Contents

Activity 1

Introduction

This dataset contains 1500 houses sold in Stockton, California, during 1996 -1998. The purpose of dataset in dataset is to examine how the sale price of houses in Stockton, California, are affected by house characteristics. There are 7 variables:

- Quantifier: sprice, livarea, beds, baths, age
- Qualitative: lgelot, pool

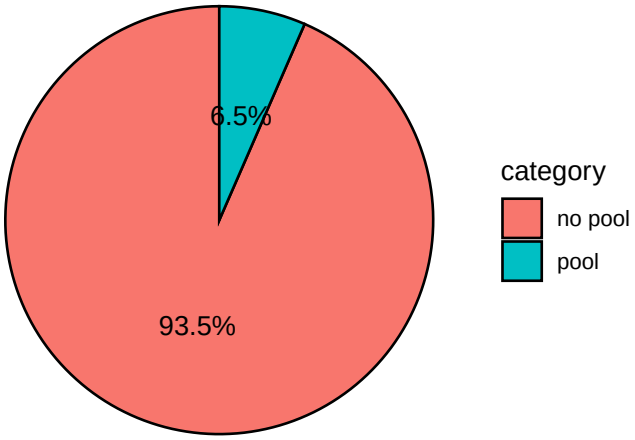
Variable	Definition
sprice	Selling price of home, in dollars
livarea	Living area, in hundreds of square feet
beds	Number of beds
baths	Number of baths
lgelot	1 if lot size is greater than 0.5 acres, 0 otherwise
age	Age of home at time of sale, in years

Variable	Definition
pool	1 if home has a pool, 0 otherwise

Data source: Dr. John Knight, Department of Finance, University of the Pacific.

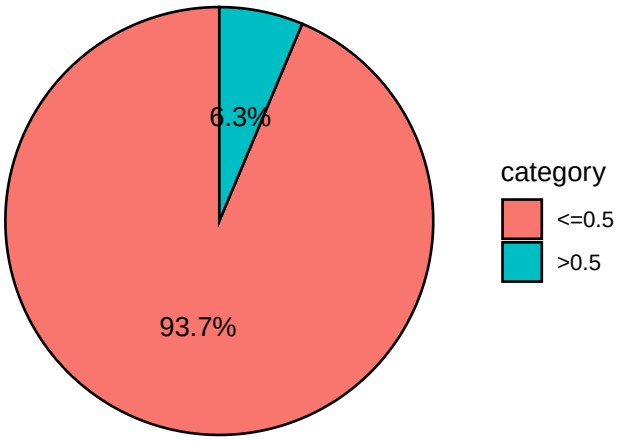
I. Analysis qualitative

a) pool:



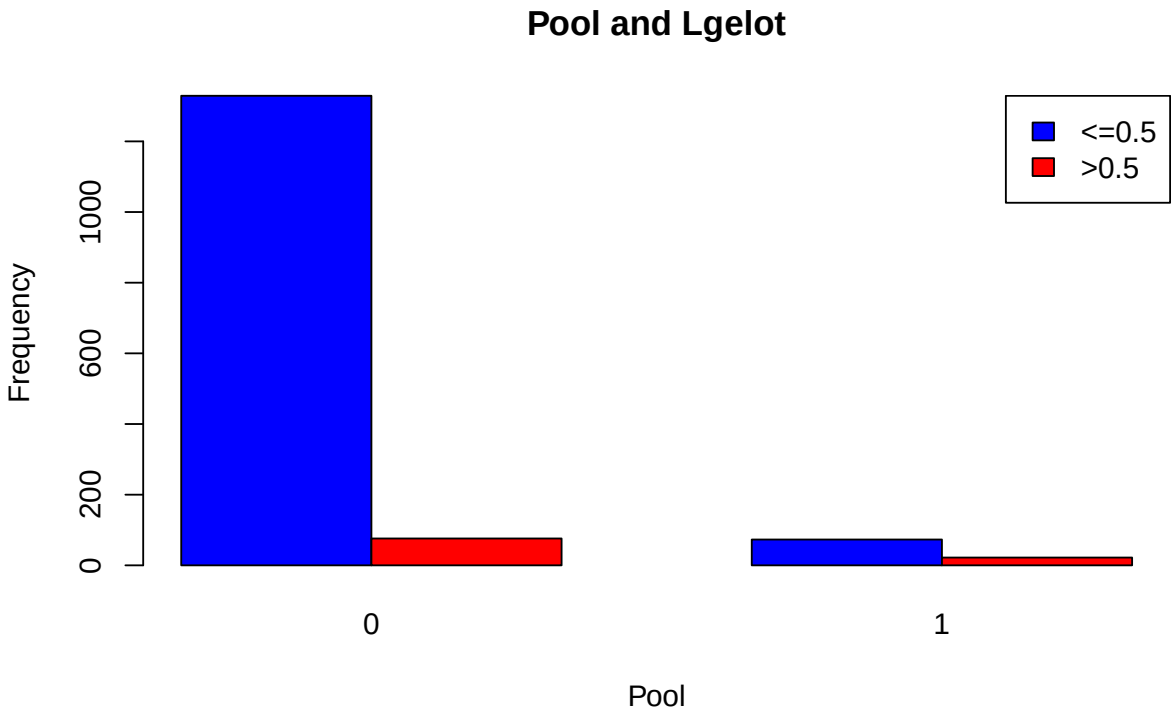
Comment: The number of houses without a pool is large (1402), while the number of houses with a pool is only a small part (98) in the collected data table. There is a significant difference between the number of houses with and without a pool. The proportion of the houses with pool in the dataset is smaller than 7%.

b) lgelot:

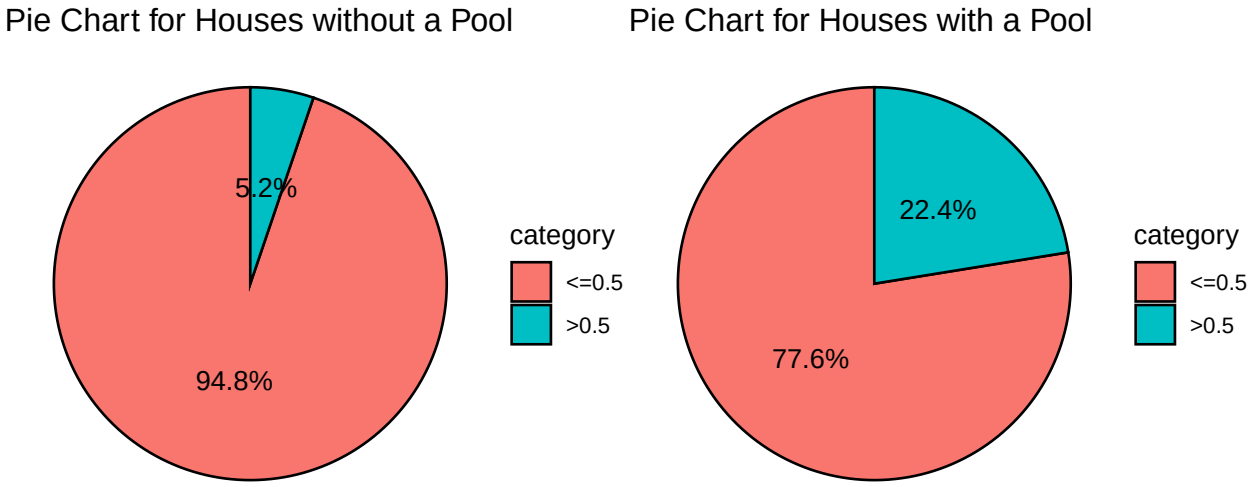


Comment: Most of the houses in the data table have a size of 0.5 acres or smaller(1405), while the number of houses with a larger size is only a very small part(95). This shows a significant difference between houses with a size larger than 0.5 acres and those from 0.5 acres down. The proportion of the houses with size larger than 0.5 acres is bigger than 6%.

c) pool and lgelot:



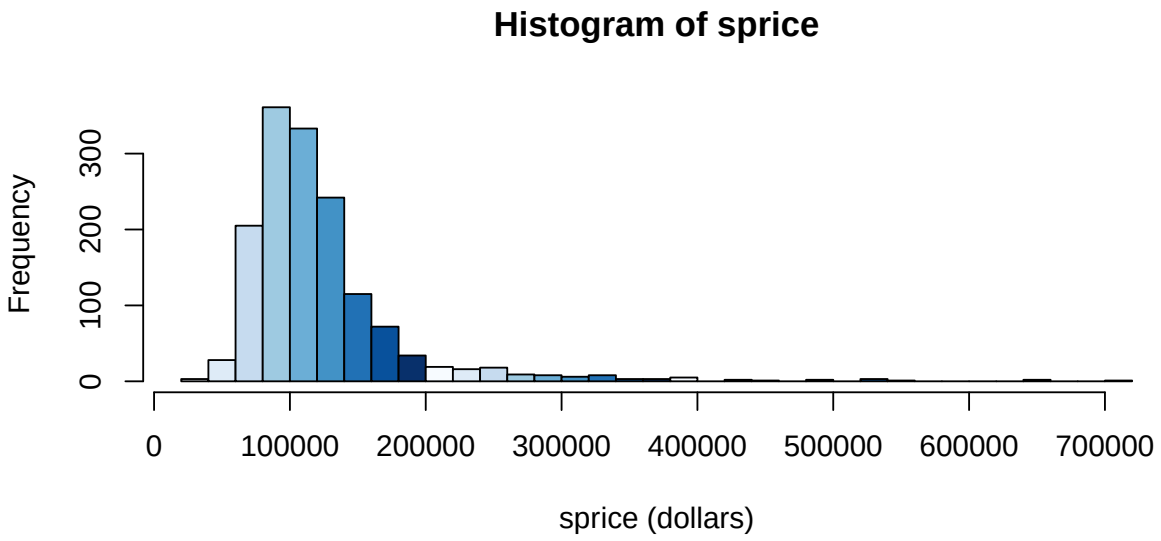
Comment: Based on the graph, among the houses with or without a pool, the number of houses with a size larger than 0.5 acres is smaller than the number of houses with a size of 0.5 acres or less.



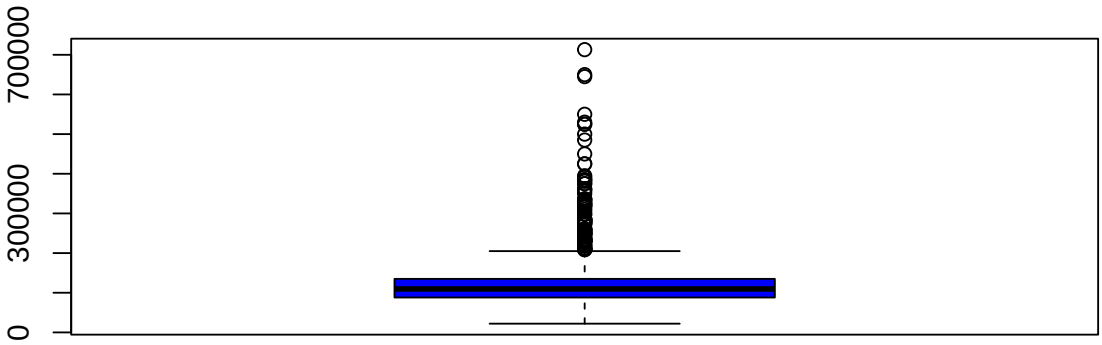
Comment: The proportion of houses with a size of larger than 0.5 acres in houses without pool is smaller than the proportion of houses with a size of larger than 0.5 acres in houses with pool.

II. Analysis Quantifier

a) sprice



Comment: The chart has a positive skew. The selling price with the highest number falls around 100,000 dollars. The selling price falls mostly in the range of 50,000 dollars to 140,000 dollars. The average selling price is about 100,000 dollars.



