IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Author: Ng Chung Pak
Date: 09th Jan 2025

Outline

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

Executive Summary: Methodology

Data Collection Integration:

- Integrated data from SpaceX API and Wikipedia
- Created comprehensive launch database
- Developed automated data collection pipeline

Data Analysis Process

- Cleaned and preprocessed launch data
- Conducted exploratory data analysis
- Built interactive visualization dashboards
- Developed predictive machine learning models

Executive Summary: Key Findings

Launch Success Factors

- Site location impacts success by up to 17%
- Optimal payload: 2000-5000 kg
- 85% success rate since 2013

Model Results

- Landing prediction accuracy: 84%
- Identified top 3 success indicators
- Cost estimation framework validated

Executive Summary: Business Impact

Strategic Value

- Competitive pricing enabled
- Launch costs optimized
- Data-driven decisions supported

Next Steps

- Real-time prediction system
- Enhanced cost modeling
- Expanded feature analysis

Introduction: Background

Commercial Space Race

- Space travel becoming more affordable
- Multiple companies entering market
- SpaceX leads in cost efficiency

SpaceX Advantage

- Falcon 9 launches: $62M vs. $165M industry standard
- Key differentiator: Reusable first stage
- 80%+ of launch costs in first stage

Introduction: Project Context

Business Challenge

- New company Space Y aims to compete with SpaceX
- Need to understand launch cost driver
- First stage recovery critical for profitability

Project Goals

1. Predict first stage landing success
2. Analyze key success factors
3. Build cost estimation framework

Section 1

Methodology

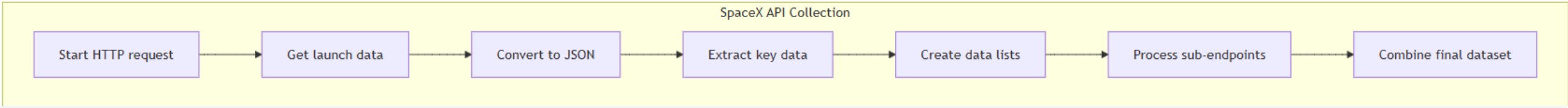
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

Flowchart of SpaceX API calls

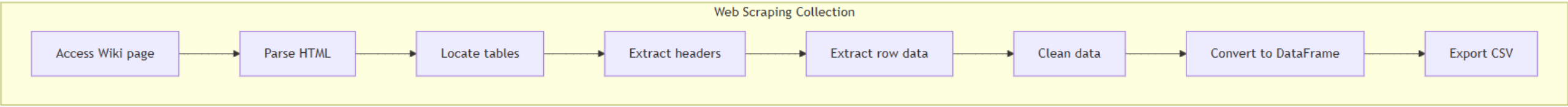


GitHub URL of the SpaceX API calls notebook:

https://github.com/kennethPakChungNg/spacex-launch-prediction/blob/main/notebooks/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Flowchart of Web Scraping Collection



GitHub URL of the SpaceX web scraping notebook:

https://github.com/kennethPakChungNg/spacex-launch-prediction/blob/main/notebooks/jupyter-labs-webscraping.ipynb

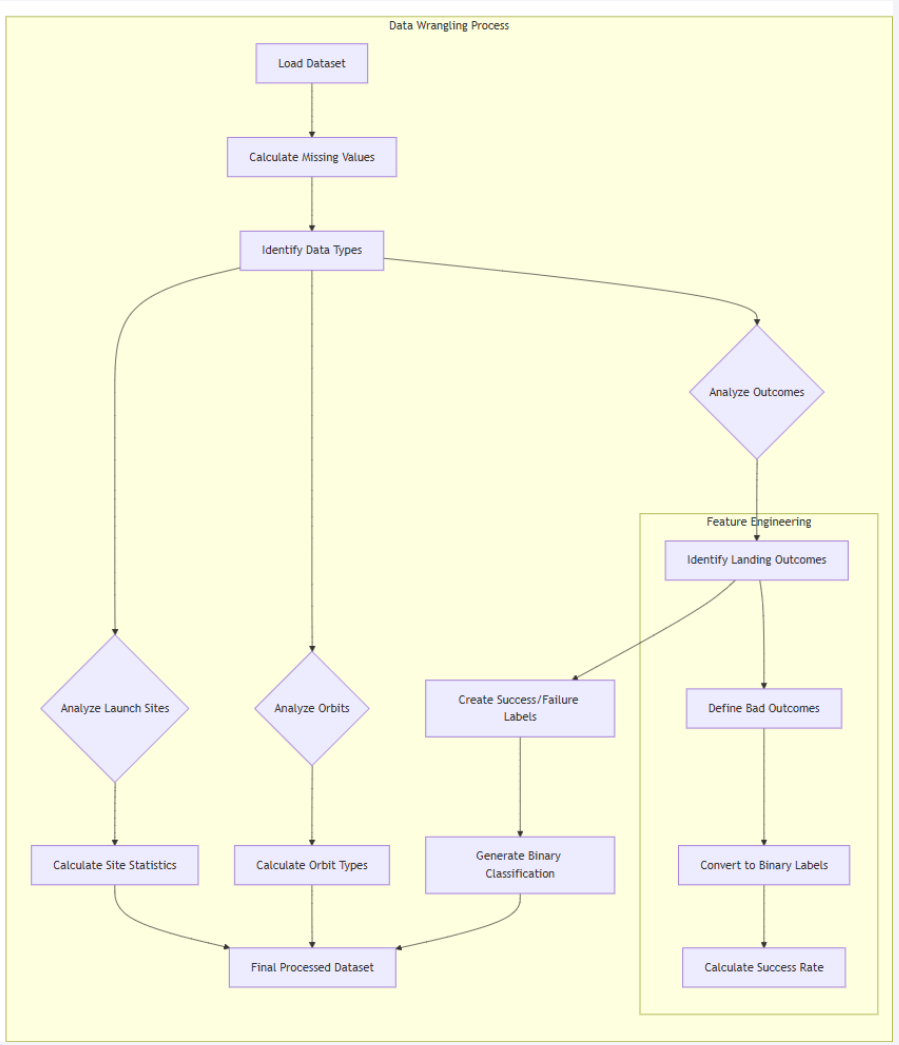
Data Wrangling

- Filtering Falcon 9 launches
- Handling missing values
- Feature engineering

GitHub URL of the SpaceX Data Wrangling notebook:

https://github.com/kennethPakChungNg/spacex-launch-prediction/blob/main/notebooks/labs-jupyter-spacex-Data%20wrangling.ipynb

Flowchart of Data Wrangling



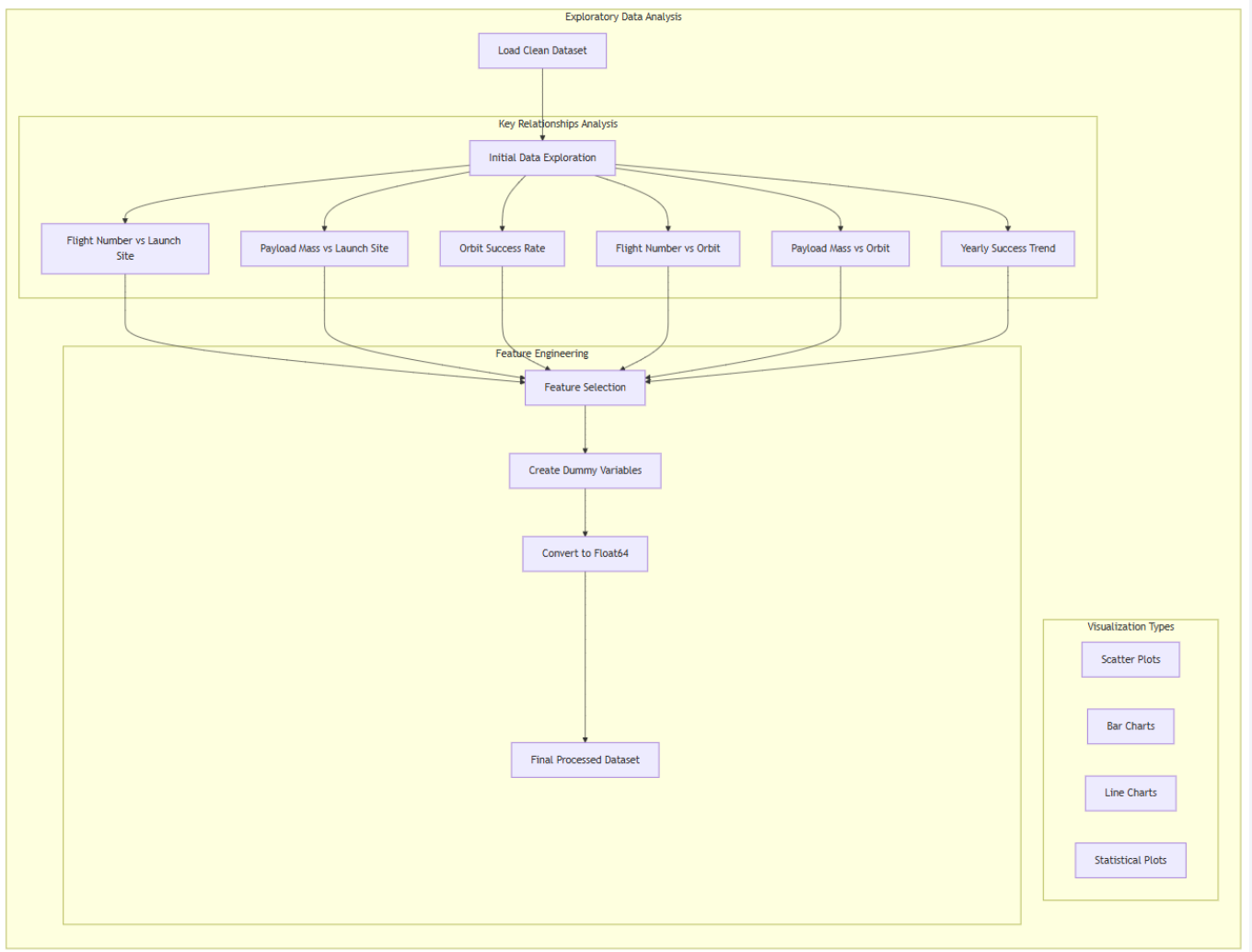
EDA with Data Visualization

- Flight Number vs Payload Mass - Shows launch success trends over time with payload impact
- Launch Site Analysis - Reveals site-specific success patterns
- Orbit Type Success Rates - Bar chart showing success rates by orbit type
- Flight Number/Orbit Relationships - Scatter plots revealing correlations
- Payload Mass/Orbit Analysis - Impact of payload on different orbits
- Yearly Success Trend - Line chart showing improvement over time

GitHub URL of the SpaceX Data Visualization notebook:

https://github.com/kennethPakChungNg/spacex-launch-prediction/blob/main/notebooks/edadataviz.ipynb

Flowchart of Data Visualization



EDA with SQL

Key SQL Queries Performed:

- **Launch Site Analysis**
  - Identified unique launch sites
  - Filtered sites starting with 'CCA'
- **Payload Analysis**
  - Total NASA payload mass
  - Average mass for F9 v1.1 boosters
  - Maximum payload carriers
- **Landing Outcomes**
  - First successful ground pad landing
  - Success/failure mission counts
  - Drone ship landing success with payload constraints
- **Temporal Analysis**
  - Monthly analysis for 2015
  - Landing outcome rankings 2010-2017
  - Success trends over time

GitHub URL of the SpaceX EDA with SQL notebook:

https://github.com/kennethPakChungNg/spacex-launch-prediction/blob/main/notebooks/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

