Explorative Data Analysis

April 20, 2023

```
In [0]: %pylab inline
```

```
In [1]: import dataiku
        from dataiku import pandasutils as pdu
        import pandas as pd
```

```
/data/dataiku/dss_data/code-envs/python/QB_HCP_propensity/lib/python3.6/site-packages/requests/_
  RequestsDependencyWarning)
```

```
In [2]: # Example: load a DSS dataset as a Pandas dataframe
        hospital_mortality_data = dataiku.Dataset("Hospital_Mortality_Dataset")
        hospital_mortality_data_df = hospital_mortality_data.get_dataframe()
```

```
In [3]: # Data Science Questions:

        # Which age group is most in the hospital?
        # Which age group of patients dies more in the hospital?
        # Which gender is the most prevalent in the hospital?
        # How many patients died in the hospital with atrial fibrillation?
        # How many patients in the hospital have depression?
        # How many patients in the hospital have depression?
        # What is the rate of non-survived patients with hypertension?
        # How many patients Alive in the hospital they are with renal failure?
        # How many patients Death in the hospital they are with Hyperlipemia ?
        # How many patients Death in the hospital they are with Anemia?
```

```
In [4]: hospital_mortality_data_df.head()
```

```
Out[4]:    group      ID  outcome  age  gendera        BMI  hypertensive  atrialfibrillation  CH
        0      1  125047      0.0   72        1  37.588179             0                   0
        1      1  139812      0.0   75        2        NaN             0                   0
        2      1  109787      0.0   83        2  26.572634             0                   0
        3      1  130587      0.0   43        2  83.264629             0                   0
        4      1  138290      0.0   75        2  31.824842             1                   0
```

```
In [6]: import pandas as pd
        import numpy as np
        import seaborn as sbn
        import seaborn as sb
        import matplotlib
        from matplotlib import pyplot as plt
        %matplotlib inline
        import seaborn as sns
        import sys
        import warnings
```

```
In [8]: hospital_mortality_data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1177 entries, 0 to 1176
Data columns (total 51 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   group                     1177 non-null   int64  
 1   ID                        1177 non-null   int64  
 2   outcome                   1176 non-null   float64
 3   age                       1177 non-null   int64  
 4   gendera                   1177 non-null   int64  
 5   BMI                       962 non-null    float64
 6   hypertensive              1177 non-null   int64  
 7   atrialfibrillation        1177 non-null   int64  
 8   CHD with no MI            1177 non-null   int64  
 9   diabetes                  1177 non-null   int64  
 10  deficiencyanemias         1177 non-null   int64  
 11  depression                1177 non-null   int64  
 12  Hyperlipemia              1177 non-null   int64  
 13  Renal failure             1177 non-null   int64  
 14  COPD                      1177 non-null   int64  
 15  heart rate                1164 non-null   float64
 16  Systolic blood pressure   1161 non-null   float64
 17  Diastolic blood pressure  1161 non-null   float64
 18  Respiratory rate          1164 non-null   float64
 19  temperature               1158 non-null   float64
 20  SP O2                     1164 non-null   float64
 21  Urine output              1141 non-null   float64
 22  hematocrit                1177 non-null   float64
 23  RBC                       1177 non-null   float64
 24  MCH                       1177 non-null   float64
 25  MCHC                      1177 non-null   float64
 26  MCV                       1177 non-null   float64
 27  RDW                       1177 non-null   float64
 28  Leucocyte                 1177 non-null   float64
 29  Platelets                 1177 non-null   float64
 30  Neutrophils               1033 non-null   float64
 31  Basophils                 918 non-null    float64
 32  Lymphocyte                1032 non-null   float64
 33  PT                        1157 non-null   float64
 34  INR                       1157 non-null   float64
 35  NT-proBNP                 1177 non-null   float64
 36  Creatine kinase           1012 non-null   float64
 37  Creatinine                1177 non-null   float64
 38  Urea nitrogen             1177 non-null   float64
 39  glucose                   1159 non-null   float64
 40  Blood potassium           1177 non-null   float64
 41  Blood sodium              1177 non-null   float64
 42  Blood calcium             1176 non-null   float64
 43  Chloride                  1177 non-null   float64
 44  Anion gap                 1177 non-null   float64
 45  Magnesium ion             1177 non-null   float64
 46  PH                        885 non-null    float64
 47  Bicarbonate               1177 non-null   float64
 48  Lactic acid               948 non-null    float64
 49  PCO2                      883 non-null    float64
 50  EF                        1177 non-null   int64
dtypes: float64(37), int64(14)
memory usage: 469.1 KB
```

```
In [9]: len(hospital_mortality_data_df)
```

```
Out[9]: 1177
```

```
In [10]: hospital_mortality_data_df.drop(['group','ID'],axis=1,inplace=True)
```

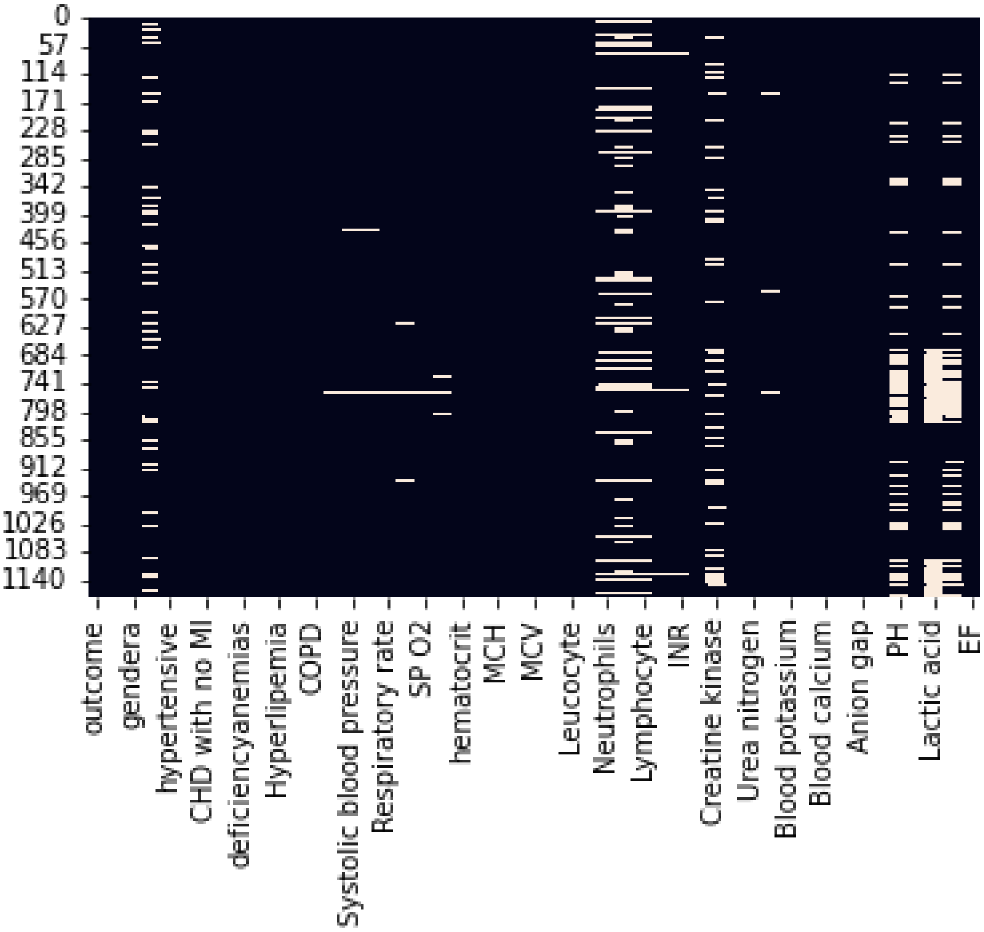
```
In [11]: hospital_mortality_data_df.isnull().sum()
```

```
Out[11]: outcome                       1
         age                           0
         gendera                       0
         BMI                         215
         hypertensive                  0
         atrialfibrillation            0
         CHD with no MI                0
         diabetes                      0
         deficiencyanemias             0
         depression                    0
         Hyperlipemia                  0
         Renal failure                 0
         COPD                          0
         heart rate                   13
         Systolic blood pressure      16
         Diastolic blood pressure     16
         Respiratory rate             13
         temperature                  19
```

