DATA SCIENCE(UCS548)- PROJECT

DASHBOARD USING TABLEAU

Submitted By:

Name: Siddharth

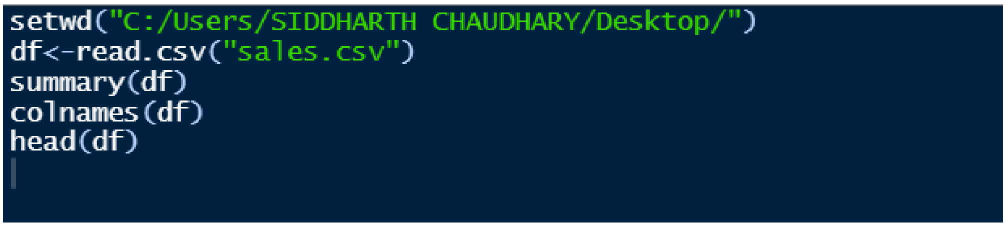
Rollno:102003707

Subgroup: 3COE28

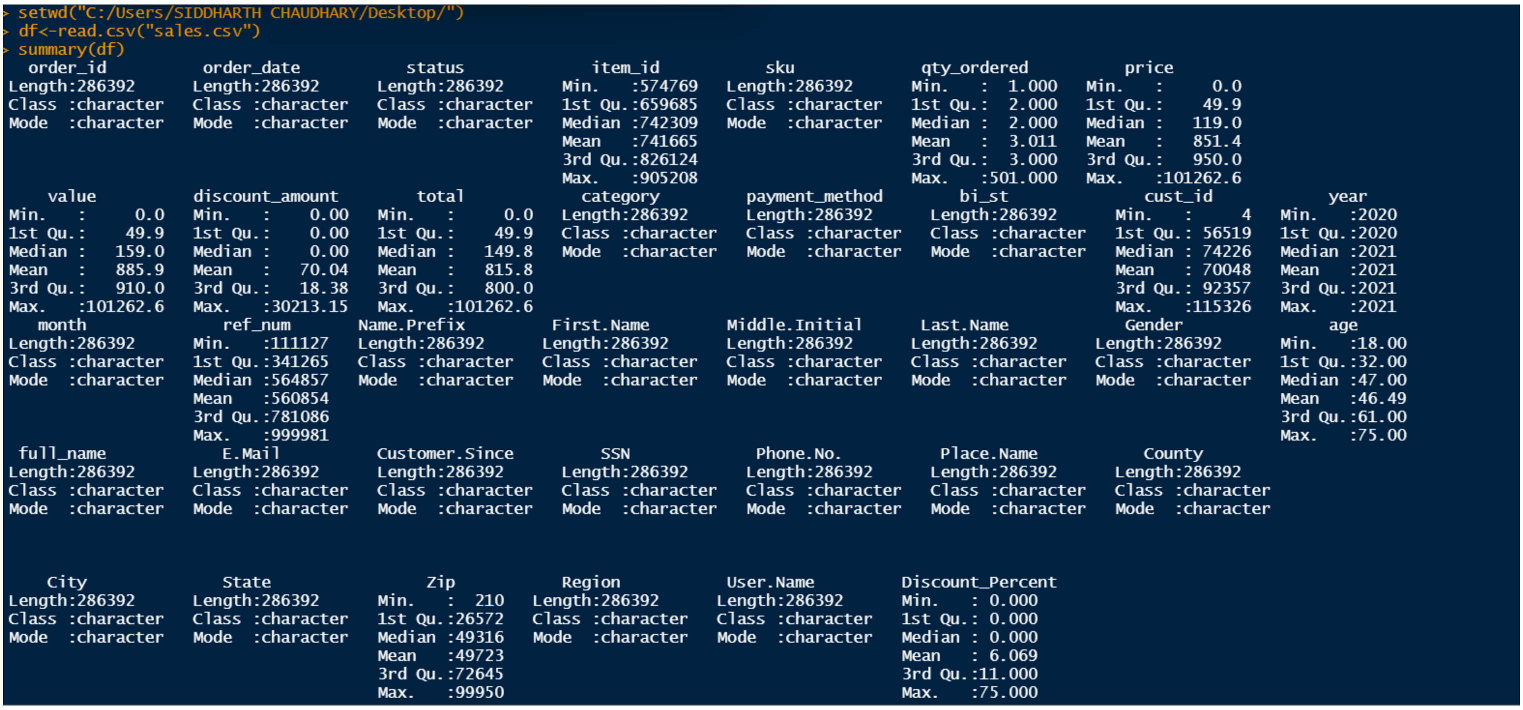
Topic:

Customer Analysis

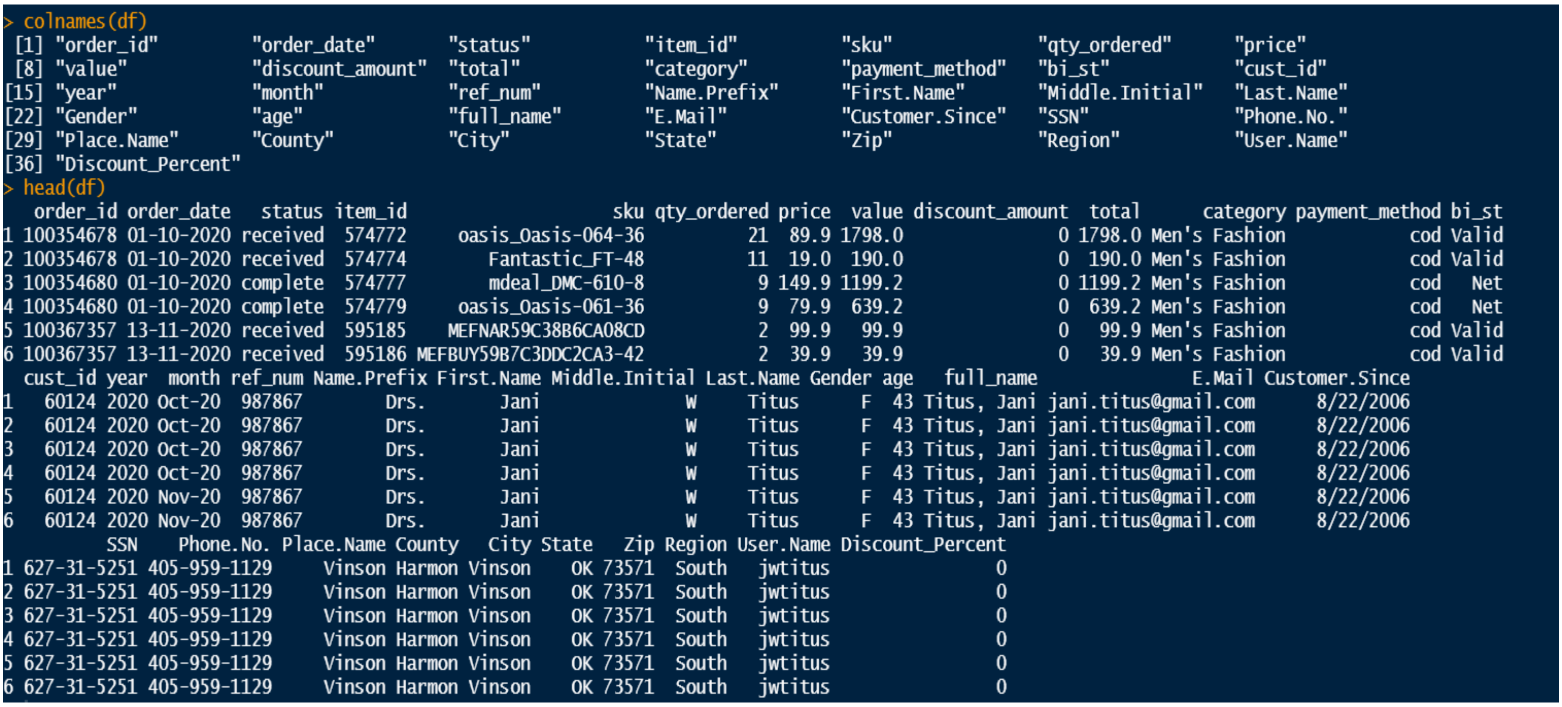
R Code:



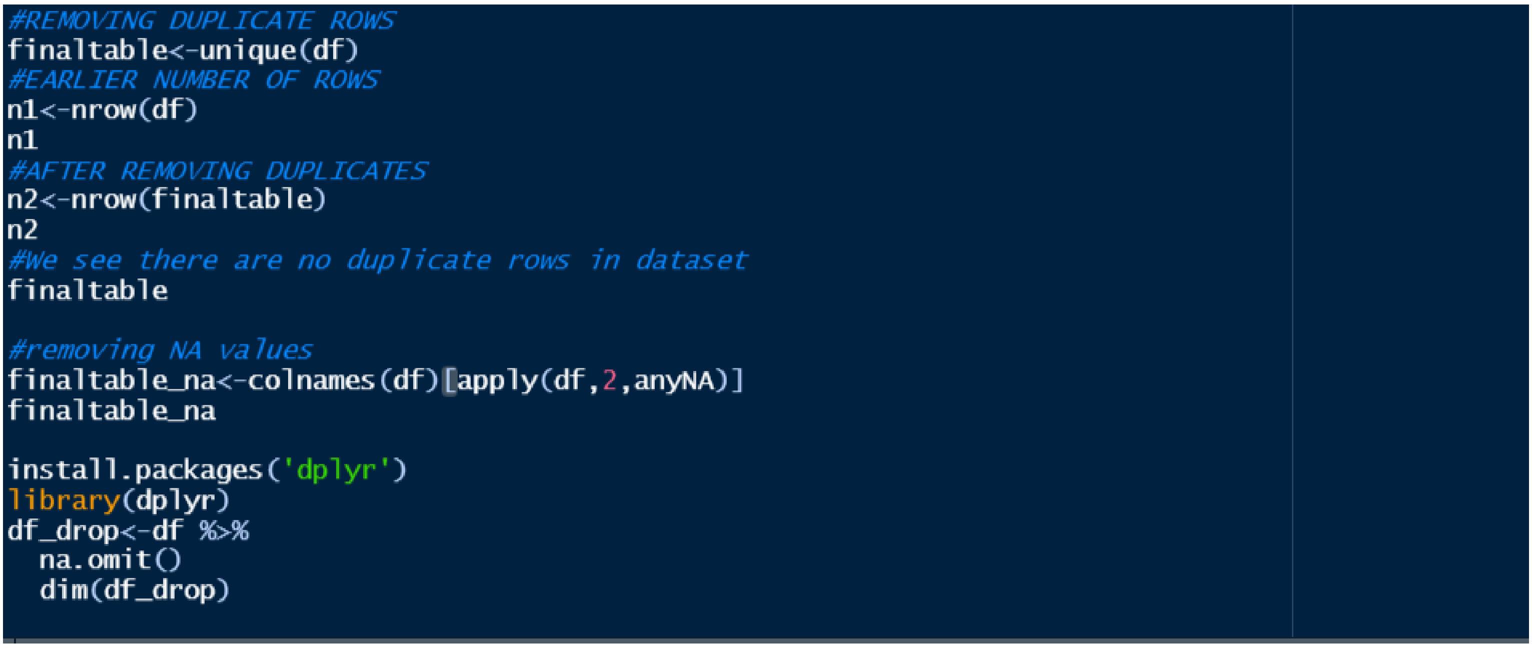
```
setwd("C:/Users/SIDDHARTH CHAUDHARY/Desktop/")
df<-read.csv("sales.csv")
summary(df)
colnames(df)
head(df)
```



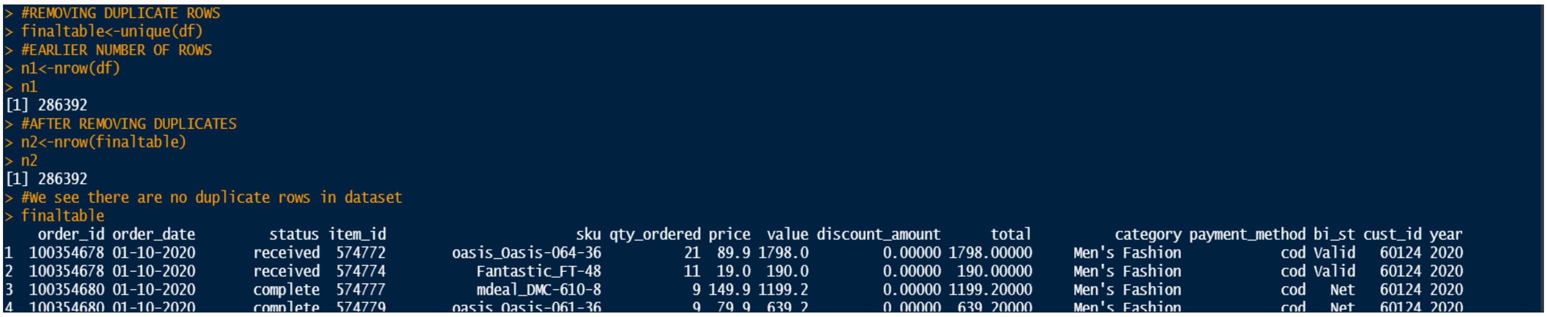
```
> setwd("C:/Users/SIDDHARTH CHAUDHARY/Desktop/")
> df<-read.csv("sales.csv")
> summary(df)
  order_id          order_date           status            item_id             sku             qty_ordered          price
Length:286392      Length:286392      Length:286392      Min.   :574769   Length:286392      Min.   :  1.000   Min.   :     0.0
Class :character   Class :character   Class :character   1st Qu.:659685   Class :character   1st Qu.:  2.000   1st Qu.:    49.9
Mode  :character   Mode  :character   Mode  :character   Median :742309   Mode  :character   Median :  2.000   Median :   119.0
                                                         Mean   :741665                      Mean   :  3.011   Mean   :   851.4
                                                         3rd Qu.:826124                      3rd Qu.:  3.000   3rd Qu.:   950.0
                                                         Max.   :905208                      Max.   :501.000   Max.   :101262.6
    value          discount_amount        total            category        payment_method         bi_st             cust_id            year
Min.   :     0.0   Min.   :    0.00   Min.   :     0.0   Length:286392      Length:286392      Length:286392      Min.   :     4   Min.   :2020
1st Qu.:    49.9   1st Qu.:    0.00   1st Qu.:    49.9   Class :character   Class :character   Class :character   1st Qu.: 56519   1st Qu.:2020
Median :   159.0   Median :    0.00   Median :   149.8   Mode  :character   Mode  :character   Mode  :character   Median : 74226   Median :2021
Mean   :   885.9   Mean   :   70.04   Mean   :   815.8                                                            Mean   : 70048   Mean   :2021
3rd Qu.:   910.0   3rd Qu.:   18.38   3rd Qu.:   800.0                                                            3rd Qu.: 92357   3rd Qu.:2021
Max.   :101262.6   Max.   :30213.15   Max.   :101262.6                                                            Max.   :115326   Max.   :2021
   month              ref_num        Name.Prefix         First.Name        Middle.Initial      Last.Name             Gender             age
Length:286392      Min.   :111127   Length:286392      Length:286392      Length:286392      Length:286392      Length:286392      Min.   :18.00
Class :character   1st Qu.:341265   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:32.00
Mode  :character   Median :564857   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :47.00
                   Mean   :560854                                                                                                  Mean   :46.49
                   3rd Qu.:781086                                                                                                  3rd Qu.:61.00
                   Max.   :999981                                                                                                  Max.   :75.00
 full_name            E.Mail          Customer.Since         SSN              Phone.No.           Place.Name          County
Length:286392      Length:286392      Length:286392      Length:286392      Length:286392      Length:286392      Length:286392
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


    City              State                Zip           Region          User.Name         Discount_Percent
Length:286392      Length:286392      Min.   :  210   Length:286392      Length:286392     Min.   : 0.000
Class :character   Class :character   1st Qu.:26572   Class :character   Class :character  1st Qu.: 0.000
Mode  :character   Mode  :character   Median :49316   Mode  :character   Mode  :character  Median : 0.000
                                      Mean   :49723                                        Mean   : 6.069
                                      3rd Qu.:72645                                        3rd Qu.:11.000
                                      Max.   :99950                                        Max.   :75.000
```



