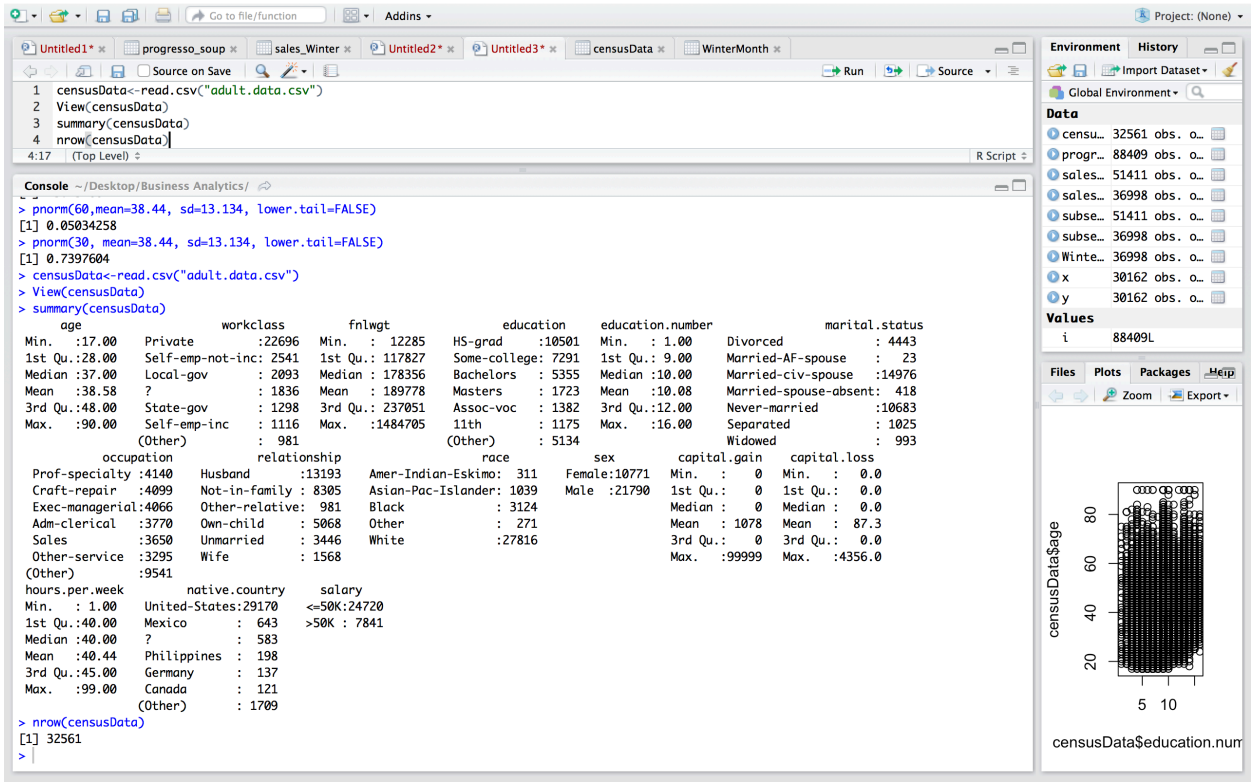
-Priyanka Thakkar | N14897740 | pbt226

Business Analytics Assignment 2

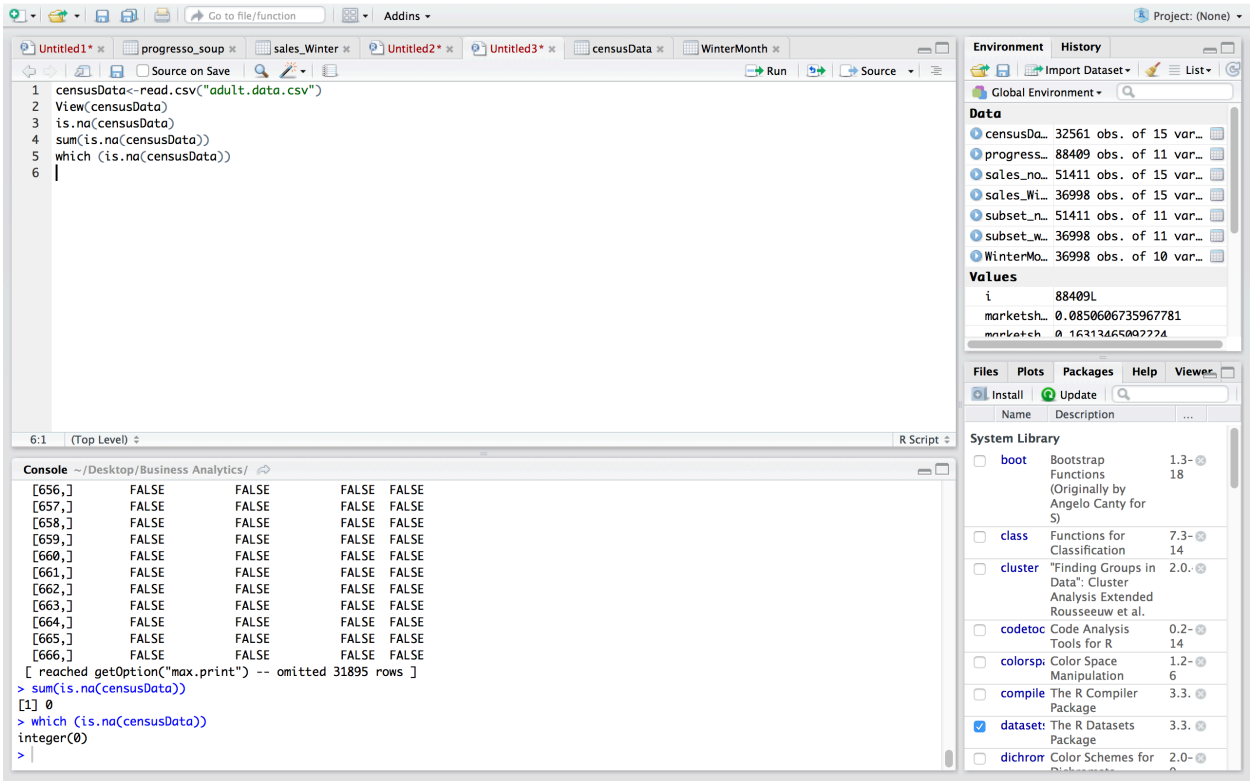
Homework Part 1

1) Census Income Data Set in the UCI Machine Learning Repository
   https://archive.ics.uci.edu/ml/datasets/Census+Income

