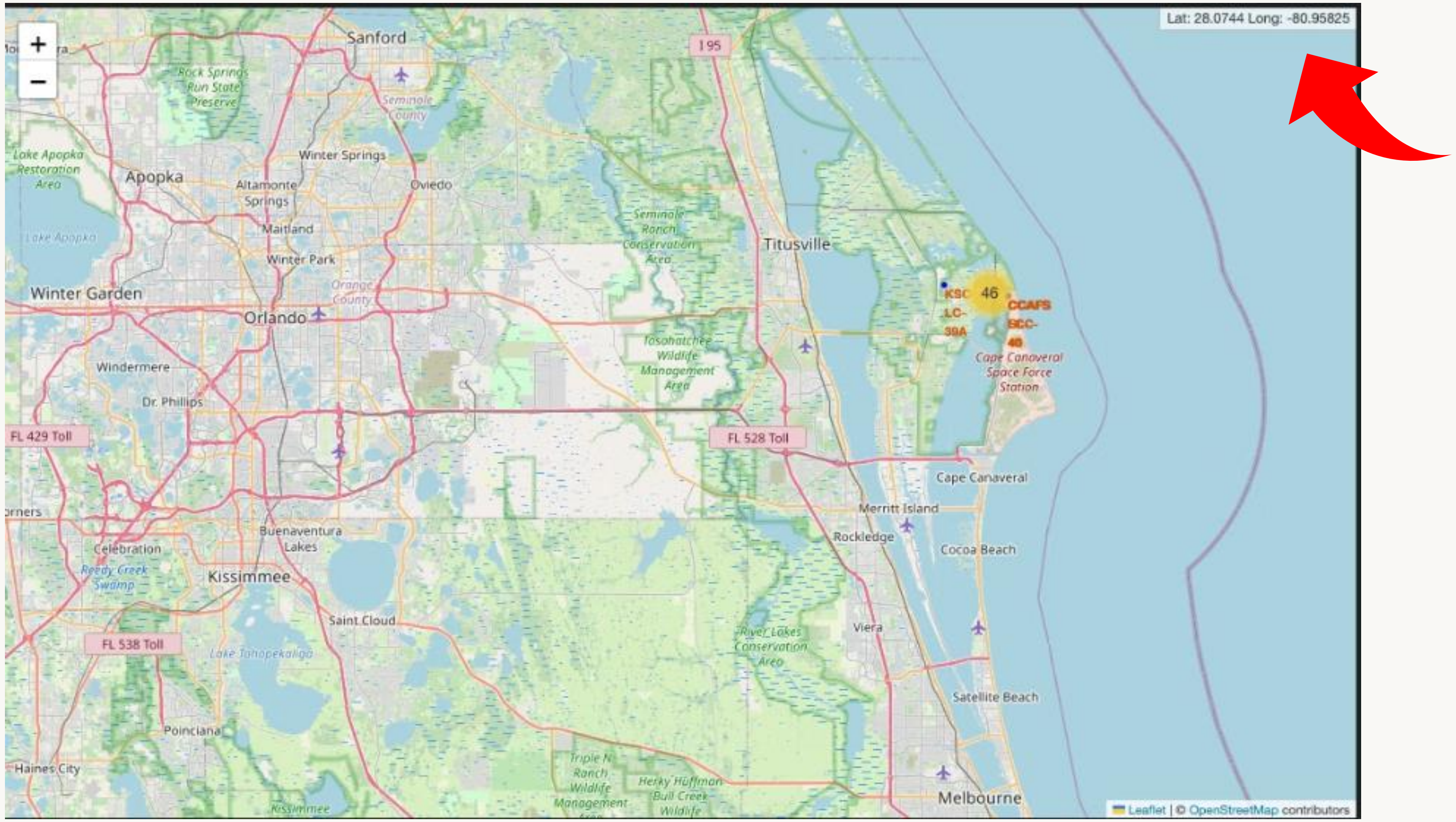
Marker clusters depict individual launches with green markers for successes and red for failures, enabling a quick assessment of site-specific performance trends.



