M6FinalProject

William Breckwoldt

10/28/2021

ALY 6000 Final Project

ALY6000: Introduction to Analytics Northeastern Univerity William Breckwoldt

Sales 2020 Final Report

Library
Data

“‘r li-
brary(readxl)
li-
brary(readr)
li-
brary(tidyverse)
li-
brary(dplyr)
li-
brary(knitr)
li-
brary(RColorBrewer)
setwd(“C:/Users/Scott/Desktop/NEU/ALY
6000
R
Project”)
sales2020
=
read_excel(“DataSets/Project6Data.xlsx”)
“‘

Introduction

Introduction General Topic

Sales analytics helps companies identify the best sales methods and enables data driven decisions for resource allocation. It can also reveal the statically significant traits of high performers. Companies that rate their analytics as effective make up for 53% of fast-growing sales companies (companies growing faster than peers and >6% per year), while 37% of slow growers and 43% of moderate growers report having effective analytics (McKinsey 2021).The report begins by making observations on the data set and variables. Then, it searches

for patterns, inconsistencies, and other insights in the data, comparing the relationships between many categorical and numerical variables, so that we can conclude the report with data-backed conclusions.

Data Set Description, Problem Identification & Plan

This report analyzes the 2020 annual sales data of an international company from a .xlsx document that contains 27 variables, 10 of which are numerical and 17 are categorical, describing the location, shipping, products, and costs of 1000 orders. The objective of this report is to improve the company's performance by analyzing its annual sales data. The analysis will begin with a broad evaluation of the data, followed by observations on the relationships between categorical and numerical data. Subsequently, the report will filter the data to identify the most profitable markets and the most and least profitable cities, utilizing analytics and visualization tools. These tools will include the creation of new variables and the display of data using pie charts, histograms, bar plots, box plots, and tables.

1. Descriptive Statistics

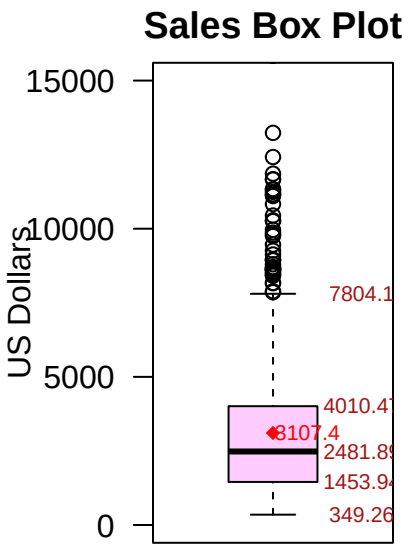
```
total_sales_mean = mean(sales2020$Sales_Total)
total_sales_sd = sd(sales2020$Sales_Total)
total_sales_range = (max(sales2020$Sales_Total)-min(sales2020$Sales_Total))
total_sales_median = median(sales2020$Sales_Total)

total_loss_mean = mean(sales2020$Total_loss)
total_loss_sd = sd(sales2020$Total_loss)
total_loss_range = (max(sales2020$Total_loss)-min(sales2020$Total_loss))
total_loss_median = median(sales2020$Total_loss)

profits_mean = mean(sales2020$Profits)
profits_sd = sd(sales2020$Profits)
profits_range = (max(sales2020$Profits)-min(sales2020$Profits))
profits_median = median(sales2020$Profits)

shipping_cost_mean = mean(sales2020$ShippingCost_Product)
shipping_cost_sd = sd(sales2020$ShippingCost_Product)
shipping_cost_range = (max(sales2020$ShippingCost_Product)-min(sales2020$ShippingCost_Product))
shipping_cost_median = median(sales2020$ShippingCost_Product)

discount_mean = mean(sales2020$Discount)
discount_sd = sd(sales2020$Discount)
discount_range = (max(sales2020$Discount)-min(sales2020$Discount))
discount_median = median(sales2020$Discount)

t1_mean = c(total_sales_mean, total_loss_mean,profits_mean,shipping_cost_mean, discount_mean)
t1_sd=c(total_sales_sd,total_loss_sd,profits_sd,shipping_cost_sd,discount_sd)
t1_range = c(total_sales_range, total_loss_range,profits_range,shipping_cost_range,discount_range)
t1_median = c(total_sales_median, total_loss_median, profits_median, shipping_cost_median,discount_media

t1_object =  c(t1_mean,t1_sd,t1_range,t1_median)

t1_matrix = matrix(t1_object, nrow = 4, byrow = TRUE)


t1_col_names = c("Total Sales","Total Loss","Profits","Shipping Cost","Discount")
t1_row_names = c("Mean","sd","Range","Median")
rownames(t1_matrix) = t1_row_names
colnames(t1_matrix) = t1_col_names
```

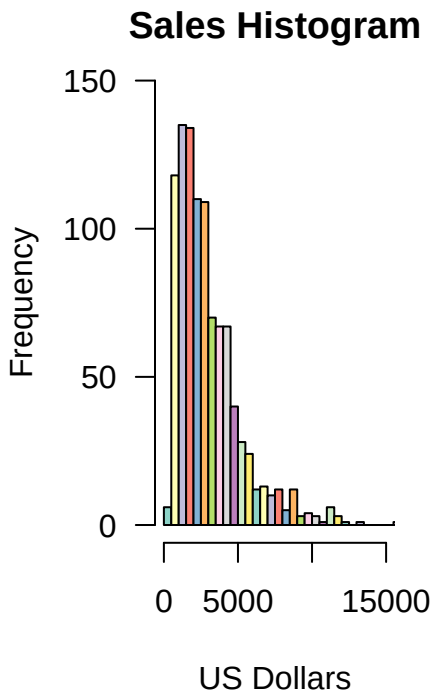
```
t1_matrix %>% knitr::kable()
```

	Total Sales	Total Loss	Profits	Shipping Cost	Discount
Mean	3107.403	823.8625	1066.5450	36.55309	0.1174600
sd	2574.062	1007.9319	901.2248	13.19651	0.1004402
Range	28934.826	11527.9200	8799.6849	83.19000	0.3000000
Median	2481.895	578.3184	834.3456	40.62500	0.1100000

2. Box Plots and Histograms of Numerical Variables 2.1

```
par(mfrow=c(1,2),mai=c(1.6,1,.4,1))

boxplot(sales2020$Sales_Total,
        main = "Sales Box Plot",
        ylab = "US Dollars",
        las=1,
        col = "#FFCCFF",
        ylim = c(0,15000))
text(y = boxplot.stats(sales2020$Sales_Total)$stats,
     labels = round(boxplot.stats(sales2020$Sales_Total)$stats,2),
     x=1.4,
     cex = 0.7,
     col = "#A11515")
points(total_sales_mean,
       pch = 18,
       col = "#F20909",
       lwd = 7)
text(y = total_sales_mean,
     labels = round(total_sales_mean,2),
     x = 1.155,
     cex = 0.7,
     col = "red")

hist(sales2020$Sales_Total,
     main ="Sales Histogram",
     breaks= 100,
     ylab= "Frequency",
     xlab = "US Dollars",
     col = brewer.pal(12,"Set3"),
     las = 1,
     ylim= c(0,150),
     xlim = c(0,15000)
     )
```



