IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Samanta Jerin
4th Nov 2024

Outline

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

Executive Summary

This capstone project predicts the likelihood of a successful Falcon 9 first-stage landing, leveraging SpaceX's launch history data.

- Summary of methodologies: Data was collected through the SpaceX REST API and web scraping, followed by data wrangling, analysis, and exploratory visualization. Key methodologies include SQL, data visualization, and machine learning.
- Summary of all results: The project developed and evaluated several machine learning models, concluding that the Decision Tree model offers the most effective predictive accuracy. This work provides insights for understanding landing success factors, enhancing predictions with data-driven machine learning strategies.

Introduction

- Project background and context: This project investigates the cost efficiency of SpaceX’s Falcon 9 rocket, which achieves substantial savings through first-stage reusability, reducing launch costs to $62 million compared to competitors' $165 million.
- Problems you want to find answers: The goal is to predict successful first-stage landings using a machine learning pipeline, which could inform competitive bids in the space launch market. Analysis will explore patterns in SpaceX launch data, factors affecting success rates, and the predictability of outcomes based on historical data.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Employed SpaceX REST API and web scraping to obtain Falcon 9 mission data..

Perform data wrangling

- Conducted data wrangling, focusing on Falcon 9 data and converting outcomes into training labels

Perform exploratory data analysis (EDA) using visualization and SQL:

- Utilized SQL and visualization tools for initial data analysis.

Perform interactive visual analytics using Folium and Plotly Dash:

- Created interactive visual analytics with Folium and Plotly Dash to examine launch sites, orbits, payloads, and mission outcomes.

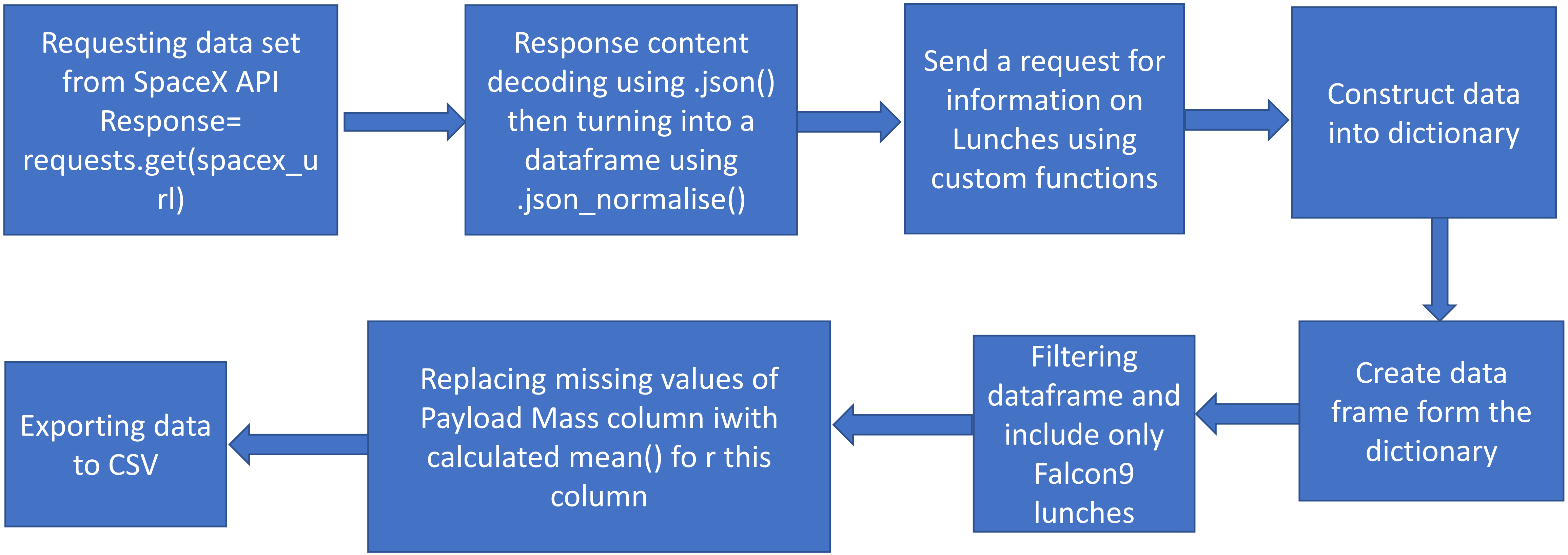
Perform predictive analysis using classification models

- Standardized data and split into training and testing sets.
- Applied classification models, with hyperparameter tuning to enhance accuracy, producing a pipeline to predict first-stage landing success.

Data Collection

- **SpaceX API**: Used to gather data on Booster Version, Launch Site, Payload Mass, and Core.
- Endpoint URL: https://api.spacexdata.com/v4/launches/past
- Data accessed via a GET request, returned in JSON format, and converted to a Pandas DataFrame with .json_normalize().
- **Web Scraping**: Collected additional data, including Flight Number, Date & Time, Booster Version & Landing, Launch Site, Payload, and Orbit Destination.

