Avocado Dataset Markdown - Statistical Analysis

06 06 2020

#### 1. Describe the Data Set statistically. What are the variables? Define the type of each variable (Nominal, ordinal, or interval). Descriptive Statistics for the total number of avocado sold in Atlanta region (Variance, Standard deviation, Range, Percentiles, Quartiles, Measure of Skew. Scatterplot, Box Plot).

We define a function to calculate percentiles first.

printf <- function(...) {print(sprintf(...))}

fractile <- function(data\_set, n) {  
 name <- "Fractile"  
 if (n == 3) name <- "Tertile"  
 if (n == 4) name <- "Quartile"  
 if (n == 5) name <- "Quintile"  
 if (n == 6) name <- "Sextile"  
 if (n == 7) name <- "Septile"  
 if (n == 8) name <- "Octile"  
 if (n == 10) name <- "Decile"  
 if (n == 12) name <- "Duo-decile"  
 if (n == 16) name <- "Hexadecile"  
 if (n == 20) name <- "Ventile"  
 if (n == 100) name <- "Percentile"  
 for (i in 2:n-1) {  
 printf("%s %s = %s", name, i, quantile(data\_set, i \* (1/n)))  
 }  
   
}

We define another function to calculate variance, standard deviation, range and skewness measures.

basic\_descriptive <- function(data\_set) {  
 printf("Variance = %s", var(data\_set))  
 printf("Standard Deviation = %s", sd(data\_set))  
 range <- range(data\_set)  
 printf("Range = %s - %s", range[1], range[2])  
 fractile(data\_set, 4)  
 fractile(data\_set, 100)  
 printf("Skewness = %s", skewness(data\_set))  
}

Now we should read our data and retrieve information about Atlanta region.

We want to define the type of each variable in our data set:

## [1] "Variables :"

## [1] "Date - Interval"

## [1] "AveragePrice - Interval"

## [1] "Total Volume - Interval"

## [1] "4046 - Interval"

## [1] "4225 - Interval"

## [1] "4770 - Interval"

## [1] "Total Bags - Interval"

## [1] "Small Bags - Interval"

## [1] "Large Bags - Interval"

## [1] "XLarge Bags - Interval"

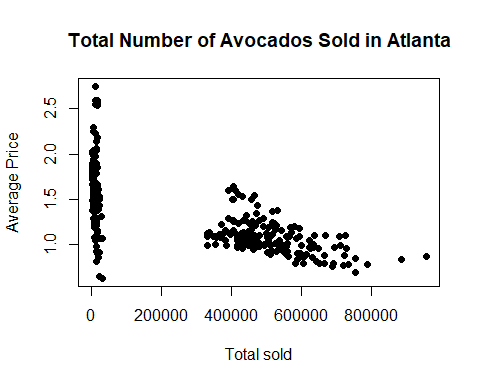
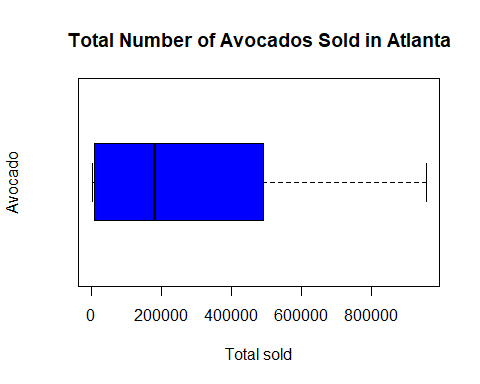
## [1] "type - Nominal"

## [1] "year - Interval"

## [1] "region - Nominal"

Now we can call our function to calculate the measures we want:

## [1] "Variance = 69225328593.3805"  
## [1] "Standard Deviation = 263107.066787231"  
## [1] "Range = 3047.38 - 957792.07"  
## [1] "Quartile 1 = 10989.8975"  
## [1] "Quartile 2 = 182673.44"  
## [1] "Quartile 3 = 490708.98"  
## [1] "Percentile 1 = 3781.8819"  
## [1] "Percentile 2 = 3975.527"  
## [1] "Percentile 3 = 4305.6964"  
## [1] "Percentile 4 = 4637.568"  
## [1] "Percentile 5 = 5069.6245"  
## [1] "Percentile 6 = 5210.6254"  
## [1] "Percentile 7 = 5332.5508"  
## [1] "Percentile 8 = 5444.1116"  
## [1] "Percentile 9 = 5503.8432"  
## [1] "Percentile 10 = 5644.792"  
## [1] "Percentile 11 = 5747.6634"  
## [1] "Percentile 12 = 6055.1916"  
## [1] "Percentile 13 = 6451.6211"  
## [1] "Percentile 14 = 6638.2494"  
## [1] "Percentile 15 = 6863.659"  
## [1] "Percentile 16 = 7267.922"  
## [1] "Percentile 17 = 7449.9799"  
## [1] "Percentile 18 = 7657.5366"  
## [1] "Percentile 19 = 7889.2354"  
## [1] "Percentile 20 = 8236.46"  
## [1] "Percentile 21 = 8908.8493"  
## [1] "Percentile 22 = 9454.8626"  
## [1] "Percentile 23 = 9808.5143"  
## [1] "Percentile 24 = 10387.498"  
## [1] "Percentile 25 = 10989.8975"  
## [1] "Percentile 26 = 11147.5846"  
## [1] "Percentile 27 = 11547.0301"  
## [1] "Percentile 28 = 11783.3892"  
## [1] "Percentile 29 = 12057.6301"  
## [1] "Percentile 30 = 12472.279"  
## [1] "Percentile 31 = 13145.3579"  
## [1] "Percentile 32 = 13298.5564"  
## [1] "Percentile 33 = 13962.0209"  
## [1] "Percentile 34 = 14377.761"  
## [1] "Percentile 35 = 14530.1255"  
## [1] "Percentile 36 = 15172.278"  
## [1] "Percentile 37 = 15490.5059"  
## [1] "Percentile 38 = 16034.435"  
## [1] "Percentile 39 = 16217.1119"  
## [1] "Percentile 40 = 16554.862"  
## [1] "Percentile 41 = 16966.9122"  
## [1] "Percentile 42 = 17302.399"  
## [1] "Percentile 43 = 17657.503"  
## [1] "Percentile 44 = 18502.5252"  
## [1] "Percentile 45 = 18843.973"  
## [1] "Percentile 46 = 19518.3636"  
## [1] "Percentile 47 = 21961.0515"  
## [1] "Percentile 48 = 23395.26"  
## [1] "Percentile 49 = 25237.234"  
## [1] "Percentile 50 = 182673.44"  
## [1] "Percentile 51 = 339104.1996"  
## [1] "Percentile 52 = 357838.7224"  
## [1] "Percentile 53 = 371433.2854"  
## [1] "Percentile 54 = 383136.1546"  
## [1] "Percentile 55 = 395127.0795"  
## [1] "Percentile 56 = 405305.3972"  
## [1] "Percentile 57 = 407299.9191"  
## [1] "Percentile 58 = 411839.0176"  
## [1] "Percentile 59 = 417680.6207"  
## [1] "Percentile 60 = 420908.082"  
## [1] "Percentile 61 = 422967.843"  
## [1] "Percentile 62 = 431288.0268"  
## [1] "Percentile 63 = 433877.3895"  
## [1] "Percentile 64 = 435582.0512"  
## [1] "Percentile 65 = 440217.9275"  
## [1] "Percentile 66 = 446562.7986"  
## [1] "Percentile 67 = 449318.2802"  
## [1] "Percentile 68 = 452372.5112"  
## [1] "Percentile 69 = 458353.7066"  
## [1] "Percentile 70 = 460848.207"  
## [1] "Percentile 71 = 465472.9366"  
## [1] "Percentile 72 = 471028.564"  
## [1] "Percentile 73 = 475816.0851"  
## [1] "Percentile 74 = 480608.1618"  
## [1] "Percentile 75 = 490708.98"  
## [1] "Percentile 76 = 496800.968"  
## [1] "Percentile 77 = 504350.9472"  
## [1] "Percentile 78 = 510941.4608"  
## [1] "Percentile 79 = 516825.3779"  
## [1] "Percentile 80 = 521869.23"  
## [1] "Percentile 81 = 528279.9721"  
## [1] "Percentile 82 = 535274.0946"  
## [1] "Percentile 83 = 542187.9532"  
## [1] "Percentile 84 = 549288.6684"  
## [1] "Percentile 85 = 556254.6645"  
## [1] "Percentile 86 = 559260.3448"  
## [1] "Percentile 87 = 563883.879"  
## [1] "Percentile 88 = 578007.4292"  
## [1] "Percentile 89 = 588487.7106"  
## [1] "Percentile 90 = 596205.628"  
## [1] "Percentile 91 = 607881.5516"  
## [1] "Percentile 92 = 624767.4348"  
## [1] "Percentile 93 = 635261.9907"  
## [1] "Percentile 94 = 648108.9696"  
## [1] "Percentile 95 = 667455.307"  
## [1] "Percentile 96 = 692084.03"  
## [1] "Percentile 97 = 713489.3058"  
## [1] "Percentile 98 = 727694.6404"  
## [1] "Percentile 99 = 755410.8328"  
## [1] "Skewness = 0.292530835521787"