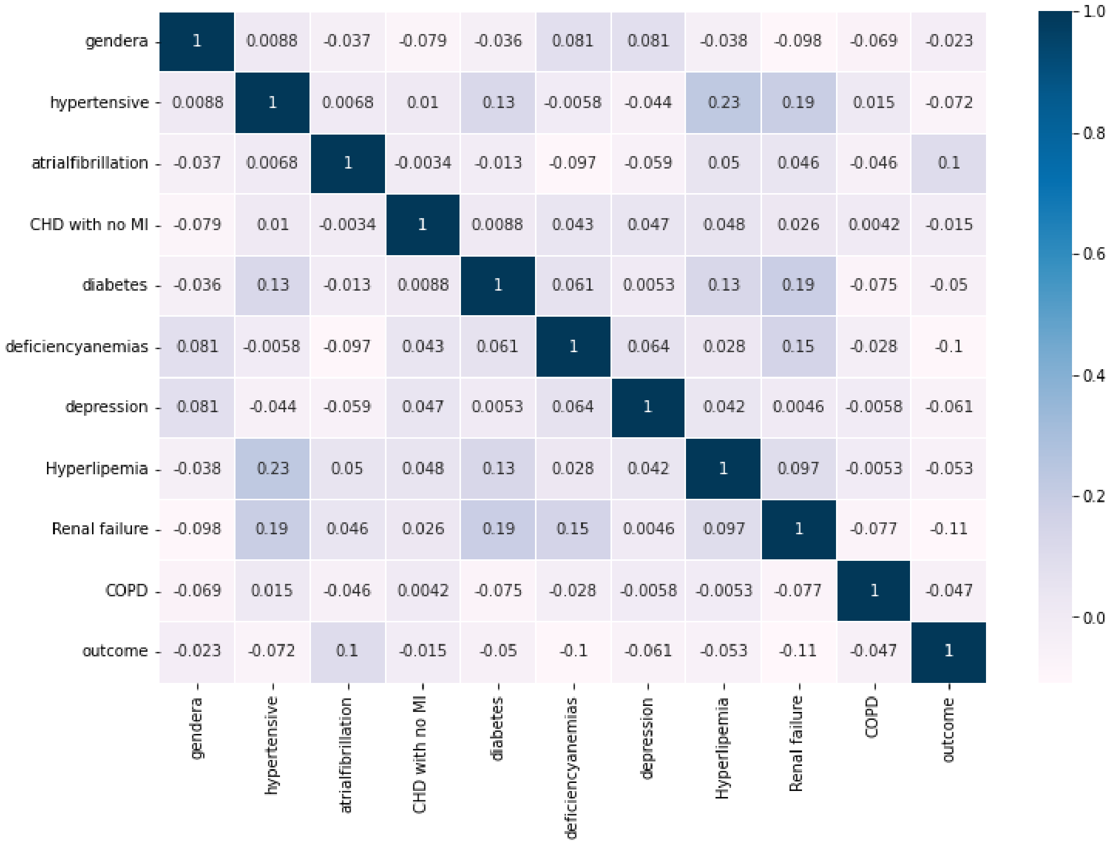
```
SP O2                        13
Urine output                 36
hematocrit                    0
RBC                           0
MCH                           0
MCHC                          0
MCV                           0
RDW                           0
Leucocyte                     0
Platelets                     0
Neutrophils                 144
Basophils                   259
Lymphocyte                  145
PT                           20
INR                          20
NT-proBNP                     0
Creatine kinase             165
Creatinine                    0
Urea nitrogen                 0
glucose                      18
Blood potassium               0
Blood sodium                  0
Blood calcium                 1
Chloride                      0
Anion gap                     0
Magnesium ion                 0
PH                          292
Bicarbonate                   0
Lactic acid                 229
PCO2                        294
EF                            0
dtype: int64
```

```
In [12]: sns.heatmap(hospital_mortality_data_df.isnull(), cbar=False)
```

```
Out[12]: <AxesSubplot:>
```

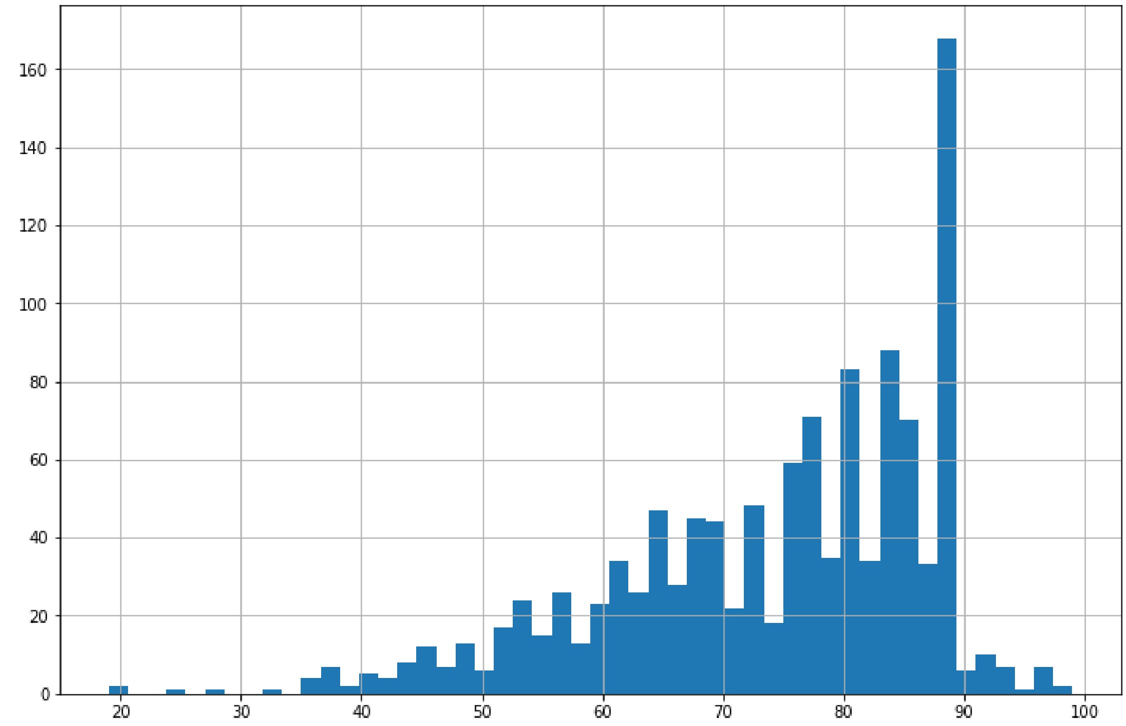


```
In [13]: col = ['gendera', 'hypertensive','atrialfibrillation', 'CHD with no MI', 'diabetes', 'd
                'depression', 'Hyperlipemia', 'Renal failure', 'COPD', 'outcome']

In [15]: corr = hospital_mortality_data_df[col].corr()

In [16]: plt.figure(figsize=(12,8))
         sns.heatmap(corr, annot=True, cmap='PuBu',linewidths=0.01,linecolor="white");
```



