JSC270 Winter 2024 Assignment 2
Exploring income and its correlated factors in 1994 America

In 1995 Kohavi and Becker extracted and cleaned this data from the 1994 US Census. It went on to be used as a benchmark dataset for many papers advancing statistical and machine learning methods. Here, we will work with this dataset to explore patterns of income and consider how time and location may alter or confound these relationships.

This assignment is to be completed individually. To submit this assignment, add the following two files to Quercus by the due date:

1. A pdf of the answers to the data analysis questions in **Part II** that includes a link to your github repo which contains the python notebook with the solutions (see **Part I**). Include the link to your repo at the top of the first page.
2. A maximum 2-page pdf report on your own linear regression analysis with the census dataset. Details are described in **Part III**.

PART I - Github (5 pts)

You will be learning about Git + GitHub in the lab this week and next. If you are unfamiliar with Git + GitHub - no problem! You can start Part II and come back to this later.

1. Create a **private** repository on your GitHub account called *JSC270_HW2_2022_<yourfirstinitiallastname>*. See instructions here.

2. Invite *UofT-JSC270* to be a collaborator of the repo from your Github account. See instructions here.

3. Create a colab (or other python notebook) and add it to your github repo. Make sure to include a link to colab at the beginning of your notebook.

4. Feel free to copy the starter code from this repo to begin the analysis.

5. Make sure your final set of solutions is in the notebook on your repo by the due date.

Note: You won't be graded on your commit history, but we highly recommend you use the assignment as a way to get practice with version control.

PART II - Data Analysis

Use section headers or comments in your notebook to indicate which pieces of code correspond to which question.

Initial data exploration (15 pts)

1. Check the columns of your data. Are they the expected data types based on their descriptions in this text file description of the data?

2. How are missing values represented in this data? Cast missing values to np.nan, if necessary. Count the number of missing values in each column.

3. Individually plot the distributions of *capital_gain* and *capital_loss*. Do you think these variables should be transformed to categorical variables? Why or why not? If yes, create a new variable(s) with your suggested transformation and plot or describe in a table the distribution of the new categorical variable(s).

4. The sampling weights in the dataset are contained in the variable *fnlwgt*. The weights indicate the share of the population that sample represents based on location (and sometimes, other factors). More information is provided in this text file description of the data.

   Plot or numerically explore the distribution of *fnlwgt*. Is the variable symmetrically distributed? Compare the distribution of this variable between men and women and comment on any trends you notice. Should outliers be excluded? If you think yes, set the *fnlwgt* values for those you deem to be outliers as missing for the remainder of your analyses.

Correlation. (10 pts)

1. Find the correlations between *age, education_num*, and *hours_per_week*.
   a. Do any of the variables appear to be correlated? How did you make your assessment?
   b. Statistically test any variable pairs with a correlation coefficient > |0.1| for its difference from 0 and report your result. Is the direction and significance of your finding as expected?
   c. How does the correlation (and its significance) between *education_num* and *age* compare between male and female participants? Is this expected?
   d. Compute the covariance matrix for *education_num* and *hours_per_week*. What conclusions can you draw from the covariance matrix?

Regression. (15 pts)

1. Fit a linear regression with *hours_per_week* as the dependent variable and *sex* as the independent variable.
    a. Do men tend to work more hours?
    b. Add *education_num* as a control variable, does the trend in hours worked by men vs women remain the same? Is the coefficient for *education_num* statistically significant? What is the 95% confidence interval?
    c. Now add *gross_income_group* as a binary variable in the model and compare this model with the models including (i) only *sex* and (ii) *sex* and *education_num*. Write down the interpretation for the coefficient for *sex* in each model. What statistic(s) can help to decide which model is the "best"? How do the three models compare?

PART III - Reporting on your own regression analysis (20 pts)

Next, think about a question you could answer using linear regression with the census dataset. Your analysis can build on the models from the previous question or you can devise an entirely new analysis. Write a (max 2 page) report providing:

1. A brief description of the dataset (eg. the dataset size, how it was obtained, where it was obtained from, why it was created, etc.)
2. The question you are interested in answering with your proposed linear regression model
3. A description of the model you fit (eg. which variables did you include and why)
4. Interpretation and description of the fitted model results using the concepts we learned in lecture and tutorial

Make sure the report is accessible for someone who doesn't already know the dataset you are working with.

Bonus question (2 pts): How does the estimate of the slope parameter in simple linear regression relate to the sample correlation coefficient? Show step-by-step calculations that demonstrate their relationship. Also, provide a written interpretation of the estimate of the slope in terms of the sample correlation coefficient.

See the definition of the sample correlation coefficient on the wiki page if you aren't familiar with it. You can also use the formulas provided in the lecture.

Answers can be submitted in any format (latex, word, image of written math) and included in the pdf containing the answers to **Part II** of this homework.