R Data Challenge: Analyzing Customer Returns Data

The coding challenge below was obtained from a friend who took the "NYC Data Science" course

R Coding Challenge Session I

For this R review section, we mainly focus on data cleaning and data visualizations.

For problems 2 and 3, you should mainly use **dplyr** and **ggplot2** to construct some plots and also provide **brief interpretations** about your findings.

Problem 1: Dataset Import & Cleaning

The data comes from a global company, including orders from 2012 to 2015. Import the dataset **Order** and do some basic EDA.

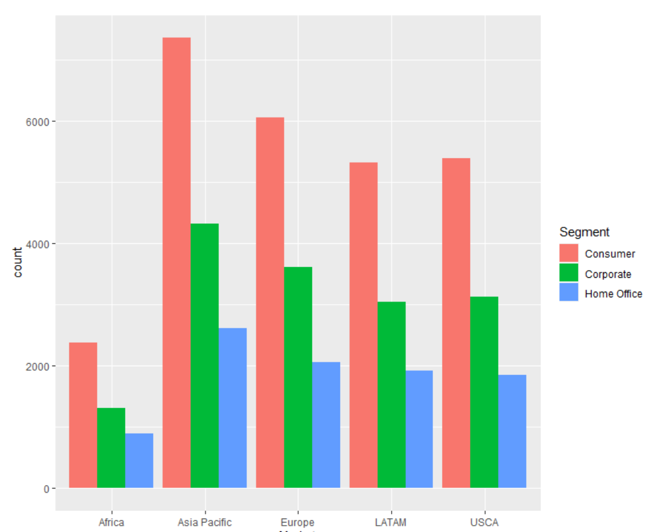
Check **"Profit"** and **"Sales"** in the dataset, convert these two columns to numeric data.

```
# Fill in your code here

#change directory
setwd("C:/Users/Acer/Desktop/projects/R Data Challenge/data/")
#read files
orders=read.csv(file="Orders.csv")
returns=read.csv("Returns.csv")

#view as table
View(orders)
View(returns)

#check the datatypes of the columns
sapply(orders,class)

#extract profit and sales and convert
profit=as.numeric(orders$Profit)
sales=as.numeric(orders$Sales)

#convert directly
orders[,"Profit"]=as.numeric(orders[,"Profit"])
orders[,"Sales"]=as.numeric(orders[,"Sales"])


#EDA
library(ggplot2)
ggplot(data=orders, aes(x=Market))+geom_bar(aes(fill=Segment),position="dodge
")

#load dplyr so we can group and summarise
library(dplyr)
by_market=group_by(orders, Market)
```

```
profit_by_market=summarise(by_market, total_profit=sum(Profit))
profit_by_market=arrange(profit_by_market,desc(total_profit))

sales_by_market=summarise(by_market, total_sales=sum(Sales))
sales_by_market=arrange(sales_by_market,desc(total_sales))

profit_by_market
sales_by_market

#create new column
orders$profit_margin=orders$Profit/orders$Sales
by_market=group_by(orders, Market)
pm_by_market=summarise(by_market, ave_pm=mean(profit_margin))
pm_by_market=arrange(pm_by_market,desc(ave_pm))
pm_by_market

#because of the margins, we now check what kind of product categpry is most p
opular in each market
#and get the average profit margin per product category
by_cat=group_by(orders, Category)
pm_by_cat=summarise(by_cat, ave_pm=mean(profit_margin))
pm_by_cat=arrange(pm_by_cat,desc(ave_pm))
pm_by_cat

by_mkt_and_cat=group_by(orders, Market,Category)
pm_by_grp=summarise(by_mkt_and_cat, ave_pm=mean(profit_margin))
pm_by_grp
sales_by_grp=summarise(by_mkt_and_cat, total_sales=sum(Sales))
sales_by_grp

ggplot(data=orders, aes(x=Market))+geom_bar(aes(fill=Category),position="fill
")
```

<u>Observations from EDA:</u>

Strongest segment is every “Market” is “Consumer” followed by “Corporate” then “Home Office.”



