IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Danilo Delibašić

June 27, 2022.

Outline

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

Executive Summary

- Summary of methodologies:
    1. Data Collection – API and Web Scrapping
    2. Data Wrangling
    3. Exploratory Data Analysis (EDA) – SQL and Visualization (Folium and Dashboard)
    4. Predictive Analysis
- Summary of all results:
    1. EDA numerical results
    2. Interactive maps and dashboards
    3. Prediction results

Introduction

- Project background and context:

The main goal is to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars, while other providers cost upward of 165 million dollars. Much of the savings at SpaceX stem from the fact that they can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems we want answered:

1. What are the main features of a successful launch?
2. How can we quantify the dependence of a launch (successful or failed) upon these features, as well as the features themselves?
3. What should be the optimal values of these features, in order to maximize the probability of a launch being successful?

Section 1

Methodology

Methodology

Executive Summary

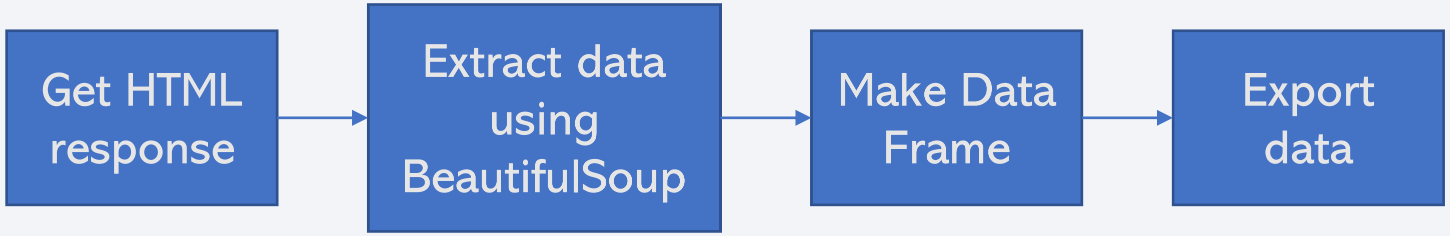
- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping Wikipedia
- Perform data wrangling:
  - One-hot encoding categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluating different classification models

Data Collection

- Data was collected in two ways:

