IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tomzine Mohamed
17 December 2024

Outline

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

Executive Summary

This project aims to predict the success of Falcon 9 first-stage landings, a critical factor in SpaceX's ability to reduce launch costs through reusability. The study uses historical launch data to uncover key factors influencing landing success, such as payload mass, orbit type, and booster reusability. Data collection was performed using JSON records, and data wrangling techniques were applied to clean and prepare the dataset for analysis.

Executive Summary

Exploratory Data Analysis (EDA) revealed trends, including higher landing success rates for lower orbit payloads and reusable boosters. SQL queries and interactive visualizations (using Folium and Plotly Dash) enabled dynamic exploration of launch site performance and spatial relationships.

Several machine learning models, including Logistic Regression, Decision Trees, K-Nearest Neighbors, and Support Vector Machines, were developed to predict landing outcomes. Logistic Regression emerged as the best-performing model with an accuracy of 94%. This model provides stakeholders with a reliable tool to estimate launch success probabilities, optimize future launches, and improve cost efficiency in the competitive space launch market.

Introduction

Space exploration has traditionally been an expensive endeavor, with significant costs associated with launching rockets. Space X, a leading aerospace manufacturer and space transportation company, has disrupted the industry by developing reusable rockets, particularly the Falcon 9. The Falcon 9's reusable first stage drastically reduces the cost of launches, enabling more frequent and affordable missions. Space X advertises Falcon 9 rocket launches at a cost of $62 million, compared to competitors whose costs exceed $165 million per launch. The ability to successfully land and reuse the first stage is the key to achieving such cost efficiency.

This project seeks to address the following key questions:

- ***What factors influence the successful landing of a Falcon 9 first stage?***
- ***Can we predict the success of a Falcon 9 first-stage landing using machine learning models?***

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

- The data used for this project was collected from publicly available SpaceX launch records using SpaceX's REST API.
- In addition to the API data, supplementary information was extracted using web scraping techniques to ensure a comprehensive dataset.

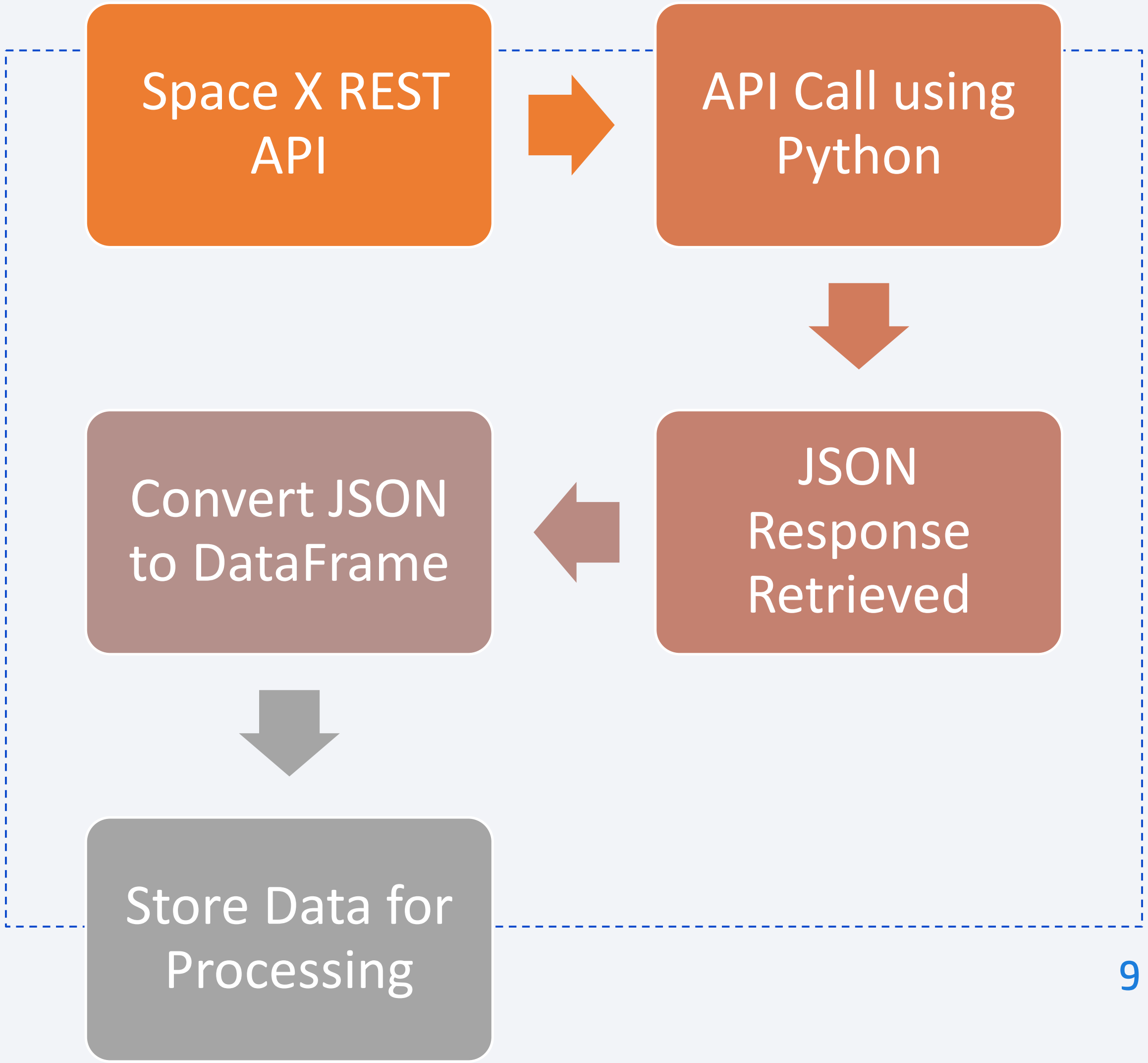
Data Collection – SpaceX API

The data for this project was collected using SpaceX's REST API calls, which provide access to detailed historical launch records.

Each API response returned structured launch data, including the following key fields:

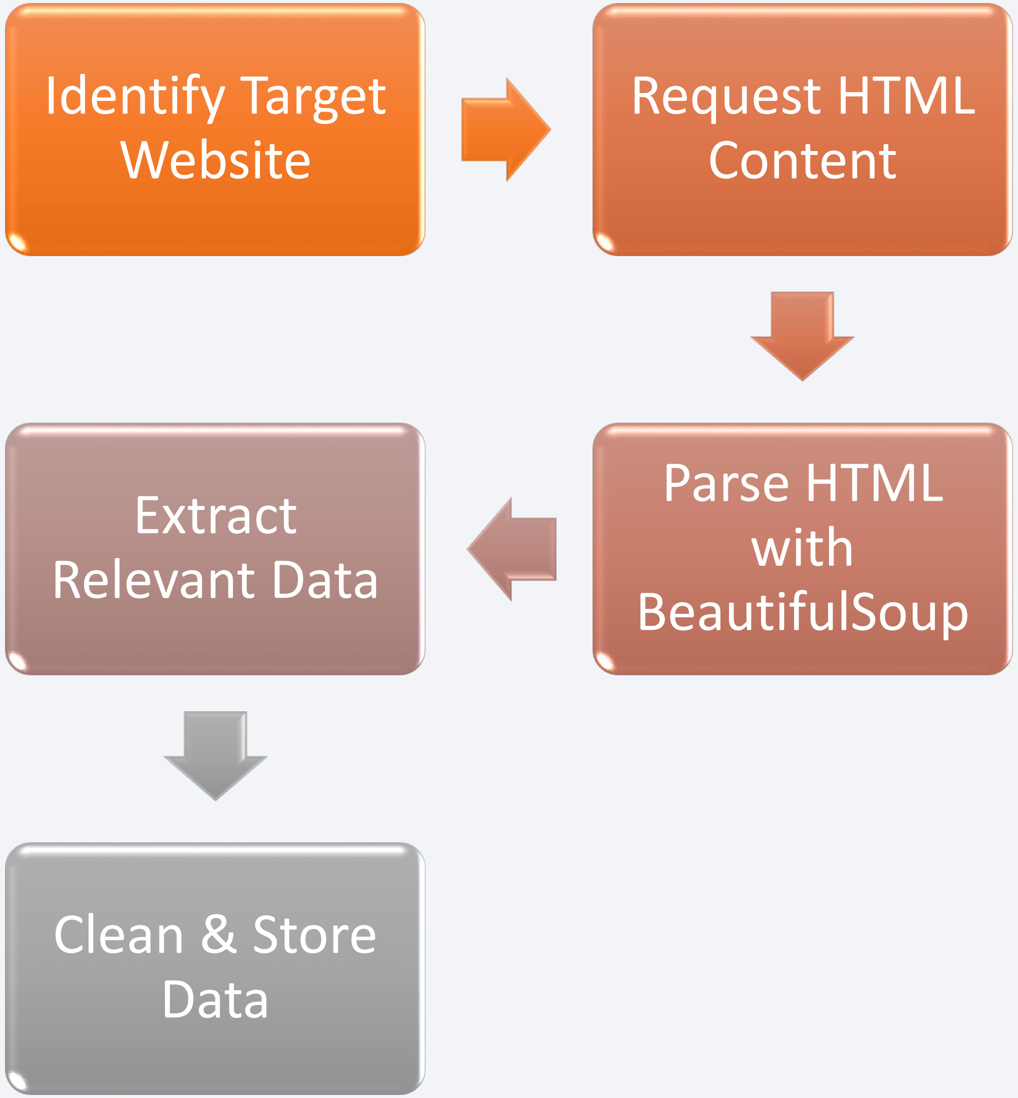
- **Flight Number**
- **Launch Site**
- **Payload Mass (kg)**
- **Orbit Type**
- **Booster Version**
- **Landing Outcome**

