COURSE: DSA-5103 – INTELLIGENT DATA ANALYTICS

SECTION: 001

SEMESTER: FALL 2023

INSTRUCTOR: DR. CHARLES NICHOLSON

TITLE: INITIAL DATA REPORT

NAME OF PROJECT: PREDICTING SMOKING AND DRINKING BEHAVIOR USING BODY SIGNAL DATA

GROUP NUMBER: 16 GROUP

MEMBERS:

Tushar Jayendra Mhatre - Tushar.Jayendra.Mhatre-1@ou.edu

Roshini Talluru - Roshini.Talluru-1@ou.edu

Rahul Kataram - Rahul.Kataram-1@ou.edu

L Srihari Boppana - L.Srihari.Boppana-1@ou.edu

Contents

PROJECT FINAL REPORT

1. Introduction to the problem:

Often, people tend to lie about their past smoking and drinking habits to the insurance companies in hopes of getting a lower rate on the premiums. In medical institutes as well, doctors need to know about the smoking and drinking habits of their patients before performing an important procedure and they can't simply rely on the patient's intuition, In such cases, there needs to be a reliable way to detect and determine a person's smoking or drinking habits based on their body data.

Moreover, Smoking and drinking are not only personal choices but also public health concerns with far-reaching implications. These behaviors are associated with a multitude of health risks, including cardiovascular diseases, cancers, and liver problems. Understanding the factors contributing to these behaviors is vital for designing targeted interventions, public health campaigns, and policy measures to reduce their prevalence.

The dataset from the National Health Insurance Service in Korea provides a unique opportunity to explore the relationships between individual attributes and smoking and drinking behaviors. This dataset includes a wide range of physiological measurements such as blood pressure, cholesterol levels, and eyesight, alongside demographic details like age and sex. These measurements offer valuable insights into the potential health effects of smoking and drinking.

2. Data Description:

The dataset consists of various individual physiological measurements such as blood pressure, cholesterol levels, and eyesight, alongside demographic details like age and sex. There are 24 columns and the "sex" column is the first one and the rest are various medical parameters that may impact the percentage of smokers and drinkers.

2.1.Details of dataset:

Column	Description
Sex	male, female
age	round up to 5 years
height	round up to 5 cm[cm]
weight	[kg]
sight_left	eyesight(left)
sight_right	eyesight(right)
hear_left	hearing left, 1(normal), 2(abnormal)
hear_right	hearing right, 1(normal), 2(abnormal)
SBP	Systolic blood pressure[mmHg]
DBP	Diastolic blood pressure[mmHg]

BLDS	BLDS or FSG(fasting blood glucose)[mg/dL]
tot_chole	total cholesterol[mg/dL]
HDL_chole	HDL cholesterol[mg/dL]
LDL_chole	LDL cholesterol[mg/dL]
triglyceride	triglyceride[mg/dL]
hemoglobin	hemoglobin[g/dL]
urine_protein	protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)
serum_creatinine	serum(blood) creatinine[mg/dL]
SGOT_AST	SGOT(Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase)[IU/L]
SGOT_ALT	ALT(Alanine transaminase)[IU/L]
gamma_GTP	y-glutamyl transpeptidase[IU/L]
SMK_stat_type_cd	Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke)
DRK_YN	Drinker or Not

3. Initial Analysis:

Initially, two data quality reports were generated, one focusing on numerical data, and the other on non-numeric (categorical) columns. The analysis of the data quality report reveals the presence of 18 numerical variables and 6 categorical variables. Notably, among the categorical variables, SMK_stat_type_cd and DRK_YN are identified as the variables to be utilized for classification purposes.

3.1 Data Report:

Following are the Data Reports that were generated.

3.1.1 Numeric Summary Report

	variable	n	missing	unique	Unique_percentage	missing_Percentage	mean	min	Q1	median	Q3	max	sd
1	age	991346	0	14	0.00141	0	47.614	20.0	35.0	45.0	60.0	85.0	14.181
2	height	991346	0	13	0.00131	0	162.241	130.0	155.0	160.0	170.0	190.0	9.283
3	weight	991346	0	24	0.00242	0	63.284	25.0	55.0	60.0	70.0	140.0	12.514
4	waistline	991346	0	737	0.07434	0	81.233	8.0	74.1	81.0	87.8	999.0	11.850
5	sight_left	991346	0	24	0.00242	0	0.981	0.1	0.7	1.0	1.2	9.9	0.606
6	sight_right	991346	0	24	0.00242	0	0.978	0.1	0.7	1.0	1.2	9.9	0.605
7	SBP	991346	0	171	0.01725	0	122.432	67.0	112.0	120.0	131.0	273.0	14.543
8	DBP	991346	0	127	0.01281	0	76.053	32.0	70.0	76.0	82.0	185.0	9.889
9	BLDS	991346	0	498	0.05023	0	100.424	25.0	88.0	96.0	105.0	852.0	24.180
10	tot_chole	991346	0	474	0.04781	0	195.557	30.0	169.0	193.0	219.0	2344.0	38.660
11	HDL_chole	991346	0	223	0.02249	0	56.937	1.0	46.0	55.0	66.0	8110.0	17.238
12	LDL_chole	991346	0	432	0.04358	0	113.038	1.0	89.0	111.0	135.0	5119.0	35.843
13	triglyceride	991346	0	1657	0.16715	0	132.142	1.0	73.0	106.0	159.0	9490.0	102.197
14	hemoglobin	991346	0	190	0.01917	0	14.230	1.0	13.2	14.3	15.4	25.0	1.585
15	serum_creatinine	991346	0	183	0.01846	0	0.860	0.1	0.7	0.8	1.0	98.0	0.481
16	SGOT_AST	991346	0	568	0.05730	0	25.989	1.0	19.0	23.0	28.0	9999.0	23.493
17	SGOT_ALT	991346	0	594	0.05992	0	25.755	1.0	15.0	20.0	29.0	7210.0	26.309
18	gamma_GTP	991346	0	940	0.09482	0	37.136	1.0	16.0	23.0	39.0	999.0	50.424

