229010645_GY7702_CourseWork2

229010645

2023-01-03

Introduction

This project utilises four data files and one look up table with LSOA or geography code as primary keys. Data manipulation and wrangling will be carried out initially using joins and dplyr library verbs(Stefano,2022; R core team,2022) to obtain a final data frame having 5 variables namely: total electricity consumption, total gas consumption, total energy consumption, total household size, and total household deprivation for all Lower Super Output Areas in Redbridge. The most recent 2020 data for electricity and gas consumption will be used since it's more beneficial for regression.
Visualization, correlation analysis, and presentation of descriptive statistics for these five variables will be conducted, followed by multiple regression analysis of three proposed models(A,B and C) for robustness.

Coding

1. Loading libraries and reading in data files

```
library(tidyverse)
library(knitr)
library(ggplot2)
library(pastecs)
library(psych)
library(GGally)
library(magrittr)
library(lmtest)
library(car)

household_deprivation <- read_csv("census2021-ts011-lsoa.csv")
household_size <- read_csv("census2021-ts017-lsoa.csv")
LSOA_LAD <- read_csv("LSOA_(2011)_to_LSOA_(2021)_to_Local_Authority_District_(2022)_Lookup_for_England_
LSOA_elec_consume <- read_csv("LSOA_domestic_elec_2010-20.csv")
LSOA_gas_consume <- read_csv("LSOA_domestic_gas_2010-20.csv")
```

2. Obtaining a table of all LSOA codes for Redbridge from LSOA_LAD file:

```
LSOA_LAD_Redbridge <- LSOA_LAD %>%
  select(LAD22NM,LAD22CD,LSOA21CD) %>%
  filter(LAD22NM=="Redbridge")
```

3. Forming left join between LSOA_LAD_Redbridge with household_size and subsequent left join with household_deprivation

```
join_Redbridge_hhSize_hhDeprivation <- LSOA_LAD_Redbridge %>%
  left_join(household_size,by=c("LSOA21CD"="geography code"))%>% left_join(household_deprivation)
```

4. Extracting Total Electricity consumption values for Redbridge in 2020 from LSOA_elec_consume.

```
Redbridge_elec_2020 <- LSOA_elec_consume %>%
  filter(`LAD name`=="Redbridge",Year==2020) %>%
  select(`LSOA code`,`Total electricity consumption (kWh)`,`LAD name`)
```

5. Extracting Total Gas consumption values for Redbridge in 2020 from LSOA_gas_consume

```
Redbridge_gas_2020 <- LSOA_gas_consume %>%
  filter(`LAD name`=="Redbridge",Year==2020) %>%
  select(`LSOA code`,`Total gas consumption (kWh)`,`LAD name`)
```

6. Forming left join between Redbridge_gas_2020 and Redbridge_elec_2020

```
Redbridge_elec_gas_2020 <- Redbridge_elec_2020 %>%
  left_join(Redbridge_gas_2020)
```

7. Forming inner join between join_Redbridge_hhSize_hhDeprivation which has 164 rows and Redbridge_elec_gas_2020 which has 161 rows

```
Redbridge_elec_gas_2020_hhDepriv_hhSize <- Redbridge_elec_gas_2020 %>% inner_join(join_Redbridge_hhSize
```

8. Forming final dataframe that will be used for exploratory statistics and multiple regression analysis. Renaming some variables that have very long names. Lastly addition of total electricity consumption and total gas consumption to form a new total energy consumption column

```
final_Redbridge <- Redbridge_elec_gas_2020_hhDepriv_hhSize %>%
  select(`LAD name`,LAD22CD,geography,`Total electricity consumption (kWh)`,`Total gas consumption (kWh
  rename(Total_elec = `Total electricity consumption (kWh)`,Total_gas=`Total gas consumption (kWh)`,Tot
  mutate(Tot_energy=Total_gas+Total_elec)
```

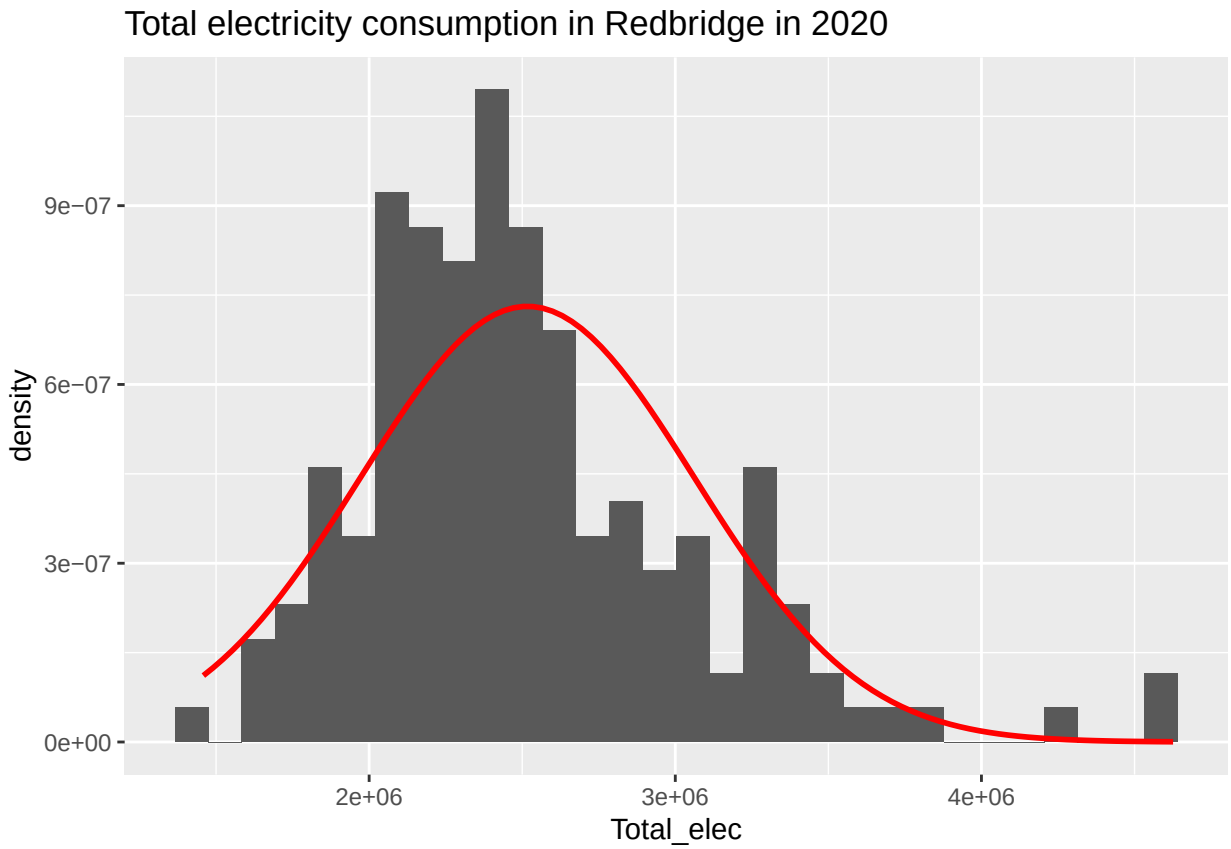
```
final_Redbridge %>% slice_head(n=10) %>% kable()
```

LAD name	LAD22CD	geography	Total_elec	Total_gas	Tot_Hhold_size	Tot_Hhold_dep	Tot_energy
Redbridge	E09000026	Redbridge 001A	3008283	11260360	617	616	14268643
Redbridge	E09000026	Redbridge 001B	3027088	14233485	477	478	17260574
Redbridge	E09000026	Redbridge 001C	3386795	12595797	673	679	15982592

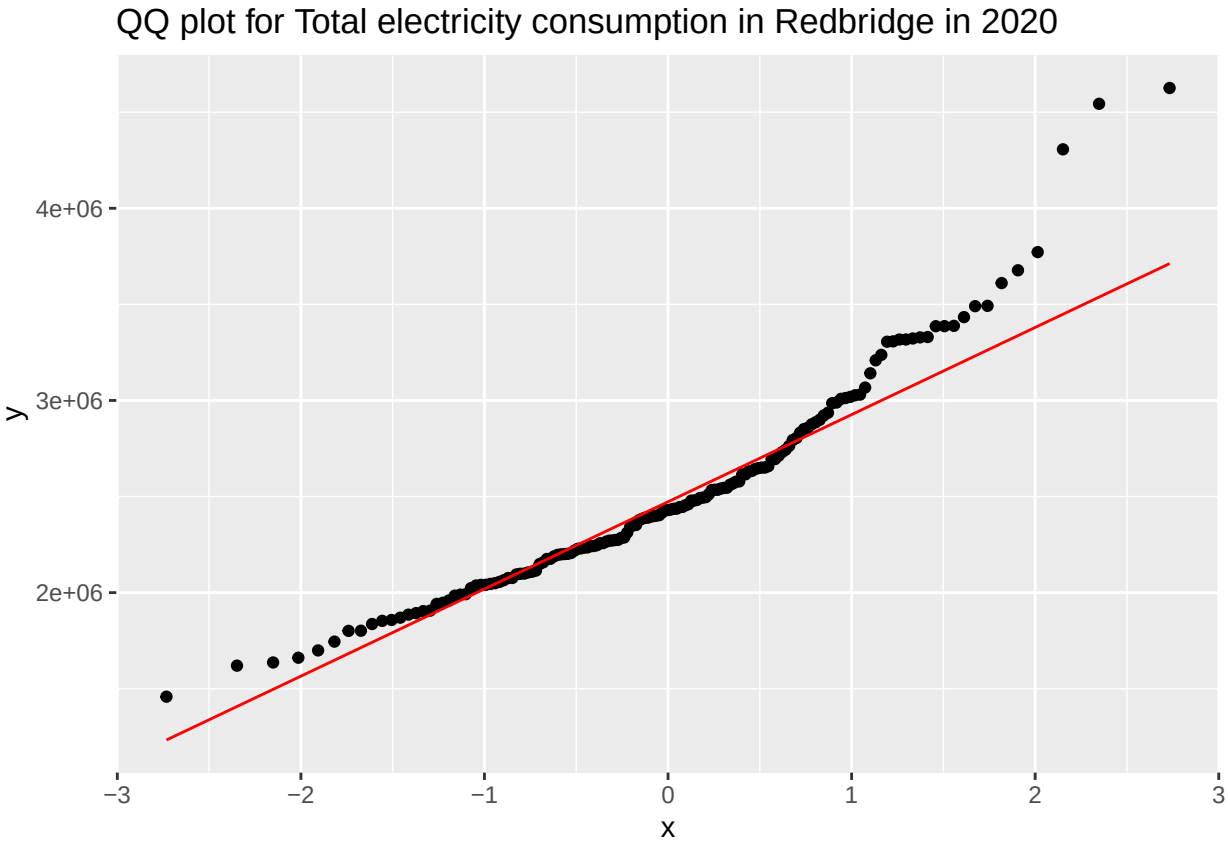
LAD name	LAD22CD	geography	Total_elec	Total_gas	Tot_Hhold_size	Tot_Hhold_dep	Tot_energy
Redbridge	E09000026	Redbridge 001D	2804464	13284670	485	484	16089134
Redbridge	E09000026	Redbridge 001E	3610810	13243433	622	625	16854244
Redbridge	E09000026	Redbridge 001F	3012784	9724629	683	686	12737413
Redbridge	E09000026	Redbridge 001G	3388232	10934128	877	878	14322360
Redbridge	E09000026	Redbridge 002A	2535222	8442133	788	787	10977355
Redbridge	E09000026	Redbridge 002B	2832522	9072917	823	818	11905438
Redbridge	E09000026	Redbridge 002C	2657884	8355014	718	720	11012898

9. Data Visualisation using Histograms,QQ-plots followed by Shapiro-Wilk normality test for Total electricity consumption(Total_elec)

```
final_Redbridge %>%
  ggplot(
    aes(
      x = Total_elec
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge %>%
        filter(!is.na(Total_elec)) %>%
        pull(Total_elec) %>%
        mean(),
      sd =
        final_Redbridge %>%
        filter(!is.na(Total_elec)) %>%
        pull(Total_elec) %>%
        sd()
    ),
    colour = "red",
    size = 1
  ) + ggtitle("Total electricity consumption in Redbridge in 2020")
```




```
final_Redbridge %>%
  ggplot(
    aes(
      sample =
        Total_elec
    )
  ) +
  stat_qq() +
  stat_qq_line(colour='red')+
  ggtitle("QQ plot for Total electricity consumption in Redbridge in 2020 ")
```



