SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

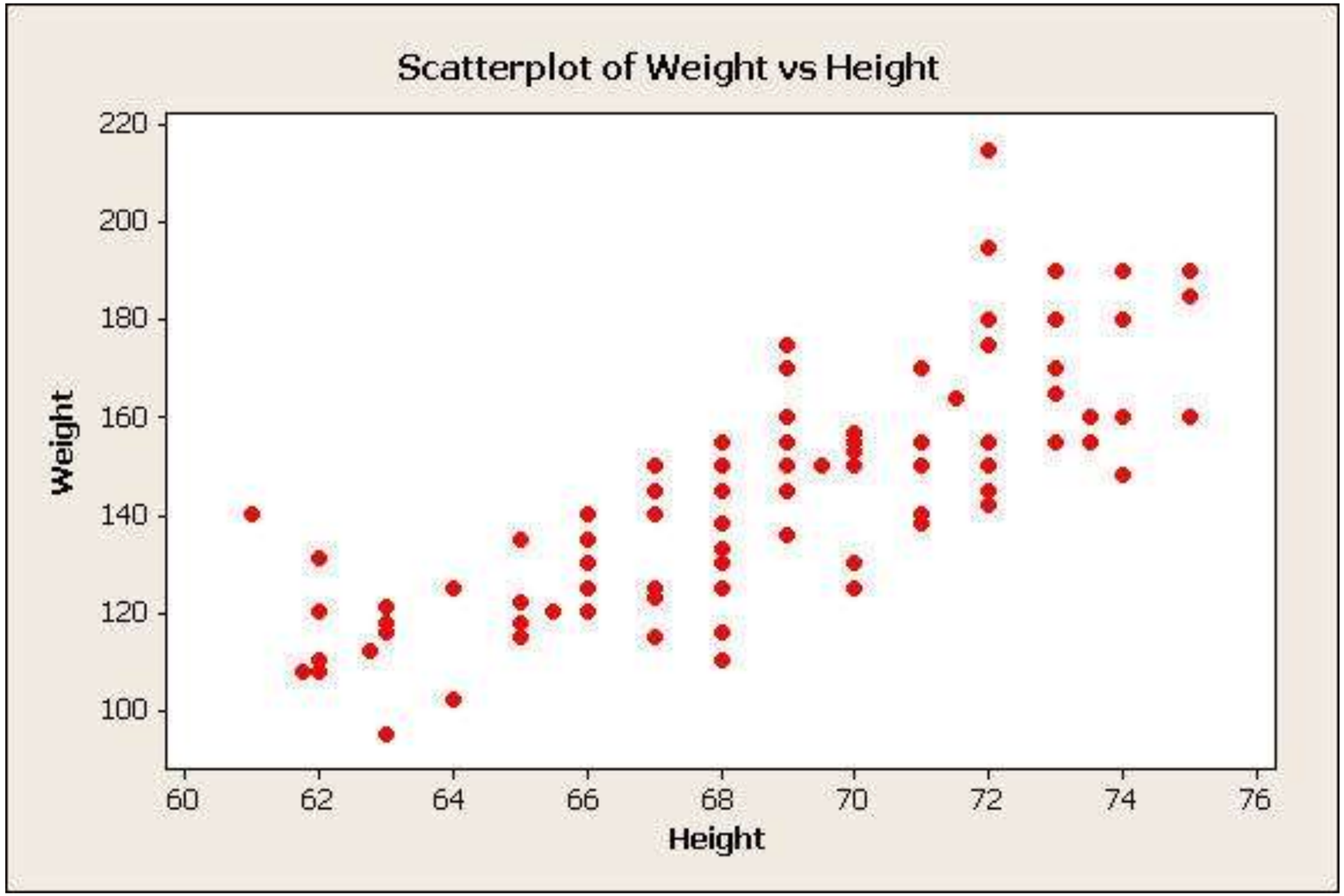
COURSE NAME :19CS407 DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit 2- GETTING INSIGHTS FROM DATA, DATA QUALITY AND PREPROCESSING

