

**TASK**

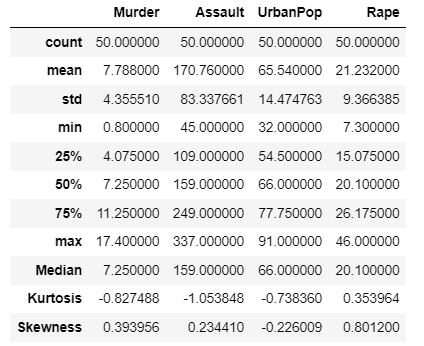
**Exploratory Data Analysis on the Arrest Dataset**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

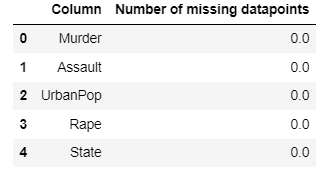
The dataset has 4 quantitative variables and 1 categorical variable called city which has been renamed to State. The table below shows the summary statistics for each of the 4 quantitative variables. Assault has the greatest rate per 100,000 compared to murder and rape. It also has the greatest standard deviation meaning the spread is quite large. Murder has the lowest standard deviation and average value out of the 3 crimes. The skewness of all fields except Rape is between -0.5 and 0.5 indicating that the data is fairly symmetrical. Rape has a moderate skewness. Furthermore, the kurtosis indicates that all 4 fields have platykurtic distributions as the kurtosis is so small and excess kurtosis would is negative. This means that there aren’t many outliers as the tails are not so heavy.



**DATA CLEANING**

* **The data types of the fields were checked**
* **Uniqueness check: Searched for duplicated rows**
* **Searched for missing cells or unconventional symbols in the data**

**MISSING DATA**

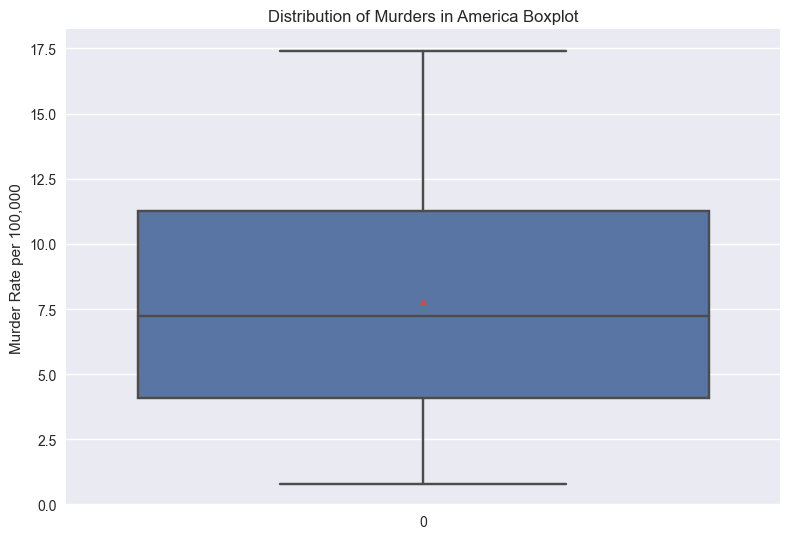
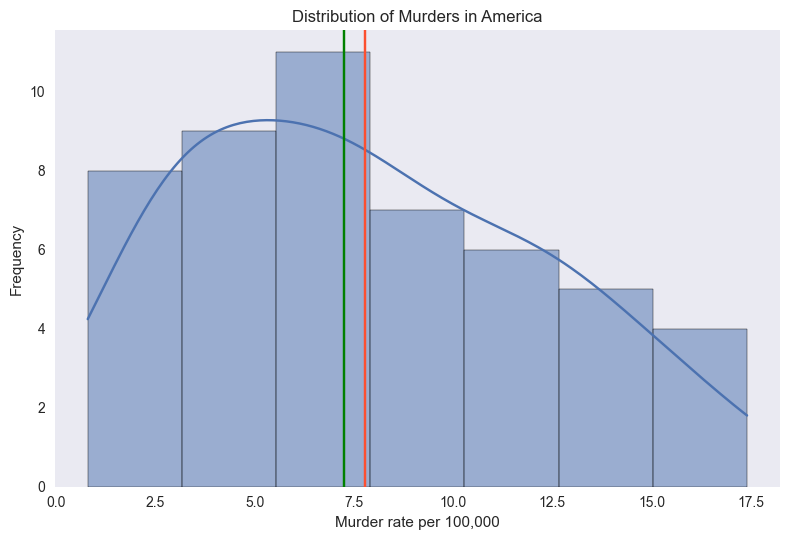


