IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joy Furbush
1/9/2024

Outline

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

Executive Summary

<u>Methodology</u>

- Data Collection included the following methodologies:
    - Web Scraping
    - SpaceX API
    - Exploratory Data Analysis (also called EDA)
    - Machine Learning Predictions
- Data collections using these methodologies involved visual analytics, data wrangling, and visual implementations to display the data in a more comprehensive manner
- As a result, these methods allowed us to do the following:
- Import and glean information from public sources like Wikipedia
- Predict failures and success with all SpaceX launches and potential site issues
- Machine learning prediction allowed better understanding of the data as a whole, predicting areas of success or improvement

Introduction

- Project Question
  - Can the company SpaceY compete with SpaceX?
- Objectives for comparison
  - Total Cost/affordability
    - Space X already has best pricing ($62 million vs. competitors $165 million)
    - Successful launches
  - Will SpaceX reuse the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Two public sources were used for the API and web scraping:
    - API: https://api.spacexdata.com/v4/rockets/
    - Webscraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Featured analysis generated landing outcome labels, enhancing the overall data value

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Per the performed lab, EDA allows attributes to determine if the first stage can be reused. From there, we can use these features with machine learning to automatically predict if the first stage can land successfully
- Perform interactive visual analytics using Folium and Plotly Dash
  - Per the performed lab, both Folium and Plotly Dash are interactive ways to manipulate, explore, and visualize the analytics in real time, such as launch sites/proximities

Methodology

Executive Summary

- Perform predictive analysis using classification models
    - After normalization, the data was divided up for training and testing, with models and multiple parameter sets evaluated for accuracy

Data Collection

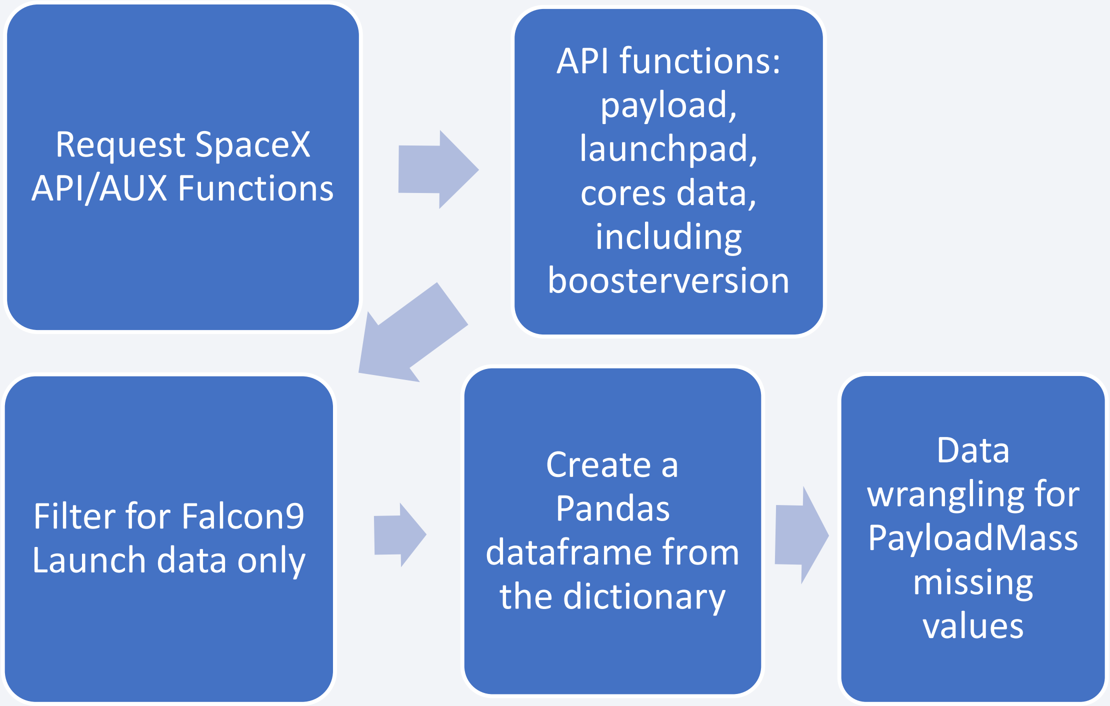
Datasets were collected from two public sources:

- SpaceX API: https://api.spacexdata.com/v4/rockets/
- Wikipedia article: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- SpaceX API

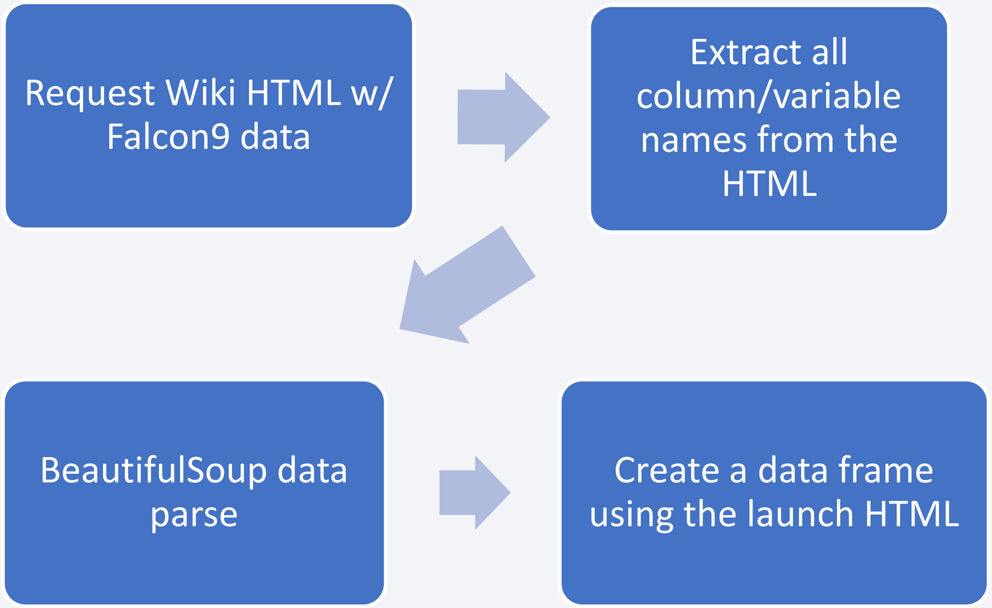
- GitHub URL: https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/71579a93b1df6b941b0a6c8ab5deca7ecba68b31/Data%20Collection%20API.ipynb



