IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Ģirts Bohvalovs>
<12/05/2025>

Outline

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

Executive Summary

- The goal was to predict SpaceX Falcon 9 rocket landings in the first stage.
- Multiple techniques have been applied in the analysis such as API, web scraping, data wrangling, EDA, visualizations, data preprocessing, and lastly machine learning task
- The resulting algorithms show similar accuracy of 83%

Introduction

- I am data scientist at SpaceY working for Allon Mask. My job is to analyze SpaceX Falcon 9 first stage rocket landing data to bid against them.
- The main problem of the study is to develop a machine learning model that can successfully predict whether a rocket is going to land or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API and Web Scraping from Wikipedia
- Perform data wrangling
  - Using pandas module data structure was transformed into a tabular format
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis has been performed using sklearn module GridSearchCV

Data Collection

- After researching SpaceX, I used Python's HTTP library to fetch data from their public API in JSON format. I converted the JSON into a DataFrame and filtered out "Falcon 1," focusing only on "Falcon 9" due to its successful landing history. I cleaned missing values early, retaining only LaunchPad nulls to indicate launches without a designated pad.
- Flowchart
  API Request -> Dataframe -> Data cleaning -> Data export

Data Collection – SpaceX API

- SpaceX, URL, Response, Json, Dataframe
- GitHub URL of the completed SpaceX API calls notebook

Spacex_url -> Response -> Response content -> Status code -> Json -> Helper Function -> Dataframe

Data Collection - Scraping

- URL, bs4, Dataframe

- GitHub URL of the completed web scraping notebook

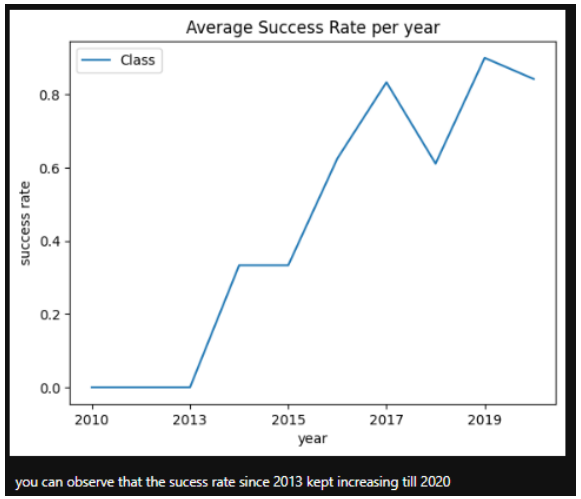
URL -> Request -> bs4 -> Tables -> Column names -> Helper function -> Dataframe

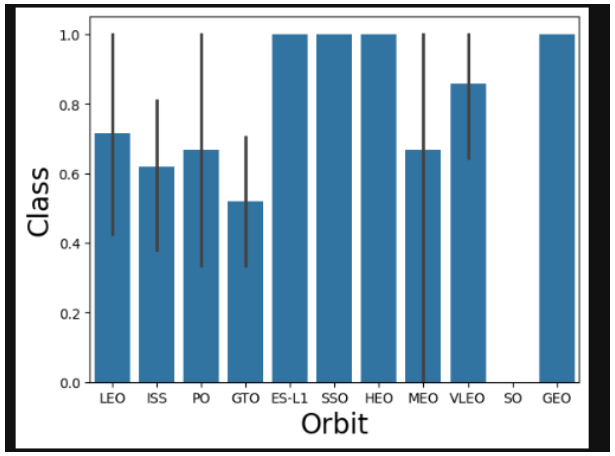
Data Wrangling

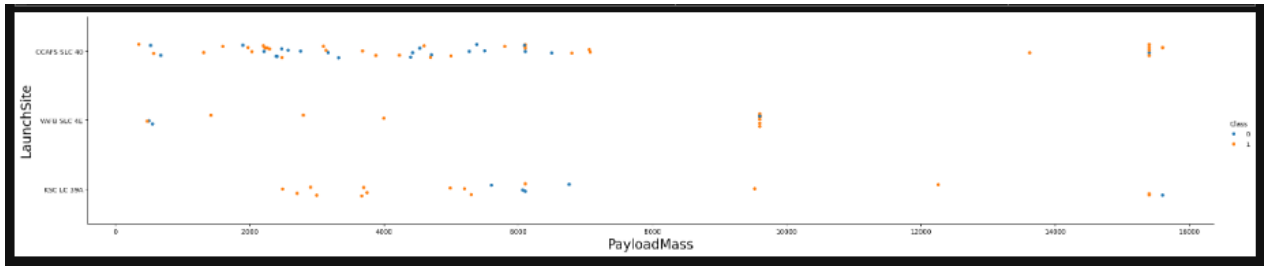
- I imported the cleaned data and checked the percentage of missing values in the LaunchingPad column, which indicated "no LaunchingPad" usage. I verified the data types (int64, object, float64, bool) and analyzed the LaunchSite value counts, finding Cape Canaveral SLC 40 and VAFB SLC 4E had the highest counts at 55.
- GitHub URL of your completed data wrangling related notebooks
- Import -> Column value counts -> Classes -> Success rate

EDA with Data Visualization

- I extensively used Seaborn and Matplotlib to create static visualizations, including:
- Catplot: To show the relationship between PayloadMass and LaunchSite.
- Barplot: To compare Orbit vs Class.
- Line Plot: To display the trend of average success rates from 2010 to 2020
- GitHub URL of your completed EDA with data visualization notebook







EDA with SQL

- Names of Unique launch sites in the space mission
- Five records where launch sites begin with string ‘CCA’
- Total payload mass by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- First successful landing outcome in ground pad
- Success in drone ship, payload mass >4000 but <6000
- Total number of successful and failure mission outcomes
- Booster_versions with maximum payload mass
- Month names of failure landing outcomes in drone ship in year 2015
- Landing outcome counts between "2010-06-04" and "2017-03-20“ in descending order
- GitHub URL of your completed EDA with SQL notebook

Build an Interactive Map with Folium

- In the Folium interactive map, I added markers to pinpoint launch site locations, such as NASA JSC Space Station, circles to highlight specific areas with text labels at precise coordinates, and lines to show the proximity between launch sites.
- GitHub URL of your completed interactive map with Folium map

Build a Dashboard with Plotly Dash – GITHUB!!!