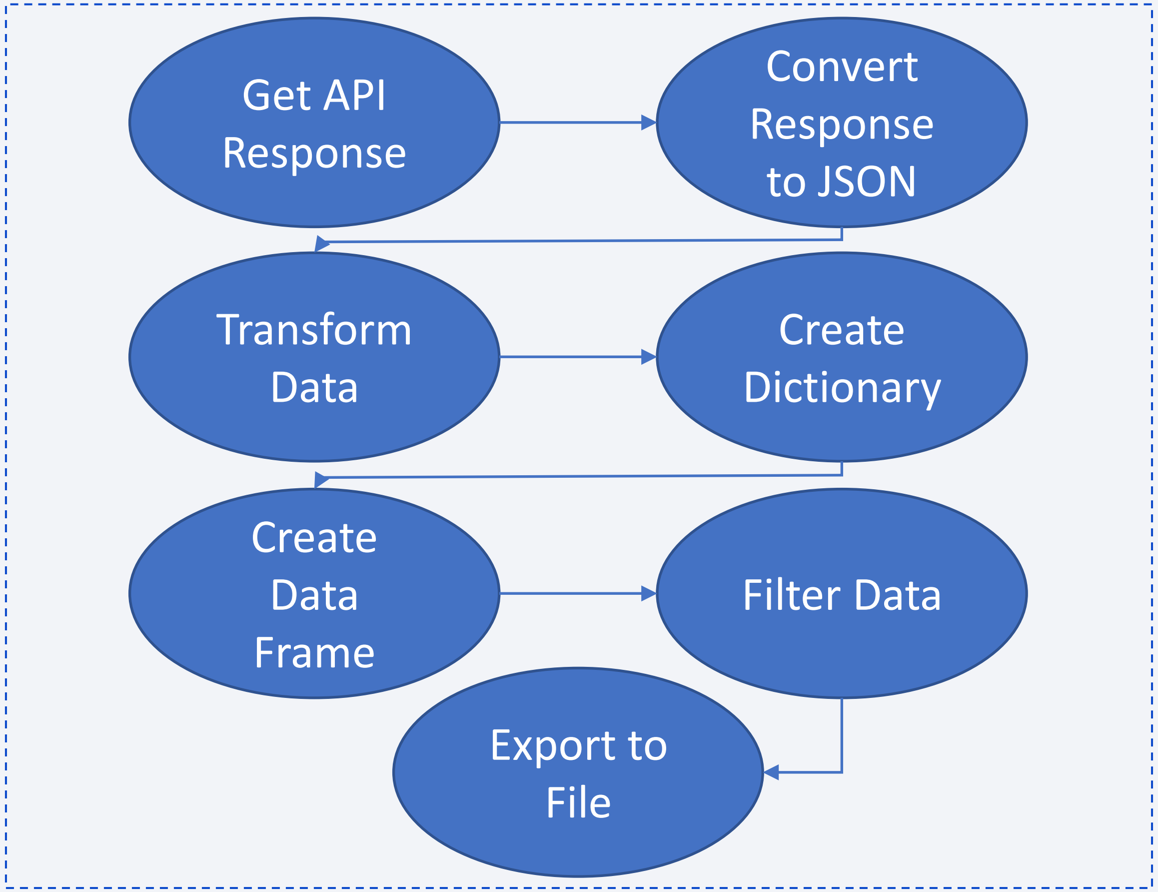
1. <u>SpaceX Rest API</u>

2. <u>Web scrapping Wikipedia</u>



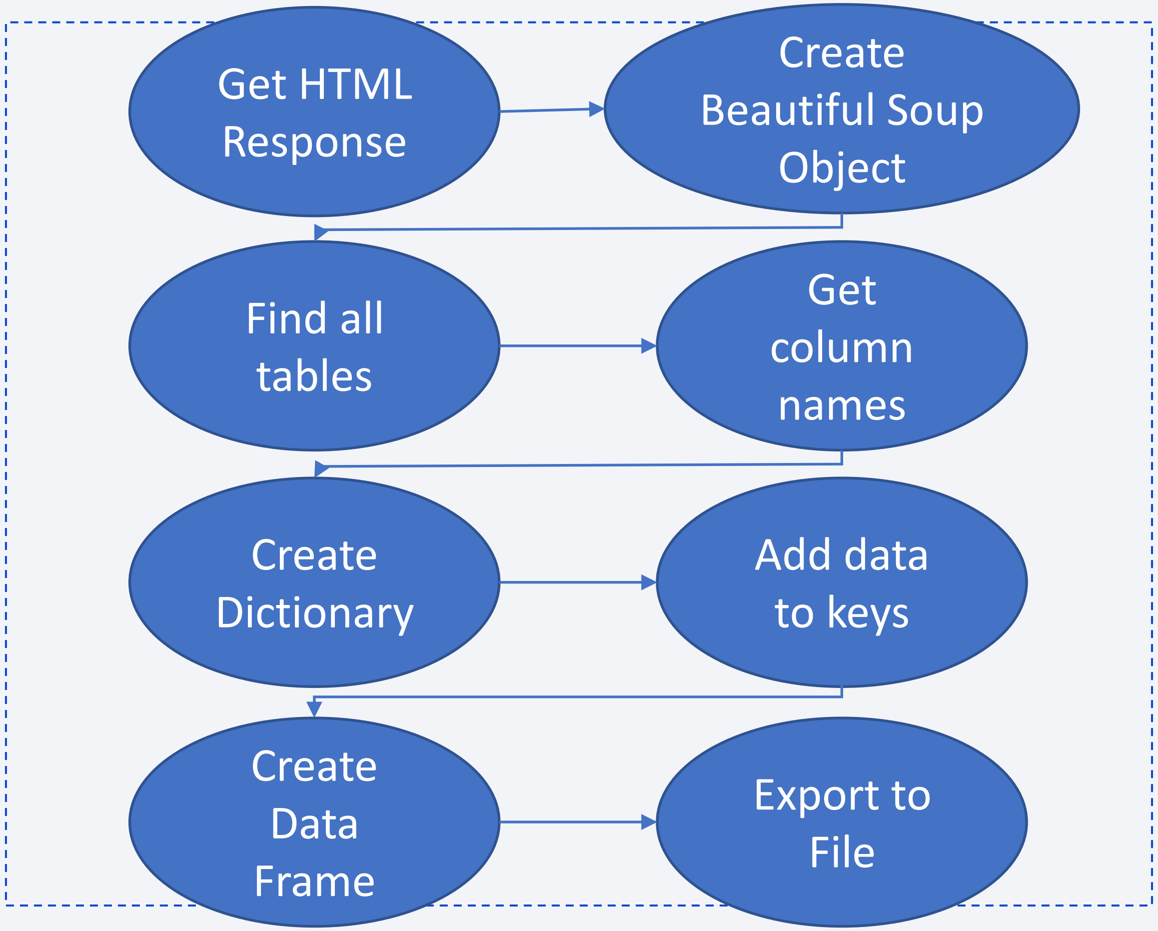
Data Collection – SpaceX API

- SpaceX REST API URL: https://api.spacexdata.com/v4/
- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/c27160020bcfd42eba11be8ab27a2187711a9212/Data%20Collection%20API.ipynb



Data Collection - Scrapping

- Wikipedia page URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/c27160020bcfd42eba11be8ab27a2187711a9212/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Some preliminary EDA was performed
- Useful insights extracted from EDA
- Categorical data (Outcome column) were one-hot encoded

- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/c27160020bcfd42eba11be8ab27a2187711a9212/Data%20Wrangling.ipynb

EDA with Data Visualization

- Scatter plots – for visualizing relationships between two chosen variables:
    1. Flight Number vs. Payload
    2. Flight Number vs. Launch Site
    3. Payload vs. Launch Site
    4. Flight Number vs. Orbit Type
    5. Payload vs. Orbit Type
- Bar plots:
    1. Success Rate vs. Orbit Type (to check how different orbit types influence the success rate)
- Line plot:
    1. Success Rate vs. Date (inspect the dependence of success rate with respect to time)
- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/c55b565245535f153db1f50667e0478f5c1ec043/EDA%20with%20Visualisation.ipynb

EDA with SQL

- The following SQL queries were performed:

    1. Display the names of the unique launch sites in the space mission
    2. Display 5 records where launch sites begin with the string 'CCA'
    3. Display the total payload mass carried by boosters launched by NASA (CRS)
    4. Display average payload mass carried by booster version F9 v1.1
    5. List the date when the first successful landing outcome in ground pad was achieved
    6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    7. List the total number of successful and failure mission outcomes
    8. List the names of the booster_versions which have carried the maximum payload mass (using a subquery)
    9. 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.
    10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/7ebba9fa642078e4b390f9ec84c968257cbb0f91/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

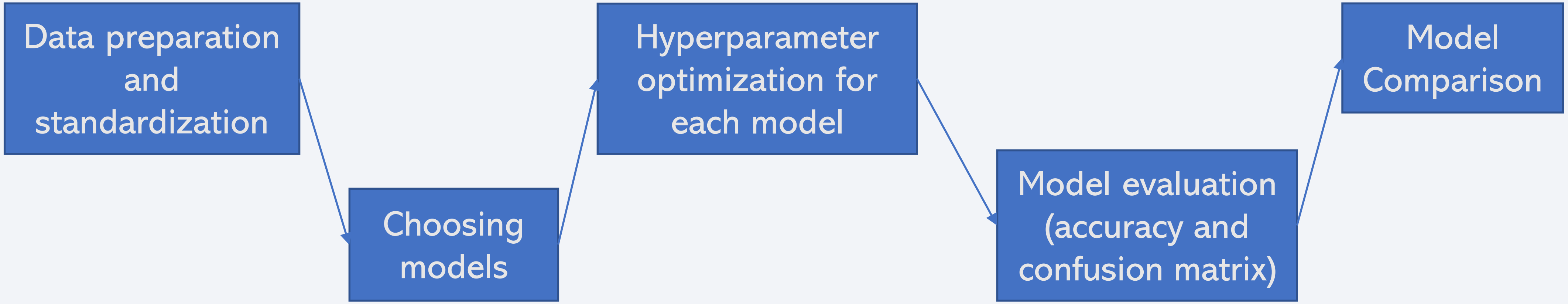
- Markers, circles, lines and marker cluster were created and added to a Folium Map:
  - Markers represent important locations, such as launch sites
  - Circles represent certain areas around important locations
  - Lines are used for indicating distance between certain points on a map
  - Marker clusters represent certain important events at a given location, such as launches
- These objects were added to better visualize the problem and potentially gain some useful insight while analyzing the geographical aspects of the problem at hand
- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/c55b565245535f153db1f50667e0478f5c1ec043/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The created dashboard contains:
  1. Dropdown – for choosing a launch site
  2. Pie chart – for showing the success/failure ratio for each launch site
  3. Rangeslider – for choosing a payload mass
  4. Scatter chart – for showing the relationship between the success rate and payload mass
- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/c55b565245535f153db1f50667e0478f5c1ec043/Dashboard%20Application%20with%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

- Four different models were developed in order to determine the one which would be the best predictor of launch success
- KNN, SVM, Logistic Regression and Decision Tree models were employed



- GitHub Notebook URL: https://github.com/DaniloDel/IBM-Applied-Data-Science-Capstone-Project/blob/d65370231e3ec4bae52a77ece973897729674b85/Machine%20Learning%20Prediction.ipynb