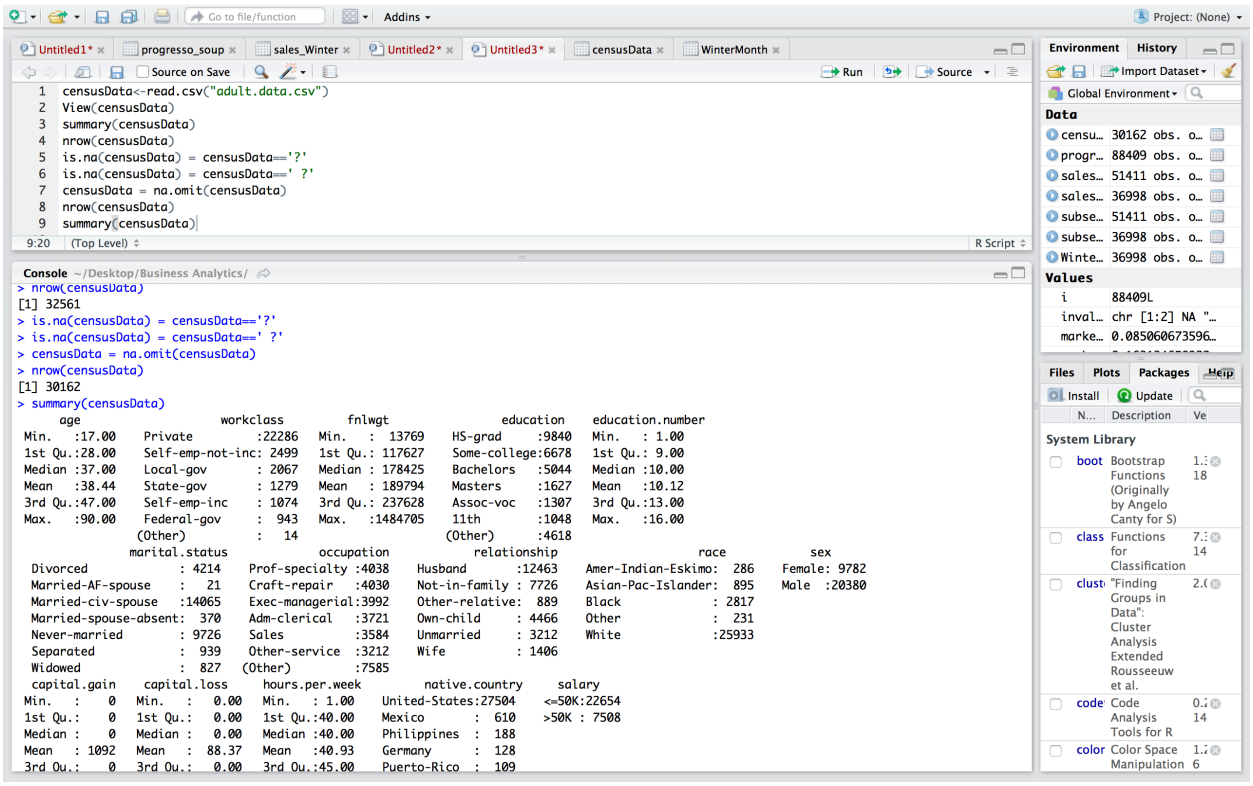


2) Data exploration
   - Missing and/or invalid values??

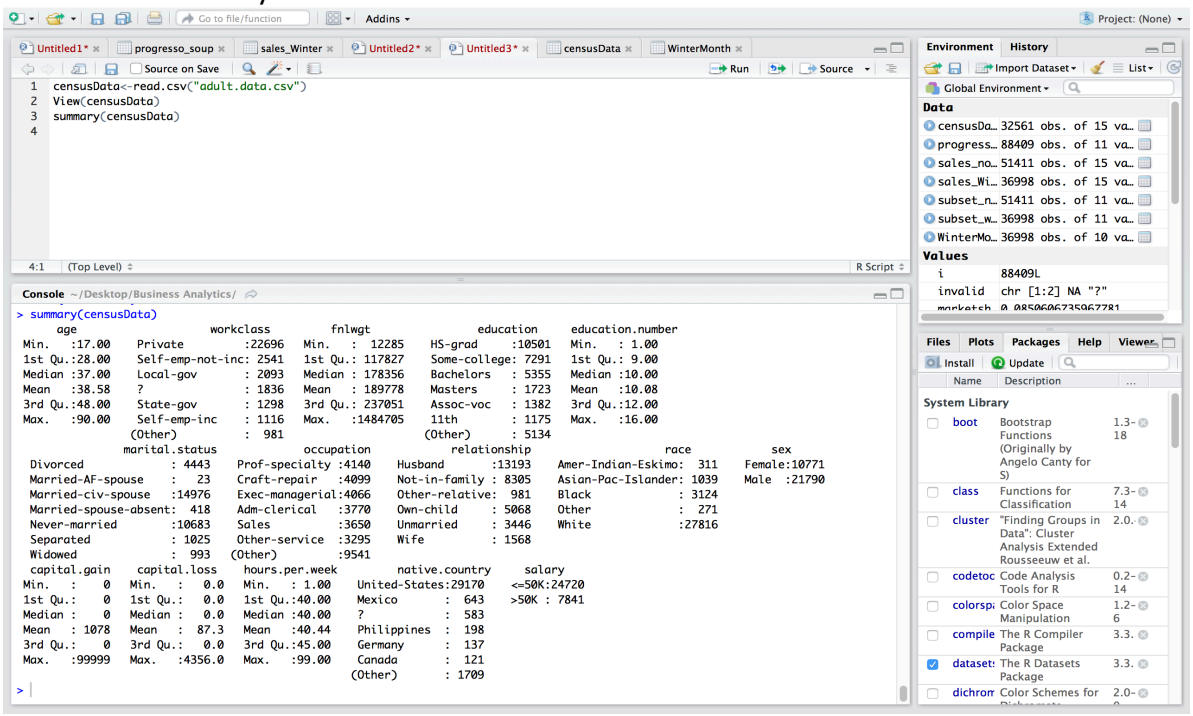
Checked for NA records



I could see missing data "?" in the summary for censusData.
Therefore, removed the records with missing value.



- Summary statistics

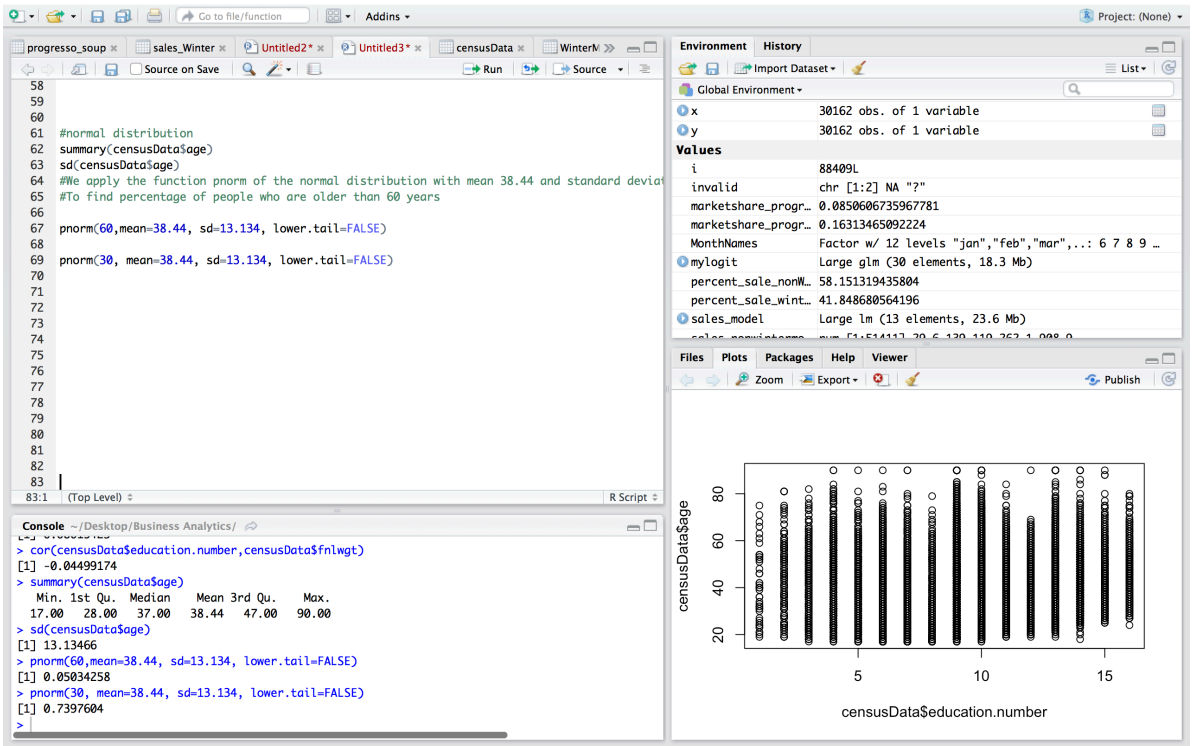


- Distributions

Normal distribution using 'pnorm'

