EXPLORATORY DATA ANALYSIS

-SALES DATASET OF A TOY STORE

DATASOURCE: Data.world

DATA DESCRIPTION:

The data shows the details of sales of toys across different countries which has 12 attributes and 2823 rows

The attributes of the dataset are mentioned as follows

Order Number- The order number of the items which is sold out

Quantity Ordered- The number of item placed as a order

Price Each-Tells about the each price of the product

Orderline Number: describes about the number of order placed at a time

Sales: Sales of the product

Status:status of the product

Month_ID-Monthlywise order placed

Year_ID-Year of the item ordered

Product Line-Types of product is sold

City- Product ordered location

Country-product ordered location by country

Deal size-The size of the deal is selling out

PROBLEM:

To import a stocks of the toy in a toy store. We need to know, which of the item is selling more out and less in cancelation and returns

ASSUMPTION:

Let assume the item is selling out based on their types of productline

considering that, the item selling out according to their monthly orders

assume that, which will not give the loss for purchasing the toys.So,considering the status of the product which is more shipped compared to other status

By the above assumption, subset the dataset into different data frames and get a knowledge about which of the product purchase will give the optimal results

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

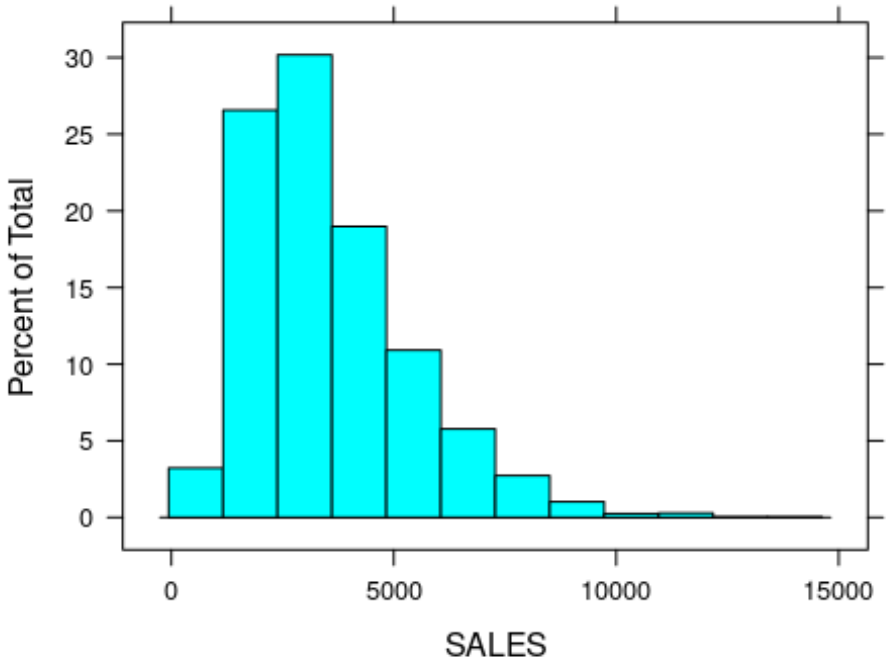
```
library(lattice)
library(ggplot2)
library(readxl)
df <- read_excel("sales_data_sample1.xlsx")
```

```
df
```

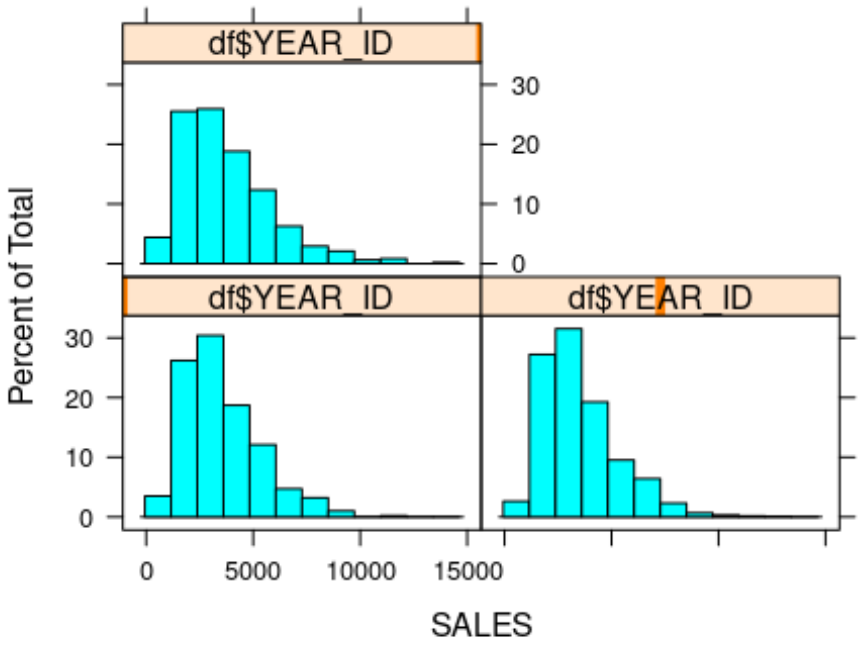
```
## # A tibble: 2,823 × 12
##    ORDERNUM…¹ QUANT…² PRICE…³ ORDER…⁴ SALES STATUS MONTH⋯⁵ YEAR_ID
PRODU…⁶ CITY
##         <dbl>   <dbl>   <dbl>   <dbl> <dbl> <chr>    <dbl>   <dbl> <chr>
<chr>
##  1      10107      30    95.7       2 2871  Shipp…       2    2003
```

```
Motorc… NYC
##  2      10121      34    81.4       5 2766. Shipp…       5    2003
Motorc… Reims
##  3      10134      41    94.7       2 3884. Shipp…       7    2003
Motorc… Paris
##  4      10145      45    83.3       6 3747. Shipp…       8    2003
Motorc… Pasa…
##  5      10159      49   100        14 5205. Shipp…      10    2003
Motorc… San …
##  6      10168      36    96.7       1 3480. Shipp…      10    2003
Motorc… Burl…
##  7      10180      29    86.1       9 2498. Shipp…      11    2003
Motorc… Lille
##  8      10188      48   100         1 5512. Shipp…      11    2003
Motorc… Berg…
##  9      10201      22    98.6       2 2169. Shipp…      12    2003
Motorc… San …
## 10      10211      41   100        14 4708. Shipp…       1    2004
Motorc… Paris
## # … with 2,813 more rows, 2 more variables: COUNTRY <chr>, DEALSIZE
<chr>, and
## #   abbreviated variable names ¹ORDERNUMBER, ²QUANTITYORDERED, ³PRICEEACH,
## #   ⁴ORDERLINENUMBER, ⁵MONTH_ID, ⁶PRODUCTLINE

df$PRODUCTLINE<-as.factor(df$PRODUCTLINE)
df$CITY<-as.factor(df$CITY)
df$COUNTRY<-as.factor(df$COUNTRY)
df$DEALSIZE<-as.factor(df$DEALSIZE)
df$STATUS<-as.factor(df$STATUS)
df$MONTH_ID=as.factor(df$MONTH_ID)
summary(df)

##   ORDERNUMBER    QUANTITYORDERED   PRICEEACH      ORDERLINENUMBER
##  Min.   :10100   Min.   : 6.00   Min.   : 26.88   Min.   : 1.000
##  1st Qu.:10180   1st Qu.:27.00   1st Qu.: 68.86   1st Qu.: 3.000
##  Median :10262   Median :35.00   Median : 95.70   Median : 6.000
##  Mean   :10259   Mean   :35.09   Mean   : 83.66   Mean   : 6.466
##  3rd Qu.:10334   3rd Qu.:43.00   3rd Qu.:100.00   3rd Qu.: 9.000
##  Max.   :10425   Max.   :97.00   Max.   :100.00   Max.   :18.000
##
##      SALES                STATUS        MONTH_ID      YEAR_ID
##  Min.   :  482.1   Cancelled :  60   11     :597   Min.   :2003
##  1st Qu.: 2203.4   Disputed  :  14   10     :317   1st Qu.:2003
##  Median : 3184.8   In Process:  41   5      :252   Median :2004
##  Mean   : 3553.9   On Hold   :  44   1      :229   Mean   :2004
##  3rd Qu.: 4508.0   Resolved  :  47   2      :224   3rd Qu.:2004
##  Max.   :14082.8   Shipped   :2617   3      :212   Max.   :2005
##                                      (Other):992
##            PRODUCTLINE             CITY           COUNTRY       DEALSIZE
##  Classic Cars    :967   Madrid       : 304   USA      :1004   Large : 157
```

