EP219: Data Analysis and Interpretation

Assignment Report 3

By Team: *Sigificantly Different*

October/01/2018 - October/08/2018

Contents

Problem Statement

Our aim is to extract data about Crime Rate (C) and Unemployment Rate (U) in different States and union teritories in year 2016.

- Find the mean and standard deviations of this data.
- Make 1D histogram of unemployment rate and crime rate and mark the mean and standard deviation on plot.
- Make the scatter plot of the pairs (U_i, C_i).
- Make 2D histogram of the pairs (U_i, C_i).
- Proof of correlation formula between two samples.
- Find estimated correlation coefficient.
- Make the Conclusion about the correlation coefficient.

Proof of Correlation Formula

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{2}$$

Now,

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N}((X_i - \mu) + (\mu - \bar{X}))((Y_i - \nu) + (\nu - \bar{Y}))$$

Where μ and ν are true mean of X and Y
So,

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N}((X_i - \mu) + (\mu - \bar{X}))((Y_i - \nu) + (\nu - \bar{Y}))$$

$$\begin{aligned}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = &\sum_{i=1}^{N}(X_i - \mu)(Y_i - \nu) \\ &+ \sum_{i=1}^{N}(X_i - \mu)(\nu - \bar{Y}) \\ &+ \sum_{i=1}^{N}(\mu - \bar{X})(Y_i - \nu) \\ &+ \sum_{i=1}^{N}(\mu - \bar{X})(\nu - \bar{Y})\end{aligned}$$

Now, Expected value of the above first term is:

$$\sum_{i=1}^{N} E((X_i - \mu)(Y_i - \nu)) = \sum_{i=1}^{N} Cov(X_i, Y_i) = NCov(X, Y)$$

And Expectation value of second and third terms are same and is equal to:

$$\sum_{i=1}^{N} E(X_i - \mu)(\nu - \bar{Y}) = -\sum_{i=1}^{N} Cov(X_i, \bar{Y})$$
$$= -\sum_{i=1}^{N} Cov(X_i, \frac{\sum_i Y_i}{N})$$
$$= -NCov(X_1, \frac{\sum_i Y_i}{N})$$
$$= -Cov(X_1, \sum_i (Y_i))$$
$$= -Cov(X, Y)$$

And Expectation value of forth term is:

$$\sum_{i=1}^{N} E(\mu - \bar{X})(\nu - \bar{Y}) = \sum_{i=1}^{N} Cov(\bar{X}, \bar{Y})$$
$$= NCov(\bar{X}, \bar{Y})$$
$$= NCov(\frac{1}{N}\sum_{i=1}^{N} X_i, \frac{1}{N}\sum_{i=1}^{N} Y_i)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} Cov(X_i, Y_j)$$
$$= \frac{1}{N}\sum_{i=1}^{N} Cov(X_i, Y_i)$$
$$= Cov(X, Y)$$

Covariance of all terms wiil be zero, except those with i = j
Therefore, on adding all four of them we will get,

$$(N-1)Cov(X, Y) = \sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$
$$Cov(X, Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Code

