IBM DATA SCIENCE CAPSTONE_SPACEX

Mohammadmahdi Safariforoushan

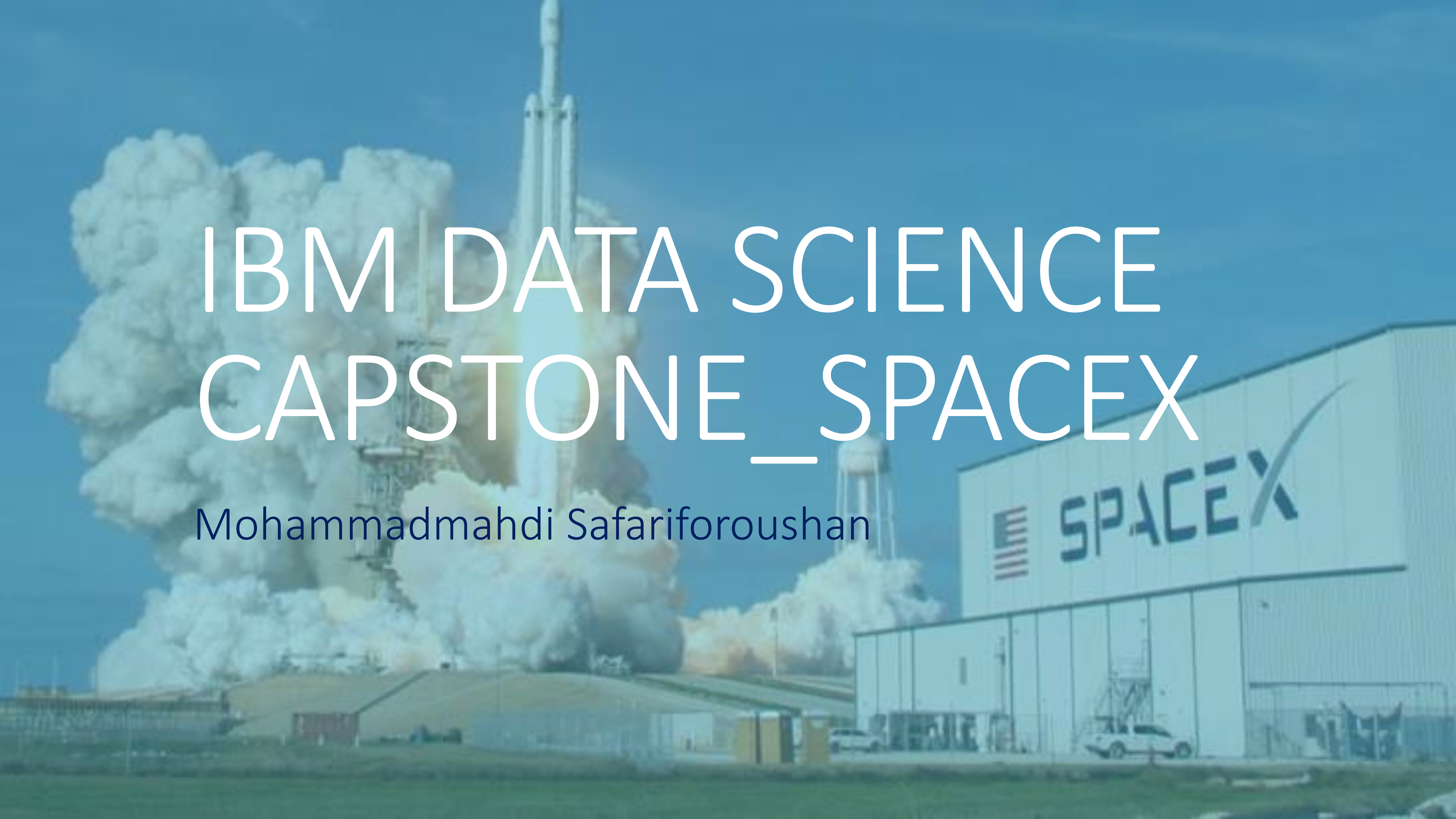


Table of contents

- Summary
- Introduction
- Methodology
- Outcomes and Results
- Conclusion

Summary

Summary of Methodology:

This research aimed to identify the factors contributing to successful rocket landings, using the following methodologies:

- Collecting data from SpaceX REST API and web scraping techniques.
- Processing the collected data to create a success/fail outcome variable.
- Exploring the data with data visualization techniques, considering factors such as payload, launch site, flight number, and yearly trends.
- Analyzing the data using SQL to calculate statistics such as total payload, payload range for successful launches, and the total number of successful and failed outcomes.
- Examining the success rates of launch sites and their proximity to geographical markers.
- Identifying the KSC LC-39A site as having the highest success rate among landing sites.
- Conducting exploratory data analysis and finding that launch success has improved over time.
- Building predictive models using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN) algorithms.

Summary of Results:

- Launch success rates have improved over time.
- Most launch sites are located near the equator and close to the coast.
- KSC LC-39A has the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.
- All of the predictive models performed similarly on the test set, with the decision tree model slightly outperforming the others.

Introduction

The space industry leader, SpaceX, is dedicated to making space travel accessible to everyone by providing affordable rocket launches. This is made possible by the reuse of the first stage of its Falcon 9 rocket, which significantly reduces launch costs compared to other providers. By determining whether the first stage will land successfully, the cost of the launch can be estimated. This study aims to use public data and machine learning models to predict the success of the first-stage landing, allowing SpaceX and competing companies to reuse the first stage efficiently. The study will explore how payload mass, launch site, number of flights, and orbits affect first-stage landing success and examine the rate of successful landings over time. Finally, the study will identify the best predictive model for successful landing through binary classification and the problems you want to find answers are:

What factors determine if the rocket will land successfully,

The interaction amongst various features that determine the success rate of a successful landing,

What operating conditions needs to be in place to ensure a successful landing program.

Methodology

Methodology

- Data collection methodology:
  - Data collected via SpaceX API and web scraping from Wikipedia
- Data wrangling:
  - Categorical features one-hot encoded
- Exploratory data analysis (EDA):
  - Visualization and SQL used
- Interactive visual analytics:
  - Folium and Plotly Dash used
- Predictive analysis:
  - Classification models employed
- Model building, tuning, and evaluation:
  - Performed to identify the optimal model.

Data Collection

- Get requests to SpaceX API were used to collect data.
- The response content was decoded as a JSON using the .json() function and transformed into a Pandas dataframe with .json_normalize().
- The data was then cleaned and missing values were checked and filled in where necessary.
- Web scraping from Wikipedia was also performed to obtain Falcon 9 launch records using BeautifulSoup.
- The launch records were extracted as an HTML table, parsed, and converted into a Pandas dataframe for future analysis.

Data Collection – SpaceX API

use the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

Here are the steps taken to obtain and prepare the rocket launch data from SpaceX API:

Request data from SpaceX API (rocket launch data)

Decode response using .json() and convert to a dataframe using .json_normalize()

Request information about the launches from SpaceX API using custom functions

Create dictionary from the data

Create dataframe from the dictionary

Filter dataframe to contain only Falcon 9 launches

Replace missing values of Payload Mass with calculated .mean()

Export data to csv file.

The link to the notebook is https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/1_Data%20_Collection_Spacex.ipynb

Data Collection – Web Scraping

apply web scrapping to webscrap Falcon 9 launch records with BeautifulSoup and parse the table and converte it into a pandas dataframe.