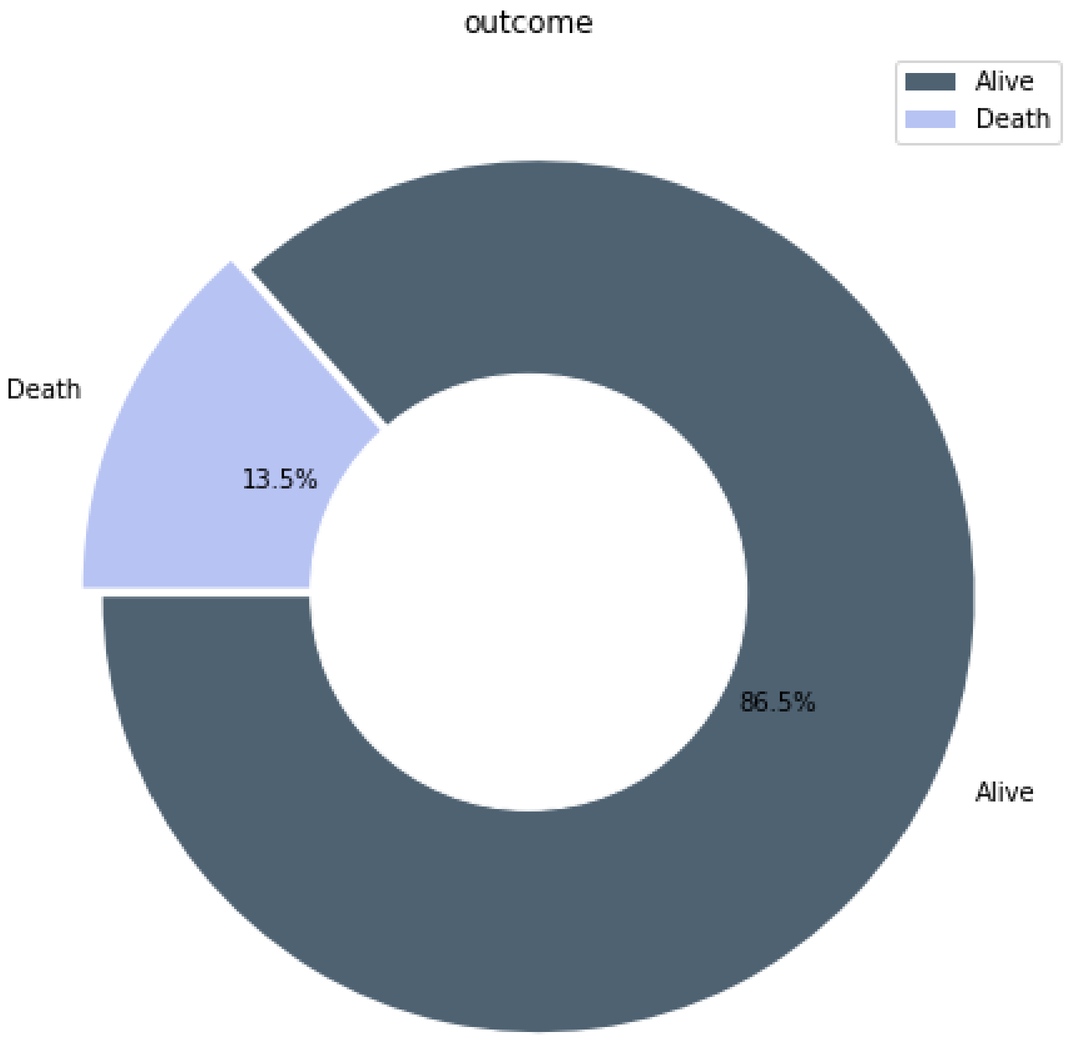
0.0.1 Which age group is most in the hospital?

```
In [18]: import matplotlib.pyplot as plt
         hospital_mortality_data_df.age.hist(bins = 50, figsize=(12,8))
         plt.show()
```



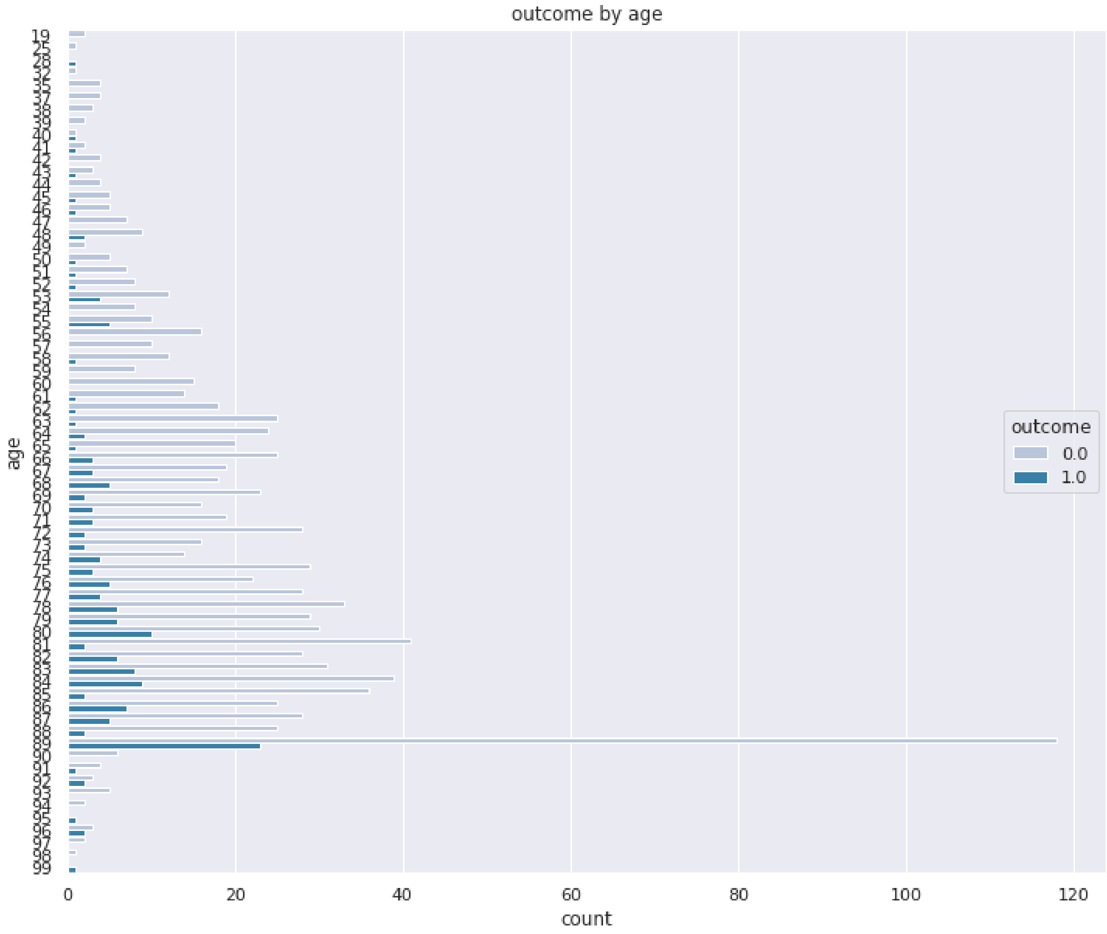
Here we can see that 89 age group are most in the hospital

```
In [21]: plt.figure(figsize=(12,8))
         plt.title("outcome")
         circle = plt.Circle((0, 0), 0.5, color='white')
         g = plt.pie(hospital_mortality_data_df.outcome.value_counts(), explode=(0.025,0.025),
         plt.legend()
         p = plt.gcf()
         p.gca().add_artist(circle)
         plt.show()
```

