SPACE X
FALCON 9 LANDING ANALYSIS

AUTHOR : YX GOH

IBM DATA SCIENCE CAPSTONE PROJECT

TABLE OF CONTENT

EXECUTIVE SUMMARY

Methodologies:

•Utilized API integration and web scraping for data collection.

•Processed and cleansed data using Python and SQL.

•Analyzed trends and applied machine learning to predict landing event outcomes.

Results:

•Identified key patterns influencing landing success.

•Developed a model with 83% accuracy in predicting successful landings, demonstrating effective forecasting ability.

INTRODUCTION

SpaceX, a leader in commercial space travel, offers Falcon 9 rocket launches at $62 million each, significantly cheaper than competitors' $165 million. This cost efficiency stems from the reusability of the rocket's first stage.

Objective:

Predict whether SpaceX will reuse the first stage of a rocket based on variables such as payload mass, launch site, and flight frequency, using machine learning.

Key Questions:

1. What factors affect the success of the first stage landing?
2. Has the success rate of landings improved over time?
3. Which binary classification algorithm is most effective for this prediction?

METHODOLOGY

Data Collection
- Using SpaceX Rest API
- Using Web Scraping from Wikipedia

Data Wrangling
- Filtering the data
- Dealing with missing values
- Using one-hot encoding to prepare the data to a binary classification

Exploratory Data Analysis (EDA)
- Manipulate dataset using SQL queries
- Visualize relationships between variables of dataset and identify pattern from the dataset

Interactive Visual Analysis
- Geospatial analysis using Folium
- Interactive dashboard using Plotly Dash

Data Modelling and Evaluation
- Predictive analysis using classification models

Data Collection

Overview

Data was gathered through API requests and web scraping. The SpaceX REST API provided detailed information on Falcon 9 launches, including flight number, date, booster version, payload mass, orbit type, launch site, and landing outcome. Additional launch details were extracted from SpaceX's Wikipedia page, such as payload, customer, and launch outcomes. This involved requesting data, parsing JSON responses, and organizing HTML content into structured DataFrames. Both data sources were merged, cleaned, and exported as CSV files for comprehensive analysis.

Process Steps

- API Data Handling:

1. Define parsing function
2. Retrieve data from API
3. Parse the data
4. Store the data

- Web Scraping:

1. Get URL
2. Create ‘soup’ object
3. Filter tables
4. Transform table into DataFrames

Data Wrangling

There are various landing outcomes in the dataset indicating success or failure:

- **True RTLS:** Successful landing on a ground pad.
- **False RTLS:** Unsuccessful landing on a ground pad.
- **True Ocean:** Successfully landed in a designated ocean region.
- **False Ocean:** Unsuccessful landing in a specific ocean region.
- **True ASDS:** Successful landing on a drone ship.
- **False ASDS:** Unsuccessful landing on a drone ship.

Data Handling Steps:

1. **Exploratory Data Analysis:** Identify variables and determine training labels for the model.
2. **Data Segmentation:** Calculate the number of launches per site, occurrences per orbit, and mission outcomes per orbit type.
3. **Outcome Labeling:** Convert mission outcomes into binary labels ('1' for successful, '0' for unsuccessful).
4. **Data Export:** Finalize the dataset and export it to CSV for further analysis.

Exploratory data analysis (EDA)

