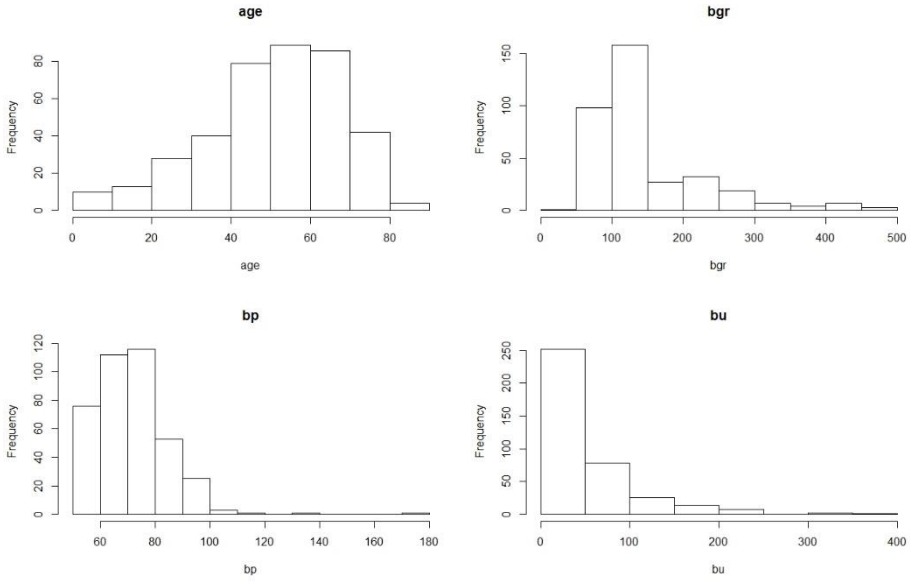
1. Data Preprocessing and Generalized Linear Model (GLM)/Logistic Regression

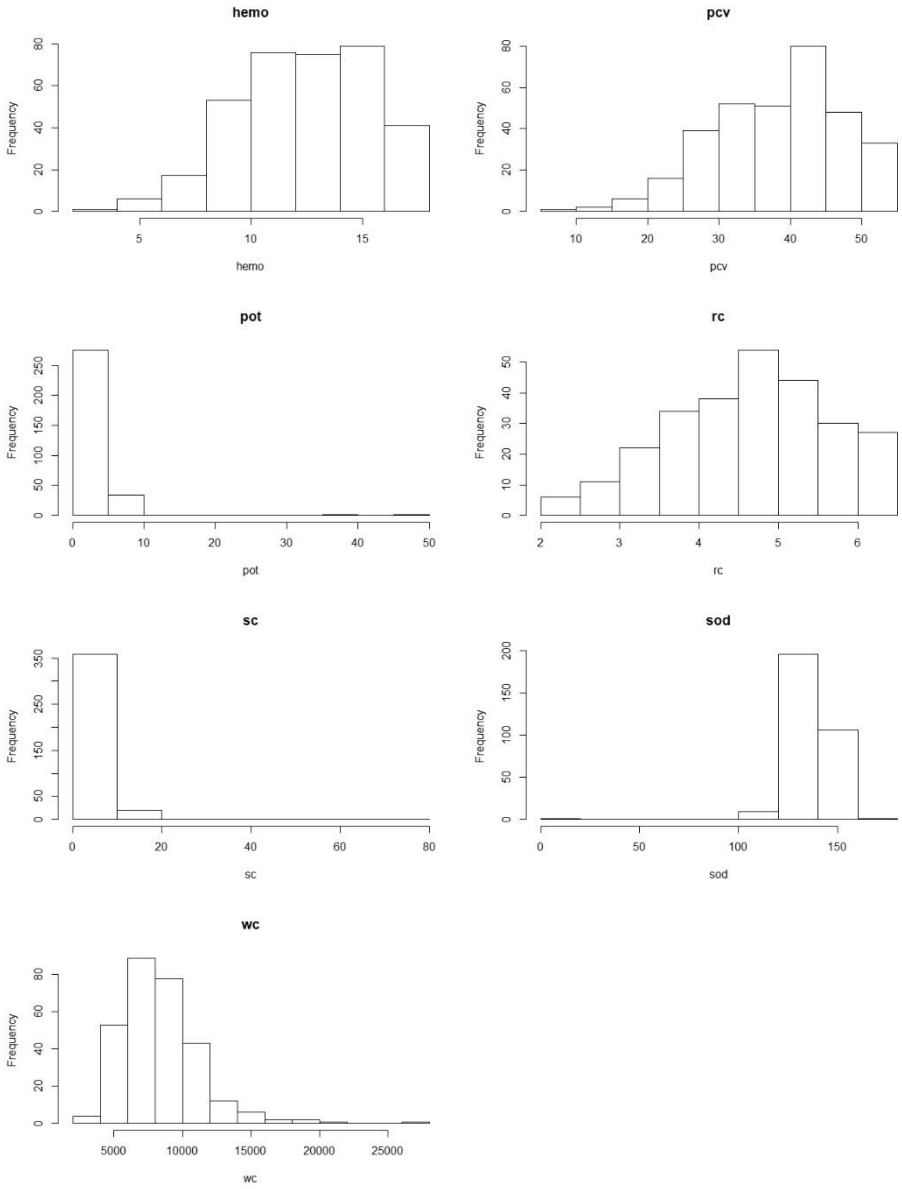
(1) Provide the descriptive statistics. Eg. mean, variance, data distribution, # of missing value, # of outlier, etc.