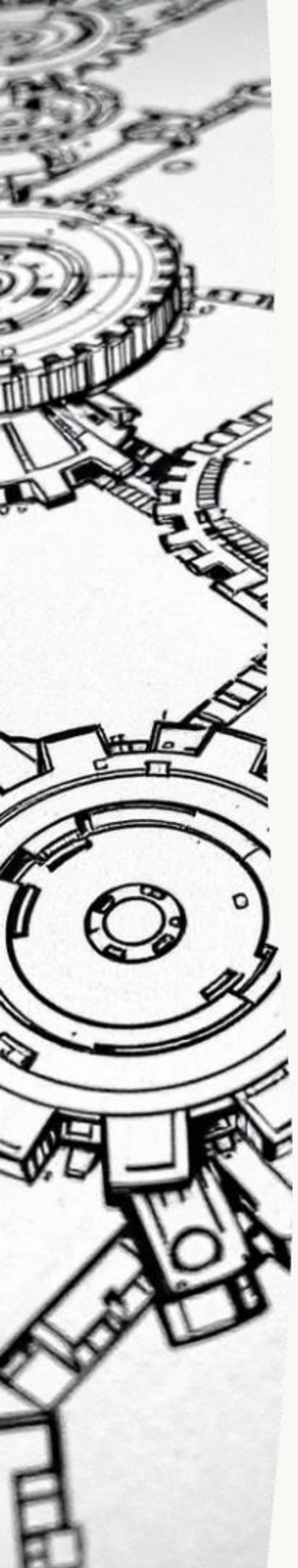
RESULT

Interactive map with *Folium*

Interactive Features



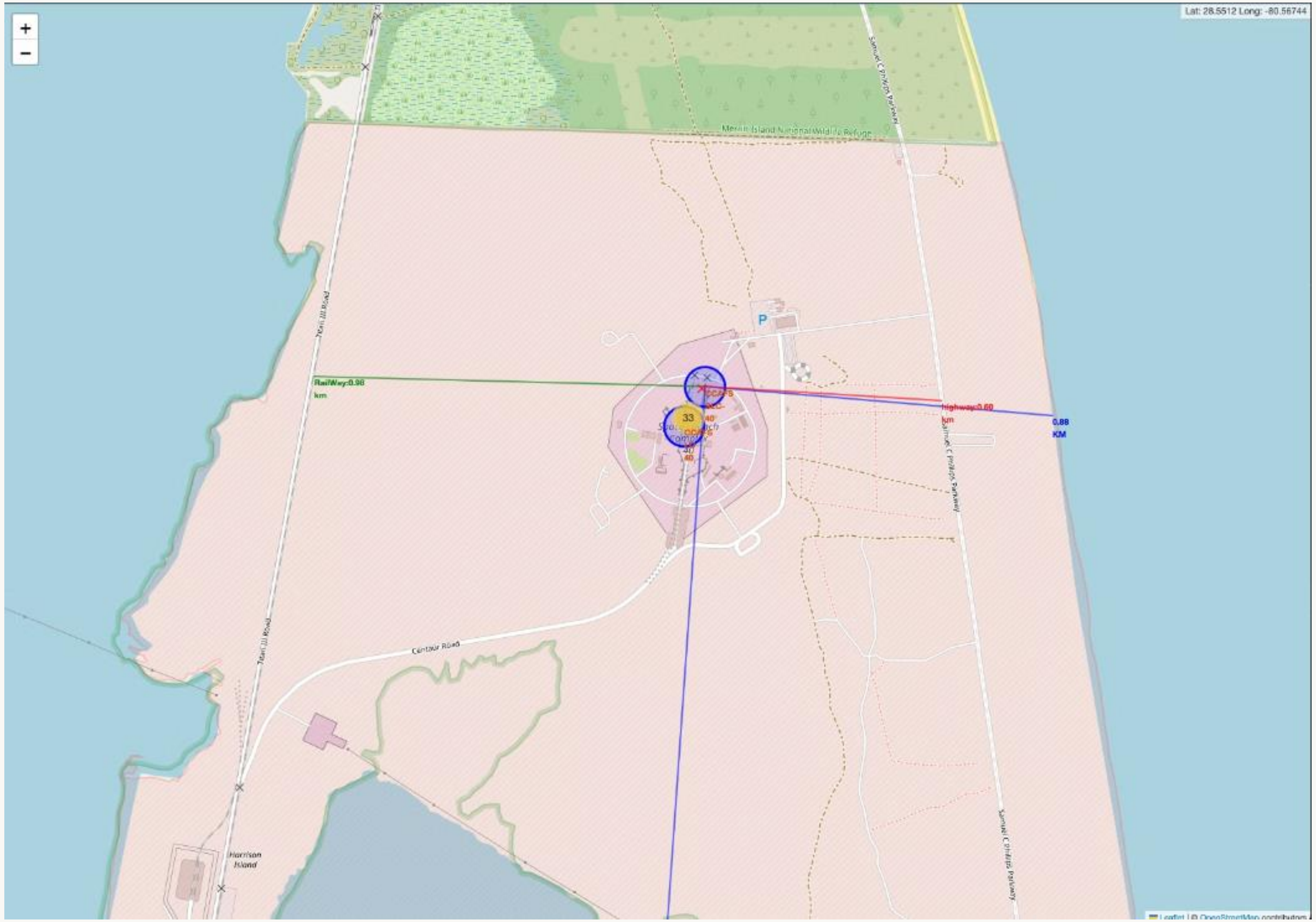
Interactive elements, such as mouseover coordinates and pop-ups with detailed site information, enhance user engagement and exploration.



