IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Duke Kojo Kongo
21st June 2024

Outline

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

Executive Summary

- Data collection (REST API, web scraping), data wrangling, exploratory data analysis (EDA), interactive visual analytics (Folium, Plotly Dash), predictive analysis (classification models).
- Results: Summary of findings from the EDA, interactive visual analytics, and predictive analysis.

Introduction

Background: Falcon 9's reusability reduces launch costs.

Problem Statement: Predicting the successful landing of Falcon 9's first stage to estimate launch costs and bid competitively.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected through the Falcon REST API and web scraping.
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data sets were collected via:

- SpaceX REST API
- Web scraping.

Data Collection – SpaceX API

- **SpaceX API Calls**

- Launchpad: https://api.spacexdata.com/v4/launchpads/
- Rocket: https://api.spacexdata.com/v4/rockets/
- Payload:
- https://api.spacexdata.com/v4/payloads/
- Core: https://api.spacexdata.com/v4/cores/
- SpaceX API: https://api.spacexdata.com/v4/launches/past

- External reference: https://github.com/poppatheduke/SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

SpaceX API Calls

TASKS

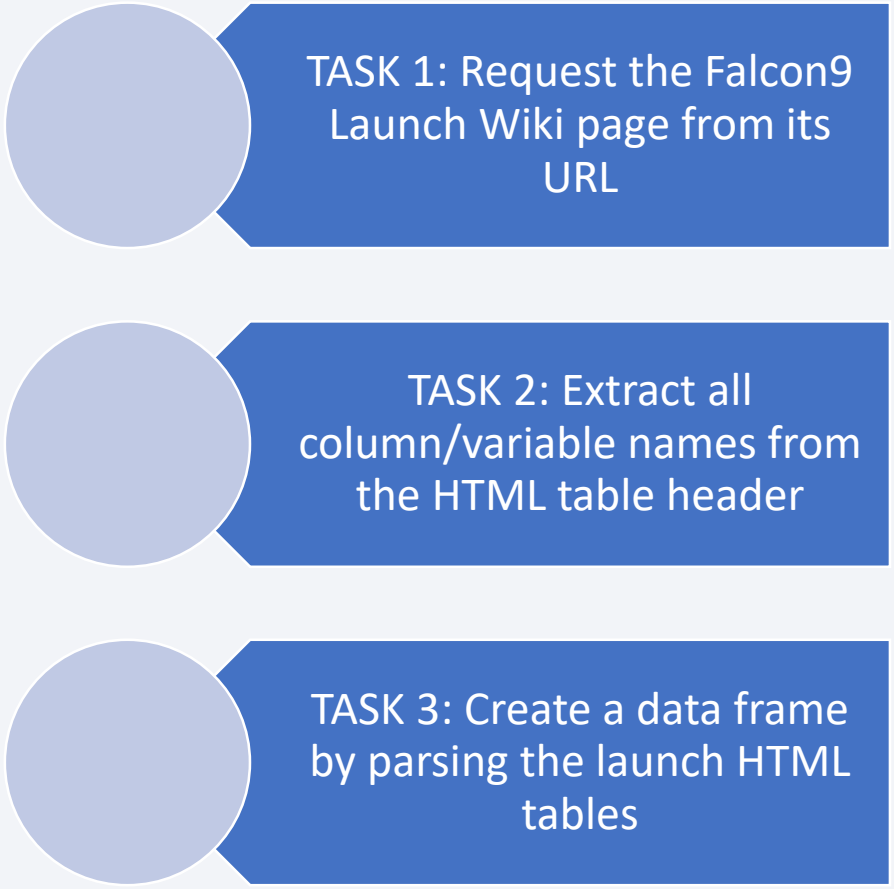
Request and parse the SpaceX launch data using the GET request

Filter the dataframe to only include Falcon 9 launches

Data Collection - Scraping

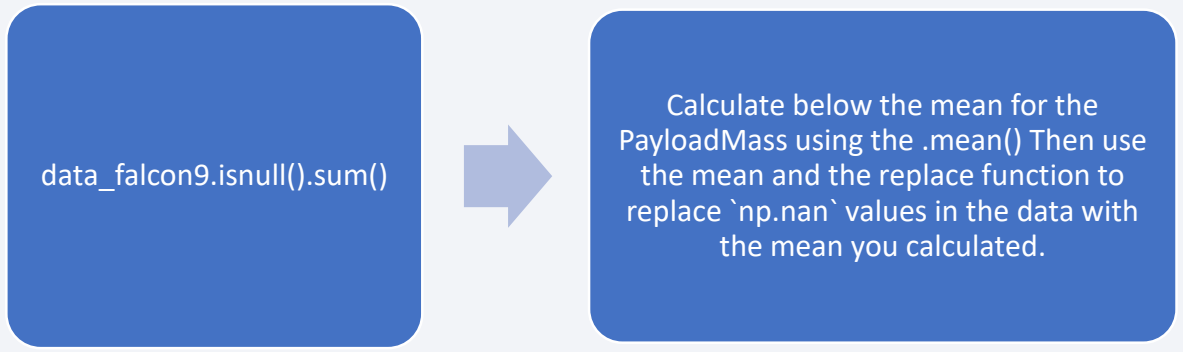
- Web scrape Falcon 9 launch records with BeautifulSoup:
- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

External reference:
https://github.com/poppathedu ke/SpaceX/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- We can see below that some of the rows are missing values in our dataset.
- Before we can continue we must deal with these missing values. The LandingPad column will retain None values to represent when landing pads were not used.
- Calculate below the mean for the PayloadMass using the .mean(). Then use the mean and the .replace() function to replace np.nan values in the data with the mean you calculated.



EDA with Data Visualization

- A scatter plot for Flight Number vs. Launch Site to
- A scatter plot for Payload vs. Launch Site
- A bar plot for Success Rate vs. Orbit Type
- A scatter plot for Flight Number vs. Orbit Type
- A scatter plot for Payload vs. Orbit Type
- A line plot for Launch Success Yearly Trend

- External reference: https://github.com/poppatheduke/SpaceX/blob/main/jupyter_labs_eda_datavi z.ipynb

EDA with SQL

- Connect to the database (SQLite was used).
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery
- 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.
- 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.
- External references: https://github.com/poppatheduke/SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Markers and Circles:** Added to highlight SpaceX launch sites, showing their exact locations and names.
- **Marker Cluster:** Implemented to efficiently display multiple markers for launch outcomes.
- **Success/Failure Markers:** Color-coded markers to indicate successful (green) and failed (red) launches.
- **Points of Interest:** Markers and lines added to show distances from launch sites to the closest city, railway, and highway.
- These objects were added to visually represent and analyse the spatial relationships and launch outcomes of SpaceX sites.
- **GitHub URL**
- Check out the interactive map on GitHub: GitHub repository for Folium

Build a Dashboard with Plotly Dash

- Drop-down Menu: Allows selection of different launch sites.
- Range Slider: Enables selection of payload range for analysis.
- Pie Chart: Displays the total success launches for all sites or a specific site.
- Scatter Plot: Shows the correlation between payload and launch success, with booster versions color-labeled.
- Check out the interactive map on GitHub: https://github.com/poppatheduke/SpaceX/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data Loading and Preparation