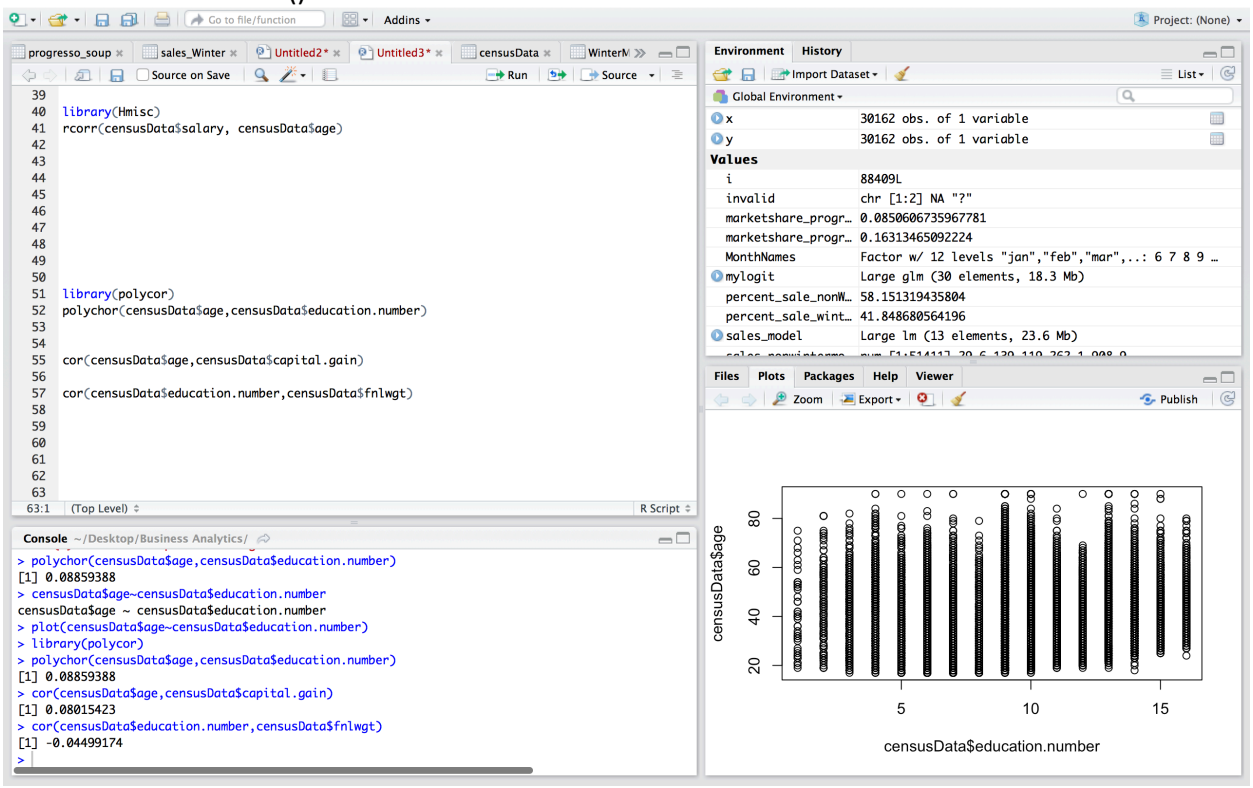
To calculate percentage of people who are older than 60 years : 5.03%

To calculate percentage of people who are older than 30 years : 73.97%

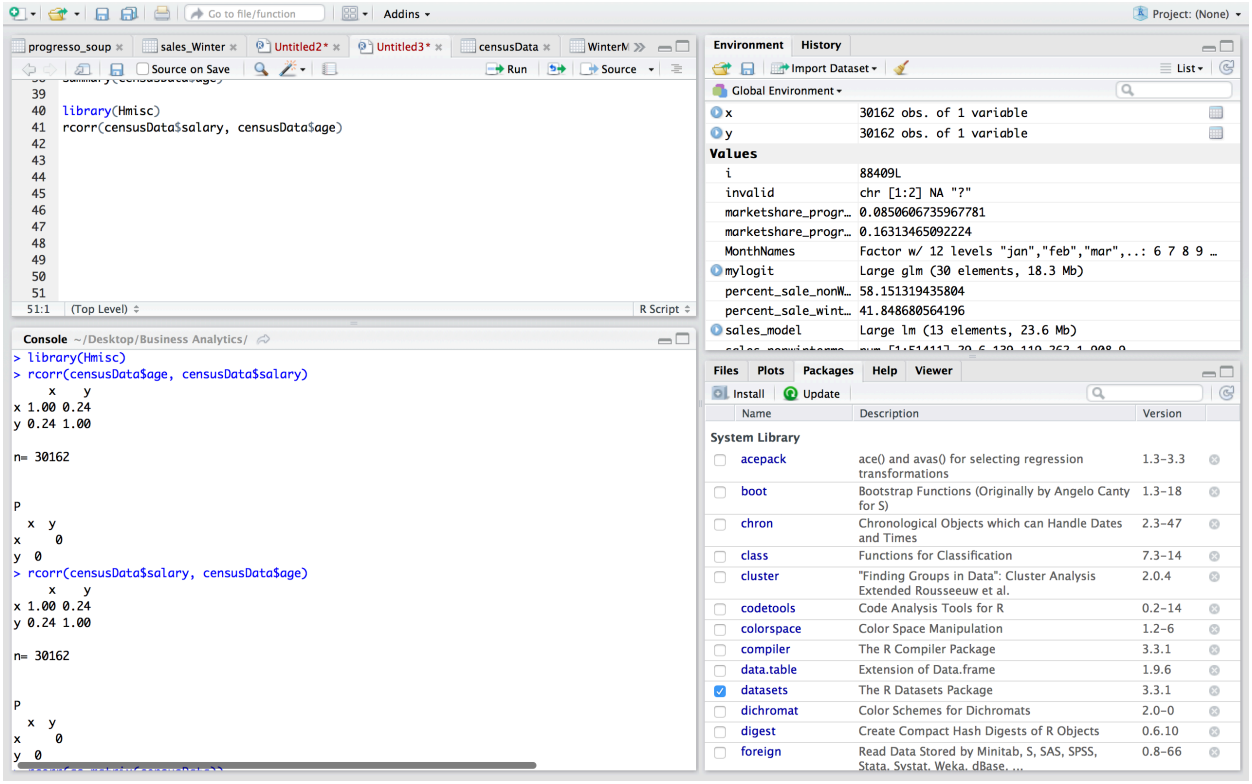


- Correlations

- Polychoric Correlation
- Performed cor()

