ELECTRONIC ASSIGNMENT COVERSHEET

Murdoch UNIVERSITY

Student Number:	34445873
Surname:	Chew
Given name:	Jian Yue
Email:	cjianyue@gmail.com

Unit Code:	ICT 513
Unit name:	Data Analytics
Enrolment mode:	External
Date:	9th Oct 2022
Assignment number:	1
Assignment name:	Assignment 1
Tutor:	Dr. Goh Zhang Hao

Student's Declaration:

- Except where indicated, the work I am submitting in this assignment is my own work and has not been submitted for assessment in another unit.
- This submission complies with Murdoch University's academic integrity commitments. I am aware that information about plagiarism and associated penalties can be found at http://our.murdoch.edu.au/Educational-technologies/Academic-integrity/. If I have any doubts or queries about this, I am further aware that I can contact my Unit Coordinator prior to submitting the assignment.
- I acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment:
  - reproduce this assignment and provide a copy to another academic staff member; and/or
  - submit a copy of this assignment to a plagiarism-checking service. This web-based service may retain a copy of this work for the sole purpose of subsequent plagiarism checking, but has a legal agreement with the University that it will not share or reproduce it in any form.
- I have retained a copy of this assignment.
- I will retain a copy of the notification of receipt of this assignment. If you have not received a receipt within three days, please check with your Unit Coordinator.

I am aware that I am making this declaration by submitting this document electronically and by using my Murdoch ID and password it is deemed equivalent to executing this declaration with my written signature.
Optional Comments to Tutor: *e.g. If this is a group assignment, list group members here*

If you can, please insert this completed form into the body of ***each*** *assignment you submit. Follow the instructions in the Unit Information and Learning Guide about how to submit your file(s) and how to name them, so the Unit Coordinator knows whose work it is.*

Start your assignment on the next page.

Question 1

Part a

The goal of graphical visualisations is clear and effective scientific communication derived from data. Visualisations should be information-rich (truthfulness) yet simplify massive datasets (like descriptive statistics).

- Seek a balance between data simplification and revealing meaningful variation without excessive generalisation (information loss).
- Use the right visualisation tool to best communicate an analyses/task
- More visualisation features (overlaying) deliver more information; limit cramming excessive information (complex figures), which creates confusion

Embrace minimalism by maximising signal-to-noise ratio and remove unjustifiable features.

- Use San-serif typefaces
- Avoid three-dimensional visualisations (affects understanding) and shadows (gimmicks)
- Consider choice of colours for:
  - Inclusivity (colour blindness)
  - Type of variable (starkly differentiable for categorical); to reduce readers' biasness

Use goal-centred graphical design to illustrate specific hypotheses to make visualisations more effective. Despite historical precedence and popular usage of some visualisation tool, search for modern visualisation tools that provide more information-rich plots.

Central tendency and data distribution

- Mean – Barplot
- Median – Boxplot

For central tendency and distribution of data, **raincloud plot** (modern descendants) is preferred. **Cluster heat maps** can be used for data changing over time across many categories.

Use jitter and partial transparency for datapoints for large datasets; chances of datapoints overlapping are high, which improves communication.

Proportions/Frequencies

- Avoid using pie charts to represent proportions/percentages as readers may inaccurately interpret proportions visually, especially when comparing proportions across multiple pie charts
- **Simple bar plot** is preferred as readers perceive lengths well; put proportions to compare adjacently
- **Stacked bar plot** for comparison within and between clusters
- **Line plot** for visualising changes of proportions across time

Relationships (correlations) between variables can be represented with scatterplots. However, improvements to information richness can be achieved through:

- Use of **improved scatterplot**: adds colour, confidence intervals and additional features such as histogram and density plot.
- Contour plot for large number of observations, rendering traditional scatterplots ineffective.
- Spaghetti plots to represent relationships between clustered data; being careful not to include too much features

In conclusion, utilise a mix of colour/transparency, size and numbers/labels to improve visibility and increase information-richness, justifying adding new features to exiting plots (to prevent confusing analyses communicated) and place the goal of scientific communication over plain aesthetics.

Part b

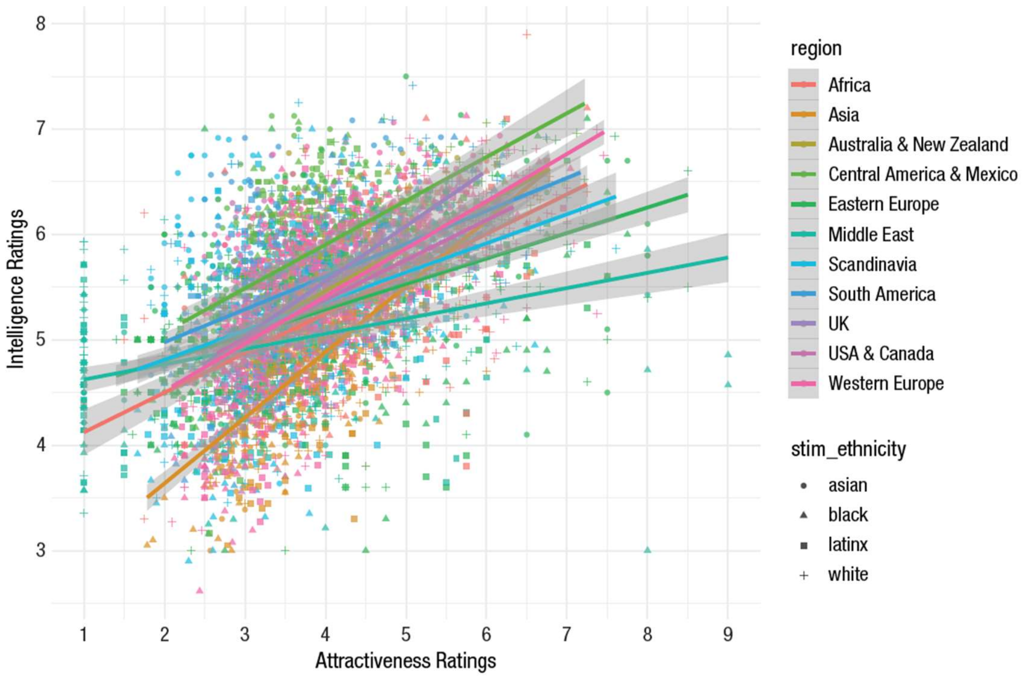


Fig. 3. An overinformationally rich visualization. This scatterplot depicts the relationship between ratings of attractiveness and ratings of intelligence made on targets across four ethnicities by perceivers from different world regions.

One of the visualisation was used to represent the relationship between intelligence ratings and attractiveness ratings in different regions and ethnicity [1]. This visualisation seems to be a scatterplot with best fit lines drawn. The is a visualisation is ineffective in communicating information to the reader. It may be aesthetically pleasing to look at, but at the same time confusing and frustrating. The reader is unsure what analysis is being communicated by the researcher. There are so many features in the single graph alone, including confidence intervals, ethnicity represented by shapes and too many regions and colours represented and overlayed on the same plot. It becomes hard to tell which region or ethnicity has stronger relationship between the two ratings. There could be unclear research objective by the author, resulting in such plots generated. There are too many overlapping points from different sub-groups, and too many overlapping best-fit lines. This violates the philosophies discussed earlier, on minimalism and information over-richness which defeats the purpose of the visualisation which is to simplify the dataset much like descriptive statistics. Furthermore, the colours used were not inclusive, and is confusing. Consider the colour difference between Western Europe and Africa, it is difficult to tell the difference between the two prominent colours in the plot and on the legend. The shapes used in the plots were too similar and small, because of the number of sub-groups for ethnicity. A quantitative analysis would have been better to illustrate an analysis, conclusion or prove a hypothesis, compared to this complex looking diagram. Possibly separating the diagram to multiple diagrams aid the reader in understanding it.

Question 2

Q2. Weights of four packets from filing machine (in g):

251.3 250.6 250.9 249.8

Question asks for Median (M) as the estimator and not mean

T(M) is defined as Median (M). T(M) jackknife is the jackknife estimator.

Eliminate 1 value of the sample to generate a subsample of size (n-1).

