PROJECT 2

REPORT

UNDER SUPERVISON OF
DR. NOR AZIZAH BINTI ALI

PROBABILITY & STATISTICAL DATA ANALYSIS COURSE

PREPARED BY
AHMED ZAKI AL-GABALY
A22EC4003

1. INTRODUCTION

The study's goal is to look at the possible link between the unemployment rate in a number of US states and the prevalence of violent occurrences. Understanding this link might provide useful insights into the social and economic issues that may contribute to violence within a specific population. The findings of this analysis could help governmental organizations to understand the substantial relationship between the unemployment rate and crime rate, which in turn will help them in to take better decisions to treat unemployment and prevent violent incidents.

At first, we will give some possible expectations before conducting the analysis. The possible expectations are stated below:

1. **Positive Relationship**: A positive relationship may be expected, implying that higher unemployment rates are related to a higher occurrence of violent incidents. This could be due to the frustration, economic strain, or a lack of job opportunities, all of which contribute to increased social stress and conflict.
2. **Negative Relationship**: Alternatively, one may predict a negative relationship, showing that higher unemployment rates are related to a lower occurrence of violent incidents. This might be due to better community cohesion, stronger social support networks, or a prioritization of economic redistribution wherein unemployment will be considered as a shared challenge which will reduce the feelings of inequality.
3. **No Relationship**: Another theory is that there is no relationship between unemployment and violent incidents. Other socioeconomic factors, such as unemployment, may have a greater influence on the aggressive conduct in such scenario.

It's crucial to emphasize that these predictions are uncertain, and that the actual results of the data analysis will give more trustworthy insights into the relationship between unemployment and violent incidents.

2. DATASET

The dataset chosen for the research includes data on some socioeconomic indicators and crime statistics from various areas and historical periods in US states. Each row in the dataset represents a US state in a certain year, while the columns indicate various factors that have been measured. However, I have modified the dataset to align my research needs. Within the analysis, I want to focus on the data of 2014 for the US states. Indeed, the modifications applied on the original dataset (as pre-processing) are as follows (Refer to Appendix to view the raw dataset and processed one):

- Renaming some of the columns for better readability.
- Converting the values of state column from state codes to state names.
- Deleting the columns that will not be used for the analysis.
- I kept only data of 2014.

Each column of the processed dataset will be briefly described below:

1. *State*: The values are the name of US states.
2. *Unemployment Rate*: The values are the unemployment rate at the specified year.
3. *Total Rate of Violent Crimes*: The values are the sum of the Murder Rate column, Rape Rate column, Robbery Rate column, and Aggravated Assault Rate.

2.1. The Chosen Variables

The Chosen Variables	
Variable name	Variable Type
Unemployment Rate	Independent Variable
Total Rate of Violent Crimes	Dependent Variable

The above table states the chosen variables and their types. With these variables we will be able to test the relationship between the unemployment rate and the total number of violent occurrences. The null hypothesis holds that there is no correlation between the unemployment rate and total rate of violent crimes in the 51 states in 2014, whereas the alternative hypothesis holds that there is a correlation between the variables. Also, with these variables, we will be able to analyze the relationship between these two variables and get some insights into probable socioeconomic aspects that may contribute to crime rates, as well as guide crime prevention.

2.2. Pre-Analysis Assumptions

Before conducting the analysis, it is important to assure the following assumption:

1. The provided data is normally distributed.
2. The relationship between the variables is linear.
3. No outliers exist in the provided data.

2.2.1. *The provided data is normally distributed.*

To determine if the data is normally distributed, I used Normal Probability Plot (Q-Q Plot) within the R script. The R code in the script generated two separate Q-Q plots, one for the unemployment rate and another for the violent crimes rate. For the unemployment rate's plot showed the points to be approximately linear following the reference line as shown in (Figure 2.1). This suggests that the unemployment rate data is reasonably close to a normal distribution. As for the total rate of violent crimes' Q-Q plot, the points also followed the reference line indicating that the total rate of violent crimes data is approximately normally distributed (Figure 2.2). Therefore, we can conclude that both variables appear to be close to normal distribution.