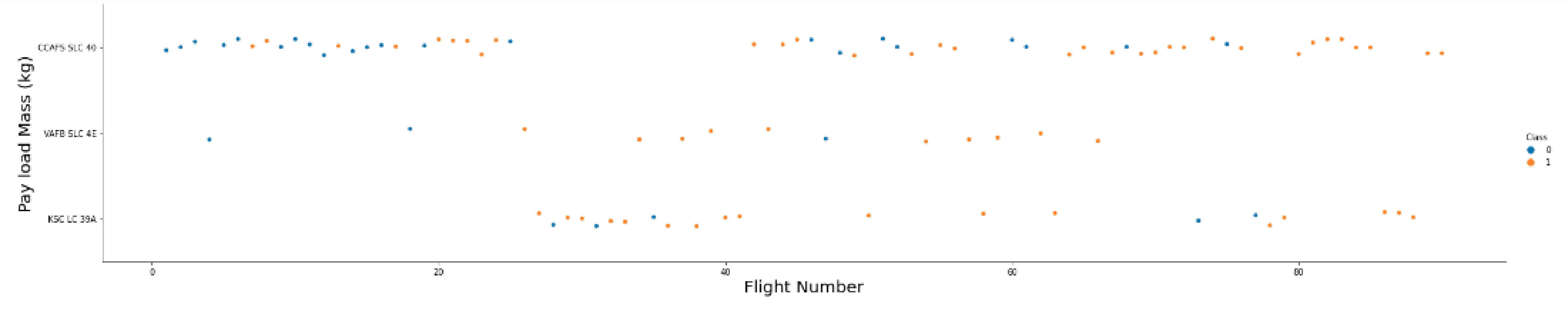
Results

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

Section 2

Insights drawn from EDA

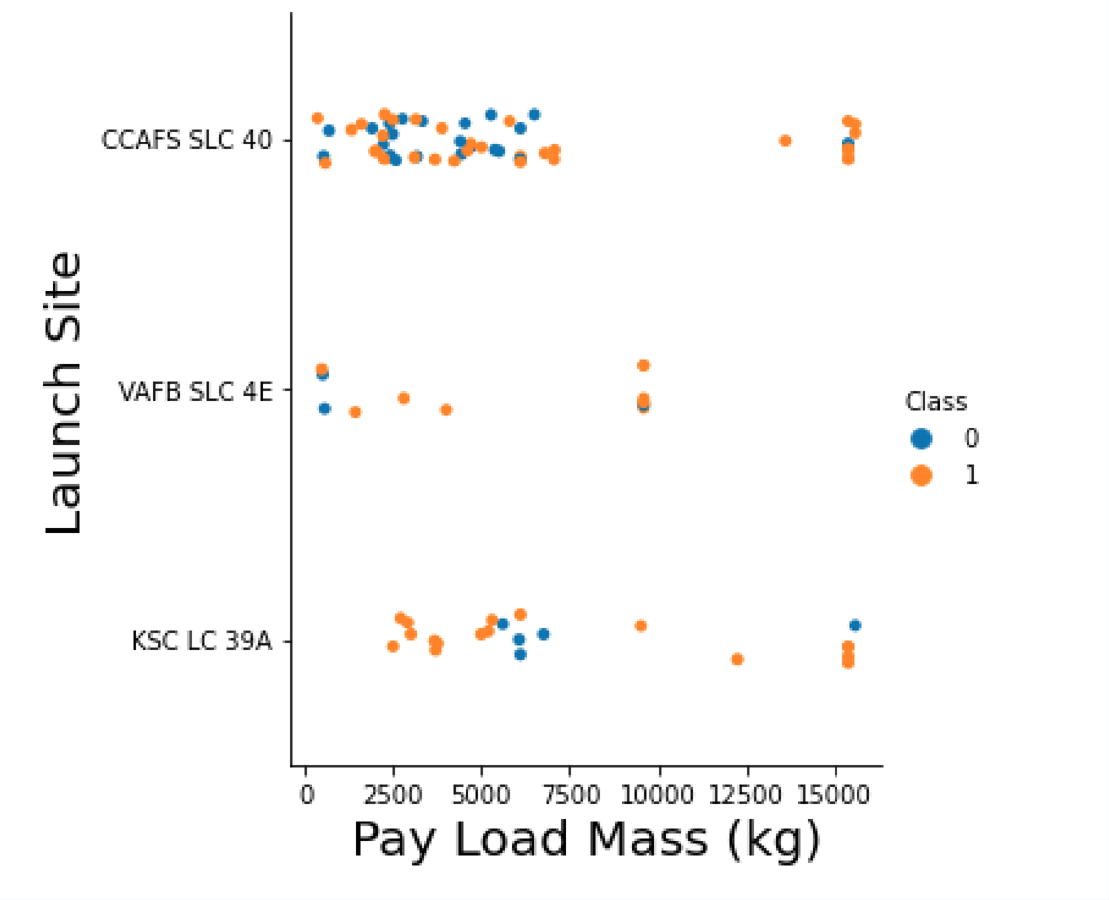
Flight Number vs. Launch Site



- The scatter plot shows an increase in the success rate with respect to the increasing flight number, for each of the displayed launch sites

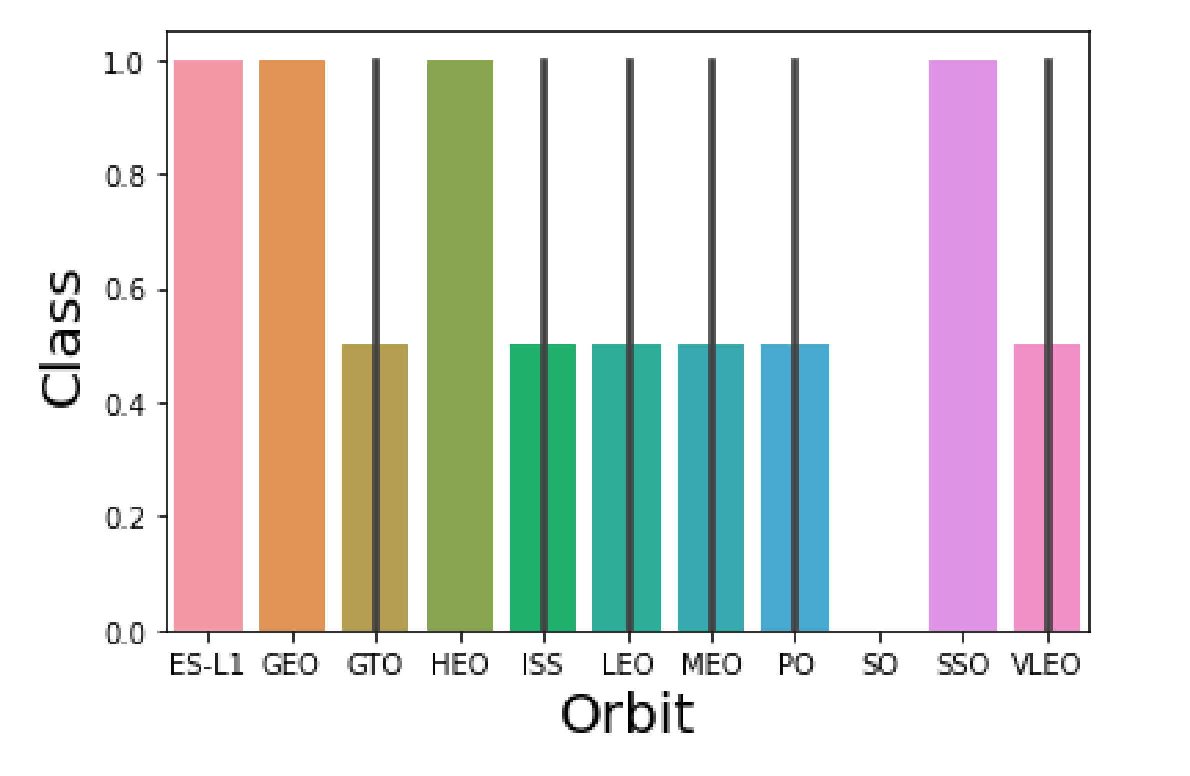
Payload vs. Launch Site

- We observe that an increase in the payload mass corresponds to an increase in the success rate for all launch sites
- The sole exception to this observation is launch site KSC LC 39A for payload mass around 6000 kg)



