IBM APPLIED DATA SCIENCE CAPSTONE

OUTLINE

1. EXECUTIVE SUMMARY

2. INTRODUCTION

3. METHODOLOGY
   - DATA COLLECTION AND DATA WRANGLING
   - EDA AND INTERACTIVE VISUAL ANALYTICS
   - PREDICTIVE ANALYSIS

4. RESULTS
   - DATA COLLECTION AND DATA WRANGLING
   - EDA AND INTERACTIVE VISUAL ANALYTICS
   - PREDICTIVE ANALYSIS

5. DISCUSSION

6. CONCLUSION

EXECUTIVE SUMMARY

This capstone project, part of the **IBM Applied Data Science Certificate**, focused on predicting the landing success of the **SpaceX Falcon 9 first stage** using data-driven techniques and machine learning. The primary goal was to develop predictive models that can assist in optimizing launch strategies, reducing costs, and advancing reusable rocket technology.

PROJECT WORKFLOW OVERVIEW

The project followed a structured approach, including:

- **Data Acquisition**: Launch data was collected from public APIs and enhanced through web scraping from Wikipedia.
- **Data Cleaning & Preparation**: The dataset was cleaned, transformed, and stored in a Db2 database for efficient querying and analysis.
- **Exploratory Data Analysis (EDA)**: Key patterns were examined to understand relationships between features such as launch site, payload mass, orbit type, and mission outcomes.
- .

•**Feature Engineering**: New variables were created and data was standardized to prepare for modeling.

•**Interactive Visualizations**: Tools like **Folium** and **Plotly Dash** were used to create dynamic dashboards and maps, offering rich insights into geographical and operational trends

•**Model Development**: Classification algorithms including **Support Vector Machines**, **Decision Trees**, and **K-Nearest Neighbors** were trained and tested using historical data.

KEY FINDINGS

•The **CCAFS LC-40** launch site recorded the highest success rate, responsible for nearly **43.7%** of all successful missions.

•The **FT booster version** showed consistent success across various payload weights, indicating superior reliability.

•There was **no strong correlation** between increased payload mass and mission failure, suggesting that factors like **booster type** and **launch site** are more impactful.

•**Interactive visualizations** significantly enhanced data interpretation, helping stakeholders identify patterns with greater clarity.

MODEL PERFORMANCE

Among the machine learning models evaluated:

- The **Decision Tree classifier** outperformed others, achieving the highest accuracy (~94.44%), making it the most effective algorithm for predicting mission outcomes.
- Both **SVM** and **K-Nearest Neighbors** models also showed reasonable accuracy (~83.33%), validating the presence of predictive patterns in the data.

This project demonstrates how machine learning and visual analytics can be leveraged to uncover critical insights in aerospace operations. By accurately predicting the success of Falcon 9 landings, the findings offer tangible value for future mission planning and cost optimization. The robust analytical framework developed here can support strategic decisions and contribute meaningfully to the future of **reusable rocket technologies**.

INTRODUCTION

The goal of this capstone project is to predict whether the **Falcon 9 first stage** will successfully land after launch. This prediction is significant due to the economic implications tied to **SpaceX's reusability model**. While a typical Falcon 9 launch is advertised at **$62 million**, other providers can charge upwards of **$165 million**.

A large portion of SpaceX's cost efficiency stems from its ability to reuse the rocket's first stage—making accurate predictions of landing success critical to assessing the true cost of a launch.

From a competitive standpoint, such predictions can also empower other companies to make informed bids against SpaceX by estimating the feasibility and cost-efficiency of their own launch strategies. Although not all unsuccessful landings are failures—many are **intentionally planned ocean landings**—understanding the conditions under which the rocket lands successfully is key.

The central question we aim to answer is:

Given a set of features about a Falcon 9 launch—such as payload mass, orbit type, and launch site—can we accurately predict whether the rocket's first stage will land successfully?

To solve this, we leverage data science and machine learning techniques, working with historical launch data to uncover patterns and build predictive models that support both strategic planning and cost estimation in the aerospace sector.

METHODOLOGY

DATA COLLECTION AND DATA WRANGLING METHODOLOGY

This project relies on two key sources of information: the **SpaceX public API** and **web scraping from Wikipedia**. These were used to compile a comprehensive dataset specifically focused on **Falcon 9 launches**, which serves as the foundation for the machine learning models developed later.

1. SpaceX API

- The primary data source was the SpaceX API at: https://api.spacexdata.com/v4/rockets/
- This API offers a wealth of technical specifications and historical data related to various rockets.
- After filtering exclusively for **Falcon 9** data, the dataset contained **90 records** and **17 relevant features**, including height, mass, engine count, and more.
- **Missing Data Handling**: Numerical features with missing values were cleaned using **mean imputation**, where missing values are replaced with the average of their respective columns.

2. Web Scraping from Wikipedia

•Additional launch data was scraped from:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

•The scraped dataset provided **121 instances** and **11 key features** such as launch site, payload mass, mission outcome, and orbit type, all specific to **Falcon 9** missions.

3. Data Wrangling

After collecting data from both sources, several steps were taken to clean, standardize, and reshape the data for analysis:

•**Column Alignment**: Datasets from the API and Wikipedia were aligned by matching column names and ensuring consistent formatting for shared features such as orbit names, launch dates, and payload descriptions.

•**Missing Value Treatment**: Beyond mean imputation for numerical data, rows with irreparable missing values in critical fields (e.g., launch outcome) were removed to preserve data integrity.

- **Categorical Data Transformation**: Features like launch site, booster version, and orbit were **converted into numerical format using one-hot encoding** to make them compatible with machine learning algorithms.

- **Feature Engineering**: A new target column named 'Class' was added to indicate whether the rocket **successfully landed (1)** or **failed (0)**. Additional features were also derived to enhance predictive power, such as grouping orbit types or simplifying payload weight bands.

- **Data Consolidation**: The final dataset, post-wrangling, consisted of **90 rows and 83 features**, including both original and engineered variables.

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLGY

Pandas & NumPy

•Derived key insights:

- Launch counts per site
- Orbit distribution
- Mission outcomes (success/failure)

SQL

•Queried structured data for:

- Unique launch site names
- Total payload mass for NASA (CRS) missions
- Avg. payload mass for F9 v1.1 boosters

📊 Matplotlib & Seaborn

•Visualized data with:

- Scatter plots, bar charts, line graphs

•Explored:

- Flight number vs. launch site
- Payload mass vs. launch site
- Success rate by orbit type

🌎 Folium

•Interactive map visualizations:

- Launch site markers
- Success & failure outcomes per site
- Distance from launch sites to nearby cities, railways, highways

📈 Dash

•Built interactive dashboard:

- Dropdown & range slider for user inputs
- Pie chart: Total successful launches per site
- Scatter plot: Payload mass vs. mission outcome

PREDICTIVE ANALYSIS METHODOLGY

🧰 **Tools Used**

•**Scikit-learn** library for building and evaluating machine learning models

⚙️ **Process Overview**

1.Data Preprocessing

1. Standardized feature data
2. Split dataset into training & testing sets

2.Model Development

1. Built classification models:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Decision Tree
    4. K-Nearest Neighbors (KNN)
2. Trained models on historical data

3. Model Tuning & Evaluation

- Used **GridSearchCV** for hyperparameter optimization
- Evaluated performance using:

