Unit - III

Visualizing Associations

Visualizing proportions

We want to show some group, entity or amounts break down into individual piece that each represent a proportion of the whole.

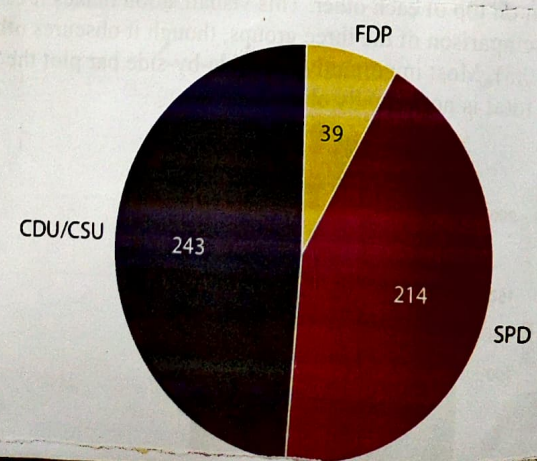
Ex:- proportions of men and women in a group, people voting for different political parties in a election etc.

For different scenarios we will different type of visualizations like. piecharts, side by side bars, stacked bars etc.

1) A case for piecharts

A pie chart breaks a circle into slices such that the area of each slice is proportional to the fraction of the total it represents.

Ex:- party composition of the 8th German Bundestag 1976-1980, visualized as a piechart. this visualization shows clearly that the ruling coalition of SPD and FDP had a small majority over the opposition CDU/CSU.



The same procedure can be performed on a rectangle and the result is a stacked barchart. Depending on whether we slice the bar vertically or horizontally, we obtain vertically stacked bars or horizontally stacked bars.

we slice the bar vertically or horizontally, we obtain vertically stacked bars (Figure 10.2a) or horizontally stacked bars (Figure 10.2b).

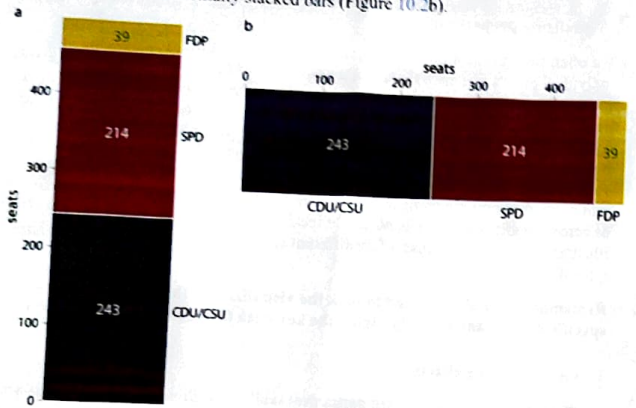


Figure 10.2: Party composition of the 8th German Bundestag, 1976–1980, visualized as stacked bars. (a) Bars stacked vertically. (b) Bars stacked horizontally. It is not immediately obvious that SPD and FDP jointly had more seats than CDU/CSU.

A case for side by side bars

consider a hypothetical scenario of five companies A, B, C, D and E who all have roughly comparable market share of approximately 20%. our hypothetical dataset lists the market share of each company for three consecutive years. when we visualize this dataset with pie charts, it is difficult to see what exactly is going on.

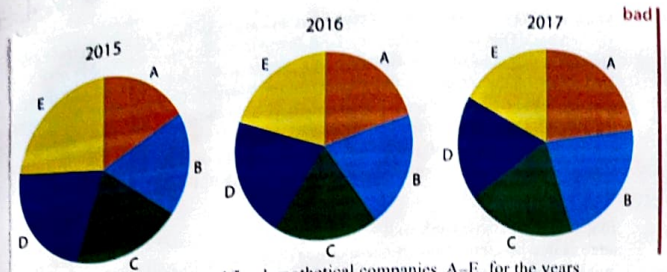


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.

This picture become little clearer when we switch to stacked bars.

