Achraf Safsafi

DSC 540

Project_Milestone2

In [1]:
```
import pandas as pd
import numpy as np
# import and read data
data = 'Table_10_Offenses_Known_to_Law_Enforcement_by_State_by_Metropolitan_and_Nonmetropoli
tan_Counties_2018.xls'
df = pd.read_excel(data, header=0, index_col=False, keep_default_na=True)

df.head()
```

Out[1]:

	State	County	Violent\ncrime	Murder and\nnonnegligent\nmanslaughter	Rape	Robbery	Aggravated\nassault	Property\ncrime
0	ALABAMA	Autauga	51.0	0.0	6.0	5.0	40.0	372.0
1	ALABAMA	Baldwin	223.0	0.0	9.0	37.0	177.0	615.0
2	ALABAMA	Blount	375.0	1.0	19.0	5.0	350.0	796.0
3	ALABAMA	Calhoun	14.0	0.0	5.0	7.0	2.0	144.0
4	ALABAMA	Elmore	68.0	4.0	30.0	12.0	22.0	669.0

In [2]:
```
# data frame shape
df.shape
```

Out[2]:
```
(2356, 12)
```

In [3]:
```
#get column names
list(df.columns)
```

Out[3]:
```
['State',
 'County',
 'Violent\ncrime',
 'Murder and\nnonnegligent\nmanslaughter',
 'Rape',
 'Robbery',
 'Aggravated\nassault',
 'Property\ncrime',
 'Burglary',
 'Larceny-\ntheft',
 'Motor\nvehicle\ntheft',
 'Arson']
```

In [4]:
```
#rename column headers
df.columns = ['State', 'County', 'Violent_crime', 'Murder',
              'Rape','Robbery','Aggravated_assault','Property_crime',
             'Burglary','Larceny_theft','Motor_vehicle_theft','Arson']
df.head()
```

Out[4]:

	State	County	Violent_crime	Murder	Rape	Robbery	Aggravated_assault	Property_crime	Burglary	Larceny_theft	M
0	ALABAMA	Autauga	51.0	0.0	6.0	5.0	40.0	372.0	92.0	240	
1	ALABAMA	Baldwin	223.0	0.0	9.0	37.0	177.0	615.0	173.0	397	
2	ALABAMA	Blount	375.0	1.0	19.0	5.0	350.0	796.0	191.0	492	
3	ALABAMA	Calhoun	14.0	0.0	5.0	7.0	2.0	144.0	49.0	95	
4	ALABAMA	Elmore	68.0	4.0	30.0	12.0	22.0	669.0	178.0	427	

In [5]:
```
#Check the data type
df.dtypes
```

Out[5]:
```
State                   object
County                  object
Violent_crime          float64
Murder                 float64
Rape                   float64
Robbery                float64
Aggravated_assault     float64
Property_crime         float64
Burglary               float64
Larceny_theft            int64
Motor_vehicle_theft    float64
Arson                  float64
dtype: object
```

In [6]:
```
#remove whitespace from the beginning and end
df.columns = [x.strip() for x in df.columns]
```

In [7]:
```
# get a statistical summary of the data
df.describe()
```

Out[7]:

	Violent_crime	Murder	Rape	Robbery	Aggravated_assault	Property_crime	Burglary	Larceny_theft
count	2296.000000	2355.000000	2301.000000	2355.000000	2353.000000	2348.000000	2349.000000	2356.000000
mean	83.155052	1.179618	11.398522	12.633121	59.467063	502.102215	108.122605	343.252971
std	297.490125	4.676039	32.092133	85.625454	198.717364	1800.823555	307.080632	1311.923551
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	5.000000	0.000000	0.000000	0.000000	4.000000	39.000000	11.000000	22.000000
50%	18.500000	0.000000	3.000000	1.000000	14.000000	123.000000	33.000000	76.000000
75%	54.000000	1.000000	9.000000	3.000000	42.000000	332.250000	93.000000	209.250000
max	5790.000000	86.000000	591.000000	2271.000000	3969.000000	41708.000000	7640.000000	28113.000000

In [8]:
```
#Checking for missing values
df.isnull().sum()
```

Out[8]:
```
State                    0
County                   0
Violent_crime           60
Murder                   1
Rape                    55
Robbery                  1
Aggravated_assault       3
Property_crime           8
Burglary                 7
Larceny_theft            0
Motor_vehicle_theft      1
Arson                  123
dtype: int64
```

In [9]:
```
#dealing with missing data using mean
df=df.fillna(df.mean().apply(np.floor))
df
```