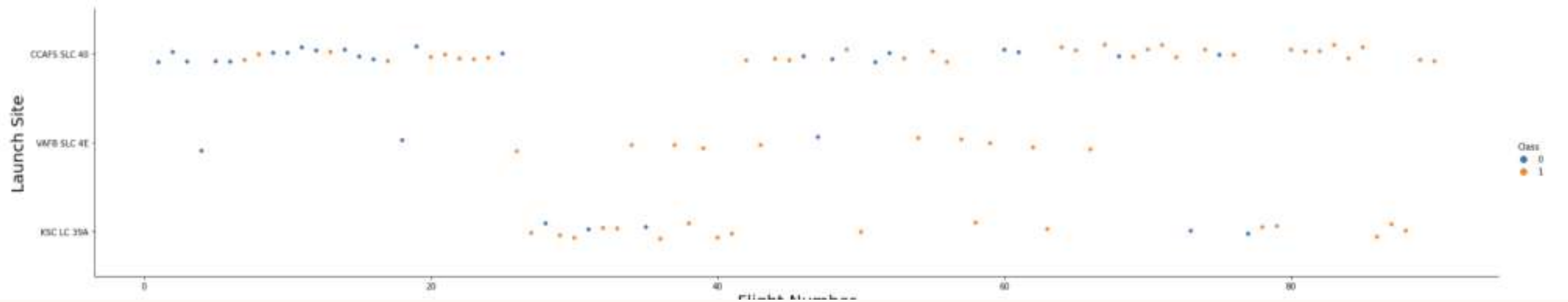
SQL

Using SQL, we had performed many queries to get better understanding of the dataset.

- Displaying the names of the launch sites.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch sites names for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and2017-03- descending order.

Data Visualization

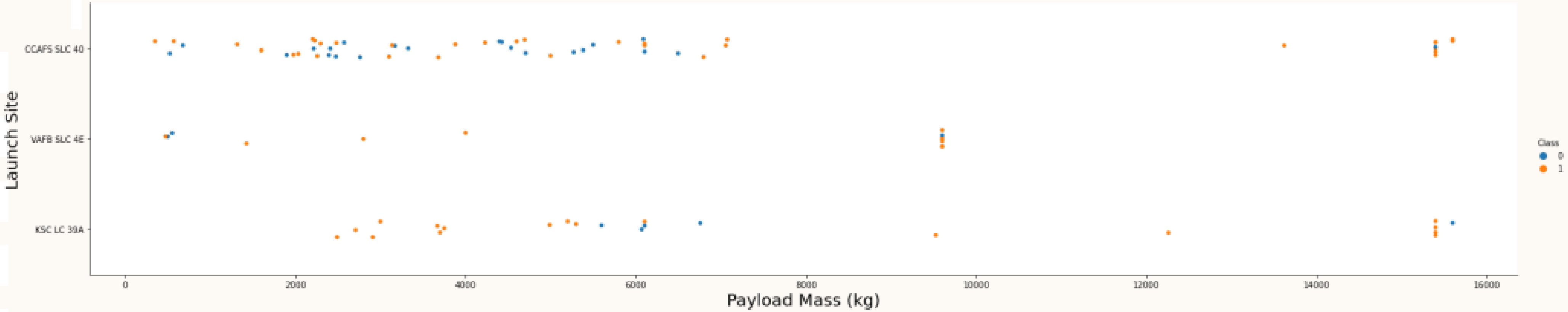
1) The relationship between Flight Number and Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Data Visualization

