Task 3. Exploratory analysis - Retail

Try to find weak areas where you can make more profit.

What business problems can you derive by exploring the data?

Download packages/libraries

```
library(readr)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(magrittr)
```

Download data

```
retail <- read.csv("/Users/macbookair/Desktop/GRIP/Task3/SampleSuperstore.csv")
summary(retail)
```

```
##   Ship.Mode          Segment           Country             City          
##  Length:9994        Length:9994        Length:9994        Length:9994       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##     State            Postal.Code       Region            Category        
##  Length:9994        Min.   : 1040   Length:9994        Length:9994       
##  Class :character   1st Qu.:23223   Class :character   Class :character  
##  Mode  :character   Median :56430   Mode  :character   Mode  :character  
##                     Mean   :55190                                        
##                     3rd Qu.:90008                                        
##                     Max.   :99301                                        
##  Sub.Category           Sales              Quantity        Discount     
##  Length:9994        Min.   :    0.444   Min.   : 1.00   Min.   :0.0000  
##  Class :character   1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.:0.0000  
##  Mode  :character   Median :   54.490   Median : 3.00   Median :0.2000  
##                     Mean   :  229.858   Mean   : 3.79   Mean   :0.1562  
##                     3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:0.2000  
##                     Max.   :22638.480   Max.   :14.00   Max.   :0.8000  
##      Profit         
##  Min.   :-6599.978  
##  1st Qu.:    1.729  
##  Median :    8.666  
##  Mean   :   28.657  
##  3rd Qu.:   29.364  
##  Max.   : 8399.976  
```

1. Data cleaning and preparation

1.1 Check missing values

```
retail.na <- is.na(retail)
summary(retail.na)
```

```
##  Ship.Mode        Segment          Country           City        
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical  
##  FALSE:9994      FALSE:9994      FALSE:9994      FALSE:9994     
##    State         Postal.Code       Region          Category      
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical  
##  FALSE:9994      FALSE:9994      FALSE:9994      FALSE:9994     
##  Sub.Category      Sales          Quantity        Discount      
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical  
##  FALSE:9994      FALSE:9994      FALSE:9994      FALSE:9994     
##    Profit       
##  Mode :logical  
##  FALSE:9994     
```

1.2 Check duplicated values

```
duplicated.retail <- duplicated(retail)
head(duplicated.retail)
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
```

There are 17 cases of duplicated values.

1.3 Drop duplicated values

```
retail.unique <- retail[!duplicated(retail), ]
```

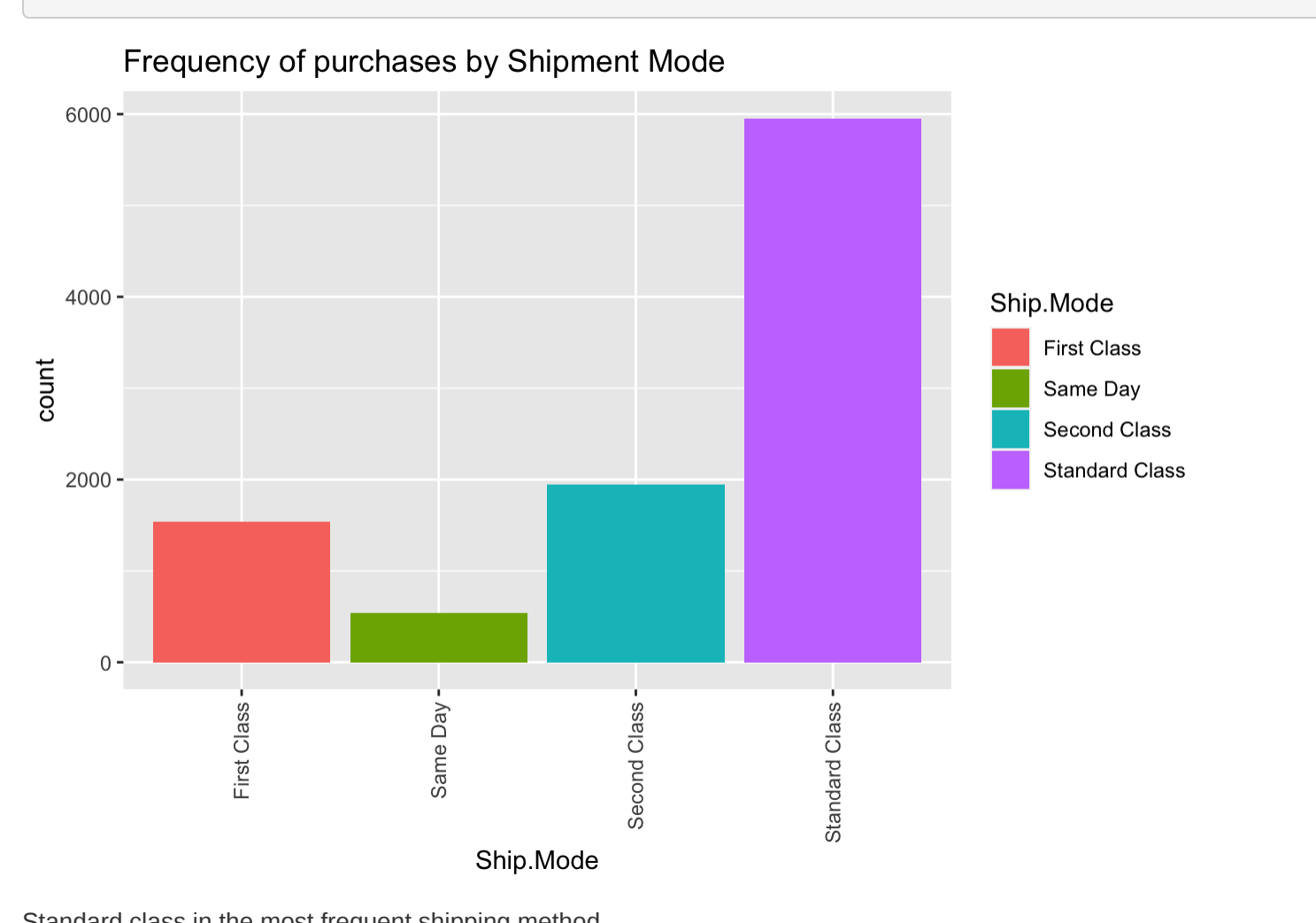
1.4 Remove the country and postal code columns

```
retail.unique <- retail.unique %>% select(-c(Country, Postal.Code))
```

