

2023-24

CMSE11432 Principles of Data Analytics

Individual coursework

A Critical Reflection on the Use of IQR Method and Box-Whisker Plots in Outlier Detection of Human Resource Datasets

B244512

Word count:

785

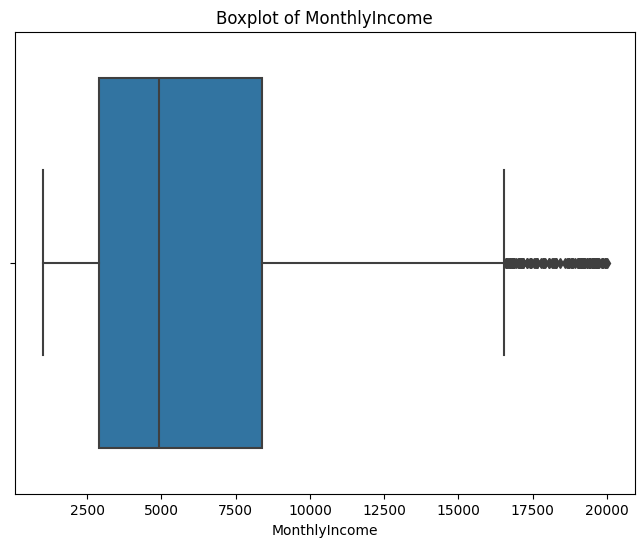
## A Critical Reflection on the Use of IQR Method and Box-Whisker Plots in Outlier Detection of Human Resource Datasets

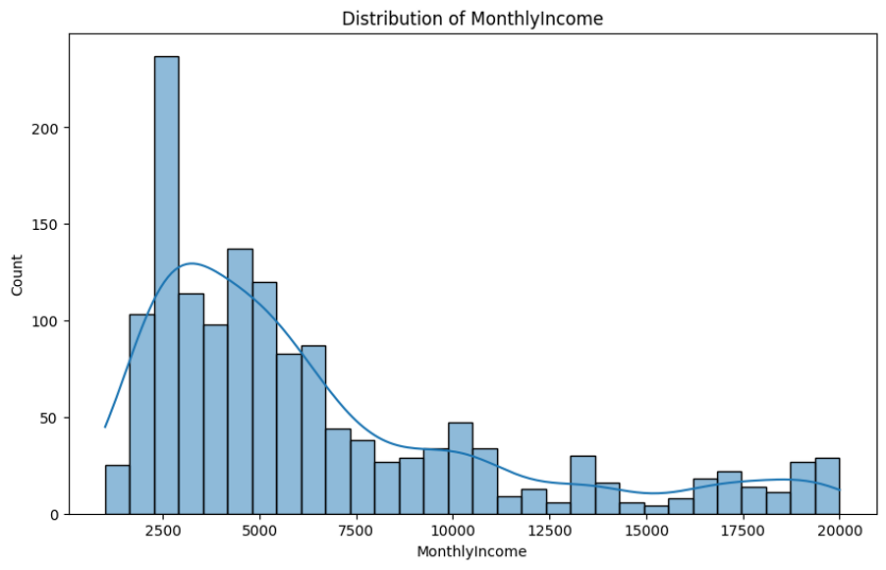
## **Introduction**

**Outliers are observations that are located far away from the rest major data points of a dataset** **impairing model performance. However, in the previous group assignment aimed at providing Human Resource (HR) management insights, as for the target variable “Monthly-Income”, removing outliers beyond 1.5 IQR (Interquartile range) in box-whisker plots resulted in a lower R2 value for my regression model. This piqued my interest in exploring the underlying reasons. In the following content, I will reflect on my use of the IQR method and box-whisker plots by combining HR dataset characteristics and comparing IQR with other outliers-detection methods, then conclude with** **their strengths and limitations and reasons for the poor performance.**

## Background

**Box-whisker plots display the 25th, 50th, and 75th percentiles, forming a box, and whiskers extending from its ends (Krzywinski and Altman, 2014, Mishra et al., 2019). They indicate whether the values fall within the IQR when measured from either end of the box. Data points outside the range are considered outliers, represented as dots beyond the whiskers (Figure 1).** In academic literature, box-whisker plots are considered convenient for identifying univariate outliers, showing their wide application in outlier detection (Mowbray et al., 2019). However, there are some scholarly debates about Box-whisker-plots’ applicability. For instance, Maximus inquired whether the IQR method remains effective for non-normally distributed data (Maximus, 2022). Given “Monthly-Income” data is right-skewed, it means that most of the data points are clustered at the lower end and a few higher values extend to the right, creating the long tail in contrast (Figure 2), this inquiry may lead to a breakthrough in explaining the poor model performance after outlier handling.

**Figure 1.** Box-Whisker Plot of Monthly Income

****

**Figure 2.** Distribution of Monthly Income

## Analysis of the Testing Result of Outlier Detection Approaches

**To diagnose IQR’s weaknesses, I tested the performance of other outlier-detection methods on handling “Monthly-Income” and then made a comparison. By combining extracurricular knowledge with methods taught in lectures, I came up with the following list of outlier-detection approaches: Z-score, Robust Z-score, IQR method, Winsorization method, DBSCAN Clustering, Isolation Forest (Bhat, 2021).**

**The results from applying these outlier-detection methods to “Monthly-Income” are listed below:**

**Table 1.** Outlier Detection Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Initial R2** | **Number of Outliers Detected** | **R2 After Removal or Replacement** |
| Z-score | 0.9433 | 0 | 0.9433 |
| Robust Z-score | 0.9433 | 118 | 0.8787 |
| IQR method (the method we used) | 0.9433 | 114 | 0.8823 |
| Winsorization method | 0.9433 | 146 | 0.9409 |
| DBSCAN Clustering | 0.9433 | 198 | 0.9324 |
| Isolation Forest | 0.9433 | 15 | 0.9385 |

We can easily notice a general decrease in model’s R2 after outlier handling, except for the Winsorization method, which sustained a higher R2. We don’t consider the Z-score although it had the highest R2 after removal, because it didn’t detect any outliers.

Next, I dived into Winsorization and further investigated its differences from IQR.

## IQR Method’s Strengths and Weaknesses When Compared with Winsorization

Winsorization is a statistical transformation that reduces the impact of possible outliers by limiting extreme values in statistical data (contributors, 2023). For example, a common practice is to replace values below the 5th percentile and above the 95th percentile with 5th and 95th percentile values. After comparing it with the IQR method, we can conclude that the IQR Method’s defects include:

### Potential Loss of Information

While IQR methods remove all the data points outside the range, Winsorization only replaces them with percentiles closer to central tendency. It avoids over-pruning and maintains the information on employees' salaries, ensuring the model has some potential explanatory capacity to extreme data points.

### Overemphasis on Central Tendency at the Expense of Variability

In HR datasets, wage disparities are important features because, in companies, senior or management employees are likely to have higher wages than the average employees due to organizational structure. Completely removing outliers may result in an underestimation of the variability, especially when the data points are right-skewed (Figure 2).

However, the IQR method still has strengths:

### Wide Adaptability

IQR method is non-parametric, so it does not require the data to follow any particular probability distribution, thus allowing them to be applied to any type of data, including those that are not normally distributed or contain outliers (Whitley and Ball, 2002).

### The Communicative Clarity When Achieved Through Box-Whisker Plots

For non-technical audiences, IQR method aligns well with Box-Whisker Plots, which visualise the distribution of the data, including medians, quartiles and outliers, making it easier to understand and interpret.

## Conclusion

In this article, we found that the IQR method, while broadly applicable, may remove vital information in right-skewed data, which is possible in HR research. A better way to handle outliers in this circumstance may be to replace them with values closer to the central tendency to reserve variability. However, the fact that methods like Winsorization, which replace extreme values instead of removing them, also failed to improve model performance (see Table 1), suggests that removing outliers is not always necessary for better model performance. Thus, it is essential to tailor our approach to outlier handling for optimal results contextually.

## References

BHAT, N. 2021. *Outlier!!! The Silent Killer* [Online]. Kaggle. Available: <https://www.kaggle.com/code/nareshbhat/outlier-the-silent-killer> [Accessed 6 November 2023].

CONTRIBUTORS, W. 2023. *Winsorizing* [Online]. Wikipedia, The Free Encyclopedia. Available:https://en.wikipedia.org/w/index.php?title=Winsorizing&oldid=1171707283 [Accessed 6 November 2023].

KRZYWINSKI, M. & ALTMAN, N. 2014. *Visualizing samples with box plots*. Nature methods*,* 11**,** 119-120.

KWAK, S. K. & KIM, J. H. 2017. *Statistical data preparation: management of missing values and outliers*. Korean journal of anesthesiology, 70, 407-411.

MAXIMUS. 2022. *Does IQR method for outliers work for non-normal data?* [Online]. Stack Exchange. Available: <https://stats.stackexchange.com/questions/580580/does-iqr-method-for-outliers-work-for-non-normal-data> [Accessed 6 November 2023].

MISHRA, P., PANDEY, C. M., SINGH, U., GUPTA, A., SAHU, C. & KESHRI, A. 2019. *Descriptive statistics and normality tests for statistical data*. Annals of cardiac anaesthesia*,* 22**,** 67.

MOWBRAY, F. I., FOX-WASYLYSHYN, S. M. & EL-MASRI, M. M. 2019. *Univariate outliers: a conceptual overview for the nurse researcher*. Canadian Journal of Nursing Research*,* 51**,** 31-37.

WHITLEY, E. & BALL, J. 2002. Statistics review 6: Nonparametric methods. *Critical care,* 6**,** 1-5.