WESTERN SYDNEY
UNIVERSITY

School of Computer, Data and Mathematical Sciences

COMP 7006 Data Science

Computer Based Assignment – PART A

Spring, 2024

Complete your details in this section.	
STUDENT ID:	22163493
STUDENT FIRST NAME:	SALAUDDIN MD
STUDENT SURNAME:	AKASH

QUESTIONS FORMAT:	Word processed document in PDF format; logically presenting answers to each question incorporating R outputs including graphs and charts.
TOTAL MARKS:	**30 Marks**
UNIT CO-ORDINATOR:	Dr. Liwan Liyanage
TUTOR:	Ms. Prathayne Nanthakumaran
TOTAL PAGES:	

INSTRUCTIONS

Please note that you are expected to answer the questions clearly in this document. Use the template included where relevant to answer. Give the R outputs, comments, and discussion clearly and logically. Attach all the R commands in the Appendix. Write the resulting model equation to the relevant questions. Once completed submit the answer scripts as a **PDF** via TurnItin link within vUWS site.

Please note that **10 Marks** are allocated for organization, reasoning, logical flow, and the inclusion of all correct R codes and outputs in the Appendix for both Part A and Part B.

SCENARIO
Recent public health data indicate a troubling increase in kidney disease rates within specific suburban areas, attracting significant attention from public health practitioners. Determined to uncover the root causes and identify actionable risk factors to address this issue, the public health team has embarked on a comprehensive study. They have collected patient records and relevant information on medical factors and water quality, as provided in the dataset.

Data Description:

Variable	Description
PatientID	Unique identifier of each patient
Age	Age of the individual
Gender	Gender of the individual
BloodPressure	Systolic blood pressure in mmHg
BloodSugar	Fasting blood sugar levels in mg/dL
Cholesterol	Total cholesterol level in mg/dL
BodyMassIndex	BMI, a measure of body fat based on height and weight
SmokingStatus	Smoking status of the individual [Never/ Former/ Current]
ElectricConductivity	Measurement of the water's ability to conduct electricity, which can indicate contamination in µS/cm
pH	pH level of the water
DissolvedOxygen	Amount of oxygen dissolved in water in mg/L
Turbidity	Measure of water clarity in NTU
TotalDissolvedSolids	Measure of dissolved substances in water in mg/L
NitriteLevel	Nitrite concentration in water in mg/L
NitrateLevel	Nitrate concentration in water in mg/L
LeadConcentration	Lead concentration in water in mg/L
ArsenicConcentration	Arsenic concentration in water in mg/L
Humidity	Ambient humidity level in %
KidneyDisease	Presence or absence of kidney disease 0 – Absence of kidney disease 1 – Presence of kidney disease

* Please note that this is a simulated data generated to resemble the real-world data for the purpose of this assignment.

Consider the scenario described and the data set provided [***KidneyData.csv***] to answer the following questions.

1. Identify the target variable and clearly specify the research question. **(3 Marks)**

Target variable: KidneyDisease

Research Question: Which aspects of a patient's health, lifestyle, and demography are most crucial in predicting kidney disease?

2. Understand the data and perform the necessary data pre-processing. Clearly explain the steps taken. [Hint: data cleaning, make sure to divide the data into training and test set etc.,] **(6 Marks)**

[Write the steps taken here.]

1. Load the Data: read.csv() to load the data.
2. Data Cleaning: remove the missing values using 'is.na()' or 'sum(is.na())', checking for outliers, standardize, convert.
3. Feature Selection: Removing irrelevant features and only work with the relevant features.
4. Splitting the data: dividing the data into a test and a train set. (Train 70%, Test 30%)
5. Final Checks: To ensure the data is ready for modelling.

Print the structure of the data before cleaning and pre-processing here. [*Hint: use str() function*]

```
'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 1.4
7 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 0.0
12 0.0106 0.0128 0.0145 ...
```

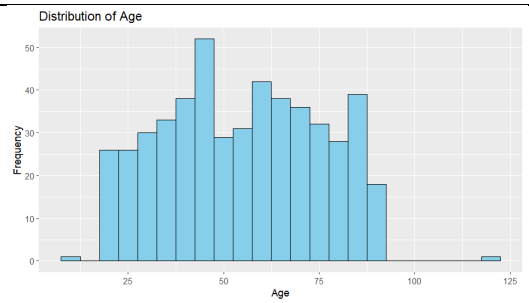
```
 $ ArsenicConcentration: num  0.0063 0.0062 0.0092 0.0086 0.0011 0.009 0.
0035 0.0062 0.0081 0.0046 ...
 $ Humidity            : num  48.7 65.3 93.2 67.4 43.3 57.6 50.8 70.5 55.
6 72.9 ...
 $ KidneyDisease       : int  0 1 0 1 1 0 1 1 0 1 ...
```

