Breckwoldt_Project2

Will Breckwoldt

10/4/2021

ALY6000 Introduction to Analytics Northeastern University William Breckwoldt Date: ‘r format(Sys.time(), "%d %B, %Y)’

Project Report

```
#import libraries
library(readxl)
library(readr)
library(tidyverse)
library(dplyr)
library(knitr)

# Data sets (add importing codes here)
M2Data = read_excel("DataSets/M2Project_V2.xlsx")
```

Introduction

As a data analyst, one must create understandable data visualizations and clearly summarize the data for executives. For this assignment, I will create an executive summary based on the dataset provided by the professor and thereby demonstrate my ability to process data, present the data visually, and calculate basic; all with explanatory analysis.

Analysis section

Task 1 Present the first 3 and last 5 records from the dataset.

```
# Codes to solve task
a = head(M2Data, 3)
b = tail(M2Data, 5)

c = rbind(a,b)
knitr::kable(head(c), "simple")
```

Region	Market	Company_Segment	Product_Category	Product_SubCategory	Price	Quantity	
Central US	USCA	Consumer	Technology	Phones	221.98	2	
Oceania	Asia Pacific	Corporate	Furniture	Chairs	3709.40	9	33
Oceania	Asia Pacific	Consumer	Technology	Phones	5175.17	9	46
Eastern Asia	Asia Pacific	Consumer	Furniture	Tables	2614.69	7	18
Western US	USCA	Corporate	Office Supplies	Appliances	69.48	1	
Oceania	Asia Pacific	Consumer	Technology	Copiers	636.78	2	1

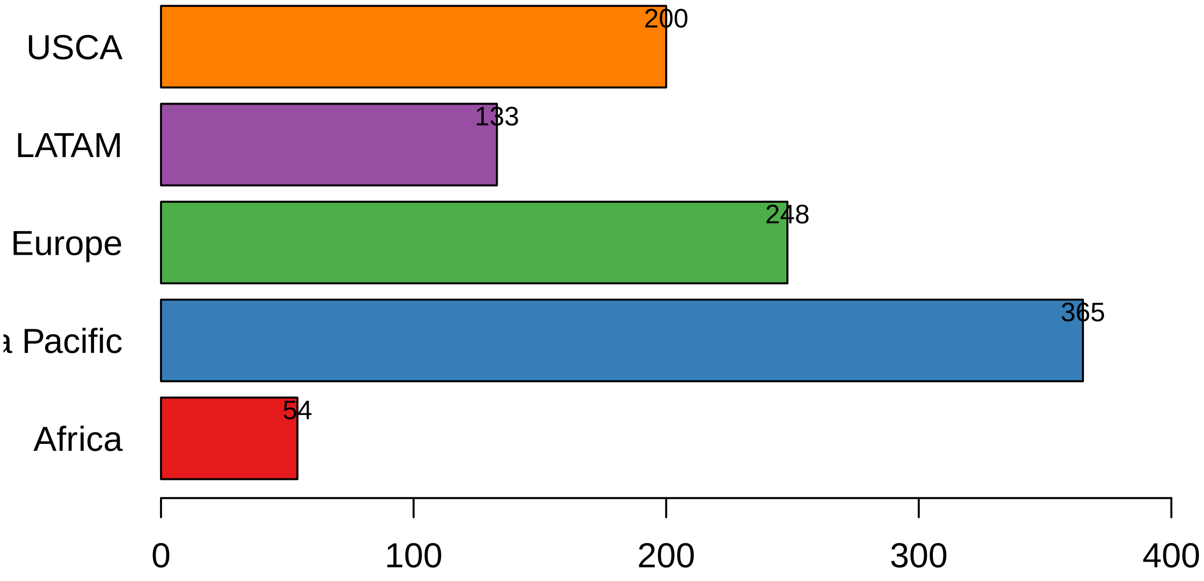
Task 2 Present a table with all categories of Market and their frequencies.

```r
table(M2Data$Market)
```

```
##
##       Africa Asia Pacific       Europe        LATAM         USCA
##           54          365          248          133          200
```

Task 3 Present the results of task 2 using a horizontal bar graph.

```r
d = table(M2Data$Market)

#Import RColorBrewer
library(RColorBrewer)

#Add values to each bar.

#Increase limits to ensure that the tallest bar is observed with its values.

e = barplot(d,
     horiz = TRUE,
     xlim = c(0,400),
     col = brewer.pal(9,"Set1"),
     las = 1,
     cex.axis = 1.1,
     cex.names = 1.1
)
text(d,
     e,
     d,
     cex=0.8,
     pos = 3)
```



