5. Summarize your results (in the file you made above or a separate document) based on the output of the model. Be sure to describe the relationship between the independent and dependent variable and your interpretation of its significance.

From the OLS summary above, it shows that the independent variable (urban_population) has a positive linear relationship with the dependent variable (tax_percent_gdp). From the coefficient for urban population (0.0887) this means for every one percent increase in the urban population, this would result in 0.0887 increase in the tax percent GDP.

The R squared of 0.189 in the model means that the model explains approximately 18.9% percent of the variation in the dependent variable (tax_percent_gdp). From the information available this is not a good predictive model.

With regards to the independent variable (urban_populaition), the P value is 0.0, meaning that the independent variable is statistically significant. This is because in the T- test performed, the null hypothesis is that the coefficient is zero and thus not helping the model. But in this case we reject the null hypothesis since the P value is less than 0.05 and this means the probability of the coefficient of the independent variable being essentially zero is very small. In relation to the data this means that the urban population has a significant positive linear relation with the tax percent GDP,so increase in the urban population would give rise to increase in the tax percent GDP.

Also the residual plot shows the spread is biased and heteroscedastic, this is because there is a systematic change in the spread of the residuals over the range of measured value as seen in the residual vs urban_population plot above.

6. Answer the free-form analysis questions below. Note that there are not inherently right or wrong answers; the questions are meant for you to showcase your ability to interpret and understand data:

a. What were some challenges you encountered in generating the dataset for analysis, if any?

Firstly, I had to derive the exact description of the columns to understand completely the dataset. Cleaning the dataset was quite straightforward except for some “%” symbol present in some of the rows in the urban population colum

b. Based on this initial model, what steps might you take next to validate or extend your analysis, if any?

I’d look to add more features to the model that would lead to a better R square, accounting for more variance in the dependent variable (tax_percent_gdb).

Also, I'll consider using a different model for this dataset. OLS assumes that the target variable is normally distributed which is not in our case, where the distribution of tax percent GDP is Right skewed. Using OLS for a target variable that is a percentage (proportions) which is bounded between 0-1 (0-100) would lead to fitted values that may be impossible(less than zero or greater than a hundred) hence predicted values that are also impossible. This is also possibly the reason for heteroscedasticity seen in the residual plot, which is not in agreement with the assumptions(that the residual plot be unbiased and homoscedastic) of an ordinary least square regression.
I'll recommend a binomial Generalized Linear Model like logistic regression.

c. If you were asked to study this relationship, how would you convey your findings?

For every one percent increase in the urban population, this would result in 0.0887 increase in the tax percent GDP.