Assignment # 1: Data Analysis Report
Earnings and Education

The data set described in the following report has been collected by the Bureau of Labor Statistics from a population of non-agricultural workers ages 25 through 64 and includes data on 55,899 individuals. This report includes numerical and graphical representations of the categorical variable education 'EDUC' and the quantitative variable earnings 'EARN' along with brief evaluations for both. Because this report involves analysis of a specified set of material, remaining independent from the data collection, it is impossible to know if the sample was taken randomly. Therefore any interpretation and description of the variables EARN and EDUC refers to the sample set alone and cannot be generalized for the population as a whole without more information on how the data was collected.

The quantitative variable, EARN, can be described as a unimodal distribution (only a slight second peak around $230,000) with a strong skew to the right (shown in Figure 1). This skew depicts that there are a select number of subjects with earnings falling considerably above the majority of the distribution. The mean amount of earnings for this distribution is $37,865 with a standard deviation of $36,158. The minimum earning level is negative at -$24,998 most likely representing an individual who ended his/her year in debt. The Maximum data value, corresponding to the highest earnings of any worker in the population, is $425,510. The first quartile is $17,000 and the third quartile is $46,505. Therefore 25% of the individuals in this sample have earnings lower than $17,000 and 25% of the workers make more than $46,505. The remaining 50% of workers have earnings between the first and third quartile. Table 1 below gives the numerical summary for the data collected for earnings in this sample of workers.

Figure 1

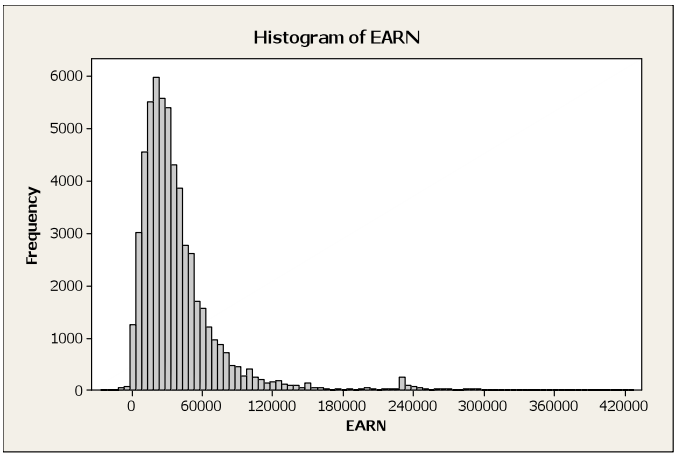


Table 1

Variable	Sample Size	Mean	Standard Deviation	Min	Q1	Median	Q3	Max
EARN	55,899	37,865	36,158	-24,998	17,000	29,717	46,505	425,510

The categorical variable, EDUC, has five possible groups, labeled numerically: EDUC 1 through 6. Group 'EDUC 1' represents workers who did not reach high school and includes 4% of the sample population. Group 'EDUC 2' represents workers who went to high school but did not get a diploma including 6.5% of the sample population. Group 'EDUC 3' represents workers with a high school diploma at 32% of the sample population. Group 'EDUC 4' represents some college without a degree including 28% of the sample population. Group 'EDUC 5' represents bachelor's degrees including 19.5% of the sample population. And lastly, Group 'EDUC 6' represents workers with postgraduate degrees including 10% of the sample population. Figure 2 gives a visual representation of how many individuals fall into a given category relative to others, while figure 3 provides a visual presentation of each groups' contribution to the whole sample population.

Figure 2

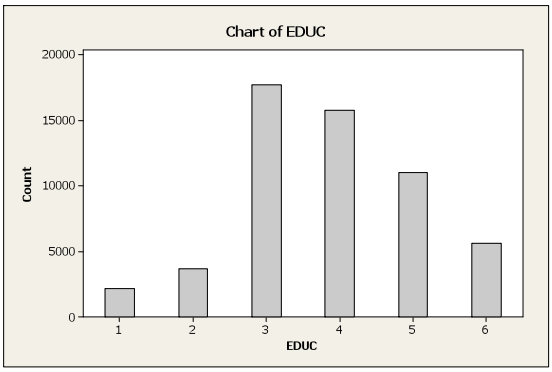
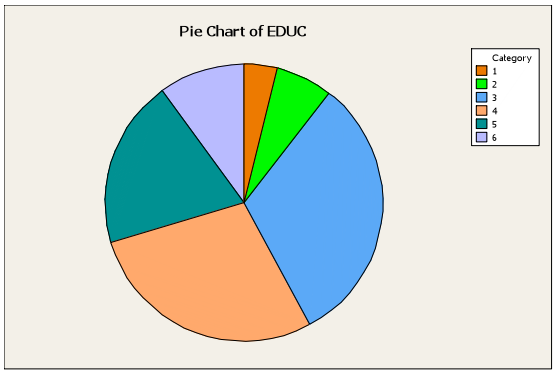


Figure 3



Assuming a perfect random sampling, this data would provide a statistical prediction describing the total working population ages 25-65. The data on earnings shows that the average of this population of workers will be an earning of around $37,865. The skew to the right shows that a very small proportion of the whole population would be making considerably more than the rest of the population. Assuming random sampling, the data on earnings when applied to the whole population allows us to predict percentages of each group for the variable education. For example, group 1 for the entire population should be roughly 4%, group 2 should be roughly 6.5%, and so on. Data collected for smaller population sizes can be analyzed in depth then generalized to the population as a whole when a random sampling has been achieved. This example for the data on workers shows how a very large data set can be interpreted and numerically represented in a brief and concise manner.