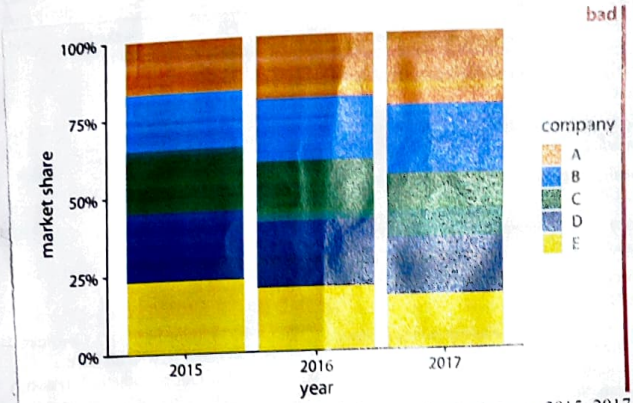


Figure 10.5: Market share of five hypothetical companies for the years 2015–2017, visualized as stacked bars. This visualization has two major problems: 1. A comparison of relative market shares within years is difficult. 2. Changes in market share across years are difficult to see for the middle companies B, C, and D, because the location of the bars changes across years.

For this hypothetical data set side by side bars are the best choice. This visualization highlights that both companies

A and B have increased their market share from 2015 to 2017 while both companies D and E have reduced theirs.

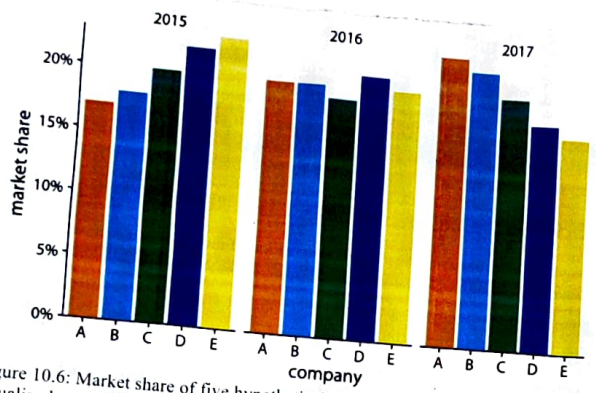
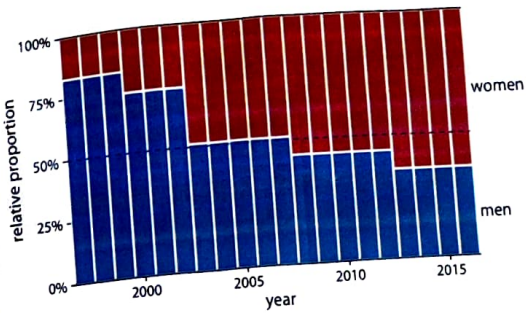


Figure 10.6: Market share of five hypothetical companies for the years 2015–2017, visualized as side-by-side bars.

A case for stacked bars and stacked densities

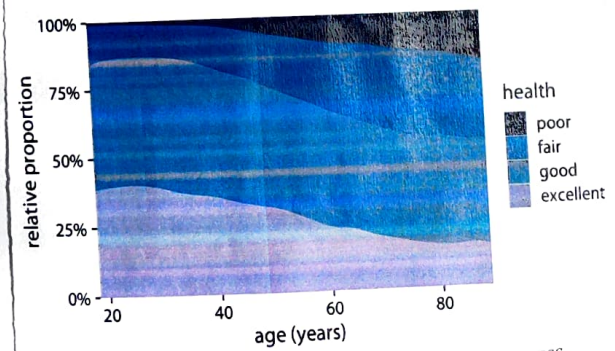
A stacked bars is a form of bar chart that shows the composition of a few variables either relative or absolute, overtime.

To visualize how the proportion of women in the parliament has changed overtime. This figure provides an immediate visual representation of the changing proportions overtime.



If we want to visualize how proportions change in response to a continuous variable, we can switch from stacked bars to stacked densities.

Ex:- consides the health status of people as a function of age. Age can be considered a continuous variable and visualizing the data in this way works reasonably well.



Health status by age, as reported by the general social survey (GSS

Visualizing proportions separately as a part of the total

Side by side bars have the problem that they don't clearly visualize the size of the individual parts relative to the whole and stacked bars have the problem that the different bars cannot be compared easily because they have different baseline.