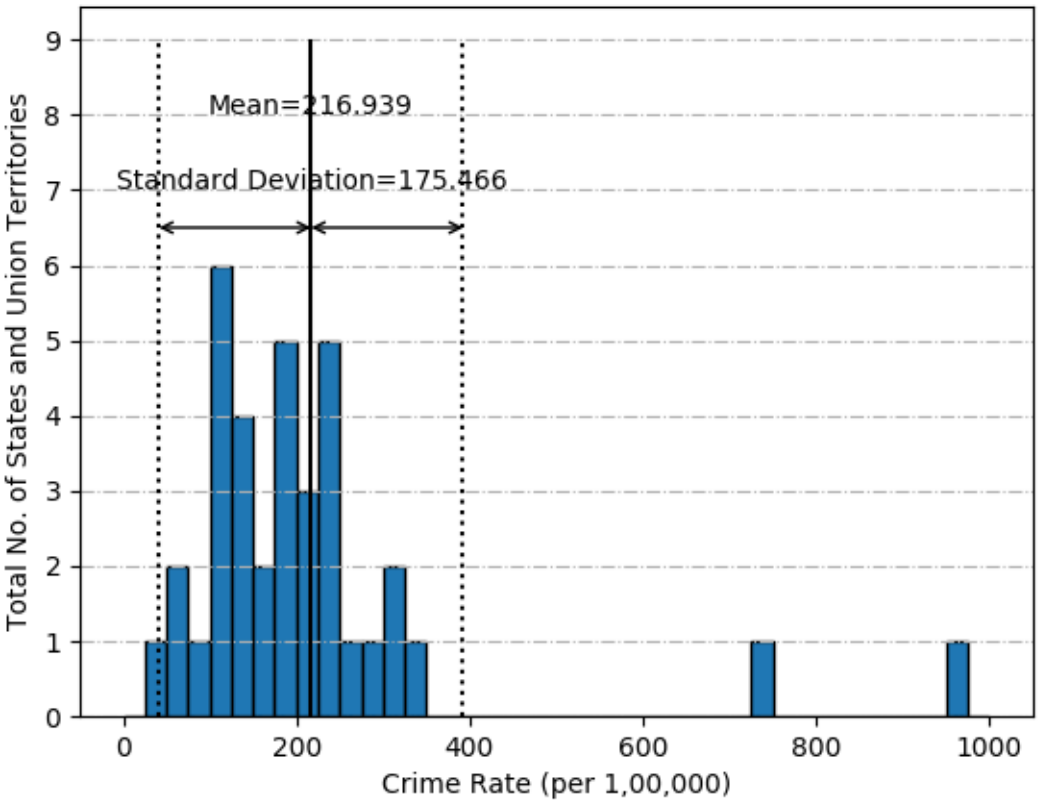
```
# importing numpy library for storing arrays and simple processes on array
import numpy as np
# importing pandas library for reading csv files
import pandas as pd
# importing pyplot from matplotlib to plot graphs
import matplotlib.pyplot as plt
# all libraries as imported as their popular short names

# defining function to find corellation between 2 data
def find_correlation(data_1,data_2):
        correlation = 0.0
        num_terms = len(data_1)

        mean_1 = np.mean(data_1)
        mean_2 = np.mean(data_2)

        for i in range(num_terms):
                correlation += ((data_1[i]-mean_1)*(data_2[i]-mean_2))/
(num_terms-1)

        return correlation

# defining number of states and uts, so ecev if number of states gets changed we
will need to change only one parameter
num_states = 29
num_uts = 7
num_states_ut = num_states + num_uts

# defining dictionaries for 2 data
names_states_ut = {}
crimerate_dict = {}
unemp_dict = {}

# defining lists for 2 data
crimerate_list = []
unemp_list = []

#defining averages and average standard deviations
mean_crimerate = 0.0
mean_unemp = 0.0
std_crimerate = 0.0
std_unemp = 0.0

# reading data from csv files
data_crime = pd.read_csv("crimerate.csv")
data_unemp = pd.read_csv("unemploymentrate.csv")

# reading names of states/uts( c(id=2) column) and crimerates ( j(id=9) column)
for i in range(num_states):
        names_states_ut[i] = data_crime.iloc[:,2][i]
        unemp_dict[data_unemp.iloc[:,1][i]] = data_unemp.iloc[:,4][i]
        crimerate_dict[data_crime.iloc[:,2][i]] = data_crime.iloc[:,9][i]
for i in range(num_uts):
        names_states_ut[i+num_states] = data_crime.iloc[:,2][i+num_states+1]
        unemp_dict[data_unemp.iloc[:,1][i+num_states]] = data_unemp.iloc[:,4][i
+num_states]
        crimerate_dict[data_crime.iloc[:,2][i+num_states+1]] = data_crime.iloc[:,9]
[i+num_states+1]

# putting all data in list
for i in range(num_states_ut):
        crimerate_list.append(crimerate_dict[names_states_ut[i]])
        unemp_list.append(unemp_dict[names_states_ut[i]])

# finding mean of data
mean_unemp = np.mean(unemp_list)
mean_crimerate = np.mean(crimerate_list)

# finding standard deviation of data
```

```
std_unemp = ((np.var(unemp_list,ddof=1)))**(0.5)
std_crimerate = ((np.var(crimerate_list,ddof=1)))**(0.5)

bins_unemp = np.linspace(0,12,37) # total 36 divisions between 0 to 12
plt.hist(unemp_list,bins=bins_unemp,ec="black") #plotting unemployment data
plt.yticks(np.arange(0,12,1)) # lines parallel to x-axis
# labels for axes
plt.ylabel('Total No. of States and Union Territories')
plt.xlabel('Unemployment Rate (in Percentage)')
# drawing lines for mean and standard deviations
plt.vlines(x = mean_unemp, ymin = 0, ymax = 7)
plt.vlines(x = mean_unemp - std_unemp, ymin = 0, ymax = 7, linestyle="dotted")
plt.vlines(x = mean_unemp + std_unemp, ymin = 0, ymax = 7, linestyle="dotted")
# annotating lines
plt.text(mean_unemp,6,'Mean='+str(round(mean_unemp,3)),ha='center')
plt.annotate(s='',xy=(mean_unemp-std_unemp,4.5),xytext=(mean_unemp,4.5),arrowprops=
{'arrowstyle':'<->','shrinkA':0,'shrinkB':0})
plt.annotate(s='',xy=(mean_unemp+std_unemp,4.5),xytext=(mean_unemp,4.5),arrowprops=
{'arrowstyle':'<->','shrinkA':0,'shrinkB':0})
plt.text(mean_unemp,5,'Standard Deviation='+str(round(std_unemp,3)),ha='center')

plt.grid(axis='y',zorder=0,ls='-.')
# saving figure
plt.savefig('Unemployment.png')

# clearing window
plt.clf()

bins_crimerate = np.linspace(0,1000,41) # total 40 divisions between 0 to 1000
plt.hist(crimerate_list, bins = bins_crimerate, ec = "black") #plotting
unemployment data
plt.yticks(np.arange(0,12,1)) # lines parallel to x-axis
# labels for axes
plt.ylabel('Total No. of States and Union Territories')
plt.xlabel('Crime Rate (per 1,00,000)')
# drawing lines for mean and standard deviations
plt.vlines(x = mean_crimerate, ymin = 0, ymax = 9)
plt.vlines(x = mean_crimerate - std_crimerate, ymin = 0, ymax = 9,
linestyle="dotted")
plt.vlines(x = mean_crimerate + std_crimerate, ymin = 0, ymax = 9,
linestyle="dotted")
# annotating lines
plt.text(mean_crimerate,8,'Mean='+str(round(mean_crimerate,3)),ha='center')
plt.annotate(s='',xy=(mean_crimerate-std_crimerate ,6.5),xytext=
(mean_crimerate,6.5),arrowprops={'arrowstyle':'<->','shrinkA':0,'shrinkB':0})
plt.annotate(s='',xy=(mean_crimerate+std_crimerate ,6.5),xytext=
(mean_crimerate,6.5),arrowprops={'arrowstyle':'<->','shrinkA':0,'shrinkB':0})
plt.text(mean_crimerate,7,'Standard Deviation='+str(round
(std_crimerate,3)),ha='center')

plt.grid(axis='y',zorder=0,ls='-.')
# saving figure
plt.savefig('Crimerate.png')

# clearing window
plt.clf()

# plotting scatter plot of crime rate vs unemployment rate
plt.plot(unemp_list,crimerate_list,'ro')
plt.xlabel('Unemployment Rate (in Percentage)')
plt.ylabel('Crime Rate (per 1,00,000)')
# saving figure
plt.savefig('ScatterPlot.png')

# clearing window
plt.clf()

bins = (bins_unemp,bins_crimerate)
plt.hist2d(unemp_list,crimerate_list, bins=bins)
```

