**Report on PURCHASE2 Data set**

**Reg no – 1822576003**

**Name – NP GANESH**

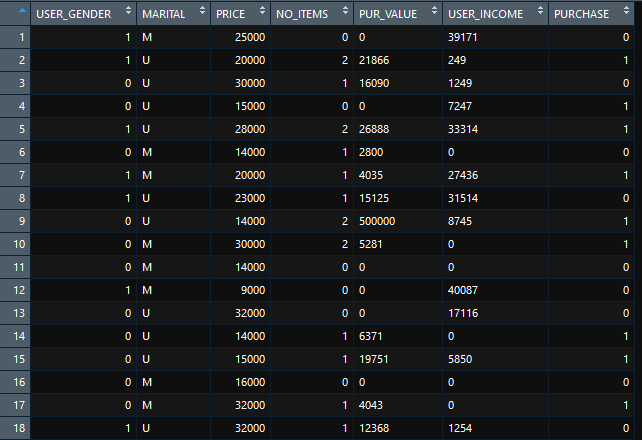
The data set is about customer and their purchases. Understanding the data set, it tells customer purchases with respect to their marital status and no. of items customers purchased and also customers not purchased any item.

**Importing Data set :**

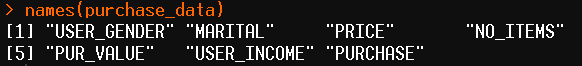
To import Data set do read the file into R studio



After reading the data set it looks like

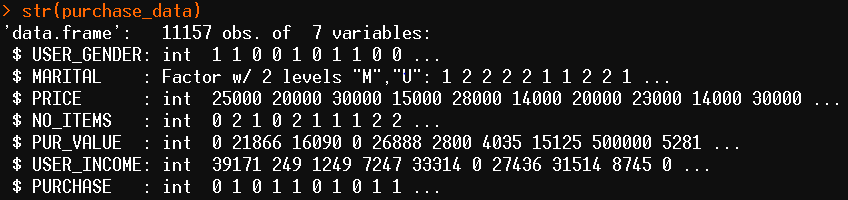


**Number of Variables in the PURCHASE2 data set :**



Based on the above image (using R), there are total of 7 variables in the available data set.

**Structure of Data set :**



Above code gives you the structure of data set. It tells us which type of data the variable consists of and showing us PURCHASE2 is having 11157 entries.

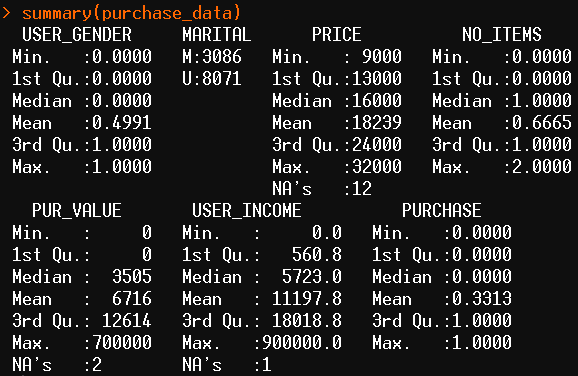
**Finding Missing Values :**

To find missing values we do ..



So there are total of 15 missing values in the PURCHASE2 data set.

**Summary of Data set before imputing the values**



Based on the summarised information of data set I can figure out the number of NA values in each column

|  |  |
| --- | --- |
| **Variable** | **Number of NA s** |
| USER\_GENDER | 0 |
| MARITAL | 0 |
| PRICE | 12 |
| NO\_ITEMS | 0 |
| PUR\_VALUE | 2 |
| USER\_INCOME | 1 |
| PURCHASE | 0 |

**Imputing Missing Values :**

By Knowing the missing values present in each variable , I am replacing those missing values with the value of previous index (ex : [ i ] = [ I – 1 ])

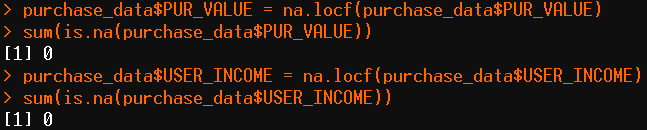
Here, imputing the missing values present in PRICE variable with the previous data point.



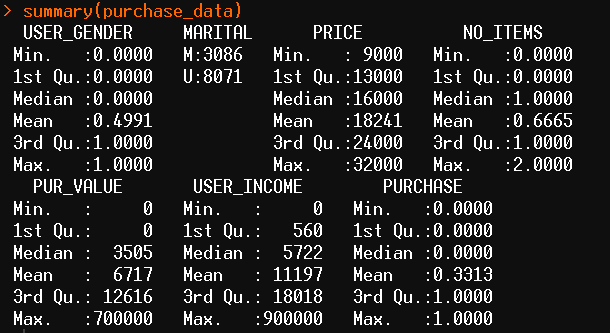
Before imputing the missing values the count is ‘12’ and Now, after imputation the count of missing values in PRICE variable should be ‘0’



In the same manner I have imputed the missing columns present in each variable as follows..