RESULT

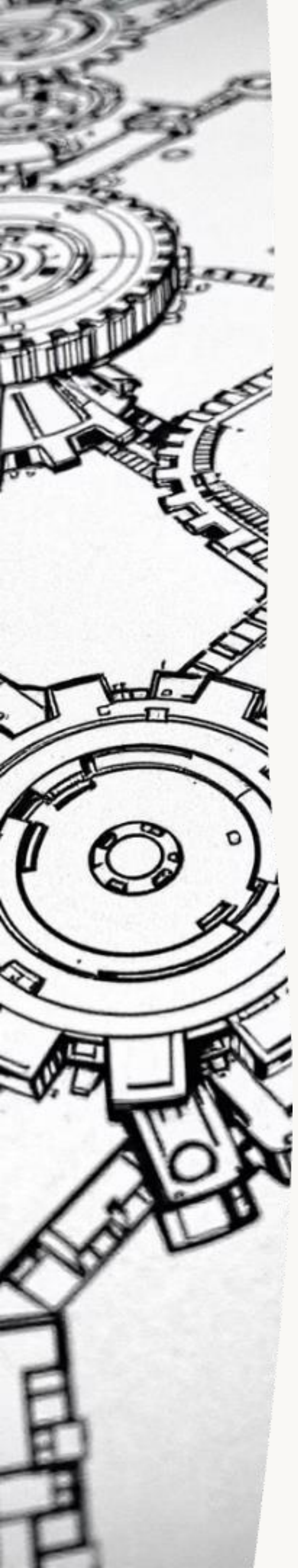
Interactive map with *Folium*

Launch Site Locations


+
−
Lat: 28.5512 Long: -80.56744
Merritt Island National Wildlife Refuge
Samuel C Phillips Parkway
P
RailWay:0.98
km
CCAFS
40
33
40
highway:0.60
km
0.88
KM
Centaur Road
Harrison Island
Leaflet | © OpenStreetMap contributors