Topic : Descriptive Bivariate Analysis

Caloric Intake X	Weight Y
3500	250lbs
2000	225lbs
1500	110lbs
2250	145lbs
4500	380lbs


Scatterplot of Weight vs Height
Weight
Height
220
200
180
160
140
120
100
60
62
64
66
68
70
72
74
76

Descriptive Bivariate Analysis

- It is organized according to the scale types of the attributes: quantitative, nominal and ordinal. When one of the attributes of the pair is qualitative – that is, nominal or ordinal – and the other is quantitative, box plots can be used

Two Quantitative Attributes

- In a data set whose objects have n attributes, each object can be represented in a n-dimensional space: a space with n axes, each axis representing one of the attributes.
- The position occupied by an object is given by the value of its attributes.
- There are several visualization techniques that can visually show the distribution of points with two quantitative attributes. One of these techniques is an extension of the histogram called a three-dimensional histogram

Two Quantitative Attributes

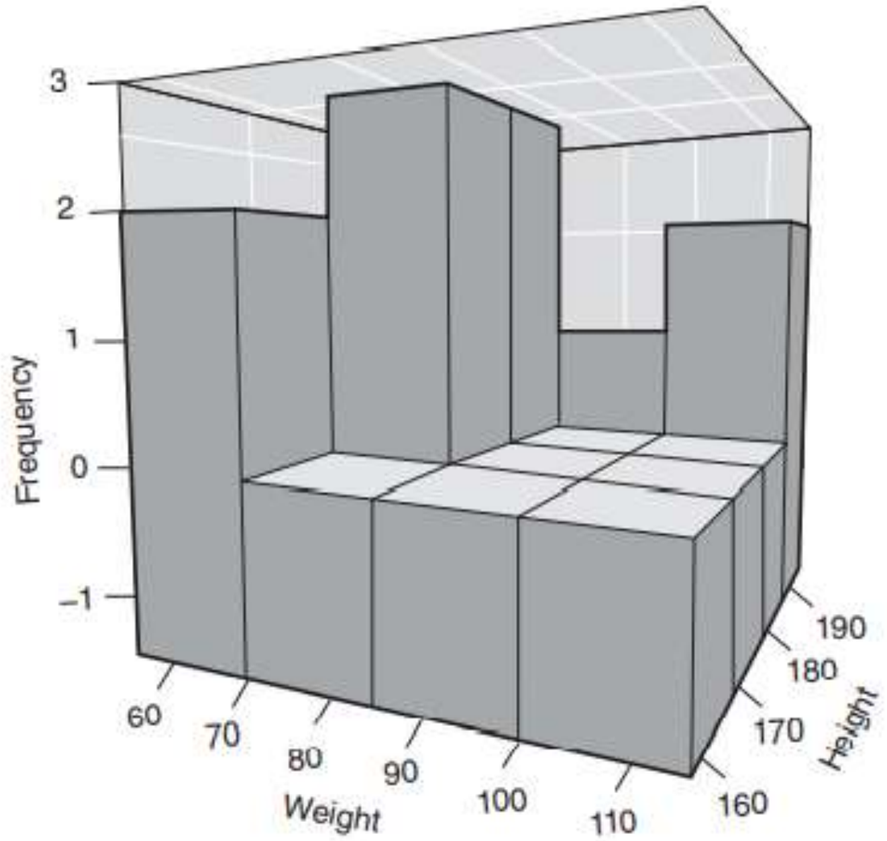


Figure 2.14 3D histogram for attributes "weight" and "height".

