Improving Space Race Outcome using Data Science

REBECA LOZANO

FEBRUARY 18, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of methodologies

- Data Collection using 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 results

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

Introduction

- Space X advertises Falcon 9 rocket launches on its website at 62 million dollars, while other providers cost over 165 million dollars each. These savings are mainly accredited to the fact that Space X can reuse the first stage, and therefore, if the outcome of the first stage can be determined (successful or unsuccessful landing), we can determine the cost of a launch. The goal here is to use a machine learning model using public information to predict if Space X will reuse the first stage.

- We would like to find:

1. What factors determine if the rocket will land successfully?
2. The interaction amongst various features that determine the success rate of a successful landing.
3. What operating conditions need to be in place to ensure a successful landing program?

Methodology

Data Collection

Our data was collected using a varied method.

1. We first collected out data using a get request to the Space X API.
2. We then decoded the response content as a Json file using .json() function and turned it into a pandas dataframe using .json_normalize()
3. We performed data cleaning, checked for missing values and filled missing values as proved essential.
4. We also gathered Falcon 9 launch records from Wikipedia using BeautifulSoup, where we extracted the launch records as an HTML table, parsed the table and converted it to a pandas dataframe for facilitated usage.

The following slides show notebook code performing the tasks above.

Data Collection – Space X API

Accessing the Space X API using link:

```
In [59]: spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
In [60]: response = requests.get(spacex_url)
```

Check the content of the response

```
In [61]: print(response.content)
```


```
b'[{"fairings":{"reused":false,"recovery_attempt":false,
{"campaign":null,"launch":null,"media":null,"recovery":r
96-spacex-inaugural-falcon-1-rocket-lost-launch.html","v
95eda69955f709d1eb","success":false,"failures":[{"time":
bb0006eeb1e1"],"launchpad":"5e9e4502f5090995de566f86","
se,"cores":[{"core":"5e9e289df35918033d3b2623","flight":
```


Here, we used the get request to access the Space X API in order to follow through with data collection, wrangling and formatting.

[Notebook: Data Collection using Space X API]()

Data Collection - Web Scraping

To keep the lab tasks consistent, you will be asked to scrape the data from a snapshot of the `Li`
updated on `9th June 2021`

In [5]:
```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and
```

Next, request the HTML page from the above URL and get a `response` object

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HT

In [6]:
```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

Create a `BeautifulSoup` object from the HTML `response`

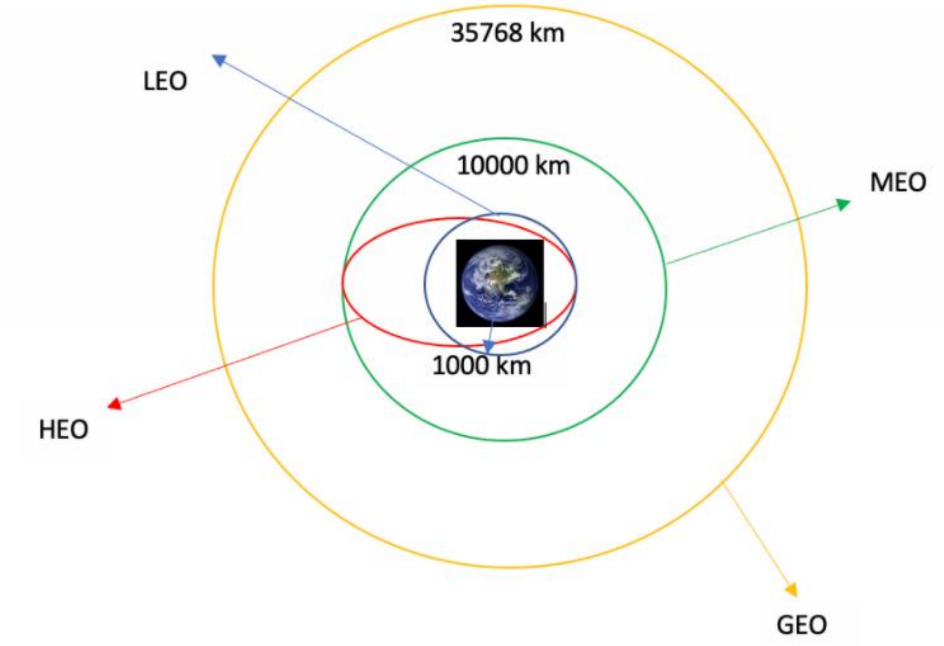
In [15]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text c
soup = BeautifulSoup(response.text, "html.parser")
```

Print the page title to verify if the `BeautifulSoup` object was created properly

Here, we used the get request and BeautifulSoup to access Falcon 9 launch records.

[Notebook: Data Collection using Web Scraping]()

Data Wrangling



- Using data wrangling processes, we determined the training tables. Here, we calculated the number of launches at each site and occurrence of each orbits.
- We also created a landing outcome label from outcome column and exported these results onto a csv file.

