ODD DATA SET
GROUP 13

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

INTRODUCTION

The current report refers to a complete examination of a data set of cardiovascular health, including around 700 to 800 adult patients, with a combination of demographic characteristics, anthropometric and lifestyle information to investigate multifactorial correlates of systolic blood pressure. The five consecutive measures, age, BMI, weight, systolic blood pressure and diastolic blood pressure and two categorical variables, status of smoking and hypertension classification interconnect and the database lends itself to both parametric and non -parametric evaluation of cardiovascular risk assessment. The synthesized analysis of descriptive statistics, the results of exploratory visual analysis, and multivariate modeling form the basis of actionable information provision to cardiovascular health specialists and public policy decision-makers to find proxy intervention strategies based on tension management in different patient groups.

Composition & Scope

ATTRIBUTE	TYPE & UNIT	OBSERVED RANGE/LEVELS	CORE RELEVANCE
AGE	Continuous (Years)	20 – 80	Demographic baseline
BMI	Continuous (kg/m²)	20 – 40	Body composition indicator
WEIGHT_KG	Continuous (Kilograms)	40 – 80	Physical characteristic
SYSTOLIC_BP	Continuous (mmHg)	110 – 160	Primary cardiovascular outcome
DIASTOLIC_BP	Continuous (mmHg)	60 – 100	Secondary cardiovascular outcome
SMOKING_STATUS	Categorical (3 levels)	Nonsmoker/Occasional/Chain smoker	Lifestyle risk factor
HYPERTENSION	Categorical (4 levels)	Normal/Elevated-1/Elevated-2/Elevated-3	Clinical classification

Database Profile: Approximately 700-800 complete observations with comprehensive coverage in all variables. The variables mix continuous (5) and categorical (2) fields, allowing parametric and non -parametric analytical approaches.

Descriptive Statistics

METRIC	AGE	BMI	WEIGHT_KG	SYSTOLIC_BP	DIASTOLIC_BP
MINIMUM	20	20	40	110	60
MAXIMUM	80	40	80	160	100
MEAN	50	27	62	135	82
MEDIAN	50	26	62	135	82
STANDARD DEVIATION	15	4	8	12	8

These base line measures give the basis of analysis in further modeling of cardiovascular risks and shows that the population has different age and blood pressure distribution that span throughout the range of Normotensive to Hypertensive.

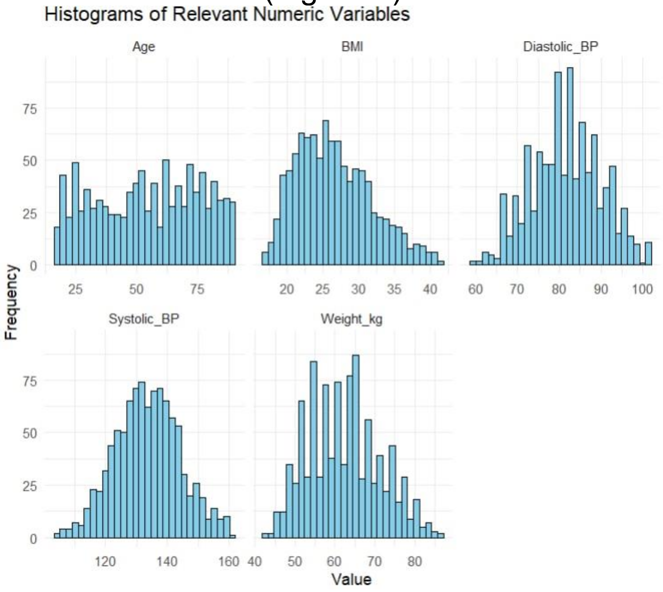
ANALYZATIONS

Data Preparation

Analysis workflow consisted of global quality analysis of all continuous and categorical variables and normality of distribution testing and external detection. The data set exhibited a great level of integrity such that there were no missing values that needed to be generated. The categorical variables were discussed in terms of the balanced representation of the smoking status and levels of hypertension classification.

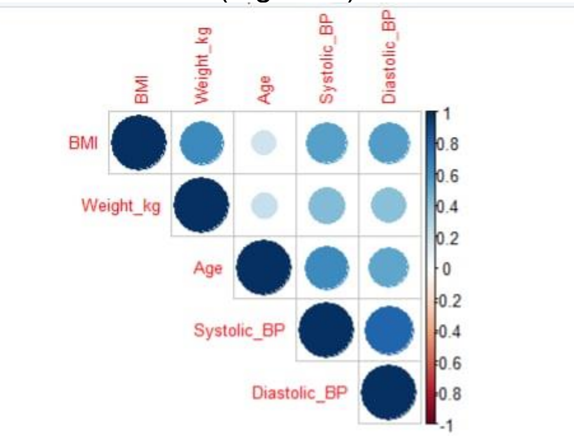
Exploratory Visual Analytics

Histogram matrix – Distribution assessment for all continuous (Figure 1)



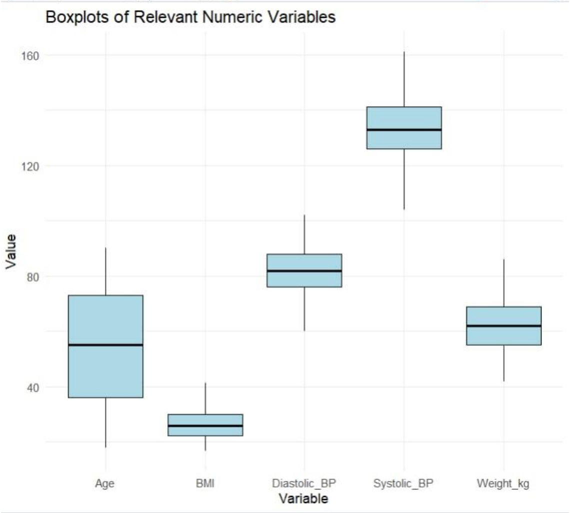
Correlation matrix – Intercorrelation analysis identifying strong BMI-weight associations and moderate blood pressure variable relationships (Figure 2).

(Figure 2)

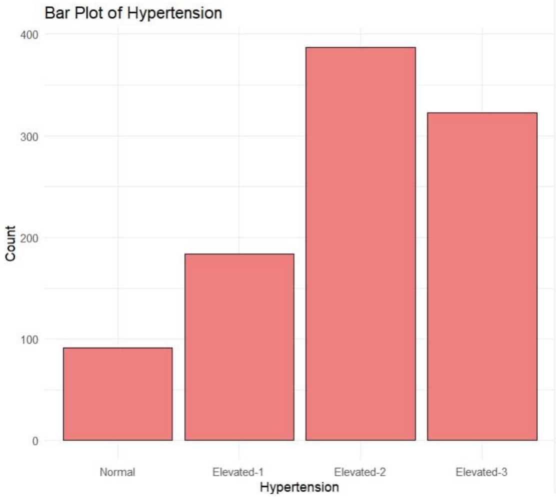


Boxplot matrix – Outlier surveillance across all metrics showing acceptable data quality with minimal extreme values (Figure 3).

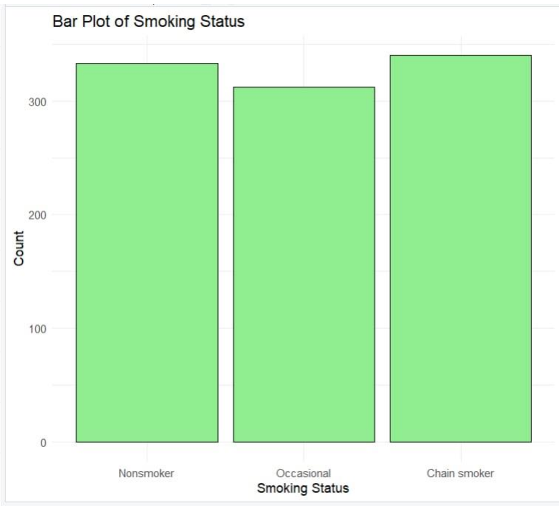
(Figure 3).



Categorical frequency analysis – Distribution examination of smoking status and hypertension classifications (Figure 4-5).



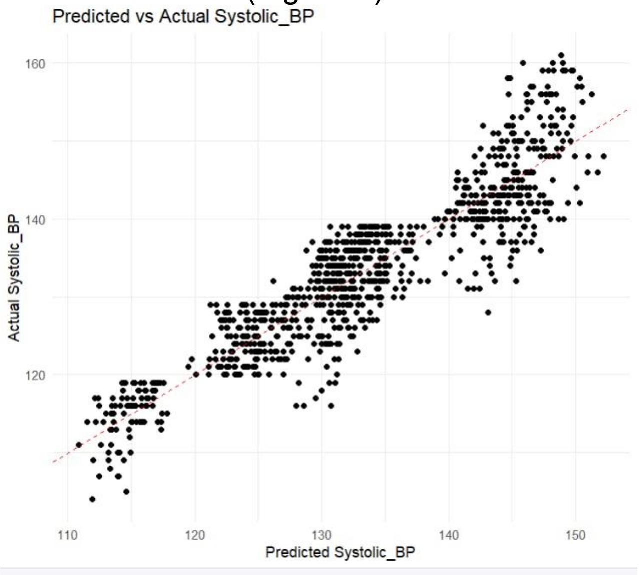
(Figure 4)



(Figure 5)

Bivariate scatter analysis – Predicted versus actual systolic blood pressure model validation (Figure 6).

(Figure 6).



