IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jennifer Vivanco
31/05/2025

Outline

Executive Summary

Summary of methodologies

- Observed Methodologies
- Data Collection
- Exploratory Data Analysis with Visualization
- Predictive Analysis

Summary of all results

- Geospatial Analytics
- Interactive Dashboard
- Predictive Analysis of Classification Models
- Exploratory Data Analysis

Introduction

Modern space missions demand substantial financial investment and are highly complex Conducting a thorough analysis will enable accurate predictions for rocket launches, helping to optimize time, cost, and resource allocation According to available data, SpaceX's Falcon 9 rockets are launched at a relatively low cost.

The primary objective of this project is to evaluate the likelihood of a successful landing of Falcon 9 ’s first stage To achieve this, a data driven model will be created to predict the success rate of future launches, using historical data The insights generated will enhance strategic planning and decision-making processes.

Section 1

Methodology

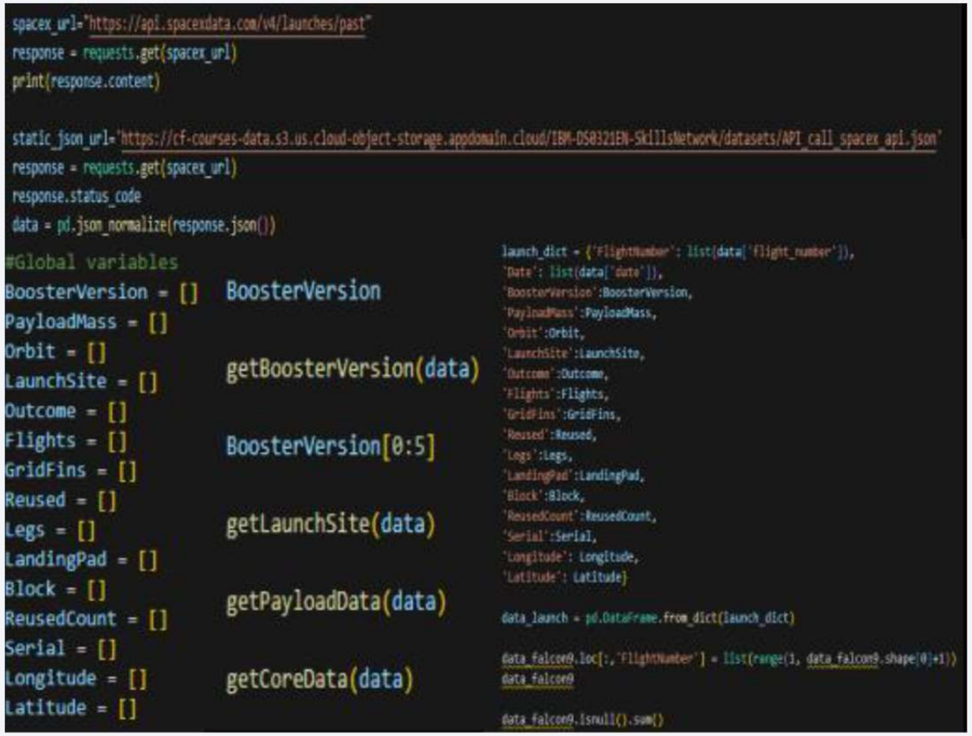
Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX rest API
  - Web Scrapping
- Perform data wrangling
  - Filtering the data
  - Conversation Method
  - Label 1 and 0 corresponding
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Manipulation and evaluation of a dataset in SQL
  - Visualization libraries to identify patterns
- Perform interactive visual analytics using Folium and Plotly Dash
  - Folium analysis is applied
  - Plotly Dash interactive panel
- Perform predictive analysis using classification models
  - Scikit-Learn libraries for standardizing and processing data, using training and testing methods
  - Confusion matrix plots for classification models

Data Collection – SpaceX API

- Data is gathered from various sources, including launch sites, payloads, rocket types, and mission outcomes.
- The collected data is then converted into JSON format and processed using Pandas for parsing and manipulation.
- Missing or incomplete data points are identified and addressed during the cleaning process.
- A new, refined dataset is generated based on the cleaned data.
- The dataset is filtered to retain only the necessary values for further analysis.



Data Collection – Scraping

- A URL containing the release history is provided, allowing for the retrieval of its HTML content.
- BeautifulSoup is utilized to facilitate navigation through and analysis of different HTML elements.
- Relevant HTML tables are extracted for further processing.
- Specific columns are selected to structure the data for storage.
- Thorough data analysis and cleaning are performed to ensure accuracy and completeness.

```python
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
html_data = requests.get(static_url)
html_data.status_code

soup = BeautifulSoup(html_data.text)
soup.title

html_tables = soup.find_all('table')

first_launch_table = html_tables[2]
print(first_launch_table)

column_names = []
for element in first_launch_table.find_all('th'):
    name = extract_column_from_header(element)
    if name is not None and len(name) > 0:
        column_names.append(name)

launch_dict= dict.fromkeys(column_names)


del launch_dict['Date and time ( )']
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]

df=pd.DataFrame(launch_dict)
```

Data Wrangling

Space launch information in a dataset contains information about launch sites and destinations.

Data analysis

- Incomplete data detection
- Data type classification
- Launch site survey
- General orbit assessment
- Mission analysis results

Data Label and Final Set Preparation

- Create Labels
- 1 = Success
- 0 = Failure
- Calculate Success Rate
- Export Final Set

EDA with Data Visualization

- Bar charts Ideal for comparing values between different categories.
- Scatter plots Ideal for visualizing possible relationships between numerical variables.
- Line Charts Ideal for observing trends over time.