Reorder by ranking numerical values : 249.8 250.6 250.9 251.3

$$\text{Sample Median } (M) = \frac{250.6 + 250.9}{2} = 250.75$$

Sub-sample 1 : ~~249.8~~ 250.6 250.9 251.3

Jackknife replicate 1 (JKM_1) = 250.9 = $M_{(-1)}$

Sub-sample 2 : 249.8 ~~250.6~~ 250.9 251.3

Jackknife replicate 2 (JKM_2) = 250.9 = $M_{(-2)}$

Sub-sample 3 : 249.8 250.6 ~~250.9~~ 251.3

$$JKM_3 = 250.6 = M_{(-3)}$$

Sub-sample 4 : 249.8 250.6 250.9 ~~251.3~~

$$JKM_4 = 250.6 = M_{(-4)}$$

Jackknife replicates : 250.6 250.6 250.9 250.9

$$T(\underset{\sim}{y})_{\text{Jackknife}} = \frac{2(250.6 + 250.9)}{4} = 250.75$$

= Mean of Jackknife estimates

= Jackknife estimator #.

$$SE\left(T\left(\underset{\sim}{M}\right)\right)_{\text{Jackknife}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(T\left(\underset{\sim}{M}\right)_{(-i)} - T\left(\underset{\sim}{M}\right)_{\text{Jackknife}}\right)^2}$$

$$= \sqrt{\frac{3}{4}\left[2\left(250.9-250.75\right)^2 + 2\left(250.6-250.75\right)^2\right]}$$

$$= \sqrt{\frac{3}{4}\left[(2)\left(0.02250\right) + 2\left(0.02250\right)\right]}$$

$$= \sqrt{\frac{3}{4}\left(0.04500 + 0.04500\right)}$$

$$= \sqrt{0.06750} = 0.25981$$

(at least 3 s.f.). #.

$$\text{Bias}\left(T\left(\underset{\sim}{M}\right)\right)_{\text{Jackknife}} = (n-1)\left(T\left(\underset{\sim}{M}\right)_{\text{Jackknife}} - T\left(\underset{\sim}{M}\right)\right).$$

$$= 3\left(250.75 - 250.75\right)$$

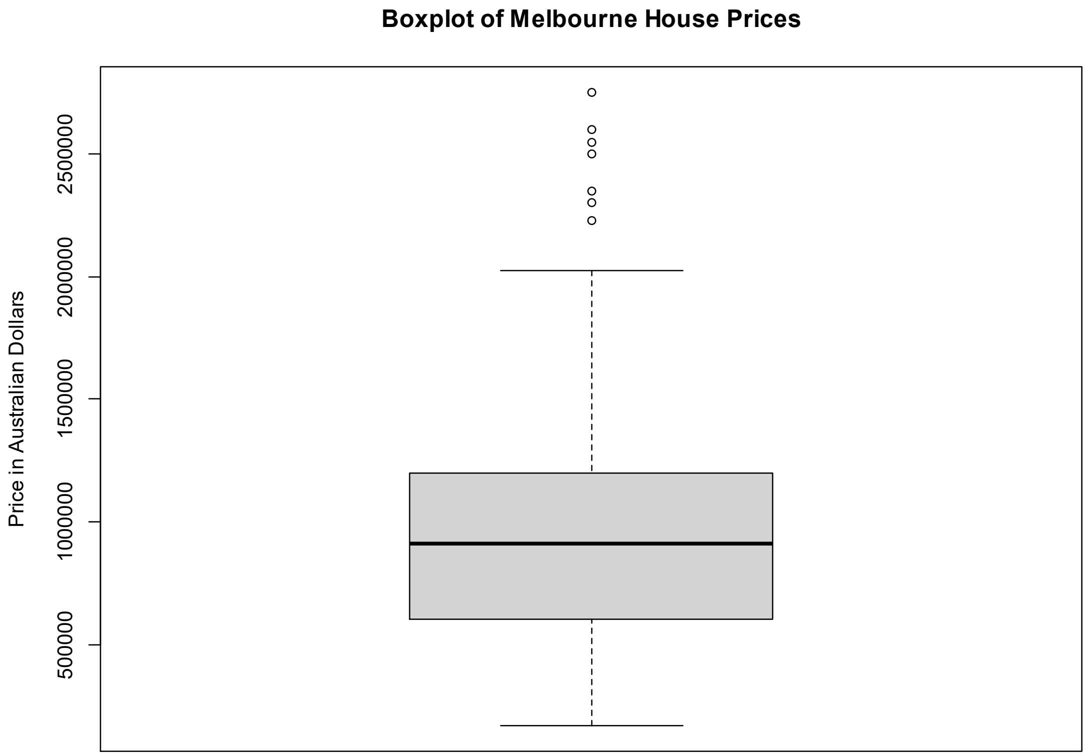
$$= 3(0) = 0 \quad \#$$

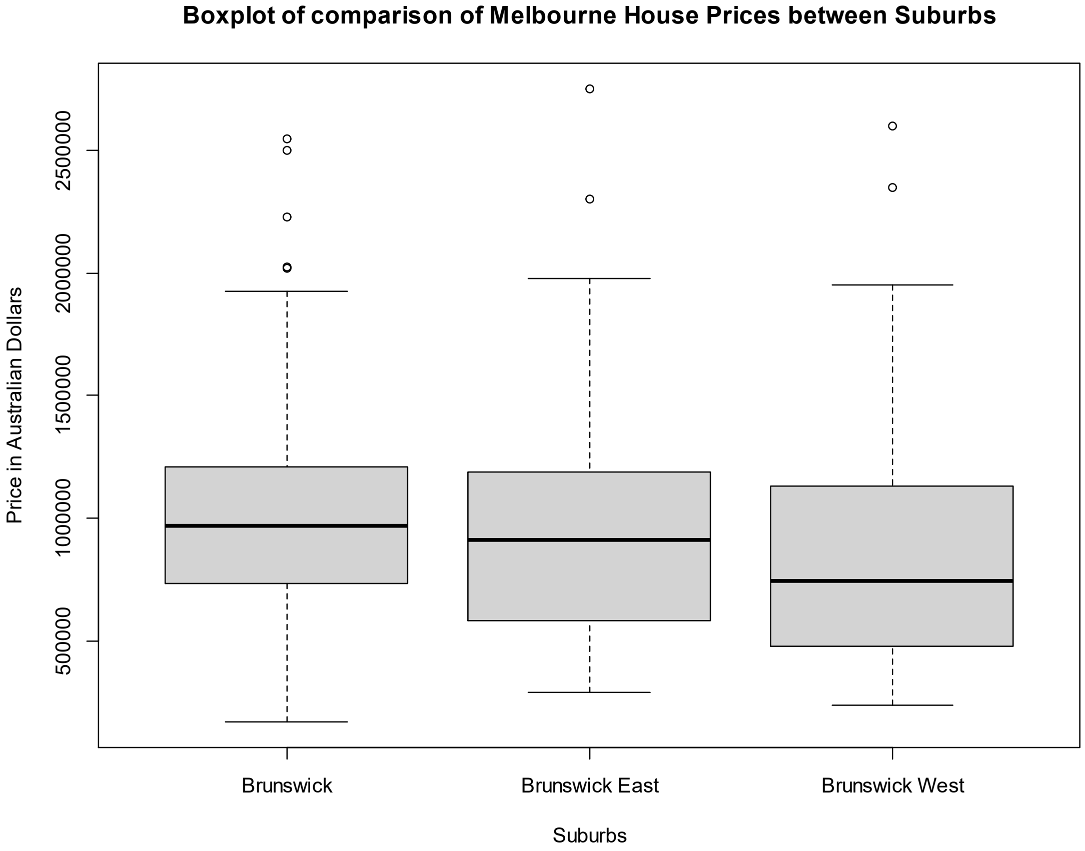
The jackknife estimate of the bias is 0. The estimator is unbiased

In conclusion, Jackknife estimator = 250.75 and standard error of the median = 0.25981. Estimator is unbiased.

