Week 10: Deliverables

Team member's details:

Group Name - Chandni

Name – Chandni

Email – cpatel521351@hotmail.com

Country – England

College/Company – NA

Specialization - Data Science

Github - https://github.com/c5678/healthcare/tree/main/4

Problem description

- ABC is a pharma company who wants to understand the persistency of drug as per the physician prescription for the patient. To solve this problem ABC pharma company approached us to automate the process of identification for the drug. This week we are going to anaylse the clean data to get valuable insight.

EDA

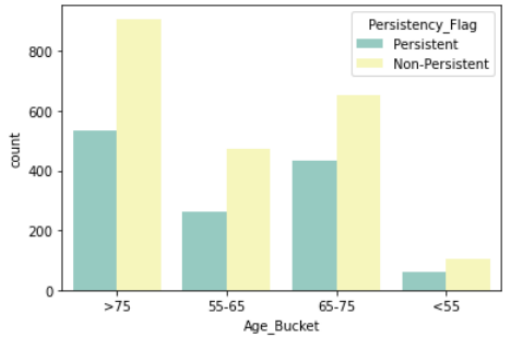
Firstly, we double checked the clean dataset making sure there are no nulls values or missing fields.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 64 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Persistency_Flag                                        3424 non-null   int64
 1   Gender                                                  3424 non-null   int64
 2   Race                                                    3424 non-null   int64
 3   Ethnicity                                               3424 non-null   int64
 4   Region                                                  3424 non-null   int64
 5   Age_Bucket                                              3424 non-null   int64
 6   Ntm_Speciality                                          3424 non-null   int64
 7   Ntm_Specialist_Flag                                     3424 non-null   int64
 8   Ntm_Speciality_Bucket                                   3424 non-null   int64
 9   Gluco_Record_Prior_Ntm                                  3424 non-null   int64
 10  Gluco_Record_During_Rx                                  3424 non-null   int64
 11  Dexa_Freq_During_Rx                                     3424 non-null   float64
 12  Dexa_During_Rx                                          3424 non-null   int64
 13  Frag_Frac_Prior_Ntm                                     3424 non-null   int64
 14  Frag_Frac_During_Rx                                     3424 non-null   int64
 15  Risk_Segment_Prior_Ntm                                  3424 non-null   int64
 16  Tscore_Bucket_Prior_Ntm                                 3424 non-null   int64
 17  Adherent_Flag                                           3424 non-null   int64
 18  Idn_Indicator                                           3424 non-null   int64
 19  Injectable_Experience_During_Rx                         3424 non-null   int64
 20  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms  3424 non-null   int64
 21  Comorb_Encounter_For_Immunization                       3424 non-null   int64
 22  Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx  3424 non-null  int64
 23  Comorb_Vitamin_D_Deficiency                             3424 non-null   int64
 24  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified    3424 non-null   int64
 25  Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx  3424 non-null  int64
```

More people age >75 will be persistent to drug compared to <55.

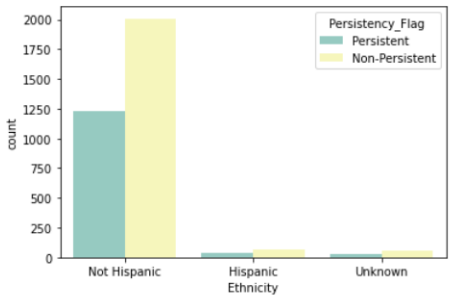
```
: age = sns.countplot(x="Age_Bucket", hue="Persistency_Flag", data=df1, palette="Set3")
```



people with Hispanic ethnicity was less persistent to the drug compared to not Hispanic ethnicity.