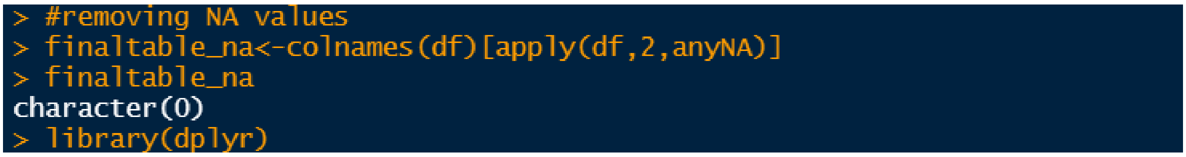
```
> colnames(df)
 [1] "order_id"         "order_date"       "status"           "item_id"          "sku"              "qty_ordered"      "price"
 [8] "value"            "discount_amount"  "total"            "category"         "payment_method"   "bi_st"            "cust_id"
[15] "year"             "month"            "ref_num"          "Name.Prefix"      "First.Name"       "Middle.Initial"   "Last.Name"
[22] "Gender"           "age"              "full_name"        "E.Mail"           "Customer.Since"   "SSN"              "Phone.No."
[29] "Place.Name"       "County"           "City"             "State"            "Zip"              "Region"           "User.Name"
[36] "Discount_Percent"
> head(df)
   order_id order_date   status item_id                    sku qty_ordered price  value discount_amount  total      category payment_method bi_st
1 100354678 01-10-2020 received  574772     oasis_Oasis-064-36          21  89.9 1798.0               0 1798.0 Men's Fashion            cod Valid
2 100354678 01-10-2020 received  574774         Fantastic_FT-48          11  19.0  190.0               0  190.0 Men's Fashion            cod Valid
3 100354680 01-10-2020 complete  574777         mdeal_DMC-610-8           9 149.9 1199.2               0 1199.2 Men's Fashion            cod   Net
4 100354680 01-10-2020 complete  574779     oasis_Oasis-061-36           9  79.9  639.2               0  639.2 Men's Fashion            cod   Net
5 100367357 13-11-2020 received  595185     MEFNAR59C38B6CA08CD           2  99.9   99.9               0   99.9 Men's Fashion            cod Valid
6 100367357 13-11-2020 received  595186 MEFBUY59B7C3DDC2CA3-42           2  39.9   39.9               0   39.9 Men's Fashion            cod Valid
  cust_id year  month ref_num Name.Prefix First.Name Middle.Initial Last.Name Gender age   full_name               E.Mail Customer.Since
1   60124 2020 Oct-20  987867        Drs.       Jani              W     Titus      F  43 Titus, Jani jani.titus@gmail.com      8/22/2006
2   60124 2020 Oct-20  987867        Drs.       Jani              W     Titus      F  43 Titus, Jani jani.titus@gmail.com      8/22/2006
3   60124 2020 Oct-20  987867        Drs.       Jani              W     Titus      F  43 Titus, Jani jani.titus@gmail.com      8/22/2006
4   60124 2020 Oct-20  987867        Drs.       Jani              W     Titus      F  43 Titus, Jani jani.titus@gmail.com      8/22/2006
5   60124 2020 Nov-20  987867        Drs.       Jani              W     Titus      F  43 Titus, Jani jani.titus@gmail.com      8/22/2006
6   60124 2020 Nov-20  987867        Drs.       Jani              W     Titus      F  43 Titus, Jani jani.titus@gmail.com      8/22/2006
          SSN    Phone.No. Place.Name County   City State   Zip Region User.Name Discount_Percent
1 627-31-5251 405-959-1129     Vinson Harmon Vinson    OK 73571  South   jwtitus                0
2 627-31-5251 405-959-1129     Vinson Harmon Vinson    OK 73571  South   jwtitus                0
3 627-31-5251 405-959-1129     Vinson Harmon Vinson    OK 73571  South   jwtitus                0
4 627-31-5251 405-959-1129     Vinson Harmon Vinson    OK 73571  South   jwtitus                0
5 627-31-5251 405-959-1129     Vinson Harmon Vinson    OK 73571  South   jwtitus                0
6 627-31-5251 405-959-1129     Vinson Harmon Vinson    OK 73571  South   jwtitus                0
```



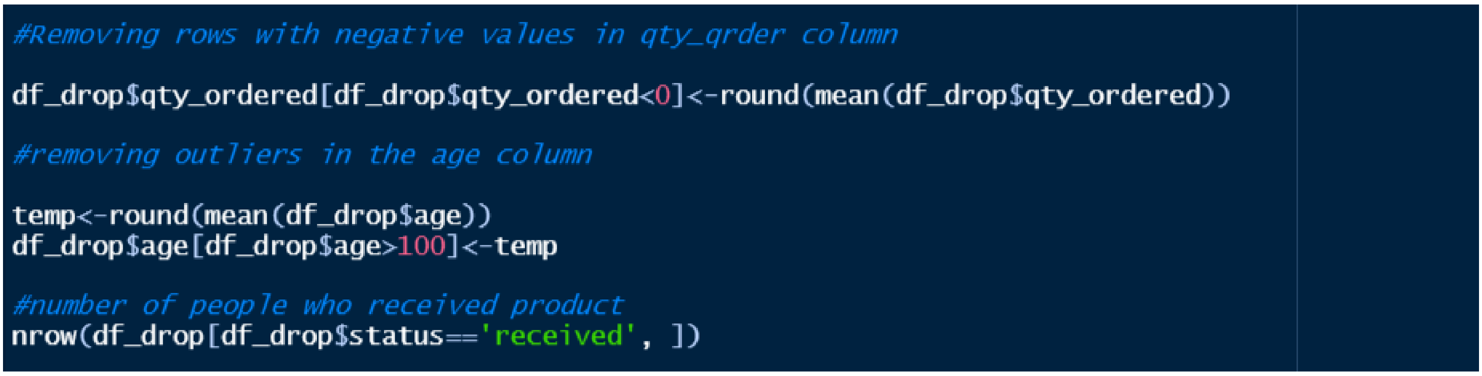
```
#REMOVING DUPLICATE ROWS
finaltable<-unique(df)
#EARLIER NUMBER OF ROWS
n1<-nrow(df)
n1
#AFTER REMOVING DUPLICATES
n2<-nrow(finaltable)
n2
#We see there are no duplicate rows in dataset
finaltable

#removing NA values
finaltable_na<-colnames(df)[apply(df,2,anyNA)]
finaltable_na

install.packages('dplyr')
library(dplyr)
df_drop<-df %>%
  na.omit()
  dim(df_drop)
```



