Skills Network X coursera

Winning Space Race with Data Science

IBM DATA SCIENCE CAPSTONE PROJECT

Nguyen Minh Tri

2022-11-05

OUTLINE

Outline for this presentation

1 Executive Summary

2 Introduction

3 Methodology

4 Results

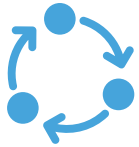
5 Conclusion

6 Appendix

1. Executive Summary

Overview

Gathering information about Space X and creating dashboards for visualization. Determine if SpaceX will reuse the first stage by training a machine learning model and use public information to predict.



Methodology

- Data was collected from SpaceX Wikipedia page.
- Data set was cleaned, analyzed, and visualized in Python and SQL.
- 4 machine learning models were built: Logistic Regression, SVM, Decision Tree, KNN.

Result

All models have similar results with accuracy rate of 83.34%. More data is needed for better model determination and accuracy.

2. Introduction

Background

- SpaceX, an American spacecraft manufacturer, space launch provider, founded in 2002 by Elon Musk. It manufactures the Falcon 9 and Falcon Heavy launch vehicles.
- SpaceY, a company that would like to compete with SpaceX, founded by Billionaire industrialist Allon Musk.
- The goal is to develop a model for SpaceY to predict whether SpaceX's Falcon 9 will successfully land in the next launch.

Problems

- The success rates heavily rely on different parameters including orbit type, payload mass, booster versions,...
- Find the classification method that performs best.

3. Methodology

Methodology Overview

Data Collection

Collecting data from SpaceX with web scraping and API.

Data Wrangling

Classifying landings as successful and unsuccessful

EDA with SQL

Exploratory Data Analysis with SQL using Google BigQuery

EDA with visualization

Exploratory Data Analysis with visualization using Matplotlib and Seaborn

Predictive Analysis

Perform predictive analysis using classification models.

Data Collection

With API

Request APIs

- Request for launch data

JSON normalize

- Decode the response content as a Json and turn it into a dataframe

Filtering

- Filter data to only include Falcon 9 launches

Model Creation

- Replace missing PayloadMass values with mean

API notebook

Data Collection

With Webscraping

Request HTML page

- Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

Soup object

- Create a BeautifulSoup object from a response text content

Extract

- Retrieve all tables and values

Data frame

- Cast dictionary to Pandas dataframe

Web Scraping notebook

Data Wrangling

Check the data type

- Identify which columns are numerical and categorical

Count values

- Determine the number of each orbit and outcome

Filter values

- Create a set of outcomes where the second stage did not land successfully

Create label

- Create a ‘Class’ column with Successful=1 & Failure=0.

Data Wrangling notebook

EDA with SQL

Perform the following SQL Queries in Google BigQuery

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string ' CCA '
3. Display NASA (CRS) total payload mass carried by boosters launched
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. Names of the boosters which have success in drone ship and have payload mass > 4000 and < 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass
9. List the failed landing outcomes in drone ship , their booster versions , and launch site names for in year 2015
10. Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 , in descending order

EDA with Visualization

Class
0
1

Please note that in the scatter plot, blue (0) indicates unsuccessful launch; orange (1) indicates successful launch.

Scatter Plot

Used to identify the type of relationship between two quantitative variables. In this case, the plot is used to view the relationship of different variables such as payload mass, orbit type, launch sites,...

Line Plot

Used to visualize a trend in data over intervals of time. In this case, the graph shows the success rate for each year.

Bar Plot

Used to show relationship between categorical variables and continuous variables. Bar plot is this project is intended to show the success rate of each orbit type.

EDA with interactive visual analytics

Pie Chart

- Show distribution of successful landings across all launch sites
- Select a particular launch site

Scatter Plot

- Show the success lauch across launch sites, payload mass, and booster version.
- Select a PauloadMass range with a slider between 0 and 10000 kg.

Predictive Analysis (Classification)

Create a NumPy array

- Split column 'Class', output a Pandas series and assign it as Y.

Standardization

- Fit and Transform features using Standard Scaler

Split the data

- Split the data into training and test data: X_train, X_test, Y_train, Y_test

Model Creation

- GridSearchCV to find best parameters

Model Evaluation

- Find the best model with accuracy and confusion matrix

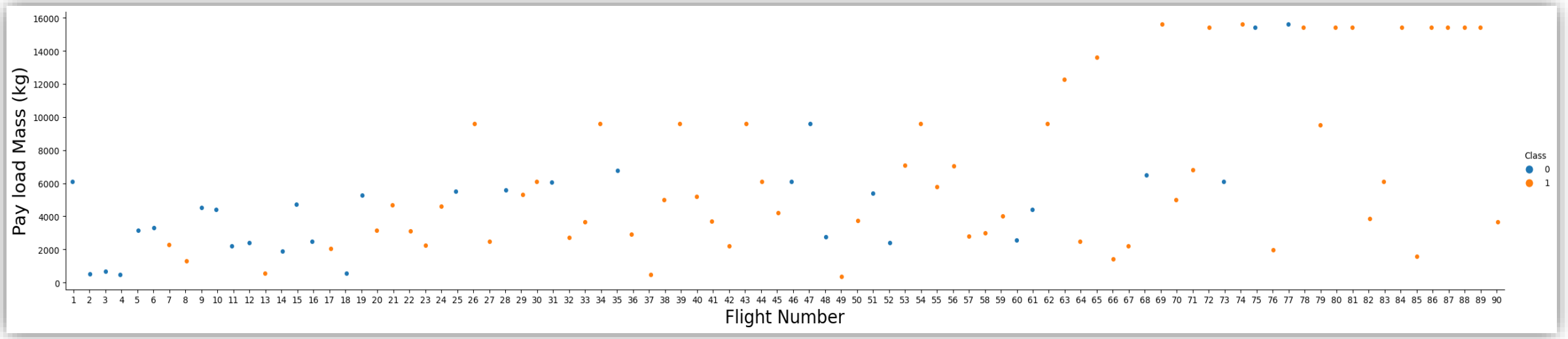
[Prediction notebook]


Resource > Your Chart
Business Chart - Visual
Business Chart
Space Usage (750 Mb)
375 Mb used
250 Mb available
Who is your audience and what are their needs? This can help you better articulate the benefits of doing business with you and deliver a smarter product or service.
Interactive User
1,505
18,321
Realtime Dashboard
37.91%
31.86%
30.23%
Marketing Chart
31.25% 43.75% 36.50% 41.25% 44.50% 50.25%