Fig 3.1.1.1 Numeric Summary Report

3.1.2 Non-Numeric Summary Report

There are 6 Factor variables in total among which the final 2 ‘SMK_stat_type_cd’ and ‘DRK_YN’ are the variables we are going to predict.

	variable	n	unique	Unique_percentage	missing	missing_percentage	1st mode	first_mode_freq	least common	least common freq
1	sex	991346	2	0.000201745909097328	0	0	Male	526415	Female	464931
2	hear_left	991346	2	0.000201745909097328	0	0	1	960124	2	31222
3	hear_right	991346	2	0.000201745909097328	0	0	1	961134	2	30212
4	urine_protein	991346	6	0.000605237727291985	0	0	1	935175	6	512
5	SMK_stat_type_cd	991346	3	0.000302618863645992	0	0	1	602441	2	174951
6	DRK_YN	991346	2	0.000201745909097328	0	0	N	495858	Y	495488

Fig 3.1.2.1 Non-Numeric Summary Report

3.2 Data Preprocessing:

The dataset at hand is characterized by high quality, devoid of any missing values. However, upon closer analysis, certain parameters have been identified that may exert an influence on the dataset. The contingency table for waistline columns was constructed using the ‘table’ (command to systematically capture counts at the intersection of factor levels. The ensuing observations from this analysis are as follows:

```
114.1 114.2 114.3 114.5 114.6 114.7 114.8 114.9   115 115.1 115.2 115.3 115.4 115.5 115.6
    5     7     5    15     3     5     2     2   180     3     6     3     4    12     2
115.7 115.8 115.9   116 116.1 116.2 116.3 116.4 116.5 116.6 116.7 116.8 116.9   117 117.1
    3     3     1   131     1     4     2     3    11     2     1     2     2   105     1
117.2 117.3 117.4 117.5 117.6 117.7 117.8 117.9   118 118.1 118.2 118.3 118.4 118.5 118.8
    5     3     1    13     1     2     3     2    73     2     7     2     2     5     1
118.9   119 119.1 119.2 119.3 119.4 119.5 119.7 119.8   120 120.1 120.2 120.5 120.6 120.7
    1    47     3     6     2     1     4     1     1    72     1     1     2     2     1
  121 121.1 121.2 121.3 121.4 121.5   122 122.3 122.4 122.5 122.6   123 123.1 123.2 123.3
   31     1     2     1     2     5    23     1     1     2     1    14     1     1     1
123.4 123.5 123.8   124 124.1 124.2 124.5   125 125.5   126 126.2 126.5 126.6   127 127.3
    2     3     1    15     1     2     2    15     1     9     1     2     1    11     1
  128   129 129.6   130 130.5   131   132   133   134   135   136 136.8   138   140   145
    7     9     1     3     1     3     2     1     3     1     2     1     1     1     1
149.1   999
    1    57
```

Fig 3.2.1 Waistline Observations

The distribution analysis of the 991,374 observations indicates a predominant range between 8 and 149.1. However, a notable anomaly is observed in 57 instances, where the recorded value is 999 cm. These instances are deemed as potentially erroneous entries, warranting consideration as missing values necessitating imputation. The distinction from outliers is crucial, as outliers may still represent valid data points deviating significantly from the norm. In contrast, the identified values of 999 cm are deemed incorrect, impossible, or inadmissible within the context of the dataset and the addressed problem. This conclusion is supported by the physical implausibility of an individual having a waistline measurement of 999 cm. Moreover, the absence of any documentation on the dataset website regarding the significance or implications of such values for the specified variables further strengthens the rationale for treating them as erroneous entries.

Initially, the histogram of the waistline column (Fig. 3.2.2) revealed an asymmetric and left-skewed distribution. However, after preprocessing, the histogram (Fig. 3.2.3) demonstrates a symmetrical and normally distributed pattern, indicating a reduced probability of outliers. The proximity of the mean, mode, and median suggests a more central and balanced distribution in the post-processing state.

