reportA.R

xerzat

2024-08-23

#Q2  
  
#Load the data: read.csv()   
#Check the data structure: str() to clarify the types and structure   
kidneydata=read.csv("KidneyData.csv")  
attach(kidneydata)  
head(kidneydata)

## PatientID Age Gender BloodPressure BloodSugar Cholesterol BMI SmokingStatus  
## 1 TIW5219 120 Female 118 155.8 165 31.7 Former  
## 2 QLJ3151 10 Female 143 162.5 214 23.9 Never  
## 3 GRL2542 58 Female 300 120.8 222 16.3 Former  
## 4 WMM4122 22 Female 20 154.2 212 21.9 Never  
## 5 LPP8404 52 Female 150 158.9 600 23.8 Current  
## 6 CIH1298 53 Female 141 131.6 199 18.3 Former  
## ElectricConductivity pH DissolvedOxygen Turbidity TotalDissolvedSolids  
## 1 336.2 7.40 9.57 1.44 455.4  
## 2 297.3 7.48 8.49 1.21 423.2  
## 3 377.9 7.49 8.18 0.88 434.3  
## 4 312.0 6.03 7.35 1.15 400.4  
## 5 222.4 6.77 7.40 0.73 349.4  
## 6 422.0 7.34 8.00 0.71 426.6  
## NitriteLevel NitrateLevel LeadConcentration ArsenicConcentration Humidity  
## 1 0.165 1.97 0.0099 0.0063 48.7  
## 2 0.075 1.74 0.0120 0.0062 65.3  
## 3 0.005 1.40 0.0173 0.0092 93.2  
## 4 0.088 0.88 0.0133 0.0086 67.4  
## 5 0.119 0.71 0.0155 0.0011 43.3  
## 6 0.076 1.00 0.0050 0.0090 57.6  
## KidneyDisease  
## 1 0  
## 2 1  
## 3 0  
## 4 1  
## 5 1  
## 6 0

dim(kidneydata)

## [1] 500 19

str(kidneydata)

## 'data.frame': 500 obs. of 19 variables:  
## $ PatientID : chr "TIW5219" "QLJ3151" "GRL2542" "WMM4122" ...  
## $ Age : int 120 10 58 22 52 53 76 45 57 30 ...  
## $ Gender : chr "Female" "Female" "Female" "Female" ...  
## $ BloodPressure : int 118 143 300 20 150 141 194 151 140 141 ...  
## $ BloodSugar : num 156 162 121 154 159 ...  
## $ Cholesterol : int 165 214 222 212 600 199 251 200 215 205 ...  
## $ BMI : num 31.7 23.9 16.3 21.9 23.8 18.3 26.2 22.2 19.5 25.7 ...  
## $ SmokingStatus : chr "Former" "Never" "Former" "Never" ...  
## $ ElectricConductivity: num 336 297 378 312 222 ...  
## $ pH : num 7.4 7.48 7.49 6.03 6.77 7.34 7.01 7.46 7.38 6.7 ...  
## $ DissolvedOxygen : num 9.57 8.49 8.18 7.35 7.4 8 9.79 8.72 8.04 6.98 ...  
## $ Turbidity : num 1.44 1.21 0.88 1.15 0.73 0.71 1.16 0.98 1.47 1.1 ...  
## $ TotalDissolvedSolids: num 455 423 434 400 349 ...  
## $ NitriteLevel : num 0.165 0.075 0.005 0.088 0.119 0.076 0.177 0.044 0.114 0.042 ...  
## $ NitrateLevel : num 1.97 1.74 1.4 0.88 0.71 1 1.13 1.13 1.13 0.82 ...  
## $ LeadConcentration : num 0.0099 0.012 0.0173 0.0133 0.0155 0.005 0.012 0.0106 0.0128 0.0145 ...  
## $ ArsenicConcentration: num 0.0063 0.0062 0.0092 0.0086 0.0011 0.009 0.0035 0.0062 0.0081 0.0046 ...  
## $ Humidity : num 48.7 65.3 93.2 67.4 43.3 57.6 50.8 70.5 55.6 72.9 ...  
## $ KidneyDisease : int 0 1 0 1 1 0 1 1 0 1 ...

