DATA SCIENCE PROJECT ON SPACEX LAUNCHES

ANALYZING SUCCESS FACTORS AND PREDICTIONS

Natalia Tkachenko
GitHub

04/12/24

EXECUTIVE SUMMARY

Why This Project Matters:

SpaceX has revolutionised space travel, slashing launch costs with its reusable *Falkon 9* rokets. 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.

But here's the catch: not every landing is successful. This project dives deep into the data to uncover the secrets behind successful landings.

This project dives deep into the data to uncover the secrets behind successful landings. 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

The Role of Data Science in SpaceX's Success

The success of *SpaceX's Falcon 9* rockets isn't just about engineering; it's also about the power of data. With every launch, massive amounts of data are generated, from payload characteristics to launch conditions and landing outcomes. This project applies **data science methodologies** to analyze these factors, **identify patterns**, and **predict** whether the *Falcon 9* first stage will land successfully.

By leveraging **machine learning**, data **visualization**, and **statistical analysis**, we aim to uncover actionable insights that can optimize operational efficiency and reduce costs further. This project highlights how data-driven decision-making plays a crucial role in modern space exploration, enabling *SpaceX* to stay ahead in the competitive commercial space industry.

Key Questions Explored

1. How do payload mass, orbit type, and launch site affect the likelihood of a successful landing?
2. Can trends in historical data reveal improvements in landing success rates over time?
3. Which machine learning algorithm is best suited for predicting first-stage landing outcomes?

Data Collection Methods

DATA COLLECTION AND PROCESSING METHODOLOGY

SpaceX REST API

- Accessed detailed launch data using endpoints like /rockets, /launchpads, /payloads, and /cores.
- Extracted booster details, payload information, orbit data, and landing outcomes for analysis.
- Transformed raw JSON responses into structured Pandas DataFrames.

Web Scraping

- Retrieved historical Falcon 9 and Falcon Heavy launch data from Wikipedia.
- Parsed HTML tables with BeautifulSoup to collect details like launch site, payload mass, orbit type, and customer information.
- Ensured consistency with preprocessed data by using a static version of the webpage.

DATA COLLECTION AND PROCESSING METHODOLOGY

2. 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).

Feature Engineering:

- Created additional columns like Class for binary classification (1 for successful landings, 0 for failures).
- Extracted and reformatted information (e.g., datetime fields, numeric conversions).

Normalization:

- Standardized categorical features using One-Hot Encoding.
- Reformatted data into 90 rows with 17 consistent features.

DATA COLLECTION AND PROCESSING METHODOLOGY

3. 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

1. 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**: For data manipulation and aggregation.
- **NumPy**: To perform mathematical computations.
- **SQL**: For complex queries and structured data analysis.

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

2. Interactive Data Visualization

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

Tools Used:

- **Matplotlib and Seaborn**: For scatter plots, bar charts, and time-series visualizations.
- **Folium**: For geographic data visualization and interactive mapping.
- **Plotly Dash**: For creating interactive dashboards.

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.
4.Interactive Dashboards were 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.s by launch site. •Analyze success probabilities visually.

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.

Key Steps

- **Data Preprocessing**:
  - Converted the Class column into a NumPy array for labels.
  - Standardized the feature dataset using StandardScaler to normalize values for better model performance.
  - Split the dataset into training (80%) and testing (20%) subsets using train_test_split.
- **Model Training and Hyperparameter Tuning**:
  - Utilized GridSearchCV for hyperparameter optimization across four machine learning algorithms:
    - **Logistic Regression**: Tuned parameters such as C, penalty, and solver.
    - **Support Vector Machine (SVM)**: Tuned kernel types, regularization C, and gamma.
    - **Decision Tree**: Optimized depth, split criteria, and minimum sample splits.
    - **K-Nearest Neighbors (KNN)**: Tuned the number of neighbors, algorithms, and distance metrics.
- **Evaluation**:
  - Measured performance using metrics like accuracy, F1-score, and confusion matrix analysis.
  - Compared results across all models to determine the best-performing method.

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

EDA with SQL results

Average and Total Payload Mass

Total_Payload_Mass
48213

Average_Payload_Mass
2928.4

The average payload mass is calculated to be 2928.4 kg, and the total payload mass across all missions is 48,213 kg

Unique Launch Sites

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

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

RESULTS

EDA with SQL results

Landing Outcomes

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

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 visualization

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.

