Blood Metrics and Chronic Kidney Disease: Exploring Key Associations

Group – 8

Eshani Hitendra Shah, Kanishka Angirish, Naga Hemasree Sravanam, Priyankitha Kandi, Raajitha Muthyala

Part 1: Introduction

Over 800 million people worldwide suffer from chronic kidney disease, a degenerative illness that affects more than 10% of the global population. It is one of the few non-communicable diseases that has seen an increase in related deaths over the past 20 years is chronic renal disease, which has become one of the primary causes of death worldwide (Kovesdy, 2022). This study aims to investigate the association between various blood parameters, comorbidities, and the presence of chronic kidney disease using a dataset obtained from kaggle.

Part 2: Dataset Description

The dataset used in this study consists of 400 observations and 26 variables, blood test results, comorbidities, and the presence or absence of chronic kidney disease. The variables of interest include blood glucose levels (bgr), blood urea levels (bu), presence of bacteria (ba), presence of ckd (class), and hypertension (htn). Additionally, the dataset contains information on other factors such as age, blood pressure, and various other blood test results.

Numerical: age - age, bp - blood pressure, sg - specific gravity, al - albumin, su - sugar, rbc - red blood cells, bgr - blood glucose random, bu - blood urea, sc - serum creatinine, sod - sodium, pot - potassium, hemo - hemoglobin, pcv - packed cell volume, wc - white blood cell count, rc - red blood cell count

Categorical: pc - pus cell, pcc - pus cell clumps, ba - bacteria, htn - hypertension, dm - diabetes mellitus, cad - coronary artery disease, appet - appetite, pe - pedal edema, ane - anemia, class - classification

Part 3: Data Cleaning

Handling Missing Values and Encoding of Categorical Variable: The original dataset contained missing values across multiple columns and categorical variables. The categorical variables were transformed into binary values of 0 or 1 based on the relevant information available, in Microsoft Excel. Furthermore, null values were checked in R which showed 1009 null values across the dataset. Eventually, we replaced the missing values. In the case of numerical variables null values were imputed with the mean values, while in categorical variables null values were replaced with the mode.

Data Type Conversion: Ensuring consistent data types across variables is essential for accurate analysis. The dataset initially contained a mix of character, numeric, and factor variables. Data type conversion was performed to standardize the types across relevant variables. Categorical variables originally encoded as character strings were converted to factor data types. Numerical variables represented as characters were converted to their appropriate numeric data types.

Outlier Detection and Removal:

Outlier Detection Process: Initially, we conducted outlier detection using the Interquartile Range (IQR) method. For each numeric column in the cleaned kidney disease dataset, outliers were identified as values that lay beyond 1.5 times the IQR from the quartiles. Subsequently, the process revealed 377 outliers across various parameters, including age, blood pressure (bp), and blood glucose random (bgr), as evidenced by the obtained output and boxplot visualization.

Winsorization Approach: To mitigate the influence of these extreme values as suggested by R: Winsorize Data (n.d.), winsorization was employed. This technique involved replacing values outside the 5th and 95th percentile range with the nearest values within it. The rationale behind winsorization is to limit the impact of outliers without completely discarding them, thus preserving the dataset's integrity. Post-winsorization, the number of outliers was substantially reduced to 140, highlighting the method's efficacy in handling extreme data points.

Outlier Replacement Strategy: However, some outliers remained resilient. Thus, we adopted a more assertive strategy where these remaining outliers were replaced with the nearest boundary values. By implementing the `winsorize_replace` function, we forced all data points to conform within the established percentile thresholds. This method ensured a more homogenous dataset by completely removing the influence of extreme outliers (R: Winsorize Data, n.d.). The boxplots generated before and after outlier treatment showcased a significant reduction in variance and extreme values. The summary statistics showcase the post-replacement dataset exhibiting tighter ranges and more central clustering of the data points.

Part 4: Exploratory Data Analysis (EDA): Data Visualization

Box Plots: Box plots were used to visualize the presence and severity of outliers in each numeric variable before and after outlier handling (Fig 1, Fig 2). Even after handling outliers, we observed few vital outliers outside the fourth interquartile. We decided to include these, as they form a significant part of the dataset.

Histograms for numeric variables: We created histograms for numerical variables, each accompanied by its own frequency distribution. This approach enabled us to observe the distribution of each variable individually, providing insights into its unique distribution characteristics (Fig4). For instance, the histogram for blood glucose random (bgr) shows a right-skewed distribution with high frequencies at lower values. Whereas the histogram for blood urea (bu) indicates a significant number of observations have elevated blood urea levels, with a long tail towards higher values.

Bar plots for categorical variables: The bar plot for hypertension (htn) depicts a binary categorical variable, representing the presence (1) or absence (0) of hypertension. The substantial difference in bar heights indicates that most observations do not have hypertension, while a smaller subset does (Fig 5). Similarly, the bar plot for diabetes mellitus (dm) indicates that many patients do not have diabetes.