#Remove irrelevant variables such as PatientID which is an index variable and need to be removed for modelling  
kidneydata=kidneydata[,-1]  
#Convert categorical variables: Convert categorical variables into factors using as.factor().  
kidneydata$Gender=as.factor(kidneydata$Gender)  
kidneydata$SmokingStatus=as.factor(kidneydata$SmokingStatus)  
kidneydata$Gender=as.factor(kidneydata$Gender)  
kidneydata$KidneyDisease=as.factor(kidneydata$KidneyDisease)  
str(kidneydata)

## 'data.frame': 500 obs. of 18 variables:  
## $ Age : int 120 10 58 22 52 53 76 45 57 30 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 2 2 1 ...  
## $ BloodPressure : int 118 143 300 20 150 141 194 151 140 141 ...  
## $ BloodSugar : num 156 162 121 154 159 ...  
## $ Cholesterol : int 165 214 222 212 600 199 251 200 215 205 ...  
## $ BMI : num 31.7 23.9 16.3 21.9 23.8 18.3 26.2 22.2 19.5 25.7 ...  
## $ SmokingStatus : Factor w/ 3 levels "Current","Former",..: 2 3 2 3 1 2 1 1 2 2 ...  
## $ ElectricConductivity: num 336 297 378 312 222 ...  
## $ pH : num 7.4 7.48 7.49 6.03 6.77 7.34 7.01 7.46 7.38 6.7 ...  
## $ DissolvedOxygen : num 9.57 8.49 8.18 7.35 7.4 8 9.79 8.72 8.04 6.98 ...  
## $ Turbidity : num 1.44 1.21 0.88 1.15 0.73 0.71 1.16 0.98 1.47 1.1 ...  
## $ TotalDissolvedSolids: num 455 423 434 400 349 ...  
## $ NitriteLevel : num 0.165 0.075 0.005 0.088 0.119 0.076 0.177 0.044 0.114 0.042 ...  
## $ NitrateLevel : num 1.97 1.74 1.4 0.88 0.71 1 1.13 1.13 1.13 0.82 ...  
## $ LeadConcentration : num 0.0099 0.012 0.0173 0.0133 0.0155 0.005 0.012 0.0106 0.0128 0.0145 ...  
## $ ArsenicConcentration: num 0.0063 0.0062 0.0092 0.0086 0.0011 0.009 0.0035 0.0062 0.0081 0.0046 ...  
## $ Humidity : num 48.7 65.3 93.2 67.4 43.3 57.6 50.8 70.5 55.6 72.9 ...  
## $ KidneyDisease : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 2 2 1 2 ...

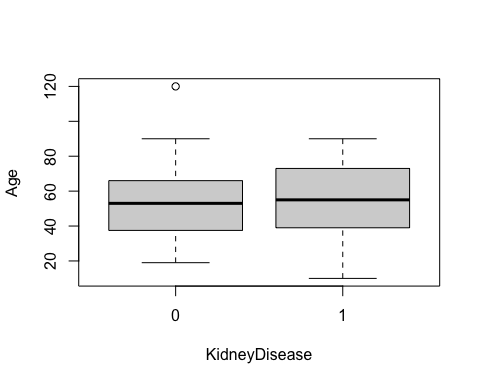
# split the data: 80% for training and 20% for testing.  
set.seed(2)  
tr.id = sample(1:nrow(kidneydata),nrow(kidneydata)\*0.8)  
train=kidneydata[tr.id, ]  
test=kidneydata[-tr.id,]  
str(train)