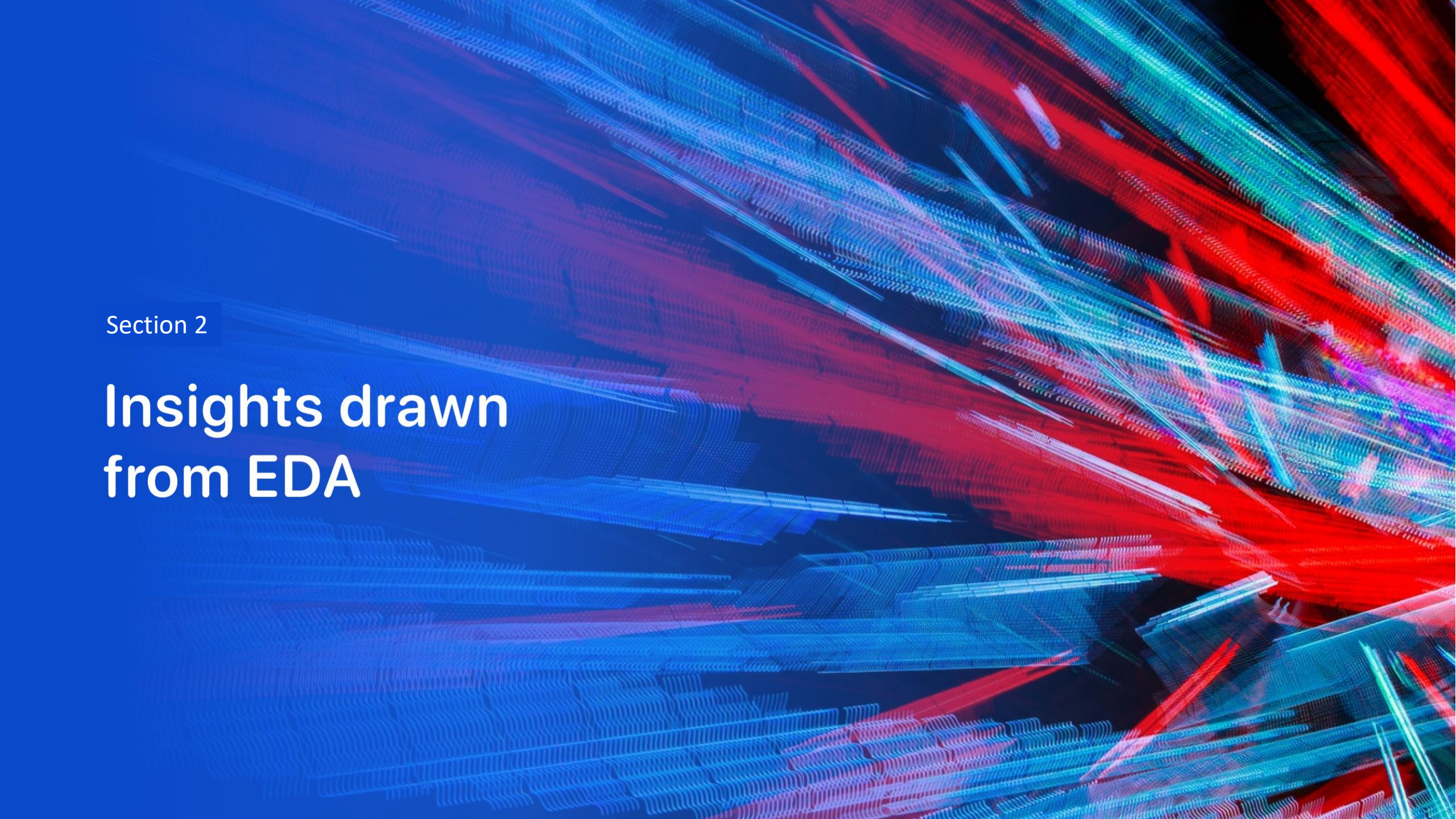
- In the Plotly Dash user application I have added a dropdown list and a range slider to allow a user to interact with a pie chart and the scatter point chart
- GitHub URL of your completed Plotly Dash lab

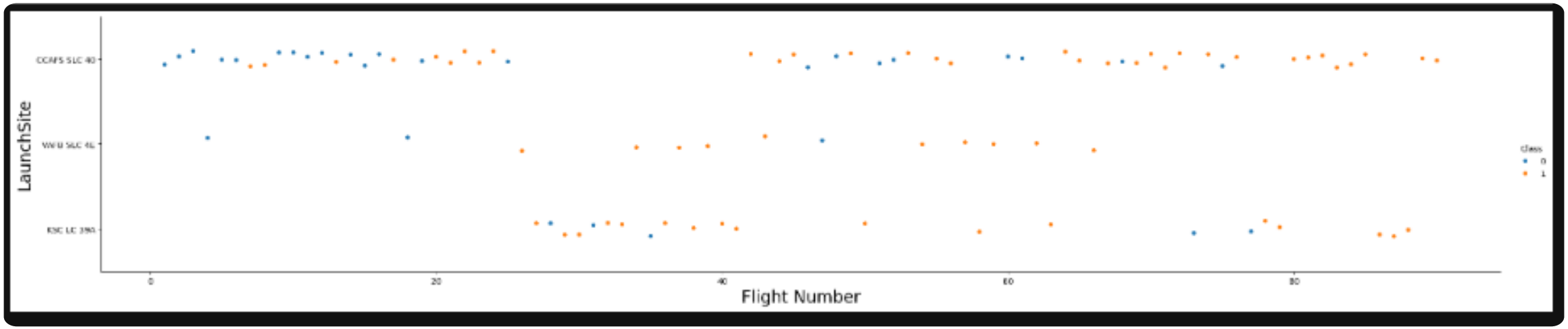
Predictive Analysis (Classification)