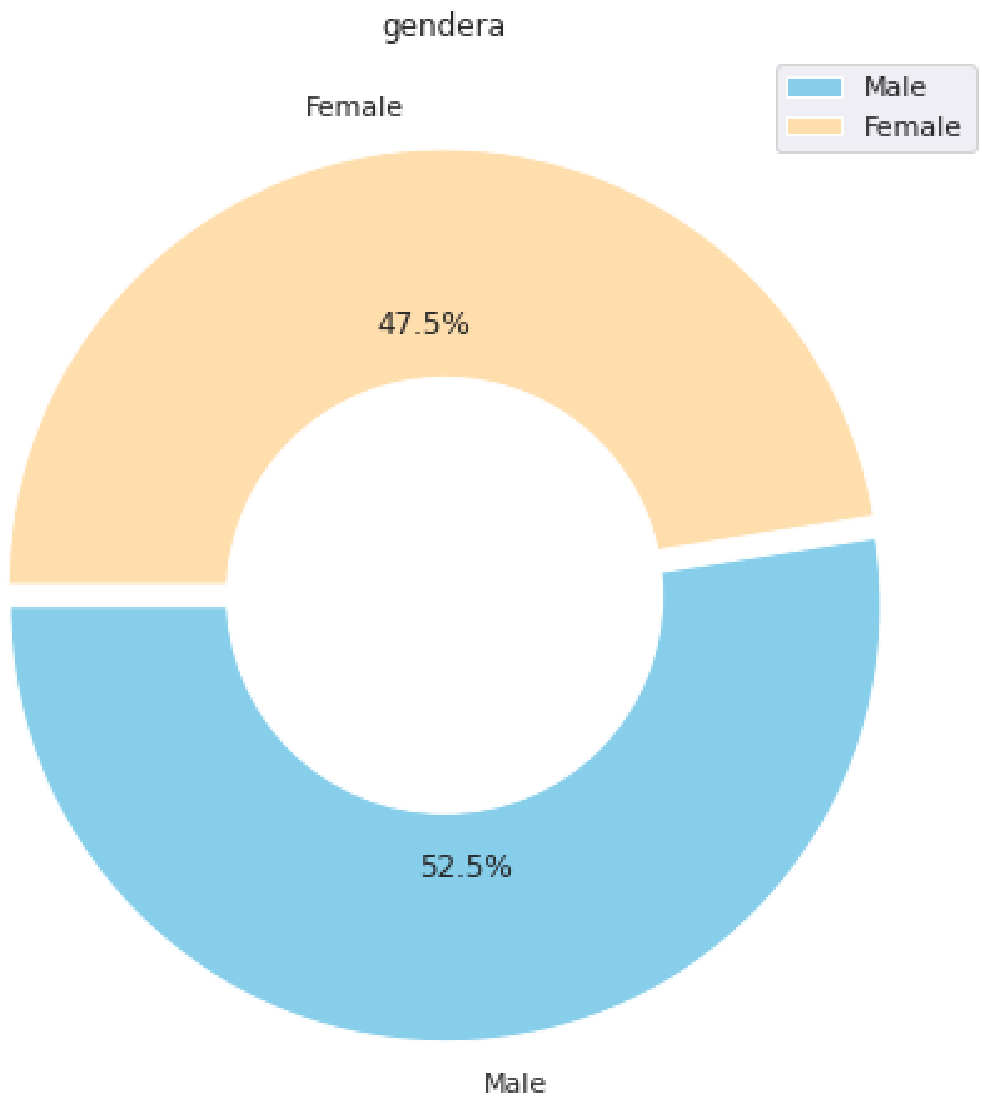


More then 15% patients are died in the hospital remaining patients were alive

```
In [23]: plt.figure(figsize=(12,10))
         sns.set_theme(style="darkgrid", color_codes=True)
         ax = sns.countplot(y="age", hue="outcome", data=hospital_mortality_data_df, palette="Pu
         plt.title("outcome by age")
         plt.show()
```

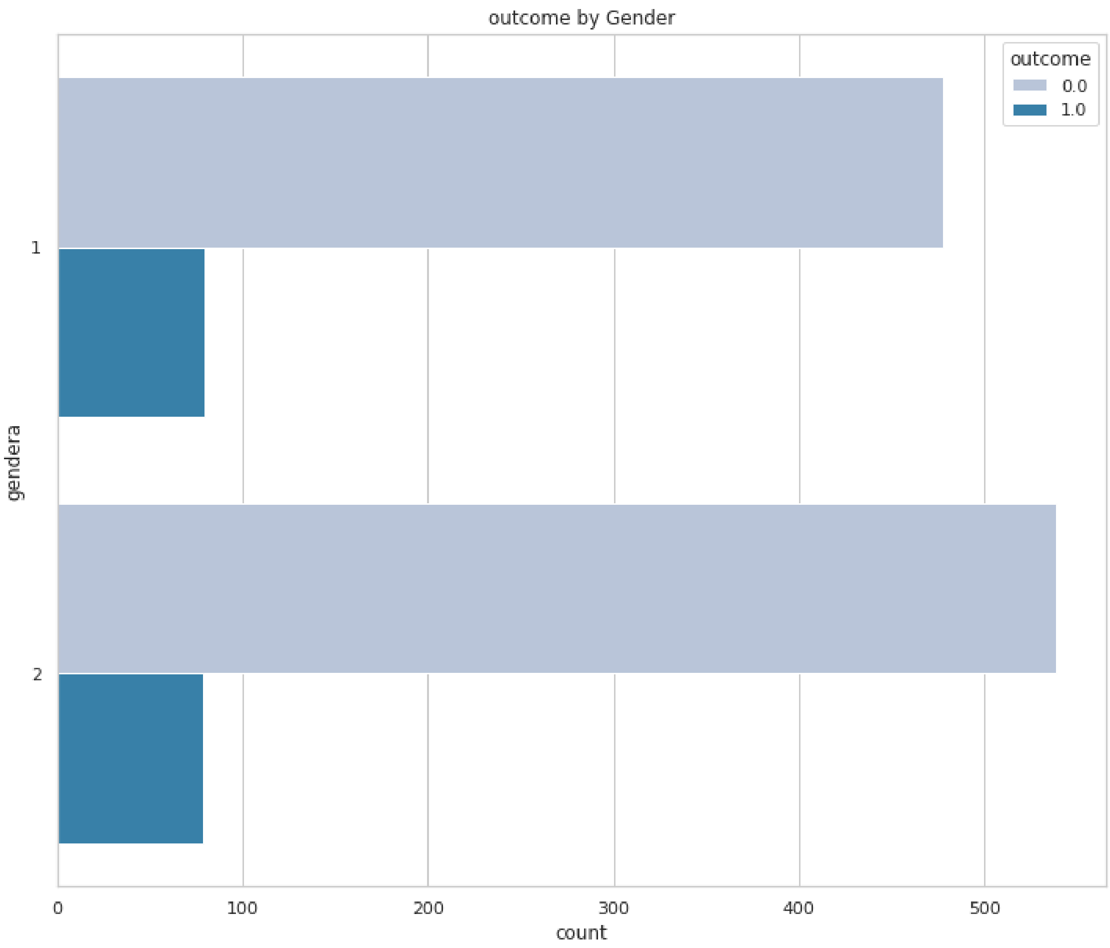


```
In [24]: plt.figure(figsize=(12,8))
         plt.title("gendera")
         circle = plt.Circle((0, 0), 0.5, color='white')
         g = plt.pie(hospital_mortality_data_df.gendera.value_counts(), explode=(0.025,0.025),
         plt.legend()
         p = plt.gcf()
         p.gca().add_artist(circle)
         plt.show()
```

