ASSIGNMENT 2: ***SUPERMARKET PRICE COMPARISIONS***

OVERVIEW

This report is an investigation of the price difference of identical products that are sold in Coles and Woolworths, two supermarkets in Australia. The report will analyse the average price differences for all those products as well as the difference between supermarkets in various categories that shows which supermarket is cheaper. The data of the stores across various categories was collected from respective online platforms and has been entered manually into a CSV-file. The sample data contains 180 observations with product name, product category and product prices for both Coles and Woolworths. To ensure the data is collected randomly, we randomly selected 1 to 3 products that contains prices information for both Coles and Woolworths on each page of websites to guarantee that sample could contains different products as possible and make the sample to be more representative. Home brands were also excluded, as it was not possible to ensure identical quality of products from each supermarket's home brand.

The data was cleaned and analysed using R. After finishing the hypothesis test, the report has found that there is no significant difference in average price between Coles and Woolworths for the products in each of the five categories. In addition, this analysis supports the fact that the average price of Coles and Woolworths are not different in each categories and all products as a whole.

LIBRARY IMPORTING AND LOADING PACKAGES:

```
# installing packages
Install.packages("dplyr")
Install.packages("readr")
Install.packages("magrittr")
Install.packages("ggplot2")
Install.packages("Car")
# dplyr included
library(dplyr)
library(readr)
library(magrittr)
library(lattice)
library(ggplot2)
library(Car)
```

```
#importing dataset
sm<-supermarket.csv
```

We have imported required packages for the data exploration and analysation.

Summary Statistics:

```
sm %>% group_by('product type') %>% summarise(Min = min(coles,na.rm = TRUE),
                                Max=max(coles,na.rm=TRUE),
                                Q1 = quantile(coles,probs = .25,na.rm = TRUE),
                               Median = median(coles, na.rm = TRUE),
                               Q3 = quantile(coles,probs = .75,na.rm = TRUE),
                               Mean = mean(coles, na.rm = TRUE),
                               SD = sd(coles, na.rm = TRUE),
```

```
                        n = n(),
                        Missing = sum(is.na(coles)))
```

```
# A tibble: 8 x 9
  `product type`   Min    Q1 Median    Q3  Mean    SD     n Missing
  <chr>          <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <int>   <int>
1 Baby             2.5  4.45   9    14.5  11.1   7.98    23       0
2 Dairy            2.1  2.8    3.7   4     3.68  1.21    21       0
3 Drinks           1.4 43     53    68.2  57.0  23.1     20       0
4 Food             0.9  3      5.55  7.5   6.09  4.78    25       0
5 Freezer          1.6  4      5     5.5   5.00  1.95    21       0
6 Household        1    3.58   5.35  8.3   6.12  3.48    42       0
7 Pantry           1    2.12   3     4.53  3.68  2.05    19       0
8 Vegetables       1.5  1.98   3     3.75  3.42  2.10     8       0
```

```
sm %>% group_by('product type') %>% summarise(Min = min(woolworths,na.rm = TRUE),
                         Max=max(woolworths,na.rm=TRUE),
                         Q1 = quantile(woolworths,probs = .25,na.rm = TRUE),
                        Median = median(woolworths, na.rm = TRUE),
                        Q3 = quantile(woolworths,probs = .75,na.rm = TRUE),
                        Mean = mean(woolworths, na.rm = TRUE),
                        SD = sd(woolworths, na.rm = TRUE),
                        n = n(),
                        Missing = sum(is.na(woolworths)))
```

```
# A tibble: 8 x 9
  `product type`   Min    Q1 Median    Q3  Mean    SD     n Missing
  <chr>          <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <int>   <int>
1 Baby            2.79  6.80  11    15.2  11.7   7.14    23       0
2 Dairy           1.75  2.65   3.5   4.2   3.50  1.22    21       0
3 Drinks          0.9  45.2   53.5  66.5  55.3  21.2     20       0
4 Food            1.15  3.79   5.55  7     6.13  4.74    25       0
5 Freezer         1.6   3.07   4.25  5.5   4.57  1.81    21       0
6 Household       0.95  3.5    4.82  8.14  6.13  3.83    42       0
7 Pantry          1     2.12   3.95  4.56  3.55  1.64    19       0
8 Vegetables      1.2   1.88   3.25  4.12  3.84  2.91     8
```

```
#create a variable of the difference between coles and woolworths

sm<-sm%>%mutate(differ=coles - woolworths)

#caculate the statistical summary of the differences between price of the same products of col
es and woolworths in different categories.

sm %>% group_by(``) %>% summarise(Min = min(differ,na.rm = TRUE),
                         Max=max(differ,na.rm=TRUE),
                         Q1 = quantile(differ,probs = .25,na.rm = TRUE),
                        Median = median(differ, na.rm = TRUE),
                        Q3 = quantile(differ,probs = .75,na.rm = TRUE),
                        Mean = mean(differ, na.rm = TRUE),
                        SD = sd(differ, na.rm = TRUE),
                        n = n(),
                        Missing = sum(is.na(differ)))
```

```
# A tibble: 8 x 9
  `product type`    Min      Q1   Median    Q3    Mean    Sd     n Missing
  <chr>           <dbl>   <dbl>    <dbl> <dbl>   <dbl> <dbl> <int>   <int>
1 Baby            -6.01 -0.06     0       0.765  0.536  2.96     23       0
2 Dairy           -2    -0.01000  0       0     -0.177  0.973    21       0
3 Drinks         -18    -7.25    -0.75    2     -1.70   7.95     20       0
4 Food            -3    -0.25    -0.01000 0.5    0.032  1.06     25       0
5 Freezer         -4.15 -1       -0.5     0     -0.428  1.45     21       0
6 Household       -2.71 -0.475    0       0.200  0.0102 1.33     42       0
7 Pantry          -4.7  -0.200   -0.01000 0.210 -0.129  1.28     19       0
8 Vegetables      -0.8   0        0.05    0.75   0.412  0.909     8       0
```