Figure 2.1: Normal Probability Plot (Q-Q Plot) – Unemployment Rate

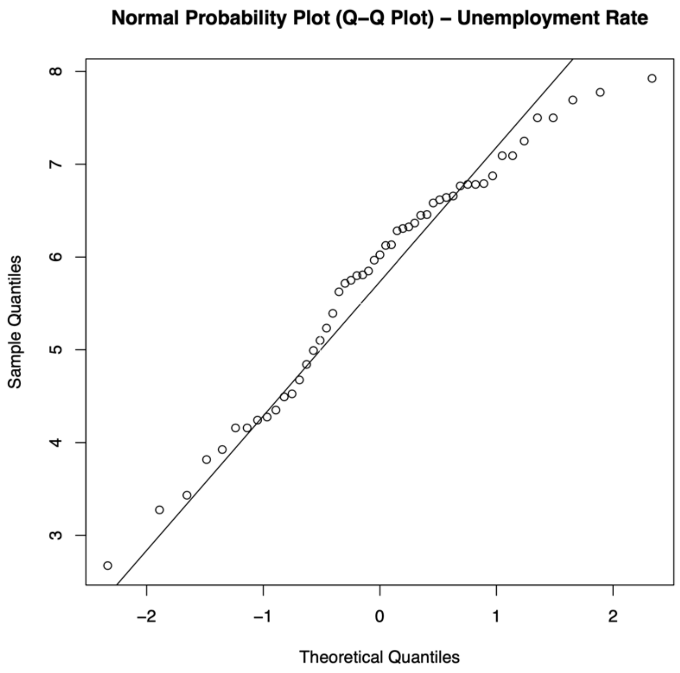
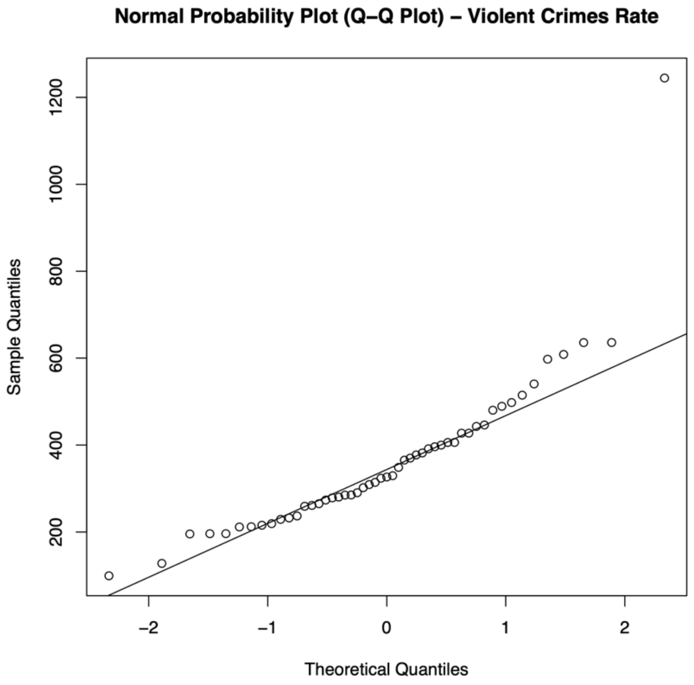


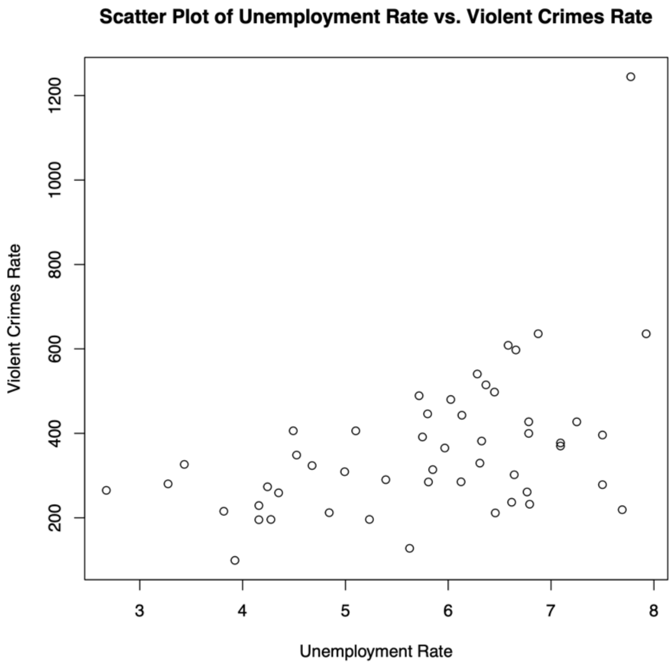
Figure 2.2: Normal Probability Plot (Q-Q Plot) – Violent Crimes Rate