Scatter plots for numeric variables: We produced scatter plots for numerical variables, each coupled with its respective frequency distribution (Fig 3). This method allowed us to analyze the distribution of each variable independently, offering insights into its distinct distribution characteristics. For example, with respect to our research question This scatter plot displays the relationship between Blood Urea (bu) and Blood Glucose Random (bgr) in a dataset of kidney disease patients. The plot suggests variability in blood glucose levels across different urea concentrations, with a dense clustering of points at lower urea levels and a broader spread as urea levels increase.

Correlation heatmap for numeric variables with outcome variable: From the heatmap, we can observe that variables like serum creatinine (sc), blood urea (bu), and red blood cell count (rc) exhibit strong positive correlations with the outcome variable (chronic kidney disease), while variables like hemoglobin (hemo) and packed cell volume (pcv) have strong negative correlations (Fig 6). This information can be valuable for feature selection and identifying potential risk factors associated with chronic kidney disease.

Normality Testing and Log Transformation: We conducted a normality testing using the Shapiro-Wilk test to evaluate the distribution of the variables across the dataset and to ascertain whether it adheres to a normal distribution. The p-values of all numeric variables are less than 0.05, thus rejecting the null hypothesis and stating that data is not normally distributed. Furthermore, we did log transformation to treat skewness of the distribution of variables but no change in distribution was observed.

Part 5: Statistical Methods

In this study, a variety of statistical methods were employed to investigate the association between blood parameters, comorbidities, and chronic kidney disease. The selection of these methods was based on the nature of the variables, the research questions, and the underlying assumptions of the techniques. Additionally, exploratory data analysis (EDA) was conducted to gain insights into the data and guide the appropriate choice of statistical methods. As the data was non-normally distributed, we selected the non-parametric tests for further evaluation.

Summary Statistics: Descriptive statistics, such as mean, median, quartiles, and standard deviation, were calculated for numerical variables to understand their central tendency and dispersion.

Spearman’s correlation Analysis: To assess the association between blood glucose levels (bgr) and blood urea levels (bu) with chronic kidney disease, Spearman's correlation coefficient was calculated. A correlation coefficient of 0.2151 between the variables "bgr" and "bu" suggests a weak positive correlation

Appropriateness: Spearman's correlation is a non-parametric measure of association, suitable for non-parametric data and non-normally distributed data, making it an appropriate choice for the given dataset (Laerd Statistics, n.d.).

Rationale: The rationale behind using Spearman's correlation is its ability to measure the monotonic relationship between two variables, without making assumptions about the normality of the data distribution (Laerd Statistics, n.d.).

Logistic Regression: The logistic regression models were employed to investigate the relationship between the predictor variables (bgr, bu, ba, and htn) and the presence of chronic kidney disease.

Appropriateness: Logistic regression is appropriate when the outcome variable is binary or categorical, and the predictor variables can be continuous, categorical, or a combination of both (Select Statistics, n.d.).

Rationale: Logistic regression model provided estimates of the coefficients associated with each predictor variable, allowing for the assessment of their individual contributions and significance in predicting the outcome that is, chronic kidney disease (Select Statistics, n.d.).

Chi-square Test:

Chi-square Test for Independence: The chi-square test is used to determine if there is a significant association or relationship between the variables htn (hypertension) and ba (bacterial infection) in the kidney_disease_cleaned dataset.

Appropriateness: Chi-square tests are appropriate when both variables are categorical, and the data meets the assumptions of independence and expected frequency counts (McHugh, 2013).

Rationale: The chi-square test provided a statistical measure of the association between the categorical variables and chronic kidney disease, enabling the evaluation of potential risk factors or indicators (McHugh, 2013).

Part 6: Findings

Research Question 1: Association between Blood Glucose Levels (bgr) and Blood Urea Levels (bu) with chronic kidney disease.

H0 – There is no association between blood glucose, blood urea levels and CKD.

H1 – There is an association between blood glucose, blood urea levels and CKD.

The Spearman correlation analysis revealed a weak positive association between blood glucose levels (bgr) and blood urea levels (bu) (correlation coefficient = 0.2236) (Statstutor, n.d.).

However, the logistic regression model demonstrated a significant positive relationship between both bgr and bu with chronic kidney disease.

- The Variance Inflation Factor (VIF) quantifies the extent of variance increase in coefficient estimates caused by multicollinearity among predictors in a regression model (Bobbitt, n.d.). A VIF of 1.01631 for "bgr" and "bu" suggests negligible multicollinearity, indicating stable variance for these coefficient estimates in the regression model.
- The intercept coefficient is -5.859607, which is statistically significant (p-value < 0.001).
- The coefficient for bgr (blood glucose levels) is 0.034044, which is statistically significant (p-value < 0.001). This positive coefficient suggests that higher blood glucose levels are associated with an increased likelihood of chronic kidney disease.
- The coefficient for bu (blood urea levels) is 0.045287, which is also statistically significant (p-value < 0.001). This positive coefficient indicates that higher blood urea levels are associated with an increased likelihood of chronic kidney disease.

The positive coefficients suggest that increasing levels of bgr and bu are strongly associated with likelyhood of CKD. Thus, we are rejecting the null hypothesis (H0).

Research Question 2: Association between Bacteria (ba) and Hypertension (htn) with chronic kidney disease.