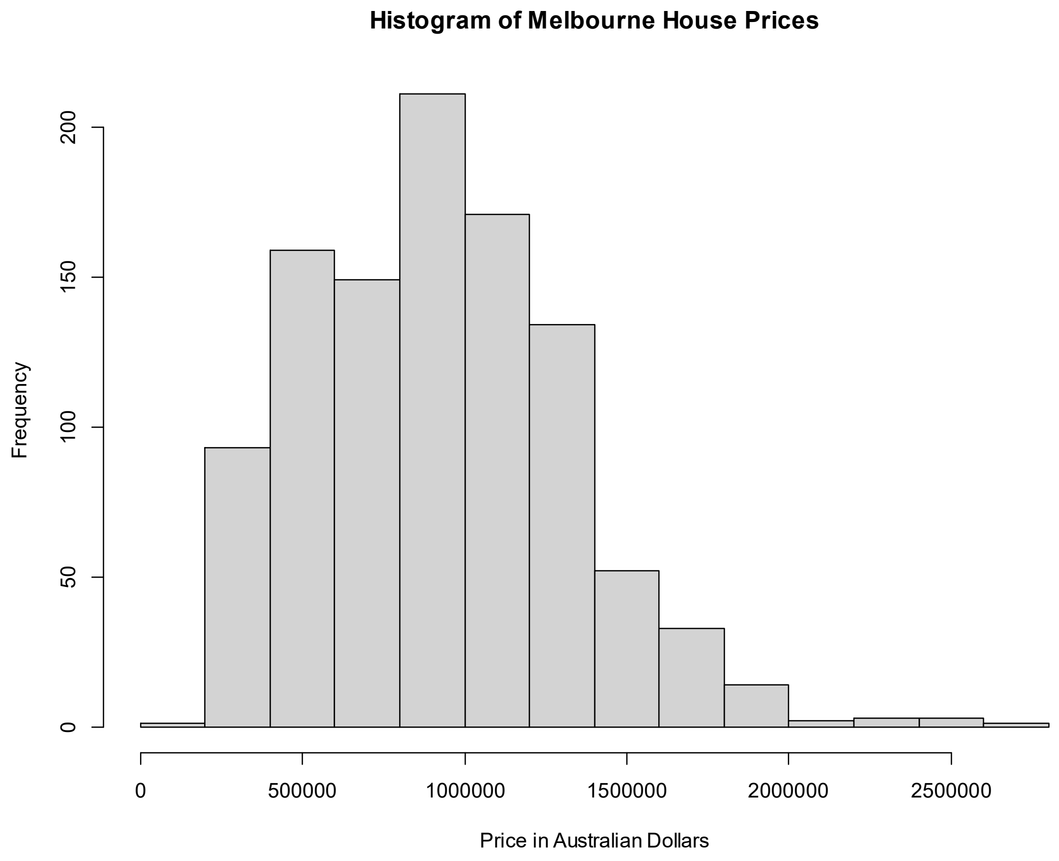
Question 3

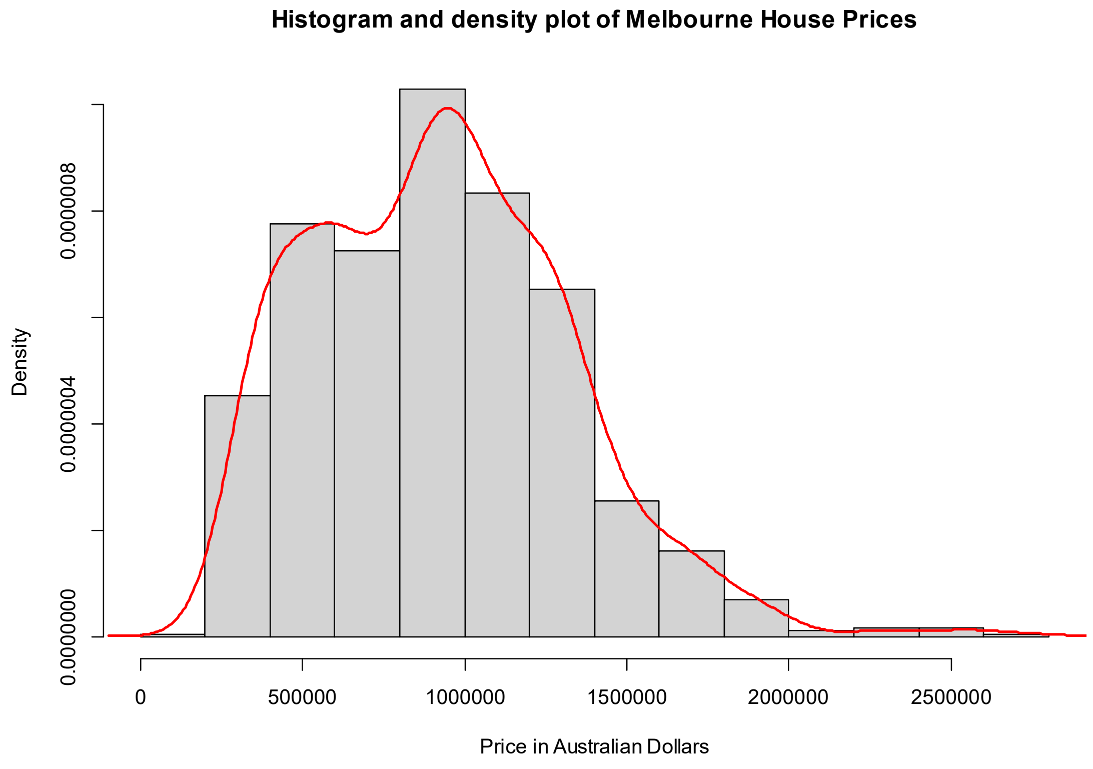
Part a



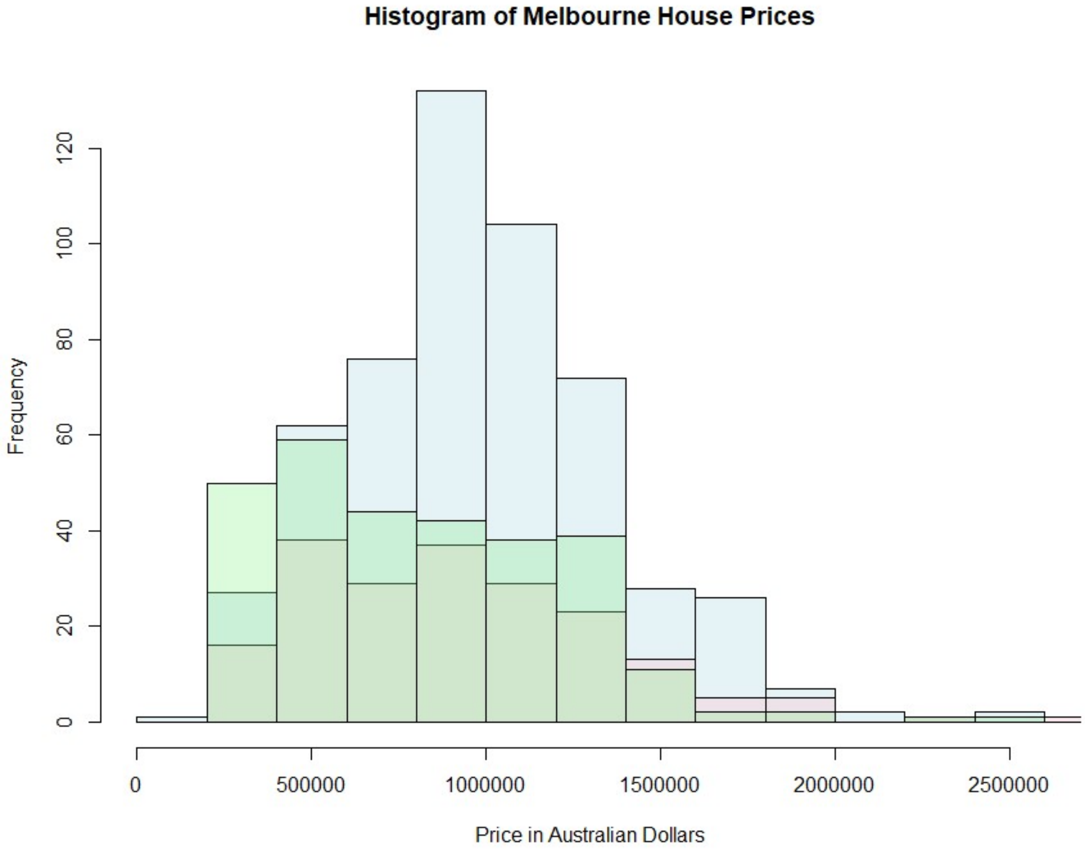


From the boxplots, it can be seen that there are many outliers on the right tail of the boxplot. Outliers are defined here as datapoints that are 1.5 times IQR above the value of Q3. Hence, there are a total of seven outliers in this distribution. It can be seen that the distribution is positively skewed. Generally, the median prices in Brunswick is higher than in other suburbs.


Histogram of Melbourne House Prices
Frequency
0
50
100
150
200
0
500000
1000000
1500000
2000000
2500000
Price in Australian Dollars


Histogram and density plot of Melbourne House Prices
Density
0.0000000
0.0000004
0.0000008
0
500000
1000000
1500000
2000000
2500000
Price in Australian Dollars

From the histogram shown, the Melbourne house prices are positively skewed or right-skewed. The peak of the distribution has a frequency of around 200 and the price of the houses are approximately AUD 1,000,000.

