Situation:

Customers in six identified locations (A, B, C, D, E & F) are confused about what are costs of treating diseases in these locations. For this purpose, we need to provide customers with some insights into costs of treating various diseases. We propose to show them a table of costs for diseases by the location they have chosen. For this we need to prepare the data from a source.

Your Assignment:

What is the percentile you would choose as representative of the ‘safe maximum’ and leave out outliers which may skew the data?

You need to smooth this data and derive the median, third quartile and Xth percentile costs (which represents a safe maximum without considering outliers) for each location.

This would also entail cleaning out any data points which may be trivial (for instance all treatments which cost less than Rs. Y) What would this Rs Y be for each disease/ ailment?  Your output should look like.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Disease | City | Median | 3rd Quartile | Xth Percentile |
|  |  |  |  |  |

If you did this on a pan India basis, i.e. did not consider the city, what would be the error for residents of each city when they query this cost?

Can you obtain the morbidity statistics for the 6 identified diseases by age group and provide an assessment the likelihood of high risk groups (gender and age group combinations). What do you define as high risk?

Data for assignment:

You have been given data for a set of 6 diseases across 6 locations over here.