IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Author: Om Sahu
Date: May 17, 2023

Outline

Executive Summary

- Summary Of Methodologies:

1. Data Collection: RESTful API and web scraping were used to collect data on Falcon 9 first-stage landings.
2. Data Wrangling: The collected data was cleaned and transformed into a dataframe using **pandas**.
3. Exploratory Data Analysis (EDA): The EDA was performed using **SQL** and data visualization techniques to gain insights into the data.
4. Interactive Visual Analytics and Dashboard: A dashboard was built using **Plotly Dash** to analyze launch records interactively, and an interactive map was built to analyze the launch site proximity using **Folium**.
5. Predictive Analysis: Machine learning algorithms were used to determine if the first stage of Falcon 9 will land successfully. **GridSearchCV** was used to tune hyperparameters of different models, including **Logistic Regression**, **SVM**, **Decision Tree**, and **KNN**.

- Summary Of Results

The results of the analysis show that the Decision Tree model performed the best with an accuracy score of **0.83**. However, all the models achieved similar accuracy scores of **0.83**, indicating that the predictive power of the models may not be significantly different. The dashboard and map visualizations provide a useful way to interactively explore the data and gain insights.

Introduction

- Project Background and Context:
  1. The Falcon 9 rocket launch is a significant event in the space industry. SpaceX, the company behind Falcon 9, has been advertising its rocket launches with a cost of 62 million dollars, which is significantly lower than other providers who charge upwards of 165 million dollars. One of the reasons for this lower cost is that SpaceX can reuse the first stage of the Falcon 9 rocket, which saves a lot of money.
  2. However, predicting whether the Falcon 9 first stage will land successfully is not an easy task. It requires a lot of data and analysis to determine the cost of a launch. Therefore, the goal of this capstone project is to predict whether the Falcon 9 first stage will land successfully and determine the cost of a launch.

- Problems to Answers:
  1. The primary problem that we want to solve is to predict whether the Falcon 9 first stage will land successfully. This information is crucial for determining the cost of a launch, and it can be used if an alternate company wants to bid against SpaceX for a rocket launch.
  2. To achieve this goal, we will use a variety of tools and techniques, including data collection, data wrangling, exploratory data analysis, interactive visual analytics, and predictive analysis using machine learning. We will collect data on Falcon 9 first-stage landings using a RESTful API and web scraping, convert it into a dataframe, perform data wrangling, and use exploratory data analysis to visualize the data. We will build a dashboard and an interactive map to analyze the launch site proximity, and use machine learning algorithms such as SVM, Classification Trees, and Logistic Regression to determine if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using a RESTful API and web scraping
- Perform data wrangling
  - Data was converted into a DataFrame and then underwent data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The classification models were built using training data, and their performance was evaluated using test data. The best hyperparameters were selected for each model using tuning techniques

Data Collection

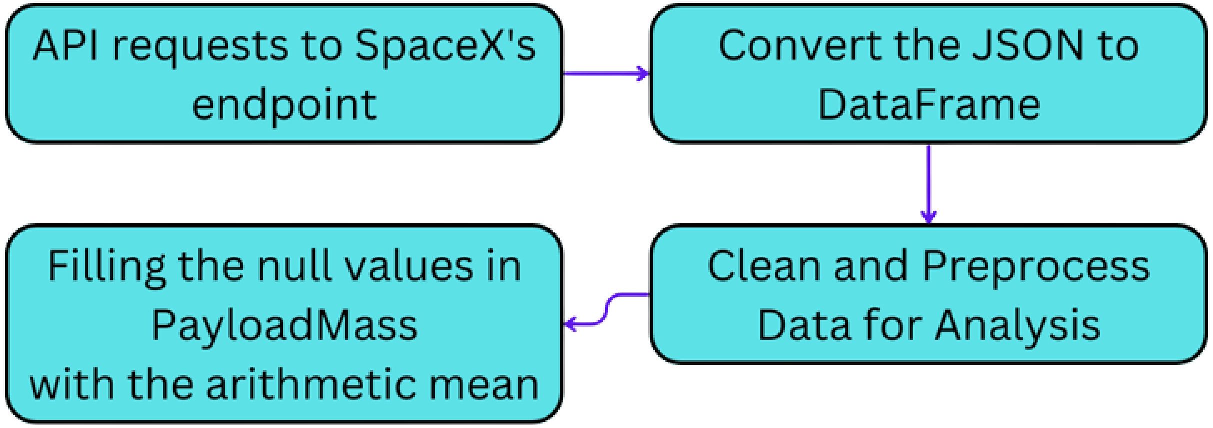
To visualize our data collection process, we can use the following flowchart:

1. Use SpaceX API to collect data on Falcon 9 launches, including mission name, launch date, and other relevant information
2. We imported Python libraries like **BeautifulSoup** and **requests** in a new notebook
3. Scrape data on each launch's first-stage landing from **wikipedia**
4. Merge data from SpaceX API and web scraping into a single dataset
5. Perform data cleaning and wrangling on merged dataset

Overall, our data collection process allowed us to gather comprehensive information on Falcon 9 launches and their corresponding first-stage landing outcomes, which was critical for our predictive analysis.

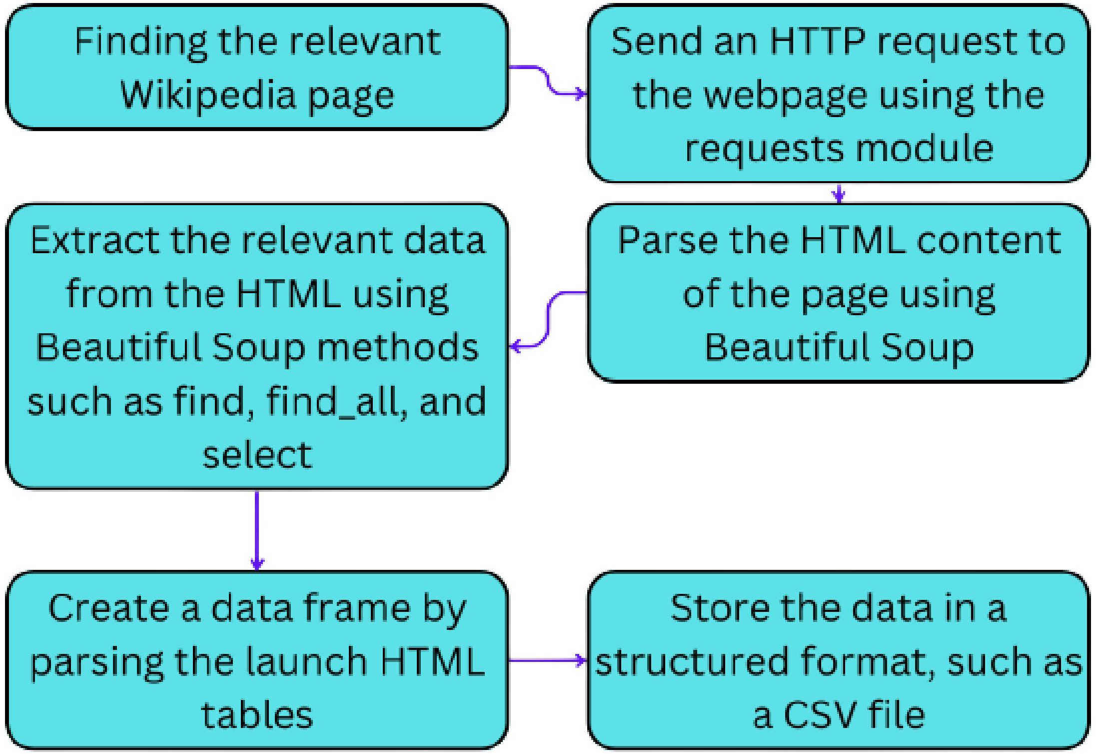
Data Collection – SpaceX API

- For our data collection process, we used SpaceX REST calls to retrieve data on Falcon 9 launches. This involved making API requests to SpaceX's servers to get information on the launch date, launch site, mission outcome, and other relevant details. We did not require an API key for this process, as the data was publicly available.
- [Notebook’s GitHub URL]



Data Collection - Scraping

- Our web scraping process involved collecting data on Falcon 9 launches from various websites, including Wikipedia. To do this, we used Python's Requests BeautifulSoup library to extract relevant information from HTML pages.
- [Notebook's GitHub URL]



Data Wrangling

- The necessary columns were selected from the original dataset.
- The missing values were handled by filling them with appropriate values.
- The data was grouped by launch site to calculate the number of launches on each site.
- The data was grouped by orbit to calculate the number and occurrence of each orbit.
- The data was grouped by orbit type and mission outcome to calculate the number and occurrence of mission outcomes per orbit type.
- A landing outcome label was created by combining the landing type and outcome columns.
- The landing outcome label was used to categorize the landing outcomes into successful and unsuccessful categories.
- The cleaned and processed data was stored in a new dataframe for further analysis.