2.2.2. The relationship between the variables is linear.

To check the linearity of the relationship between the two variables, I used the scatter plot as shown below (Figure 2.3) using R script. Based on that graph, the data points were scattered across the plot without forming a distinct linear pattern. Despite the scatter plot is not indicating a clear linear pattern between the variables, I will proceed to examine the correlation between the two variables in the subsequent steps. Specifically, I will calculate the **Pearson correlation coefficient** to assess the strength of their association (in the Analysis section).

Figure 2.3: Scatter Plot – Unemployment Rate Vs. Violent Crimes Rate



2.2.3. No outliers exist in the provided data.

In fact, there is one outlier in the Total rate of violent crimes as it is shown in figure 2.3. However, I have chosen not to remove this outlier as it represents a valid and an accurate value. This value is 1,244.4, which belongs to District of Colombia, widely known for its significantly high crime rates (Meola, 2014). Therefore, it is unlikely to be an error. Furthermore, the primary research question focuses on examining the correlation between the two variables. To address this, I calculated the Pearson correlation coefficient within the Analysis section both with and without the outliers. In both cases, the coefficient revealed a moderately positive relationship between the variables. These findings will be discussed in greater detail in the analysis section.

3. ANALYSIS

3.1. Hypothesis Testing

Null Hypothesis (H_0): There is no correlation between the unemployment rate and total rate of violent crimes in the 51 states in 2014.

Alternative Hypothesis (H_1): There is a correlation between the unemployment rate and total rate of violent crimes in the 51 states in 2014.

The Significance Level (α) is chosen to be 0.05 for this test.

H0: $\rho = 0$ (Null Hypothesis)

H1: $\rho \neq 0$ (Alternative Hypothesis)

3.2. Correlation Test

Now, to perform the correlation test between the two variables, I used the Pearson Correlation Coefficient using R script. The following table is the summary of the output generated by R script.

No.	Variables	Value
A	Correlation Coefficient (r)	0.4723111
B	t-Value	3.7509
C	Degrees of Freedom (df)	49
D	p-Value	0.0004664
E	Confidence Interval	(0.2261654, 0.6617594)

3.2.1. Correlation Coefficient (r)

The correlation coefficient found to be 0.4723111. It was calculated in R script using cor() function. Since this value lies between 0.3 and 0.7, this suggests that there is a positive relationship between the variables, but the association is not strong, instead, it is a moderate positive correlation ($0.3 \leq r < 0.7$).

NOTE: I calculated the Pearson Correlation Coefficient with and without the outlier. The correlation coefficient with the outlier r=0.4723, while without the outlier it was found to be r=0.4552. So, in both cases, we can interpret the result as a moderate positive linear relationship between the variables. Accordingly, the presence of an outlier in this case did not have a significant effect.

3.2.2. t-Value

t-value found to be 3.7509. This value measures the strength and the direction of the correlation considering the sample size. Since the t-value is positive and significantly different from zero, it indicates a statistical positive correlation between the unemployment rate and the total rate of violent crimes.

3.2.3. Degrees of Freedom (df)

Degrees of freedom found to be 49. Which is calculated by n-2. It represents the number of independent observations used in the correlation test.

3.2.4. p-Value

p-Values found to be 0.0004664. It measures the strength of evidence against the null hypothesis. In our case, the calculated p-value is small, which suggests strong evidence against the null hypothesis of no correlation. In other words, there is a significant correlation between the unemployment rate and total rate of violent crimes.

3.2.5. Confidence Interval

The confidence interval found to be between 0.2261654 and 0.6617594. This suggests that with 95% confidence, the true correlation value falls between 0.2261654 and 0.6617594.

