ASR Analytics: Data Analysis Assessment

Overview:

This exercise is meant to be an assessment of both your skills as a data scientist and your ability to convey the findings of your analysis.

You will be assessed on the following parameters:

- Did the candidate complete the exercise as per the instructions given?
- Can a reviewer replicate the candidate's results?
- Did the candidate demonstrate a clear understanding of the data they were analyzing?
- Was the candidate able to explain the results of the analysis?
- Does the candidate's explanation demonstrate clear insight and understanding of analytical concepts?
- Does the candidate demonstrate an ability to convey technical information to a broad audience?

While most of the analytical work at our firm is conducted in Python, R, and SQL, you may use whatever language (or combination of languages) you feel most comfortable with to solve the problem.

Instructions:

Attached are the following CSV files (Original Source: Our World In Data):

- *share-of-population-urban.csv*
- *taxes-on-incomes-of-individuals-and-corporations-gdp.csv*

With the given data, please complete the following tasks:

1. Merge the two datasets such that the resultant dataset contains **only the intersection of the rows present in both files**.
2. Export the results to a new file called *merged_social_data.csv*.
3. Using the merged results, generate an ordinary least squares regression showing the effect of the independent variable *Urban_Population* on the dependent variable *Tax_Percent_GDP*. Ensure that your results contain at a minimum:
    a. The R squared of the results
    b. The *t*-statistic and *p*-value of the coefficient and intercept
    c. Degrees of freedom
    d. The spread of the residuals.
4. Output your results. This can be in a multitude of formats:
    a. Jupyter Notebook
    b. R Markdown
    c. PDF/PS
    d. Other formats are acceptable, as long as they contain all of the information requested in task 3 above, and enable an independent reviewer to reproduce the steps used to generate the output.
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.
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?
    b. Based on this initial model, what steps might you take next to validate or extend your analysis, if any?
    c. If you were asked to study this relationship, how would you convey your findings?
7. Put the following materials into a zip file named DSC_YOURLASTNAME_YOURFIRSTNAME.zip
    a. The merged data set
    b. The results/output from task 4
    c. All code used to generate your data set and results
    d. Your answers to the free-form questions
8. E-mail this zip file to your reviewer.