**Summary of Data set after imputing the missing values :**

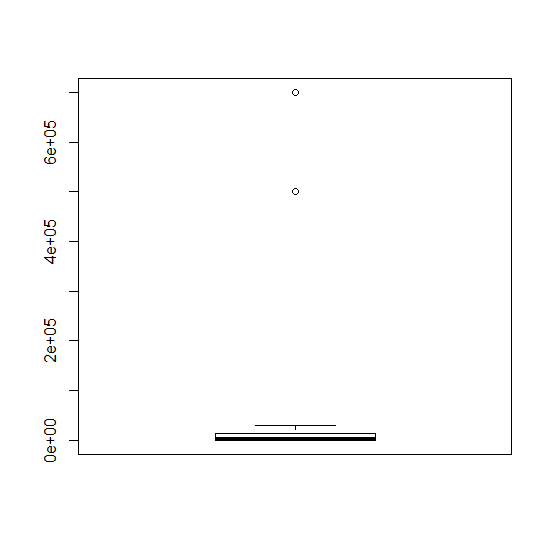
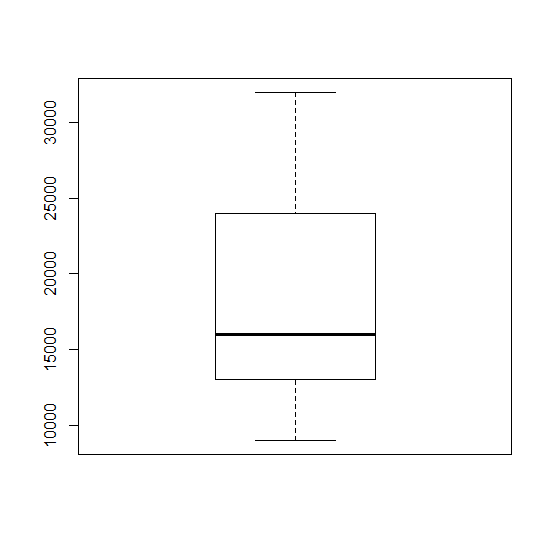


From this summary I ensure all the missing values are imputed.

