Title: Estimates of Location

Introduction

In this assignment there are calculations of estimates of location like Trimmed Mean, Weighted Mean and Weighted Median. Below will be shown how to find these estimates with examples. As the main dataset there were chosen business_15.csv file. In this file for about 100 records of employee data with 5 types of features: EmployeeId, Salary, HoursWorked, PerformanceWorked and Department.

Findings

Firstly, finding the mean of the feature Salary will be the most correct way. Summarizing all values and dividing it by the length, the mean is 73883.97108970863. So, the median is 61919.21252595773.

Looking at the dataset I decided that the weight feature for the salary column is the HoursWorked feature.

Trimmed Mean: In our dataset there are some outlier values especially on the end, so to trim them will be the right decision. Trim them, let's say for 5% from both sides. After calculation how many values need to trim, the trimmed mean of salary feature is 62322.09792493505

Weighted Mean: If trimmed mean is not enough there also will be able to find weighted mean. The weighted value for our salary feature is the HoursWorked column. The calculation is firstly just multiplying values itself to their weighted ones and summarizing them. Lastly divide this value to sum of weighted values. So, the weighted mean of the salary feature is 62694.4705479452. This value is less sensitive even if there are any outliers.

Weighted Median: And the least sensitive and the hardest to find among before ones is Weighted Median. There also firstly need to sort by Salary and then divide by two the sum of weighted values. After, finding cumulative values of each row and finding the closer value to our divided by two value. So, after before like calculations there found that weighted median salary is 61935.0

Conclusion

Summarizing my actions, it is clear that the values of simple mean and median are a bit different from weighted mean and median. It is because our datasets outliers not so much and if even exists, not so big difference between normal values. Though the weighted values are less sensitive and more robust and more clearer.