Data Collection - Scraping

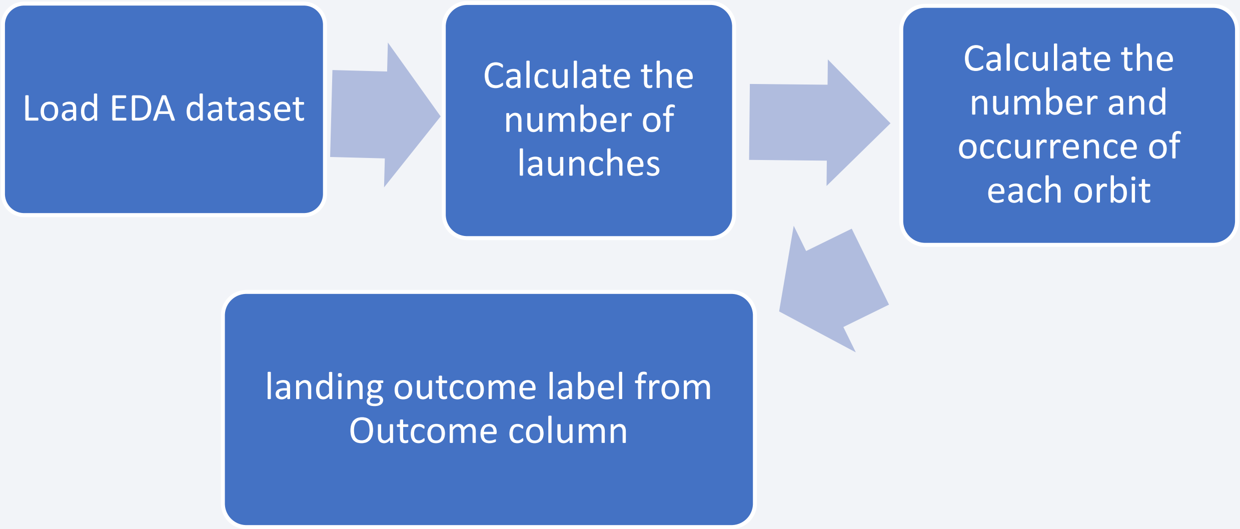
- Data Webscraping (collected from Wiki page)

- GitHub URL: https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/71579a93b1df6b941b0a6c8ab5deca7ecba68b31/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- EDA was performed on the dataset, then calculated summaries of launches per site, orbit occurrences, and mission outcome per orbit type. Lastly, a landing outcome label was created from the 'Outcome' column, which includes both mission outcome and landing location.
- GitHub URL: https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/9def8d4183e6065c61aee4fe24e7b2eaa536cc60/SpaceXData%20wrangling.ipynb



EDA with Data Visualization

- EDA explored Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year.
- Scatter plots, line charts, and bar charts compared variables to identify relationships for machine learning
- GitHub URL for both notebook/dataset https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/59aa3f7d3db51bbea57b001c6ecc355c11d4da6d/EDAVisualization.ipynb

and https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/f646157ab430ee04fc42e2e5c88b05377f95448d/EDAVisualizationDataset.csv

EDA with SQL

- EDA using SQL queries as follows:
  - Dataset loaded to IBM DB2
  - Top 5 launch sites starting with "CCA"
  - Payload mass analysis: total for NASA (CRS) boosters; average for booster F9 v1.1
  - Landing outcomes:
    - Date of first successful ground pad landing
    - Successful drone ship landings (4000-6000 kg payload): boosters, launch sites
    - Failed drone ship landings (2015): booster versions, launch sites
    - Ranking of successful/failed counts (2010-06-04 to 2017-03-20)
- GitHub URL: https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/f646157ab430ee04fc42e2e5c88b05377f95448d/EDAwithSQL.ipynb

