IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Jesús Arévalo López

24/10/2024

Outline

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

Executive Summary

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward 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, we can determine the cost of a launch.

Introduction

- The process followed is the data science methodology, which involves data collection, data management, exploratory data analysis, data visualization, model development, model evaluation, and communication of their results to stakeholders.

- What methods are best for collecting data relevant to predicting rocket landings?
- How to effectively clean and preprocess data to ensure accurate analysis?
- Which machine learning models are most effective?

Section 1

Methodology

Methodology

Executive Summary

- **Introduction**

  The objective is to predict whether SpaceX will attempt to land a rocket or not.

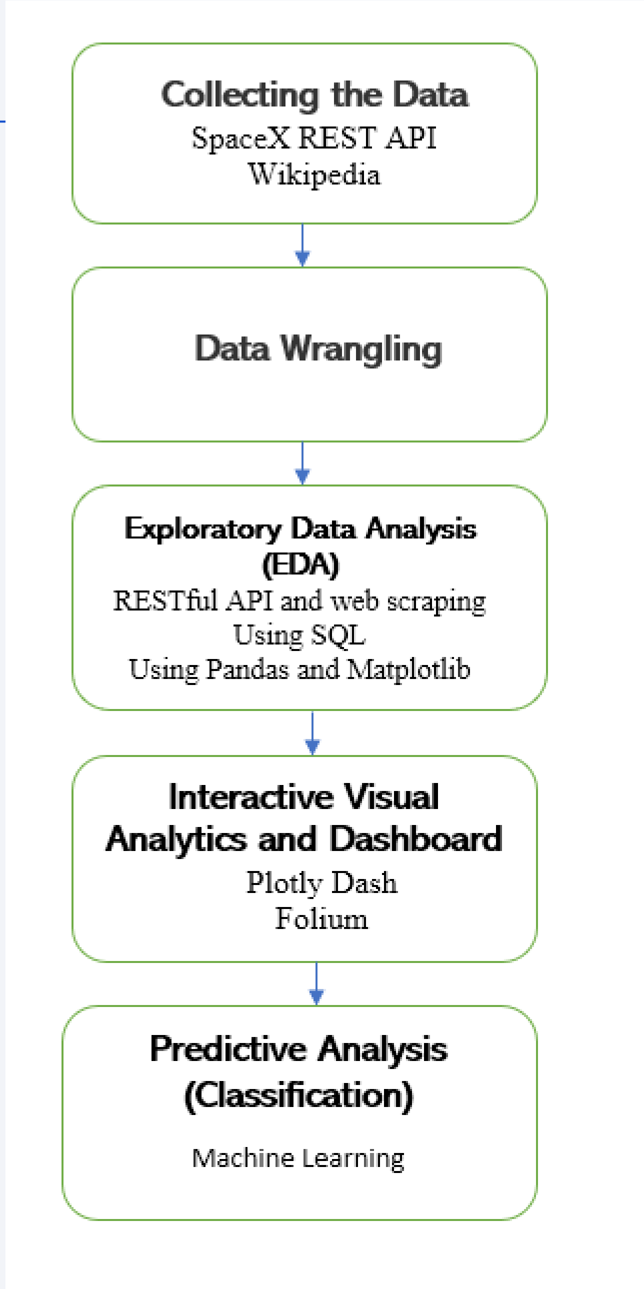
  - **Collecting the Data**
    - Data collection using SpaceX REST API to obtain data about rocket launches.
    - Using the Python library called "requests" to make a GET request to the API and obtain the launch data. This data will be in JSON format, which will then be converted into a dataframe using the "json_normalize" function.
    - Using web scraping using BeautifulSoup library to obtain data from HTML tables related to Falcon 9 launches.
  - **Data Wrangling**
    - Data manipulation: Flight Number, Date, Booster version, Payload mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing pad, Block, Reused count, Serial, and Longitude and latitude of launch.
- **Exploratory Data Analysis (EDA)**

  Collect data on Falcon 9 first stage landings. Use a RESTful API and web scraping. Convert the data into a data frame and then perform data manipulation.

  - **Exploratory Analysis Using SQL**
    - Create scatter plots and bar charts with Python to analyze data in a Pandas data frame.
    - Perform exploratory data analysis using Python by manipulating data in a Pandas data frame.
    - Create and execute SQL queries to select and sort data.
  - **Exploratory Analysis Using Pandas and Matplotlib**
    - Visualize data and extract meaningful patterns to guide the modeling process.
- **Interactive Visual Analytics and Dashboard**
  - Dashboard to analyze launch records interactively with Plotly Dash.
  - Interactive map to analyze the launch site proximity with Folium.
  - Interactive dashboard containing pie charts and scatter plots to analyze data with the Plotly Dash Python library.
  - Distances on an interactive map by writing Python code with the Folium library.
  - Interactive maps, plot coordinates, and mark clusters by writing Python code with the Folium library.
  - Dashboard to interactively analyze launch logs with Plotly Dash.
  - Interactive map to analyze launch site proximity with Folium.
- **Predictive Analysis (Classification)**
  - Use Machine Learning to determine if the Falcon 9 first stage will land successfully. Split data into training data and test data to find the best hyperparameter for SVM, classification trees, and logistic regression. Find the method that works best using the test data.
  - Split data into training and test data.
  - Train different classification models.
  - Optimize hyperparameter grid search.
  - Create a predictive model that helps a business run more efficiently.

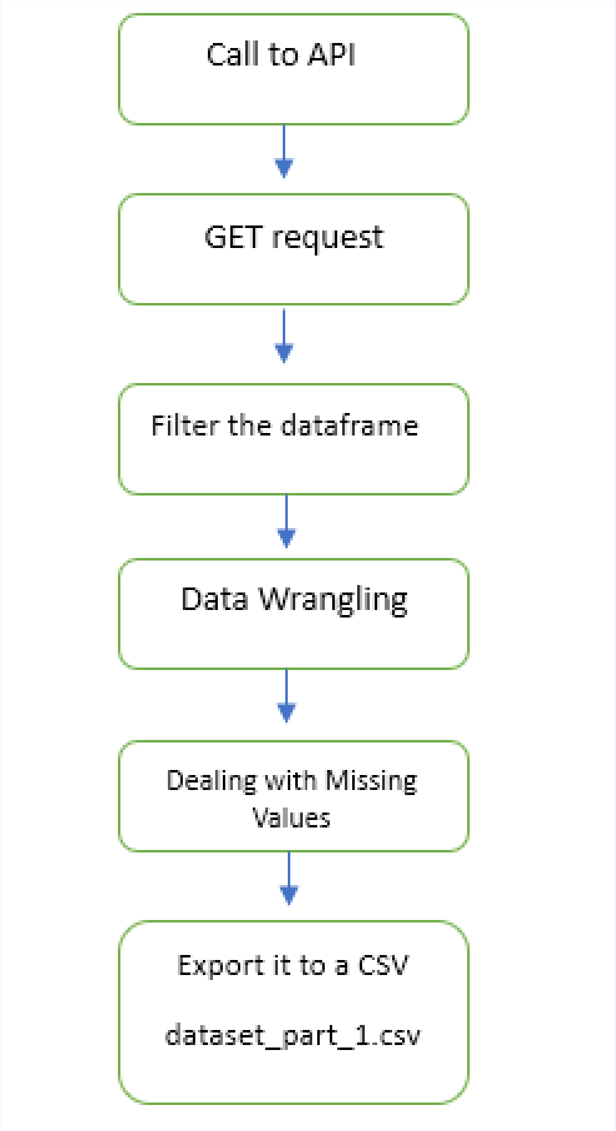
Data Collection

- SpaceX launch data collected through an API, specifically the SpaceX REST API.
- [Wikipedia]



