Data Dive — Sampling and Drawing Conclusions

Amritha Prakash

2023-09-15

ASSIGNMENT 3

Read the Data

```r
# Load tidyverse
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)

Superstore_data=read.csv("SampleSuperstore_final.csv")
head(Superstore_data)
```

```
##        Ship.Mode   Segment       Country           City      State Postal.Code
## 1   Second Class  Consumer United States       Henderson   Kentucky       42420
## 2   Second Class  Consumer United States       Henderson   Kentucky       42420
## 3   Second Class Corporate United States     Los Angeles California       90036
## 4 Standard Class  Consumer United States Fort Lauderdale    Florida       33311
## 5 Standard Class  Consumer United States Fort Lauderdale    Florida       33311
## 6 Standard Class  Consumer United States     Los Angeles California       90032
##   Region        Category Sub.Category    Sales Quantity Discount    Profit
## 1  South       Furniture    Bookcases 261.9600        2     0.00   41.9136
## 2  South       Furniture       Chairs 731.9400        3     0.00  219.5820
## 3   West Office Supplies       Labels  14.6200        2     0.00    6.8714
## 4  South       Furniture       Tables 957.5775        5     0.45 -383.0310
## 5  South Office Supplies      Storage  22.3680        2     0.20    2.5164
## 6   West       Furniture  Furnishings  48.8600        7     0.00   14.1694
```

Task(s)

(The purpose of this week's data dive is for you to think critically about what might go wrong when it comes time to make conclusions about your data.)

- Part 1: A collection of 5-10 random samples of data (with replacement) from at least 6 columns of data
  - Each subsample should be as long as roughly 50% percent of your data.We are simulating the act of collecting data from a population where the "population" is represented by the data set you already have.
  - Store each sample set in a separate data frame (e.g., df_i might contain m rows from columns 1-6)
  - These subsamples should include both categorical and continuous (numeric) data
- Part 2: Scrutinize these subsamples.
  - How different are they?
  - What would you have called an anomaly in one sub-sample that you wouldn't in another?
  - Are there aspects of the data that are consistent among all sub-samples?
- Part 3: Consider how this investigation affects how you might draw conclusions about the data in the future.

---

1. Part 1 - Collecting 5/6 random samples of data (with replacement) from at least 6 columns of data

Population : -

```
# Population - count :
# Rows of data in the data set -
nrow(Superstore_data)
```

```
## [1] 9994
```

Sample size :-

```
# Sample - should be as long as roughly 50% percent of your data.
# 50 % of 9994
sample_size <- 0.5 * 9994
sample_size
```

```
## [1] 4997
```

- sample 1

```
set.seed(10)
df_sample_1 <- Superstore_data |> sample_frac(0.5, replace = TRUE) |> select("Ship.Mode","Segment","Reg
nrow(df_sample_1)
```

```
## [1] 4997
```

```
Random 20 rows from 1st sample
```

```r
df_sample_1 |> sample_n(20)
```

```
##         Ship.Mode     Segment  Region         State        Category
## 1  Standard Class Home Office   South   Mississippi       Furniture
## 2  Standard Class   Corporate Central      Michigan       Furniture
## 3  Standard Class   Corporate    West       Arizona      Technology
## 4  Standard Class    Consumer    West    California      Technology
## 5  Standard Class    Consumer    West    Washington Office Supplies
## 6  Standard Class    Consumer    West    California      Technology
## 7        Same Day    Consumer    East      New York Office Supplies
## 8  Standard Class    Consumer Central      Michigan Office Supplies
## 9    Second Class Home Office    East      New York       Furniture
## 10   Second Class    Consumer Central         Texas       Furniture
## 11 Standard Class    Consumer    West       Arizona Office Supplies
## 12   Second Class    Consumer    West    California Office Supplies
## 13 Standard Class   Corporate Central      Michigan Office Supplies
## 14 Standard Class   Corporate Central      Illinois Office Supplies
## 15 Standard Class   Corporate   South      Virginia Office Supplies
## 16 Standard Class    Consumer    East Massachusetts Office Supplies
## 17 Standard Class    Consumer    West    California      Technology
## 18 Standard Class   Corporate    East      New York Office Supplies
## 19    First Class    Consumer Central      Illinois Office Supplies
## 20 Standard Class   Corporate   South       Georgia      Technology
##    Sub.Category    Sales    Profit Discount
## 1   Furnishings   18.920    7.3788      0.0
## 2        Tables  801.960  200.4900      0.0
## 3   Accessories   62.352  -10.9116      0.2
## 4   Accessories  179.950   37.7895      0.0
## 5         Paper   41.860   18.8370      0.0
## 6        Phones  470.376   52.9173      0.2
## 7     Envelopes   68.460   31.4916      0.0
## 8    Appliances  283.140   72.3580      0.1
## 9   Furnishings   82.640    7.4376      0.0
## 10  Furnishings   30.560  -19.8640      0.6
## 11      Binders   19.194  -12.7960      0.7
## 12      Binders 1016.792  381.2970      0.2
## 13   Appliances  207.144   48.3336      0.1
## 14        Paper   23.520    8.5260      0.2
## 15      Storage   67.900    0.6790      0.0
## 16        Paper   19.440    9.3312      0.0
## 17       Phones  333.576   25.0182      0.2
## 18        Paper   68.520   31.5192      0.0
## 19      Binders   96.784 -145.1760      0.8
## 20       Phones  206.100   55.6470      0.0
```