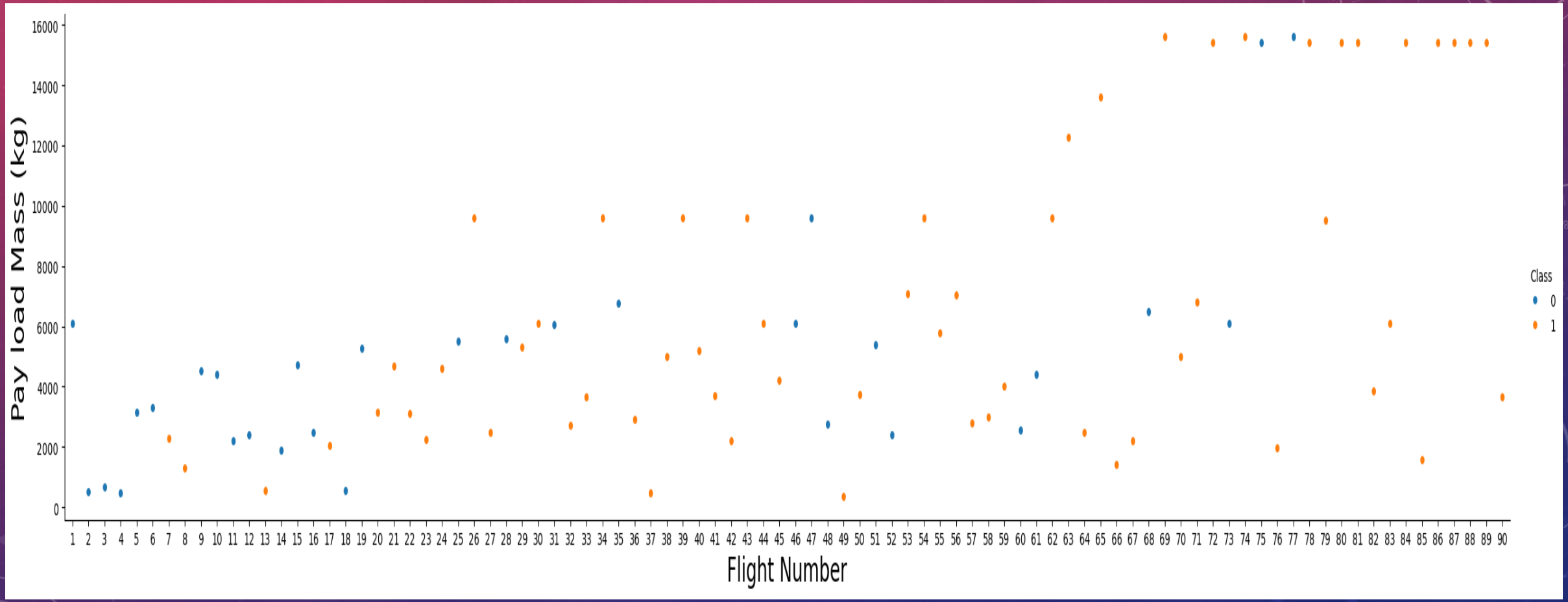
  - **Accuracy Score**: Measures how often the model correctly predicts outcomes
  - **Confusion Matrix**: Displays counts of:
  - **True Positives (TP)**: Correct positive predictions
  - **True Negatives (TN)**: Correct negative predictions
  - **False Positives (FP)**: Incorrectly predicted positive outcomes
  - **False Negatives (FN)**: Incorrectly predicted negative outcomes

These models help us uncover patterns in launch data and offer a reliable way to predict the success of Falcon 9 first-stage landings, aiding future mission planning and cost estimation

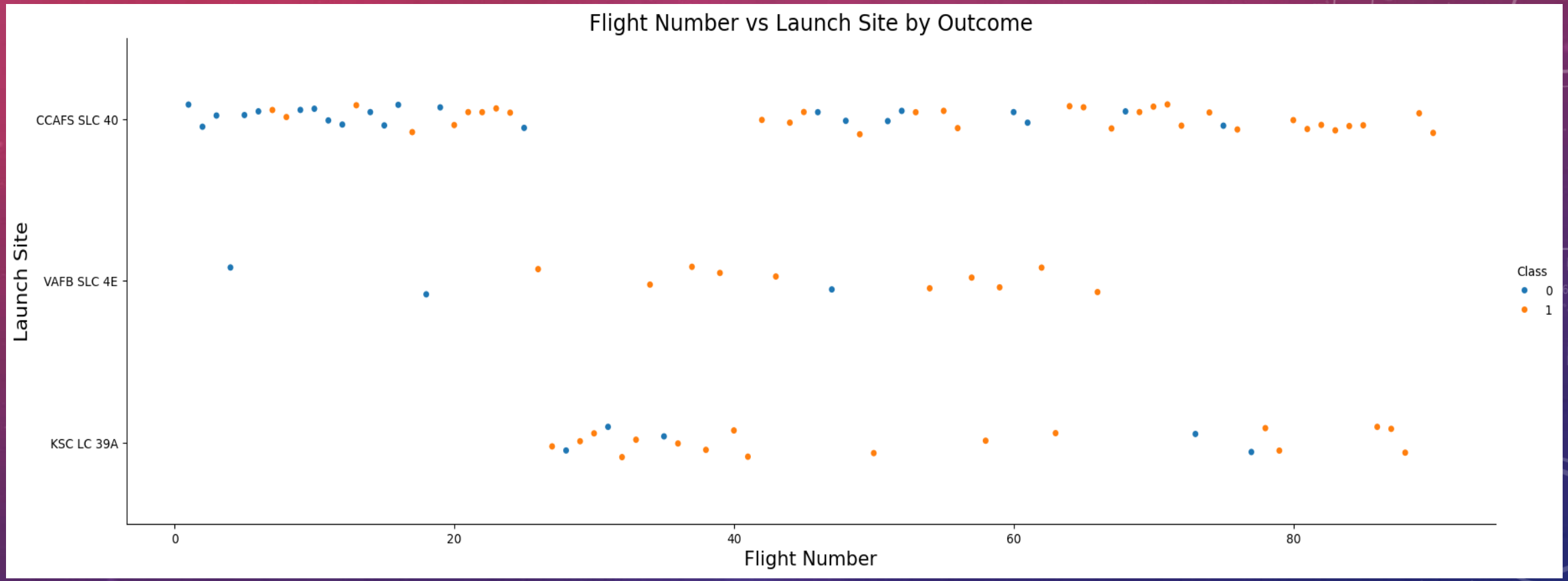
RESULTS

EDA WITH VISUALIZATION

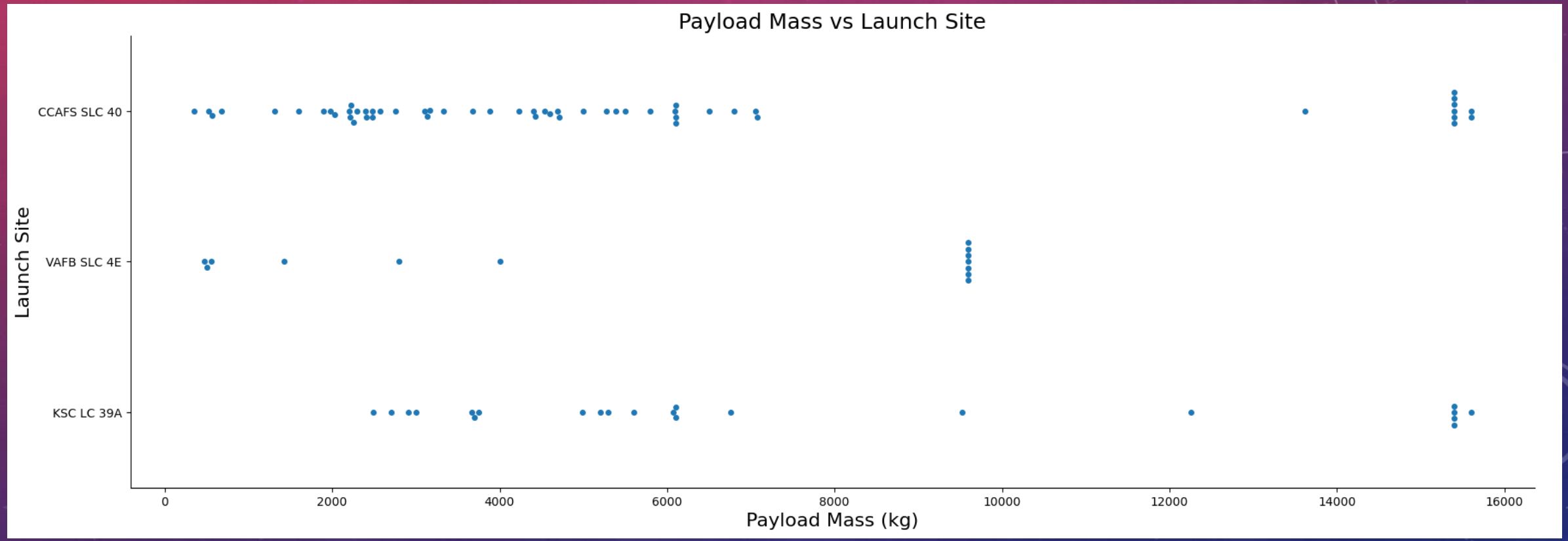
The relation between FlightNumber vs. PayloadMass