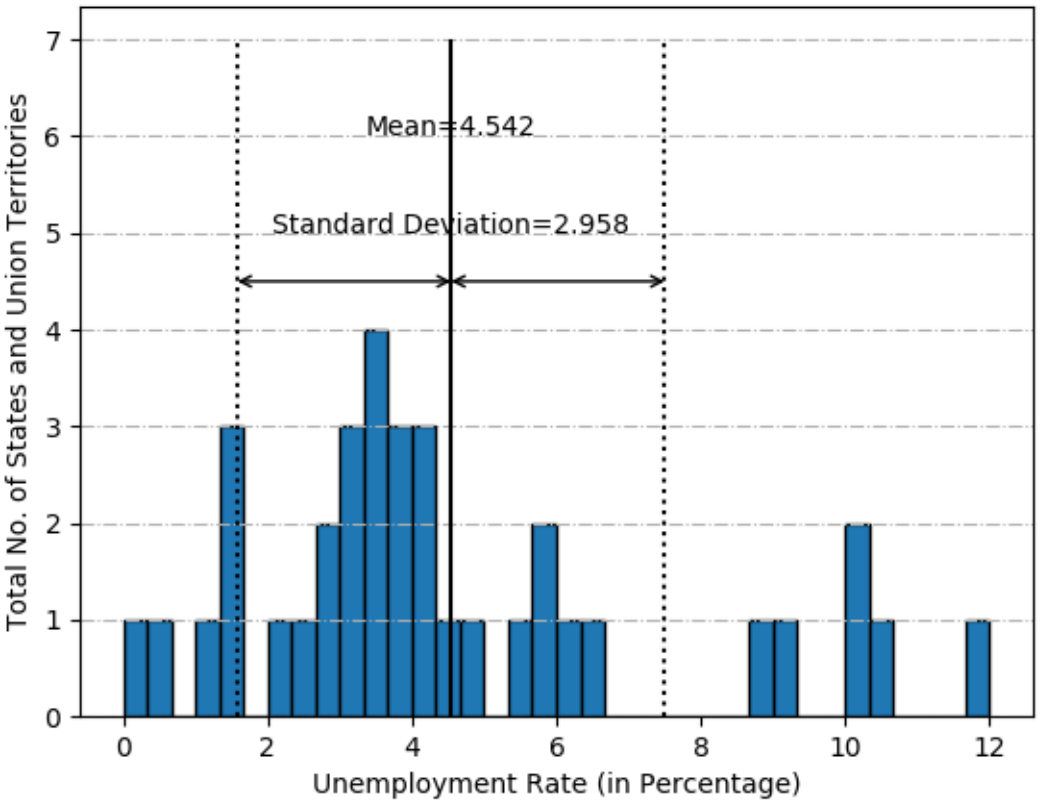
```
plt.xlabel('Unemployment Rate (in percentage)')
plt.ylabel('Crime Rate (per 100,000)')
# Drawing ColorBar for Information
cbar = plt.colorbar()
cbar.ax.set_ylabel('Counts')

# saving figure
plt.savefig('2D-histogram.png')

# clearing window
plt.clf()

# finding correation using function define in the start of the code
correlation = find_correlation(unemp_list,crimerate_list)

# finding correlation coefficient
correlation_co_eff = (correlation)/((std_unemp*std_crimerate))


# printing some important values to be observed
print("mean_unemp = " + str(mean_unemp))
print("std_unemp = " + str(std_unemp))

print("mean_crimerate = " + str(mean_crimerate))
print("std_crimerate = " + str(std_crimerate))

print("correlation = " + str(correlation))
print("correlation_co_eff = " + str(correlation_co_eff))
```