We can address these two issues by working at a separate plot for each part in each plot showing the respective part relative to the whole.

Ex: Health status by age, shown as proportion of the total number of people in the survey. The colored areas show the density estimates of the ages of people with the respective health status and the gray areas show the overall age distributions.

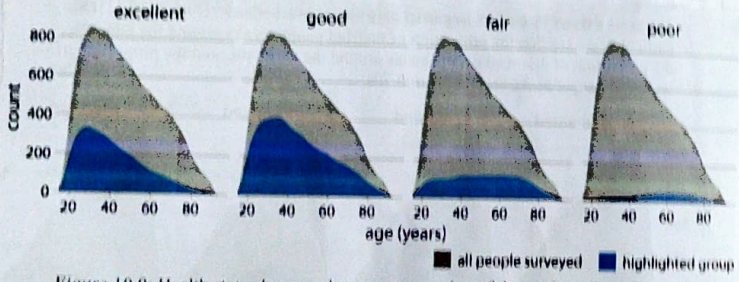
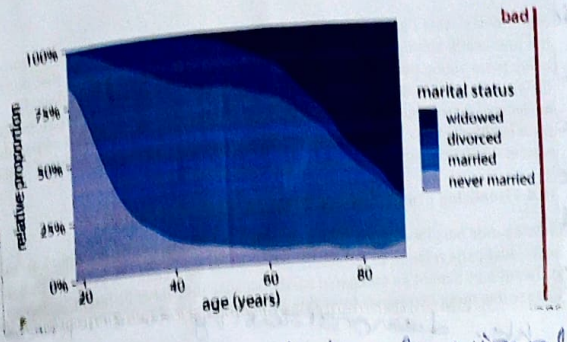


Figure 10.9: Health status by age, shown as proportion of the total number of people in the survey. The colored areas show the density estimates of the ages of people with the respective health status and the gray areas show the overall age distribution.

Each panel displays a different variable from the same survey. Marital status changes much more drastically with age than does health status.



The same dataset visualized as partial densities is much clearer.

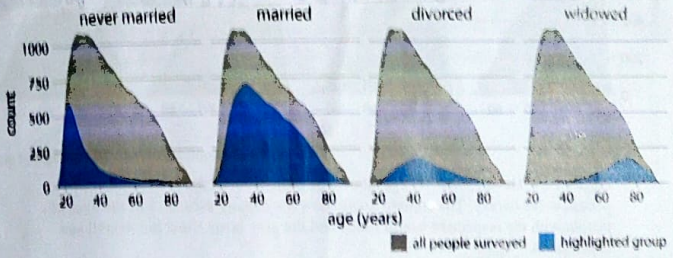


Figure 10.11: Marital status by age, shown as proportion of the total number of people in the survey. The colored areas show the density estimates of the ages of people with the respective marital status, and the gray areas show the overall age distribution.

Mosaic plots and Treemaps.

A mosaic plot is a special type of stacked barchart that shows percentages of data in groups. It is used to show relationships and to provide a visual comparison of groups.

Whenever we have categories that overlap. it is best to show clearly how they relate to each other. This can be done with a mosaic plot.

To draw a mosaic plot we begin by placing one categorical variable along the x-axis (era of construction) and subdivides the x-axis by the relative proportions that make up the categories. We then place the other categorical variable along the y-axis (building material) and within each category along the x-axis, subdivide the y-axis by the relative proportions that make up the categories of the y variable.

The result is a set of rectangles whose areas are proportional to the number of ~~classes~~ cases representing each possible combination of the two categorical variables.

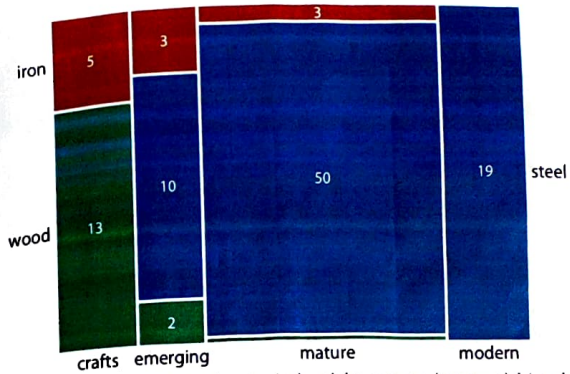


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of

Tree Maps

A Tree map is a visual method for displaying hierarchical data that uses nested rectangles to represent the branches of a tree diagram.