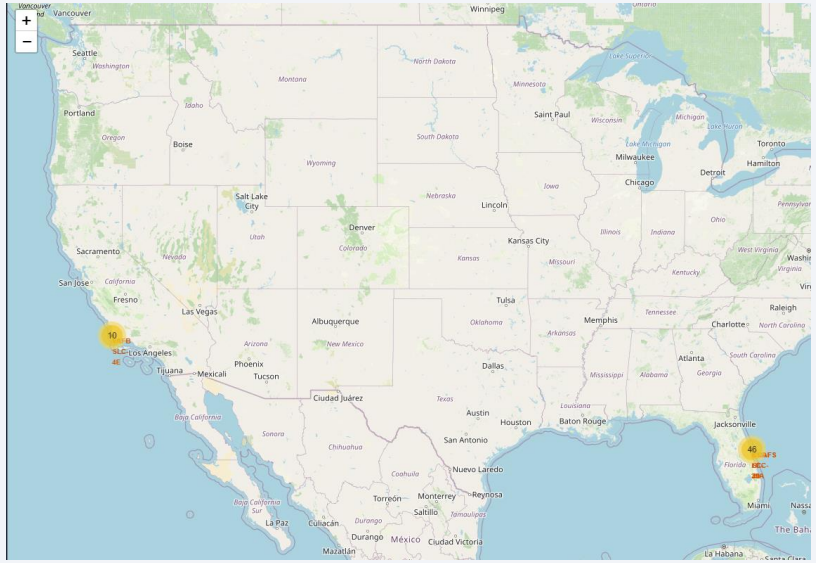
Launch Site Locations Overview

Key Features:

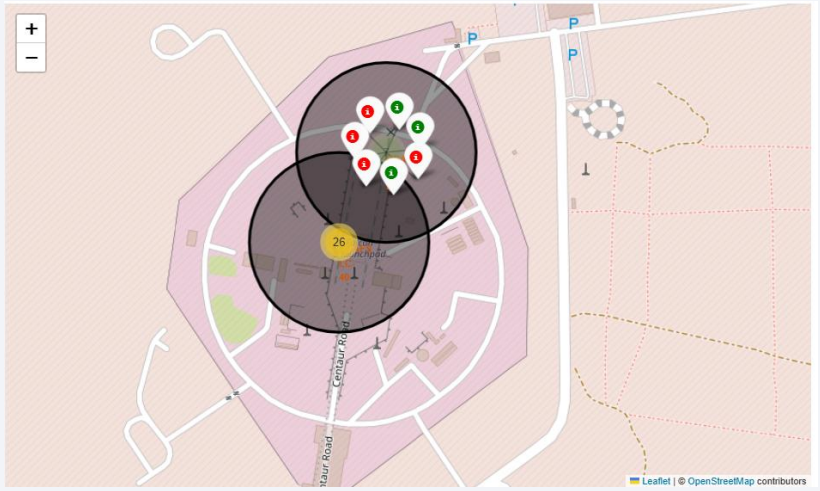
- Launch sites marked with circle markers
- Color-coded launch outcomes (green=success, red=failure)
- Marker clusters showing launch frequency
- Interactive zoom capabilities

Findings:

- Strategic coastal positioning
- Concentrated in Florida and California
- Clear success rate patterns by location



Map 1: Overview Map



Map 2: Detailed View

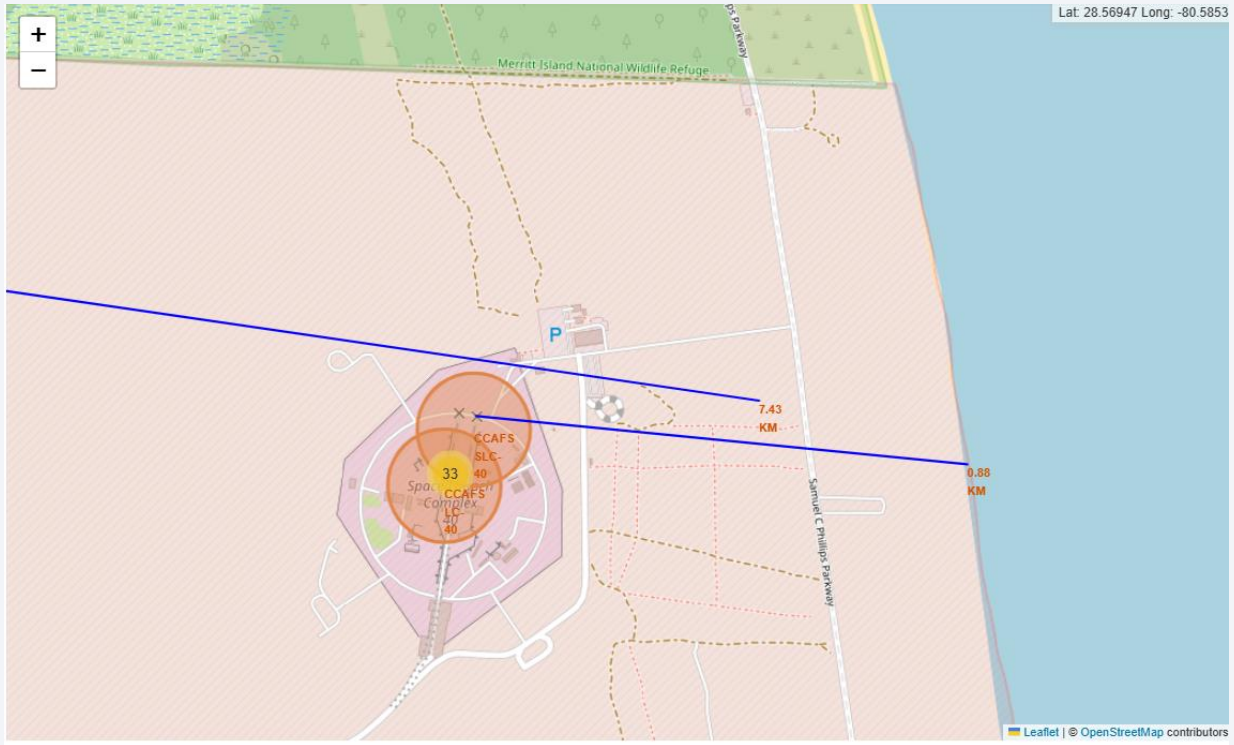
Build an Interactive Map with Folium

Infrastructure Access:

- Coastline: 5-8 km
- Railways: ~10 km
- Highways: ~12 km
- Urban areas: Safe distance maintained

Impact on Success:

- Coastal proximity enables safe booster returns
- Transportation infrastructure supports operations
- Urban distance ensures safety compliance



Build a Dashboard with Plotly Dash

Summary of Plots and Interactions in the SpaceX Launch Prediction Dashboard

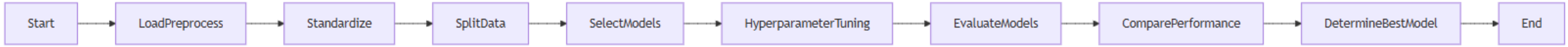
- **Launch Site Dropdown:**
  - **Purpose:** Allows users to select a specific launch site or view all sites.
  - **Why Added:** Provides flexibility in analyzing success rates across different launch sites.
- **Pie Chart:**
  - **Purpose:** Displays the proportion of successful launches by site or success vs. failure for a selected site.
  - **Why Added:** Effectively shows the distribution of successful and failed launches, highlighting proportions at a glance.
- **Payload Slider:**
  - **Purpose:** Enables users to filter data based on payload mass range.
  - **Why Added:** Payload mass significantly impacts launch success, allowing users to explore this relationship dynamically.
- **Scatter Plot:**
  - **Purpose:** Illustrates the correlation between payload mass and launch success, colored by booster version category.
  - **Why Added:** Provides a deeper analysis of how payload mass and booster versions influence launch outcomes.

Explanation of Choices:

- **Interactivity:** The dropdown and slider empower users to explore data from various perspectives, enhancing engagement and insights.
- **Data Visualization:**
  - The pie chart offers an overview of success rates, making it easy to compare sites.
  - The scatter plot allows for detailed analysis of the relationship between payload mass and success, with additional insights from booster version categories.

Github: *https://github.com/kennethPakChungNg/spacex-launch-prediction/blob/main/spacex_dash_app.py*

Predictive Analysis (Classification)

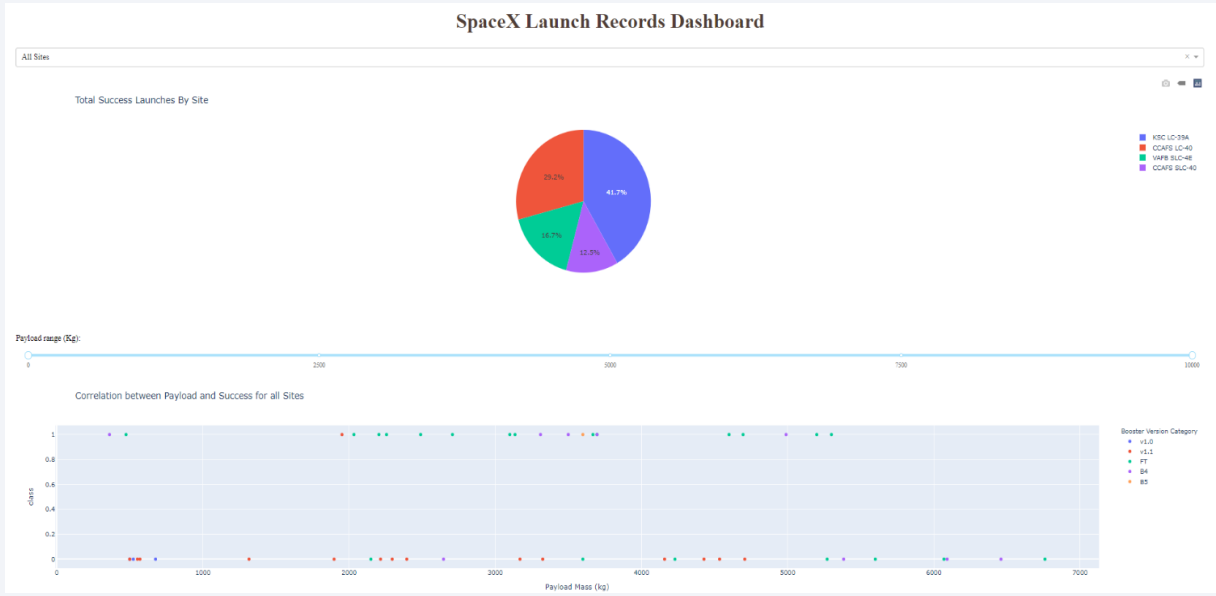


Model Development Process

- **Data Loading and Preprocessing**: Loaded and preprocessed the dataset to ensure it is ready for modeling.
- **Feature Standardization**: Standardized the features to normalize the data distribution.
- **Data Splitting**: Split the data into training and test sets to evaluate model performance.
- **Model Selection**: Selected four models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- **Hyperparameter Tuning**: Tuned hyperparameters using GridSearchCV to optimize model performance.
- **Model Evaluation**: Evaluated the models on the test data to assess their accuracy.
- **Performance Comparison**: Compared the performances of the models to identify the best one.
- **Best Model Selection**: Selected the Decision Tree model as the best performing model based on accuracy.

Based on the evaluations, I determined that the Decision Tree model performed the best with the highest cross-validation and test accuracy. This structured approach ensured that I selected the most effective model for the task.