[Notebook's GitHub URL]

EDA with Data Visualization

1. Scatter plot of FlightNumber vs. PayloadMass with outcome overlay to observe how the flight number and payload mass affect the launch outcome.
2. Bar chart of success rates for each launch site to compare the success rates of different launch sites.
3. Scatter plot of PayloadMass vs. Launch Site to observe the relationship between payload mass and launch site.
4. Bar chart of success rates for each orbit type to compare the success rates of different orbits.
5. Scatter plot of FlightNumber vs. Orbit to observe the relationship between flight number and orbit type.
6. Scatter plot of PayloadMass vs. Orbit to observe the relationship between payload mass and orbit type.

- These charts were used to understand the relationships between different variables and the launch outcome, identify patterns in the data, and find insights that could help improve the launch success rate. For example, the charts helped identify that the success rate of launches depends on the launch site, payload mass, and orbit type. The insights gained from these charts can help optimize the launch process and increase the success rate.Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- Notebook's GitHub URL

EDA with SQL

- Display unique launch sites
- Display 5 records of launch sites starting with 'CCA'
- Display total payload mass of NASA (CRS) boosters launched
- Display average payload mass of booster version F9 v1.1
- Display date of first successful ground pad landing outcome
- Display names of boosters with success in drone ship and payload mass between 4000 and 6000
- Display total number of successful and failed mission outcomes
- Display names of booster versions with maximum payload mass using subquery
- Display month names, failure landing outcomes in drone ship, booster versions, and launch sites for months in 2015
- Rank count of successful landing outcomes between 04-06-2010 and 20-03-2017 in descending order.
- Notebook’s GitHub URL

Build an Interactive Map with Folium

- In this lab, the following map objects were created and added to a folium map:
- folium.Circle and folium.Marker for each launch site on the site map to mark all launch sites
- folium.Marker for each launch result in the spacex_df data frame to mark the success/failed launches for each site
- MousePosition to get coordinate for a mouse over a point on the map
- PolyLine between a launch site and the selected coastline point to calculate the distance between them
- Lines between a launch site and its closest city, railway, highway, etc. to explore its proximity
- These map objects were added to the map to visualize the launch sites and their surroundings, as well as to calculate the distances between them and their proximities. The markers were used to identify the launch sites and the success/failed launches, while the circles were used to highlight the areas around the launch sites. The PolyLine and Lines were used to draw lines between the launch sites and their closest points of interest to calculate the distances. The MousePosition was used to find the coordinates of points of interest on the map.
- Notebook's GitHub URL

Build a Dashboard with Plotly Dash

- The dashboard has three plots/graphs and two interactions.
    1. A dropdown list is added to enable Launch Site selection, with a default select value for ALL sites.
    2. A pie chart is added to show the total successful launches count for all sites. If a specific launch site was selected, the chart shows the Success vs. Failed counts for the site.
    3. A slider is added to select payload range. The range of payload is set from 0 to 10000 kg with a step of 1000 kg.
    4. A scatter chart is added to show the correlation between payload and launch success.
- Two interactions are added:
    1. A callback function is added for **site-dropdown** as input and **success-pie-chart** as output.
    2. A callback function is added for **site-dropdown** and **payload-slider** as inputs and **success-payload-scatter-chart** as output.
- The dropdown list and slider are added to enable users to filter data based on Launch Site and Payload Mass (kg). The pie chart shows the total success launch count for all sites, and the scatter chart shows the correlation between payload and launch success. These plots and interactions are added to provide a comprehensive overview of the SpaceX Launch Records data.
- Notebook's GitHub URL – contains the screenshot of the dashboard in pdf file

Predictive Analysis (Classification)

- To build, evaluate, improve, and find the best performing classification model, we started by performing exploratory data analysis and determining training labels. Then, we created a column for the class and standardized the data. Next, we split the data into training and test sets. We used **GridSearchCV** to find the best hyperparameters for four different models: **logistic regression**, **support vector machine**, **decision tree classifier**, and **k-nearest neighbors**. Finally, we calculated the accuracy of each model on the test data and determined which one performed the best. The flowchart for this process can be represented as follows:
- Data -> EDA and training labels -> Class column creation -> Standardization -> Splitting -> GridSearchCV for hyperparameters -> Accuracy calculation -> Model performance comparison -> Best performing model
- Notebook's GitHub URL

Results

Exploratory Data Analysis Results

- The exploratory data analysis revealed several interesting patterns in the data. Firstly, the FlightNumber and PayloadMass variables appeared to be important factors in determining the success of the rocket launches. As the flight number increased, the success rate of the first stage landing also increased. However, as the payload mass increased, the success rate of the first stage landing decreased.

- The analysis also revealed that different launch sites had different success rates, with CCAFS LC-40 having a success rate of 60%, while KSC LC-39A and VAFB SLC 4E had a success rate of 77%. Further analysis of each launch site showed that the VAFB-SLC launch site had no rockets launched for heavy payload masses (greater than 10000).

- The analysis also showed that there was a relationship between the success rate and the orbit type, with LEO orbit having a positive relationship between success and flight number, while GTO orbit showed no clear relationship. The Payload vs. Orbit scatter point charts showed that heavy payloads had a higher success rate for Polar, LEO, and ISS orbits, while the success rate for GTO orbits was unclear, with both positive and negative landing rates present.

Predictive Analysis Results

- Based on the provided results, it seems that all four methods perform similarly well on the test data, with an accuracy of 0.833. However, the DecisionTreeClassifier method had the highest accuracy on the training data, with a score of 0.875, which may suggest that it has better generalization performance than the other methods.

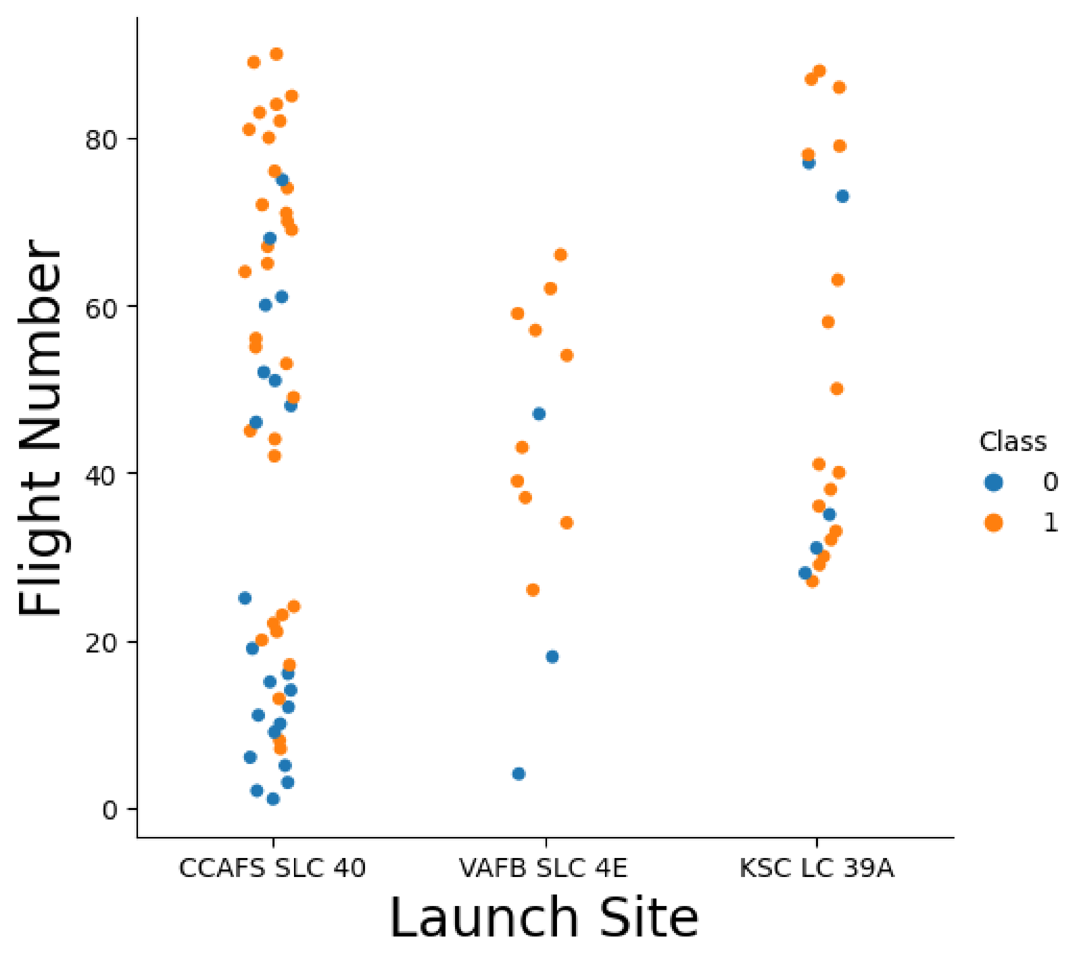
- Therefore, based on the given results, the DecisionTreeClassifier method appears to perform the best. However, it is important to note that this conclusion is based solely on the provided results and may change depending on the specific data set and problem at hand.

