HOMEWORK 2
Submitted by: SHARANG BIRADAR N11407631

Part 1.

Data Exploration.

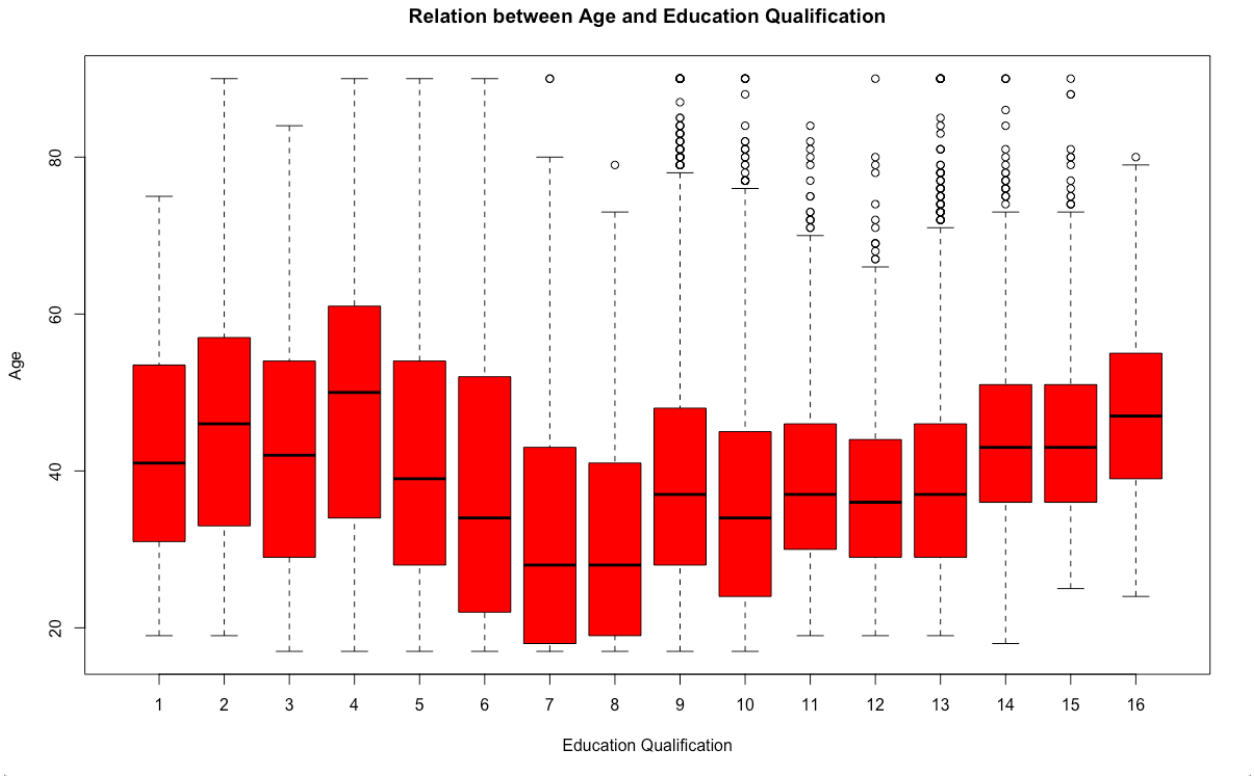
- Missing Values & Summary

```
> adult[adult ==' ?']=NA
> sum(is.na(adult))
[1] 4262
> summary(adult)
      age                  workclass         fnlwgt               education     education.number
 Min.   :17.00    Private         :22696   Min.   :  12285    HS-grad     :10501   Min.   : 1.00
 1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827    Some-college: 7291   1st Qu.: 9.00
 Median :37.00    Local-gov       : 2093   Median : 178356    Bachelors   : 5355   Median :10.00
 Mean   :38.58    State-gov       : 1298   Mean   : 189778    Masters     : 1723   Mean   :10.08
 3rd Qu.:48.00    Self-emp-inc    : 1116   3rd Qu.: 237051    Assoc-voc   : 1382   3rd Qu.:12.00
 Max.   :90.00    (Other)         :  981   Max.   :1484705    11th        : 1175   Max.   :16.00
                  NA's            : 1836                      (Other)     : 5134
               marital.status             occupation             relationship                    race           sex
  Divorced             : 4443    Prof-specialty : 4140    Husband       :13193    Amer-Indian-Eskimo:  311    Female:10771
  Married-AF-spouse    :   23    Craft-repair   : 4099    Not-in-family : 8305    Asian-Pac-Islander: 1039    Male  :21790
  Married-civ-spouse   :14976    Exec-managerial: 4066    Other-relative:  981    Black             : 3124
  Married-spouse-absent:  418    Adm-clerical   : 3770    Own-child     : 5068    Other             :  271
  Never-married        :10683    Sales          : 3650    Unmarried     : 3446    White             :27816
  Separated            : 1025    (Other)        :10993    Wife          : 1568
  Widowed              :  993    NA's           : 1843
  capital.gain    capital.loss    hours.per.week         native.country     salary
 Min.   :    0   Min.   :   0.0   Min.   : 1.00    United-States:29170    <=50K:24720
 1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00    Mexico       :  643    >50K : 7841
 Median :    0   Median :   0.0   Median :40.00    Philippines  :  198
 Mean   : 1078   Mean   :  87.3   Mean   :40.44    Germany      :  137
 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00    Canada       :  121
 Max.   :99999   Max.   :4356.0   Max.   :99.00    (Other)      : 1709
                                                   NA's         :  583
```