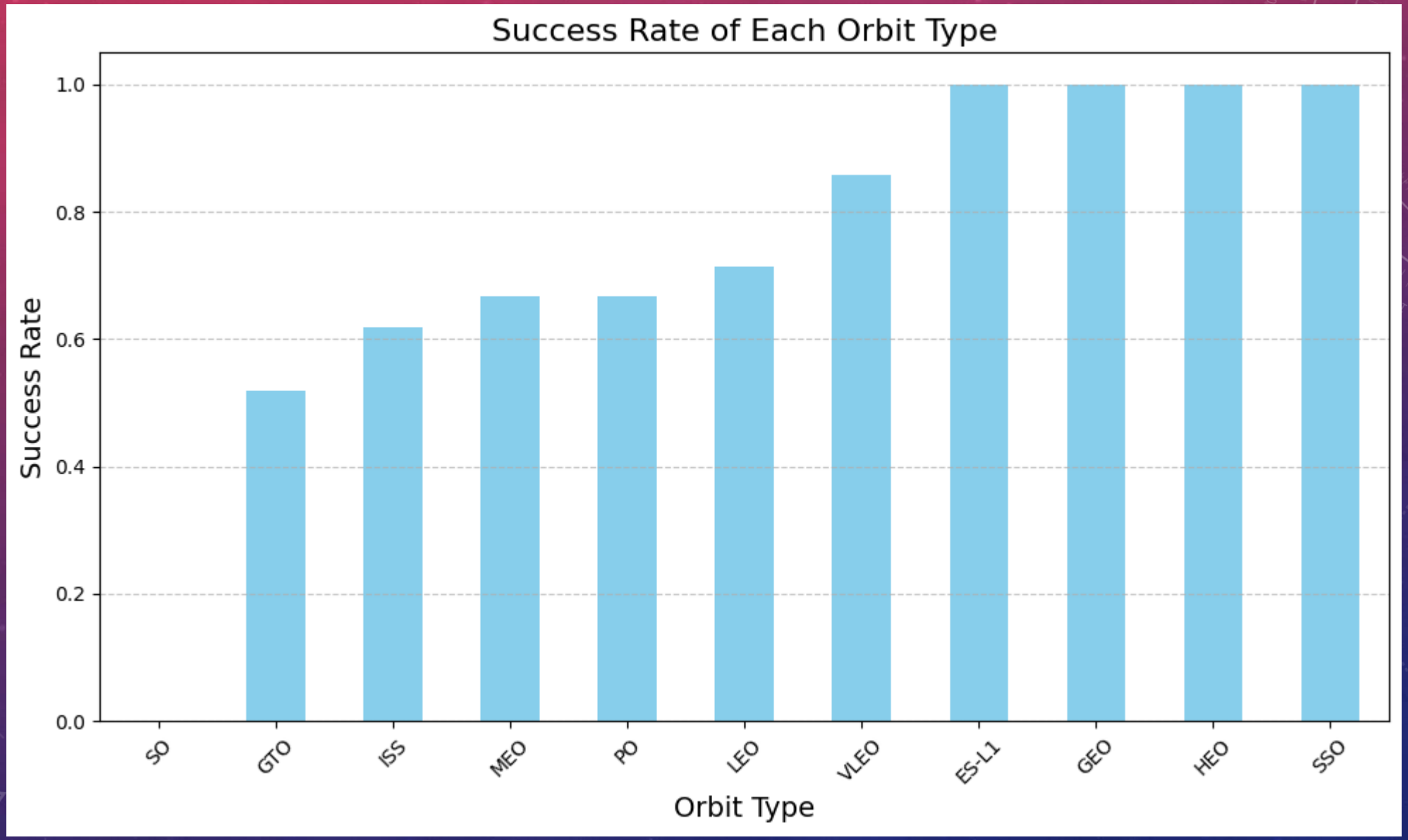


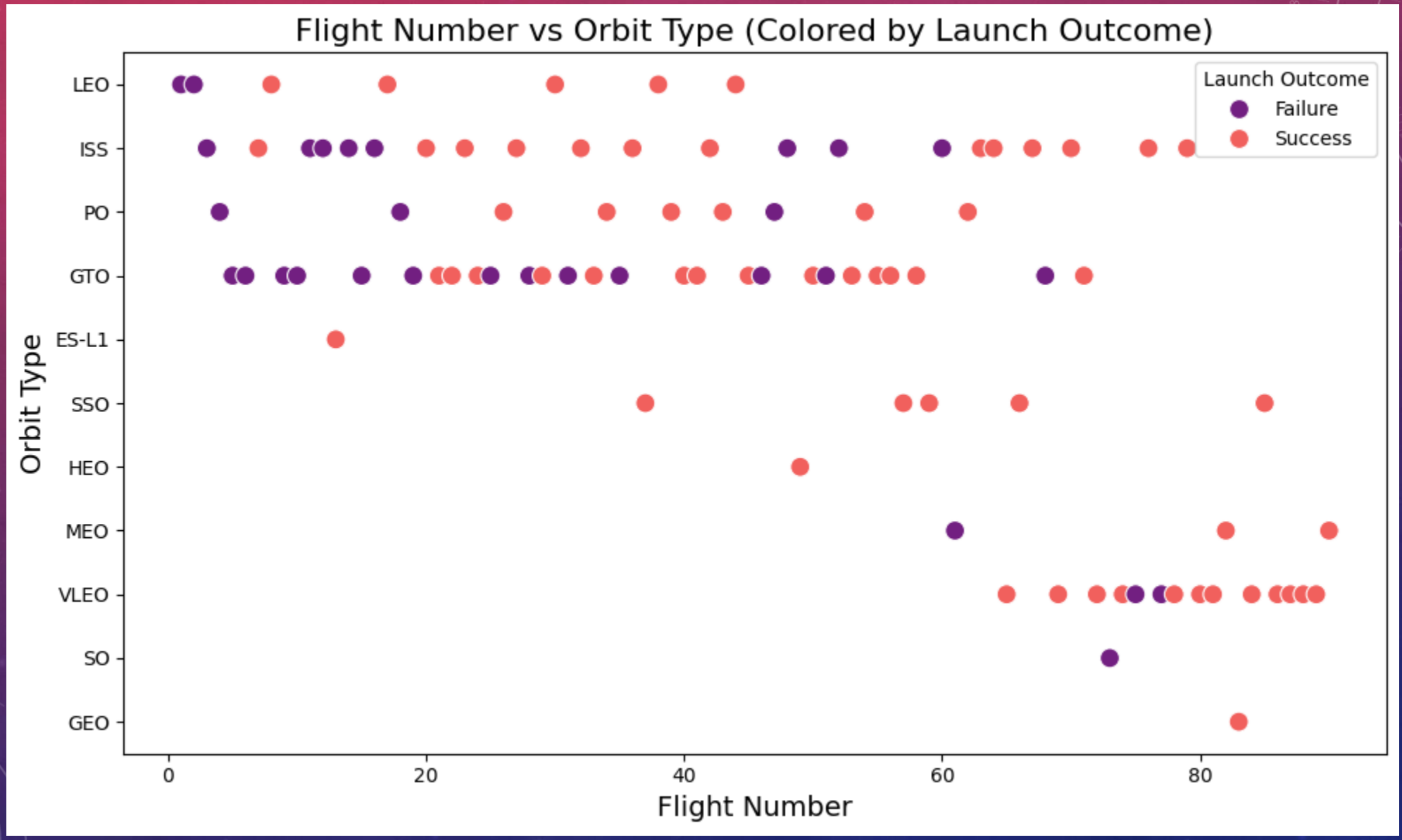
The relation between FlightNumber vs. Launch Site by Outcome