There are no missing datapoints in any of the fields in arrest dataset.

**DATA STORIES AND VISUALISATIONS**

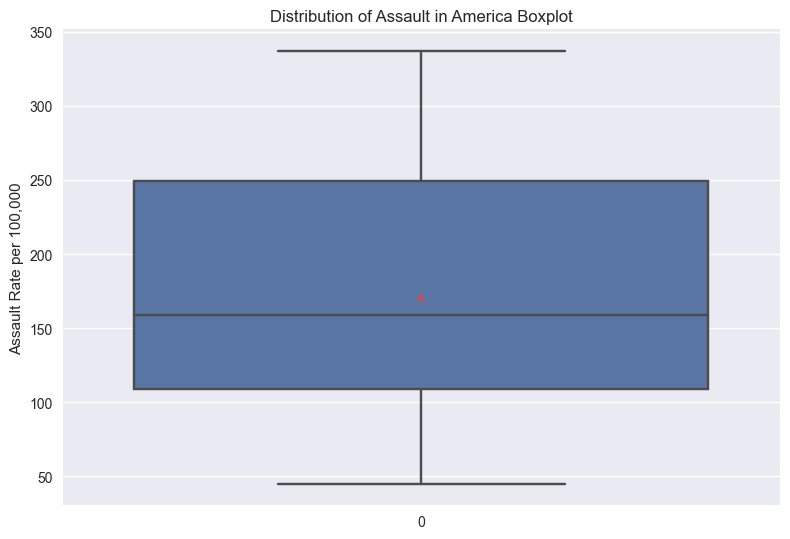
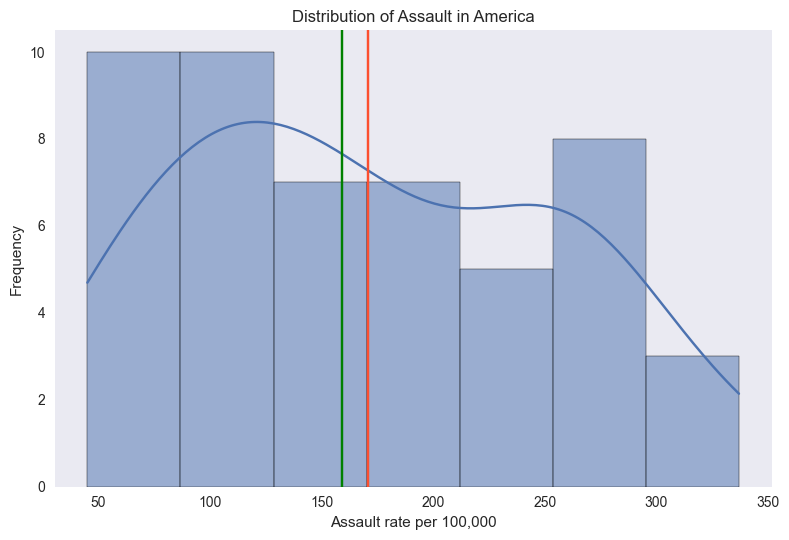
Univariate Analysis:

* Murder Rates



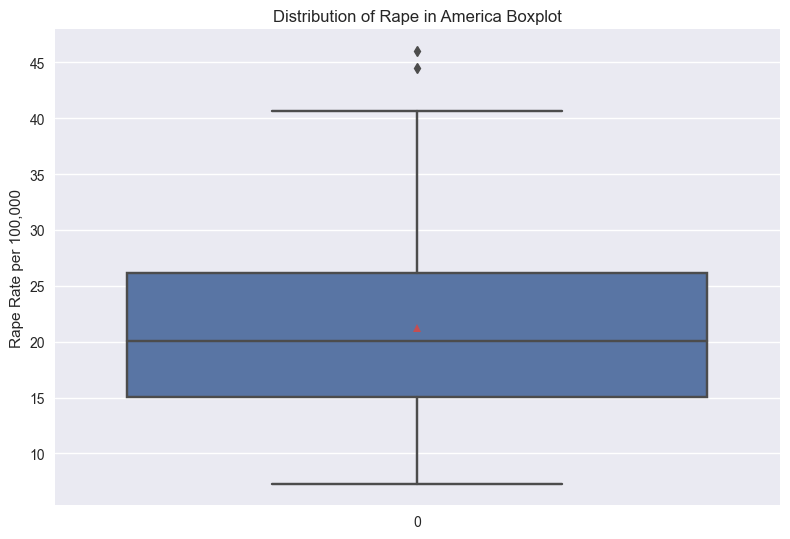
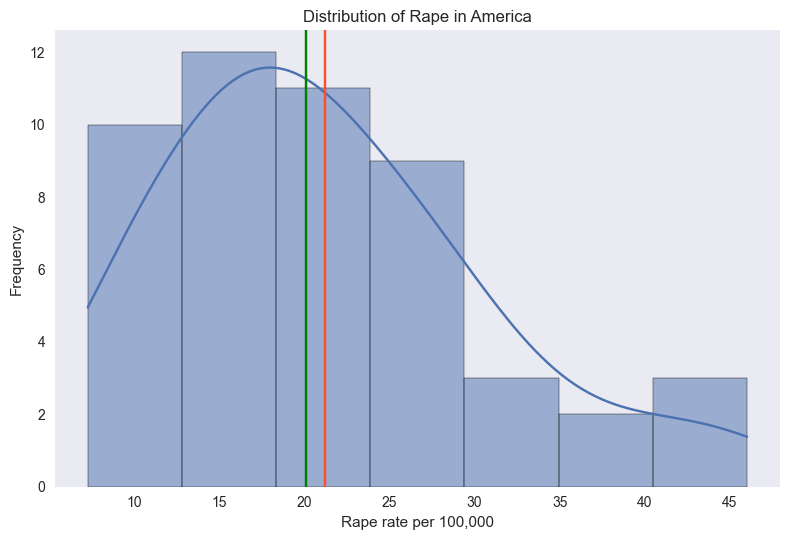
The diagrams above display the univariate distribution of Murder rates per 100,000 in America. We use a histogram with kernel density estimates. Secondly we use a boxplot to see the spread of the data. The histogram indicates that the distribution is skewed slightly to the left. Notice that the mean value is greater than the median value (green) meaning the skew is positive. The boxplot shows us that 50% of the data lies between 11 and 4 murders per 100,000 which means that the variation in the data of murders is not very large at the country level.

* Assault rates:

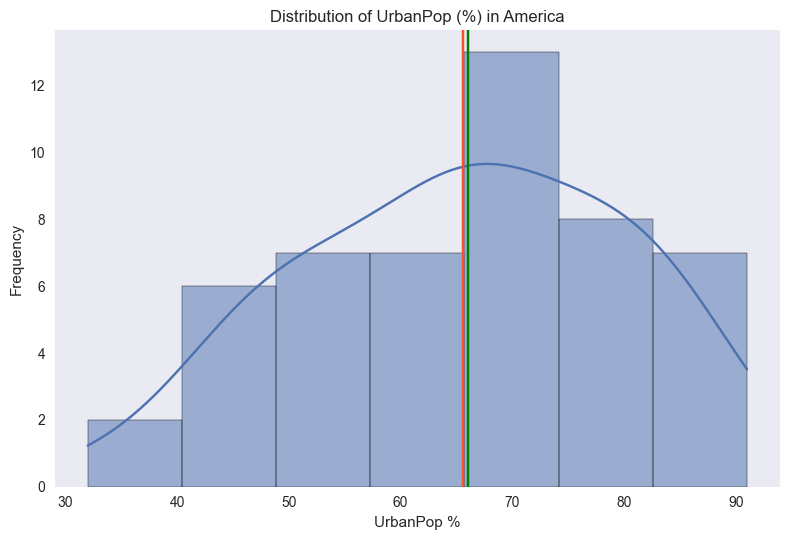
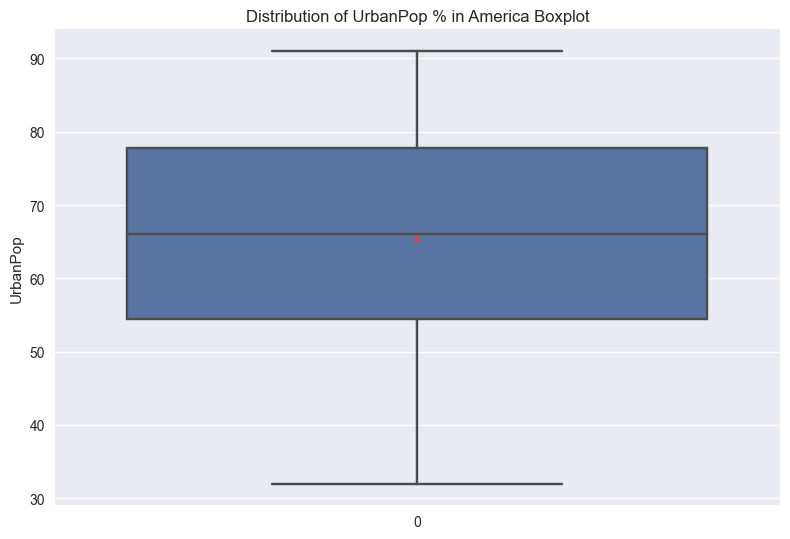


The graphs above display the univariate distribution of assaults per 100,000 in America. The shape of the data is bimodal. The red line displays the average value of assault rates per 100,000 and it is greater than the median which is why the distribution has a positive skew. The boxplot shows that 50% of the data lies between roughly 100 and 250 assaults per 100,000 people and therefore the spread of the data is large. However, there are no datapoints greater than 1.5\* the interquartile range hence there are no outliers.

* Rape:



The figure above shows the distribution of rape per 100,000 people in America. We can see a highly positively skewed distribution indicating that the rape rates across the different states of America tends to be on the lower range. Furthermore the median value of rape rates is lower than the mean. On average, there are around 21 rape cases per 100,000 people. The interquartile range for this distribution is 10 rapes per 100,000 people indicating that there is less variability in the data. In other words, the middle 50% of the data tightly clusters together. The boxplot also shows us that there are two outliers with rape rate of over 40 per 100,000 people.



The distribution of urban population in America is negatively skewed with majority of the states having higher urban population %. This could be attributed to the fact that America is a large developed country and it is no longer an agrarian economy thus most states have developed cities. The mean and median are almost equal at around 65 %. The interquartile range is about 20% which means that the middle values are tightly clustered together. Overall the range is quite large with some states having 90% urban populations whilst rural states having only 30% urban populations.

