ML LAB 3

Explore and implement logistic regression algorithm in a given business scenario and comment on its efficiency and performance.

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
from sklearn.preprocessing import PolynomialFeatures, StandardScaler
from warnings import filterwarnings
filterwarnings('ignore')
```

In [3]:
```
data = pd.read_csv('E:\DS\Datasets\drug200.csv')
```

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

Out[4]:

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	DrugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	DrugY

In [5]:
```
data.isnull().sum()
```

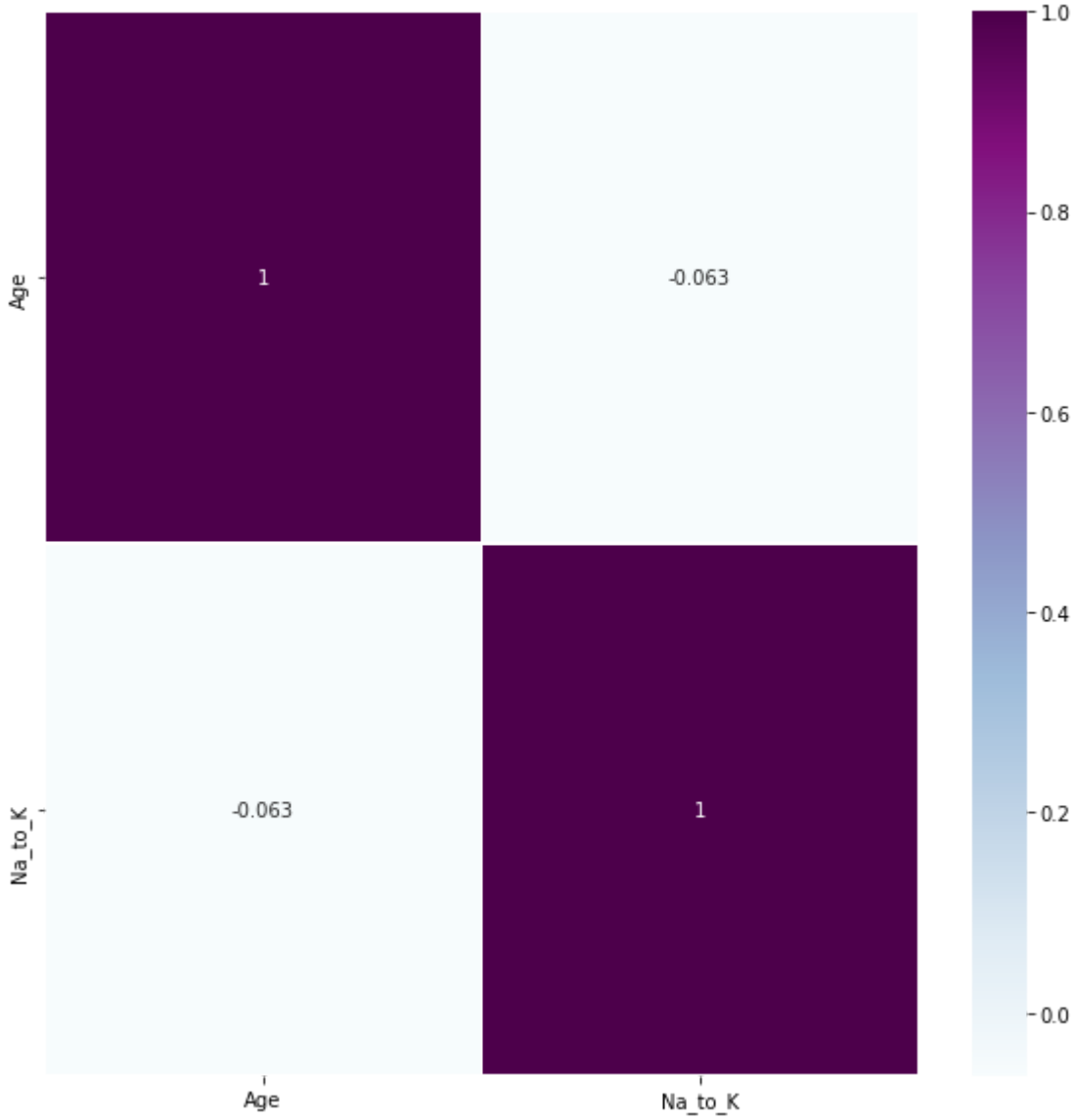
Out[5]:
```
Age            0
Sex            0
BP             0
Cholesterol    0
Na_to_K        0
Drug           0
dtype: int64
```

In [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
Age            200 non-null int64
Sex            200 non-null object
BP             200 non-null object
Cholesterol    200 non-null object
Na_to_K        200 non-null float64
Drug           200 non-null object
dtypes: float64(1), int64(1), object(4)
memory usage: 9.5+ KB
```