## 'data.frame': 400 obs. of 18 variables:  
## $ Age : int 71 71 61 83 87 42 50 52 46 77 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 1 2 1 1 1 ...  
## $ BloodPressure : int 175 172 147 170 145 144 157 161 155 172 ...  
## $ BloodSugar : num 62.8 86.5 89 110.8 39.1 ...  
## $ Cholesterol : int 243 227 196 253 200 179 204 215 204 240 ...  
## $ BMI : num 14.7 22.4 31 15.7 32.2 23.4 19.8 19.2 16.2 22.1 ...  
## $ SmokingStatus : Factor w/ 3 levels "Current","Former",..: 3 1 3 3 1 3 3 1 2 3 ...  
## $ ElectricConductivity: num 232 246 365 290 256 ...  
## $ pH : num 6.76 7.06 6.95 7.54 7.51 7.92 6.39 6.56 6.72 6.16 ...  
## $ DissolvedOxygen : num 8.36 8.29 8.85 8.96 8.25 7.79 6.37 9.11 8.06 8.27 ...  
## $ Turbidity : num 1.01 0.94 0.67 0.77 0.78 0.72 1.28 1.18 1.26 0.89 ...  
## $ TotalDissolvedSolids: num 391 422 469 411 355 ...  
## $ NitriteLevel : num 0.078 0.043 0.08 0.052 0.1 0.076 0.076 0.055 0.019 0.119 ...  
## $ NitrateLevel : num 0.48 0.77 0.33 0.67 1.41 0.73 1.94 0.7 1.16 0.28 ...  
## $ LeadConcentration : num 0.0189 0.0081 0.011 0.0123 0.0105 0.0146 0.0093 0.0056 0.0091 0.0116 ...  
## $ ArsenicConcentration: num 0.0061 0.0015 0.0075 0.0057 0.0067 0.0036 0.0063 0.0053 0.0038 0.0072 ...  
## $ Humidity : num 70.2 56.7 67.4 57.9 52.2 66.4 61.4 42.5 48.7 45.1 ...  
## $ KidneyDisease : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 2 ...

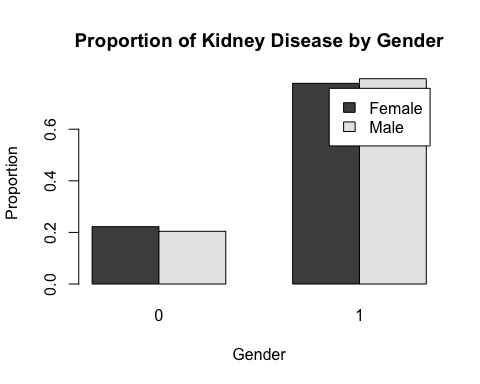
str(test)

## 'data.frame': 100 obs. of 18 variables:  
## $ Age : int 30 73 66 44 44 57 69 30 38 43 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 1 1 1 2 2 2 2 2 2 2 ...  
## $ BloodPressure : int 141 165 172 121 138 154 163 119 150 122 ...  
## $ BloodSugar : num 74.6 52.6 111.6 62.9 77.3 ...  
## $ Cholesterol : int 205 200 235 181 241 210 215 271 181 167 ...  
## $ BMI : num 25.7 22.4 24.4 26.8 27.5 20.4 16.6 24.7 26.5 21.7 ...  
## $ SmokingStatus : Factor w/ 3 levels "Current","Former",..: 2 2 2 1 2 3 3 1 1 2 ...  
## $ ElectricConductivity: num 252 317 188 228 360 ...  
## $ pH : num 6.7 7.96 6.87 7.06 7.43 6.11 6.56 7.06 6.57 8.19 ...  
## $ DissolvedOxygen : num 6.98 7.35 7.02 7.55 7.89 8.75 9.2 7.39 6.48 7.03 ...  
## $ Turbidity : num 1.1 0.9 1.07 1.27 1.08 0.76 1.06 0.72 1.31 1.22 ...  
## $ TotalDissolvedSolids: num 372 407 325 373 445 ...  
## $ NitriteLevel : num 0.042 0.09 0.065 0.14 0.126 0.09 0.134 0.094 0.159 0.124 ...  
## $ NitrateLevel : num 0.82 0.95 1.34 1.08 1.6 0.77 0.56 0.92 1.1 1.57 ...  
## $ LeadConcentration : num 0.0145 0.0088 0.0075 0.0121 0.0155 0.0137 0.0193 0.0164 0.0059 0.0045 ...  
## $ ArsenicConcentration: num 0.0046 0.006 0.0045 0.0066 0.0036 0.0051 0.0036 0.0021 0.0071 0.0035 ...  
## $ Humidity : num 72.9 55.2 64.8 52.5 58.4 60.3 37.2 41.2 59.4 63.5 ...  
## $ KidneyDisease : Factor w/ 2 levels "0","1": 2 1 2 2 1 2 2 2 2 2 ...