We can plot scatter and boxplot for total number of avocado sold in Atlanta region:  

#### 2. Hypothesis testing for mean (Hint: You need determine test value for the total amount of avocado sold in Atlanta region.)

#### 2.1. State the hypothesis.

H0(Null hypothesis): mean of total amount of avocado sold (in volume) in Atlanta region is equal to 200,000 (test value)

H1(Alternative hypothesis): mean of total amount of avocado sold (in volume) in Atlanta region is greater than 200,000 (test value). We define our test value:

test\_mu <- 200000

#### 2.2 Compute the test statistic.

t.test(atlanta\_total, mu = test\_mu, alternative = "greater", var.equal = T, conf.level = 0.95)

##   
## One Sample t-test  
##   
## data: atlanta\_total  
## t = 4.3424, df = 337, p-value = 0.000009332  
## alternative hypothesis: true mean is greater than 200000  
## 95 percent confidence interval:  
## 238540.7 Inf  
## sample estimates:  
## mean of x   
## 262145.3

Here’s our t-statistic value:

t\_value\_for\_atlanta <- (mean(atlanta\_total) - test\_mu)/ sqrt(var(atlanta\_total)/nrow(atlanta))

## [1] 4.342445

#### 2.3. Interpret the results.

t\_value = 4.3424 with df = 337 and p-value = 0.000009 which means 200000 (conf. level = 0.95 then alpha = 0.05)

Reject null hypothesis in favor of alternative hypothesis.

Mean of total amount of avocado sold in Atlanta region is greater than 200000.

#### 3. Hypothesis testing for standard deviation (Hint: You need determine test value for the total amount of avocado sold in Atlanta region).

##### 3.1. State the hypothesis.

H0(Null hypothesis): standard deviation of total amount of avocado sold (in volume) in Atlanta region is equal to 200000(test value)

H1(alternative hypothesis): standard deviation of total amount of avocado sold (in volume)in Atlanta region is greater than 200000(test value)

We define a test value:

test\_sigma <- 200000

#### 3.2.Compute the test statistic.

We can apply sigma test from Teaching Demos library:

sigma.test(atlanta\_total, sigma = test\_sigma, sigmasq = test\_sigma^2, alternative = "greater", conf.level = 0.95)

##   
## One sample Chi-squared test for variance  
##   
## data: atlanta\_total  
## X-squared = 583.22, df = 337, p-value = 0.000000000000001887  
## alternative hypothesis: true variance is greater than 40000000000  
## 95 percent confidence interval:  
## 61261460846 Inf  
## sample estimates:  
## var of atlanta\_total   
## 69225328593

We also define a Chi-squared value:

chisquare\_value <- (nrow(atlanta) - 1) \* (var(atlanta\_total)) / ((test\_sigma)^2)

## [1] 583.2234

#### 3.3. Interpret the results.

chisquare\_value =583.22 with df = 337 and p-value = 0.000000000000001887 (conf. level = 0.95 then alpha = 0.05)

Reject null hypothesis in favor of alternative hypothesis.

Standard deviation of total amount of avocado sold in Atlanta region is greater than 200000 (it is 263107.1)

#### 4. Two sample confidence intervals for the means of total amount of avocado sold in Boston region and Atlanta region.

Let us obtain the variances and means for Atlanta and Boston regions.