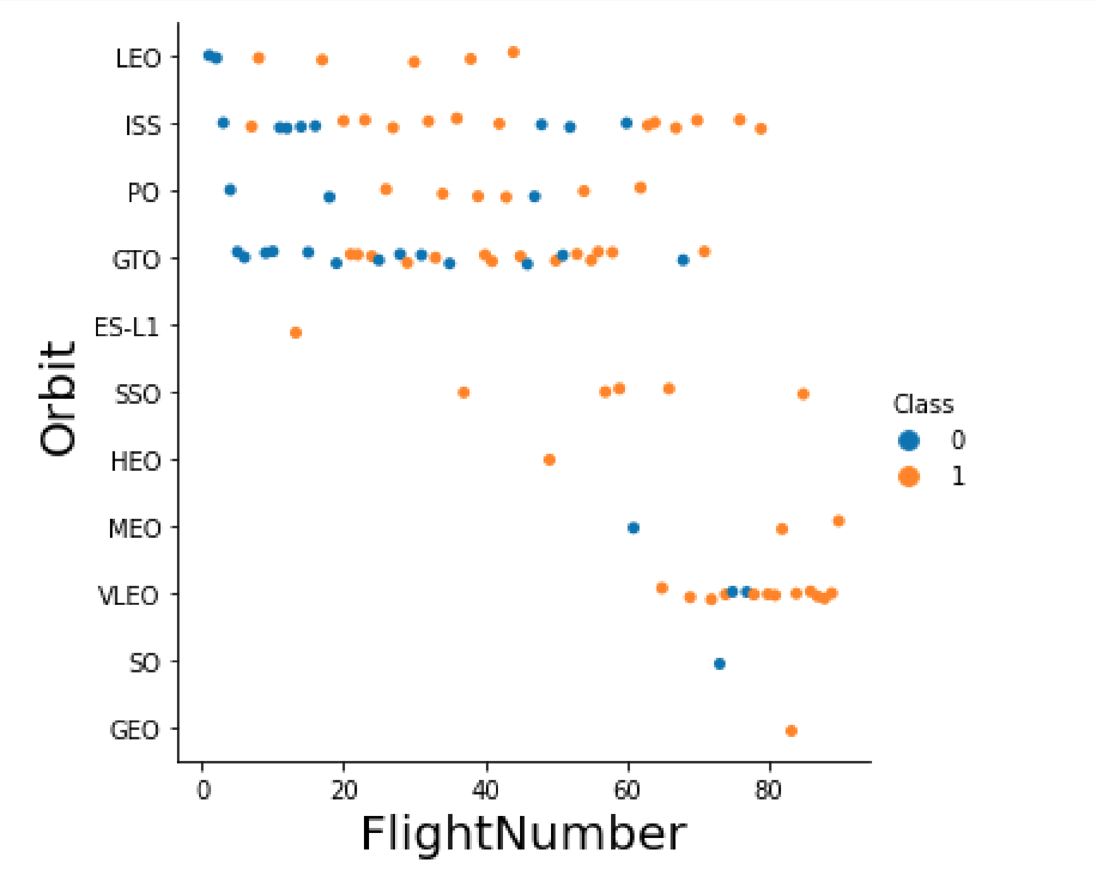
Success Rate vs. Orbit Type

- We observe significantly different success rates for different orbits
- The best success rates correspond to the following orbits: ES-L1, GEO, HEO and SSO



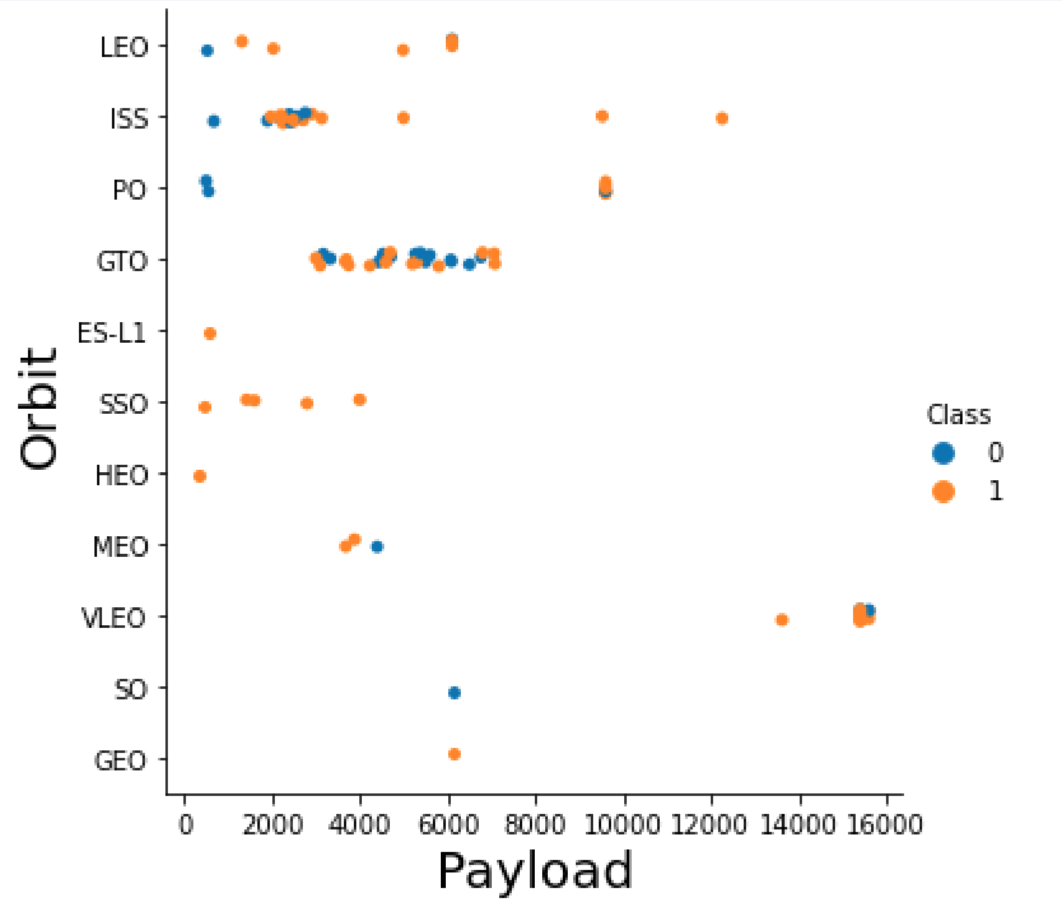
Flight Number vs. Orbit Type

- Most orbits show an increasing trend of success with rising flight number
- We can probably attribute this to an accumulated experience with each new launch, resulting in a higher success rate



