IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akanksha Mathpati
18 April 2024

Outline

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

- **Project background and context**

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage. Questions

- **Problems you want to find answers**

➢ What factors might affect the decision?

➢ What is the relations among various factors which will help determine successful launch?

➢ Does the rate of successful landings increase over the years?

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data was collected from SpaceX API
  - Web scraping from Wikipedia
- Perform data wrangling
  - Filtered the data
  - Dealt with missing values
  - Performed one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

SpaceX API Integration

- Utilized GET requests to access the SpaceX API.
- Decoded the response content, converting it into JSON format using the **.json()** function call.
- Transformed the JSON data into a structured Pandas dataframe using the **.json_normalize()** function.

Data Cleaning

- Conducted data cleaning procedures to ensure data integrity.
- Checked for missing values and implemented strategies to handle them effectively.

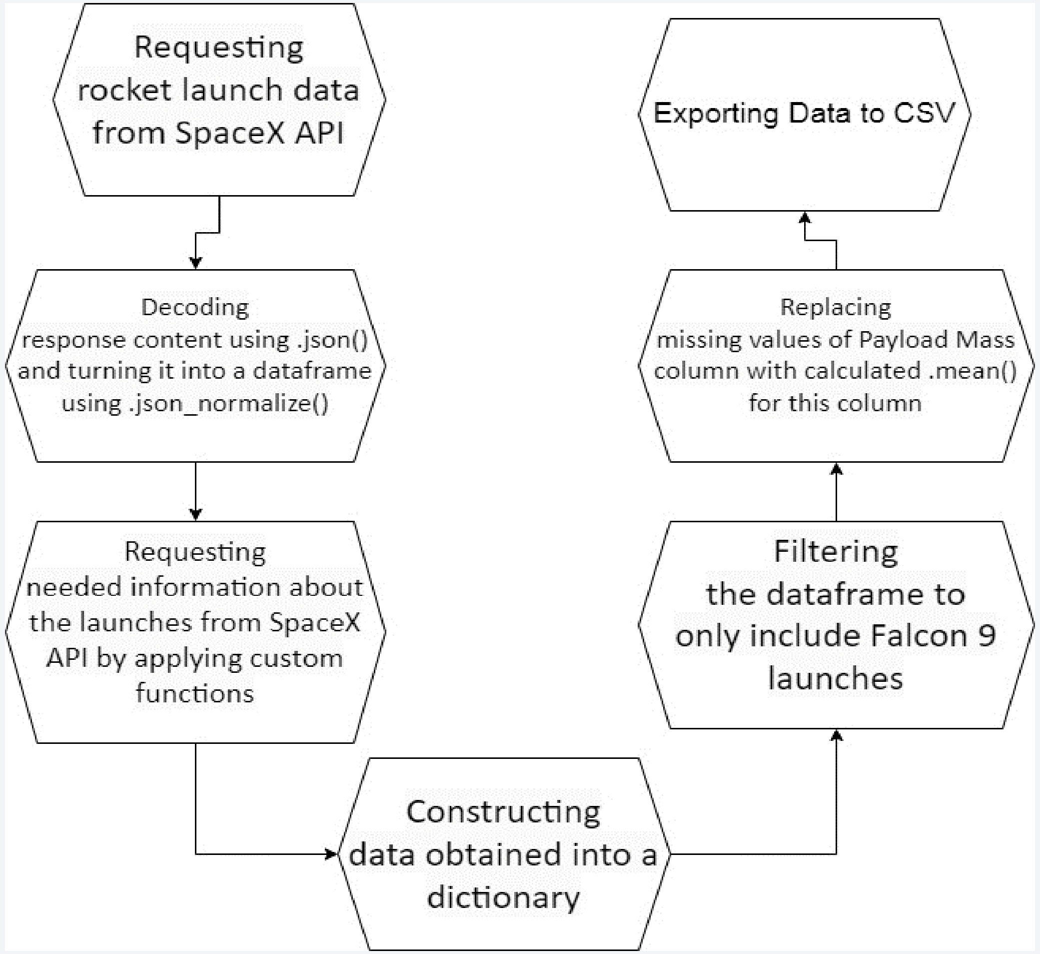
Web Scraping from Wikipedia

- Employed web scraping

- GitHub link Data Collection- SpaceX API

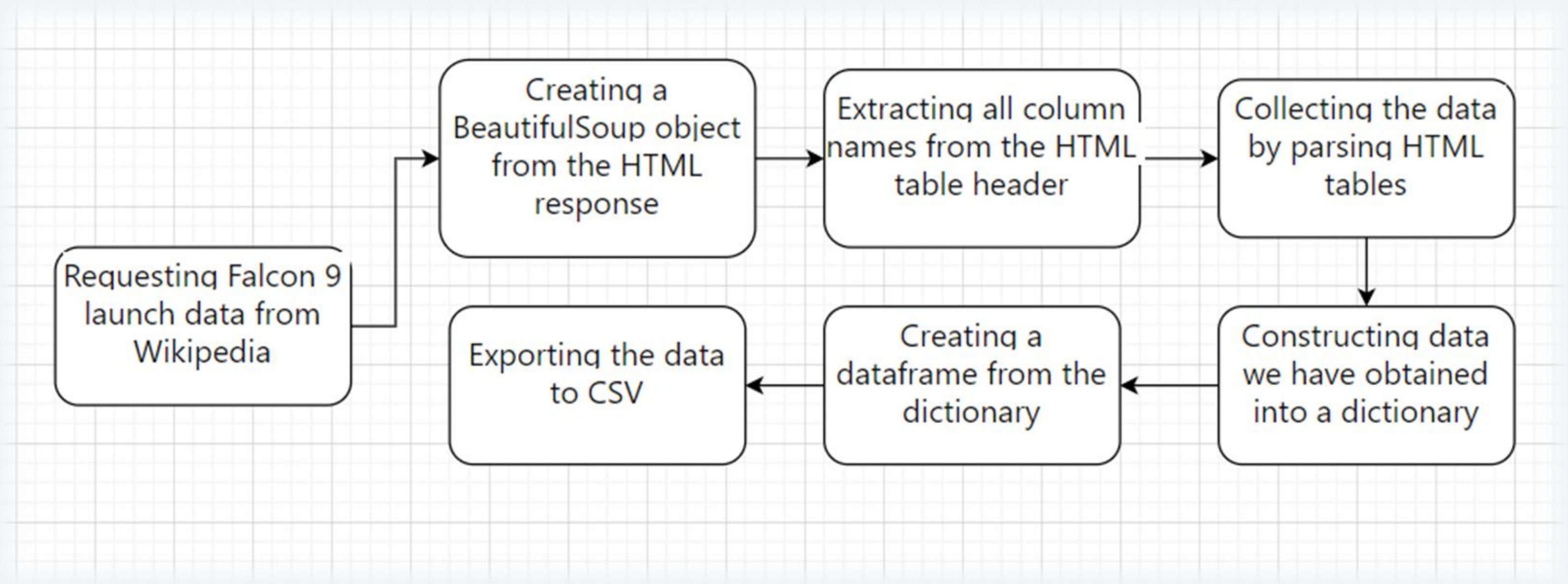
Data Collection – SpaceX API

- GitHub link Data Collection- SpaceX API



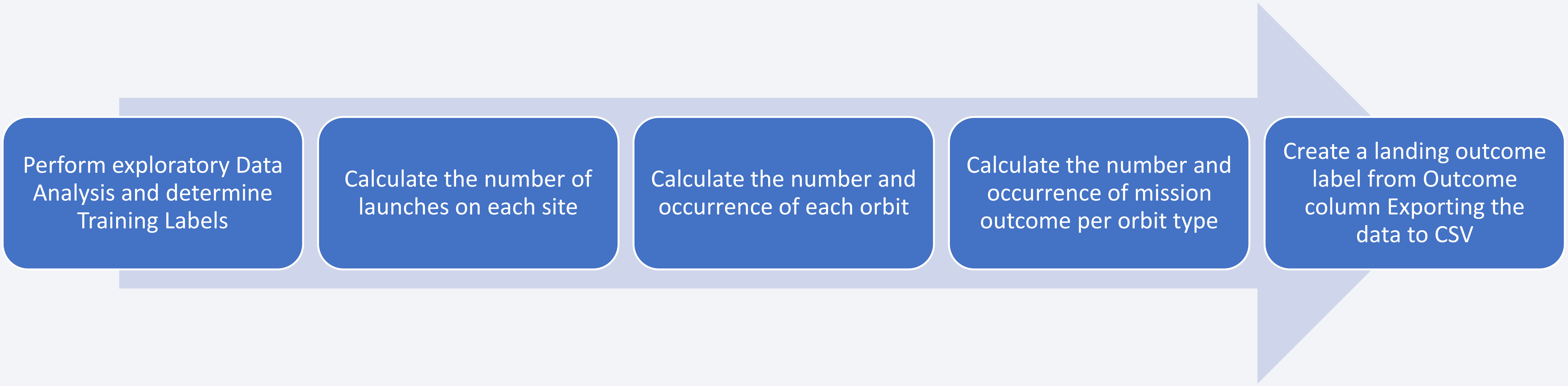
Data Collection - Scraping

- We applied web scrapping to web scrap Falcon 9 launch records with BeautifulSoup
- •We parsed the table and converted it into a pandas dataframe.
- •The link to the notebook is GitHub Link.