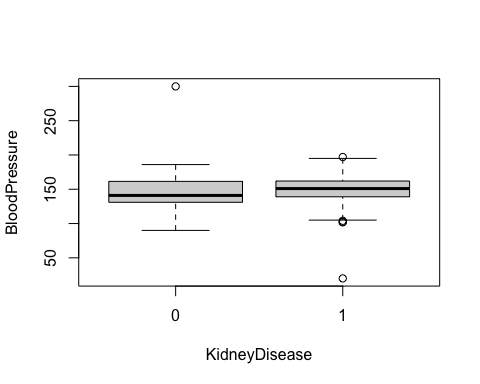
#visulization   
  
boxplot(Age~KidneyDisease, data = kidneydata)



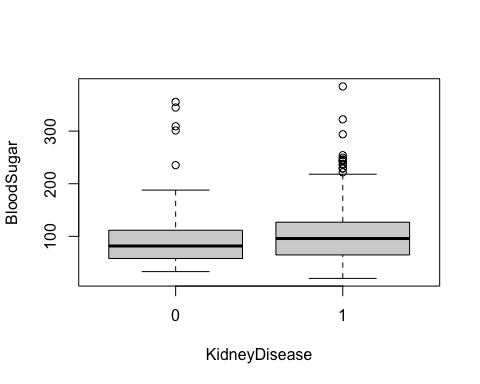
#The median age for individuals without or with kidney disease has no big difference,  
#but There is a slightly wider range of younger people ages among people with kidney disease.  
  
  
gender\_table <- table(kidneydata$Gender, kidneydata$KidneyDisease)  
gender\_prop <- prop.table(gender\_table, 1)  
barplot(gender\_prop, beside = TRUE,  
 legend=TRUE,  
 xlab = "Gender",  
 ylab = "Proportion",  
 main = "Proportion of Kidney Disease by Gender")



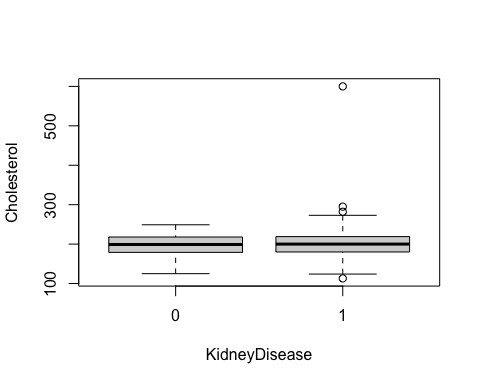
#We can see that there is no notable difference in gender proportions with regard to kidney disease.  
boxplot(BloodPressure~KidneyDisease, data = kidneydata)



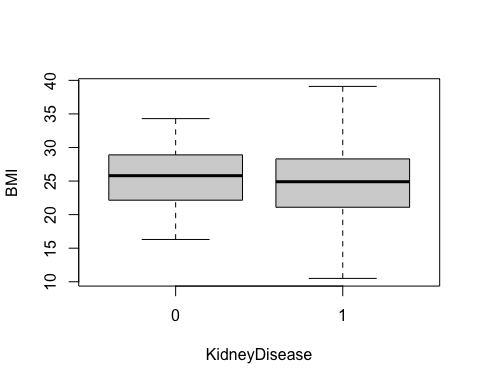
#those who have KidneyDiseases tend to have higher blood pressure median and lower range  
  
boxplot(BloodSugar~KidneyDisease, data = kidneydata)



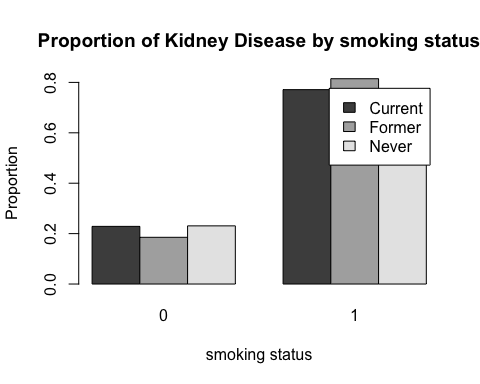
#those who have KidneyDiseases tend to have higher BloodSugar median and wider range ,also more oulier on the higher bloodsugar side.  
  
boxplot(Cholesterol~KidneyDisease, data = kidneydata)



