ASSIGNMENT 2

QUESTION 1

According to a study by Gaspar et al. (2011), 59.3% of researchers use statistical techniques to detect outliers in administrative medical data as opposed to clustering techniques (14.8%), nearest neighbour techniques (14.8%) and classification techniques (11.1%). In a study on outlier detection in blood pressure data of patients, it was found that the Quartile and Hampel tests gave the best performance (Kuppusamy et al., 2013). Both of these methods, being formal statistical methods, also showed the best performance in a study of outlier detection in medical data, namely patients' body mass index (BMI) (AL.Astal, 2018). It was also found that the informal method of detecting outliers using boxplots, showed the best performance among other informal methods.

Thus, I have decided to use the boxplot method to get an initial overview of outliers in the patient kidney disease dataset. Then, I will employ the Quartile test which is similar to the boxplot method with a usage of statistical values, followed by the Hampel test. The number of outliers identified for each variable by each method are listed down and the most sensitive detection method is chosen to identify the outliers.

The boxplot method is a graphical technique which gives us a visualisation of the median, first quartile (Q_1), third quartile (Q_3), lower and upper inner fences of the dataset. Outliers are represented as small circles in boxplots, where observations are considered outliers if they lie outside of the lower or upper inner fences. Next, we have the Quartile test, in which we need to compute the first quartile, third quartile and interquartile range (IQR). We can then identify outliers as mild outliers if they have a value lower than $Q_1 - 1.5(IQR)$ or higher than $Q_3 + 1.5(IQR)$. Extreme outliers on the other hand, either have values lower than $Q_1 - 3(IQR)$ or higher than $Q_3 + 3(IQR)$. The Hampel's test is a robust outlier detector which involves median and standard deviations values. In this case, outliers are observations that lie multiple standard deviations away from the median.

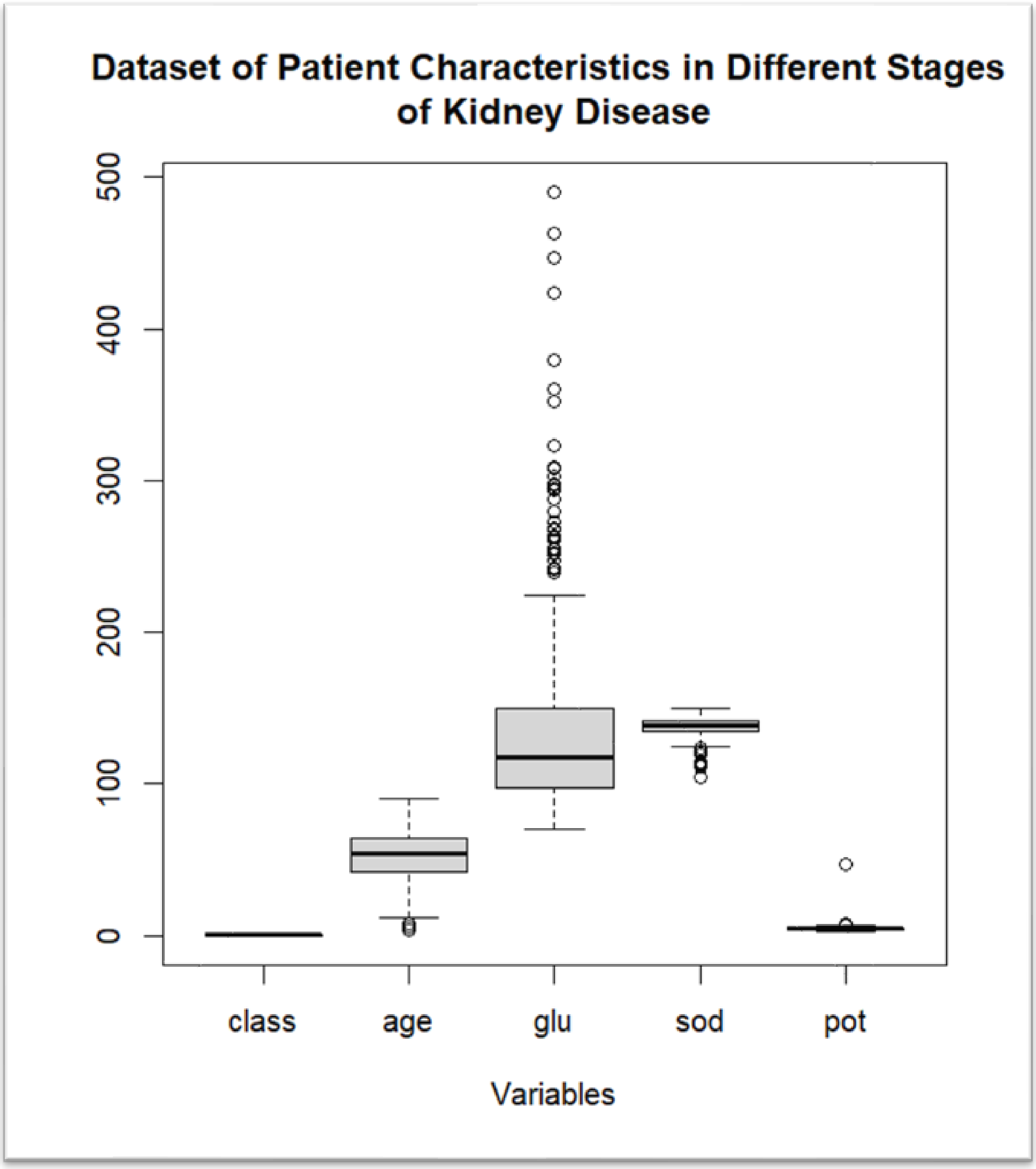
(a)

