Winning Space Race with Data Science

Lyle Davis
21 June 2024

Outline

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

Executive Summary

Data science tools and techniques were used to predict whether or not a first stage booster landing would be successful. These predictions are useful to competitors of SpaceX competing for space launch contracts; specifically, to understand how to bid launch costs against SpaceX. The results of the analysis will show which model is the best predictor or successful booster landings.

Executive Summary

- Summary of methodologies
  - Data Collection and Data Wrangling
  - Exploratory Data Analysis
  - Machine Learning
  - Predictive Modeling
  - Advanced Data Visualization
  - Working With Big Data
  - Capstone Project
  - Soft Skills

Executive Summary

- Summary of all results
  - Data Collection and Data Wrangling
    - Extracted and prepared SpaceX datasets; cleaned, pre-processed for effective use (including missing data)
  - Exploratory Data Analysis
    - Identified key patterns, trends, and anomalies within datasets using descriptive statistics and data visualization
  - Machine Learning
    - Built & evaluated supervised/unsupervised models; performed model evaluation/selection (cross-validation & hyperparameter tuning)
  - Predictive Modeling
    - Models validated using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC
  - Advanced Data Visualization
    - Developed interactive visualizations using tools like Matplotlib, Seaborn, and Plotly
  - Working With Big Data
  - Capstone Project
  - Soft Skills

Introduction

- Project background and context
  - Key Details:
    - Objective: Predict Falcon 9 first stage landing success
    - Data Source: SpaceX launch records and relevant datasets
    - Tools & Techniques: Data wrangling, EDA, machine learning models, and evaluation metrics
    - Outcome: Insight into launch cost efficiency and competitive bidding strategies
- Problems you want to find answers
  - Predict whether the Falcon 9 first stage will land successfully. This prediction is crucial because SpaceX's cost efficiency—$62 million per launch compared to $165 million by other providers—largely depends on reusing the first stage. Accurate predictions will help determine launch costs, benefiting companies bidding against SpaceX

Section 1

Methodology

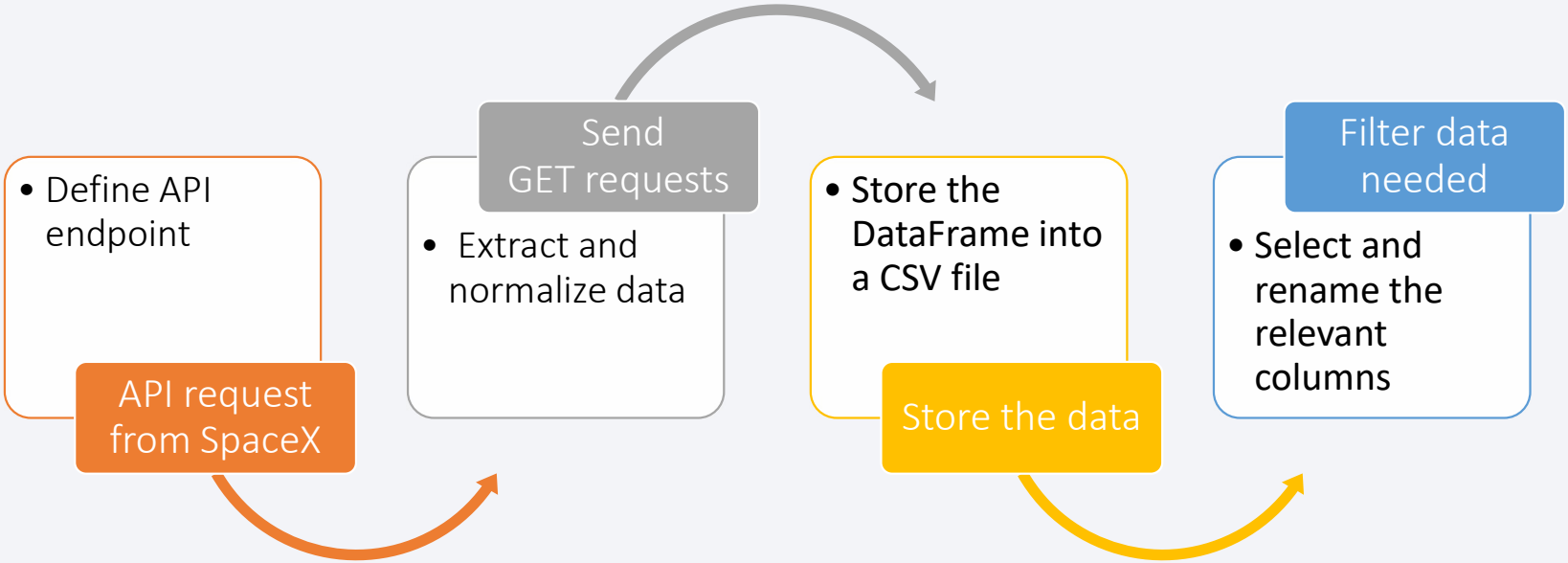
Methodology

Executive Summary

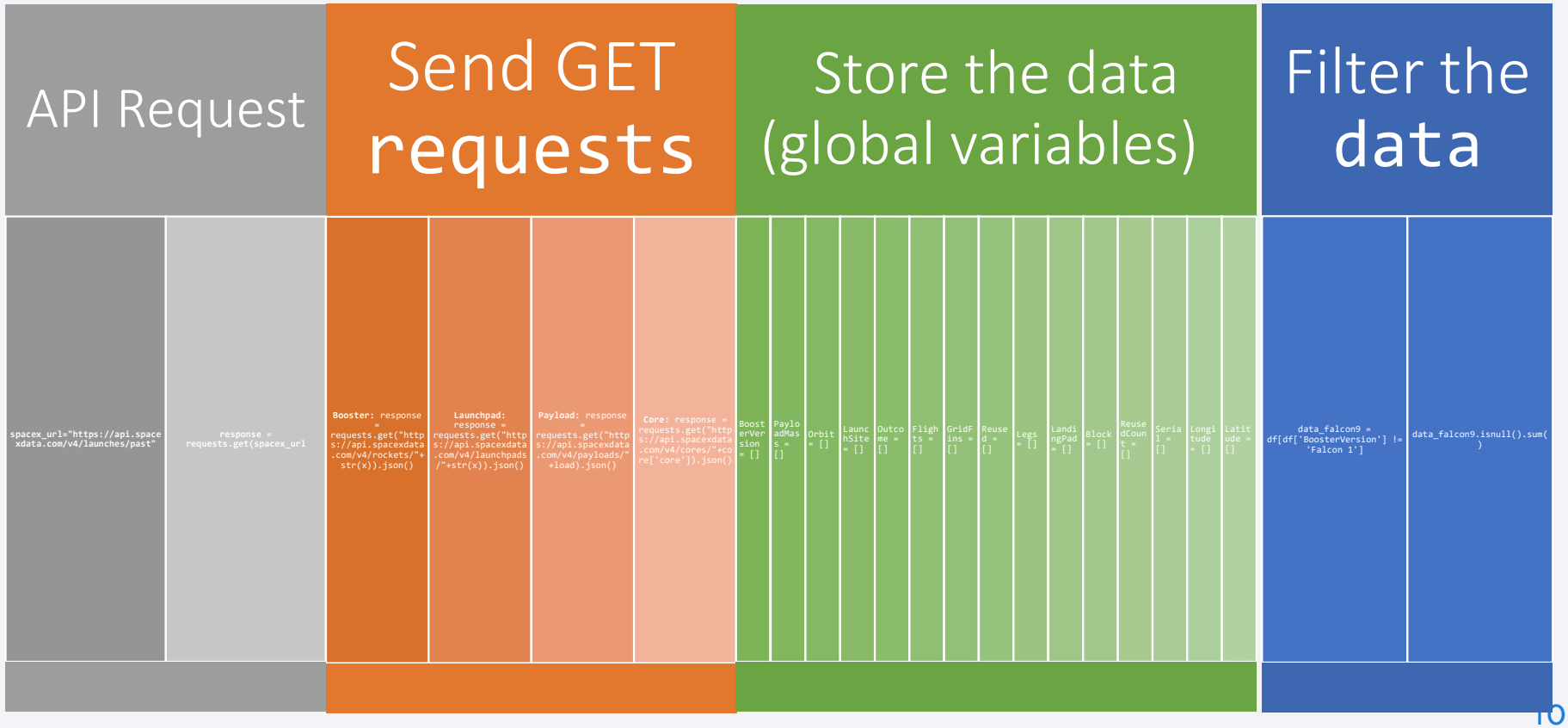
- Data collection methodology:
  - Web scraping, API usage from SpaceX, Data Wrangling and Cleaning
- Perform data wrangling
  - Data cleaning, data transformation, data integration, data preparation
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Descriptive statistics, data visualization, correlation analysis
- Perform interactive visual analytics using Folium and Plotly Dash
  - Interactive maps, trajectory mapping, dashboards and comparative analysis
- Perform predictive analysis using classification models
  - Feature engineering, data splitting, model selection, training and evaluation, and interpretation

Data Collection - API

- Data was collected using SpaceX API to extract information from 4 different datasets
  - Rocket, launchpad, payload, and cores dataset
  - Datasets were combined as needed and filtered for Falcon 9 launches