Build an Interactive Map with Folium

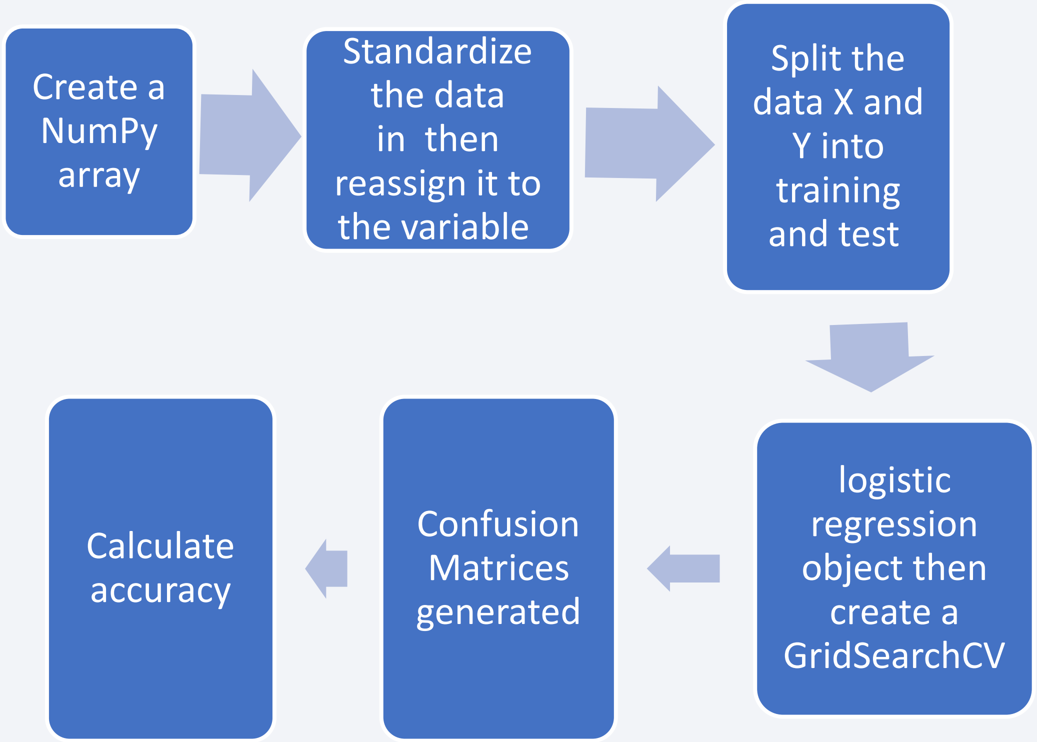
- Folium maps pinpoint launch sites, successful landings, and key locations like railways, highways, coasts, and cities, revealing why launch sites are where they are and how successful landings are geographically connected. Markers (points), circles (highlights), lines (distances), and clusters (grouped events) paint a clear picture on the map
- GitHub URL: https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/f646157ab430ee04fc42e2e5c88b05377f95448d/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Plotly Dash is an interactive dashboard to visualize data, including launch site distribution (pie chart) and launch success rates by site and payload mass (scatter plot)
- This helps analyze launch site distribution and success rates/best launch sites by payload mass. The pie chart lets you explore overall and individual site success, while the scatter plot reveals how success varies across factors
- GitHub URL:https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/f646157ab430ee04fc42e2e5c88b05377f95448d/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models:
  - logistic regression
  - support vector machine
  - decision tree
  - K-nearest neighbors
  - GitHub URL: https://github.com/Bleufly/IBM-Applied-Data-Science-Capstone/blob/f646157ab430ee04fc42e2e5c88b05377f95448d/SpaceX_Machine_Learning_Prediction.ipynb



Results

- Exploratory data analysis results
- SpaceX launches from 4 sites, with early missions for itself and NASA.
- The first successful landing occurred in 2015, and many Falcon 9 versions have since landed with above-average payloads.
- Nearly all missions are successful, with only two booster versions failing in 2015 (F9 v1.1 B1012 and B1015). Landing success has steadily improved over time

Results

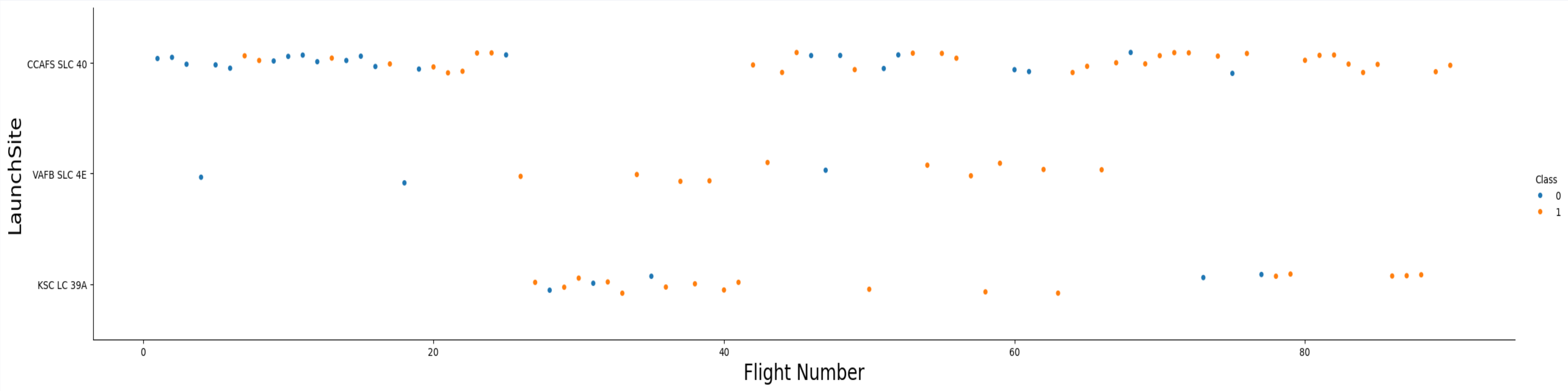
- Interactive analytics revealed that historical launch sites prioritized safety through proximity to the sea, with a bias towards the east coast