[link to lab]



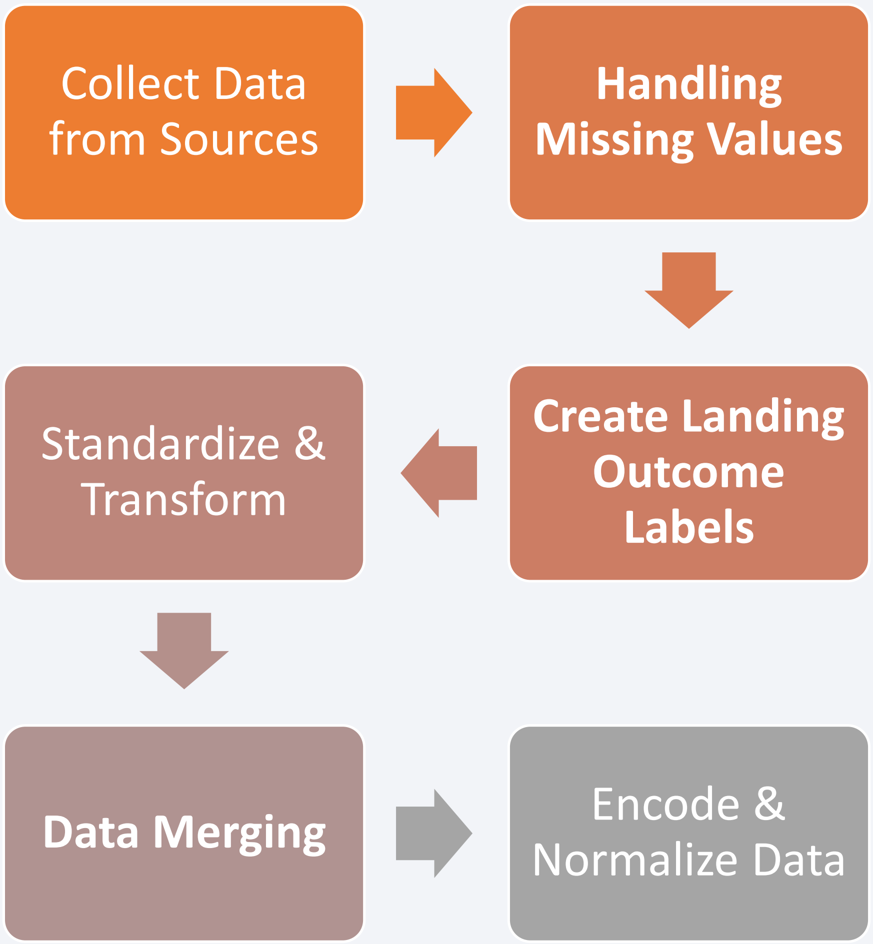
Data Collection - Scraping

- In this lab, you will be performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches [(Wikipedia).]()
- Required information such as ***mission names, payload masses, booster landing outcomes, booster versions, and date and time*** were extracted.
- [lab]()



Data Wrangling

- Data wrangling was performed to clean and prepare the data for analysis. Detailed steps and examples are documented in my Jupyter Notebook available at [Data Wrangling Notebook].
- The process is visualized in the following flowchart:



EDA with Data Visualization

Visualization Techniques:

- **Scatter Plots:** Examined relationships between payload mass vs orbit type and orbit type vs flight number and payload mass vs launch site
- **Bar Charts** : Illustrated launch success rates by orbit types.
- **Line chart :** Illustrated launch success rates by years of lunch.
- EDA Lab

EDA with SQL

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

%sql select distinct Launch_Site from SPACEXTABLE

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

%sql select Payload from SPACEXTABLE where Customer = 'NASA (CRS)'

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

%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'

- List the total number of successful and failure mission outcome :

%sql select count(Mission_Outcome), Mission_Outcome from SPACEXTABLE group by Mission_Outcome

- For more visit [SQL Notebook]

Build an Interactive Map with Folium

Interactive visualizations were created using Folium Library:

- Generated interactive maps to visualize launch site locations and their proximities to landing zones.
- Cluster markers were used to highlight key spatial relationships.
- PolyLine were used to Draw line between launch site and the location
- For more visit **Interactive Visual Analytics with Folium Lab**

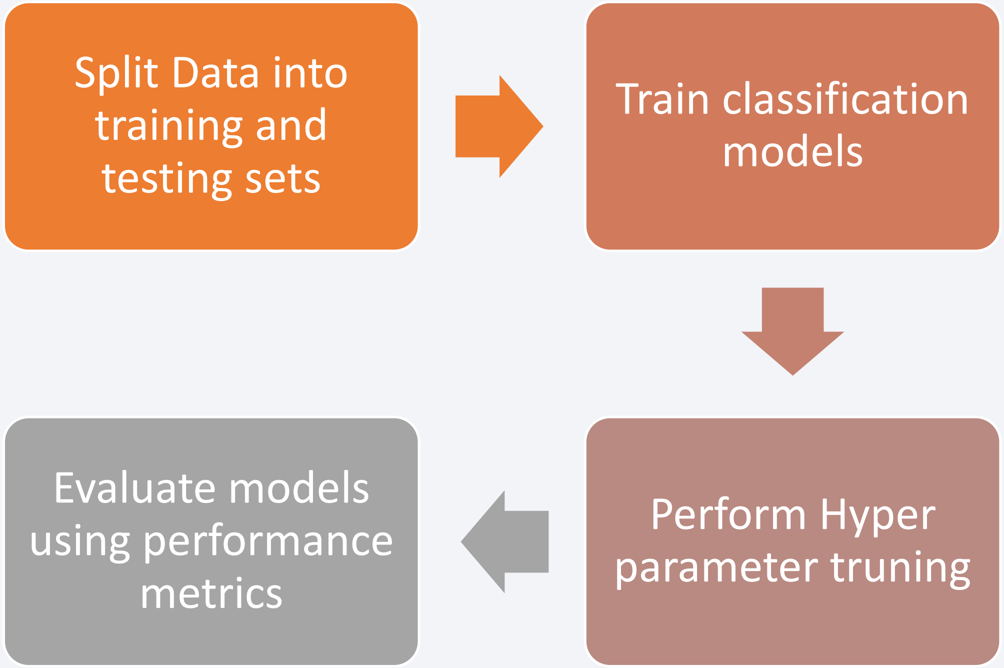
Build a Dashboard with Plotly Dash

- Built interactive dashboards with dynamic pie chart and scatter chart.
- Enabled stakeholders to explore launch success rates across multiple dimensions.
- For more visit **Dashboard with Ploty Dash**

Predictive Analysis (Classification)

To predict landing success, the following classification models were developed:

Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM)



1. **Model Building:**

- Models were trained on the pre-processed dataset using the training subset.
- Target variable: Landing Outcome (Success/Failure).