H0- There is no association between the presence of hypertension, bacteria and CKD.
H1 – There is an association between the presence of hypertension, bacteria and CKD.

Chi-square test was done for all categorical variables, and htn and ba are the variables that didn’t show multi co-linearity. The chi-square test for independence between ba and htn with chronic kidney disease yielded a p-value of 0.1203, indicating no significant association between the two variables. Furthermore, the logistic regression model revealed that neither ba nor htn were statistically significant predictors of chronic kidney disease in the studied dataset.

- The coefficient for htn (presence of hypertension) is 19.904, but it is not statistically significant (p-value = 0.981602).
- The coefficient for ba (presence of bacterial infection) is 18.741, but it is also not statistically significant (p-value = 0.991566).

  Thus, the null hypothesis (H0) cannot be rejected.

<u>Part 6: Limitations</u>

- **Sample Size and Generalizability**: With only 400 observations, the dataset may not be representative of the entire population, potentially limiting the applicability of the findings to different or broader groups.
- **Presence of Missing Values**: Imputing missing values can introduce biases, especially if the missing data patterns are not random. This method relies on the assumption that missing values resemble the observed data.
- **Potential for Residual Confounding**: The study could be limited by unmeasured variables that may influence the outcomes, such as dietary habits, lifestyle choices, or genetic factors, which are not included in the dataset.
- **Data Processing and Outlier Management**: Outlier management techniques used, like winsorization and replacement, may not fully address their impact on the analysis and can be subjective, leading to potential data distortion.

Part 7: Conclusion

We explored the associations between various blood parameters, comorbidities, and chronic kidney disease using statistical methods and data visualization techniques. The findings suggest that elevated blood glucose levels (bgr) and blood urea levels (bu) are significantly associated with an increased risk of chronic kidney disease. However, the presence of bacteria (ba) and hypertension (htn) did not exhibit a significant association with chronic kidney disease in the studied dataset.

These results highlight the importance of monitoring and managing blood glucose and blood urea levels to potentially mitigate the risk of developing chronic kidney disease.

References:

Bobbitt, Z. (n.d.). *How to calculate Variance Inflation Factor (VIF) in R*. Statology. https://www.statology.org/variance-inflation-factor-r/

Kovesdy C. P. (2022). Epidemiology of chronic kidney disease: an update 2022. *Kidney international supplements*, *12*(1), 7–11. https://doi.org/10.1016/j.kisu.2021.11.003

Laerd Statistics. (n.d.). *Spearman's rank-order correlation using SPSS Statistics*.

https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica,* 143–149.

https://doi.org/10.11613/bm.2013.018

R: Winsorize data. (n.d.). *R-project.org.* https: //search.r-project.org/ CRAN/ refmans/data

wizard/html/winsorize.html

Select Statistics. (n.d.). *Analysing categorical data using logistic regression models*. https://select-statistics.co.uk/blog/analysing-categorical-data-using-logistic-regression-models/#:~:text=Logistic%20

regression%20models%20are%20a,and%20plan%20for%20future%20scenarios.

Statstutor. (n.d.). https://www.statstutor.ac.uk/

Appendix:

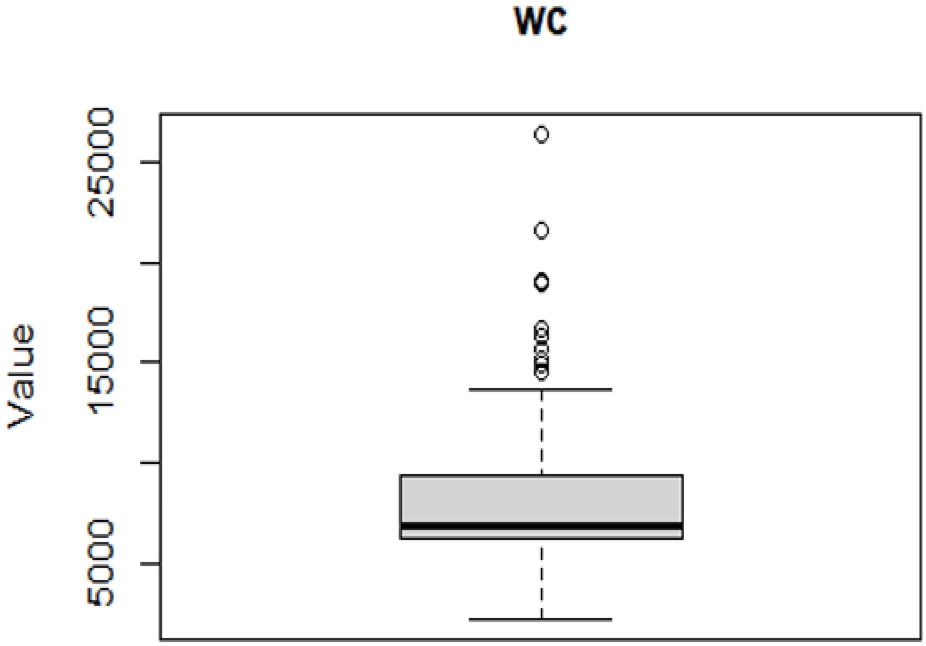


Figure 1- Data Visualization of outlier detection and treatment before winsorization.