Outlier Detection Method	Variable				
	class	age	glu	sod	pot
	Number of Outliers Detected				
Boxplot	0	3	35	12	3
Quartile	0	3	35	12	3
Hampel		1	38	8	2

Note that the class variable only contains classes of 0 and 1, indicating not chronic kidney disease and with chronic kidney disease respectively. Since the values are 0 for the boxplot and Quartile methods, it can be said that no observations have incorrect entries. The outliers for class variable cannot be found using the Hampel test because the lower and upper bounds both amount to 1, leading all 0 entries to be outliers, which is incorrect.

Observe that the boxplot and Quartile methods are more sensitive to outliers than the Hampel Test as more outliers are detected in most of the variables. These two methods also have the same exact outliers since they both depend on quartiles and essentially use the same algorithm. Therefore, the **boxplot is used for visualisation** and the **Quartile test is utilised** as the appropriate analysis to detect the presence of outliers in the dataset.

In the following diagram, a visualisation of the outliers can be seen through the boxplots.



Listed below are the outliers for the age, glu, sod and pot variables respectively:

age : 4, 6, 8

glu : 239, 241, 242, 248, 252, 253, 253, 255, 256, 261, 263, 264, 268, 269, 273, 280, 288, 294, 295, 297, 298, 303, 303, 308, 309, 323, 352, 360, 360, 380, 424, 424, 447, 463, 490

sod : 104, 111, 113, 114, 114, 115, 120, 120, 122, 122, 124, 124

pot : 6.6, 7.6, 47.0

Listed below are the outliers that are considered extreme outliers as mentioned earlier:

age	: -
glu	: 323, 352, 360, 360, 380, 424, 424, 447, 463, 490
sod	: 104, 111, 113
pot	: 47

(b)

Age Variable

The 3 outliers for the age variable are retained. If we look at the dataset, all three of these patients aged 4, 6 and 8 have chronic kidney disease (CKD). This is not unusual as CKD can also prevail in children as shown by a study that enrolled 586 children aged between 1 and 16 who have CKD (Wong et al., 2012). CKD may not be as common in children as they are in adults, leading them to be identified as outliers. However, it must be retained as they are valid data showing three child patients having CKD.

Glucose Variable

Retain all 35 outliers. The dataset shows that all 35 patients with outlier values have CKD. Extremely high glucose levels could indicate that a person has diabetes. Those with glucose levels over 240 mg/dl, need to test for ketones which are access levels of acid produced from the break down of fat for fuel by the liver (*Diabetic Ketoacidosis | Diabetes | CDC*, n.d.). High levels of ketones, constantly over 300 mg/dl could indicate diabetic ketoacidosis (DKA). According to a health care provider, DKA is a very serious problem and can have glucose levels over 500 mg/dl (*If Your Blood Glucose Is Too High or Too Low - Lahey Health*, n.d.). It is a well-known fact that diabetes can cause kidney disease, so it does not come as a shock that some CKD patients have highly elevated glucose levels. Also noteworthy, the highest glucose level ever recorded by a survivor is 2656 mg/dl, which is a very abnormal finding (*Highest Blood Sugar Level | Guinness World Records*, n.d.). Patients with the outlier values might have serious health conditions but the values are possible to happen and will not be discarded.

Sodium Variable

Retain all 12 outliers. All 12 patients with the outlier values have CKD. From a medical perspective, sodium levels in blood between 135 and 145 mEq/L are considered normal (*Hyponatremia - Symptoms and Causes - Mayo Clinic*, n.d.). Hyponatremia, a condition that occurs when sodium levels in blood fall below 135 mEq/L can be unsafe. Hyponatremia can be caused by kidney disease where fluids accumulate and dilute the sodium in one's body, lowering the sodium levels. As can be seen, the outlier values range from 104 to 124 mEq/L, which is a severe condition. These patients with outlier values are also aged from 48 to 83, coinciding with the fact that older adults are more likely to face hyponatremia. However, one of the lowest values of sodium level ever recorded was by a 54-year-old woman at 99 mEq/L (Gupta et al., 2015). Even the lowest extreme outlier of 104 mEq/L is above 99 mEq/L. Thus, the outlier values represent serious health conditions but will not be discarded as the values are realistic to happen.

