Problem :

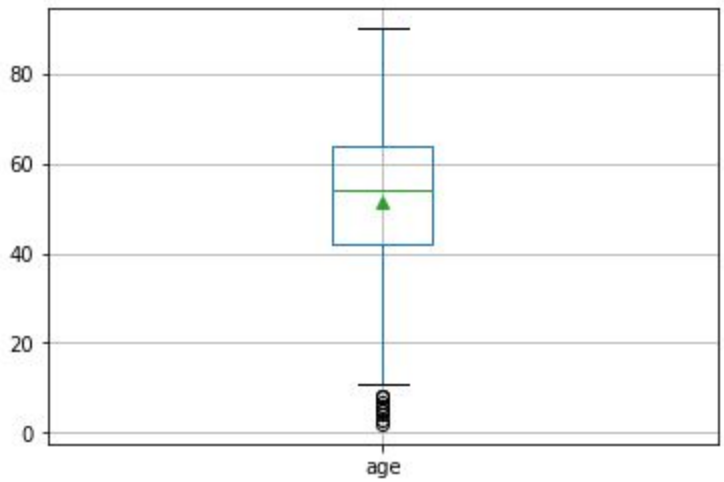
Chronic Kidney Disease

Abstract: This dataset can be used to predict chronic kidney disease and it has been collected at a hospital for a period of nearly 2 months.

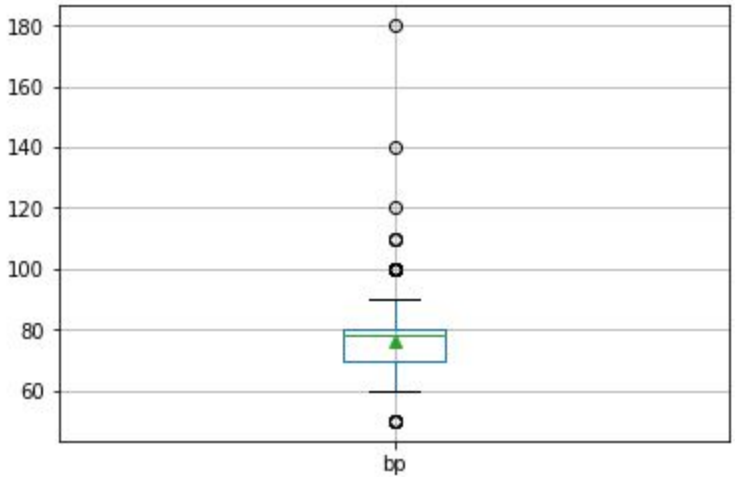
Understanding of data:
In the given data set we have been given two different data type numerical and nominal. We do see that some of the data have been wrongly typed, presence of missing values and coerced data

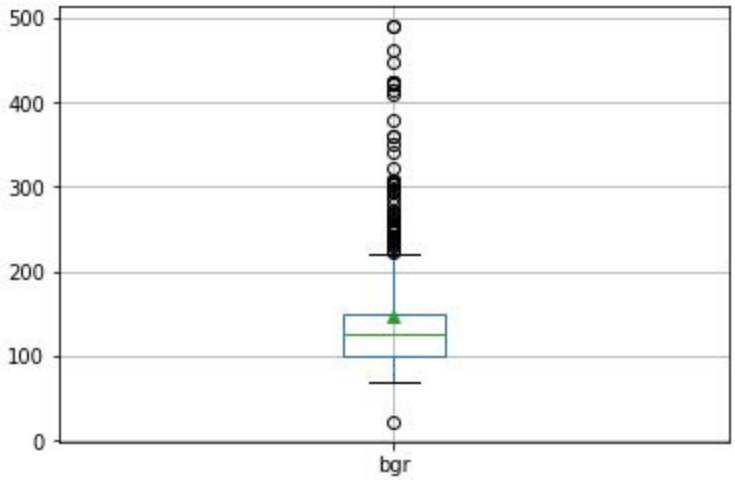
Pre-processing techniques:
In order to fix the wrongly type words, presence of missing values. For identifying the missing value we have used the panda's read_csv method passing in the missing value identifier i.e., "?" and for the rest of the pre-processing technique we have used different approach for the numerical and nominal data type.