In a treemap we recursively nest rectangles inside each other. In the case of pittsburgh bridges, we can first subdivide the total area into three parts representing the three building materials wood, iron and steel. Then we subdivide each of those areas further to represent the construction eras represented for each building material.

Nested pies

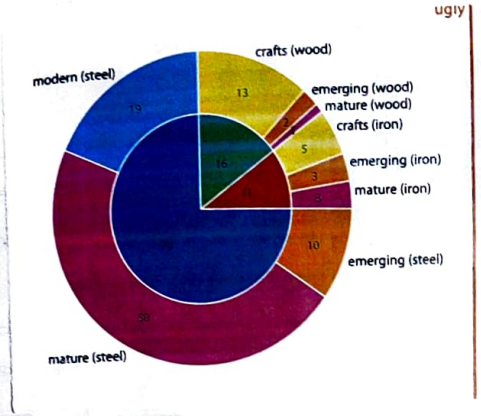
Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine

Nested pies:

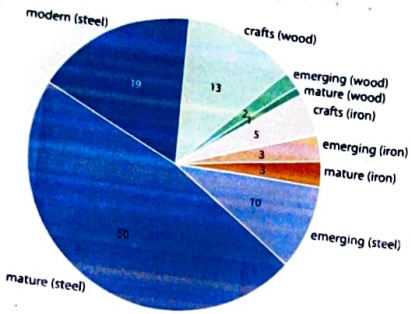
The Nested pie chart is a special type of chart that allows you to show symmetrical and assymmetrical tree structures in a consolidated pie-like structure.

The chart is visualized as a series of concentric circles are arranged like a pie. The circles are divided into segments that represent each of the data values, the ratio of each segment is determined by the corresponding data value.

These are two possibilities, we can draw a piechart composed of a inner and an outer circle. The inner circle shows the breakdown of the data by one variable (building material) and the outer circle shows the breakdown of each slice of the inner circle by the second variable (here, era of bridge construction)



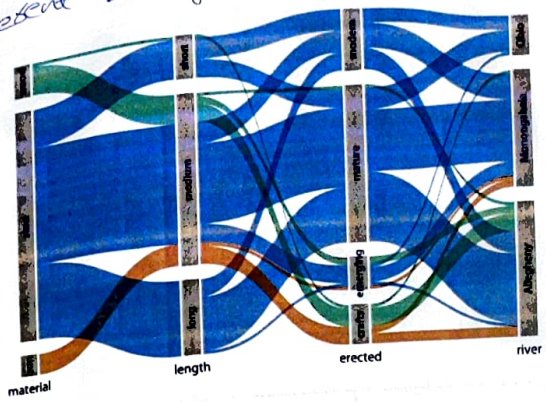
Otherone is we can first slice the pie into pieces representing the proportions according to one variable (material) and then subdivide these slices further according to the other variable (construction era)



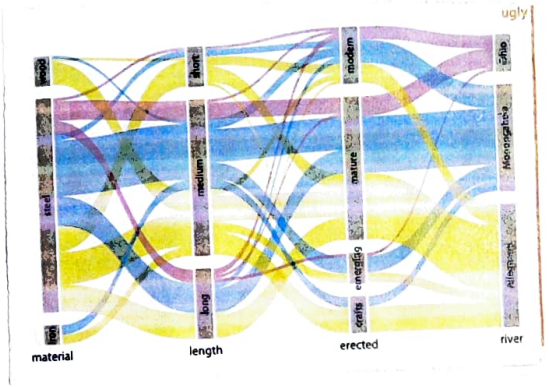
Parallel sets

A parallel sets plot is a representation of various dimensions with relationships. Each dimension has a horizontal line which divides into the numbers of categories of that dimension. The width of the bar denotes the higher value of that category. Each flow path can be colocoded to show and compare the distribution between different categories.