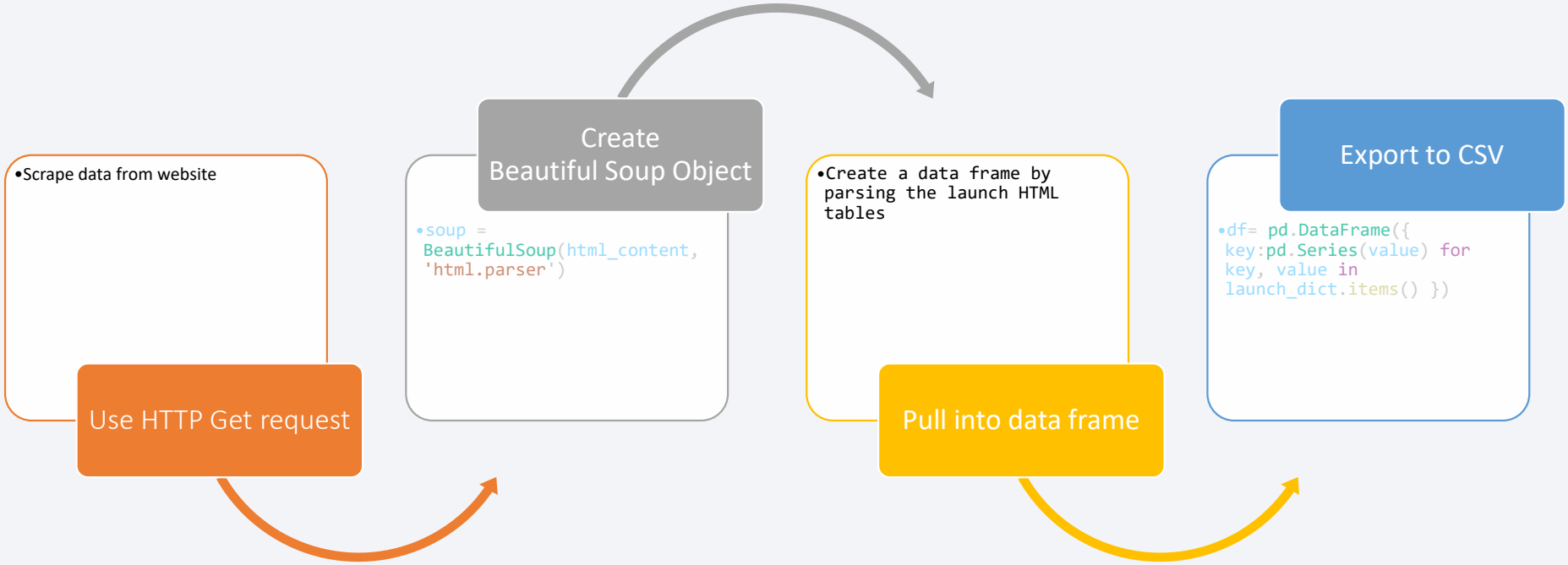


Data Collection – SpaceX API

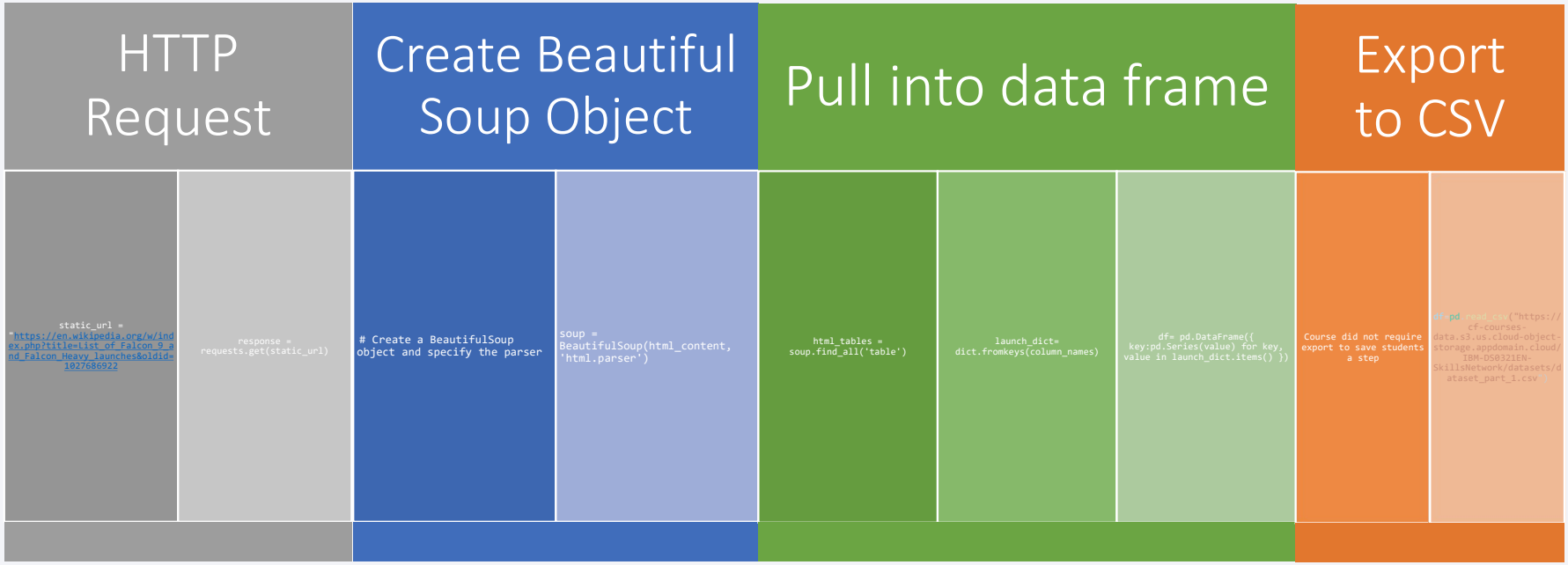


- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Data Collection - Web Scraping



Data Collection - Scraping



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Data Wrangling

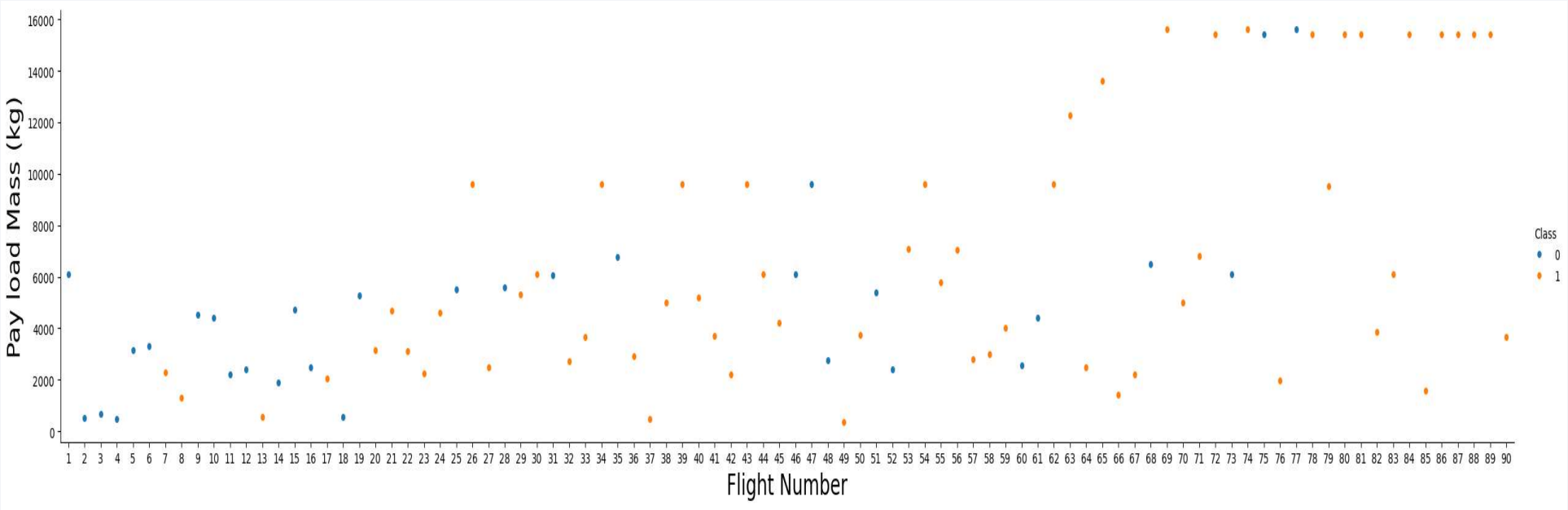
- Describe how data were processed
  - Import Necessary Libraries
  - Load Data (from CSV file)
  - Inspect Data (for missing values and data types)
  - Handle Missing Values (fill or drop)
  - Convert Data Types (to appropriate formats)
  - Create New Features (from existing data)
  - Normalize or Scale Numerical Features
  - Save Cleaned Data (to new CSV file)

- `df.isnull().sum()/len(df)*100`
- `# landing_outcomes = values on Outcome column`
  - `landing_outcomes = df['Outcome'].value_counts()`
- `df['Outcome'].value_counts()`
  - `for i,outcome in enumerate(landing_outcomes.keys()):`
  - `print(i,outcome)`
- `bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])`
- `# landing_class = 0 if bad_outcome`
- `# landing_class = 1 otherwise`
  - `landing_class = [0 if outcome in bad_outcomes else 1 for outcomes in df['Outcome']]`

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

EDA with Data Visualization

- This chart presents an overview of the flights and payload mass
- Aids in seeing patterns between payload mass and successful landings



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

EDA with Data Visualization

- The following charts were plotted to identify relationships between different variables
  - Scatter Plots
    - Illustrates relationship between Payload Mass and Flight Number
    - Illustrates relationship between Lauch Site and Flight Number
    - Illustrates relationship between Launch Site and Payload Mass
    - Illustrates relationship between Orbit Type and Flight Number by classification
    - Illustrates relationship between Orbit Type and Payload Mass by classification
  - Bar Chart
    - Success Launch and Orbit Type
  - Line Chart
  - Summarize what charts were plotted and why you used those charts

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

EDA with SQL

The following queries were performed on the SpaceX launch datasets:

- 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 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.

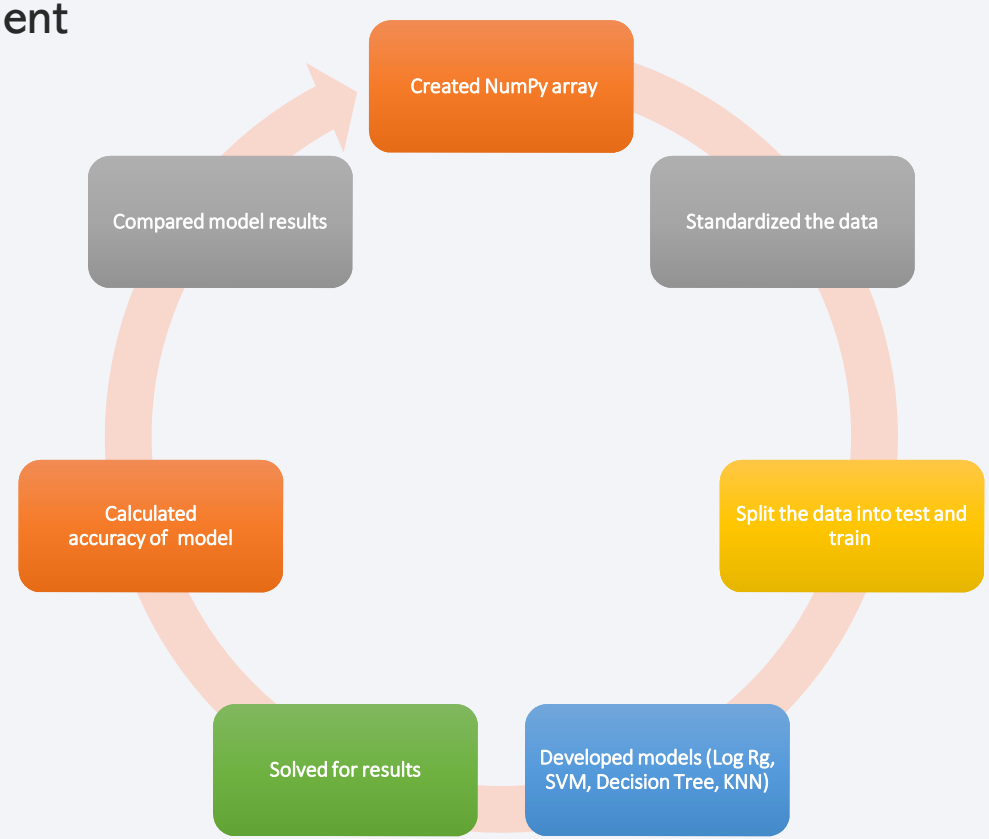
- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