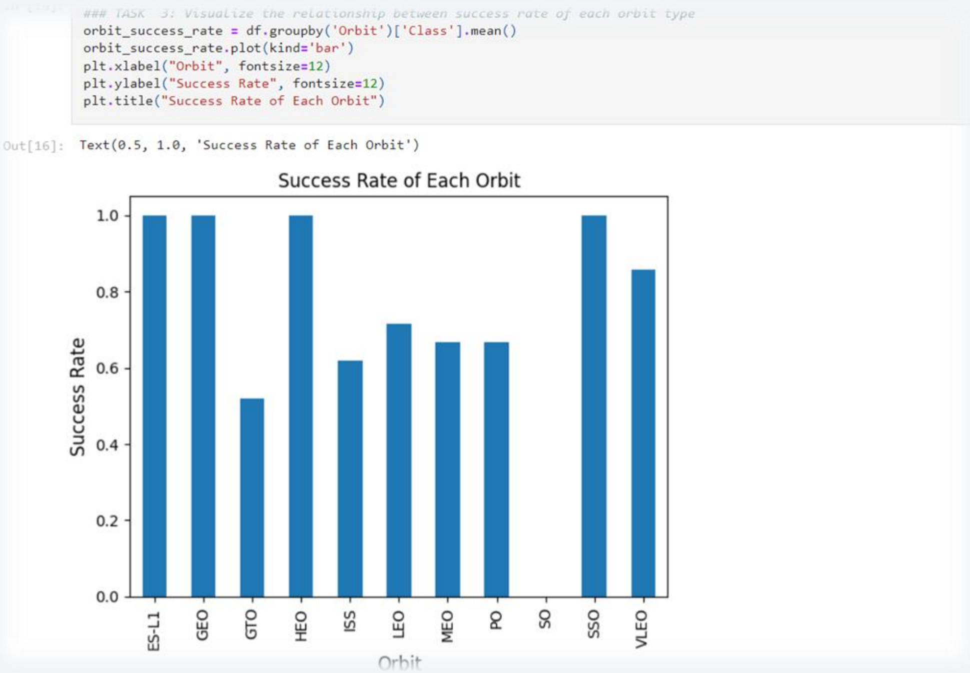
Data Wrangling

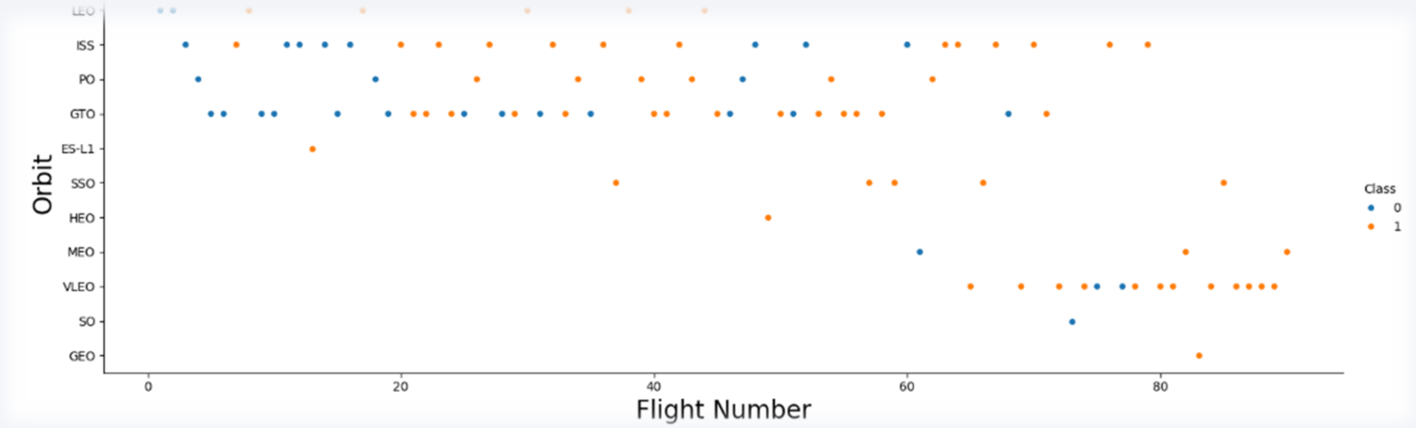
- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.
- •The link to the notebook is GitHub Link: Data Wrangling.



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.





Eda with Data Visualization GitHub Link

EDA with SQL

- **Queries were performed for the below tasks:**
- total by NASA (CRS) payload mass carried by boosters launched
- average payload mass carried by booster version F9 v1.1
- successful landing outcome in drone ship was achieved.
- To List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass.
- List the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- 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,