and…

- Distribution

```
> boxplot(adult$age~adult$education.number, xlab="Education Qualification", ylab="Age", ylim=c(17,90), col='red', main="Relation betwee
n Age and Education Qualification")
```



- Correlation

```
> cor(adult[c(1,3,5,13)])
                      age      fnlwgt education.number hours.per.week
age            1.00000000 -0.07664587       0.03652719     0.06875571
fnlwgt        -0.07664587  1.00000000      -0.04319463    -0.01876849
education.number  0.03652719 -0.04319463       1.00000000     0.14812273
hours.per.week    0.06875571 -0.01876849       0.14812273     1.00000000
```

- **Logistics Regression**

```
> mylogit<- glm(salary~ age+education.number+race+sex+hours.per.week+native.country, data=adult, family="binomial", na.action(adult))
> mylogit

Call:  glm(formula = salary ~ age + education.number + race + sex +
    hours.per.week + native.country, family = "binomial", data = adult,
    weights = na.action(adult))

Coefficients:
                              (Intercept)                                       age                            education.number
                                  -8.45563                                   0.04491                                   0.35255
                    race Asian-Pac-Islander                        race Black                                race Other
                                   0.42021                                   0.18723                                  -0.01685
                                race White                                sex Male                           hours.per.week
                                   0.58172                                   1.14003                                   0.03531
                     native.country Canada                     native.country China                   native.country Columbia
                                  -0.99893                                  -1.46261                                  -3.60036
                       native.country Cuba           native.country Dominican-Republic                    native.country Ecuador
                                  -1.01652                                  -2.54453                                  -1.52275
                native.country El-Salvador                native.country England                     native.country France
                                  -1.58913                                  -1.01452                                  -0.92537
                    native.country Germany                     native.country Greece                  native.country Guatemala
                                  -0.84508                                  -1.46767                                  -1.55944
                     native.country Haiti           native.country Holand-Netherlands                   native.country Honduras
                                  -1.41295                                 -11.06708                                  -1.88385
                       native.country Hong                    native.country Hungary                     native.country India
                                  -0.69226                                  -1.52163                                  -1.20776
                       native.country Iran                    native.country Ireland                     native.country Italy
                                  -1.16640                                  -1.28394                                  -0.52740
                    native.country Jamaica                      native.country Japan                      native.country Laos
                                  -1.17885                                  -0.72894                                  -1.42817
                     native.country Mexico                  native.country Nicaragua  native.country Outlying-US(Guam-USVI-etc)
                                  -1.83682                                  -2.07616                                 -12.25542
                       native.country Peru                native.country Philippines                    native.country Poland
                                  -2.10735                                  -0.82782                                  -1.53516
                    native.country Portugal                native.country Puerto-Rico                   native.country Scotland
                                  -1.34622                                  -1.43040                                  -1.19158
                     native.country South                      native.country Taiwan                  native.country Thailand
                                  -1.75831                                  -0.93254                                  -1.66677
          native.country Trinadad&Tobago              native.country United-States                    native.country Vietnam
                                  -1.19536                                  -1.11572                                  -2.35080
                native.country Yugoslavia
                                  -0.72474

Degrees of Freedom: 31977 Total (i.e. Null);  31929 Residual
  (583 observations deleted due to missingness)
Null Deviance:      35290
Residual Deviance: 27230        AIC: 27330
```

PART 2.

Q1. Create a dummy variable for “Winter” months defined as Oct, Nov, Dec, Jan & Feb. Use the “Month” variable to create this.

```
data$Month=factor(data$Month)
data$Month_names<-factor(data$Month, levels=1:12, labels=c("January","February", "March", "April",
                                                           "May", "June", "July", "August",
                                                           "September", "October", "November", "December"))
summary(data$Month)
data$Winter<- as.logical(0)
data$Non_Winter<- as.logical(0)
for (i in 1:nrow(data)){
             if (data$Month[i]=="October")
                        data$Winter[i]<-as.logical(1)
                   else if (data$Month[i]=="November")
                              data$Winter[i]<-as.logical(1)
                         else if (data$Month[i]=="December")
                                    data$Winter[i]<-as.logical(1)
                                else if (data$Month[i]=="January")
                                           data$Winter[i]<-as.logical(1)
                                      else if (data$Month[i]=="February")
                                                 data$Winter[i]<-as.logical(1)
                                            else
                                                       data$Non_Winter[i]<-as.logical(1)
                                                 }
```