4. Results

Insights drawn from EDA with visualization

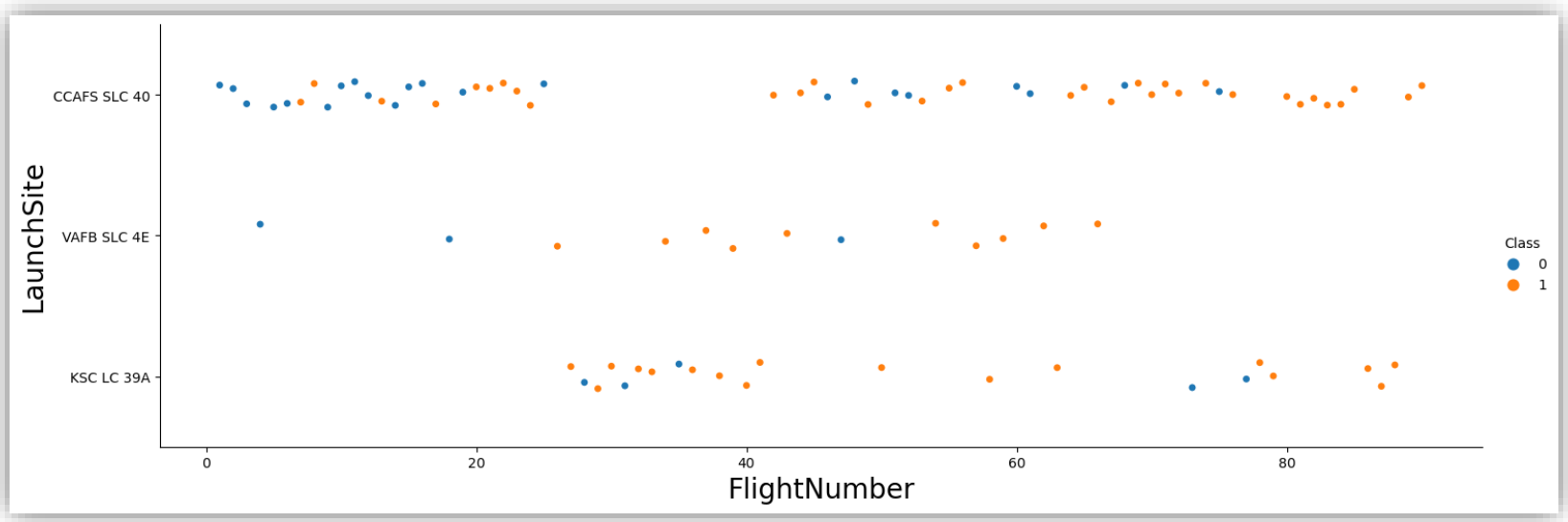
Flight Number vs Payload Mass



There was an increase in success rate over time (indicated in Flight Number).

The higher the flight number, the heavier the payload mass.

