PREDICTING SPACEX LAUNCH SUCCESS WITH DATA SCIENCE

Fabian Pulgar

EXECUTIVE SUMMARY

OBJECTIVE

- Predict the success of Falcon 9 first-stage landings to optimize launch costs and provide strategic insights for competitive bidding.
- Apply data science methodologies to address a real-world business problem.

METHODOLOGY

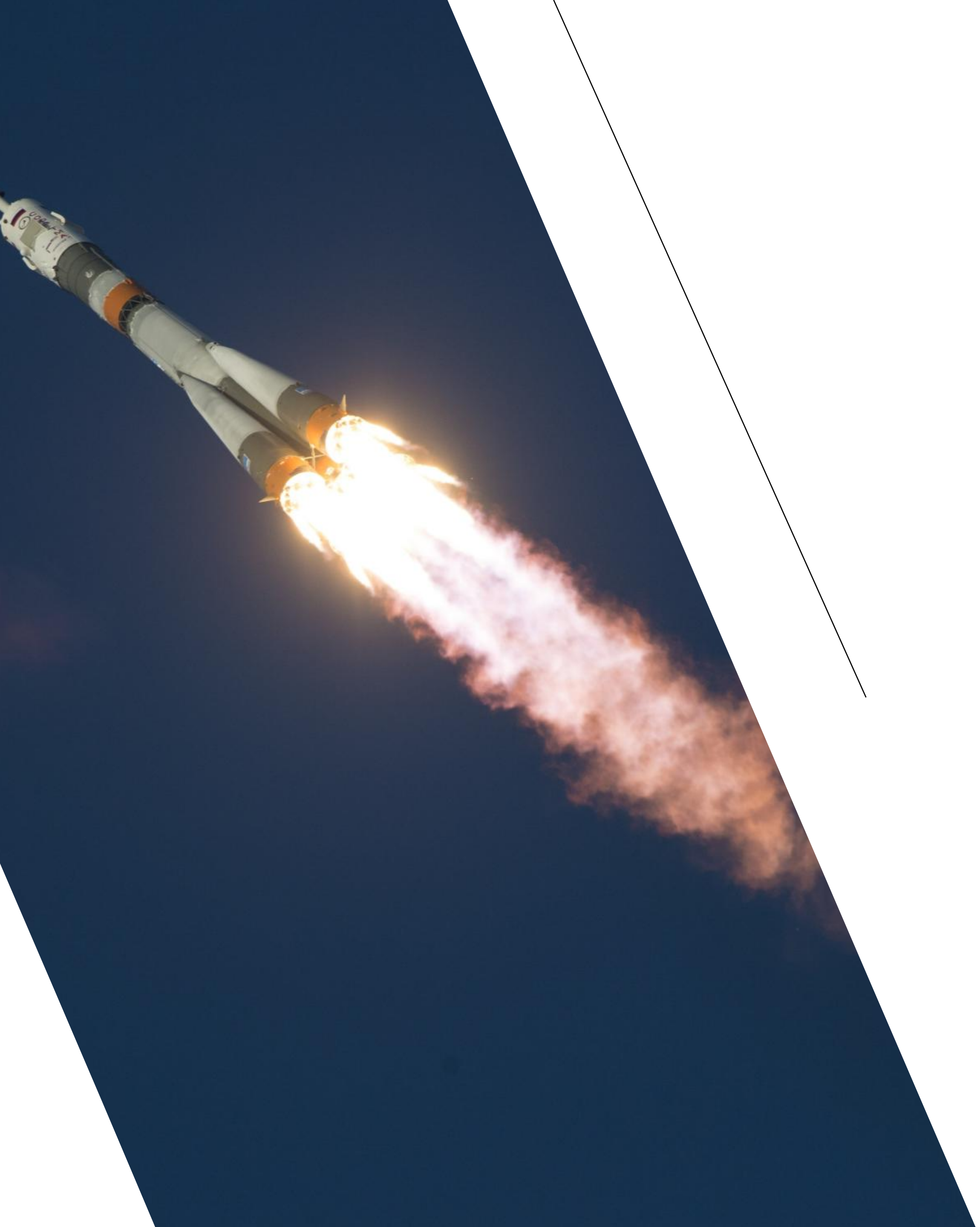
- Data collection using RESTful APIs and web scraping.
- Exploratory Data Analysis with created visualizations and SQL queries.
- Interactive tools using dashboards and interactive maps.
- Predictive modeling for landing success.

INSIGHTS

- Payload mass significantly influences the likelihood of a successful landing.
- Specific launch sites show higher success rates, revealing operational advantages.
- Interactive maps and dashboards provide actionable insights for decision-makers.

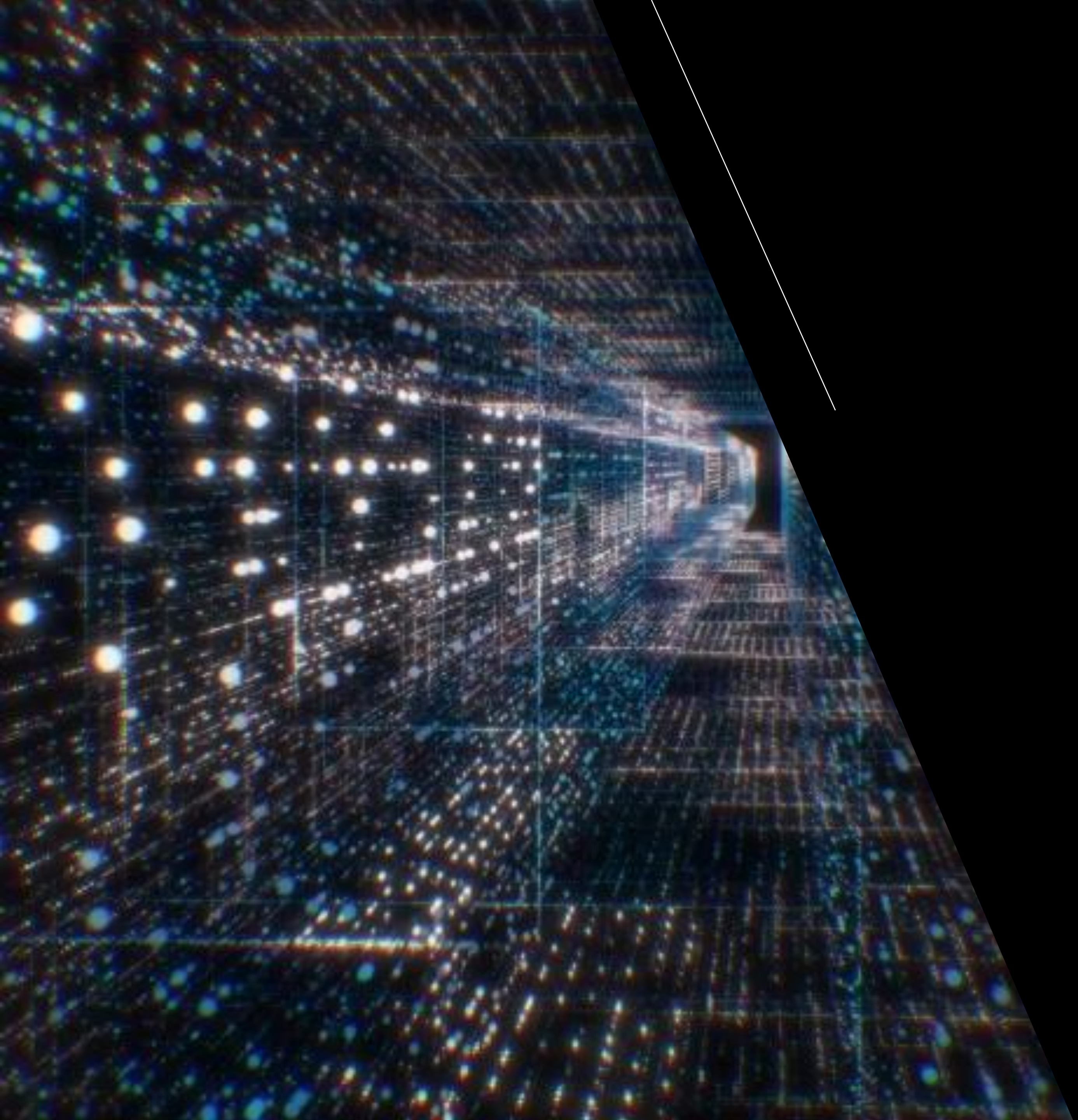
IMPACT

- Supports SpaceX and competitors in optimizing launch operations and reducing costs.
- Enhances strategic decision-making through accessible, interactive visualizations.
- Demonstrates how data science can tackle complex business challenges effectively.



INTRODUCTION

- This project focuses on analyzing and predicting the success of SpaceX's Falcon 9 first-stage landings.
- SpaceX revolutionized the aerospace industry by significantly reducing launch costs through reusable rocket technology.
- The primary goal of this project is to leverage data science techniques to understand key factors influencing launch success and develop predictive models to estimate the likelihood of successful landings.
- To achieve this, we utilized a structured approach, starting with data collection and cleaning, followed by exploratory analysis, interactive visualizations, and predictive modeling.



DATA COLLECTION AND DATA WRANGLING METHODOLOGY

DATA SOURCES

Space RESTful API

- Provided detailed data on SpaceX launches, including launch date and time, launch site location, first-stage landing outcome and payload details

Web Scraping

- Collected supplementary information such as launch sites and historical data using BeautifulSoup to extract data from HTML web pages.

TOOLS AND TECHNIQUES

Python libraries

- Pandas for data wrangling
- Requests for API calls
- BeautifulSoup for web scraping

Data transformation

- Converted JSON to DataFrame
- Cleaned missing values

FlightNumber	Date	BoosterVersion
1	2010-06-04	Falcon 9
2	2012-05-22	Falcon 9
3	2013-03-01	Falcon 9
4	2013-09-29	Falcon 9
5	2013-12-03	Falcon 9
6	2014-01-06	Falcon 9
7	2014-04-18	Falcon 9
8	2014-07-14	Falcon 9
9	2014-08-05	Falcon 9
10	2014-09-07	Falcon 9
11	2014-09-21	Falcon 9
12	2015-01-10	Falcon 9
13	2015-02-11	Falcon 9