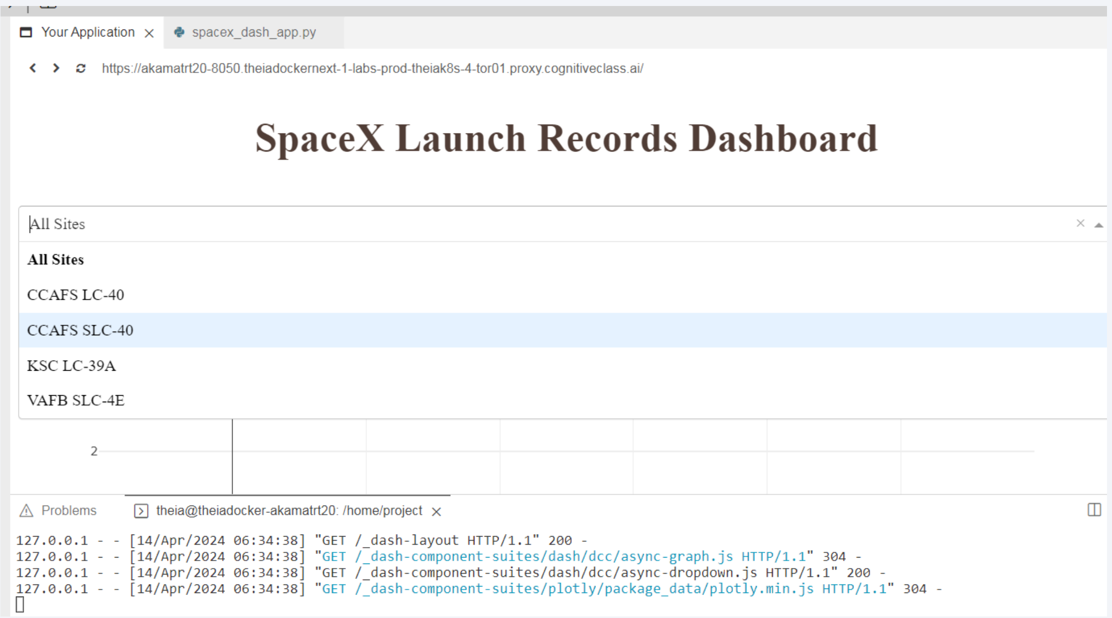
GitHub Link: EDA with SQL

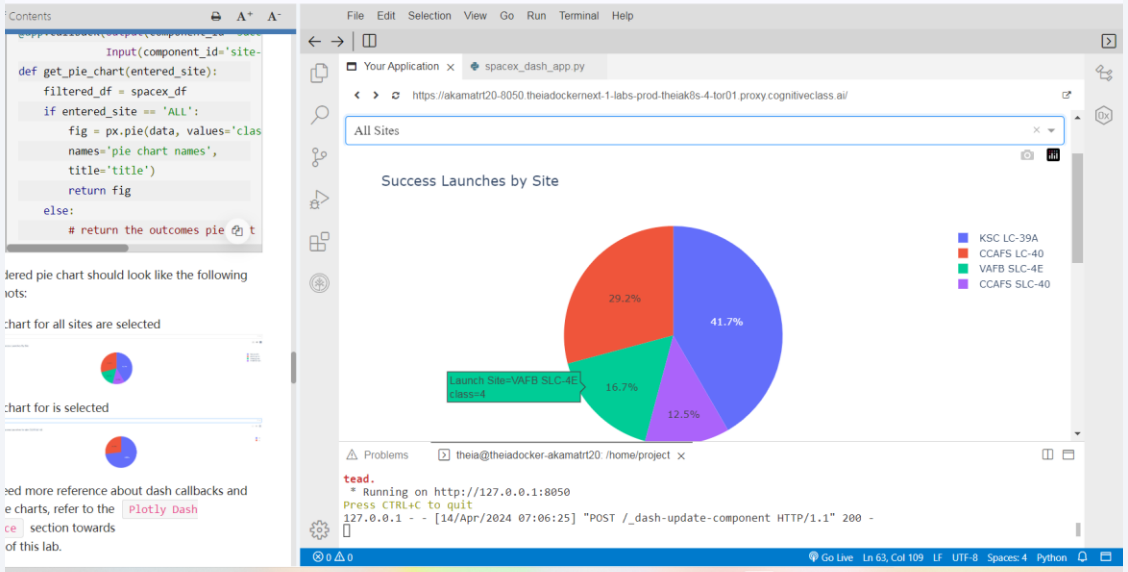
Build an Interactive Map with Folium

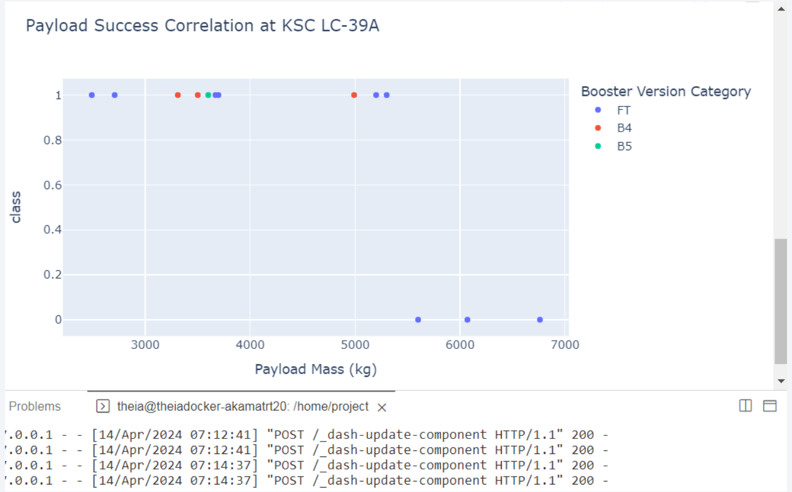
- Utilized markers, circles, and lines on a Folium map to visually represent launch sites and their proximities.
- Categorized launch outcomes as class 0 for failure and class 1 for success, facilitating data analysis.
- Employed color-labeled marker clusters to identify launch sites with high success rates, aiding in strategic decision-making.
- Conducted distance calculations between launch sites and nearby features such as railways, highways, coastlines, and cities.
- Addressed questions regarding the proximity of launch sites to various infrastructures and urban areas, enhancing understanding of site selection criteria.

Interactive Map with Folium

Build a Dashboard with Plotly Dash







We built an interactive dashboard with Plotly dash

•We plotted pie charts showing the total launches by a certain sites

•We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

•The link to the notebook is Ploty Dash GitHub Link.

Predictive Analysis (Classification)

- loaded the data using numpyand pandas.
- Standardizing the data with StandardScaler, then fitting and transforming it.
- split our data into training and testing.
- built different machine learning models and tune different hyperparameters using GridSearchCV.
- used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- found the best performing classification model.

[Predictive Analysis GitHub Link]

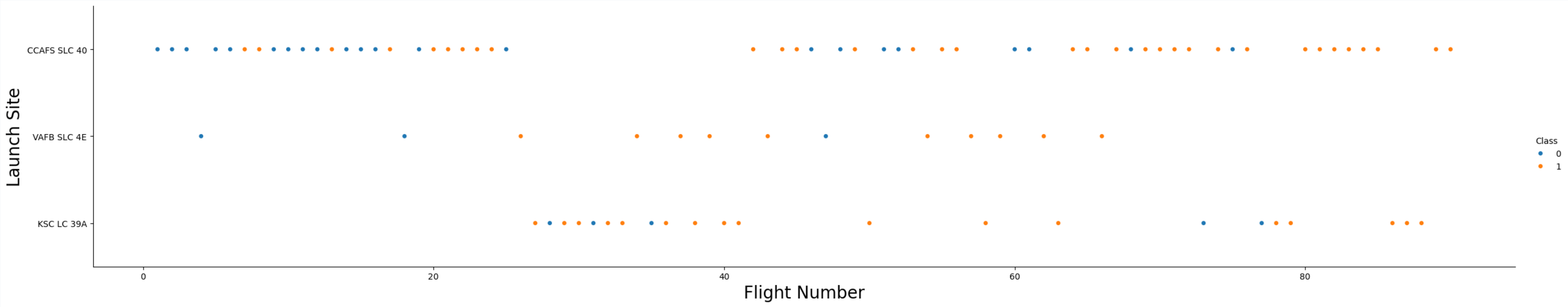
Results

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

Section 2

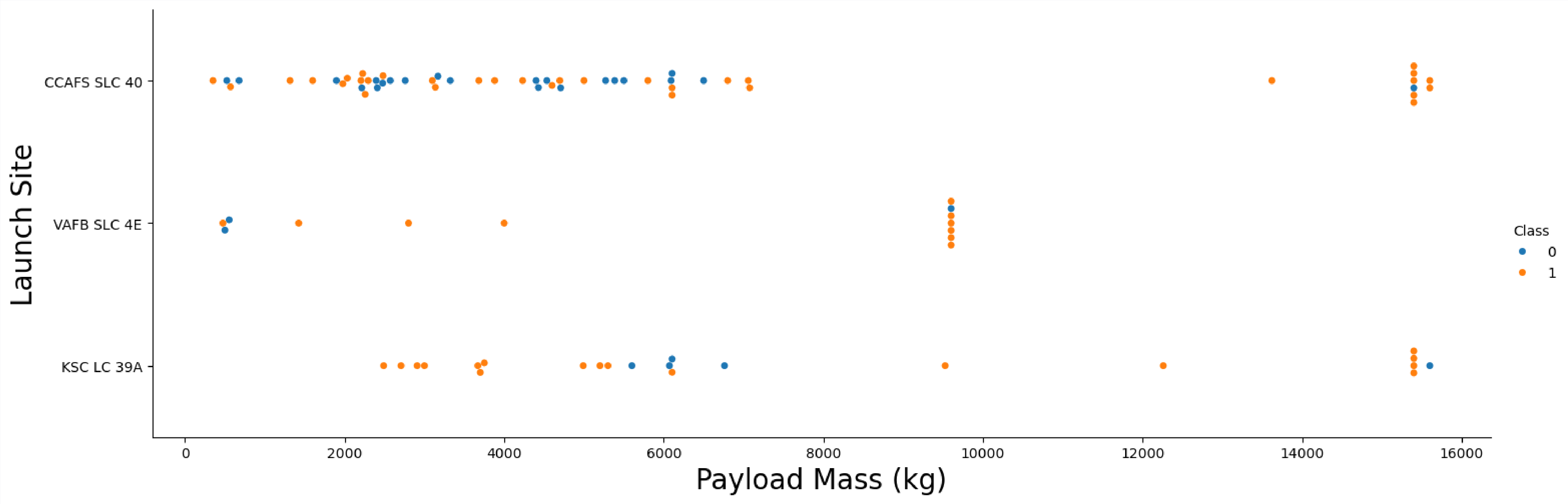
Insights drawn from EDA

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.