**Relationships between variables:**

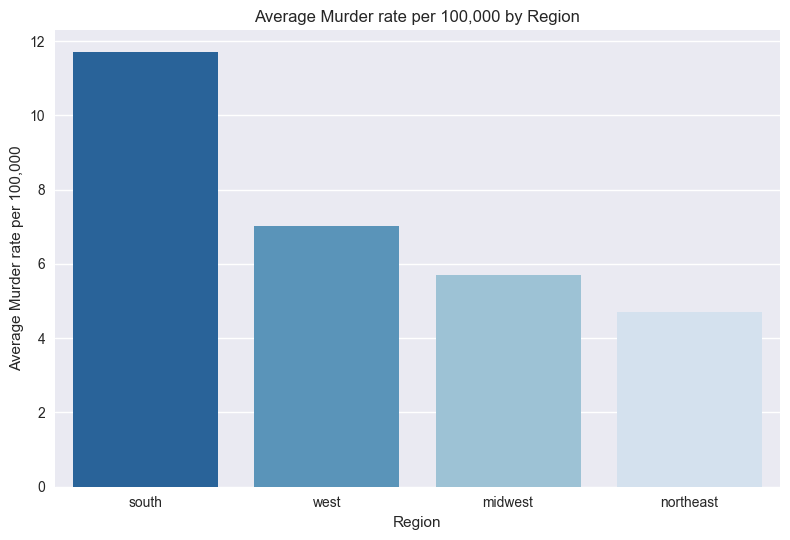
* **Crimes By State & Region**

1. **Murder**

The figure below shows the murder rate per 100,000 people across the 50 states of America. We can see that Georgia has the highest rate of 17.5 murders per 100,000 people whilst North Dakota has the lowest rate. We can also see that the average murder rate is the highest in Southern states of America (12), whilst the northeast has the lowest average murder rate of 5 per 100,000 people.

Chart

Description automatically generated



1. **Rape**

The diagram above shows the rape cases per 100,000 people across different states of America. Nebraska and Alaska have the highest rate of rape cases over 40 per 100,000 people which is quite high. North Dakota again has the lowest rape cases as well as the lowest murder rate. Generally, western states have the highest rape cases at around 30 per 100,000 people and northeast has the lowest at around 15 per 100,000.

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

1. **Assault**

The assault rates in America can be as high as around 325 per 100,000 in North Carolina and Florida and as low as 50 per 100,000 in North Dakota. There is quite a lot of variation in the data however, based on the 3 metrics, North Dakota seems to have the lowest rates in all 3 types of crimes and could be considered a safer place to live. In general, Assault rates are highest in the southern states of America whilst the Midwest has the lowest average rate of assault per 100,000 people.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

From the descriptive statistics, we can conclude that on average, southern states tend to have the largest non-sexual crimes whereas the western states e.g. Alaska have the highest sexual crimes. The north-eastern states and midwestern states tend to be generally safer. Although region seems to be a contributing factor in explaining variation in these crime rates, we would need more information such as incomes, culture, policing, and other socio-economic factors to understand why there are regional differences in different types of crimes.