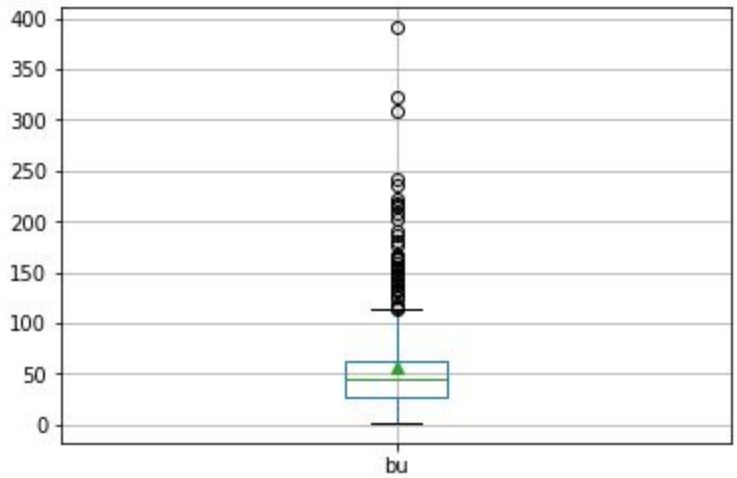
Numerical data type:
Inorder to better understand the numerical data we have used boxplot technique to get a graphical representation of our data. Below are plot of the numeric data



