IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jacob Carr
2025-08-22

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with Web Scraping
  - Data wrangling
  - Exploratory Data using SQL
  - Exploratory Data Analysis for Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Visual Analytics with Plotly Dash
  - Machine Learning Predictive Analysis for Classification
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics
  - Predictive Analysis Results

Introduction

- Project background and context
    - In this capstone project, we seek to understand the factors affecting the successful landing of SpaceX Falcon 9 rocket launches. The cost of each launch is above 60 million dollars, but the reusability of the first stage nonetheless provides a significant cost savings compared to alternatives. As such, it is desirable to have a model that provides accurate prediction of whether a launch will be successful. This information can be used to improve launch design and to increase the likelihood of successful launches in the future.
- Problems you want to find answers
    - For a given set of features of Falcon 9 rocket launches, including payload mass, orbit type, launch site, etc., will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through web scraping and using the SpaceX API
- Perform data wrangling
  - Data filtering was performed, missing values were replaced with mean values for certain variables, and One Hot encoding was used to prepare data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were built, trained, and tuned using a variety of methods, including Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors

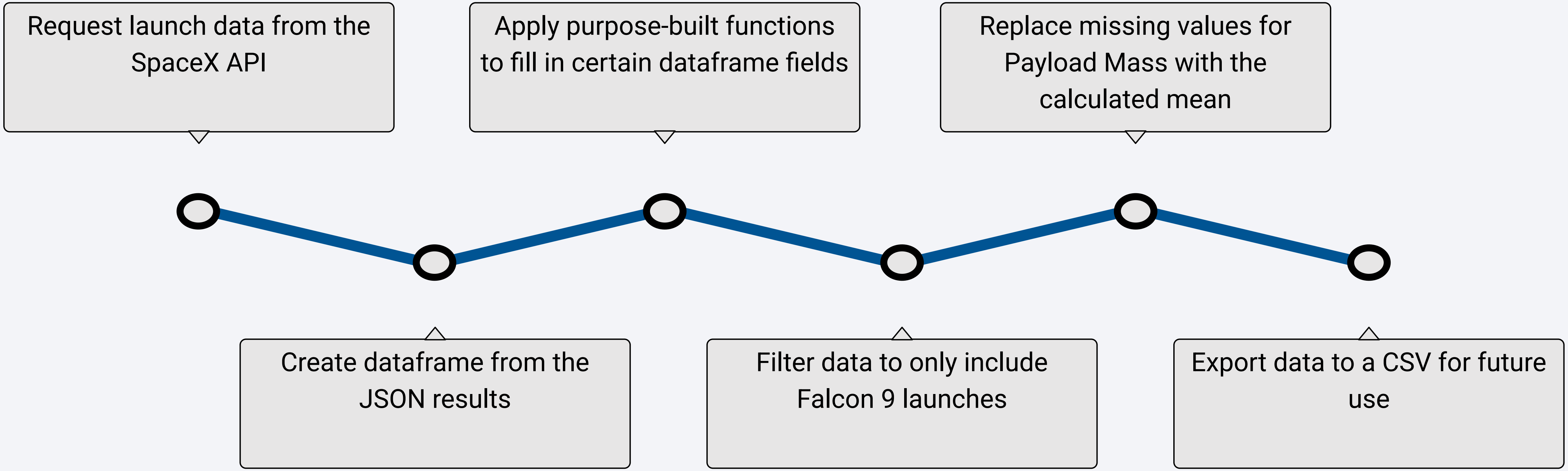
Data Collection

Data were collected through web scraping and using the SpaceX API

- SpaceX API Data Collection
  - Data on SpaceX launches can be accessed through the API here: https://api.spacexdata.com/v4/launches/past
  - For more consistent results in this capstone project, data was collected using JSON results from this static URL: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- Web Scraping Data Collection
  - Data on SpaceX launches was also collected using web scraping from the static Wikipedia page here: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- REST calls from the SpaceX API was used to collect data including Flight Number, Date, Booster Version, Payload Mass, Orbit Type, Launch Site, Outcome, etc.
- The notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraping was performed using the Wikipedia page on SpaceX launches to obtain data including Flight Number, Launch Site, Payload, Payload Mass, etc.
- The notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/2_jupyter-labs-web scraping.ipynb

Request launch data from Wikipedia page

Create BeautifulSoup object from the HTML response

Extract column names from the HTML Headers

Parse collected data and construct a dictionary

Create a dataframe from the dictionary

Export data to a CSV for future use

Data Wrangling

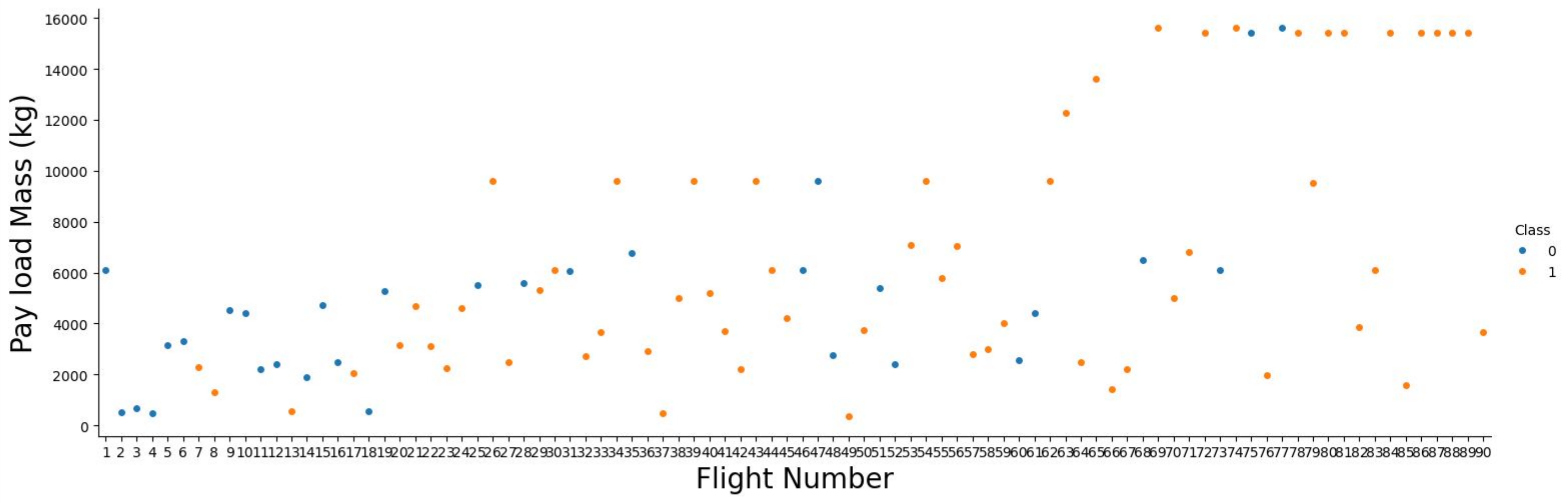
- Data wrangling was completed through the following tasks:
    - Calculating the number of launches on each site
    - Calculating the number and occurrence of each orbit
    - Calculating the number and occurrence of mission outcome of the orbits
    - Creating a landing outcome label from the Outcome column
    - Exporting the processed data to a CSV
- The completed notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

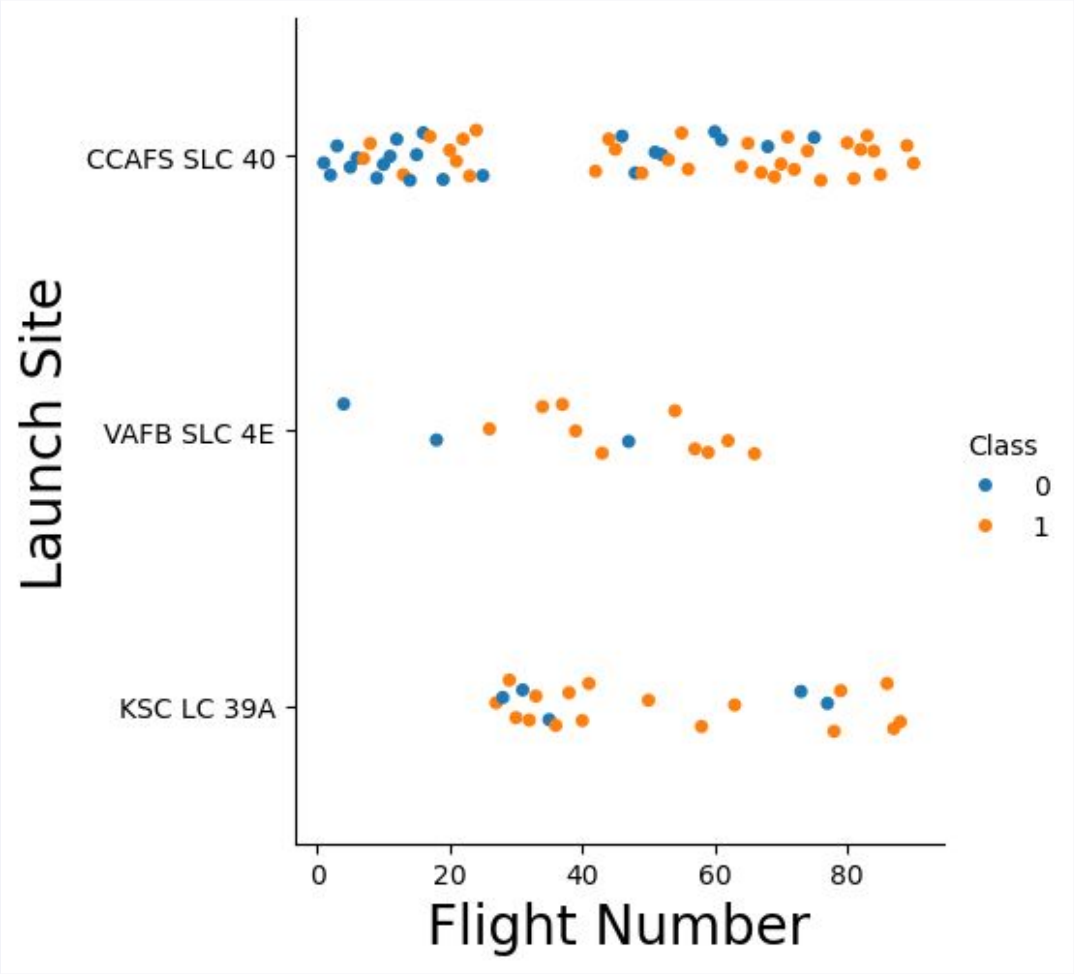
- Exploratory data analysis was performed, and plots, including scatter plots, bar charts, and line plots were created to explore relationships between flight number, payload mass, launch site, orbit type, year, and success rate.
- Example plots are shown on the following slides.