```
In [26]: sns.countplot(x="Ethnicity", hue="Persistency_Flag", data=df1, palette="Set3")
```

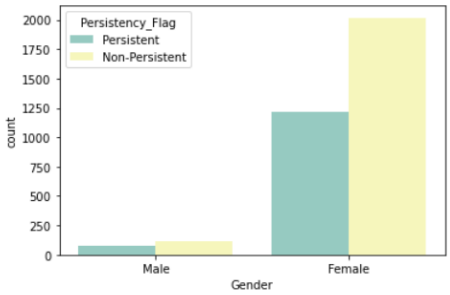
```
Out[26]: <AxesSubplot:xlabel='Ethnicity', ylabel='count'>
```



females was more persistent to the drug compared to male.

```
In [27]: sns.countplot(x="Gender", hue="Persistency_Flag", data=df1, palette="Set3")
```

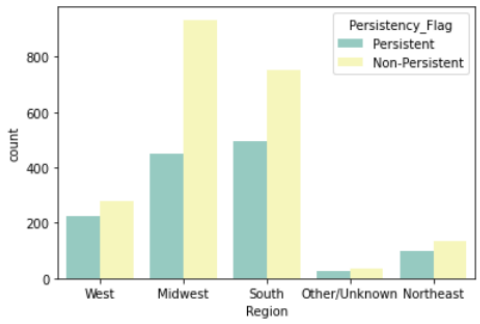
```
Out[27]: <AxesSubplot:xlabel='Gender', ylabel='count'>
```



People living in South Region are more persistent to drug compared to people living in the Midwest

```
]: sns.countplot(x="Region", hue="Persistency_Flag", data=df1, palette="Set3")
```

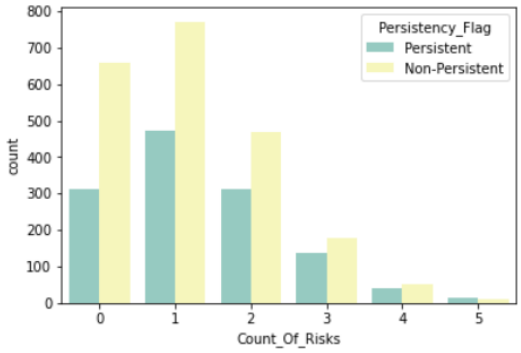
```
]: <AxesSubplot:xlabel='Region', ylabel='count'>
```



Looking at the plot below we can see as the number of Risks increase the ratio of Persistent increases compared to Non-Persistent.

```
sns.countplot(x="Count_Of_Risks", hue="Persistency_Flag", data=df1, palette="Set3")
```

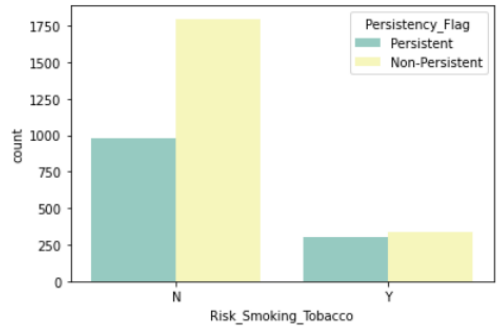
```
<AxesSubplot:xlabel='Count_Of_Risks', ylabel='count'>
```



There are more Non-Persistent people who don’t smoke compared to Non-Persistent Smokers.

```
sns.countplot(x="Risk_Smoking_Tobacco", hue="Persistency_Flag", data=df1, palette="Set3")
```

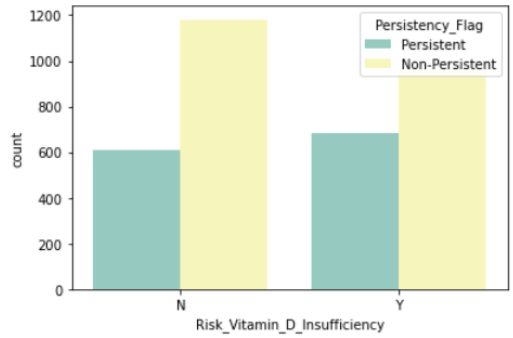
```
<AxesSubplot:xlabel='Risk_Smoking_Tobacco', ylabel='count'>
```



There is a slight increase in persistency for drug if a person has Vitamin D insufficiency compared to Persistency for drug is a person does not have Vitamin D insufficiency.

