IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Antuan Vayisqui
25-02-2024

Outline

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

Executive Summary

- Summary of methodologies

- Data Collection using web scraping and SpaceX API.

- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics with Dash and Matplotlib.

- Machine Learning Prediction: Decision Tree Classifier, Support Vector Machine (SVM), Logistic Regression and K-Nearest Neighbors (KNN).

- Summary of all results

- Valuable data was gathered from public sources.

- Exploratory Data Analysis (EDA) pinpointed the most predictive features for launch success.

- Machine Learning Prediction determined the optimal model for predicting key characteristics that drive opportunities most effectively, leveraging the entire dataset.

Introduction

- Project background and context

- The aim is to assess the feasibility of the new company Space Y in competing with Space X.

Problems you want to find answers

- Determining the most accurate method for estimating total launch costs by forecasting successful first-stage rocket landings.

- Identifying the optimal launch location.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data from Space X was obtained from 2 sources:

  - Space X API (https://api.spacexdata.com/v4/rockets/)

  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling

  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models

  - The data gathered up to this point underwent normalization, division into training and testing datasets, and evaluation through four distinct classification models. Each model's accuracy was assessed using various parameter combinations.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts.

  - Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping.

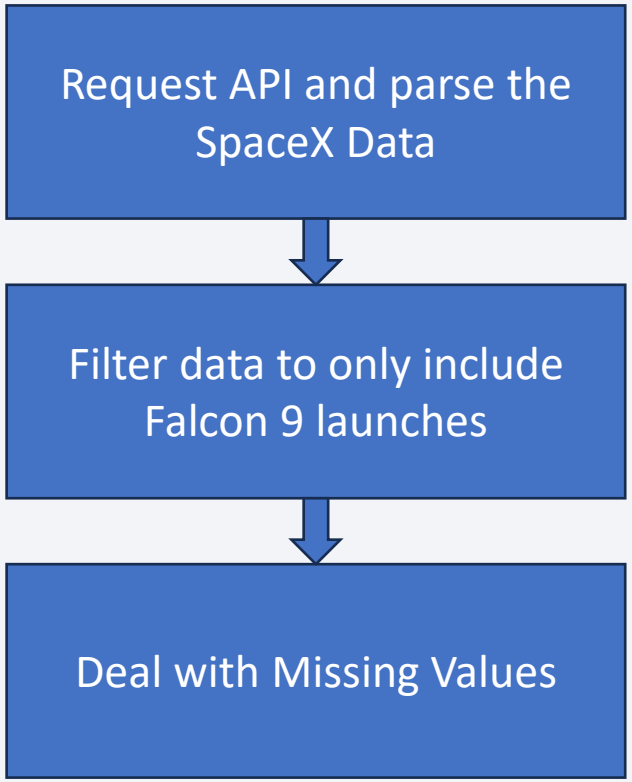
Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

  - SpaceX provides a public API for accessing data, which is utilized following the outlined flowchart.

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

  - https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



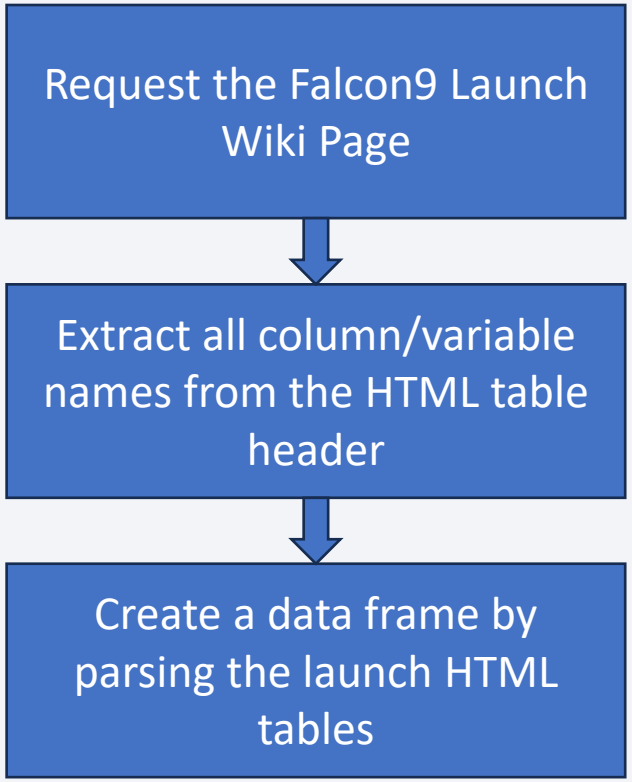
Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Data is downloaded from Wikipedia in accordance with the provided flowchart. Information regarding SpaceX launches is also accessible through Wikipedia.

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Describe how data were processed

- Initially, Exploratory Data Analysis (EDA) in the dataset.

- Summaries of launches per site, occurrences of each orbit, and occurrences of mission outcomes per orbit type were computed.

- Finally, the landing outcome label was generated from the Outcome column.

- You need to present your data wrangling process using key phrases and flowcharts

- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

- Visualizing the relationship between pairs of features involved utilizing scatterplots and barplots to explore the data. Specifically, the following relationships were examined: Payload Mass versus Flight Number, Launch Site versus Flight Number, Launch Site versus Payload Mass, Orbit versus Flight Number and Payload versus Orbit.

- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose.

- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Listing the unique launch site names involved in the space mission.
  - Identifying the top 5 launch sites whose names begin with the string 'CCA.'
  - Calculating the total payload mass carried by boosters launched by NASA (CRS).
  - Determining the average payload mass carried by booster version F9 v1.1.
  - Noting the date of the first successful landing outcome on a ground pad.
  - Listing the names of boosters that successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kg.
  - Summarizing the total number of successful and failed mission outcomes.
  - Identifying the booster versions that carried the maximum payload mass.
  - Detailing failed landing outcomes on a drone ship, including their booster versions and launch site names, specifically for the year 2015.
  - Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017.
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

- Folium Maps utilized markers, circles, lines, and marker clusters for visualization.

- Explain why you added those objects

- Markers were employed to pinpoint locations such as launch sites.
- Circles were utilized to highlight areas around specific coordinates, such as the NASA Johnson Space Center.
- Marker clusters were used to group events occurring at each coordinate, such as launches at a particular launch site.
- Lines were employed to illustrate distances between two coordinates.

- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard

- The following graphs and plots were used to visualize data: Percentage of launches by site and Payload range.

- Explain why you added those plots and interactions

- This combination facilitated a rapid analysis of the relationship between payloads and launch sites, aiding in the identification of the optimal launch locations based on payload considerations.

- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

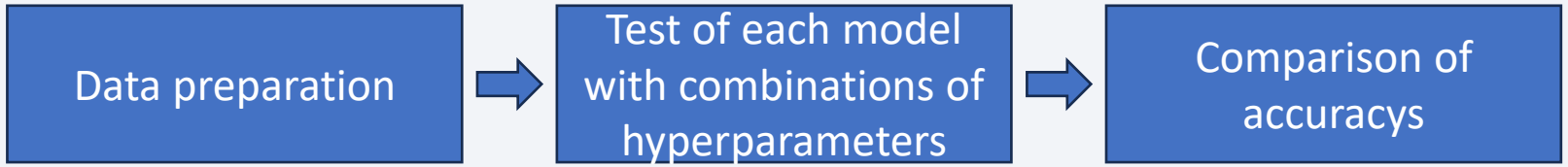
- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/spacex_dash_app.ipynb

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

- I prepared the data to compared classification models including logistic regression, support vector machine, decision tree, and k-nearest neighbors. I utilized GridSearchCV to find the optimal combination of hyperparameters. Afterward, I assessed the accuracy of each model to determine the best performing classification model.

