IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christoph
21.04.2024

Outline

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

Executive Summary

- Various methodologies were employed to examine the data:
  - Data was gathered through web scraping and the SpaceX API.
  - Exploratory Data Analysis (EDA) encompassed data manipulation, visualization, and interactive analytics.
  - Machine Learning was utilized for predictive analysis.
- A comprehensive summary of all findings was compiled:
  - Valuable data was acquired from publicly available sources.
  - EDA facilitated the identification of key features for predicting launch success.
  - Machine Learning Prediction determined the optimal model for discerning crucial characteristics that influence success, leveraging all available data

Introduction

- The aim is to assess the potential of the emerging company Space Y in competing with Space X:
- Desired outcomes include:
  - Determining the most effective method to estimate the overall launch costs by forecasting the successful landings of the initial stage of rockets.
  - Identifying the optimal location for conducting launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was acquired from two distinct sources:
  - The SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_)
- Perform data wrangling
  - The gathered data was enhanced by generating a landing outcome classification derived from the outcome data subsequent to feature summarization and analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - The data collected up to this stage underwent normalization, segmentation into training and test datasets, and evaluation through four distinct classification models. The accuracy of each model was assessed using various parameter combinations.

Data Collection

- Datasets were gathered from two sources:
  - the Space X API (https://api.spacexdata.com/v4/rockets/)
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) by utilizing web scraping techniques.

Data Collection – SpaceX API

- SpaceX provides a public API for accessing data.
- The API was utilized following the flowchart depicted alongside, and the obtained data was subsequently stored.

Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Retrieve data from the SpaceX API and parse the information regarding SpaceX launches

↓

Refine the dataset to exclusively encompass Falcon 9 launches.

↓

Handle any missing values in the dataset

Data Collection - Scraping

- Information regarding SpaceX launches is also available on Wikipedia.
- Data is retrieved from Wikipedia following the outlined flowchart and then stored for further use.

Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Retrieve the Falcon 9 Launch Wikipedia page

↓

Extract the names of all columns/variables from the HTML table header

↓

Generate a dataframe by parsing the HTML tables containing launch data

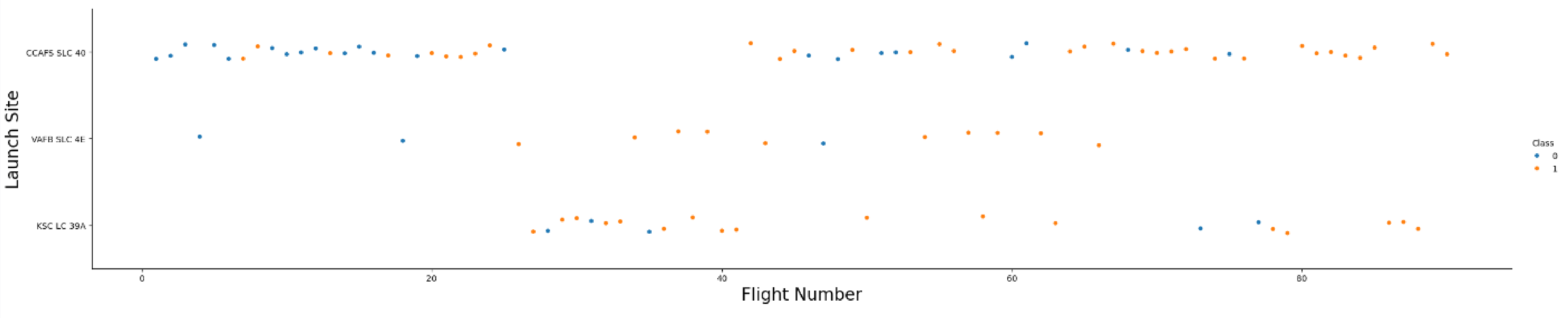
Data Wrangling

- Initially, exploratory data analysis (EDA) was conducted on the dataset.
- Subsequently, summaries were generated for launches per site, occurrences of each orbit, and occurrences of mission outcomes per orbit type.
- Finally, the landing outcome label was derived from the Outcome column.

Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatterplots and bar plots were employed to visualize the relationships between pairs of features during the data exploration phase
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- The following SQL queries were executed:
  - Retrieve the names of unique launch sites in the space mission.
  - Identify the top 5 launch sites whose names begin with the string 'CCA'.
  - Calculate the total payload mass carried by boosters launched by NASA (CRS).
  - Determine the average payload mass carried by booster version F9 v1.1.
  - Find the date when the first successful landing outcome on a ground pad was achieved.
  - List the names of boosters that achieved success on a drone ship and carried a payload mass between 4000 and 6000 kg.
  - Count the total number of successful and failed mission outcomes.
  - Identify the booster versions that carried the maximum payload mass.
  - Retrieve information on failed landing outcomes on a drone ship, including booster versions and launch site names, for the year 2015.
  - Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20.

Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Various visual elements were utilized with Folium Maps:
  - Markers: Represent points such as launch sites
  - Circles: Highlight specific areas around particular coordinates, such as NASA Johnson Space Center
  - Marker clusters: Group events at each coordinate, such as launches at a launch site
  - Lines: Indicate distances between two coordinates

Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/6-jupyter_launch_site_location_folium.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 facilitated a rapid analysis of the relationship between payloads and launch sites, aiding in the identification of the optimal launch location based on payload requirements

Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models—logistic regression, support vector machine, decision tree, and k-nearest neighbors—were evaluated and compared

Preparation and standardization of the data → Testing each model with various combinations of hyperparameters → Comparing the results obtained from each model.

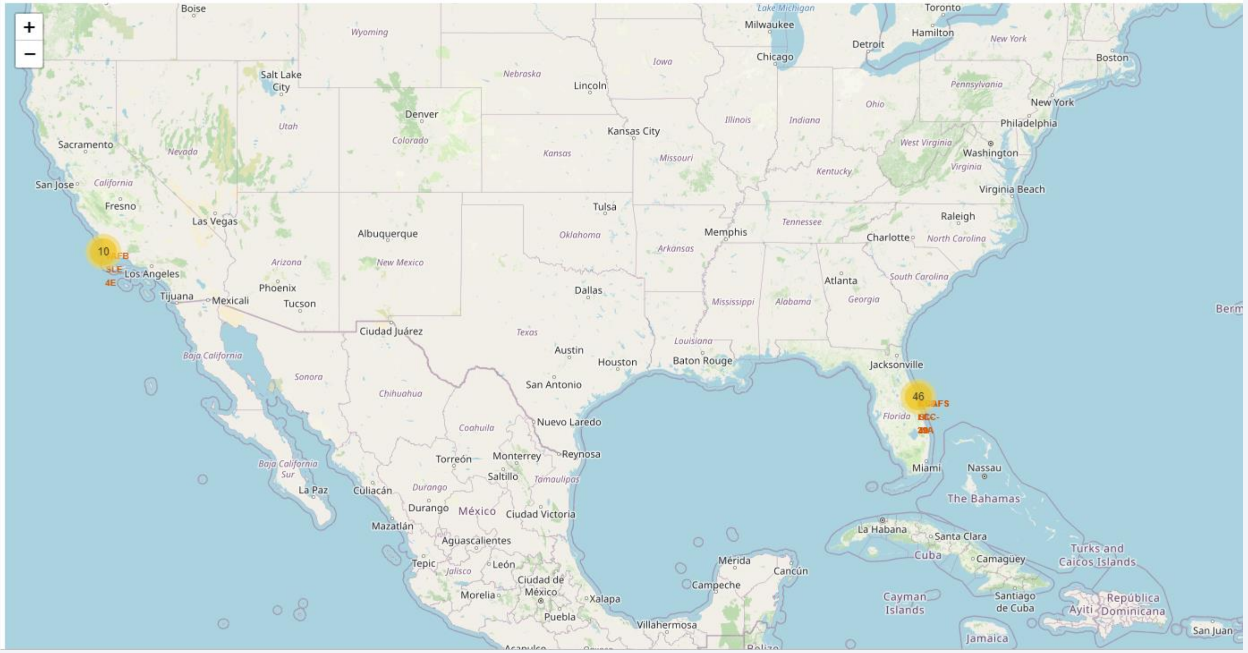
Github Url: https://github.com/CPblue93/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:
  - SpaceX operates from 4 distinct launch sites
  - Initial launches were conducted for SpaceX itself and NASA
  - The average payload of the F9 v1.1 booster is 2,928 kg
  - The first successful landing outcome occurred in 2015, five years after the first launch
  - Many Falcon 9 booster versions achieved successful landings on drone ships, particularly those with payloads exceeding the average
  - Nearly 100% of mission outcomes were successful
  - Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015
  - The success rate of landing outcomes improved over the years

Results

- Interactive analytics demo in screenshots:
  - Through interactive analytics, it was observed that launch sites tend to be situated in secure locations, often near the sea, and boast strong logistical infrastructure
  - The majority of launches occur at launch sites along the east coast.



