The Sparks Foundation: Graduate Rotational Internship Program

Raphael Frimpong Marfo

19 March, 2022

Exploratory Data Analysis on Super Store

Introduction

The Super Store is a small retail business. They sell Furniture, Office Supplies and Technology products and their customers are the mass Consumer, Corporate and Home Offices. The data set contains sales, profit and geographical information of individual orders

The task is to determine weak areas and opportunities for Super Store to boost business growth.

The analysis will answer these questions:

1. What are the strongest to weakest categories in terms of revenue and profit
2. What are the strongest to weakest sub-categories
3. Are discounts given in each category and sub-categories worth it
4. Which region is the most profitable

There are several factors that can affect the profitability of a good but the one provided in the data is **discounts**. There will be a major focus on how discounts affects profitability.

Preparing The Environment

Setting up my R environment by loading the required R packages and Super Store dataset to aid with the analysis

```
library('tidyverse')
library('tidyr')
library('dplyr')
library("here")
library("skimr")
library("janitor")
library("ggplot2")
library("patchwork")
```

```
SampleSuperstore <- read_csv("SampleSuperstore.csv")
```

Data Exploration

Let's have a preview at the data set

```
 head(SampleSuperstore)
```

```
## # A tibble: 6 x 13
##   `Ship Mode`    Segment   Country     City  State `Postal Code` Region Category
##   <chr>          <chr>     <chr>       <chr> <chr>         <dbl> <chr>  <chr>
## 1 Second Class   Consumer  United Sta~ Hend~ Kent~         42420 South  Furnitu~
## 2 Second Class   Consumer  United Sta~ Hend~ Kent~         42420 South  Furnitu~
```

```
## 3 Second Class   Corporate United Sta~ Los ~ Cali~         90036 West   Office ~
## 4 Standard Class Consumer  United Sta~ Fort~ Flor~         33311 South  Furnitu~
## 5 Standard Class Consumer  United Sta~ Fort~ Flor~         33311 South  Office ~
## 6 Standard Class Consumer  United Sta~ Los ~ Cali~         90032 West   Furnitu~
## # ... with 5 more variables: `Sub-Category` <chr>, Sales <dbl>, Quantity <dbl>,
## #   Discount <dbl>, Profit <dbl>
```

```
tail(SampleSuperstore)
```

```
## # A tibble: 6 x 13
##   `Ship Mode`    Segment   Country     City  State `Postal Code` Region Category
##   <chr>          <chr>     <chr>       <chr> <chr>         <dbl> <chr>  <chr>
## 1 Standard Class Corporate United Sta~ Athe~ Geor~         30605 South  Technol~
## 2 Second Class   Consumer  United Sta~ Miami Flor~         33180 South  Furnitu~
## 3 Standard Class Consumer  United Sta~ Cost~ Cali~         92627 West   Furnitu~
## 4 Standard Class Consumer  United Sta~ Cost~ Cali~         92627 West   Technol~
## 5 Standard Class Consumer  United Sta~ Cost~ Cali~         92627 West   Office ~
## 6 Second Class   Consumer  United Sta~ West~ Cali~         92683 West   Office ~
## # ... with 5 more variables: `Sub-Category` <chr>, Sales <dbl>, Quantity <dbl>,
## #   Discount <dbl>, Profit <dbl>
```

Data Cleaning

Cleaning data set off any "dirt"(duplicate values, null values). Uncleaned data sets can produce biased analysis

```
is.null(SampleSuperstore)
```

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

```
SampleSuperstore<-unique(SampleSuperstore)
```

Data Manipulation

Filtering out data for each category for further analysis

```
SampleSuperstore_furniture <- filter(SampleSuperstore, Category == 'Furniture')
SampleSuperstore_OfficeSupllies <- filter(SampleSuperstore, Category == 'Office Supplies')
SampleSuperstore_Technology <- filter(SampleSuperstore, Category == 'Technology')
```