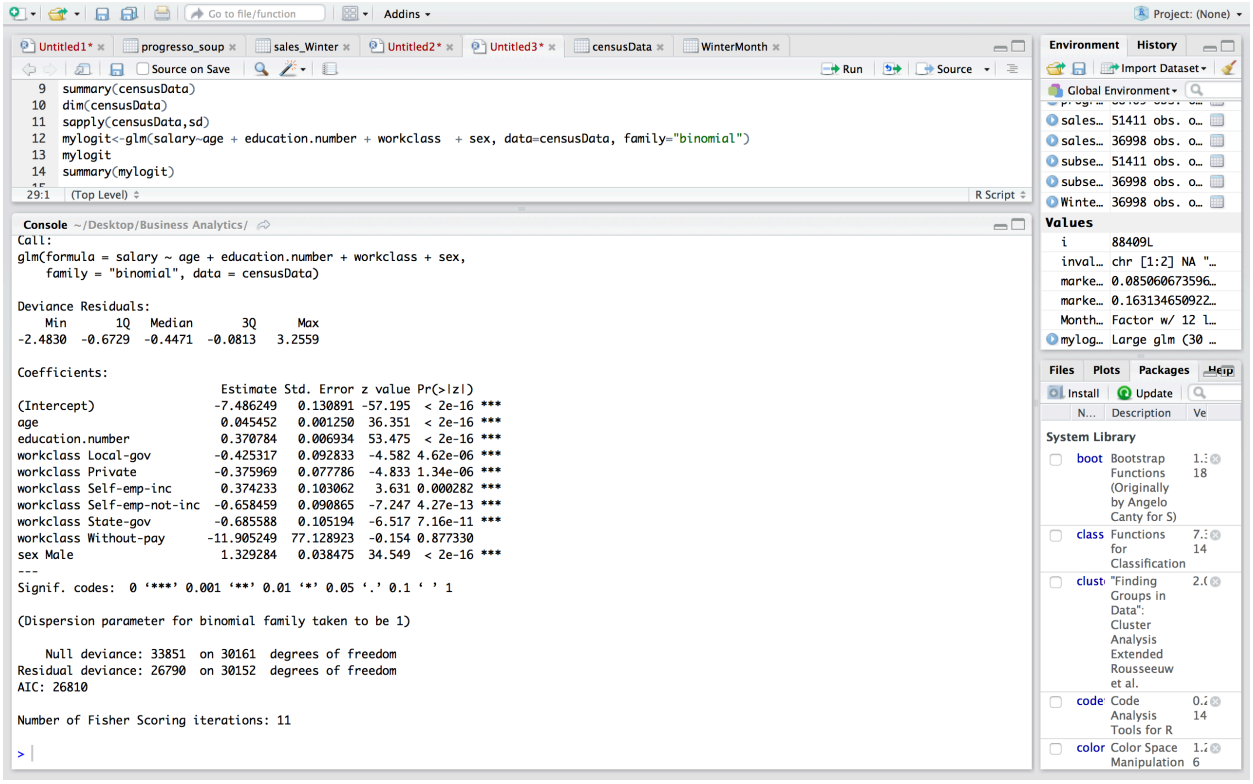


- Used Library Hmisc to implement rcorr
  (To find correlation between variables that are not numeric)



3) Prediction
   - Run a logistic regression

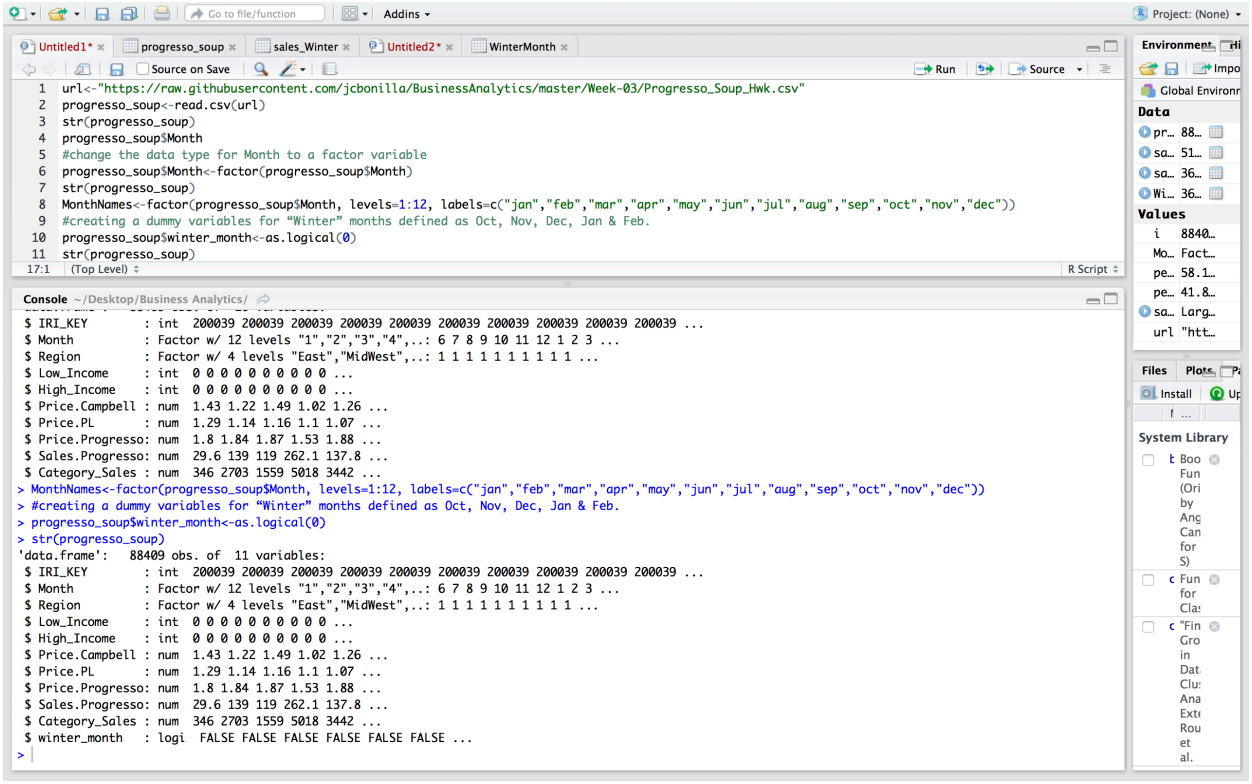
Since there are two possible outcomes i.e. salary >50k or salary<=50k, I made use of Logistic Regression
Age, sex and education Number contributes a lot to determine the salary of a person.
Implication:
Salary increases by 0.045 if the age increases by one year



Homework Part 2

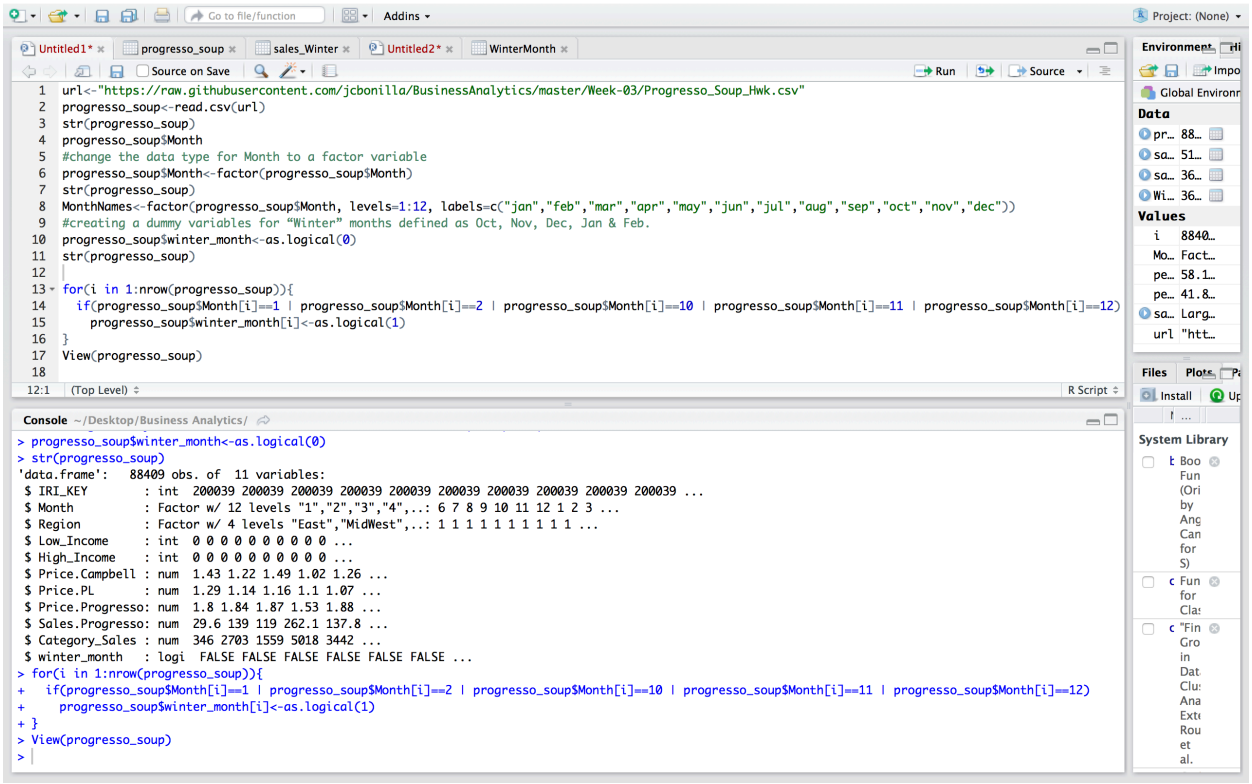
- Create a dummy variable for “Winter” months defined as Oct, Nov, Dec, Jan & Feb. Use the “Month” variable to create this.

