Probability Theory and Statistics

ALY6010, WINTER 2022

Week -4

Module-2- R Practice Output

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Prof. Tom Beur

Date: 03/20/2022

Introduction to Avocado:

Avocado is produced by a tree that originated in America and is primarily found in the highlands of south-central Mexico and Guatemala. Avocado is the name of the fruit (or alligator pear or pear). Various cultivation locations such as Mexico, California, Peru, and Chile produce the largest avocado in the world, with a total area occupied by avocado of 188,723 hectares (415520 acres) in 2013 and total harvesting of 2.03 million tons in 2017. Avocados were introduced to California from Mexico in the 19th century and are regarded for being the most profitable cash crop in the United States, covering 24k hectares (59k acres). Avocado output is 95 percent in Southern California, and 60 percent in San Diego County..

It has certain elements that aid in the preservation of the raw avocado's 73 % water content, 2 % protein, 15% fat, and 9% carbohydrate content. Per 100g of raw avocado, there is the nutritional value (3.5 oz). Vitamin B-28 percent, vitamin K-20 percent, vitamin C-(10-19) percent, vitamin E, and potassium are all present.

Let’s discuss more the descriptive analysis and various functions in R generating tables and plots like a scatterplot, jitter chart, and boxplot. However, implementing the t-test on the avocado dataset with a two-sample hypothesis test.

Avocado sales data available from 2015 to 2018 indicate the average price per unit cost, with avocados distributed in bags. The PLU (Product lookup codes) are just to the size of avocados, with conventional and organic varieties like small, large available. Portland and Raleigh are popular in many areas. Greensboro, RichmondNorfolk, Roanoke, Roanoke, SanDiego, SanFrancisco, Seattle, and then the total number of bags sold in each region (small bags, large bags, extra-large bags).

Task 1: In this task import the dataset of avocado which has average price, total volume, as per the PLU 4046,4225,4770, type of avocado with the price or amount of for conventional or organic

Whereas:

X4046 is the number of small avocados sold

X4225 is the number of medium avocados sold

X4770 is the number of large avocados sold

X	Date	AveragePrice	Total.Volume	X4046	X4225	X4770	Total.Bags	Small.Bags	Large.Bags	XLarge.Bags	type	year	region
0	12/27/2015	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.00	conventional	2015	Albany
1	12/20/2015	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.00	conventional	2015	Albany
2	12/13/2015	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.00	conventional	2015	Albany
3	12/6/2015	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.00	conventional	2015	Albany
4	11/29/2015	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.00	conventional	2015	Albany
5	11/22/2015	1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0.00	conventional	2015	Albany
6	11/15/2015	0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0.00	conventional	2015	Albany
7	11/8/2015	0.98	109428.33	703.75	101815.36	80.00	6829.22	6266.85	562.37	0.00	conventional	2015	Albany
8	11/1/2015	1.02	99811.42	1022.15	87315.57	85.34	11388.36	11104.53	283.83	0.00	conventional	2015	Albany
9	10/25/2015	1.07	74338.76	842.40	64757.44	113.00	8625.92	8061.47	564.45	0.00	conventional	2015	Albany
10	10/18/2015	1.12	84843.44	924.86	75595.85	117.07	8205.66	7877.86	327.80	0.00	conventional	2015	Albany
11	10/11/2015	1.28	64489.17	1582.03	52677.92	105.32	10123.90	9866.27	257.63	0.00	conventional	2015	Albany
12	10/4/2015	1.31	61007.10	2268.32	49880.67	101.36	8756.75	8379.98	376.77	0.00	conventional	2015	Albany
13	9/27/2015	0.99	106803.39	1204.88	99409.21	154.84	6034.46	5888.87	145.59	0.00	conventional	2015	Albany
14	9/20/2015	1.33	69759.01	1028.03	59313.12	150.50	9267.36	8489.10	778.26	0.00	conventional	2015	Albany
15	9/13/2015	1.28	76111.27	985.73	65696.86	142.00	9286.68	8665.19	621.49	0.00	conventional	2015	Albany
16	9/6/2015	1.11	99172.96	879.45	90062.62	240.79	7990.10	7762.87	227.23	0.00	conventional	2015	Albany
17	8/30/2015	1.07	105693.84	689.01	94362.67	335.43	10306.73	10218.93	87.80	0.00	conventional	2015	Albany
18	8/23/2015	1.34	79992.09	733.16	67933.79	444.78	10880.36	10745.79	134.57	0.00	conventional	2015	Albany
19	8/16/2015	1.33	80043.78	539.65	68666.01	394.90	10443.22	10297.68	145.54	0.00	conventional	2015	Albany