Renaming SampleSuperstore to df for analysis

```
df <- SampleSuperstore
```

Descriptive Summary

Descriptive summary is to have an initial statistical preview about the data set

```
summary(SampleSuperstore)
```

```
##   Ship Mode           Segment            Country             City
##  Length:9977        Length:9977        Length:9977        Length:9977
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     State            Postal Code       Region            Category
```

```
##  Length:9977        Min.   : 1040   Length:9977        Length:9977
##  Class :character   1st Qu.:23223   Class :character   Class :character
##  Mode  :character   Median :55901   Mode  :character   Mode  :character
##                     Mean   :55155
##                     3rd Qu.:90008
##                     Max.   :99301
##  Sub-Category           Sales              Quantity         Discount
##  Length:9977        Min.   :    0.444   Min.   : 1.000   Min.   :0.0000
##  Class :character   1st Qu.:   17.300   1st Qu.: 2.000   1st Qu.:0.0000
##  Mode  :character   Median :   54.816   Median : 3.000   Median :0.2000
##                     Mean   :  230.149   Mean   : 3.791   Mean   :0.1563
##                     3rd Qu.:  209.970   3rd Qu.: 5.000   3rd Qu.:0.2000
##                     Max.   :22638.480   Max.   :14.000   Max.   :0.8000
##      Profit
##  Min.   :-6599.978
##  1st Qu.:    1.726
##  Median :    8.671
##  Mean   :   28.690
##  3rd Qu.:   29.372
##  Max.   : 8399.976
```

```
summary(SampleSuperstore_furniture)
```

```
##   Ship Mode           Segment            Country             City
##  Length:2118        Length:2118        Length:2118        Length:2118
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     State            Postal Code       Region            Category
##  Length:2118        Min.   : 1040   Length:2118        Length:2118
##  Class :character   1st Qu.:22801   Class :character   Class :character
##  Mode  :character   Median :60505   Mode  :character   Mode  :character
##                     Mean   :55716
##                     3rd Qu.:90032
##                     Max.   :99301
##  Sub-Category           Sales              Quantity         Discount
##  Length:2118        Min.   :   1.892   Min.   : 1.000   Min.   :0.000
##  Class :character   1st Qu.:  47.060   1st Qu.: 2.000   1st Qu.:0.000
##  Mode  :character   Median : 182.103   Median : 3.000   Median :0.200
##                     Mean   : 350.003   Mean   : 3.787   Mean   :0.174
##                     3rd Qu.: 435.168   3rd Qu.: 5.000   3rd Qu.:0.300
##                     Max.   :4416.174   Max.   :14.000   Max.   :0.700
##      Profit
##  Min.   :-1862.312
##  1st Qu.:  -12.871
##  Median :    7.782
##  Mean   :    8.698
##  3rd Qu.:   33.727
##  Max.   : 1013.127
```

```
summary(SampleSuperstore_OfficeSupllies)
```

```
##   Ship Mode           Segment            Country             City
```

```
##  Length:6012        Length:6012        Length:6012        Length:6012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     State            Postal Code       Region            Category
##  Length:6012        Min.   : 1453   Length:6012        Length:6012
##  Class :character   1st Qu.:23223   Class :character   Class :character
##  Mode  :character   Median :55123   Mode  :character   Mode  :character
##                     Mean   :54835
##                     3rd Qu.:90004
##                     Max.   :99301
##  Sub-Category          Sales             Quantity         Discount
##  Length:6012        Min.   :   0.444   Min.   : 1.000   Min.   :0.0000
##  Class :character   1st Qu.:  11.760   1st Qu.: 2.000   1st Qu.:0.0000
##  Mode  :character   Median :  27.536   Median : 3.000   Median :0.0000
##                     Mean   : 119.550   Mean   : 3.803   Mean   :0.1574
##                     3rd Qu.:  79.960   3rd Qu.: 5.000   3rd Qu.:0.2000
##                     Max.   :9892.740   Max.   :14.000   Max.   :0.8000
##      Profit
##  Min.   :-3701.893
##  1st Qu.:    2.099
##  Median :    6.882
##  Mean   :   20.353
##  3rd Qu.:   19.423
##  Max.   : 4946.370
```

```
summary(SampleSuperstore_Technology)
```

```
##   Ship Mode           Segment            Country             City
##  Length:1847        Length:1847        Length:1847        Length:1847
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     State            Postal Code       Region            Category
##  Length:1847        Min.   : 1841   Length:1847        Length:1847
##  Class :character   1st Qu.:23392   Class :character   Class :character
##  Mode  :character   Median :59601   Mode  :character   Mode  :character
##                     Mean   :55552
##                     3rd Qu.:90008
##                     Max.   :99207
##  Sub-Category          Sales             Quantity         Discount
##  Length:1847        Min.   :    0.99   Min.   : 1.000   Min.   :0.0000
##  Class :character   1st Qu.:   68.02   1st Qu.: 2.000   1st Qu.:0.0000
##  Mode  :character   Median :  166.16   Median : 3.000   Median :0.2000
##                     Mean   :  452.71   Mean   : 3.757   Mean   :0.1323
##                     3rd Qu.:  448.53   3rd Qu.: 5.000   3rd Qu.:0.2000
##                     Max.   :22638.48   Max.   :14.000   Max.   :0.7000
##      Profit
##  Min.   :-6599.978
##  1st Qu.:    5.204
##  Median :   25.018
```

```
##  Mean   :   78.752
##  3rd Qu.:   74.895
##  Max.   : 8399.976
```