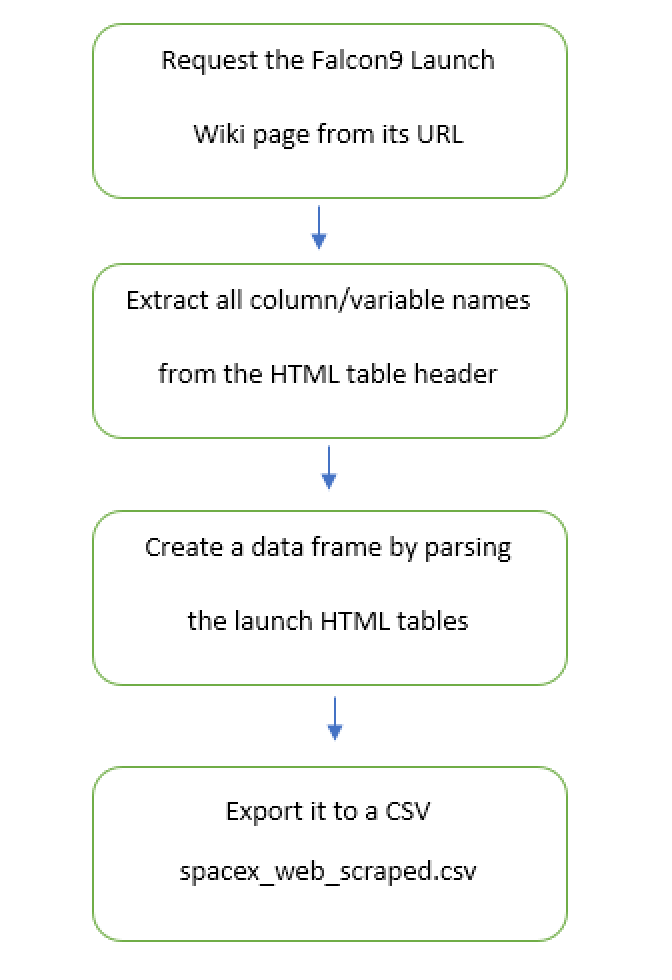
Data Collection – SpaceX API

- GitHub URL: [link here]



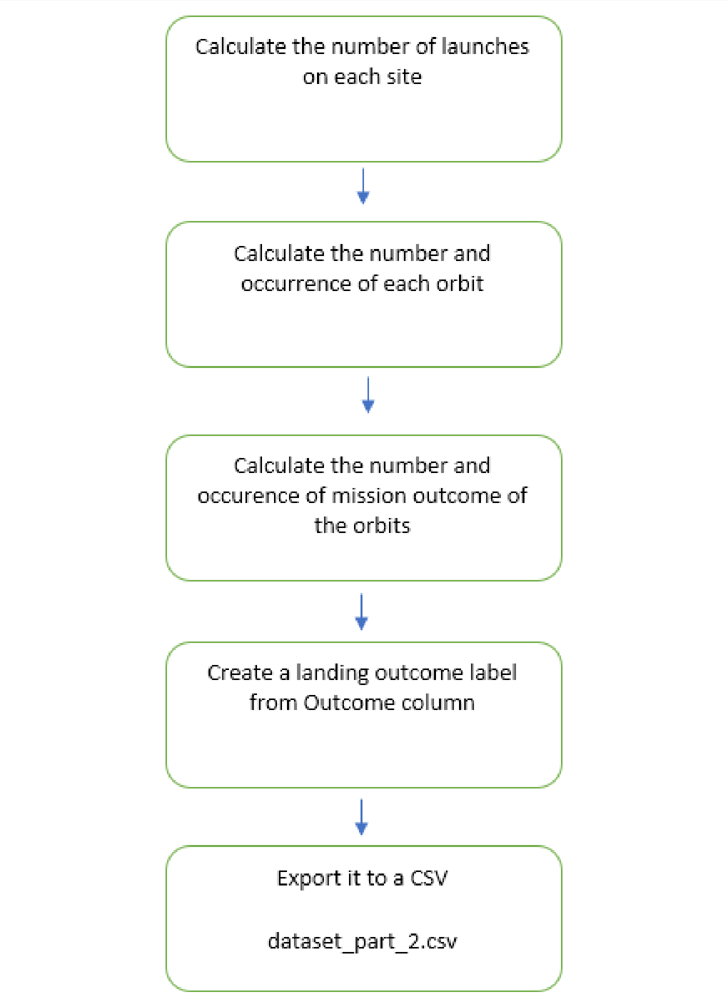
Data Collection - Scraping

- GitHub URL: link here



Data Wrangling

- GitHub URL: link here



EDA with Data Visualization

- Scatter plot: Visualize the relationship between Flight Number and Launch Site
- Scatter plot: Visualize the relationship between Payload Mass and Launch Site
- Bar chart: Visualize the relationship between success rate of each Orbit type
- Scatter plot: Visualize the relationship between FlightNumber and Orbit type
- Scatter plot: Visualize the relationship between Payload Mass and Orbit type
- Line chart: Visualize the launch success yearly trend

- GitHub URL: link here

EDA with SQL

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first succesful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. 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.
10. 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.

- GitHub URL: link here

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Markers for all launch sites and for Nasa Johnson Space Center.
  - Circles for each launch site and blue circle at NASA Johnson Space Center.
  - Mark successful/failed launches for each site:
    - If class = 1, the value of marker_color will be green.
    - If class = 0, the value of marker_color will be red.
  - Lines to calculate distances between a launch site and its surroundings, whether coast, city, railway, highway, etc.
- GitHub URL: link here

Build a Dashboard with Plotly Dash

Dataset obtained from spacex_launch_dash.csv.

There are four launch sites and we want to see which one is more successful.

We want to select a specific site and check its detailed success rate (class = 0 vs class = 1). To do this, we use a dropdown menu that allows us to select different launch sites. A pie chart, a completed payload range slider, and a scatter plot will be displayed.

By selecting all sites, the pie chart will show the percentage of launches that have been successful for each site.

If we select a site, we will be shown the percentage of successful and failed launches for that site.