EDA with SQL

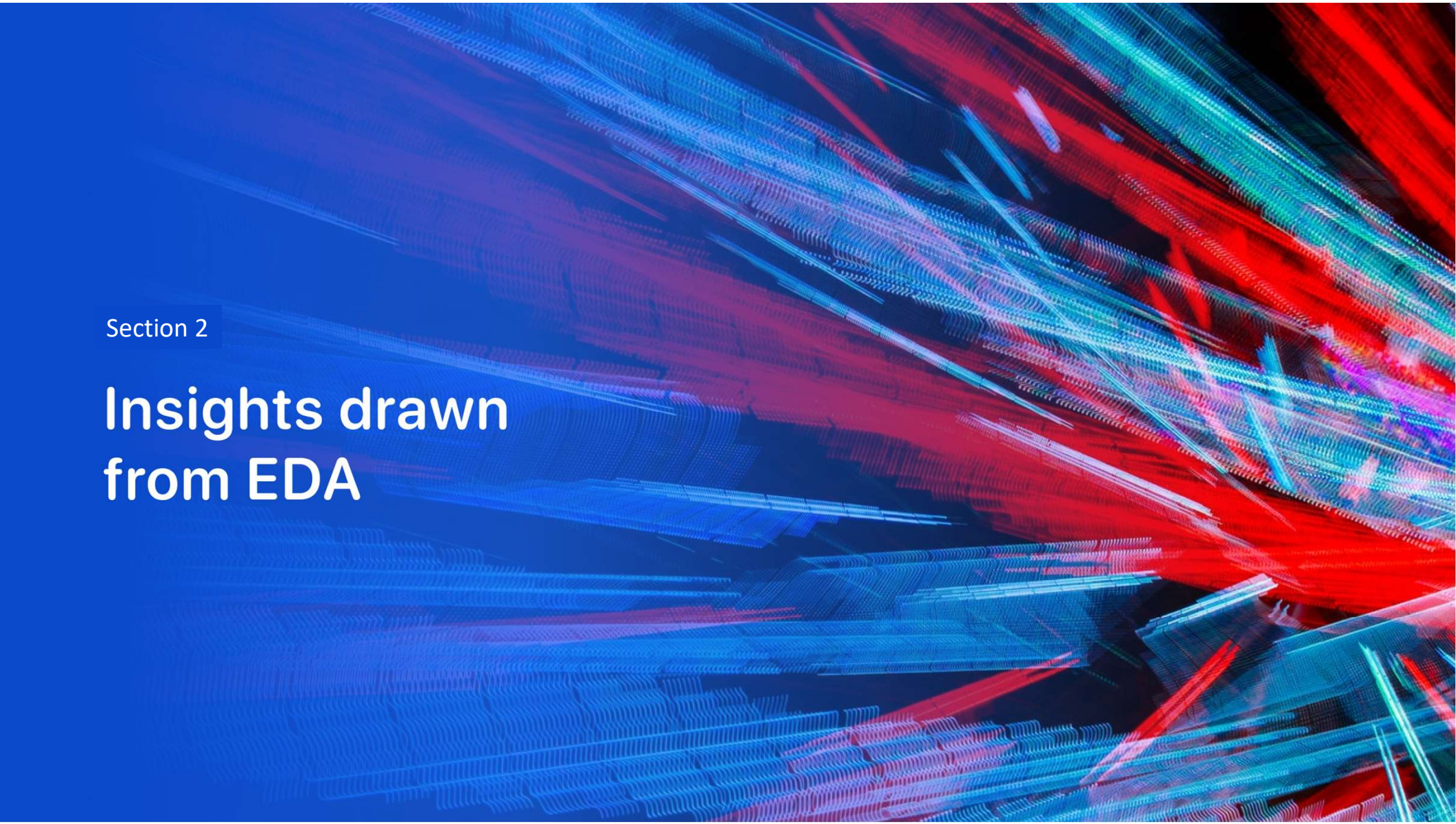
- Found launch sites: ['CCAFS LC-40' 'VAFB SLC-4E' 'KSC LC-39A' 'CCAFS SLC-40']
- Found the total payload mass carried by boosters launched by NASA: 45596 kg.
- Found the average payload mass by Falcon 9: version F9 v1.1 is 2928.4 kg.
- Found the 1st successful landing outcome on a ground pad: achieved on 2015-12-22.
- Found boosters with successful drone ship landing withing a range of payload mass: < 4000 but < 6000 are:
  - F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2
- Found outcomes of launches: Mission Outcome - Success 9, 8 Failure (in flight), 1 Success (payload status unclear), 1 Success 1
- Found boosters teat carried the max payload
  - ('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 ',)
- 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.
  - ('01', 'F9 v1.1 B1012', 'CCAFS LC-40', 'Failure (drone ship)') ('04', 'F9 v1.1 B1015', 'CCAFS LC-40', 'Failure (drone ship)')
- Found landing outcomes between a range of dates:
  - ('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)
- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Predictive Analysis (Classification)

Model Development



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

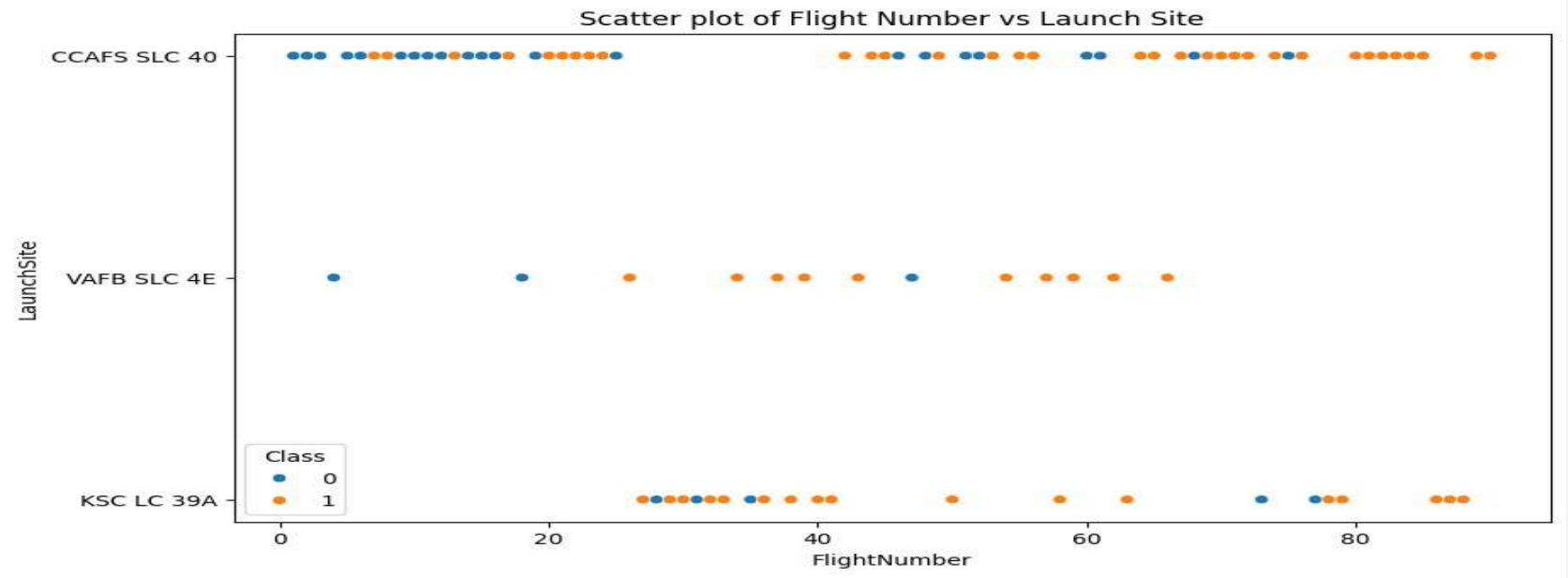


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

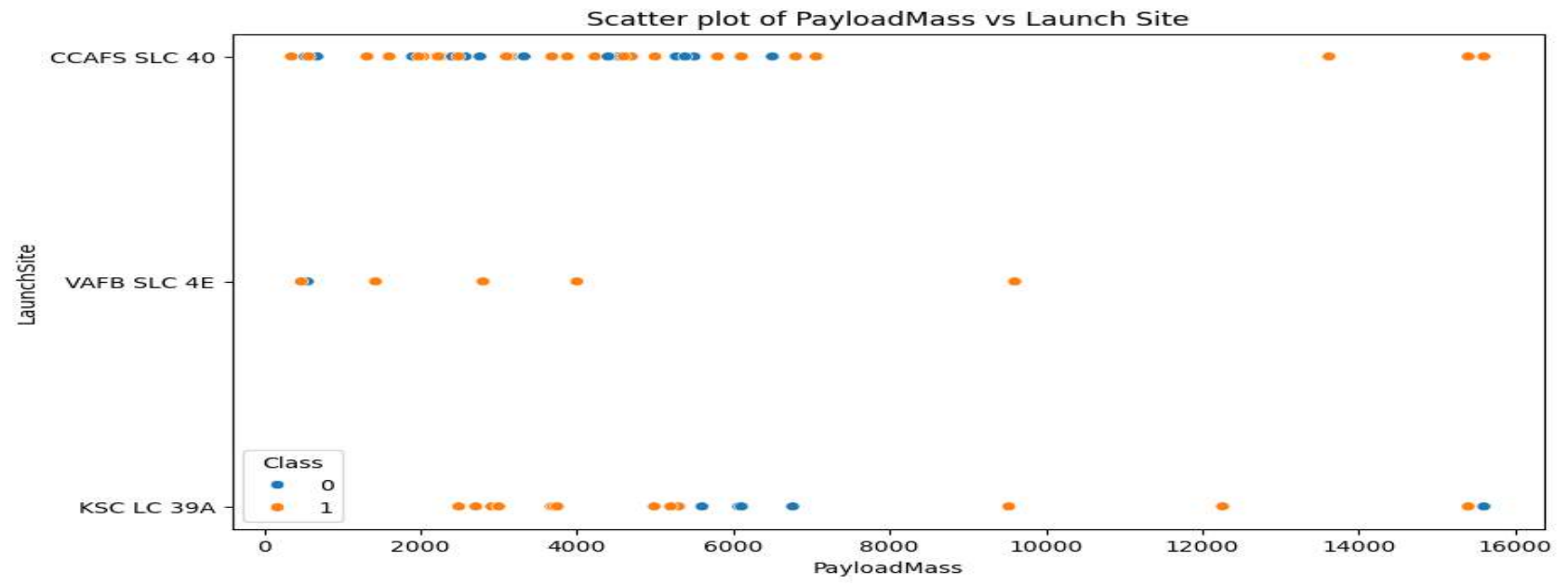
- This chart presents the launch sites, flight number and classification
- Aids in identifying patterns of successful launches by site and flight number



