


DATA SCIENCE CAPSTONE PROJECT

Marija Comarla
6/3/2024



OUTLINE

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

EXECUTIVE SUMMARY

- Summary of methodologies
 - Data Collection
 - Data wrangling
 - Exploratory Data Analysis with data Visualization
 - Exploratory Data Analysis with SQL
 - Building an interactive map with Folium
 - Building a dashboard with Plotly Dash
 - Predictive Analysis
- Summary of results
 - Exploratory Data Analysis results
 - Interactive maps and dashboards
 - Predictive results

INTRODUCTION

- Project background and context

In this capstone, we will 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; 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success on the first stage landing?
- Does the rate of success increase over time?

METHODOLOGY

- Data Collection Methodology
- SpaceX REST API
- Web Scrapping from Wikipedia
- Data Wrangling
- Filtering the data
- Using One Hot Encoding for classification models
- Perform Exploratory Data Analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

METHODOLOGY

DATA COLLECTION

Datasets are collected from SpaceX REST API and Web Scraping Wikipedia.

Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

DATA COLLECTION SPACEX API

Requesting rocket launch data from SPACEX API

Decoding the response as Json using `.json()` and turning it into Pandas dataframe using `.json_normalize`

Requesting information about the launches using the custom functions

Construct dataset using the data obtained into a dictionary

Filter the dataframe to only include Falcon 9 launches

Replacing the missing values of Payload Mass column with calculated mean using the `.mean()`

Export the data to CSV

DATA COLLECTION – WEB SCRAPPING

Requesting the
Falcon9 launch data
from Wikipedia

Create a
BeautifulSoup
object from the
HTML response

Extract column
names from the
HTML table
header

Create a data
frame by parsing
the launch HTML
tables

Constructing data
into a dictionary

Creating a data
frame from the
dictionary

Exporting the
data to CSV

DATA WRANGLING

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad. False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship. False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome of the orbits

Create a landing outcome label from Outcome column

Export data to CSV

EDA WITH DATA VISUALIZATION

Scatter Graphs:

Flight number vs. Payload Mass

Flight number vs. Launch Site

Launch Sites vs. Payload Mass

Flight Number vs. Orbit Type

Payload vs. Orbit Type

Scatter plots show relationship between variables.

Bar Graphs:

Success Rate vs. Orbit

Bar graphs show the relationship between numeric and categoric variables.

Line Graph:

Success Rate vs. Year

Line charts show trends in data over time.

EDA WITH SQL

Queries were written to extract information about:

- Launch sites
 - Payload masses
 - Dates
 - Booster types
 - Mission outcomes
-
- GitHub URL (EDA with SQL):

BUILT AND INTERACTIVE MAP WITH FOLIUM

Markers were added for launch sites and for the NASA Johnson Space Center

- Circles were added for the launch sites.
- Lines were added to show the distance to the nearby features:
 - Distance from CCAFS LC-40 to the coastline
 - Distance from CCAFS LC-40 to the rail line
 - Distance from CCAFS LC-40 to the perimeter road

BUILD A DASHBOARD WITH PLOTLY DASH

This dashboard application contains input components such as dropdown list and range slider to interact with a pie chart and scatter point chart.

- Added a Launch Site Drop-down Input Component to enable Launch Site selection.
- Added a Callback function to render success-pie-chart based on selected site dropdown to get the selected launch site from success-pie-chart and render a pie chart visualizing launch success counts.
- Added a Range Slider to select Payload to find if the variable payload is correlated to mission outcome.
- Added a callback function to render the success-payload-scatter-chart scatter plot with x axis to be the payload and y axis to be the launch outcome.

PREDICTIVE ANALYSIS (CLASSIFICATION)

Creating a Numpy array from the column the Class in data



Standardizing the data with Standard Scaler then fitting and transforming it



Using the function `train_test_split` to split the data into training and testing data.



Creating a logistic regression object and create a `GridSearchCV` object `logreg_cv` with `cv=10`.

