

2023-24

CMSE11432 Principles of Data Analytics

Group coursework

A Reflective Essay on Dealing with Outliners in Human Resource Management Themed Datasets

B244512

Word count:

# A Reflective Essay on Dealing with Outliners in Human Resource Management Themed Datasets

## Introduction

**Outliers are observations that locate far away from the rest major data points of a dataset, which will probably cause poor performance of the predicting model. However, in previous group assignment aimed at providing Human Resource (HR) management insights, when I brutally removed the outliners identified by the Box-whisker-plots and the 1.5 Interquartile range (IQR), the performance of my linear regression model turned out to be worse with a lower R2 value. This phenomenon sparked my interests in further investigating the reasons behind this, and I’ll try my best to do a thorough analysis from perspectives including analytical approaches and their Limit**ations, the characteristics of HR datasets, etc then present you with summarizing conclusions.

## Analysis on the Testing Result of Outliers Detecting Approaches

### Testing Outliers Detecting Approaches

**By combining extracurricular knowledge and methods taught in lectures, I came up with the following list of outliers detecting(tackling) approaches:**

1. **Z-score**
2. **Robust Z-score**
3. **IQR method**
4. **Winsorization method**
5. **DBSCAN Clustering**
6. **Isolation Forest**

**I tested out all these approache**s to handle outliers on the dataset used in my group assignment, and the contrastive results reflecting model performance are listed below:

Figure 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Initial R2** | **Number of Outliers Detected** | **R2 After Removal or Replacement** |
| Z-score | 0.9376 | 0 | 0.9376 |
| Robust Z-score | 0.9376 | 0 | 0.9376 |
| IQR method | 0.9376 | 49 | 0.8694 |
| Winsorization method | 0.9376 | 28 | 0.9377 |
| DBSCAN Clustering | 0.9376 | 0 | 0.9376 |
| Isolation Forest | 0.9376 | 74 | 0.9350 |

We can easily notice two interesting things:

1. Z-score, Robust Z-score and DBSCAN Clustering failed to detect any outliers, while IQR method(the method we used in group assignment), Winsorization and Isolation Forest succeeded.
2. after processing the outliers, the model R2 values of most of the methods decreased compared to the pre-processing period, but the Winsorization method succeeded in increasing model’s R2 value.

### Comparation between Winsorization and IQR Method

Next, let’s focus on Winsorization and IQR method to campare the observations tagged as outliers by them and dig deeper into the reason for their different performances.

Winsorization

以下是根据箱线图（使用 I.Q.R 方法的边界值）被识别为异常值的前几行数据：

Age: 这些员工的年龄范围从 41 到 59 岁。

Department: 大多数员工来自研发部门，但也有来自销售和人力资源部门的员工。

MonthlyIncome: 这些员工的月收入都相对较高。

YearsAtCompany: 这些员工在公司的工作年限也相对较长，例如，有些员工在公司工作了超过 20 年。

分析：

这些异常值可能代表了公司中的高级员工或管理层员工，他们的工资、年资和其他福利可能都比平均员工要高。

这些异常值可能并不真的是“异常”，而是代表了公司的一个特定子群体。因此，直接移除这些异常值可能会导致我们丢失有价值的信息。

## Experience from the Group Assignment

Des**cribe how outliers were identified in the group assignment.**

**Reflect on the possible origins of these outliers and the specific impact they might have on your analysis res**ults.

## Choice and Application of Treatment Methods

Introduce specific methods or techniques chosen to handle outliers.

Reflect on the reasons for choosing these methods and the challenges and difficulties encountered when applying them.

Describe how the results after treatment differ from expectations and possible reasons for these differences.

## Lessons Learned

Reflect on the impact of this experience on your learning and future strategies for handling outliers.

Describe the different choices you might make if you encounter similar problems again.

## Conclusion

Sum up the importance of handling outliers and the deep insights gained from this experience regarding data analysis.

## How We Detected Outliers Group Assignment

### Basic Descriptive Statistics of the Dataset

**Our research question was to identify the features with the strongest correlation with Monthly Income, and build a linear regression model for companies to predict an appropriate wage for a certain employee based on his or her personal information. So, we focused on our target variable which is “Monthly Income” and only detected outliers within it. At this point, I’ve already found it problematic to not detect outliners by**

Used Box-whisker-plots and 1.5 Interquartile Range to Tackle Outliners in the