## [1] "Mean for Boston region: 287792.854526627"

## [1] "Mean for Atlanta region: 262145.32204142"

## [1] "Variance for Boston region: 80890679511.5635"

## [1] "Variance for Boston region: 69225328593.3805"

We define v, t-statistic, standard error and estimate variables to calculate confidence interval.

v <- ((var\_one / n\_one + var\_two / n\_two)^2) / (((var\_one / n\_one)^2) / (n\_one - 1) + ((var\_two / n\_two)^2) / (n\_two - 1))  
  
t <- - qt((1 - 0.95) / 2 , df=v)  
  
std\_err <- sqrt((var\_one / n\_one) + (var\_two / n\_two))  
  
estimate <- mean\_one - mean\_two

Now we can calculate confidence interval:

## [1] "Confidence interval:"

## [1] "25647.5324852071 - 1.96351122600851 \* 21074.3968239643 < mu\_boston - mu\_atlanta < 25647.5324852071 + 1.96351122600851 \* 21074.3968239643"

## [1] "25647.5324852071 - 41379.8147452121 < mu\_boston - mu\_atlanta < 25647.5324852071 + 41379.8147452121"

#### 5. Two sample hypothesis testing for means of total amount of avocado sold in Boston region and Atlanta region.

We will apply t-test. Assume that each groups are APPROXIMATELY NORMALLY DISTRIBUTED WITH EQUAL VARIANCES.

##### 5.1. State the hypothesis.

H0: There is NO significant difference between total sales (in volume) for Boston and Atlanta regions.

H1: There is a significant difference between total sales (in volume) for Boston and Atlanta regions.

Notice that this test is two-tailed.

##### 5.2. Compute the test statistic.

t.test(atlanta\_total, boston\_total, paired = F, alternative = "two.sided", var.equal = T, conf.level = 0.95)

##   
## Two Sample t-test  
##   
## data: atlanta\_total and boston\_total  
## t = -1.217, df = 674, p-value = 0.224  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -67026.90 15731.83  
## sample estimates:  
## mean of x mean of y   
## 262145.3 287792.9

t\_value\_for\_two\_sample <- (mean\_two - mean\_one) / sqrt(var\_two / nrow(atlanta) + var\_one / nrow(boston))

Our t-value is:

## [1] -1.217

We could also obtain this value from t.test function output.

##### 5.3. Interpret the results.

t\_value = -1.217 with df = 674 and p = 0.224. p > 0.025 (two-tailed test, conf. level = 0.95, then alpha = 0.05)

Reject null hypothesis in favor of alternative hypothesis.

There is a significant difference between means for Boston and Atlanta regions.

#### 6. Regression Analysis: ‘4046’ and ‘Total bags’ in Boston region.

##### 6.1. Obtain the equation of the regression line.

model <- lm(Total.Bags ~ X4046, data = atlanta )

##   
## Call:  
## lm(formula = Total.Bags ~ X4046, data = atlanta)  
##   
## Coefficients:  
## (Intercept) X4046   
## 20554.7389 0.4376

Our formula should be stated as:

#### Total Bags = 20554.7389 + 0.4376 (Avocados with PLU sticker 4046)

##### 6.2. Graph the line on a scatter diagram.

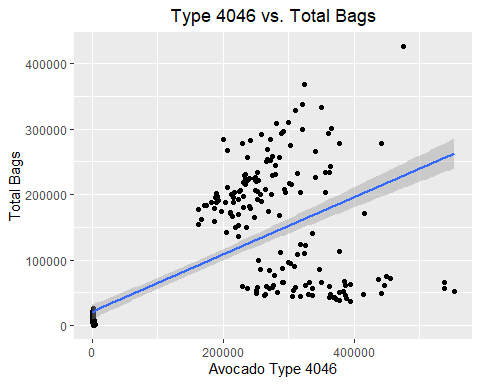
Let us first call ggplot2 library for a better visiualization.

library(ggplot2)

Now we can plot the data using the following code:

ggplot(atlanta, aes(X4046, Total.Bags)) + labs(title = "Type 4046 vs. Total Bags", x= "Avocado Type 4046", y= "Total Bags") + theme(plot.title = element\_text(hjust = 0.5)) + geom\_point() + stat\_smooth(method = lm)

## `geom\_smooth()` using formula 'y ~ x'



##### 6.3. Interpret R-squared value

We call the summary function:

summary(model)

##   
## Call:  
## lm(formula = Total.Bags ~ X4046, data = atlanta)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -210952 -20903 -14433 56724 206215   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20554.73888 5604.40216 3.668 0.000285 \*\*\*  
## X4046 0.43762 0.02638 16.591 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 74800 on 336 degrees of freedom  
## Multiple R-squared: 0.4503, Adjusted R-squared: 0.4487   
## F-statistic: 275.3 on 1 and 336 DF, p-value: < 0.00000000000000022

R-squared value is used to interpret how well our model fits the data.

R-squared takes a value between 0 and 1. Here the summary gives us two different values for R-squared: multiple and adjusted.