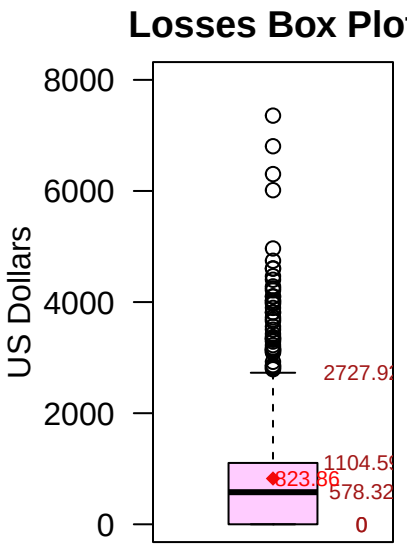


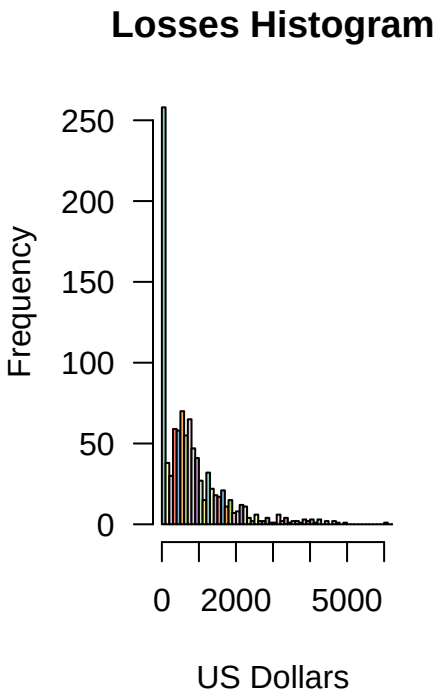
The Sales Box Plot provides us with valuable insights into the sales data. It reveals that the mean sales amount is $3,107.4 and the median sales amount is $2,481.89, indicating that the data is skewed to the right or positively skewed, which is further confirmed by the sales histogram. The box plot also shows that 50% of the data lies between the first quartile (Q1) of $1,453.94 and the third quartile (Q3) of $4,010.47, giving us an interquartile range (IQR) of $2,556.53. The majority of the data falls between the minimum value of $349.26 and the maximum value of $7,804.1. However, there are several outliers above the maximum value, with the highest one exceeding $29,000.

2.2

```
par(mfrow=c(1,2),mai=c(1.6,1,.4,1))

boxplot(sales2020$Total_loss,
        main = "Losses Box Plot",
        ylab = "US Dollars",
        las=1,
        col = "#FFCCFF",
        ylim = c(0,8000))
text(y = boxplot.stats(sales2020$Total_loss)$stats,
     labels = round(boxplot.stats(sales2020$Total_loss)$stats,2),
     x=1.4,
     cex = 0.7,
     col = "#A11515")
points(total_loss_mean,
       pch = 18,
       col = "#F20909",
       lwd = 7)
```

```
text(y = total_loss_mean,
     labels = round(total_loss_mean,2),
     x = 1.155,
     cex = 0.7,
     col = "red")

hist(sales2020$Total_loss,
     main ="Losses Histogram",
     breaks= 100,
     ylab= "Frequency",
     xlab = "US Dollars",
     col = brewer.pal(12,"Set3"),
     las = 1,
     ylim= c(0,275),
     xlim=c(0,6000)
     )
```

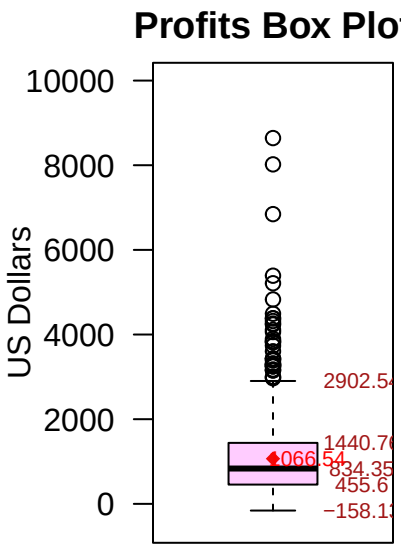


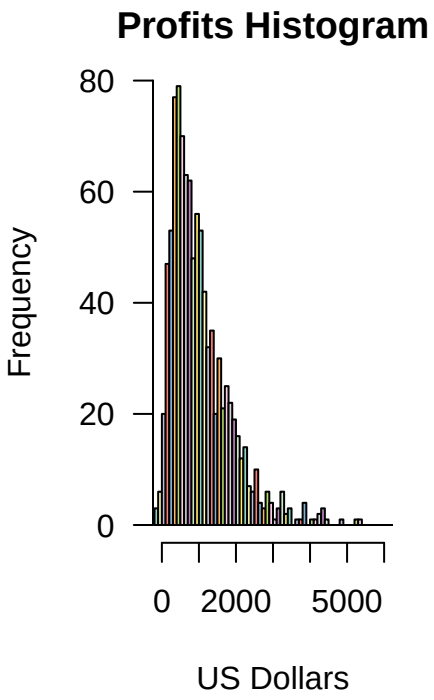


The Losses Box Plot shows that the mean is $823.86 and the median is $578.32, indicating positive skewness, which is also confirmed by the Losses Histogram. The box plot also reveals that 50% of the data falls between 0 (Q1) and $1,104.59 (Q3), resulting in an interquartile range (IQR) of $1,104.59. The majority of data falls between 0 (the minimum) and $2,727.92 (the maximum). However, there are many outliers above the maximum value, with the highest one being almost $12,000.

2.3

```r
par(mfrow=c(1,2),mai=c(1.6,1,.4,1))


boxplot(sales2020$Profits,
        main = "Profits Box Plot",
        ylab = "US Dollars",
        las=1,
        col = "#FFCCFF",
        ylim = c(-500,10000))
text(y = boxplot.stats(sales2020$Profits)$stats,
     labels = round(boxplot.stats(sales2020$Profits)$stats,2),
     x=1.4,
     cex = 0.7,
     col = "#A11515")
points(profits_mean,
       pch = 18,
       col = "#F20909",
       lwd = 7)
text(y = profits_mean,
     labels = round(profits_mean,2),
     x = 1.155,
     cex = 0.7,
     col = "red")

hist(sales2020$Profits,
     main ="Profits Histogram",
     breaks= 100,
     ylab= "Frequency",
     xlab = "US Dollars",
     col = brewer.pal(12,"Set3"),
     las = 1,
     ylim= c(0,80),
     xlim=c(0,6000)
     )
```