- sample 2

```r
set.seed(50)
df_sample_2 <- Superstore_data |> sample_frac(0.5, replace = TRUE)|> select("Ship.Mode","Segment","Regi
nrow(df_sample_2)
```

```
## [1] 4997
```

Random 20 rows from 2st sample

```
df_sample_2 |> sample_n(20)
```

```
##          Ship.Mode     Segment  Region        State        Category Sub.Category
## 1  Standard Class    Consumer Central        Texas Office Supplies      Binders
## 2  Standard Class   Corporate    West   California      Technology  Accessories
## 3    Second Class    Consumer Central     Illinois       Furniture       Chairs
## 4  Standard Class Home Office    East Pennsylvania Office Supplies          Art
## 5  Standard Class   Corporate    West   Washington       Furniture  Furnishings
## 6  Standard Class   Corporate Central        Texas Office Supplies        Paper
## 7  Standard Class    Consumer    West   California Office Supplies      Storage
## 8  Standard Class   Corporate   South      Georgia       Furniture  Furnishings
## 9    Second Class Home Office   South     Kentucky      Technology  Accessories
## 10 Standard Class    Consumer    West   California Office Supplies   Appliances
## 11   Second Class    Consumer    East         Ohio Office Supplies    Envelopes
## 12 Standard Class   Corporate Central        Texas Office Supplies   Appliances
## 13 Standard Class   Corporate   South     Kentucky      Technology       Phones
## 14 Standard Class   Corporate Central        Texas      Technology       Phones
## 15    First Class    Consumer    East Pennsylvania Office Supplies   Appliances
## 16 Standard Class    Consumer    West     Colorado Office Supplies      Binders
## 17   Second Class Home Office    East     New York Office Supplies        Paper
## 18 Standard Class   Corporate    West   California       Furniture    Bookcases
## 19    First Class    Consumer    West   California Office Supplies          Art
## 20    First Class    Consumer Central        Texas Office Supplies      Storage
##       Sales    Profit Discount
## 1     1.248   -1.9344     0.80
## 2    27.880    3.9032     0.00
## 3   602.651 -163.5767     0.30
## 4     5.248    0.5904     0.20
## 5   137.540   55.0160     0.00
## 6    15.552    5.6376     0.20
## 7   139.040   38.9312     0.00
## 8    39.920   11.1776     0.00
## 9    18.000    3.2400     0.00
## 10  160.960   48.2880     0.00
## 11   46.720   17.5200     0.20
## 12   34.176  -87.1488     0.80
## 13   36.990    9.9873     0.00
## 14   21.072    1.5804     0.20
## 15  434.352   43.4352     0.20
## 16    8.736   -6.1152     0.70
## 17   30.440   14.3068     0.00
## 18  308.499  -18.1470     0.15
## 19   16.020    4.4856     0.00
## 20   18.160    1.8160     0.20
```

- sample 3

```
set.seed(100)
df_sample_3 <- Superstore_data |> sample_frac(0.5, replace = TRUE)|> select("Ship.Mode","Segment","Regi
nrow(df_sample_3)
```

```
## [1] 4997
```

Random 20 rows from 3rd sample

```
df_sample_3 |> sample_n(20)
```

```
##         Ship.Mode     Segment  Region        State        Category Sub.Category
## 1  Standard Class Home Office Central     Michigan       Furniture  Furnishings
## 2    Second Class    Consumer    West         Utah      Technology       Phones
## 3  Standard Class    Consumer Central        Texas Office Supplies   Appliances
## 4  Standard Class    Consumer Central        Texas       Furniture  Furnishings
## 5        Same Day Home Office Central     Illinois      Technology       Phones
## 6  Standard Class   Corporate Central     Illinois      Technology       Phones
## 7  Standard Class    Consumer    East Pennsylvania Office Supplies    Fasteners
## 8     First Class    Consumer    East  Connecticut Office Supplies        Paper
## 9  Standard Class    Consumer    East     New York Office Supplies      Binders
## 10 Standard Class Home Office    West   Washington Office Supplies      Binders
## 11 Standard Class   Corporate    East   New Jersey      Technology     Machines
## 12 Standard Class Home Office Central    Minnesota      Technology      Copiers
## 13       Same Day    Consumer    West   Washington      Technology  Accessories
## 14 Standard Class    Consumer    West   Washington      Technology      Copiers
## 15   Second Class    Consumer    West   Washington Office Supplies       Labels
## 16    First Class    Consumer    East     New York Office Supplies        Paper
## 17 Standard Class    Consumer    West      Arizona Office Supplies       Labels
## 18 Standard Class Home Office    West     Colorado Office Supplies          Art
## 19 Standard Class   Corporate   South      Alabama Office Supplies        Paper
## 20 Standard Class Home Office    West   California      Technology       Phones
##       Sales    Profit Discount
## 1    33.480    8.7048      0.0
## 2   399.960   34.9965      0.2
## 3    58.924 -153.2024      0.8
## 4    66.112  -84.2928      0.6
## 5    34.360   -7.3015      0.2
## 6   239.976   26.9973      0.2
## 7    10.584   -2.3814      0.2
## 8    27.120   12.4752      0.0
## 9    49.536   17.3376      0.2
## 10    6.096    2.1336      0.2
## 11 9099.930 2365.9818      0.0
## 12  549.990  274.9950      0.0
## 13  118.000   20.0600      0.0
## 14  999.980  449.9910      0.0
## 15   87.710   41.2237      0.0
## 16   46.760   22.4448      0.0
## 17    5.040    1.7640      0.2
## 18   14.576    2.3686      0.2
## 19   23.920   11.7208      0.0
## 20  271.960   27.1960      0.2
```