You need present your model development process using key phrases and flowchart



- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

- https://github.com/AntuanVayisqui/Final-Project-Data-Science/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results

- Space X uses 4 different launch sites.

- The first launches were done to Space X itself and NASA.

- The average payload of F9 v1.1 booster is 2,928 kg.

- The first success landing outcome happened in 2015 fiver year after the first launch.

- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.

- Almost 100% of mission outcomes were successful.

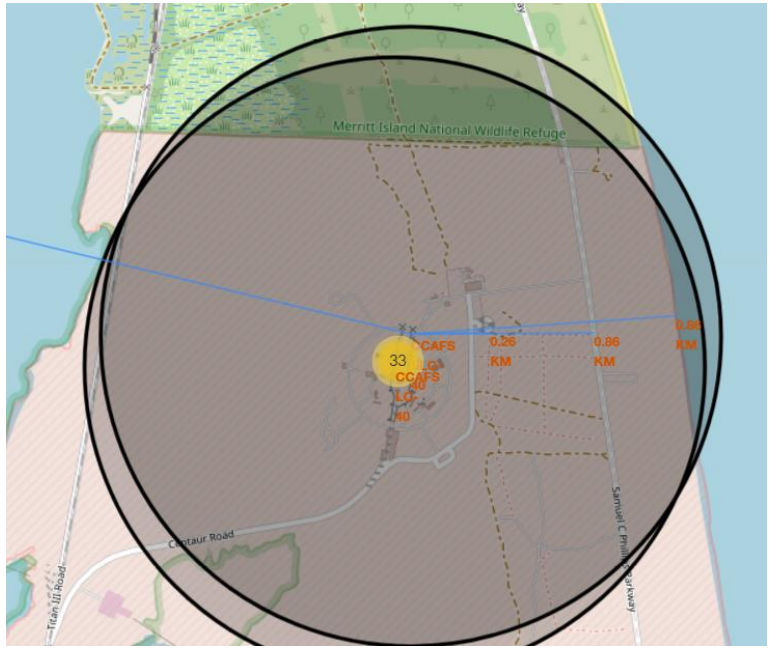
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.

- The number of landing outcomes became as better as years passed.

Results

- Interactive analytics demo in screenshots

- With interactive analytics, launch sites are near railways, highways, and the coastline, indicating convenient access to transportation networks and potentially advantageous coastal locations. However, they maintain a significant distance from cities, likely due to safety regulations or considerations regarding population density and associated risks.