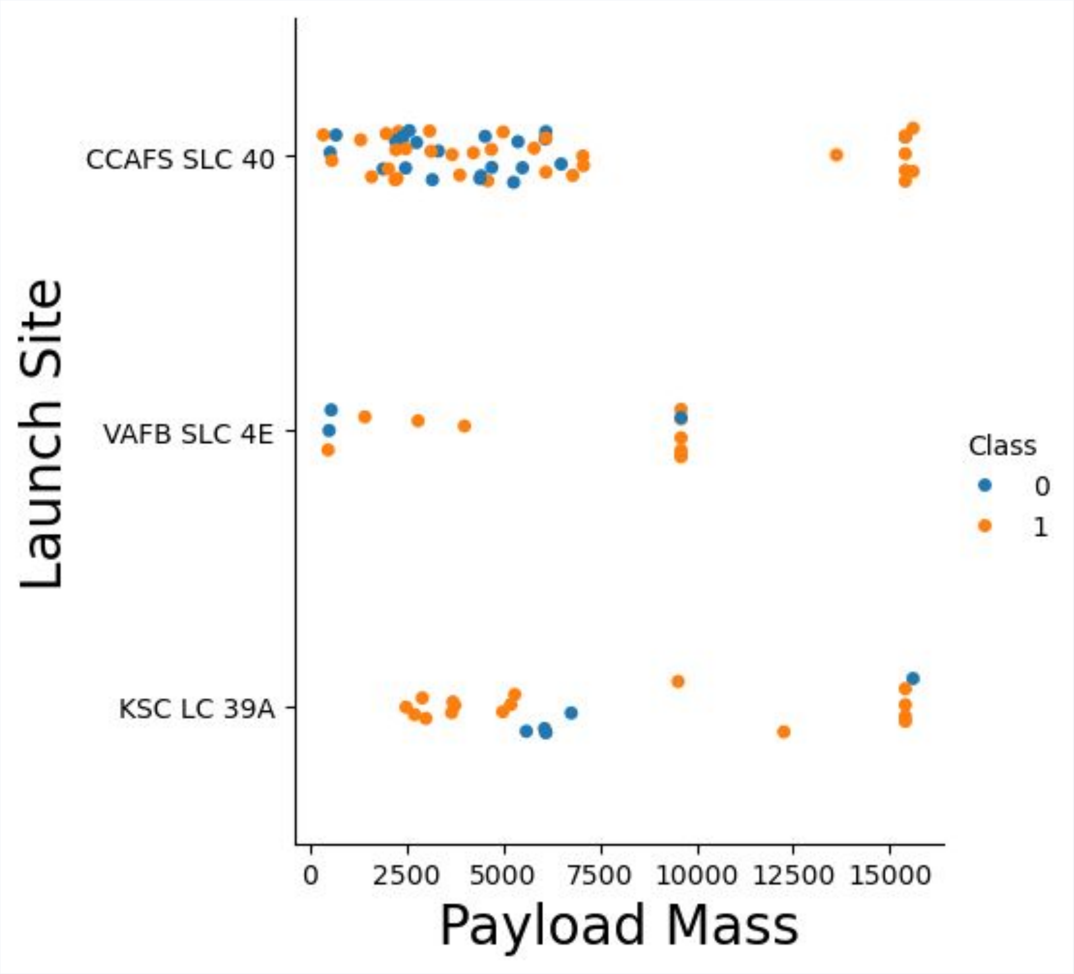
- The completed notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/5_edadataviz.ipynb

