IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

HUANG Junxin
9/3/2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through API: Utilizing Application Programming Interfaces (APIs) to gather structured data.
- Data Collection with Web Scraping: Employing techniques to extract data from websites.
- Data Wrangling: Cleaning, structuring, and enriching raw data into a desired format for better decision making in less time.
- Exploratory Data Analysis with SQL: Using Structured Query Language (SQL) to explore and analyze data, identifying patterns and insights.
- Exploratory Data Analysis with Data Visualization: Creating visual representations of data to uncover hidden patterns and convey information clearly.
- Interactive Visual Analytics with Folium: Using Folium library for visualizing geospatial data to create interactive maps for in-depth analysis.
- Machine Learning Prediction: Applying machine learning algorithms to predict future events or outcomes based on historical data.

Executive Summary

Summary of all results

- Exploratory Data Analysis result: Insightful patterns and trends identified through initial data analysis.
- Interactive analytics in screenshots: Captured visual analytics that allow for interactive exploration of data, often saved as screenshots.
- Predictive Analytics result: Outcomes of machine learning predictions, stating the accuracy, effectiveness, and potential applications.

Introduction

Project background and context

- SpaceX has disrupted the space industry with its Falcon 9 rocket, which is offered at a cost of $62 million per launch on its website. This price point is significantly lower than that of other providers, who charge upwards of $165 million. A significant portion of SpaceX's cost savings derives from its ability to reuse the first stage of the Falcon 9. As such, predicting whether the first stage will land successfully is not only a technical challenge but also a financial one. If we can predict successful landings, we can estimate the cost of a launch more accurately. This prediction could serve as a crucial piece of information for alternate companies aiming to compete with SpaceX in the launch services market. The project's goal is to develop a machine learning pipeline that can predict the successful landing of the rocket's first stage.

Introduction

Problems you want to find answers

- On what factors does the successful landing of a rocket depend?
- What are the interactions between the various factors that determine the success of a landing?
- What operational conditions are required to ensure a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using the SpaceX REST API, which provides a wealth of information about SpaceX launches. The specific endpoint used for this capstone assignment was api.spacexdata.com/v4/launches/past, which returns data on past SpaceX launches.
- Perform data wrangling
  - Through Filtering Data and Resolving Missing IDs

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data collected prior to this step were normalised into training and test datasets and evaluated by four different classification models, each of which was evaluated for accuracy by different combinations of parameters.

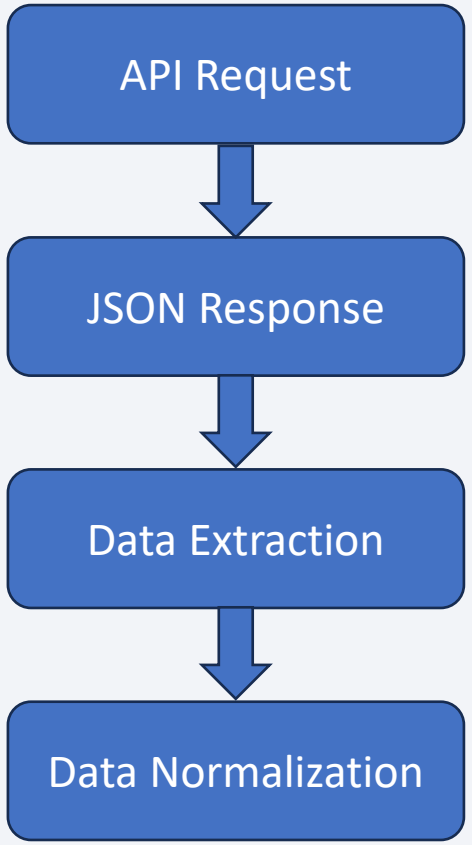
Data Collection

- The data was collected using the SpaceX REST API, which provides a wealth of information about SpaceX launches. The specific endpoint used for this capstone assignment was api.spacexdata.com/v4/launches/past, which returns data on past SpaceX launches.

Data Collection – SpaceX API

Steps for Data Collection – SpaceX API:

- API Request: A GET request was made to the SpaceX API endpoint using the requests library in Python.
- JSON Response: The data returned from the API was in JSON format, a list of JSON objects where each object represented a past launch.
- Data Extraction: The .json() method was called on the response object to extract the JSON data.
- Data Normalization: To transform the structured JSON data into a flat table suitable for analysis, the json_normalize function was employed, which converts nested JSON into a pandas DataFrame.

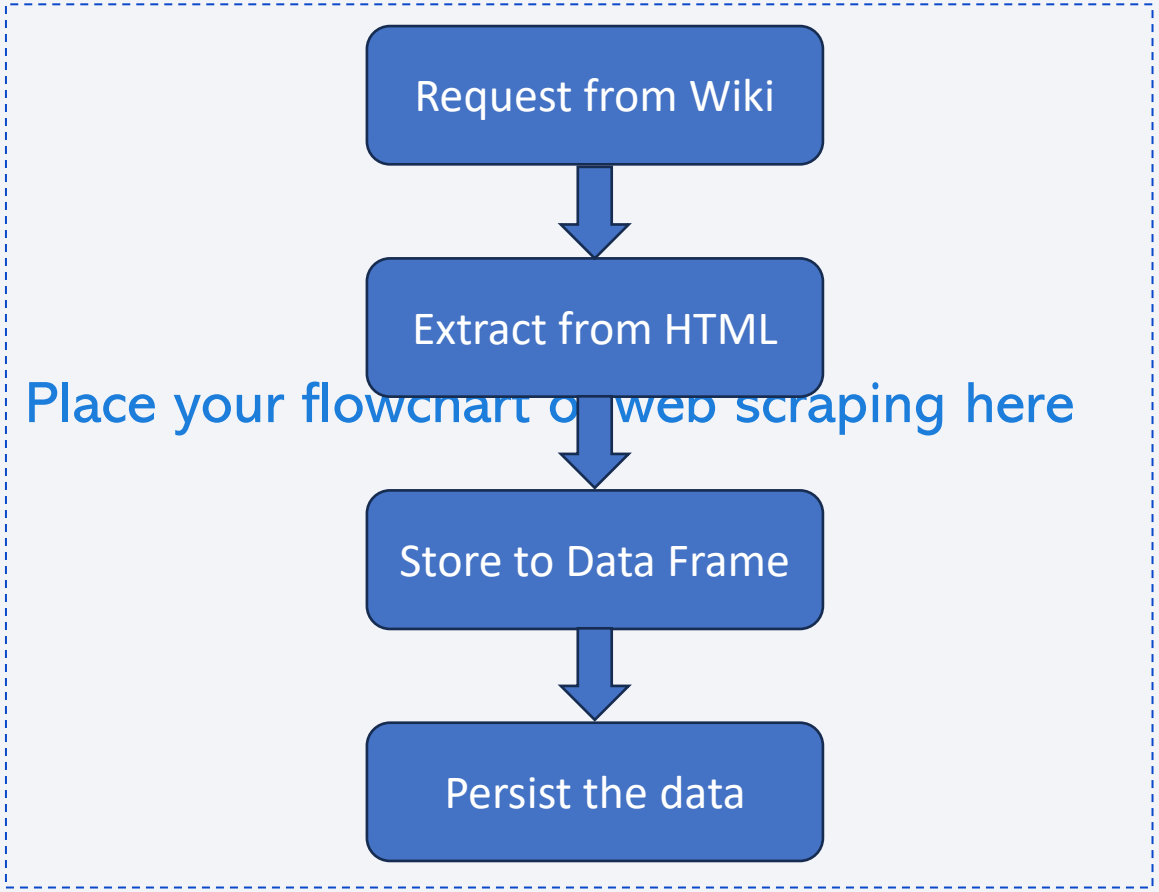


Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%201/Hands-on%20Lab%20Complete%20the%20Data%20Collection%20API%20Lab/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Steps for Data Collection – Scriping:

- SpaceX launch data is also available from Wikipedia;
- Download the data from Wikipedia according to the flowchart and then persist it.



Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%201/Hands-on%20Lab%20Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab/jupyter-labs-webscraping.ipynb

Data Wrangling

- Filtering Data: The dataset initially included launches for both Falcon 1 and Falcon 9 boosters. The data needed to be filtered to exclude Falcon 1 launches to focus solely on Falcon 9 data.
- Resolving Missing IDs: Some columns contained identification numbers rather than direct data. By targeting specific endpoints on the API (e.g., Booster, Launchpad, Payload, Core), detailed information for each ID was retrieved and compiled into lists to enrich the dataset.

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%201/Hands-on%20Lab%20Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Catplot of Payload Mass vs. Flight Number colored by class (launch outcome) - This chart helps visualize the relationship between flight number, payload mass, and launch success. It shows that as flight number increases, success rate improves, and heavier payloads are associated with lower success rates.
- Catplot of Flight Number vs. Launch Site colored by class - This scatter plot depicts the distribution of flight numbers across different launch sites, with the outcome indicated by color. It allows identifying which launch sites have higher flight numbers and success rates.
- Scatter plot of Payload Mass vs. Launch Site colored by class - Similar to the previous plot, this one shows the distribution of payload mass launched from each site, along with the outcome. It reveals that certain sites handle heavier payloads and may have differing success rates.

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%202/EDA%20with%20Visualization%20Lab/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

- Bar chart of success rate for each orbit type - By grouping flights by orbit and calculating the mean success rate, this bar chart easily compares the success rates across different orbit types. It identifies which orbits have the best historical success.
- Scatter plot of Flight Number vs. Orbit colored by class - This plot analyzes the potential relationship between flight number, orbit type and mission outcome. It can highlight if certain orbits have improving success rates as they are attempted more.
- Scatter plot of Payload Mass vs. Orbit colored by class - Plotting payload mass against orbit type with outcome identified by color can reveal if heavier payloads tend to be associated with specific orbits and if there is any link between payload mass, orbit and success rate.
- Line plot of yearly average success rate over time - By calculating the mean success rate for each year and plotting it as a line chart, the trend in success rates over the years can be visualized. An improving trend would indicate increasing reliability.

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%202/EDA%20with%20Visualization%20Lab/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome on a ground pad was achieved
- Listed the names of the boosters which had successful drone ship landings and payload mass between 4000 kg and 6000 kg
- Listed the total number of successful and failed mission outcomes
- Listed the names of the booster versions which carried the maximum payload mass, using a subquery
- Listed the records showing the month names, failed drone ship landing outcomes, booster versions, and launch sites for the months in 2015
- Ranked the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%202/Hands-on%20Lab%20Complete%20the%20EDA%20with%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and clusters of markers are used with Folium maps
- Markers indicate points such as launch sites
- Circles indicate highlighted areas around specific coordinates, such as the NASA Johnson Space Centre
- Marker clusters represent clusters of events at each coordinate, such as launches at a launch site
- Lines indicate the distance between two coordinates

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%203/Hands-on%20Lab%20Interactive%20Visual%20Analytics%20with%20Folium%20lab/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Created an interactive dashboard with Plotly dash
- Pie charts were created showing the total number of launches for a given site
- Plotted scatter plots showing the relationship between different booster versions and outcomes and payload mass in kilograms.

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%203/Hands-on%20Lab%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was loaded using numpy and pandas, transformed and split into training and testing parts.
- Different machine learning models were built and tuned with different hyperparameters using GridSearchCV.
- Accuracy was used as a measure of the model and feature engineering and algorithm tuning was used to improve the model.
- Finally the best performing classification model was found.

Source code: https://github.com/huangjunxin/IBM-DS-Capstone/blob/main/Week%204/Hands-on%20Lab%20Complete%20the%20Machine%20Learning%20Prediction%20lab/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

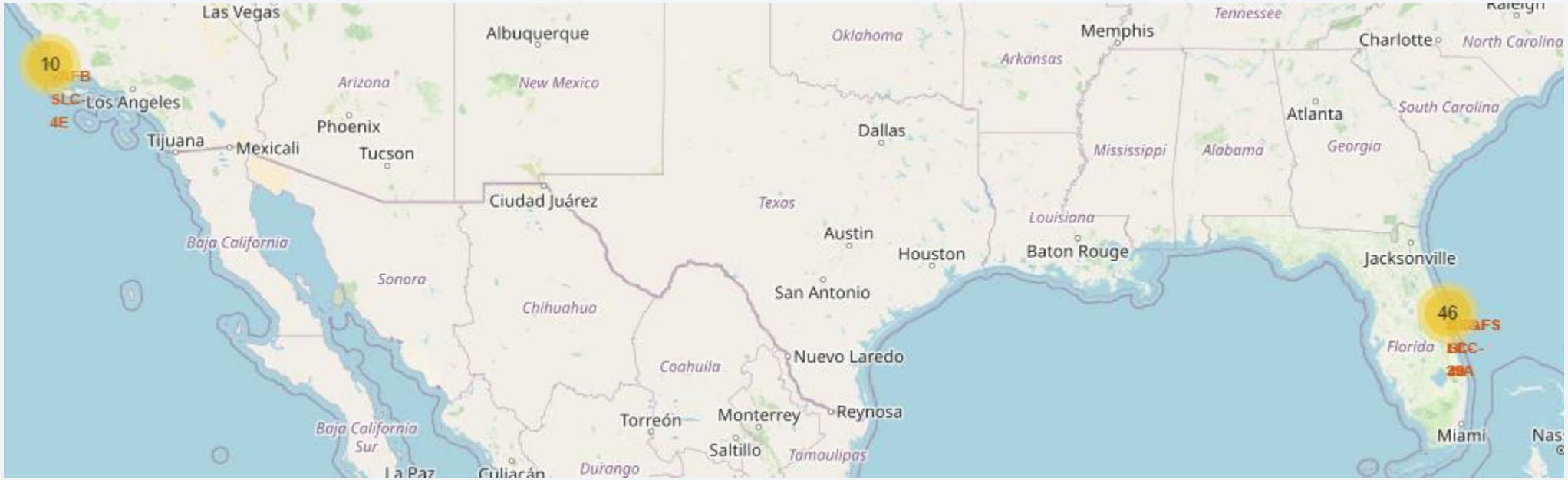
Exploratory data analysis results

- SpaceX operates from four distinct launch sites.
- The company's early launches served both its own and NASA's needs.
- The average payload capacity of the Falcon 9 version 1.1 booster is 2,928 kg.
- A milestone was reached in 2015 with the first successful booster landing, five years after the first launch.
- Several versions of the Falcon 9 booster have successfully landed on drone ships, even with payloads above the average.
- The mission success rate is close to 100%.
- Two booster versions, specifically F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015.
- The rate of successful landings has shown improvement over the years.

