Data Mining
HW 2

Omid Naeej Nejad
610301189

October 18, 2024

1 Exercise 1

Replacing the missing value with the field mean (for numeric variables) or the mode (for categorical variables) is most likely to result in an underestimation of the variable's dispersion. Suppose many missing values are in the data set, replacing them by mean would increase density around the mean. It means that measures of spread (like standard deviation) will be artificially reduced.

Advantages of Mean Imputation (Mode for categorical predictors):

1. Simplicity: Easy to implement, understand, and automate.
2. Maintains Overall Mean(/ Mode): The mean(/ Mode) of the data set remains unchanged.
3. Maintains Sample Size: No loss of data due to missing values.

2 Exercise 2

- **A:** 145 is outlier.
- **B:** Verify outlier using the Z-score method:
  A data value is an outlier if it has a Z-score that is either less than -3 or greater than 3.

$|z| > 3 \rightarrow |\frac{x-\bar{x}}{sd}| > 3 \rightarrow |x - \bar{x}| > 3sd$
$Mean : \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{348}{12} = 29$
$Variance : \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = 1300$
$SD : sd = \sqrt{Var} = 36.05$

$|x - \bar{x}| > 3sd \rightarrow |x - \bar{x}| > 108.16 \rightarrow x > 137.16$ or $x < -79.16 \Rightarrow$ **145 is outlier.**

- **C:** Verify outlier using the IQR method: robust statistical methods for outlier detection, which are less sensitive to the presence of the outliers themselves.
  A data value is an outlier if it is located 1.5(IQR) or more below Q1 or It is located 1.5(IQR) or more above Q3.
  sort data points $\rightarrow 3, 9, 11, 13, 15, 19, 19, 23, 25, 31, 35, 145 \rightarrow$
  $Q1 = \frac{11+13}{2} = 12$, $Q2 = 19$, $Q3 = \frac{25+31}{2} = 28$
  $IQR = Q3 - Q1 = 16 \rightarrow x \leq Q1 - 1.5IQR$ or $x \geq Q3 + 1.5IQR \rightarrow$
  $x \leq -12$ or $x \geq 52 \Rightarrow$ **145 is outlier.**

3 Exercise 3

- **A:** with the outlier:
  $Mean : \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{348}{12} = 29$
  $Median : Q2 = 19 \leftarrow 3, 9, 11, 13, 15, \underline{19, 19}, 23, 25, 31, 35, 145$

  without the outlier:
  $Mean : \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{203}{11} = 18.45$
  $Median : Q2 = 19 \leftarrow 3, 9, 11, 13, 15, \underline{19}, 19, 23, 25, 31, 35$

- **B: An outlier affects the mean more than the median.**

  The mean is calculated by adding up all values and dividing by the number of values. Outliers, which are extreme values, can significantly alter this sum.In other words, If an outlier is extremely high or low, it can pull the mean towards it.

The median is the middle value when data is ordered. Because it depends on position rather than magnitude, a few outliers (even until a half of data) won't change the median much.

4 Exercise 4

- **A:** Using equal width binning, we partition X into the following categories of equal width:
  First: $3 \leq x < 38.75$, which contains all the data values except one.
  Second: $38.75 \leq x < 74.5$, which contains no data values at all.
  Third: $74.5 \leq x < 110.25$, which contains no data values at all.
  Fourth: $110.25 \leq x < 146$, which contains a single outlier.

- **B:** Using equal frequency binning, we have n = 12, k = 3, and n/k = 4.
  Low: X = {3,9,11,13}
  Medium: X = {15,19,19,23}
  High: X = {25,31,35,145}

- **C:** Both equal width binning and equal frequency binning have their own limitations, which can make them less effective in certain scenarios.

  Equal Width Binning is easy to implement and understand but it creates bins of uniform size, which can leave many bins empty and others overpopulated if your data isn't evenly distributed. In other words, ***Equal Width Binning is not efficient with Skewed Data.***

  In Equal Frequency Binning each bin contains approximately the same number of records, which can be useful for visualizing and analyzing data distributions but it's not the best method for binning data because of:

  1. ***Irregular Intervals:*** **If the data is unevenly distributed, this can lead to intervals that don't make sense**. For instance, one bin might span a wide range while another only spans a narrow range. Outlier and others might end up together in one bin, making interpretation and comparison more difficult.

2. ***Similar data in different bins!*** Suppose there are 5 equal data points but $\frac{n}{k} < 5$, so they are not in a similar bin that is not clearly logical.

Therefore, both of the above methods are not the most effective solution.