BHARATIYA VIDYA BHAVAN'S

SARDAR PATEL INSTITUTE OF TECHNOLOGY

MUNSHI NAGAR, ANDHERI (WEST), MUMBAI – 400 058

Date:15/09/22

Computer Engineering Minors
Machine Learning : Lab 1

Name: Divija Pankaj Shringarpure
Roll No.:56
Class:BE ETRX
Batch: C

Objective : Import the dataset and perform EDA such as number of data samples, number of features,number of classes, number of data samples per class, removing missing values, conversion to numbers, explore dimensionality, type the mean or average value, and using seaborn library to plot different graphs.

.

Software used : Google Colab Notebook

Dataset considered : "NASA Patents"
Available on : https://data.nasa.gov/Raw-Data/NASA-Patents/gquh-watm
This dataset shows information pertaining to NASA held and pending patents giving us details about the title of the patent , its status , the center name, patent number , patent expiration date, application number , case number.

Code Snippets and Output :

```
[3] #After downloading the dataset from NASA's website I uploaded it on the mounted drive and pasted the location
    import pandas as pd
    import numpy as np
    df=pd.read_csv("/content/drive/MyDrive/NASA_Patents.csv")
    df.head()
```

	Center	Status	Case Number	Patent Number	Application SN	Title	Patent Expiration Date
0	NASA Kennedy Space Center	Application	KSC-12871	0	13/033,085	Polyimide Wire Insulation Repair System	NaN
1	NASA Ames Research Center	Issued	ARC-14048-1	5694939	08/543,093	Autogenic-Feedback Training Exercise Method & ...	10/03/2015
2	NASA Ames Research Center	Issued	ARC-14231-1	6109270	09/017,519	Multimodality Instrument For Tissue Characteri...	02/04/2017
3	NASA Ames Research Center	Issued	ARC-14231-2DIV	6976013	10/874,003	Metrics For Body Sensing System	06/16/2024
4	NASA Ames Research Center	Issued	ARC-14231-3	6718196	09/652,299	Multimodality Instrument For Tissue Characteri...	02/04/2017

```
[6] #Finding total number of null entities in each column
    df.isnull().sum()
```

```
Center                    0
Status                    0
Case Number               0
Patent Number           274
Application SN            7
Title                     0
Patent Expiration Date  350
dtype: int64
```

```
[11] #Dropping all entries with patent number 0
     df = df.dropna(subset=['Patent Number'])
```

```
[12] len(df)

     1119
```

```
#Checking null count
df.isnull().sum()

Center                    0
Status                    0
Case Number               0
Patent Number             0
Application SN            0
Title                     0
Patent Expiration Date    220
dtype: int64
```

```
#Extracting only issued patents
df_issue
```

	Center	Status	Case Number	Patent Number	Application SN	Title	Patent Expiration Date
1	NASA Ames Research Center	Issued	ARC-14048-1	5694939	08/543,093	Autogenic-Feedback Training Exercise Method & ...	10/03/2015
2	NASA Ames Research Center	Issued	ARC-14231-1	6109270	09/017,519	Multimodality Instrument For Tissue Characteri...	02/04/2017
3	NASA Ames Research Center	Issued	ARC-14231-2DIV	6976013	10/874,003	Metrics For Body Sensing System	06/16/2024
4	NASA Ames Research Center	Issued	ARC-14231-3	6718196	09/652,299	Multimodality Instrument For Tissue Characteri...	02/04/2017
5	NASA Ames Research Center	Issued	ARC-14275-1	6445390	09/226,673	Automated Triangle Geometry Processing For Sur...	12/24/2018

```
[32] df_corr.columns

     Index(['Center', 'Patent Number', 'Patent Expiration Date'], dtype='object')
```

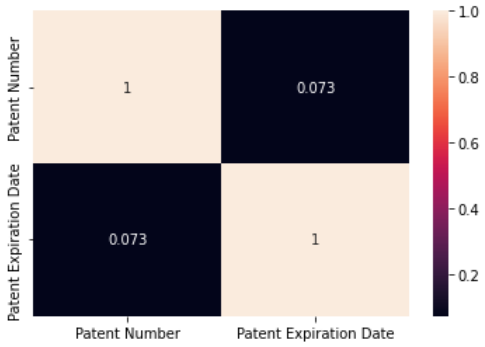
```
[33] df_corr.dtypes

     Center                   object
     Patent Number            object
     Patent Expiration Date   object
     dtype: object
```

```
correlation = df_corr.corr()
correlation
#Here we infer that the patent number goes on increasing as the date increases(goes ahead). There's a positive correlation
```

	Patent Number	Patent Expiration Date
Patent Number	1.000000	0.073074
Patent Expiration Date	0.073074	1.000000