Potassium Variable

Retain the outliers 6.6 and 7.6 but discard 47.0. The three patients with outlier values have CKD. It is a fact that, normal potassium levels in blood are 3.5 to 5.0 mEq/L, high levels are from 5.1 to 6.0 mEq/L, whereas dangerously high levels are above 6.0 mEq/L (*7 Things to Know About Potassium in Your Diet | National Kidney Foundation*, n.d.). Potassium levels below 2.5 mEq/L can be life threatening but it is observed that the lowest potassium level for patients in this dataset is 2.5 mEq/L. CKD can cause high potassium levels in patients as damaged kidneys are not able to remove excess potassium, causing it to accumulate in the blood. It was found in a study of extreme hyperkalemia (fatal potassium levels) that the highest recorded potassium level of a patient who survived is 14.0 mEq/L (Tran, 2005). Even a potassium level above 10.0 mEq/L is fatal unless treated immediately. This means that only the extreme outlier, 47 mEq/L, must be removed from the dataset as it is impossible for a person to have this potassium level and still survive. It could have been a recording error where the value intended to be entered was 4.7 but 47 was accidentally recorded.

QUESTION 2

(a)

The original variable set is $\{X_1, X_2, X_3, X_4\}$. However, for a forward selection, the procedure starts with an empty subset of variables as the reduced set. The best of the original variables is determined and added to the reduced set one by one based on their AIC values. Since a lower AIC value indicates a better model, we need to keep adding variables to the model as long as the new addition reduces the current AIC value. Stop the procedure when the AIC value can no longer decrease, that is, there is no more significant improvement in the model's performance. This means that for each iteration, the excluded variable that lowers the AIC value the most is added to the model, provided that this AIC value is lower than the current AIC value of the model. Variables that have been added to the model, remain in the model.

Step 1: $\{\quad\}$

Start with an empty subset of variables.

Step 2: $\{X_3\}$ with an AIC value of 73.2174

Add the variable X_3 to the model as it contributes the lowest AIC value, indicating the best model by far compared to adding other variables. This variable cannot be eliminated from the model from here forth. Next, we will only consider models that contain the variable X_3.

Step 3: $\{X_2, X_3\}$ with an AIC value of 63.1980

The choices were either $\{X_1, X_3\}$, $\{X_2, X_3\}$ or $\{X_3, X_4\}$, since only these three models contained X_3. We choose $\{X_2, X_3\}$ because this model provides the lowest AIC value, which is lower than the value in Step 2. The models considered next must contain the variables X_2 and X_3.

Step 4: No more variables are added to the model. We can only consider $\{X_1, X_2, X_3\}$ and $\{X_2, X_3, X_4\}$ as these two models have the variables X_2 and X_3. However, both these models give a higher AIC value than the current model in Step 3. Even though, the model $\{X_1, X_2, X_3, X_4\}$ gives an AIC value of 63.0223, which is lower than the AIC value of the model in Step 3, we cannot choose this model as it involves the addition of two variables, X_1 and X_2, in one go instead of adding the variables one by one. The procedure stops with the model in Step 3.

The best subset of variables using forward selection is $\{X_2, X_3\}$.

(b)

The original variable set is $\{X_1, X_2, X_3, X_4\}$. In backward elimination, this model with the full set of variables is used at the start of the procedure. At each iteration, the worst variable in the current set is deleted one by one until deletions of variables no longer reduces the current model's AIC value. The worst variable in each iteration is the variable that when removed, provides a model with the lowest AIC value. Similar to forward selection, the procedure is stopped when the AIC value can no longer decrease and there is no more significant improvement in the model's performance. Variables that have been dropped from the model, cannot be added back to it.

Step 1: $\{X_1, X_2, X_3, X_4\}$ with an AIC value of 63.0223.

Start with the full set of variables.

Step 2: $\{X_1, X_3, X_4\}$ with an AIC value of 61.3073.

The variable X_2 is removed from the model as the removal of this variable results in the lowest AIC score compared to the removal of X_1 or X_3 or X_4. The variable X_2 cannot be added back to the model, indicating that the next model considered must not contain X_2.