Section 2

Insights drawn from EDA

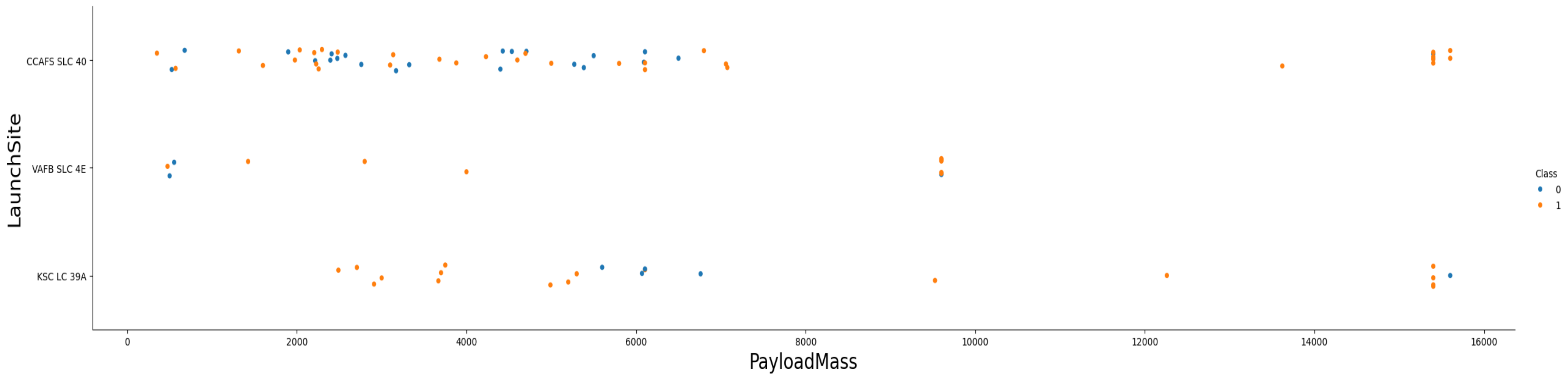
Flight Number vs. Launch Site

- Comparing recent launches, CCAF5 SLC 40 emerges as the clear leader, followed by VAFB SLC 4E and KSC LC 39A. Notably, the overall success rate has also been on the rise



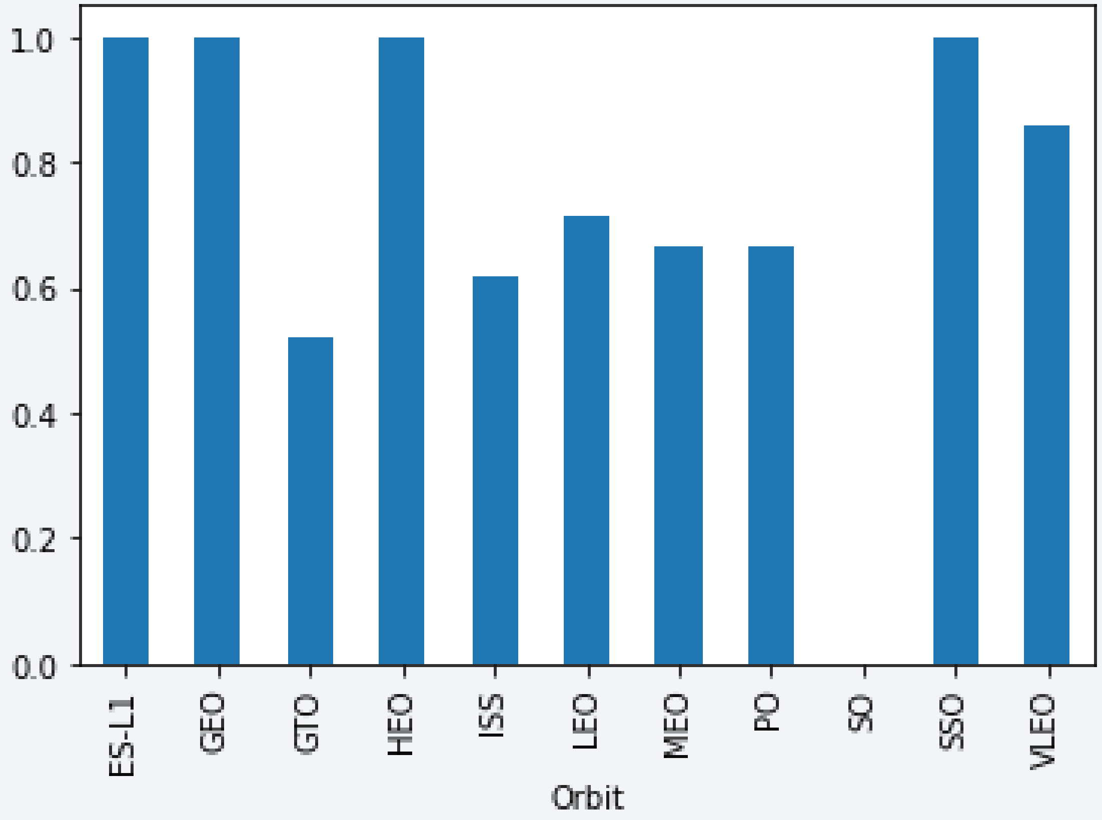
Payload vs. Launch Site

- Heavier payloads (over 9,000kg) boast excellent success rates, while those exceeding 12,000kg appear feasible only at CCAFS SLC 40 and KSC LC 39A launch sites



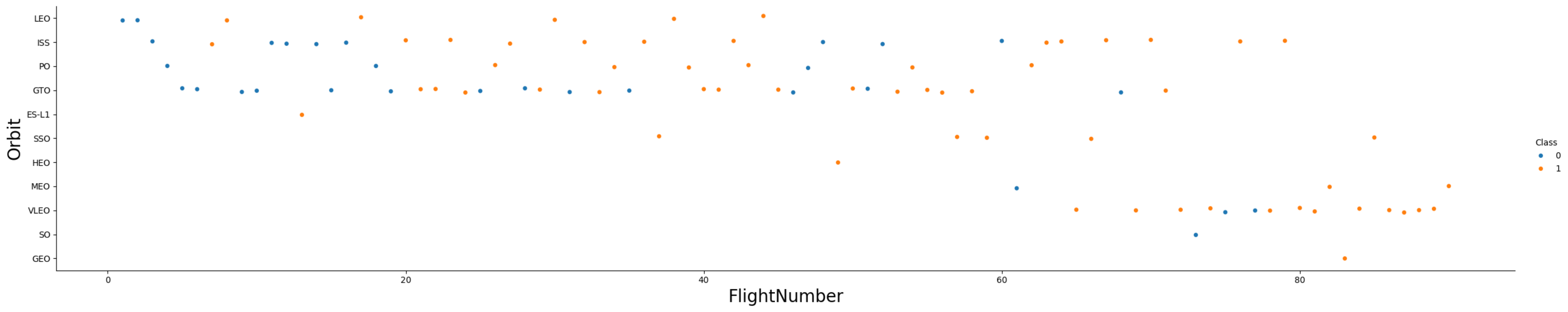
Success Rate vs. Orbit Type

- Orbits like ES-L1, GEO, HEO, and SSO deliver the best success rates, followed by VLEO (above 80%) and LFO (above 70%)