Calculating the accuracy on the test data using the method `score`



Create a support vector machine object and `GridSearchCV` object `svm_cv` with `cv = 10`



Calculating the accuracy on the test data using the method `score`



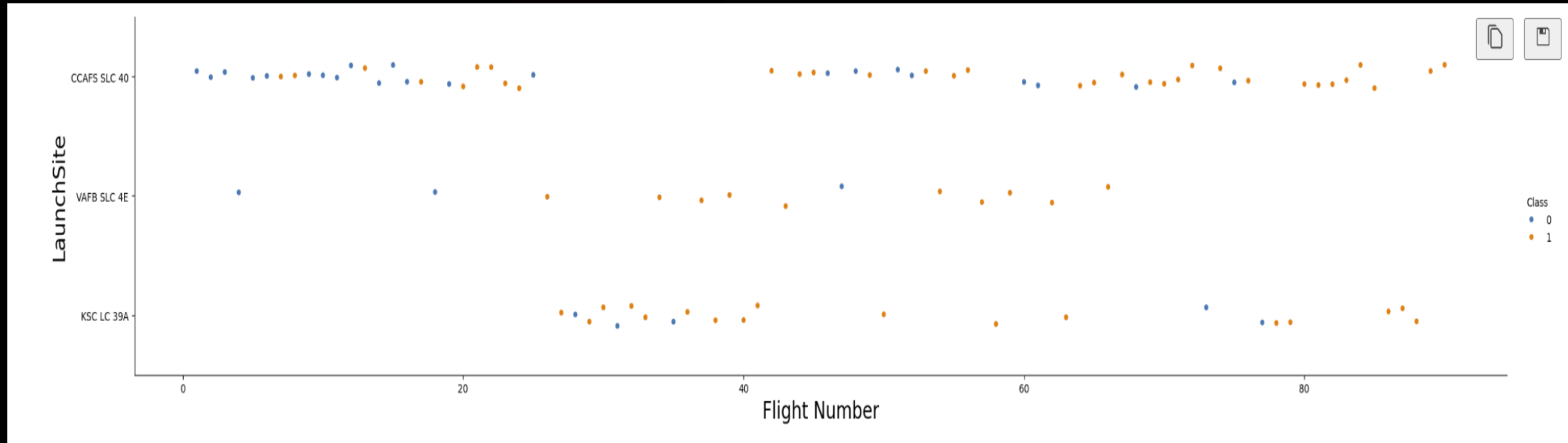
Create a decision tree classifier object and `GridSearchCV` object `tree_cv` with `CV=10`

RESULTS

- Exploratory Data Analysis Results
- Interactive analytics demo in screenshots
- Predictive analysis results

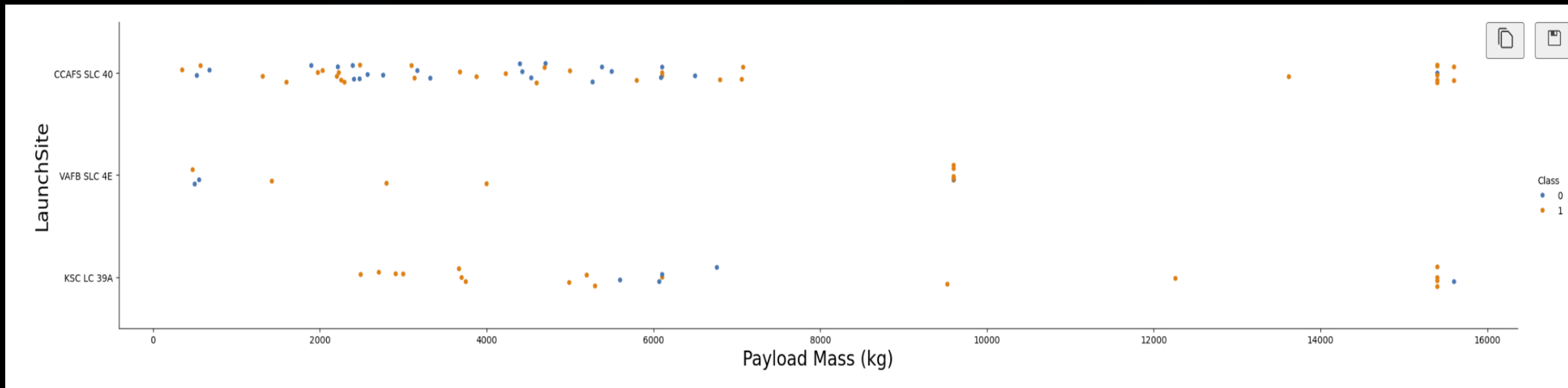
EDA WITH VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE



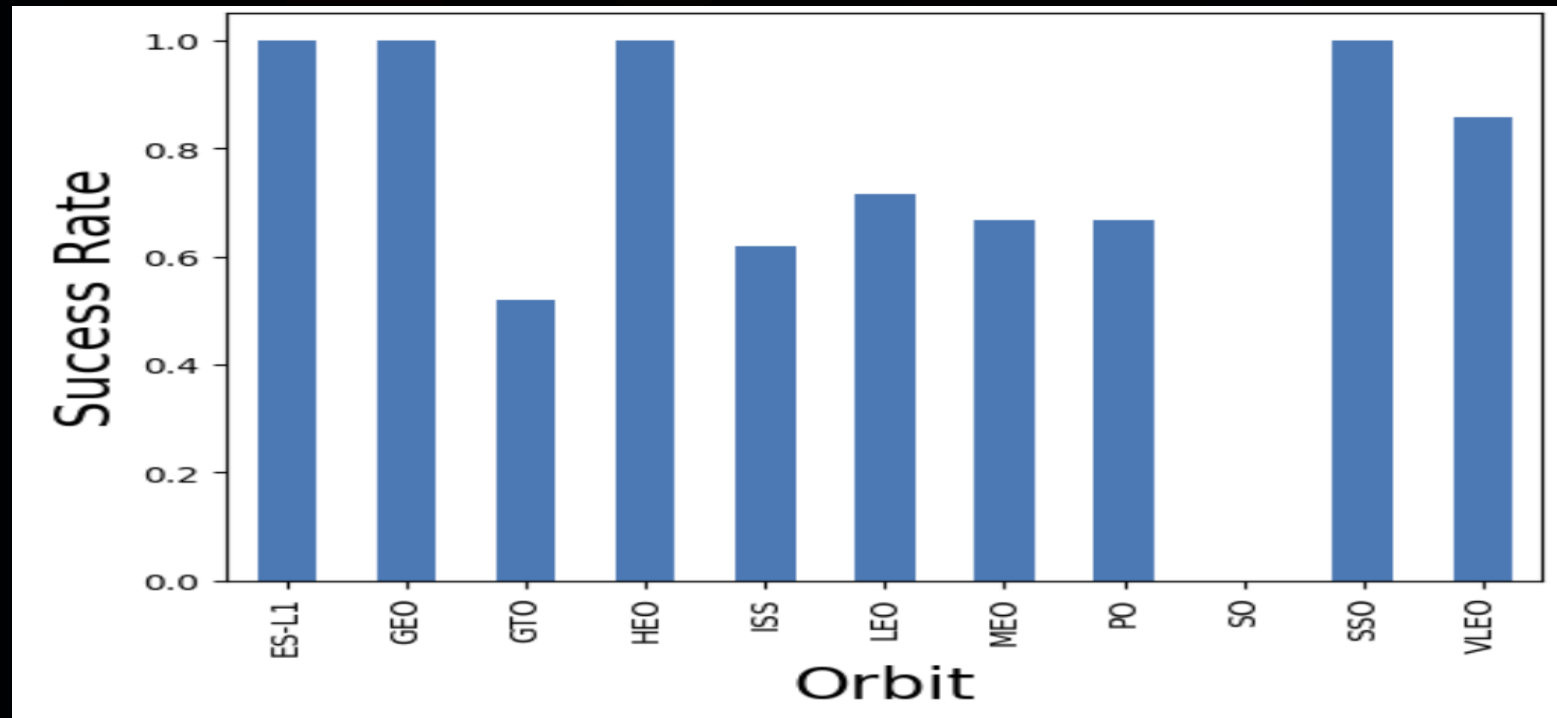
Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases.

PAYLOAD MASS VS. LAUNCH SITE



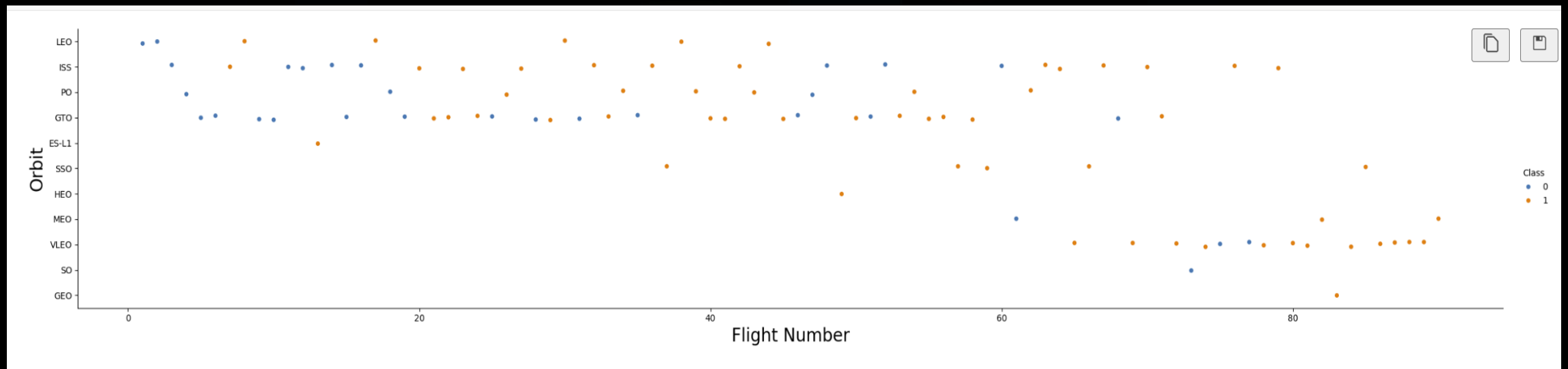
We observed, that for every launch site the higher the payload mass , the higher the success rate.

