**Multivariate Linear Regression**

**Author:** Harshita Singh

**Background**

The data has been covered to determine the education factors of two groups- Treat and Control, where Treat were provided additional education and Control were not. This analysis will help to discover the different factors involve in survey of 2021 residents of Canada.

This analysis will help to determine the key factors which are associated to the Political Awareness.

**Data Source**

Data was obtained by survey of 2021 residents of Canada which includes information such as, nationality, housing, age, marital status, political affiliation, number of children’s, annual household income, food, score on political awareness and the attempts.

**Data Transformation and Cleaning (Description)**

**Dummy Variables:**

**Nation:**

The data identifying the nationality of the surveyor for the particular group: Treat and Control was transformed to 4 dummies variable: nationAsia, nationEurope, nationNorthAmerica, nationSouthern

I have included all the code used for transformations in the Appendix.

**Treat/Control Field:** Treat\_HS, Control\_HS

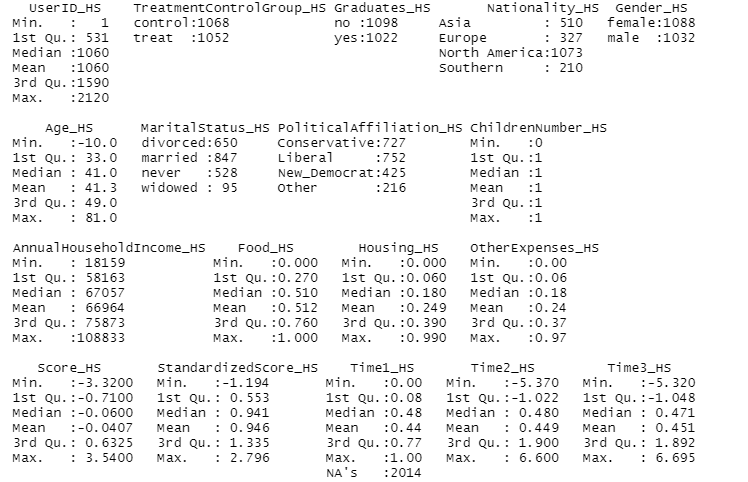
**Graduate:** No\_Graduates\_HS; Yes\_Gradauates\_HS

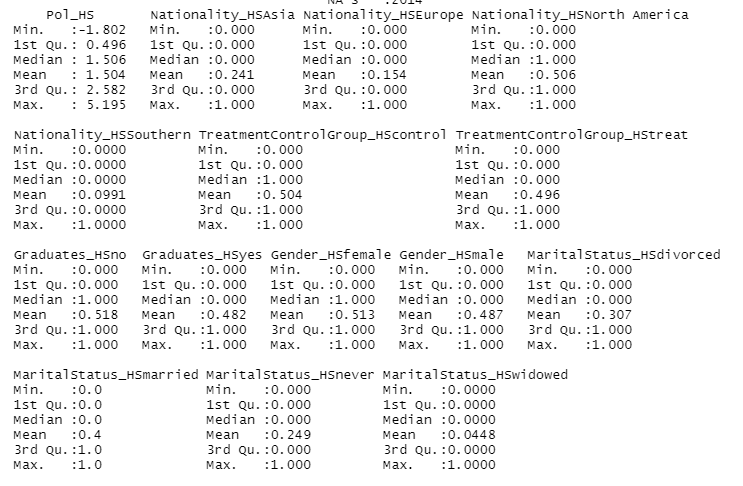
**Gender:** Female\_HS; Male\_HS

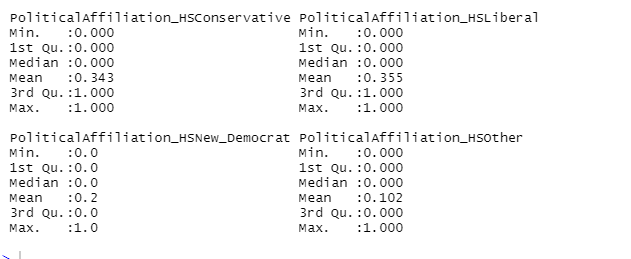
**Martial:** Married\_HS;Never\_HS,Divorced\_HS, Other\_HS