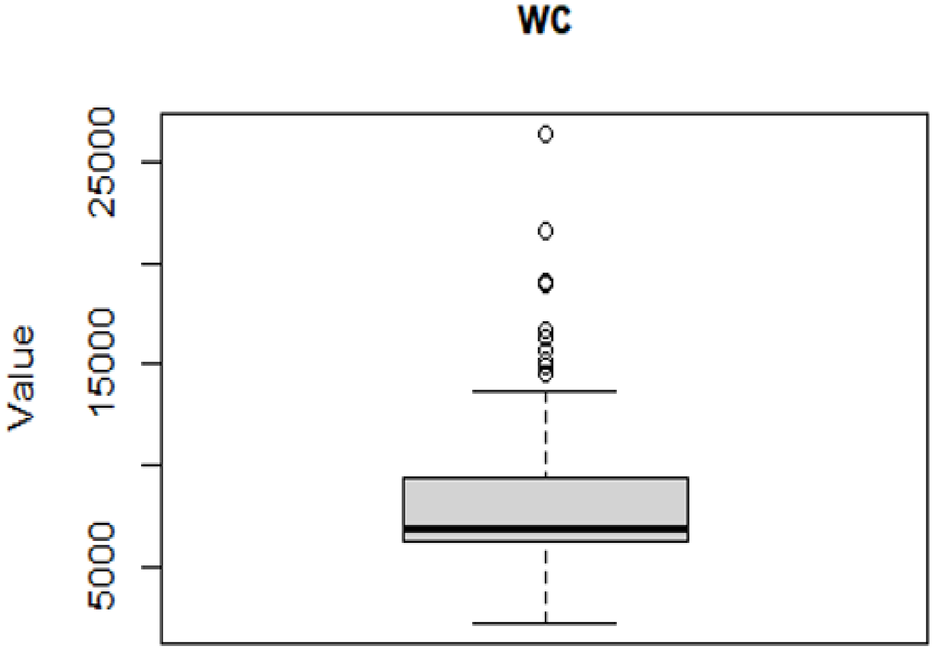


Figure 2- Data Visualization of outlier detection and treatment after winsorization.

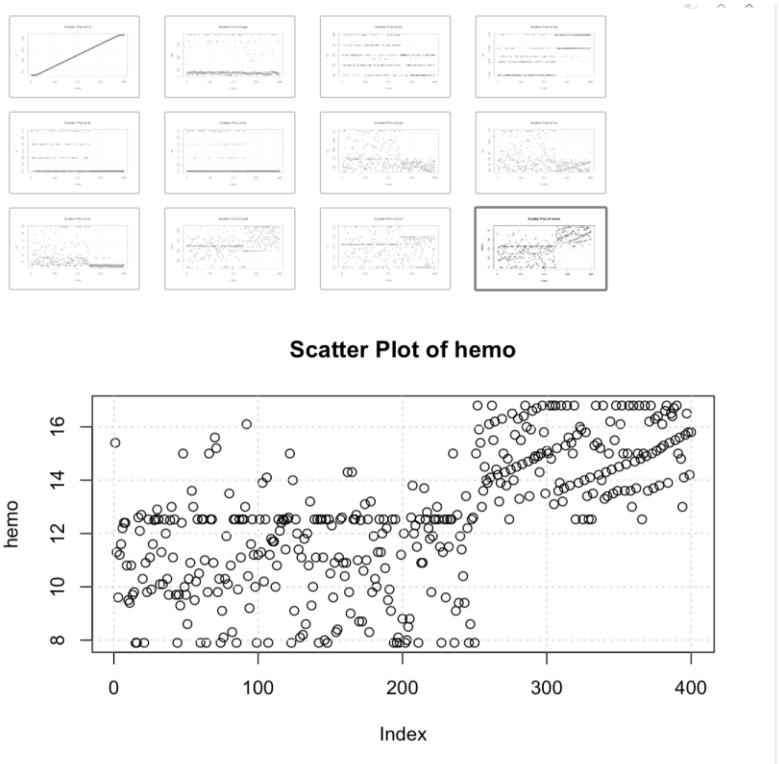


Figure 3- Scatter plots for numeric variables to understand the distribution of the variables.

