SMARTWATCH HEATH DATA ANALYSIS AND EXPLORATION

Name : Gauransh Garg

Enrollment No : 23114027

Data Set Type :

The data set being analyzed here has been taken from a 10000 smartwatches and contains information about the Heart Rate (bpm) , Blood Oxygen Level (%) , Sleep Duration (hours) , Activity Level , Steps taken and Stress Level of the individuals,

Report :

With the initial reference to the dataset , it had 10000 rows and 7 columns . The 1st column “User Id” does not have much of a impact on the data analysis and so has been removed from the data set to remove redundant attributes

Data set is cleaned using data cleaning techniques, like first checking how many rows are missing data .

These rows are then imputed with data , for numerical data mean is used for imputation while for categorical data mode is used for imputation.

The duplicates too are removed

The outliers are also removed by first checking their existence through box plots and then removing them by IQR method

The Categorical variable “Stress Level” also had a different data category “Verh High” which was removed as the no. of data elements which had this were few in number and all other rows had numbers to denote Stress Level

The Variable “Activity Level” also presented problems in the form that the same category was represented in different ways.
For eg: it had both actve and active for active,

Sedentary and sedentary for sedentary

These were also taken into account

Summary of the statistics and the skewness are provided for the numerical data. Frequency distribution of the categorical variables are also shown

Histogram and Box plots were also made for the numerical variables to show their distribution. It helped to understand what was the level of heart rate , blood oxygen level and other criterions that was most frequent in individuals.

Bivariate analysis was also done using heatmaps which were created for all pairs of numerical variables.

It was found through them that the “Heart Rate” and the “Steps count ” had the most correlation between them.

Box, Violin and Bar charts were also made to understand the relationship between the numerical and categorical variables .
It was revealed through them the relation between Stress Level and Step Count had the most number of outliers between them.

A scatter plot to study the relationship between Heart Rate and Sleep Duration (wrt to Step count) was also made (could taken as a multivariate analysis step [Grouped analysis] but performed under bivariate analysis)

Now for multivariate analysis , the large amount of data made it very difficult to plot pair plots and get any sort of distinction between them .

So the data was sampled to get clear distinction between the various kinds of plots.

Conclusion :

This analysis provided valuable insights into smartwatch health data, including data inconsistencies, relationships between health metrics, and behavioral trends among users.

Significant differences were observed between active and sedentary individuals in terms of step count and sleep duration.

The findings can be useful for further predictive modeling or personalized health recommendations.