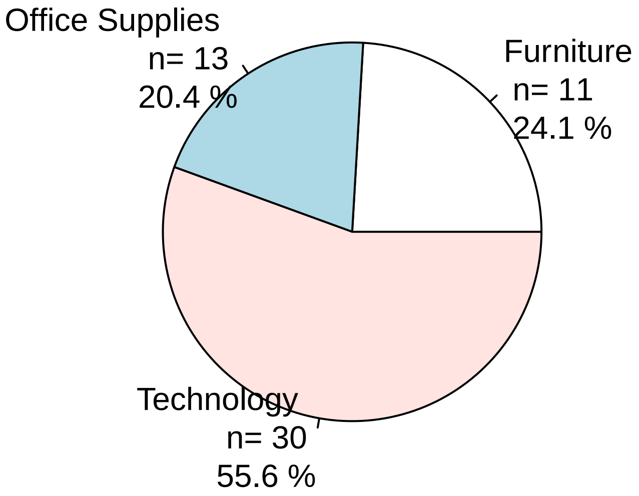
Task 4 Using code filter(), from library(dplyr), filter all observations for the African Market.

```
t4Africa = dplyr::filter(M2Data, Market=="Africa")
t4Africa
```

```
## # A tibble: 54 x 10
##    Region       Market Company_Segment Product_Category Product_SubCatego~ Price
##    <chr>        <chr>  <chr>           <chr>            <chr>              <dbl>
##  1 Western Afr~ Africa Consumer        Technology       Copiers            2833.
##  2 Eastern Afr~ Africa Consumer        Office Supplies  Appliances         3410.
##  3 Central Afr~ Africa Corporate       Technology       Phones             3817.
##  4 Eastern Afr~ Africa Corporate       Technology       Phones             2582.
##  5 Central Afr~ Africa Consumer        Furniture        Chairs             3809.
##  6 North Africa Africa Corporate       Technology       Copiers            5301.
##  7 North Africa Africa Home Office     Office Supplies  Appliances         2266.
##  8 Central Afr~ Africa Consumer        Technology       Phones             3856.
##  9 Eastern Afr~ Africa Consumer        Technology       Phones             3834
## 10 Southern Af~ Africa Home Office     Office Supplies  Storage            2785.
## # ... with 44 more rows, and 4 more variables: Quantity <dbl>, Sales <dbl>,
## #   Profits <dbl>, ShippingCost <dbl>
```