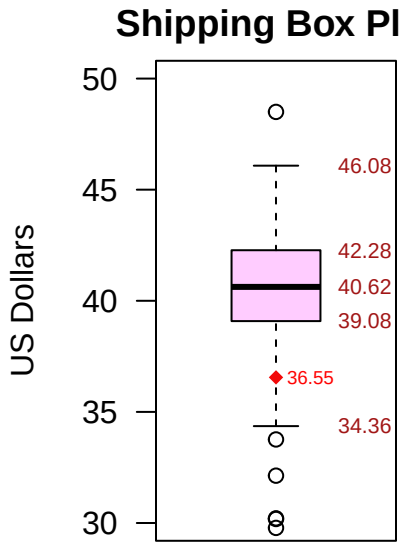


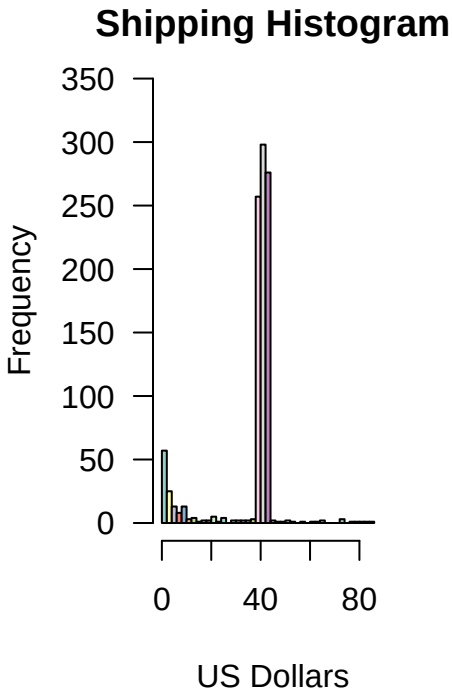
The Profits Box Plot indicates that the mean profit is $1,066.54 and the median profit is $834.35, suggesting that the data is right-skewed (or has positive skewness), which is confirmed by the Profits Histogram. The box plot also shows that 50% of our data is between $455.6 (Q1) and $1,440.76 (Q3), giving us an interquartile range (IQR) of $985.16. The majority of our data falls between -$158.13 (the minimum) and $2,727.92 (the maximum).

2.4

```
par(mfrow=c(1,2),mai=c(1.6,1,.4,1))

boxplot(sales2020$ShippingCost_Product,
        main = "Shipping Box Plot",
        ylab = "US Dollars",
        las=1,
        col = "#FFCCFF",
        ylim = c(30,50))
text(y = boxplot.stats(sales2020$ShippingCost_Product)$stats,
     labels = round(boxplot.stats(sales2020$ShippingCost_Product)$stats,2),
     x=1.4,
     cex = 0.7,
     col = "#A11515")
points(shipping_cost_mean,
       pch = 18,
       col = "#F20909",
       lwd = 7)
text(y = shipping_cost_mean,
     labels = round(shipping_cost_mean,2),
```

```
     x = 1.155,
     cex = 0.6,
     col = "red")

hist(sales2020$ShippingCost_Product,
     main ="Shipping Histogram",
     breaks= 50,
     ylab= "Frequency",
     xlab = "US Dollars",
     col = brewer.pal(12,"Set3"),
     las = 1,
     ylim= c(0,350),
     xlim=c(0,90)
     )
```

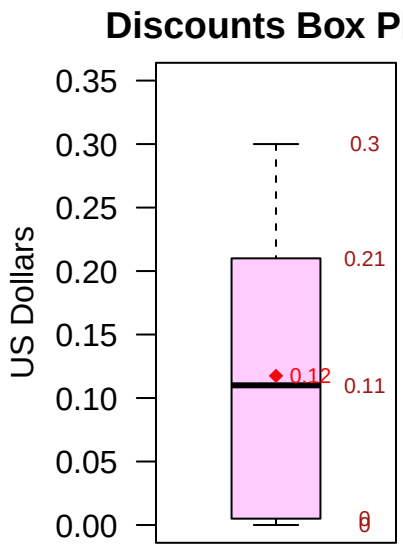


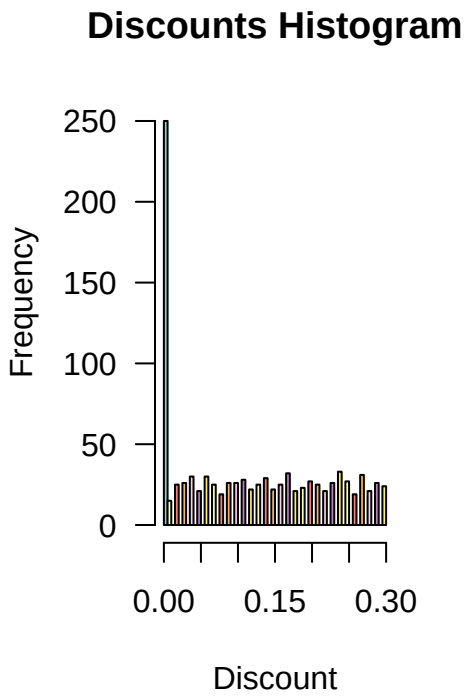


The Shipping Box Plot indicates that the data is left-skewed or negatively skewed, as evidenced by the mean of $36.55 and the median of $40.62, which is further supported by the Shipping Histogram. Additionally, the box plot illustrates that 50% of the data falls between the first quartile (Q1) of $39.08 and the third quartile (Q3) of $42.28, resulting in an interquartile range (IQR) of 3.2. This is an interesting finding since the mean of $36.55 and the majority of orders, which are between $34.36 and $46.08, are significantly different from the IQR. However, this can be explained by the large number of orders with shipping costs that are equal to or close to 0.

2.5

```
par(mfrow=c(1,2),mai=c(1.6,1,.4,1))

boxplot(sales2020$Discount,
        main = "Discounts Box Plot",
        ylab = "US Dollars",
        las=1,
        col = "#FFCCFF",
        ylim = c(0,.35))
text(y = boxplot.stats(sales2020$Discount)$stats,
     labels = round(boxplot.stats(sales2020$Discount)$stats,2),
     x=1.4,
     cex = 0.7,
     col = "#A11515")
points(discount_mean,
       pch = 18,
       col = "#F20909",
       lwd = 7)
text(y = discount_mean,
     labels = round(discount_mean,2),
     x = 1.155,
     cex = 0.7,
     col = "red")

hist(sales2020$Discount,
     main ="Discounts Histogram",
     breaks= 50,
     ylab= "Frequency",
     xlab = "Discount",
     col = brewer.pal(12,"Set3"),
     las = 1,
     ylim= c(0,275),
     xlim=c(0,.3)
     )
```

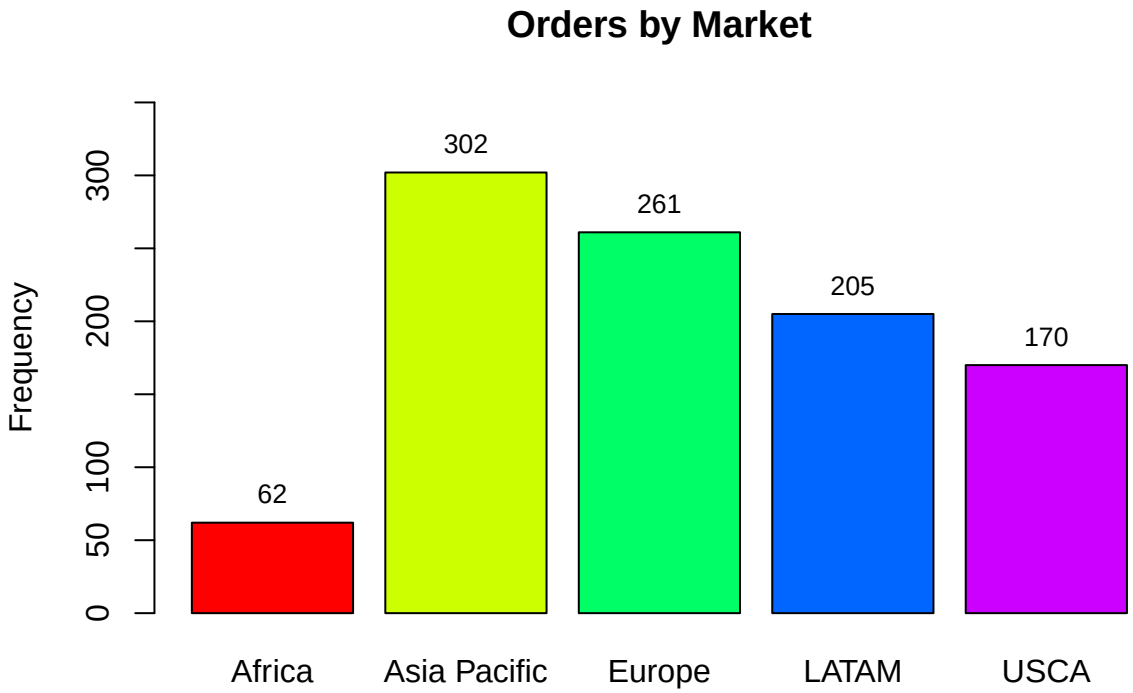




The Discounts Box Plot tells us that our mean is 0.12 and our median is 0.11, suggesting the data is right-skewed, which appears true from the Discounts Histogram. It also demonstrates that 50% of our data is between 0 (Q1) and 0.21 (Q3), giving us an IQR of 0.21. The majority of orders have discounts between 0 and 0.3.

