

Vital Signs Diagnosis

Group 13

*SIMPAO, CHARIZE R.
TANYAG, LORD EXZEL
JHONNE L.*

INTRODUCTION

The report entails the exploratory data analysis of the dataset that holds a number of different health indicators and lifestyle factors of the cohort of several people. The data is obtained in file, 1 Vital signs diagnosis data.csv and contains patient details like age, sex, vital signs, and health condition and pattern indicators. The main goal of the given analysis is to comprehend the distributions of critical variables, discover their possible relationship with each other, and investigate the presence of any structures by reducing the dimensionality.

METHODOLOGY

The analysis involved:

- Data Cleaning: Removal of "Medication" column and rows with missing values.
- Summary Statistics: Calculation of min, max, mean, median, SD, and count for numeric variables.
- Visualizations:
 - Scatter Plot: Age vs. BMI.
 - Histograms: Distributions of numeric variables.
 - Boxplots: Spread and outliers of numeric variables.
 - Bar Plots: Frequencies of "Smoking_Status" and "Hypertension".
- Principal Component Analysis (PCA): Dimensionality reduction on scaled numeric variables, with a summary, biplot, and scatter plot of the first two principal components.

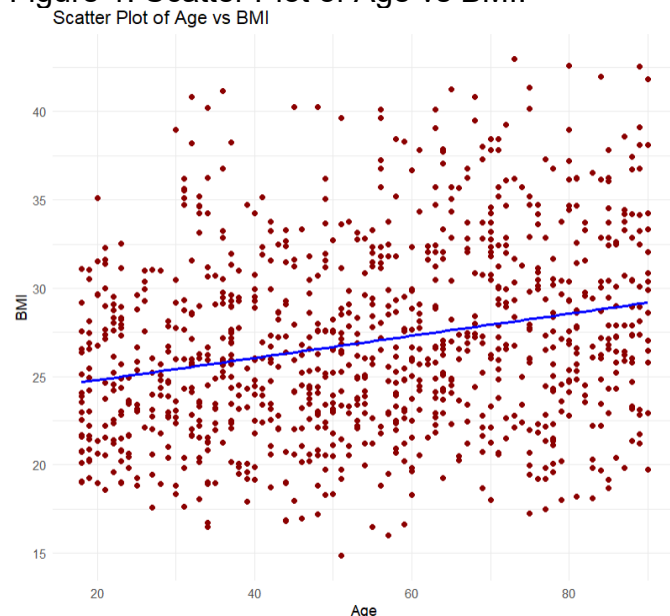
RESULTS AND FIGURES

Table 1. Summary Statistics for Numeric Variables after Data Cleaning.

Statistic	Patient.ID	Age	Sex	Weight_kg	Height_cm	BMI	Heart_rate	Smoking_Status
Min	1	18	0	40	130	14.86	50	0
Max	1000	90	1	86	175	42.96	154	2
Mean	502.03	54.01	0.49	62.39	153.3	26.93	100.15	1.01
Median	502	54	0	62	153	26.35	100	1
SD	284.97	21.12	0.5	9.16	11.89	5.3	19.16	0.82
Count	981	981	981	981	981	981	981	981

Physical_Activity Hours_Week	Stress_Level	Daily_Sleeping_hours	Glucose_mg.dL	Cholesterol_mg.dL
0	1	4	80	101
16	10	9	196	303
7.91	5.37	5.43	129.32	189.67
8	5	5	127	188
4.81	2.25	1.21	22.43	35.18
981	981	981	981	981

Figure 1. Scatter Plot of Age vs BMI.



The Age versus Body Mass Index (BMI) vote can be seen in figure 1. There seems to be a minor positive linear trend; however, this is an indication that BMI grows slightly over the years. But, the high dispersion of the data values show that there is a significant difference in BMI in all age groups and it is not just age that might determine the BMI of this group.

This scatter plot is a graph representing the Body Mass Index (BMI) and Age of the entire population of the data. Blue line means the fit of the linear regression and it expresses an overall trend in the two variables. Age and BMI measurement are on red dots of each person.

Figure 2. Histograms of Relevant Numeric Variables.

