Dataset’s name: Retail Supermarket (Retrieved on 18th November 2022)
Dataset’s link: https://www.kaggle.com/datasets/roopacalistus/superstore

About the Dataset

The dataset contains sales details of different stores of a supermarket chain that has multiple stores in different parts of the US.

Explanation for each column:

- Ship Mode : Shipping class. There are 4 different types of ‘Ship Mode’, which are:
  Same Day : Shipping on same day.
  First Class : Shipping within 1 day.
  Second Class : Shipping within 3 days.
  Standard Class : Shipping within 6 days.
- Segment : Customer type. There are 3 segments, which are ‘Consumer’, ‘Corporate’, and ‘Home Office’.
- Country : The country where the sales happen (will be dropped later as the dataset specifically said that the sales are made in the US).
- City : The city, in USA, where the stores located. There are 531 different cities in this dataset.
- State : The state, in USA, where the stores located. There are 49 different states in this dataset.
- Postal Code : The postal code of the area.
- Region : The region, in USA, where the stores located. There are 4 different regions, which are ‘Central’, ‘East’, ‘West’, and ‘South’
- Category : Category of the item. There are 3 categories for the items, namely ‘office Supplies’, ‘Furniture’, and ‘Technology’.
- Sub-Category : Sub-category of the item. There are 17 sub-categories for the items.
- Sales : Total sale of the items (US$).
- Quantity : Number of items purchased.
- Discount : Discount of the items.
- Profit : The profit of the items saled (US$).

Assuming the price is in US$ since the stores are in US.

Type of Data in the Dataset

Data	Type
Ship Mode	Ordinal
Segment	Nominal
Country	Nominal
City	Nominal
State	Nominal
Postal Code	Nominal
Region	Nominal
Category	Nominal
Sub-Category	Nominal
Sales	Numerical
Quantity	Numerical
Discount	Numerical
Profit	Numerical

Changing Data

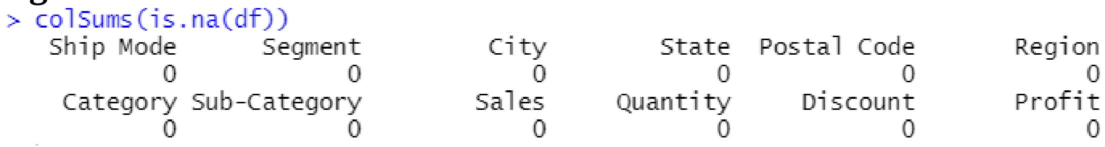
Changing the data type of the Postal Code to categorical because it is not numerical and should not be treated as such.

Because the Discount data is still in decimal form, it needs to be converted to percentage form by multiplying it by 100. The Discount is converted because we are more familiar with the percentage form.

Dropping Column

Data ‘Country’ is dropped since the only value of the data is ‘United States’, the country of stores, which is quite obvious.

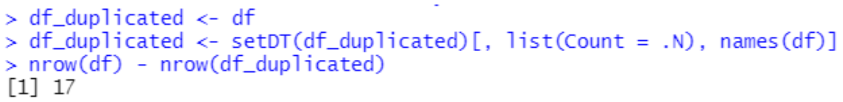
Missing Value



```
> colSums(is.na(df))
   Ship Mode      Segment         City        State  Postal Code       Region 
           0            0            0            0            0            0 
    Category Sub-Category        Sales     Quantity     Discount       Profit 
           0            0            0            0            0            0 
```

As can be seen, there is no missing value in this dataset.

Duplicated Row



```
> df_duplicated <- df
> df_duplicated <- setDT(df_duplicated)[, list(Count = .N), names(df)]
> nrow(df) - nrow(df_duplicated)
[1] 17
```

As can be seen, there are 17 duplicated rows in this dataset.

Because there is no unique identifier in this dataset, duplicated rows are unavoidable. It's nearly impossible to determine whether this is a duplicated row or maybe just a coincidence. As a result, dropping any duplicated row is not a wise decision.

Univariate Plot (Numerical)

(a) Sales

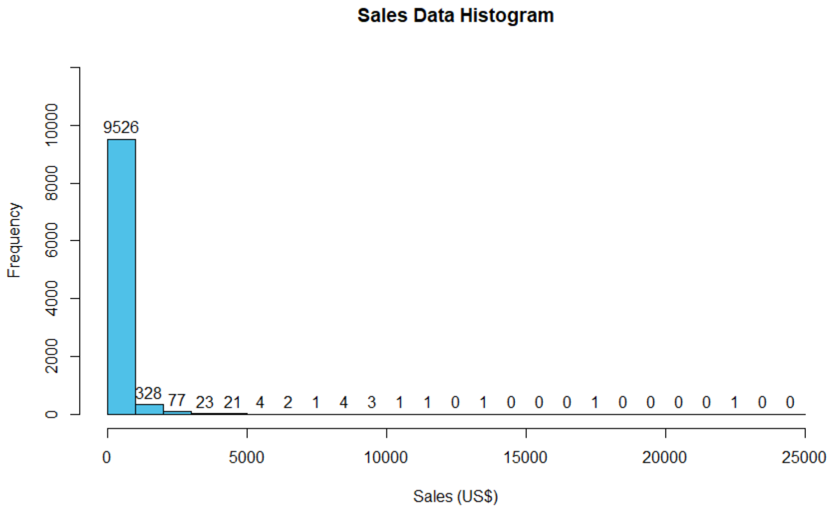


Figure 1. Sales Data Histogram with the class interval of $1000

As illustrated in Figure 1, more than 95% of transactions are under $1000 and more than 99% are under $5000. The histogram also clearly shows that the number of worth sales drops dramatically from the first to the second class, then gradually decreases with some fluctuation between $5000 and $25000 worth sales. The maximum value for Sales data is between $22000 and $23000.

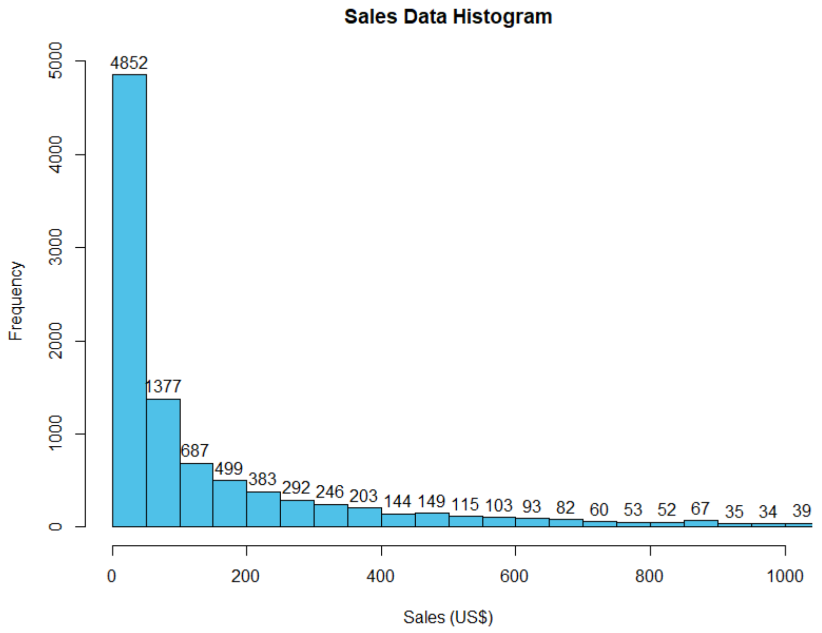


Figure 2. Sales Data Histogram with the class interval of $50

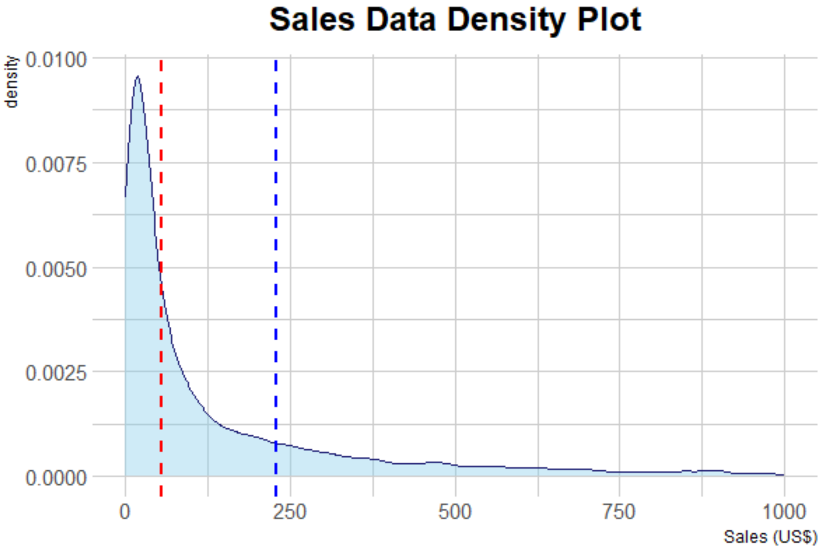


Figure 3. Sales Data Density Plot

Since it is difficult to analyze Sales data with such a wide range, let us concentrate on the sales under $1000.

As can be seen in the Figure 1, Figure 2, and Figure 3, the Sales data is heavily skewed to the right. The position of the mean (blue-dashed) and median (red-dashed) line in Figure 2 further demonstrate the right-skewedness of the distribution.

Figure 2 reveals that about 50% of sales are $50 or less and nearly 60% are $100 or less, indicating that the item purchased is more likely to be used for daily activities. Similar to Figure 1, the worth sales in Figure 2 are also drop significantly from the first class to the second class and then gradually decrease to the end.

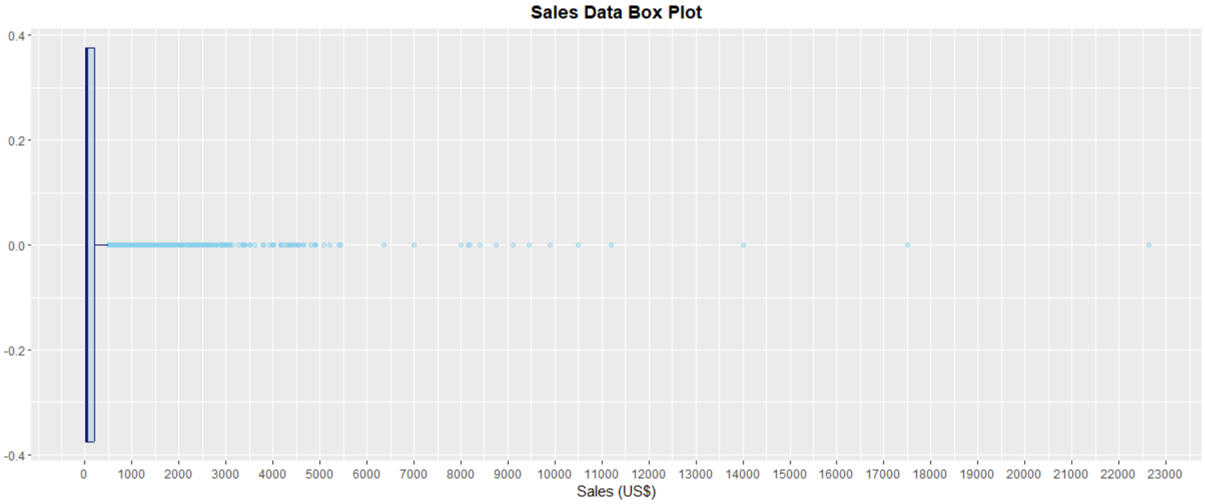


Figure 4. Sales Data Box Plot

The box plot depicts that there are numerous amounts of outliers in the Sales data. The fact that the Sales data is highly concentrated inside a small range while yet having a significant amount of data outside that range is not surprising, as it has been previously illustrated by the histogram and density plot.

Although the initial value of the outliers is difficult to determine, based on the grid line, it should be close to $500, which is also the boxplot's maximum value. The value of Q3 is around $200. While the minimum, Q1, and Q2/median values are difficult to pinpoint, they are unquestionably under $100 for Q2 and under $25 for both Q1 and the minimum value.