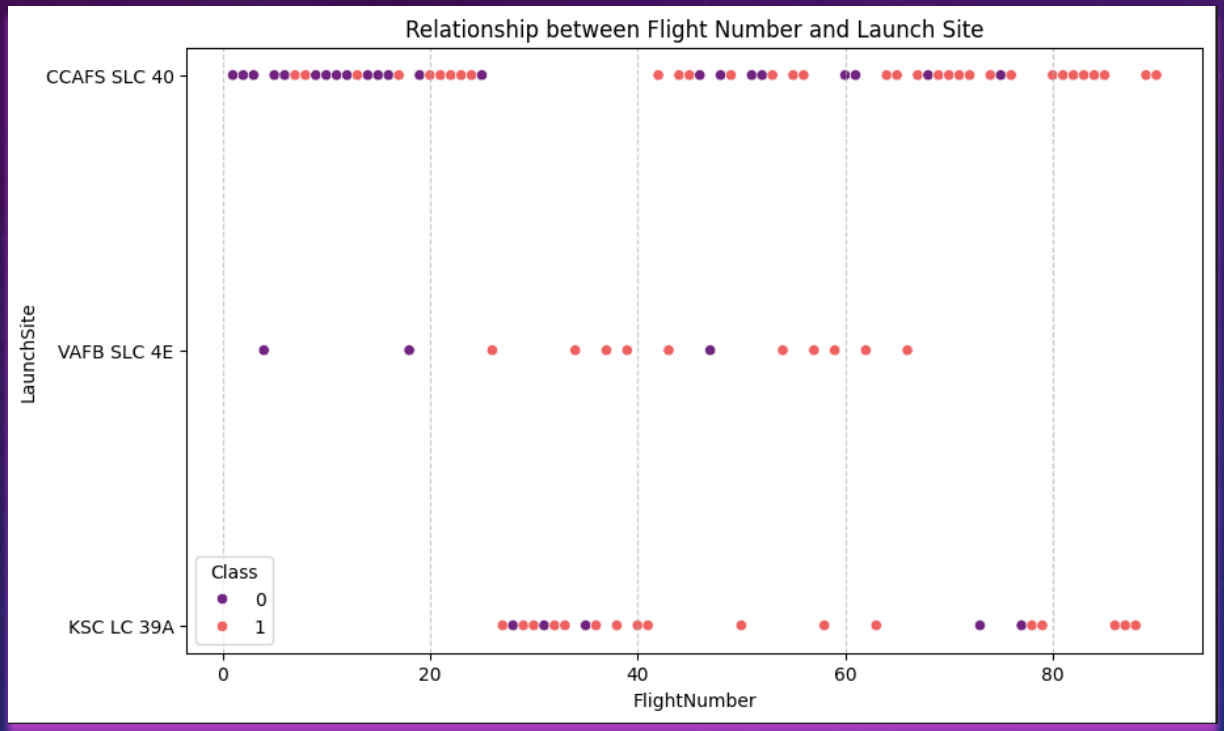
Key findings

- 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.

RESULTS

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

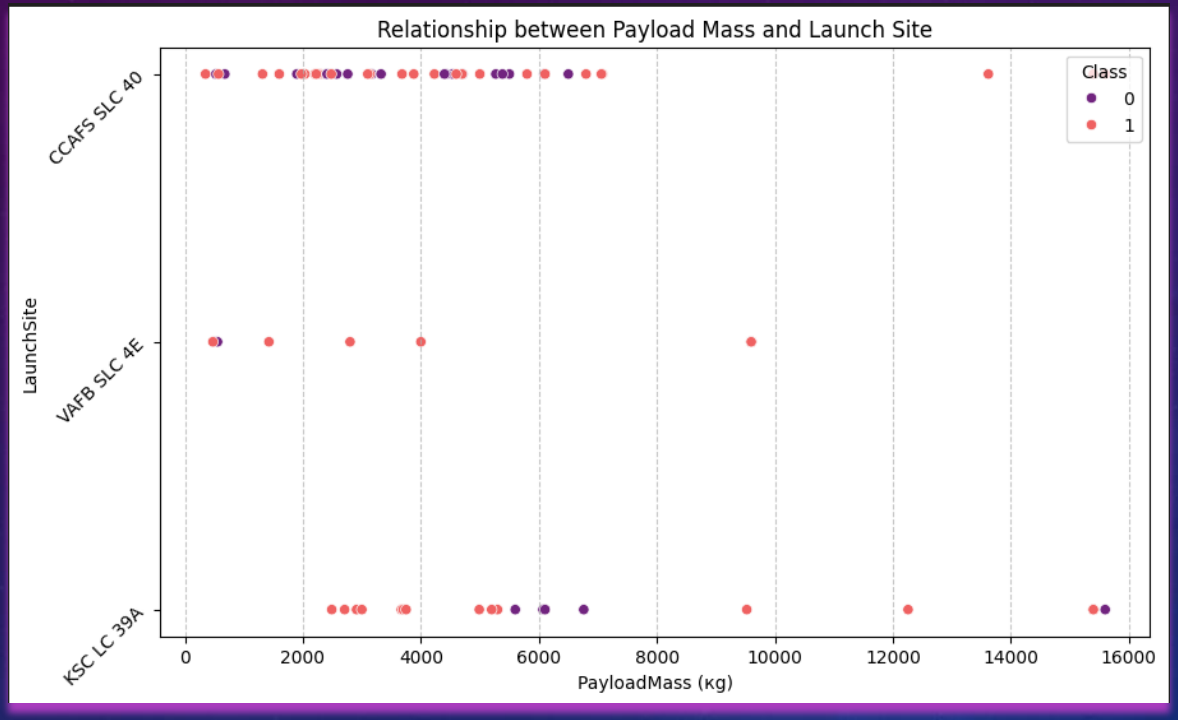
CCAFS SLC-40.**KSC LC-39A**: This site has consistently successful outcomes for most launches, especially at higher flight numbers.

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

RESULTS

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.

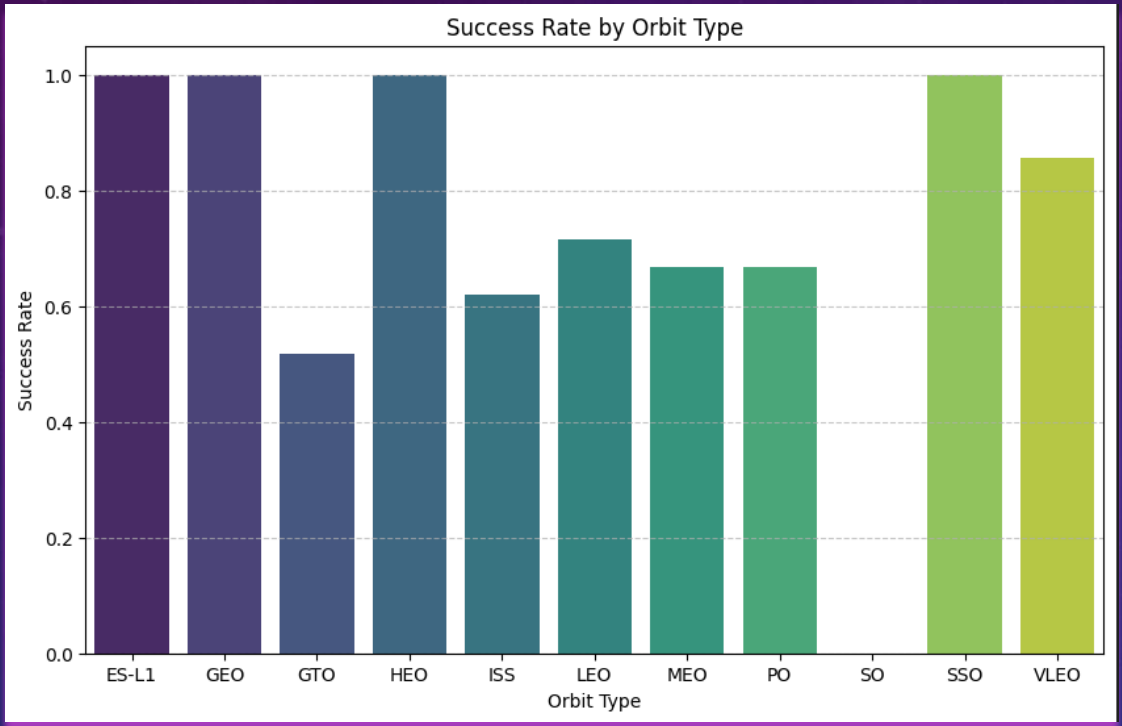
KSC LC-39A: This site is notable for handling larger payloads (up to 16,000 kg). Successful outcomes (1) dominate, especially for higher payload masses.