EDA with Data Visualization

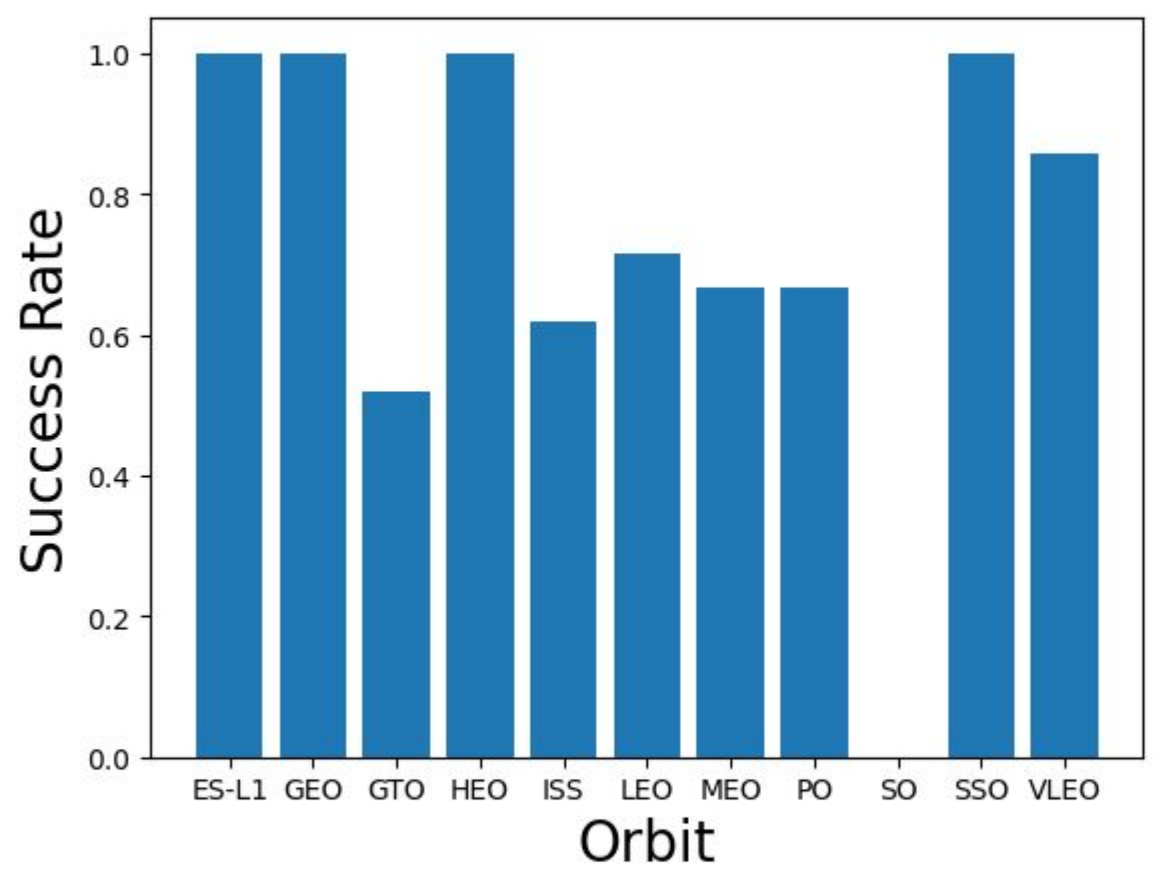


EDA with Data Visualization

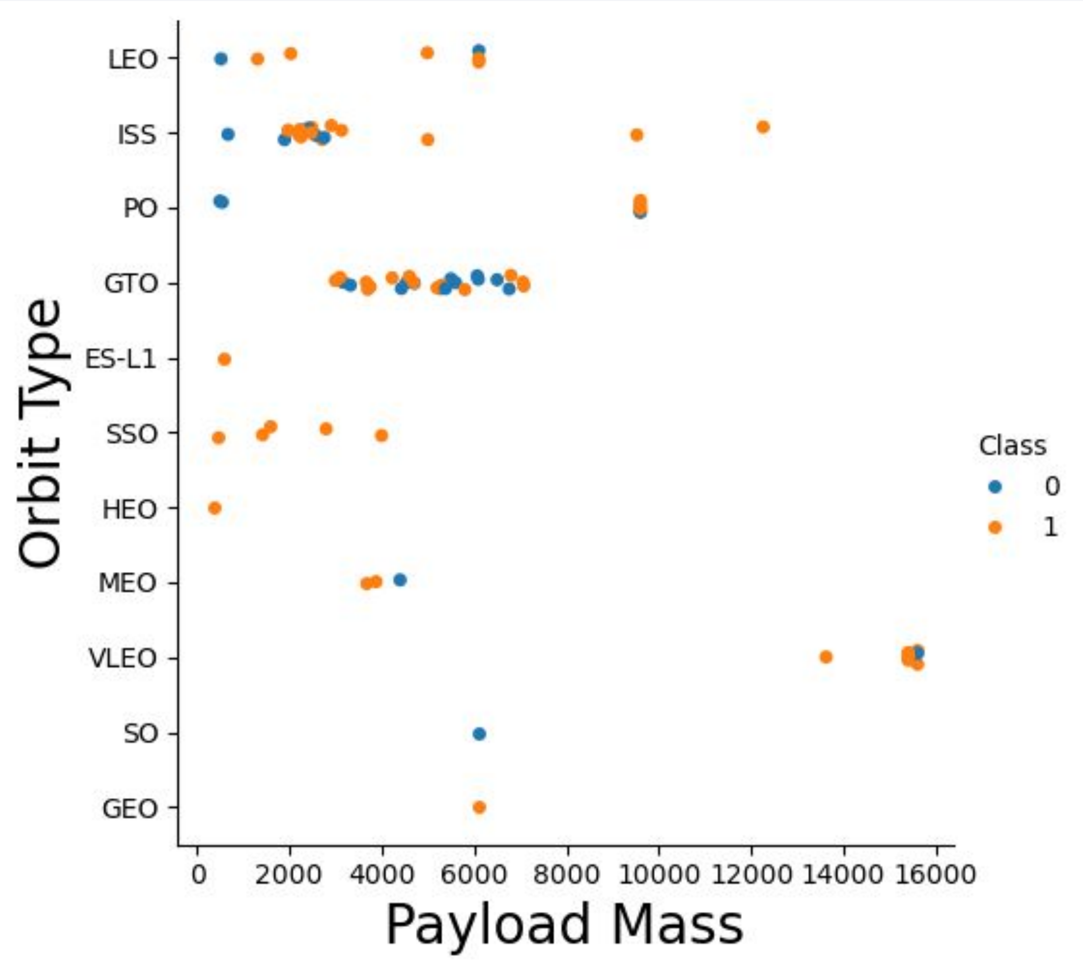


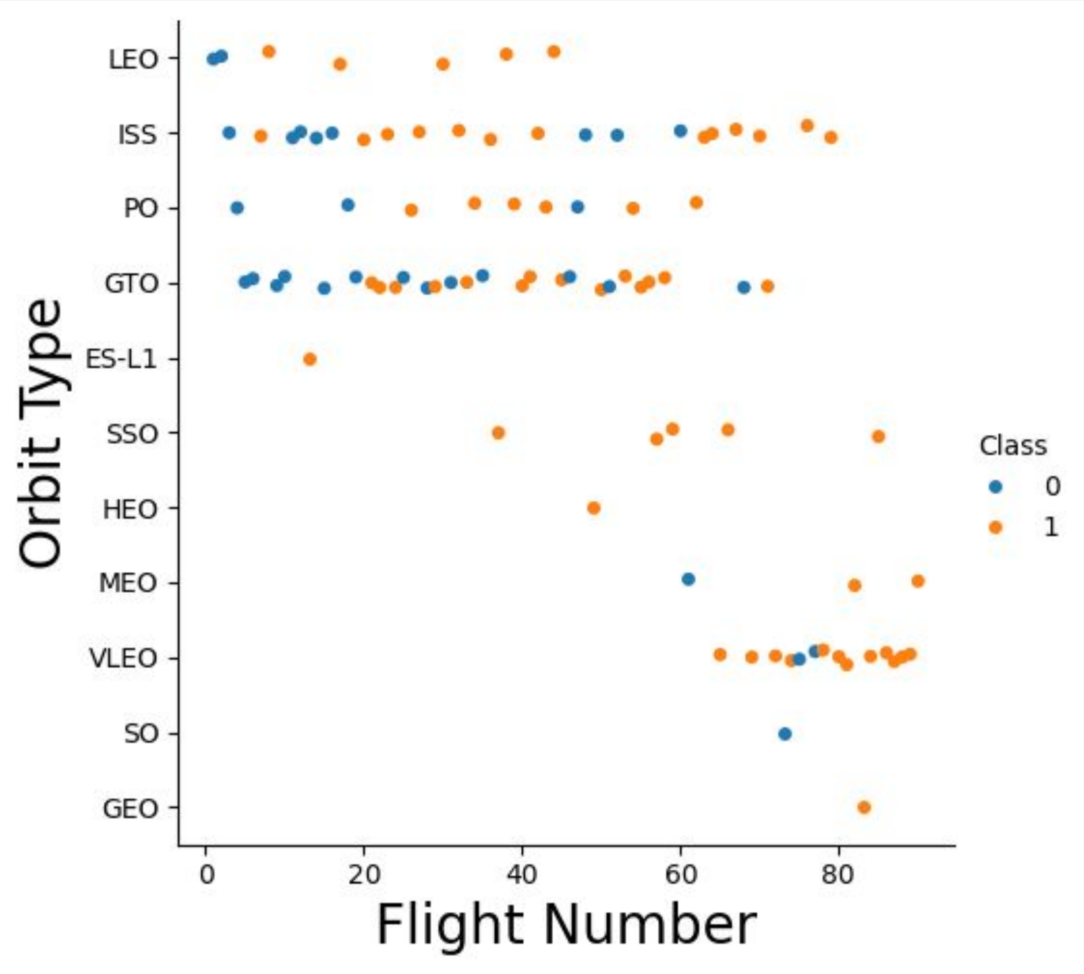


EDA with Data Visualization

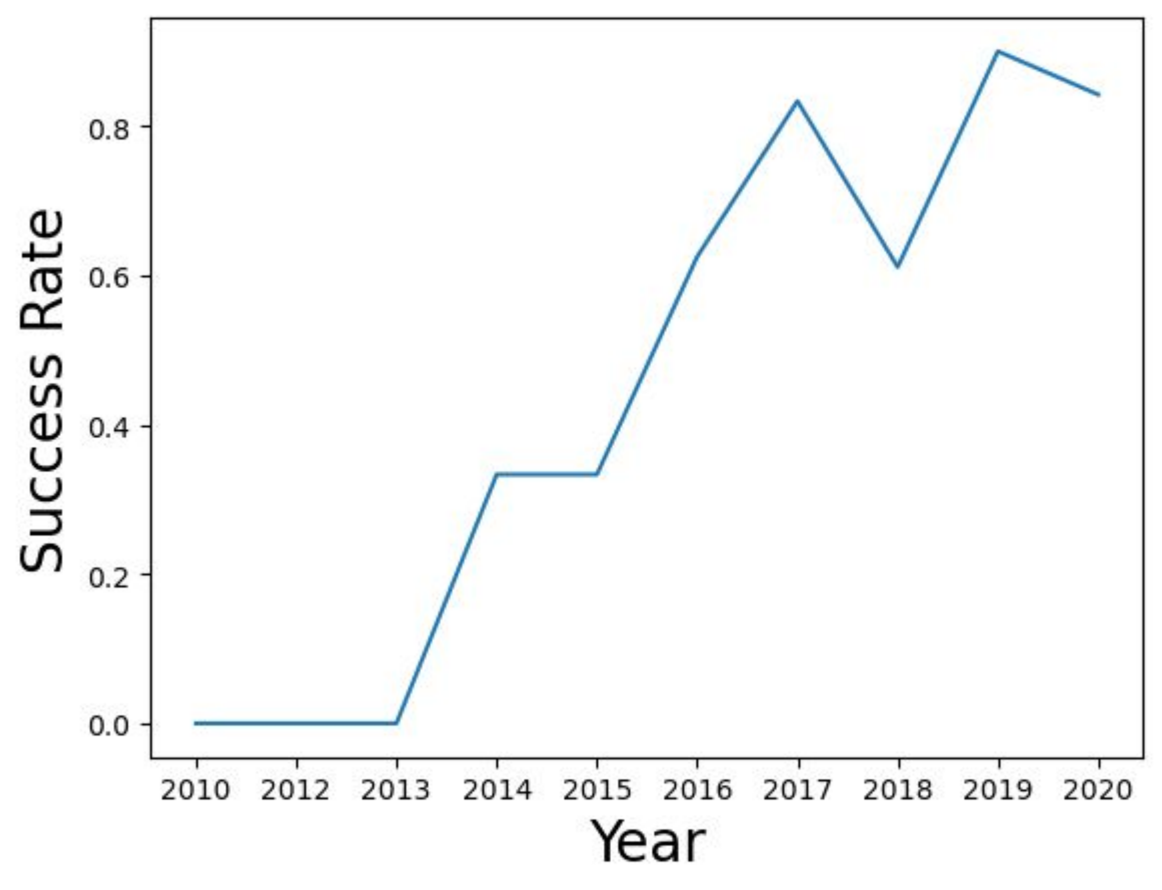


EDA with Data Visualization





EDA with Data Visualization



EDA with SQL

- SQL queries were performed for the following purposes:
    - 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 all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function
    - 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

