Team Member Details:

Group Name: Data Hacks

Name: Suhas Yogeshwara

Email: suhas.gys1996@gmail.com

Country: Germany

College: SRH University of Applied Sciences Berlin

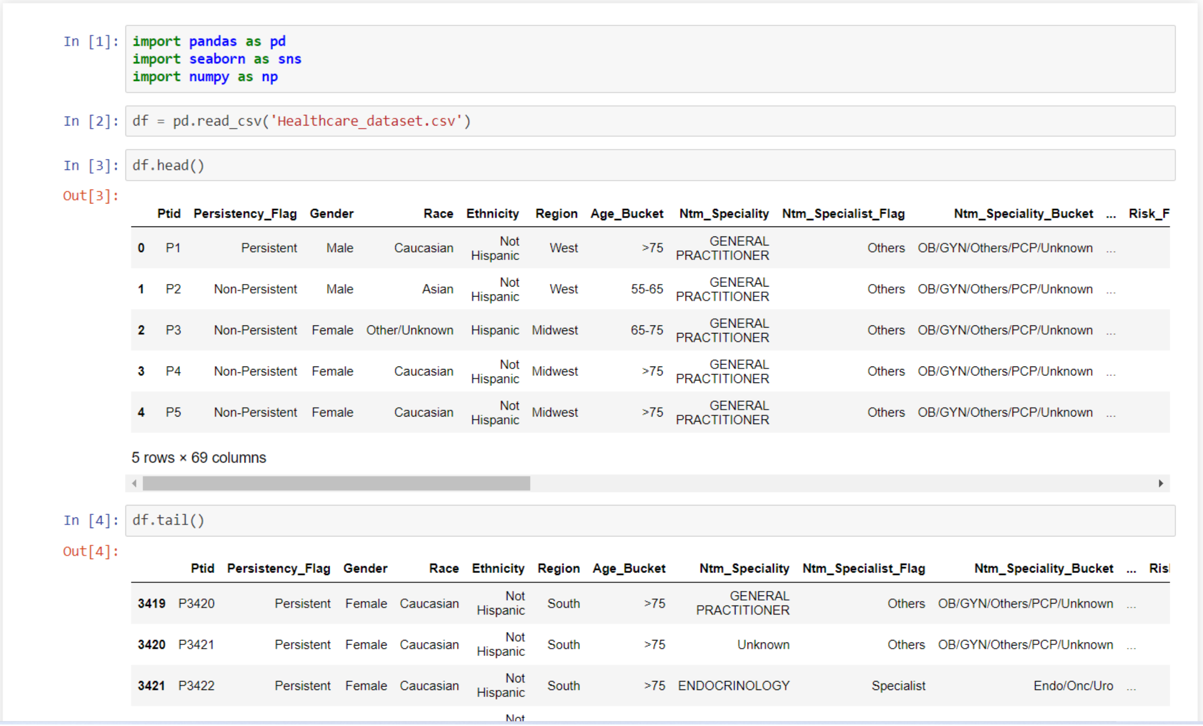
Specialization: Data Science

Problem Description:

The pharmaceutical industry is currently having trouble keeping track of whether a prescription remains to be applied in practice advised by a physician. Classification must be required in order to automate the procedure in to address this problem.

GITHUB repo link: https://github.com/SuhasYogeshwara1996/healthcare

Ipynb Notebook: EDA Repo Link



```
In [1]: import pandas as pd
        import seaborn as sns
        import numpy as np
```

```
In [2]: df = pd.read_csv('Healthcare_dataset.csv')
```

```
In [3]: df.head()
```

Out[3]:

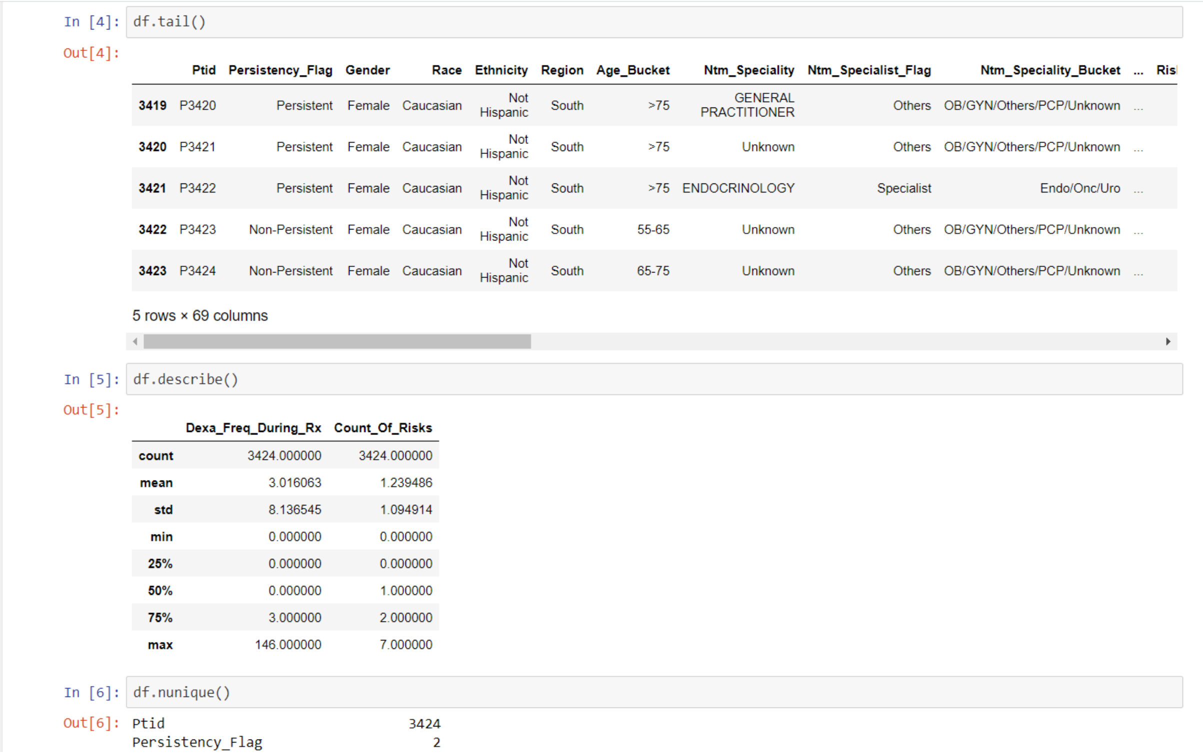
	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Risk_F
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	

5 rows × 69 columns

```
In [4]: df.tail()
```

Out[4]:

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Ris
3419	P3420	Persistent	Female	Caucasian	Not Hispanic	South	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
3420	P3421	Persistent	Female	Caucasian	Not Hispanic	South	>75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	
3421	P3422	Persistent	Female	Caucasian	Not Hispanic	South	>75	ENDOCRINOLOGY	Specialist	Endo/Onc/Uro	...	
					Not							



```
In [4]: df.tail()
Out[4]:
```