The main purpose of the descriptive summary is to throw an early analytical light on the data set. It gives an initial hint about the weakest areas in sample Superstore and the rest of the analysis are built on this foundation.

1. Technology has the highest average profit despite having the lowest quantity bought per transaction and the lowest average discount.
2. Office supplies has the highest quantity bought per transaction, second in discounts and lowest in sales revenue per transaction.
3. Furniture has the highest discounts, lowest average profit, second in quanity and second in average sales revenue.

The descriptive summary makes it obvious that Furniture and Office Supplies are the least profitable categories and therefore the potential weak links.

Exploratory Data Analysis

```
df_grp_category = df %>% group_by(Category) %>%
  summarise(total_sales = sum(Sales), total_profit = sum(Profit),
    total_quantity = sum(Quantity), total_discount = sum(Discount),
    average_sales = mean(Sales), average_profit = mean(Profit),
    average_quantity = mean(Quantity), average_discount = mean(Discount)*100,
    .groups = "drop")
 view(df_grp_category)
```

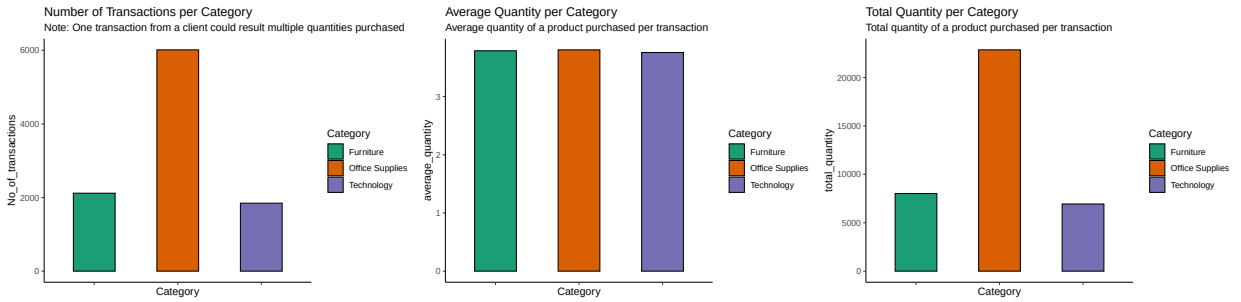
Number of transactions per each category

```
# counting the number of transactions per each category
df_grp_category_count = df %>% group_by(Category) %>% count()
view(df_grp_category_count)

#creating a new column with the column name of No_of_transactions
df_grp_category_count = mutate(df_grp_category_count, No_of_transactions = n)

#dropping column n from the df_grp_category_count table
df_grp_category_count = select(df_grp_category_count, -n)
view(df_grp_category_count)
```