Results

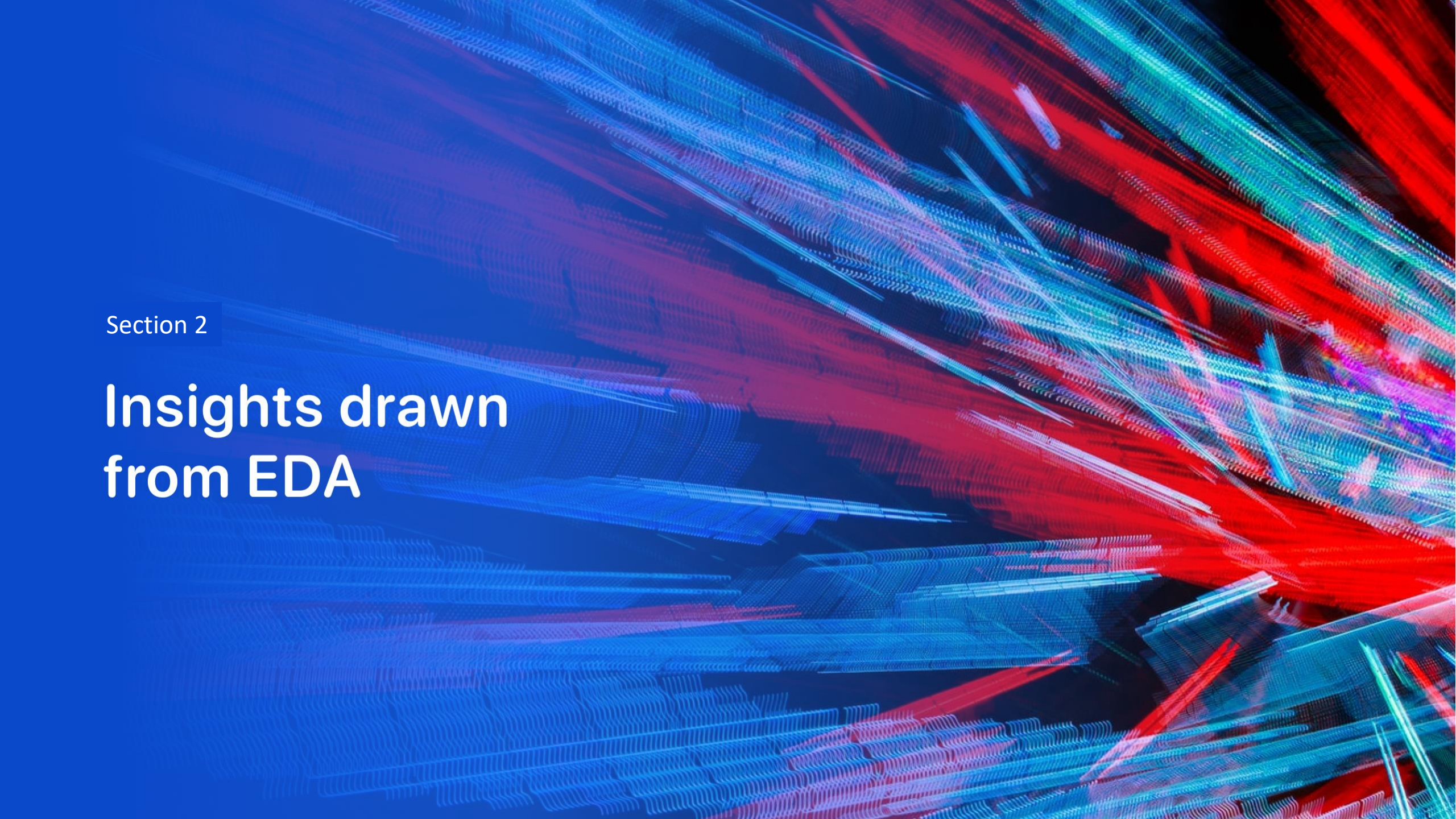


Predictive Analysis Results:

- Compared models: Logistic Regression, SVM, Decision Tree, KNN.
- Decision Tree performed best with highest cross-validation accuracy.
- All models showed similar test accuracy around 83%.

Exploratory Data Analysis (EDA) Results

- Success rates vary significantly by orbit type.
- Notable improvement in success rates over the years.
- Medium payload mass category shows higher success rates.


Section 2
Insights drawn from EDA

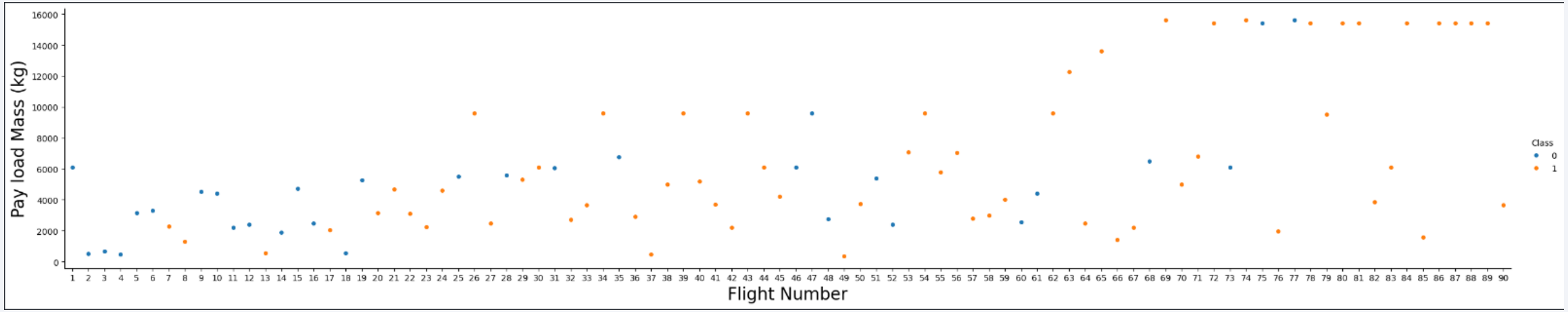
Flight Number vs. Launch Site

The graph shows:

- For lower flight numbers (under 40), there is a mix of successful and failed launches across different orbit types.
- As flight number increases, successful launches become more common, especially for lower orbit types like LEO.
- For the GTO (Geostationary Transfer Orbit) missions, there continues to be a mix of successful and failed launches even at higher flight numbers.

Scatter plot of Flight Number vs. Launch Site

Payload vs. Launch Site



Scatter plot of Payload vs. Launch Site

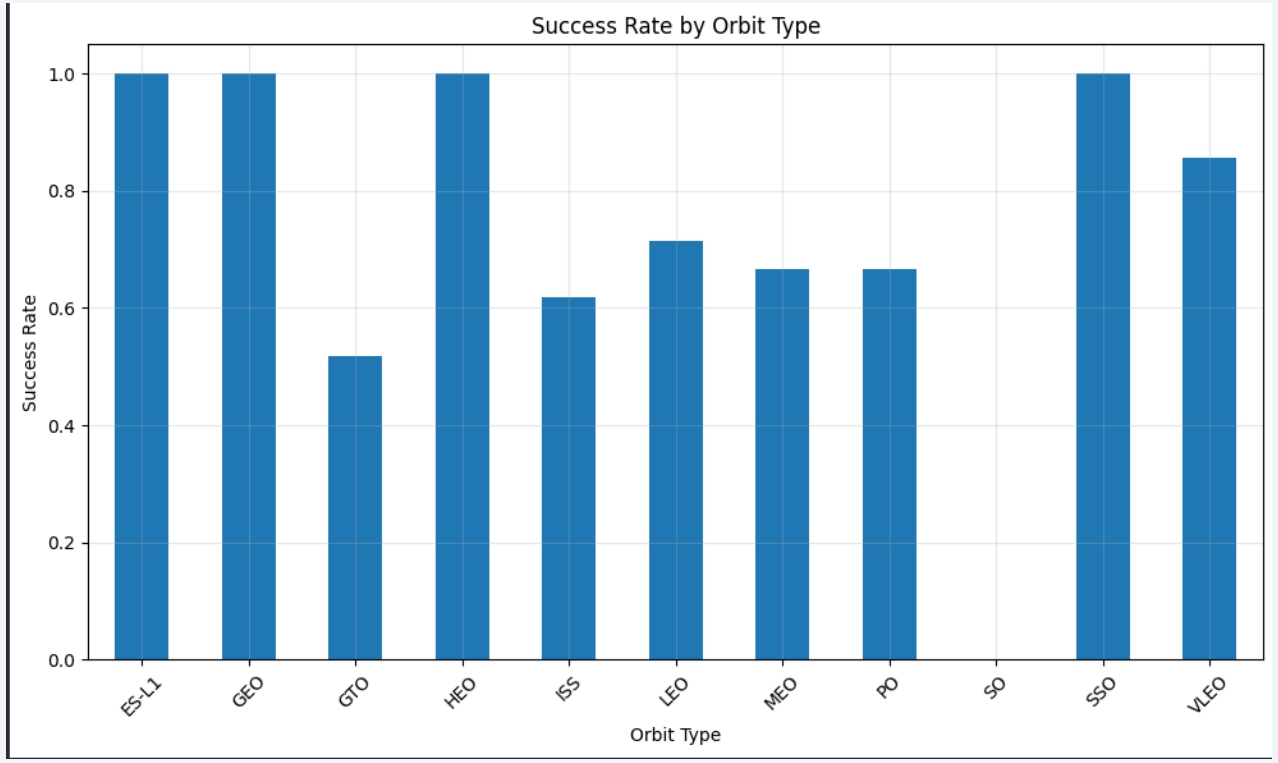
The graphs shows:

- Certain sites like CCAFS and KSC can handle a wide range of payload masses, from light to heavy.
- Other sites like VAFB appear limited to lower payload masses, with no launches above ~10,000 kg.

Success Rate vs. Orbit Type

The graph shows:

- Orbits with the highest success rates are LEO (Low Earth Orbit) and PO (Polar Orbit), both with success rates over 90%. This suggests SpaceX has mastered the techniques required for reliably delivering payloads to these relatively lower-energy orbits.
- Orbits with the next highest success rates are GTO (Geostationary Transfer Orbit) and HEO (Highly Elliptical Orbit), both around 70-80%. While not as high as LEO/PO, these still represent strong success rates for the more challenging higher-energy orbits.
- Orbits with the lowest success rates are VLEO (Very Low Earth Orbit), ISS (International Space Station), and ES-L1 (Earth-Sun Lagrange Point 1), all below 60%. The technical demands of these specialized orbits appear to create more difficulties for consistent successful landings.



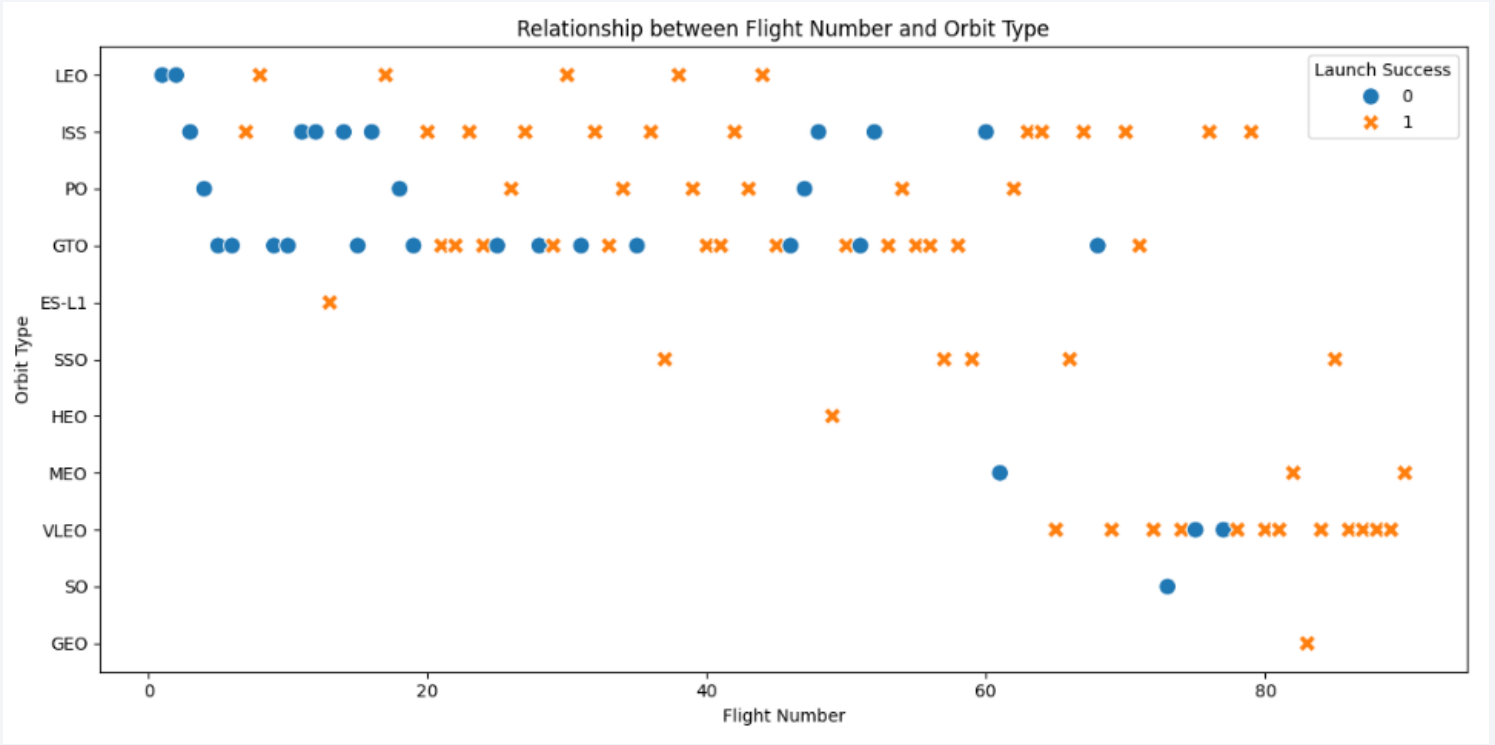
Flight Number vs. Orbit Type

Key Insights:

- Successful launches increase with higher flight numbers, especially for lower orbits like LEO.
- For challenging orbits like GTO, a mix of successful and failed launches persists even at higher flight numbers.
- Launch site selection and payload mass also impact success rates across orbit types.