Here are the steps:

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create dataframe from the dictionary
- Export data to csv file

The link to the notebook is : https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/2_Web_Scraping_Spacex.ipynb

Data Wrangling

In order to gain insights and determine the training labels, we conducted an exploratory data analysis. This involved calculating the number of launches that took place at each site, as well as the frequency and occurrence of each orbit. Additionally, we created a landing outcome label from the outcome column and exported the resulting data to a CSV file.

The link to the notebook is: https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/3_Data_Wrangling_Spacex.ipynb

EDA with Data Visualization

By visualizing the flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, and the yearly trend of launch success, we conducted an analysis of the data

Charts

Flight Number vs. Payload

Flight Number vs. Launch Site

Payload Mass (kg) vs. Launch Site

Payload Mass (kg) vs. Orbit type

EDA with Visualization

Analysis

By utilizing scatter plots, examine the correlation between variables. If a correlation is found, these variables may be valuable for machine learning purposes.

Employ bar charts to compare discrete categories. Bar charts display the connections between categories and their corresponding measured values.

The link to the notebook is : https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/5_Eda_Data_Visual_Spacex.ipynb

EDA with SQL

- utilized SQL queries to retrieve data from the dataset, in response to various inquiries we received. These inquiries prompted us to seek specific information within the dataset, which we obtained through the use of SQL queries.
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

The link to the notebook is : https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/4_Eda_Sql_Spacex.ipynb

Interactive Map With Folium

- Labeled All Launch Sites And Incorporated Various Map Objects, Such As Markers, Circles, And Lines, On The Folium Map To Indicate The Success Or Failure Of Launches At Each Site.
- Classified The Launch Outcomes As Either Success Or Failure, Assigning A Value Of 1 To Success And 0 To Failure.
- By Utilizing Color-coded Marker Clusters, We Identified The Launch Sites With A Relatively High Success Rate.
- Calculated The Distances Between Each Launch Site And Its Surrounding Areas To Answer Specific Questions, Such As:
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

The link to the notebook is : https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/6_Visual_Analytics_Folium_Spacex.ipynb

Build a Dashboard with Plotly Dash

- Built An Interactive Dashboard With Plotly Dash
- Plotted Pie Charts Showing The Total Launches By A Certain Sites
- Plotted Scatter Graph Showing The Relationship With Outcome And Payload Mass (Kg) For The Different Booster Version.

The link to the notebook is : https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/7_Visual_Analytics_Plotly_Spacex.ipynb

Predictive Analysis (Classification)

BUILDING MODEL

Load our dataset into NumPy and Pandas

- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

EVALUATING MODEL

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

IMPROVING MODEL

- Feature Engineering
- Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

- The model with the best accuracy score wins the best performing model

The link to the notebook is: https://github.com/m-mahdisafari/spacexproject/blob/c0226fb73df7d7b62960815f729a4658644cf77f/8_Predictive_Analytics___Spacex.ipynb

Outcomes and Results

Exploratory Data Analysis Results

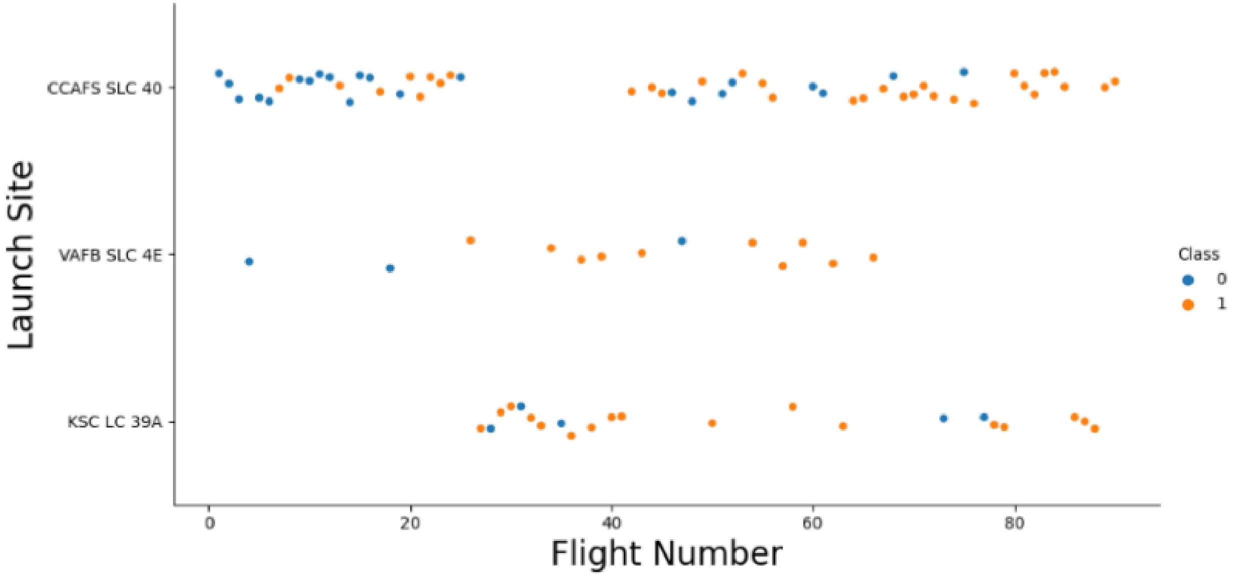
Interactive Analytics Results

Predictive Analysis Results

EDA With Visualization

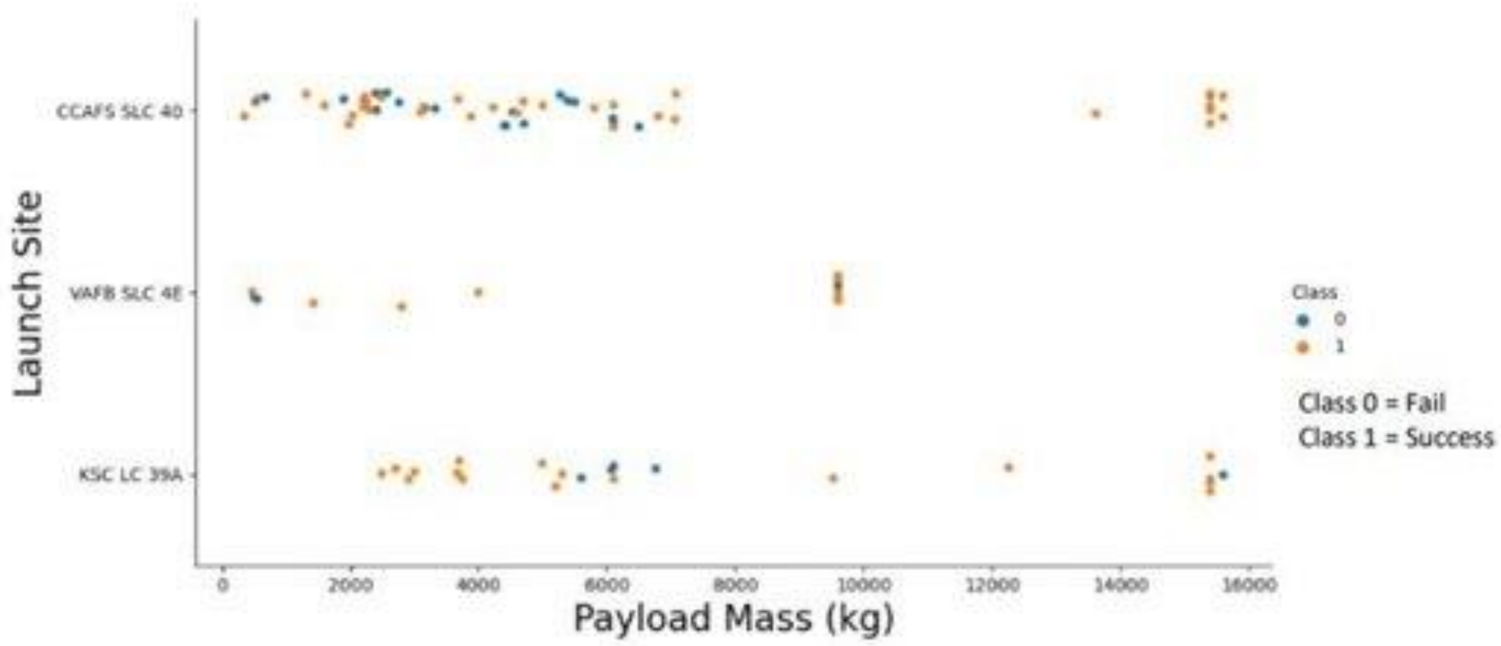
Flight Number vs. Launch Site

There is a positive correlation between the number of flights at a launch site and its success rate.