- sample 4

```r
set.seed(120)
df_sample_4 <- Superstore_data |> sample_frac(0.5, replace = TRUE) |> select("Ship.Mode","Segment","Reg
nrow(df_sample_4)
```

```
## [1] 4997
```

Random 20 rows from 4th sample

```r
df_sample_4 |> sample_n(20)
```

```
##         Ship.Mode     Segment  Region        State        Category Sub.Category
## 1  Standard Class   Corporate    West       Oregon       Furniture       Tables
## 2  Standard Class Home Office    East     New York Office Supplies        Paper
## 3        Same Day    Consumer Central        Texas       Furniture  Furnishings
## 4  Standard Class    Consumer   South  Mississippi Office Supplies   Appliances
## 5  Standard Class   Corporate Central        Texas Office Supplies        Paper
## 6  Standard Class    Consumer    West   California Office Supplies   Appliances
## 7  Standard Class   Corporate    West       Nevada Office Supplies          Art
## 8        Same Day    Consumer    West   California Office Supplies      Storage
## 9  Standard Class   Corporate Central        Texas      Technology     Machines
## 10 Standard Class    Consumer Central     Illinois Office Supplies        Paper
## 11 Standard Class   Corporate    East     New York      Technology       Phones
## 12   Second Class Home Office   South     Virginia       Furniture  Furnishings
## 13    First Class Home Office    East Pennsylvania Office Supplies    Fasteners
## 14       Same Day   Corporate    West   California Office Supplies    Fasteners
## 15   Second Class    Consumer Central        Texas      Technology  Accessories
## 16    First Class Home Office   South    Tennessee Office Supplies          Art
## 17 Standard Class    Consumer    East     New York Office Supplies      Binders
## 18 Standard Class    Consumer Central    Minnesota Office Supplies          Art
## 19    First Class   Corporate    West   Washington Office Supplies      Binders
## 20 Standard Class   Corporate    West   California Office Supplies        Paper
##        Sales    Profit Discount
## 1   177.225 -120.5130      0.5
## 2    99.900   47.9520      0.0
## 3    25.160  -11.3220      0.6
## 4   320.640   89.7792      0.0
## 5    98.376   35.6613      0.2
## 6     8.670    2.3409      0.0
## 7     3.640    1.6380      0.0
## 8    31.440    8.4888      0.0
## 9   559.710 -121.2705      0.4
## 10  143.856   48.5514      0.2
## 11  307.980   89.3142      0.0
## 12   47.980   11.0354      0.0
## 13    3.168   -0.7128      0.2
## 14   17.900    8.9500      0.0
## 15 1399.944   52.4979      0.2
## 16   67.920    6.7920      0.2
## 17  106.344   37.2204      0.2
## 18    8.800    2.5520      0.0
## 19  895.920  302.3730      0.2
## 20   38.880   18.6624      0.0
```

- sample 5

```
set.seed(150)
df_sample_5 <- Superstore_data |> sample_frac(0.5, replace = TRUE)|> select("Ship.Mode","Segment","Regi
nrow(df_sample_5)
```

```
## [1] 4997
```

Random 20 rows from 5th sample

```
df_sample_5 |> sample_n(20)
```

```
##         Ship.Mode     Segment  Region          State        Category
## 1  Standard Class   Corporate Central          Texas Office Supplies
## 2  Standard Class    Consumer    West     California Office Supplies
## 3     First Class   Corporate    East       New York Office Supplies
## 4  Standard Class Home Office Central       Michigan      Technology
## 5  Standard Class    Consumer Central       Oklahoma Office Supplies
## 6    Second Class   Corporate    East       New York Office Supplies
## 7  Standard Class    Consumer    West     California       Furniture
## 8  Standard Class Home Office Central      Minnesota Office Supplies
## 9  Standard Class Home Office   South        Florida       Furniture
## 10 Standard Class   Corporate   South        Florida      Technology
## 11 Standard Class    Consumer    East       New York       Furniture
## 12       Same Day   Corporate    East    Connecticut Office Supplies
## 13       Same Day    Consumer    East           Ohio       Furniture
## 14 Standard Class   Corporate    West     Washington      Technology
## 15 Standard Class   Corporate   South South Carolina Office Supplies
## 16 Standard Class   Corporate Central        Indiana Office Supplies
## 17   Second Class    Consumer    West     California       Furniture
## 18 Standard Class   Corporate Central          Texas Office Supplies
## 19    First Class   Corporate    East    Connecticut Office Supplies
## 20 Standard Class    Consumer    West        Arizona       Furniture
##    Sub.Category    Sales   Profit Discount
## 1       Storage   32.232   2.4174      0.2
## 2       Storage  777.210  54.4047      0.0
## 3       Storage   83.920  20.1408      0.0
## 4   Accessories 1928.780 829.3754      0.0
## 5        Labels   14.620   6.8714      0.0
## 6        Labels    8.670   4.0749      0.0
## 7        Chairs  194.352  19.4352      0.2
## 8           Art   29.790  12.5118      0.0
## 9   Furnishings  258.072   0.0000      0.2
## 10       Phones  100.792  10.0792      0.2
## 11  Furnishings   28.440  11.3760      0.0
## 12      Binders   23.200  10.4400      0.0
## 13  Furnishings   51.264   7.6896      0.2
## 14       Phones   71.960  25.1860      0.2
## 15      Storage  628.810  12.5762      0.0
## 16        Paper   14.940   7.3206      0.0
## 17  Furnishings   24.140   7.9662      0.0
```

```
## 18        Paper   36.288  12.7008      0.2
## 19     Supplies   30.690   7.9794      0.0
## 20  Furnishings  206.112  48.9516      0.2
```