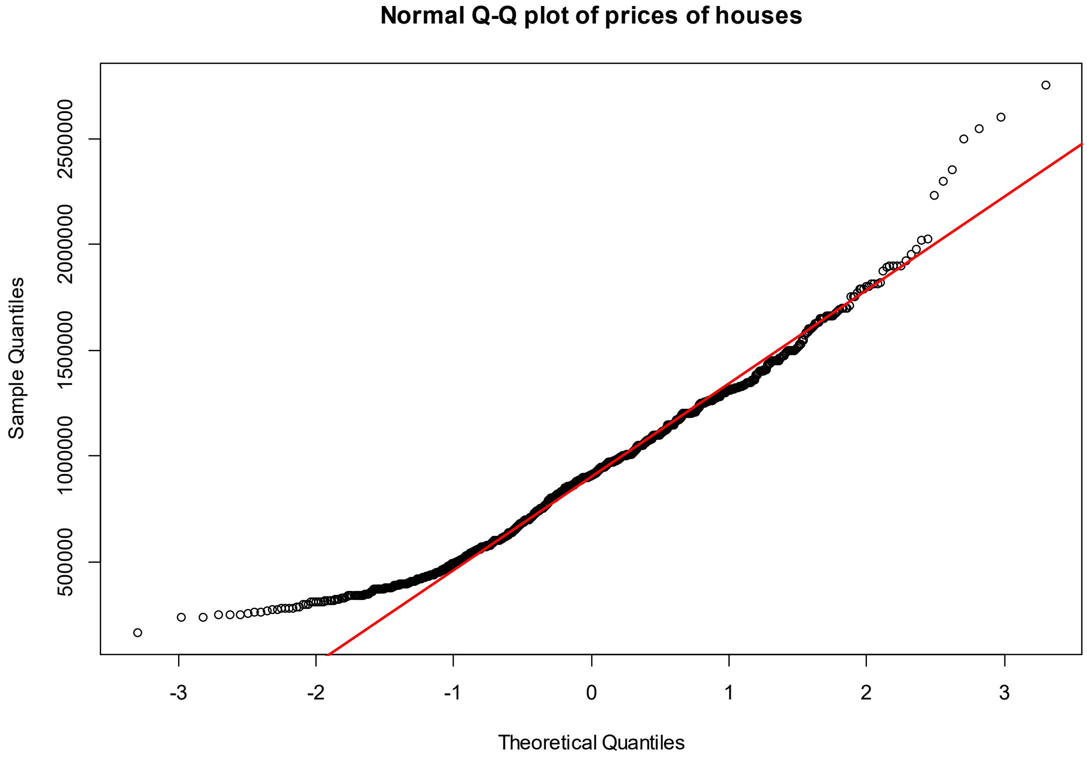


The histogram above shows the distribution separated into Surburb sub-groups, in colours.

The colour mapping is as follows, and can be reflected from the R source code:

Colour	Surburb Sub-group
Blue	Brunswick
Pink	Brunswick East
Green	Brunswick West

From the above histogram, the Melbourne house prices for the “Brunswick” area seems to influence the central tendency of the entire dataset.



Quantile-quantile plot provides an avenue to compare the quantiles of the empirical distribution with the known normal distribution. There are many areas of the Q-Q plot having observed points deviating from the normal line. This suggests that the distribution is unlikely to be normally distributed.

The plausible explanation for this is that most of the Melbourne houses are located in Brunswick, for this subset of data, compared to Brunswick East and Brunswick West. The least number of houses are possibly located at Brunswick West. On average, most of the houses in the world, and including in Melbourne are priced at around the peak frequency, while some houses which could be very expensive are owned by the wealthy elites. The shape of the distribution is rather bell-shaped, which is a characteristic shared by the prices of houses around the world. The majority of people who seek houses that are likely of average size, which costs the average price (assuming the price per square feet around the same area is the same). The majority of people in the area would generally have similar incomes to afford houses that represents around 68% of the distribution of data. The lower priced houses may be less desired for any average individual with an average-sized family. Similarly, those that require larger houses (possibly more expensive) are rarer, and those who are afford very expensive houses are lesser.

Brunswick West seems to have cheaper houses, represented by the frequency peaking at a lower price, although the number of datapoints for Brunswick West is lesser than for Brunswick.

For the frequency distribution of Brunswick East, there are two represented peaks (bimodal data).

Since the data is slightly skewed, I would not consider this distribution as a whole to be normally distributed. There are outliers that may affect the results of central tendency, hence mean may not be a robust and appropriate measure for central tendency. Hence, for centrality measure, I have decided to use Median instead. From the density plot overlaying on the histogram, it is more clearly describing that the distribution is unlikely to be normally distributed, as the bell-shape is not very obvious. The right tail extends quite far to the right. The possible reason for some unevenness in the density plot may be due to noise.

The interquartile range is the most robust measure for dispersion, compared to other measures of dispersions such as range, mean absolute deviation, variance, and standard deviation.

Hence, choosing the measure of centrality (median) and measure of dispersion (interquartile range).

Median	AUD 910000
Interquartile range (Q3 – Q1)	AUD 595000

Part b

Measure of Centrality: *Median (Value of Median when applied to original data = AUD 910000)*

(in AUD)	Jackknife	Bootstrap
Estimator	910000	912449
Standard error	0	14523.03
Bias	0	-2615.25

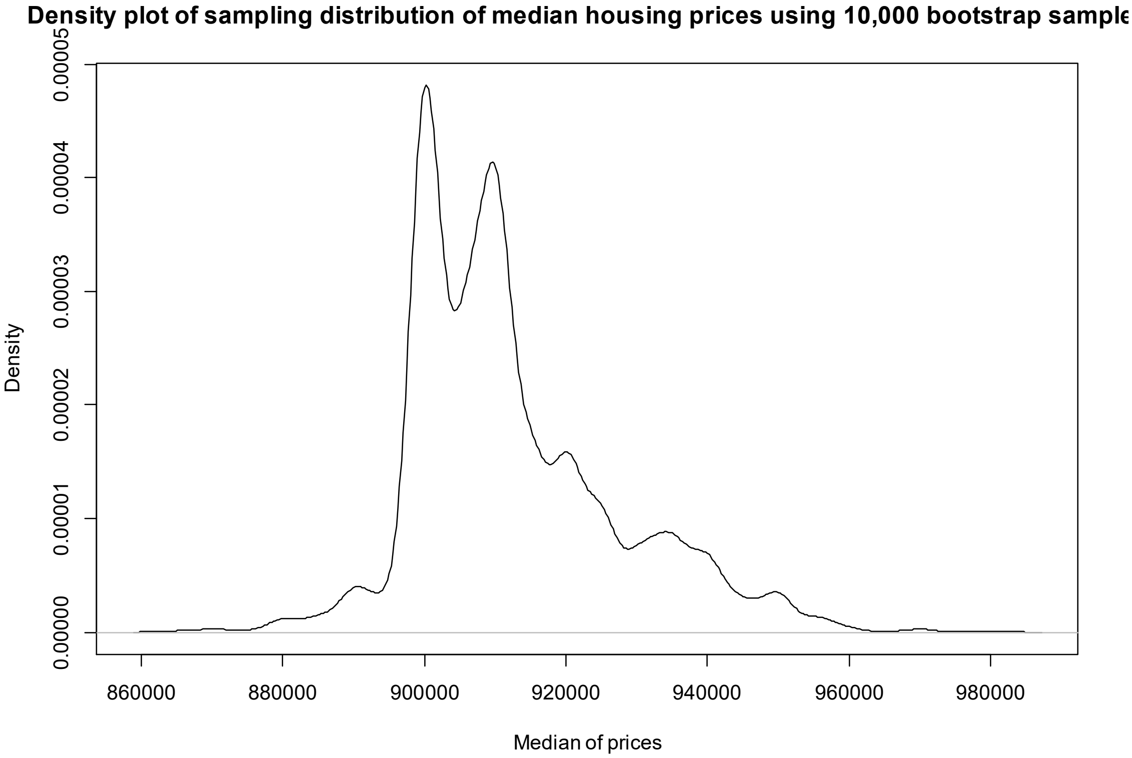
Measure of Dispersion: *Interquartile Range (Value of IQR when applied to original data = AUD 595000)*

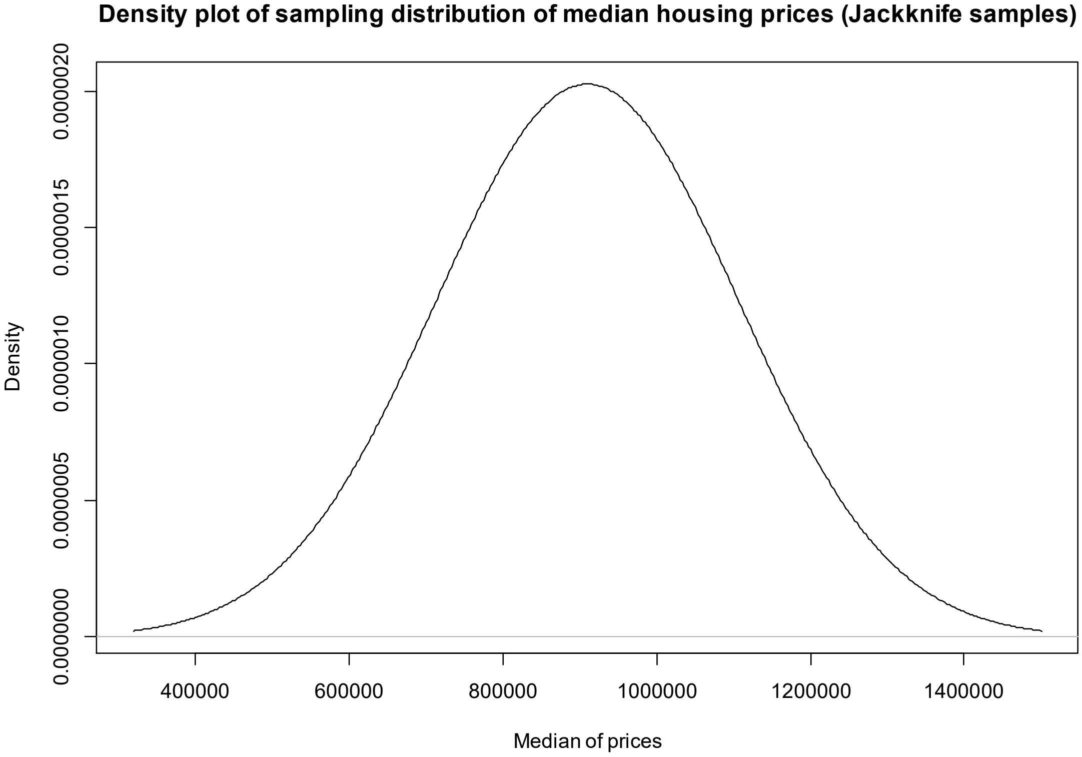
(in AUD)	Jackknife	Bootstrap
Estimator	595000	586069.8
Standard error	0	18527.6
Bias	0	-317449

For jackknife estimator of median and interquartile range, the measures are unbiased because bias = 0. For the bootstrap estimator of median and interquartile range, the measures appear to be biased estimators.