The company likely has higher profit margins in USCA since sales in LATAM>USCA but profits in USCA>LATAM.

```
> profit_by_market
# A tibble: 5 x 2
  Market       total_profit
  <fct>               <dbl>
1 Asia Pacific    121343408
2 Europe          115103688
3 USCA            102004204
4 LATAM            92860488
5 Africa           43311015
```

```
> sales_by_market
# A tibble: 5 x 2
  Market       total_sales
  <fct>              <dbl>
1 Asia Pacific   143490630
2 Europe         118882295
3 LATAM          105980976
4 USCA           102771926
5 Africa          44953738
```

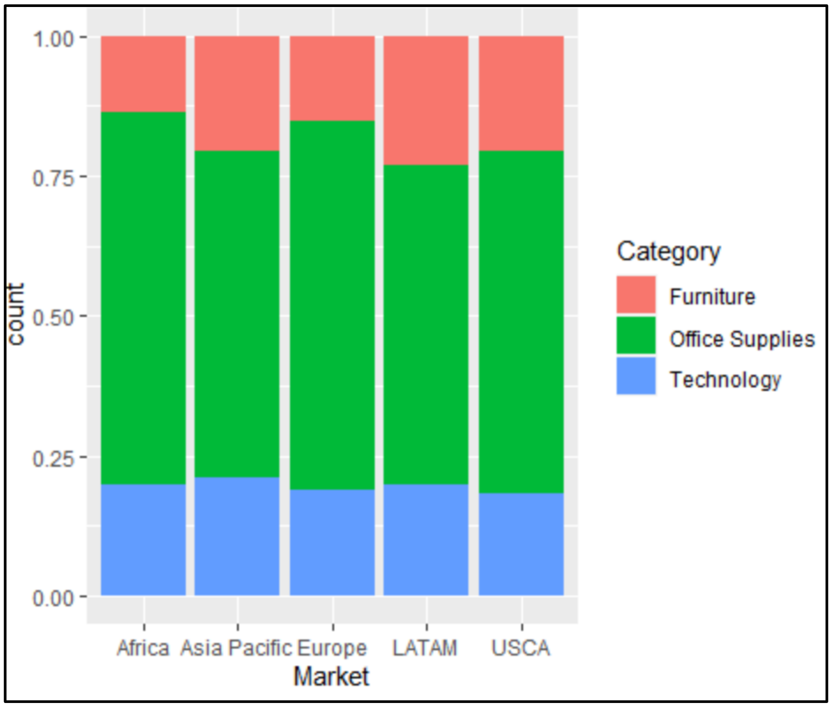
We confirm this by getting the average profit margin per market. The highest margins are indeed in USCA market and may be dependent on the types of products that are more popular in that market.

```
> pm_by_market
# A tibble: 5 x 2
  Market       ave_pm
  <fct>         <dbl>
1 USCA           2.95
2 Europe         2.93
3 Africa         2.67
4 Asia Pacific   2.53
5 LATAM          2.48
```

The product category with the highest margins are Technology products.

```
> pm_by_cat
# A tibble: 3 x 2
  Category        ave_pm
  <fct>            <dbl>
1 Technology        4.26
2 Furniture         3.25
3 Office Supplies   2.04
```

However, the distribution of product orders in USCA does not really reflect this.




```
> sales_by_grp
# A tibble: 15 x 3
# Groups:   Market [5]
   Market       Category        total_sales
   <fct>        <fct>                 <dbl>
 1 Africa       Furniture           6379573
 2 Africa       Office Supplies    29770607
 3 Africa       Technology          8803558
 4 Asia Pacific Furniture          29012525
 5 Asia Pacific Office Supplies    84392643
 6 Asia Pacific Technology         30085462
 7 Europe       Furniture          17598036
 8 Europe       Office Supplies    78655486
 9 Europe       Technology         22628773
10 LATAM        Furniture          24767034
11 LATAM        Office Supplies    60274855
12 LATAM        Technology         20939087
13 USCA         Furniture          21690751
14 USCA         Office Supplies    61571842
15 USCA         Technology         19509333
```

Problem 2: Inventory Management

Retailers that depend on seasonal shoppers have a particularly challenging job when it comes to inventory management. Your manager is making plans for next year's inventory.