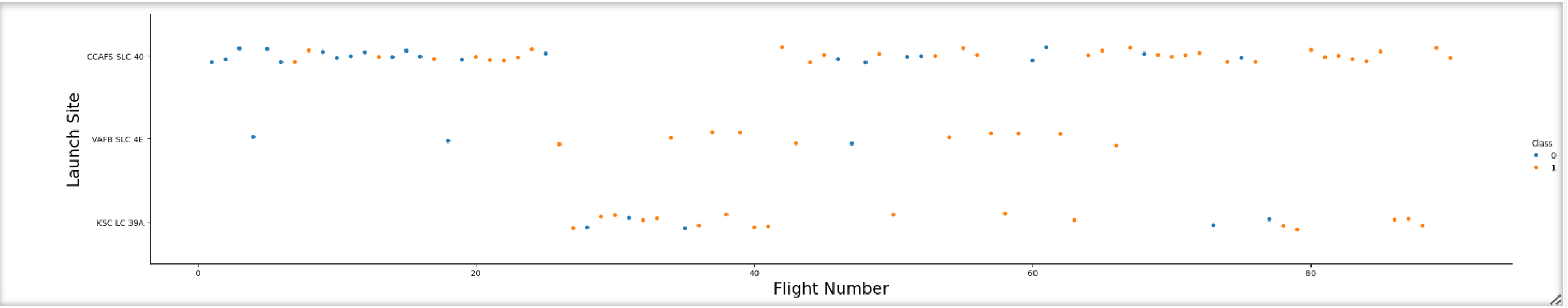
Results

- Predictive analysis results:
  - Predictive analysis indicated that the Decision Tree Classifier emerged as the optimal model for predicting successful landings, boasting an accuracy rate exceeding 87%. Moreover, its accuracy for test data surpassed 94%.

Section 2

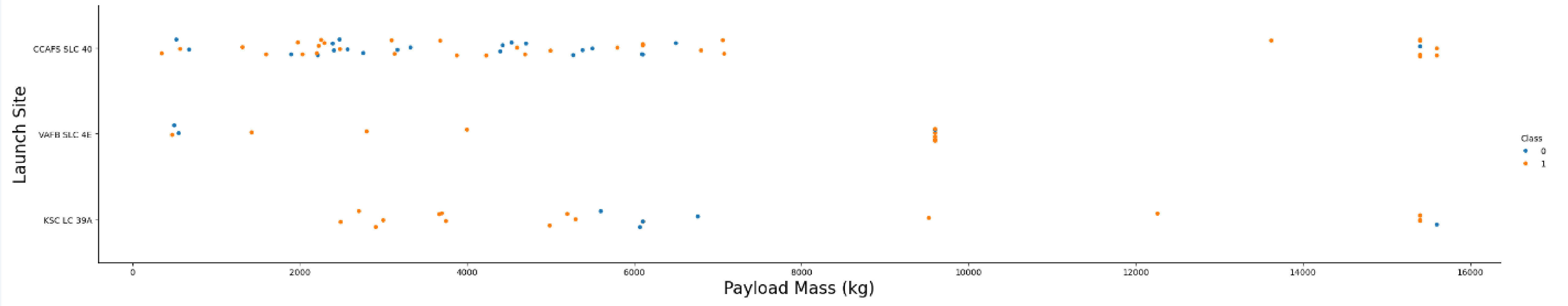
Insights drawn from EDA

Flight Number vs. Launch Site



- Based on the plot above, it's evident that the premier launch site presently is CCAF5 SLC 40, with the majority of recent launches culminating in success
- Following closely behind are VAFB SLC 4E in second place and KSC LC 39A in third place
- Additionally, there's a discernible trend showcasing an overall improvement in the success rate over time.