- The completed notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

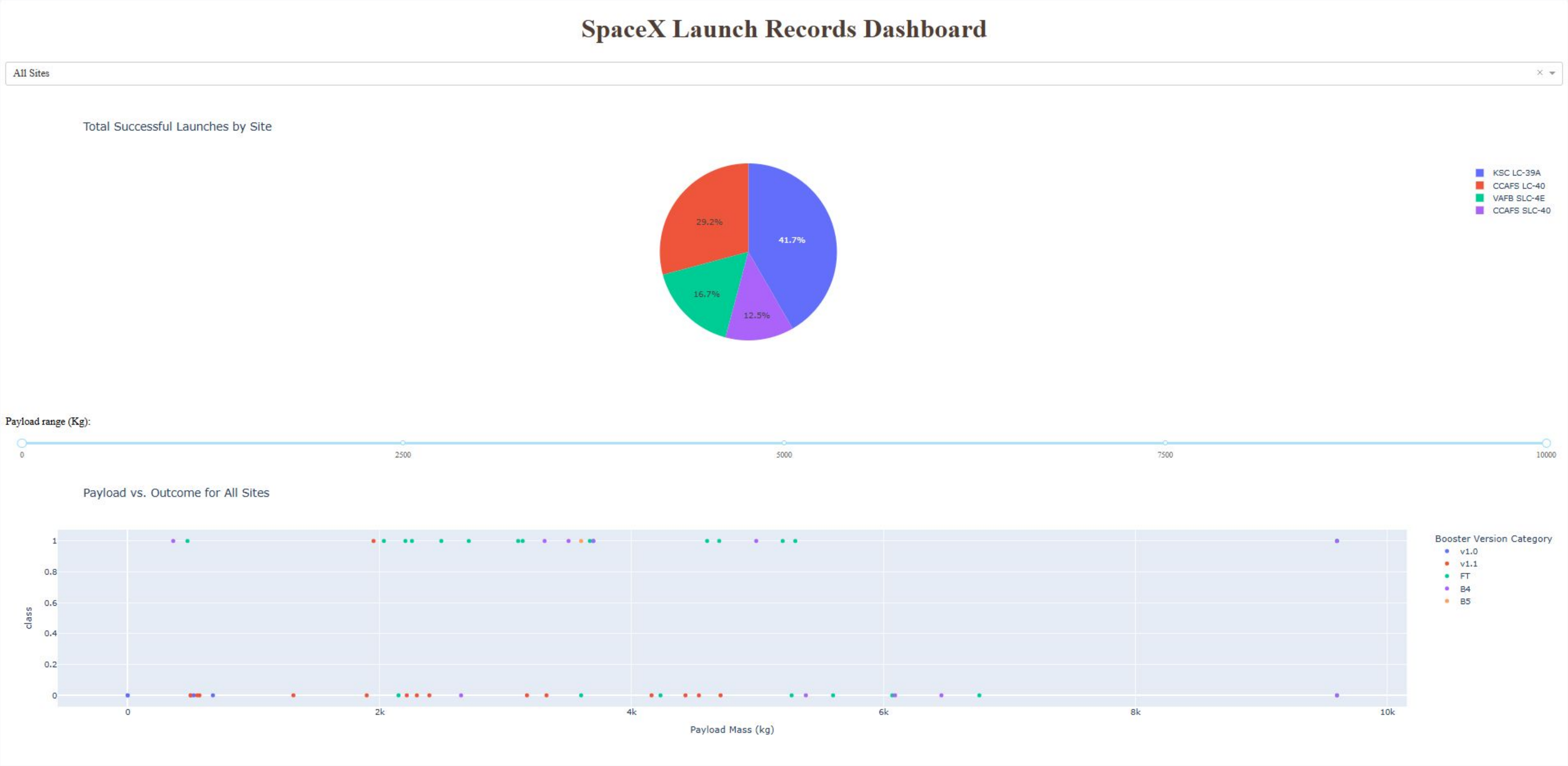
Build an Interactive Map with Folium

- An interactive folium map was created to show:
  - A circle marker at each launch site for the successful and failed launches
  - A polyline showing the distance to the nearest coastline
- The completed notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/6_lab_jupyter_launch_site_location.ipynb

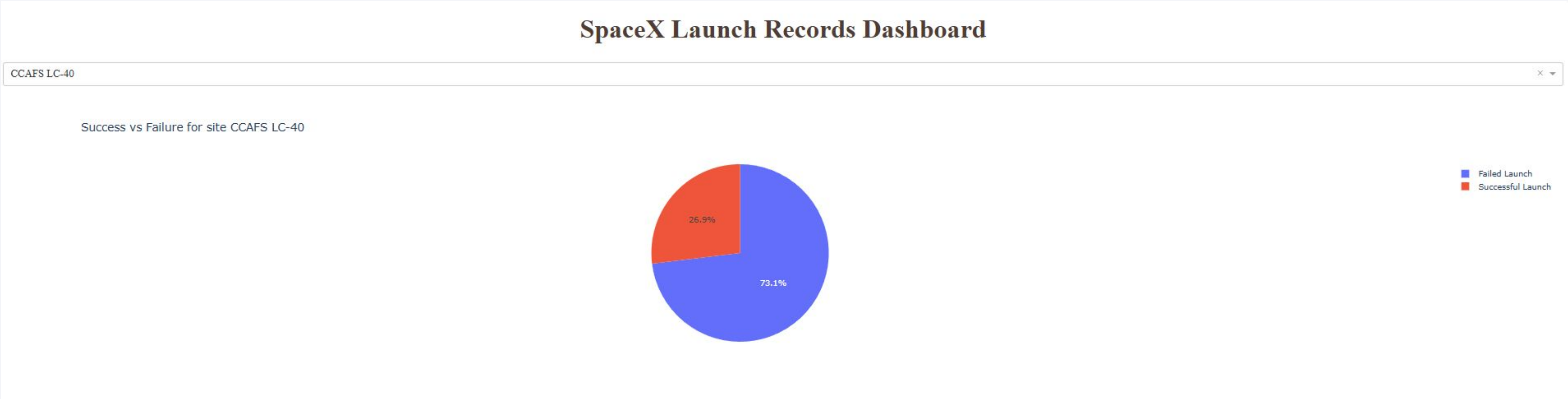
Build a Dashboard with Plotly Dash

- An interactive dashboard was created to show the following:
  - A pie chart of total successful launches by site
  - Selectable pie charts for successful launched from each site
  - A plot of payload mass vs.outcome for all sites with color coding by booster version
  - A plot of payload mass vs.outcome for a selected site with color coding by booster version
- Screenshots of the dashboard are shown on the following slides
- The completed code for the plotly dash dashboard can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/7_spacex-dash-app.py

