IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohamed Yassine Cherif
04/08/2024

Outline

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

Executive Summary

- The applied methodologies in this project involve collecting, web scraping and wrangling data about Falcon 9 launch operations of the SpaceX company, performing exploratory data analysis (EDA) using visualization methods and SQL queries, interactive visual analytics using Folium and Plotly Dash and predictive analysis using machine learning algorithms for building classification models.
- The phase of collecting and preprocessing data was successfully implemented based on the provided templates and the predictive models were well performing giving an accuracy score equal to 0.83.

Introduction

- This project consists in a practical application of the previously acquired data science knowledges and technologies, based on collected and processed data about the Falcon 9 launching and landing operations performed by the SpaceX company.
- The purpose of the project is to predict the possibility of Falcon 9 launching and landing success depending on different circumstances and characteristics of the launch operations.
- The implementation and results of different project phases are available in the following Github repository:

  https://github.com/yassine89/Applied-Data-Science-Capstone

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX API and web-scraped from the Wikipedia page about the list of Falcon 9 heavy launches.
- Perform data wrangling:
  - Performing exploratory data analysis (EDA) to find data patterns.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a column for the class, standardize data and test different classification algorithms with train and test data to find the best performing method.

Data Collection

- The collection of the data sets was based on using the SpaceX API and and web scraping tools to retrieve information from the following Wikipedia page:
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- To retrieve data using the SpaceX API, a ‘GET’ type request is used to receive the data in a JSON file response, while the web scraping task is performed using the beautifulsoup library to extract the data from the web page HTML source code.

Data Collection – SpaceX API

- The data were collected from the SpaceX API by sending a ‘GET’ request, as represented in the figure code.
- The following GitHub URL represents the completed SpaceX API calls notebook:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Data Collection - Scraping

- The data web scraping process was performed using the beautifulsoup library to extract data from HTML page source code, as represented in the figure code.
- The following GitHub URL represents the completed web scraping notebook:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

```
# use requests.get() method with the provided static_url
# assign the response to a object
page = requests.get(static_url)
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(page.text, 'html.parser')
```

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables=soup.find_all('table')
```

Data Wrangling

- Data wrangling consists of performing some exploratory data analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Pandas library is used to transform the data into a dataframe and apply the required statistical operations.
- The following GitHub URL represents the completed data wrangling related notebooks:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory data analysis and feature engineering in this phase were performing by visualizing different charts consisting of scatter point charts to represent the relationship between flight number and launch site, the relationship between payload mass and launch site, the relationship between flight number and orbit type and the relationship between payload mass and orbit type, a bar chart to represent the relationship between success rate of each orbit type and a line chart to represent the launch success yearly trend.
- The following GitHub URL represents the completed EDA with data visualization notebook:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- The exploratory data analysis was performed applying SQL queries to:
- 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 having success in drone ship and a specific payload mass.
- List the total number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the maximum payload mass.
- 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 between two specific dates in descending order.

- The following GitHub URL represents the completed EDA with SQL notebook:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In this phase, some map objects belonging to the folium library such as markers, circles, lines were created and added to the interactive map.
- Circles and markers were used to locate launch sites on the map, markers with different colors were used to locate the successful and failed launch operations for each site on the map, while lines were used to represent the distances between a launch site to its proximities such as coastlines, railways or highways.
- The following GitHub URL represents the completed interactive folium map:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The dashboard contains a drop-down list to enable the launch site selection, a pie chart to show the total count of successful launches for all sites, a slider to select payload mass range and a scatter chart to show the correlation between payload mass, launch success and the booster version category.
- Those plots and interactions were added to the dashboard to provide interactive visual analytics enabling users to explore and manipulate data in an interactive and real-time way.
- The following GitHub URL represents the completed Plotly Dash lab:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The aim of this phase is to build classification models to predict the SpaceX Falcon 9 first stage landing success.
- The used data set was preprocessed by creating the class column, standard-scaling the data and splitting it into train and test subsets, then the predictive models were built testing many classification algorithms while using the grid search method to select the best parameters and evaluating the result by calculating the accuracy score and generating the confusion matrix.