Bar charts to compare the Sales, Profit and Discount of all SampleSuperStore categories

Comparing number of transactions with quantity purchased per category

```
transactions + quantity_average + quantity_total
```



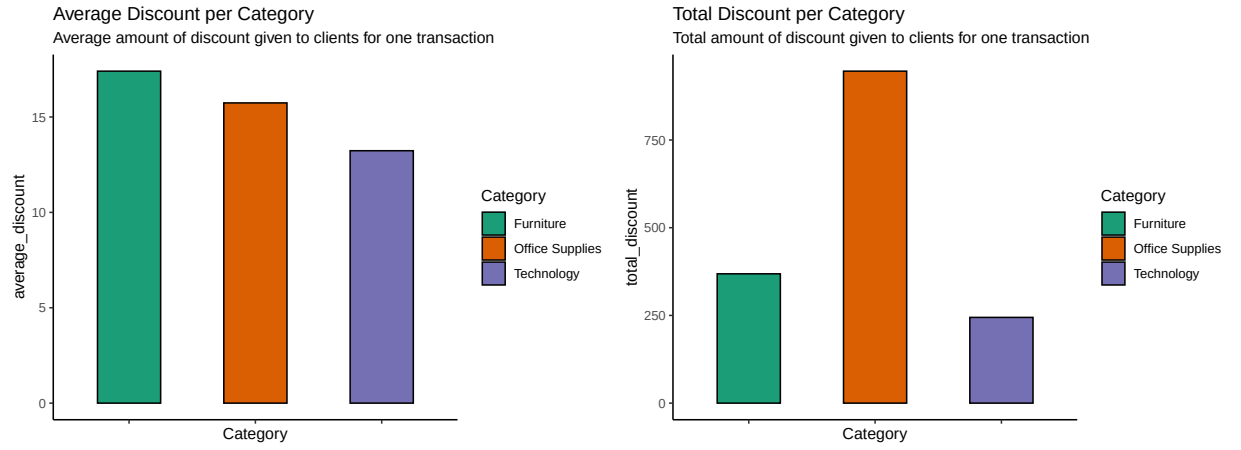
There are several factors that could affect how much a good is bought. In the SampleSuperstore data, discount is potentially a major factor. When companies offers discounts it decreases the average sales revenue and average profits, therefore, discounts are given to substantially increase the number of transactions and items purchased in order to cover for the drop in price of the good due to the discount.

1. The charts above show that despite not having the highest discount rate, office supply substantially has more transactions and quantities purchased than the other categories.
2. Furniture despite having the highest discount rate is significantly behind Office supplies.
3. Technology has the lowest transactions but not significantly lower than furniture despite having the lowest discount rates.

Furniture not pulling a lot of transactions despite the high discount rates shines light on it as potentially the weakest area. The 'type of good', 'price' and the consumers 'marginal utility' could be potentially causes why Furniture does not generate a lot of transactions.