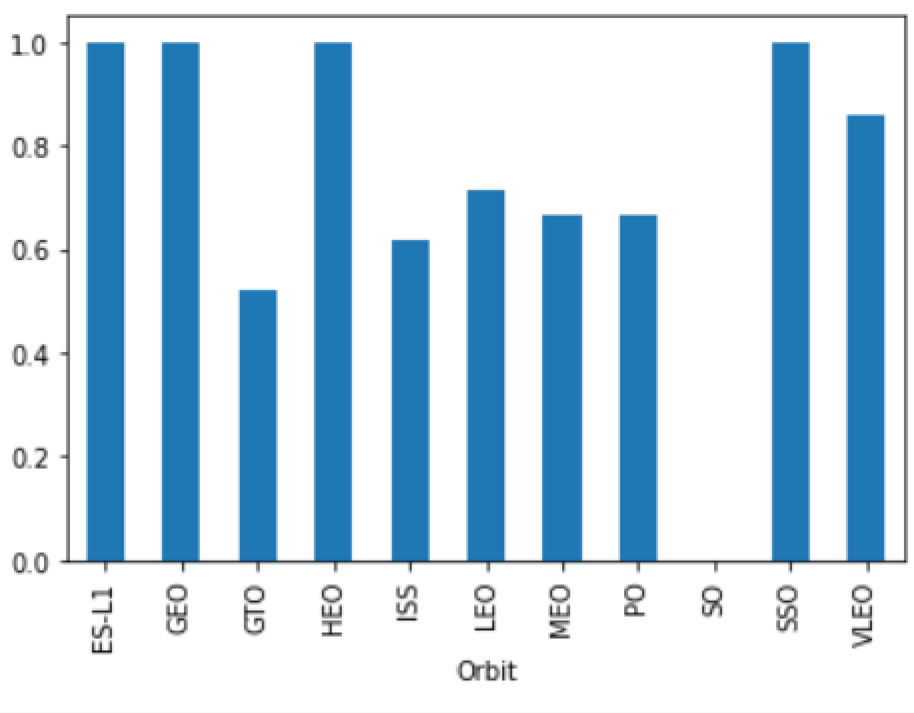
Payload vs. Launch Site



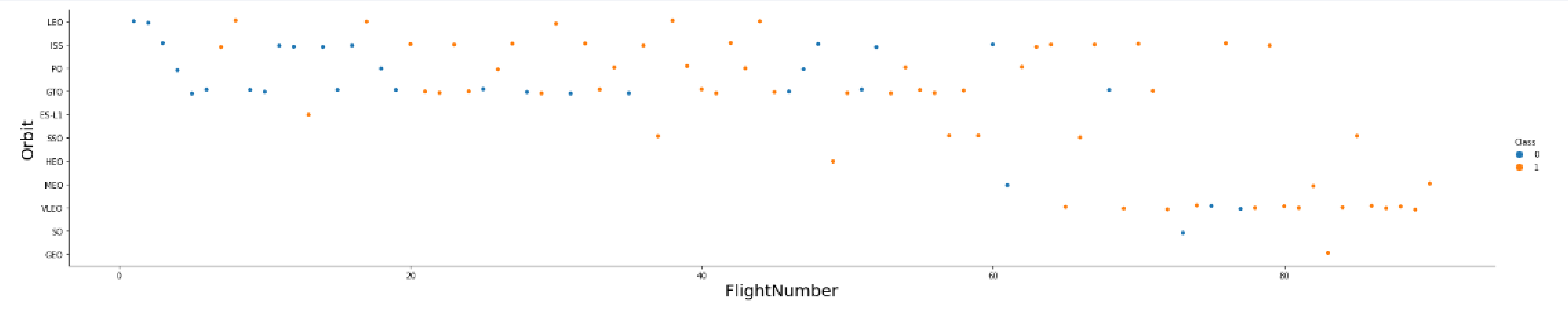
- Payloads exceeding 9,000 kg (roughly equivalent to the weight of a school bus) exhibit an outstanding success rate
- Payloads surpassing 12,000 kg appear feasible primarily at CCAFS SLC 40 and KSC LC 39A launch sites

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES L1
  - GEO
  - HEO
  - SSO
- In addition:
  - VLEO missions boast a success rate exceeding 80%.
  - LEO missions exhibit a success rate above 70%.

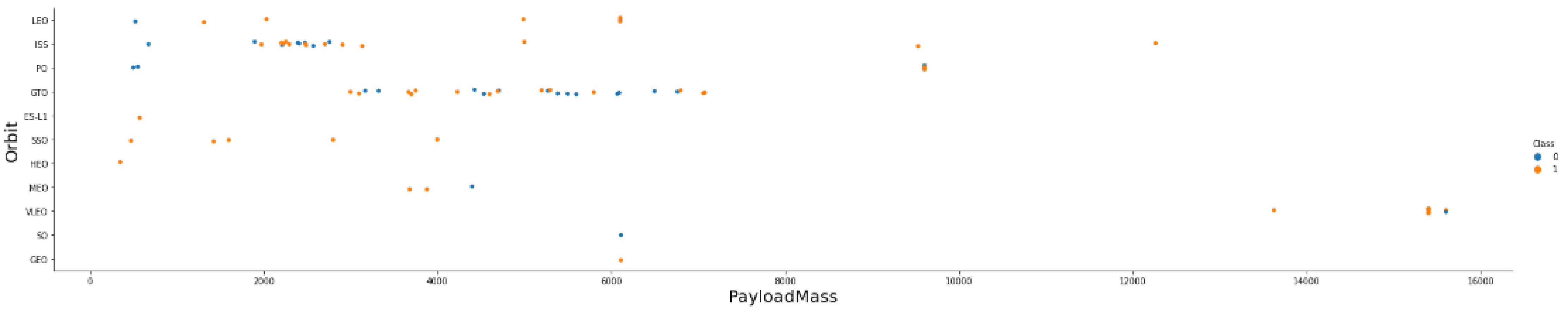


Flight Number vs. Orbit Type



- It appears that the success rate has improved over time for all orbits
- The recent uptick in the frequency of VLEO missions suggests it as a burgeoning business opportunity