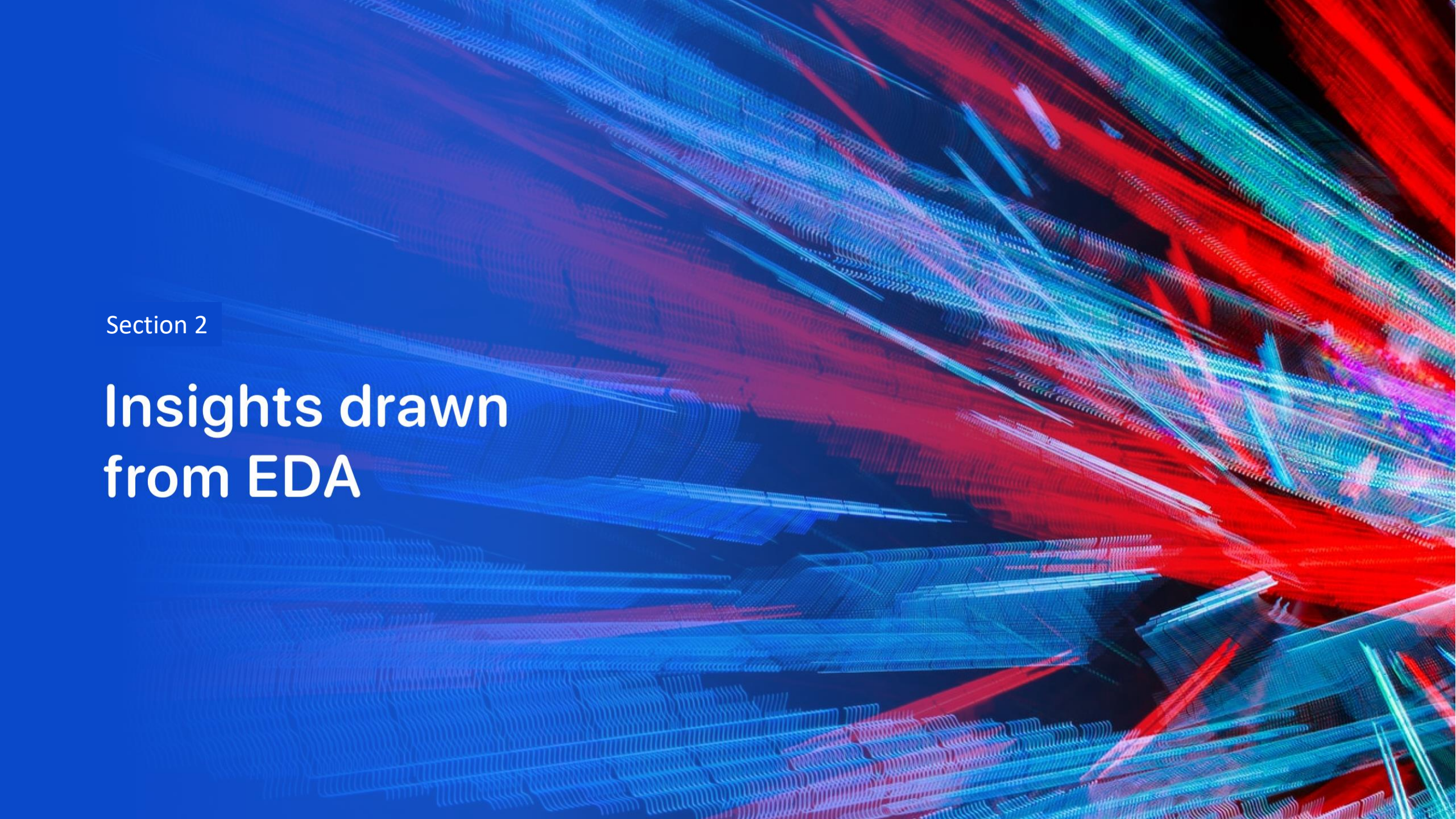


Results

- Predictive analysis results

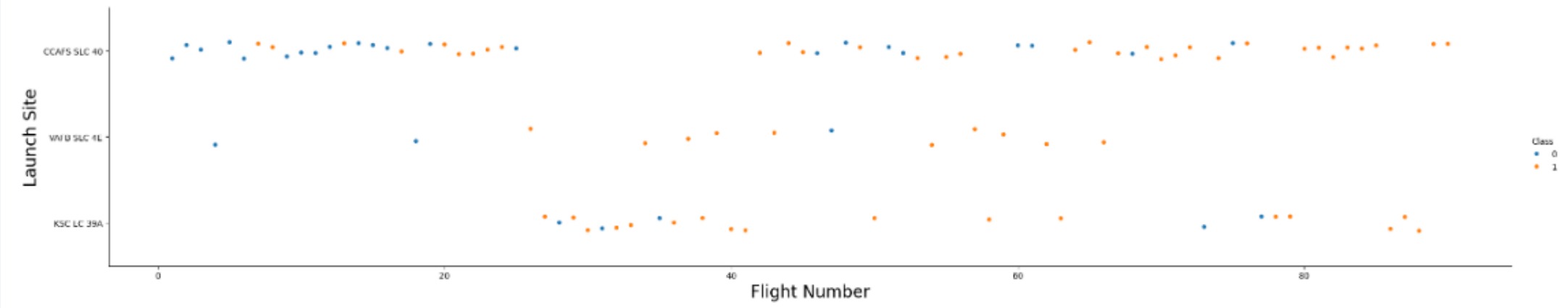
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87%.

```
Logistic Regression - Accuracy: 0.8464285714285713
Support Vector Machine - Accuracy: 0.8482142857142856
Decision Tree - Accuracy: 0.8767857142857143
K Nearest Neighbors - Accuracy: 0.8482142857142858
```


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

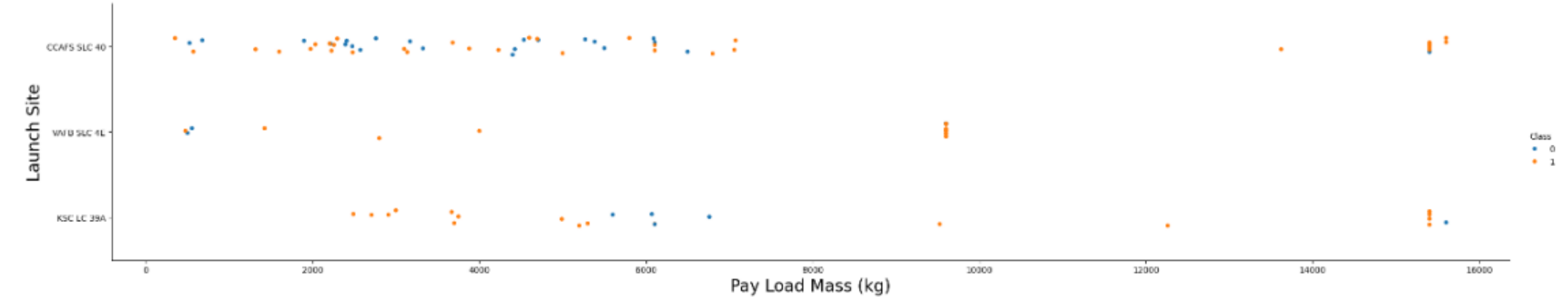


- Show the screenshot of the scatter plot with explanations

- Based on the displayed data, CCAF5 SLC 40 emerges as the top-performing launch site presently, boasting a notable success rate in recent launches.
- Following closely in second position is VAFB SLC 4E, with KSC LC 39A securing the third spot.
- Furthermore, a discernible trend reveals a progressive enhancement in the overall success rate as time has advanced.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

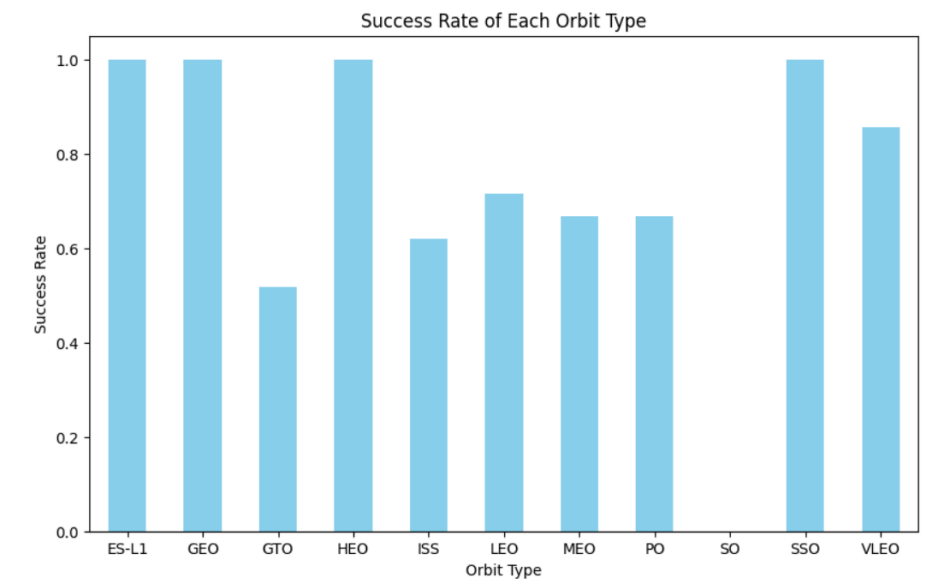


- Show the screenshot of the scatter plot with explanations

  - Payloads exceeding 9,000kg exhibit an impressive success rate.
  - Payloads surpassing 12,000kg appear to be feasible primarily at CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

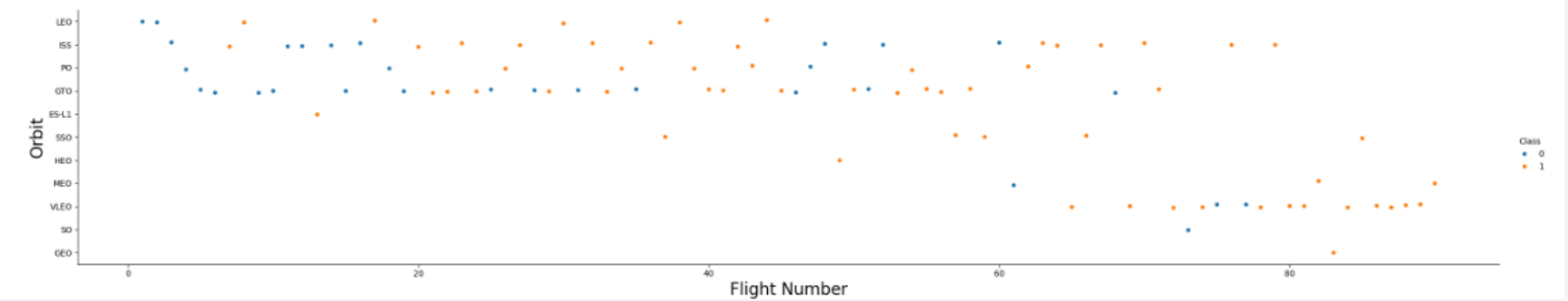


- Show the screenshot of the scatter plot with explanations

- The biggest success rates happens to orbits: ES-L1, GEO, HEO, SSO.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