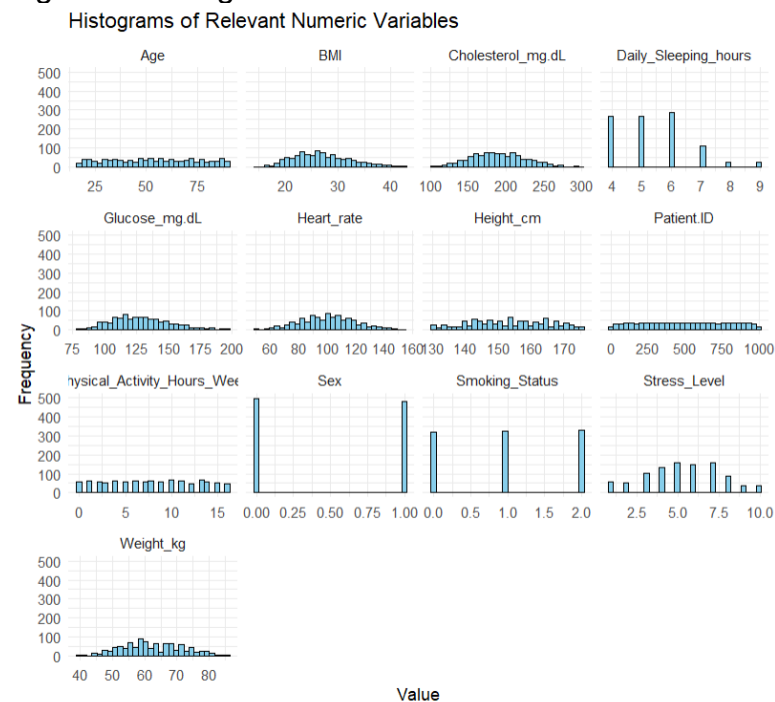


Figure 2 shows histograms of different numeric health and lifestyle variables. These plots demonstrate the distribution nature of each variable. As an example, in Daily_Sleeping_hours, clear peaks are observed which means preferred or popular sleeping times. The BMI shows a little with normal distribution whereas Cholesterol_mg.dL and Glucose_mg.dL take right skewed distribution.

Figure 3. Boxplots of Relevant Numeric Variables.

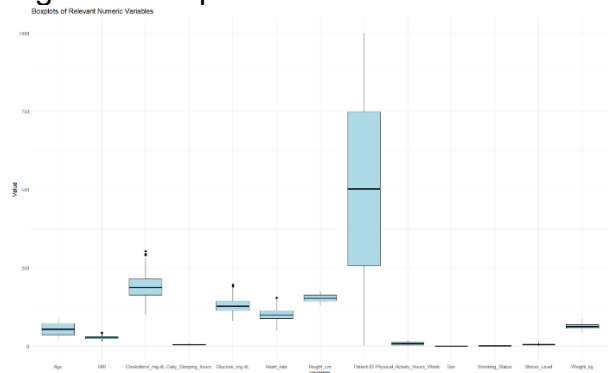


Figure 4. Bar Plot of Smoking Status.

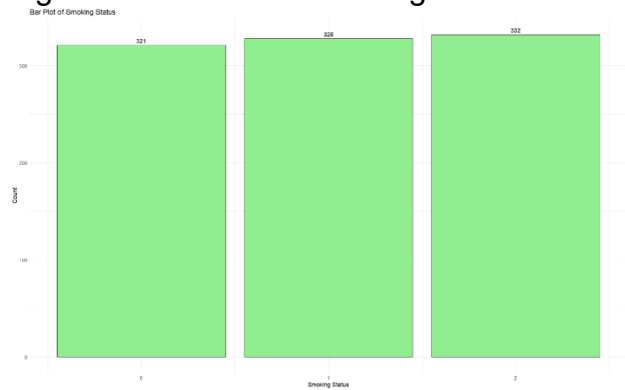
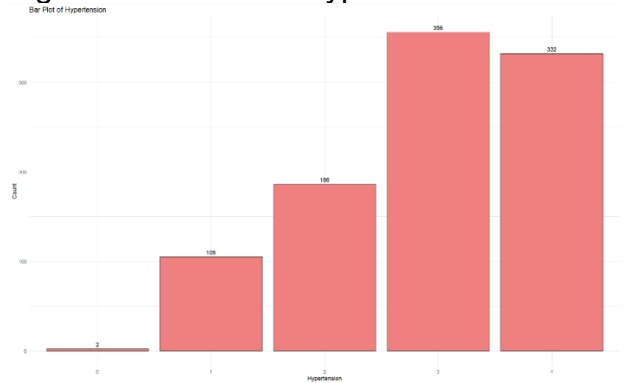


Figure 5. Bar Plot of Hypertension.



Principal Component Analysis (PCA)

Table 2. Importance of Principal Components.

This table gives the amount of variance explained by individual principal component (PC1-PC13), standard deviation, percentage of variance, and corresponding cumulative percentage.