(b) Quantity

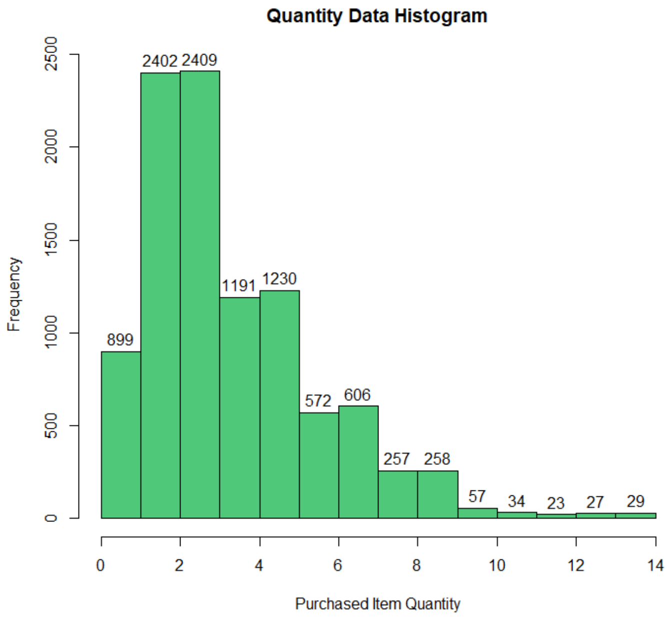


Figure 5. Quantity Data Histogram (Discrete)

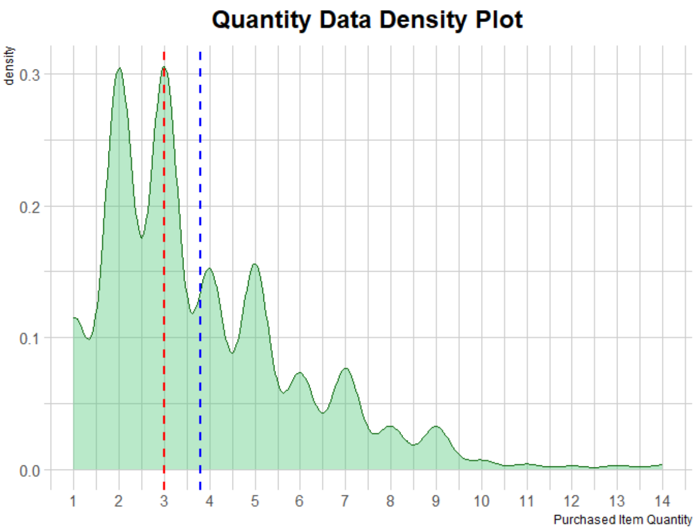


Figure 6. Quantity Data Density Plot (Discrete)

The density plot suggest that Quantity data is distribute in the form of bimodal distribution (distribution with 2 mode), with the main peak at 3 and the lower peak at 2. Since 2 and 3 are close to one another, let's treat them as a single peak. Similarly, there are several that are strangely adjacent and have a similar frequency, such as 4 and 5, 6 and 7, and 8. It's important to do more research to be sure, although this may or may not have any practical significance.

Figure 5 shows that the highest frequency of amounts of item purchased is 3 items and not so far off is 2 items by only a difference of 7. On the other hand, the least frequency of amounts of item purchased is 12 with the frequency of only 23. Shown in Figure 6, the Quantity data is positively skewed which additionally supported by the fact that the mean value (blue-dashed line) is more than the median value (red-dashed line).

There frequency of the Quantity data jumps to almost 300% at the beginning before falling to only half of it and eventually decline considerably until the end. There are also some points where the data remains stable.

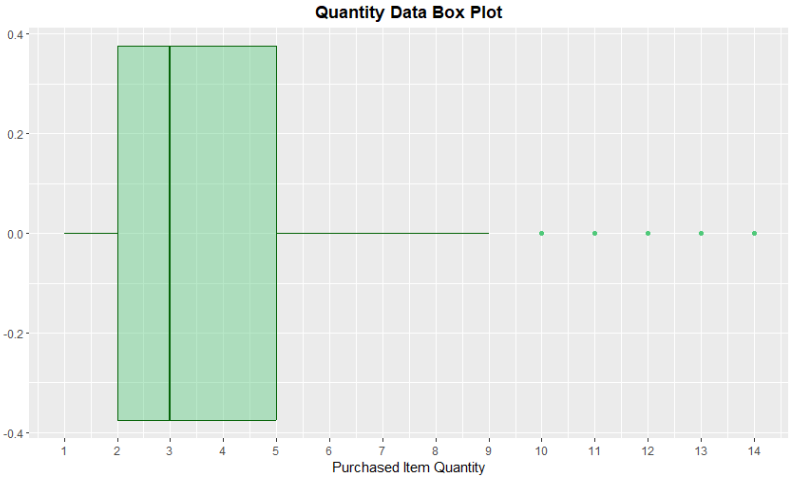


Figure 7. Quantity Data Box Plot (Discrete)

The box plot suggests that the there are few outliers, start from the value of 10 to 14. The minimum value, Q1, Q2/median, Q3, and maximum value of the boxplot consecutively are 1, 2, 3, 5, and 9.

(c) Discount

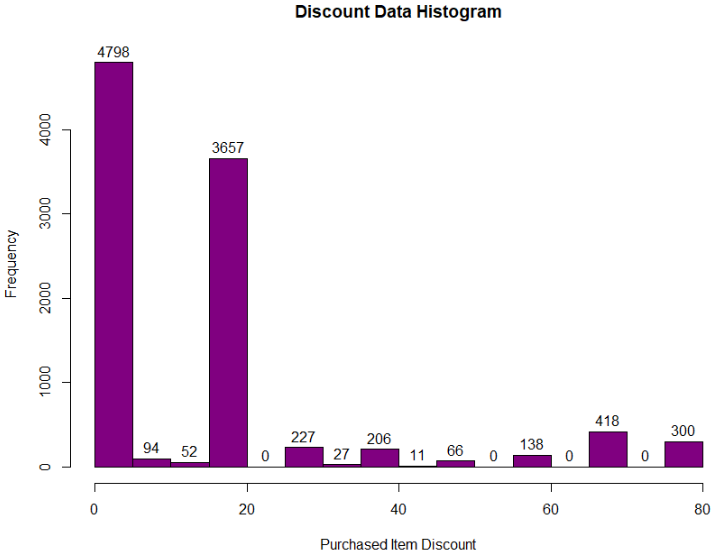


Figure 8. Discount Data Histogram (Discrete)

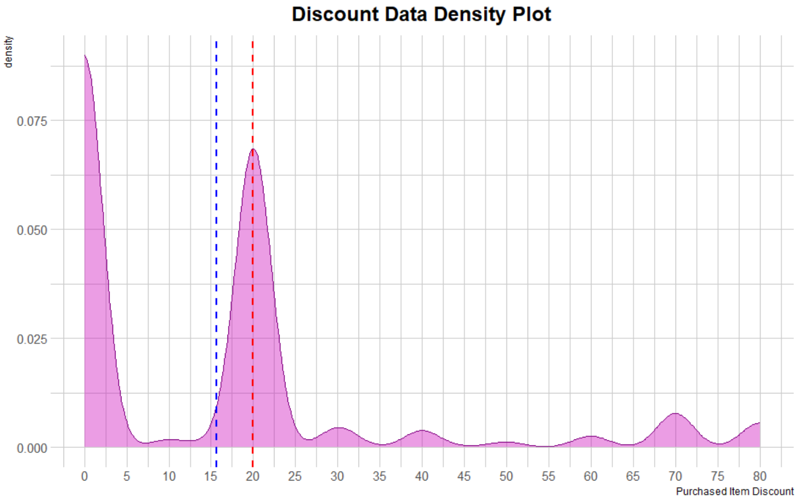
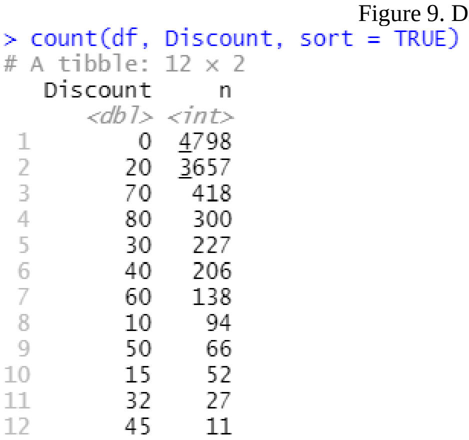


Figure 9. Discount Data Density Plot (Discrete)



```
> count(df, Discount, sort = TRUE)
# A tibble: 12 x 2
   Discount     n
      <dbl> <int>
 1        0  4798
 2       20  3657
 3       70   418
 4       80   300
 5       30   227
 6       40   206
 7       60   138
 8       10    94
 9       50    66
10       15    52
11       32    27
12       45    11
```

The density plot suggest that Quantity data is distribute in the form of bimodal distribution, with the main peak at 0 and the lower peak at 20. The double peaks are worth investigating to be certain about the reason behind.

From the histogram (and the console of RStudio) we can see that 0% discount's frequency exceeds 4500 and 20% discount's frequency exceeds 3500, while the others stay under 500.

The frequency is at its peak at the start, then drops significantly before rocketing back to the second peak at 20, Later the frequency falls again for the second time and then remain stable until the end.

The Discount data's mean value, shown by blue-dashed line, is around 15.5% while the median value, shown by red-dashed line, is 20%. Even though the mean value is under the median value, caused by extreme frequency of the two peaks, but we can be sure that the data distribution is positively skewed.

Other than both peaks, the Discount data distribution produced several "waves" which relatively low and stable.

There is strange value of discount which is the only one that is not the multiply of 5 (32%). There are also several abnormally huge value of discounts, such as 60%, 70%, and 80% which may or may not be caused by Black Friday (needed further investigation).

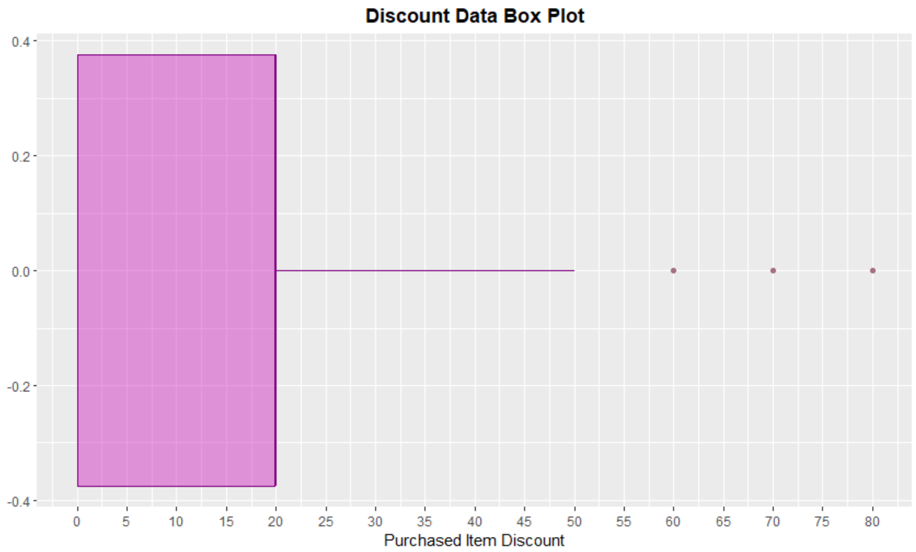


Figure 10. Discount Data Box Plot (Discrete)

The boxplot suggests there are three values considered as outliers, such as 60, 70, and 80. The minimum value and Q1 of the boxplot is coincide at 0 while the Q2/median and Q3 of the boxplot is also coincide at 20. The maximum value of the boxplot is 50.