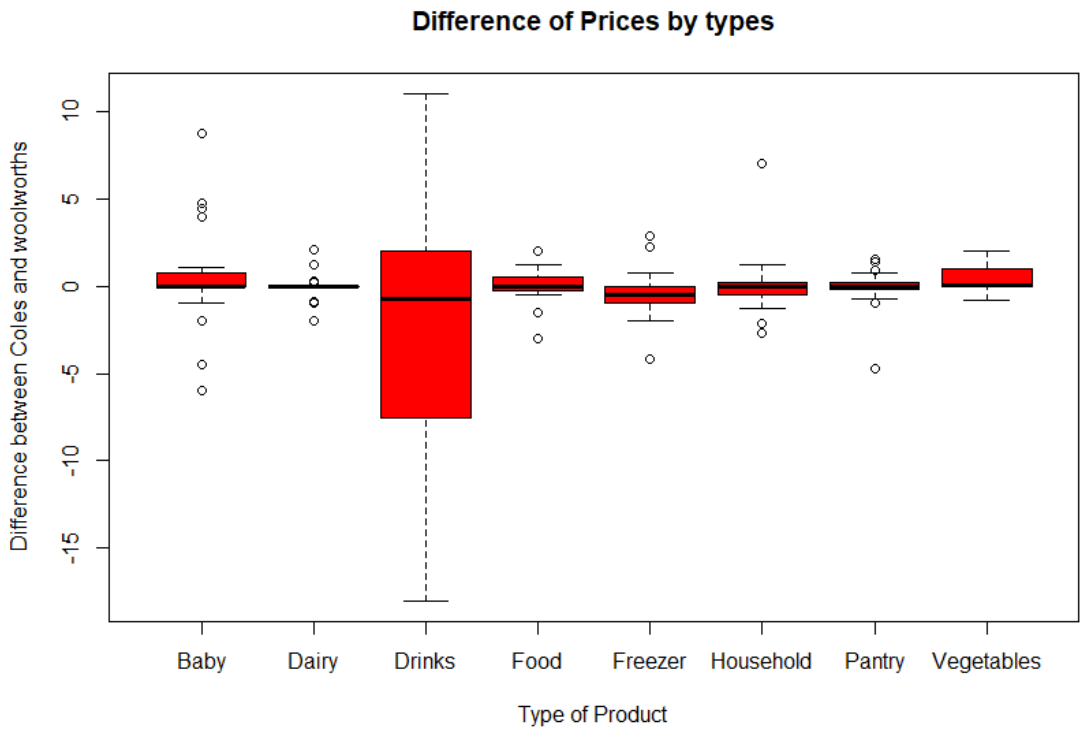
Visualising the results of a paired samples t-test using R studio. The first plot is a boxplot with all the product ty pes for both the supermarkets

```
# This is a chunk for your summary statistics and visualisation code

am %>% boxplot(pricediff ~ `product type`, data = .,ylab="difference between coles and woolwor
th", xlab="category of products",

                  col="red",main="difference of price by categroies",ylim=c(-4,4))

grid()
```