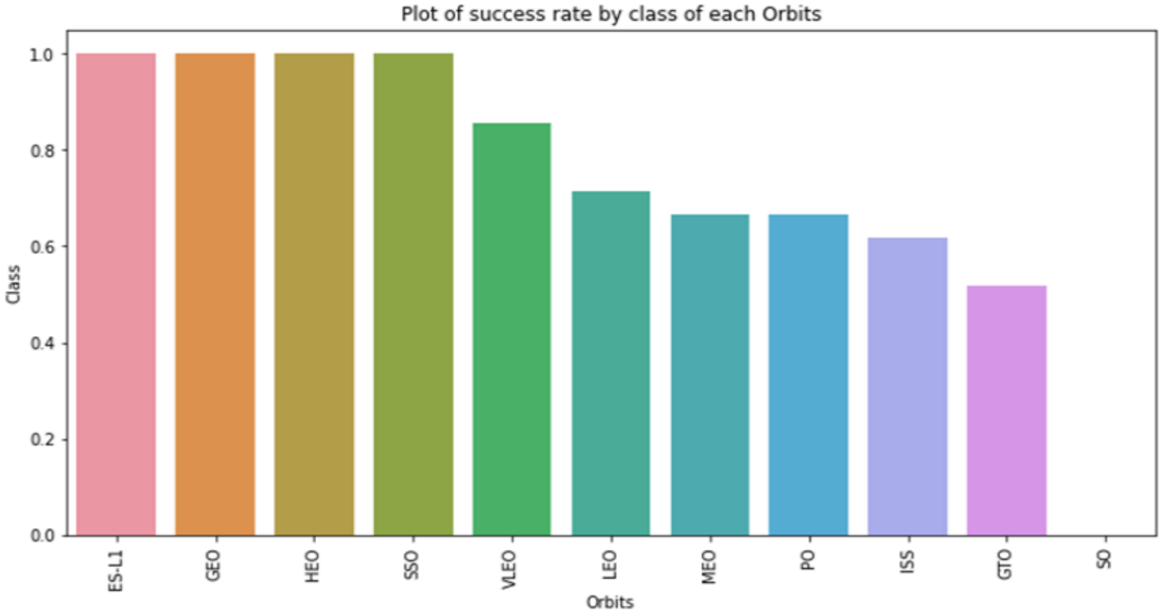
Payload Mass vs. Launch Site

Although there is a positive relationship between the payload mass and success rate for rockets launched from Launch Site CCAFS SLC 40, the visualization does not provide a clear indication of whether the launch site's success is dependent on the payload mass.



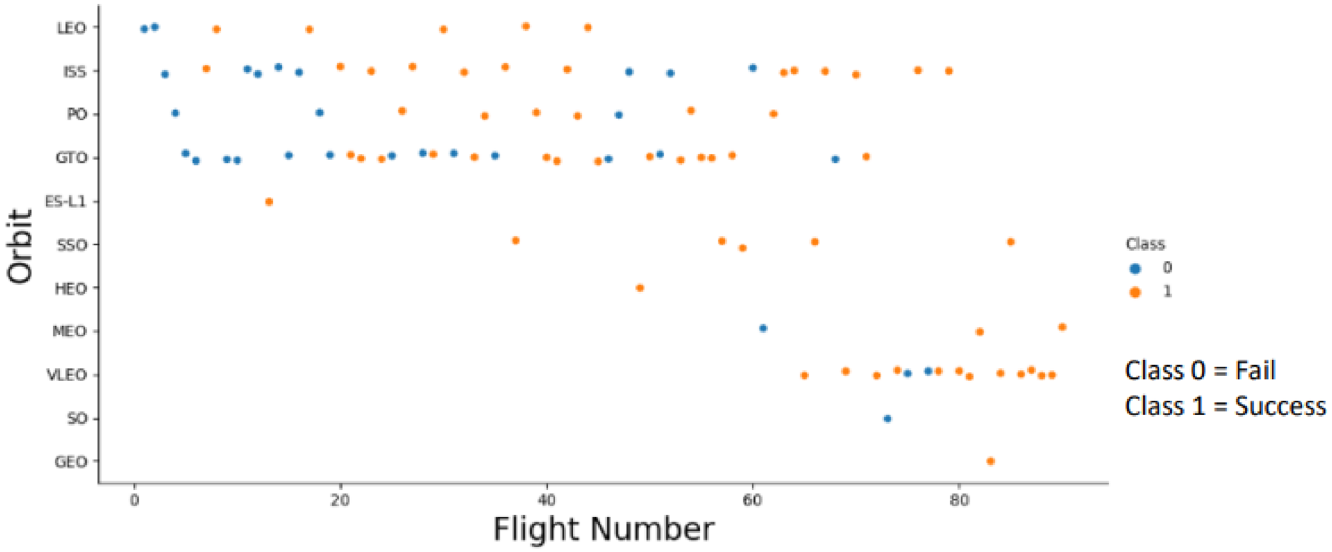
Success rate vs. Orbit type

Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate



Flight Number vs. Orbit type

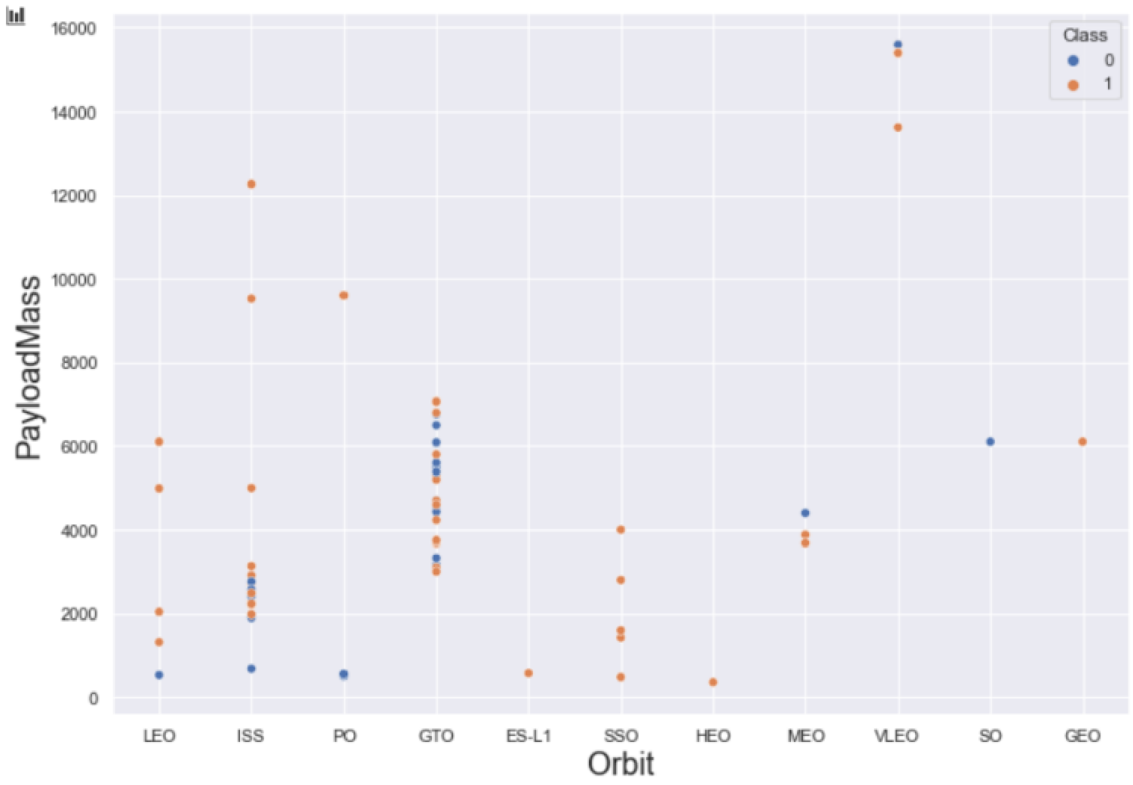
the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Class 0 = Fail
Class 1 = Success