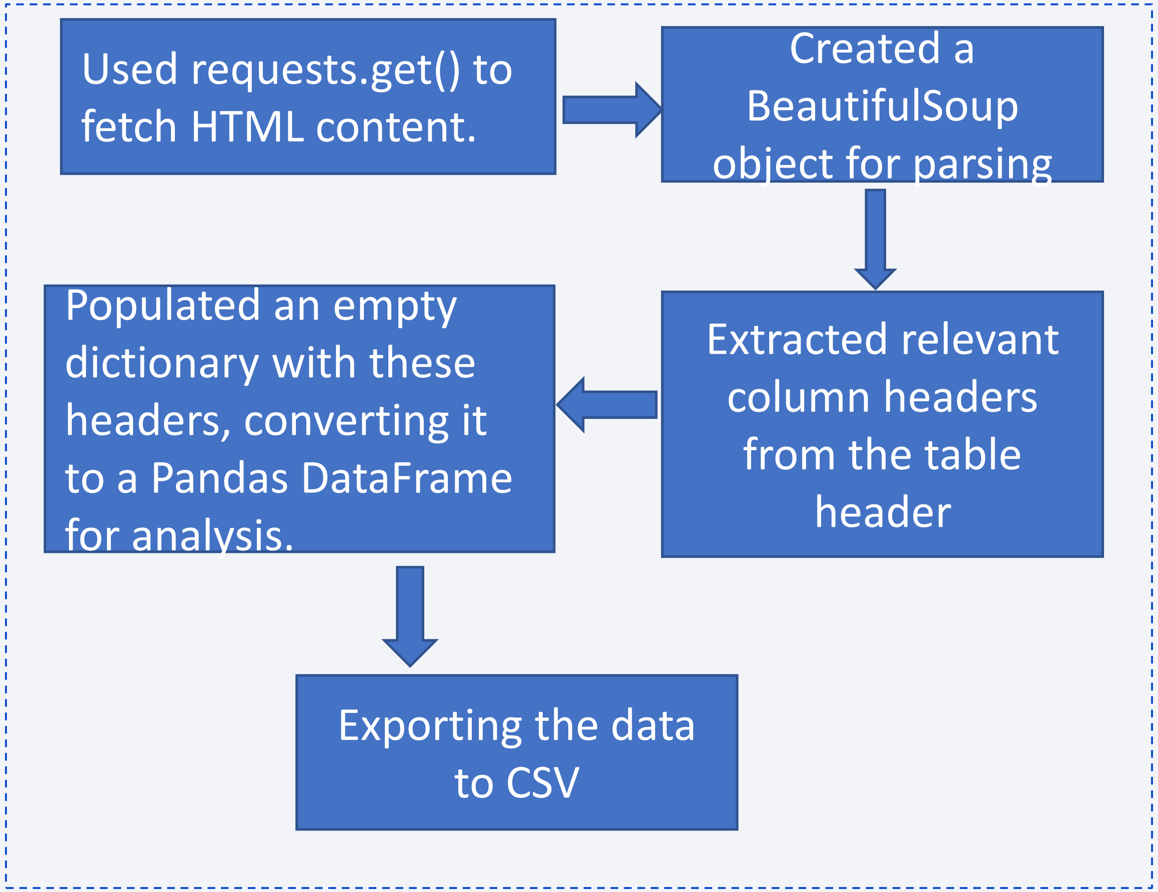
Data Collection-Flow chart



Git Repository: https://github.com/Arony09/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

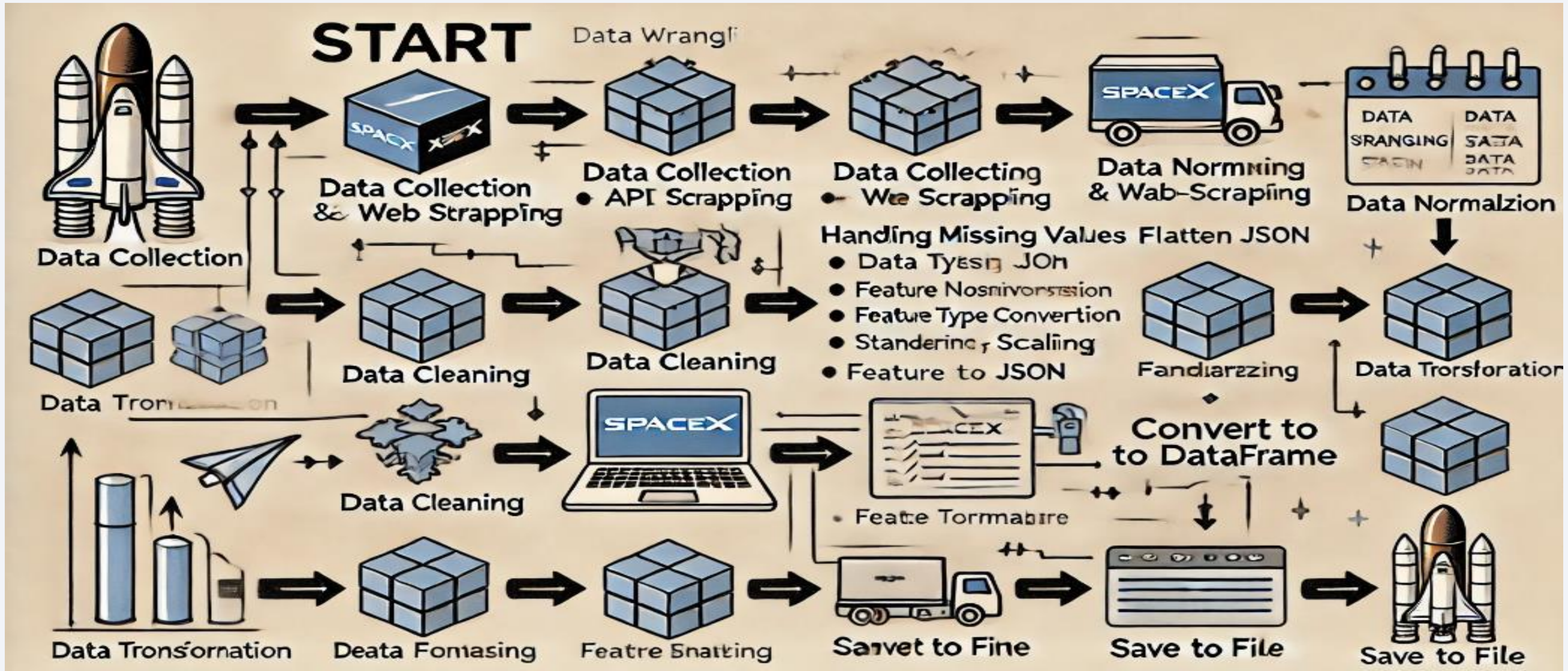
- **Data Source**: Wikipedia table for Falcon 9 launches.
- **URL**: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Git Repository: https://github.com/Arony09/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Wrangling

The data wrangling process began with data collection, using the SpaceX API to gather specific fields (e.g., Booster Version, Launch Site, Payload Mass) and supplementing it with additional data through web scraping (e.g., Flight Number, Date & Time). Next, data cleaning was performed, which included handling missing values, converting data types, and standardizing formats. The JSON data was then normalized using json_normalize() for a flat structure, and any necessary scaling or transformations were applied. Feature engineering was used to create new, insightful variables, and the API and scraped data were merged into a single dataset. Finally, the data was loaded into a Pandas DataFrame, cleaned, and stored in a suitable format for predictive modeling and analysis.



Git Repository: https://github.com/Arony09/DATA-WRANGLING-LAB/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization

Tools Used:

- **Seaborn**: Explored feature relationships with scatter plots for:
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- **Matplotlib**: Visualized data trends using:
- **Bar Charts** for comparing success rates across orbit types.
- **Line Charts** for observing yearly trends in launch successes.

Key Visualizations:

- **Scatter Chart**: Shows feature correlations.
- **Bar Chart**: Illustrates success rate comparisons by orbit.
- **Line Chart**: Displays the annual trend in launch outcomes.