先利用資料給的說明敘述，區分出類別變數及非類別變數，之後透過 Summary 函式得到基本的敘述統計，輸出結果如圖。

```
> summary(Kidney_data)
      age              bp              sg               al              su              rbc              pc
 Min.   : 2.00   Min.   : 50.00   1.005:  7   0   :199   0   :290   abnormal: 47   abnormal: 76
 1st Qu.:42.00   1st Qu.: 70.00   1.01 : 84   1   : 44   1   : 13   normal  :201   normal  :259
 Median :55.00   Median : 80.00   1.015: 75   2   : 43   2   : 18   NA's    :152   NA's    : 65
 Mean   :51.48   Mean   : 76.47   1.02 :106   3   : 43   3   : 14
 3rd Qu.:64.50   3rd Qu.: 80.00   1.025: 81   4   : 24   4   : 13
 Max.   :90.00   Max.   :180.00   NA's : 47   5   :  1   5   :  3
 NA's   :9       NA's   :12                   NA's: 46   NA's: 49
       pcc              ba             bgr              bu              sc              sod
 notpresent:354   notpresent:374   Min.   : 22   Min.   :  1.50   Min.   : 0.400   Min.   :  4.5
 present   : 42   present   : 22   1st Qu.: 99   1st Qu.: 27.00   1st Qu.: 0.900   1st Qu.:135.0
 NA's      :  4   NA's      :  4   Median :121   Median : 42.00   Median : 1.300   Median :138.0
                                   Mean   :148   Mean   : 57.43   Mean   : 3.072   Mean   :137.5
                                   3rd Qu.:163   3rd Qu.: 66.00   3rd Qu.: 2.800   3rd Qu.:142.0
                                   Max.   :490   Max.   :391.00   Max.   :76.000   Max.   :163.0
                                   NA's   :44    NA's   :19       NA's   :17       NA's   :87
      pot             hemo            pcv              wc              rc           htn
 Min.   : 2.500   Min.   : 3.10   Min.   : 9.00   Min.   : 2200   Min.   :2.100   no  :247
 1st Qu.: 3.800   1st Qu.:10.30   1st Qu.:32.00   1st Qu.: 6500   1st Qu.:3.900   yes :145
 Median : 4.400   Median :12.65   Median :40.00   Median : 8000   Median :4.800   NA's:  8
 Mean   : 4.627   Mean   :12.53   Mean   :38.87   Mean   : 8426   Mean   :4.696
 3rd Qu.: 4.900   3rd Qu.:15.00   3rd Qu.:45.00   3rd Qu.: 9800   3rd Qu.:5.400
 Max.   :47.000   Max.   :17.80   Max.   :54.00   Max.   :26400   Max.   :6.500
 NA's   :88       NA's   :52      NA's   :72      NA's   :109     NA's   :134
    dm         cad         appet          pe          ane          class
  yes:  1   no  :351   good:304   no  :312   no  :326   ckd   :237
 no  :256   yes : 34   poor: 82   yes : 74   yes : 60   notckd:150
 yes :130   NA's: 15   NA's: 14   NA's: 14   NA's: 14   NA's  : 13
 NA's: 13
```

另外，針對數值型的資料畫出 data distribution，以下為 age、bgr、bp、bu、hemo、pcv、pot、rc、sc、sod、wc 的頻率分布圖。