We want to find if the variable payload is correlated with the mission outcome. To do this, we will use the slider to filter the payload range (Kg) for the scatter plot

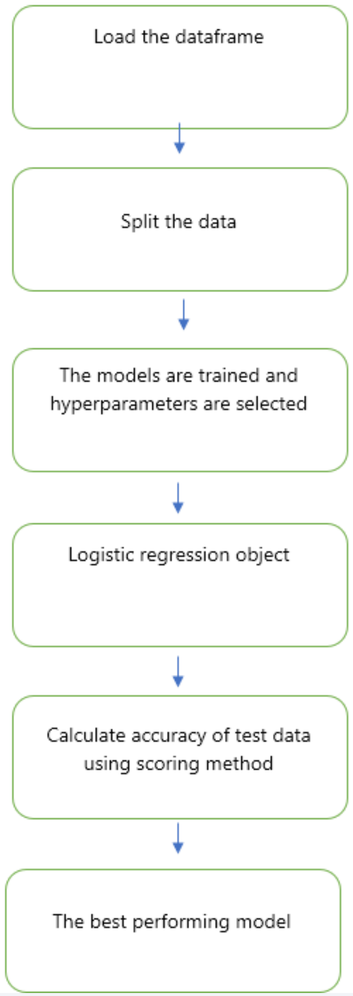
Below, the scatter plot visually shows how the payload can be correlated with the mission outcomes for the selected sites. We also color-labeled the Booster version at each spread point to see the mission results with different boosters.

- GitHub URL: link here

Predictive Analysis (Classification)

Load the dataframe
Create a NumPy array by applying the method to_numpy()
Standardize the data
We split the data into training and testing data.
The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function GridSearchCV.
Use the function train_test_split to split the data X and Y into training and test data.
Create a logistic regression object then create a GridSearchCV object logreg_cv with cv = 10.
We output the GridSearchCV object for logistic regression.
Calculate the accuracy on the test data using the method score.
Examining the confusion matrix, we see that logistic regression can distinguish between the different classes.
We see that the problem is false positives.
Create a support vector machine object then create a GridSearchCV object svm_cv with cv = 10.
Calculate the accuracy on the test data using the method score.
Create a decision tree classifier object then create a GridSearchCV object tree_cv with cv = 10.
Calculate the accuracy of tree_cv on the test data using the method score.
Create a k nearest neighbors object then create a GridSearchCV object knn_cv with cv = 10.
Calculate the accuracy of knn_cv on the test data using the method score.
Find the method performs best.

- GitHub URL: [link here]()