- Import Data
- Extract Features (X) and Labels (Y)
- Standardize Features

Data Splitting

- Split Data into Training and Testing Sets

Model Building and Hyperparameter Tuning

- Logistic Regression: Define Parameters and Grid Search
- SVM: Define Parameters and Grid Search
- Decision Tree: Define Parameters and Grid Search
- KNN: Define Parameters and Grid Search

Model Evaluation on Test Data

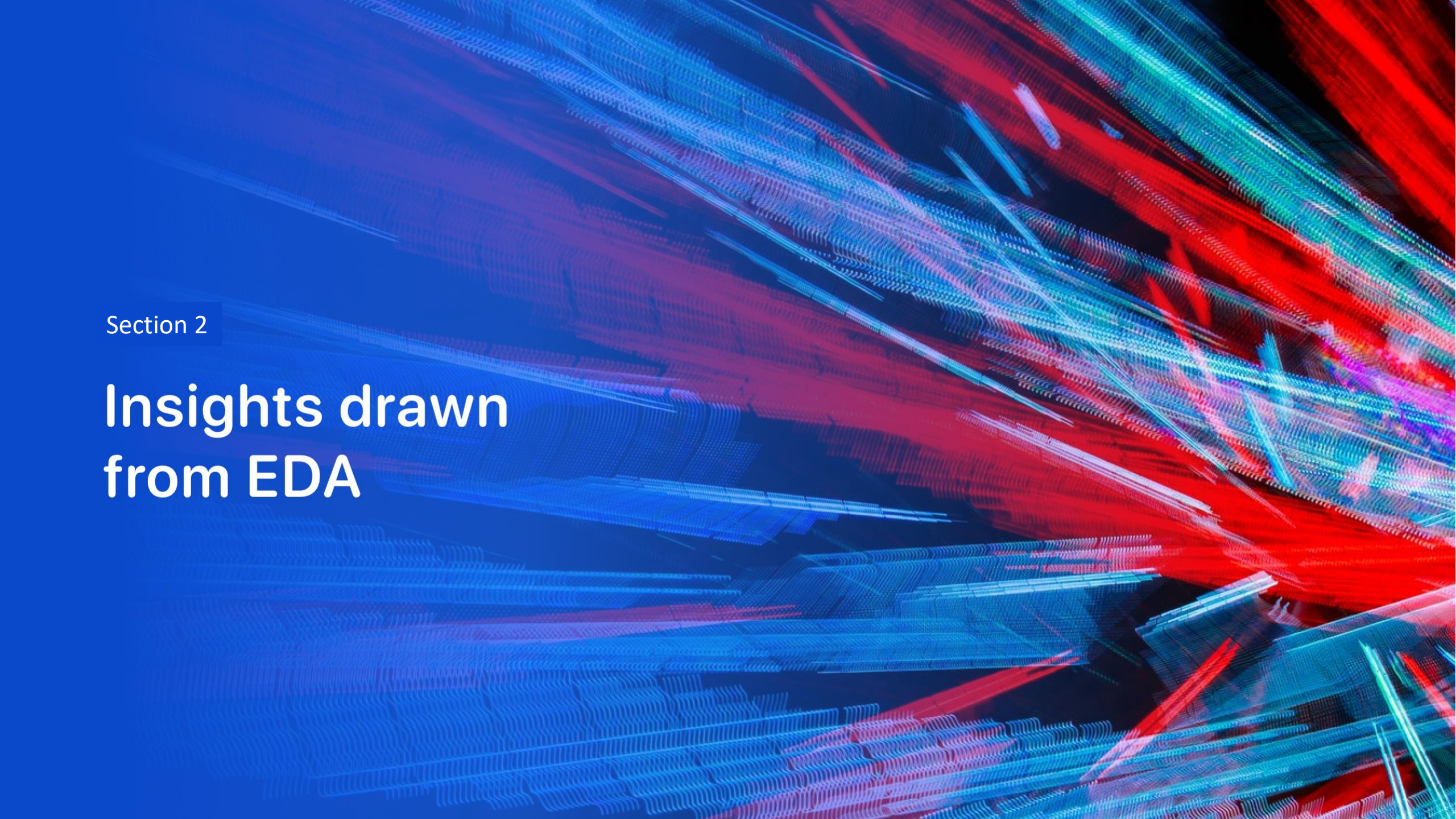
- Calculate Accuracy
- Plot Confusion Matrices

Best Model Selection

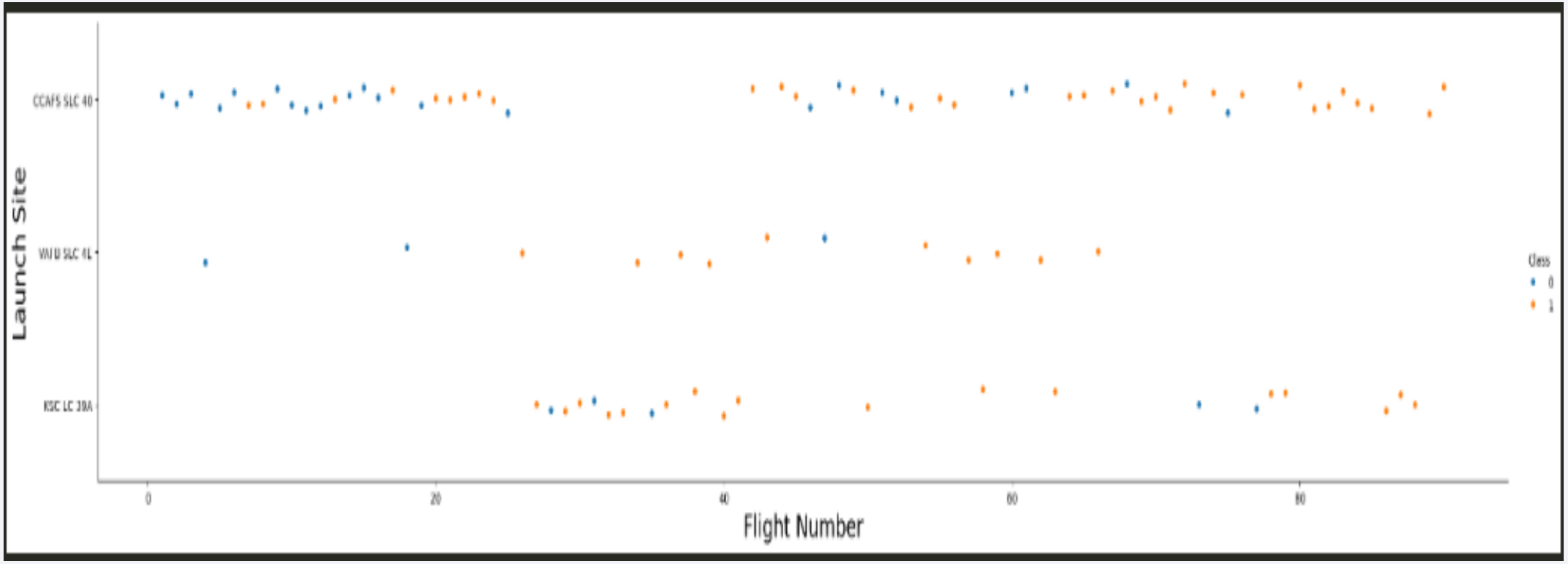
- Compare Test Accuracies
- Select Best Model