1. **Hyperparameter Tuning:**

- Grid Search was used to optimize hyper parameters for each model to achieve the best performance.

2. **Model Evaluation Metrics:**

- Accuracy: Overall correctness of predictions.
- confusion matrix

For more visit [Machine Learning Prediction lab]

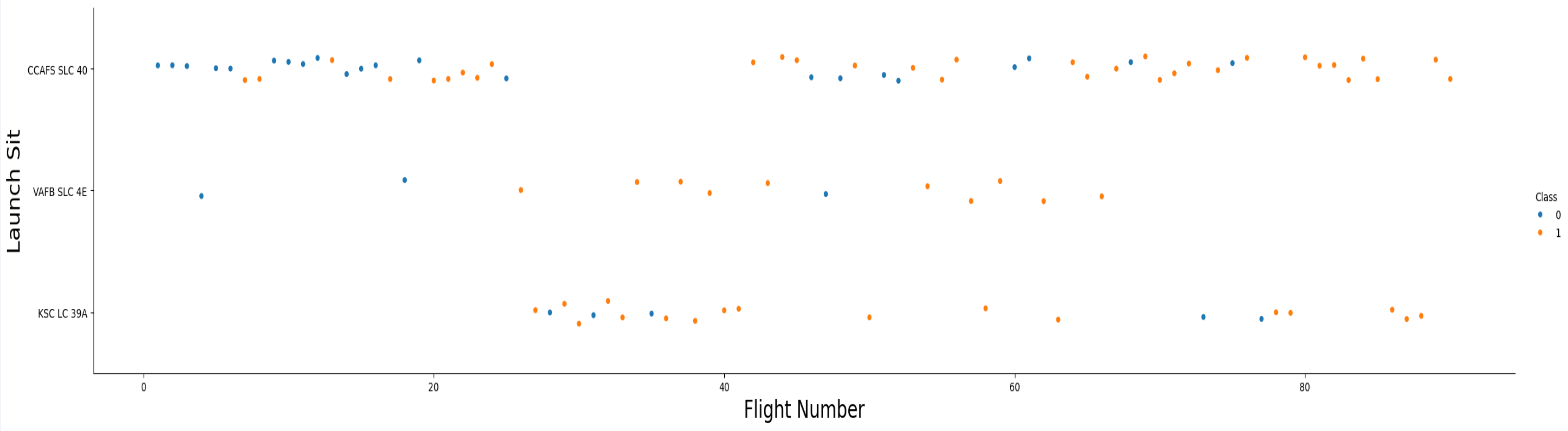
Results

- Launch Site CCAFS SLC-40 conducted the highest number of launches, followed by KSC LC-39A.
- Success rates for LEO-targeted launches were notably higher compared to other orbits.
- Success rates for missions varied by orbit type, indicating a strong dependency on mission goals and parameters.
- Logistic Regression achieved the highest accuracy of 94%, closely followed by Decision Trees and SVM. This superior performance is attributed to Logistic Regression's ability to model linear relationships between input variables, such as payload mass and orbit type, and the binary target outcome.

Section 2

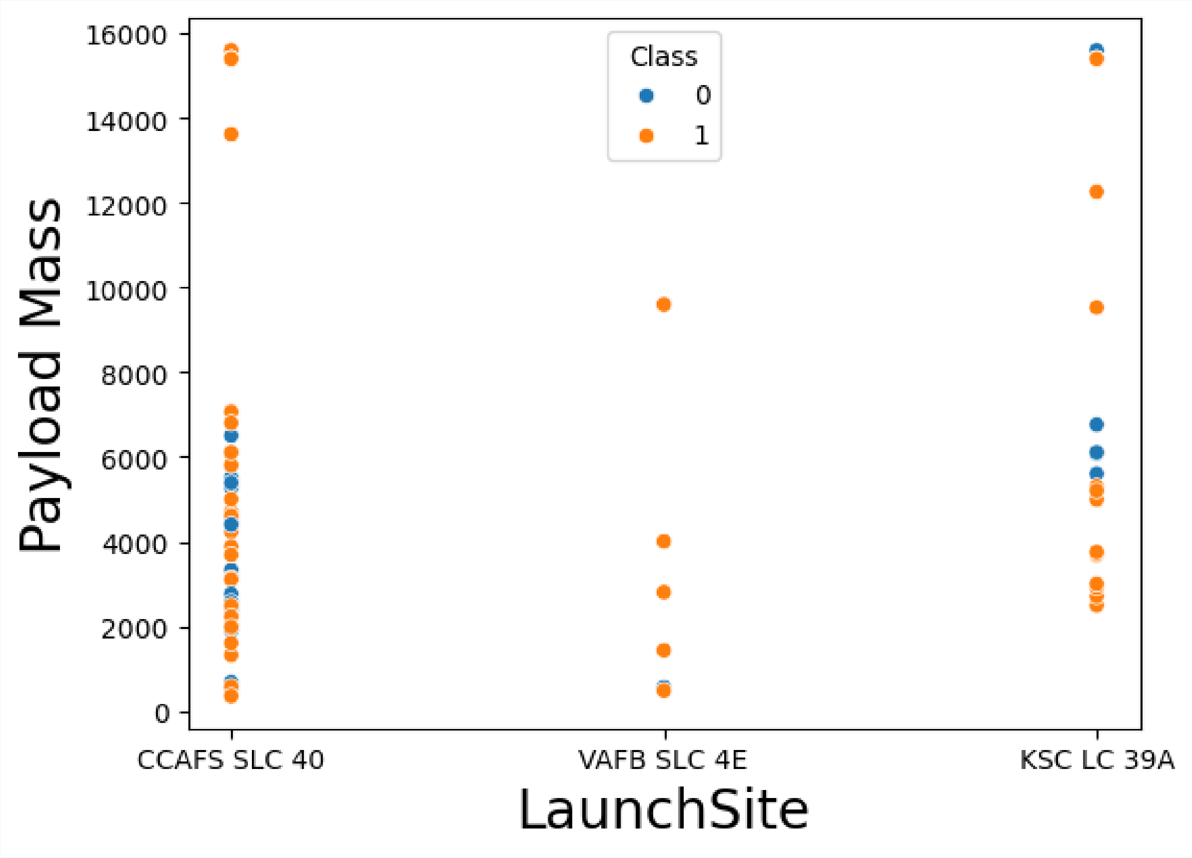
Insights drawn from EDA

Flight Number vs. Launch Site



- **CCAFS SLC 40** has the highest number of launches.
- Over time, the number of successful missions (Class 1) increases while failures (Class 0) become less frequent, indicating improvement in launch success rates.