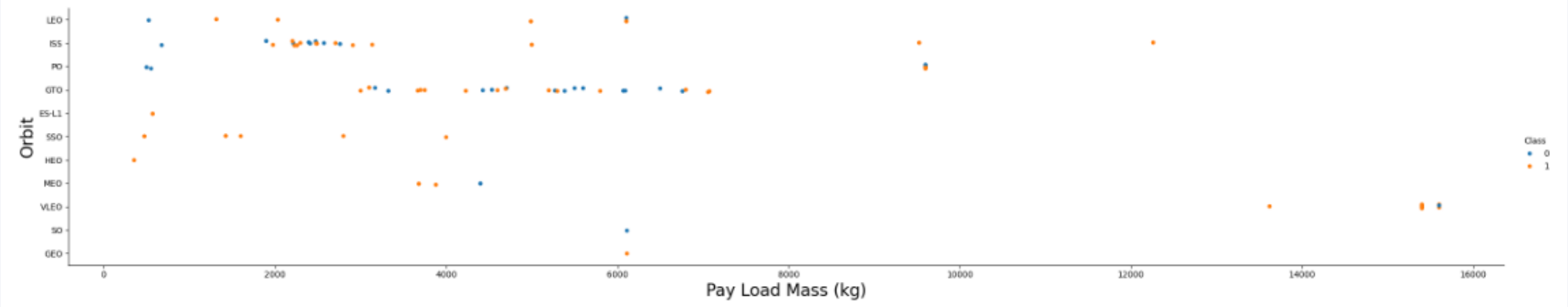


- Show the screenshot of the scatter plot with explanations

- Apparently, there has been an improvement in success rates across all orbits over time.- The Very Low Earth Orbit (VLEO) appears to present a new business opportunity, indicated by its recent increase in frequency.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

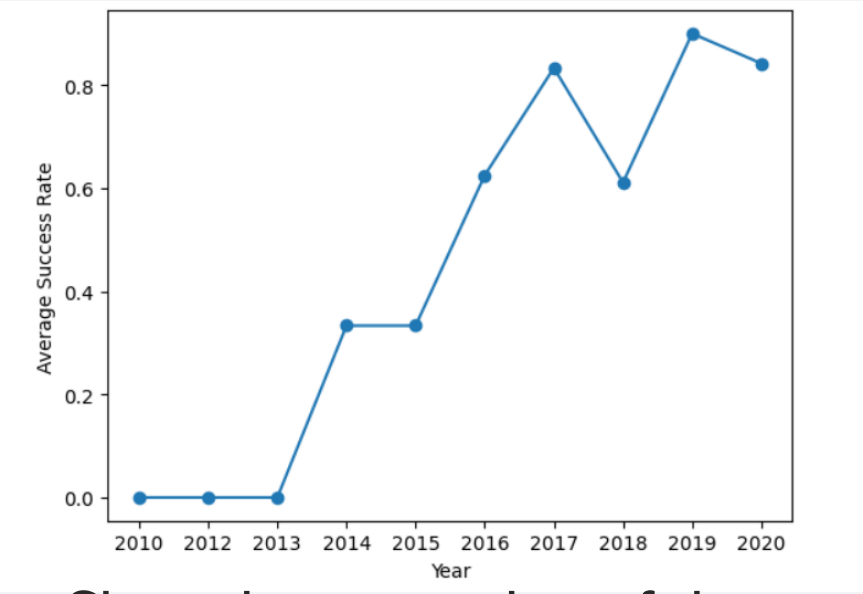


- Show the screenshot of the scatter plot with explanations

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- Show the screenshot of the scatter plot with explanations

- We can observe that the sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

- Find the names of the unique launch sites

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

- Present your query result with a short explanation here

- These are derived by extracting unique instances of "launch_site" values from the dataset.

Launch Site Names Begin with 'CCA'

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Present your query result with a short explanation here

- Here, we have five examples of launches from Cape Canaveral.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total_Payload_Mass
45596

- Present your query result with a short explanation here

- The total payload calculated above is obtained by summing all payloads with codes containing 'CRS', which correspond to NASA.

Average Payload Mass by F9 v1.1

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

Average_Payload_Mass
2928.4

- Present your query result with a short explanation here

- After filtering the data based on the booster version mentioned earlier and calculating the average payload mass, we obtained a value of 2,928 kg.

First Successful Ground Landing Date

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

First_Successful_Landing_Date
2015-12-22

- Present your query result with a short explanation here

- By filtering the data based on a successful landing outcome on the ground pad and obtaining the minimum date value, we can identify the first occurrence, which happened on 12/22/2015.

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

- Present your query result with a short explanation here

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- Present your query result with a short explanation here

- By grouping mission outcomes and tallying records for each group, we arrived at the summary provided above for the successful and failure.

Boosters Carried Maximum Payload

- List the names of the booster which 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

- Present your query result with a short explanation here

- These are the boosters that have carried the maximum payload mass recorded in the dataset.

2015 Launch Records

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

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Present your query result with a short explanation here

- The list above has the only two occurrences.

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

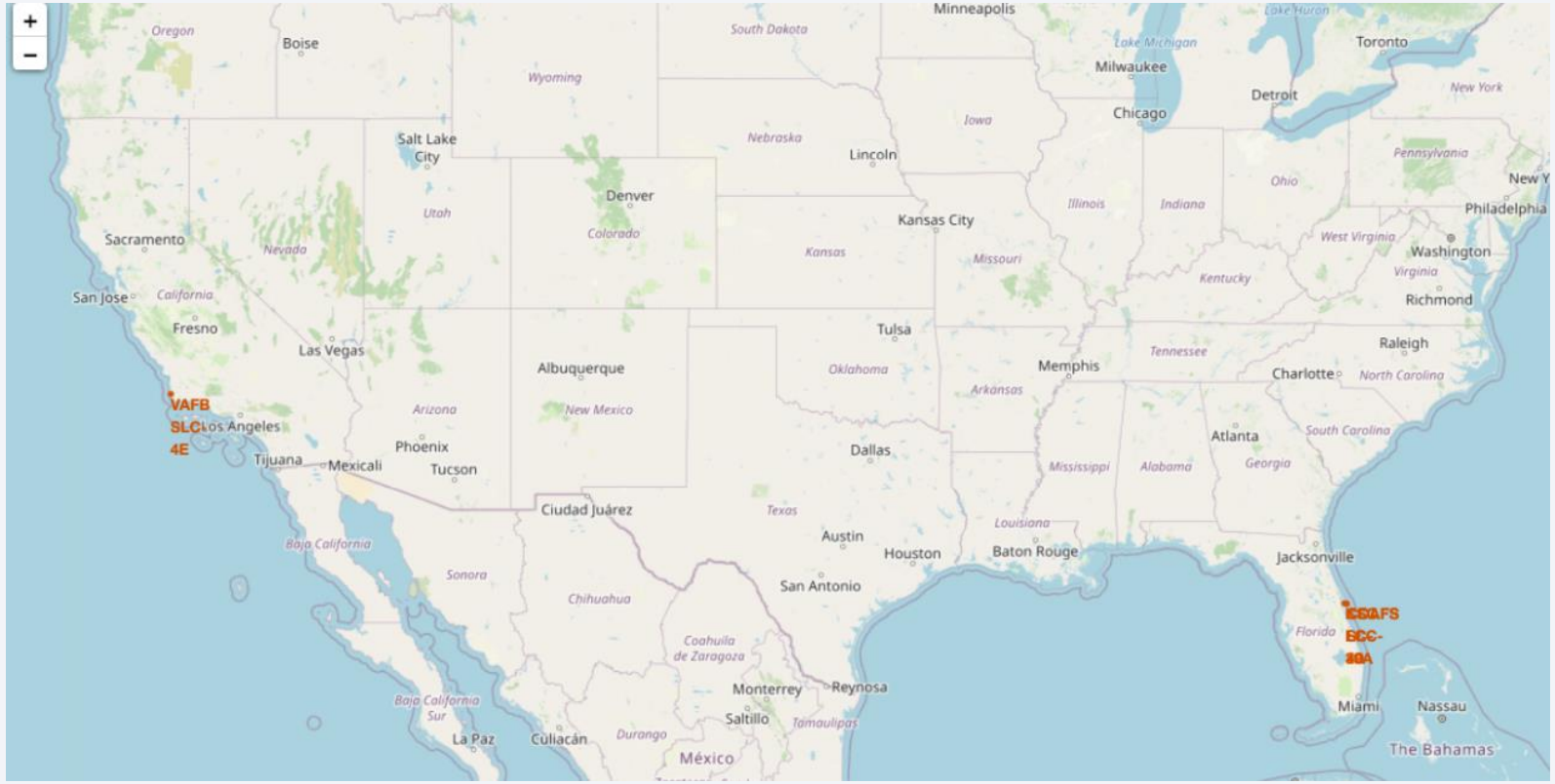
- Present your query result with a short explanation here