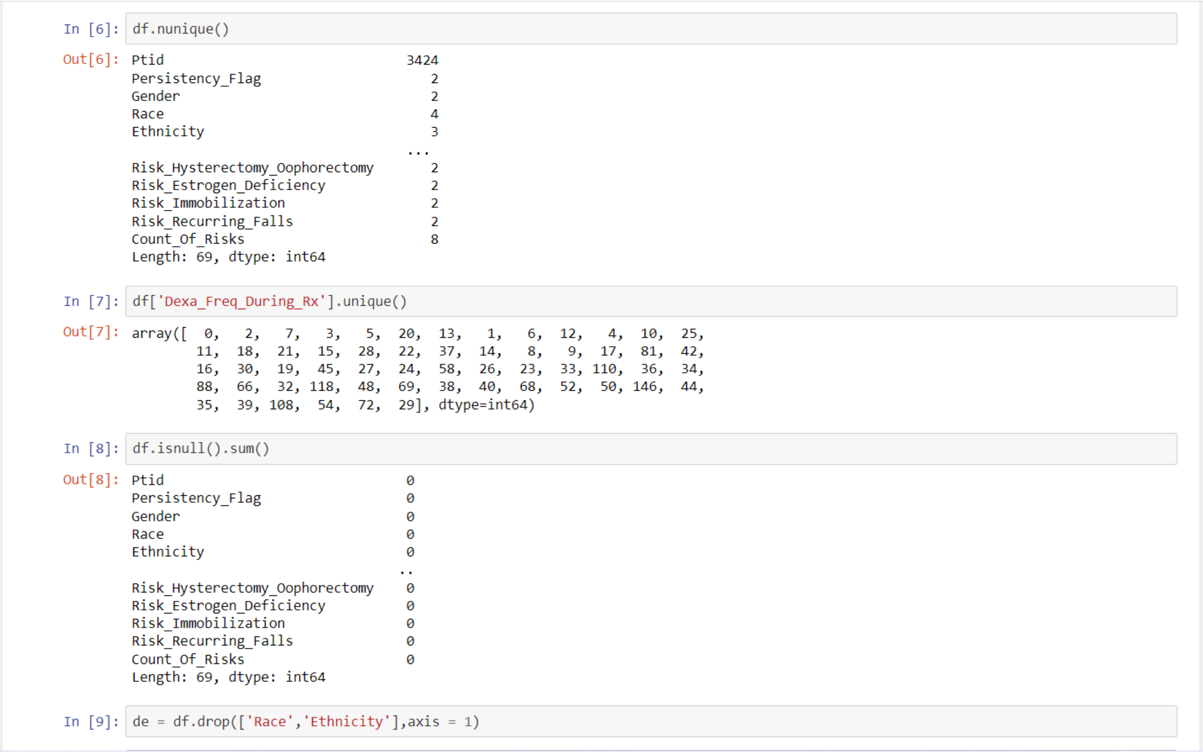
	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Ris
3419	P3420	Persistent	Female	Caucasian	Not Hispanic	South	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
3420	P3421	Persistent	Female	Caucasian	Not Hispanic	South	>75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	
3421	P3422	Persistent	Female	Caucasian	Not Hispanic	South	>75	ENDOCRINOLOGY	Specialist	Endo/Onc/Uro	...	
3422	P3423	Non-Persistent	Female	Caucasian	Not Hispanic	South	55-65	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	
3423	P3424	Non-Persistent	Female	Caucasian	Not Hispanic	South	65-75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	

5 rows × 69 columns

```
In [5]: df.describe()
Out[5]:
```

	Dexa_Freq_During_Rx	Count_Of_Risks
count	3424.000000	3424.000000
mean	3.016063	1.239486
std	8.136545	1.094914
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	1.000000
75%	3.000000	2.000000
max	146.000000	7.000000

```
In [6]: df.nunique()
Out[6]: Ptid                                  3424
        Persistency_Flag                         2
```