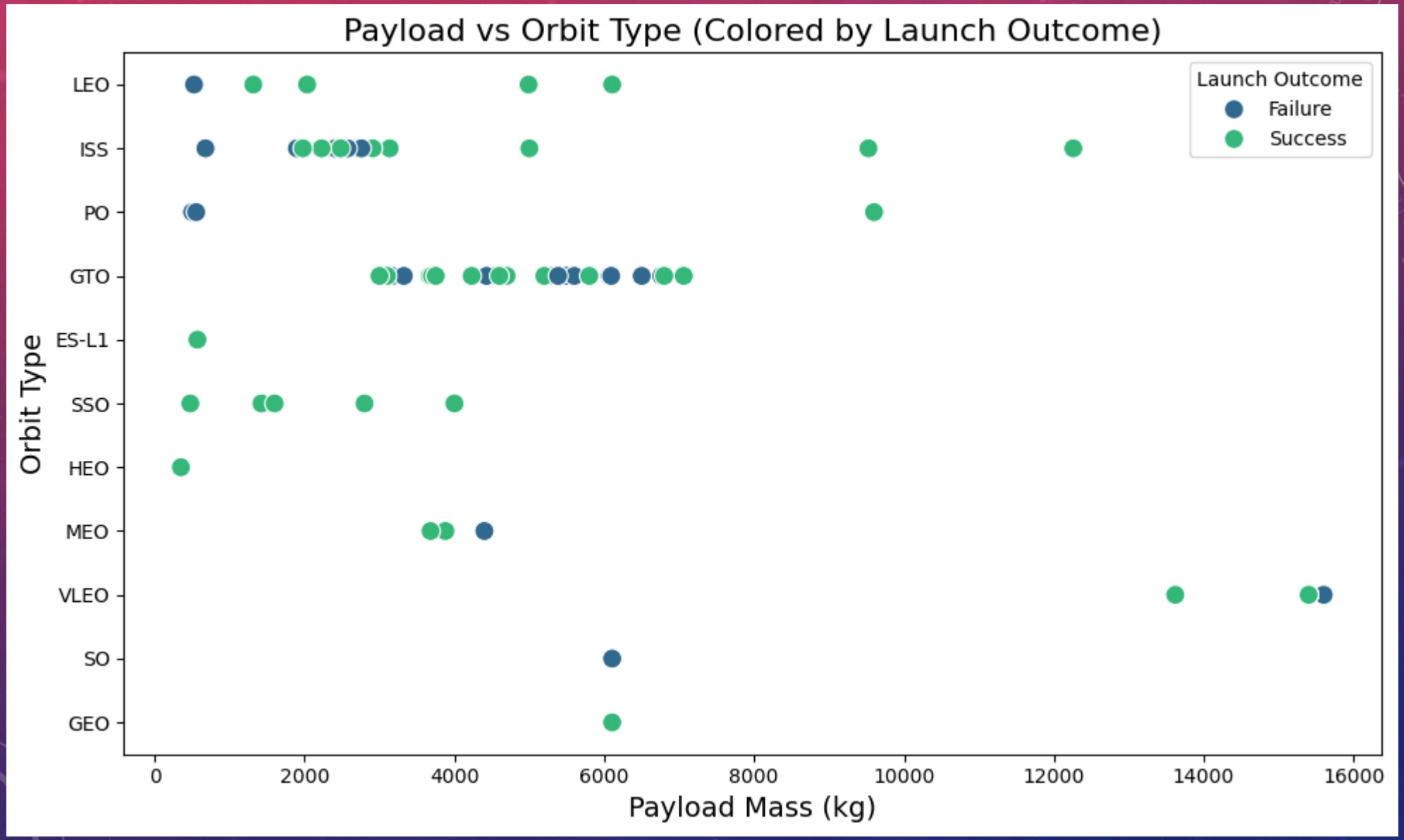


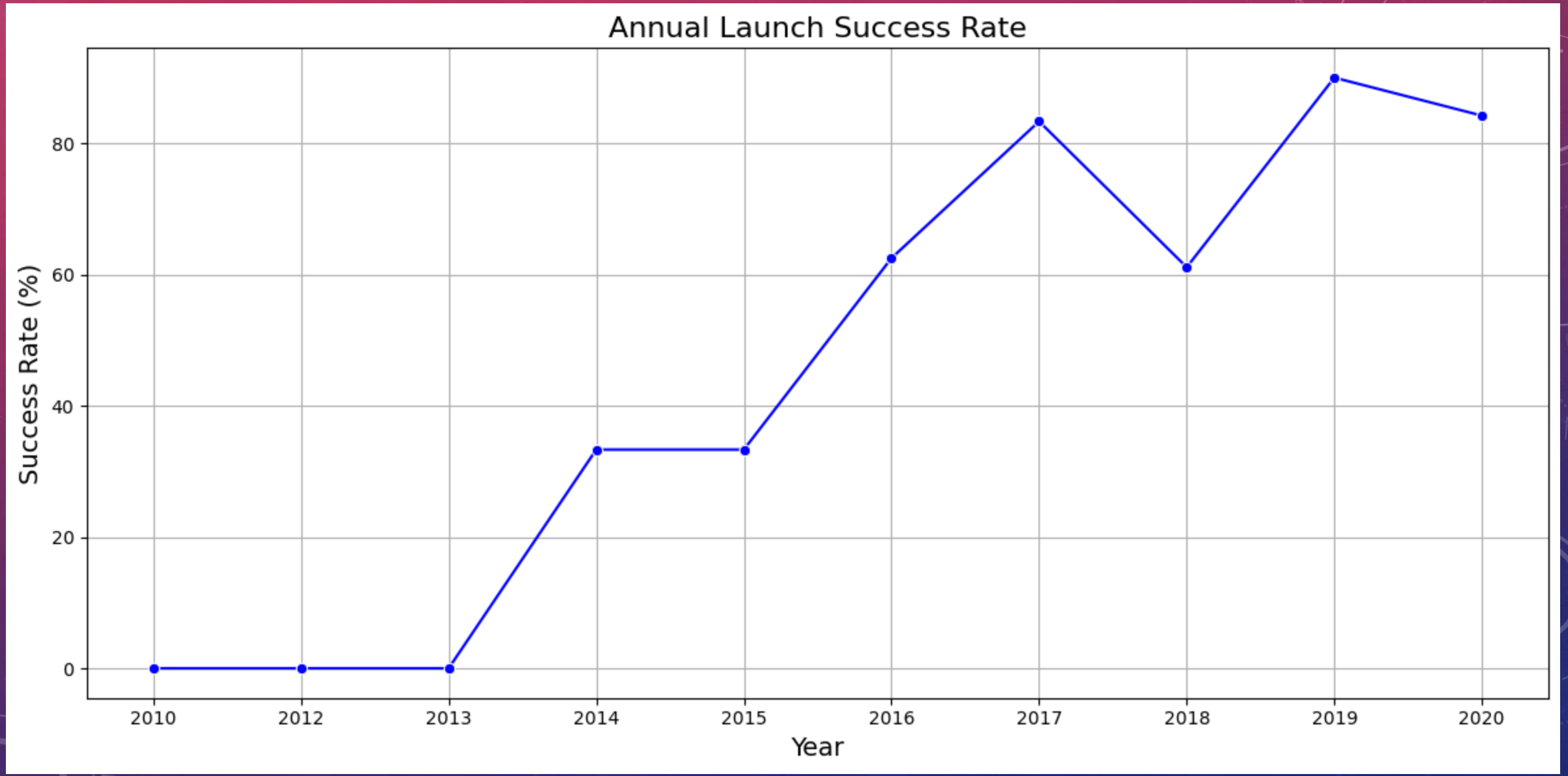
The relation between PayloadMass Vs Launch Site




Success Rate of Each Orbit Type
Success Rate
1.0
0.8
0.6
0.4
0.2
0.0
SO
GTO
ISS
MEO
PO
LEO
VLEO
ES-L1
GEO
HEO
SSO
Orbit Type


Flight Number vs Orbit Type (Colored by Launch Outcome)
Launch Outcome
Failure
Success
Orbit Type
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
20
40
60
80
Flight Number


Payload vs Orbit Type (Colored by Launch Outcome)
Launch Outcome
Failure
Success
Orbit Type
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
2000
4000
6000
8000
10000
12000
14000
16000
Payload Mass (kg)


Annual Launch Success Rate
Success Rate (%)
0
20
40
60
80
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

EDA WITH SQL

The names of the unique launch sites in the space mission

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

5 records where launch sites begin with the string '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 (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

The total payload mass carried by boosters launched by NASA (CRS)

SUM("PAYLOAD_MASS__KG_")
45596

Average payload mass carried by booster version F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2928.4

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

MIN("Date")
2015-12-22

The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

The total number of successful and failure mission outcomes

Mission_Outcome	Total
Success	98

All the booster_versions that have carried the maximum payload mass

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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