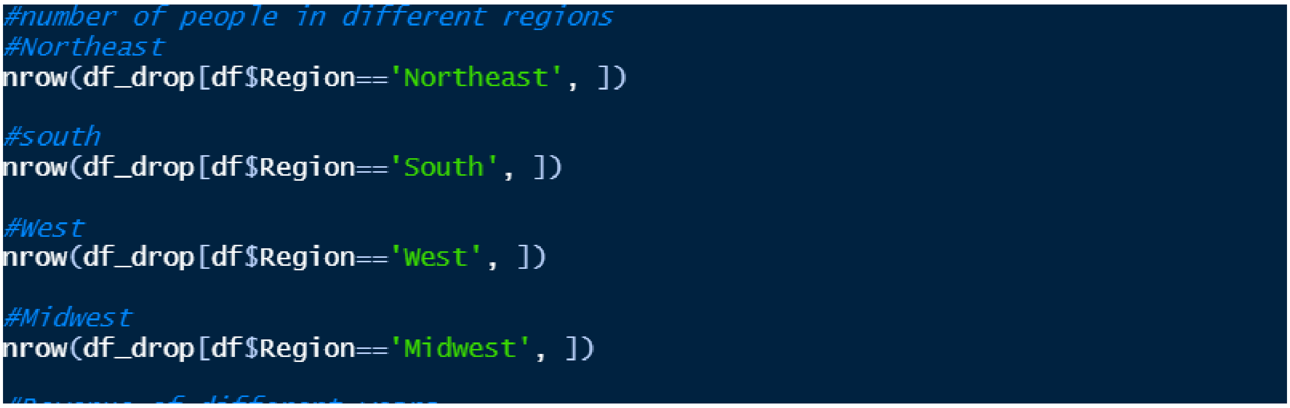
```
> #REMOVING DUPLICATE ROWS
> finaltable<-unique(df)
> #EARLIER NUMBER OF ROWS
> n1<-nrow(df)
> n1
[1] 286392
> #AFTER REMOVING DUPLICATES
> n2<-nrow(finaltable)
> n2
[1] 286392
> #We see there are no duplicate rows in dataset
> finaltable
   order_id order_date         status item_id                   sku qty_ordered price  value discount_amount      total         category payment_method bi_st cust_id year
1  100354678 01-10-2020       received  574772   oasis_Oasis-064-36          21  89.9 1798.0         0.00000 1798.00000    Men's Fashion            cod Valid   60124 2020
2  100354678 01-10-2020       received  574774      Fantastic_FT-48          11  19.0  190.0         0.00000  190.00000    Men's Fashion            cod Valid   60124 2020
3  100354680 01-10-2020       complete  574777      mdeal_DMC-610-8           9 149.9 1199.2         0.00000 1199.20000    Men's Fashion            cod   Net   60124 2020
4  100354680 01-10-2020       complete  574779   oasis_Oasis-061-36           9  79.9  639.2         0.00000  639.20000    Men's Fashion            cod   Net   60124 2020
```



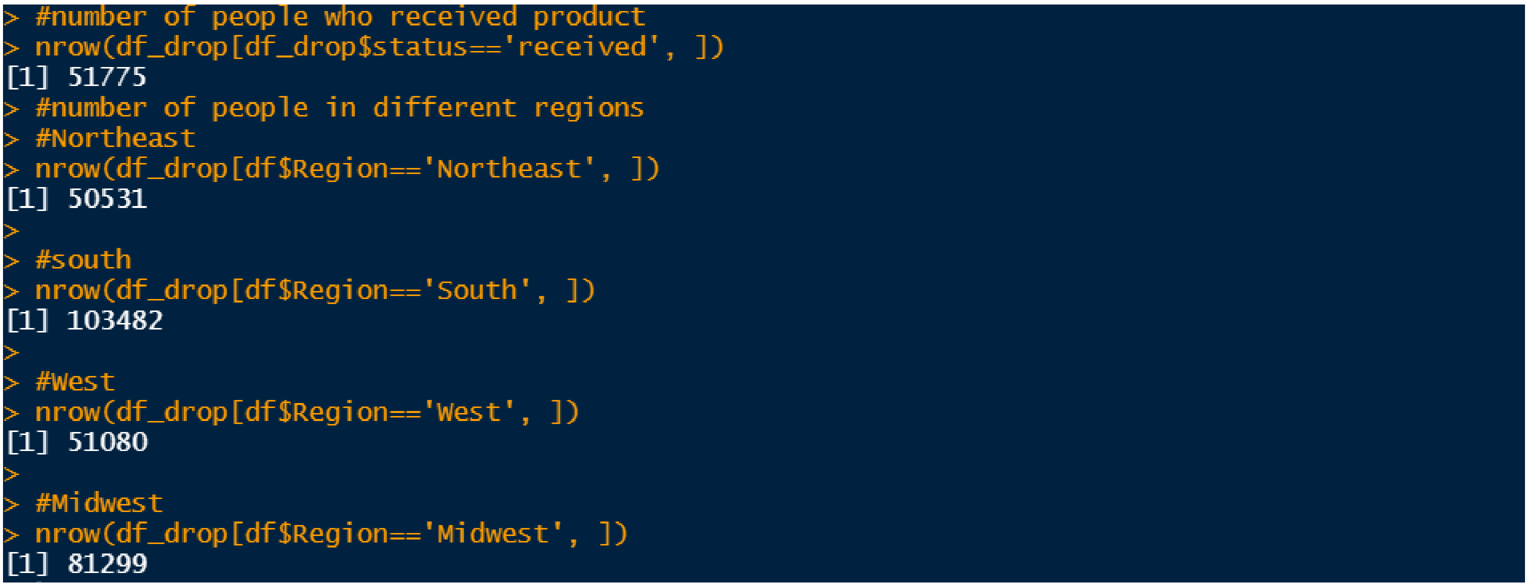
```
> #removing NA values
> finaltable_na<-colnames(df)[apply(df,2,anyNA)]
> finaltable_na
character(0)
> library(dplyr)
```