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

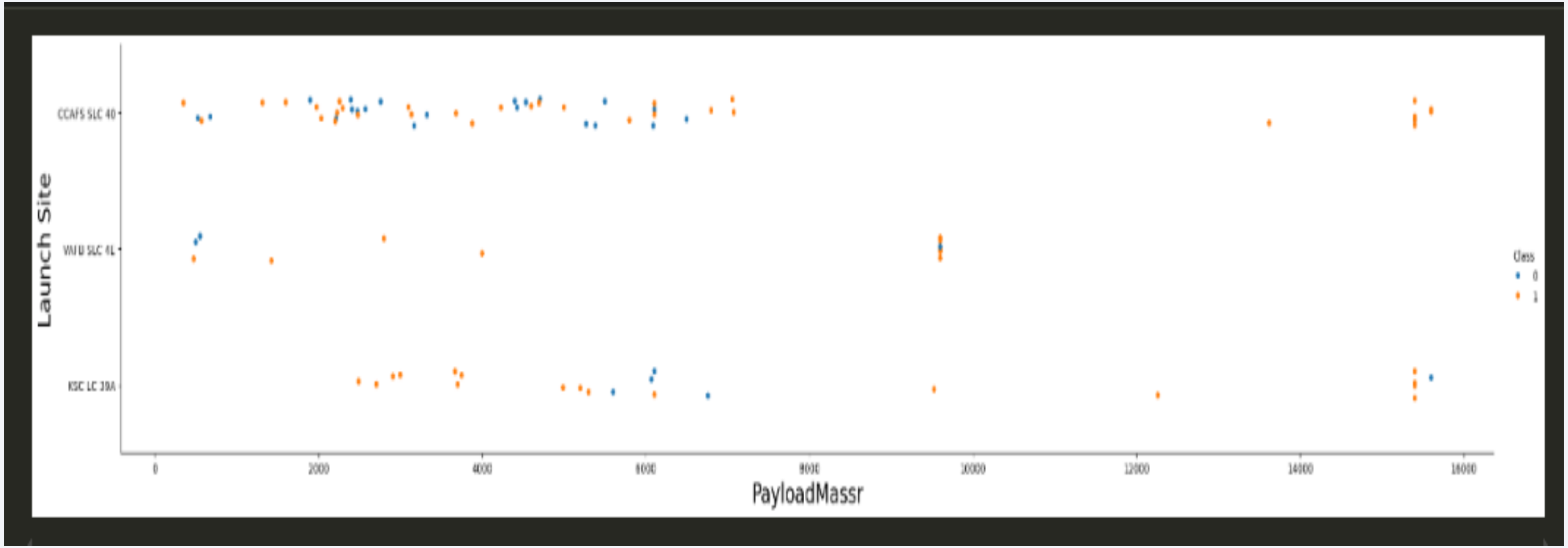

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



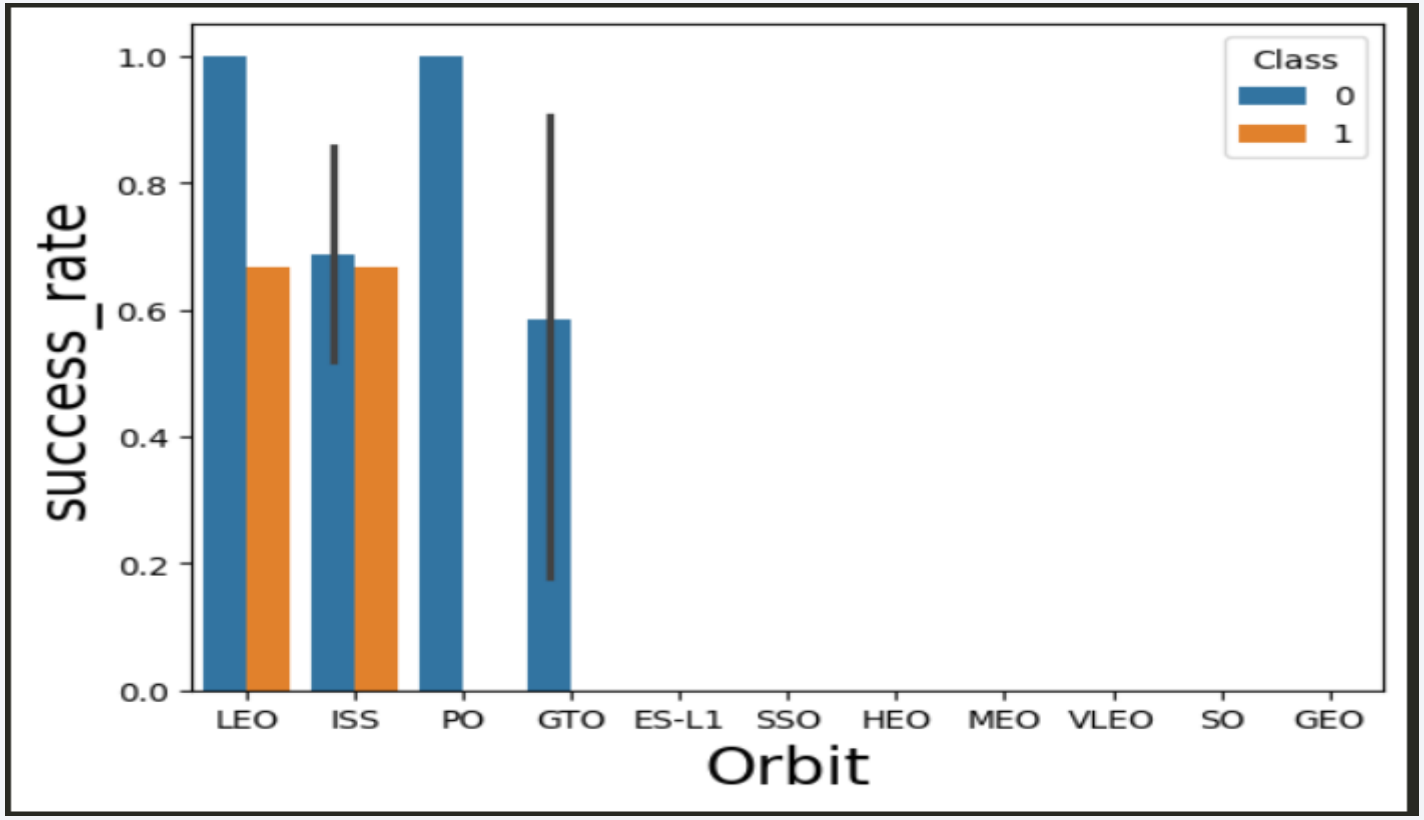
The spread of points across the y-axis shows which launch sites are used more frequently. This can highlight primary and secondary sites based on the density of flight numbers.