2) The relationship between Payload and Launch Site

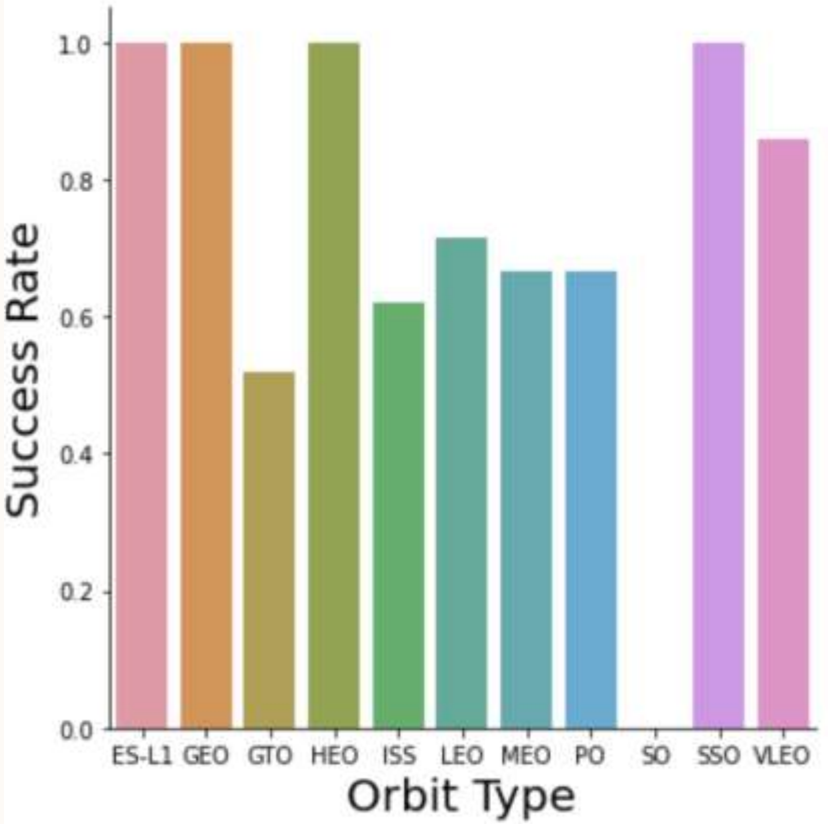


- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successfull.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Data Visualization

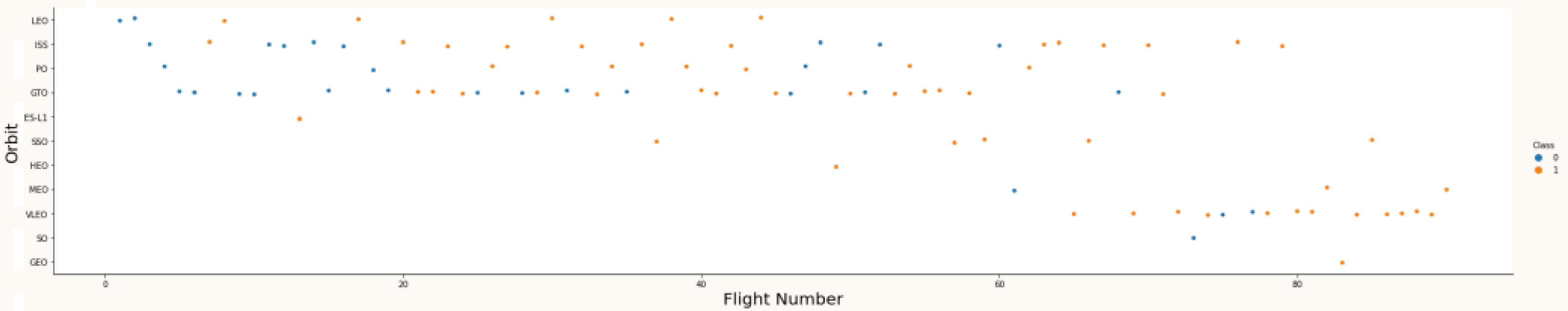
3) The relationship between Success Rate and Orbit Type

- Orbits with 100% success rate are ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate are: SO
- Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO



Data Visualization

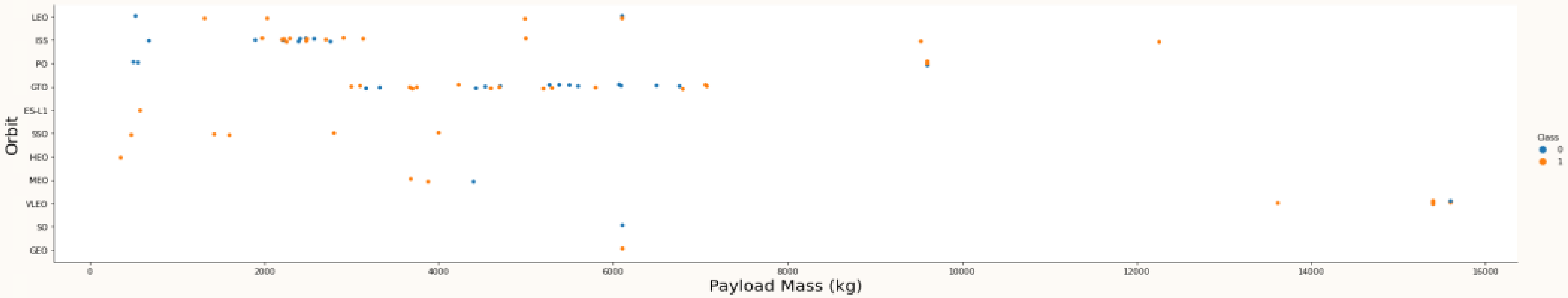
4) The relationship between Flight Number and Orbit Type



- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Data Visualization

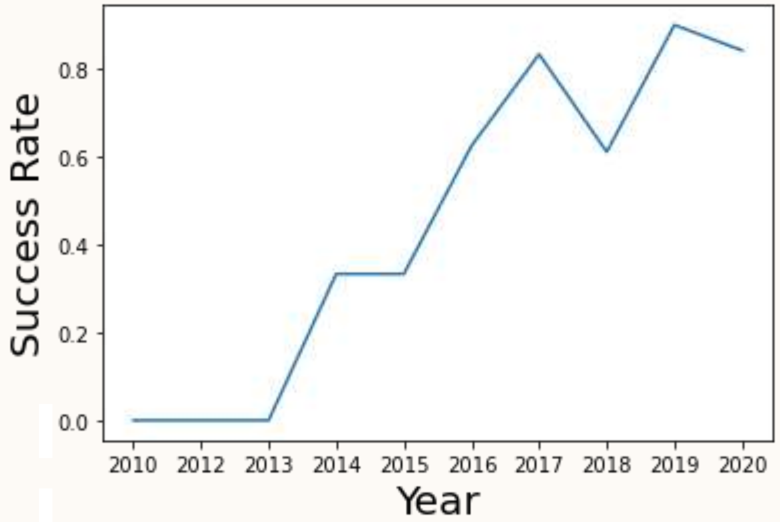
5) The relationship between Payload and Orbit Type



- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Data Visualization

6) The launch success yearly trend



- the sucess rate since 2013 kept increasing till 2020

Interactive Visual Analysis

Geospatial Analysis Using Folium

All launch sites global map markers

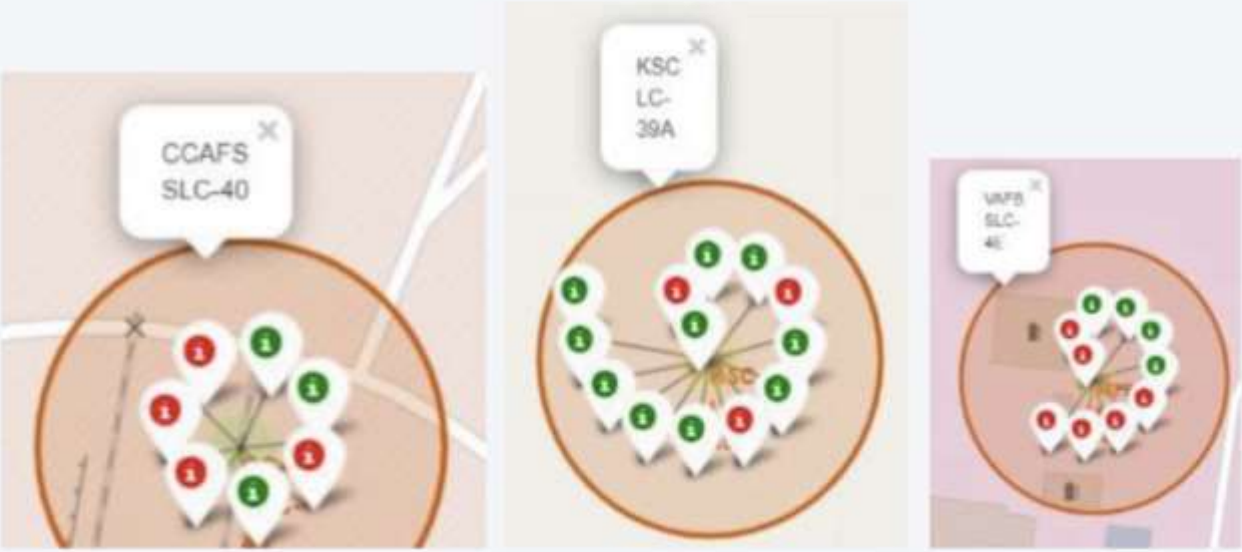
We can see that the Space launch sites are in the United States of America coasts. Florida and California