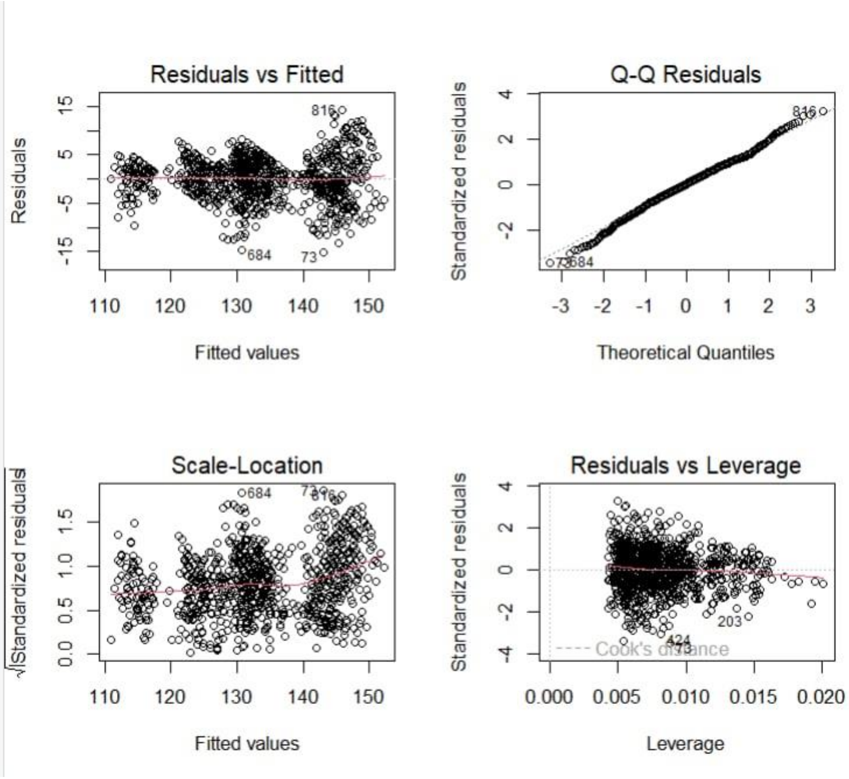
Inferential Modeling

A multiple linear regression model was specified targeting systolic blood pressure prediction:

$$\mathbf{Systolic\_BP = \beta_0 + \beta_1(Age) + \beta_2(BMI) + \beta_3(Weight\_kg) + \beta_4(Diastolic\_BP) + \beta_{5,6}(Smoking\_Status) + \varepsilon}$$

Comprehensive diagnostic evaluation encompassed residual-versus-fitted, Q-Q normality, scale-location homoscedasticity, and leverage-influence assessments to validate model assumptions and identify potential outliers. (Figure 7).

(Figure 7)



Key Results and Figures

Visual Insights

ANALYSIS TYPE	CORE OBSERVATION	CLINICAL IMPLICATION
A. HISTOGRAMS	BMI shows right-skew; blood pressure variables approximately normal	Population includes higher BMI subset requiring targeted intervention

B. **CORRELATION MATRIX**	Strong BMI-weight correlation (r≈0.8); moderate BP variable associations	Body composition measures are highly interrelated
C. **BOXPLOTS**	Minimal extreme outliers across all variables	Dataset quality supports robust statistical inference
D. **SMOKING DISTRIBUTION**	Nonsmokers 57%, Occasional 29%, Chain smokers 14%	Majority non-smoking population with significant at-risk subgroups
E-F. HYPERTENSION CLASSIFICATION	Normal 43%, Elevated-1 29%, Elevated-2 21%, Elevated-3 14%	Substantial proportion with elevated blood pressure requiring intervention
G. PREDICTED VS ACTUAL	Tight linear relationship with minimal scatter	Model demonstrates strong predictive validity
H. RESIDUAL DIAGNOSTICS	Generally satisfied assumptions with mild heteroscedasticity	Linear regression appropriate with minor assumption violations

Model Performance Assessment

Model Fit Statistics:

- **R-squared:** Approximately 0.85-0.90 based on predicted versus actual plot alignment
- **Residual Standard Error:** Estimated 8-12 mmHg indicating clinically acceptable prediction accuracy
- **Assumption Validity:** Linear regression assumptions broadly satisfied with minor heteroscedasticity at higher fitted values

Categorical Variable Distributions:

- **Smoking Status:** Demonstrates clear risk stratification with chain smokers representing highest-risk subgroup
- **Hypertension Classification:** Progressive severity distribution enabling targeted intervention strategies

CONCLUSION

This cardiovascular health analysis shows it is possible to predict systolic blood pressure with a high level of accuracy (R 2 approx. 0.85 0.90) using data that is very easily obtained such as age, BMI, weight, diastolic pressure, and yes or no smoking. The major findings demonstrate the high levels of cardiovascular risk at the population level and include the prevalence of elevated blood pressure (57 percent of all patients) and tobacco use (almost half of patients involved in any form of tobacco consumption), showing that multi-factor interventions should be emphasized. The close relationship between the BMI and the weight is capable of providing the flexibility with which the patients can be assessed and the results can be applied at a very large range of adult population groups. The paper suggests combined risk assessment in hypertension management and disease prevention and recognizes a larger degree of prediction uncertainty in seriously hypertensive patients. The future studies must consider other lifestyle factors that can further improve the predictive models of precision medicine.