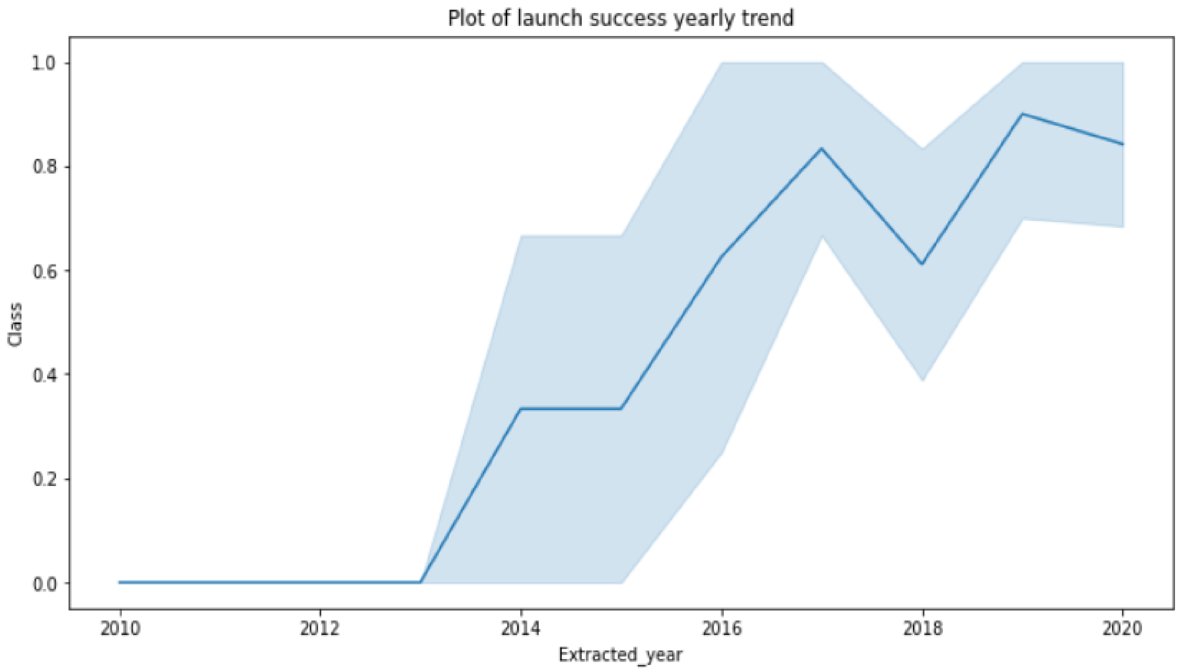
Payload Vs. Orbit Type

Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits



Launch Success Yearly Trend

success rate since 2013 kept on increasing till 2020.

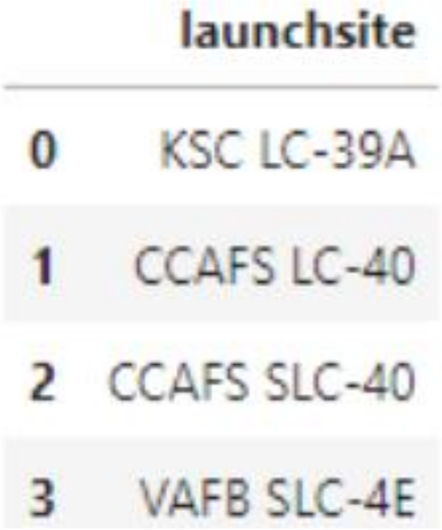


EDA with SQL

All Launch Site Names

SQL QUERY : SELECT DISTINCT Launch_Site FROM tblSpaceX

EXPLAINATION : Using the word DISTINCT in the query means that it will only show Unique values in the Launch_Site column from tblSpaceX

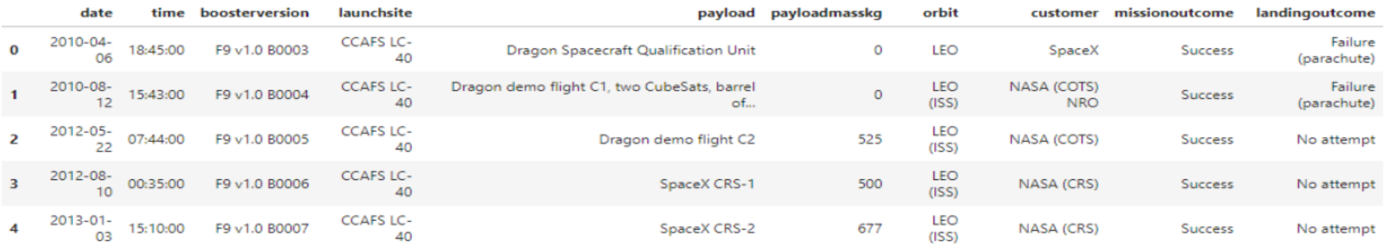


	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch site names begin with `CCA`

SQL QUERY : select TOP 5 * from tblSpaceX WHERE Launch_Site LIKE 'KSC%'

EXPLANATION : Using the word TOP 5 in the query means that it will only show 5 records from tblSpaceX and LIKE keyword has a wild card with the words ‘KSC%’ the percentage in the end suggests that the Launch_Site name must start with KSC.