Geospatial Analysis Using Folium

Markers showing launch sites with color labels

Green marker showing **Successful** Launches.

Red marker showing **Unsuccessful** Launches.



Geospatial Analysis Using Folium

Launch Site distance to landmarks

Are lunch sites is close proximity to railways?

No

Are lunch sites is close proximity to highway?

No

Are lunch sites is close proximity to coastline?

Yes

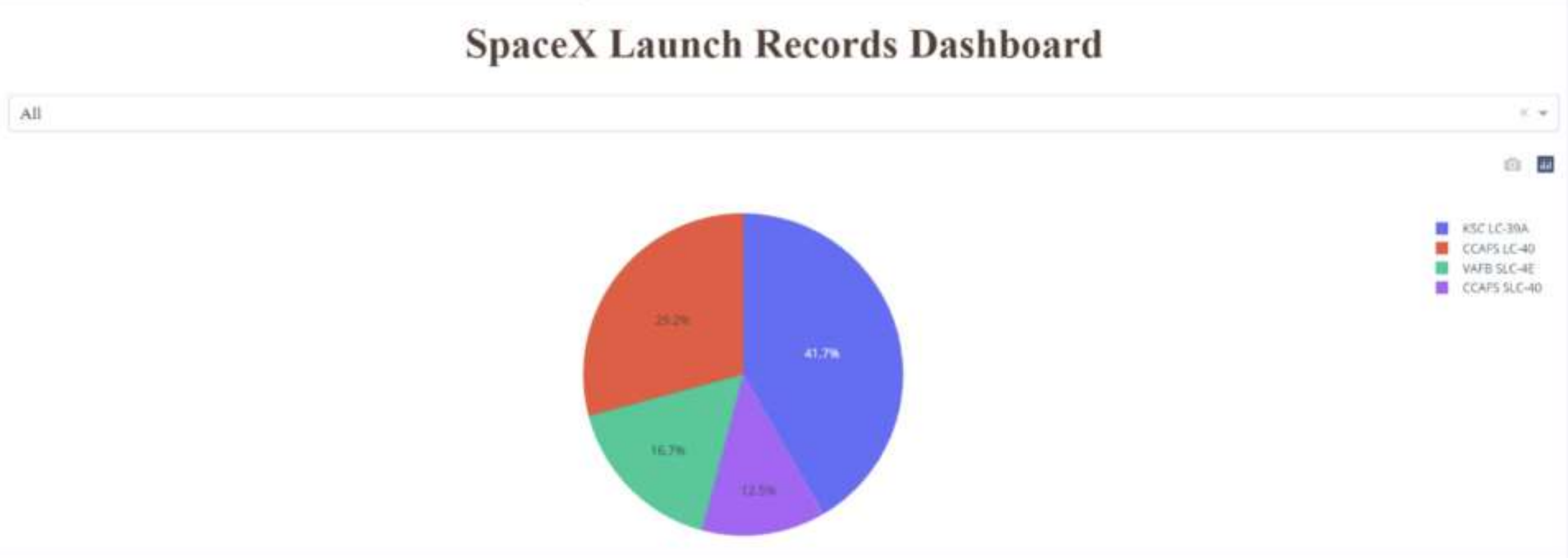
Do launch sites keep certain distance away from cities?