**Political Affiliation:** Conservative\_HS; Liberal\_HS; NewDemocrat\_HS

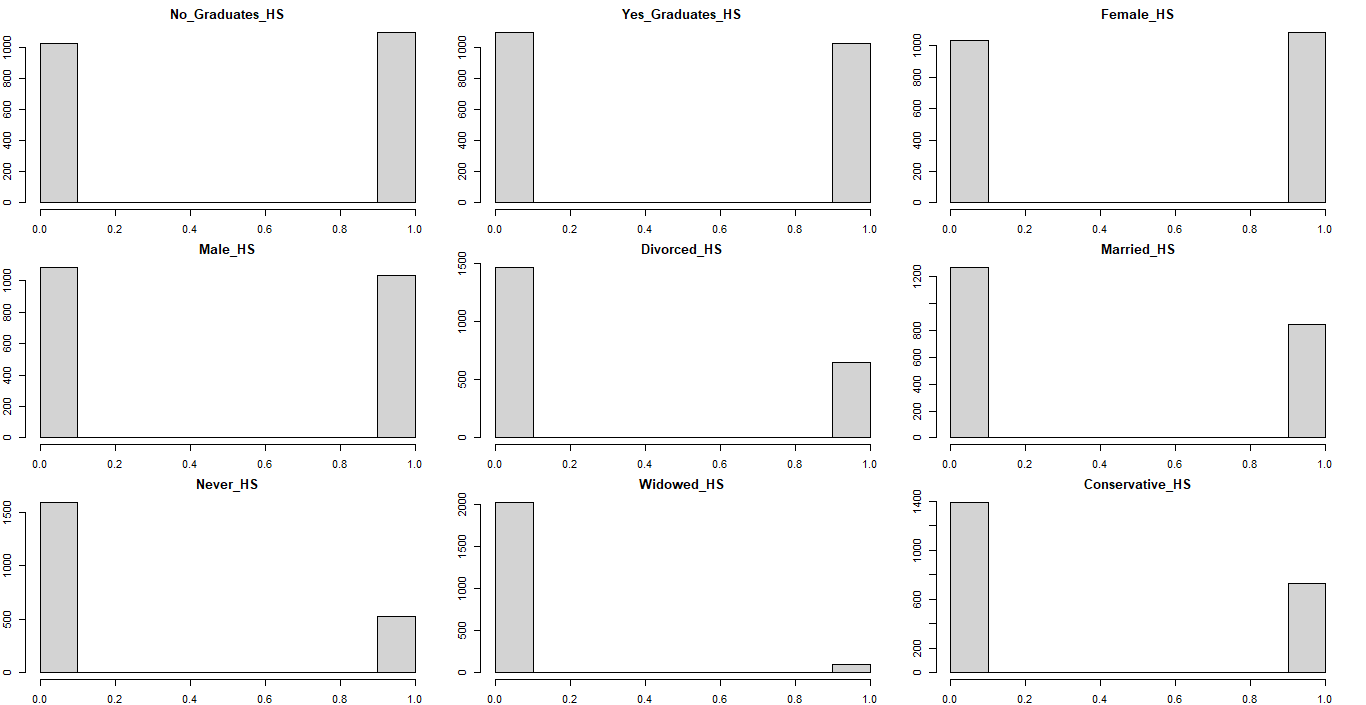
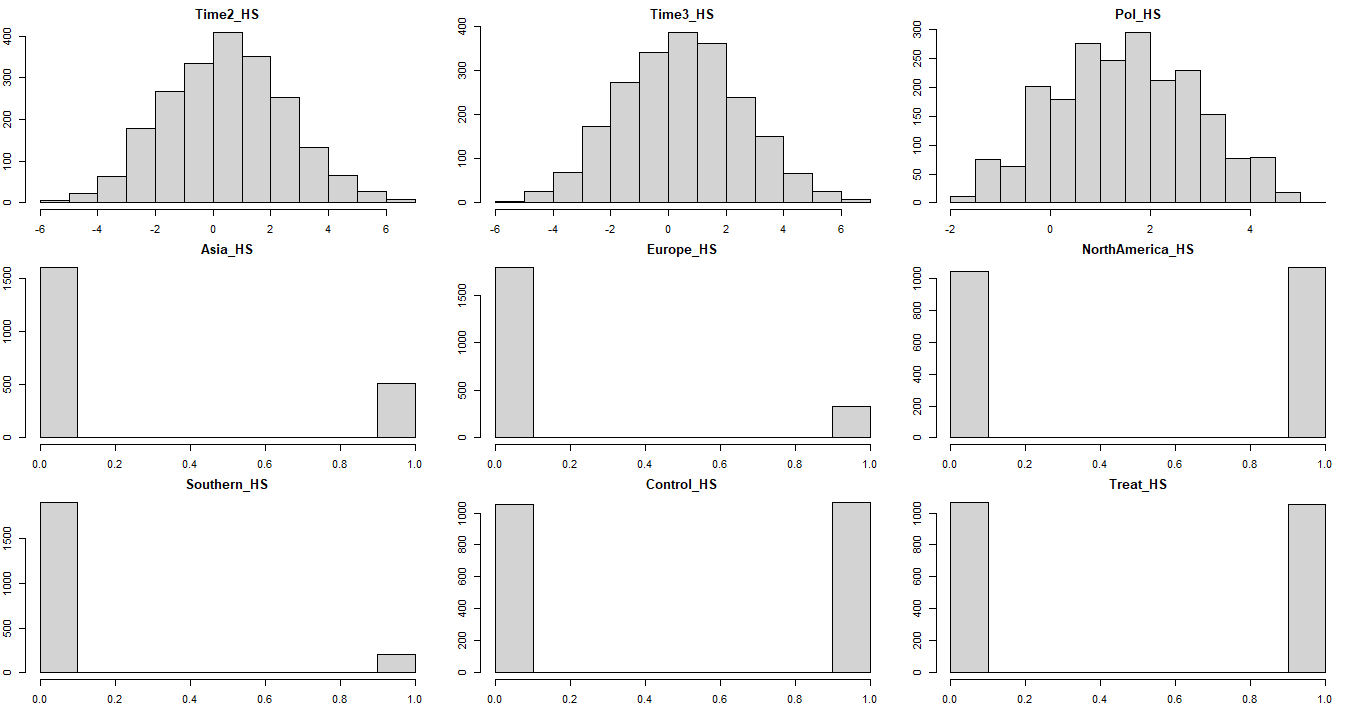
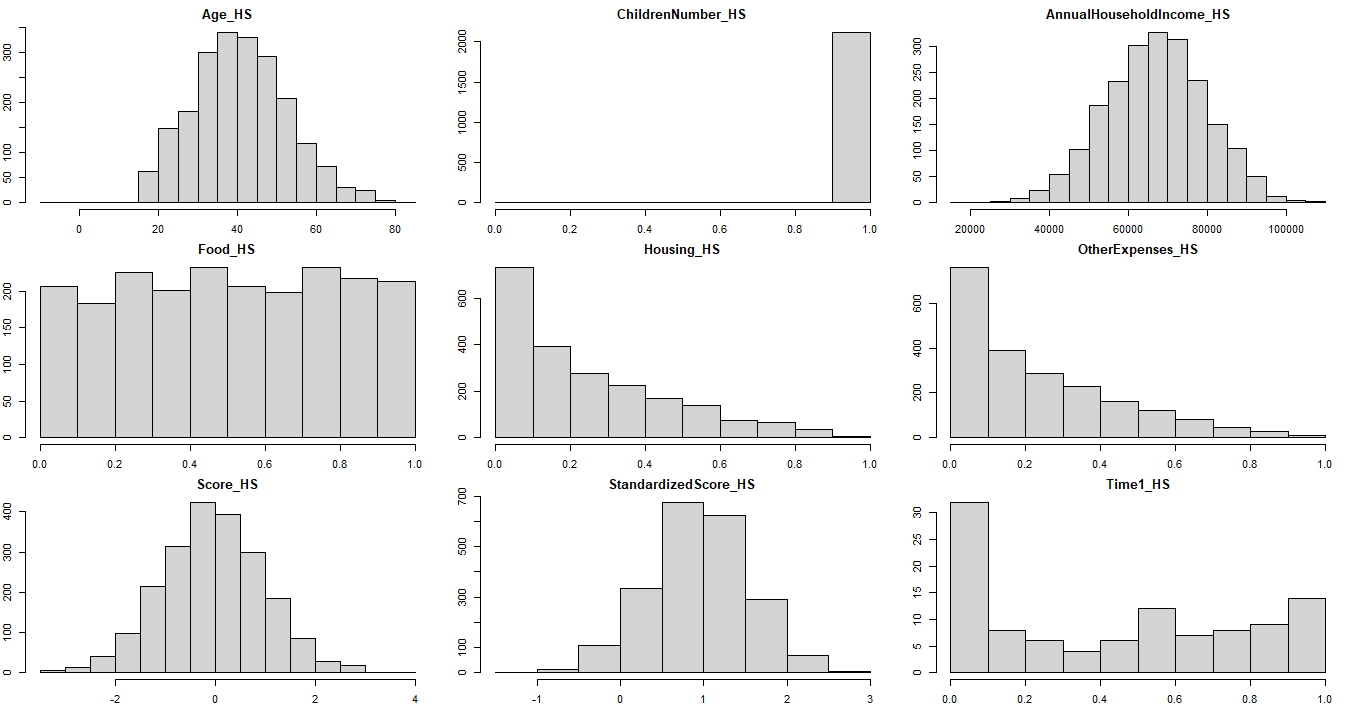
**Descriptive Data Analysis**