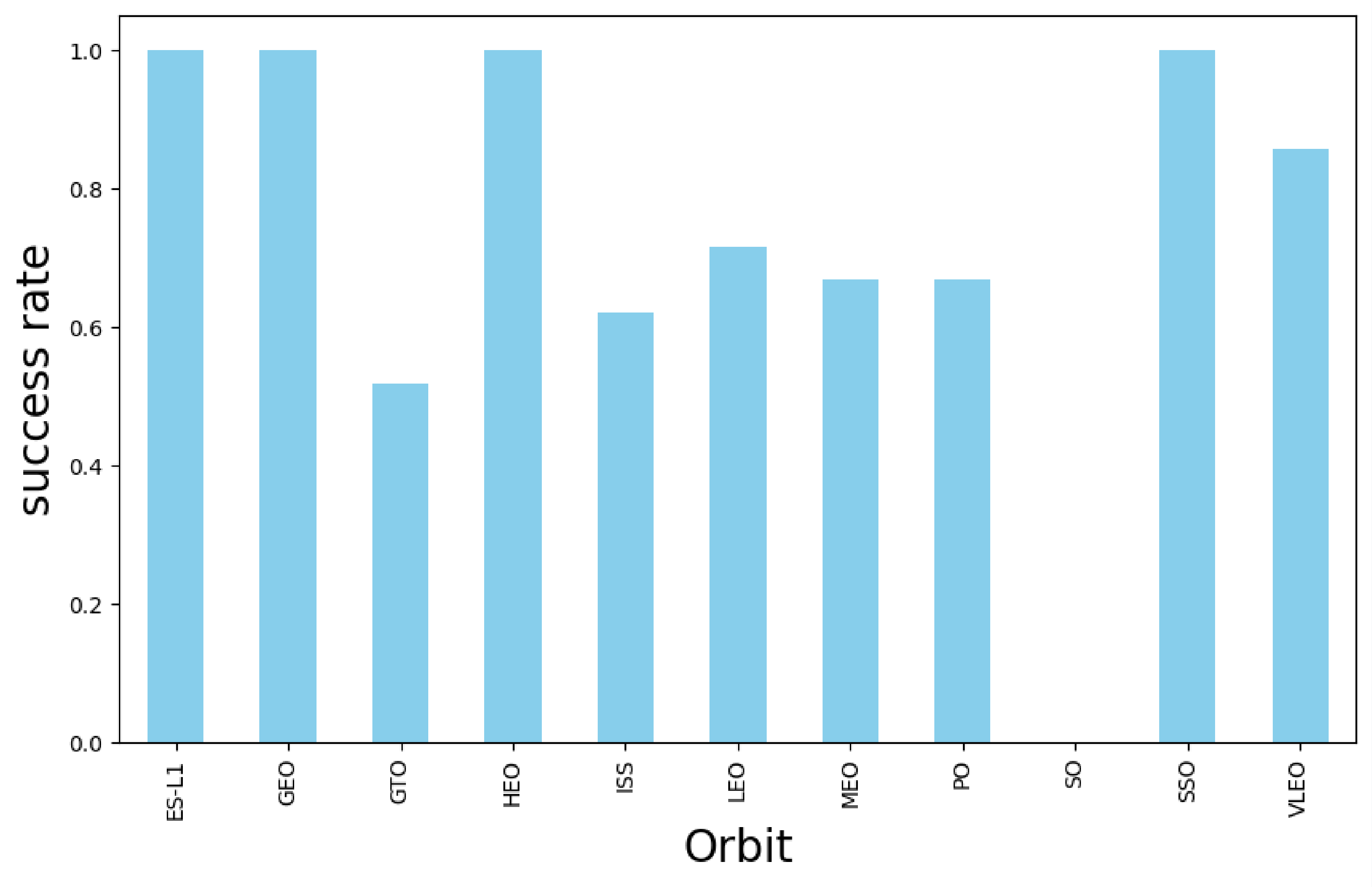
Payload vs. Launch Site



- Launch success seems to improve for higher payloads at CCAFS SLC 40 and KSC LC 39A, highlighting advancements or efficiencies at this launch site.

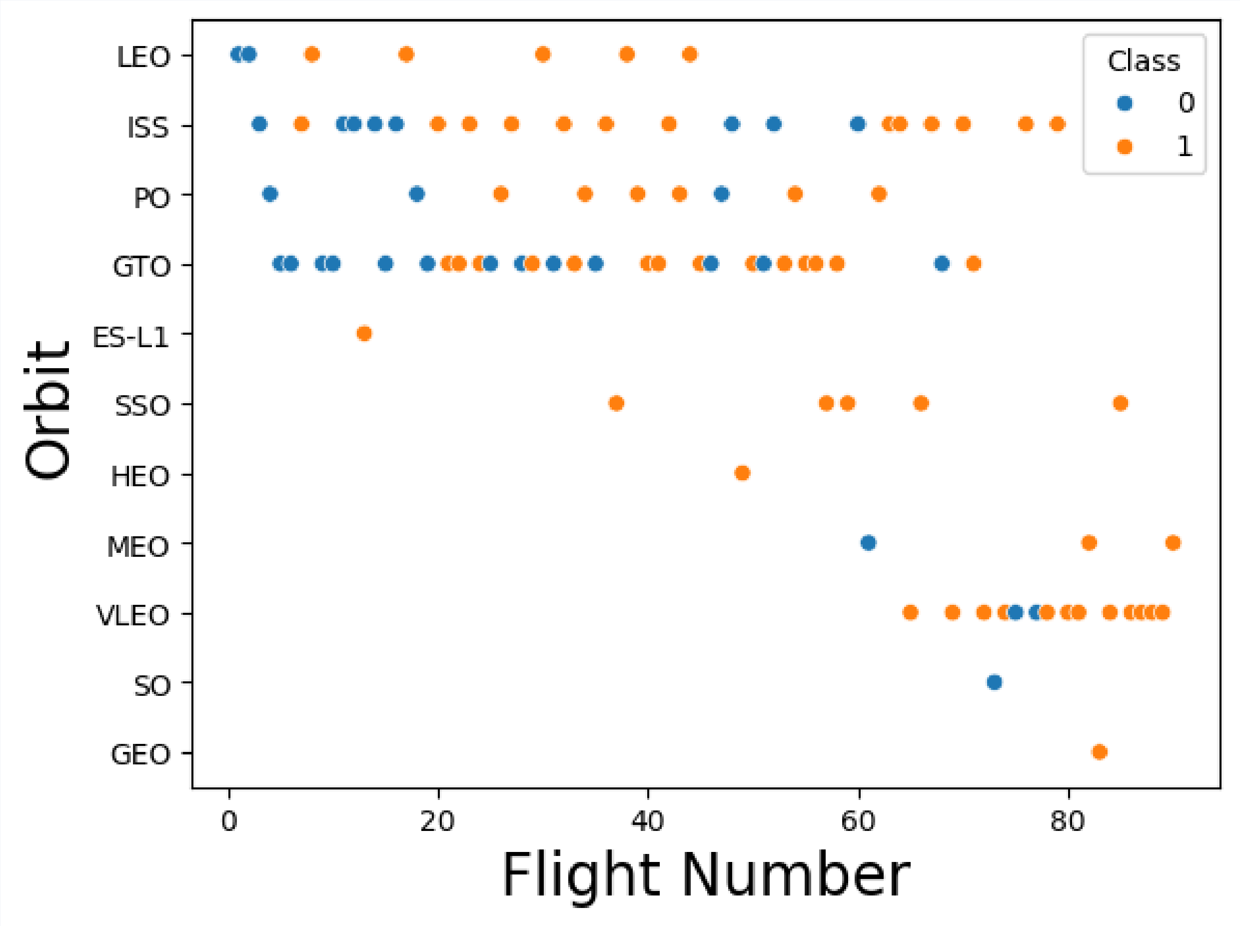
Success Rate vs. Orbit Type

- The orbits SSO , HEO, GEO and ES-L1 have the highest success rates.