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Payload vs. Launch Site

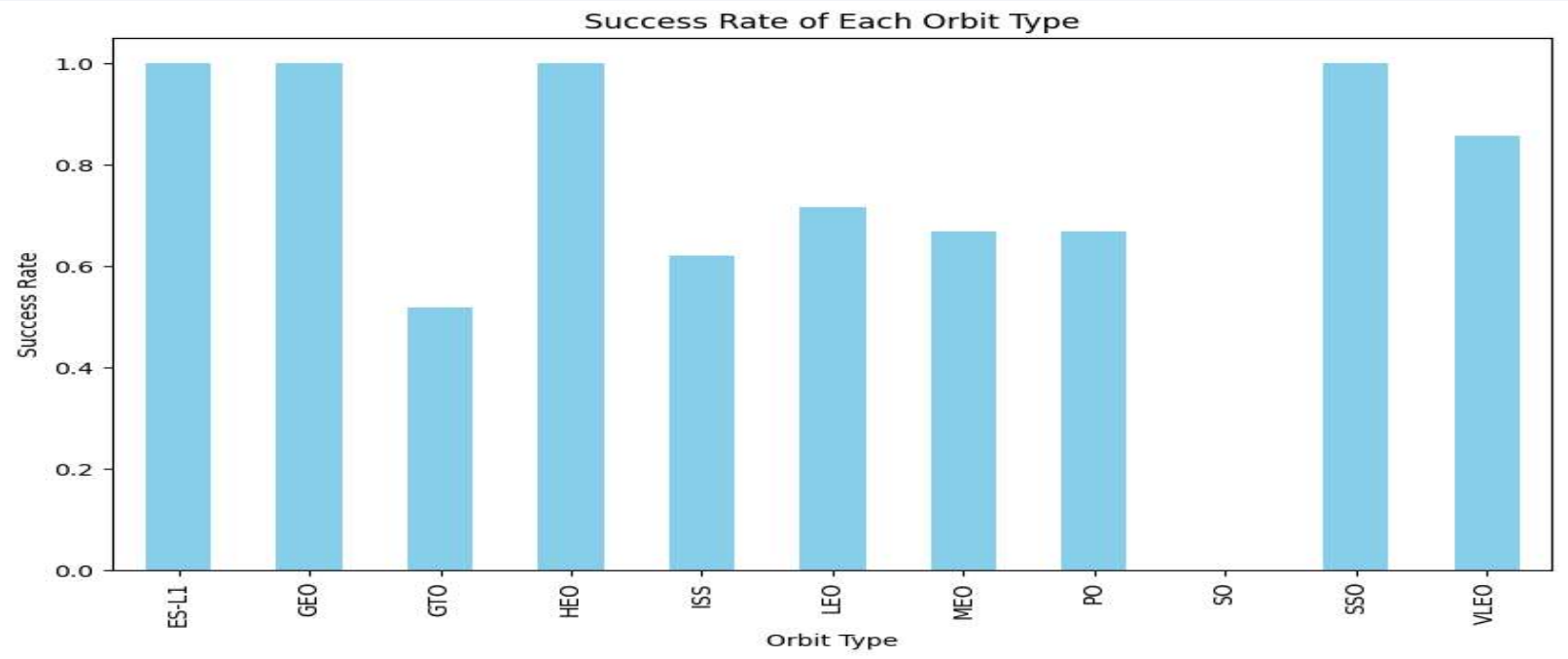
- This chart presents payload mass by launch site with classification
- Aids in identifying patterns of successful launches by site and payload mass



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Succes Rate vs. Orbit Type

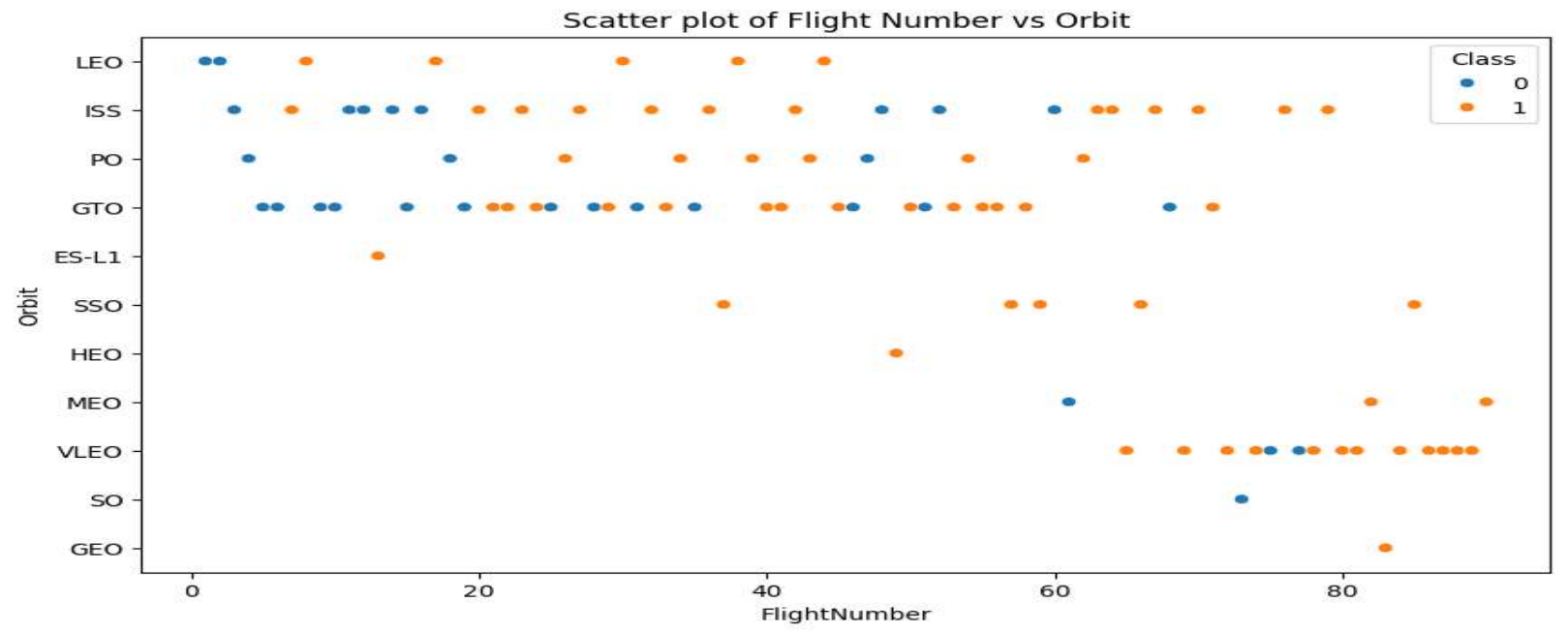
- This chart presents success rate of the launch by orbit type
- Aids in identifying patterns of successful launches by orbit



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Flight Number vs. Orbit Type

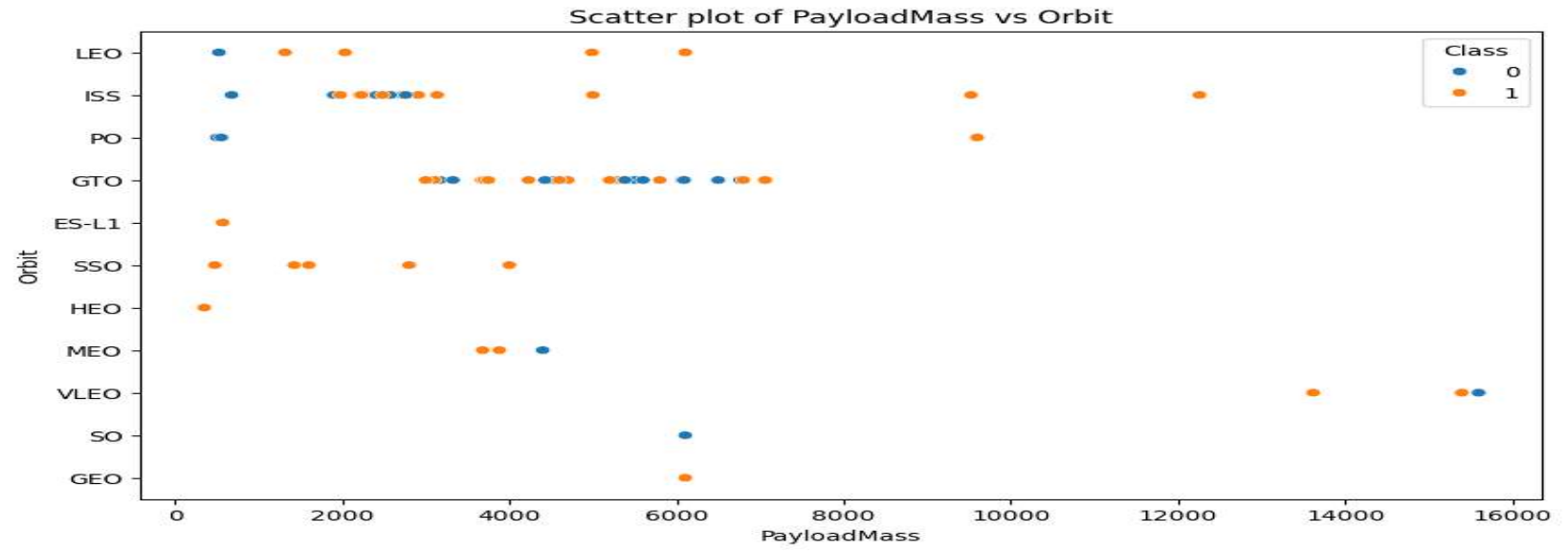
- This chart presents launches by orbit and flight number
- Aids in identifying patterns of successful launches by orbit and flight