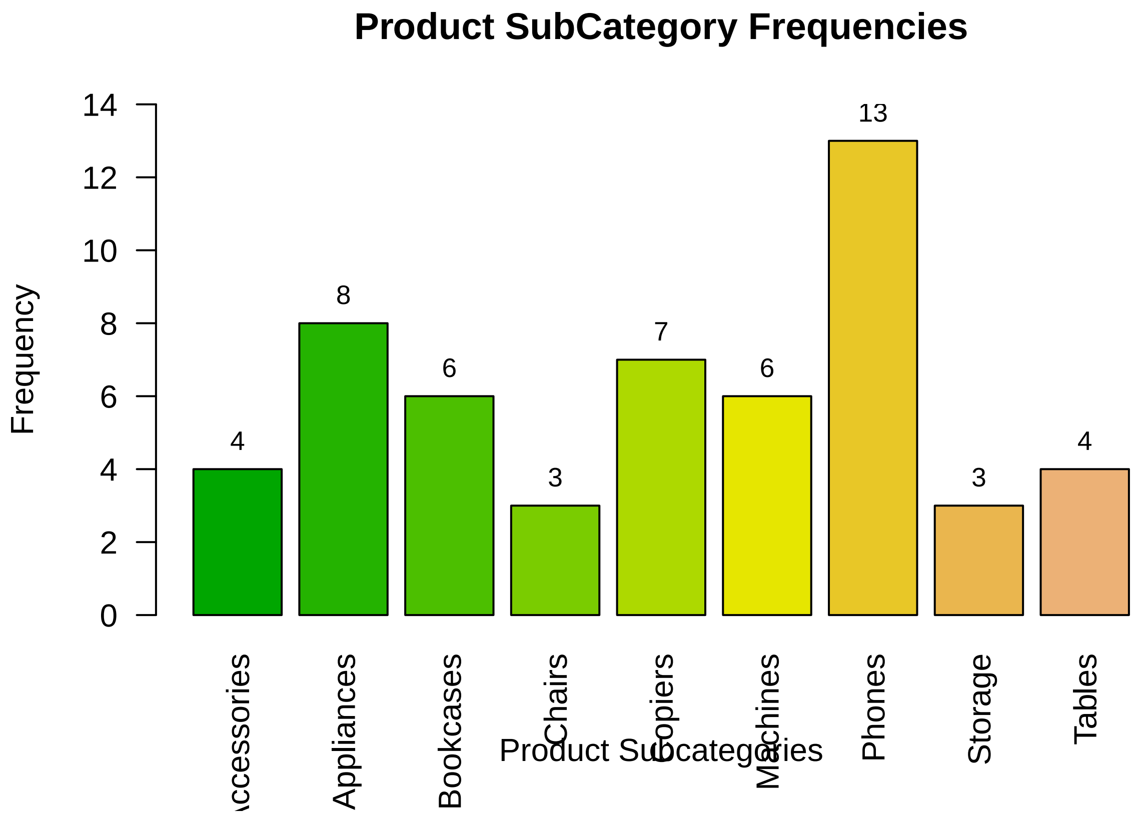
Use the name of the object (t4Africa) to create a pie chart displaying product category and their frequencies. Remember that you need to create a table with the data, otherwise the code pie() will not create the right graph.

```r
#first table
f = table(t4Africa$Product_Category)

# Table calculating percentages
g = f/nrow(t4Africa)*100

#Define labels
pieLabels = paste(unique(sort(t4Africa$Product_Category)),
                  "",
                  "\n",
                  "n=",
                  sort(f),
                  "\n",
                  round(g,1),
                  "%")

#Basic Plot of Pie Chart
pie1 = pie(f,
           labels= pieLabels,
           radius = 0.8)
```



Task 5 Improve that graph with labels, title, and colors. Make sure that all names are displayed. You can create a horizontal or vertical bar plot.

```
t5bar = barplot(table(t4Africa$Product_SubCategory),
                main = "Product SubCategory Frequencies",
                xlab = "Product Subcategories",
                ylab = "Frequency",
                col = terrain.colors(12),
                ylim = c(0, 14),
                las = 2
                )

text(y= table(t4Africa$Product_SubCategory),
     t5bar,
     table(t4Africa$Product_SubCategory),
     cex=0.8,
     pos=3
     )
```



Task 6 What are the mean sales per subcategory in the African Market?

```
meanSales = tapply(X = t4Africa$Sales, INDEX = t4Africa$Product_SubCategory, FUN = mean, na.rm= T)
knitr::kable(meanSales)
```

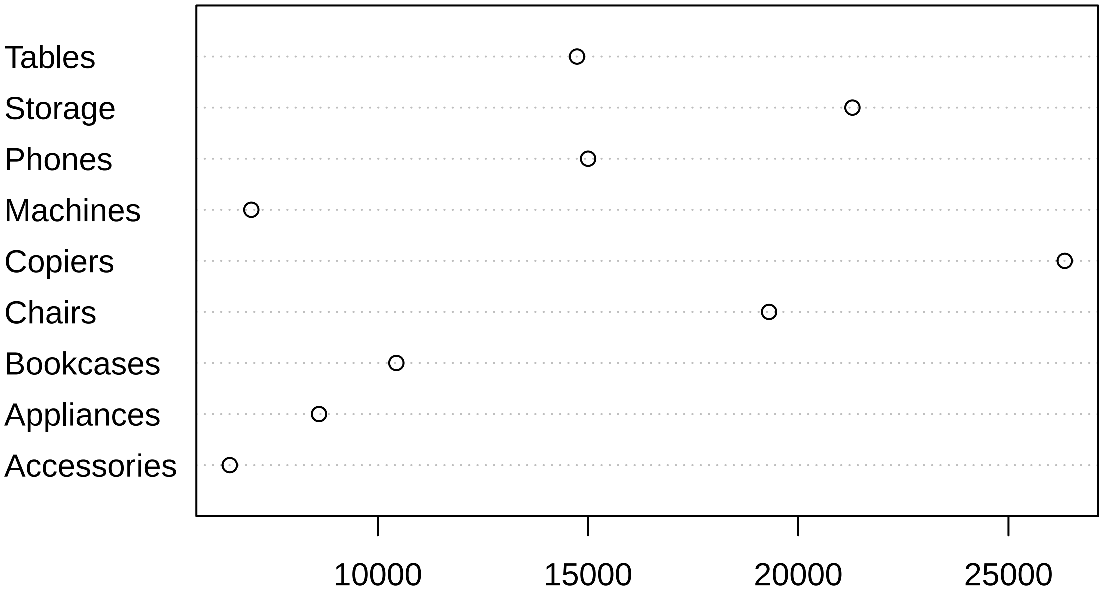
	x
Accessories	6478.980
Appliances	8601.975