2. Exploratory data analysis (EDA)

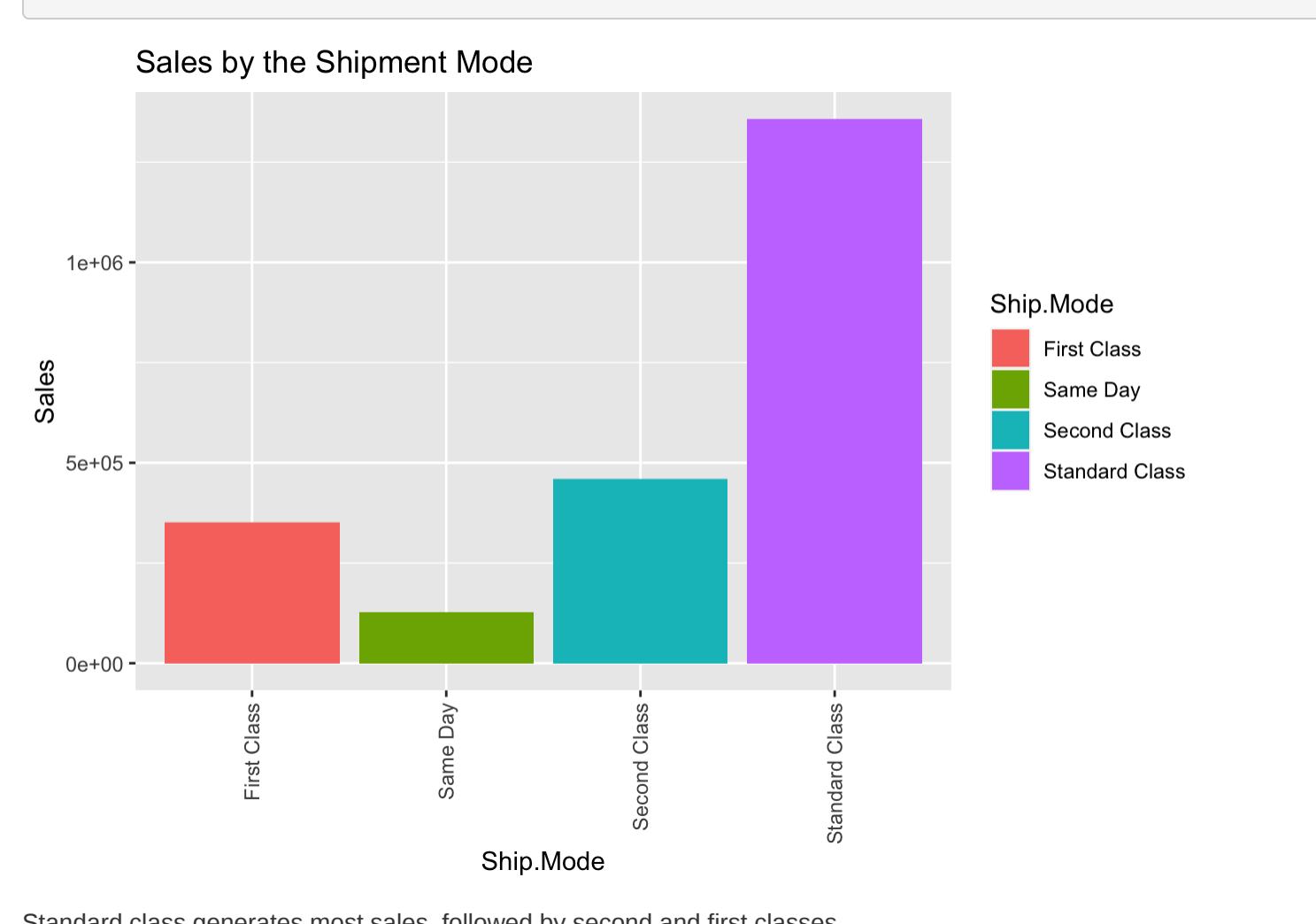
2.1 Shipment mode analysis

```
ggplot(retail.unique, aes(Ship.Mode, fill = Ship.Mode)) +
  geom_bar()+
  ggtitle("Frequency of purchases by Shipment Mode")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Standard class in the most frequent shipping method.

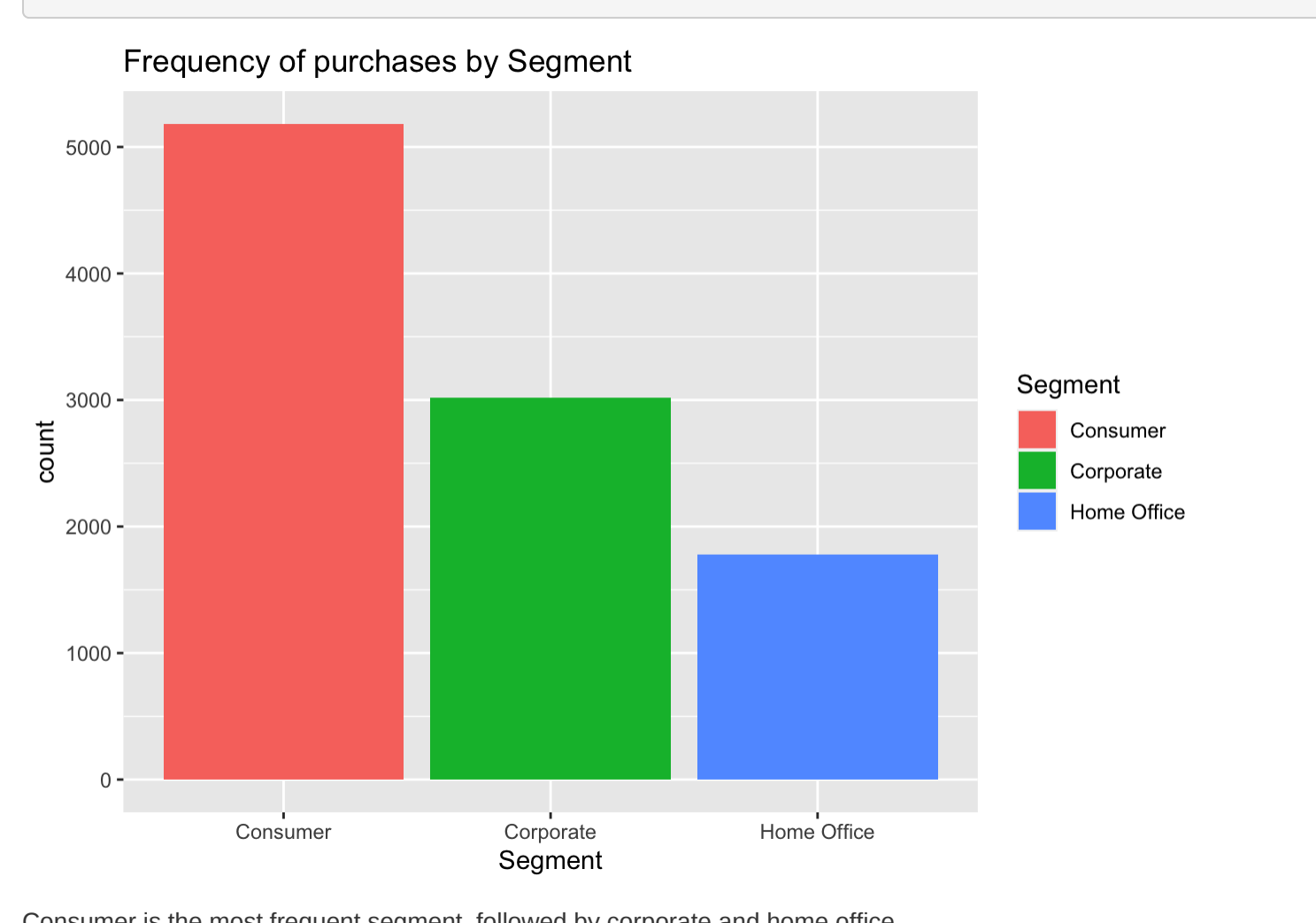
```
ggplot(retail.unique, aes(Ship.Mode, Sales, fill = Ship.Mode)) +
  geom_col() +
  ggtitle("Sales by the Shipment Mode")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Standard class generates most sales, followed by second and first classes.

2.2 Segment analysis

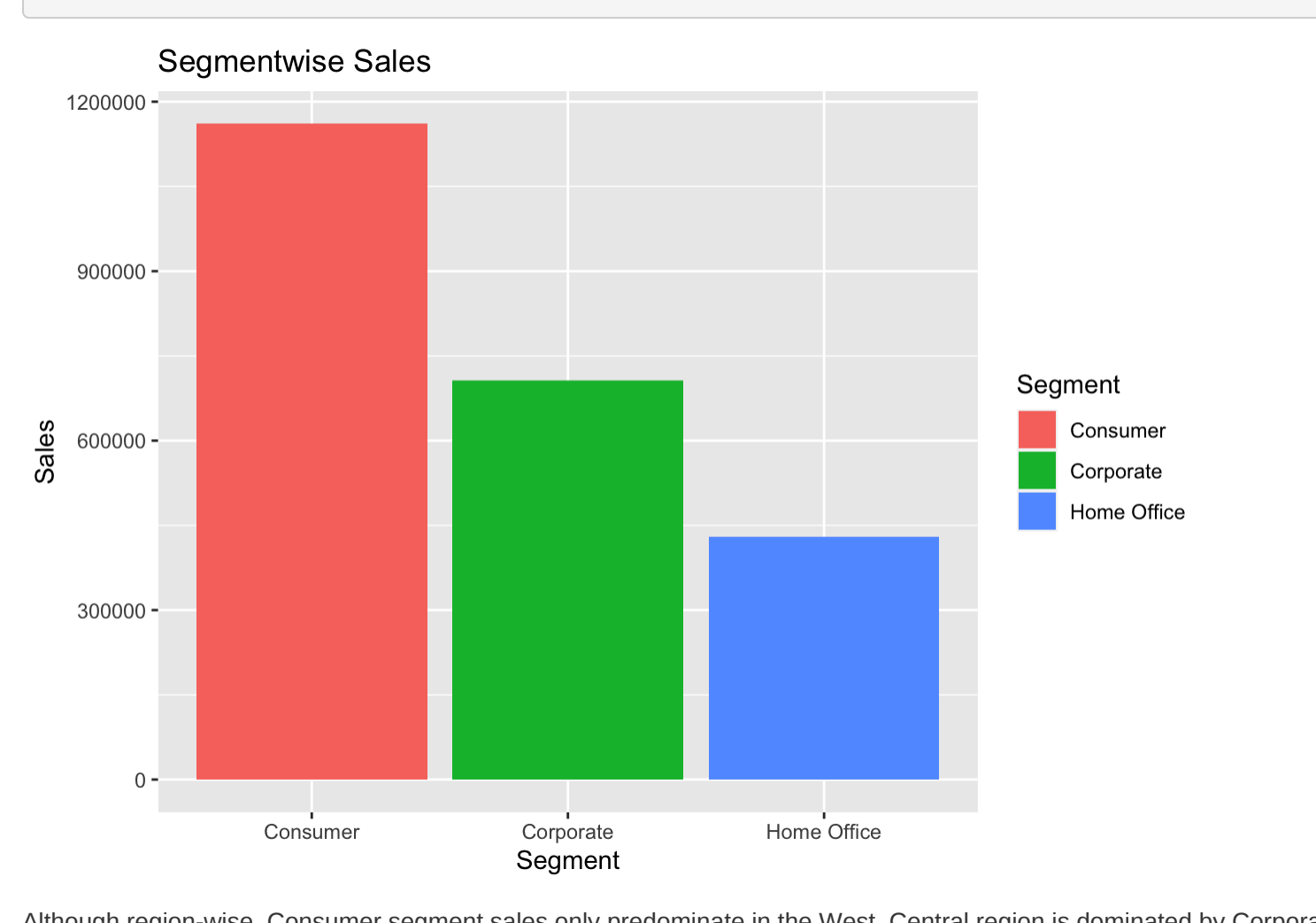
```
ggplot(retail.unique, aes(Segment, fill = Segment)) +
  geom_bar()+
  ggtitle("Frequency of purchases by Segment")
```