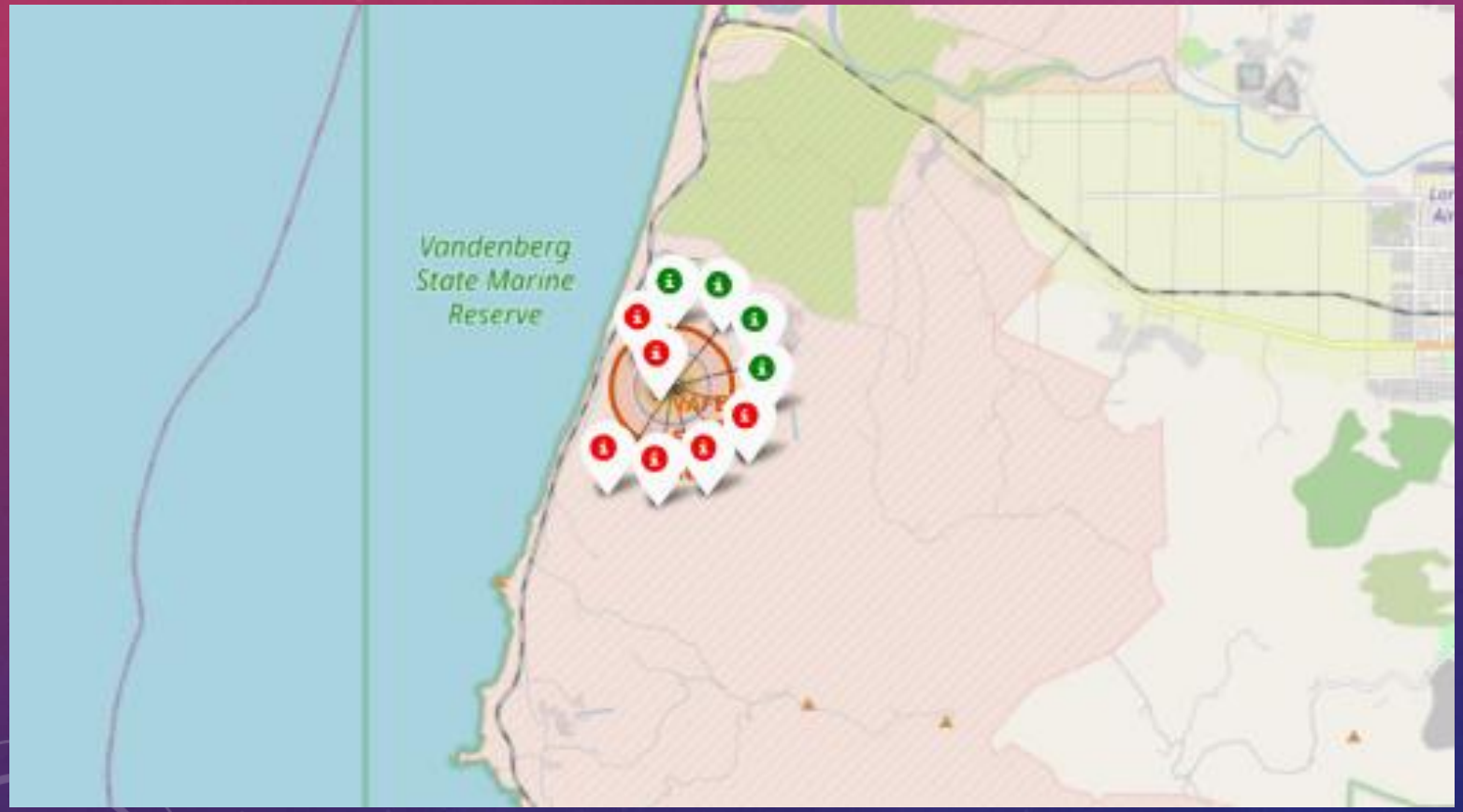
Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	CCAFS LC-40

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

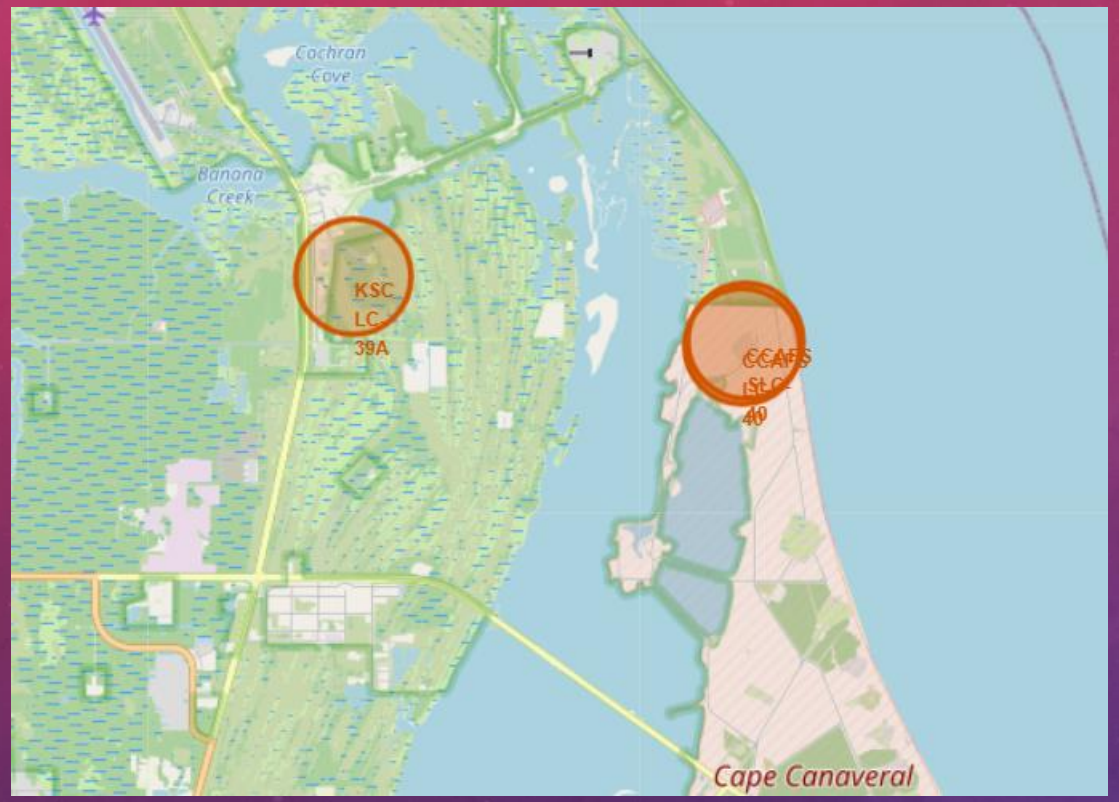
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

INTERACTIVE MAP WITH FOLIUM

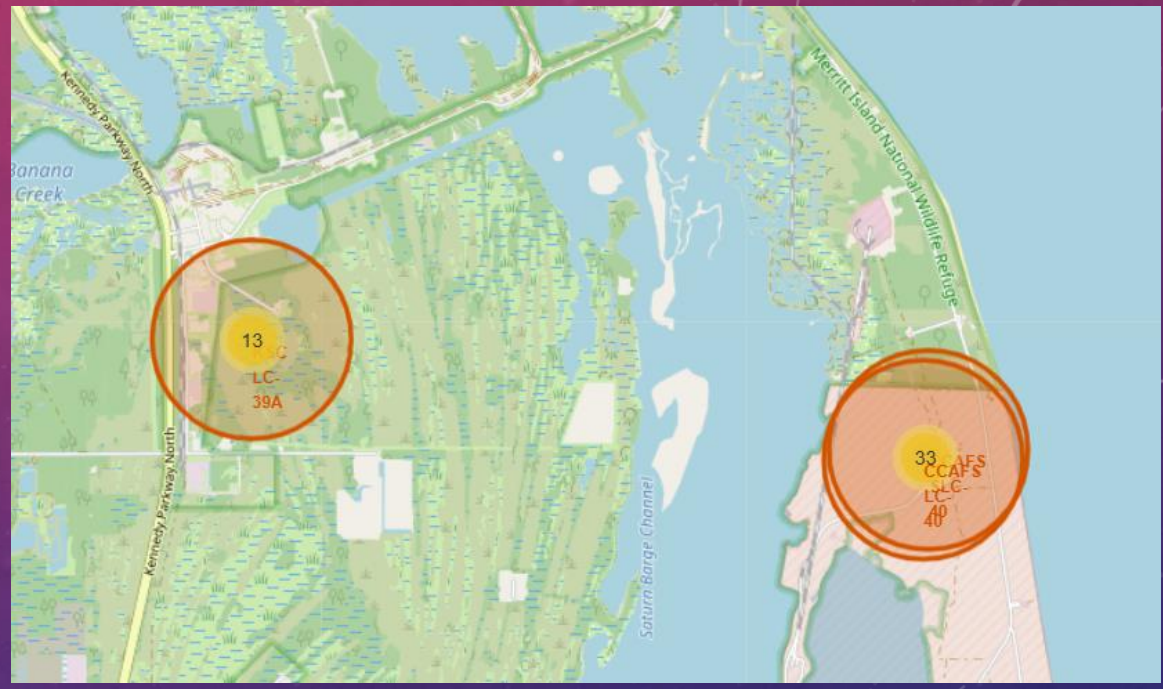
The succeeded launches and failed launches for each site on map each green tag represents a successful launch while each red tag represents a failed launch