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Payload vs. Orbit Type

- This chart presents launches by orbit and payload mass
- Aids in identifying patterns of successful launches by orbit and payload mass



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

All Launch Site Names

- ['CCAFS LC-40' 'VAFB SLC-4E' 'KSC LC-39A' 'CCAFS SLC-40']

```
# Display the unique launch sites
unique_launch_sites = df['Launch_Site'].unique()
print(unique_launch_sites)
```

- Query uses Launch_Site field to identify all launch site names

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Launch Site Names Begin with 'CCA'

- First 5 records where launch sites begin with `CCA`

```
# Filter the dataframe
filtered_df = df[df['Launch_Site'].str.startswith('CCA')]

# Display the first 5 records
print(filtered_df.head(5))
```

```
         Date Time (UTC) Booster_Version  Launch_Site  \
0  2010-06-04   18:45:00  F9 v1.0  B0003  CCAFS LC-40
1  2010-12-08   15:43:00  F9 v1.0  B0004  CCAFS LC-40
2  2012-05-22    7:44:00  F9 v1.0  B0005  CCAFS LC-40
3  2012-10-08    0:35:00  F9 v1.0  B0006  CCAFS LC-40
4  2013-03-01   15:10:00  F9 v1.0  B0007  CCAFS LC-40

                                             Payload  PAYLOAD_MASS__KG_  \
0                    Dragon Spacecraft Qualification Unit                 0
1  Dragon demo flight C1, two CubeSats, barrel of...                      0
2                                  Dragon demo flight C2                525
3                                             SpaceX CRS-1              500
4                                             SpaceX CRS-2              677

       Orbit         Customer Mission_Outcome      Landing_Outcome
0        LEO           SpaceX         Success  Failure (parachute)
1  LEO (ISS)  NASA (COTS) NRO         Success  Failure (parachute)
2  LEO (ISS)      NASA (COTS)         Success           No attempt
3  LEO (ISS)       NASA (CRS)         Success           No attempt
4  LEO (ISS)       NASA (CRS)         Success           No attempt
```

- All at the LC-40 site and Orbit

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Total Payload Mass

- Total payload mass by boosters launched by NASA (CRS) is 45596 kg.

```
# Filter the dataframe for rows where the 'Customer' is 'NASA (CRS)'
nasa_crs_df = df[df['Customer'] == 'NASA (CRS)']

# Calculate the total payload mass
total_payload_mass = nasa_crs_df['PAYLOAD_MASS__KG_'].sum()

print(f"The total payload mass carried by boosters launched by NASA (CRS) is {total_payload_mass} kg.")
```

- The filtered dataframe runs payloads launched by NASA

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

```
# Filter the dataframe for rows where the 'Booster Version' is 'F9 v1.1'
f9_v1_1_df = df[df['Booster_Version'] == 'F9 v1.1']

# Calculate the average payload mass
average_payload_mass = f9_v1_1_df['PAYLOAD_MASS__KG_'].mean()

print(f"The average payload mass carried by booster version F9 v1.1 is {average_payload_mass} kg.")
```

- The filtered dataframe for the F9 booster calculates the mean payload mass

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved on 2015-12-22.

```
# Filter the dataframe for rows where the 'Landing Outcome' is 'Success (ground pad)'
success_ground_pad_df = df[df['Landing_Outcome'] == 'Success (ground pad)']

# Find the earliest date
first_success_date = success_ground_pad_df['Date'].min()

print(f"The first successful landing outcome in ground pad was achieved on {first_success_date}.")
```

- The filtered data frame for "Success (ground pad)" uses a minimum date with the defined criteria

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

```
#Filter the dataframe for rows where 'Booster_Version' is >

# Filter the dataframe for rows where the 'Landing Outcome' is 'Success (drone ship)'
# and 'Payload Mass (kg)' is greater than 4000 but less than 6000
filtered_df = df[(df['Landing_Outcome'] == 'Success (drone ship)') &
                 (df['PAYLOAD_MASS__KG_'] > 4000) &
                 (df['PAYLOAD_MASS__KG_'] < 6000)]

# Get the unique booster names
unique_booster_names = filtered_df['Booster_Version'].unique()

print("The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are:")
for name in unique_booster_names:
    print(name)
```


- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Total Number of Successful and Failure Mission Outcomes

- Mission Outcomes:
  - Success 98
  - Failure (in flight) 1
  - Success (payload status unclear) 1
  - Success 1
  - Name: Mission_Outcome, dtype: int64

```
# Get the counts of each unique value in the 'Mission Outcome' column
mission_outcomes = df['Mission_Outcome'].value_counts()

# Print the counts
print(mission_outcomes)
```

- The query asks for value counts for mission outcomes

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Boosters Carried Maximum Payload

```
#List booster names which have carried the maximum payload

# Write your SQL query
query = """
SELECT "Booster_Version"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTBL
)
"""

# Execute the query
cur.execute(query)

# Fetch all the results
results = cur.fetchall()

# Print the results
for result in results:
    print(result)
```

- ('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 ',)
- Queries by Booster Version and maximum payload mass

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

2015 Launch Records

- ('01', 'F9 v1.1 B1012', 'CCAFS LC-40', 'Failure (drone ship)')
- ('04', 'F9 v1.1 B1015', 'CCAFS LC-40', 'Failure (drone ship)')
- Queries by year failed outcome, drone ship, and booster version

```
# Write your SQL query
query = """
SELECT substr(Date, 6, 2) AS Month, "Booster_Version", "Launch_Site", "Landing_Outcome"
FROM SPACEXTBL
WHERE substr(Date, 0, 5) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)'
"""

# Execute the query
cur.execute(query)

# Fetch all the results
results = cur.fetchall()

# Print the results
for result in results:
    print(result)
```


- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

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

```
# Write your SQL query
query = """
SELECT "Landing_Outcome", COUNT(*) as Count
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY Count DESC
"""

# Execute the query
cur.execute(query)

# Fetch all the results
results = cur.fetchall()

# Print the results
for result in results:
    print(result)
```

- ('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)
- Query uses landing outcome and date ranges

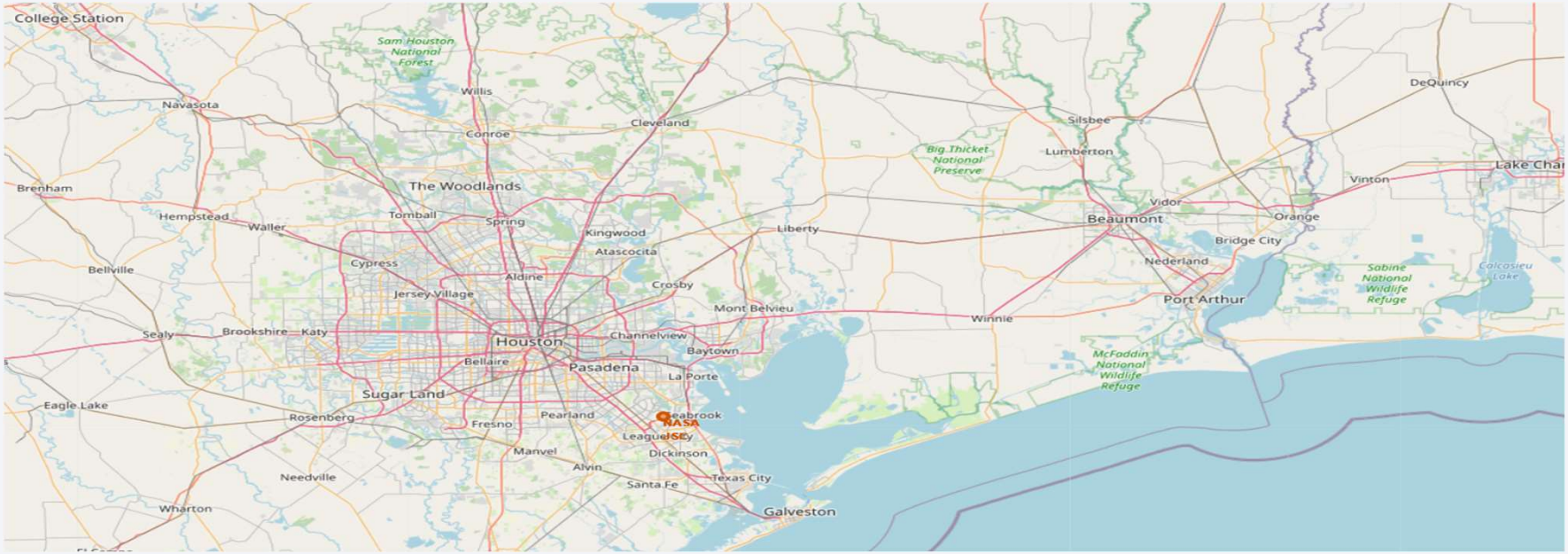
- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Section 3

Launch Sites Proximities Analysis

Folium Interactive Map

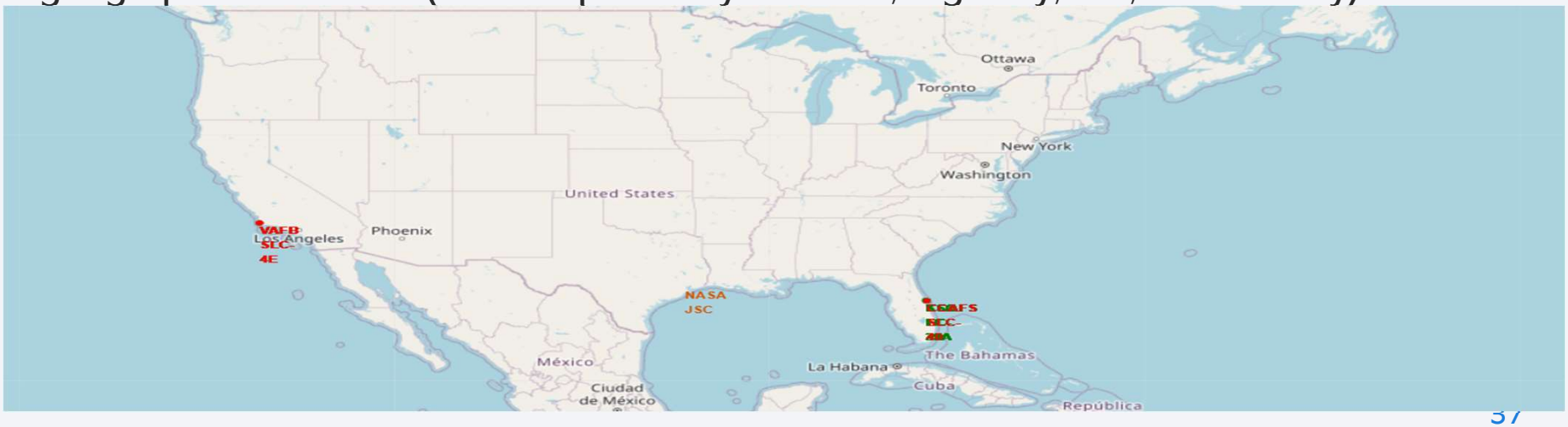
- Marker added to see the location of the NASA flights in TX