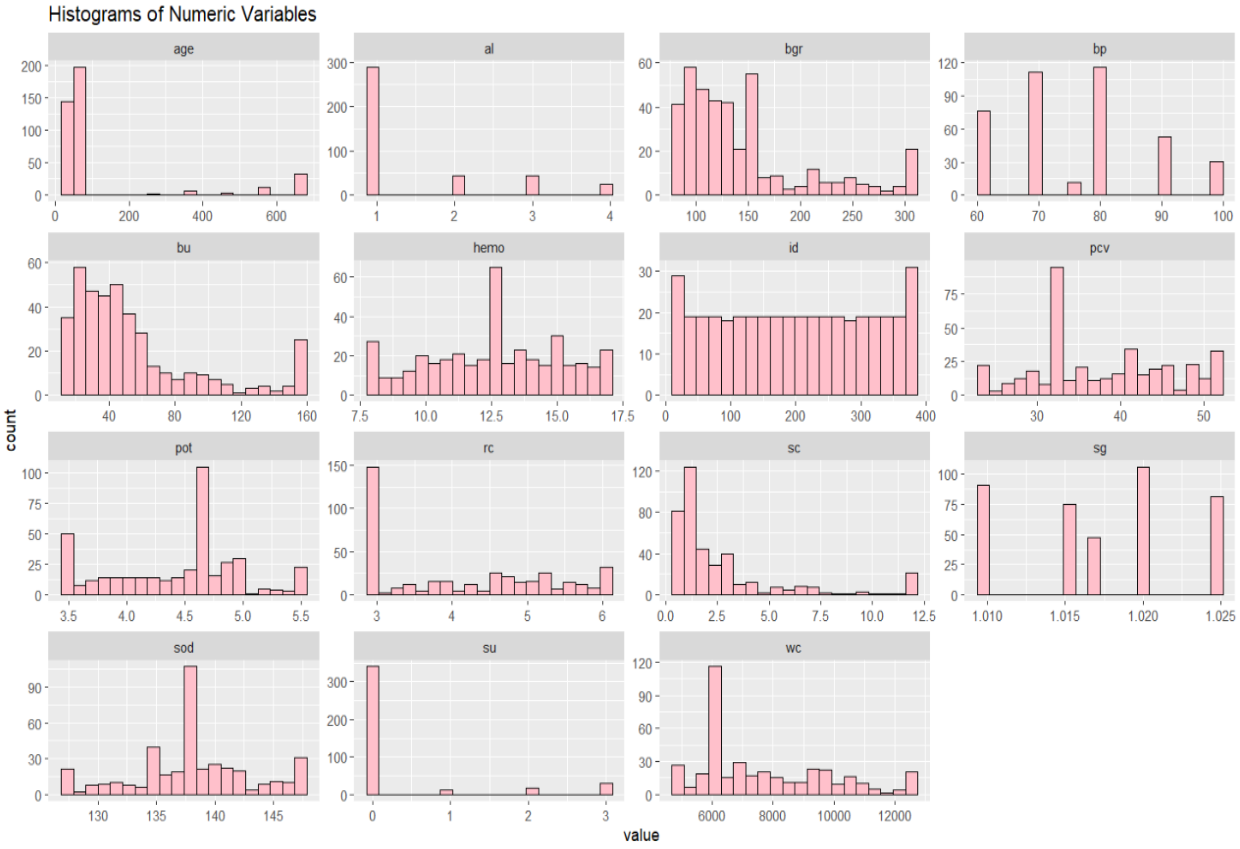


Figure 4- Histograms for numeric variables to understand the distribution of the variables.