Numeric variables were subjected to PCA to find out significant dimensions of variability. Table 2 shows the significance of each of the principal components (PC), standard deviation, percent of variance explained and cumulative percent of variance explained. Table 2 shows that PC1 explains 32.69 percent of the total variance and the first two principal components (PC1 and PC2) contribute 47.37 percent to the total variance. This implies that much of the variability of the data can be summarized to a smaller number of dimensions.

Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	2.0615	1.3816	1.2566	1.11991	1.05034	0.98384	0.87751	0.7758	0.70221	0.23958	0.10491	0.06028	0.0088
Proportion of Variance	0.3269	0.1468	0.1215	0.09648	0.08486	0.07446	0.05923	0.0463	0.03793	0.00442	0.00085	0.00028	0.00001
Cumulative Proportion	0.3269	0.4737	0.5952	0.69167	0.77653	0.85099	0.91022	0.9565	0.99445	0.99887	0.99971	0.99999	1

Figure 6. Variance Explained by Each Principal Component. This bar chart shows the percentage of total variance explained by each principal component.

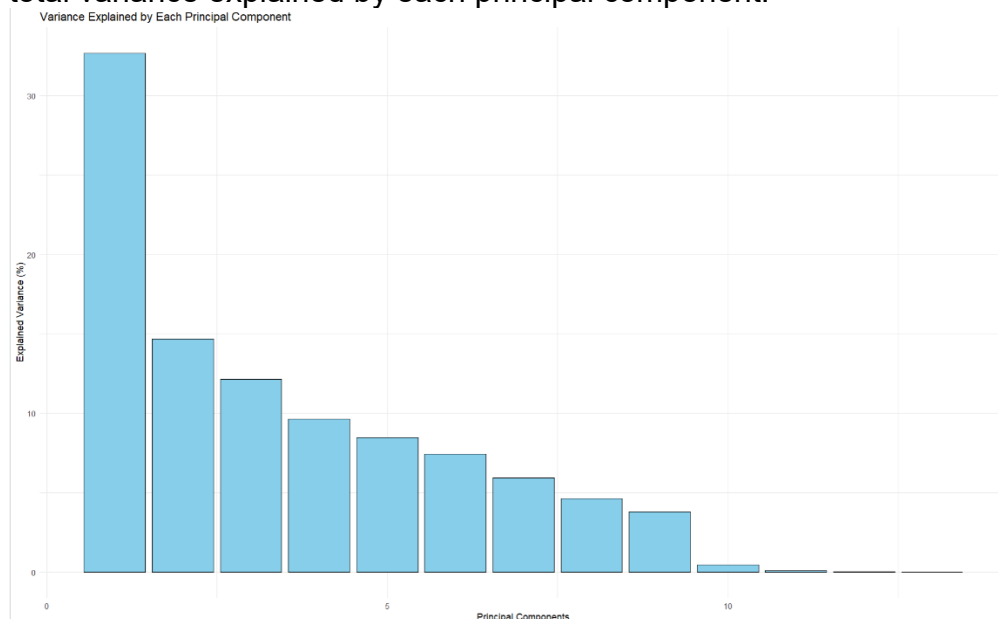
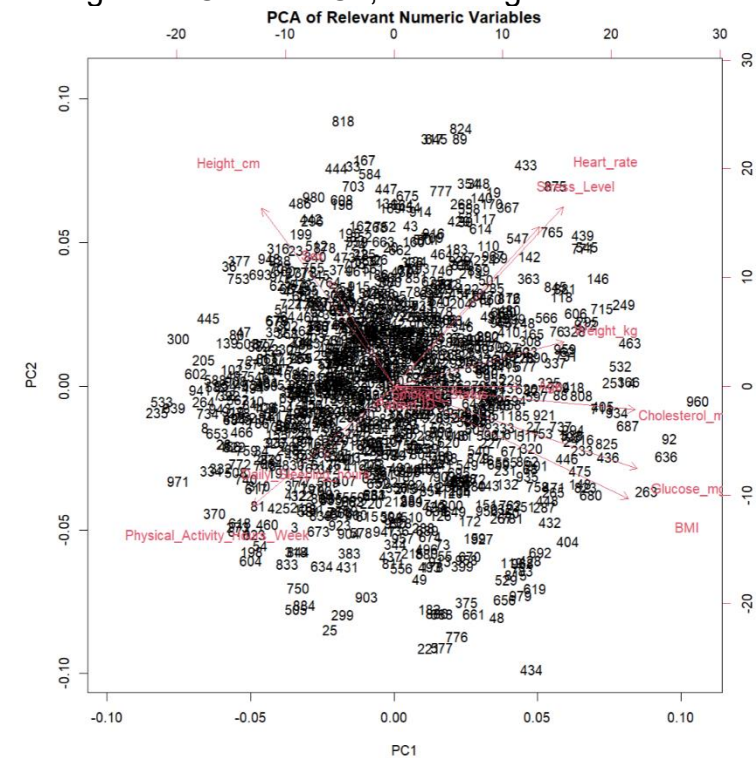


Figure 7. PCA of Relevant Numeric Variables (Biplot). This biplot visualizes data points and variable loadings on PC1 and PC2, indicating variable contributions to these components.