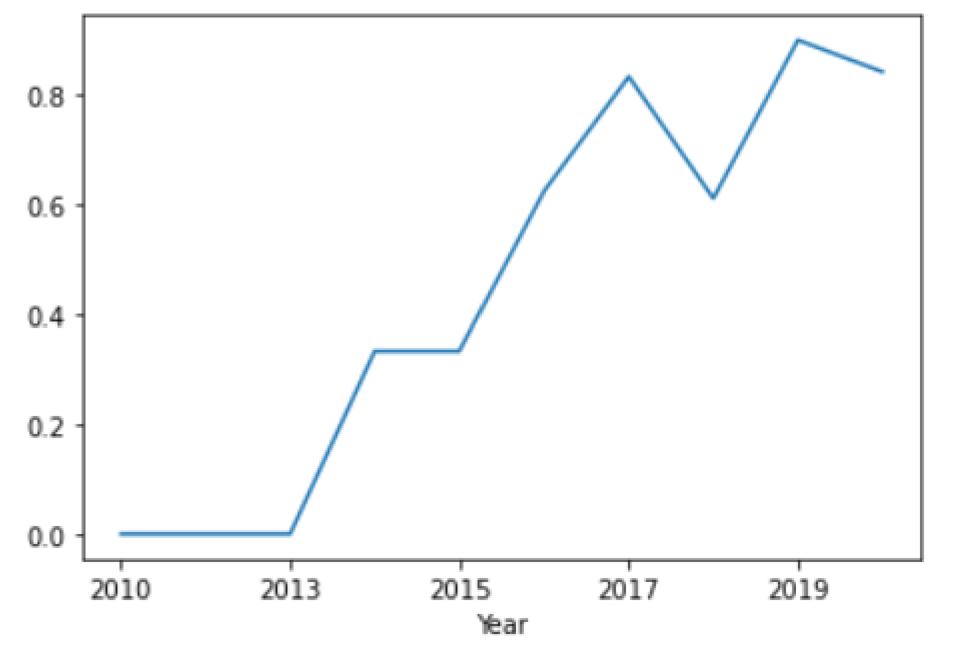
Payload vs. Orbit Type



- There seems to be no discernible correlation between payload and success rate for GTO missions.
- The ISS orbit boasts the widest range of payloads and a commendable success rate.
- There are relatively few launches to SO and GEO.

Launch Success Yearly Trend

- The success rate began its ascent in 2013 and continued to rise until 2020
- It appears that the initial three years served as a period of adjustment and technological refinement



All Launch Site Names

- Based on the data, there are four distinct launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They are derived by selecting unique instances of "launch_site" values from the dataset

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attemp

- Here are five samples of Cape Canaveral launches.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- The total payload was calculated by summing all payloads associated with codes containing 'CRS', which corresponds to NASA missions

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- By filtering the data using the booster version provided and calculating the average payload mass, we obtained a value of 2,928 kg

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad:

Min Date
2015-12-22

- Filtering the data by successful landing outcomes on a ground pad and retrieving the minimum date value, we can identify the first occurrence, which happened on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- List 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 B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Applying the filters mentioned above and selecting distinct booster versions yields the following four results

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes:

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

- Grouping mission outcomes and counting the records for each group resulted in the summary provided above

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

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

- The following are the boosters that have carried the maximum payload mass recorded in the dataset