Results

Interactive analytics demo in screenshots

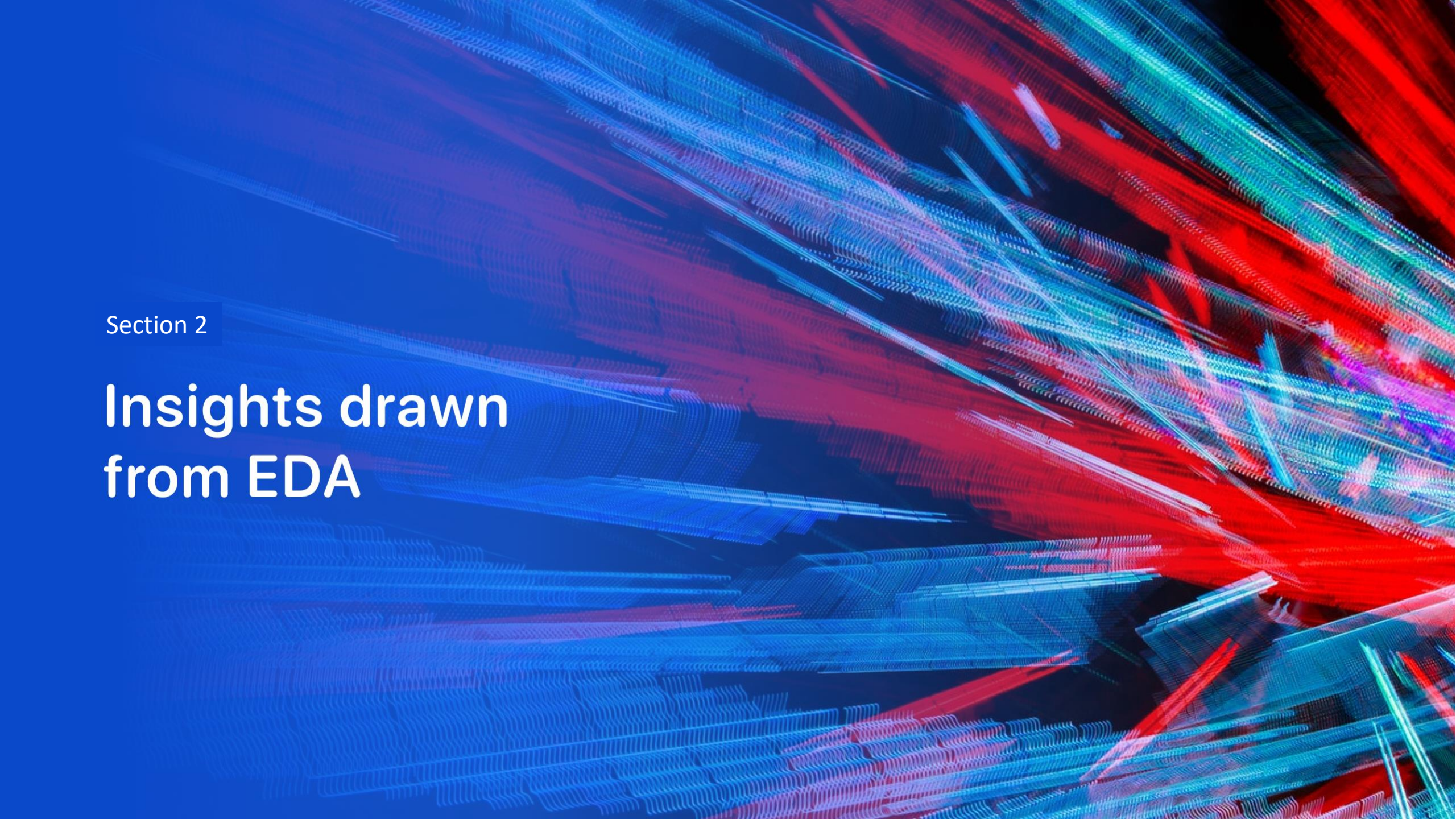
- Interactive analytics revealed that launch sites are typically situated in safe locations, often near the sea.
- These sites also benefit from robust logistic infrastructures in their vicinity.
- A majority of launches are conducted from launch sites located on the east coast.



Results

Predictive analysis results

- In terms of accuracy, the decision tree has the highest accuracy of 0.875 among all the classification models.
- The main problem shown in the confusion matrix is false positives, where the classifier marks unsuccessful landings as successful landings.

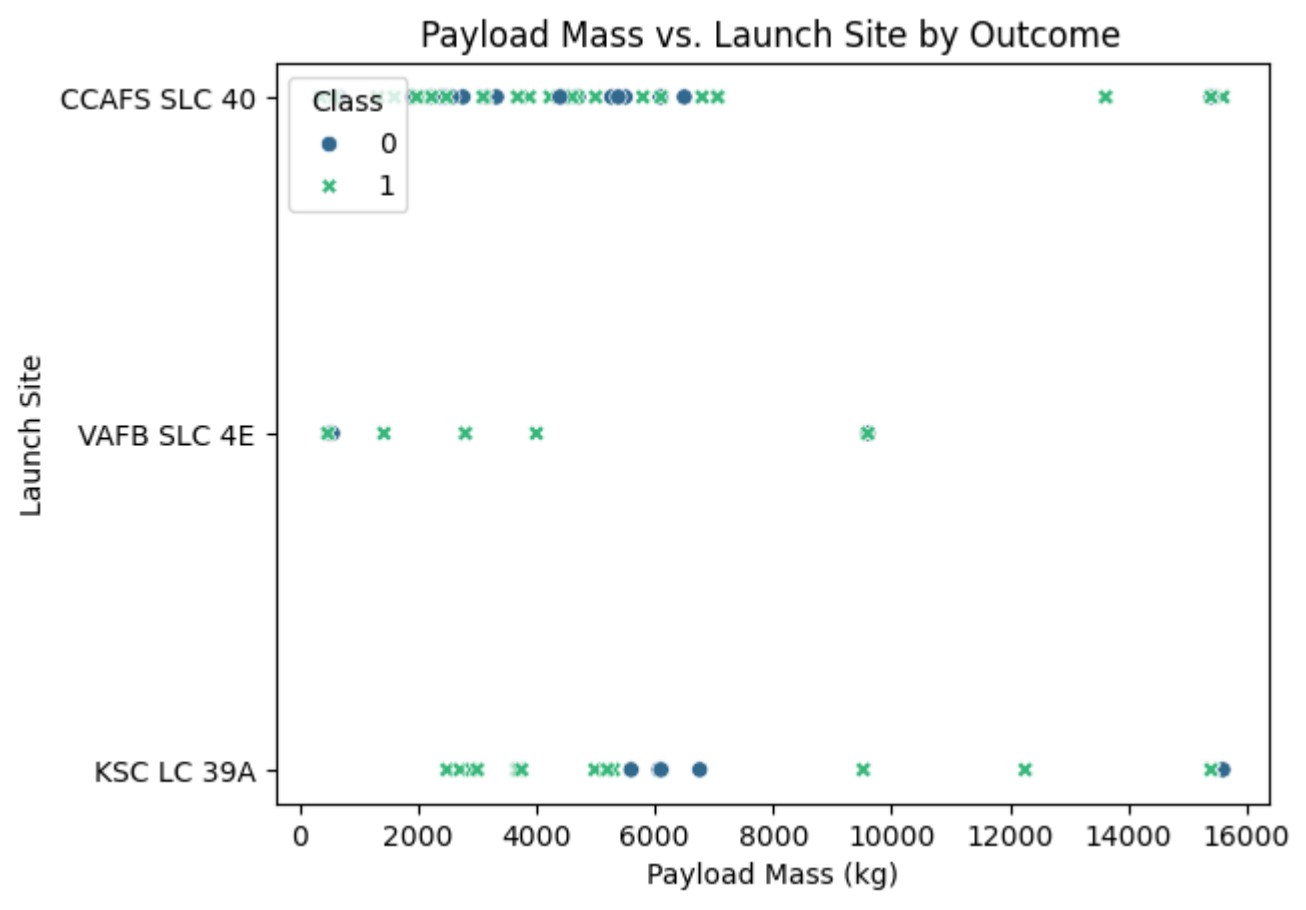

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- The more flights, the higher the success rate at the launch site.