Two Quantitative Attributes

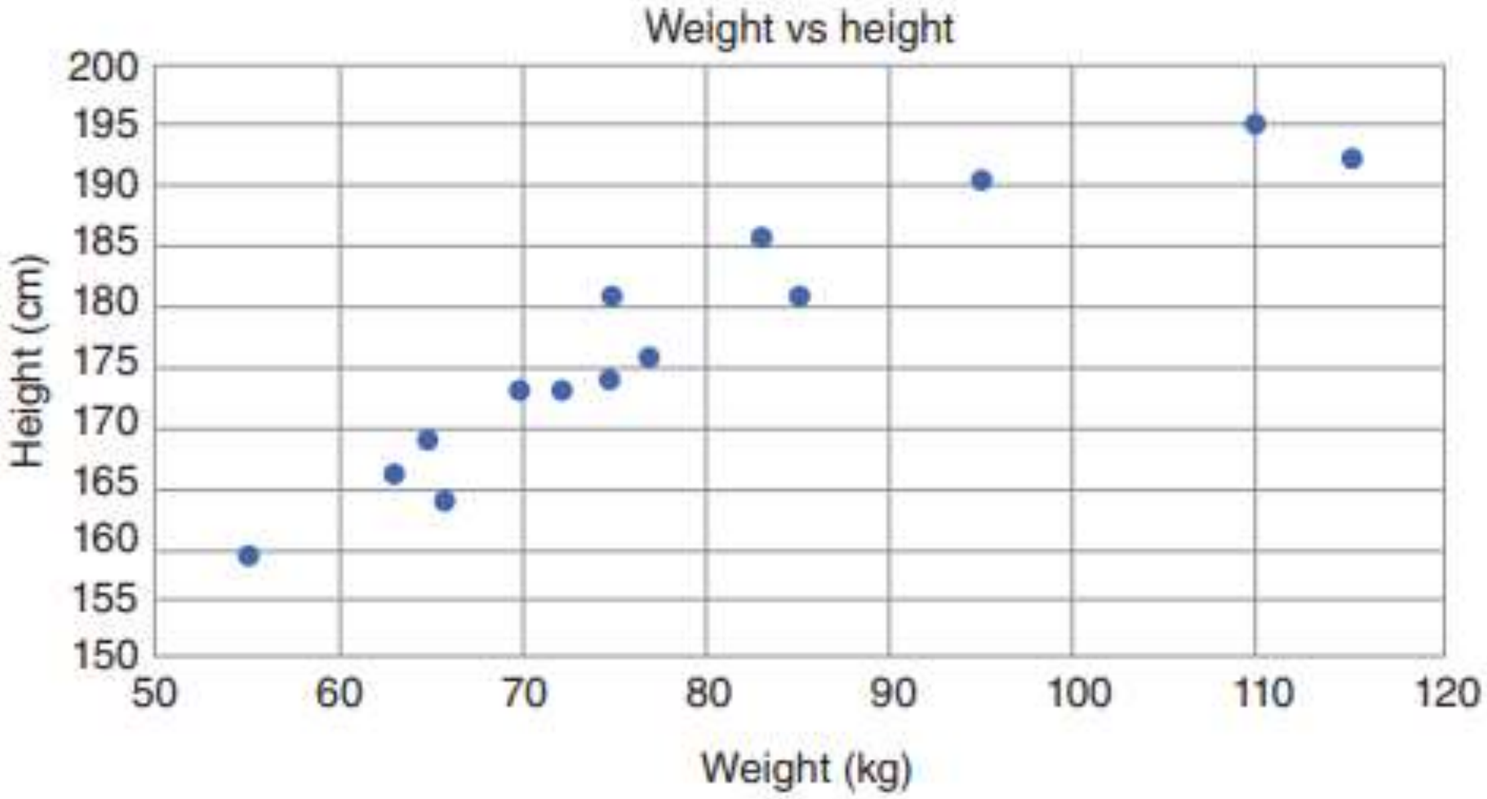


Figure 2.15 Scatter plot for the attributes "weight" and "height".

Two Quantitative Attributes

- how an attribute varies when a second attribute is changed – is measured by the covariance between them.
- When two attributes have a similar variation, the covariance has a positive value. If the two attributes vary in the opposite way, the covariance is negative.
- The value depends on the magnitude of the attributes. If they seem to have independent variation, the covariance value tends to zero. It must be observed that only linear relations are captured.
- The variance can be seen as a special case of covariance: it is the covariance of an attribute with itself.

$$s_{ij} = \text{cov}(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

correlations between two attributes

- A and B: a positive correlation, a negative correlation and a lack of correlation.

