Data Visualisation

Chapter 1

Data Types and Representations

Dr. Tony Lyons

Autumn Semester 2019

B Sc (Hons) Multimedia Applications Development

B Sc (Hons) Creative Computing

Waterford Institute of Technology

Contents

1 Introduction to Data Types

In statistical analysis we encounter four data-types which are :

Nominal

Nominal data-types are usually non-numeric, although not necessarily. Examples of nominal data we often encounter include Name, Gender, Nationality, Occupation and so on.

Ordinal

Ordinal data-types are numeric, but they do not provide a scale by which to measure individual data points. An example of an ordinal data type is age brackets, for instance we may categorise a group of people according to the age brackets $0-18$, $18-34$, $35-64$, $65+$. This allows us to order the people in a population according to age bracket, however it does not tell us the age difference between people from different brackets.

Interval

Interval variables are also numerical variables whose **differences** can be compared, however it makes no sense to compare the scales of each data point. A good example of interval data is temperature. For instance if Dublin is 15°C and Madrid is 30°C, it makes sense to say Madrid is 15°C hotter than Dublin. However it does not make sense to say Madrid is twice as hot as Dublin.

Ratio

Like interval variables, ratio variables can be compared and it is meaningful to speak of the difference between various values of these variables. In contrast to interval variables, ratio variables have an absolute value in the scale, while it also makes sense to compare the individual values of these variables. As an example, the temperature, measured in Kelvin, of a filament subject to various currents has an absolute scale. The Kelvin scale possesses an absolute zero point, while it also makes since to say the filament at temperature 450 K is 1.5 times the temperature of a filament at 300 K.

1.1 Statistical Analysis

The aim of this course is to apply various simple statistical tools to these data sets and to represent these data using appropriate graphical representations of the data. The purpose of these representations and statistical measures is to obtain or illustrate various relationships between variables in this set. The variables in a data set are related if there is a clear and consistent pattern of behaviour among the data. As an example, the height versus weight of individuals in a sample are related, since taller people tend to be heavier than shorter people.

Statistical analysis can be broken into two broad categories **descriptive** and **inferential** analysis.

Descriptive Analysis

In descriptive statistics we concentrate on representing the data in a given data-set. For this purpose, some tried and tested tools include

- Histograms
- Box plots
- Pareto
- Pie-chart

Inferential Analysis

In inferential statistics we use mathematical analysis of a data-set to deduce certain properties of the population from the sample data. Such data analyses include

- Significance testing
- Multiple regression
- Cluster analyses
- $t-$tests

Data visualisation is concerned primarily with descriptive methods. The goal of every data visualisation it to convey the information in a given data set in a clear and efficient manner, which allows the reader to quickly and easily deduce key observations from this data, without having to employ any mathematical tools to make these observations.

1.2 Frequency and Relative Frequency Tables

Before we implement any visualisation on an qualitative or quantitative set of data, in most cases we must first divide the data into various categories to determine the ***frequency*** and ***relative frequency*** of each data class. These terms are defined as follows:

Definition 1.1: Frequency of a Data Class

Given a data set with N data points, which is to be divided into q data classes

$$Class_1,\ Class_2,\ \dots,\ Class_q,$$

then the frequency f_m of $Class_m$ is the number of data points in this class.

Remark 1.1

Since every data point must belong to exactly one data class, it follows that adding all the frequencies should return the total number of data points, that is to say

$$f_1 + f_2 + \dots + f_q = N$$

Example 1.2.1

The date of birth of each student in a college class of 50 students is collected. The students are categorised according to the month they were born in, with the following frequencies

January	**February**	**March**	**April**
5	3	4	2
May	**June**	**July**	**August**
4	7	3	2
September	**October**	**November**	**December**
4	6	3	7

The example given here is an example of a frequency table, with the categories corresponding to the months of the year, and the frequency of each category being the number of students being born in that month. In this case we see that the sum of the frequencies satisfies

$$5 + 3 + 4 + 2 + 4 + 7 + 3 + 2 + 4 + 6 + 3 + 7 = 50, \tag{1.1}$$

that is to say, the sum of the frequencies matches the number of data points.