Git Repository: https://github.com/Arony09/Exploring-and-Preparing-Data

EDA with SQL

Key Queries

- **DISTINCT**: Retrieve unique launch site names.
- **SUM(), AVG()**: Calculate total and average payload mass for specific boosters (e.g., NASA(CRS) and F9 v1.1).
- **COUNT()**: Determine the number of successful and failed missions.
- **GROUP BY**: Categorize records by factors like launch sites.
- **ORDER BY DESC**: Sort successful landings in descending order between 04-06-2010 and 20-03-2017.

Specific Data Extracted:

- Launch sites starting with "CCA"
- Records of successful drone ship landings with payloads between 4000-6000 kg
- First successful ground landing date
- Booster versions with maximum payload
- Git Repository: https://github.com/Arony09/Exploring-and-Preparing-Data

Build an Interactive Map with Folium

Launch Site Visualization:

- Created markers and circles to represent all launch sites, highlighting their national distribution.
- Each site features a cluster of markers indicating successful and failed launch locations, providing insights into the success rates at each site.

Distance Analysis:

- Drew two lines from each launch site:
- One line to the nearest coastline marker.
- A second line to a major highway intersection.
- This visual representation illustrates the distance from launch sites to hazardous areas, such as rocket launch activities.

Map Features:

- **Circle Object**: Added highlighted circles at specific coordinates with descriptive popups using folium.Popup().
- **Marker Object**: Utilized markers to signify important elements on the map, functioning as signposts for user navigation.
- **MousePosition Object**: Enabled coordinate retrieval for user interaction when hovering over specific points on the map.
- **PolyLine Object**: Implemented lines connecting launch sites to selected points for better spatial understanding.
- This interactive mapping project enhances the visualization and analysis of launch site data, facilitating informed decision-making regarding safety and operational efficiency.

Git Repository: https://github.com/Arony09/Interactive-Visual-Analytics-with-Folium

Build a Dashboard with Plotly Dash

Dashboard Overview

- Developed an interactive dashboard that allows users to select different launch sites and payload masses, providing a comprehensive view of launch data.

Key Visualizations:

➢ **Pie Chart**:

- Displays the number of launches per site.
- Dynamically updates to show the success rate based on the selected launch site from the dropdown list.

➢ **Scatter Plot**:

- Visualizes the relationship between payload mass and booster usage.
- Updates in real time as users adjust the payload range using a slider.

User Interaction Components:

➢ **Dropdown List**:

- Enables users to choose specific launch sites or view data for all sites combined.
- Affects the pie chart to display success rates pertinent to the selected site.

➢ **Range Slider**:

- Allows users to define a payload mass range.
- Interacts with the scatter plot to filter data on booster landings based on the selected payload range, categorized by booster versions.
- This Plotly Dash dashboard offers a user-friendly interface for analyzing launch data, facilitating insights into launch success rates and booster performance relative to payload masses.

Predictive Analysis (Classification)

- **Data Preparation**:

•Created an array using NumPy from the "Class" column in the data frame to define the target variable.
•Standardized and transformed features for prediction, assigning them to variable.

- **Data Splitting:**

•Utilized train test_split() to divide the data into training and test sets, using a test size of 20% (0.2) and a random state of 2, resulting in 18 test samples.

- **Model Evaluation:**

•Evaluated multiple classification algorithms:

- **Logistic Regression**

•Support Vector Machine (SVM)
•K-Nearest Neighbors (KNN)
•Decision Trees
•Computed accuracy scores and confusion matrices for each model to identify the best performer.

- **Hyperparameter Optimization:**

•Created a GridSearchCV object for each model to tune hyperparameters.
•Fitted models using parameter grids in dictionary format to determine the optimal hyperparameters.
•Retrieved the best parameters using the best_params_ attribute.

- **Performance Assessment:**

•Identified the best-performing model by analyzing the best_score_ attribute for accuracy.
•Visualized the model's performance on the test dataset using confusion matrices.
•This predictive analysis framework leverages various classification algorithms to ascertain the most effective model for classification tasks, enhancing predictive capabilities through rigorous evaluation and optimization of model parameters.

Git Repository: **https://github.com/Arony09/Space-X-Falcon-9-First-StageLandingPredictSpace-X-FalconComplete-the-Machine-Learning-Prediction-lab/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb**