SUCCESS RATE VS. ORBIT TYPE



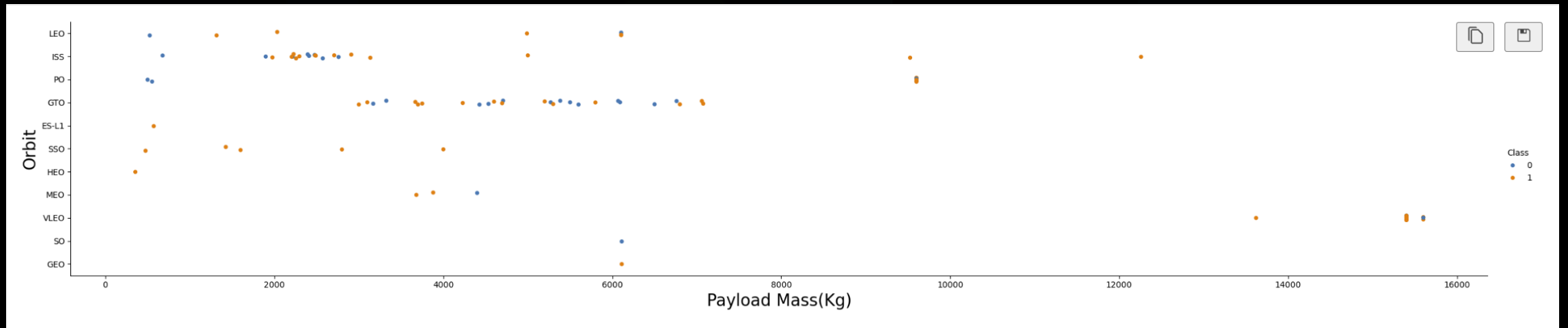
We observed that ES-L1, GEO, HEO, SSO have the highest success rate.

FLIGHT NUMBER VS. ORBIT TYPE



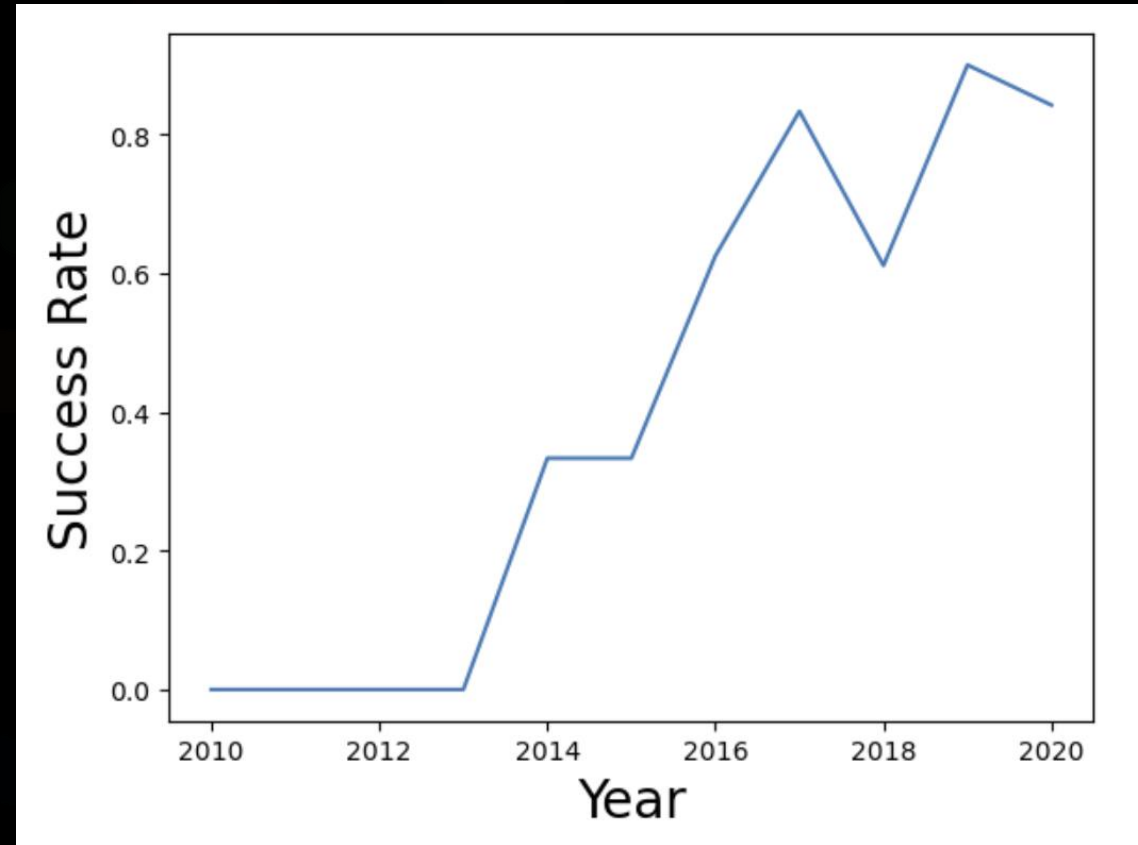
We observed, that the success rate increases with the number of flight for the LEO orbit.
On the other hand, there is no relationship between flight number when in GTO orbit.