Folium Circle on map for each site



Cluster with numeric value



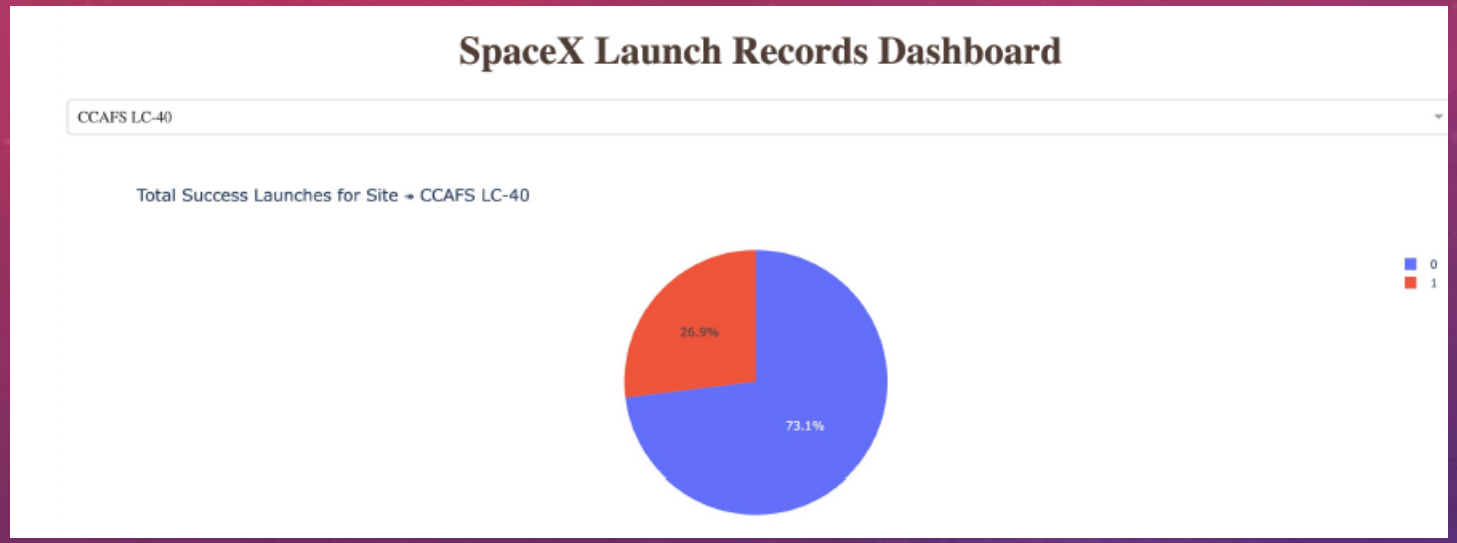
Proximity Analysis of Launch Sites

•The **distances** from the **VAFB SLC-4E** launch site to nearby geographical and infrastructure features such as:

- **Nearest city**: *Lompoc*
- **Nearest railway**: *Santa Barbara Subdivision MT1*
- **Nearest highway**: *Agena Way*
- **Nearest coastline**

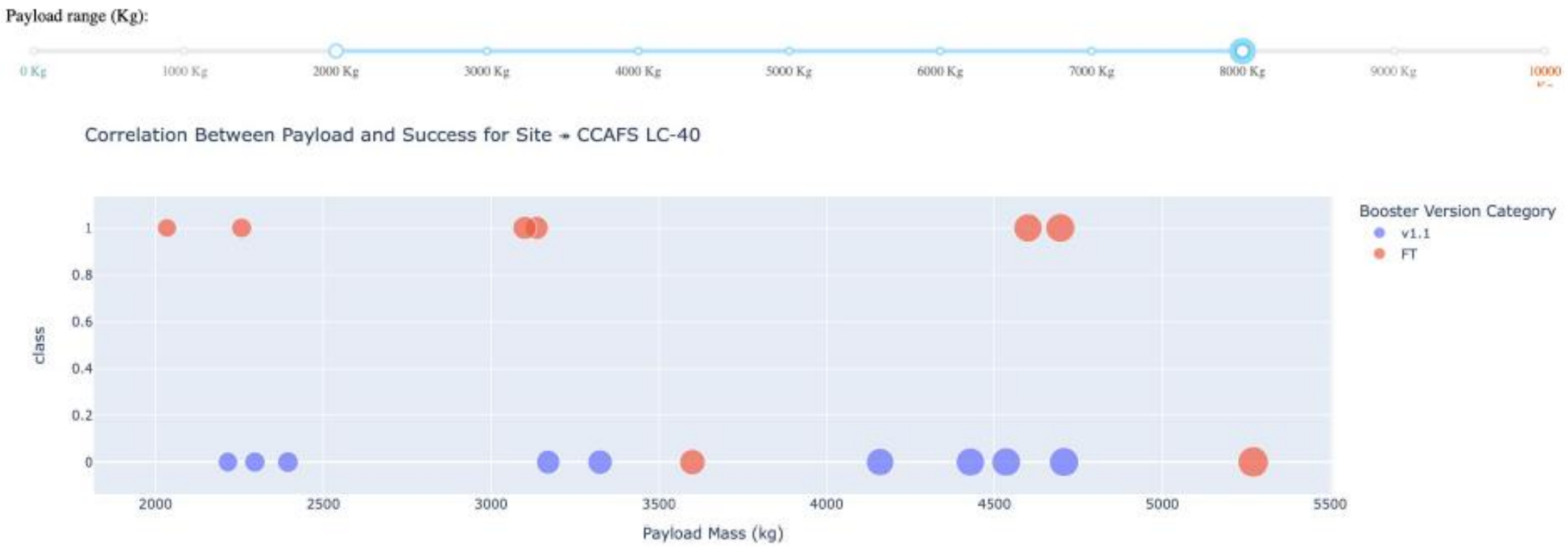
These proximities are visualized using interactive maps, allowing clear representation of the launch site's surroundings. Such spatial context can be useful for logistical planning and safety analysis.

PLOTLY DASH DASHBOARD



The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.