Payload vs. Launch Site



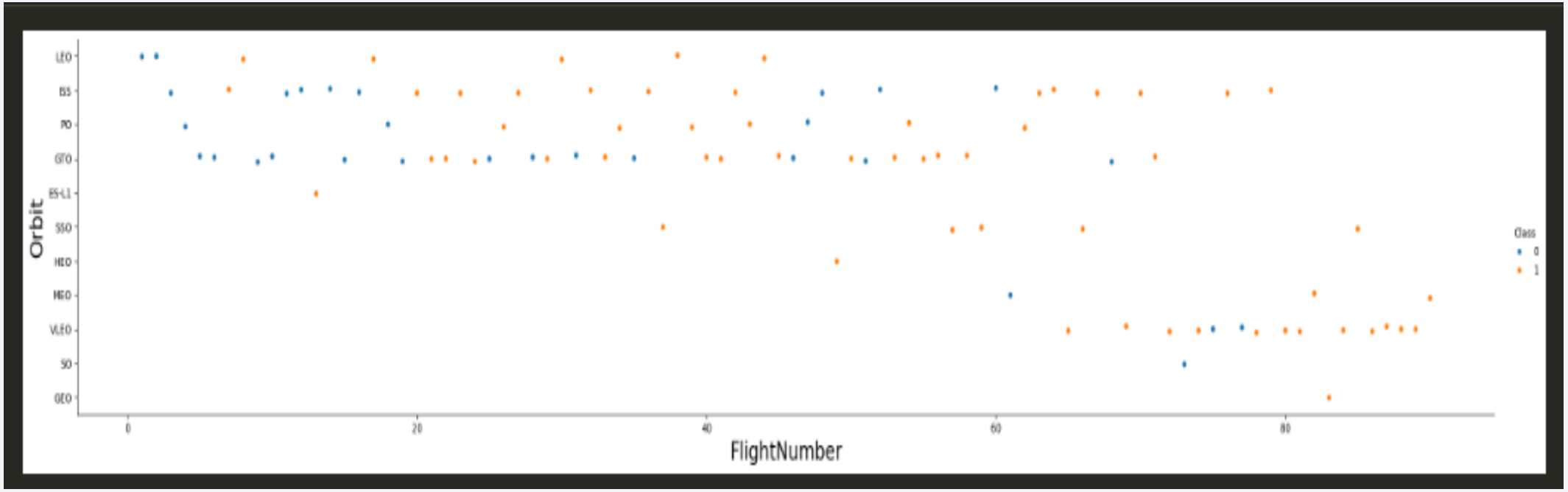
- Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



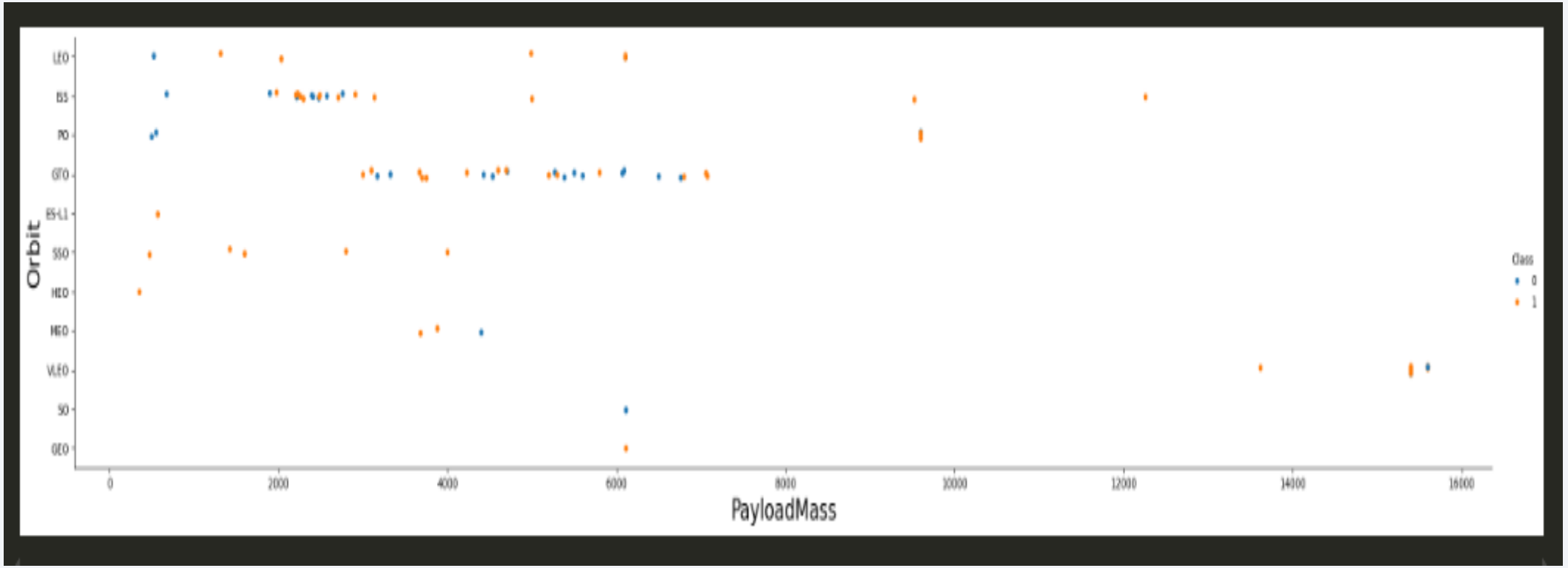
Both PO and LEO have the highest success rate.

Flight Number vs. Orbit Type



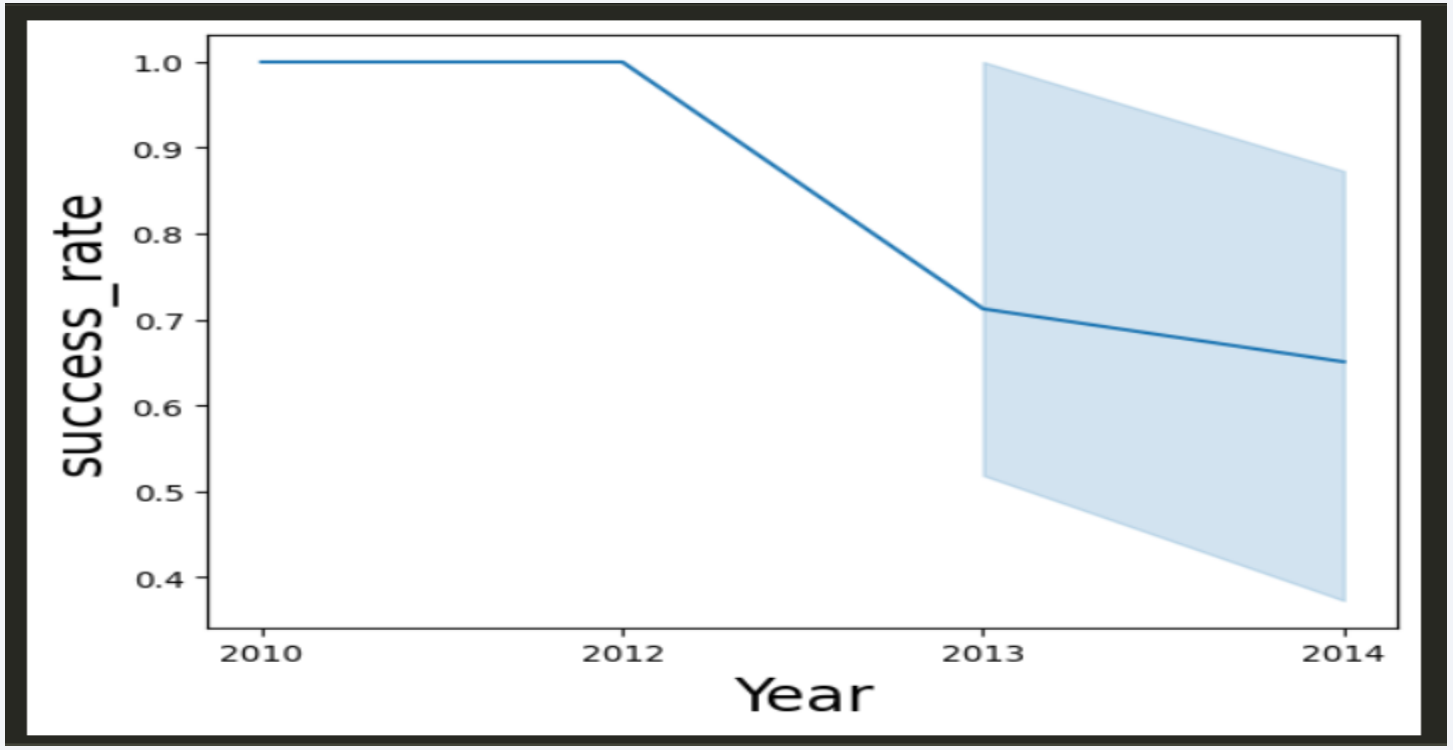
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- 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