Yes

Interactive dashboard using Plotly Dash

Pie chart showing the success percentage achieved by each launch site

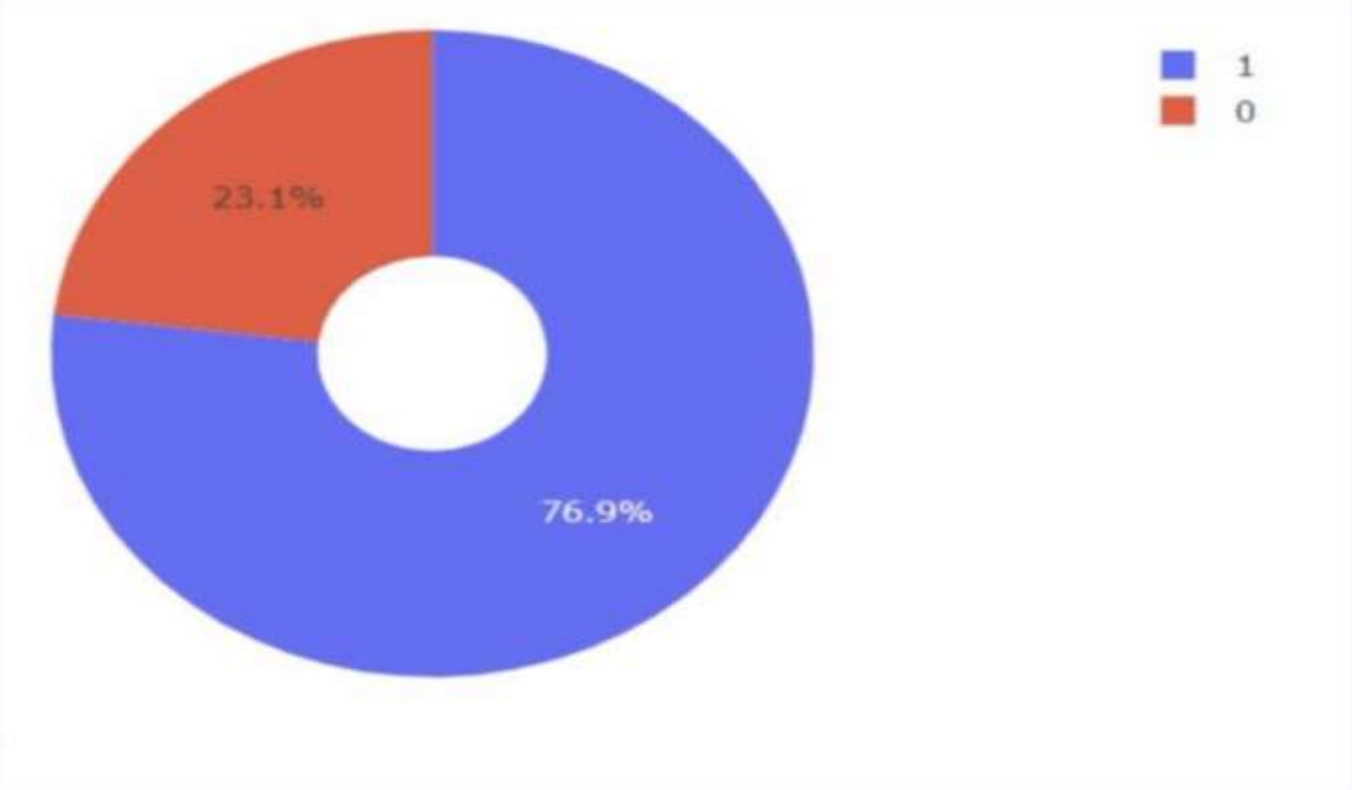
Total success launches by all sites



We can see that KSC LC-39A had the most successful launches from all the sities.

