Danielle Senechal

MAT-374: Data Analytics

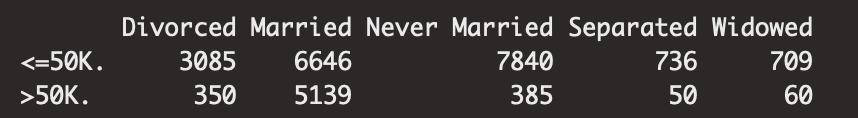
Project 1

March 10, 2020

1. Summary of variables:
   1. There are 1,399 missing variables in the work class variable. This is the only category with missing variables
   2. The range of the age variable is between 17 and 90 years old. The mean of this variable is about 39 years old, and the median is 37 years old.
   3. The mode for the work class variable shows that the majority of adults are private workers, with 17,385 adults be private workers. For education, the mode is 8120, with most adults being high school graduates. The majority of adults are also married, with 11,785 married adults. Lastly, the majority of adults (19,016 adults) make greater than or equal to $50,000 as their income.
   4. The “typical” person in the data set is around the age of 37-39 and is most likely a private worker. They are likely to be a high school graduate with an income of at least $50,000. It is also likely this “typical” person is also married.
2. There are no outliers with a z-score less than -3, but there are 100 outliers with a z-score above 3. The ages of the adults who fall into the z-scores above 3 are:

90 80 81 90 88 90 90 80 90 81 82 81 80 83 90 90 81 90 90 80 90 90 84 90 80 90 81 83 84 81 85 82 80 90 90 90 84 80 90 90 84 90 90 90 90 82 81 90 84 81 82 81 80 90 80 84 82 90 84 90 83 81 80 80 80 90 90 80 90 90 81 83 82 90 90 81 80 80 90 80 82 85 80 90 81 80 81 82 88 90 82 88 84 83 86 90 90 82 83 81

1. Contingency table of income and marital status:



Percentage table:

A screenshot of a cell phone

Description automatically generated

1. Marital status and income (stacked):

A screenshot of a cell phone

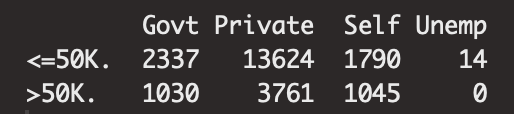
Description automatically generated

Marital status and income (normalized):

A screenshot of a cell phone

Description automatically generated

1. Of the people who are divorced, 89.81% (3085) have incomes less than or equal to $50,000, while 10.19% (350) make more than $50,000. For those who are married, 56.39% (6646) have lower incomes, and 43.61% (5139) have higher incomes. Of those who have never been married, 95.32% (7840) have lower incomes, and 4.68% (385) have higher incomes. In the separated category, 93.64% (736) have low incomes, and 6.36% (50) have higher incomes. In the last category, those who are widowed, 92.2% (709) have low incomes and 7.8% (60) have higher incomes.
2. Marital status is a decent predictor of income. This is because of those who are married, only 56.39% have lower incomes. In the other categories, the likelihood of having a lower income is usually around 90%. When comparing married versus non-married, this is a good predictor, but when comparing the other categories (divorced, separated, never married, and widowed) to each other, this is not a good predictor. This is because these categories all have very similar percentages to each other.
3. Skipped
4. Contingency table of income and work class:



Percentage table:

A screenshot of a cell phone

Description automatically generated

1. Income and work class (stacked):

A screenshot of a cell phone

Description automatically generated

Income and work class (normalized):

A screenshot of a cell phone

Description automatically generated

1. Of the people who are government employees, 69.4% (2337) have lower incomes and 30.59% (1030) have higher incomes. 78.37% (13624) of private employees have lower incomes, and 21.63% (3761) are in the high-income bracket. Of those who are self-employed, 63.14% (1790) have lower incomes and 36.86% (1045) have higher incomes. 100% (14) of unemployed people make less than or equal to $50,000.
2. This is not a good predictor of income, because the percentages are all similar to each other. Although, it is a good predictor when comparing unemployed versus employed, because 100% of people who are unemployed make less than or equal to $50,000, while those who are employed have a range between 21.63% and 36.86% of having higher income.
3. Age and income (stacked):

A close up of a map

Description automatically generated

Age and income (normalized):

A picture containing screenshot

Description automatically generated

1. Across the age variable in the stacked version, there is a right skewed unimodal trend. In the normalized version of this graph, the graph is more symmetrical, but is still unimodal.
2. The typical age range for a person with higher income is about 40-50 years old. They are most likely to be married, since married individuals tend to have the higher incomes. They are also most likely to be self-employed.
3. For those who are in the lower income bracket, they are most likely to be 20 to 30 years old. They are also most likely to be never married, and unemployed.
4. To go about this, I would separate out those who are in their 20s, unemployed, and have never been married, and the recombine them to make a lower income group. I would then separate those in their 40s, who are married, and are self-employed, recombine them, and make that the higher income group.