You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

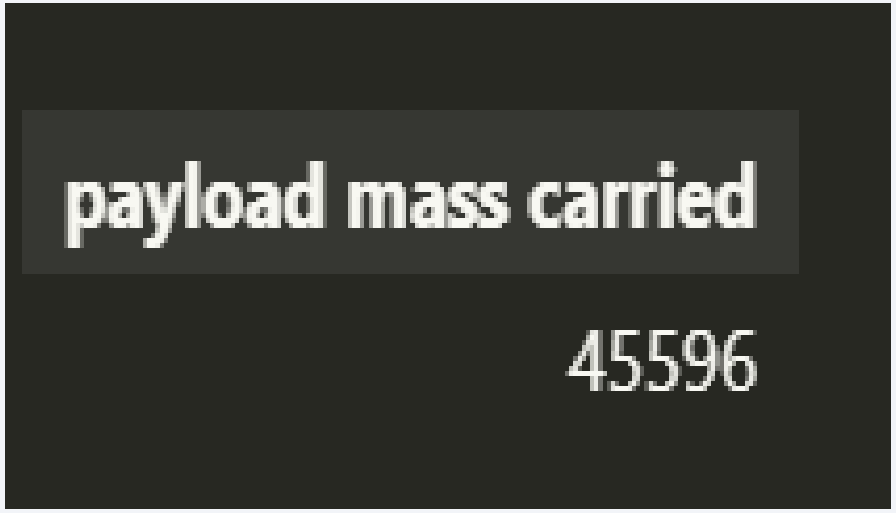
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

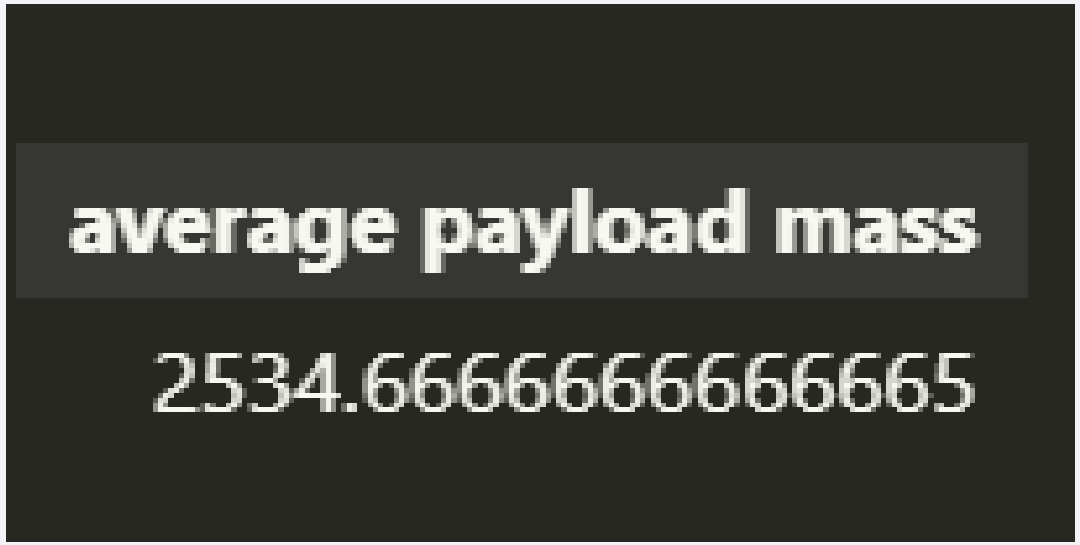
5 records where launch sites begin with the string 'CCA

Total Payload Mass



The total payload mass carried by boosters launched by NASA (CRS) was 45596

Average Payload Mass by F9 v1.1



The average payload mass carried by booster version F9 v1.1 was approximately 23534.67

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad was 22nd July 2018

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

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

2015 Launch Records

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

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

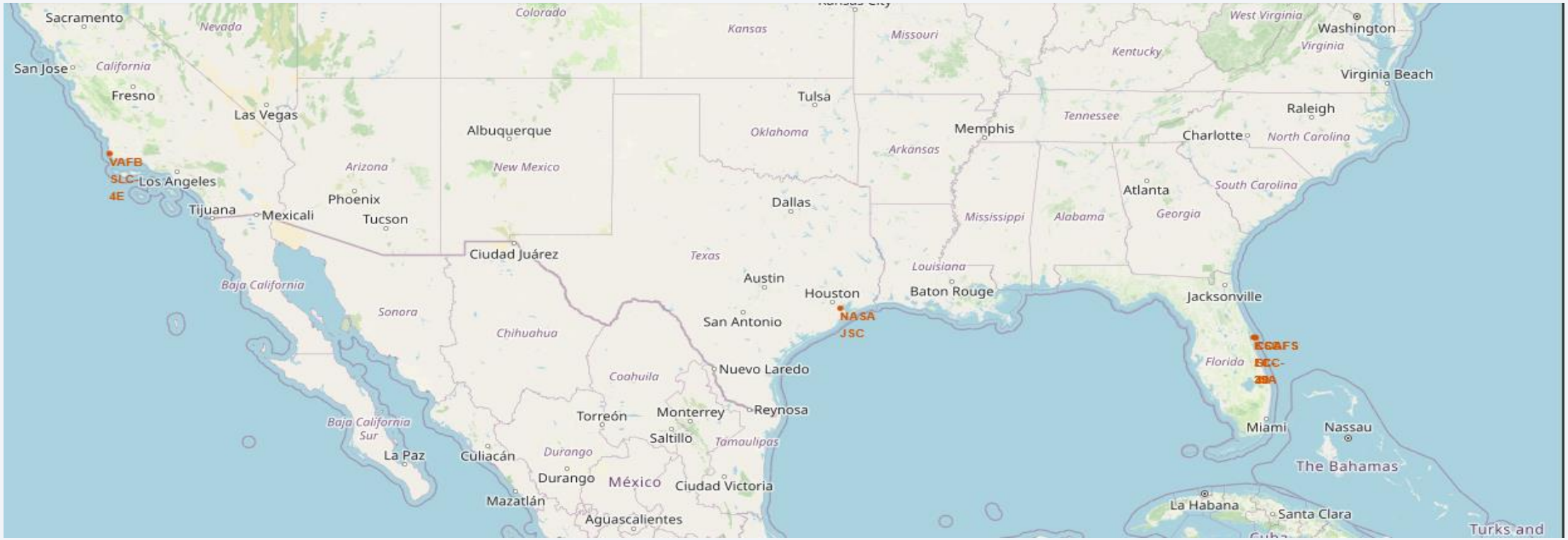
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

