

labs.cognitiveclass.ai

File Edit View Run Kernel Tabs Settings Help

+ G

Filter files by name

/

Name Modified

M README.md 3mo ago

Launcher Eda Data Visualization.ipynb edadataviz.ipynb +

Python (Pyodide) ⚙

## Objectives

Perform exploratory Data Analysis and Feature Engineering using `Pandas` and `Matplotlib`

- Exploratory Data Analysis
- Preparing Data Feature Engineering

## Import Libraries and Define Auxiliary Functions

We will import the following libraries the lab

```
[1]: import pipelite
await pipelite.install(['numpy'])
await pipelite.install(['pandas'])
await pipelite.install(['seaborn'])

[2]: # pandas is a software library written for the Python programming language for data manipulation and analysis.
import pandas as pd
#NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices.
import numpy as np
# Matplotlib is a plotting library for python and pyplot gives us a MatLab like plotting framework. We will use this to visualize our data.
import matplotlib.pyplot as plt
#Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
import seaborn as sns
```

<ipython-input-2-cde6ab162d36>:2: DeprecationWarning:  
Pyarrow will become a required dependency of pandas in the next major release of pandas (pandas 3.0),  
(to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)  
but was not found to be installed on your system.  
If this would cause problems for you,  
please provide us feedback at <https://github.com/pandas-dev/pandas/issues/54466>

```
import pandas as pd
```

## Exploratory Data Analysis

First, let's read the SpaceX dataset into a Pandas dataframe and print its summary

```
[3]: from js import fetch
import json
```

5:55 PM Thu 14 Nov

41%

Applied Data Science Cap X Exercise: Preparing Your X IBM Machine Learning Pr X Skills Network Labs X can i right click in an ipad X +

File Edit View Run Kernel Tabs Settings Help

+ Filter files by name

Name Modified

M README.md 3mo ago

Launcher Eda Data Visualization.ipynb edadataviz.ipynb

Python (Pyodide)

# Exploratory Data Analysis

First, let's read the SpaceX dataset into a Pandas dataframe and print its summary

```
[3]: from js import fetch
import io

URL = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/"
resp = await fetch(URL)
dataset_part_2_csv = io.BytesIO(await resp.arrayBuffer()).to_py()
df=pd.read_csv(dataset_part_2_csv)
df.head(5)
```

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

First, let's try to see how the `FlightNumber` (indicating the continuous launch attempts.) and `Payload` variables would affect the launch outcome.

We can plot out the `FlightNumber` vs. `PayloadMass` and overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass also appears to be a factor; even with more massive payloads, the first stage often returns successfully.

```
[4]: sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Pay load Mass (kg)", fontsize=20)
```

labs.cognitiveclass.ai

File Edit View Run Kernel Tabs Settings Help

+ Filter files by name

Name Modified

M README.md 3mo ago

[4]:

```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Pay load Mass (kg)", fontsize=20)
plt.show()
```

Next, let's drill down to each site visualize its detailed launch records.

### TASK 1: Visualize the relationship between Flight Number and Launch Site

Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber`, set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

[5]:

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be t
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```

Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

### TASK 2: Visualize the relationship between Payload Mass and Launch Site

labs.cognitiveclass.ai

File Edit View Run Kernel Tabs Settings Help

+ G

Launcher Eda Data Visualization.ipynb edadataviz.ipynb +

Python (Pyodide)

## TASK 2: Visualize the relationship between Payload Mass and Launch Site

We also want to observe if there is any relationship between launch sites and their payload mass.

```
[6]: # Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass(kg)", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```

Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

## TASK 3: Visualize the relationship between success rate of each orbit type

Next, we want to visually check if there are any relationship between success rate and orbit type.

Let's create a bar chart for the success rate of each orbit