	x
Bookcases	10441.840
Chairs	19306.760
Copiers	26338.286
Machines	6991.880
Phones	15001.698
Storage	21289.200
Tables	14738.970

Use a dot plot to display the information.

```r
dotchart(meanSales)
```

```
## Warning in dotchart(meanSales): 'x' is neither a vector nor a matrix: using
## as.numeric(x)
```



Task 7 What are the total sales per Region in the African Market?

```r
totSales = tapply(X = t4Africa$Sales, INDEX = t4Africa$Region, FUN = sum, na.rm= T)
knitr::kable(totSales)
```

	x
Central Africa	205523.8
Eastern Africa	96575.4
North Africa	178792.3
Southern Africa	161749.4
Western Africa	116827.0

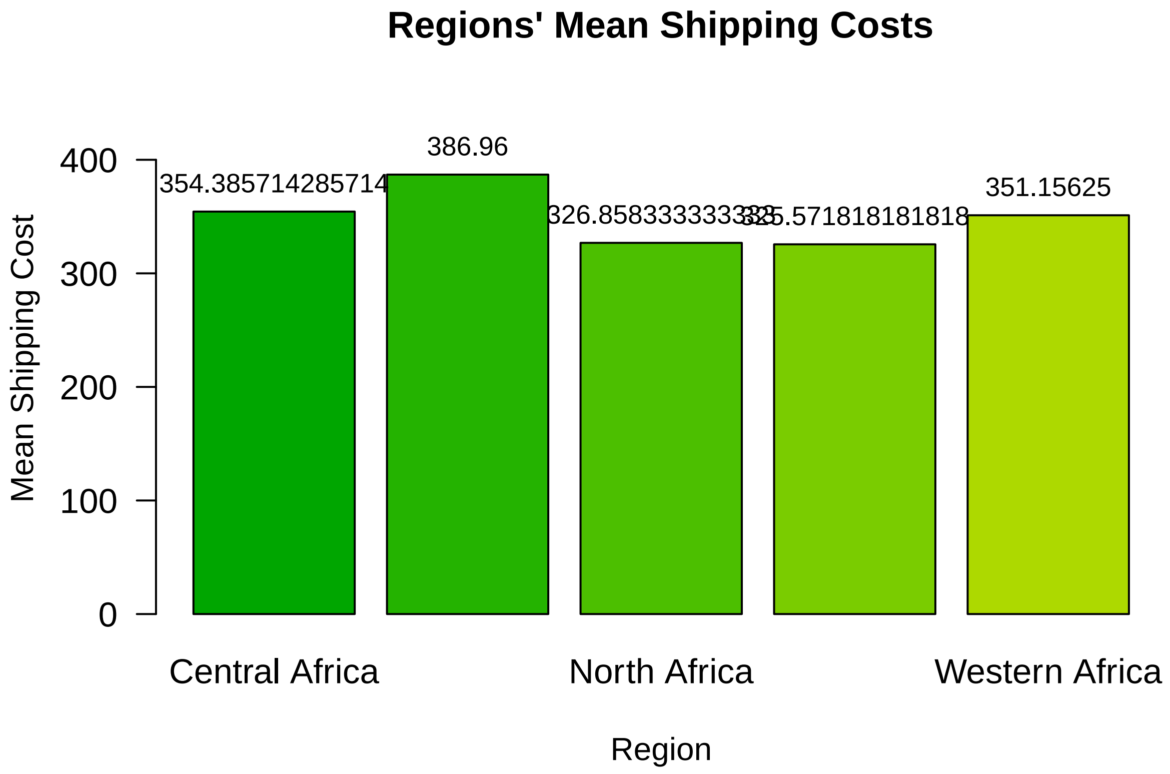
Task 8 What are the mean shipping costs per Region in the African Market?

```
meanShipping = tapply(X = t4Africa$ShippingCost, INDEX = t4Africa$Region, FUN = mean, na.rm= T)
knitr::kable(meanShipping)
```

	x
Central Africa	354.3857
Eastern Africa	386.9600
North Africa	326.8583
Southern Africa	325.5718
Western Africa	351.1562