- The cleaned data was imported, with features assigned to variable X and the target to Y. The features were scaled using StandardScaler() and split using train_test_split(), allocating 20% for testing. Models were instantiated, and hyperparameters were optimized using GridSearchCV. After fitting on the training set and evaluating on the test set, the DecisionTreeClassifier achieved the best performance with an accuracy of 88.88%.
- GitHub URL of your completed predictive analysis lab
- Data import -> Feature assigning -> Train/Validation split -> Model -> Evaluation

Results

- The exploratory analysis revealed intriguing findings, including NASA (CRS) having a total payload mass of 45,596, an average payload mass of 2,928.4 for booster version F9 v1.1, and the first successful ground pad landing on "2015-12-22." Additionally, five booster versions achieved successful drone ship landings with payload masses between 4,000 and 6,000, and SpaceX's Falcon 9 recorded 99 successful mission outcomes. Visualizations highlighted a higher success rate for CCAFS SLC 40 as flight numbers increased, a 100% success rate for orbit types ES-L1, SSO, HEO, and GEO, and a significant improvement in Falcon 9's first-stage landing success from 2010 to 2020.


Section 2
Insights drawn from EDA

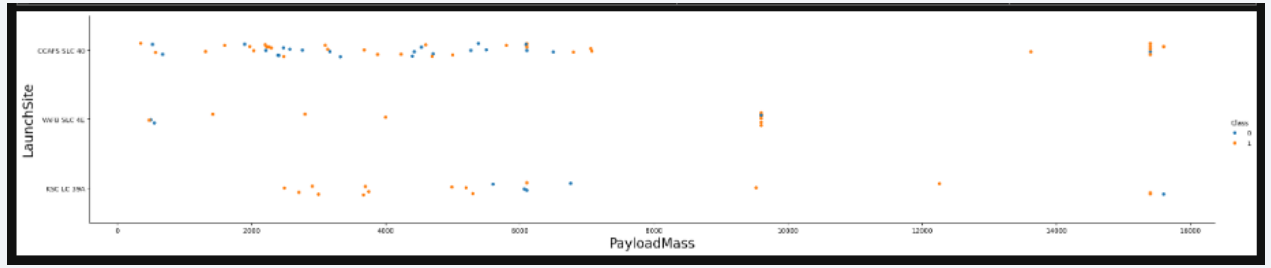


Flight Number vs. Launch Site

- The relationship between Flight Number and LaunchSite shows that lower flight numbers (around 20) resulted in no success or failure metrics for "KSC LC 39A" but two failures for "VAFB SLC 4E." However, as flight numbers increased beyond 80, "CCAFS SLC 40" achieved more successes, while "VAFB SLC 4E" recorded none.

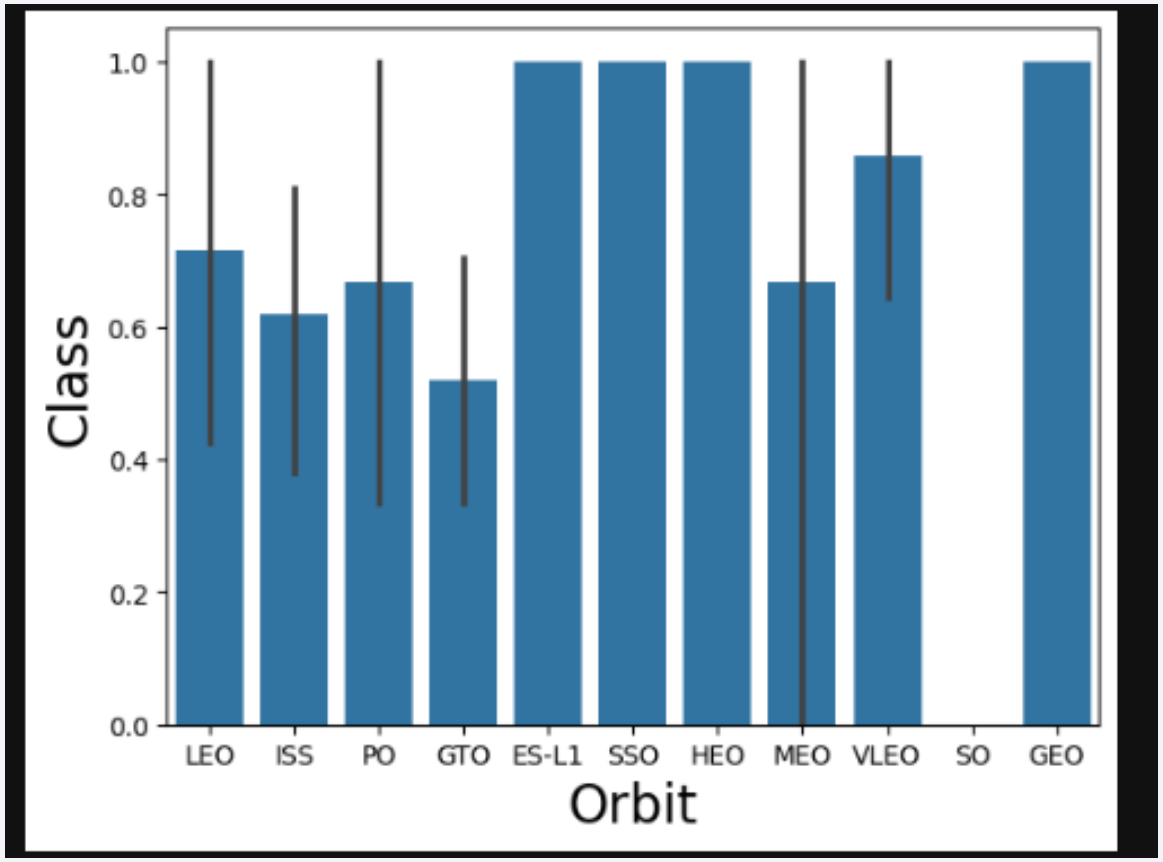
Payload vs. Launch Site

- The relationship between Flight Number and LaunchSite reveals that at lower flight numbers (around 20), "KSC LC 39A" had no success or failure metrics, while "VAFB SLC 4E" recorded two failures and "CCAFS SLC 40" had more failures than successes. However, as flight numbers increased beyond 80, "CCAFS SLC 40" achieved more successes, and "VAFB SLC 4E" recorded none.