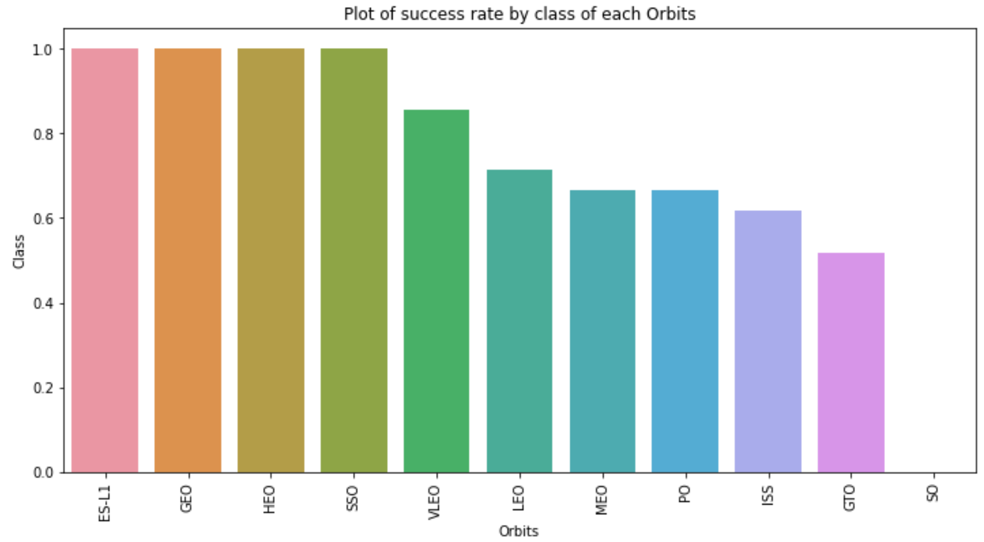
Notebook: Data Wrangling

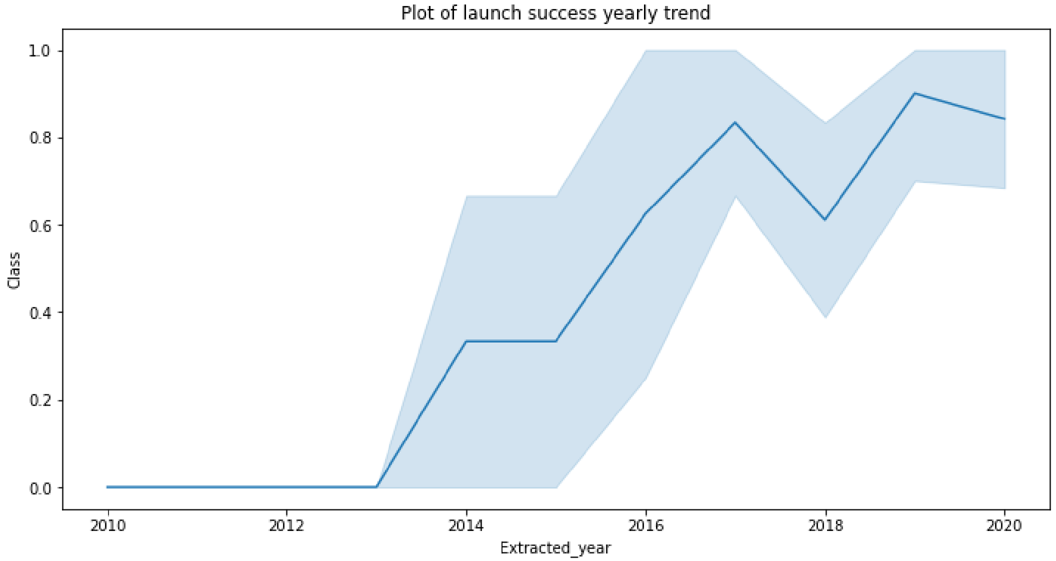
Exploratory Data Analysis with Visualization

We used our data to explore different visualizations that pinpoint crucial information of our dataset. We visualized the relationship between flight number and launch site, payload and launch site, success rate and orbit type, flight number and orbit type, payload and orbit type, and the launch success yearly trend.

To the right you can find the visualizations showing the relationship between success rate and orbit type, and the launch success yearly trend.

[Notebook: EDA with Visualization]





Exploratory Data Analysis with SQL

We loaded the Space X dataset into a PostgreSQL database using Jupyter Notebooks.

We conducted exploratory data analysis with SQL to gain further insight and gathered information such as:

- The names of the unique launch sites in the space mission
- The total payload mass carried by boosters launches by NASA
- The average payload mass carried by booster version F9 v1.1
- The total number of successful and failure mission outcomes
- The ranks of landing outcomes between specific dates

Notebook: EDA with SQL

Interactive Visual Analytics – Building an Interactive Map with Folium

- We built an interactive visual analytics map using Folium. We marked launch sites and added map objects to highlight the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to classes 0 and 1, where 0 signifies a failure and a 1 a success.
- We calculated the distances between a launch site to its proximities, that allowed us to answer questions such as:
  - Are launch sites near railways, highways, and coastlines?
  - Do launch sites keep certain distance away from cities?

[Notebook: Interactive Map with Folium]

Building a Dashboard with Plotly Dash

- We constructed a dashboard using Plotly Dash.
- Here, we created pie charts that show the total launches by sites
- We plotted the scatter graph showing the relationship between outcome and payload mass for the different booster version.

Predictive Analysis

- Lastly, we built different machine learning models and applied different parameters using GridSearchCV.
- Using accuracy as the metric for our model, we improved the model with feature engineering and algorithm tuning.
- We were able to find the best classification model.