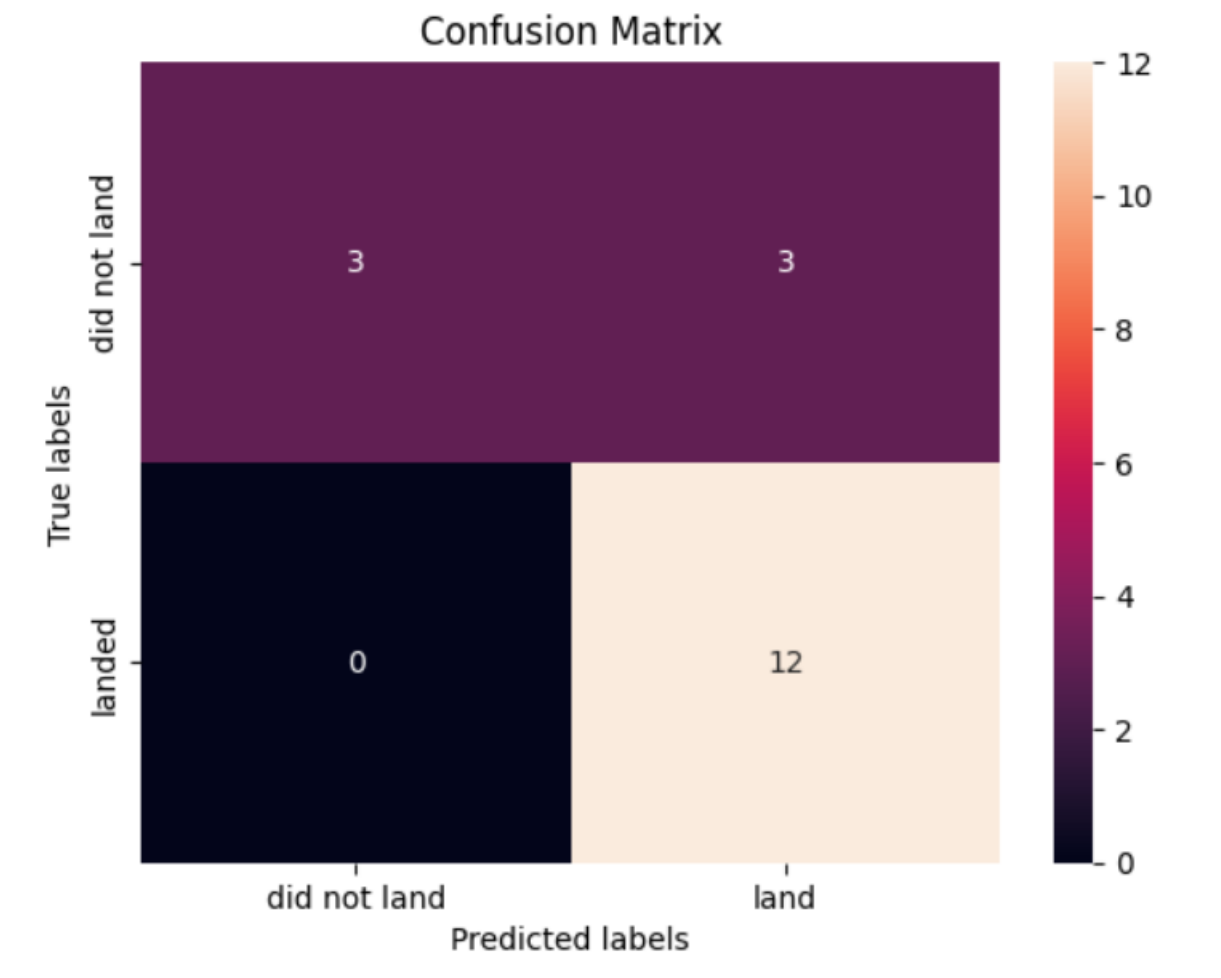
• 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches



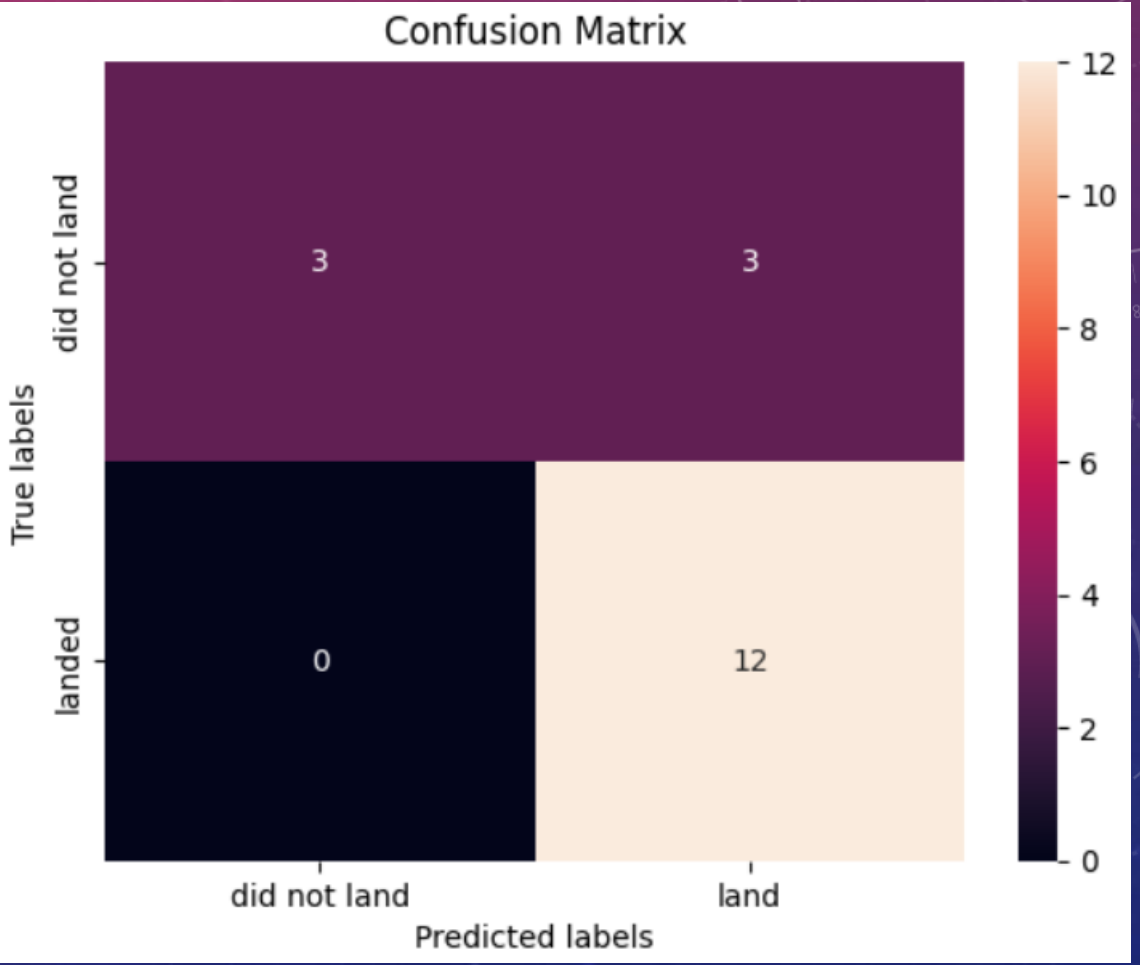
The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg. Class 0 represents failed launches while class 1 represents successful launches.

PREDICTIVE ANALYSIS (CLASSIFICATION)

