Winning Space Race with Data Science

Mirwais Parsa
6/10/2025

Outline

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

Executive Summary

Methodologies

- **Data Collection:** using SpaceX API and web scraping from Wikipedia.
- **Data Wrangling:** using the Pandas library and One-hot encoding for categorical variables.
- **Exploratory Data Analysis (EDA):** using SQL and interactive visualization techniques, including Folium and Plotly Dash.
- **Predictive Analysis:** Machine Learning Prediction using classification models.

Summary of all results

- The larger the flight amount at a launch site, the greater the success rate at that launch site.
- The greater the payload mass for launch site CCAFS SLC 40, the higher the success rate of the rocket.
- The decision tree classifier is the model with the highest classification accuracy.
- The launch success rate has increased from 2013 to 2020.

Introduction

Project background and context

- Space Y is a new entrant in the space industry, competing with SpaceX, a dominant force in space exploration.
- SpaceX offers Falcon 9 launches at $62M, while competitors charge $ 165M+, mainly due to non-reusable first-stage rockets.
- SpaceX's hardware reusability significantly lowers costs, giving them a market advantage.
- To remain competitive, Space Y must develop cost-saving strategies, including reusable hardware.

Problems and Objectives

- Space Y needs cost-effective solutions to remain competitive in the commercial space industry.
- This project aims to estimate launch costs and analyze SpaceX's strategies, providing data-driven insights for Space Y and other industry competitors.
- By extracting, cleaning, and analyzing SpaceX's data, the project will uncover strategic insights through exploratory data analysis (EDA), interactive visualization, and predictive modeling.
- Predicting first-stage landings will enhance pricing estimates and improve Space Y's market positioning.

Section 1

Methodology

Methodology

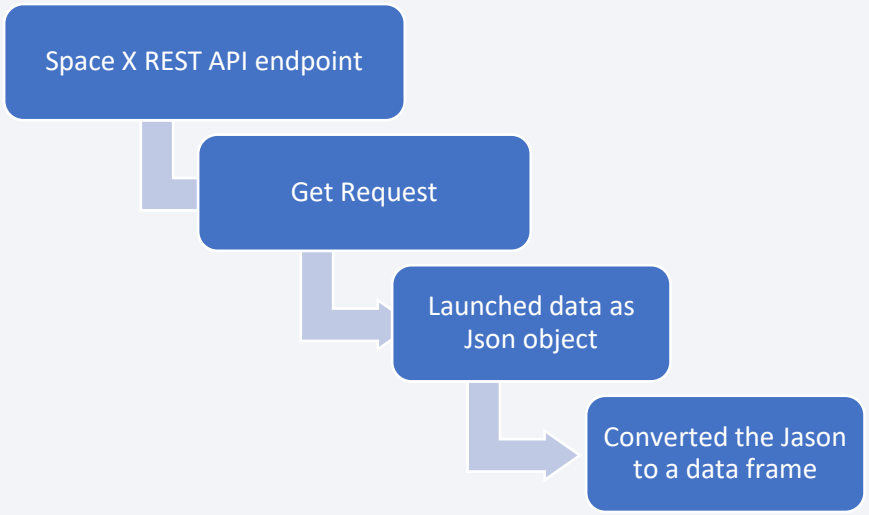
Executive Summary

- Data collection methodology:
  - Web scraping and API extraction techniques were used to collect SpaceX's historical launch data, including flight outcomes, launch costs, and environmental conditions.
- Perform data wrangling
  - The extracted raw data was processed by handling missing values, correcting inconsistencies, standardizing formats, and transforming variables using the Pandas library.
- Data Processing
  - Exploratory Data Analysis (EDA): Statistical techniques and visualizations, such as Folium and Plotly, were applied to uncover trends and correlations.
  - Predictive Modeling: Machine learning techniques such as classification, K-nearest mean, etc., were employed to forecast first-stage landing success based on historical patterns.

Data Collection

- The project gathered SpaceX launch data directly from SpaceX's website using the REST API, as well as scraped from the open-source Wikipedia content.
- We applied the following key steps in fetching the data:

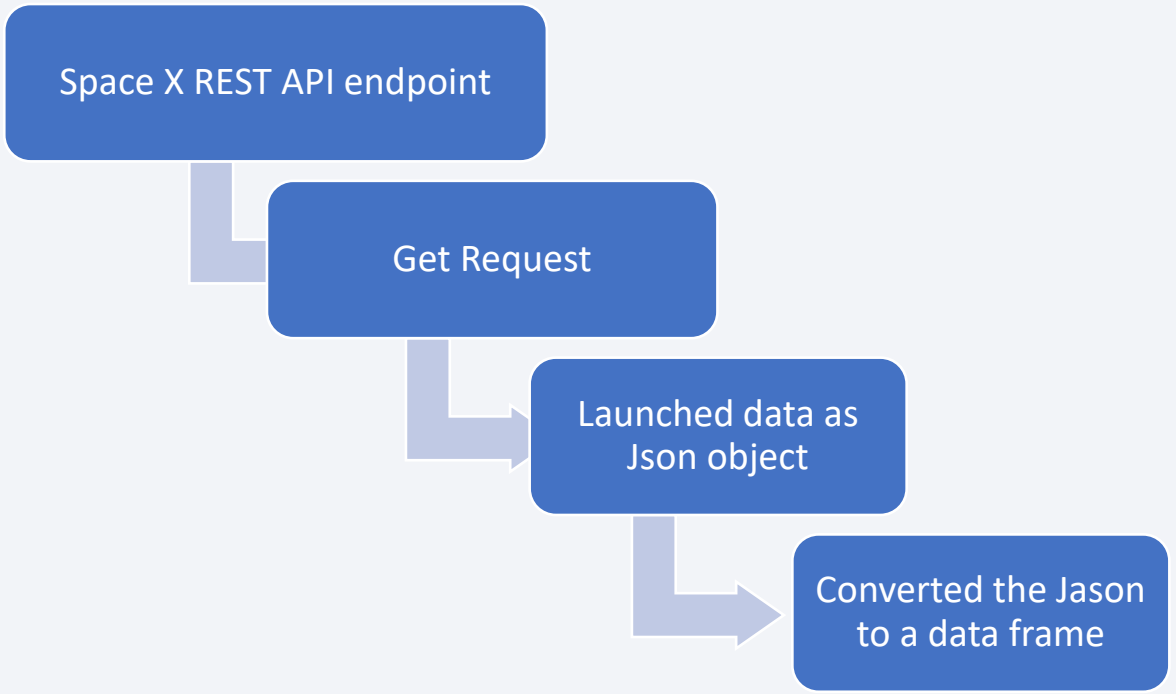
REST API



Web Scraping

Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Collecting%20SpaceX.ipynb



Data Collection - Scraping

- GitHub URL of the completed SpaceX Web Scrapping notebook: https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Collecting%20SpaceX.ipynb