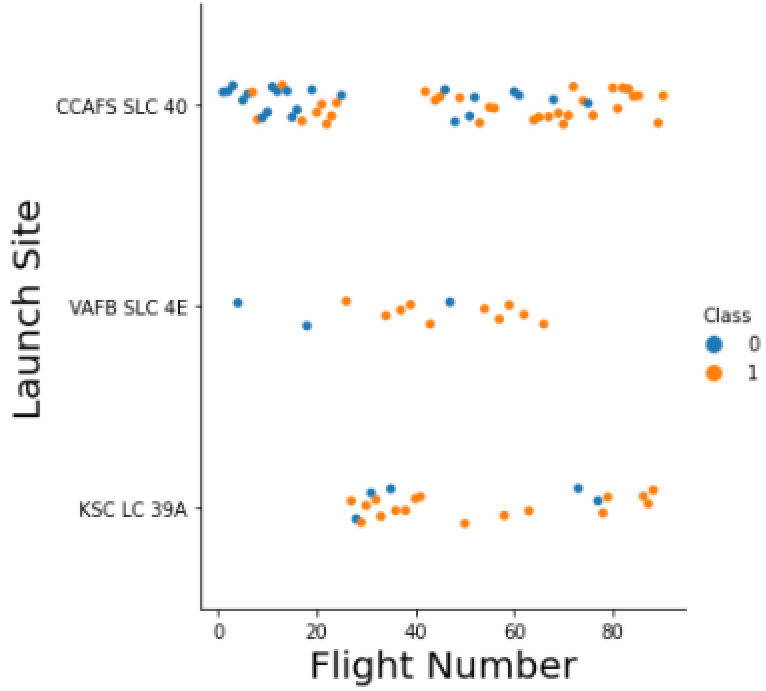
Notebook: Machine Learning Predictions

Results

EDA Results

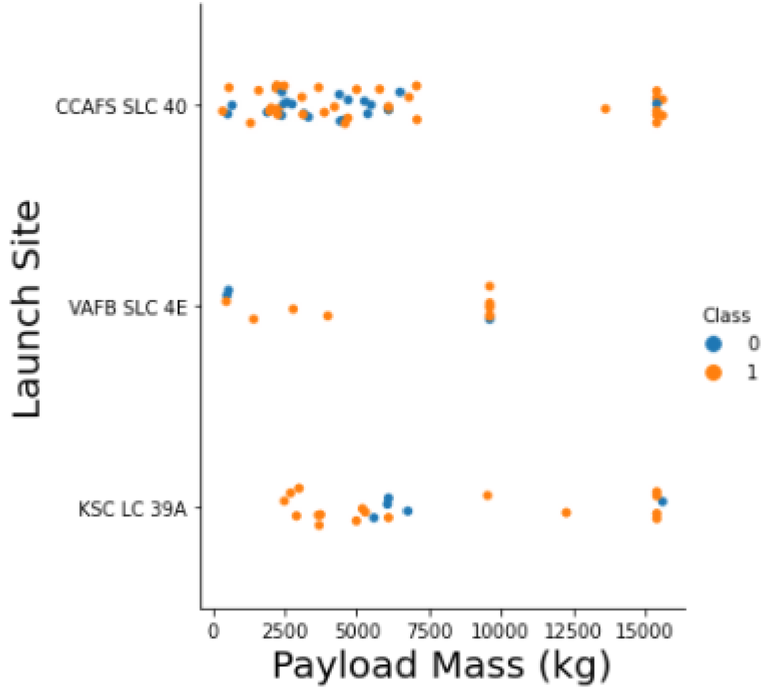
Flight Number and Launch Site

Our plot shows that the greater the flight amount at a launch site, the greater the success rate at a launch site.