A table which is also useful in descriptive statistics is the relative frequency table. The relative frequency of a data class is defined as follows

Definition 1.2: Relative Frequency of a Data Class

Given a data set with N data points, which is to be divided into q data classes

$$\text{Class}_1,\ \text{Class}_2,\ \ldots,\ \text{Class}_q,$$

then the relative frequency of Class_m, which we denote by r_m, is the frequency of this class, f_q, divided by the number of data points N, that is to say,

$$r_m = \frac{f_m}{N}.$$

Remark 1.2

Since the sum of frequencies in a data set always equals the number of data points, it follows that the sum of relative frequencies in a categorised data set should always equal 1, that is to say

$$\begin{aligned} r_1 + r_2 + \ldots + r_q &= \frac{f_1}{N} + \frac{f_2}{N} + \ldots + \frac{f_q}{N} \\ &= \frac{f_1 + f_2 + \ldots + f_q}{N} \\ &= \frac{N}{N} \\ &= 1 \end{aligned} \tag{1.2}$$

Using Example 1.2.1, we can also write the relative frequency by dividing each frequency by 50, the total number of data points, to give

Example 1.2.2

The date of birth of each student in a college class of 50 students is collected. The students are categorised according to the month they were born in, with the following relative frequencies

January	**February**	**March**	**April**
0.1	0.06	0.08	0.04
May	**June**	**July**	**August**
0.08	0.14	0.06	0.04
September	**October**	**November**	**December**
0.08	0.12	0.06	0.14

We also note that the sum of the relative frequencies adds to 1, that is

$$\begin{aligned} &0.1 + 0.06 + 0.08 + 0.04 + 0.08 + 0.14 \\ &\qquad + 0.06 + 0.04 + 0.08 + 0.12 + 0.06 + 0.14 = 1.00 \end{aligned} \tag{1.3}$$

2 Descriptive Statistics - Data Plotting

2.1 Bar charts

On of the simplest graphical representations of statistical data is a **bar chart**. Bar charts are useful when we want to represent categories of data, where the categories are represented along the x-axis of the chart, and the number of data points that appear in that particular category. As an example, if we go the the WIT car park and record the make of each car present, then the make of each car would be the categories, while the number of cars made by a particular car manufacturer would be represented by the height of each bar. Generally, we can make the bars any width we choose, however, to convey the data clearly it useful to make each bar of equal width and only vary the height.

Example 2.1.1: Car makes in a car-park

The cars in a car-park were counted according the car-make, and the following data were collected:

Audi	4	**Ford**	21	**Mercedes**	1
BMW	2	**Hyundai**	18	**Opel**	14
Citroen	12	**Kia**	11	**Peugeot**	10

Use a bar-chart to represent this data. Can you see a way that the information in this bar chart might be conveyed to the observer more effectively? Hint: The categories do not necessarily have to be arranged alphabetically.

Solution. In this example, the independent variable is the car make (i.e. the categories) while the dependent variable is the number of each make is the dependent variable. In most graphs, we place the independent variable on the x-axis and the dependent variable on the $y-$axis. A number of features of this graph we wish to point out in relation to the visualisation of the data:

- The horizontal labels are oriented make them more legible
- The step-size of the vertical label is chosen so as to make the axis uncluttered without loosing information in the plot. If we chose to label