Consumer is the most frequent segment, followed by corporate and home office.

Home office segment is the weak area, let's look if sales-wise it is the same.

```
ggplot(retail.unique,aes(Segment,Sales, fill = Segment)) +
  geom_col()+
  ggtitle("Segmentwise Sales")
```



Although region-wise, Consumer segment sales only predominate in the West. Central region is dominated by Corporate segment, East and South by Home Office.

When approaching Segment-wise sales, focus on the region.

2.3 City analysis

Identify cities with most frequent sales.

```
city_freq <- table(retail.unique$City)
sorted_freq <- sort(city_freq, decreasing = TRUE)
top_cities <- names(head(sorted_freq, 10))
print(top_cities)
```

```
##  [1] "New York City" "Los Angeles"   "Philadelphia"  "San Francisco"
##  [5] "Seattle"       "Houston"       "Chicago"       "Columbus"     
##  [9] "San Diego"     "Springfield"  
```

New York City, LA, Philadelphia, San Francisco, Seattle, Houston, Chicago, Columbus, San Diego and Springfield are the top cities sales-wise.

2.4 State analysis

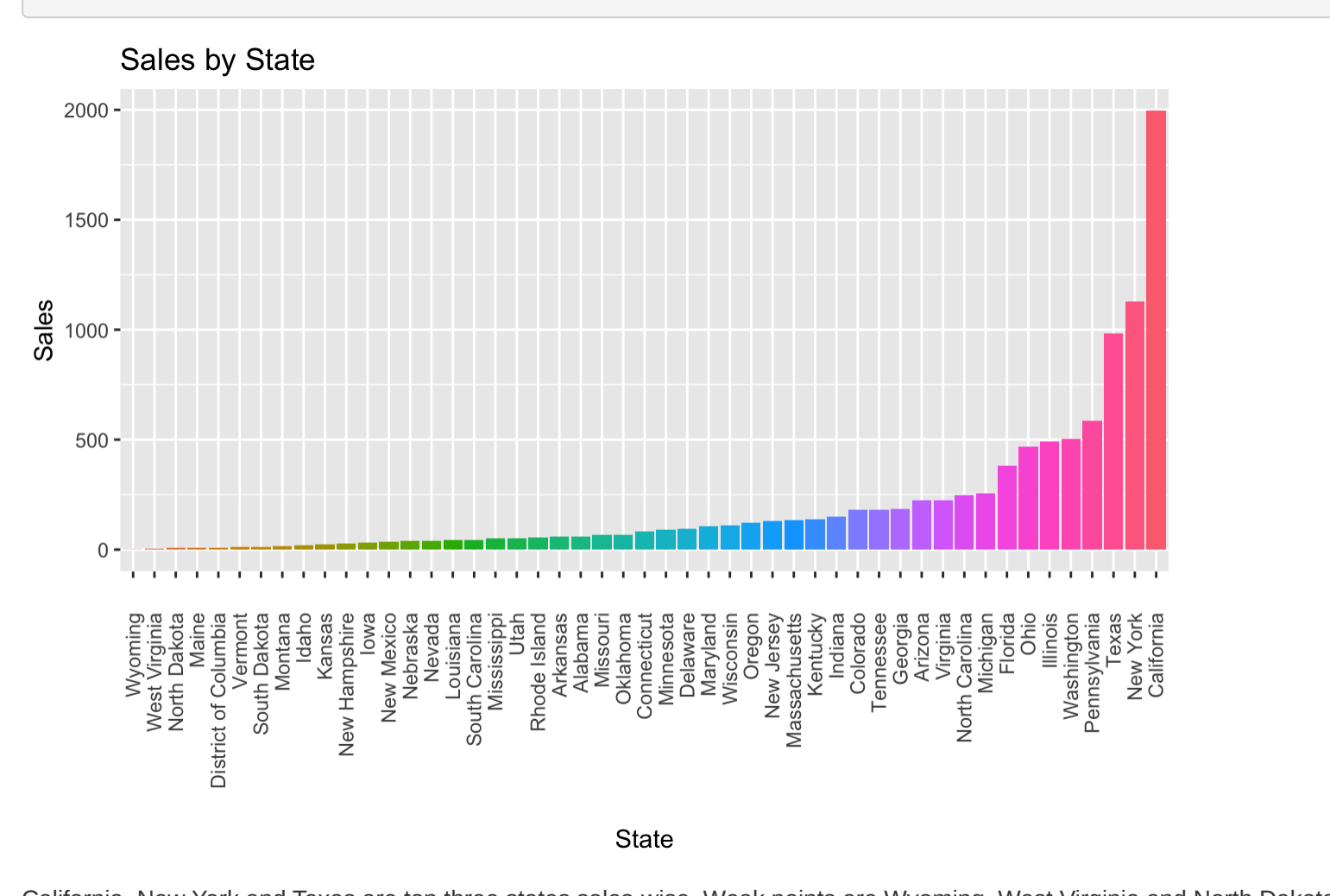
Calculate the frequency of each state

```
state_frequency <- count(retail.unique, State)
```

Reorder the levels of the State factor based on the frequency of Sales

```
retail.unique$State <- factor(retail.unique$State, levels = state_frequency$State[order(state_frequency$n)])
```

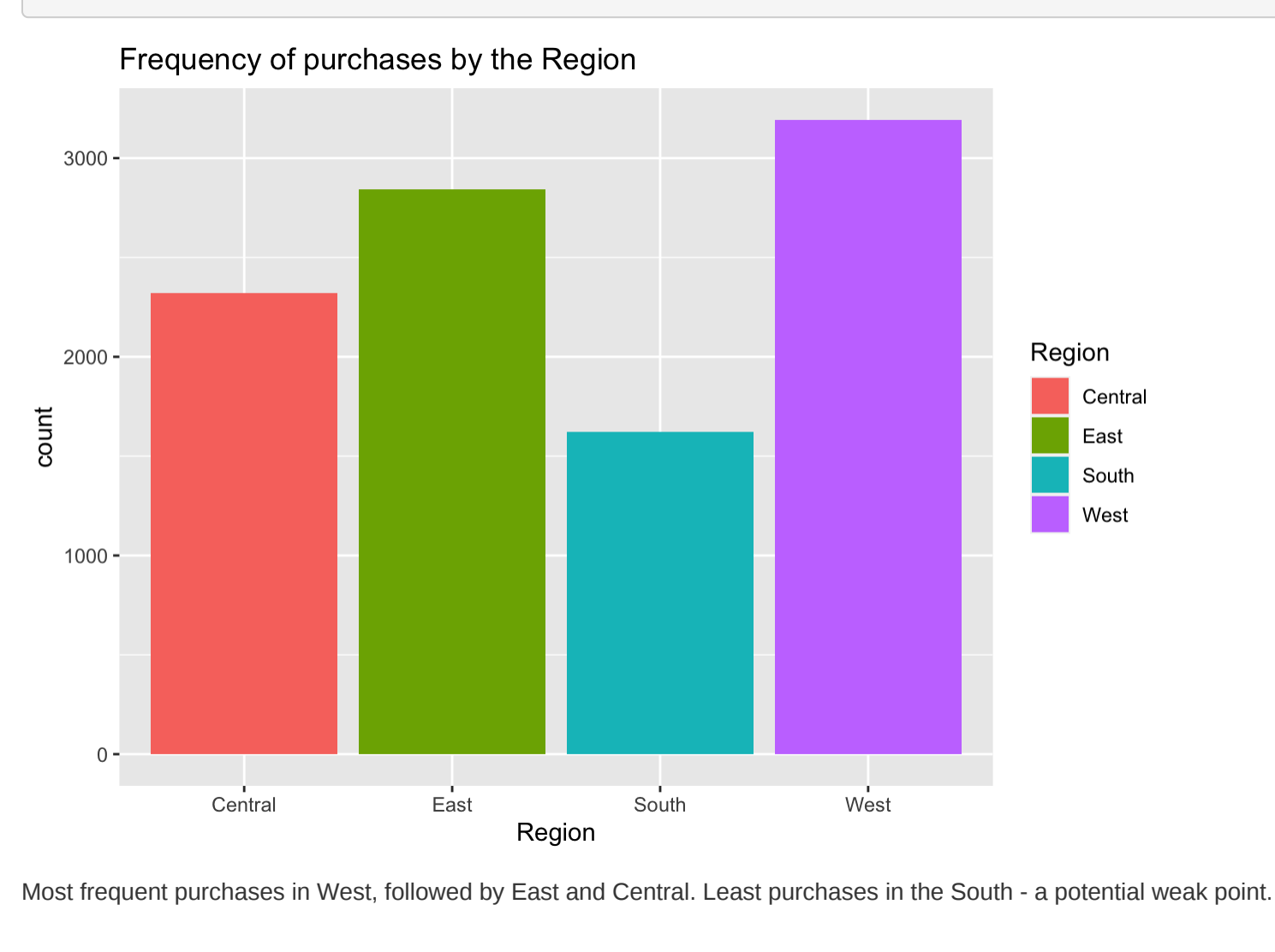
```
ggplot(retail.unique, aes(State, fill = State)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, margin = margin(15, 0, 15, 0)),
        legend.position = "none")+
  xlab("State") +
  ylab("Sales") +
  ggtitle("Sales by State")
```