Code source for generating histograms and QQ plots:(Stefano,2022; R core team,2022)

```
final_Redbridge %>%
  pull(Total_elec) %>%
  shapiro.test()
```

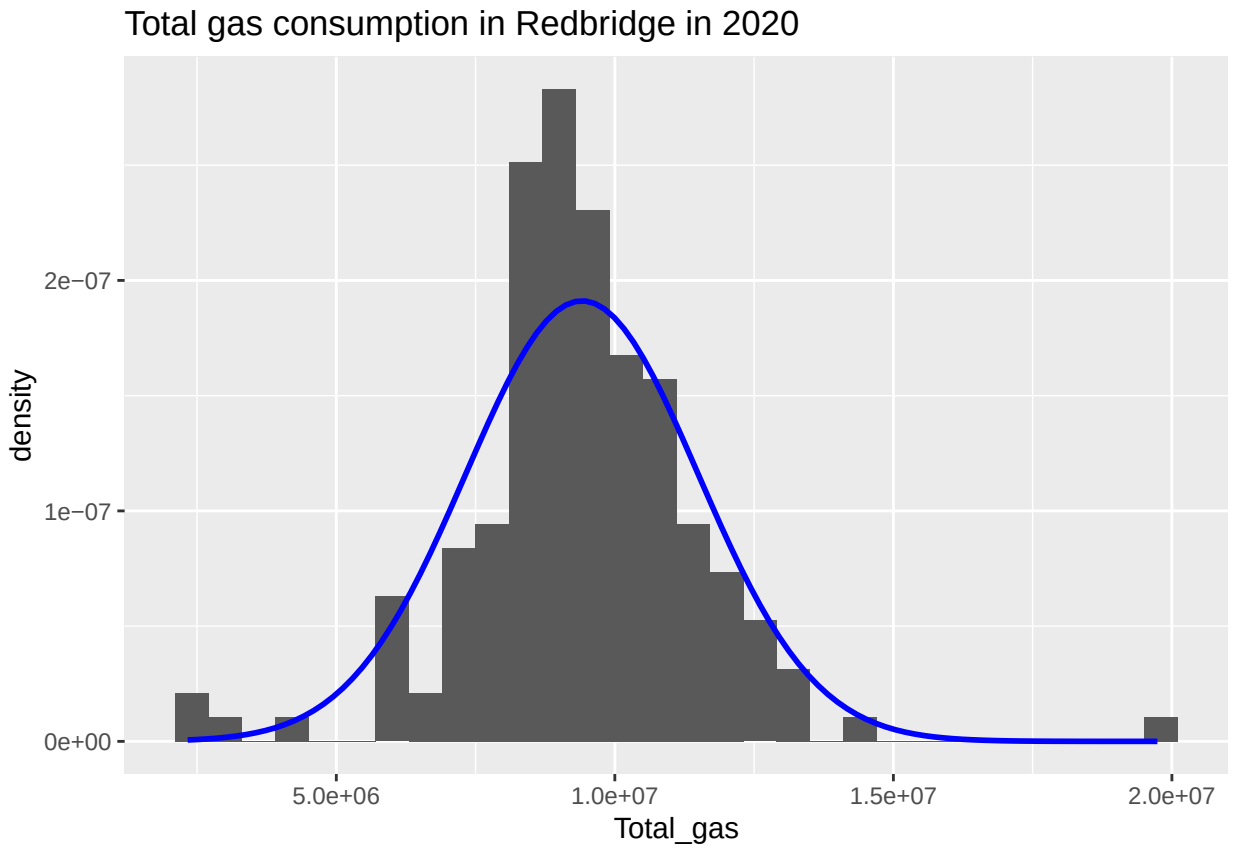
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.93248, p-value = 7.895e-07
```

Conclusion:The histogram and QQ-plot show that the distribution of 'total electricity consumption' does not follow a normal distribution. In addition, the p-value of 7.895e-07 from the Shapiro-Wilk normality test is less than 0.05, hence the variable,'total electricity consumption' is not normally distributed.

10. Data Visualisation using Histograms,QQ-plots followed by Shapiro_Wilk Normality test for Total_gas

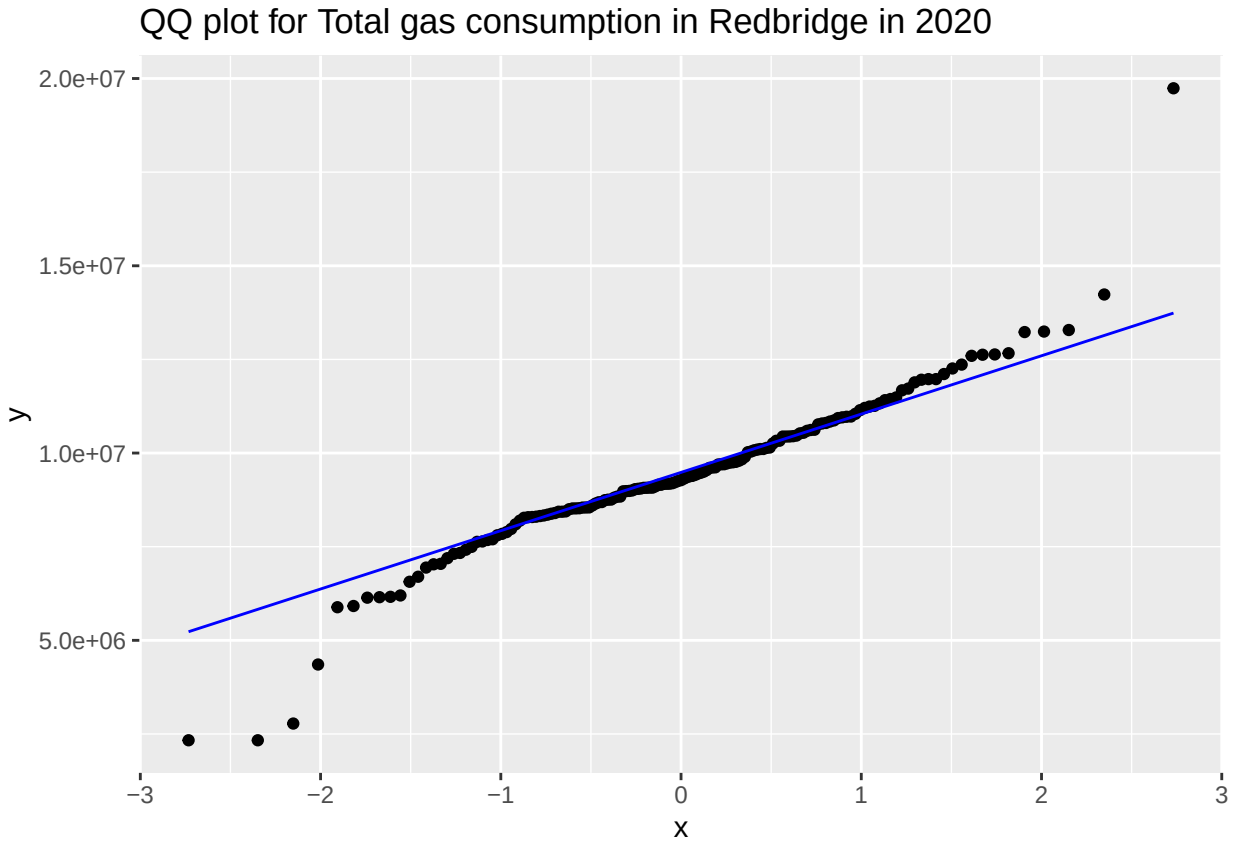
```
final_Redbridge %>%
  ggplot(
    aes(
      x = Total_gas
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
```

```
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge %>%
        filter(!is.na(Total_gas)) %>%
        pull(Total_gas) %>%
        mean(),
      sd =
        final_Redbridge %>%
        filter(!is.na(Total_gas)) %>%
        pull(Total_gas) %>%
        sd()
    ),
    colour = "blue",
    size = 1
  )+ ggtitle("Total gas consumption in Redbridge in 2020")
```




```
final_Redbridge %>%
  ggplot(
    aes(
      sample =
        Total_gas
    )
  ) +
```

```
  stat_qq() +
  stat_qq_line(colour='blue')+
  ggtitle("QQ plot for Total gas consumption in Redbridge in 2020 ")
```




```
final_Redbridge %>%
  pull(Total_gas) %>%
  shapiro.test()
```

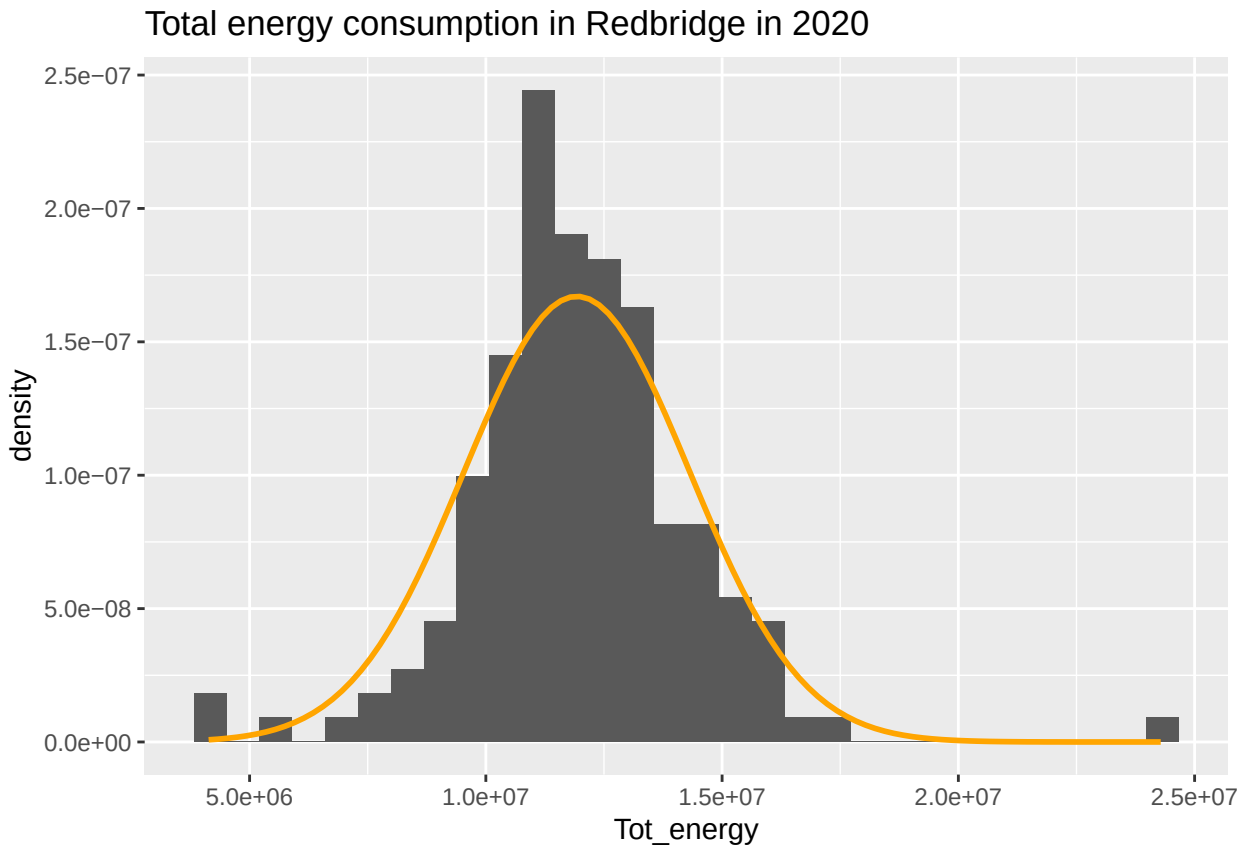
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.93462, p-value = 1.135e-06
```

Conclusion:The histogram and QQ-plot show that the distribution of 'total gas consumption' does not follow a normal distribution. In addition, the p-value of 1.135e-06 from the Shapiro-Wilk normality test is less than 0.05, hence the variable,'total gas consumption' is not normally distributed.

