Presentation Assignment

TEAM 13

4/27/2022

Import the data and load the required packages

suppressPackageStartupMessages(library(ggplot2))  
suppressPackageStartupMessages(library(tidyverse))  
suppressPackageStartupMessages(library(corrplot))

## Warning: package 'corrplot' was built under R version 4.1.3

suppressPackageStartupMessages(library(readxl))  
#df = read.csv(file.choose(),header = T,stringsAsFactors = T)  
farmdat= read\_excel("stat\_380\_studentFarm dataFinal.xlsx")  
farmdat2 = read\_excel("stat\_380\_studentFarm dataFinal.xlsx")

## INTRODUCTION

The primary goal of this study is to identify factors that impact the quantity and unit pricing of tomatoes. To find the factors that influence the quantity and unit price of the tomatoes, we must perform an analysis of variance, Pearson’s correlation, and exploratory data analysis. The following are the research questions:

1. Does tomato sub-unit have any impact on the tomato quantity?
2. Is there a relationship between tomato sub-unit and unit price of the tomato?
3. Is there a relationship between unit price and variety
4. Is there a relationship between tomato variety and tomato quantity

# Description of the data

str(farmdat)

## tibble [314 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ Date : chr [1:314] "7.30.21" "9.3.21" "10.5.21" "42593" ...  
## $ Invoice# : chr [1:314] "21048" "21114" "21160" "16016" ...  
## $ Client : chr [1:314] "HFS" "HFS" "HFS" "HFS" ...  
## $ Sub\_Unit : chr [1:314] "Pollock" "Cafe Laura" "BJC" "Pollock" ...  
## $ Crop : chr [1:314] "Tomatoes" "Tomatoes" "Tomatoes" "Tomatoes" ...  
## $ Variety : chr [1:314] "Cherry" "Cherry" "Slicers" "Heirloom" ...  
## $ Quantity : num [1:314] 5 5 5 5 5 5 5 5 5.2 5.7 ...  
## $ Units : chr [1:314] "pounds" "pounds" "pounds" "pounds" ...  
## $ Unit\_Price: num [1:314] 2.1 2.53 1.25 0.69 2.54 0.61 1.48 1.87 1.1 2.1 ...  
## $ Line\_Total: num [1:314] 10.5 12.65 6.25 3.45 12.68 ...

The majority of the variables in this data frame are categorical, according to the structure of the data frame. The data set has 331 observations and 9 variables. The unit price, quantity, and invoice are all numerical values.

First 6 observation of the data}

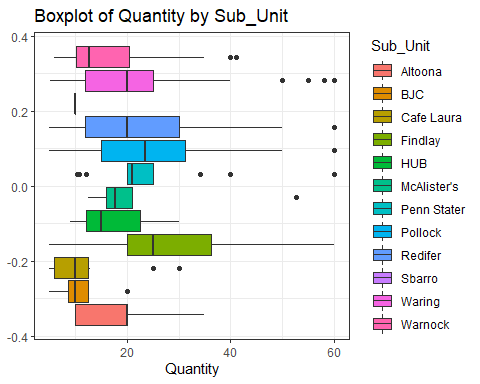
head(farmdat)

## # A tibble: 6 x 10  
## Date `Invoice#` Client Sub\_Unit Crop Variety Quantity Units Unit\_Price  
## <chr> <chr> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl>  
## 1 7.30.21 21048 HFS Pollock Tomatoes Cherry 5 pounds 2.1   
## 2 9.3.21 21114 HFS Cafe Laura Tomatoes Cherry 5 pounds 2.53  
## 3 10.5.21 21160 HFS BJC Tomatoes Slicers 5 pounds 1.25  
## 4 42593 16016 HFS Pollock Tomatoes Heirloom 5 pounds 0.69  
## 5 42607 16032 HFS Cafe Laura Tomatoes Heirloom 5 pounds 2.54  
## 6 42607 16032 HFS Cafe Laura Tomatoes Slicers 5 pounds 0.61  
## # ... with 1 more variable: Line\_Total <dbl>

