# ASSIGNMENT 2 DATA MINING (STQD6414)

**Name:** Nur Diana Binti Abdul Kahar
**No matrix:** P137263

**1.**

```
data<-read.csv(file.choose())

str(data)          # To check the structure of data.

summary(data)          # To check summary of data

par(mfrow = c(2, 3))          # To look at multiple plots at once
```

**a)**

- Classes

```
boxplot (data$class, main = "Class of patients that are not and with Chronic Kidney Disease",
col = "blue")       # Boxplot for Class of Patients
```

- Age

```
boxplot(data$age, main = "Age of patients with Chronic Kidney Disease", ylab = "Years", col =
"green")        # Boxplot for Age of Patients
```

- Glucose

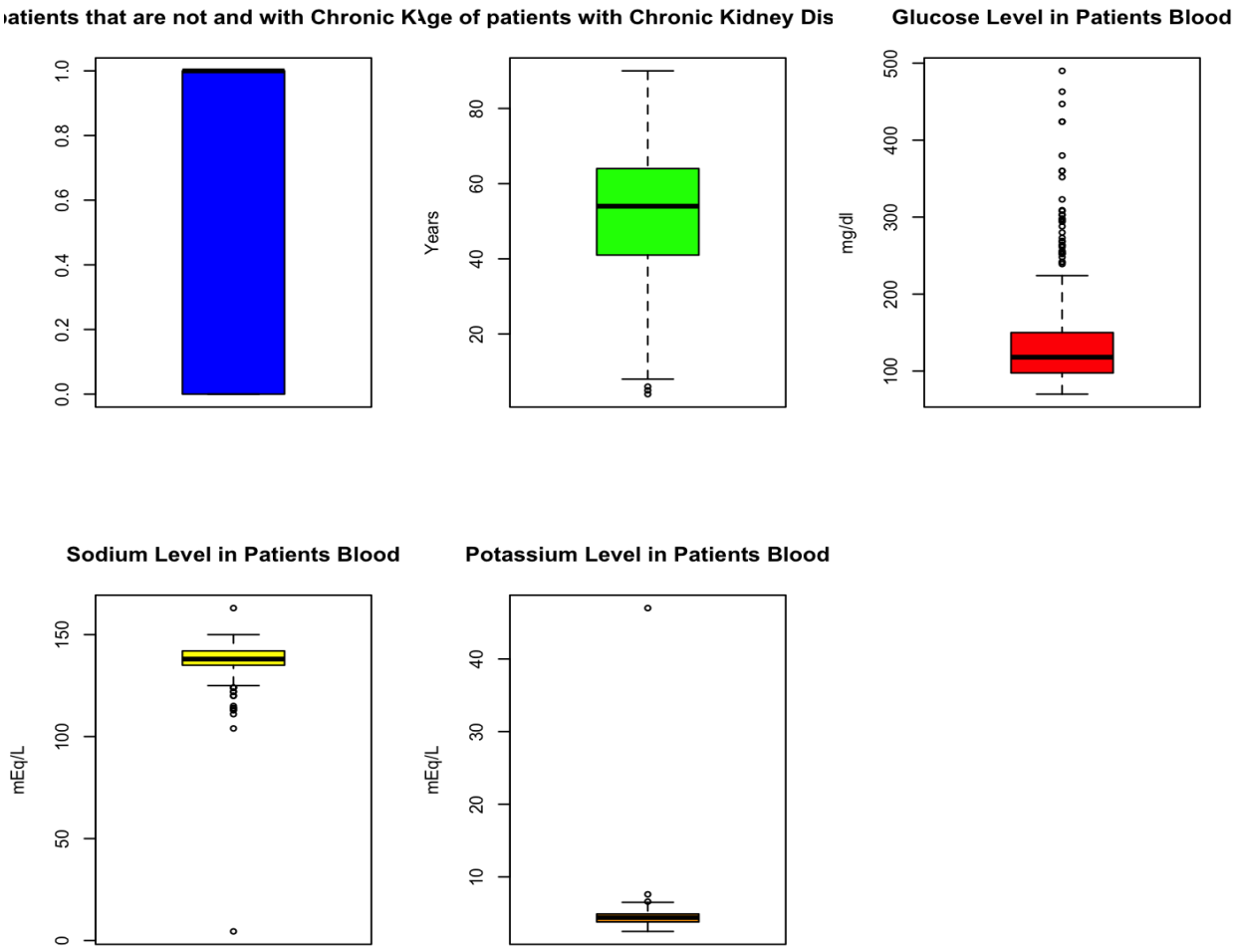
```
boxplot(data$glu, main = "Glucose Level in Patients Blood", ylab = "mg/dl", col = "red")
# Boxplot for Glucose Level
```