there is no missing values in the date we have 6 coulmns and 200 rows

In [7]:
```
fig, ax = plt.subplots(figsize  = (10, 10))
sns.heatmap(data.corr(), cmap = 'BuPu', cbar = True, linewidth = 0.5, annot = Tru
plt.show()
```

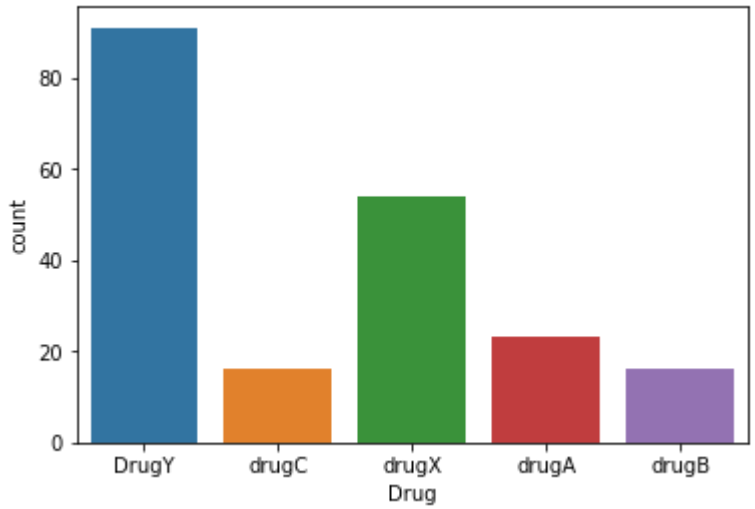


In [8]:
```
data['Drug'].value_counts()
```

Out[8]:
```
DrugY    91
drugX    54
drugA    23
drugC    16
drugB    16
Name: Drug, dtype: int64
```

In [9]:
```
sns.countplot(x = 'Drug', data= data)
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee10d437f0>
```