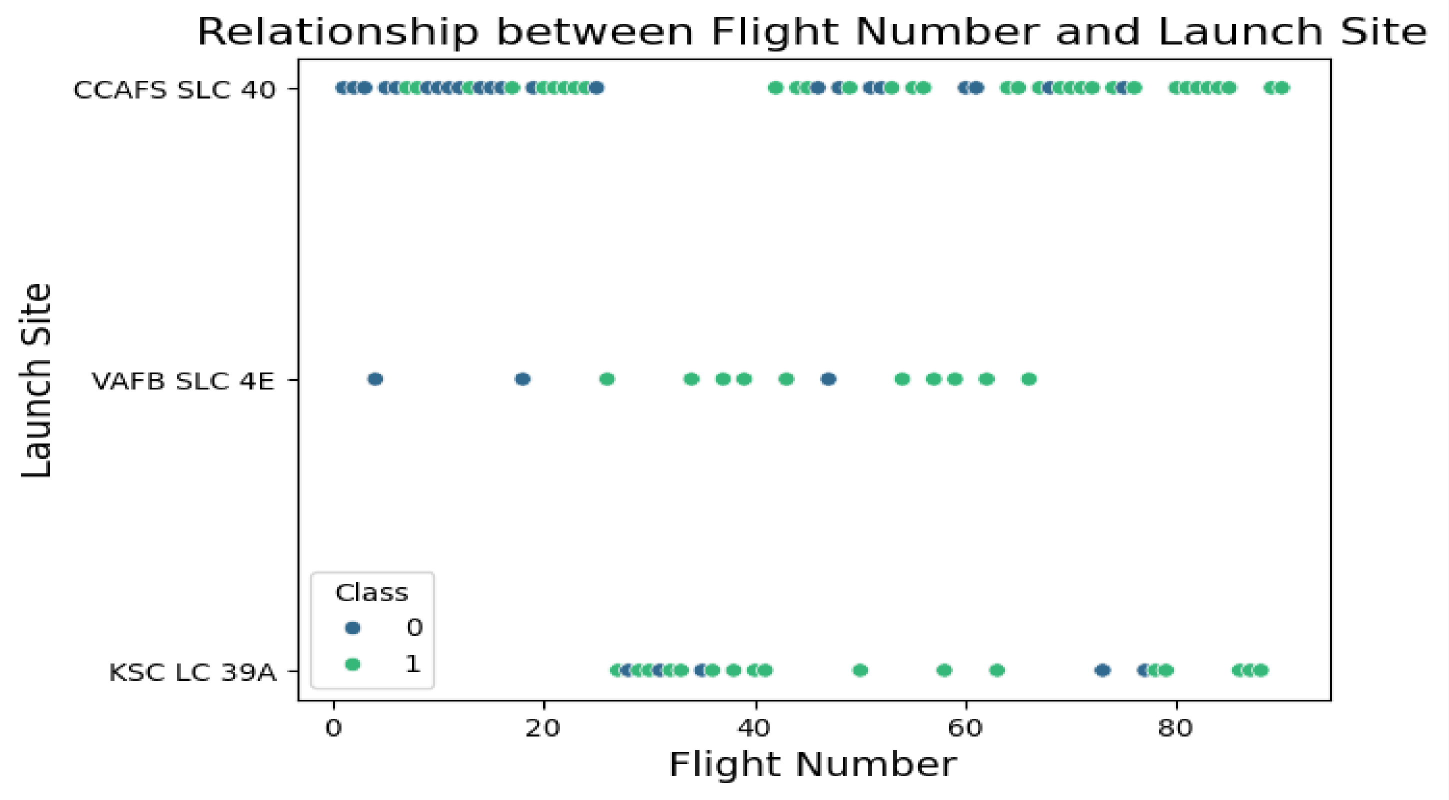
Results

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

Section 2

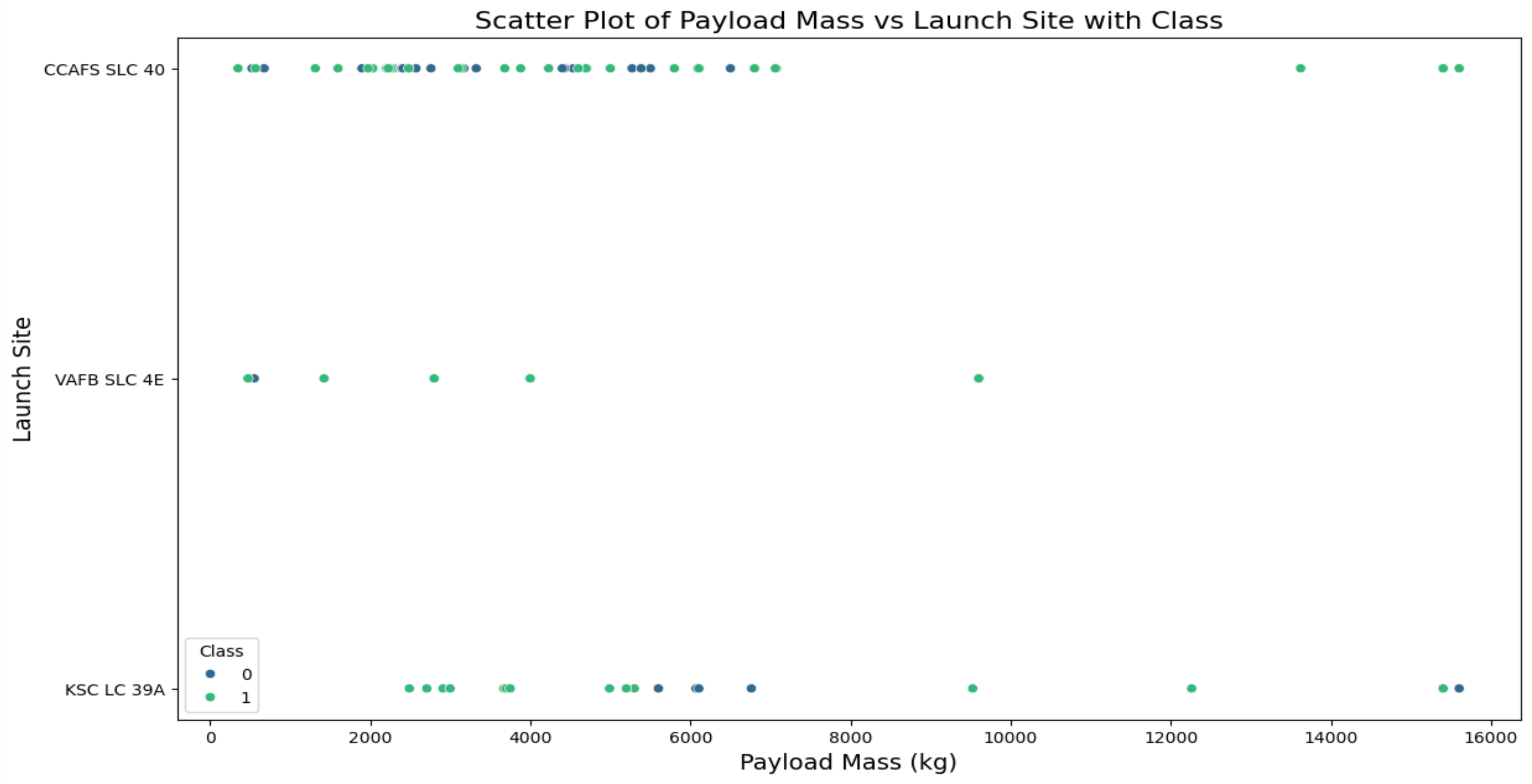
Insights drawn from EDA

Flight Number vs. Launch Site



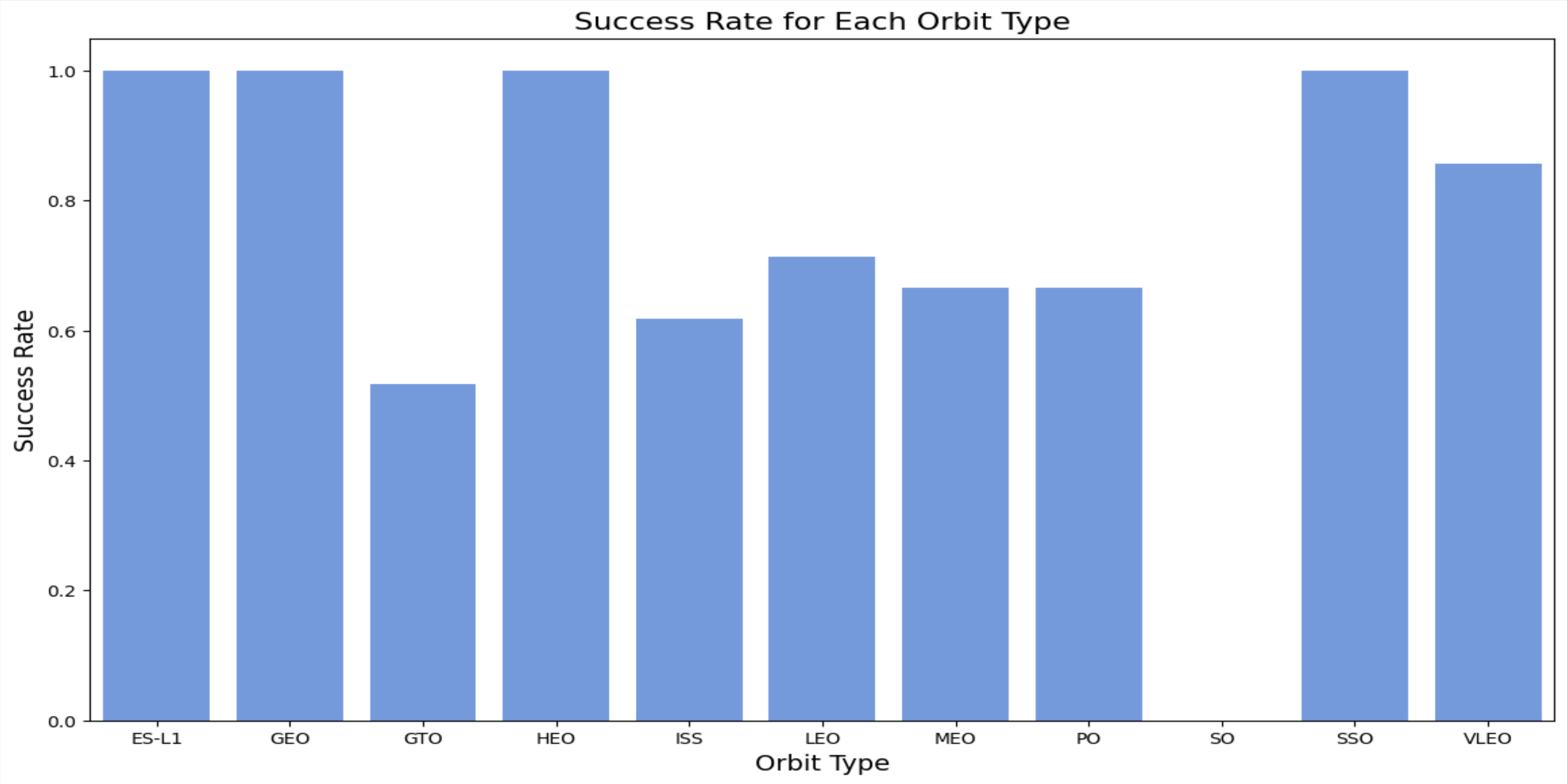