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.

RESULTS

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.

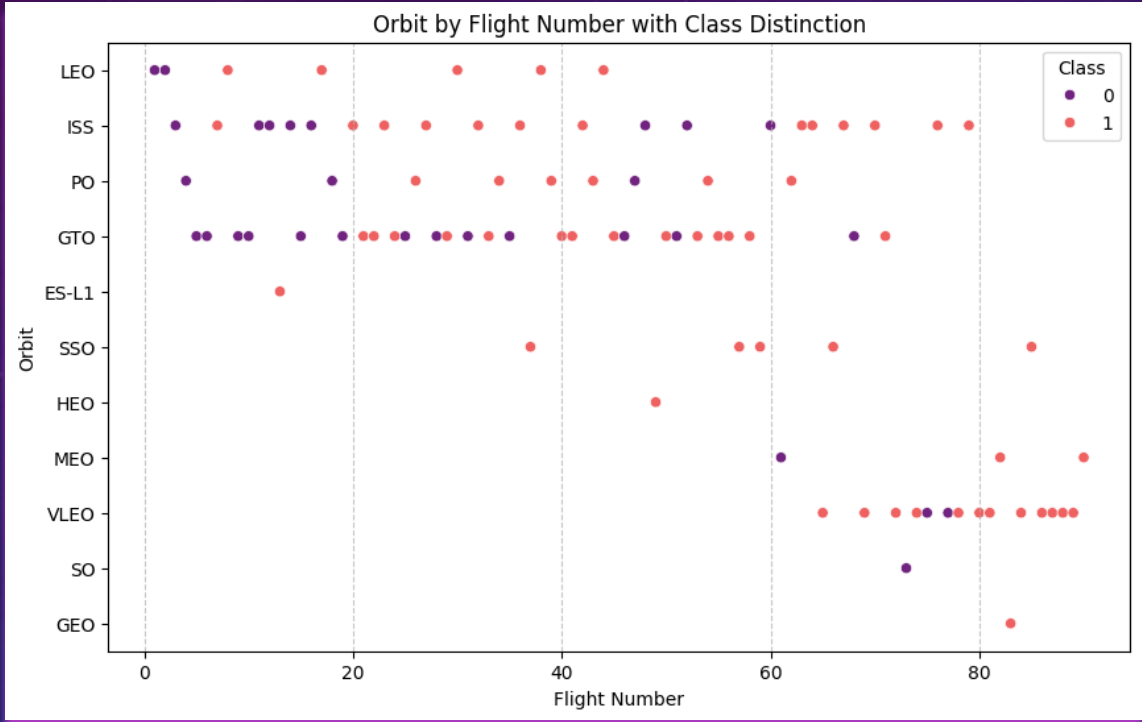
ISS and VLEO Orbits: These orbits exhibit a slightly better-than-average success rate but do not reach the reliability of orbits like GEO or SSO.

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

RESULTS

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:

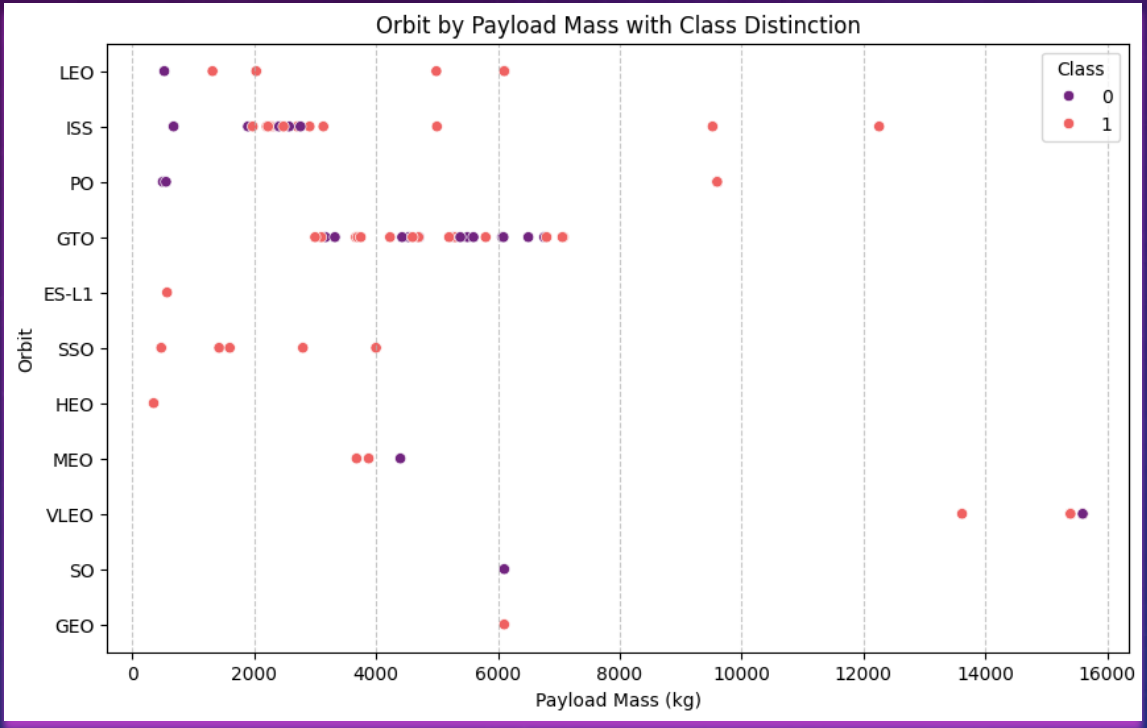
- **LEO and ISS**: Mixed results with increasing success in later flights.
- **GTO**: Balanced mix of successes and failures, showing challenges in this orbit.
- **Rare orbits (GEO, ES-L1, SSO)**: Mostly successful, reflecting specialized missions.