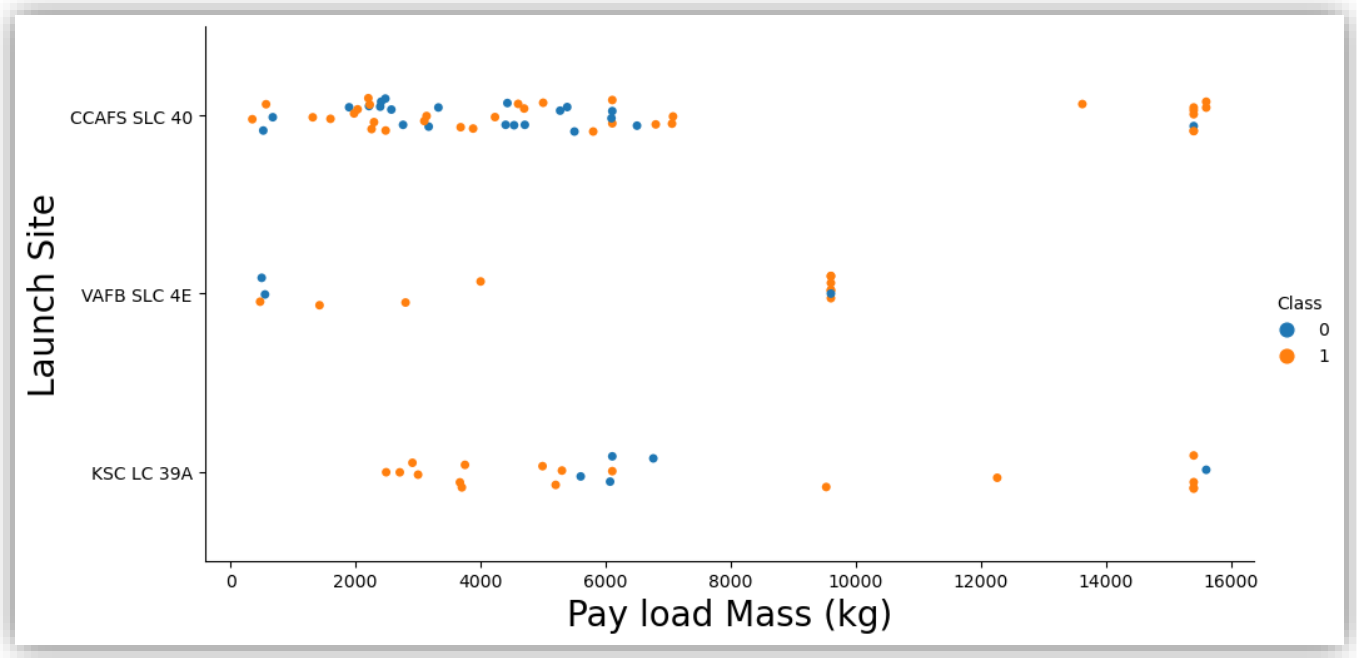
Flight Number vs Launch Site



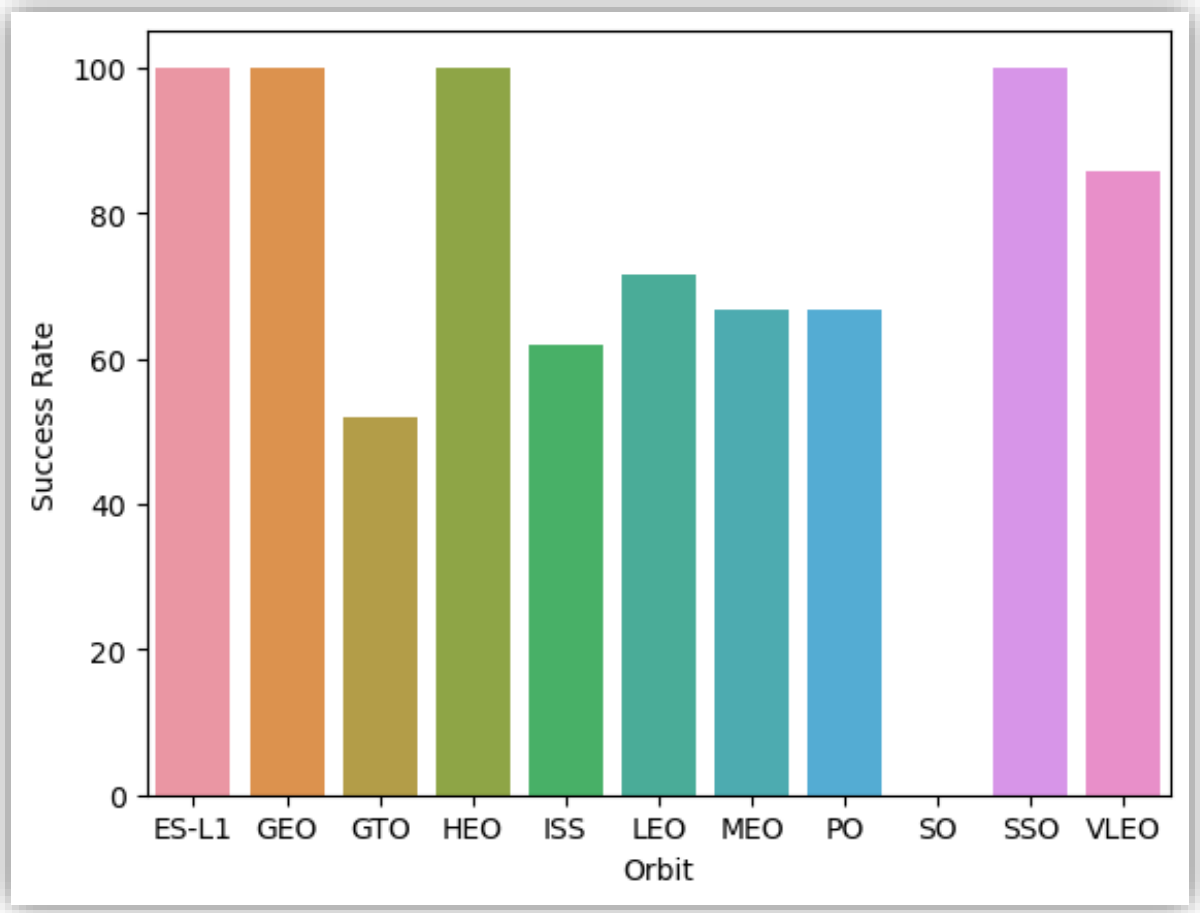
It can be seen from the scatter plot that for each launch site, the higher the flight number, the more likely it will land successfully.

Payload Mass vs Launch Site

- Payload Mass is seen to be between 0 and 6000 kg, although there are some outliers from CCAFS and KSC with almost 16000kg.
- We can assume from the plot that the heavier the payload mass, the more likely it will land successfully.



Success Rate vs Orbit Type



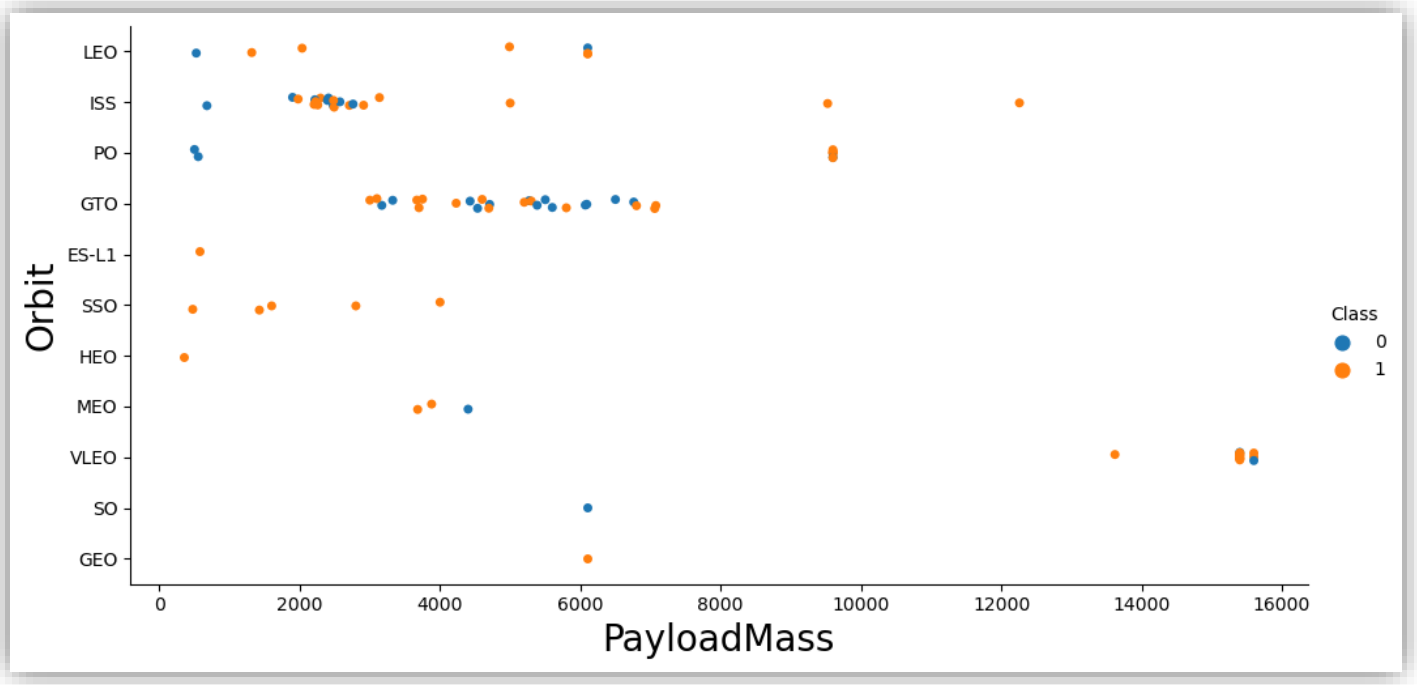
4 Orbit types which are ES-L1, GEO, HEO, and SSO have the highest success rate of 100%.

While VLEO, LEO, MEO, and PO also have a good rate (>60%), GEO and ISS don’t seem to be so (<=60%).

The SO type is apart from the others with a 0% rate of success.