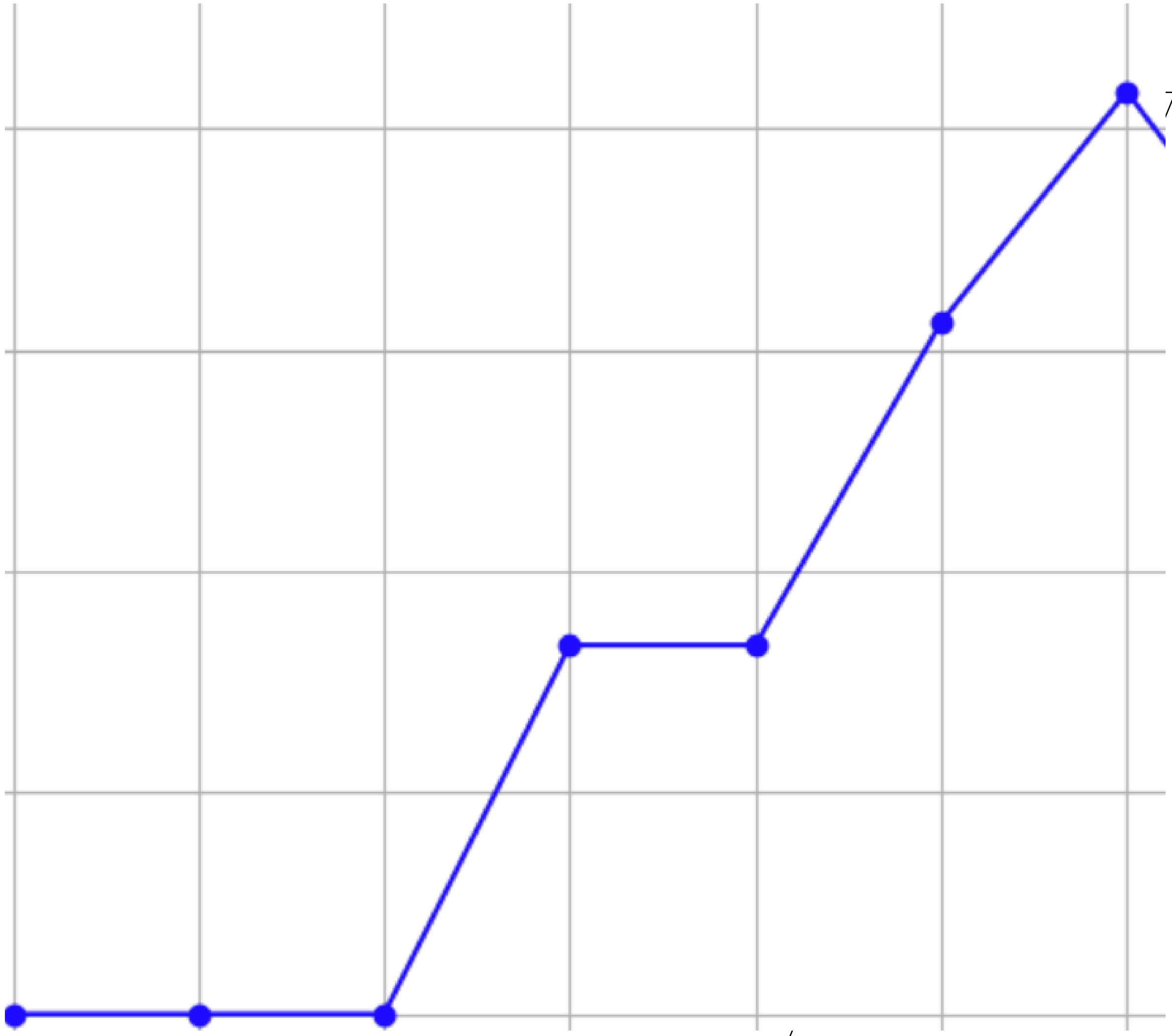
KEY STEPS

Extraction	Extract data from APIs and web scraping
Analysis	Format datasets and data understanding
Cleaning	Clean, prepare and filter data

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

TOOLS AND TECHNIQUES

- **SQL**: for exploratory queries and data aggregation
- **Pandas**: for data analysis and manipulation
- **Plotly express**: for creating scatter plots and pie charts
- **Plotly Dash**: for developing interactive charts
- **Folium**: for creating interactive maps to visualize launch sites and proximities.



