Spacex Presentation on Capstone IBM Data Science

Samuel Breen

Presentation Contents

Executive summary and Data Methodology

- The methods used in this project were
- 1. API Collection
- 2. Data Wrangling
- 3 Web Scraping
- 4.Explortory Data Analysis
- 5. Exploratory Data Analysis with visualisation
- 7 Ploty Dash
- 6. Predictive Analysis with Machine Learning
- In this project on Space x , it was studies from various aspects , such as location of site, failure and success rates of landing , map visualisation relating to said site maps.
- The Data was analysed through various methods.
- These will be discussed in each section and given detail on each aspect of data analysis from each method.
- I will explain each method type and then go into detail how each affected SpaceX.

Introduction

- In this project on Space x , it was studies from various aspects , such as location of site, failure and success rates of landing , map visualisation relating to said site maps.
- The Data was analysed through various methods.
- These will be discussed in each section and given detail on each aspect of data analysis from each method.
- I will explain each method type and then go into detail how each affected SpaceX.
- SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX –or a competing company –can reuse the first stage.

API Data Collection

Steps

Requestdata from SpaceX API (rocket launch data)

Decoderesponse using .json() and convert to a dataframeusing .json_normalize()

Request informationabout the launches from SpaceX API using custom functions

Create dictionaryfrom the data

Create dataframefrom the dictionary

Filter dataframe to contain only Falcon 9 launches

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

Export data to csv file

Web Scraping

- Steps
- Request data(Falcon 9 launch data) from Wikipedia
- Create BeautifulSoupobject 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

Data Wrangling

Steps

Perform EDA and determine data labels

Calculate:# of launches for each site

and occurrence of orbit

and occurrence of mission outcome per orbit type]

Create binary landing outcome column (dependent variable)

Export data to csv file

False Ocean:represented an unsuccessful landing to a specific region of ocean

True RTLS:meant the mission had a successful landing on a ground pad

False RTLS: represented an unsuccessful landing on a ground pad

True ASDS: meant the mission outcome had a successful landing on a drone ship

False ASDS: represented an unsuccessful landing on drone ship

Outcomes converted into 1 for a successful landing and 0 for an unsuccessful landing

Landing Outcome Cont.

Landing Outcome

Landing was not always successful

True Ocean :mission outcome had a successful landing to a specific region of the ocean

Exploratory Data Analysis

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

View relationship by using **scatter plots**. The variables could be useful for machine learning if a relationship exists

Show comparisons among discrete categories with **bar charts**. Bar charts show the relationships among the categories and a measured value.

EDA with SQL

Queries → Display: → Names of unique launch sites → 5 records where launch site begins with 'CCA'

→ Total payload mass carried by boosters launched by NASA (CRS) → Average payload mass carried by booster version F9 v1.1. → List: → Date of first successful landing on ground pad

→ Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000 → Total number of successful and failed missions → Names of booster versions which have carried the max payload → Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015

Interactive Maps with Folium

Markers Indicating Launch Sites

Added **blue circle** at **NASA Johnson Space Centre's coordinate** with a **popup label** showing its name using its latitude and longitude coordinates

Added **red circles** at **all launch sites coordinates** with a **popup label** showing its name using its name using its latitude and longitude coordinates

Map with Folium

Colored Markers of Launch Outcomes

Added **coloured markers** of **successful**(**green**) and **unsuccessful**(**red**) **launches** at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

Added **coloured lines** to **show distance between** launch site **CCAFS SLC-40 and** its proximity to the **nearest coastline, railway, highway, and city**

Plotly Dash Dashboard

Dropdown List with Launch Sites

Allow user to select all launch sites or a certain launch site

Dashboard with Plotly Dash

Slider of Payload Mass Range

Allow user to select payload mass range

Pie Chart Showing Successful Launches

Allow user to see successful and unsuccessful launches as a percent of the total

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

Allow user to see the correlation between Payload and Launch Success

Predictive Analytics with Machine Learning

Charts

Create NumPy array from the Class column

Standardize the data with StandardScaler. Fit and transform the data.