EDA with SQL

This process is executed using SQL to explore and extract specific information from a space mission database.

1. Extract the names of the launch sites used .
2. Select 5 records named "CCA“.
3. Calculate the total payload mass per rocket launched under the NASA mission (CRS).
4. Average payload weight of the F9 v1.1 booster.
5. Identify the date of the first successful landing.
6. Locate the names of boosters that achieved successful landings on unmanned missions (payloads between 4000 kg and 6000 kg).
7. Total number of failed and successful missions.
8. Determine which booster versions carried the heaviest payloads.
9. List failed landings on unmanned missions during 2015, including booster versions and launch sites.
10. Classify landings as successful or unsuccessful from June 4, 2010, to June 4, 2013.

Build an Interactive Map with Folium

Creation and entry of different objects on an interactive map to visualize launch data

- All launch sites are marked.

1. Each site has a marker and a circle.
2. Events are grouped because they share the same coordinates.
3. Each marker is assigned a color (red = 0, green = 1).
4. To estimate the proximity between launch sites, distances between points (latitude and longitude) are calculated.
5. Markers are created to show distances, and lines are added to visualize connections.

Build a Dashboard with Plotly Dash

- Pie Chart

The px.pie() function is used to represent the number of successful launches for each site, allowing us to visualize locations with the highest number of launches.

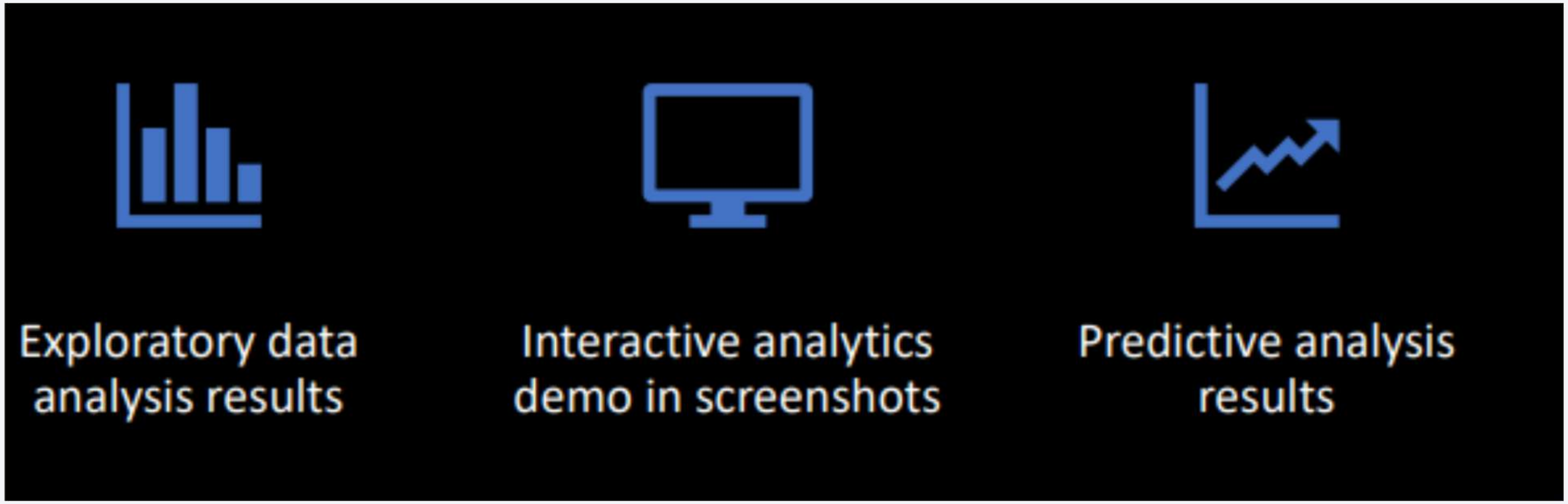
- Scatter Chart

Using px.scatter(), a graph is applied showing the relationship between the results of each mission (success or failure) and the payload transported. This tool allows us to analyze whether there is a correlation between these factors.

Predictive Analysis (Classification)

1. Load data with Numpyand Pandas, analyze the data to process what is necessary, group data for testing and training.
2. Model development, different machine learning algorithms are selected, GridSearchCVand hyperparameters are used for tuning.
3. Each model is evaluated and hyperparameters, accuracy metrics, graphics, and confusion matrix analysis are reviewed to see the performance.
4. Selecting the model to use after comparing accurate scores from other tested models.