Analysis

Analyzed trends and success rates by launch site

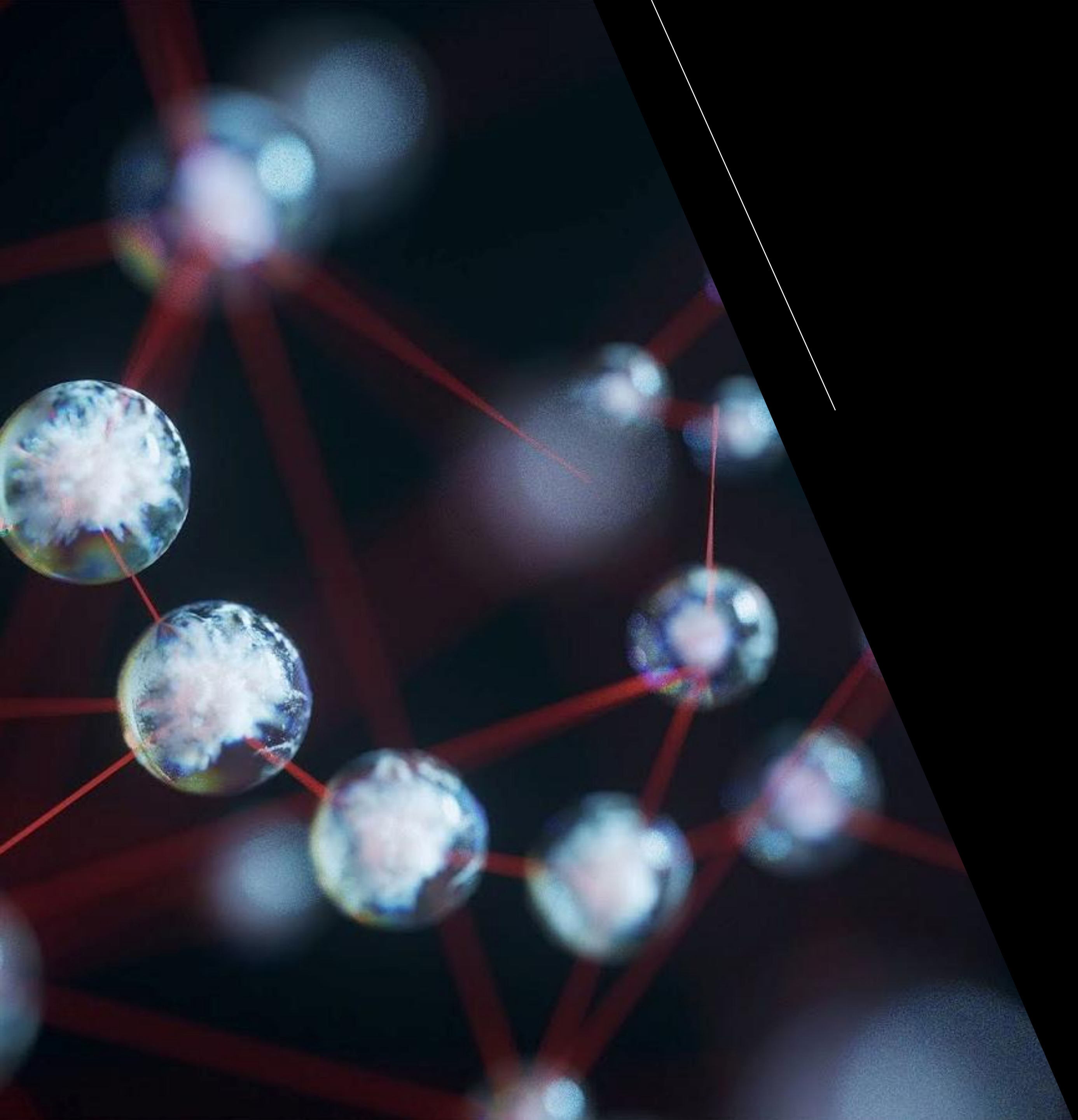
Visualizations

Dashboard for dynamic success rate exploration

Map visualizations

Highlight spatial relationships and proximities

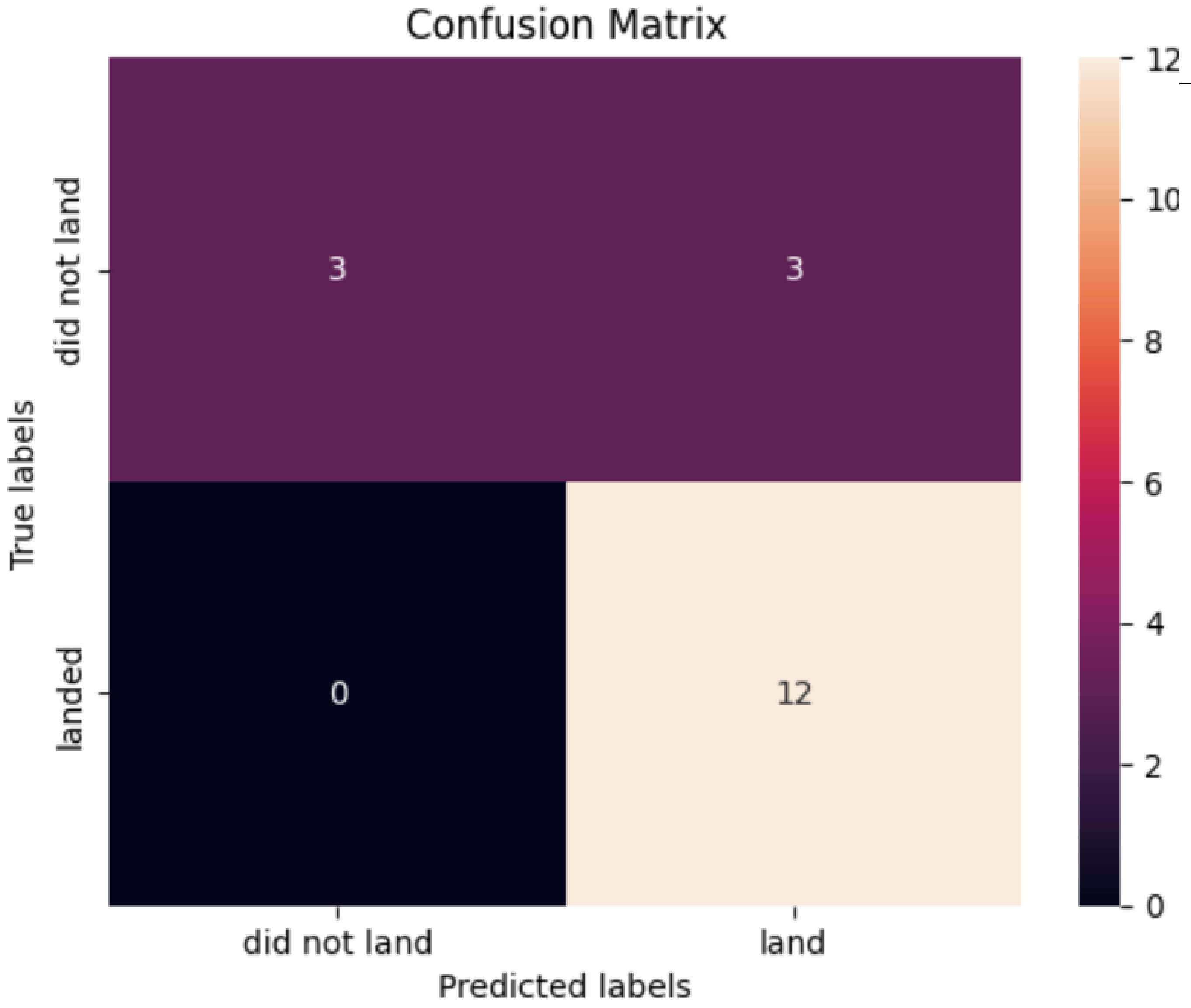
KEY STEPS



PREDICTIVE ANALYSIS METHODOLOGY

TOOLS AND TECHNIQUES

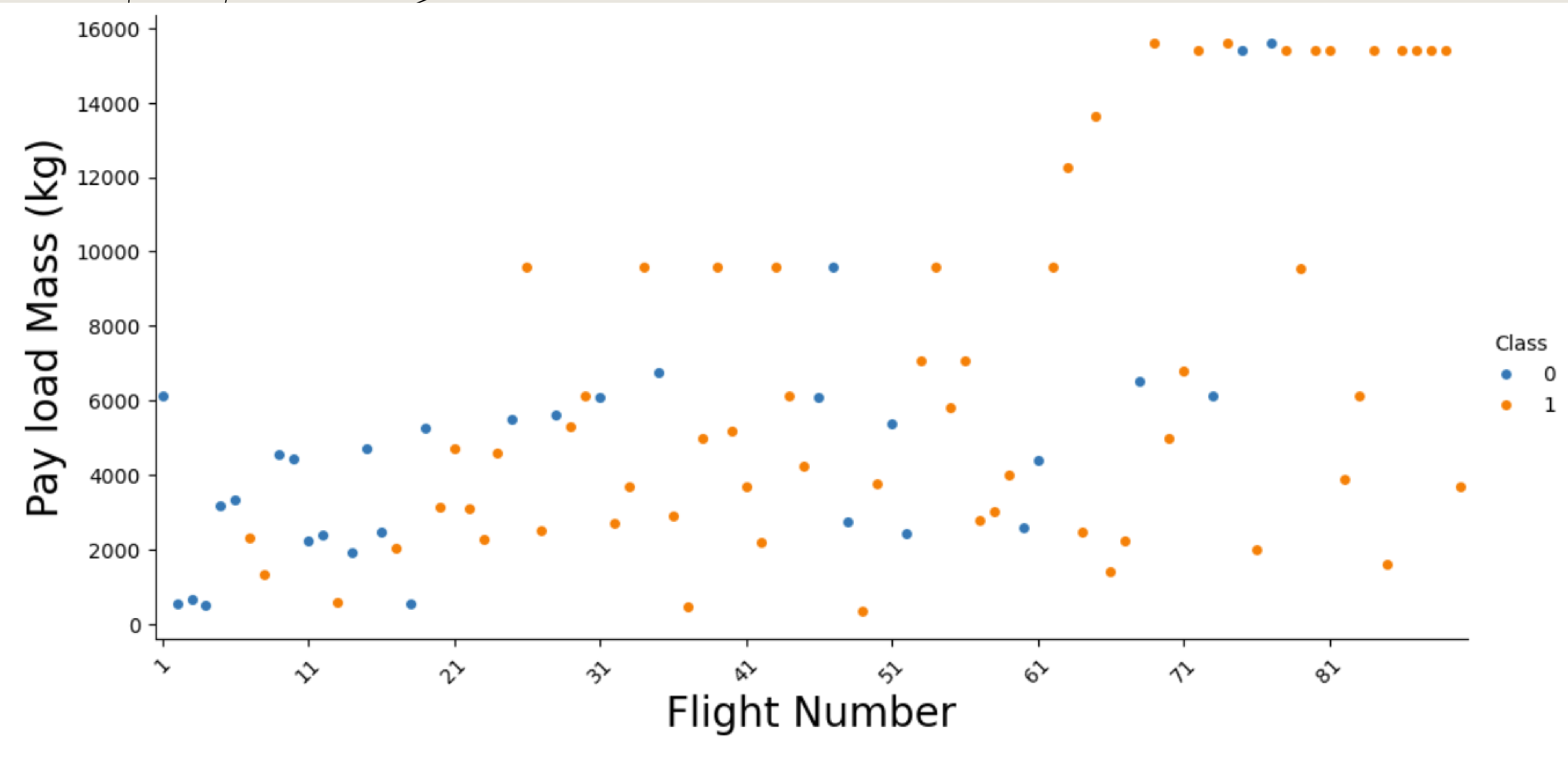
- **Scikit-learn**: used for building and evaluating models.
- **Grid search**: for hyperparameter optimization
- **Python**: for data preprocessing and analysis
- **Classification Models**: Logistic Regression, SVM and decision trees



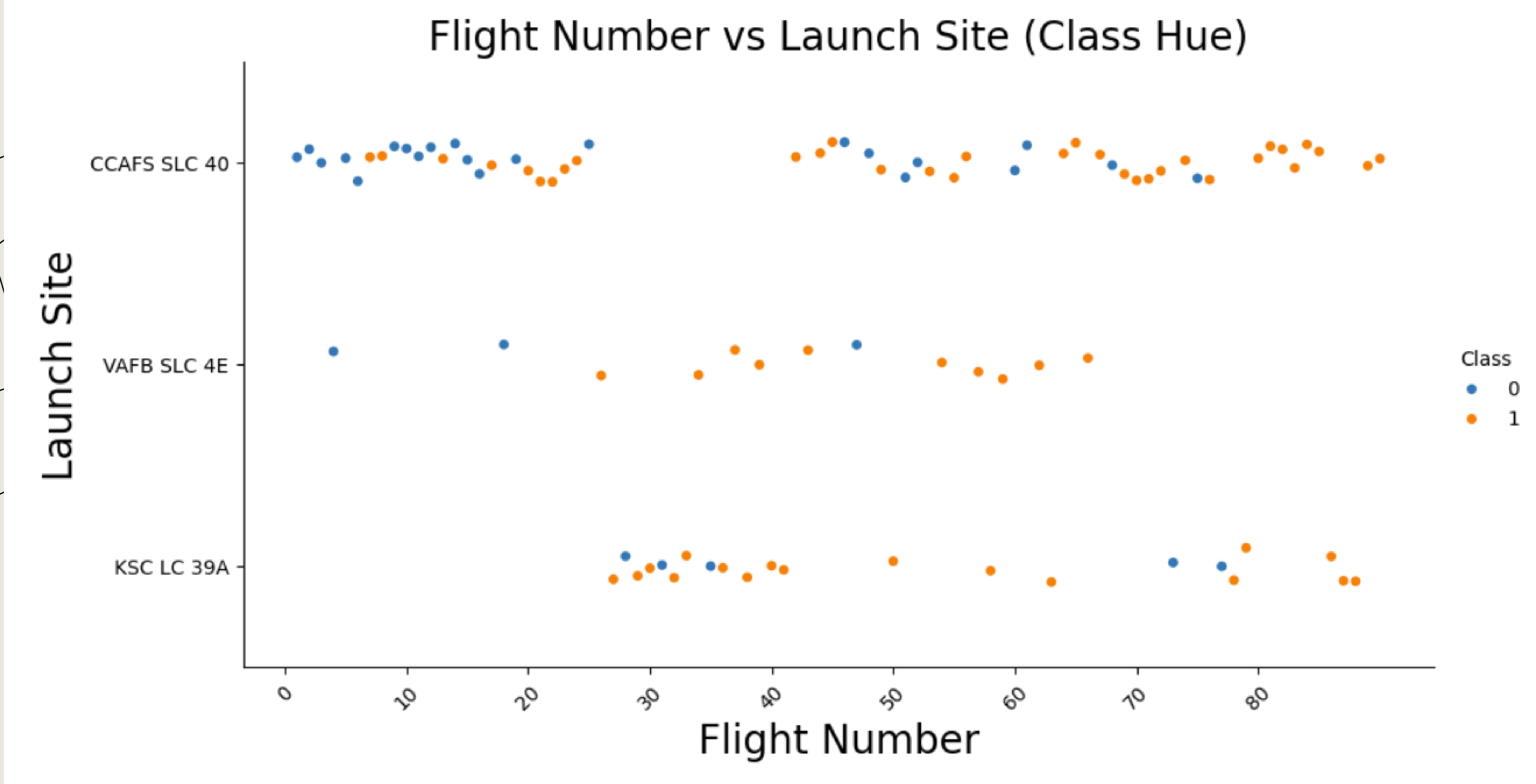
KEY STEPS FOR PREDICTIVE WORKFLOW

Data processing		Split data into training and testing sets.
Model training		Trained using SVM, decision trees and logistic regression
Hyperparameter Tuning		Used grid search to find optimal parameters
Model evaluation		Based on metrics like accuracy, precision, recall and f1-score