Task2: Str() In this task analyzing the structure of the dataset which has 18249 observation and 14 variables which consist of various information about the avocados sold with the average price, total volume, x4046, x4225, x4770, total bags which contain small bags, large bags, x large bags, the type of avocado whether is conventional or organic, sales per year from 2015-2018 and region avocado sold

```
> str(df)
'data.frame':   18249 obs. of  14 variables:
 $ X           : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Date        : chr  "12/27/2015" "12/20/2015" "12/13/2015" "12/6/2015" ...
 $ AveragePrice: num  1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...
 $ Total.Volume: num  64237 54877 118220 78992 51040 ...
 $ X4046       : num  1037 674 795 1132 941 ...
 $ X4225       : num  54455 44639 109150 71976 43838 ...
 $ X4770       : num  48.2 58.3 130.5 72.6 75.8 ...
 $ Total.Bags  : num  8697 9506 8145 5811 6184 ...
 $ Small.Bags  : num  8604 9408 8042 5677 5986 ...
 $ Large.Bags  : num  93.2 97.5 103.1 133.8 197.7 ...
 $ XLarge.Bags : num  0 0 0 0 0 0 0 0 0 0 ...
 $ type        : chr  "conventional" "conventional" "conventional" "conventional" ...
 $ year        : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ region      : chr  "Albany" "Albany" "Albany" "Albany" ...
>
```

Task3: st() In this task installing with vtable library to extract the mean, median, sd, N overview about the which clearly shows that we have the type of avocado available i.e. conventional and organic which has 50% and 50% observation

sumtable {vtable}

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X	18249	24.232	15.481	0	10	38	52
AveragePrice	18249	1.406	0.403	0.44	1.1	1.66	3.25
Total.Volume	18249	850644.013	3453545.355	84.56	10838.58	432962.29	62505646.52
X4046	18249	293008.425	1264989.082	0	854.07	111020.2	22743616.17
X4225	18249	295154.568	1204120.401	0	3008.78	150206.86	20470572.61
X4770	18249	22839.736	107464.068	0	0	6243.42	2546439.11
Total.Bags	18249	239639.202	986242.399	0	5088.64	110783.37	19373134.37
Small.Bags	18249	182194.687	746178.515	0	2849.42	83337.67	13384586.8
Large.Bags	18249	54338.088	243965.965	0	127.47	22029.25	5719096.61
XLarge.Bags	18249	3106.427	17692.895	0	0	132.5	551693.65
type	18249						
... conventional	9126	50%					
... organic	9123	50%					
year	18249	2016.148	0.94	2015	2015	2017	2018