This visualization suggests SpaceX's learning curve and technical advancements as flight numbers increase, leading to improved reliability and success rates across most orbits.

RESULTS

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.

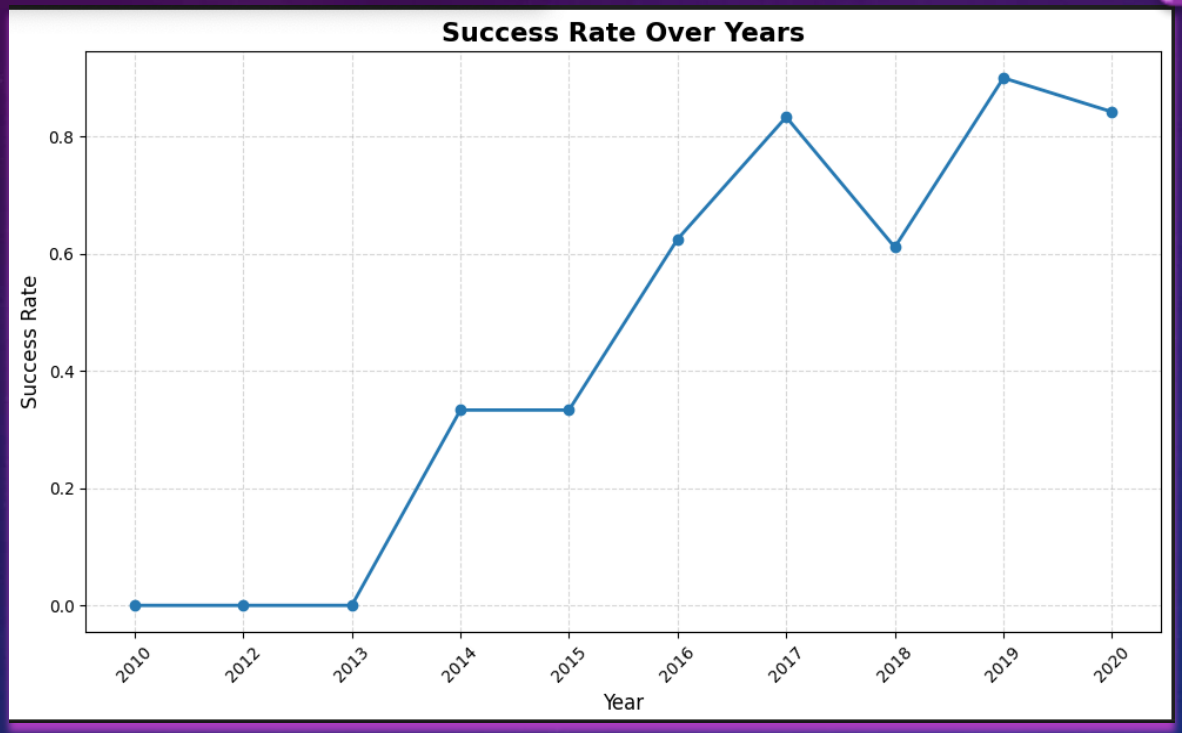
Specialized orbits (SSO, ES-L1, VLEO): Predominantly successful, likely due to specific mission designs.

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

RESULTS

EDA with visualization

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



2010-2012: No successful landings, with a 0% success rate.

2013-2014: Gradual progress, reaching a 40% success rate by 2014.

2015-2016: Steady growth, achieving 60% success by 2016.

2017-2018: Temporary dip in 2017, followed by recovery and over 80% success in 2018.

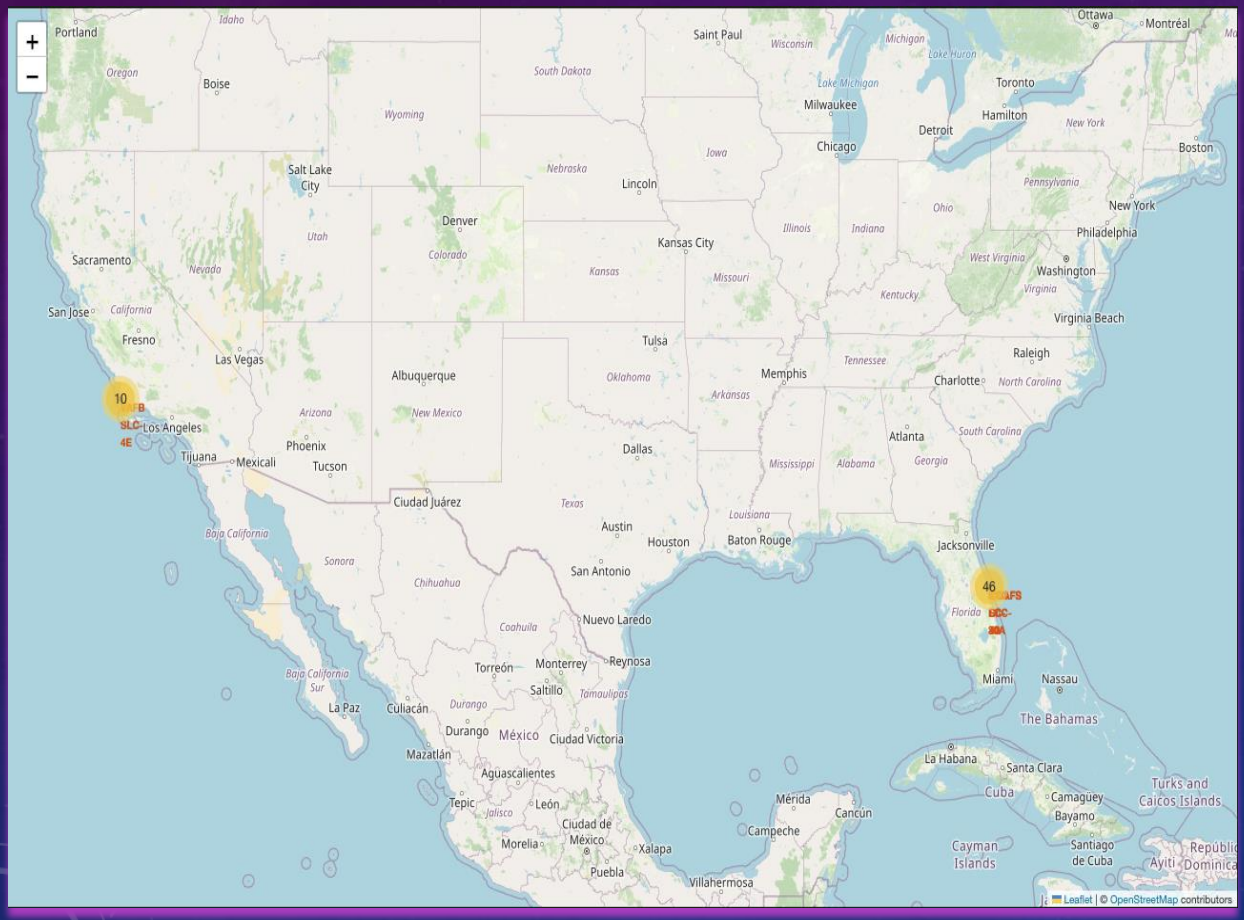
2019-2020: Consistently high success rates above 80%, reflecting operational maturity.