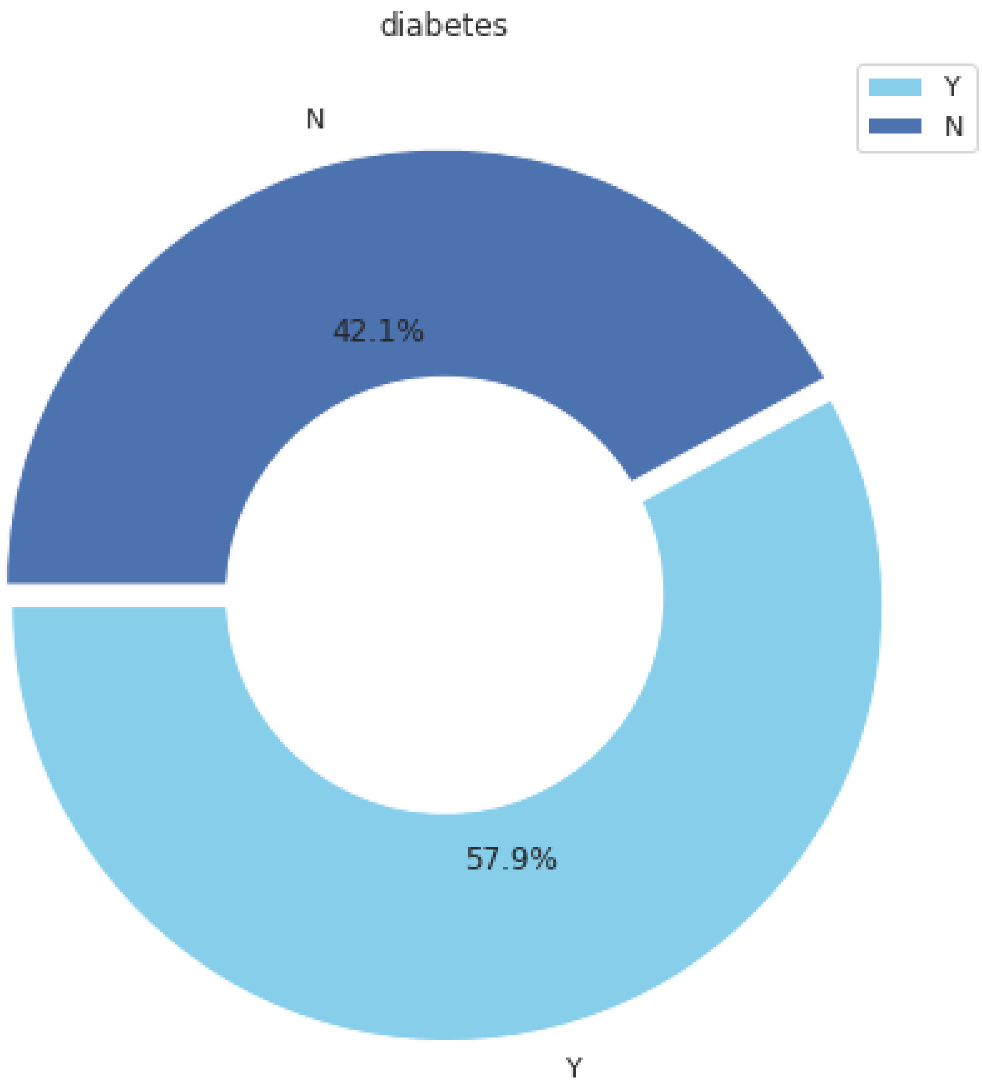


```
In [25]: df = hospital_mortality_data_df
```

```
In [26]: plt.figure(figsize=(12,10))
         sns.set_theme(style="whitegrid", color_codes=True)
         ax = sns.countplot(y="gendera", hue="outcome", data=df, palette="PuBu")
         plt.title("outcome by Gender")
         plt.show()
```

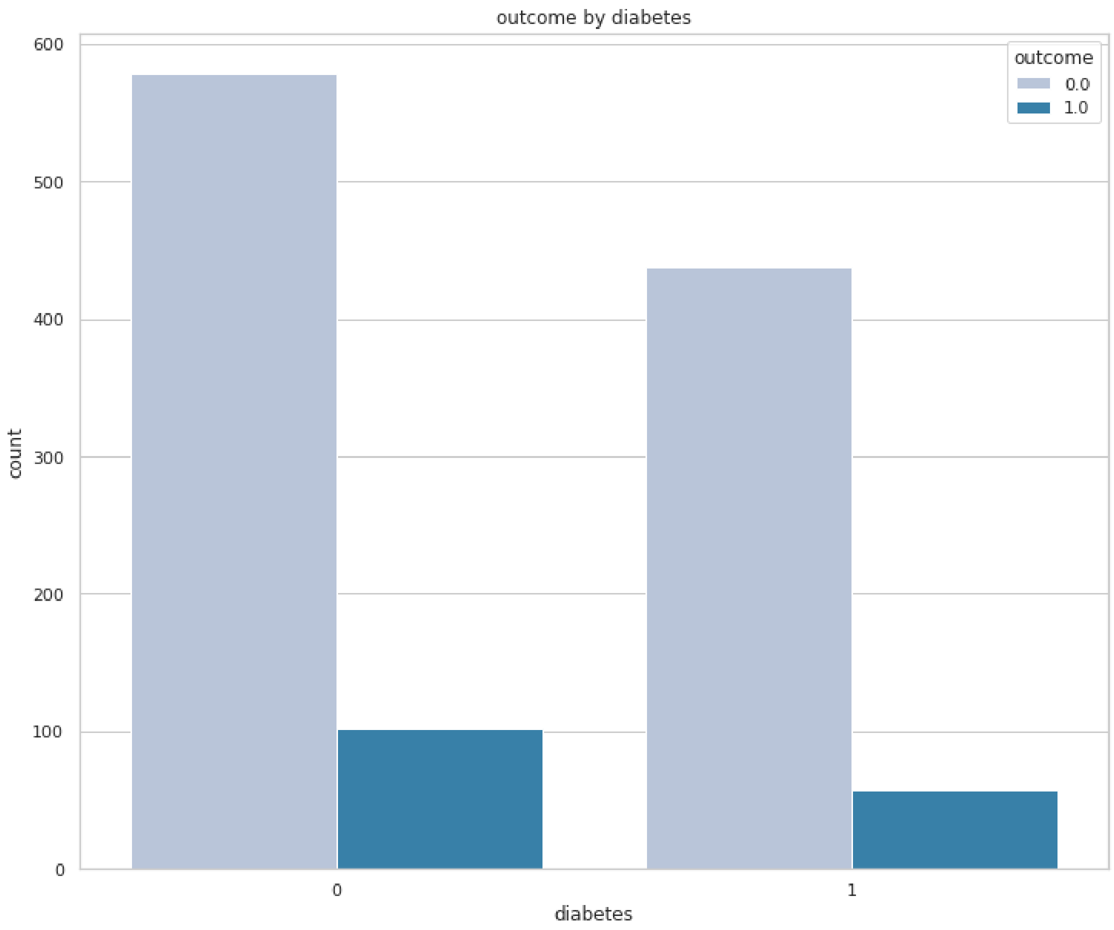


```
In [27]: plt.figure(figsize=(12,8))
         plt.title("diabetes")
         circle = plt.Circle((0, 0), 0.5, color='white')
         g = plt.pie(df.diabetes.value_counts(), explode=(0.025,0.025), labels=['Y','N'], colors
         plt.legend()
         p = plt.gcf()
         p.gca().add_artist(circle)
         plt.show()
```