- Sodium

```
boxplot(data$sod, main = "Sodium Level in Patients Blood", ylab = "mEq/L", col = "yellow")
# Boxplot for Sodium Level
```

- Potassium

```
boxplot(data$pot, main = "Potassium Level in Patients Blood", ylab = "mEq/L", col = "orange")
# Boxplot for Potassium Level
```




**b)**

- Age

For the age of patients it shows that there are 3 outliers under the lower inner fence of the boxplot, meaning there are patients that are under 10 that have developed chronic kidney disease. These outliers will be kept due to several possibilities for children under 10 to develop this disease. For example, birth defects like posterior urethral valves and diseases found in families, such as polycystic kidney disease, are the most common causes of chronic kidney disease in children. Also, urinary tract abnormalities could potentially lead to chronic kidney disease.

- Glucose

There are multiple outliers above the upper inner fence for glucose level in patients' blood. The multiple outliers show that some of the patients with chronic kidney disease may also have hyperglycemia or diabetes. According to the Centers for Disease Control and Prevention (CDC) website, it is uncommon for people with Chronic Kidney Disease (CKD) to have elevated levels of glucose in their blood. The expected values for normal fasting blood glucose concentration are between 70 mg/dL (3.9 mmol/L) and 100 mg/dL (5.6 mmol/L). A fasting blood sugar level from 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes. If it's 126 mg/dL (7 mmol/L) or higher on two separate tests, you have diabetes. The highest glucose level ever reported was 2000mg/dl. Chronic Kidney Disease can impact the body's ability to regulate glucose due to various factors, such as insulin resistance, reduced insulin production, and impaired kidney function. With all these retaining facts, i feel like these outliers will be retained and kept.

- Sodium

Based on the boxplot, it shows that there are multiple outliers under the lower inner fence and only one outlier above the upper inner fence. The normal range for blood sodium level is 137-143 mEq/L. Anything under 137 mEq/L is considered Hyponatremia and readings above 143 mEq/L is Hypernatremia. These extreme cases often present as medical emergencies and are associated with life-threatening symptoms such as seizures, coma, and respiratory arrest. As it is mentioned on the National Kidney Foundation website, that Hyponatremia and Hypernatremia is a common sequelae of chronic kidney disease (CKD). However, the outlier under 100 mEq/L will be discarded as there are no known cases reported of someone with a sodium level under 10 mEq/L. This data may be a possible typo. Also, the outlier above the upper inner fence that is above 150 mEq/L will also be discarded due to another possible typo.

- Potassium

The boxplot shows that there are 3 outliers above the upper inner fence. The normal range for potassium levels in blood is 3.5-5.2 mEq/L. Readings above 6.0 mEq/L are life threatening and called Hyperkalemia. Hyperkalemia is commonly associated with patients with Chronic Kidney Disease. Extreme hyperkalemia (such as levels around 8.5 mEq/L) is relatively uncommon and is typically considered a medical emergency due to the severe risks it poses to heart function. It requires immediate intervention to lower potassium levels and stabilize the individual. The extreme outlier above the upper inner fence (above 40 mEq/L) will be discarded as a possible wrongly recorded data and the other outliers will be kept.

