IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tumedi Madihlaba
12 March 2024

Outline

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

Executive Summary

- In this capstone project, we use various machine learning classification algorithms to determine whether or not SpaceX Falcon 9 first stage will land successfully.
- The major steps in this project include the following:
  - Data collection, wrangling and formatting
  - Exploratory data analysis (EDA)
  - Interactive data visualization
  - Machine learning prediction
- Our tables and graphs illustrate that some features of rocket launches have correlation with the launch outcomes.
- Our results prove that decision tree may be the best machine learning algorithm to predict whether or not the Falcon 9 first stage will land successfully.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if we can determine if the first stage will land, then we can determine the cost of a launch.
- SpaceX Falcon 9 tends to launch numerous rockets, most landings are planned and occur in a controlled environment.
- This information is very useful when an alternative company wants to bit against SpaceX for a rocket launch.
- Given a set of features about SpaceX Falcon 9 rocket launch such as payload mass, orbit type, launch site, etc., will the first stage of the rocket land successfully?

Section A

Methodology

Methodology

Process Summary

- Data collection, wrangling and formatting using:
  - SpaceX API
  - Web Scraping
  - Wrangling
- Performing exploratory data analysis (EDA) using:
  - NumPy and Pandas
  - Matplotlib and Seaborn
  - Structured Query Language (SQL)

Methodology (cont...)

Process Summary

- Performing interactive visual analytics using:
  - Folium
  - Plotly Dash
- Performing predictive analysis using classification models:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K Nearest Neighbors (KNN)

Data Collection with SpaceX API

- The API used is https://api.spacexdata.com/v4/rockets/
- Collecting the Data with an API, especially the SpaceX REST API
- The API provides data about many types of rocket launches done by SpaceX, this data is therefore filtered to include only Falcon 9 launches
- Every missing value in the data is replaced by the mean of that particular column
- In the end, we remain with 90 rows or instances and 17 columns or features
- Here is the GitHub URL of the completed SpaceX API calls notebook testrepo/jupyter-labs-spacex-data-collection-api.ipynb at main · TumediMadihlaba/testrepo (github.com)

Data Collection with Web Scraping

- We perform web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records
- Parsing the data from those tables and convert them into a Pandas data frame for further visualization and analysis
- Here is the GitHub URL of our completed data wrangling related notebook testrepo/jupyter-labs-webscraping.ipynb at main · TumediMadihlaba/testrepo (github.com)

Data Wrangling

- We perform wrangling to find some patterns in the data and determine what would be the label for training supervised models
- Use the value counts method on the Launch Site column to determine the number of launches on each site
- Use the value counts method to determine the number and occurrences of each orbit in the Orbit column
- Create a list where the element is zero if the corresponding row in Outcome column is in the set Bad Outcome; otherwise, it is one
- Here is the GitHub URL of our completed data wrangling related notebook testrepo/labs-jupyter-spacex-Data wrangling.ipynb at main · TumediMadihlaba/testrepo (github.com)

EDA with SQL

- Performing some EDA using IBM DB2 database
- Establishing a connection between the SQL extension and database
- Determining what attributes are correlated with successful landings
- Using one hot encoding to convert the categorical values
- Preparing the data for a machine learning model that will predict if the first stage of Falcon 9 will successfully land
- Here is the GitHub URL of our completed EDA with SQL notebook [testrepo/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · TumediMadihlaba/testrepo (github.com)]

EDA with Visualization

- Using functions from NumPy and Pandas libraries to derive basic information about data collected, including:
  - The number of launches on each launch side
  - The number of occurrences of each orbit
  - The number and occurrence of each mission outcome
- Using functions form Matplotlib and Seaborn libraries to visualize the data through plots and charts, including:
  - The scatter plot relationship between flight number and launch site
  - The bar chart to visually check if there is any relationship between success rate and orbit type
  - Plotting a line chart to display the average launch success trend
- Here is the GitHub URL of our completed EDA with data visualization notebook testrepo/jupyter-labs-eda-dataviz.ipynb at main · TumediMadihlaba/testrepo (github.com)

Build an Interactive Map with Folium

- Functions from Folium library are used to visualize the data through interactive maps
- This library is used to:
  - Mark all the launch sides on the map
  - Mark the succeeded and failed launches for each site on the map
  - Mark the difference distances between the launch site to its proximities such as the nearest city, railway or highway
- Here is the GitHub URL of our completed interactive map with Folium map testrepo/lab_jupyter_launch_site_location.ipynb at main · TumediMadihlaba/testrepo (github.com)

Build a Dashboard with Plotly Dash

- Functions from Plotly Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and range slider
- Using a pie chart and scatter plot interactive sets show that:
  - The total success launches for each launch site
  - The correlation between payload mass and mission outcome
  - The total failure launches for each launch site
- Here is the GitHub URL of our completed Plotly Dash lab testrepo/spacex_dash_app.py at main · TumediMadihlaba/testrepo (github.com)

Predictive Analysis (Classification)

- Functions from Scikit-Learn library are used to create our machine learning models
- The machine learning phase is a process consisting of several steps:
  - Standardizing the data
  - Splitting the data into training and testing data sets
  - Creating machine learning models, including:
    - Logistic regression
    - Support Vector Machine (SVM)
    - Decision tree
    - K Nearest Neigbors (KNN)
  - Fitting the models on the training set
  - Finding the best combination of the hyperparameters for each model
  - Evaluating the models based on their accuracy score and confusion matrix
- Here is the GitHub URL of our completed predictive analysis lab testrepo/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · TumediMadihlaba/testrepo (github.com)

Results

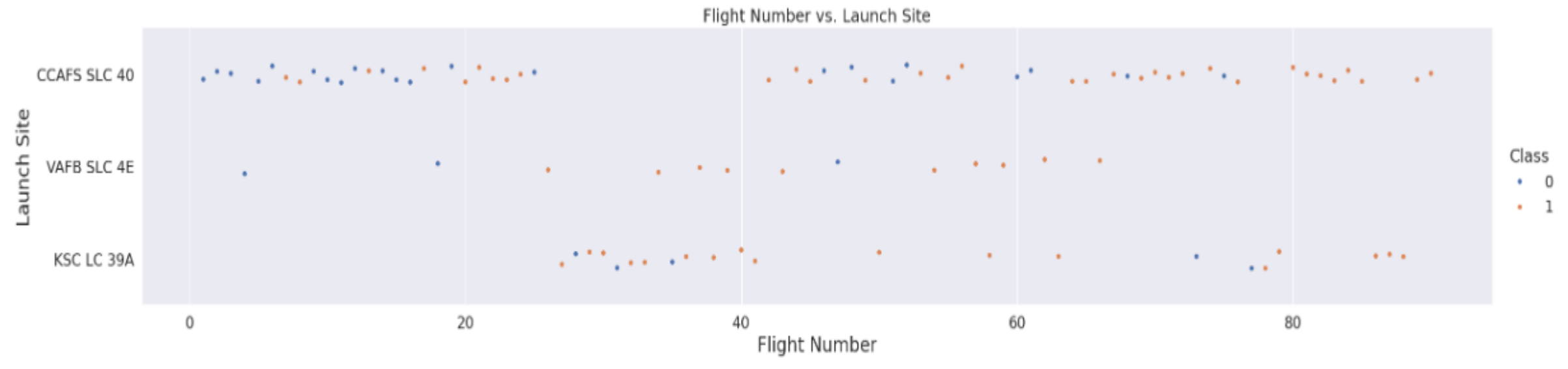
- Exploratory data analysis results
  - SQL
  - Visualization
- Interactive analytics demo in pictures
  - Folium
  - Plotly Dash
- Predictive analysis results
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K Nearest Neighbors (KNN)