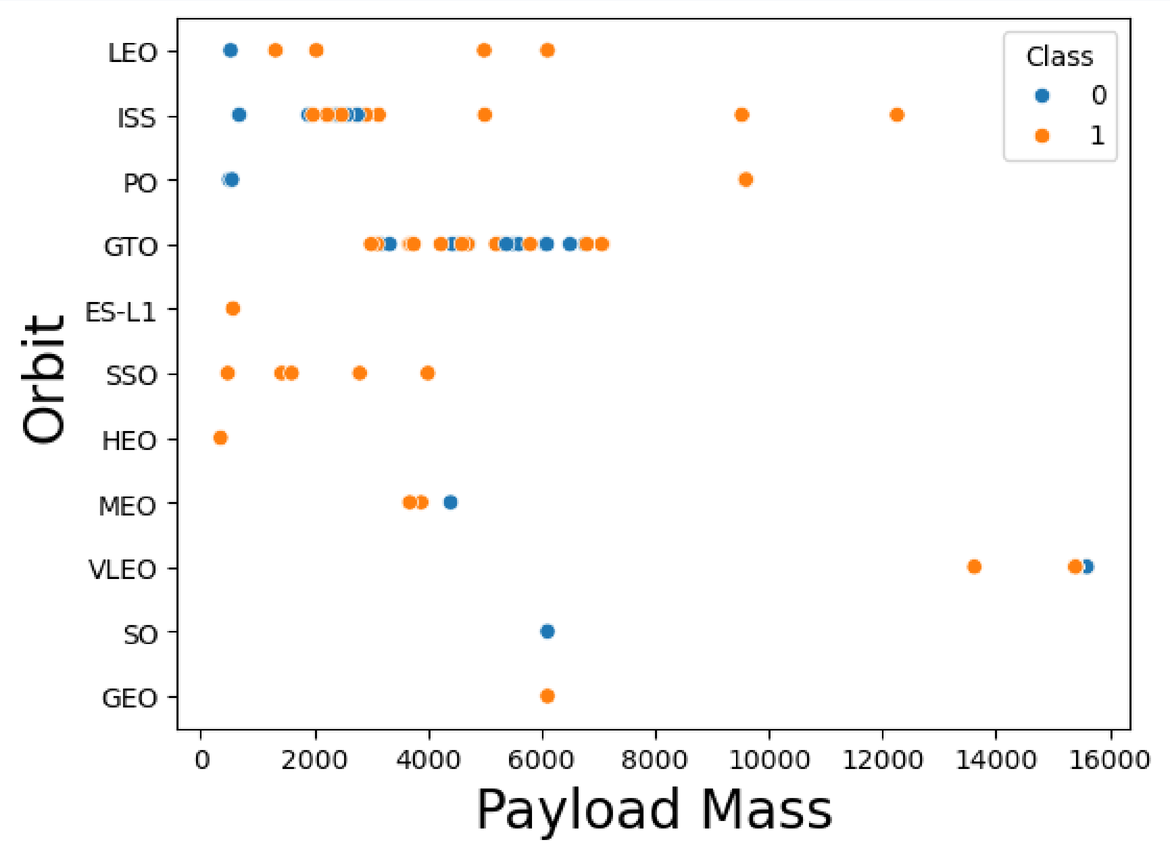
Flight Number vs. Orbit Type

- the LEO orbit, success improve with increasing in number of flights.



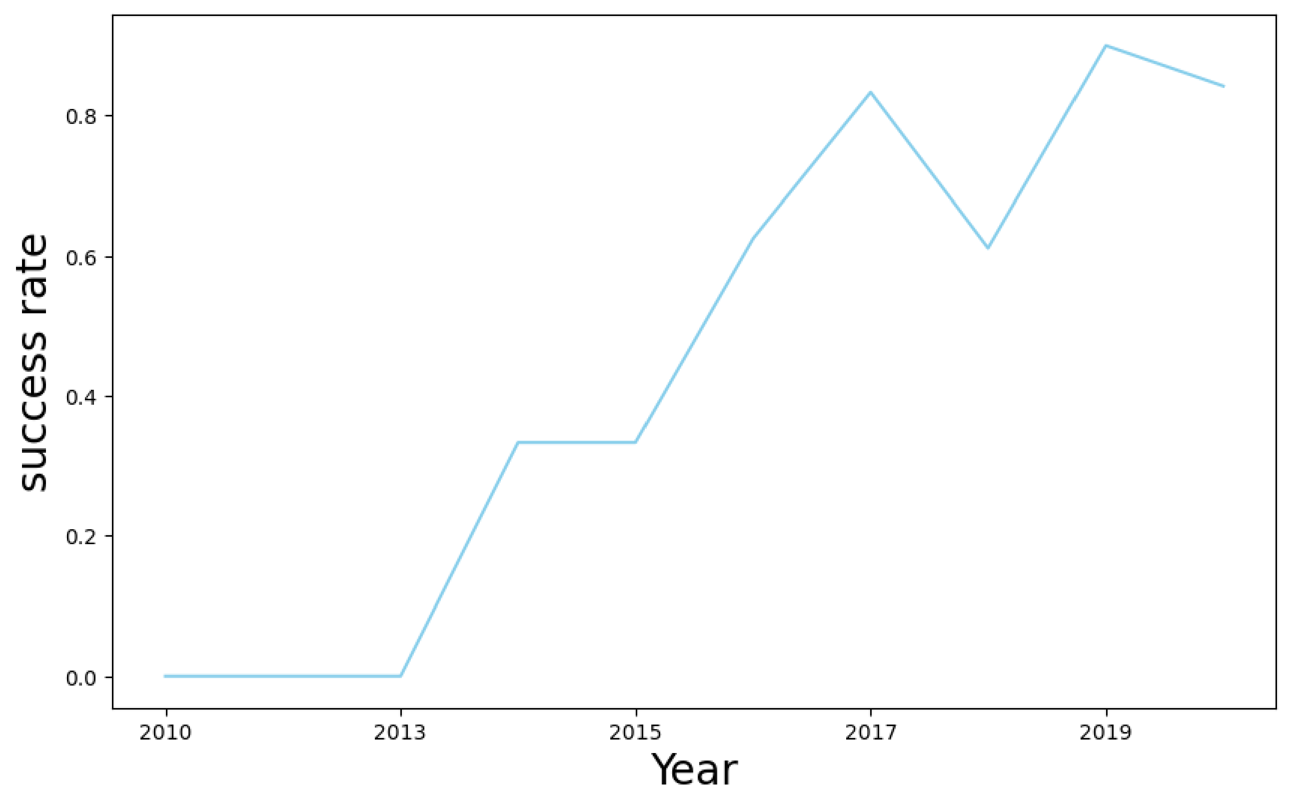
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.



Launch Success Yearly Trend

- the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- The names of the unique launch sites :

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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 (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

Total Payload Mass

- The total payload carried by boosters from NASA

total payload mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had 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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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

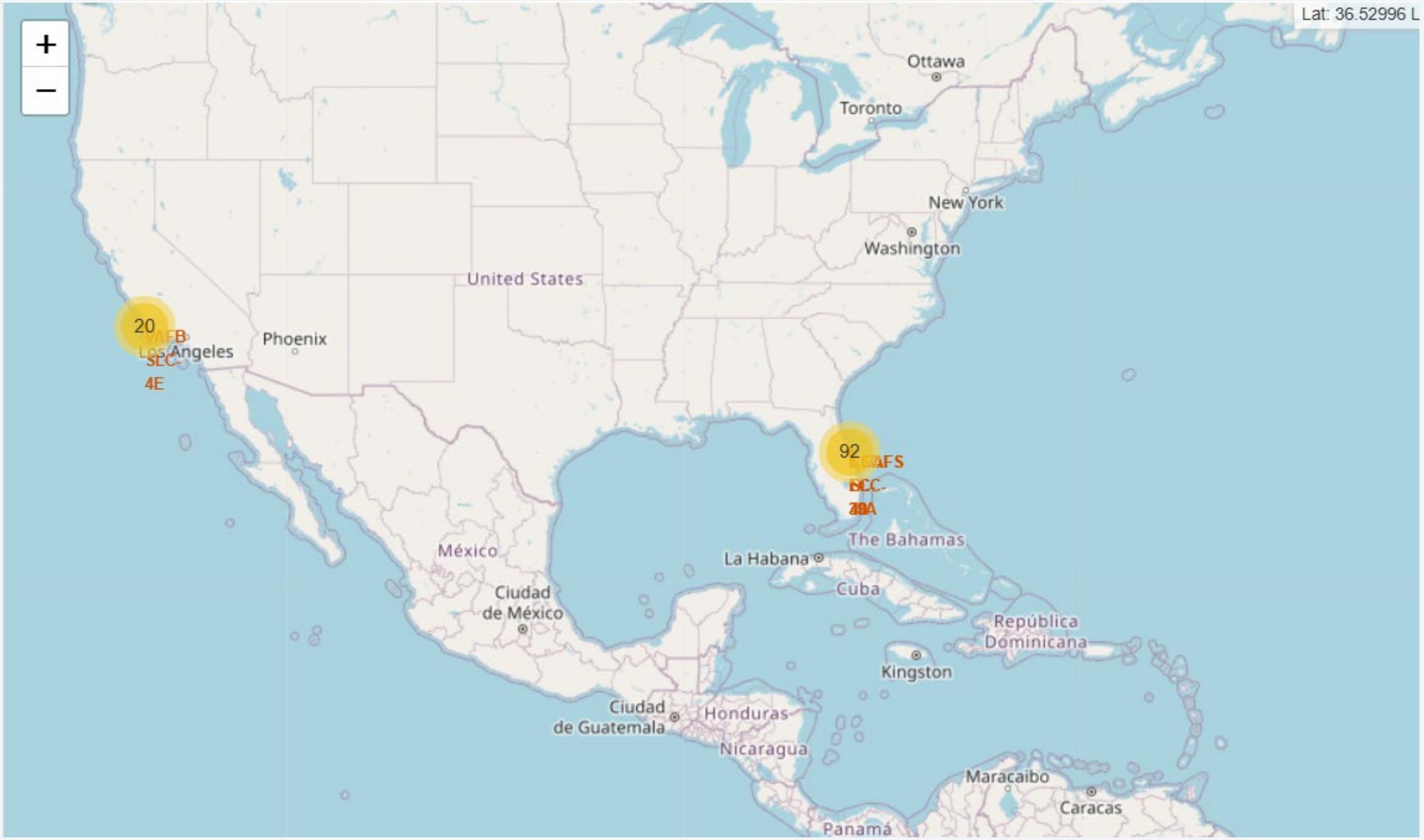
Landing_Outcome	Outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