- All these sub-samples contain both categorical and continuous (numeric) data.
- Check for replacement and if there are common data inbetween the samples

```
nrow(intersect(df_sample_1, df_sample_2))
```

```
## [1] 1568
```

```
-    1568 records are common between sample 1 and 2
```

```
nrow(intersect(df_sample_2, df_sample_3))
```

```
## [1] 1572
```

```
-    1572 records are common between sample 2 and 3
```

```
nrow(intersect(df_sample_3, df_sample_4))
```

```
## [1] 1547
```

```
-    1547 records are common between sample 3 and 4
```

```
nrow(intersect(df_sample_4, df_sample_5))
```

```
## [1] 1582
```

```
-    1582 records are common between sample 3 and 4
```

Part 2:- Scrutinize these sub-samples

1. Lets take into consideration Column - Segment :

- Segment :-
    1. check the various segments in each sample -
        - Sample 1 -

    ```
    count_df_sample_1 <- df_sample_1 |> group_by(Segment) |>
      summarise(total_count_segment=n(),
                .groups = 'drop')
    count_df_sample_1
    ```

```
## # A tibble: 3 x 2
##   Segment     total_count_segment
##   <chr>                     <int>
## 1 Consumer                   2596
## 2 Corporate                  1535
## 3 Home Office                 866
```

- Sample 2 -

```
count_df_sample_2 <- df_sample_2 |> group_by(Segment) |>
  summarise(total_count_segment=n(),
            .groups = 'drop')
count_df_sample_2
```

```
## # A tibble: 3 x 2
##   Segment     total_count_segment
##   <chr>                     <int>
## 1 Consumer                   2574
## 2 Corporate                  1515
## 3 Home Office                 908
```

- Sample 3 -

```
count_df_sample_3 <- df_sample_3 |> group_by(Segment) |>
  summarise(total_count_segment=n(),
            .groups = 'drop')
count_df_sample_3
```

```
## # A tibble: 3 x 2
##   Segment     total_count_segment
##   <chr>                     <int>
## 1 Consumer                   2560
## 2 Corporate                  1511
## 3 Home Office                 926
```

- Sample 4 -

```
count_df_sample_4 <- df_sample_4 |> group_by(Segment) |>
  summarise(total_count_segment=n(),
            .groups = 'drop')
count_df_sample_4
```

```
## # A tibble: 3 x 2
##   Segment     total_count_segment
##   <chr>                     <int>
## 1 Consumer                   2592
## 2 Corporate                  1487
## 3 Home Office                 918
```

- Sample 5 -

```r
count_df_sample_5 <- df_sample_5 |> group_by(Segment) |>
  summarise(total_count_segment=n(),
            .groups = 'drop')
count_df_sample_5
```

```
## # A tibble: 3 x 2
##   Segment     total_count_segment
##   <chr>                     <int>
## 1 Consumer                   2563
## 2 Corporate                  1576
## 3 Home Office                 858
```

- From all the above samples, for categorical value - SEGMENT
  - we see 3 types of Segment,the data is somewhat spread out and the count of "Home Office" in all samples is seen to be around 900.
  - This indicates that for the whole population the Home Office is the least purchased Segment.
  - This is same in the case of other 2 segments.Corporate (around 1500) and Consumer (around 2600) segments see a similar count on all samples.
  - This indicates that the data is spread evenly.