Calculating the statistical summary of the price difference of Coles and Woolworth's across all the products

```
allprices %>% group_by %>% summarise(Min = min(differ,na.rm = TRUE),
                                     Max=max(differ,na.rm=TRUE),
                                     Q1 = quantile(differ,probs = .25,na.rm = TRUE),
                                    Median = median(differ, na.rm = TRUE),
                                    Q3 = quantile(differ,probs = .75,na.rm = TRUE),
                                    Mean = mean(differ, na.rm = TRUE),
                                    SD = sd(differ, na.rm = TRUE),
                                    n = n(),
```

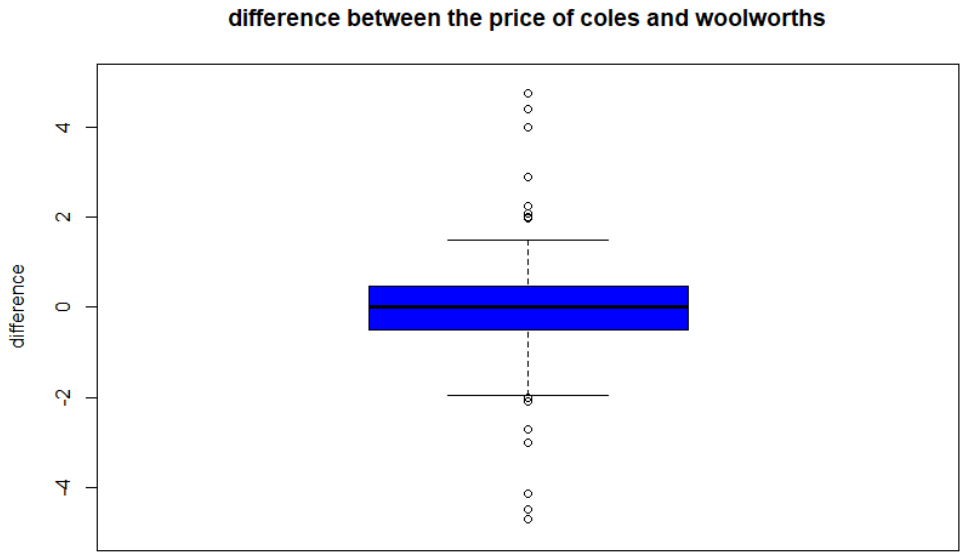
```
                    Missing = sum(is.na(differ)))
```

```
# A tibble: 1 x 8
    Min    Q1 Median    Q3   Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl> <int>   <int>
1   -18  -0.5      0  0.47 -0.180  3.05   179       0
```

The boxplot of the difference of prices between the products in same products in Coles and Woolworths

```
# the boxplot of the difference of price between the same products in coles and woolworths
sm$pricediff%>%boxplot(.,ylab="difference",
                  col="blue",main="difference between the price of coles and woolworth",ylim=c(-4,4))
grid()
```

From the boxplot and scatter plot of graphs, it can be seen that difference of price between Coles and Woolworths is slightly and positively above zero, which indicates that the price of the products in Coles might be slightly higher than that of in Woolworths. The mean of the differences of the 200 observations for sample is 0.180 with standard deviation of 3.05.

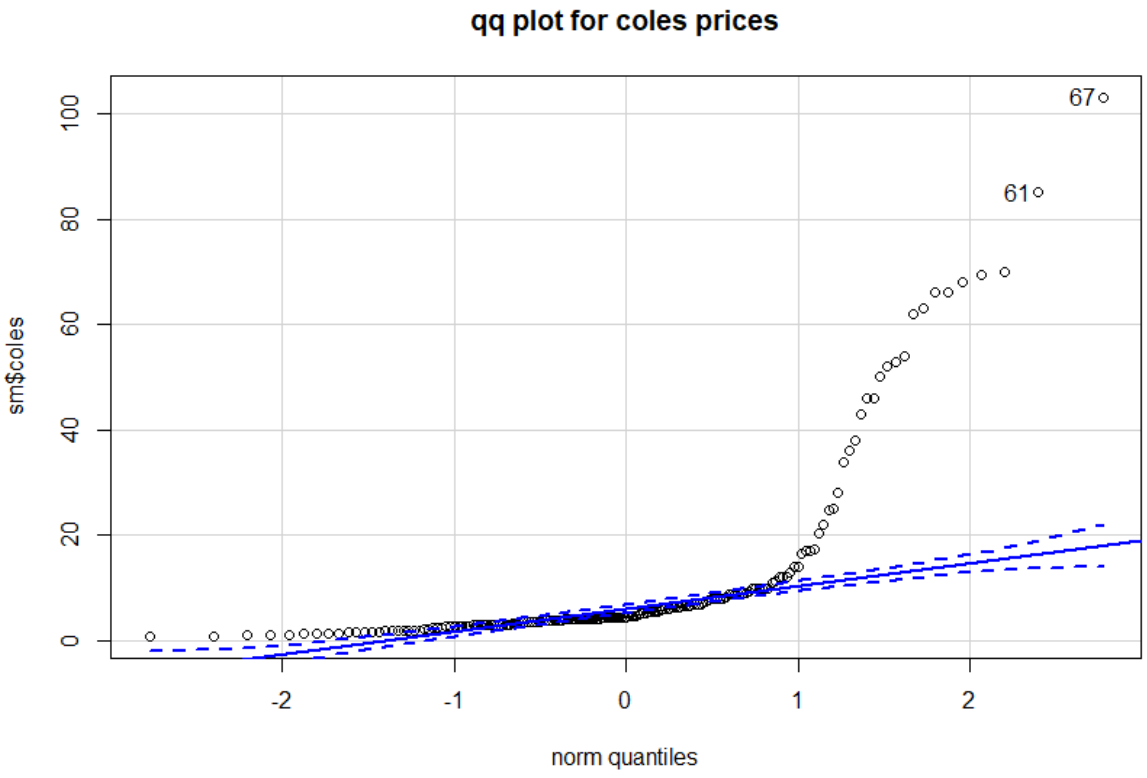


HYPOTHESIS TEST:

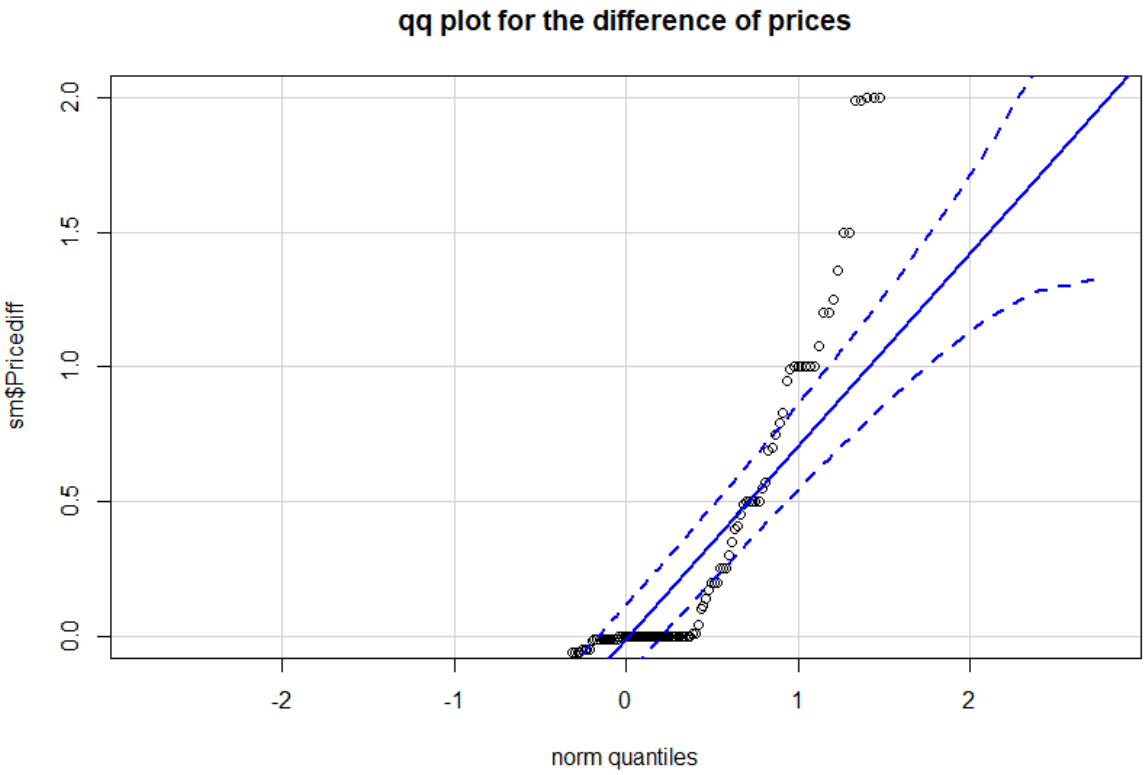
The analysis conducted a hypothesis test to check if there is a price difference between Coles and Woolworths for all sample products. Also, the analysis has conducted a hypothesis test to check if there is average price difference of products between Coles and Woolworths for each of the 5 categories, and if so which supermarket is cheaper. The test is conducted under the assumption that the variances of prices for Coles and Woolworths are unequal and the sample is approximately normally distributed.

```
qqPlot(sm$woolworths, dist = "norm", main = "qq plot for woolworths prices", ylim = c(0, 10))
```


```
qqPlot(sm$coles, dist = "norm", main = "qq plot for coles prices", ylim = c(0, 10))
```



```
qqPlot(sm$Pricediff,dist="norm", main="qqPlot Visulisation for Price mean difference", ylim =
c(0,2))
```




```
t.test(sm$woolworths,sm$coles, var.equal = FALSE, alternative = "less")
```

```
        Welch Two Sample t-test

data:  sm$coles and sm$woolworths
t = -0.094764, df = 355.41, p-value = 0.4623
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.95617
sample estimates:
mean of x mean of y
 11.46296  11.64318
```

```
t.test(sm$woolworths,sm$coles, var.equal = FALSE, alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  sm$coles and sm$woolworths
t = -0.094764, df = 355.41, p-value = 0.9246
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.920463  3.560016
sample estimates:
mean of x mean of y
 11.46296  11.64318
```

```
fridge<-sm%>%filter(`product type`=="Fridge")

t.test(fridge$woolworths,fridge$coles, var.equal = FALSE, alternative = "two.sided")
```

```
	Welch Two Sample t-test

data:  freezer$coles and freezer$woolworths
t = -0.73762, df = 39.791, p-value = 0.4651
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6012612  0.7450707
sample estimates:
mean of x mean of y
 4.574286  5.002381
```

```
food<-sm%>%filter(`product type`=="food")

t.test(food$woolworths,food$coles, var.equal = FALSE, alternative = "two.sided")
```

```
	Welch Two Sample t-test

data:  food$coles and food$woolworths
t = 0.023771, df = 47.998, p-value = 0.9811
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.674728  2.738728
sample estimates:
mean of x mean of y
   6.1256    6.0936
```

```
baby<-allprices%>%filter(`product type`=="baby")

t.test(baby$woolworths,baby$coles, var.equal = FALSE, alternative = "two.sided")
```

```
	Welch Two Sample t-test

data:  baby$coles and baby$woolworths
t = 0.24004, df = 43.461, p-value = 0.8114
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.966421  5.038595
sample estimates:
mean of x mean of y
 11.65565  11.11957
```

```
household<-sm%>%filter(`product type`=="Household")

t.test(household$woolworths,household$coles, var.equal = FALSE, alternative = "two.sided")
```

```
	Welch Two Sample t-test

data:  household$coles and household$woolworths
t = 0.01283, df = 81.25, p-value = 0.9898
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.577409  1.597885
sample estimates:
mean of x mean of y
 6.129286  6.119048
```

```
pantry<-sm%>%filter(`product type`=="Pantry")

t.test(pantry_drinks$woolworths,pantry_drinks$coles, var.equal = FALSE, alternative = "two.sid
ed")
```

```
        Welch Two Sample t-test

data:  pantry$coles and pantry$woolworths
t = -0.21419, df = 34.32, p-value = 0.8317
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.351972  1.094077
sample estimates:
mean of x mean of y
 3.552105  3.681053
```

It has been test the price difference between Coles and Woolworths for all sample products and products in each of the categories.