California, New York and Texas are top three states sales-wise. Weak points are Wyoming, West Virginia and North Dakota. Cannot conclude a relationship between most profitable states and regions, as all of these are located in different parts of the country.

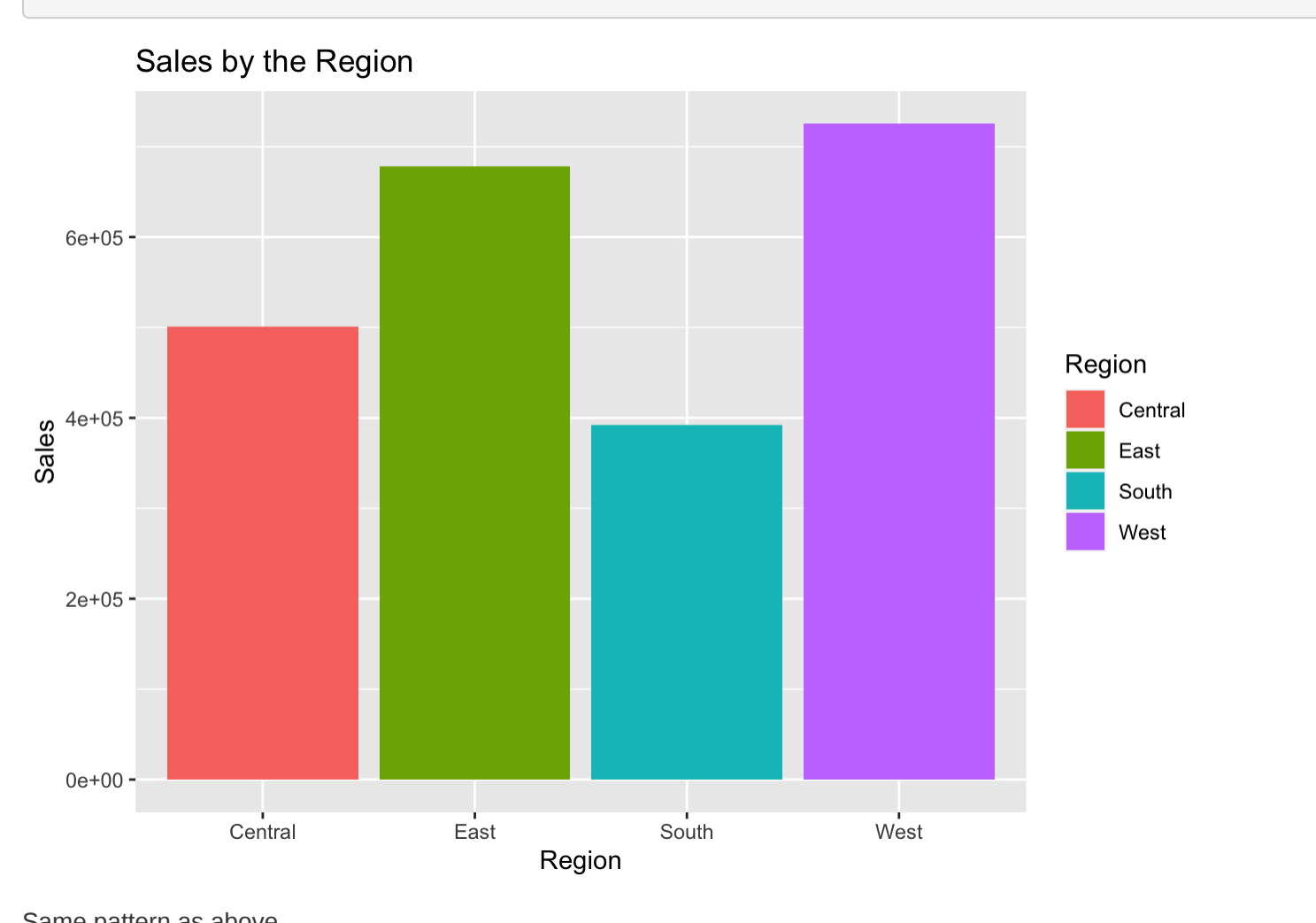
2.5 Region analysis

```
ggplot(retail.unique, aes(Region, fill = Region))+
  geom_bar()+
  ggtitle("Frequency of purchases by the Region")
```



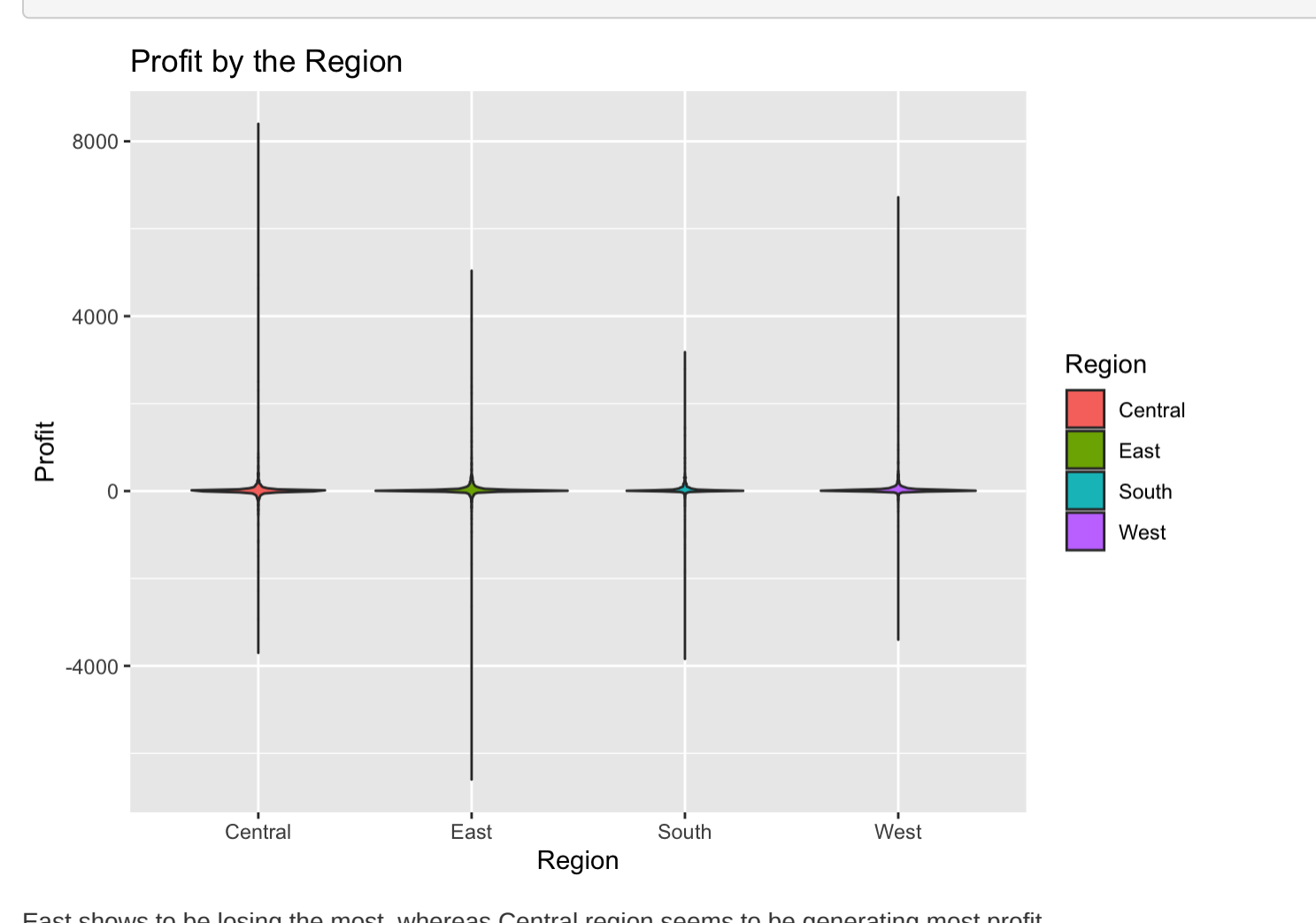
Most frequent purchases in West, followed by East and Central. Least purchases in the South - a potential weak point.

```
ggplot(retail.unique, aes(Region, Sales, fill = Region))+
  geom_col()+
  ggtitle("Sales by the Region")
```