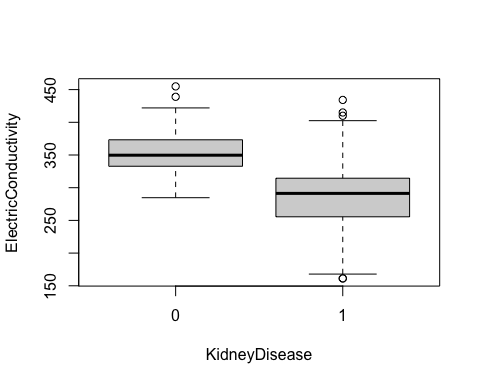
#We can see that there is almost no difference in Cholesterol median with regard to kidney disease.  
boxplot(BMI~KidneyDisease, data = kidneydata)



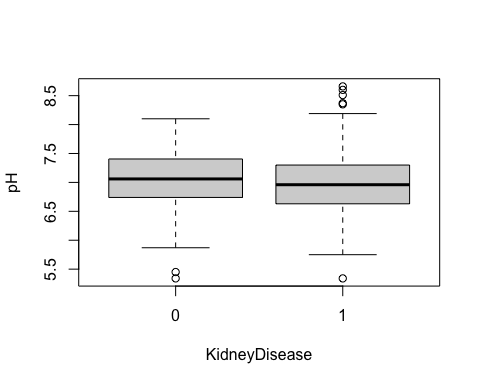
#There appears to be no significant difference in kidney disease between people with low BMI and those with higher BMI.  
  
  
smoking\_table <- table(kidneydata$SmokingStatus, kidneydata$KidneyDisease)  
smoking\_prop <- prop.table(smoking\_table, 1)  
barplot(smoking\_prop, beside = TRUE,  
 legend=TRUE,  
 xlab = "smoking status",  
 ylab = "Proportion",  
 main = "Proportion of Kidney Disease by smoking status")



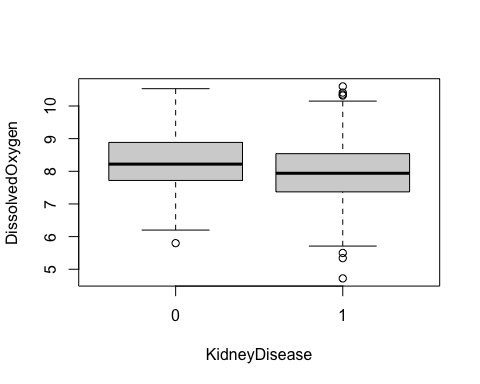
#People who have never smoked and those who currently smoke have similar percentages of developing or not developing kidney disease.Therefore smoking status has no significant relationship with kidneydiseas  
  
boxplot(ElectricConductivity~KidneyDisease, data = kidneydata)



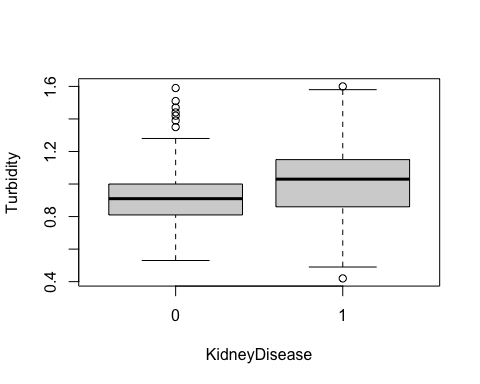
#those who have KidneyDiseases tend to have significantly lower ElectricConductivity median and wider range, those who does not have kidneay diseas higher median of ElectricConductivity and lower range.  
  
  
boxplot(pH~KidneyDisease, data = kidneydata)



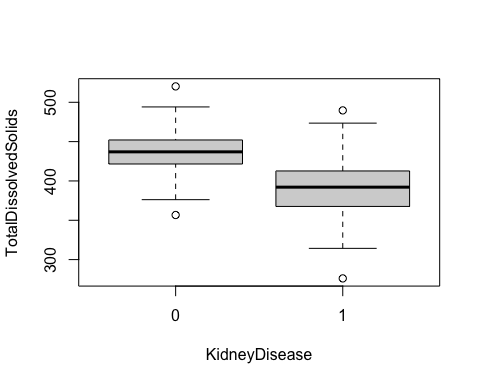
#there is no significant difference with pH median with kidneyKidneyDiseases status  
  
boxplot(DissolvedOxygen~KidneyDisease, data = kidneydata)