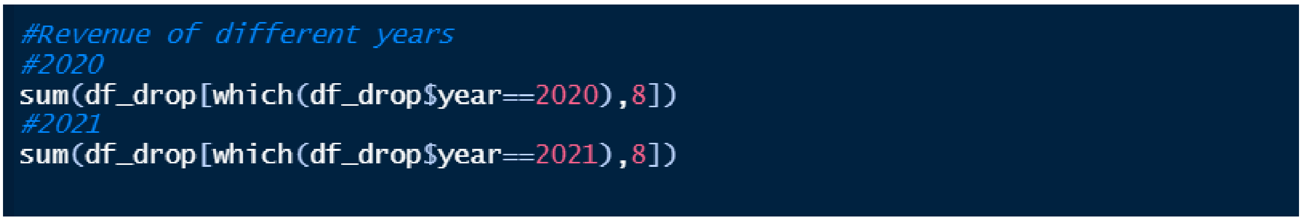
```
#Removing rows with negative values in qty_qrder column

df_drop$qty_ordered[df_drop$qty_ordered<0]<-round(mean(df_drop$qty_ordered))

#removing outliers in the age column

temp<-round(mean(df_drop$age))
df_drop$age[df_drop$age>100]<-temp

#number of people who received product
nrow(df_drop[df_drop$status=='received', ])
```



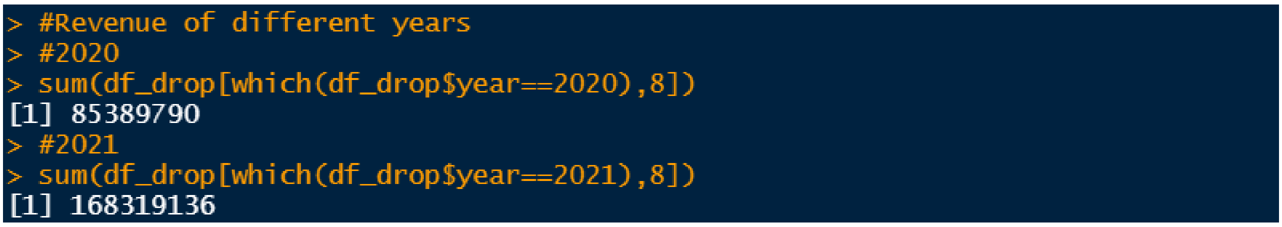
```
#number of people in different regions
#Northeast
nrow(df_drop[df$Region=='Northeast', ])

#south
nrow(df_drop[df$Region=='South', ])

#West
nrow(df_drop[df$Region=='West', ])

#Midwest
nrow(df_drop[df$Region=='Midwest', ])

#Revenue of different years
```



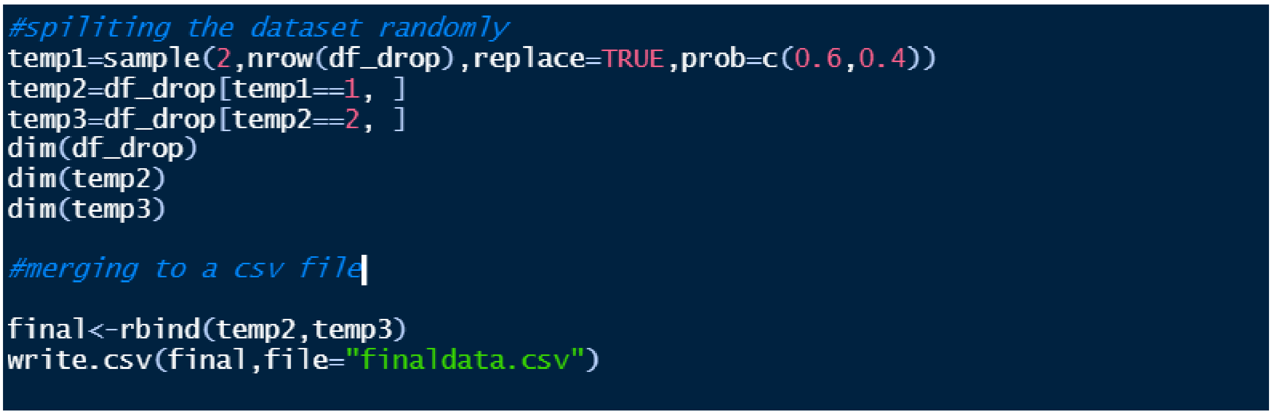
```
> #number of people who received product
> nrow(df_drop[df_drop$status=='received', ])
[1] 51775
> #number of people in different regions
> #Northeast
> nrow(df_drop[df$Region=='Northeast', ])
[1] 50531
>
> #south
> nrow(df_drop[df$Region=='South', ])
[1] 103482
>
> #West
> nrow(df_drop[df$Region=='West', ])
[1] 51080
>
> #Midwest
> nrow(df_drop[df$Region=='Midwest', ])
[1] 81299
```



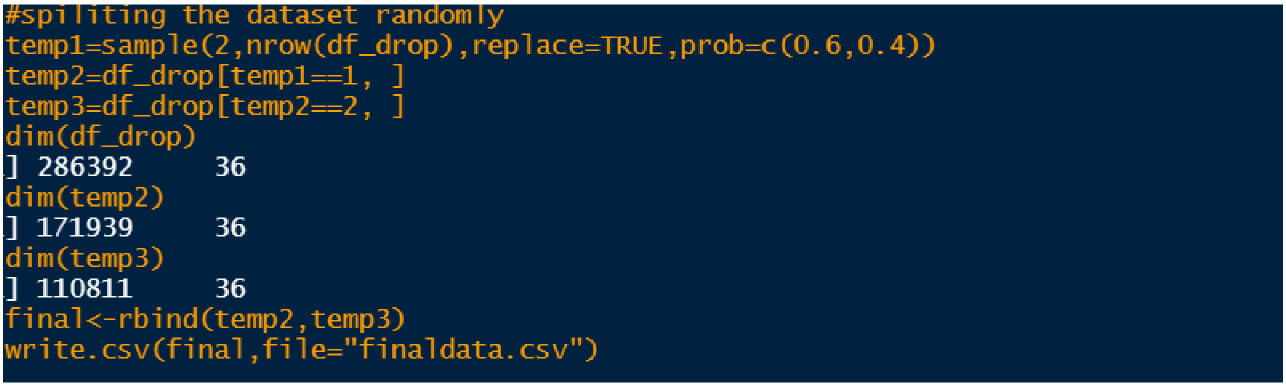
```
#Revenue of different years
#2020
sum(df_drop[which(df_drop$year==2020),8])
#2021
sum(df_drop[which(df_drop$year==2021),8])
```



```
> #Revenue of different years
> #2020
> sum(df_drop[which(df_drop$year==2020),8])
[1] 85389790
> #2021
> sum(df_drop[which(df_drop$year==2021),8])
[1] 168319136
```