CCAFS SLC 40 lunch site shows significantly higher lunches compared to other sites.

Payload vs. Launch Site



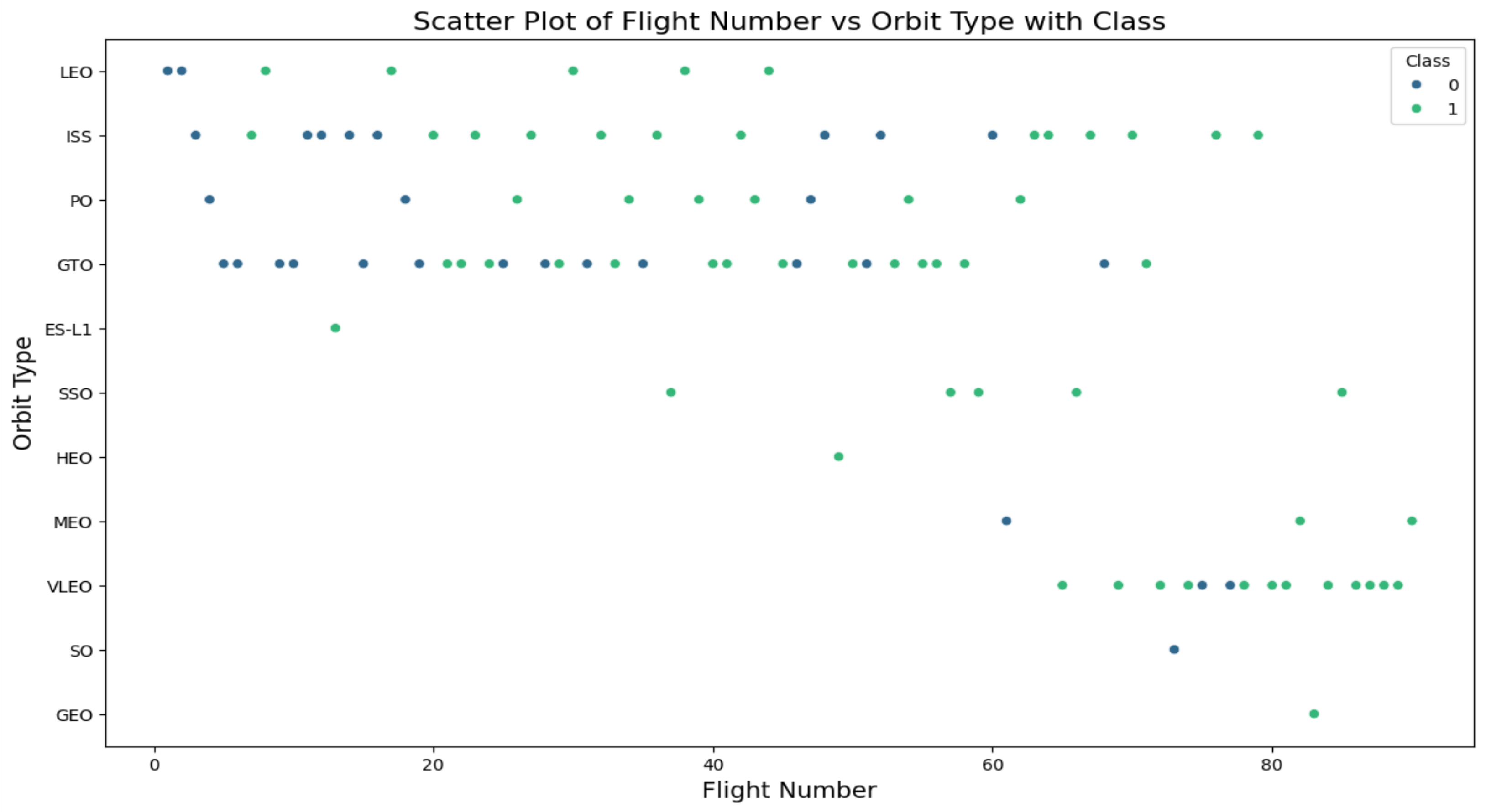
There are significantly higher number of lunch site in CCAFS SLC 40 sites with lower Pay Load Mass(KG)