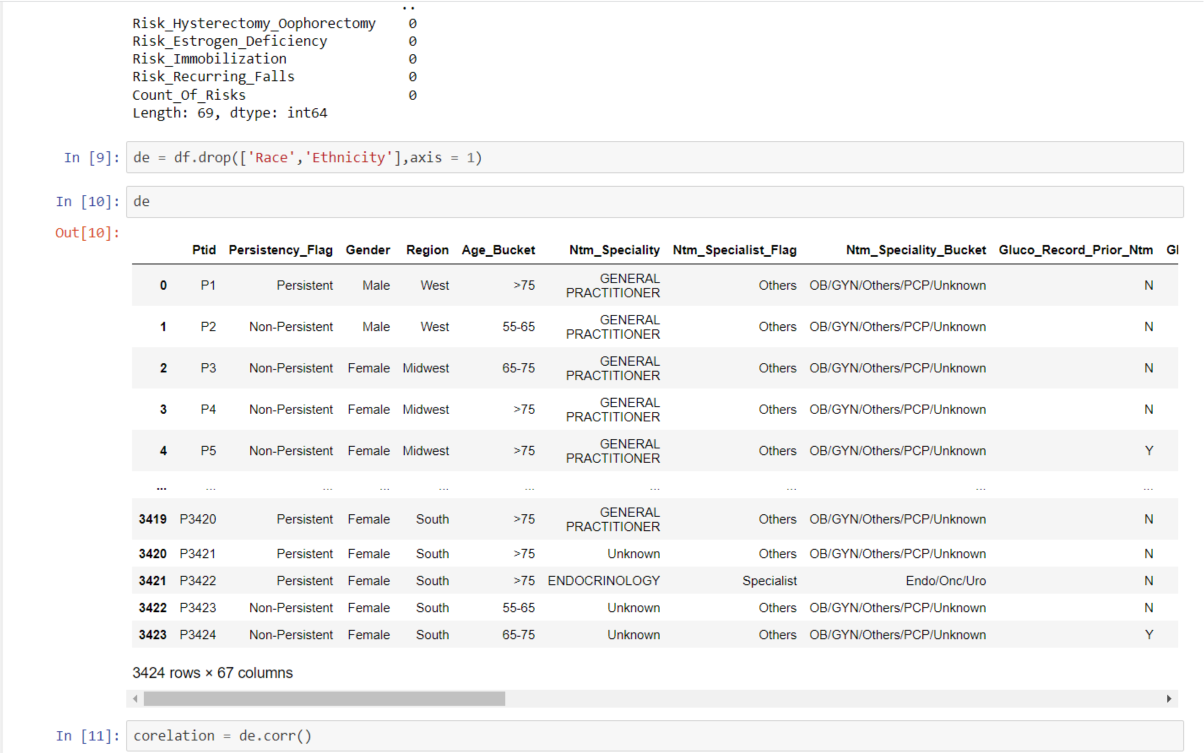


```
In [6]: df.nunique()
Out[6]: Ptid                                  3424
        Persistency_Flag                         2
        Gender                                   2
        Race                                     4
        Ethnicity                                3
                                              ...
        Risk_Hysterectomy_Oophorectomy           2
        Risk_Estrogen_Deficiency                 2
        Risk_Immobilization                      2
        Risk_Recurring_Falls                     2
        Count_Of_Risks                           8
        Length: 69, dtype: int64

In [7]: df['Dexa_Freq_During_Rx'].unique()
Out[7]: array([  0,   2,   7,   3,   5,  20,  13,   1,   6,  12,   4,  10,  25,
                11,  18,  21,  15,  28,  22,  37,  14,   8,   9,  17,  81,  42,
                16,  30,  19,  45,  27,  24,  58,  26,  23,  33, 110,  36,  34,
                88,  66,  32, 118,  48,  69,  38,  40,  68,  52,  50, 146,  44,
                35,  39, 108,  54,  72,  29], dtype=int64)

In [8]: df.isnull().sum()
Out[8]: Ptid                              0
        Persistency_Flag                  0
        Gender                            0
        Race                              0
        Ethnicity                         0
                                         ..
        Risk_Hysterectomy_Oophorectomy    0
        Risk_Estrogen_Deficiency          0
        Risk_Immobilization               0
        Risk_Recurring_Falls              0
        Count_Of_Risks                    0
        Length: 69, dtype: int64

In [9]: de = df.drop(['Race','Ethnicity'],axis = 1)
```



```
                                        ..
Risk_Hysterectomy_Oophorectomy     0
Risk_Estrogen_Deficiency           0
Risk_Immobilization                0
Risk_Recurring_Falls               0
Count_Of_Risks                     0
Length: 69, dtype: int64
```

```
In [9]: de = df.drop(['Race','Ethnicity'],axis = 1)
```