He wants you to answer the following questions:

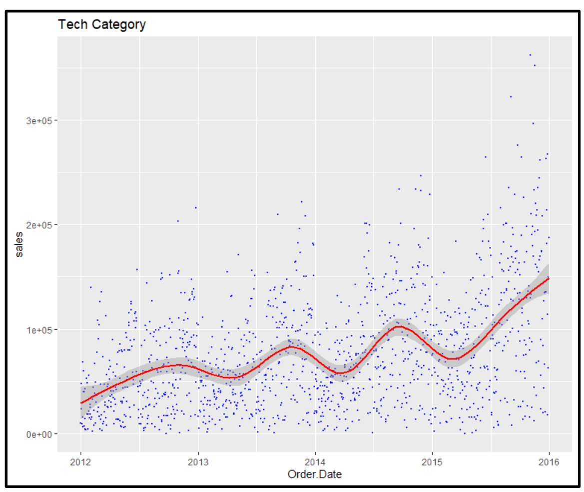
1. Is there any seasonal sales trend in your company?
2. Is there any seasonal trend of **different categories** of products?

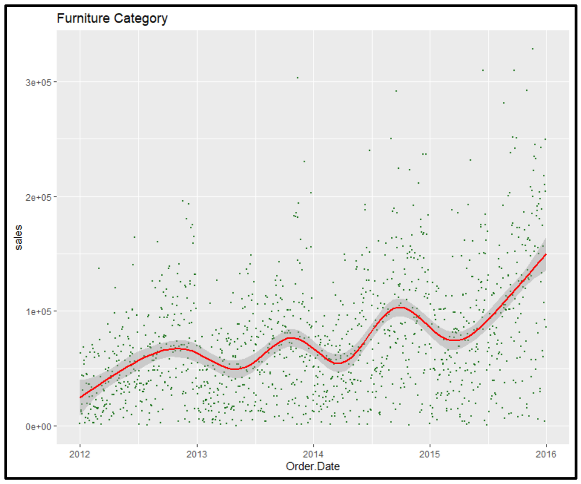
Note: Each order has a column called Quantity.

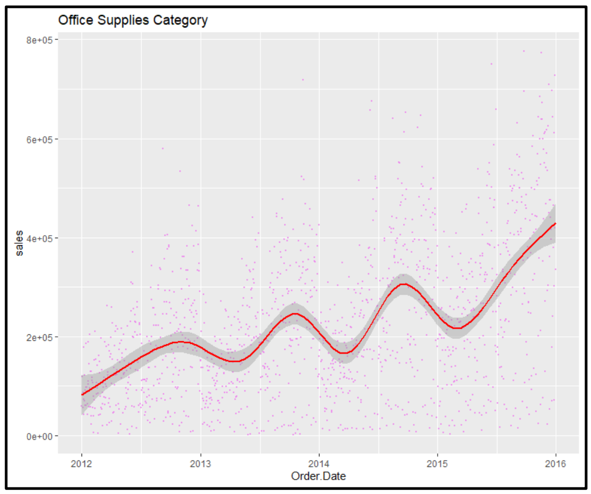
```
# Fill in your code here


#First converted dates to Date Format (for both order and shipping dates).
orders$Order.Date=as.Date(orders$Order.Date,"%m/%d/%y")
orders$Ship.Date=as.Date(orders$Order.Date,"%m/%d/%y")
#From class "Factor" to class "Date"

#sum sales by Order.Date then plot
by_date=group_by(orders, Order.Date)
sales_by_date=summarise(by_date, sales=sum(Sales))
g=ggplot(sales_by_date, aes(x=Order.Date,y=sales))
g+geom_point(alpha=0.7,color="orange")+geom_smooth(method=lm)

#check if there is seasonality per category
tech=filter(orders, Category=="Technology")
by_date=group_by(tech, Order.Date)
sales_by_date=summarise(by_date, sales=sum(Sales))
g=ggplot(sales_by_date, aes(x=Order.Date,y=sales))
g+geom_point(size=0.5,alpha=0.7,color="blue")+geom_smooth(color="red")+ggtitl
e("Tech Category")

furniture=filter(orders, Category=="Furniture")
by_date=group_by(furniture, Order.Date)
sales_by_date=summarise(by_date, sales=sum(Sales))
g=ggplot(sales_by_date, aes(x=Order.Date,y=sales))
g+geom_point(size=0.5,alpha=0.7,color="darkgreen")+geom_smooth(color="red")+g
gtitle("Furniture Category")

office_supplies=filter(orders, Category=="Office Supplies")
by_date=group_by(office_supplies, Order.Date)
sales_by_date=summarise(by_date, sales=sum(Sales))
g=ggplot(sales_by_date, aes(x=Order.Date,y=sales))
g+geom_point(size=0.5,alpha=0.7,color="violet")+geom_smooth(colo r="red")+ggti
tle("Office Supplies Category")
```