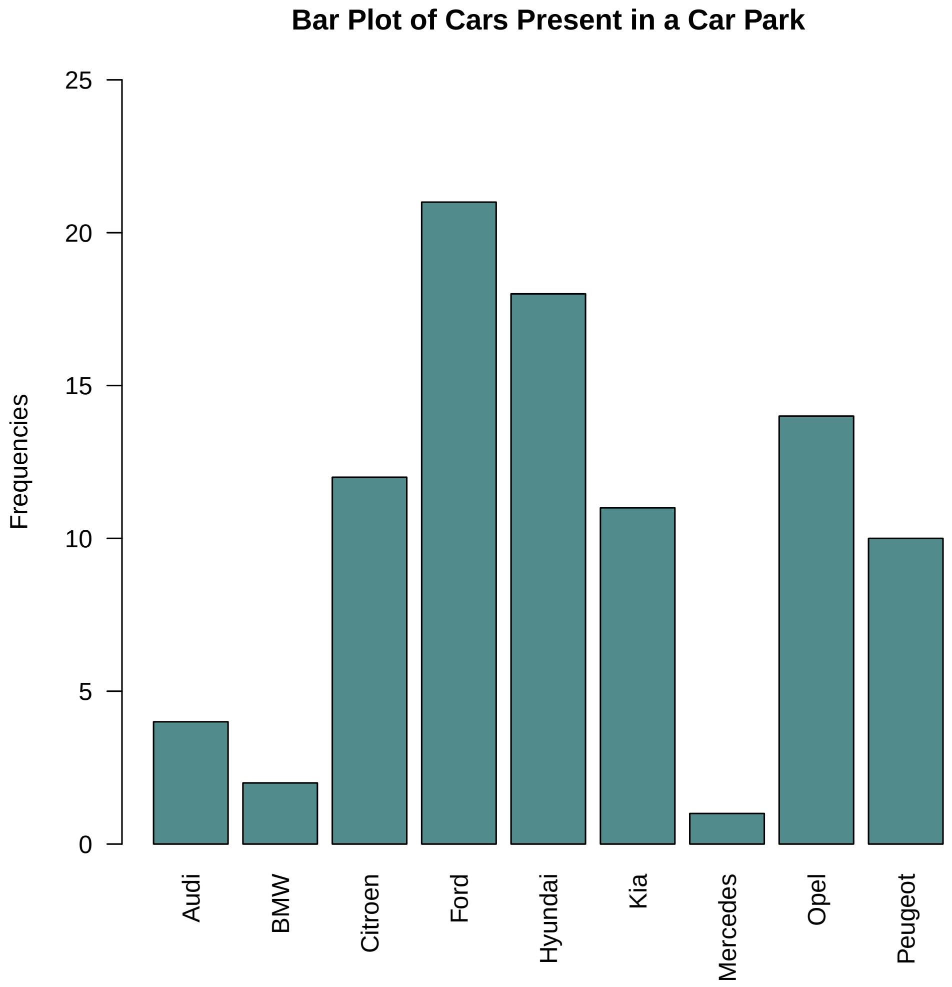

Figure 1: A bar chart for data representing the number of cars be make in a car-park.

the axis 0,1, 2, 3,...,22, the axis would be cluttered and visually unappealing. If on the other hand we chose to label the axis 0,5,10,15,25, then it would not be immediately clear to the reader how many cars of each make are present.

- When preparing a box-plot, it is crucial that you take time to consider the best approach to labeling your axes

- Each axis also tells us what the labels represent, i.e. car make on the horizontal and number of cars present on the vertical axis
- Each bar is of equal width and equal spacing and all have the same colouring. While strictly speaking it is not necessary that this is the case, making the bars of varying width or colouring would add no new information to the graphic and would ultimately make the representation more confusing

Since the categorical labels have no strict order, that is to say, we are not forced to impose an alphabetical ordering on the data, then it is we may arrange the box-plot in order of increasing values, which make the data available in the set much more obvious to the reader. To see this we re-plot the bar chart to give the following representation of the same data:

2.2 Histograms

Given a collection of **ordinal data** a histogram is often an effective way of representing this data. As an example, we could collect data on the test scores in a class exam, and arrange the test scores in a series of ranges. One of the central questions when drawing a histogram is how large one should make the intervals into which the data set is categorised. While there is no definitive rule to set this number of intervals, there are some guidelines to help improve the appearance and accuracy of a histogram:

(1) The number of bars in the bar chart should be between 5 and 20

(2) If the maximum (x_{max}) and minimum (x_{min}) values of the data set are not whole numbers we proceed as follows

- Round the minimum value to next integer below this value $x_{min} \to n_{min}$
- Round the maximum value to the next integer above the value $x_{max} \to n_{max}$