- The following GitHub URL represents the completed predictive analysis lab:

  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

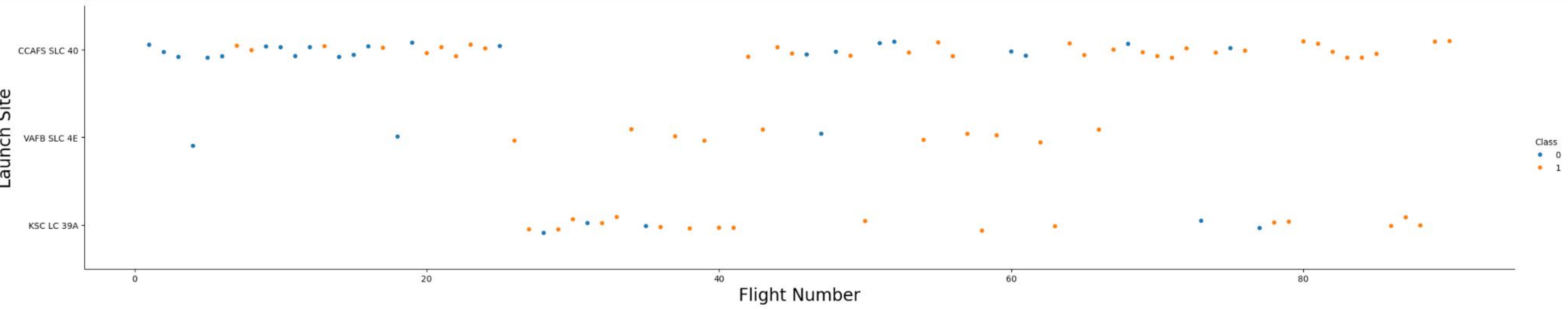
- Exploratory data analysis enabled successfully to find the data patterns and perform the feature engineering task.
- Interactive analytics enabled successfully to explore and manipulate data in an interactive and real-time way.
- Predictive analysis has given successful result with an accuracy score equal to 0.83 using the logistic regression, support vector machine and k nearest neighbors algorithms.

Section 2

Insights drawn from EDA

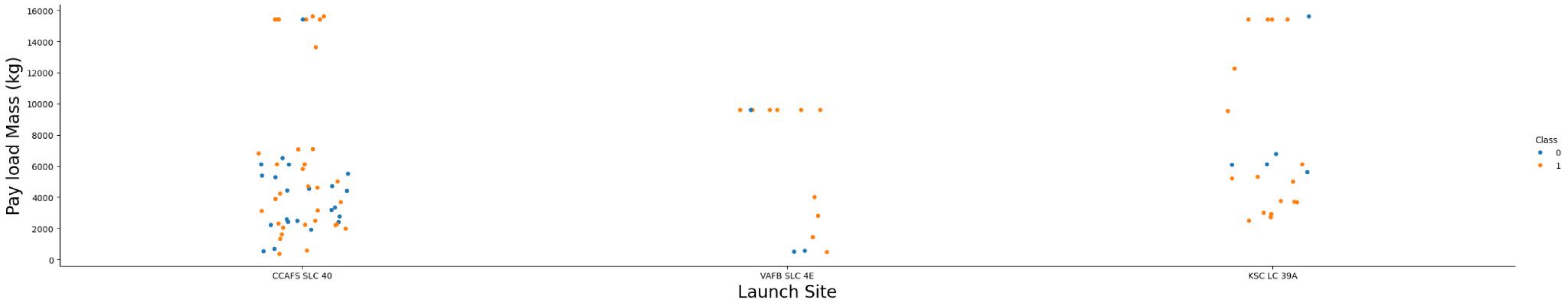
Flight Number vs. Launch Site

- The scatter plot of Flight Number vs. Launch Site



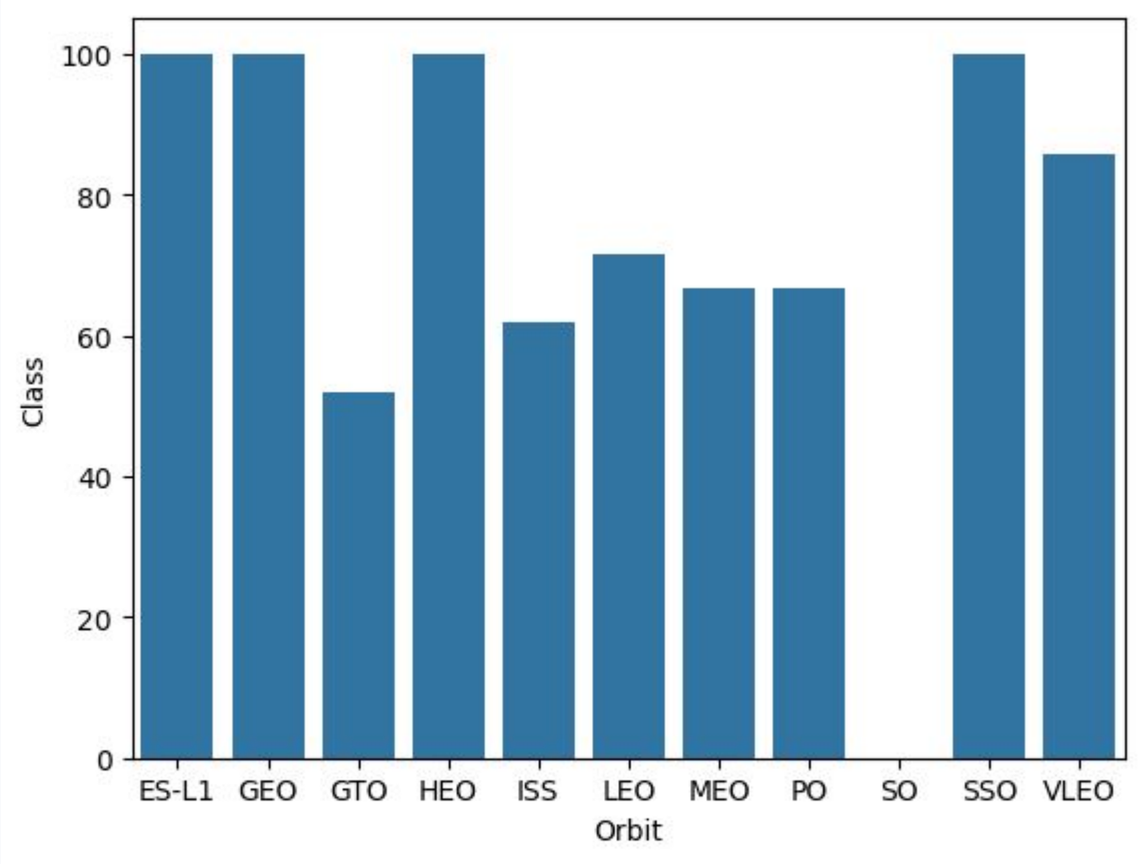