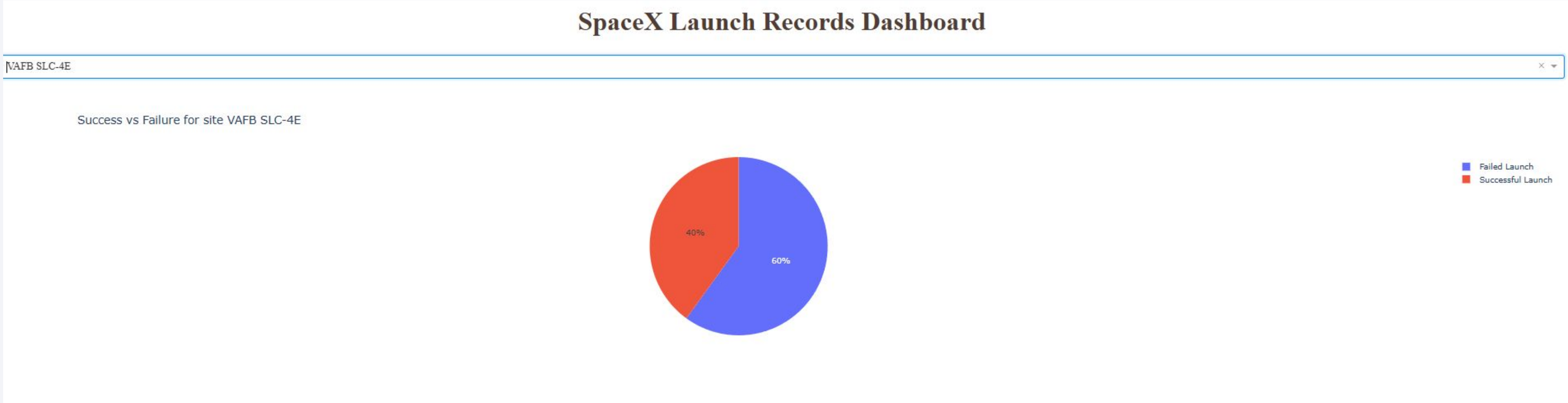
Build a Dashboard with Plotly Dash



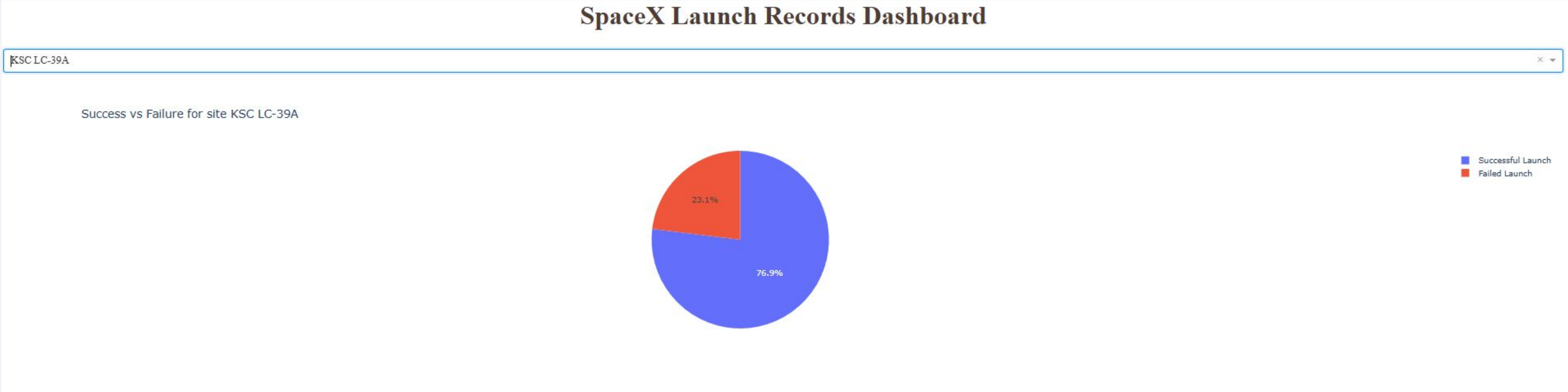
Build a Dashboard with Plotly Dash



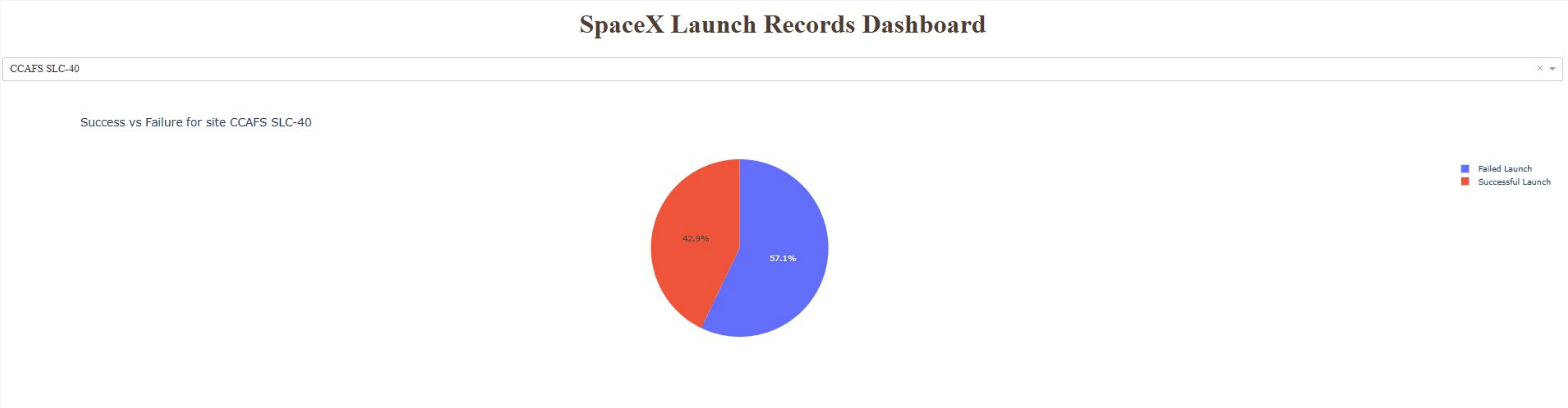
Build a Dashboard with Plotly Dash



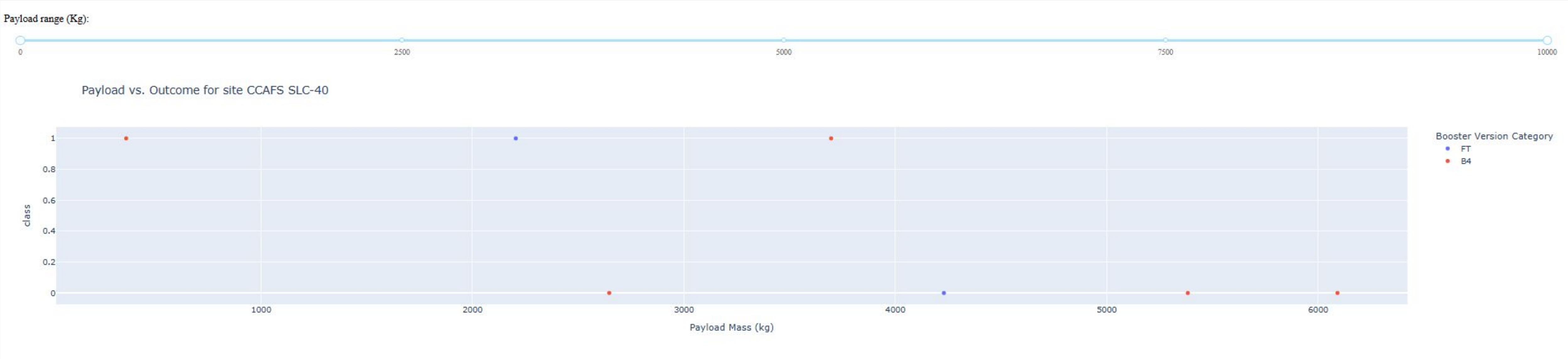
Build a Dashboard with Plotly Dash



Build a Dashboard with Plotly Dash



Build a Dashboard with Plotly Dash



Predictive Analysis (Classification)

- Models were developed using the following methods:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
- For each model, tuned parameters were determined, and the model was evaluated using a confusion matrix and the method score
- The completed notebook for this section can be found here: https://github.com/JCarr412/Coursera-DataScience/blob/main/8_completed_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

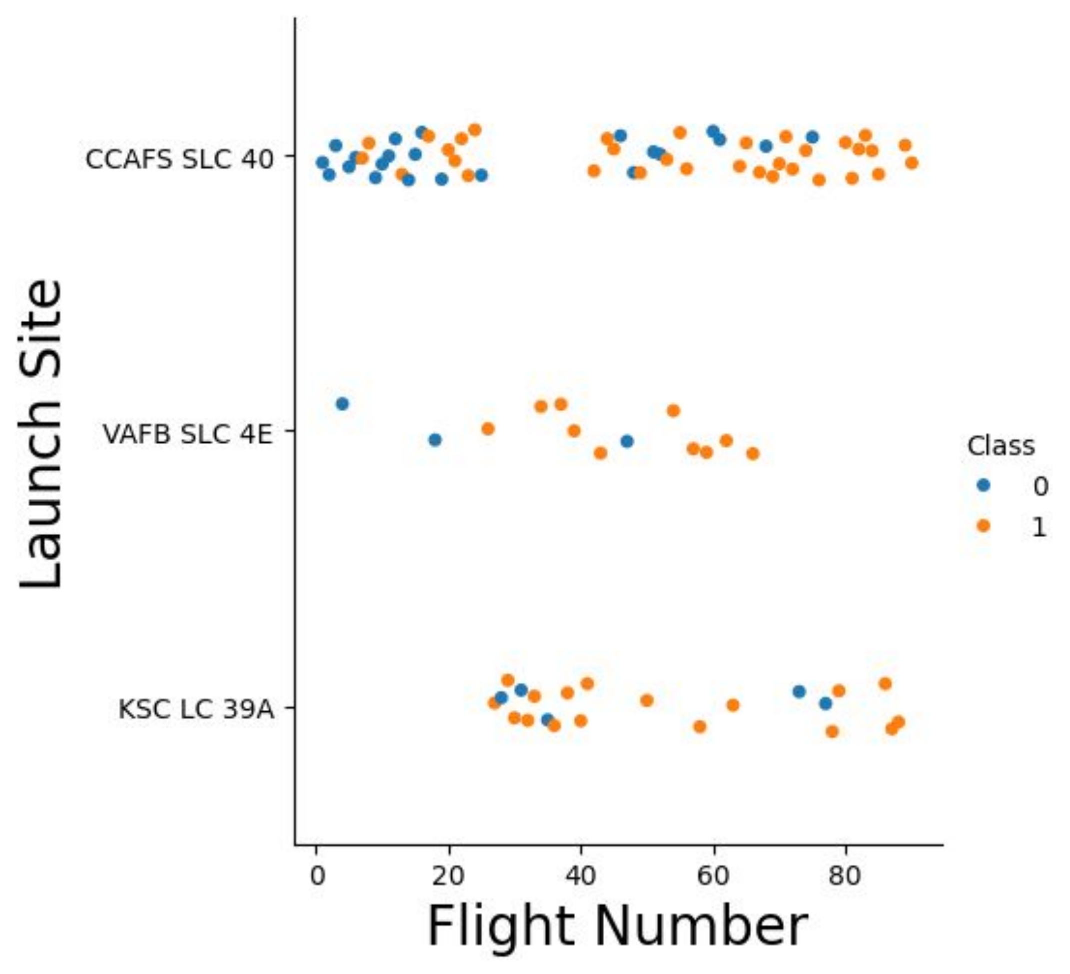
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

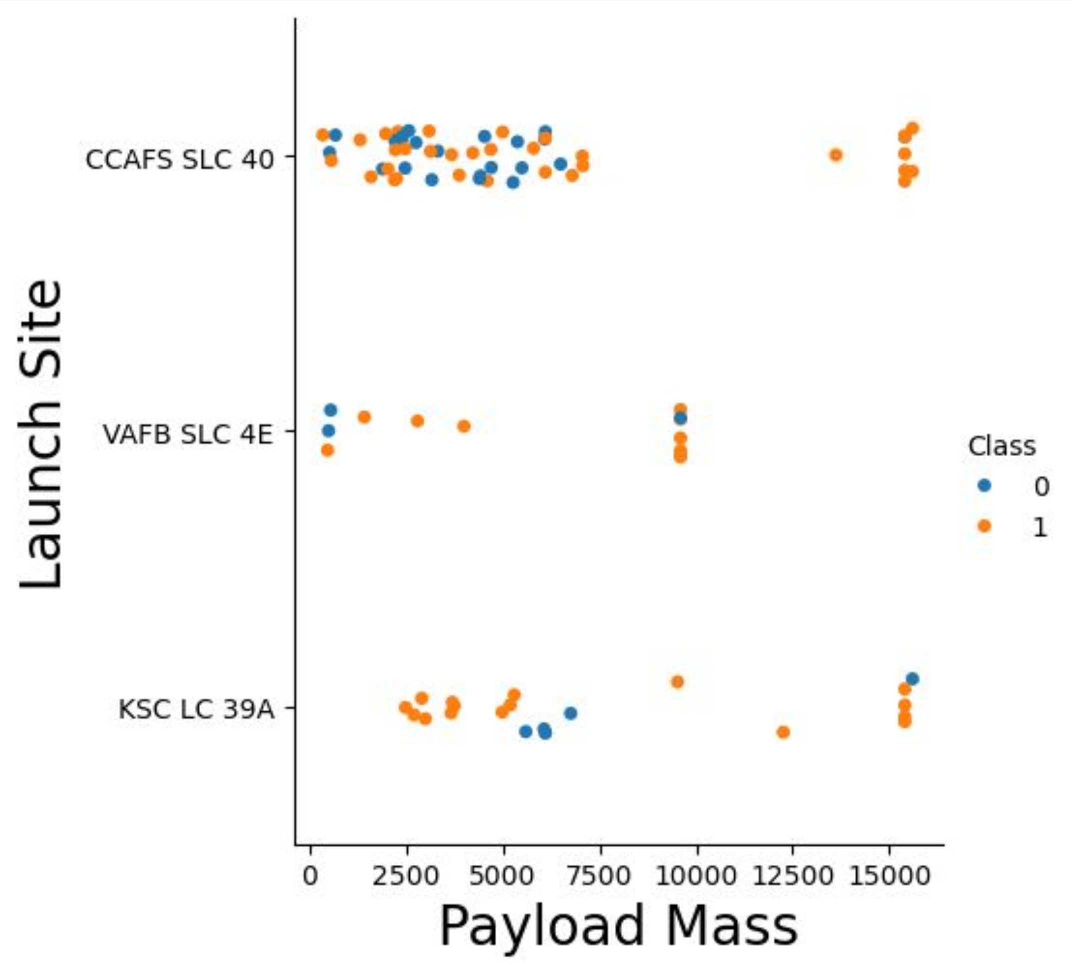
Flight Number vs. Launch Site

- The plot shows increasing success rates with the later flights.
- The plot shows that more launches occurred at the CCAFS SLC 40 site than at either of the other two sites
- The plot does not show a clear trend in where launches occur as a function of flight number