Split the data using train_test_split

Create a GridSearchCV object with cv=10 for parameter optimization

different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor

Calculate accuracy on the test data using .score() for all models

Assess the confusion matrix for all models

Identify the best model using Jaccard_Score, F1_Score and Accuracy

Results Exploratory Data Analysis

Launch success has improved over time

KSC LC-39A has the highest success rate among landing sites

Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Results Summary

Visual Analytics

Most launch sites are near the equator, and all are close to the coast

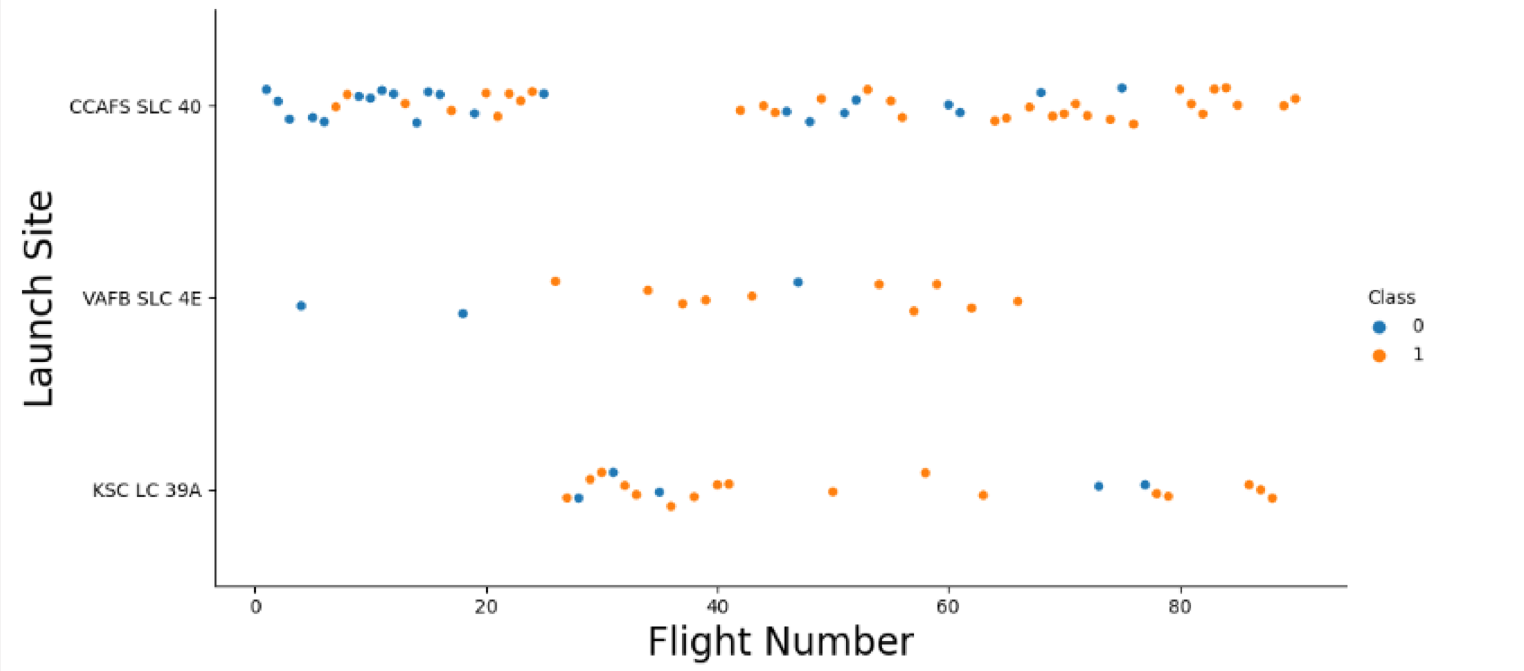
Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

