Falcon 9 Success: Predictive Data Insights

<DATE>

IBM Developer
SKILLS NETWORK

Richard Van Buren
February 17th 2024

Outline

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

Executive Summary

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

Introduction

❏ **Project Overview and Relevance**

❏ SpaceX offers Falcon 9 rocket launches at a cost of 62 million dollars on its website, significantly lower than the upwards o f 165 million dollars charged by competing providers. A substantial portion of these cost savings comes from SpaceX's innovative ability to reuse the rocket's first stage. Understanding whether the first stage will successfully land is crucial for estimating the cost of each launch. This knowledge is particularly valuable for any competing firm aiming to challenge SpaceX in the rocket launch market. The project's primary aim is to develop a machine learning pipeline capable of predicting the successful landing of the Falcon 9's first stage.

❏ **Research Objectives and Questions**

❏ The project seeks to uncover insights into the following key areas:

❏ Determining Success Factors: Identifying which variables play a pivotal role in ensuring the successful landing of the rocket 's first stage. This involves understanding the weight each factor carries in the predictive model.

❏ Feature Interactions and Success Rates: Exploring how different variables interact with each other and how these interactions affect the likelihood of a successful landing. This analysis will help in understanding the complexity and dynamics of rocket landings.

❏ Optimal Operating Conditions: Investigating the specific conditions required to maximize the probability of a successful land ing. This includes environmental factors, rocket configurations, and any pre-launch procedures critical to the mission's success.

❏ By addressing these questions, the project aims to not only enhance the predictive accuracy of the machine learning model but also provide actionable insights for improving the efficiency and reliability of space missions.

Methodology

Methodology

Executive Overview

❑ **Methodology for Data Acquisition**: Data was sourced through the SpaceX API alongside web scraping techniques employed on Wikipedia.

❑ **Data Preparation**: Categorical variables underwent one-hot encoding to transform them into a machine-readable format.

❑ **Exploratory Data Analysis (EDA)**: Visualization tools and SQL queries were utilized for an in-depth analysis of the data.

❑ **Interactive Data Visualization**: Tools such as Folium and Plotly Dash were implemented for dynamic data exploration.

❑ **Predictive Analysis**: The project focused on employing various classification models to forecast outcomes.

❑ **Model Development and Optimization**: Guidance on constructing, fine-tuning, and assessing the performance of classification models.

Data Collection

The data was collected using various methods

❑ Data collection was done using get request to the SpaceX API.

❑ Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

❑ We then cleaned the data, checked for missing values and fill in missing values where necessary.

❑ In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

❑ The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

Data was gathered via a GET request to the SpaceX API, followed by cleaning and initial data wrangling and formatting tasks.

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

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

The link to the notebook is
https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/
Data%20Collection%20API.ipynb/blob/main/Data%20Collection%20API.ipynb

Data Collection – Scraping

Web scraping was performed to extract Falcon 9 launch records using BeautifulSoup. The data from the table was then parsed and transformed into a pandas DataFrame.

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

```
200
```

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

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

Print the page title to verify if the `BeautifulSoup` object was created properly

```python
# Use soup.title attribute
soup.title
```

The link to the notebook is
https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/Data%20Collection%20with%20Web%20Scrapping.ipynb