Success Rate vs. Orbit Type



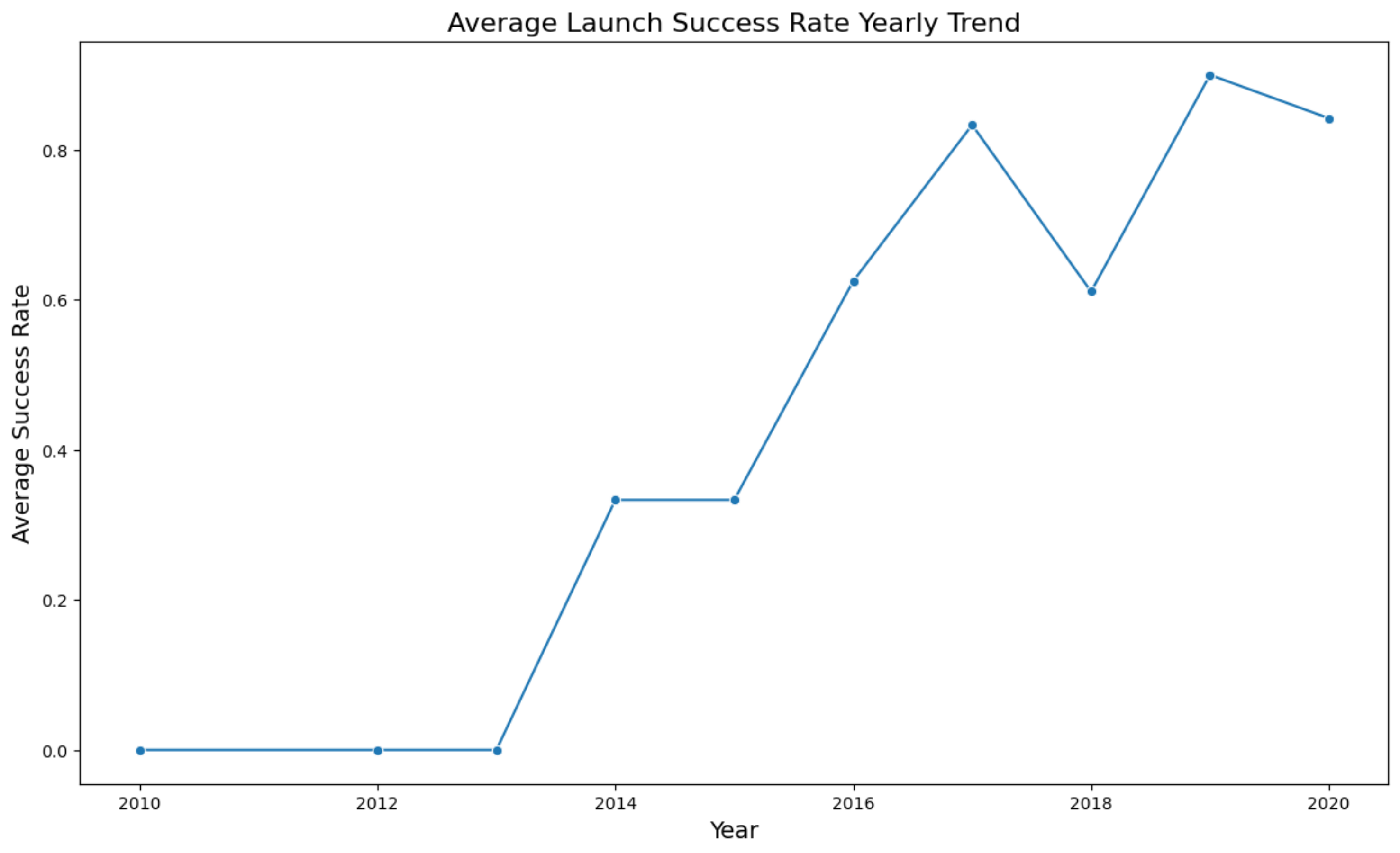
ES-L-1, GEO, HEO ,SSO and VLEO orbits show the highest success rate.

Flight Number vs. Orbit Type



There is a trend of shifting to VLEO lunches in recent years.

Launch Success Yearly Trend



Success rate for lunching has increase significantly between 2023-2020

All Launch Site Names

```
%sql select DISTINCT Launch_Site from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL WHERE Launch_Site LIKE "%CCA%" LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

ate	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
10--04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
10--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)
12--22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
12--08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
13--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
45596

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [14]:
```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Booster_Version = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by booster version F9v1.1 is 2928.4

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(Date) from SPACEXTBL WHERE Landing_Outcome LIKE "%Success%"
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

The first successful ground landing occurred on 22 December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

30]:
```
%sql select Booster_Version from SPACEXTBL WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MAS
```

```
* sqlite:///my_data1.db
Done.
```

30]:

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

List of the Boosters which have successful drone shipping with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [17]:
```
%sql select COUNT(Mission_Outcome) from SPACEXTBL GROUP BY "Mission_Outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

Out[17]:

COUNT(Mission_Outcome)
1
98
1
1

List of Mission Outcome with total number of successful and failure missions

Boosters Carried Maximum Payload

```
[18]: %sql select Booster_Version from SPACEXTBL  WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_

 * sqlite:///my_data1.db
Done.
```

t[18]:

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

List of Booster Versions that carried Maximum Payload

2015 Launch Records

```
%sql select substr(Date, 6,2) as month, substr(Date,0,5)=='2015',Landing_Outcome, Booster_'
```