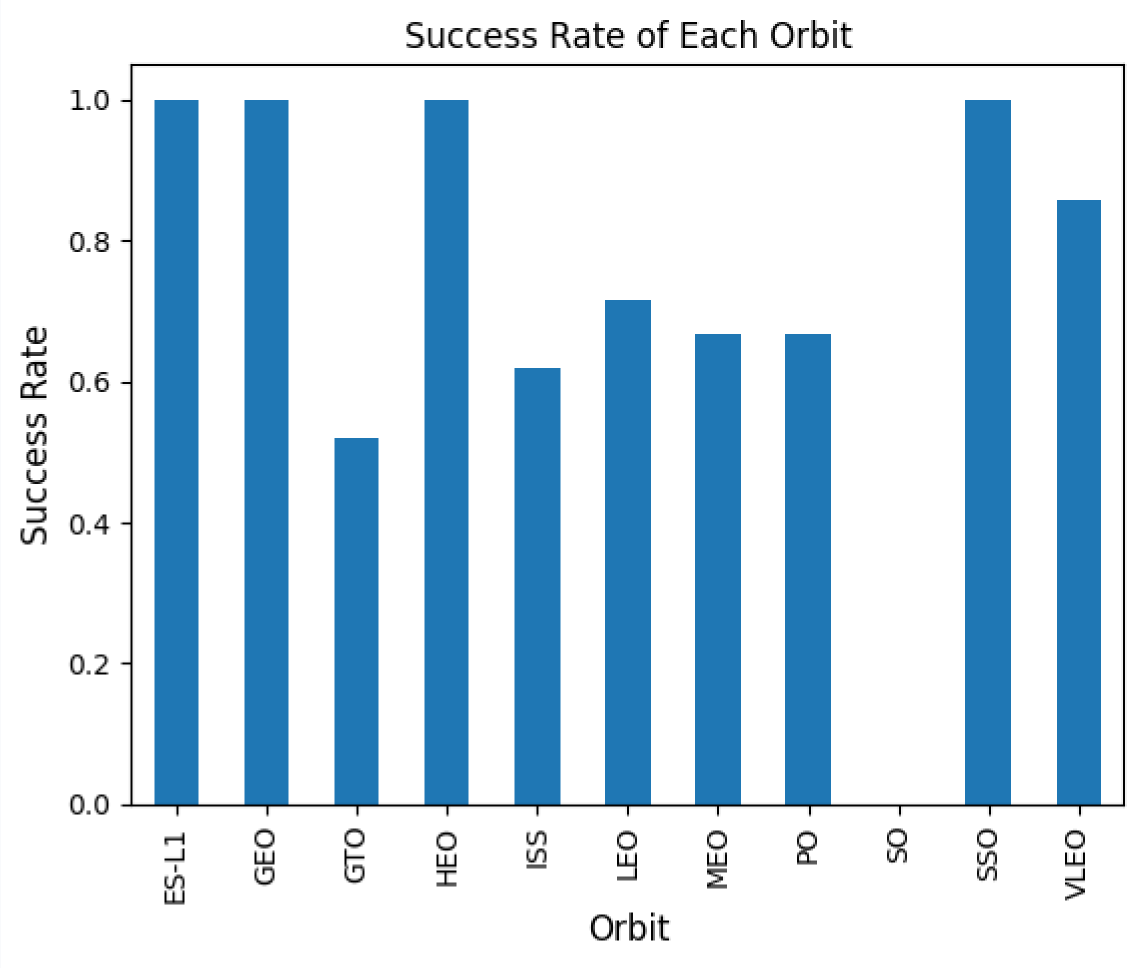
Payload vs. Launch Site



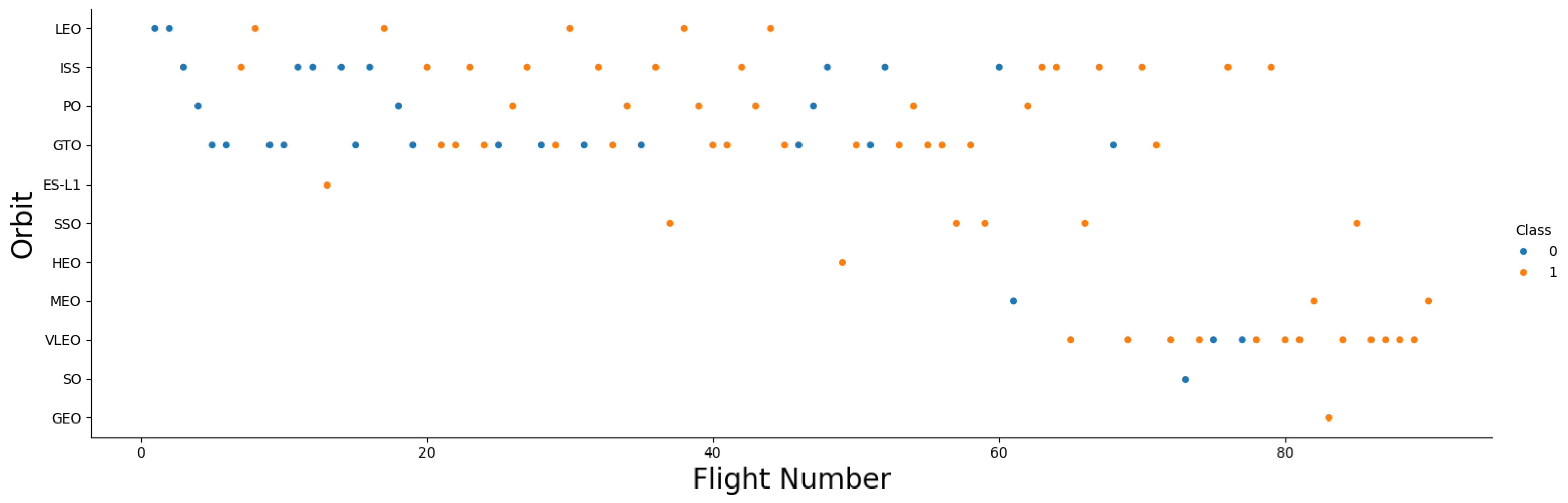
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.

Success Rate vs. Orbit Type

- Orbits with 100% success rate are ES-L1, GEO, HEO, SSO, VLEO.
- Orbit SO has 0% success rate.