11. Data Visualisation using Histograms,QQ-plots followed by Shapiro-Wilk Normality test for Tot_energy

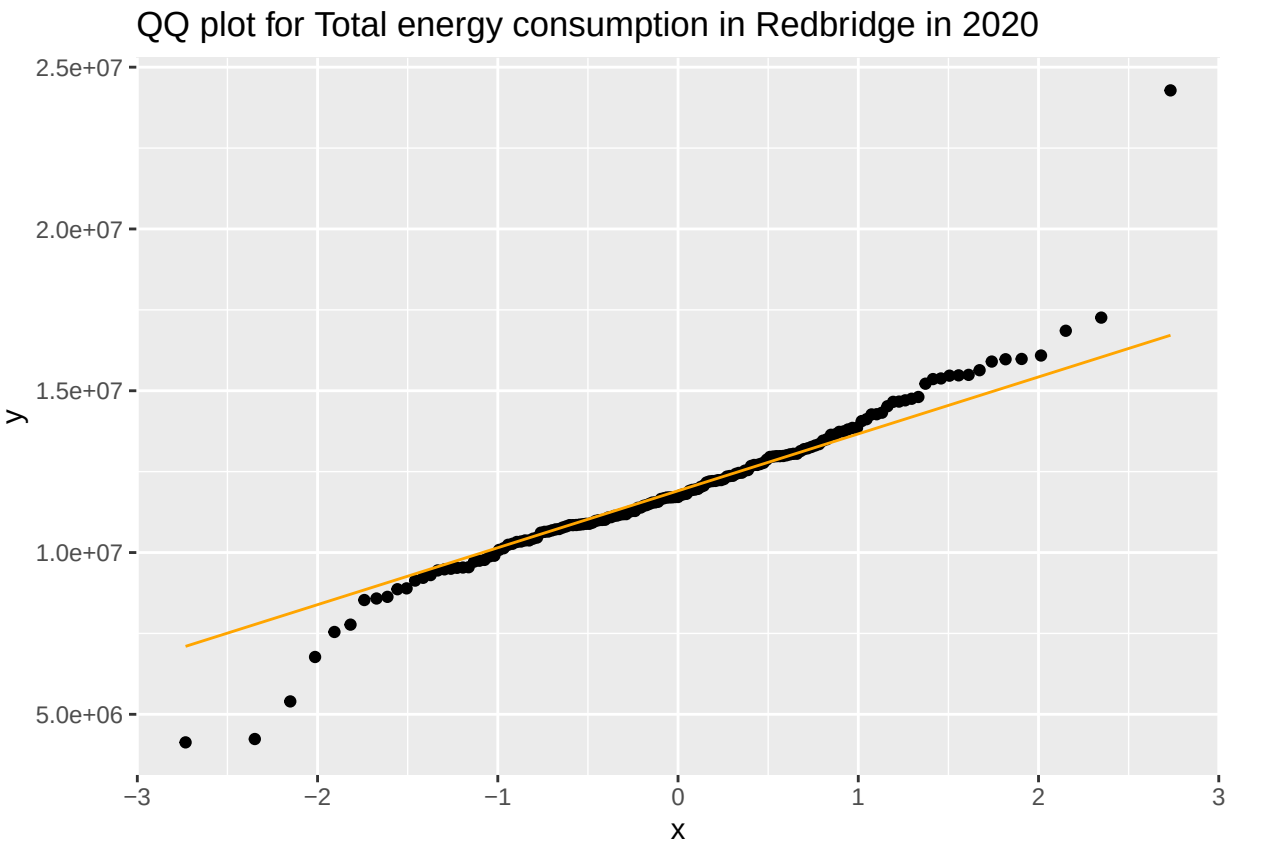
```
final_Redbridge %>%
  ggplot(
    aes(
      x = Tot_energy
    )
  ) +
```

```
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge %>%
        filter(!is.na(Tot_energy)) %>%
        pull(Tot_energy) %>%
        mean(),
      sd =
        final_Redbridge %>%
        filter(!is.na(Tot_energy)) %>%
        pull(Tot_energy) %>%
        sd()
    ),
    colour = "orange",
    size = 1
  ) + ggtitle("Total energy consumption in Redbridge in 2020")
```




```
final_Redbridge %>%
  ggplot(
    aes(
      sample =
```

```
      Tot_energy
    )
  ) +
  stat_qq() +
  stat_qq_line(colour='orange')+
  ggtitle("QQ plot for Total energy consumption in Redbridge in 2020 ")
```




```
final_Redbridge %>%
  pull(Tot_energy) %>%
  shapiro.test()
```

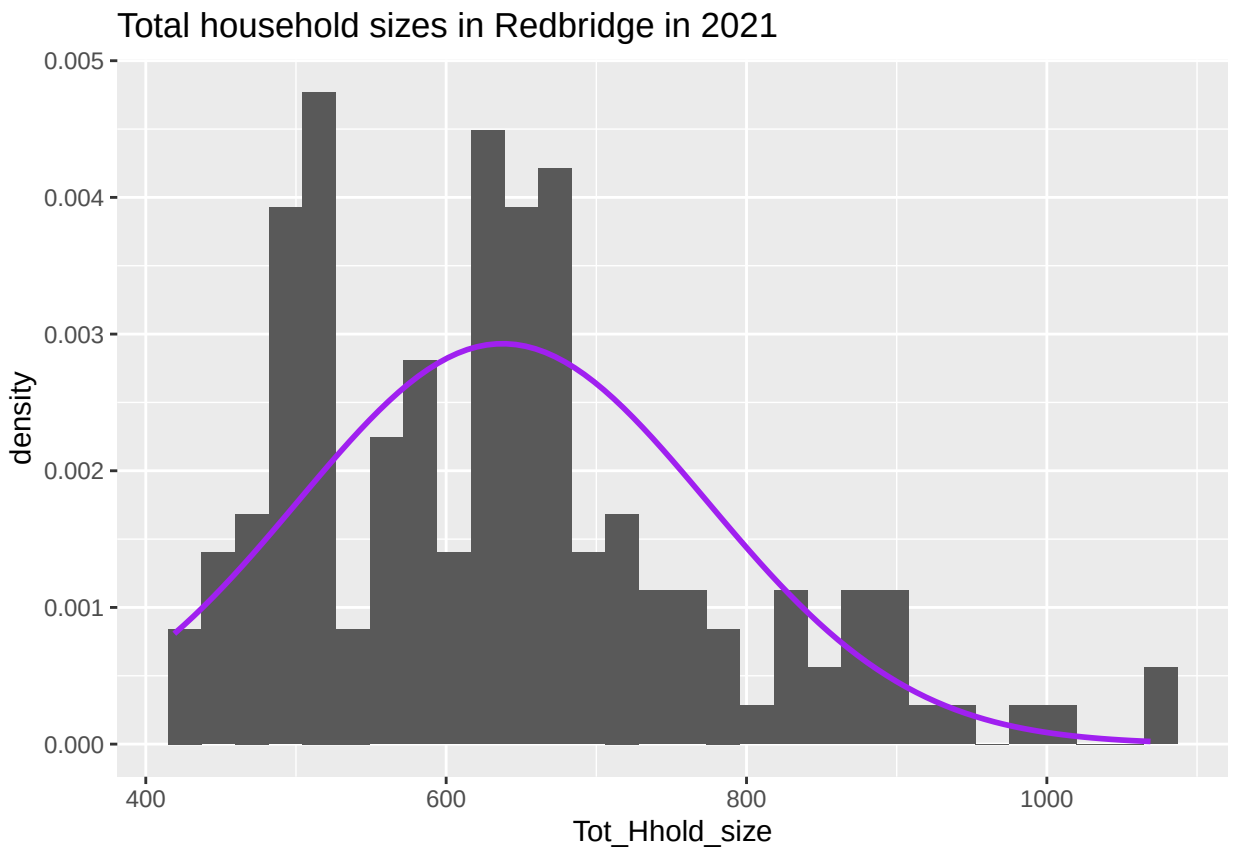
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.94048, p-value = 3.177e-06
```

Conclusion:The histogram and QQ-plot show that the distribution of 'total energy consumption' does not follow a normal distribution. In addition, the p-value of 3.177e-06 from the Shapiro-Wilk normality test is less than 0.05, hence the variable,'total energy consumption' is not normally distributed.

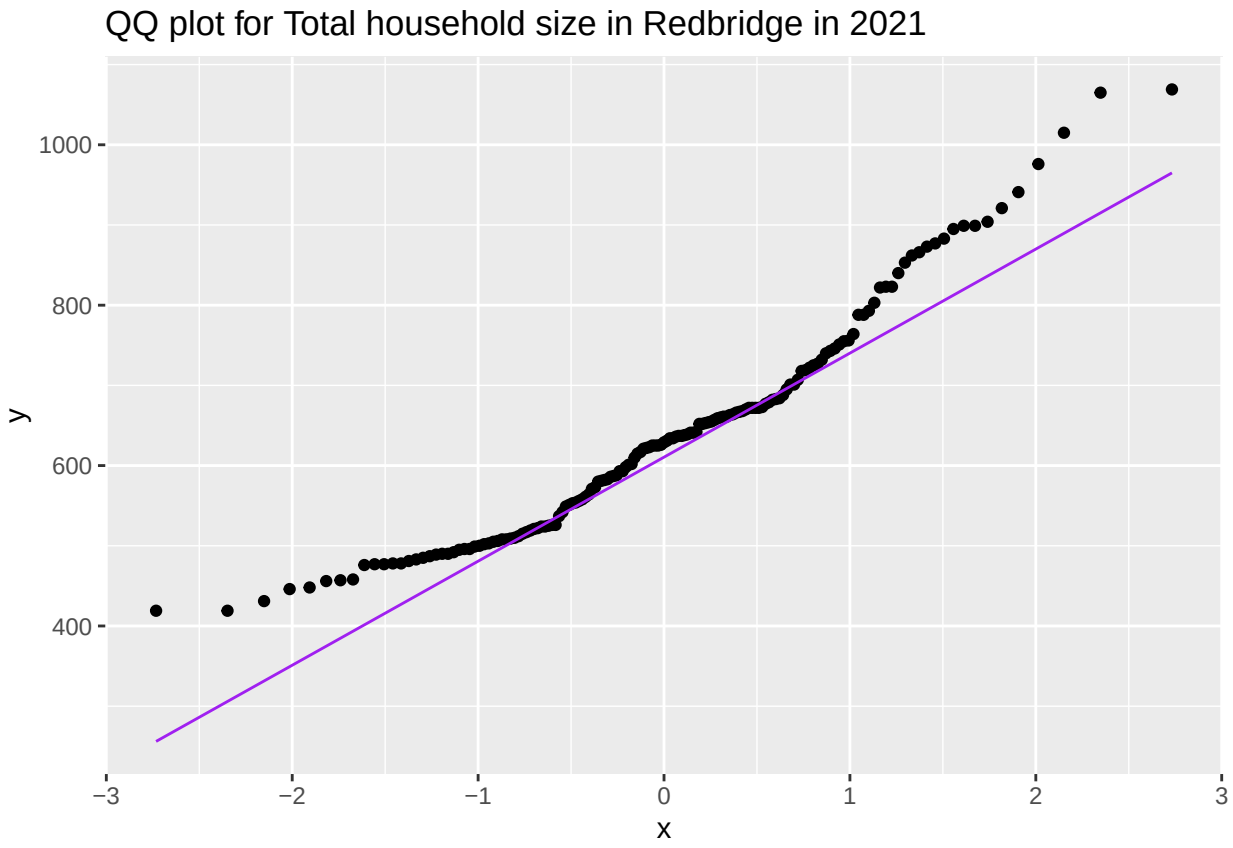
12. Data Visualisation using Histograms,QQ-plots followed by Shapiro-Wilk Normality test for Tot_Hhold_size