Created a Dummy Variable ‘winter_month’ representing months ‘oct, nov, dec, jan, feb’



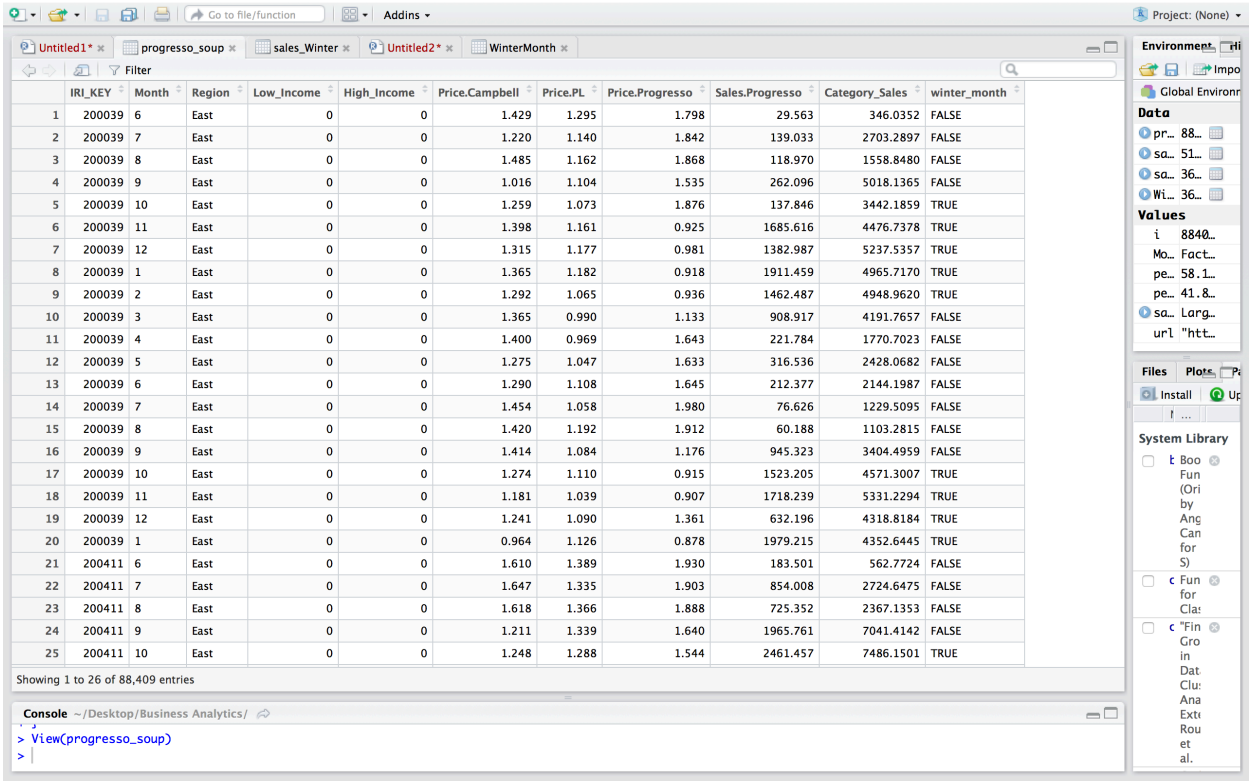
```
url<-"https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/Week-03/Progresso_Soup_Hwk.csv"
progresso_soup<-read.csv(url)
str(progresso_soup)
progresso_soup$Month
#change the data type for Month to a factor variable
progresso_soup$Month<-factor(progresso_soup$Month)
str(progresso_soup)
MonthNames<-factor(progresso_soup$Month, levels=1:12, labels=c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"))
#creating a dummy variables for "Winter" months defined as Oct, Nov, Dec, Jan & Feb.
progresso_soup$winter_month<-as.logical(0)
str(progresso_soup)
```

```
 $ IRI_KEY        : int  200039 200039 200039 200039 200039 200039 200039 200039 200039 200039 ...
 $ Month          : Factor w/ 12 levels "1","2","3","4",..: 6 7 8 9 10 11 12 1 2 3 ...
 $ Region         : Factor w/ 4 levels "East","MidWest",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Low_Income     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ High_Income    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Price.Campbell : num  1.43 1.22 1.49 1.02 1.26 ...
 $ Price.PL       : num  1.29 1.14 1.16 1.1 1.07 ...
 $ Price.Progresso: num  1.8 1.84 1.87 1.53 1.88 ...
 $ Sales.Progresso: num  29.6 139 119 262.1 137.8 ...
 $ Category_Sales : num  346 2703 1559 5018 3442 ...
> MonthNames<-factor(progresso_soup$Month, levels=1:12, labels=c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"))
> #creating a dummy variables for "Winter" months defined as Oct, Nov, Dec, Jan & Feb.
> progresso_soup$winter_month<-as.logical(0)
> str(progresso_soup)
'data.frame':   88409 obs. of  11 variables:
 $ IRI_KEY        : int  200039 200039 200039 200039 200039 200039 200039 200039 200039 200039 ...
 $ Month          : Factor w/ 12 levels "1","2","3","4",..: 6 7 8 9 10 11 12 1 2 3 ...
 $ Region         : Factor w/ 4 levels "East","MidWest",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Low_Income     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ High_Income    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Price.Campbell : num  1.43 1.22 1.49 1.02 1.26 ...
 $ Price.PL       : num  1.29 1.14 1.16 1.1 1.07 ...
 $ Price.Progresso: num  1.8 1.84 1.87 1.53 1.88 ...
 $ Sales.Progresso: num  29.6 139 119 262.1 137.8 ...
 $ Category_Sales : num  346 2703 1559 5018 3442 ...
 $ winter_month   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
>
```



```
url<-"https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/Week-03/Progresso_Soup_Hwk.csv"
progresso_soup<-read.csv(url)
str(progresso_soup)
progresso_soup$Month
#change the data type for Month to a factor variable
progresso_soup$Month<-factor(progresso_soup$Month)
str(progresso_soup)
MonthNames<-factor(progresso_soup$Month, levels=1:12, labels=c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"))
#creating a dummy variables for "Winter" months defined as Oct, Nov, Dec, Jan & Feb.
progresso_soup$winter_month<-as.logical(0)
str(progresso_soup)

for(i in 1:nrow(progresso_soup)){
  if(progresso_soup$Month[i]==1 | progresso_soup$Month[i]==2 | progresso_soup$Month[i]==10 | progresso_soup$Month[i]==11 | progresso_soup$Month[i]==12)
    progresso_soup$winter_month[i]<-as.logical(1)
}
View(progresso_soup)
```

```
> progresso_soup$winter_month<-as.logical(0)
> str(progresso_soup)
'data.frame':   88409 obs. of  11 variables:
 $ IRI_KEY        : int  200039 200039 200039 200039 200039 200039 200039 200039 200039 200039 ...
 $ Month          : Factor w/ 12 levels "1","2","3","4",..: 6 7 8 9 10 11 12 1 2 3 ...
 $ Region         : Factor w/ 4 levels "East","MidWest",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Low_Income     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ High_Income    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Price.Campbell : num  1.43 1.22 1.49 1.02 1.26 ...
 $ Price.PL       : num  1.29 1.14 1.16 1.1 1.07 ...
 $ Price.Progresso: num  1.8 1.84 1.87 1.53 1.88 ...
 $ Sales.Progresso: num  29.6 139 119 262.1 137.8 ...
 $ Category_Sales : num  346 2703 1559 5018 3442 ...
 $ winter_month   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
> for(i in 1:nrow(progresso_soup)){
+   if(progresso_soup$Month[i]==1 | progresso_soup$Month[i]==2 | progresso_soup$Month[i]==10 | progresso_soup$Month[i]==11 | progresso_soup$Month[i]==12)
+     progresso_soup$winter_month[i]<-as.logical(1)
+ }
> View(progresso_soup)
>
```