**2.**

| Variables | AIC |
|---|---|
| $\{X_1\}$ | 103.6723 |
| $\{X_2\}$ | 76.4944 |
| $\{X_3\}$ | 73.2174 |
| $\{X_4\}$ | 111.7761 |
| $\{X_1, X_2\}$ | 74.4070 |
| $\{X_1, X_3\}$ | 75.2171 |
| $\{X_1, X_4\}$ | 82.7902 |
| $\{X_2, X_3\}$ | 63.1980 |
| $\{X_2, X_4\}$ | 77.1644 |
| $\{X_3, X_4\}$ | 63.9084 |
| $\{X_1, X_2, X_3\}$ | 65.1772 |
| $\{X_1, X_2, X_4\}$ | 75.9472 |
| $\{X_1, X_3, X_4\}$ | 61.3073 |
| $\{X_2, X_3, X_4\}$ | 63.3784 |
| $\{X_1, X_2, X_3, X_4\}$ | 63.0223 |

a) Forward Selection

| **STEP 1:** | **AIC** | **EXPLANATION** |
|---|---|---|
| (X3) | 73.2174 | Choose ‘X3’ due to it having the lowest AIC value. |
| **STEP 2:** | | In step 2, only choose the AIC with ‘X3’ model. |
| (X1, X3) | 75.4.2171 | By adding X1 to the model, it increases the AIC which makes it worse. ‘X1, X3’ won't be chosen |
| (X2, X3) | 63.1980 | This value of AIC will be chosen by adding ‘X2” with the previous AIC value ‘X3’, it decreases the value, making it a better model. |
| (X3, X4) | 63.9084 | This AIC model won't be chosen because it isn't the smallest value. |
| **STEP 3:** | | In step 3, AIC value with ‘X2, X3’ will be chosen. |
| (X1, X2, X3) | 65.1772 | By adding ‘X1’, it increases the AIC value. So it won't be chosen. |
| (X2, X3, X4) | 63.3784 | This is the lowest AIC model in step 3 by adding ‘X4’, However, it increases the AIC value in step 2. So this AIC model will also not be chosen. |

= The final AIC model chosen in forward selection is (X2, X3) with 63.1980 with the lowest value.

b) Backward Selection

| STEP 1 | AIC | EXPLANATION |
|---|---|---|
| (X1, X2, X3, X4) | 63.0223 | Start with this AIC model. |
| **STEP 2** | | In step 2, Choose the AIC value with the lowest/smallest value. |
| (X1, X3, X4) | 61.3073 | Subtracts 'X2' to get the lowest AIC value. This AIC model will be chosen because this is the smallest AIC value among the rest. |
| **STEP 3** | | In step 3, will only choose the AIC model that has 'X1','X3','X4' only, since 'X2' has been subtracted. |
| (X1, X3) | 75.217 | This AIC model will not be chosen because it makes the AIC value increase. |
| (X1, X4) | 82.7902 | This AIC model will also not be chosen due to the increased AIC value. |
| (X3, X4) | 63.9084 | This is the smallest AIC value among the 3 AIC models in step 3, However it increases the AIC value from the AIC value in step 2. So this AIC model will also not be chosen. |

= The final AIC model chosen from backward selection is (X1, X3, X4) with 61.3073 because it has the lowest AIC value.

c) In Forward Selection the AIC model chosen is (X2, X3) with 63.1980 and in Backward Selection the AIC model chosen is (X1, X3, X4) with 61.3073. These 2 answers differ because in forward selection 'X2' was included and chosen in step 2, while in backward selection 'X2' was subtracted and was not included in the further steps. With 'X2' the best model is 'X2, X3' with the lowest value that includes 'X2'. However the best answer between forward selection and backward selection is backward selection with (X1, X3, X4) with 61.3073 due to it having the lowest AIC value.