Section B

Insights drawn from EDA

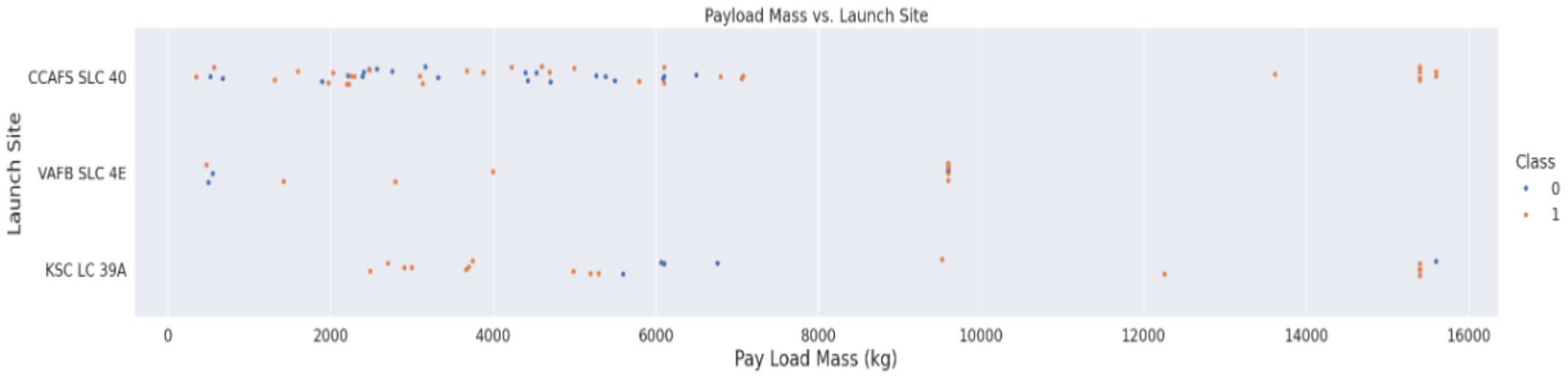
EDA with Visualization

- Showing a scatter plot of Flight Number vs. Launch Site



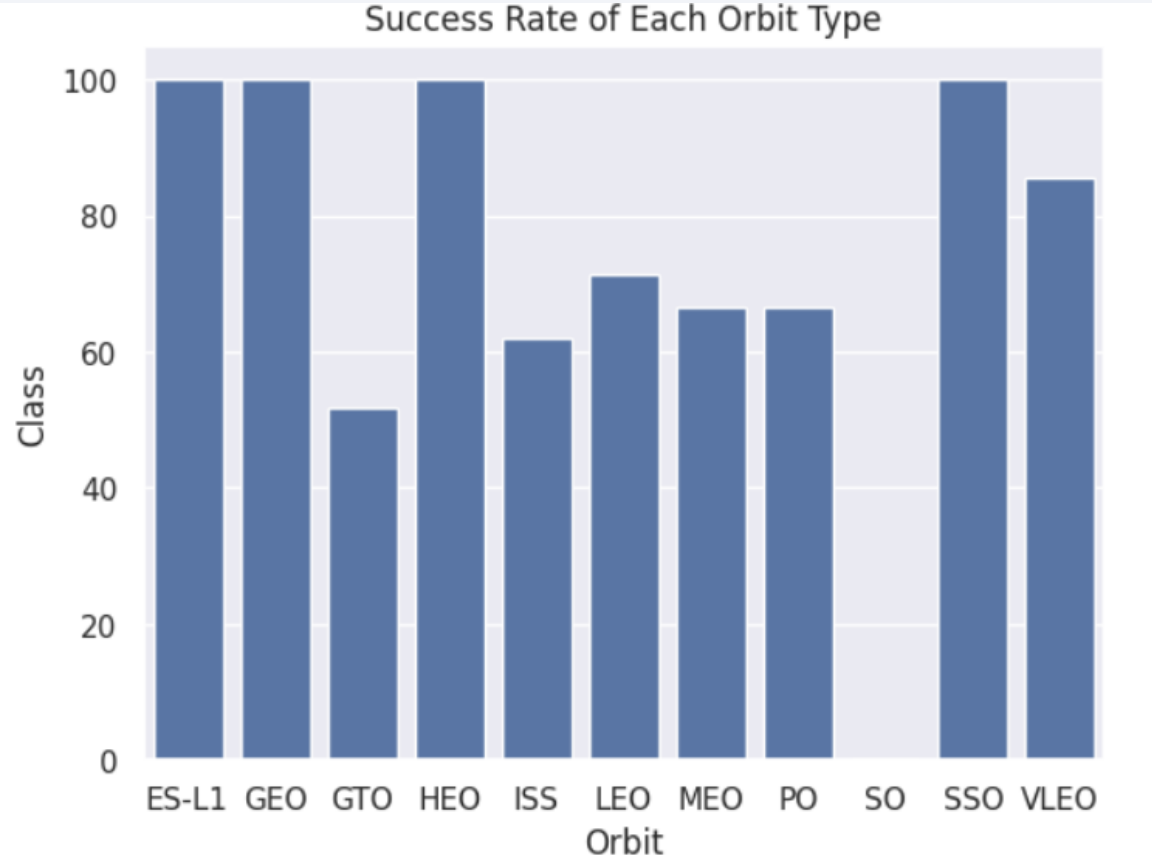
EDA with Visualization

- Showing a scatter plot of Payload vs. Launch Site



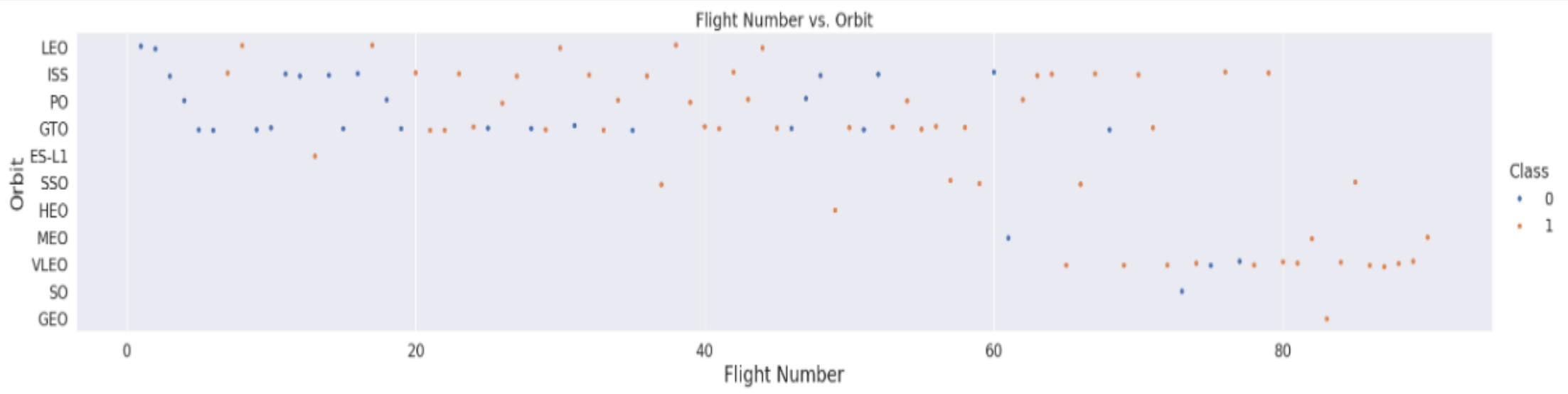
EDA with Visualization

- Showing a bar chart for the success rate of each Orbit Type



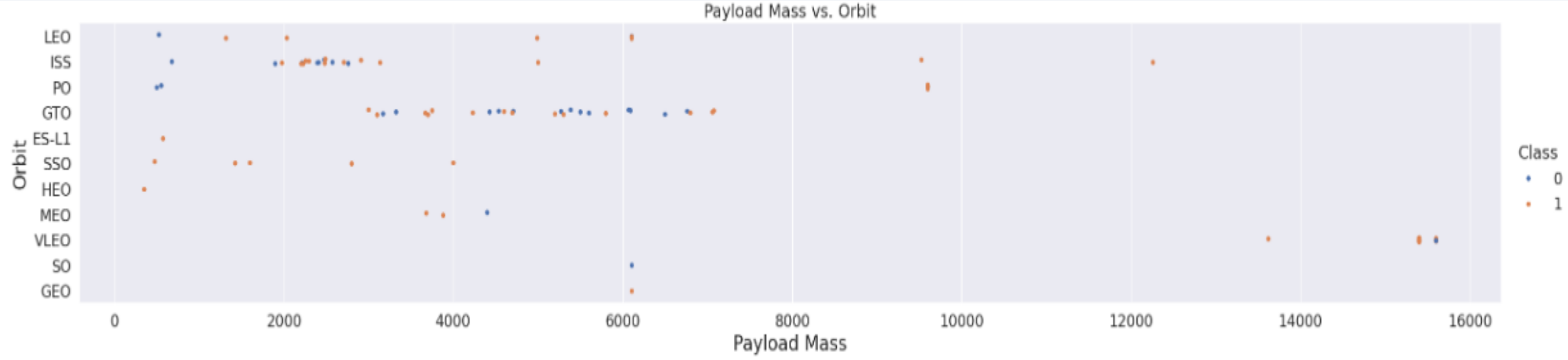
EDA with Visualization

- Showing a scatter point of Flight Number vs. Orbit type



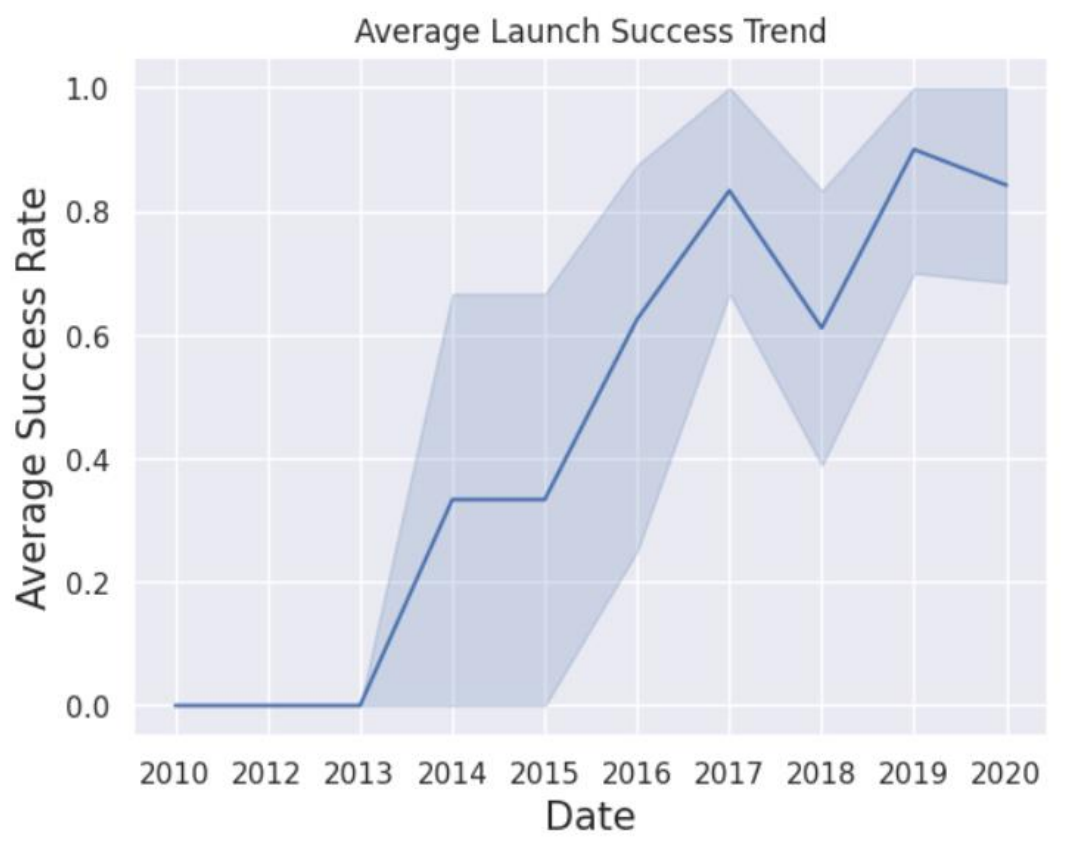
EDA with Visualization

- Showing a scatter point of Payload Mass vs. Orbit Type



EDA with Visualization

- Showing a line chart of yearly Average Success Rate



EDA with SQL

- Displaying the names of the unique launch sites in the space mission

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

- Listing the dates of the first successful landing outcome on ground pad

date
2010-06-04