- This perspective on the data highlights the significance of factoring in occurrences labeled as 'No attempt'.

Section 3

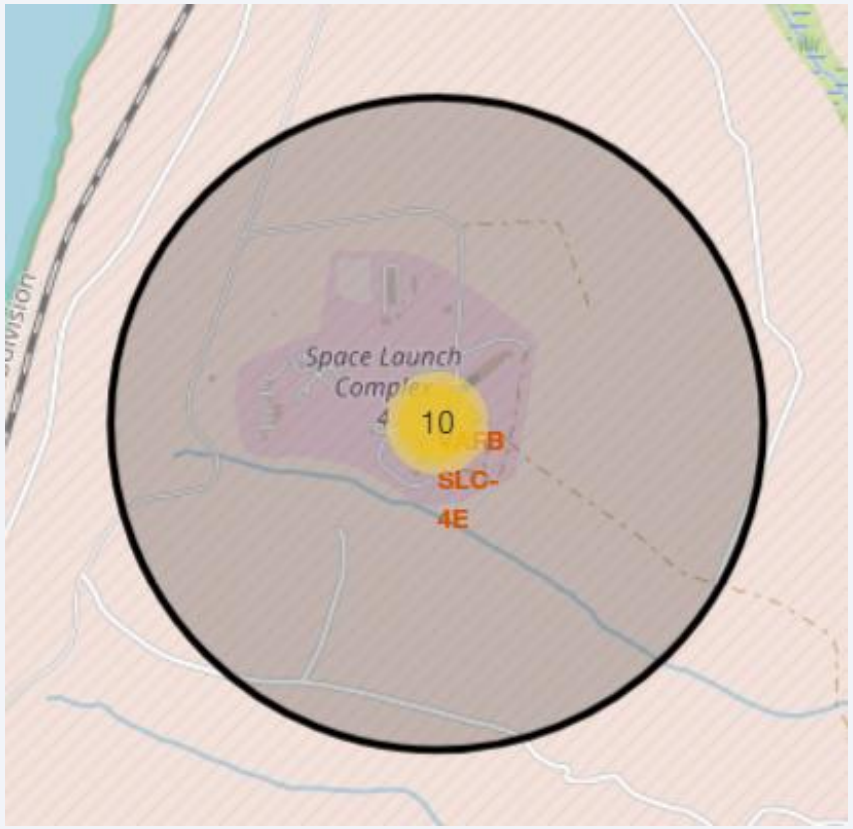
Launch Sites Proximities Analysis

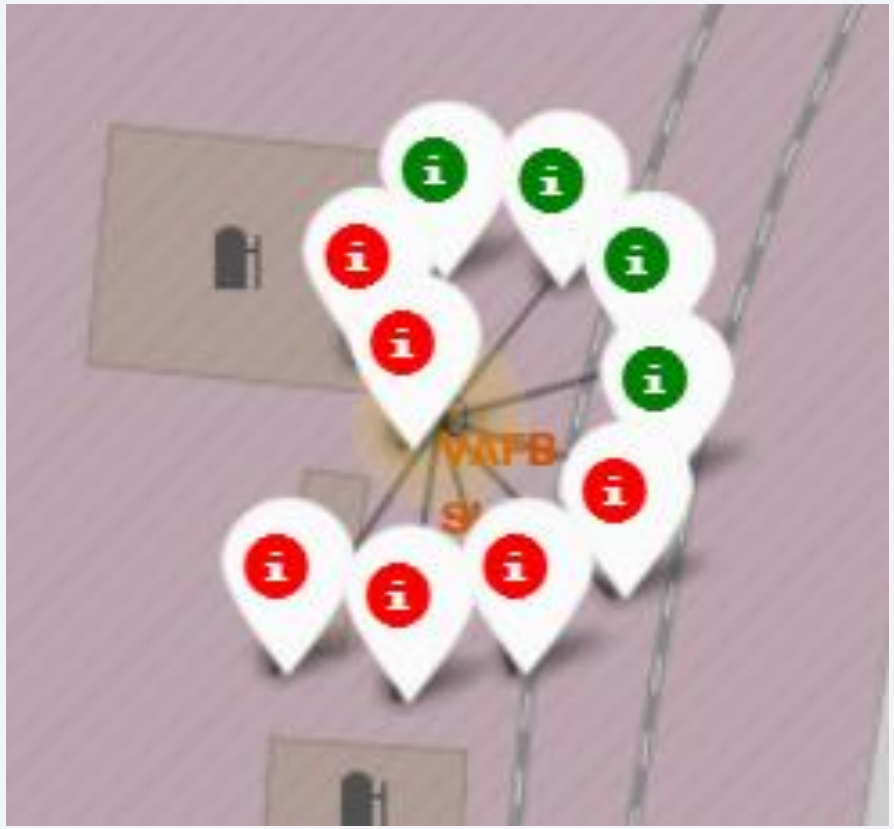
Launch Sites



Launch sites are near sea, probably for security.