(d) Profit

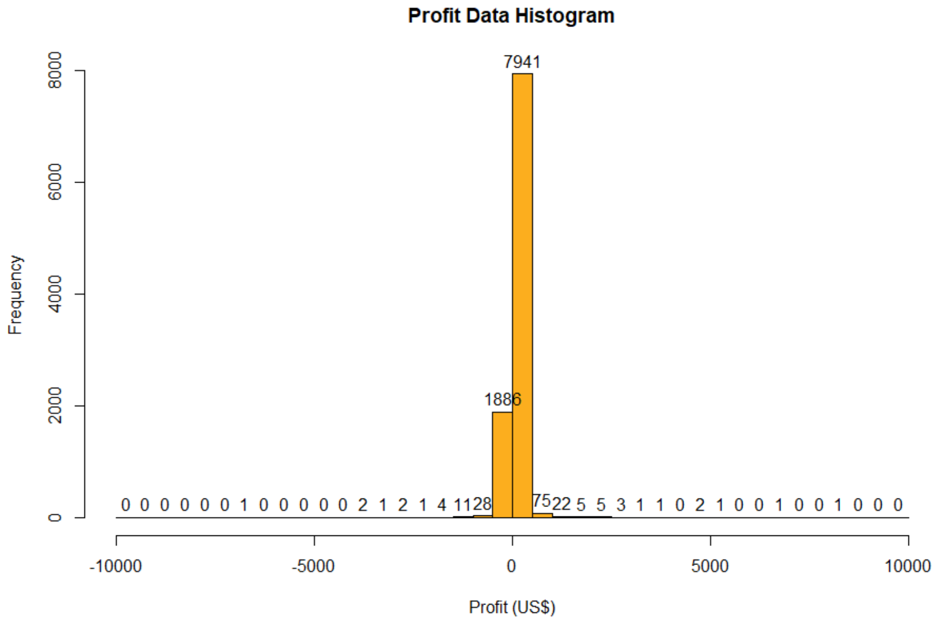


Figure 11. Profit Data Histogram with the class interval of $500

Shown by Figure 11, almost 80% of the Profit data range between $0 and 500$ and around 19% of the Profit data range between -$500 and $0, which combined produce more than 98% of the data. The frequency of the data outside -$500 and $500 range are relatively low and stable and gradually decrease as they go further away.

The maximum profit is lying inside the interval between $8000 and $8500 and the maximum loss is lying inside the interval between -$7000 and -$6500.

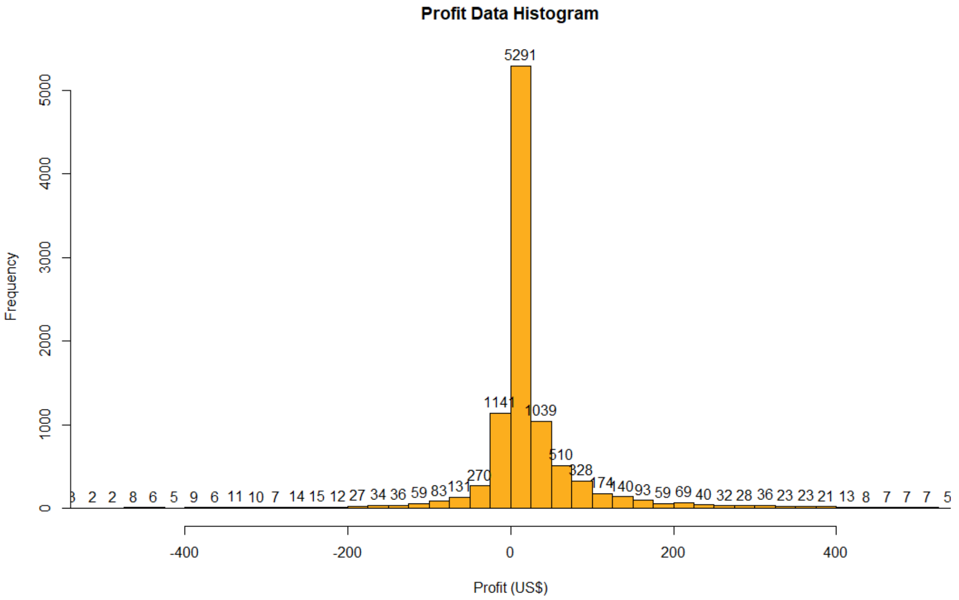


Figure 12. Profit Data Histogram with the class interval of $25

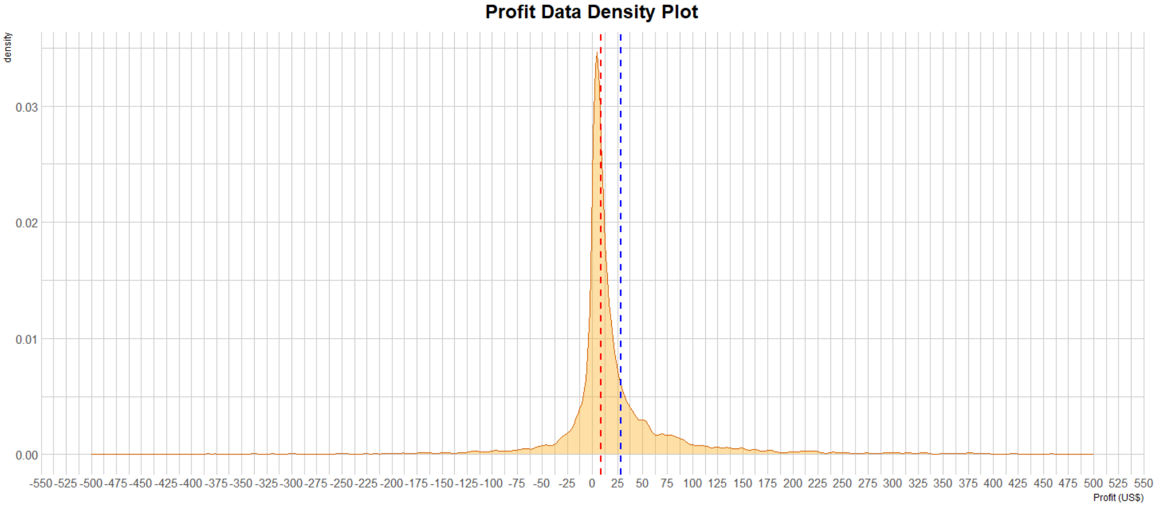


Figure 13. Profit Data Density Plot

Since it is difficult to analyze Sales data with such a wide range, let us concentrate on the profit between -$500 and $500.

There also considerable number of losses, which some are very high, as shown by the graph above while not as much as the profits.

Figure 12 shows that 53% of the Profit data range between $0 and 25$, 11% of the Profit data range between -$25 and $0, and 18% of the Profit data range between $25 and $100, which combined make more than 80% of the data.

As shown in the Figure 12 and Figure 13, the Profit data is slightly positively skewed, which also supported due to the position of the mean (blue-dashed) line which is on the right side of the median (red-dashed) line in Figure 13.

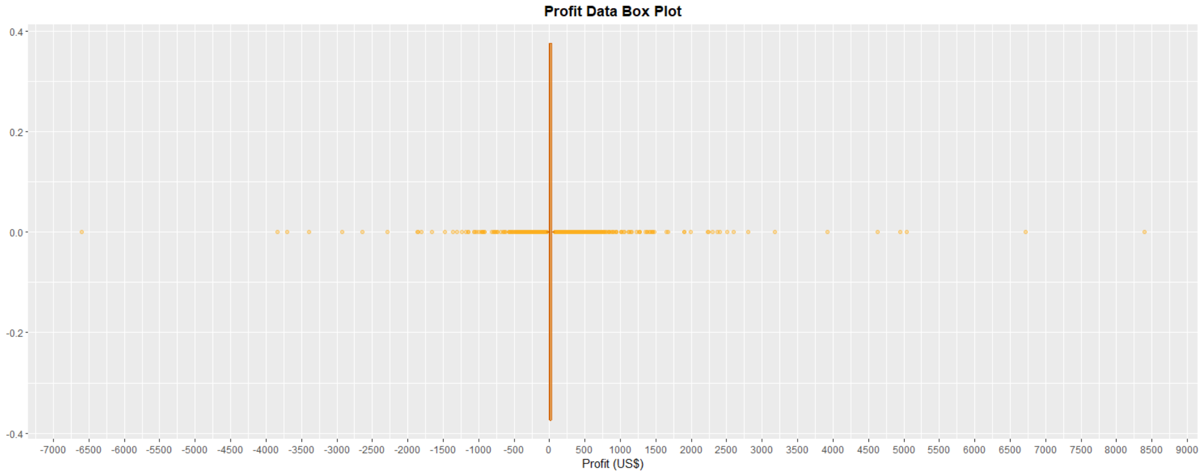


Figure 14. Profit Data Box Plot

It is impossible to get meaningful insight from this boxplot. Henceforth, let us create another boxplot with shorter range.

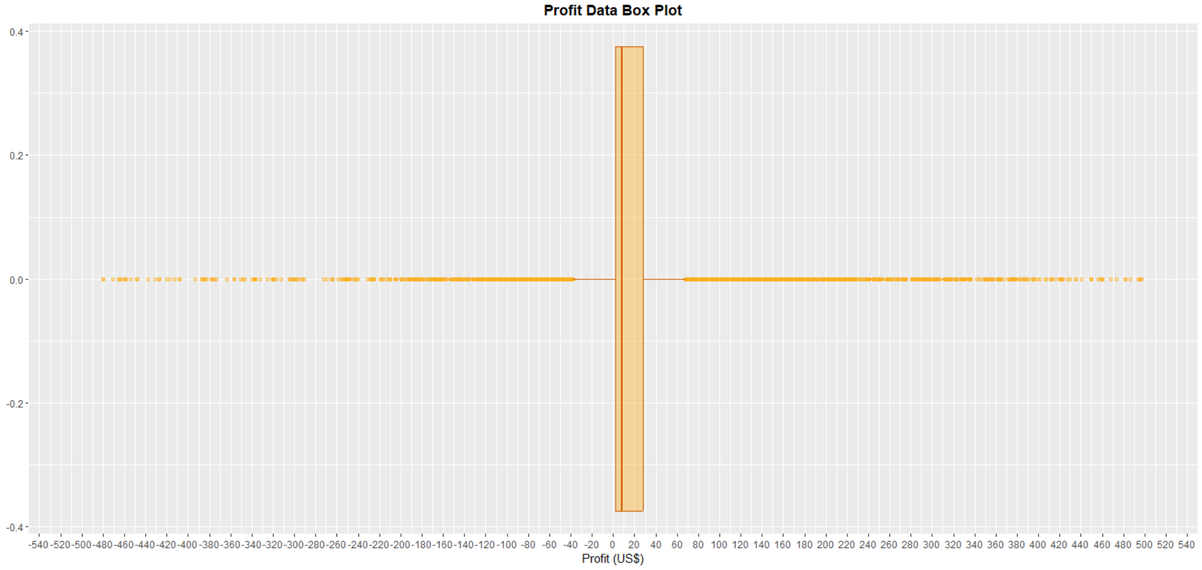


Figure 15. Profit Data Box Plot

There may be some data that is converted to outlier as we omit some data from the graph. As a result, the value of max, min, Q1, Q2, and Q3 may be shifted.

The box plot depicts that there are numerous amounts of outliers in the Profit data. However, it doesn't come as a surprise as both histogram and density plot has shown that the Profit data is very concentrated inside such a short range while also have a numerous amount of data outside that range as well.

It is hard to tell, but the minimum value of the boxplot should be around -$38, Q1 value around $1, Q2/median value around $9, Q3 value around 29$, and the max value of the box plot around 68$.

Univariate Plot (Categorical)

(a) Ship Mode

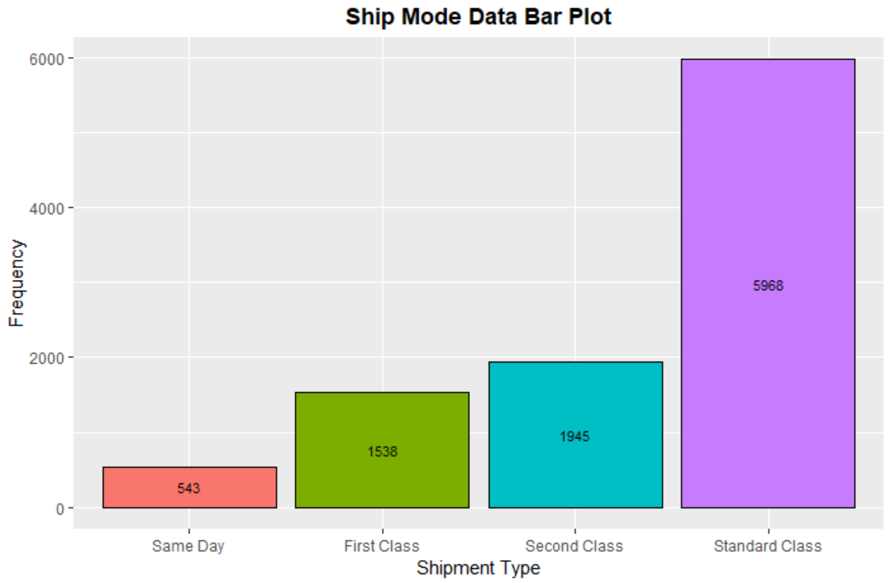


Figure 16. Shipment Type Data Bar Plot

The bar graph shows that the “Standard Class” is the mostly preferred by customers as it covers 60% of the Ship Mode data, 3 time as much as the second highest. On the other hand, the least common choice is “Same Day”, which only 5% of the total data.

This bar graph is to be expected as the data’s frequency is inversely proportional as it comes to price.

(b) Segment

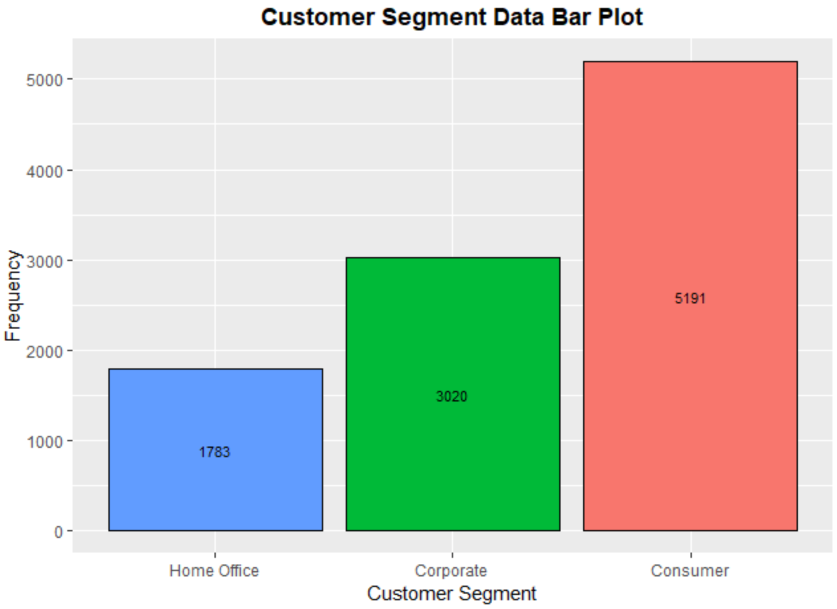


Figure 17. Segment Data Bar Plot

“Consumer” is the most common type of buyer as it dominates the data, with the frequency of 52% of the total data. “Corporate” is the second highest with the frequency of 30% of the total data and “Home Office” is the least of all with only 18% of the total data.

(c) City

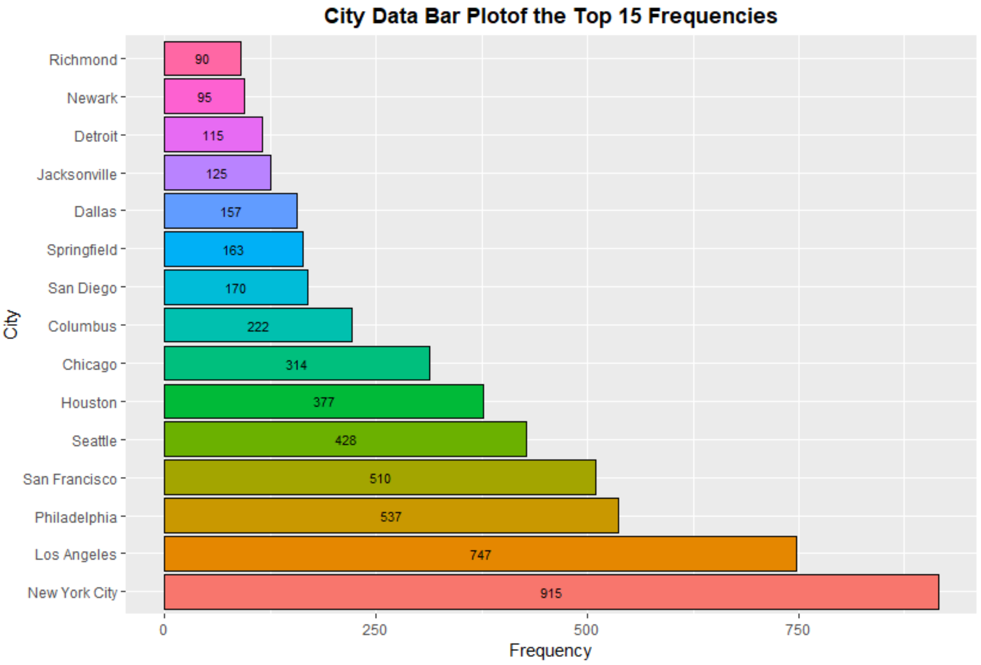


Figure 18. City Data Bar Plot

There are 531 different cities in this dataset, it is impossible to plot them all. Henceforth, I will only take top 15 cities.

As shown by Figure 18, the city with the utmost frequency is New York City, with the frequency of 915, followed by Los Angeles and Philadelphia on the second and third place, with the frequency of 747 and 537.

(d) State

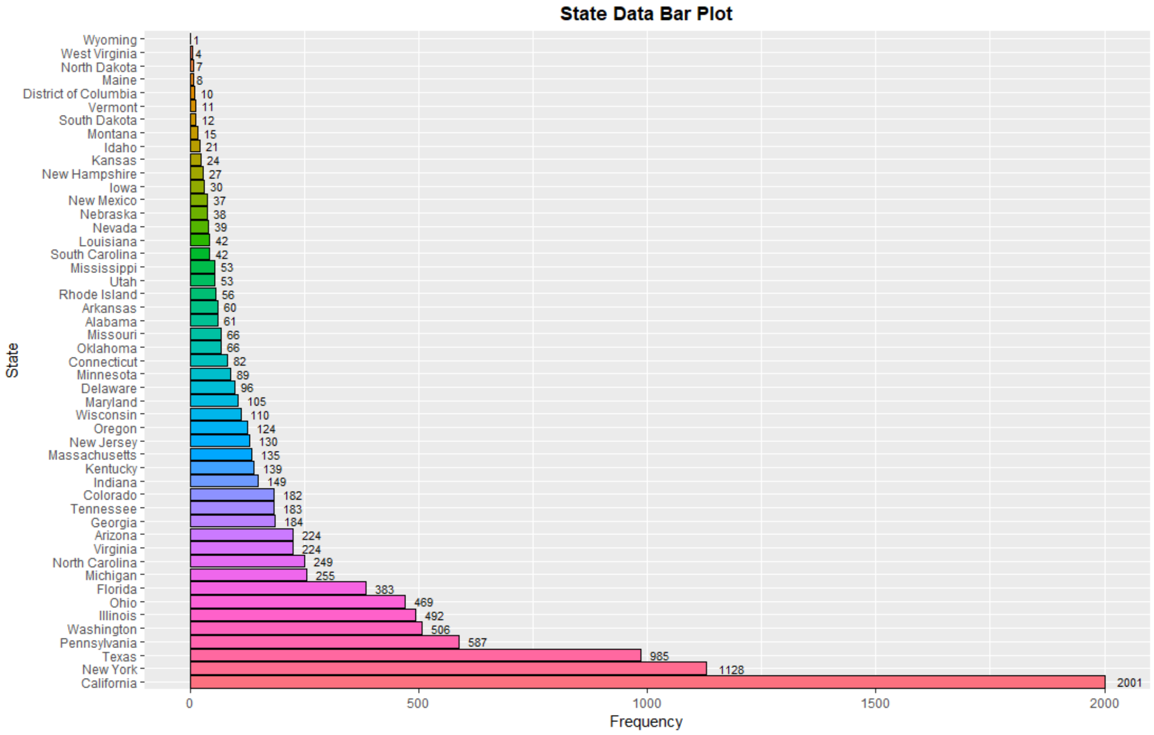


Figure 19. State Data Bar Plot

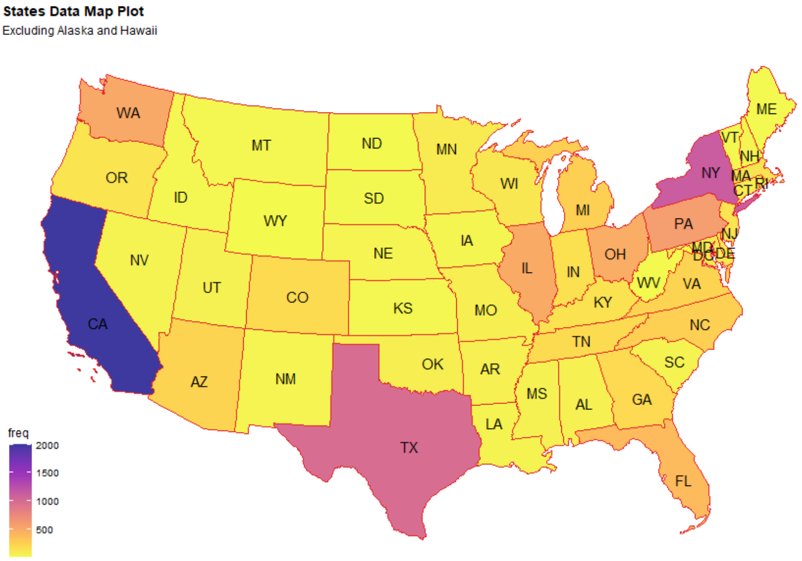


Figure 20. State Data Map Plot

There is total 49 States included in this dataset. Alaska (AK) and Hawaii (HI) are not included for unknown reason(s).

If you look at Figure 19 you will see that the State with utmost frequency is California, with the frequency of 2001 (around 20% of the total data), almost as much as the second (New York) and third (Texas) utmost frequency combined.

On the other hand, the state with the lowest frequency is Wyoming, with only 1 frequency, followed by West Virginia and North Dakota, with frequency of 4 and 7 respectively.

Figure 20 depicts that the location of the states has little to no effect on the frequency of the state.

(e) Postal Code

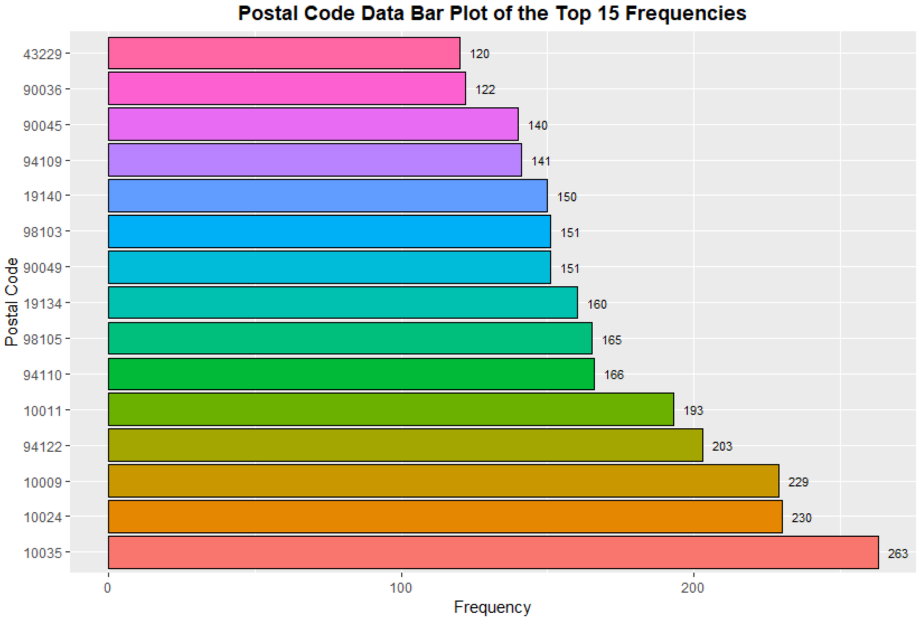


Figure 21. Postal Code Bar Plot

This dataset contains 631 areas with different postal codes; it is impossible to plot them all. I will only take the top 15 from now on.

Figure 21 depicts the postal codes 10035, 10024, and 10009 as having the highest frequency in this dataset. There is one thing they all have in common: they all begin with "100". This may or may not mean something.

It is difficult to get meaningful insight from this bar chart unless you are familiar with the geography and administration of the United States.

(f) Region

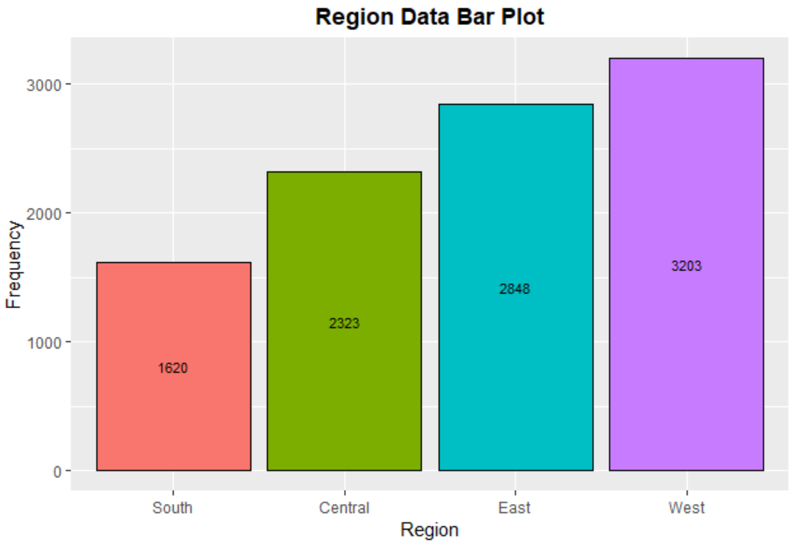


Figure 22. Region Data Bar Plot

Figure 22 present the distribution of the data frequency among the four regions. The West region hold the utmost frequency with almost a third of the data, followed by East Region in the second place, with the frequency of 2848, and Central Region in the third place, with the frequency of 2323. The least frequency is possessed by the South region, with the frequency of only 1620.

(g) Category

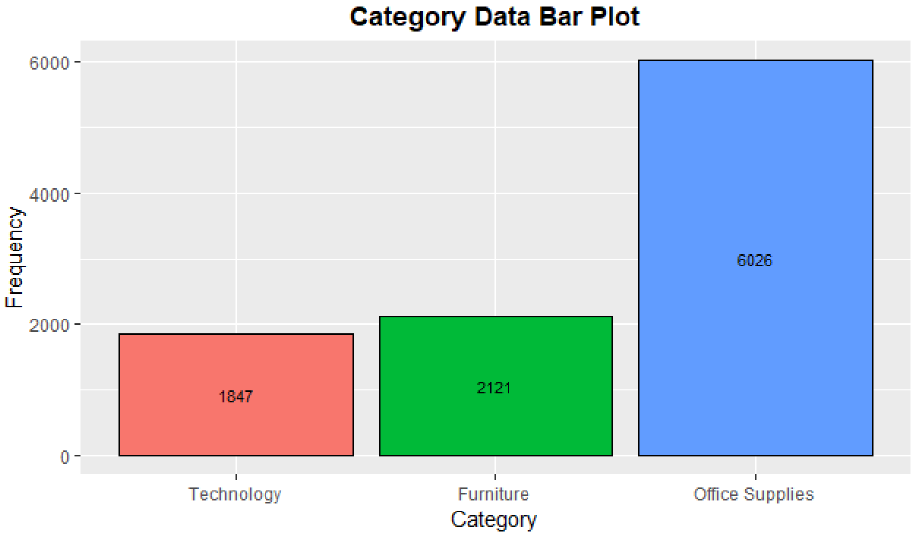


Figure 23. Category Data Bar Plot

According to Figure 23, the category "Office Supplies" dominates Category Data in the data set, accounting for 60% of the total Category data. The category "Furniture" comes in second with a frequency of 2121 and the category "Technology" comes in third with a frequency of 1847, not far behind.

(h) Sub-Category

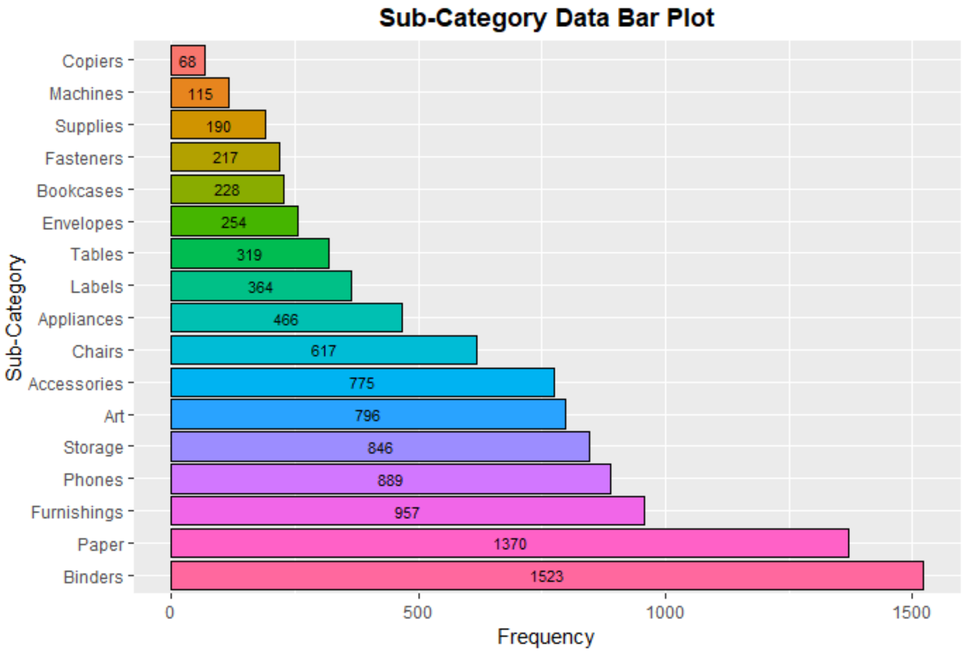


Figure 24. Sub-Category Data Bar Plot

Based on Figure 24, “Binders” and “Paper” are the most common sub-category of the item purchased with both exceed 1000 data frequency. On the other hand, “Copier” is the least common with only 68 of data frequency.

Bivariate Graph

(a) Ship Mode and Segment

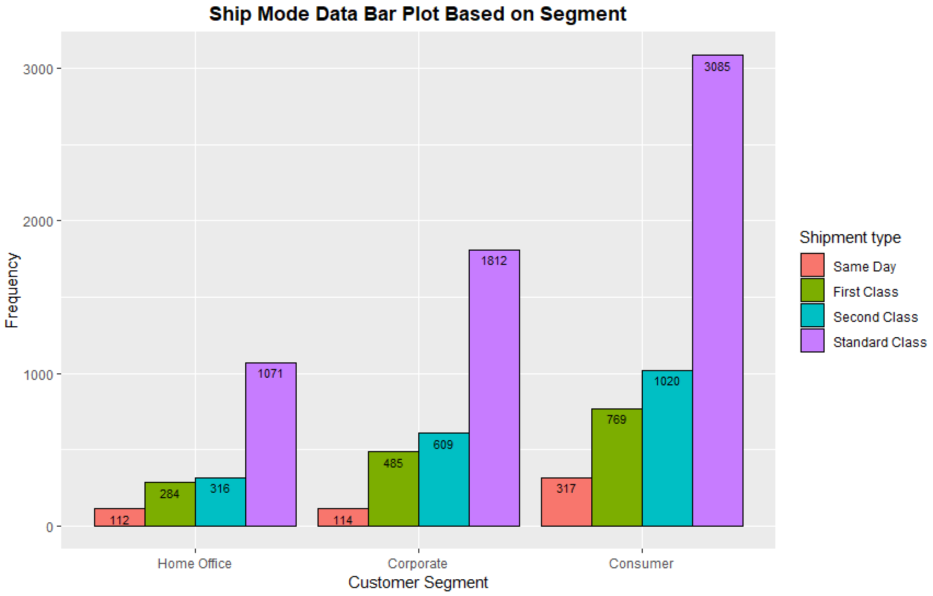


Figure 25. Ship Mode-Segment Data Bar Plot

As can be seen from Figure 25, The order of frequency of Segment data appears to be the same for each Shipment Mode. There appears to be a big gap between the frequency of “Standard Class” and the frequency of the other Shipment Mode as even the lowest segment frequency (Home Office) of the Standard Class is still higher than the highest segment frequency of the other Shipment Mode.

Even though the frequency of Corporate is over 1.7 times more than that of Home Office, there is only a very slight difference in data frequency between Home Office and Corporate when "Same Day" is chosen as the shipment method. The frequency of Home Office on the “First Class” and “Second Class” is also relatively similar.

(b) Category and Segment

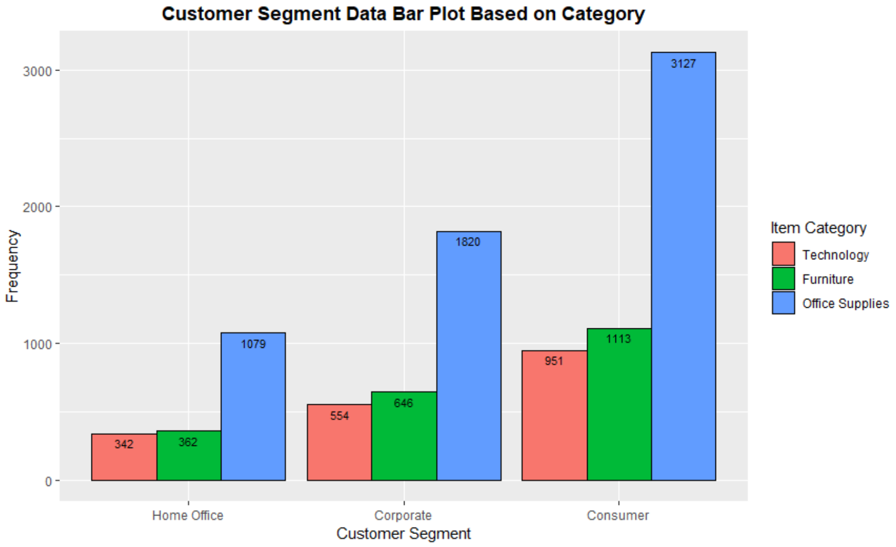


Figure 26. Category-Segment Data Bar Plot

Similar to Figure 25, the frequency order of the segment data appears to be the same for each Category. Each Segment's percentage of each category is similarly compared to one another. There are no additional noteworthy occurrences to interpret.