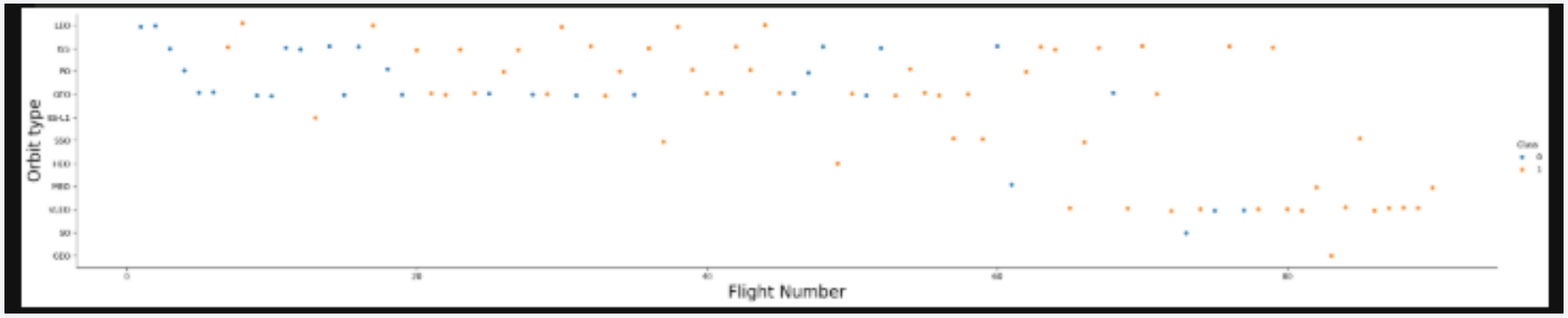
Success Rate vs. Orbit Type

- Four orbit types—ES-L1, SSO, HEO, and GEO—achieved the highest success rate of 100%.



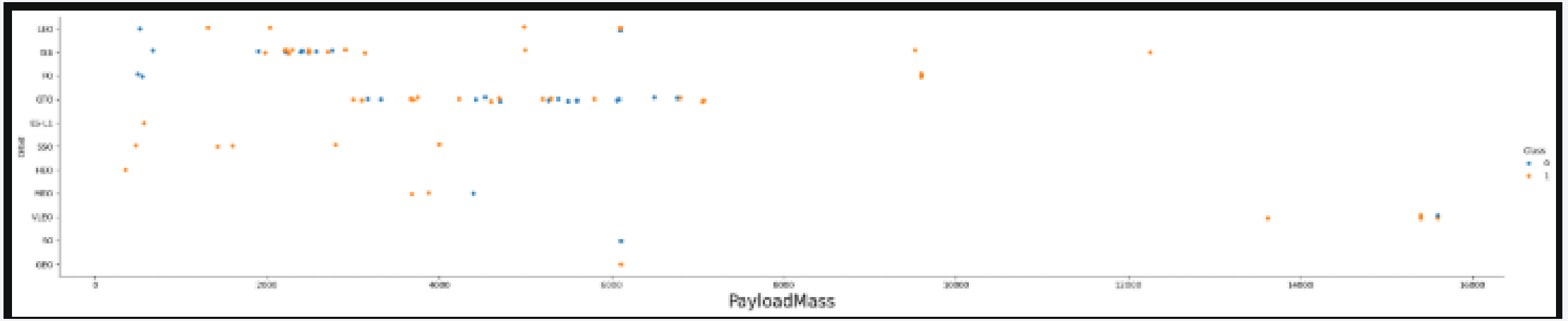
Flight Number vs. Orbit Type

- In VLEO, success occurs only for flight numbers above 60, while GTO shows no correlation.