Task4: In this task analysing the dataset using psych() to understand the comparison among the variables

```
> psych::describe(df)
              vars     n       mean         sd    median   trimmed       mad     min         max       range  skew kurtosis       se
X                1 18249      24.23      15.48     24.00     23.96     20.76    0.00       52.00       52.00  0.11    -1.25     0.11
Date*            2 18249      85.00      48.79     85.00     85.00     62.27    1.00      169.00      168.00  0.00    -1.20     0.36
AveragePrice     3 18249       1.41       0.40      1.37      1.38      0.42    0.44        3.25        2.81  0.58     0.32     0.00
Total.Volume     4 18249  850644.01 3453545.36 107376.76 232479.38 152652.16   84.56 62505646.52 62505561.96  9.01    92.07 25564.99
X4046            5 18249  293008.42 1264989.08   8645.30  58604.80  12775.10    0.00 22743616.17 22743616.17  8.65    86.78  9364.13
X4225            6 18249  295154.57 1204120.40  29061.02  80079.89  42285.68    0.00 20470572.61 20470572.61  8.94    91.91  8913.54
X4770            7 18249   22839.74  107464.07    184.99   3375.02    274.27    0.00  2546439.11  2546439.11 10.16   132.51   795.51
Total.Bags       8 18249  239639.20  986242.40  39743.83  67480.14  55300.92    0.00 19373134.37 19373134.37  9.75   112.23  7300.69
Small.Bags       9 18249  182194.69  746178.51  26362.82  49175.03  37953.80    0.00 13384586.80 13384586.80  9.54   106.97  5523.61
Large.Bags      10 18249   54338.09  243965.96   2647.71  12057.05   3925.49    0.00  5719096.61  5719096.61  9.79   117.95  1805.97
XLarge.Bags     11 18249    3106.43   17692.89      0.00    243.54      0.00    0.00   551693.65   551693.65 13.14   233.51   130.97
type*           12 18249       1.50       0.50      1.00      1.50      0.00    1.00        2.00        1.00  0.00    -2.00     0.00
year            13 18249    2016.15       0.94   2016.00   2016.10      1.48 2015.00     2018.00        3.00  0.22    -1.03     0.01
region*         14 18249      27.50      15.58     27.00     27.50     19.27    1.00       54.00       53.00  0.00    -1.20     0.12
>
```

Task5: In this task used Group_by () Comparison between the type and year comparison with averageprice of the avocado to get the observation with the precision to get the information about the dataset which help in analyzing the graphs for visualization. Whereas, we can define the Mean, Median, SD, Min, Max

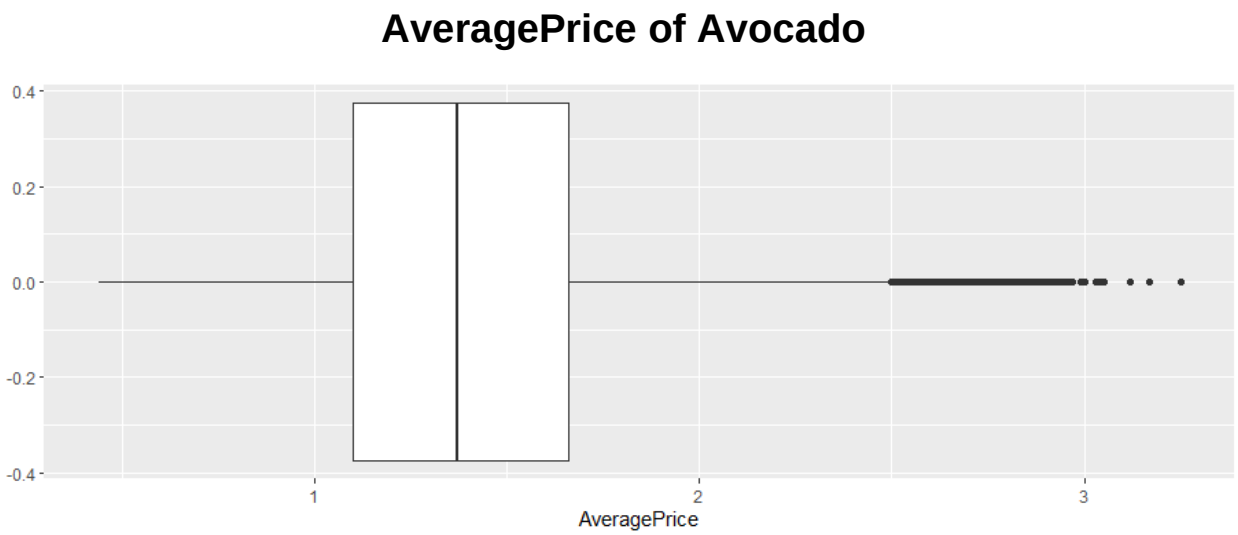
```
> #Type by average price group_by function
> df1<-df %>%
+   group_by(type) %>%
+   summarise_at(vars(AveragePrice), list( Mean = mean, Median = median, SD = sd, Min = min, Max = max))
> df1
# A tibble: 2 x 6
  type          Mean Median    SD   Min   Max
  <chr>        <dbl>  <dbl> <dbl> <dbl> <dbl>
1 conventional  1.16   1.13 0.263  0.46  2.22
2 organic       1.65   1.63 0.364  0.44  3.25
> #yearly average price
> df2<-df %>%
+   group_by(year) %>%
+   summarise_at(vars(AveragePrice), list( Mean = mean, Median = median, SD = sd, Min = min, Max = max))
> df2
# A tibble: 4 x 6
   year  Mean Median    SD   Min   Max
  <int> <dbl>  <dbl> <dbl> <dbl> <dbl>
1  2015  1.38   1.3  0.376  0.49  2.79
2  2016  1.34   1.3  0.394  0.51  3.25
3  2017  1.52   1.49 0.433  0.44  3.17
4  2018  1.35   1.35 0.306  0.56  2.3
> 
```