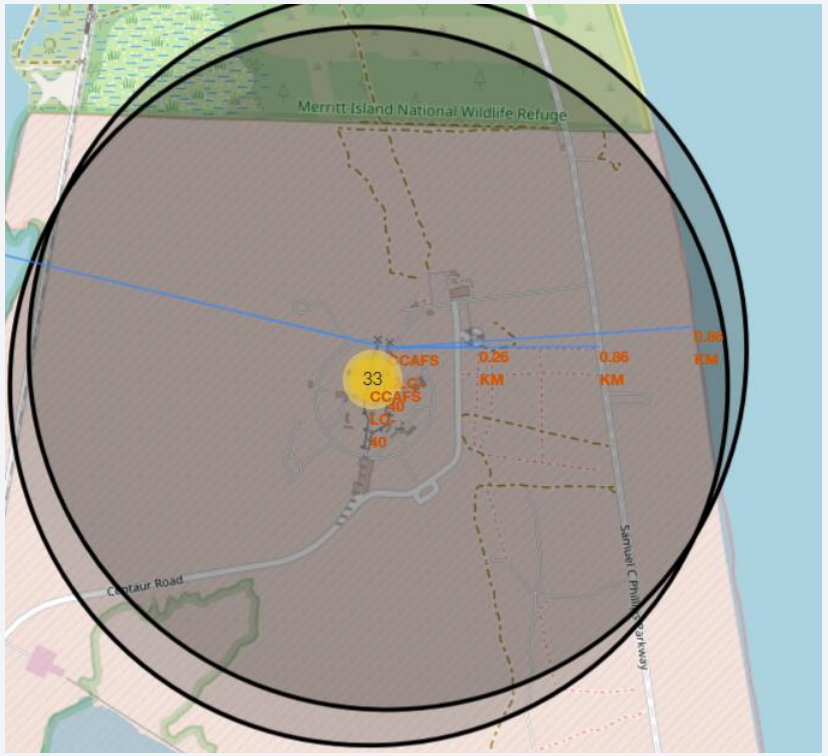
Launch Outcomes



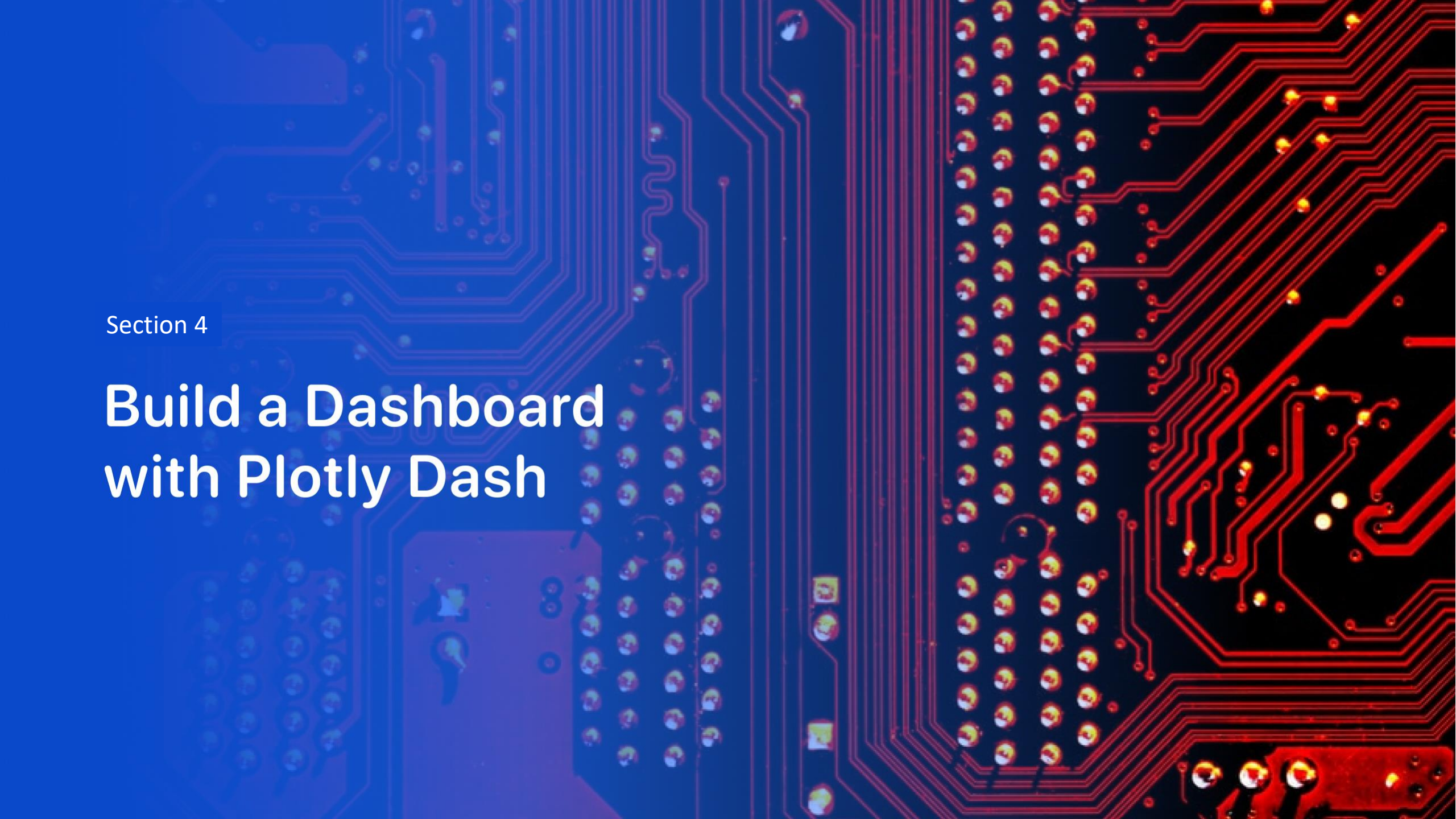


Green markers indicate successful and red ones indicate failure.