PAYLOAD MASS VS. ORBIT TYPE



We observed, that heavy payloads have a negative influence on GTO orbits and positive influence on GTO and ISS orbits

LAUNCH SUCCESS YEARLY TREND



We observed, that since 2013 the success rate kept increasing until 2020.

EDA WITH SQL

ALL LAUNCH SITE NAMES

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

```
%sql select distinct Launch_Site from SPACEXTBL
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

| Launch_Site |
|-------------|
|-------------|

| |
|-------------|
| CCAFS LC-40 |
|-------------|

| |
|-------------|
| VAFB SLC-4E |
|-------------|

| |
|------------|
| KSC LC-39A |
|------------|

| |
|--------------|
| CCAFS SLC-40 |
|--------------|

LAUNCH SITE NAMES BEGIN WITH 'CCA'

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

```
[9]: %sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE 'CCA%' LIMIT 5
```

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

Done.

| [9]: | 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 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| | 2012-10-08 | 0:35:00 | 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

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

```
[10]: %sql SELECT SUM("PAYLOAD_MASS_KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

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

Done.

```
[10]: SUM("PAYLOAD_MASS_KG_")
```

45596

AVERAGE PAYLOAD MASS BY F9 V1.1

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

```
[11]: %sql SELECT AVG("PAYLOAD_MASS_KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

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

Done.

```
[11]: AVG("PAYLOAD_MASS_KG_")
```

2534.6666666666665

FIRST SUCCESSFUL GROUND LANDING DATE

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

Hint: Use min function

```
%sql select min(Date) from SPACEXTBL where "Landing_Outcome" = "Success (ground pad)"
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

```
min(Date)
```

```
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

```
%sql select Booster_Version from SPACEXTBL where "Landing_Outcome"="Success (drone ship)"and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000
```

```
* sqlite:///my\_data1.db
```

Done.

| Booster_Version |
|-----------------|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOME

```
%sql select count(*) from SPACEXTBL where "Mission_Outcome" like "Success%"
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

```
count(*)
```

```
100
```

```
%sql select count(*) from SPACEXTBL where "Mission_Outcome" like "Failure%"
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

```
count(*)
```

```
1
```

BOOSTER CARRIED MAXIMUM PAYLOAD

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

```
[21]: %sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS_KG_=(select max(PAYLOAD_MASS_KG_) from SPACEXTBL);
* sqlite:///my_data1.db
Done.
```

```
[21]: boosterversion
```

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

```
%sql select substr(Date, 6,2) as Month, Booster_Version, Launch_Site from SPACEXTBL where substr(Date,0,5)='2015' and "Landing_Outcome" = "Failure (drone ship)"
```

* [sqlite:///my_data1.db](#)

Done.

| Month | Booster_Version | Launch_Site |
|-------|-----------------|-------------|
| 01 | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | F9 v1.1 B1015 | CCAFS LC-40 |

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

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.