Q2. Compute the “Market Share” for Progresso (as percentage of total sales) in the Winter vs. non-Winter months using the variable created in (1)

```
Sales_Winter_Month<-(data$Sales.Progresso[data$Winter=="TRUE"])
Sales_Non_Winter_Month<-(data$Sales.Progresso[data$Non_Winter=="TRUE"])

Category_Sales_Winter_Month<-(data$Category_Sales[data$Winter=="TRUE"])
Category_Sales_Non_Winter_Month<-(data$Category_Sales[data$Non_Winter=="TRUE"])

Market_Sales_Winter_Month<-(sum(Sales_Winter_Month)/sum(Category_Sales_Winter_Month))
Market_Sales_Non_Winter_Month<-(sum(Sales_Non_Winter_Month)/sum(Category_Sales_Non_Winter_Month))

Market_Sales_Winter_Month
Market_Sales_Non_Winter_Month

```

Market_Sales_Winter_Month= 0.2846215
Market_Sales_Non_Winter_Month= 0.1992817

Q3. Develop a linear regression model to predict Progresso sales. Explain the results of the regression model (model strength, variable importance, relationship between the predictor and dependent variables). Use 1st tab in file.

PTO..

```
> model<- lm(Sales.Progresso~ Month+Region+Price.Campbell+Price.PL+Price.Progresso
+            +Category_Sales, data=data )
> model

Call:
lm(formula = Sales.Progresso ~ Month + Region + Price.Campbell +
    Price.PL + Price.Progresso + Category_Sales, data = data)

Coefficients:
    (Intercept)    MonthFebruary       MonthMarch       MonthApril
       -780.9227        -164.2100        -327.8280        -296.2152
        MonthMay        MonthJune        MonthJuly      MonthAugust
       -309.7912        -216.4101        -296.6425        -360.1521
 MonthSeptember     MonthOctober    MonthNovember    MonthDecember
       -361.2857        -161.3227        -424.2446        -452.5144
  RegionMidWest      RegionSouth       RegionWest   Price.Campbell
      -1374.4564        -664.8532        -743.6411        2062.3648
       Price.PL  Price.Progresso   Category_Sales
        913.3372       -1594.5058           0.3174
```

and….

```
> summary(model)

Call:
lm(formula = Sales.Progresso ~ Month + Region + Price.Campbell +
    Price.PL + Price.Progresso + Category_Sales, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-13559.9   -478.6    -23.3    427.1  30335.3

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)     -7.809e+02  3.531e+01  -22.114   <2e-16 ***
MonthFebruary   -1.642e+02  1.851e+01   -8.870   <2e-16 ***
MonthMarch      -3.278e+02  1.861e+01  -17.619   <2e-16 ***
MonthApril      -2.962e+02  1.910e+01  -15.506   <2e-16 ***
MonthMay        -3.098e+02  1.936e+01  -15.999   <2e-16 ***
MonthJune       -2.164e+02  1.892e+01  -11.437   <2e-16 ***
MonthJuly       -2.966e+02  1.883e+01  -15.751   <2e-16 ***
MonthAugust     -3.602e+02  1.844e+01  -19.528   <2e-16 ***
MonthSeptember  -3.613e+02  1.793e+01  -20.145   <2e-16 ***
MonthOctober    -1.613e+02  1.776e+01   -9.081   <2e-16 ***
MonthNovember   -4.242e+02  1.774e+01  -23.919   <2e-16 ***
MonthDecember   -4.525e+02  1.770e+01  -25.568   <2e-16 ***
RegionMidWest   -1.374e+03  1.159e+01 -118.613   <2e-16 ***
RegionSouth     -6.649e+02  1.069e+01  -62.214   <2e-16 ***
RegionWest      -7.436e+02  1.133e+01  -65.608   <2e-16 ***
Price.Campbell   2.062e+03  1.941e+01  106.247   <2e-16 ***
Price.PL         9.133e+02  2.027e+01   45.060   <2e-16 ***
Price.Progresso -1.595e+03  1.299e+01 -122.787   <2e-16 ***
Category_Sales   3.174e-01  8.834e-04  359.252   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1064 on 88390 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7554
F-statistic: 1.517e+04 on 18 and 88390 DF,  p-value: < 2.2e-16
```

Explanation:
The model explains the role of independent variables in predicting the value of the dependent variables in the Sales.

- With increase in the Price in Campbell and Price in PL the Revenue is increased by 2062.3 and 9133.2 respectively.
- With increase in the Price of Progresso by 1 dollar the Revenue decreases by 1595.3.
- If the Category Sales increase by 1 dollar the Revenue increases by about 31 cents.

Overall the model explains about 75.55% of variability in the system which is good.

All the Independent variables are statistically significant when it comes to predicting the values with Dependent Variables.