3. Counting of Categorical Variable 3.1

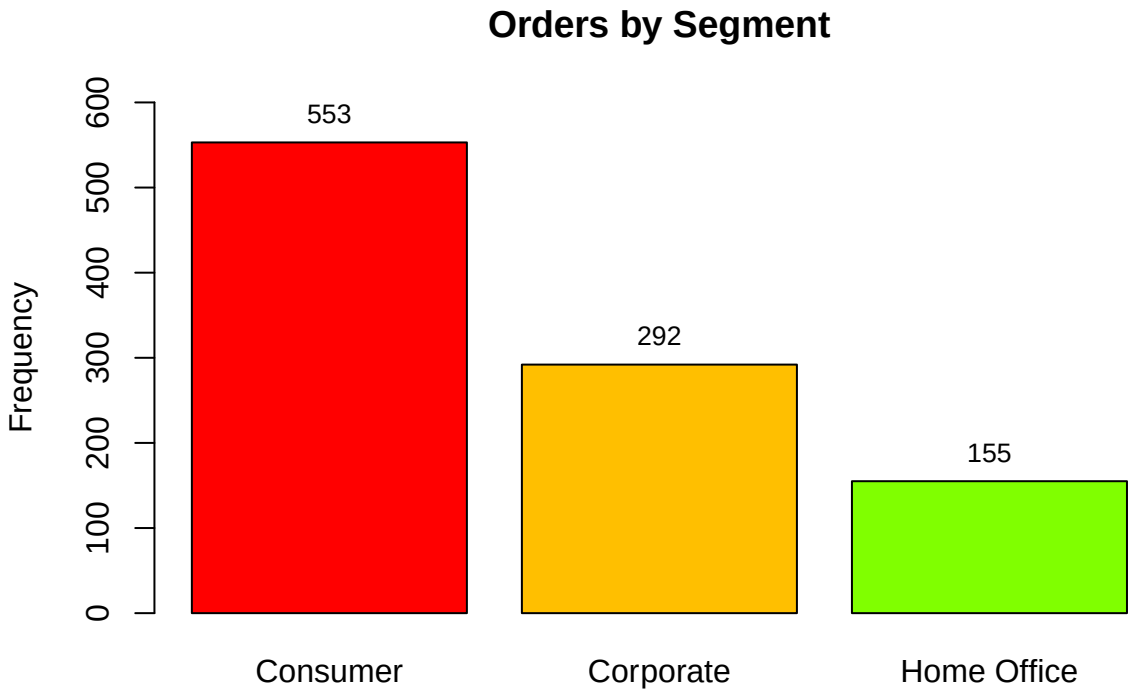
```
bp31 = barplot(table(sales2020$Market),
         main= "Orders by Market",
         ylab="Frequency",
         col=rainbow(5),
         ylim=c(0,350))
text(y=table(sales2020$Market),
     bp31,
     table(sales2020$Market),
     cex=0.8,
     pos=3)
```



The market with the highest number of orders is Asia Pacific, while the market with the lowest number of orders is Africa. Further analysis can be done to identify the factors contributing to the difference between the most and least popular markets, and to explore ways to improve the performance of the African market.

3.2

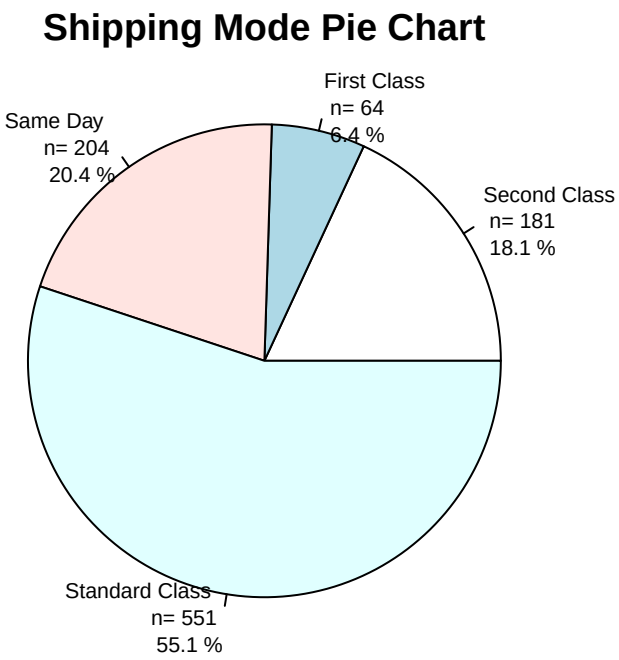
```
bp32 = barplot(table(sales2020$Segment),
        main= "Orders by Segment",
        ylab="Frequency",
        col=rainbow(8),
        ylim=c(0,600))
text(y=table(sales2020$Segment),
     bp32,
     table(sales2020$Segment),
     cex=0.8,
     pos=3)
```



The majority of orders are from consumers, but a significant number of orders also come from corporate and home office.

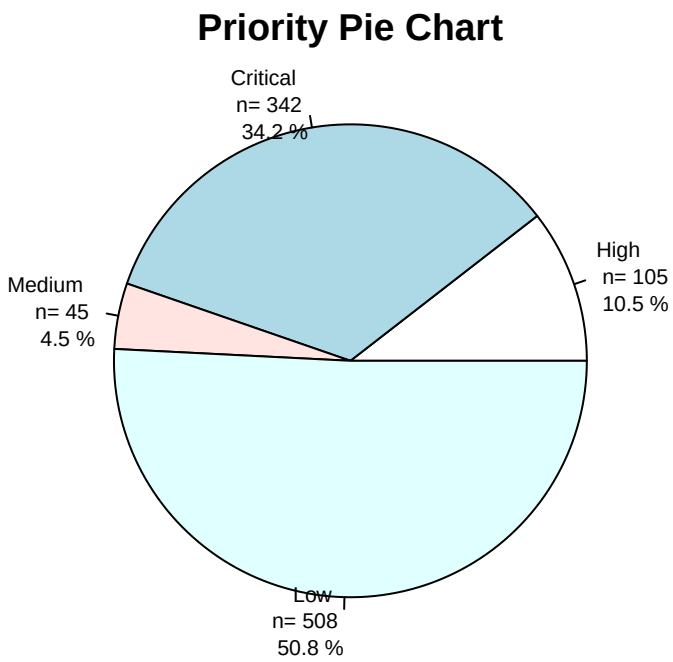
3.3

```
f = sales2020$ShipMode
f= table(f)


g = (f/1000)*100

pieLabels = paste(unique(sales2020$ShipMode),
                  "",
                  "\n",
                  "n=",
                  f,
                  "\n",
                  round(g,1),
                  "%")

pie1 = pie(f,
           main="Shipping Mode Pie Chart",
           labels= pieLabels,
           radius = 1,
           cex = .7)
```