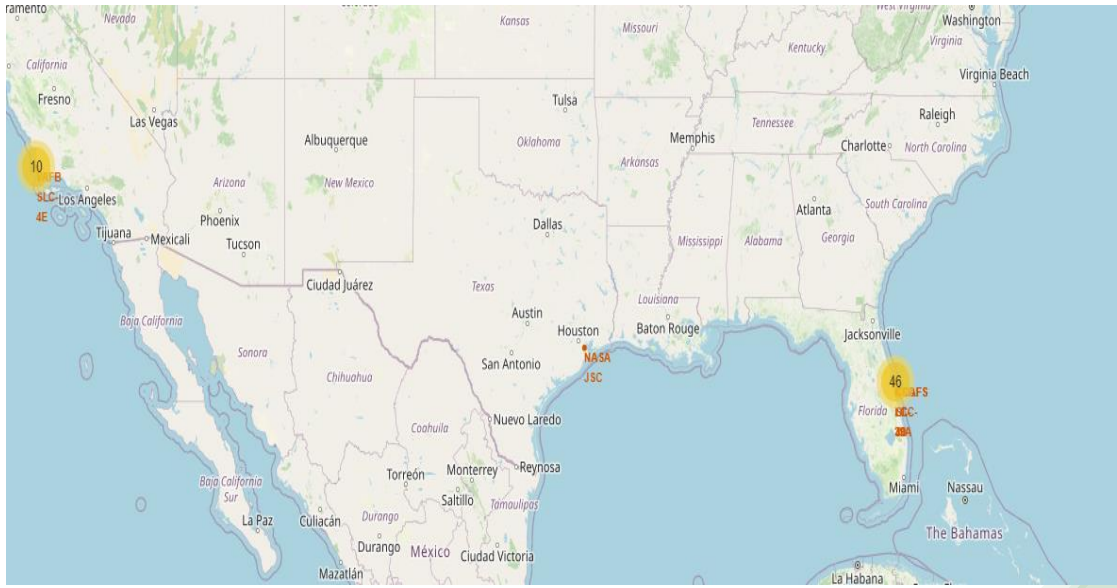
Interactive Map of SpaceX Launch Sites and Outcomes

- The map clearly shows the geographic distribution of SpaceX launch sites across the United States, with sites located in Florida, California, and Texas.



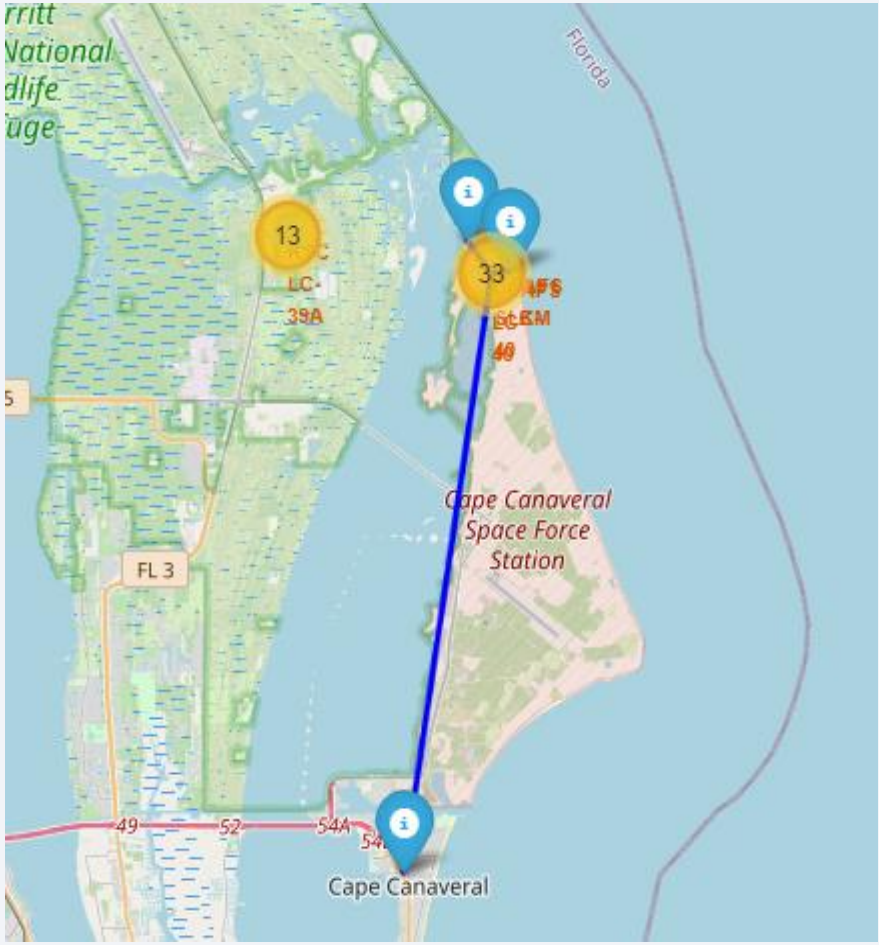
Color-Coded Launch Outcomes on SpaceX Launch Sites Map