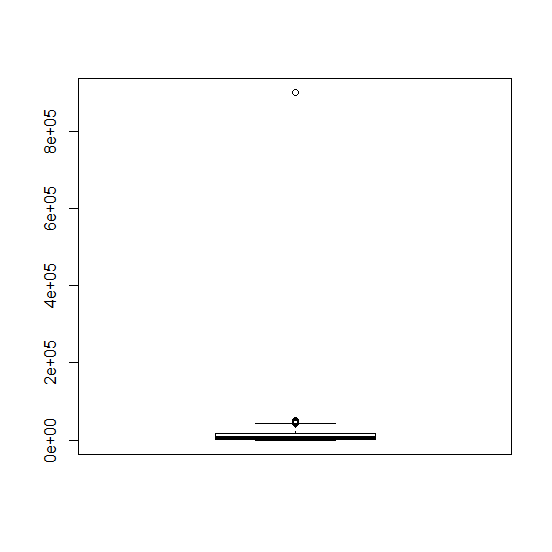
**Outliers and Box plot :**

**Before Removing Outliers**

**Box plot of Variable Purchase Box plot of Variable PUR\_VALUE**

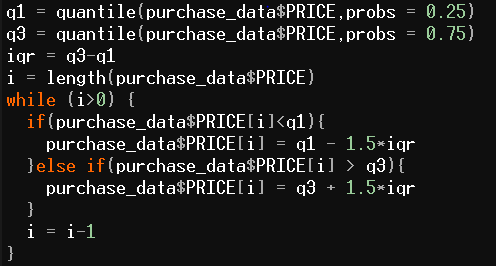


**Box plot of Variable USER\_INCOME**

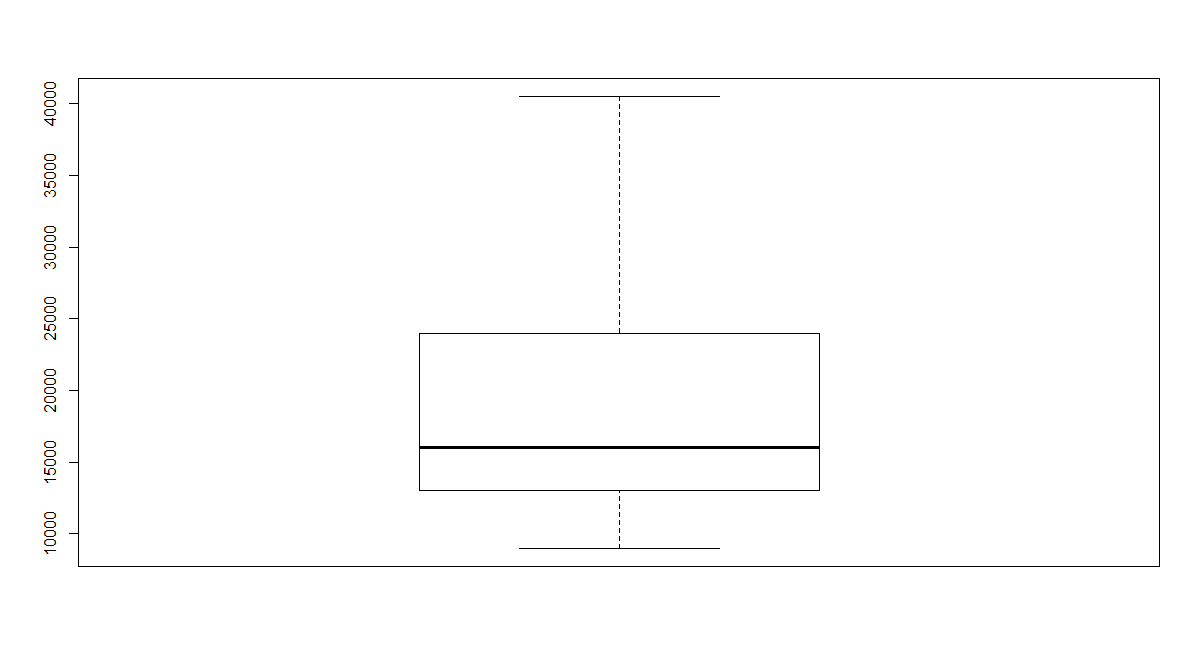