(3) Find the range of this interval, i.e. $R = x_{max} - x_{min}$ or else $R = n_{max} - n_{min}$

(4) If possible, choose a number of bars which divides evenly into the range R

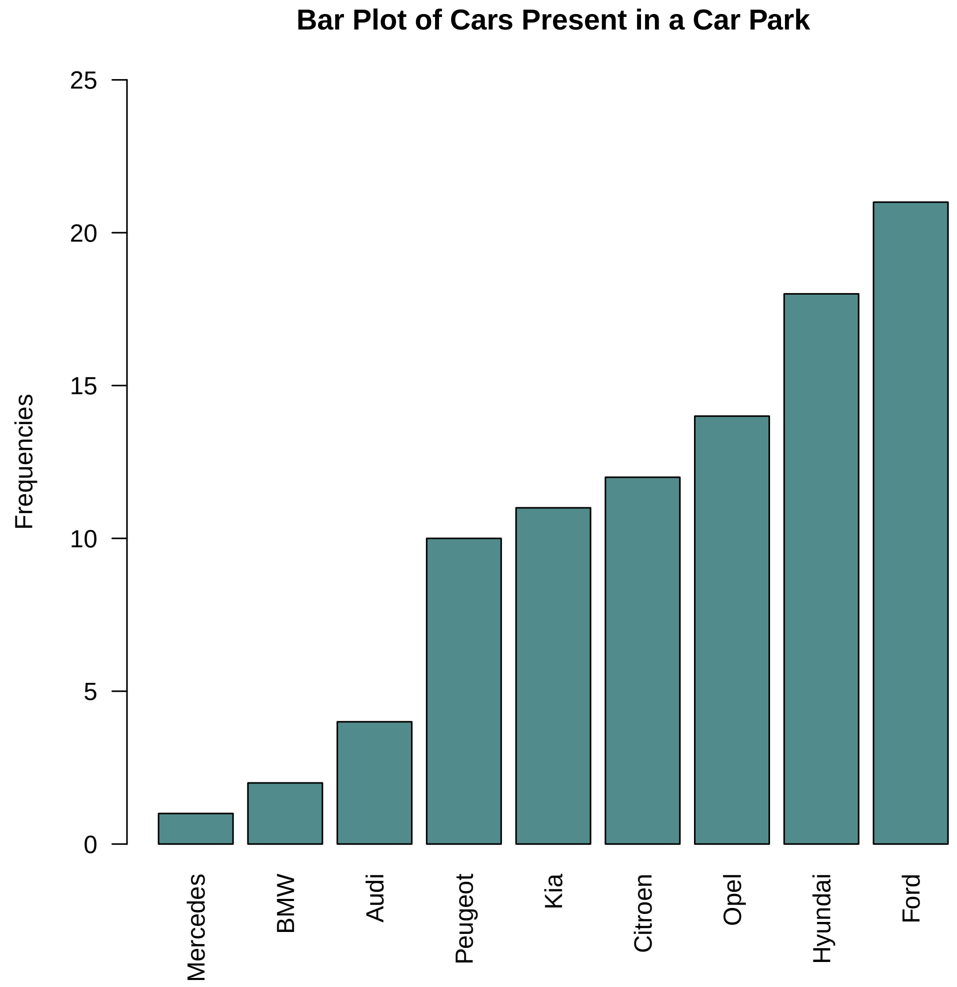

Figure 2: A box-plot for data representing the number of cars be make in a car park.