Sales has trended upwards throughout the years and appears to have seasonality with peaks towards the end of the year and decline come start of the year.

The same seasonal trend can be observed for each product category.







Problem 3: Why did customers make returns?

Your manager required you to give a brief report (**Plots + Interpretations**) on returned orders from the **Returns** dataset.

1. How much profit did we lose for each year?
2. How many customer returned more than once? more than 10 times?
3. Which regions are more likely to return orders?
4. Which categories (sub-categories) of products are more likely to be returned?

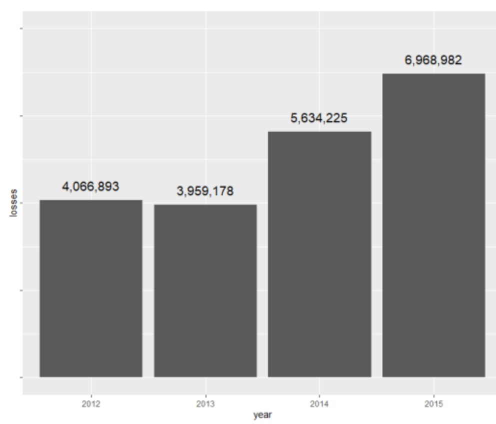
Hint:

1. Import **Returns.csv**
2. Merge the **Returns** dataframe you imported with the **Orders** dataframe.

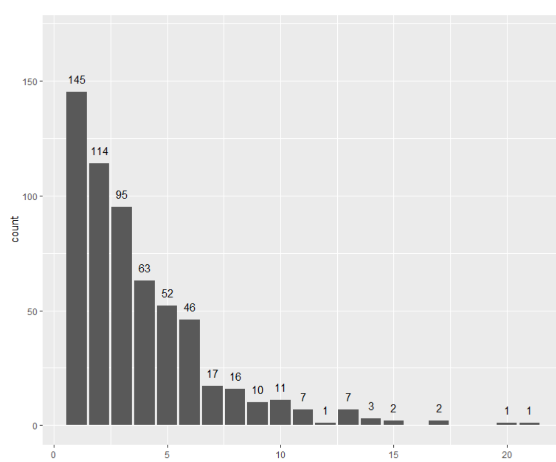
```r
# Fill in your code here

#merge returns and orders files by Order.ID
joined=inner_join(orders,returns,by="Order.ID")

#compute profit losses
sum(joined$Profit)

#add year to compute losses per year
joined$year=format(joined$Order.Date,"%Y")
by_year=group_by(joined,year)
losses_by_year=summarise(by_year, losses=sum(Profit))
losses_by_year

#plot losses by year
ggplot(losses_by_year,aes(x=year,y=losses))+geom_col()+geom_text(aes(label=fo
rmat(losses,big.mark=",")),vjust=-1,size=4)+theme(axis.text.y=element_blank()
)+ylim(0,8000000)


#how many customers return more than once?
joined$count=1
by_customer=group_by(joined,Customer.Name)
summary1=summarise(by_customer,reps=sum(count))
max(summary1$reps)
ggplot(summary1, aes(x=reps))+geom_bar()+geom_text(stat='count',aes(label=..c
ount..),vjust=-1)+ylim(0,170)

#more than once
summary2=summary1[summary1$reps>1,]
nrow(summary2)

#more than 10x