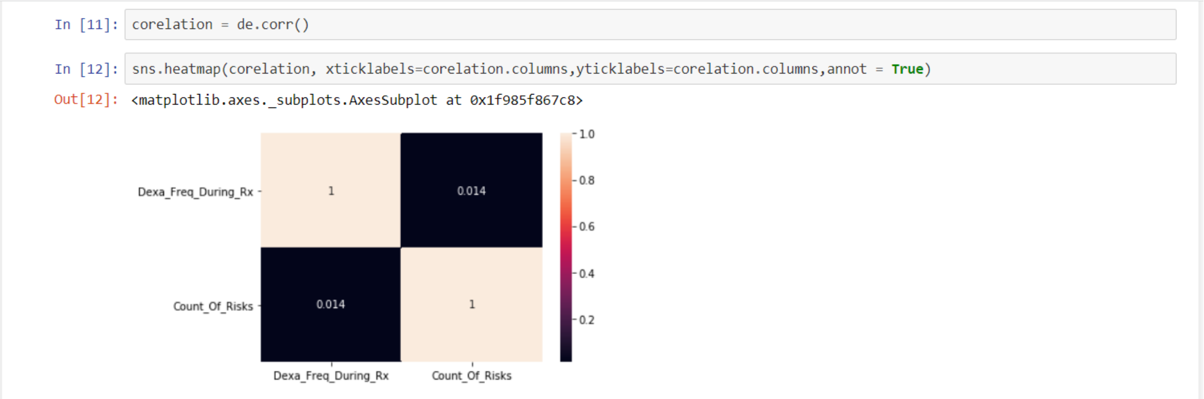
```
In [10]: de
```

Out[10]:

	Ptid	Persistency_Flag	Gender	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	Gluco_Record_Prior_Ntm	Gl
0	P1	Persistent	Male	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	
1	P2	Non-Persistent	Male	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	
2	P3	Non-Persistent	Female	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	
3	P4	Non-Persistent	Female	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	
4	P5	Non-Persistent	Female	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	Y	
...	...	...	...	...	...	...	...	...	...	
3419	P3420	Persistent	Female	South	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	
3420	P3421	Persistent	Female	South	>75	Unknown	Others	OB/GYN/Others/PCP/Unknown	N	
3421	P3422	Persistent	Female	South	>75	ENDOCRINOLOGY	Specialist	Endo/Onc/Uro	N	
3422	P3423	Non-Persistent	Female	South	55-65	Unknown	Others	OB/GYN/Others/PCP/Unknown	N	
3423	P3424	Non-Persistent	Female	South	65-75	Unknown	Others	OB/GYN/Others/PCP/Unknown	Y	