```
[24]: %sql SELECT "Landing_Outcome", count(*) as count_outcomes FROM SPACEXTBL WHERE DATE between '04-06-2010' and '20-03-2017' group by "
```

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

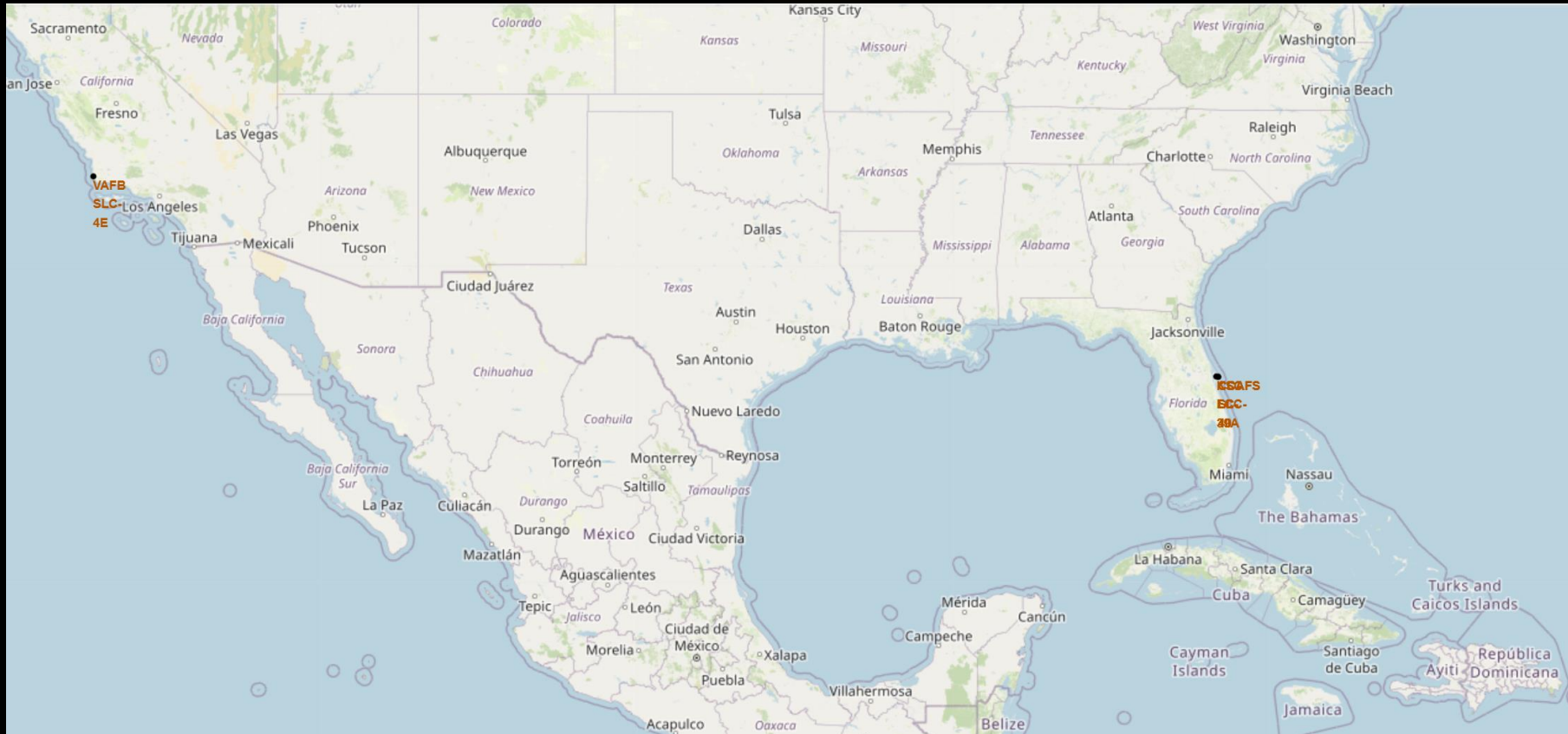
Done.