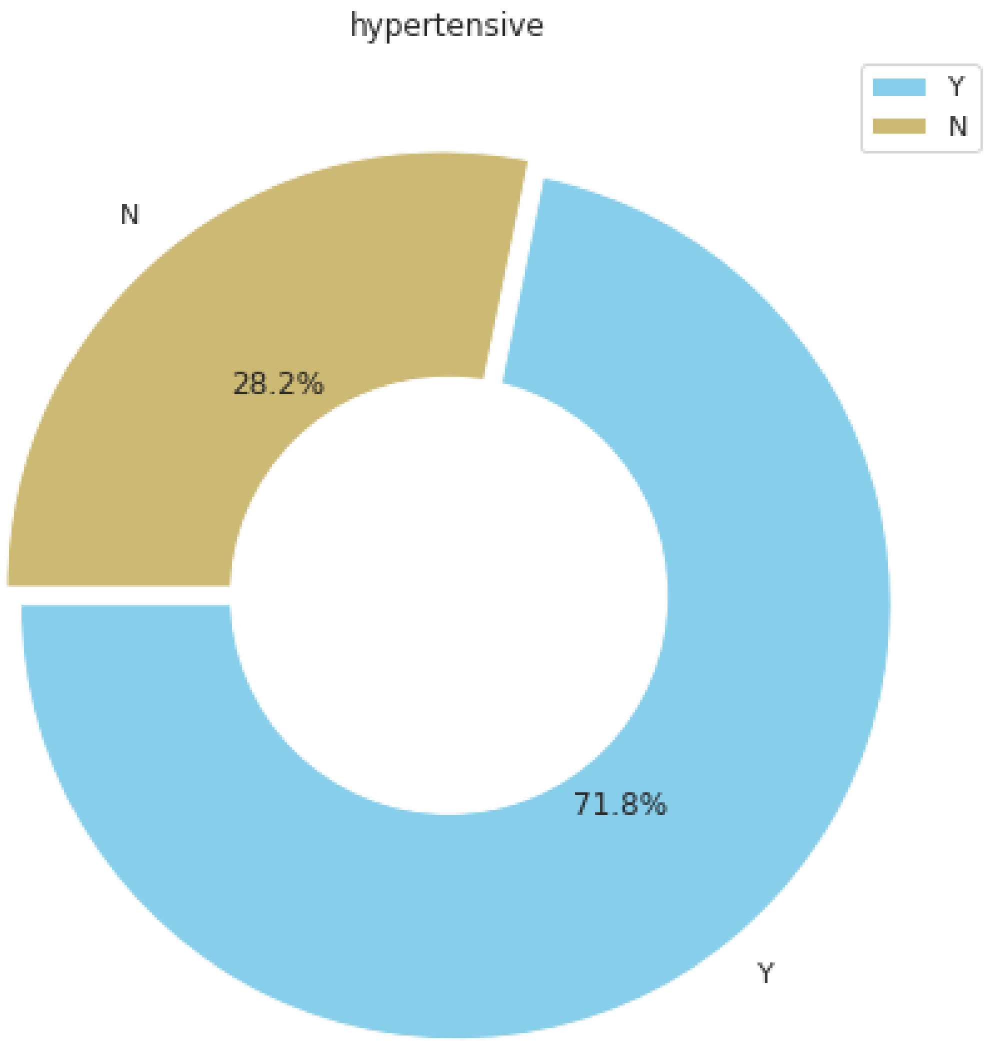


```
In [28]: plt.figure(figsize=(12,10))
         sns.countplot(df['diabetes'],hue=df['outcome'],palette="PuBu")
         plt.title("outcome by diabetes")
         plt.show()
```