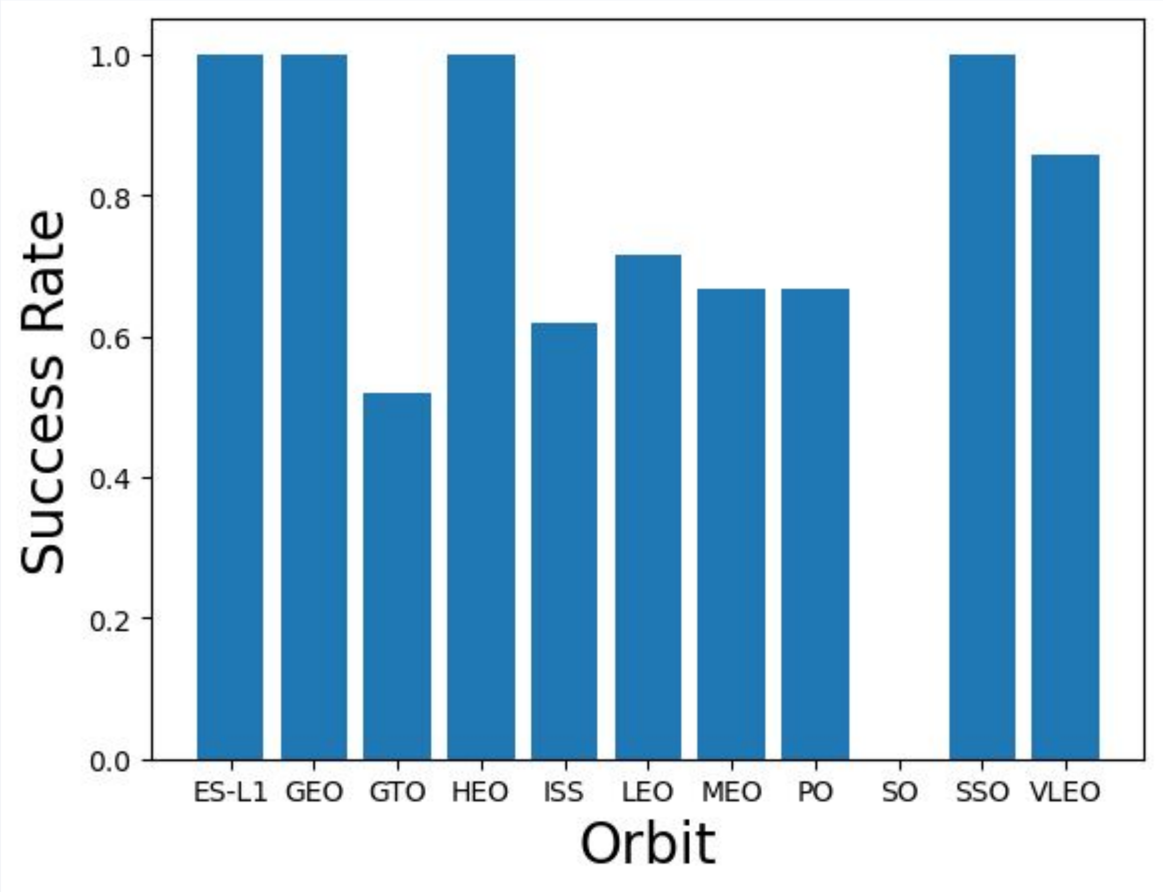
Payload vs. Launch Site

- The plot shows the range of payload masses that were handled at each launch site
- The highest payload masses appear to have greater success rates
- The highest payload masses were only launched from the CCAFS SLC 40 and KSC LC 39A sites



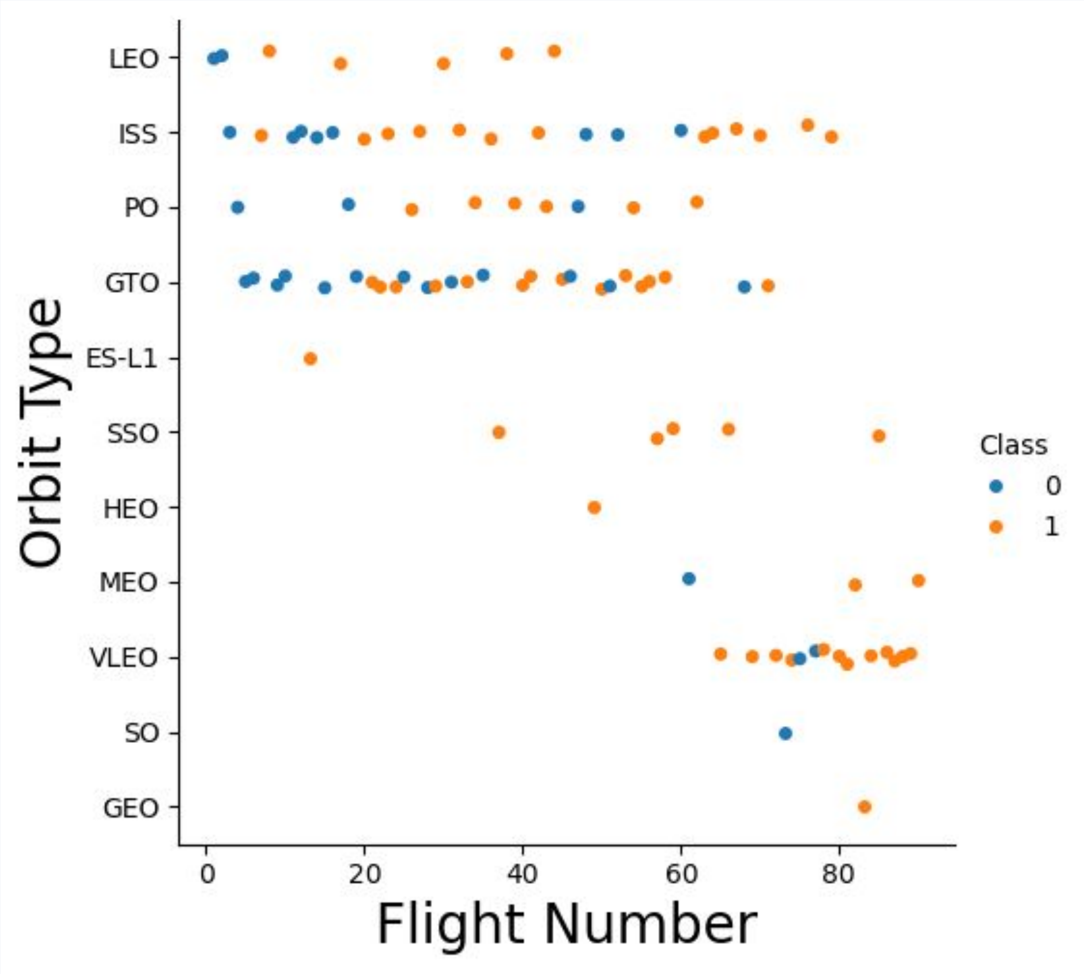
Success Rate vs. Orbit Type

- Several orbit types have 100% success rates:
    - ES-L1, GEO, HEO, and SSO
- One orbit type has a 0% success rate
    - SO
- The success rate for other sites ranges from about 50% to about 85%