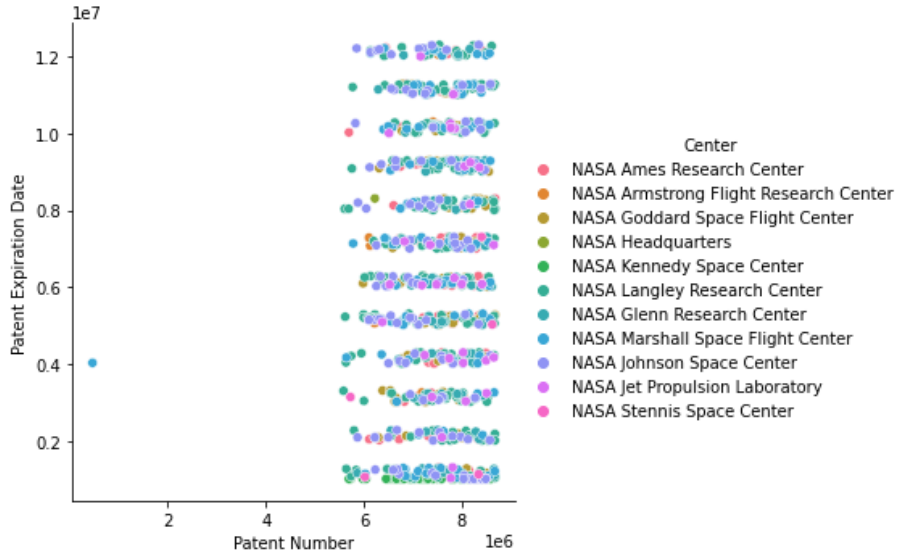
```
#Getting a clarity of the correlation using heatmap
import seaborn as sns
import matplotlib.pyplot as plt
sns.heatmap(correlation,xticklabels=correlation.columns,yticklabels=correlation.columns,annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd55ad2f650>
```



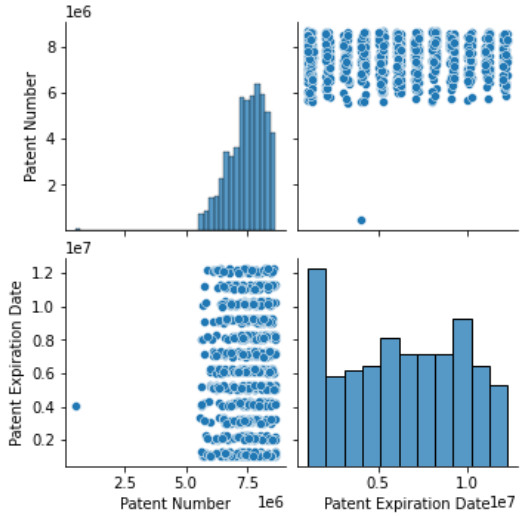

```
#The plot below shows the number of patents by every center over a period years
sns.relplot(x='Patent Number',y='Patent Expiration Date',hue='Center',data=df_corr)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fd55ac4b490>
```



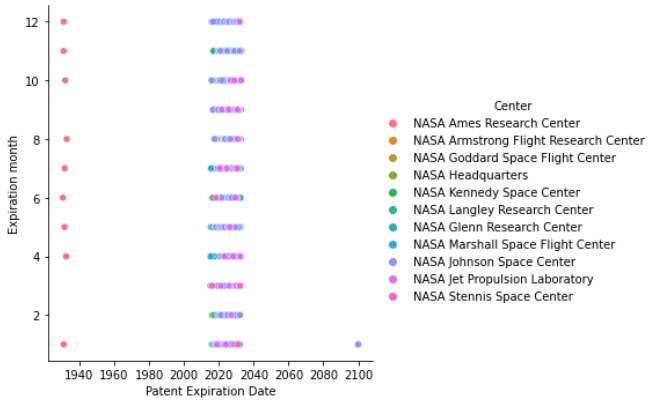
```
sns.pairplot(df_corr)
```

```
<seaborn.axisgrid.PairGrid at 0x7fd557f06110>
```




```
sns.relplot(x='Patent Expiration Date',y='Expiration month',hue='Center',data=df_corr)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fd5564d0e10>
```

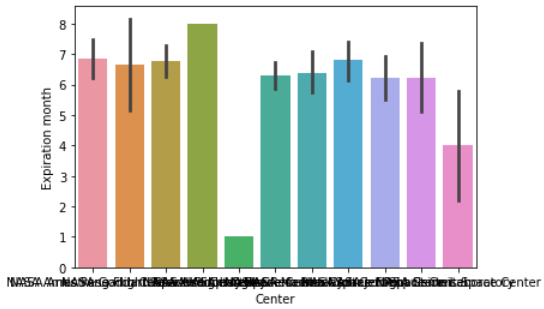



```
#Understanding statistics for a specific center
a.describe()
```