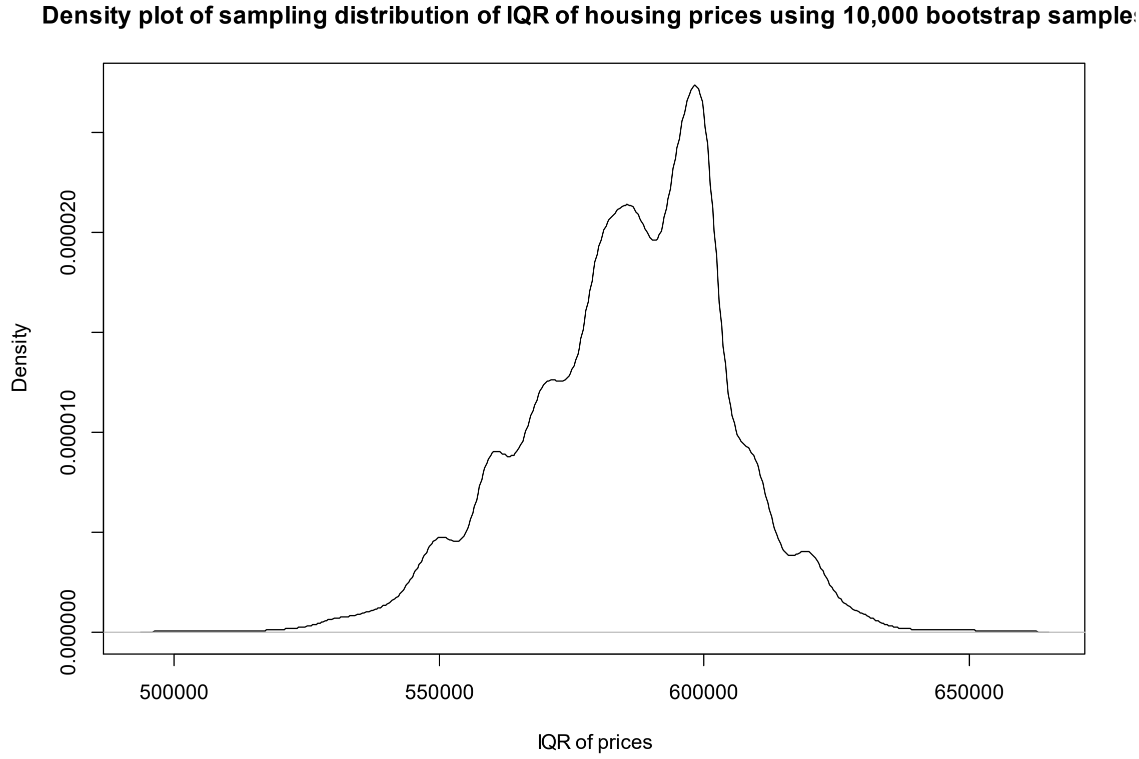
Part c

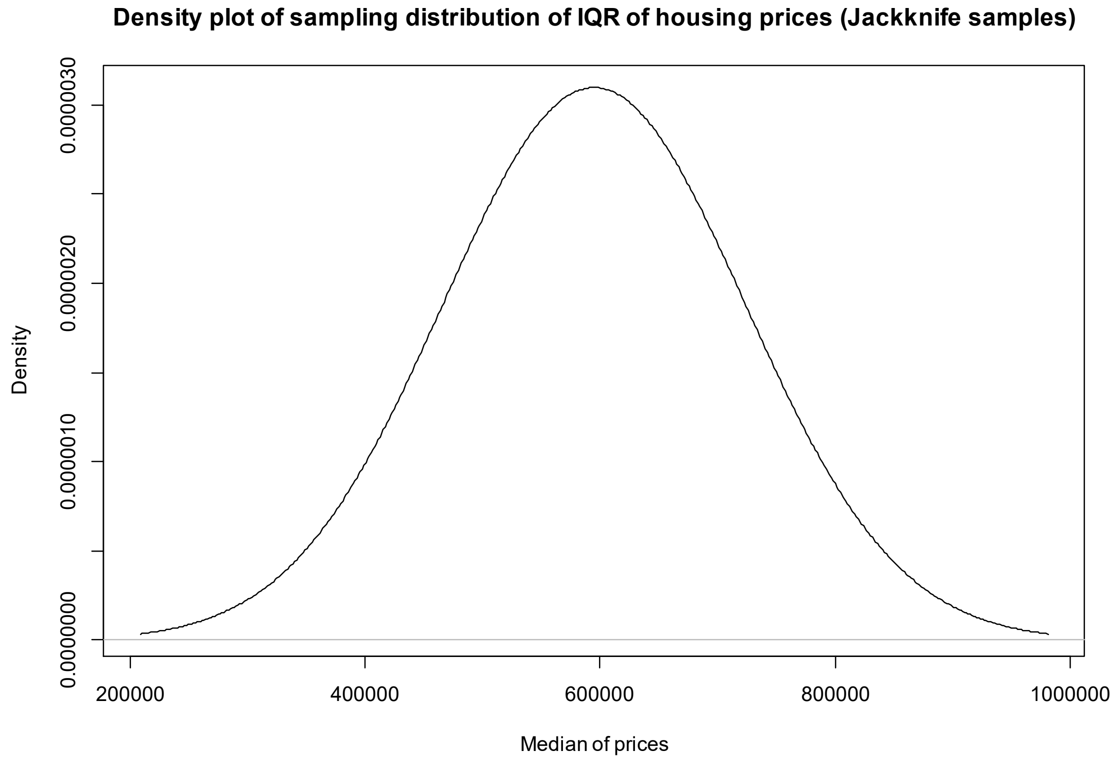
For centrality:





For dispersion:





The shape of both the density plots for bootstrap samples do not look normally distributed. There are many jagged areas to the curves for both bootstrap estimates.

For the jackknife estimators for median and IQR, the distribution now looks normally distributed complying with the central limit theorem.

The confidence interval for measure of centrality (median): (890000, 950000)

2.5%	97.5%
890000	950000

This means that we are 95% confident that most of the density of the sampling distribution (median) falls between 890000 and 950000. The median of the sample falls within the confidence interval.

Confidence interval for measure of dispersion (IQR): (548000, 620000)