```
final_Redbridge %>%
  ggplot(
    aes(
```

```
      x = Tot_Hhold_size
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge %>%
        filter(!is.na(Tot_Hhold_size)) %>%
        pull(Tot_Hhold_size) %>%
        mean(),
      sd =
        final_Redbridge %>%
        filter(!is.na(Tot_Hhold_size)) %>%
        pull(Tot_Hhold_size) %>%
        sd()
    ),
    colour = "purple",
    size = 1
  ) + ggtitle("Total household sizes in Redbridge in 2021")
```



```
final_Redbridge %>%
  ggplot(
    aes(
      sample =
        Tot_Hhold_size
    )
  ) +
  stat_qq() +
  stat_qq_line(colour='purple')+
  ggtitle("QQ plot for Total household size in Redbridge in 2021 ")
```



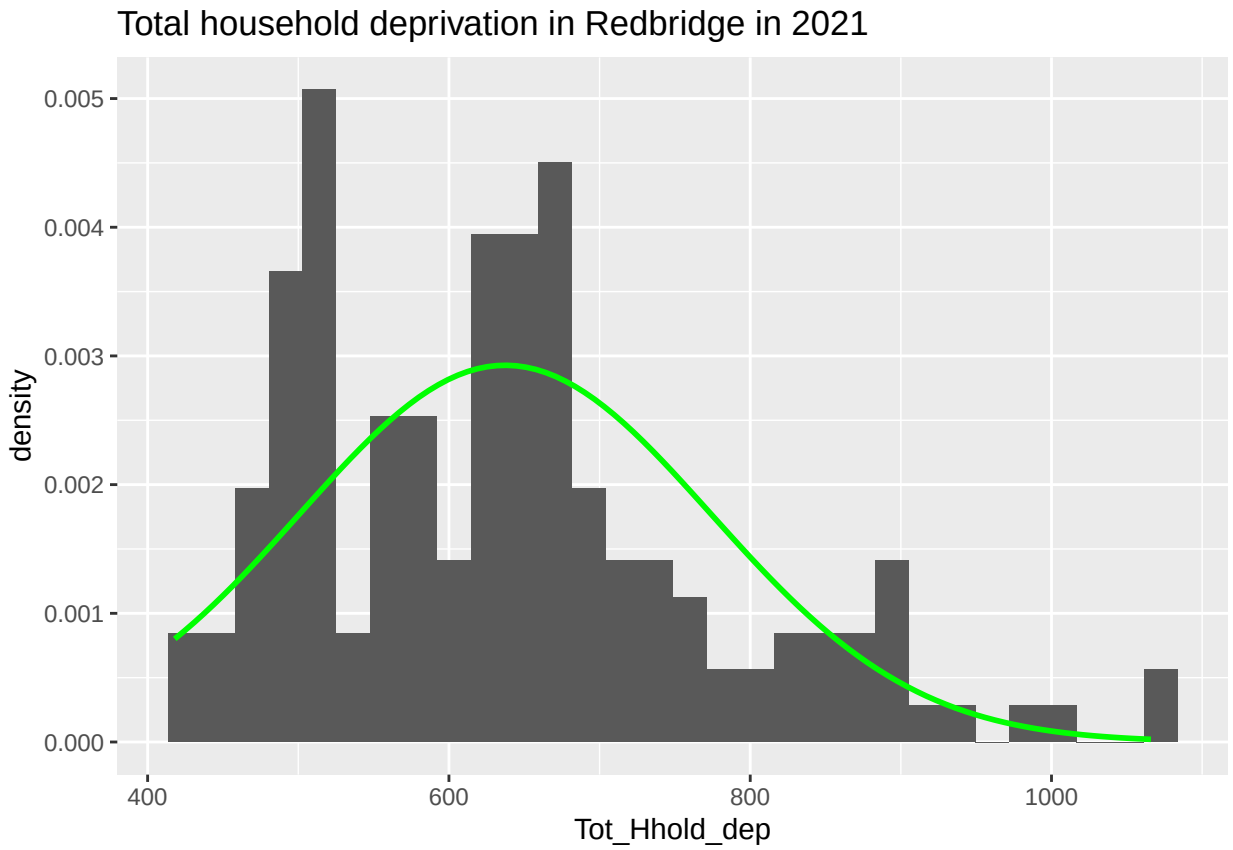

```
final_Redbridge %>%
  pull(Tot_Hhold_size) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.94167, p-value = 3.937e-06
```

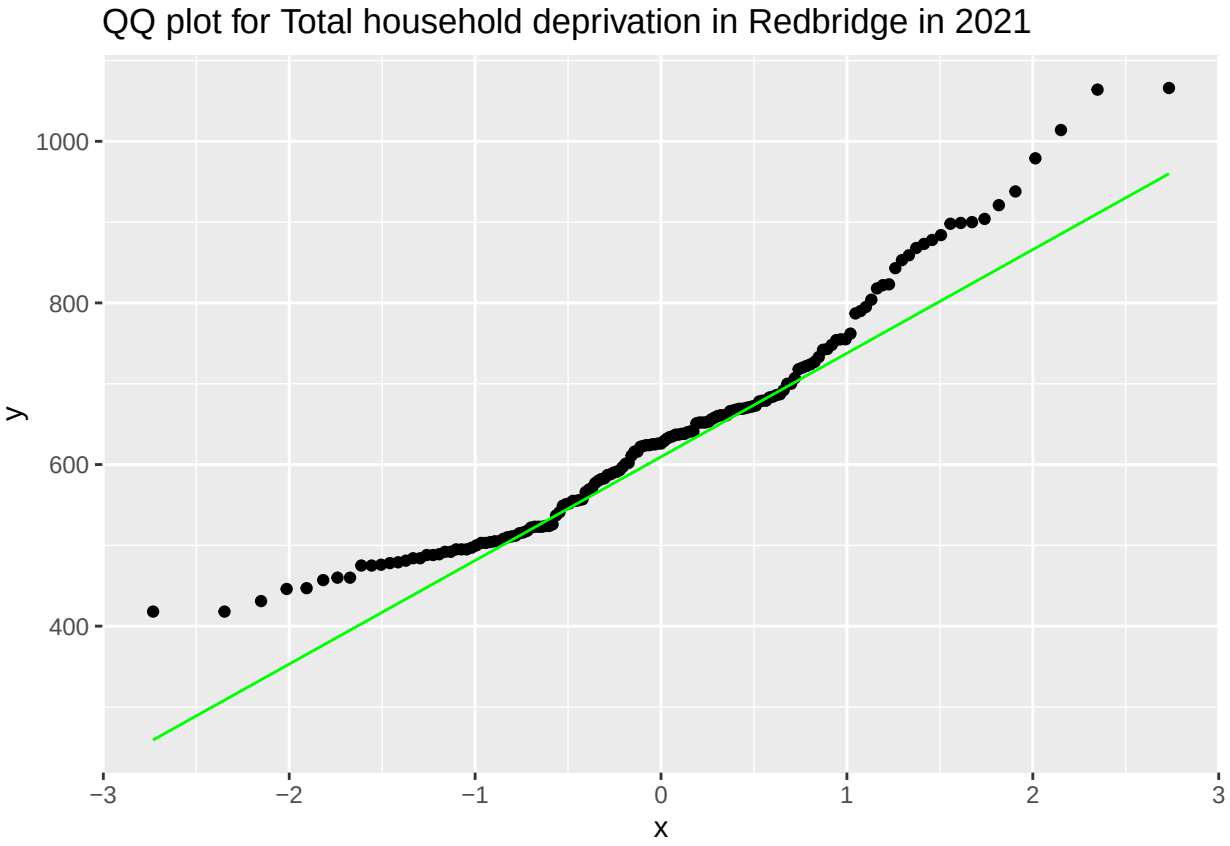
Conclusion:The histogram and QQ-plot show that the distribution of 'total household size' does not follow a normal distribution. In addition, the p-value of 3.937e-06 from the Shapiro-Wilk normality test is less than 0.05, hence the variable,'total household size' is not normally distributed.

13. Data Visualisation using Histograms,QQ-plots followed by Shapiro-Wilk Normality test for Tot_Hhold_dep

```
final_Redbridge %>%
  ggplot(
    aes(
      x = Tot_Hhold_dep
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge %>%
        filter(!is.na(Tot_Hhold_dep)) %>%
        pull(Tot_Hhold_dep) %>%
        mean(),
      sd =
        final_Redbridge %>%
        filter(!is.na(Tot_Hhold_dep)) %>%
        pull(Tot_Hhold_dep) %>%
        sd()
    ),
    colour = "green",
    size = 1
  )+ ggtitle("Total household deprivation in Redbridge in 2021")
```




```
final_Redbridge %>%
  ggplot(
    aes(
      sample =
        Tot_Hhold_dep
    )
  ) +
  stat_qq() +
  stat_qq_line(colour='green')+
  ggtitle("QQ plot for Total household deprivation in Redbridge in 2021 ")
```




```
final_Redbridge %>%
  pull(Tot_Hhold_dep) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.94157, p-value = 3.867e-06
```

Conclusion:The histogram and QQ-plot show that the distribution of 'total household deprivation' does not follow a normal distribution. In addition, the p-value of 3.867e-06 from the Shapiro-Wilk normality test is less than 0.05, hence the variable,'total household deprivation' is not normally distributed.

14. Calculating Descriptive Statistics for Total electricity consumption,Total gas consumption,Total energy consumption,Total household size and deprivation using final_Redbridge

```
final_Redbridge %>%
  select(Total_elec,Total_gas,Tot_energy,Tot_Hhold_size,Tot_Hhold_dep) %>%
  stat.desc() %>%
  kable(digits = c(2,2,2,2,2))
```

	Total_elec	Total_gas	Tot_energy	Tot_Hhold_size	Tot_Hhold_dep
nbr.val	1.590000e+02	1.590000e+02	1.590000e+02	159.00	159.00
nbr.null	0.000000e+00	0.000000e+00	0.000000e+00	0.00	0.00
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.00	0.00

	Total_elec	Total_gas	Tot_energy	Tot_Hhold_size	Tot_Hhold_dep
min	1.458554e+06	2.331585e+06	4.132960e+06	419.00	418.00
max	4.625942e+06	1.973901e+07	2.428267e+07	1069.00	1066.00
range	3.167388e+06	1.740743e+07	2.014971e+07	650.00	648.00
sum	4.001964e+08	1.496042e+09	1.896239e+09	101372.00	101345.00
median	2.429538e+06	9.276043e+06	1.171712e+07	629.00	626.00
mean	2.516959e+06	9.409071e+06	1.192603e+07	637.56	637.39
SE.mean	4.327838e+04	1.655068e+05	1.893934e+05	10.80	10.81
CI.mean.0.95	8.547879e+04	3.268912e+05	3.740695e+05	21.33	21.35
var	2.978099e+11	4.355410e+12	5.703310e+12	18543.92	18573.83
std.dev	5.457196e+05	2.086962e+06	2.388160e+06	136.18	136.29
coef.var	2.200000e-01	2.200000e-01	2.000000e-01	0.21	0.21

```
final_Redbridge %>%
  select(Total_elec,Total_gas,Tot_energy,Tot_Hhold_size,Tot_Hhold_dep) %>%
  stat.desc(basic = FALSE, desc = FALSE, norm = TRUE)%>%
  kable()
```

	Total_elec	Total_gas	Tot_energy	Tot_Hhold_size	Tot_Hhold_dep
skewness	1.1119431	0.1662089	0.4890622	0.8464919	0.8436296
skew.2SE	2.8887979	0.4318062	1.2705703	2.1991628	2.1917268
kurtosis	1.8390668	4.4122881	4.5361685	0.4456669	0.4243303
kurt.2SE	2.4032436	5.7658608	5.9277444	0.5823856	0.5545036
normtest.W	0.9324837	0.9346209	0.9404822	0.9416653	0.9415666
normtest.p	0.0000008	0.0000011	0.0000032	0.0000039	0.0000039

15. Correlation analysis

Since all variables are not normally distributed, we use Spearman's rank correlation coefficient if no ties exist. If ties exist, we shall use Kendall's tau

```
cor.test(final_Redbridge$Total_elec, final_Redbridge$Total_gas, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  final_Redbridge$Total_elec and final_Redbridge$Total_gas
## S = 328460, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5097027
```