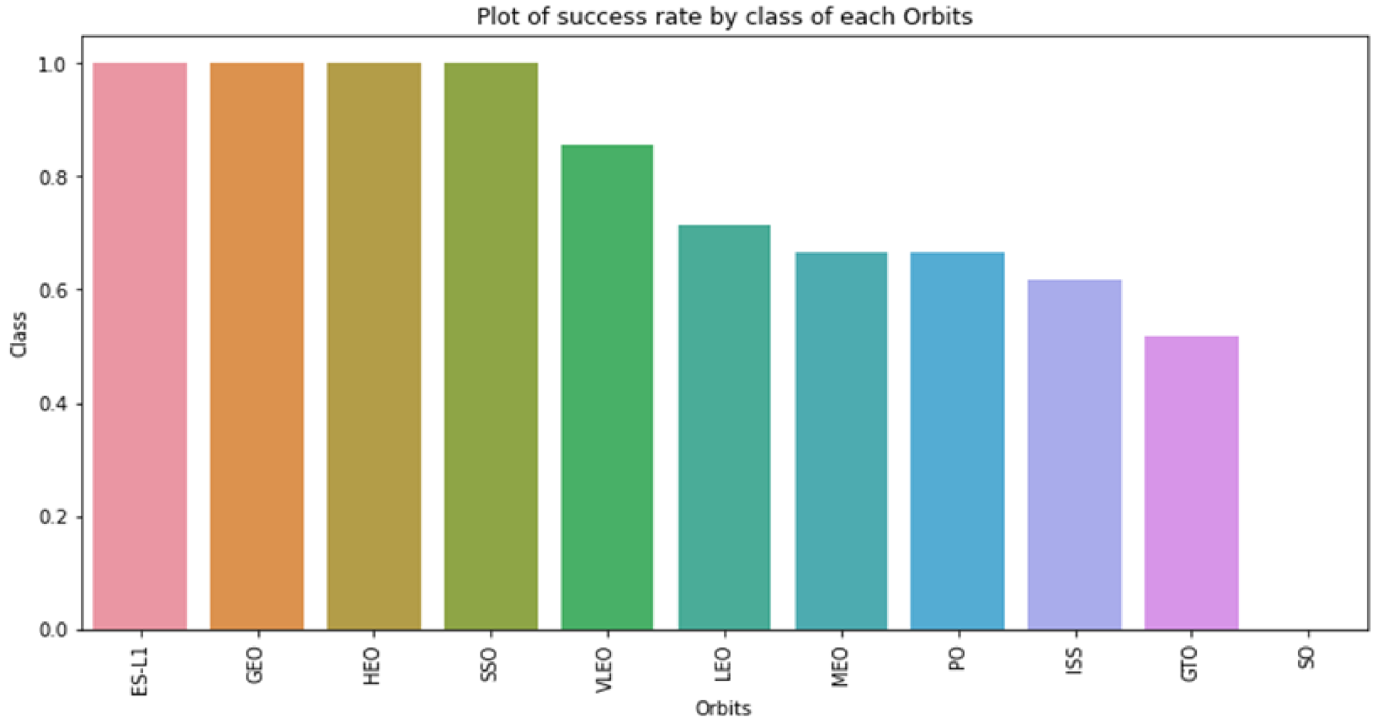
Payload and Launch Site

Our plot shows that the greater the payload mass for the launch site CCAFS SLC 40, the higher the success rate for the rocket.



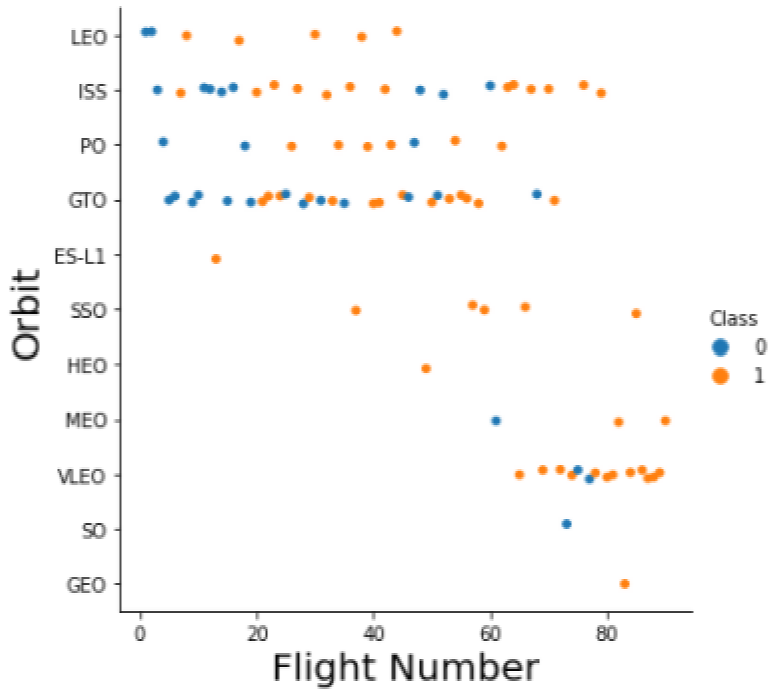
Success Rate and Orbit Type

Our plot shows that ES-L1, GEO, HEO, SSO, and VLEO had the most success rate.



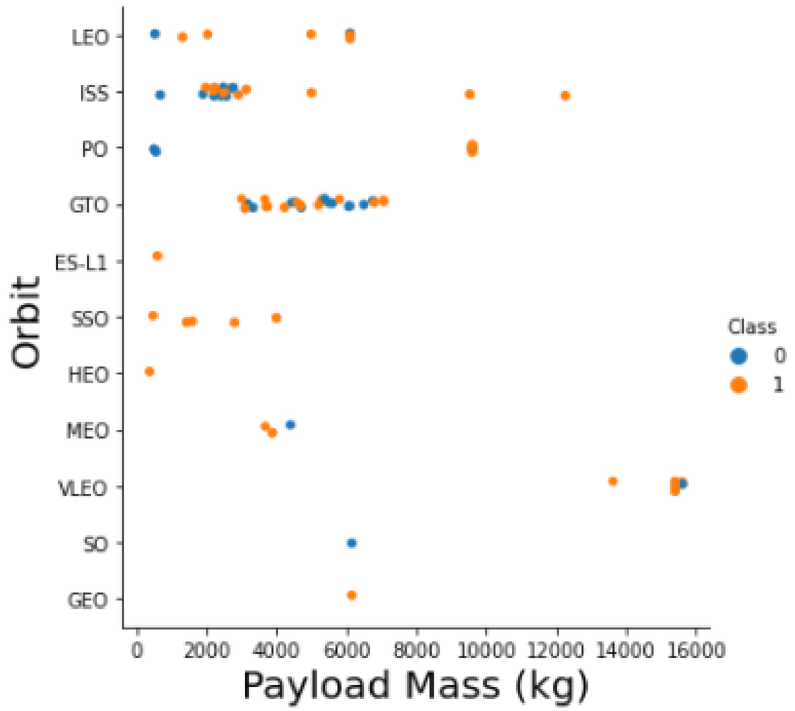
Flight Number and Orbit Type

Our plot shows Flight Number against Orbit Type. We can see that with the LEO orbit, success is related to the flight number while with the GTO orbit, there is no relationship between flight number and the orbit.