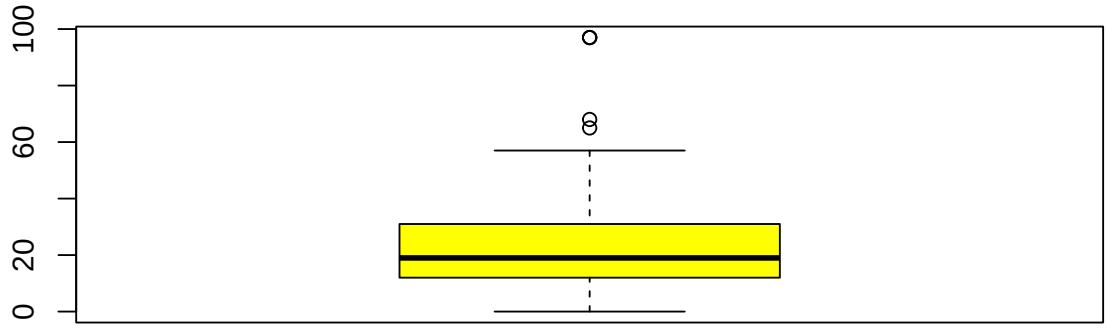

```
## [1] "Percent of outliers: 0.0673333333333333"
```

```
## sd =  63250.89
```

Comment: The data has about 101 outliers, which is accounted for 6.7% in dataset. The data does not have large fluctuations. The highest price is around over 700,000 dollars and the lowest price is about 22,000 dollars. The price difference is about 63000 dollars.

b) age

Comment: Houses that are about 18 years occupies the largest number. Houses are distributed from when they were built to about 50 years, with houses for the most from 10 to 22 years old. Houses that are above 60 years are rare and almost non-existent. The average age of the house is around 18

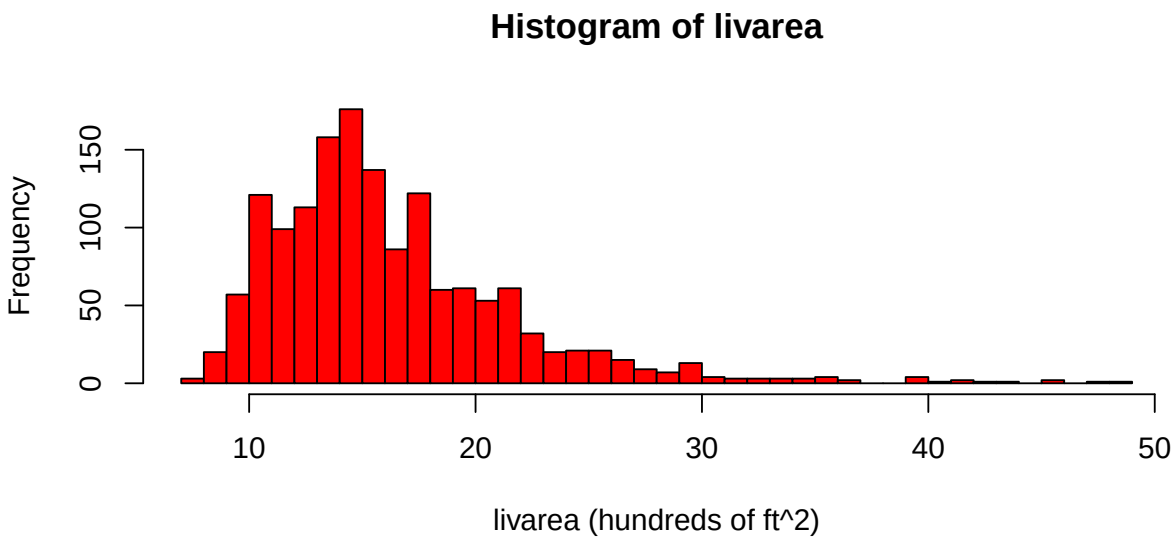



```
## [1] "Percent of outliers: 0.00333333333333333"
```

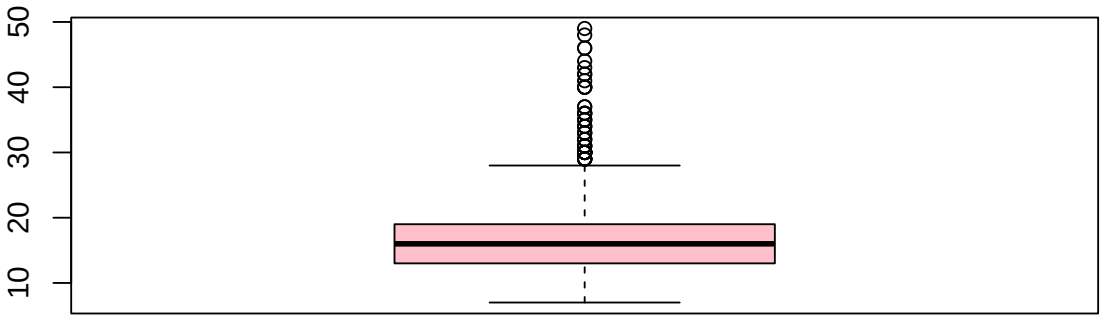
```
## sd =  13.11464
```

Comment: The data has outliers, which is not a large proportion compared to the dataset. The data has a wide fluctuation. The median is skewed downwards. The oldest house is over 95 years and the age difference between the houses is about 13 years.

c) livarea



Comment: Houses with an area of around 15 hundreds of square feet occupy the largest number. The chart has a positive skew, indicating that houses with small areas occupy a large number. The number of houses is distributed from 10 hundreds of square feet to 18 hundreds of square feet. The average house area is around 17 hundreds of square feet.



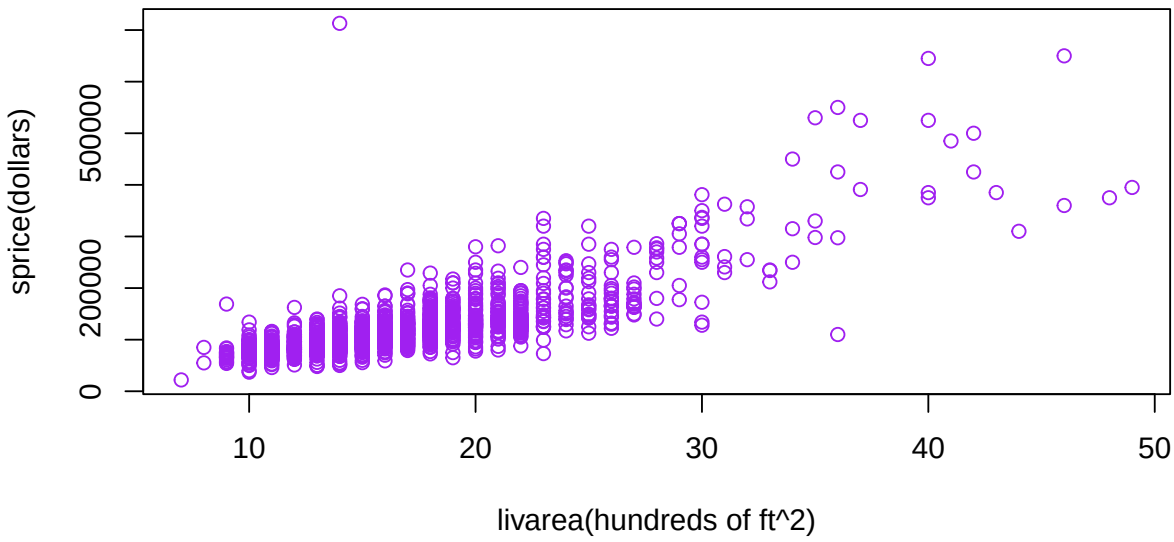
```
## [1] "Percent of outliers: 0.0366666666666667"
```

```
## sd =  5.461963
```

```
## [1] 55
```

Comment: The number of outliers is not significant compared to the dataset. The data has fluctuations but not large. The largest area is close to 50 hundreds of square feet and the smallest is around 5 hundreds of square feet. The area difference is about 5 hundreds of square feet.

d) sprice and livarea




```
## [1] 0.7928769
```

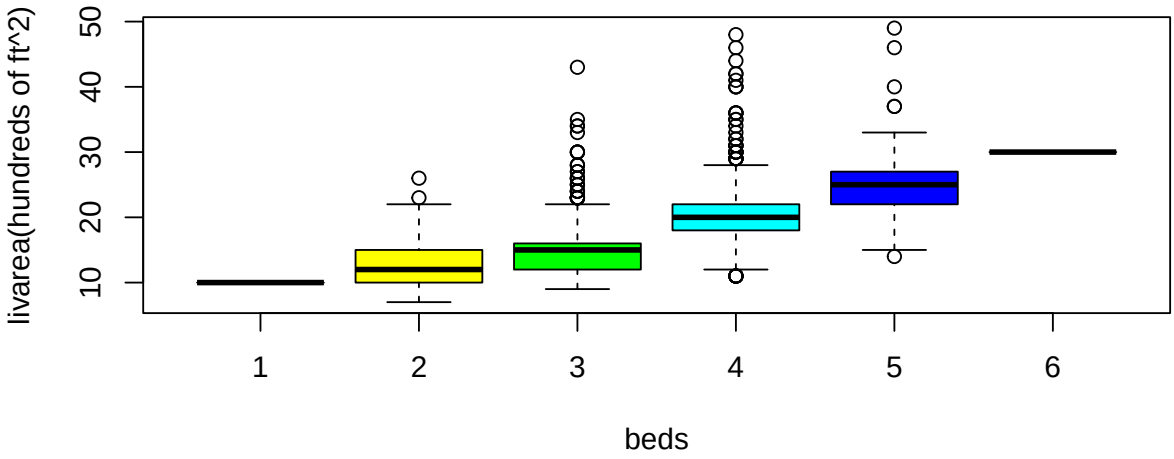
Comment: Based on the graph, the higher the area of the house, the higher the selling price of the house. The correlation rate is nearly 80%.

e) livarea and baths

```
## [1] 0.7233931
```

Comment: Drawing like above, based on the graph, the more baths there are, the higher the area of the house. The correlation between baths and livarea is about 72%

f) sprice and beds

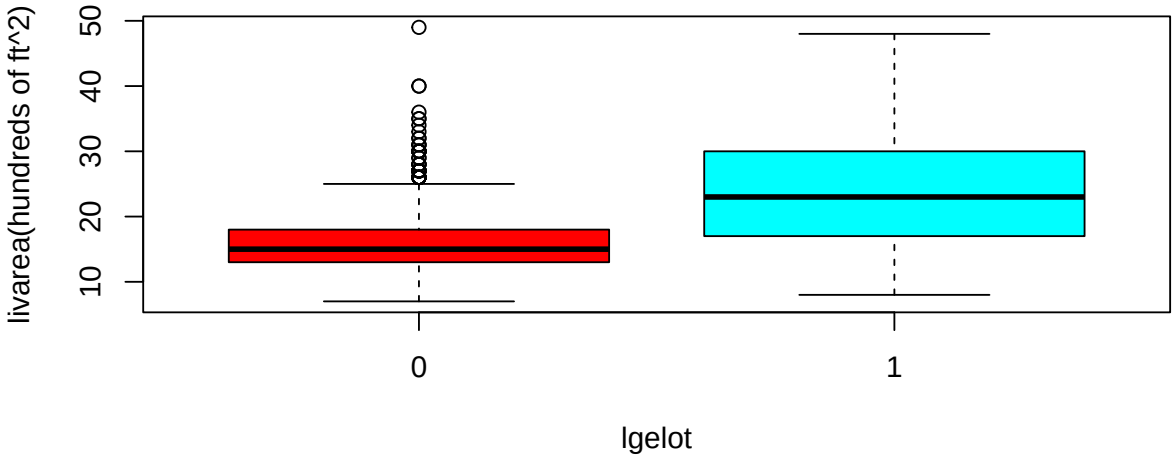



```
## [1] 0.5752849
```

Comment: Continue with beds, based on the boxplot graph, the more bedrooms a house has, the higher the area of the house. On average, the area of a house with 2 bedrooms is smaller than that of a house with 3 bedrooms.There is a relatively correlation of about 58%.