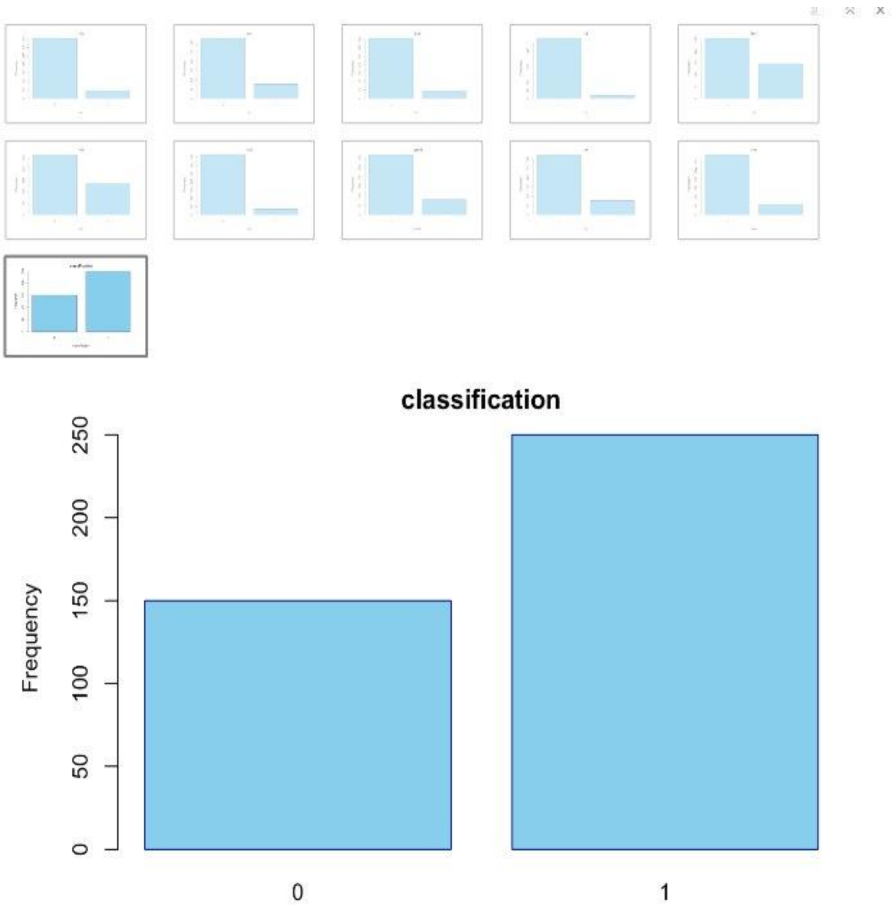
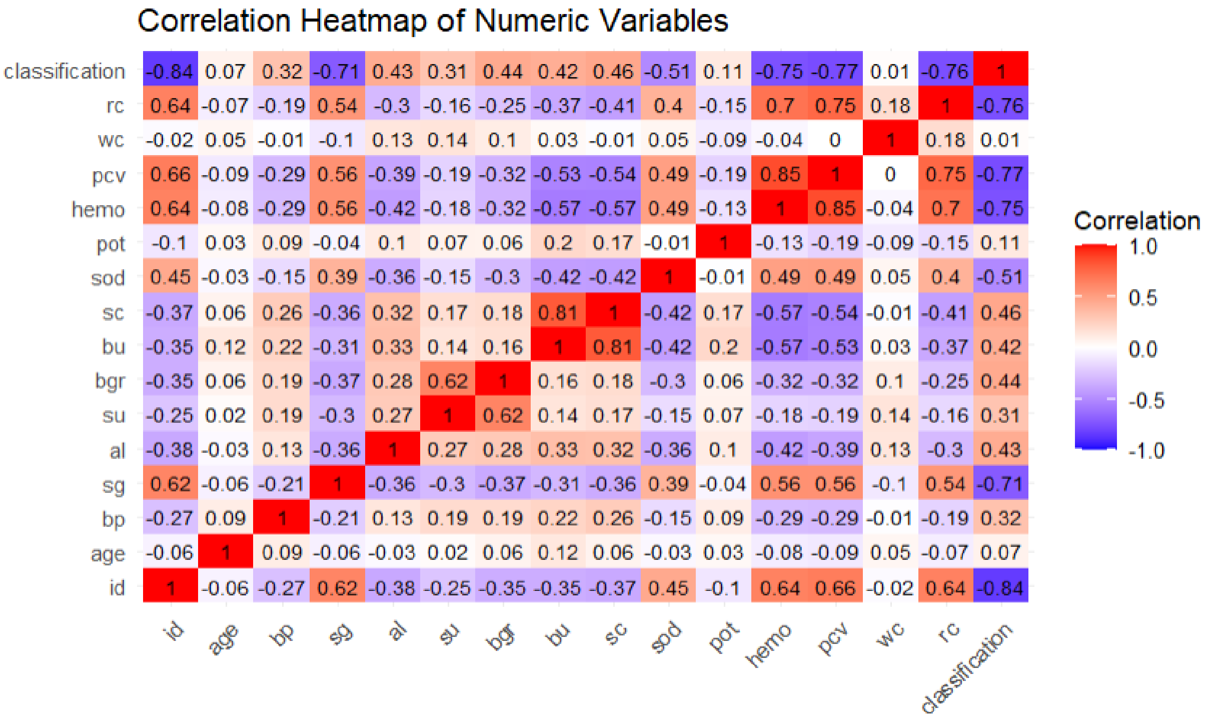


Figure 5- Bar plots for the categorical variables to understand the frequency of each variable.



	id	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	classification
classification	-0.84	0.07	0.32	-0.71	0.43	0.31	0.44	0.42	0.46	-0.51	0.11	-0.75	-0.77	0.01	-0.76	1
rc	0.64	-0.07	-0.19	0.54	-0.3	-0.16	-0.25	-0.37	-0.41	0.4	-0.15	0.7	0.75	0.18	1	-0.76
wc	-0.02	0.05	-0.01	-0.1	0.13	0.14	0.1	0.03	-0.01	0.05	-0.09	-0.04	0	1	0.18	0.01
pcv	0.66	-0.09	-0.29	0.56	-0.39	-0.19	-0.32	-0.53	-0.54	0.49	-0.19	0.85	1	0	0.75	-0.77
hemo	0.64	-0.08	-0.29	0.56	-0.42	-0.18	-0.32	-0.57	-0.57	0.49	-0.13	1	0.85	-0.04	0.7	-0.75
pot	-0.1	0.03	0.09	-0.04	0.1	0.07	0.06	0.2	0.17	-0.01	1	-0.13	-0.19	-0.09	-0.15	0.11
sod	0.45	-0.03	-0.15	0.39	-0.36	-0.15	-0.3	-0.42	-0.42	1	-0.01	0.49	0.49	0.05	0.4	-0.51
sc	-0.37	0.06	0.26	-0.36	0.32	0.17	0.18	0.81	1	-0.42	0.17	-0.57	-0.54	-0.01	-0.41	0.46
bu	-0.35	0.12	0.22	-0.31	0.33	0.14	0.16	1	0.81	-0.42	0.2	-0.57	-0.53	0.03	-0.37	0.42
bgr	-0.35	0.06	0.19	-0.37	0.28	0.62	1	0.16	0.18	-0.3	0.06	-0.32	-0.32	0.1	-0.25	0.44
su	-0.25	0.02	0.19	-0.3	0.27	1	0.62	0.14	0.17	-0.15	0.07	-0.18	-0.19	0.14	-0.16	0.31
al	-0.38	-0.03	0.13	-0.36	1	0.27	0.28	0.33	0.32	-0.36	0.1	-0.42	-0.39	0.13	-0.3	0.43
sg	0.62	-0.06	-0.21	1	-0.36	-0.3	-0.37	-0.31	-0.36	0.39	-0.04	0.56	0.56	-0.1	0.54	-0.71
bp	-0.27	0.09	1	-0.21	0.13	0.19	0.19	0.22	0.26	-0.15	0.09	-0.29	-0.29	-0.01	-0.19	0.32
age	-0.06	1	0.09	-0.06	-0.03	0.02	0.06	0.12	0.06	-0.03	0.03	-0.08	-0.09	0.05	-0.07	0.07
id	1	-0.06	-0.27	0.62	-0.38	-0.25	-0.35	-0.35	-0.37	0.45	-0.1	0.64	0.66	-0.02	0.64	-0.84