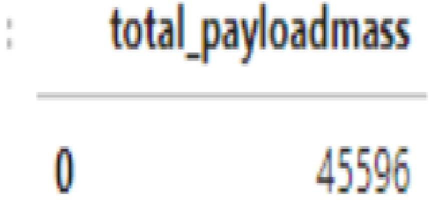


	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass by Customer NASA (CRS)

SQL QUERY : select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass from tblSpaceX where Customer = 'NASA (CRS)'"",'TotalPayloadMass

EXPLANATION : Using the function SUM summates the total in the column PAYLOAD_MASS_KG_ The WHERE clause filters the dataset to only perform calculations on Customer NASA (CRS)



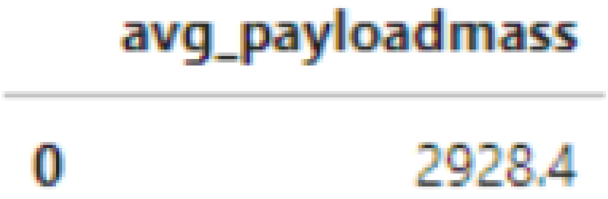
```
     total_payloadmass
0                45596
```

Average Payload Mass carried by booster version F9 v1.1

SQL QUERY : select AVG(PAYLOAD_MASS_KG_)

AveragePayloadMass from tblSpaceX where Booster_Version = 'F9 v1.1

EXPLANATION : Using the function AVG works out the average in the column PAYLOAD_MASS_KG_ The WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1

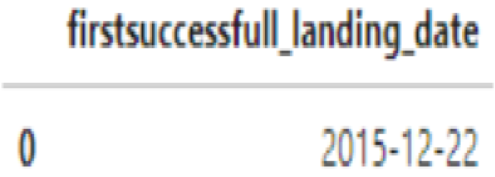


	avg_payloadmass
0	2928.4

The date where the successful landing outcome in drone ship was achieved

SQL QUERY : select MIN(Date) SLO from tblSpaceX where Landing_Outcome = ”Success (drone ship)”

EXPLANATION : Using the function MIN works out the minimum date in the column Date The WHERE clause filters the dataset to only perform calculations on Landing_Outcome Success (drone ship)

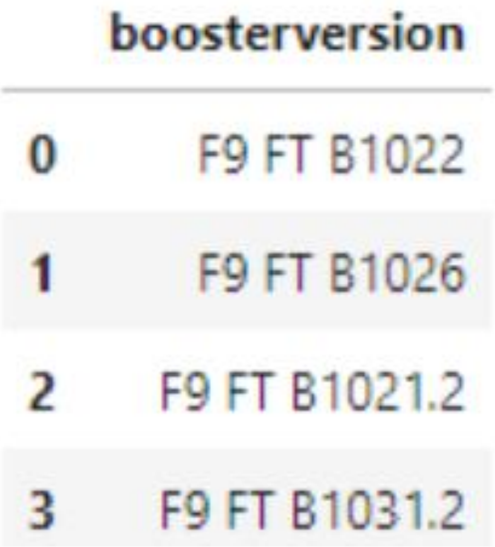


	firstsuccessfull_landing_date
0	2015-12-22

Successful drone ship landing with payload between 4000 and 6000

select Booster_Version from tblSpaceX where Landing_Outcome = 'Success (ground pad)’ AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

Selecting only Booster_Version The WHERE clause filters the dataset to Landing_Outcome = Success (drone ship) The AND clause specifies additional filter conditions Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

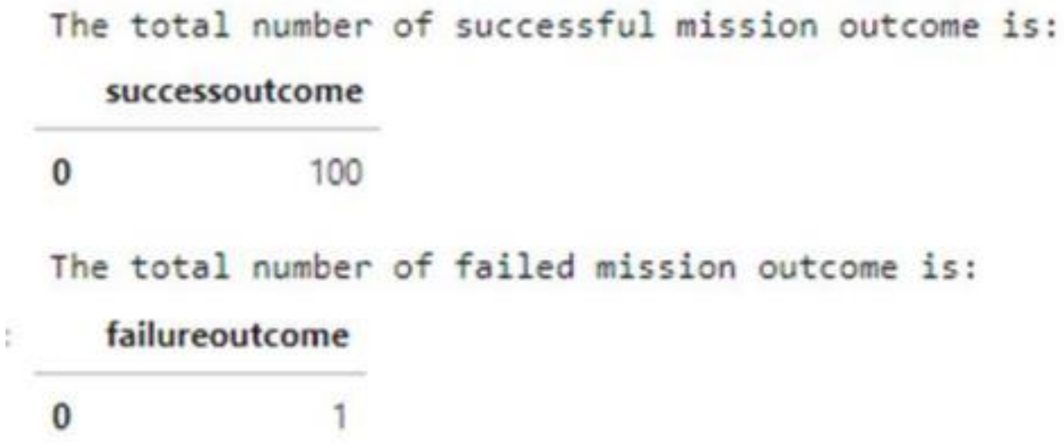


	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SELECT(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Success%') as Successful_Mission_Outcomes, (SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Failure%') as Failure_Mission_Coutcomes

PHRASE "(Drone Ship was a Success)" LIKE '%Success%' Word 'Success' is in the phrase the filter will include it in the dataset

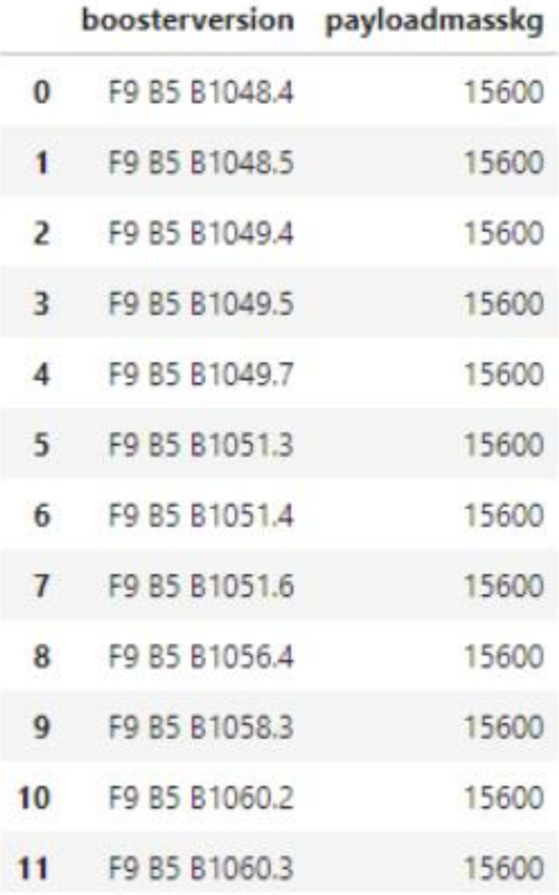


```
The total number of successful mission outcome is:

     successoutcome
0               100

The total number of failed mission outcome is:

     failureoutcome
0                 1
```

Boosters Carried Maximum Payload

SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS _KG_) AS [Maximum Payload Mass] FROM tblSpaceX GROUP BY Booster_Version ORDER BY [Maximum Payload Mass] DESC

Using the word DISTINCT in the query means that it will only show Unique values in the Booster_Version column from tblSpaceX GROUP BY puts the list in order set to a certain condition. DESC means its arranging the dataset into descending order

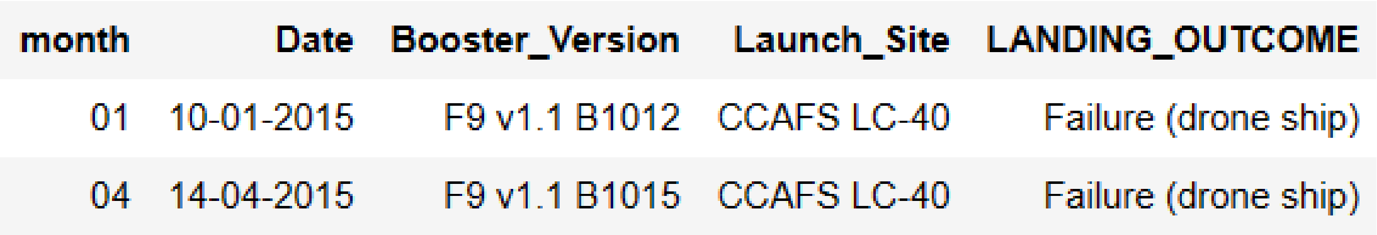


	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

Launch Records

```
SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION,
LAUNCH_SITE, [Landing_Outcome] \FROM SPACEXTBL \where
[Landing_Outcome] = 'Failure (drone ship)' and
substr(Date,7,4)='2015';
```