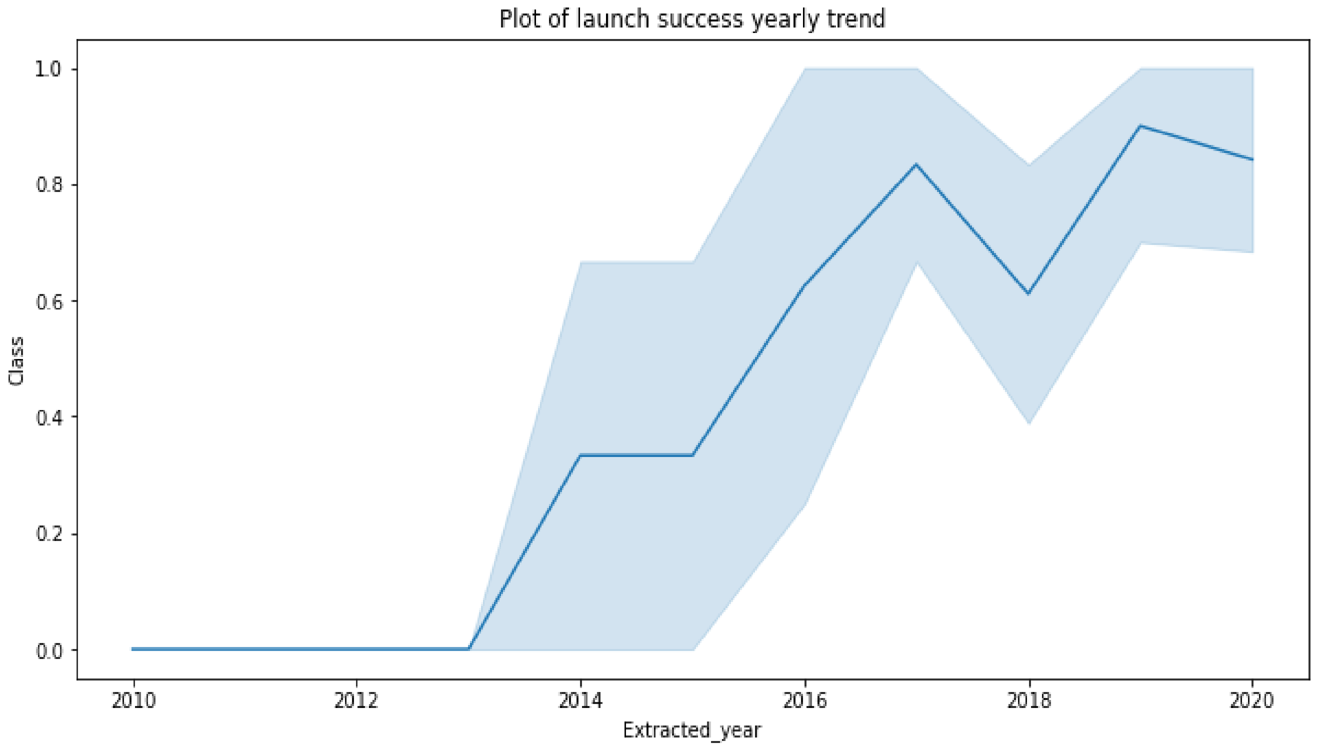
Payload and Orbit Type

Our plot to the right shows that heavy payloads are associated with more successful landings for PO, LEO, and ISS.



Launch Success Yearly Trend

Our plot shows that there has been a substantial overall increase in success rate from 2013 to 2020.



Launch Site Names

We used **DISTINCT** to show the unique launch sites from the Space X data.

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

In [46]:
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

Out[46]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

We used the query below to display 5 records where launch sites begin with `CCA`.

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

Rebeca Lozano (rebecalozano48@gma

In [47]:

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

Out[47]:

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

Total Payload Mass

We calculated the total payload carried by boosters from NASA as 45596 using the query below.

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

In [50]:
```
%sql SELECT SUM(PAYLOAD_MASS_KG) AS Total_PayloadMass FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA (CRS)';
```

Out[50]:

total_payloadmass
45596

Average Payload Mass by F9 v1.1

We calculated the average payload mass carried by booster version F9 v1.1 to be 2928.

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

In [51]:
```
%sql SELECT AVG(PAYLOAD_MASS_KG) AS Avg_PayloadMass FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Out[51]:

avg_payloadmass
2928

First Successful Ground Landing Date

We gathered that the dates of the first successful landing outcome on ground pad was December 22, 2015.

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

Hint:Use min function

In [52]:

```
%sql SELECT MIN(DATE) AS FirstSuccessfull_landing_date FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'Success (ground pad)';
```

Out[52]:

firstsuccessfull_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

We gathered that the list of boosters which have success drone ship landing and payload mass between 4000 to 6000 consists of F9 FT B1022, F9 FT B1026, F9 FT B1021 2, and F9 FT B1031 2.

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

In [53]:
```
%sql SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS_KG > 4000  AND PAYLOAD_MASS_KG < 6000;
```

Out[53]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

We gathered that the total number of successful and failure mission outcomes were 100 and 1 respectively.

List the total number of successful and failure mission outcomes

In [54]:
```
%sql SELECT COUNT(MISSION_OUTCOME)
AS SuccessOutcome
FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'Success%';
```

Out[54]:

successoutcome
100

In [55]:
```
%sql SELECT COUNT(MISSION_OUTCOME)
AS FailureOutcome
FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Failure%';
```

Out[55]:

failureoutcome
1

Boosters Carried Maximum Payload

We gathered the list of booster versions that have carried the maximum payload mass using a subquery.

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

In [58]:
```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE PAYLOAD_MASS_KG = (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTBL)
ORDER BY BOOSTER_VERSION;
```

Out[58]:

booster_version	payload_mass_kg
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

We gathered the failed landing outcomes in drone ship, their booster versions, and launch sites for 2015 to be F9 v1.1 B1012 and F9 v1.1 B1015.

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [59]:

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME
FROM SPACEXTBL
WHERE LANDING_OUTCOME
LIKE 'Failure (drone ship)' AND DATE BETWEEN '2015-01-01' AND '2015-12-31';
```

