Assignment 1 – Part B

In this assignment, you will perform exploratory analysis of demographic (census) data from the United States. The goal is to compute statistics on the data, identify interesting correlations, and investigate hypotheses. You will also practice using Pandas for data analysis.

The assignment consists of the following steps. You should implement them primarily using methods from the Pandas library (along with your own functions applied using apply), as well as the plotting and statistical analysis libraries covered in the tutorials.

Step 0. The file you will work with is Census_Data.csv. The file contains the following data about residents of the United States:

- age: Integer representing age
- CoW (Category of Work): Type of employment (government employee, private employee, etc.)
- education: Education level
- marital: Marital status
- occupation: Occupation
- PoB (Place of Birth): Place of birth
- hours: Hours worked per week
- sex: Sex
- race: Race
- state: State of residence
- income: Annual income

The values of the attributes are currently numerically encoded. For example, the state of Alabama is represented by the number 1, and Male is represented by the value 1, etc. Before starting the analysis, convert the data into a format that is easier to read and present. You are provided with the file Attribute_Values.csv, which contains triples of the form (Attribute Name, Attribute Numeric Value, Attribute Value). For each attribute, it lists the different numerical values it can take and their corresponding string values. Use this file to modify the values in the DataFrame containing the original data as follows:

- CoW: Group the records into the following employment categories: Private Employee (1,2), Government Employee (3–5), Self-Employed (6,7), No pay (8), and Unemployed (9).
- education: Keep the numerical value, but you may also need to create a new column with education categories for Steps 5 and 6. Group categories 1–15 as no diploma, 16–17 as high-school, and 18–19 as post-high-school; keep the remaining categories as they are.
- marital: Replace the values with the corresponding marital status.
- occupation: Keep the numerical value, but you may need to replace it for Step 6. You may group occupations into different employment sectors (e.g., finance, education, etc.).

- PoB: Convert the values into US (numerical values between 1 and 99, excluding 60) and Not US (values greater than 99 and the value 60).
- sex: Replace the values with the corresponding sex.
- race: Replace the values with the corresponding race.
- state: Replace the values with the state names.

Use Pandas functions to transform the values. It will be useful to load the Attribute_Values file into a DataFrame and perform some transformations using operations between DataFrames.

The result of Step 0 is provided in the file Census_Data_cleaned.csv as a reference.

Step 1. In this step, we want to understand the distribution of income. Create the following plots:

- Create a histogram of income with 100 bins using a built-in function from the Pandas library.
- Create a new column, log_income, containing the **logarithm** of income, and create a histogram with 100 bins using the Pandas library again. To handle zero incomes, use log(1 + income).
- Create a histogram with bins whose widths increase exponentially (i.e., double). The first bin should have width 5,000 (incomes in [0, 5,000]), the second width 10,000 (incomes in [5,000, 15,000]), the third width 20,000 (incomes in [15,000, 35,000]), and so on. Create a point plot of the number of records in each bin against the right endpoint of the bin, using logarithmic scale on both axes.
- Compute the cumulative frequency vector. For each income value (from smallest to largest), the vector stores the number of records with **at least** that income (e.g., for the smallest income, this is all records). Plot the cumulative frequency as a function of income, using logarithmic scale on both axes.
- Create the Zipf plot for income. In a Zipf plot, the Y-axis contains the values (income in this case), and the X-axis contains the rank of the values. For example, the highest income has rank 1, the second-highest has rank 2, and so on. The rank should be assigned to individual records, not to distinct income values. Use logarithmic scale on both axes.

Present all your plots in a single grid and comment on the distribution.

Note: For steps 3–5, you may implement the computations by converting the data to lists.

Step 2. In this step, create a new attribute, income_category, derived from income, which takes a binary value of 0 or 1 depending on whether income is below 50,000 or not. You will investigate whether there is an association between race and income_category. Create the contingency table for these two variables (you may represent it as a heatmap). Use the chi-square test to determine whether there is a statistically significant association. Evaluate the results by examining the p-value. Also present the contingency table of the expected frequencies (provided by the chi-square test) and inspect individual pairs of values with a large lift (the ratio of the observed frequency for a pair to its expected frequency). Comment on the results and explain the null hypothesis of the test.

Perform the same calculations for sex and income_category, using the appropriate statistical test as discussed in class. Explain the choice of test and comment on the results.

Step 3. In this step, you will examine whether there is a statistically significant difference in mean income between the two sexes. Perform the analysis for both income and log_income.

For this purpose, create a point plot (linestyle='none') showing the mean for each value of the attribute and the 95% confidence interval. First, try to draw conclusions from the plot. Then, use the appropriate statistical test to

examine whether there is a statistically significant difference between the mean incomes of the two groups. When choosing the test, take into account the observations from Step 1. Your plots should be presented in a 1×2 grid.

Explain your choice of test and describe your results and conclusions.

Step 4. In this step, you will examine the association between income and attributes that take continuous values.

In the first part, you will examine the association between hours and income. Create scatter plots of hours versus income and log_income for the state of California. Present your plots in a 1×2 grid.

For each pair, compute the Pearson Correlation Coefficient and the Spearman Rank Correlation Coefficient, along with the corresponding p-values. Comment on the results and on the differences between using income and log_income, and between Pearson and Spearman correlation.

In the second part, you will examine how income varies with education level. Create plots of the mean income and log_income as a function of education (for California), with 95% confidence intervals (use Seaborn's lineplot). Present your plots in a 1×2 grid. Comment on the results.

Step 5. Formulate your own hypothesis and investigate it using the data. Your hypothesis should be non-trivial and different from those examined in the previous steps. State the null hypothesis and test it using an appropriate statistical test.

Step 6. In this step, you will attempt to relate demographic data to crime data across U.S. states. First, search online for statistics on the violent crime rate by state, expressed as the number of crimes per 100,000 residents. Make sure that the crime statistics and the demographic data you use refer to the same time period (around 2020).

Next, select one of the demographic characteristics and compute an aggregate value for each state (e.g., the mean, the median, or the percentage of residents with a particular value). Formulate a hypothesis about the relationship between this characteristic and crime. Create plots to examine the relationship visually, and then test your hypothesis using an appropriate statistical test. Justify your choice of test and explain the null hypothesis. Comment on your results.

Report the source of the data you used and how you found it, and describe the data used in your experiments.

Note: If you examine multiple associations, take into account the issues that arise from multiple hypothesis testing.

Submission: Submit a Notebook containing:

- The code for each step, with the individual questions clearly separated using bold headers.
- The requested plots and tables.
- A brief discussion of the results immediately after each analysis.
- For each statistical test, the null and alternative hypotheses, the effect size, the significance level, the p-value, and the conclusion.
- A clear description of any choices or assumptions made during data processing.
- A description of the assistance you received from AI tools, web searches, or other sources.

Export the notebook to HTML and submit the HTML file as well.