IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

CRISP-DM (Cross-Industry Standard Process for Data Mining): This is a widely used methodology that consists of six major phases:

- Business Understanding: Define the problem, the business goals, and the objectives you want to achieve using data analysis.
- Data Understanding: Collect and explore the data, understand its structure, quality, and completeness.
- Data Preparation: Clean, preprocess, and transform the data into a suitable format for analysis.
- Modeling: Choose and apply appropriate data analysis techniques, such as machine learning algorithms, to build predictive or descriptive models.
- Evaluation: Assess the performance and effectiveness of the models using various metrics and validation techniques.
- Deployment: Integrate the models into the business process and make them accessible to stakeholders.

The results of this project contain final probability of spaceX willreuse their stage 1 rocket and the price of their single rocket launch

Introduction

This project is proposed to help the Space Y in making such new rocket launches competing the company's competitor SpaceX owned by Allon Musk. SpaceX is the most successful rocket launcher company by current time. So it is so useful for us to study from them about rocket launch to make ours as good as theirs.

This project will help readers about the price of a single rocket launch and the probability of weather SpaceX will reuse their stage 1 rocket or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

We use GET request and web scraping for data collection

- Perform data wrangling

We remove NaN value and replace them with the average value of the column. then we only take the valuable variable and make dataframe of them.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

we use KNN, Decision tree, SVM, and Logistic regression as the predictive model

Data Collection – SpaceX API

first we use GEt request and normalize the data with ‘json_normalize’ function. the return is data frame with value of dictionaries. to get the value of variable we want, we have to use GET request again to the considered link and dont forget to add ‘rocket’ value to the url.

github link:

https://github.com/arsya-ux/Capstone-DS/blob/9597a9ecb7667b2a45bdea1004ba71b3d361b380/SpaceX%20API.ipynb

Data Collection - Scraping

- Web scraping is used to data collection, first we use GET request and parse the data with BeatifulSoup function. then we get the first table by ‘find_all’ function which return with the name of columns. to get the table value, we call the second table and make the dataframe.

- Github URL:

https://github.com/arsya-ux/Capstone-DS/blob/9597a9ecb7667b2a45bdea1004ba71b3d361b380/SpaceX%20Scraping.ipynb

Data Wrangling

form dataset_part_1.csv file, we make new column called ‘Class’ represent the success or not the rocket launch. this column was made by the launch result column from the dataset file

- The Data Wrangling notebook can be accessed with this link:

https://github.com/arsya-ux/Capstone-DS/blob/9597a9ecb7667b2a45bdea1004ba71b3d361b380/Data%20Wrangling.ipynb

EDA with Data Visualization

- In this stage, we will visualize many of variable and see the correlation of each variable. such as launchsite vs payload, launchsite vs flight number and much more.

github url:

https://github.com/arsya-ux/Capstone-DS/blob/9597a9ecb7667b2a45bdea1004ba71b3d361b380/EDA%20Data%20Visualization.ipynb

EDA with SQL

- This stage queried dataframe which is used to gain information such as first successful rocket launch, maximum payload, and rocket boosters type.

- github url:

https://github.com/arsya-ux/Capstone-DS/blob/9597a9ecb7667b2a45bdea1004ba71b3d361b380/EDA%20SQL.ipynb

Predictive Analysis (Classification)

- We use KNN, SVM, Decision tree, and logistic regression. we test our model by using ‘score’ function to see its accuracy and rebuilt the model with different parameters until we get the highest accuracy

github url:

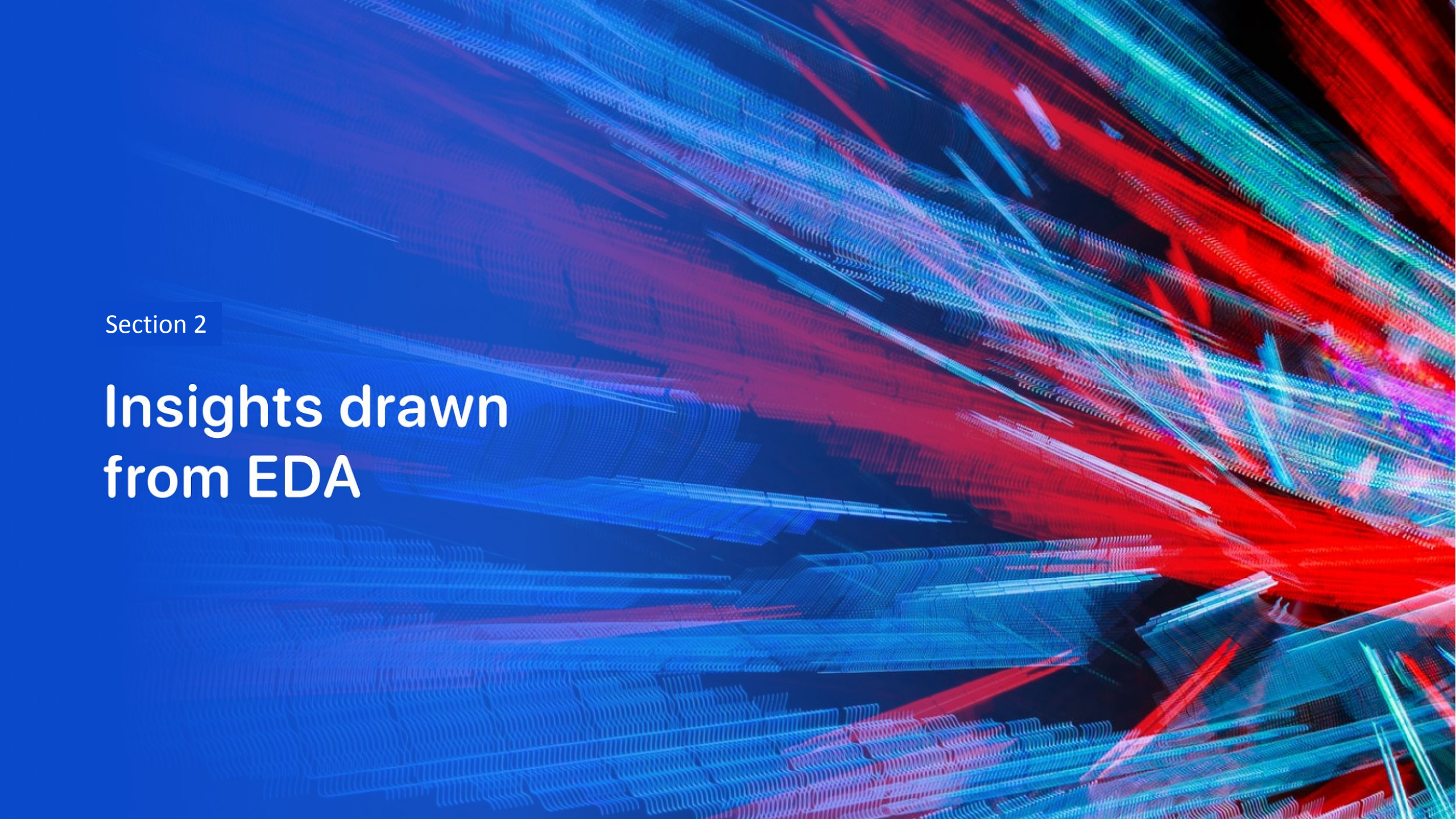
https://github.com/arsya-ux/Capstone-DS/blob/9597a9ecb7667b2a45bdea1004ba71b3d361b380/Predictive%20Analysis.ipynb

Results

- The results of these stages are spacex data frame which contain important variable like class, serial, launchsite, flight number and much more
- data visualization of each variable to see the correlation of variables
- predictive analysis using KNN, decision tree, Logistic regression, and SVM to predict the probability of spaceX reuse their rocket boosters.
- queried dataframe which is used to gain information such as first successful rocket launch, maximum payload, and rocket boosters type.

Section 2

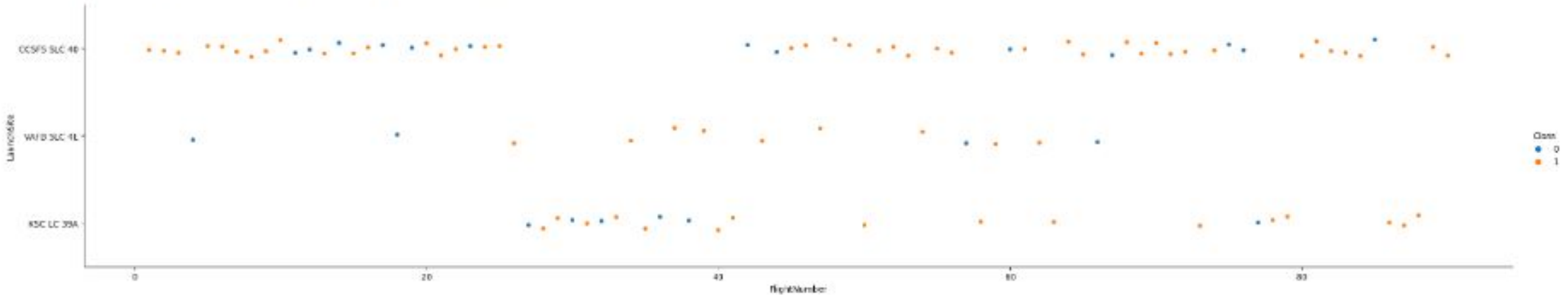
Insights drawn from EDA



Flight Number vs. Launch Site

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
```

```
]: <seaborn.axisgrid.FacetGrid at 0x60c9b10>
```

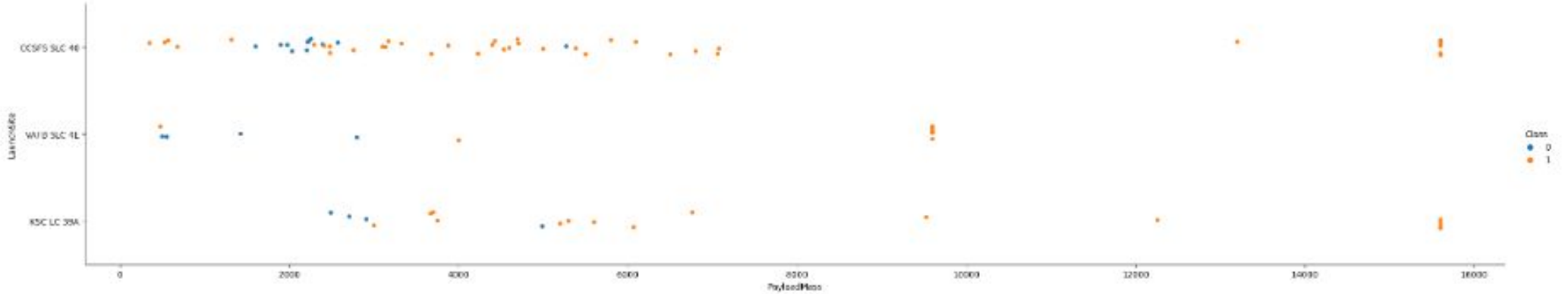


CCSFS SLC-40 is used to do more launches than any other launchsite

Payload vs. Launch Site

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
```