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.

RESULTS

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.

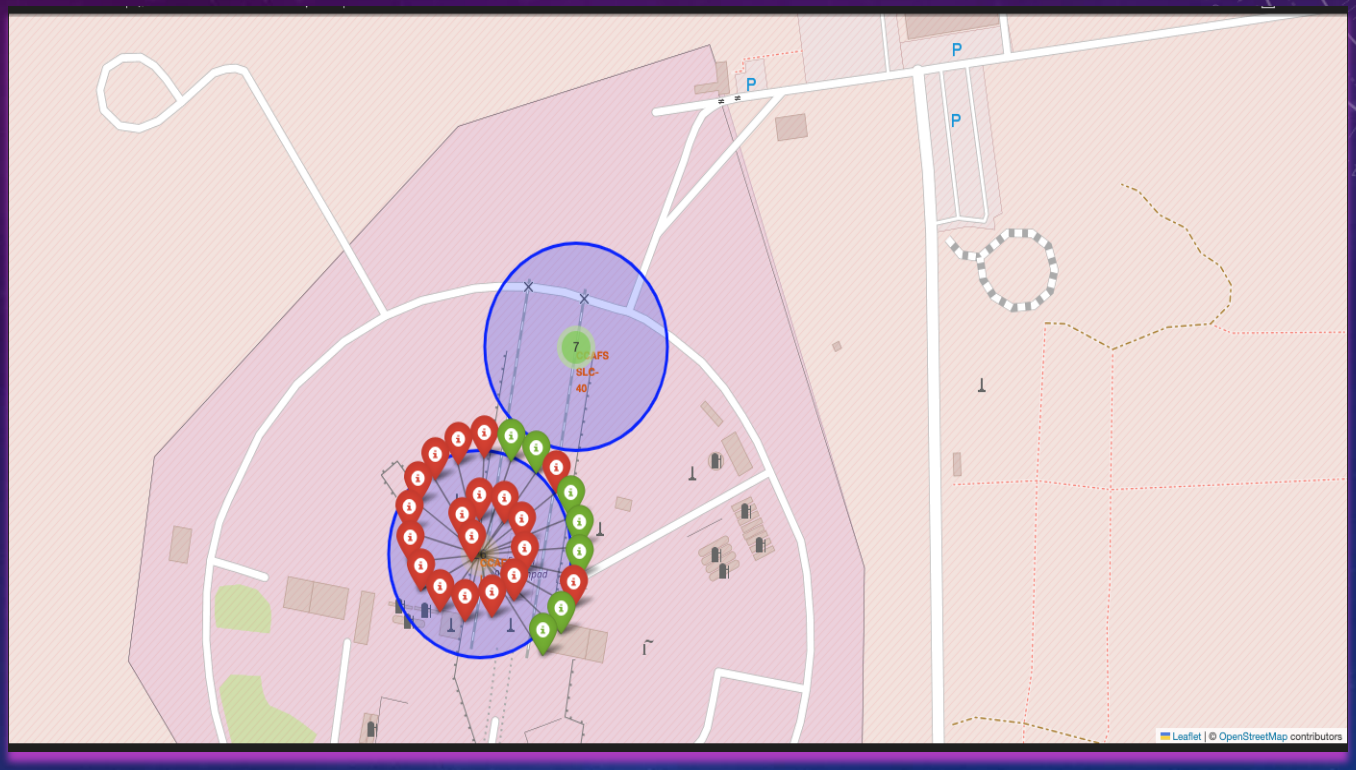
The map reflects the functional differentiation of sites, such as VAFB for polar orbits and Florida sites for equatorial orbits, showcasting SpaseX's strategic considerations for optimal operations.