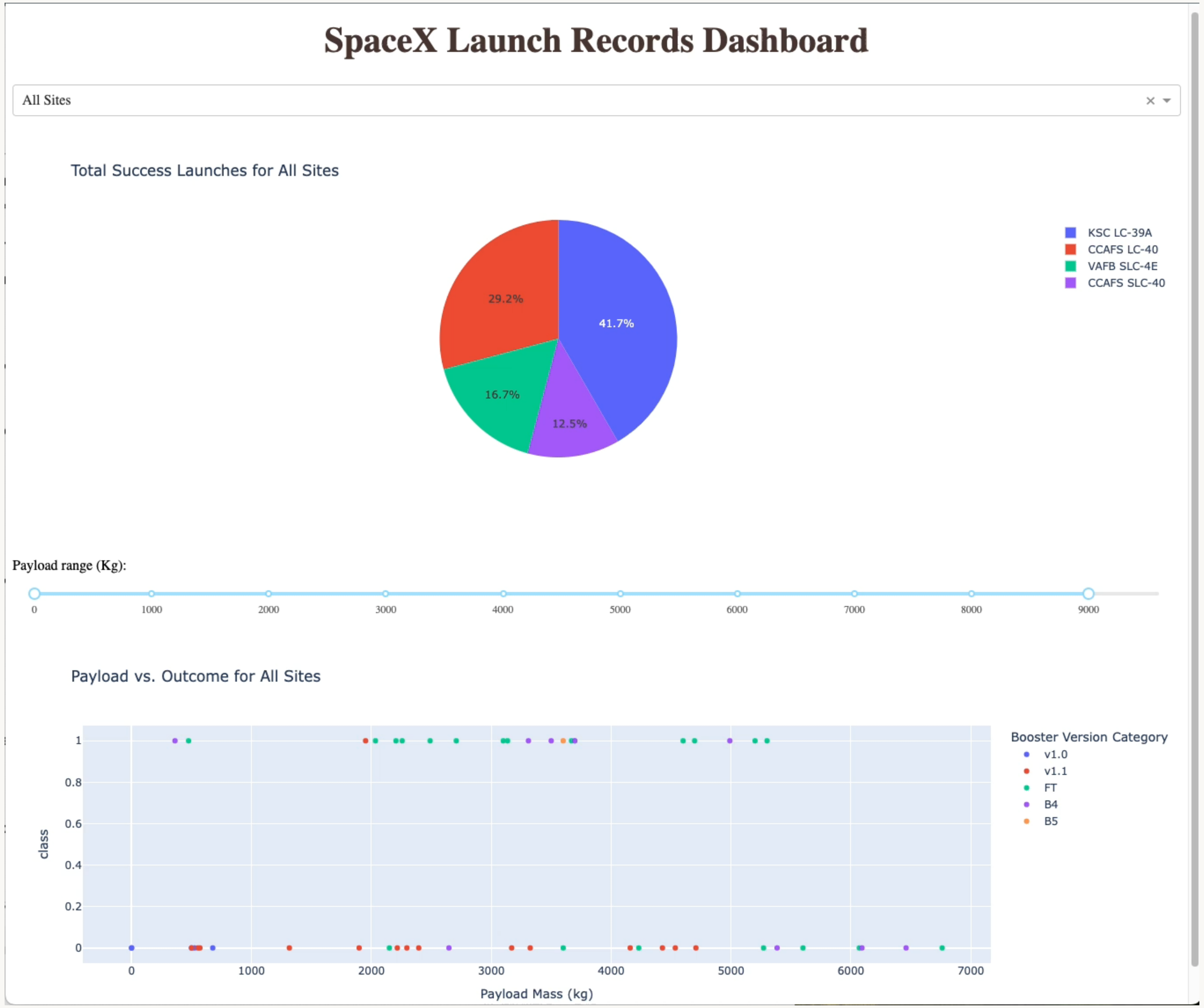
Blue, green, and red lines show calculated distances to nearby features, emphasizing the strategic design and operational planning of each site.

Distances from launch sites to coastlines, highways, railways, and cities are illustrated, demonstrating the sites' accessibility and logistical support.



RESULT

Plotly Dash dashboard



SpaceX Launch Records Dashboard

Launch Success Analysis:

A pie chart shows total successes by site or for all sites combined.

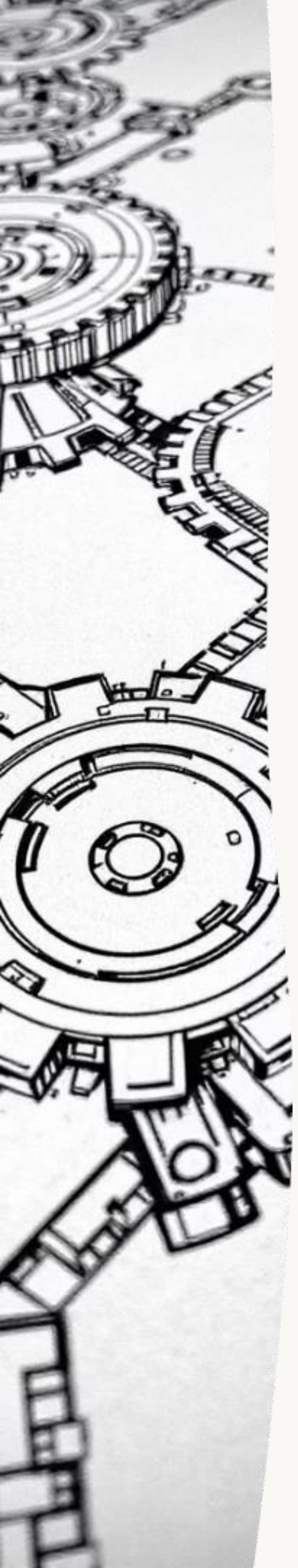
Payload vs Success:

A scatter plot highlights the relationship between payload mass and launch outcomes, with filtering by site and payload range.