Interactive dashboard using Plotly Dash

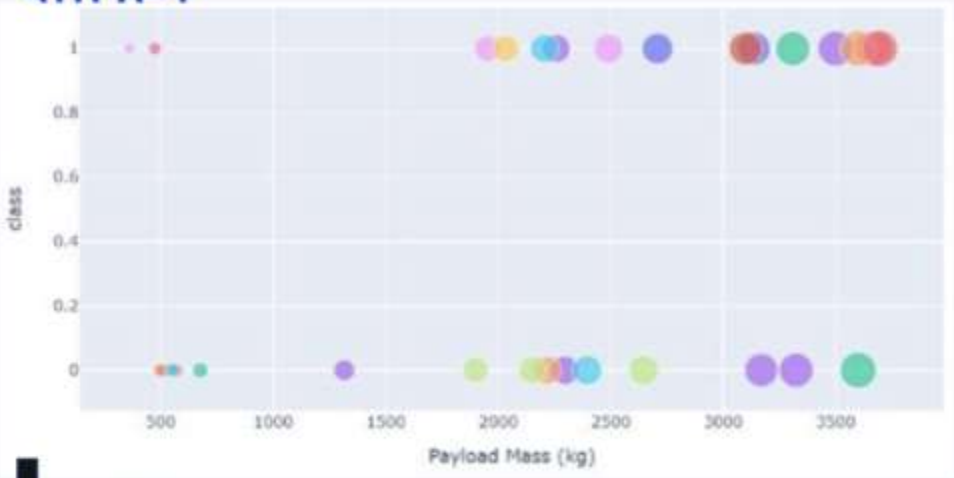
Pie chart showing the Launch site with the highest launch success ratio



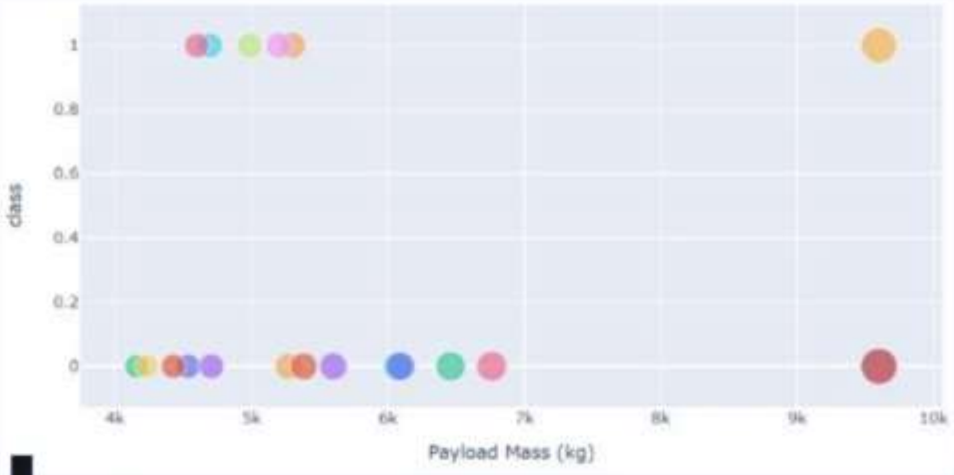
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Interactive dashboard using Plotly Dash

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



Low Weighted Payload 0kg - 4000kg



Heavy Weighted Payload 4000kg - 10000kg

We can see the success rates for low weighted payloads is higher than the heavy weighted payloads.

Classification Accuracy

The small test sample size (18 samples) led to similar performance across models in the test set, with all models achieving the same Jaccard score, F1 score, and accuracy.

To resolve the uncertainty due to the small sample size, the models were evaluated using the entire dataset.

Decision Tree Model stood out with the highest scores when evaluated across the entire dataset.