Use a bar plot to display the information. Add colors, labels, extend the y-axis limits, and display the vlues on top of each bar.

```
meanShipping_table = table(meanShipping)

t8bar = barplot(meanShipping,
                main = "Regions’ Mean Shipping Costs",
                xlab = "Region",
                ylab = "Mean Shipping Cost",
                ylim = c(0, 450),
                col = terrain.colors(12),
                las = 1,
                cex.axis = 1.1,
                cex.names = 1.1
                 )


text(y= tapply(X = t4Africa$ShippingCost, INDEX = t4Africa$Region, FUN = mean, na.rm= T),
     t8bar,
     tapply(X = t4Africa$ShippingCost, INDEX = t4Africa$Region, FUN = mean, na.rm= T),
     cex= 0.8,
     pos= 3
     )
```



Task 9 Make a summary of tasks 3 to 8. What is the data analysis process you just followed and what information you were able to obtain? Task 3: The boxplot tells us that the Asian Pacific market is the most popular market with the most observations and the African market is the least popular with the least amount of observations.

Task 4: I filtered observations in the African market by Product Category and presented by findings with a piechart.

Task 5: I made a histogram, which displayed Product Subcategories and their frequency in the African market.

Task 6: I found the mean sales of Product Subcategories in the African Region.

Task 7: I displayed the total sales per Region in Africa.

Task 8: I found the mean shipping cost per Region in Africa.

Task 10 Explain the differences on data type designations used in R: integer, factor, double, numeric.

Integer: a type of numeric data without decimals. For example, the number of people in a course.

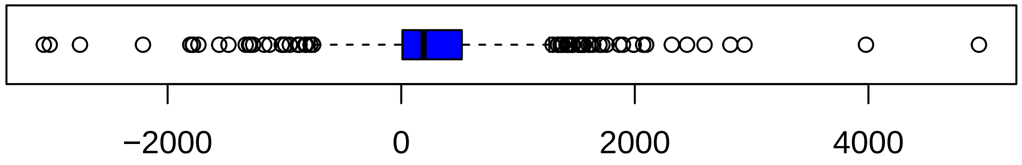
Factor: categorical variables that can be either numeric or string variables.

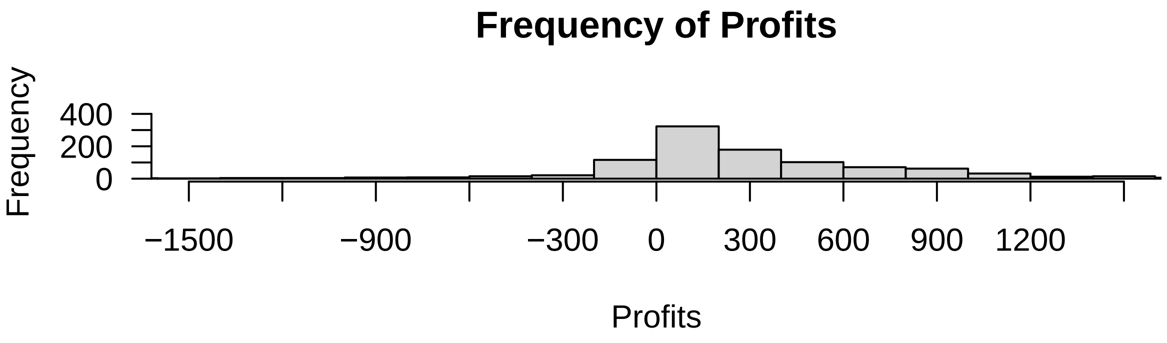
Double: stands for “double-precision floating-point”; numbers with decimals.

Numeric: continuous variables that can be measured with numbers; includes integer and double numbers.