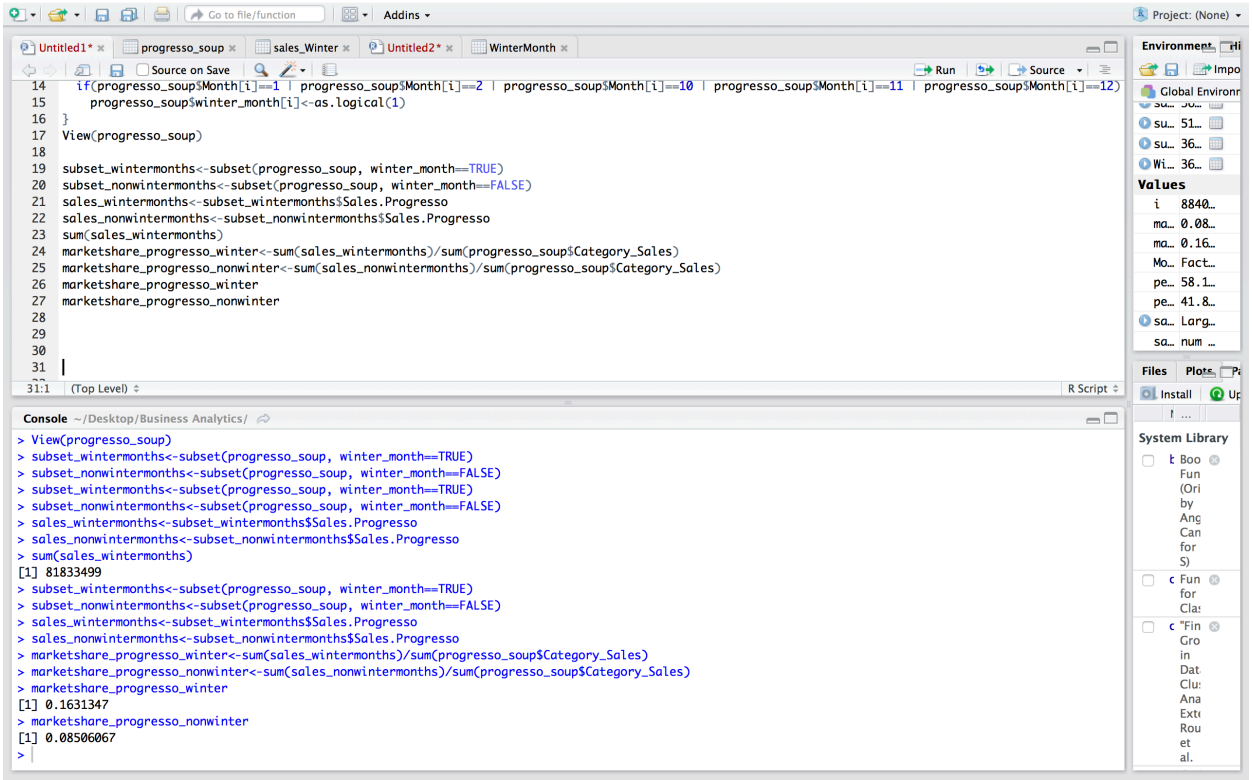


	IRI_KEY	Month	Region	Low_Income	High_Income	Price.Campbell	Price.PL	Price.Progresso	Sales.Progresso	Category_Sales	winter_month
1	200039	6	East	0	0	1.429	1.295	1.798	29.563	346.0352	FALSE
2	200039	7	East	0	0	1.220	1.140	1.842	139.033	2703.2897	FALSE
3	200039	8	East	0	0	1.485	1.162	1.868	118.970	1558.8480	FALSE
4	200039	9	East	0	0	1.016	1.104	1.535	262.096	5018.1365	FALSE
5	200039	10	East	0	0	1.259	1.073	1.876	137.846	3442.1859	TRUE
6	200039	11	East	0	0	1.398	1.161	0.925	1685.616	4476.7378	TRUE
7	200039	12	East	0	0	1.315	1.177	0.981	1382.987	5237.5357	TRUE
8	200039	1	East	0	0	1.365	1.182	0.918	1911.459	4965.7170	TRUE
9	200039	2	East	0	0	1.292	1.065	0.936	1462.487	4948.9620	TRUE
10	200039	3	East	0	0	1.365	0.990	1.133	908.917	4191.7657	FALSE
11	200039	4	East	0	0	1.400	0.969	1.643	221.784	1770.7023	FALSE
12	200039	5	East	0	0	1.275	1.047	1.633	316.536	2428.0682	FALSE
13	200039	6	East	0	0	1.290	1.108	1.645	212.377	2144.1987	FALSE
14	200039	7	East	0	0	1.454	1.058	1.980	76.626	1229.5095	FALSE
15	200039	8	East	0	0	1.420	1.192	1.912	60.188	1103.2815	FALSE
16	200039	9	East	0	0	1.414	1.084	1.176	945.323	3404.4959	FALSE
17	200039	10	East	0	0	1.274	1.110	0.915	1523.205	4571.3007	TRUE
18	200039	11	East	0	0	1.181	1.039	0.907	1718.239	5331.2294	TRUE
19	200039	12	East	0	0	1.241	1.090	1.361	632.196	4318.8184	TRUE
20	200039	1	East	0	0	0.964	1.126	0.878	1979.215	4352.6445	TRUE
21	200411	6	East	0	0	1.610	1.389	1.930	183.501	562.7724	FALSE
22	200411	7	East	0	0	1.647	1.335	1.903	854.008	2724.6475	FALSE
23	200411	8	East	0	0	1.618	1.366	1.888	725.352	2367.1353	FALSE
24	200411	9	East	0	0	1.211	1.339	1.640	1965.761	7041.4142	FALSE
25	200411	10	East	0	0	1.248	1.288	1.544	2461.457	7486.1501	TRUE