Flight Number vs. Orbit Type

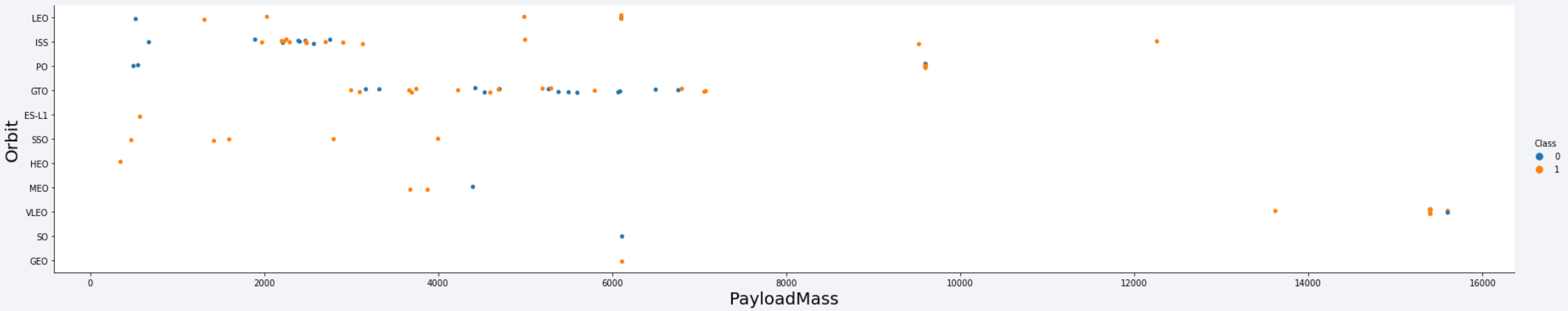
- SpaceX's Launch Orbit preferences and Outcomes exhibited a dynamic relationship across flight numbers.
- Initial focus on LEO yielded moderate success, followed by a shift to VLEO in recent launches. This may indicate higher performance in lower or Sun-synchronous orbits. However, success rates have improved over time for all orbits, with VLEO's increased frequency suggesting possible market expansion



Payload vs. Orbit Type

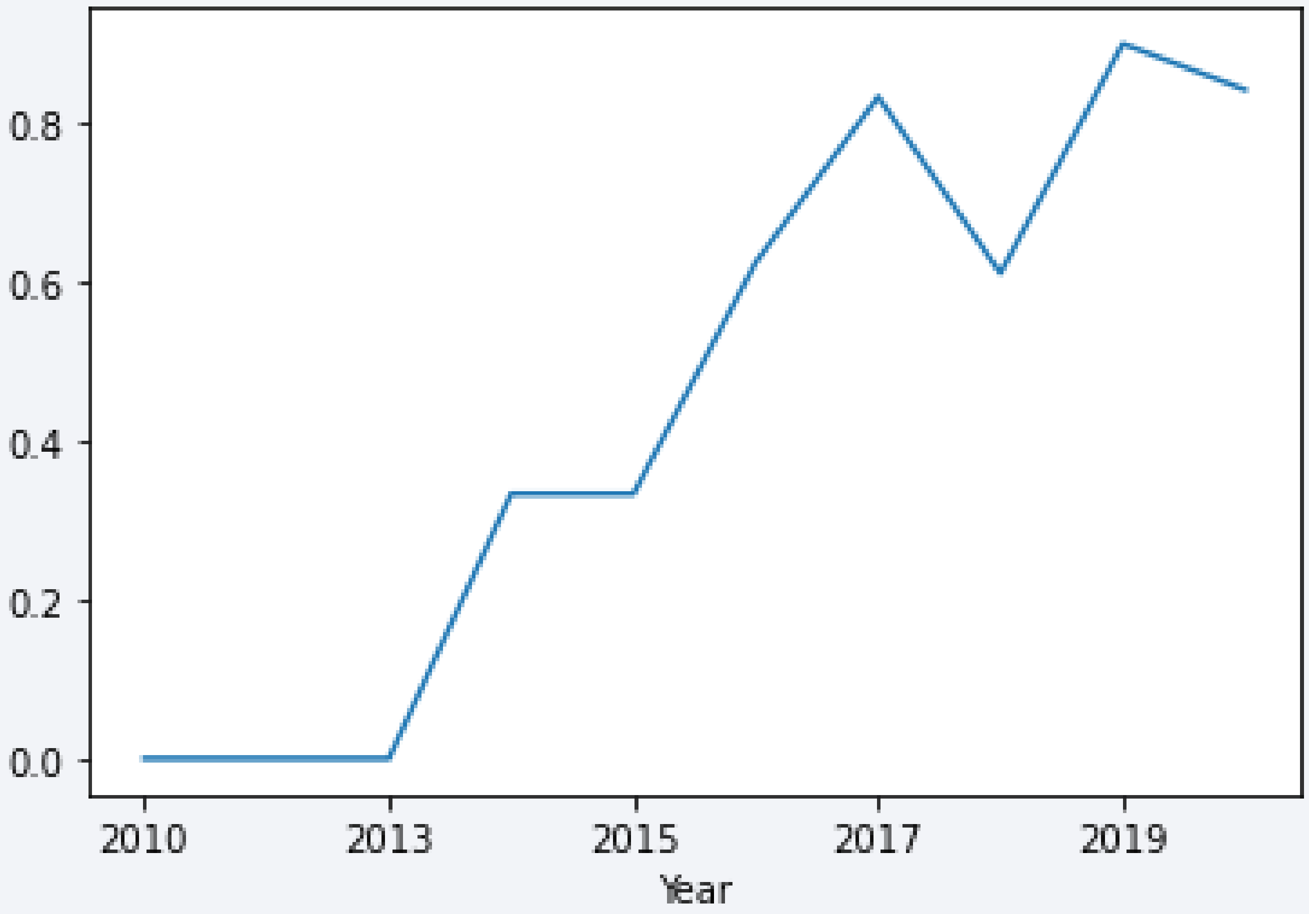
- While ISS orbit displays the widest payload range and high success, LEO and SSO show lower masses, and VLEO favors heavier payloads. Launches to SO and GEO appear infrequent.