再來，將數值型資料的 variance 算出，以下為輸出結果。

```
> var(Kidney_data$age,na.rm=TRUE)
[1] 294.7991
> var(Kidney_data$bp,na.rm=TRUE)
[1] 187.2419
> var(Kidney_data$bgr,na.rm=TRUE)
[1] 6285.59
> var(Kidney_data$bu,na.rm=TRUE)
[1] 2550.554
> var(Kidney_data$sc,na.rm=TRUE)
[1] 32.96053
> var(Kidney_data$sod,na.rm=TRUE)
[1] 108.3421
```

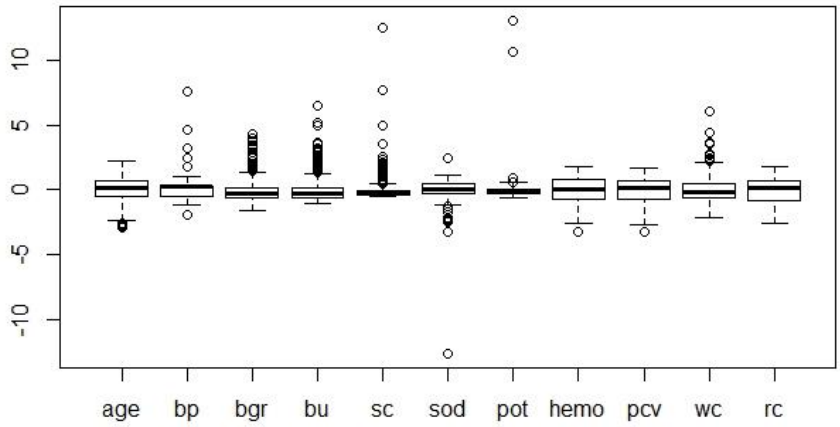
```
> var(Kidney_data$pot,na.rm=TRUE)
[1] 10.20102
> var(Kidney_data$hemo,na.rm=TRUE)
[1] 8.483161
> var(Kidney_data$pcv,na.rm=TRUE)
[1] 81.01719
> var(Kidney_data$wc,na.rm=TRUE)
[1] 8700058
> var(Kidney_data$rc,na.rm=TRUE)
[1] 1.019341
```

(2) Remove the observation without “class” value

透過計算可以得知總共有 13 筆資料沒有 class 的值，因此利用 complete.cases()函式將資料移除。

(3) How to identify the outlier? How to impute the missing value?

可以透過標準化分數或是盒鬚圖來判斷離群值。標準化分數部分，透過將資料轉為標準化分數，當資料落於 Z 分數正負 3（亦可自訂其他分數數值）之外，就判定其為離群值。盒鬚圖部分，則是畫出盒鬚圖取出 Q1、Q3，當資料落於 3 倍 IQR 的距離之外，即可判定為離群值。以下為標準化過後的盒鬚圖。



填補遺失值部分可以透過平均值、第一四分位數、KNN 或 MICE 方法來填補，此次我是使用 KNN 來填補，具體作法如附檔程式碼。

(4) How to transform the categorical variable to dummy variable?

在 R 裡面可以透過 dummies 套件轉換，具體作法如附檔程式碼。

(5) How to “randomly” split the dataset into training dataset and testing dataset (eg. 80% vs. 20%)?

隨機切割資料可以透過 sample 函式指定資料範圍與大小之後產出 index，再利用 index 去分割 training 與 test 資料集。

(6) Please use the Generalized Linear Model (GLM)/Logistic Regression to predict the “Class” in the testing dataset.

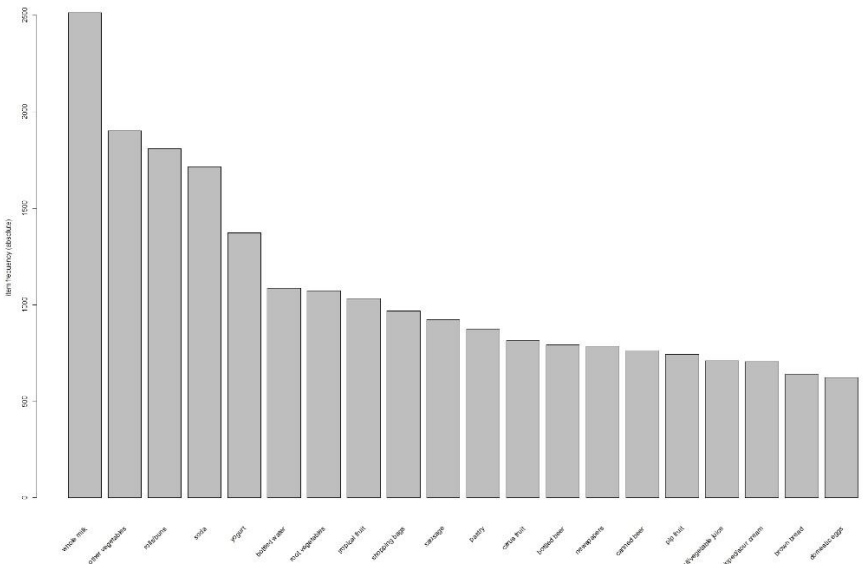
透過 glm(formula, family, data)函式來實作 Logistic 迴歸並預測，具體作法如附檔程式碼。最後預測的準確結果與否如圖。

```
pred  0  1
   0 27  3
   1  0 47
```