```
[24]: # HINT use groupby method on Orbit column and get the mean of Class column
SR_orbit = df.groupby(['Orbit']).mean('Class').reset_index()

sns.barplot(x='Orbit', y='Class', data=SR_orbit)
plt.show()
```



File Edit View Run Kernel Tabs Settings Help



Filter files by name

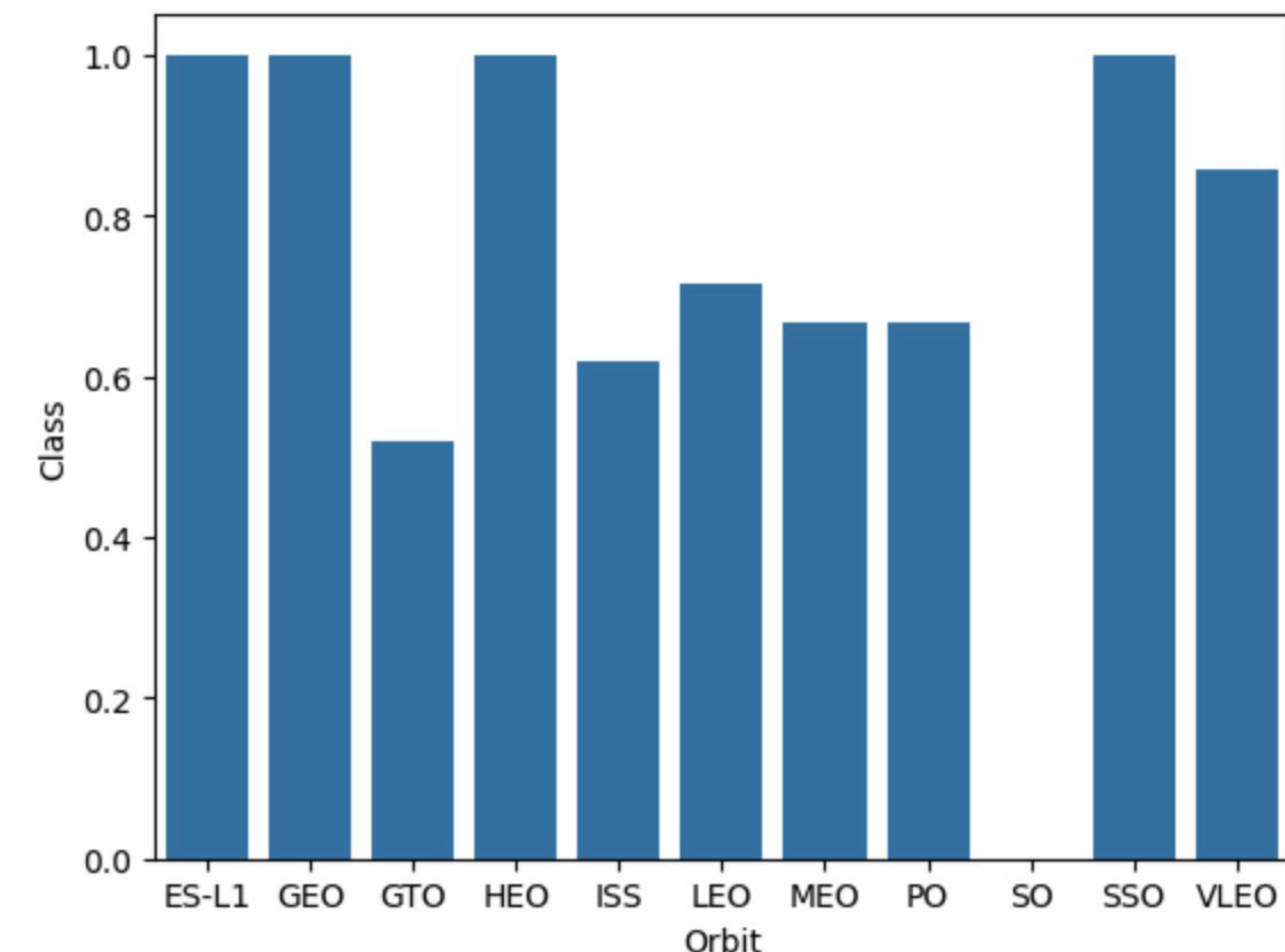


Name / Modified

M README.md 3mo ago

Launcher Eda Data Visualization.ipynb edadataviz.ipynb

Python (Pyodide)



Analyze the plotted bar chart to identify which orbits have the highest success rates.

#### TASK 4: Visualize the relationship between FlightNumber and Orbit type

For each orbit, we want to see if there is any relationship between FlightNumber and Orbit type.