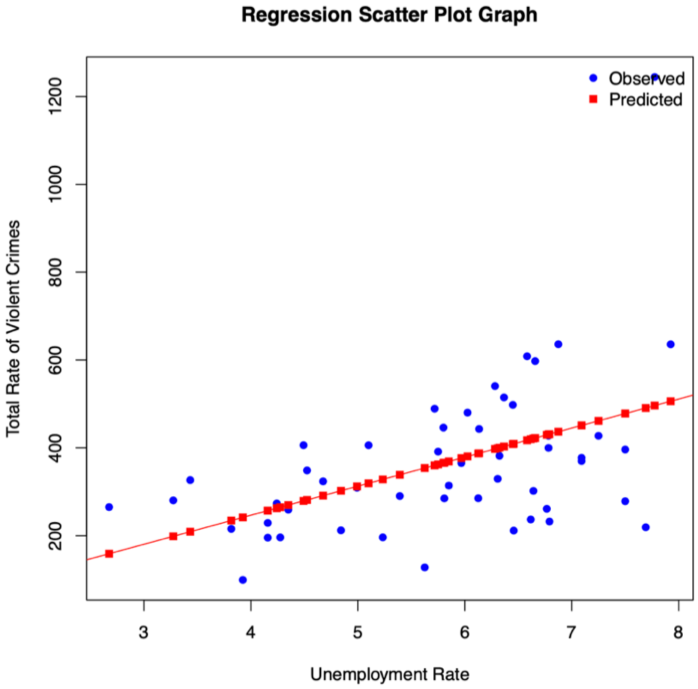
Overall, the output confirms a moderately positive correlation between the unemployment rate and the total rate of violent crimes. The correlation test indicates that this correlation is statistically significant, with a low p-value, supporting the presence of a meaningful relationship between the variables.

3.3. Regression Analysis (Including Goodness of Fit Test)

The scatter plot in Figure 2.3 can be used to assess the linearity, identify outliers, and observe patterns and variability between the two variables, which is a crucial step before conducting the regression model.

To fit a linear regression model, I used the R script. The below figure shows the Regression model (Figure 3.1).

Figure 3.1: Regression Scatter Plot Graph



3.3.1. Residuals

The following tables represents the output of R script. The below table displays the summary statistics of the residuals (differences between the observed and the predicted values).

Residuals				
Min	1Q	Median	3Q	Max
-271.49	-81.57	-18.51	87.85	748.22

3.3.2. Coefficients

The below table presents the estimated coefficients for the intercept and the unemployment rate predicted variable. The **estimate** column provides the estimated values. The **Std. Error** column represents the standard errors. The t-value is the ratio of the estimated coefficient to its standard error, and the Pr(>|t|) column shows the p-values for each coefficient.

Coefficients				
	Estimate	Std.Error	t-value	Pr(>|t|)
(intercept)	-17.87	104.33	-0.171	0.864726
Unemployment Rate	66.12	17.63	3.751	0.000466

In this case the intercept coefficient is not statistically significant (P= 0.865), while the unemployment rate coefficient is significant (p=0.000466).

3.3.3. *The Goodness of Fit*

The Goodness of Fit	
Metrics	Value
Residual Standard Error	159.4
Multiple R-Squared	0.2231
Adjusted R-squared	0.2072
F-statistic	14.07
p-value	0.0004664

The residual standard error is 159.4, which indicates the standard. Deviation of the residuals providing a measure of the model's goodness effect.

The multiple R-squared and Adjusted R-squared were found to be 0.2231, 0.2072 respectively. These two metrics are used to assess the goodness of fit regression model and they represent the proportion of variance in the dependent variable explained by the independent variable. The multiple R-squared indicates that 22.31% of the variance in the total rate of violent crimes can be explained by the unemployment rate. Since the Adjusted R-squared is less than 1, it suggests that there is still a significant portion of the variation in the dependent variable that is not accounted for by the unemployment rate alone. This indicates that there might be other factors or variables that also contribute to the violent crimes rate.

F-statistic represents the overall significant of the model. It tests the null hypothesis that all coefficients are zero. The f-statistic is 14.07 with 1 and 49 degrees of freedom, and the associated p-value is 0.0004664, indicating strong evidence against the null hypothesis.

Overall, the regression analysis suggests that the unemployment rate is a significant predictor of the total rate of violent crimes, with a positive coefficient of 66.12. The model explains 22.31% of the variance, and the f-statistic indicates the model's overall significance.

4. Conclusion

To sum up, we can reject the null hypothesis of no correlation between the two variable and conclude that there is evidence of a relationship between the two variables, therefore we accept the alternative hypothesis.