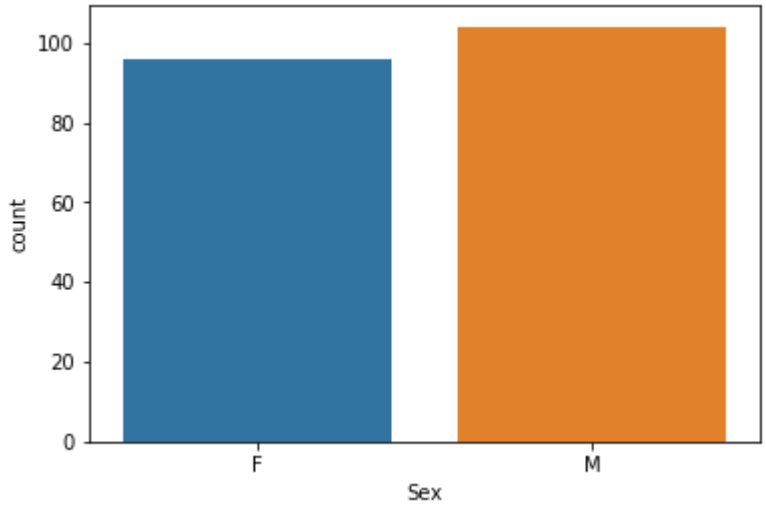


In [10]:
```
data['Sex'].value_counts()
```

Out[10]:
```
M    104
F     96
Name: Sex, dtype: int64
```

In [11]:
```
sns.countplot(x = 'Sex', data= data)
```

Out[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee10ebe400>
```

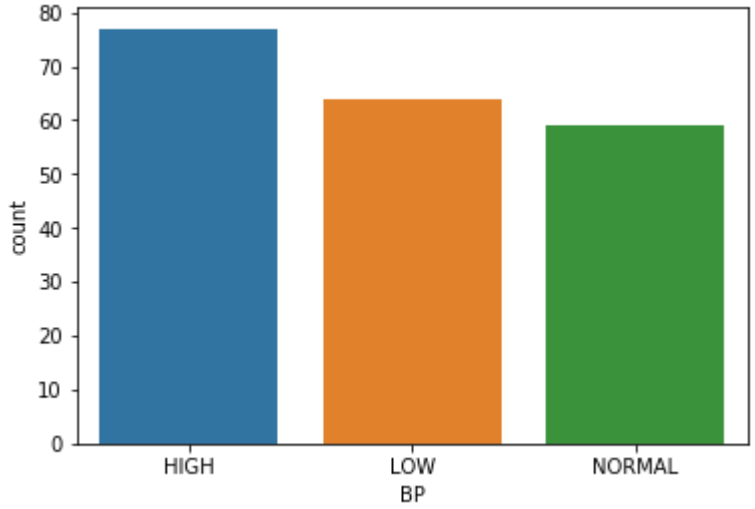


In [12]:
```
data['BP'].value_counts()
```

Out[12]:
```
HIGH      77
LOW       64
NORMAL    59
Name: BP, dtype: int64
```

In [13]:
```
sns.countplot(x = 'BP', data= data)
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee10ecfa58>
```

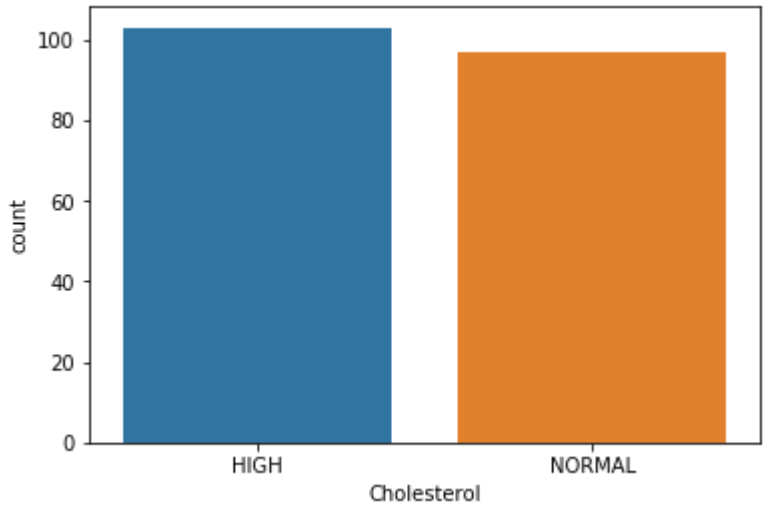


In [14]:
```
data['Cholesterol'].value_counts()
```

Out[14]:
```
HIGH      103
NORMAL     97
Name: Cholesterol, dtype: int64
```

In [15]:
```
sns.countplot(x = 'Cholesterol', data= data)
```

Out[15]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee10ebe1d0>
```



In [16]:
```
data['Na_to_K'].describe()
```

Out[16]:
```
count    200.000000
mean      16.084485
std        7.223956
min        6.269000
25%       10.445500
50%       13.936500
75%       19.380000
max       38.247000
Name: Na_to_K, dtype: float64
```

