Capstone Project: Statistical Analysis

The first step in analyzing the data in my capstone dataset was to look at the visualizations and observations that I had made from those visualizations, to determine what factors were most effective at determining financial need. Factors that stuck out to me in this analysis were education level, floor type, and ownership of a computer and television. Based on the visualizations it appeared as though these were factors that were strongly determinant of income level. I decided to conduct a statistical analysis to look for the pearson coefficient for between these factors and the income level. For the floor type and education level groups, there were many different values that varied in their representation in the dataset and in their apparent effectiveness at determining the target value. As a result, I chose to only use the factors for these categories that showed a visual indication of correlation and that had at least 20% of the data as a one in at least one of the four aid categories. I ran a pearson correlation analysis on the following columns, “no education”, “incomplete primary education”, “undergraduate higher education”, “mosaic, ceramic, or terrazo floor”, “cement floor”, “owning a computer”, and “owning a television” as they relate to the financial need. Of these variables “no education”, “incomplete primary education”, and “cement floor” had a negative correlation with the need category. As a result, I can conclude that people who are in any of these categories are more likely to be in more need. All of the remaining variables that were tested had a positive correlation with the need category. Meaning that anyone in these categories is more likely to be in less need. I then used bootstrap replicates to resample the data, to ensure that any correlation found was not due to chance. The p value received from all the bootstrap replicates was zero or very close to zero, so it can be safely assumed that the correlation was not due to chance.