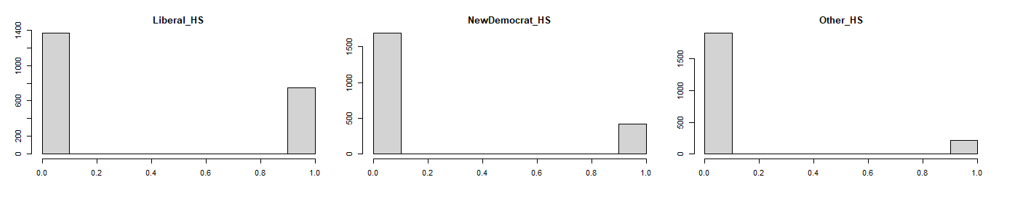




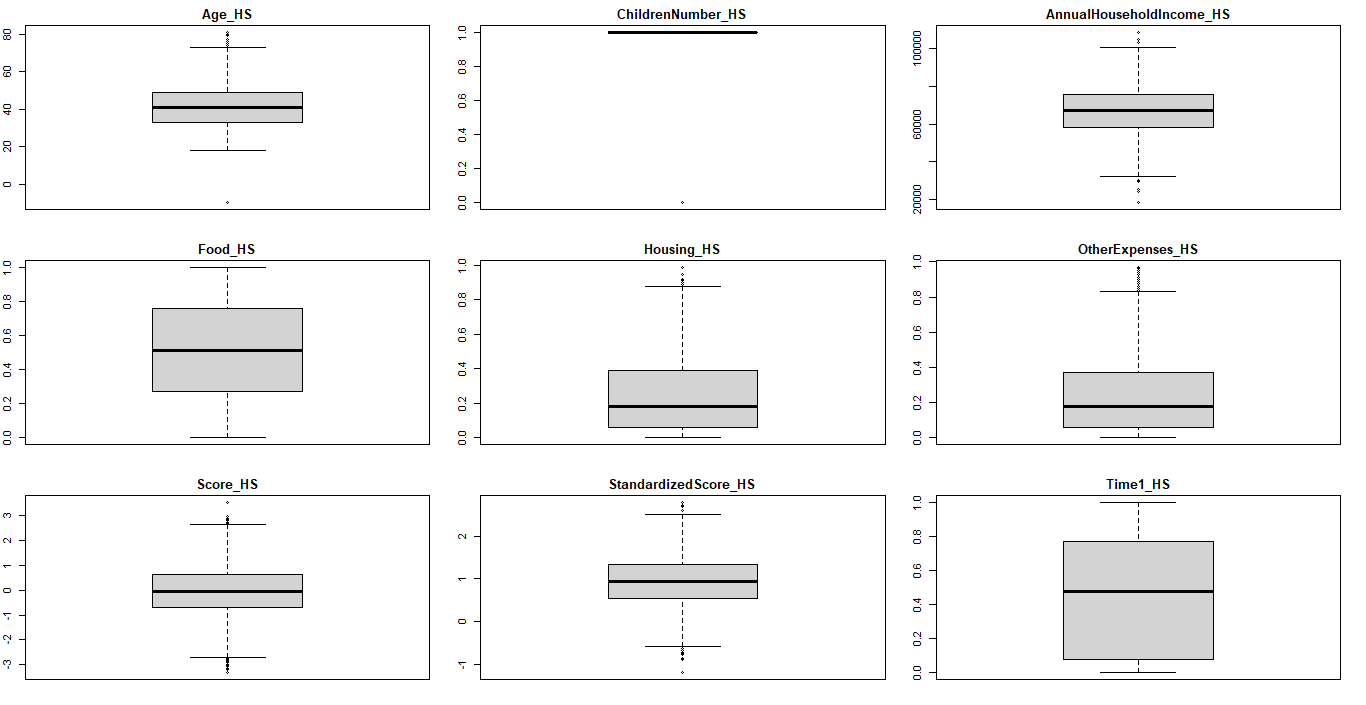


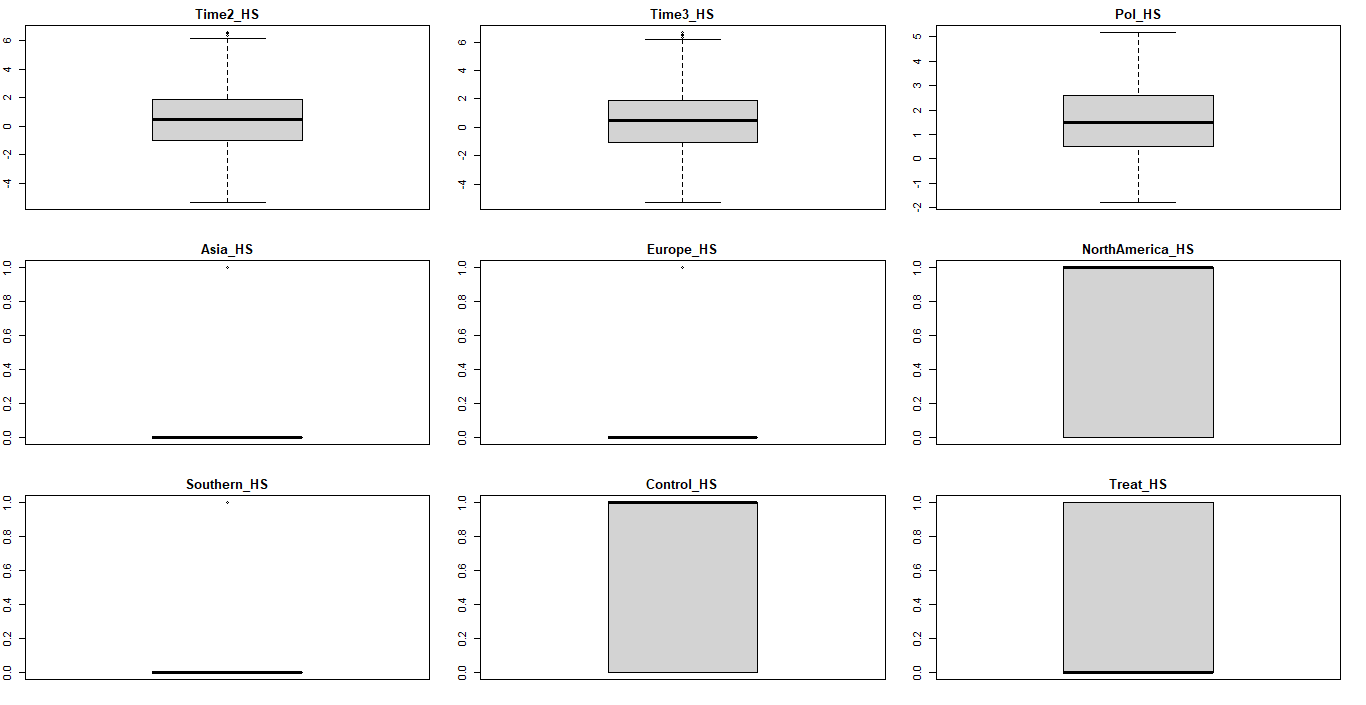
From the summary statistics we conclude that the transformation of nation worked properly. Also, all of the data look reasonable. Political awareness has tightly packed data.

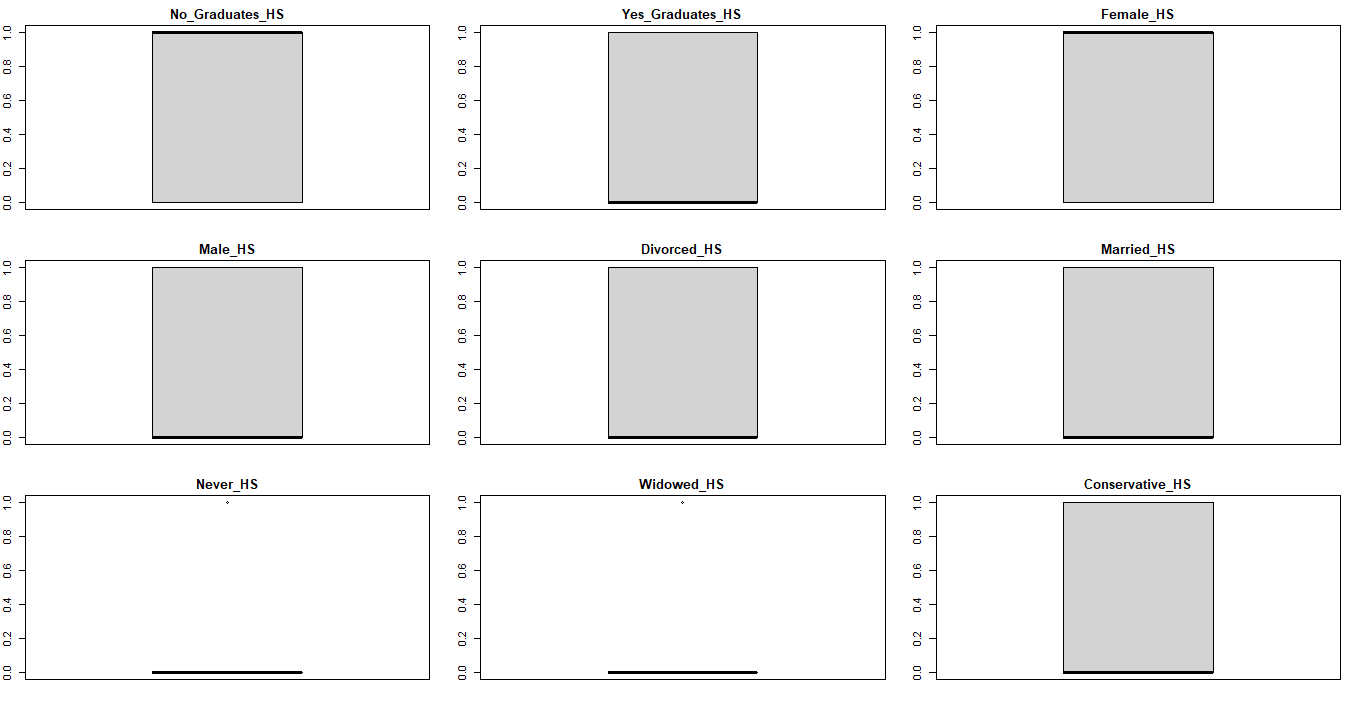
****

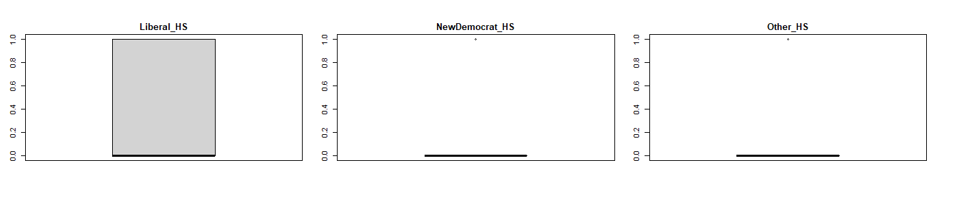


The histogram represents the fitted normal distribution for: Age\_HS, AnnualHouseholdIncome\_HS, Score\_HS, StandardizedScore\_HS. It is normally distributed and graph is symmetric about the mean that indicates graph is normality. Normal Probability plot is graphical technique for normality testing.