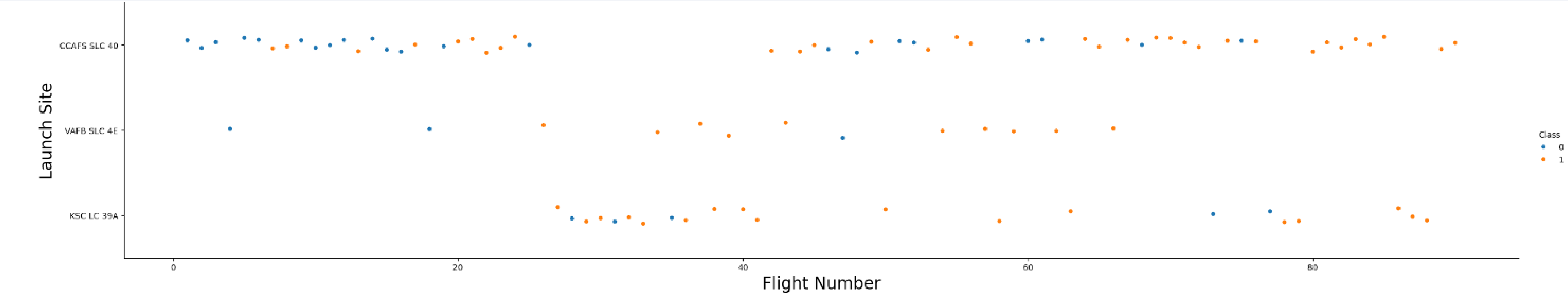
Results

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

Section 2

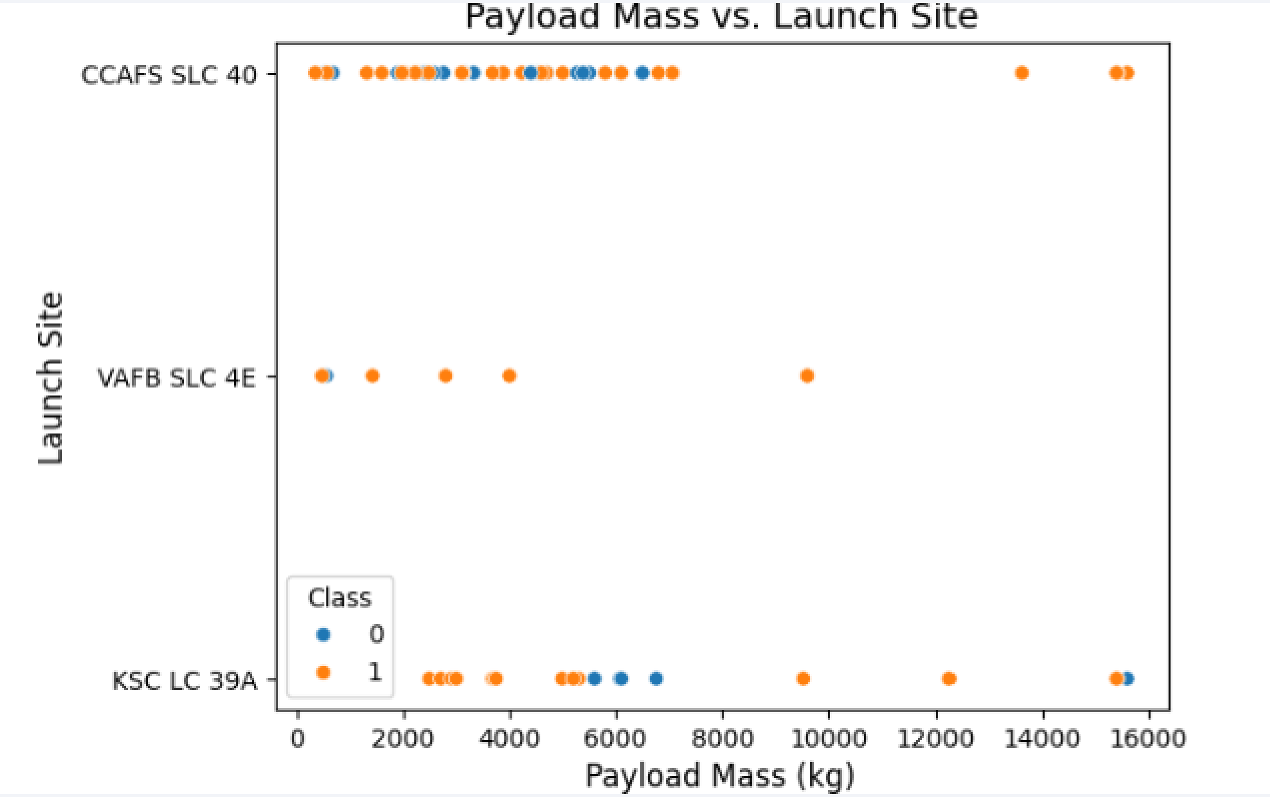
Insights drawn from EDA

Flight Number vs. Launch Site



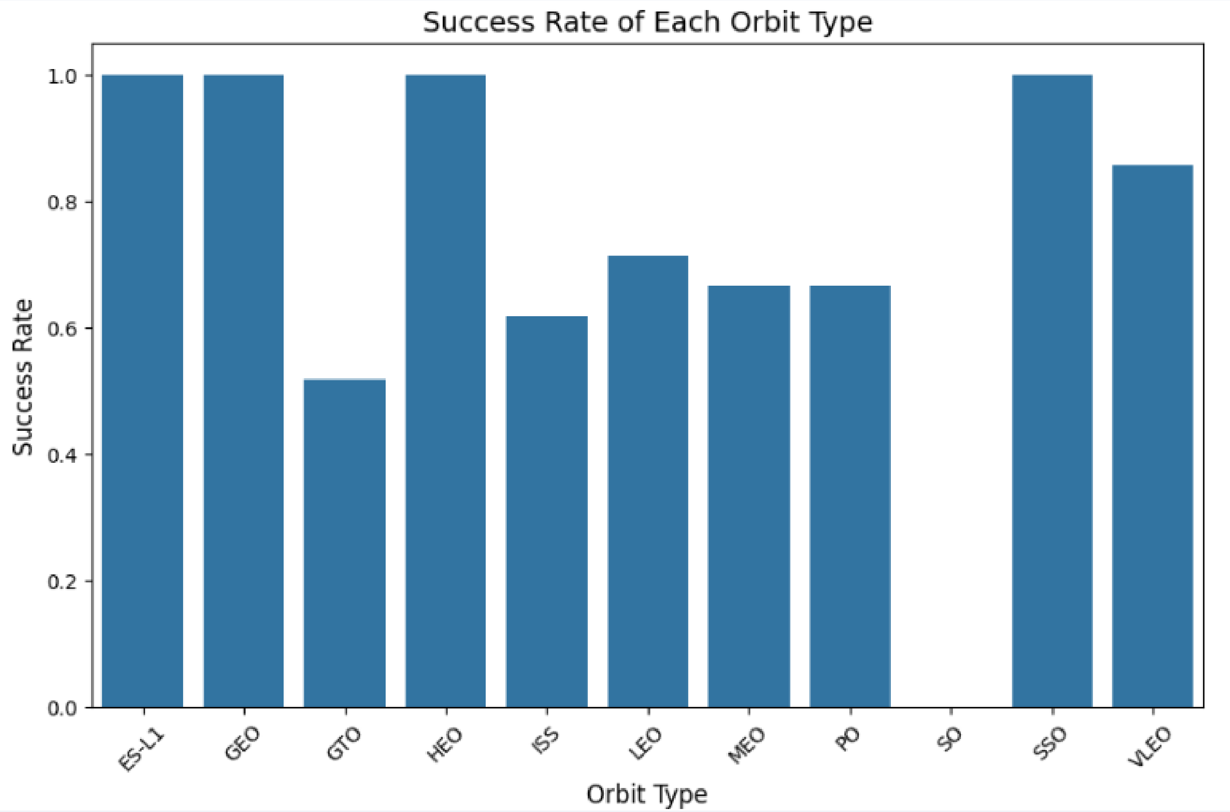
- FlightNumber on the x-axis.
- Launch Site on the y-axis.
- Color indicating success (Class 1) or failure (Class 0).
- Highest success rate: CCSFS SLC 40.
- As the number of flights increases, the landing rate is more successful.

Payload vs. Launch Site



- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

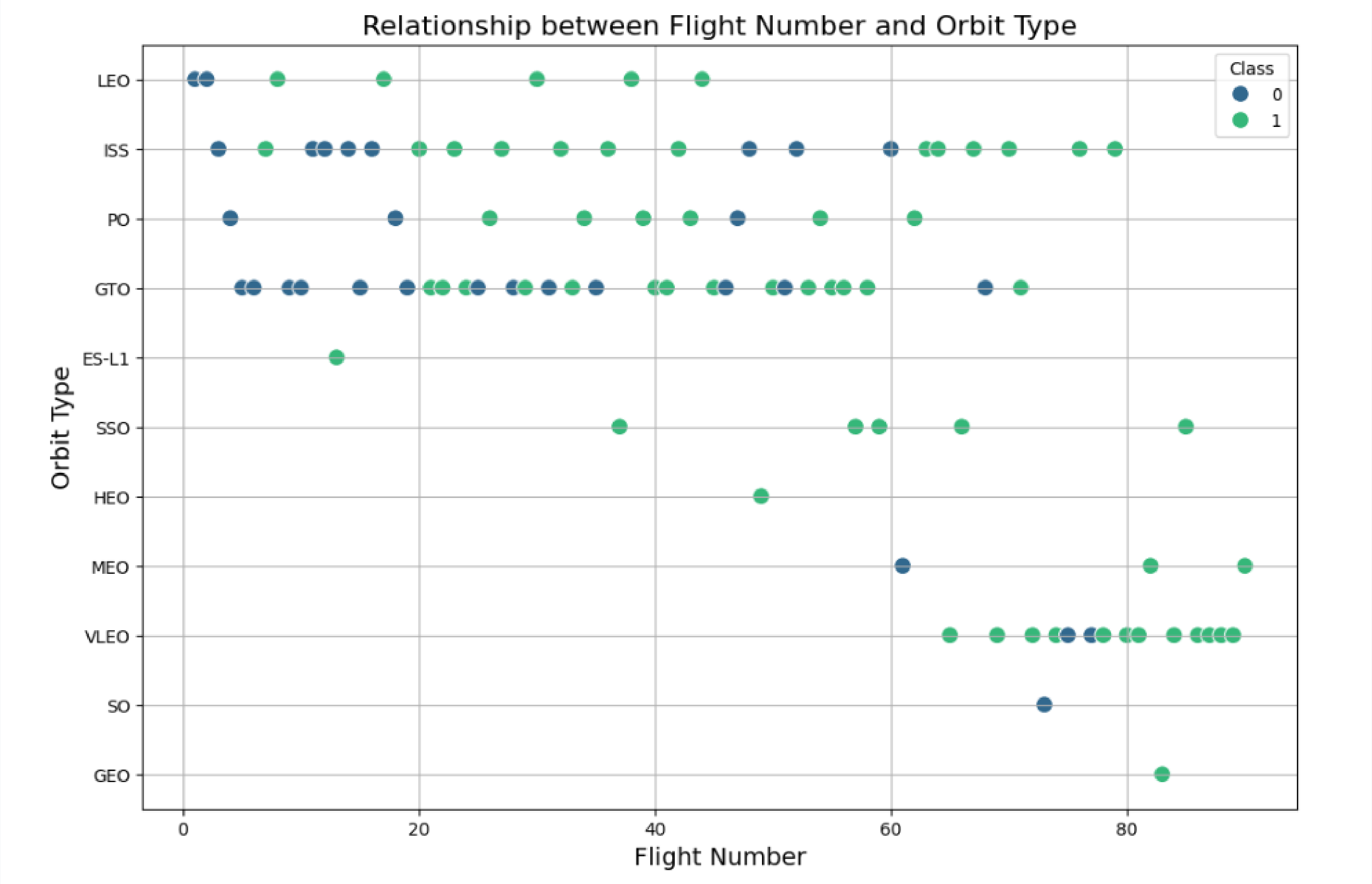
Success Rate vs. Orbit Type



- A bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO and SSO orbits with no failed first stage landings.
- SO orbit with no successful first stage landings.