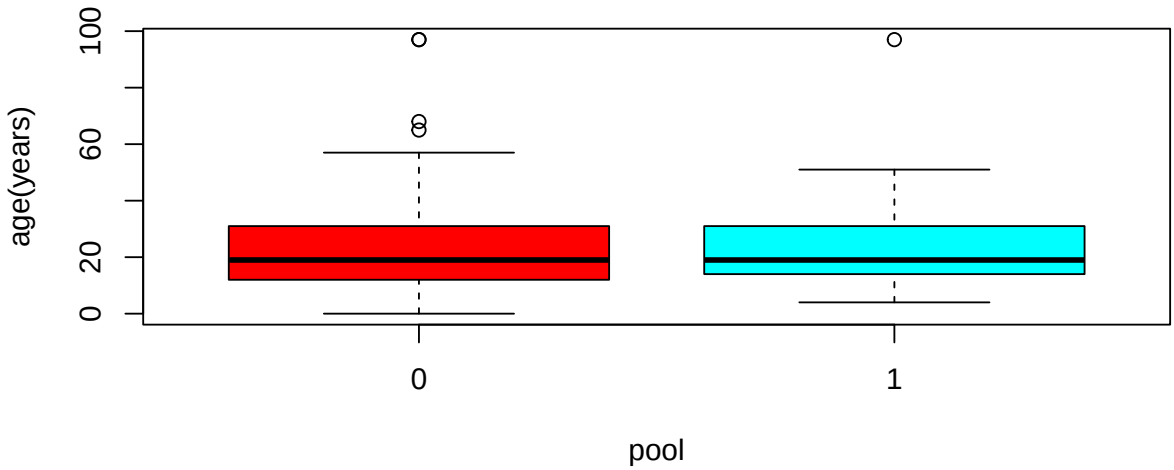
III. Quantifier and Qualitative

a) livarea and lgelot



Comment: For houses with a size of <=0.5 acres, there are some outliers but the number is not significant, while houses with a size larger than 0.5 acres have no outliers. Based on the boxplot graph, we see that the average area of houses with a size of <=0.5 acres is smaller than the average area of houses with a size of >0.5 acres. However, houses with a size of <=0.5 acres do not vary much, while houses with a size >0.5 acres have large and wide variations.

b) age and pool



Comment: Houses with and without swimming pools both have some outliers but the number is not significant. Based on the boxplot, both types of houses have relative variations in the age of the house. It seems that the average age of houses with swimming pools is equal to the average age of houses without swimming pools.

c) sprice and lgelot

Comment: Let's do exactly the same as two case above. Base on **Figure 1** For houses with a size of <=0.5 acres and >0.5 acres, there are some outliers where the number of outliers for <=0.5 acres houses is quite high.The selling price of >0.5 acres houses has wide and large variations, while <=0.5 acres houses have insignificant variations. Houses with a size >0.5 acres have a higher average selling price compared to houses with a size <=0.5 acres.

d) sprice and pool

Comment: Continue with sprice and pool,base on **Figure 2** for houses with and without a pool, there are some outliers, with a higher number of outliers for houses without a pool. The selling price of houses with a pool has wide and large variations, while houses without a pool have insignificant variations. Houses with a pool have a higher average selling price compared to houses without a pool.

IV. Hypothesis Testing

I.Qualitative

a) The proportion of the houses with pool in the dataset is not more than 7%.

p is the proportion of the houses with pool in the dataset.

$H_0 : p = 7\%$

$H_a : p < 7\%$

```
##
##  1-sample proportions test without continuity correction
##
## data:  bangpool[2] out of length(pool), null probability 0.07
## X-squared = 0.50179, df = 1, p-value = 0.2394
## alternative hypothesis: true p is less than 0.07
## 95 percent confidence interval:
##  0.00000000 0.07663052
## sample estimates:
##          p
## 0.06533333
```

Because $p_{value} = 0.2394 > 0.05 \rightarrow$ *Accept* H_0. Therefore, we cannot conclude that the proportion of the houses with pool in the dataset is not more than 7% at risk level $\alpha = 5\%$

b) The proportion of the houses with size larger than 0.5 acres is bigger than 6%.

p is the proportion of the houses with size larger than 0.5 acres

$H_0 : p = 6\%$

$H_a : p > 6\%$

```
##
##  1-sample proportions test without continuity correction
##
## data:  banglge[2] out of length(lgelot), null probability 0.06
## X-squared = 0.29551, df = 1, p-value = 0.2934
## alternative hypothesis: true p is greater than 0.06
## 95 percent confidence interval:
##  0.05375494 1.00000000
## sample estimates:
##          p
## 0.06333333
```

Because $p_{value} = 0.2934 > 0.05 \rightarrow$ *Accept* H_0. Therefore, we cannot conclude that the proportion of the houses with size larger than 0.5 acres is bigger than 6%. at risk level $\alpha = 5\%$

c) The proportion of houses with a size of larger than 0.5 acres in houses without pool is smaller than the proportion of houses with a size of larger than 0.5 acres in houses with pool.
p_1 is the proportion of houses with a size of larger than 0.5 acres in houses without pool

p_2 is the proportion of houses with a size of larger than 0.5 acres in houses with pool

$$H_0 : p_1 = p_2$$

$$H_a : p_1 < p_2$$

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(tbl[3], tbl[4]) out of c(length(pool[pool == "0"]), length(pool[pool == "1"]))
## X-squared = 45.904, df = 1, p-value = 0.000000000006211
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.1024101
## sample estimates:
##     prop 1     prop 2
## 0.05206847 0.22448980
```

Because $p_{value} < 0.05 \rightarrow Reject\ H_0$. Therefore, we conclude that the proportion of houses with a size of larger than 0.5 acres in houses without pool is smaller than the proportion of houses with a size of larger than 0.5 acres in houses with pool at risk level $\alpha = 5\%$

II.Quantifier

a) The average selling price is about 100,000 dollars

$$H_0 : \mu = 100000$$

$$H_a : \mu \neq 100000$$

```
##
##  One Sample t-test
##
## data:  sprice
## t = 14.508, df = 1499, p-value < 0.00000000000000022
## alternative hypothesis: true mean is not equal to 100000
## 95 percent confidence interval:
##  120490.4 126897.3
## sample estimates:
## mean of x
##  123693.9
```

Because $p_{value} < 0.05 \rightarrow Reject\ H_0$. Therefore, we conclude the average selling price is not about 100,000 dollars at risk level $\alpha = 5\%$

b) The average house area is around 17 hundreds of square feet

$H_0 : \mu = 17$

$H_a : \mu \neq 17$

```
##
##  One Sample t-test
##
## data:  livarea
## t = -1.7963, df = 1499, p-value = 0.07264
## alternative hypothesis: true mean is not equal to 17
## 95 percent confidence interval:
##  16.47003 17.02330
## sample estimates:
## mean of x
##  16.74667
```

Because $p_{value} = 0.07264 > 0.05 \rightarrow$ *Accept* H_0. Therefore, we conclude the average house area is around 17 hundreds of square feet at risk level $\alpha = 5\%$

c) The average area of houses with a size of <=0.5 acres is smaller than the average area of houses with a size of >0.5 acres μ_1 : the average area of houses with a size of <=0.5 acres

μ_2 : the average area of houses with a size of >0.5 acres

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 < \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  livarea[lgelot == "0"] and livarea[lgelot == "1"]
## t = -8.3742, df = 96.846, p-value = 0.0000000000002158
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -6.74862
## sample estimates:
## mean of x mean of y
##  16.21352  24.63158
```

Because $p_{value} = 2.158e^{-13} < 0.05 \rightarrow$ *Reject* H_0. Therefore, we conclude the average area of houses with a size of <=0.5 acres is smaller than the average area of houses with a size of >0.5 acres at risk level $\alpha = 5\%$

d) The average age of houses with swimming pools is equal to the average age of houses without swimming pools. μ_1 : the average age of houses with swimming pools

μ_2 : the average age of houses without swimming pools

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  age[pool == "1"] and age[pool == "0"]
## t = 0.57344, df = 109.82, p-value = 0.5675
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.003482  3.634971
## sample estimates:
## mean of x mean of y
##  22.62245  21.80670
```

Because $p_{value} = 0.5675 > 0.05 \rightarrow$ *Accept* H_0. Therefore, we conclude the average age of houses with swimming pools is equal to the average age of houses without swimming pools at risk level $\alpha = 5\%$

e) The houses with a size >0.5 acres have a higher average selling price compared to houses with a size <=0.5 acres. μ_1 : the average selling price of houses with a size >0.5 acres

μ_2 : the average selling price of houses with a size <=0.5 acres

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 > \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  sprice[lgelot == "1"] and sprice[lgelot == "0"]
## t = 10.206, df = 95.619, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  112023.6      Inf
## sample estimates:
## mean of x mean of y
##  249017.4  115220.0
```

Because $p_{value} < 2.2e^{-16} < 0.05 \rightarrow$ *Reject* H_0. Therefore, we conclude the houses with a size >0.5 acres have a higher average selling price compared to houses with a size <=0.5 acres at risk level $\alpha = 5\%$

f) Houses with a pool have a higher average selling price compared to houses without a pool
μ_1 : the average selling price of houses with a pool

μ_2 : the average selling price of houses without a pool

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 > \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  sprice[pool == "1"] and sprice[pool == "0"]
## t = 5.9191, df = 100.48, p-value = 0.00000002262
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  48184.14      Inf
## sample estimates:
## mean of x mean of y
##  186285.4  119318.7
```

Because $p_{value} = 2.262e^{-08} < 0.05 \rightarrow Reject\ H_0$. Therefore, we conclude the houses with a pool have a higher average selling price compared to houses without a pool at risk level $\alpha = 5\%$

V. Prediction model

a) Eliminate outlier from dataset

As above mention, sprice has 101 outliers acounted for 6.7% in the dataset. We will remove it.

```
## [1] "Length of outliers: "
```

```
## [1] 0.01429593
```

```
## [1] "dimension of current data: "
```

```
## [1] 1399    7
```

After remove the outliers, we have some new outliers acounted for 1.4% which is not significant. Therefore, we can ignore it.

b) Split training and validate set

```
## [1] "train set:"
```

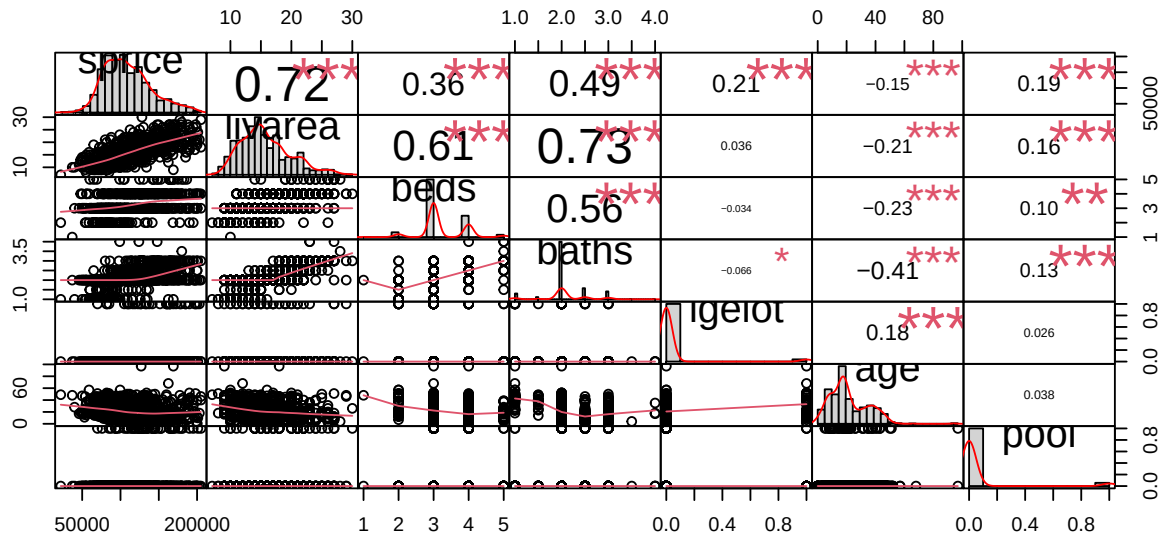
```
##      sprice livarea beds baths lgelot age pool
## 1419  83000      15    3   2.5      0  18    0
## 496   93000      14    3   2.0      0  33    0
## 726   98000      20    4   3.0      0  22    0
## 228   99950      14    3   2.0      0  38    0
## 650   87500      12    3   1.5      0  36    0
## 1088  85000      12    3   2.0      0  16    0
```