```
 * sqlite:///my_data1.db
Done.
```

month	substr(Date,0,5)=='2015'	Landing_Outcome	Booster_Version	Launch_Site
01	1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
01	0	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
03	0	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	0	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

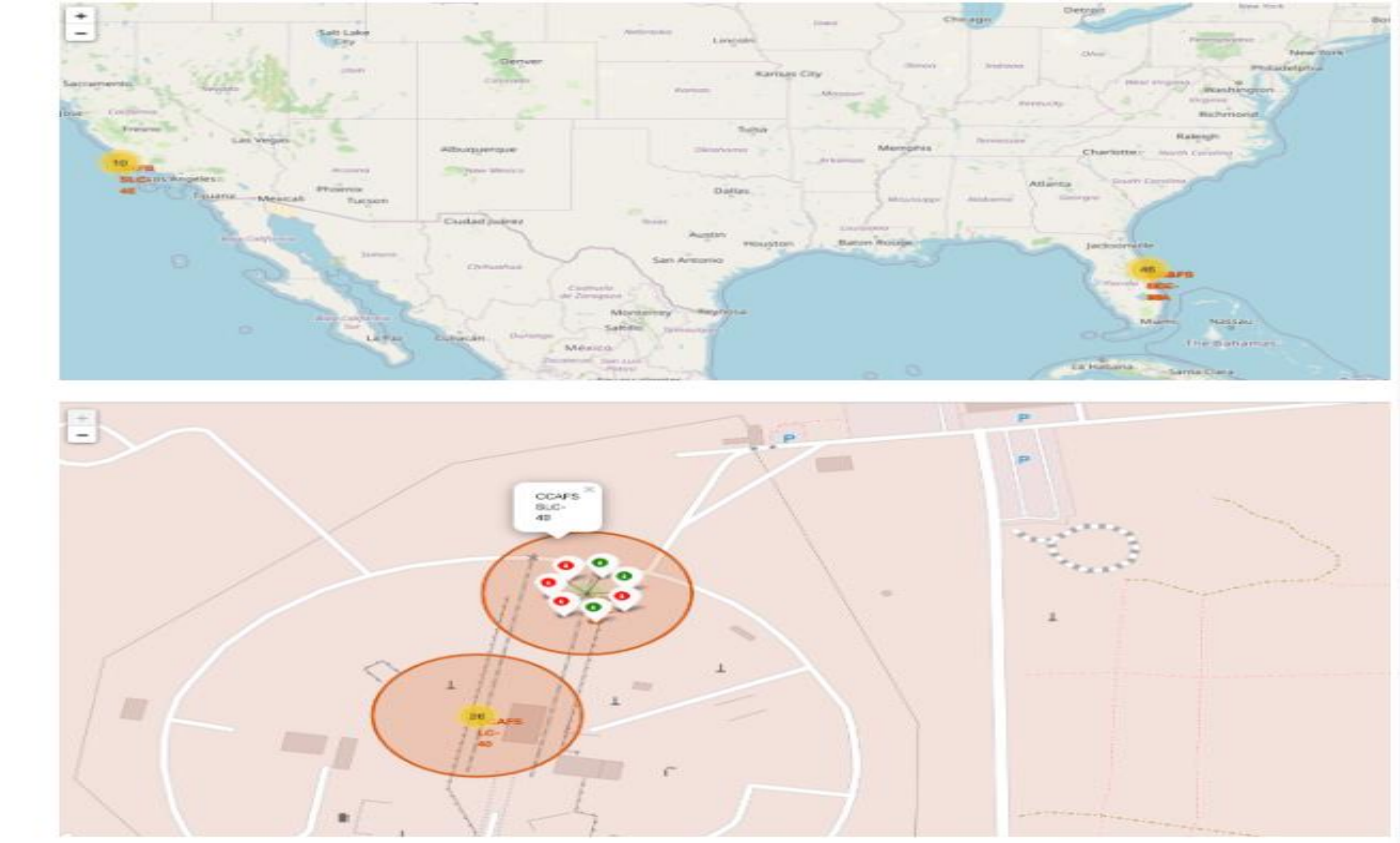
Record of 2015 Landing outcome of 5 different Booster_Version from CCAFS LC-40 and VAFB SLC-4E lunch site

Section 3

Launch Sites Proximities Analysis

Folium Map Screenshot 1

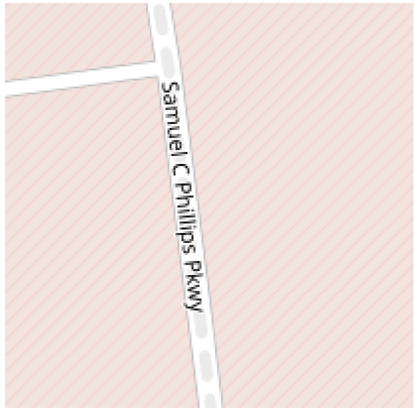
Folium Map Screenshot 2



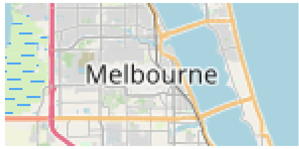
<Folium Map Screenshot 3>

A railway map symbol may look like this:

A highway map symbol may look like this:



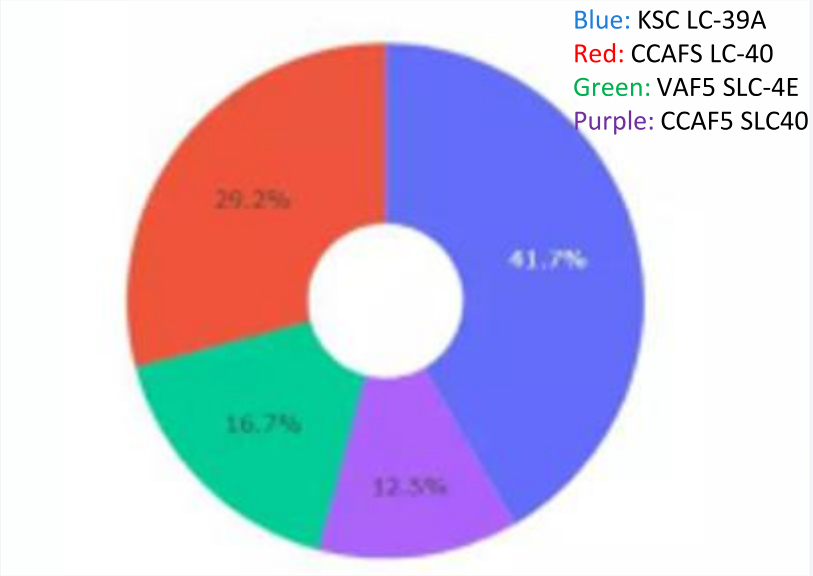
A city map symbol may look like this:



Section 4

Build a Dashboard with Plotly Dash

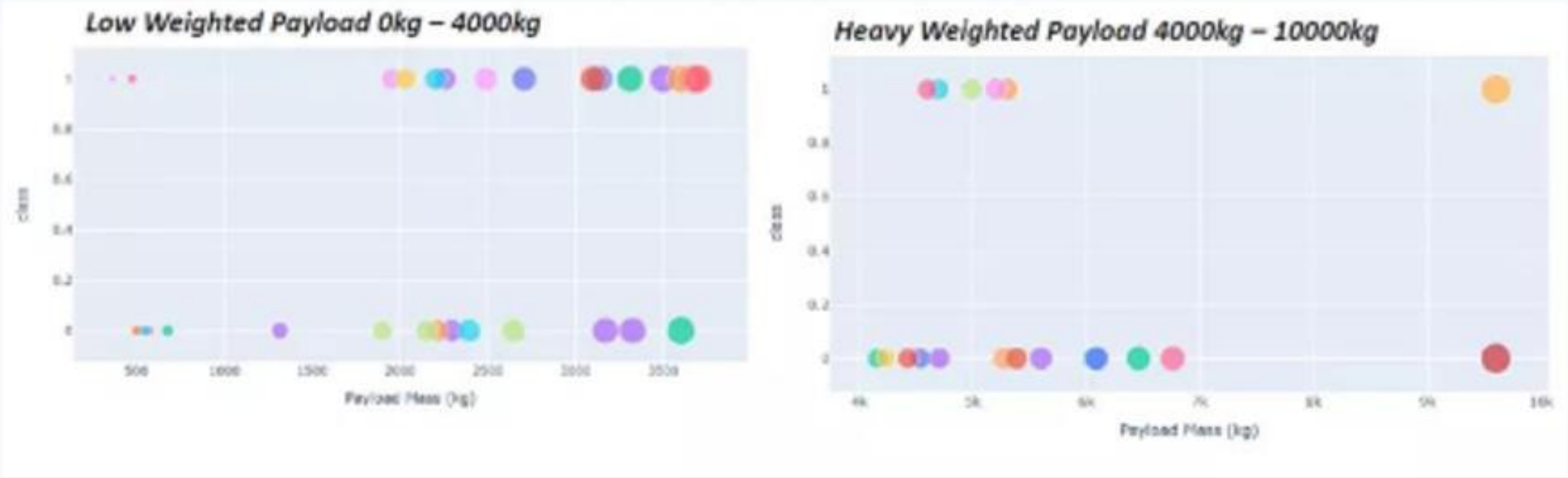
Dashboard Screenshot :Total Success Lunches by all sites



KSC LC-39A showed the most successful launches in all sites.

<Dashboard Screenshot 2>

<Dashboard Screenshot 3>



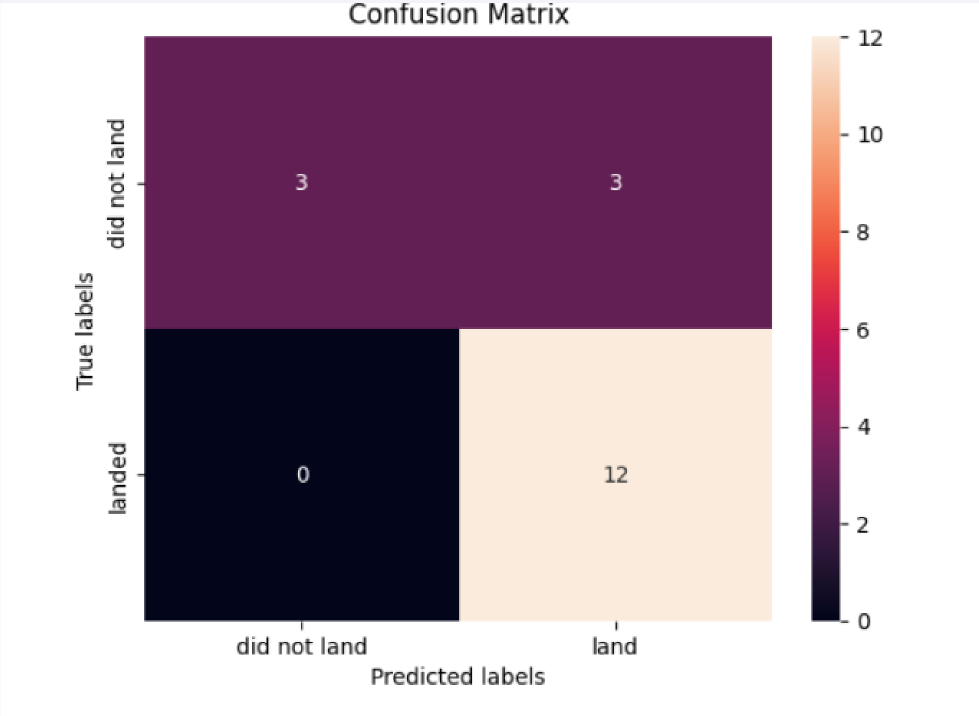
Success rate for lower weighted payloads are higher than that of the heavier ones.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Confusion Matrix



Observations:

- The confusion matrix indicates a **very high True Positive rate of 12**, demonstrating that the model accurately predicted all successful landings.
- There is **no False Positive**, suggesting that there were no incorrect predictions of landings when they did not occur.
- However, the model faced a **False Negative issue**, where 3 instances that actually landed successfully were incorrectly predicted not to land. This highlights a critical area for improvement in the model’s predictive accuracy.