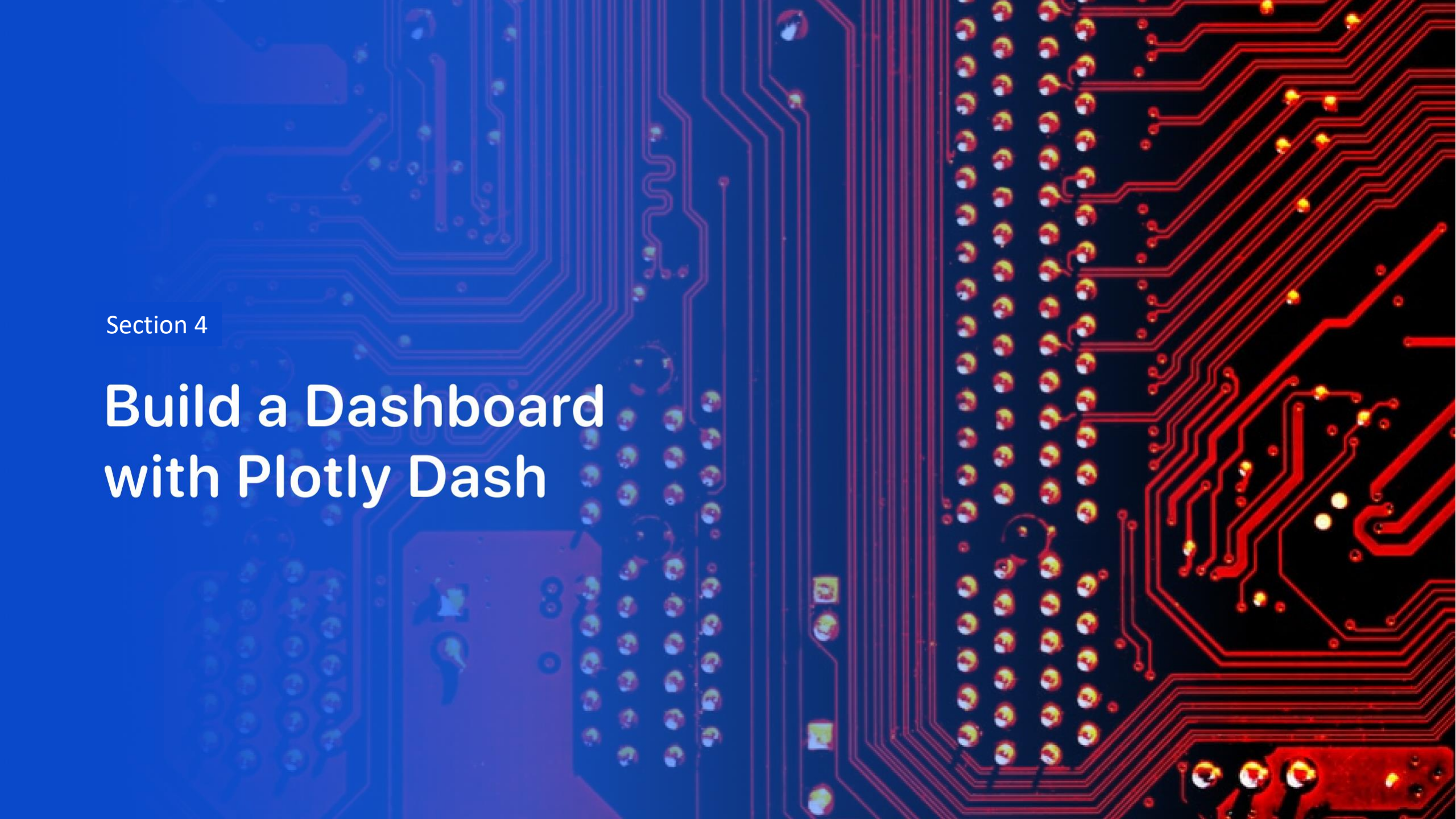
- **Important Elements and Findings**
- **Color-Coded Markers:**
  - **Green Markers:** Indicate successful launches.
  - **Red Markers:** Indicate failed launches.
- **Marker Clusters:**
  - Efficiently group multiple markers to avoid clutter.
  - Clicking on a cluster zooms in to show individual markers.
- **Launch Sites:**
  - Markers and circles denote specific launch site locations.
  - Includes sites such as CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.
- **Findings**
- The color-coded markers provide a clear visual differentiation between successful and failed launches.
- The distribution and outcomes of launches at different sites are easily analyzed at a glance.



Proximities of SpaceX Launch Site to Railway, Highway, and Coastline

- Proximity Analysis: The map shows the spatial relationship between the launch site and nearby infrastructure. Distances to key proximities such as railways, highways, and coastlines are calculated and displayed, providing insights into the site's accessibility and logistics.
- Logistical Insights: Understanding these proximities is crucial for planning and executing space missions, as they impact transportation of materials and personnel. This detailed proximity analysis aids in evaluating the strategic location of SpaceX launch sites in relation to essential infrastructure.




Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Success Count for All Sites

- The pie chart provides a quick visual summary of the total number of successful launches, helping to gauge the overall success rate across all sites.

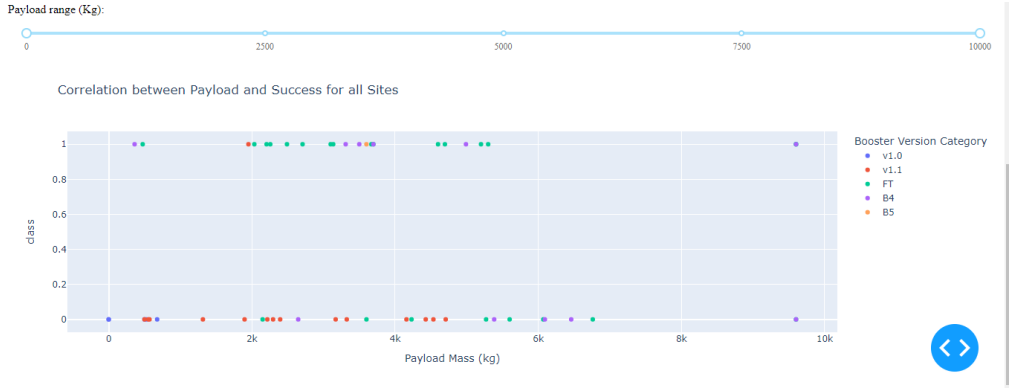
Total Success Launches by Site

Pie Chart for Launch Site with Highest Launch Success Ratio

- The pie chart demonstrates that the selected launch site has a very high success ratio, as evidenced by the dominant green segment.

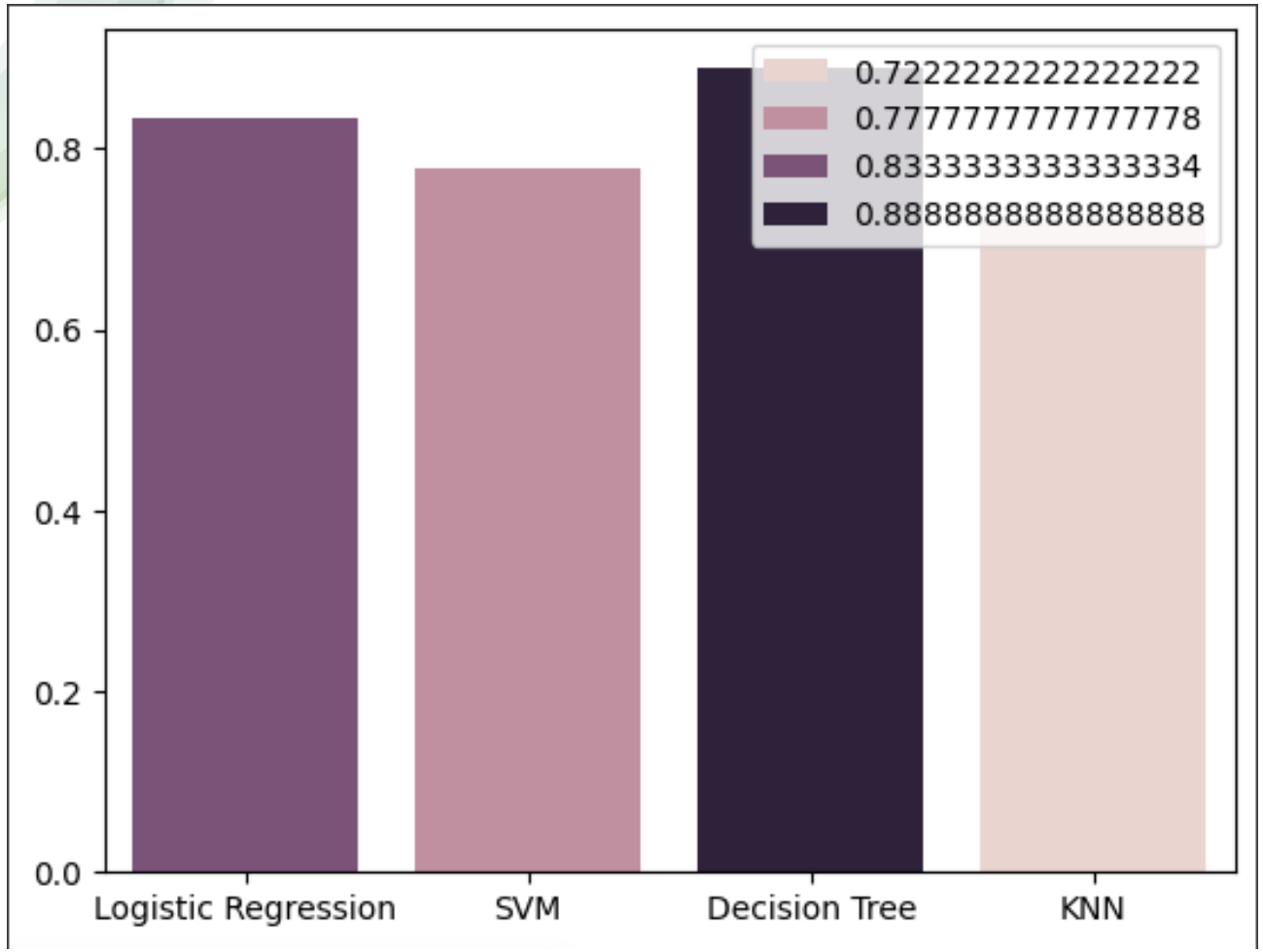
Payload vs. Launch Outcome Scatter Plot for All Sites