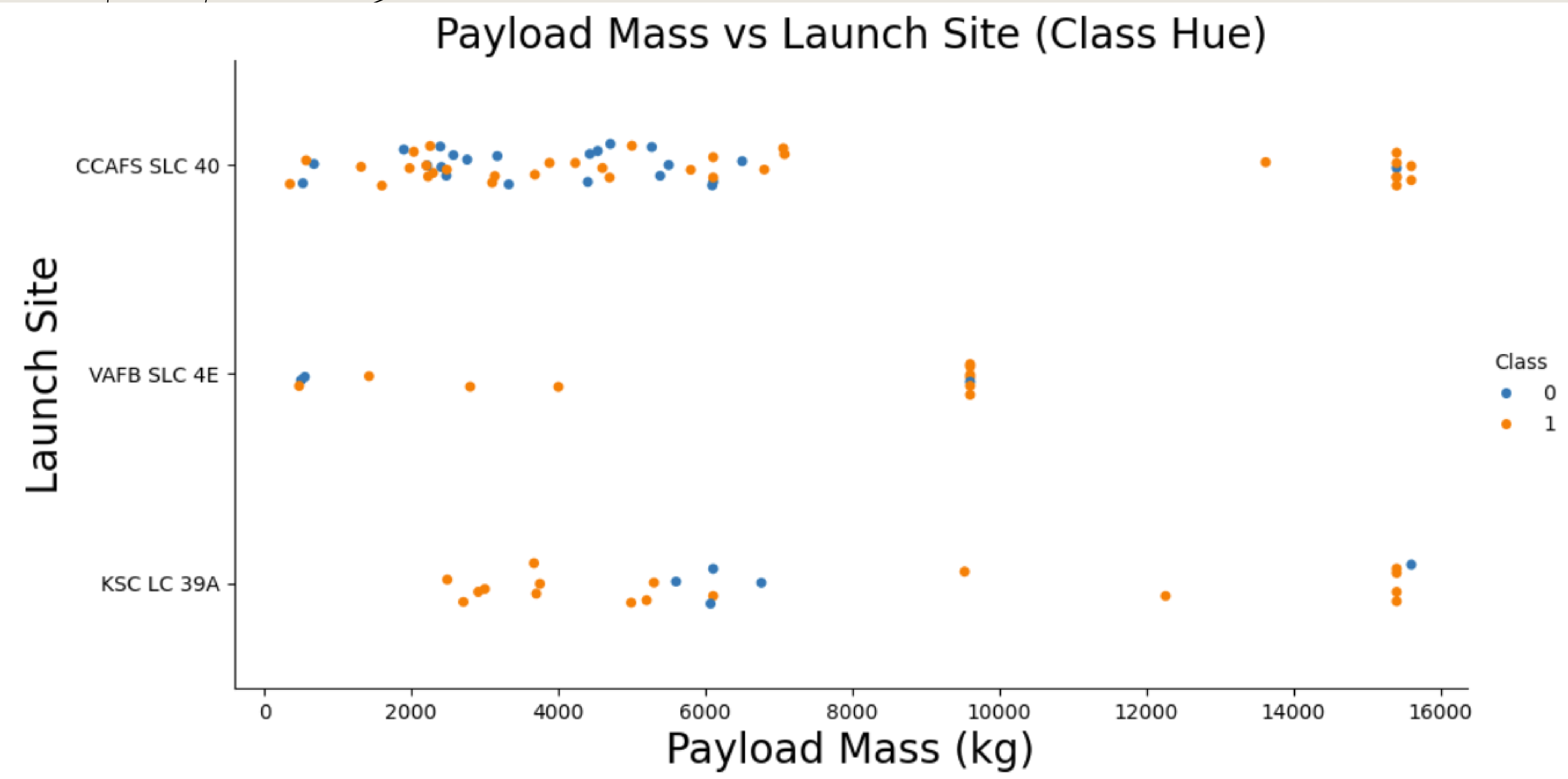
EDA VISUALIZATIONS AND SQL RESULTS



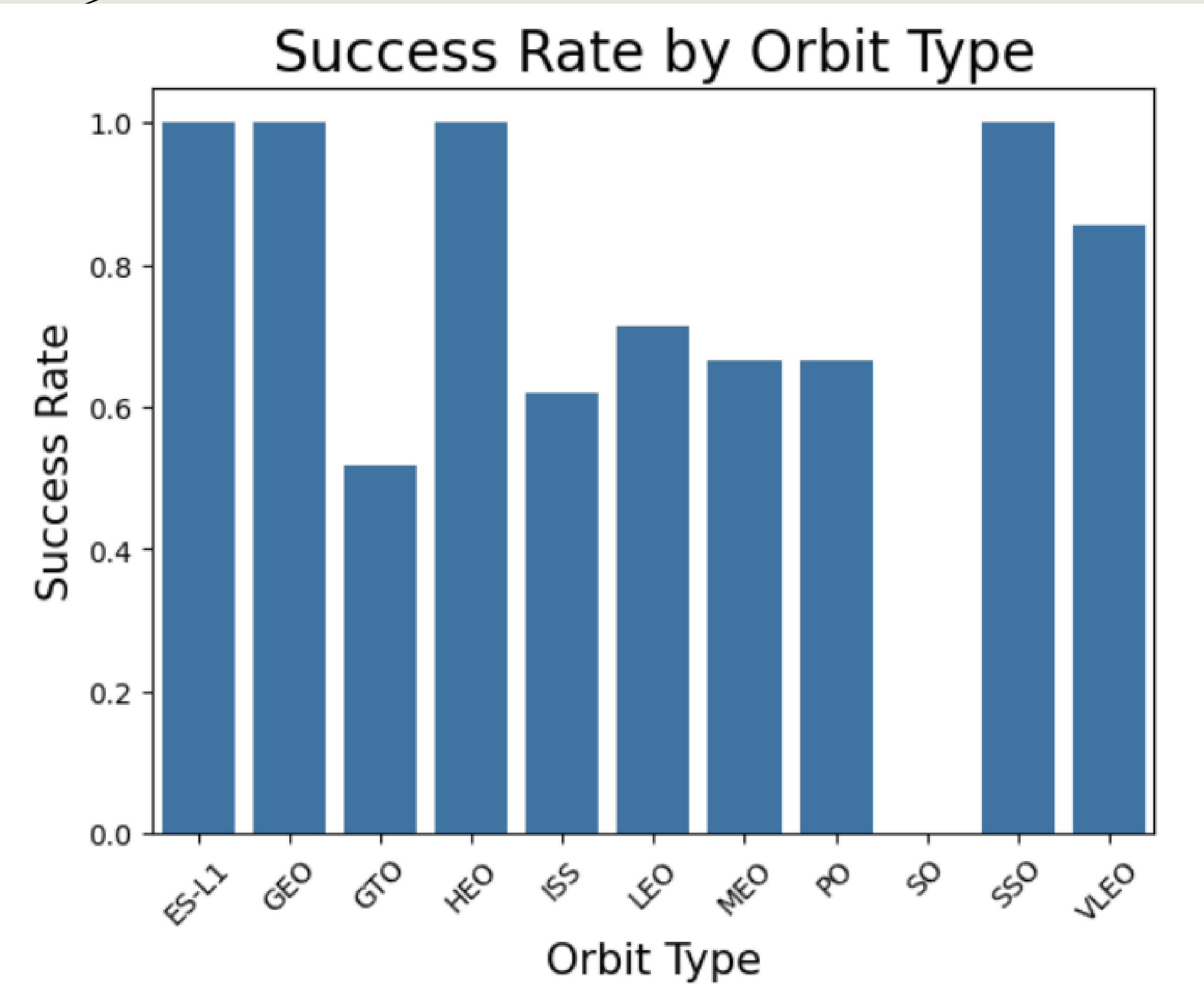
- The chart highlights a potential correlation between flight number and launch success.
- An optimal payload mass range can be observed, where launches are more likely to succeed, which is significant for future planning.



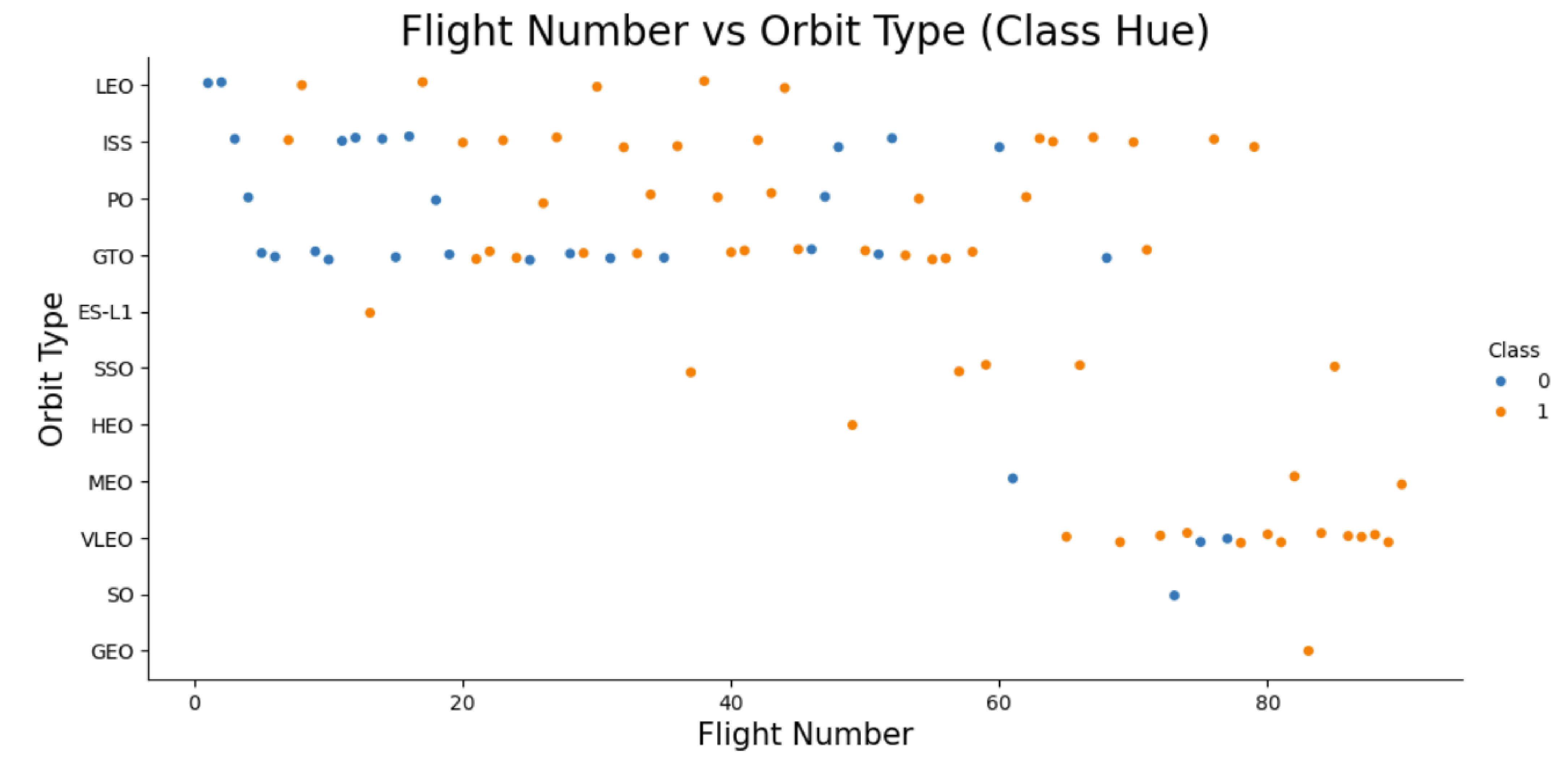
- CCAFS SLC 40: handles the majority of launches, showing an increasing number of successes over time.
- VAFB SLC 4E: results are more varied, but a trend toward increased success is noticeable in later flights.
- KSC LC 39A: Prominent in more recent flights and exhibits a high proportion of successes.