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

- The list above contains only two occurrences

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

- This data view alerts us that "No attempt" must be taken into account

Section 3

Launch Sites Proximities Analysis

All launch sites

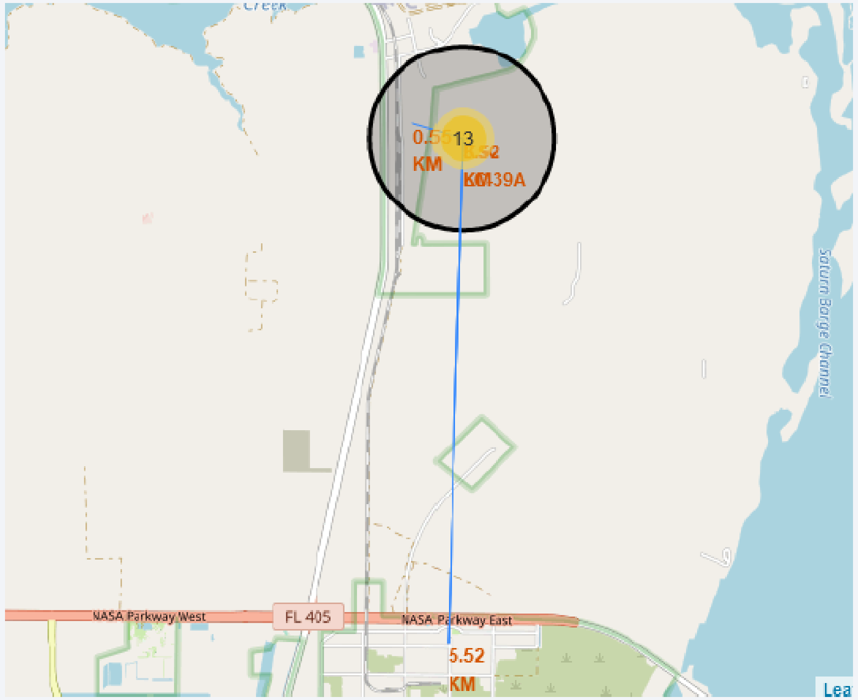
- The launch sites are situated near the sea, likely for safety reasons, while also being reasonably close to roads and railroads

Launch Outcomes by Site

- Example of KSC LC 39A launch site launch outcomes:

- Green markers indicate successful and red ones indicate failure

Logistics and Safety

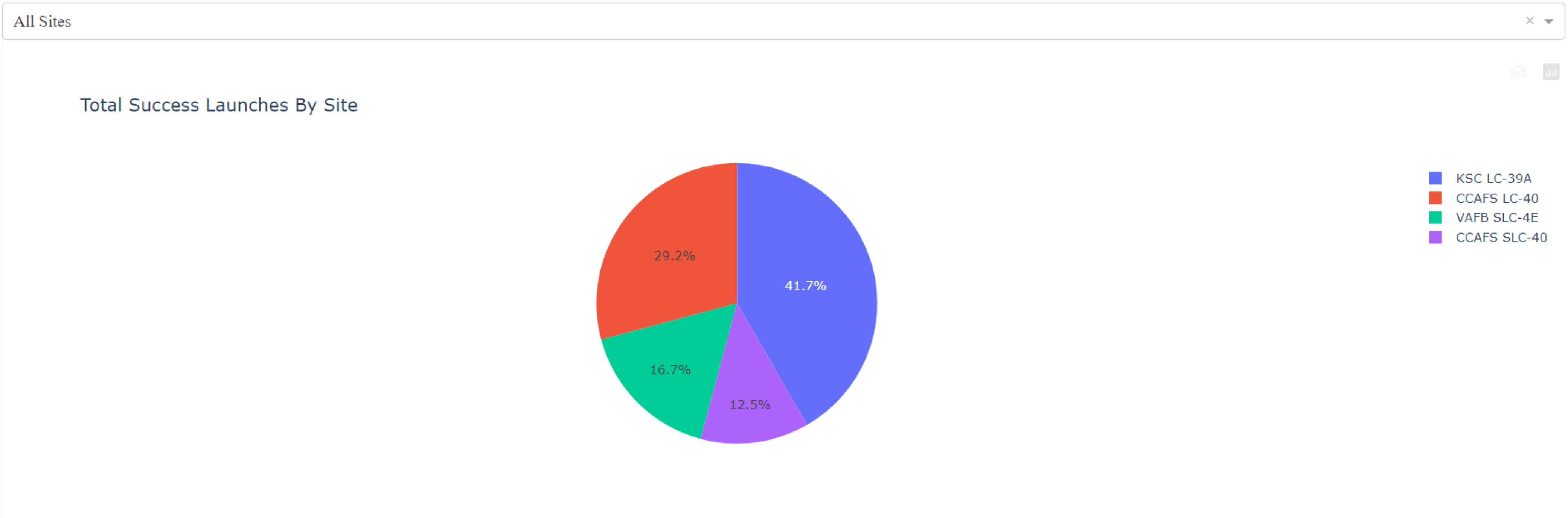


- The launch site KSC LC 39A boasts favorable logistics aspects, as it is situated near both a railroad and a road, while also being relatively distant from inhabited areas