(c) Sales and Profit

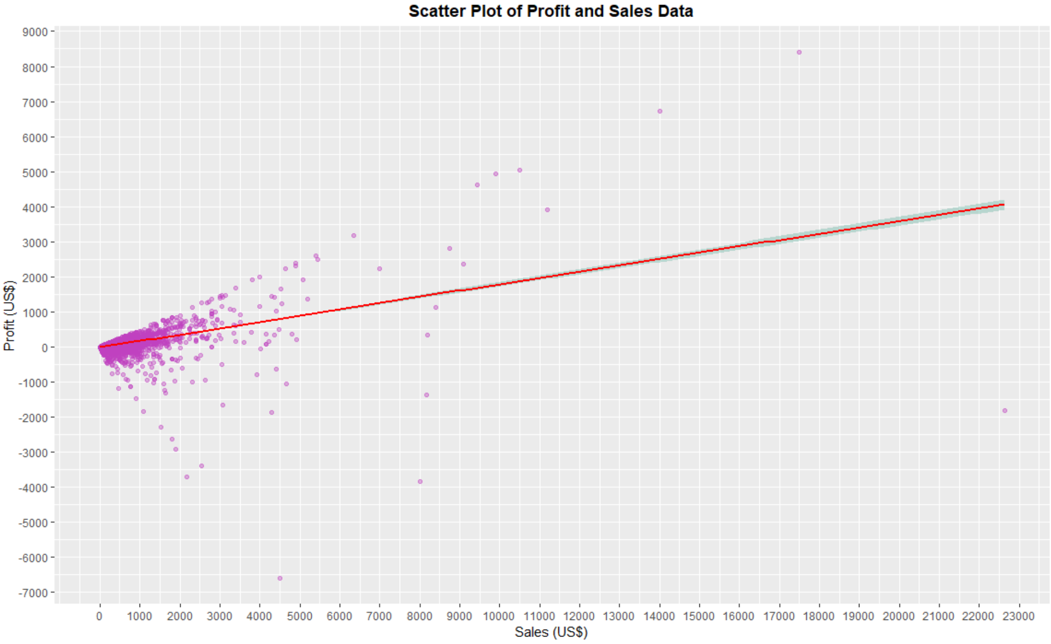


Figure 27. Sub-Category-Segment Data Bar Plot

By looking at figure, we can see that there is a lot of data point that is neither in regression line nor in the interval confidence, some data point is even so far away from the trend line. The data a point tends to create a big cluster with vague boundaries around point (0, 0). The cluster itself resembles the trendline in one way or another. This could mean that Sales data and Profit data may have some correlation to some degree, though not highly correlated.

(d) Category and Sub-Category

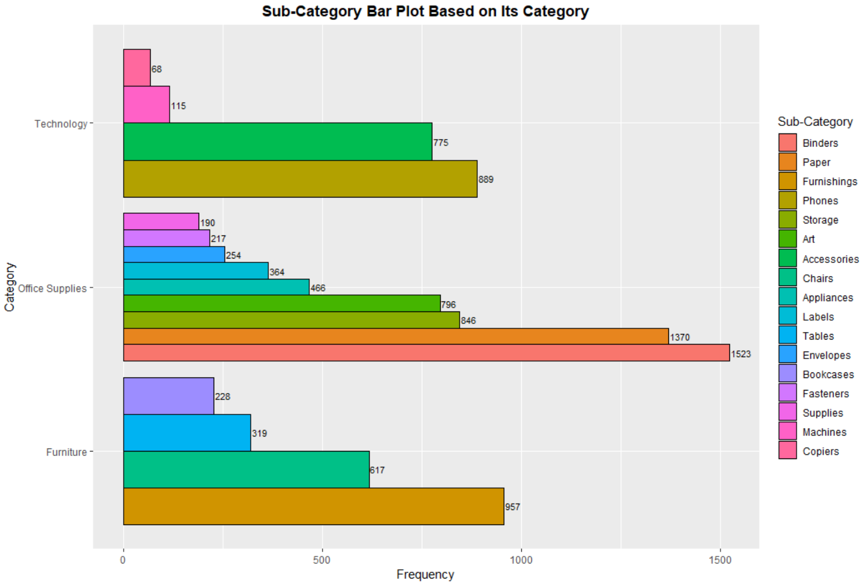


Figure 28. Sub-Category-Segment Data Bar Plot

Figure 28 shows classification of sub-category to its corresponding category. By doing this, we will be able to see the distinguish the sales of each category easily.

“Copiers”, “Machines”, “Accessories”, and “Phones” belong to “Technology” category.

“Supplies”, “Fasteners”, “Envelopes”, “Labels”, “Appliances”, “Art”, “Storage”, “Paper”, and “Binders” belong to “Office Supplies” category.

“Bookcases”, “Tables”, “Chairs”, and “Furnishings” belong to “Furniture” category.

“Phones”, “Binders”, and “Furnishings” is the sub-category with the top frequency to its respective category. Contradictorily, “Copiers”, “Supplies”, and “Bookcases” is the sub-category with the lowest frequency to its respective category.

(e) Sales and State

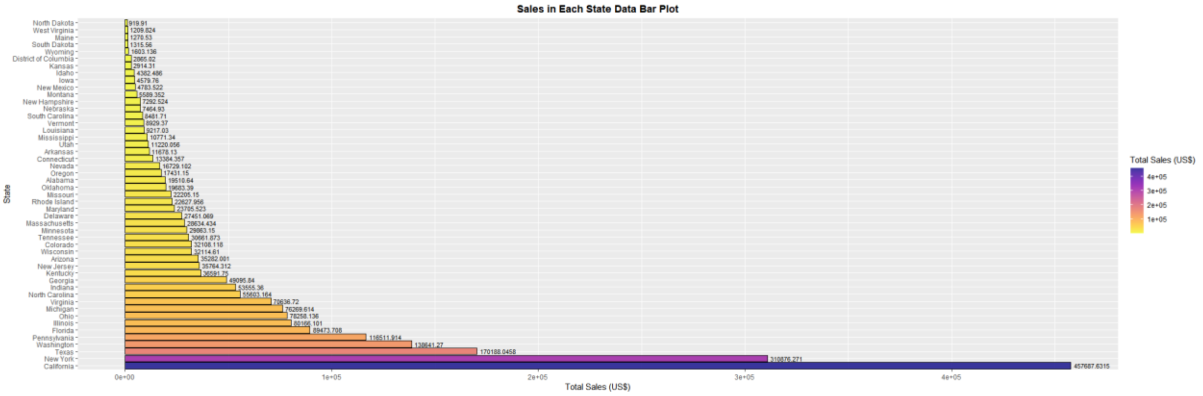


Figure 29. Total Sales in Each State Data Bar Plot

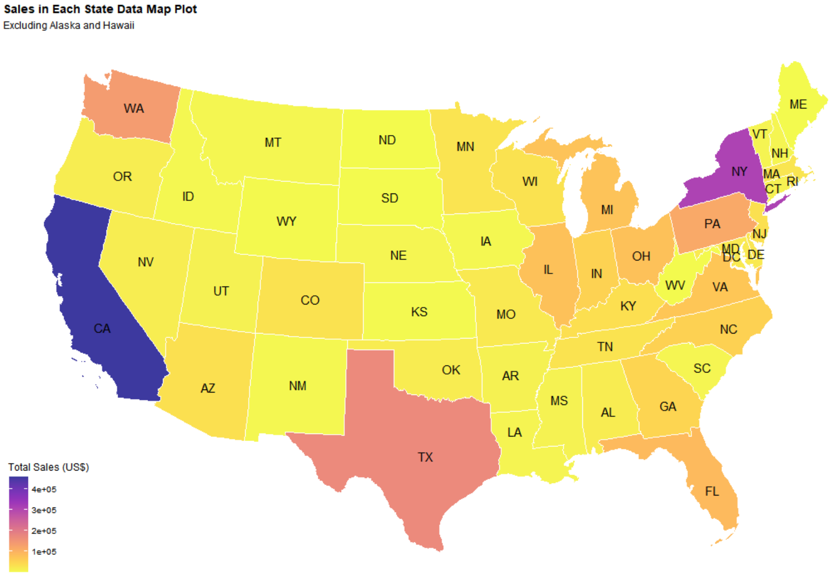


Figure 30. Total Sales in Each State Data Map Plot

As can be seen in Figure 30 where these states have the most noticeable color among the others, California, New York, Texas, and Washington are the four states with the highest sales. Additionally, you'll see that California and New York have the most distinctive colors (darker colors) in comparison to the others. This is given the fact that New York's total sales are nearly twice as high as Texas' total sales, which are ranked third, and California's total sales are 1.5 times greater than New York's total sales.

Figure 30 shows that states in the eastern, some southern, and some central regions are orange-yellow in color, signifying larger overall sales than the remaining.

(f) Profit and State

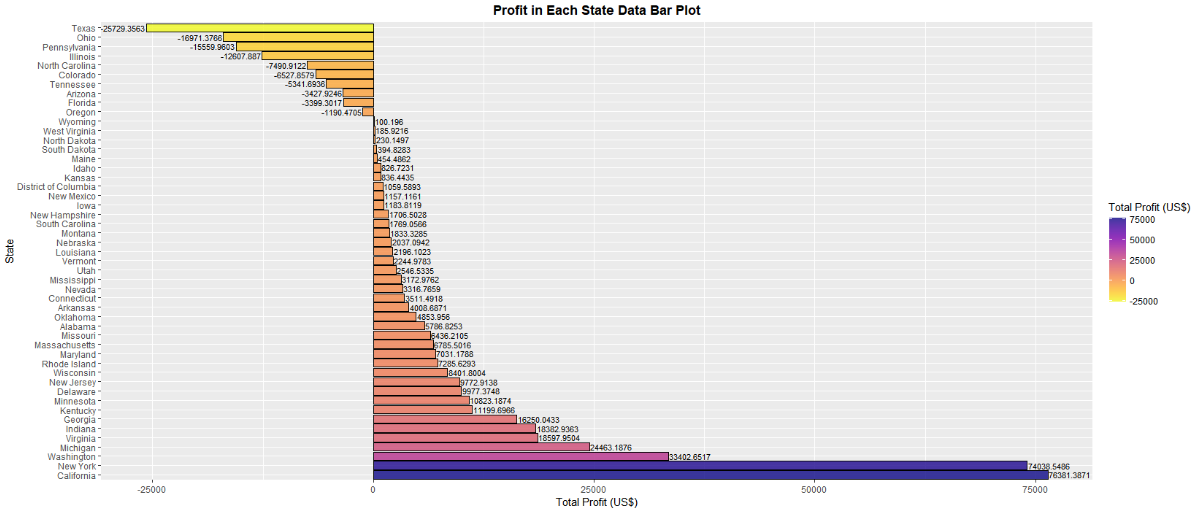


Figure 31. Total Profit in Each State Data Bar Plot

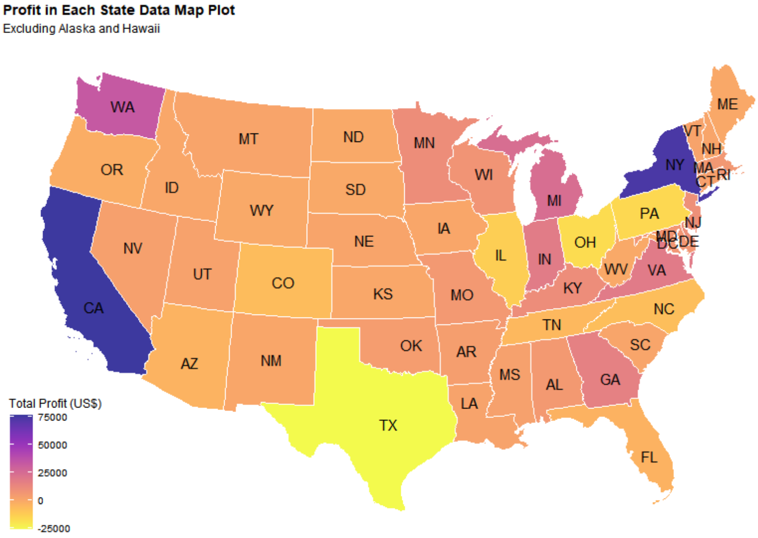


Figure 32. Total Profit in Each State Data Map Plot

Each state's overall profit and loss are shown in Figure 31. It seems that there are some states that suffer a loss, such as Oregon, Florida, Arizona, Tennessee, Colorado, North Carolina, Illinois, Pennsylvania, Ohio, and especially Texas with its noticeable bright yellow color (in Figure 32). Even though Texas is third in overall sales, it is currently last with a deficit of $25729.3563. Washington, which has a deep purple color, climbs up to third place with a $33402.6517 profit. New York and California, in particular, have profits that are nearly same and far higher than the others (more than $74 k).