1. Hypothesis test for price difference between Coles and Woolworths for all sample product prices. Based on unequal variance and the fact that average price of Woolworths for all products is lower than that of Coles, we conduct Welch two-sample test to check if average price of Woolworths is lower than that of Coles. Null hypothesis HO: the average price of products in Coles is equal to that of in Woolworths Alternative hypothesis H1: the average price of products in Woolworths is less than that of in Coles. Significant level, a = 0.05. p-value = 0.4623 > a=0.05, there is a probability of 46.23% that the null hypothesis is true when we reject null hypothesis. There is insufficient evidence to reject the null hypothesis, and it is reasonable to conclude that average price of all the products for Coles is equal to the average price of Woolworths.

2. Hypothesis test for price difference between Coles and Woolworths for each of the categories Conducting Welch Two Sample t-test to check if there is a difference for price of sample products between Coles and Woolworths in each category.

a. Category of fridge, which include Fridge products in the two supermarket. Null hypothesis HO: the average price of products in Woolworths for fridge category is equal to that of in Coles Alternative hypothesis H1: the average price of products in Woolworths in fridge category is unequal to that of in Coles. Significant level, a = 0.05. p-value = 0.4651 > a=0.05, there is a probability of 46.51% that the null hypothesis is true when we reject null hypothesis. Therefore, there is insufficient evidence to reject the null hypothesis, and it is reasonable to conclude that average product price of Woolworths in fridge category is equal to average product price of Coles.

b. Category of baby, which include relative baby health care and adult skin care products in the two supermarket. Null hypothesis HO: the average price of products in Woolworths for baby health category is equal to that of in Coles Alternative hypothesis H1: the average price of products in Woolworths in baby health category is unequal to that of in Coles. Significant level, a = 0.05. p-value = 0.8114 > a=0.05, there is a probability of 81.84% that the null hypothesis is true when we reject null hypothesis. Therefore, there is insufficient evidence to reject the null hypothesis, and it is reasonable to conclude that average product price of Woolworths in baby health category is equal to average product price of Coles.

c. Category of household, which include relative entertainment clothing, household and pet products in the two supermarket. Null hypothesis HO: the average price of products in Woolworths for household category is equal to that of in Coles Alternative hypothesis H1: the average price of products in Woolworths in household category is unequal to that of in Coles. Significant level, a = 0.05. p-value = 0.9898 > a=0.05, there is a probability of 98.98% that the null hypothesis is true when we reject null hypothesis. Therefore, there is insufficient evidence to reject the null hypothesis, and it is reasonable to conclude that average product price of Woolworths in household category is equal to average product price of Coles.

d. Category of Pantry, which include relative pantry drinks products in the two supermarket. Null hypothesis HO: the average price of products in Woolworths for pantry and drinks category is equal to that of in Coles Alternative hypothesis H1: the average price of products in Woolworths in pantry and drinks category is unequal to that of in Coles. significant level, a = 0.05. p-value =0.8317 > a=0.05, there is a probability of 83.17% that the null hypothesis is true when we reject null hypothesis. Therefore, there is insufficient evidence to reject null hypothesis, and it is reasonable to conclude the average product price of Woolworths in Pantry category is equal to the average product price of Coles.

e. Category of food in the two supermarket. Null hypothesis HO: the average price of products in Woolworths for food category is equal to that of in Coles Alternative hypothesis H1: the average price of products in Woolworths in foods category is unequal to that of in Coles. significant level, a = 0.05. p-value =0.9811 > a=0.05, there is a probability of 98.11% that the null hypothesis is true when we reject null hypothesis.

In summary, there is no significant difference in average price between Coles and Woolworths for both all sample products and the products in each of the 5 categories. And this analysis shows that fact that the average price of Coles and Woolworths are not different in each categories and all products as a whole and maintain competitive prices across the stores.

INTERPRETATION:

The investigation shows the difference between the products of the two supermarkets is slightly different and not deviated towards any supermarket. Limitations of the investigation included the number of samples collected as well as the narrow scope with a primary focus on only few product categories. Additionally, the data sets are not available publicly and so data had to be collected and entered manually. The investigation could be improved by considering price changes over time and inclusion of data from different days. Given more time and resources, ideally the investigation would compare all categories of products, with products randomly selected for analysis. The availability of data sets would also improve the process of the investigation.