The Shipping Mode Pie Chart shows that standard class shipping is the most popular shipping mode, accounting for 55.1% of all orders, followed by same day shipping (20.4%), second class shipping (18.1%), and first class shipping (6.4%). 3.4

```
h = sales2020$OrderPriority
h= table(h)


i = (h/1000)*100

pieLabels2 = paste(unique(sales2020$OrderPriority),
                   "",
                   "\n",
                   "n=",
                   h,
                   "\n",
                   round(i,1),
                   "%")

pie2 = pie(h,
           main ="Priority Pie Chart",
           labels= pieLabels2,
           radius = 1,
           cex = .7)
```

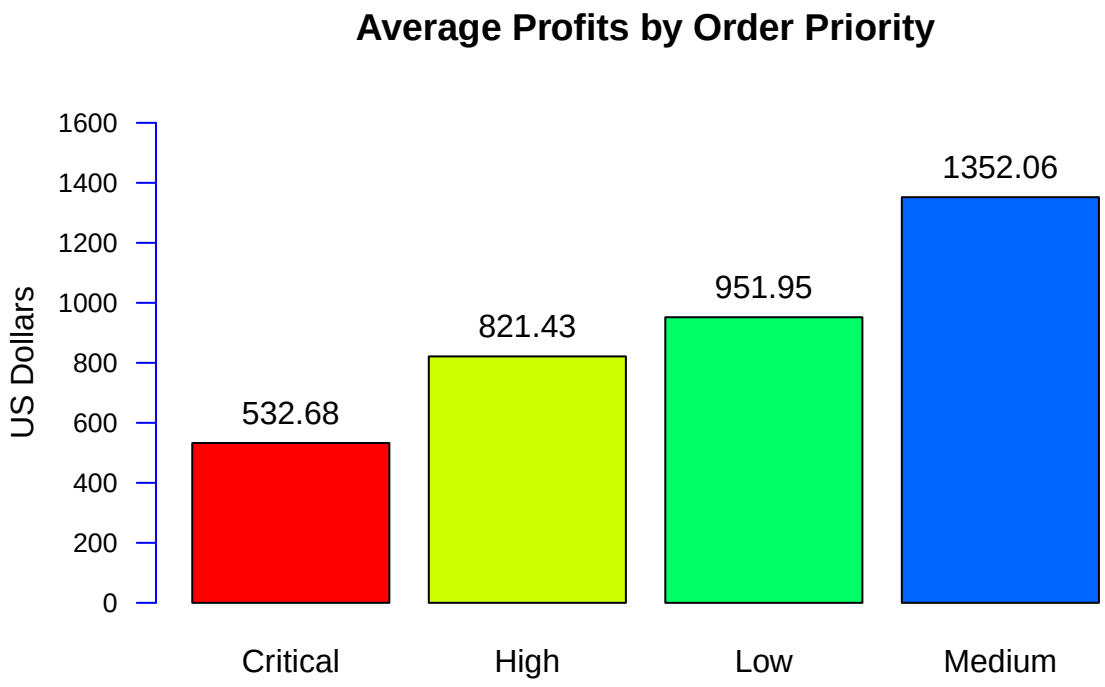


The Priority Pie Chart shows the proportion and frequency of orders based on their priority. The majority of orders (50.8%, n=503) are classified as Low priority, followed by Critical priority (34.2%, n=342), High priority (10.5%, n=105), and Medium priority (4.5%, n=45). These differences between the frequency of orders in different priority levels may suggest that there is a need to reevaluate the system of measuring order priority and make sure that the priority level accurately reflects the urgency of the order.

4. Sub-Category Analysis 4.1 Average Profits by Order Priority This will allow us to better understand if order priority affects profits, as it appeared from the last set of graphs that the Order Priority system may need improvement.

```
avg_prof_per_order = tapply(sales2020$Profits, INDEX = sales2020$OrderPriority, FUN= mean)

par(mai=c(1,1,1,0.4))
plot19= barplot(avg_prof_per_order,
        main = "Average Profits by Order Priority",
        axes=FALSE,
        col=rainbow(5),
        ylim=c(0,1600))
axis(side= 2,
     labels = T,
     at = c(0,200,400,600,800,1000,1200,1400,1600),
     tick=8,
     col.axis = 'black',
       col = "blue",
     cex.axis = .8,
     cex.axis = .8,
     las = 1
```

```
       )
mtext("US Dollars",
      side=2,
      col = 'black',
      line = 3,
      cex.lab = .8)
text(y= avg_prof_per_order,
     plot19,
     round(avg_prof_per_order,2),
     cex=1,
     pos=3
     )
```

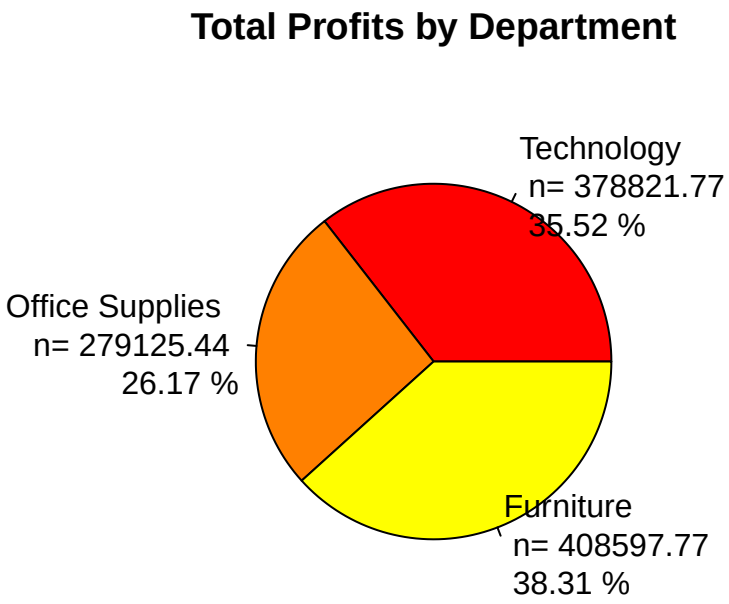


The Average Profits by Order Priority Bar Plot indicates that Medium priority orders generate the highest average profits at $1,352.,06, followed by High priority orders at $1,200.83, Low priority orders at $1041.31, and Critical priority orders at $532.68. This stark difference in average profits between Critical priority orders and the other priority levels may indicate that the company's order priority system needs to be re-evaluated to ensure that the priority level accurately reflects the urgency of the order and its impact on profitability.