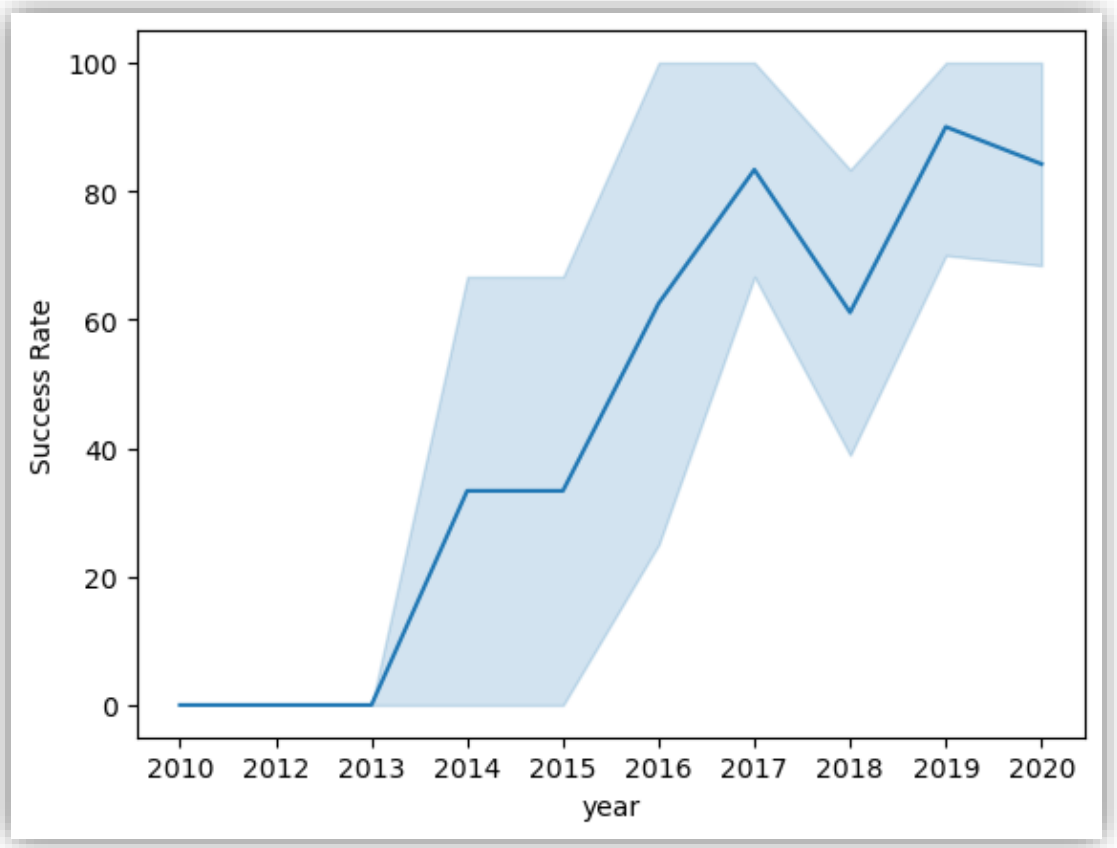
Payload Mass vs Orbit Type

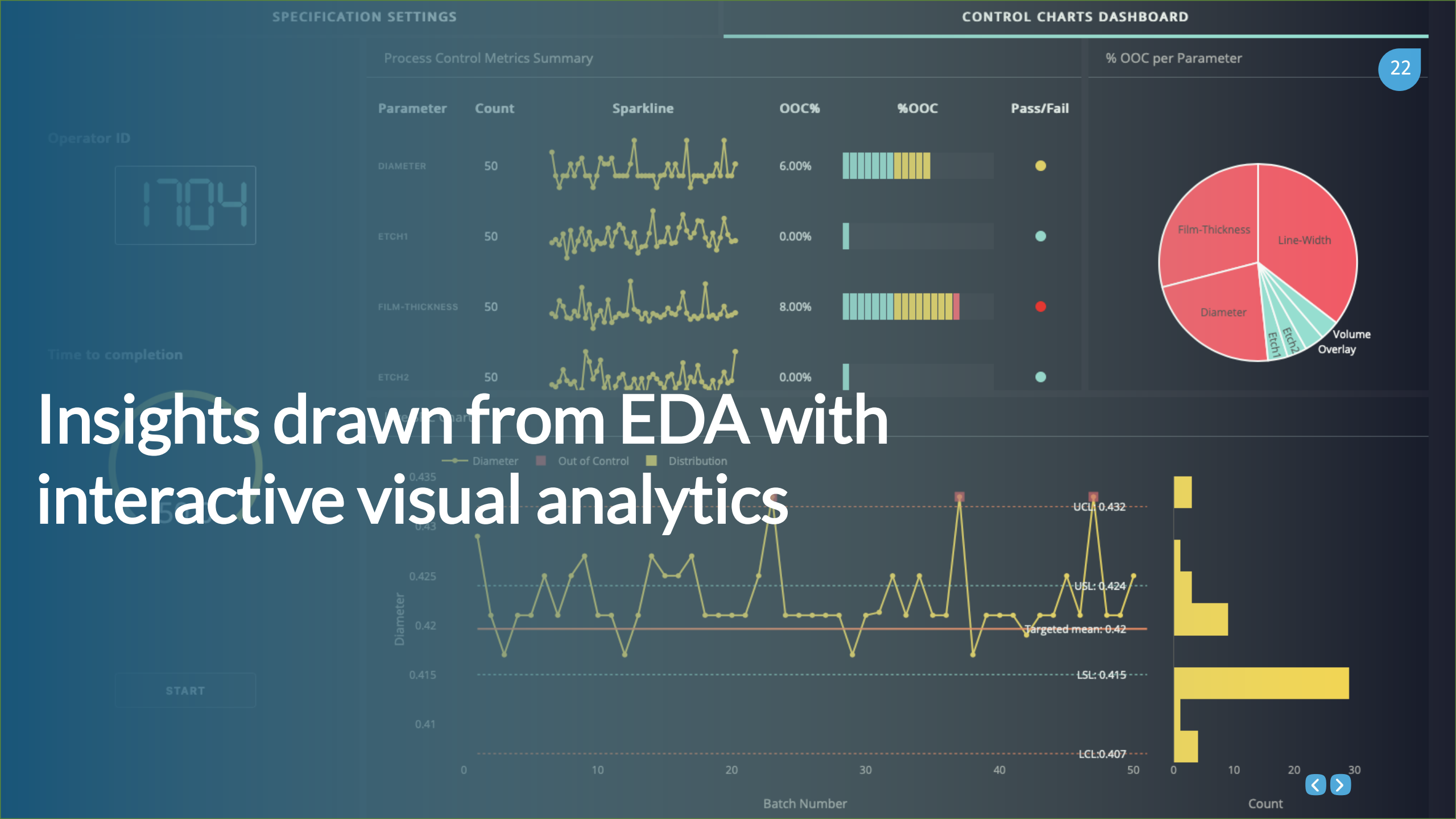
We can see from the plot that most orbit types have a payload mass of < 8.000kg with the exception of VLEO have a payload mass of ~16.000



Launch Success Yearly Trend

Overall, the success rate generally increased over time since 2013 with a slight drop in 2018. Whereas the rate in recent years reached a peak of 90%, the percentage from 2010 to 2013 remained the same with no flight landed.