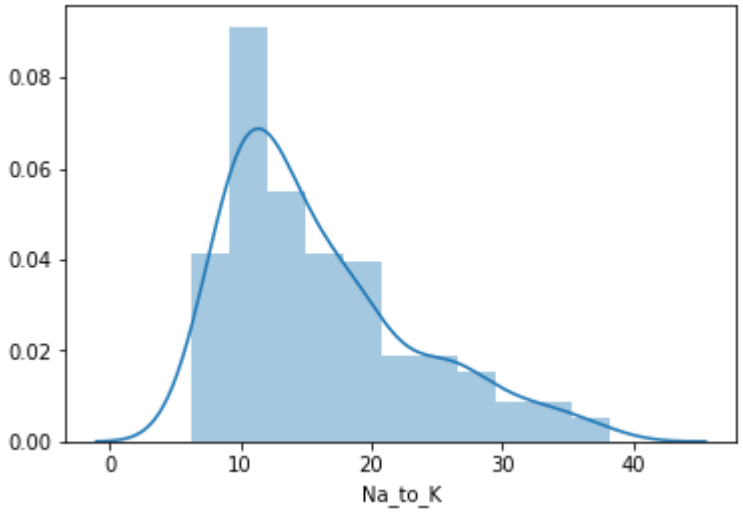
In [21]:
```
!pip install seaborn --upgrade
```

```
Requirement already satisfied: seaborn in c:\users\pranav\anaconda3\lib\site-pa
ckages (0.8.1)
Collecting seaborn
  Using cached seaborn-0.11.2-py3-none-any.whl (292 kB)
Collecting matplotlib>=2.2
  Using cached matplotlib-3.3.4-cp36-cp36m-win_amd64.whl (8.5 MB)
Requirement already satisfied: numpy>=1.15 in c:\users\pranav\anaconda3\lib\sit
e-packages (from seaborn) (1.19.5)
Requirement already satisfied: scipy>=1.0 in c:\users\pranav\anaconda3\lib\site
-packages (from seaborn) (1.5.4)
Collecting pandas>=0.23
  Using cached pandas-1.1.5-cp36-cp36m-win_amd64.whl (8.7 MB)
Requirement already satisfied: cycler>=0.10 in c:\users\pranav\anaconda3\lib\si
te-packages (from matplotlib>=2.2->seaborn) (0.10.0)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\pranav\anaconda
3\lib\site-packages (from matplotlib>=2.2->seaborn) (2.8.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\u
sers\pranav\anaconda3\lib\site-packages (from matplotlib>=2.2->seaborn) (2.4.7)
Requirement already satisfied: pillow>=6.2.0 in c:\users\pranav\anaconda3\lib\s
ite-packages (from matplotlib>=2.2->seaborn) (8.3.1)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\pranav\anaconda3\l
ib\site-packages (from matplotlib>=2.2->seaborn) (1.3.1)
Requirement already satisfied: six in c:\users\pranav\anaconda3\lib\site-packag
es (from cycler>=0.10->matplotlib>=2.2->seaborn) (1.11.0)
Requirement already satisfied: pytz>=2017.2 in c:\users\pranav\anaconda3\lib\si
te-packages (from pandas>=0.23->seaborn) (2017.3)
Installing collected packages: pandas, matplotlib, seaborn
  Attempting uninstall: matplotlib
    Found existing installation: matplotlib 2.1.2
    Uninstalling matplotlib-2.1.2:


WARNING: Value for scheme.headers does not match. Please report this to <http
s://github.com/pypa/pip/issues/9617>
distutils: c:\users\pranav\anaconda3\Include\UNKNOWN
sysconfig: c:\users\pranav\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
```

```
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
    WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anacond
a3\lib\site-packages)
    WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\li
b\site-packages)
    WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\li
b\site-packages)
ERROR: Could not install packages due to an OSError: [WinError 5] Access is den
ied: 'c:\\users\\pranav\\anaconda3\\lib\\site-packages\\matplotlib\\backends\\_
backend_agg.cp36-win_amd64.pyd'
Consider using the `--user` option or check the permissions.

WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: You are using pip version 21.1; however, version 21.3.1 is available.
You should consider upgrading via the 'c:\users\pranav\anaconda3\python.exe -m
pip install --upgrade pip' command.
```

In [23]:
```
sns.distplot(data['Na_to_K'])
```

Out[23]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee11365d68>
```



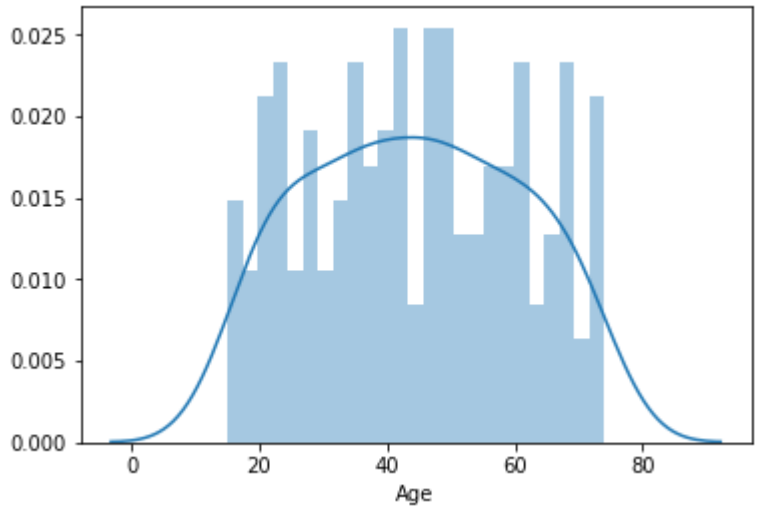
In [25]:
```
!pip install -U seaborn
```