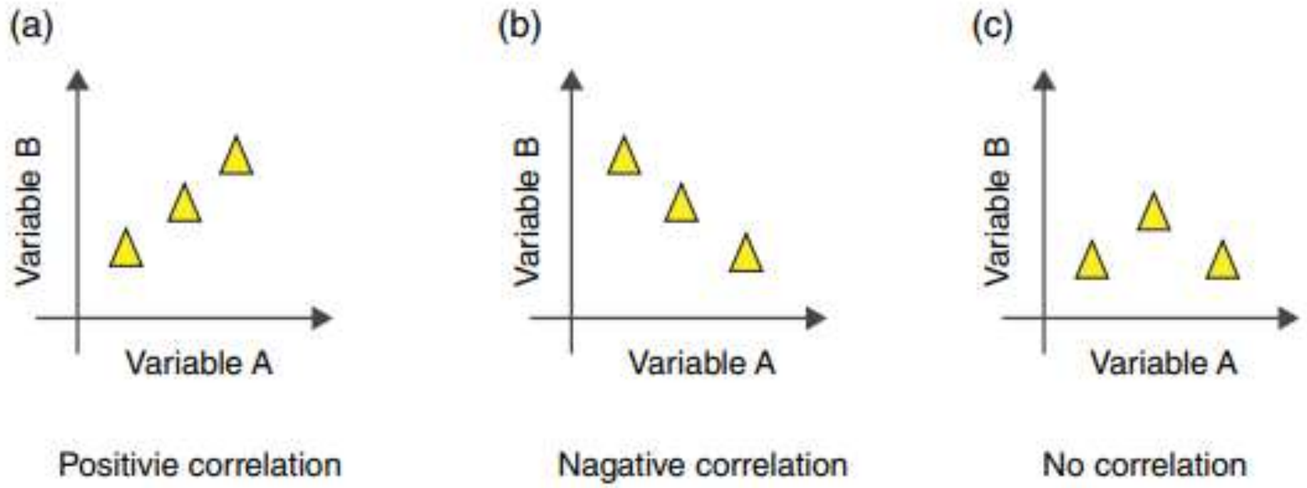


Figure 2.16 Three examples of correlation between two attributes.

correlations between two attributes

- The Pearson correlation evaluates the linear correlation between the attributes.
- If the points are in an increasing line, the Pearson correlation coefficient will have a value of 1.
- If the points are in a decreasing line, its value will be −1.
- A value of 0 is when the points form a horizontal line or a cloud without any increasing or decreasing tendency, meaning the nonexistence of a Pearson correlation between the two attributes.
- Positive values mean the existence of a positive tendency between the two attributes

$$r_{ij} = \text{cor}(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{s_i s_j}$$

correlations between two attributes

Table 2.8 The rank values for the attributes "weight" and "height".

Weight	Height
1.0	1.0
4.0	2.0
2.0	3.0
3.0	4.0
5.0	5.5
6.0	5.5
7.5	7.0
9.0	8.0
7.5	9.5
11.0	9.5
10.0	11.0
12.0	12.0
14.0	13.0
13.0	14.0