2. Lets take into consideration Column - Sales :

- Sales :-

  1. Mean Sales in each sample -

     - Sample 1 -
       Mean of Sales for sample 1 :-

```r
mean_Sample_1 <- df_sample_1 |> pluck("Sales") |> mean(na.rm=TRUE)
mean_Sample_1
```

```
## [1] 246.4534
```

```
-  Sample 2 - \
Mean of Sales for sample 2 :-
```

```r
mean_Sample_2 <- df_sample_2 |> pluck("Sales") |> mean(na.rm=TRUE)
mean_Sample_2
```

```
## [1] 221.7272
```

```
-  Sample 3 -
Mean of Sales for sample 3 :-
```

```r
mean_Sample_3 <- df_sample_3 |> pluck("Sales") |> mean(na.rm=TRUE)
mean_Sample_3
```

```
## [1] 255.5451
```

```
-  Sample 4 -
Mean of Sales for sample 4 :-
```

```
mean_Sample_4 <- df_sample_4 |> pluck("Sales") |> mean(na.rm=TRUE)
mean_Sample_4
```

```
## [1] 217.4075
```

```
-  Sample 5 -
Mean of Sales for sample 5 :-
```

```
mean_Sample_5 <- df_sample_5 |> pluck("Sales") |> mean(na.rm=TRUE)
mean_Sample_5
```

```
## [1] 228.1538
```

```
Mean_of_Sample_average <- mean(mean_Sample_1, mean_Sample_2, mean_Sample_3, mean_Sample_4, mean_Sample_5
Mean_of_Sample_average
```

```
## [1] 246.4534
```

- From all the above samples, for Continuous value - SALES
  - we see the mean of sales in each sample to be somewhat similar. The average of the same comes out to be 246.45. There are no anomalies observed there.

  2. Max Sales in each sample -

    - Sample 1 -
      Max of Sales for sample 1 :-

```
max_Sample_1 <- df_sample_1 |> pluck("Sales") |> max(na.rm=TRUE)
max_Sample_1
```

```
## [1] 22638.48
```

```
-  Sample 2 - \
Max of Sales for sample 2 :-
```

```
max_Sample_2 <- df_sample_2 |> pluck("Sales") |> max(na.rm=TRUE)
max_Sample_2
```

```
## [1] 9099.93
```

```
-  Sample 3 -
Max of Sales for sample 3 :-
```

```
max_Sample_3 <- df_sample_3 |> pluck("Sales") |> max(na.rm=TRUE)
max_Sample_3
```

```
## [1] 22638.48
```

```
-  Sample 4 -
Max of Sales for sample 4 :-
```

```r
max_Sample_4 <- df_sample_4 |> pluck("Sales") |> max(na.rm=TRUE)
max_Sample_4
```

```
## [1] 13999.96
```

```
-  Sample 5 -
Max of Sales for sample 5 :-
```

```r
max_Sample_5 <- df_sample_5 |> pluck("Sales") |> max(na.rm=TRUE)
max_Sample_5
```

```
## [1] 17499.95
```

- From all the above samples, for Continuous value - SALES
  - we see the maximum to be varying in each sample, 1st and 3rd sample have a record with Maximum sale of 22638.48, when compared to rest.
  - Almost all other samples have various max values. Sample 2 sees a max value of 9099.93, which wouldnt be true if we considered that sample alone. So that max value would have been incorrect if considered max for the entire population. And then there are other max values too in each of the other samples (like 17499.95, 13999.96)
  - Overall considering the samples the max seems to common value of 22638.48, which was observed in 2 samples.

  3. Minimum Sales in each sample -

     - Sample 1 -
       Min of Sales for sample 1 :-

```r
min_Sample_1 <- df_sample_1 |> pluck("Sales") |> min(na.rm=TRUE)
min_Sample_1
```

```
## [1] 0.444
```

```
-  Sample 2 - \
Min of Sales for sample 2 :-
```

```r
min_Sample_2 <- df_sample_2 |> pluck("Sales") |> min(na.rm=TRUE)
min_Sample_2
```

```
## [1] 0.836
```

```
-  Sample 3 -
Min of Sales for sample 3 :-
```

```r
min_Sample_3 <- df_sample_3 |> pluck("Sales") |> min(na.rm=TRUE)
min_Sample_3
```

```
## [1] 0.556
```

```
-  Sample 4 -
Min of Sales for sample 4 :-
```

```r
min_Sample_4 <- df_sample_4 |> pluck("Sales") |> min(na.rm=TRUE)
min_Sample_4
```

```
## [1] 0.852
```

```
-  Sample 5 -
Min of Sales for sample 5 :-
```

```r
min_Sample_5 <- df_sample_5 |> pluck("Sales") |> min(na.rm=TRUE)
min_Sample_5
```

```
## [1] 0.444
```

- From all the above samples, for Continuous value - SALES
  - we see the minimum to be varying in each sample. The least of all was 0.444 which is seen in 2 of the samples.
  - Rest of the samples have other minimum values like 0.852, 0.556, 0.836. But for the populatin seems like 0.444 the minimum value for sales.