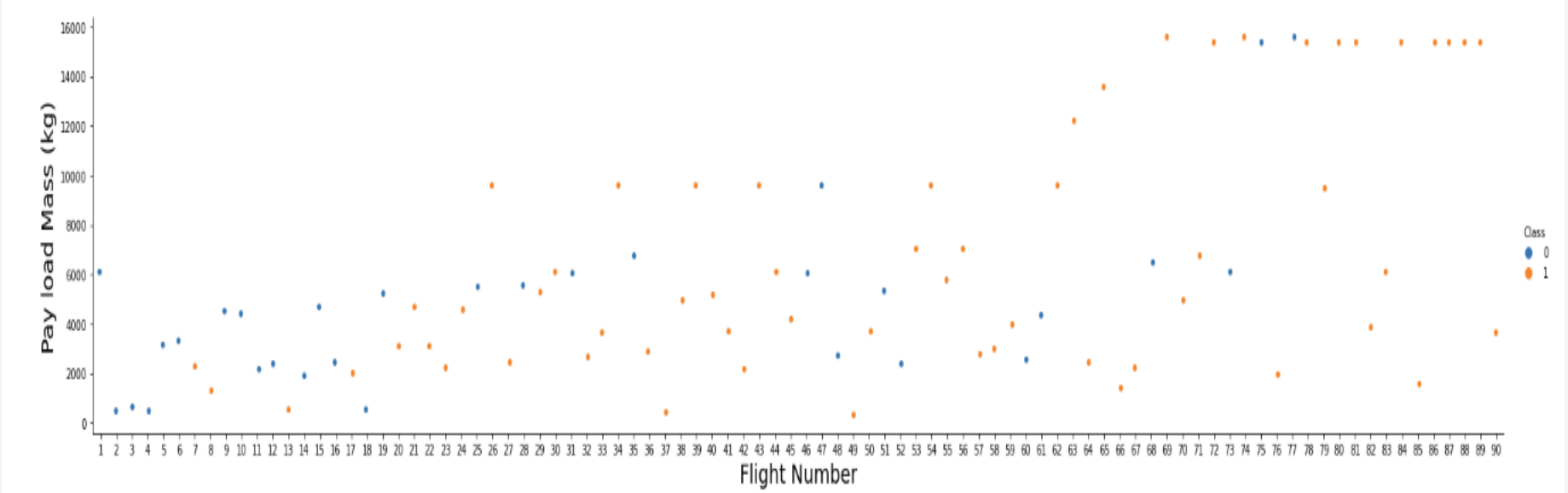
Data Wrangling

- Handled Missing Values using df.isnull() and replaced missings with averages.
- Duplicate Removal: Used df.drop_duplicates() to eliminate redundant rows
- Standardization & Normalization: Used MinMaxScaler() for scaling numerical data
- Categorical Encoding: Applied pd.get_dummies() and LabelEncoder() for non-numeric values
- Selected relevant columns using df[['column1', 'column2']]
- Stored cleaned data into .csv format using df.to_csv()

Here is the GitHub link to the notebook: https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Data%20Rangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and bar plots were used to visualize the relationship between a pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose names begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA(CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in the ground pad was achieved;
  - Names of the boosters that have been successful in drone ship and have a payload mass between 4000 and 6000 kg;
  - Total number of successful and failed mission outcomes;
  - Names of the booster versions that have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and• Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

- Markers, circles, lines, and marker clusters were used with Folium Maps.
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like the NASA Johnson Space Center;
  - Marker clusters indicate groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.

GitHub Link: https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Folium%20Map.ipynb

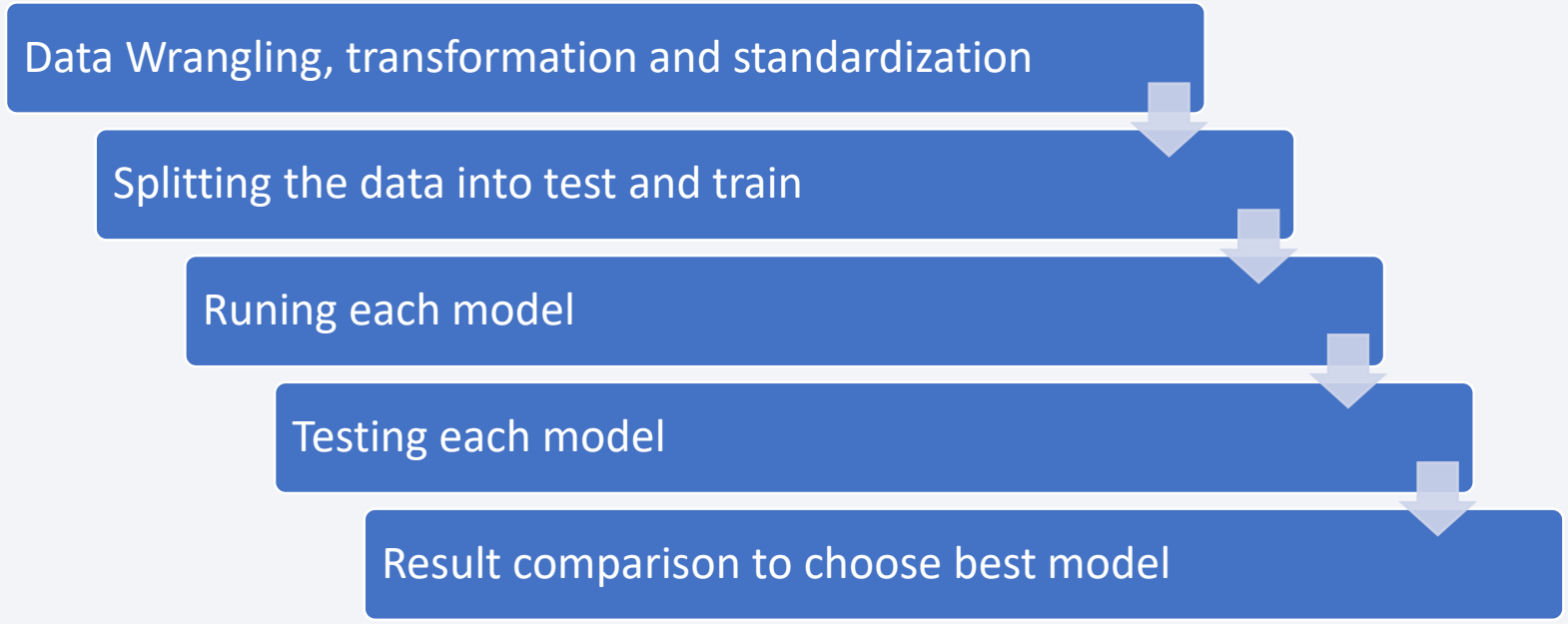
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed us to quickly analyze the relation between payloads
- and launch sites, helping to identify where is best place to launch according to payloads.

GitHub Link: https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Plotly_dashboard.ipynb

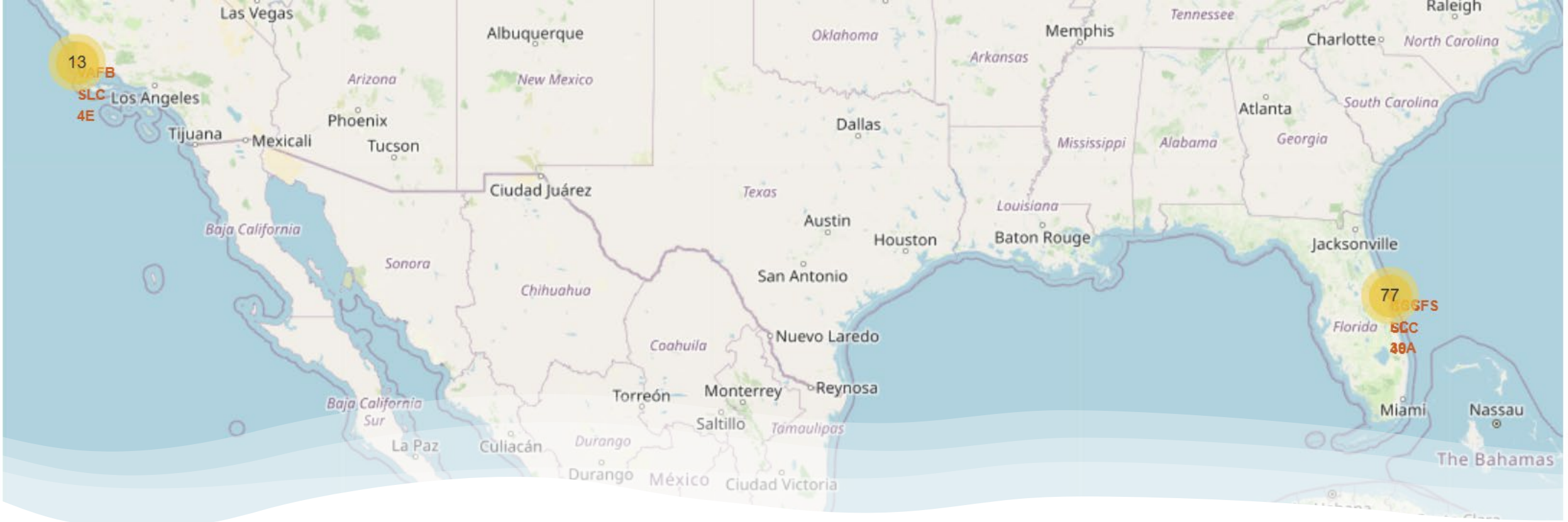
Predictive Analysis (Classification)

- In this project, we used four classification models for a comparative analysis: logistic regression, support vector machine, decision tree, and k-nearest neighbors.