Let us consider the dataset of pittsbourgh bridges. The Breakdown of bridges in pittsburgh by constuction material, length, era of construction and the river they span shown in parallelset plot. The colouring of the bands highlights the construction material of the different bridges.



The same visualization looks quite different if we color by different criterion, for example by rivers.

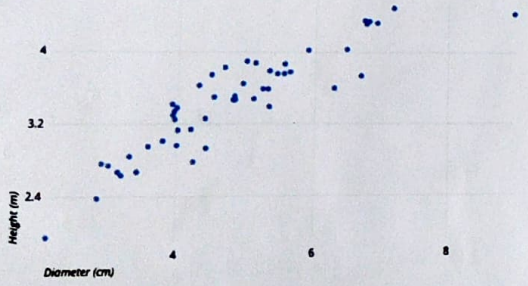


Visualizing associations among two or more quantitative variables

1) Scatter plots:

A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual dot points. they are used to observe relationships between variables.

For example scatter plot below shows the diameter and heights for a sample of fictional trees. Each dot represents a single tree, each points horizontal position indicates that trees height.



Types of correlation

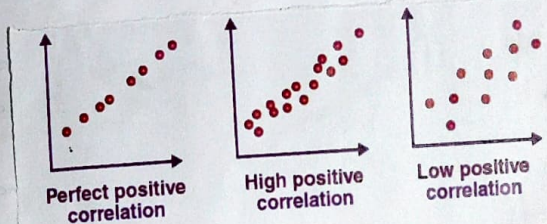
The scatter plots explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There are three types of correlation.

1. positive correlation
2. Negative correlation
3. No correlation.

Positive correlation.

When the points in the graph are rising, moving from left to right then the scatter plot shows a positive correlation

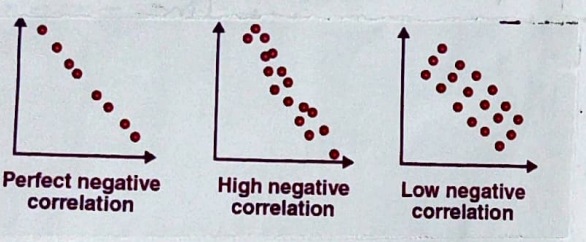
1. perfect positive - which represents a perfectly straight line
2. High positive - All points are nearby
3. Low positive - when all the points are scattered.



Negative correlation

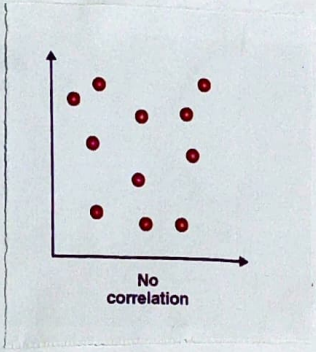
when the points in the scatter graph fall while moving left to right then it is called a negative correlation. It means the value of one variable are decreasing with respect to another.

1. perfect Negative - It is almost a straight line
2. High Negative - when points are near to one another
3. Low Negative - when points are in scattered form.



No correlation:

When the points are scattered all over the graph and it is difficult to conclude whether the values are increasing or decreasing then there is no correlation between the variables.



The blue Jay dataset contains both male and female birds and we may want to know whether the overall relationship between head length and body mass holds up separately for each sex.

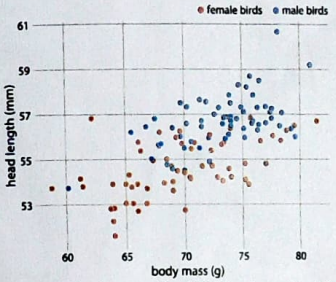


Figure 12.2: Head length versus body mass for 123 blue jays. The birds' sex is indicated by color. At the same body mass, male birds tend to have longer heads (and specifically, longer bills) than female birds.

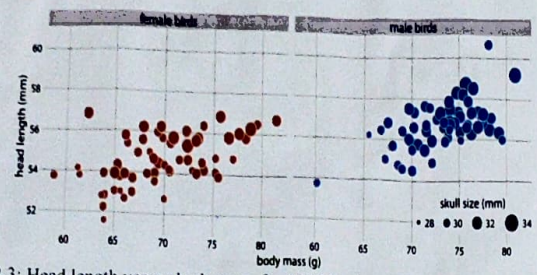


Figure 12.3: Head length versus body mass for 123 blue jays. The birds' sex is indicated by color, and the birds' skull size by symbol size. Head-length measurements include the length of the bill while skull-size measurements do not.