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Folium Interactive Map

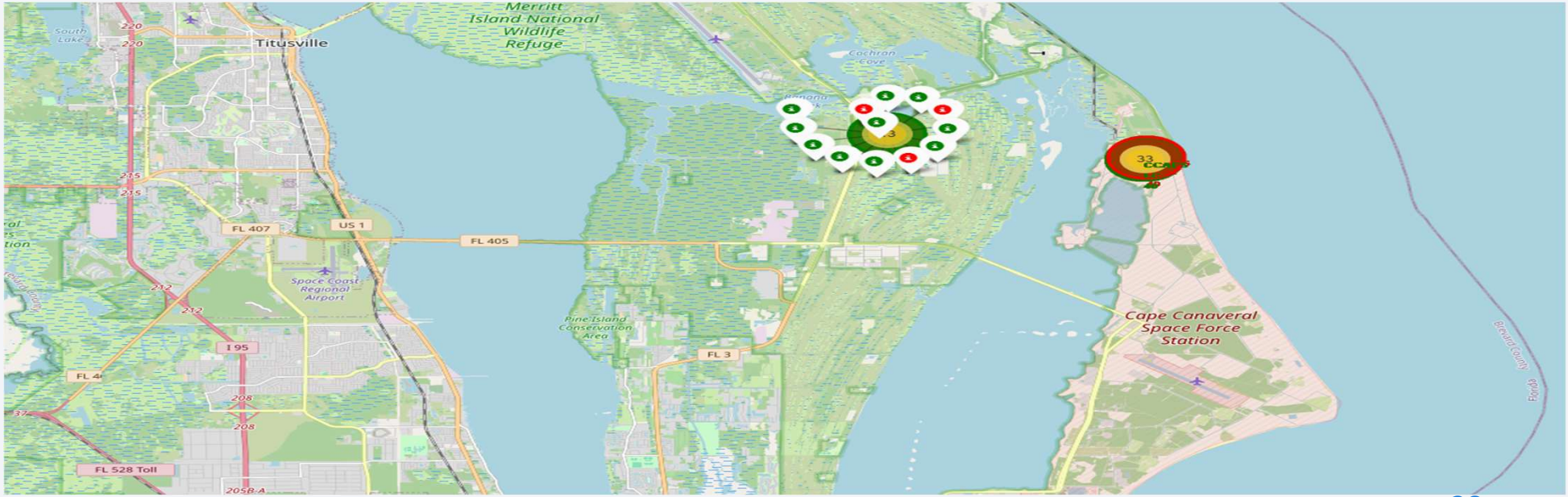
- The map below presents the multiple launch sites for the Falcon 9 rockets across the US
- These were added to see the locations for each launch and proximity to local geographical features (such as proximity to coast, highway, rail, nearest city)



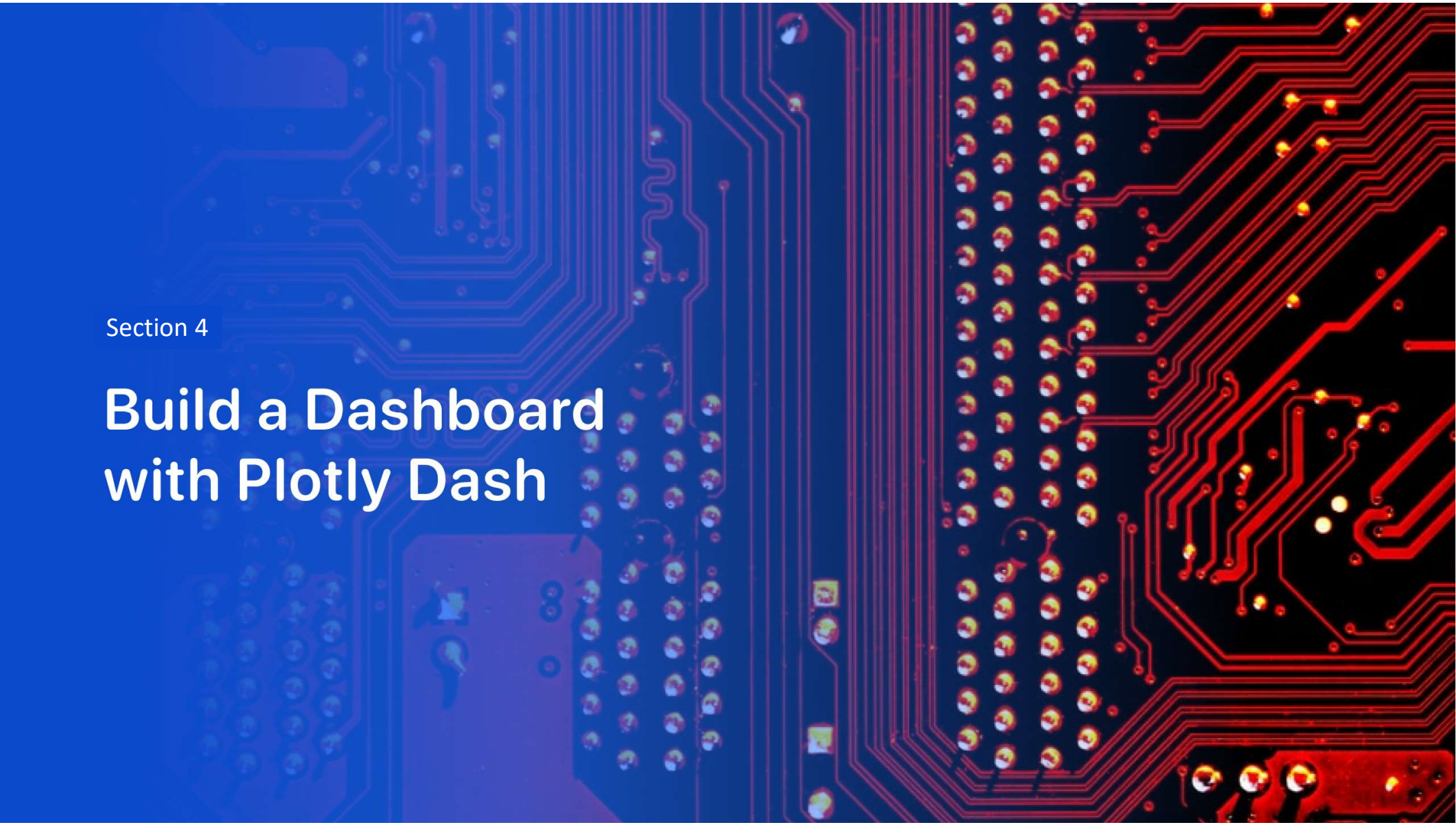
- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Folium Interactive Map

- The map below shows the launch sites off the coast of FL and the number of launches at each site with indicator for successful launches