```
Requirement already satisfied: seaborn in c:\users\pranav\anaconda3\lib\site-pa
ckages (0.8.1)
Collecting seaborn
  Using cached seaborn-0.11.2-py3-none-any.whl (292 kB)
Requirement already satisfied: numpy>=1.15 in c:\users\pranav\anaconda3\lib\sit
e-packages (from seaborn) (1.19.5)
Collecting matplotlib>=2.2
  Using cached matplotlib-3.3.4-cp36-cp36m-win_amd64.whl (8.5 MB)
Requirement already satisfied: scipy>=1.0 in c:\users\pranav\anaconda3\lib\site
-packages (from seaborn) (1.5.4)
Requirement already satisfied: pandas>=0.23 in c:\users\pranav\anaconda3\lib\si
te-packages (from seaborn) (1.1.5)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\u
sers\pranav\anaconda3\lib\site-packages (from matplotlib>=2.2->seaborn) (2.4.7)
Requirement already satisfied: cycler>=0.10 in c:\users\pranav\anaconda3\lib\si
te-packages (from matplotlib>=2.2->seaborn) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\pranav\anaconda3\l
ib\site-packages (from matplotlib>=2.2->seaborn) (1.3.1)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\pranav\anaconda
3\lib\site-packages (from matplotlib>=2.2->seaborn) (2.8.2)
Requirement already satisfied: pillow>=6.2.0 in c:\users\pranav\anaconda3\lib\s
ite-packages (from matplotlib>=2.2->seaborn) (8.3.1)
Requirement already satisfied: six in c:\users\pranav\anaconda3\lib\site-packag
es (from cycler>=0.10->matplotlib>=2.2->seaborn) (1.11.0)
Requirement already satisfied: pytz>=2017.2 in c:\users\pranav\anaconda3\lib\si
te-packages (from pandas>=0.23->seaborn) (2017.3)
Installing collected packages: matplotlib, seaborn
  Attempting uninstall: matplotlib
    Found existing installation: matplotlib 2.1.2
    Uninstalling matplotlib-2.1.2:

WARNING: Value for scheme.headers does not match. Please report this to <http
s://github.com/pypa/pip/issues/9617>
distutils: c:\users\pranav\anaconda3\Include\UNKNOWN
sysconfig: c:\users\pranav\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
```

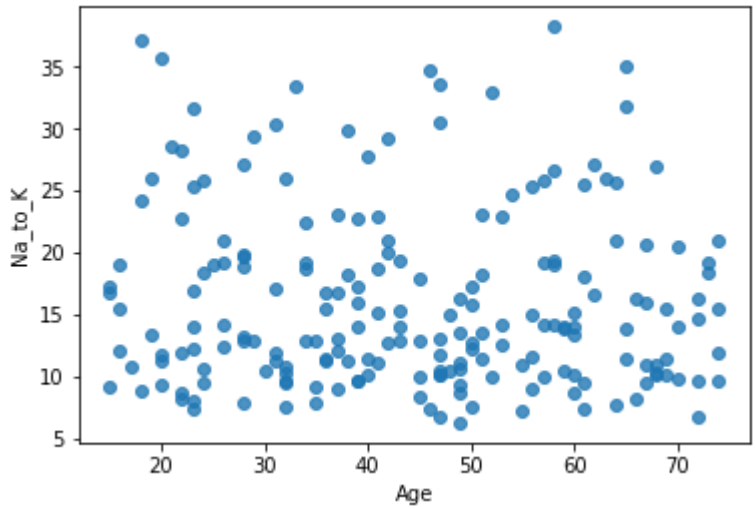
```
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
    WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anacond
a3\lib\site-packages)
    WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\li
b\site-packages)
    WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\li
b\site-packages)
ERROR: Could not install packages due to an OSError: [WinError 5] Access is den
ied: 'c:\\users\\pranav\\anaconda3\\lib\\site-packages\\matplotlib\\backends\\_
backend_agg.cp36-win_amd64.pyd'
Consider using the `--user` option or check the permissions.

WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -tatsmodels (c:\users\pranav\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: Ignoring invalid distribution -andas (c:\users\pranav\anaconda3\lib\si
te-packages)
WARNING: You are using pip version 21.1; however, version 21.3.1 is available.
You should consider upgrading via the 'c:\users\pranav\anaconda3\python.exe -m
pip install --upgrade pip' command.
```

In [32]:
```
sns.distplot(data['Age'], hist=True,kde=True, bins = 25)
```

Out[32]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee11455f98>
```



In [42]:
```
sns.regplot(x=data["Age"], y=data["Na_to_K"], fit_reg=False, scatter=True);
```

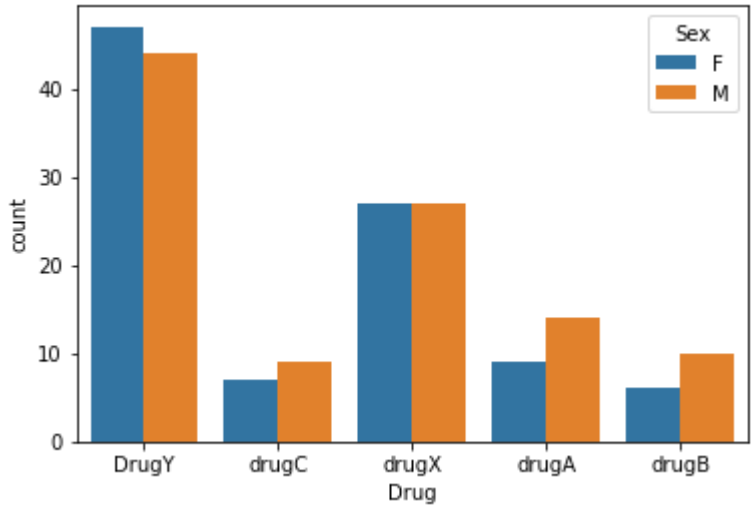


In [43]:
```
data_sex_drug = data.groupby(['Drug','Sex']).size().reset_index(name = 'count')
print(data_sex_drug)
```

```
    Drug Sex  count
0  DrugY   F     47
1  DrugY   M     44
2  drugA   F      9
3  drugA   M     14
4  drugB   F      6
5  drugB   M     10
6  drugC   F      7
7  drugC   M      9
8  drugX   F     27
9  drugX   M     27
```

In [44]:
```
sns.countplot(x = 'Drug', data= data, hue = 'Sex')
```

Out[44]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee123dc2b0>
```

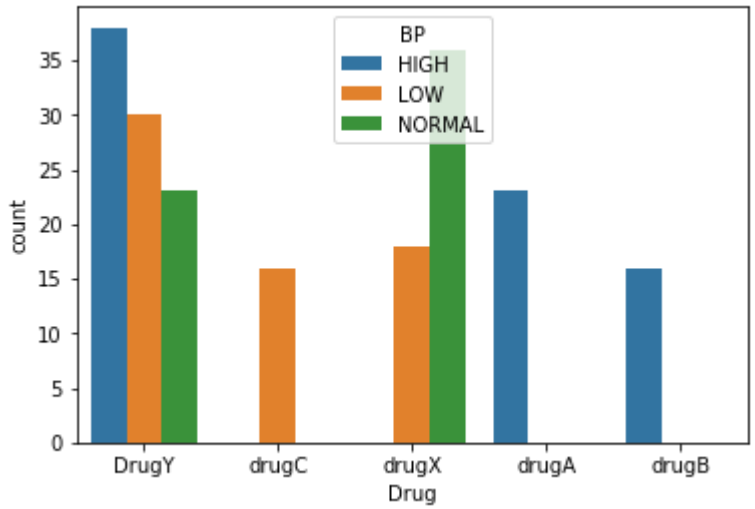


In [45]:
```
data_BP_drug = data.groupby(['Drug','BP']).size().reset_index(name = 'count')
print(data_BP_drug)
```

```
    Drug      BP  count
0  DrugY    HIGH     38
1  DrugY     LOW     30
2  DrugY  NORMAL     23
3  drugA    HIGH     23
4  drugB    HIGH     16
5  drugC     LOW     16
6  drugX     LOW     18
7  drugX  NORMAL     36
```

In [46]:
```
sns.countplot(x = 'Drug', data= data, hue = 'BP')
```

Out[46]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee12417080>
```