EDA with SQL

- Displaying 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

EDA with SQL

- Displaying the total payload carried by boosters from NASA

sum
45596

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

average
2534.6666666666665

EDA with Visualization

- Listing 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

EDA with SQL

- Displaying 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

EDA with SQL

- Listing 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

EDA with SQL

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

MONTH	YEAR	Booster_Version	Landing_Outcome	Launch_Site
01	2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	2015	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

EDA with SQL

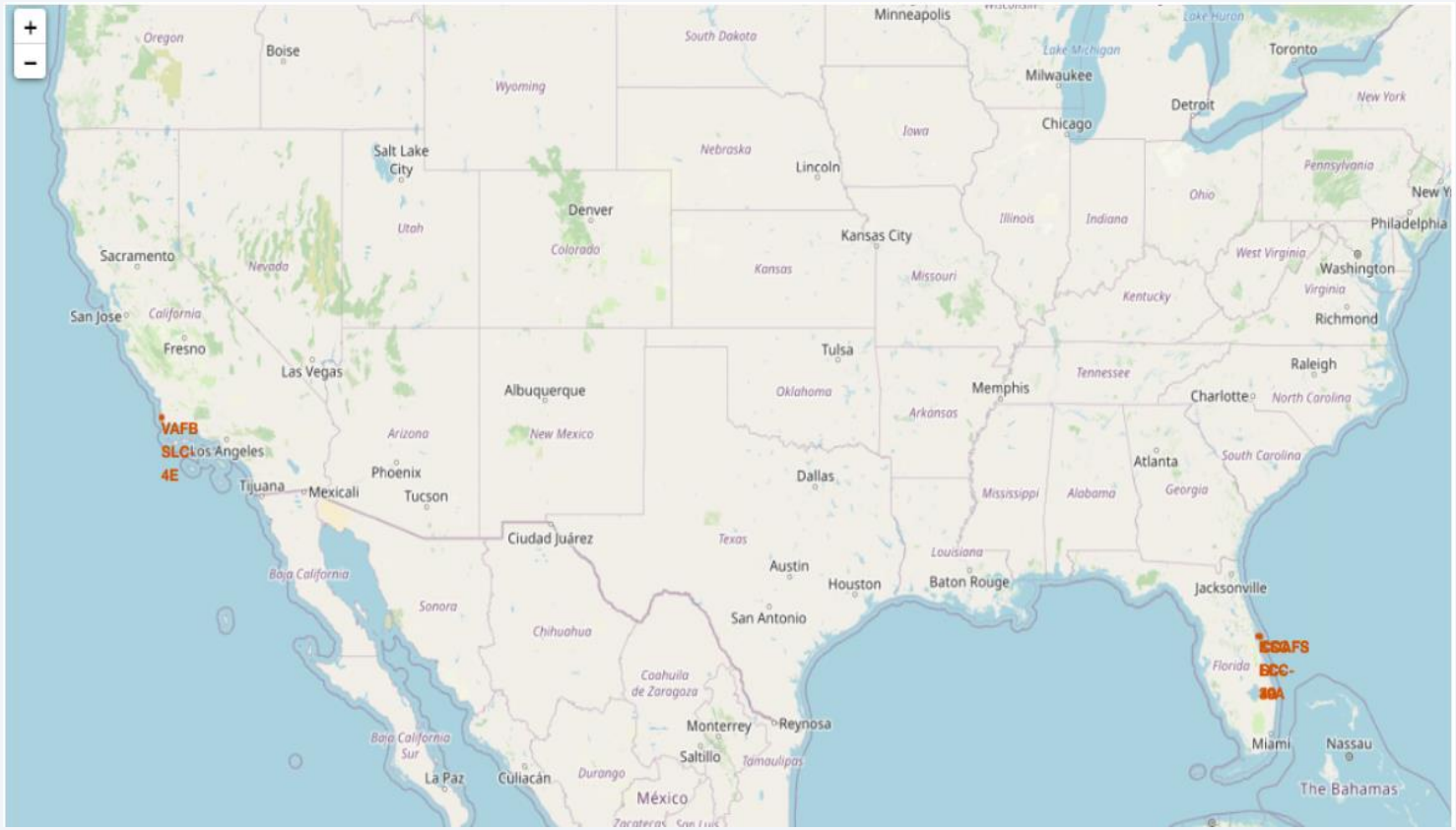
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing_Outcome	count_outcomes
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 C
Launch Sites
Proximities Analysis