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

RESULTS

Interactive map with *Folium*

Success and Failure Visualization

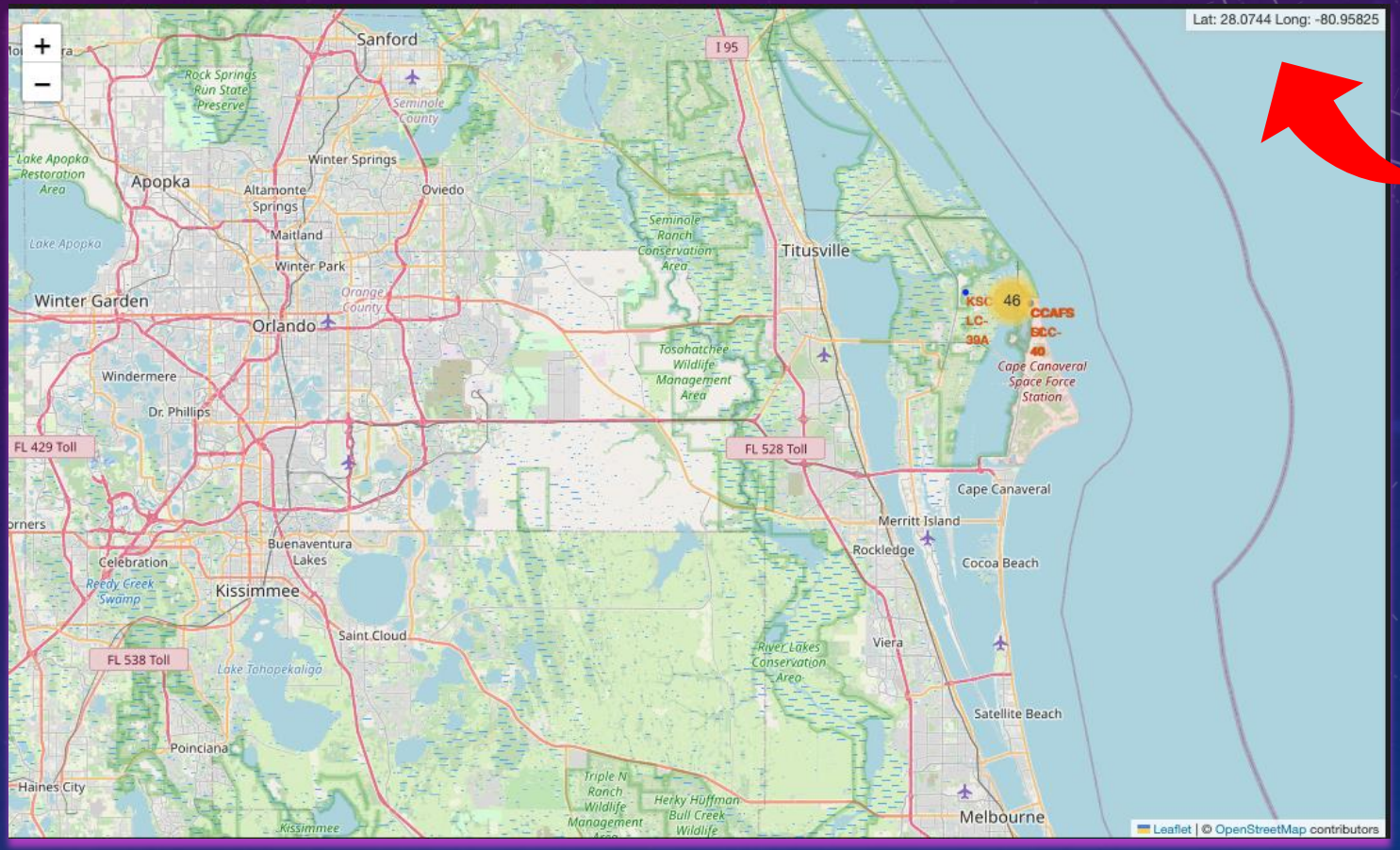


Marker clusters depict individual launches with green markers for successes and red for failures, enabling a quick assessment of site-specific performance trends.

RESULTS

Interactive map with *Folium*

Interactive Features

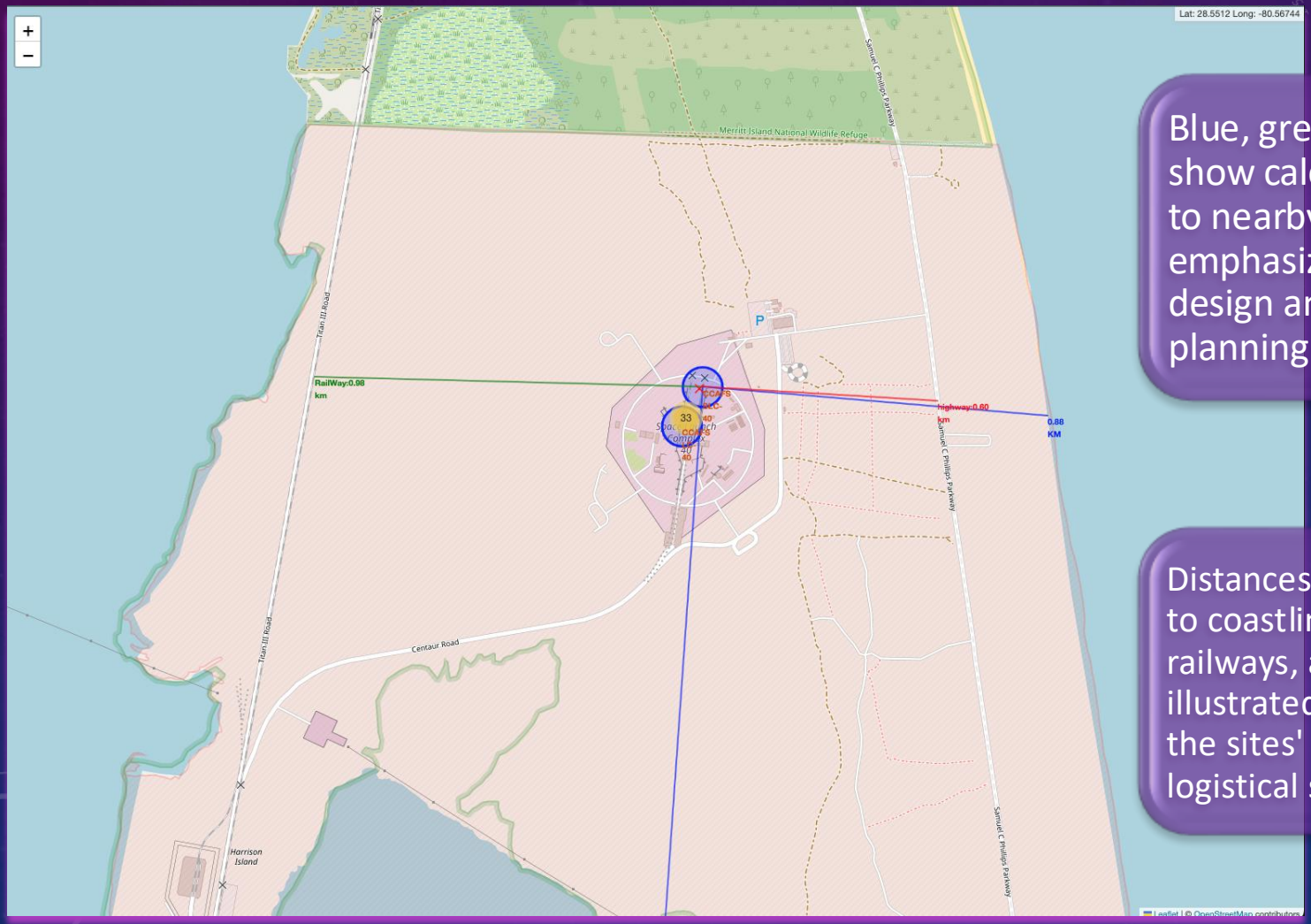


Interactive elements, such as mouseover coordinates and pop-ups with detailed site information, enhance user engagement and exploration.