Flight Number vs. Orbit Type

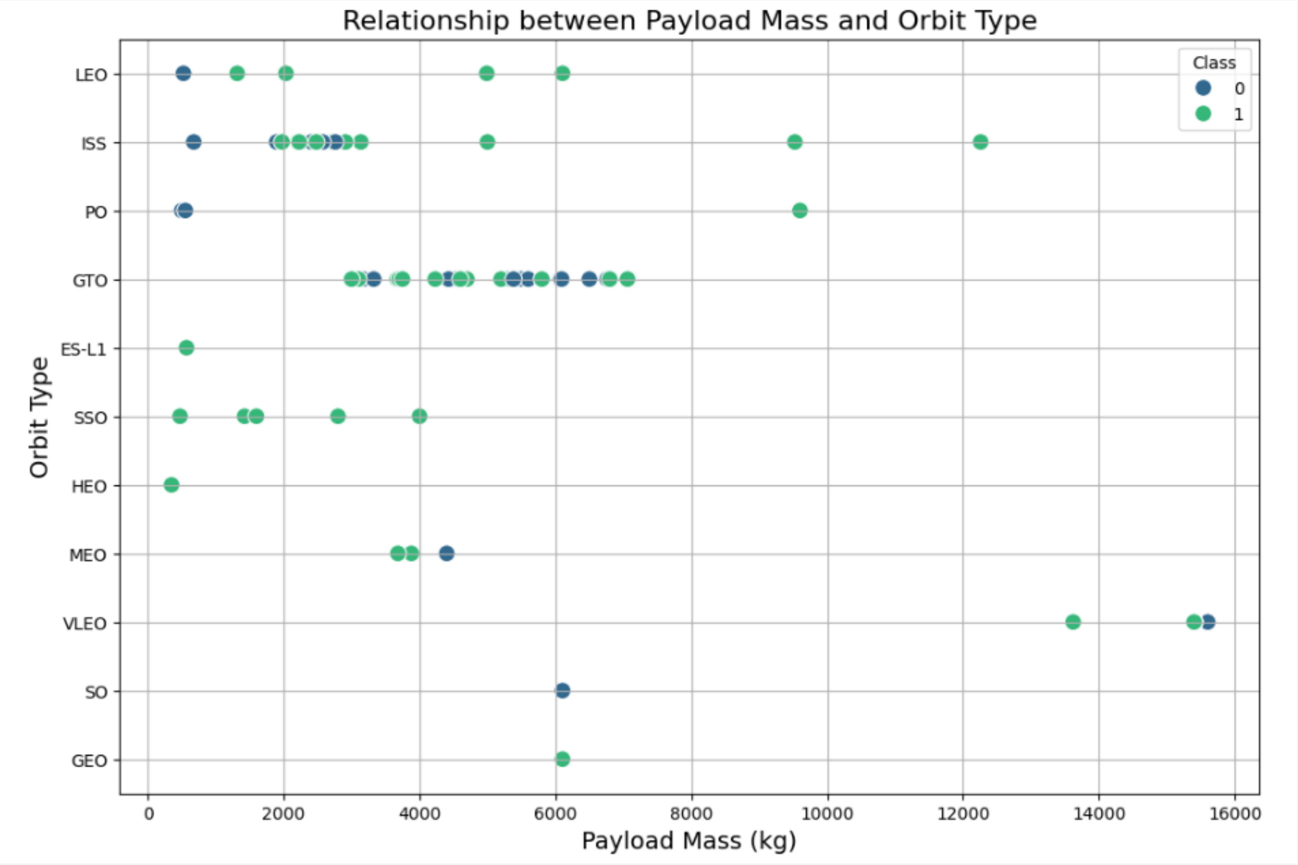
- Scatter plot of Flight number vs. Orbit type.
- LEO: the greater the number of flights, the greater the success.
- GTO: there is no clear correlation between the number of flights and success.



Payload vs. Orbit Type

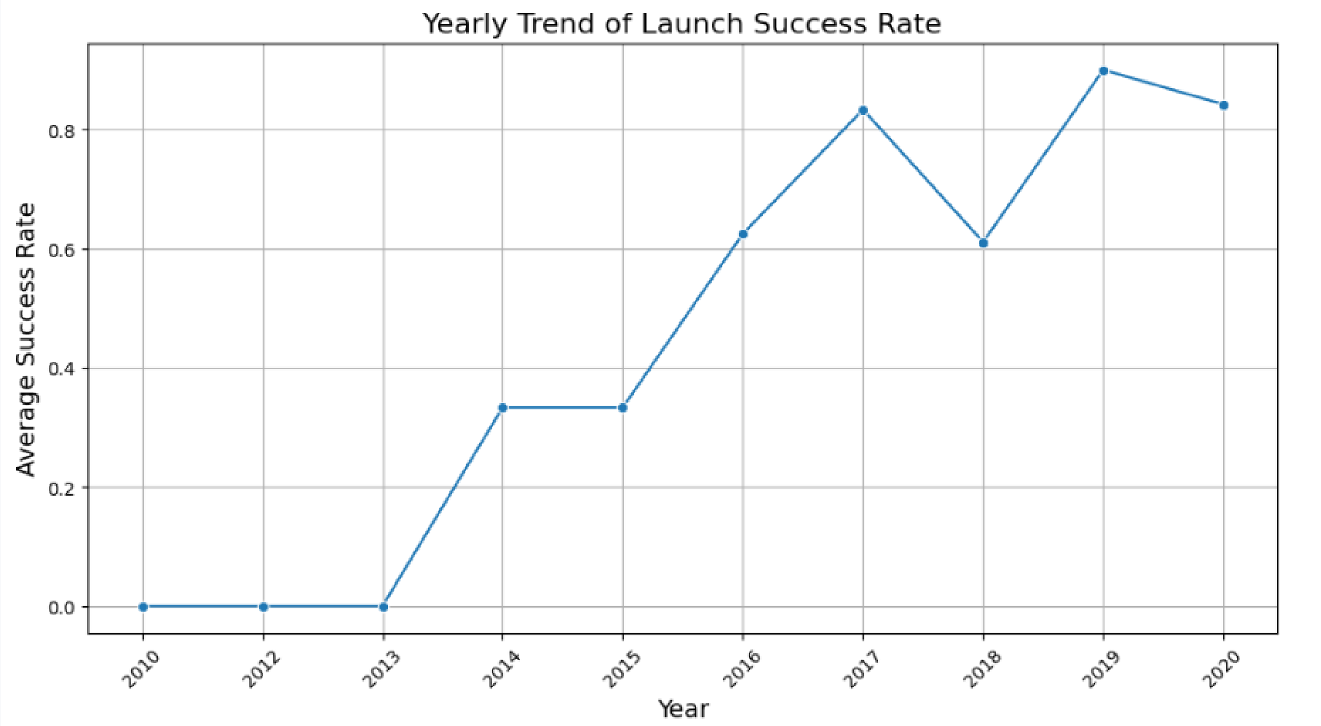
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- Line chart of yearly average success rate.
- It can be observed that the success rate since 2013 continued to increase until 2017 with a stable period between 2014 and 2015 to decline in 2018 and recover between 2019-2020.