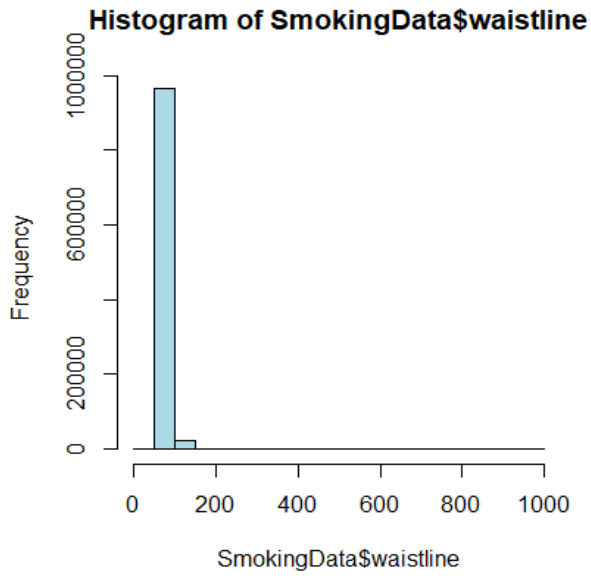

Fig 3.2.2 Before Data-Preprocessing

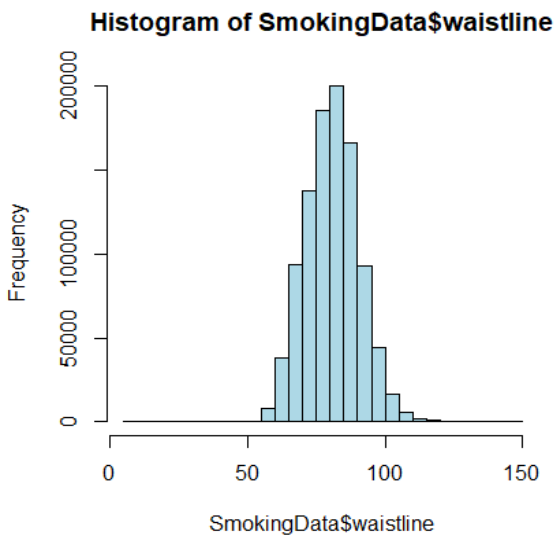

Fig 3.2.3 After Data-Preprocessing

We analyzed the data and noticed the same trend was there in the following variables as well

1. HDL_chole
2. waistline
3. LDL_chole
4. triglyceride
5. SGOT_AST

After replacing missing values with "NA," a Predictive Mean Matching approach was employed for imputation, resulting in a noteworthy improvement in the distribution of data for the respective variables.

Now, things are a bit different for these two variables

1. HDL_leftSight
2. HDL_rightSight

These two variables, HDL_leftSight and HDL_rightSight, serve as indicators of eyesight strength, with higher values indicating weaker eyesight. The standard range for both variables is typically between 0 and 2.5. However, there are instances where a value of 9.9 is recorded, which falls significantly outside this expected range. Notably, the dataset author has clarified on the website that these particular values denote blindness in the respective individuals. Despite this clarification, the inclusion of 9.9 values contributes to a highly skewed distribution, as illustrated in the histogram below. The skewness arises from the presence of these extreme values, and it is important to consider their impact on the analysis and interpretation of the dataset, especially in the context of eyesight strength assessment.