Section 4

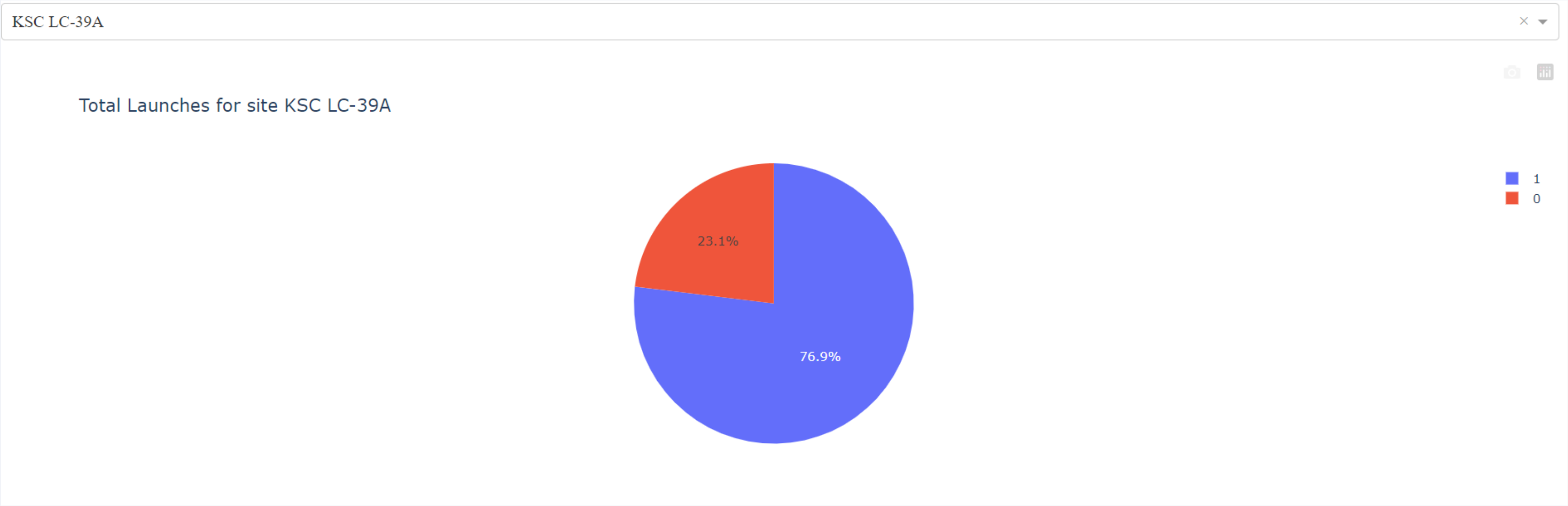
Build a Dashboard with Plotly Dash

Successful Launches by Site



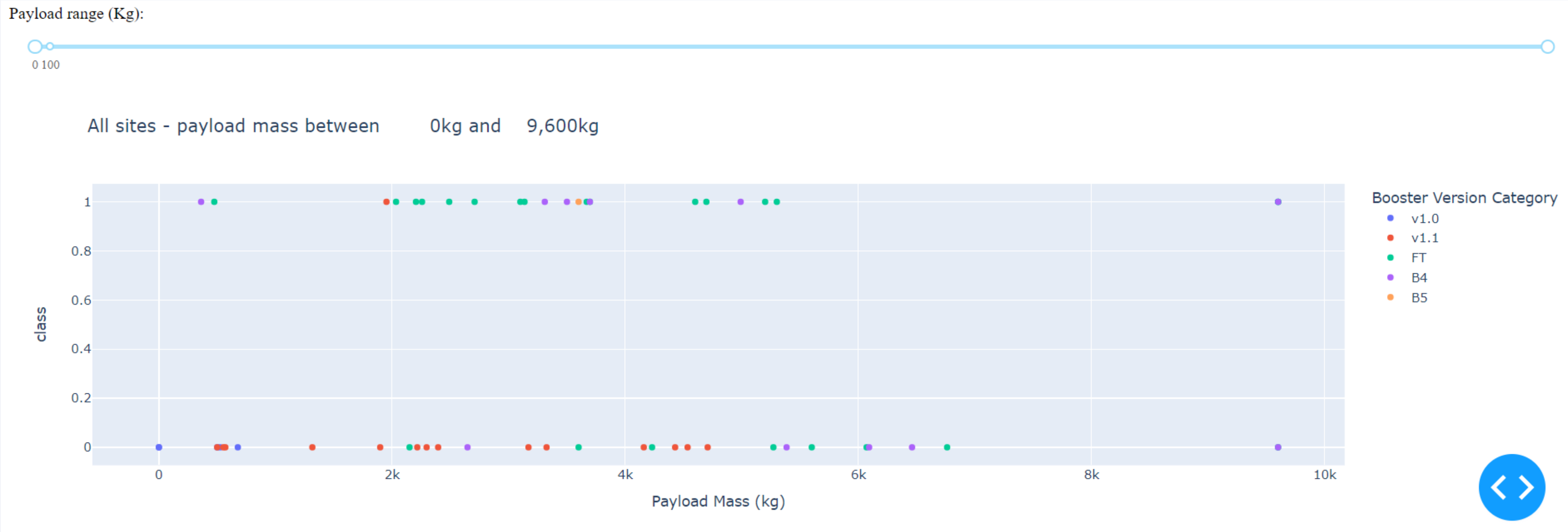
- The location from which launches are conducted appears to be a crucial factor influencing the success of missions