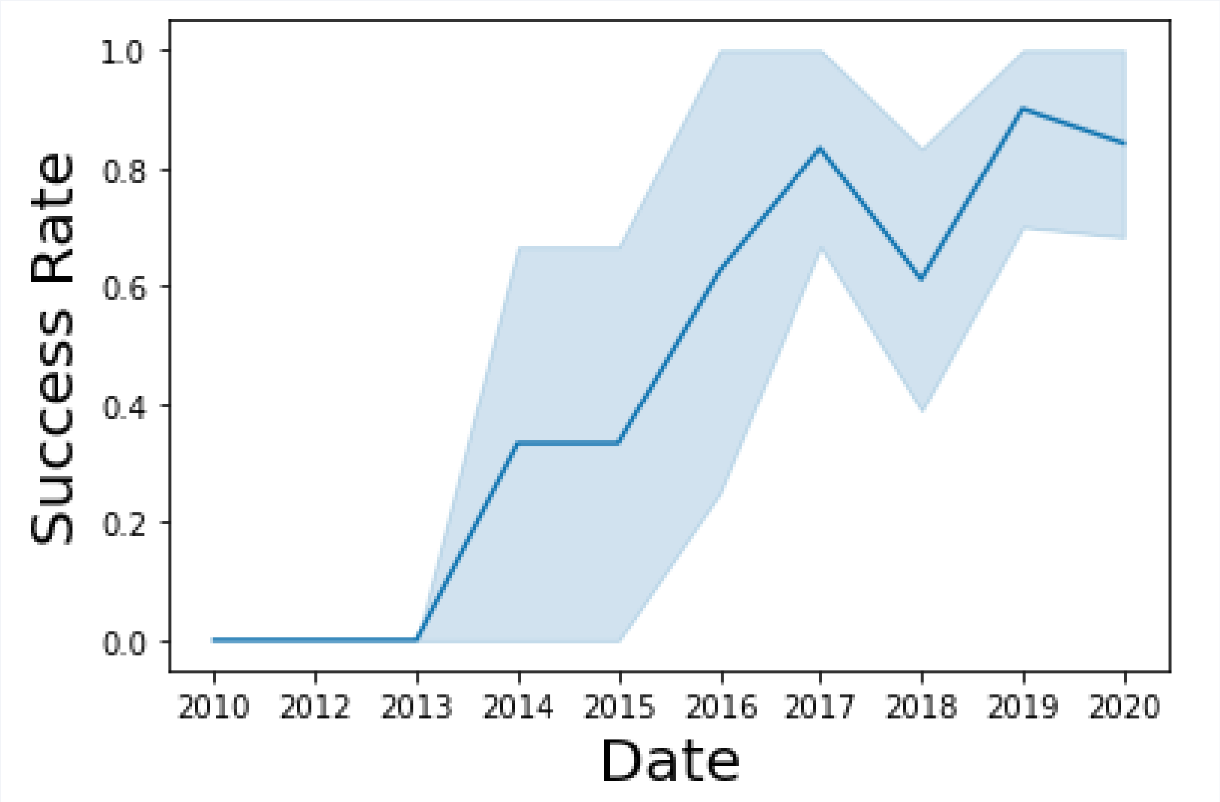
Payload vs. Orbit Type

- No clear and general relationships between the orbit type and payload mass can be observed (in most cases)
- LEO and ISS do seem to exhibit a rise in success rate with increasing payload mass



Launch Success Yearly Trend

- After an initial few years of failed launches (2010-2013), we observe a steady rise in the success rate with each following year
- The exceptions are a dip in 2018, as well as a slight decrease in 2020



All Launch Site Names

- We use DISTINCT to omit duplicates

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

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- We use LIKE ‘CCA%’ to find each entry beginning with the string ‘CCA’ (if we did not use % we would obtain only those exactly named CCA)
- We use LIMIT 5 to display only 5 entries

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

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Total Payload Mass

- Sum of payload mass only for boosters launched by NASA (CRS)

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

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- We use AVG to average out the payload mass carried by booster version F9 v1.1

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

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- We use MIN on Date to find the lowest (earliest) date when a successful landing was achieved

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

Hint:Use min function

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- We use AND in the WHERE clause to put to conditions (>4000 and <6000) for the payload mass

Total Number of Successful and Failure Mission Outcomes

- We independently extract the number of successful and failed missions by separate queries, and display them

List the total number of successful and failure mission outcomes

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful missions is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed missions is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful missions is:
```

	successoutcome
0	100

```
The total number of failed missions is:
```

	failureoutcome
0	1

Boosters Carried Maximum Payload

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

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

- We use a subquery to extract only the maximum payload mass, and the main query to extract the booster versions carrying this maximum payload mass

2015 Launch Records

- We use WHERE to extract only the correct year, as well as LIKE to filter out only failed launches

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.