```
[25]: # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class

sns.catplot(x='FlightNumber', y='Orbit', hue='Class', data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```





File Edit View Run Kernel Tabs Settings Help

+

Filter files by name



/

Name Modified

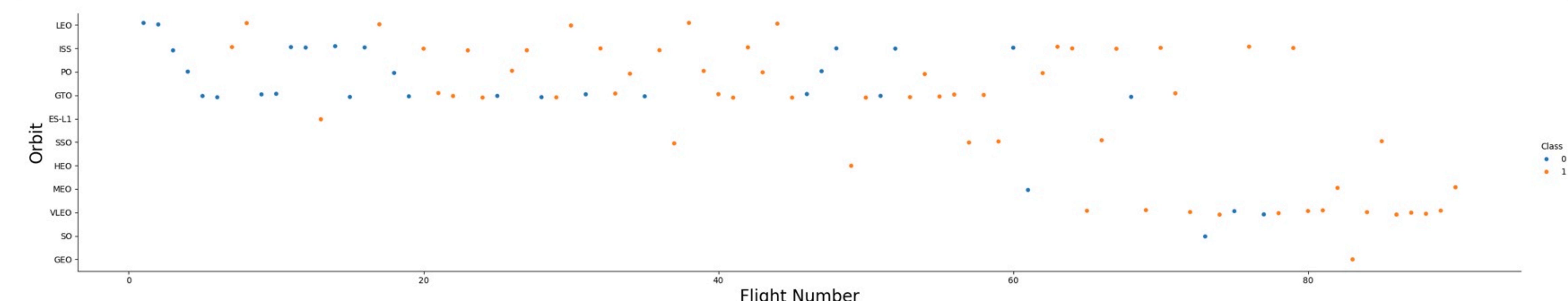
M README.md 3mo ago

Launcher Eda Data Visualization.ipynb edadataviz.ipynb +

Python (Pyodide)

[25]: *# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the clas*

```
sns.catplot(x='FlightNumber', y='Orbit', hue='Class', data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```



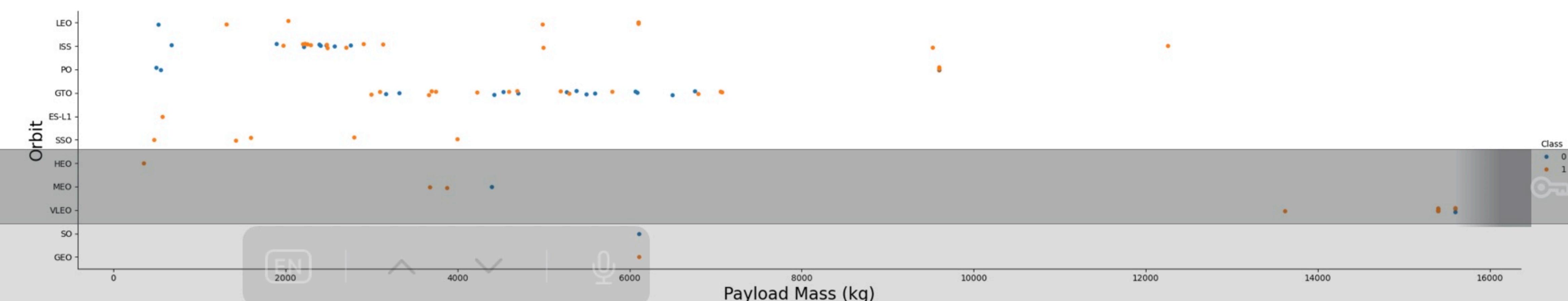
You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

## TASK 5: Visualize the relationship between Payload Mass and Orbit type

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

[26]: *# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, and hue to be the clas*