Step 3: No more variables are removed from the model. We can only consider $\{X_1, X_3\}$, $\{X_1, X_4\}$ and $\{X_3, X_4\}$ as these three models do not contain the variable X_2. However, all three of these models give a higher AIC value than the current model in Step 2, providing no improvement to the current model. It is also noticed that none of the models provide a better performance than the model in Step 2, as it is the model with the lowest AIC value.

The best subset of variables using backward elimination is $\{X_1, X_3, X_4\}$.

(c)

The best subset of variables using forward selection and backward elimination provide different answers. The best subset of variables using forward selection is $\{X_2, X_3\}$ (with an AIC value of 63.1980), whereas using backward elimination is $\{X_1, X_3, X_4\}$ (with an AIC value of 61.3073). The answers using both methods may differ because it is possible to miss the optimal model, as variables are added or dropped one at a time. We know that forward selection starts with smaller models or an empty set, whereas backward elimination starts with a full model and progressively eliminates variables. The difference in their algorithms cause their ‘optimal models’ to differ.

Here, forward selection is not required to consider the full model, which could be advantageous in certain conditions such as when the number of variables considered are very large. In the case of this question, it misses a better model with lower AIC value. Besides that, when a new variable is added to the model, it could cause other existing variables to become non-significant (Chowdhury & Turin, 2020). However, variables that were added cannot be removed anymore.

On the other hand, in backward elimination where the procedure starts with a full model, the effects of all variables are considered simultaneously. Noteworthily, when there is correlation between variables in a model, this method assesses them as a whole, dropping variables only when it is beneficial to do so (Mantel, 2016). The least significant variables can be dropped early in the procedure so that only significant ones are left. On the contrary, in forward selection, this model might not be reached. However, since dropped variables cannot be added back to the model in backward selection, dropped variables that become significant later on cannot be considered anymore.

Both methods have their own advantages and disadvantages. In this case, **the best answer is the model obtained using backward elimination, which is $\{\boldsymbol{X_1}, \boldsymbol{X_3}, \boldsymbol{X_4}\}$**. This is because it provides a lower AIC value compared to the model obtained using forward selection.

References

7 Things to Know About Potassium in Your Diet | National Kidney Foundation. (n.d.). Retrieved December 9, 2022, from https://www.kidney.org/atoz/content/potassium#what-safe-level-potassium-my-blood

AL.Astal, J. A. (2018). *Comparison of methods for detecting outliers in medical data*. [Master's thesis, Al-Azhar University]. http://www.alazhar.edu.ps/arabic/He/files/20154009.pdf

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1). https://doi.org/10.1136/fmch-2019-000262

Diabetic Ketoacidosis | Diabetes | CDC. (n.d.). Retrieved December 9, 2022, from https://www.cdc.gov/diabetes/basics/diabetic-ketoacidosis.html

Gaspar, J., Catumbela, E., Marques, B., & Freitas, A. (2011). *A Systematic Review of Outliers Detection Techniques in Medical Data-Preliminary Study. Clinical data modeling for EHR exchange View project*. https://www.researchgate.net/publication/221334605

Gupta, E., Kunjal, R., & Cury, J. D. (2015). Severe Hyponatremia Due to Valproic Acid Toxicity. *Journal of Clinical Medicine Research*, *7*(9), 717–719. https://doi.org/10.14740/JOCMR2219W

Highest blood sugar level | Guinness World Records. (n.d.). Retrieved December 9, 2022, from https://www.guinnessworldrecords.com/world-records/highest-blood-sugar-level

Hyponatremia - Symptoms and causes - Mayo Clinic. (n.d.). Retrieved December 9, 2022, from https://www.mayoclinic.org/diseases-conditions/hyponatremia/symptoms-causes/syc-20373711

If Your Blood Glucose is Too High or Too Low - Lahey Health. (n.d.). Retrieved December 9, 2022, from https://www.lahey.org/article/if-your-blood-glucose-is-too-high-or-too-low/

Kuppusamy, M., Kannan Kaliyaperumal, S., & Kannan, S. K. (2013). *Comparison of Methods for detecting Outliers Comparative analysis of community discovery methods in social networks View project Comparison of methods for detecting outliers*. http://www.ijser.org

Mantel, N. (2016). *Why Stepdown Procedures in Variable Selection* (Vol. 12, Issue 3). http://about.jstor.org/terms

Tran, H. A. (2005). Extreme hyperkalemia. *Southern Medical Journal*, *98*(7), 729–732.

Wong, C. J., Moxey-Mims, M., Jerry-Fluker, J., Warady, B. A., & Furth, S. L. (2012). CKiD (CKD in Children) prospective cohort study: A review of current findings. In *American Journal of Kidney Diseases* (Vol. 60, Issue 6, pp. 1002–1011). W.B. Saunders. https://doi.org/10.1053/j.ajkd.2012.07.018