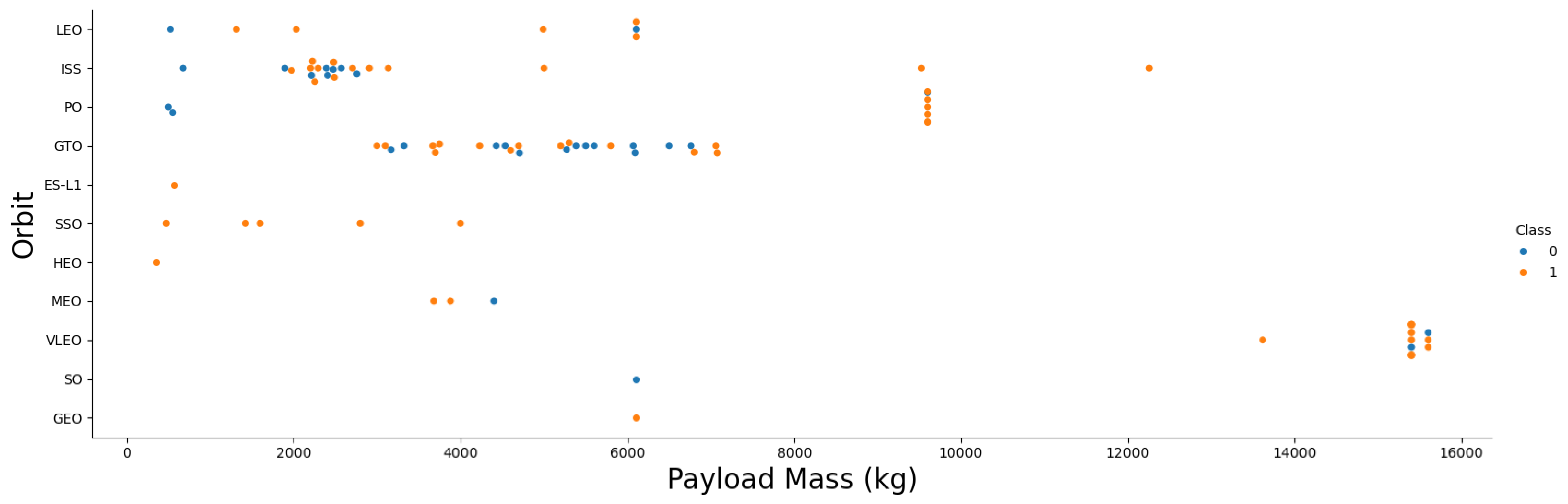


Flight Number vs. Orbit Type



- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

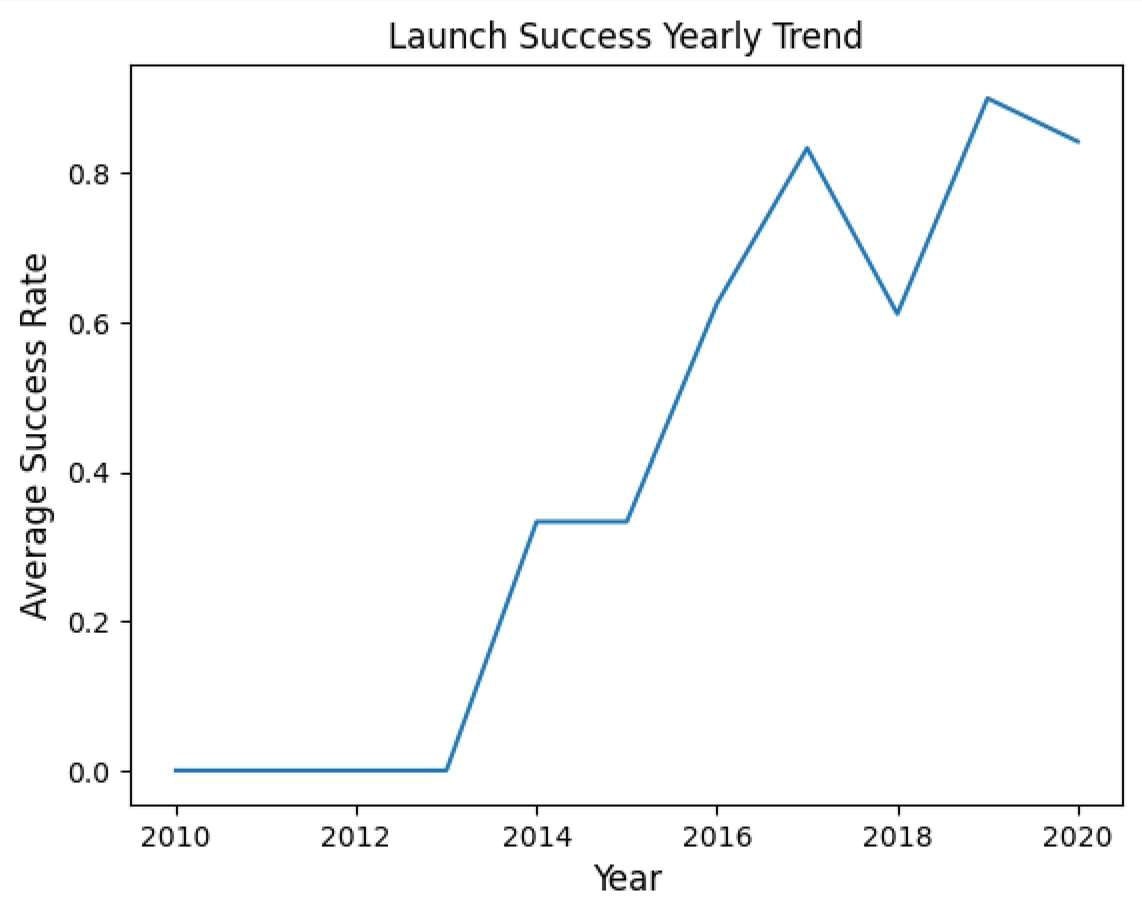
Payload vs. Orbit Type



- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend

- The average success rate kept increasing since 2013.
- There was a drop in the year 2018, however that rate picked up soon in 2019.



All Launch Site Names

- We used the key word distinct to Displaying the names of the unique launch sites in the space mission.

Task 1

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

```
%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

Task 2

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

Launch Site Names Begin with 'KSC'

- Displaying 5 records where launch sites' names start with `KSC`

Task 2

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

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'KSC%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

Task 3

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE 'NASA%';
```

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

Total_Payload_Mass
99980

- We used the query SELECT SUM.
- Total Payload Mass was 99980.

Average Payload Mass by F9 v1.1

Task 4

Display 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';
```

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

Average_Payload_Mass
2928.4

We used the query SELECT AVG.

Total Payload Mass was 2928.4

First Successful Ground Landing Date

- dates of the first successful landing outcome on drone ship.
- Used the min function.

Task 5

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) AS Successful_Landing_Date FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)';
```

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

Successful_Landing_Date
2016-04-08

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ > 4000 ANI
```

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

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- Listing the 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

- Listing the total number of successful and failure mission outcomes.

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS Total_Outcomes FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

Task 8

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass.

Task 8

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

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

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

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

Task 9

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Note: SQLLite does not support monthnames. So you need to use substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2017' for year.

```
%sql SELECT substr(Date,6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

- Ranking 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.