Interactive Visual Analytics with Folium

- Displaying the generated Folium map and making a proper picture to include all launch sites' location markers on a global map

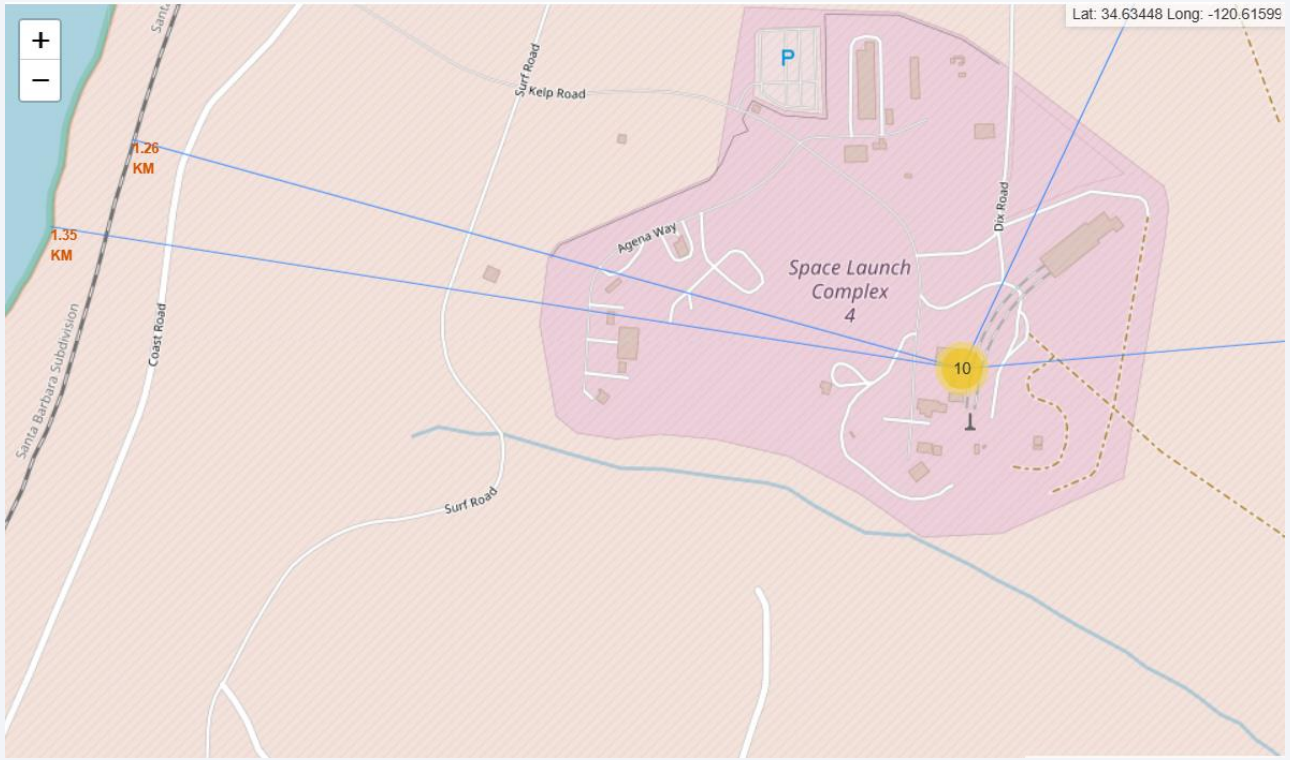


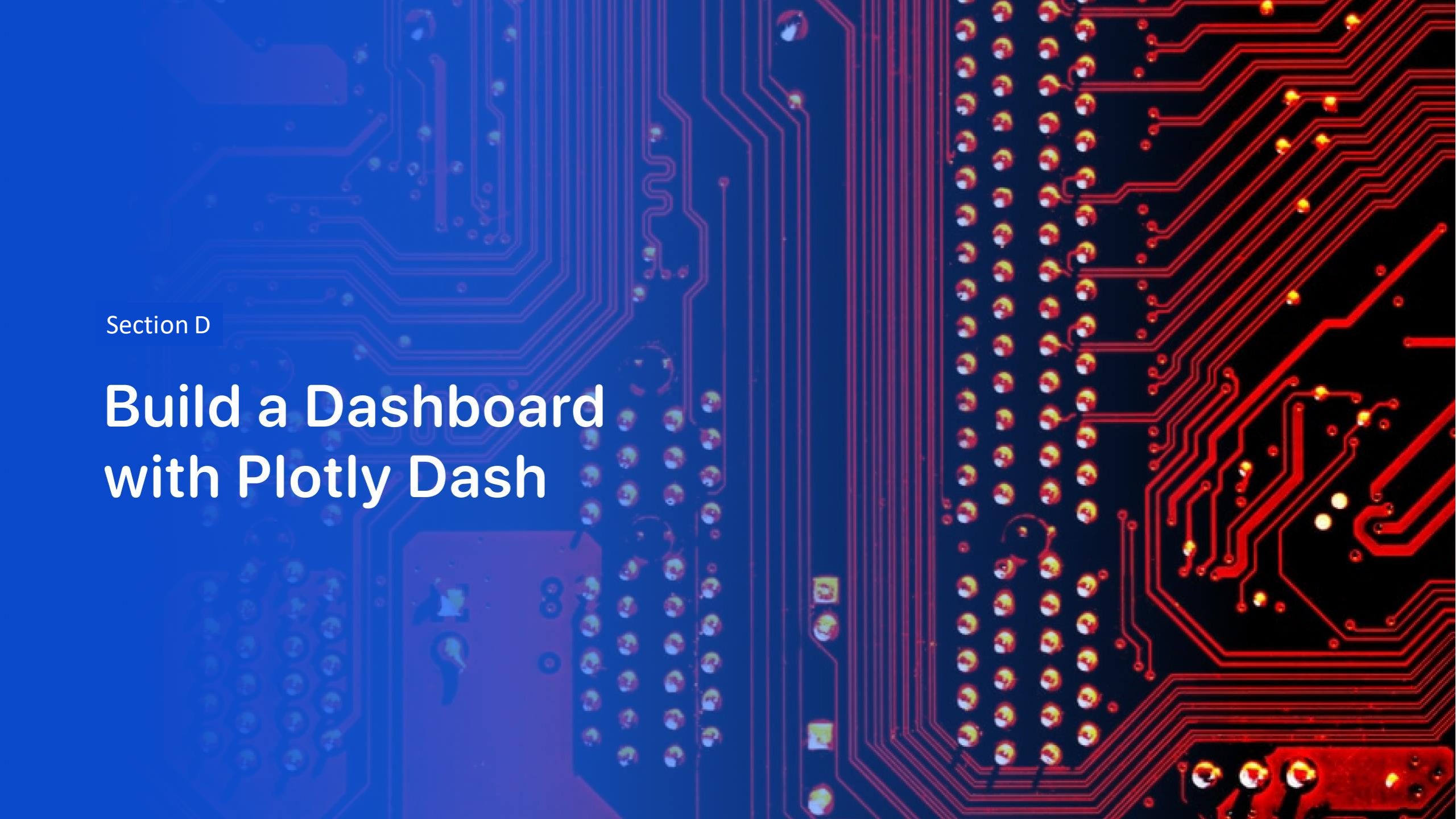
Interactive Visual Analytics with Folium