```
## [1] "validate set:"
```

```
##    sprice livarea beds baths lgelot age pool
## 1  138000      17    3   2.0      1  97    0
## 2  105700      21    4   2.5      0  18    0
## 9  125000      14    3   2.0      0   3    0
## 12 160000      19    4   2.5      0   4    0
## 13 151000      17    3   2.5      0   0    0
## 14 166000      19    4   2.0      0   0    0
```

We split data into train and test with ratio of 7/3. Almost always the training set is greater than the research set. 70:30 is the most common split ratio used by data scientists. A 70:30 split ratio means that 70% of the knowledge will go to the training set and 30% of the dataset will go to the validation set.

c) Multiple linear regression



Based on the above graph, the correlation of sprice with other variables is not strong. Meanwhile, livarea has a relatively high correlation with other variables. The variables sprice, baths, and beds which are quantifiers have the most linear relationship with the livarea variable. Therefore, we choose living area (hundreds of square feet) is our research purposes. We want to build a model to predict living area (hundreds of square feet).

Model: $\hat{Y} = \hat{B}_0 + \hat{B}_1 sprice + \hat{B}_2 beds + \hat{B}_3 baths$

Three variables have positive correlations with livarea. Among of them, the correlation of beds is the weakest. We will choose these variables to explain livarea. Then we will test if $\hat{B}_2 = 0$.

$H_0 : B_1 = B_2 = B_3 = 0$

$H_a : At\ least\ one\ B_i \neq 0$

```
##
## Call:
## lm(formula = livarea ~ sprice + beds + baths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4930 -1.3901 -0.0508  1.2243  8.8828
##
## Coefficients:
##                  Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) -2.797025150  0.393496495  -7.108     0.00000000000227 ***
## sprice       0.000057242  0.000002418  23.669 < 0.0000000000000002 ***
## beds         1.684164603  0.135532688  12.426 < 0.0000000000000002 ***
## baths        3.283547829  0.189519405  17.326 < 0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.079 on 976 degrees of freedom
## Multiple R-squared:  0.7402, Adjusted R-squared:  0.7394
## F-statistic: 927.1 on 3 and 976 DF,  p-value: < 0.00000000000000022
```

```
## Analysis of Variance Table
##
## Response: livarea
##            Df Sum Sq Mean Sq F value                Pr(>F)
## sprice      1 8390.2  8390.2 1941.90 < 0.00000000000000022 ***
## beds        1 2329.6  2329.6  539.18 < 0.00000000000000022 ***
## baths       1 1297.0  1297.0  300.18 < 0.00000000000000022 ***
## Residuals 976 4216.9     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because $p_{value} < 2.2e^{-16} \rightarrow Reject\ H_0$. This is a highly significant result. The null hypothesis should be rejected at any reasonable significance level. Therefore at least one variable can be used to explain. Moreover, we can see that at risk level $\alpha = 5\%$, sprice , beds, baths have the meaning to explain livarea in this model. Among that p_{value} of beds is smaller than 0.5 then Reject H_0, we can use to explain livarea.

Moreover, $R^2 = 74.02\%$ means that 74.02% of the observed variation in livarea can be explained by the linear regression relationship based on sprice, beds, baths.

After that we have the model: $\hat{Y} = -2.797e^{+00} + 5.724e^{-05}sprice + 1.684e^{+00}beds + 3.284e^{+00}baths$. This model shows that when increasing 1 dollar then living area increases $5.724e^{-05}$ hundreds of square feet; increasing 1 number of beds then living area increases $1.684e^{+00}$ hundreds of square feet; increasing 1 number of baths then living area increases $3.284e^{+00}$ hundreds of square feet.

d) Exam multicollinearity of model:

```
##   Variables Tolerance      VIF
## 1    sprice 0.7443339 1.343483
## 2      beds 0.6784026 1.474051
## 3     baths 0.5902340 1.694243
```

The VIF indexes of three variables sprice,beds, baths are not significant. Therefore we can ignore the multicollinearity of this model.

e) Predict validation data

```
## Root mean square error:  2.326186
```

After having prediction values, we can see that root mean square error is equal to 2.326186. There are still some predicted values that differ significantly from the actual values. However, the model can be considered acceptable for now.

f) Exam the independence of the model

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01417278      2.026274   0.672
##  Alternative hypothesis: rho != 0
```

Because $p_{value} = 0.672 > 0.05 \rightarrow$ *Accept* H_0. Therefore, there is no correlation among the residuals at rist level $\alpha = 5\%$

g) Exam stability of model

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 9.9009, df = 3, p-value = 0.01943
```

Because $p_{value} = 0.01943 < 0.05 \rightarrow$ *reject* H_0. Therefore the variance of the residuals is not constant.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.98954, p-value = 0.000001885
```

When observing above plot, a fan or cone shape indicates the presence of heteroskedasticity. This is seen as a problem because linear regression assumes that the spread of residuals is constant across the plot. If there is an unequal scatter of residuals, the population used in the regression contains unequal variance, and therefore the analysis results may be invalid. As we can see, $p_{value} = 1.885e^{-06} < 0.05 \rightarrow$ *Reject* H_0. Therefore the residuals do not adhere to normal distribution at risk level $\alpha = 5\%$.

VI. Solving heteroskedasticity

a) Transforming the outcome variable

We will transform the livarea by using a log transformation.

```
##
## Call:
## lm(formula = log(livarea) ~ sprice + beds + baths)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50160 -0.08441  0.00522  0.08497  0.56704
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) 1.5969125000 0.0252103386   63.34 <0.0000000000000002 ***
## sprice      0.0000035846 0.0000001549   23.13 <0.0000000000000002 ***
## beds        0.0986148648 0.0086832411   11.36 <0.0000000000000002 ***
## baths       0.2002734697 0.0121420354   16.49 <0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1332 on 976 degrees of freedom
## Multiple R-squared:  0.7223, Adjusted R-squared:  0.7214
## F-statistic: 846.2 on 3 and 976 DF,  p-value: < 0.00000000000000022
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitfix1
## BP = 14.355, df = 3, p-value = 0.002459
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitfix1$residuals
## W = 0.99661, p-value = 0.03317
```

This seems like log model is better than normal model. However, the log transformation did not solve the problem in this case because $p_{value} = 0.002459 < 0.05 \rightarrow Reject\ H_0$ when using studentized Breusch-Pagan test. Therefore problem that the variance of residuals is not constant is still here with the risk level at $\alpha = 5\%$. When using Shapiro-Wilk normality test, $p_{value} = 0.03317 < 0.05 \rightarrow RejectH0$. The residuals do not adhere to normal distribution at risk level $\alpha = 5\%$. However, the residuals adhere to normal distribution at risk level $\alpha = 3\%$

b) Weighted least squares regression

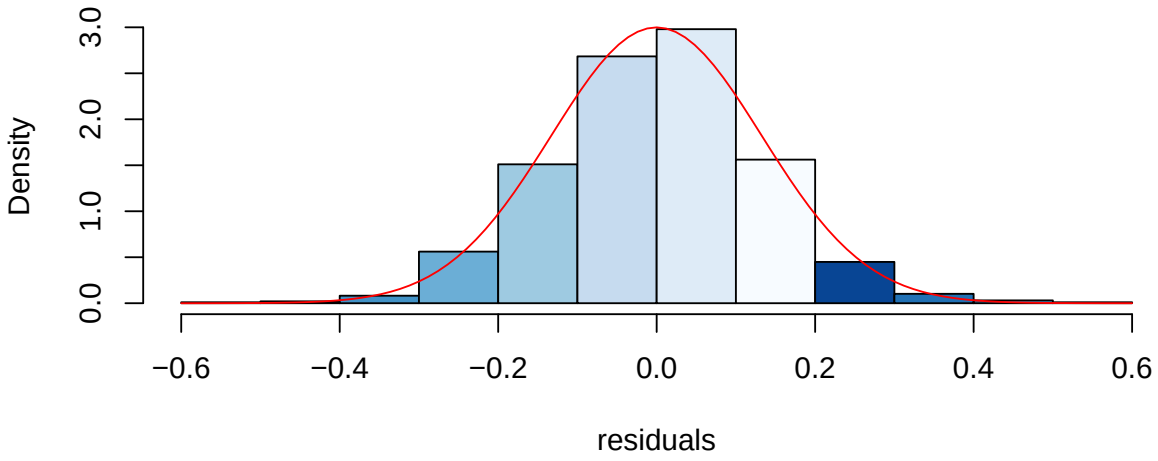
Weighted least squares regression is a generalization of ordinary least squares (OLS) and linear regression in which the unequal variance of observations (heteroscedasticity) is incorporated into the regression. In WLS, each observation is weighted by the reciprocal of its variance. This means that observations with smaller variances, which contain more information, are given more weight in the regression. The weights are used to construct a weight matrix, which is used to modify the normal equations of the OLS method to obtain the WLS estimates.

```
##
##  studentized Breusch-Pagan test
##
## data:  fitfix2
## BP = 3.2439, df = 3, p-value = 0.3555
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitfix2$residuals
## W = 0.99649, p-value = 0.02724
```

```
##
## Call:
## lm(formula = log(livarea) ~ sprice + beds + baths, weights = 1/wt)
##
## Weighted Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.053929 -0.008931  0.000680  0.008554  0.071435
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) 1.5899593067 0.0261919493   60.70 <0.0000000000000002 ***
## sprice      0.0000037038 0.0000001625   22.80 <0.0000000000000002 ***
## beds        0.0989562307 0.0089557488   11.05 <0.0000000000000002 ***
## baths       0.1968430531 0.0121988256   16.14 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01408 on 976 degrees of freedom
## Multiple R-squared:    0.7,  Adjusted R-squared:  0.6991
## F-statistic: 759.1 on 3 and 976 DF,  p-value: < 0.00000000000000022
```



As we can see when using studentized Breusch-Pagan test, $p_{value} = 0.3555 > 0.05 \rightarrow Accept\ H_0$. Therefore we can think that the variance of the residuals is constant at risk level $\alpha = 5\%$. However, when using Shapiro-Wilk normality test, $p_{value} = 0.02724 < 0.05 \rightarrow Reject\ H_0$. The residuals do not adhere to normal distribution at risk level $\alpha = 5\%$. However, the residuals adhere to normal distribution at risk level $\alpha = 2\%$. As we can see, the histogram is like having the normal distribution.