Payload vs. Launch Site

- The scatter plot of Payload vs. Launch Site



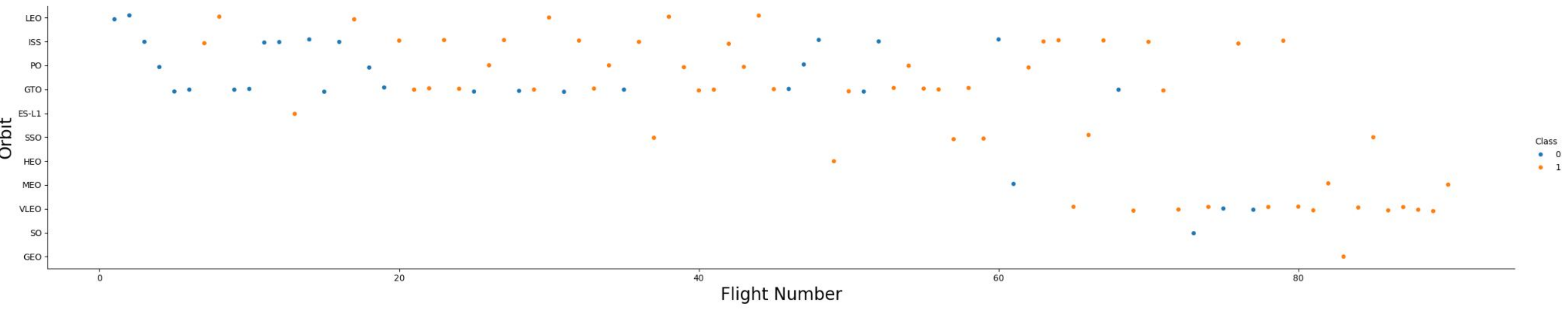
Success Rate vs. Orbit Type

- The bar chart for the success rate of each orbit type



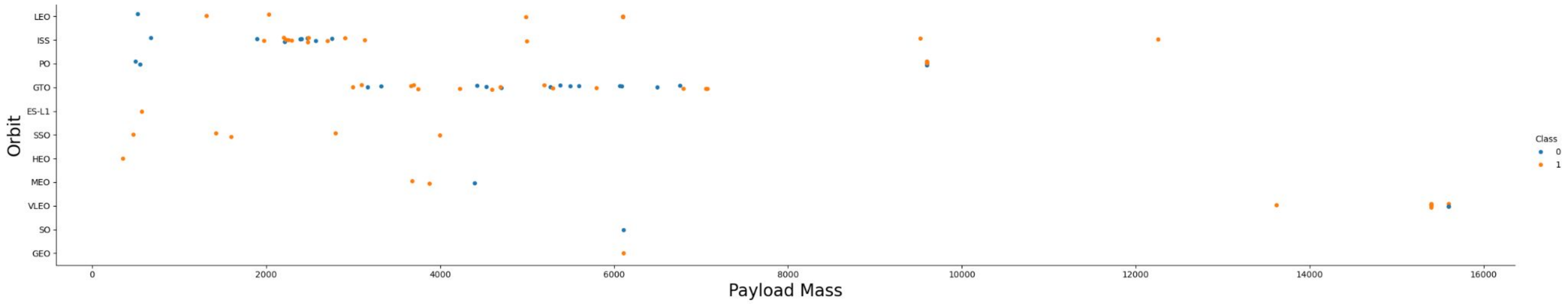
Flight Number vs. Orbit Type

- The scatter point of Flight number vs. Orbit type



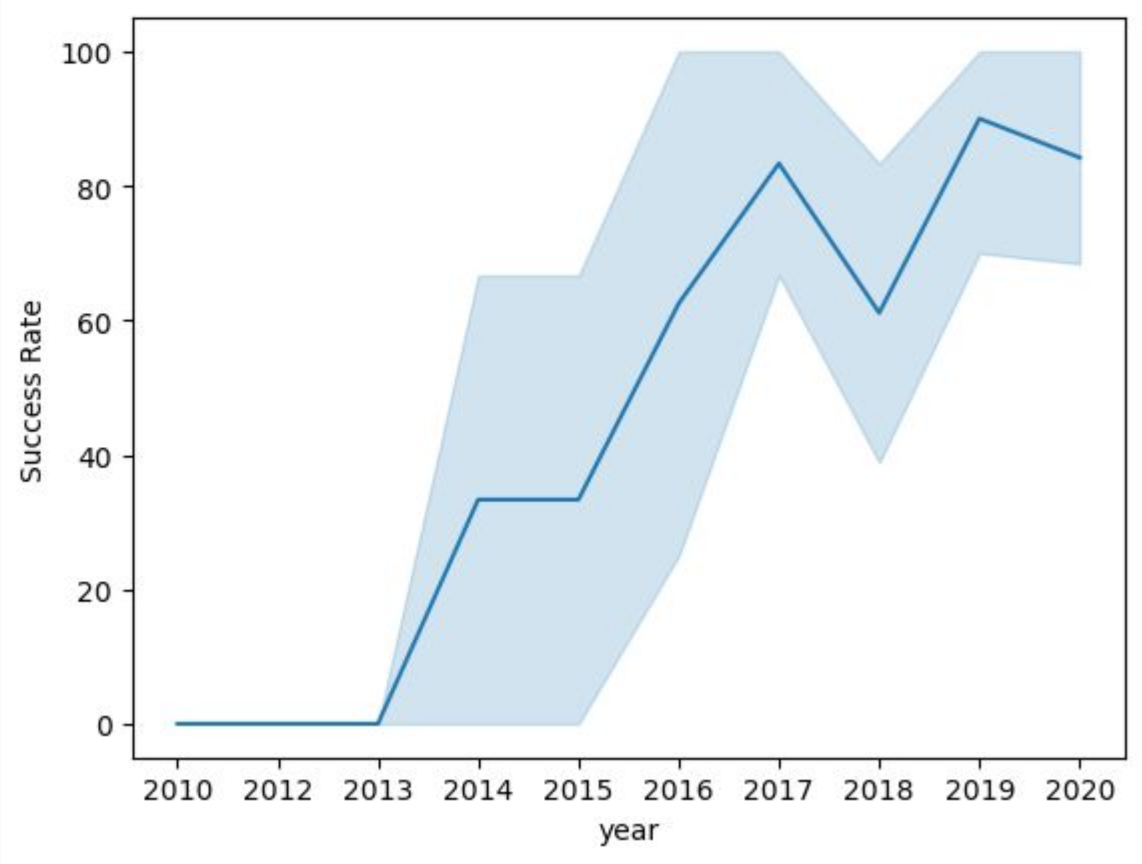
Payload vs. Orbit Type

- The scatter point of payload vs. orbit type



Launch Success Yearly Trend

- The line chart of yearly average success rate



All Launch Site Names

- The result of the query giving the names of the unique launch sites

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- The result of the query giving 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	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