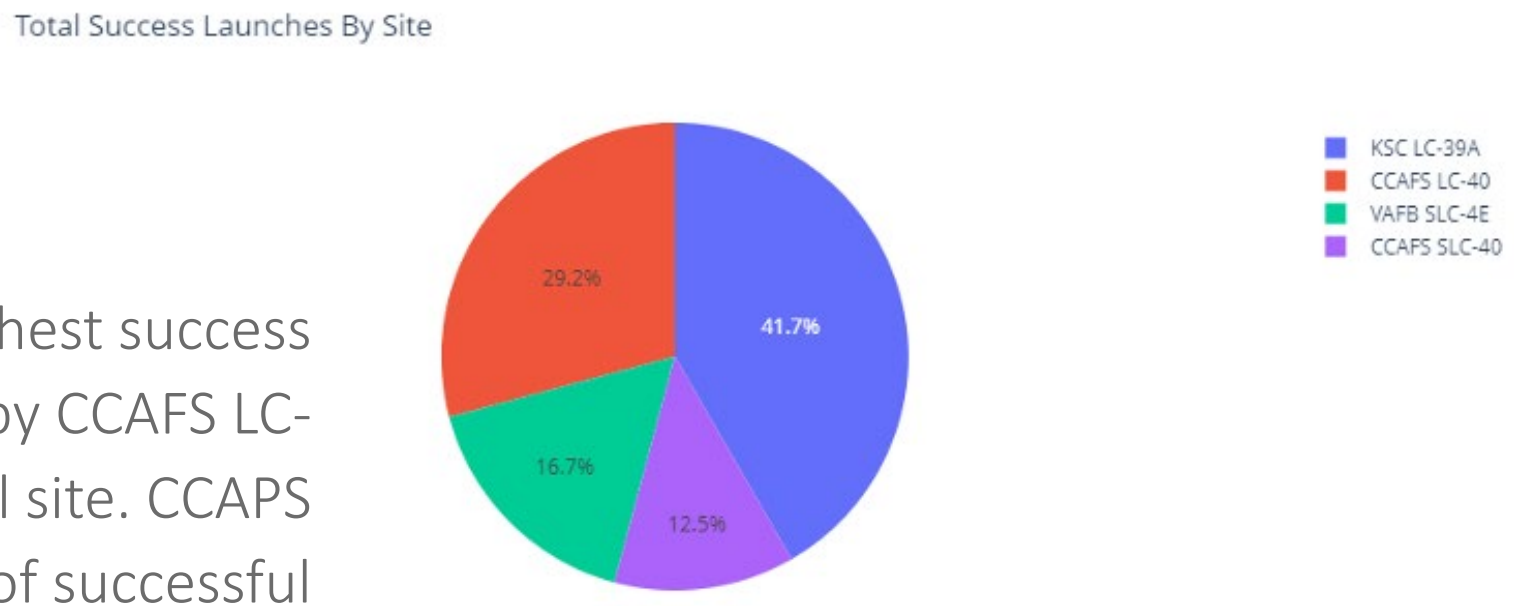




Insights drawn from EDA with interactive visual analytics

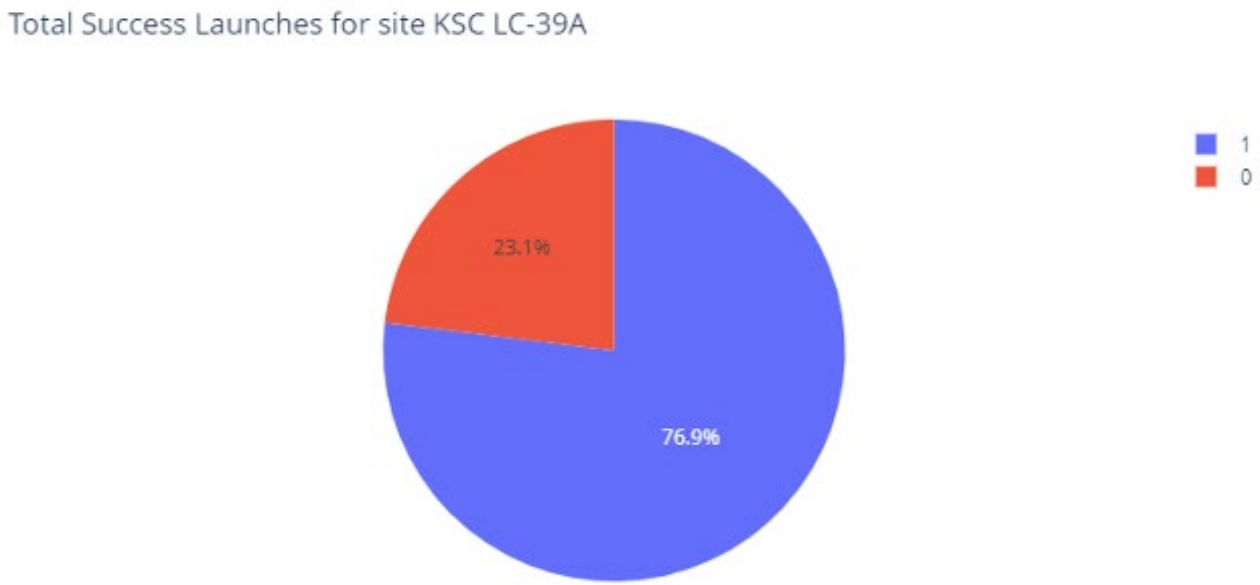
Success Rate across Launch Sites

KSC LC-39A has the highest success launches with 41.7% followed by CCAFS LC-40 with over a quarter by all site. CCAPS SLC-40 has the smallest share of successful landings with only over a one tenth.