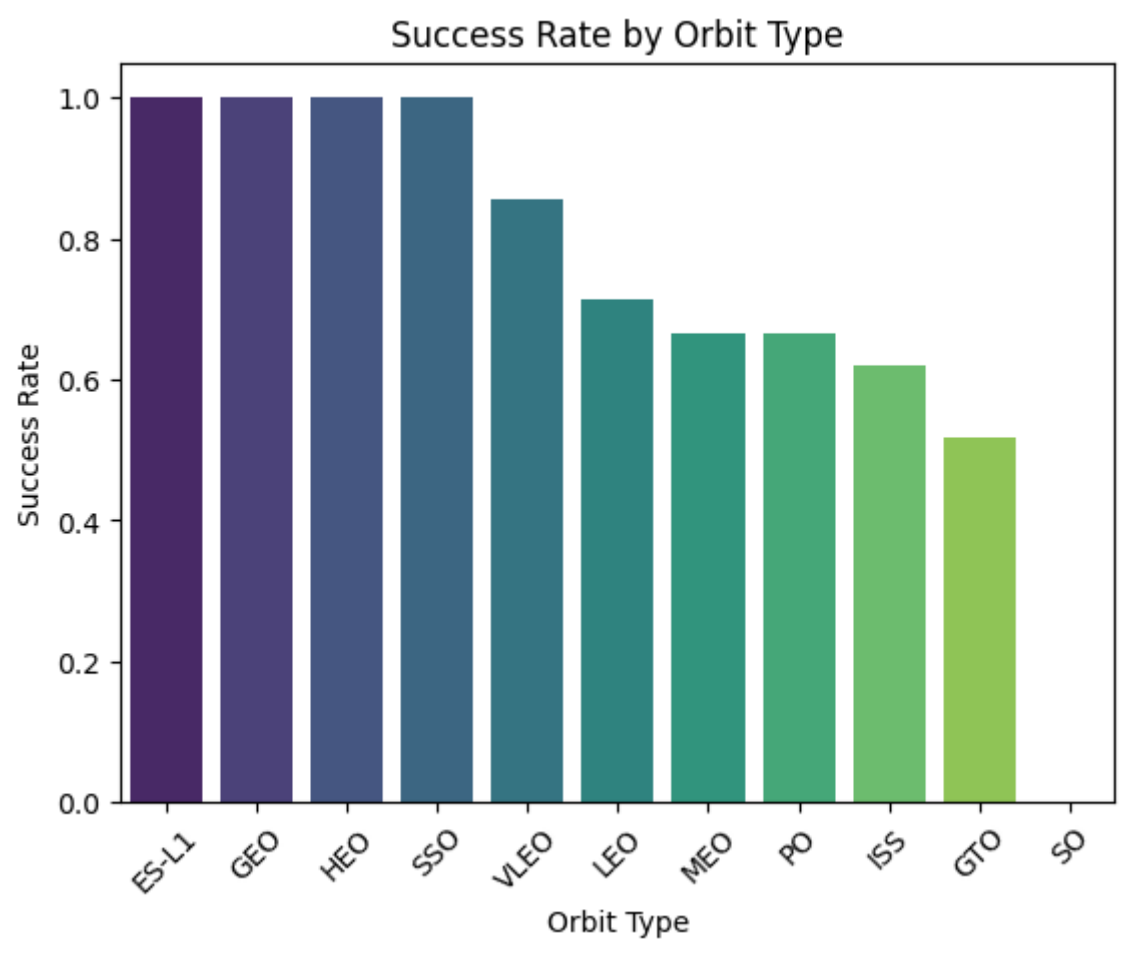
Payload vs. Launch Site

- The larger the payload mass of the CCAFS SLC 40 at the launch site, the higher the success rate of the rocket.



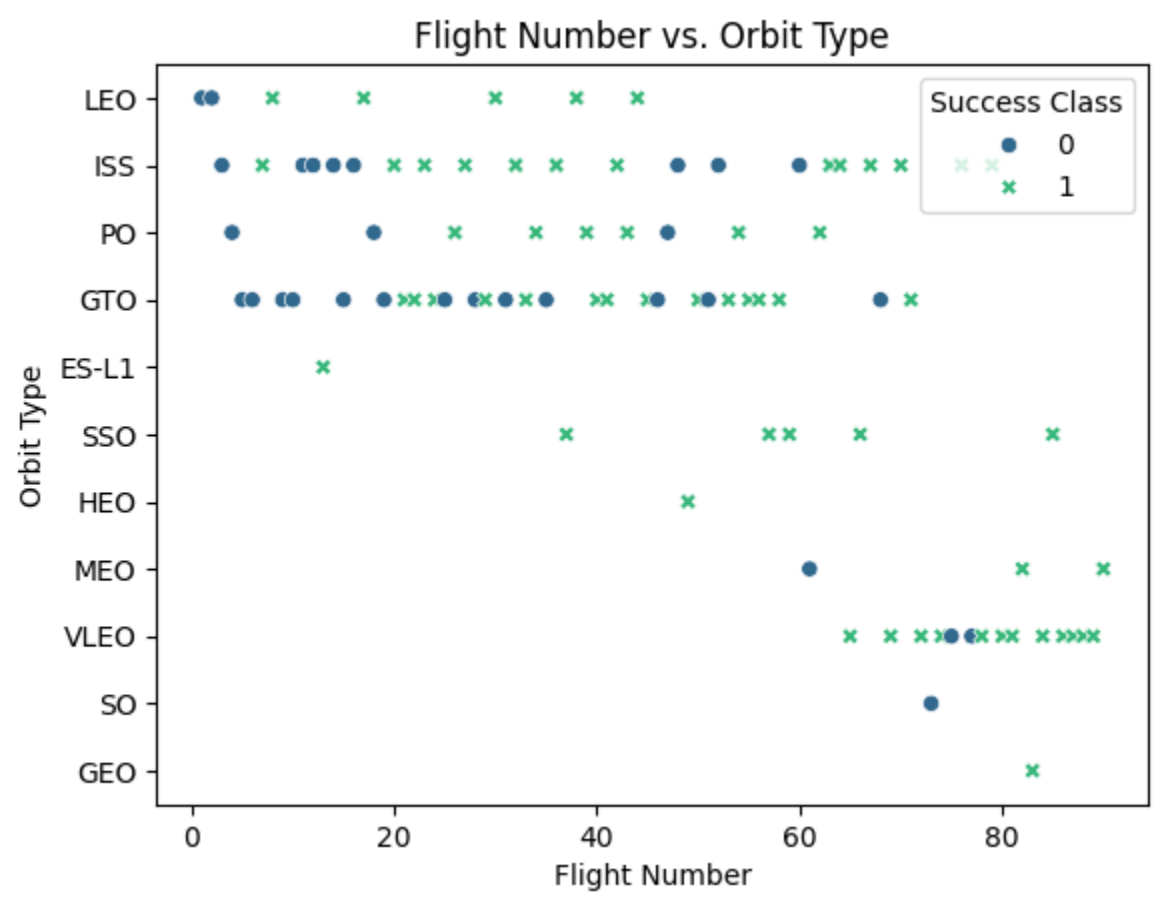
Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, SSO had the highest success rate.