- Displaying the Folium map and making a proper picture to showcase the color-labeled launch outcomes on the map
- Each green tag represents a successful launch while each red tag represents a failed launch

Interactive Visual Analytics with Folium

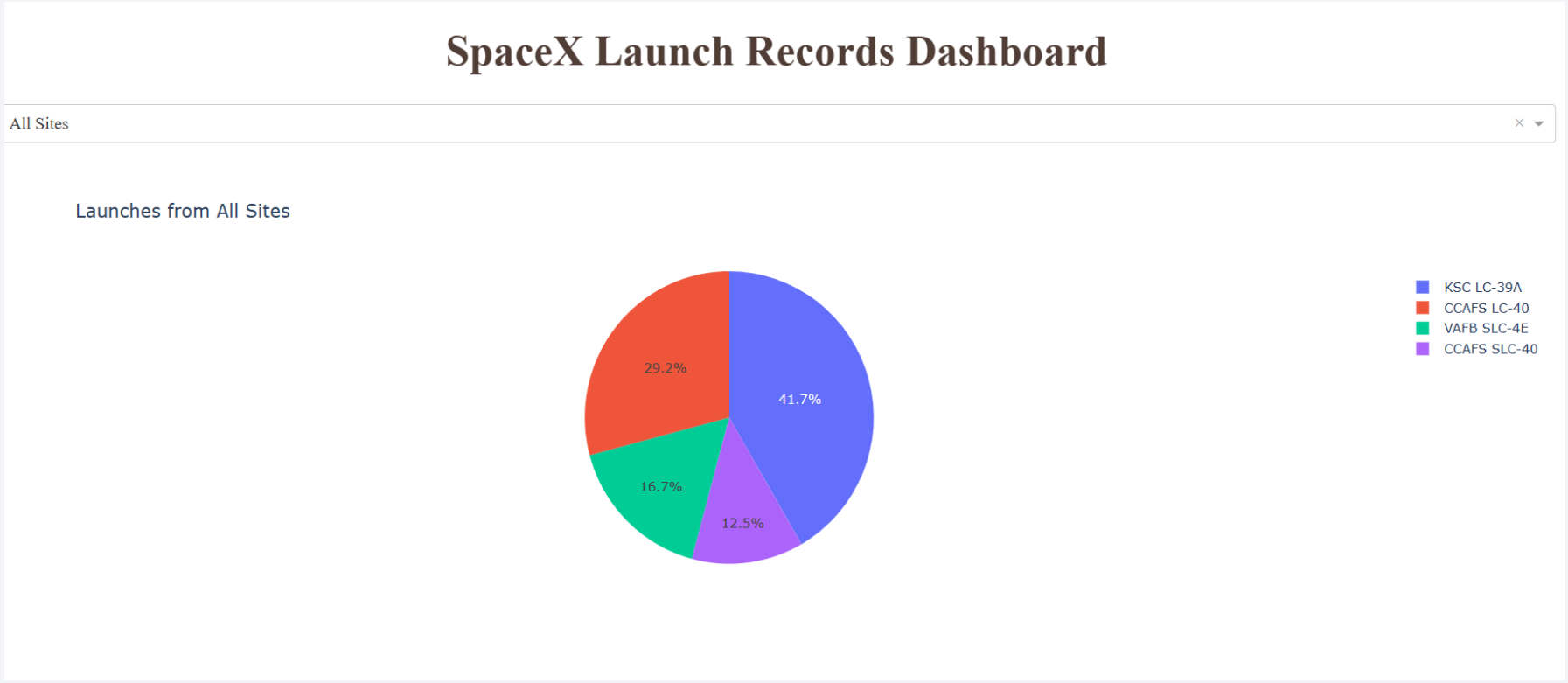
- Displaying the generated Folium map and demonstrating the distances from a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed




Section D
Build a Dashboard with Plotly Dash

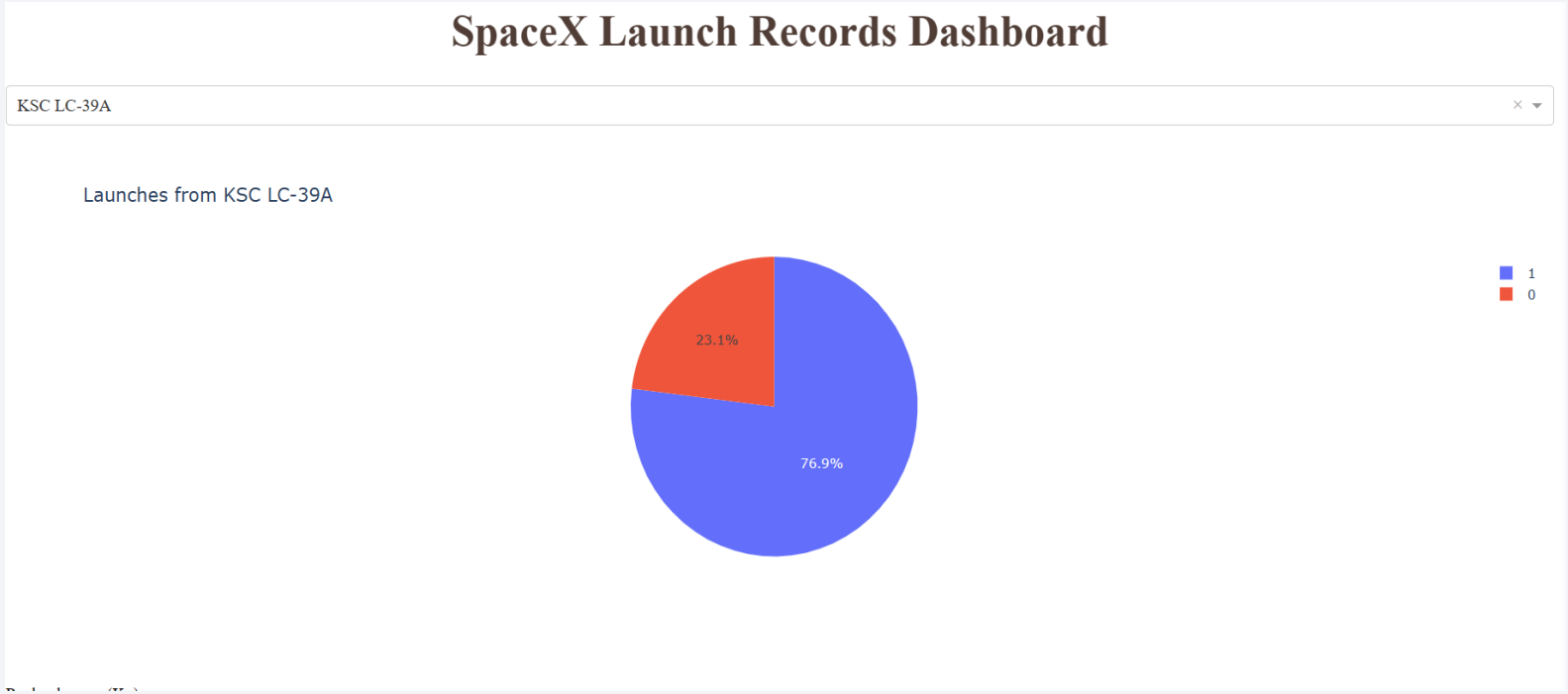
Interactive Dashboard with Plotly Dash

- Displaying the picture of launch success count for all sites, in a pie chart
- The CCAFS SLC-40 site has the lowest success count (12.5%) while KSC LC-39A side has the highest success count (41.7%)



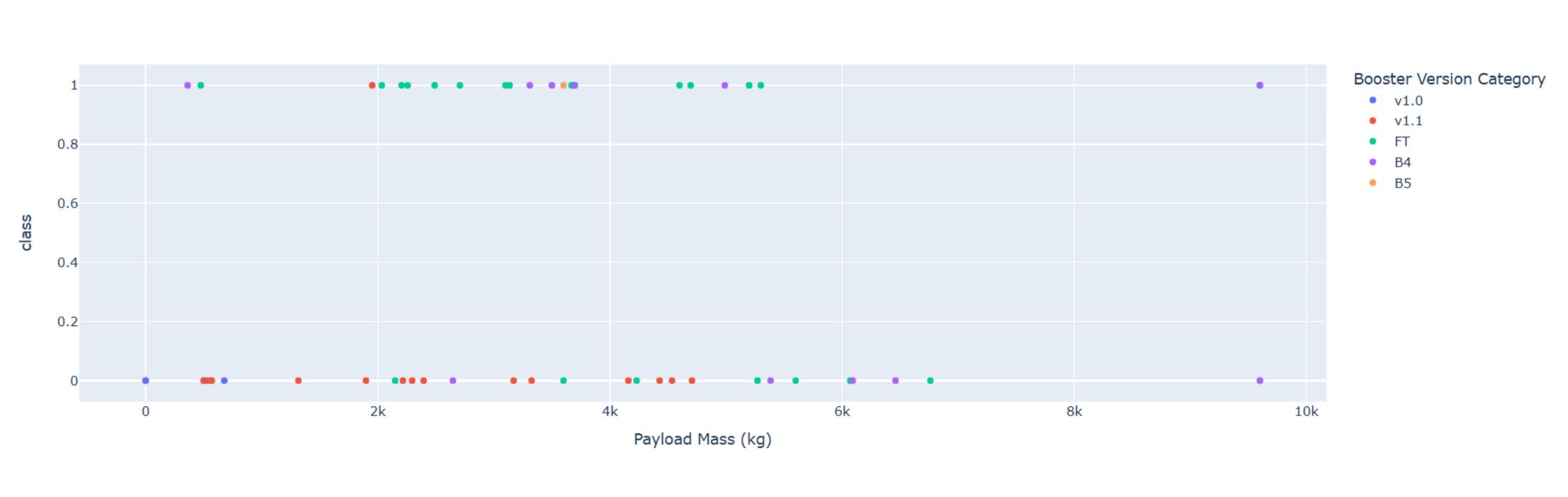
Interactive Dashboard with Plotly Dash

- Displaying the picture of the pie chart for the launch site with highest launch success ratio
- In the picture below, zero represents failed launches while one represents successful launches. We can clearly see that 76.9% of launches at KSC LC-39A have been successful



Interactive Dashboard with Plotly Dash

- Displaying of Payload Mass vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Payload range 2000-5750 and booster version FT have the largest success rate

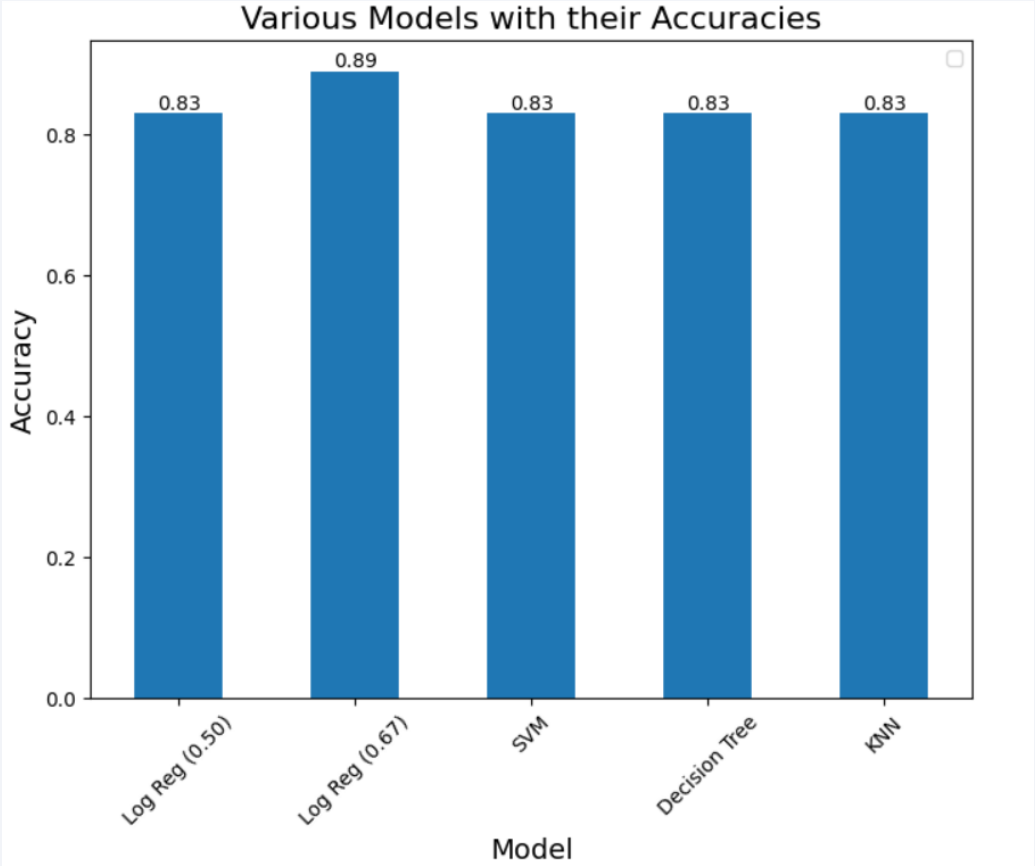


Section E

Predictive Analysis (Classification)

