Suman Gurung

CIS335- Data mining

1. Go to https://www.kaggle.com/datasets
   The data set I found in Kaggle is Crime.csv
2. Download a dataset to your liking (make sure the dataset can be used for a classification task for the later assignments) (5)
   I used PyCharm to and panda to implement the tasks.
3. Identify the different types of attributes (numeric, nominal etc.) (15)
   I dataset has 9 attributes and 550 rows.

```
Attributes are: State,City,Year,Population,Violent Crime,Murder,Rape,Robbery
```

4. Identify the mean,median,mode of the numeric attributes, if any. Also identify which attributes are positively/negatively skewed. (do for at least three numeric attributes) (20)

```
------------------Mean, Median, Mode Robbery Attributes Robbery-----------
mean for the attribute Robbery :  266.3157894736842
median for the attribute Robbery :  85.0
mode for the attribute Property Robbery :  0    44
dtype: int64

------------------Mean, Median, Mode Robbery Attributes Murder------------
mean for the attribute Murder :  12.88021778584392
median for the attribute Murder :  4.0
mode for the attribute Property Murder :  0    1
dtype: int64

------------------Mean, Median, Mode Robbery Attributes Rape--------------
mean for the attribute Robbery :  88.72140221402213
median for the attribute Robbery :  39.5
mode for the attribute Property Robbery :  0    24.0
dtype: float64
```

5. Compute the IQR for the numerical attributes. Based on the IQR, determine the outliers. Decide whether in the context of the dataset, if you are going to keep the outliers or not. Provide arguments. (do for at least three numeric attributes) (20)
   First I figure out percentile of one attribute and compute IQR in python

6. Use scatter plots to determine if there's correlation between the numeric attributes (do for at least two pairs of numeric attributes) (20) 7. Finally, draw boxplots for the numeric attributes (do for at least three numeric attributes)

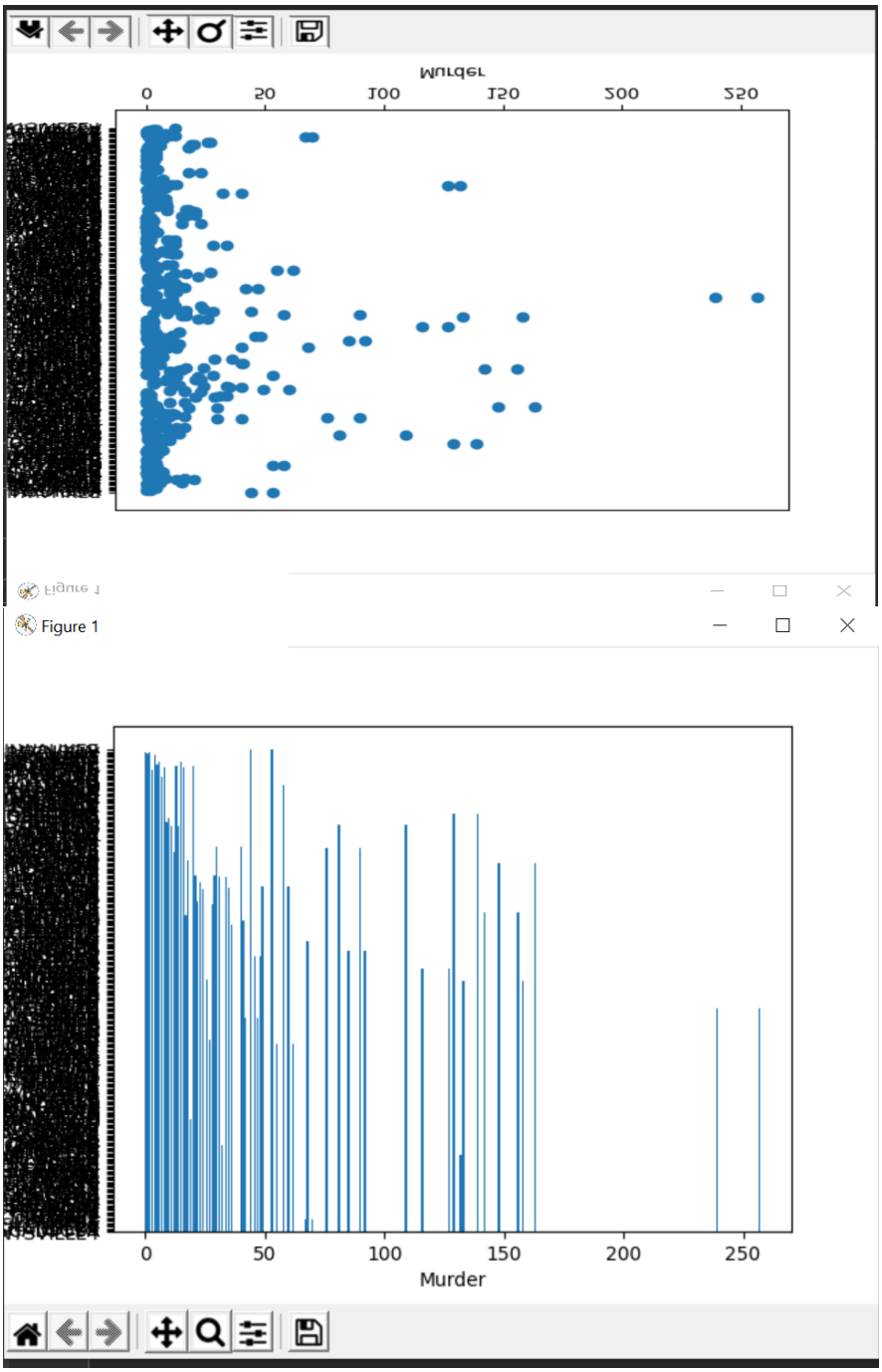
Figure 1
0
50
100
150
200
250
Murder