**Urban Population**Chart

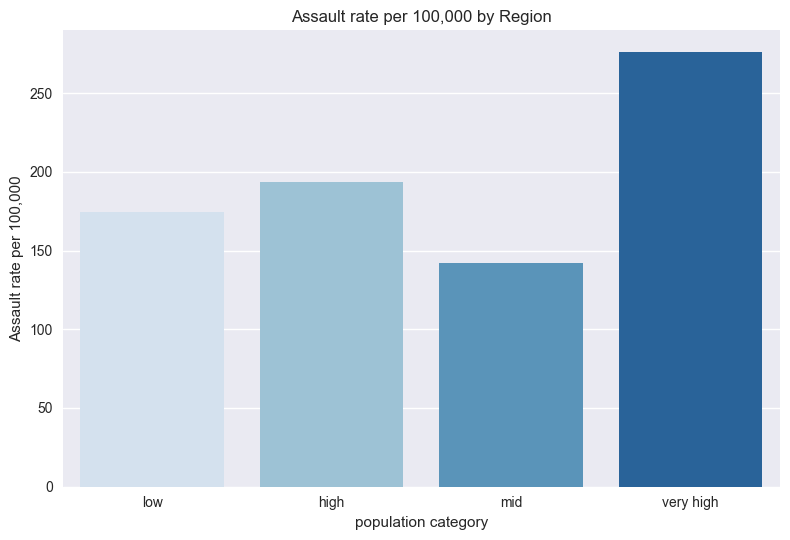
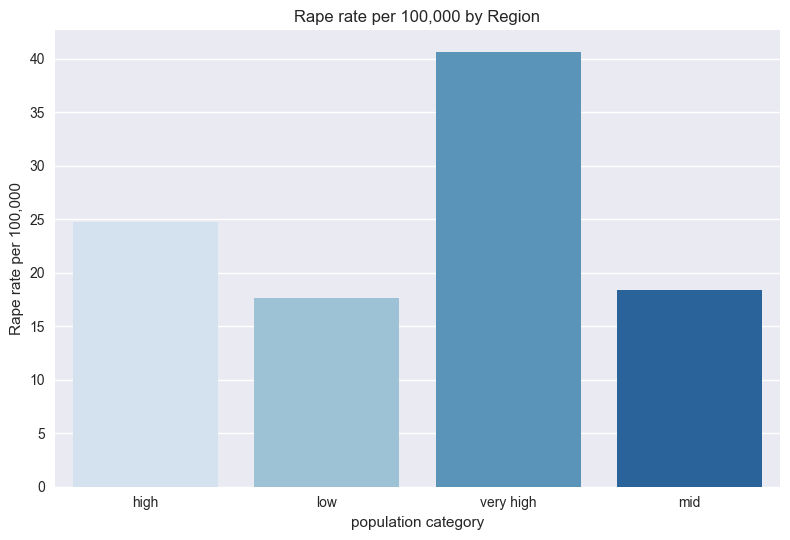
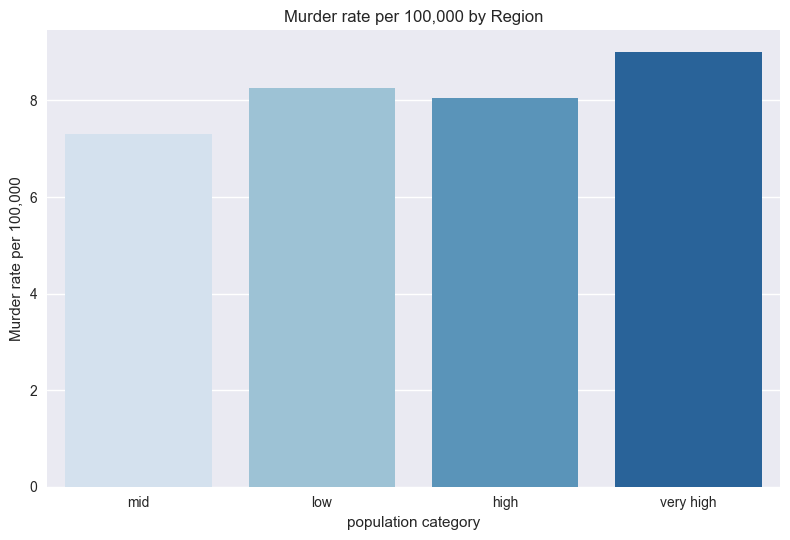
Description automatically generated

Chart, bar chart

Description automatically generated

The graph above shows the % of urban population per state. California has the highest urban population of around 90% followed by New Jersey, Rhode Island and New York. On the other hand, Vermont, West Virginia and Mississippi have some of the lowest urban population % at around 35-40%. Overall, the western and northeastern states have the highest average urban population % whereas the Midwest and south have the lowest urban population %.

* **Average Crime Rates by Population % category**



The charts above indicate that population clusters play a contributing role in explaining the variation in some crimes but not in others. For example, the higher population cluster, the higher the average number of rape cases and assault rates per 100,000 people.

On the other hand, very high urban population clusters in states result in an average murder rate of over 8 per 100,000 but states with low urban population clusters have a greater average murder rate than those with just “high” urban population clusters. Also, population clusters may be a contributing factor to why western states, which generally have very high population, tend to have high rape cases. However, the story changes if we look based on urban population clusters for assault rates as states with very high population clusters have a higher average rate of assault than those with lower population rates even if individually states like North Carolina have the highest assault rates because they are being put in a basket together with other low urban population.

