DATA ANALYSIS

- **#Load libraries**

  library(dplyr)

- **#Load data set**

  data <- read.csv("sales_data_sample.csv",stringsAsFactors = TRUE)

- **#Basic information**

  str(data)

  output:

  data.frame': 2823 obs. of 25 variables:
  $ ORDERNUMBER : int 10107 10121 10134 10145 10159 10168 10180 10188 10201 10211 ...
  $ QUANTITYORDERED : int 30 34 41 45 49 36 29 48 22 41 ...
  $ PRICEEACH : num 95.7 81.3 94.7 83.3 100 ...
  $ ORDERLINENUMBER : int 2 5 2 6 14 1 9 1 2 14 ...
  $ SALES : num 2871 2766 3884 3747 5205 ...
  $ ORDERDATE : Factor w/ 252 levels "1/10/2003 0:00",..: 113 186 205 227 24 40 49 57 83 5 ...
  $ STATUS : Factor w/ 6 levels "Cancelled","Disputed",..: 6 6 6 6 6 6 6 6 6 6 ...
  $ QTR_ID : int 1 2 3 3 4 4 4 4 4 1 ...
  $ MONTH_ID : int 2 5 7 8 10 10 11 11 12 1 ...
  $ YEAR_ID : int 2003 2003 2003 2003 2003 2003 2003 2003 2003 2004 ...
  $ PRODUCTLINE : Factor w/ 7 levels "Classic Cars",..: 2 2 2 2 2 2 2 2 2 2 ...
  $ MSRP : int 95 95 95 95 95 95 95 95 95 95 ...
  $ PRODUCTCODE : Factor w/ 109 levels "S10_1678","S10_1949",..: 1 1 1 1 1 1 1 1 1 1 ...

$ CUSTOMERNAME : Factor w/ 92 levels "Alpha Cognac",..: 47 68 48 87 24 81 27 42 58 10 ...
$ PHONE : Factor w/ 91 levels "(02) 5554 67",..: 49 55 17 77 78 80 41 22 79 4 ...
$ ADDRESSLINE1 : Factor w/ 92 levels "Rambla de Catalu\xa4a, 23",..: 60 43 24 57 54 61 10 69 40 20 ...
$ ADDRESSLINE2 : Factor w/ 10 levels "","2nd Floor",..: 1 1 1 1 1 1 1 1 1 1 ...
$ CITY : Factor w/ 73 levels "Aaarhus","Allentown",..: 49 57 53 54 60 13 29 5 60 53 ...
$ STATE : Factor w/ 17 levels "","BC","CA","CT",..: 11 1 1 3 3 3 1 1 3 1 ...
$ POSTALCODE : Factor w/ 74 levels "","10022","10100",..: 2 29 43 51 1 56 33 66 1 42 ...
$ COUNTRY : Factor w/ 19 levels "Australia","Austria",..: 19 7 7 19 19 19 7 12 19 7 ...
$ TERRITORY : Factor w/ 3 levels "APAC","EMEA",..: NA 2 2 NA NA NA 2 2 NA 2 ...
$ CONTACTLASTNAME : Factor w/ 77 levels "Accorti","Ashworth",..: 77 29 18 76 9 31 58 53 49 54 ...
$ CONTACTFIRSTNAME: Factor w/ 72 levels "Mart\xa1n","Adrian",..: 38 55 13 33 33 34 45 66 33 16 ...
$ DEALSIZE : Factor w/ 3 levels "Large","Medium",..: 3 3 2 2 2 2 3 2 3 2 ...

summary(data)

output:

ORDERNUMBER QUANTITYORDERED
Min. :10100 Min. : 6.00
1st Qu.:10180 1st Qu.:27.00
Median :10262 Median :35.00
Mean :10259 Mean :35.09

```
3rd Qu.:10334   3rd Qu.:43.00
Max.   :10425   Max.   :97.00

  PRICEEACH     ORDERLINENUMBER
Min.   : 26.88   Min.   : 1.000
1st Qu.: 68.86   1st Qu.: 3.000
Median : 95.70   Median : 6.000
Mean   : 83.66   Mean   : 6.466
3rd Qu.:100.00   3rd Qu.: 9.000
Max.   :100.00   Max.   :18.000

    SALES                ORDERDATE
Min.   :  482.1   11/14/2003 0:00:  38
1st Qu.: 2203.4   11/24/2004 0:00:  35
Median : 3184.8   11/12/2003 0:00:  34
Mean   : 3553.9   11/17/2004 0:00:  32
3rd Qu.: 4508.0   11/4/2004 0:00 :  29
Max.   :14082.8   10/16/2004 0:00:  28
                  (Other)       :2627
     STATUS        QTR_ID
Cancelled :  60   Min.   :1.000
Disputed  :  14   1st Qu.:2.000
In Process:  41   Median :3.000
On Hold   :  44   Mean   :2.718
Resolved  :  47   3rd Qu.:4.000
Shipped   :2617   Max.   :4.000

   MONTH_ID        YEAR_ID
Min.   : 1.000   Min.   :2003
1st Qu.: 4.000   1st Qu.:2003
Median : 8.000   Median :2004
Mean   : 7.092   Mean   :2004
3rd Qu.:11.000   3rd Qu.:2004
Max.   :12.000   Max.   :2005
```

```
        PRODUCTLINE      MSRP
Classic Cars    :967   Min.   : 33.0
Motorcycles     :331   1st Qu.: 68.0
Planes          :306   Median : 99.0
Ships           :234   Mean   :100.7
Trains          : 77   3rd Qu.:124.0
Trucks and Buses:301   Max.   :214.0
Vintage Cars    :607
  PRODUCTCODE
S18_3232:  52
S10_1949:  28
S10_4962:  28
S12_1666:  28
S18_1097:  28
S18_2432:  28
(Other) :2631
                CUSTOMERNAME
Euro Shopping Channel      : 259
Mini Gifts Distributors Ltd.: 180
Australian Collectors, Co.  :  55
La Rochelle Gifts           :  53
AV Stores, Co.              :  51
Land of Toys Inc.           :  49
(Other)                    :2176
         PHONE
(91) 555 94 44: 259
4155551450    : 180
03 9520 4555  :  55
40.67.8555    :  53
(171) 555-1555:  51
6175558555    :  51
(Other)       :2174
                ADDRESSLINE1
C/ Moralzarzal, 86          : 259
5677 Strong St.             : 180
```

```
636 St Kilda Road         :  55
67, rue des Cinquante Otages:  53
Fauntleroy Circus         :  51
897 Long Airport Avenue    :  49
(Other)                   :2176
  ADDRESSLINE2          CITY
       :2521   Madrid      : 304
Level 3  :  55   San Rafael   : 180
Suite 400:  48   NYC         : 152
Level 15 :  46   Singapore    :  79
Level 6  :  46   Paris        :  70
2nd Floor:  36   San Francisco:  62
(Other)  :  71   (Other)     :1976
    STATE       POSTALCODE
      :1486   28034  : 259
CA     : 416   97562  : 205
MA     : 190   10022  : 152
NY     : 178   94217  :  89
NSW    :  92          :  76
Victoria:  78   50553  :  61
(Other) : 383   (Other):1981
    COUNTRY    TERRITORY    CONTACTLASTNAME
USA     :1004   APAC : 221   Freyre : 259
Spain   : 342   EMEA :1407   Nelson : 204
France  : 314   Japan: 121   Young  : 115
Australia: 185   NA's :1074   Frick  :  91
UK      : 144                Brown  :  88
Italy   : 113                Yu     :  80
(Other) : 721                (Other):1986
CONTACTFIRSTNAME   DEALSIZE
Diego  : 259    Large : 157
Valarie: 257    Medium:1384
Julie  : 117    Small :1282
Michael:  84
Sue    :  84
```

Juri : 60
(Other):1962

head(data)

output:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE	MSRP
1	10107	30	95.70	2	2871.00	2/24/2003 0:00	Shipped	1	2			
2	10121	34	81.35	5	2765.90	5/7/2003 0:00	Shipped	2	5			
3	10134	41	94.74	2	3884.34	7/1/2003 0:00	Shipped	3	7			
4	10145	45	83.26	6	3746.70	8/25/2003 0:00	Shipped	3	8			
5	10159	49	100.00	14	5205.27	10/10/2003 0:00	Shipped	4	10			
6	10168	36	96.66	1	3479.76	10/28/2003 0:00	Shipped	4	10			

1	2003	Motorcycles	95
2	2003	Motorcycles	95
3	2003	Motorcycles	95
4	2003	Motorcycles	95
5	2003	Motorcycles	95
6	2003	Motorcycles	95

	PRODUCTCODE
1	S10_1678
2	S10_1678
3	S10_1678
4	S10_1678
5	S10_1678
6	S10_1678

	CUSTOMERNAME
1	Land of Toys Inc.
2	Reims Collectables
3	Lyon Souveniers
4	Toys4GrownUps.com
5	Corporate Gift Ideas Co.
6	Technics Stores Inc.

	PHONE
1	2125557818
2	26.47.1555
3	+33 1 46 62 7555
4	6265557265
5	6505551386
6	6505556809

	ADDRESSLINE1
1	897 Long Airport Avenue
2	59 rue de l'Abbaye
3	27 rue du Colonel Pierre Avia
4	78934 Hillside Dr.
5	7734 Strong St.
6	9408 Furth Circle

	ADDRESSLINE2	CITY

1	NYC
2	Reims
3	Paris
4	Pasadena
5	San Francisco
6	Burlingame

	STATE	POSTALCODE	COUNTRY
1	NY	10022	USA
2		51100	France
3		75508	France
4	CA	90003	USA
5	CA		USA
6	CA	94217	USA

	TERRITORY	CONTACTLASTNAME
1	None	Yu
2	EMEA	Henriot
3	EMEA	Da Cunha
4	None	Young
5	None	Brown
6	None	Hirano

	CONTACTFIRSTNAME	DEALSIZE
1	Kwai	Small
2	Paul	Small
3	Daniel	Medium
4	Julie	Medium
5	Julie	Medium
6	Juri	Medium

- **# Check the missing values**

  sum(is.na(data))

  output:

  ```
  [1] 1074
  ```

```
missing_values <- colSums(is.na(data))

missing_values[missing_values > 0]
```

output:

```
TERRITORY
 1074
```

```
sapply(data$TERRITORY,class)
```

- **# Missing values handling**

```
# Step 1: Convert factor column to character
data$TERRITORY <- as.character(data$TERRITORY)

# Step 2: Replace missing values with "None"
data$TERRITORY[is.na(data$TERRITORY)] <- "None"

# Step 3: Convert back to factor
data$TERRITORY <- as.factor(data$TERRITORY)

sum(is.na(data))
```

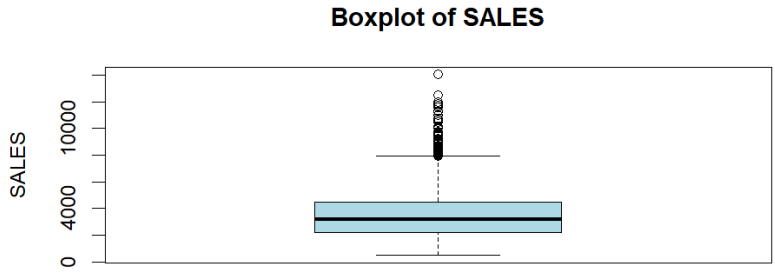
output:

```
[1] 0
```

- **# check out layers**

```
for (col in names(data)){
 if(is.numeric(data[[col]])){
  boxplot(data[[col]],
      main = paste("Boxplot of", col),
      ylab = col,
      col = "lightblue")
 }
}
```

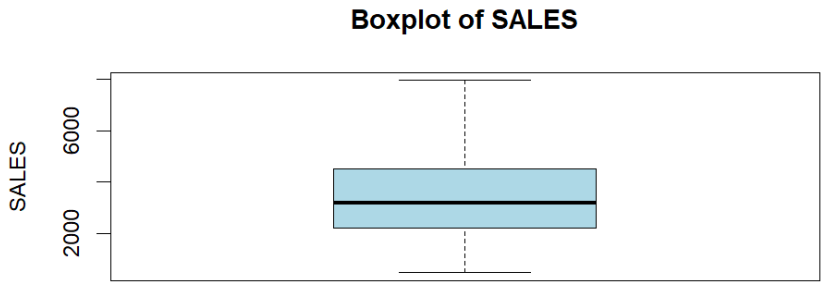
Output:



- **# Handling out layers**

```
for (col in names(data)) {
 if (is.numeric(data[[col]])) {

   Q1 <- quantile(data[[col]],0.25,na.rm = TRUE)
   Q3 <- quantile(data[[col]],0.75,na.rm = TRUE )
   IQR <- Q3 - Q1

   lower_bound <- Q1 - 1.5*IQR
   upper_bound <- Q3 + 1.5*IQR


   data[[col]] <- ifelse(data[[col]]< lower_bound, lower_bound,
                         ifelse(data[[col]] > upper_bound, upper_bound,
                                data[[col]]))
 }
}
```

Output:


Boxplot of SALES
SALES
6000
2000

DATA VISUALIZATION

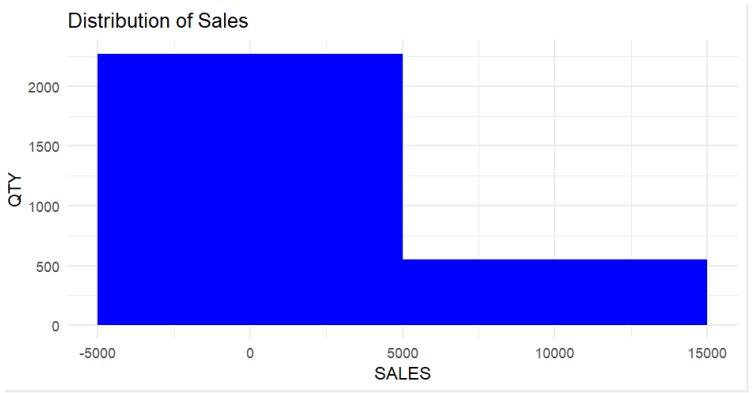
- **#Load libraries**

```
library(ggplot2)
```

- **Exploratory Data Analysis**

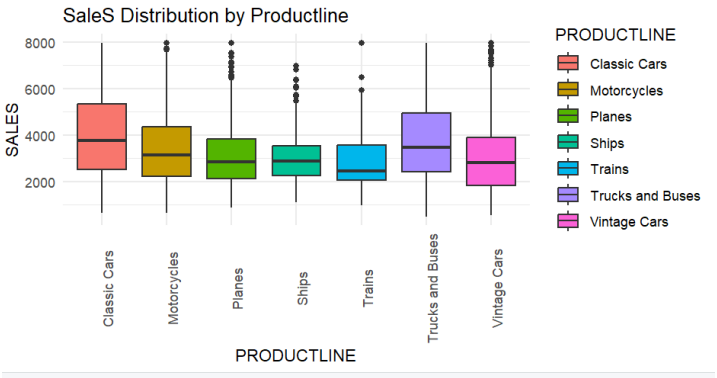
```
# Histogram of Sales
ggplot(data, aes(x = SALES)) +
 geom_histogram(binwidth = 10000, fill = "blue") +
 theme_minimal() +
 labs(title = "Distribution of Sales", x = "SALES", y = "QTY")
```

output:



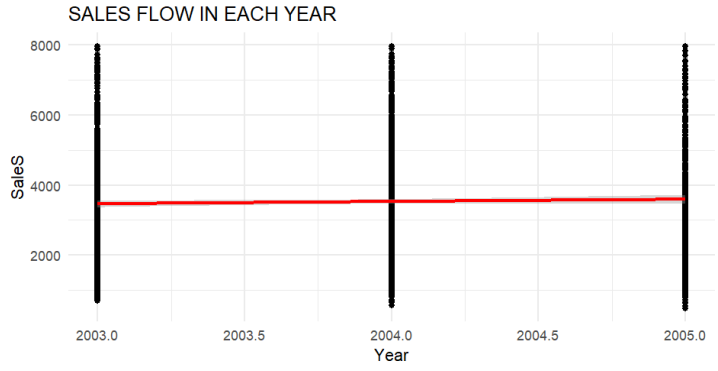

```
# Boxplot of Sale by productline
ggplot(data, aes(x = PRODUCTLINE, y = SALES, fill = PRODUCTLINE)) +
 geom_boxplot() +
 theme_minimal() +
 labs(title = "SaleS Distribution by Productline") +
 theme(axis.text.x = element_text(angle = 90))
```

output:




```
# Line plot of sales over the years
ggplot(data, aes(x = YEAR_ID, y = SALES)) +
 geom_point() +
 geom_smooth(method = "lm", col = "red") +
 theme_minimal() +
 labs(title = "SALES FLOW IN EACH YEAR", x = "Year", y = "SaleS")
```

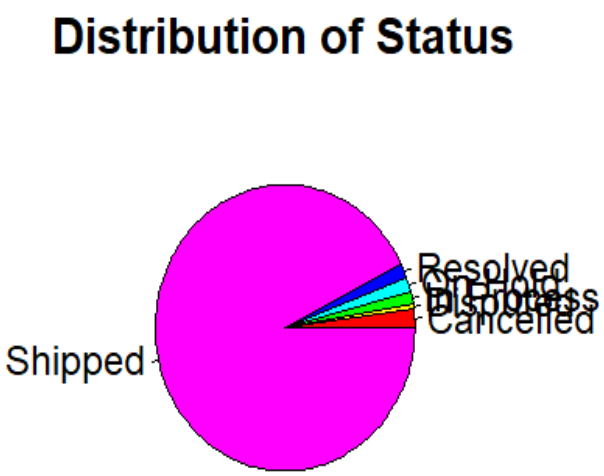
output:



- **# Count the frequency of each category in MSZoning**

```
status_counts <- table(data$STATUS)

# Create a simple pie chart

pie(status_counts, main = "Distribution of Status", col = rainbow(length(status_counts)))

data$DEALSIZE
```

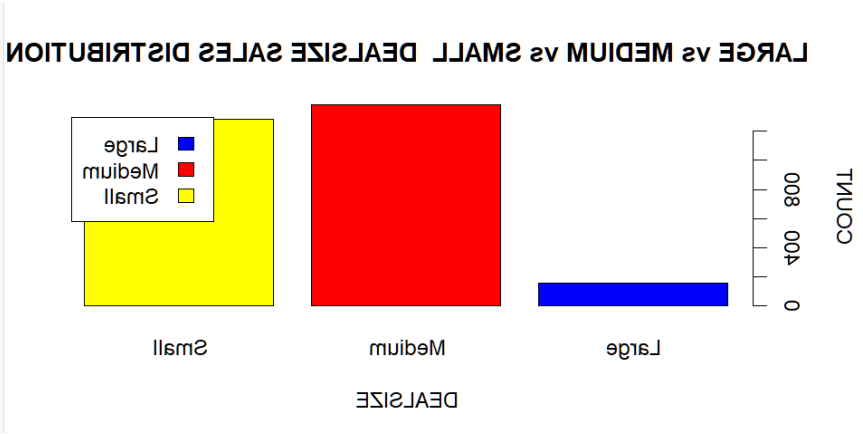
output:



- **# Count the number of High and Low priced houses**

```
DEALSIZE_COUNTS <- table(data$DEALSIZE)

# Create a bar chart

barplot(DEALSIZE_COUNTS,

    main = "LARGE vs MEDIUM vs SMALL  DEALSIZE SALES DISTRIBUTION",

    col = c("blue", "red","yellow"),

    xlab = "DEALSIZE",

    ylab = "COUNT",

    legend = TRUE)

DEALSIZE_COUNTS
```

Output:

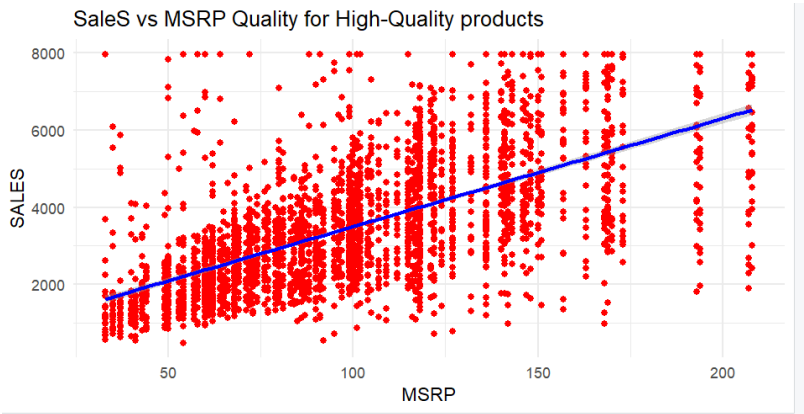


- **DEALSIZE_COUNTS**

```
Large Medium  Small
  157   1384   1282
```

```
ggplot(data , aes(x = MSRP, y = SALES)) +
 geom_point(color = "red") +
 geom_smooth(method = "lm", col = "blue") +
 labs(title = "SaleS vs MSRP Quality for High-Quality products",
    x = "MSRP",
    y = "SALES") +
 theme_minimal()
```

output:



MODEL BUILDING

- **#Load libraries**

```
library(caret)
```

- **# step1:Drop unwanted columns**

```
data_model <-data%>%select(-ORDERNUMBER,-
ORDERLINENUMBER,-PRODUCTCODE,-CUSTOMERNAME,-
CONTACTLASTNAME,-CONTACTFIRSTNAME,-PHONE)
```

- **# step2:partition data**

```
set.seed(123)
trainIndex <- createDataPartition(data_model$SALES, p = 0.8, list =
FALSE)
train_data <- data_model[trainIndex, ]
test_data  <- data_model[-trainIndex, ]
train_control <- trainControl(
 method = "cv", # cross-validation
 number = 5,   #5-fold CV
 verboseIter = TRUE
)
```

- **# step3:Train a model, Start with Linear Regression**

```
set.seed(123)
lm_model <- train(
 SALES ~ .,
 data = train_data,
 method = "lm",
 preProcess = c("center", "scale", "zv", "nzv"),
 trControl = train_control
)
print(lm_model)
```

output:

+ Fold1: intercept=TRUE

- Fold1: intercept=TRUE

+ Fold2: intercept=TRUE

- Fold2: intercept=TRUE

+ Fold3: intercept=TRUE

- Fold3: intercept=TRUE

+ Fold4: intercept=TRUE

- Fold4: intercept=TRUE

+ Fold5: intercept=TRUE

- Fold5: intercept=TRUE

Aggregating results

Fitting final model on full training set

- **# Evaluation on Test data**

  pred_lm <- predict(lm_model, newdata = test_data)

- **# RMSE and R^2**

  postResample(pred = pred_lm, obs = test_data$SALES)

  output:

```
    RMSE          Rsquared
592.6248155     0.8821663
     MAE
425.3852273
```