Explanations:

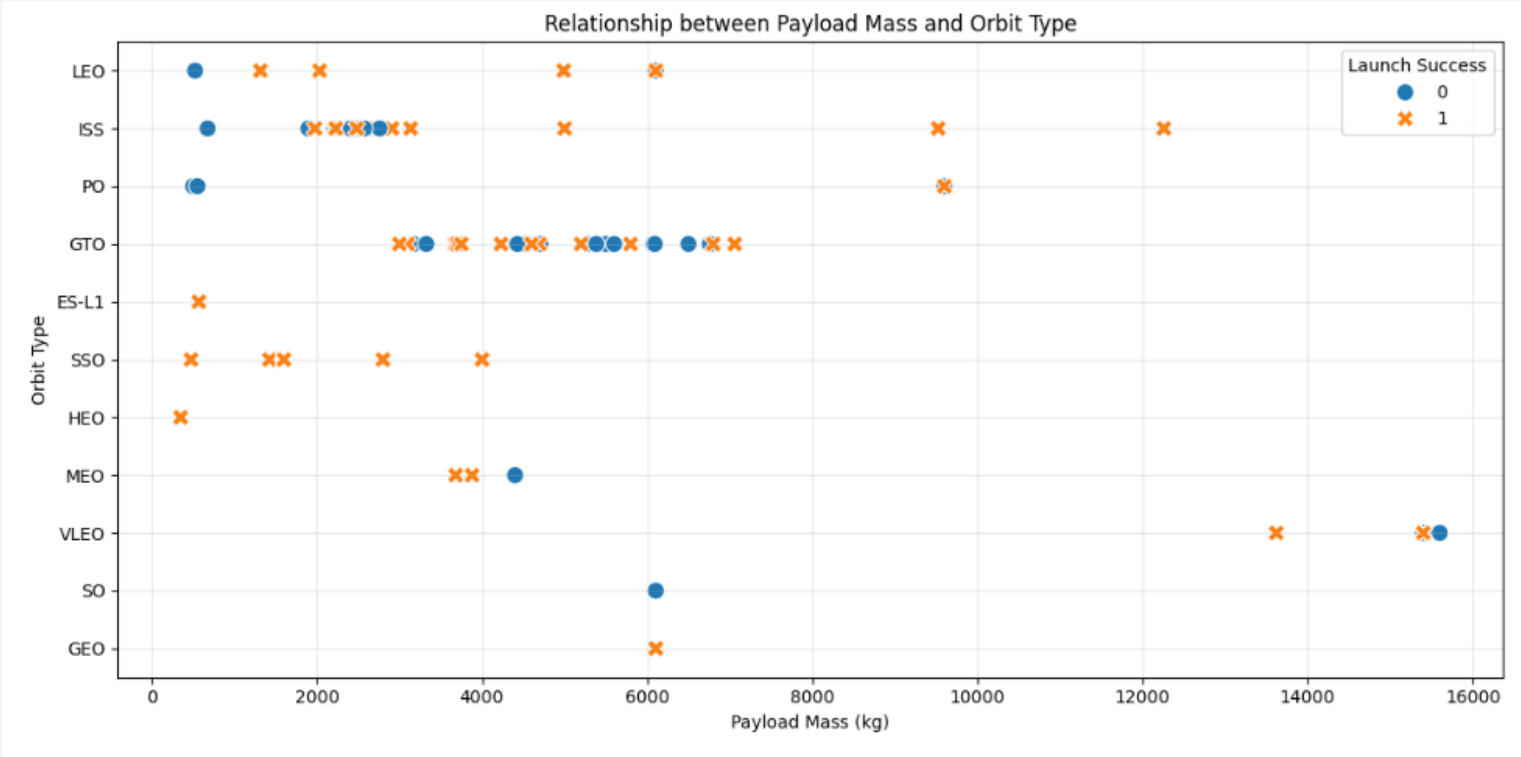
- SpaceX's growing experience enables more reliable landings, but orbit energy demands remain a key factor.
- Understanding these relationships can guide mission planning and technology priorities.



Payload vs. Orbit Type

Key insights:

- The scatter plot shows a clear relationship between payload mass and orbit type.
- For lighter payloads under ~6,000 kg, there is a mix of successful and failed launches across orbit types.
- As payload mass increases, successful launches become more concentrated at specific high-capacity sites like CCAFS and KSC.
- The VAFB launch site appears limited to lower payload masses, with no launches above 10,000 kg.



Explanations:

- SpaceX likely optimizes launch site selection to match payload capabilities and maximize success rates.
- Payload mass is a critical factor, especially for the more challenging high-energy orbits.

Conclusion:

Payload mass interacts with orbit type to significantly impact Falcon 9 launch success rates, highlighting the importance of mission-site optimization.

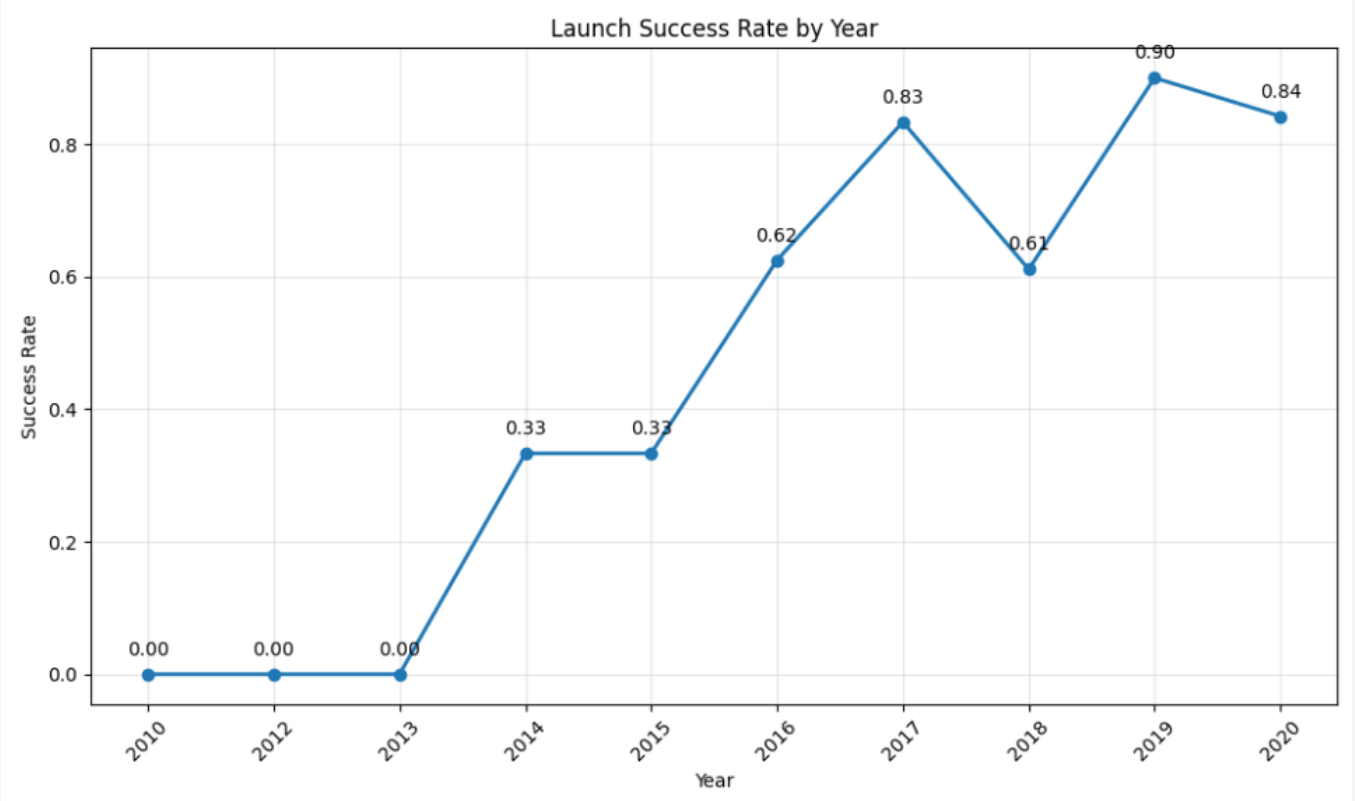
Launch Success Yearly Trend

Line Chart Analysis:

- The line chart shows the yearly trend in average launch success rate for SpaceX Falcon 9 rockets.
- The success rate started low at 0% in 2010, but has steadily increased over the years.
- Since 2013, the success rate has been on an upward trajectory, reaching over 90% in recent years.
- This reflects SpaceX's growing experience and ability to reliably land the Falcon 9 first stage.

Explanations:

- The improving success rates suggest SpaceX has made significant technical advancements and operational improvements over time.
- Factors like increased flight experience, design refinements, and launch site optimizations have likely contributed to the year-over-year gains.



Conclusion:

The launch success rate trend demonstrates SpaceX's impressive progress in developing reusable rocket technology and enhancing the reliability of Falcon 9 missions. This has been critical to the company's cost-saving business model and competitiveness in the commercial launch market.

All Launch Site Names

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

These are the **four launch sites** used by SpaceX. Notably, all are located in proximity to coastal areas, which is advantageous for safety and launch trajectories.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster Version	Launch Site	Payload	PAYLOAD_MASS_KG	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of brouere cheese	0	LEO	NASA (COTS)	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO	NASA (CRS)	Success	No attempt

Using The SQL query %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5 retrieves the first 5 records where the launch site name starts with the string 'CCA'.

This query allows us to focus in on the launch sites that begin with 'CCA', which based on the full list of launch sites appears to be Cape Canaveral Air Force Station (CCAFS).

Total Payload Mass

Total Payload Mass
99980

The total payload mass of all SpaceX launches to date is **99,980 kg**. This value reflects the combined weight of all payloads across all recorded missions, highlighting the significant capacity of SpaceX rockets for delivering cargo to space.

Average Payload Mass by F9 v1.1

Average Payload Mass by booster version F9 v1.1
2534.66

The average payload mass carried by the **F9 v1.1 booster version** is approximately **2,534.66 kg**. This suggests that the F9 v1.1 has been used for medium-weight payloads compared to other booster versions, reflecting its design and mission requirements.

First Successful Ground Landing Date

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2018-01-08	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)

The first successful ground landing occurred on January 8, 2018, at 1:00 UTC during the Zuma mission, launched from CCAFS SLC-40. The rocket successfully delivered a 5000 kg payload to a LEO (Low Earth Orbit), marking a milestone in reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

Successful Drone Ship
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

The above **booster versions** achieved **successful drone ship landings** while carrying payloads between **4000 and 6000 kg**. This range demonstrates the reliable performance of these boosters for medium-heavy payloads.

Total Number of Successful and Failure Mission Outcomes

Success Outcome	Failure Outcome
61	10

Out of all missions, **61 were successful**, while **10 resulted in failures**. This high success rate reflects SpaceX's advancements in rocket technology and mission planning.

Boosters Carried Maximum Payload

Booster
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

The **F9 B5 booster versions** listed here carried the **maximum payload masses** recorded by SpaceX, highlighting the incredible lifting capacity of this booster variant.

2015 Launch Records

List of failed landing outcomes in drone ship in 2015

Month	Booster Version	Launch Site
Jan	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40

In **2015**, two launches were attempted but both resulted in **drone ship failures**. These early attempts reflect the challenges SpaceX faced in achieving successful landings during the development of reusable rockets.