## Exploratory Data Analysis

Boxplot of Quantity by Sub\_unit

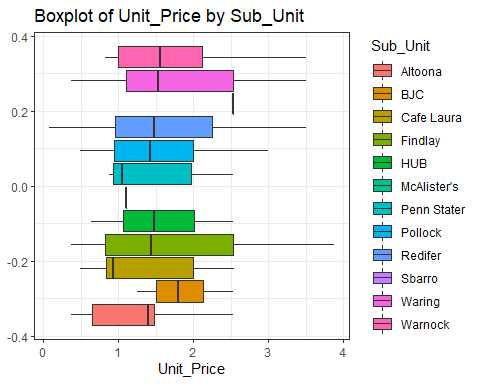
farmdat%>%  
 ggplot(aes(Quantity,fill=Sub\_Unit))+  
 geom\_boxplot()+  
 labs(title = "Boxplot of Quantity by Sub\_Unit")+  
 theme\_bw()



There are just a few outliers, according to the boxplot. The majority of the values based on the sub-units do not follow a normal distribution. The quantities for sub-units do not have an equal mean, indicating that there is a considerable variance There are just a few outliers, according to the boxplot. The majority of the values based on the sub-units do not follow a normal distribution. The quantities for sub-units do not have an equal mean, indicating that there is a considerable variance between sub-units.

Boxplot of Unit\_Price by Sub\_unit

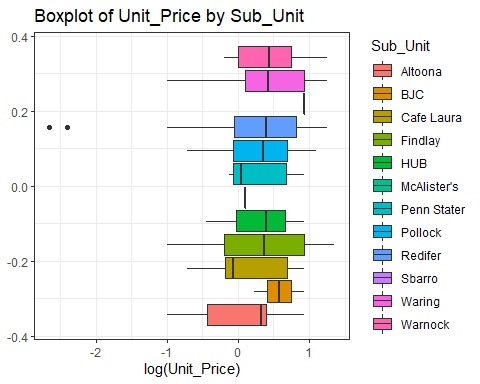
farmdat%>%  
 ggplot(aes(Unit\_Price,fill=Sub\_Unit))+  
 geom\_boxplot()+  
 labs(title = "Boxplot of Unit\_Price by Sub\_Unit")+  
 theme\_bw()



According to the boxplot, the variable Unit Price is not normally distributed, hence there is no obvious evidence of a relationship between Sub Unit and Unit Price. This variable must be transformed using the log-function.

Boxplot of log(Unit\_Price) by Sub\_unit

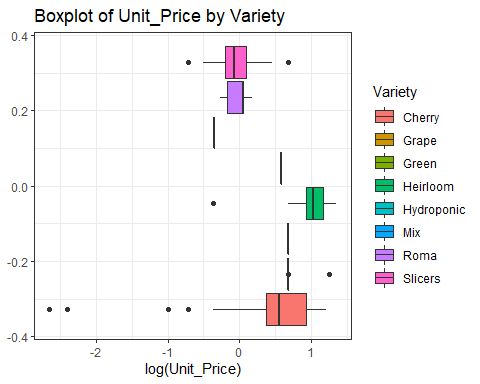
farmdat%>%  
 ggplot(aes(log(Unit\_Price),fill=Sub\_Unit))+  
 geom\_boxplot()+  
 labs(title = "Boxplot of Unit\_Price by Sub\_Unit")+  
 theme\_bw()



After we transformed the Unit Price, there is clear evidence that the mean of the Unit Price is not equal for all sub-units, implying that there is a relationship between Sub Unit and Unit Price.

Boxplot of log(Unit\_Price) by Variety

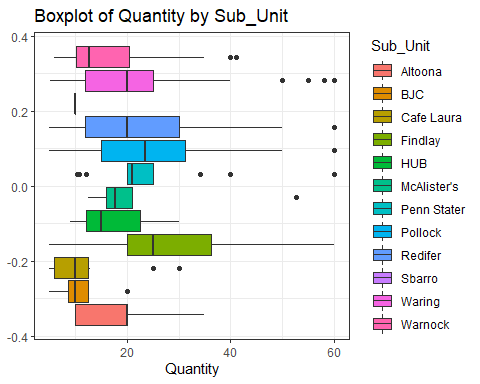
farmdat%>%  
 ggplot(aes(log(Unit\_Price),fill=Variety))+  
 geom\_boxplot()+  
 labs(title = "Boxplot of Unit\_Price by Variety")+  
 theme\_bw()



The boxplot above shows that tomato varieties have a significant difference Unit\_Price on average.