```
## Root mean square error: 2.500965
```

Compare to model 1, RMSE is approximately equal.

c) Conclude

All three models does not have residuals that adhere to normal distribution at risk level $\alpha = 5$. Among of them, model 3 is the best model that satisfied no correlation among the residuals, the variance of the residuals is constant, residuals that adhere to normal distribution at risk level lower. The problem may arise due to the influence of outliers in the variables even though they have been processed, or simply because the initial assumption of using a linear regression model to handle this dataset is not appropriate.

d) Choose other models with AIC standard

The best model is $\hat{livarea} = B_0 + B_1 sprice + B_2 beds + B_3 baths + B_4 lgelot + B_5 age$ according to AIC standard. As shown by above plot, age and lgelot do not have correlations with livarea. Let's exam we can drop it.

$H_0 : \hat{B_4} = \hat{B_5} = 0$

$H_a : \exists B_i \neq 0$

```
## Analysis of Variance Table
##
## Model 1: livarea ~ sprice + beds + baths
## Model 2: livarea ~ sprice + beds + baths + lgelot + age
##   Res.Df    RSS Df Sum of Sq     F      Pr(>F)
## 1    976 4216.9
## 2    974 4105.0  2    111.94 13.28 0.000002041 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because $p_{value} = 0.000002041 < 0.05 \rightarrow Reject\ H_0$. Therefore, we can drop age and lgelot variables for this case.

e) Exam AIC model

```
##
## Call:
## lm(formula = train$livarea ~ train$sprice + train$beds + train$baths +
##     train$lgelot + train$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0995 -1.3897 -0.0748  1.2140  8.4905
##
## Coefficients:
##                   Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  -3.995036902  0.463922547  -8.611 < 0.0000000000000002 ***
## train$sprice  0.000058334  0.000002494  23.392 < 0.0000000000000002 ***
## train$beds    1.681351070  0.133937491  12.553 < 0.0000000000000002 ***
## train$baths   3.540180630  0.200511214  17.656 < 0.0000000000000002 ***
## train$lgelot -0.929242160  0.369184659  -2.517                0.012 *
## train$age     0.026609863  0.005505657   4.833          0.00000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.053 on 974 degrees of freedom
## Multiple R-squared:  0.7471, Adjusted R-squared:  0.7458
## F-statistic: 575.6 on 5 and 974 DF,  p-value: < 0.00000000000000022
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01297797      2.024146   0.682
##  Alternative hypothesis: rho != 0
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bestmodel
## BP = 72.47, df = 5, p-value = 0.00000000000003135
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bestmodel$residuals
## W = 0.99042, p-value = 0.000005267
```

```
## [1] 5.246857
```

$p_{value} = 0.682 > 0.05 \rightarrow Accept\ H_0$ when using Durbin Watson Test. Therefore, there is no correlation among the residuals at rist level $\alpha = 5\%$

$p_{value} = 0.00000000000003135 < 0.05 \rightarrow Reject\ H_0$ when using studentized Breusch-Pagan test, so the variance of the residuals is not constant.

$p_{value} = 0.000005267 < 0.05 \rightarrow Reject\ H_0$ when using Shapiro-Wilk normality test, so the residuals does not adhere to normal distribution. Finally, root mean square error is equal to 5.246857 worse than above model.

f) AIC model with weight

```
##
##  studentized Breusch-Pagan test
##
## data:  bestmodel3
## BP = 68.275, df = 5, p-value = 0.0000000000002341
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bestmodel3$residuals
## W = 0.99061, p-value = 0.000006667
```

```
## Root mean square error:  5.20272
```

$p_{value} = 0.0000000000002341 < 0.05 \rightarrow Reject\ H_0$ when using studentized Breusch-Pagan test, so the variance of the residuals is not constant.

$p_{value} = 0.000006667 < 0.05 \rightarrow RejectH_0$ when using Shapiro-Wilk normality test, so the residuals does not adhere to normal distribution. Finally, root mean square error is equal to 5.20272 worse than above model.

g) log(model)

```
##
## Call:
## lm(formula = log(train$livarea) ~ train$sprice + train$beds +
##     train$baths + train$lgelot + train$age)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51873 -0.08711  0.00258  0.08529  0.53674
##
## Coefficients:
##                   Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)   1.5433412187  0.0298425517  51.716 < 0.0000000000000002 ***
## train$sprice  0.0000036954  0.0000001604  23.036 < 0.0000000000000002 ***
## train$beds    0.0981066952  0.0086157410  11.387 < 0.0000000000000002 ***
## train$baths   0.2093325280  0.0128982010  16.230 < 0.0000000000000002 ***
## train$lgelot -0.0735046903  0.0237483872  -3.095             0.002023 **
## train$age     0.0012159632  0.0003541601   3.433             0.000621 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1321 on 974 degrees of freedom
## Multiple R-squared:  0.7275, Adjusted R-squared:  0.7261
## F-statistic:   520 on 5 and 974 DF,  p-value: < 0.00000000000000022


##
##  studentized Breusch-Pagan test
##
## data:  bestmodel1
## BP = 61.283, df = 5, p-value = 0.0000000000066


##
##  Shapiro-Wilk normality test
##
## data:  bestmodel1$residuals
## W = 0.99672, p-value = 0.03995


## Root mean square error:  5.320361
```

$p_{value} = 0.0000000000066 < 0.05 \rightarrow Reject\ H_0$ when using studentized Breusch-Pagan test, so the variance of the residuals is not constant.

$p_{value} = 0.03995 < 0.05 \rightarrow Reject\ H_0$ when using Shapiro-Wilk normality test, so the residuals does not adhere to normal distribution. Finally, root mean square error is equal to 5.320361 worse than above model.

h) log(model) with weight

```
##
## Call:
## lm(formula = log(train$livarea) ~ train$sprice + train$beds +
##     train$baths + train$lgelot + train$age, weights = wt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8708 -0.8453  0.0330  0.8347  4.4240
##
## Coefficients:
##                   Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)   1.5677783610  0.0284084055  55.187 < 0.0000000000000002 ***
```

```
## train$sprice  0.0000035896  0.0000001542  23.275 < 0.0000000000000002 ***
## train$beds    0.0973499822  0.0083626076  11.641 < 0.0000000000000002 ***
## train$baths   0.2065153456  0.0125159767  16.500 < 0.0000000000000002 ***
## train$lgelot -0.0694002316  0.0231447659  -2.999              0.00278 **
## train$age     0.0010167395  0.0003446769   2.950              0.00326 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.277 on 974 degrees of freedom
## Multiple R-squared:  0.7444, Adjusted R-squared:  0.7431
## F-statistic: 567.3 on 5 and 974 DF,  p-value: < 0.00000000000000022


##
##  studentized Breusch-Pagan test
##
## data:  bestmodel2
## BP = 93875, df = 5, p-value < 0.00000000000000022


##
##  Shapiro-Wilk normality test
##
## data:  bestmodel1$residuals
## W = 0.99672, p-value = 0.03995


## Root mean square error:  5.276937
```

$p_{value} < 0.00000000000000022 < 0.05 \rightarrow Reject\ H_0$ when use studentized Breusch-Pagan test, so the variance of the residuals is not constant. $p_{value} = 0.03995 < 0.05 \rightarrow Reject\ H_0$ when using Shapiro-Wilk normality test, so the residuals does not adhere to normal distribution. Finally, root mean square error is equal to 5.276937 worse than above model.

i) Boxcox transformation

Box-Cox transformation is a statistical technique that involves transforming your target variable so that your data follows a normal distribution. Box-Cox transformation helps to improve the predictive power of your analytical model because it cuts away white noise.The basic idea behind this method is to find some value for λ such that the transformed data is as close to normally distributed as possible, using the following formula:

$$\begin{cases} y(\lambda) = \frac{y^\lambda - 1}{\lambda}, \lambda \neq 0 \\ y(\lambda) = log(y), \lambda = 0 \end{cases}$$

We examed with $\lambda = 0$, now let's check with $\lambda \neq 0$:

```
## Model with good correlation and lambda =   1.272727


##
##  studentized Breusch-Pagan test
##
## data:  model1lam
## BP = 18.993, df = 3, p-value = 0.0002743
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1lam$residuals
## W = 0.98464, p-value = 0.00000001248

## Root mean square error:  0.02537233
```

Although the root mean square error is quite small, indicating that the model has a good fit. However, two issues, the variance of the residuals is not constant and residuals do not adhere to normal distribution, still remain, leading to the possibility that the model's results may not be accurate.

Conclusion for the final model

Model 3 is the best model that satisfied no correlation among the residuals, the variance of the residuals is constant, residuals that adhere to normal distribution at risk level $\alpha = 2\%$. The remaining models violate one or both of the above conditions.

```
##
## Call:
## lm(formula = log(livarea) ~ sprice + beds + baths, weights = 1/wt)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.053929 -0.008931  0.000680  0.008554  0.071435
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) 1.5899593067 0.0261919493   60.70 <0.0000000000000002 ***
## sprice      0.0000037038 0.0000001625   22.80 <0.0000000000000002 ***
## beds        0.0989562307 0.0089557488   11.05 <0.0000000000000002 ***
## baths       0.1968430531 0.0121988256   16.14 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01408 on 976 degrees of freedom
## Multiple R-squared:    0.7,  Adjusted R-squared:  0.6991
## F-statistic: 759.1 on 3 and 976 DF,  p-value: < 0.00000000000000022
```

$R^2 = 70\%$ means that 70% of the observed variation in livarea can be explained by the model based on sprice, beds, baths.

The model 3: $\hat{livarea} = e^{1.5899593067} e^{0.0000037038 sprice} e^{0.0989562307 beds} e^{0.1968430531 baths}$. When the selling price increases by 1 dollar, the living area increases by 1.000003704 times, when the number of bedrooms increases by 1, the living area increases by 1.104017976 times, when the number of bathrooms increases by 1, the living area increases by 1.217552935 times.

Activity 2

Introduction

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made

available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. This dataset is to be used for Machine Learning and Data Visualization purposes.

About data: This dataset contains 10 columns in 22 columns of the big dataset, including Status, Life expectancy, Adult Mortality, Alcohol, Total expenditure, Hepatitis B, BMI, Income composition of resources, Schooling, Population.

Dataset includes 2938 observations of 10 variables:

- Quantifier: Life_expectancy, Adult_Mortality, Alcohol, Total_expenditure, Hepatitis_B, BMI, Income_composition_of_resources, Schooling, Population
- Qualitative: Status

Variable	Definition
Life_expectancy	Life Expectancy in age (years)
Adult_Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)(%)
Alcohol	recorded per capita (15+) consumption (in litres of pure alcohol)
Total_expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Hepatitis_B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
BMI	Average Body Mass Index of entire population
Income_composition_of_resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of Schooling(years)
Population	Population of the country
Status	Developed or Developing status

Dataset Source: Life Expectancy WHO is taken from Kaggle and helps in predicting life expectancy with the help of various factors for a period of 15 years.

Purposes of research:

- Testing whether the average life expectancy of developed countries is greater than that of developing countries
- Build a model to predict average life expectancy based on factors affecting life expectancy

I. Clean dataset

```
## [1] "NA value:"
```

```
## Status 0  -  Life_expectancy 10  -  Adult_Mortality 10  -  Alcohol 194
```

```
## Hepatitis_B 553  -  BMI 34  -  Income_composition_of_resources 167
```

```
## Schooling 163  -  Total_expenditure 226  -  Population 652
```

There are many rows containing NA-value in the dataset. Among them, Hepatitis B and Population have the highest NA numbers, respectively 553 and 652. Therefore we will replace them by mean value before going to the next step. After that, we have the new dataset not containing any NA-value.

II. Preprocess outliers

```
## [1] "Outliers:"
```

```
## Life_expectancy  17  -  Adult_Mortality  86  -  Alcohol  3
```

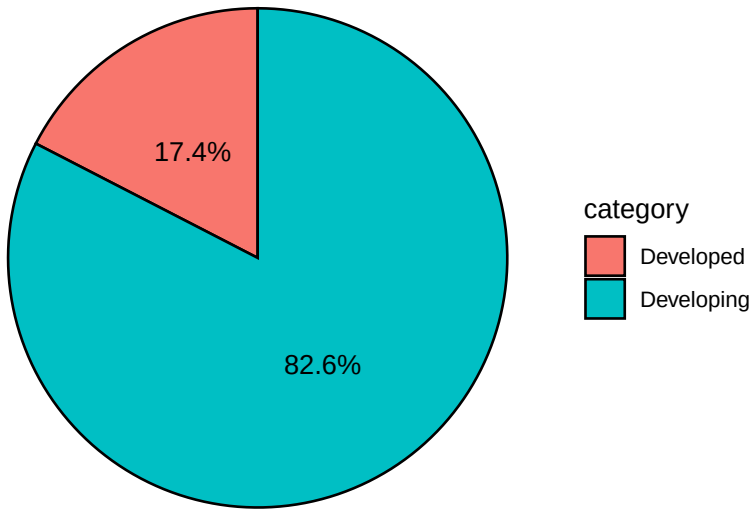
```
## Hepatitis_B  316  -  BMI  1  -  Income_composition_of_resources  130
```

```
## Schooling  77  -  Total_expenditure  51  -  Population 194
```

It can be seen that each variable has outliers. Among them, the variable serving the research purpose, life expectancy, has only 17 outliers, accounting for a very small number compared to the size of the data. The variables Income composition of resources, Hepatitis B, Population have outlier numbers exceeding 100, respectively 130, 316, 194. We will solve it after Descriptive Statistics section.