All Launch Site Names

- Names of the unique launch sites
- **%sql SELECT DISTINCT** "Launch_Site" **FROM** SPACEXTABLE;
- Launch_Site:
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

- **%sql SELECT * FROM** SPACEXTABLE **WHERE** "Launch_Site" **LIKE** 'CCA%' **LIMIT** 5;

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 carried by boosters from NASA **(CRS)**

- %sql SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" LIKE '%NASA (CRS)%';

- Total_Payload_Mass:
  - None

Average Payload Mass by F9 v1.1

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

- **%sql SELECT** AVG("Payload_Mass__kg_") **AS** Average_Payload_Mass **FROM** SPACEXTABLE **WHERE** "Booster_Version" = 'F9 v1.1';
- Average_Payload_Mass 2928.4
- **Display average payload mass carried by booster version F9 v1.1**

First Successful Ground Landing Date

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

- **%sql SELECT** MIN("Date") **AS** First_Successful_Landing **FROM** SPACEXTABLE **WHERE** "Landing_Outcome" = 'Success (ground pad)';
- First_Successful_Landing 2015-12-22
- **Date when the first succesful landing outcome in ground pad was acheived.**

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
- **%sql SELECT DISTINCT** "Booster_Version" **FROM** SPACEXTABLE **WHERE** "Landing_Outcome" **=** 'Success (drone ship)' **AND** "Payload_Mass__kg_" **>** 4000 **AND** "Payload_Mass__kg_" **<** 6000;
- Booster_Version
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- **Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- **%sql SELECT** "Mission_Outcome", **COUNT**(*) **AS** Total_Count **FROM** SPACEXTABLE **GROUP BY** "Mission_Outcome";
- Mission_Outcome

  Failure (in flight) 1

  Success 98

  - Success 1

  Success (payload status unclear) 1
- **Total number of successful and failure mission outcomes**

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- **%sql SELECT** "Booster_Version" **FROM** SPACEXTABLE **WHERE** "Payload_Mass__kg_" = (**SELECT** MAX("Payload_Mass__kg_") **FROM** SPACEXTABLE);
- 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
- **Names of the booster_versions which have carried the maximum payload mass.**

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- **%sql SELECT** SUBSTR("Date", 6, 2) **AS Month**, "Landing_Outcome", "Booster_Version", "Launch_Site" **FROM** SPACEXTABLE **WHERE** "Landing_Outcome" = 'Failure (drone ship)' **AND** SUBSTR("Date", 1, 4) = '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

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

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

- **Landing_Outcome**

  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

- 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

Section 3

Launch Sites Proximities Analysis

Launch sites Falcon 9

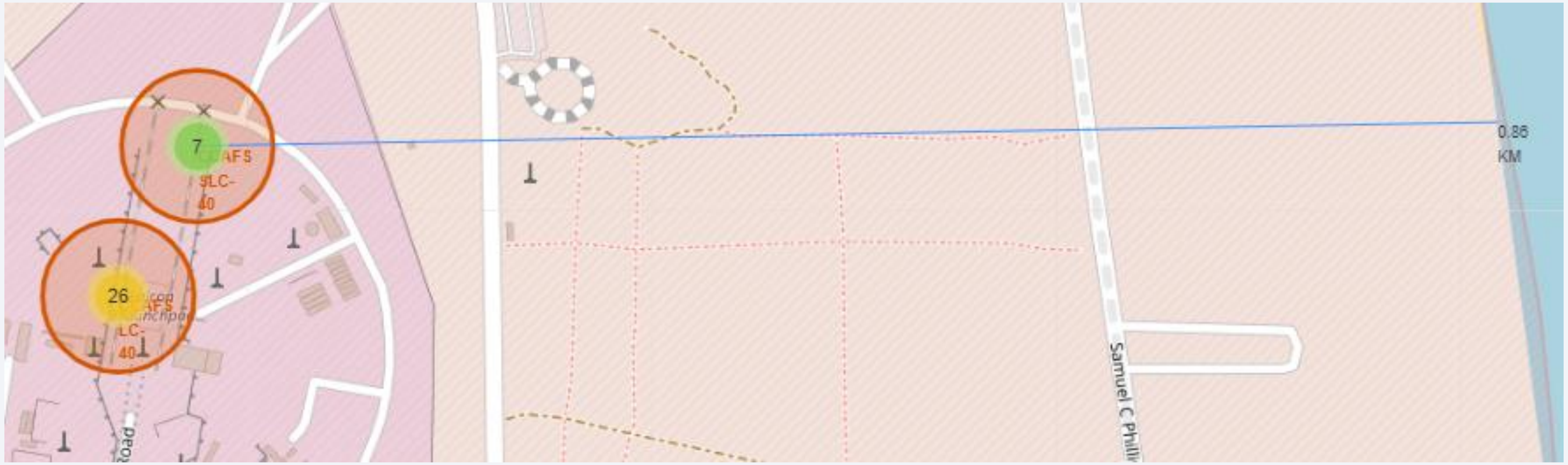
West Coast: Vandenberg Space Force Base (VAFB SLC-4E).

East Coast: Kennedy Space Center (KSC LC-39A) and Cape Canaveral Space Force Station (CCAFS SLC-40, CCAFS LC-40).

The success/failed launches for each site on the map

- From the color-labeled markers in the marker clusters, launch sites with relatively high success rates are identified.