Showing 1 to 26 of 88,409 entries

```
> View(progresso_soup)
>
```

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



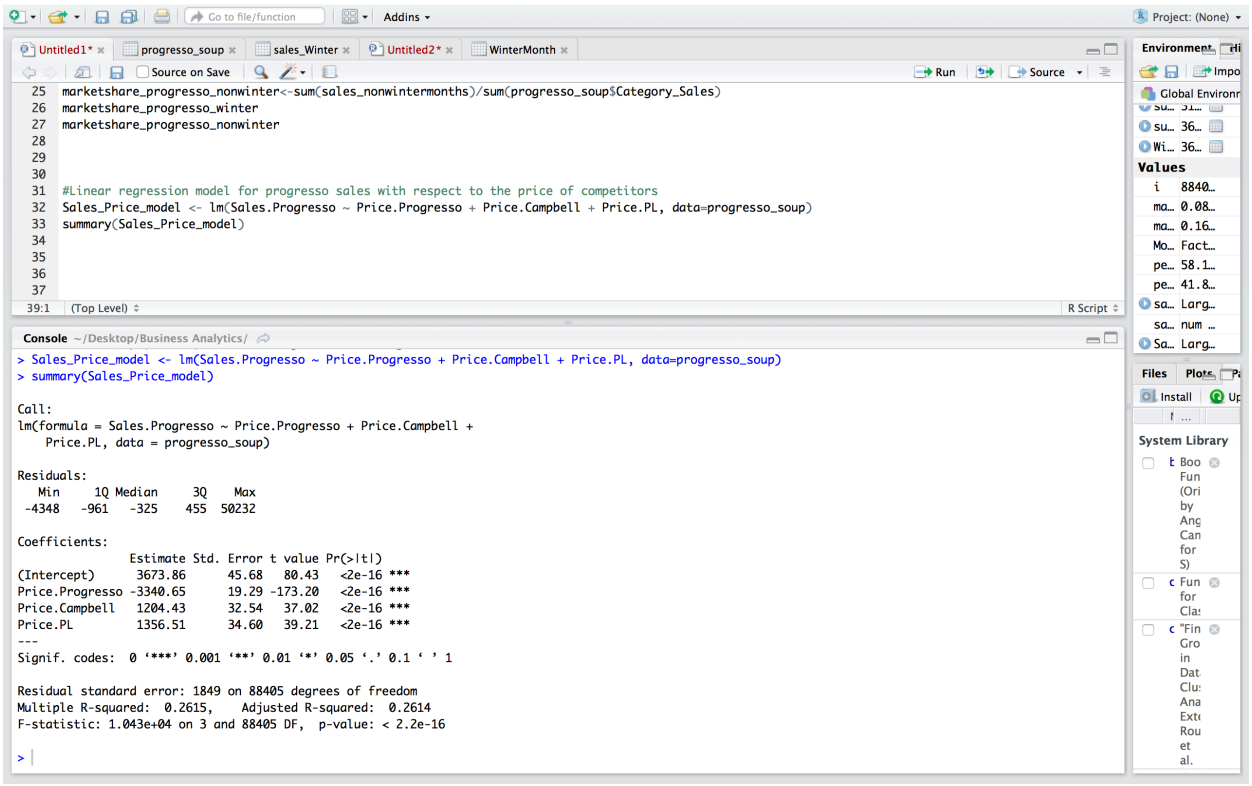
```
  if(progresso_soup$Month[i]==1 | progresso_soup$Month[i]==2 | progresso_soup$Month[i]==10 | progresso_soup$Month[i]==11 | progresso_soup$Month[i]==12)
    progresso_soup$winter_month[i]<-as.logical(1)
}
View(progresso_soup)

subset_wintermonths<-subset(progresso_soup, winter_month==TRUE)
subset_nonwintermonths<-subset(progresso_soup, winter_month==FALSE)
sales_wintermonths<-subset_wintermonths$Sales.Progresso
sales_nonwintermonths<-subset_nonwintermonths$Sales.Progresso
sum(sales_wintermonths)
marketshare_progresso_winter<-sum(sales_wintermonths)/sum(progresso_soup$Category_Sales)
marketshare_progresso_nonwinter<-sum(sales_nonwintermonths)/sum(progresso_soup$Category_Sales)
marketshare_progresso_winter
marketshare_progresso_nonwinter
```

```
> View(progresso_soup)
> subset_wintermonths<-subset(progresso_soup, winter_month==TRUE)
> subset_nonwintermonths<-subset(progresso_soup, winter_month==FALSE)
> subset_wintermonths<-subset(progresso_soup, winter_month==TRUE)
> subset_nonwintermonths<-subset(progresso_soup, winter_month==FALSE)
> sales_wintermonths<-subset_wintermonths$Sales.Progresso
> sales_nonwintermonths<-subset_nonwintermonths$Sales.Progresso
> sum(sales_wintermonths)
[1] 81833499
> subset_wintermonths<-subset(progresso_soup, winter_month==TRUE)
> subset_nonwintermonths<-subset(progresso_soup, winter_month==FALSE)
> sales_wintermonths<-subset_wintermonths$Sales.Progresso
> sales_nonwintermonths<-subset_nonwintermonths$Sales.Progresso
> marketshare_progresso_winter<-sum(sales_wintermonths)/sum(progresso_soup$Category_Sales)
> marketshare_progresso_nonwinter<-sum(sales_nonwintermonths)/sum(progresso_soup$Category_Sales)
> marketshare_progresso_winter
[1] 0.1631347
> marketshare_progresso_nonwinter
[1] 0.08506067
>
```

Market Share for progresso is 16.31% in Winter months and 8.5% in non Winter Months.

- 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.

  Initial analysis to see how price structuring and price of competition impacts Progresso sales.

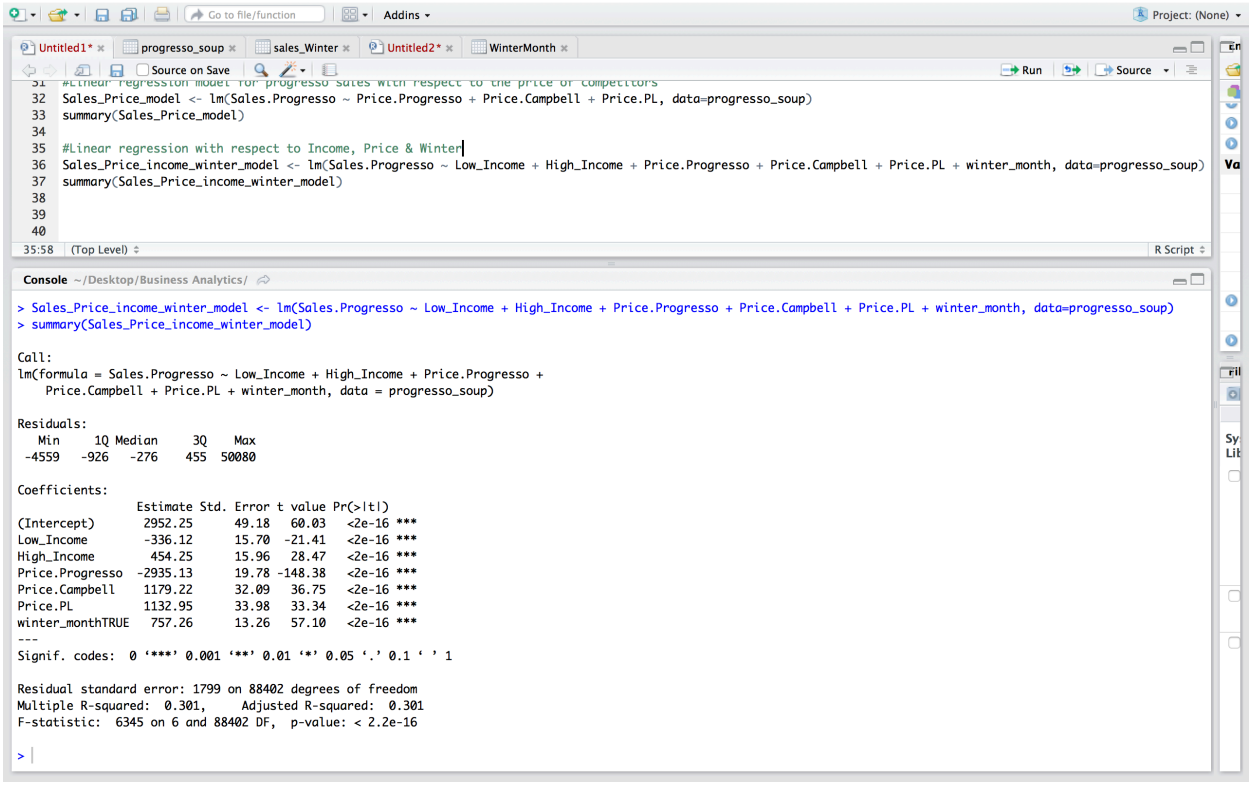


```
marketshare_progresso_nonwinter<-sum(sales_nonwintermonths)/sum(progresso_soup$Category_Sales)
marketshare_progresso_winter
marketshare_progresso_nonwinter

#Linear regression model for progresso sales with respect to the price of competitors
Sales_Price_model <- lm(Sales.Progresso ~ Price.Progresso + Price.Campbell + Price.PL, data=progresso_soup)
summary(Sales_Price_model)
```

```
> Sales_Price_model <- lm(Sales.Progresso ~ Price.Progresso + Price.Campbell + Price.PL, data=progresso_soup)
> summary(Sales_Price_model)

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

Residuals:
   Min     1Q Median     3Q    Max
 -4348   -961   -325    455  50232

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      3673.86      45.68   80.43   <2e-16 ***
Price.Progresso -3340.65      19.29 -173.20   <2e-16 ***
Price.Campbell   1204.43      32.54   37.02   <2e-16 ***
Price.PL         1356.51      34.60   39.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1849 on 88405 degrees of freedom
Multiple R-squared:  0.2615,    Adjusted R-squared:  0.2614
F-statistic: 1.043e+04 on 3 and 88405 DF,  p-value: < 2.2e-16
```