4.2 Total Profits by Department This will enable us to assess the performance of each department, identify any areas that require improvement, and recognize those departments that are excelling in creating value for the company.

```
prof_per_dept = tapply(sales2020$Profits, INDEX = sales2020$Department, FUN= sum)

par(mai=c(1,1,1,0.4))
```

```
w=(prof_per_dept/1066544.98)*100

pieLabels3 = paste(unique(sales2020$Department),
                   "",
                   "\n",
                   "n=",
                   round(prof_per_dept,2),
                   "\n",
                   round(w,2),
                   "%")
pie(prof_per_dept,
        main = "Total Profits by Department",
    labels=pieLabels3,
        col=rainbow(12),
        ylim=c(0,1500))
```

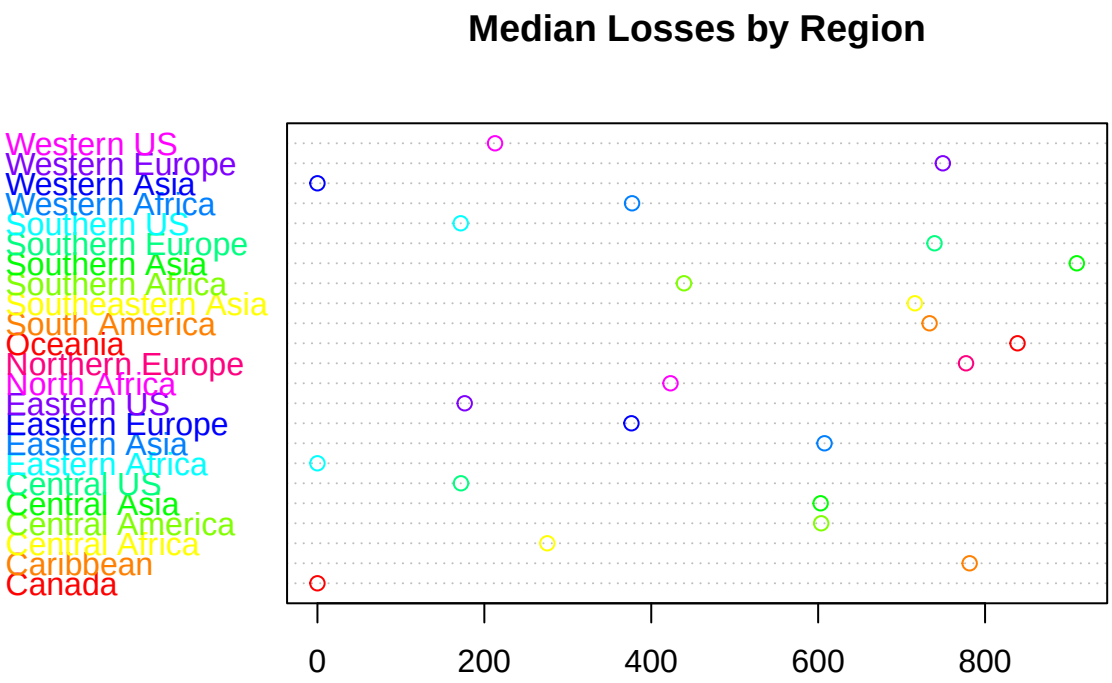


The Profits by Department Bar Plot shows that profits are distributed relatively evenly between the three departments. Furniture is responsible for the most profits (38.31%, $408,597.77), followed closely by Technology (35.52%, $378,821.77). Office Supplies earned the least profits (26.17%, $279,125.44).

4.3 Median Losses by Region The median Losses per region will enable us to identify the regions with the highest and lowest median losses. This information will help the company determine which regions may be more or less risky for conducting business.

```
md_loss_reg = tapply(sales2020$Total_loss, INDEX = sales2020$Region, FUN= median)
```

```
par(mai=c(1,1,1,0.4))
dotchart(md_loss_reg,
         main = "Median Losses by Region",
         col=rainbow(12),
         ylim=c(0,50))
```



Southern Asia, Oceania, and the Caribbean are the regions with the highest median losses. Further research may be required to identify the underlying causes of these losses and to develop effective strategies to mitigate them.

5. Filter Data: What are the total profits per Market? 5.1 Massachusetts Sales Data

```
ma_sales2020 = filter(sales2020, Region == "Eastern US" & State =="Massachusetts")
ma_sales2020
```

```
## # A tibble: 4 x 27
##   OrderID       OrderDate ShipDate ProductID CustomerID CustomerName City  State
##   <chr>         <chr>     <chr>    <chr>     <chr>      <chr>        <chr> <chr>
## 1 CA-2015-CL11~ 9/15/2020 9/19/20~ OFF-AR-3~ CL-118901~ Carl Ludwig  Ever~ Mass~
## 2 CA-2015-CL11~ 9/22/2020 9/26/20~ TEC-PH-4~ CL-118901~ Carl Ludwig  Ever~ Mass~
## 3 CA-2015-CL11~ 9/19/2020 9/21/20~ OFF-BI-6~ CL-118901~ Carl Ludwig  Ever~ Mass~
## 4 CA-2015-CL11~ 9/22/2020 9/24/20~ FUR-BO-3~ CL-118901~ Carl Ludwig  Ever~ Mass~
## # i 19 more variables: Country <chr>, Region <chr>, Market <chr>,
## #   Segment <chr>, Department <chr>, Division <chr>, ProductName <chr>,
## #   OrderPriority <chr>, ShipMode <chr>, Price <dbl>, Quantity <dbl>,
## #   Discount <dbl>, ShippingCost_Product <dbl>, Sales_Total <dbl>,
```

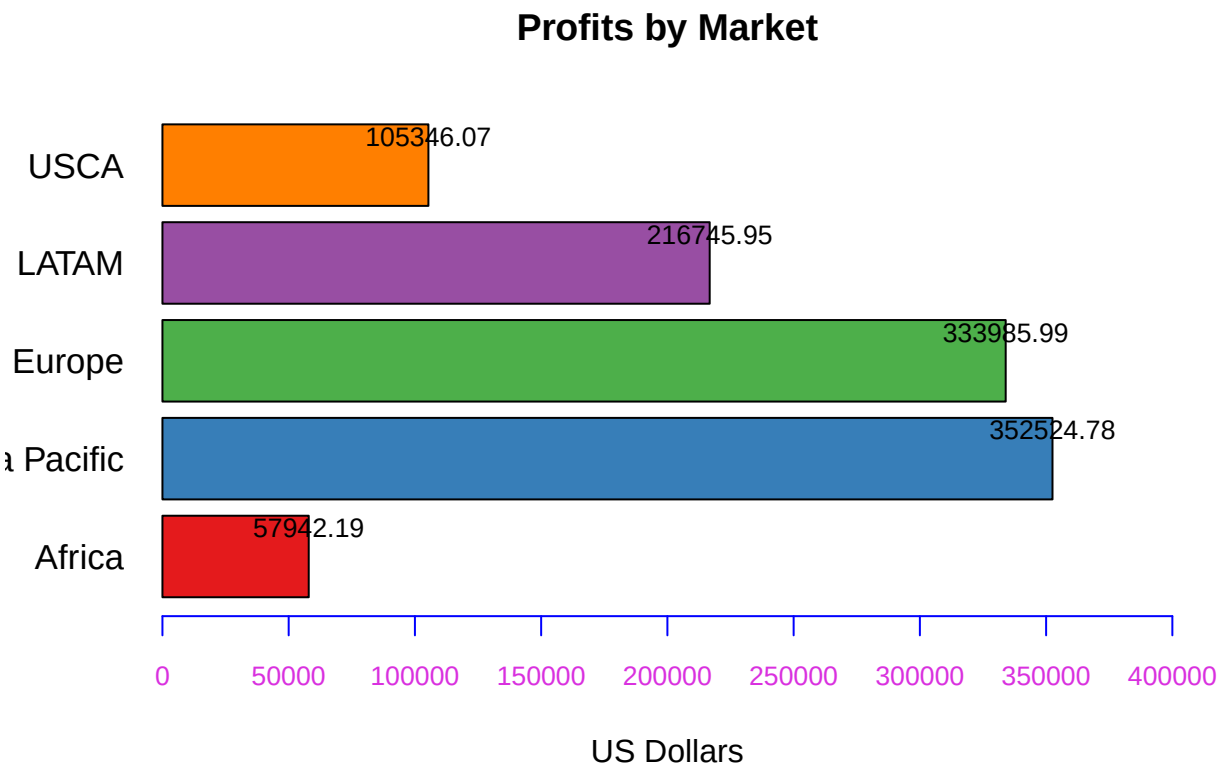
```
## #   Returns <dbl>, LossPerReturn <dbl>, Total_loss <dbl>, Net_Sale <dbl>,
## #   Profits <dbl>
```