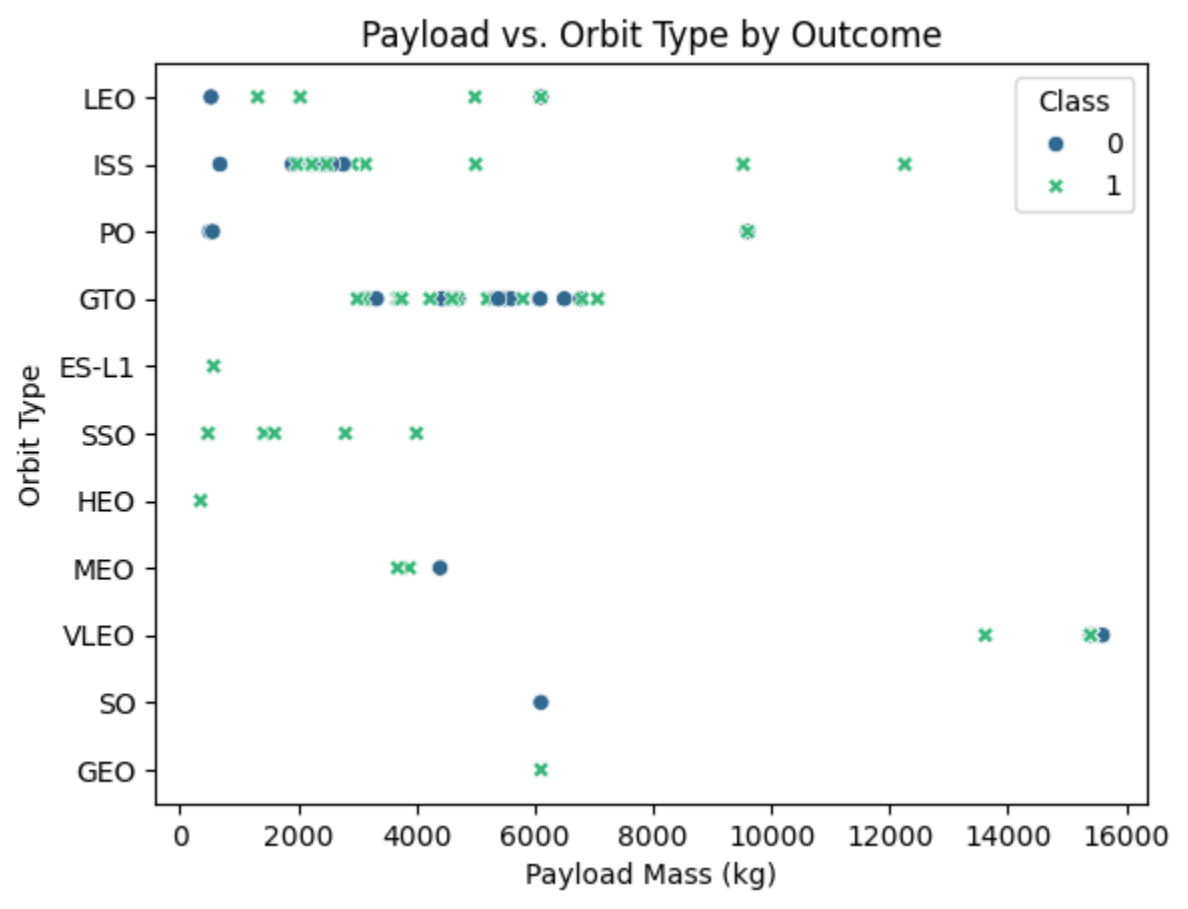
Flight Number vs. Orbit Type

- The data points are distributed randomly, and there is no clear relationship between Flight Number and Orbit.



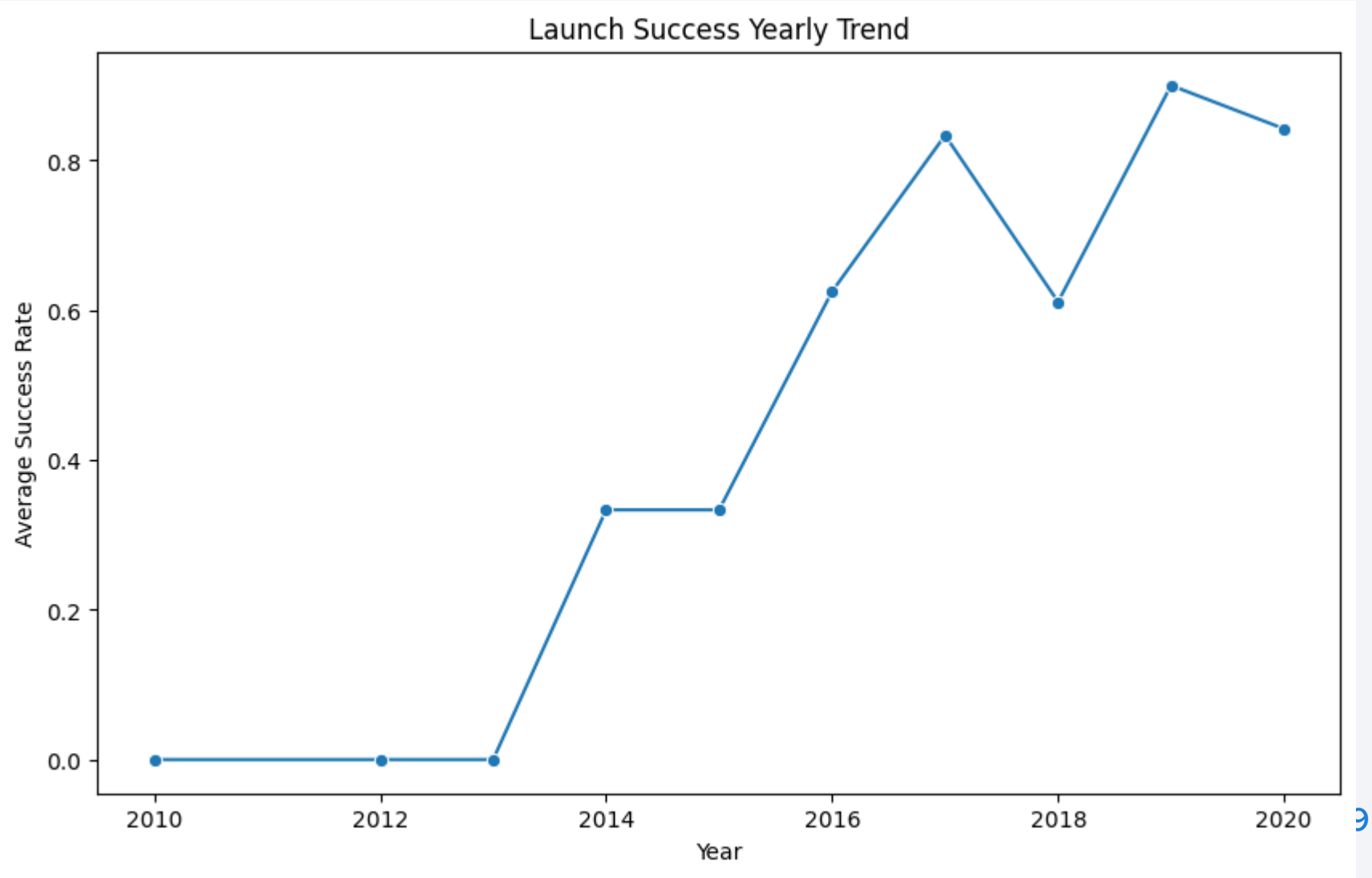
Payload vs. Orbit Type

- LEO, ISS and PO orbits are more likely to have successful landings due to larger payloads.



Launch Success Yearly Trend

- Since 2013, the success rate has been increasing until 2020.



All Launch Site Names

- Find the names of the unique launch sites using DISTINCT

Task 1

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

```
[11] %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA` using LIMIT

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

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

Total Payload Mass

- Calculate the total payload carried by boosters from NASA using SUM()

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS total_payload_mass FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1 using AVG()

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload_mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

average_payload_mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad using MIN()

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

Hint:Use min function

```
%sql SELECT MIN("Date") AS first_successful_landing_date FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

first_successful_landing_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

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

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MA
```

Python

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS number_of_missions FROM SPACEXTABLE GROUP BY "Mission_Outcome" HAVING "Mission_Outcome" LIKE 'Succe
```

Python

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

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

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

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

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

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' fo year.

```
%sql SELECT SUBSTR("Date", 6, 2) AS month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE
```

Python

```
 * sqlite:///my_data1.db
Done.
```

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

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

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.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing
```

Python