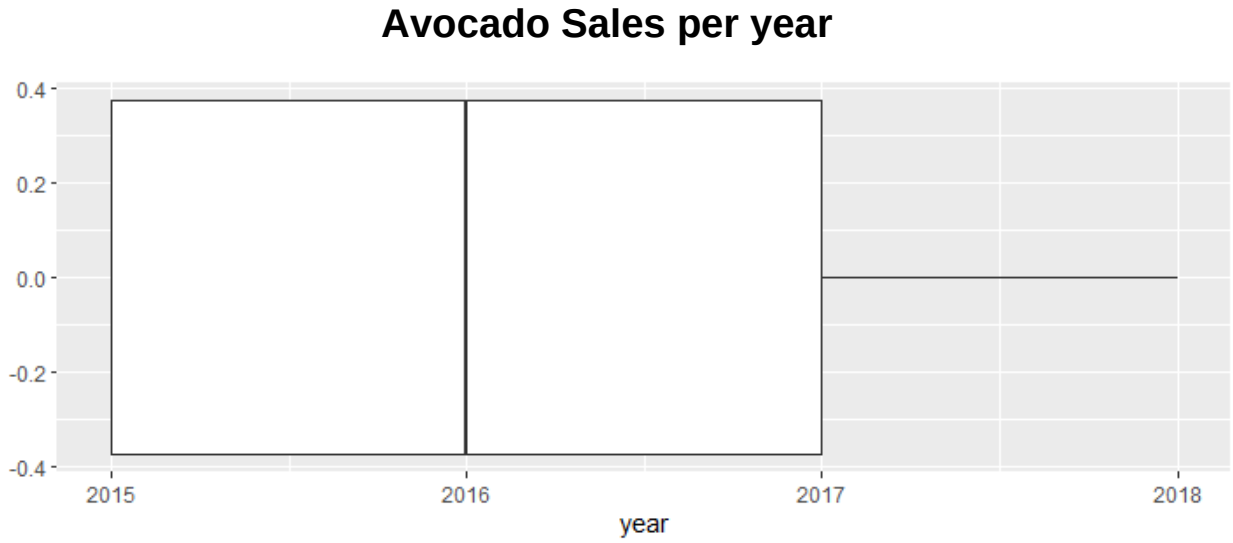
Task 6: **Two Sample t-test** -In this task implementing the t-test after grouping between the year and the average price which helps in determining the p value, which is 2.2, whereas the degree of freedom is 6, which show that the alternative hypothesis true is the difference in the mean is not equal to 0

```
> # t-test by group
> df_t_test <- t.test(df6$year, df6$AveragePrice, var.equal = TRUE)
> df_t_test

        Two Sample t-test

data:  df6$year and df6$AveragePrice
t = 3120.9, df = 6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2014.419 2017.581
sample estimates:
mean of x mean of y
   2016.5       0.5
```