```
: sns.countplot(x="Risk_Vitamin_D_Insufficiency", hue="Persistency_Flag", data=df1, palette="Set3")
```

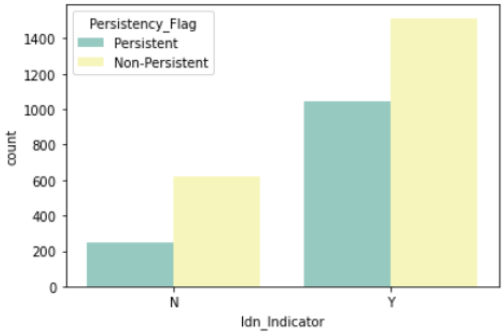
```
: <AxesSubplot:xlabel='Risk_Vitamin_D_Insufficiency', ylabel='count'>
```



Patient mapped to IDN show more non-persistency to drug.

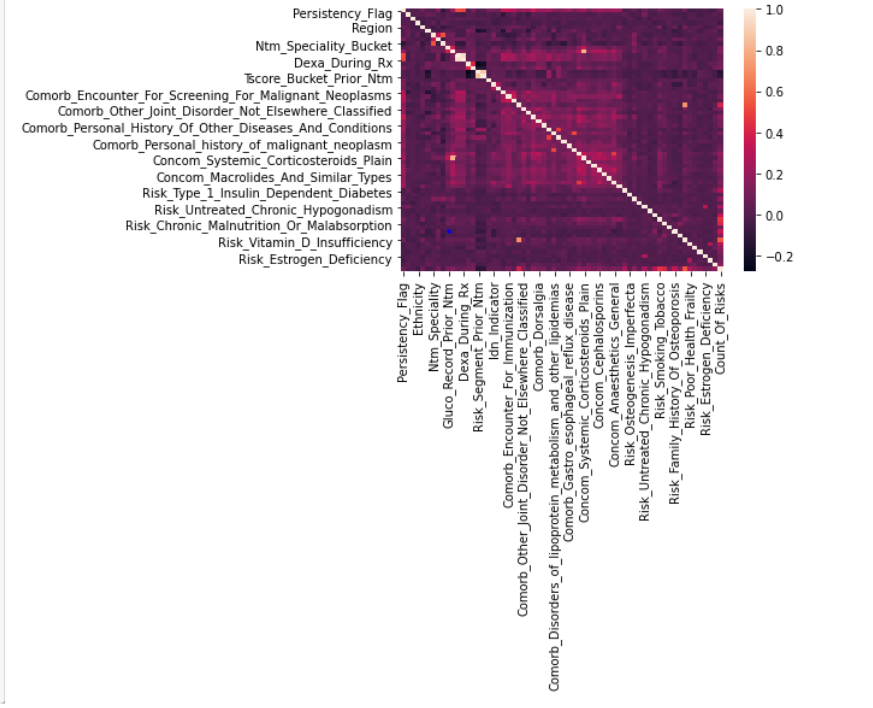
```
sns.countplot(x='Idn_Indicator',hue='Persistency_Flag',data=df1,palette='Set3')
```

```
<AxesSubplot:xlabel='Idn_Indicator', ylabel='count'>
```



We also created a heat map to figure out which feature collerated with one another.

```
corrMatrix = df.corr()
sns.heatmap(corrMatrix)
plt.show()
```



On initial analysis Dexa_Freq_During_Rx seemed to be the most relevant feature related to Persistency_Flag.

```
In [57]: cor = df.corr()
         cor_target = abs(cor["Persistency_Flag"])

In [47]: relevant_features = cor_target[cor_target>0.5]

In [48]: relevant_features

Out[48]: Persistency_Flag       1.000000
         Dexa_Freq_During_Rx    0.517337
         Name: Persistency_Flag, dtype: float64
```