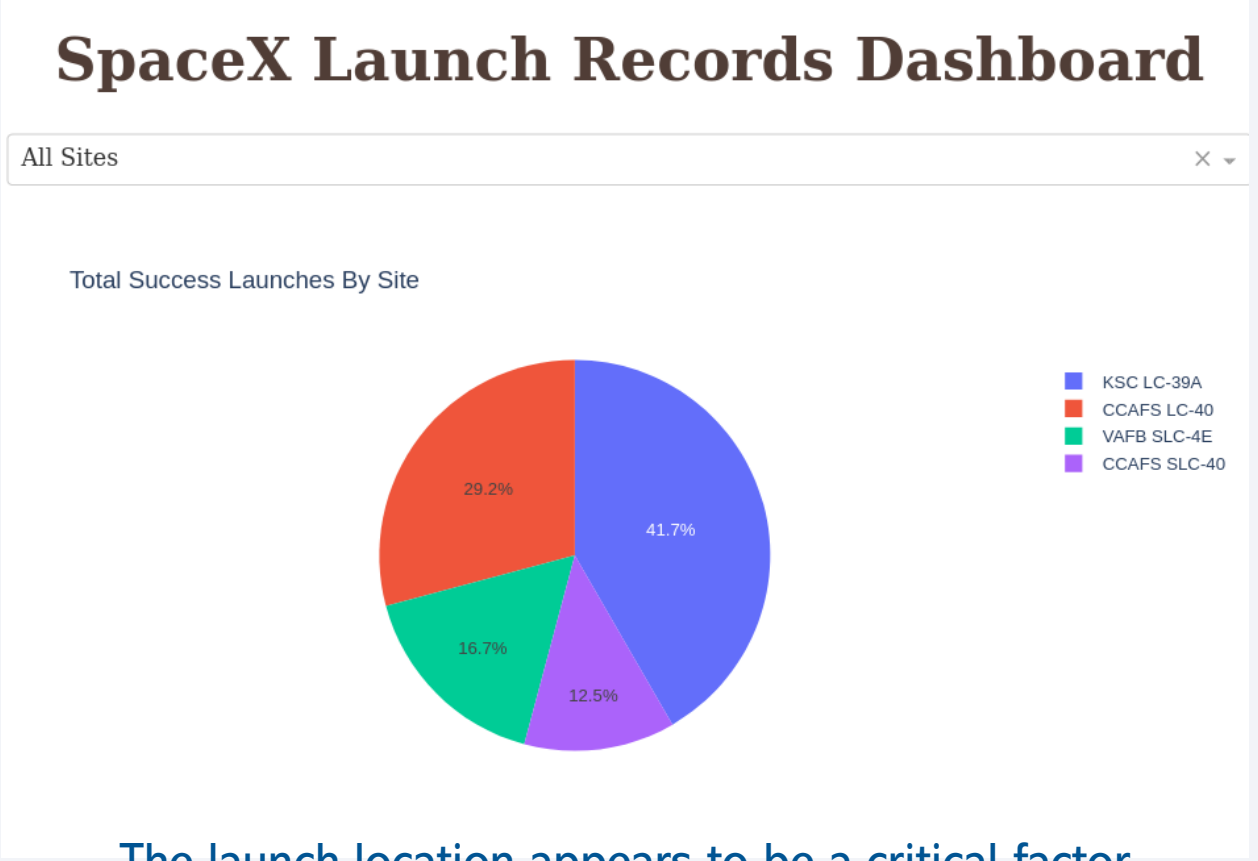
Reasons for ubications



Launch sites are near sea, probably for security. being near railroad and road and relatively far from inhabited areas.