```
cor.test(final_Redbridge$Tot_Hhold_size, final_Redbridge$Tot_Hhold_dep, method = "spearman")
```

```
## Warning in cor.test.default(final_Redbridge$Tot_Hhold_size,
## final_Redbridge$Tot_Hhold_dep, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  final_Redbridge$Tot_Hhold_size and final_Redbridge$Tot_Hhold_dep
## S = 239.01, p-value < 2.2e-16
```

```
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9996432
```

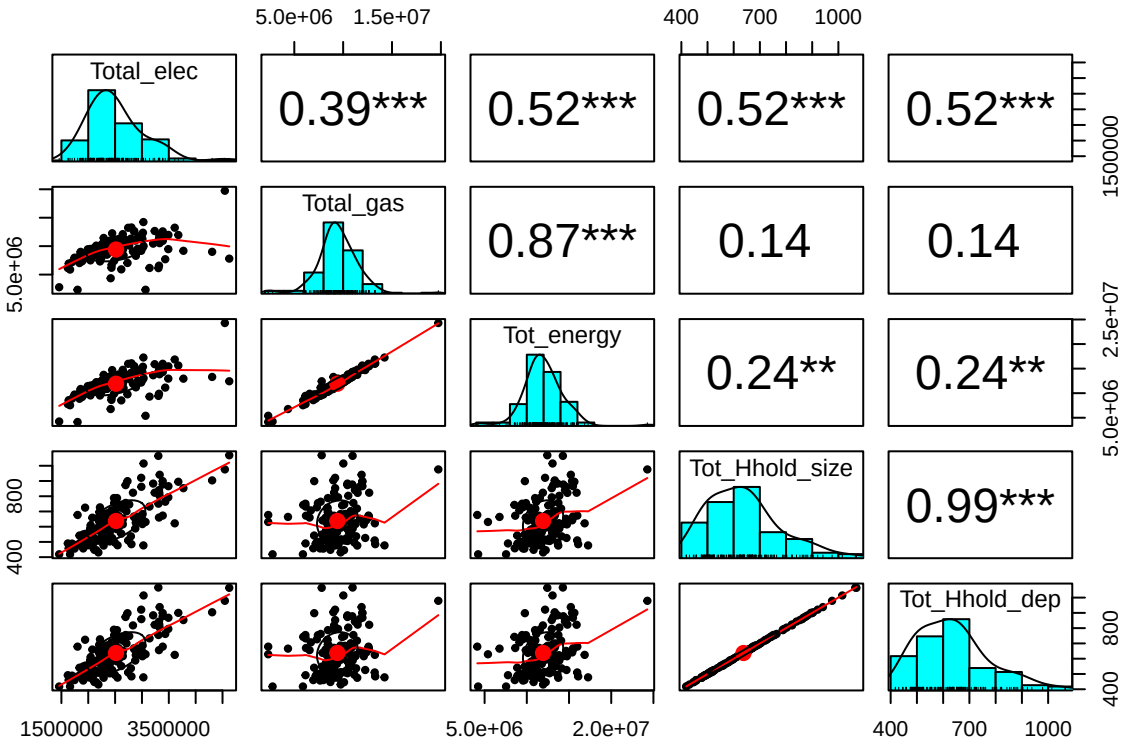
Due to ties in the data for Tot_Hhold_size and Tot_Hhold_dep, switch to using Kendall's tau correlation cofficient

```
cor.test(final_Redbridge$Tot_Hhold_size, final_Redbridge$Tot_Hhold_dep, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  final_Redbridge$Tot_Hhold_size and final_Redbridge$Tot_Hhold_dep
## z = 18.491, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.9906632
```

We can also present a pair panels plot to investigate multicollinearity between the predictor(independent) variables namely;Tot_Hhold_size and Tot_Hhold_dep. In the prior,analysis,all our variables were not normally distributed, hence the pairs panel plot must use the Kendall method.

```
final_Redbridge %>%
  select(Total_elec,Total_gas,Tot_energy,Tot_Hhold_size,Tot_Hhold_dep) %>%
  pairs.panels(method = "kendall",stars = TRUE)
```



Interpretation of pairs panel plot

From the pairs-panel plot above,there is a moderate, positive correlation of 0.39 between total electricity consumption and total gas consumption. There exists a strong,positive correlation of 0.52 between total electricity consumption and total energy consumption. A strong, positive correlation of 0.52 exists between

total electricity consumption and total household size. A strong,positive correlation of 0.52 exists between total electricity consumption and total household deprivation. A very strong, positive correlation of 0.87 exists between total gas consumption and total energy consumption. A weak, positive correlation of 0.14 exists between total gas consumption and total household size. A weak, positive correlation of 0.14 exists between total gas consumption and total household deprivation. A weak, positive correlation of 0.24 exists between total energy consumption and total household size. A weak, positive correlation of 0.24 exists between total energy consumption and total household deprivation. The plot above shows a very strong positive correlation of 0.99 between 'Total household size' and the 'Total household deprivation' meaning there is multicollinearity between our two predictor variables that we plan to use in the multiple regression. This is most likely to render our model unrobust. The model is likely not to work since the multiple regresion assumption of no multicollinearity between predictors is violated.

16. Multiple regression analysis for model_A

model_A has Total energy consumption i.e Tot_energy as the dependent variable with total household size and total household deprivation as independent variables. This variable is the sum of Total electricity consumption and Total gas consumption.

```
model_A <- final_Redbridge %$% lm(Tot_energy~Tot_Hhold_size+Tot_Hhold_dep)
model_A %>% summary()
```

```
##
## Call:
## lm(formula = Tot_energy ~ Tot_Hhold_size + Tot_Hhold_dep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8277713 -1148680  -118115  1138174  9890832
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7953839     851318   9.343   <2e-16 ***
## Tot_Hhold_size  -103321      90764  -1.138    0.257
## Tot_Hhold_dep    109580      90691   1.208    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2231000 on 156 degrees of freedom
## Multiple R-squared:  0.1386, Adjusted R-squared:  0.1276
## F-statistic: 12.55 on 2 and 156 DF,  p-value: 8.821e-06
```

From the output above, both predictor variables are not significant since they have p-values greater than 0.01. Tot_Hhold_size has a slope of -103321 and Tot_Hhold_dep has a slope of 109580. The intercept of 7953839 is significant. The model is significant with p-value of 8.821e-06 for F(2,156)=12.55 hence p-value is less than 0.01. The model has a very low R-squared value of 0.1276 meaning the predictors(total house hold size and total household deprivation) can account for only 12.76% variation in the total energy consumed. Hence model_A is;

```
Tot_energy = 7953839-103321(Tot_Hhold_size)+ 109580(Tot_Hhold_dep)
```

16.1 Checking for Linearity, Outliers and Influential cases of model_A

```
final_Redbridge_outputA <- final_Redbridge %>%
  mutate(modelA_stdres=model_A %>%
           rstandard(),modelA_cook_dist=model_A %>%
```

```
          cooks.distance())
final_Redbridge_outputA %>%
  select(Tot_energy,modelA_stdres,modelA_cook_dist) %>%
  filter(abs(modelA_stdres)>2.58|modelA_cook_dist>1)
```

```
## # A tibble: 5 x 3
##   Tot_energy modelA_stdres modelA_cook_dist
##        <dbl>         <dbl>            <dbl>
## 1  17260574.          2.81           0.0477
## 2  24282673.          4.57           0.442
## 3   4132960.         -3.74           0.0681
## 4   4236534.         -2.83           0.0637
## 5   5398657.         -2.83           0.0320
```

The values for Cooks distance are not close to 1 meaning there are no strong outliers and the relationship is linear.

16.2 Checking the normality of standard residuals of model_A

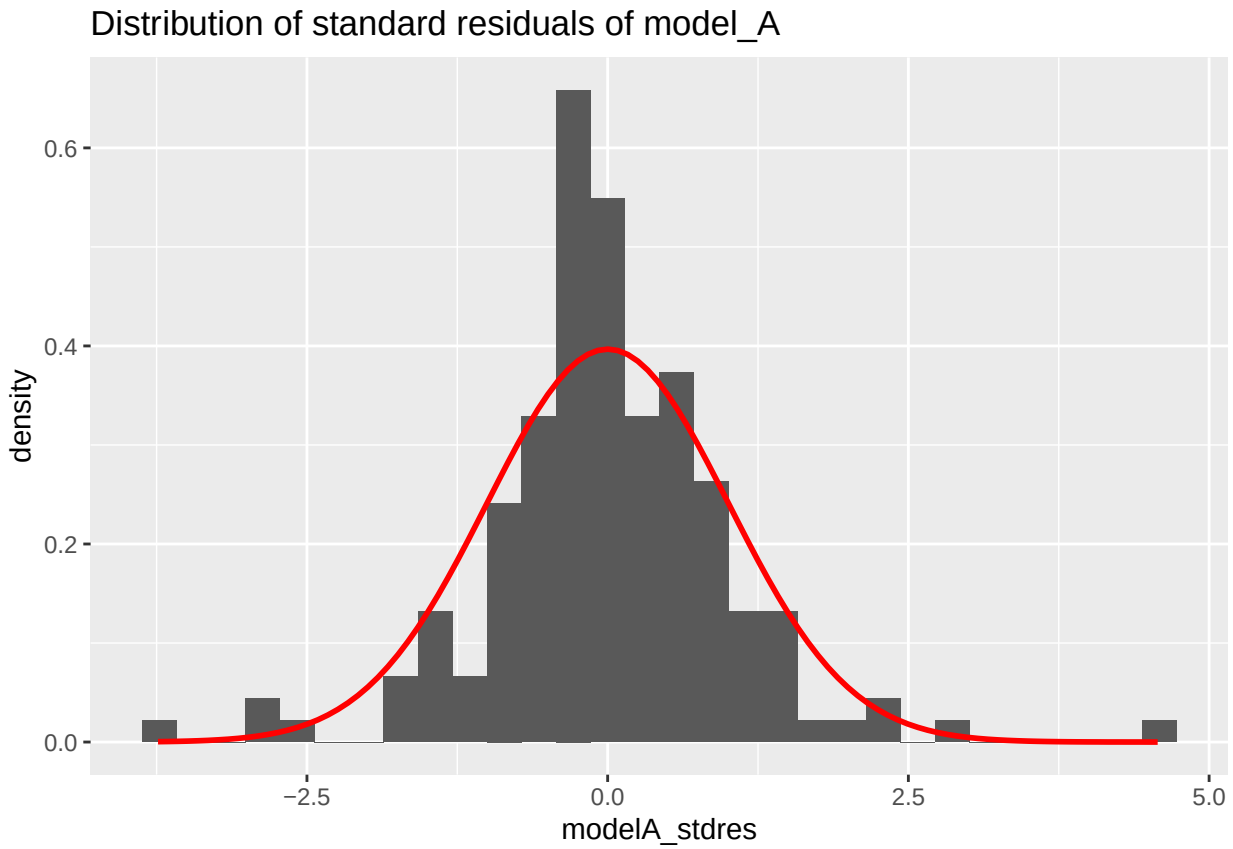
```
final_Redbridge_outputA %$%
  shapiro.test(modelA_stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelA_stdres
## W = 0.94739, p-value = 1.15e-05
```

The p-value from the Shapiro-Wilk test is less than 0.05, hence the standard residuals are not normally distributed. The plot below however shows that the distribution of the standard residuals is not far away from a normal distribution.

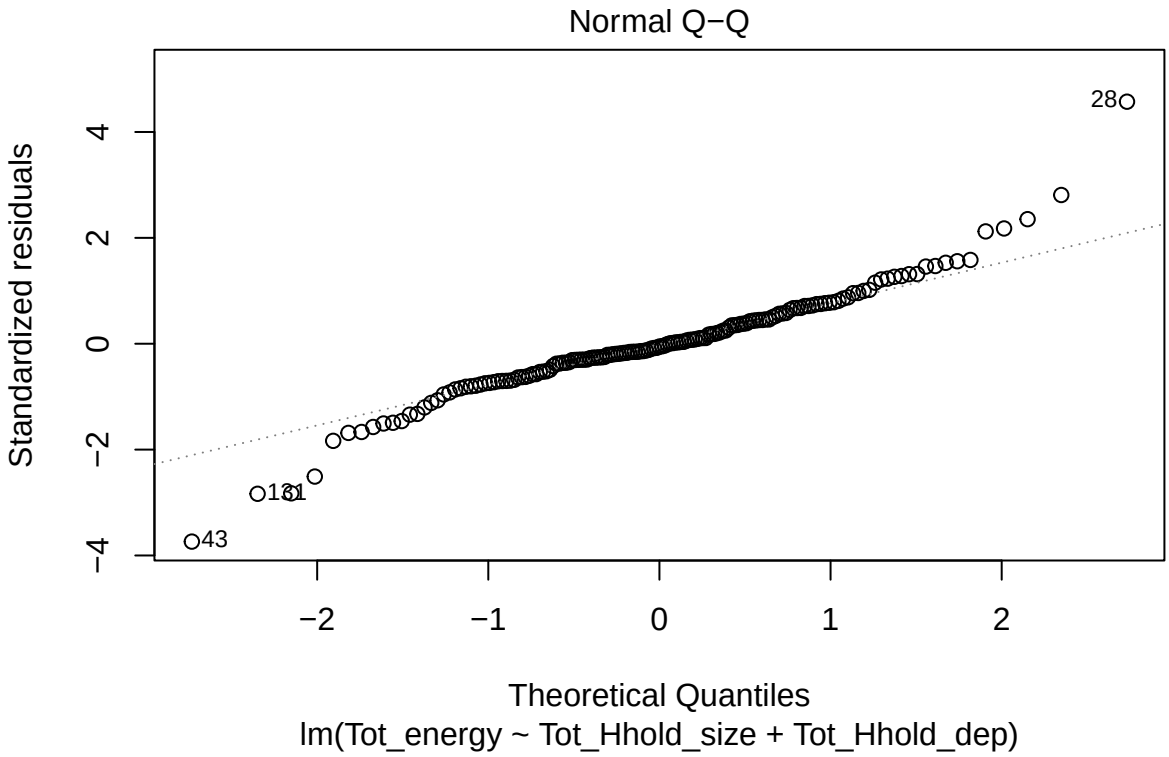
```
final_Redbridge_outputA %>%
  ggplot(
    aes(
      x = modelA_stdres
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge_outputA %>%
        filter(!is.na(modelA_stdres)) %>%
        pull(modelA_stdres) %>%
        mean(),
      sd =
        final_Redbridge_outputA %>%
        filter(!is.na(modelA_stdres)) %>%
        pull(modelA_stdres) %>%
```