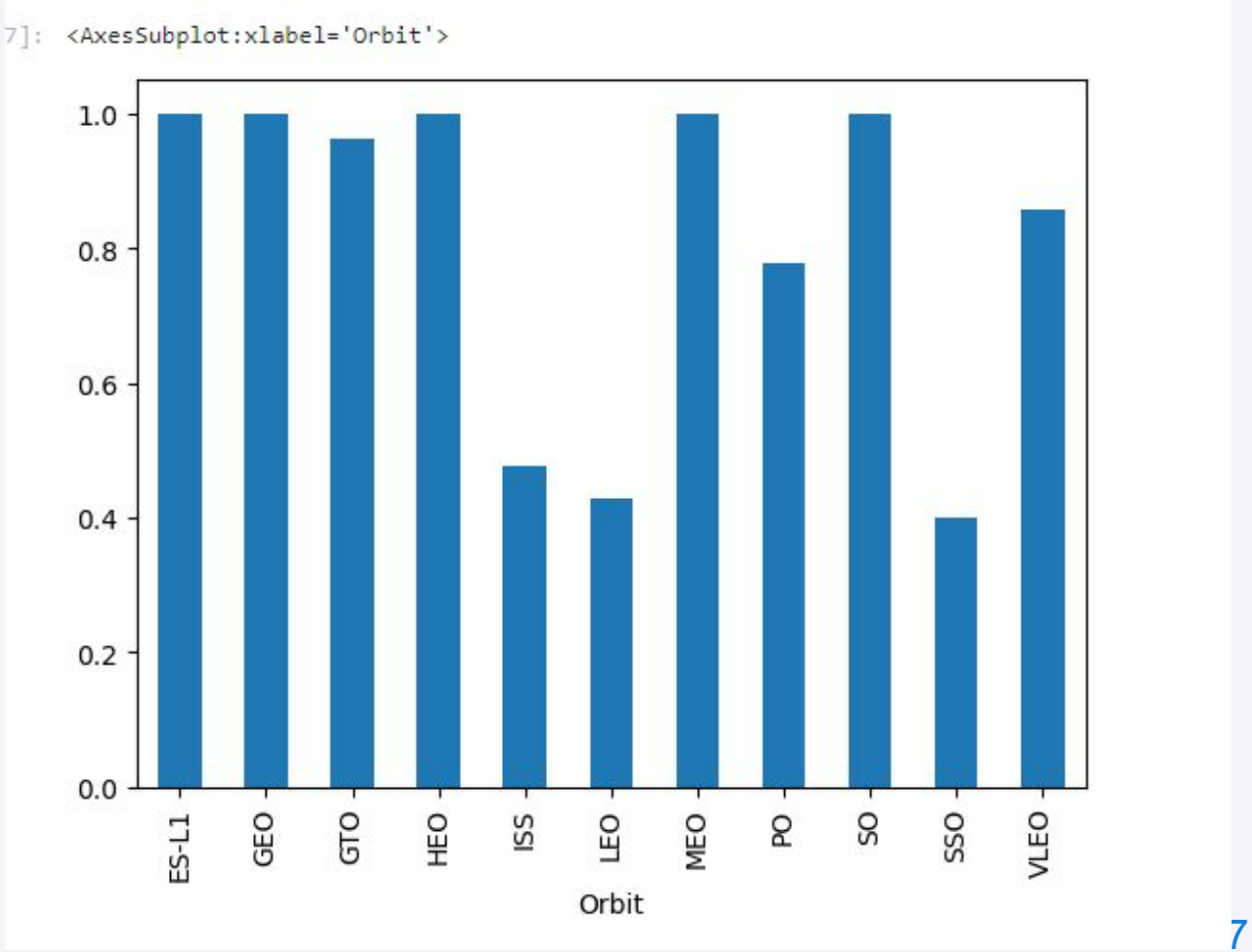
```
[15]: <seaborn.axisgrid.FacetGrid at 0x70ff510>
```



CCSFS SLC-40 have the highest number of total launch that carried the maximum payload mass

Success Rate vs. Orbit Type

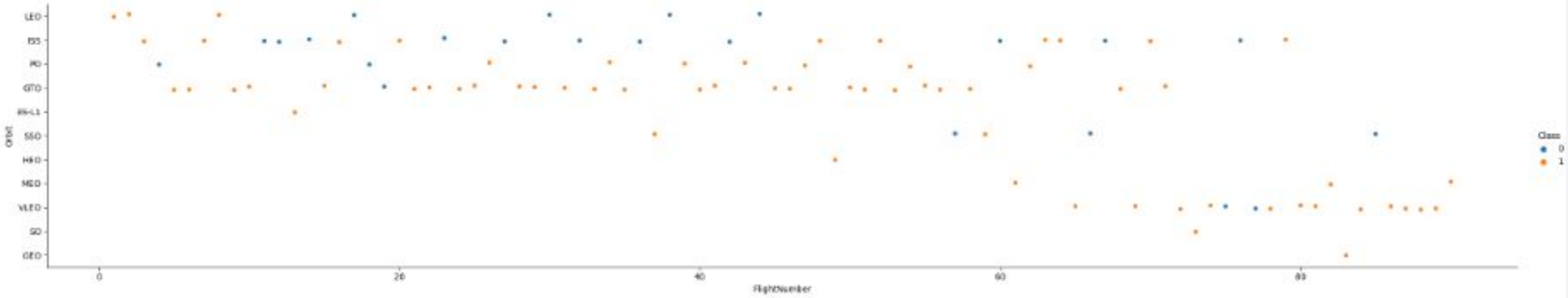
- There are multiple of orbit type that has the highest success rate (100%)



