Giovanni Villasenor

Midterms 607

1. Age: Numerical variable representing the ages of individuals in the dataset.

Age column ranges from 23 yrs to 130 years on the initial review of the dataset. The data age variable is currently right skewed (shown w the bar plot) with a lot of outliers shown in the box plot.

A screenshot of a computer

Description automatically generated

A graph of age distribution

Description automatically generated

A diagram of a box plot

Description automatically generated

Since the Age column is skewed, it is best to impute the missing value with the median. By imputing the data with the median, we capped the Age column based on the upper and the lower bound of the age using the IQR method. This made our Age column more normally distributed and have removed the outliers. The standard deviation also was dropped from 14 to 9.

A screenshot of a computer

Description automatically generated

A graph of age distribution

Description automatically generated

A diagram of a box plot

Description automatically generated

1. gender: Categorical variable with two possible categories of Male and Female.

Frequency distribution for 'gender' before imputation has 52% female compared to 48% males:

A graph of a number of blue rectangular bars

Description automatically generated with medium confidence

A graph of a person and person

Description automatically generated

1. smoking\_habit: Categorical variable with three possible categories:
   1. Non-Smoker
   2. Light Smoker
   3. Heavy Smoker

A graph of smoking habit distribution

Description automatically generated

A graph of smoking habit

Description automatically generated

There is relatively no structure change with the data after imputation of the missing values for smoking habits.

1. work\_out\_habit: Categorical variable with three possible categories
   1. Blank/None/Nan
   2. Regularly
   3. Occasionally

A graph of a number of blue bars

Description automatically generated

A graph of blue bars

Description automatically generated

There is relatively no structure change with the data after imputation of the missing values for work out habits.

1. heart\_attack: Categorical variable with two possible categories:
   1. (No heart attack)
   2. 1 (Had a heart attack)

A graph of heart attack distribution

Description automatically generated

A graph with blue rectangles

Description automatically generated

There is relatively no structure change with the data after imputation of the missing values for heart attack.

1. salary: Numerical variable representing the annual salary in dollars.

Salary before imputations and outliers:

Initial review of the salary variable shows right skew with salaries ranging from $87481 to $376253.

A screenshot of a computer

Description automatically generated

A graph of a salary distribution

Description automatically generated

A diagram of a box plot

Description automatically generated

Salary also is skewed to justify the imputation method of using the median rather than the mean. We decided to not remove the outliers due to the dataset having a high kurtosis. After imputation, the structure of the dataset remained relatively the same.A screenshot of a computer

Description automatically generatedA graph of salary distribution

Description automatically generatedA diagram of a box plot

Description automatically generated

1. education: Categorical variable with four possible categories:
   1. High School
   2. Bachelor's
   3. Master's
   4. PhD

A graph of a number of blue rectangular bars

Description automatically generated

There is relatively no structure change with the data after imputation of the missing values for Education.

A graph of a number of people

Description automatically generated with medium confidence