5.2 Total Profits Per Market

```
prof_per_mark = tapply(sales2020$Profits, INDEX = sales2020$Market, FUN= sum)

knitr::kable(prof_per_mark)
```

	x
Africa	57942.19
Asia Pacific	352524.78
Europe	333985.99
LATAM	216745.95
USCA	105346.07

```
prof_bp = barplot(prof_per_mark,
                  main = "Profits by Market",
                  xlab = "US Dollars",
     horiz = TRUE,
     axes = FALSE,
     xlim = c(0,400000),
     col = brewer.pal(9,"Set1"),
     las = 1,
     cex.axis = 1.1,
     cex.names = 1.1
)

axis(side= 1,
     labels = T,
     at = c(0,50000,100000,150000,200000,250000,300000,350000, 400000),
     tick=8,
     col.axis = '#D930DF',
       col = "blue",
     cex.axis = .8,
     cex.axis = .8,
     las = 1
       )
text(prof_per_mark,
     prof_bp,
     round(prof_per_mark,2),
     cex=0.8,
     pos = 3)
```



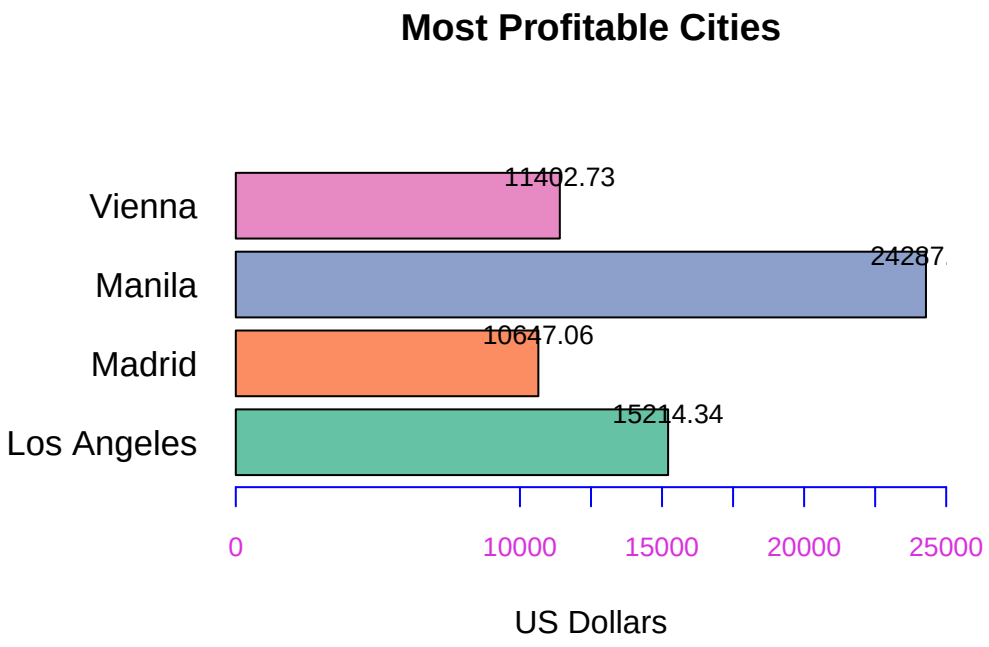
The Profits by Market Bar Plot displays the total profits earned by each market. The Asia Pacific market has the highest total profits with $352,524.78, followed by Western Europe ($252,477.20) and North America ($223,295.73). On the other hand, the Africa market has the lowest total profits with $57,942.19. This aligns with the Orders by Market Bar Plot (3.1) from earlier, which displayed the total number of orders. However, it should be noted that the United States and Central America market bar is closer to the Africa bar and further from the others, suggesting that this market may have lower profits per order. Further analysis may be necessary to identify the factors contributing to this trend.

6. The Most and the Least Profitable Cities

```
prof_per_city = tapply(sales2020$Profits, sales2020$City, FUN= sum)
```

```
dfprof_city = data.frame(prof_per_city)

most_profitable_cities = filter(dfprof_city, prof_per_city > 10000)

least_profitable_cities = filter(dfprof_city, prof_per_city < 0)

par(mai=c(1.4,1.4,1.4,1.4))
mostprof_bp = barplot(most_profitable_cities$prof_per_city,
                  main = "Most Profitable Cities",
                  xlab = "US Dollars",
     horiz = TRUE,
     axes = FALSE,
     xlim = c(0,25000),
     col = brewer.pal(9,"Set2"),
     las = 1,
```

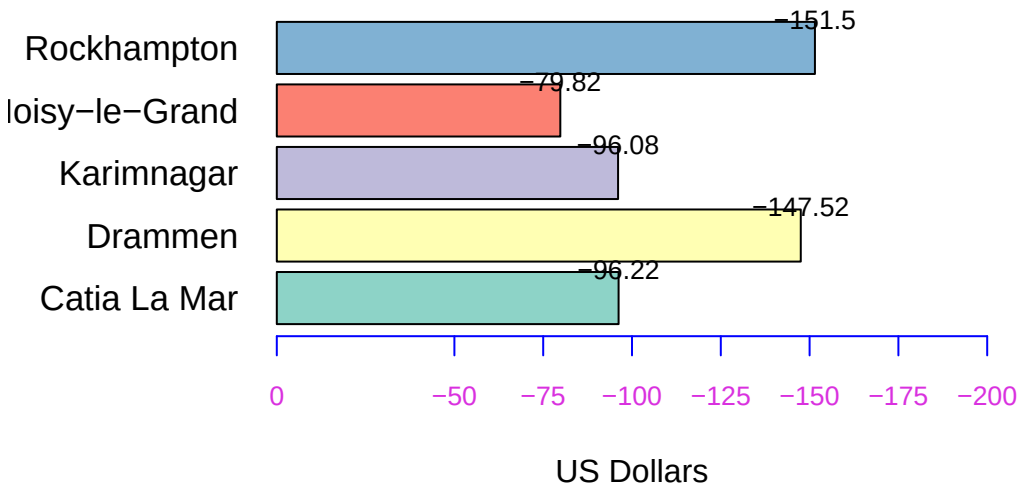
```
     cex.axis = 1.1,
     cex.names = 1.1,
     names.arg = c("Los Angeles","Madrid","Manila","Vienna")
)

axis(side= 1,
     labels = T,
     at = c(0,10000,12500,15000, 17500, 20000, 22500,25000),
     tick=8,
     col.axis = '#D930DF',
       col = "blue",
     cex.axis = .8,
     cex.axis = .8,
     las = 1
       )
text(most_profitable_cities$prof_per_city,
     mostprof_bp,
     round(most_profitable_cities$prof_per_city,2),
     cex=0.8,
     pos = 3)
```