Task 7: **Boxplot-** Sales of the Avocado is less above the 2.5 which we can see in the boxplot with the outliers

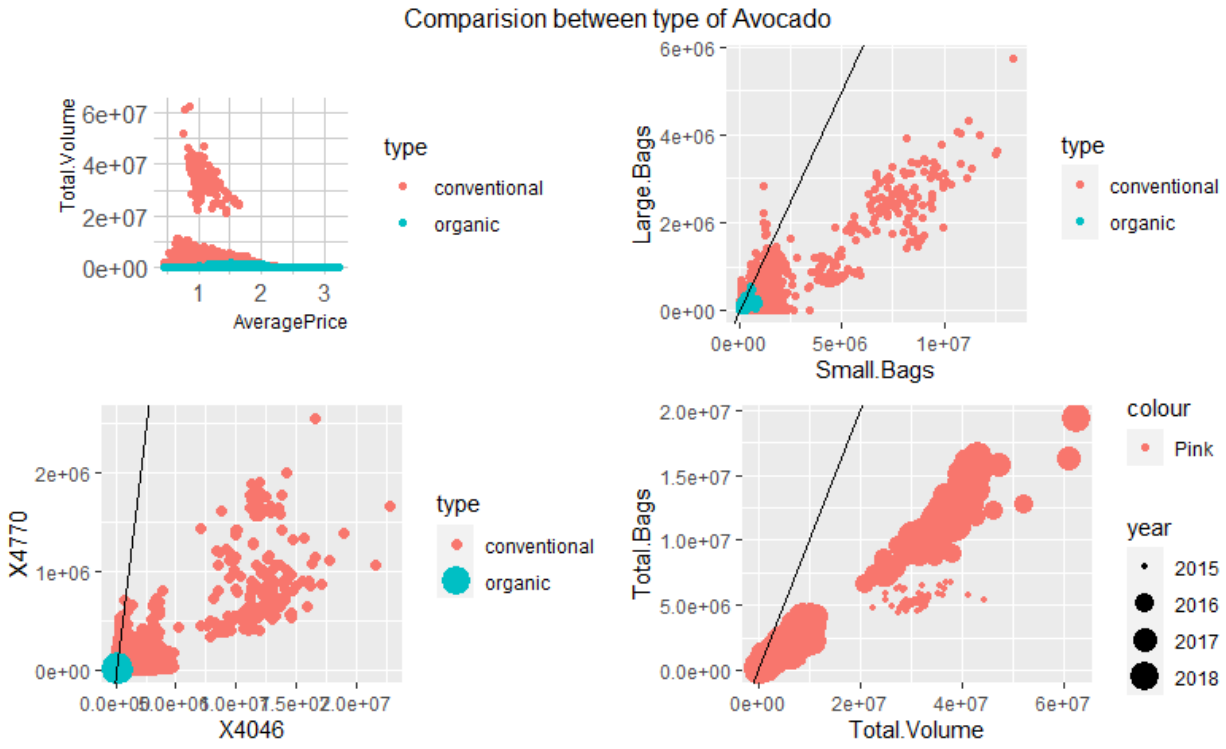
```
#BoxPlot
ggplot(df, aes(AveragePrice))+
  geom_boxplot()
ggplot(df, aes(year))+
  geom_boxplot()
```