Figure 6- Heatmap to visualize the correlation.

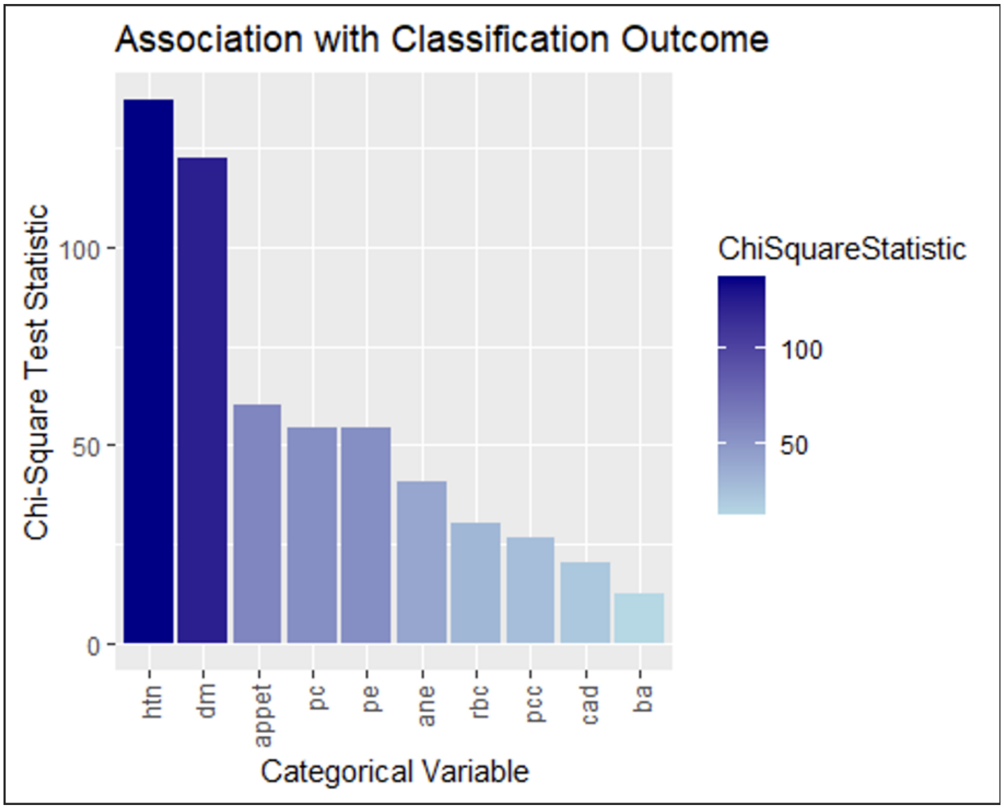


Figure 7: Data Visualization for categorical variables

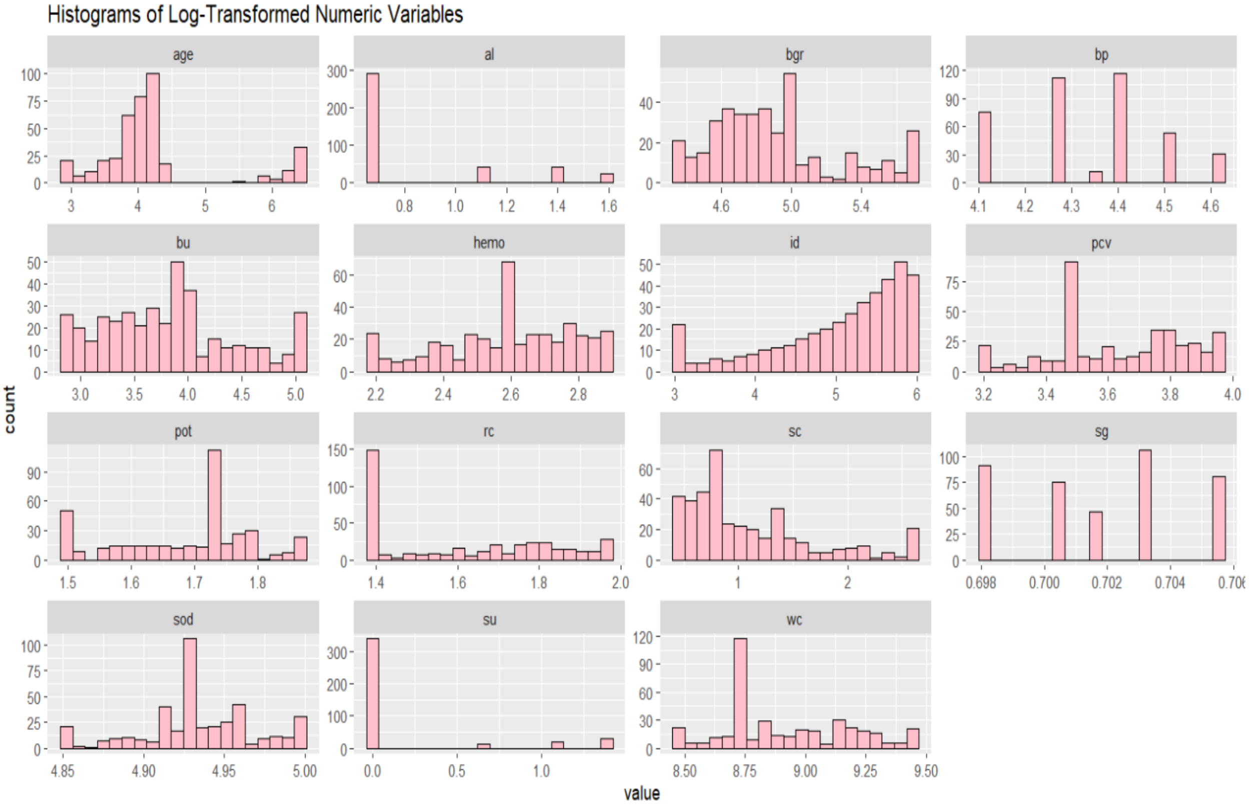