Decision Tree model is the best predictive model for the dataset

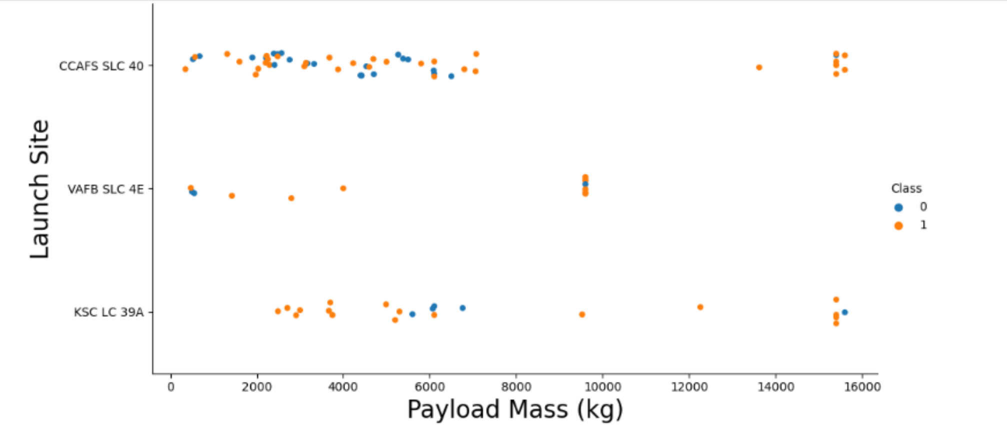
Flight Number vs. Launch Site

- **Exploratory Data Analysis**
- **Earlier flights** had a **lower success rate (blue = fail)**
- **Later flights** had a **higher success rate (orange = success)**
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



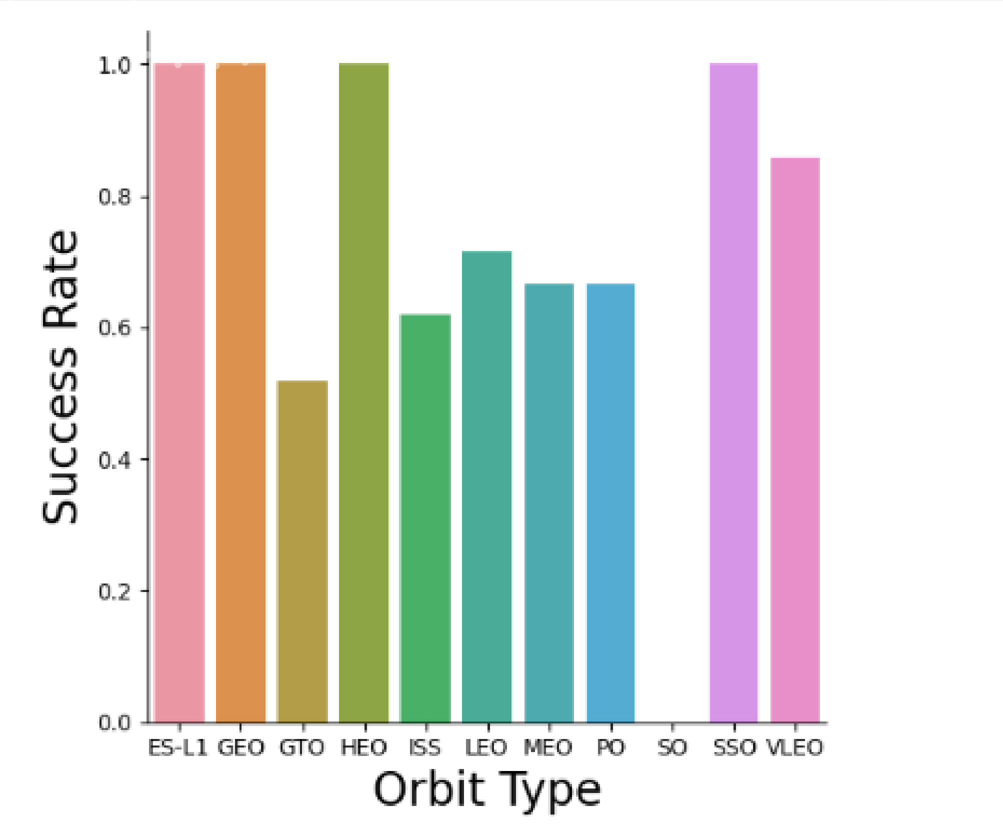
Payload vs. Launch Site

- **Exploratory Data Analysis**
- Typically, the **higher**the **payload mass** (kg), the **higher** the **success rate**
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg

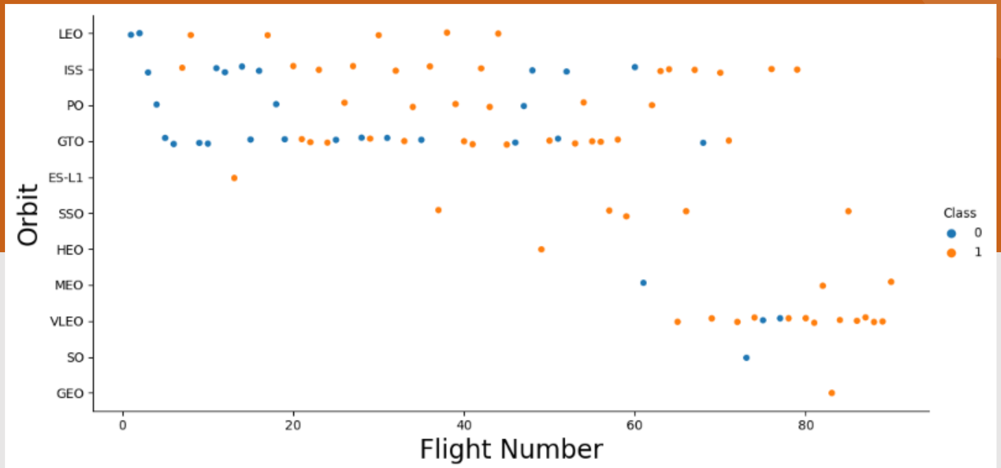


Success Rate by Orbit

- **Exploratory Data Analysis**
- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



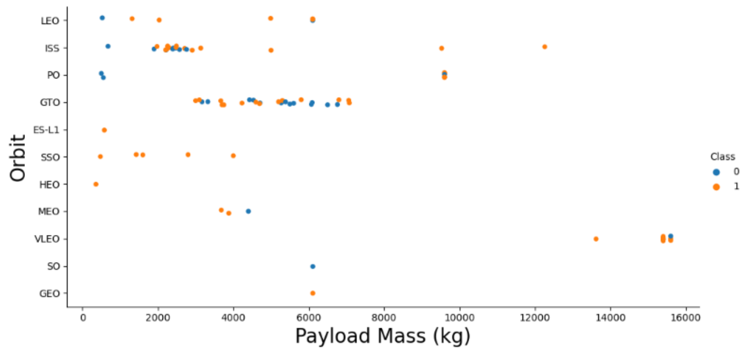
Flight Number vs. Orbit



- **Exploratory Data Analysis**
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend

Payload vs. Orbit

- **Exploratory Data Analysis**
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success over Time

- **Exploratory Data Analysis**
- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013

```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   sqlite:///my_data1.db
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Information

- **Launch Site Names**
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

- **Landing Outcome Cont.**
- **Records with Launch Site Starting with CCA**
- Displaying 5 records below

Payload Mass

- **Total Payload Mass**
- **45,596 kg** (total) carried by boosters launched by NASA (CRS)
- **2,928 kg** (average) carried by booster version F9 v1.1
- **Average Payload Mass**
- **2,928 kg** (average) carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
2928

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4l
   sqlite:///my_data1.db
Done.
```

1
45596

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

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

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Landing & Mission Info

- **1stSuccessful Landing in Ground Pad**
- 12/22/2015
- **Booster Drone Ship Landing**
- Booster mass greater than 4,000 but less than 6,000
- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105
- **Total Number of Successful and Failed Mission Outcomes**