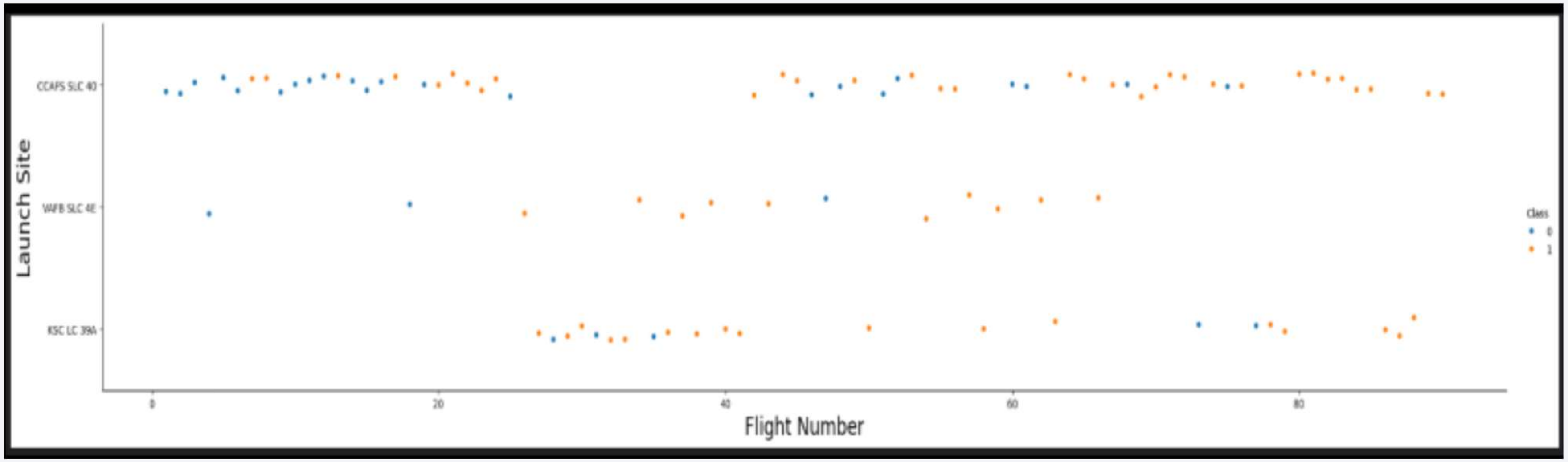
Results


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

Section 2

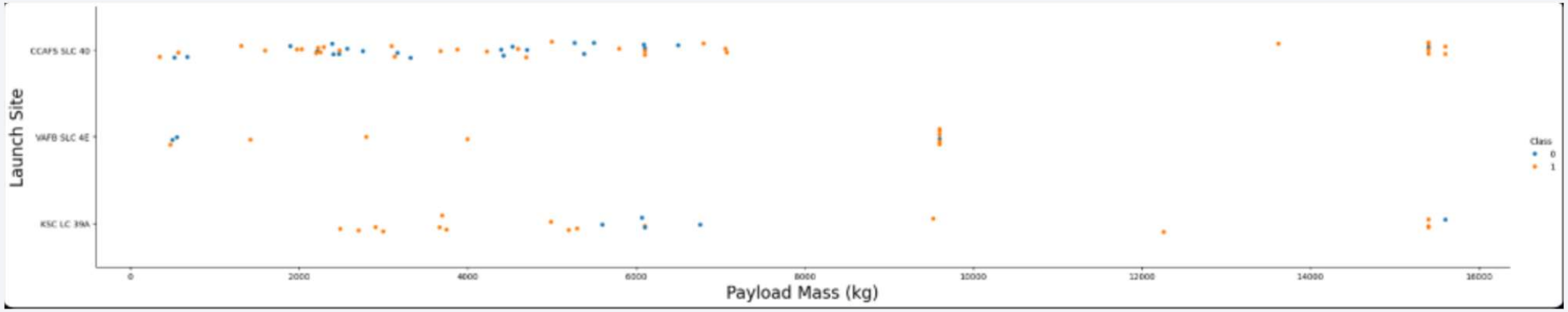
Insights drawn from EDA

Flight Number vs. Launch Site



- The first flights were unsuccessful, while the most recent were successful.
- CCAFS launch sites SLC 40 account for approximately half of all recorded launches.
- VAFB sites SLC 4E and KSC LC 39A have higher success rates compared to others.
- It can be argued that as time passes and new launches are conducted, this generates a continuously increasing probability of success.