1D Histograms

A) Histogram of crime Rate in different States and Union Teritories.

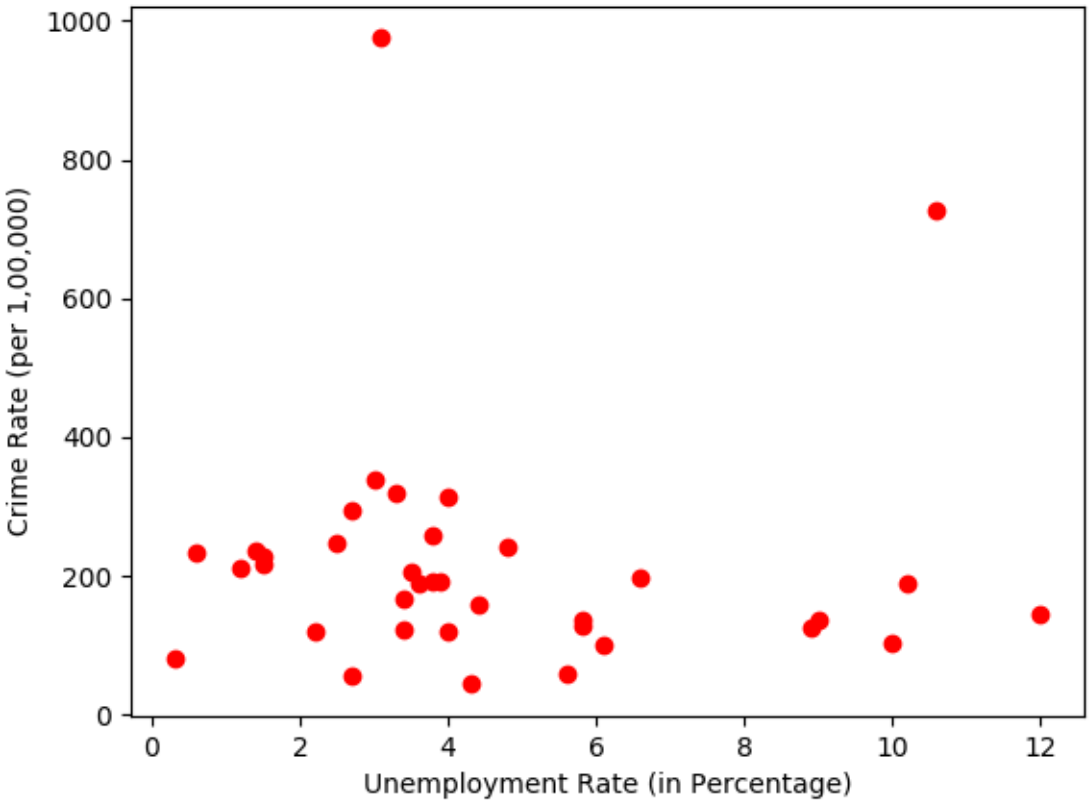


B) Histogram of Unemployment Rate in different states and Union Teritories.