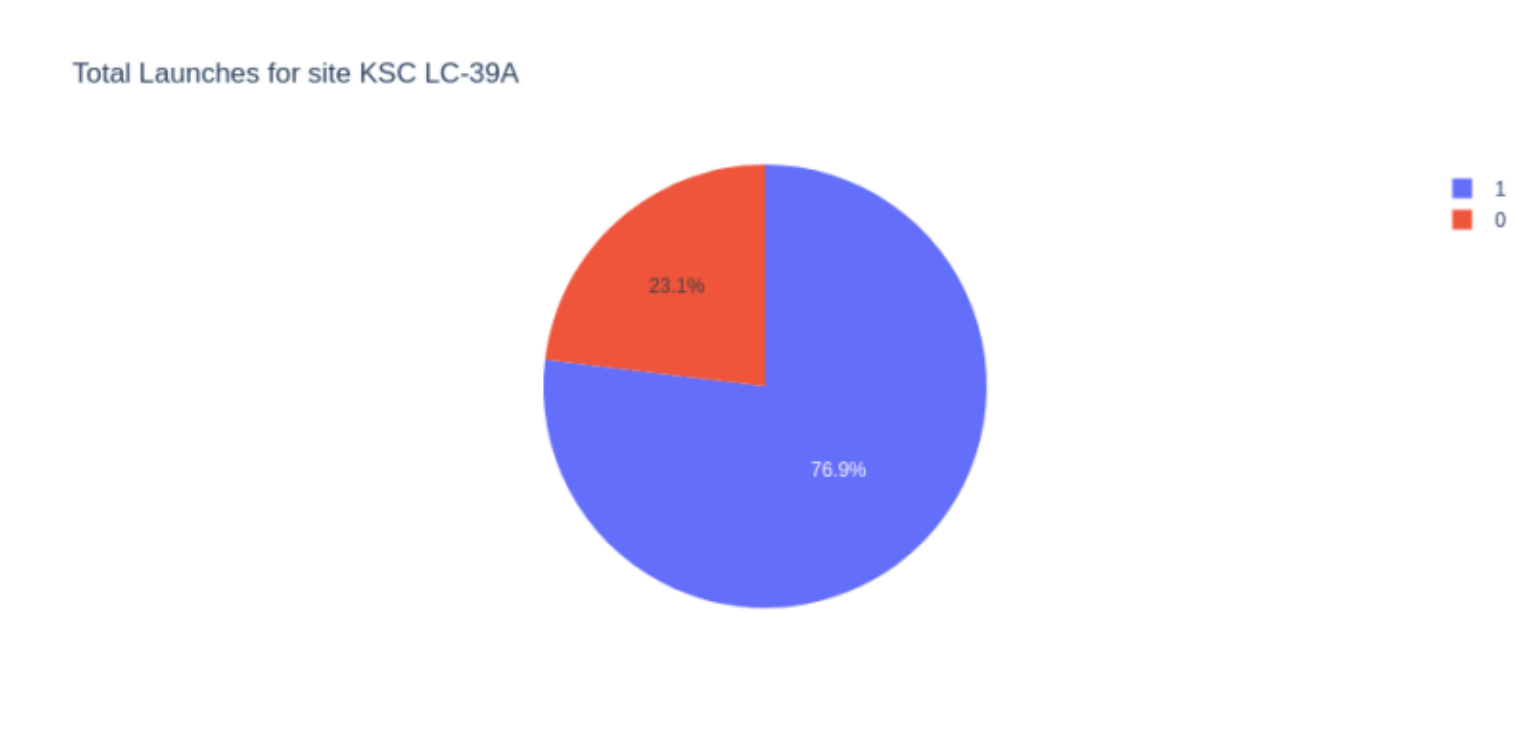

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



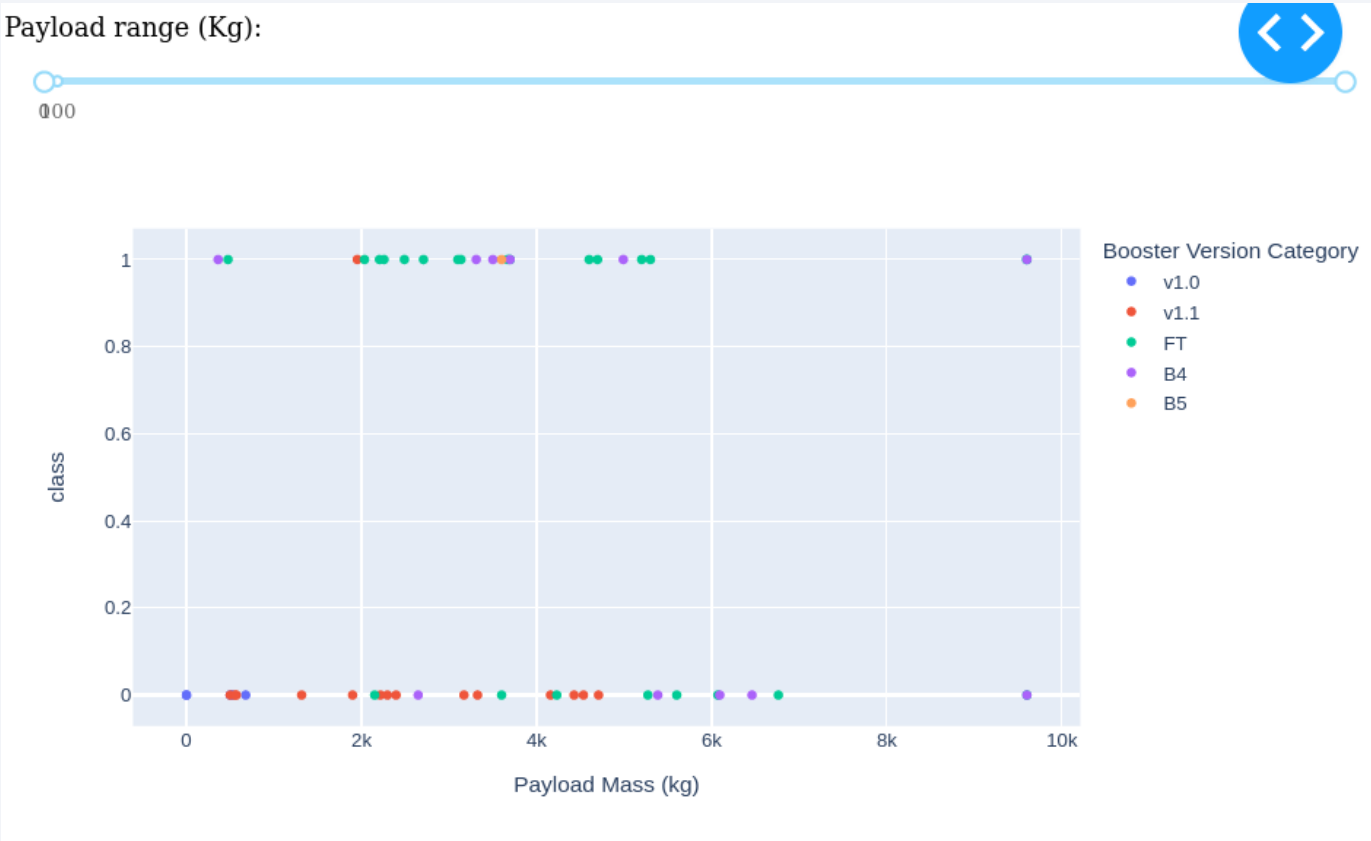
The launch location appears to be a critical factor influencing the success of missions.

Launch Success Ratio for KSC LC-39A



76.9% of launches are successful in this site.

<Dashboard Screenshot 3>



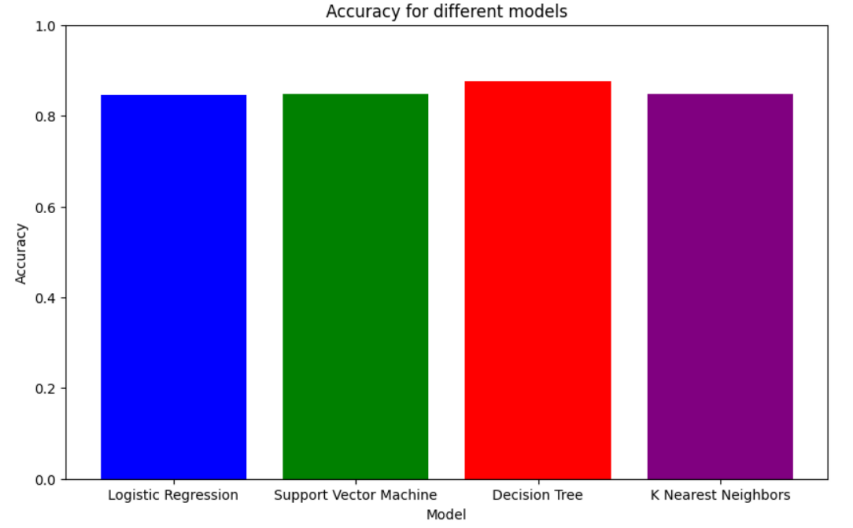
The most successful combination appears to be payloads under 6,000kg paired with FT boosters.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

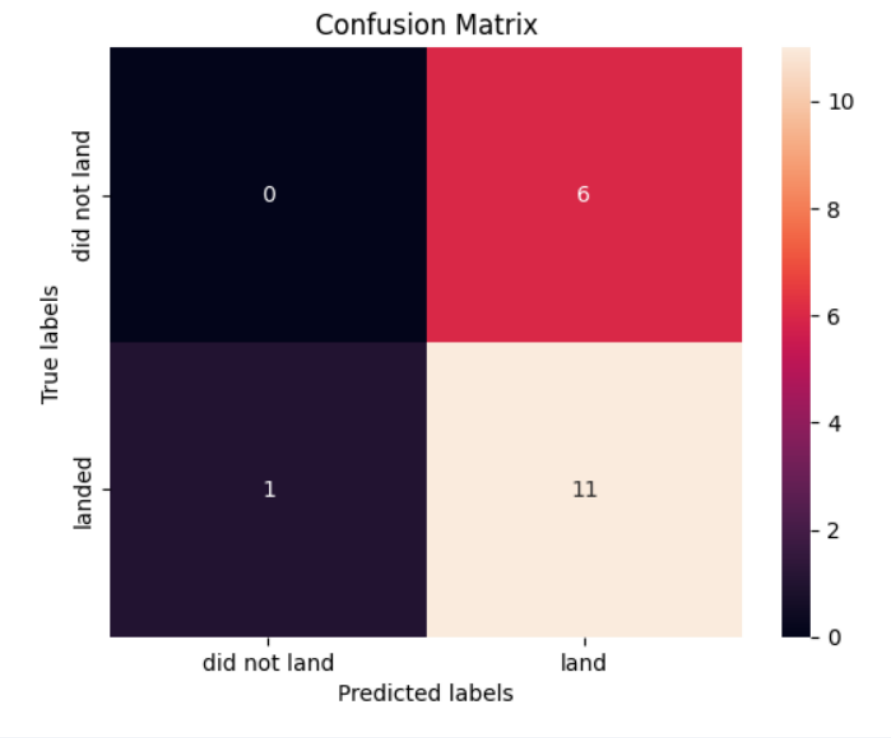


- Find which model has the highest classification accuracy

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having an Accuracy over 87%.

Confusion Matrix of Decision Tree Classifier

- Show the confusion matrix of the best performing model with an explanation



- The model performs relatively well in predicting cases that did land but struggles to accurately predict cases that did not land.

Conclusions

- KSC LC-39A stands out as the optimal launch site.

- Launches exceeding 7,000kg entail lower risks.

- While the majority of missions succeed, the likelihood of successful landings appears to grow with advancements in processes and rocket technology.

- Utilizing a Decision Tree Classifier can forecast successful landings, potentially bolstering profits.

- The most successful combination appears to be payloads under 6,000kg paired with FT boosters.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project.

```
accuracies = {
    "Logistic Regression": logreg_cv.best_score_,
    "Support Vector Machine": svm_cv.best_score_,
    "Decision Tree": tree_cv.best_score_,
    "K Nearest Neighbors": knn_cv.best_score_
}

plt.figure(figsize=(10, 6))
plt.bar(accuracies.keys(), accuracies.values(), color=['blue', 'green', 'red', 'purple'])

plt.title('Accuracy for different models')
plt.xlabel('Model')
plt.ylabel('Accuracy')
plt.ylim(0, 1)  # Establecer el rango del eje y de 0 a 1
plt.xticks(rotation=360)
plt.show()
```

Thank you!