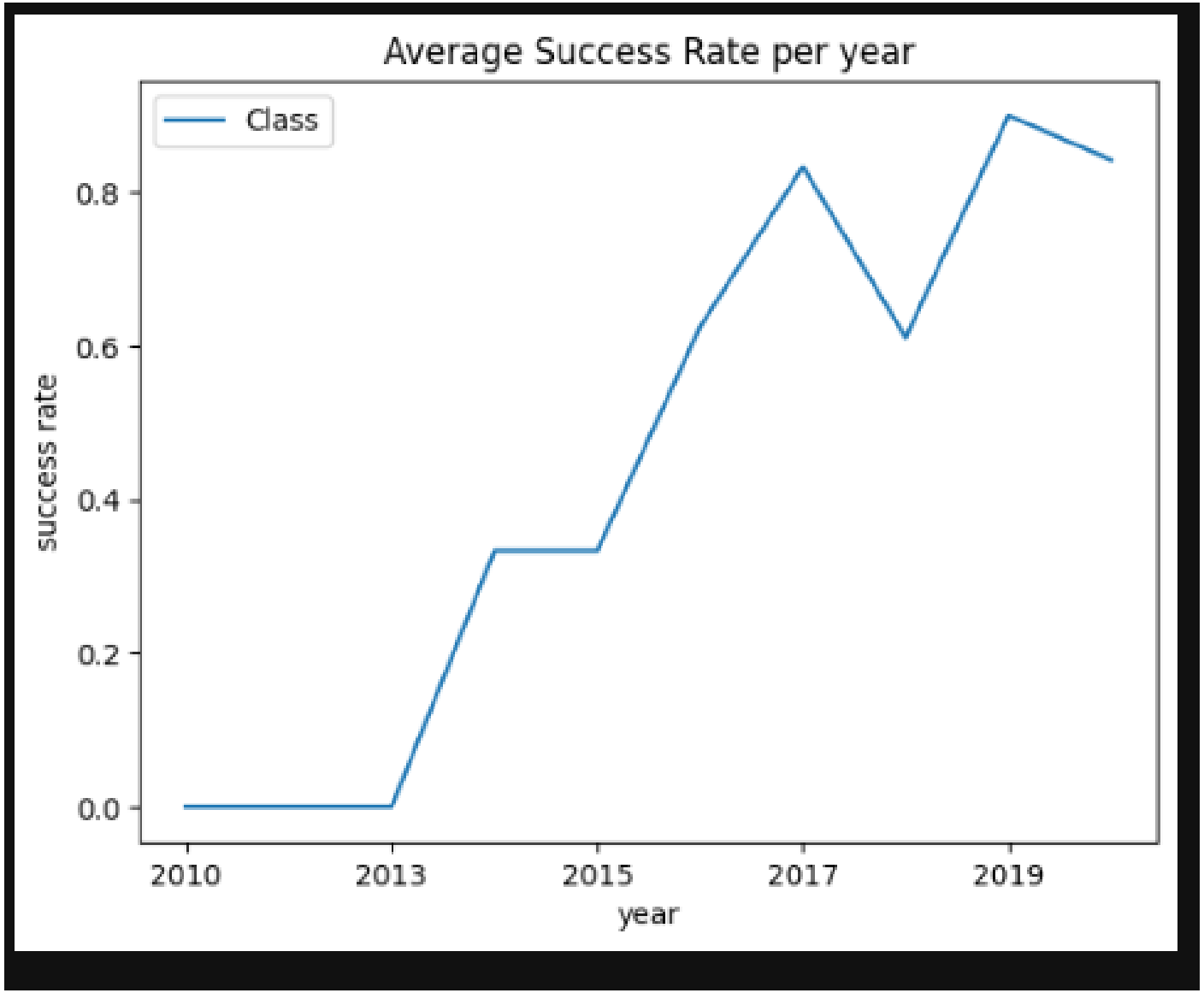
Payload vs. Orbit Type

- Polar, LEO, and ISS have higher landing success with heavy payloads, while GTO shows mixed results.



Launch Success Yearly Trend

- Falcon 9's landing success rate increased yearly from 2013 to 2020.



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

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

Total Payload Mass

- SUM (PAYLOAD_MASS__kg_)
- None

Average Payload Mass by F9 v1.1

- AVG(PAYLOAD_MASS__KG_)
- 2928.4

First Successful Ground Landing Date

- First_Successful_Landing
- 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 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

- SuccessfulMissions FailureMissions
- 100 1

Boosters Carried Maximum Payload

- Booster Versions which carried the Maximum Payload Mass
- 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

- Booster_Version Launch_Site
- F9 v1.1 B1012 CCAFS LC-40
- F9 v1.1 B1015 CCAFS LC-40

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

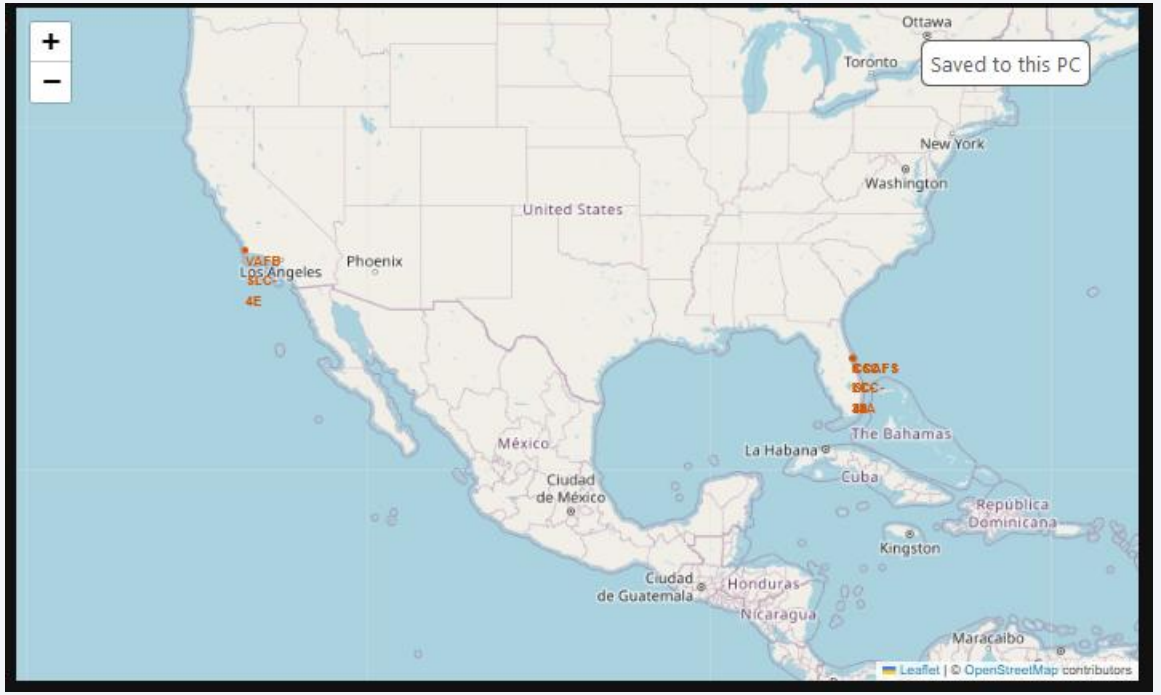
- Landing Outcome Total Count
- No attempt 10
- Success (drone ship) 5
- Failure (drone ship) 5
- Success (ground pad) 3
- Controlled (ocean) 3
- Uncontrolled (ocean) 2
- Failure (parachute) 2
- Precluded (drone ship) 1