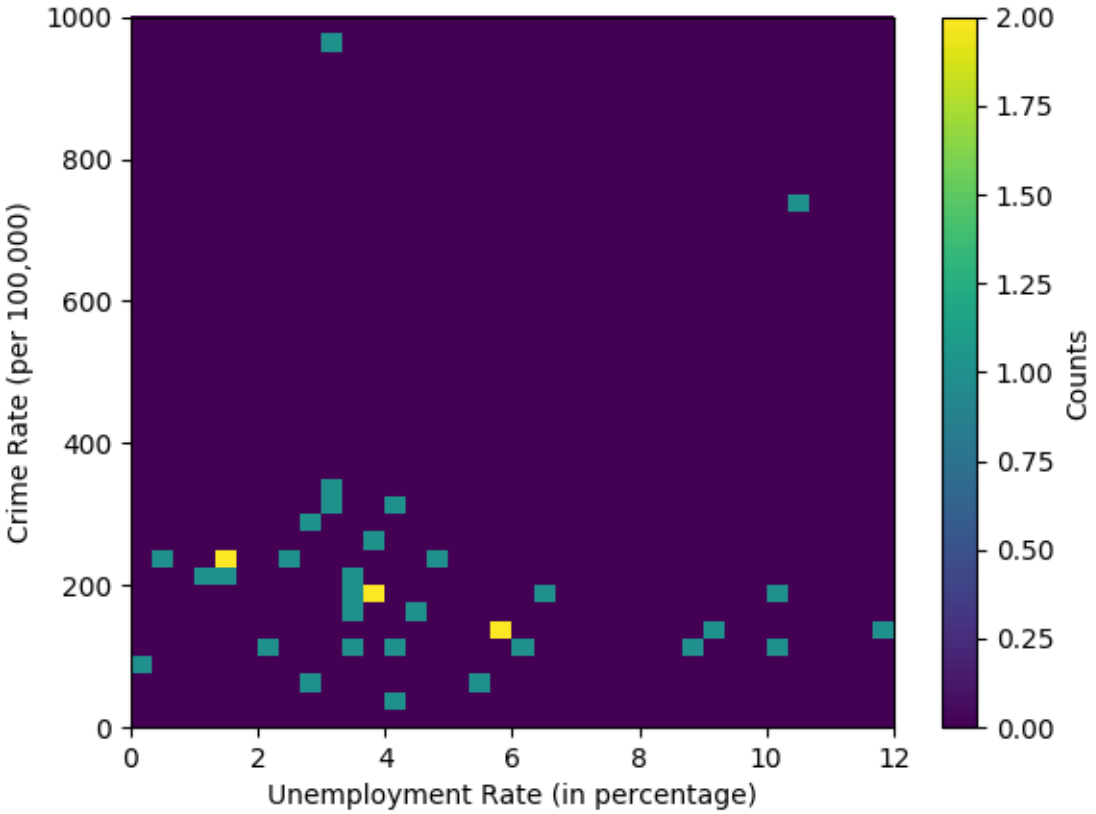
Scatter Plot

Scatter Plot of the pairs (U_i, C_i)



2D Histogram

2D Histogram of the Pairs (U_i, C_i)



Conclusion

By ploting this data we conclude that:

a) Mean and Standard deviation of Crime Rate are 216.939 and 175.466 respectively.

b) Mean and Standard deviation of Unemployment Rate are 4.542 and 2.958 respectively.

c) Correlation of Crime rate and Unemployment rate is 1.917×10^{-1}

d) Correlation coefficient is 3.69×10^{-4}

e) Correlation coefficient is Positive but very small. So Unemployment rate and Crime rate are nearly Independent.

Team Contribution

a) **Vashishtha Kochar** - Report writer 25%

b) **Nihal Barde** - Programmer 25%

c) **Adeem Jassani** - Team Leader 25%

d) **Ram** - Web Developer 25%