Section 2

Insights drawn from EDA

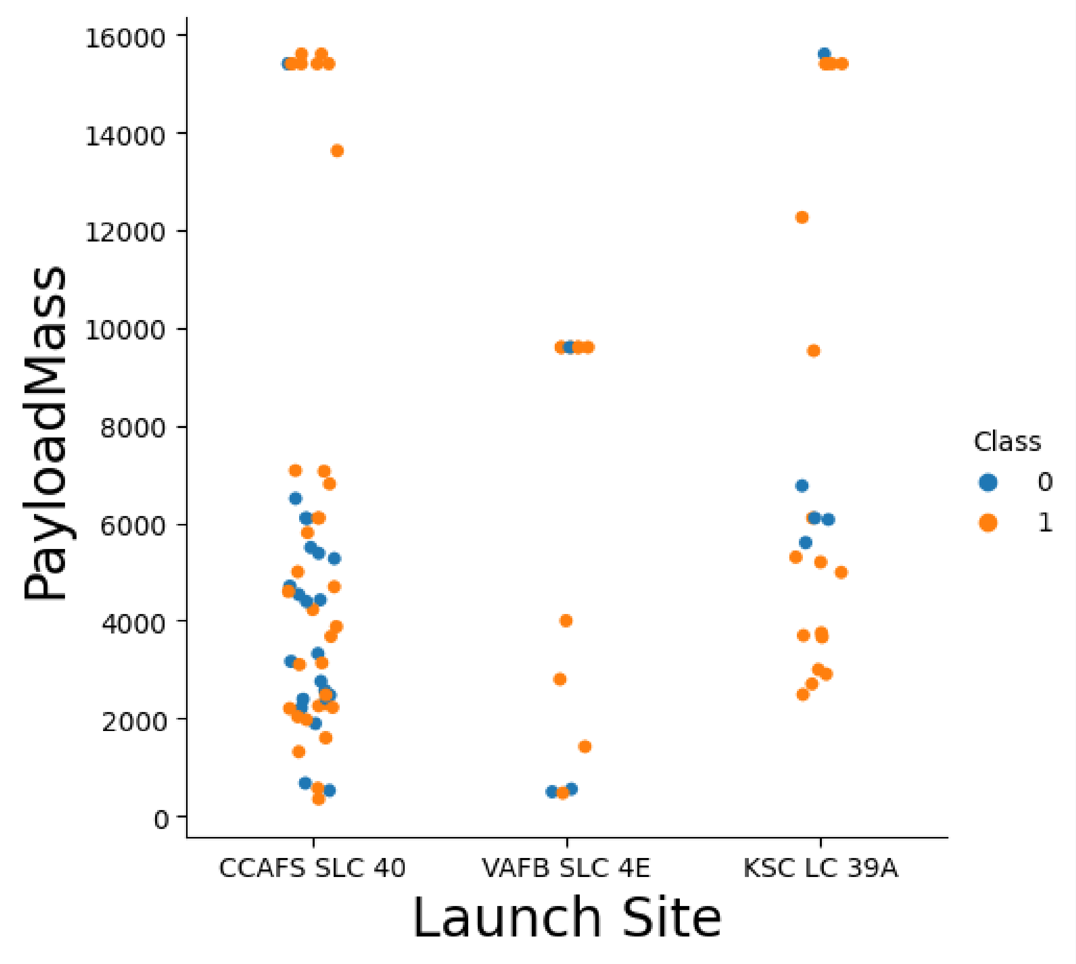
Flight Number vs. Launch Site

From this scatter plot, we can analyze the relationship between the number of flights launched from a particular site and the success rate of those launches. Based on the scatter plot, we can make the observation that as the number of flights launched from a site increases, the success rate also tends to increase. This information can be valuable in analyzing the performance of different launch sites and can inform decisions about where to focus resources and investment in the future.



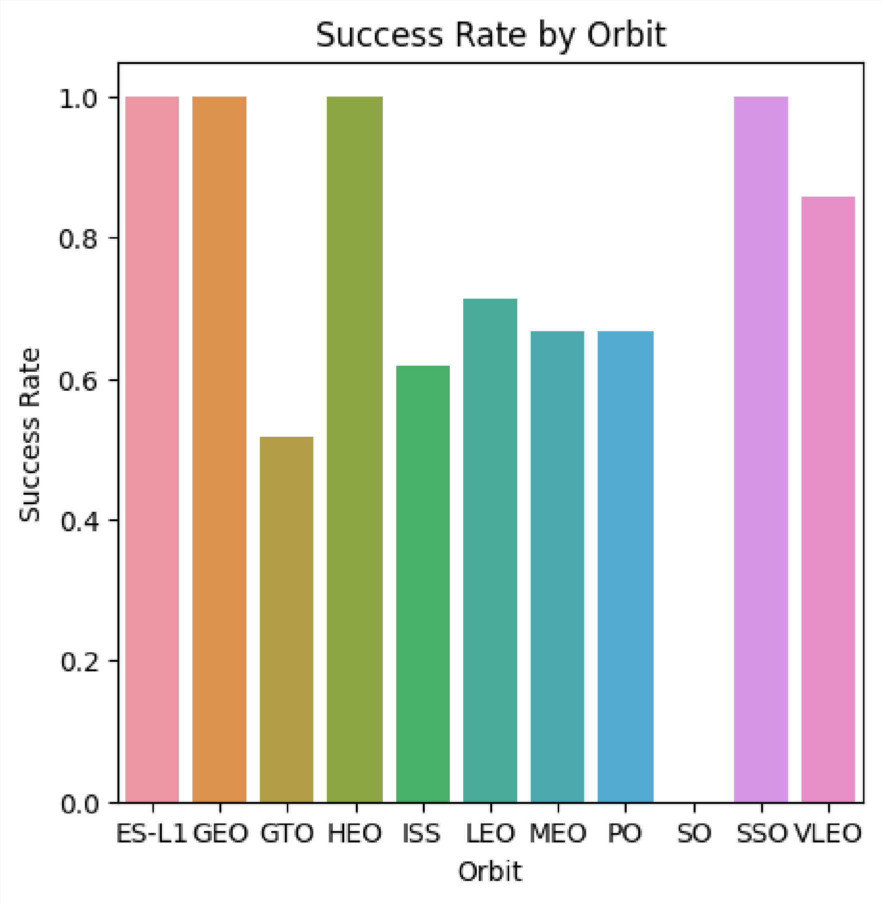
Payload vs. Launch Site

- From this scatter plot, we can analyze the relationship between the payload mass of the rocket and the success rate of launches from each site. We can observe that for the VAFB-SLC launch site, there are no rockets launched for heavy payload mass, indicating a limitation in capabilities at that site.
- However, overall, there appears to be a weak correlation between payload mass and launch site. Therefore, it may not be a useful metric for making decisions regarding launch site selection or investment in launch site infrastructure.