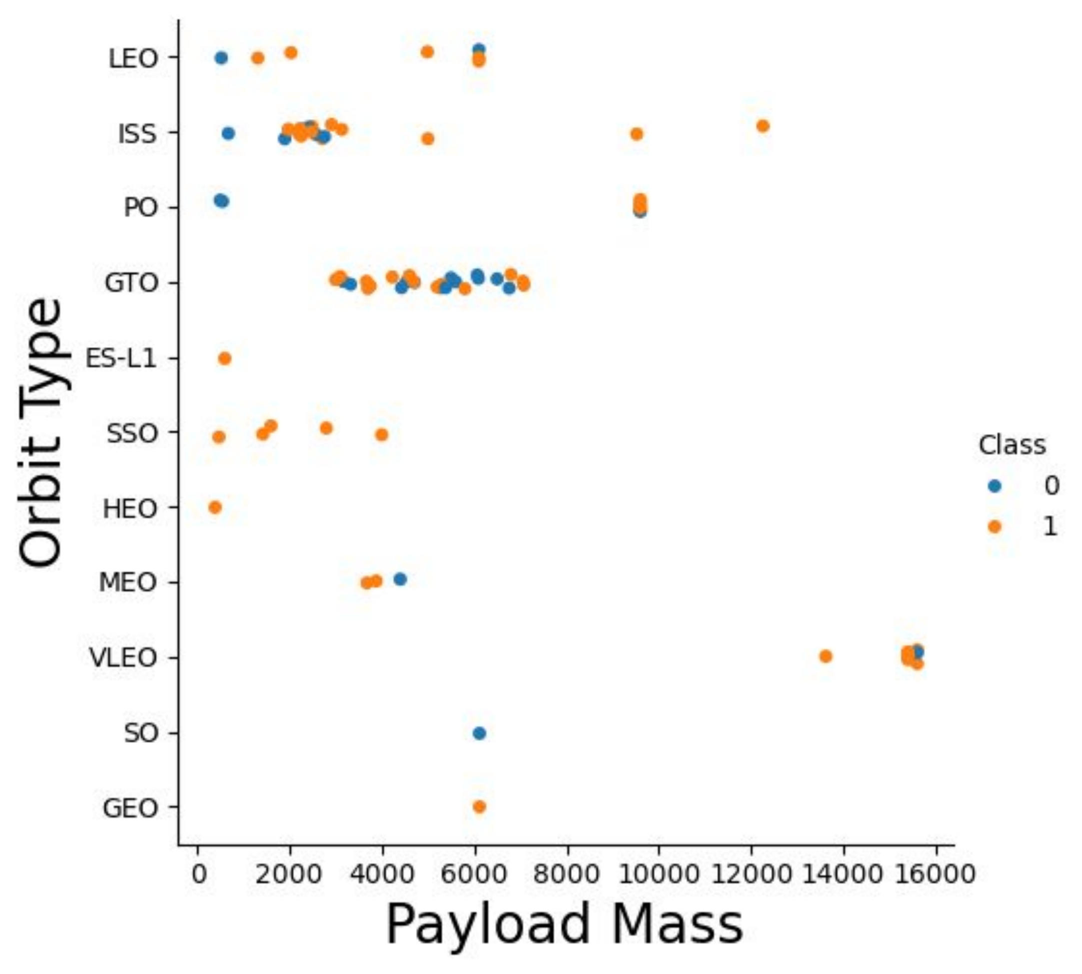
Flight Number vs. Orbit Type

- The chart shows a trend in changes over time with the types of orbits that are being launched.
- LEO, ISS, PO, GTO, and ES-L1 launches have been conducted since fairly early.
- More recently, SSO, HEO, MEO, VLEO, SO, and GEO launches have been conducted.
- Success rates for each type of launch can also be visualized on the plot



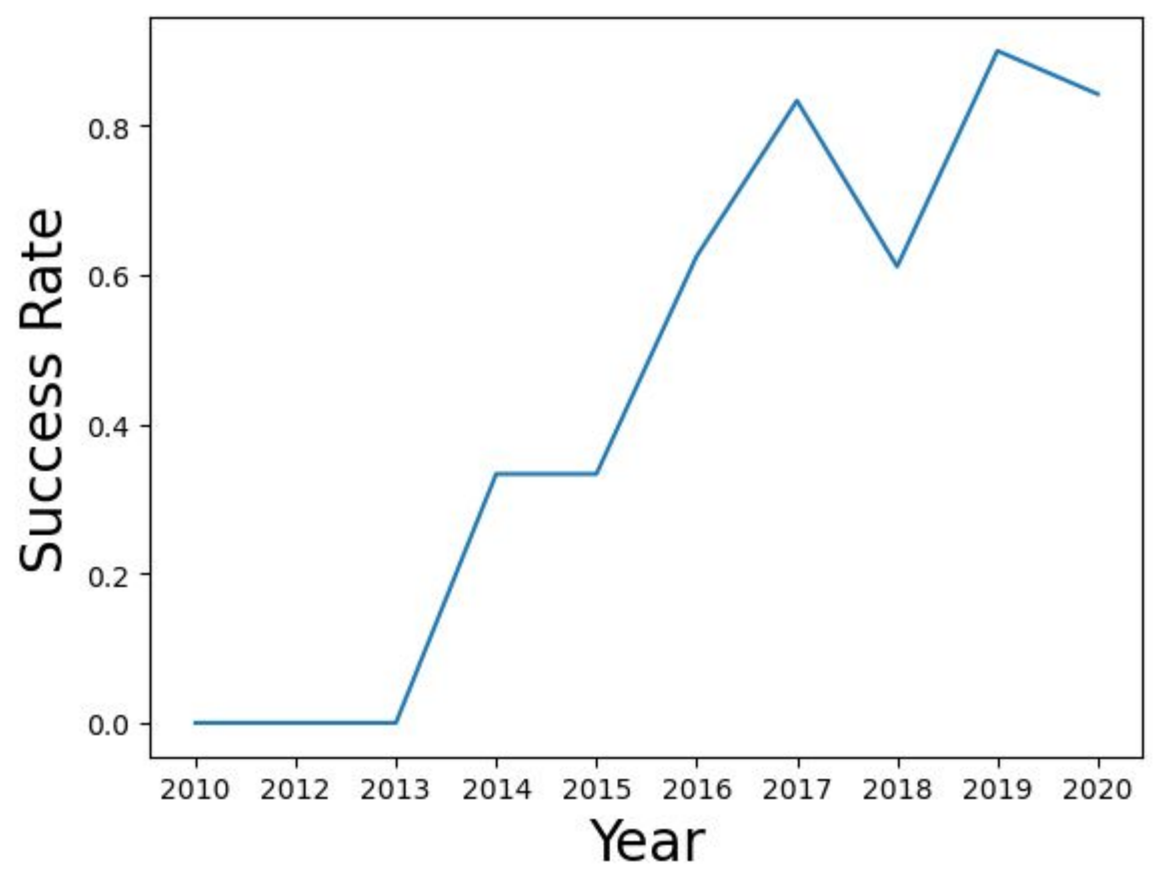
Payload vs. Orbit Type

- The orbit type as a function of payload mass is shown on the plot.
- The heaviest payloads are handled only by the VLEO orbit type.
- The success rate as a function of orbit type and payload mass can also be seen



Launch Success Yearly Trend

- The plot shows a fairly steady increase in success rate
- Success rate in the early years was 0%
- Success in the most recent years was above 80%



All Launch Site Names

- Unique launch site names were found with the following query:

  %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL

- Results:

  Launch_Site

  CCAFS LC-40

  VAFB SLC-4E

  KSC LC-39A

  CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- The sites beginning with CCA were found with this query:

  %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

- Result:

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

Total Payload Mass

- Total payload mass was found with this query:

  %sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
  FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';

- Result: 45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 V1.1 was found with this query:

  %sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';

- Result: 2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was found with this query:

  %sql SELECT MIN(Date) AS First_Successful_Ground_Pad_Landing FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';

- Result: 2015-12-22

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 were found with this query:

  %sql SELECT DISTINCT Landing_Outcome FROM SPACEXTBL

- Result shown to the right

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes was found with the following query:

  %sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)'   AND PAYLOAD_MASS__KG_ > 4000   AND PAYLOAD_MASS__KG_ < 6000;

- Result:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass was found with this query:

  %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

-

- The results are shown to the right

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 were found with this query:

  %sql SELECT substr(Date, 6, 2) AS Month,Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date, 1, 4) = '2015' AND Landing_Outcome = 'Failure (drone ship)';

- Result:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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, were ranked in descending order using this query:

  %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

- Result shown to the right

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

Folium map sreenshots

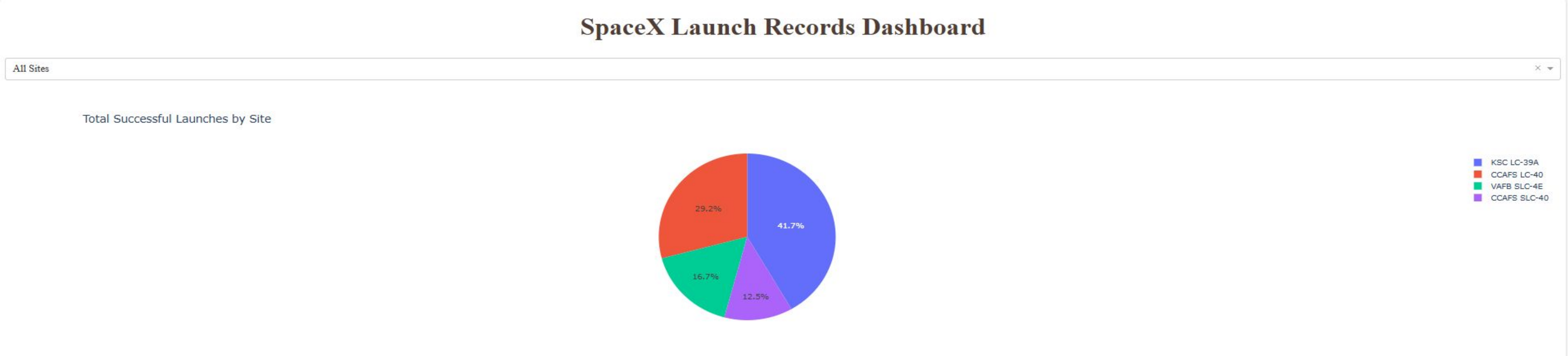
- Unfortunately, screenshots for the folium map were not obtained.

Section 4

Build a Dashboard with Plotly Dash

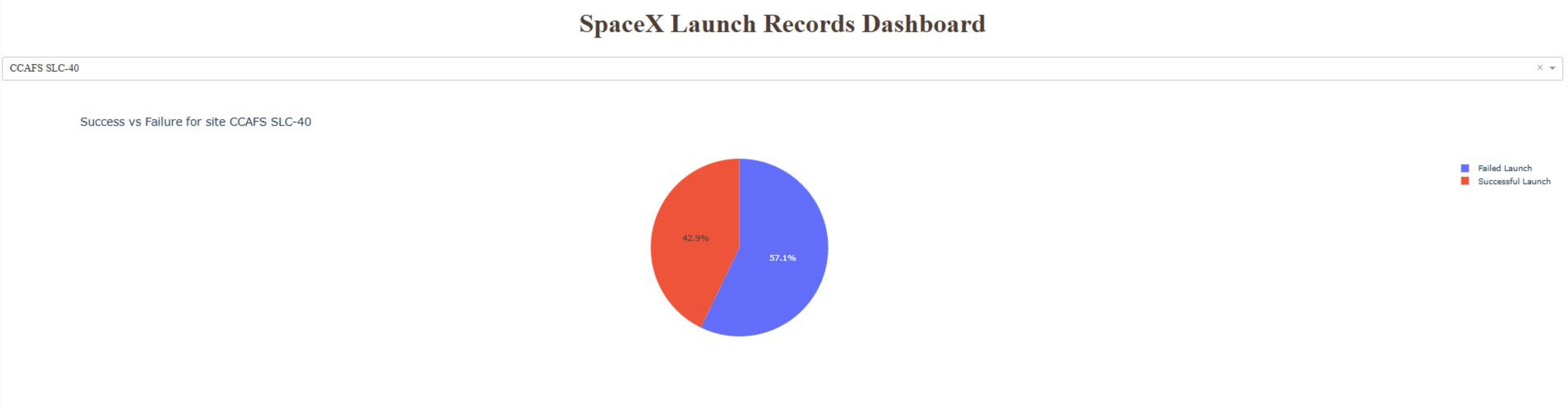
Dashboard Screenshot: Launch Success Count for All Sites