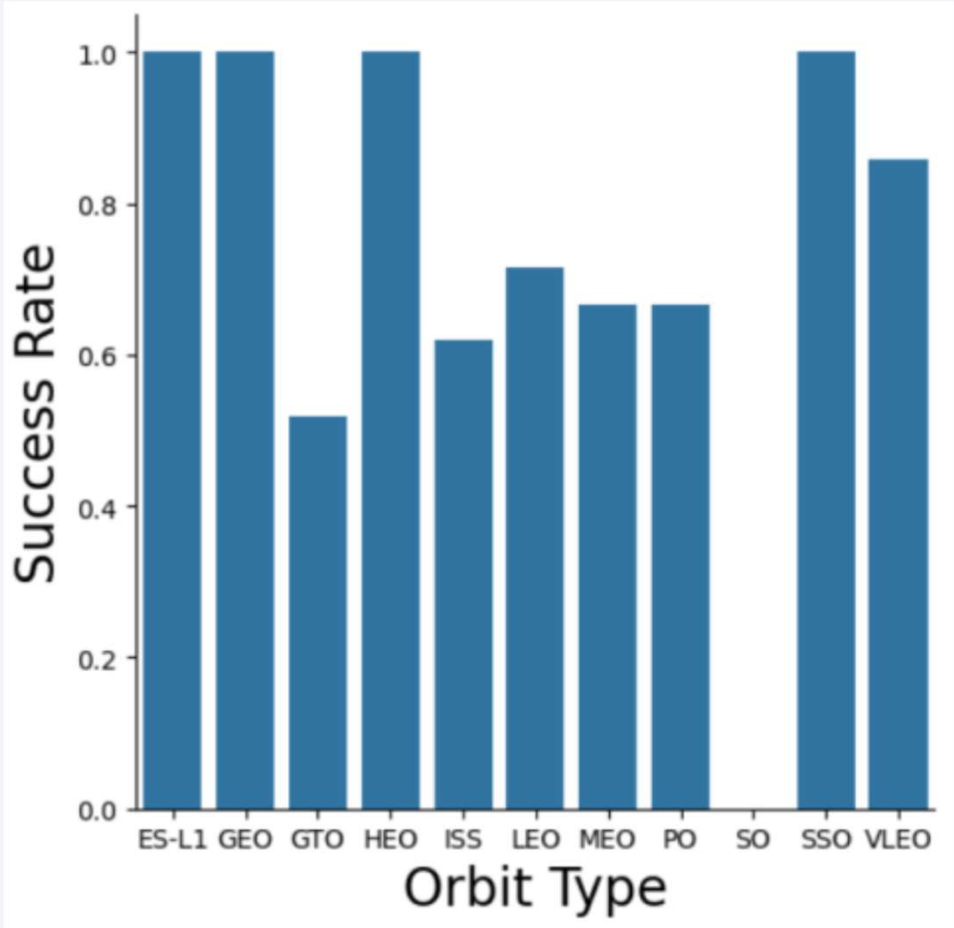
Payload vs. Launch Site



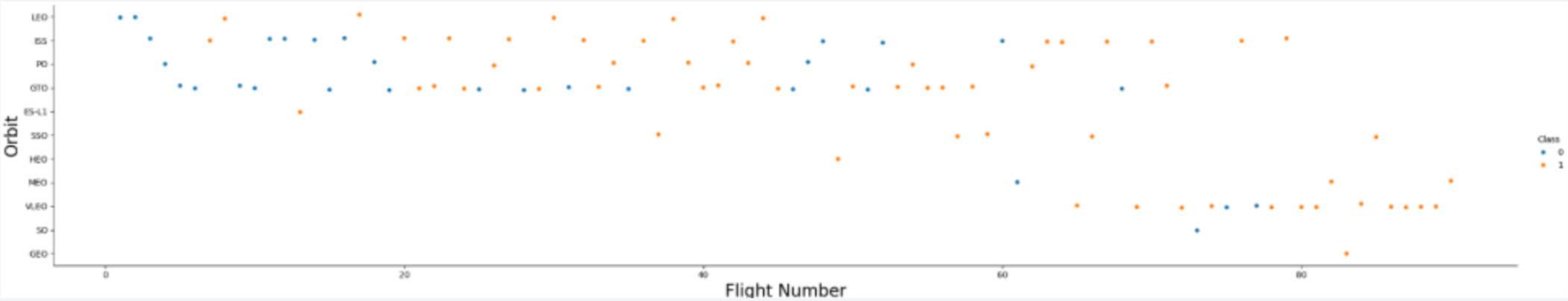
- The heaviest payloads (7,000 kg) generate higher success rates but are otherwise rare.
- There is no clear correlation between mass, payload, and success rate.
- Higher payloads at CCAFS SLC 40 are prone to greater landing success.
- CCAFS SLC 40 displays lighter payloads compared to other sites.

Success Rate vs. Orbit Type

Comparison of success rates between different types of orbits Orbits with a 100% success rate are observed VLEO showed over 80% SO remained with 0% success



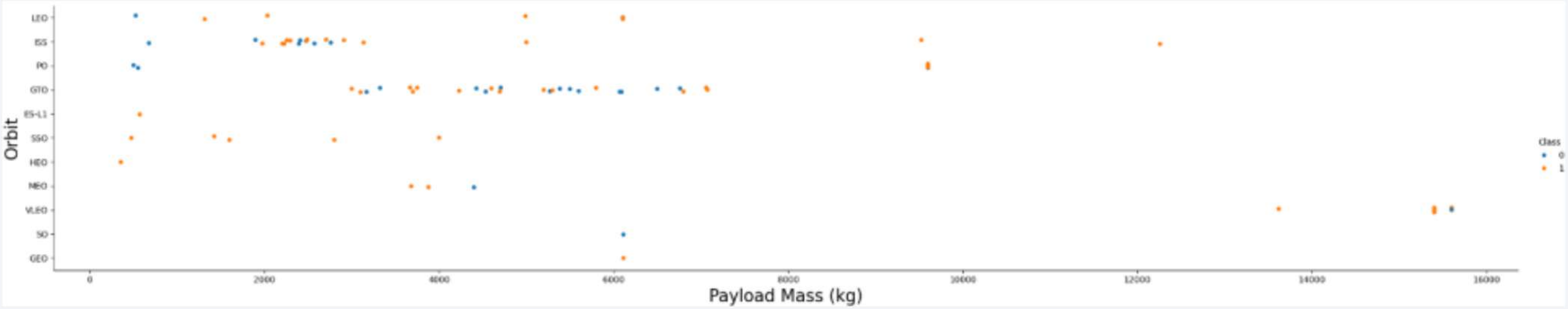
Flight Number vs. Orbit Type



Flight numbers affect success in different orbit types.

- The SSO orbit shows consistently successful flights.
- The GTO orbit shows no strong relationship between flights and success.
- HEO, GEO, and ES-L1 show high success rates, but only with one launch.