Launch Success Ratio for KSC LC-39A



- 76.9% of launches are successful in this site

Payload vs. Launch Outcome



- The combination of payloads under 6,000 kg and FT boosters appears to yield the highest success rate

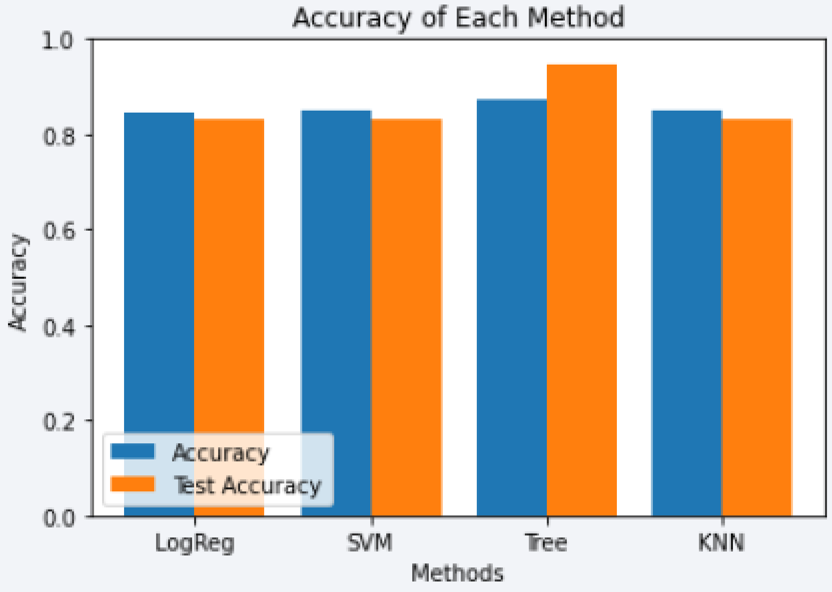
Section 5

Predictive Analysis (Classification)

Classification Accuracy

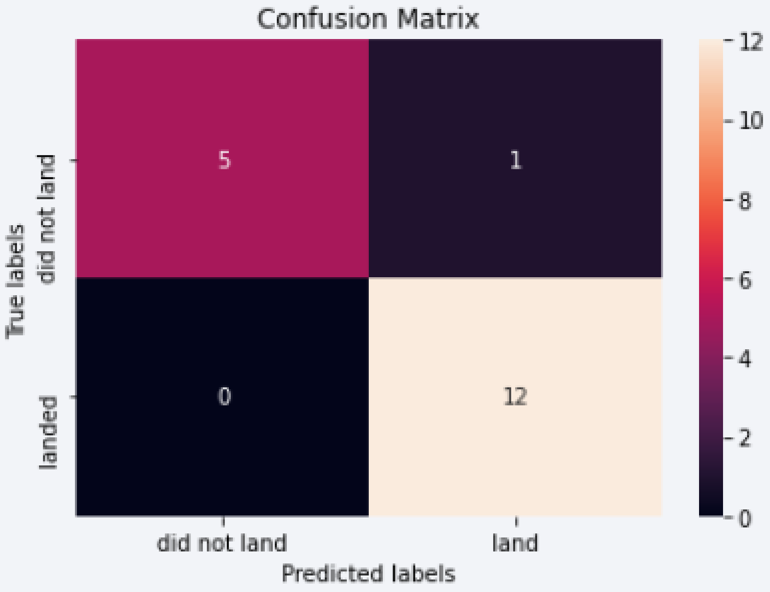
- Four classification models were tested, and their respective accuracies are plotted alongside each other

- The Decision Tree Classifier emerges as the model with the highest classification accuracy, boasting accuracies exceeding 87%



Confusion Matrix

- The confusion matrix of the Decision Tree Classifier validates its accuracy, as evidenced by the large numbers of true positives and true negatives compared to the false ones



Conclusions

- Various data sources were analyzed, with conclusions refined throughout the process.
- The optimal launch site is determined to be KSC LC 39A.
- Launches exceeding 7,000 kg pose less risk.
- While most mission outcomes are successful, successful landing outcomes appear to improve over time, reflecting advancements in processes and rocket technology.
- The Decision Tree Classifier stands out as a reliable tool for predicting successful landings and potentially increasing profits

Appendix

Thank you!