summary3=summary1[summary1$reps>10,]
nrow(summary3)

#region-level analysis
by_region=group_by(joined,Region.x)
by_region_count=summarise(by_region,reps=sum(count))
by_region_count=by_region_count[order(-by_region_count$reps),]
top10=by_region_count[1:10,]
ggplot(top10,aes(x=Region.x,y=reps))+geom_col()
ggplot(top10,aes(x=reorder(Region.x,-reps),y=reps))+geom_col()+theme(text = e
lement_text(size=12),axis.text.x = element_text(angle=90, hjust=1))+xlab("Reg
ion")+ylab("# of Customers w/ Returns")

#categories most likely to be returned
by_cat=group_by(joined,Category)
by_cat_count=summarise(by_cat,reps=sum(count))
ggplot(by_cat_count,aes(x=Category,y=reps))+geom_col()+ylab("# of Returns")

#filter for top Regions
by_cat=joined %>% select('Region.x','Category','Sub.Category','count') %>% fi
lter(Region.x %in% c('Central America','Western Europe','Western US'))
ggplot(by_cat, aes(x=Region.x))+geom_bar(aes(fill=Category))+xlab("Region")+y
lab("# of Returns")
```

```
#sub-categories most likely to be returned
ggplot(by_cat, aes(x=Region.x))+geom_bar(aes(fill=Sub.Category),position="fil
l")+xlab("Region")+ylab("# of Returns")

by_subcat=group_by(joined,Sub.Category)
by_subcat_count=summarise(by_subcat,reps=sum(count))
by_subcat_count=by_subcat_count[order(-by_subcat_count$reps),]
top10_subcat=by_subcat_count[1:10,]
ggplot(top10_subcat,aes(x=reorder(Sub.Category,-reps),y=reps))+geom_col()+xla
b("SubCategory")+ylab("# of Returns")+theme(axis.text.x=element_text(angle=90
))
top10_subcat

g=ggplot(by_cat,aes(Region.x, Sub.Category))
g+geom_count(alpha=0.5,color='red')+scale_size_area(max_size=10)
```

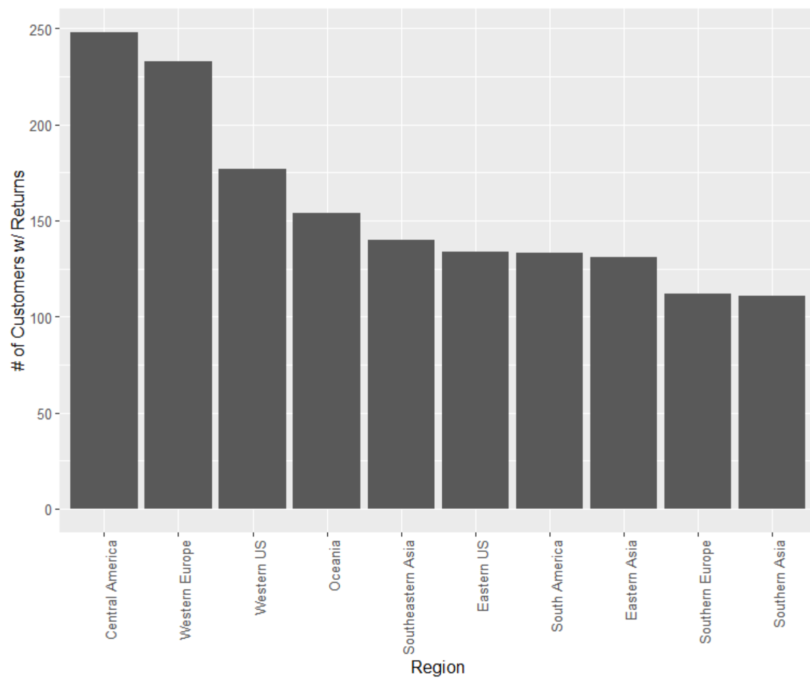
Total profit losses from returns are $20,629,278 from 2012 to 2015, and losses have been increasing annually.