2) correlogram.

A correlogram is a visualization between two or more variables, to show correlation. If on a plot, both the variables move towards their positive side then it is called positive correlation.
→ If both the variables moves towards negative then it is called negative correlation.
→ If no pattern detected between the variables, then it is called no correlation.
→ We describe correlations with a unit-free measure called a correlation coefficient which ranges from -1 to +1 and is denoted by r.
→ Statistical significance is ~~dar~~ indicated with a p-value. A p-value is a measure of probability used for hypothesis testing

→ If r=0, the weakest the linear relationship
→ If r=1, positive correlation
→ If r=-1, Negative correlation.
→ Visualization of correlation coefficients are called correlograms.
→ The sign of the correlation coefficient indicates whether the variables are correlated or anticorrelated.

Let us show randomly generated sets of points that differs widely in the degree to which the x and y values are correlated.

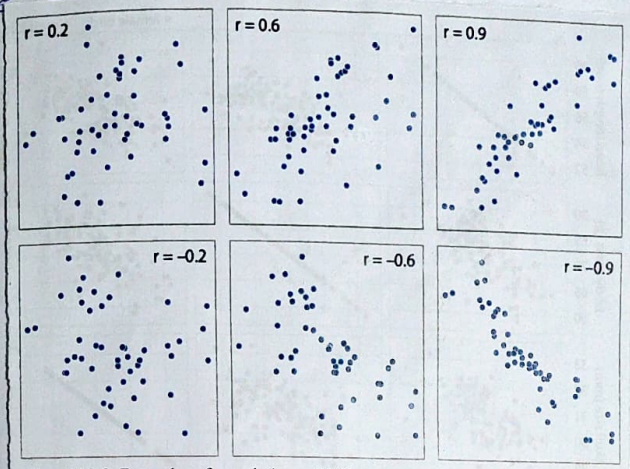


Figure 12.5: Examples of correlations of different magnitude and direction, with associated correlation coefficient *r*. In both rows, from left to right correlations go from weak to strong. In the top row the correlations are positive (larger values for one quantity are associated with larger values for the other) and in the bottom row they are negative (larger values for one quantity are associated with smaller values for the other). In all six panels, the sets of *x* and *y* values are identical, but the pairings between individual *x* and *y* values have been reshuffled to generate the specified correlation coefficients.

The correlation coefficient is defined as

To illustrate the use of a correlogram we will consider a data set of over 200 glass fragments obtained during forensic work.

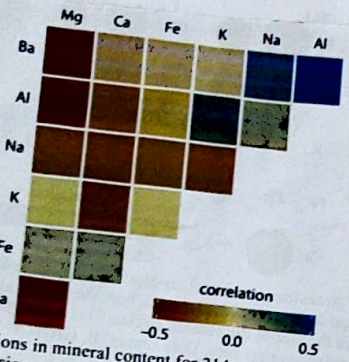


Figure 12.6: Correlations in mineral content for 214 samples of glass fragments obtained during forensic work. The dataset contains seven variables measuring the amounts of magnesium (Mg), calcium (Ca), iron (Fe), potassium (K), sodium (Na), aluminum (Al), and barium (Ba) found in each glass fragment. The colored tiles represents the correlations between pairs of these variables. Data source: B. German

(3) Dimension Reduction.

There are many techniques for dimension reduction. We will take one technique here, the most widely used one called principal component Analysis (PCA)

It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the principal components

It is a technique to draw strong patterns from the given dataset by reducing the variances. PCA generally tries to find the lower dimensional surface to project the high dimensional data.

When we perform PCA, we are generally interested in two pieces of information

1) The composition of the PCs
2) The location of the individual data points in the principal components space

Lets look at these two pieces in a PC analysis of the forensic glass dataset. component one (PC1) measures primarily the amount of aluminum, barium, sodium and magnesium contents in a glass fragment, whereas component two (PC2) measures primarily the amount of calcium and potassium

content and to some extent the amount of aluminum and magnesium.