Out[59]:

booster_version	launch_site	landing_outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

We gathered the count of landing outcomes between 2010-06-04 and 2017-03-20 by ranks as shown below.

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

In [60]:
```
%sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME)
FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME ORDER BY COUNT(LANDING_OUTCOME) DESC;
```

Out[60]:

landing_outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Launch Sites Proximities

Launch Sites with Global Map Markers

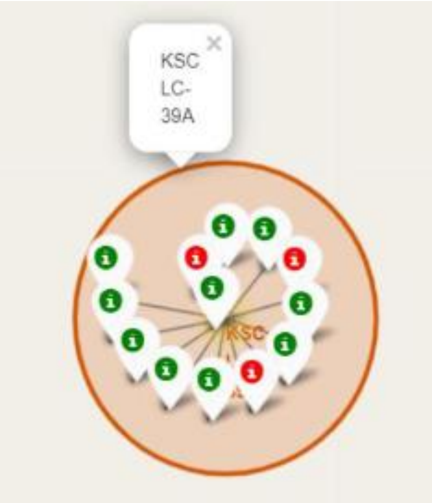
Markers showing launch sites with color labels

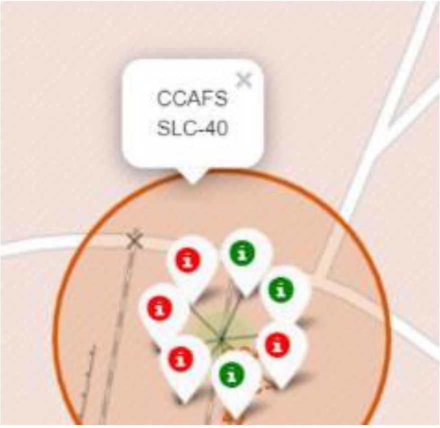
California Launch Site

Red Markers: Failure Launches
Green Markers: Successful Launches

Markers showing launch sites with color labels

Florida Launch Sites



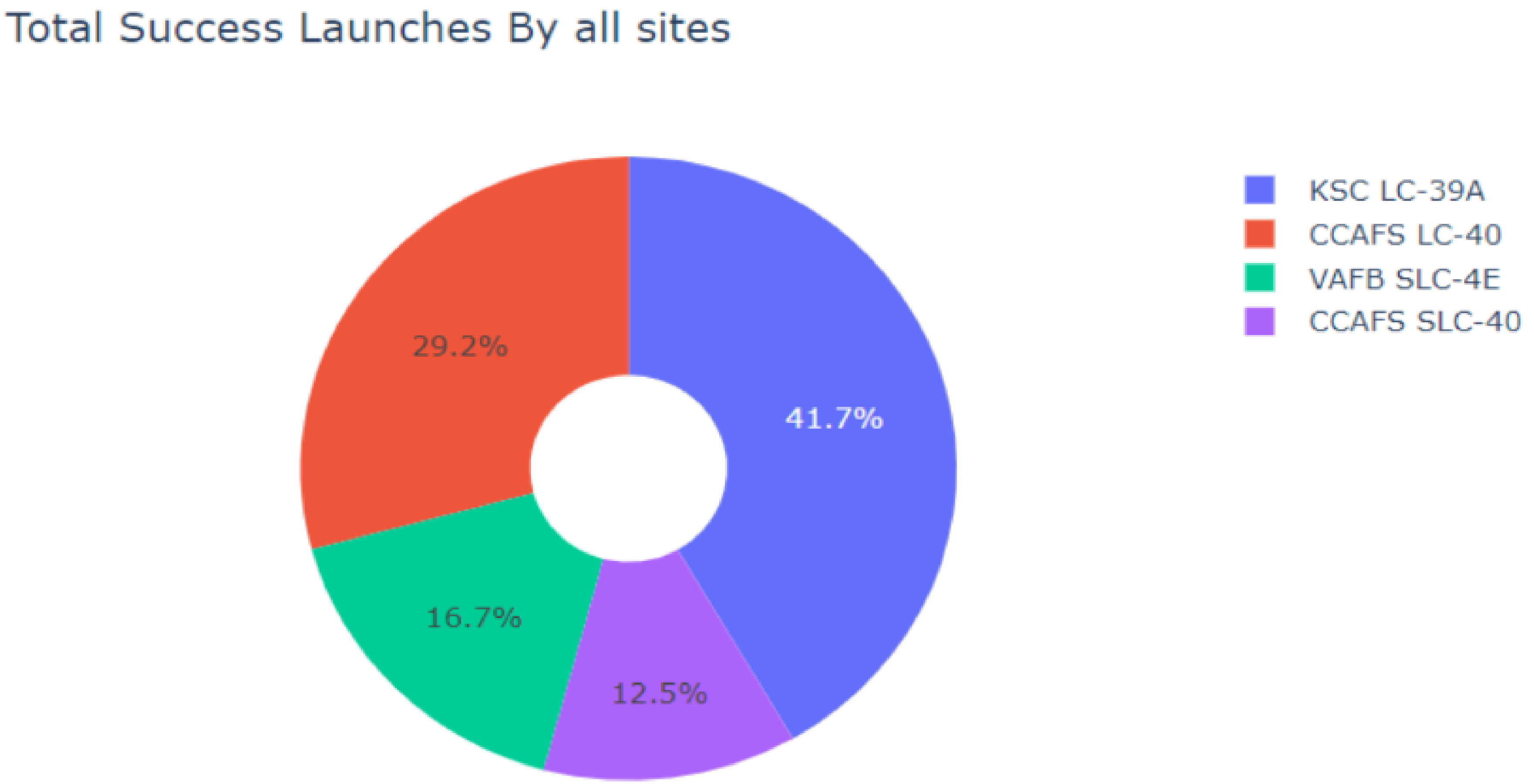


Red Markers: Failure Launches
Green Markers: Successful Launches

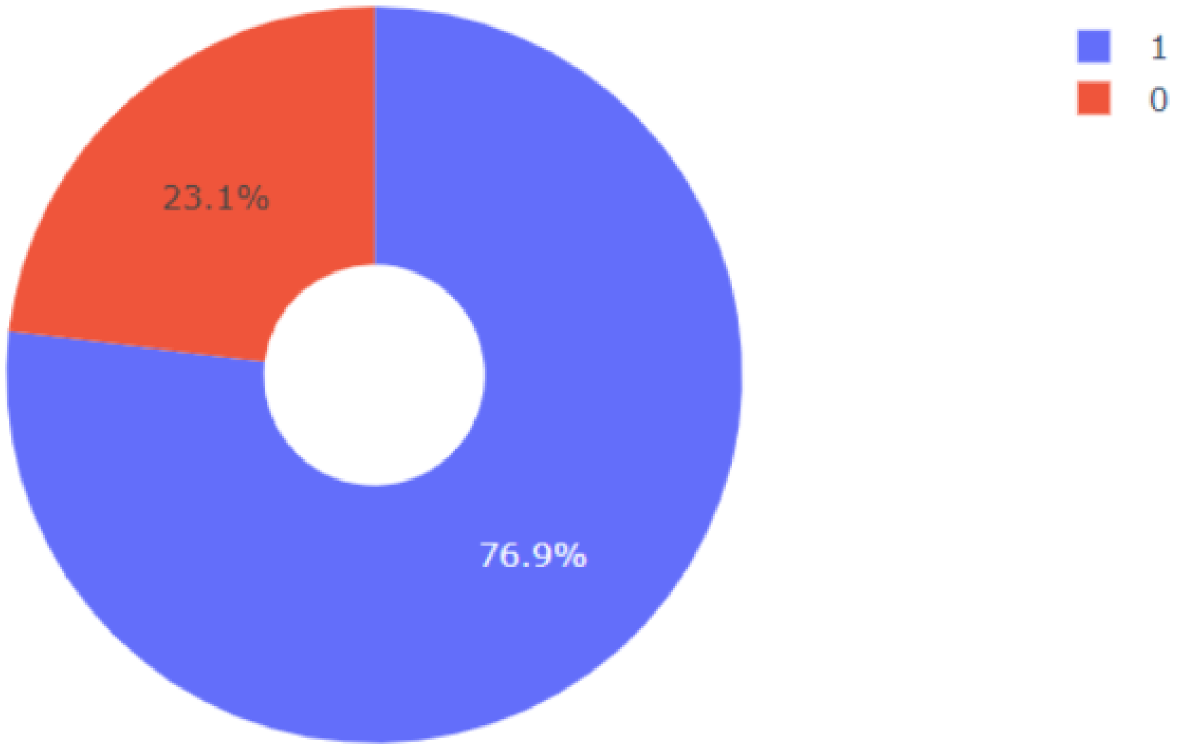
Questions Answered using Launch Site Markers

1. Are launch sites in close proximity to railways? No.
2. Are launch sites in close proximity to highways? No.
3. Are launch sites in close proximity to coastlines? Yes.
4. Do launch sites keep certain distance away from cities? Yes.