Data Wrangling

Exploratory data analysis was conducted to identify the training labels. We tallied the frequency of launches per site and analyzed the number and frequency of orbits. A landing outcome label was generated from the outcome column, and the findings were exported to a CSV file.

```
# Apply # Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```
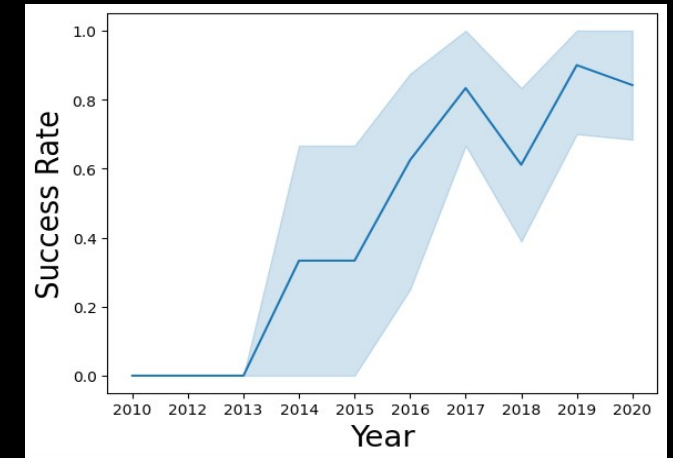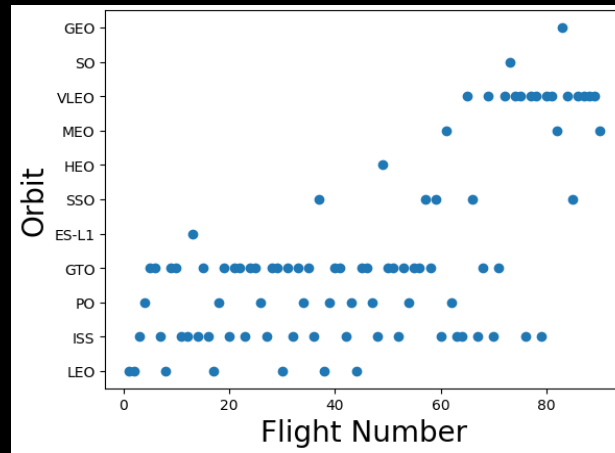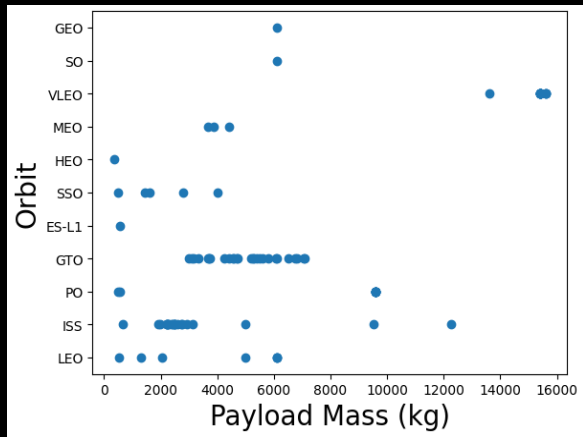
```
[6]

...    LaunchSite
       CCAFS SLC 40    55
       KSC LC 39A      22
       VAFB SLC 4E     13
```

The link to the notebook is https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.

The link to the notebook is
https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

The SpaceX dataset was seamlessly loaded into a database from a Jupyter Notebook. Utilizing SQL for exploratory data analysis (EDA), valuable insights were gleaned from the dataset. SQL queries were developed to discover:

❑ The unique launch sites engaged in space missions.

❑ The overall payload mass carried on NASA (CRS) missions.

❑ The average payload mass delivered by the F9 v1.1 booster version.

❑ The total count of mission outcomes, distinguishing successes from failures.

❑ Specifics on failed drone ship landings, including booster versions and names of launch sites.

The link to the notebook is
https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

We annotated all launch sites on the folium map, incorporating map elements like markers, circles, and lines to indicate the launch outcomes (success or failure) at each location. Launch outcomes were categorized into classes 0 and 1, with 0 representing failure and 1 indicating success. Through the use of color-coded marker clusters, we were able to discern which launch sites boasted higher success rates. Distances from each launch site to nearby features were calculated to address questions such as the proximity of launch sites to railways, highways, coastlines, and the distance maintained from urban areas.
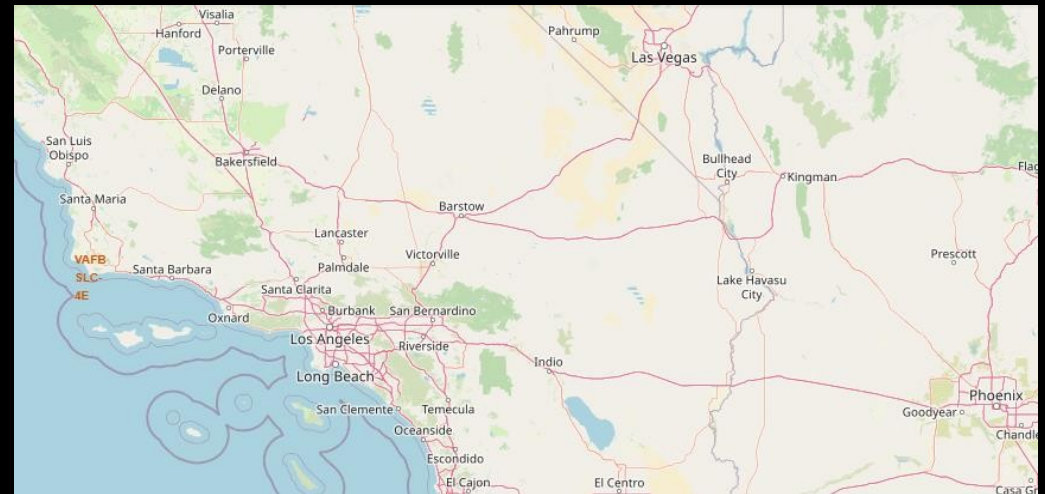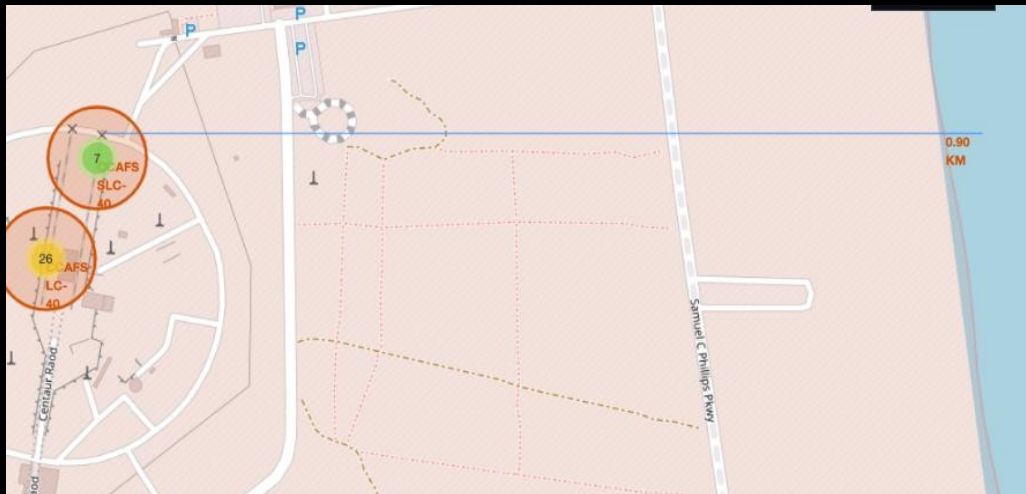
Build a Dashboard with Plotly Dash

An interactive dashboard was created using Plotly Dash, featuring pie charts to visualize the distribution of total launches from various sites. Scatter plots were also designed to explore the association between launch outcomes and payload mass (Kg) across various booster versions.

The link to the notebook is
https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/app.py

Predictive Analysis (Classification)

Data was imported utilizing numpy and pandas, followed by data transformation and division into training and testing sets. Various machine learning models were developed, with hyperparameters optimized through GridSearchCV. The model's performance was evaluated based on accuracy, enhancing it further with feature engineering and tuning of algorithms. The most effective classification model was identified.

The link to the notebook is

https://github.com/Ulfvaldr/Applied_Data_Science_Capstone_SpaceX_IBM/blob/main/Machine%20Learning%20Prediction.ipynb

Results

- ❑ Exploratory data analysis results
- ❑ Interactive analytics demo in screenshots
- ❑ Predictive analysis results
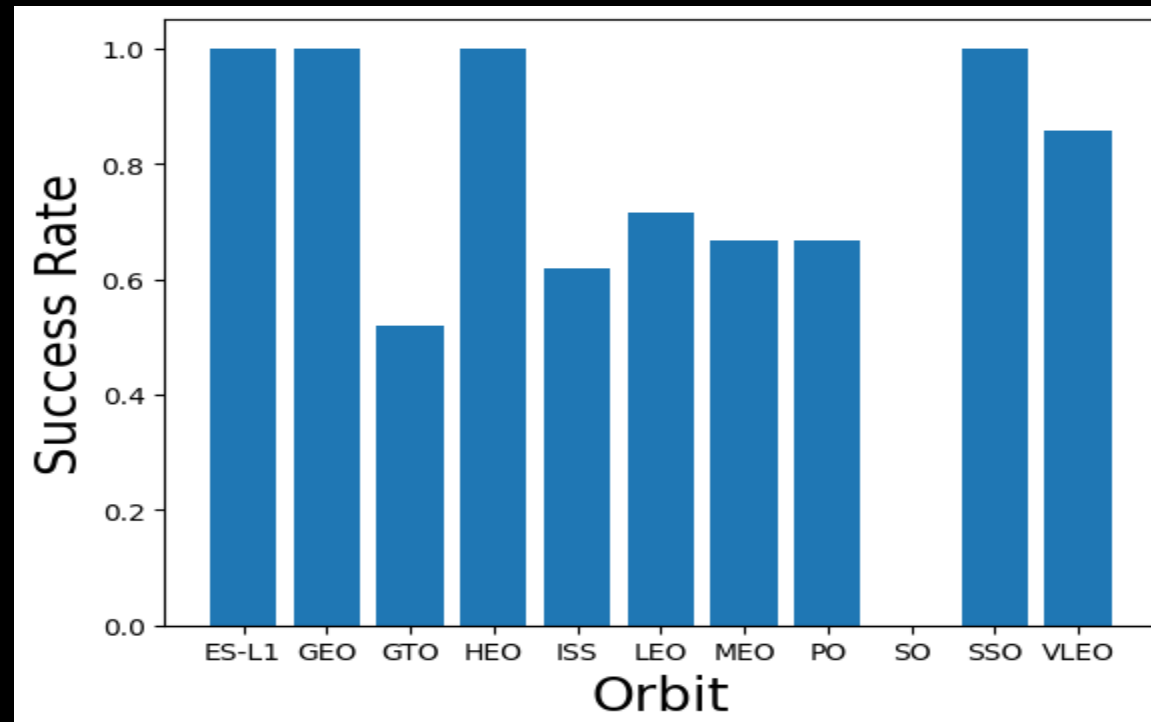
Key Insights from Data Exploration

Section 2

Flight Number vs. Launch Site

❑ From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.
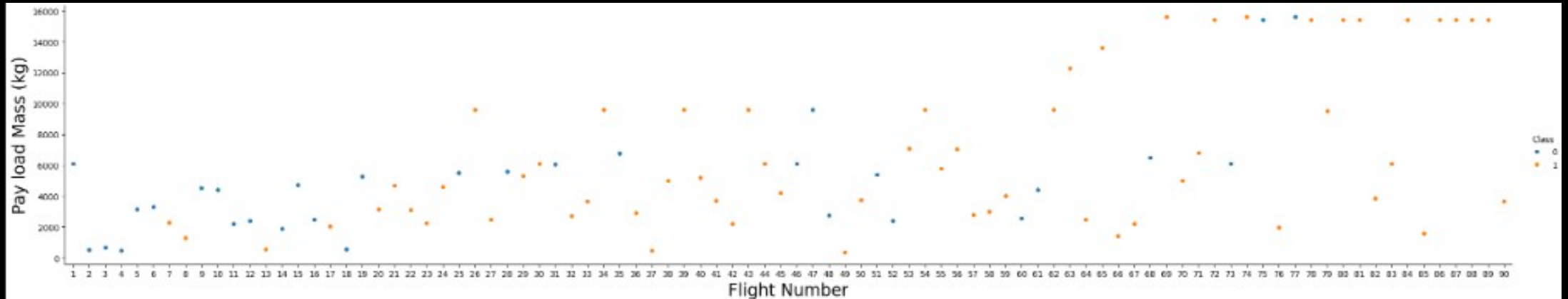
Success Rate vs. Orbit Type

❑ From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.
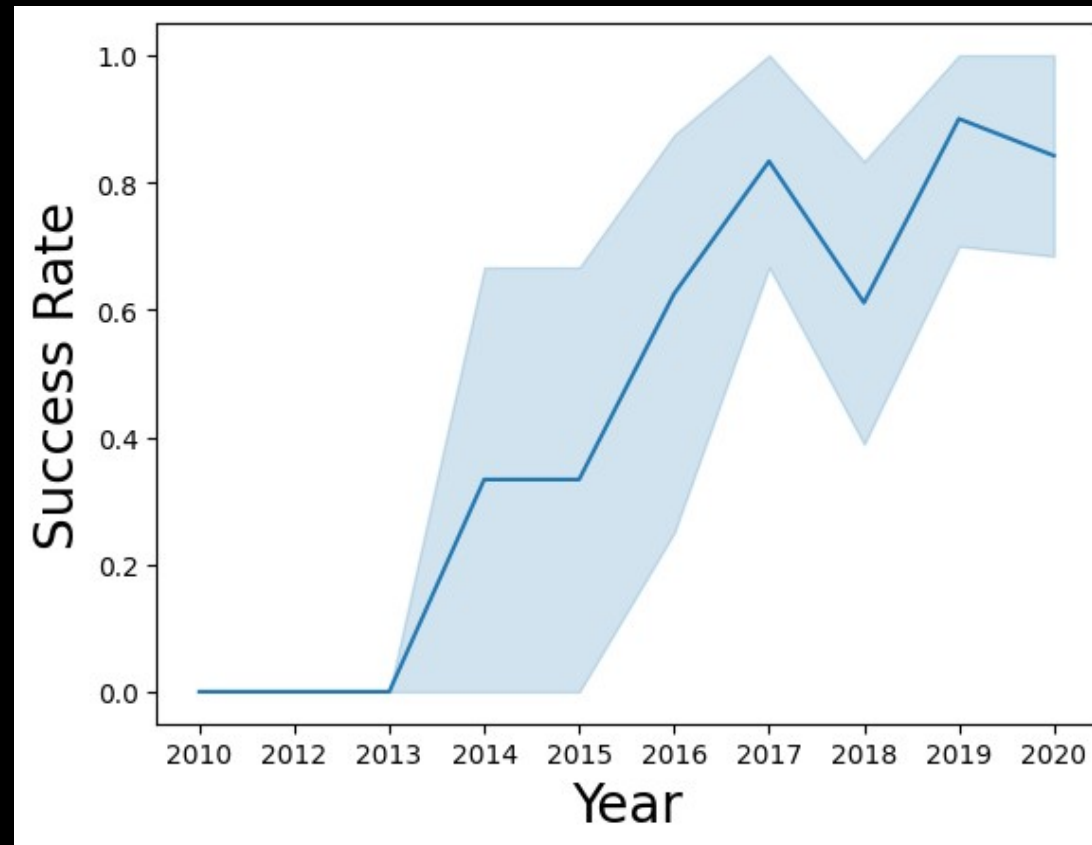
Flight Number vs. Payload

❏ Below, the plot illustrating Flight Number against Orbit type indicates that success in the LEO orbit correlates with flight frequency, while in the GTO orbit, flight number appears unrelated to success.
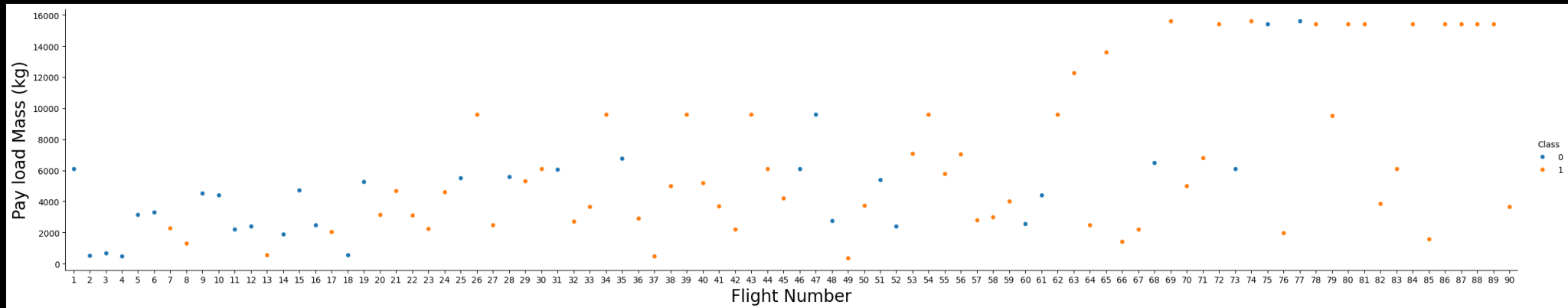
Launch Success Yearly Trend

❑ From the plot, we can observe that success rate since 2013 kept on increasing till 2020.

All Launch Site Names

❏ Now, let's examine the coordinates for each launch site.