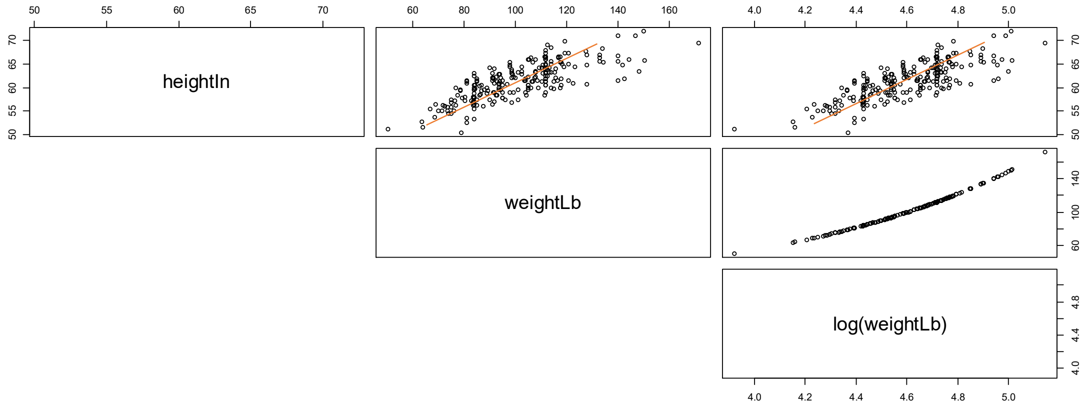
2.5%	97.5%
548000	620000

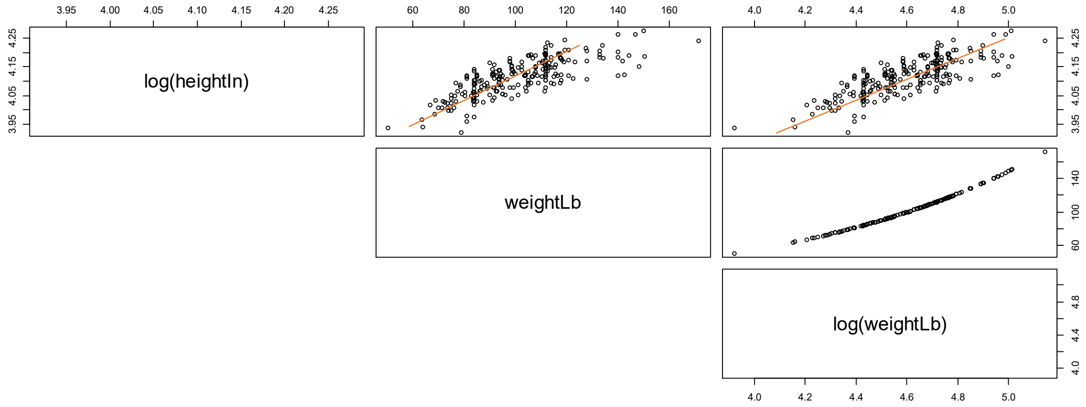
This means that we are 95% confident that most of the density of the sampling distribution (median) falls between 548000 and 620000. The IQR of the sample falls within the confidence interval.

Question 4

The two variables particularly under analysis is height in inches (heightIn) and weight in pounds (weightLb) available in the R package *gcookbook*.

The assumptions for linear regression must be met for the results of linear regression to be valid. The four key assumptions of the linear regression model are linearity, homoscedasticity, normality and independence of observations. We are assuming here that height predicts the weight of school children. Both variables are numerical variables, linearity check is required.

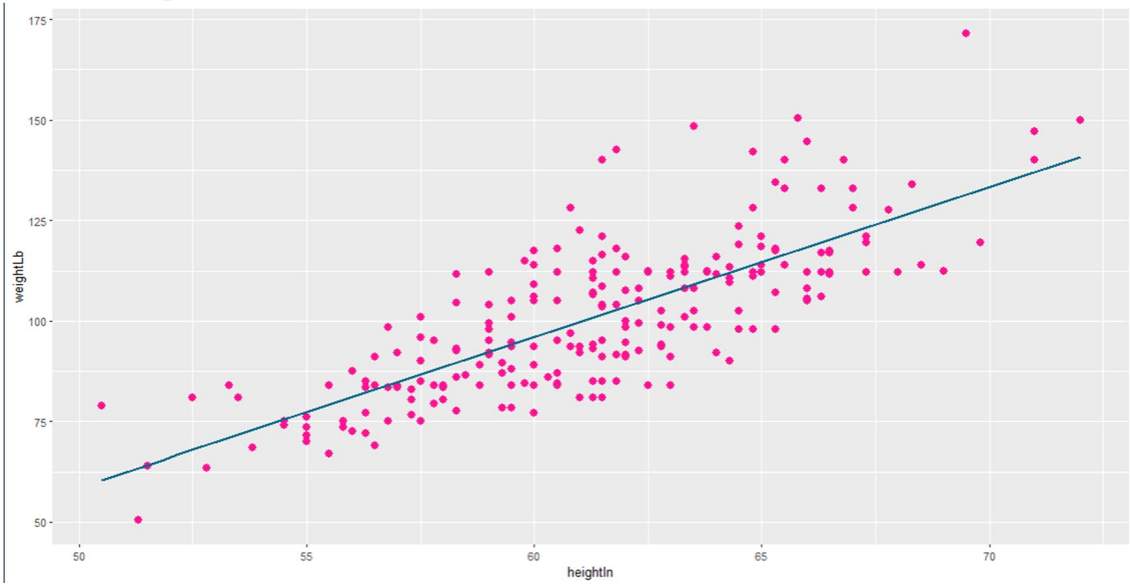




There is no major difference between the linearity assumption (not violated) when any of the variables are subjected to log transformation.

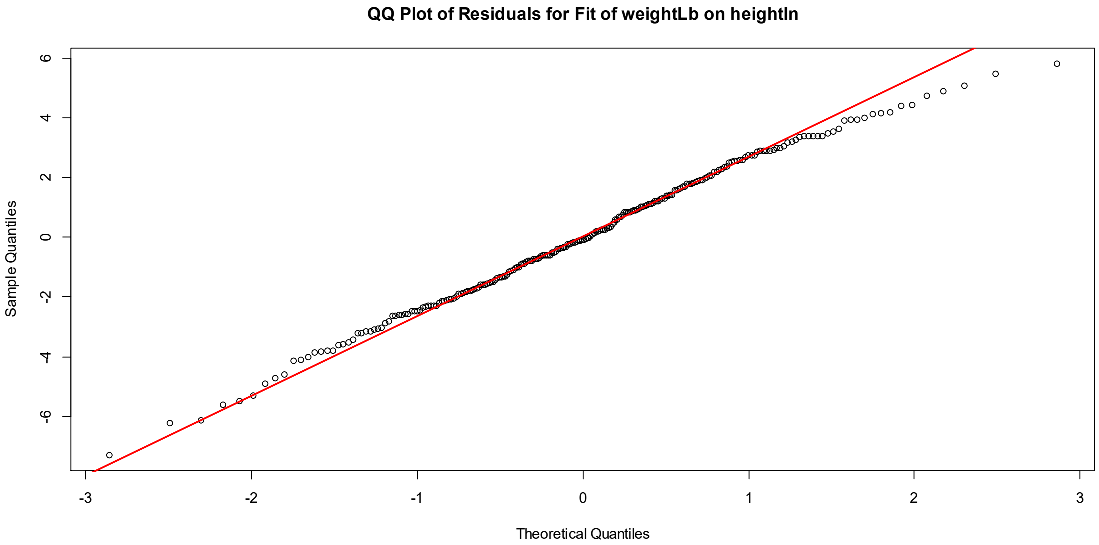
(a) Linearity check

With scatterplot



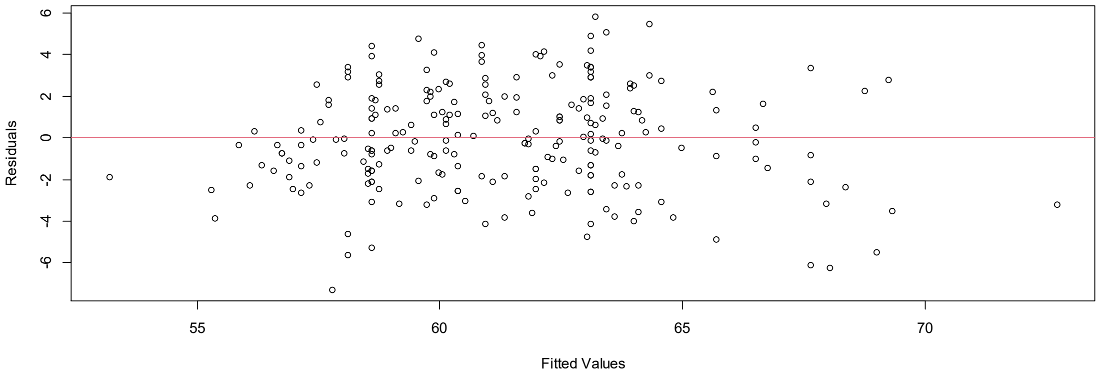
The relationship between the two variables is rather linear, as seen on the first scatterplot. Hence the linearity assumption is not violated.