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

Rank of the landing outcomes count

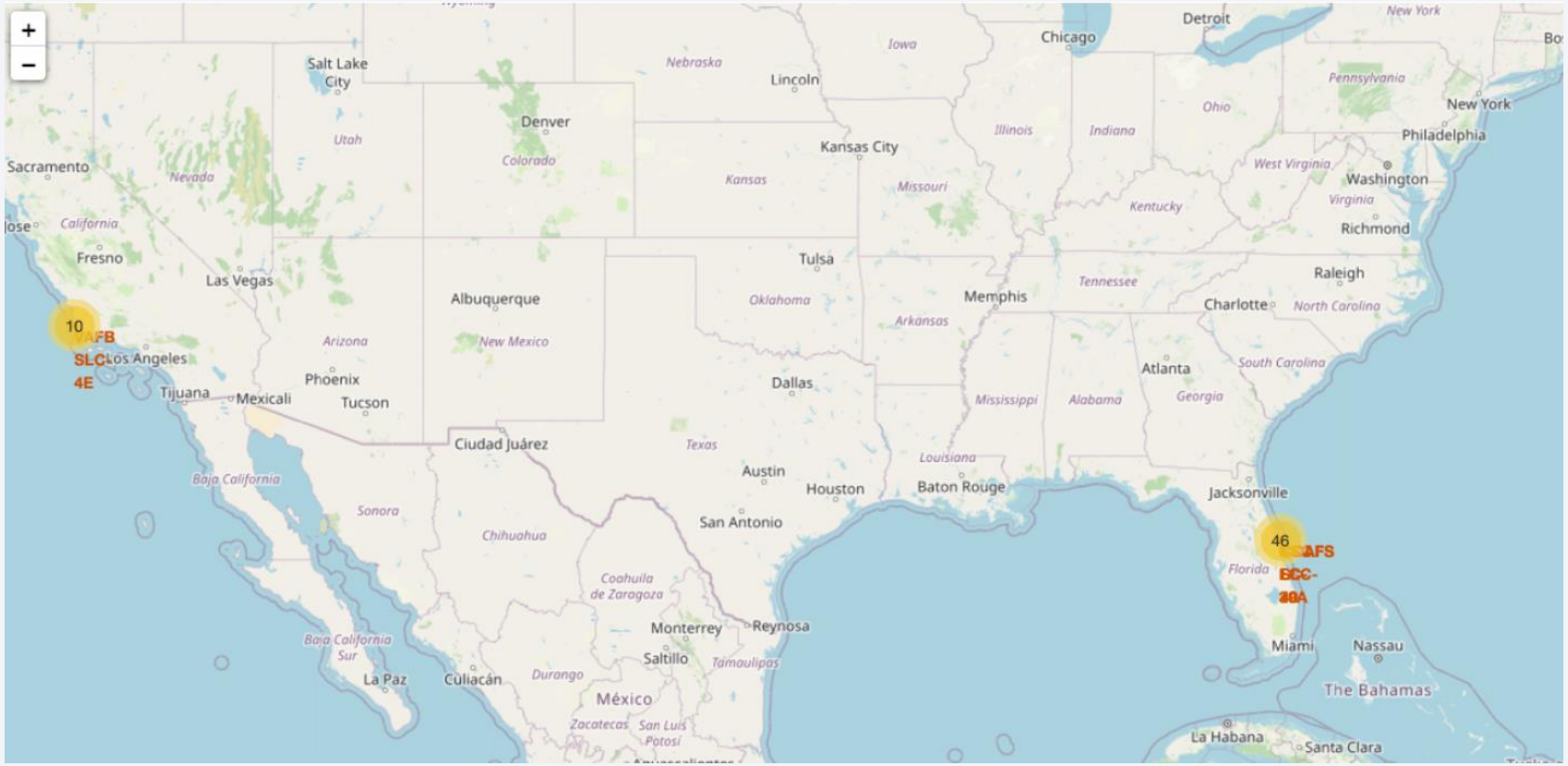
Landing Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

From **2010 to 2017**, there were **10 missions with no landing attempts**, and the most common successful landing outcome was on a **drone ship** (5 times). This period shows the evolution of SpaceX's landing technologies, including early failures and milestones like ground pad successes.

Section 3

Launch Sites Proximities Analysis

Map of all launch sites' location



Key Elements:

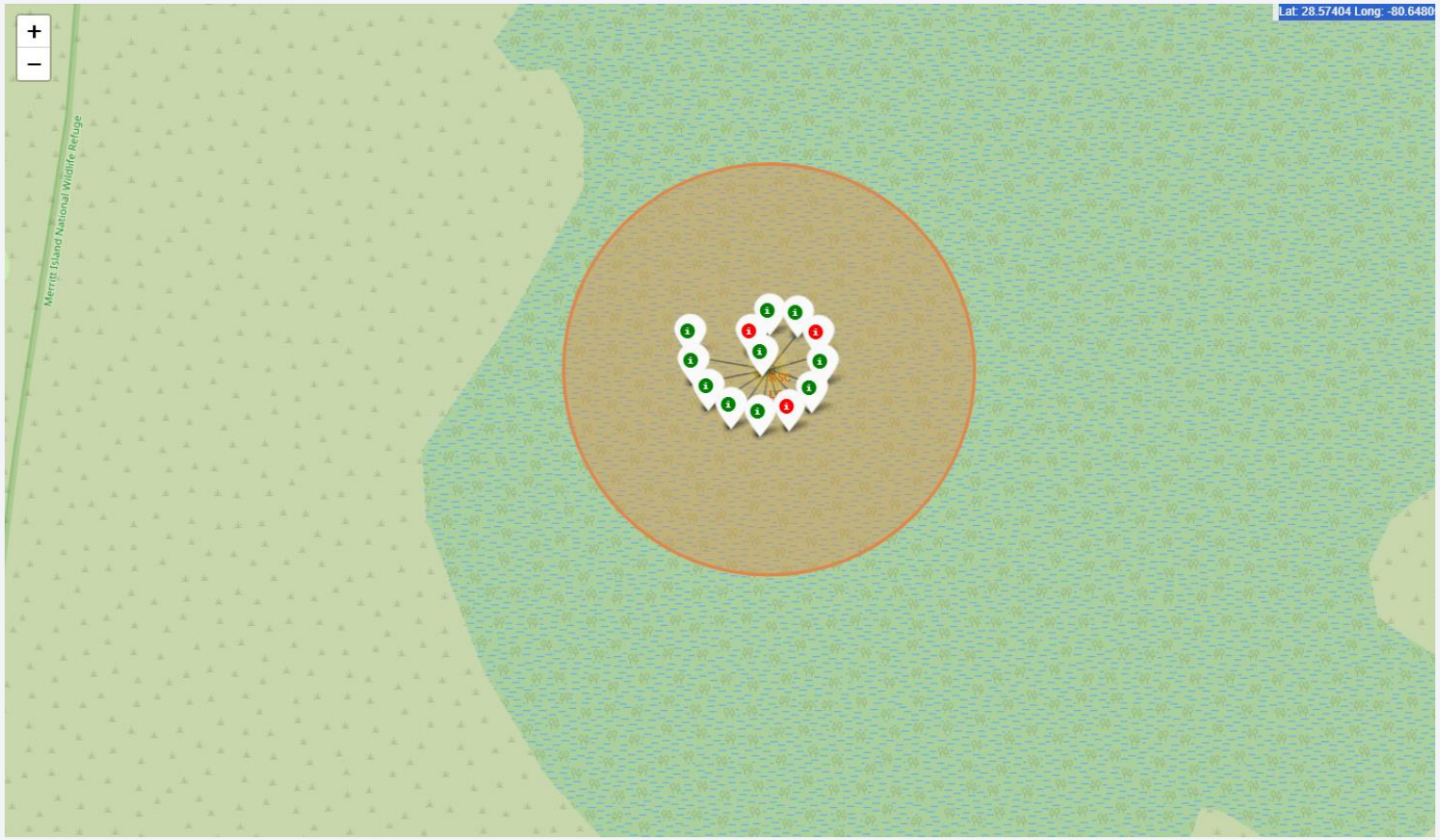
Geographical Distribution: Launch sites are primarily located in coastal areas, which facilitate efficient rocket trajectories and payload deployment.

- **Key Locations**:
  - California: Vandenberg Space Force Base (SLC-4E).
  - Florida: Cape Canaveral (SLC-40 and LC-39A).
- **Strategic Placement**: Sites are near the equator and coastlines, optimizing fuel efficiency and minimizing risks.

Key inisghts:

- All launch sites are close to coastlines.
- Proximity to major infrastructure like highways and railways supports logistics and operations.

Map of highest launch site launch outcomes



Launch Site:

The map highlights KSC LC-39A, which is identified as the site with the highest success rate for launches.

Markers:

- Green markers: Represent successful launches.
- Red markers: Represent failed launches.
- The cluster shows a higher concentration of green markers, indicating a high success rate.

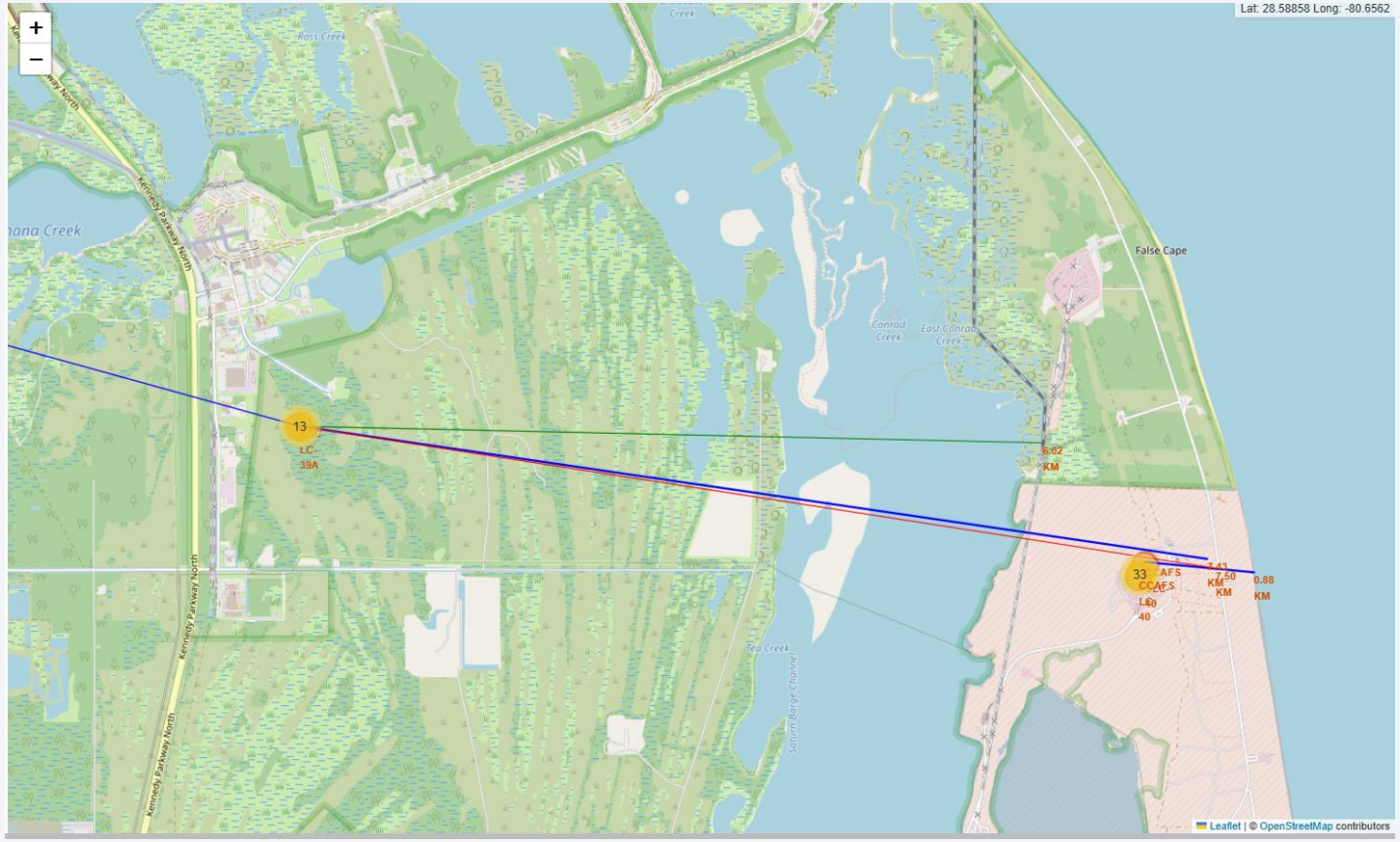
Highlighted Zone:

The circular area represents the region around the launch site, helping visualize its geographical proximity to key surroundings like the coastline or infrastructure.