Correlation:

Chart, scatter chart

Description automatically generated

* Murder has a strong positive correlation with Assault however it has a weak positive correlation with Rape and no correlation with Urbanpop.
* Assault has weak positive correlation with urban population but a stronger positive correlation with Rape.
* Rape has a strong weak positive correlation with urban population

**Regression Analysis: Understanding relationships between these variables**

**OLS Regression Results**

**==============================================================================**

**Dep. Variable: Murder R-squared: 0.808**

**Model: OLS Adj. R-squared: 0.781**

**Method: Least Squares F-statistic: 30.14**

**Date: Mon, 30 Jan 2023 Prob (F-statistic): 6.97e-14**

**Time: 09:20:23 Log-Likelihood: -102.77**

**No. Observations: 50 AIC: 219.5**

**Df Residuals: 43 BIC: 232.9**

**Df Model: 6**

**Covariance Type: nonrobust**

**==============================================================================**

**coef std err t P>|t| [0.025 0.975]**

**------------------------------------------------------------------------------**

**const 0.7335 1.506 0.487 0.629 -2.304 3.771**

**Assault 0.0225 0.006 3.853 0.000 0.011 0.034**

**UrbanPop -0.0187 0.025 -0.756 0.454 -0.068 0.031**

**Rape 0.1875 0.056 3.340 0.002 0.074 0.301**

**South 3.1577 0.949 3.329 0.002 1.245 5.071**

**West -2.0506 0.912 -2.249 0.030 -3.890 -0.212**

**Northeast -0.1535 0.967 -0.159 0.875 -2.103 1.796**

**==============================================================================**

**Omnibus: 1.584 Durbin-Watson: 1.601**

**Prob(Omnibus): 0.453 Jarque-Bera (JB): 0.810**

**Skew: 0.237 Prob(JB): 0.667**

**Kurtosis: 3.406 Cond. No. 1.12e+03**

**==============================================================================**

**Notes:**

**[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.**

**[2] The condition number is large, 1.12e+03. This might indicate that there are**

**strong multicollinearity or other numerical problems.**

The table above displays the multiple regression of explanatory variables such as rape, assault, urban population and geographical dummy variables on the murder rate per 100,000 across the United States. Firstly, we can test the validity of the model by performing the hypothesis test of whether

H0: Assault coefficient = UrbanPop coeff = Rape coeff = South coeff = West coeff = Northeast coeff = 0

H1: Not H0

**In other words, H0 says that we are better off simply taking the average of murder as an estimator of murder. The F-statistic of this model is very large at around 30. This means that we are able explain more variation in murder with this restricted model than we are by simply using the average of murder. Moreover, the adjusted R squared is very high which means that we are able to explain a lot of variation in the murder rate using our explanatory variables.**

**Secondly, although the text in the summary indicates that there may be multicollinearity, if we look at the variance inflation factor table below, we can clearly see that none of the explanatory variables have a large VIF and therefore we can ignore the multicollinearity in the model.**

Table

Description automatically generated

**We also took care of the dummy variable trap by removing Midwest because otherwise we would have variables that can be expressed as a linear combination of one another causing coefficients to be unstable and standard errors to inflate.**

**We can interpret our results as follows:**

**Holding all other variables constant, a 1 unit increase in the number of assaults per 100,000 people will cause the murders to increase by 0.0225 per 100,000 people. This is statistically significant at the 1% level as there is a 0% chance of observing a coefficient value of 0.0225 or larger has the true effect of coefficient of assault been zero.**

**Holding all other variables constant, a 1 unit increase in the number of rape cases per 100,000 people will cause the murders to increase by 0.1875 per 100,000 people. This is statistically significant at the 5% level as indicated by the p value as there is a 0.2% chance of observing a coefficient value of 0. 1875 or larger has the true effect of coefficient of assault been zero.**

**Holding all other variables constant, the effect of living in a Southern state means that, on average, the murder rate is 3.15 people per 100,000 more than if a person were to live in a midwestern state. This is statistically significant at the 5% level as indicated by the p value as there is a 0.2% chance of observing a coefficient value of 3.1577 or larger has the true effect of coefficient of assault been zero.**

**Holding all other variables constant, the effect of living in a Western state means that, on average, the murder rate is 2.05 people per 100,000 less than if a person were to live in a midwestern state. This is statistically significant at the 5% level as indicated by the p value as there is a 3% chance of observing a coefficient value of 3.1577 or larger has the true effect of coefficient of assault been zero.**

# ENSURE THIS DOCUMENT IS NEAT AND ADD IT TO YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : Zafar Hassan**