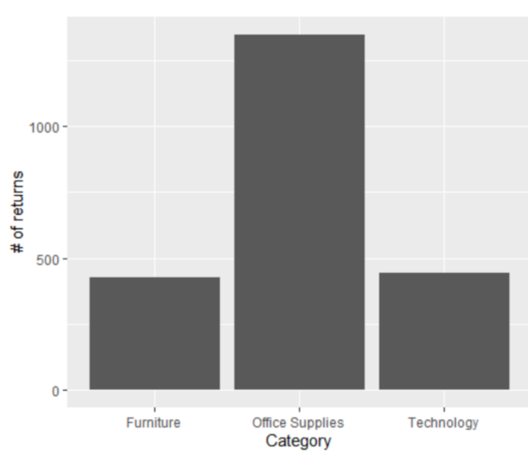
The most number of times a customer has returned an item is 21x from 2012-2015. This is an alarming frequency. However, majority of returns are 5x and below. There are 448 customers that have returned an item more than once, and there are 24 customers that have returned an item more than 10x

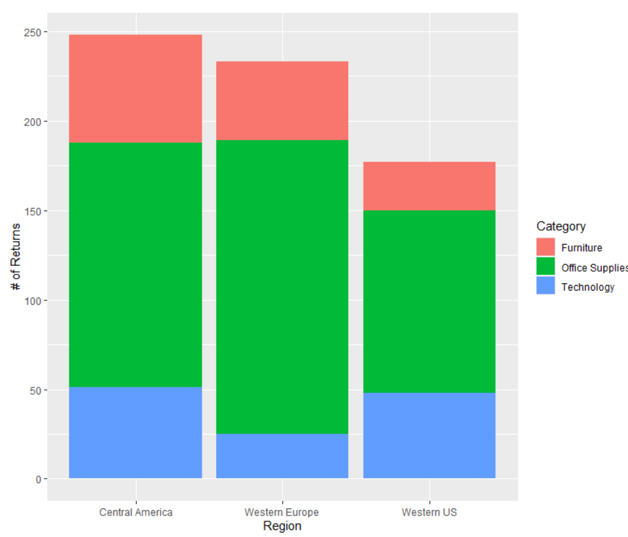


The following regions are the most likely to have return orders: Central America, Western Europe, and Western US. As such an investigation on the products or delivery handling will have to be done.

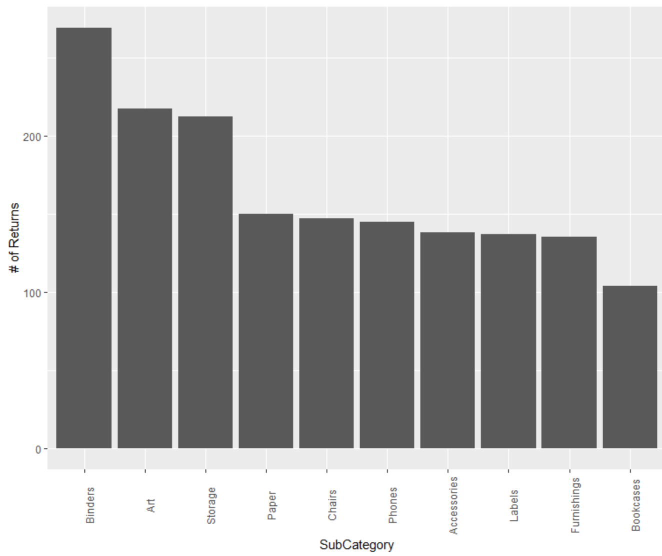


For categories, office supplies experience the most returns. This is the case for the regions with the most returns.