```
[24]: "Landing_Outcome" count_outcomes
```

INTERACTIVE MAP WITH FOLIUM

FALCON 9 LAUNCH SITE LOCATIONS

- VAFB SLC-4E (California, USA)
 - Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A (Florida, USA)
 - Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 (Florida, USA)
 - Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 (Florida, USA)
 - Cape Canaveral Air Force Station Space Launch Complex 40

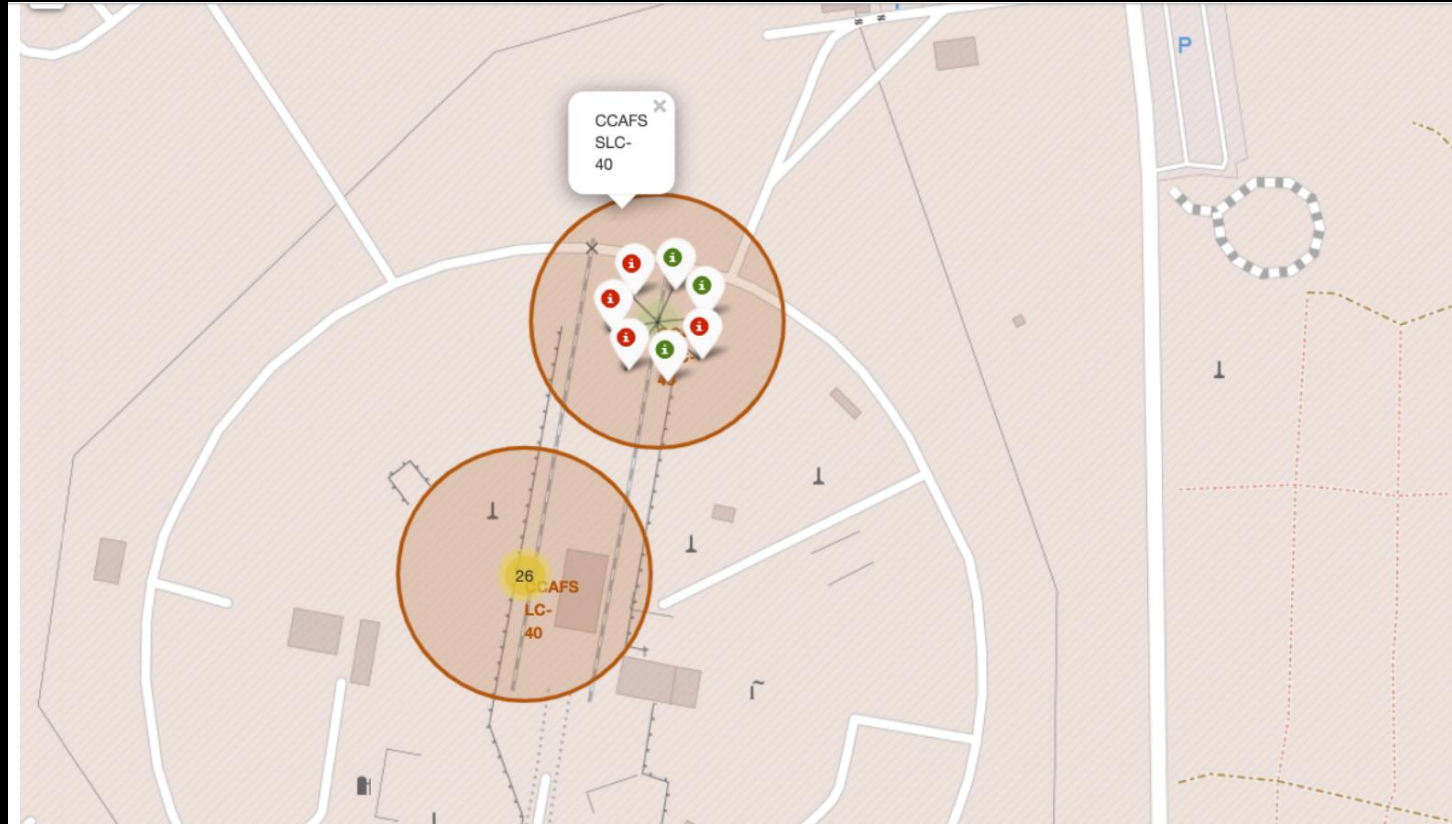


MAP MARKERS OF SUCCESS/FAILED LANDINGS

The markers display the mission outcome (success/failure) for Falcon 9 first stage landings.

Green marker = Successful launch

Red marker = Failed launch



BUILD A DASHBOARD WITH PLOTLY DASH

LAUNCH SUCCESS COUNT FOR ALL SITES

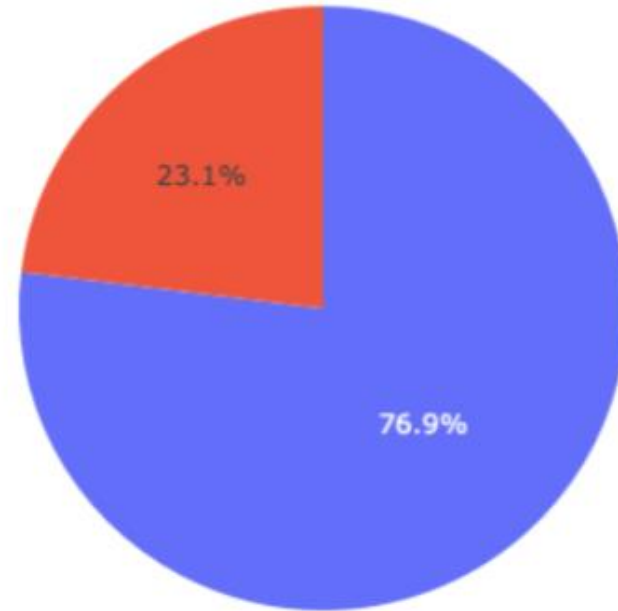
Total Success Launches by Site



Based on the chart, KSC LC-39A has the most successful launches.

LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

Total Success Launches for Site KSC LC-39A