Results

- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites;
  - The first launches were done by SpaceX itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first successful landing outcome happened in 2015, five years after the first launch;
  - Many Falcon 9 booster versions were successful at landing on drone ships, having a 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.

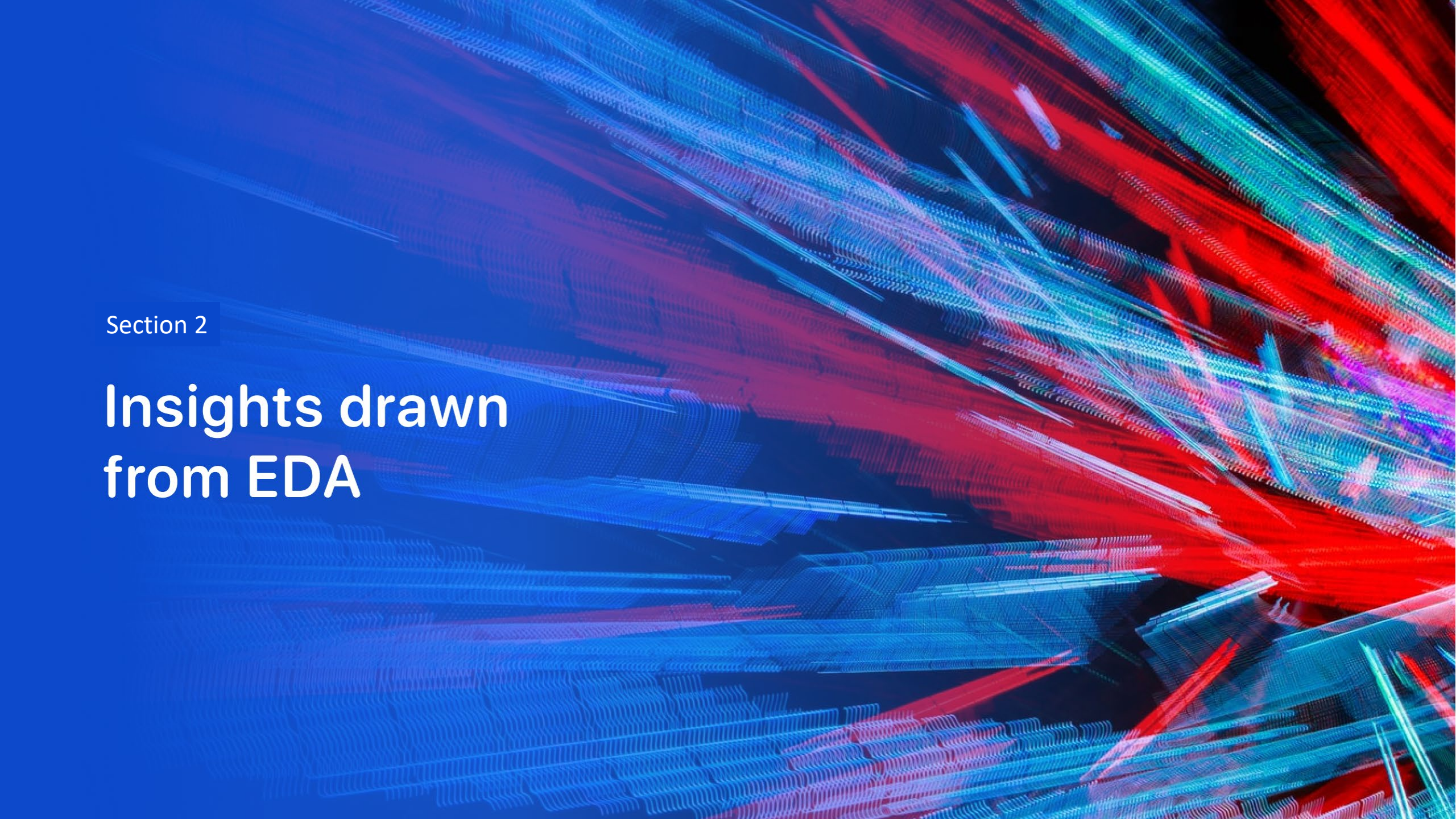


Results

- Using interactive analytics, as shown below, we were able to find out that launch sites were in:
  - Safety places, near the sea, for example, and have a good logistics infrastructure around.
  - Most launches happen at East Coast launch sites.

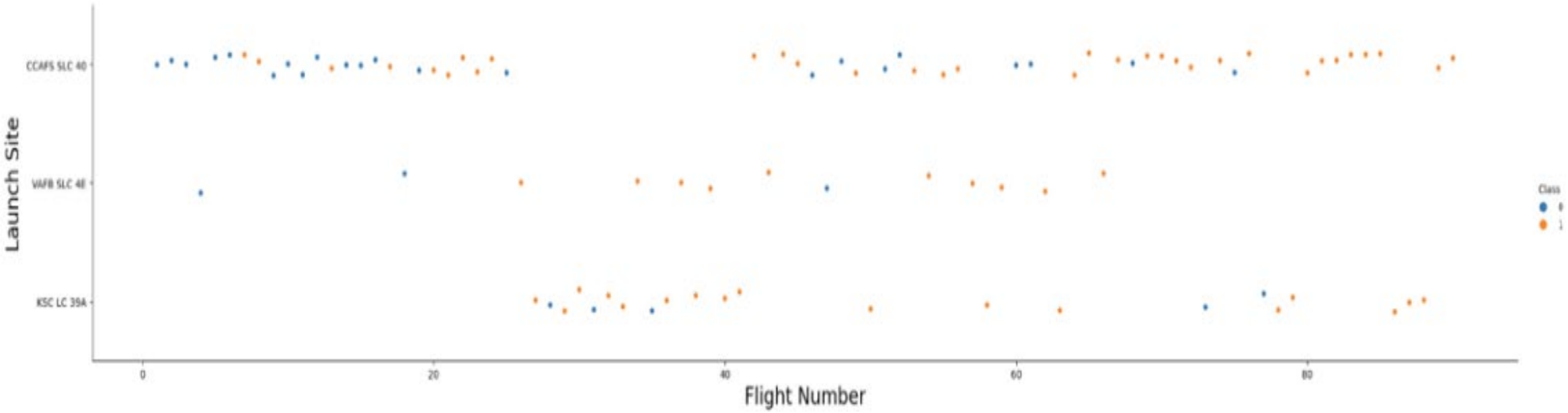
Results

- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites;
  - The first launches were done by SpaceX itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first successful landing outcome happened in 2015, five years after the first launch;
  - Many Falcon 9 booster versions were successful at landing on drone ships, having a 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.


Section 2
Insights drawn from EDA

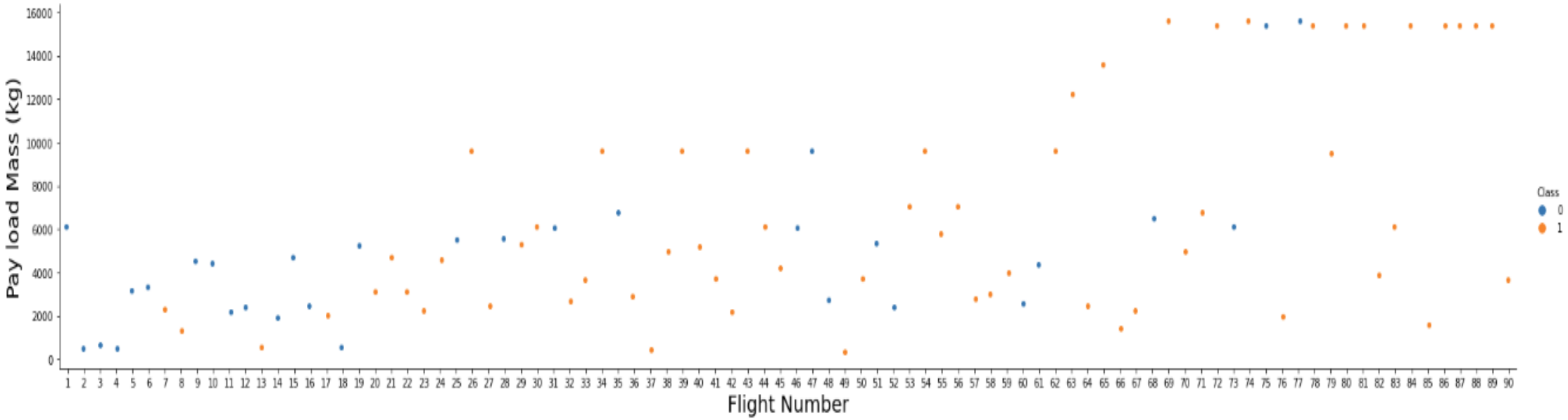
Flight Number vs. Launch Site

- As shown below, the best launch site is CCAF5 SLC 40, where most of the recent launches were successful;
- In second place comes VAFB SLC 4E and in the third is the KSC LC 39A;
- The overall success rate improved over time.



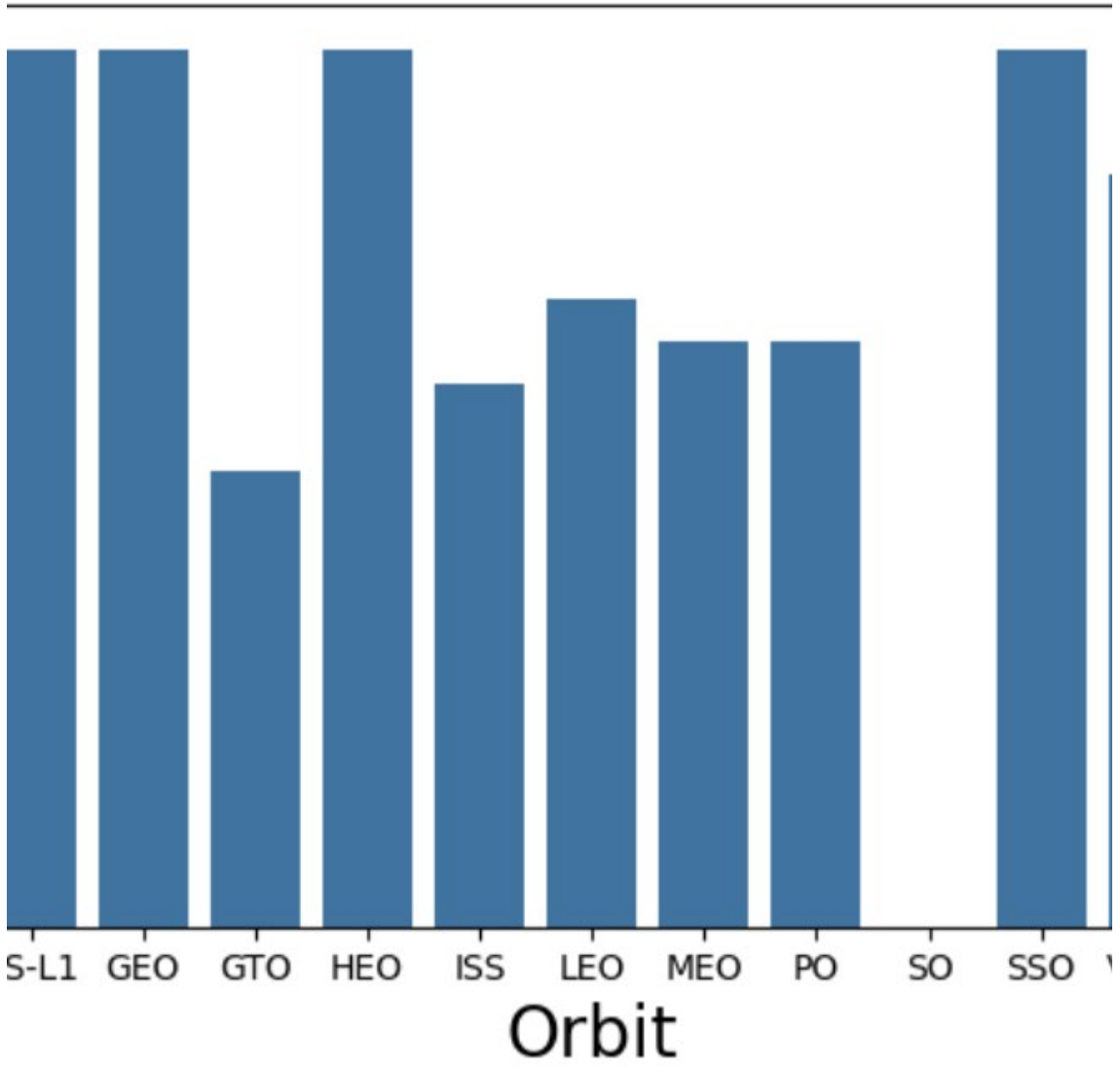
Payload vs. Launch Site

- As indicated in the plot, the number of launches from the CCAFS SLC 40 site is notably higher than those from other launch sites.



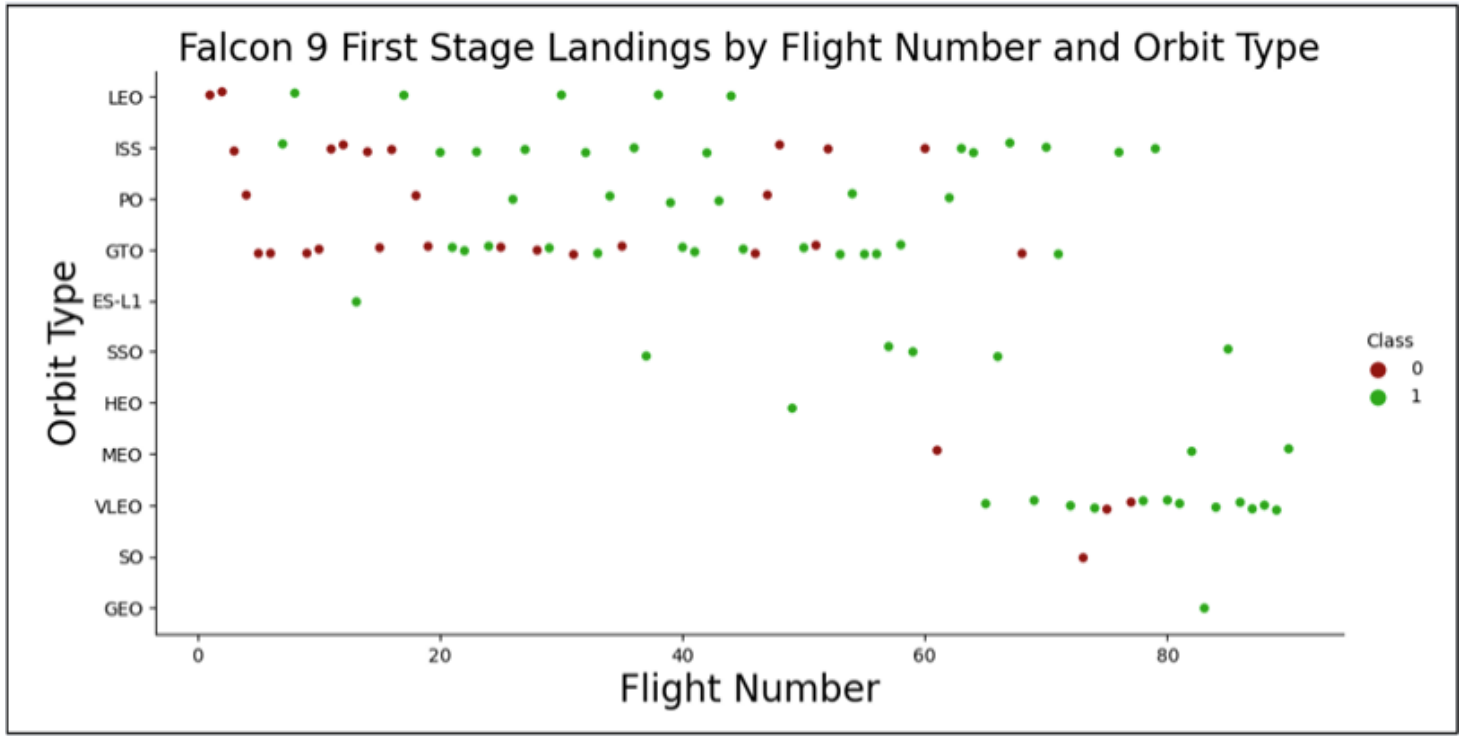
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO, and GEO orbits have no failed first-stage landings.
- SO orbits have no successful first-stage landings.



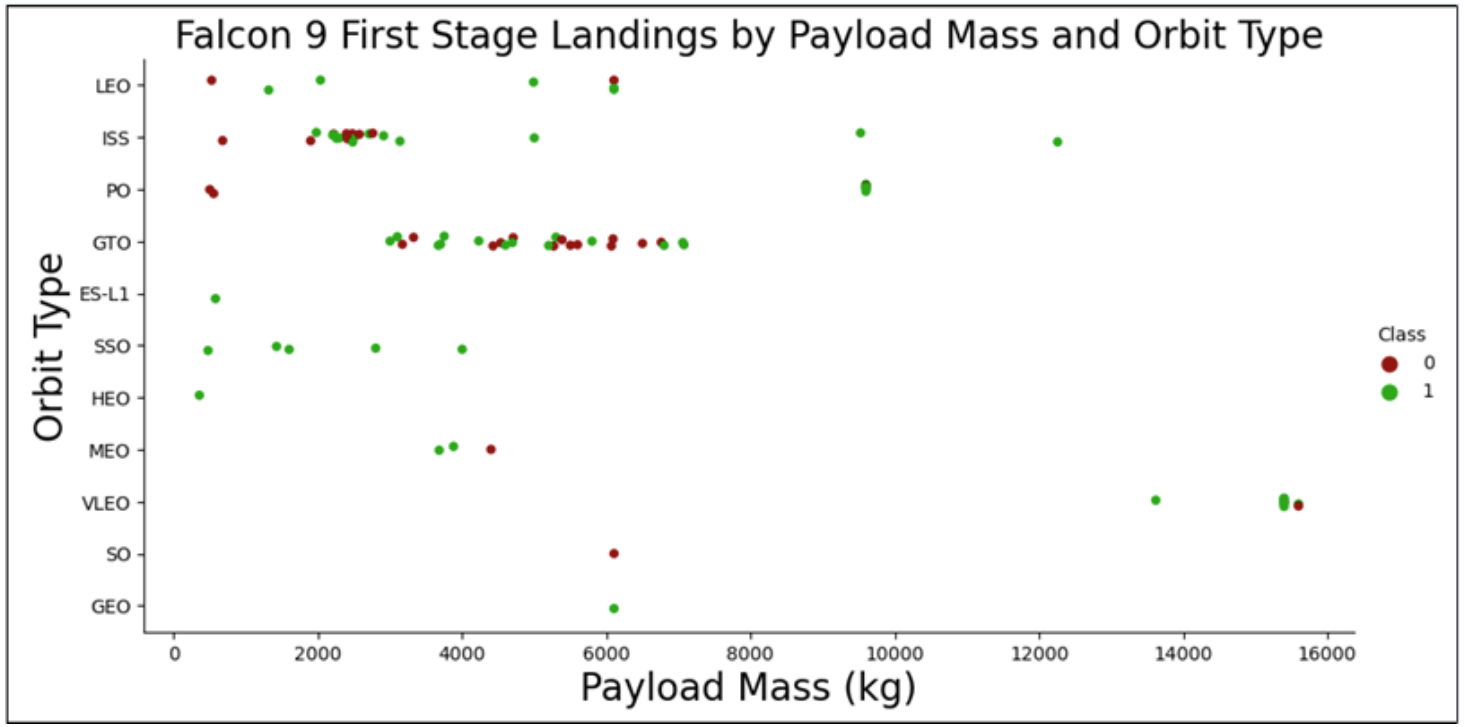
Flight Number vs. Orbit Type

- There is a positive correlation between flight number and success rate. (I.e., Larger flight numbers were associated with higher success rates.)



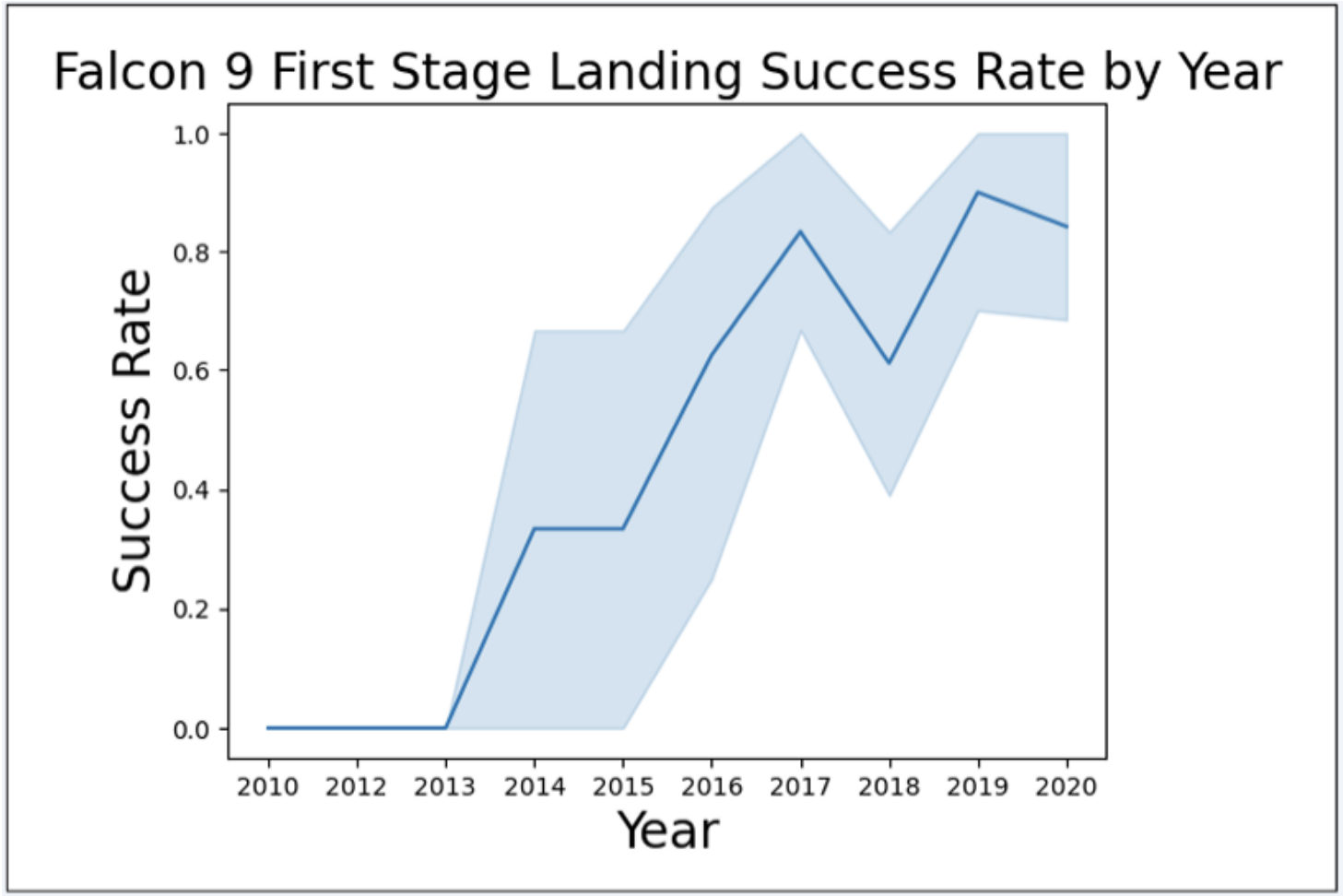
Payload vs. Orbit Type

- Some orbit types showed higher success rates than others.
- Success rate appeared to have no obvious correlation with payload mass.



Launch Success Yearly Trend

- The success rate of the Falcon 9 first-stage landings has increased significantly over the selected interval of years.



All Launch Site Names

I. Launch Site

II. CCAFS LC-40

III. CCAFS SLC-40

IV. KSC LC-39A

V. VAFB SLC-4E

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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Total payload carried by boosters from NASA: Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Total Payload (kg)
111.268

Average Payload Mass by F9 v1.1

- The Average payload mass carried by booster version F9 v1.1:
- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

Avg Payload (kg)
2.928

First Successful Ground Landing Date

- By filtering data by successful landing outcome on the ground pad and getting the minimum value for the date, it’s possible to identify the first occurrence, which happened on 12/22/2015.

Min Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Selecting distinct booster versions according to the following filters: payload mass>4000 & <6000.

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

Total Number of Successful and Failure Mission Outcomes

Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

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

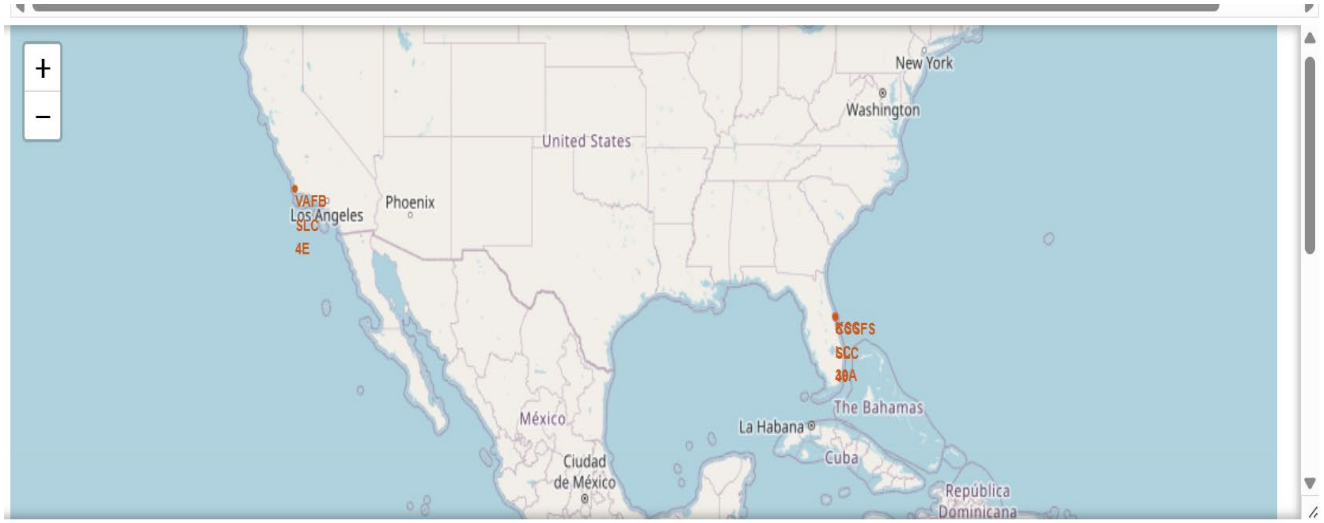
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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