Interactive Filters:

Dropdown menu for site selection and a slider for payload range adjustment.

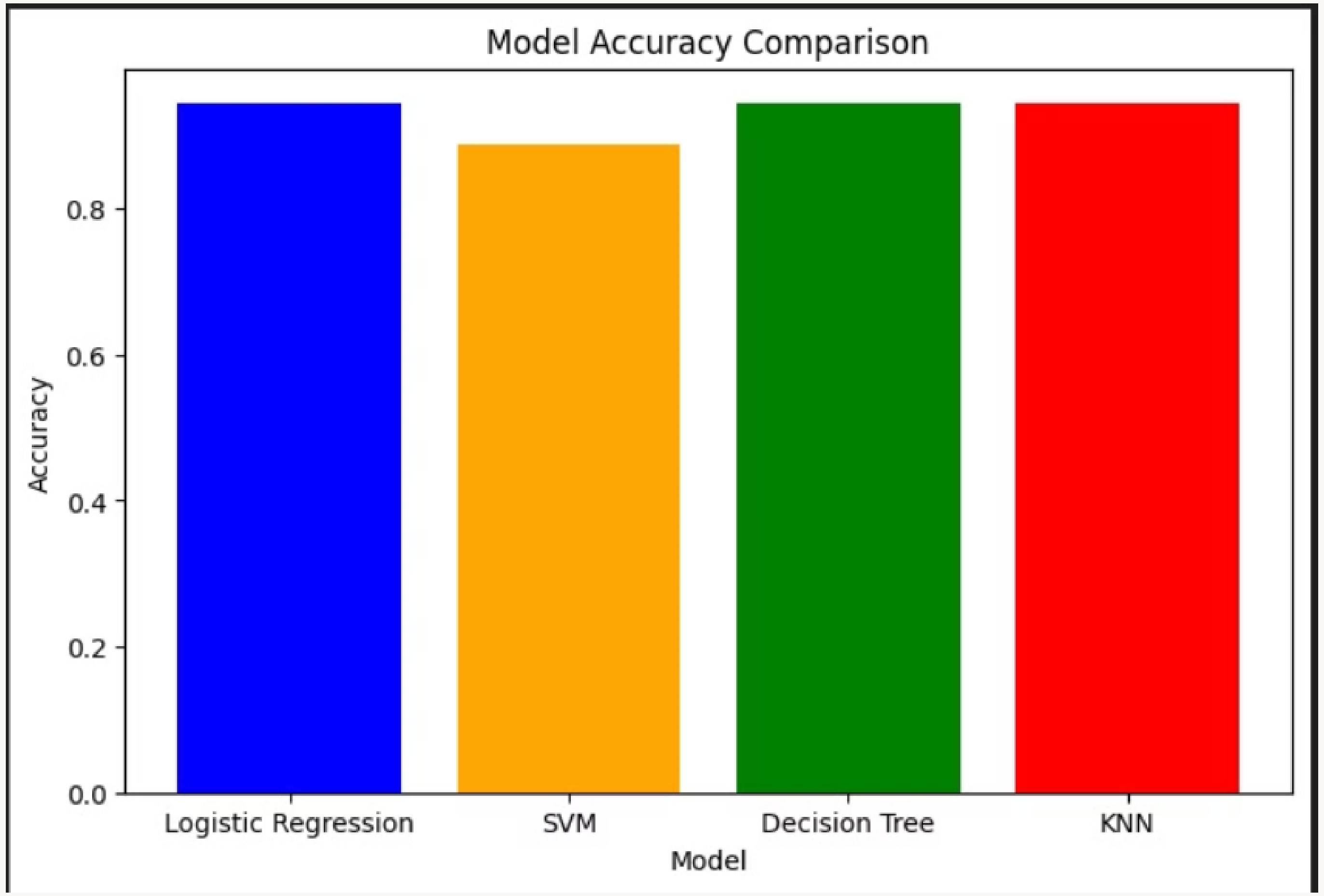
The dashboard provides a user-friendly interface for exploring SpaceX's operational trends and booster performance, offering clear insights into payload success rates and site efficiencies.