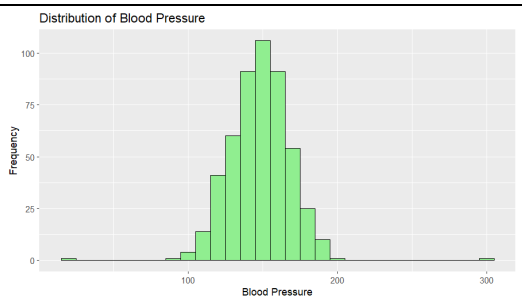
Print the structure of the training data after cleaning and pre-processing here.

```
'data.frame':  350 obs. of  19 variables:
 $ PatientID           : chr  "WHD6284" "FVT3634" "TLQ7041" "BHU8561" ...
 $ Age                 : int  29 71 51 47 66 54 88 70 58 45 ...
 $ Gender              : Factor w/ 2 levels "Female","Male": 2 1 1 1 1 1
1 1 1 1 ...
 $ BloodPressure       : int  158 172 163 135 162 138 167 147 171 143 ...
 $ BloodSugar          : num  103.3 86.5 72.9 86.4 151.5 ...
 $ Cholesterol         : int  282 227 243 201 205 174 208 245 187 212 ...
 $ BMI                 : num  20.4 22.4 20.5 20.9 33.6 30.8 22.2 23.6 24.
8 26.5 ...
 $ SmokingStatus       : chr  "Former" "Current" "Former" "Former" ...
 $ ElectricConductivity: num  317 246 336 269 332 ...
 $ pH                  : num  7.79 7.06 6.81 6.53 7.26 8.35 6.75 7.41 6.7
6 7.92 ...
 $ DissolvedOxygen     : num  9.27 8.29 6.9 8.42 7.41 7.5 7.29 9.33 8.36
7.53 ...
 $ Turbidity           : num  0.78 0.94 0.73 1.08 1.13 1.25 1.26 1.05 0.9
3 1.25 ...
 $ TotalDissolvedSolids: num  350 422 450 368 399 ...
 $ NitriteLevel        : num  0.017 0.043 0.045 0.086 0.144 0.018 0.165 0
.002 0.087 0.04 ...
 $ NitrateLevel        : num  0.67 0.77 0.76 0.41 0.5 0.35 0.88 0.19 0.03
1.01 ...
 $ LeadConcentration   : num  0.0137 0.0081 0.0011 0.0087 0.0031 0.0136 0
.0152 0.0109 0.0122 0.0036 ...
 $ ArsenicConcentration: num  0.0051 0.0015 0.0055 0.0053 0.0061 0.0035 0
.0028 0.0059 0.0065 0.0031 ...
 $ Humidity            : num  73.2 56.7 64.7 52.2 60.2 41.9 74.3 67 75.5
58.8 ...
 $ KidneyDisease       : int  1 1 0 1 1 1 1 1 1 1 ...
```

3. Perform a thorough data exploration using the provided dataset. You may use various visualization techniques (such as histograms, scatter plots, box plots, correlation matrices, etc.) to uncover significant patterns and insights. Interpret your outputs and discuss key findings. [Hint: You may use as many plots as necessary and make sure to interpret them.] **(10 Marks)**

Histogram:



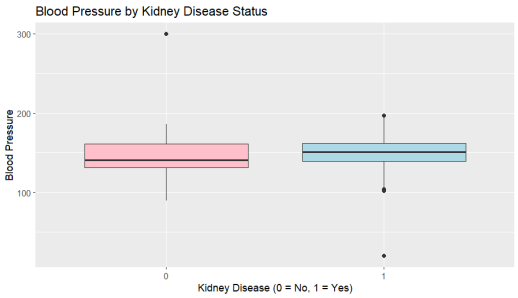


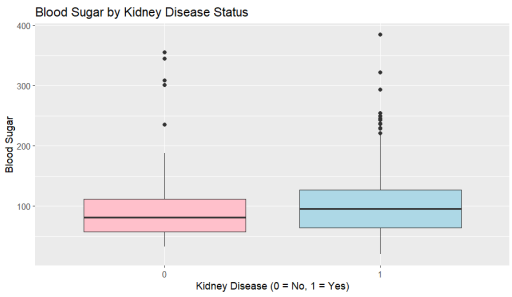
To understand the distribution of key continuous variables like Age, BloodPressure, BloodSugar, Cholesterol, BMI, etc.