Section 3

Launch Sites Proximities Analysis

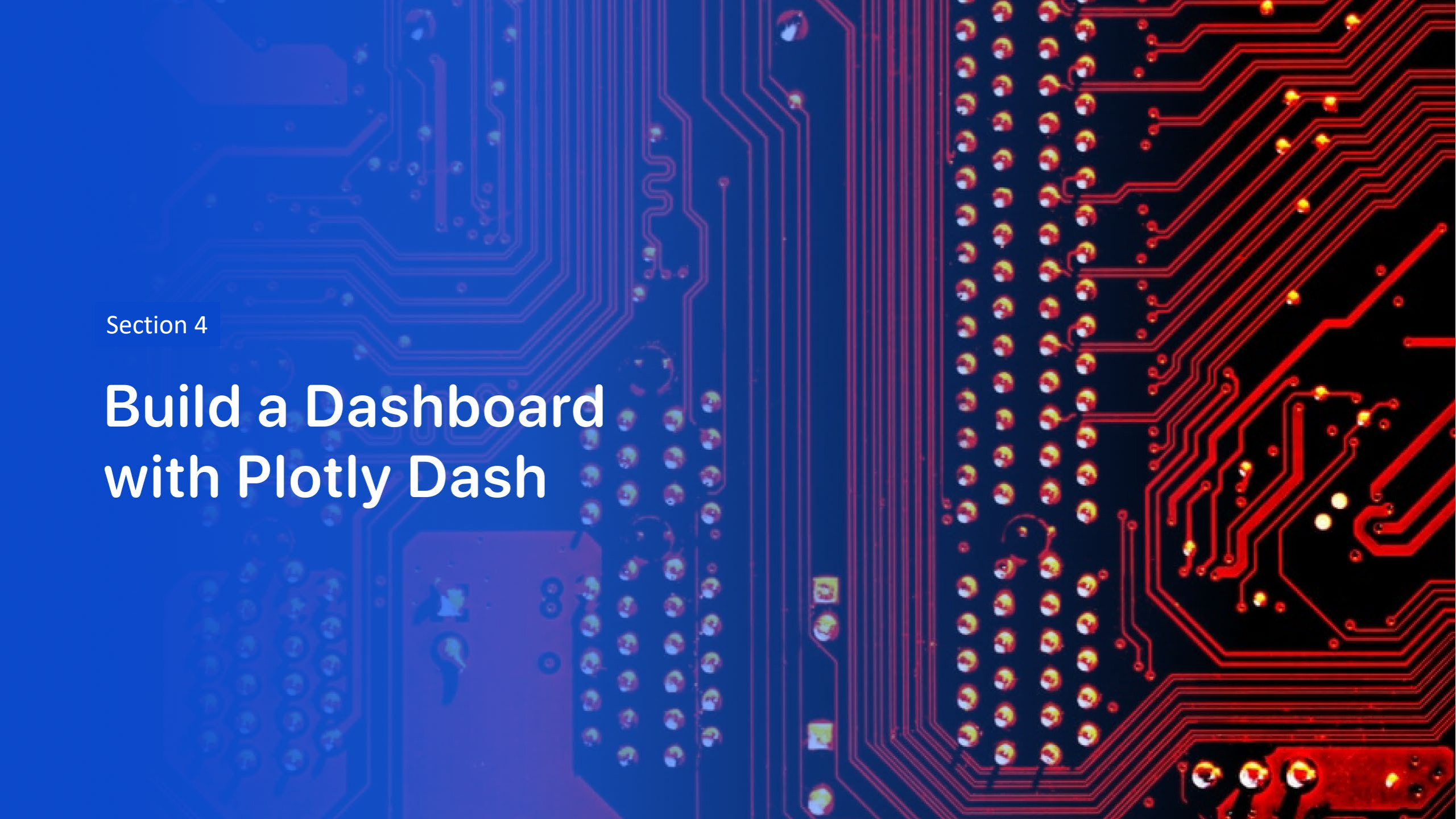
Falcon 9 Launch Site Locations



- California, USA
  - VAFB SLC-4E | Vandenberg Air Force Base Space Launch Complex 4E.
- Florida, USA
  - KSC LC-39A | Kennedy Space Center Launch Complex 39A.
  - CCAFS LC-40 | Cape Canaveral Air Force Station Launch Complex 40.
  - CCAFS SLC-40 | Cape Canaveral Air Force Station Space Launch Complex 40.

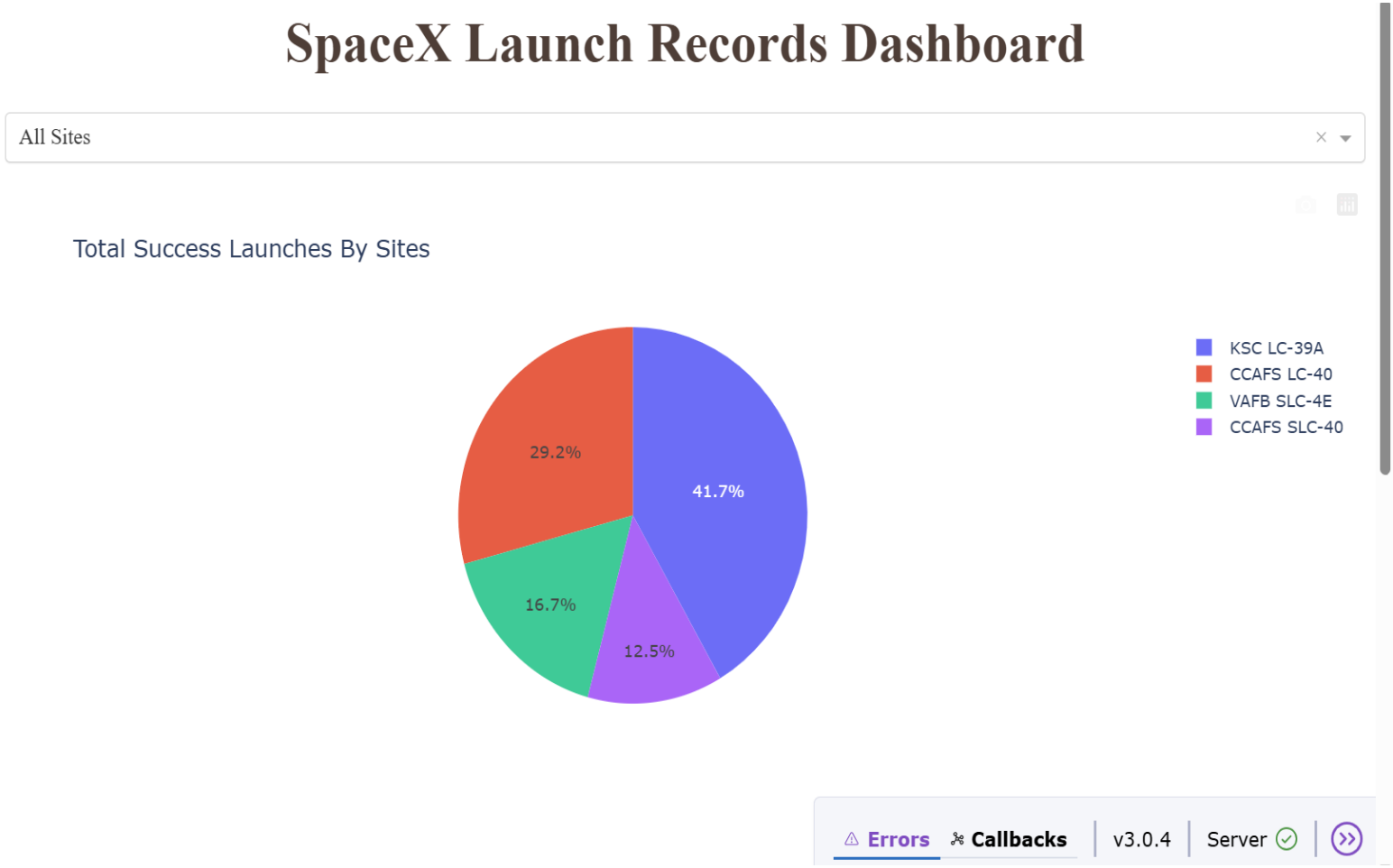
Falcon 9 Launch Site Locations With Circles

- California, USA
  - VAFB SLC-4E | Vandenberg Air Force Base Space Launch Complex 4E.
- Florida, USA
  - KSC LC-39A | Kennedy Space Center Launch Complex 39A.
  - CCAFS LC-40 | Cape Canaveral Air Force Station Launch Complex 40.
  - CCAFS SLC-40 | Cape Canaveral Air Force Station Space Launch Complex 40.


Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site

- The Pie chart below displays the distribution of successful Falcon 9 first-stage landing outcomes across different launch sites.
- The highest share of successful Falcon 9 first-stage landing outcomes (@ 41.7% of the total) occurred at KSC LC-39A.