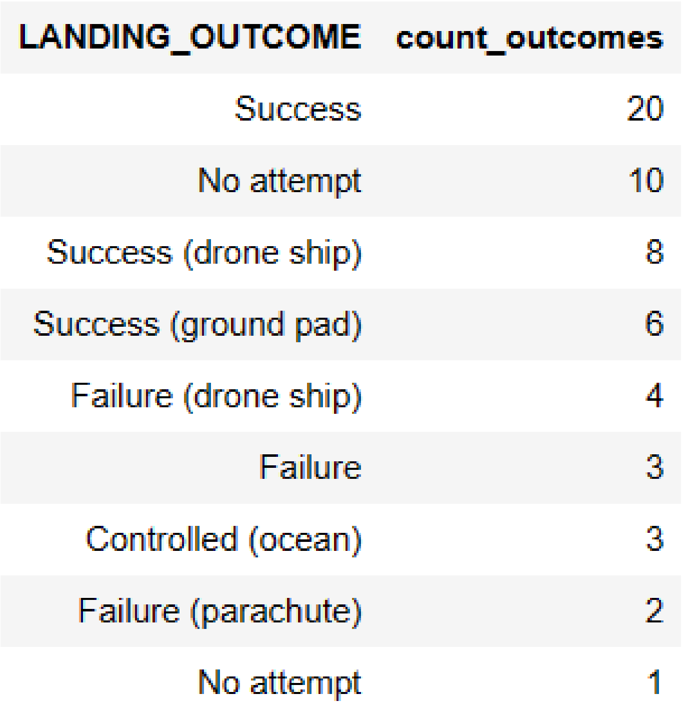
We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015



month	Date	Booster_Version	Launch_Site	LANDING_OUTCOME
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Success Count Between 2010-06-04 And 2017-03-20

```
SELECT [Landing_Outcome], count(*) as count_outcomes \

FROM SPACEXTBL \

WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing_Outcome] order by count_outcomes DESC;
```

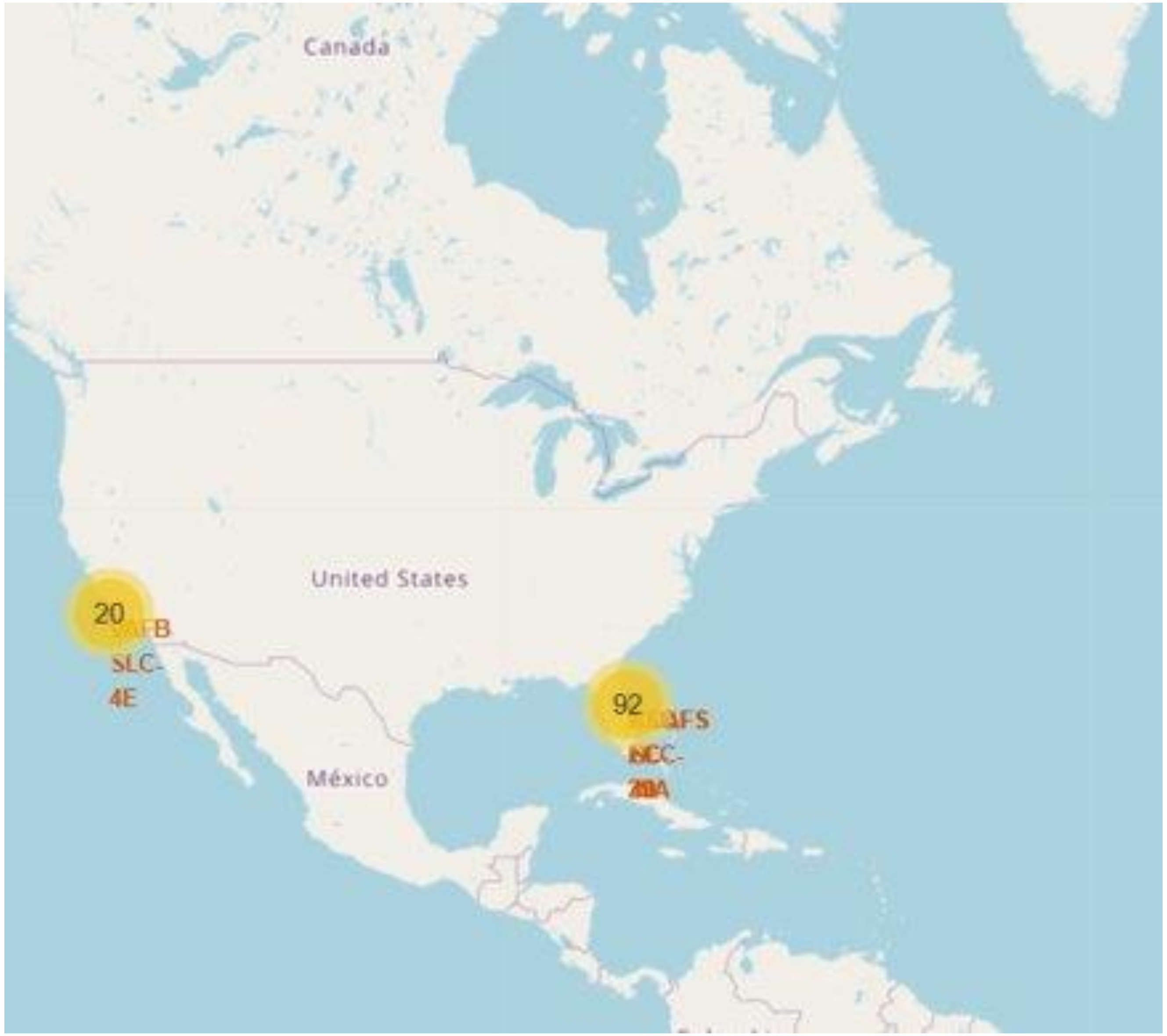


LANDING_OUTCOME	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.

- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

Interactive Analytics Results



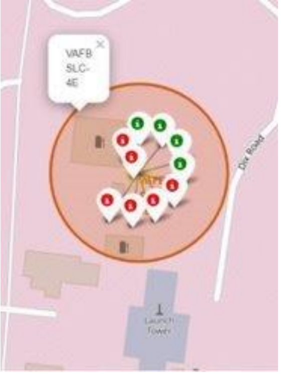
Interactive Map With Folium

the SpaceX launch sites are in the United States of America coasts. Florida and California