Payload vs. Orbit Type

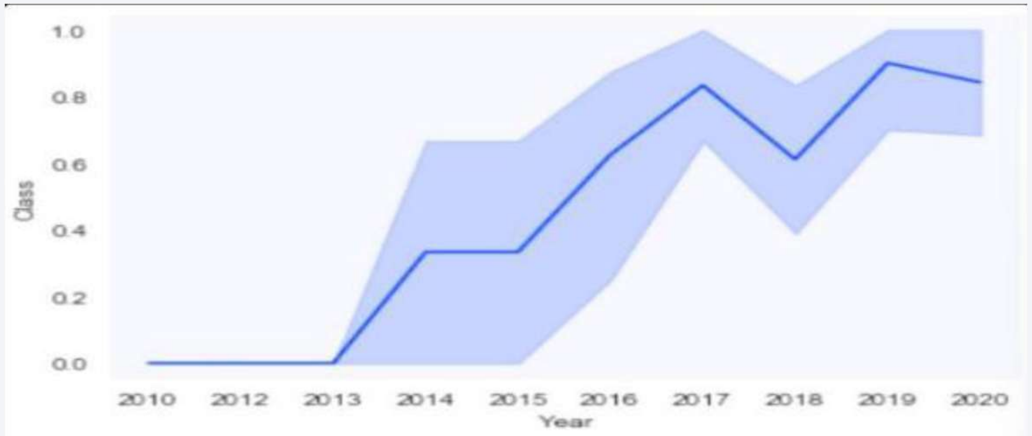


- It indicates, through analysis, the relationship between payload mass and orbit type success.
- The heaviest payloads (6,000 kg) are launched to PO, ISS, and LEO.
- Heavy payloads tend to have greater landing success.
- VLEO involves heavy payloads; this aligns with mission needs.

Launch Success Yearly Trend

This graph charts annual landing success trends over time.

- Between 2010 and 2010, no landings were successful.
- After 2013, the success rate increased considerably, although with some declines in 2018 and 2020.
- After 2016, the success probability remained within 50%.



All Launch Site Names

We use SQL to obtain the names of the launch sites using the DISTINCT function. This will allow us to extract information from the LAUNCH_SITE column.

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

```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

An SQL query is executed to find 5 records with launch site names starting with "CCA" by applying the LIKE wildcard to filter and also using LIMIT to display only the first 5 records.

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

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculating the Total Payload Carried by NASA Rockets
- The SUM variable is used in the payload mass column, then filtered by NASA missions, yielding a total of 45,596

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by the F9 v1.1 booster version is calculated.
- The AVG key is used, and after analysis, an average of 2928 is obtained.

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

average_payload_mass
2928

First Successful Ground Landing Date

The date of successful landing of a Land Platform is determined.

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

first_successful_ground_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The data was filtered to find the boosters that successfully landed with a payload of 4000 and 6000 kg, using keywords such as WHERE , AND , and BETWEEN.

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failed mission outcomes was calculated using keywords such as COUNT and GROUPBY. Groupings were then performed.

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- A list is created of the boosters that carried the maximum payload. A subquery is applied to retrieve the unique booster versions that achieved this goal.

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

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

2015 Launch Records

- Retrieved results for failed drone landings in 2015.WHERE, LIKE, AND, and BETWEEN clauses were used to filter out drone failures.

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (EXTRACT(YEAR FROM DATE) = '2015');
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The results of different landings between June 4, 2010, and March 20, 2021, are classified.
- WHERE and BETWEEN are used to filter time intervals, and GROUPBY and ORDER BY are applied to group and sort.