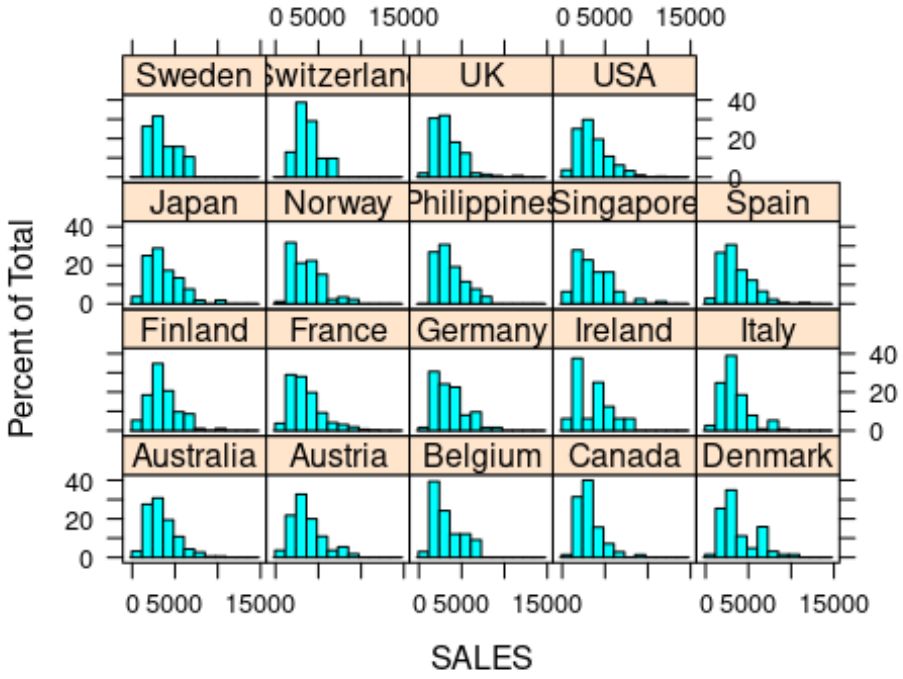
```
##  Motorcycles     :331   San Rafael    : 180   Spain    : 342   Medium:1384
##  Planes          :306   NYC           : 152   France   : 314   Small :1282
##  Ships           :234   Singapore     :  79   Australia: 185
##  Trains          : 77   Paris         :  70   UK       : 144
##  Trucks and Buses:301   San Francisco:  62   Italy    : 113
##  Vintage Cars    :607   (Other)       :1976   (Other)  : 721

colSums(is.na(df))

##     ORDERNUMBER QUANTITYORDERED       PRICEEACH ORDERLINENUMBER
SALES
##               0               0               0               0
0
##          STATUS        MONTH_ID         YEAR_ID     PRODUCTLINE
CITY
##               0               0               0               0
0
##         COUNTRY        DEALSIZE
##               0               0

histogram(~SALES,data = df)
```



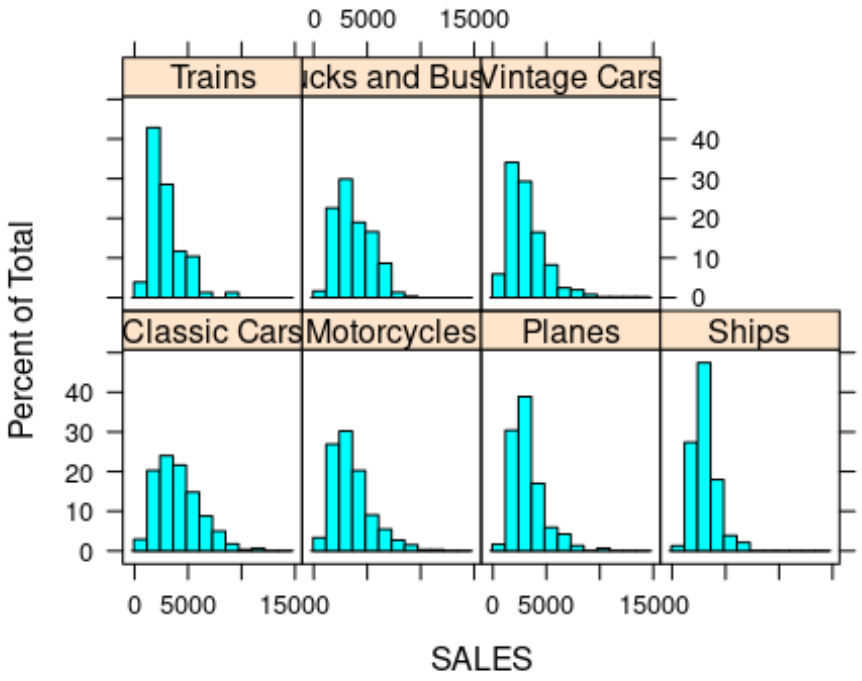

```
histogram(~SALES|df$YEAR_ID,data=df)
```