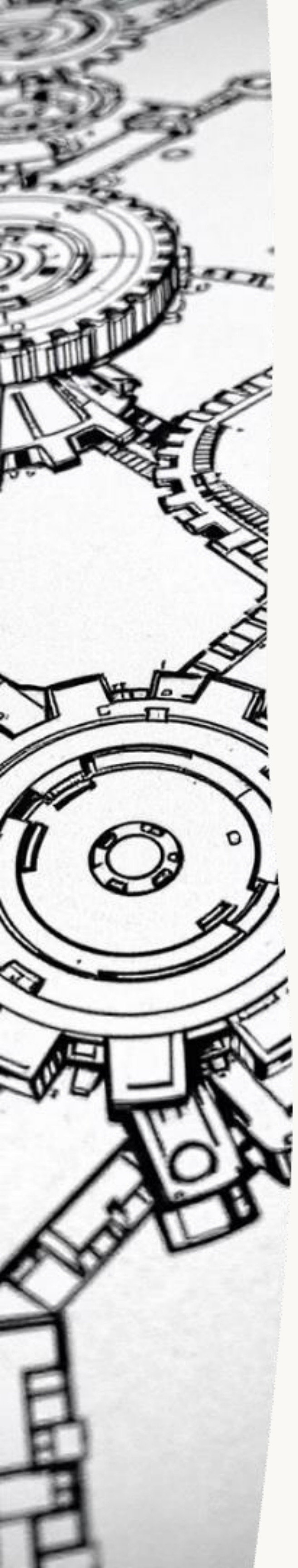
RESULT

Predictive Analysis (classification)

Bar Chart for Model Accuracy Comparison



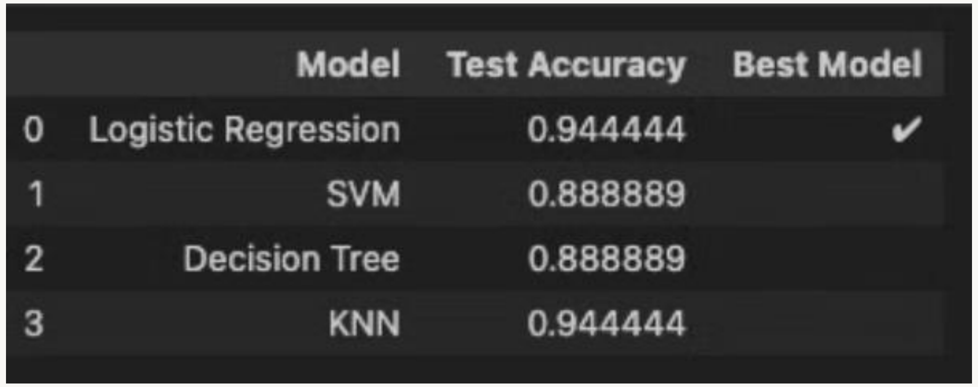
Emphasizes the strong performance of Logistic Regression and KNN, with consistent results across all metrics.



RESULT

Predictive Analysis (classification)

Test Accuracies and Best Model



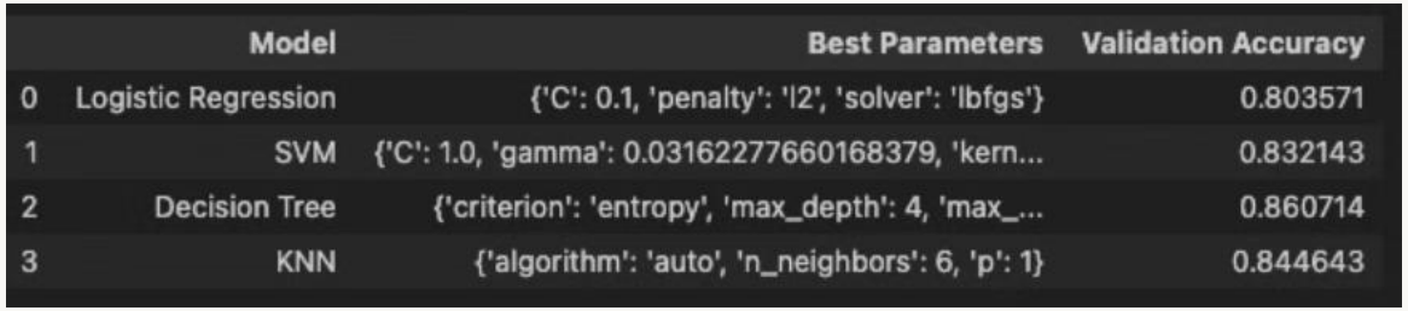
	Model	Test Accuracy	Best Model
0	Logistic Regression	0.944444	✔
1	SVM	0.888889	
2	Decision Tree	0.888889	
3	KNN	0.944444	

The table showcases test accuracy for four models: Logistic Regression, SVM, Decision Tree, and KNN.