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

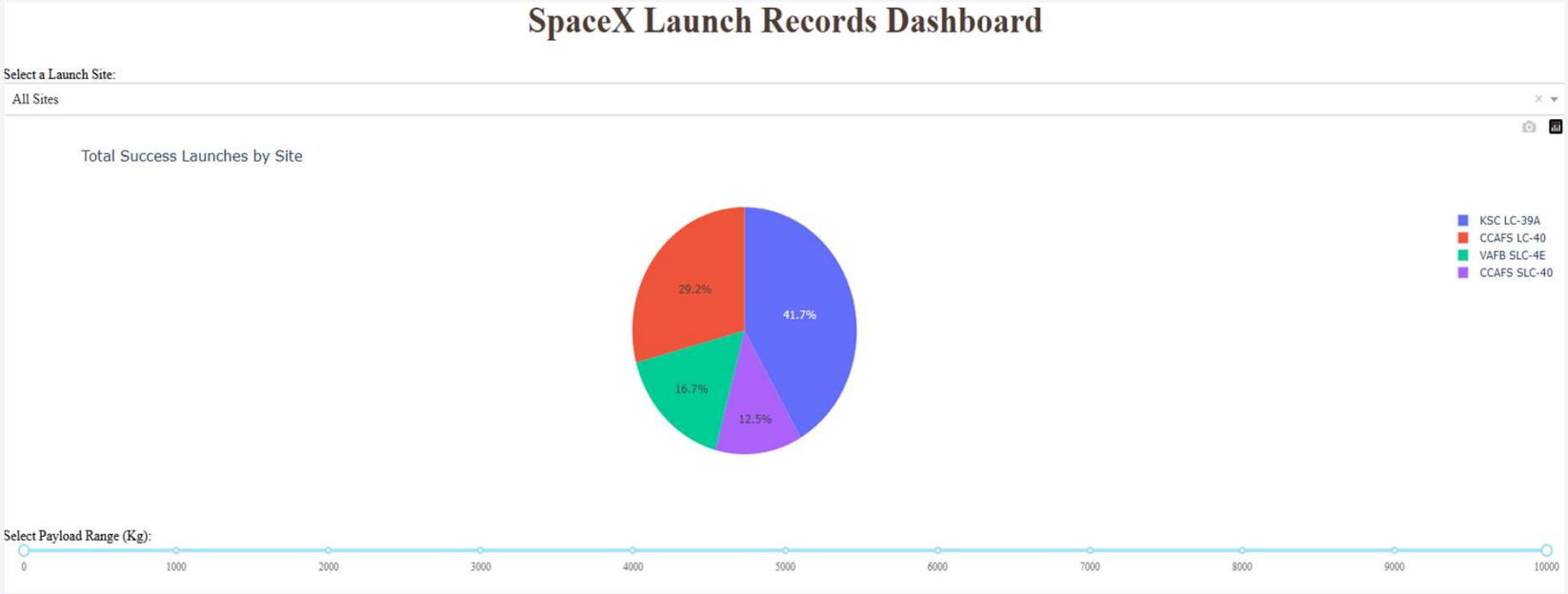


Section 4

Build a Dashboard with Plotly Dash

Dashboard with Plotly Dash

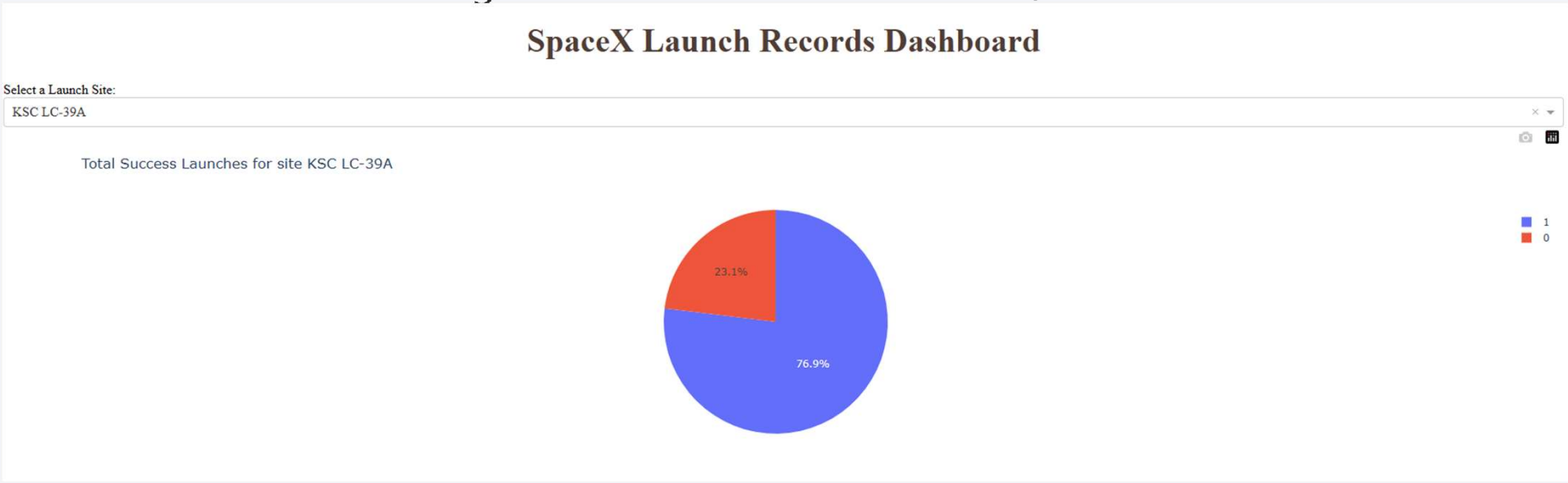
- The interactive dashboard presented below allows the user to sort by launch site and use the slider to adjust for payload range



- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

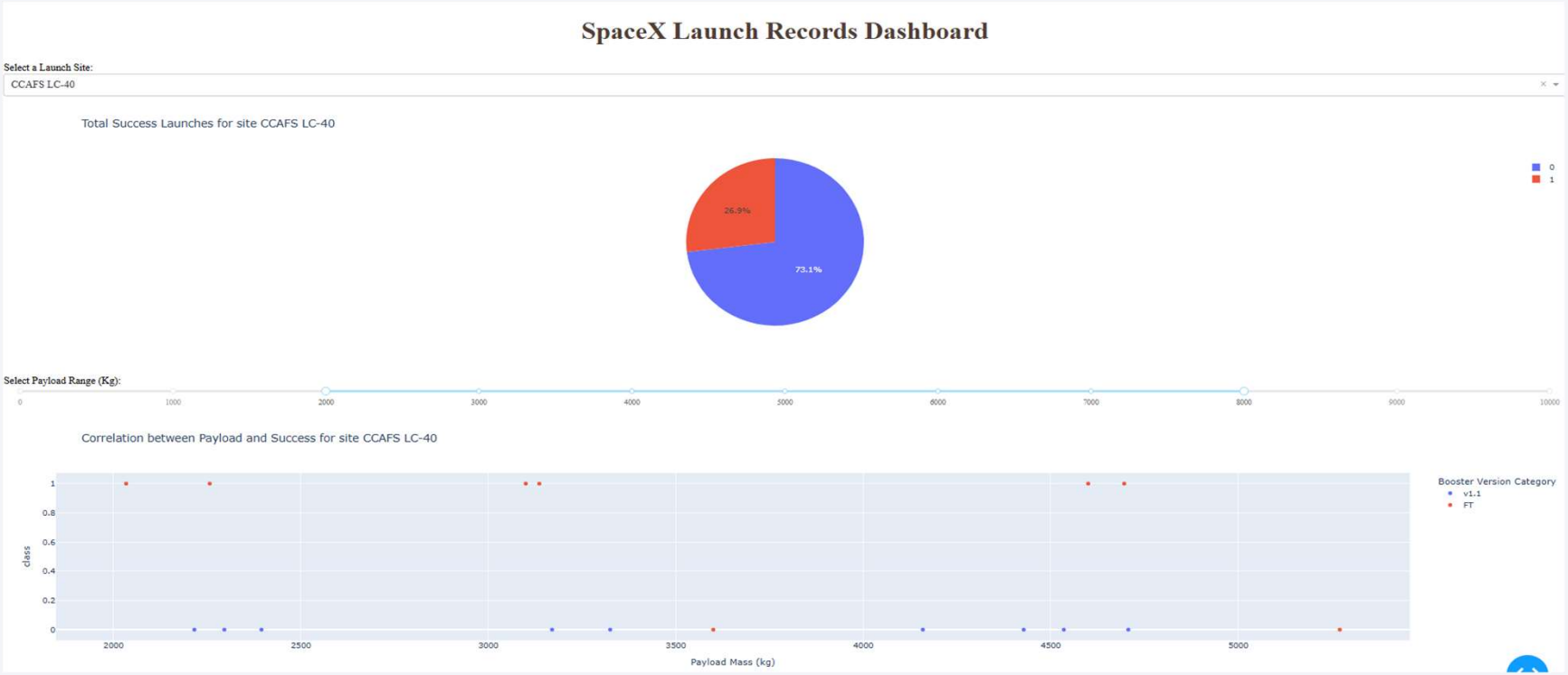
Launch Site Dashboard – KSC LC-39A

- The site below has the higher launch success rate of 76.9%



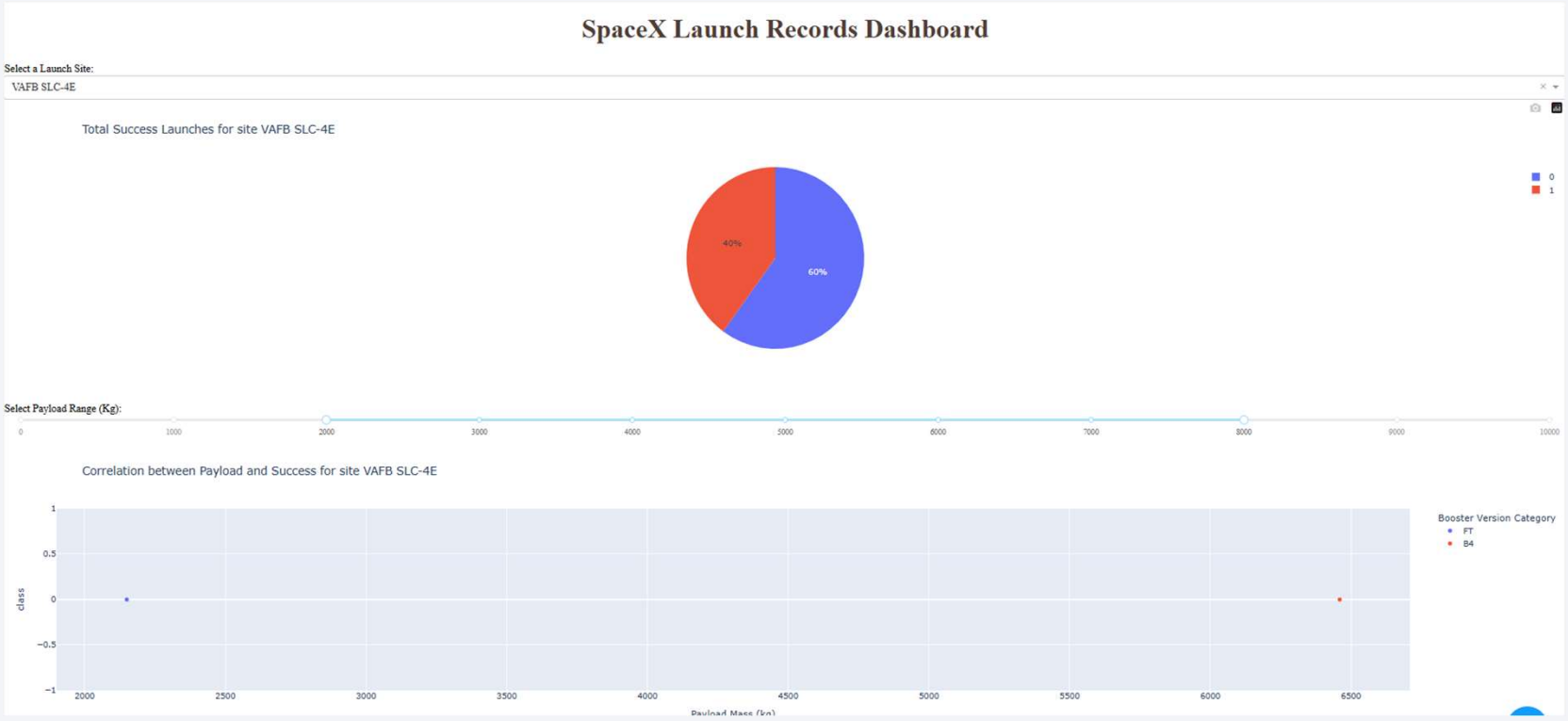
Launch Site Dashboard – CCAFS LC-40

- The results below show the launch success rate at the site below with the payload range from 2,000 to 8,000Kg