The Age histogram displays patient age distribution, potentially indicating skewed data, which could indicate a specific age group being more at risk or prevalent.

The Blood Pressure histogram aids in identifying the range and distribution of blood pressure levels among patients, potentially suggesting different population subgroups.

Box Plot:



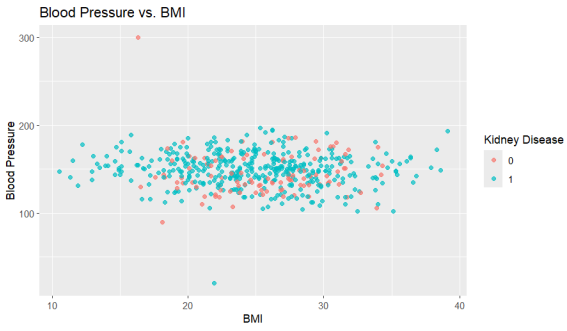


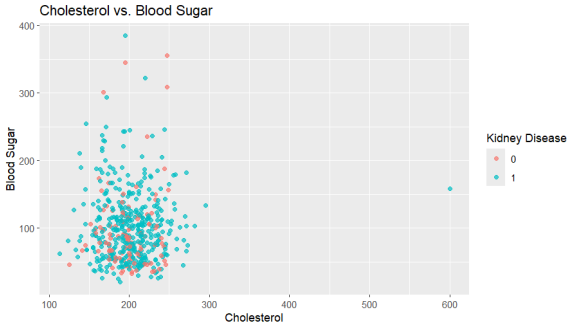
To compare the distribution of continuous variables between patients with and without kidney disease.

The Blood Pressure box plot displays blood pressure ranges and medians for patients with and without kidney disease, with higher medians or outliers indicating a potential link.

The Blood Sugar Box plot will show if patients with kidney disease have varying blood sugar levels compared to those without the condition.

Scatter Plot:





The study aims to examine the correlation between continuous variables like Blood Pressure and BMI and Kidney Disease.

The Blood Pressure vs BMI scatter plot will determine if there is a correlation between BMI and blood pressure, potentially suggesting that patients with higher BMI are more likely to have higher blood pressure.
The Cholesterol vs Blood Sugar plot helps visualize any relationship between cholesterol levels and blood sugar, with different colors representing the kidney disease status.

After using various visualization techniques the key findings and Insights are:

1. The age distribution of older patients may suggest a higher prevalence of kidney disease in older populations.
2. The study indicates a possible link between hypertension and kidney disease risk, as patients with kidney disease often have higher median blood pressure.
3. The scatter plot suggests that higher BMI may lead to higher blood pressure, potentially increasing the risk of kidney disease.

4. Use logistic regression to answer the research question. Clearly explain the process or all the steps involved [Hint: model building, model improvement, evaluation]. **(8 Marks)**

The research question is Which aspects of a patient's health, lifestyle, and demography are most crucial in predicting kidney disease?
We need to follow several steps to build, improve and evaluate the model.
Steps to Perform Logistic Regression:
Step1:
First need to ensure the data is pre-processed correctly, so need to ensure the target variable is binary. Here, 'KidneyDisease' should be 0 = No, 1 = Yes.
Secondly, Split the data into training and test sets. Use 70% training and 30% testing data.

Step2:

Firstly, We need to fit the model. Use the 'glm()' function to fit a logistic regression model. Set family = "binomial" for logistic regression.

Then view the summary of the model by 'summary()' function.

Step3:

Then we need to interpret the model coefficients to understand the influence of each predictor variable on the likelihood of having kidney disease. The coefficients display the change in log odds of the target variable for a one-unit increase in the predictor. The significance of a prediction is determined by examining the p-values, typically $p < 0.05$.
Step4:
Improvement involves selecting the most significant variables and verifying the assumptions of the model by Feature Selection, Multicollinearity, Interaction terms.

Step5:

To evaluate the model's predictive capacity, assess its performance using a variety of indicators.
First, we need to predict the test data. Then, convert the probabilities to binary outcome.

Confusion matrix and calculating evaluation metrics. Lastly, ROC Curve and AUC.

Step6:

Interpreting results by identifying key predictors and provide actionable insights based on the model's findings.