Logistic Regression and KNN achieved the highest test accuracy (94.4%).

Logistic Regression is highlighted as the best model based on accuracy and consistency.

Hyperparameter Tuning Results



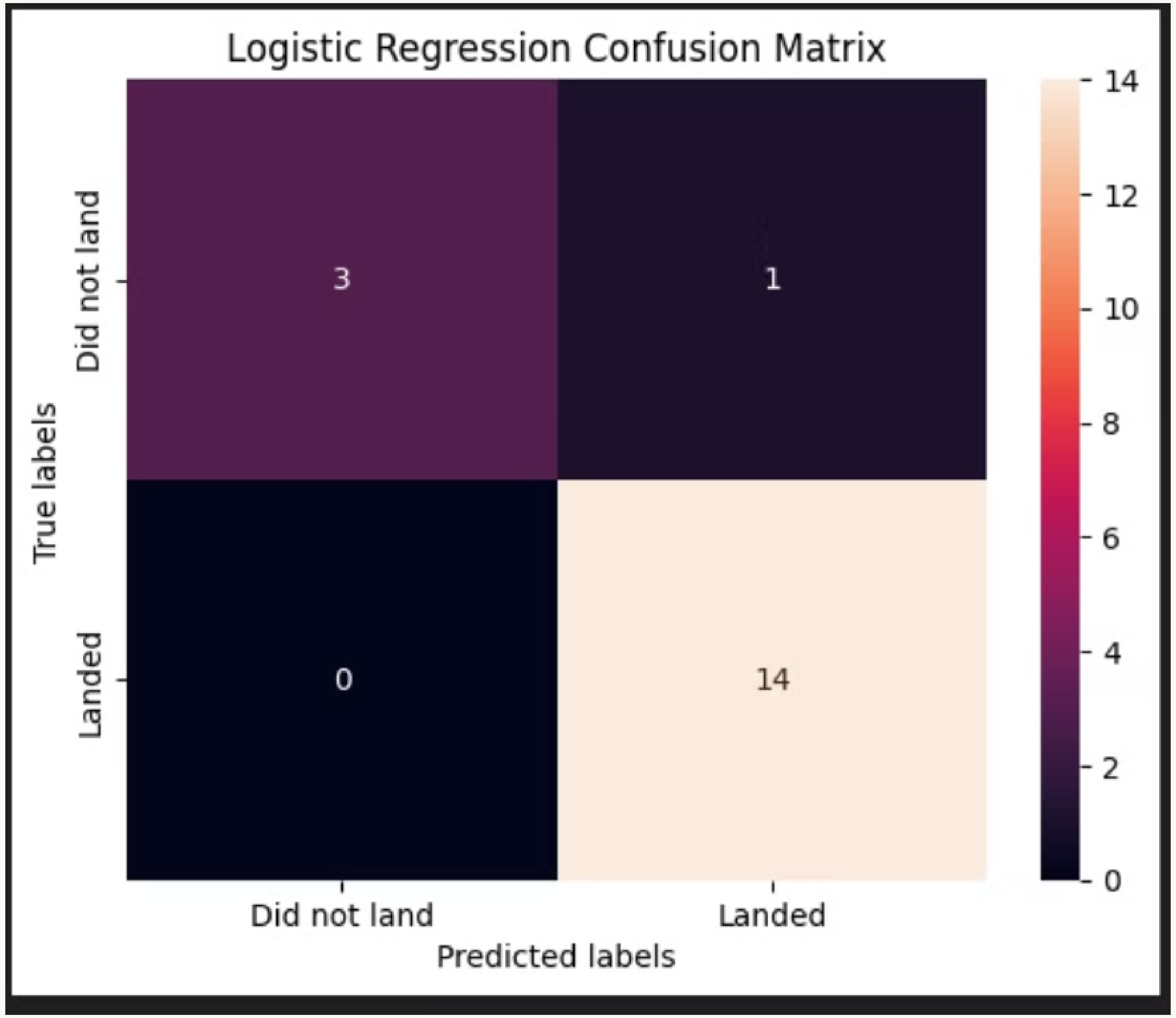
	Model	Best Parameters	Validation Accuracy
0	Logistic Regression	{'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}	0.803571
1	SVM	{'C': 1.0, 'gamma': 0.03162277660168379, 'kern...	0.832143
2	Decision Tree	{'criterion': 'entropy', 'max_depth': 4, 'max_...	0.860714
3	KNN	{'algorithm': 'auto', 'n_neighbors': 6, 'p': 1}	0.844643

Displays the best hyperparameters and validation accuracies achieved using GridSearchCV for each model.Decision Tree had the highest validation accuracy, but Logistic Regression proved more reliable on test data.