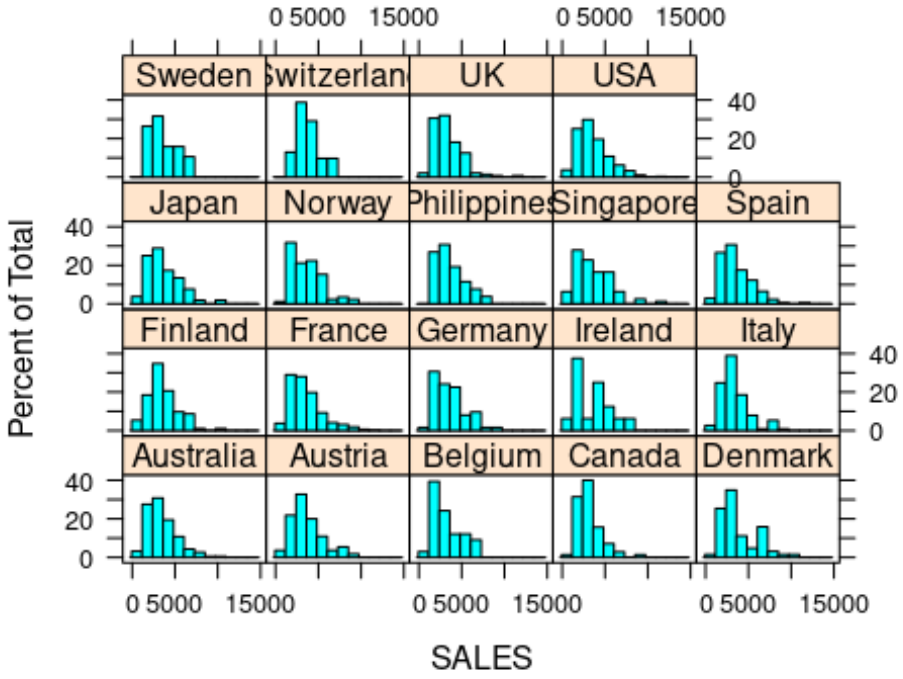

```
histogram(~SALES|COUNTRY,data = df)
```



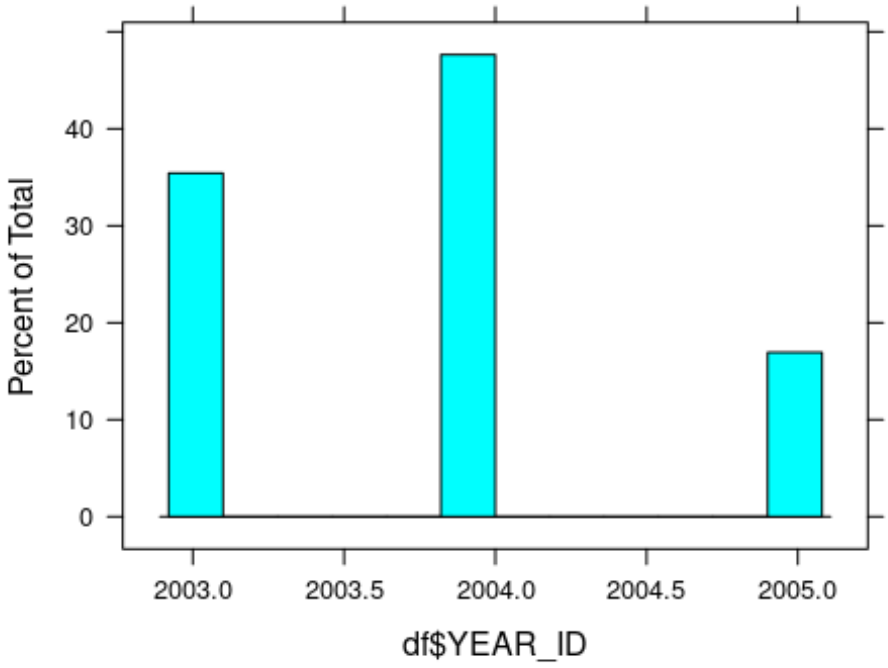

```
histogram(~SALES|PRODUCTLINE,data = df)
```



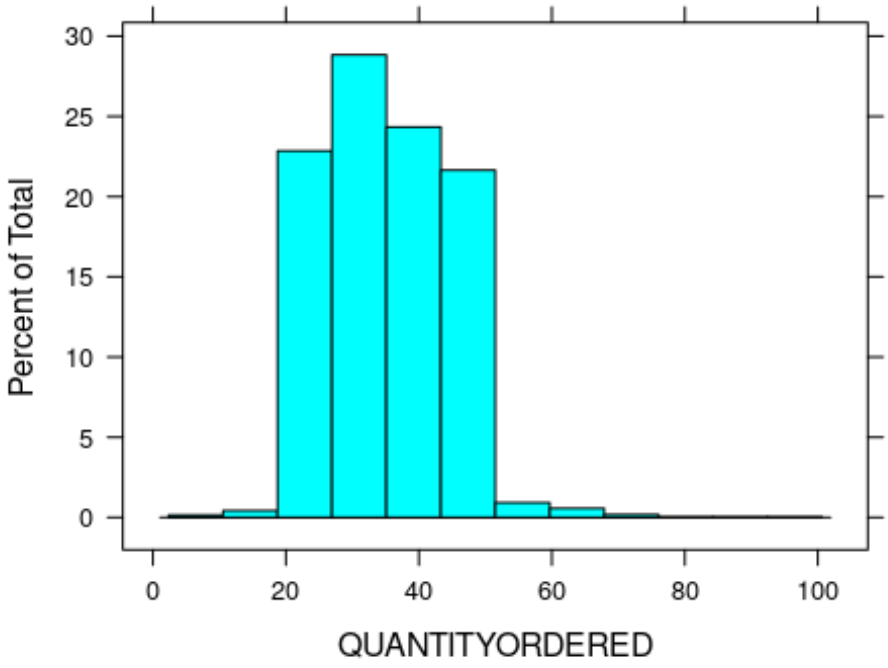

```
histogram(~SALES|COUNTRY,data = df)
```




```
histogram(~df$YEAR_ID,data = df)
```