Task 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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GR(
```

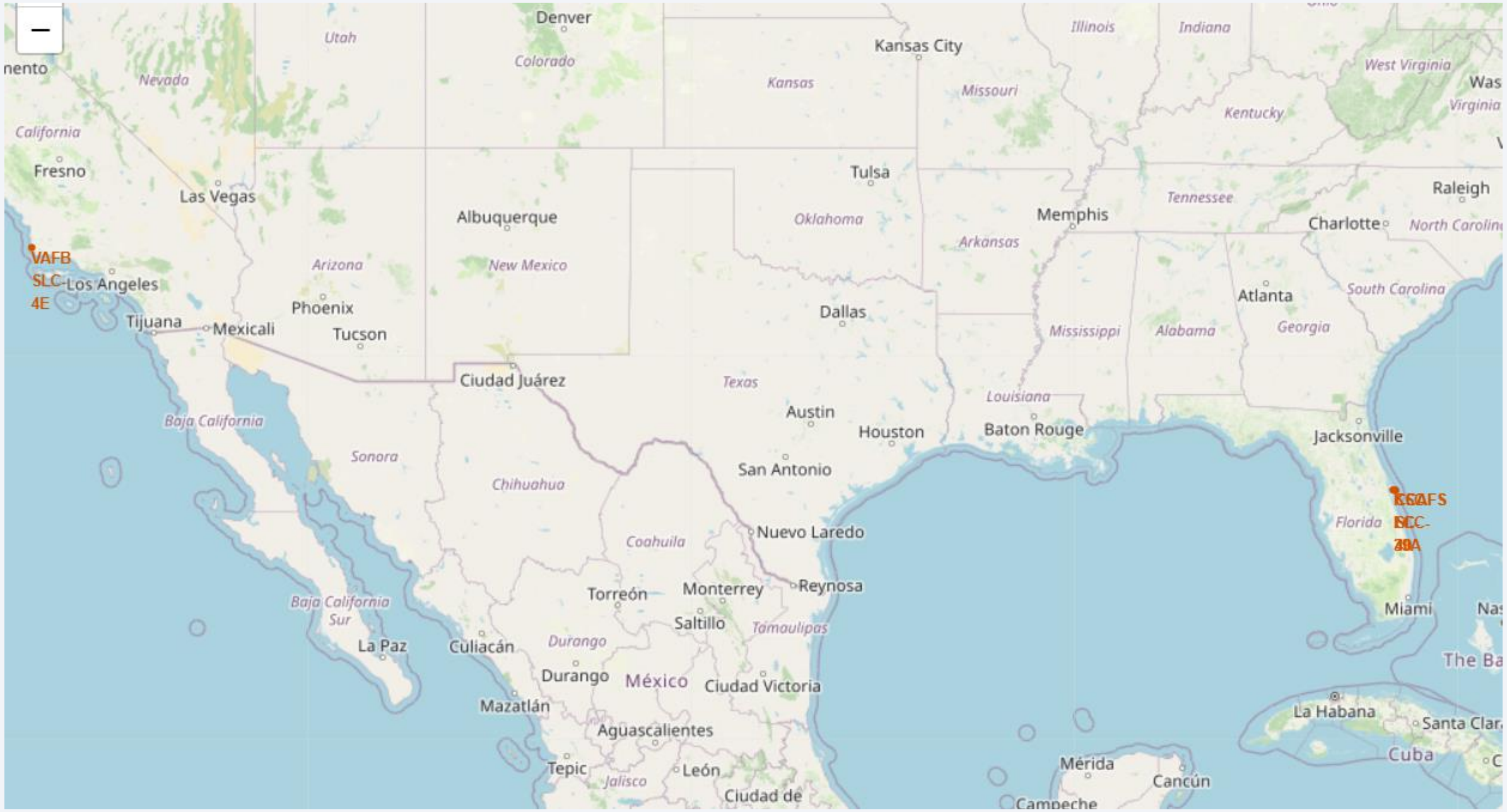
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Outcome_Count
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

Section 3

Launch Sites Proximities Analysis

Launch sites' location markers on a global map

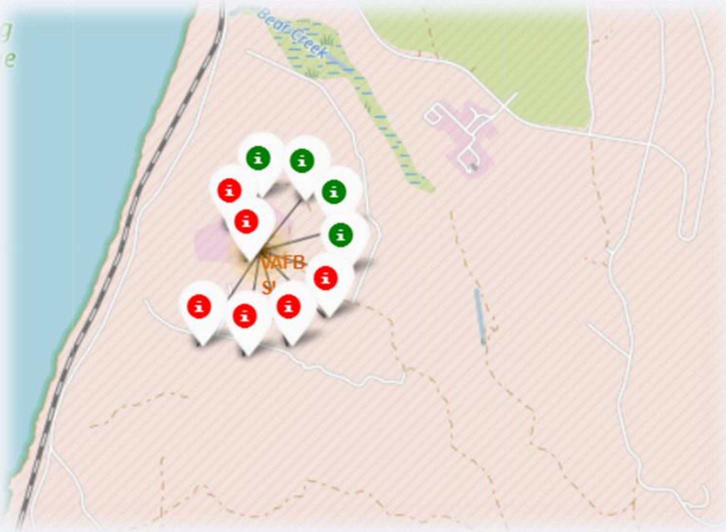


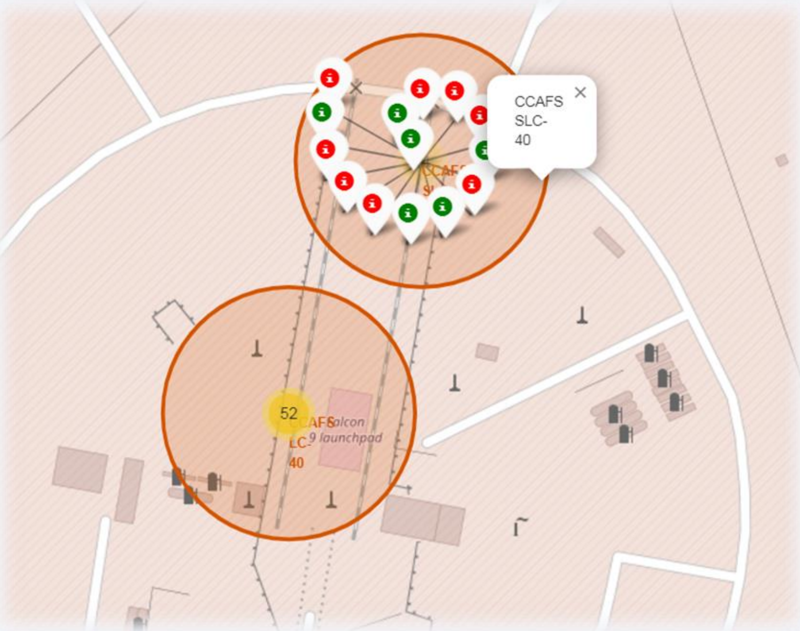
- Most of Launch sites are in proximity to the Equator line.

Markers showing launch sites with colored labels

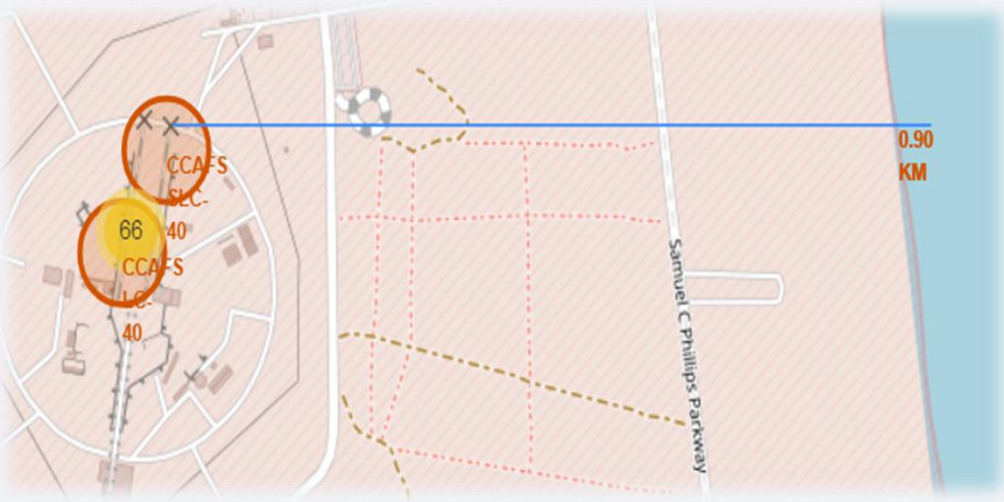
From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

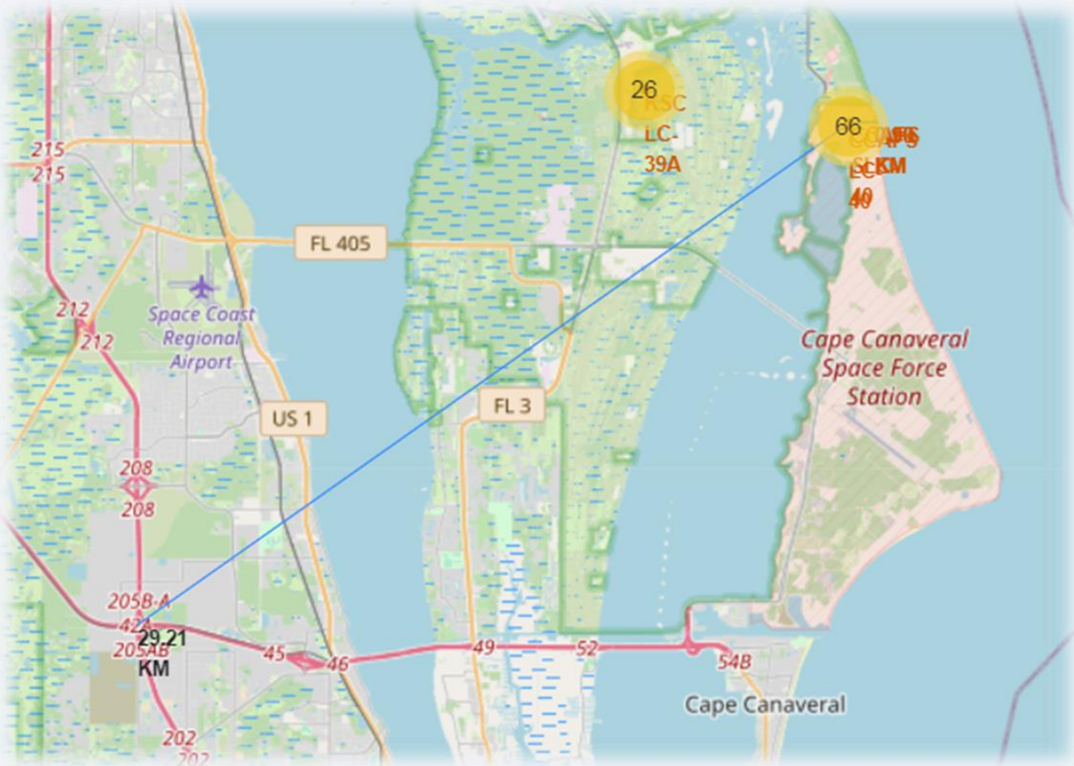
- Green Marker = Successful Launch
- Red Marker = Failed Launch