Out[9]:

	State	County	Violent_crime	Murder	Rape	Robbery	Aggravated_assault	Property_crime	Burglary	Larceny_tl
0	ALABAMA	Autauga	51.0	0.0	6.0	5.0	40.0	372.0	92.0	
1	ALABAMA	Baldwin	223.0	0.0	9.0	37.0	177.0	615.0	173.0	
2	ALABAMA	Blount	375.0	1.0	19.0	5.0	350.0	796.0	191.0	
3	ALABAMA	Calhoun	14.0	0.0	5.0	7.0	2.0	144.0	49.0	
4	ALABAMA	Elmore	68.0	4.0	30.0	12.0	22.0	669.0	178.0	
...	...	...	...	...	...	...	...	...	...	
2351	WYOMING	Sublette	4.0	0.0	0.0	0.0	4.0	49.0	11.0	
2352	WYOMING	Sweetwater	22.0	0.0	8.0	0.0	14.0	77.0	17.0	
2353	WYOMING	Uinta	7.0	1.0	1.0	1.0	4.0	53.0	12.0	
2354	WYOMING	Washakie	0.0	0.0	0.0	0.0	0.0	17.0	4.0	
2355	WYOMING	Weston	5.0	0.0	1.0	0.0	4.0	0.0	0.0	

2356 rows × 12 columns

In [10]:
```
##Checking  again for missing values
df.isnull().sum()
```

Out[10]:
```
State                  0
County                 0
Violent_crime          0
Murder                 0
Rape                   0
Robbery                0
Aggravated_assault     0
Property_crime         0
Burglary               0
Larceny_theft          0
Motor_vehicle_theft    0
Arson                  0
dtype: int64
```

In [11]:
```
#check outliers for Violent_crime col.
outliers_Violent_crime = df[df['Violent_crime'] > df['Violent_crime'].mean() + 3 * df['Viole
nt_crime'].std()]
outliers_Violent_crime.shape
```

Out[11]:
```
(32, 12)
```

In [12]:
```
#check outlier for Murder col.
outliers_Murder = df[df['Murder'] > df['Murder'].mean() + 3 * df['Murder'].std()]
outliers_Murder.shape
```

Out[12]:
```
(28, 12)
```

In [13]:
```
#check outlier for Rape col.
outliers_Rape = df[df['Rape'] > df['Rape'].mean() + 3 * df['Rape'].std()]
outliers_Rape.shape
```

Out[13]:
```
(46, 12)
```

In [14]:
```
#check outlier for Robbery col.
outliers_Robbery = df[df['Robbery'] > df['Robbery'].mean() + 3 * df['Robbery'].std()]
outliers_Robbery.shape
```

Out[14]:
```
(23, 12)
```

In [15]:
```
#check outlier for Aggravated_assault col.
outliers_Aggravated_assault = df[df['Aggravated_assault'] > df['Aggravated_assault'].mean()
+ 3 * df['Aggravated_assault'].std()]
outliers_Aggravated_assault.shape
```

Out[15]:
```
(31, 12)
```

In [16]:
```
#check outlier for Property_crime col.
outliers_Property_crime = df[df['Property_crime'] > df['Property_crime'].mean() + 3 * df['Pr
operty_crime'].std()]
outliers_Property_crime.shape
```

Out[16]:
```
(31, 12)
```

In [17]:
```
#check outlier for Burglary col.
outliers_Burglary = df[df['Burglary'] > df['Burglary'].mean() + 3 * df['Burglary'].std()]
outliers_Burglary.shape
```

Out[17]:
```
(35, 12)
```

In [18]:
```
#check outlier for Larceny_theft col.
outliers_Larceny_theft = df[df['Larceny_theft'] > df['Larceny_theft'].mean() + 3 * df['Larce
ny_theft'].std()]
outliers_Larceny_theft.shape
```

Out[18]:
```
(32, 12)
```

In [19]:
```
#check outlier for Motor_vehicle_theft col.
outliers_Motor_vehicle_theft = df[df['Motor_vehicle_theft'] > df['Motor_vehicle_theft'].mean
() + 3 * df['Motor_vehicle_theft'].std()]
outliers_Motor_vehicle_theft.shape
```

Out[19]:
```
(26, 12)
```

In [20]:
```
#check outlier for Arson col.
outliers_Arson = df[df['Arson'] > df['Arson'].mean() + 3 * df['Arson'].std()]
outliers_Arson.shape
```

Out[20]:
```
(27, 12)
```