- The result of the query calculating the total payload carried by boosters from NASA that was given equal to 45596

Sum
45596

Average Payload Mass by F9 v1.1

- The result of the query calculating the average payload mass carried by booster version F9 v1.1 that was given equal to 2534.667

Average
2534.6666666666665

First Successful Ground Landing Date

- The result of the query finding the date of the first successful landing outcome on ground pad that was given as: 04-06-2010

Date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- The result of the query giving the list of 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 B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The result of the query calculating the total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

- The result of the query giving the list of 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 result of the query giving the list of the failed landing_outcomes in drone ship, their booster versions and launch site names for the year 2015

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 result of the query ranking 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	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

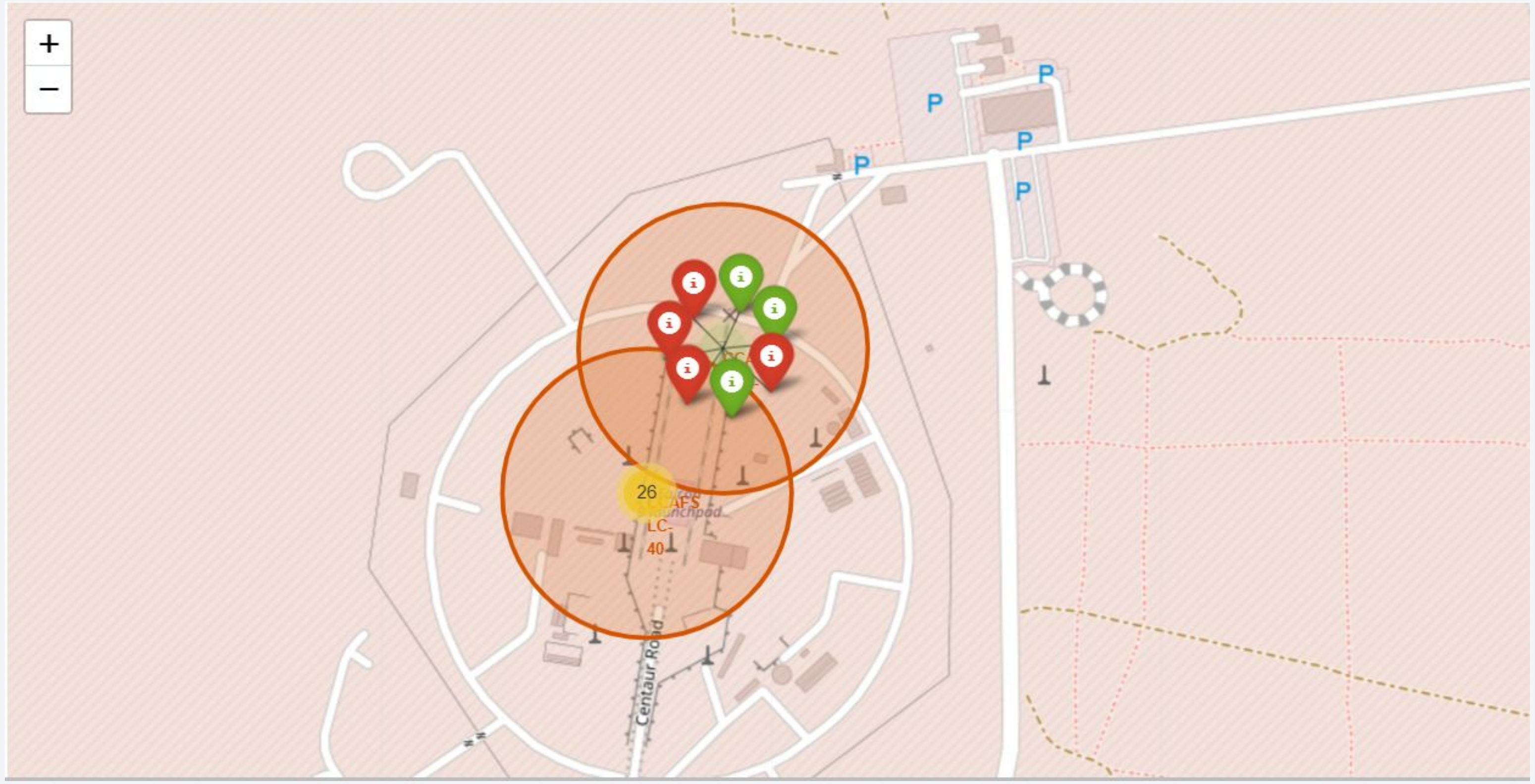
Launch Sites Proximities Analysis

Folium map with launch sites' location markers

- The following figure represents the global folium map including all launch sites' location markers.

Folium map with color-labeled launch outcomes

- The following figure represents the folium map showing the color-labeled launch outcomes.