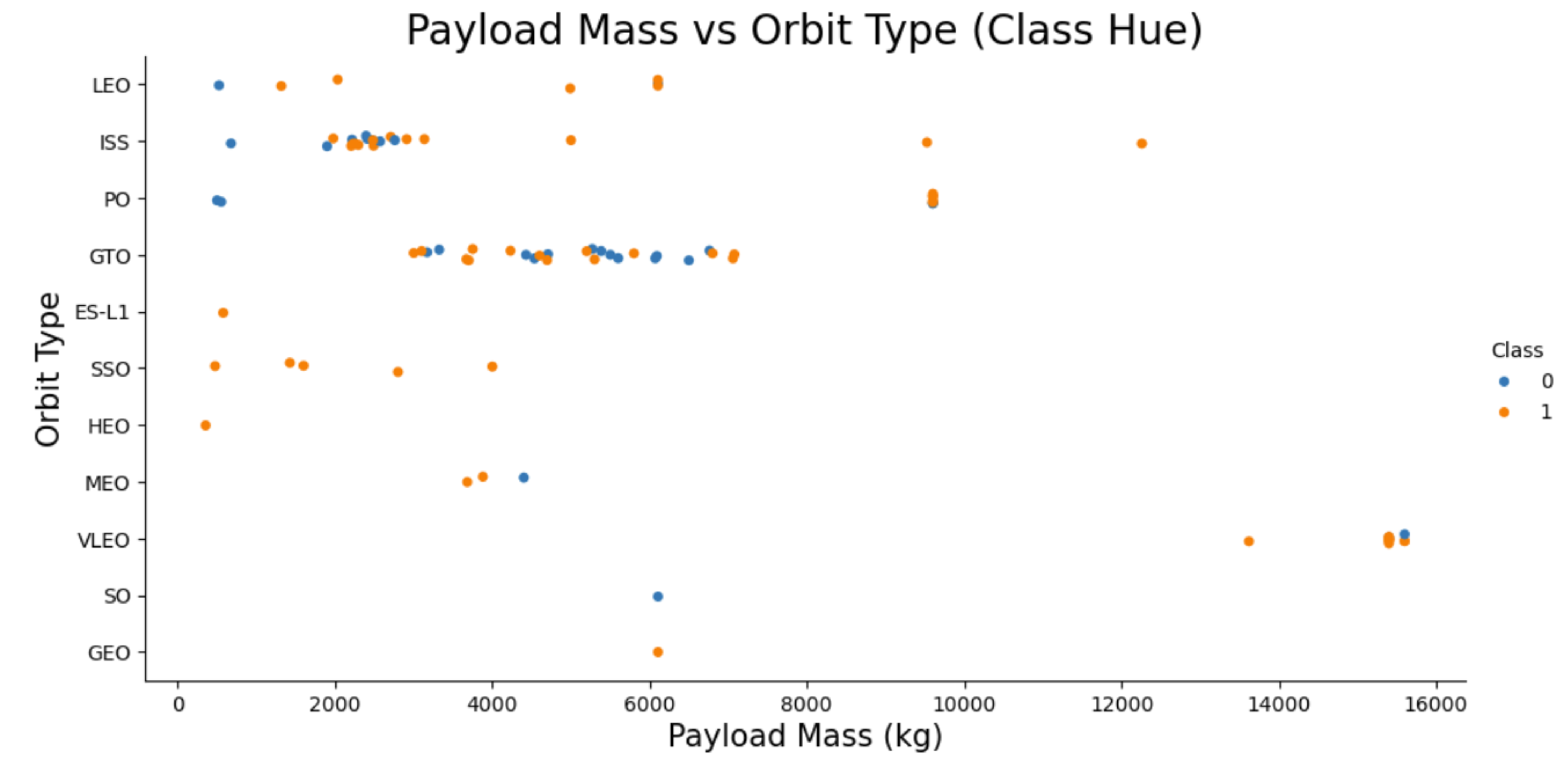
- There is a correlation between payload mass and launch success: larger payloads (>10,000 kg) are more likely to succeed.
- Certain launch sites, such as KSC LC 39A, appear more reliable for handling heavier payloads.
- While data is limited, in VAFB SLC 4E larger payloads seem to have a higher success rate.



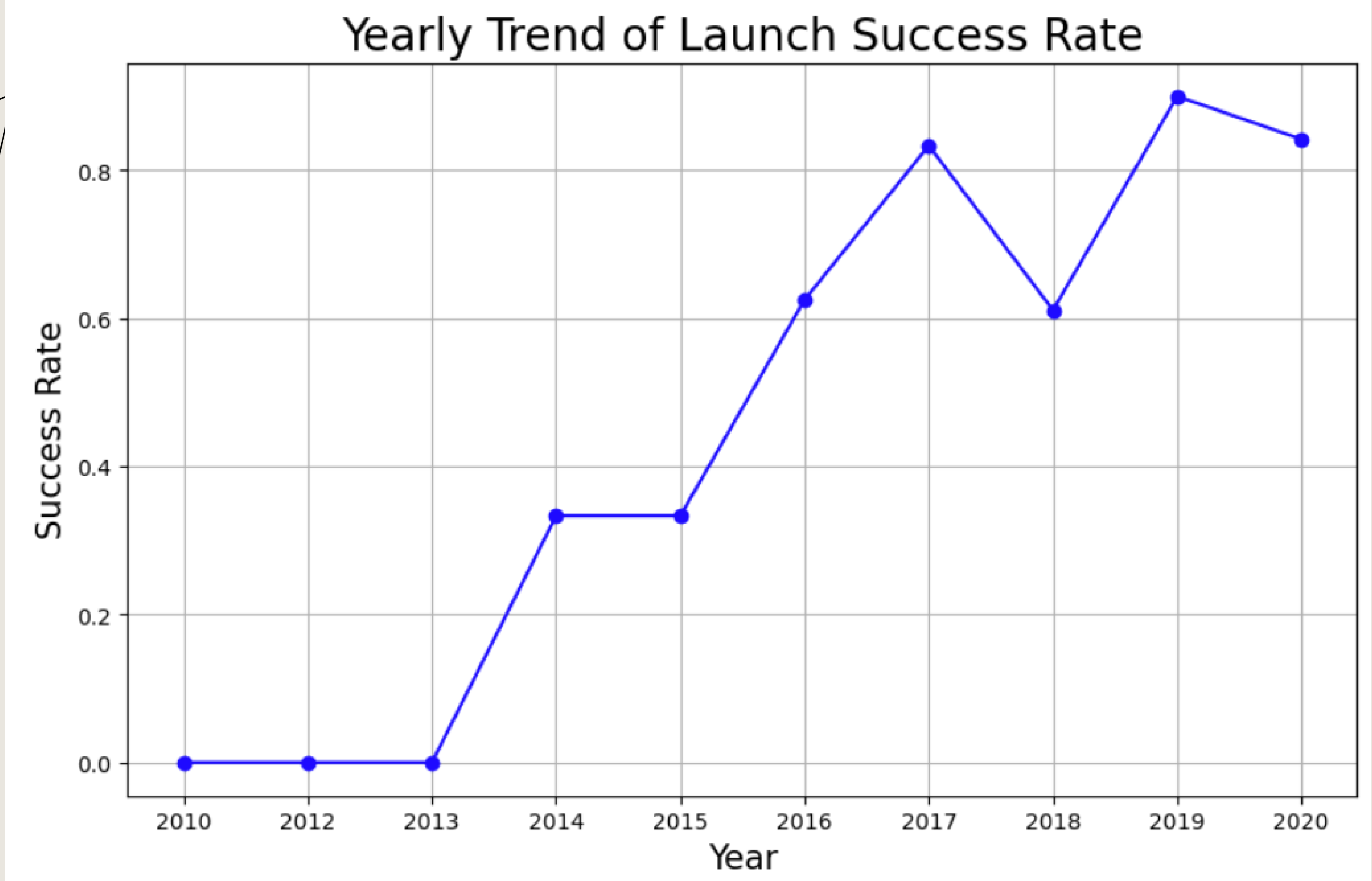
- Orbits such as ES-L1, GEO, HEO and SSO are the most reliable, with a 100% success rate.
- Missions to GTO require improvements due to their lower success rate.
- ISS, LEO, MEO y PO orbits show consistent but not perfect success rates, potentially due to specific mission challenges.



- GTO and PO orbits show more failures, which could suggest higher technical challenges or complexities associated with these types of missions.
- More recent flights appear to focus on VLEO and ISS, with an increasing number of successes as flights progress.



- Orbits like SSO, HEO, and GEO are more reliable, with high success rates even for larger payloads.
- Orbits like GTO present significant challenges, as reflected in a higher failure rate.



- There were no successful launches before 2014. Starting in 2014, the success rate increased significantly.
- Since 2018, the success rate has remained consistently high, with minor fluctuations year over year.

SQL RESULTS

Display the names of the unique launch sites in the space mission

```
select distinct Launch_Site from SPACEXTABLE;
```

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

SQL RESULTS

Display 5 records where launch sites begin with the string 'CCA'

```
select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;
```

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

SQL RESULTS

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

```
select sum(PAYLOAD_MASS__KG_) as total_payload_mass
from SPACEXTABLE where Customer = 'NASA (CRS)';
```

total_payload_mass
45596

SQL RESULTS

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

```
select avg(PAYLOAD_MASS__KG_) as average_payload_mass
from SPACEXTABLE where Booster_Version like 'F9 v1.1%';
```

average_payload_mass
2534.6666666666665

SQL RESULTS

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

```
select min(Date) from SPACEXTABLE
where Landing_Outcome like '%ground pad%';
```

min(Date)
2015-12-22

SQL RESULTS

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

```
select distinct Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTABLE where Landing_Outcome like '%success%drone%'
and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

List the total number of successful and failure mission outcomes

```
select Mission_Outcome, count(*) as total
from SPACEXTABLE group by Mission_Outcome;
```

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

SQL RESULTS

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
select distinct Booster_Version
from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (
select
       max(PAYLOAD_MASS__KG_)
from SPACEXTABLE
);
```

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

SQL RESULTS

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.

```
select substr(Date, 6,2) as month, Booster_Version, Launch_Site
from SPACEXTABLE where substr(Date,0,5)='2015'
and Landing_Outcome = 'Failure (drone ship)';
```

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

SQL RESULTS

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

```
select Landing_Outcome, count(*)
from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20’
group by Landing_Outcome order by count(*) desc;
```

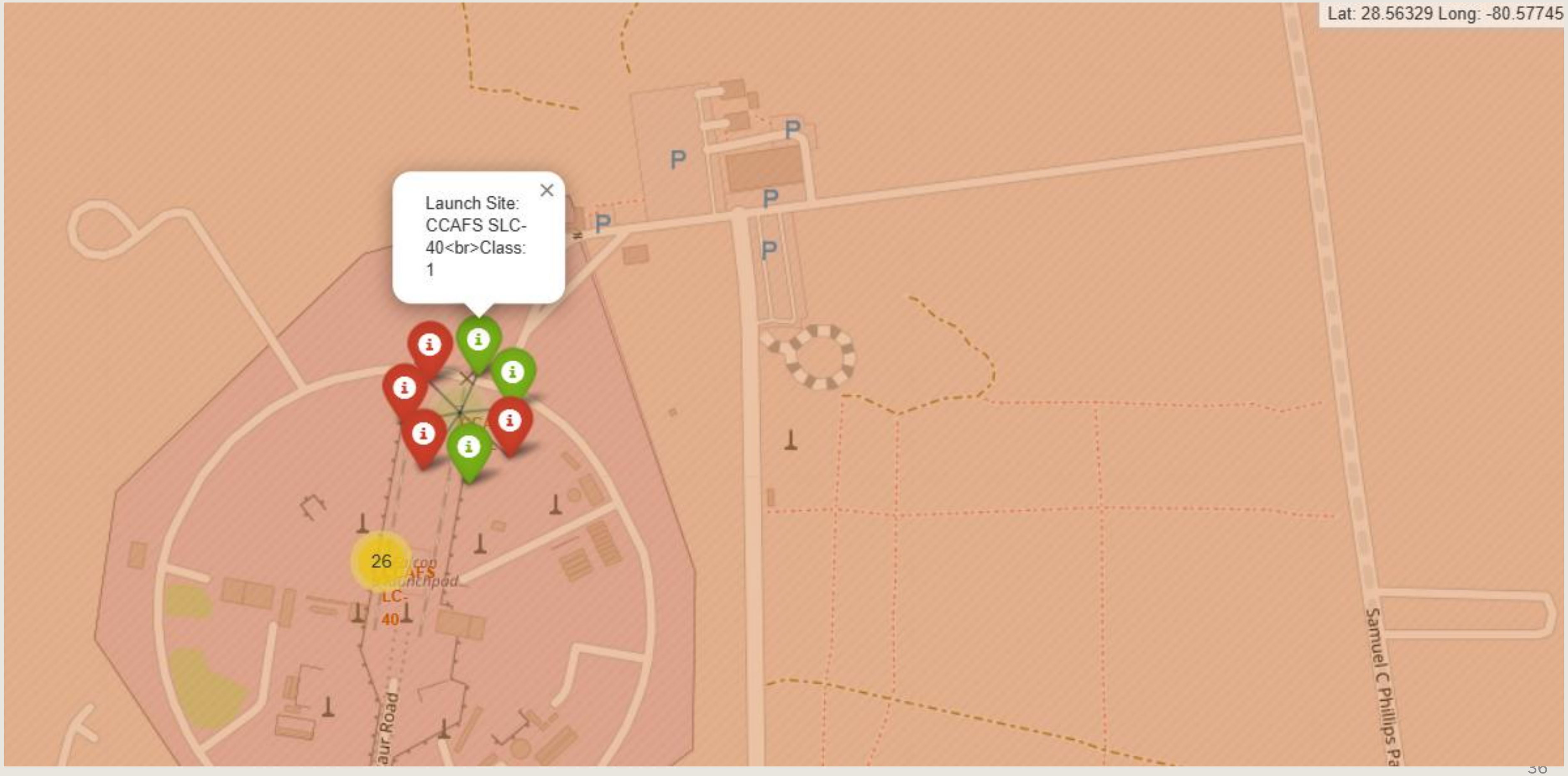
Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