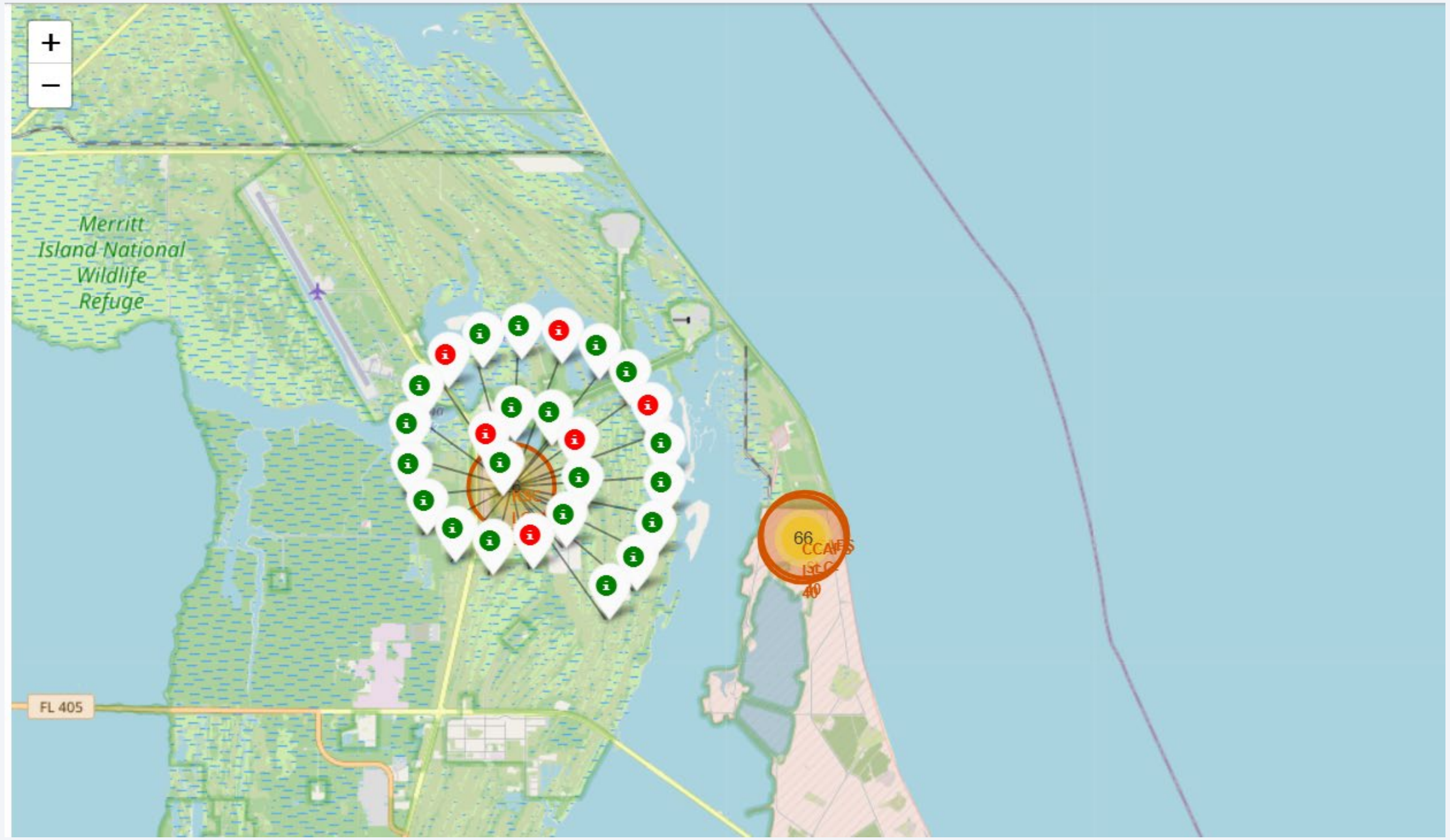
All launch sites location

The **East Coast** (Florida) hosts the majority of launches (92)

