The geographic area of context:

Washington State

Research question:

1. In the state of Washington, how do earnings vary by educational attainment?
2. Does the premium for higher education vary by race and ethnicity?

To answer the two research questions, we tried to understand the dataset and refer to the dictionary. Then we chose the following variables to remain in our research dataset:

‘SERIALNO','SPORDER': as the identifier of each row.

'WAGP'- Wages or salary income in past 12 months: as the dependent variable.

'SCHL'-educational attainment: as the independent variable.

'RAC1P'-recorded detailed race code: as the independent variable.

'ESR'- Employment status: as the reference for sample selection.

Sample selection criteria:

1. The data should be complete without missing values in the chosen columns.
2. The research object should be the people who are able to work and earn money.
3. The ratio in the sample should be the same in the population.

Rationale for the sample selection criteria:

Firstly, we need to ensure the integrity of the data in the sample. We have a large enough dataset of the population whose number of rows is 72383. If we run the whole dataset in R, it will return an error indicating the size of the data is too large to run. Thus we need to take a sample from the original dataset. Since the dataset is so large, we can simply abandon the rows with missing data and remain the ones whose values in all columns are complete and valid. After excluding the missing value rows, our dataset shrinks to 59113. In following analysis, we do not need to worry about the NA any more.

Secondly, we need to define the research object. We hope to explore the relationship between earnings and educational attainment. So we should not include the people who are still in education and could not make money due to the age limit. So we exclude the rows whose “ESR” value is NA. At the same time, the people who is not in labor force are not included in our research, either. They might be too old to work or physically disabled. These factors are not of our interest, so we decided to exclude this group in our study. Thus the rows whose “ESR” value is 6 are eliminated too. Then our dataset now contains 35688 rows.

Thirdly, we need to draw a sample from the current dataset. Since our second research question is about the race and ethnicity, we should always keep the population’s characteristics in race and ethnicity in our sample. Thus we computed the ratio of each race and randomly choose the same ratio of each race from the dataset and the sample size is 10% of the valid dataset, which is 3570 rows.