Section 3

Launch Sites Proximities Analysis

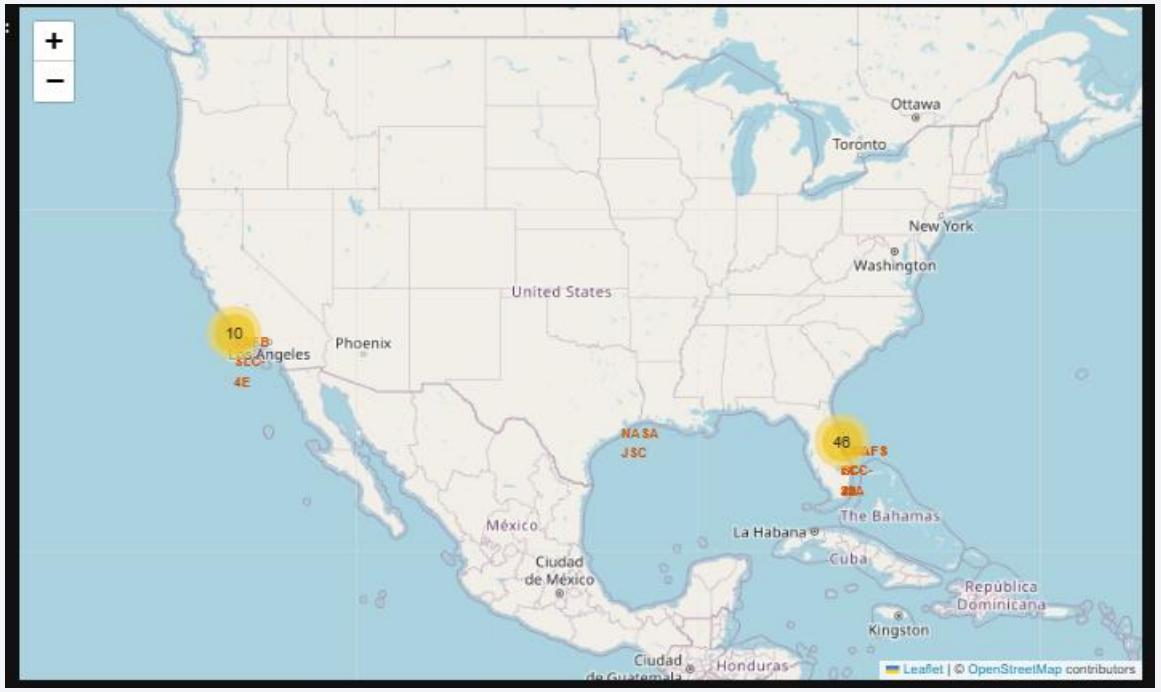
Launch site locations

- Launch sites are in California and Florida

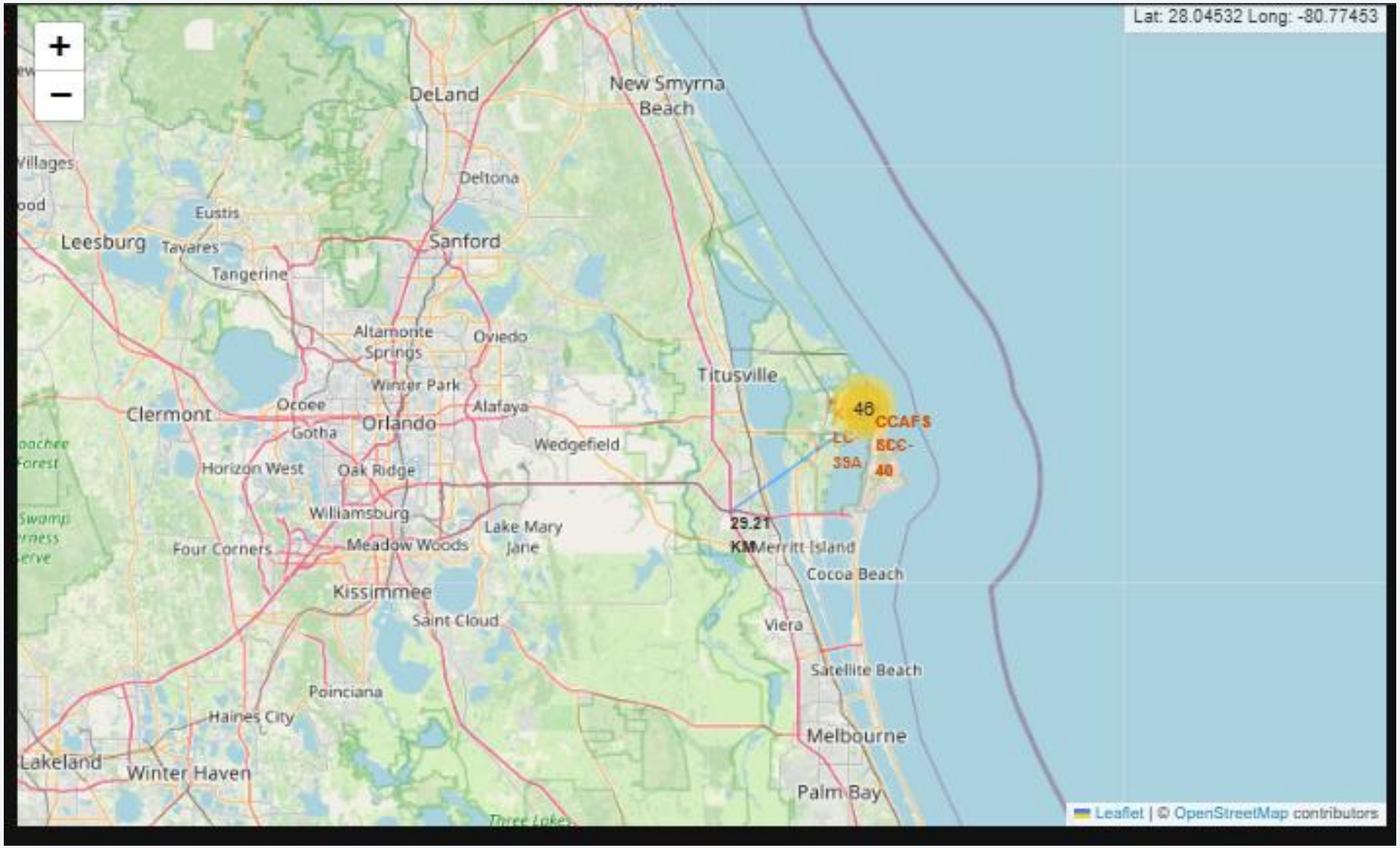


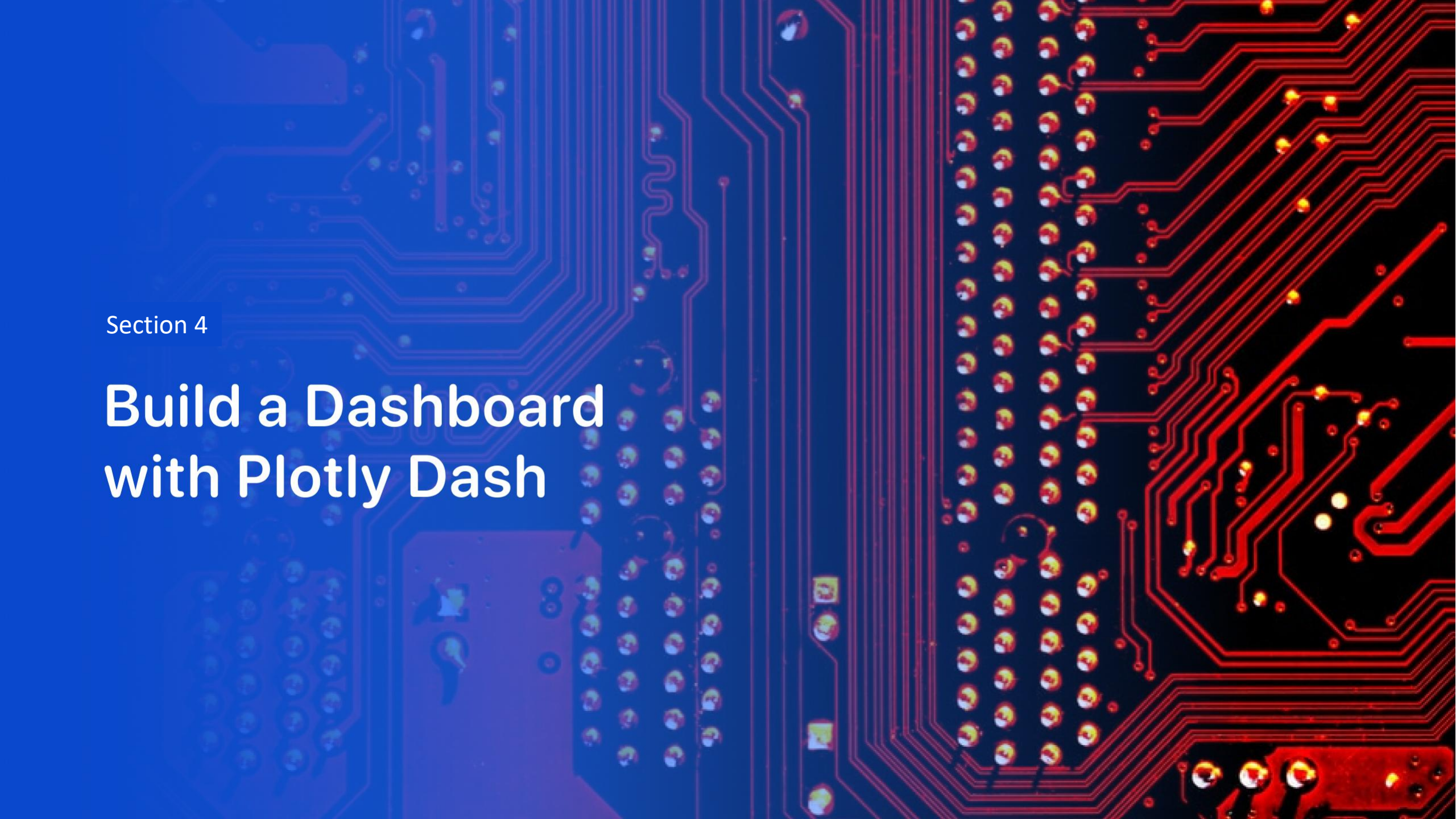
Launch outcomes

- 10 Launches in California and 46 in Florida



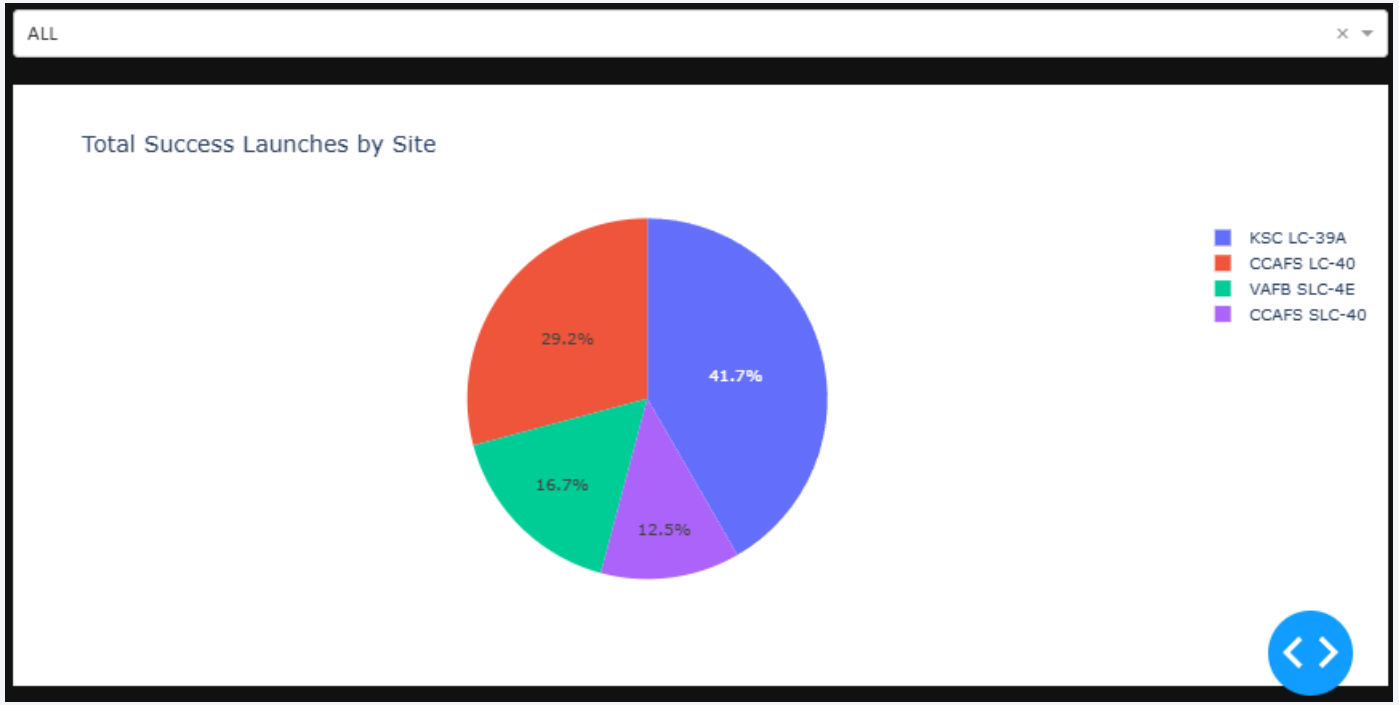
Proximity to highway




Section 4
Build a Dashboard with Plotly Dash

Pie chart