RESULTS

Interactive map with *Folium*

Interactive Features

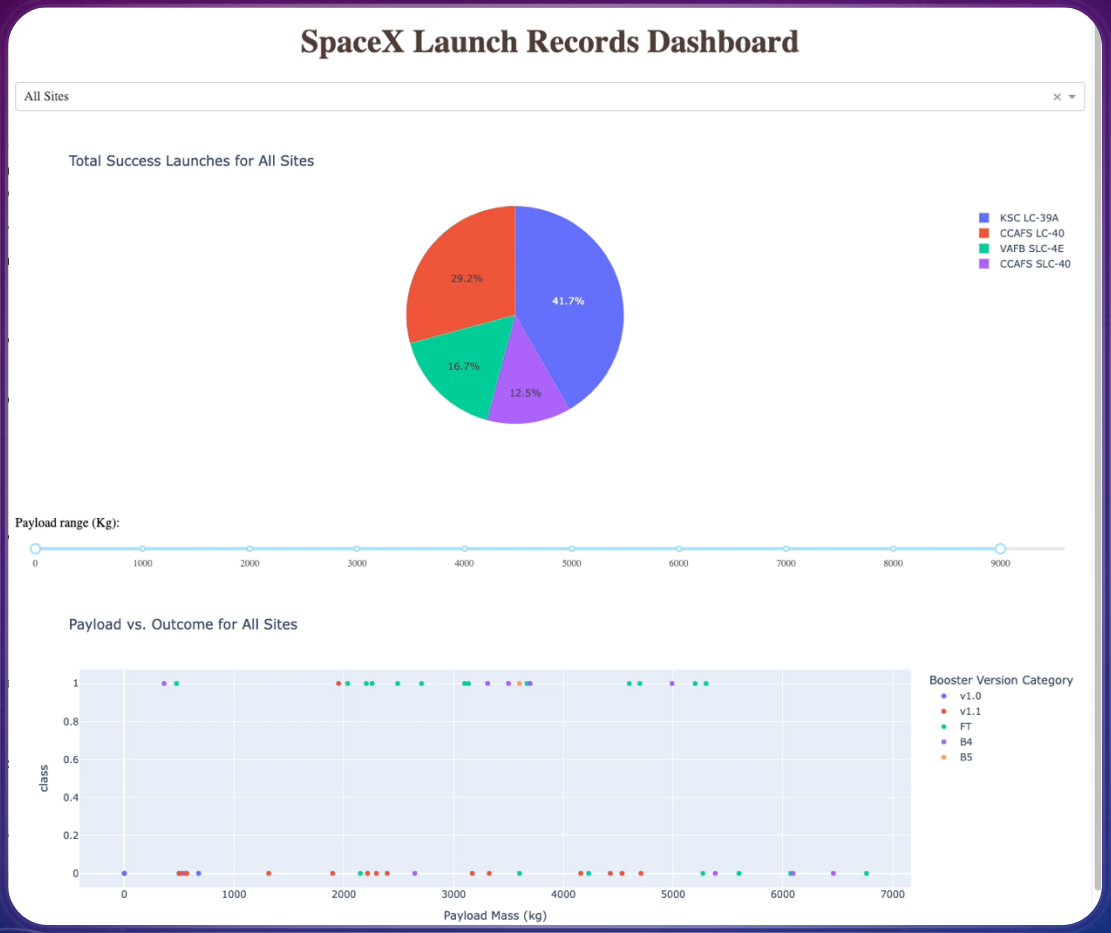


Blue, green, and red lines show calculated distances to nearby features, emphasizing the strategic design and operational planning of each site.

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

RESULTS

Plotly Dash 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:

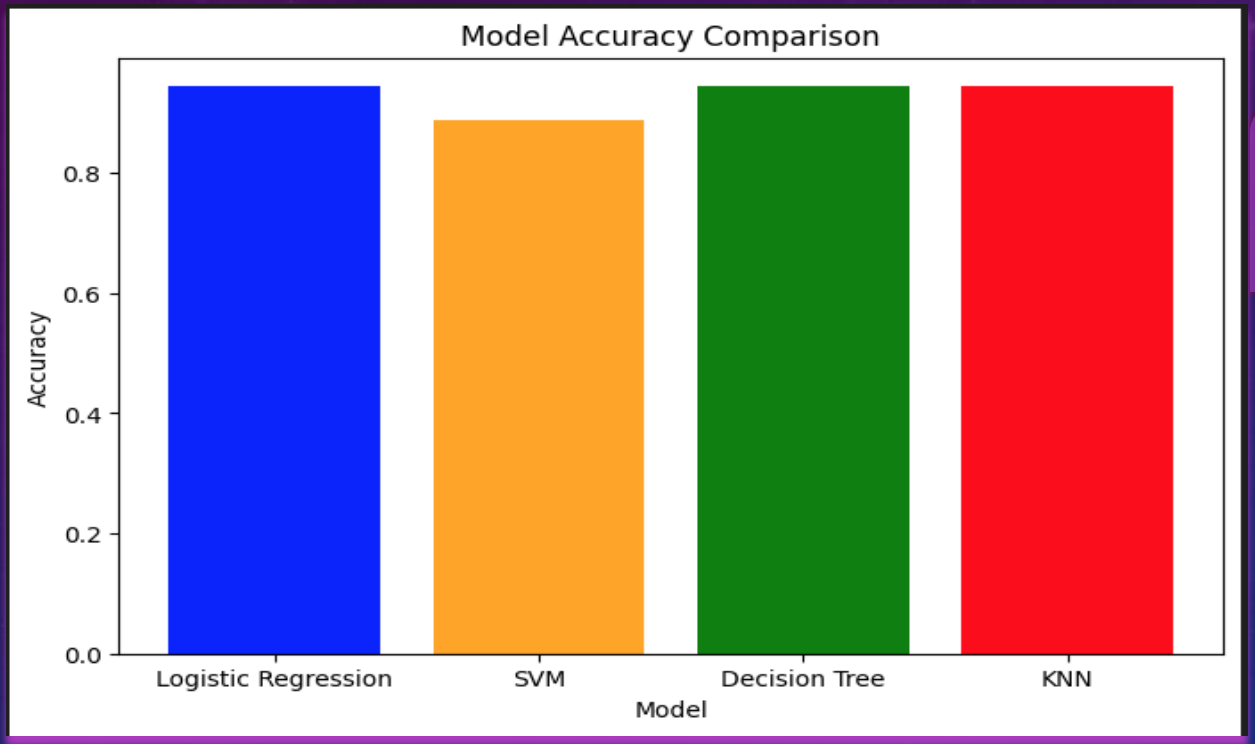
Dropdown menu for site selection and a slider for payload range adjustment.