Launch Site distance to different landmarks



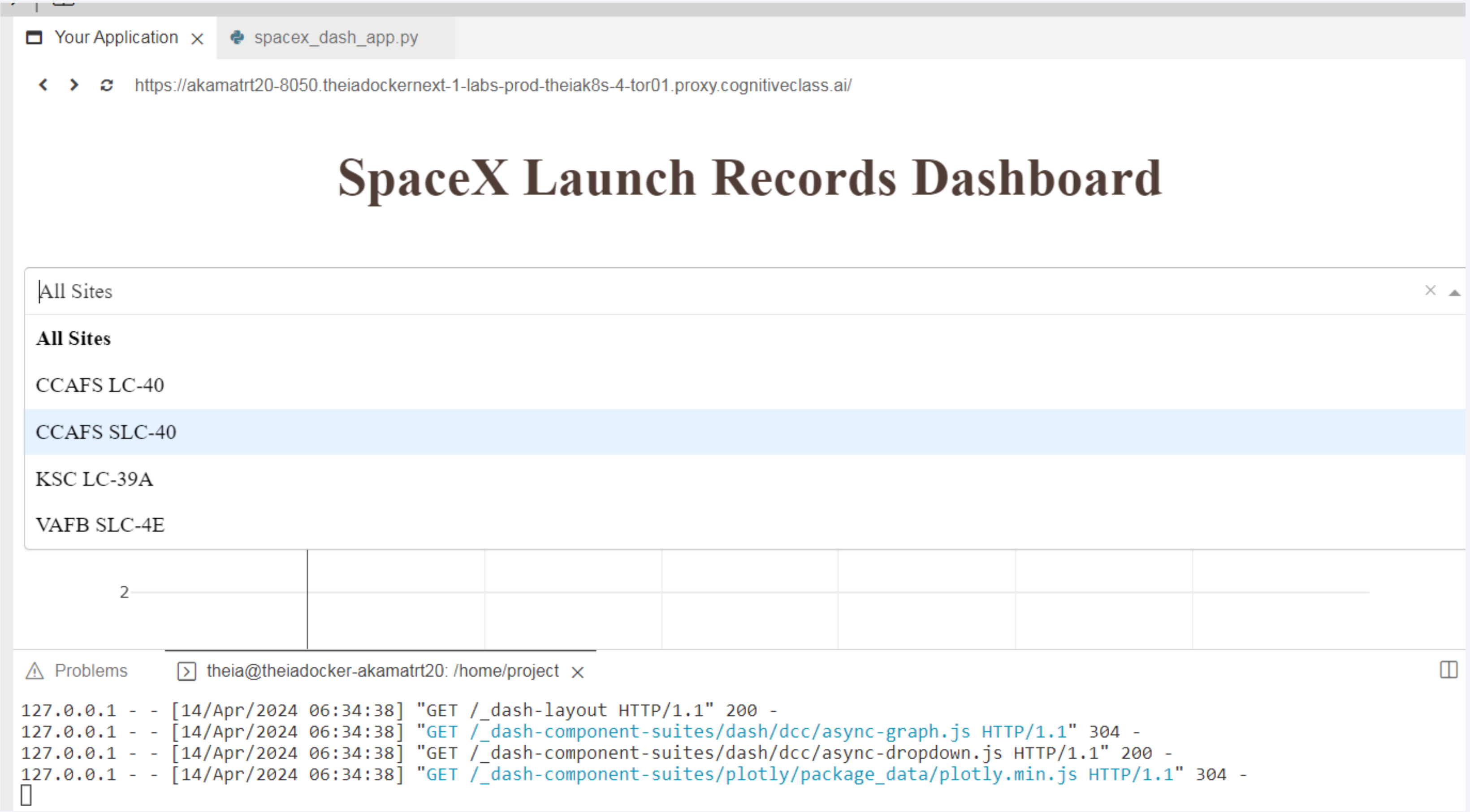


The distance from the coastal in above image is approx. 0.90km.

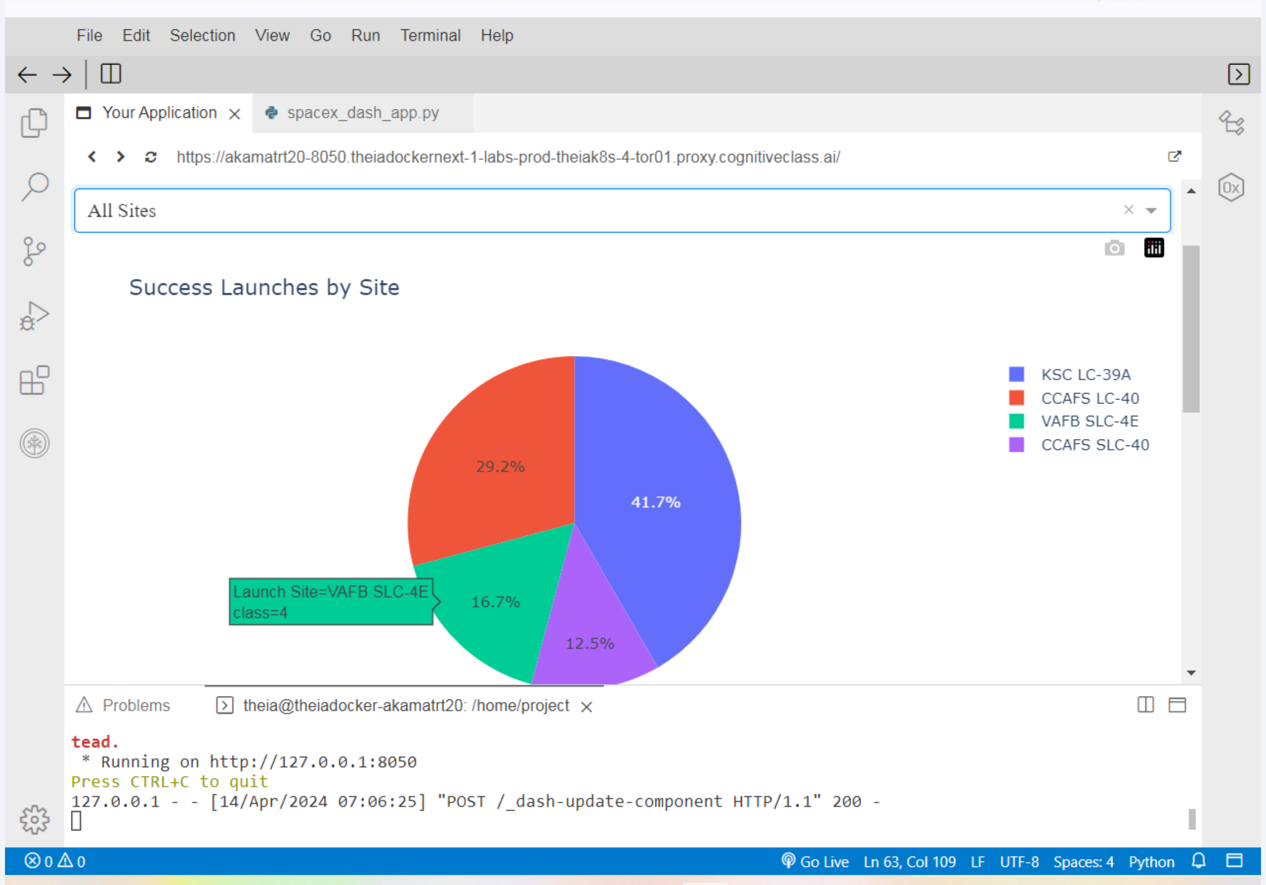
Section 4

Build a Dashboard with Plotly Dash

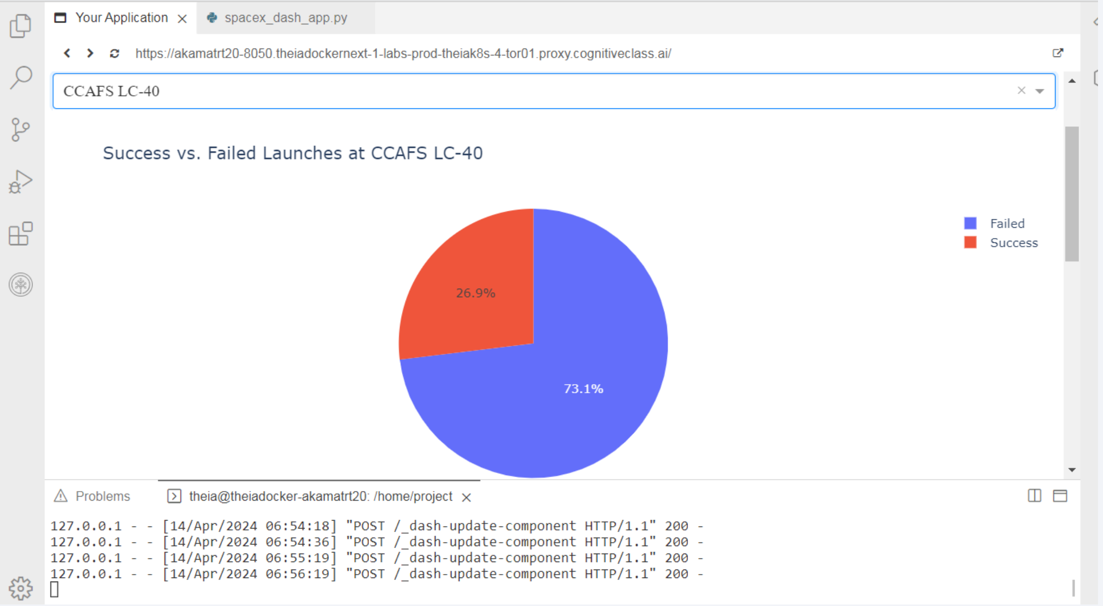
Listing all the sites



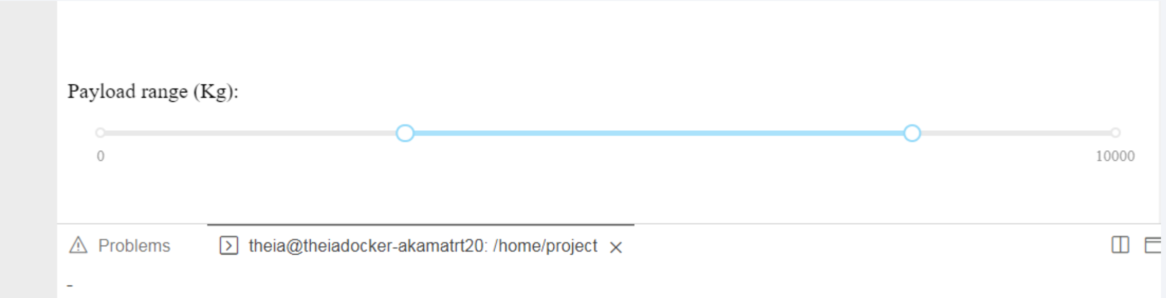
Launch success counts for all sites



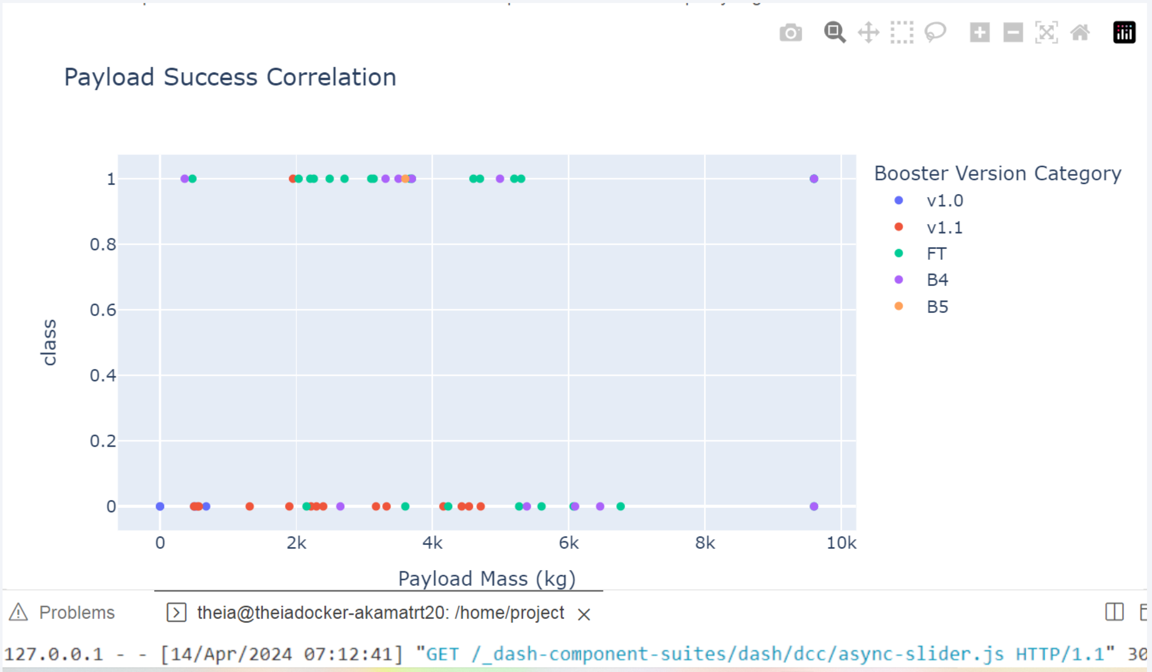
KSC LC-39 A has the highest success count.

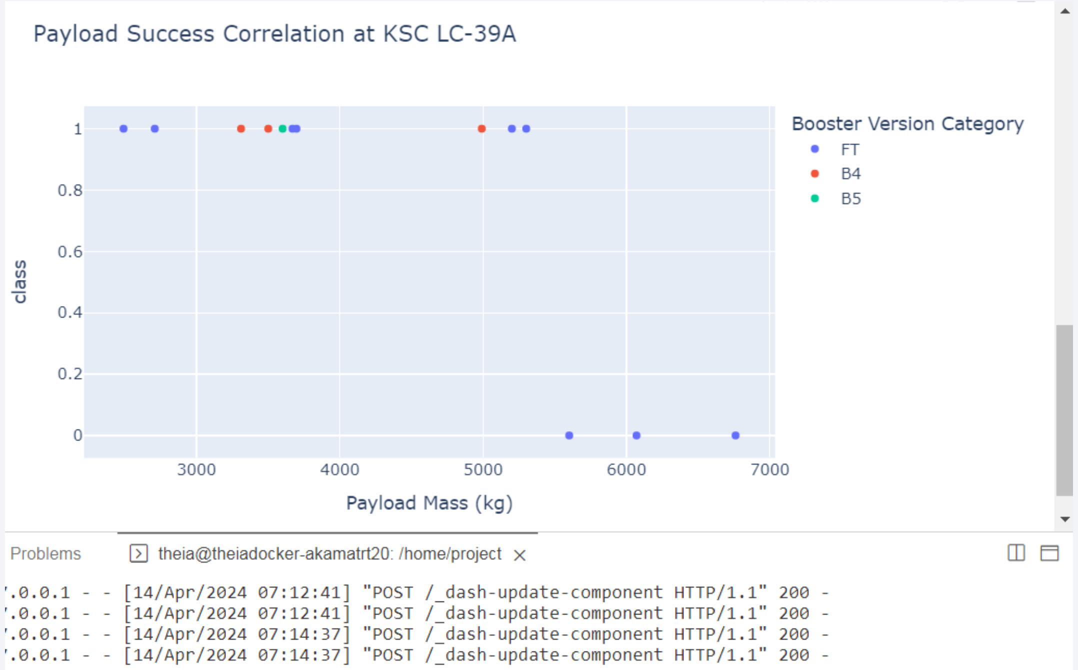


Payload Mass vs. Launch Outcome for all sites



The charts show that payloads between 2000 and 5500 kg have the highest success rate.





Section 5

Predictive Analysis (Classification)

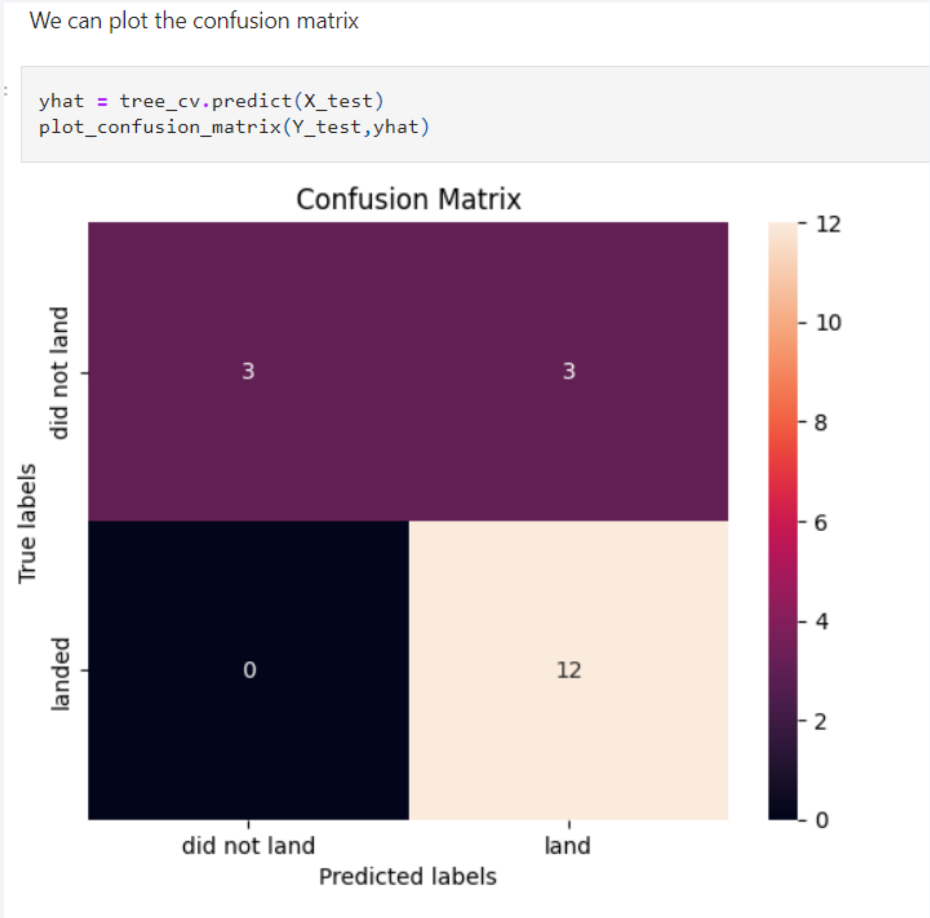
Classification Accuracy