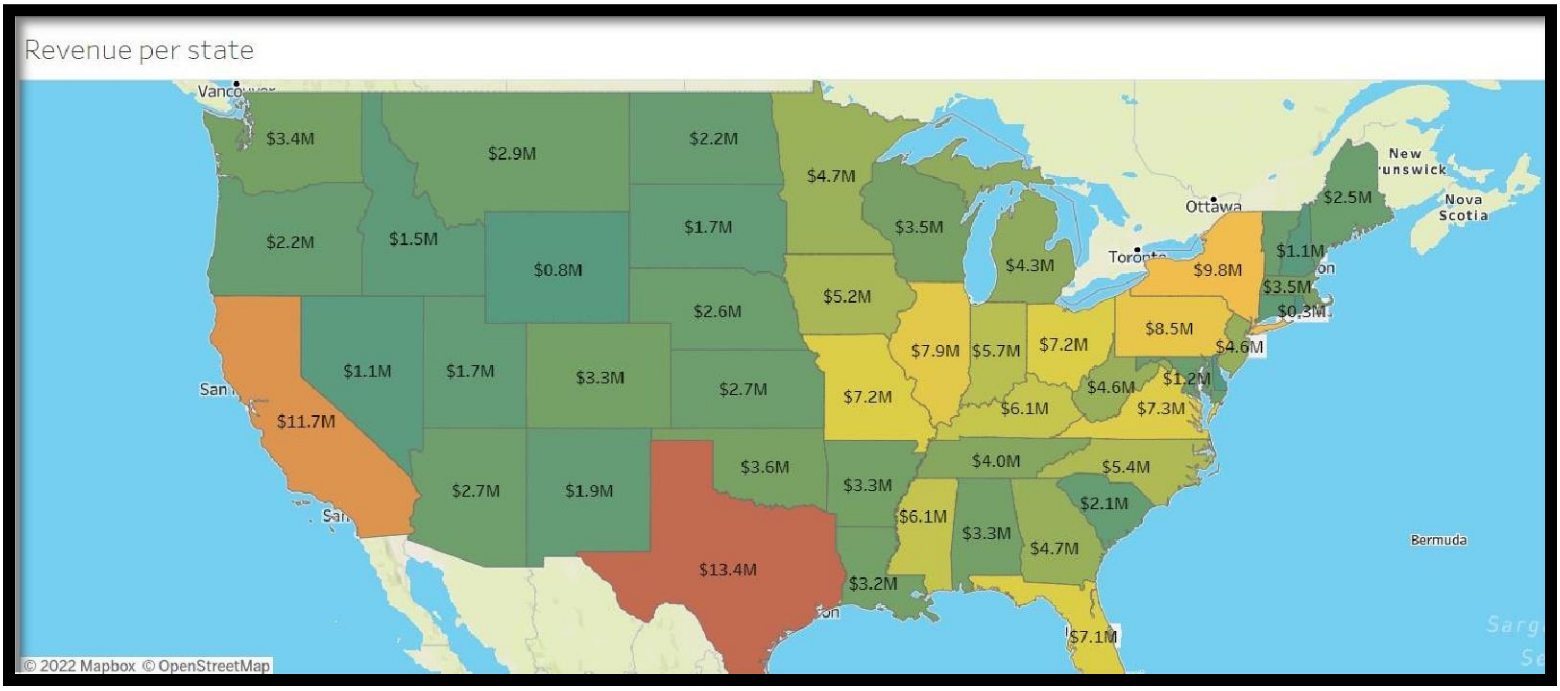


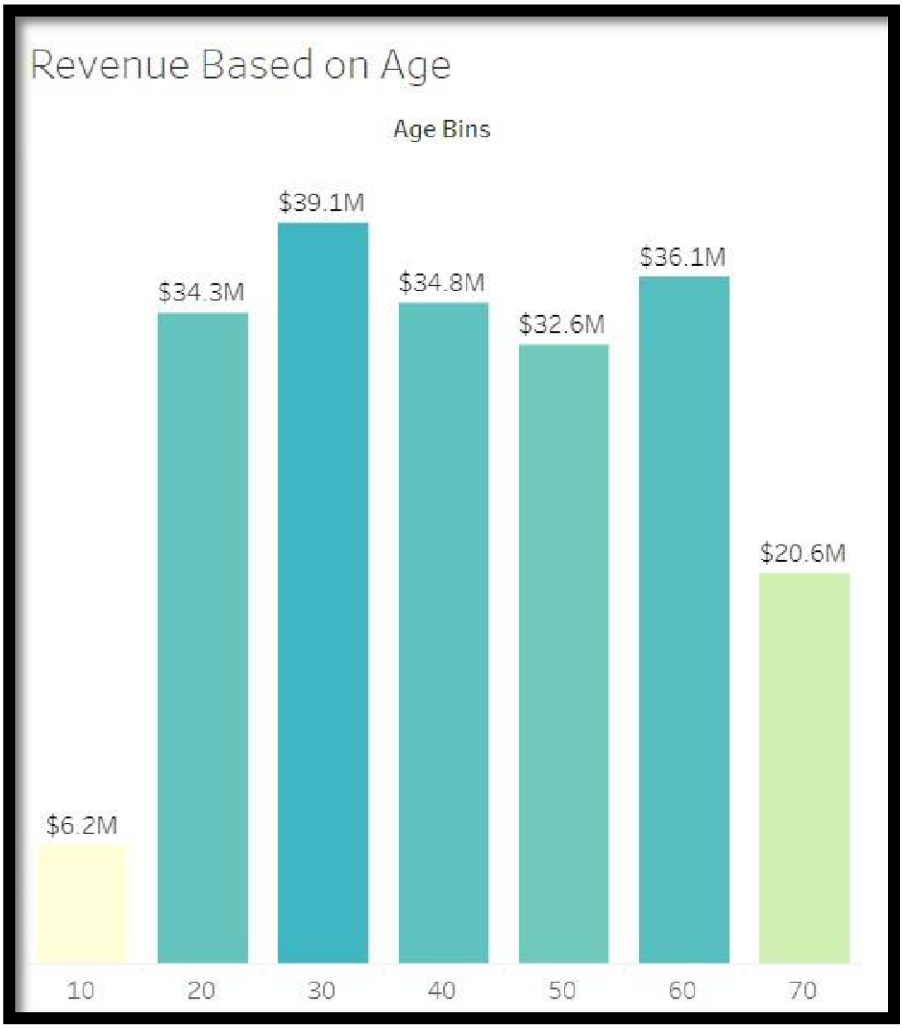
```
#spiliting the dataset randomly
temp1=sample(2,nrow(df_drop),replace=TRUE,prob=c(0.6,0.4))
temp2=df_drop[temp1==1, ]
temp3=df_drop[temp2==2, ]
dim(df_drop)
dim(temp2)
dim(temp3)

#merging to a csv file

final<-rbind(temp2,temp3)
write.csv(final,file="finaldata.csv")
```

