**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TOPIC** | **Pg.No** |
| **1** | **DESCRIPTIVE ANALYSIS USING R** | **3** |
| **2** | **EXPLORING THE DATASET** | **4** |
| **3** | **DESCRIPTIVE STATISTICS** | **8** |
| i) | MEAN | **8** |
| ii) | STADARD DEVIATION | **9** |
| iii) | RANGE | **10** |
| iv) | QUANTILE | **11** |
| **4** | **DATA VISUALIZATION** | **12** |
| i) | SCATTER PLOT | **12** |
| ii) | HISTOGRAM | **13** |
| **4** | **SYNTAX** | **14** |

**1.DESCRIPTIVE ANALYSIS USING R**

**DESCRIPTIVE ANALYSIS**

**Descriptive analytics** is the process of using current and historical data to identify trends and relationships. It’s sometimes called the simplest form of data analysis because it describes trends and relationships but doesn’t dig deeper.

Descriptive analytics is especially useful for communicating change over time and uses trends as a springboard for further analysis to drive decision-making.

**DESCRIPTIVE ANALYSIS USING R**

In Descriptive analysis, we are describing our data with the help of various representative methods like using charts, graphs, tables, excel files, etc. In the descriptive analysis, we describe our data in some manner and present it in a meaningful way so that it can be easily understood. Most of the time it is performed on small data sets and this analysis helps us a lot to predict some future trends based on the current findings. Some measures that are used to describe a data set are measures of central tendency and measures of variability or dispersion.

**DATASET**

The chosen dataset is **Imports of Major Petrochemical**

Data Figure Quantity in MT and Values in Rs Lakh. Data source in respect of Imports of Chemicals is Directorate General of Commercial Intelligence and Statistics (DGCIS), Kolkata. Product - It includes those chemical products which are being monitored by SM Division of D/o Chemicals Petrochemicals.

**2.EXPLORING THE DATASET**

**IMPORTING**

Before doing any computation, first of all, we need to prepare our data, save our data in external .txt or .csv files and it’s a best practice to save the file in the current directory. After that import, our data into R as follow:

**>mydata=read.csv(file.choose())**

**# To read the dataset**

**>mydata**

**>ncol(mydata)**

**# To know the number of columns in my data set**

## [1] 18

**>nrow(mydata)**

**# To know the number of rows in my dataset**

## [1] 70

**>str(mydata)**

**# To see the structure of my data**

'data.frame': 70 obs. of 18 variables:

$ Group : chr "Synthetic Fibres" "Synthetic Fibres" "Synthetic Fibres" "Synthetic Fibres" ...

$ Product : chr "Acrylic Fibre" "Nylon Filament Yarn" "Nylon Industrial Yarn/Tyre Cord" "Polyester Filament Yarn" ...

$ X2014.15.QTY : int 36339 27813 3750 72718 81864 716 329 12061 97388 32276 ...

$ X2014.15..VAL: int 60959 59448 7035 89105 71552 1002 461 56643 118956 43571 ...

$ X2015.16.QTY : int 39238 28972 4411 72605 99719 697 1885 16247 160202 44577 ...

$ X2015.16..VAL: int 53809 54485 6803 86166 73357 1001 1997 68750 121692 45299 ...

$ X2016.17.QTY : int 36771 31063 7741 84367 99809 695 2016 17531 139709 52906 ...

$ X2016.17..VAL: int 44394 53352 10247 92492 70113 850 2181 69364 108192 50601 ...

$ X2017.18.QTY : int 32184 22731 6566 82784 88990 1343 3281 21753 159778 58413 ...

$ X2017.18..VAL: int 46549 43317 10070 100756 71053 1715 3378 86241 172707 75965 ...

$ X2018.19.QTY : int 40475 23259 8108 83954 91775 1362 3333 23351 182234 66497 ...

$ X2018.19..VAL: int 71158 52088 16323 118045 89685 2007 3863 100187 244043 99336 ...

$ X2019.20.QTY : int 56316 24476 6893 115170 118176 670 2778 22121 175867 68049 ...

$ X2019.20..VAL: int 84841 49669 11802 129162 94729 954 3097 86629 200713 75256 ...

$ X2020.21.QTY : int 41928 17058 9459 238680 84273 755 2164 8956 135230 57649 ...

$ X2020.21..VAL: int 53586 34533 15174 197017 64616 1418 2316 42041 121075 55553 ...

$ X2021.22.QTY : int 35164 25307 11445 295709 78310 315 2986 11141 175763 60061 ...

$ X2021.22..VAL: int 76074 60951 31261 343248 78489 975 4218 83304 298158 94382 ...