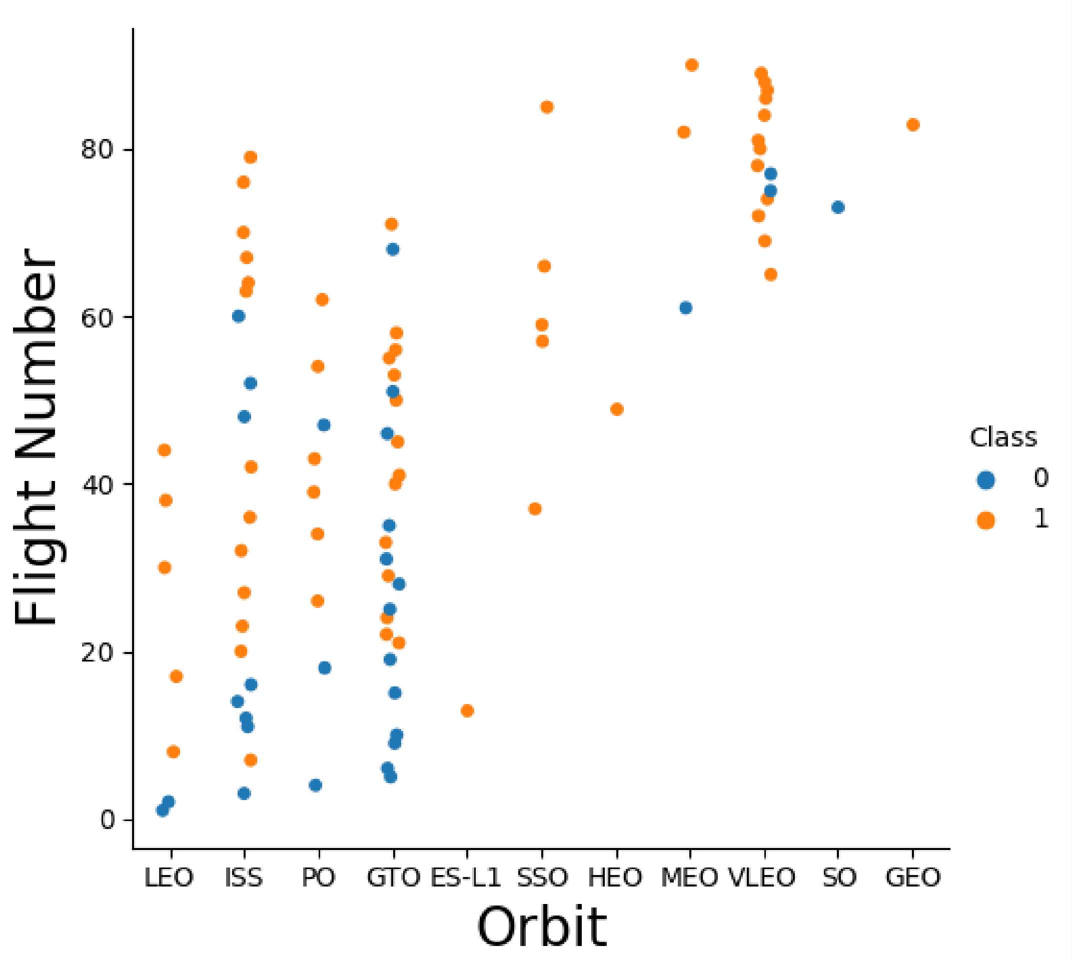
Success Rate vs. Orbit Type

Based on the bar chart, we can observe which orbit types have the highest success rates. The orbit types SSO (Sun-synchronous orbit), HEO (Highly elliptical orbit), GEO (Geostationary orbit), and ES-L1 (Earth-Sun Lagrange point 1) appear to have the highest success rates, while other orbit types may have lower success rates.



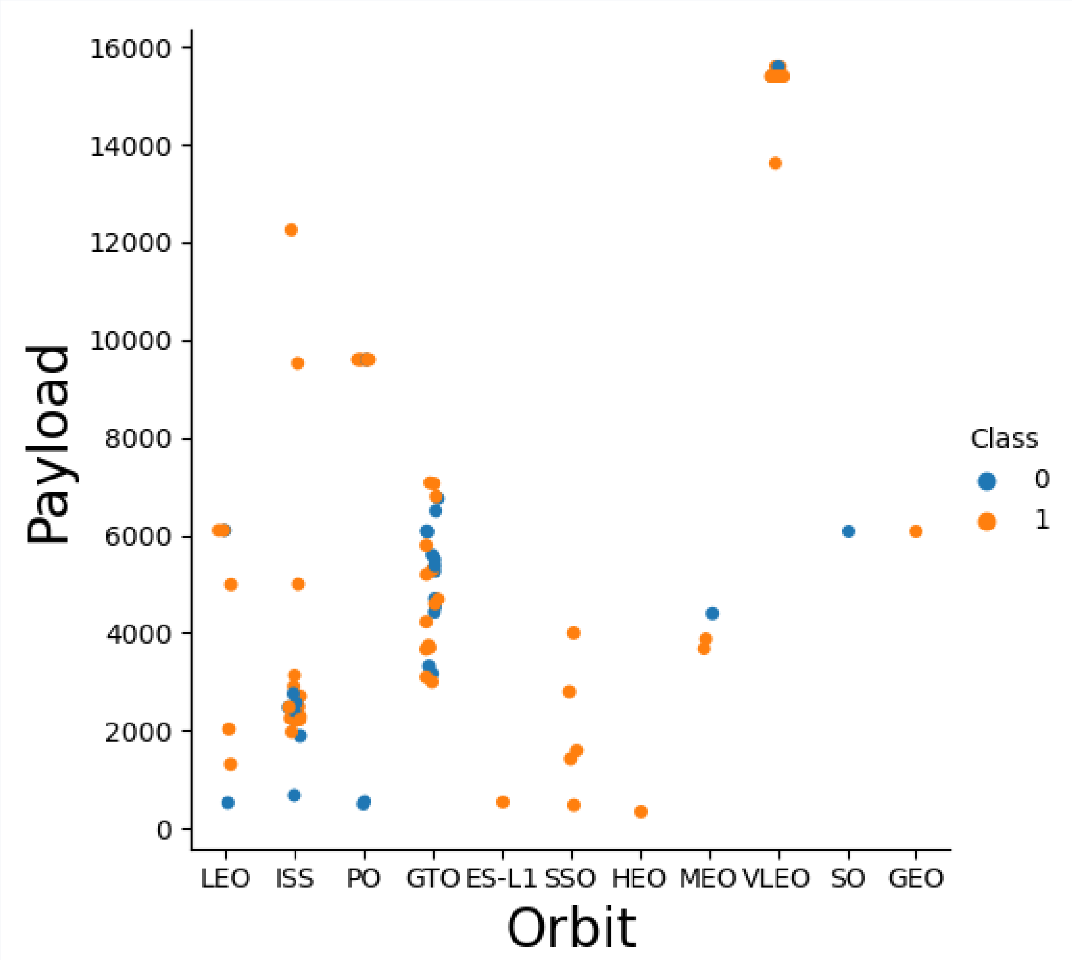
Flight Number vs. Orbit Type

- By looking at the scatter plot, we can see that there are several clusters of launches for certain orbit types, such as LEO and GTO. Additionally, we can see that some orbit types, like SSO and HEO, have a higher proportion of successful launches compared to others, like PO and MEO.
- We can also identify any outliers or unusual patterns, such as the single launch in the SO orbit type, which appears to be a failure.



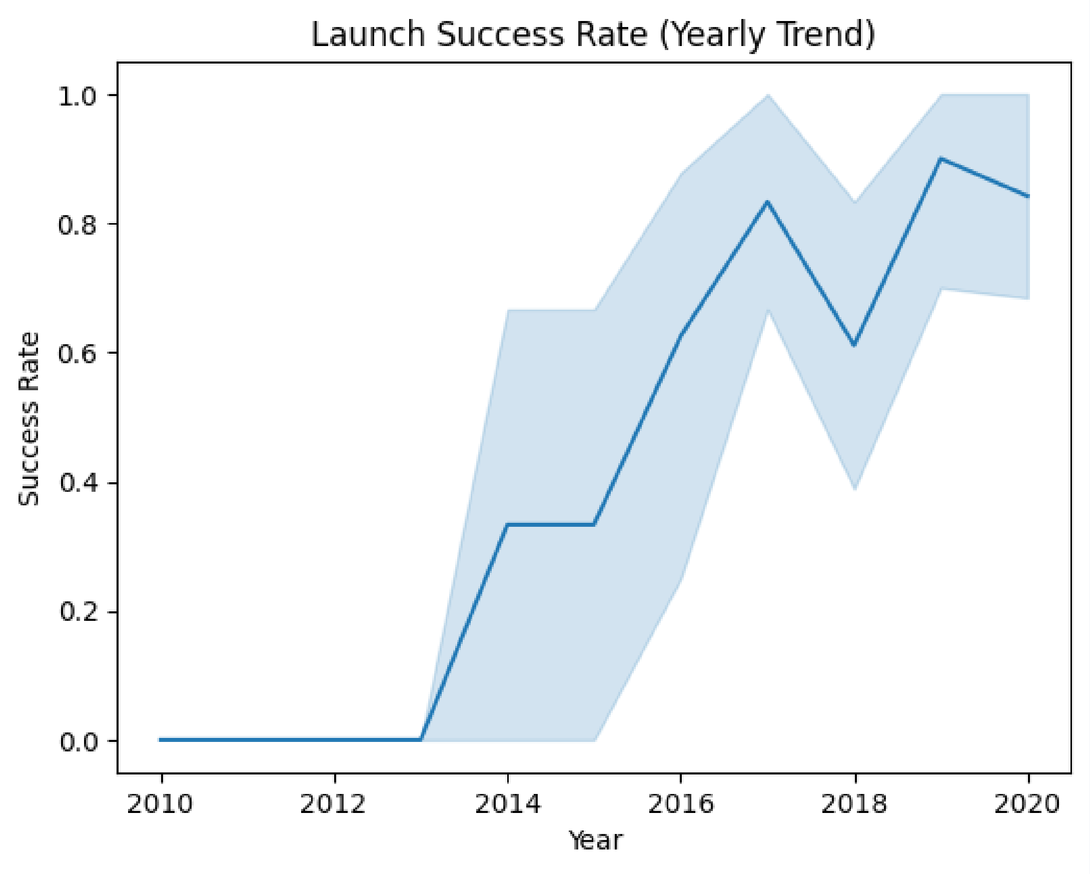
Payload vs. Orbit Type

From the scatter plot, it appears that there is a relationship between the payload mass and the success of a mission based on the orbit type. Specifically, the successful landing rates appear to be higher for Payload Masses greater than 5,000 kg for orbits LEO, PO, and ISS. On the other hand, payloads greater than 6,000 kg seem to be less successful for orbits such as GTO. Overall, it seems that the higher the payload mass, the lower the success rate of landing, except for certain orbits such as LEO, PO, and ISS.



Launch Success Yearly Trend

The success rate appears to have steadily increased since 2013, with a notable jump in 2014 and another in 2016. It then appears to have plateaued somewhat between 2017 and 2019 before seeing another increase in 2020.



All Launch Site Names

The query **SELECT DISTINCT Launch_Site FROM SPACEXTBL** was used to retrieve the unique launch sites from the SPACEXTBL table. The result of the query shows that there are four unique launch sites in the table: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Display the names of the unique launch sites in the space mission

In [7]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

Out[7]:

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

Launch Site Names Begin with 'CCA'

This SQL query uses the LIKE operator and a wildcard % to find launch sites that begin with the letters "CCA". The DISTINCT keyword ensures that only unique launch sites are returned, and the LIMIT keyword is used to limit the output to 5 records. The resulting output shows the launch sites that meet the specified criteria, which are CCAFS LC-40 and CCAFS SLC-40.

In [8]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Out[8]:

Launch_Site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass

The query calculates the total payload mass carried by boosters for NASA's Commercial Resupply Services (CRS) program using the SUM() function and filtering for the Customer column with the value 'NASA (CRS)'. The result is 45596 kg.

Display the total payload mass carried by boosters launched by NASA (CRS)

In [9]:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer=='NASA (CRS)';
```

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

Out[9]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The query calculates the average payload mass carried by boosters of the version "F9 v1.1" using the AVG() function and filtering for the Booster_Version column with the value starting with 'F9 v1.1'. The result is approximately 2534.67 kg.

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

In [10]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Out[10]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

The query retrieves the date of the first successful landing outcome on a ground pad by filtering the records where the Mission_Outcome is 'Success' and the Date is equal to the minimum date in the dataset. The result is '01-03-2013', indicating the date when the first successful landing on a ground pad was achieved.

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

Hint:Use min function

In [11]:

```
%sql SELECT Date FROM SPACEXTBL WHERE Mission_Outcome=='Success' AND Date==(SELECT MIN(Date) FROM SPACEXTBL);
```

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

Out[11]:

Date
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

The query selects the Booster_Version from the dataset where the Landing_Outcome is 'Success (drone ship)' and the PAYLOAD_MASS__KG_ is between 4000 and 6000. The result of the query will be the names of boosters that successfully landed on a drone ship and had a payload mass greater than 4000 but less than 6000.

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

```
In [12]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome"=='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Out[12]:

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 query selects the Mission_Outcome column from the dataset and counts the occurrences of each unique mission outcome. The result of the query provides the total number of successful and failure mission outcomes. The count is grouped by the mission outcome, and the result shows the mission outcome along with the corresponding count.

List the total number of successful and failure mission outcomes

```
In [13]: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome

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

Out[13]:

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

Boosters Carried Maximum Payload

The query selects the Booster_Version column from the dataset where the PAYLOAD_MASS__KG_ is equal to the maximum payload mass achieved. It uses a subquery to find the maximum payload mass from the same dataset. The result of the query provides the names of the booster versions that have carried the maximum payload mass.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [14]:
```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Out[14]:

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 query selects the month names (using substr(Date, 4, 2)), booster versions, and launch site names from the dataset for the records where the landing outcome is a failed landing on a drone ship, and the year is 2015 (using substr(Date,7,4)='2015'). The result displays the month names, corresponding booster versions, and launch site names for the failed landing outcomes on a drone ship in the year 2015.

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [15]:
```
%sql SELECT substr(Date, 4, 2), Booster_Version, Launch_Site FROM SPACEXTBL WHERE "Landing _Outcome" == 'Failure (drone ship)' AND substr(Date,7,4)='20
```

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

Out[15]:

substr(Date, 4, 2)	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

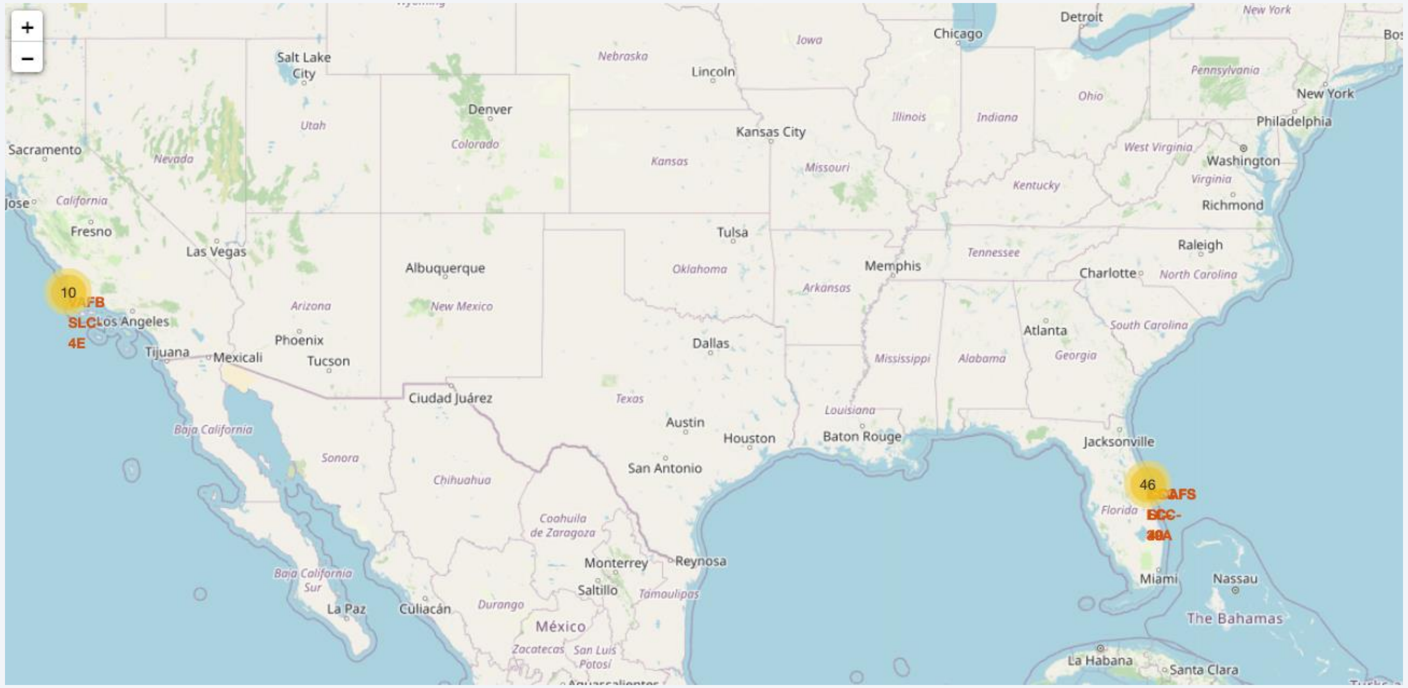
This query selects the landing outcome and counts the occurrences of each landing outcome within the specified date range. It groups the results by the landing outcome and sorts them in descending order based on the count. The result will display the landing outcomes along with their corresponding counts, with the highest count appearing first.

Section 3

Launch Sites Proximities Analysis

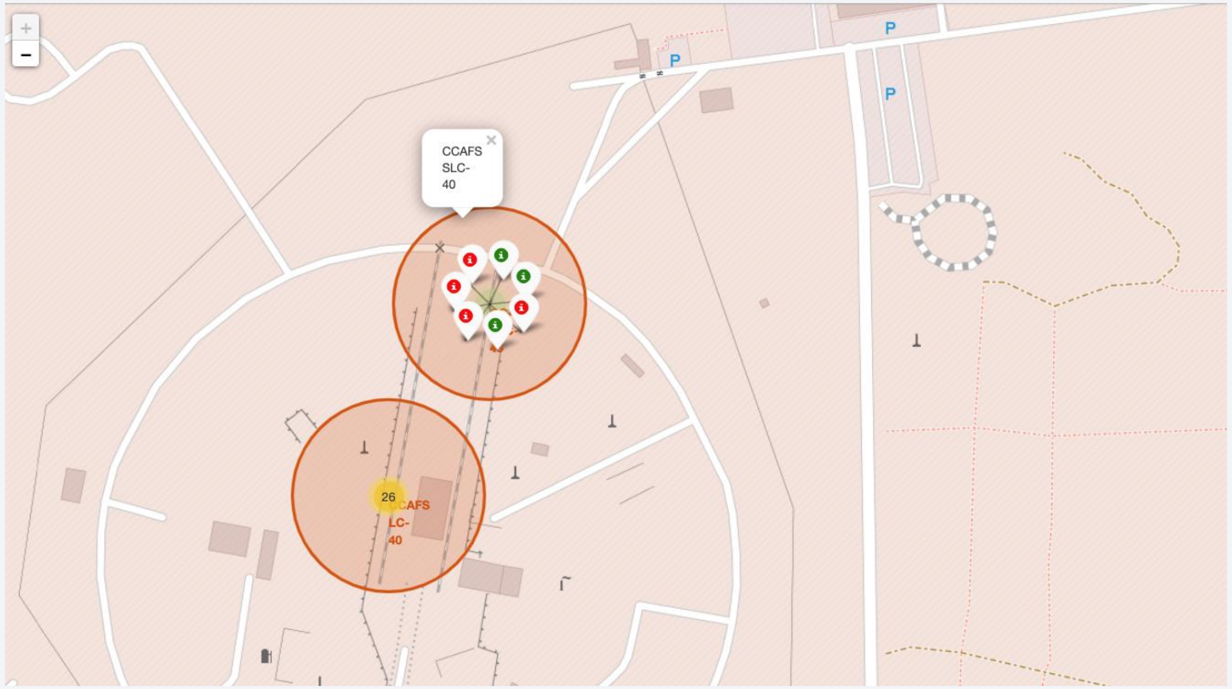
SpaceX Launch Sites around the USA

Launch sites are strategically located near the Equator to leverage the Earth's rotation. The Equator provides a higher initial speed advantage for spacecraft. Objects at the Equator are already moving at a significant speed due to the Earth's rotation. When a spacecraft is launched from the Equator, it benefits from this higher initial speed, enabling it to achieve and maintain the necessary velocity for successful orbit. This proximity to the Equator offers a valuable advantage, reducing fuel requirements and enhancing the efficiency of space missions.



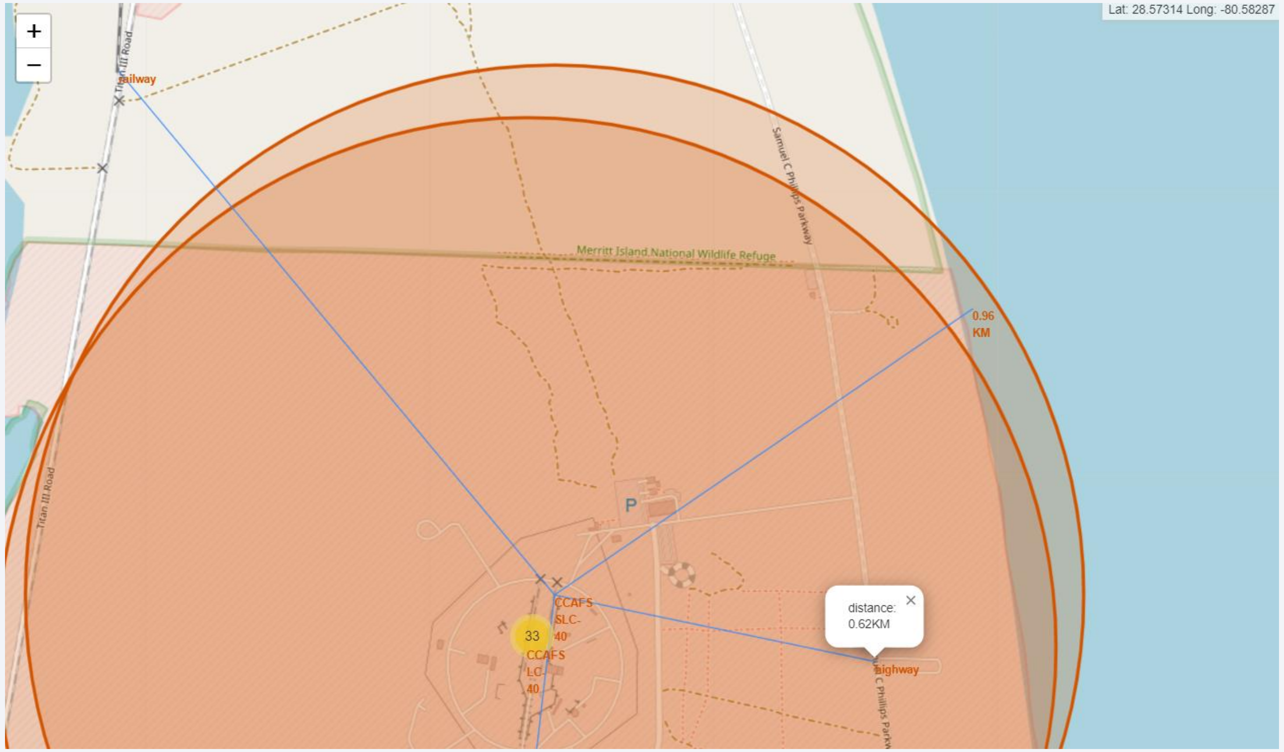
SpaceX Launch Success by Location

The color-labeled markers provide a clear visual distinction between successful and failed launches. The predominance of green markers indicates high success rates, while red markers represent launch failures. Notably, the launch site KSC LC-39A stands out with a consistently high success rate, as evidenced by the abundance of green markers. This information helps us quickly identify successful launch sites and highlights the impressive performance of KSC LC-39A.



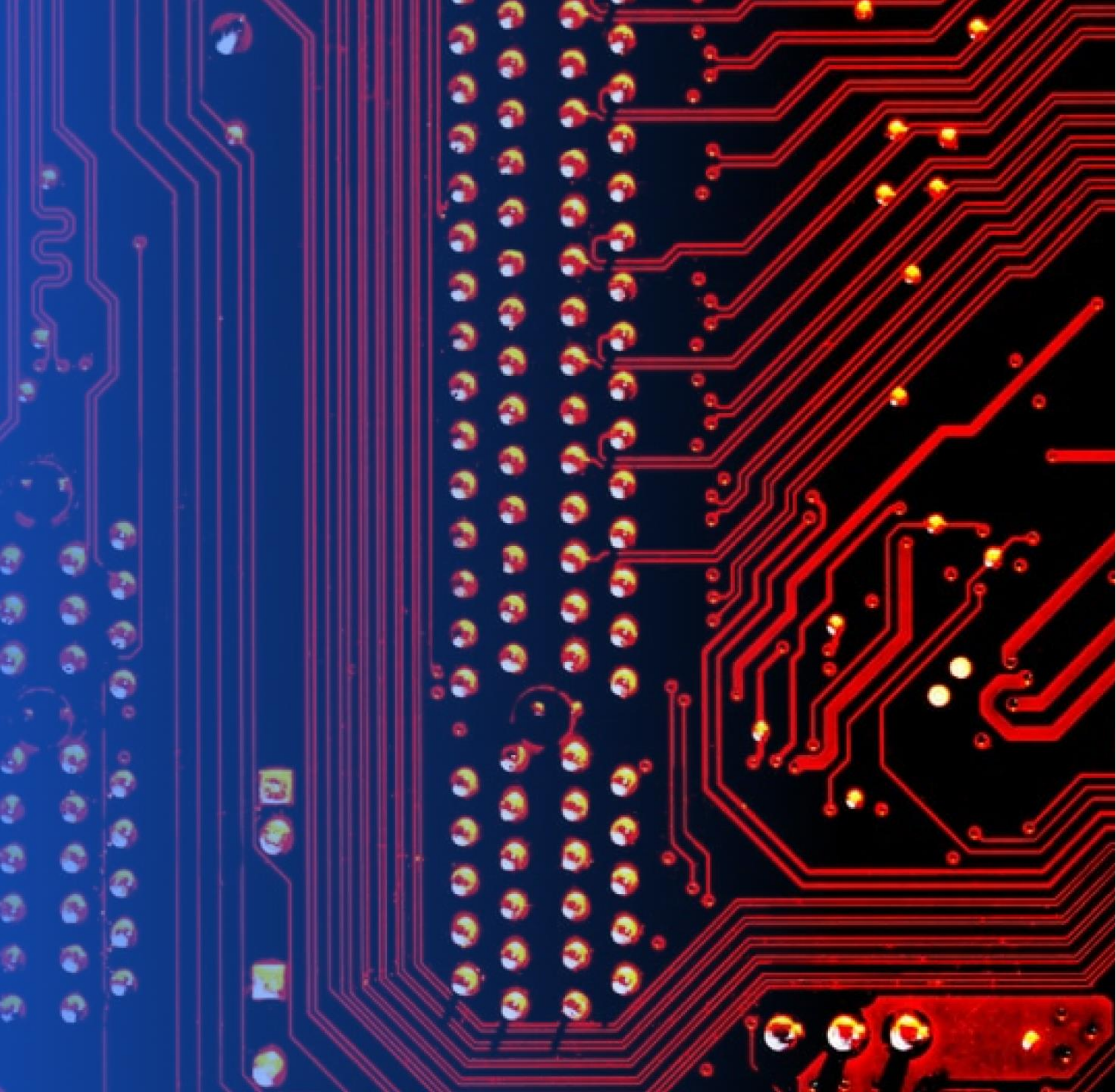
SpaceX's KSC LC-39A Proximities

From the analysis of the launch site KSC LC-39A, it is evident that it has strategic proximity to various transportation modes. It is located approximately 1.29 km from the nearest railway, 0.62 km from the nearest highway, and 0.96 km from the coastline. These close connections to key transportation routes provide logistical advantages for efficient operations and accessibility.