comparing discounts offered per category

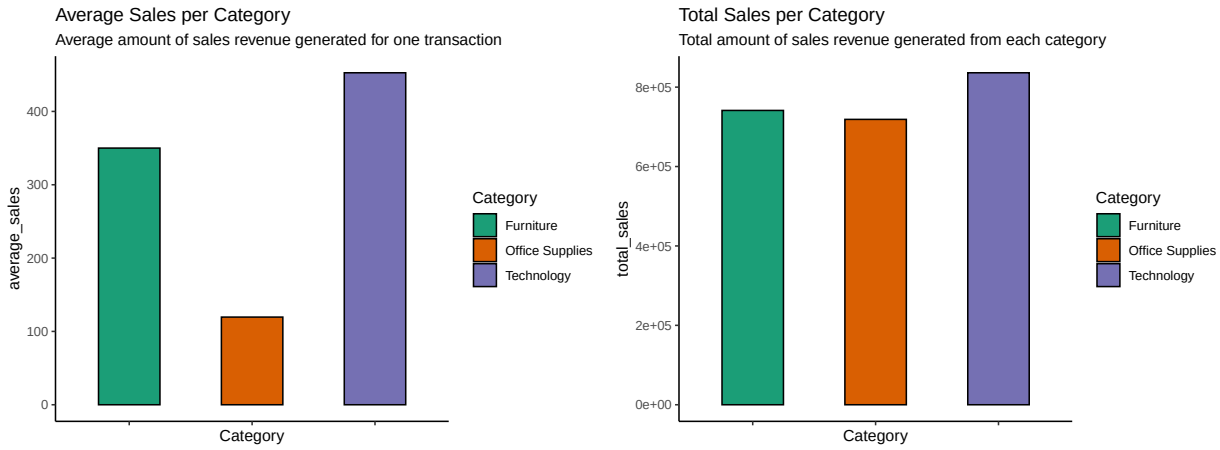
```
discount_average +discount_total
```



Per transaction, furniture had the highest discount rate however Office Supplies has the highest in total. This might be because discounts given generated a lot more discounted transactions.

comparing sales revenue per category

```
sales_average + sales_total
```



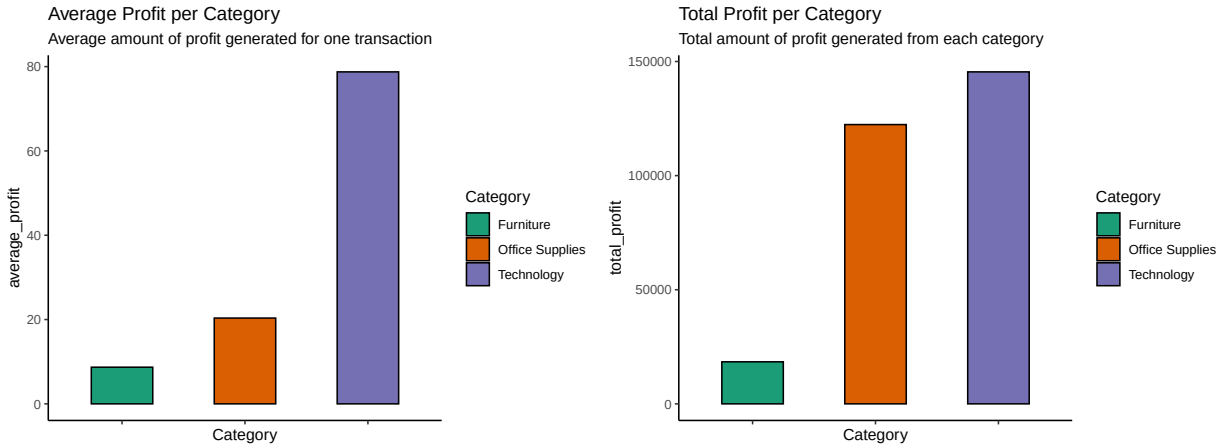
Office Supplies has lowest sales revenue generated despite having the highest number of transactions. This is probably due to the low prices of office supplies goods and to an extent discounts. Discount is a not a major factor because Furniture had more discounts on average but still has significantly higher revenue generated.

1. Technology is the most profitable category and furniture is significantly the least profitable.

In general, it is expected that the most discounted categories have the least profit per transactions. However, this disadvantaged can be curbed by substantial increase in demand for the product. Office Supplies satisfy this requirement more than Furniture.

comparing profit per category

```
profit_average + profit_total
```



Scatter plots to determine the relationship between Profit, Sales Revenue and Discount from Office Supplies sales

Scatter plots are being used to know the exact level of discounts rate was more beneficial. It could potentially provide a guide on what good to apply discounts on and by how much.