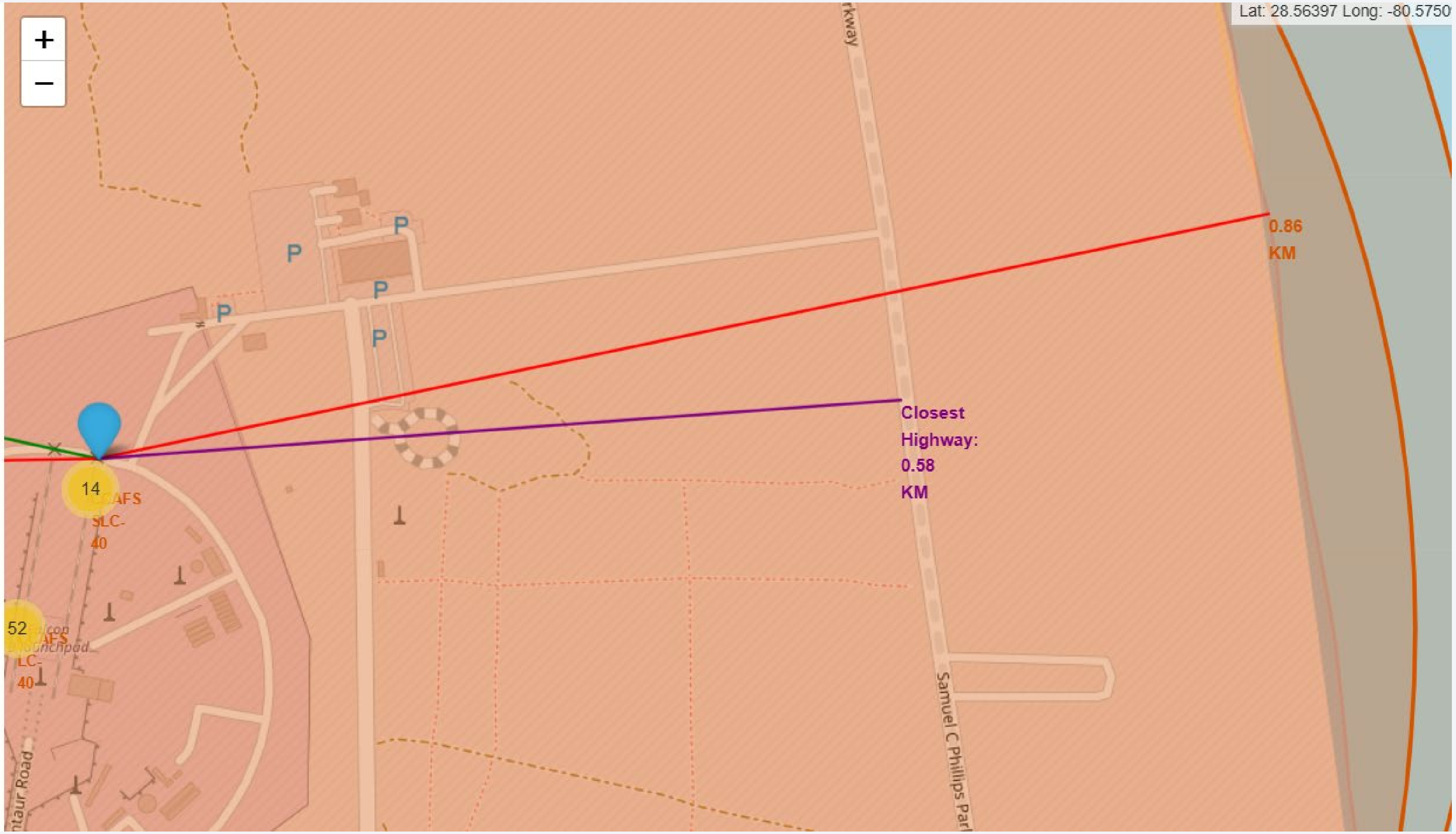


SpaceX Launch at KSC LC-39A

The majority of launches have been successful



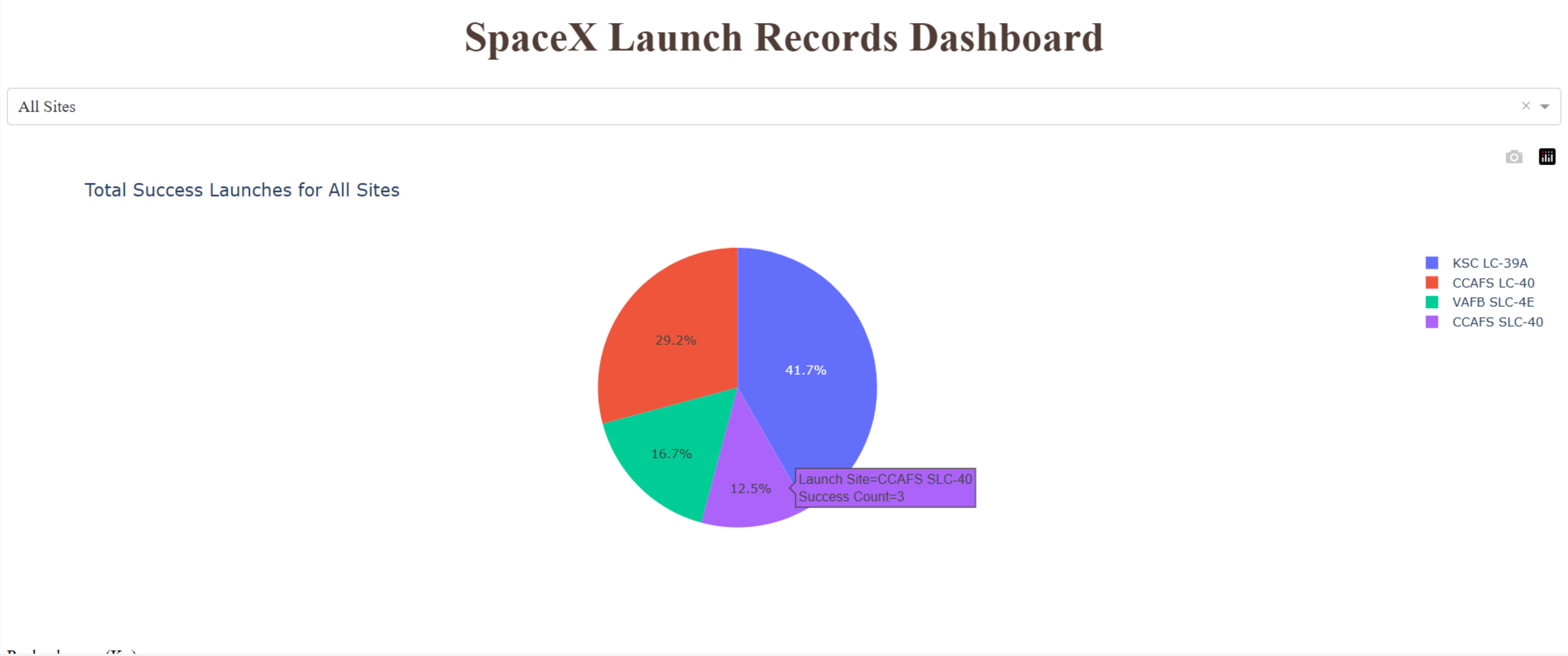
Analyzing Proximities of SpaceX Launch Sites with Folium



Section 4

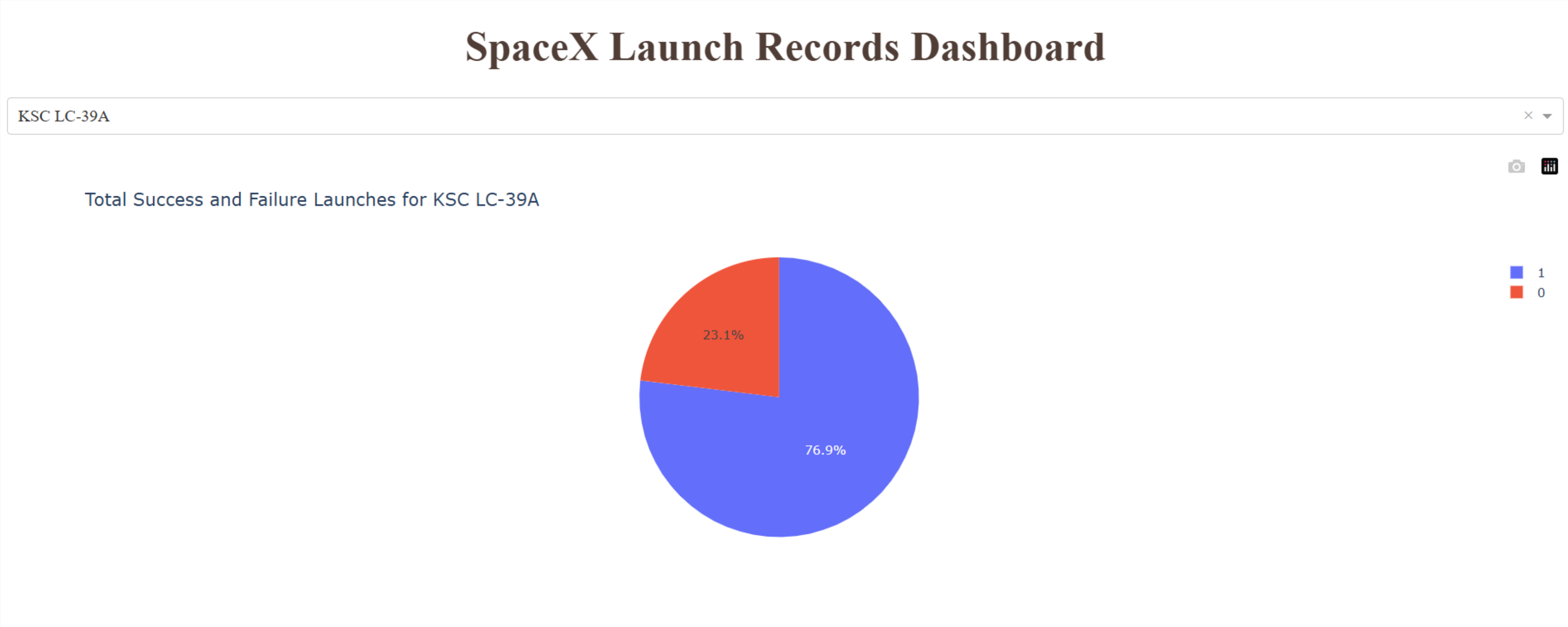
Build a Dashboard with Plotly Dash

Total success launches for all sites



Total success and failure for KSC LC-39A

- KSC LC-39A has highest launch success ratio



- KSC LC-39A has achieved 76 successful launches.