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

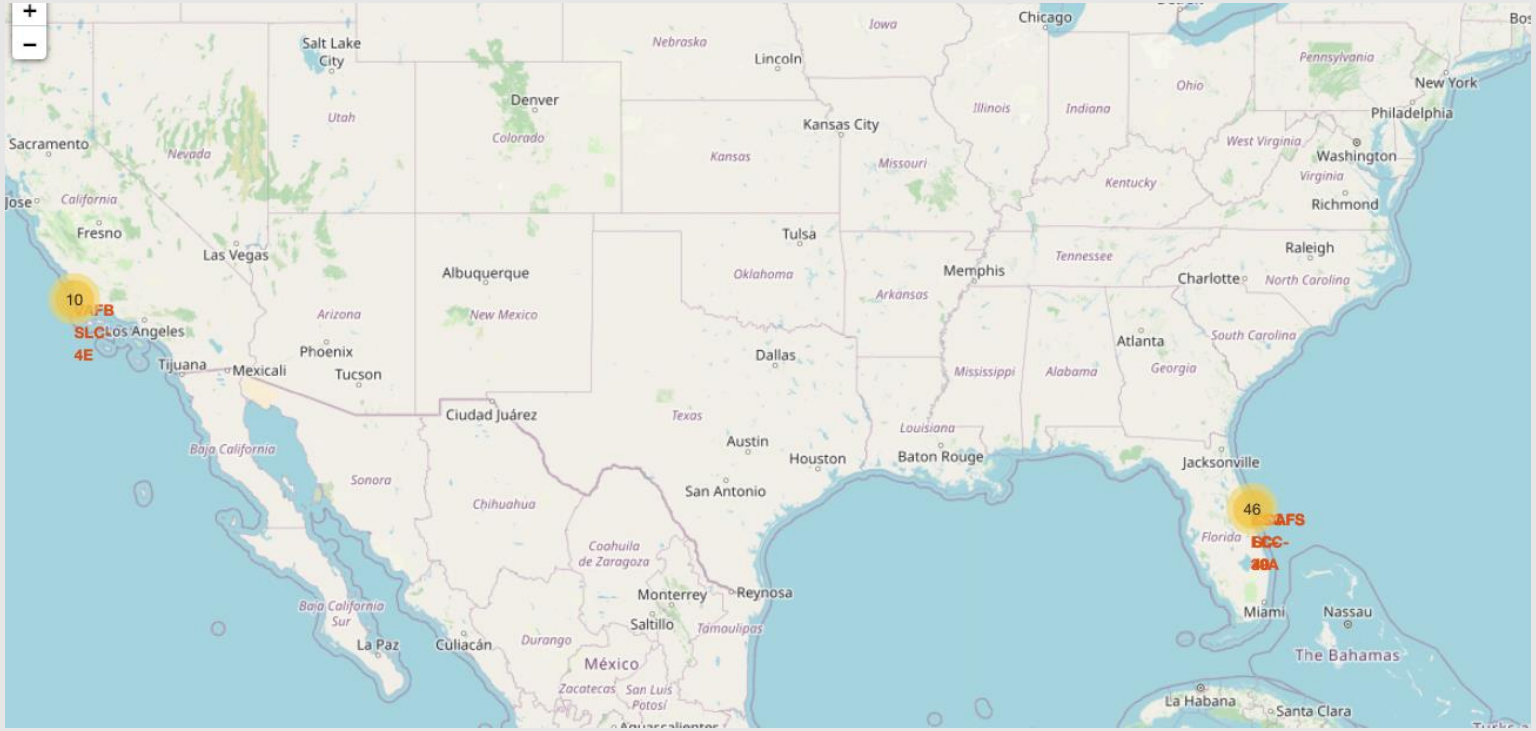
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Boosters

- **Carrying Max Payload**
- 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

Launch sites analysis

- We can see that all the SpaceX launch sites are located inside the United States
- After you plot distance lines to the proximities, you can answer the following questions easily:
- Are launch sites in close proximity to railways?
- Are launch sites in close proximity to highways?
- Are launch sites in close proximity to coastline?
- Do launch sites keep certain distance away from cities?
- Launch sites are in close proximity to highways, which allows for easily transport required people and property. Launch sites are in close proximity to railways, which allows transport for heavy cargo. Launch sites are not in close proximity to cities, which minimizes danger to population dense areas



Predictive analysis

- Classification Accuracy As we can see, by using the code as below: we could identify that the best algorithm to be the Tree Algorithm which have the highest classification accuracy
-

```
parameters_lr ={"C":[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}# l1 lasso l2 ridge

# define the model
lr = LogisticRegression(random_state = 12345)

# define the grid search object
grid_search_lr = GridSearchCV(
    estimator = lr,
    param_grid = parameters_lr,
    scoring = 'accuracy',
    cv = 10
)
# execute search
logreg_cv = grid_search_lr.fit(X_train,Y_train)
```