Scores and Accuracy of the Test Dataset

Metric	LogReg	SVM	Tree	KNN
Jaccard Score	0.800000	0.800000	0.800000	0.800000
F1 Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire Dataset

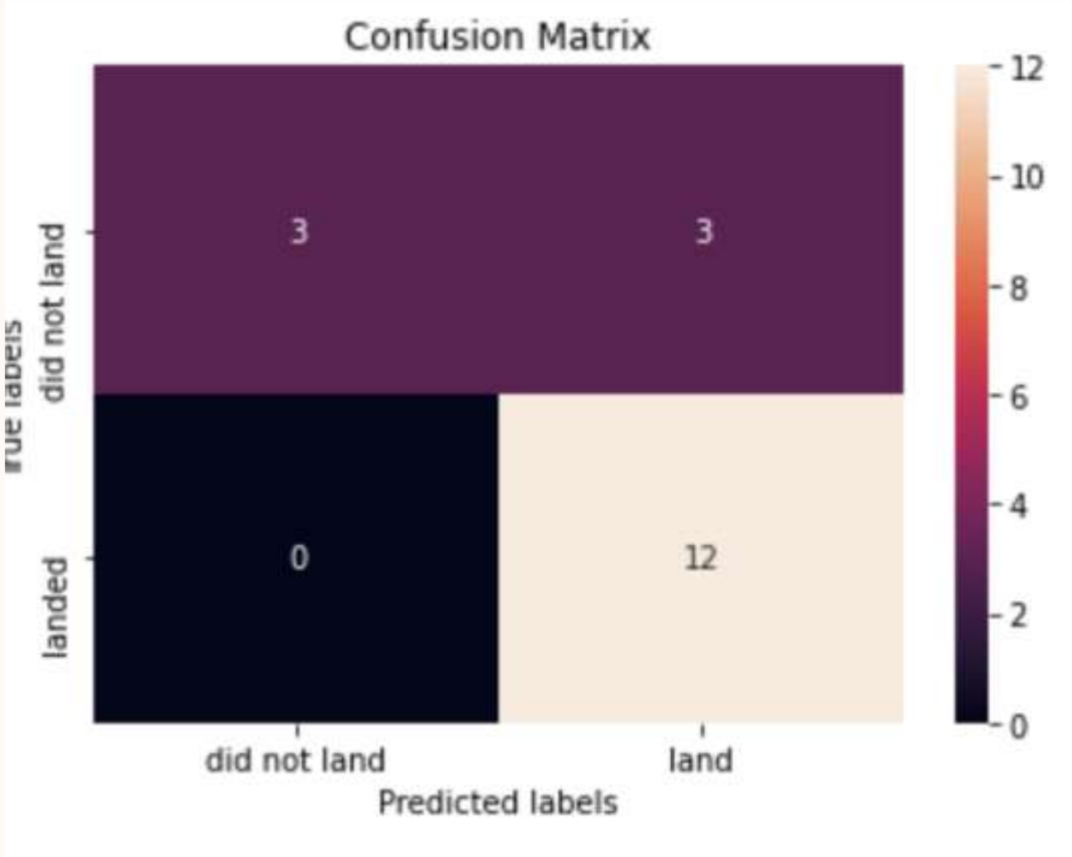
Metric	LogReg	SVM	Tree	KNN
Jaccard Score	0.833333	0.845070	0.882353	0.819444
F1 Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Classification Accuracy

Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.

The major problem is the false positive, i.e. unsuccessful landing marked as successful landing by the classifier.



Predictive Analysis (Classification)

Conclusions

Launch Success Trends

•Success rates have increased at launch sites as flight numbers grow, particularly notable from 2013 onwards.

•KSC LC-39A is the most successful launch site, with a success rate of 76.9%.

Orbit Success Analysis

•Orbits ES-L1, GEO, HEO, and SSO achieved the highest success rates at 100%, though it's important to note that some of these have fewer flights.

•The success rate for SSO is notable across five successful flights.

•Heavier payloads are more commonly associated with VLEO launches, which show higher success rates.

Payload Insights

•Larger payloads tend to have lower success rates compared to smaller ones.

Machine Learning Insights

•The Decision Tree model proved most effective for this analysis, achieving an accuracy of 94.44%.

Appendix

Credit and Acknowledgments:

Coursera – Applied Data Science Capstone