2. Association Rule Market Basket Analysis

(1) How to handle the raw dataset via data preprocessing?

透過 read.transaction 將資料讀進 R，並且看看最常被買的產品有哪些，並且針對各產品的頻率畫出長條圖，可以看出最常被購買的產品前五名分別是 whole milk、other vegetables、rolls/buns、soda、yogurt。



(2) What’s the top 5 association rules? Show the support, confidence, and lift to each specific rule, respectively?

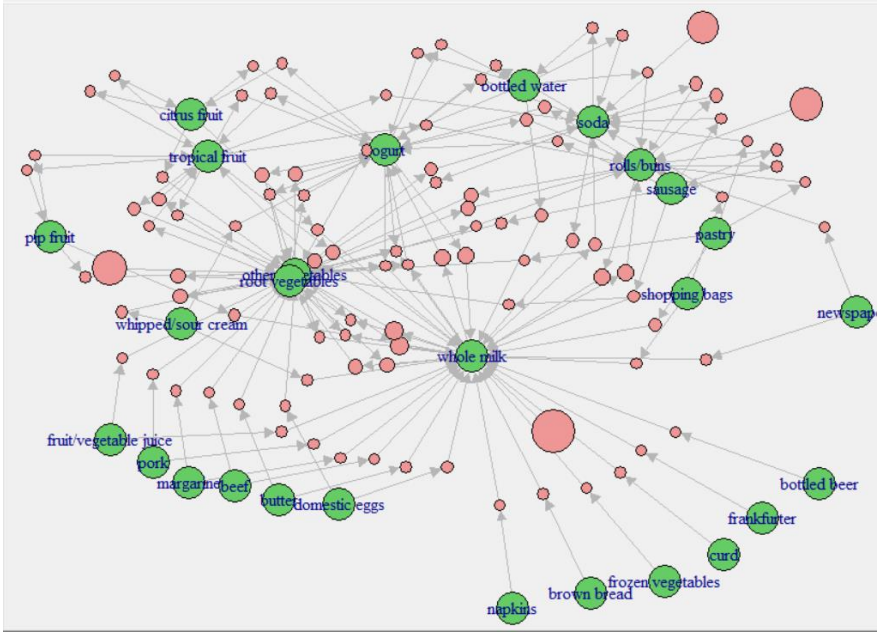
前五名的關聯規則分別是{bottled beer, red/blush wine}配{liquor}、{hamburger meat, soda}配{Instant food products}、{ham, white bread}配{processed cheese}、{other vegetables, root vegetables, whole milk, yogurt}配{rice}以及{bottled beer, liquor}配{red/blush wine}。各自的 support, confidence, lift 如下圖。