Success percentage achieved by each launch site

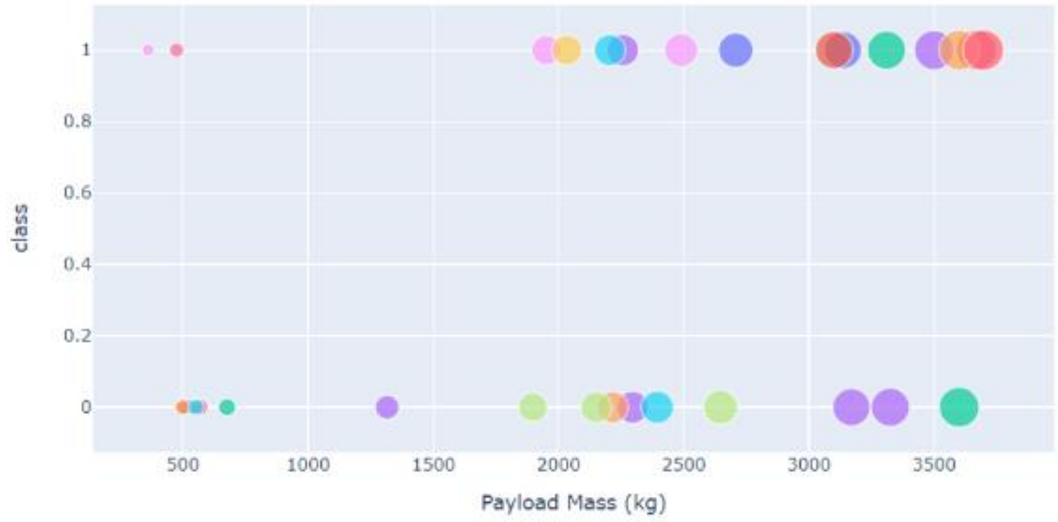


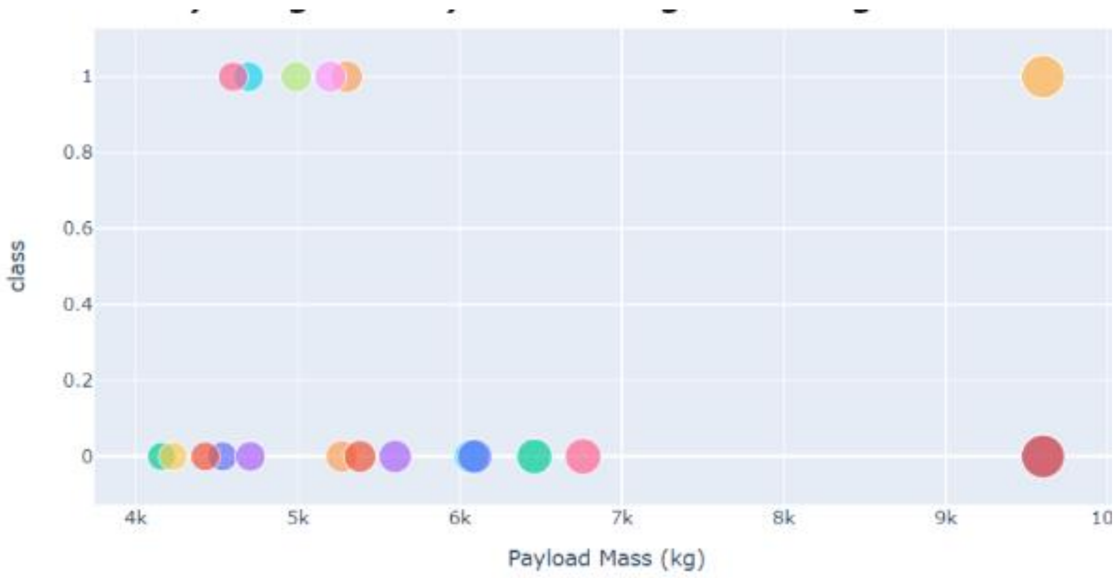
Launch site with the highest launch success ratio



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

The success rates for low payload mass is higher than the heavy payload mass.





Predictive Analytics Results

Classification Accuracy

Out of our four models, using Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Decision Tree, the **Decision Tree classifier model proved the highest classification accuracy** of approximately 0.8768.

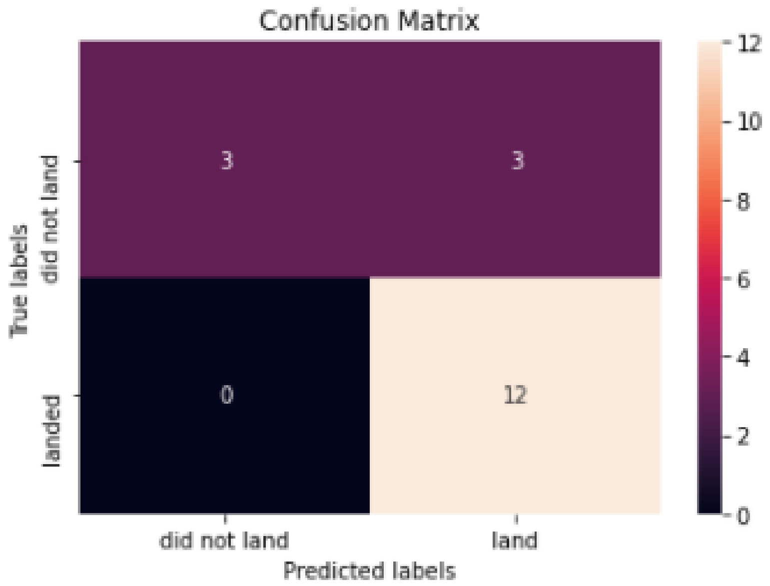
In [31]:
```python
models = {'LogisticRegression':lr_cv.best_score_,
          'SupportVector': svm_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'KNeighbors':knn_cv.best_score_}

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

```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'min_samples_leaf':
2, 'min_samples_split': 10, 'splitter': 'best'}
```

Confusion Matrix

The confusion matrix to the right of the decision tree classifier allows us compare the actual target values with those predicted by the machine learning model. Our confusion matrix tells us that there is some concern in the number of false positives, which are the number of failed landings marked as successful landings.



Conclusion

Based upon our analysis and predictions, we conclude that the larger the flight amount at a launch site, the greater the success rate at a launch site.

We also found that:

- Launch success rate had an overall increase from 2013 to 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the highest success rates.
- KSC LC-39A had the highest amount of successful launches of any sites.
- The Decision Tree classifier proved to be the best machine learning model for this task with a score of 0.8768.