Applied Data Science Capstone

The Applied Data Science Capstone project was a course I completed to earn my IBM Professional Certification in Data Science on Coursera. This project allowed me to apply the skills I learned over the previous seven months of courses, utilizing real-world datasets for extraction, cleaning, exploration, and analysis to derive meaningful insights. Today, I successfully completed the certification that I began in the first quarter of the year, and I'm excited to share more about my capstone project.

Problem

SpaceX is an aerospace manufacturer and space transportation services and communications corporation, much like Tesla, both founded by Elon Musk. Despite being less than 20 years old, SpaceX has managed to reduce launch costs by more than 50% compared to other companies, with projections suggesting a 99% reduction once the Starship project is completed. This reduction is largely attributed to SpaceX's development of technology to land the first stage booster, which constitutes 70% of the rocket's cost. By safely landing and reusing the booster, SpaceX significantly cuts down on launch expenses. Reusing boosters reduces costs by 50% compared to using new boosters, solidifying SpaceX's dominance in the market. In this capstone project, we will analyze data extracted from Wikipedia through web scraping and the SpaceX API to gain insights and predict safe booster landings onto drone ships.

Data Collection

Data collection through web scraping is available on this GitHub here
SpaceX API data collection is available on this GitHub here

Data Wrangling

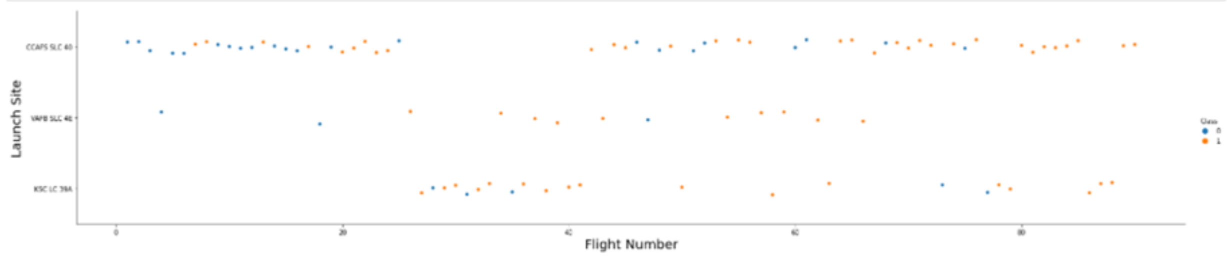
After collecting the data, we check the missing data and data types and do on of the following to clean the data:

a. *Replace the missing data with one-Using mean or so.*
b. *Change data type of the data.*
c. *Represent categorical data using integer or float dummy numbers-one hot encoding*

Exploratory Data Analysis

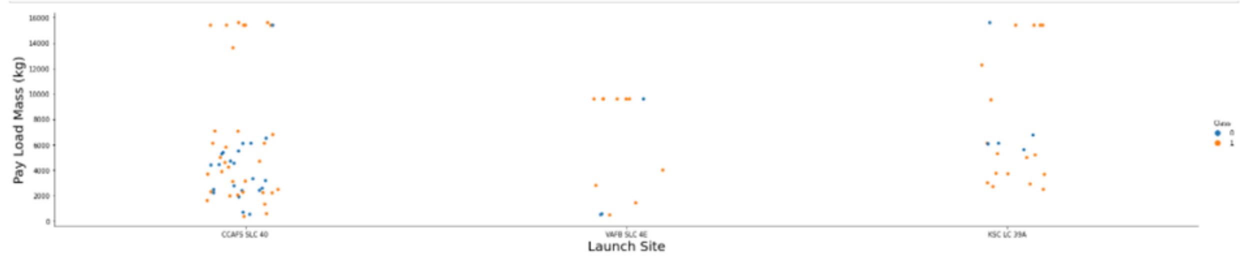
After Data cleaning the we can proceed to Analyzing the data using visualization to get some insights of the launches.

Here are some of the screenshots



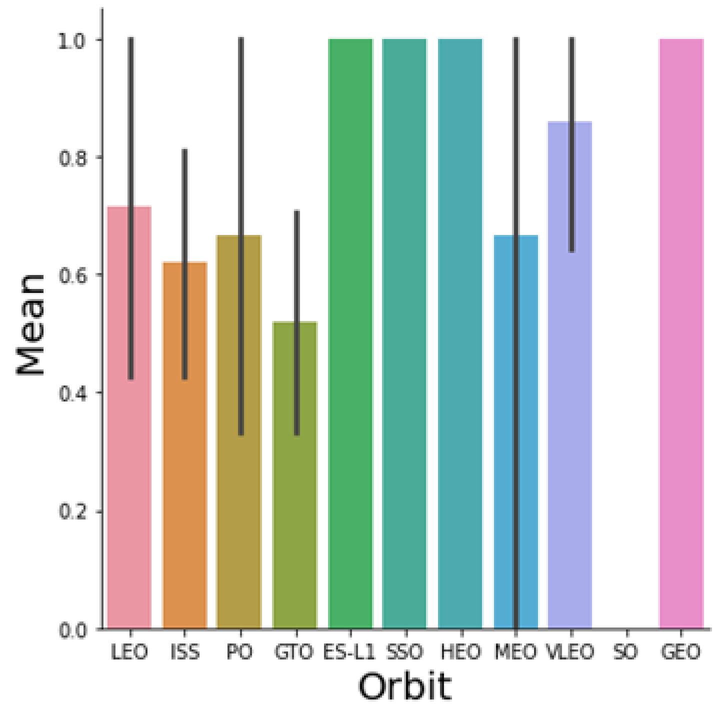
From the Visualization we can conclude that:

1. Earlier flights launch was from CCAFS-SLC-40 site, Followed by KSC-LC-39A
2. Most Launches are Launched from CCAFS-SLC-40
3. Fewer Launches from VAFB SLC 4E site



From the Visualization we can concluded that:

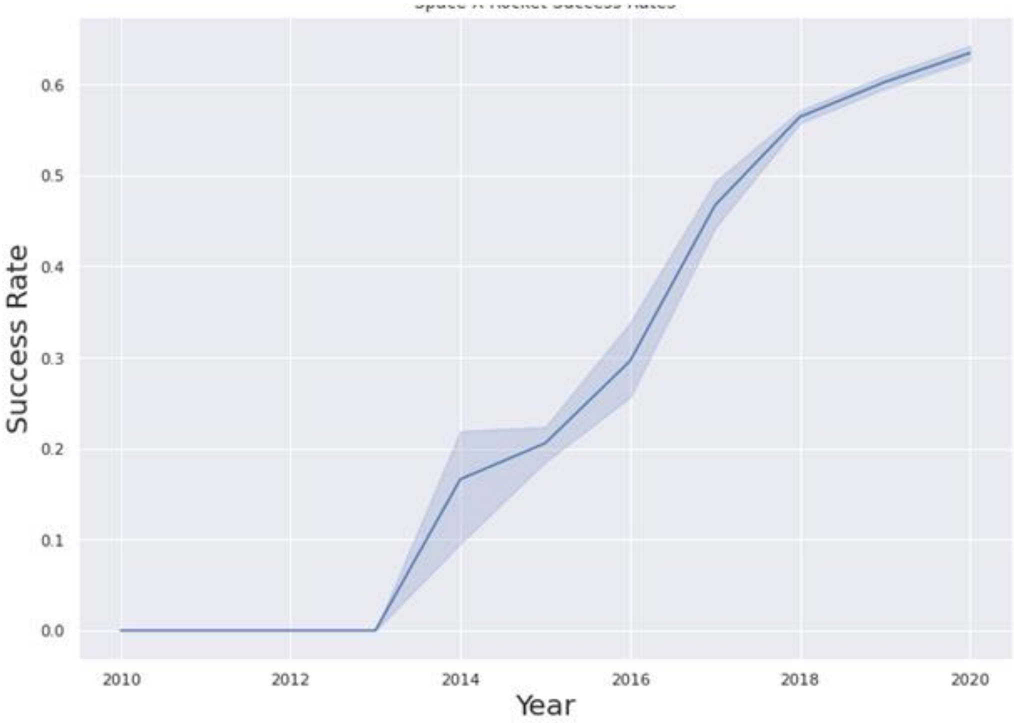
- VAFB SLC 4E has Low Payload launches
- CCAFS SLC 40 has more Higher Payload Launches and Low Payload Launches.



From the visualization, we can conclude that:

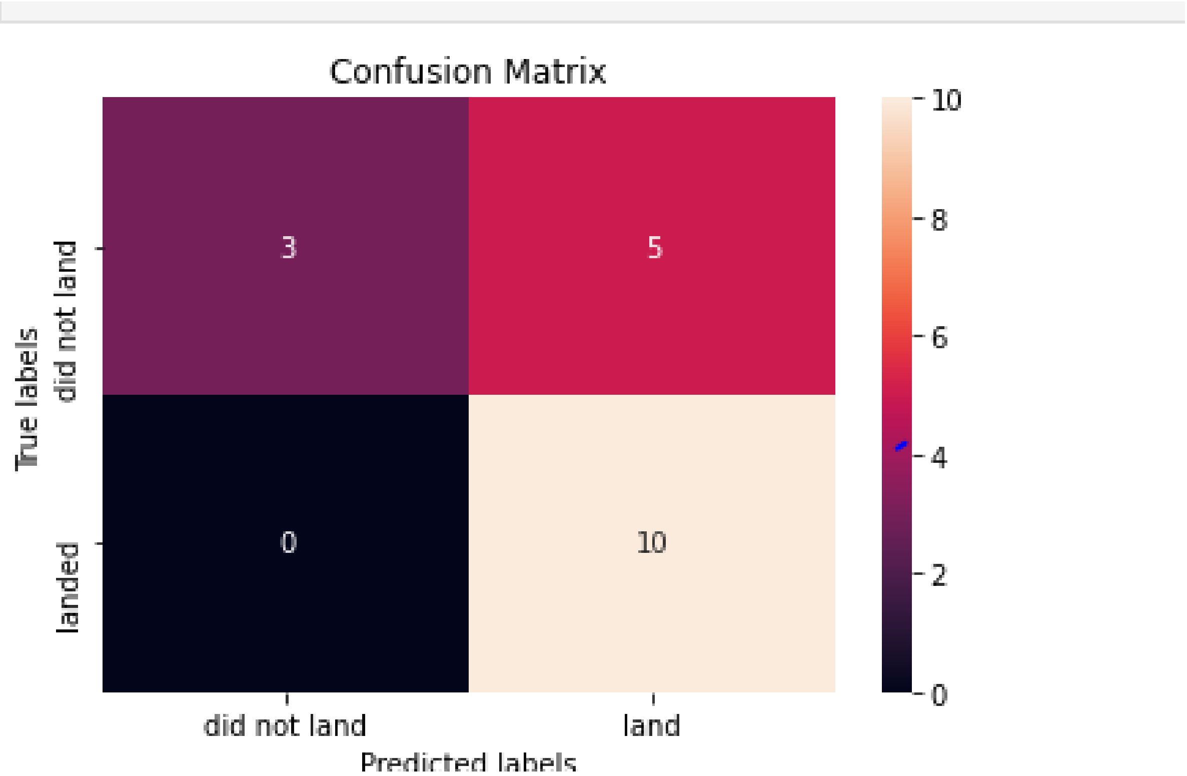
❖ GEO, HEO & ES-L1, SSO) have high success rate.

Other analysis includes Exploratory analysis of data from db2 database using sql statement to get insights here



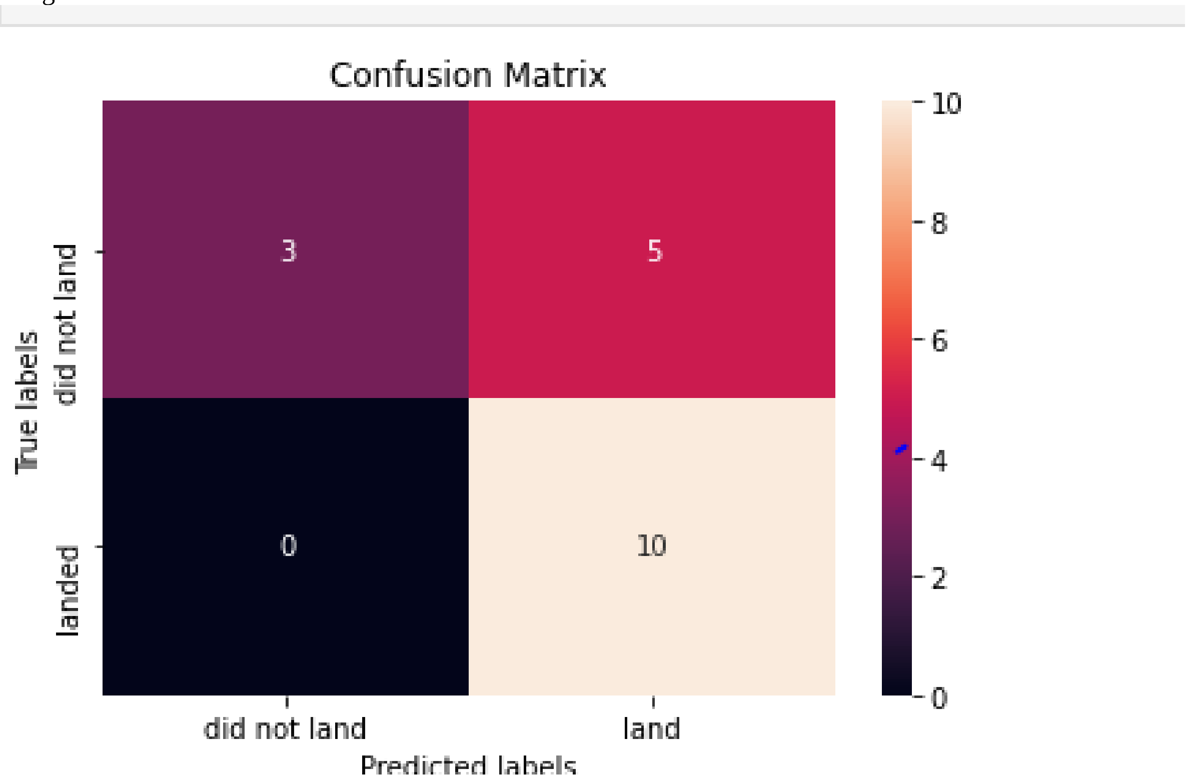
We analyze the data and see that their success rate which shows increase in landing success probability.

Predictive Analysis



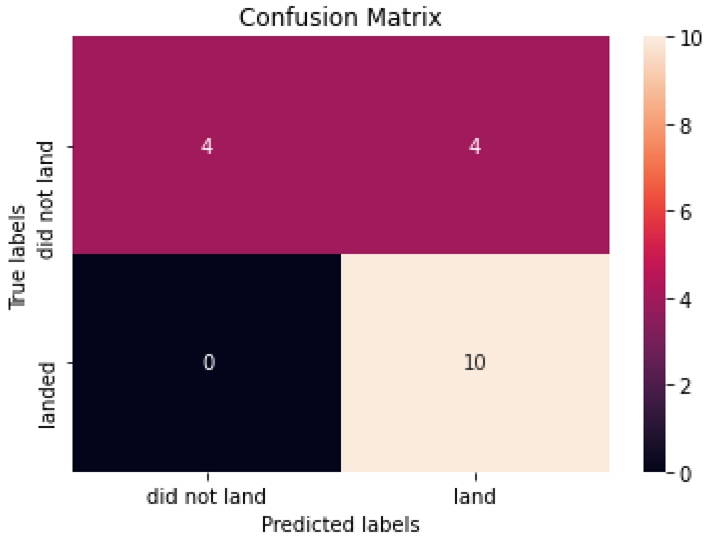
Using the data I trained the Machine learning models such as:

- ❖ KNeighborsClassifier

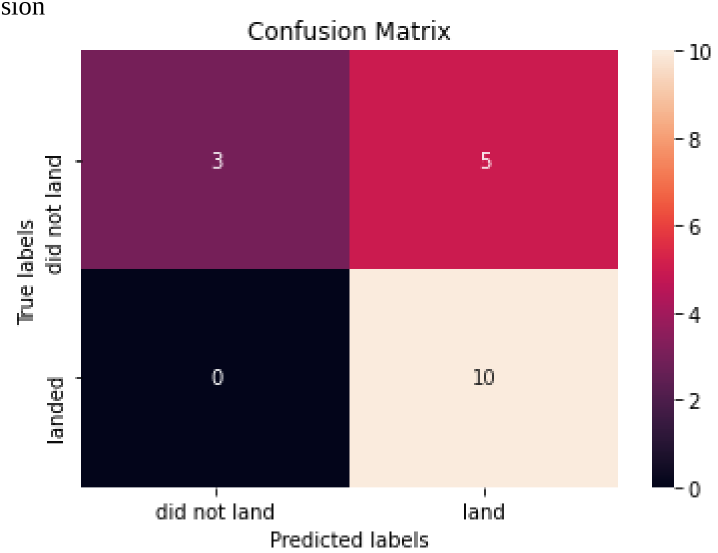


KNN confusion Matrix

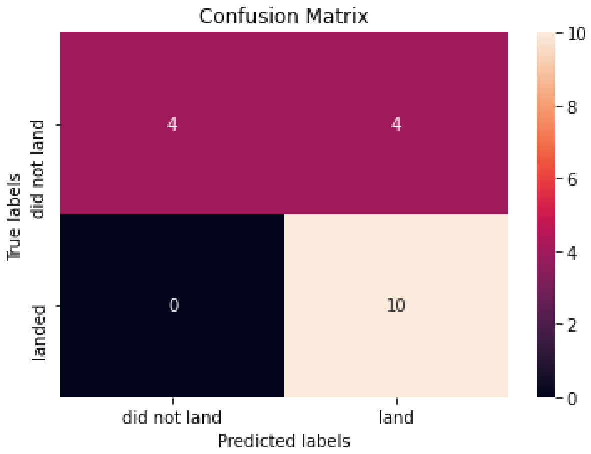
- Decision Tree classifier



- Logistic regression

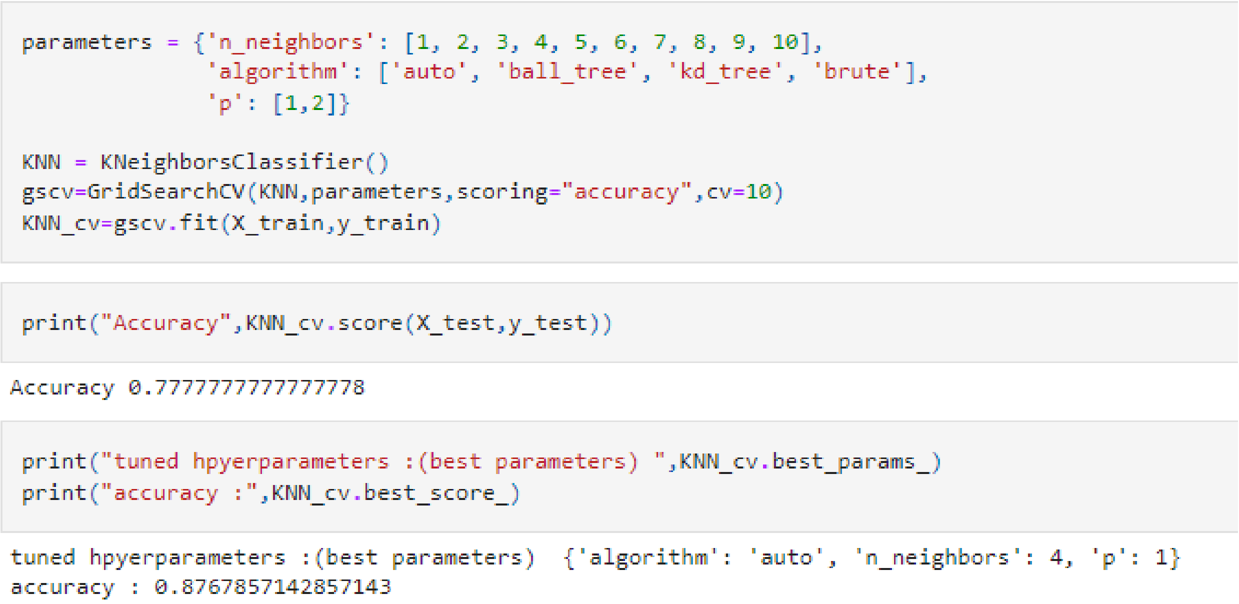


- Support Vector Machine



svm confusion matrix

Furthermore, I tuned the models to obtain the most accurate model of all of these evaluating their score, best score and confusion matrix plot and concluded that KNN model has the best score, accuracy and least bias confusion matrix.



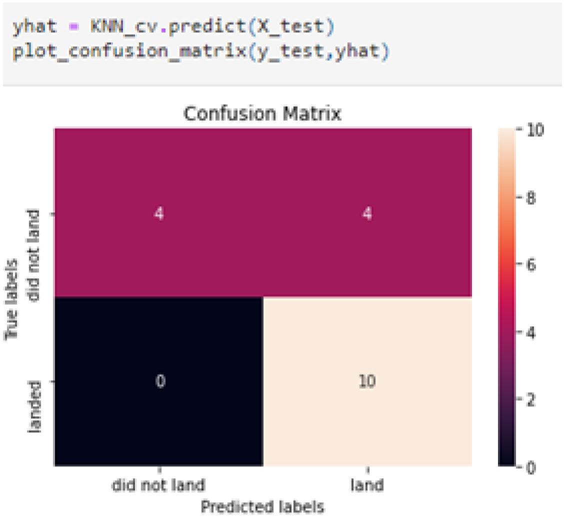
```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
gscv=GridSearchCV(KNN,parameters,scoring="accuracy",cv=10)
KNN_cv=gscv.fit(X_train,y_train)
```

```
print("Accuracy",KNN_cv.score(X_test,y_test))
```

```
Accuracy 0.7777777777777778
```

```
print("tuned hpyerparameters :(best parameters) ",KNN_cv.best_params_)
print("accuracy :",KNN_cv.best_score_)
```

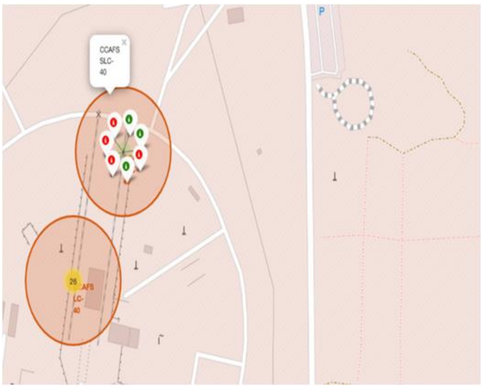
```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 4, 'p': 1}
accuracy : 0.8767857142857143
```



```
yhat = KNN_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat)
```