Section 4

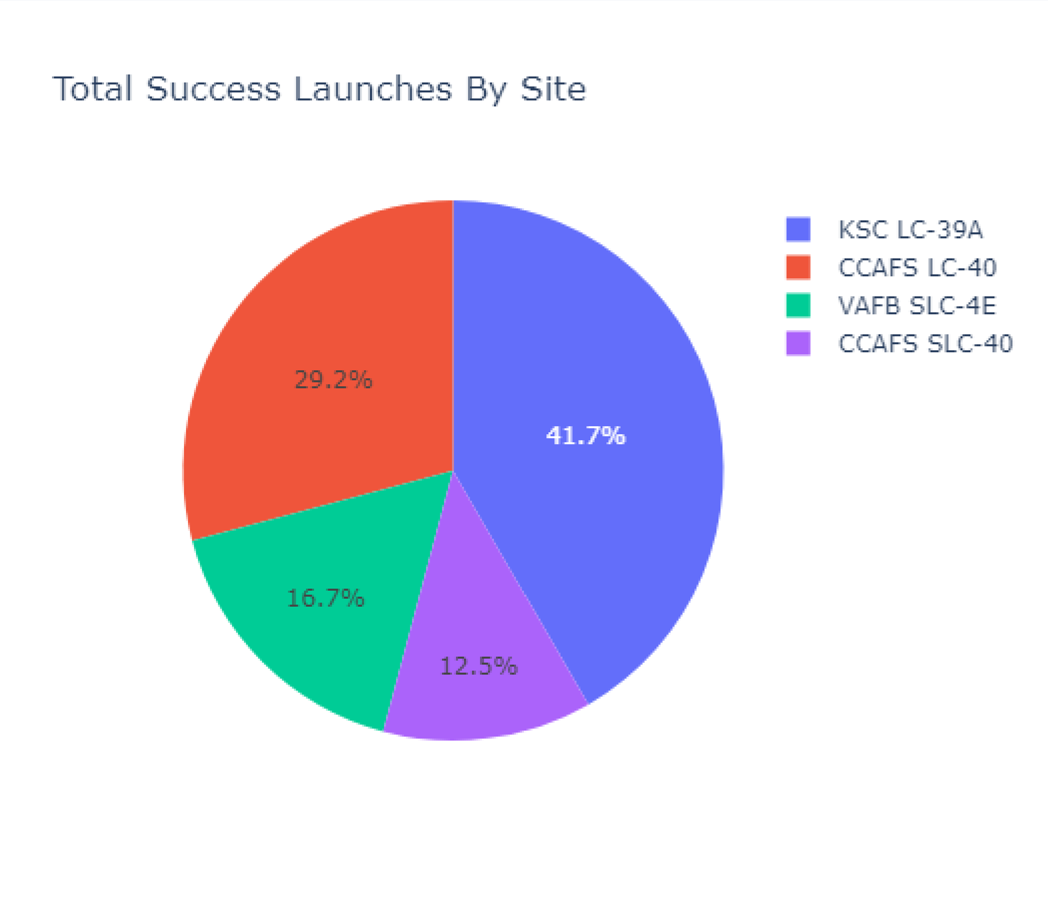
Build a Dashboard with Plotly Dash



Launch Site Success Percentage Distribution

- The pie chart showcases the success percentages achieved by various launch sites:

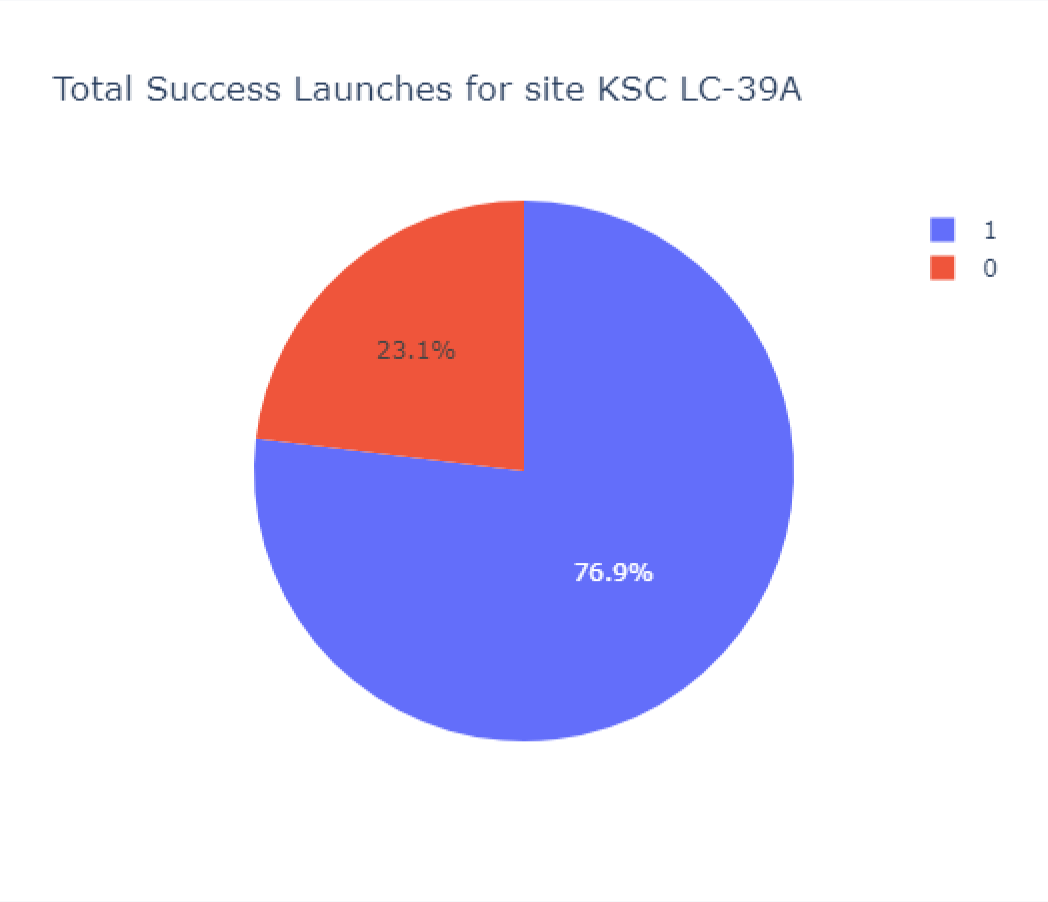
- Cape Canaveral Air Force Station (CCAFS) LC-40: **29.2%**
- Kennedy Space Center (KSC) LC-39A: **41.7%**
- Vandenberg Air Force Base (VAFB) SLC-4E: **16.7%**
- Cape Canaveral Air Force Station (CCAFS) SLC-40: **12.5%**



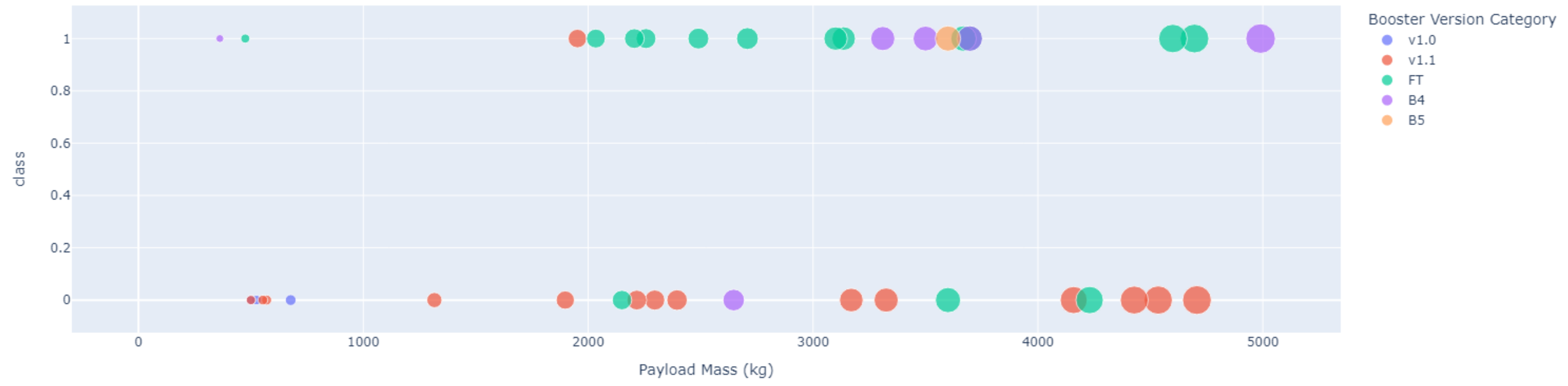
Top Performing Launch Site based on Launch Success Ratio