This biplot shows the outcome of Principal Component Analysis on the corresponding numerical variables. The plot demonstrates the projected data points (indicated with the numbers equal to patient IDs) on the plane of the first two principal components (PC1 and PC2). Red arrows show the directions of original variables in this new coordinate system, and their length shows the strength of their correlation with the principal components.

Figure 8. Scatter Plot of the First Two Principal Components. This scatter plot displays data points projected onto the first two principal components.

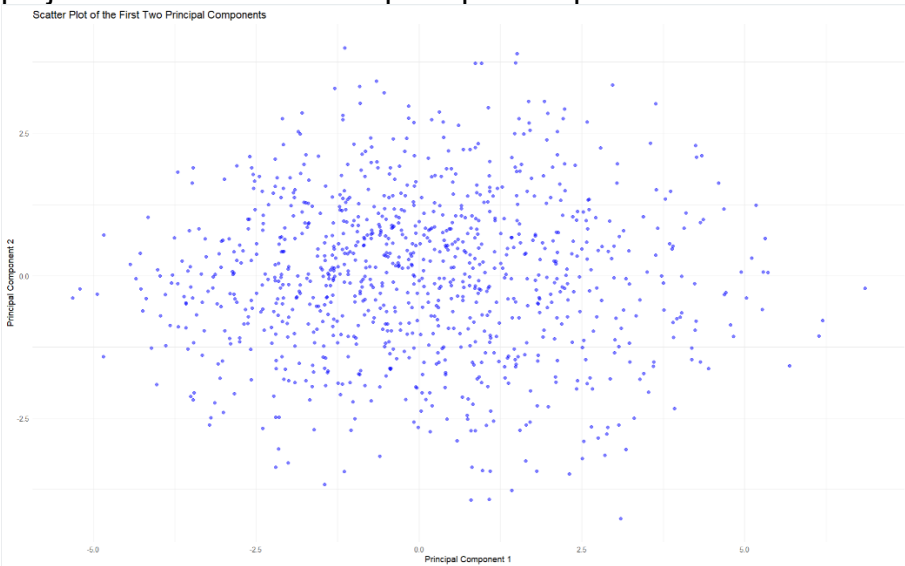


Figure 8 presents a scatter plot of first 2 principal components. There is no staunch clustering of the data points in the plane and they seem to be random. It means that although PC1 and PC2 can explain

a major part of the total variance, they might fail to show unified groupings or powerful structures underlying the data through only the two dimensions.

The blue points are the data in an individual projected on to these components. We can see the density of data through the alpha transparency of 0.5. This plot aids in either seeing whether or not there are any apparent clusters or tendencies to the data relative to the 2 most influential principal components.

Discussion

This descriptive study of vital signs and lifestyle information provides some important evidence on the health status of the analyzed cohort. Preliminary descriptive statistics identified the wide age span (18-90 years) and average BMI indicating the existence of an overweight population. These results were also elaborated on through visualizations: a positive correlation between age and BMI of slight magnitude, uneven distribution of numeric health indicators, and the high occurrence of high hypertension categories were notable. PCA was effective in dimensionality reduction where the first eight components took over 95 percent variance of the data. The PCA scatter plot however, did not show any specific patient clusters indicating that there are complex interactions of factors as opposed to specific subgroups. Such basic analysis highlights how complicated health data is, creating additional research directions through focusing on the investigation of certain health outcomes or risks.

Interpretation and Conclusion

The discussion gives an overarching knowledge of the data set. Whereas certain tendencies, such as a small positive correlation between age and BMI, are present, data is highly fluctuating. The differences in distributions between variables are very high and higher levels of hypertension are prevalent. In PCA, it is stated that only a small group of the major components can explain the bulk of the variance in the data in which case there is a possibility of reducing the dimensions. Nevertheless, the absence of distinct groupings in the PCA scatter plot suggests a complicated structure of the data which can potentially necessitate additional analysis to find out certain subgroups or connections.