Multiple R-squared: 0.4503, Adjusted R-squared: 0.4487

We are interested in adjusted R-squared value. Multiple R-squared includes a penalty for multiple linear regression models.

0.4487 value means that our model is not able to explain the variability (only partially). A value closer to 1 would be preferred.

We can observe that from Residual standard error value as well: 74,800. RSE value is relatively high, which means that the linear model fits our data poorly. The values deviate from our regression line by 74,800 units.

Also the percentage error for RSE is calculated by

sigma(model) \* 100 / mean(atlanta$Total.Bags)

## [1] 88.52812

##### 6.4. Hypothesis testing for the slope and intercept.

H0: The coefficient is equal to zero (no correlation between Total Bags and Avocado type 4046.)

H1: The coefficient is NOT equal to zero. (There is a correlation between Total Bags and Avocado type 4046.)

Let us define variables to obtain our regression t-value.

regression <- summary(model)   
coefficients <- regression$coefficients  
beta\_estimate <-coefficients["X4046", "Estimate"]  
stderror <- coefficients["X4046", "Std. Error"]  
regression\_t <- beta\_estimate / stderror  
regression\_p <- 2\*pt(-abs(regression\_t), df=nrow(atlanta)-ncol(atlanta))

## [1] 16.59131

Again, using t-value for our hypothesis: 16.59131.

It simply tells us that the coefficient cannot be equal to zero.

Checking our summary model, we can see that P-value is extremely small and falls into critical region ( <alpha = 0.05) We can reject H0.

#### 7. Analysis of Variance (Is there a difference between the mean of total amount of avocado sold: in the following regions: Atlanta, Boston and Detroit.

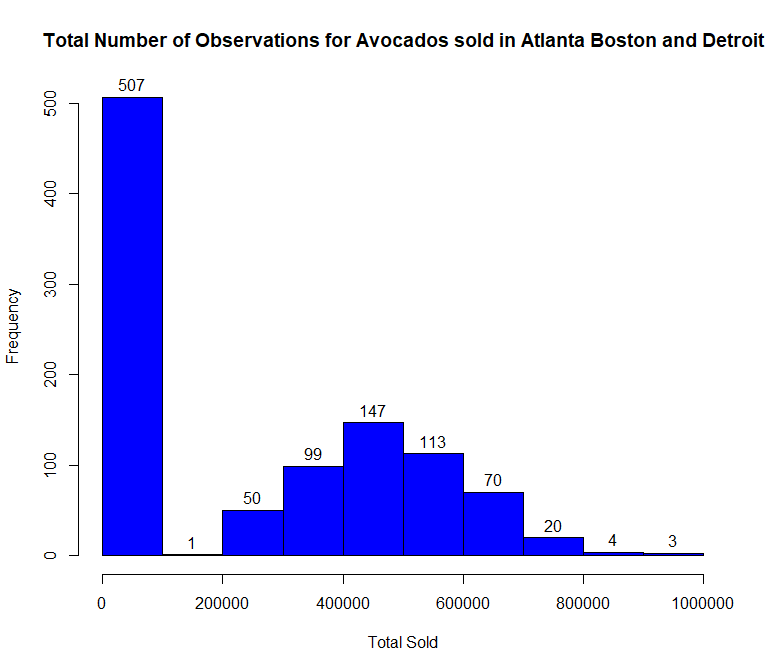
We use “car” library for Levene Test.

## Loading required package: carData

##### 7.1. Perform the diagnostic tests for the assumptions on ANOVA and test the following assumptions:

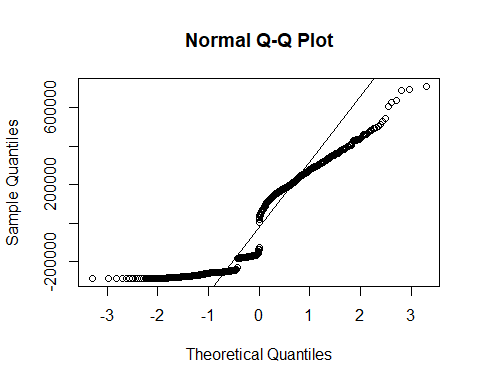
##### (i) Normality of residuals,

hist(new\_dt$Total.Volume, main = paste("Total Number of Observations for Avocados sold in Atlanta Boston and Detroit"), xlab = "Total Sold", labels = T, col = "blue")



Let’s do one-way ANOVA test:

anova\_one\_way <- aov(Total.Volume ~ region, data = new\_dt)  
qqnorm(anova\_one\_way$residuals)  
qqline(anova\_one\_way$residuals)



Observations are incompatible with theoretical values so there is not a normality.