Relationship between quantity purchased and discount per category

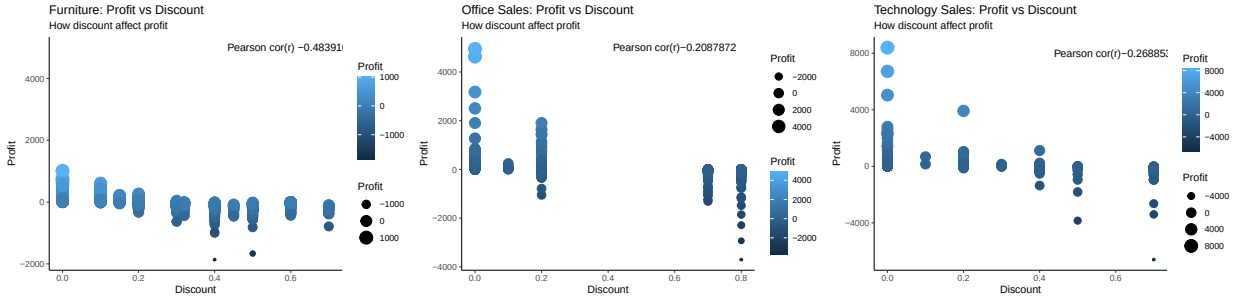
```
furniture_quantity + OfficeSupplies_quantity + Technology_quantity
```

Relationship between sales revenue and discount per category

```
furniture_sales + OfficeSupplies_sales + Technology_sales
```

Relationship between profits and discount per category

```
furniture_profit + OfficeSupplies_profit + Technology_profit
```



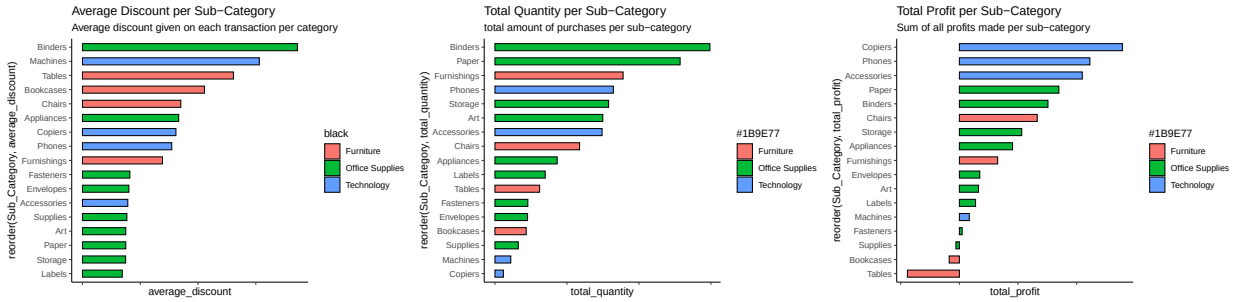
1. Furniture profits, sales and quantity all have a negative correlation with discount.
2. Office Supplies quantity had a positive correlation with discount but the other profit and sales had a negative correlation
3. With Technology profit and quantity had a negative correlation with discount except sales. The positive relationship with sales was due to the big outlier at 50% discount.

The analysis so far has proved that furniture sales exhibits a downward trend with discount in all aspects and it is clearly a weak area. Discounts given on furniture must not be more than 20% Office Supplies satisfies the aim of discounts as it significantly increases the demand for it. Negative profits for extreme discounts rate is expected since it decrease sales revenue per unit sold by a significant margin.

Further analysis will be taken on specific sub-categories that don't respond well to discounts and relevant suggestion will be made.