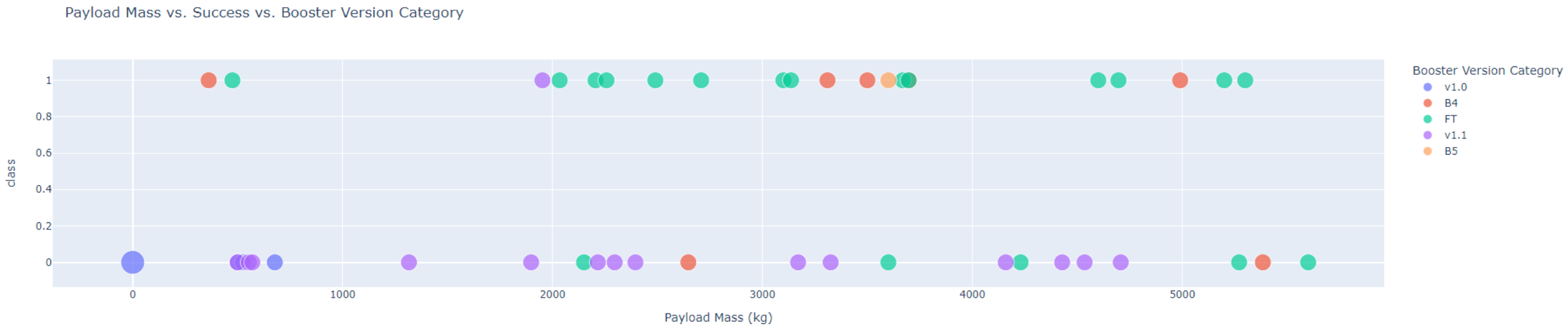


KSC LC-39A

KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings

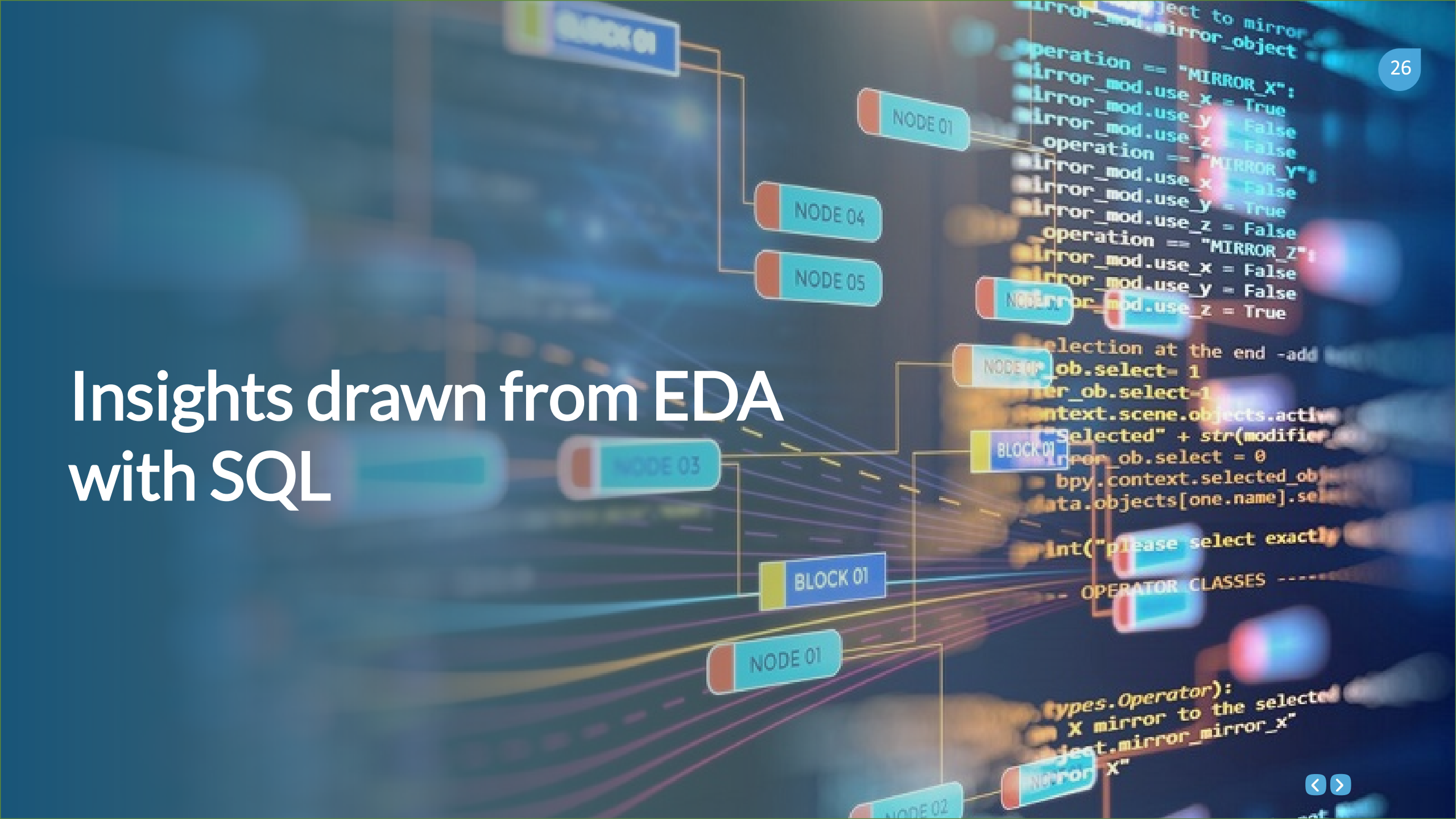


Payload Range vs Orbit Type



Scatter plot for all sites with <6.000 kg. In this range, the plot contain the majority of unsuccess launches.

Insights drawn from EDA with SQL



All Unique Launch Sites

This query outputs unique names in the Launch Site column.