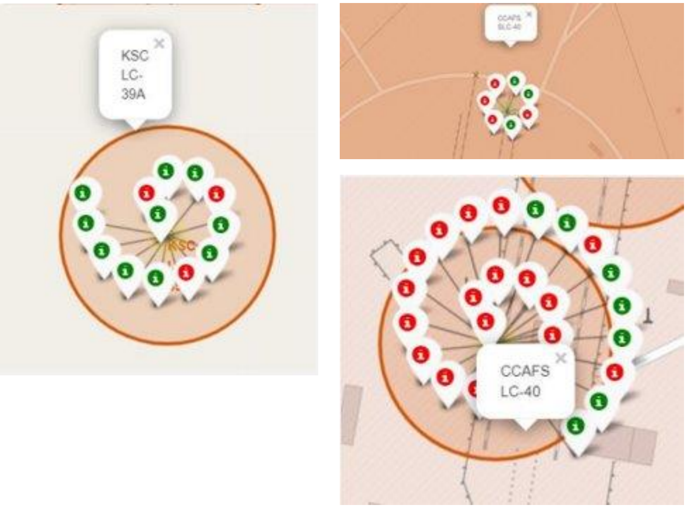
Color Labelled Markers

Green Marker shows successful Launches and Red Marker shows Failures

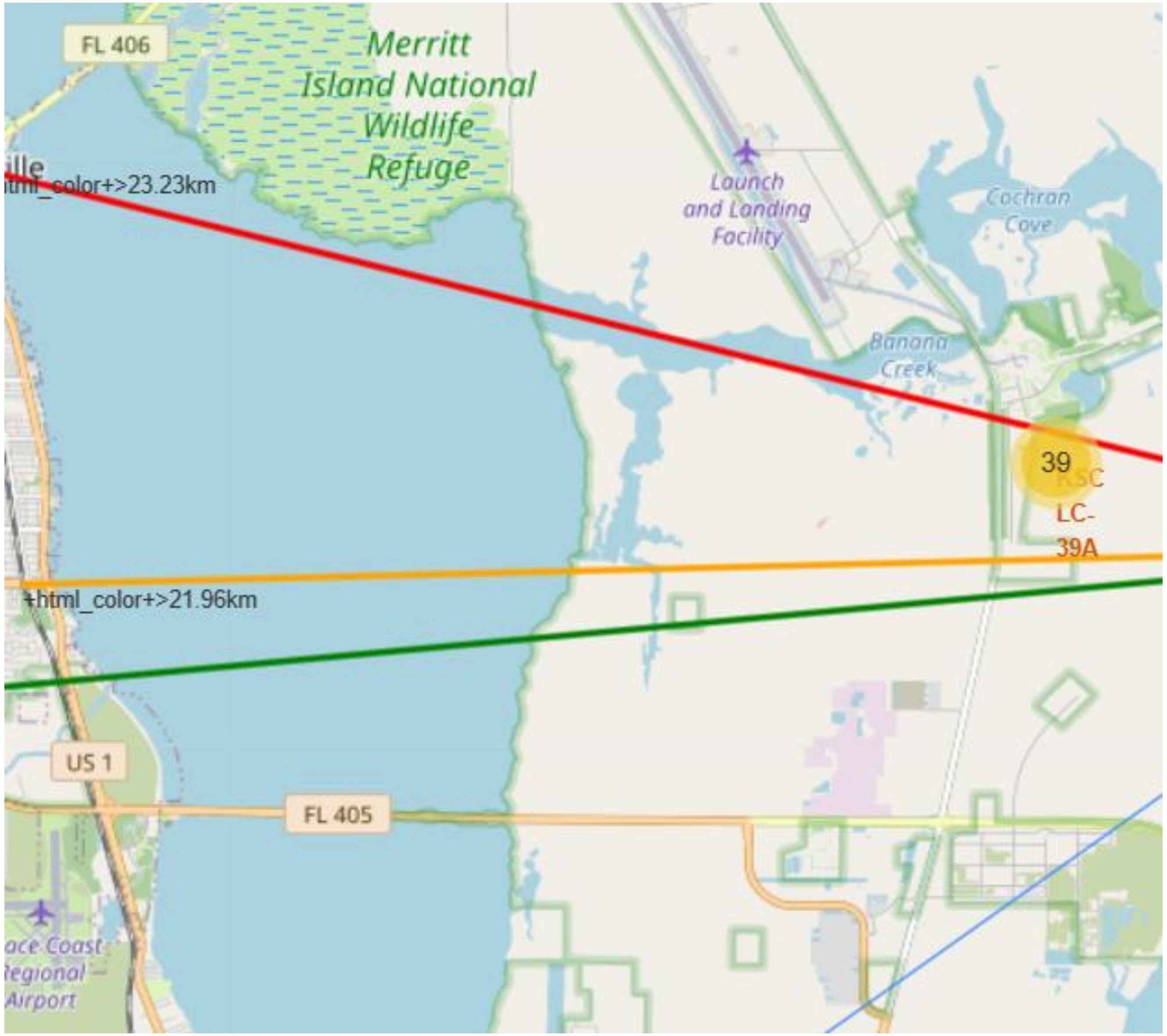
CCAFS SLC-40 has 42.9 % success



California launch sites



Florida launch sites



Launch Site Distance To Landmarks

CCAFS SLC-40

0.86 km from nearest coastline

21.96 km from nearest railway

23.23 km from nearest city

26.88 km from nearest highway

Are launch sites in close proximity to railways? No

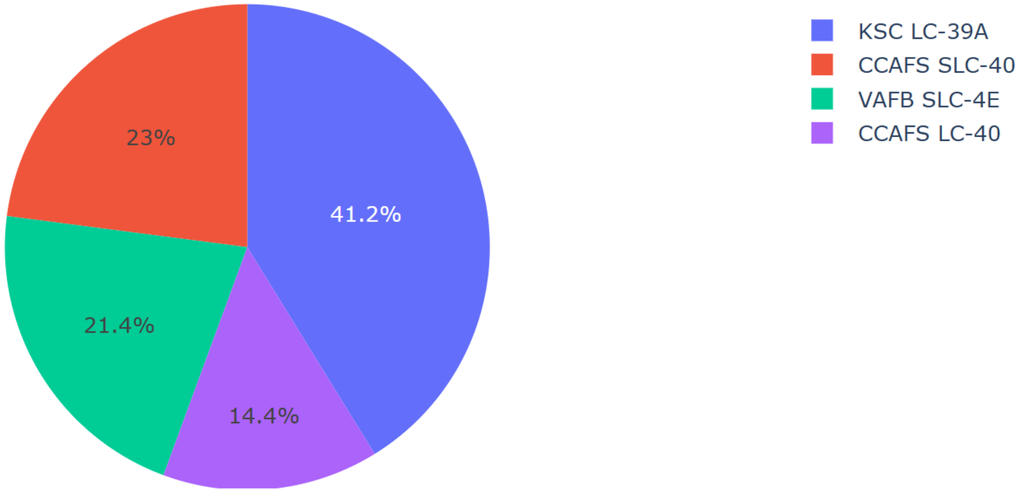
Are launch sites in close proximity to highways? No

Are launch sites in close proximity to coastline? Yes

Do launch sites keep certain distance away from cities? Yes

Dashboard with Plotly

KSC LC-39A has the most successful launches from all the sites



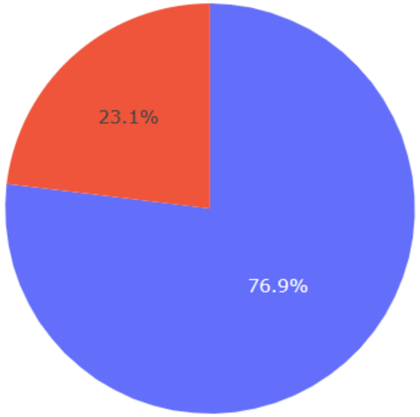
Pie chart showing the success percentage achieved by each launch site