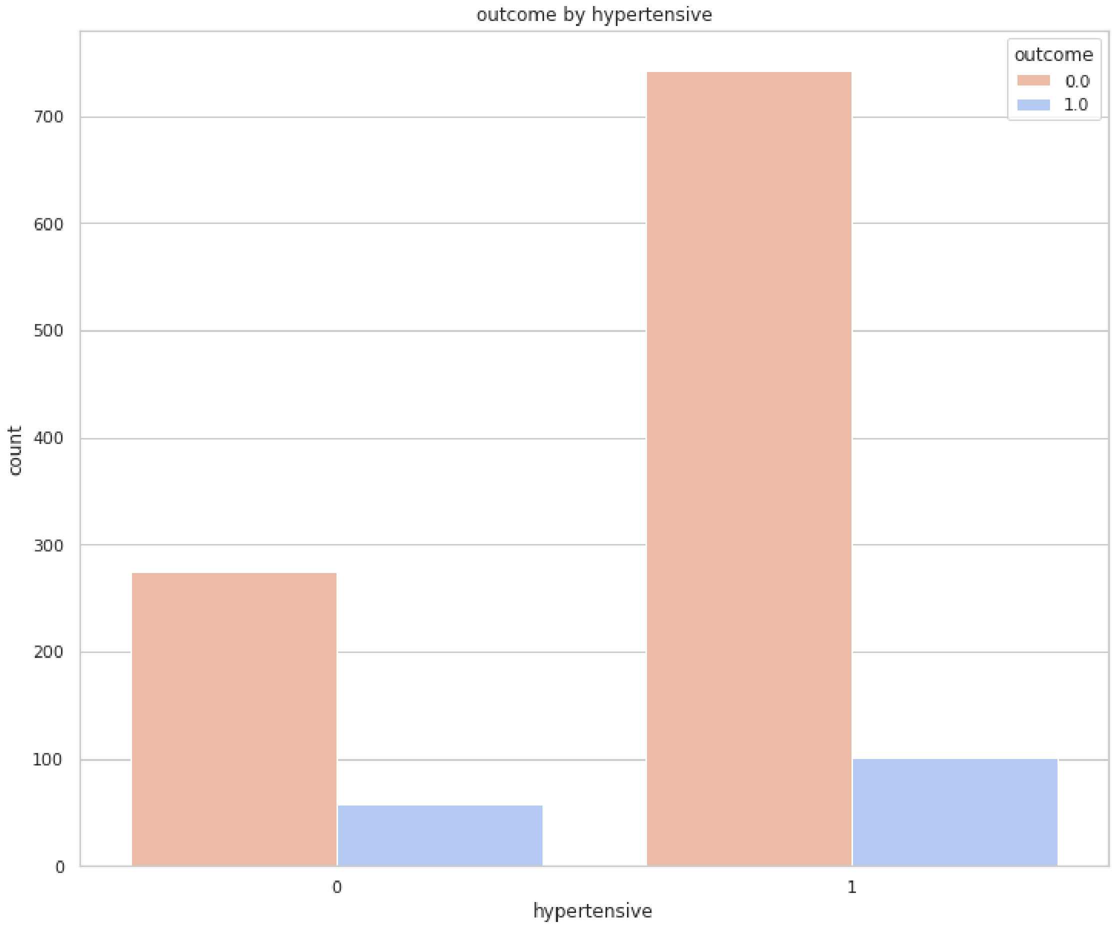
```
/data/dataiku/dss_data/code-envs/python/QB_HCP_propensity/lib/python3.6/site-packages/seaborn/_d
  FutureWarning
```



```
In [29]: plt.figure(figsize=(12,8))
         plt.title("hypertensive")
         circle = plt.Circle((0, 0), 0.5, color='white')
         g = plt.pie(df.hypertensive.value_counts(), explode=(0.025,0.025), labels=['Y','N'], co
         plt.legend()
         p = plt.gcf()
         p.gca().add_artist(circle)
         plt.show()
```

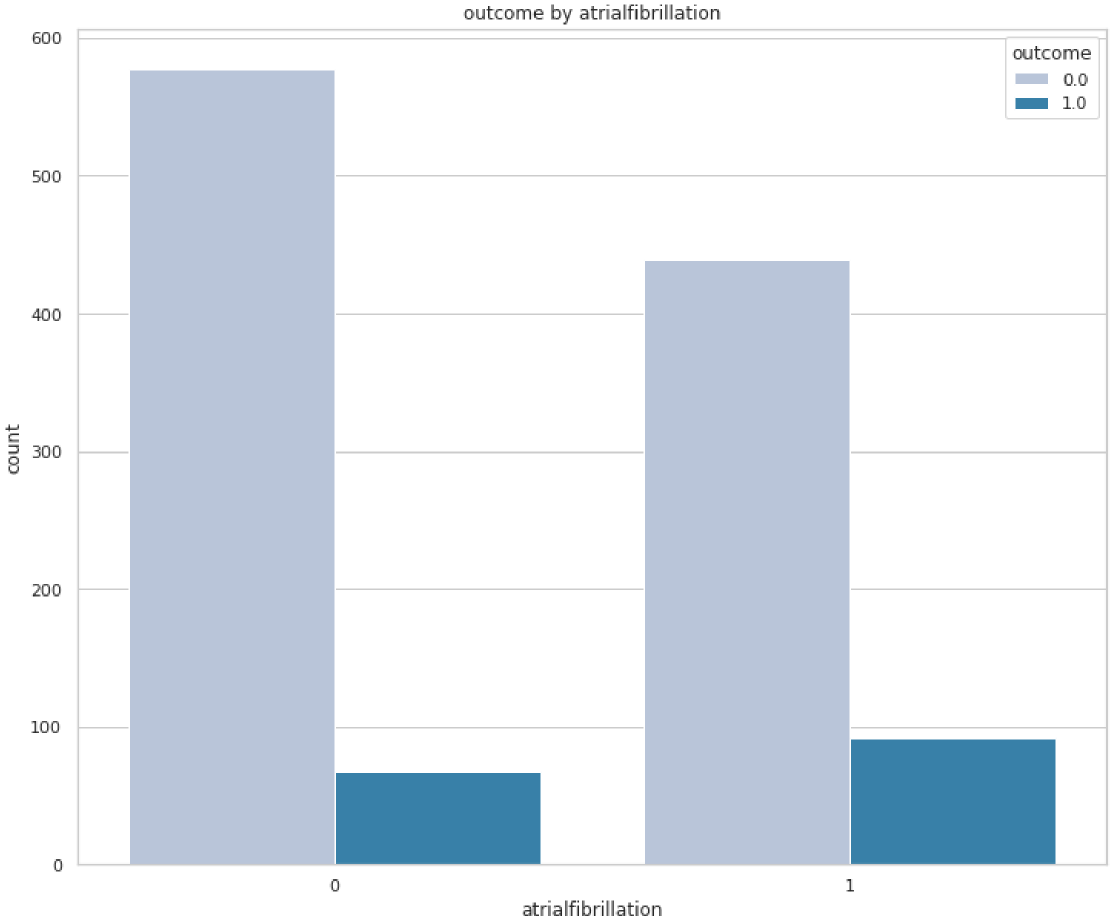


```
In [30]: plt.figure(figsize=(12,10))
         sns.countplot(df['hypertensive'],hue=df['outcome'],palette="coolwarm_r")
         plt.title("outcome by hypertensive")
         plt.show()

/data/dataiku/dss_data/code-envs/python/QB_HCP_propensity/lib/python3.6/site-packages/seaborn/_d
  FutureWarning
```