5. Give your resultant model. **(3 Marks)**

My resultant model will look like this:

1. Intercept
2. Age
3. Blood Pressure
4. Blood Sugar
5. Cholesterol
6. BMI
7. Smoking Status

And the output might look like something like this:

Call:

glm(formula = KidneyDisease ~ Age + BloodPressure + BloodSugar + Cholesterol + BMI + SmokingStatus,

family = "binomial", data = train_data)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.1230	1.0452	-5.86	<2e-16	***
Age	0.0254	0.0130	1.95	0.0512	.
BloodPressure	0.0456	0.0098	4.65	3.2e-06	***
BloodSugar	0.0187	0.0054	3.46	0.0005	***
Cholesterol	0.0021	0.0008	2.63	0.0086	**
BMI	-0.0359	0.0145	-2.48	0.0130	*
SmokingStatus1	0.7870	0.3250	2.42	0.0155	*

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 200.23 on 199 degrees of freedom

Residual deviance: 167.35 on 193 degrees of freedom

AIC: 181.35

Number of Fisher Scoring iterations: 4

--- End of questions for Part A. Part B will be available soon ---

APPENDIX

[Attach all your R codes and outputs here.]

#Question 2:

```
# Print the structure of the data before cleaning and pre-processing here.
my_data <- read.csv("KidneyData.csv")
head(my_data)
str(my_data)

# Print the structure of the training data after cleaning and pre-processing here.
# Example data cleaning steps
my_data$BloodPressure <- ifelse(my_data$BloodPressure > 300, NA, my_data$BloodPressure)
# Handling outliers
my_data$Gender <- as.factor(my_data$Gender)  # Convert categorical to factor
my_data <- na.omit(my_data)  # Remove rows with missing values

# Split the data into training and test sets
set.seed(123)
sample_index <- sample(1:nrow(my_data), size = 0.7 * nrow(my_data))
train_data <- my_data[sample_index, ]
test_data <- my_data[-sample_index, ]
str(train_data)
```

#Question 3:

```
library(ggplot2)
# Histogram for Age
ggplot(my_data, aes(x = Age)) +
```

```
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Frequency")
# Histogram for Blood Pressure
ggplot(my_data, aes(x = BloodPressure)) +
  geom_histogram(binwidth = 10, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Blood Pressure", x = "Blood Pressure", y = "Frequency")


# Box plot for Blood Pressure by Kidney Disease status
ggplot(my_data, aes(x = as.factor(KidneyDisease), y = BloodPressure)) +
  geom_boxplot(fill = c("pink", "lightblue")) +
  labs(title = "Blood Pressure by Kidney Disease Status", x = "Kidney Disease (0 = No, 1 =
Yes)", y = "Blood Pressure")

# Box plot for Blood Sugar by Kidney Disease status
ggplot(my_data, aes(x = as.factor(KidneyDisease), y = BloodSugar)) +
  geom_boxplot(fill = c("pink", "lightblue")) +
  labs(title = "Blood Sugar by Kidney Disease Status", x = "Kidney Disease (0 = No, 1 = Yes)",
y = "Blood Sugar")

# Scatter plot for Blood Pressure vs. BMI
ggplot(my_data, aes(x = BMI, y = BloodPressure, color = as.factor(KidneyDisease))) +
  geom_point(alpha = 0.7) +
  labs(title = "Blood Pressure vs. BMI", x = "BMI", y = "Blood Pressure", color = "Kidney
Disease")

# Scatter plot for Cholesterol vs. Blood Sugar
ggplot(my_data, aes(x = Cholesterol, y = BloodSugar, color = as.factor(KidneyDisease))) +
  geom_point(alpha = 0.7) +
  labs(title = "Cholesterol vs. Blood Sugar", x = "Cholesterol", y = "Blood Sugar", color =
"Kidney Disease")
```

#Question 4 & 5:

#Task 4

```
my_data$KidneyDisease <- as.factor(my_data$KidneyDisease)
```

```
set.seed(123)  # For reproducibility
sample_index <- sample(1:nrow(my_data), size = 0.7 * nrow(my_data))
train_data <- my_data[sample_index, ]
test_data <- my_data[-sample_index, ]
model <- glm(KidneyDisease ~ Age + BloodPressure + BloodSugar + Cholesterol + BMI +
SmokingStatus + ElectricConductivity + pH + DissolvedOxygen + Turbidity +
TotalDissolvedSolids + NitriteLevel + NitrateLevel + LeadConcentration +
ArsenicConcentration + Humidity,
             data = train_data,
             family = "binomial")
summary(model)
model_improved <- step(model)
summary(model_improved)

model_interactions <- glm(KidneyDisease ~ Age + BloodPressure + BMI +
BloodPressure:SmokingStatus,
                          data = train_data,
                          family = "binomial")
summary(model_interactions)
# Fit the improved logistic regression model
model_improved <- glm(KidneyDisease ~ Age + BloodPressure + BloodSugar + Cholesterol +
BMI + SmokingStatus,
                      data = train_data,
                      family = "binomial")

# Display the summary of the improved model
summary(model_improved)
```