Boxplot of Quantity based on Sub\_Unit.

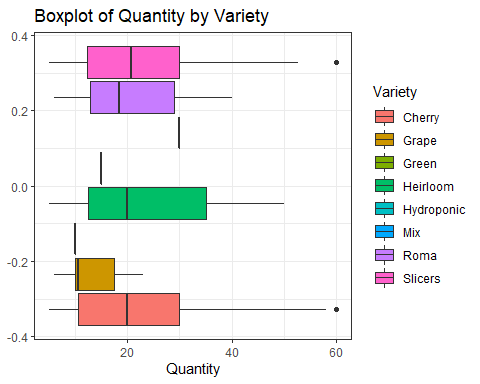
farmdat%>%  
 ggplot(aes(Quantity,fill=Sub\_Unit))+  
 geom\_boxplot()+  
 labs(title = "Boxplot of Quantity by Sub\_Unit")+  
 theme\_bw()



According to this boxplot, the tomato sub-units show a substantial difference in average quantity.

Boxplot of Quantity based on Variety.

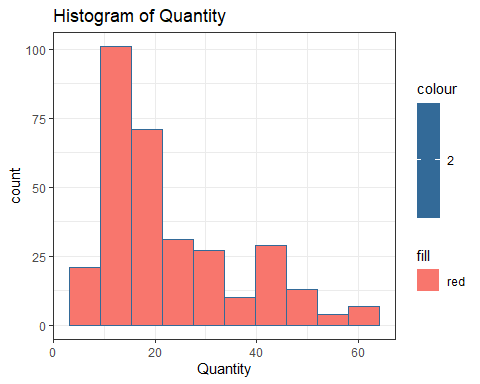
farmdat%>%  
 ggplot(aes(Quantity,fill=Variety))+  
 geom\_boxplot()+  
 labs(title = "Boxplot of Quantity by Variety")+  
 theme\_bw()



According to the boxplot, the tomato varieties show a substantial difference in average quantity.

Histogram of Quantity

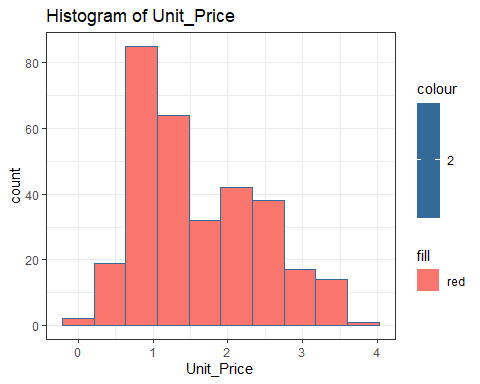
farmdat%>%  
 ggplot(aes(Quantity))+  
 geom\_histogram(bins = 10,aes(fill="red",col=2))+  
 labs(title = "Histogram of Quantity")+  
 theme\_bw()



According to the histogram, the majority of the quantity falls between 10 and 20. There is a positive skewness or a long right tail. As a result, the quantity does not follow normal distribution.

Histogram of Unit\_Price

farmdat%>%  
 ggplot(aes(Unit\_Price))+  
 geom\_histogram(bins = 10,aes(fill="red",col=2))+  
 labs(title = "Histogram of Unit\_Price")+  
 theme\_bw()



Unit Price’s histogram has a positive skewness and hence does not follow the normal distribution.

## Correlation

The Pearson’s correlation measures the direction and strength of the linear relationship among the numerical data.

Correlation heat map

Corr = cor(farmdat[sapply(farmdat, is.numeric)])  
corrplot(Corr,method = "number")

 According to the correlation findings, there is no substantial association between the numerical data. The predictor variable has a slight negative association with the target variables.The correlation results indicate that linear regression cannot fit the data.

## ANOVA

In this part, we will use ANOVA to determine whether or not there is a significant relationship between the variables of interest, as well as which variables have a significant influence on our target variables.

1. Hypothesis testing on research question

H0: There is no significant relationship between tomato quantity and tomato variety.

H1: There is a significant relationship between tomato quantity and tomato variety.

Result:

summary(aov(log(Quantity)~Variety,data=farmdat2))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Variety 7 5.86 0.8378 2.401 0.021 \*  
## Residuals 306 106.76 0.3489   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

There is no significant relationship between tomato quantity and tomato variety since the p-value of 0.0686 is bigger than the significance level of 0.05. At a 95% confidence level, we found that tomato varieties have no effect on quantity.