Task 11 Go back to the original data set M2Data. Add colors to both graphs, change orientation of y-axis labels in the histogram, improve x-axis label on the histogram.

```r
# Use par() code to organize graphs in a matrix of 2 rows ad 1 column
par(mfcol=c(2,1))

#Present a box plot of all profits
boxplot(M2Data$Profits,
        col = "blue",
        horizontal = T)

#Present a histogram of all profits

hist(M2Data$Profits,
     ylim= c(0,450),
     xlim= c(-1500,1500),
     xlab= "Profits",
     breaks = 50,
     xaxp = c(-1500, 1500, 10),
     main = "Frequency of Profits",
     las = 1
     )
```





Task 12 Make observations of the two graphs you obtained on task 11. Each graph tells a different story of your data. Notice the negative values, comment on them According to the box plot, Quartile Group 1 or the lower whisker are negative values, therefore 25% of profits are negative, suggesting a loss. Fortunately 75% of our profits are greater than 0.The box plot also suggest that there are many outliers in our data set. According to the histogram, we seen our highest frequency of profits are between 0 and 200.

Task 13 Using the strategies you learn above, analyze the profits from the Latin American market.

a. Filter the data for Latin America: Code filter(). Name this new data subset as t13LATAM. Do not present this new data subset.

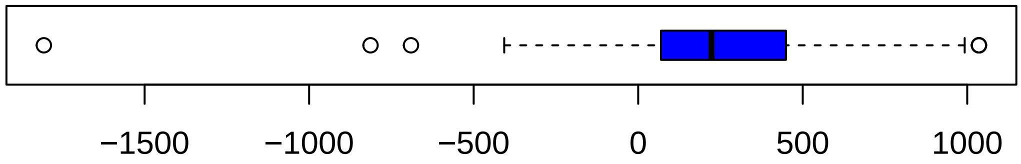
```
filter(M2Data, Market=="LATAM")
```

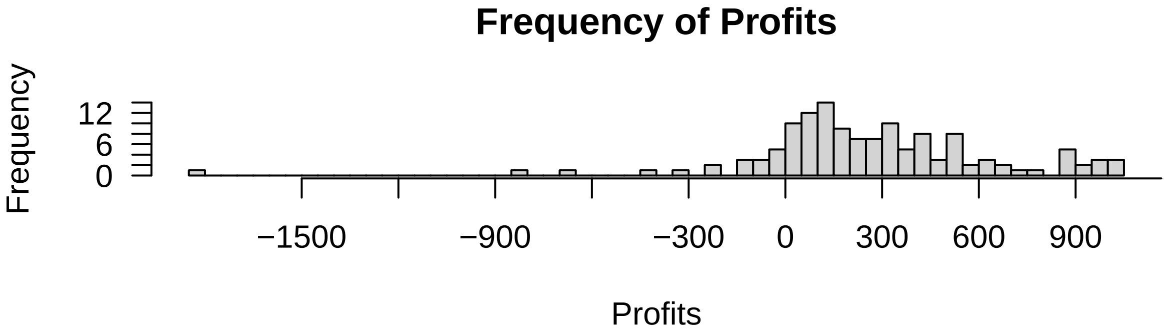
```
## # A tibble: 133 x 10
##    Region       Market Company_Segment Product_Category Product_SubCatego~ Price
##    <chr>        <chr>  <chr>           <chr>            <chr>              <dbl>
##  1 South Ameri~ LATAM  Home Office     Furniture        Chairs             2222.
##  2 Central Ame~ LATAM  Consumer        Technology       Phones             1714.
##  3 Central Ame~ LATAM  Consumer        Furniture        Tables             2106.
##  4 Caribbean    LATAM  Corporate       Technology       Phones             1697.
##  5 South Ameri~ LATAM  Consumer        Furniture        Chairs             3473.
##  6 Central Ame~ LATAM  Home Office     Technology       Phones             1704
##  7 Central Ame~ LATAM  Consumer        Office Supplies  Appliances         2443.
##  8 Central Ame~ LATAM  Corporate       Technology       Phones             2556
##  9 South Ameri~ LATAM  Home Office     Furniture        Bookcases          2473.
## 10 South Ameri~ LATAM  Consumer        Technology       Copiers            1213.
## # ... with 123 more rows, and 4 more variables: Quantity <dbl>, Sales <dbl>,
## #   Profits <dbl>, ShippingCost <dbl>
```

```
t13LATAM = filter(M2Data, Market=="LATAM")
```

b. Similar to the codes used on task 11, present a box plot and a histogram to display profit information only from the Latin American market (t13LATAM).

```
# Use par() code to organize graphs in a matrix of 2 rows ad 1 column
par(mfcol=c(2,1))

#Present a box plot of all profits
boxplot(t13LATAM$Profits,
        col = "blue",
        horizontal = T)

#Present a histogram of all profits

hist(t13LATAM$Profits,
     xlab= "Profits",
     breaks = 50,
     xaxp = c(-1500, 1500, 10),
     main = "Frequency of Profits",
     las = 1
     )
```





Make observations about your two graphs.