##### (ii) The errors have a mean of 0,

##### (iii)The errors are independent,

Independency can only be achieved if you have set your experiment up correctly.

There is no way to test the independency;

##### (iv) The errors have equal variance among each treatment levels.

leveneTest(Total.Volume~region, data = new\_dt)

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to  
## factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 2 80.306 < 0.00000000000000022 \*\*\*  
## 1011   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p value = 0.00000000000000022 and p << 0.05 so there is not a homogeneity of variance.

We can illustrate this difference with boxplot.

One-way ANOVA is not a valid test to determine any significant difference between group means. Conclusion of ANOVA is not certain.

Because there are not homogeneity of variance and normality. ##### 7.2. State the hypotheses and identify the claim. H0: There is not a difference between the mean of total amount of avocado sold: in the following regions: Atlanta, Boston and Detroit.

H1: At least two of the means are not equal.

##### 7.3. Compute the test statistic.

anova\_one\_way <- aov(Total.Volume~region, data = new\_dt)  
summary(anova\_one\_way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## region 2 1829631236086 914815618043 14.81 0.000000457 \*\*\*  
## Residuals 1011 62443296826426 61763893993   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### 7.4. Make the decision and interpret the result.

p\_value = 0.000000457 and p << 0.05. Region is a significantly important factor for sell.

Reject null hypothesis in favor of alternative hypothesis.

At least two of the means are not equal.

##### 7.5. If the ANOVA is significant, conduct the Post-hoc tests.

TukeyHSD(anova\_one\_way)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Total.Volume ~ region, data = new\_dt)  
##   
## $region  
## diff lwr upr p adj  
## Boston-Atlanta 25647.53 -19223.64 70518.70 0.3724433  
## Detroit-Atlanta -74505.03 -119376.20 -29633.86 0.0003052  
## Detroit-Boston -100152.56 -145023.73 -55281.39 0.0000006

For Boston-Atlanta, p\_value is 0.3724433 and more than 0.05 so there is not a regional sales difference between Bostan and Atlanta.

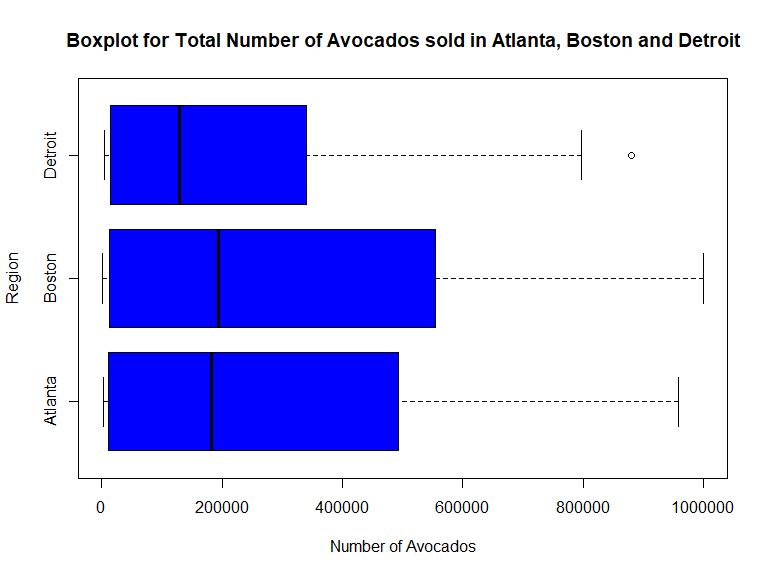
But for Detroit-Boston and Detroit-Atlanta which p\_values is smaller than 0.05,regional sales differences are important.

p\_value=0.0003052 and p < 0.05 for Detroit-Atlanta.

p\_value=0.0000006 and p < 0.05 for Detroit-Boston.

##### 7.6. Plot the boxplot and interpret it.

boxplot(Total.Volume~region,data = new\_dt, main = "Boxplot for Total Number of Avocados sold in Atlanta, Boston and Detroit", xlab = "Number of Avocados",ylab = "Region", col= "blue", horizontal = T)



Boxplot is always a good idea for understanding which groups have the significantly different variance.

Atlanta and Boston have a larger variance than Detroit’s variance.

Atlanta and Boston have close variances.