```
[35]: models = {
          'K Nearest Neighbors': knn_cv.best_score_,
          'Decision Tree': tree_cv.best_score_,
          'Logistic Regression': logreg_cv.best_score_,
          'Support Vector': svm_cv.best_score_
      }

      best_algorithm = max(models, key=models.get)
      print('Best model is', best_algorithm, 'with a score of', models[best_algorithm])

      if best_algorithm == 'Decision Tree':
          print('Best params is:', tree_cv.best_params_)
      elif best_algorithm == 'K Nearest Neighbors':
          print('Best params is:', knn_cv.best_params_)
      elif best_algorithm == 'Logistic Regression':
          print('Best params is:', logreg_cv.best_params_)
      elif best_algorithm == 'Support Vector':
          print('Best params is:', svm_cv.best_params_)

      Best model is Decision Tree with a score of 0.8607142857142858
      Best params is: {'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'rand
      om'}
```

- Decision tree has the highest accuracy.

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.



Conclusions

- Lower payload masses correlate with higher launch success rates, suggesting the importance of optimizing payloads for efficiency.
- Launch sites tend to be situated near the Equator and coastal regions, highlighting the strategic advantage of proximity to these locations for successful launches.
- Over time, there is a noticeable improvement in launch success rates, indicating advancements in technology, processes, and understanding of spaceflight dynamics.
- KSC LC-39A stands out with the highest success rate among launch sites, underscoring the significance of site-specific factors in achieving successful launches.
- Orbits ES-L1, GEO, HEO, and SSO demonstrate a perfect success rate, suggesting robustness in launching missions to these specific orbits.

Thank you!