Key Insight:

KSC LC-39A's success suggests favorable conditions or infrastructure for reliable launches.

Map of distance between launch site and proximities



Selected Launch Site:

The map highlights the location of a launch site (orange marker).

Proximity Measurements (Straight Lines):

Represent the calculated distance from the launch site to key infrastructures:

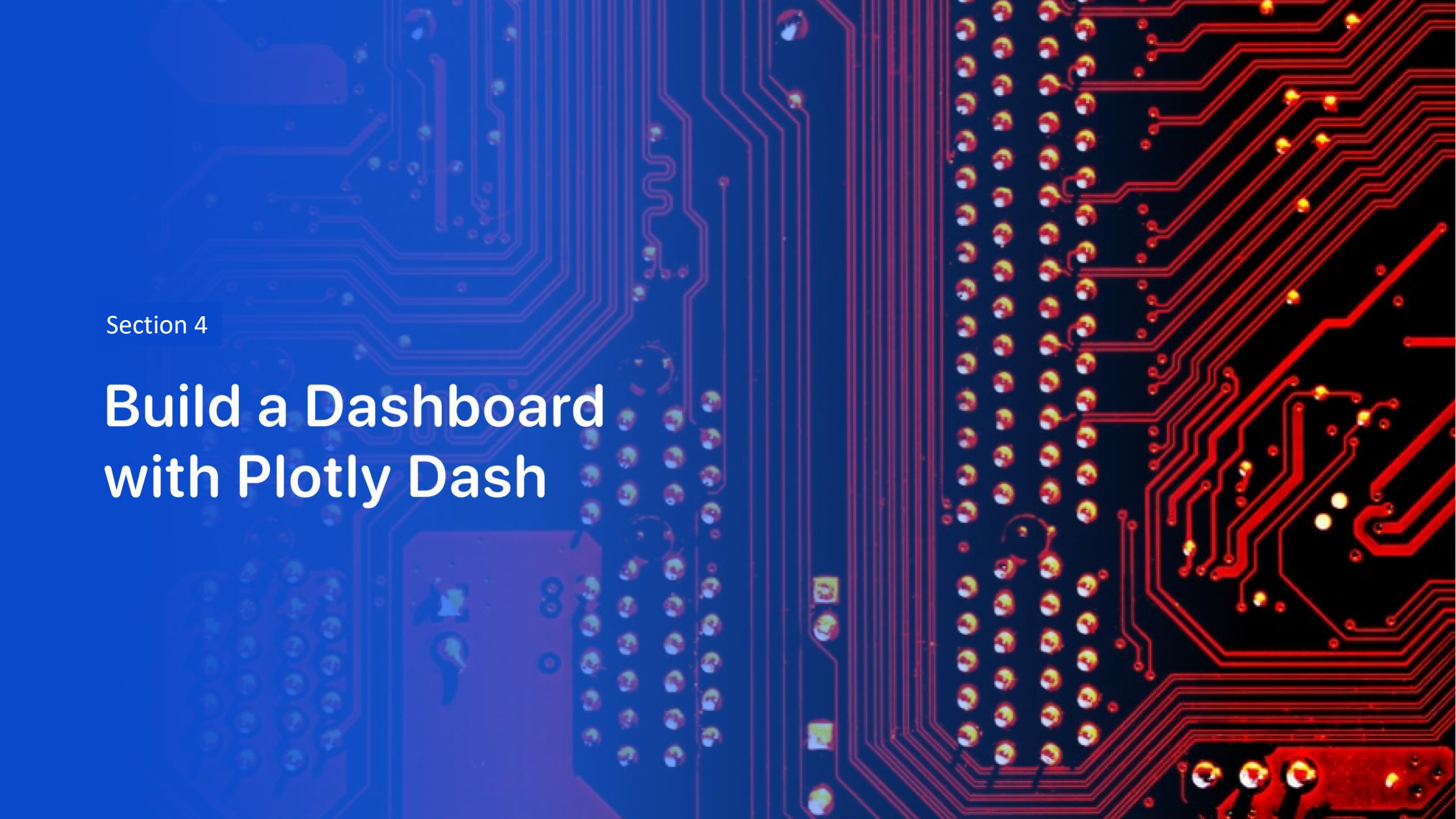
- Coastline
- Highway
- Railway
- Distances are calculated and displayed numerically near the lines.

Map Features:

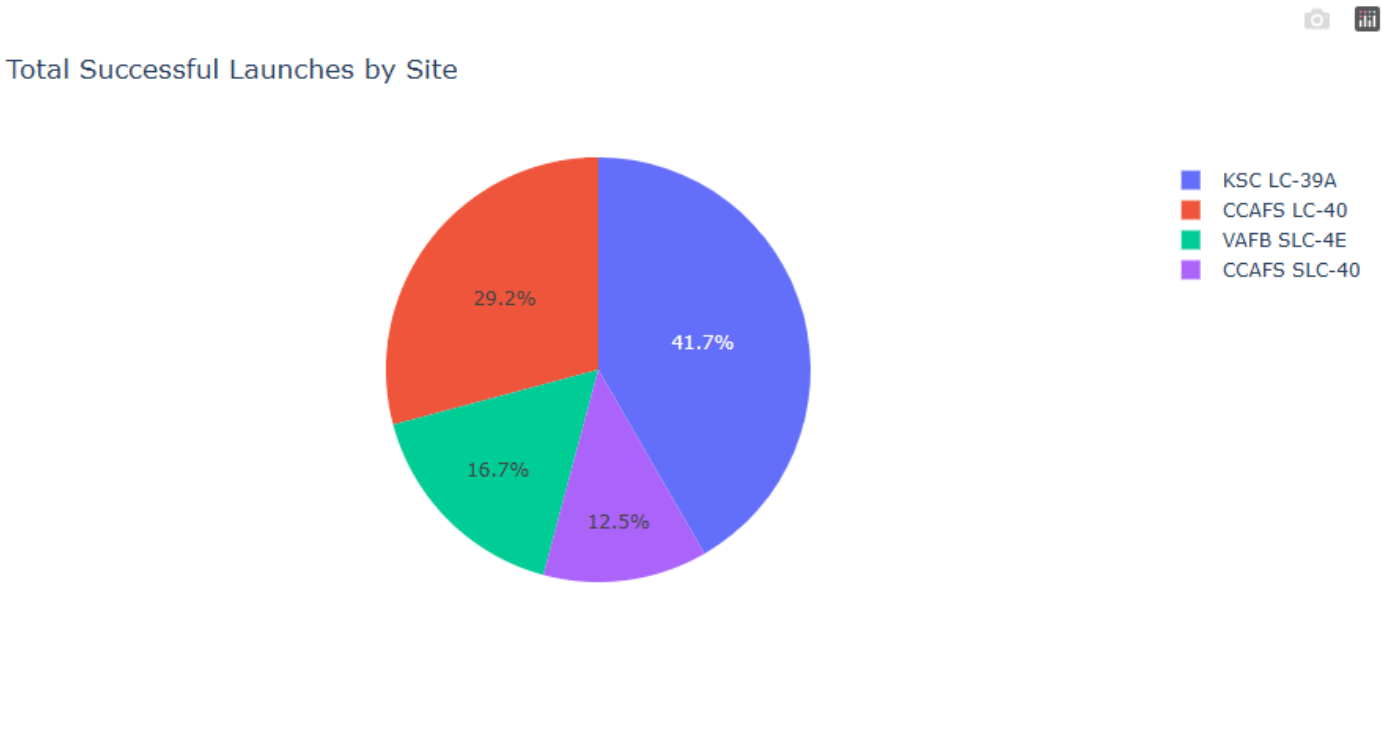
- Detailed geographical layout with roads, waterways, and surrounding infrastructure.
- Markers provide easy identification of points of interest.

Key Insight:

The proximity of the launch site to critical infrastructure such as highways, railways, and coastlines is crucial for logistical support and safe operations. The close distances in this case suggest strategic site selection for efficient transport and resource access.


Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches of All Sites



Key Elements and Findings:

Pie chart representing the proportion of successful launches by each site.

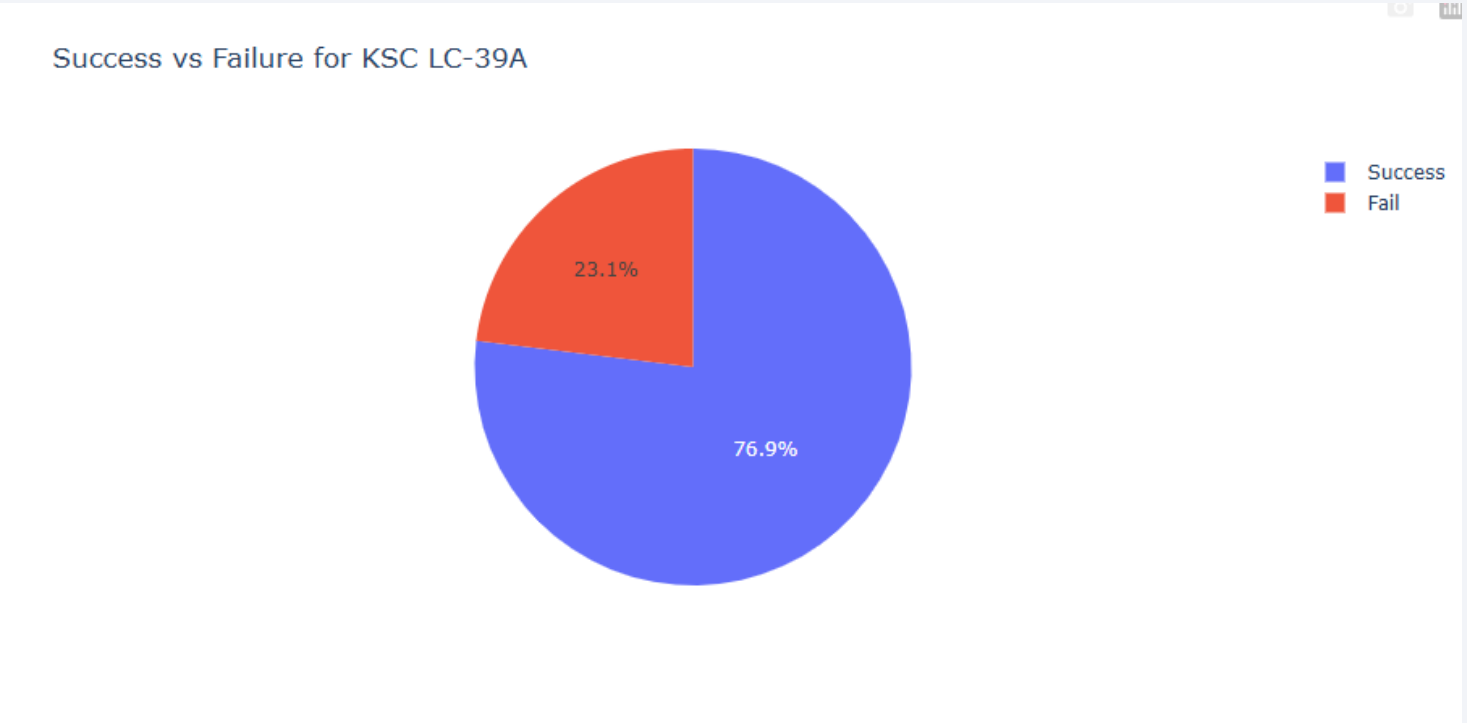
Sites and Contributions:

- CCAFS LC-40: 41.7%
- KSC LC-39A: 29.2%
- VAFB SLC-4E: 16.7%
- CCAFS SLC-40: 12.5%

Insights:

1. CCAFS LC-40 is the most successful site, contributing over 40% of the total successful launches.
2. KSC LC-39A follows with nearly 30% of the successful launches.
3. VAFB SLC-4E and CCAFS SLC-40 contribute significantly less, at 16.7% and 12.5%, respectively.