Correlation between Success and Payload: All sites

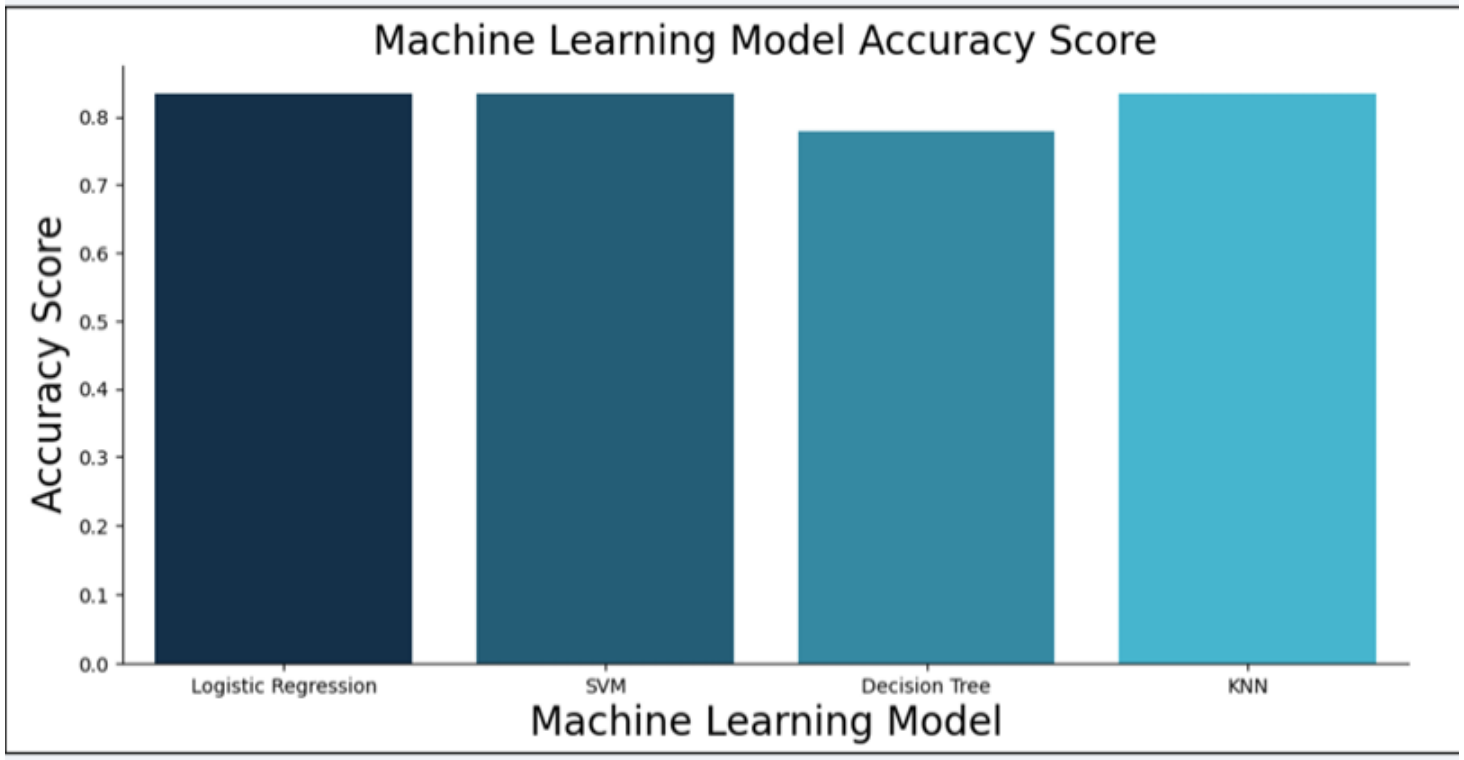
- Scatterplots of payloads for all sites, with different payloads selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The ‘FT’ booster version category has the largest success rate.

Section 5

Predictive Analysis (Classification)

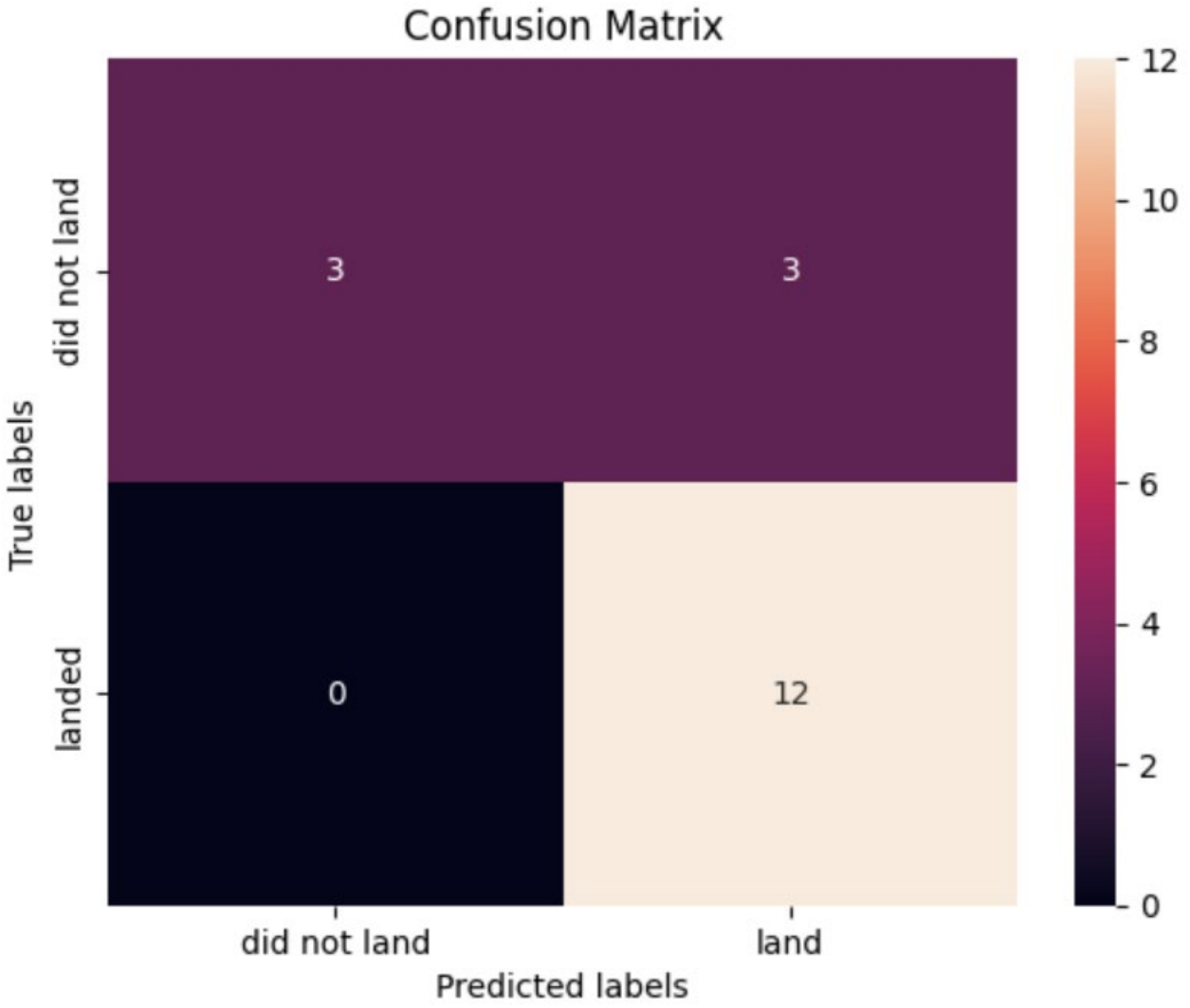
Classification Accuracy

- The result does not indicate any significant difference in statistical accuracy among other models.



Confusion Matrix

- As indicated in the confusion matrix, logistic regression had perfect precision (no false positives) but moderate recall (some false negatives).
- So the model is great at identifying positives correctly, but misses some true positive cases.



Conclusions

Data Collection: Retrieved SpaceX rocket launch data via API and auxiliary functions.

Data Wrangling: Filtered dataset to include only Falcon 9 launches.

- Handled missing values in PayloadMass using mean imputation.
- Created landing outcome labels from the Outcome column.

Exploratory Data Analysis (EDA): Analyzed launch frequency across features.

- Calculated landing success rate.
- Identified trends and correlations via visualizations.

Interactive Visual Analytics

- Built Folium maps to visualize launch sites and success rates.
- Developed a Plotly Dash interactive dashboard with input controls and visualizations.

Predictive Modeling

- Preprocessed data: One-hot encoding for categorical variables, normalization with StandardScaler.
- Model training & evaluation using classification models from Scikit-learn.
- Hyperparameter tuning with GridSearchCV.
- Performance - assessment using confusion matrix, tables, and charts.

Appendix

1. Data References

- SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json • Wikipedia (Webpage): https:// en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- SpaceX (CSV): https ://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321ENSkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetworkChannel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- Launch Dash (CSV): https:// cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

2. Python Jupyter Notebooks

- https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Collecting%20SpaceX.ipynb
- https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb
- https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/EDA%20and%20visualization.ipynb
- https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Folium%20Map.ipynb
- https://github.com/M-Parsa/IBM-Data-Science-Capstone/blob/main/Plotly_dashboard.ipynb

Thank you!