INTERACTIVE MAPS WITH FOLIUM

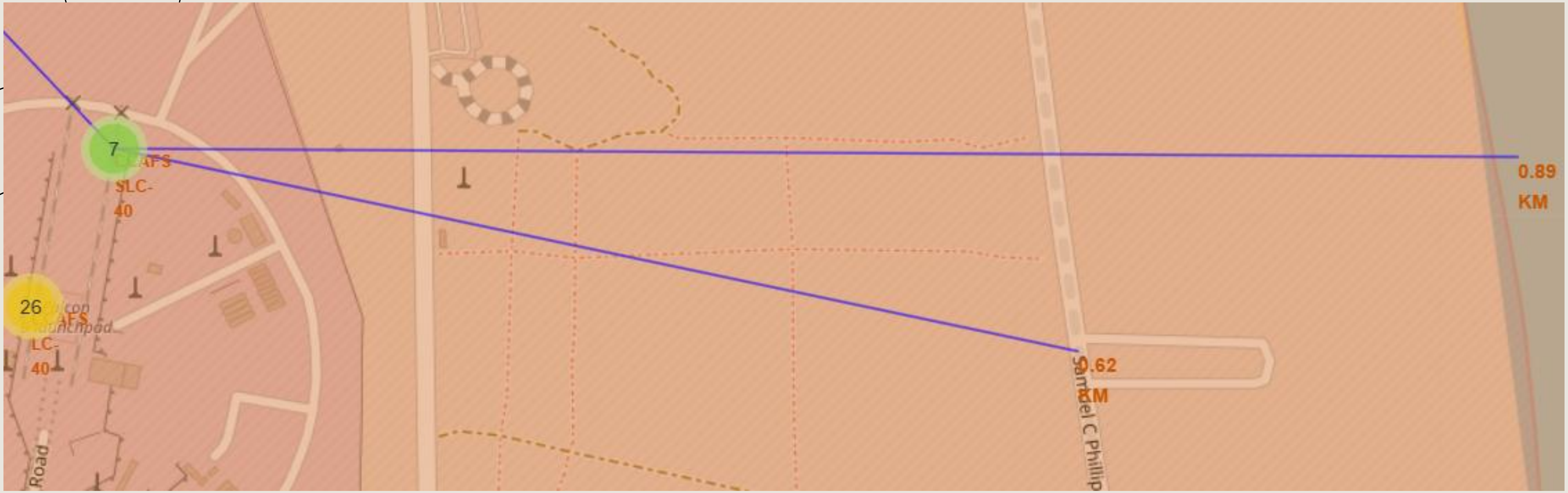
This map showcases the strategic locations utilized by SpaceX for its launches, emphasizing the importance of their positioning on Florida's east coast, which optimizes trajectories to various orbits.

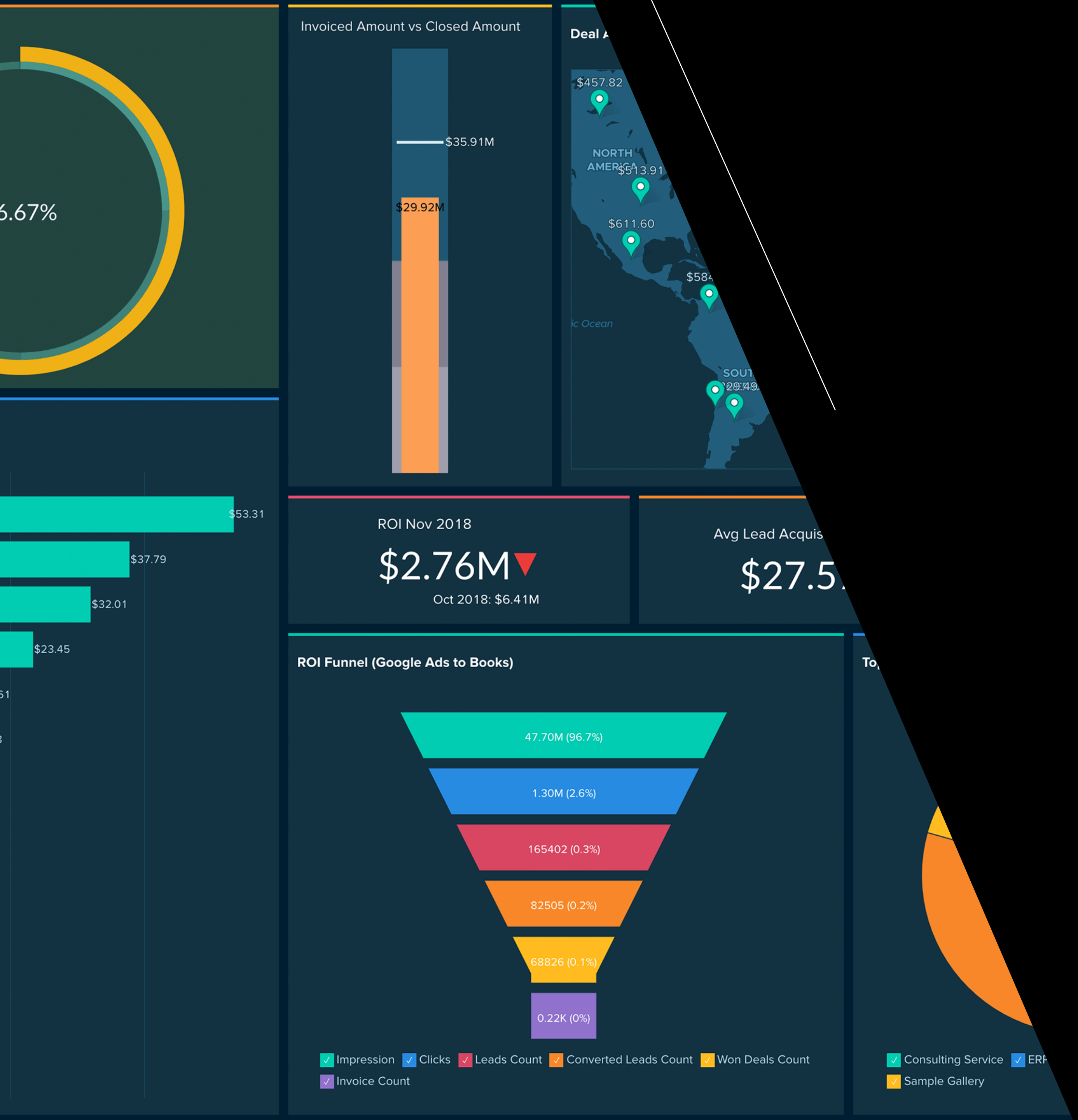
Now, map incorporates marker clusters to group launches performed from each site. The clusters emphasize the high activity at CCAFS SLC-40 compared to KSC LC-39A, revealing clear usage patterns between launch sites

In this updated map visualization, the launch clusters were expanded to display individual markers for each launch within the cluster. This level of detail facilitates identifying specific trends in launch outcomes at a particular site, supporting data-driven decision-making



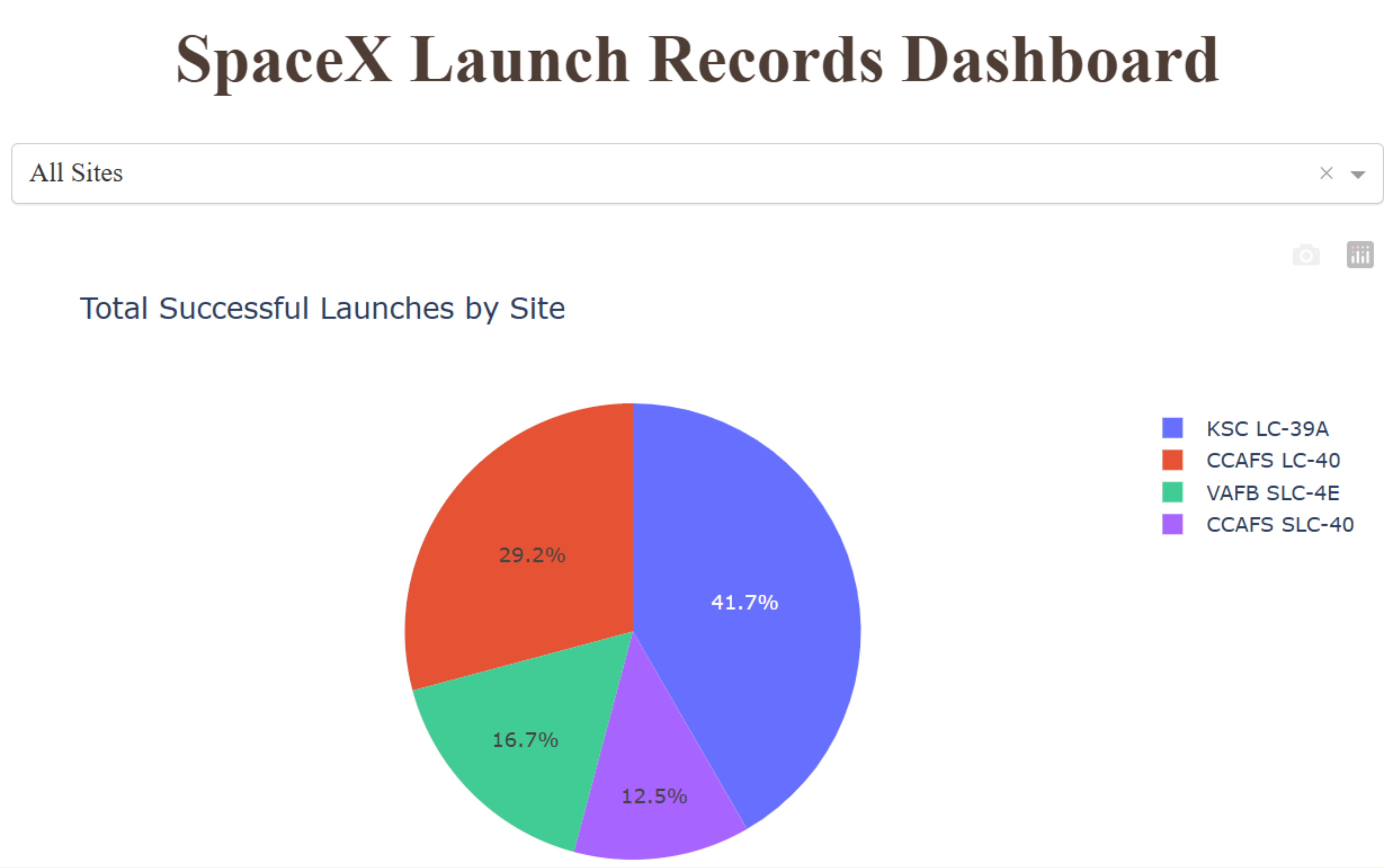
This version of the map includes lines and distance labels connecting the launch site to nearby key locations.
In conclusion, the site's proximity to the coastline and terrestrial infrastructure facilitates the transportation of materials and components.