(5) Divide the range R by this number to give the bar width I

(6) Starting at x_{min} or n_{min}, mark of segments of length I along the horizontal axis until reaching x_{max} or n_{max}

(7) If a data value falls into two different categories, it is normal to place it in the upper category

Example 2.2.1

The test scores of a class of 32 students are given by the following data

4	17	28	32	33	35	38	41
42	42	48	50	52	52	56	59
61	68	68	69	70	71	74	75
79	80	80	84	89	90	91	94

Using this data answer the following

(i) Identify the data type given

(ii) Find an appropriate range and intervals to catergorise this data

(iii) Construct the resulting frequency table for this data

(iv) Draw a histogram to represent this data

Solution. In this case the minimum value is $x_{min} = 4$ and the maximum value is $x_{max} = 94$. These are both whole numbers so we do not need to round off. However, for the purposes of our histogram, it may be useful to start our range at $x_min = 0$ and finish the range at $x_{max} = 100$ giving a range $R = 100$. A good choice for the number of bars would be 10, giving a bar width of $100/10 = 10$, giving us the following frequency table:

0-10 %	**10-20 %**	**20-30 %**	**30-40 %**	**40-50 %**
1	1	1	4	4
50-60 %	**60-70 %**	**70-80 %**	**80-90 %**	**90-100 %**
5	4	5	4	3.

With this categorisation of the scores we obtain the following histogram

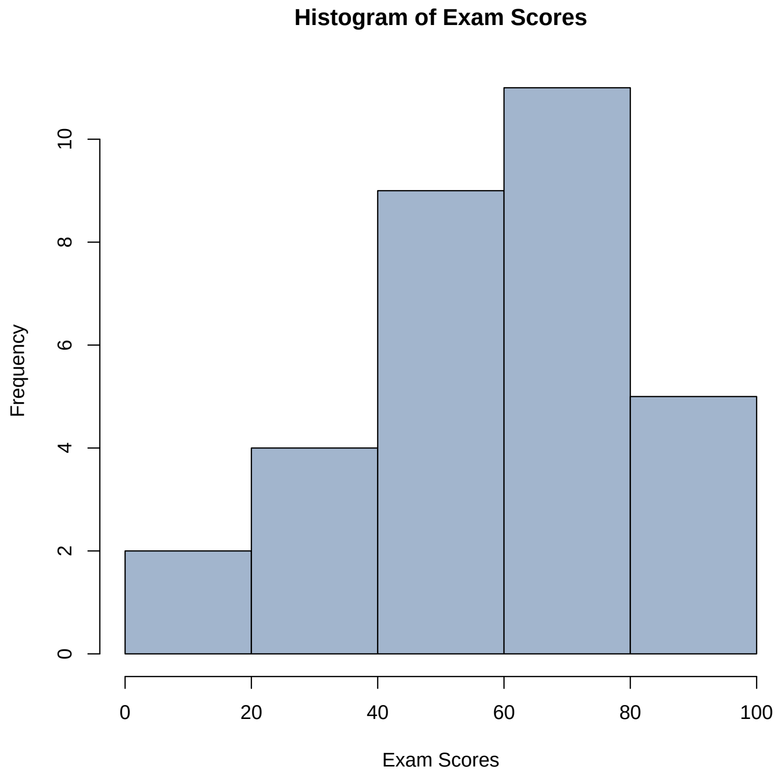

Figure 3: A histogram depicting the exam socres obtained by a class of 32 students.

Example 2.2.2: Example form lectures

The height of each tree in an orchard was measured and found to be

3.5m 4.4m 4.7m 5.1m 5.2m 5.3m 5.9m 6.0m 6.4m 8.7m

Using this data answer the following

(i) Identify the data type given

(ii) Find an appropriate range and intervals to catergorise this data

(iii) Construct the resulting frequency table for this data

(iv) Draw a histogram to represent this data

Solution. The data type is ratio, since there is an absolute zero, i.e. a tree height cannot be less than 0m, and the data points can be meaningfully compared.

Since $x_{min} = 3.5$ we move to a new minimum value $x_{min} \to 3.0$, while the maximum data value is $x_{max} = 8.7$m which changes to $x_{max} \to 9.0$m. This makes the new data range $R = 9.0 - 3.0 = 6.0$m. Next we should choose a bar number between 5 and 20, which also divides into this range if possible. An obvious choice for this example would be 6 bars. This means the width of each interval should be $\frac{6.0}{6} = 1.0$m.