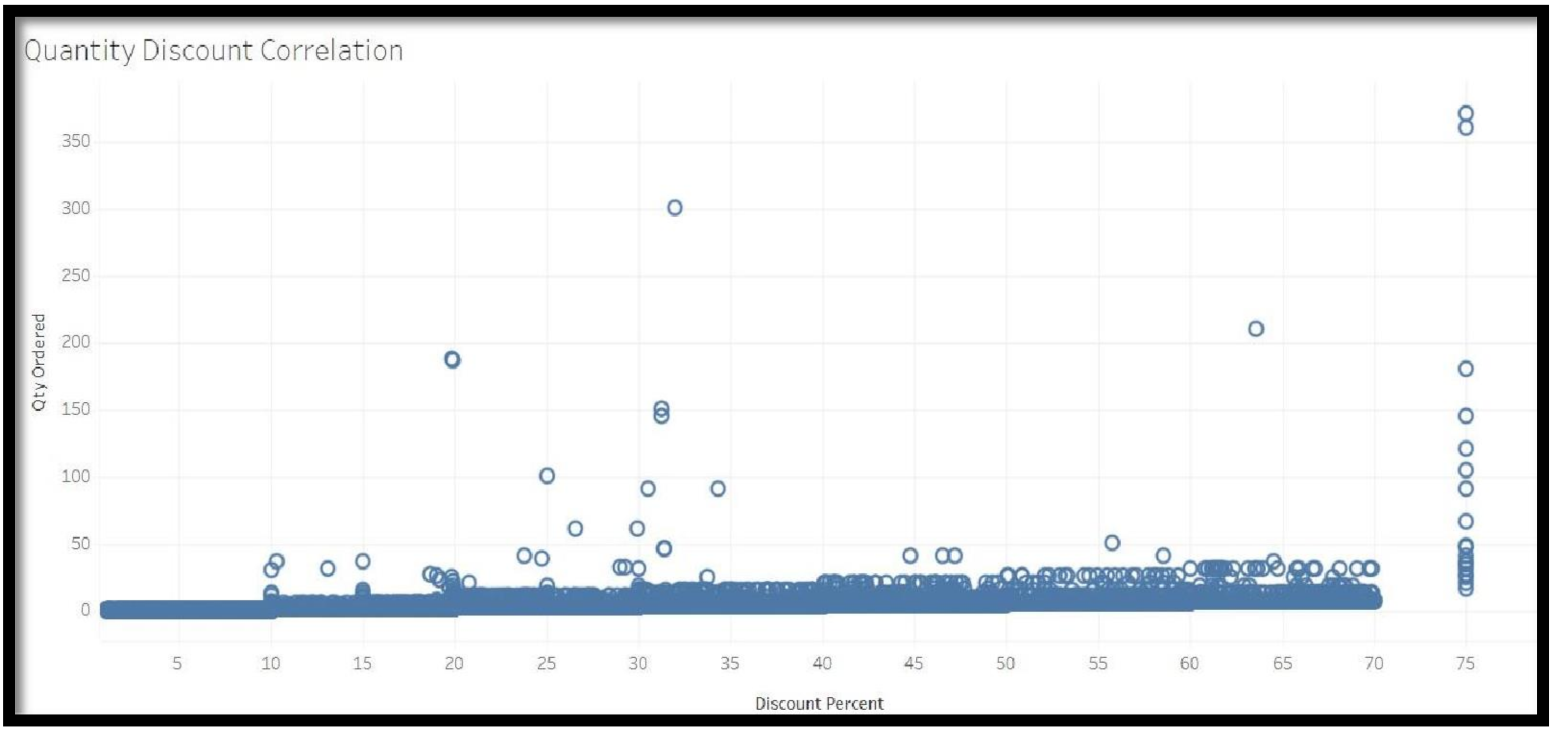


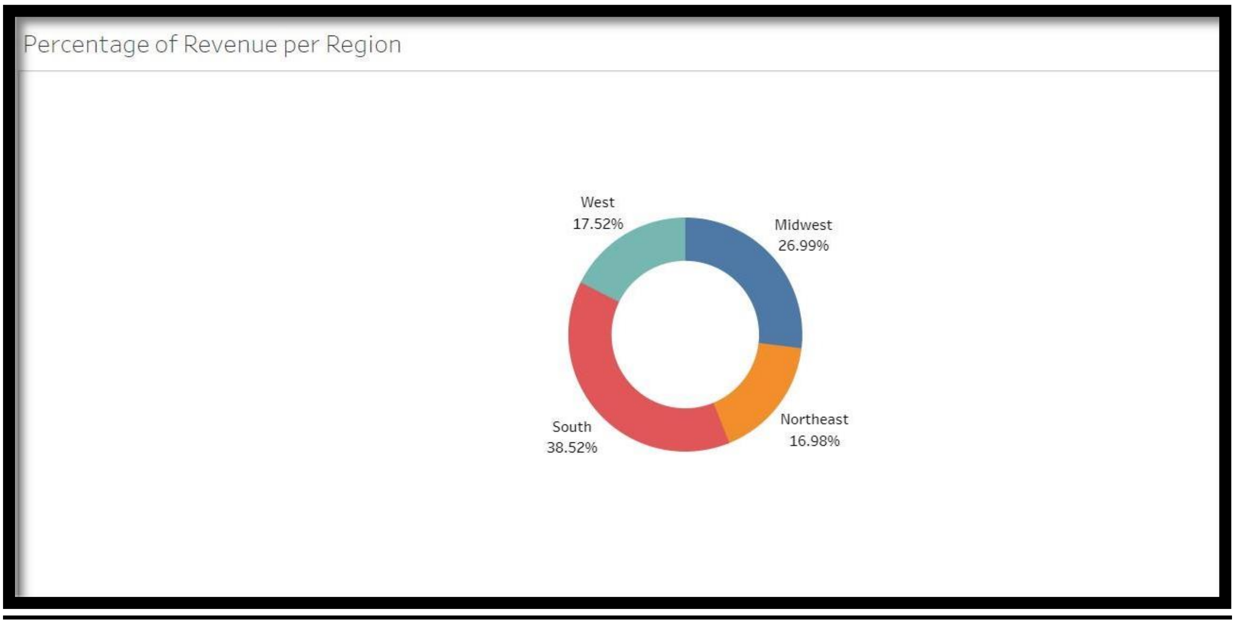
```
#spiliting the dataset randomly
temp1=sample(2,nrow(df_drop),replace=TRUE,prob=c(0.6,0.4))
temp2=df_drop[temp1==1, ]
temp3=df_drop[temp2==2, ]
dim(df_drop)
] 286392     36
dim(temp2)
] 171939     36
dim(temp3)
] 110811     36
final<-rbind(temp2,temp3)
write.csv(final,file="finaldata.csv")
```