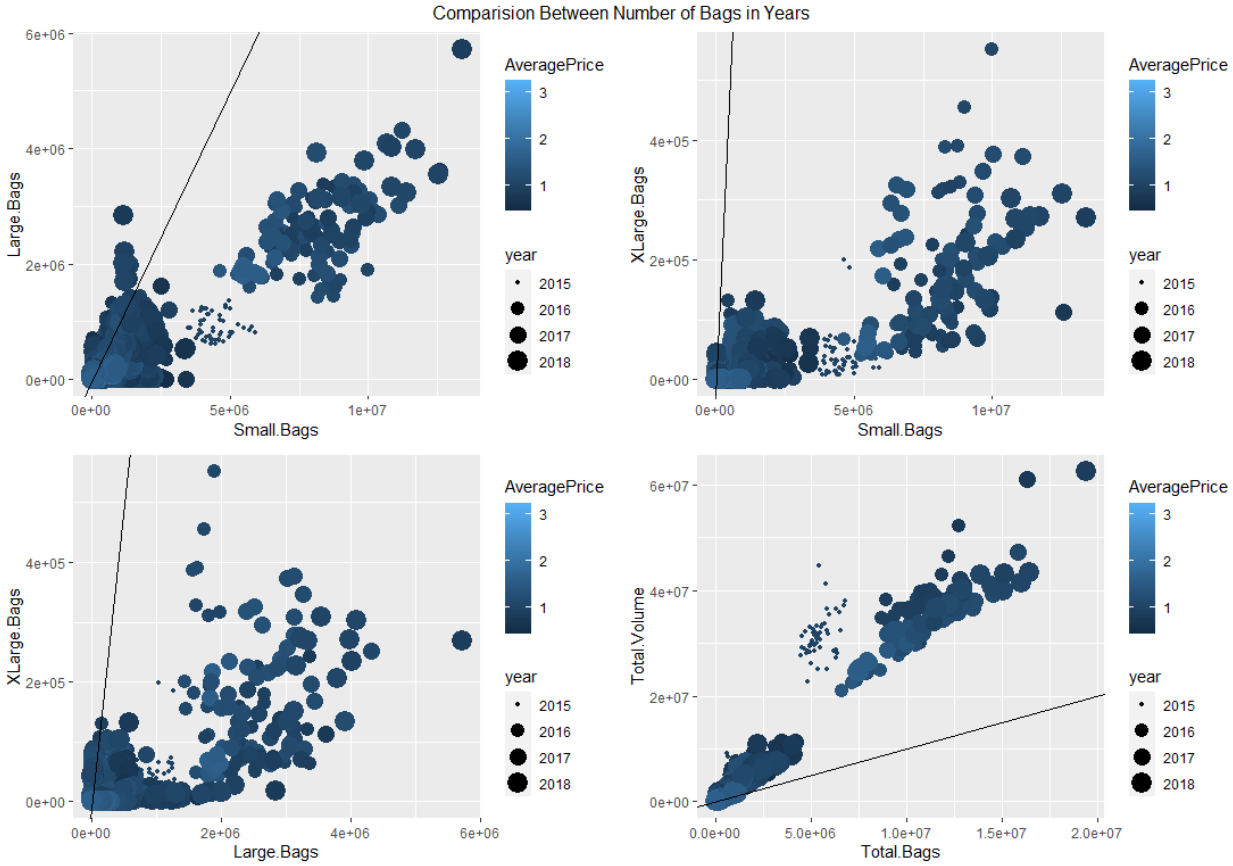
Task8: **Scatterplot** – In this task, analyzing the comparison of type of avocado with various factors and which will tell about the relationship between the variables affected about the sales of avocado.

1. Top left graph shows the comparison between average price and total volume of the sales with the different types of avocados like conventional and organic
2. Right top graph gives the understanding of about the small bags and large bags of with the conventional and organic avocado.
3. Left bottom gives the clear observation about the demand of the conventional and organic avocado with PLU x4770 and x4046.
4. Bottom right of the graph shows the total volume vs total bags sold in a particular in the year 2018, 2017, 2016,2015



Task9:JitterPlot - In this task comparing the average bags sold of Avocados in different years. Understading about the demand of avocado in particular year with compariso with the average price depending on the large bags, small bags, xlarge bags

1. Top left jitter plot explain about the comparison between small bags and large bags with abline.
2. Bottom left jitter plot clearly shows the comparison between the largebags and the xlarge bags with the abline.
3. Top right shows the comparison between about the small bags and x-large bags sold in particular year in average price.
4. Bottom right jitter plot gives the understanding of the total bags and total volume sold in particular year with averageprice



Task10: t-test- In this task performing the t-test gives the understanding of statistics. It helps in analyzing the mean of two groups are similarly giving the results. T- test of the two sample gives the normal distribution with equal variances.