The boxplot suggests that variance for LATAM is higher than the variance of all of the markets. They also have fewer outliers in proportion to the total amount of data. Furthermore, the location of the 25th percentile is positive, suggesting that this region is performing above global standards.

The histogram tells use that there are few instances when profits are negative and that most of our profits from LATAM are between 0 and 1,000.

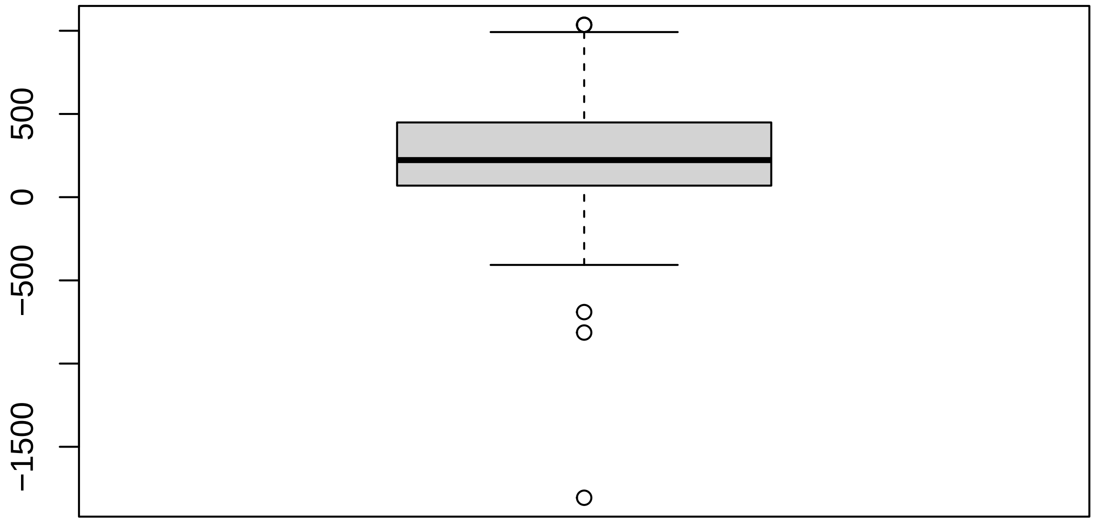
Task 14 Using the Latin American data, present a table with the total sales per region: codes tapply() and knitr::kable().

```
t14 = tapply(X = t13LATAM$Sales, INDEX = t13LATAM$Region, FUN = sum, na.rm= T)
knitr::kable(t14)
```

	x
Caribbean	196775.2
Central America	924226.2
South America	457623.3

Task 15 Present a box plot displaying the distribution of profits from the 3 regions in the Latin American market (t13LATAM).

```
boxplot(t13LATAM$Profits
        )
```



Task 16 Make observations of the results you obtained on tasks 13 and 14. Task 14: The output tells us that Central America has the most sales in the Latin America market, followed by South America, than the Caribbean.
Task 15: Our LATAM boxplot demonstrates that more than 75% of our data is between 0 and 500 and that there are four outliers.

CONCLUSIONS In conclusion, this executive summary has identified the most popular and profitable markets and regions. We learned that the Asian Pacific is the most popular market with the most observations, and the African market is the least popular with the least observations. Observations in the African market were filtered by Product Category and findings were presented with a pie chart. A histogram was created that displays Product SubCategories and their frequencies in the African market. The total sales per region in Africa is displayed, as well as the mean shipping cost per region in Africa. I defined the different data designation used in R. The frequency of total profits in the entire data set is shown. Observations were made based on box plots that represented profits among total observations. Profits were analyzed in the Latin American market. Total sales per region in Latin America is displayed. Displayed are also box plots that demonstrate the distribution of profit in Latin America regions.

Throughout the executive summary, I presented easily understandable data visualizations. The code is such that I can easily go back and make addition inferences on profits and sales per region and market. This assignment has demonstrated my ability to analyze data sets with R, make observations, and present them to executives.