****

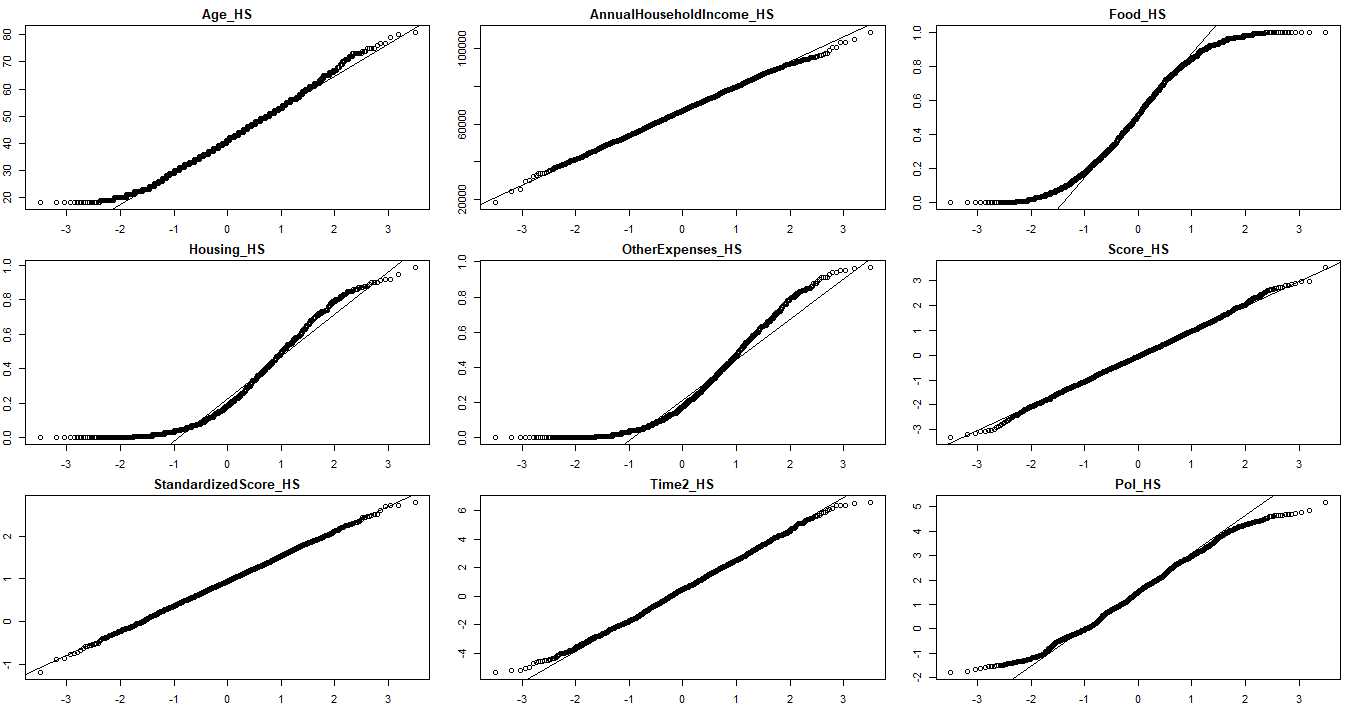
****

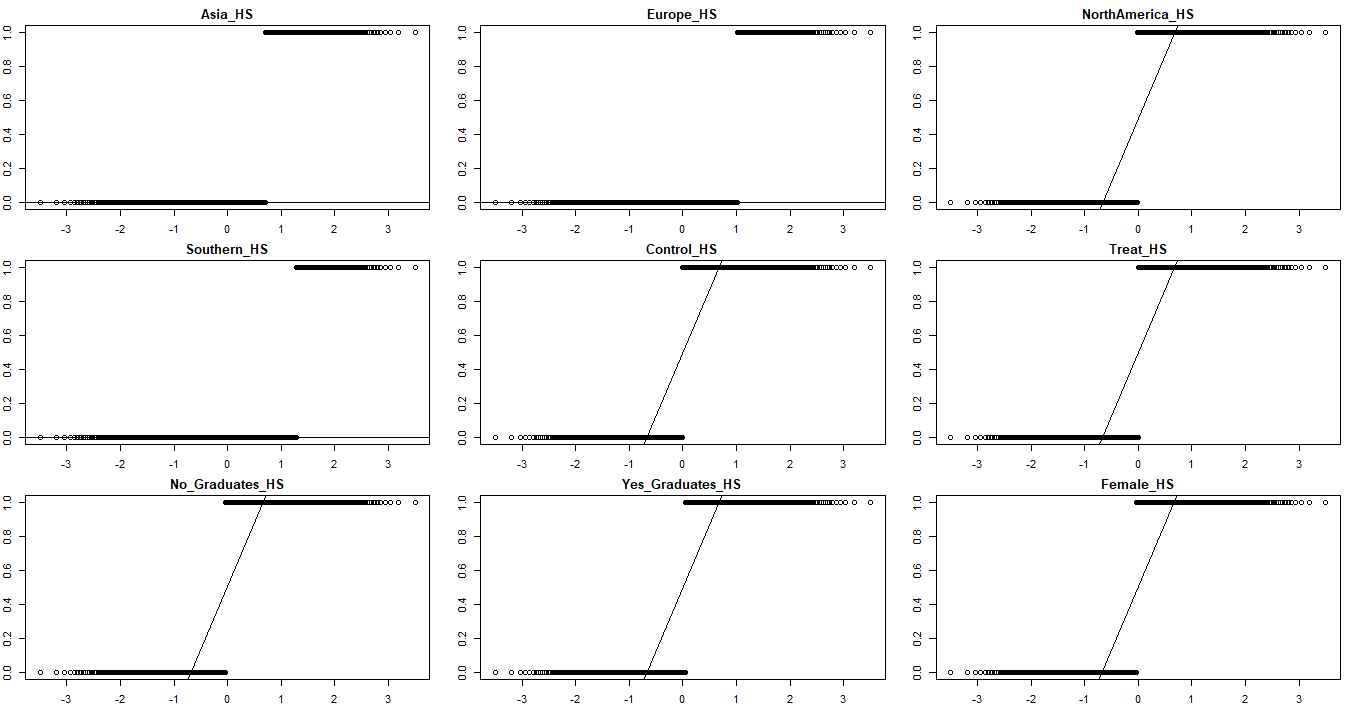


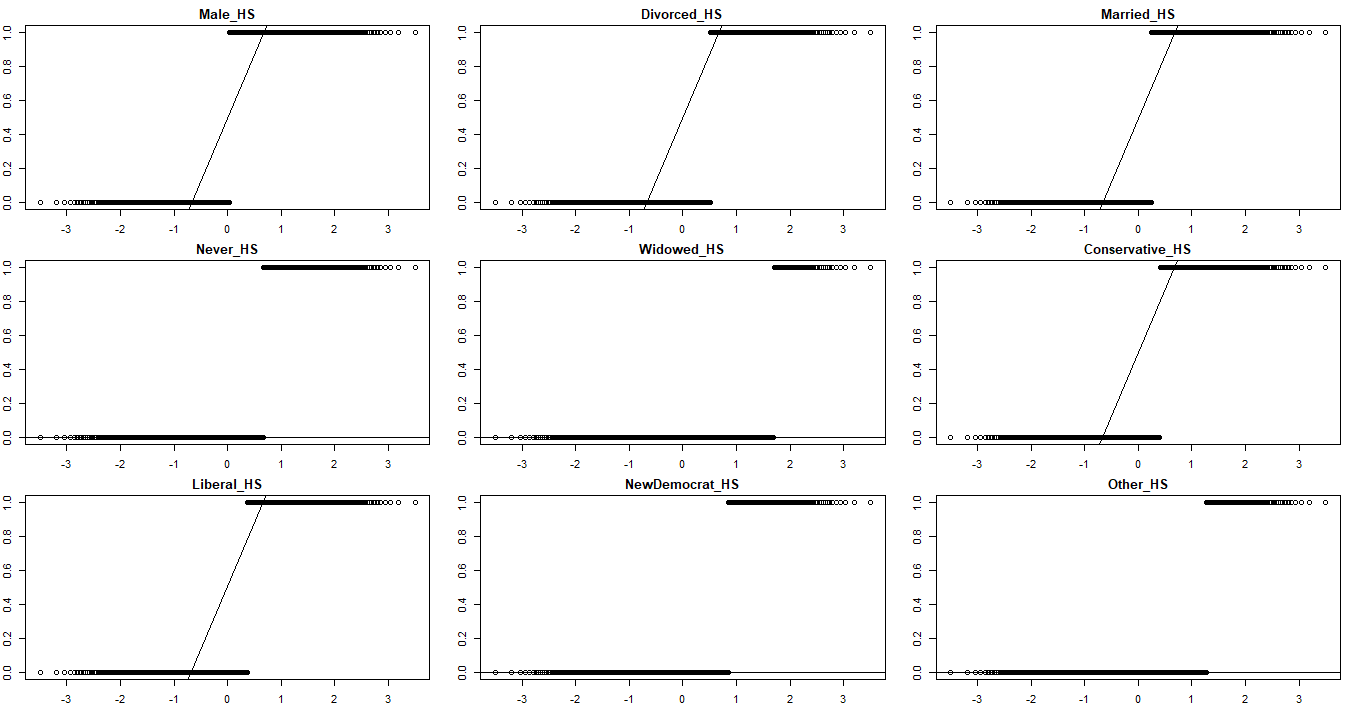


Boxplot represents the visual distribution of the quantitate data. Outlier observed: Age\_HS, Housing\_HS, Time\_HS, StandardizedScore\_HS, OtherExpenses\_HS , Score\_HS, they have 1.5 interquartile ranges away from Q1 or Q3 percentiles. Outlier helps to figure out the central tendency, sense of variability, range of depression. For example in Housing\_HS, range of depression is from 0 to 1 symptoms of depression which is pretty high; non-outliers: 0 to 0.8; most data follows under lower part which means most of the users have low to no expenses, it indicates we have a distribution with the positive skew and it tails off towards higher score.