```
task_9 = '''
        SELECT EXTRACT(MONTH FROM Date) AS "Month", MissionOutcome, BoosterVersion, LaunchSite
        FROM SpaceX
        WHERE EXTRACT(YEAR FROM Date)='2015'
            AND MissionOutcome LIKE '%Failure%';
        '''
create_pandas_df(task_9, database=conn)
```

	Month	missionoutcome	boosterversion	launchsite
0	6.0	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40

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

- We rank the count of successful landing outcomes in the sought interval
- We use DESC to make the order descending

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
            AND LandingOutcome LIKE '%Success%'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

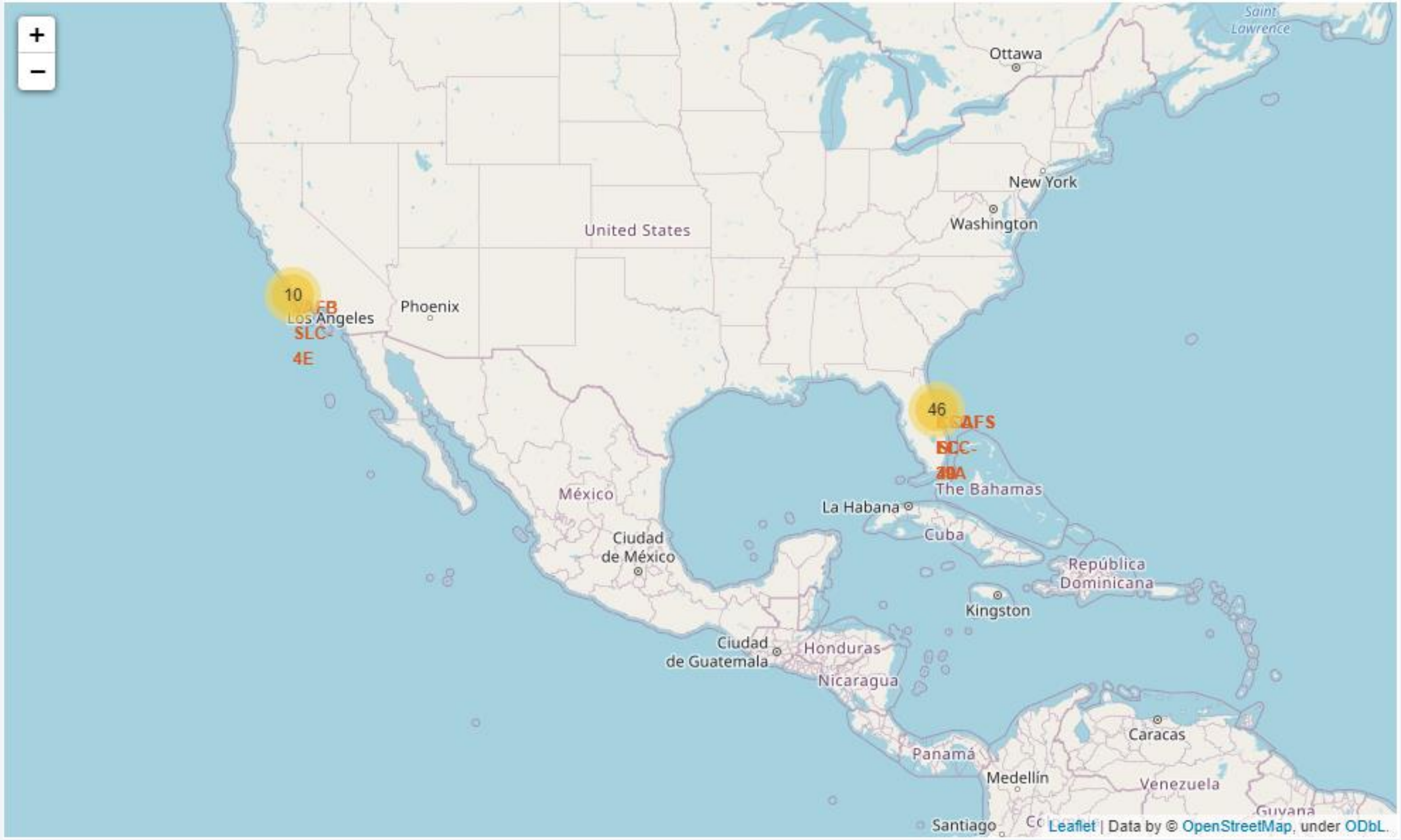
	landingoutcome	count
0	Success (drone ship)	6
1	Success (ground pad)	5

Section 3

Launch Sites Proximities Analysis

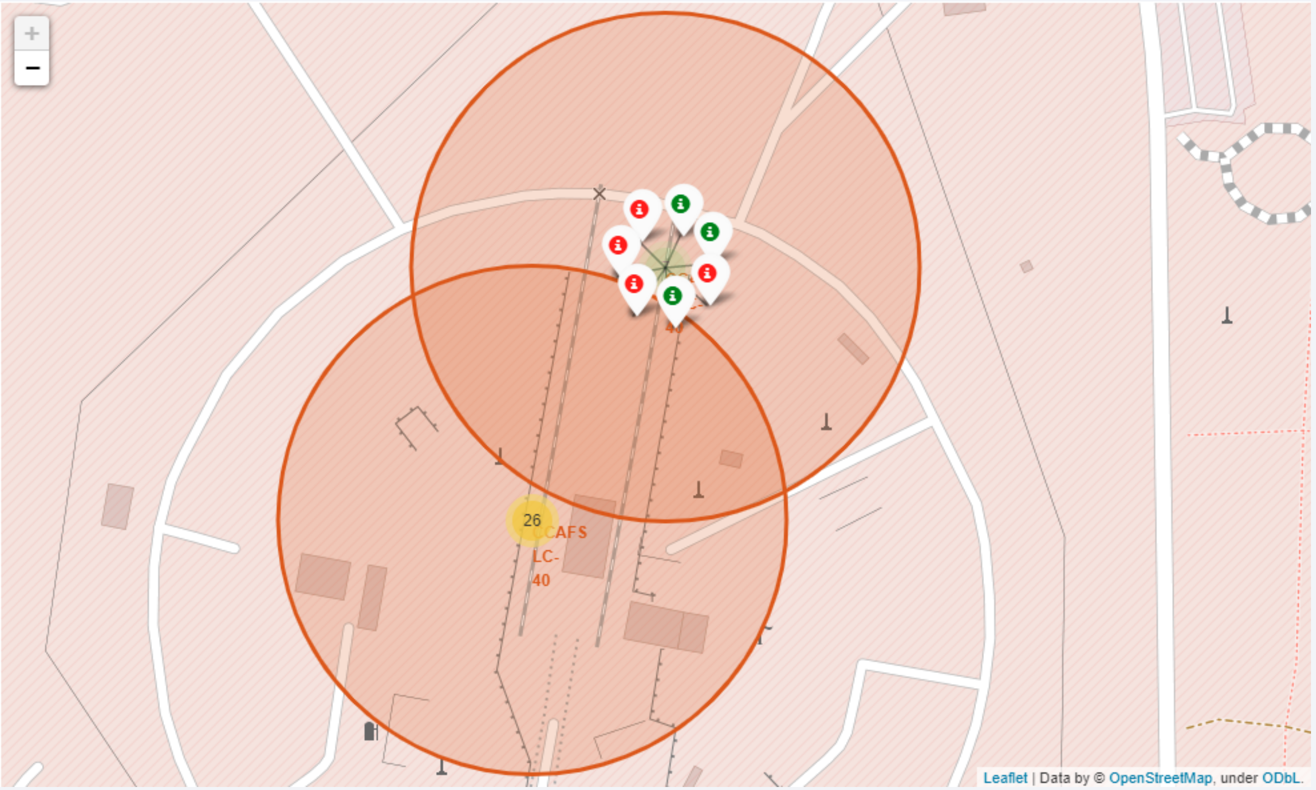
Launch Sites

- We observe that there are two clusters of launch sites located on the east and west US coast



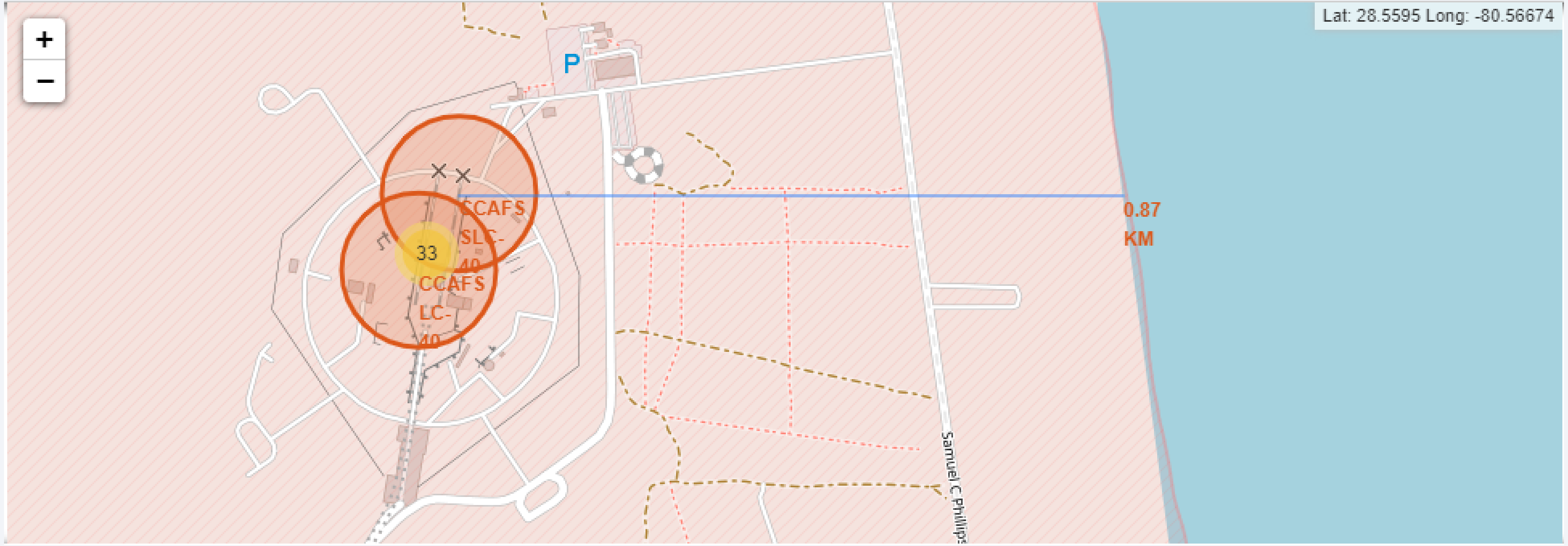
Launch Outcomes

- We these maps we can observe the number of successful/failed launches at a given loation