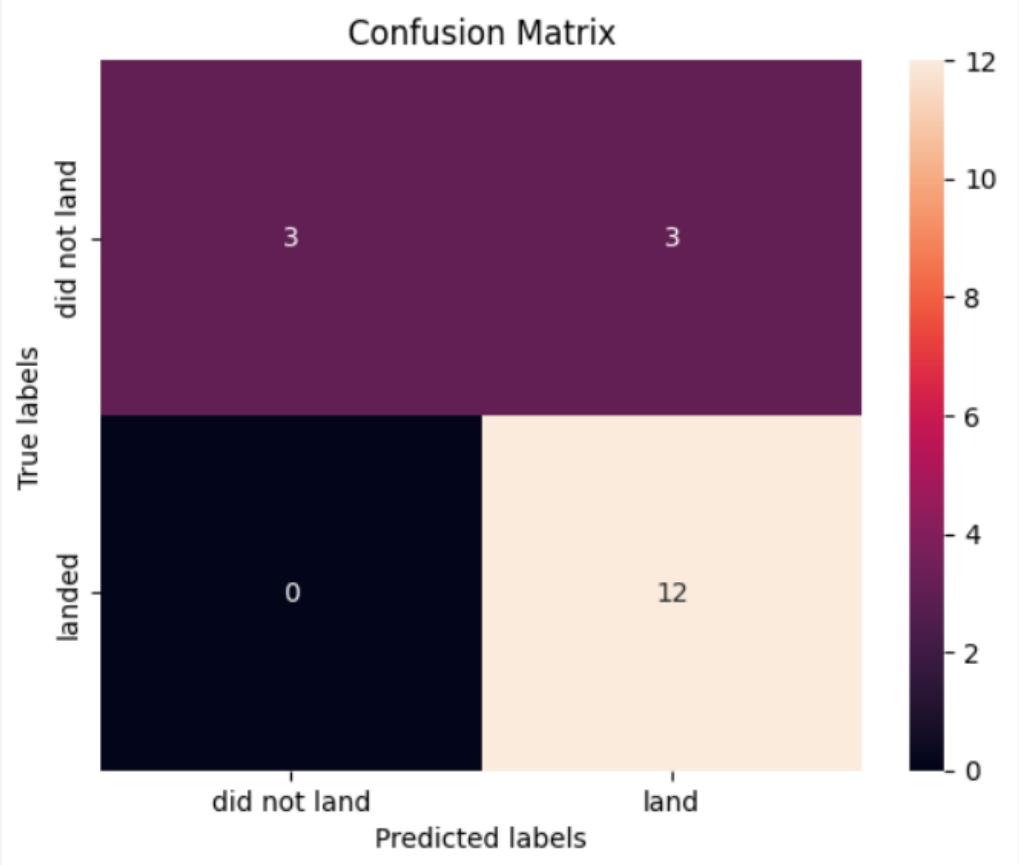
Classification Accuracy

- All predictive models employed have the same classification accuracy of 0.8333
- The best method to predict the outcome of the first stage landing is logistic regression with the optimized threshold of 0.67



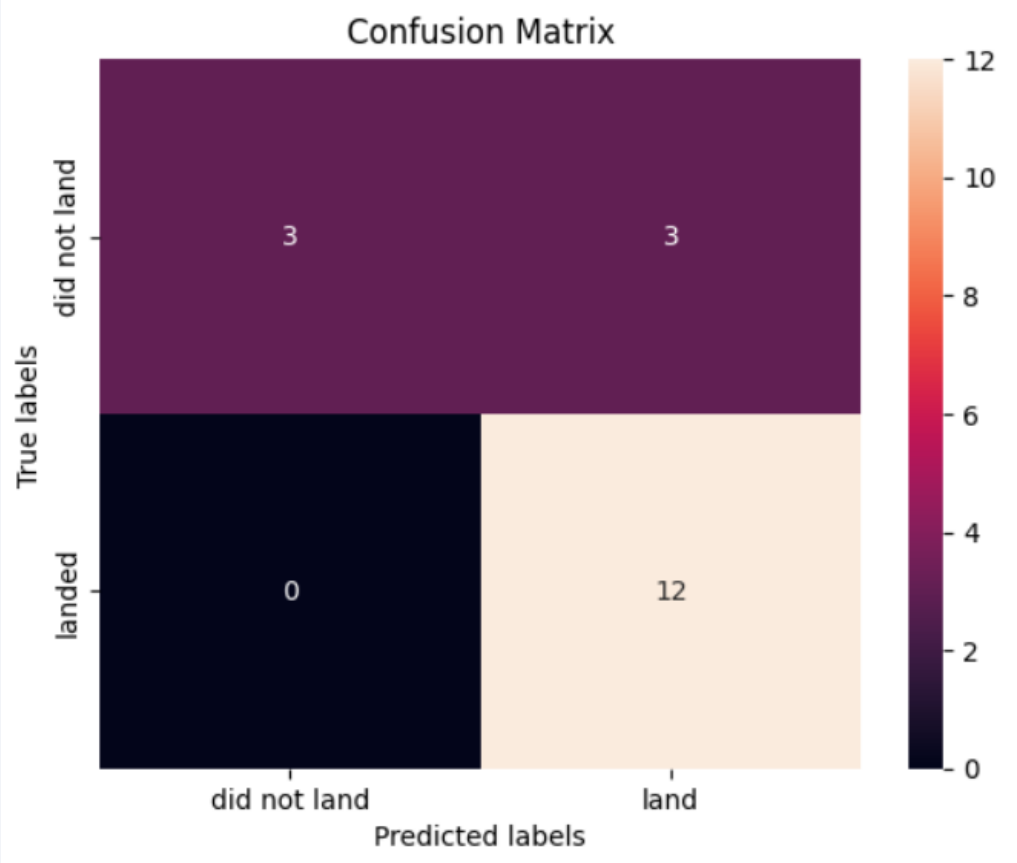
Predictive Analysis

- Logistic Regression Confusion Matrix
  - GridSearchCV best score: 0.8464
  - Accuracy Score: 0.8333



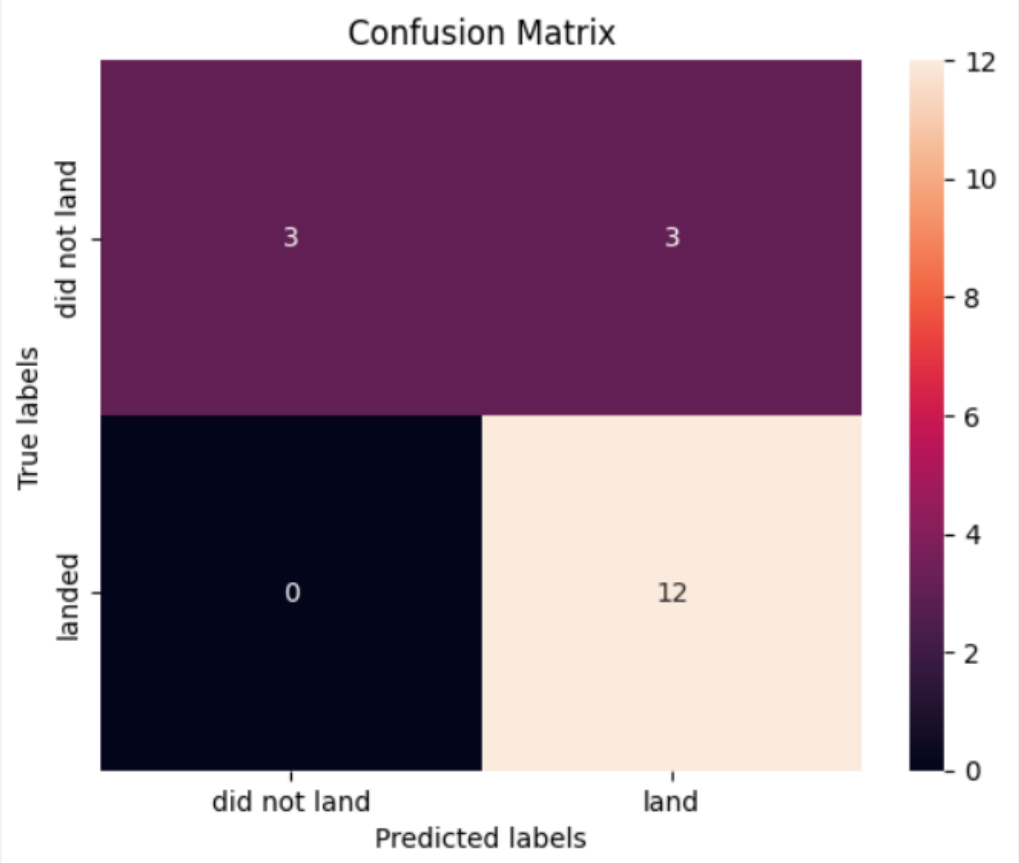
Predictive Analysis

- Support Vector Machine (SVM) Confusion Matrix
  - GridSearchCV best score: 0.8482
  - Accuracy Score: 0.8333