```
 * sqlite:///my_data1.db
Done.
```

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

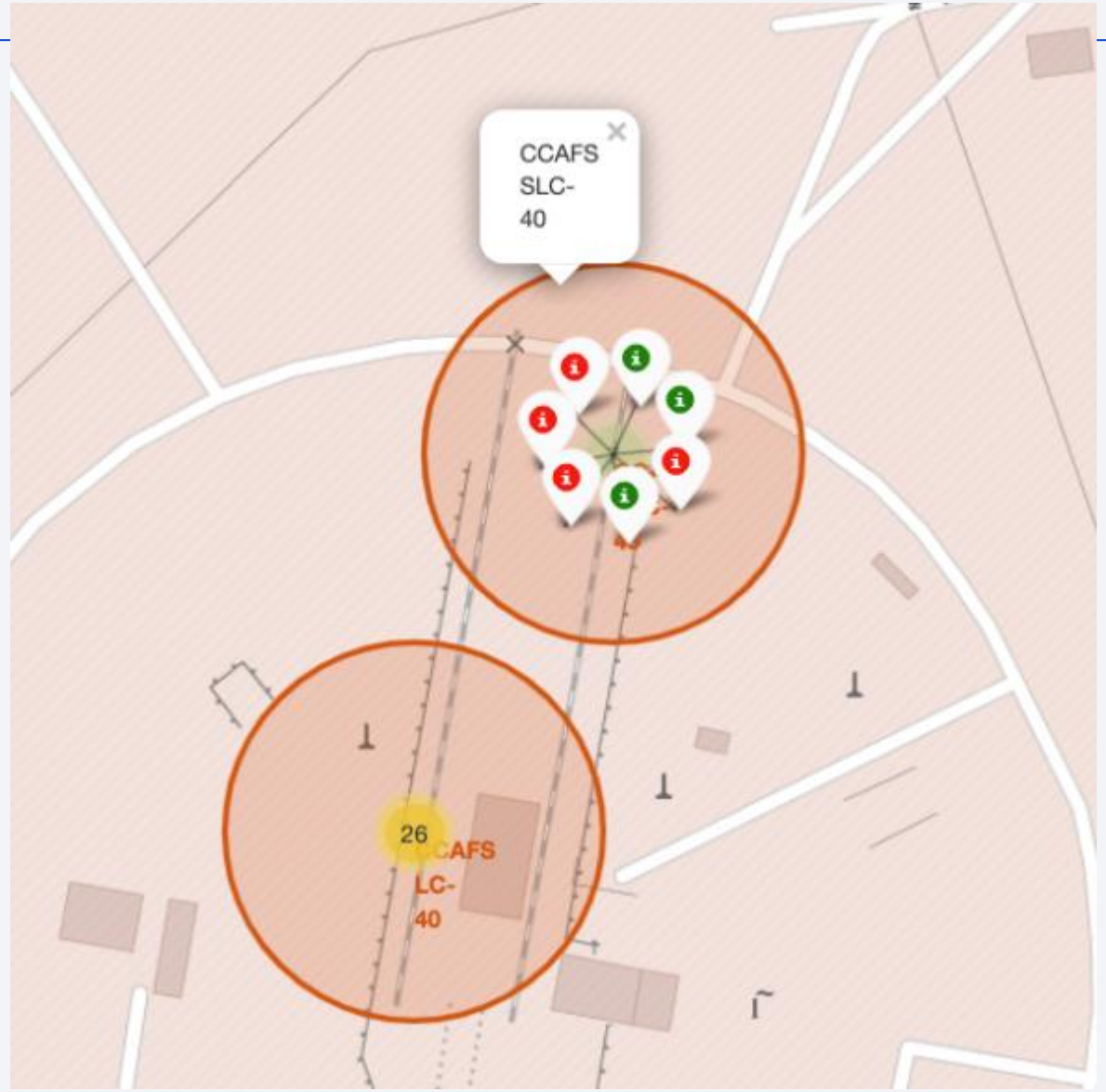
Launch Sites Proximities Analysis

Launch Locations

- The launch locations are in Florida and California.

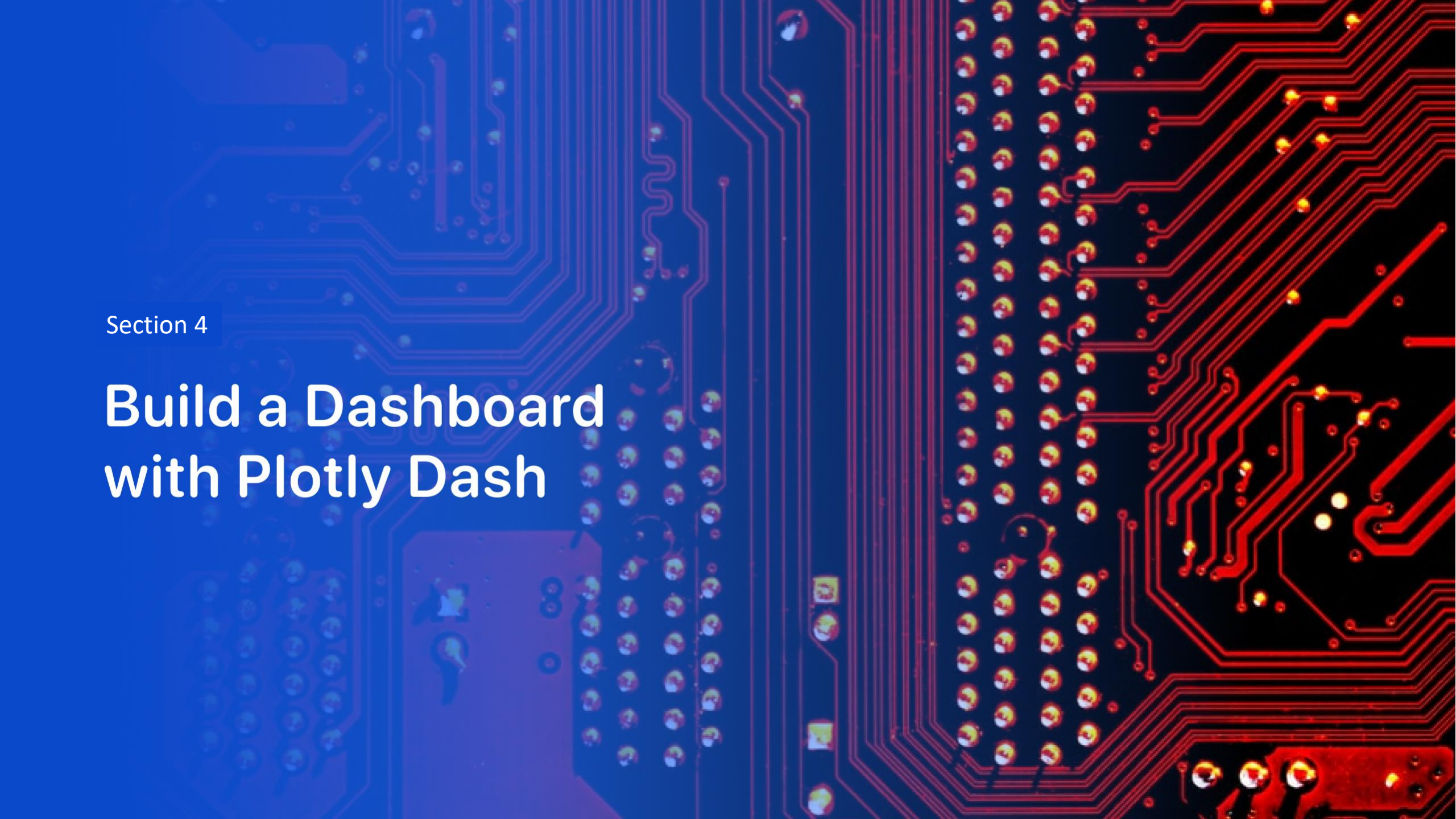
Use of colours to mark successful launches

- The red markers indicates failure and the green markers indicates success.



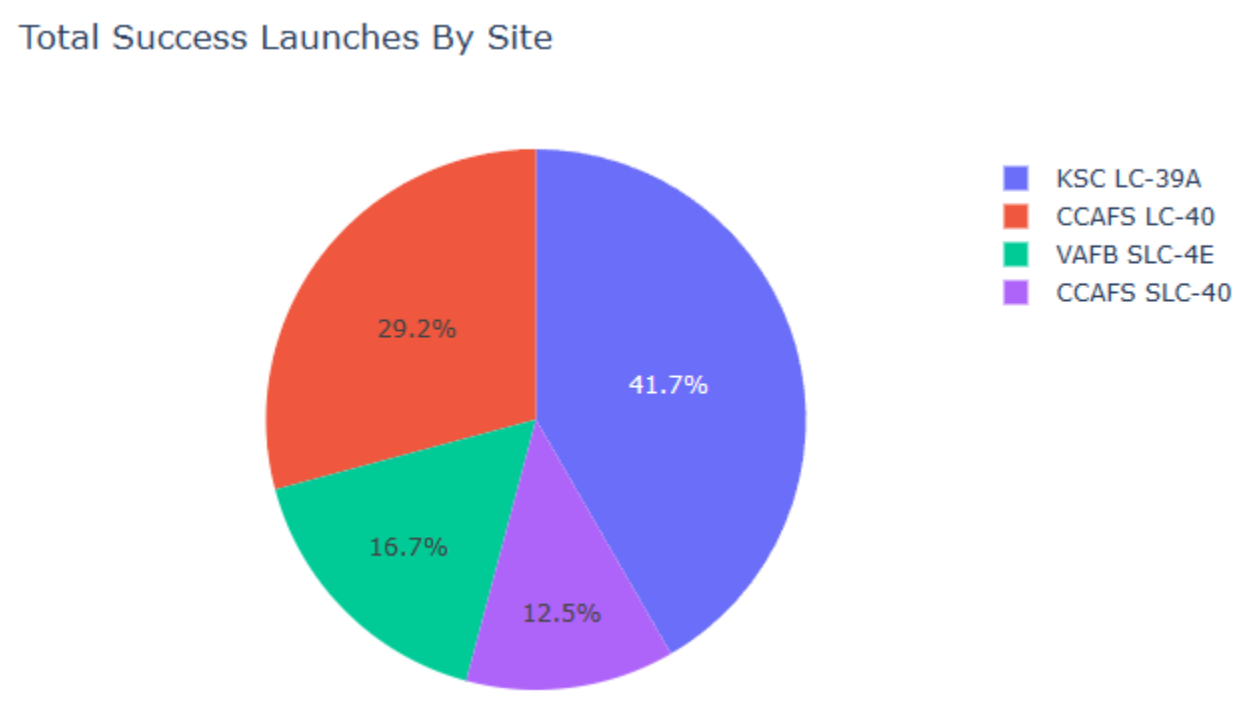
Launch sites are close to the coastline

- It was found that the launch sites were all very close to the coastline.


Section 4
Build a Dashboard with Plotly Dash

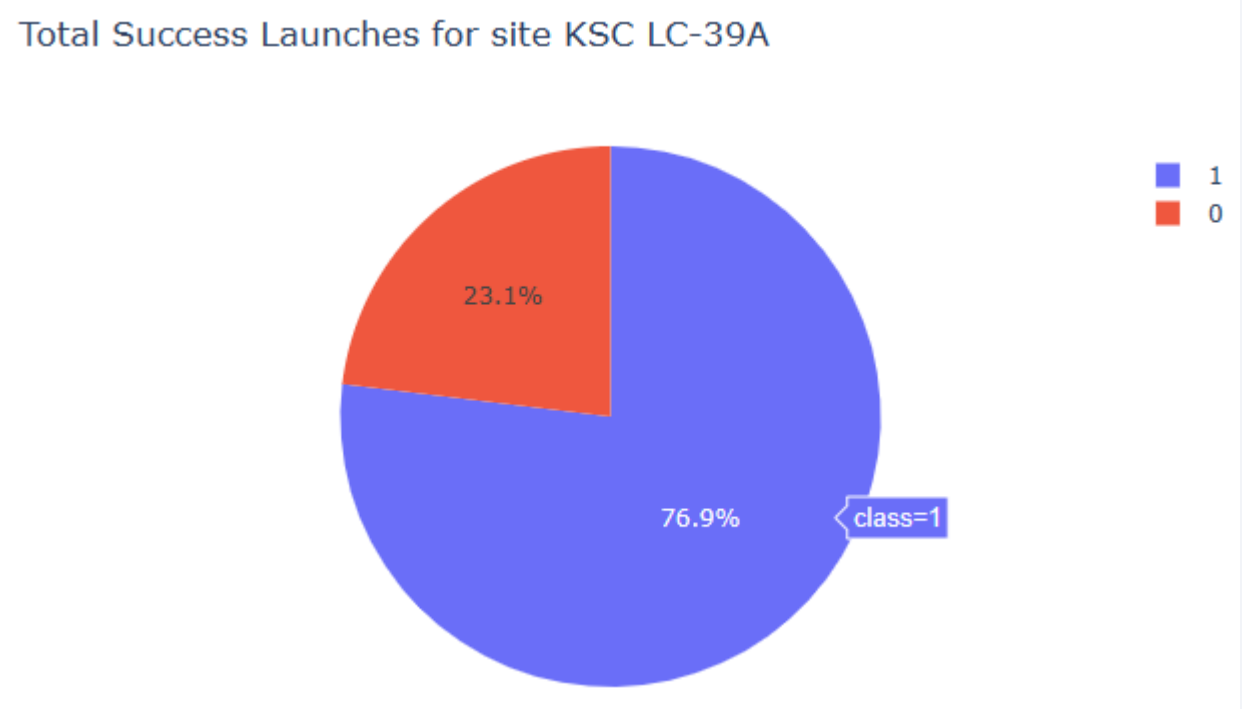
Total Success Launches by Sites