predictor variables: Price.Progresso, Price.Combell, Price.PL (Competitor Prices)
dependent variable: Sales.Progressor
Model Strength: R-squared: 0.26 indicates moderate strength of the model
Results represent progresso price has the strongest influence on sales (coefficent -3340 and t test = -173 > 1.96 for two tailed).
Price of competition also impacts Progesso Sales (increased price of competition = increase in Progresso sales)

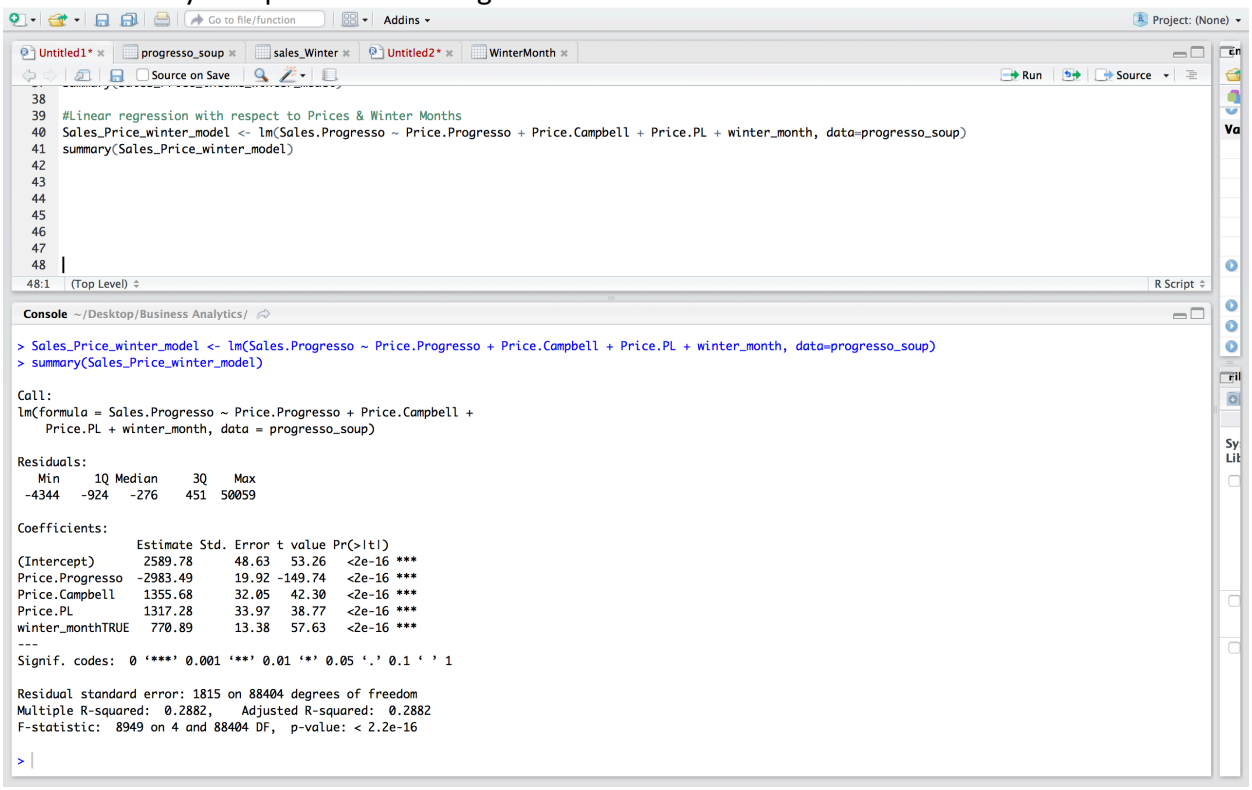
The second Linear regression includes income, Price winter months
It indicates Low income shoppers do not buy Progresso



```
#Linear regression model for progresso sales with respect to the price of competitors
Sales_Price_model <- lm(Sales.Progresso ~ Price.Progresso + Price.Campbell + Price.PL, data=progresso_soup)
summary(Sales_Price_model)

#Linear regression with respect to Income, Price & Winter
Sales_Price_income_winter_model <- lm(Sales.Progresso ~ Low_Income + High_Income + Price.Progresso + Price.Campbell + Price.PL + winter_month, data=progresso_soup)
summary(Sales_Price_income_winter_model)
```

```
> Sales_Price_income_winter_model <- lm(Sales.Progresso ~ Low_Income + High_Income + Price.Progresso + Price.Campbell + Price.PL + winter_month, data=progresso_soup)
> summary(Sales_Price_income_winter_model)

Call:
lm(formula = Sales.Progresso ~ Low_Income + High_Income + Price.Progresso +
    Price.Campbell + Price.PL + winter_month, data = progresso_soup)

Residuals:
   Min     1Q Median     3Q    Max
 -4559   -926   -276    455  50080

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2952.25      49.18   60.03   <2e-16 ***
Low_Income        -336.12      15.70  -21.41   <2e-16 ***
High_Income        454.25      15.96   28.47   <2e-16 ***
Price.Progresso  -2935.13      19.78 -148.38   <2e-16 ***
Price.Campbell    1179.22      32.09   36.75   <2e-16 ***
Price.PL          1132.95      33.98   33.34   <2e-16 ***
winter_monthTRUE   757.26      13.26   57.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1799 on 88402 degrees of freedom
Multiple R-squared:  0.301,     Adjusted R-squared:  0.301
F-statistic:  6345 on 6 and 88402 DF,  p-value: < 2.2e-16
```

The third analysis represent that Progresso Sales increases in winter



```
#Linear regression with respect to Prices & Winter Months
Sales_Price_winter_model <- lm(Sales.Progresso ~ Price.Progresso + Price.Campbell + Price.PL + winter_month, data=progresso_soup)
summary(Sales_Price_winter_model)
```

```
> Sales_Price_winter_model <- lm(Sales.Progresso ~ Price.Progresso + Price.Campbell + Price.PL + winter_month, data=progresso_soup)
> summary(Sales_Price_winter_model)

Call:
lm(formula = Sales.Progresso ~ Price.Progresso + Price.Campbell +
    Price.PL + winter_month, data = progresso_soup)

Residuals:
   Min     1Q Median     3Q    Max
 -4344   -924   -276    451  50059

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2589.78      48.63   53.26   <2e-16 ***
Price.Progresso  -2983.49      19.92 -149.74   <2e-16 ***
Price.Campbell    1355.68      32.05   42.30   <2e-16 ***
Price.PL          1317.28      33.97   38.77   <2e-16 ***
winter_monthTRUE   770.89      13.38   57.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1815 on 88404 degrees of freedom
Multiple R-squared:  0.2882,    Adjusted R-squared:  0.2882
F-statistic:  8949 on 4 and 88404 DF,  p-value: < 2.2e-16
```