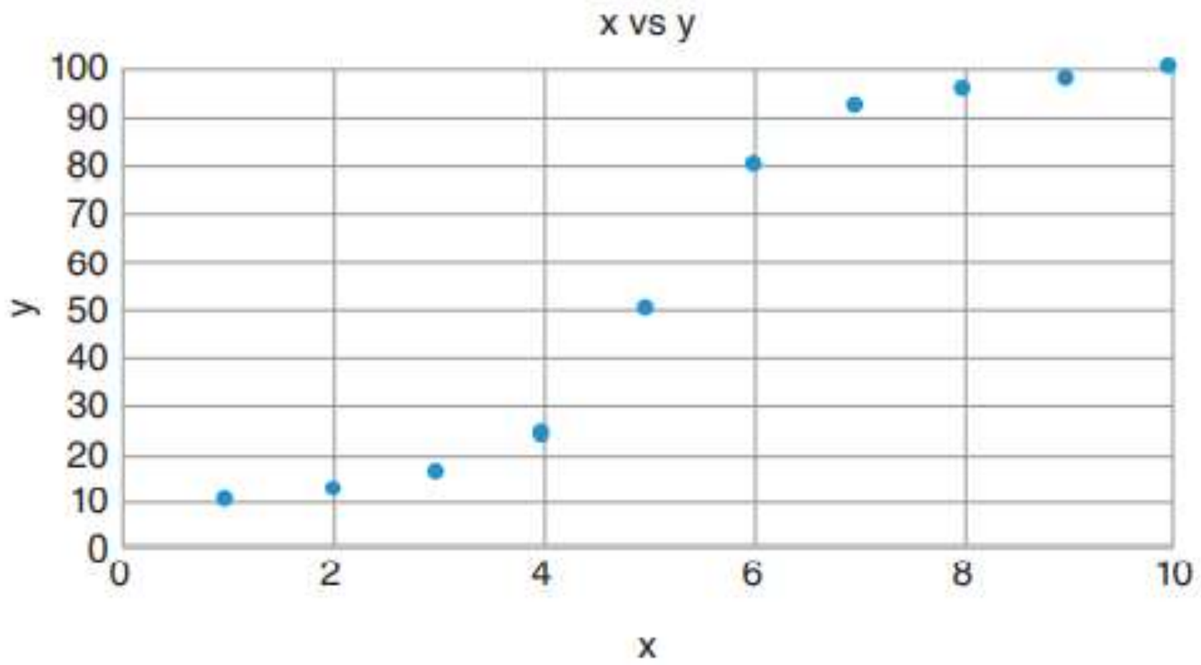


Figure 2.17 The scatter plot for the attributes *x* and *y*.

Two Qualitative Attributes, at Least one of them Nominal

- When the attributes are both qualitative with at least one nominal, contingency tables are used.
- Contingency tables present the joint frequencies, facilitating the identification of interactions between the two attributes.
- They have a matrix-like format, with cells in a square and labels at the left and top.
- On the right most column are the totals per row while in the bottom most row are the totals per column.
- The bottom right-hand corner has the total number of values

Two Qualitative Attributes, at Least one of them Nominal

		Company		
		Good	Bad	
Gender	Male	6	2	8
	Female	1	5	6
		7	7	14

Figure 2.18 Contingency table with absolute joint frequencies for "company" and "gender".

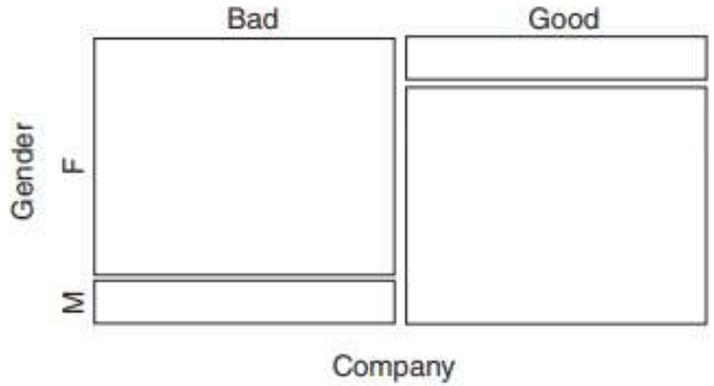


Figure 2.19 Mosaic plot for "company" and "gender".

Two Ordinal Attributes

- ✓ Spearman's rank correlation should be used instead of the Pearson correlation.
- ✓ Scatter plots with ordinal attributes usually have the problem that there are many values falling at the same point, making it impossible to evaluate the number of values per point. In order to avoid this problem, some software packages use a jitter effect which add a random deviation to the values, making it possible to evaluate how large the cloud is.
- ✓ Contingency tables can be used and mosaic plots too. The values should be in increasing order

Assessment 1

What are the most appropriate scales for the following examples?

- university students' exam marks
- level of urgency in the emergency room of a hospital
- classification of the animals in a zoo
- carbon dioxide levels in the atmosphere

References

1. João Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

Thank You