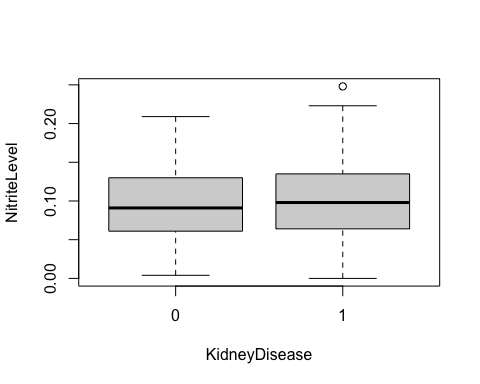
#those who have KidneyDiseases tend to have lower DissolvedOxygen median and more outlier  
  
boxplot(Turbidity ~KidneyDisease, data = kidneydata)



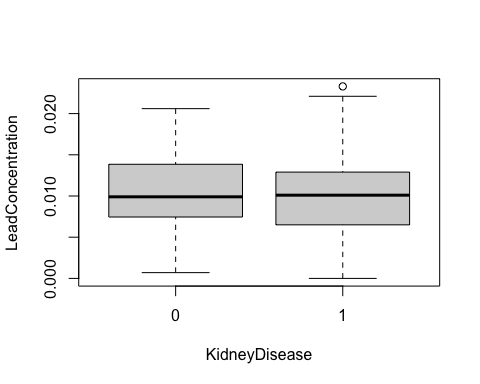
#those who have KidneyDiseases tend to have higher Turbidity median and wider range,those who don'n have KidneyDiseases have lower median and more outlier.  
  
boxplot(TotalDissolvedSolids ~KidneyDisease, data = kidneydata)



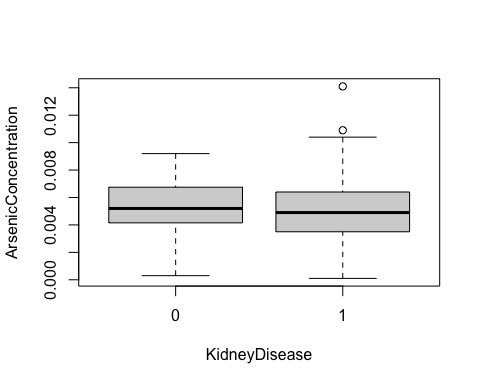
#those who have KidneyDiseases tend to have lower TotalDissolvedSolids median and wider range  
  
boxplot(NitriteLevel ~KidneyDisease, data = kidneydata)



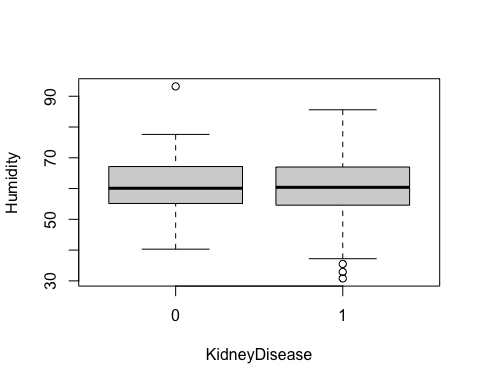
#there is no significant difference with NitriteLevel median with kidneyKidneyDiseases status  
  
boxplot(LeadConcentration ~KidneyDisease, data = kidneydata)



#there is no significant difference with LeadConcentration median with kidneyKidneyDiseases status  
  
boxplot(ArsenicConcentration ~KidneyDisease, data = kidneydata)



#there is no significant difference with ArsenicConcentration median with kidneyKidneyDiseases status  
  
boxplot(Humidity ~KidneyDisease, data = kidneydata)



#there is no significant difference with Humidity median with kidneyKidneyDiseases status  
  
#4. Use logistic regression to answer the research question. Clearly explain the process or all the steps involved   
   
#Model Building:  
 model1 <- glm(KidneyDisease ~ ., data = train, family = binomial)  
 summary(model1)