Nearby objects

- The blue lines and the red numerical values indicate the distance to the selected object (in this particular case, the coastline)

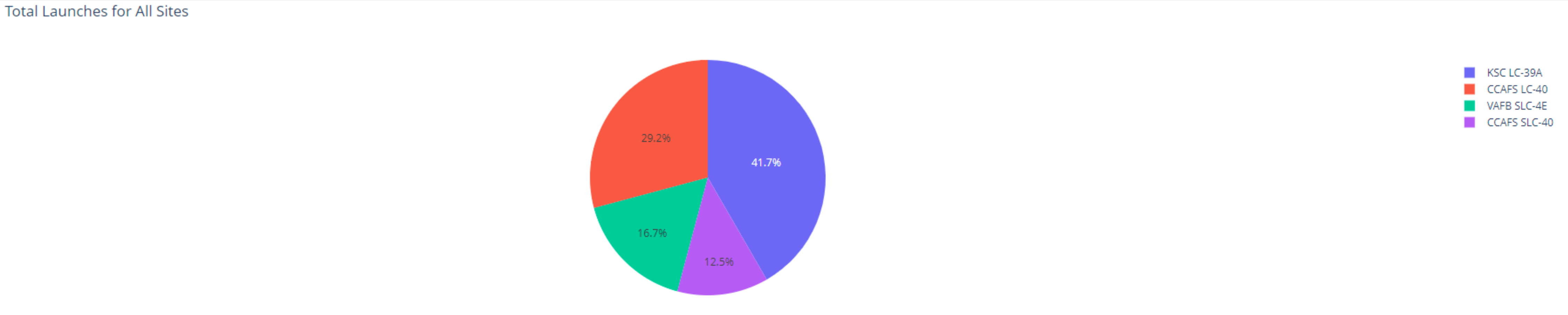


Section 4

Build a Dashboard with Plotly Dash

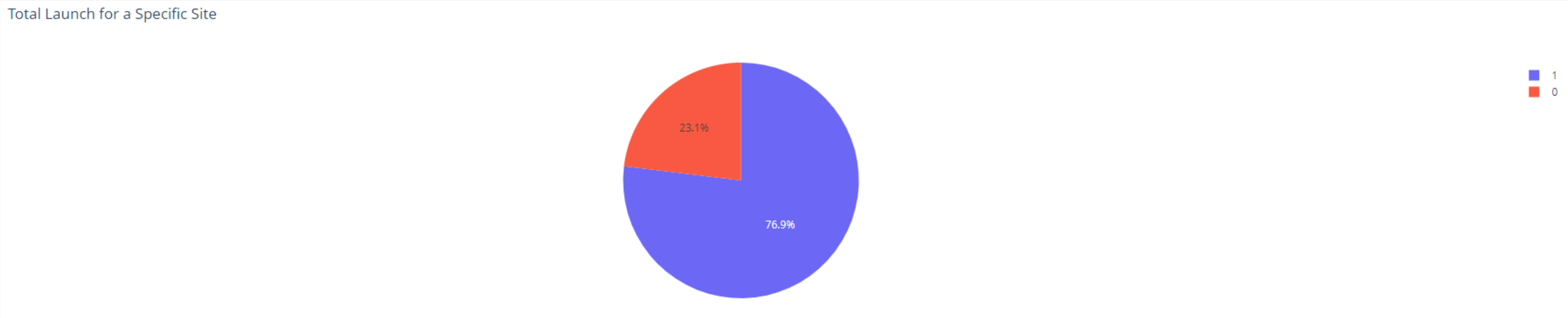
Total Launches by Site

- On this pie chart we can observe the distribution of total launches with respect to different sites



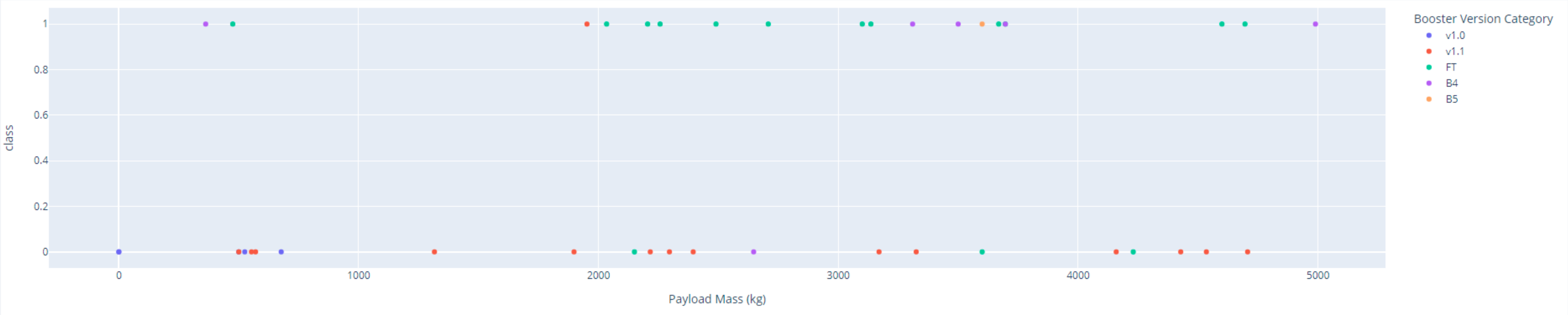
Most Successful Site

- This is the pie chart for the site with the highest success rate
- It is the KSC LC-39A site, with 76.9% successful launches

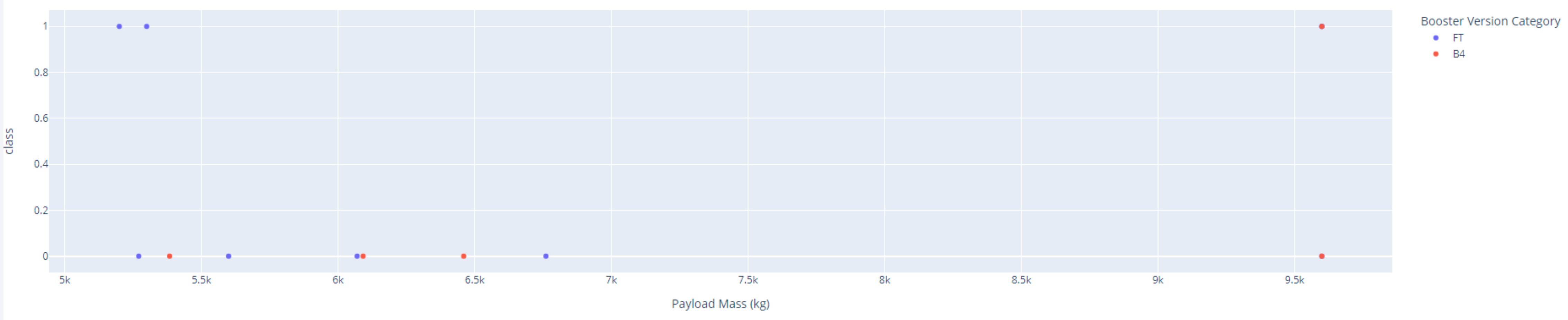


Payload vs Launch Outcome

- Payload vs. Launch Outcome for payload mass between 0 and 5000



- Payload vs. Launch Outcome for payload mass between 5000 and 10000



- Lower weight payloads seem to be more successful

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The accuracy for all built classification models can be compared via a bar chart

- The best predictor is the decision tree, with an accuracy of about 89% (for a hyperparameter optimized model)

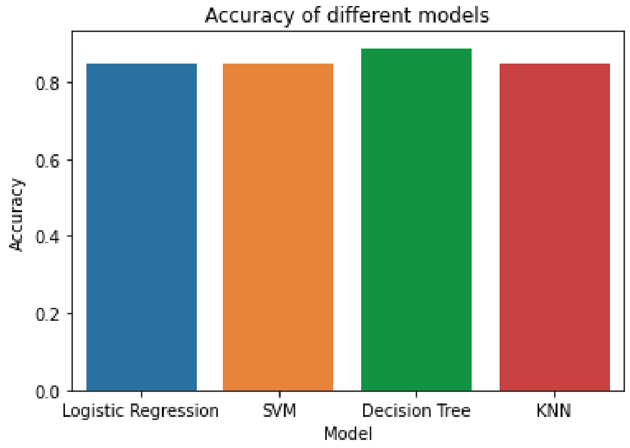
Find the method performs best:

```
scores = {'Logistic Regression': logreg_cv.best_score_, 'SVM': svm_cv.best_score_,
          'Decision Tree': tree_cv.best_score_, 'KNN': knn_cv.best_score_}
print("The best model is: ", max(scores, key=scores.get), "with the accuracy of ", max(scores.values())*100, "%")
```

```
The best model is:  Decision Tree with the accuracy of  88.75 %
```

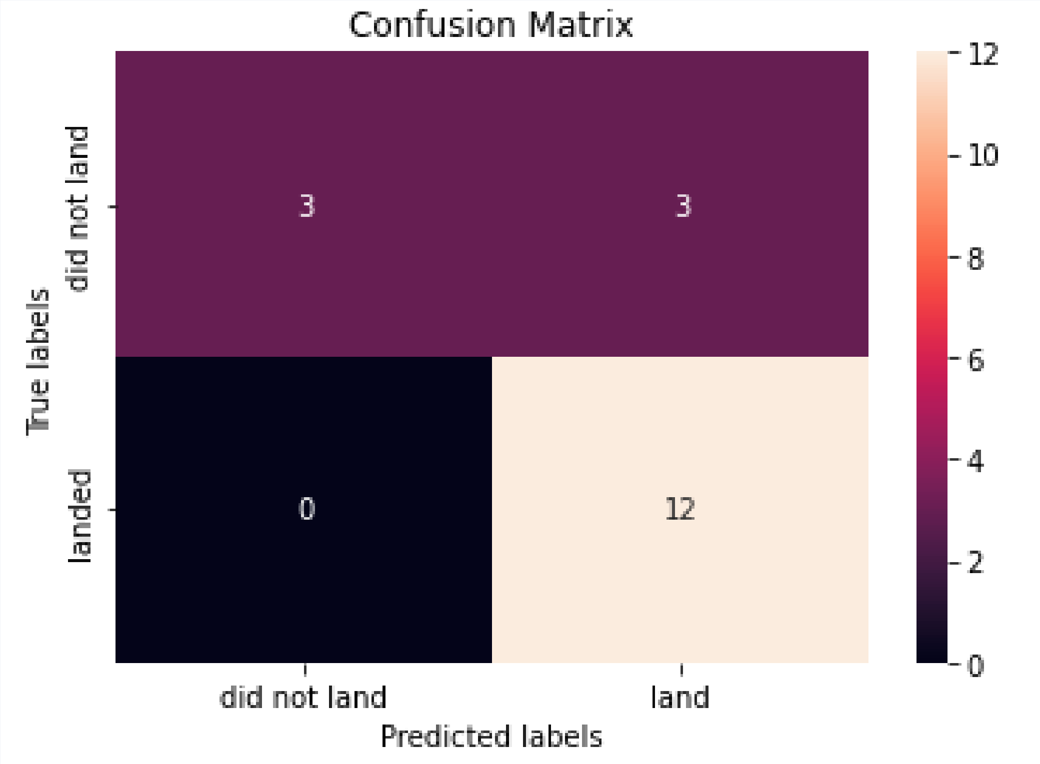
```
sns.barplot(x=list(scores.keys()), y=list(scores.values()))
plt.title("Accuracy of different models")
plt.xlabel("Model")
plt.ylabel("Accuracy")
```

```
Text(0, 0.5, 'Accuracy')
```



Confusion Matrix

- The confusion matrix of the best performing model



- We see that the model perfectly (100%) predicts true positives ("landed"), however, it predicts the true negatives ("did not land") with an accuracy of only 50%

Conclusions

- Several features were identified as important predictors for mission outcomes
- The orbits with the best success rate are GEO, HEO, SSO, ES-L1
- KSC LC-39A is the most successful site
- Most payloads above 5000 kg result in a failed mission
- We observed an improved mission outcome with time, probably due to technological advancements
- Negative outcomes (fail) are harder to predict than positive (success) ones
- The best prediction model is a Decision Tree, with an accuracy of about 89%

Thank you!