```
sns.catplot(x='PayloadMass', y='Orbit', hue='Class', aspect=5, data=df)
plt.xlabel("Payload Mass (kg)", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```





File Edit View Run Kernel Tabs Settings Help



Filter files by name



/

Name

Modified

M README.md

3mo ago

Launcher Eda Data Visualization.ipynb edadataviz.ipynb

Python (Pyodide)



## TASK 6: Visualize the launch success yearly trend

You can plot a line chart with x axis to be Year and y axis to be average success rate, to get the average launch success trend.

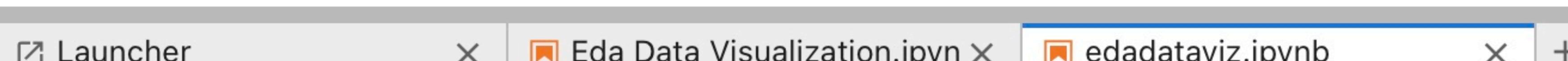
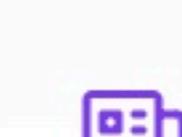
The function will help you get the year from the date:

```
[27]: # A function to Extract years from the date
year=[]
def Extract_year():
    for i in df["Date"]:
        year.append(i.split("-")[0])
    return year
Extract_year()
df['Date'] = year
df.head()
```

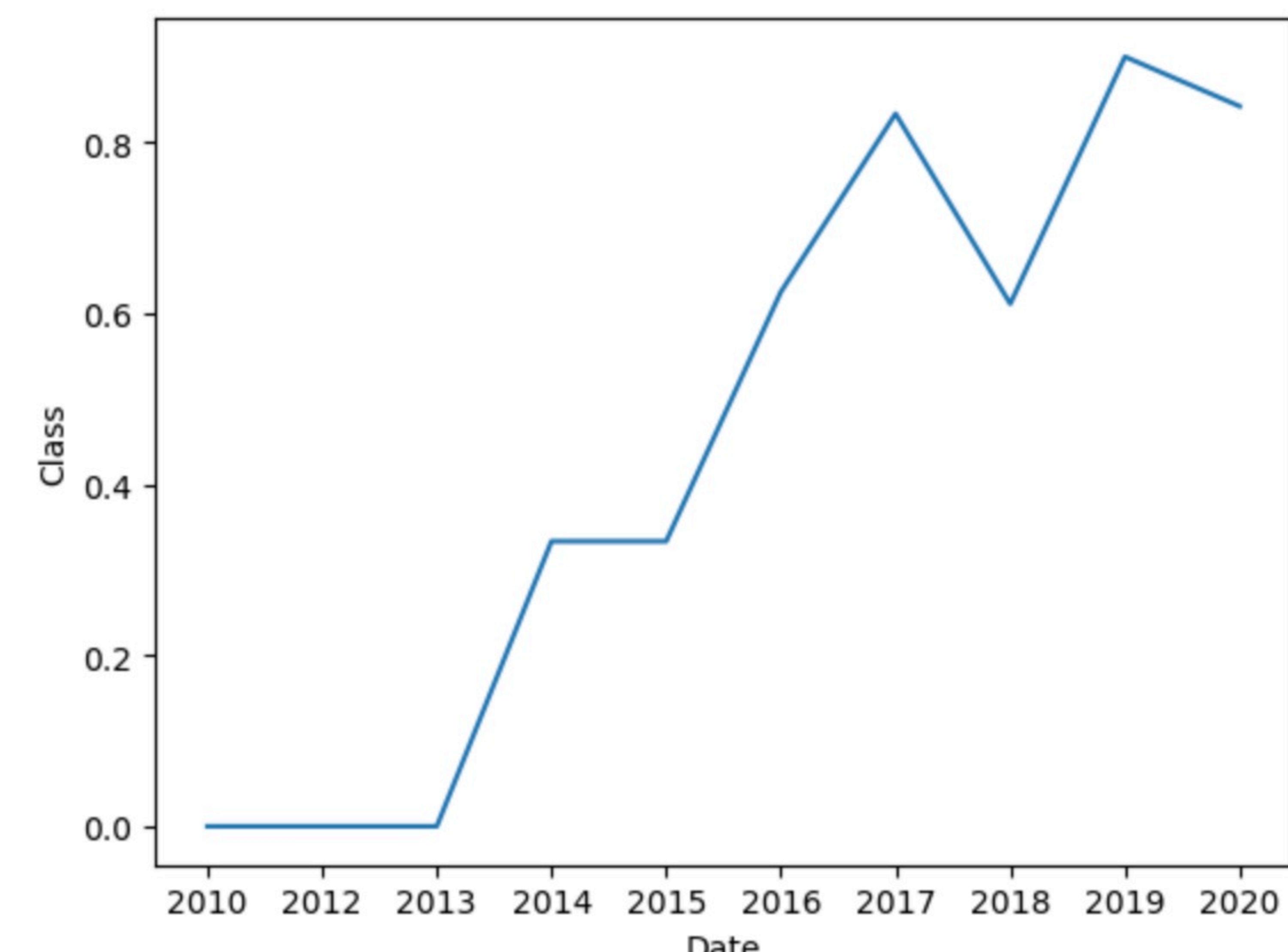
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	Landing
0	1	2010	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	False
1	2	2012	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	False
2	3	2013	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	False
3	4	2013	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	False
4	5	2013	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	False

```
[39]: # Plot a line chart with x axis to be the extracted year and y axis to be the success rate
```

```
SR_year=df.groupby(['Date']).mean(['Class']).reset_index()
sns.lineplot(data=SR_year,x='Date', y='Class')
plt.show()
```



```
sns.lineplot(data=SR_year, x='Date', y='Class')
plt.show()
```