We can see in Figure 32 that states in eastern region, some of the southern region and some of the central region have purplish orange color indicating higher total profit compared to the rest.

Multivariate Graph

(a) Heatmap of the Numeric Data

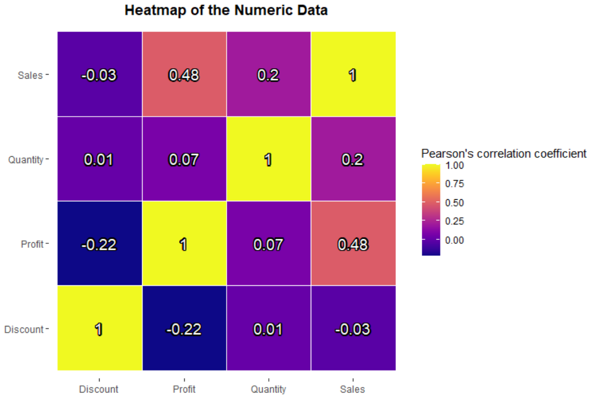


Figure 33. Sub-Category-Segment Data Bar Plot

By simplifying all the process: Sales data should be strongly correlated to Quantity data and Profit data positively and Discount data negatively; Quantity data should be strongly corelated to Profit data positively and may have some correlation with Discount data to some extent; And Profit data should be strongly corelated to Discount data negatively.

But in fact, Figure 33 shows that Sales data has a little to no correlation to Discount data, moderate correlation to Profit data and weak correlation to Quantity data; Quantity data has little to no correlation to Discount data and Profit data; and Profit data is weakly correlated to Discount data negatively.

The major deviation to the simple but reasonable logic may be caused by several reasons, such as:

1. To my knowledge, different state has different tax regulation which may or may not affect the correlation between Sales data and Profit data and with the other numerical data. Furthermore, the difference of purchasing power parity from state to state or city to city may cause the vary in price.
2. Different sub-category can cause the vary in price, e.g., the price of paper is not comparable to the price of phone.

In the end, further investigation and advanced analysis is needed to look at this problem.

(b) Profit, Sales, Segment, and Category

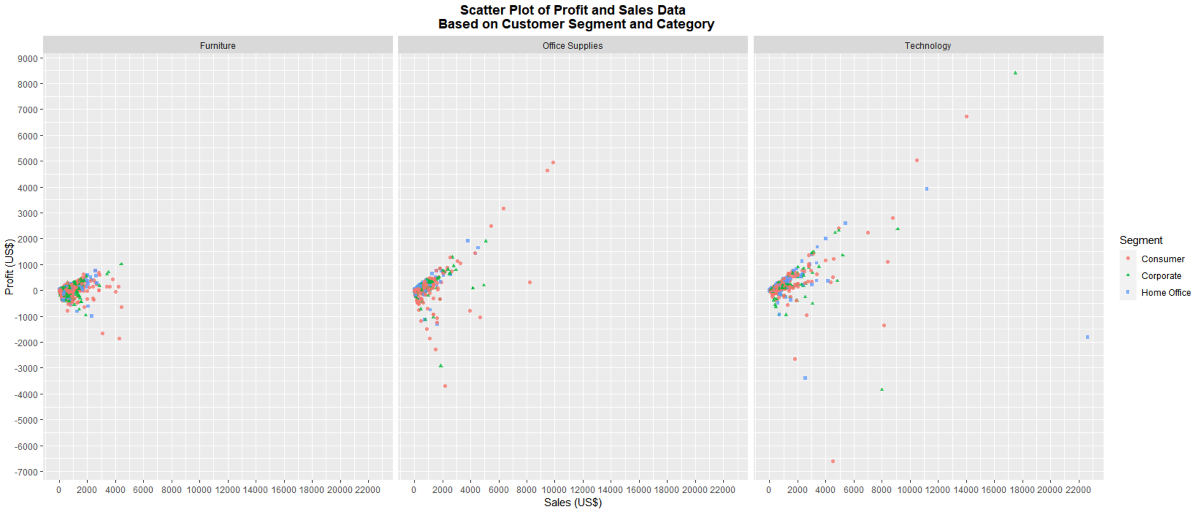


Figure 34. Scatter Plot of Profit and Sales Data Based on Customer Segment and Category

From the scatter plot above we can see that data point in Furniture category is more collected than the other two for the reason that there are no “wild” data in the plot. There are consumer data point that vary a little far but that’s it. It also easy to see the boundaries of the cluster in Furniture category. Most of the data point vary with the sales from the range $0 until $3000 and profit from the range -$1000 until $1000.

On the other hand, Technology category has the least collected plot on the grounds that it has a lot of outliers which some become the min or max value of the Sales or Profit Data. Most of the data point vary with the sales from the range $0 until $3000 and profit from the range -$600 until $800.

Office Supplies category has no outstanding feature to interpret. Most of the data point vary with the sales from the range $0 until $3000 and profit from the range -$600 until $1000.

(c) Segment, Sub-Category, and Quantity

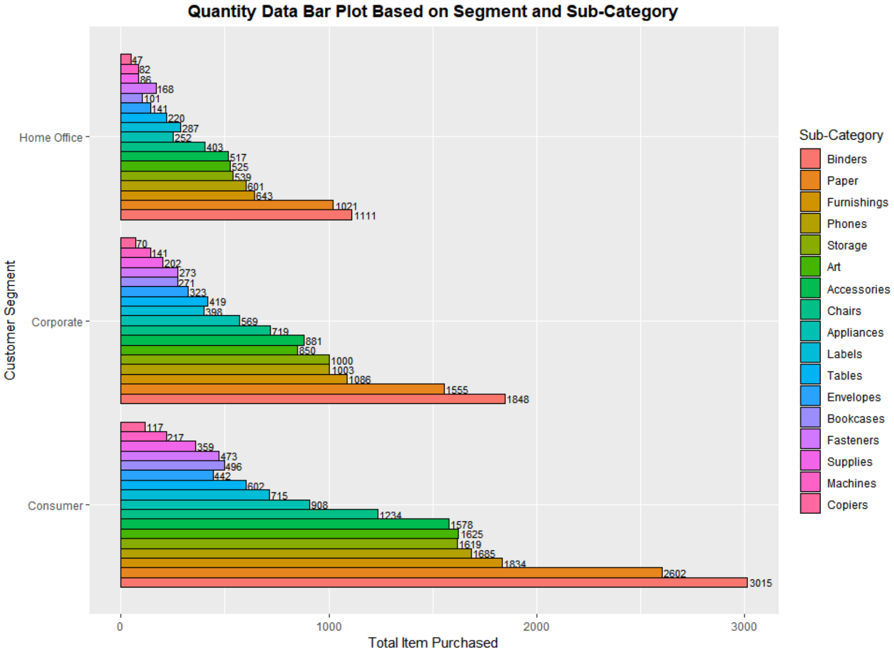


Figure 35. Quantity Data Bar Plot Based on Segment and Sub-Category

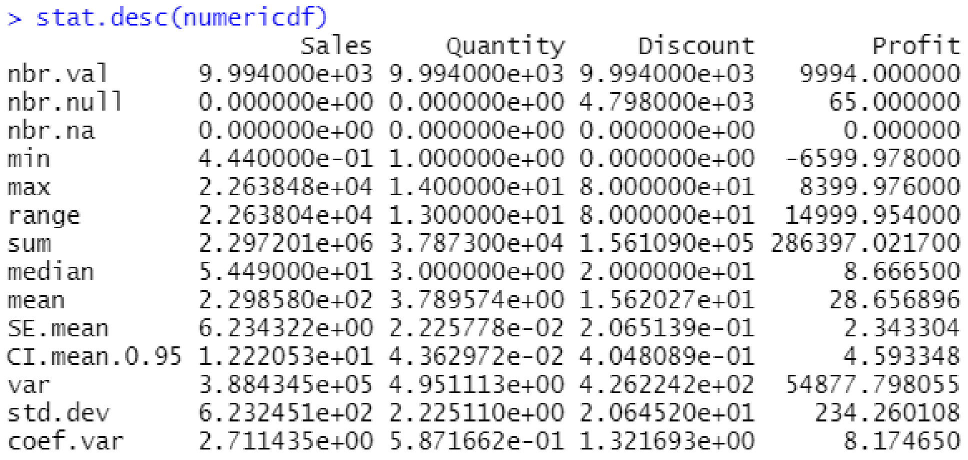
It seems that there are some irregularities of the order of the bar in each segment. The bar is sorted so each sub-category is appeared in the same way as Figure 24.

In “Home Office” segment, there are more “Fasteners” sold than “Bookcases” or “Envelopes”. Also, there are more “Labels” sold than “Appliances” which is also happening in “Corporate” segment. This may be caused by the bigger margin in the “Consumer” segment, since consumers need more appliance than label resulting the frequency of “Appliances” to be higher than the frequency of “Labels” in Figure 24. The number of “Machines” sold is relatively close to the number of “Supplies” sold, as well as “Art” and “Accessories” in “Home Office” segment. The number of “Paper” and “Binder” sold is exceptionally high compared to the other sub-category which also happens in the other segment.

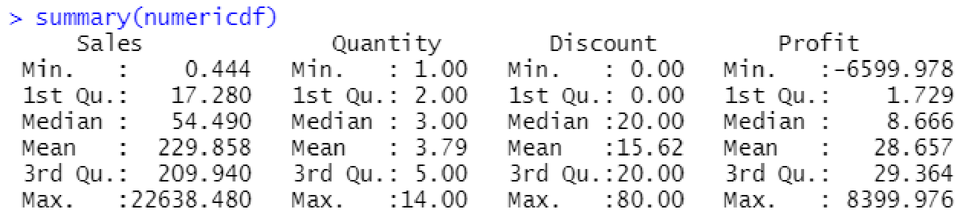
In “Corporate” segment, the number of “Accessories” sold is higher than the number of “Art” sold. Beside that, the number of items sold of “Fasteners” and “Bookcases”, and “Phones” and “Storage” is relatively close.

In “Consumer” segment, there are more “Fasteners” and “Bookcases” sold than the number of “Envelopes” sold. There also more “Art” than “Storage” sold.