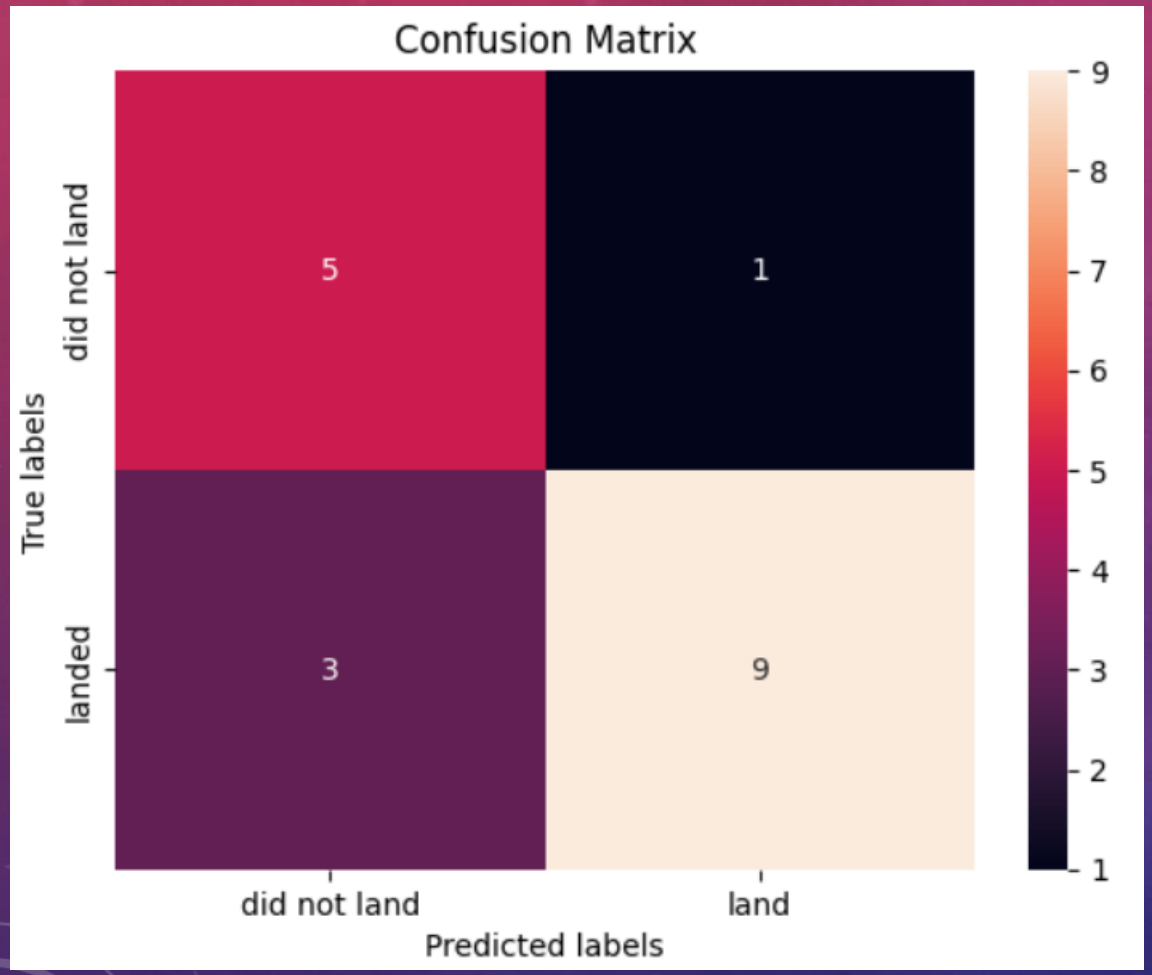
LOGISTIC REGRESSION



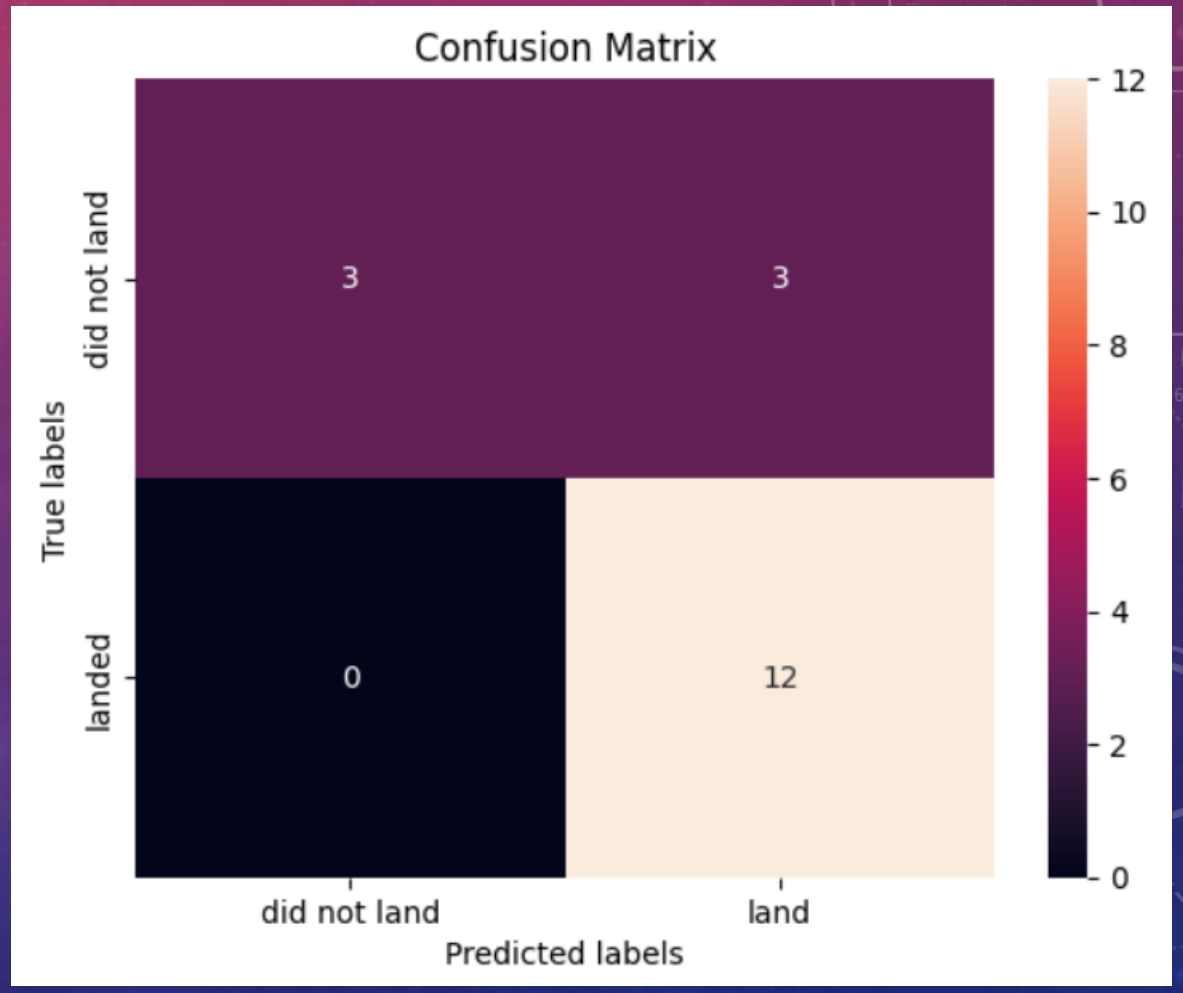
SVM



DECISION TREE



KNN



Test Set Evaluation

- All four models showed **identical accuracy scores** and **confusion matrices** on the test set.
- Therefore, **GridSearchCV best scores** were used for final ranking.

Model Ranking (Based on GridSearchCV Best Scores)

1. **Decision Tree** — *Best Score: 87.53%* ✅
2. **K-Nearest Neighbors (KNN)** — *Best Score: 84.82%*
3. **Support Vector Machine (SVM)** — *Best Score: 84.82%*
4. **Logistic Regression** — *Best Score: 84.64%*

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.875000
KNN	0.848214

Decision Tree outperformed the others, showing the highest predictive capability based on cross-validation.

DISCUSSION

Feature Impact and Predictive Approach

Data visualizations suggest that certain features may be correlated with mission outcomes. For instance, missions carrying heavier payloads tend to have higher landing success rates when targeting orbits such as Polar, LEO, and ISS. In contrast, for missions directed toward GTO (Geostationary Transfer Orbit), the distinction is less clear, as both successful and unsuccessful landings are observed.

These observations indicate that individual features likely influence mission success to some degree. However, the exact nature and strength of these relationships can be complex and challenging to interpret manually. To better understand these patterns and make reliable predictions, machine learning algorithms can be employed. These models can learn from historical data to identify hidden patterns and provide predictions about mission outcomes based on key input features.

CONCLUSION

This project aimed to predict whether the first stage of a Falcon 9 rocket would successfully land, an important factor in estimating launch costs and enhancing the efficiency of future missions. Each feature related to a Falcon 9 launch—such as payload mass, orbit type, booster version, and launch site—was analyzed to assess its influence on mission success.

Our findings highlight several key insights:

•**Launch Site Performance**: The **CCAFS LC-40** launch site demonstrated the highest success rate, accounting for approximately **43.7%** of all successful missions. This suggests that this location may offer advantageous conditions or more refined operational procedures that contribute to its higher performance.

•**Booster Version Impact**: The **FT booster version** stood out for its consistent success across a wide range of payloads, indicating strong reliability. This insight supports the idea that utilizing this version in future missions could enhance mission success rates.

•**Payload Mass Observations**: Contrary to potential assumptions, there was **no strong correlation** between heavier payloads and reduced success rates. This suggests that other factors—such as the launch site and booster technology—may play a more pivotal role in influencing mission outcomes.

•**Visualization and Insights**: The use of **interactive visualizations** through **Folium and Plotly Dash**
enabled a deeper exploration of launch patterns and geographical factors. These tools provided stakeholders with clear and actionable insights, enhancing their understanding of the data.

From a modeling perspective, several machine learning algorithms were implemented to identify patterns in historical launch data. Among them, the **Decision Tree classifier** delivered the most accurate predictions, outperforming other models such as SVM and KNN. This model demonstrated its potential as a valuable tool for predicting mission success based on launch features.

THANK YOU