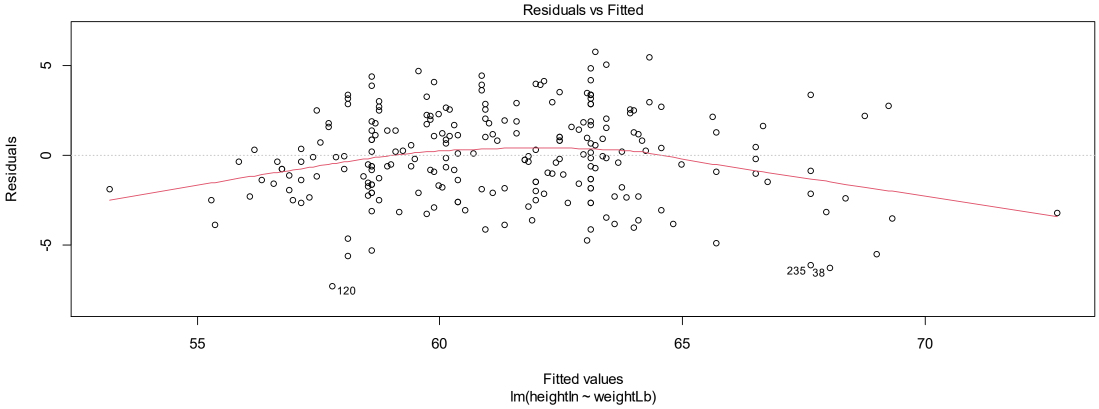
Q-Q Plot for normality check



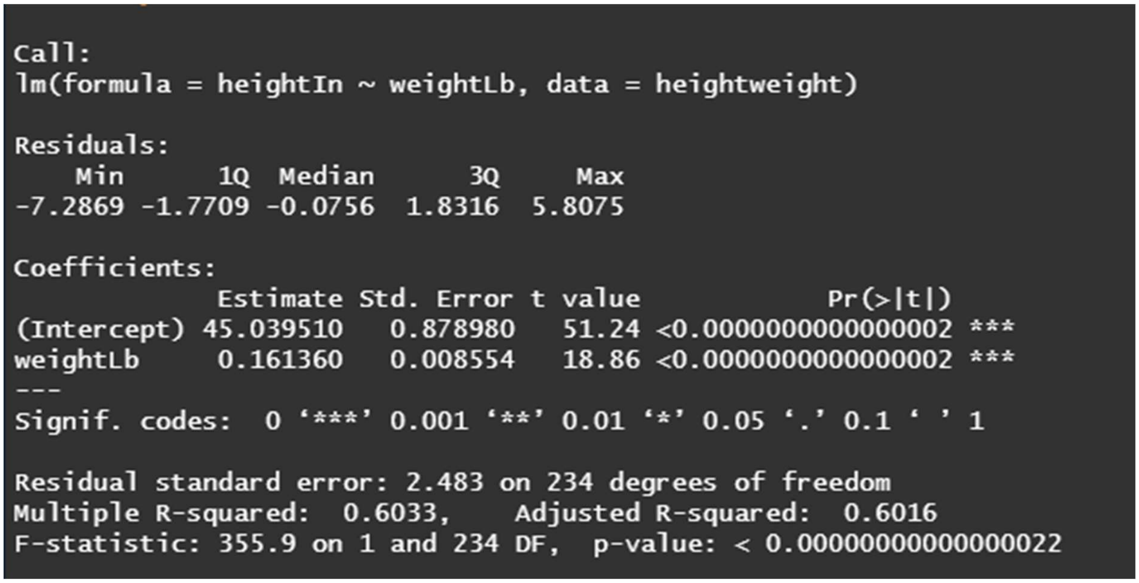
Most of the data points fit closely to the normal line plotted on the Q-Q plot except for larger theoretical quantiles. Normality assumption may be violated due to that.

Homoscedasticity assumption (residuals vs. fitted values scatterplot)



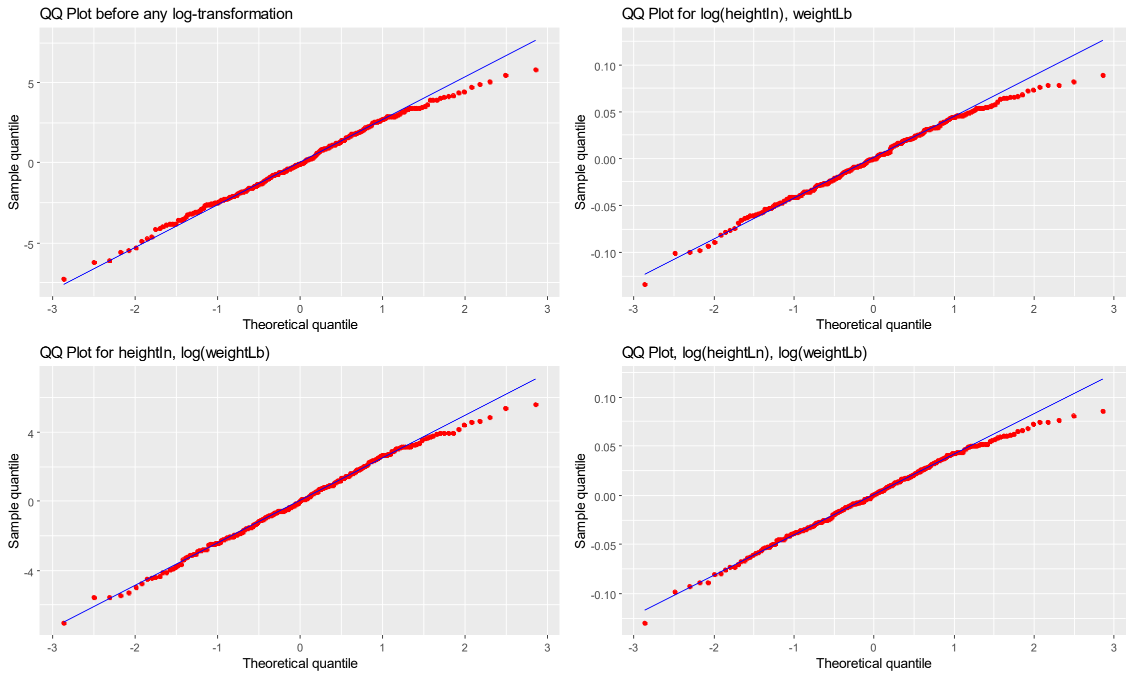


There seems to be constant variance throughout the fitted values. The variation around the horizontal line at residual = 0, are rather equal. There is no funnelling shape from the observations seen. From the new plot, it is more obvious that the red line is not horizontal and seems curvilinear, hence the homoscedasticity assumption may be violated. The variance may not be as constant as thought. The independence of observations is not violated as there is no obvious funnelling of datapoints seen.

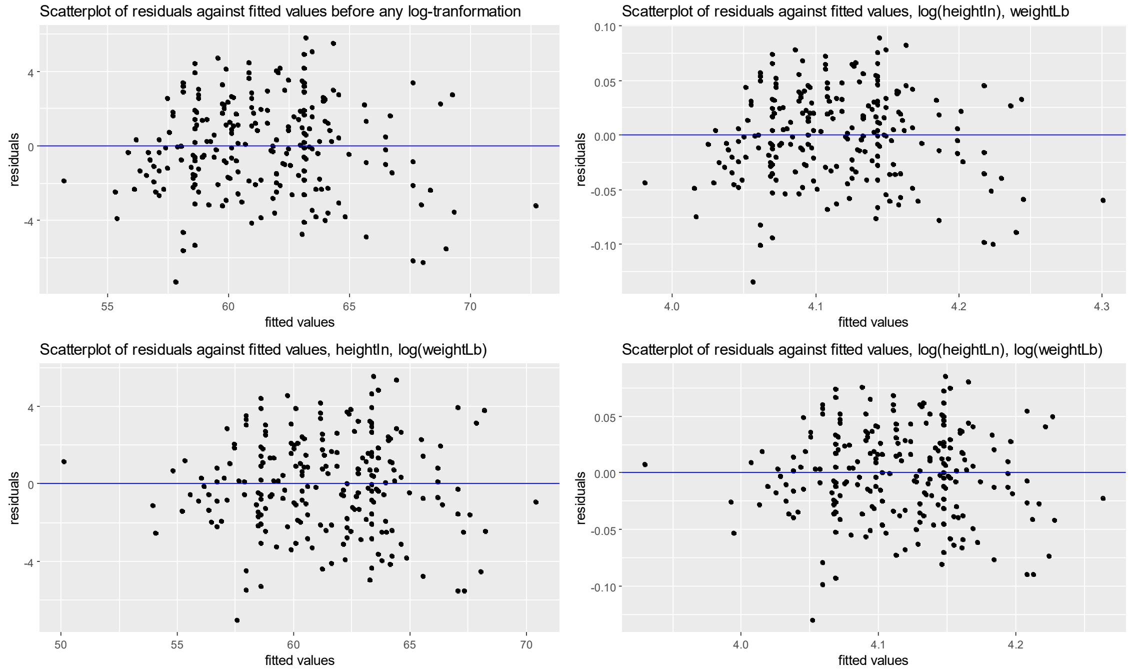


```
Call:
lm(formula = heightIn ~ weightLb, data = heightweight)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2869 -1.7709 -0.0756  1.8316  5.8075 

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 45.039510   0.878980   51.24 <0.0000000000000002 ***
weightLb     0.161360   0.008554   18.86 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.483 on 234 degrees of freedom
Multiple R-squared:  0.6033,	Adjusted R-squared:  0.6016 
F-statistic: 355.9 on 1 and 234 DF,  p-value: < 0.00000000000000022
```

The p-value shows statistical significance (<0.05), hence the weight variable is a statistically significant variable predicting the response variable, height. There is a relationship between the two variables. However, based on the earlier assumptions which are not fully met, these results are not valid.