3424 rows × 67 columns

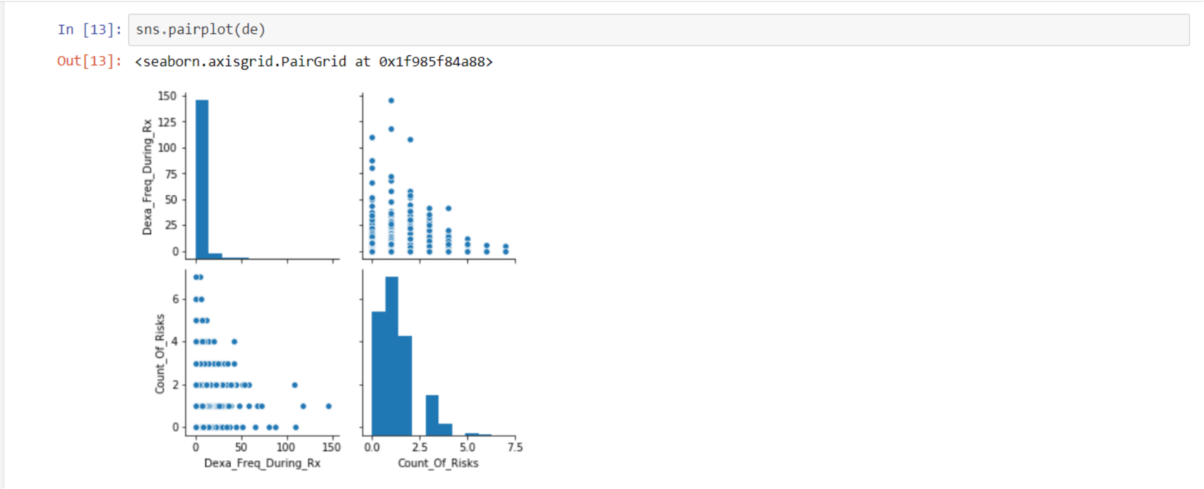
```
In [11]: corelation = de.corr()
```



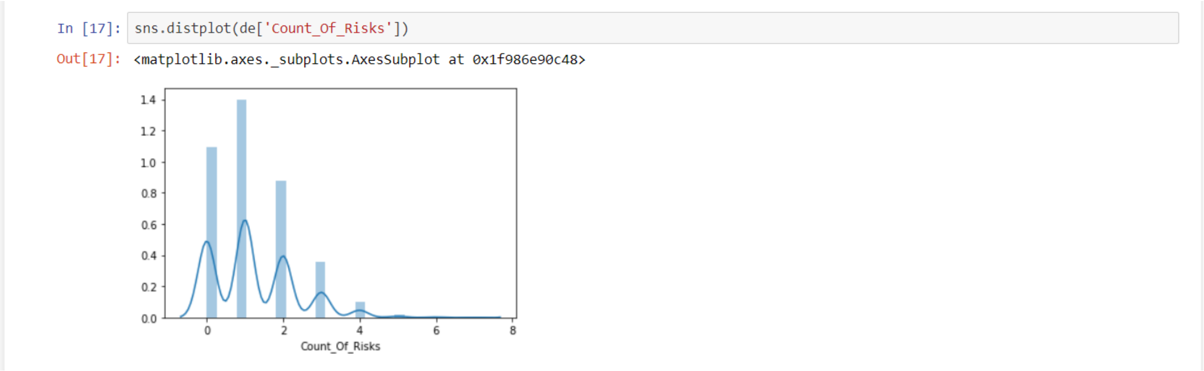
```
In [11]: corelation = de.corr()
```

```
In [12]: sns.heatmap(corelation, xticklabels=corelation.columns,yticklabels=corelation.columns,annot = True)
```

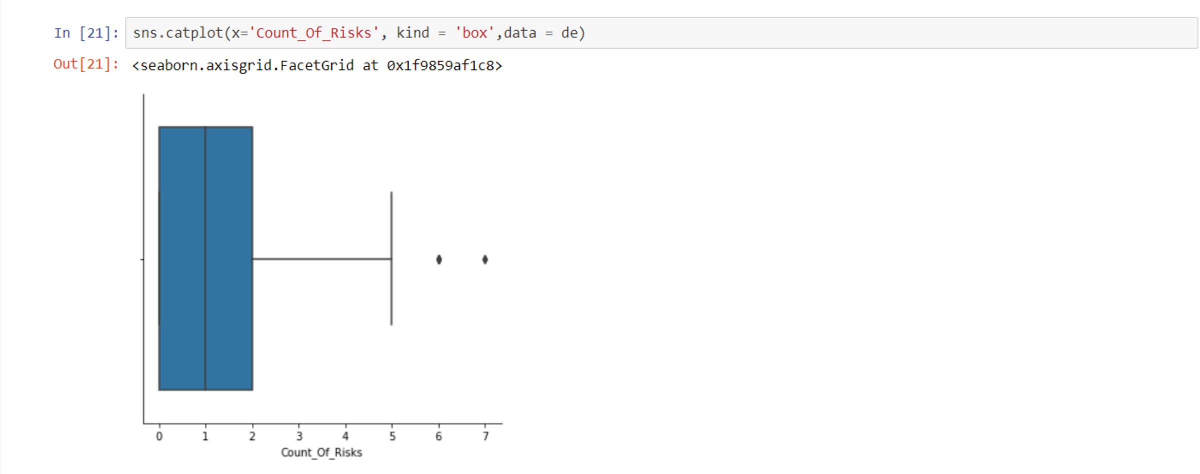
```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x1f985f867c8>
```



```
In [13]: sns.pairplot(de)
Out[13]: <seaborn.axisgrid.PairGrid at 0x1f985f84a88>
```



```
In [17]: sns.distplot(de['Count_Of_Risks'])
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x1f986e90c48>
```



```
In [21]: sns.catplot(x='Count_Of_Risks', kind = 'box',data = de)
Out[21]: <seaborn.axisgrid.FacetGrid at 0x1f9859af1c8>
```

Final Recommendation: Distplot and Catplot are the best Recommendations for the Data Analysis