Through the process of choosing the dataset, I realized how important it is to select the right data for research and analysis. I discovered that picking a dataset that matches my research question and goals is vital for getting useful and significant findings. As for the pre-processing, I had committed multiple modifications on the original data set as mentioned previously to help me in conducting the analysis within the required range of data.

The analysis itself was interesting. My ability to officially analyze the relationship between variables was made possible by hypothesis testing, specifically the Pearson correlation coefficient. I was able to evaluate the strength and importance of associations thanks to my understanding of the interpretation of test findings.

I learned about using linear regression to predict one variable based on another in the regression analysis portion. I learned more about the connection between the rate of violent crimes and the unemployment rate by constructing a regression model. The interpretation of the regression coefficients revealed important details regarding the nature and strength of this relationship.

The substantial correlation between the rate of violent crimes and the unemployment rate was one of my research's interesting findings. Although I recognize that other factors may also play a role, learning about how unemployment may affect a community's stability and safety was interesting. This result underlines the pressing need to address unemployment as a significant societal issue and establish all-encompassing solutions to foster social well-being.

Overall, I feel that this project has given me essential analytical abilities and expanded my understanding of statistical tests and their uses. I now have a greater knowledge of complicated social processes and am more equipped for upcoming data analysis projects. I'm amazed at how raw data can be transformed into meaningful insights that help solve societal issues.

5. Limitations and Assumptions

It is important to consider the limitations and assumptions, which will be stated. As follows:

- The presence of an outlier generally can affect the regression analysis results, however as previously mentioned in section ??, the influence of the outlier in this present analysis was found not significant.
- Additional variables and factors could have influenced the total rate of violent crimes such as the socioeconomic factors, demographics, and society characteristics. Further understanding of the impact of these factors should be examined.
- The dataset was specified to one year which 2014 to simplify the analysis. Therefore, it may not be possible to generalize these findings to other time periods.
- The relationship in the scatter plot as explained previously was not distinctly linear, however the Pearson coefficient still indicated a significant relationship.
- The regression model assumes a linear relationship between the variables and assumes that the data values are normally distributed.

6. Appendix

6.1. Raw Dataset

- https://www.kaggle.com/datasets/lydiavasil/crime-rate-and-unemployment-rate-by-state (Also can be found as CSV in the folder named as original_dataset.csv)

6.2. Pre-processed Dataset

- The following is the processed dataset. (Also can be found as CSV in the folder named as processed_dataset.csv)

State	Unemployment_Rate	Total_Rate_of_Violent_Crimes
Alabama	6.783	427.4
Alaska	6.875	635.8
Arizona	6.783	399.9
Arkansas	6.025	480.1
California	7.5	396.1
Colorado	4.992	309.1
Connecticut	6.617	236.9
Delaware	5.717	489.1
District of Columbia	7.775	1244.4
Florida	6.283	540.5
Georgia	7.092	377.3
Hawaii	4.35	259.2
Idaho	4.842	212.2
Illinois	7.092	370
Indiana	5.967	365.3
Iowa	4.242	273.5
Kansas	4.525	348.6
Kentucky	6.458	211.6
Louisiana	6.367	514.7
Maine	5.625	127.8
Maryland	5.8	446.1
Massachusetts	5.75	391.4
Michigan	7.25	427.3
Minnesota	4.158	229.1
Mississippi	7.5	278.5
Missouri	6.133	442.9
Montana	4.675	323.7
Nebraska	3.275	280.4

Nevada	7.925	635.6
New Hampshire	4.275	196.1
New Jersey	6.767	261.2
New Mexico	6.658	597.4
New York	6.325	381.8
North Carolina	6.308	329.5
North Dakota	2.675	265.1
Ohio	5.808	284.9
Oklahoma	4.492	406
Oregon	6.792	232.3
Pennsylvania	5.85	314.1
Rhode Island	7.692	219.2
South Carolina	6.45	497.7
South Dakota	3.433	326.5
Tennessee	6.583	608.4
Texas	5.1	405.9
Utah	3.817	215.6
Vermont	3.925	99.3
Virginia	5.233	196.2
Washington	6.125	285.2
West Virginia	6.642	302
Wisconsin	5.392	290.3
Wyoming	4.158	195.5

6.3. Other References

- Meola, Andrew ANDREW. “The 10 Most Dangerous States in America.” *The Street*, 29 Nov. 2014, www.thestreet.com/markets/the-10-most-dangerous-states-in-america-12968105.