The above is the Q-Q plots for after individual log transformations on either or both of the variables. There are no major differences in their normality assumption, although for log transformation of weight without log transformation of height, the data points seem to have the least deviation from the normal line.


Scatterplot of residuals against fitted values before any log-tranformation

Scatterplot of residuals against fitted values, log(heightIn), weightLb

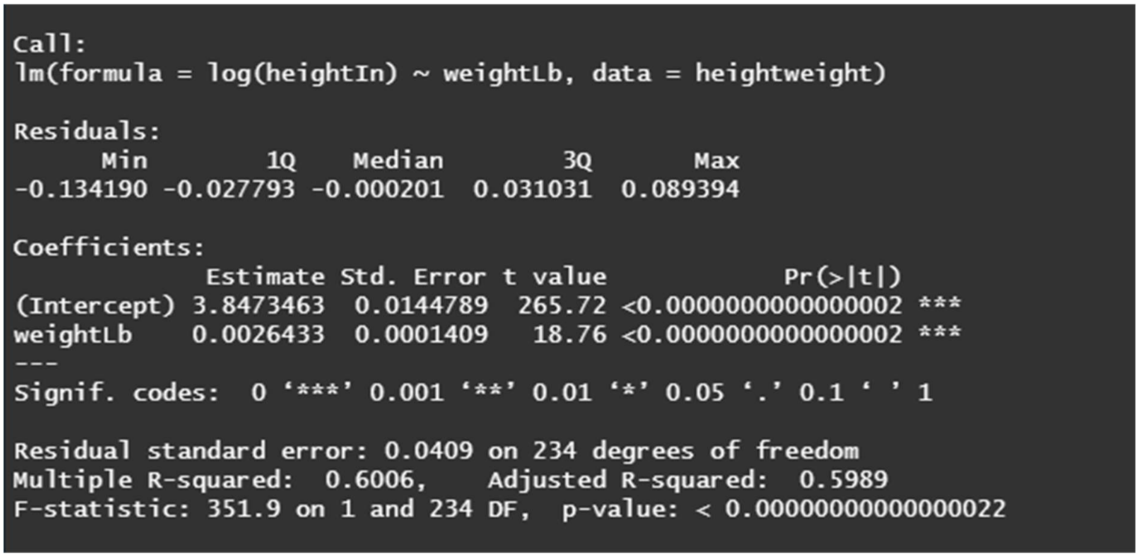
Scatterplot of residuals against fitted values, heightIn, log(weightLb)

Scatterplot of residuals against fitted values, log(heightLn), log(weightLb)

There is no obvious funnelling of the points and variance seems rather constant according to the plot above. Although the results seem to be better for heightIn and log(weightLb). I believe the linear model for this would be the best although the difference may be marginal.

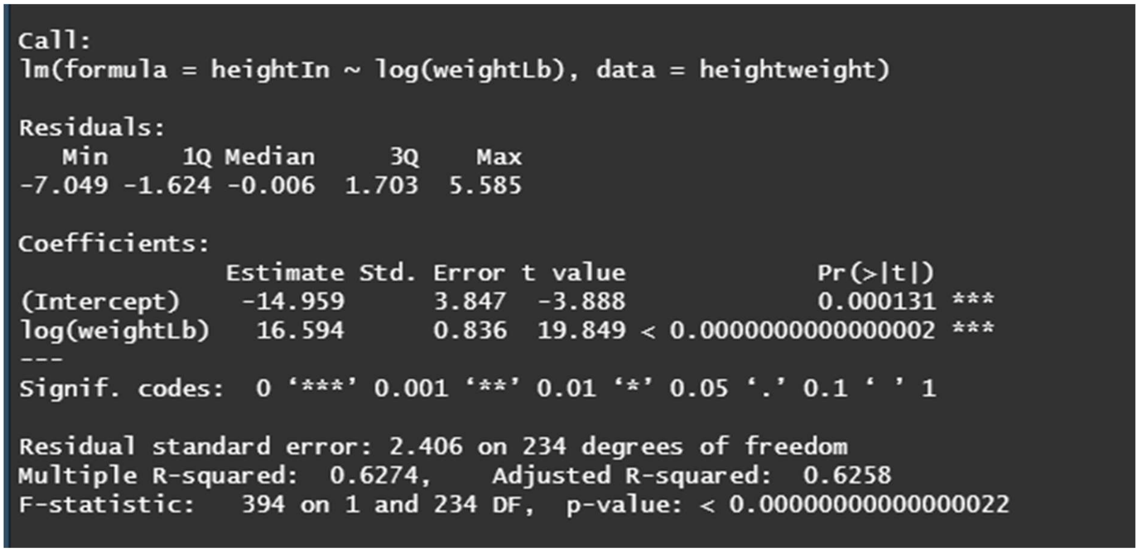
Model summaries:

Model 2:



```
Call:
lm(formula = log(heightIn) ~ weightLb, data = heightweight)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.134190 -0.027793 -0.000201  0.031031  0.089394 

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 3.8473463  0.0144789  265.72 <0.0000000000000002 ***
weightLb    0.0026433  0.0001409   18.76 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0409 on 234 degrees of freedom
Multiple R-squared:  0.6006,	Adjusted R-squared:  0.5989 
F-statistic: 351.9 on 1 and 234 DF,  p-value: < 0.00000000000000022
```

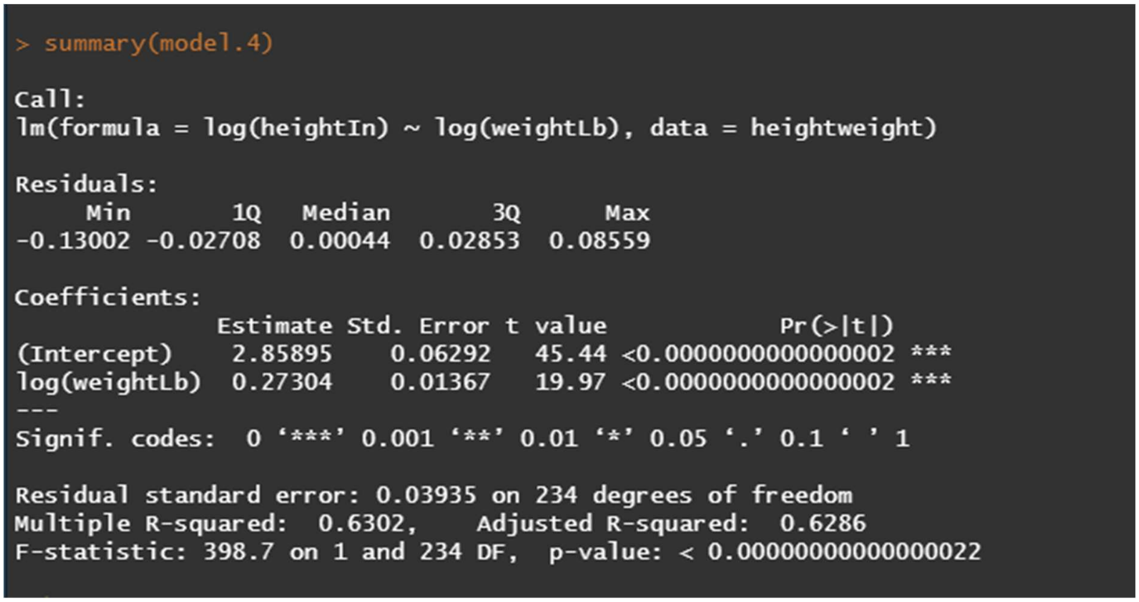
Model 3:



```
Call:
lm(formula = heightIn ~ log(weightLb), data = heightweight)

Residuals:
   Min     1Q Median     3Q    Max 
-7.049 -1.624 -0.006  1.703  5.585 

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)    -14.959      3.847  -3.888             0.000131 ***
log(weightLb)   16.594      0.836  19.849 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.406 on 234 degrees of freedom
Multiple R-squared:  0.6274,	Adjusted R-squared:  0.6258 
F-statistic:   394 on 1 and 234 DF,  p-value: < 0.00000000000000022
```

Model 3’s log(weightLb) is statistically significant, hence there is a relationship between the predictor and the response variable.

Model 4:



```
> summary(model.4)

Call:
lm(formula = log(heightIn) ~ log(weightLb), data = heightweight)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13002 -0.02708  0.00044  0.02853  0.08559 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept)    2.85895    0.06292   45.44 <0.0000000000000002 ***
log(weightLb)  0.27304    0.01367   19.97 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03935 on 234 degrees of freedom
Multiple R-squared:  0.6302,	Adjusted R-squared:  0.6286 
F-statistic: 398.7 on 1 and 234 DF,  p-value: < 0.00000000000000022
```

References

[1] E. Hehman and S. Y. Xie, ‘Doing better data visualization’, *Adv. Methods Pract. Psychol. Sci.*, vol. 4, no. 4, p. 25152459211045336, Oct. 2021, doi: 10.1177/25152459211045334.