- KSC LC-39A is the most successful launches from all sites



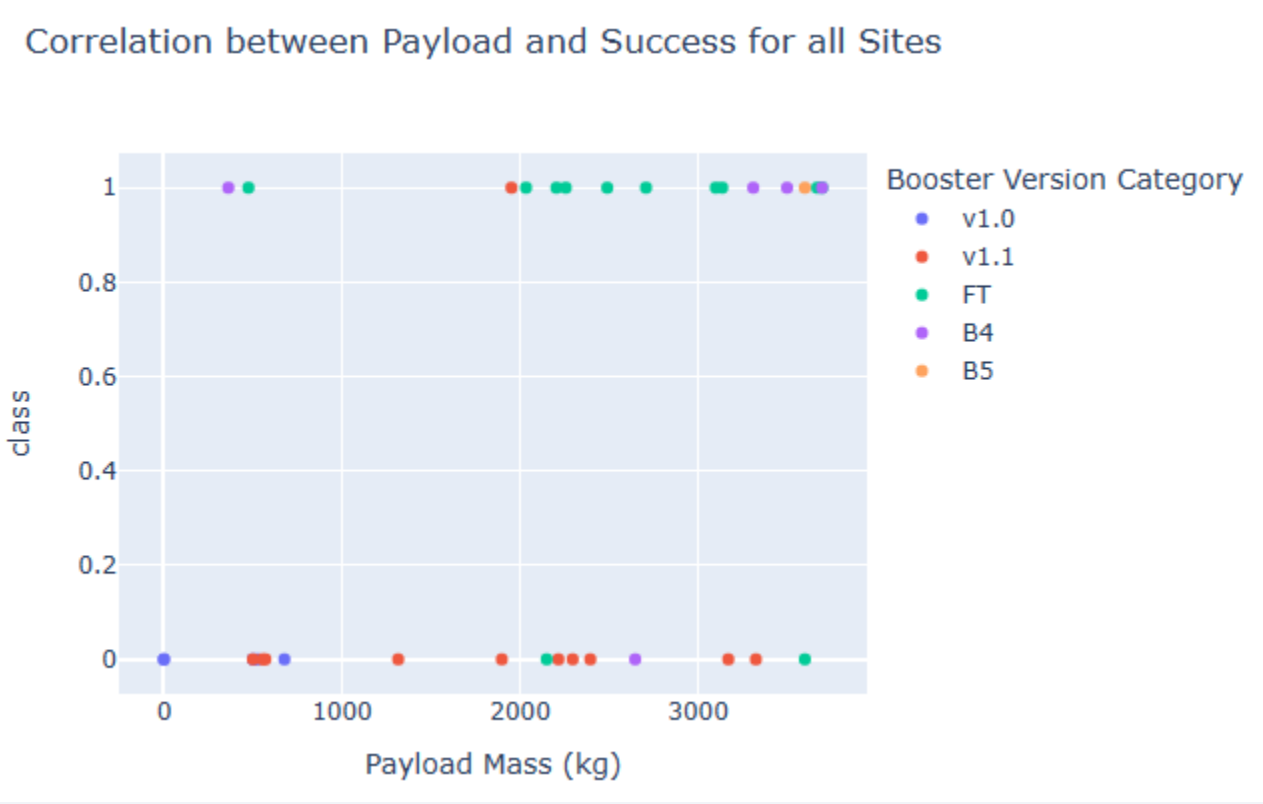
Launch Success Rate of KSC LC-39A

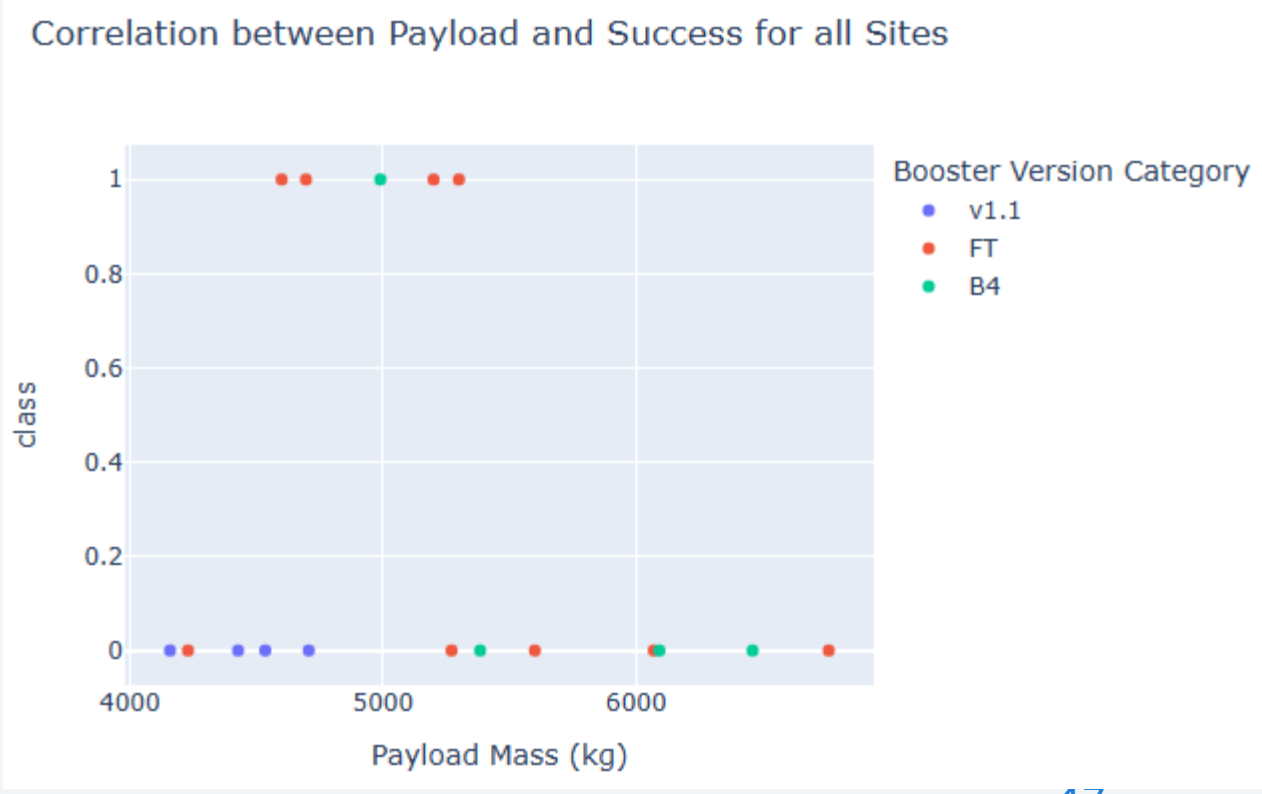
- The launch success rate of KSC LC-39A was 76.9%.



Effect of Different Payloads on the Success Rate

- Low-weight payloads have a higher success rate than heavy payloads





Section 5

Predictive Analysis (Classification)

Classification Accuracy

- We can see the results on the right-side.
- In terms of accuracy, the decision tree has the highest accuracy of 0.875.

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solve
accuracy : 0.8464285714285713
```

```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168
accuracy : 0.8482142857142856
```