**Exploratory Data Analysis**

****

****

****

|  |  |  |
| --- | --- | --- |
|  | **statistic** | **p.value** |
| Age\_HS | 0.98923 | 1.71E-11 |
| AnnualHouseholdIncome\_HS | 0.99897 | 0.26355 |
| Food\_HS | 0.9573 | 1.86E-24 |
| Housing\_HS | 0.89778 | 2.05E-35 |
| OtherExpenses\_HS | 0.89357 | 5.67E-36 |
| Score\_HS | 0.99923 | 0.53475 |
| StandardizedScore\_HS | 0.99967 | 0.98824 |
| Time2\_HS | 0.9982 | 0.019055 |
| Pol\_HS | 0.99129 | 5.73E-10 |
| Asia\_HS | 0.52999 | 1.06E-59 |
| Europe\_HS | 0.4325 | 3.74E-63 |
| NorthAmerica\_HS | 0.63656 | 3.78E-55 |
| Southern\_HS | 0.34039 | 5.56E-66 |
| Control\_HS | 0.63659 | 3.80E-55 |
| Treat\_HS | 0.63659 | 3.80E-55 |
| No\_Graduates\_HS | 0.63613 | 3.61E-55 |
| Yes\_Graduates\_HS | 0.63613 | 3.61E-55 |
| Female\_HS | 0.63635 | 3.70E-55 |
| Male\_HS | 0.63635 | 3.70E-55 |
| Divorced\_HS | 0.57959 | 1.06E-57 |
| Married\_HS | 0.62183 | 7.71E-56 |
| Never\_HS | 0.53783 | 2.14E-59 |
| Widowed\_HS | 0.20848 | 1.71E-69 |
| Conservative\_HS | 0.59978 | 7.83E-57 |
| Liberal\_HS | 0.60525 | 1.37E-56 |
| NewDemocrat\_HS | 0.48993 | 3.47E-61 |
| Other\_HS | 0.34591 | 8.03E-66 |

Threshold p-value = 0.05; According to Shapiro Wilk Test:

1. If p-value >0.05 then it implies that the distribution of the data are not significantly different from normal distribution and it can be considered as normality; statistically significant. Example: Score\_HS; StandardizedScore\_HS
2. If p-value <0.05 then we do not consider as normality; not statistically significant. Example: Yellow highlighted p-value of the variables

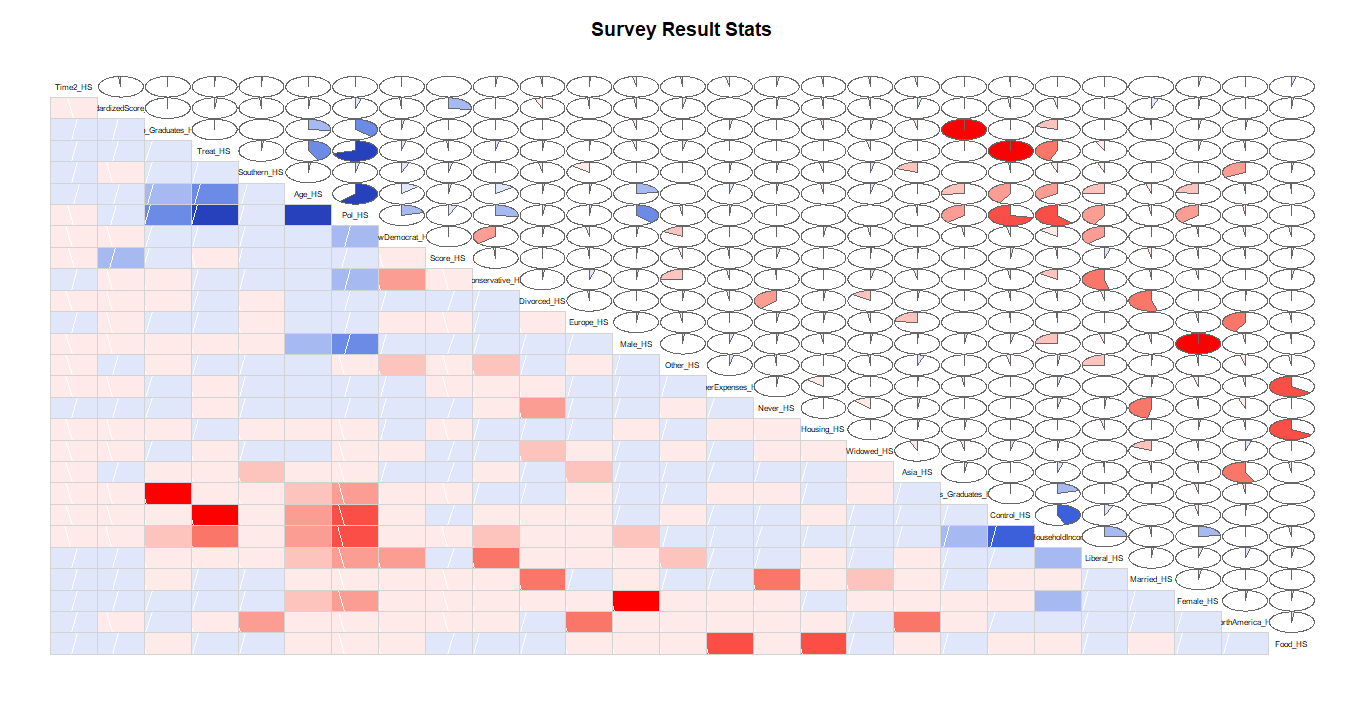
QQ Plot Graph:

1. Normal QQ Plot: Age\_HS, AnnualHouseholdIncome\_HS this is showing that distribution of data against the expected normal distribution and symmetric and no skew which indicates that the mean= median
2. Bimodel Plot: Food\_HS, reduces the number of outlier and it’s thinner tail than normal distribution
3. Left tailed Plot: Housing\_HS, OtherExpenses\_HS this is showing that distribution of data that is left-skewed and negative exponential distribution(negative skew)

All the dummy variable are not normally distributed because of 0 and 1 values.





****

**Notes from Correlation:**

1. **Strong Correlation:** 0.70≤|r|<1.00
2. Pol\_HS and Treat\_HS: 0.73
3. Pol\_HS and Control\_HS: 0.73
4. **Moderate Linear Relationship:** 0.50≤|r|<0.70
5. Pol\_HS and Age\_HS
6. **Weak Linear Relationship:** 0.25≤|r|<0.50
7. Age\_HS and No\_Graduate\_HS(0.26)
8. Pol\_HS and Conservative\_HS(0.26)
9. **No linear Relationship:** 0.00≤|r|<0.25
10. Score\_HS and StandarizedScore\_HS (0.23)
11. AnnualHouseholdIncome\_HS and Female\_HS (0.24)
12. Age\_HS and Male\_HS (0.23)

**Question 5: Simple Linear Regression**

**1. Create a simple linear regression model using Pol as the dependent variable and age as the independent.**

Call:

lm(formula = surveyResultNum\_HS$Pol ~ surveyResultNum\_HS$Age)

Residuals:

Min 1Q Median 3Q Max

-3.068 -0.776 -0.020 0.762 3.582

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.51023 0.08688 -17.4 <2e-16 \*\*\*

surveyResultNum\_HS$Age 0.07288 0.00202 36.1 <2e-16 \*\*\*

---

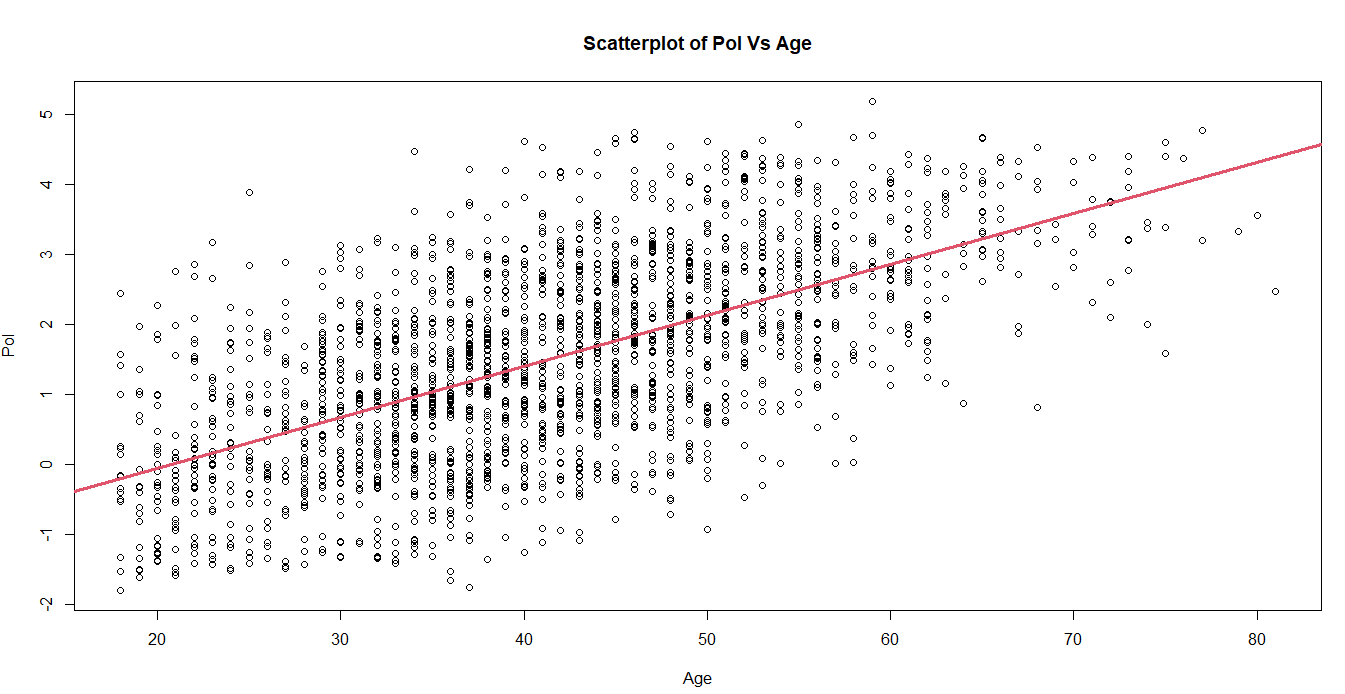
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.1 on 2117 degrees of freedom

Multiple R-squared: 0.381, Adjusted R-squared: 0.381

F-statistic: 1.3e+03 on 1 and 2117 DF, p-value: <2e-16

**Create a scatter plot of the two variables and overlay the regression line.**



**2. Create a simple linear regression model using Pol as the dependent variable and income as the independent.**

Call:

lm(formula = surveyResultNum\_HS$Pol ~ surveyResultNum\_HS$AnnualHouseholdIncome\_HS)

Coefficients:

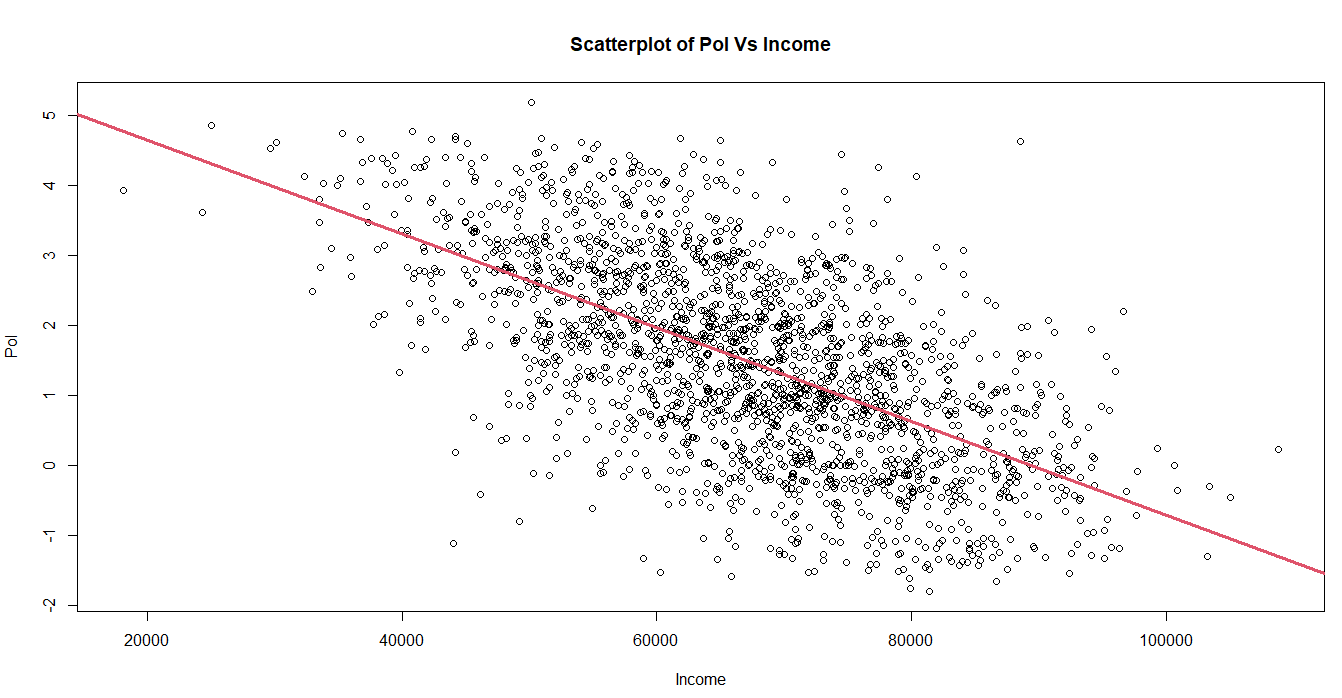
(Intercept)

5.991859

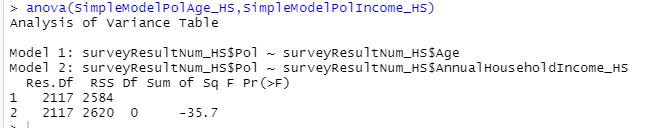
surveyResultNum\_HS$AnnualHouseholdIncome\_HS

-0.000067

**Create a scatter plot of the two variables and overlay the regression line.**



3. **Compare the models. Which model is superior? Why?**



* Df value of both the models are same
* RSS(Residual Sum of Squares): Mode1 1 < Model 2, the lower the RSS value the better regression model can explain the data; whereas the higher value of RSS indicates the poorly combined data set.
* Sum of square for Model 2: -35.7 which indicates that data does not fluctuate much from mean value

On the basis of RSS value and Sum of Square, I choose **Model 1 (Pol VS Age).**

**Models**

**Model 1: All Variables included**

1. F-Stat- Comparing the model and the error (Overall model is statistical significant)

* Bigger F score is better for the model; Numerator: 19; Denominator: 2099
* Compared the p value of the model to the significance level to access the null hypothesis
* P-Value = <2e-16; F-Stat= 2.72e+03 which means the p-value is extremely close to 0 but not 0. Alternative hypothesis: true probability of success is not equal to 0.5; p-value of F-stat is less than 0.05.

2. R-Squared value:

* Value=0.9601 which indicates that it takes into the account of how many exploratory variables are in the model
* 96.01% of the exploratory variables are in the model are elaborated.