- The scatter plot reveals which payload ranges have the highest success rates. For example, payloads in the mid-range (e.g., 2000-5000 kg) might show a higher density of successful launches compared to very low or very high payloads.
- Booster Version Performance: Certain booster versions (e.g., FT or B5) may show a higher concentration of successful launches. This information is crucial for selecting the right booster for a given payload to maximize success chances.
- Critical Payload Thresholds: The scatter plot may reveal critical payload thresholds beyond which the success rate drops. Identifying these thresholds helps in planning and optimizing payloads for future launches.
- Overall Success Trends: A visual analysis of the scatter plot helps in understanding overall success trends across different sites and payload ranges. This can guide strategic decisions in launch planning and booster development.



Section 5

Predictive Analysis (Classification)



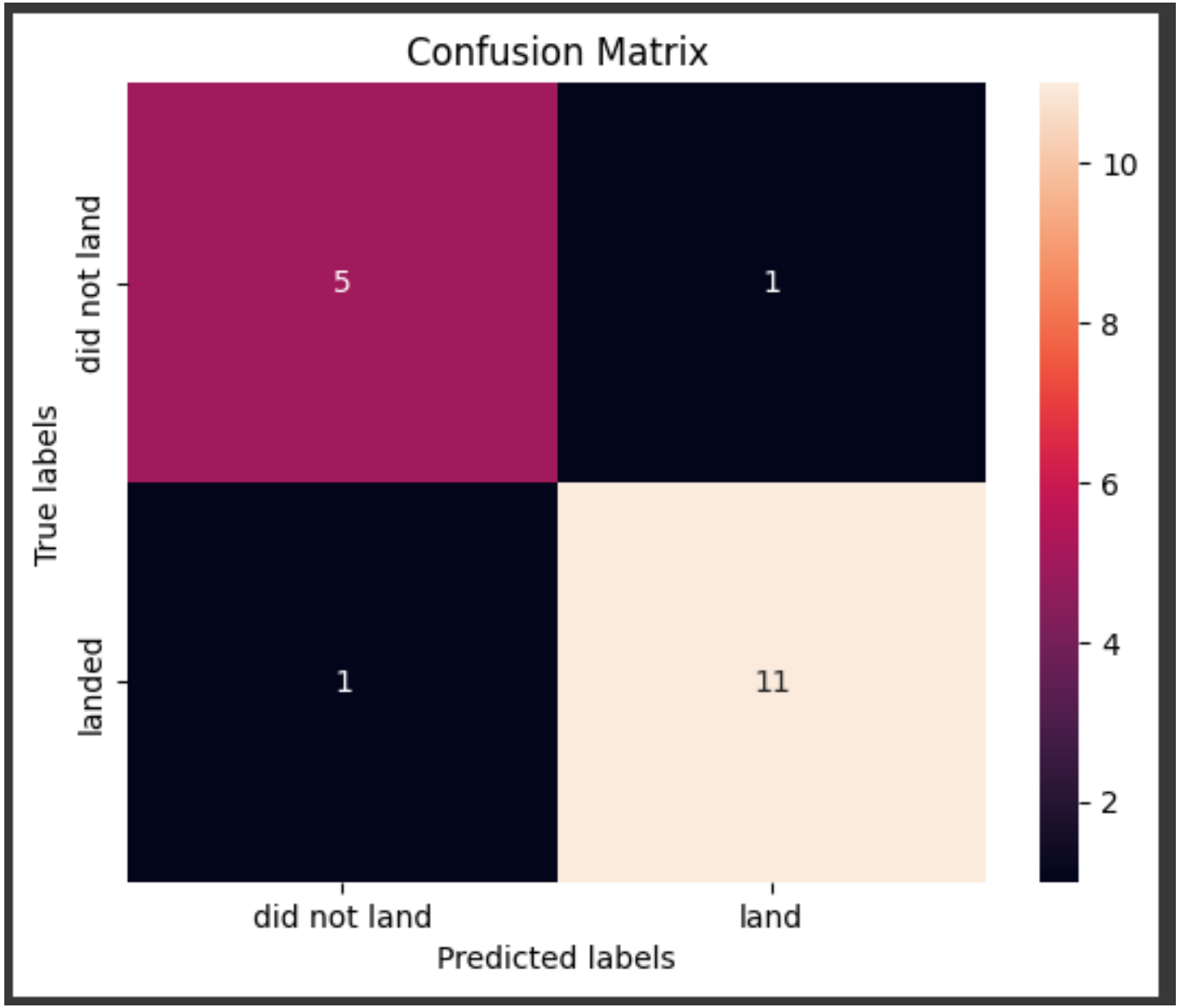
Classification Accuracy

Decision Tree has the highest accuracy with its accuracy at 0.89

Confusion Matrix

The confusion matrix of the best performing model

Decision Trees with an accuracy of 0.89



Conclusions

Data Wrangling and Acquisition:	• Successfully web scraped data and utilized the SpaceX API to gather comprehensive launch data. • Performed extensive data wrangling to clean and pre-process the dataset, ensuring accuracy and consistency.
Feature Engineering and Analysis:	• Conducted detailed feature extraction to identify crucial factors influencing launch success. • Utilized SQL for data analysis, providing insights into historical launch patterns and success rates.
Interactive Dash Application:	• Developed a Dash application to visualize key metrics and trends, offering an interactive platform for data exploration. • Enabled real-time predictions and analysis through a user-friendly interface.
Model Building and Evaluation	
Hyperparameter Tuning:	• Implemented GridSearchCV to fine-tune parameters for Logistic Regression, SVM, Decision Tree, and KNN models. • Achieved optimal model configurations through cross-validation, enhancing prediction accuracy.
Performance Metrics:	• Evaluated models using accuracy scores and confusion matrices, with clear visualization in the Dash application. • The [selected model] demonstrated superior accuracy, making it the most reliable predictor of launch success.

Thank you!