We can perform t- test on two sample test with independent variables by using the t-test function in R which will help in understanding the hypothesis testing about the group of sample.

In this dataset we have performed the following t-test with the variables in the group total volume of the avocado sold and the averageprice of the avocados.

```
> #t-test
> df_t_test2 <- t.test(df$Total.Volume, df$AveragePrice, var.equal = FALSE, conf.level = .95)
> df_t_test2

        Welch Two Sample t-test

data:  df$Total.Volume and df$AveragePrice
t = 33.274, df = 18248, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 800532.8 900752.4
sample estimates:
   mean of x    mean of y
8.506440e+05 1.405978e+00
```

```
> #t-test
> df_t_test2 <- t.test(df$Total.Volume, df$AveragePrice, var.equal = TRUE, conf.level = .95)
> df_t_test2

        Two Sample t-test

data:  df$Total.Volume and df$AveragePrice
t = 33.274, df = 36496, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 800534.5 900750.7
sample estimates:
   mean of x    mean of y
8.506440e+05 1.405978e+00
```

In conclusion, of the above t-test for hypothesis testing we have assumed the two samples as average price and total volume sold. The t value is 33.274, degree of freedom is 18248, p-value is 2.2 whereas the significance level of the two samples is 2.2 in exponential which suggest the null hypothesis is accepted as the p- value is not less than 0.5. there is no need to do the alternative hypothesis in the two sample test. While findin the variables invariable true and false the degree of freedom changes which depend on the sample of test results whether the sample is above 30 the critical value is close to the standard normal critical value.

Reference:

[1]ggplot2 - Easy Way to Mix Multiple Graphs on The Same Page

Kassambara et al.

http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/

[2]Side-by-side plots with ggplot2

Christopher DuBoisChristopher DuBois 40.2k2323 gold badges6868 silver badges9292 bronze badges et al.

https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2

[3]Marginal distribution with ggplot2 and ggExtra

Holtz

https://r-graph-gallery.com/277-marginal-histogram-for-ggplot2.html

[4]Boxplot

Holtz

https://r-graph-gallery.com/scatterplot.html

[5]Unpaired Two-Samples T-test in R

http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

[6]How to Use the Jitter Function in R for Scatterplots

Zach

https://www.statology.org/jitter-function-r/

[7]Scatter plot by group in R

https://r-charts.com/correlation/scatter-plot-group/

[8]Group by one or more variables - group_by

https://dplyr.tidyverse.org/reference/group_by.html

[9]dplyr.com is for sale

https://www.hugedomains.com/domain_profile.cfm?d=dplyr.com

[10]ggplot2 histogram plot : Quick start guide - R software and data visualization