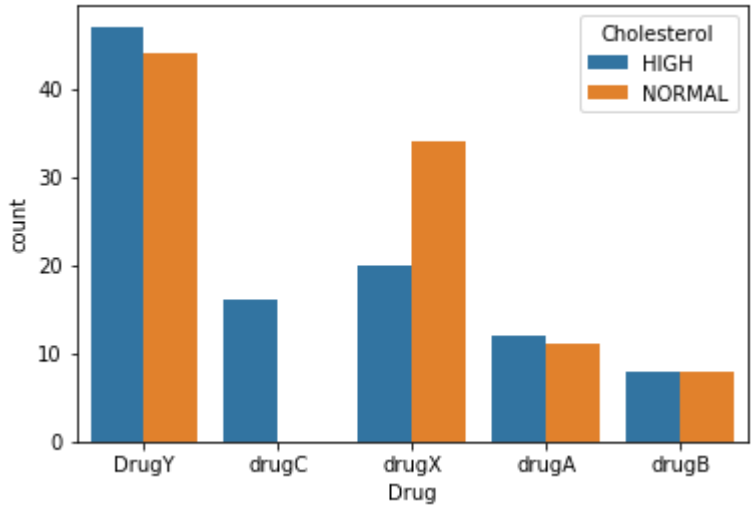


In [47]:
```
data_Cholesterol_drug = data.groupby(['Drug','Cholesterol']).size().reset_index(n
print(data_Cholesterol_drug)
```

```
    Drug Cholesterol  count
0  DrugY        HIGH     47
1  DrugY      NORMAL     44
2  drugA        HIGH     12
3  drugA      NORMAL     11
4  drugB        HIGH      8
5  drugB      NORMAL      8
6  drugC        HIGH     16
7  drugX        HIGH     20
8  drugX      NORMAL     34
```