The four most profitable cities are Manila ($24,287.16), Los Angeles ($15,214.34), Vienna ($11,402.73), and Madrid ($10,647.06).

```
par(mai=c(1.4,1.4,1.4,1.4))
```

```
lprof_bp = barplot(least_profitable_cities$prof_per_city,
                   main = "Least Profitable Cities",
                   xlab = "US Dollars",
     horiz = TRUE,
     axes = FALSE,
     xlim = c(0,-200),
     col = brewer.pal(9,"Set3"),
     las = 1,
     cex.axis = 1.1,
     cex.names = 1.1,
     names.arg = c("Catia La Mar","Drammen","Karimnagar","Noisy-le-Grand","Rockhampton")
)

axis(side= 1,
     labels = T,
     at = c(0,- 50, -75, -100, -125, -150, -175, -200),
     tick=8,
     col.axis = '#D930DF',
       col = "blue",
     cex.axis = .8,
     cex.axis = .8,
     las = 1
       )
text(least_profitable_cities$prof_per_city,
     lprof_bp,
     round(least_profitable_cities$prof_per_city,2),
     cex=0.8,
     pos = 3)
```



The least profitable cities in the dataset are Rockhampton with a loss of -$151.5, followed by Drammen with a loss of -147.52, Catia La Mar with a loss of -$96.22, Karimnagar with a loss of -$96.08, and Noisy-le-Grand with a loss of -$79.82.

7. Conclusions What has the report told us?

This report provides an overview of sales data analytics, emphasizing its significance and potential applications. We presented a data set consisting of 1000 orders and 27 variables, including 17 categorical and 10 numerical variables. Additionally, we outlined our approach for analyzing the data set. Subsequently, we conducted a comprehensive analysis of the data to derive valuable insights.

First, we obtained the descriptive statistics of 5 numerical variables (Sales Total, Total Loss, Profits, Shipping Cost, and Discounts) and created a table. Total Sales, Total Loss, and Profits all had relatively high standard deviations, suggesting high variance, compared to Shipping Cost and Discount. Our Box Plots and Histograms also supported this conclusion. The Sales Box Plot and Histogram indicated a right-skewed distribution, with the majority of data between $349.26 and $7,804.1 and many outliers above the maximum (the highest is slightly over $29,000). The Losses Box Plot and Histogram showed a similar right-skewed distribution, with over a quarter of orders experiencing $0 or close to $0 losses, and the majority of Total Losses falling between $0 and $2,727.92. However, there were many outliers above this range.

The Profits Box Plot and Histogram also show that Profits are right-skewed. Half of the orders have Profits between $455.6 and $1,440.76, and the majority of orders fall between -$158.13 and $2,902.54. There are many outliers above the maximum. The Shipping Box Plot and Histogram demonstrate that the majority of orders are between $34.36 and $46.08. However, over 50 orders had 0 or close to 0 Shipping Costs. The Discounts Bar Plot and Histogram indicate that the data is normally distributed between 0 and 0.3. It should be noted that roughly 250 or a quarter of the orders had no discounts applied.

When examining the categorical variables of Market, Segment, Shipping Mode, and Order Priority, we found that the Asia Pacific had the highest frequency of orders at 302, followed by Europe (261), Latin America

(205), US and Central America (170), and Africa (62). For the Segment variable, the Consumer segment had the most orders with 553, followed by Corporate (292) and Home Office (155). Standard Class shipping was the most popular shipping mode, accounting for 551 orders (55.1%), followed by Same Day (204; 20.4%), Second Class (181; 18.1%), and First Class (64; 6.4%). In terms of order priority, we observed that Low was the most frequent priority level with 508 orders (50.8%), followed by Critical (342; 34.2%), High (105; 10.5%), and Medium (45; 4.5%).

Next, we combined three categorical and three numerical variable to create data visualizations that demonstrated Average Profits by Order Priority, Total Profits by Department, and the Median Losses by Region. The Average Profits by Order Priority Bar Plot showed us that, surprisingly, Medium priority orders average the highest profits at $1,352.06, followed by Low ($951.95), High ($821.43), and Critical ($532.68). The Total Profits by Department Pie Chart demonstrates that Profits were pretty even distributed between the three Departments: Furniture ($408,597.77; 38.31%), Technology ($378,821.77; 35.52%), and Office Supplies ($279,125.44; 26.17%). The Median Losses by Region Dot Chart demonstrated that the regions with the most median Losses (>$800) were Southern Asia and Oceania.

We then filtered data. First we created a Massachusetts Sales Data Set, which only consisted of four orders. Then, we calculated the Total Profits Per Market. We found that the market with the highest profits is Asia Pacific ($352,524.78), followed by Europe ($333,985.99), LATAM ($216,745.95), USCA ($105,346.07), and Africa ($57,942.19). Lastly, we looked at profits per city and identified the four most profitable cities: Manila ($24,287.60), Los Angeles ($15,214.34), Vienna ($11,402.73), and Madrid ($10,647.06). We also identified the five least profitable cities: Rockhampton (-$151.50), Drammen (-$147.52), Catia La Mar (-$96.22), Karimnagar (-$96.08), and Noisy-le-Grand (-$79.82).

Observations and Recommendations

An important observation to note is that despite the USCA having a much larger number of orders (170) compared to Africa (62), they are relatively close in terms of Total Profits (USCA = $105,346.07 and Africa = $57,942.19). This is interesting, especially since the US has the second most profitable city (Los Angeles) and African regions experience high median losses per order (3 out of 5 African regions experience between $400 and $500 median total losses). Based on this, I recommend that the company researches this further and considers dedicating more resources to African sales orders, as mitigating losses in the African market could potentially result in higher company profits. However, it's also worth noting that another reason for this could be the extremely high number of median losses (~$800) in the Caribbean.

Another major observation is that there appear to be issues with the Order Priority Level System. The distribution of orders and the average profits by order priority level seems odd. Low priority accounts for 50.8% of orders, Critical accounts for 34.2%, High accounts for 10.5%, and Medium accounts for 4.5%. On the other hand, order priority with the highest average profits is Medium ($1,352.05), followed by Low ($951.95), High ($821.43), and Critical ($532.68). The high number of Critical orders, which average the least amount of profits, may put unnecessary pressure on company employees. I recommend that the company research alternatives or adjust their order priority system to avoid this issue.

Lastly, by examining populations and Average Profits in the most profitable cities (Manila (population: 14,158,573), Los Angeles (3,983,540), Vienna (1,920,949), and Madrid (6,668,865)) and the least profitable cities (Rockhampton (population: 80,665), Drammen (64,776), Catia La Mar (106,822), Karimnagar (406,466), and Noisy-le-Grand (71,233)), it appears that population size may be a factor in profitability. This suggests that it may be more profitable for the company to focus on orders in larger cities.

8. References Atkins, C., Valdivieso De Uster, M.,Mahdavian,M. & Yee, L. (14 Dec 2016) Unlocking the power of data in sales. McKinsey & Company. https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/unlocking-the-power-of-data-in-sales.

9. Appendix M6Final Project.Rmd was the file used by William Breckwoldt to create this HTML report. Instructions courtesy of Professor Dee Chiluiza, PhD., Northeastern University.