III. Descriptive Statistics

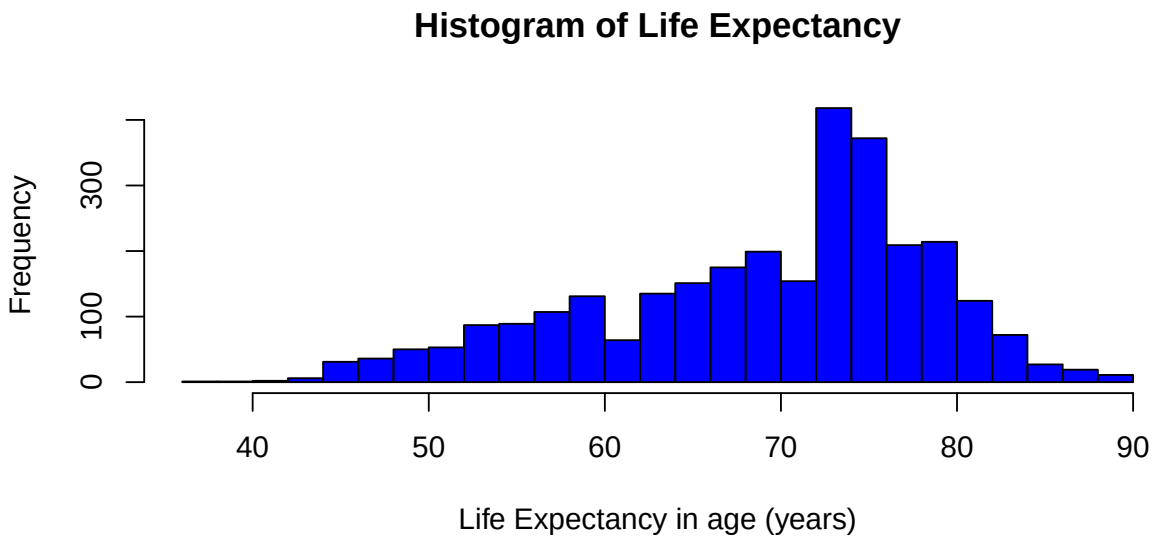
a) Status



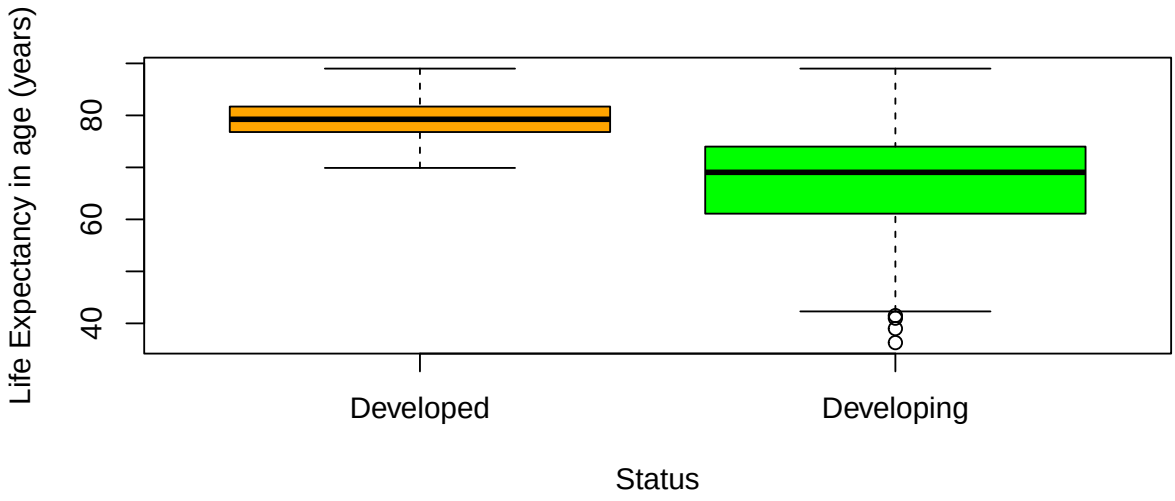
Comment: Developing countries make up the majority of the data (2409), while developed countries (512) make up a relatively portion.

We will set Status variable with 1 for Developed and 0 for Developing. Converting the Status variable into a dummy variable will help build the model later.

b) Life Expectancy in age (years)



Comment: The lifespan falls mostly in the range of 65 to 80 years, with the most common age being around 73 years old. The graph appears to be left-skewed, indicating that the majority have a high average age.



Comment: Based on the graph, the average lifespan of developing countries is lower than that of developed countries. The average lifespan of developing countries has a large fluctuation, while the average lifespan of developed countries is narrow and not significantly fluctuating.

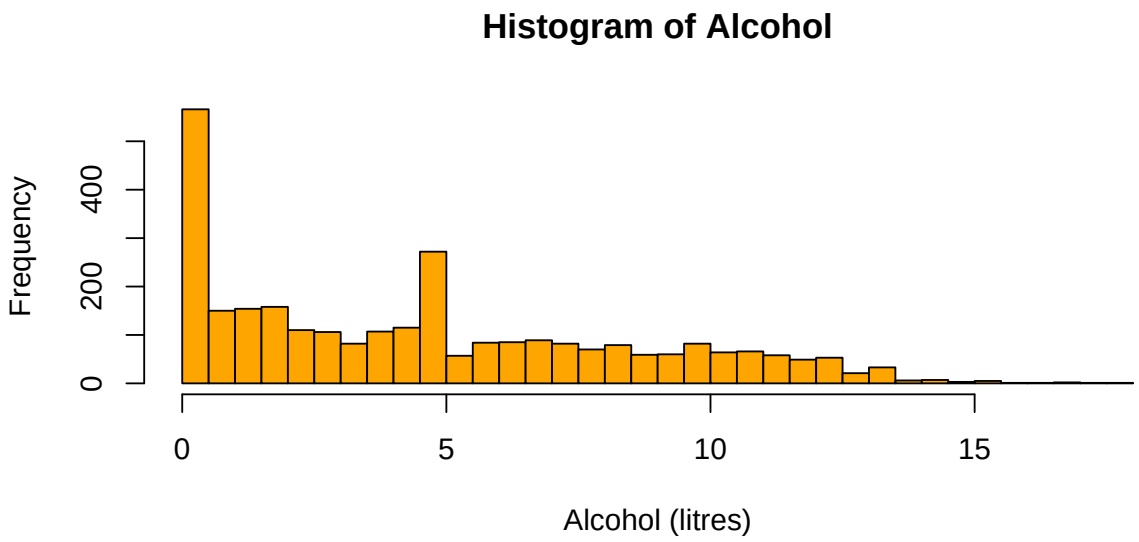
c) Adult Mortality

Draw the same as above one, based on **Figure 3** the Adult Mortality Rates of both sexes are widely distributed and skewed to the right, meaning that the small Adult Mortality Rates of both sexes account for

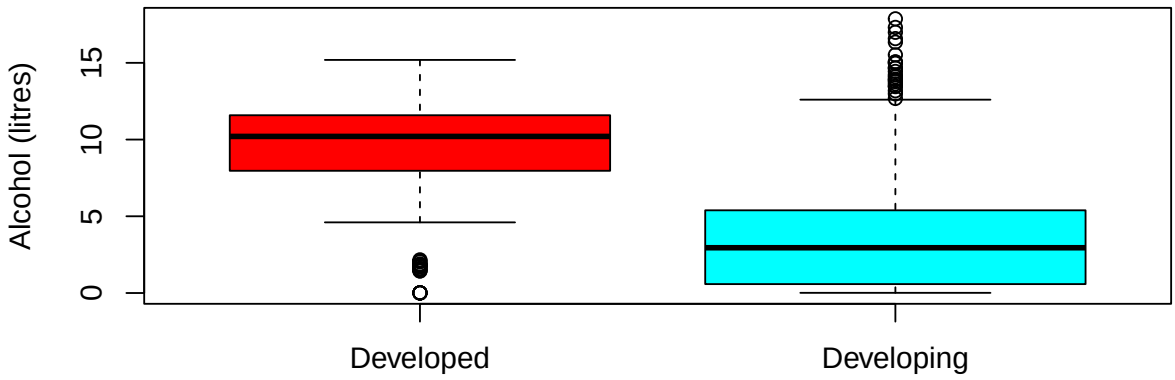
a large number. In particular, the range from 0 to 200% accounts for a large number but also fluctuates, not evenly high. From about 200%, the distribution is sparse and the number decreases as the Adult Mortality Rates of both sexes increase.

Base on **Figure 4** The Adult Mortality Rates of both sexes have outliers in both developed and developing countries. Developed countries have narrow Adult Mortality Rates of both sexes and do not fluctuate much, while developing countries fluctuate a lot. Looking at the figure, the Adult Mortality Rates of both sexes in developed countries are smaller than those in developing countries.

d) Alcohol(litres)



The histogram is quite evenly distributed, with only two peaks of consumption (in litres of pure alcohol) being almost non-consumption or consuming about 5 liters.



Both consumption alcohol in liters of developing countries and developed countries have outliers. Especially, consumption (in litres of pure alcohol) in developing countries is lower than consumption (in litres of pure alcohol) in developed countries, although both fluctuate quite a lot.

e) BMI

Based on **Figure 5** , the histogram of BMI is widely distributed but not even, with many peaks but the highest peak is around 58 to 60.

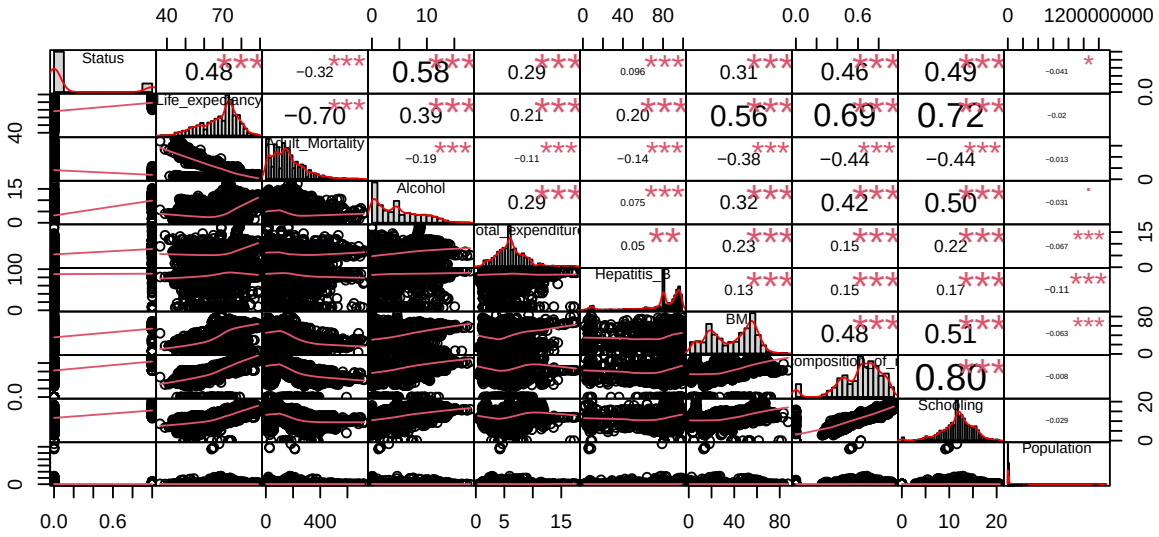
Based on **Figure 6**, the Average Body Mass Index of the entire population of developing countries does not have outliers, but the Average Body Mass Index of the entire population of developed countries has a large number of outliers. However, the Average Body Mass Index of the entire population of developing countries fluctuates greatly and widely, while the Average Body Mass Index of the entire population of developed countries fluctuates narrowly and insignificantly.

f) Schooling

Based on **Figure 7**, the highest number of years of schooling is about 12 years. The chart has a number of countries with almost 0 years of schooling. The highest number of years of schooling falls between 11 and 14 years.