Barcharts comparing average discount, total profits and quantity per sub category in descending order

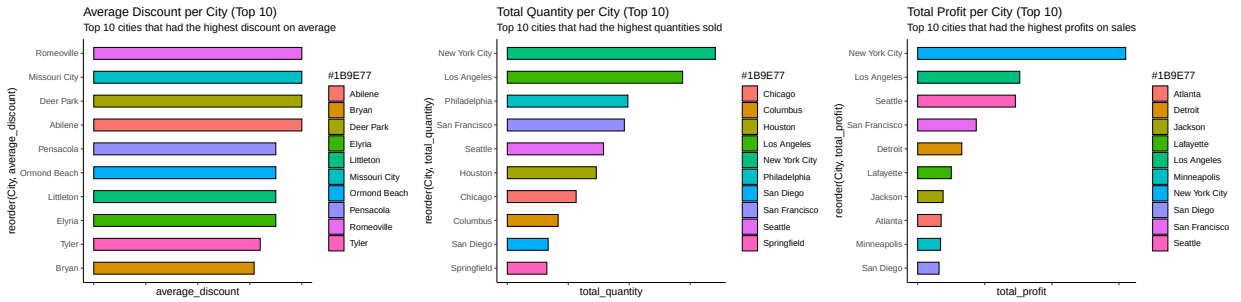
```
Sub_category_average_discount + sub_category_total_quantity + sub_category_total_profit
```



1. Furniture had 3 out of the top 5 sub-categories with the highest average discount but 1 in the top 5 for quantity and zero for profits. Tables and bookcases are part of the most discounted categories but part of the least demanded products and also had the greatest losses. Furnishings however responded positively to discounts.
2. Office Supplies dominates the top 5 in quantity demanded. Binders has had the most success with the discounts as it ranks top five in all three sections.
3. Technology generally is the most profitable category but discounts given to machines must be reconsidered as it ranks very low in quantity demanded and profits despite high rates of discounts given.

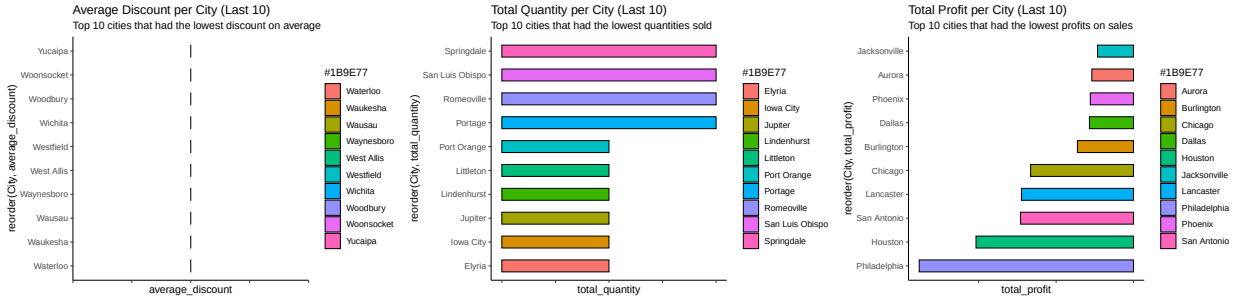
Bar charts comparing average discount, total profits and quantity per City in descending order

```
average_discount_city_top + total_quantity_city_top + total_profit_city_top
```



Bar charts comparing average discount, total profits and quantity per City in descending order **

```
average_discount_city_last + total_quantity_city_last + total_profit_city_last
```



1. None of the top 10 cities that had the most demand and profit shows up in the highest and lowest discounted cities.
2. Discount programs for Littleton and Elyria must be reconsidered as it ranks among the cities with the lowest quantities purchased.

Solutions

1. Discounts given on furniture products must be reduced in order to make it more profitable. Furniture's generally are needed in specific quantities and it is also durable making it less attractive for customers to buy more of it even if high discounts are offered.
2. Higher discounts rate can be explored for furnishings. Even though it is in the furniture category it responds very well to high discounts.
3. Although on average Office Supplies is second in the the most discounted category due to binders, most of its products are in the lower tier of discounts rates given. Higher discount programs must be explored for office supplies since it's product type responds well to higher discounts.
4. On average Techonology sales is flourishing and there should not be much change. However discounts given on machines must be reduced as it ranks among the lowest quantities purchased and least profitable products despite being the second highest discounted product.
5. Discount programs for Littleton City and Elyria City must be reduced because it is failing to drive demand in the city.