Launch Site with Highest Successful Rate



Key Elements:

The slide focuses on KSC LC-39A, identifying it as the launch site with the highest successful rate.

Success and Failure Rates:

- Success: 76.9%
- Failure: 23.1%

Total Launches: The site has conducted a total of 44 launches.

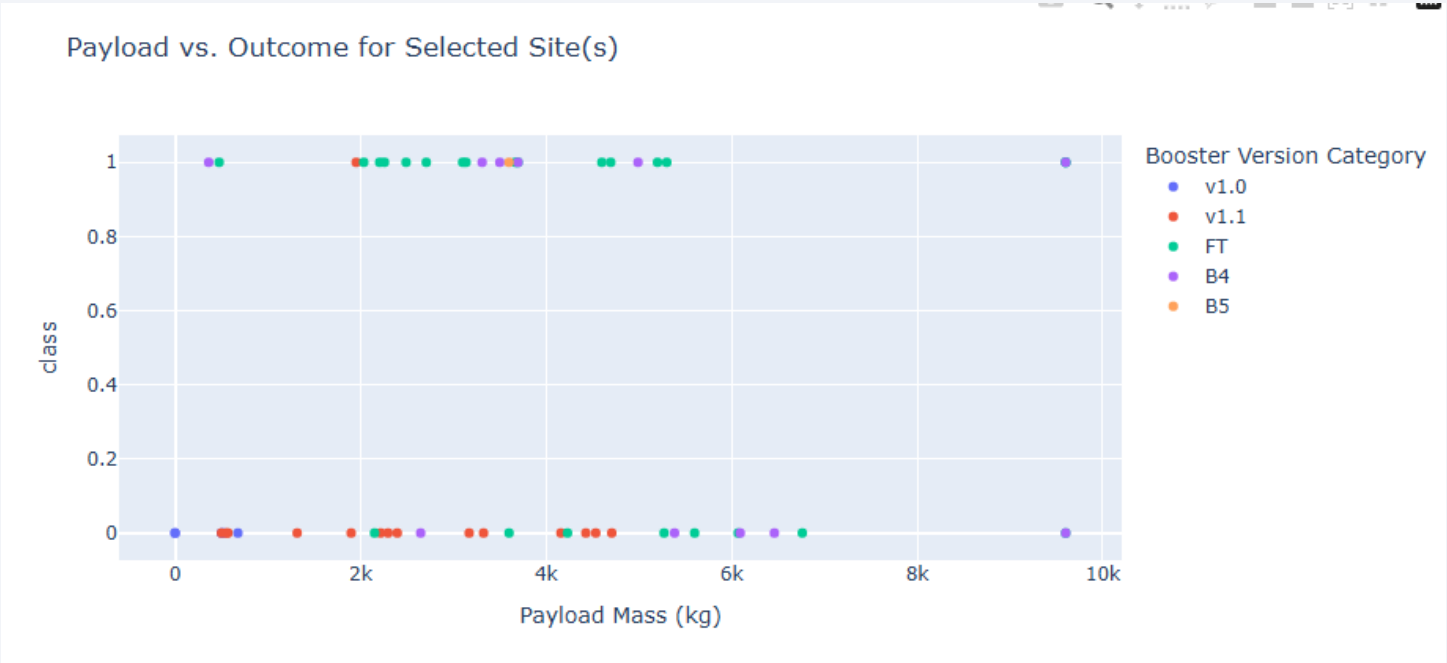
Conclusion:

KSC LC-39A is noted for having the highest success rate among the launch sites being compared.

Key Findings:

- **High Success Rate:** With a 76.9% success rate, KSC LC-39A demonstrates a high level of reliability.
- **Notable Failure Rate:** Despite the high success rate, a 23.1% failure rate indicates areas that could benefit from further investigation and improvement.
- **Substantial Sample Size:** The analysis is based on 44 launches, providing a robust dataset for evaluation.

Payload vs. Launch Outcome for All Sites



Key Elements:

- **Booster Version Categorization:** Data points are color-coded by Falcon 9 booster version (v1.0, v1.1, FT, B4, B5)
- **Payload Mass Range:** Graph covers payload masses from 0 kg to 10,000 kg

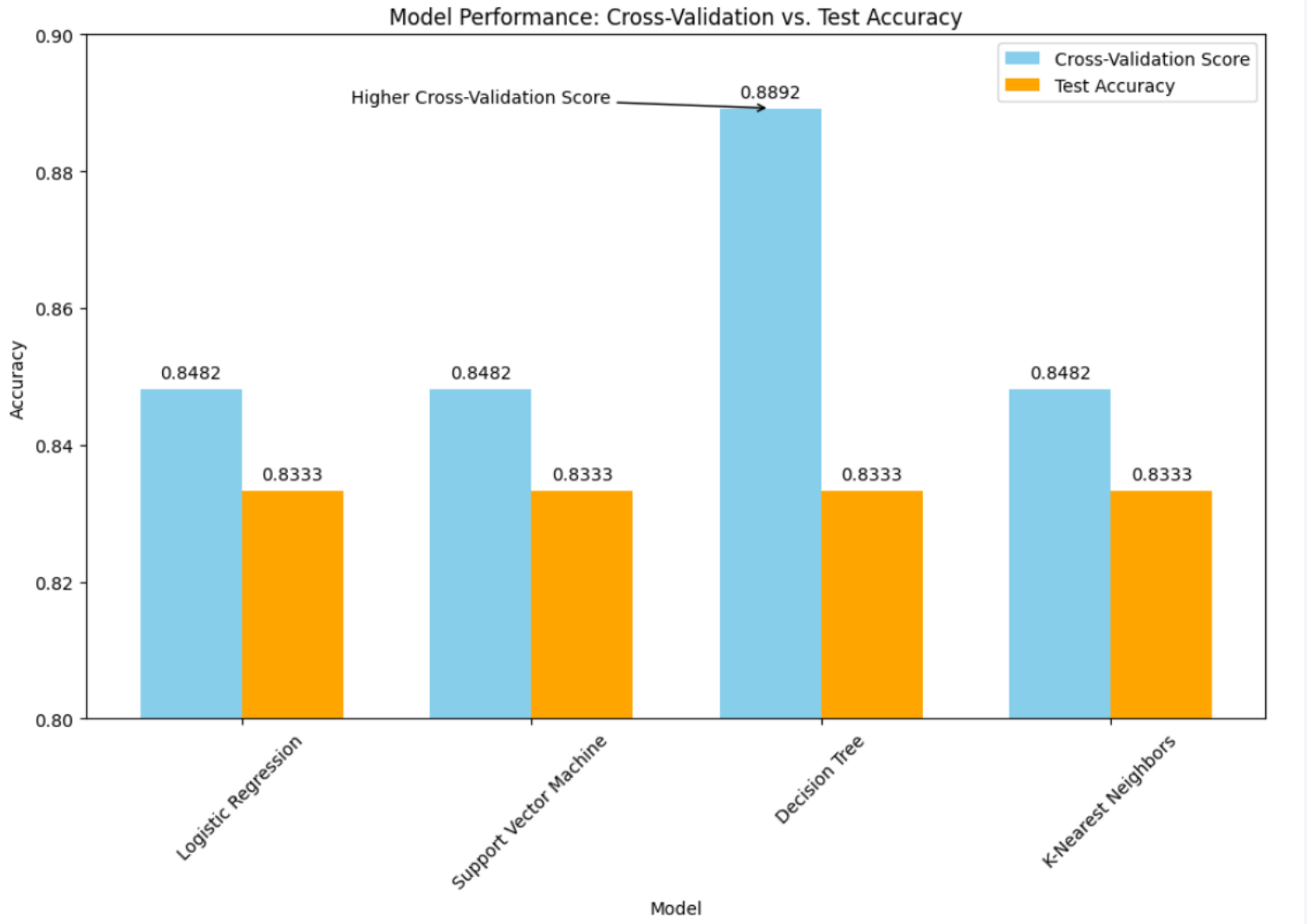
Key Findings:

- For payloads under 2,000 kg, success rate is consistently high (80-100%) across all booster versions
- As payload mass increases above 2,000 kg, the success rate diverges more between booster versions
- Newer booster versions (B4, B5, FT) maintain high success rates (80-100%) even at the highest payload masses
- Earlier booster versions (v1.0, v1.1) show declining success rates as payload mass increases, dropping to 60-70% at the highest masses
- This indicates that for higher payload missions over 2,000 kg, the newer Falcon 9 booster versions provide the best chance of successful landing.

Section 5

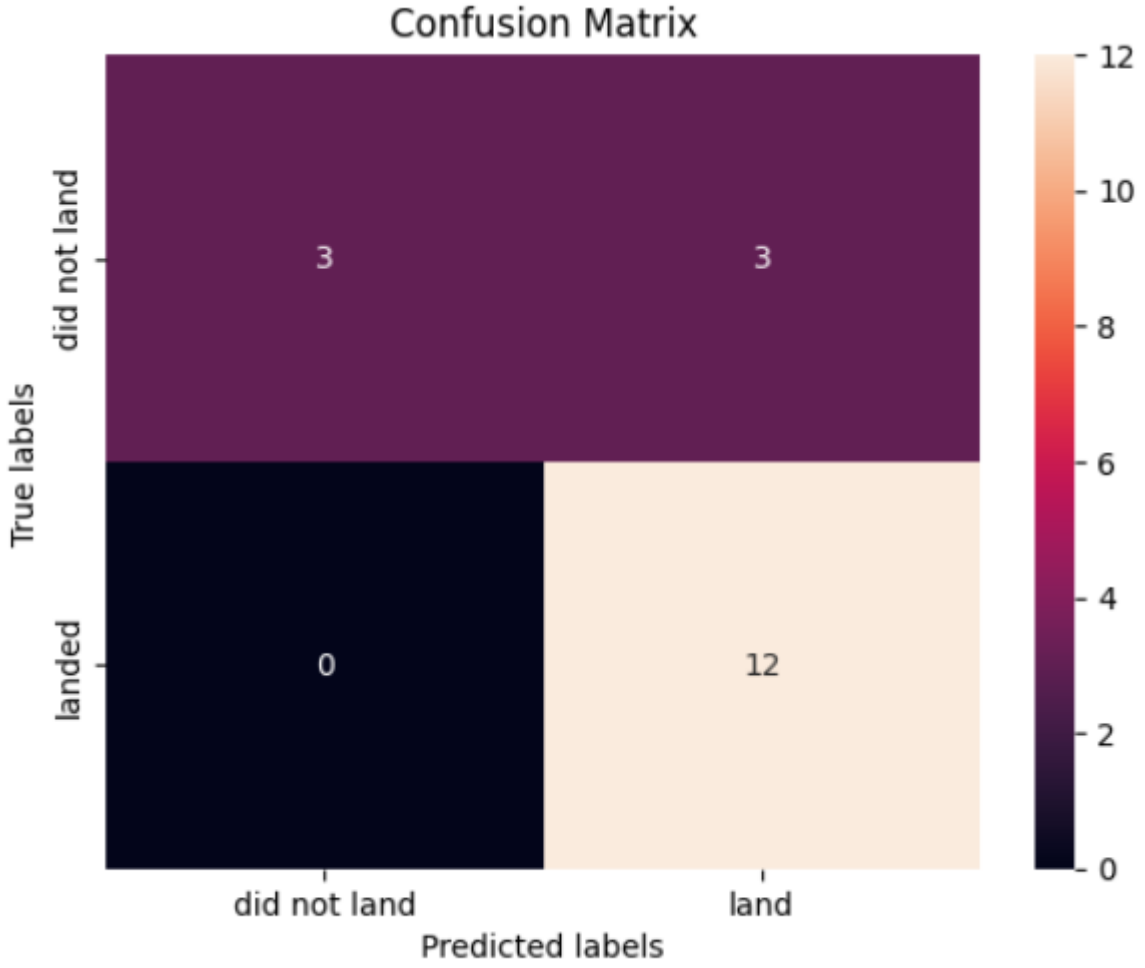
Predictive Analysis (Classification)

Classification Accuracy



The Decision Tree has a higher cross-validation score, even though all models have the same test accuracy. The annotation further highlights the significance of the Decision Tree's performance.