Based on **Figure 8**, as predicted, developing countries have a lower average number of years of schooling than the average number of years of schooling in developed countries. The number of years of schooling in both developed and developing countries has outliers and insignificant fluctuations, but the fluctuations in developing countries are greater.

g) Analysis correlation



Comment: Based on the graph, life expectancy has a strong correlation with Adult_Mortality, Income_composition_of_resources, and Schooling, and has a moderate correlation with BMI. Among these, Adult_Mortality has a negative correlation while the rest have positive correlations. However, Schooling and Income_composition_of_resources have a strong correlation with each other. We will check whether Schooling and Income_composition_of_resources are two independent variables or not. Furthermore,

although Hepatitis_B and Population have a large number of outliers, they have almost no correlation with other variables. Therefore, we do not need to remove the outliers of Hepatitis_B and Population. Because Adult_Mortality, Income_composition_of_resources, Schooling have relative number of outliers and they have good correlation to predict life expectancy, we also replace them by its mean value.

IV. Testing hypothesis

a) The average lifespan of developing countries is lower than that of developed countries

μ_1 : The average lifespan of developing countries

μ_2 : The average lifespan of developed countries

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 < \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  data2$Life_expectancy[data2$Status == 0] and data2$Life_expectancy[data2$Status == 1]
## t = -47.935, df = 1799.3, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -11.66302
## sample estimates:
## mean of x mean of y
##  67.12018  79.19785
```

Because $p_{value} < 2.2e^{-16} \rightarrow RejectH_0$. We can reject null hypothesis at risk level 5%. Therefore, we can conclude that the average lifespan of developing countries is lower than that of developed countries.

b) The average consumption alcohol in developing countries is lower than the consumption alcohol in developed countries

μ_1 : The average consumption alcohol in developing countries

μ_2 : The average consumption alcohol in developed countries

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 < \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  data2$Alcohol[data2$Status == 0] and data2$Alcohol[data2$Status == 1]
## t = -41.055, df = 796.1, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -5.740503
## sample estimates:
## mean of x mean of y
##  3.560670  9.541055
```

Because $p_{value} < 2.2e^{-16} \rightarrow RejectH_0$. We can reject null hypothesis at risk level 5%. Therefore, we can conclude that the average consumption alcohol in developing countries is lower than the consumption alcohol in developed countries.

c) Schooling and Income_composition_of_resources are two independent variables

H_0: two variables are independent

H_a: two variables are not independent

```
##
##  Pearson's Chi-squared test
##
## data:  data2$Schooling and data2$Income_composition_of_resources
## X-squared = 157080, df = 108125, p-value < 0.00000000000000022
```

Because $p_{value} < 2.2e^{-16} \rightarrow Reject\ H_0$. We can reject null hypothesis at risk level 5%. Therefore we can conclude that Schooling and Income_composition_of_resources are not two independent variables.

V. Prediction model

a) Proccess outliers after splitting

We removed the outliers of life expectancy, and replaced outliers of Adult_Mortality, Income_composition_of_resources, Schooling, Hepatitis_B, Population by their means. After that, we remain 2921 observations. Then, we splitting with ratio of 7/3 because this is the ratio commonly recommended for data analysis

b) Model with predictors having good correlations

We will choose Adult_Mortality, BMI, Schooling to be the predictors. Schooling and Income_composition_of_resources are not independent variables, so we only choose one of them. We will choose Schooling because it has stronger correlation than Income_composition_of_resources.Moreover, I want use dummy variable Status to the model.

```
##
## Call:
## lm(formula = train2$Life_expectancy ~ Status + Adult_Mortality +
##     BMI + Schooling, data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.5416  -2.3994   0.6814   3.4132  16.8419
##
## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     52.777271   0.725342  72.762 < 0.0000000000000002 ***
## Status           2.878414   0.366483   7.854  0.00000000000000645 ***
## Adult_Mortality -0.027476   0.001359 -20.222 < 0.0000000000000002 ***
## BMI              0.074892   0.007407  10.110 < 0.0000000000000002 ***
## Schooling        1.434586   0.057009  25.164 < 0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.474 on 2040 degrees of freedom
## Multiple R-squared:  0.6557, Adjusted R-squared:  0.655
## F-statistic: 971.1 on 4 and 2040 DF,  p-value: < 0.00000000000000022
```

```
##         Variables Tolerance      VIF
## 1          Status 0.7425452 1.346719
## 2 Adult_Mortality 0.7450639 1.342167
## 3             BMI 0.6666680 1.499997
## 4       Schooling 0.5489763 1.821572
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04793705      1.904081   0.026
##  Alternative hypothesis: rho != 0
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelbestnew
## BP = 236.92, df = 4, p-value < 0.00000000000000022
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelbestnew$residuals
## W = 0.93584, p-value < 0.00000000000000022
```

```
## Root mean square error:  0.169167
```

Because $p_{value} < 2.2e^{-16} \rightarrow Reject\ H_0$.So the null hypothesis that all four $B_i's$ corresponding to predictors have value 0 is resoundingly rejected. There appears to be a useful relationship between the dependent variable and at least one of the predictors. We can see p_{value} of 4 variables are too small, so they can be use to explain life expantancy.

Moreover, $R^2 = 65.57\%$ means that 65.57% of the observed variation in life_expectancy can be explained by the linear regression relationship based on Status, Adult_Mortality, BMI, Schooling.

The VIF indexes of three variables Status, Adult_Mortality, BMI, Schooling are not significant. Therefore we can ignore the multicollinearity of this model.

Because $p_{value} = 0.026 < 0.05 \rightarrow Reject\ H_0$ when using Durbin Watson Test. Therefore, there is correlation among the residuals at rist level $\alpha = 5\%$

Because $p_{value} < 2.2e^{-16} \rightarrow\ reject\ H_0$ with studentized Breusch-Pagan test. Therefore the variance of the residuals is not constant at any risk level.

Because $p_{value} < 2.2e^{-16} \rightarrow\ reject\ H_0$ with Shapiro-Wilk normality test. Therefore the residuals of model do not adhere to normal distribution at any risk level.

After having prediction values, we can see that root mean square error is equal to 0.169167. There are still some predicted values that differ significantly from the actual values.

c) Variable Selection

The best model by AIC criterion is $\hat{Life_expectancy} = B_0 + B_1\hat{Status} + B_2\hat{Adult\_Mortality} + B_3\hat{Alcohol} + B_4\hat{Total\_expenditure} + B_5\hat{Hepatitis\_B} + B_6\hat{BMI} + B_7\hat{Income\_composition\_of\_resources} + B_8\hat{Schooling} + B_9\hat{Population}$

Then we exam the hypothesis that:

$H_0 : B_3 = B_4 = B_5 = B_7 = B_9 = 0$

$H_a : \exists B_i \neq 0$

```
## Analysis of Variance Table
##
## Model 1: train2$Life_expectancy ~ Status + Adult_Mortality + BMI + Schooling
## Model 2: train2$Life_expectancy ~ Status + Adult_Mortality + Alcohol +
##     Total_expenditure + Hepatitis_B + BMI + Income_composition_of_resources +
##     Schooling + Population
##   Res.Df   RSS Df Sum of Sq      F               Pr(>F)
## 1   2040 61136
## 2   2035 44981  5     16156 146.18 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because $p_{value} < 2.2e^{-16} \rightarrow RejectH_0$. Therefore we can choose full model to test.

```
##
## Call:
## lm(formula = train2$Life_expectancy ~ Status + Adult_Mortality +
##     Alcohol + Total_expenditure + Hepatitis_B + BMI + Income_composition_of_resources +
##     Schooling + Population, data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8746  -2.0372   0.3155   2.7163  17.5768
##
## Coefficients:
##                                    Estimate     Std. Error t value
## (Intercept)                    38.78887541143  1.18764579615  32.660
## Status                          1.94531506043  0.36498899155   5.330
## Adult_Mortality                -0.01857857493  0.00121966568 -15.233
## Alcohol                        -0.12917773621  0.03586498833  -3.602
## Total_expenditure               0.27877673670  0.04731864508   5.891
## Hepatitis_B                     0.06295219411  0.01234678731   5.099
## BMI                             0.03331404603  0.00660521576   5.044
## Income_composition_of_resources 40.53173563644  1.59909193957  25.347
## Schooling                      -0.12846744781  0.07842273688  -1.638
## Population                      0.00000004024  0.00000001588   2.535
##                                           Pr(>|t|)
## (Intercept)                    < 0.0000000000000002 ***
## Status                                 0.00000010922 ***
## Adult_Mortality                < 0.0000000000000002 ***
## Alcohol                                    0.000324 ***
```

```
## Total_expenditure                      0.00000000447 ***
## Hepatitis_B                            0.00000037373 ***
## BMI                                    0.00000049737 ***
## Income_composition_of_resources < 0.0000000000000002 ***
## Schooling                                   0.101547
## Population                                  0.011326 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.701 on 2035 degrees of freedom
## Multiple R-squared:  0.7467, Adjusted R-squared:  0.7455
## F-statistic: 666.4 on 9 and 2035 DF,  p-value: < 0.00000000000000022
```

```
##                         Variables Tolerance      VIF
## 1                          Status 0.5521572 1.811078
## 2                 Adult_Mortality 0.6819744 1.466331
## 3                         Alcohol 0.5505628 1.816323
## 4               Total_expenditure 0.8232627 1.214679
## 5                     Hepatitis_B 0.8627911 1.159029
## 6                             BMI 0.6183991 1.617079
## 7 Income_composition_of_resources 0.1789530 5.588060
## 8                       Schooling 0.2139667 4.673625
## 9                      Population 0.9848607 1.015372
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.03200149      1.935791    0.17
##  Alternative hypothesis: rho != 0
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelbestnew
## BP = 199.4, df = 9, p-value < 0.00000000000000022
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelbestnew$residuals
## W = 0.94109, p-value < 0.00000000000000022
```

```
## Root mean square error:  0.1476864
```

Because $p_{value} < 2.2e^{-16} \rightarrow Reject\ H_0$.So the null hypothesis that all nine $B_i's$ corresponding to predictors have value 0 is resoundingly rejected. There appears to be a useful relationship between the dependent variable and at least one of the predictors. We can see p_{value} of Schooling $= 0.101547 > 0.5$. Therefore Schooling is not significant in this model to explain life expectancy at risk level $\alpha = 5\%$.

Moreover, $R^2 = 74.67\%$ means that 74.67% of the observed variation in life_expectancy can be explained by the linear regression relationship based on Status, Adult_Mortality, Alcohol, Total_expenditure, Hepatitis_B, BMI, Income_composition_of_resources, Schooling, Population.

The VIF indexes of these variables are not significant except for Income_composition_of_resources and Schooling as above explain.

Because $p_{value} = 0.017 < 0.05 \rightarrow Reject\ H_0$ when using Durbin Watson Test. Therefore, there is correlation among the residuals at risk level $\alpha = 5\%$

Because $p_{value} < 2.2e^{-16} \rightarrow\ reject\ H_0$ with studentized Breusch-Pagan test. Therefore the variance of the residuals is not constant at any risk level.

Because $p_{value} < 2.2e^{-16} \rightarrow\ reject\ H_0$ with Shapiro-Wilk normality test. Therefore the residuals of model do not adhere to normal distribution at any risk level.

After having prediction values, we can see that root mean square error is equal to 0.1476864. There are still some predicted values that differ significantly from the actual values but better than above model.

We will drop Schooling:

```
##
## Call:
## lm(formula = train2$Life_expectancy ~ Status + Adult_Mortality +
##     Alcohol + Total_expenditure + Hepatitis_B + BMI + Income_composition_of_resources +
##     Population, data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.0189  -2.0886   0.2891   2.7851  17.5726
##
## Coefficients:
##                                      Estimate     Std. Error t value
## (Intercept)                     38.69644839711  1.18679516855  32.606
## Status                           1.94338904934  0.36513796480   5.322
## Adult_Mortality                 -0.01864939394  0.00121940310 -15.294
## Alcohol                         -0.13545573652  0.03567438287  -3.797
## Total_expenditure                0.26605231373  0.04669603876   5.698
## Hepatitis_B                      0.06236427183  0.01234667139   5.051
## BMI                              0.03274085878  0.00659866793   4.962
## Income_composition_of_resources 38.57576207668  1.06411699525  36.251
## Population                       0.00000004065  0.00000001588   2.560
##                                              Pr(>|t|)
## (Intercept)                     < 0.0000000000000002 ***
## Status                                 0.0000001137 ***
## Adult_Mortality                 < 0.0000000000000002 ***
## Alcohol                                0.000151 ***
## Total_expenditure                     0.0000000139 ***
## Hepatitis_B                           0.0000004784 ***
## BMI                                   0.0000007566 ***
## Income_composition_of_resources < 0.0000000000000002 ***
## Population                             0.010546 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.703 on 2036 degrees of freedom
## Multiple R-squared:  0.7463, Adjusted R-squared:  0.7453
## F-statistic: 748.7 on 8 and 2036 DF,  p-value: < 0.00000000000000022
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.03269154      1.934435   0.138
##  Alternative hypothesis: rho != 0
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelbestnew
## BP = 198.89, df = 8, p-value < 0.00000000000000022


##
##  Shapiro-Wilk normality test
##
## data:  modelbestnew$residuals
## W = 0.94126, p-value < 0.00000000000000022


## Root mean square error:  0.1552736
```

$R^2 = 74.63\%$ means that 74.63% of the observed variation in life_expectancy can be explained by the linear regression relationship based on Status, Adult_Mortality, Alcohol, Total_expenditure, Hepatitis_B, BMI, Income_composition_of_resources, Population.

Root mean square error is equal to 0.1552736. This model does not differ much from the above model.

Clearly, after removing, this model is better than above model. However two problems (the variance of the residuals is not constant and the residuals of model do not adhere to normal distribution) remain. While $p_{value} = 0.138 > 0.5 \rightarrow Accept\ H_0$, when using Durbin Watson Test. Therefore, there is no correlation among the residuals at rist level $\alpha = 5\%$.

VI. Recommend different model

a) log(model) with the highest correlation

Because Adult_Mortality and Schooling have the highest correlation with life expectancy and there are two kind of country in this dataset. We will build the model from this.

```
##
## Call:
## lm(formula = log(train2$Life_expectancy) ~ Status + Adult_Mortality +
##     Schooling, data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45473 -0.03247  0.01795  0.05248  0.24246
##
## Coefficients:
##                    Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)      3.98261888  0.01188173 335.188 < 0.0000000000000002 ***
## Status           0.03542545  0.00600959   5.895        0.00000000438 ***
## Adult_Mortality -0.00045350  0.00002186 -20.742 < 0.0000000000000002 ***
## Schooling        0.02557722  0.00085361  29.964 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08982 on 2041 degrees of freedom
## Multiple R-squared:  0.6044, Adjusted R-squared:  0.6038
## F-statistic:  1039 on 3 and 2041 DF,  p-value: < 0.00000000000000022
```

```
##         Variables Tolerance      VIF
## 1          Status 0.7434347 1.345108
## 2 Adult_Mortality 0.7746403 1.290922
## 3       Schooling 0.6592086 1.516971
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04926737      1.901461    0.02
##  Alternative hypothesis: rho != 0
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelbestnew2
## BP = 150.26, df = 3, p-value < 0.00000000000000022
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelbestnew2$residuals
## W = 0.88329, p-value < 0.00000000000000022
```

```
## Root mean square error:  0.002763509
```

Because $p_{value} < 2.2e^{-16} \rightarrow Reject\ H_0$.So the null hypothesis that all three $B_i's$ corresponding to predictors have value 0 is resoundingly rejected. There appears to be a useful relationship between the dependent variable and at least one of the predictors. We can see p_{value} of 3 variables are too small, so they can be use to explain log(life_expantancy).

The VIF indexes of three variables Status, Adult_Mortality, Schooling are not significant. Therefore we can ignore the multicollinearity of this model.

Moreover, $R^2 = 60.44\%$ means that 60.44% of the observed variation in log(life_expectancy) can be explained by the linear regression relationship based on Status, Adult_Mortality, Schooling.

After having prediction values, we can see that root mean square error is equal to 0.002763509 that is near 0.

This model is better than above model. However three problems (there is correlation among the residuals, the variance of the residuals is not constant and the residuals of model do not adhere to normal distribution) remain.

b) log(model) with full model

We can see that Population does not have correlation with life expectancy. Although processing outliers, Total_expenditure has the most outliers. Therefore, we will remove them and Schooling as correlation with Income_composition_of_resources.

```
##
## Call:
## lm(formula = log(train2$Life_expectancy) ~ Status + Adult_Mortality +
##     Alcohol + Hepatitis_B + BMI + Income_composition_of_resources,
##     data = train2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.36295 -0.03059  0.00690  0.04443  0.24195
##
## Coefficients:
##                                   Estimate  Std. Error t value
## (Intercept)                     3.77448386  0.01854327 203.550
## Status                          0.02470471  0.00586201   4.214
## Adult_Mortality                -0.00027642  0.00001972 -14.014
## Alcohol                        -0.00194775  0.00056957  -3.420
## Hepatitis_B                     0.00111065  0.00020009   5.551
## BMI                             0.00065968  0.00010558   6.248
## Income_composition_of_resources  0.58010385  0.01710893  33.906
##                                            Pr(>|t|)
## (Intercept)                    < 0.0000000000000002 ***
## Status                               0.000026140383 ***
## Adult_Mortality                < 0.0000000000000002 ***
## Alcohol                                   0.000639 ***
## Hepatitis_B                          0.000000032161 ***
## BMI                                  0.000000000503 ***
## Income_composition_of_resources < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0763 on 2038 degrees of freedom
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.7141
## F-statistic: 851.8 on 6 and 2038 DF,  p-value: < 0.00000000000000022

##                          Variables Tolerance      VIF
## 1                           Status 0.5638536 1.773510
## 2                  Adult_Mortality 0.6868741 1.455871
## 3                          Alcohol 0.5750266 1.739050
## 4                      Hepatitis_B 0.8653605 1.155588
## 5                              BMI 0.6376106 1.568355
## 6 Income_composition_of_resources 0.4117911 2.428416

##  lag Autocorrelation D-W Statistic p-value
##    1      0.02647534      1.946694   0.232
##  Alternative hypothesis: rho != 0

##
##  studentized Breusch-Pagan test
##
## data:  modelbestnew3
## BP = 239.45, df = 6, p-value < 0.00000000000000022

##
##  Shapiro-Wilk normality test
##
## data:  modelbestnew3$residuals
## W = 0.91743, p-value < 0.00000000000000022

## Root mean square error:  0.001374664
```

Because $p_{value} < 2.2e^{-16} \rightarrow Reject\ H_0$.So the null hypothesis that all six $B_i's$ corresponding to predictors have value 0 is resoundingly rejected. There appears to be a useful relationship between the dependent variable and at least one of the predictors. We can see p_{value} of 6 variables are too small, so they can be use to explain log(life_expantancy).

The VIF indexes of six variables are not significant. Therefore we can ignore the multicollinearity of this model.

Moreover, $R^2 = 71.49\%$ means that 71.49% of the observed variation in log(life_expectancy) can be explained by the linear regression relationship based on Status, Adult_Mortality, Alcohol, Hepatitis_B, BMI, Income_composition_of_resources.

After having prediction values, we can see that root mean square error is equal to 0.001374664.

This model is better than above model. However two problems (the variance of the residuals is not constant and the residuals of model do not adhere to normal distribution) remain. While $p_{value} = 0.236 > 0.5 \rightarrow$ *Accept* H_0, when using Durbin Watson Test. Therefore, there is no correlation among the residuals at rist level $\alpha = 5\%$.

c) Box-Cox transformation

Firsly, we find the optimal λ for two model in V by using boxcox function:

```
## Model with good correlation and lambda =   2

##
##  studentized Breusch-Pagan test
##
## data:  model1lam
## BP = 215.24, df = 4, p-value < 0.00000000000000022

##
##  Shapiro-Wilk normality test
##
## data:  model1lam$residuals
## W = 0.96056, p-value < 0.00000000000000022

## Root mean square error:  0.4156363

## Full model after drop schooling and  lambda =   2

##
##  studentized Breusch-Pagan test
##
## data:  model2lam
## BP = 155.99, df = 8, p-value < 0.00000000000000022

##
##  Shapiro-Wilk normality test
##
## data:  model2lam$residuals
## W = 0.96086, p-value < 0.00000000000000022

## Root mean square error:  0.3249085
```

Although apply to use Box-Cox transformation, the residuals of two models do not adhere to normal distribution and variances are not constants. Moreover two models are not better than log(model) with full model when compared by root mean square error.

VII. Explain the problems

Formal normality tests, such as the Shapiro-Wilk test, can be highly sensitive to sample size. With very large samples, even small deviations from normality can result in a significant result, leading to the rejection of the null hypothesis even if the distribution is reasonably normal for practical purposes. It is a fact that formal normality tests always reject on the huge sample sizes we work with today. And since every dataset has some degree of randomness, no single dataset will be a perfectly normally distributed sample. As we can see p_{value} of studentized Breusch-Pagan test and Shapiro-Wilk normality test are very small, leading to reject null hypothesis. Because the assumption of normality of residuals in the regression model is violated, the models here may not be very reliable.

Conclusion

We use compare_performance() function to compute indices of model performance for different models at once and hence allow comparison of indices across models. Any of "all", "common", "AIC", "AICc", "BIC", "WAIC", or "LOOIC are requested in metrics to give the Performance_score.

Name	R2_adjusted	RMSE	Performance_Score
model3	0.7140717	0.07617358	0.9131815
model1	0.7453261	4.69301667	0.5373205
model2lam	0.7622222	303.14228631	0.3264693
model2	0.6037861	0.08973462	0.2857034
model1lam	0.6768602	353.73841218	0.1316785

After considering, model 3 has the highest Performance_score 0.9131815 consistent with the above analysis. Although other models have better parameters when compared to model 3, if we consider the factors mentioned above, model 3 is the best model.

$\hat{y} = e^{3.77448386} e^{0.02470471 status} e^{-0.00027642 Adult\_Mortality} e^{-0.00194775 Alcohol} e^{0.00111065 Hepatitis\_B} e^{0.00065968 BMI}$ $e^{0.58010385 Income\_composition\_of\_resources}$ is the best model. For developed countries, life expectancy will increase by 1.0250124 compared to developing countries. When the Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) increases by 1%, then life expectancy decreases by a factor of 1.000276. When consumption (in liters of pure alcohol) increases by 1 liter, then life expectancy decreases by a factor of 1.00195. When Hepatitis B (HepB) immunization coverage among 1-year-olds increases, then life expectancy increases by a factor of 1.001111. When the Average Body Mass Index of the entire population increases by 1, then life expectancy increases by a factor of 1.00066. When the Human Development Index in terms of income composition of resources is 1, then life expectancy increases by a factor of 1.786224.

— The End —