Figure 8: Histograms of Log- transformed variables

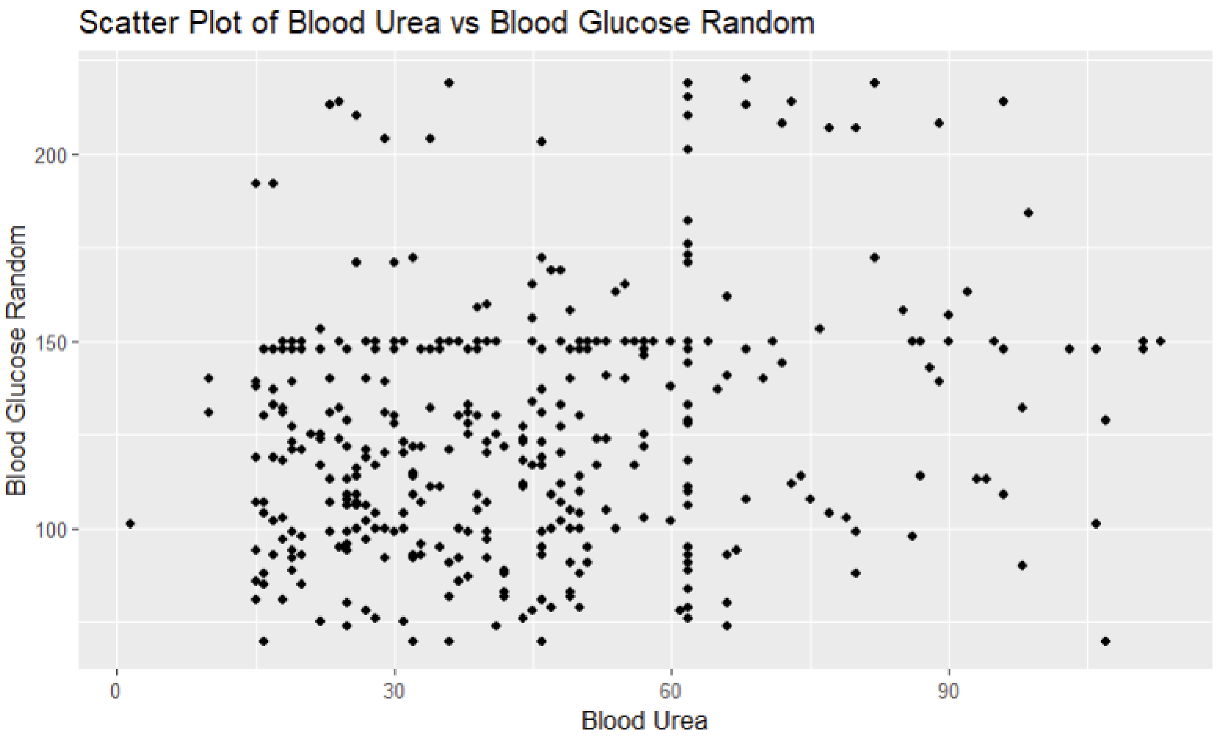


Figure 9: Scatter plot of BloodUrea vs Blood Glucose Random