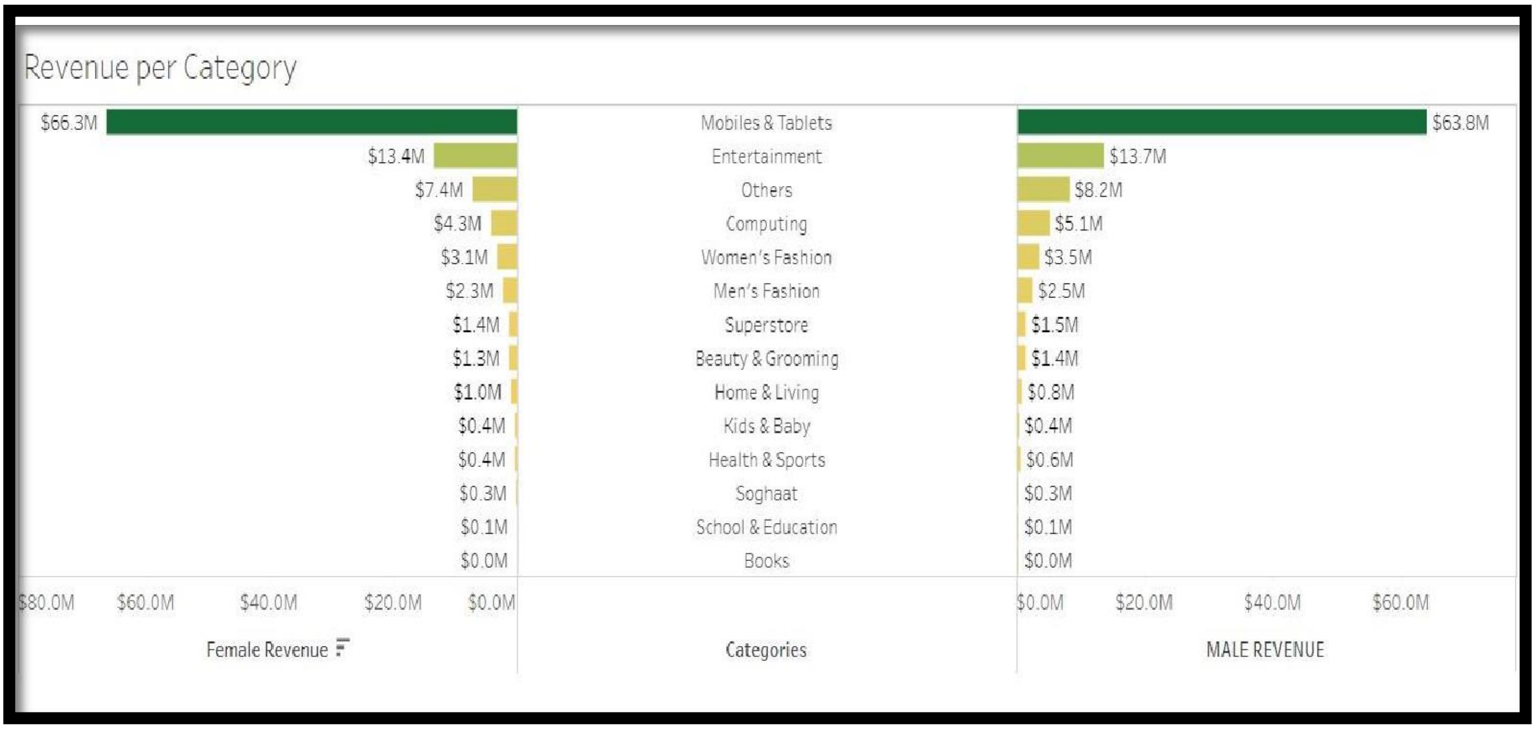
Tableau Sheets:



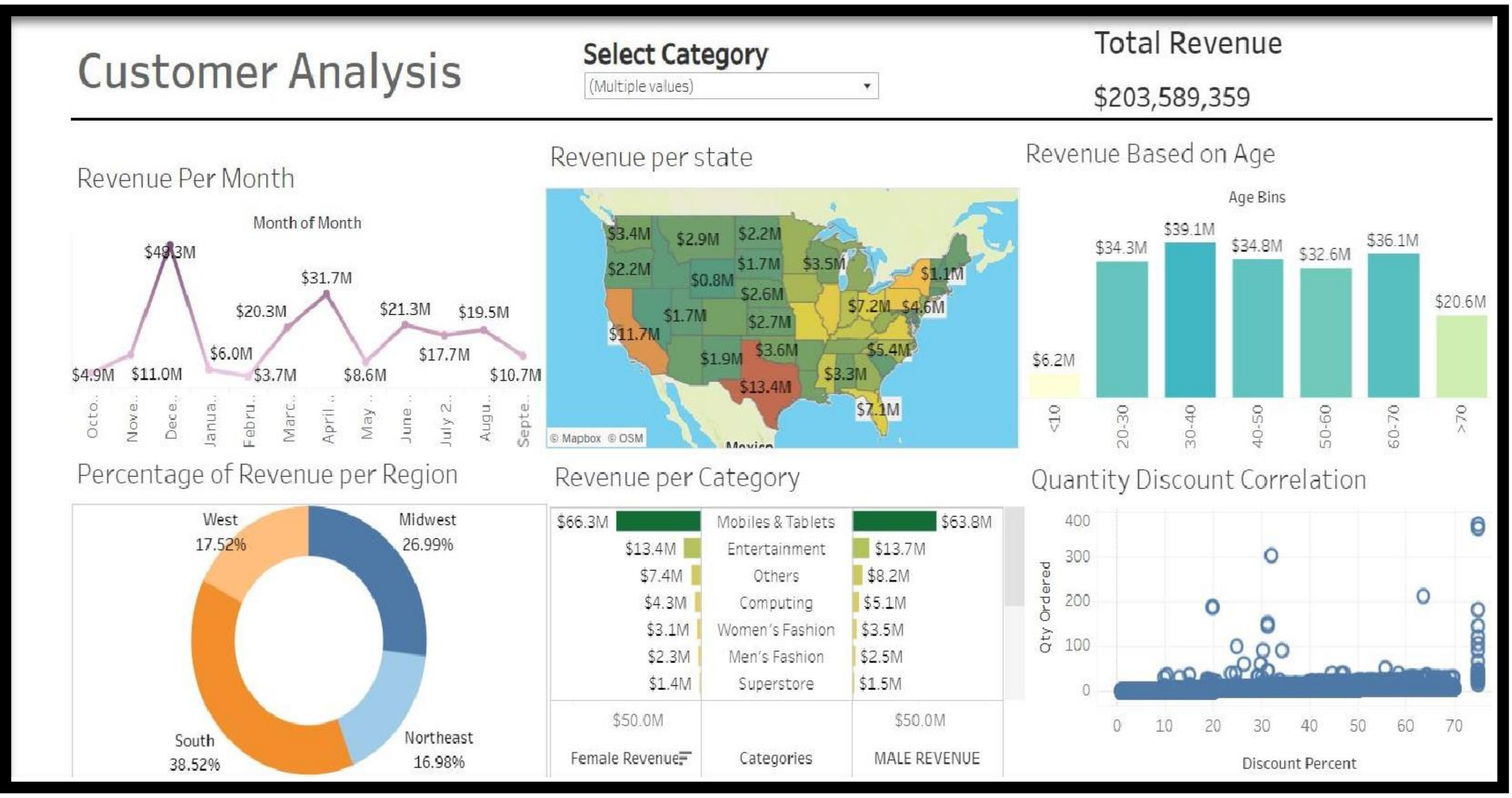

Revenue Based on Age


Quantity Discount Correlation


Percentage of Revenue per Region
West
17.52%
Midwest
26.99%
Northeast
16.98%
South
38.52%


Revenue per Category
$66.3M
Mobiles & Tablets
$63.8M
$13.4M
Entertainment
$13.7M
$7.4M
Others
$8.2M
$4.3M
Computing
$5.1M
$3.1M
Women's Fashion
$3.5M
$2.3M
Men's Fashion
$2.5M
$1.4M
Superstore
$1.5M
$1.3M
Beauty & Grooming
$1.4M
$1.0M
Home & Living
$0.8M
$0.4M
Kids & Baby
$0.4M
$0.4M
Health & Sports
$0.6M
$0.3M
Soghaat
$0.3M
$0.1M
School & Education
$0.1M
$0.0M
Books
$0.0M
$80.0M
$60.0M
$40.0M
$20.0M
$0.0M
$0.0M
$20.0M
$40.0M
$60.0M
Female Revenue
Categories
MALE REVENUE

DASHBOARD



Customer Analysis

Select Category
(Multiple values)

Total Revenue
$203,589,359