- The **True Negative count is 3**, which is consistent across the other models, indicating reliable performance in identifying unsuccessful landings

Conclusions

Visualization and SQL analysis revealed significant patterns in the data:

- **KSC LC-39A** boasts the highest overall success rate among all launch sites throughout the study period.
- **CCAFS SLC 40** has the highest success rate for more recent launch attempts.
- The **Falcon 9 first stage** is more likely to land successfully when the payload mass ranges between **2,000 kg and 4,000 kg**.
- **Booster Version FT** shows a higher likelihood of successful landings.
- For the orbit type **'SO'**, the Falcon 9 first stage has never achieved a successful landing.
- **Model Performance**:
- The **Decision Tree model** emerged as the most accurate predictor for whether the Falcon 9 first stage will land successfully, achieving an accuracy score of **0.85**.

General Observations:

- SpaceX data is highly accessible, providing ample details for in-depth analysis.
- Most successful missions had payloads ranging between **2,000 kg and 6,000 kg**.
- All analyzed launch sites are strategically located at a safe distance from major urban areas.
- The Falcon 9 missions demonstrate a strong track record of success, regardless of payload mass, orbit type, or landing site—including drone ships in the ocean.
- The reusability of the Falcon 9 significantly reduces costs compared to competitors, coupled with its high success rate, bolstering SpaceX’s competitive position.
- In summary, considering the analyzed data, successful mission characteristics, and model performance, we confidently conclude that the **Falcon 9 rocket will land successfully again** in future missions.

Thank you!