```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING__OUTCOME \
    ORDER BY TOTAL_NUMBER DESC;
```

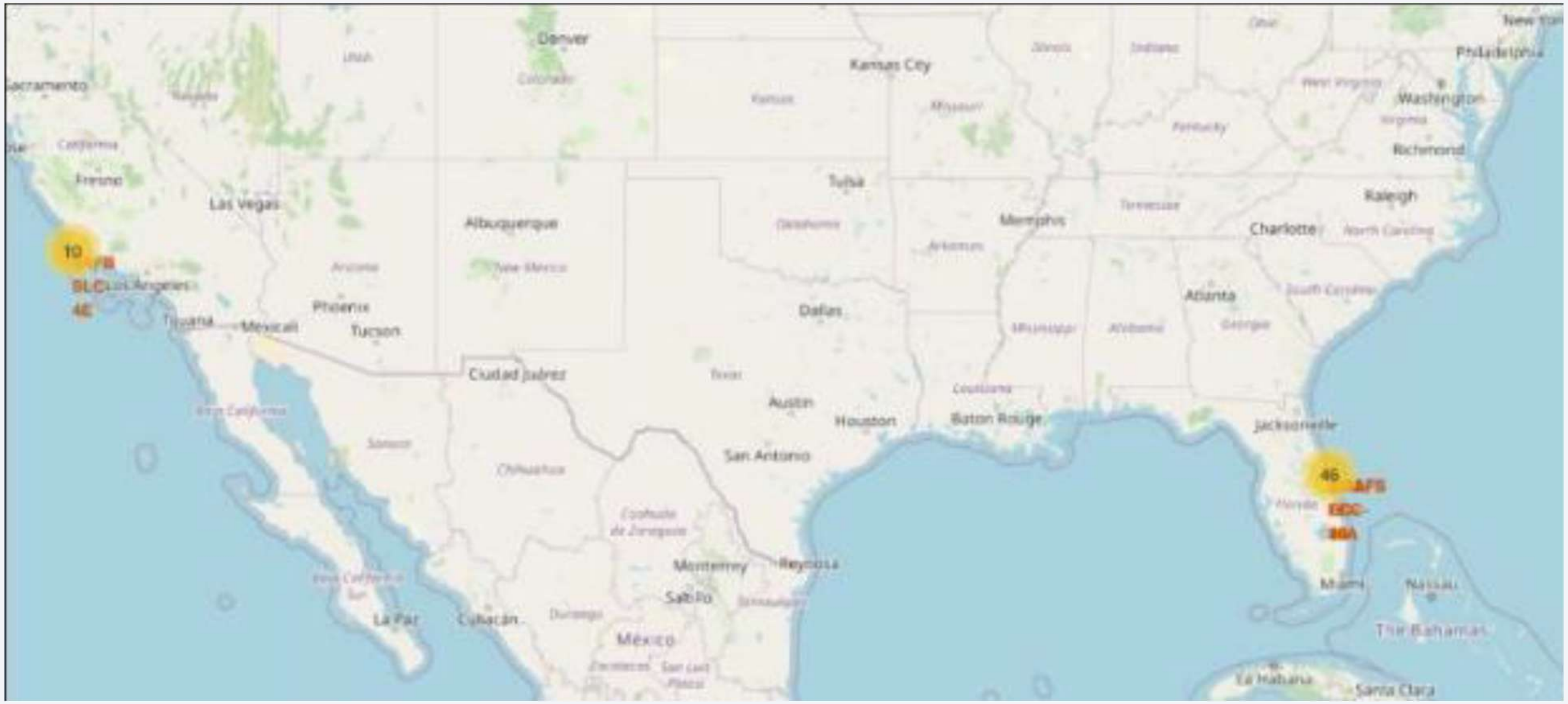
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch sites

This map shows SpaceX launch sites. (California and Florida)

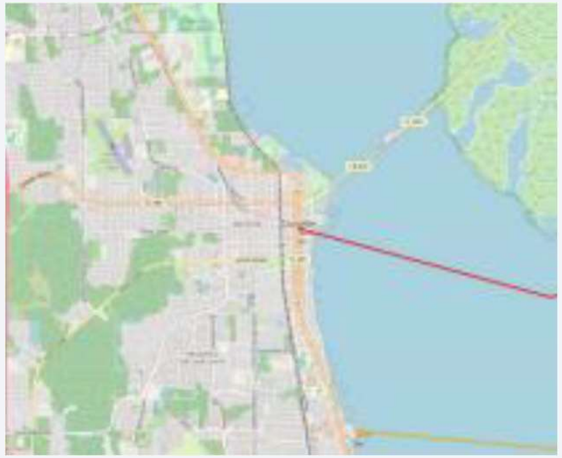


Launch sites with color labels

SLC-40 CCAFS launches, successful launches marked in green and unsuccessful launches in red

Launch point to other places of interest

Florida is a used launch site, 20.28km away with roads, 16.32km near cities, and 14.99km near seacoasts.

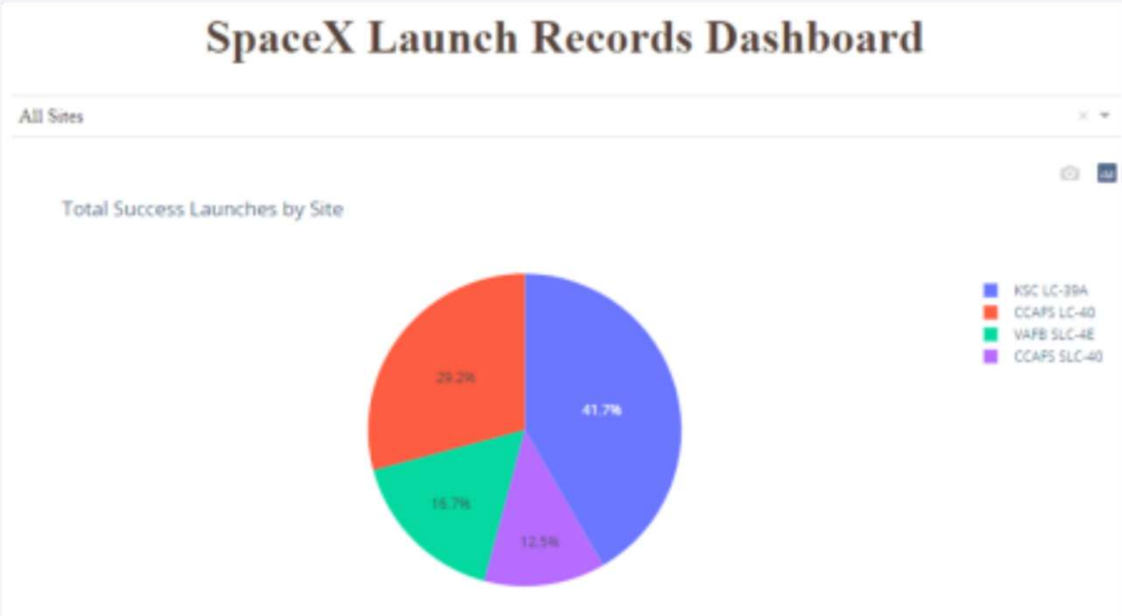


Section 4

Build a Dashboard with Plotly Dash

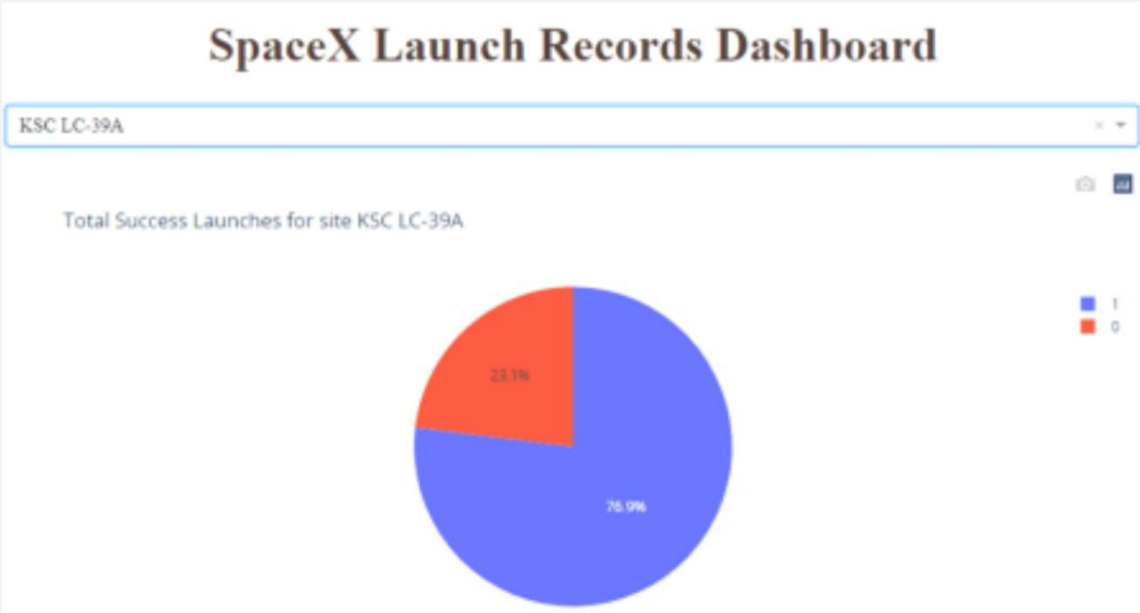
Successful launches across all sites

It can be seen from the graph that KSC LC 39 launch sites achieved the highest launch success with a total of 41.7%.



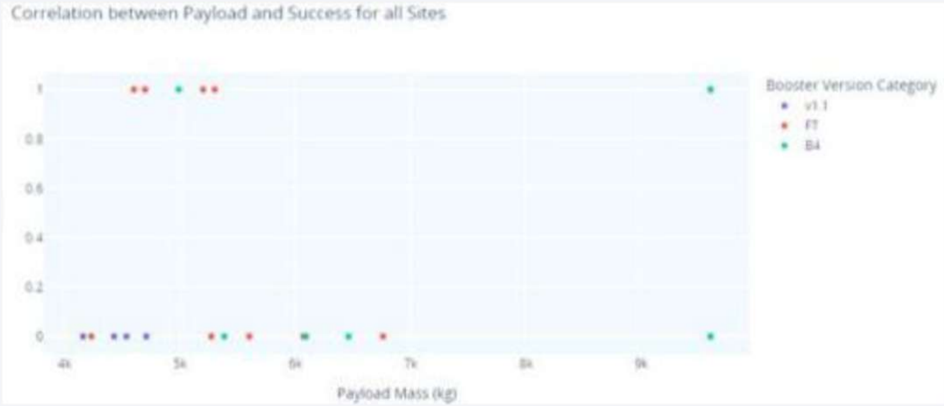
Successful Launches

- The KSC LC-39 A launch site also had the highest rate of successful launches, with a
- success rate of 76.9%. While its launch failure rate is 23.1%.



Scatter plot vs. launch result

In relation to success rates, it is analyzed that low-weight payloads are greater than large weight payloads; V1.0 and B5 type boosters have not had launches with massive payloads.

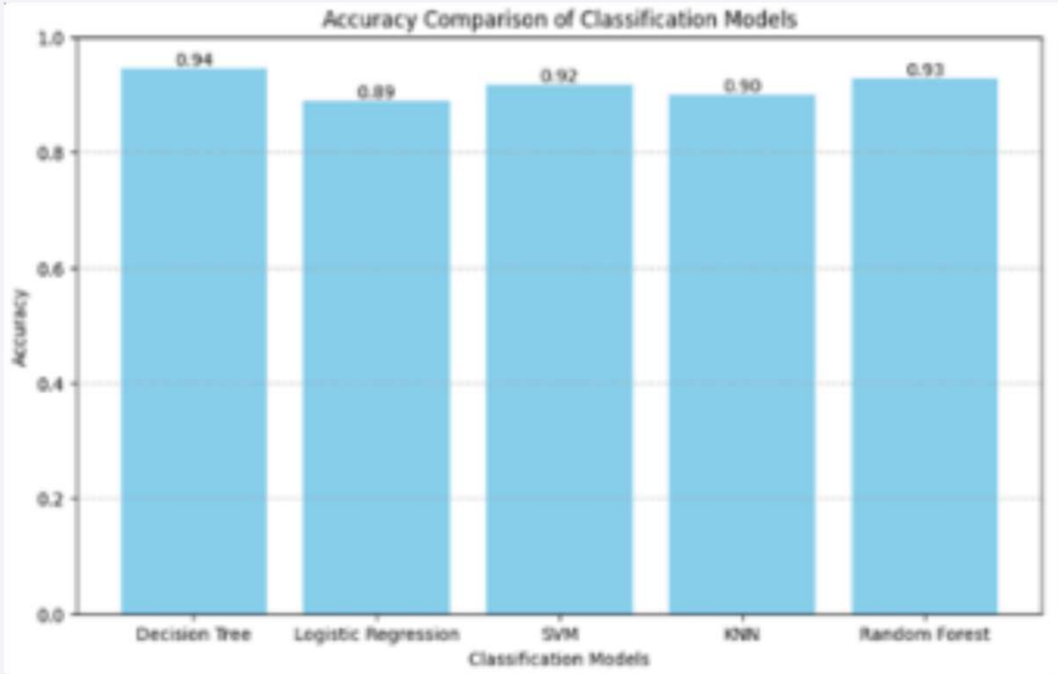


Section 5

Predictive Analysis (Classification)

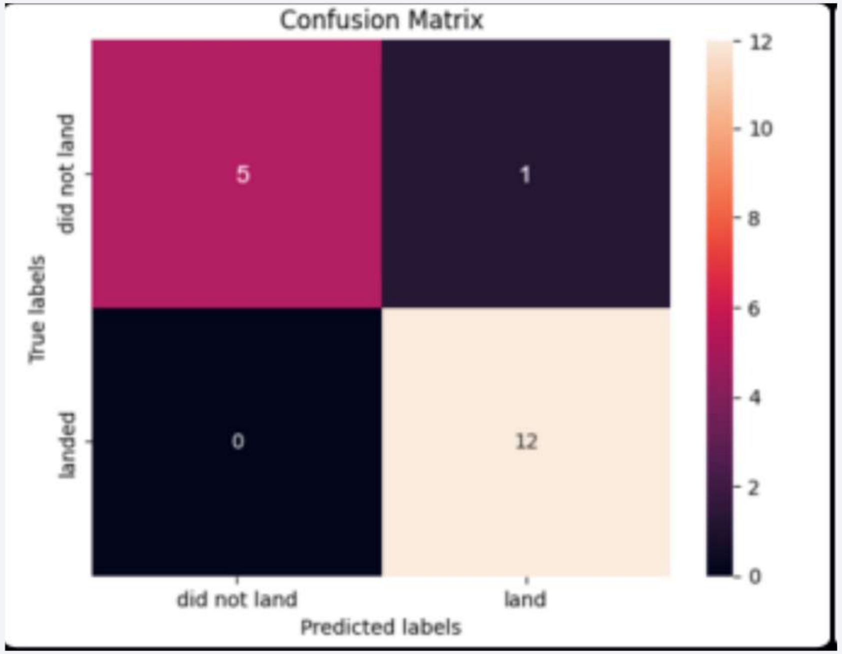
Classification Accuracy

Decision trees are the model with the highest classification accuracy; this model achieved an accuracy score of 94.44%.



Confusion Matrix

- We applied a decision tree as a performance classifier with a prior accuracy of 94.44%.
- (VP) 12 successful landings predicted.
- (VN) 5 failed landings predicted.
- (FP) No failed landings were predicted as successful.
- (FN) Only 1 landing was false negative.
- After analysis, only one misclassification was determined out of a total of 18 predictions and no false positives, indicating that it is a reliable model.



Conclusions

- We developed a comprehensive data analysis pipeline that combined web scraping, Python processing, exploratory analysis with interactive visualizations, and SQL queries to create a seamless workflow.
- Using tools like Plotly Dash and Folium, we were able to clearly and dynamically represent the data, helping to uncover patterns and insights that are easily interpretable.
- Through exploratory analysis and visualizations, we uncovered significant correlations between flight variables, orbit types, and launch success rates.
- We built and evaluated machine learning models to predict the likelihood of a successful launch, achieving high accuracy and identifying factors with strong predictive power.
- Beyond the technical analysis, this project offers valuable insights into how launch performance influences the company's logistics, reputation, and overall business strategy.

Appendix

In [5]:
```
from js import fetch
import io