	Expiration month
count	100.000000
mean	6.850000
std	3.239201
min	1.000000
25%	4.750000
50%	7.000000
75%	9.000000
max	12.000000

```
#We are plotting the various centers and their patent's average expiration month in the plot below
import seaborn as sns
import matplotlib.pyplot as plt

sns.barplot(data=frame, x="Center", y="Expiration month")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd5563b0c10>
```



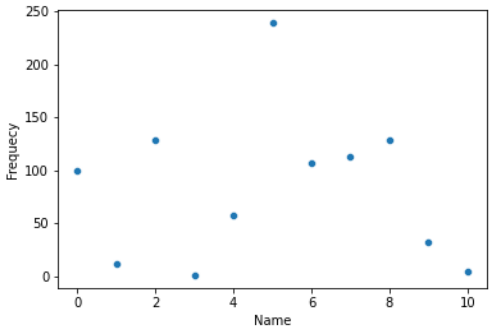

```
df_issue['Center'].replace(['NASA Ames Research Center','NASA Armstrong Flight Research Center','NASA Goddard Space Flight Center', 'NASA Headquarters',
       'NASA Kennedy Space Center', 'NASA Langley Research Center',
       'NASA Glenn Research Center', 'NASA Marshall Space Flight Center',
       'NASA Johnson Space Center', 'NASA Jet Propulsion Laboratory',
       'NASA Stennis Space Center'],[0,1,2,3,4,5,6,7,8,9,10],inplace=True)
df_issue
```

```
#Getting the Center number and its number of patents
cf
```

	Name	Frequecy
0	0	100
1	1	12
2	2	129
3	3	1
4	4	58
5	5	239
6	6	107
7	7	113
8	8	128
9	9	32
10	10	5

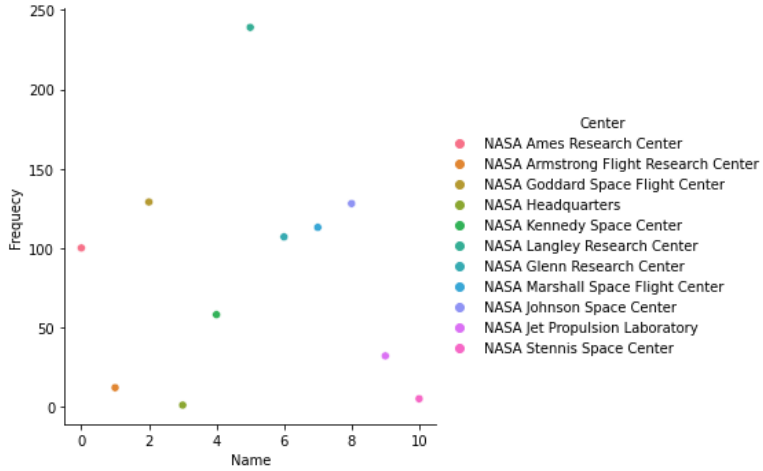
```
#Scatter plot for the same center and frequency
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure()
sns.scatterplot(data=cf,x='Name',y='Frequecy')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd5560a2fd0>
```



```
#Relation plot explaining how many patents each of the center has issued
sns.relplot(x='Name',y='Frequecy',hue='Center',data=cf)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fd56dad7e90>
```



Conclusion :

We got a raw data set here of NASA Patents which gave us details about the title of the patent , its status , the center name, patent number , patent expiration date, application number , case number. Out of all the attributes present I concluded that Center name , Status , Patent number, Patent Expiration date are some of the attributes that could be analyzed. So, initially I cleaned out the dataset by removing entries with no patent numbers and replaced the NaN dates of patent expiration with some far off dates. Thus, eliminating all the null values.

Now, depending on the status I extracted the data with Patent status "Issued". Then I found out the correlation between Expiration date and Patent Number and observed a positive correlation stating that the patent number increased as the date went ahead. I also plotted the heat map for the same for better understanding.

Then I moved to a plot for understanding the patents vs the date for various hues for each of the 11 centers mentioned here and also obtained a pairplot for the same.

Similarly in order to understand the center wise which year ranges have the highest number of issued patents across which expiration months , there is a plot for understanding the same.

To understand the average month of patent expiration date I plotted all the centers vs the expiration month on a bar plot , automatically all the months in which a particular center has patent expiry was averaged out to give one highest bar for each center.

Then we moved on to understand how many issued patents do each of the centers have for which I found out the center name allotted it a number and then found out the frequency occurrence of it in the data frame.Further, I plotted two graphs that showed the center vs no.of patents issued (multi-coloured and homogenous) which gave a clear understanding of the same.

All of the above cleaning and analysis happened on the "Issued" Patents segregated from the raw data.

Thus, through this experiment I understood the Exploratory Data Analysis that's required to analyze , clean and have a clear understanding for further processing.