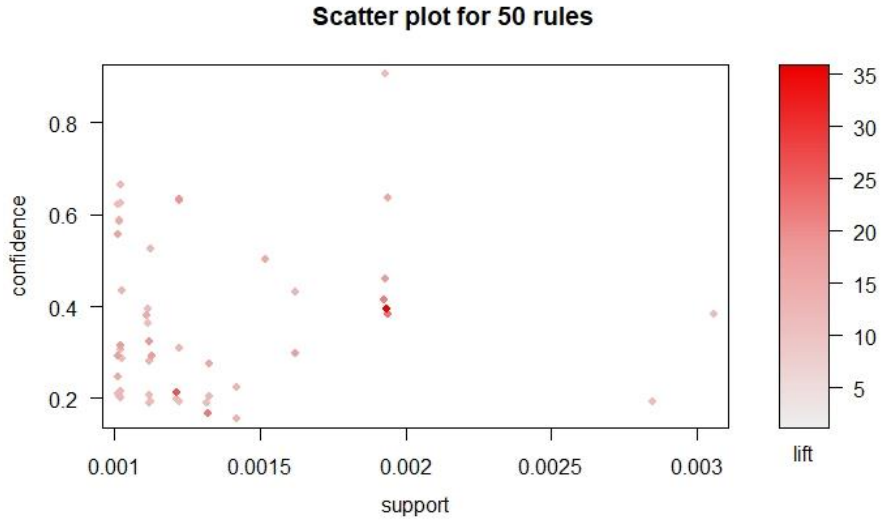
	lhs		rhs	support	confidence	lift	count
[1]	{bottled beer,red/blush wine}	=>	{liquor}	0.001931876	0.3958333	35.71579	19
[2]	{hamburger meat,soda}	=>	{Instant food products}	0.001220132	0.2105263	26.20919	12
[3]	{ham,white bread}	=>	{processed cheese}	0.001931876	0.3800000	22.92822	19
[4]	{other vegetables,root vegetables,whole milk,yogurt}	=>	{rice}	0.001321810	0.1688312	22.13939	13
[5]	{bottled beer,liquor}	=>	{red/blush wine}	0.001931876	0.4130435	21.49356	19

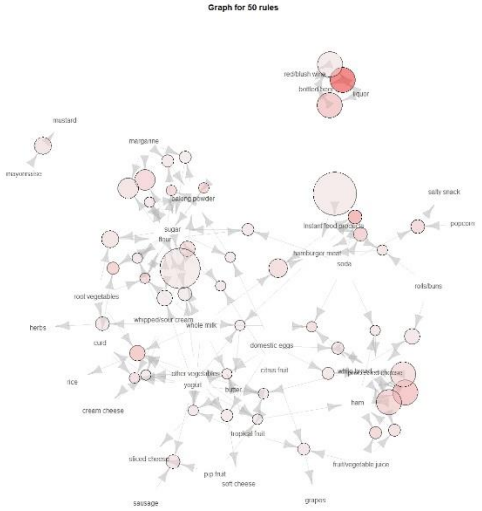
(3) Please provide/guess the “story” to interpret one of top 5 rules you are interested in.

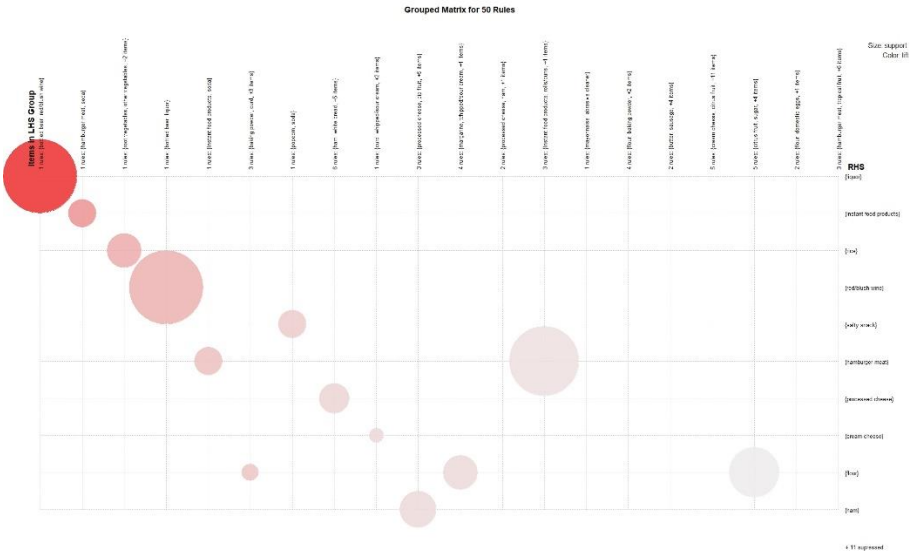
我比較感興趣的是第三名關聯規則{ham, white bread}配{processed cheese}，這個關聯應該是有消費者想要自己動手做漢堡。因為漢堡的組成最基本的就是白麵包、火腿以及起司，所以消費者一起買這些東西還蠻合理的。

(4) Give a visualization graph of your association rules.








Grouped Matrix for 50 Rules
Size: support
Color: lift
Items in LHS Group
RHS
{bars}
{instant food products}
{rice}
{pot/pan/dish wash}
{whole wheat}
{hamburger meat}
{processed cheese}
{cream cheese}
{flour}
{ham}
+ 11 suppressed