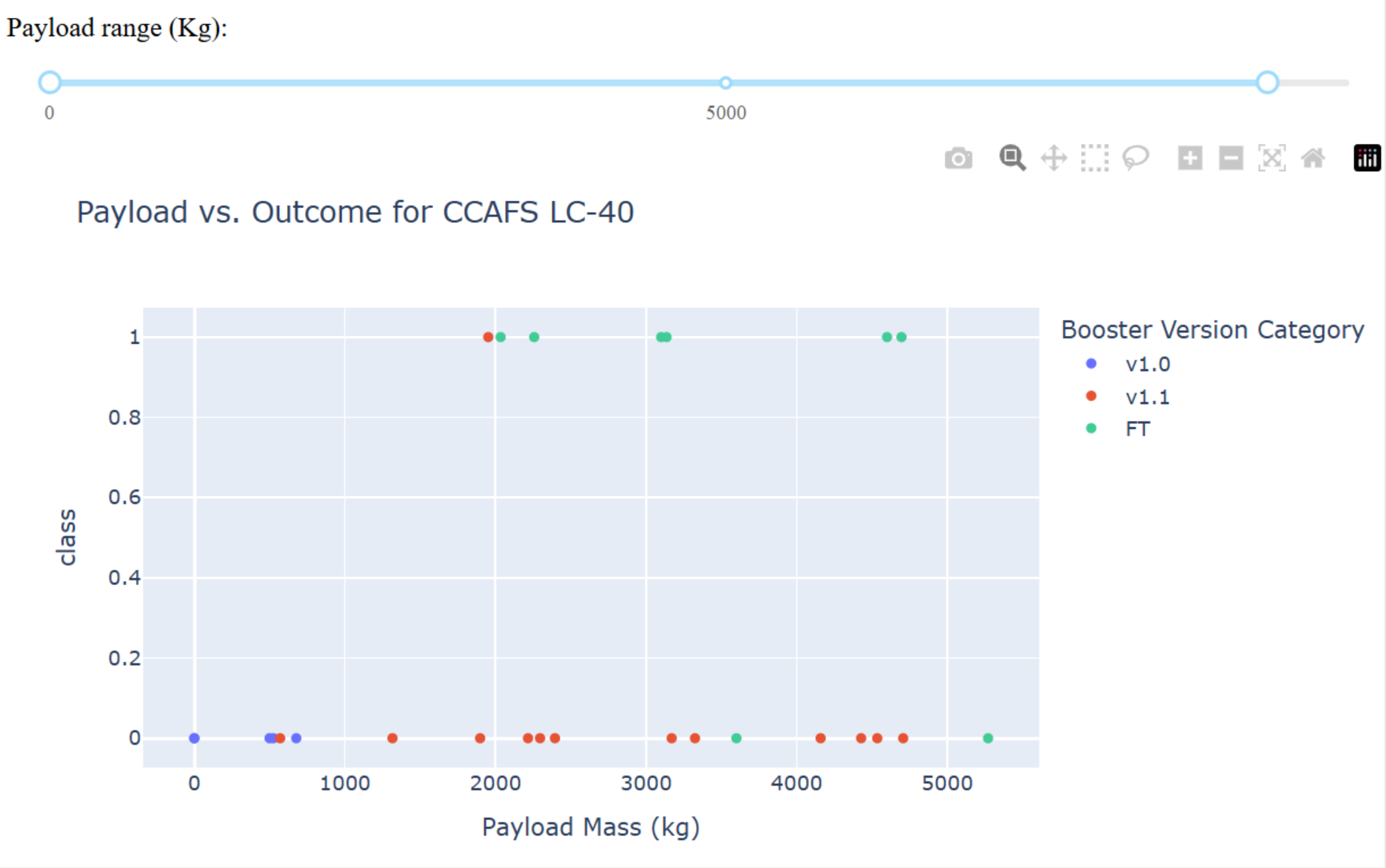
PLOTLY DASH DASHBOARD RESULTS

This pie chart shows the distribution of total successful launches by launch site for SpaceX. The percentages are divided among the various launch sites.



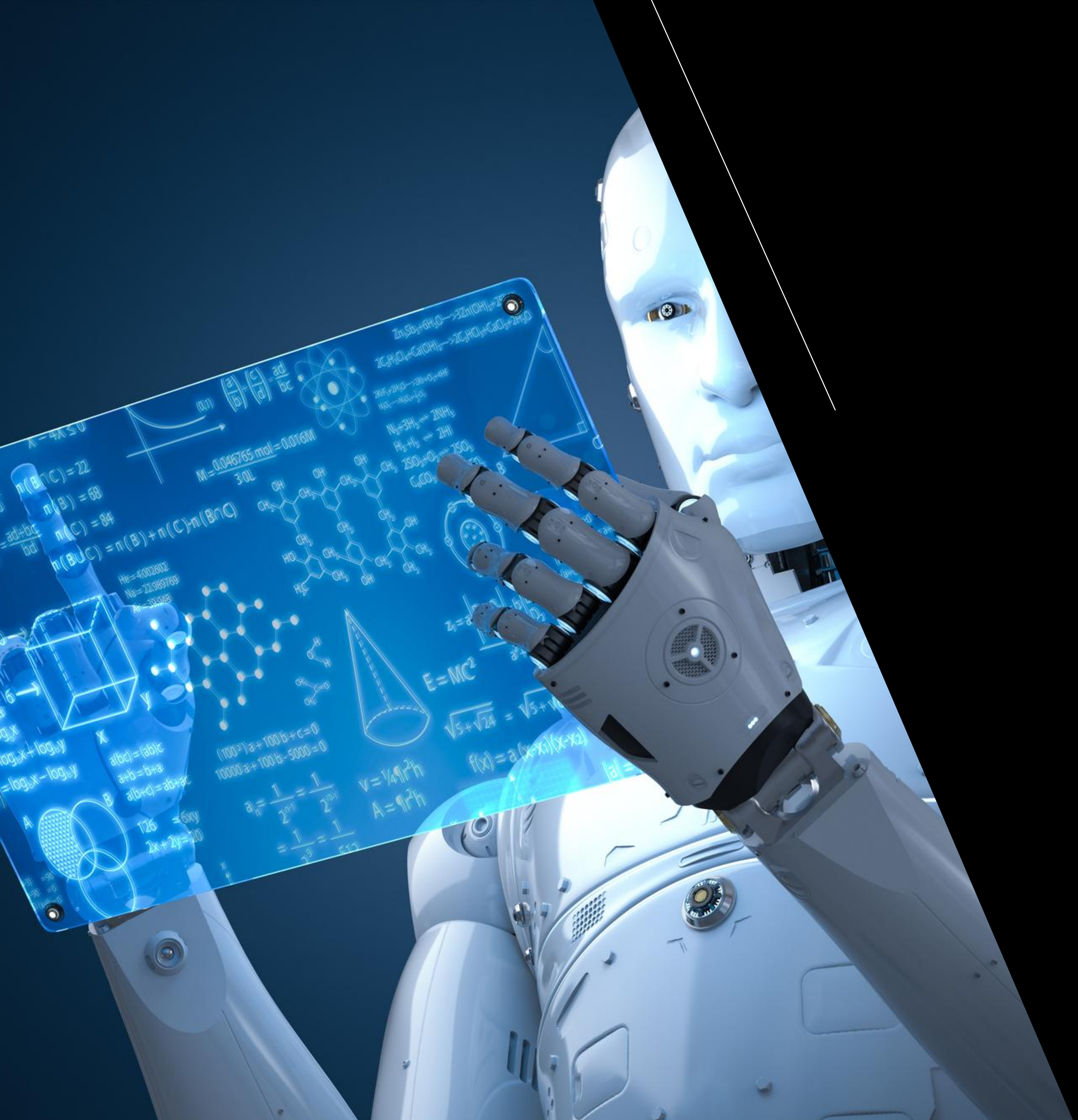
The majority of successful launches are concentrated at KSC LC-39A and CCAFS LC-40, reflecting their advanced infrastructure and high activity.

This scatter plot illustrates the relationship between Payload Mass and launch outcome (class, where 1 is successful and 0 is failed) for the CCAFS LC-40 site, categorized by the Booster Version Category.



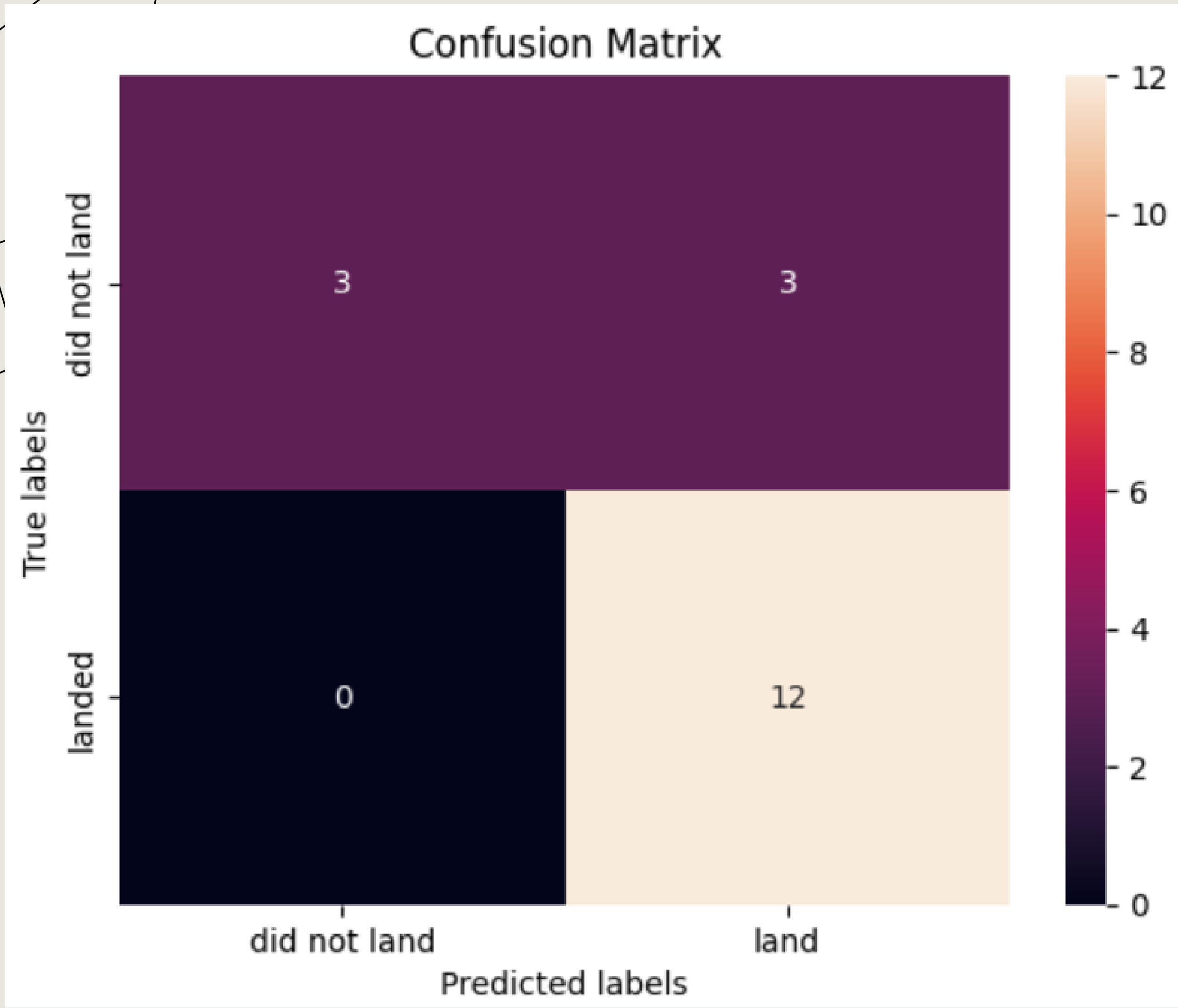
The FT booster version demonstrates superior performance in launch success, particularly for larger payloads.