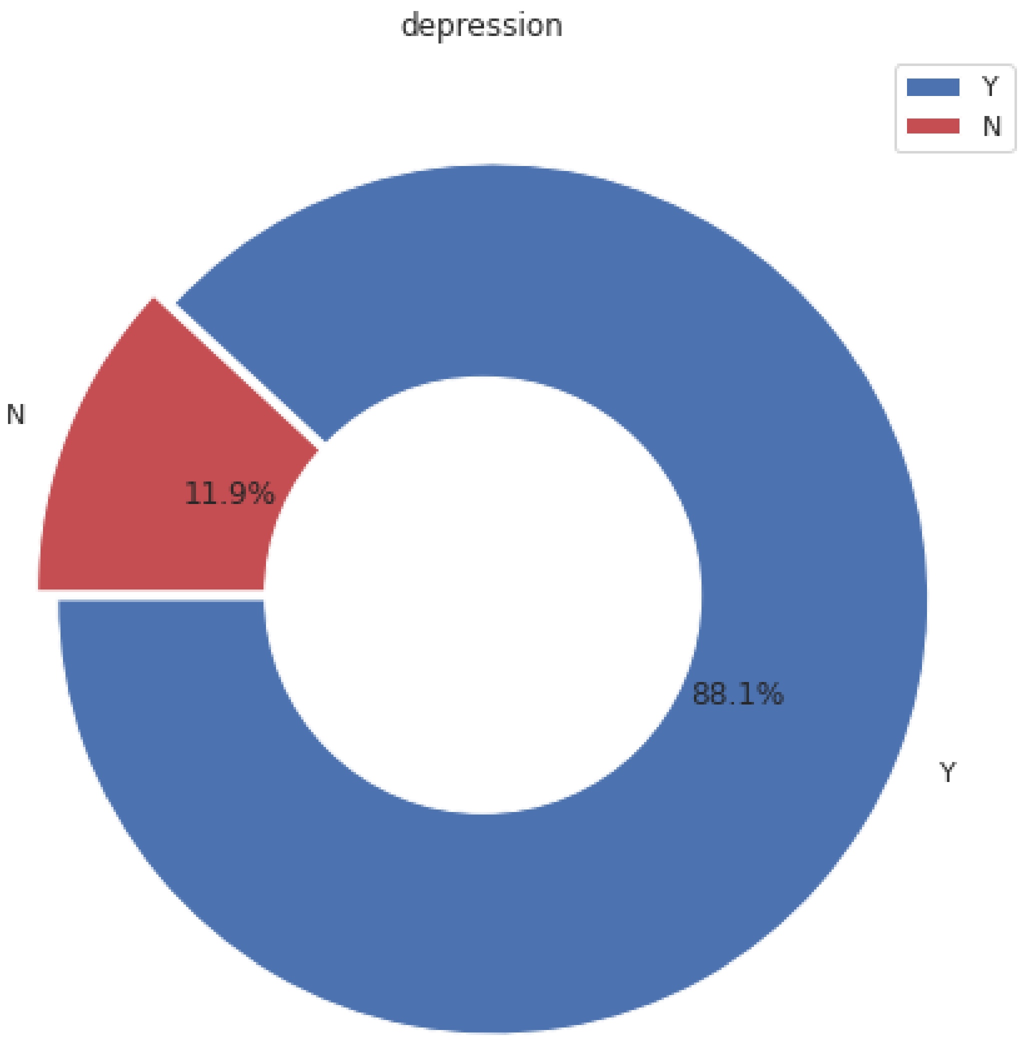


```
In [31]: plt.figure(figsize=(12,10))
         sns.countplot(df['atrialfibrillation'],hue=df['outcome'],palette="PuBu")
         plt.title("outcome by atrialfibrillation")
         plt.show()
```

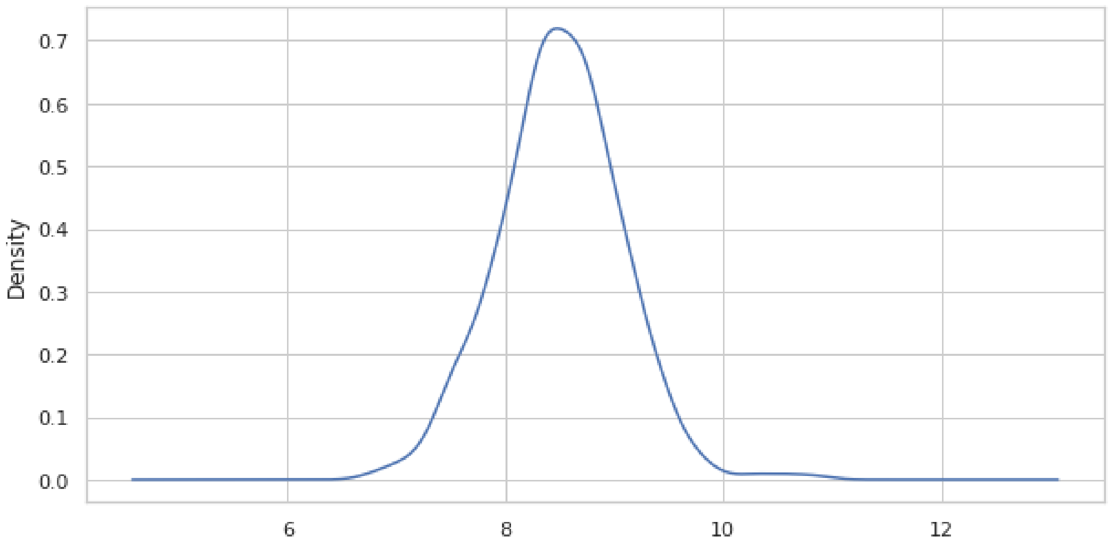
```
/data/dataiku/dss_data/code-envs/python/QB_HCP_propensity/lib/python3.6/site-packages/seaborn/_d
  FutureWarning
```



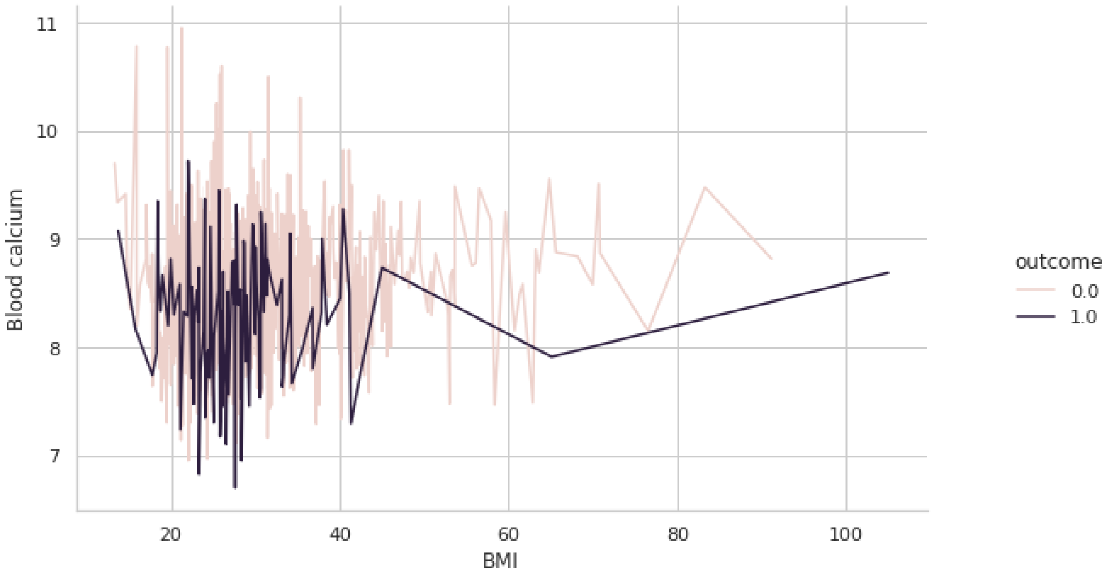
```
In [32]: plt.figure(figsize=(12,8))
         plt.title("depression")
         circle = plt.Circle((0, 0), 0.5, color='white')
         g = plt.pie(df.depression.value_counts(), explode=(0.025,0.025), labels=['Y','N'], colo
         plt.legend()
         p = plt.gcf()
         p.gca().add_artist(circle)
         plt.show()
```



```
In [33]: plt.figure(figsize=(10,5))
         df['Blood calcium'].plot(kind='kde')

Out[33]: <AxesSubplot:ylabel='Density'>
```



```
In [34]: from seaborn.relational import relplot
         f= sns.relplot(data=df, x="BMI", y="Blood calcium", hue="outcome",kind="line")
         f.fig.set_figwidth(10)
         f.fig.set_figheight(5)
```



```
In [35]: from seaborn.relational import relplot
         f= sns.relplot(data=df, x="heart rate", y="Blood calcium", hue="outcome",kind="line")
         f.fig.set_figwidth(10)
         f.fig.set_figheight(5)
```