Flight Number vs. Orbit Type

```
[16]: # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
      sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)

[18]: <seaborn.axisgrid.FacetGrid at 0x7248ff8>
```



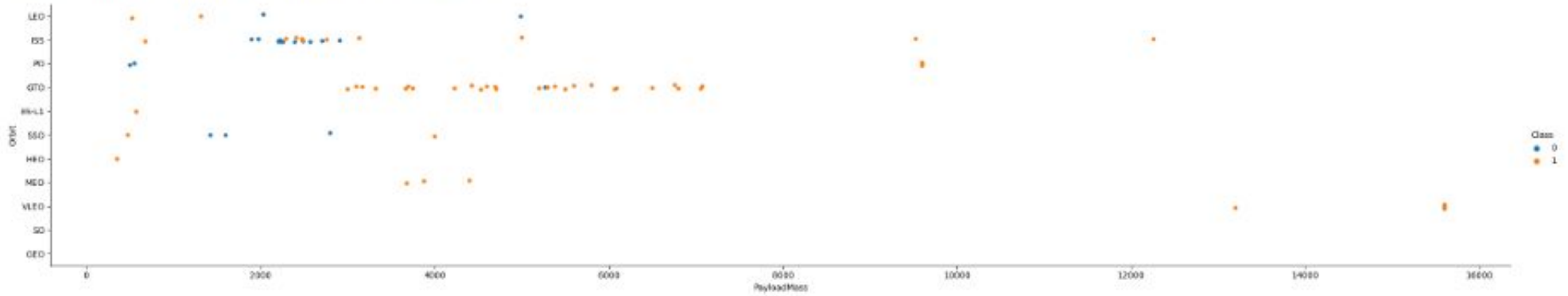
The highest number of orbit target of all launches are from GTO to LEO

Payload vs. Orbit Type

Similarly, we can plot the Payload vs. Orbit scatter point charts to reveal the relationship between Payload and Orbit type

```
[20]: # Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
      sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
```

```
[20]: <seaborn.axisgrid.FacetGrid at 0x5986520>
```



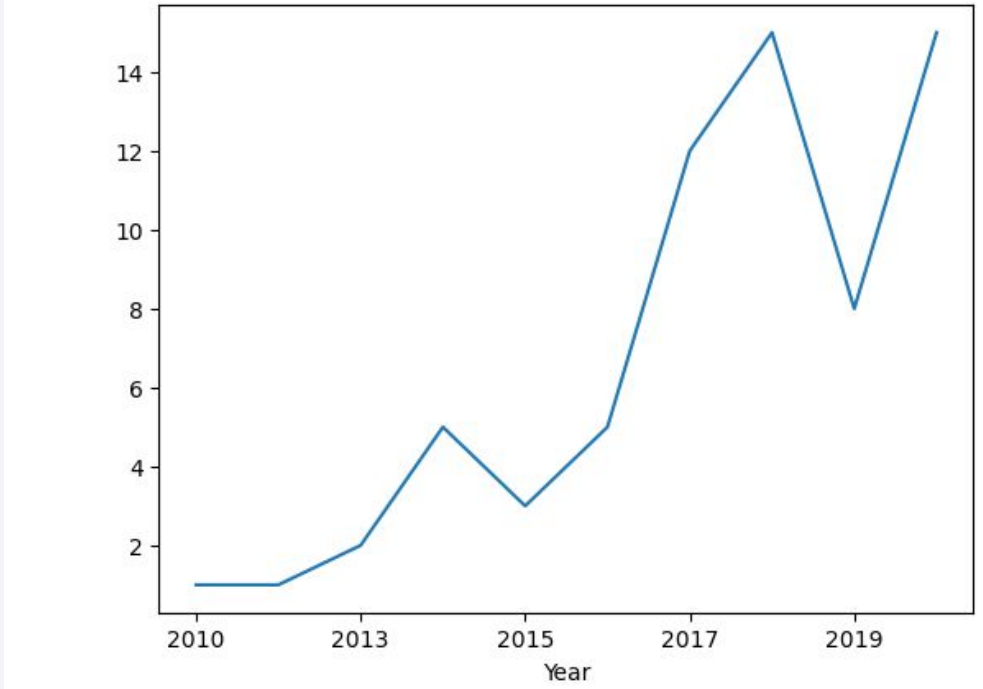
- The maximum payload carried is to VLEO Orbit