3. Lets take into consideration Column - State :

- State :-

  1. check the various state in each sample -

     - Sample 1 -
       Top 10 states where the purchases were done the most.

```r
count_df_sample_1 <- df_sample_1 |> group_by(State) |>
  summarise(total_count_state=n(),
            .groups = 'drop') |>   arrange(desc(total_count_state))
head(count_df_sample_1, 10)
```

```
## # A tibble: 10 x 2
##    State          total_count_state
##    <chr>                      <int>
##  1 California                  1016
##  2 New York                     562
##  3 Texas                        503
##  4 Pennsylvania                 288
##  5 Washington                   263
##  6 Illinois                     259
##  7 Ohio                         258
##  8 Florida                      193
##  9 Michigan                     137
## 10 North Carolina               134
```

10 states where the purchases were done the least.

```
tail(count_df_sample_1,10)
```

```
## # A tibble: 10 x 2
##    State                total_count_state
##    <chr>                            <int>
##  1 Iowa                                10
##  2 Idaho                                7
##  3 South Dakota                         7
##  4 Vermont                              7
##  5 District of Columbia                 6
##  6 Montana                              5
##  7 North Dakota                         5
##  8 Maine                                4
##  9 West Virginia                        1
## 10 Wyoming                              1
```

- Sample 2 - \

Top 10 states where the purchases were done the most.

```
count_df_sample_2 <- df_sample_2 |> group_by(State) |>
  summarise(total_count_state=n(),
            .groups = 'drop') |>   arrange(desc(total_count_state))
head(count_df_sample_2, 10)
```

```
## # A tibble: 10 x 2
##    State          total_count_state
##    <chr>                      <int>
##  1 California                   986
##  2 New York                     592
##  3 Texas                        490
##  4 Pennsylvania                 299
##  5 Washington                   272
##  6 Illinois                     251
##  7 Ohio                         231
##  8 Florida                      178
##  9 Michigan                     127
## 10 North Carolina               125
```

10 states where the purchases were done the least.

```
tail(count_df_sample_2,10)
```

```
## # A tibble: 10 x 2
##    State                total_count_state
##    <chr>                            <int>
##  1 New Mexico                          12
##  2 Kansas                              11
##  3 Nevada                              10
##  4 North Dakota                         5
```

```
##  5 District of Columbia                 3
##  6 Vermont                              3
##  7 South Dakota                         2
##  8 West Virginia                        2
##  9 Wyoming                              2
## 10 Maine                                1
```

- Sample 3 - \

Top 10 states where the purchases were done the most.

```r
count_df_sample_3 <- df_sample_3 |> group_by(State) |>
  summarise(total_count_state=n(),
            .groups = 'drop') |>   arrange(desc(total_count_state))
head(count_df_sample_3, 10)
```

```
## # A tibble: 10 x 2
##    State          total_count_state
##    <chr>                      <int>
##  1 California                   994
##  2 New York                     541
##  3 Texas                        528
##  4 Pennsylvania                 289
##  5 Illinois                     261
##  6 Ohio                         249
##  7 Washington                   243
##  8 Florida                      197
##  9 Michigan                     134
## 10 North Carolina               130
```

10 states where the purchases were done the least.

```r
   tail(count_df_sample_3,10)
```

```
## # A tibble: 10 x 2
##    State                total_count_state
##    <chr>                            <int>
##  1 South Carolina                      13
##  2 South Dakota                        13
##  3 Iowa                                12
##  4 Kansas                              11
##  5 Montana                             10
##  6 Idaho                                8
##  7 Vermont                              8
##  8 District of Columbia                 5
##  9 Maine                                1
## 10 Wyoming                              1
```

```
-  Sample 4 - \
Top 10 states where the purchases were done the most.\
```

```r
count_df_sample_4 <- df_sample_4 |> group_by(State) |>
  summarise(total_count_state=n(),
            .groups = 'drop') |>   arrange(desc(total_count_state))
head(count_df_sample_4, 10)
```

```
## # A tibble: 10 x 2
##    State          total_count_state
##    <chr>                      <int>
##  1 California                   983
##  2 New York                     570
##  3 Texas                        492
##  4 Pennsylvania                 298
##  5 Illinois                     253
##  6 Washington                   236
##  7 Ohio                         233
##  8 Florida                      209
##  9 North Carolina               131
## 10 Virginia                     125
```

10 states where the purchases were done the least.\

```r
   tail(count_df_sample_4,10)
```

```
## # A tibble: 10 x 2
##    State                total_count_state
##    <chr>                            <int>
##  1 Kansas                              13
##  2 Vermont                             12
##  3 Montana                              8
##  4 Idaho                                7
##  5 South Dakota                         5
##  6 North Dakota                         4
##  7 West Virginia                        4
##  8 District of Columbia                 3
##  9 Maine                                3
## 10 Wyoming                              1
```

- Sample 5 - \
Top 10 states where the purchases were done the most.\

```r
count_df_sample_5 <- df_sample_5 |> group_by(State) |>
  summarise(total_count_state=n(),
            .groups = 'drop') |>   arrange(desc(total_count_state))
head(count_df_sample_5, 10)
```

```
## # A tibble: 10 x 2
##    State          total_count_state
##    <chr>                      <int>
##  1 California                  1027
##  2 New York                     546
##  3 Texas                        480
##  4 Pennsylvania                 270
```

```
##  5 Ohio                          260
##  6 Washington                    256
##  7 Illinois                      242
##  8 Florida                       187
##  9 North Carolina                135
## 10 Arizona                       119
```

10 states where the purchases were done the least.\

```
tail(count_df_sample_5,10)
```

```
## # A tibble: 10 x 2
##    State                total_count_state
##    <chr>                            <int>
##  1 Kansas                              11
##  2 Vermont                             10
##  3 Idaho                                9
##  4 Montana                              9
##  5 District of Columbia                 7
##  6 South Dakota                         6
##  7 North Dakota                         5
##  8 West Virginia                        3
##  9 Maine                                2
## 10 Wyoming                              1
```

- From all the above samples, for categorical value - STATE

  - We have calculated the top 10 states which purchase the products.It has been observed that top 5 states are always constant in each of the sample. Even the order is somewhat same.