The evolution of booster versions suggests significant technological progress that improves success rates



PREDICTIVE ANALYSIS RESULTS

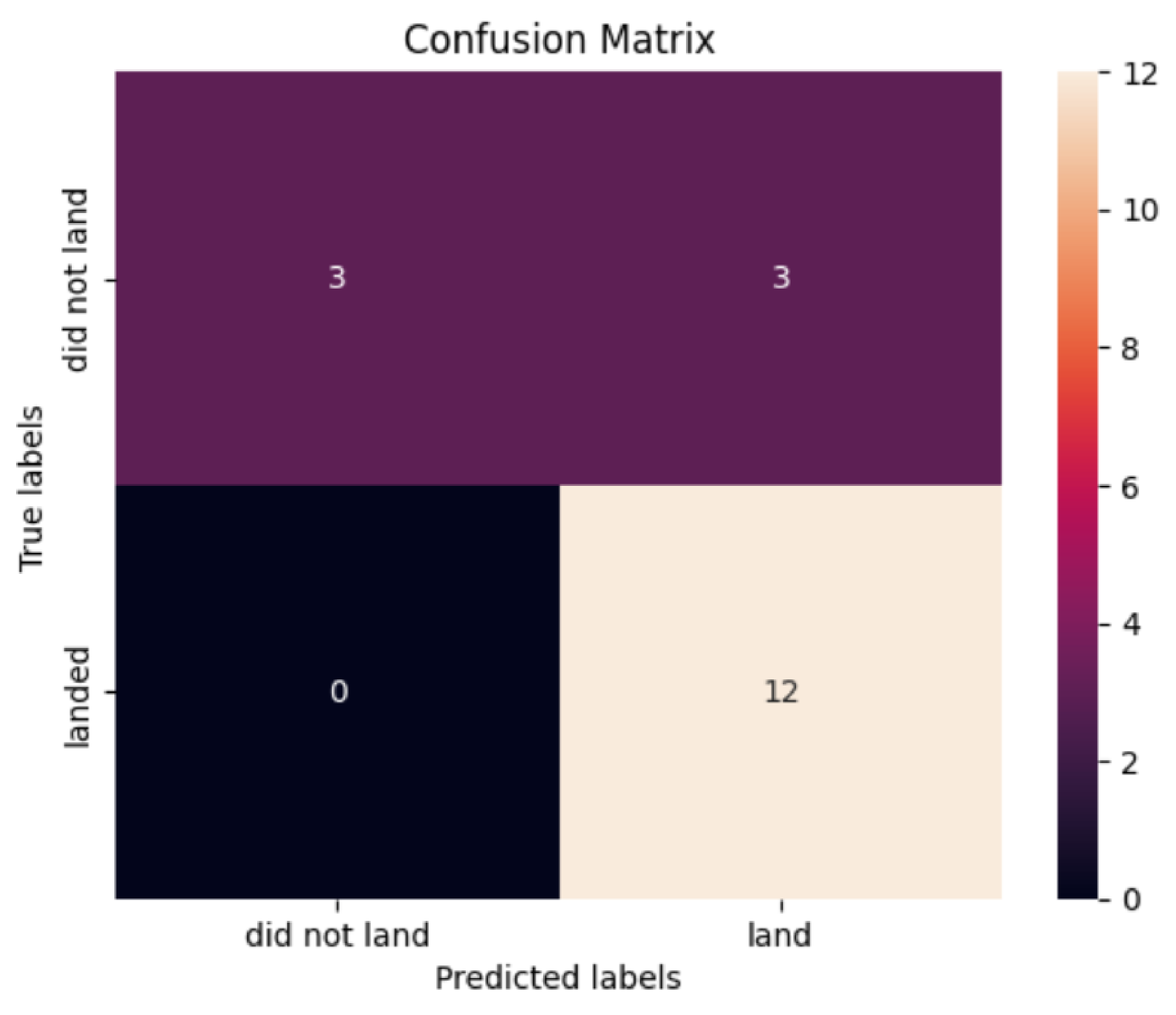
LOGISTIC REGRESSION



This confusion matrix illustrates the performance of a Logistic Regression model in predicting whether a landing was successful or not.

The model performs well overall, achieving an accuracy of 83.3%

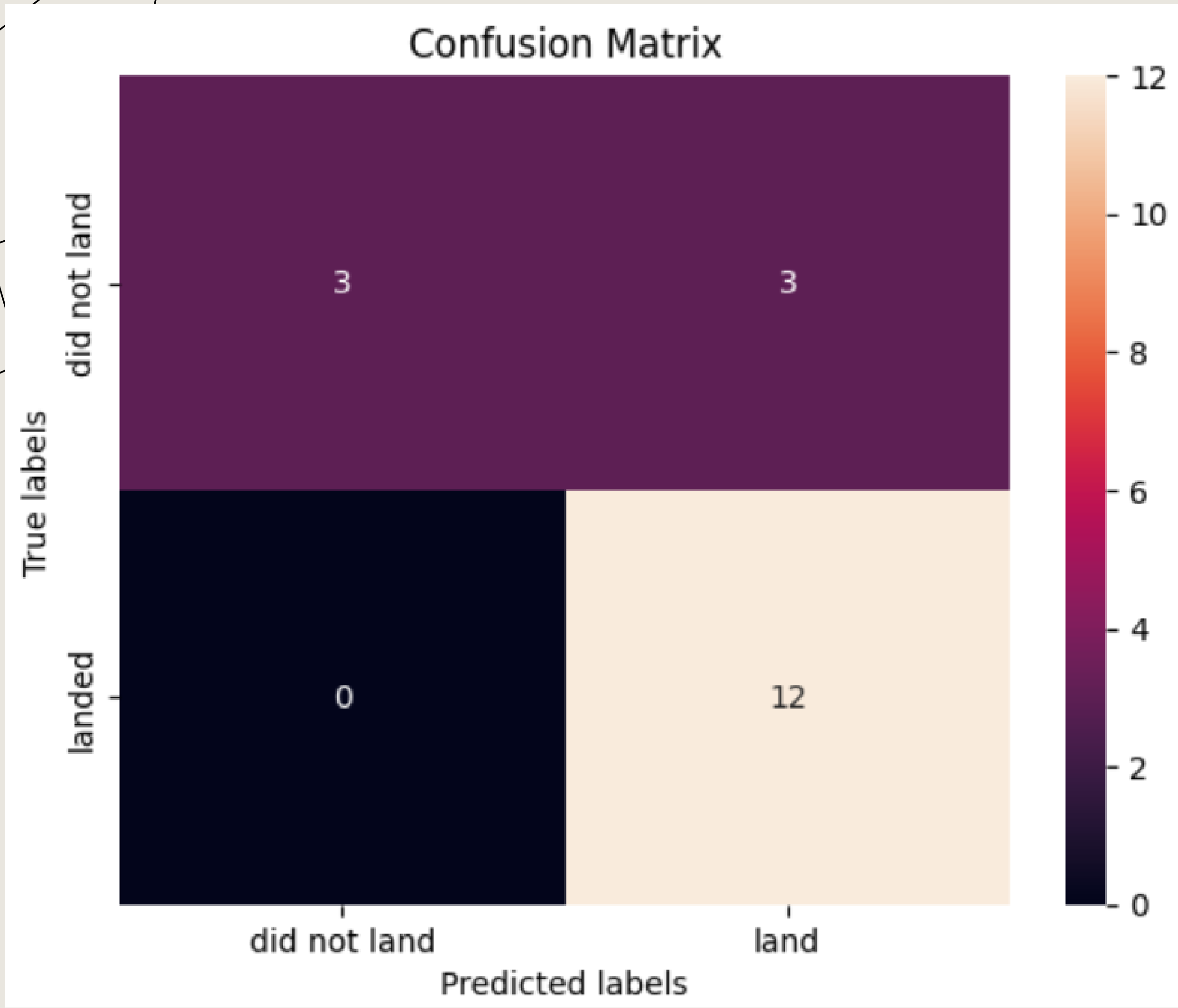
SVM



This confusion matrix shows the performance of a Support Vector Machine (SVM) model in predicting whether a landing was successful or not.

The SVM model performs well overall but lacks balance in predicting both classes.

DECISION TREE



This confusion matrix illustrates the performance of a decision tree model in predicting whether a landing was successful or not.

All the models present the same test accuracy, but the difference has on the best score accuracy

COMPARISON

	Validation Accuracy	Test Accuracy
Logistic Regression	0.8464	0.8333
SVM	0.8482	0.8333
Decision Tree	0.8750	0.8333
KNN	0.8482	0.8333

While Decision Tree shows better performance during validation, it does not improve test accuracy compared to the other models.

Logistic Regression, SVM, and KNN exhibit consistent performance between validation and test sets, making them more stable models for generalization.

The choice of model should consider factors such as interpretability, training time, and specific problem requirements, as all models have similar performance in terms of test accuracy

CONCLUSION

This project tackled the challenge of predicting whether the first stage of the Falcon 9 rocket will successfully land, using a comprehensive data science approach that includes data collection, cleaning, exploratory analysis, visualization, and predictive modeling.

- Payloads above 10,000 kg have a high likelihood of success
- Orbits like **SSO** and **HEO** are more reliable, with success rates above 90%
- Logistic Regression is a strong choice due to its simplicity and interpretability.
- SpaceX has achieved remarkable optimization in its operations, reflected in the high performance of recent launches.

THANK YOU FOR YOUR TIME

Fabian Pulgar

f.pulgarlopez@uandresbello.edu