```
        sd()
    ),
    colour = "red",
    size = 1
  )+ ggtitle("Distribution of standard residuals of model_A")
```



The Normal QQ plot below further shows that the standard residuals are not normally distributed

```
model_A %>%  plot(which = c(2))
```



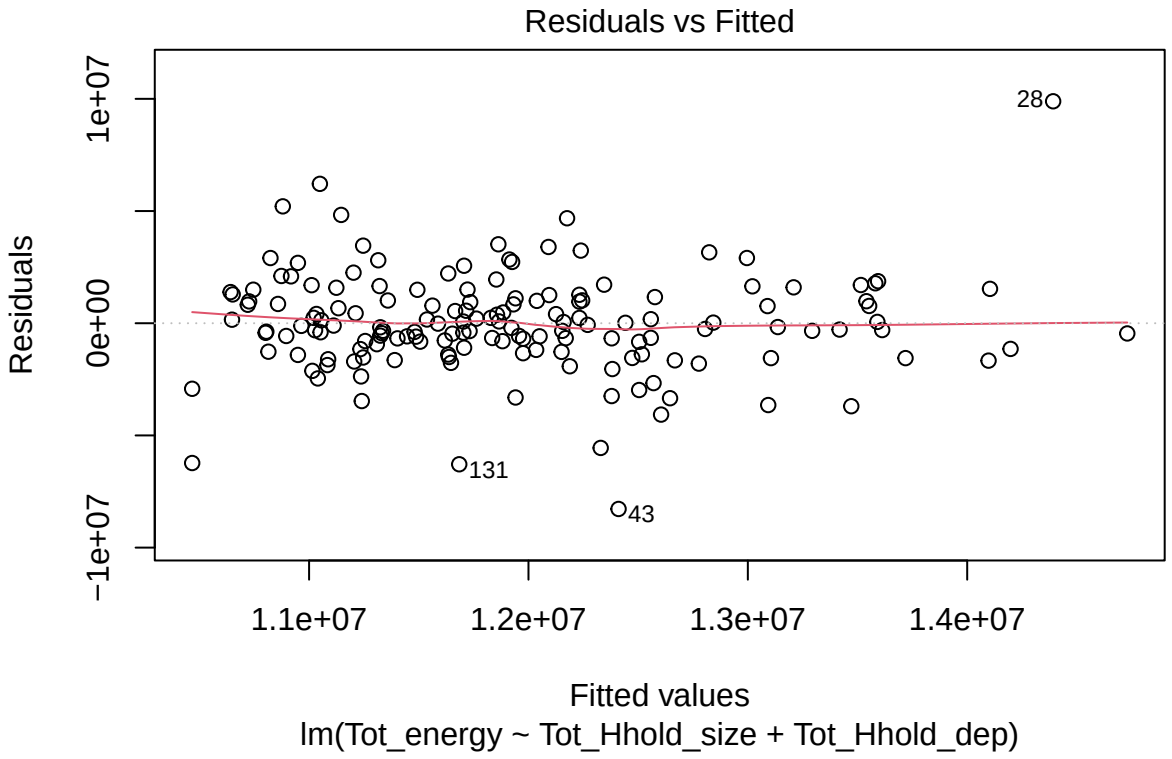
16.3 Checking Homoscedasticity of standard residuals of model_A

```
model_A %>% bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 4.7619, df = 2, p-value = 0.09246
```

From the Breusch-pagan test above,the p-value is greater than 0.05,hence the standard residuals are homoscedastic or have constant variance. The Residuals vs Fitted plot below further provides an insight into the homoscedasticity of the residuals.

```
model_A %>%  plot(which = c(1))
```



16.4 Checking Independence of standard residuals of model_A

The Durbin-Watson test below has a value of 1.6271 that is between 1 and 3 but the p-value is less than 0.05, hence the standard residuals are not independent.

```
model_A %>% dwtest()
```

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.6271, p-value = 0.008469
## alternative hypothesis: true autocorrelation is greater than 0
```

16.5 Checking multicollinearity between predictors in model_A

The Variance Inflation Factor(VIF) values of 4850.955 are much bigger than 10. Hence, there is multicollinearity between the Total household size and Total household deprivation as earlier illustrated by the pairs panel plot.

```
model_A %>% vif()
```

```
## Tot_Hhold_size  Tot_Hhold_dep
##       4850.955       4850.955
```

16.6 Conclusion for model_A

The model satisfies the assumption of linearity. In addition its standard residuals are homoscedastic, not normally distributed and not independent. Also, there exists multicollinearity between its two predictors. In conclusion, therefore,the model_A is not robust and does not work. Another model_B is proposed.

17. Multiple regression analysis for model_B

model_B has Total electricity consumption i.e Tot_elec as the dependent variable with total household size and total household deprivation as independent variables.

```
model_B <- final_Redbridge %$% lm(Total_elec~Tot_Hhold_size+Tot_Hhold_dep)
model_B %>% summary()
```

```
##
## Call:
## lm(formula = Total_elec ~ Tot_Hhold_size + Tot_Hhold_dep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -863103 -242424  -59806  197931 1098288
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      606900     138086   4.395 2.04e-05 ***
## Tot_Hhold_size   -10324      14722  -0.701    0.484
## Tot_Hhold_dep     13324      14710   0.906    0.366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361800 on 156 degrees of freedom
## Multiple R-squared:  0.566,  Adjusted R-squared:  0.5604
## F-statistic: 101.7 on 2 and 156 DF,  p-value: < 2.2e-16
```

From the output above, both predictor variables are not significant since they have p-values that are greater than 0.01. Tot_Hhold_size has a slope of -10324 and Tot_Hhold_dep has a slope of 13324. The intercept of 606900 is significant. The model is significant with p-value of 2.2e-16 for F(2,156)=101.7 since the p-value is less than 0.01. The model has an Adjusted R-squared value of 0.5604 meaning the predictors(total house hold size and total household deprivation) can account for 56.04% variation in the total electricity consumed. Hence model_B is;

```
Total_elec = 606900-10324(Tot_Hhold_size)+ 13324(Tot_Hhold_dep)
```

17.1 Checking for Linearity, Outliers and Influential cases of model_B

```
final_Redbridge_outputB <- final_Redbridge %>%
  mutate(modelB_stdres=model_B %>%
           rstandard(),modelB_cook_dist=model_B %>%
           cooks.distance())
final_Redbridge_outputB %>%
  select(Total_elec,modelB_stdres,modelB_cook_dist) %>%
  filter(abs(modelB_stdres)>2.58|modelB_cook_dist>1)
```

```
## # A tibble: 5 x 3
##   Total_elec modelB_stdres modelB_cook_dist
##        <dbl>         <dbl>            <dbl>
## 1   3027089.          2.72           0.0448
## 2   3610811.          3.07           0.0743
## 3   4543660.          2.76           0.161
## 4   4306821.          2.77           0.0808
## 5   4625942.          2.58           0.284
```

The values for Cooks distance are not close to 1 meaning there are no strong outliers and the relationship is linear.

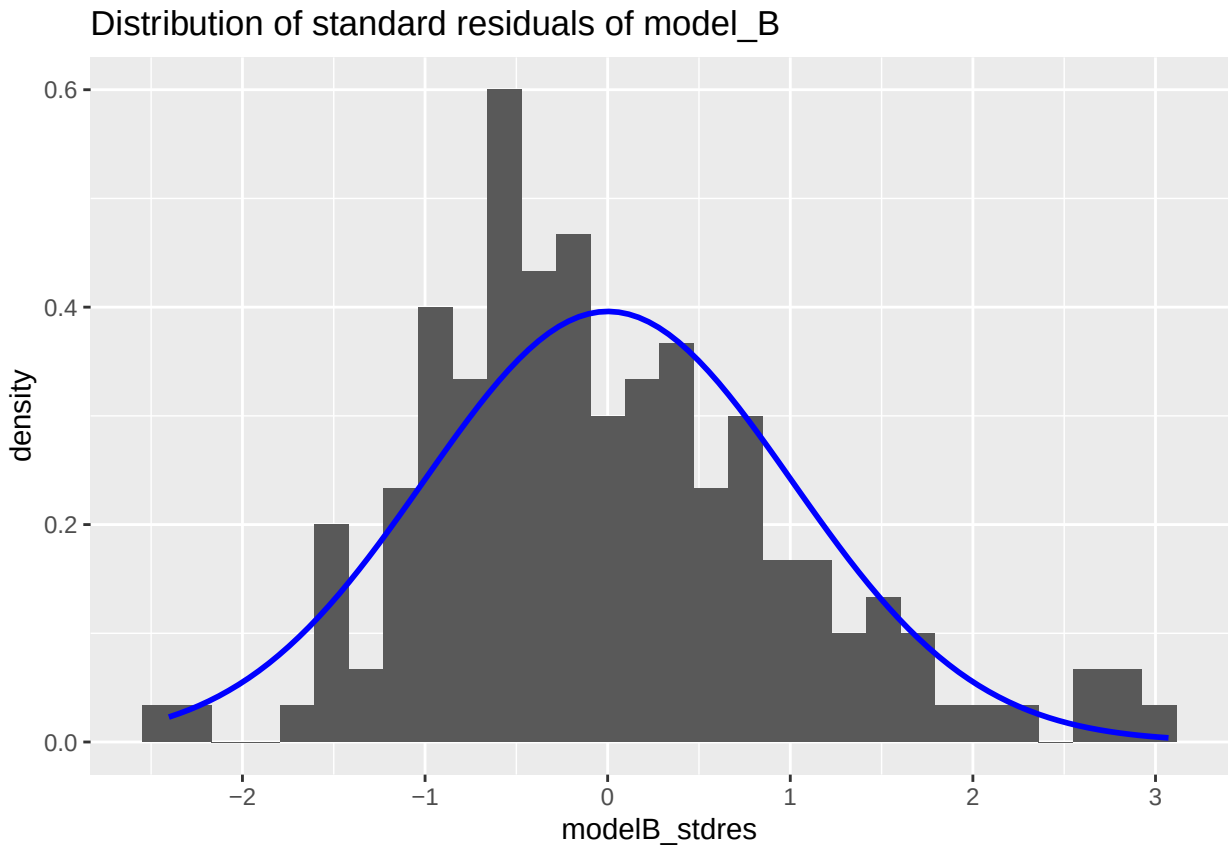
17.2 Checking the normality of standard residuals of model_B

```
final_Redbridge_outputB %$%
  shapiro.test(modelB_stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelB_stdres
## W = 0.96882, p-value = 0.001168
```