- The plot shows that the site with the most successful launches (41.7% of all successful launches) was KSC LC-39A
- The site with the fewest successful launches was CCAFS SLC-40



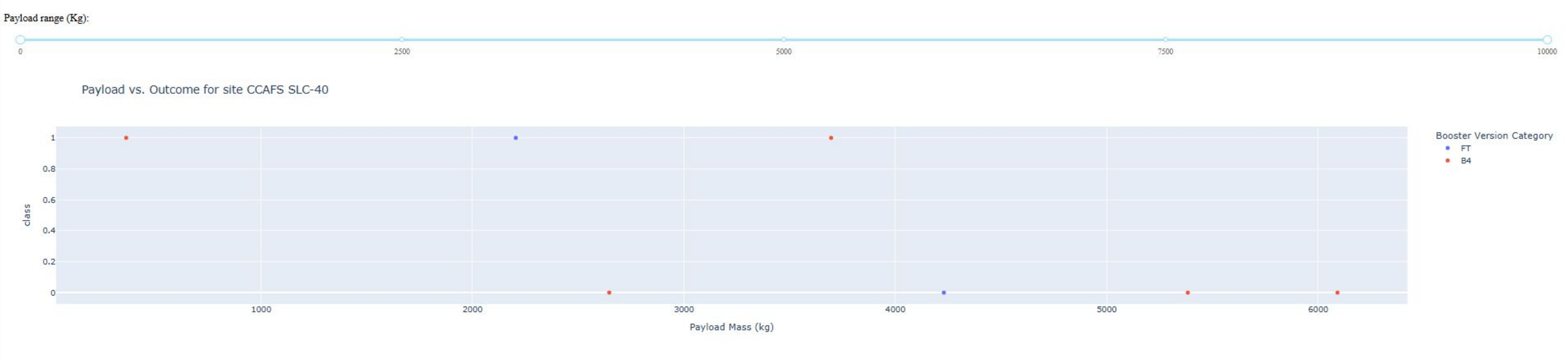
Dashboard Screenshot: Site with Highest Success Ratio

- The site with the highest success rate is CCAFS SLC-40, with a success rate of 42.9%



Dashboard Screenshot:
Payload vs. Launch Outcome scatter plot for CCAFS SLC-40

- The plot shows the payload vs. outcome for CCAFS SLC-40 for the full range of payload masses
- Payload range can be selected with the slider
- A screenshot of the plot for all sites was not taken

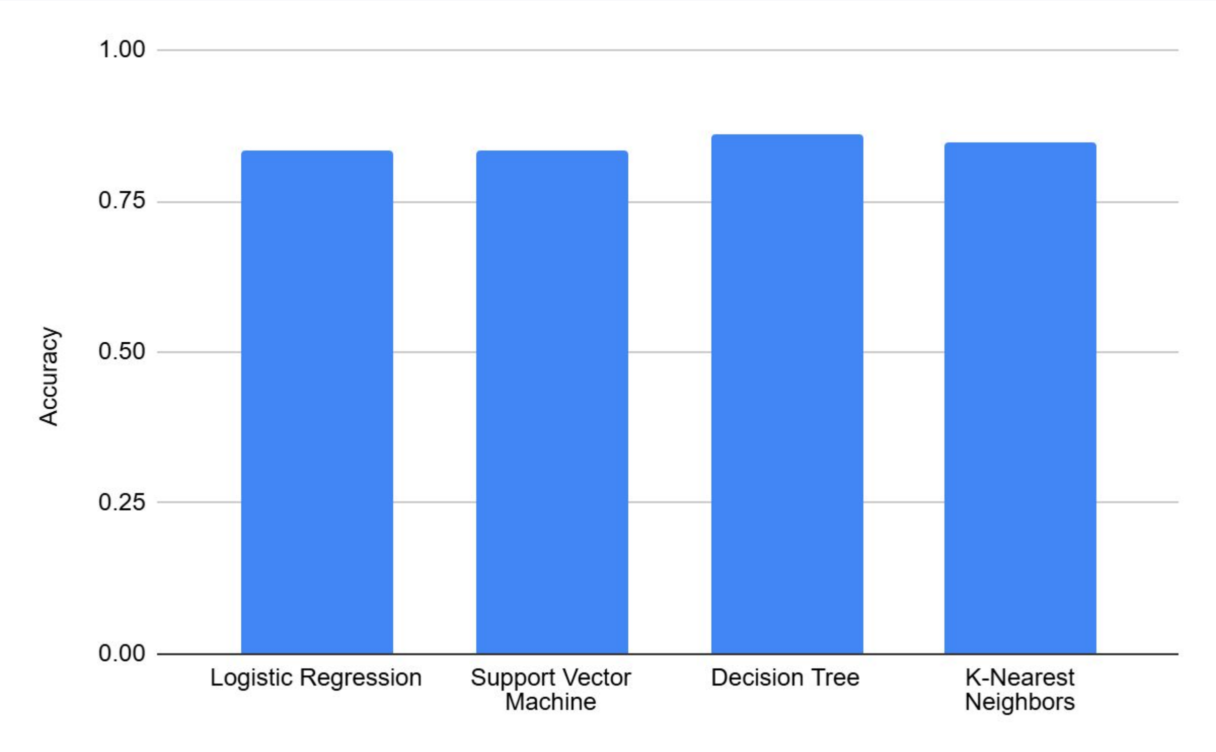


Section 5

Predictive Analysis (Classification)

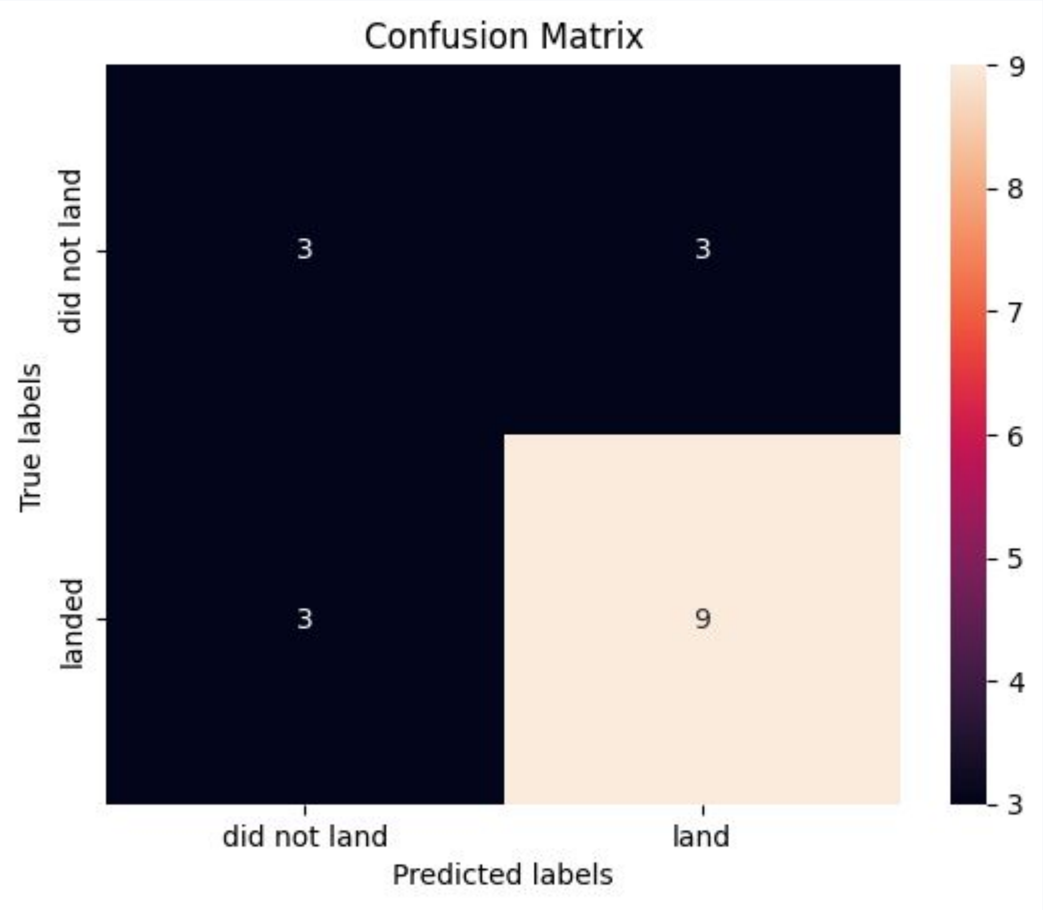
Classification Accuracy

- The model accuracy for each tuned model is shown to the right.
- The Decision Tree model has the highest accuracy.



Confusion Matrix

- The confusion matrix for the Decision Tree model is shown to the right.
- The matrix shows relatively low errors for both Type 1 and Type 2 errors with 3 each.
- The model successfully predicts the correct outcome in 12 of the cases for the test set (9 landed and 3 did not land)



Conclusions

- The Decision Tree Model performs best for this dataset
- Launches with higher payload mass show greater success rates
- Most launches are from sites close to the equator and close to the coastline
- Success rates have increased over time

Appendix

- All relevant code and notebooks can be found here: https://github.com/JCarr412/Coursera-DataScience

Thank you!