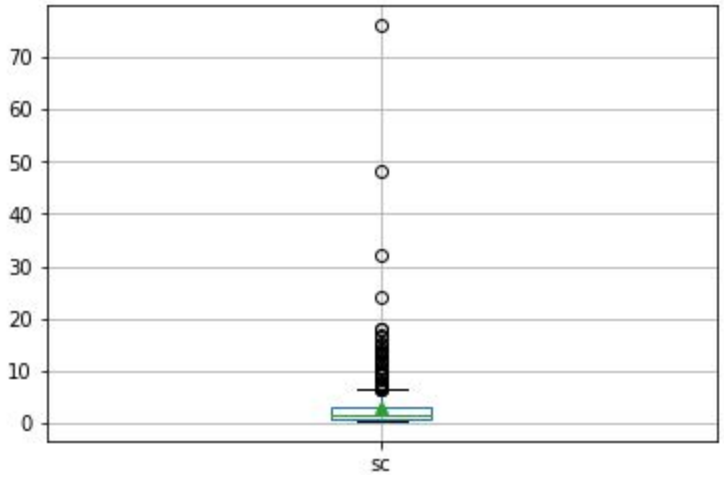
Chronic Kidney Disease

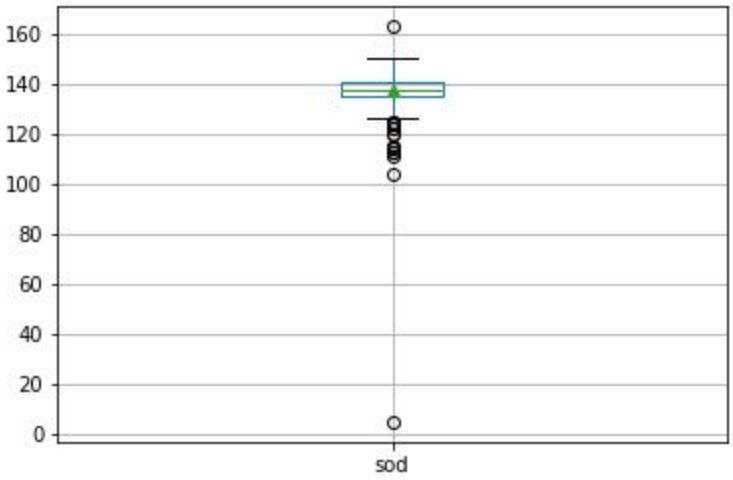


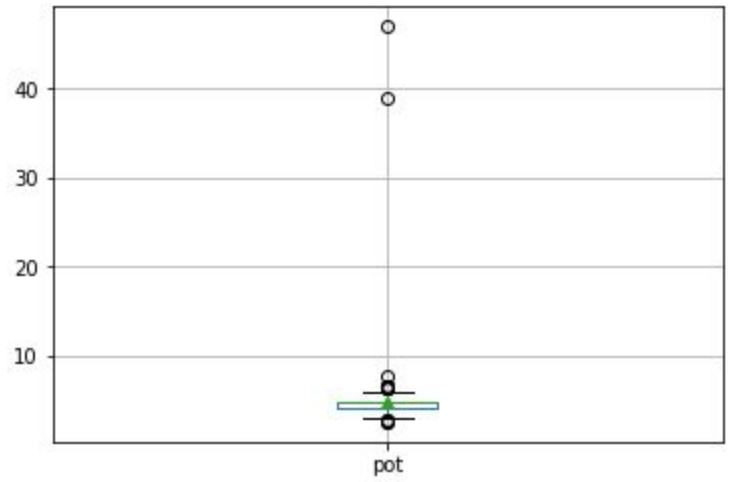




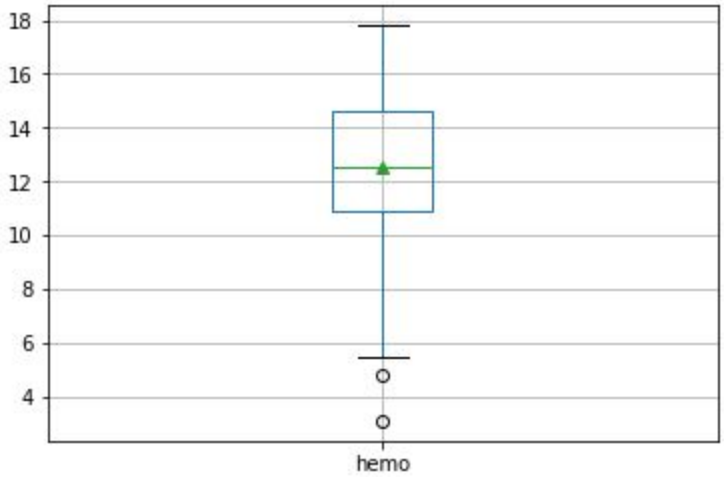
Chronic Kidney Disease

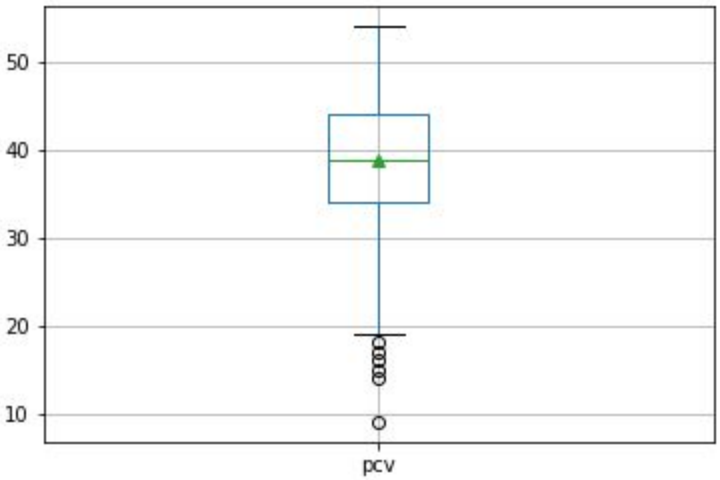


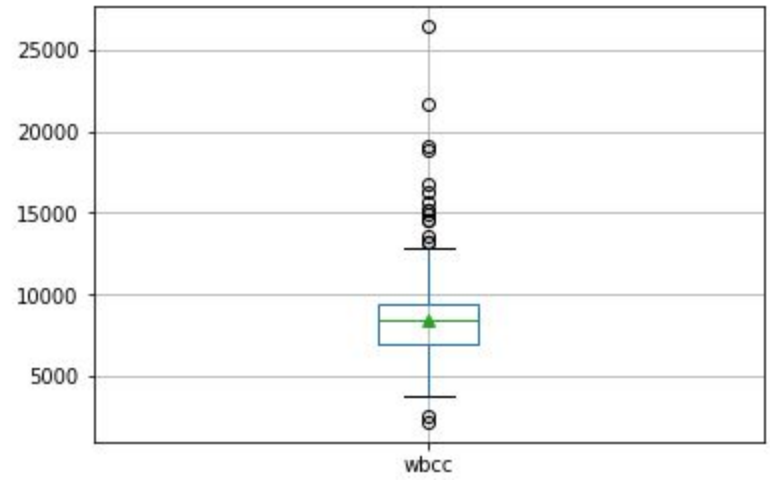


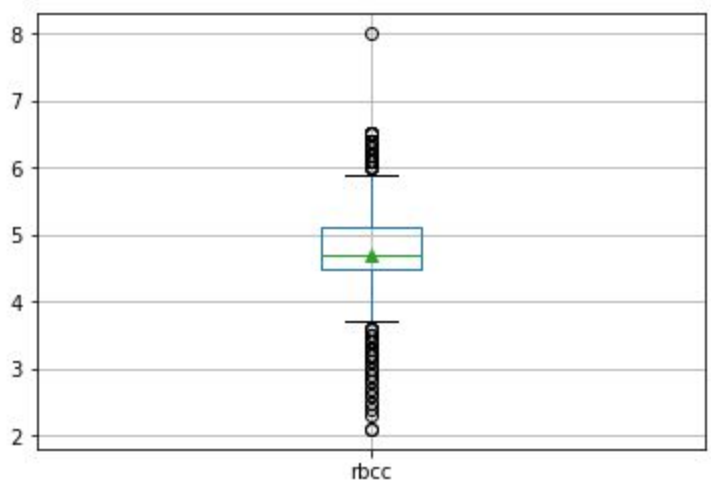


Chronic Kidney Disease









From the box plot we can see that most of the data's mean and median are almost similar except for sc which contains lot of outliners. But by going through some research https://www.researchgate.net/publication/310952883_Missing_Data_Analysis_Using_Multiple_Imputation_in_Relation_to_Parkinson's_Disease, https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779 . We can see that for the given dataset filling the missing value with the help of mean gives fairly accurate result and it is most time efficient and does fairly decent job for small dataset

Nominal Data set:

Once we have sorted out numerical data we need to fill missing entry in nominal data type with most occurring value. Before filling the missing value check the values. We see that htn and dm has bad values such has tab and spaces, we rectified it manually. Inorder to fill missing value for nominal data type we use mode of column. We chose this technique because of the simplicity and gives a good result.

We then proceed with Separating feature variable and target variable and Handling categorical columns in feature variable. Here we convert the nominal data type into categorical columns so that it will be easier to make classifiers. So cad is converted into cad_no cad_yes with their respective values

Target Data set:
The target data set was completely fine and didn't require any modification.

Intuition behind Algorithm selection for building model:
Since the problem here is to classify for a given patient record whether he is susceptible to chronic kidney disease or not. So, this is a typical classification technique and as discussed in class we chose two of classification algorithm which is **Decision Tree Classifier** and **Random Forest Classifier**. We separated our data set into two categories: one for training and other for testing with 80 % and 20% of data from our data set.

Discussion of results:

Conclusion, In this chronic kidney classification problem we see that, there are many problems in the real world data set. Some of the problem can be missing values, garbled character and in order to fill missing values we have used the mean and mode technique. Once our data set is clean and all the missing values is filled we used two classification techniques to compare its accuracy, we used Random Forest Classifier and Decision Tree Classifier having accuracy of 0.9625 and 0.896875 and we have chosen Random Forest Classifier for our classification technique.