Same pattern as above.

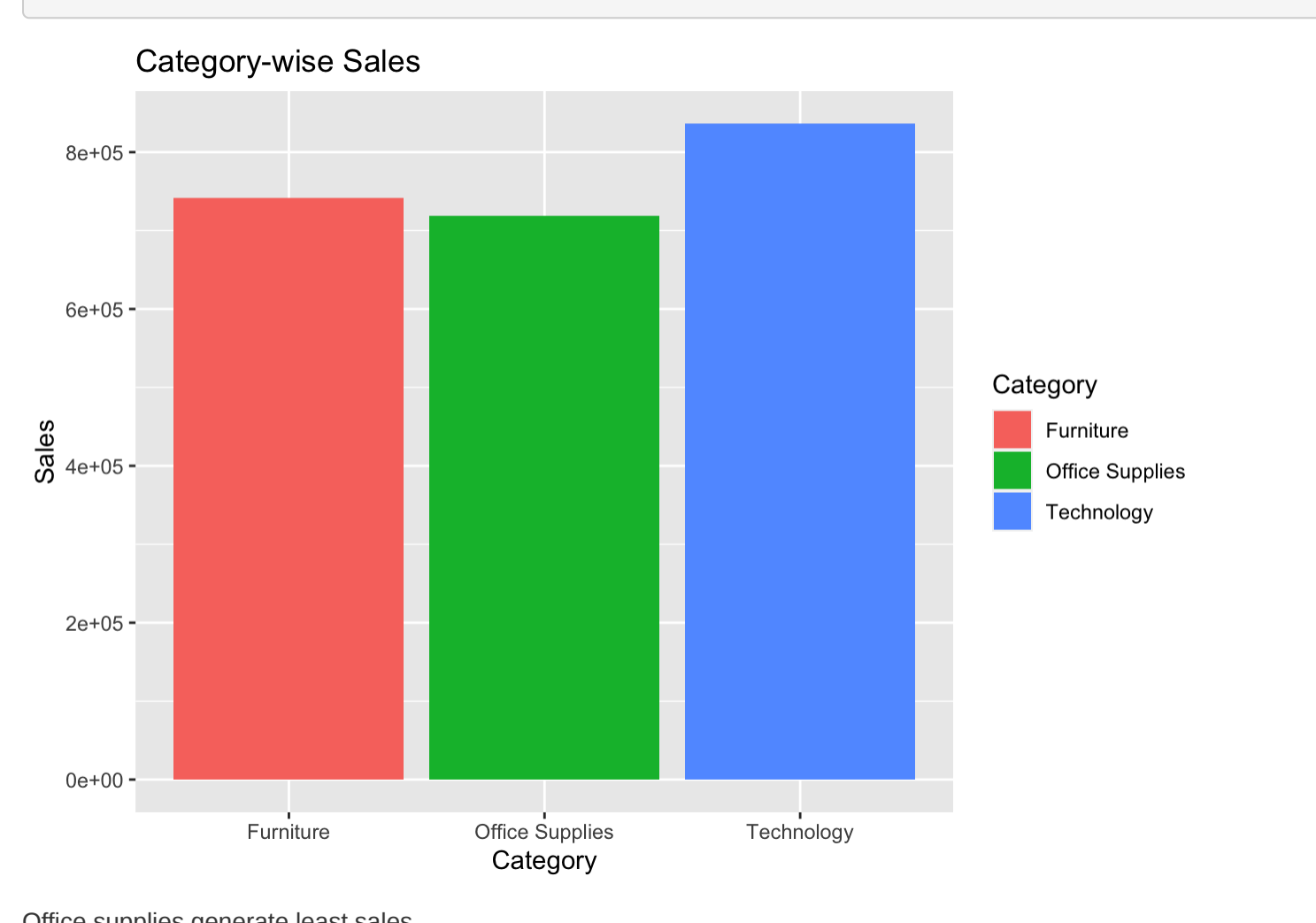
```
ggplot(retail.unique, aes(Region, Profit, fill = Region))+
  geom_violin()+
  ggtitle("Profit by the Region")
```



East shows to be losing the most, whereas Central region seems to be generating most profit.

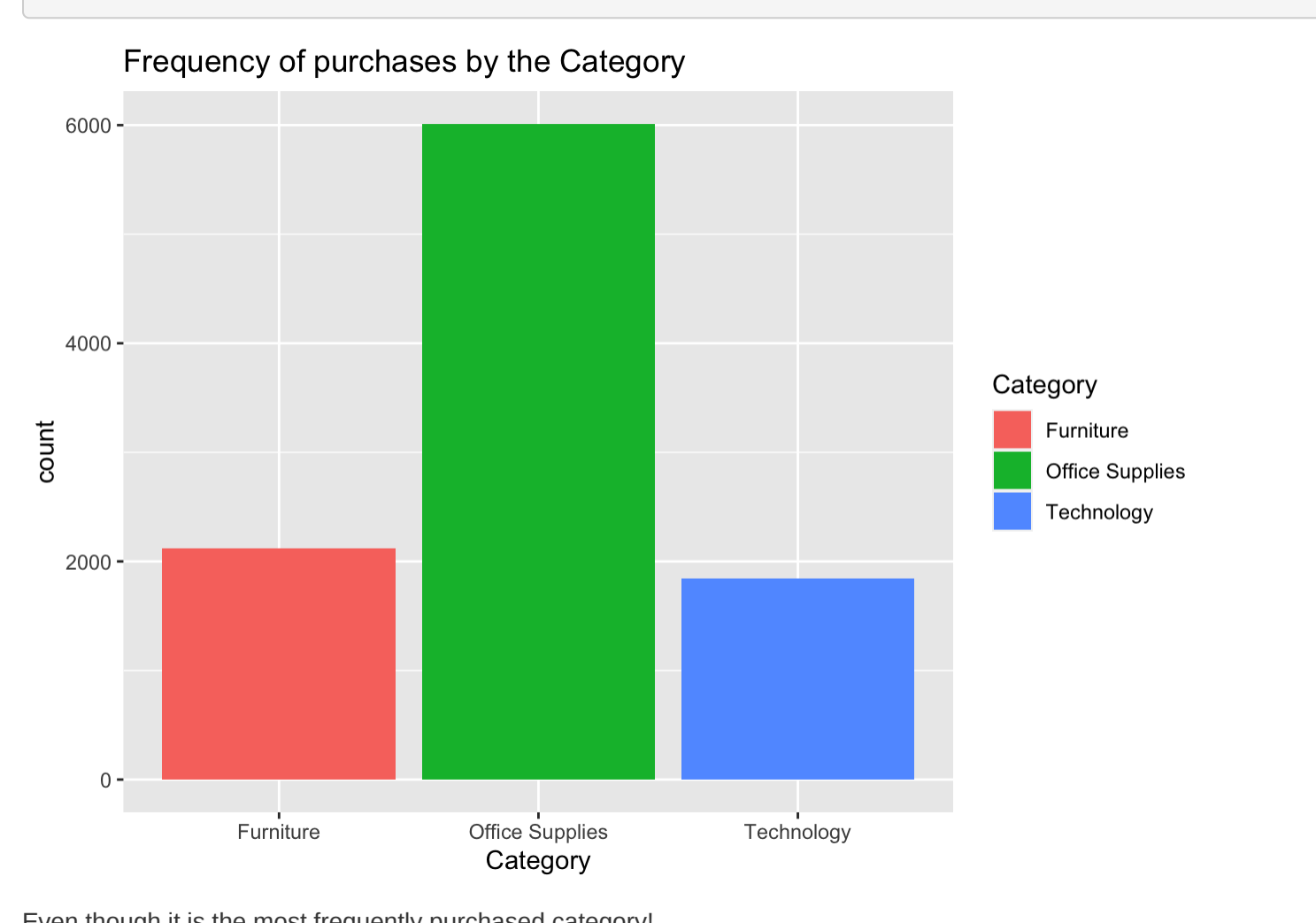
2.6 Category analysis

```
ggplot(retail.unique,aes(Category,Sales, fill = Category))+
  geom_col()+
  ggtitle("Category-wise Sales")
```