Folium map with distance of launch site and its proximities

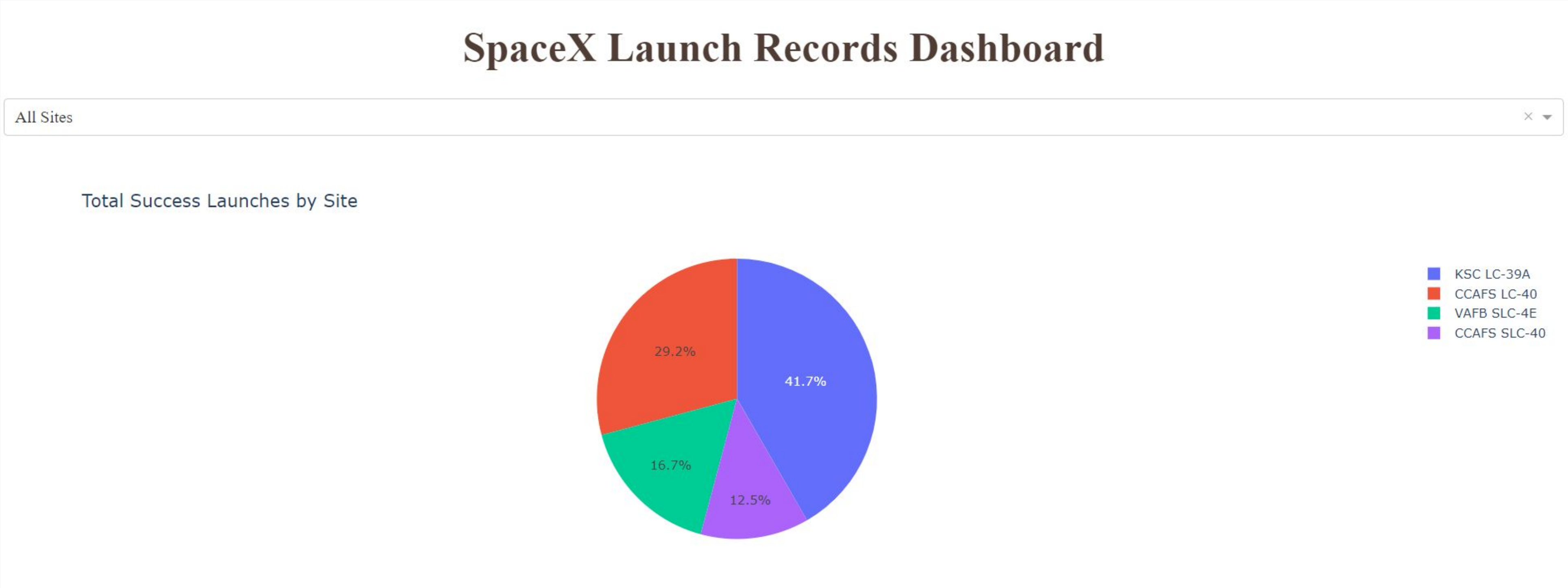
- The following figures represent the generated folium map showing the distance of a selected launch site to its proximities such as a railway (in the left) and a coastline (in the right).

Section 4

Build a Dashboard with Plotly Dash

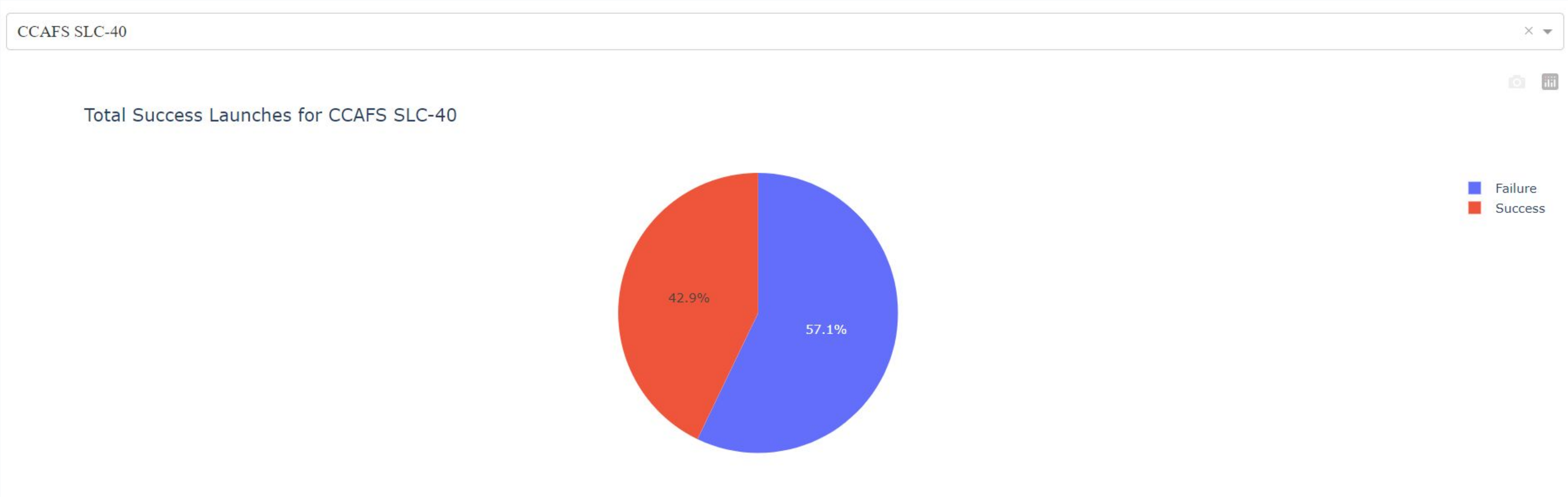
Pie chart of launch success count for all sites

- The following figure contains the pie chart representing the launch success count for all sites.