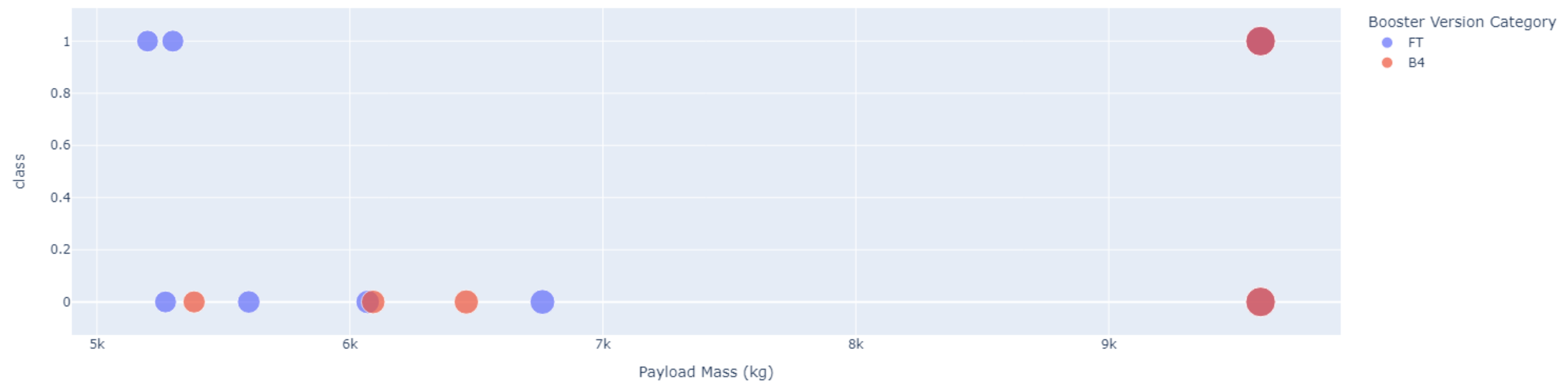
- Success: 76.9% - This represents the percentage of launches that have been successful. It indicates a significantly high success rate, implying that the majority of launches have been completed successfully.

- Fail: 23.1% - This represents the percentage of launches that have ended in failure. While this percentage is relatively smaller compared to the success rate, it signifies that there have been instances where launches did not achieve the desired objectives.



Payload vs. Launch Outcome Scatter Plot Across All Sites





Scatter plot reveals higher success rate for lower payload mass and lower success rate for higher payload.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

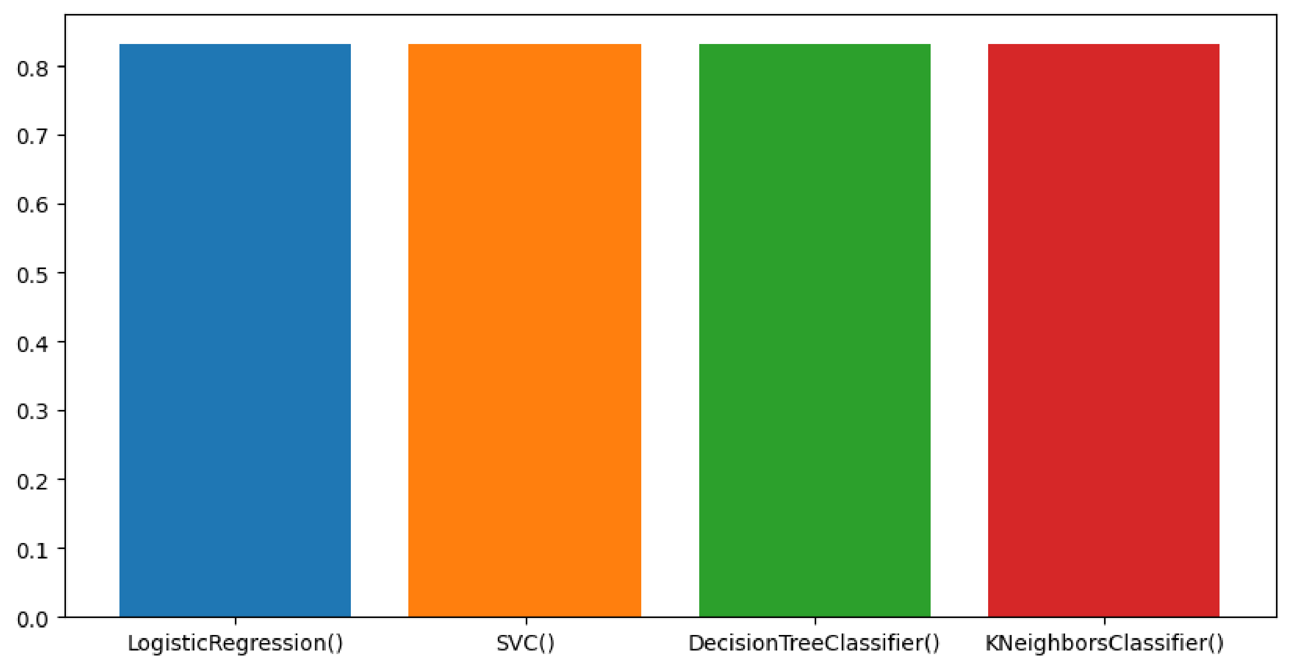
Based on the provided information, the models have the following classification test accuracy:

Logistic Regression: 0.8333

SVM: 0.8333

Decision Tree: 0.8333

K-Nearest Neighbors: 0.8333



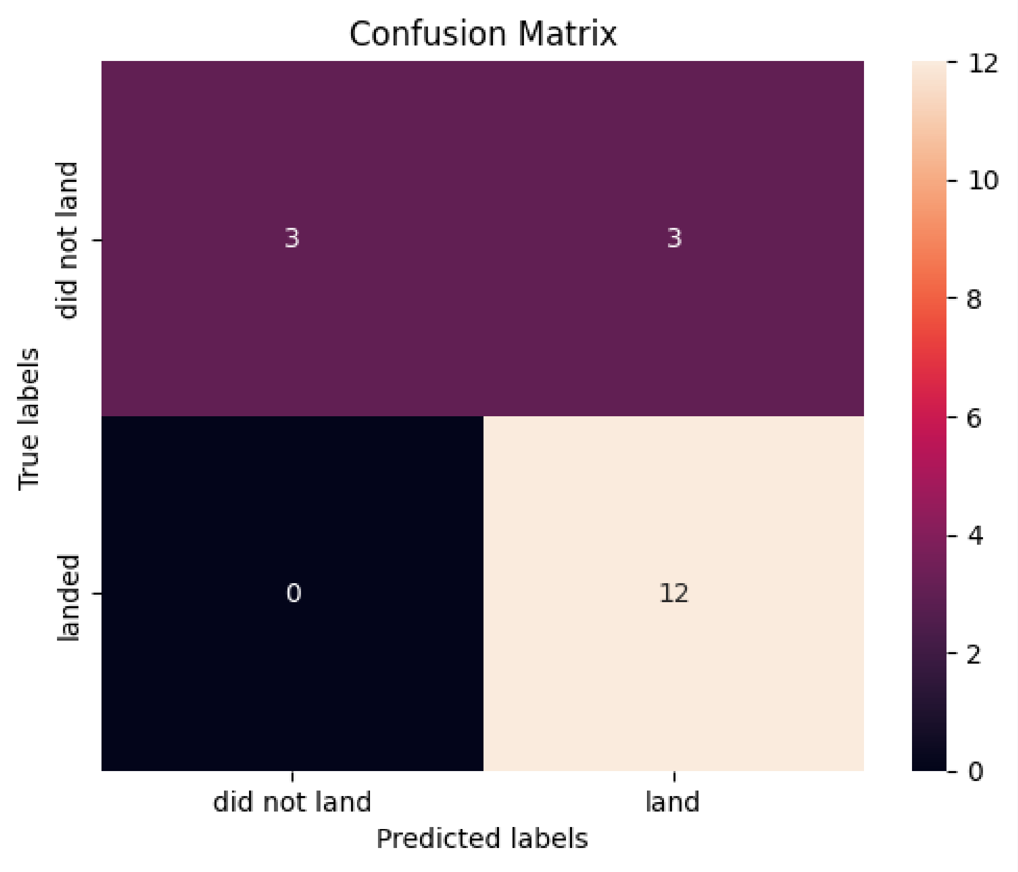
All four models have the same highest classification test accuracy of 0.8333. Therefore, there is a tie among these models for the highest test accuracy.

Confusion Matrix

- Decision Tree Classifier appears to be the best performing model among the given options. Here's the relevant information:

- Accuracy on train: 0.90
- Accuracy on test: 0.83

- Tuned hyperparameters: {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'random'}



Conclusions

- Orbits ES-L1, GEO, HEO, SSO, VLEO showed the highest success rates for Falcon 9 launches.
- KSC LC-39A was the most successful launch site, with a success rate of 76.9%.
- Launch success rate for Falcon 9 increased consistently from 2013 to 2020.
- All SpaceX launch sites are strategically located near the coast, away from populated areas. Additionally, the sites are situated near the Equator to leverage the Earth's rotation, enabling higher initial speeds and more efficient space missions.
- Machine learning models achieved an accuracy of 83.33% in predicting the landing outcome of rocket launches, with Decision Tree Classifier performing the best among the options.

Appendix

- GitHub Repository :- https://github.com/TheStrange-007/IBM-DS-Project

Thank you!