**Removing outliers**

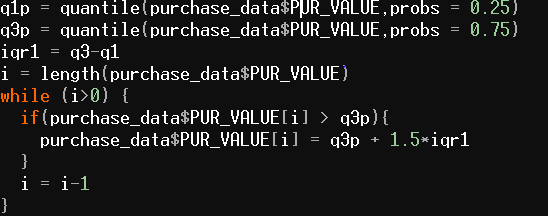
**For PRICE variable**



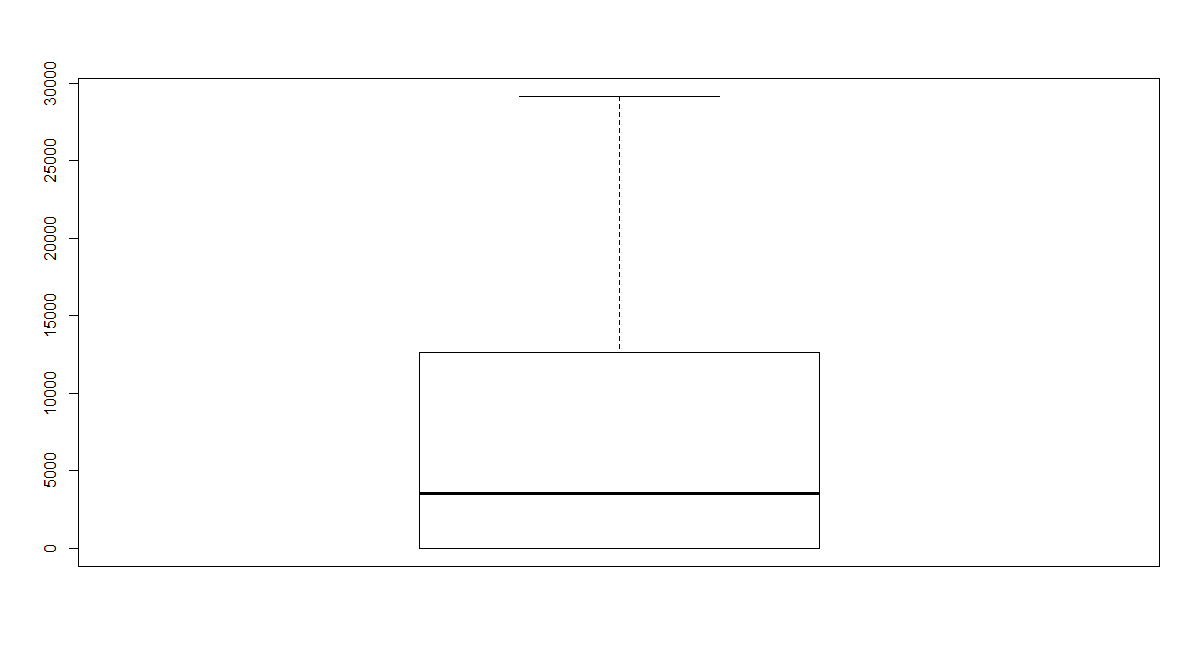
**Box plot after removing outlier**



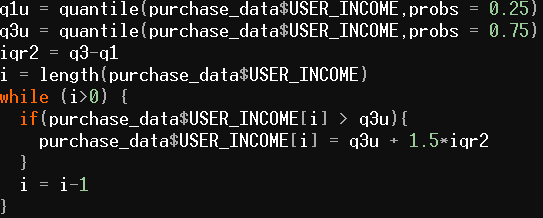
**For PUR\_INCOME variable**



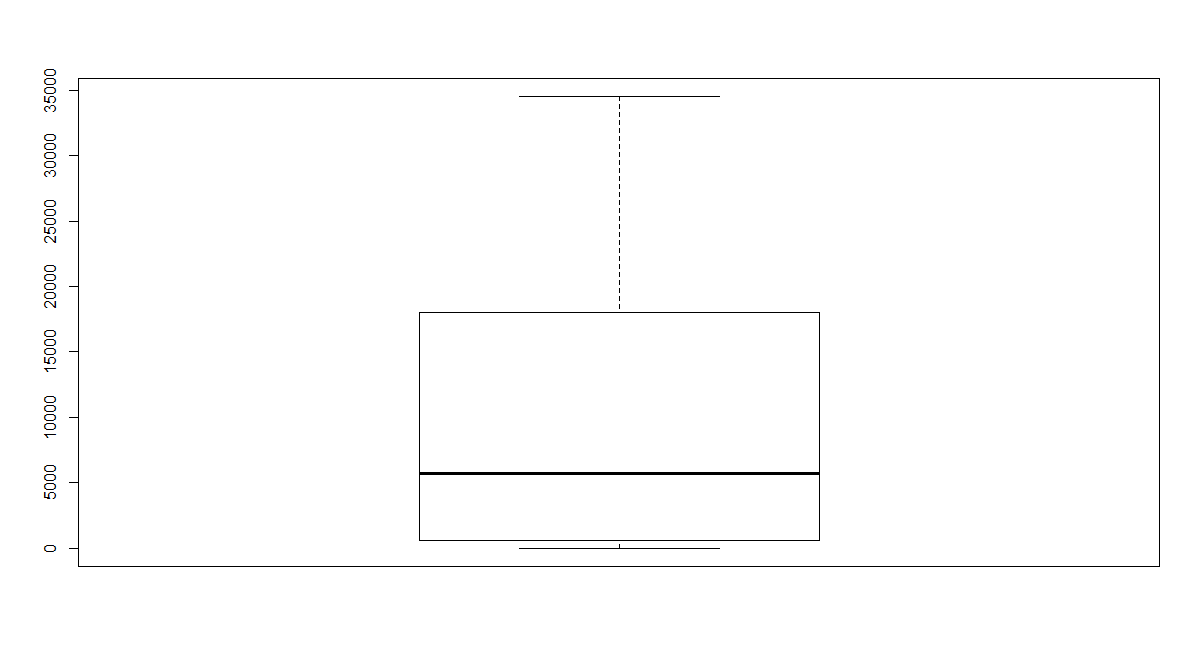