The distances between a launch site to its proximities



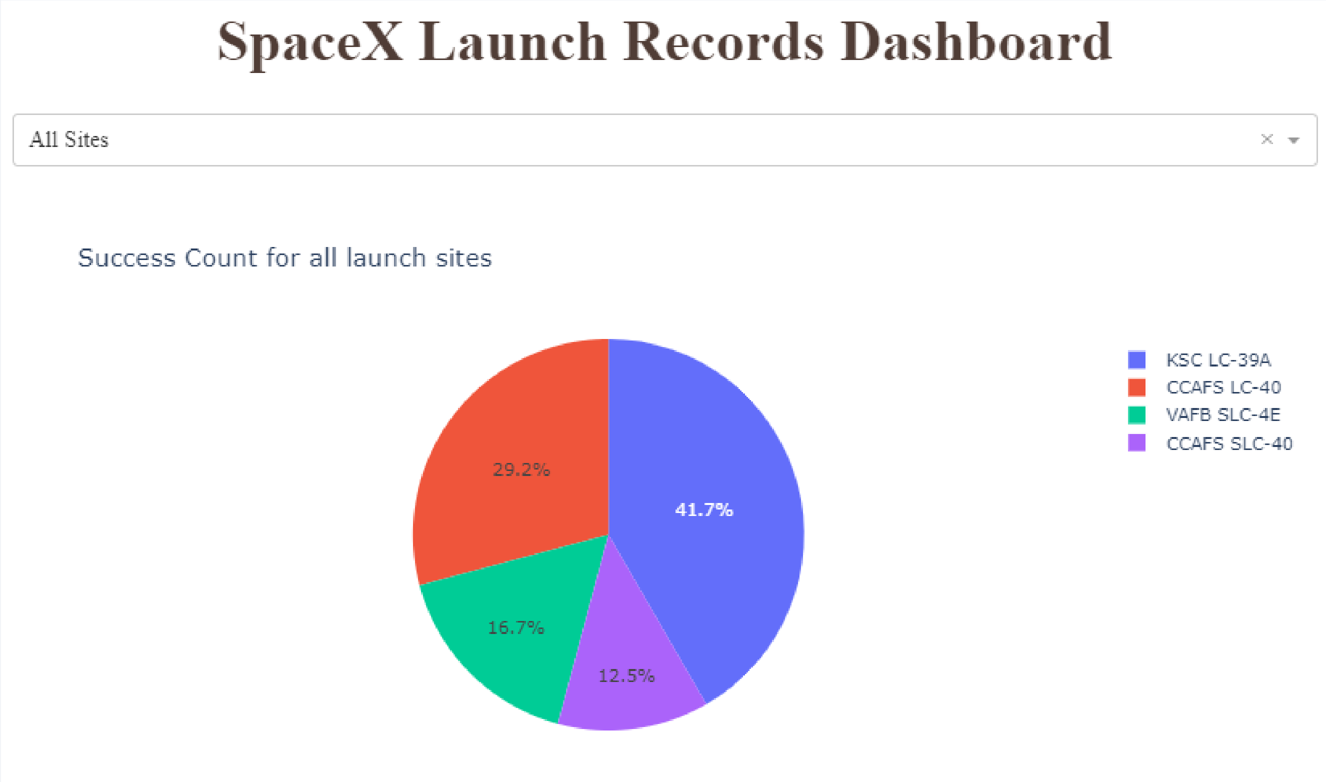
- **CCAFS SLC-40**

- distance_coastline = 0.8627671182499878 km
- distance_highway = 0.5834695366934144 km
- distance_railroad = 1.2845344718142522 km
- distance_city = 51.434169995172326 km

Section 4

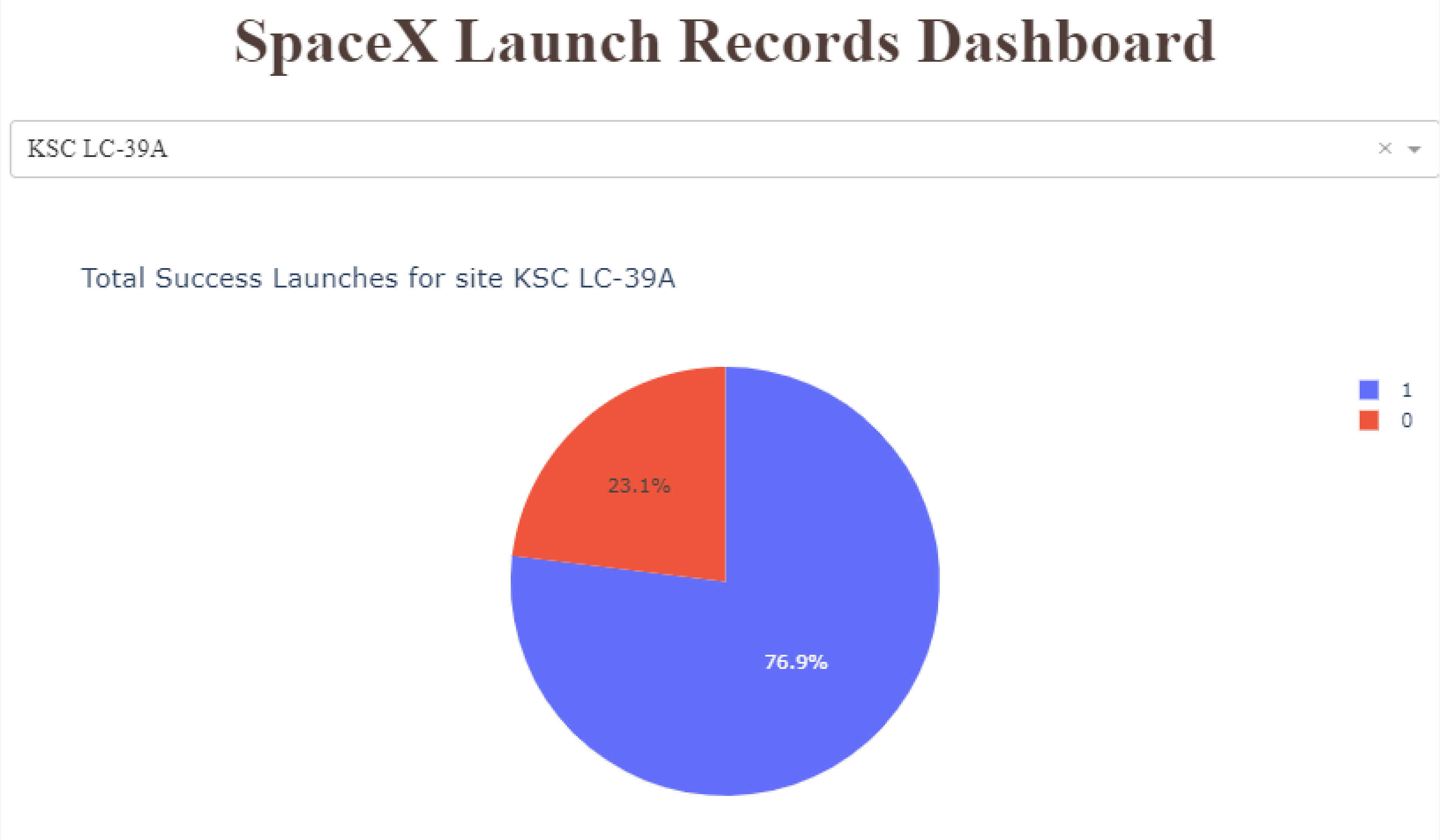
Build a Dashboard with Plotly Dash

Launch success count for all sites



- Proportion of successful rocket launches from different launch sites.