3. Residuals (Residual Standard Error):

* It is an average error of the model.
* Our all variable model is giving us an average error of 0.279 when predicting individual model
* Adding variables to the model will increase the value of Residuals

4. Significant variables:

* While considering the t-values, variables that are below 0.05 are highly significant variables **StandardizedScore\_HS, Age\_HS, AnnualHouseholdIncome\_HS, StandardizedScore\_HS, Conservative\_HS, NewDemocrat\_HS, Liberal\_HS**

5. Variable Co-Efficients:

* By looking at the co-efficient(Intercept Value/Estimates)Negative correlation of variable :AnnualHouseholdIncome with Pol

\*Yellow highlights are the NA data because of singularities

Call:

lm(formula = Pol\_HS ~ Age\_HS + AnnualHouseholdIncome\_HS + Food\_HS + Housing\_HS + OtherExpenses\_HS + Score\_HS + StandardizedScore\_HS + Time2\_HS + Asia\_HS + Europe\_HS + NorthAmerica\_HS + Southern\_HS + Control\_HS + Treat\_HS + No\_Graduates\_HS + Yes\_Graduates\_HS + Female\_HS + Male\_HS + Divorced\_HS + Married\_HS + Never\_HS +Widowed\_HS + Conservative\_HS + Liberal\_HS + NewDemocrat\_HS + Other\_HS, data = surveyResultNum\_HS, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-0.9552 -0.1764 -0.0079 0.1767 1.0655

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients: (7 not defined because of singularities)** |  |  |  |  |
|  | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| **(Intercept)** | 2.278479 | 0.066867 | 34.07 | < 2e-16 \*\*\* |
| **Age\_HS** | 0.006513 | 0.000636 | 10.23 | < 2e-16 \*\*\* |
| **AnnualHouseholdIncome\_HS** | -4.6E-06 | 5.93E-07 | -7.74 | 1.6e-14 \*\*\* |
| **Food\_HS** | 0.003094 | 0.02844 | 0.11 | 0.91339 |
| **Housing\_HS** | 0.005646 | 0.036996 | 0.15 | 0.87871 |
| **OtherExpenses\_HS** | NA | NA | NA | NA |
| **Score\_HS** | 0.105292 | 0.006174 | 17.05 | < 2e-16 \*\*\* |
| **StandardizedScore\_HS** | 0.039473 | 0.010824 | 3.65 | 0.00027 \*\*\* |
| **Time2\_HS** | -0.00219 | 0.002923 | -0.75 | 0.45376 |
| **Asia\_HS** | 0.007926 | 0.023001 | 0.34 | 0.73042 |
| **Europe\_HS** | -0.00627 | 0.024748 | -0.25 | 0.80007 |
| **NorthAmerica\_HS** | -0.00145 | 0.021128 | -0.07 | 0.94534 |
| **Southern\_HS** | NA | NA | NA | NA |
| **Control\_HS** | -1.83187 | 0.015193 | -120.58 | < 2e-16 \*\*\* |
| **Treat\_HS** | NA | NA | NA | NA |
| **No\_Graduates\_HS** | 0.97132 | 0.013128 | 73.99 | < 2e-16 \*\*\* |
| **Yes\_Graduates\_HS** | NA | NA | NA | NA |
| **Female\_HS** | -1.03083 | 0.013244 | -77.83 | < 2e-16 \*\*\* |
| **Male\_HS** | NA | NA | NA | NA |
| **Divorced\_HS** | -0.01327 | 0.030758 | -0.43 | 0.66627 |
| **Married\_HS** | -0.00739 | 0.03031 | -0.24 | 0.80735 |
| **Never\_HS** | -0.00227 | 0.031203 | -0.07 | 0.94198 |
| **Widowed\_HS** | NA | NA | NA | NA |
| **Conservative\_HS** | 0.643547 | 0.021851 | 29.45 | < 2e-16 \*\*\* |
| **Liberal\_HS** | -0.47352 | 0.021763 | -21.76 | < 2e-16 \*\*\* |
| **NewDemocrat\_HS** | 0.660261 | 0.023571 | 28.01 | < 2e-16 \*\*\* |
| **Other\_HS** | NA | NA | NA | NA |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.279 on 2099 degrees of freedom

Multiple R-squared: 0.961, Adjusted R-squared: 0.961

F-statistic: 2.72e+03 on 19 and 2099 DF, p-value: <2e-16

**Model 2: Backward Selection**

Eliminated the variables that are less significant in backward selection; currently we have 10 variables whereas in Full baseline selection model, we had 26 variables; AIC works: the lower the number, the better result we get. **Last AIC drop: AIC=-5409.7**

1. F-Stat- Comparing the model and the error (Overall model is statistical significant)

* Bigger F score is better for the model; Numerator: 10; Denominator: 2108
* Compared the p value of the model to the significance level to access the null hypothesis
* P-Value = <2e-16; F-Stat= 5.18e+03 which means the p-value is extremely close to 0 but not 0. Alternative hypothesis: true probability of success is not equal to 0.5; p-value of F-stat is less than 0.05.

2. R-Squared value:

* Value=0.9601 which indicates that it takes into the account of how many exploratory variables are in the model
* 96.01% of the exploratory variables are in the model are elaborated.

3. Residuals (Residual Standard Error):

* It is an average error of the model.
* Our all variable model is giving us an average error of 0.278 when predicting individual model
* Adding variables to the model will increase the value of Residuals

4. Significant variables:

* Almost all of them because we have dropped the insignificant variables in backward selection

5. Variable Co-Efficient:

* By looking at the co-efficient(Intercept Value/Estimates)Negative correlation of variable :AnnualHouseholdIncome and Control\_HS with Pol

Call:

lm(formula = Pol\_HS ~ Age\_HS + AnnualHouseholdIncome\_HS + Score\_HS +

StandardizedScore\_HS + Control\_HS + No\_Graduates\_HS + Female\_HS +

Conservative\_HS + Liberal\_HS + NewDemocrat\_HS, data = surveyResultNum\_HS,

na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-0.9477 -0.1741 -0.0089 0.1735 1.0737

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 2.271522 | 0.052596 | 43.19 | < 2e-16 \*\*\* |
| Age\_HS | 0.006515 | 0.000634 | 10.28 | < 2e-16 \*\*\* |
| AnnualHouseholdIncome\_HS | -4.6E-06 | 5.9E-07 | -7.73 | 1.6e-14 \*\*\* |
| Score\_HS | 0.105351 | 0.006154 | 17.12 | < 2e-16 \*\*\* |
| StandardizedScore\_HS | 0.040125 | 0.010766 | 3.73 | 0.0002 \*\*\* |
| Control\_HS | -1.83177 | 0.015155 | -120.87 | < 2e-16 \*\*\* |
| No\_Graduates\_HS | 0.971231 | 0.013098 | 74.15 | < 2e-16 \*\*\* |
| Female\_HS | -1.03119 | 0.013199 | -78.13 | < 2e-16 \*\*\* |
| Conservative\_HS | 0.642671 | 0.021762 | 29.53 | < 2e-16 \*\*\* |
| Liberal\_HS | -0.47418 | 0.021675 | -21.88 | < 2e-16 \*\*\* |
| NewDemocrat\_HS | 0.659858 | 0.02351 | 28.07 | < 2e-16 \*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.278 on 2108 degrees of freedom

Multiple R-squared: 0.961, Adjusted R-squared: 0.961

F-statistic: 5.18e+03 on 10 and 2108 DF, p-value: <2e-16

**Model Evaluation: Full Baseline Model**

**Verifying Assumptions**

1. **Independence of Predictors:** The lower the correlation matrix value and less least predictor are the variables:StandardizedScore\_HS;
2. **Influential Value:** While considering the t-values, variables that are below 0.05 are highly significant variables **StandardizedScore\_HS, Age\_HS, AnnualHouseholdIncome\_HS, StandardizedScore\_HS, Conservative\_HS, NewDemocrat\_HS, Liberal\_HS**
3. **Distribution of Error Terms:** It seems data is normally distributed because both the p-value < 0.05.

> shapiro.test(surveyMRes)

Shapiro-Wilk normality test

data: surveyMRes

W = 0.998, p-value = 0.013

> shapiro.test(surveyBSMRes\_HS)