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.

RESULTS

Predictive Analysis (classification)

Bar Chart for Model Accuracy Comparison



A visual comparison of test accuracies for all models.

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

RESULTS

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

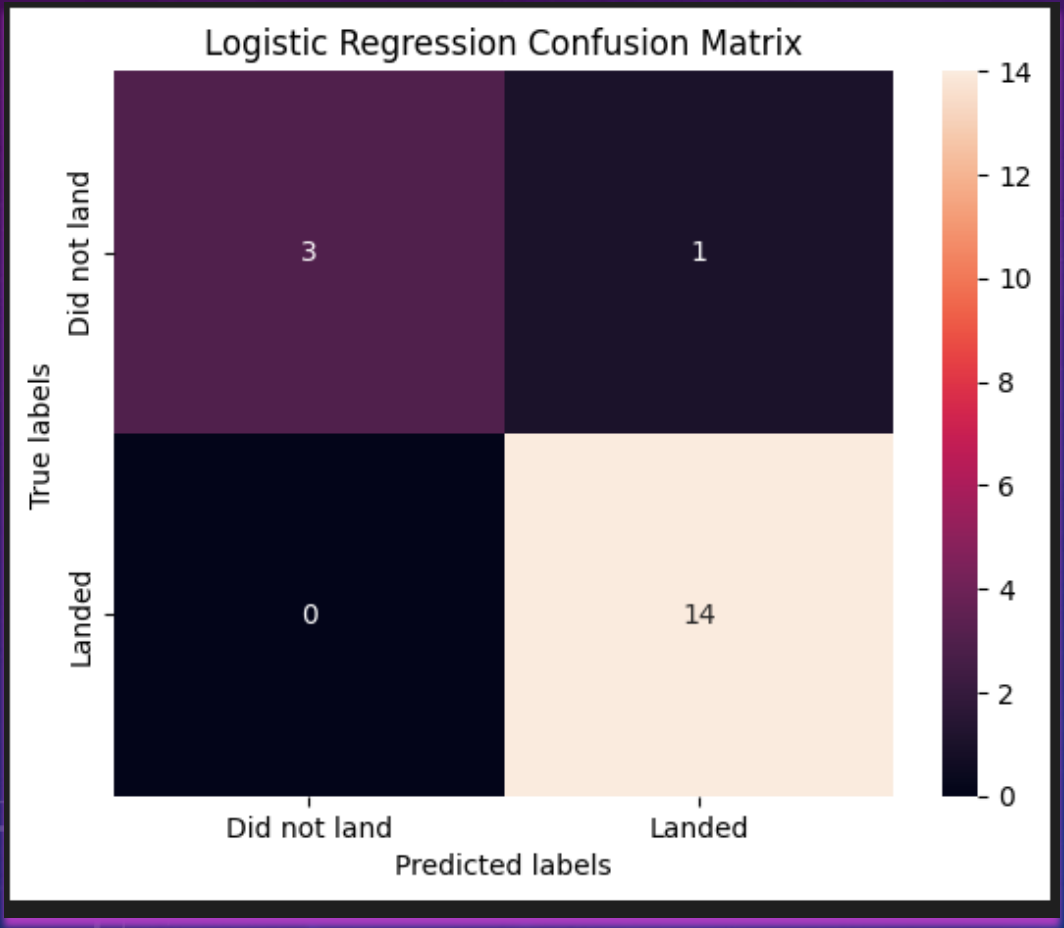
	Model	Best Parameters	Validation Accuracy
0	Logistic Regression	{'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}	0.803571
1	SVM	{'C': 1.0, 'gamma': 0.03162277660168379, 'kern...	0.832143
2	Decision Tree	{'criterion': 'entropy', 'max_depth': 4, 'max_...	0.860714
3	KNN	{'algorithm': 'auto', 'n_neighbors':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

RESULTS

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

This project gave us some fascinating insights into SpaceX launches and how data can help predict success:

What We Learned:

- Payload mass, orbit type, and launch sites play a big role in launch outcomes.
- SpaceX has clearly gotten better over time, with higher success rates showing how much they've improved.

How We Showed It:

- Interactive maps and dashboards helped us see where and why launches succeeded or failed.
- Visualizations made it easy to spot trends and patterns across different missions.

Modeling Success:

- Logistic Regression and KNN came out on top, with both achieving a test accuracy of 94.4%.
- Hyperparameter tuning showed how small tweaks can make a big difference in model performance.

Takeaways:

- SpaceX's focus on learning from each launch is paying off in reliability and cost savings.
- Predictive models like these are great tools for planning future launches and staying ahead of the competition.

In short, this project was a great example of how data science and machine learning can help solve real-world challenges—even ones as big as space exploration!

APPENDIX

Resources Used in the Project

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

2.Dataset Sources:

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

3.Dash and Plotly for Visualization:
https://dash.plotly.com
Guide to interactive dashboards used in the project.

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

Python Libraries Documentation:

Pandas: https://pandas.pydata.org

NumPy: https://numpy.org

Seaborn: https://seaborn.pydata.org

Author Profile:

- LinkedIn Profile
- GitHub Repository