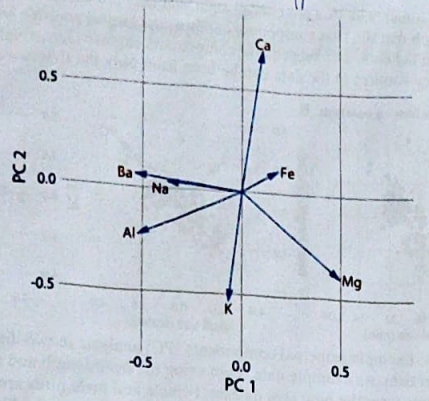


Figure 12.9: Composition of the first two components in a principal components analysis (PCA) of the forensic glass dataset. Component one (PC 1) measures primarily the amount of aluminum, barium, sodium, and magnesium contents in a glass fragment, whereas component two (PC 2) measures primarily the amount of calcium and potassium content, and to some extent the amount of aluminum and magnesium.

4) Paired data

A special case of multivariate quantitative data is paired data. Data where there are two or more measurements of the same quantity under slightly different conditions.

For paired data, it is reasonable to assume that the two measurements belonging to a pair are more similar to each other than to the measurements belonging to other pairs. We need to choose visualizations that highlight any differences between the paired measurements.

Ex:- In a slopegraph (may be a better choice) we draw individual measurements as dots arranged into two columns and indicate pairings by connecting the paired dots with a line. The slope of each line highlights the magnitude and direction of the change.

By using this approach to show the ten countries with the largest difference in CO_2 emission per person from 2000 to 2010 is shown below.

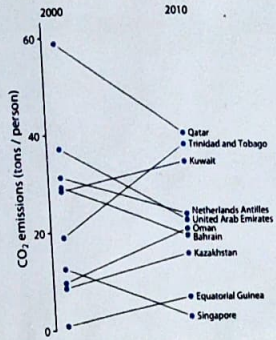


Figure 12.12: Carbon dioxide (CO2) emissions per person in 2000 and 2010, for the ten countries with the largest difference between these two years. Data source: Carbon Dioxide Information Analysis Center

Slopegraphs have one important advantage that it we can be used to compare more than two measurements at a time. For example CO_2 emissions per person in 2000, 2005 and 2010 for the ten countries with the largest difference between the years 2000 and 2010

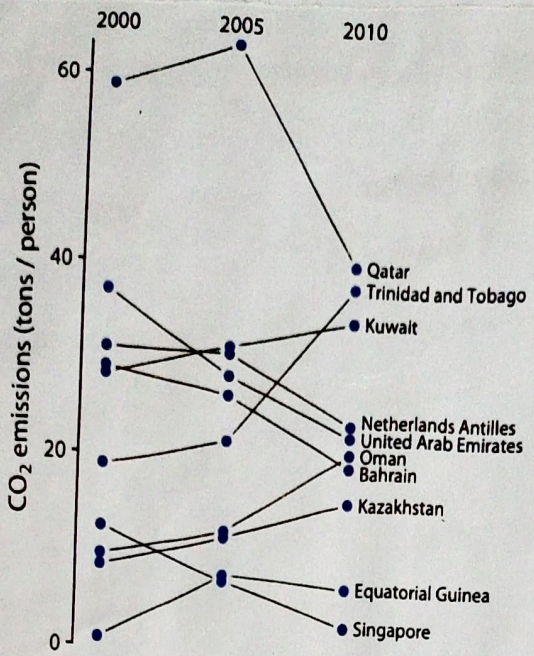


Figure 12.13: CO2 emissions per person in 2000, 2005, and 2010, for the ten countries with the largest difference between the years 2000 and 2010. Data source Carbon Dioxide Information Analysis Center

Assignment

1) Explain the casestudy of side by side bars and piecharts
2) Explain the nested proportions, tree maps and mosaic plots.
3) Explain Nested pies and parallel sets.
4) Explain visualizing associations among two or more quantitative variables.