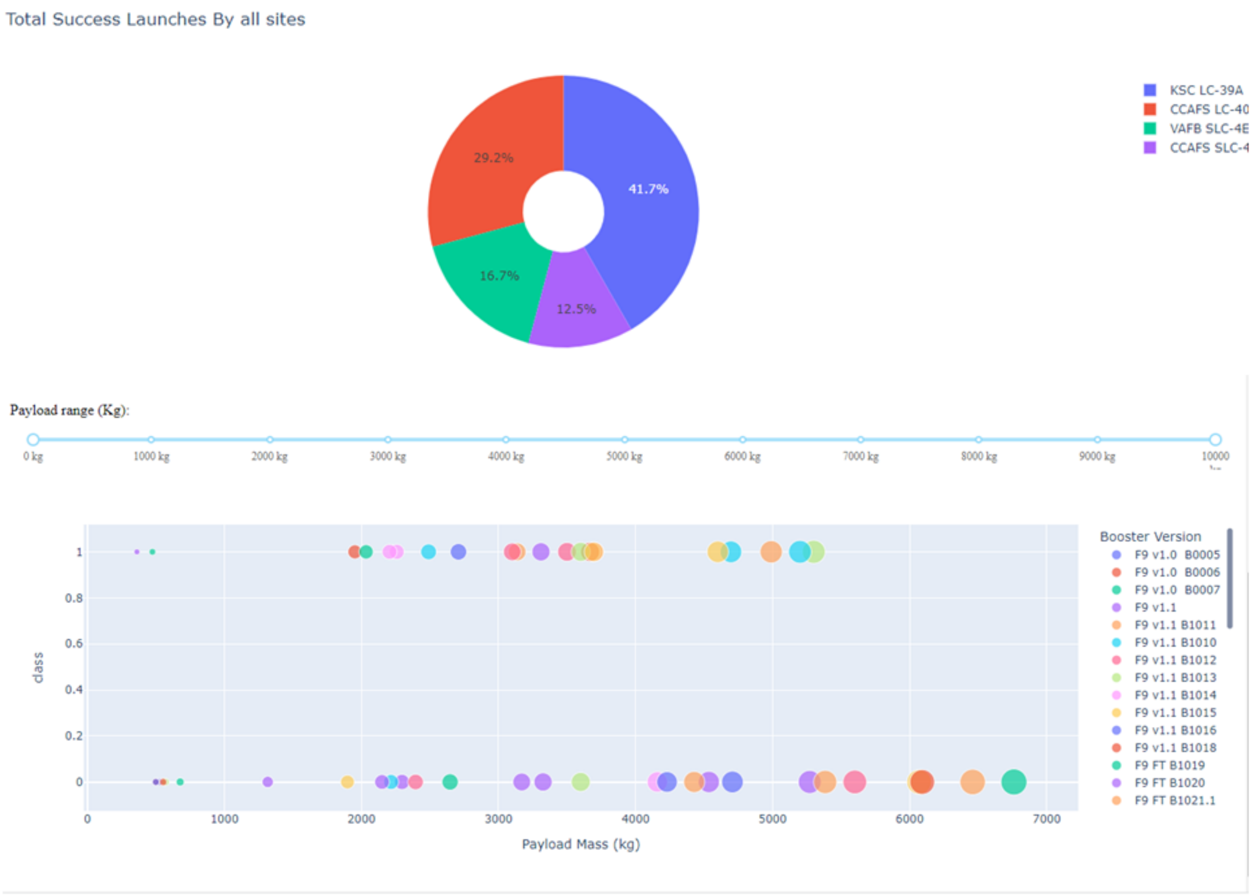
Interactive Map with Folium

Since SpaceX launches come from different launch sites, I displayed the information of failed and successful launches as a cluster on the map. Through zooming in and out you can observe the clusters of success launches and failed launches.



Interactive Dashboard with Plotly Dash

Plotly Dash is Python library that makes it easier to create a dashboard for us as Data Scientist. With a simple interactive dashboard one can change the inputs to see representation of values in graphs.



Results

My insights from the analysis show that most launches originated from the Kennedy Space Center (KSC). This is primarily due to its proximity to SpaceX's production facility. The majority of launches occurred from KSC PAD 39A, as many of these missions targeted Very Low Earth Orbit (VLEO), Geostationary Orbit (GEO), or the International Space Station (ISS), making it an ideal launch site. Falcon Heavy launches typically carry full payloads to maximize the Falcon's payload capacity. The probability of successful booster landings increases over time, using data collected from previous attempts. SpaceX achieved its first successful booster landing on 06/05/2016.

Conclusion

By utilizing existing data and analyzing it, SpaceX and other rocket companies can identify the most effective strategies to reduce launch costs and evolve their operations. This proactive approach is essential to avoid traditional costly launches that could potentially lead to obsolescence and loss of client.