```python
# Select relevant sub-columns: `Launch Site`, `Lat(Latitude)`, `Long(Longitude)`, `class`
spacex_df = spacex_df[['Launch Site', 'Lat', 'Long', 'class']]
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
launch_sites_df
```

[4]

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

# Launch Sites and Proximities Analysis

Section 3

Global Map Markers for Launch Sites

SpaceX's launch sites located along the coasts of Florida, Texas and California in the United States.

Launch Site distance to landmarks

This analysis delves into measuring and understanding the spatial relationship between launch sites and nearby significant landmarks, geographical features, and infrastructure.
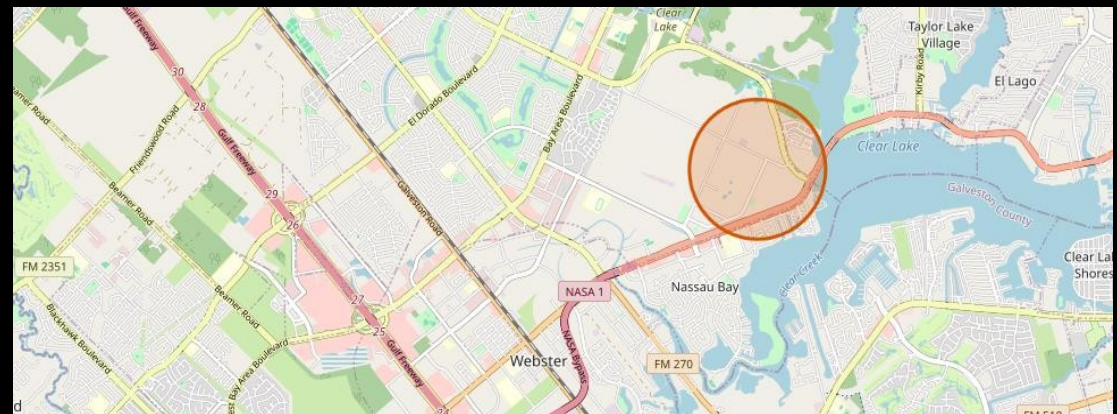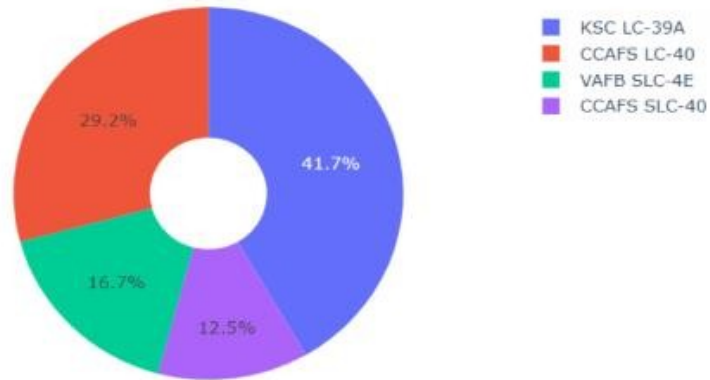
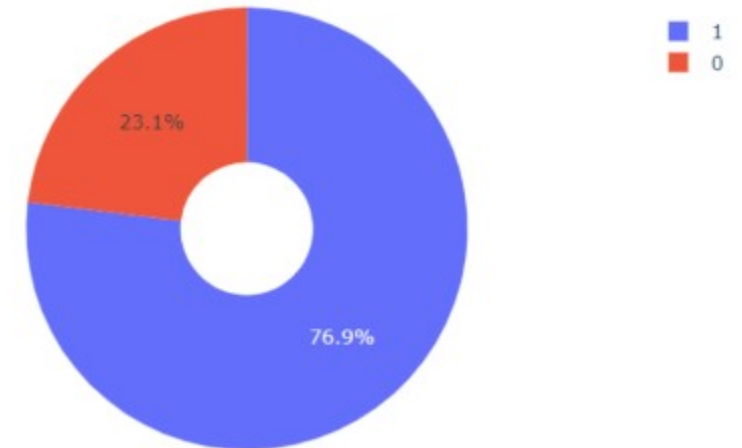Building Dashboard with Plotly Dash

Section 4
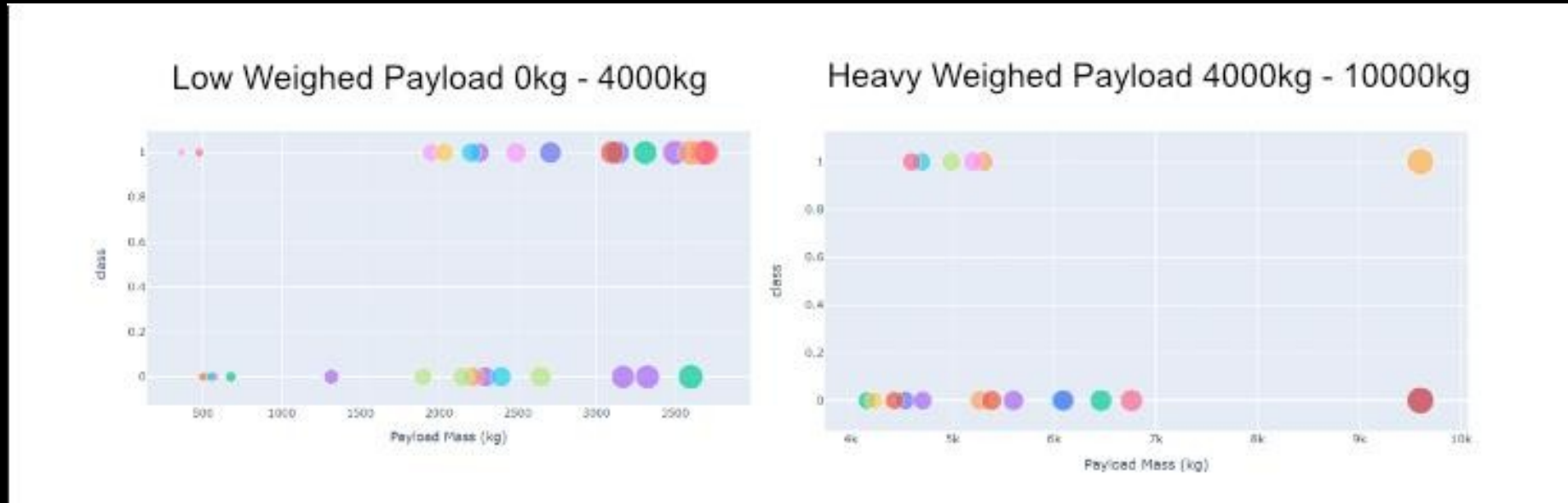
Pie and Scatter Plots: Launch Site Success

KSC-LC-39A had the most successful launches

KSC-LC-39A achieved a 76.9% success and 23.1% failure rate

Payload, Outcome Scatter Plot: Range Selection

The success rates for lighter payloads exceed those of heavier payloads.

Predictive Analysis (Classification)

Highest Accuracy Model