http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

[11]http://www.bobbywires.com/plu-1.php?S=P&L=A&V=AVOCADOS

Appendix:

```
#============================
# R_Practise Module 2 Week 4
#============================

install.packages("ggplot2")
install.packages("tidyverse")
install.packages("dplyr")
install.packages("plyr")
install.packages("vtable")

library(ggplot2)
library(tidyverse)
library(dplyr)
library(plyr)
library(vtable)
library(hrbrthemes)
library(ggExtra)
library(epiDisplay)
library(gridExtra)
df <- read.csv("C:\\Abhinav _ NEU BOSTON\\ALY6010 Probability Theory & Intro Statistics\\Module 4\\avocado_prices_20220315.csv")
df

str(df)
st(df)
psych::describe(df)

#Type by average price group_by function
df1<-df %>%
  group_by(type) %>%
  summarise_at(vars(AveragePrice), list( Mean = mean, Median = median, SD = sd, Min = min, Max = max))
df1

#yearly average price
df2<-df %>%
  group_by(year) %>%
  summarise_at(vars(AveragePrice), list( Mean = mean, Median = median, SD = sd, Min = min, Max = max))
df2

# t-test by group
df_t_test <- t.test(df1$year, df1$AveragePrice, var.equal = TRUE)
df_t_test

#Type by average price group_by function
df1<-df %>%
  group_by(type) %>%
  summarise_at(vars(AveragePrice), list( AveragePrice = mean ))
df1
df3<-df %>%
  group_by(type) %>%
  summarise_at(vars(AveragePrice), list( AveragePrice = sd ))
df3
```

```
#yearly average price
df4<-df %>%
  group_by(year)%>%
  summarise_at(vars(AveragePrice), list( AveragePrice = mean)) %>%
  filter(AveragePrice == max(AveragePrice))
df4

#Maximum
df5<-df %>%
  group_by(year)%>%
  summarise_at(vars(AveragePrice), list( AveragePrice = max))
df5

#Minimum
df6<-df %>%
  group_by(year)%>%
  summarise_at(vars(AveragePrice), list( AveragePrice = min))
df6

# t-test by group
df_t_test <- t.test(df6$year, df6$AveragePrice, var.equal = TRUE)
df_t_test

#BoxPlot
ggplot(df, aes(AveragePrice))+
  geom_boxplot()
ggplot(df, aes(year))+
  geom_boxplot()+

#Scatter plot
ggplot1<-ggplot(df, aes(AveragePrice, Total.Volume, color =type))+
  geom_point()+
  geom_abline()


#Scatter Plot with abline
ggplot2<-ggplot(df, aes(Small.Bags, Large.Bags, color = type))+
  geom_point()+
  geom_abline()

#Scatter Plot with type
ggplot3<-ggplot(df, aes(X4046, X4770, size = type, color = type))+
  geom_point()+
  geom_abline()

#Scatter Plot with year

ggplot4<-ggplot(df, aes(Total.Volume, Total.Bags, color = type))+
  geom_point()+
  geom_abline()
grid.arrange(ggplot1, ggplot2, ggplot3, ggplot4, ncol =2, nrow =2, top = 'Comparision between type of Avocado')
```

```
#Scatter Plot with abline
ggplot2<-ggplot(df, aes(Small.Bags, Large.Bags, color = type))+
  geom_point()+
  geom_abline()

#Scatter Plot with type
ggplot3<-ggplot(df, aes(X4046, X4770, size = type, color = type))+
  geom_point()+
  geom_abline()

#Scatter Plot with year

ggplot4<-ggplot(df, aes(Total.Volume, Total.Bags, color = type))+
  geom_point()+
  geom_abline()
grid.arrange(ggplot1, ggplot2, ggplot3, ggplot4, ncol =2, nrow =2, top = 'Comparision between type of Avocado')


#Jitterplot
ggplot5<-ggplot(df, aes(Small.Bags, Large.Bags, size = year, color = AveragePrice))+
  geom_jitter()+
  geom_abline()
ggplot5

ggplot6<-ggplot(df, aes(Small.Bags, XLarge.Bags, size = year, color = AveragePrice))+
  geom_jitter()+
  geom_abline()
ggplot6

ggplot7<-ggplot(df, aes(Large.Bags, XLarge.Bags, size = year, color = AveragePrice))+
  geom_jitter()+
  geom_abline()
ggplot7

ggplot7<-ggplot(df, aes(Total.Bags, Total.Volume, size = year, color = AveragePrice))+
  geom_jitter()+
  geom_abline()
ggplot7

grid.arrange(ggplot5, ggplot6, ggplot7, ggplot8, ncol =2, nrow =2, top = "Comparision Between AveragePrice with Number of Bags")
#t-test
df_t_test2 <- t.test(df$year, df$AveragePrice, var.equal = TRUE)
df_t_test2
```