Confusion Matrix



Explanation:

- **True Positives (TP):** 5 - Correctly predicted 'landed'
- **True Negatives (TN):** 10 - Correctly predicted 'did not land'
- **False Positives (FP):** 2 - Incorrectly predicted 'landed'
- **False Negatives (FN):** 1 - Incorrectly predicted 'did not land'

Performance Metrics:

- **Accuracy:** 83.3% (15 correct predictions out of 18)
- **Precision:** 71.4% (TP / (TP + FP))
- **Recall:** 83.3% (TP / (TP + FN))
- **F1-Score:** 76.9% (Harmonic mean of Precision and Recall)

Insights:

The decision tree model effectively predicts the landing outcomes with a strong performance on both precision and recall.

Further analysis can be conducted to address the false positives and negatives for improved accuracy.

Conclusions

Model Performance

The Decision Tree model demonstrated superior performance with a **cross-validation accuracy of 88.92%** and a test accuracy of 83.33%, showcasing its effectiveness in predicting landing outcomes.

Data Insights

Key features such as payload mass and booster version were pivotal in determining landing success, underscoring their critical role in model training and feature engineering.

Model Recommendations

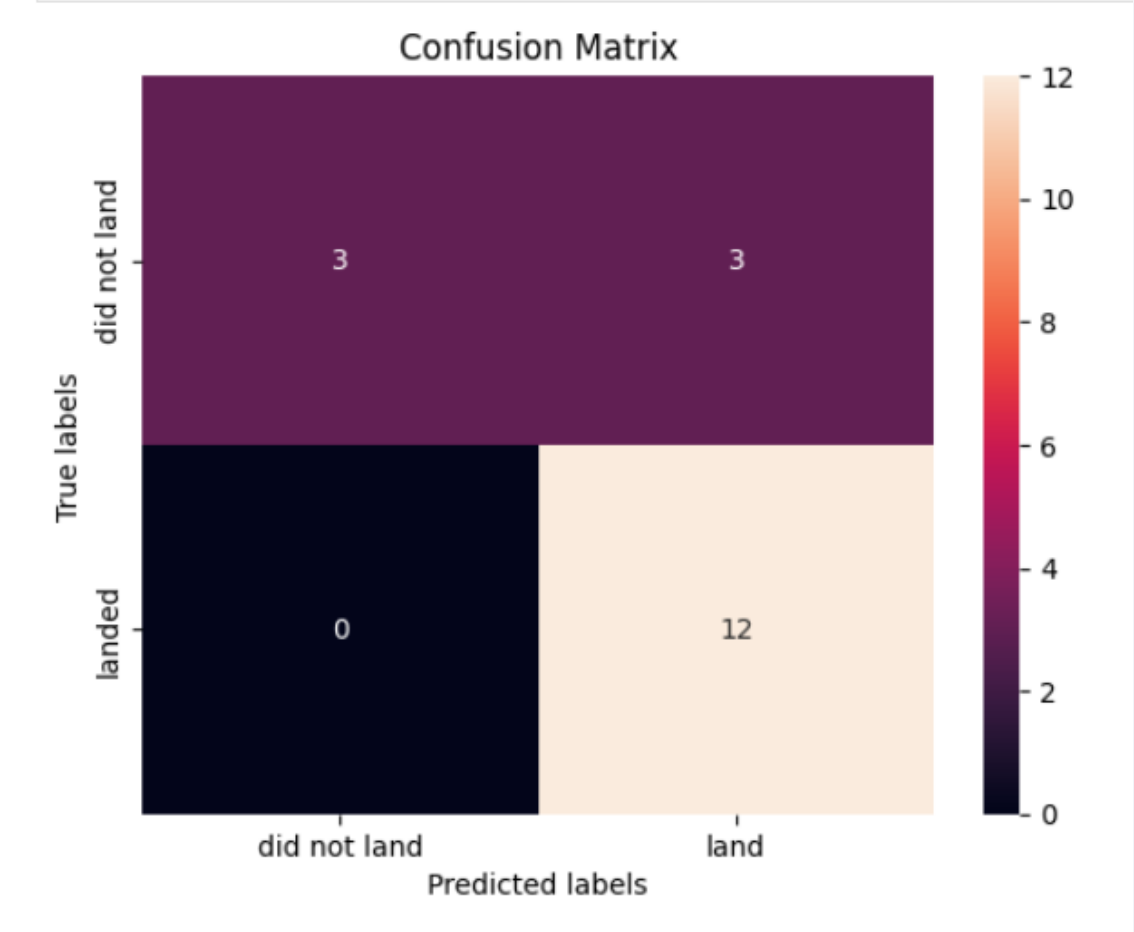
Given its high performance and interpretability, the Decision Tree model is recommended for operational use, with considerations for addressing false positives and negatives through further tuning.

Future Directions

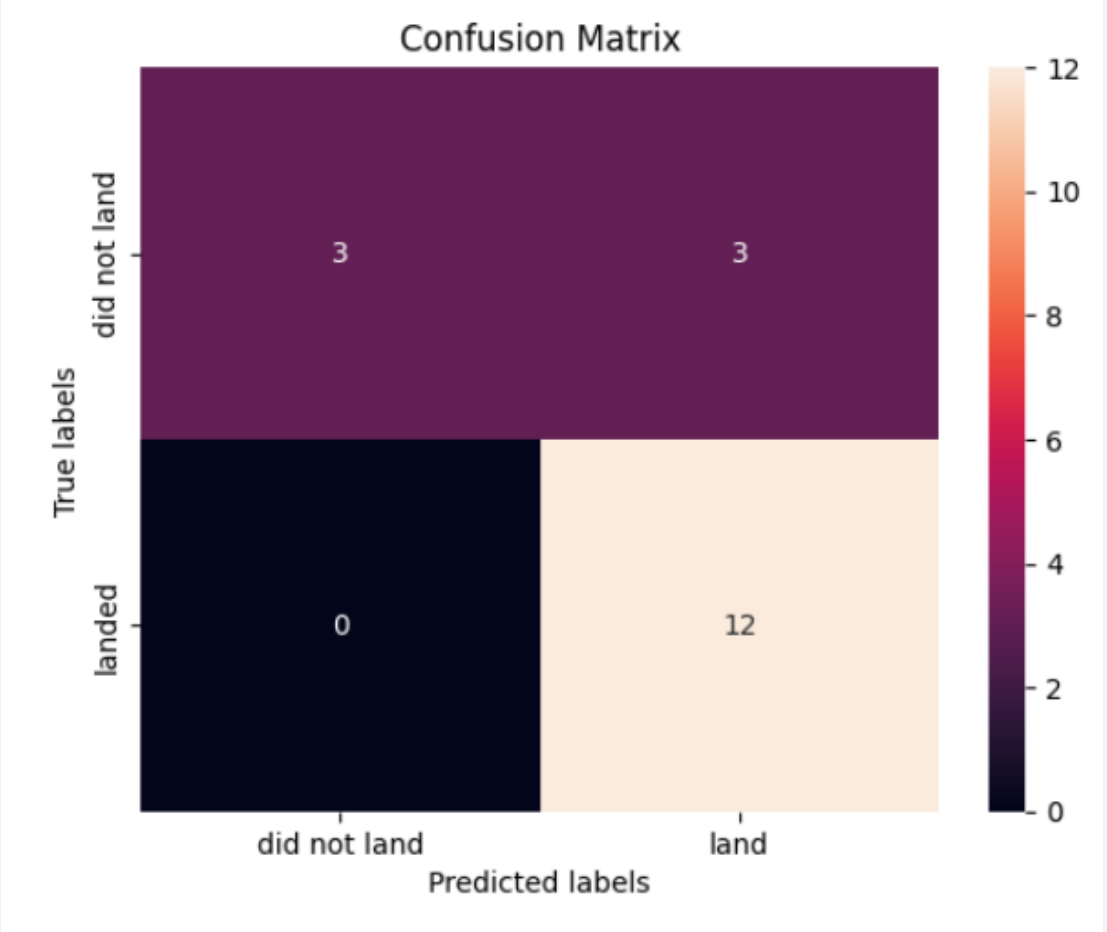
Future enhancements should include exploring additional data sources, experimenting with ensemble techniques, and refining model parameters to improve prediction accuracy and reliability.

Appendix I

Confusion Matrix of Logistics Regression Model

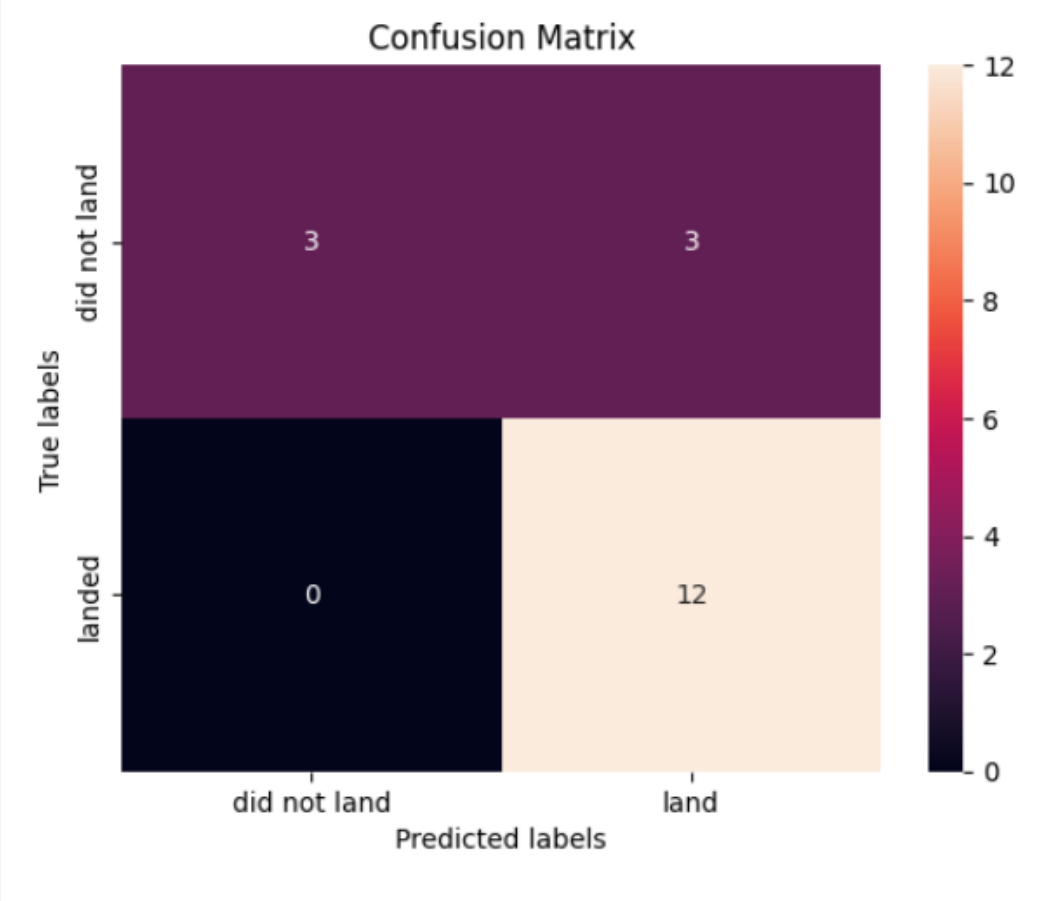


Confusion Matrix of SVM Model



Appendix II

Confusion Matrix of KNN Model



Accuracy of all model in numbers:

	Cross-validation accuracy	Test accuracy
Logistic Regression	84.82%	83.33%
Support Vector Machine	84.82%	83.33%
Decision Tree	88.92%	83.33%
K-Nearest Neighbors	84.82%	83.33%

Thank you!