Launch Success Yearly Trend

From 2013, the success rate kee rising until 2020

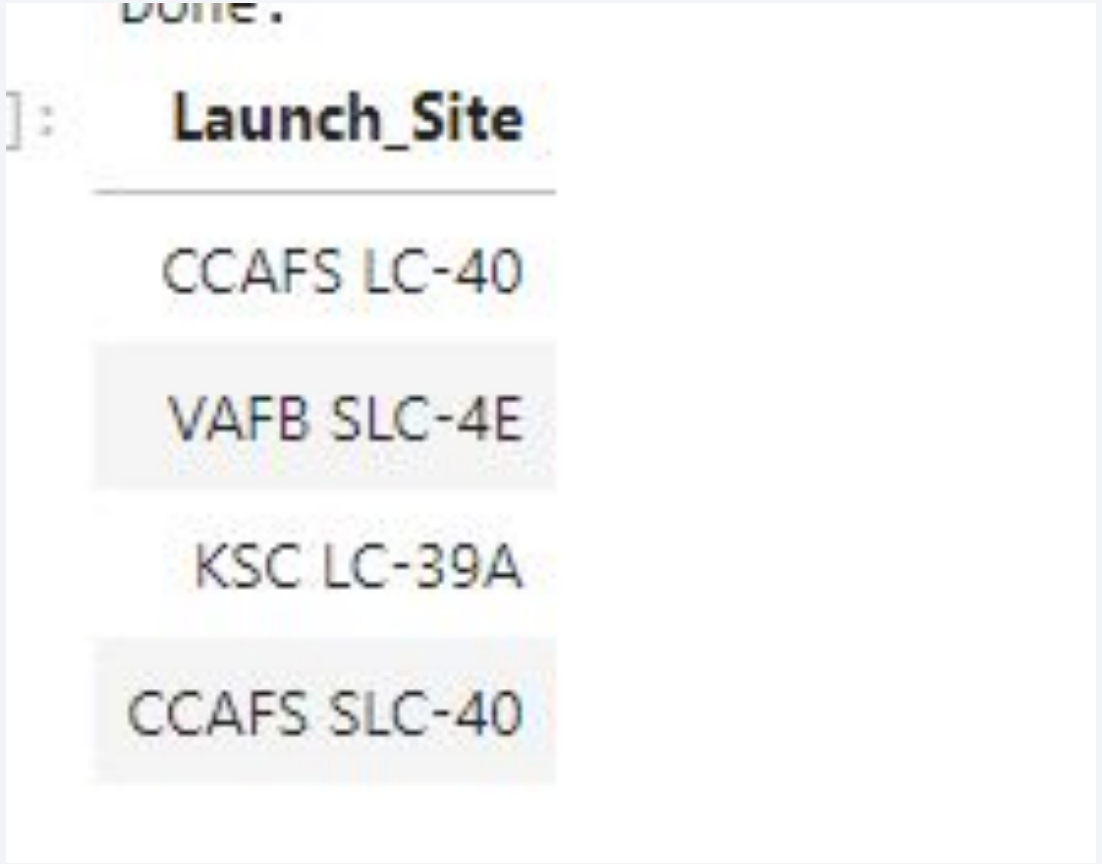
```
c=df.groupby('Year')['Class'].sum()
c.plot()
```

[32]: <AxesSubplot:xlabel='Year'>



All Launch Site Names

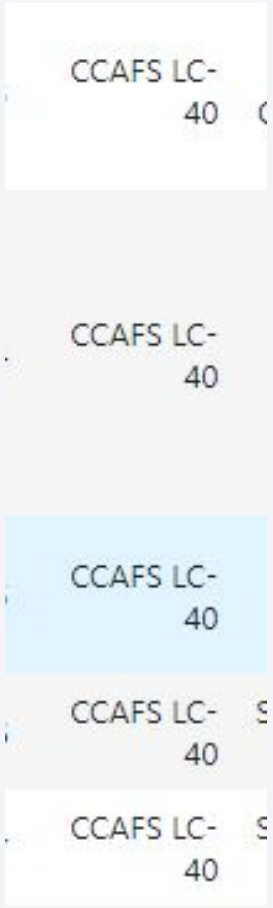
There is only 4 Launch Site name



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

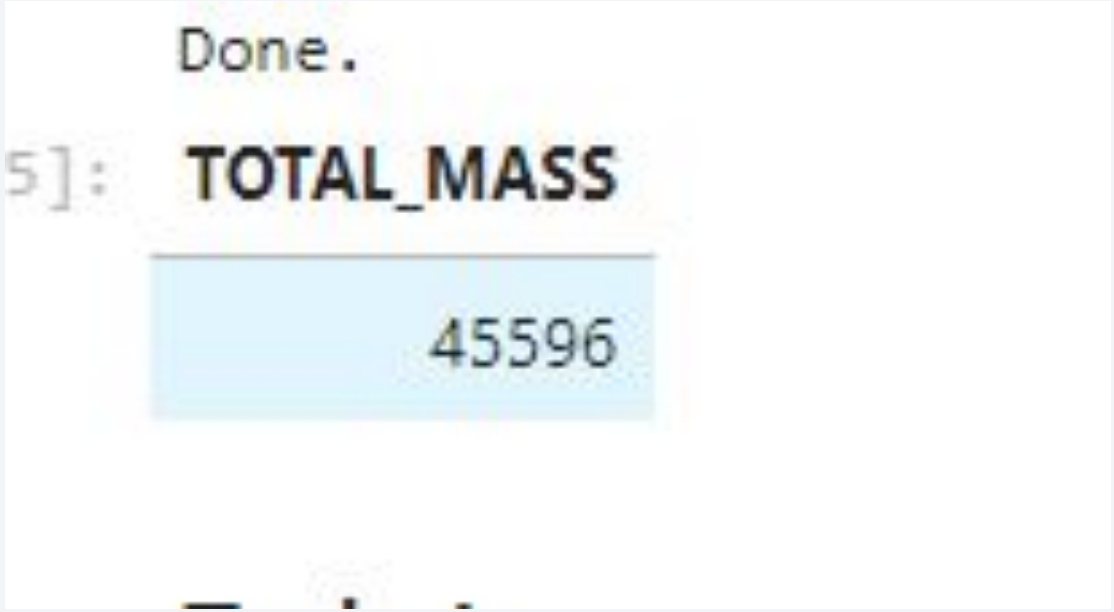
Launch Site Names Begin with 'CCA'