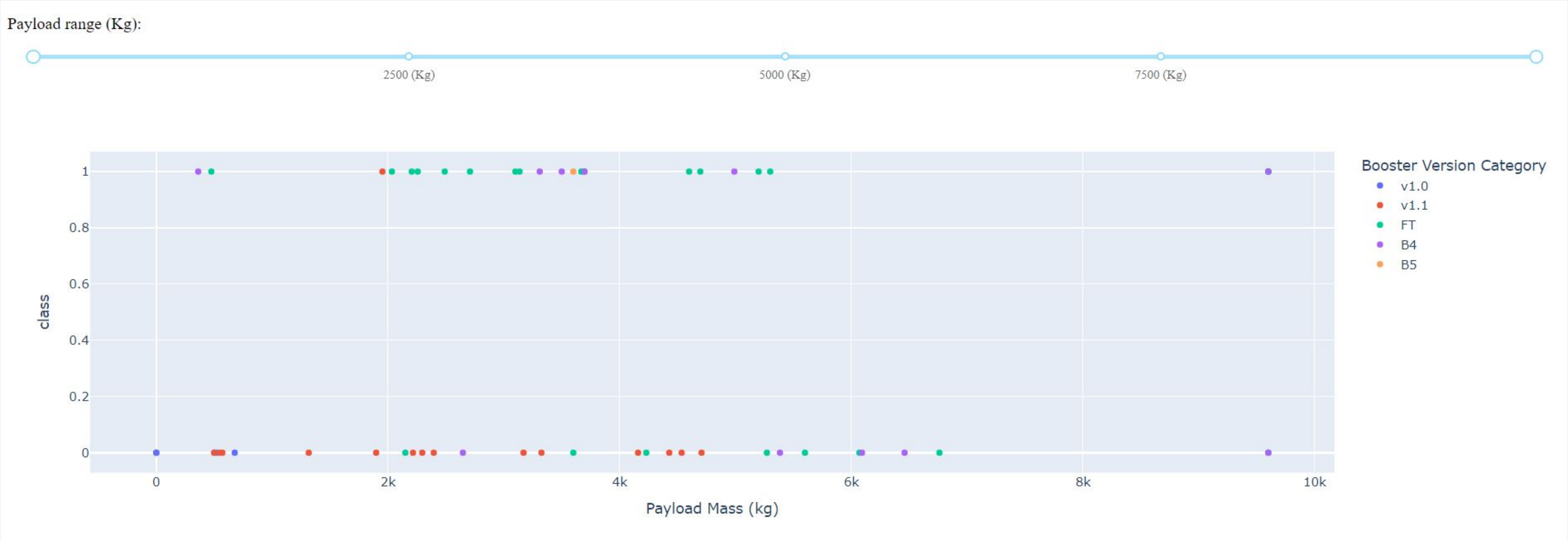
Pie chart of launch site with highest success ratio

- The following figure contains the pie chart corresponding for the launch site CCAFS SLC-40 with highest launch success ratio equal to 42.9%.



Scatter plot of Payload vs. Launch outcome

- The following figure contains the scatter plot representing the relationship of the payload mass with the launch outcome for all sites, with different payload intervals selected in the range slider, and also the booster version category.
- The payload range between 2500 kg and 5000 kg and the booster version FT correspond to the largest success rate.

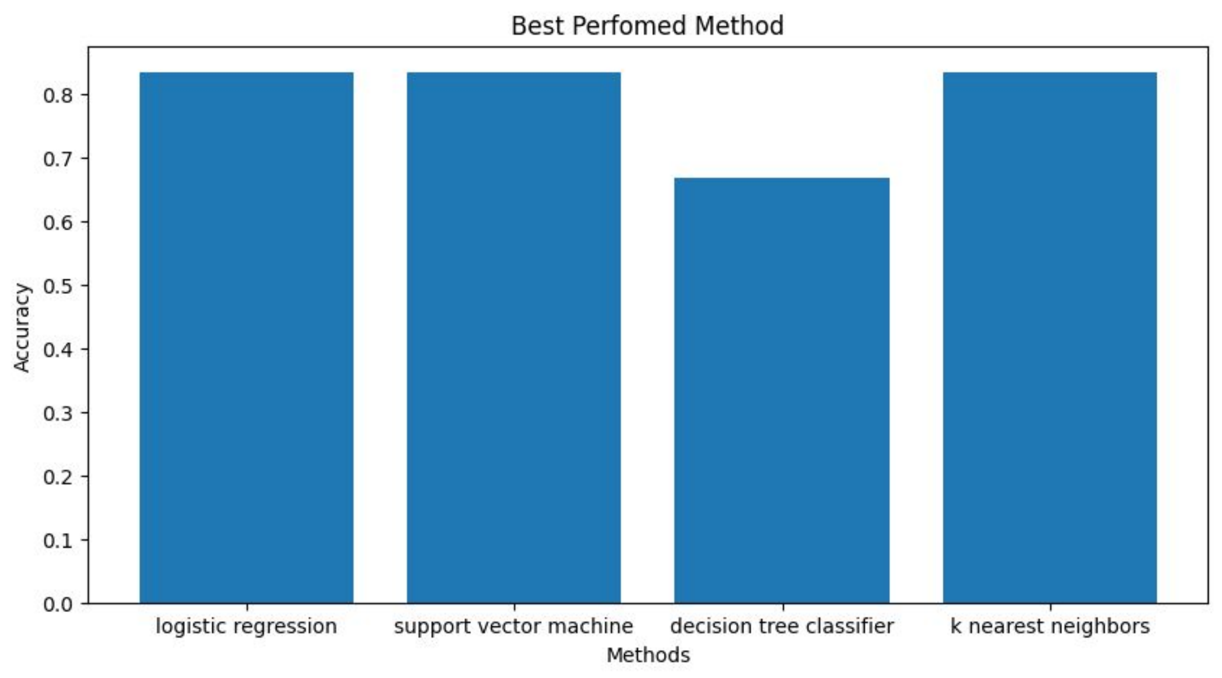


Section 5

Predictive Analysis (Classification)

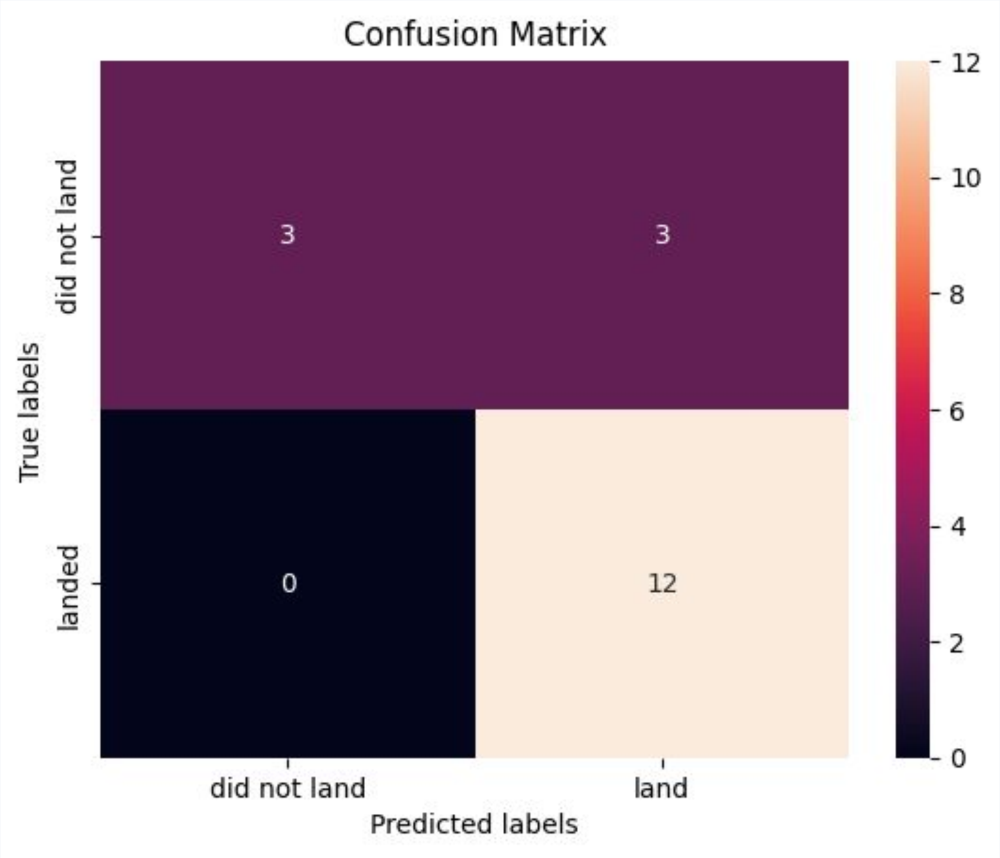
Classification Accuracy

- The following figure contains the bar chart representing the built model accuracy for all built classification models.
- The logistic regression, support vector machine and k nearest neighbors models have the highest classification accuracy equal to 0.83.



Confusion Matrix

- The following figure contains the confusion matrix of the result of the logistic regression, support vector machine and k nearest neighbors methods corresponding to the best performing models maximizing the precision (the ratio of the true positive and the total predicted positive) and the recall (the ratio of the true positive and the total real positive).



Conclusions

- The aim of this project was to apply the data science tools, using the collected and processed data corresponding to the Falcon 9 launching and landing operations performed by the SpaceX company, to predict the landing success cases.
- The exploratory data analysis (EDA) enabled to find data patterns and perform feature engineering.
- The application of different machine learning algorithms enabled to predict the landing operation success, giving an accuracy score equal to 0.83.

Appendix

- The following GitHub URL corresponds to the CSV file containing the used data to create the interactive dashboard for exploratory data analysis:
  https://github.com/yassine89/Applied-Data-Science-Capstone/blob/main/spacex_launch_dash.csv

Thank you!