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

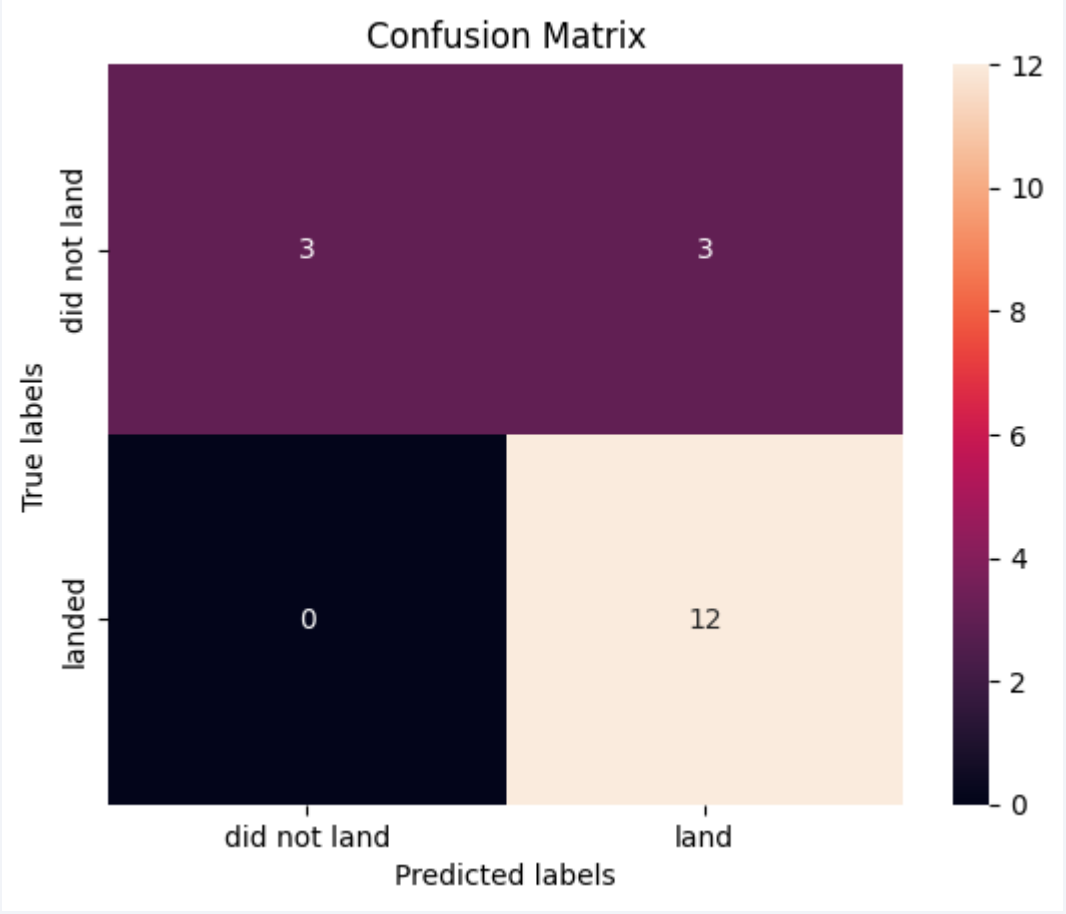
```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 8,
'min_samples_split': 10, 'splitter': 'random'}
accuracy : 0.875
```

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbor
accuracy : 0.8482142857142858
```

Confusion Matrix

- The confusion matrix of the decision tree classifier shows that the classifier can distinguish between different categories.
- The main problem is false positives, where the classifier marks unsuccessful landings as successful landings.



Conclusions

We can conclude that:

- There is a correlation indicating that a higher volume of flights at a launch site tends to be associated with a higher success rate at that site.
- The overall launch success rate began to climb in 2013 and continued to rise through 2020.
- The orbits designated as ES-L1, GEO, HEO, SSO, and VLEO have experienced the highest success rates.
- Kennedy Space Center's Launch Complex 39A (KSC LC-39A) has recorded the highest number of successful launches compared to other sites.
- The Decision Tree classifier has been identified as the most effective machine learning algorithm for analyzing and predicting these outcomes.

Thank you!