1. Hypothesis testing on research question

Hypothesis

H0: There is no relationship between tomato quantity and tomato sub-unit.

H1: There is a relationship between tomato quantity and tomato sub-unit.

Result:

summary(aov(log(Quantity)~Sub\_Unit,data=farmdat2))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Sub\_Unit 11 13.82 1.2563 3.84 3.42e-05 \*\*\*  
## Residuals 302 98.80 0.3272   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Because the p-value of 0.000171 is smaller than the significance level of 0.05, there is a significant relationship between tomato quantity and tomato sub-unit, implying that sub-units have a major impact on tomato quantity at a 95 percent level of significance.

TukeyHSD Post hoc

TukeyHSD(aov(log(Quantity)~Sub\_Unit,data=farmdat2))

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = log(Quantity) ~ Sub\_Unit, data = farmdat2)  
##   
## $Sub\_Unit  
## diff lwr upr p adj  
## BJC-Altoona -4.658149e-01 -1.56588848 0.63425867 0.9638641  
## Cafe Laura-Altoona -4.219402e-01 -1.16984641 0.32596603 0.7837869  
## Findlay-Altoona 4.083486e-01 -0.23672959 1.05342674 0.6337961  
## HUB-Altoona 7.750355e-03 -1.21943310 1.23493381 1.0000000  
## McAlister's-Altoona 2.701769e-01 -0.74602759 1.28638130 0.9992957  
## Penn Stater-Altoona 3.230598e-01 -0.44880290 1.09492246 0.9668060  
## Pollock-Altoona 2.914101e-01 -0.35767819 0.94049841 0.9453311  
## Redifer-Altoona 2.530660e-01 -0.34201720 0.84814928 0.9627210  
## Sbarro-Altoona -4.658149e-01 -2.43368635 1.50205653 0.9997680  
## Waring-Altoona 1.676793e-01 -0.49949159 0.83485021 0.9995866  
## Warnock-Altoona -5.465871e-02 -0.68701903 0.57770161 1.0000000  
## Cafe Laura-BJC 4.387472e-02 -1.01636390 1.10411333 1.0000000  
## Findlay-BJC 8.741635e-01 -0.11622435 1.86455131 0.1435757  
## HUB-BJC 4.735653e-01 -0.96543489 1.91256541 0.9951685  
## McAlister's-BJC 7.359918e-01 -0.52789656 1.99988008 0.7470722  
## Penn Stater-BJC 7.888747e-01 -0.28839696 1.86614634 0.4011148  
## Pollock-BJC 7.572250e-01 -0.23577943 1.75022946 0.3366265  
## Redifer-BJC 7.188809e-01 -0.23969418 1.67745608 0.3628549  
## Sbarro-BJC 4.440892e-16 -2.10648053 2.10648053 1.0000000  
## Waring-BJC 6.334942e-01 -0.37142330 1.63841174 0.6399449  
## Warnock-BJC 4.111562e-01 -0.57099541 1.39330781 0.9667580  
## Findlay-Cafe Laura 8.302888e-01 0.25577167 1.40480585 0.0001873  
## HUB-Cafe Laura 4.296905e-01 -0.76191479 1.62129588 0.9895838  
## McAlister's-Cafe Laura 6.921170e-01 -0.28082464 1.66505872 0.4484123  
## Penn Stater-Cafe Laura 7.450000e-01 0.03105596 1.45894398 0.0321562  
## Pollock-Cafe Laura 7.133503e-01 0.13433418 1.29236641 0.0035929  
## Redifer-Cafe Laura 6.750062e-01 0.15725329 1.19275917 0.0014012  
## Sbarro-Cafe Laura -4.387472e-02 -1.98975808 1.90200864 1.0000000  
## Waring-Cafe Laura 5.896195e-01 -0.00959754 1.18883653 0.0584365  
## Warnock-Cafe Laura 3.672815e-01 -0.19291815 0.92748110 0.5802546  
## HUB-Findlay -4.005982e-01 -1.53050325 0.72930682 0.9909385  
## McAlister's-Findlay -1.381717e-01 -1.03448508 0.75814164 0.9999970  
## Penn Stater-Findlay -8.528879e-02 -0.69066319 0.52008562 0.9999988  
## Pollock-Findlay -1.169385e-01 -0.55514137 0.32126445 0.9992707  
## Redifer-Findlay -1.552825e-01 -0.50859576 0.19803070 0.9528466  
## Sbarro-Findlay -8.741635e-01 -2.78288666 1.03455970 0.9373899  
## Waring-Findlay -2.406693e-01 -0.70523700 0.22389847 0.8641730  
## Warnock-Findlay -4.630073e-01 -0.87602756 -0.04998701 0.0137609  
## McAlister's-HUB 2.624265e-01 -1.11352082 1.63837382 0.9999729  
## Penn Stater-HUB 3.153094e-01 -0.89147620 1.52209506 0.9994022  
## Pollock-HUB 2.836598e-01 -0.84853951 1.41585902 0.9995989  
## Redifer-HUB 2.453157e-01 -0.85681118 1.34744255 0.9998729  
## Sbarro-HUB -4.735653e-01 -2.64912900 1.70199847 0.9998980  
## Waring-HUB 1.599290e-01 -0.98273310 1.30259101 0.9999989  
## Warnock-HUB -6.240907e-02 -1.18510187 1.06028374 1.0000000  
## Penn Stater-McAlister's 5.288293e-02 -0.93859264 1.04435850 1.0000000  
## Pollock-McAlister's 2.123326e-02 -0.87797051 0.92043703 1.0000000  
## Redifer-McAlister's -1.711081e-02 -0.87814267 0.84392105 1.0000000  
## Sbarro-McAlister's -7.359918e-01 -2.79991274 1.32792922 0.9904953  
## Waring-McAlister's -1.024975e-01 -1.01484003 0.80984494 0.9999999  
## Warnock-McAlister's -3.248356e-01 -1.21203981 0.56236868 0.9882078  
## Pollock-Penn Stater -3.164967e-02 -0.64129543 0.57799609 1.0000000  
## Redifer-Penn Stater -6.999374e-02 -0.62178754 0.48180006 0.9999996  
## Sbarro-Penn Stater -7.888747e-01 -2.74409088 1.16634150 0.9747470  
## Waring-Penn Stater -1.553805e-01 -0.78424401 0.47348306 0.9996489  
## Warnock-Penn Stater -3.777185e-01 -0.96952244 0.21408545 0.6215063  
## Redifer-Pollock -3.834407e-02 -0.39892695 0.32223881 0.9999999  
## Sbarro-Pollock -7.572250e-01 -2.66730720 1.15285717 0.9779142  
## Waring-Pollock -1.237308e-01 -0.59385096 0.34638935 0.9993597  
## Warnock-Pollock -3.460688e-01 -0.76532474 0.07318710 0.2221223  
## Sbarro-Redifer -7.188809e-01 -2.61129274 1.17353085 0.9841606  
## Waring-Redifer -8.538673e-02 -0.47758746 0.30681399 0.9998979  
## Warnock-Redifer -3.077248e-01 -0.63724558 0.02179607 0.0929415  
## Waring-Sbarro 6.334942e-01 -1.28280830 2.54979674 0.9949773  
## Warnock-Sbarro 4.111562e-01 -1.49330644 2.31561883 0.9999060  
## Warnock-Waring -2.223380e-01 -0.66907835 0.22440231 0.8930710

According to the Tukey post-hoc test, the following sub-units have a significant influence on tomato quantity: Laura - Penn Stater’s Cafe, Pollock’s Cafe, Redifer’s Cafe, and Findlay’s Cafe

1. Hypothesis testing on research question

H0: There is no significant relationship between tomato unit price and tomato variety.

H1: There is a significant relationship between tomato unit price and tomato variety.

Result:

summary(aov(log(Unit\_Price)~Variety,data=farmdat2))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Variety 7 42.17 6.024 29.78 <2e-16 \*\*\*  
## Residuals 306 61.90 0.202   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Because the p-value of 1.17e-14 is less than the significance level of 0.05, there is a significant relationship between tomato unit price and tomato variety at a 95 percent level of significance, showing that tomato varieties have a substantial influence on the price.