```python
tree = DecisionTreeClassifier()

parameters = {
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
}

tree_cv = GridSearchCV(tree, parameters, cv=10)

tree_cv.fit(X_train, Y_train)
accuracy = tree_cv.score(X_test, Y_test)
print("Accuracy of the DecisionTree on Test Data:", accuracy)
```

```
Accuracy of the DecisionTree on Test Data: 0.8888888888888888
```

```python
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier()

parameters = {
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

model_cv = GridSearchCV(model, parameters, cv=10)

model_cv.fit(X_train, Y_train)

print("Best parameters:", model_cv.best_params_)
best_model = model_cv.best_estimator_
```

```
Best parameters: {'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 5}
```

The decision tree classifier is the model with the highest classification accuracy

Confusion Matrix

The confusion matrix of the decision tree classifier indicates its capability to differentiate among the classes. However, its primary issue lies in the false positives, where unsuccessful landings are incorrectly identified as successful by the classifier.
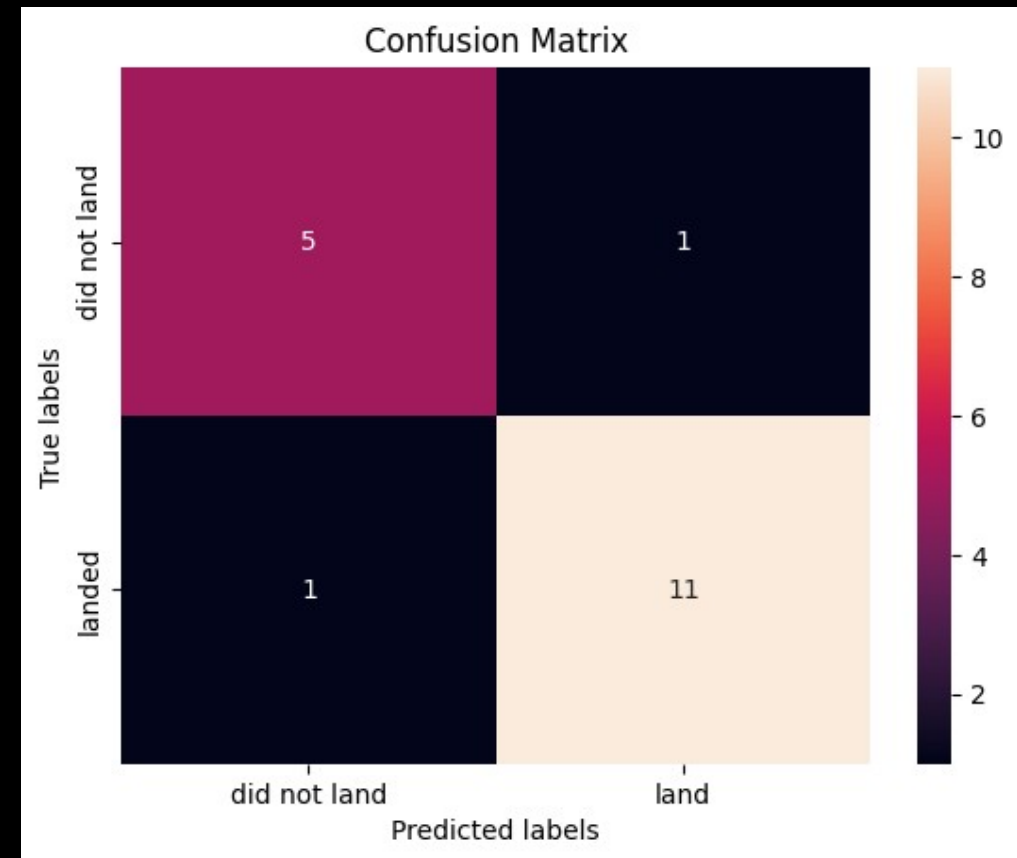
Conclusions

In Conclusion:

❑ A higher number of flights at a launch site correlates with an increased success rate there.
❑ The success rate of launches has been on an upward trend from 2013 to 2020.
❑ The orbits with the highest success rates include ES-L1, GEO, HEO, SSO, and VLEO.
❑ KSC LC-39A boasts the highest number of successful launches among all sites.
❑ The decision tree classifier emerges as the most effective algorithm for this analysis.

Thank you!