    1. California
    2. New York
    3. Texas
    4. Pennsylvania
    5. Illinois/Washington/Ohio

    The remaining states(6 to 10) have certain similarities with the top 5 while occasionally changing their order. The top-performing states within each sample are, nevertheless, largely stable.

  - We have calculated the least 10 states which purchase the products.Here it can be observed that there are certain differences in the state with count at the bottom within the samples.

    1. Sample_1 has the following order for last 5 (Montana 5 > North Dakota 5 > Maine 4 > West Virginia 1 > Wyoming 1)

    2. Sample_2 has the following order for last 5 (Vermont 3 > South Dakota 2 > West Virginia 2 > Wyoming 2 > Maine 1)

    3. Sample_3 has the following order for last 5 (Idaho 8 > Vermont 8> District of Columbia 5 > Maine 1 > Wyoming 1)

    4. Sample_4 has the following order for last 5 (North Dakota 4 > West Virginia 4 > District of Columbia 3 > Maine 3 >Wyoming 1)

5. Sample_5 has the following order for last 5 (South Dakota 6 > North Dakota 5 > West Virginia 3 > Maine 2 > Wyoming 1)

From above samples and their last count on products purchased can see that, West Virginia, Maine and Wyoming is having the least count in all the samples. In the 2nd sample it is seen that Vermont is present in the bottom 5 for one of the samples, a case where the least of all in counts of products are purchased. Also, in 1st and 3rd sample can see Montana and Idaho state present in the sample of least products, which wasnt the case in other samples. But overall, certain states are seen to be similar in case of being the least. No major anomallies detected.

4. Lets take into consideration Column - Profit :

- Profit :-
    1. Mean Profit in each sample -
        - Sample 1 -
          Mean of Profit for sample 1 :-

```r
mean_Sample_1 <- df_sample_1 |> pluck("Profit") |> mean(na.rm=TRUE)
mean_Sample_1
```

```
## [1] 24.96756
```

```
-  Sample 2 - \
Mean of Profit for sample 2 :-
```

```r
mean_Sample_2 <- df_sample_2 |> pluck("Profit") |> mean(na.rm=TRUE)
mean_Sample_2
```

```
## [1] 29.75024
```

```
-  Sample 3 -
Mean of Profit for sample 3 :-
```

```r
mean_Sample_3 <- df_sample_3 |> pluck("Profit") |> mean(na.rm=TRUE)
mean_Sample_3
```

```
## [1] 36.66652
```

```
-  Sample 4 -
Mean of Profit for sample 4 :-
```

```r
mean_Sample_4 <- df_sample_4 |> pluck("Profit") |> mean(na.rm=TRUE)
mean_Sample_4
```

```
## [1] 25.37763
```

```
-  Sample 5 -
Mean of Profit for sample 5 :-
```

```r
mean_Sample_5 <- df_sample_5 |> pluck("Profit") |> mean(na.rm=TRUE)
mean_Sample_5
```

```
## [1] 30.72369
```

- From all the above samples, for Continuous value - PROFIT
  - we see the mean of Profit in each sample to be somewhat similar, within the range of 24 to 36.
  - We can say that the profit for all samples depicts that the populations also has a similar average on profit achieved through each sale.

2. Max Profit in each sample -

   - Sample 1 -
     Max of Profit for sample 1 :-

```
max_Sample_1 <- df_sample_1 |> pluck("Profit") |> max(na.rm=TRUE)
max_Sample_1
```

```
## [1] 6719.981
```

```
-  Sample 2 - \
Max of Profit for sample 2 :-
```

```
max_Sample_2 <- df_sample_2 |> pluck("Profit") |> max(na.rm=TRUE)
max_Sample_2
```

```
## [1] 2591.957
```

```
-  Sample 3 -
Max of Profit for sample 3 :-
```

```
max_Sample_3 <- df_sample_3 |> pluck("Profit") |> max(na.rm=TRUE)
max_Sample_3
```

```
## [1] 6719.981
```

```
-  Sample 4 -
Max of Profit for sample 4 :-
```

```
max_Sample_4 <- df_sample_4 |> pluck("Profit") |> max(na.rm=TRUE)
max_Sample_4
```

```
## [1] 6719.981
```

```
-  Sample 5 -
Max of Profit for sample 5 :-
```

```
max_Sample_5 <- df_sample_5 |> pluck("Profit") |> max(na.rm=TRUE)
max_Sample_5
```

```
## [1] 8399.976
```

- From all the above samples, for Continuous value - PROFIT
  - we see the maximum profit to be somewhat 8399.976.

- 3 samples, seems to have the max around 6719.981. This would not entirely claim to be an anomaly, but if that sample is considered alone then the assumption would be that products were not sold with a higher profit to the Superstore. But that is not the case.
- Sample 2 seems to have max of 2591.957, which would not be considered as a max of Profit, when compared to rest of the samples. Hence it can be considered as an anomaly.