All of the Launch Site that start with ‘CCA’ is CCAFS LC-40



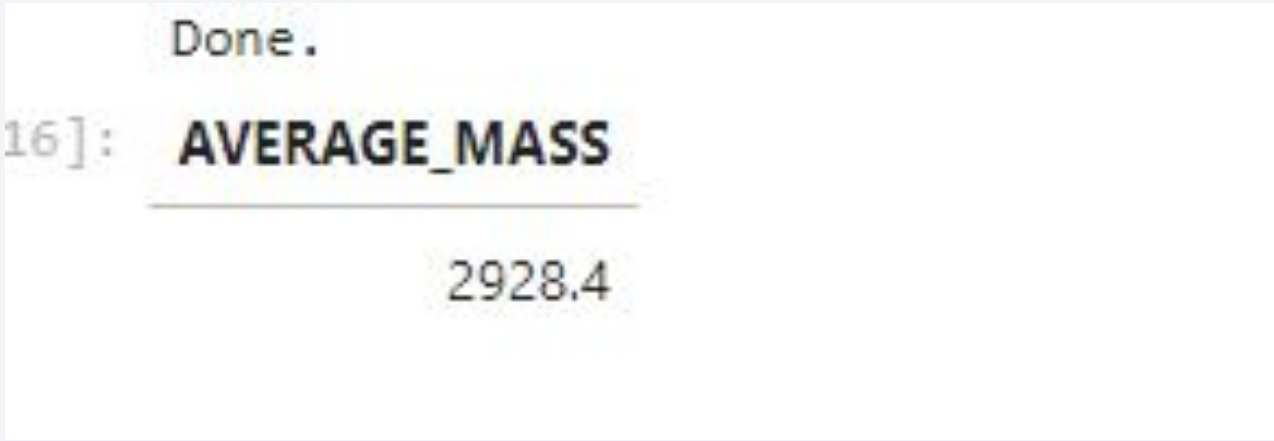
Total Payload Mass

There is total of 45.6 tons of mass that carried by SpaceX rockets



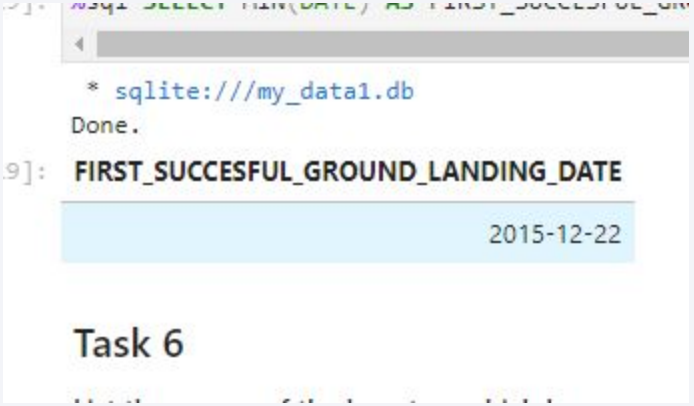
Average Payload Mass by F9 v1.1

The average mass for each F9 v1.1 is 3 tons



First Successful Ground Landing Date

The date is 22 December 2015



Successful Drone Ship Landing with Payload between 4000 and 6000

Here is the list of booster that landed successfully with payload between

4000 and 6000 kg

[21]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

There is almost 98% in successful landing

Done.

[32]:

COUNT(CASE WHEN Mission_Outcome = 'Success' THEN 1 ELSE NULL END)	COUNT(CASE WHEN Mission_Outcome LIKE 'Failure%%' THEN 1 ELSE NULL END)
98	1

Boosters Carried Maximum Payload

There is multiple of boosters type that carried the maximum amount of payload

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

[63]:

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

2015 Launch Records

There is only two launch which all of that fail

[75]:

YEAR	MONTH	Landing_Outcome	Booster_Version	Launch_Site
2015	5	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	5	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

The highest rank is no attemp and the lowest is failure(parachute)