- All payload masses align with target orbit



Launch Success Yearly Trend

- Success steadily increased since 2013, barring a temporary fall in 2018. It's now hovering around 80%
- 2013-2020 saw a clear upward trend



All Launch Site Names

- 3 unique "launch_site" values are present in the dataset:
  - CCAFS SLC-40
  - KSC LC-39A,
  - VAFB SLC-4E

Section 3

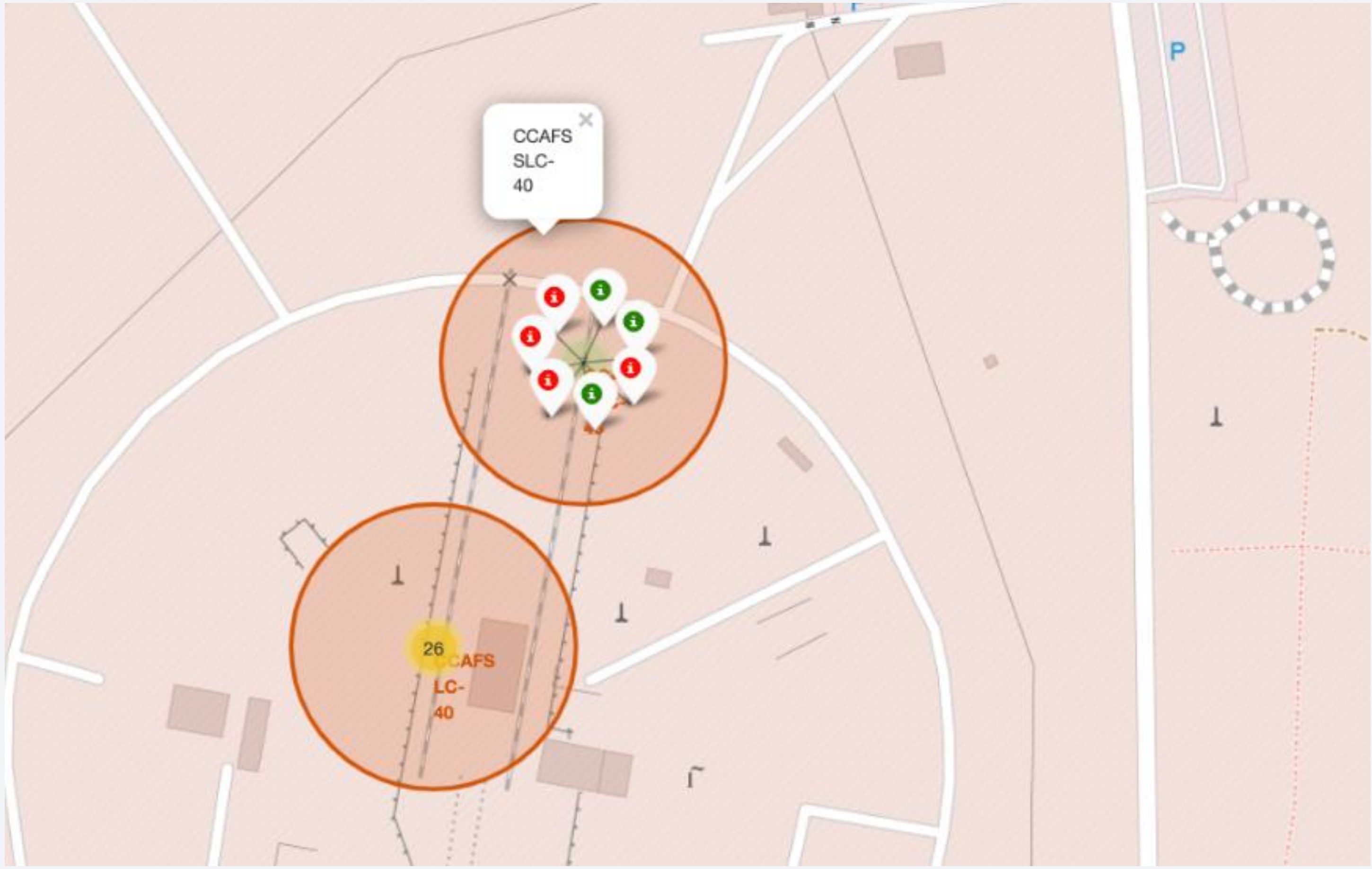
Launch Sites Proximities Analysis

Map Launch Sites

- Below is the U.S map displaying all launch sites, notably by seas. Two are bundled in Florida,