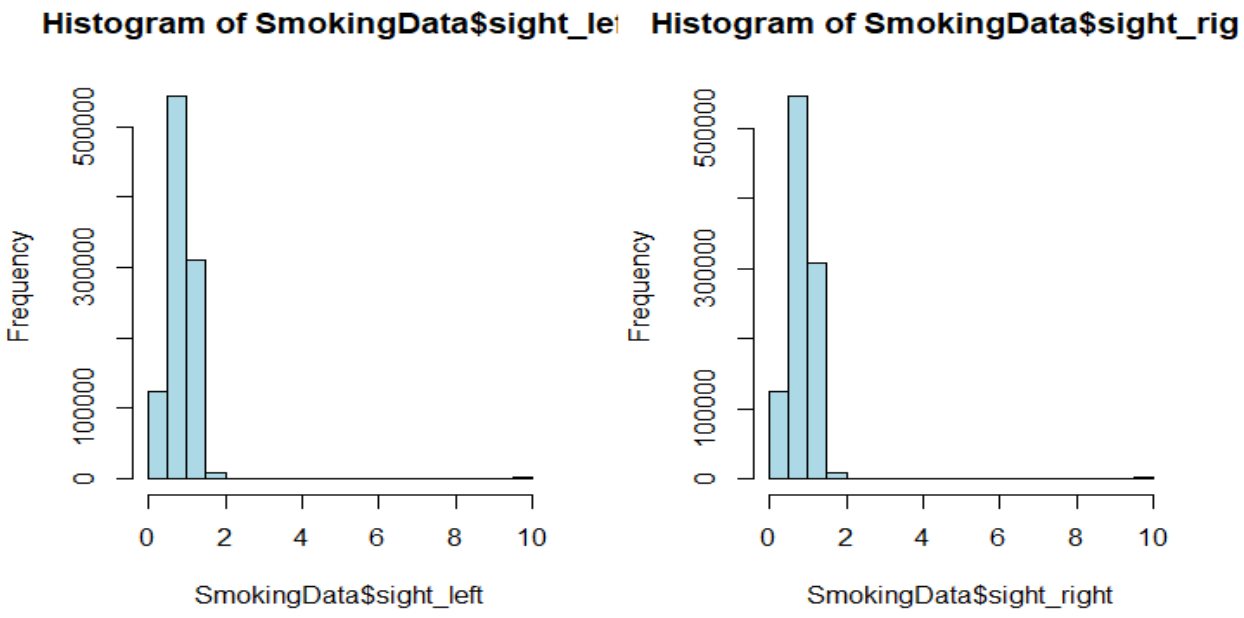

Fig 3.2.4 Histogram of Smoking data sight left and sight right before Data Preprocessing

The presence of outlier values, particularly the instances with a value of 9.9 in HDL_leftSight and HDL_rightSight variables, was successfully mitigated to achieve a more normally distributed dataset. A threshold value close to 4 was chosen, considering that none of the values in other observations exceeded 2.5. Subsequently, all values greater than this threshold, specifically the observations with a value of 9.9, were set equal to the chosen threshold. This strategic adjustment significantly improved the overall distribution of data in these two variables, as illustrated in the figure (3.2.5) below.

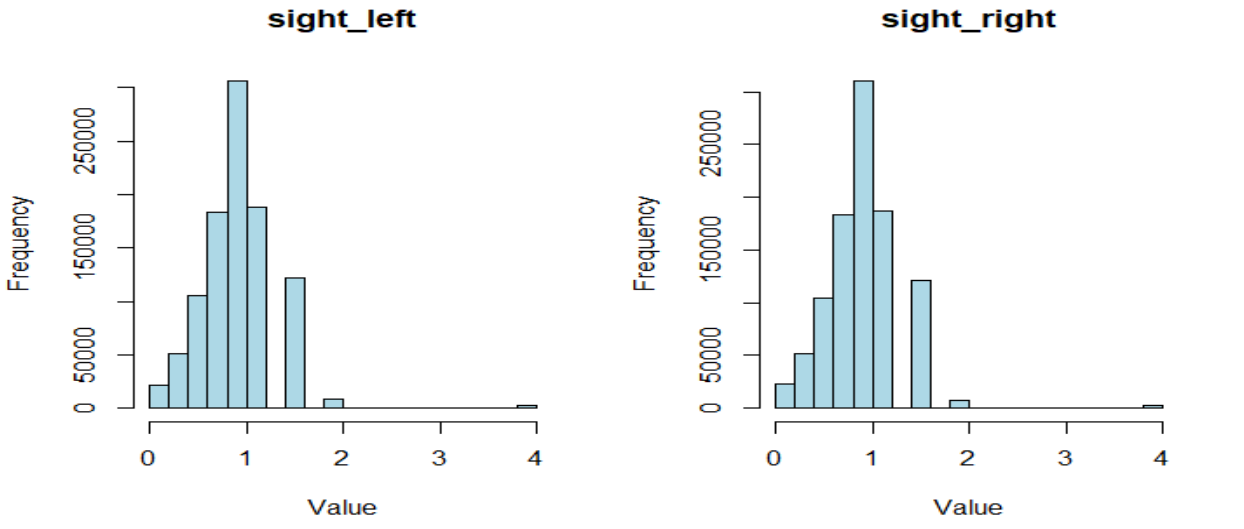

Fig 3.2.5 Histogram of Smoking data sight left and sight right after Data Preprocessing

3.3 Exploratory Data Analysis:

3.3.1 Visualization

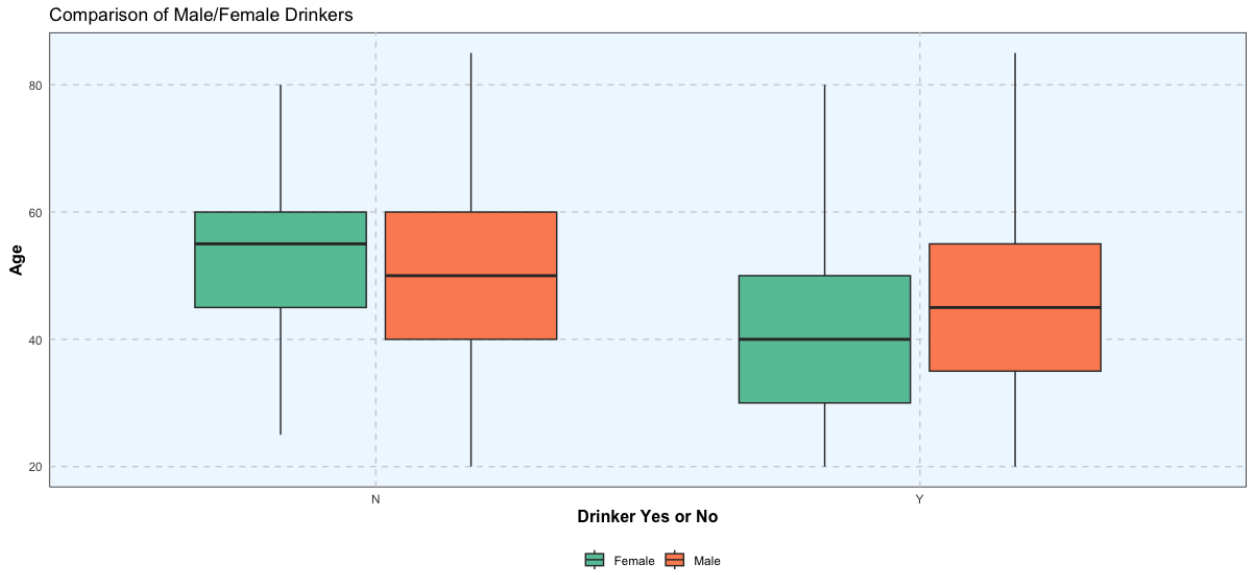

Fig 3.3.1.1 Boxplot Visualization

Based on the depicted graph (Fig 3.3.1.1), it is observed that within the age range of 20-30, the prevalence of female drinkers exceeds that of males. Conversely, the demographic of male drinkers is more prominent within the age bracket of 30-50. Consequently, it can be inferred that, on average, the life cycle exhibits a higher proportion of male drinkers, with a greater concentration of young females.

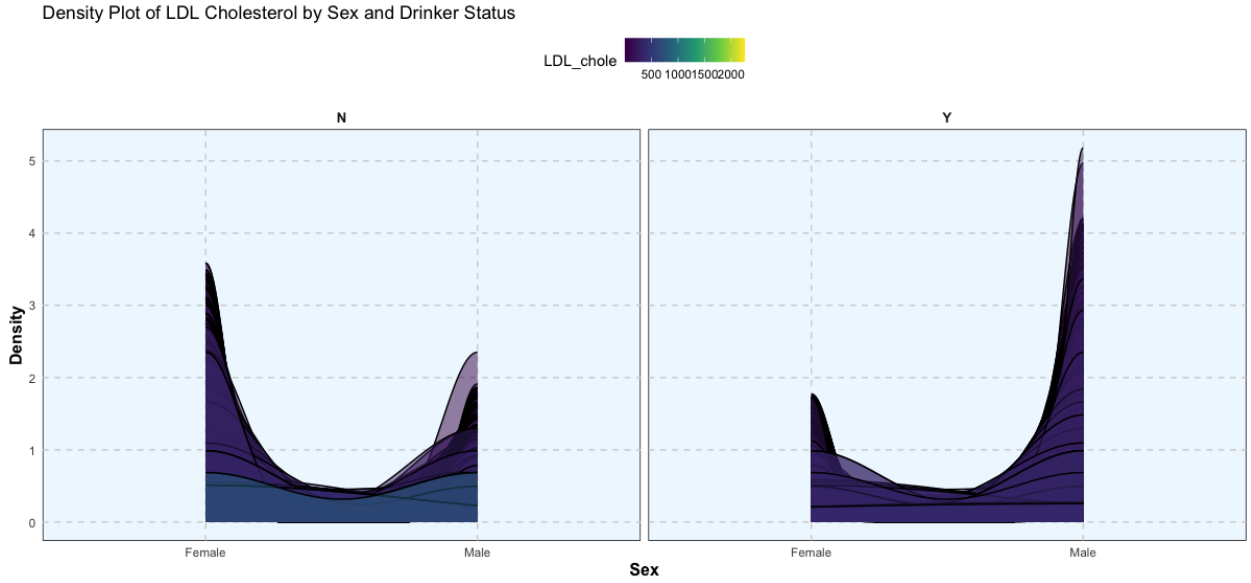

Fig 3.3.1.2 Density-plot Visualization

Upon examination of the above-density plot (Fig 3.3.1.2), it becomes apparent that male drinkers exhibit higher levels of bad cholesterol. This observation leads to the inference that alcohol consumption may have an impact on individual health, particularly on cholesterol levels.

3.4 Feature Selection

3.4.1 Correlation Matrix

A correlation matrix was used for feature selection to identify highly correlated features, which can be candidates for removal to avoid redundancy and avoid multicollinearity.

Observing the below heatmap (Fig 3.4.1.1), it is evident that waistline and age exhibit a robust negative correlation, as evidenced by the light coloration at their intersection. Conversely, a strong positive correlation is observed between age and height, substantiated by the darker coloration at their intersection.