you can observe that the sucess rate since 2013 kept increasing till 2020

# Features Engineering

By now, you should obtain some preliminary insights about how each important variable would affect the success rate, we will select the features that will be used in success prediction in the future module.

```
[40]: features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs',  
features.head()
```

File Edit View Run Kernel Tabs Settings Help

+ + ↑ C

Filter files by name

/

Name Modified

README.md 3mo ago

Launcher Eda Data Visualization.ipynb edadataviz.ipynb +

Python (Pyodide) ○

```
[40]: features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial']]
features.head()
```

	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

## TASK 7: Create dummy variables to categorical columns

Use the function `get_dummies` and `features` dataframe to apply OneHotEncoder to the column `Orbits`, `LaunchSite`, `LandingPad`, and `Serial`. Assign the value to the variable `features_one_hot`, display the results using the method `head`. Your result dataframe must include all features including the encoded ones.

```
[49]: # HINT: Use get_dummies() function on the categorical columns
features_one_hot=pd.get_dummies(features[['Orbit','LaunchSite', 'LandingPad', 'Serial']])
features_one_hot.head()
```

	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	Orbit_LEO	Orbit_MEO	Orbit_PO	Orbit_SO	Orbit_SSO	...	Serial
0	False	False	False	False	False	False	True	False	False	False	...	...
1	False	False	False	False	False	False	True	False	False	False	...	...
2	False	False	False	False	False	True	False	False	False	False	...	...

5 rows x 72 columns

**TASK 8:** Cast all numeric columns to `float64`

Now that our `features_one_hot` dataframe only contains numbers, cast the entire dataframe to variable type `float64`

```
[51]: # HINT: use astype function  
features_one_hot.astype(float)
```