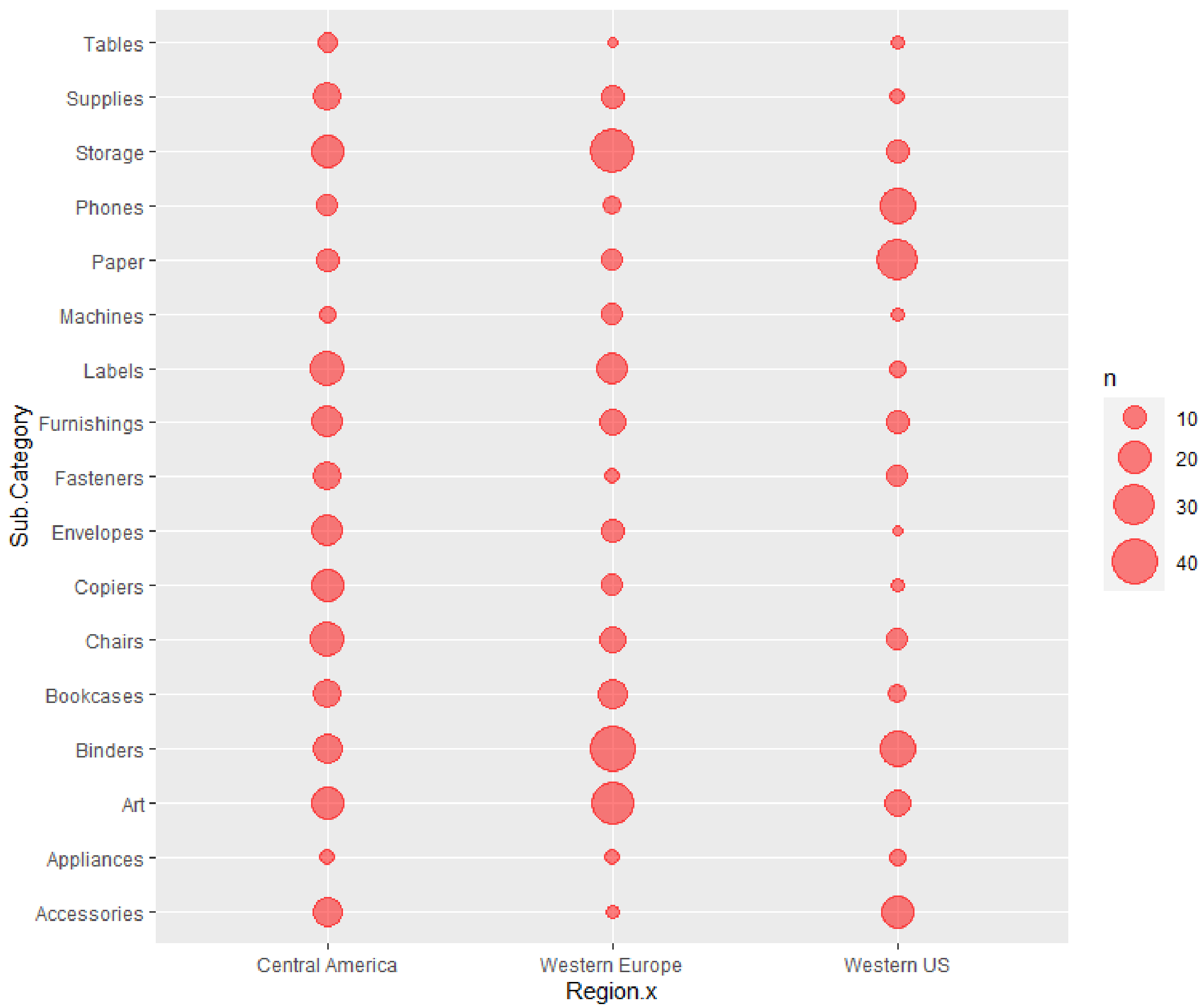




Drilling in further to sub-categories, "Binder", "Art", and "Storage" are the top 2 returned items. On a regional level, it is "Western Europe" that returns most of these types of items. There may be an issue with the supplier of these types of materials there.



# of Returns
200
100
0
Binders
Art
Storage
Paper
Chairs
Phones
Accessories
Labels
Furnishings
Bookcases
SubCategory



Sub.Category
Tables
Supplies
Storage
Phones
Paper
Machines
Labels
Furnishings
Fasteners
Envelopes
Copiers
Chairs
Bookcases
Binders
Art
Appliances
Accessories
Central America
Western Europe
Western US
Region.x
n
10
20
30
40