This results in the following frequency table for the data:

3.0-4.0 m	**4.0-5.0 m**	**5.0-6.0 m**	**6.0-7.0 m**	**7.0-8.0 m**	**8.0-9.0 m**
1	2	4	2	0	1

The histogram is given by the following

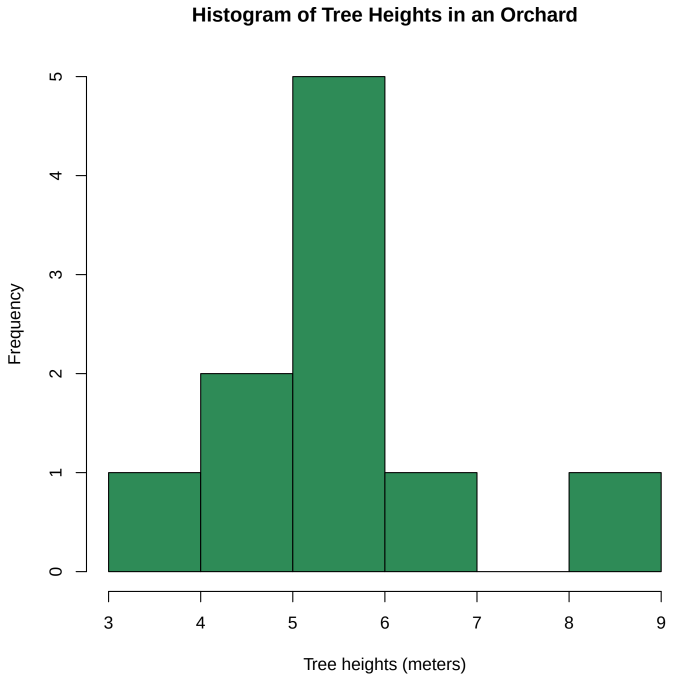

Figure 4: A histogram of the heights of trees in orchard.

2.3 Pareto Charts

2.3.1 Cumulative Relative Frequency Tables

Suppose we are given a nominal data set so that the first class $Class_1$ has relative frequency f_1, the next category $Class_2$ has relative frequency f_2 and so on, up to class $Class_N$ with frequency f_N. Since the categories are nominal, we can arrange them any way we choose, and so we arrange them from highest to lowest frequency

$$f_1 > f_2 > f_3 > \ldots > f_N.$$

The cumulative relative frequency table of the data set is given by

Class	**Cumulative Relative Frequency**
$Class_1$	f_1
$Class_2$	$f_1 + f_2$
$Class_3$	$f_1 + f_2 + f_3$
$\vdots$	$\vdots$
$Class_N$	$f_1 + f_2 + \ldots + f_N = 1$

2.3.2 Pareto Charts

A Pareto chart is a combination of a cumulative frequency chart and a bar chart, where the bars are arranged in decreasing order from left to right.

Example 2.3.1

A survey of 120 employees in a company asked the main reasons for late arrivals at work. The reasons given were as follows

Reason	**Frequency**
Child Care	22
Emergency	8
Overslept	12
Public Transport	15
Traffic	36
Weather	27

Construct a cumulative relative frequency table and a Pareto chart for this data.

Solution. To begin, we organise the data from highest frequency to lowest frequency, and divide each frequency by the total number of respondents (120) to find the relative frequency table:

Reason	**Frequency**	**Relative Frequency**
Traffic	36	0.3
Weather	27	0.225
Child Care	22	0.183
Public Transport	15	0.125
Overslept	12	0.1
Emergency	8	0.067

Using this we construct the cumulative relative frequency table. The cumulative relative frequency of a particular class is given by the relative frequency of that class plus the sum of relative frequencies of all the previous classes:

Reason	**Cumulative Relative Frequency**
Traffic	0.3
Weather	0.525
Child Care	0.708
Public Transport 15	0.833
Overslept	0.933
Emergency	1.0

In ***any*** cumulative relative frequency table, the final class will always have a cumulative relative frequency of 1.0, as is the case in this table. The left hand axis of each Pareto chart represents the frequency of each class, the right hand side represents the cumulative frequency of each class. The cumulative frequency of any class is plotted above the center or the right hand edge of each bar, as in Figure 5.