**>names(mydata)**

**# To list the variables of the dataset**

[1] "Group" "Product" "X2014.15.QTY" "X2014.15..VAL" "X2015.16.QTY"

[6] "X2015.16..VAL" "X2016.17.QTY" "X2016.17..VAL" "X2017.18.QTY" "X2017.18..VAL"

[11] "X2018.19.QTY" "X2018.19..VAL" "X2019.20.QTY" "X2019.20..VAL" "X2020.21.QTY"

[16] "X2020.21..VAL" "X2021.22.QTY" "X2021.22..VAL"

**>dim(mydata)**

**# To see the dimension of my data**

# [1] 70 18

**>head(mydata)**

**# To display first 6 rows of dataset**

Group Product X2019.20.QTY X2019.20..VAL X2020.21.QTY

1 Synthetic Fibres Acrylic Fibre 56316 84841 41928

2 Synthetic Fibres Nylon Filament Yarn 24476 49669 17058

3 Synthetic Fibres Nylon Industrial Yarn/Tyre Cord 6893 11802 9459

4 Synthetic Fibres Polyester Filament Yarn 115170 129162 238680

5 Synthetic Fibres Polyester Staple Fibre 118176 94729 84273

6 Synthetic Fibres Polypropylene Filament Yarn 670 954 755

X2020.21..VAL X2021.22.QTY X2021.22..VAL

1 53586 35164 76074

2 34533 25307 60951

3 15174 11445 31261

4 197017 295709 343248

5 64616 78310 78489

6 1418 315 975

**SUMMARY**

Summary statistics are the first figures used to represent nearly every dataset. They also form the foundations for more complicated computations and analyses. Therefore, they are essential to the analysis process.

**>summary(mydata)**

**# To get the summary of the dataset**

Group Product X2019.20.QTY X2019.20..VAL X2020.21.QTY

Length:70 Length:70 Min. : 0 Min. : 0 Min. : 0

Class :character Class :character 1st Qu.: 10536 1st Qu.: 8225 1st Qu.: 4129

Mode :character Mode :character Median : 57590 Median : 59784 Median : 49246

Mean : 349209 Mean : 260038 Mean : 291739

3rd Qu.: 189220 3rd Qu.: 209919 3rd Qu.: 176708

Max. :12222311 Max. :9101336 Max. :10210873

X2020.21..VAL X2021.22.QTY X2021.22..VAL

Min. : 0 Min. : 0 Min. : 0

1st Qu.: 6056 1st Qu.: 4922 1st Qu.: 8862

Median : 53079 Median : 48488 Median : 78648

Mean : 204559 Mean : 335191 Mean : 343124

3rd Qu.: 151418 3rd Qu.: 178159 3rd Qu.: 265710

Max. :7159554 Max. :11731702 Max. :12009342

**3.DESCRIPTIVE STATISTICS**

**i) MEAN**

It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

The mean is used **to summarize a data set**. It is a measure of the center of a data set.

* **Mean of variable 2019.2020**

**>mean(mydata$X2019.20.QTY)**

**# To calculate the mean of the variable 2019.20.QTY**

[1] 349208.9

* **Mean of variable 2020.21**

**>mean(mydata$X2020.21.QTY)**

**# To calculate the mean of the variable 2020.21 Qty**

[1] 291739.2

* **Mean of variable 2021.22**

**>mean(mydata$X2021.22.QTY)**

**# To calculate the mean of the variable 2021.22**

[1] 335191.5

**ii) Standard deviation**

Standard deviation **calculates the extent to which the values differ from the average**. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore, a change in even one value affects the value of standard deviation. It is independent of origin but not of scale.

* **Standard of deviation of 2019.20**

**>sd(mydata$X2019.20.QTY)**

**# To calculate the standard deviation of 2019.20**

[1] 1465101

* **Standard of deviation of 2020.21**.

**>** **sd(mydata$X2020.21.QTY)**

**# To calculate the standard deviation of 2020.21**

[1] 1222887

* **Standard of deviation of 2021.22**.

**>** **sd(mydata$X2021.22.QTY)**

**# To calculate the standard deviation of 2021.22**

[1] 1412593

**iii) RANGE of EACH VARIABLE: MINIMUM & MAXIMUM**

Continuing in the same trajectory, minimum can be computed on a single variable using the min(VAR) formula. In the same token, max(VAR) operates similarly. Minimum and Maximum give the min and max of individual variables in the dataset.

* **Minimum and maximum imports of 2019.20**

**>** **min(mydata$X2019.20.QTY);max(mydata$X2019.20.QTY)**

**# To get the minimum and maximum imports of 2019.20**

[1] 0

[1] 12222311

* **Minimum and maximum imports of 2020.21**

**>** **min(mydata$X2020.21.QTY);max(mydata$X2020.21.QTY)**

**# To get the minimum and maximum imports of 2020.21**

[1] 0

[1] 10210873

* **Minimum and maximum imports of 2021.22**

**>** **min(mydata$X2021.22.QTY);max(mydata$X2021.22.QTY)**

**# To get the minimum and maximum imports of 2021.22**

[1] 0

[1] 1173170

**iv) Quantile**

We can get more insight into the distribution of a set of observations by examining quantiles. A quantile is a value computed from a collection of numeric measurements that indicates an observation’s rank when compared to all other present observations. Alternatively, quantile can be expressed as a percentile, this is identical but on a percent scale of 0 to 100.

* **Quantile of 2019.20**

**>quantile(mydata$X2019.20.QTY)**

**# To find the quantiles of 2019.20**

0% 25% 50% 75% 100%

0.00 10535.75 57590.50 189220.00 12222311.00

* **Quantile of 2020.21**

**>** **quantile(mydata$X2020.21.QTY)**

**# To find the quantiles of 2020.21**

0% 25% 50% 75% 100%

0.0 4129.0 49246.5 176708.0 10210873.0

* **Quantile of 2021.22**

**>** **quantile(mydata$X2021.22.QTY)**

**# To find the quantiles of 2021.22**

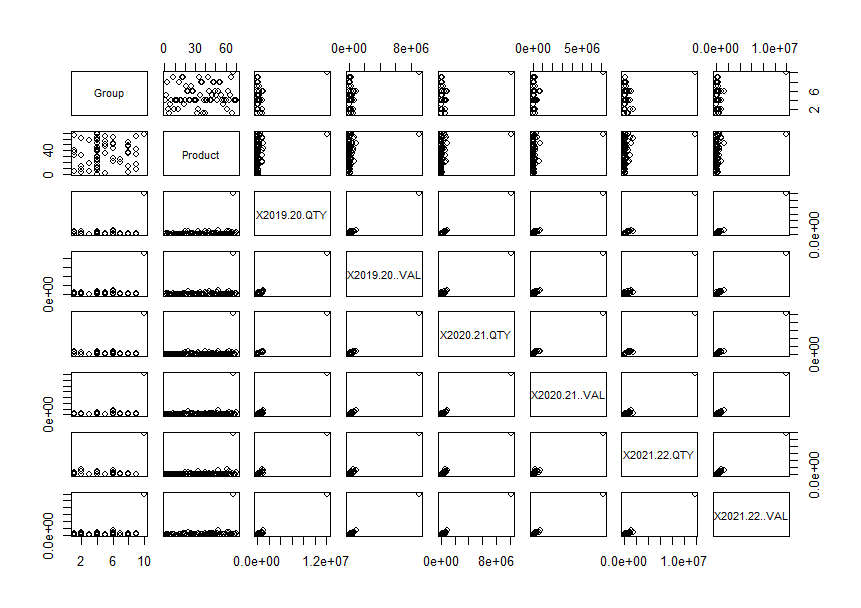
0% 25% 50% 75% 100%

0.00 4922.25 48488.50 178159.25 11731702.00

**4. DATA VISUALIZATION**

**i) MATRIX OF PLOTS**

The single type of planar scatterplot is really useful only when comparing two numeric-continuous variables. When there are more continuous variables of interest, it is possible to display this information satisfactorily on a single plot. A simple and common solution is to generate a two-variable scatterplot for each pair of variables and show them together in a structured way; this is referred to as a Scatterplot Matrix.

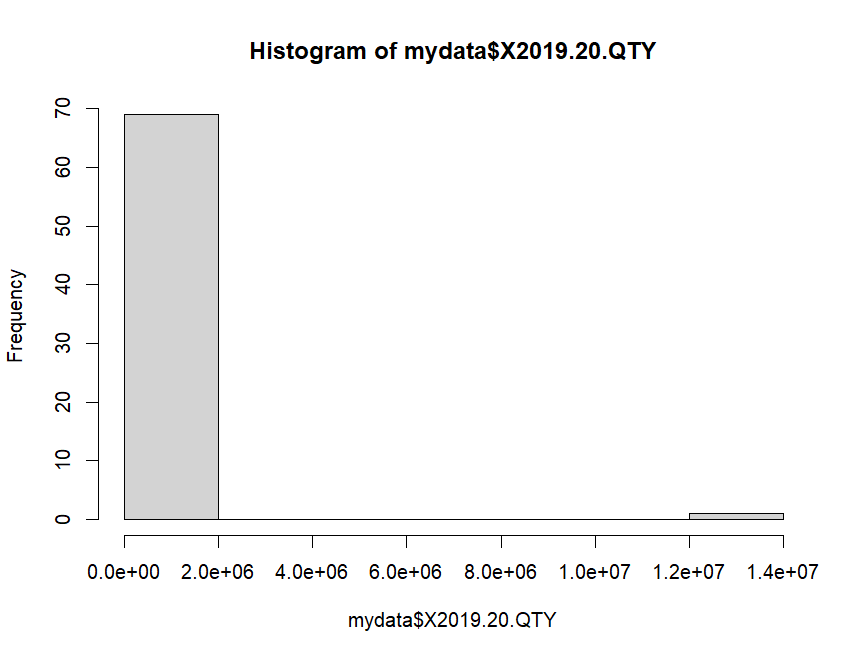
**>plot(mydata)**

**ii).HISTOGRAM**

A histogram represents the frequencies of values of a variable bucketed into ranges. Histogram is similar to bar chat but the difference is it groups the values into continuous ranges. Each bar in histogram represents the height of the number of values present in that range.

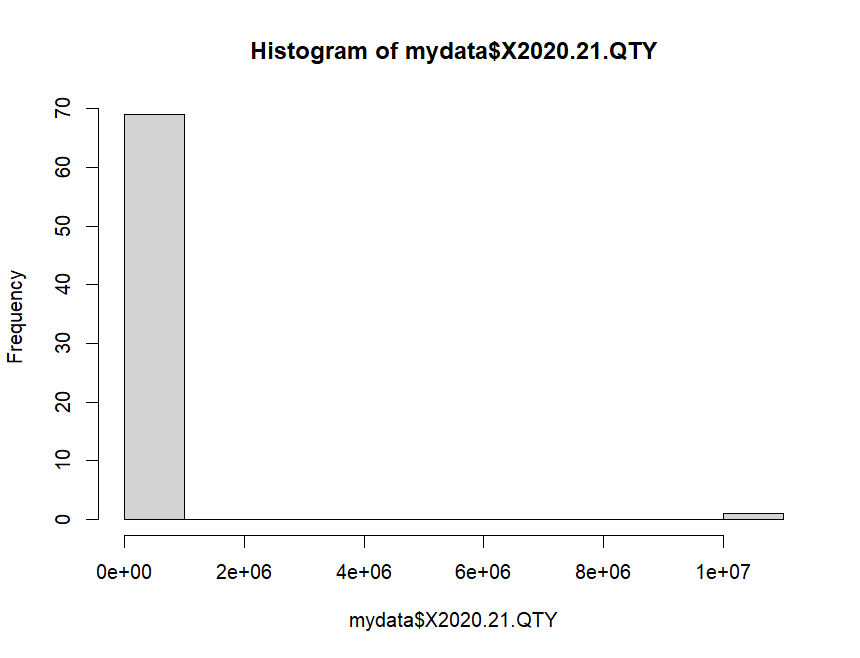
**>hist(mydata$X2019.20.QTY)**

**# To get histogram for the year 2019.20**



**>hist(mydata$X2020.21.QTY)**

**# To get histogram for the year 2019.20**



**5. SYNTAX USED IN THIS REPORT FOR THE DESCRIPTIVE ANALYSIS OF THE DATASET**

* mydata=read.csv(file.choose())
* mydata
* ncol(mydata)
* nrow(mydata)
* str(mydata)
* names(mydata)
* dim(mydata)
* head(mydata)
* summary(mydata)
* mean(mydata$X2019.20.QTY)
* mean(mydata$X2020.21.QTY)
* mean(mydata$X2021.22.QTY)
* sd(mydata$X2019.20.QTY)
* sd(mydata$X2020.21.QTY)
* sd(mydata$X2021.22.QTY)
* min(mydata$X2019.20.QTY);max(mydata$X2019.20.QTY)
* min(mydata$X2020.21.QTY);max(mydata$X2020.21.QTY)
* min(mydata$X2021.22.QTY);max(mydata$X2021.22.QTY)
* quantile(mydata$X2019.20.QTY)
* quantile(mydata$X2020.21.QTY)
* quantile(mydata$X2021.22.QTY)
* plot(mydata$X2020.21.QTY)
* plot(mydata$X2021.22.QTY)
* plot(mydata)
* hist(mydata$X2019.20.QTY)