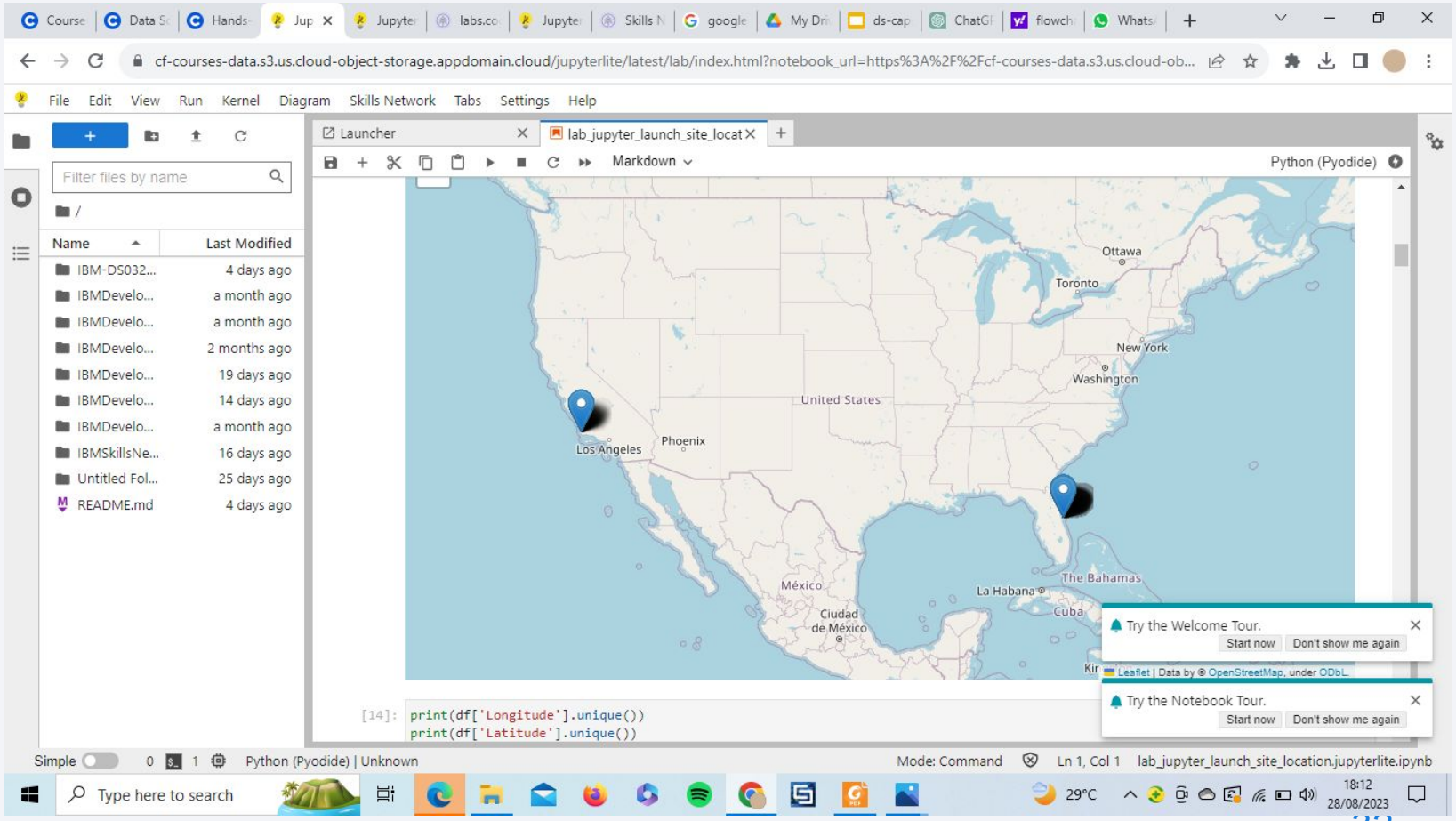
[83]:

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


Section 3
Launch Sites
Proximities Analysis

Launch site location all over the globe

The launch site is in florida and california

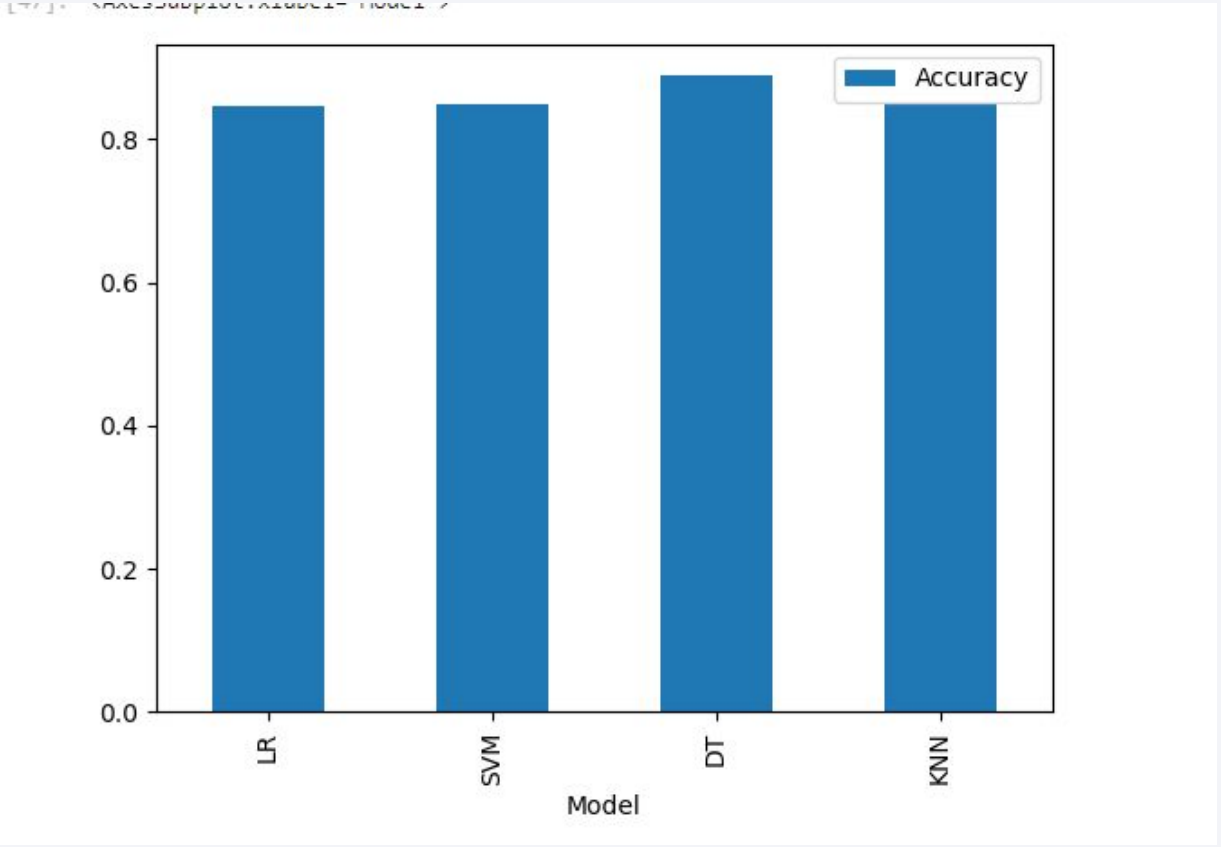


Section 5

Predictive Analysis (Classification)

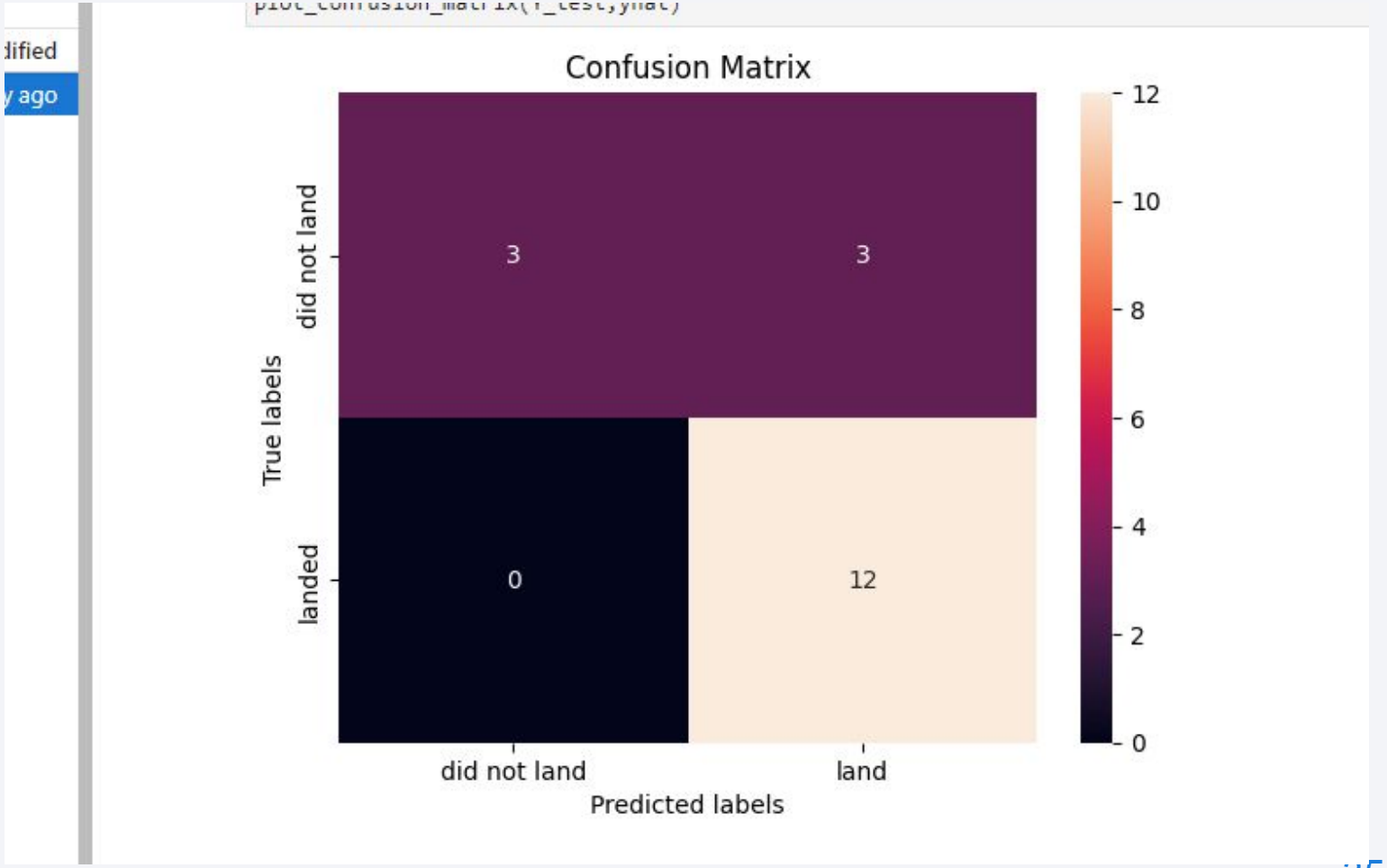
Classification Accuracy

KNN and decision tree has the highest accuracy with 0.8482 accuracy value



Confusion Matrix

There are such errors on land predictions. there are 3 value that supposed to be ‘did not land’ but our prediction predicted it as landed



Conclusions

SpaceX has multiple of boosters launchers that all of them has high success launch value. the cost of spaceX launcher are only in range of $60M in price while other space exploration company’s launchers price is almost touching $160M. Most of spaceX launcher have the high probability of reusable rocket and success rate of launch

Appendix

dataset_part_1.csv
dataset_part_2.csv
dataset_part_3.csv

Thank you!