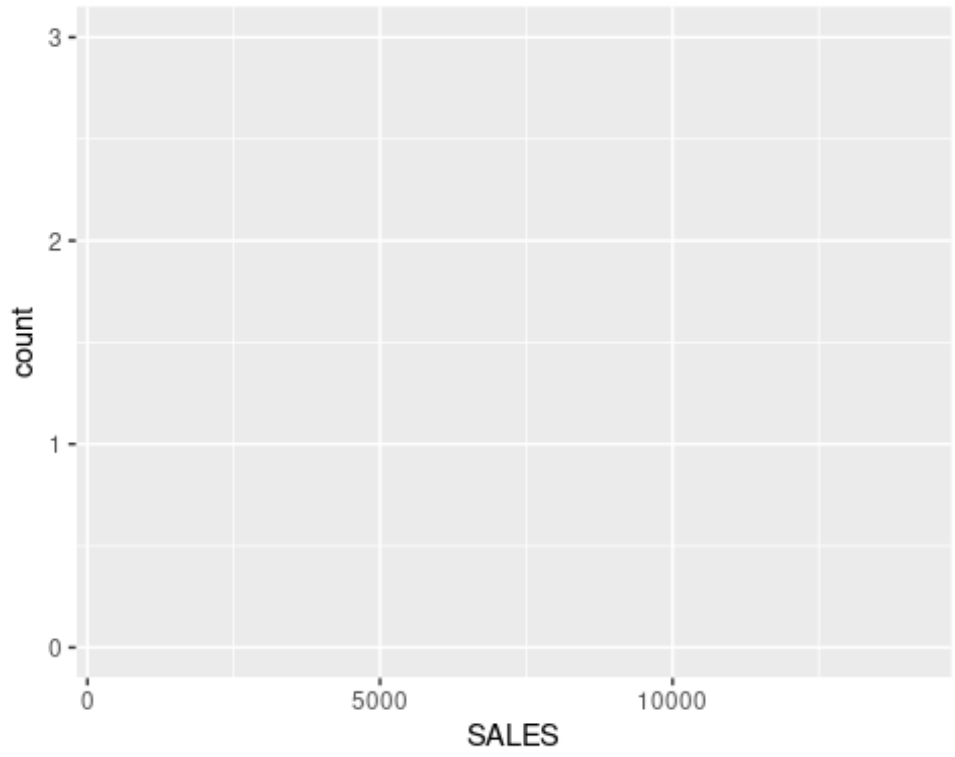

```
histogram(~QUANTITYORDERED,data = df)
```



```r
dfc=subset(df,COUNTRY=="USA")
summary(dfc)
```

```
##   ORDERNUMBER    QUANTITYORDERED   PRICEEACH      ORDERLINENUMBER
##  Min.   :10100   Min.   : 6.00   Min.   : 27.22   Min.   : 1.000
##  1st Qu.:10182   1st Qu.:27.00   1st Qu.: 69.94   1st Qu.: 3.000
##  Median :10257   Median :36.00   Median : 96.88   Median : 6.000
##  Mean   :10256   Mean   :35.52   Mean   : 83.82   Mean   : 6.344
##  3rd Qu.:10325   3rd Qu.:44.00   3rd Qu.:100.00   3rd Qu.: 9.000
##  Max.   :10422   Max.   :85.00   Max.   :100.00   Max.   :18.000
##
##      SALES                STATUS       MONTH_ID      YEAR_ID
##  Min.   :  541.1   Cancelled : 14   11     :205   Min.   :2003
##  1st Qu.: 2246.2   Disputed  :  0   10     :119   1st Qu.:2003
##  Median : 3236.1   In Process:  4   8      :112   Median :2004
##  Mean   : 3613.5   On Hold   : 38   5      : 85   Mean   :2004
##  3rd Qu.: 4583.9   Resolved  : 13   12     : 76   3rd Qu.:2004
##  Max.   :14082.8   Shipped   :935   3      : 71   Max.   :2005
##                                     (Other):336
##            PRODUCTLINE             CITY          COUNTRY       DEALSIZE
##  Classic Cars    :329   San Rafael   :180   USA      :1004   Large : 64
##  Motorcycles     :149   NYC          :152   Australia:   0   Medium:505
##  Planes          : 95   San Francisco: 62   Austria  :   0   Small :435
##  Ships           : 70   New Bedford  : 61   Belgium  :   0
##  Trains          : 25   Brickhaven   : 47   Canada   :   0
##  Trucks and Buses:112   Boston       : 44   Denmark  :   0
##  Vintage Cars    :224   (Other)      :458   (Other)  :   0
```