Payload vs. Launch Success Across All Sites

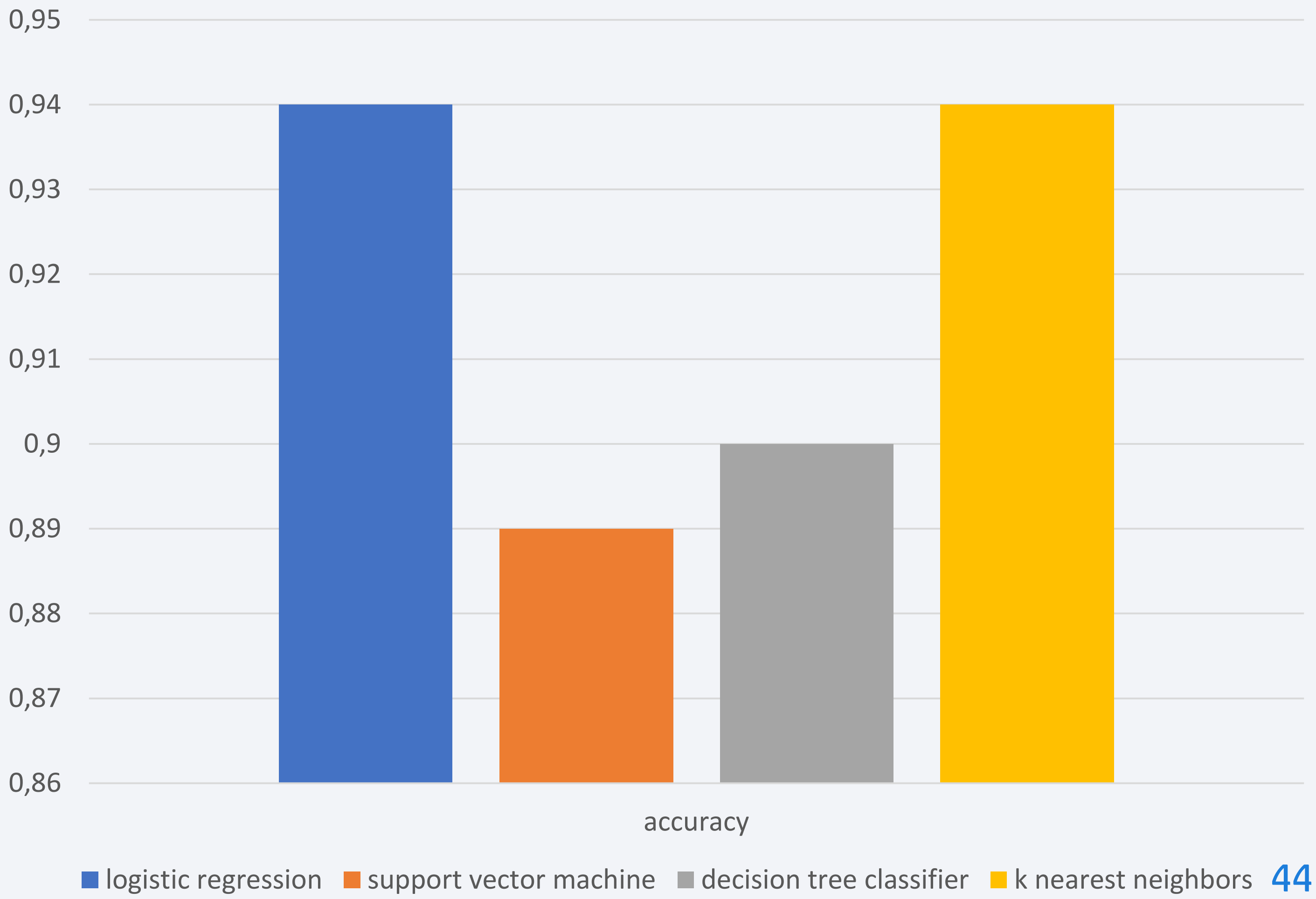
- The FT booster version has the highest success rate.
- The payload range with the highest success rate is between 2000 and 5000 Kg.

Section 5

Predictive Analysis (Classification)

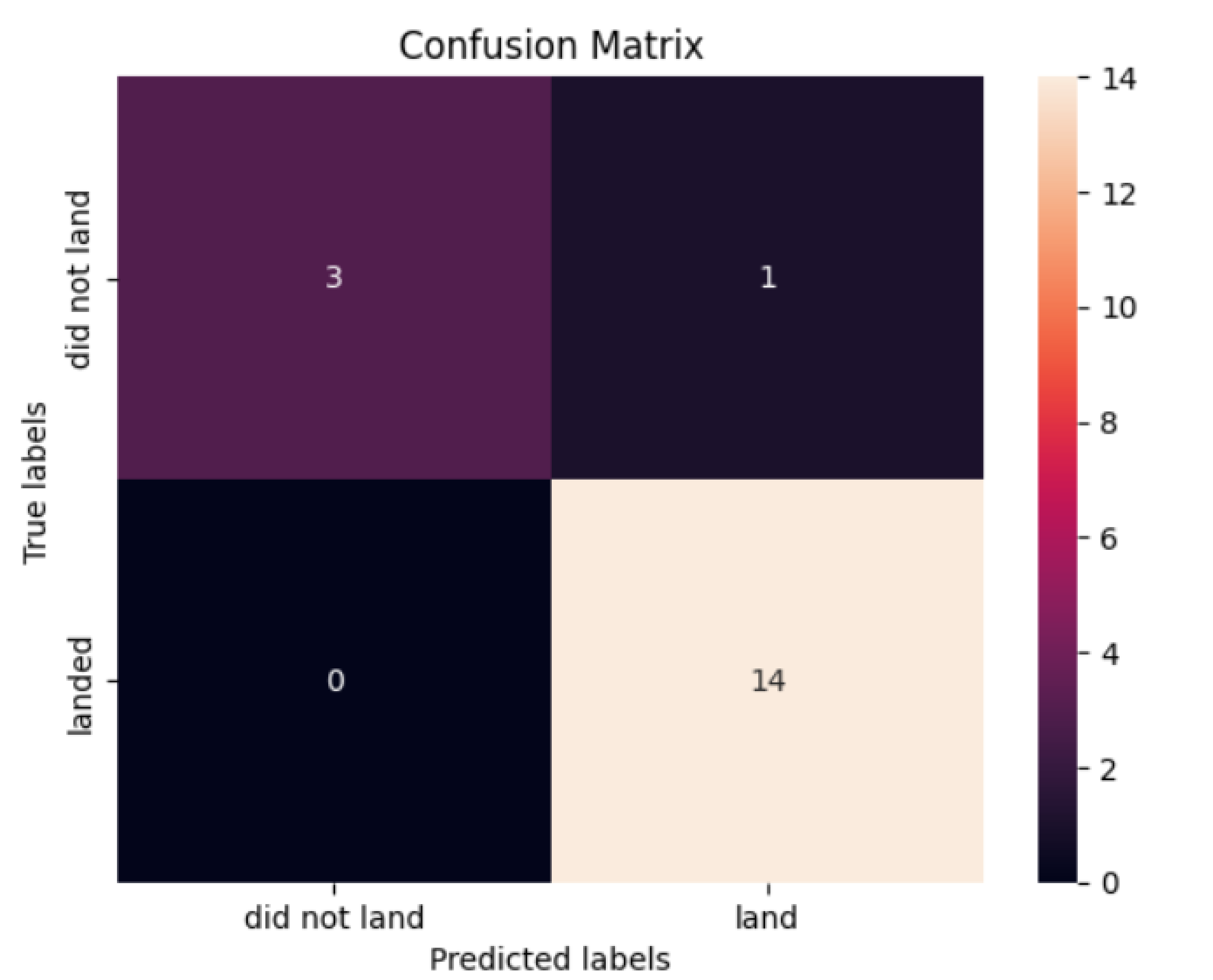
Classification Accuracy

- The models who has the highest classification accuracy are logistic regression and KNN



Confusion Matrix

There is only one prediction classified as 'land' that was actually 'did not land



Conclusions

- Logistic Regression outperformed other models with an accuracy of 94%, making it a reliable choice for predicting landing success.
- Payload mass, orbit type, and launch site emerged as the most influential factors
- The results of this study can guide SpaceX and its stakeholders in optimizing mission planning, launch site selection, and payload configurations.
- Further research is recommended to expand the dataset and incorporate additional variables such as weather conditions and booster modifications to refine predictions.

Appendix

- Dataset collected from the SpaceX API
- Dataset collected through web scraping
- Dataset after data wrangling

Thank you!