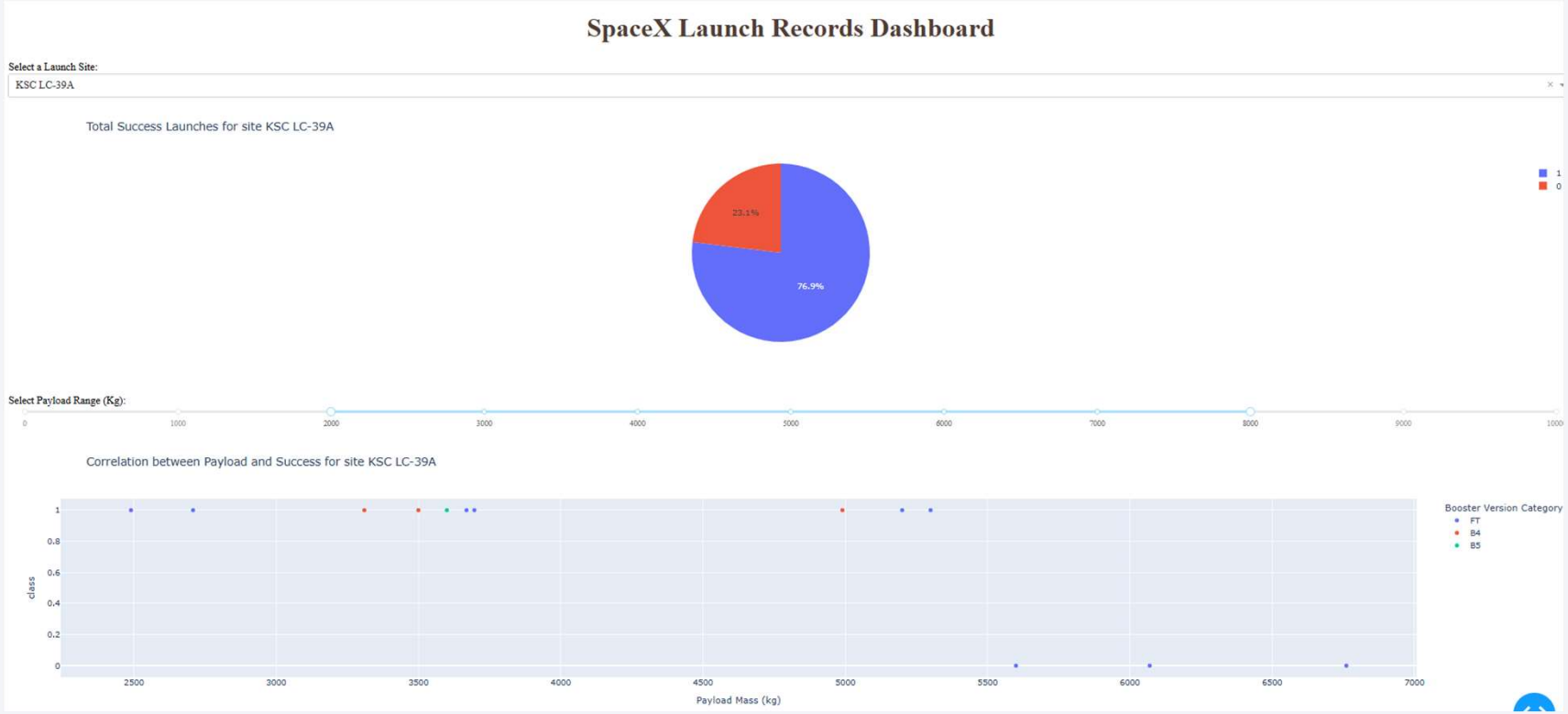
Launch Site Dashboard – VAFB SLC-4E

- The results below show the launch success rate at the site below with the payload range from 2,000 to 8,000Kg



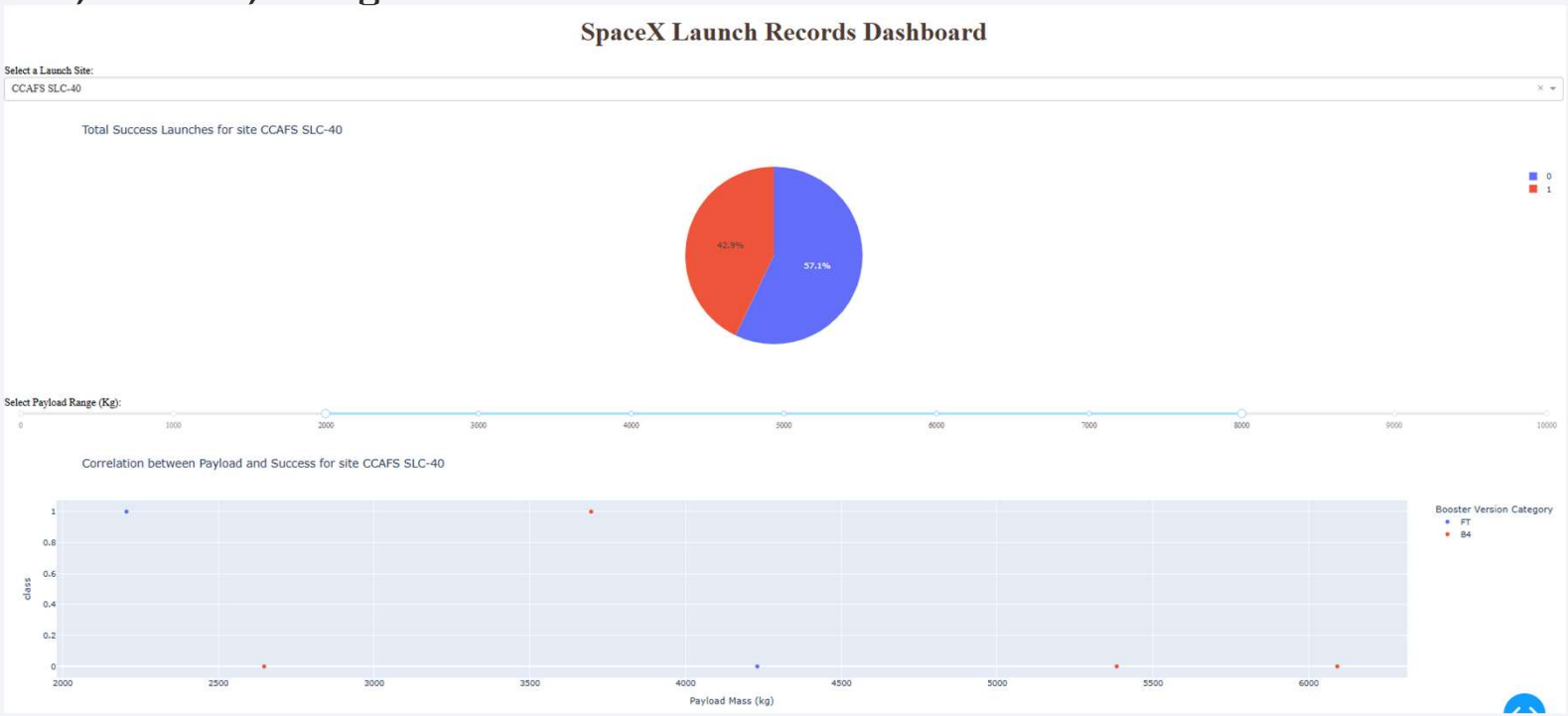
Launch Site Dashboard – KSC LC-39A

- The results below show the launch success rate at the site below with the payload range from 2,000 to 8,000Kg



Launch Site Dashboard – CCAFS SLC-40

- The results below show the launch success rate at the site below with the payload range from 2,000 to 8,000Kg



Section 5

Predictive Analysis (Classification)

Predictive Analysis (Classification)

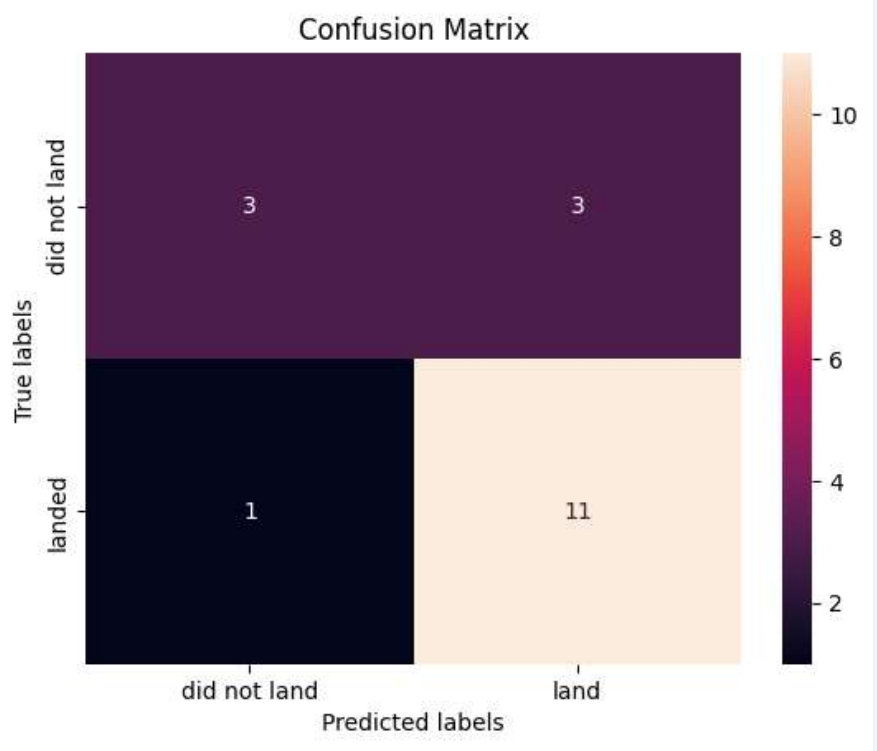
Model Results:

- KNN
  - Best parameters found: {'algorithm': 'auto', 'n_neighbors': 1, 'p': 1}
  - Best score: 0.7535714285714286
  - Accuracy: 0.7222222222222222
- Decision Tree
  - Best parameters found: {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}
  - Best score: 0.7928571428571428
  - Accuracy: 0.6111111111111112
- Logistic Regression
  - Best parameters found: {'C': 1, 'penalty': 'l2', 'solver': 'lbfgs'}
  - Best score: 0.7446428571428572
  - Accuracy: 0.8333333333333334
- SVM
  - Best parameters found: {'C': 1000.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
  - Best score: 0.8321428571428571
  - Accuracy : 0.7777777777777778

- Applied-Data-Science-Capstone/Capston_Wk1_Data.ipynb at main · osmllc/Applied-Data-Science-Capstone (github.com)

Confusion Matrix

- SVM Confusion Matrix
  - 11 True positives for land
  - 3 True positives for did not land
- Other considerations:
  - Accuracy: 0.78 measures total number of true results
  - Precision: 0.79
  - Sensitivity: 0.92
  - F1 Score: 0.85



Conclusions

- To assist SpaceX competitors in bidding against SpaceX for launch contracts the following the recommendations are made:
  - The model with the best score was the SVM: 0.83, accuracy of 0.77
    - Best parameters found: {'C': 1000.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
- The following methodologies were used:
  - Data Collection and Data Wrangling
  - Exploratory Data Analysis
  - Machine Learning
  - Predictive Modeling
  - Advanced Data Visualization
  - Working With Big Data
  - Capstone Project
  - Soft Skills
- The skills, tools and data science techniques to solve the problem presented were appropriate and effective in finding the needed data for Space X competitors to bid against SpaceX for launch contracts

Thank you!