TukeyHSD Post hoc

TukeyHSD(aov(log(Unit\_Price)~Variety,data=farmdat2))

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = log(Unit\_Price) ~ Variety, data = farmdat2)  
##   
## $Variety  
## diff lwr upr p adj  
## Grape-Cherry 0.34268449 0.03143388 0.65393509 0.0195895  
## Green-Cherry 0.24492743 -0.73337014 1.22322500 0.9947423  
## Heirloom-Cherry 0.54571774 0.28344093 0.80799456 0.0000000  
## Hydroponic-Cherry 0.13900841 -1.23909830 1.51711513 0.9999871  
## Mix-Cherry -0.79988215 -2.17798886 0.57822456 0.6399037  
## Roma-Cherry -0.47557077 -0.86227684 -0.08886470 0.0050853  
## Slicers-Cherry -0.51365694 -0.69191963 -0.33539425 0.0000000  
## Green-Grape -0.09775705 -1.10970308 0.91418897 0.9999904  
## Heirloom-Grape 0.20303326 -0.16541993 0.57148645 0.6990888  
## Hydroponic-Grape -0.20367607 -1.60586962 1.19851747 0.9998459  
## Mix-Grape -1.14256664 -2.54476018 0.25962691 0.2049634  
## Roma-Grape -0.81825526 -1.28356183 -0.35294868 0.0000044  
## Slicers-Grape -0.85634143 -1.17058052 -0.54210233 0.0000000  
## Heirloom-Green 0.30079031 -0.69718110 1.29876173 0.9839962  
## Hydroponic-Green -0.10591902 -1.78709002 1.57525198 0.9999995  
## Mix-Green -1.04480958 -2.72598058 0.63636142 0.5541971  
## Roma-Green -0.72049820 -1.75813947 0.31714307 0.4048822  
## Slicers-Green -0.75858437 -1.73783685 0.22066810 0.2629049  
## Hydroponic-Heirloom -0.40670933 -1.79885118 0.98543252 0.9866644  
## Mix-Heirloom -1.34559990 -2.73774174 0.04654195 0.0665022  
## Roma-Heirloom -1.02128852 -1.45536500 -0.58721203 0.0000000  
## Slicers-Heirloom -1.05937469 -1.32519116 -0.79355821 0.0000000  
## Mix-Hydroponic -0.93889056 -2.88013962 1.00235850 0.8196839  
## Roma-Hydroponic -0.61457918 -2.03542801 0.80626964 0.8907274  
## Slicers-Hydroponic -0.65266536 -2.03145010 0.72611939 0.8355726  
## Roma-Mix 0.32431138 -1.09653744 1.74516020 0.9970458  
## Slicers-Mix 0.28622521 -1.09255954 1.66500996 0.9983845  
## Slicers-Roma -0.03808617 -0.42720165 0.35102930 0.9999895

According to the Tukey post-hoc test, the following varieties have a significant influence on unit price of the tomatoes: Slicers-Cherry, Slicers-Grape, Roma-Heirloom, and Slicers-Heirloom.

1. Hypothesis testing on research question

H0: There is no significant relationship between tomato unit price and tomato sub-unit.

H1: There is a significant relationship between tomato unit price and tomato sub-unit.

Result:

summary(aov(log(Unit\_Price)~Sub\_Unit,data=farmdat2))

## Df Sum Sq Mean Sq F value Pr(>F)  
## Sub\_Unit 11 3.03 0.2755 0.824 0.617  
## Residuals 302 101.04 0.3346

There is no significant relationship between tomato unit price and tomato sub-unit because p-value of 0.383 is greater than the significance level of 0.05. We conclude that the sub-units of the tomatoes do not influence the price of the tomatoes at a 95% level of confidence.

## Conclusion

To summarise all that has been stated, the varieties that impact tomato prices include Slicers-Cherry, Slicers-Grape, Roma-Heirloom, and Slicers-Heirloom. As a result, while growing tomatoes, we should examine the variety because it has a considerable impact on the price. The tomato’s sub-units have a significant impact on its quantity. The following sub-units have a substantial influence on the quantity of the tomato: Laura - Penn Stater’s Cafe, Pollock’s Cafe, Redifer’s Cafe, and Findlay’s Cafe. However, the sub-units have no impact on the price of the tomatoes, and the variety of the tomato has no affect on the quantity of the tomato.