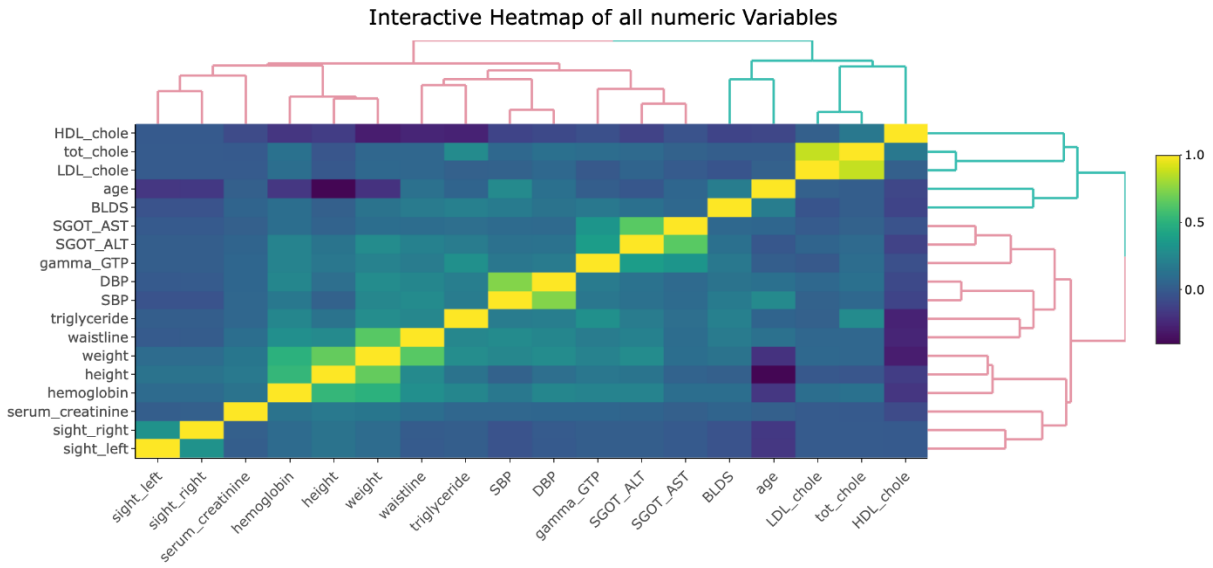

Fig 3.4.1.1 Heatmap

Furthermore, parameters such as sight_left, sight_right, hemoglobin, and weight exhibit moderate correlations with age. This heatmap serves as a valuable tool for identifying significant features influencing model performance, aiding in the informed selection of variables for further analysis and modeling.

3.4.2 Decision Tree

Decision tree was used for feature selection due to their intrinsic ability to identify and rank features based on their importance or contribution to predicting the target variable.

3.4.2.1 Feature Selection for DRK_YN Variable (Binary Classification)

The presented graph(Fig 3.4.2.1.1) illustrates the key variables influencing drinking behavior. The identified variables include sex, height, hemoglobin, gamma_GTP, age, and HDL_cholesterol.

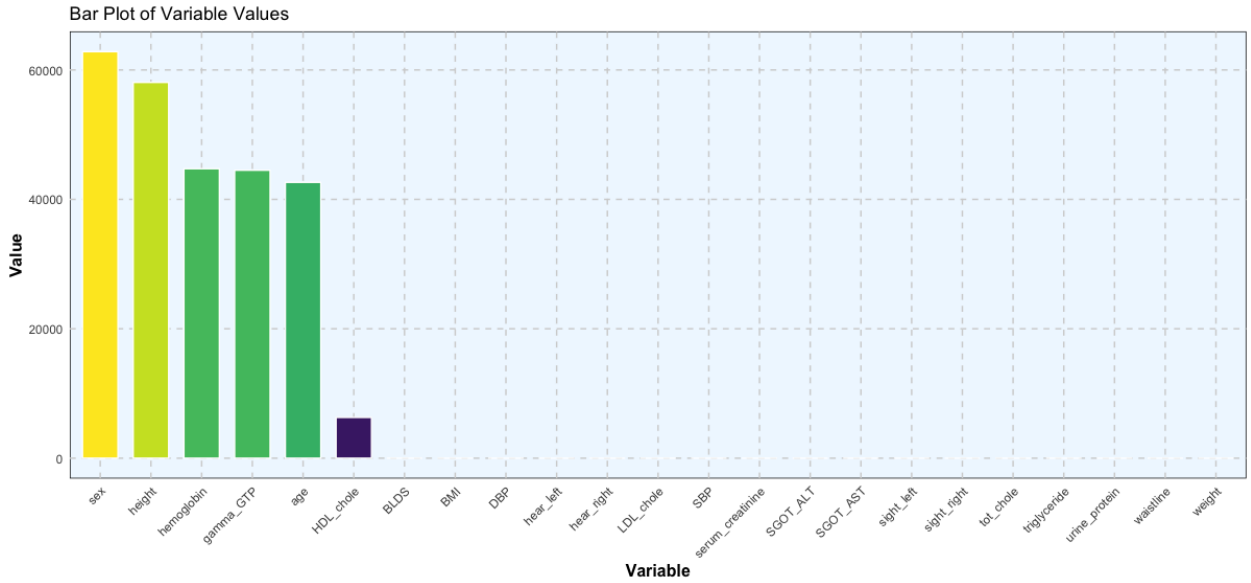

Fig 3.4.2.1.1 Bar plot for feature selection Drinking

Employing these critical variables, a binary classification was conducted to determine an individual's drinking status, categorized as either Y (Drinker) or N (Non-Drinker).