There are 4 unique launch sites by SpaceX.

```
1  /* -------Task 1- Display the names of the unique launch
2  sites in the space mission */
3  SELECT DISTINCT LAUNCH_SITE
4  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DET

Row	LAUNCH_SITE
1	KSC LC-39A
2	CCAFS LC-40
3	VAFB SLC-4E
4	CCAFS SLC-40

Launch Sites Begin with 'CCA'

This query outputs 5 records where launch site begin with ‘CCA’

```
/* -------Task 2- Display 5 records where launch sites
begin with the string 'CCA'  */
 SELECT *
 FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
 WHERE launch_site LIKE 'CCA%'
 LIMIT 5;
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Date	Time__UTC_	Booster_Version	Launch_Site	Payload
1	2013-12-03	22:41:00	F9 v1.1	CCAFS LC-40	SES-8
2	2014-01-06	22:06:00	F9 v1.1	CCAFS LC-40	Thaicom 6
3	2014-08-05	08:00:00	F9 v1.1	CCAFS LC-40	AsiaSat 8
4	2014-09-07	05:00:00	F9 v1.1 B1011	CCAFS LC-40	AsiaSat 6
5	2015-03-02	03:50:00	F9 v1.1 B1014	CCAFS LC-40	ABS-3A Eutelsat 115 West B

Total Payload Mass by NASA

This query displays the sum of payload mass launched by NASA

```
9  /* -------Task 3- Display the total payload mass carried
10 by boosters launched by NASA (CRS) */
11 SELECT SUM(payload_mass__kg_) AS sum
12 FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
13 WHERE customer='NASA (CRS)'
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DET

Row	sum
1	45596

Average Payload Mass F9 v1.1 booster vers.

This query displays the average of payload mass with F9v1.1 booster version

```
13  /* -------Task 4 Display average payload mass carried by booster version F9 v1.1
14  SELECT AVG(payload_mass__kg_) AS average
15  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
16  WHERE booster_version LIKE 'F9 v1.1%'
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GR

Row	average
1	2534.666666666665

First Successful Landing Date

This query displays the date of the first successful landing outcome

```
17  /* -------Task 5- List the date when the first succesful
18  landing outcome in ground pad was achieved.  */
19  SELECT MIN(date) AS date
20  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
21  WHERE mission_outcome= 'Success'
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETA

Row	date
1	2010-06-04

Successful Drone Ship Version with Payload between 4000 and 6000

This query returns the booster versions of successful drone ship with payload between 4000 and 6000

```
/* -------Task 6- List the names of the boosters which have
success in drone ship and have payload mass greater than 4000
but less than 6000 */
SELECT booster_version
FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
WHERE (mission_outcome= 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
AND (landing__outcome= 'Success (drone ship)')
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	booster_version
1	F9 FT B1021.2
2	F9 FT B1031.2
3	F9 FT B1022
4	F9 FT B1026

Total Number of Success and Failure

This query lists the number of successful and failure mission outcomes

```
25  /* -------Task 7- List the total number of successful
26  and failure mission outcomes  */
27  SELECT mission_outcome,
28         count(*) AS count
29  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
30  GROUP BY mission_outcome
31  ORDER BY count DESC
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION

Row	mission_outcome	count
1	Success	98
2	Failure (in flight)	1
3	Success (payload status uncle...	1
4	Success	1

Booster versions with max payload

This query displays booster versions with max payload

```
31  /* -------Task 8- List the names of the booster_versions which have carried the maximum
32  payload mass. Use a subquery */
33  SELECT booster_version
34  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
35  WHERE payload_mass__kg_=
36    (SELECT max(PAYLOAD_MASS__KG_) FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET)
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

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

Failed drone ship in 2015

This query returns information of failed drone ship in 2015

```
36  /* -------Task 9- List the records which will display the month names, failure landing_outcomes
37  in drone ship, booster versions, launch_site for the months in year 2015. */
38  SELECT EXTRACT(month FROM date) as Month,
39         landing__outcome,
40         booster_version,
41         launch_site
42  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
43  WHERE (CAST(date AS string) LIKE '2015%') AND landing__outcome='Failure (drone ship)'
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

Row	Month	landing__outcome	booster_version	launch_site
1	1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2	4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Count of successful landing outcomes

This query counts successful landing outcomes in descending order where date between 04-06-2010 and 20-03-2017

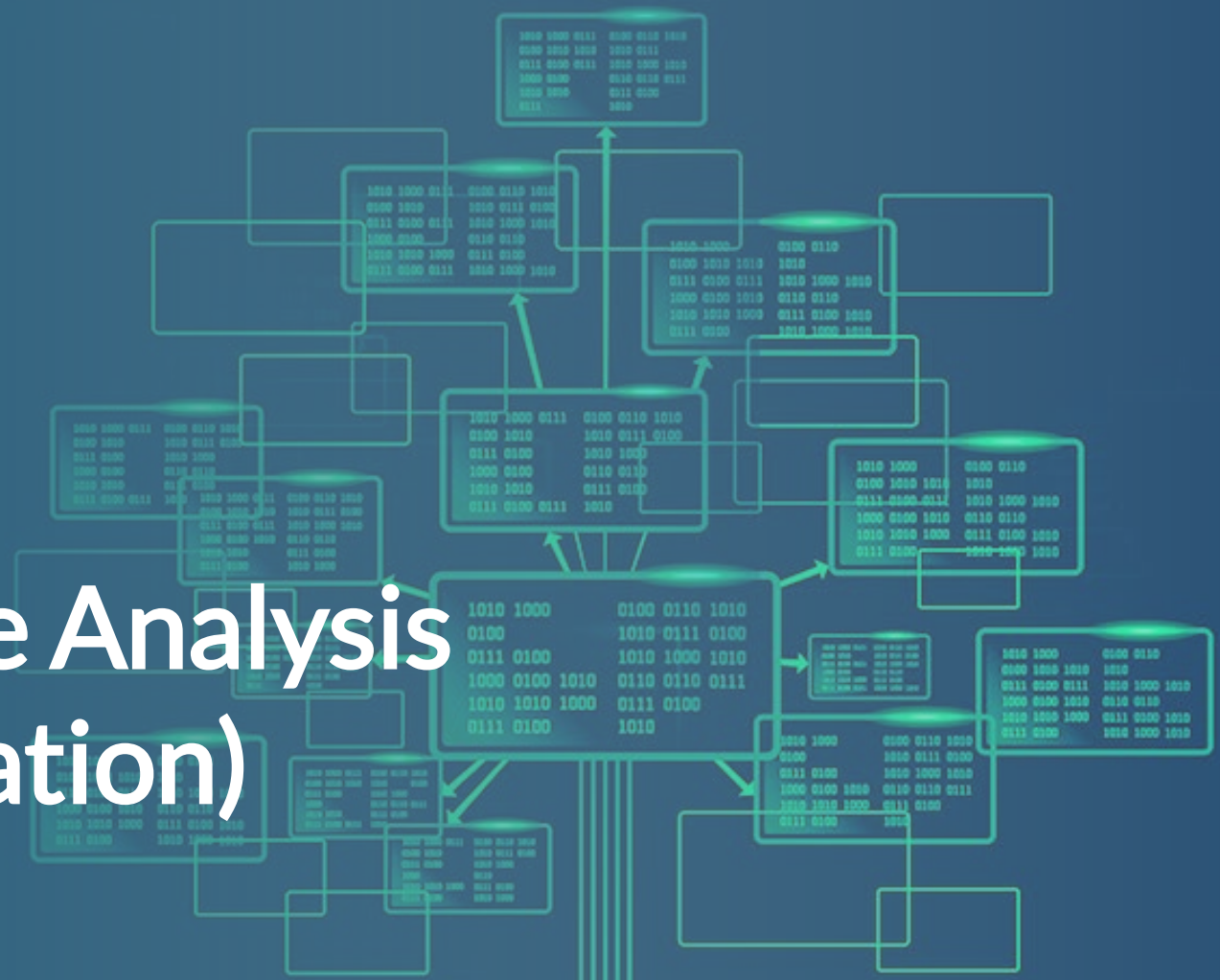
```
43  /* -------Task 10- Rank the count of successful landing_outcomes between the date
44  04-06-2010 and 20-03-2017 in descending order.*/
45  SELECT landing__outcome,
46         COUNT(*) as count
47  FROM processingtimewith-sql.CAPSTONE_01.SPACEXDATASET
48  WHERE (date BETWEEN '2010-06-04' AND '2017-03-20')
49  GROUP BY landing__outcome
50  ORDER BY count DESC
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	landing__outcome	count
1	No attempt	10
2	Failure (drone ship)	5
3	Success (drone ship)	5
4	Success (ground pad)	3
5	Controlled (ocean)	3
6	Failure (parachute)	2
7	Uncontrolled (ocean)	2
8	Precluded (drone ship)	1

Predictive Analysis (Classification)



Classification Accuracy

All models had the same accuracy on the test set with 83.34% accuracy.

It should be noted that the test size only has 18 shapes, which means the accuracy can be misleading. So we may need more data to predict and determine the best-performed methods.

Find the method performs best:

```
accu=[]
methods=[]
accu.append(logreg_cv.score(X_test,Y_test))
methods.append('Logistic Regression')
accu.append(svm_cv.score(X_test,Y_test))
methods.append('SVM')
accu.append(tree_cv.score(X_test,Y_test))
methods.append('Decision Tree Classifier')
accu.append(knn_cv.score(X_test,Y_test))
methods.append('K Nearest Neighbors')
```

```
print(accu)
print(methods)
```

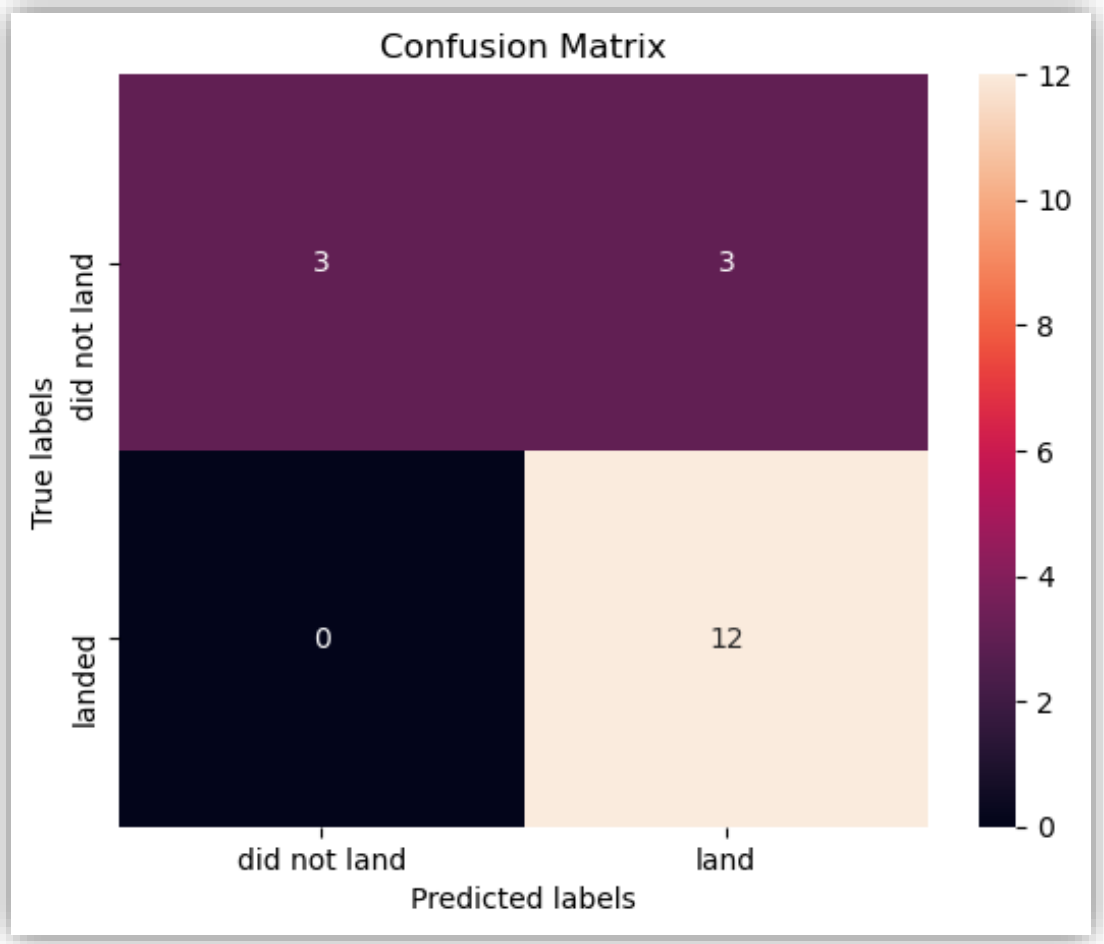
```
[0.8333333333333334, 0.8333333333333334, 0.8333333333333334, 0.8333333333333334]
['Logistic Regression', 'SVM', 'Decision Tree Classifier', 'K Nearest Neighbors']
```

Confusion Matrix

Since all models performed the same accuracy for the test set, the confusion matrix is the same across all models.

We can see from the confusion matrix that:

- **True Positive:** 12 successful landings are classified accurately.
- **False Positive:** 3 unsuccessful landings are classified as landings.
- **True Negative:** 3 unsuccessful landings are classified accurately.
- **False Negative:** No successful landings are classified as unsuccessful landings.



5. Conclusion

Conclusion

- We can conclude that the more Falcon 9 flights, the more successful it is in the last flights.
- Payload Mass, Launch Site, and Orbit Type are key factors that affect the result of successful landings.
- While there are only 18 test samples to build the models, more data is needed to determine which model perform best.

6. Appendix

SpaceX wikipedia page: List of Falcon 9 and Falcon Heavy launches

Github repository

SQL Queries in Google BigQuery

Thank you!