In [48]:
```
sns.countplot(x = 'Drug', data= data, hue = 'Cholesterol')
```

Out[48]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee115854e0>
```



In [49]:
```
data['Sex'] = data['Sex'].map({'M': 1, 'F': 0})
data['Cholesterol'] = data['Cholesterol'].map({'HIGH' : 1, 'NORMAL' : 0})
data['Drug'] = data['Drug'].map({'DrugY':1, 'drugC':2, 'drugX':3, 'drugA':4, 'dru
data.head()
```

Out[49]:

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	0	HIGH	1	25.355	1
1	47	1	LOW	1	13.093	2
2	47	1	LOW	1	10.114	2
3	28	0	NORMAL	1	7.798	3
4	61	0	LOW	1	18.043	1

In [50]:
```
data.shape
```

Out[50]:
```
(200, 6)
```

In [51]:
```
data = pd.get_dummies(data)
data.head()
```

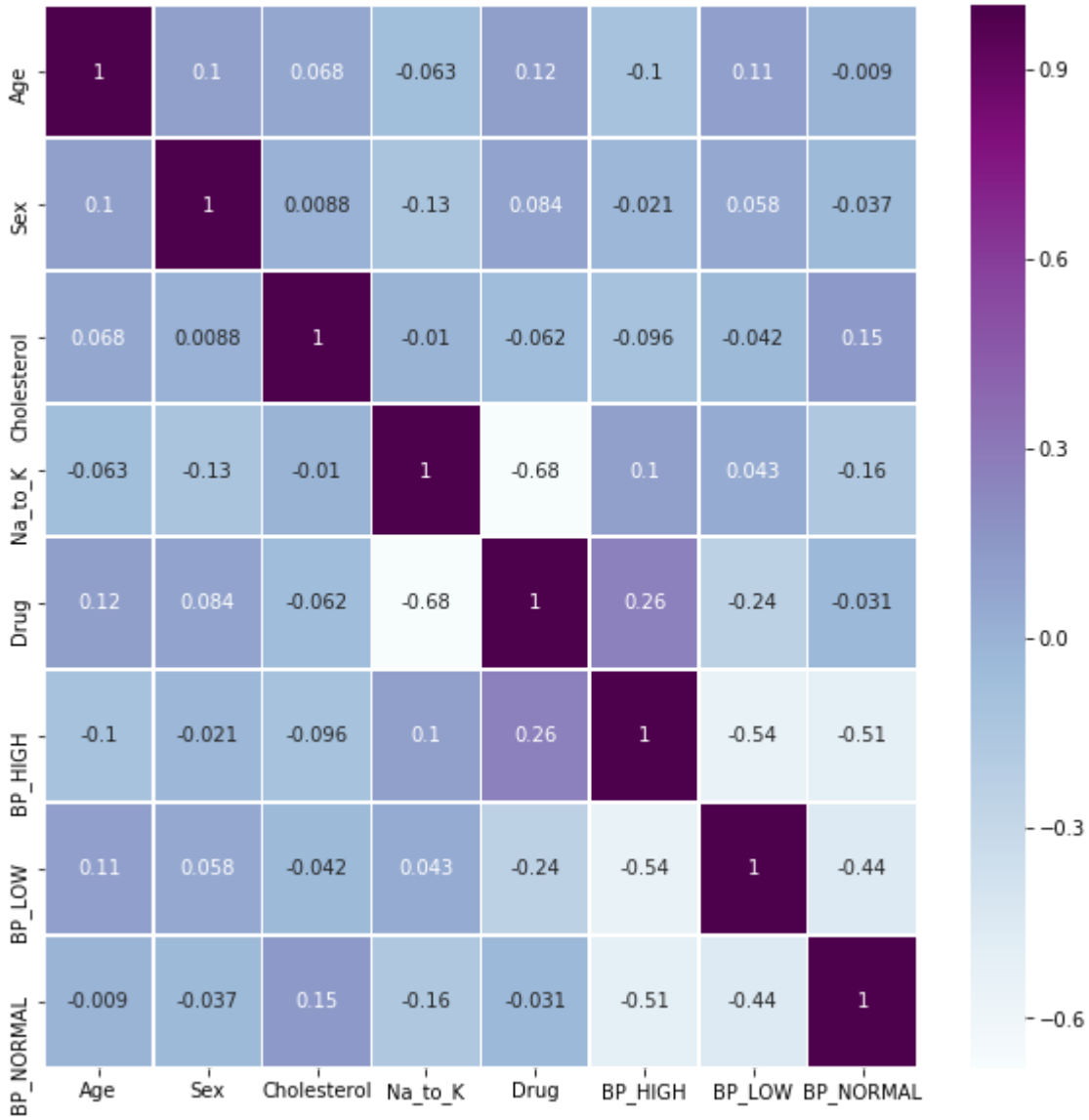
Out[51]:

	Age	Sex	Cholesterol	Na_to_K	Drug	BP_HIGH	BP_LOW	BP_NORMAL
0	23	0	1	25.355	1	1	0	0
1	47	1	1	13.093	2	0	1	0
2	47	1	1	10.114	2	0	1	0
3	28	0	1	7.798	3	0	0	1
4	61	0	1	18.043	1	0	1	0

In [52]:
```
data.shape
```

Out[52]:
```
(200, 8)
```

In [53]:
```
fig, ax = plt.subplots(figsize  = (10, 10))
sns.heatmap(data.corr(), cmap = 'BuPu', cbar = True, linewidth = 0.5, annot = Tru
plt.show()
```



In [54]:
```
X = data.drop('Drug', axis = 1).values
y = data['Drug'].values.reshape((-1,1))
```

In [55]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_
print('x train shape {}'.format(X_train.shape))
print('x test shape  {}'.format(X_test.shape))
print('y train shape {}'.format(y_train.shape))
print('y test shape  {}'.format(y_test.shape))
```

```
x train shape (160, 7)
x test shape  (40, 7)
y train shape (160, 1)
y test shape  (40, 1)
```

In [56]:
```
from sklearn.linear_model import LogisticRegression

logistic_model = LogisticRegression(C = 2 ,solver = 'liblinear', tol = .001)
```

In [58]:
```
from sklearn.metrics import confusion_matrix, accuracy_score, classification_repo
```

In [64]:
```
logistic_model.fit(X_train, y_train)
y_pred = logistic_model.predict(X_test)
print(logistic_model.score(X_train,y_train)*100)
logistic_score = accuracy_score(y_test, y_pred)
print(logistic_score*100)
```

```
97.5
95.0
```

In [65]:
```
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[17  0  0  0  0]
 [ 0  4  0  0  0]
 [ 0  0 13  0  0]
 [ 1  0  0  2  1]
 [ 0  0  0  0  2]]
             precision    recall  f1-score   support

          1       0.94      1.00      0.97        17
          2       1.00      1.00      1.00         4
          3       1.00      1.00      1.00        13
          4       1.00      0.50      0.67         4
          5       0.67      1.00      0.80         2

avg / total       0.96      0.95      0.94        40
```

Interpretation:

The drugs have been classified using Logistic Regression, with 95% accuracy.

In [ ]:
```
```