##   
## Call:  
## glm(formula = KidneyDisease ~ ., family = binomial, data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.295e+01 5.871e+00 5.613 1.99e-08 \*\*\*  
## Age -8.358e-03 1.038e-02 -0.805 0.420712   
## GenderMale -2.444e-01 3.731e-01 -0.655 0.512501   
## BloodPressure 3.352e-02 9.444e-03 3.550 0.000386 \*\*\*  
## BloodSugar -4.042e-05 3.360e-03 -0.012 0.990400   
## Cholesterol 4.676e-03 6.246e-03 0.749 0.454059   
## BMI 1.564e-02 3.596e-02 0.435 0.663568   
## SmokingStatusFormer 2.291e-01 4.724e-01 0.485 0.627684   
## SmokingStatusNever -3.113e-01 4.755e-01 -0.655 0.512658   
## ElectricConductivity -3.376e-02 6.050e-03 -5.579 2.42e-08 \*\*\*  
## pH -1.069e+00 3.766e-01 -2.839 0.004532 \*\*   
## DissolvedOxygen -5.221e-01 2.124e-01 -2.459 0.013949 \*   
## Turbidity 3.541e+00 8.958e-01 3.952 7.74e-05 \*\*\*  
## TotalDissolvedSolids -4.245e-02 8.655e-03 -4.905 9.36e-07 \*\*\*  
## NitriteLevel 5.311e+00 3.855e+00 1.378 0.168246   
## NitrateLevel 3.127e-02 3.926e-01 0.080 0.936526   
## LeadConcentration 5.542e+00 3.679e+01 0.151 0.880251   
## ArsenicConcentration -6.033e+00 9.294e+01 -0.065 0.948241   
## Humidity -1.691e-02 2.075e-02 -0.815 0.415067   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 408.50 on 399 degrees of freedom  
## Residual deviance: 204.82 on 381 degrees of freedom  
## AIC: 242.82  
##   
## Number of Fisher Scoring iterations: 7

#Using the hypothesis testing, it can be seen clearly that BloodPressure,ElectricConductivity,pH,DissolvedOxygen,Turbidity,TotalDissolvedSolids have significant relationship with kidneydiseas .  
  
#model improvement  
model2=glm(KidneyDisease~BloodPressure+ElectricConductivity+pH+DissolvedOxygen+Turbidity+TotalDissolvedSolids, data = train,family =binomial)  
  
#Use summary(model) to evaluate significance.  
summary(model2)

##   
## Call:  
## glm(formula = KidneyDisease ~ BloodPressure + ElectricConductivity +   
## pH + DissolvedOxygen + Turbidity + TotalDissolvedSolids,   
## family = binomial, data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 31.128553 4.899149 6.354 2.10e-10 \*\*\*  
## BloodPressure 0.029560 0.008208 3.602 0.000316 \*\*\*  
## ElectricConductivity -0.033188 0.005779 -5.743 9.29e-09 \*\*\*  
## pH -0.969937 0.364296 -2.662 0.007756 \*\*   
## DissolvedOxygen -0.528639 0.198907 -2.658 0.007867 \*\*   
## Turbidity 3.416778 0.861470 3.966 7.30e-05 \*\*\*  
## TotalDissolvedSolids -0.037734 0.007722 -4.887 1.03e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 408.50 on 399 degrees of freedom  
## Residual deviance: 211.21 on 393 degrees of freedom  
## AIC: 225.21  
##   
## Number of Fisher Scoring iterations: 7

#To predict KidneyDisease probability by model2;  
testdata=test[-18]  
glm\_prob =predict(model2, type="response",newdata=testdata)  
  
#Evaluate model performance on the test set using confusion matrix .  
glm\_pred=rep("0",100)  
glm\_pred[glm\_prob>0.5]="1"  
table(glm\_pred,test$KidneyDisease)

##   
## glm\_pred 0 1  
## 0 17 5  
## 1 7 71

misclassification\_rate=(5+7)/100  
misclassification\_rate

## [1] 0.12

#Give your resultant model  
coef(model2)

## (Intercept) BloodPressure ElectricConductivity   
## 31.12855285 0.02955996 -0.03318803   
## pH DissolvedOxygen Turbidity   
## -0.96993682 -0.52863860 3.41677763   
## TotalDissolvedSolids   
## -0.03773373

#logit(P(KidneyDisease))= 31.12855285 + 0.02955996 \*BloodPressure -0.03318803 \*ElectricConductivity -0.96993682 \*pH− -0.52863860 \*DissolvedOxygen+ 3.41677763 \*Turbidity -0.03773373 \*TotalDissolvedSolids