RESULT

•Predictive Analysis (classification)

Confusion Matrix for Logistic Regression

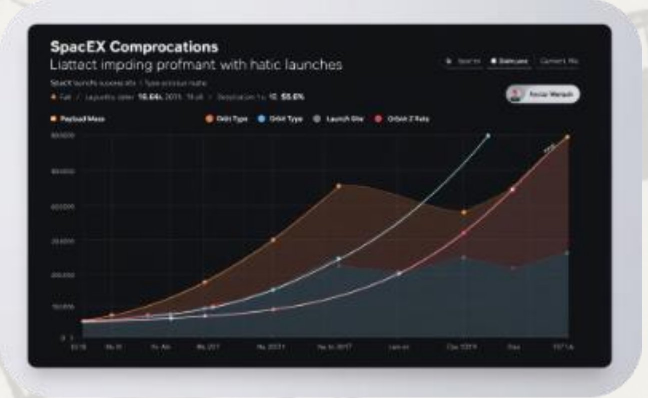


Visual representation of true positives, false positives, true negatives, and false negatives for the Logistic Regression model.

Highlights its reliability, with minimal false positives and high classification accuracy.

Conclusion: Data Science Insights and Impact

This project provided valuable insights into SpaceX launches and the predictive power of data:

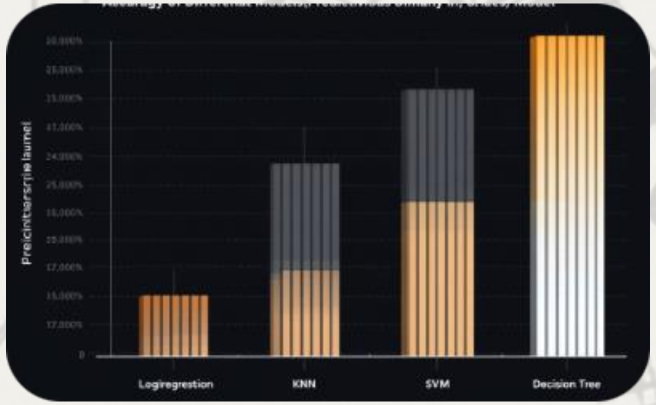


Key Findings

Payload mass, orbit type, and launch sites significantly influence launch outcomes. SpaceX's success rate has steadily improved, showcasing continuous operational enhancements.

Visualization Techniques

Interactive dashboards and maps facilitated the identification of launch success and failure patterns and trends across missions.



Predictive Modeling

Logistic Regression and KNN models achieved 94.4% test accuracy. Hyperparameter tuning significantly impacted model performance.

Project Takeaways

SpaceX's iterative learning from each launch enhances reliability and reduces costs. Predictive models are valuable for planning future missions and maintaining a competitive edge.

This project demonstrates the effectiveness of data science and machine learning in addressing complex challenges in space exploration.

APPENDIX

SpaceX Official Website

https://www.spacex.com For understanding SpaceX's mission, launches, and reusability efforts.

Dataset Sources

- Launch Data: SpaceX Launch Data on Kaggle (or the exact link if provided in your project source).
- Supplementary Data: IBM Skills Network Datasets

Dash and Plotly for Visualization

https://dash.plotly.com Guide to interactive dashboards used in the project.

Scikit-learn Documentation

https://scikit-learn.org For information on GridSearchCV, hyperparameter tuning, and classification algorithms.

SpaceX Falcon 9 Reusability

Why Reusability is Game-Changing

Predictive Analytics in Space Industry

Using Machine Learning for Space Exploration

Python Libraries Documentation

- Pandas: https://pandas.pydata.org
- NumPy: https://numpy.org
- Seaborn: https://seaborn.pydata.org

Author Profile

LinkedIn GitHub Repository

Thank You!

This presentation explored the use of data science in analyzing SpaceX launches, from data collection and processing to visualization and predictive modeling.