3.4.2.2 Feature Selection for SMK_stat_type_cd (Multi-class Classification)

The following graph (Fig 3.4.2.2.1) showcases the pivotal variables influencing smoking behavior. The identified variables encompass hemoglobin, height, gamma_GTP, weight, serum_creatinine, age, triglyceride, and waistline. Leveraging this set of crucial variables, a multi-level classification was executed to ascertain an individual's smoking status.

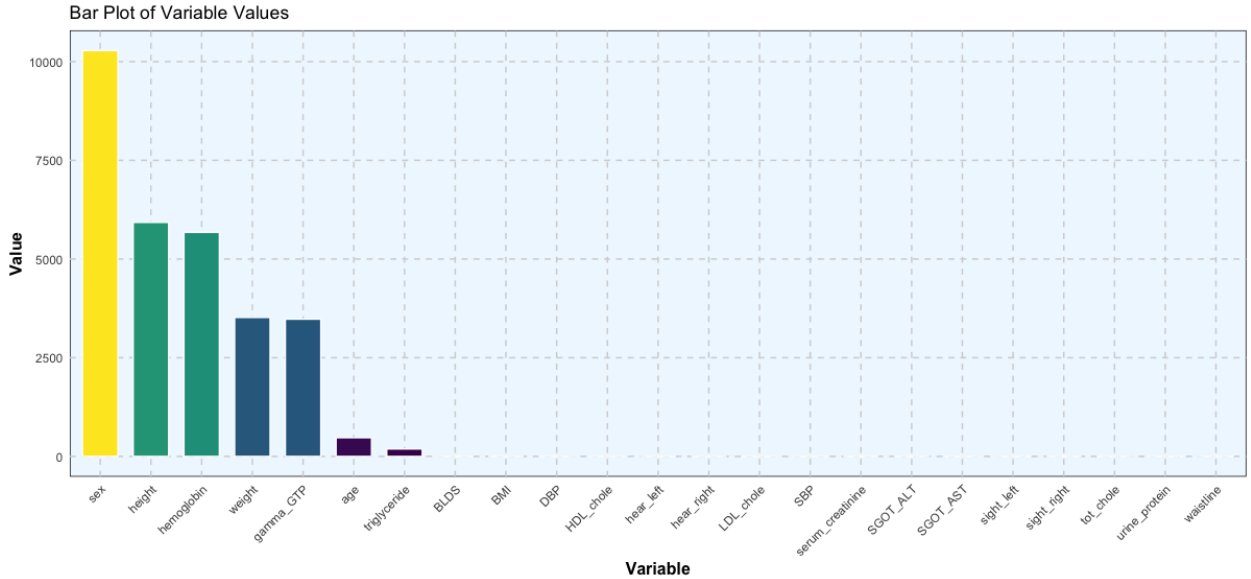

Fig 3.4.2.2.1 Bar plot for feature selection Drinking

3.5. Clustering

- An elbow diagram (Fig 3.5.1) was used to find out the optimal number of clusters which was k=2.
- The motive behind using clustering was to ensure that our training set is representative of the entire dataset by using stratified sampling. This means sampling a proportional number of observations from each cluster.
- We used the createDataPartition function to create a stratified random sample of indices based on the cluster assignment.

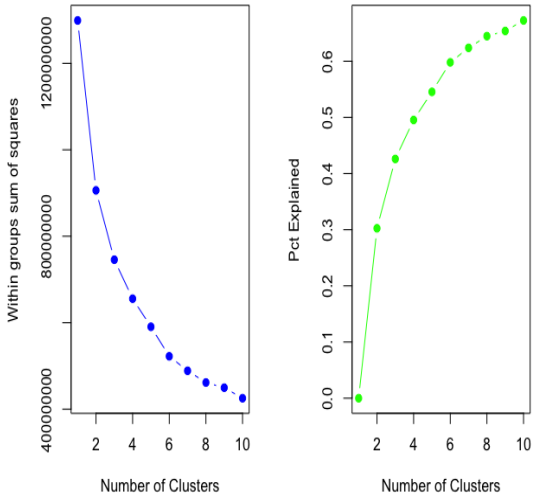

Fig 3.5.1 Elbow Diagram

Fig 3.5.2 Clustering

4. Modeling:

4.1 Expected approach for modeling:

Our predictive analysis focuses on two variables:

1. **SMK_stat_type_cd (Smoking Status):**
    a. Levels: 1 (never), 2 (used to smoke but quit), 3 (still smoke)
    b. Total Levels: 3
2. **DRK_YN (Drinker Status):**
    a. Levels: Y (Drinker), N (Non-Drinker)
    b. Total Levels: 2

For the Smoking Status variable (SMK_stat_type_cd), we conducted multiclass classification, while for the Drinker Status variable (DRK_YN), we performed binary classification.