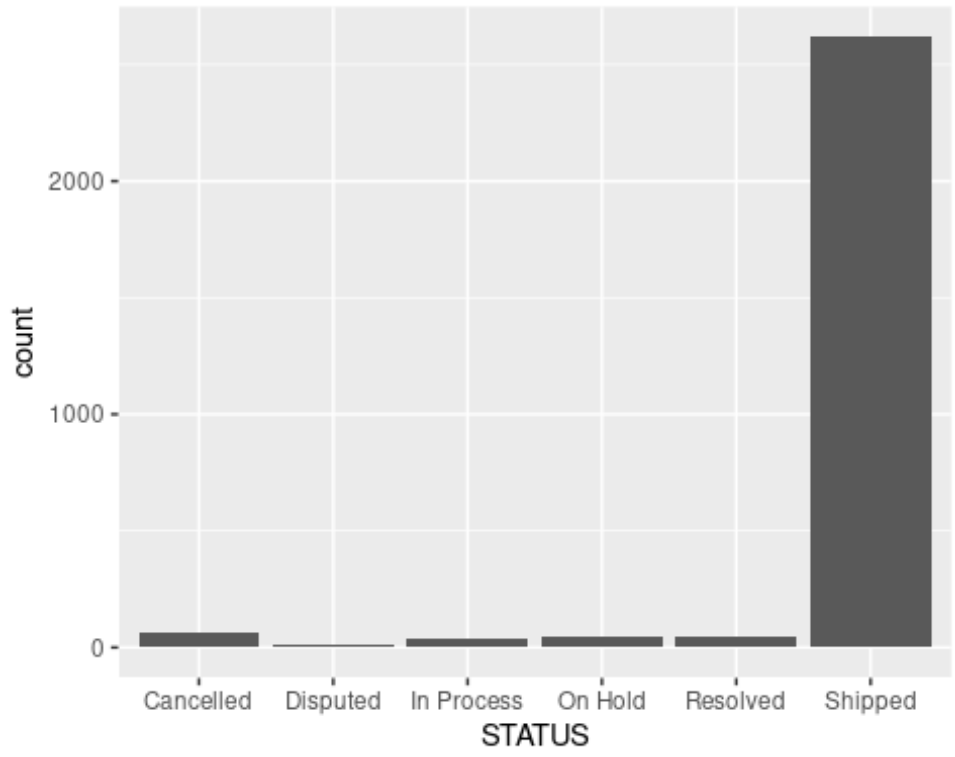
```r
df %>% ggplot(aes(SALES))+geom_bar(stat = "count")
```



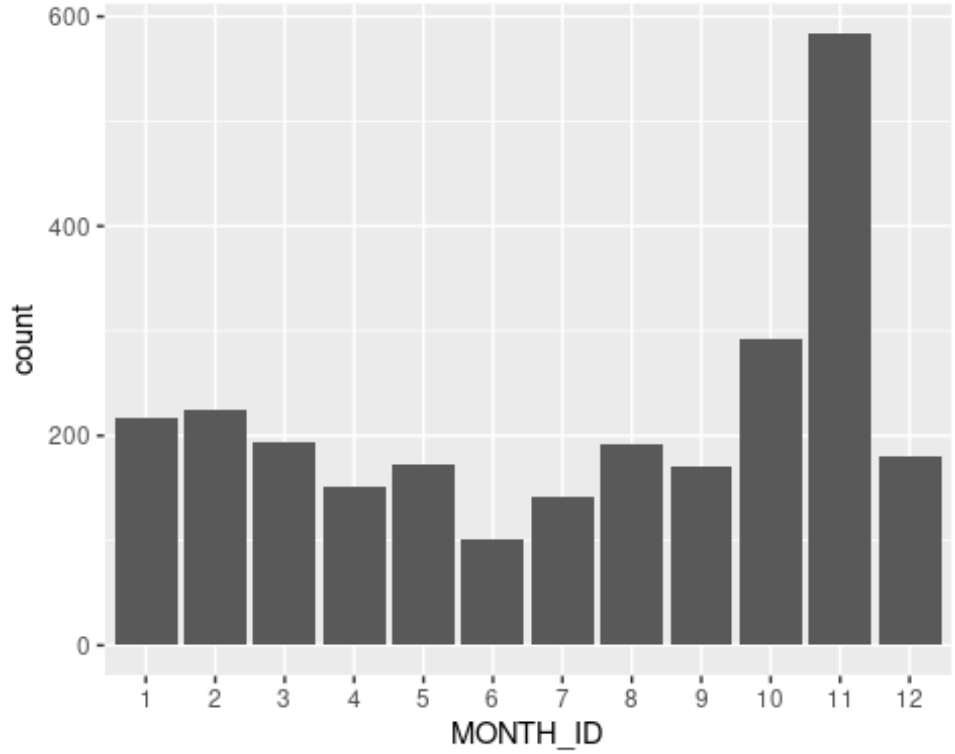

```
summary(df$SALES)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   482.1  2203.4  3184.8  3553.9  4508.0 14082.8
```

```
df %>% ggplot(aes(STATUS))+geom_bar(stat = "count")
```



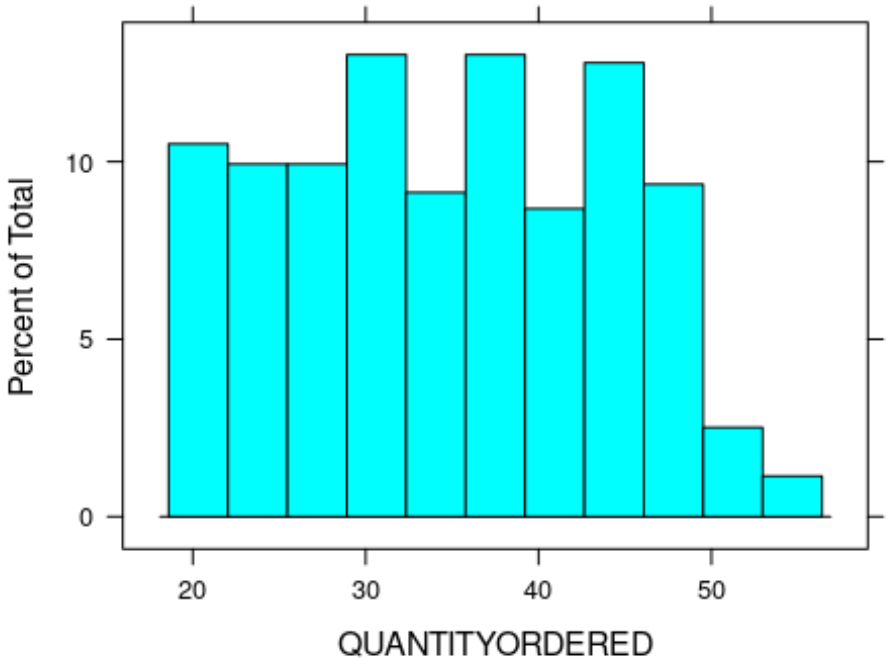

```
dfst=subset(df,STATUS=="Shipped")
dfst %>% ggplot(aes(MONTH_ID))+geom_bar(stat = "count")
```