**Box plot after removing outlier**



**For USER\_INCOME variable**

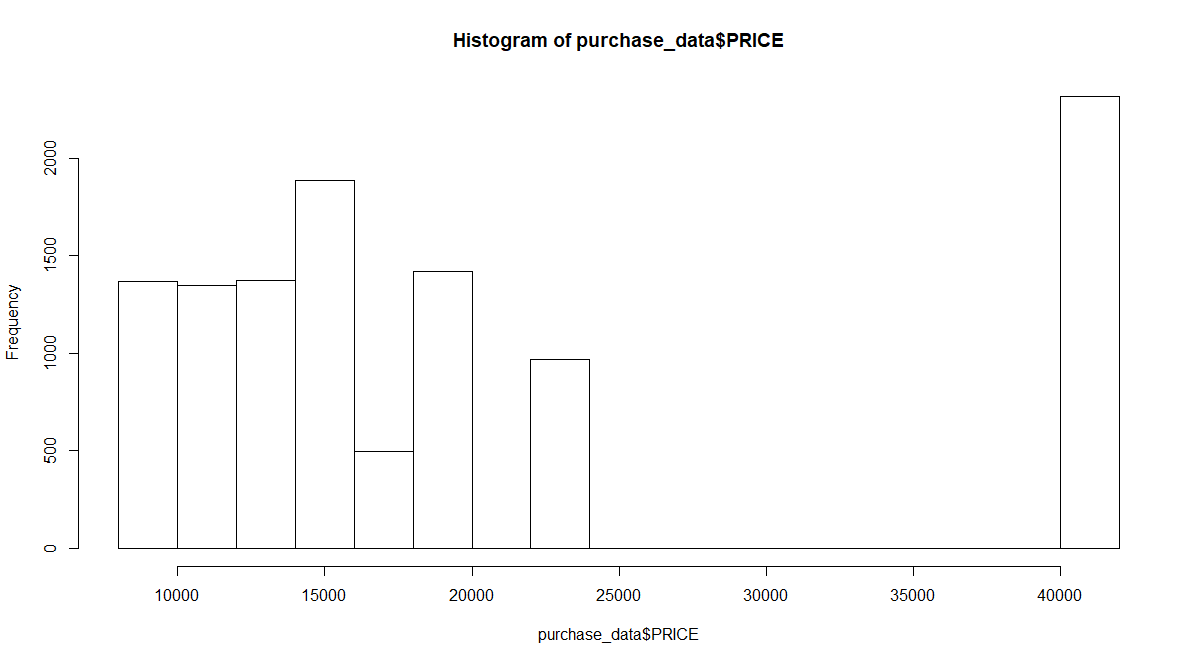


**Box plot after removing outlier**

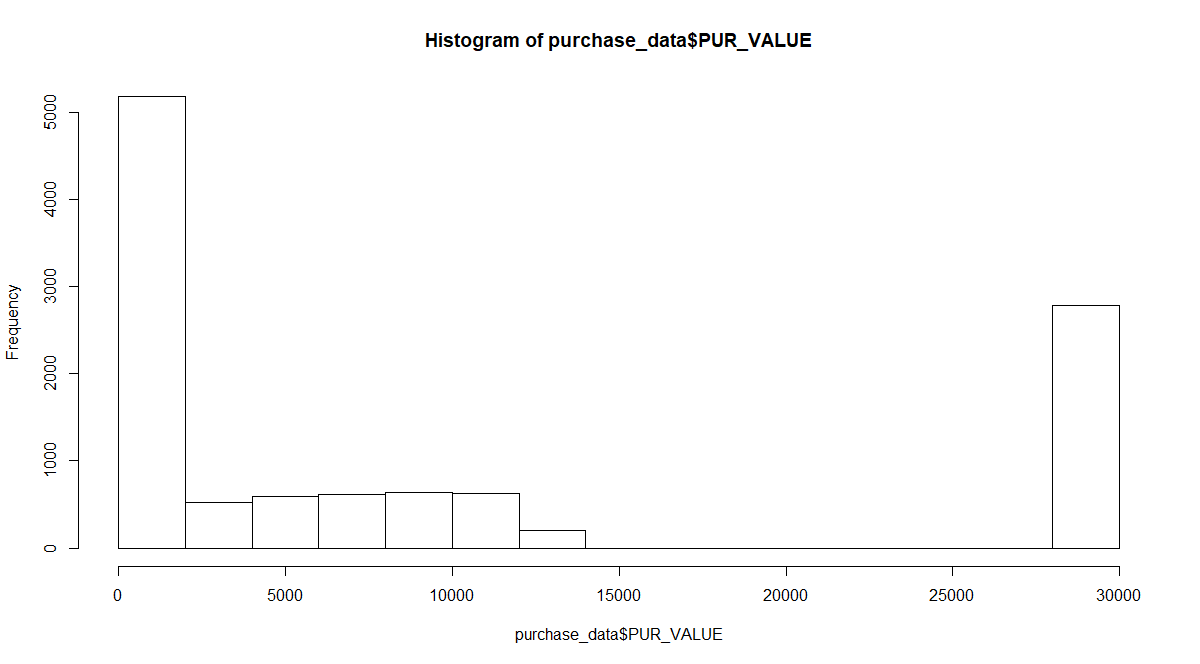


**Distribution Graphs of PRICE, PUR\_VALUE, USER\_INCOME:**

**PRICE :**



**PUR\_VALUE :**



**USER\_INCOME:** 