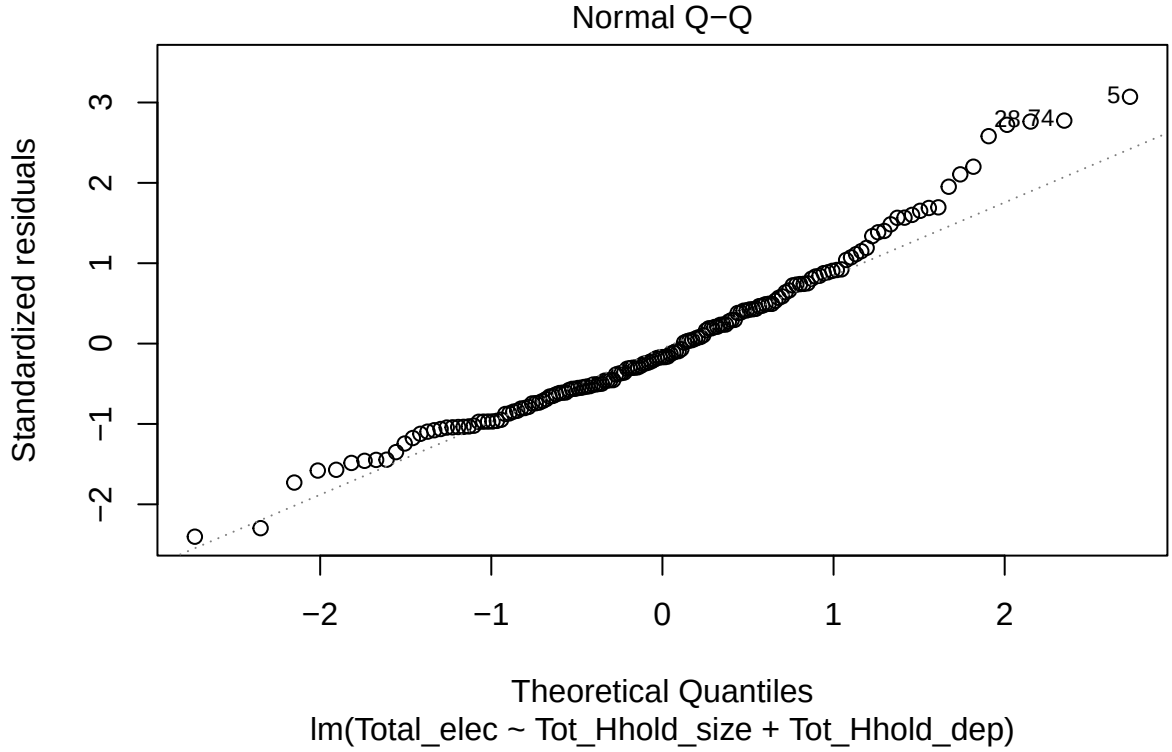
The p-value from the Shapiro-Wilk test is less than 0.05, hence the standard residuals are not normally distributed. The plot below however shows that the distribution of the standard residuals is not far away from a normal distribution.

```
final_Redbridge_outputB %>%
  ggplot(
    aes(
      x = modelB_stdres
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge_outputB %>%
        filter(!is.na(modelB_stdres)) %>%
        pull(modelB_stdres) %>%
        mean(),
      sd =
        final_Redbridge_outputB %>%
        filter(!is.na(modelB_stdres)) %>%
        pull(modelB_stdres) %>%
        sd()
    ),
    colour = "blue",
    size = 1
  )+ ggtitle("Distribution of standard residuals of model_B")
```



The Normal QQ plot below further shows that the standard residuals are not normally distributed

```
model_B %>%  plot(which = c(2))
```



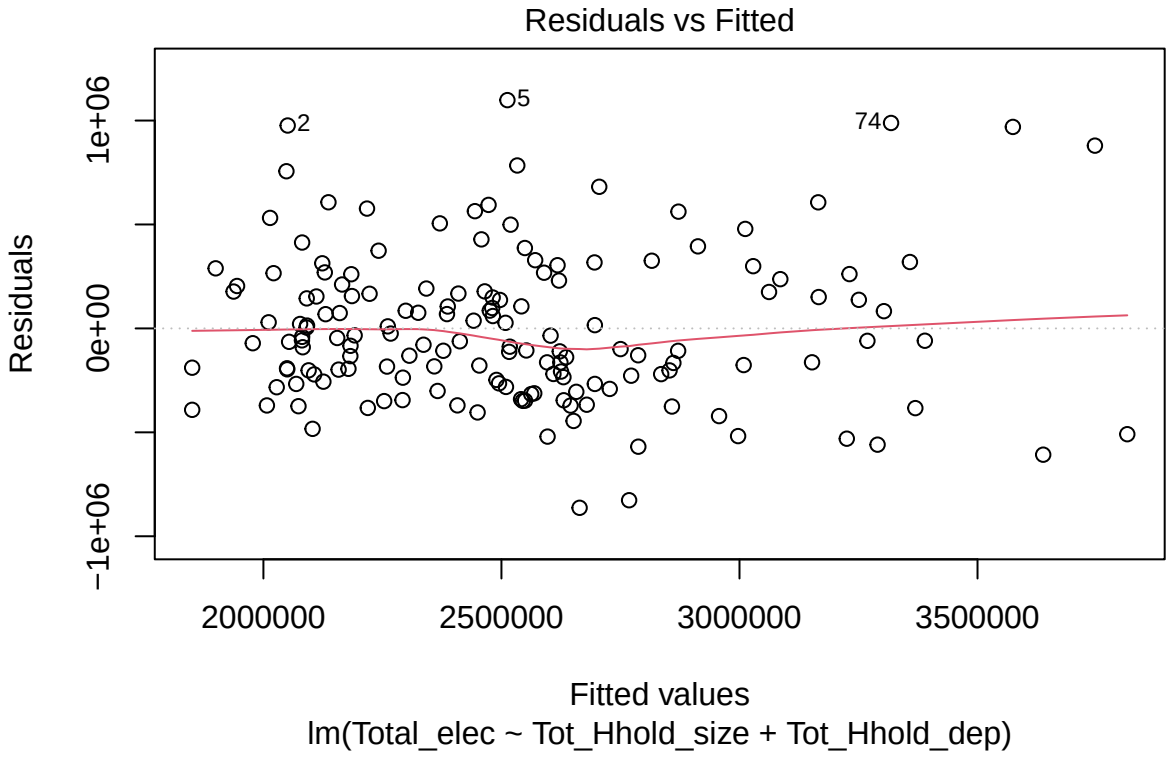
17.3 Checking Homoscedasticity of standard residuals of model_B

```
model_B %>% bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 12.797, df = 2, p-value = 0.001664
```

From the Breusch-pagan test above,the p-value is less than 0.05,hence the standard residuals are not homoscedastic or have no constant variance. The Residuals vs Fitted plot below further illustrates this.

```
model_B %>%  plot(which = c(1))
```



17.4 Checking Independence of standard residuals of model_B

The Durbin-Watson test below has a value of 1.6528 that is between 1 and 3 but the p-value is less than 0.05, hence the standard residuals are not independent.

```
model_B %>% dwtest()
```

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.6528, p-value = 0.01304
## alternative hypothesis: true autocorrelation is greater than 0
```

17.5 Checking multicollinearity between predictors in model_B

The Variance Inflation Factor(VIF) values of 4850.955 are much bigger than 10. Hence, there is multicollinearity between the Total household size and Total household deprivation as earlier illustrated by the pairs panel plot.

```
model_B %>% vif()
```

```
## Tot_Hhold_size  Tot_Hhold_dep
##       4850.955       4850.955
```

17.6 Conclusion for model_B

The model satisfies the assumption of linearity. In addition its standard residuals are not homoscedastic, not normally distributed and not independent. Also, there exists multicollinearity between its two predictors. In conclusion, therefore,model_B too is not robust and does not work. Another model_C is proposed.

18. Multiple regression analysis for model_C

model_C has Total gas consumption i.e Tot_gas as the dependent variable with total household size and total household deprivation as independent variables.

```
model_C <- final_Redbridge %$% lm(Total_gas~Tot_Hhold_size+Tot_Hhold_dep)
model_C %>% summary()
```

```
##
## Call:
## lm(formula = Total_gas ~ Tot_Hhold_size + Tot_Hhold_dep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7414610  -951911   -53844  1205101  8921495
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7346938     779097    9.43   <2e-16 ***
## Tot_Hhold_size   -92996      83064   -1.12    0.265
## Tot_Hhold_dep     96256      82997    1.16    0.248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2041000 on 156 degrees of freedom
## Multiple R-squared:  0.05529,    Adjusted R-squared:  0.04318
## F-statistic: 4.565 on 2 and 156 DF,  p-value: 0.01183
```

From the output above, both predictor variables are not significant since they have p-values that are greater than 0.01. Tot_Hhold_size has a slope of -92996 and Tot_Hhold_dep has a slope of 96256. The intercept of 7346938 is significant. The model is not significant with p-value of 0.01183 for F(2,156)=4.565 since p-value is greater than 0.01. The model has a very low Adjusted R-squared value of 0.04318 meaning the predictors(total house hold size and total household deprivation) can account for only 4.318% variation in the total gas consumed. Hence model_C is;

```
Total_gas = 7346938-92996(Tot_Hhold_size)+ 96256(Tot_Hhold_dep)
```

18.1 Checking for Linearity, Outliers and Influential cases of model_C

```
final_Redbridge_outputC <- final_Redbridge %>%
  mutate(modelC_stdres=model_C %>%
           rstandard(),modelC_cook_dist=model_C %>%
           cooks.distance())
final_Redbridge_outputC %>%
  select(Total_gas,modelC_stdres,modelC_cook_dist) %>%
  filter(abs(modelC_stdres)>2.58|modelC_cook_dist>1)
```