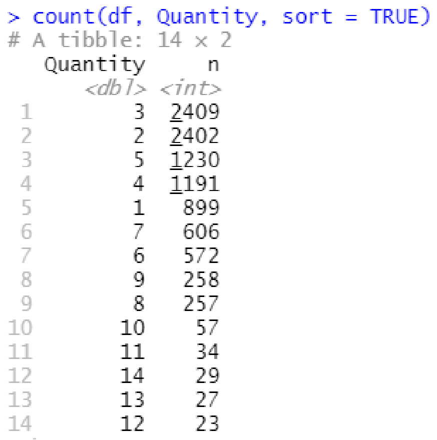
Summary Statistics and Interpretation



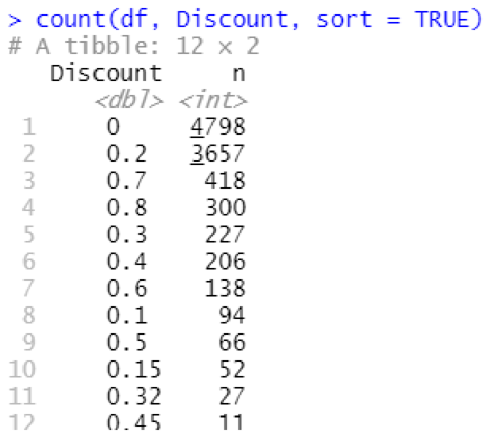
```
> stat.desc(numericdf)
                    Sales     Quantity     Discount         Profit
nbr.val      9.994000e+03 9.994000e+03 9.994000e+03    9994.000000
nbr.null     0.000000e+00 0.000000e+00 4.798000e+03      65.000000
nbr.na       0.000000e+00 0.000000e+00 0.000000e+00       0.000000
min          4.440000e-01 1.000000e+00 0.000000e+00   -6599.978000
max          2.263848e+04 1.400000e+01 8.000000e+01    8399.976000
range        2.263804e+04 1.300000e+01 8.000000e+01   14999.954000
sum          2.297201e+06 3.787300e+04 1.561090e+05  286397.021700
median       5.449000e+01 3.000000e+00 2.000000e+01       8.666500
mean         2.298580e+02 3.789574e+00 1.562027e+01      28.656896
SE.mean      6.234322e+00 2.225778e-02 2.065139e-01       2.343304
CI.mean.0.95 1.222053e+01 4.362972e-02 4.048089e-01       4.593348
var          3.884345e+05 4.951113e+00 4.262242e+02   54877.798055
std.dev      6.232451e+02 2.225110e+00 2.064520e+01     234.260108
coef.var     2.711435e+00 5.871662e-01 1.321693e+00       8.174650
```



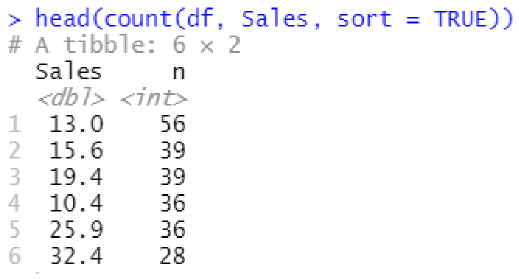
```
> summary(numericdf)
     Sales              Quantity        Discount         Profit
 Min.   :    0.444   Min.   : 1.00   Min.   : 0.00   Min.   :-6599.978
 1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.: 0.00   1st Qu.:    1.729
 Median :   54.490   Median : 3.00   Median :20.00   Median :    8.666
 Mean   :  229.858   Mean   : 3.79   Mean   :15.62   Mean   :   28.657
 3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:20.00   3rd Qu.:   29.364
 Max.   :22638.480   Max.   :14.00   Max.   :80.00   Max.   : 8399.976
```



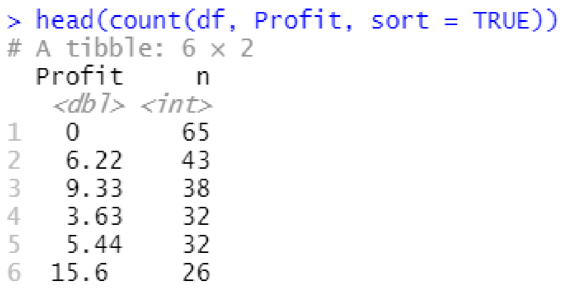
```
> count(df, Quantity, sort = TRUE)
# A tibble: 14 x 2
   Quantity     n
      <dbl> <int>
 1        3  2409
 2        2  2402
 3        5  1230
 4        4  1191
 5        1   899
 6        7   606
 7        6   572
 8        9   258
 9        8   257
10       10    57
11       11    34
12       14    29
13       13    27
14       12    23
```



```
> count(df, Discount, sort = TRUE)
# A tibble: 12 x 2
   Discount     n
      <dbl> <int>
 1     0     4798
 2     0.2   3657
 3     0.7    418
 4     0.8    300
 5     0.3    227
 6     0.4    206
 7     0.6    138
 8     0.1     94
 9     0.5     66
10     0.15    52
11     0.32    27
12     0.45    11
```



```
> head(count(df, Sales, sort = TRUE))
# A tibble: 6 x 2
  Sales     n
  <dbl> <int>
1  13.0    56
2  15.6    39
3  19.4    39
4  10.4    36
5  25.9    36
6  32.4    28
```



```
> head(count(df, Profit, sort = TRUE))
# A tibble: 6 x 2
  Profit     n
   <dbl> <int>
1   0       65
2   6.22    43
3   9.33    38
4   3.63    32
5   5.44    32
6  15.6     26
```

The important and noteworthy interpretations are all highlighted.

(a) Number of Values

Number of values (indicated by “nbr.val” in stat.desc()) is the total number of observations in the dataset. Each variable has exactly 9994 observations.

(b) Number of Null Values

Number of null values (indicated by “nbr.null” in stat.desc()) is the total number of observations in dataset that hold NULL (categorical) or 0 (numerical) value.

Sales variable and Quantity variable have no null value. However, Discount variable and Profit variable contains some null values (a.k.a., no discount / profit), especially Discount variable which almost 50% of the total number of observations are null value.

(c) Number of Missing Values

Number of missing values (indicated by “nbr.na” in stat.desc()) is the total number of observations in dataset that hold missing value. It is different from nbr.null which is known for holding null value rather than unknown. There is no missing value in this dataset.

(d) Minimum Value (min)

Minimum value (indicated by “min” in stat.desc() or “Min.” in summary()) is the lowest value of all the observations.

The lowest sale is 44.4¢ (approximately Rp7000), which is quite low for a transaction. The lowest number of items purchased (Quantity) is 1, which is also the lowest possible number of items purchased. The lowest discount given is 0%, as is variable Discount has null values, this is quite expected. The lowest “profit” is -$6599.978. This is a major loss for the store.

(e) Maximum Value

Maximum Value (indicated by “max” in stat.desc() or “Max.” in summary()) is the highest value of all the observations.

The highest sale is $22638.48 (around 50000 times as much as the lowest sale). It is an exceptionally big sale compared to the other sale. The highest quantity of item purchased is 14. The maximum of discount given is 80%, it is overly high and nearly impossible to see in daily practice. The maximum profit is $8399.976.

(f) Range

Range (indicated by “range” in stat.desc()) is the difference between the maximum value and the minimum value (a.k.a., max – min).

The range of Sales variable is $22638.04, the range of Quantity variable is 13, the range of Discount variable is 80%, and the range of Profit variable is $14999.954.

(g) Q1

Quartile is a way to divide the number of total observations into four fairly equal size (Hence, get the name of quarter – fourth). Q1 / 1st quartile (indicated by “1st Qu.” in summary()) is a value in which roughly 25% of the data value is less than or equal to.

The Q1 value of Sales variable is $17.28. This is quite expected value for Sales variable. The Q1 value of Quantity is 2. Again, this is quite expected value since we already know that the value distribution of Quantity data is positively skewed. The Q1 value of Discount variable is 0%. The Q1 value of Profit variable is $1.729. As we have seen in Figure 11, there are quite number of losses in the observations, therefore it doesn't come as a surprise to get such a low Q1 value.

(h) Median / Q2

Median / Q2 / 2nd quartile (indicated by "median" in stat.desc() or "Median" in summary()) is a value in which 50% of the data value is less than or equal to.

The median value of Sales variable is $54.490. It is considerably high value in IDR due to low purchasing power parity but may be considered as normal amount for US citizen. The median value of Quantity variable is 3, which is reasonable. The median value of Discount variable is 20%, which is also reasonable. The median value of Profit variable is $8.666. It is reasonable profit for low-medium purchasing.

(i) Q3

Q3 / 3rd quartile (indicated by "3rd Qu." In summary()) is a value in which 25% of the data value is more than or equal to.

The Q3 value of Sales variable is $209.94. It is increasing exponentially by comparing it to the min, Q1, and Q2 values. The Q3 value of Quantity variable is 3, which is to no surprise since we already have looked over the histogram before. The Q3 value of Discount variable is 20%. The Q3 value of Profit variable is $29.364, which is quite low if compared the Q3 value of Sales variable (only one-seventh or 14%).

(j) Sum

Sum (indicated by "sum" in stat.desc()) is the total value of all observations in the dataset.

The sum of Sales variable is $2297201. The sum of Quantity variable is 37873, which is the number indicating the total number of items purchased. The sum of Discount variable is 156109%. The sum of Profit variable is $286397.0217. This could imply that the store-type business in US is growing because, despite a significant number of losses, there is more profit than loss in the end.

(k) Mean

Mean (indicated by "mean" in stat.desc() or "Mean" in summary()) is the average value of the data (sum divided by number of observation).

The mean value of Sales variable is $229.858, which is higher than Q3 value due the appearance of a lot of extreme values / outliers. The mean value of Quantity variable is 3.79. The mean value of Discount variable is 15.62%. The mean value of Profit variable is $28.657. This could imply that the store-type business in US is growing in moderate rate.

(l) Standard Deviation

Standard deviation (indicated by "std.dev" in stat.desc()) is a measurement of how "differing" one data might be to another. It is the square root of variance.

The standard deviation of Sales Data is $623.2451, while the standard deviation of Sales data is only $234.260108. It means that Sales Data is more differ than Sales data, which is previously known by logic. It also quite a high number but is inevitable because of the existence of numerous outliers. Therefore, it needs some adjustment in calculation to see how differ it really is. Meanwhile, the standard deviation of Quantity data is 2.22511. Discount data has standard deviation of 20.6452%.

(m) Variance

Variance (indicated by "var" in stat.desc()) is a measurement of how "differing" one data might be to another. It is the square of standard deviation.

The variance of Sales Data is $623.2451, while the standard deviation of Sales data is only $234.260108. The variance of Quantity data is 2.22511. Discount data has variance of 20.6452%. The variance of Profit Data is. The interpretation is identical standard deviation.

(n) Standard Error of the Mean

Standard error of the mean (abbreviated as standard error) (indicated by "SE.mean" in stat.desc()) is an expression of how likely that the sample means vary from sample to sample. It is proportional to the standard deviation in the population and inversely proportional to the square root of the sample size. It also described as standard deviation of all possible sample means. It is important to remember that this dataset contain a sample not population.

The standard error of Sales data is 6.234322$, Quantity data is 0.02225778, discount data is 0.2065139%, 2.343304$. Considering the value of the other numerical measurement, the standard error of all the numerical data in this data set is low.

(o) Confidence Interval of the mean with p = 0.95

Confidence Interval of the mean (indicated by "CI.mean" in stat.desc()) is just like its name, is a prediction of the interval of where the population mean might be by looking on the sample mean while p-value (indicated by "0.95" after "CI.mean" in stat.desc()) is the accuracy level of the calculation.

The CI of Sales data is $12.22053, which means the population mean (μ) is somewhere between $217.63747 and $242.07853 with the "confidence" of 95%. It can be said "You are 95% confident that the mean Sales of population is somewhere between interval $217.63747 and $242.07853 is correct".

The CI of Quantity data is 0.04362972, which means $3.74637028 \leq \mu_{Quantity} \leq 3.83362972$ with the "confidence" of 95%.

The CI of Discount data is 0.4048089%, which means $15.2151911\% \leq \mu_{Discount} \leq 16.0248089$ with the "confidence" of 95%.

The CI of Profit data is $4.593348, which means $\$24.063652 \leq \mu_{Profit} \leq \$33.250348$ with the "confidence" of 95%.

(p) Coefficient of Variation

Coefficient of Variation (Indicated by "coef.var" in stat.desc()) is relative measure that compare the standard deviation to the mean.

The CV of Sales data is 2.711435%, Quantity data is 0.5871662%, Discount data is 1.321693% and Profit data is 8.17465%. This means that Profit data is the most variated data among the others, Sales data is second most variated, Discount data is the third, and Quantity data is the least variated data.

(q) Mode

Mode is the value that has the utmost occurrence frequency (Calculated using count(sort = TRUE)). The mode of Sales variable is $13.0, with the frequency of 56 (0.6% of total data), mode of Quantity variable is 3, with the frequency of 2409 (24% of total data), mode of Discount variable is 0%, with the frequency of 4798 (48% of total data), and mode of Profit variable is $0, with the frequency of 65 (0.7% of total data).