Remark 2.1

Pareto charts are extremely useful when trying to eliminate deficiencies or diagnose faults in various contexts. For instance, in this example it is easily seen that to eliminate 60% or more late arrivals, to company should address the effects of traffic, weather and child care.

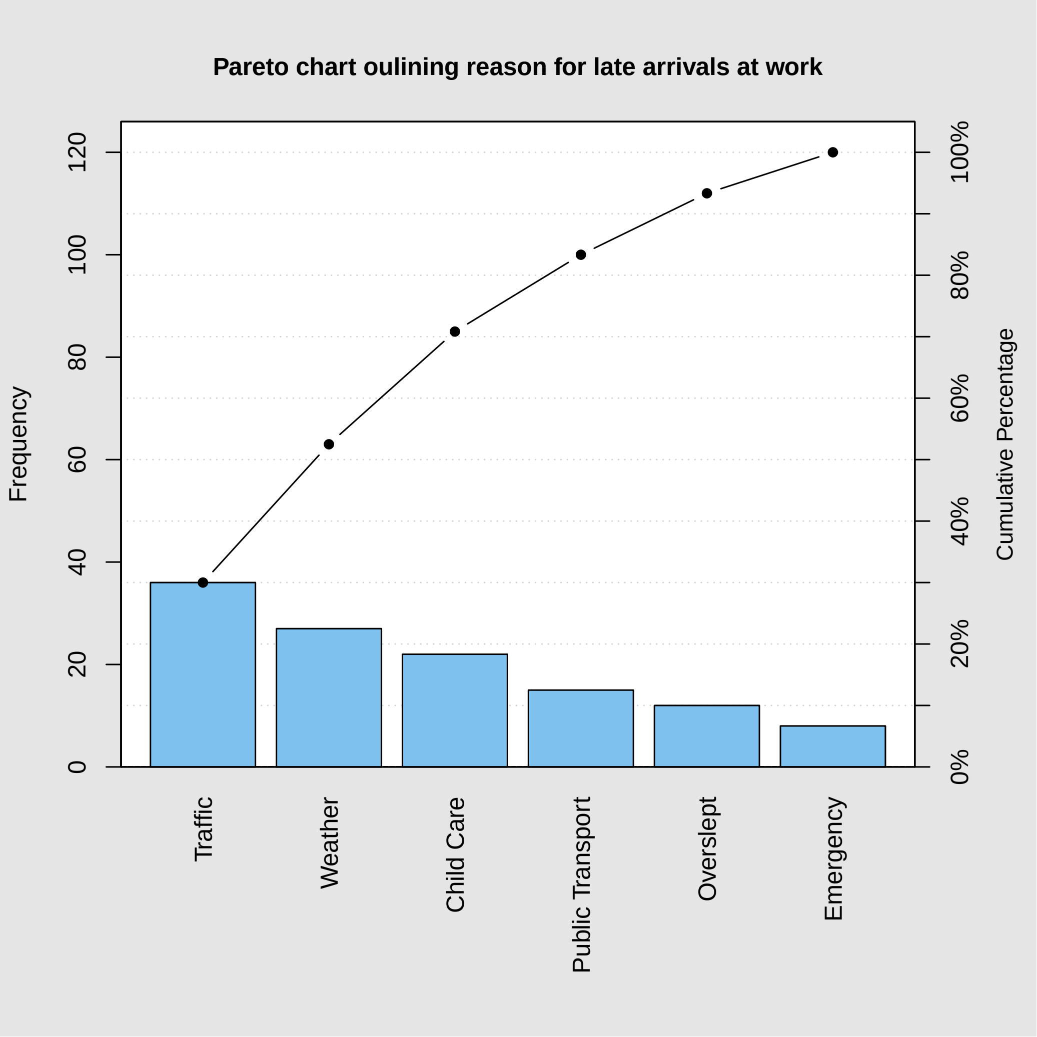

Figure 5: A Pareto chart for the results of a survey relating to late arrivals at work