3. Minimum Profit in each sample -

   - Sample 1 -
     Min of Profit for sample 1 :-

```
min_Sample_1 <- df_sample_1 |> pluck("Profit") |> min(na.rm=TRUE)
min_Sample_1
```

```
## [1] -6599.978
```

```
-  Sample 2 - \
Min of Profit for sample 2 :-
```

```
min_Sample_2 <- df_sample_2 |> pluck("Profit") |> min(na.rm=TRUE)
min_Sample_2
```

```
## [1] -3399.98
```

```
-  Sample 3 -
Min of Profit for sample 3 :-
```

```
min_Sample_3 <- df_sample_3 |> pluck("Profit") |> min(na.rm=TRUE)
min_Sample_3
```

```
## [1] -3839.99
```

```
-  Sample 4 -
Min of Profit for sample 4 :-
```

```
min_Sample_4 <- df_sample_4 |> pluck("Profit") |> min(na.rm=TRUE)
min_Sample_4
```

```
## [1] -3839.99
```

```
-  Sample 5 -
Min of Profit for sample 5 :-
```

```
min_Sample_5 <- df_sample_5 |> pluck("Profit") |> min(na.rm=TRUE)
min_Sample_5
```

```
## [1] -3839.99
```

- From all the above samples, for Continuous value min Profit or even Loss can be figured out
  - we see the minimum to be a loss in all of the samples.From those can incur, the products bought all over US are sold with a minimum loss of 6599.978. Negative indicates Loss, I believe.
  - Also, in one of the sample minimum loss is obtained to be around 3399.98.
  - Rest of the 3 samples have a common loss of 3839.99

5. Lets take into consideration Column - Region and Sales :

- Region and Quantity :-
  1. check the various Region and Sales in each sample

     – Sample 1 -

```
count_df_sample_1 <- df_sample_1 |> group_by(Region) |>
  summarise(total_max_region_sales=max(Sales),
            .groups = 'drop') |>
  arrange(desc(total_max_region_sales),.by_group= TRUE)
count_df_sample_1
```

```
## # A tibble: 4 x 2
##   Region  total_max_region_sales
##   <chr>                    <dbl>
## 1 South                   22638.
## 2 West                    14000.
## 3 East                    11200.
## 4 Central                  8160.
```

```
We can see that when grouping by Region and Segment, products bought in the southern region see the
```

```
-  Sample 2 -
```

```
count_df_sample_2 <- df_sample_2 |> group_by(Region) |>
  summarise(total_max_region_sales=max(Sales),
            .groups = 'drop') |>
  arrange(desc(total_max_region_sales),.by_group= TRUE)
count_df_sample_2
```

```
## # A tibble: 4 x 2
##   Region  total_max_region_sales
##   <chr>                    <dbl>
## 1 East                     9100.
## 2 West                     8188.
## 3 Central                  5444.
## 4 South                    3080
```

```
We can see that when grouping by Region and Segment, products bought in the Eastern region have pro
```

```
-  Sample 3 -


‘‘‘r
count_df_sample_3 <- df_sample_3 |> group_by(Region) |>
  summarise(total_max_region_sales=max(Sales),
            .groups = ’drop’) |>
  arrange(desc(total_max_region_sales),.by_group= TRUE)
count_df_sample_3
```

```
## # A tibble: 4 x 2
##   Region  total_max_region_sales
##   <chr>                    <dbl>
## 1 South                   22638.
## 2 West                    14000.
## 3 East                    11200.
## 4 Central                  9893.
```

```
From above grouping can see products from southern region have the highest sale.

-  Sample 4 -
```

```
count_df_sample_4 <- df_sample_4 |> group_by(Region) |>
  summarise(total_max_region_sales=max(Sales),
            .groups = 'drop') |>
  arrange(desc(total_max_region_sales),.by_group= TRUE)
count_df_sample_4
```

```
## # A tibble: 4 x 2
##   Region  total_max_region_sales
##   <chr>                    <dbl>
## 1 West                    14000.
## 2 East                    11200.
## 3 Central                  9893.
## 4 South                    8000.
```

```
From above grouping can see products from Western region have the highest sale. But southern is at


-  Sample 5 -
```

```
count_df_sample_5 <- df_sample_5 |> group_by(Region) |>
  summarise(total_max_region_sales=max(Sales),
            .groups = 'drop') |>
  arrange(desc(total_max_region_sales),.by_group= TRUE)
count_df_sample_5
```

```
## # A tibble: 4 x 2
##   Region  total_max_region_sales
##   <chr>                    <dbl>
## 1 Central                 17500.
## 2 East                    11200.
## 3 South                    8750.
## 4 West                     3611.
```

```
From above grouping can see products from central region have the highest sale.Followed by Eastern
```

- From all the above samples, for categorical value - Region and Max_Sales :
  - we see for about 2 samples Southern region has the max sales cost for the products purchased. Also, the value of sales is 22638.480 in each.
  - But in rest of the samples, South is seen to be the region having mid or even lowest at sale value. This seemed like an anomaly considering the samples showing a different picture.We cant entirely rely on any sample for knowing the max sales in regions.