Shapiro-Wilk normality test

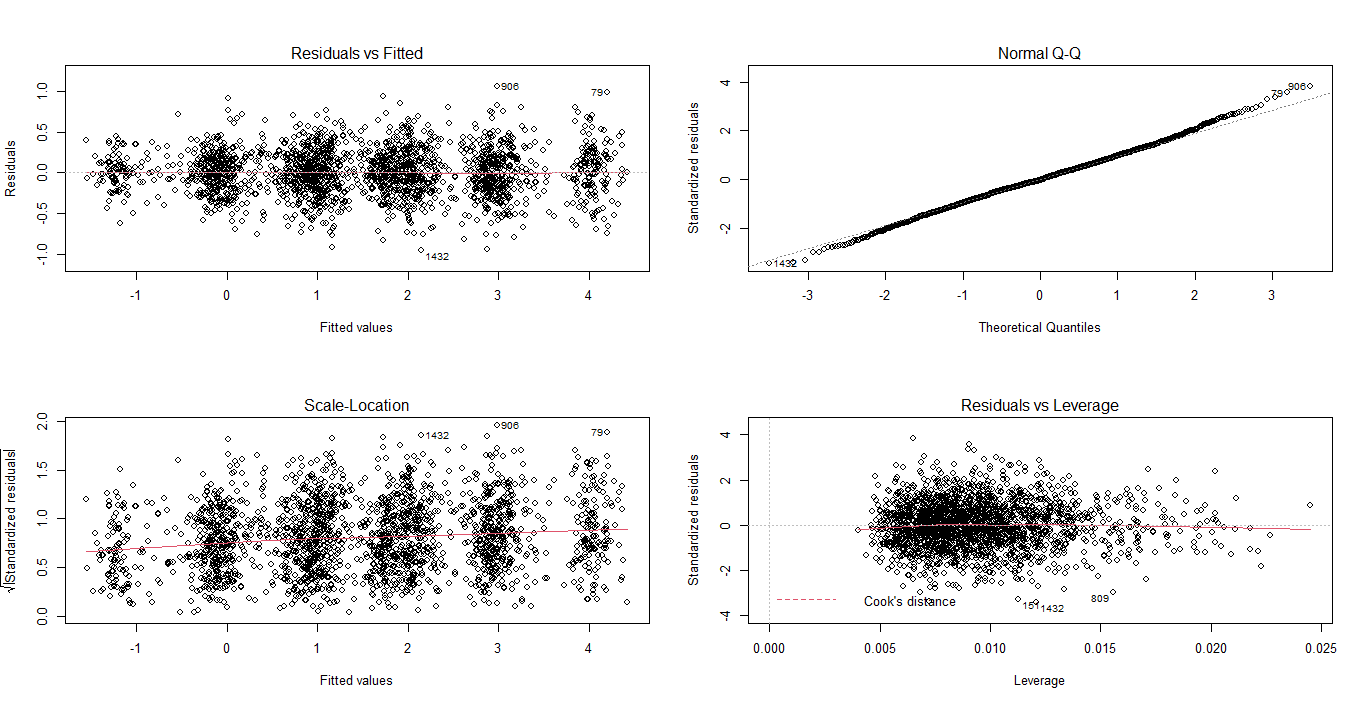
data: surveyBSMRes\_HS

W = 0.998, p-value = 0.011

1. **Non-Auto Correlation and Homoscedasticity:**

* Residual vs Fitted in full baseline model, red line indicates that the hypothesis of Homoscedasticity is fulfilled for a given model. Spread Residuals are almost equal to fitted values; there seems to be no obvious patter of spread residuals. Hence, no autocorrelation.
* Residual vs Leverage and Cook’s distance by using cooks.distance(modelName)- 1432 in Full Baseline Model and it is one of the influential outliers in a set of predictor variables

**Full Baseline Model:**



**Model Evaluation: Backward Baseline Model**

**Verifying Assumptions**

1. **Independence of Predictors:** The lower the correlation matrix value and less least predictor are the variables: **Control\_HS; Liberal\_HS; Female\_HS**
2. **Influential Value:** While considering the t-values, variables that are below 0.05 are highly significant variables **StandardizedScore\_HS, Age\_HS, AnnualHouseholdIncome\_HS, StandardizedScore\_HS, Conservative\_HS, NewDemocrat\_HS, Liberal\_HS**
3. **Distribution of Error Terms:** It seems data is normally distributed because both the p-value < 0.05.

> shapiro.test(surveyBSMRes\_HS)

Shapiro-Wilk normality test

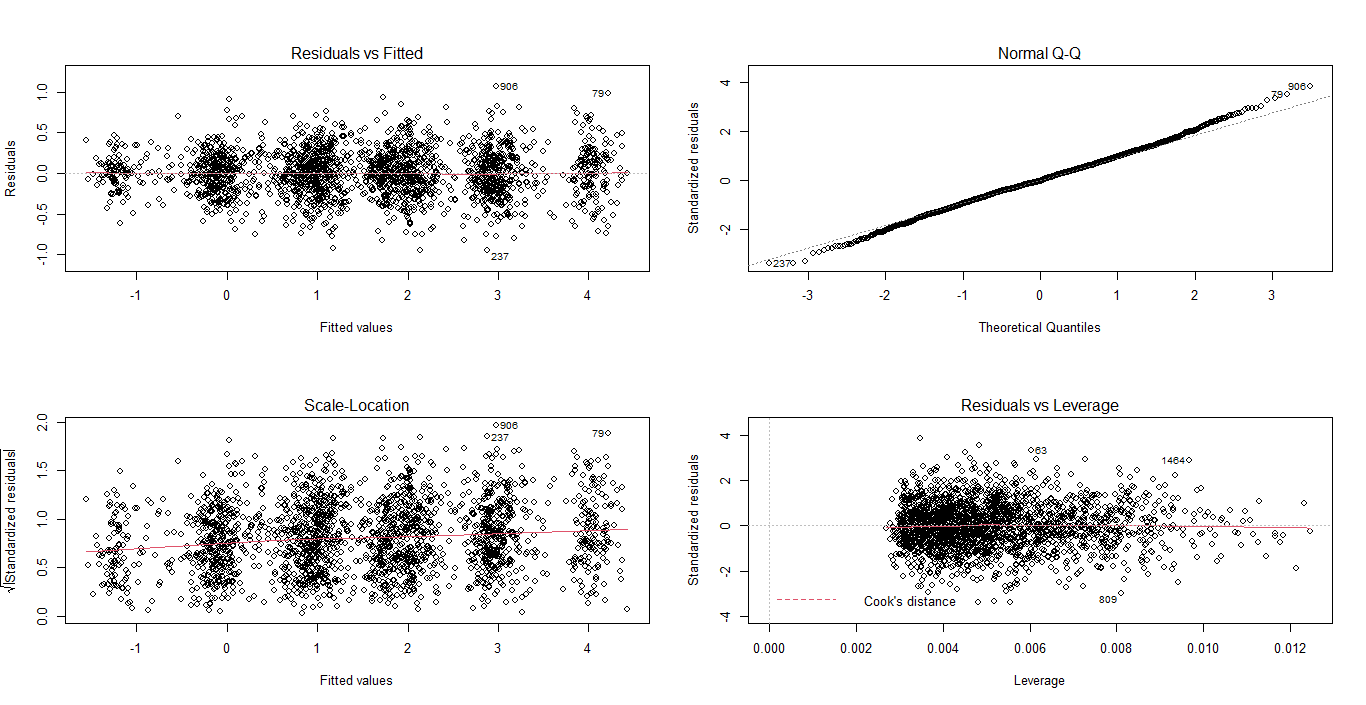
data: surveyBSMRes\_HS

W = 0.998, p-value = 0.011

1. **Non-Auto Correlation and Homoscedasticity:**

* Residual vs Fitted in full baseline model, red line indicates that the hypothesis of Homoscedasticity is fulfilled for a given model. Spread Residuals are almost equal to fitted values; there seems to be no obvious patter of spread residuals. Hence, no autocorrelation.
* Residual vs Leverage and Cook’s distance by using cooks.distance(modelName)- 1464 in Full Baseline Model and it is one of the influential outliers in a set of predictor variables

**Backward Baseline Model:**



**Final Model, Recommendation and Interpretation**

After reviewing the above points, I would go the **Full Baseline Model**:

* .F stat value for the full baseline model is lower than backward baseline model;
* Residual and p value remains the same and we will not consider it

Factors determined which contributed to Political awareness:

Pol = 0.006513\* Age + (-0.000004589)\* Annual Household + -------------------(remaining field)

**APPENDIX 1: Data Transformation**

NationDummy\_HS <- model.matrix(~Nationality\_HS -1, data=surveyDataSet\_HS)

head(NationDummy\_HS)

#Nation Field - Convert to index (Dummy) variable

TreatmentControlGroupDummy\_HS <- model.matrix(~TreatmentControlGroup\_HS -1, data=surveyDataSet\_HS)

head(TreatmentControlGroupDummy\_HS)

#Graduate Field - Convert to index (Dummy) variable

GraduateDummy\_HS <- model.matrix(~Graduates\_HS -1, data=surveyDataSet\_HS)

head(GraduateDummy\_HS)

#Gender Field - Convert to index (Dummy) variable

GenderDummy\_HS <- model.matrix(~Gender\_HS -1, data=surveyDataSet\_HS)

head(GenderDummy\_HS)

#Martial Status Field - Convert to index (Dummy) variable

MaritalStatusDummy\_HS <- model.matrix(~MaritalStatus\_HS -1, data=surveyDataSet\_HS)

head(MaritalStatusDummy\_HS)

#Political Affiliation Field - Convert to index (Dummy) variable

PoliticalAffiliationDummy\_HS <- model.matrix(~PoliticalAffiliation\_HS -1, data=surveyDataSet\_HS)

head(PoliticalAffiliationDummy\_HS)

#Combine the datasets again

surveyDataSet\_HS <- cbind(surveyDataSet\_HS, NationDummy\_HS, TreatmentControlGroupDummy\_HS, GraduateDummy\_HS,GenderDummy\_HS, MaritalStatusDummy\_HS,PoliticalAffiliationDummy\_HS)

head(surveyDataSet\_HS)

str(surveyDataSet\_HS)

summary(surveyDataSet\_HS)