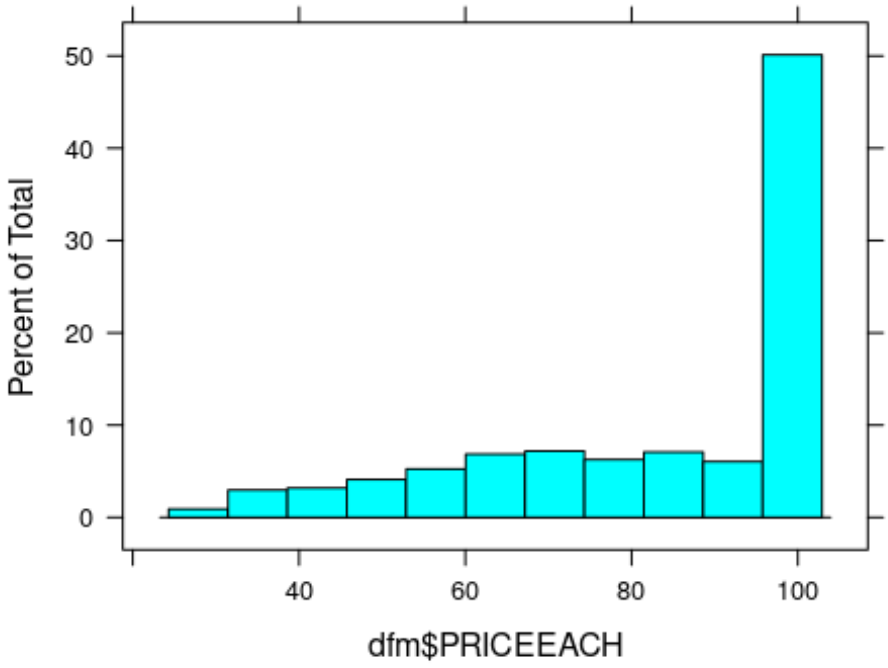
```
dfm=subset(dfst,MONTH_ID==10|dfst$MONTH_ID==11)
summary(dfm)
```

```
##   ORDERNUMBER    QUANTITYORDERED   PRICEEACH      ORDERLINENUMBER
##  Min.   :10154   Min.   :20.00   Min.   : 27.22   Min.   : 1.000
##  1st Qu.:10181   1st Qu.:27.00   1st Qu.: 69.56   1st Qu.: 3.000
##  Median :10304   Median :34.00   Median : 95.91   Median : 6.000
##  Mean   :10255   Mean   :34.87   Mean   : 83.73   Mean   : 6.776
##  3rd Qu.:10322   3rd Qu.:43.00   3rd Qu.:100.00   3rd Qu.:10.000
##  Max.   :10348   Max.   :55.00   Max.   :100.00   Max.   :18.000
##
##      SALES                STATUS       MONTH_ID      YEAR_ID
##  Min.   :  694.6   Cancelled :  0   11     :583   Min.   :2003
##  1st Qu.: 2219.3   Disputed  :  0   10     :293   1st Qu.:2003
##  Median : 3215.4   In Process:  0   1      :  0   Median :2004
##  Mean   : 3553.1   On Hold   :  0   2      :  0   Mean   :2004
##  3rd Qu.: 4513.4   Resolved  :  0   3      :  0   3rd Qu.:2004
##  Max.   :12536.5   Shipped   :876   4      :  0   Max.   :2004
##                                      (Other):  0
##            PRODUCTLINE             CITY          COUNTRY      DEALSIZE
##  Classic Cars    :326   NYC          : 53   USA      :324   Large : 45
##  Motorcycles     :102   San Rafael   : 37   France   : 87   Medium:432
##  Planes          : 79   Madrid       : 36   UK       : 80   Small :399
##  Ships           : 61   San Francisco: 36   Spain    : 73
##  Trains          : 24   Manchester   : 35   Australia: 57
##  Trucks and Buses: 86   Philadelphia : 35   Norway   : 52
##  Vintage Cars    :198   (Other)      :644   (Other)  :203
```

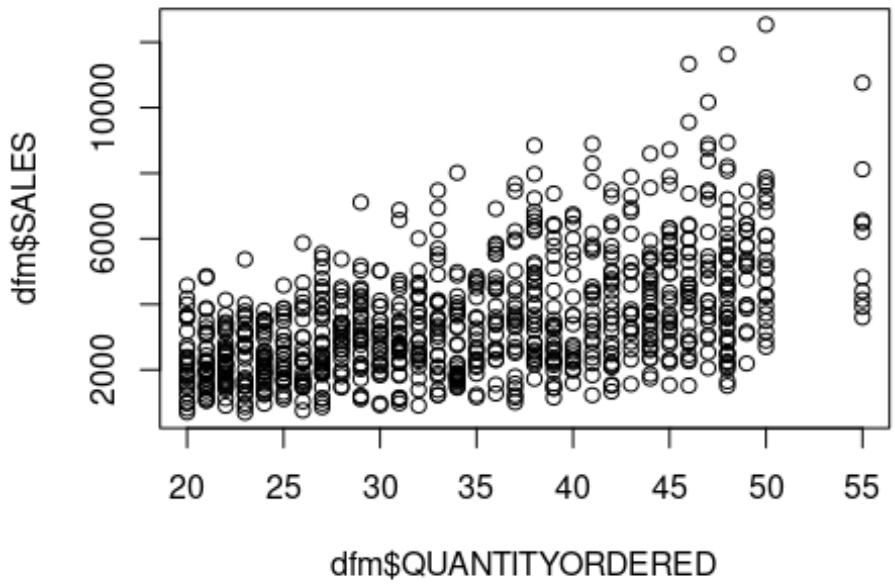
```
histogram(~QUANTITYORDERED,data=dfm)
```



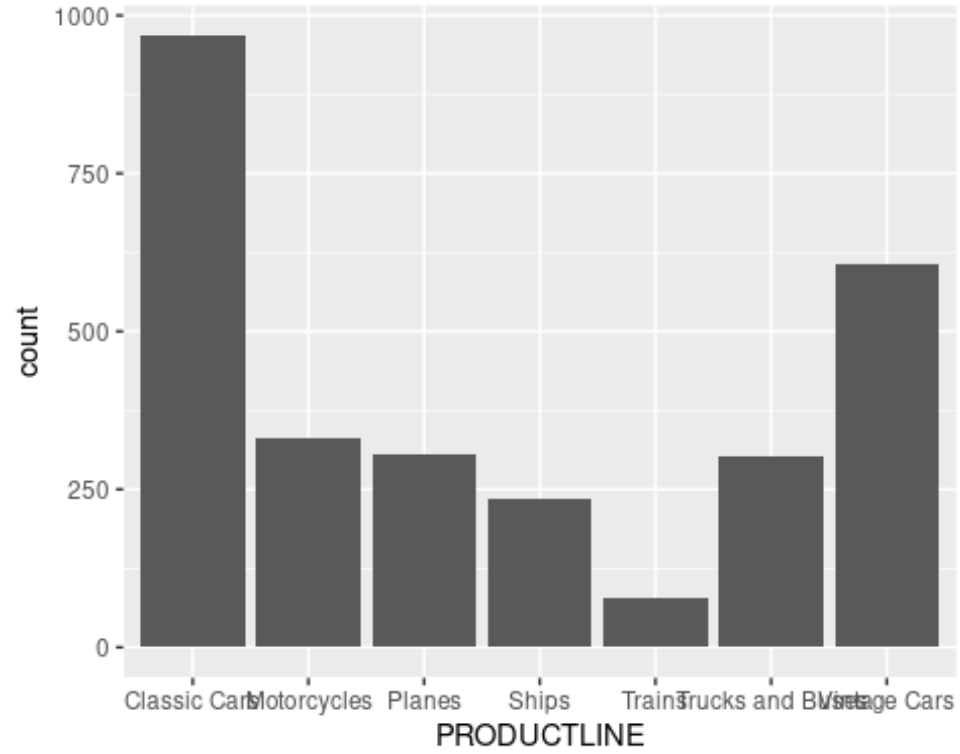

```
histogram(dfm$PRICEEACH)
```



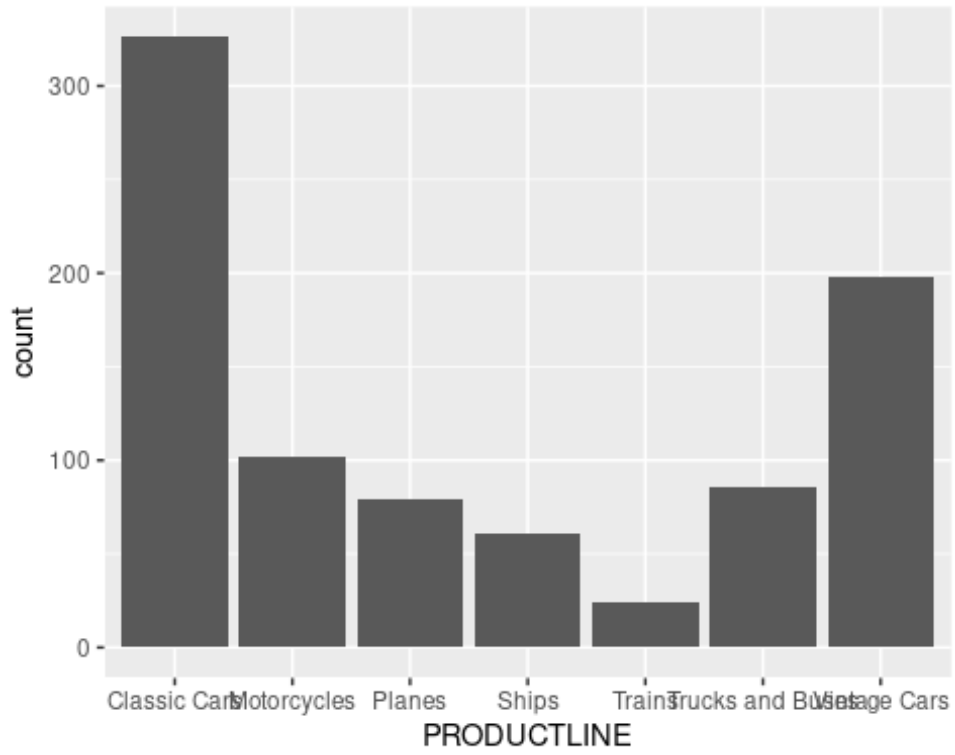

```
plot(x=dfm$QUANTITYORDERED,y=dfm$SALES)
```




```
df %>% ggplot(aes(PRODUCTLINE))+geom_bar(stat = "count")
```




```
dfm%>% ggplot(aes(PRODUCTLINE))+geom_bar(stat = "count")
```




```
dfsi=subset(dfm,DEALSIZE=="Medium")

dfst1=subset(df,STATUS !="Shipped")
summary(dfst1)
```

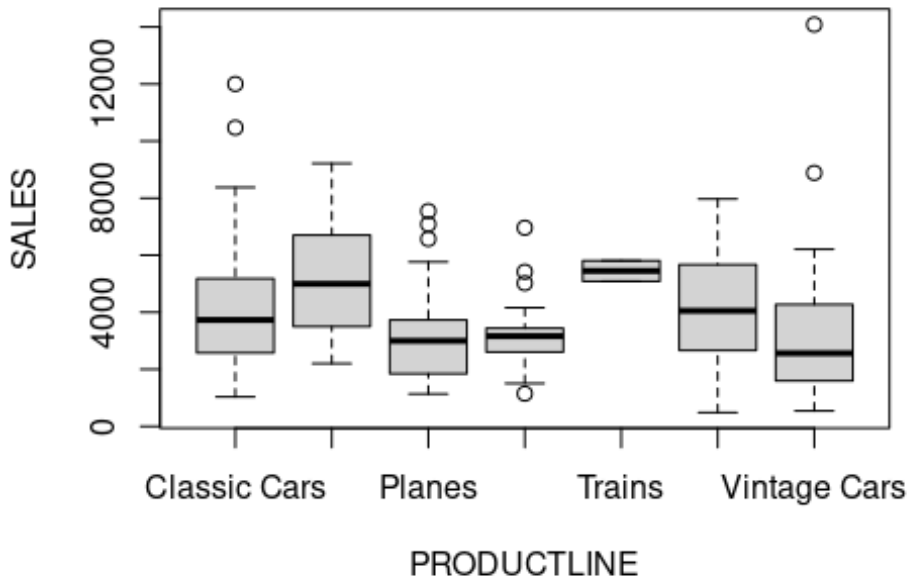
```
##   ORDERNUMBER    QUANTITYORDERED   PRICEEACH      ORDERLINENUMBER
##  Min.   :10164   Min.   : 6.0    Min.   : 26.88   Min.   : 1.000
##  1st Qu.:10255   1st Qu.:27.0    1st Qu.: 70.17   1st Qu.: 3.000
##  Median :10386   Median :37.0    Median : 90.37   Median : 6.000
##  Mean   :10341   Mean   :37.2    Mean   : 82.98   Mean   : 6.621
##  3rd Qu.:10415   3rd Qu.:45.0    3rd Qu.:100.00   3rd Qu.:10.000
##  Max.   :10425   Max.   :85.0    Max.   :100.00   Max.   :18.000
##
##      SALES                STATUS      MONTH_ID     YEAR_ID
##  Min.   :  482.1   Cancelled :60   5      :80   Min.   :2003
##  1st Qu.: 2187.4   Disputed  :14   6      :30   1st Qu.:2004
##  Median : 3194.3   In Process:41   4      :27   Median :2005
##  Mean   : 3597.7   On Hold   :44   10     :24   Mean   :2004
##  3rd Qu.: 4701.0   Resolved  :47   3      :18   3rd Qu.:2005
##  Max.   :14082.8   Shipped   : 0   11     :14   Max.   :2005
##                                    (Other):13
##            PRODUCTLINE        CITY         COUNTRY     DEALSIZE
##  Classic Cars    :53   Madrid   :46   USA      :69   Large : 14
##  Motorcycles     : 7   Boras    :16   Spain    :46   Medium:106
##  Planes          :35   Boston   :14   Sweden   :22   Small : 86
##  Ships           :39   Liverpool:14   Australia:18
```

```
##  Trains          : 2   NYC       :14   UK        :14
##  Trucks and Buses:20   Chatswood:13   France    :13
##  Vintage Cars    :50   (Other)  :89   (Other)   :24
```

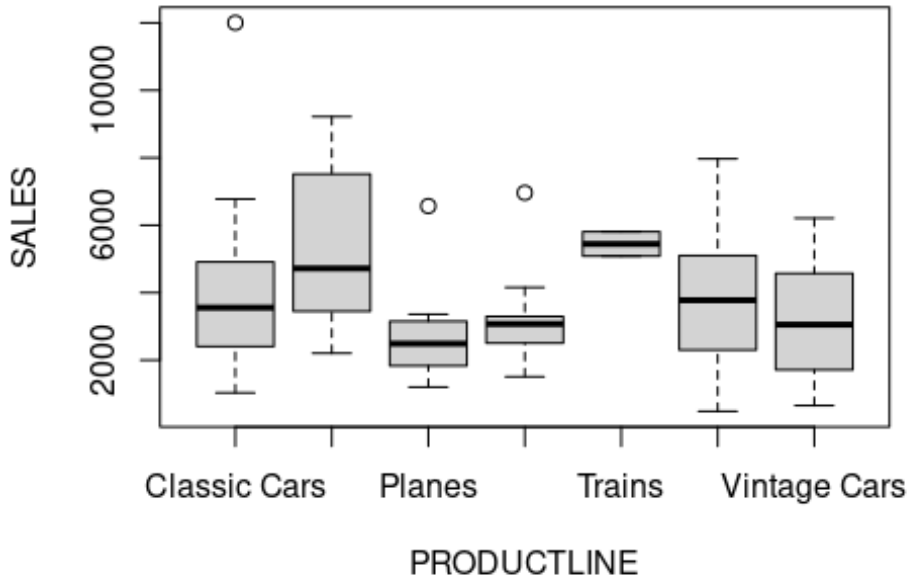
```
dfm1=subset(dfst1,MONTH_ID==5|dfst1$MONTH_ID==6)
summary(dfm1)
```

```
##   ORDERNUMBER    QUANTITYORDERED   PRICEEACH      ORDERLINENUMBER
##  Min.   :10248   Min.   :10.00   Min.   : 26.88   Min.   : 1.000
##  1st Qu.:10255   1st Qu.:26.00   1st Qu.: 70.53   1st Qu.: 3.000
##  Median :10414   Median :35.00   Median : 91.85   Median : 6.000
##  Mean   :10354   Mean   :35.52   Mean   : 83.54   Mean   : 6.464
##  3rd Qu.:10421   3rd Qu.:44.00   3rd Qu.:100.00   3rd Qu.:10.000
##  Max.   :10425   Max.   :66.00   Max.   :100.00   Max.   :16.000
##
##      SALES                STATUS      MONTH_ID     YEAR_ID
##  Min.   :  482.1   Cancelled :44   5      :80   Min.   :2004
##  1st Qu.: 2204.1   Disputed  :11   6      :30   1st Qu.:2004
##  Median : 3167.3   In Process:41   1      : 0   Median :2005
##  Mean   : 3500.2   On Hold   :14   2      : 0   Mean   :2005
##  3rd Qu.: 4426.0   Resolved  : 0   3      : 0   3rd Qu.:2005
##  Max.   :12001.0   Shipped   : 0   4      : 0   Max.   :2005
##                                    (Other): 0
##            PRODUCTLINE        CITY         COUNTRY     DEALSIZE
##  Classic Cars    :30   Madrid   :28   USA       :32   Large : 5
##  Motorcycles     : 6   Boston   :14   Spain     :28   Medium:57
##  Planes          :11   Liverpool:14   Australia:18   Small :48
##  Ships           :19   NYC      :14   UK        :14
##  Trains          : 2   Chatswood:13   France    :13
##  Trucks and Buses:11   Nantes   :13   Belgium   : 5
##  Vintage Cars    :31   (Other)  :14   (Other)   : 0
```

```
boxplot(SALES~PRODUCTLINE,data = dfst1)
```




```
boxplot(SALES~PRODUCTLINE,data = dfm1)
```



INFERENCE:

- we get a idea of the most of the orders are selling in the month of 10 and 11
- Most of the shipped products are having the medium deal size which is get ordered for the product of classic cars
- The classic cars and vintage cars holds the high sales in a certain month
- Most of the items are sold to the country USA is concluded

INSIGHTS:

- By comparing with the every month in the barplot ,the maximum of product sold in the month of 10 and 11
- In the month of 10 and 11 the classic and vintage cars are sold out to nearly
- 300 and 200 respectively
- From the scatter plot the sales gradually increase when the quantity of the product is increased.
- The unshipped product for vintage cars and classic cars are low as compared to other product_line

CONCLUSION:

So,the purchase of classic and vintage cars are more important for the next stock purchase

By

Sachin.k

22CSEG26

Ist M.sc Data Analytics