Office supplies generate least sales…

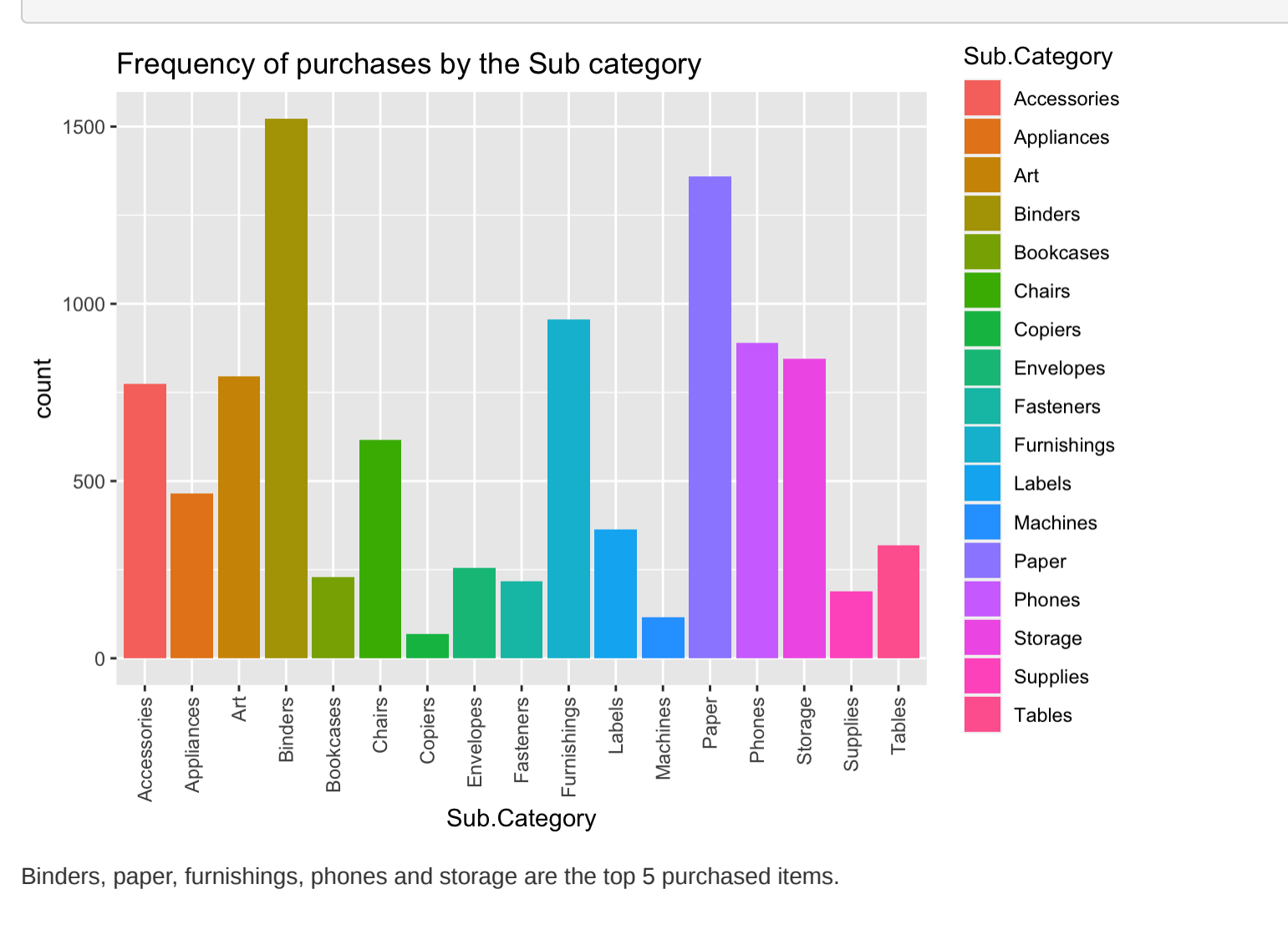
```
ggplot(retail.unique, aes(Category, fill = Category)) +
  geom_bar()+
  ggtitle("Frequency of purchases by the Category")
```



Even though it is the most frequently purchased category!

2.7 Sub category analysis

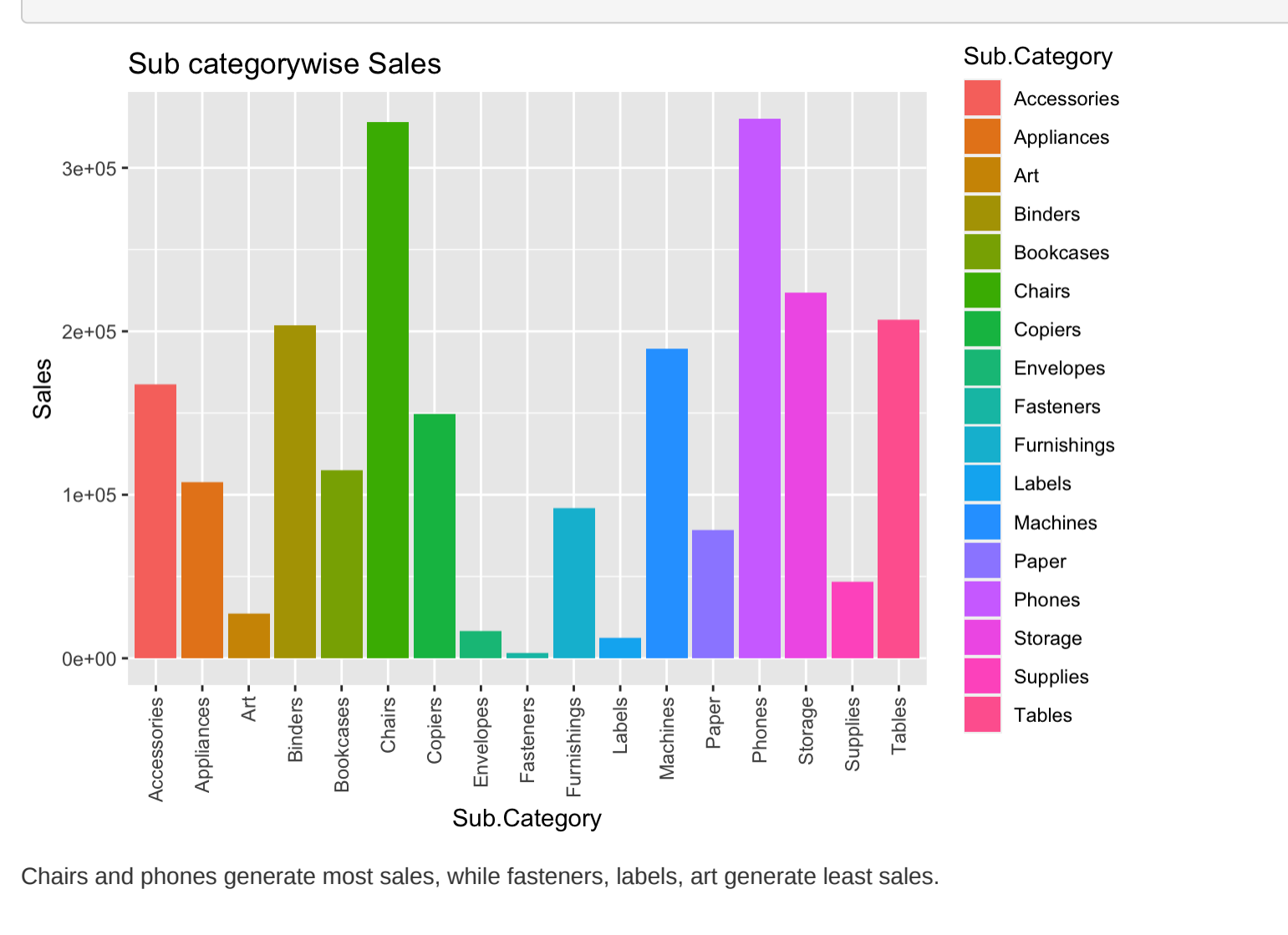
```
ggplot(retail.unique, aes(Sub.Category, fill = Sub.Category)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ggtitle("Frequency of purchases by the Sub category")
```



Binders, paper, furnishings, phones and storage are the top 5 purchased items.