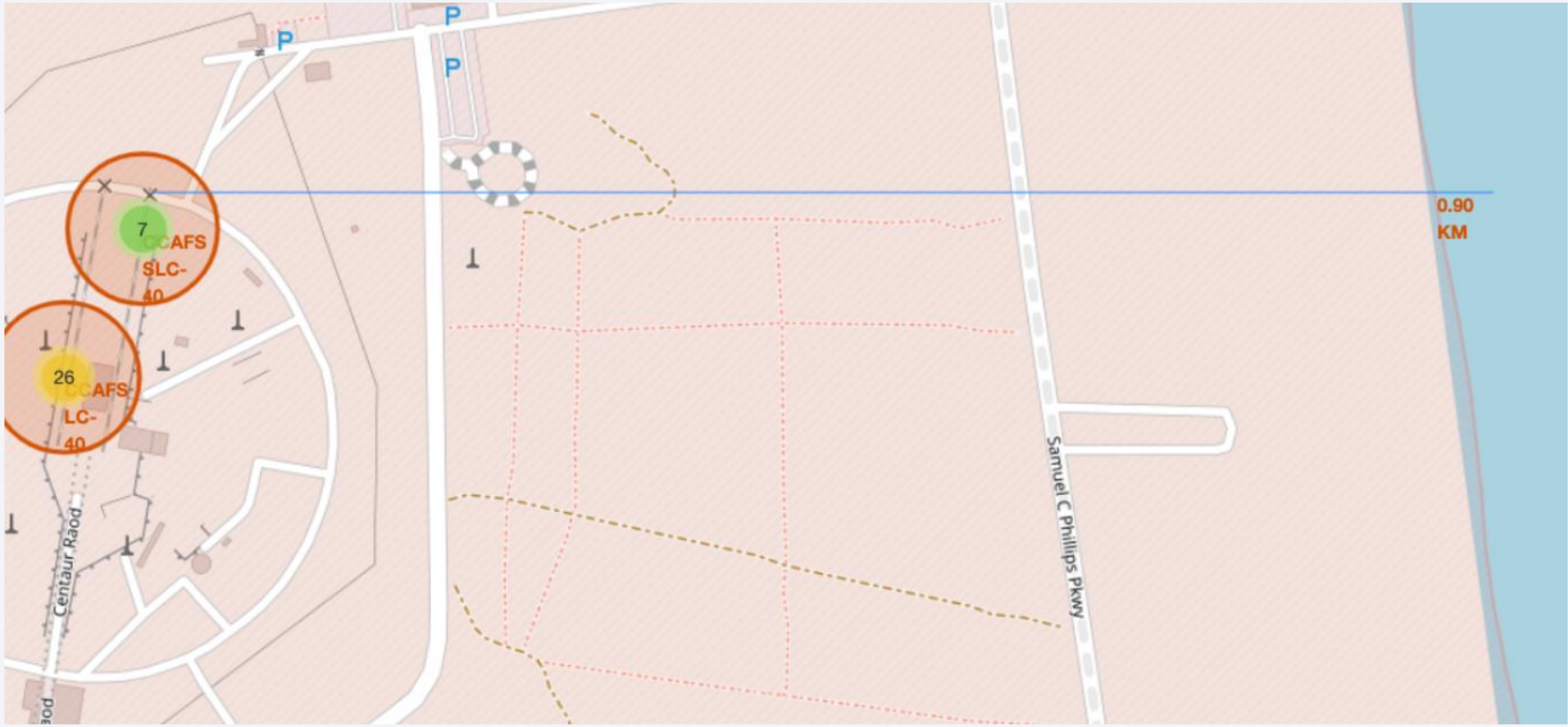
Site Launch Outcomes

- Example: CCAFS SLC-40 launch site launch outcomes, some of which were successful (in green) while others were failed (in red)



Proximities (railways,etc)

- Both Florida sites hugged highways for easy human access, and coastlines for safe emergency landings away from cities (<1KM from water bodies)