URL1 = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_
resp1 = await fetch(URL1)
text1 = io.BytesIO((await resp1.arrayBuffer()).to_py())
data = pd.read_csv(text1)
```

In [6]:
```
data.head()
```

Out[6]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.C
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.C
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.C
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.C
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.C

In [7]:
```
URL2 = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_
resp2 = await fetch(URL2)
text2 = io.BytesIO((await resp2.arrayBuffer()).to_py())
X = pd.read_csv(text2)
```

In [8]:
```
X.head(100)
```

Out[8]:

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...	Serial_B1058
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0

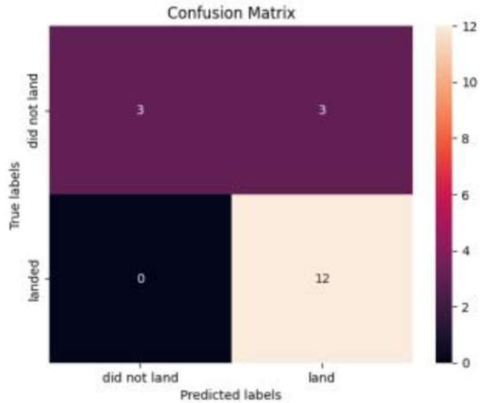
Calculate the accuracy on the test data using the method `score`:

In [18]:
```
logreg_accuracy = logreg_cv.score(X_test, Y_test)
logreg_accuracy
```

Out[18]: 0.8333333333333334

Lets look at the confusion matrix:

In [20]:
```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

True Postive - 12 (True label is landed, Predicted label is also landed)

Thank you!