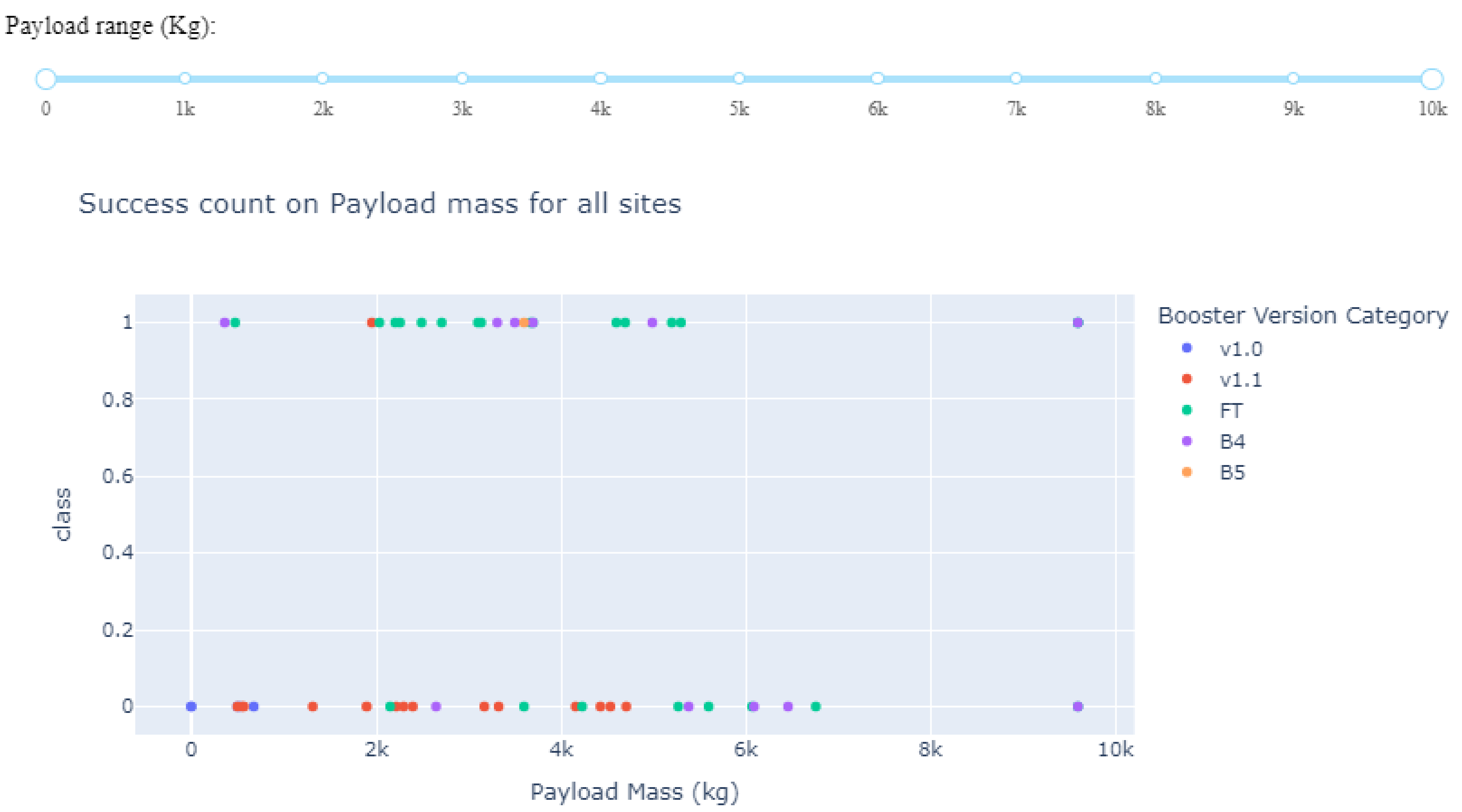
The launch site with highest launch success ratio



- Show the piechart for the launch site with highest launch success ratio

Payload vs. Launch Outcome

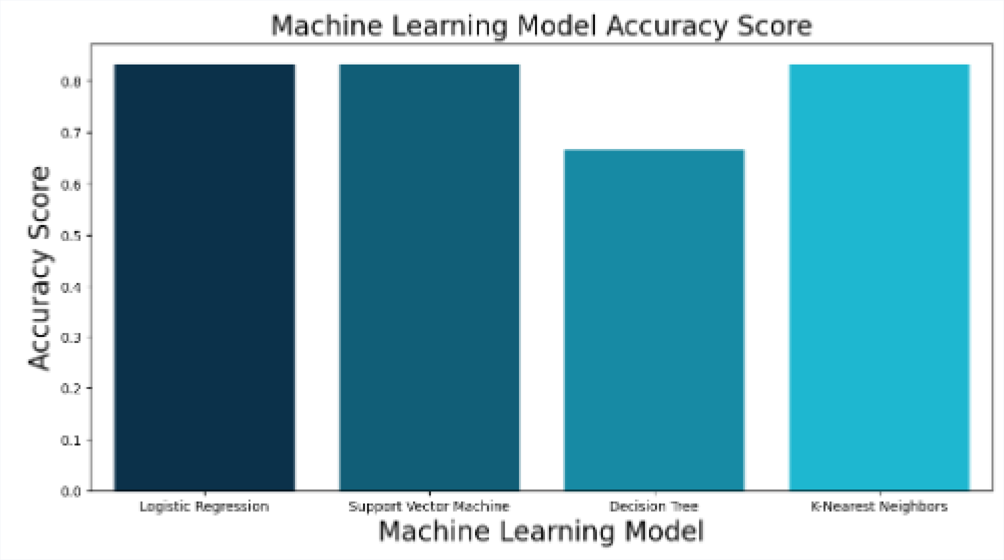
- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



Section 5

Predictive Analysis (Classification)

Classification Accuracy

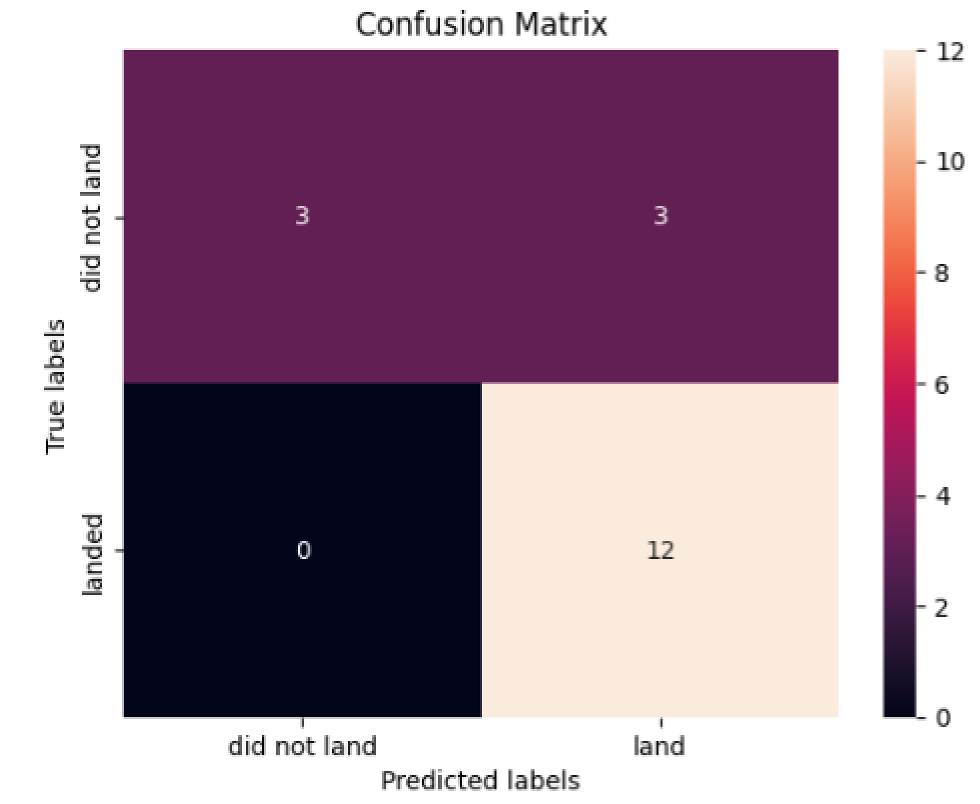


- The models had the same accuracy except for the Decision Tree model which was lower.

Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

- Overview:
  - True Postive - 12 (True label is landed, Predicted label is also landed)
  - False Postive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- Machine learning models can be used to predict SpaceX Falcon 9 first stage landing results and will become more successful as more launches are conducted.

Appendix

- **GitHub:** IBM-Applied-Data-Science-Capstone-SpaceX

Thank you!