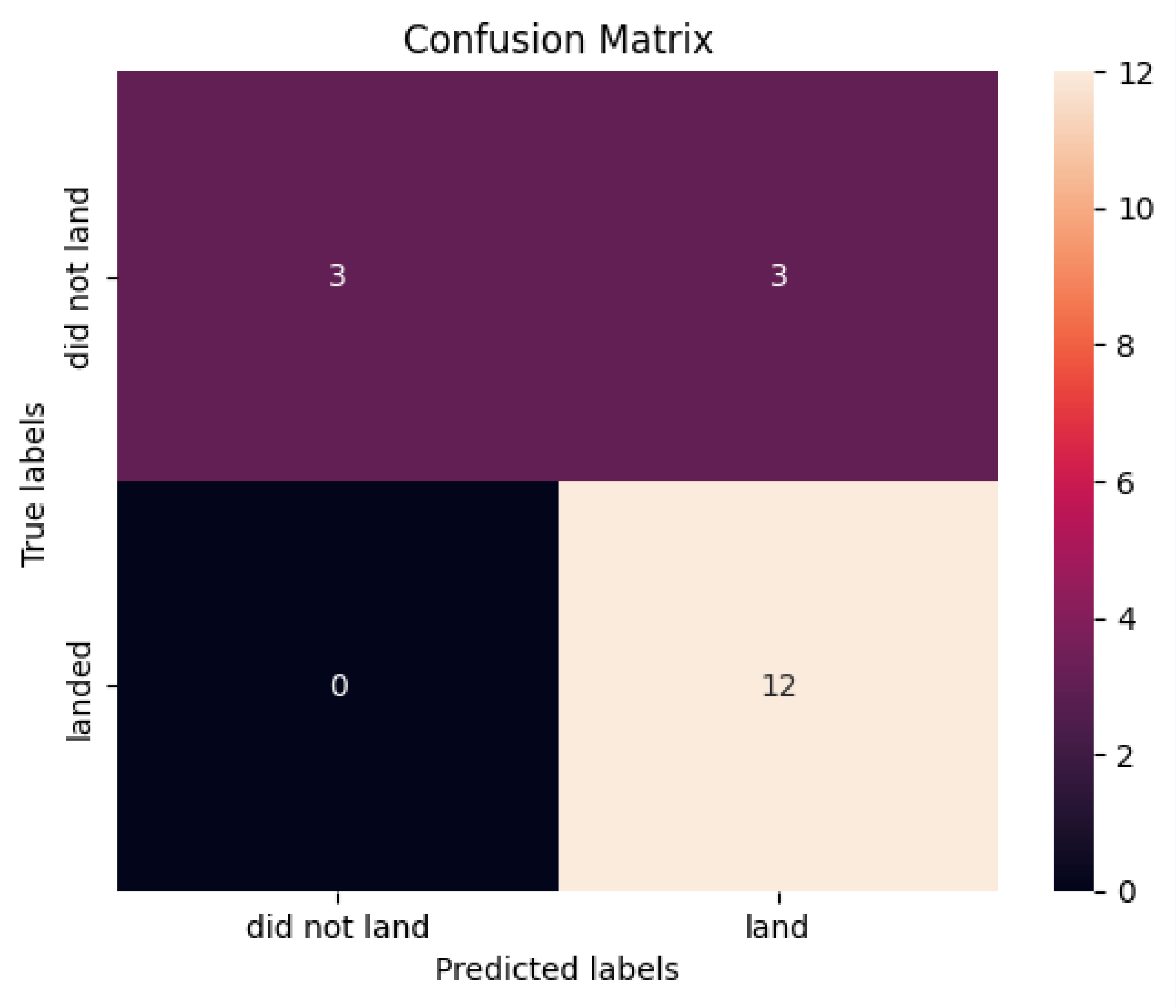


Section 4

Predictive Analysis (Classification)

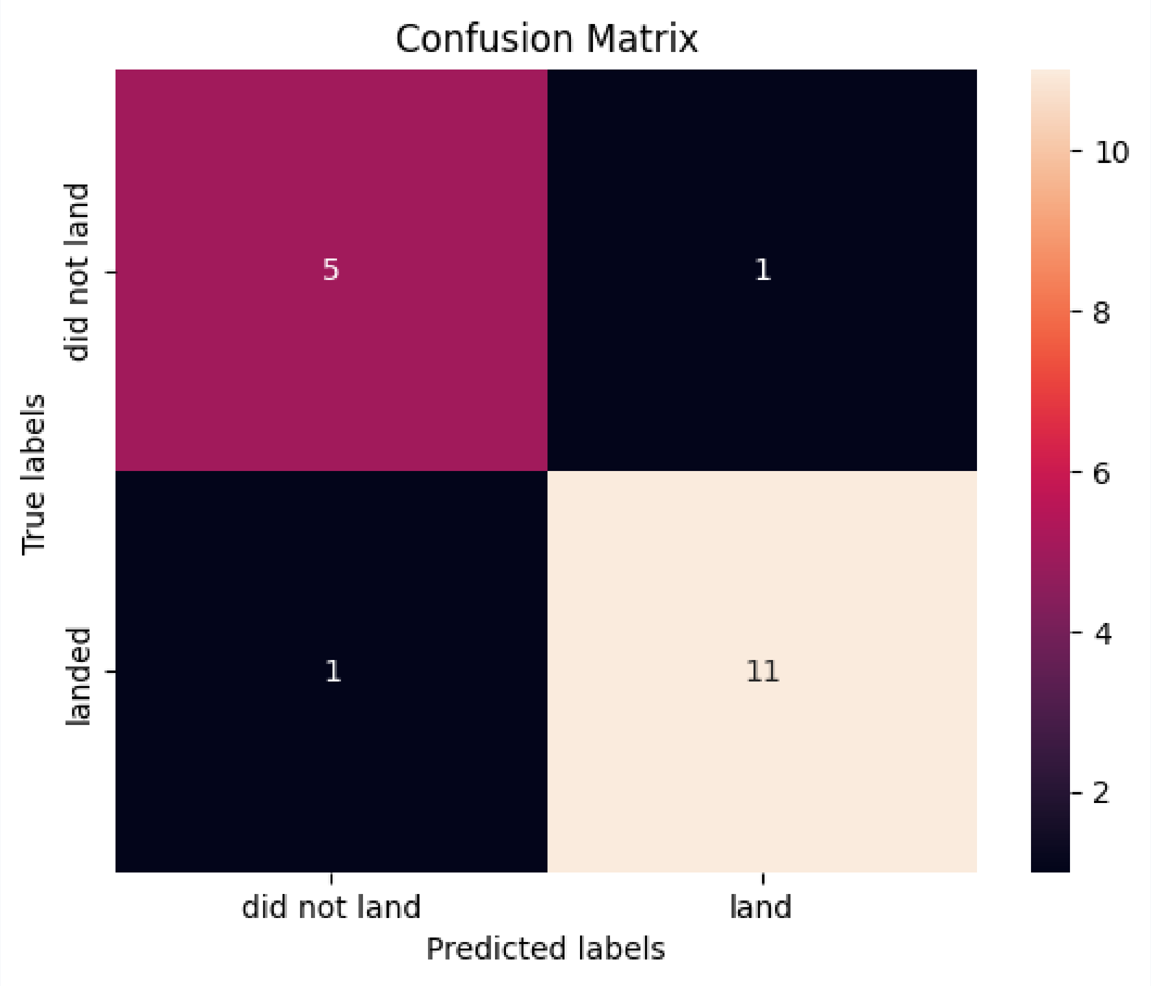
Confusion Matrix

- The models had high accuracy for successful landings (12 correct).
- However, they wrongly predicted successful landings 3 times for unsuccessful attempts (false positives). This shows a tendency to over-predict success



Confusion Matrix

- The decision tree confusion matrix turned out differently
- The Decision Tree accuracy stands out: its confusion matrix shows a strong dominance of true positives and true negatives over false results, setting it apart from other model matrices



Conclusions

- Analyzing diverse public data sources:
  - Lower risk for >7,000kg launches
  - Mission success improves in time with streamlining process/tech advancements
- Decision Tree Classifiers: predict successful landings & boost profits
- SpaceX's model reliably predicts Stage 1 landing success, guiding launch decisions

Appendix

- Issues
  - SQL screenshot crashes occurred during coding; no screenshots to provide other than GitHub link to SQL queries used
  - Plotly Application/Dashboard never refreshed for proper screenshots even after coding was implemented properly; GitHub link provided along with .csv file generated

Thank you!