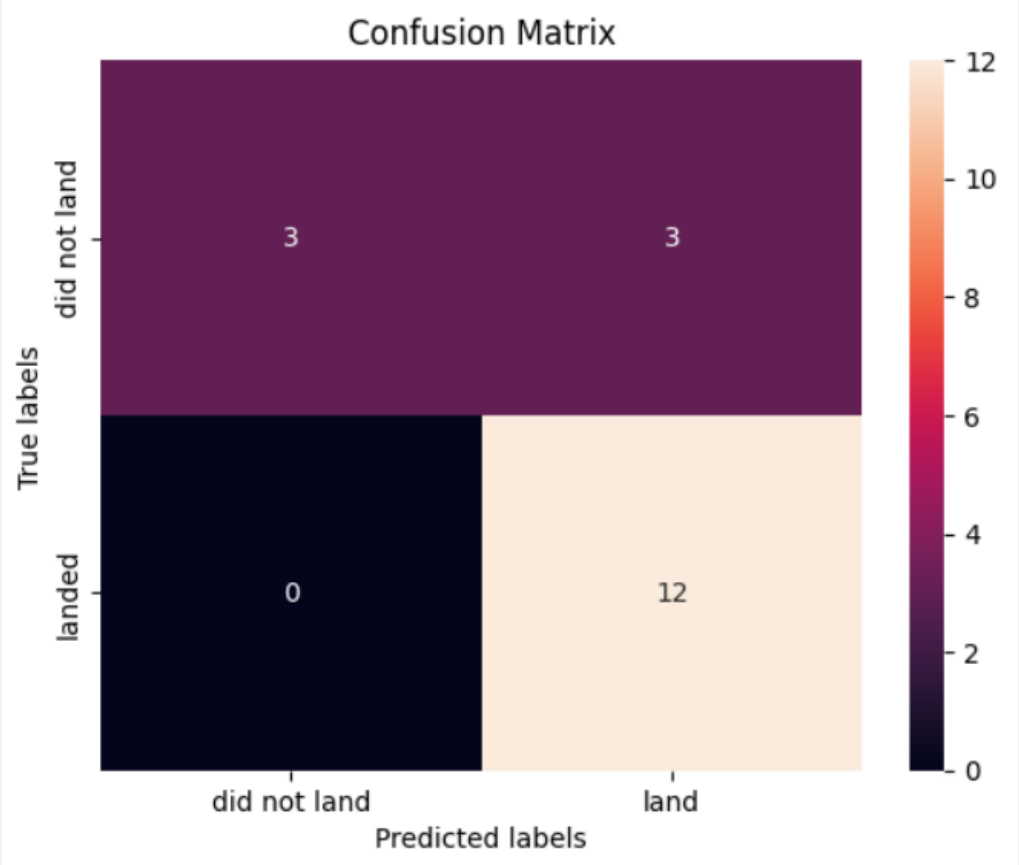
Predictive Analysis

- Decision Tree Confusion Matrix
  - GridSearchCV best score: 0.8768
  - Accuracy Score: 0.8333
- Decision Tree Model has the highest GridSearchCV best score of 0.8768



Predictive Analysis

- K Nearest Neighbors Confusion Matrix
  - GridSearchCV best score: 0.8482
  - Accuracy Score: 0.8333



Discussion

- As shown in data visualization section, some features may have correlation with the mission outcome in several ways.
- With heavier payload masses, the successful landing rates are more for orbit types Polar, LEO and ISS. However, for GTO orbit type, we cannot distinguish this well as both successful and unsuccessful rates are equally represented.
- It is very difficult to differentiate the exact way of how each of these features impact the mission outcome. Therefore, each feature may have a certain impact on the final mission outcome. However, we have a privilege of using machine learning algorithms to learn the patterns of the past data and predict whether a mission will be successful or not based on the given features.

Conclusion

- In this capstone project, we are tasked to predict whether or not the first stage of SpaceX Falcon 9 will land in order to determine the cost of the launch.
- Each feature of Falcon 9 launch, such as payload mass or orbit type, may affect the mission outcome a particular way.
- Various machine learning algorithms are employed to learn the patterns of the past Falcon 9 launch data to produce predictive models that are used to predict the desired outcome of Falcon 9 launch.
- All the given predictive models yield the same accuracy scores, implying similar performance by the employed machine learning algorithms.
- Based on GridSearchCV best scores, the predictive model produced by decision tree performed the best among other machine learning algorithms employed.

Appendix

- Include any relevant assets that have been created during this project, such as:
  - The picture showing the first few rows of the SpaceX API data
  - The picture showing the first few rows of the Web Scraping data
  - The picture showing the first few rows of the Data Wrangling data
  - The picture showing the scatter plot of Flight Number vs. Payload Mass

Data Collection – SpaceX API

- The picture below shows the first few rows of the SpaceX API data

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCou
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	

Data Collection – Web Scraping

- The picture below shows the first few rows of the Web Scraping data

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

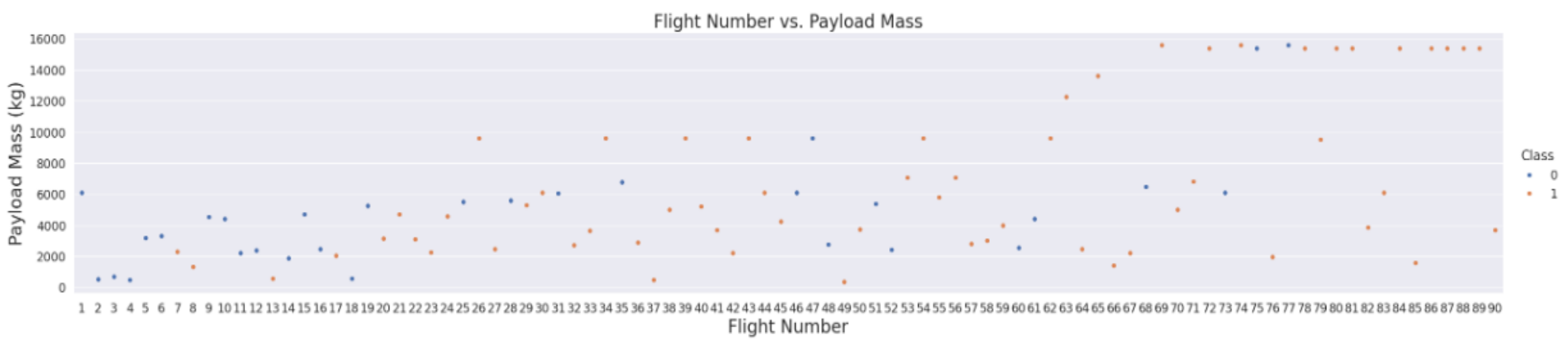
Data Collection – Data wrangling

- The picture below shows the first few rows of the Data Wrangling data

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCou
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	

EDA with Visualization

- Showing a scatter plot of Flight Number vs. Payload Mass



Thank you!