- KSC LC-39A leads with 41.7% of launches, followed by CCAFS LC-40 at 29.2%.



Payload mass and success rate

- For booster FT, success rate increases with payload mass, while v1.1 sees more failures.

Section 5

Predictive Analysis (Classification)

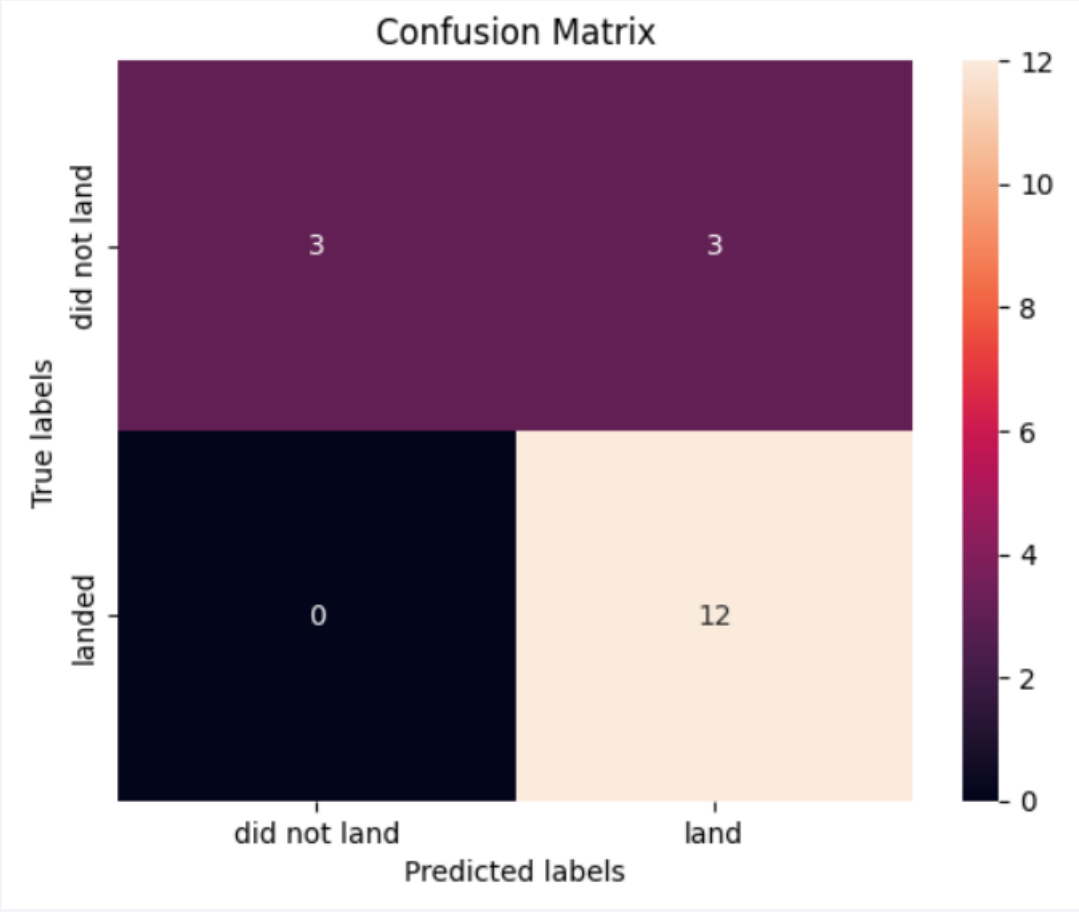
Classification Accuracy

- Model Comparison:
- Logistic Regression: 0.8196
- SVM: 0.8482
- Decision Tree: 0.8732
- KNN: 0.8482

- The best model is Decision Tree with an accuracy of 0.8732

Confusion Matrix

- Decision Tree with an accuracy of 0.8732



Conclusions

- Higher flight numbers and payload mass increase success at CCAFS SLC-40.ES-L1
- SSO, HEO, and GEO have 100% landing success.
- Falcon 9 landing success steadily rose from 2010 to 2020.
- The model predicts landing outcomes with 87.32% accuracy.

Appendix

- github repository - https://github.com/girts-bohvalovs/applied_data_science_capstone_spacex

Thank you!