Let's look at sales within sub categories.

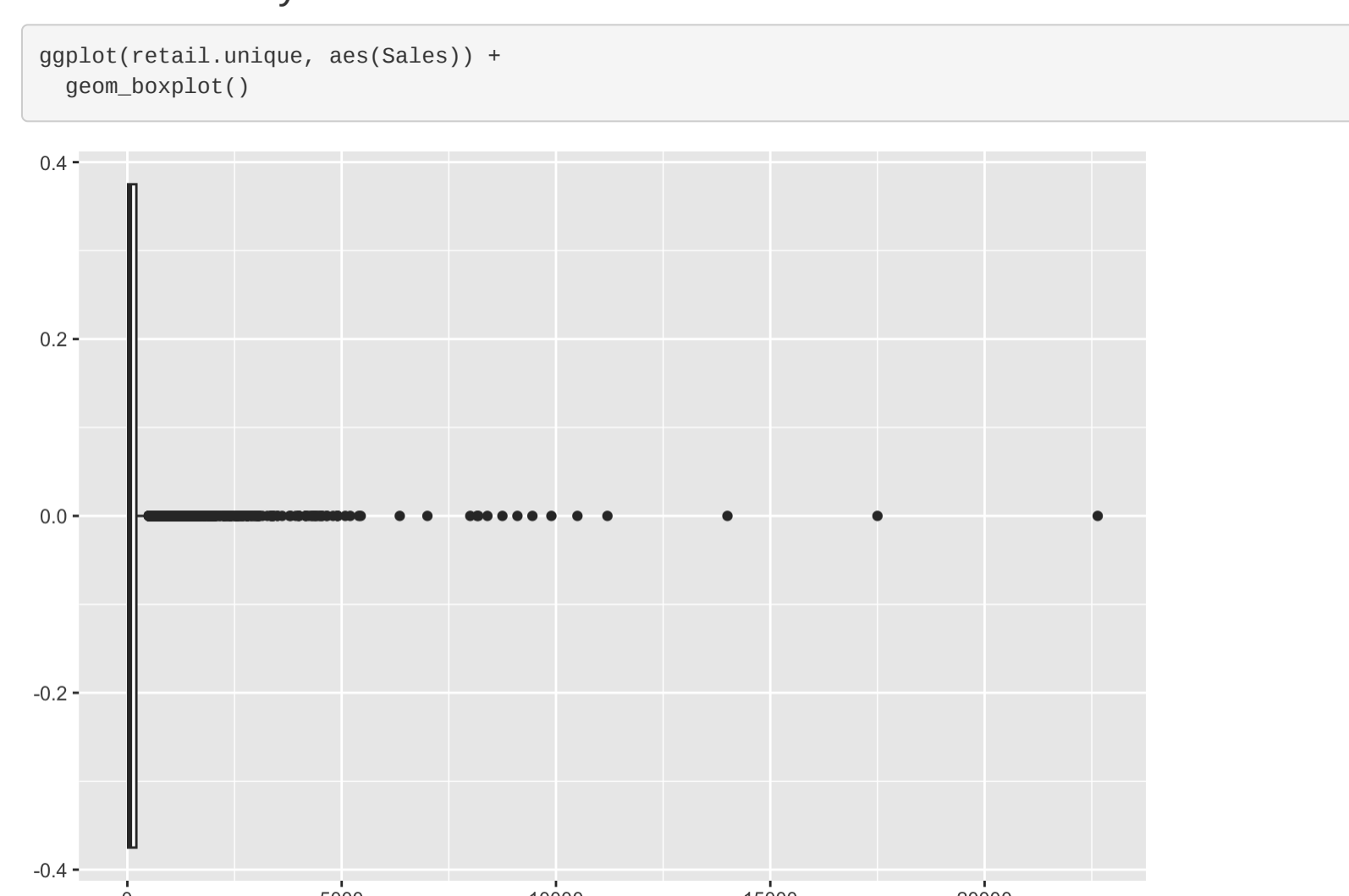
```
ggplot(retail.unique,aes(Sub.Category,Sales, fill = Sub.Category))+
  geom_col()+
  ggtitle("Sub categorywise Sales")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Chairs and phones generate most sales, while fasteners, labels, art generate least sales.

2.8 Sales analysis

```
ggplot(retail.unique, aes(Sales)) +
  geom_boxplot()
```



Sales has some abnormally high value.

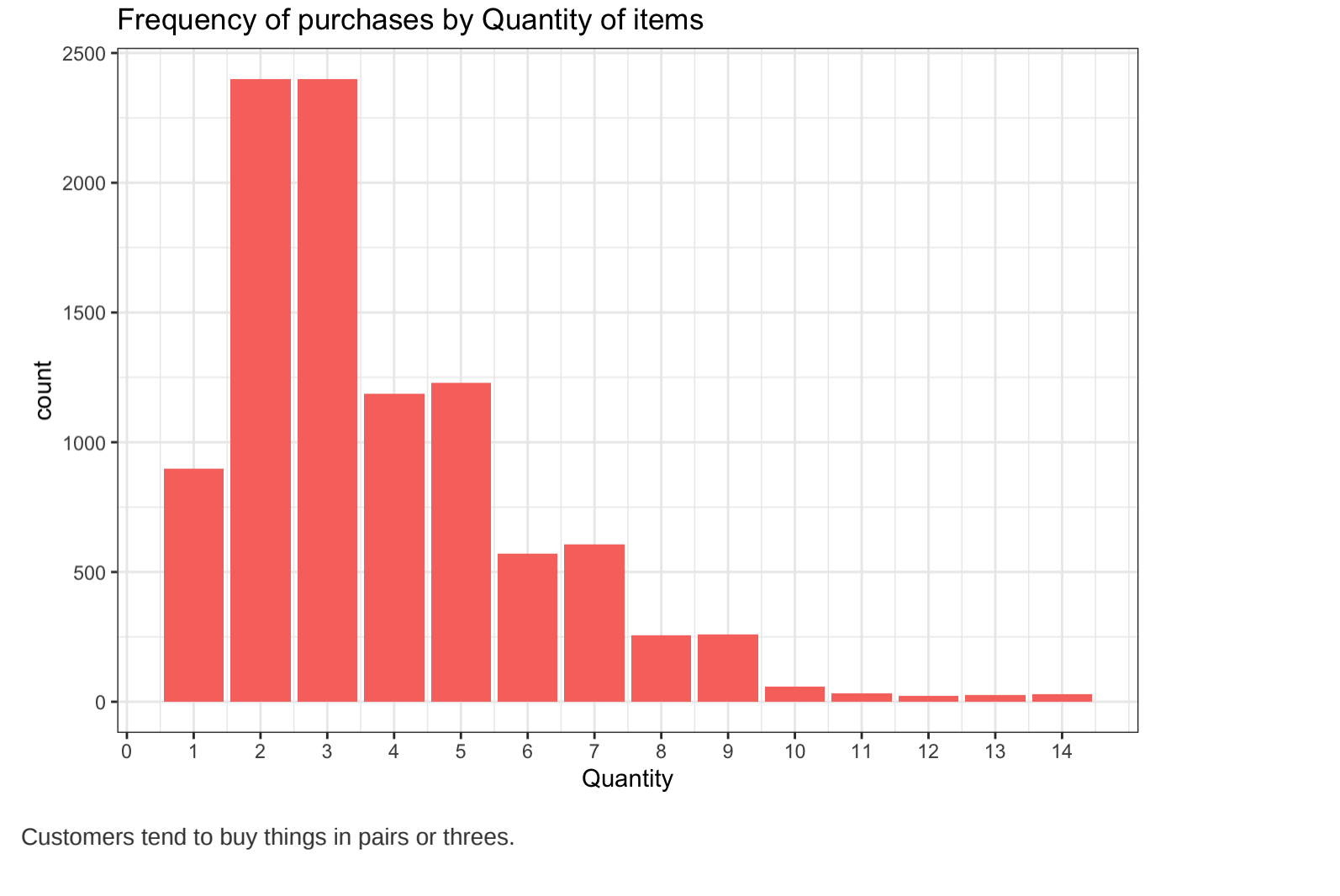
Replace that value with an average sales value.