```
## # A tibble: 6 x 3
##   Total_gas modelC_stdres modelC_cook_dist
##       <dbl>         <dbl>            <dbl>
## 1 14233485.          2.59           0.0405
## 2 19739013.          4.51           0.429
## 3  2331896.         -3.66           0.0652
## 4  2777980.         -2.89           0.0667
## 5  4354220.         -2.64           0.0334
## 6  2331585.         -3.39           0.0458
```

The values for Cooks distance are not close to 1 meaning there are no strong outliers and the relationship is linear.

18.2 Checking the normality of standard residuals of model_C

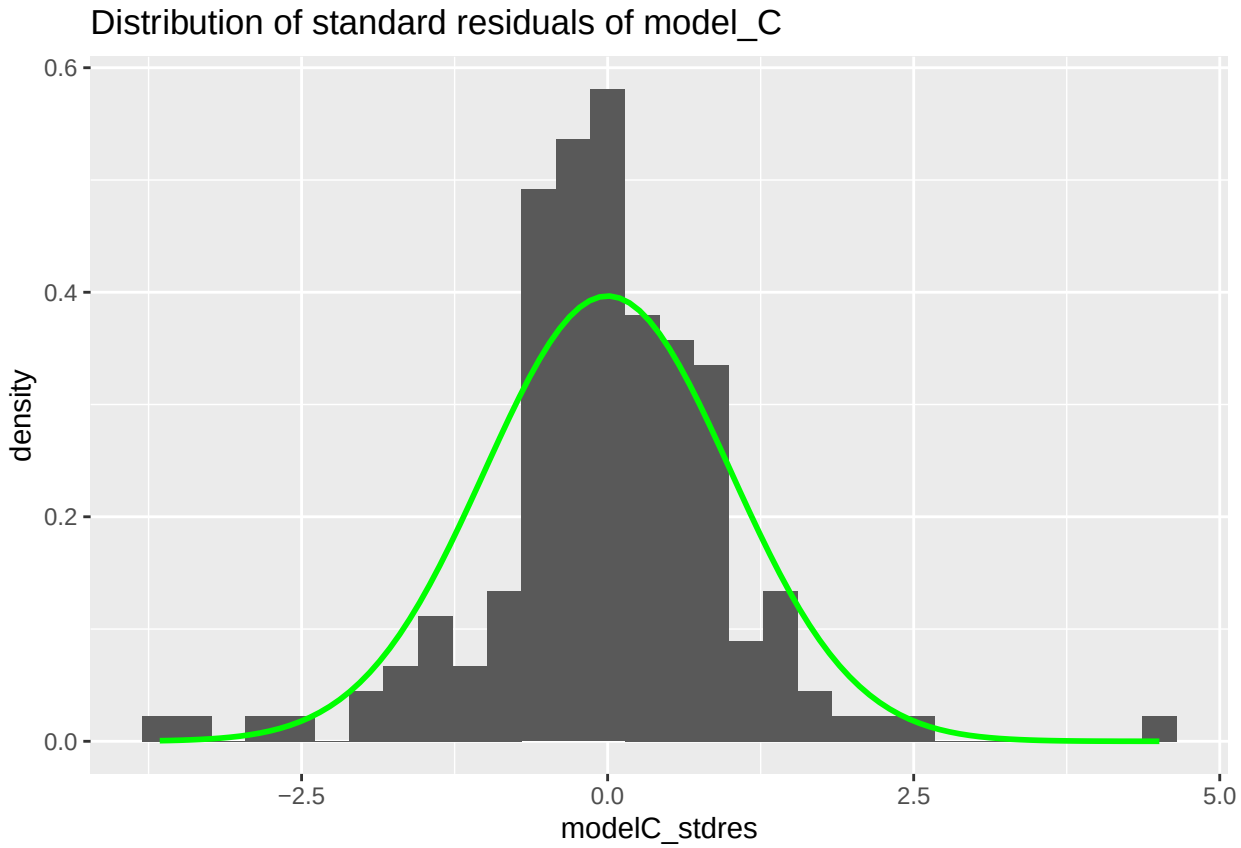
```
final_Redbridge_outputC %$%
  shapiro.test(modelC_stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelC_stdres
## W = 0.94234, p-value = 4.454e-06
```

The p-value from the Shapiro-Wilk test is less than 0.05, hence the standard residuals are not normally distributed. The plot below however shows that the distribution of the standard residuals is not far away from a normal distribution.

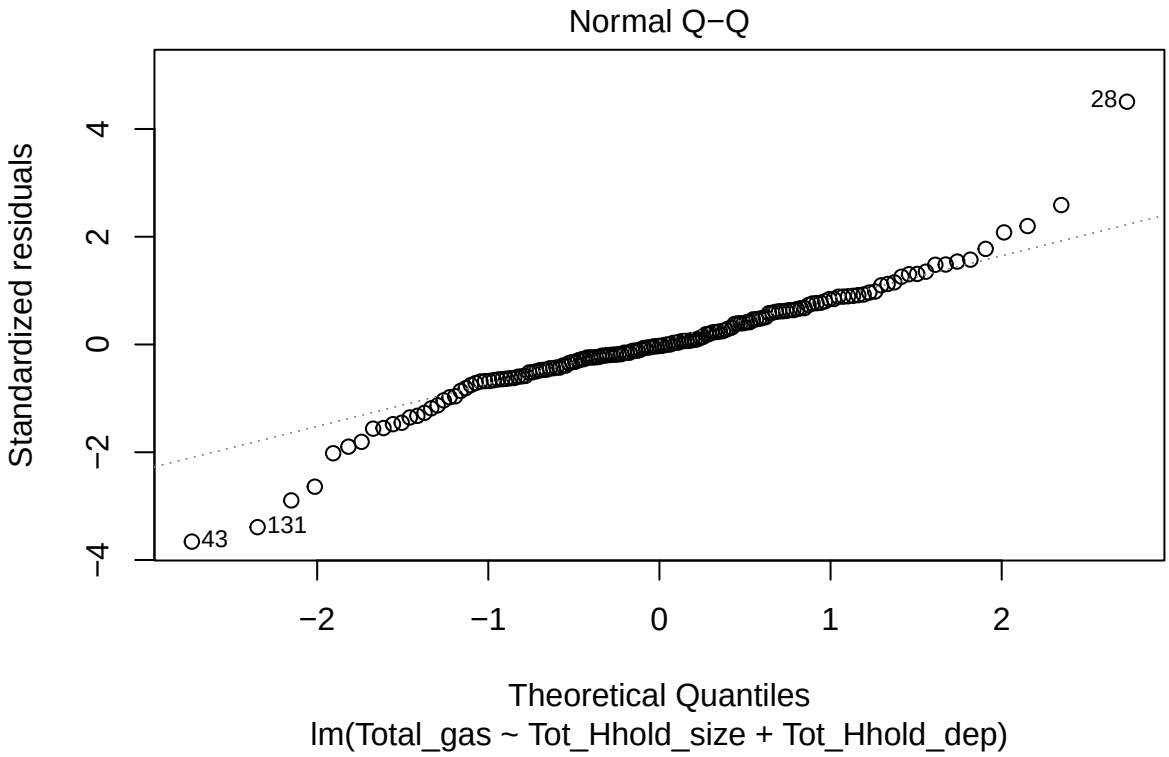
```
final_Redbridge_outputC %>%
  ggplot(
    aes(
      x = modelC_stdres
    )
  ) +
  geom_histogram(
    aes(
      y =..density..
    ),
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean =
        final_Redbridge_outputC %>%
        filter(!is.na(modelC_stdres)) %>%
```

```
      pull(modelC_stdres) %>%
      mean(),
    sd =
      final_Redbridge_outputC %>%
      filter(!is.na(modelC_stdres)) %>%
      pull(modelC_stdres) %>%
      sd()
  ),
  colour = "green",
  size = 1
)+ ggtitle("Distribution of standard residuals of model_C")
```



The Normal QQ plot below further shows that the standard residuals are not normally distributed

```
model_C %>%  plot(which = c(2))
```



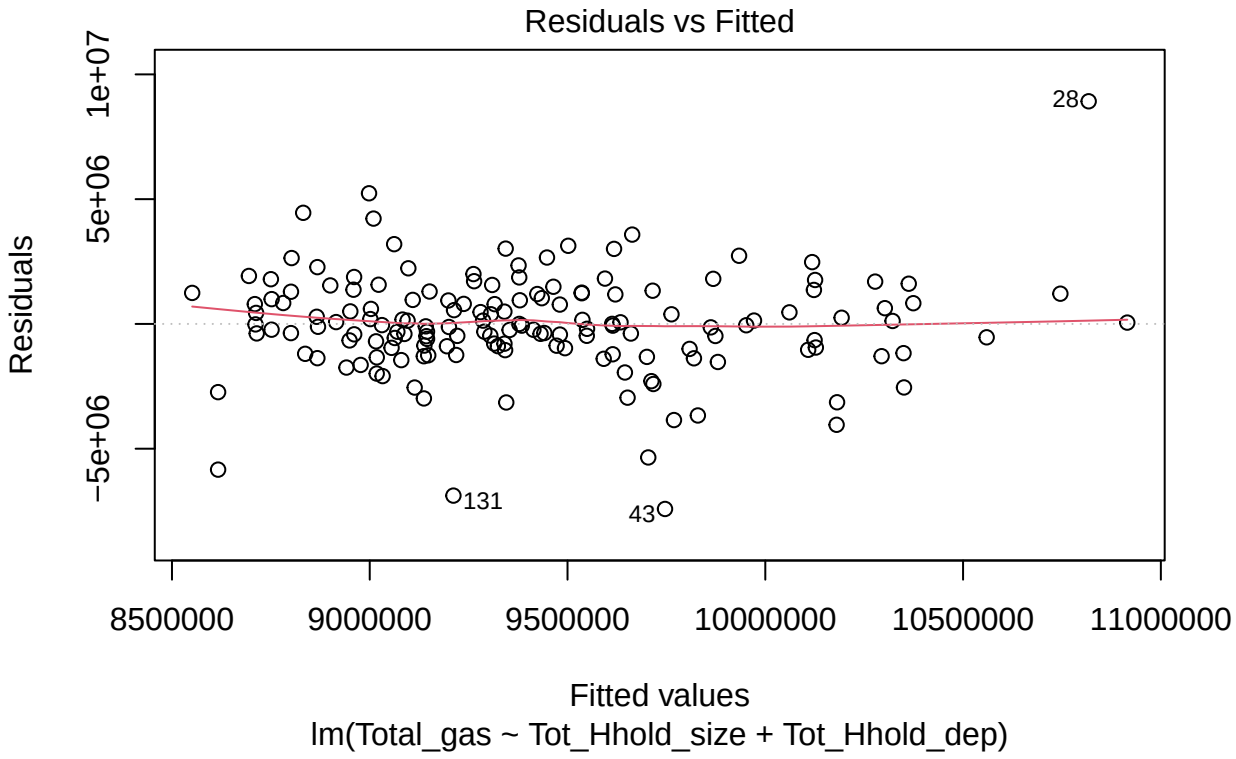
18.3 Checking Homoscedasticity of standard residuals of model_C

```
model_C %>% bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 3.8626, df = 2, p-value = 0.145
```

From the Breusch-pagan test above,the p-value is greater than 0.05,hence the standard residuals are homoscedastic or have constant variance. The Residuals vs Fitted plot below further provides an insight into the homoscedasticity of the residuals.

```
model_C %>%  plot(which = c(1))
```



18.4 Checking Independence of standard residuals of model_C

The Durbin-Watson test below has a value of 1.6576 that is between 1 and 3 but the p-value is less than 0.05, hence the standard residuals are not independent.

```
model_C %>% dwtest()
```

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.6576, p-value = 0.0141
## alternative hypothesis: true autocorrelation is greater than 0
```

18.5 Checking multicollinearity between predictors in model_C

The Variance Inflation Factor(VIF) values of 4850.955 are much bigger than 10. Hence, there is multicollinearity between the Total household size and Total household deprivation as earlier illustrated by the pairs panel plot.

```
model_C %>% vif()
```

```
## Tot_Hhold_size  Tot_Hhold_dep
##       4850.955       4850.955
```

18.6 Conclusion for model_C

The model satisfies the assumption of linearity. In addition its standard residuals are homoscedastic, not normally distributed and not independent. Also, there exists multicollinearity between its two predictors. In conclusion, therefore,the model_C is not robust and does not work.

19. Comparison of model_A, model_B and model_C

Though the models are not robust, the Akaike Information Criterion(AIC) is used below to identify which of the 3 models is a better one, much as they are all not robust. From the analysis below, model_B which has the lowest AIC value of 4526.241 and highest adjusted R squared value of 0.5604 would have been a better model to use of the 3, had they been robust.

```
      Tot_energy = 7953839-103321(Tot_Hhold_size)+ 109580(Tot_Hhold_dep)
```

```
AIC(model_A)
```

```
## [1] 5104.654
```

```
      Total_elec = 606900-10324(Tot_Hhold_size)+ 13324(Tot_Hhold_dep)
```

```
AIC(model_B)
```

```
## [1] 4526.241
```

```
      Total_gas = 7346938-92996(Tot_Hhold_size)+ 96256(Tot_Hhold_dep)
```

```
AIC(model_C)
```

```
## [1] 5076.463
```

Conclusion

The three models A, B and C generally have low adjusted R squared values and each model fails to satisfy ALL the five assumptions of linearity, normality,homoscedasiticity, independence and mullticollinearity. Hence all 3 models are not robust and cannot be used for applicable regression.

References

R Core Team.(2022)'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing*, Vienna, Austria, https://www.R-project.org/

Stefano,D.S.(2022) 'R for Geographic Data Science', *github*, https://sdesabbata.github.io/r-for-geographic-data-science/index.html