4.2 Model Performance for Binary Classification:

Below is the various model performance summary:

					CV Performance	
Model	Method	Package	Hyperparameter	Selection	Accuracy	Kappa
Logistic Regression	glm	stats	NA	NA	0.722	0.444
Decision Tree	rpart	rpart	cp	0.00154	0.703	0.405
MARS	earth	earth	nprune	8	0.727	0.453
XGBoost	Xgbtree	`xgboost	nrounds,max_depth, eta, gamma,colsample_bytree,min_child_weight,subsampl	400,3, 0.1,0.01, 0.6, 0,0.75	0.732	0.464

4.2.1 XGBoost Model Performance

The XGBoost model emerged as the top-performing model in comparison to others.

Below is the confusion matrix (Fig 4.2.1.1) of the model:

```
Confusion Matrix and Statistics

          Reference
Prediction      N      Y
         N 248719  85826
         Y 101198 264257

               Accuracy : 0.733
                 95% CI : (0.732, 0.734)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.466

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.711
            Specificity : 0.755
         Pos Pred Value : 0.743
         Neg Pred Value : 0.723
              Precision : 0.743
                 Recall : 0.711
                     F1 : 0.727
             Prevalence : 0.500
         Detection Rate : 0.355
   Detection Prevalence : 0.478
      Balanced Accuracy : 0.733

       'Positive' Class : N

>
```

Fig 4.2.1.1 Confusion Matrix

4.2.2 XGBoost Confusion Matrix Summary:

- The model demonstrates a moderate overall accuracy of 73.3%, indicating its ability to correctly classify instances.
- The balanced accuracy of 73.3% suggests a fair trade-off between sensitivity and specificity, especially relevant when dealing with imbalanced datasets.
- The model achieves a precision of 74.3%, indicating that when it predicts a positive outcome, it

is correct approximately three-quarters of the time.

- The recall (or sensitivity) of 71.1% indicates the model's effectiveness in capturing a substantial proportion of actual positive instances.
- The Kappa statistic of 0.466 suggests moderate agreement beyond random chance, providing a more nuanced evaluation than accuracy alone.
- The p-value (less than 0.0000000000000002) for the comparison of accuracy to the No Information Rate (NIR) indicates a statistically significant improvement over random chance.

4.3 Model Performance for Multi-level Classification:

					CV Performance	
Model	Method	Package	Hyperparameter	Selection	Accuracy	Kappa
Logistic Regression	Multinom	nnet	NA	NA	0.664	0.343
Decision Tree	rpart	rpart	cp	0.00154	0.649	0.331
MARS Model	earth	earth	nprune	18	0.631	0.284
XGBoost model	Xgbtree	xgboost	Nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsampl	400	0.686	0.417

4.3.1 XgBoost Model Performance:

Here as well, XGboost had an overall better accuracy value, and we decided to pick that as our best-performing model.

```
> XGBoost_Model
eXtreme Gradient Boosting

56001 samples
   24 predictor
    3 classes: 'X1', 'X2', 'X3'

Pre-processing: centered (24), scaled (24)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 50400, 50401, 50401, 50401, 50400, 50402, ...
Resampling results:

  logLoss  AUC   prAUC  Accuracy  Kappa  Mean_F1
  0.685    0.83  0.616  0.688     0.42   0.574
  Mean_Sensitivity  Mean_Specificity
  0.571             0.814
  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
  0.586                0.824
  Mean_Precision  Mean_Recall  Mean_Detection_Rate
  0.586           0.571        0.229
  Mean_Balanced_Accuracy
  0.692

Tuning parameter 'nrounds' was held constant at
 at a value of 0
Tuning parameter 'subsample'
 was held constant at a value of 0.75
```

Fig 4.3.1.1 XGBoost Model Performance

4.3.2 XGboost model Summary:

- **Strong Discrimination:** Excellent AUC score of 0.83 indicates the model effectively distinguishes between classes.
- **Good Overall Accuracy:** 69% accuracy shows the model correctly classifies a majority of instances.
- **Balanced Performance:** Similar balanced accuracy (0.69) suggests fair performance across all classes, avoiding bias.
- **High Precision:** 74.3% precision means most positive predictions are accurate, reducing false positives.
- **Moderate Sensitivity:** 71.1% recall indicates the model captures a substantial portion of actual positive cases, minimizing false negatives.
- **Significant Improvement over Random Chance:** Statistically significant p-value demonstrates the model's performance surpasses mere guessing.

5. Conclusion and Insights:

- The combined results of both classification models can be collectively used to evaluate the parameters that help to detect smoking and drinking habits in a person.
- Some common medical parameters will be utilized to know whether a person is a smoker or a drinker. The following are the few common variables: Hemoglobin, gamma_GTP, and height.
- While predicting both drinking and smoking behavior collectively, attention should be given to the common key variables: hemoglobin, gamma_GTP, and age.
- These variables seem to have a notable impact on both behaviors, indicating potential shared factors influencing drinking and smoking.