```
maxSales <- max(retail.unique$Sales)
retail.unique$Sales <- replace(retail.unique$Sales, retail.unique$Sales==maxSales, mean(retail.unique$Sales))
summary(retail.unique$Sales)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##     0.444    17.300    54.816   227.903   209.970 17499.950 
```

2.9 Quantity analysis

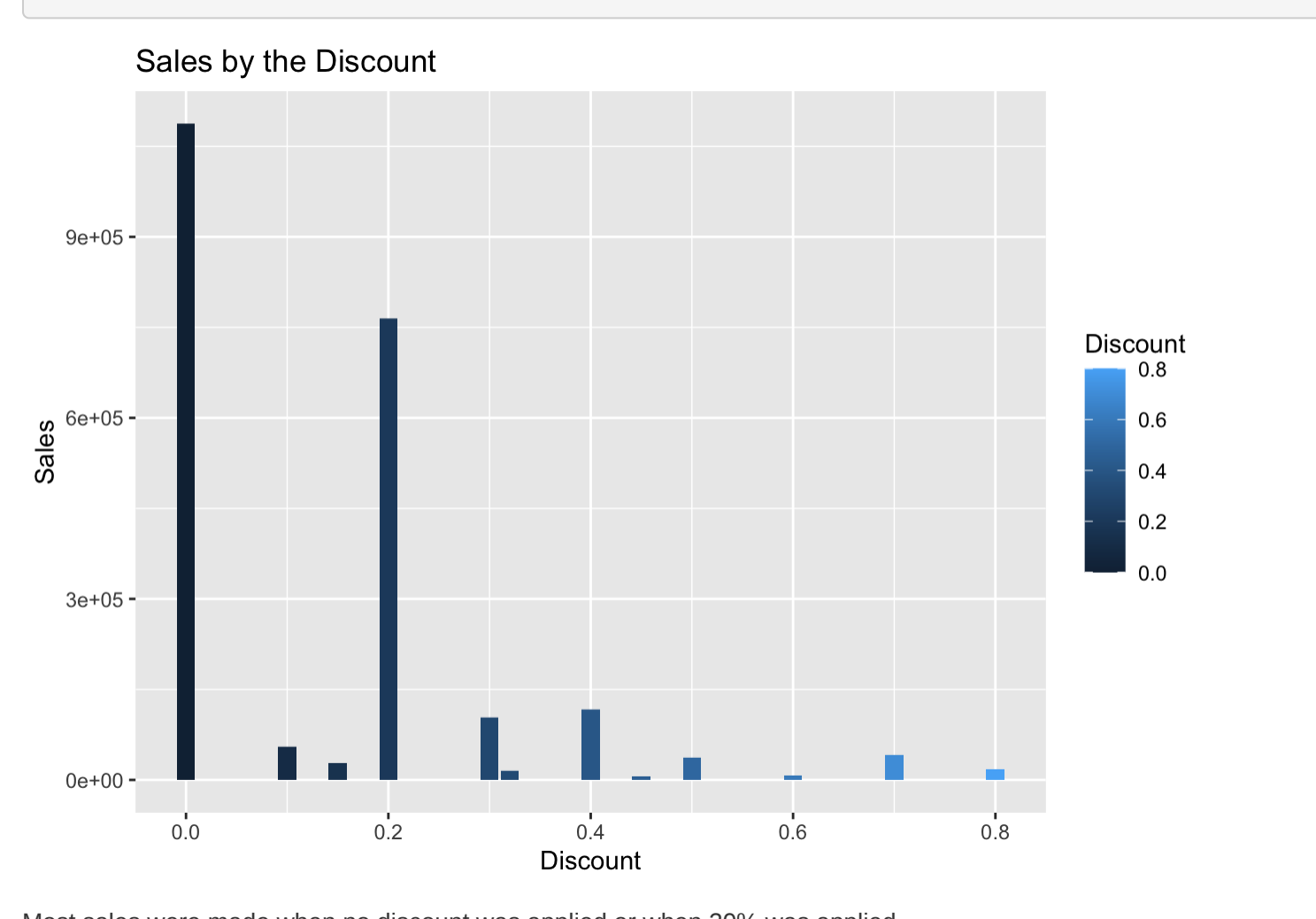
```
ggplot(retail.unique, aes(Quantity, fill = "pink"))+
  geom_bar()+
  scale_x_continuous(breaks = seq(0, max(retail.unique$Quantity), by = 1))+
  theme_bw()+
  ggtitle("Frequency of purchases by Quantity of items")+
  theme(legend.position = "none")
```



Customers tend to buy things in pairs or threes.

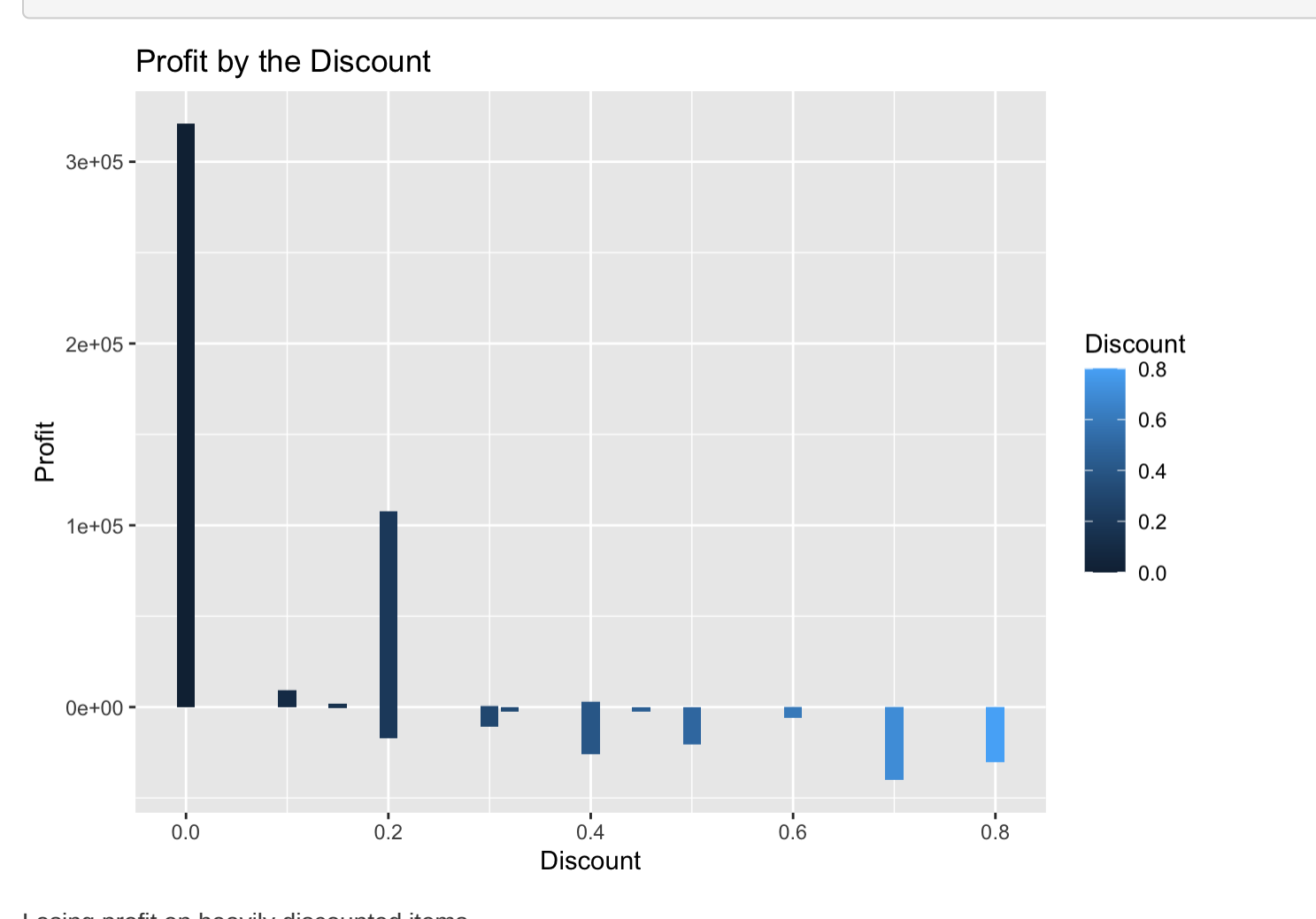
2.10 Discount analysis

```
ggplot(retail.unique, aes(Discount, Sales, fill = Discount))+
  geom_col()+
  ggtitle("Sales by the Discount")
```



Most sales were made when no discount was applied or when 20% was applied.

```
ggplot(retail.unique, aes(Discount, Profit, fill = Discount))+
  geom_col()+
  ggtitle("Profit by the Discount")
```



Losing profit on heavily discounted items.