Based on the chart, KSC LC-39A has the highest launch success rate.

PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

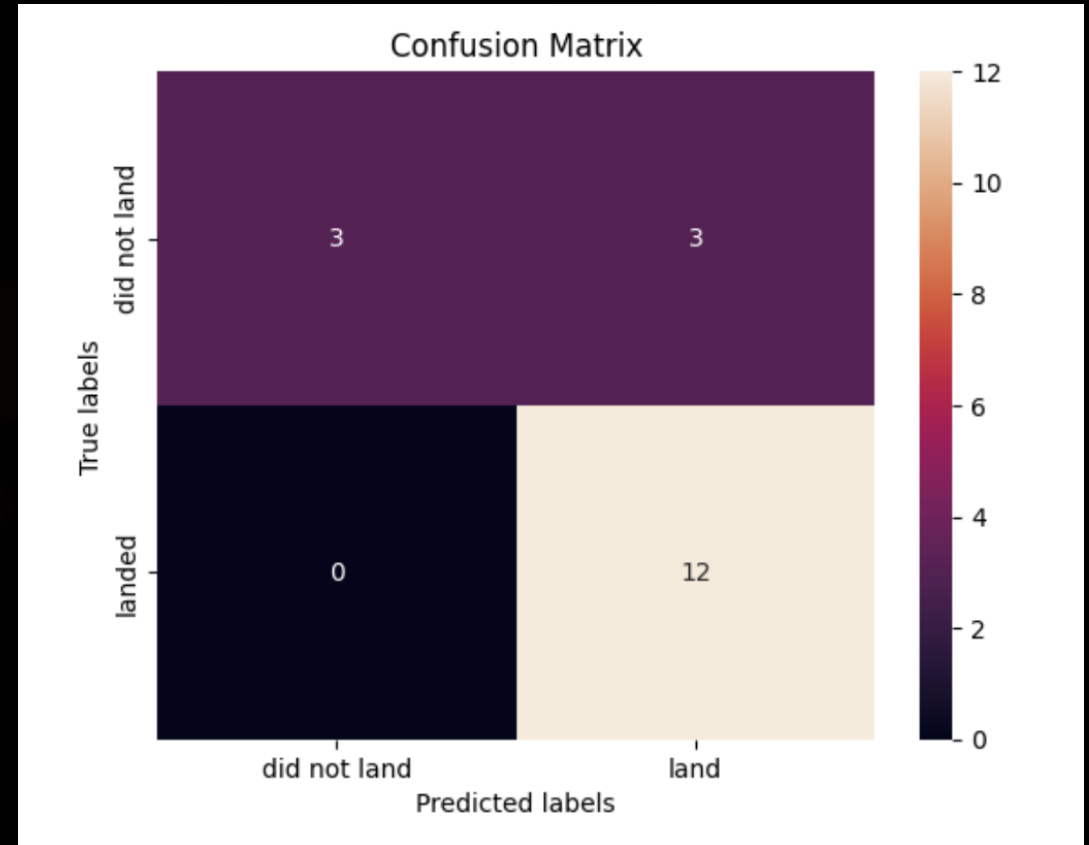
```
Accuracy for Logistics Regression method: 0.8333333333333334  
Accuracy for Support Vector Machine method: 0.8333333333333334  
Accuracy for Decision tree method: 0.8333333333333334  
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

All models performed equally well.

CONFUSION MATRIX

Shown here is the confusion matrix for the Logistic Regression model.

- Confusion matrices can be read as:
- Prediction Breakdown:
- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives



CONCLUSION

- SpaceX Falcon 9 first stage landing outcomes success rate increases over the years.
- Launches with a low payload mass show better results than launches with a larger payload mass
- The machine learning models can be used to predict future SpaceX Falcon 9 first stage landing outcomes.

THANK YOU!