Pie chart showing the Launch site with the highest launch success ratio

Pie chart

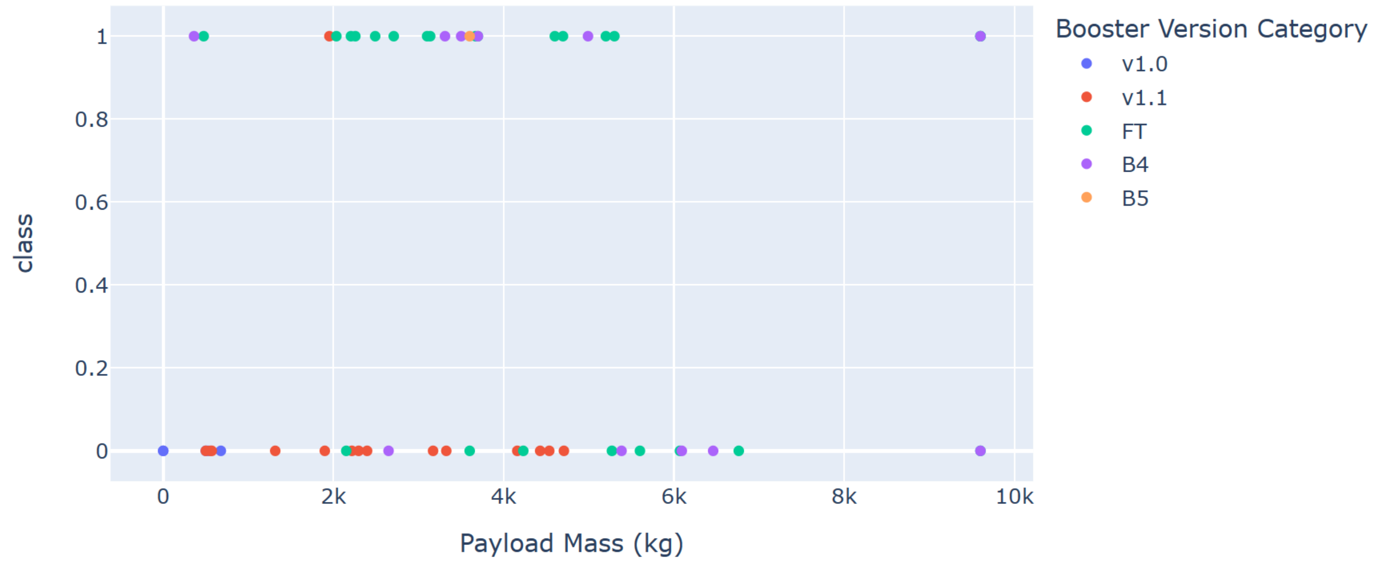
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Total Success Launches for Site KSC LC-39A



Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

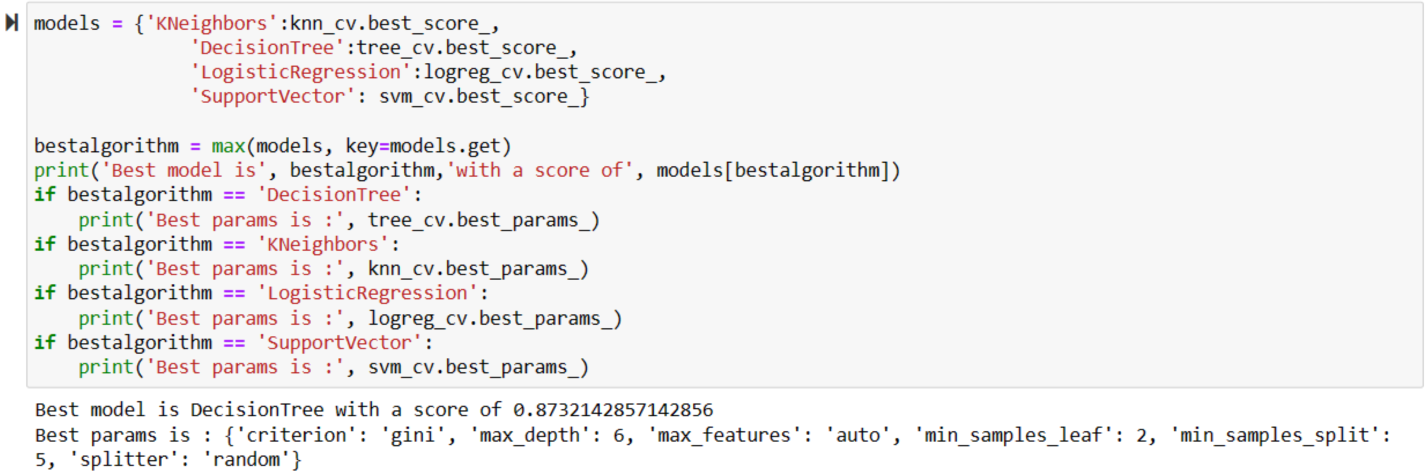
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads



Predictive Analysis Results

Classification Accuracy

The models had similar performance with equivalent scores and accuracy, possibly due to the limited dataset. However, upon examining the ".best_score_" metric, the Decision Tree model exhibited slightly superior performance compared to the others.

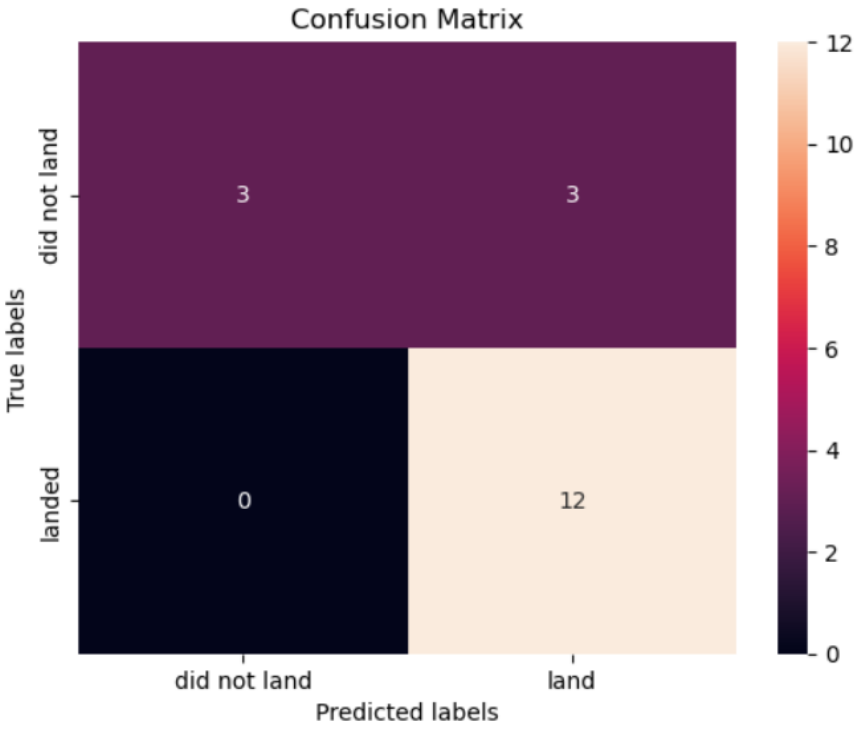


```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split':
5, 'splitter': 'random'}
```

Confusion Matrix

Upon inspecting the confusion matrix of the decision tree classifier, it is evident that the classifier is capable of distinguishing among the distinct classes. Nevertheless, a notable issue is the occurrence of false positives, which indicates instances where the classifier incorrectly identifies unsuccessful landings as successful landings.



Conclusion

The Machine Learning dataset suggests that the Tree Classifier Algorithm is the optimal choice.

- It appears that lower weighted payloads exhibit superior performance compared to heavier payloads.
- The success rates of SpaceX launches appear to improve proportionally with the passage of time, implying that they will eventually perfect their launches.
- Upon analyzing the launch sites, KSC LC-39A had the highest success rate.
- Success rates are highest for orbits GEO, HEO, SSO, and ES-L1.