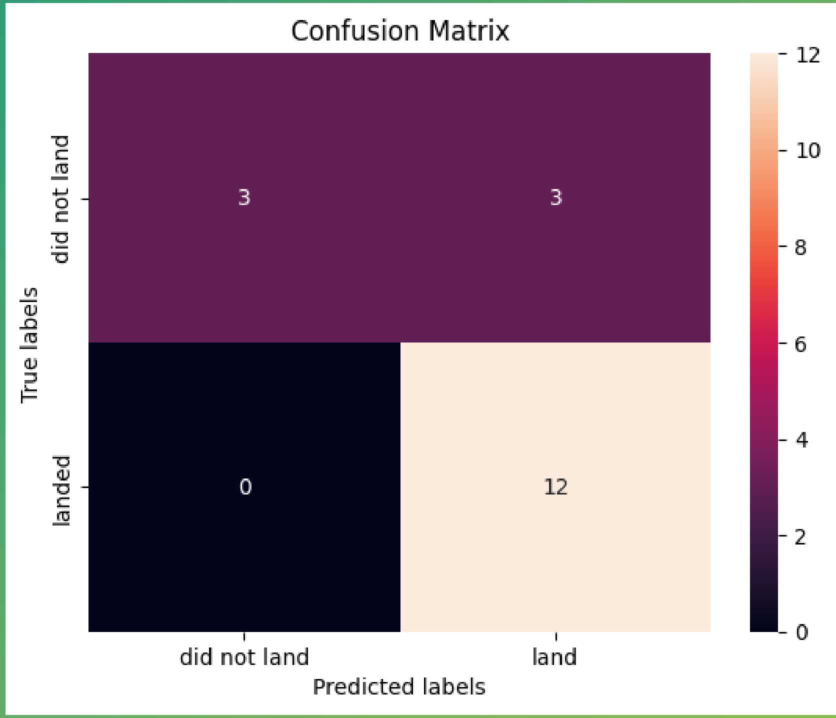
We output the `GridSearchCV` object for logistic regression. We display the best parameters using the data attribute `best_params_` and the accuracy on the validation data using the data attribute `best_score_`.

Ploty Dash

Built a ploty dashboard to play around with the data

Made plots in order to visualise it on a dashboard

Scatter plots and pie charts were made.



Confusion matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

support vector machine, decision tree and k nearest neighbour

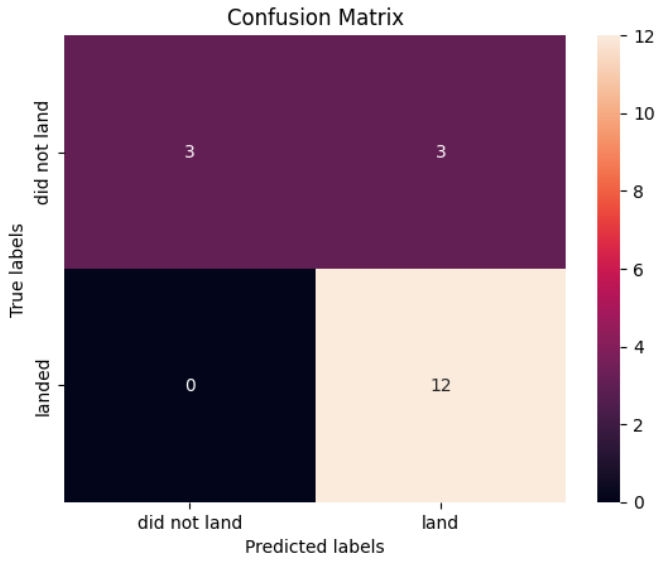
TASK 12

Find the method performs best:

In [32]:
```
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': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```



- **Performance Summary**
- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:12 True positive
- 3 True negative
- **3 False positive**
- 0 False Negative
- **Precision**= TP / (TP + FP)12 / 15 = .80
- **Recall**= TP / (TP + FN)12 / 12 = 1
- **F1 Score**= 2 * (Precision * Recall) / (Precision + Recall)2 * (.8 * 1) / (.8 + 1) = .89
- **Accuracy**= (TP + TN) / (TP + TN + FP + FN) = .833
- Best method: decision tree

Conclusion

Research

Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming

Equator: Most of the launch sites are near the equator for an additional natural boost -due to the rotational speed of earth - whichhelps save the cost of putting in extra fuel and boosters

Coast: All the launch sites are close to the coast

Launch Success: Increases over time

KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg

Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate

Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Things to Consider

Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set

Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy