Import the data

Clean the data

Process the data

Visualize the data

Import the files using read csv

Combine them using rbind

involves exploring a dataset in three ways

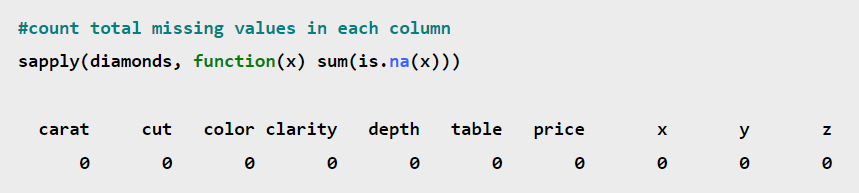
1. Summarizing a dataset using descriptive statistics – use summary() function, dim()
2. Visualizing a dataset using charts – using geom() functions

#create correlation matrix of (rounded to 2 decimal places)

round(cor(diamonds[c('carat', 'depth', 'table', 'price', 'x', 'y', 'z')]), 2)

1. Identifying missing values

We can use the following code to count the total number of missing values in each column of the dataset



Using IBM’s Explore procedure, you can:

* Screen data
* Identify outliers
* Check assumptions
* Characterize differences among groups of cases

Notes

Insights

1. About ratings

4 to 6 – low

6.1 to 8 – medium

8.1 to 10 – high

1. Which customers buy the most – members, normal, male or female
2. How many product lines do they have & which one earned the most revenue
3. Which month had highest revenue
4. What hour do they shop most
5. Which mode of payment is most/least common so they can improve

hour payment invoice.id

1 10 Card 35

2 11 Card 25

3 12 Card 30

4 13 Card 27

5 14 Card 25

6 15 Card 30

7 16 Card 21

8 17 Card 27

9 18 Card 31

10 19 Card 36

11 20 Card 24

12 10 Cash 39

13 11 Cash 34

14 12 Cash 29

15 13 Cash 40

16 14 Cash 27

17 15 Cash 35

18 16 Cash 23

19 17 Cash 29

20 18 Cash 30

21 19 Cash 35

22 20 Cash 23

23 10 Epay 27

24 11 Epay 31

25 12 Epay 30

26 13 Epay 36

27 14 Epay 31

28 15 Epay 37

29 16 Epay 33

30 17 Epay 18

31 18 Epay 32

32 19 Epay 42

33 20 Epay 28

1. What are your recommendations?

The data was obtained from a retail chain stores that sell different products to various clientele. Ms Perera recently joined as a new business partner and she together with management are interested to know the business’ position before she can be fully invested.

The exploratory data analysis will also be useful in making some decisions which could better the business operations. Tools used for analysis are R & Excel.

y=total revenues

x= product line

fill=customer type

110731.46

"" "Fashion accessories" "Electronic accessories"

## [4] "Sports and travel" "Home and lifestyle" "Health and beauty"