# FIT3152 Data analytics. Tutorial 02: Visualizing data

1. Try and reproduce the graphics from Lecture 2. Note – the 'iris' data set comes as part of the base R installation. To reproduce the lattice plots, you will need to load lattice. To reproduce the ggplot2 graphics you will need to install and load ggplot2 – this will then give you access to the 'diamonds' data which are required for question 2. Commands are below:

```
library(lattice)
install.packages("ggplot2")
library(ggplot2)
# note help site for ggplot2 is http://docs.ggplot2.org/current/
```

2. The 'diamonds' data set comes packaged with ggplot2 and contains data about the price of diamonds as well as information on size as well as the 4 Cs affecting diamond price: carat (size), cut, colour and clarity. The diagram below, copied from Wickham, *Ggplot2: Elegant graphics for data analysis*, gives you the details.

| carat | cut | color | clarity | depth | table | price | x | y | z |
|---|---|---|---|---|---|---|---|---|---|
| 0.2 | Ideal | E | SI2 | 61.5 | 55.0 | 326 | 3.95 | 3.98 | 2.43 |
| 0.2 | Premium | E | SI1 | 59.8 | 61.0 | 326 | 3.89 | 3.84 | 2.31 |
| 0.2 | Good | E | VS1 | 56.9 | 65.0 | 327 | 4.05 | 4.07 | 2.31 |
| 0.3 | Premium | I | VS2 | 62.4 | 58.0 | 334 | 4.20 | 4.23 | 2.63 |
| 0.3 | Good | J | SI2 | 63.3 | 58.0 | 335 | 4.34 | 4.35 | 2.75 |
| 0.2 | Very Good | J | VVS2 | 62.8 | 57.0 | 336 | 3.94 | 3.96 | 2.48 |

Table 2.1: `diamonds` dataset. The variables depth, table, x, y and z refer to the dimensions of the diamond as shown in Figure 2.1

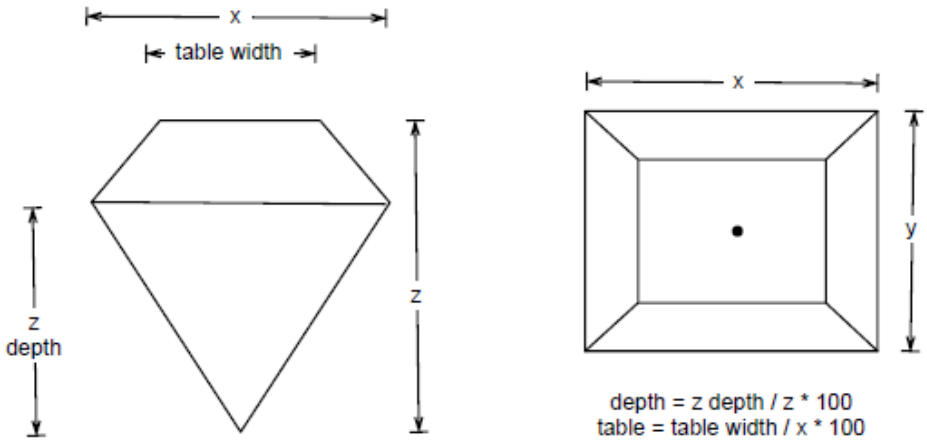


Fig. 2.1: How the variables x, y, z, table and depth are measured.

(a) Taking a random sample using the code below, create a subset of the diamonds data set: 'dsmall' to use in the following analysis.

```
set.seed(9999) # Random seed to make subset reproducible
dsmall <- diamonds[sample(nrow(diamonds), 1000), ] # sample of 1000 rows
```

(b) Using the data 'dsmall' investigate the factors affecting diamond price. Using a variety of graphs and/or tables, show systematically the effect of the 4 Cs on diamond price. Which single variable has the greatest effect on price? Which has the least? *Use ggplot2 for your graphics.*

3 The file "body.dat.csv" contains data from a study on the relationship between body dimensions. The study measured 500+ active individuals. A legend to the data is below.

| Column | Measuring (cm unless stated) |
|---|---|
| ShoulderWidth | Biacromial diameter |
| Pelvis | Pelvic Breadth |
| Hips | Bitrochanteric diameter |
| ChestDepth | Chest depth at nipple level, full expiration |
| ChestDiam | Chest diameter at nipple level, mid-expiration |
| ElbowDiam | Elbow diameter, sum of two elbows |
| WristDiam | Wrist diameter, sum of two wrists |
| KneeDiam | Knee diameter, sum of two knees |
| AnkleDiam | Ankle diameter, sum of two ankles |
| ShoulderGirth | Shoulder girth over deltoid muscles |
| Chest | Chest girth |
| Waist | Waist girth, narrowest part of torso below the rib cage |
| Abdomen | Navel (or "Abdominal") girth |
| HipGirth | Hip girth at level of bitrochanteric diameter |
| ThighGirth | Thigh girth below gluteal fold |
| Bicep | Bicep girth, flexed |
| Forearm | Forearm girth, extended, palm up |
| KneeGirth | Knee girth over patella, slightly flexed position |
| CalfGirth | Calf maximum girth |
| AnkleGirth | Ankle minimum girth |
| WristGirth | Wrist minimum girth |
| Age | Age (years) |
| Weight | Weight (kg) |
| Height | Height (cm) |
| Gender | Male, Female |

The data was obtained from http://www.amstat.org/publications/jse/jse_data_archive.htm
A related article is http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html

Using the data, investigate the following:

(a) Which variables are the best predictors of height? Does this vary between men and women? For examples, are some variables better at predicting height in one gender over the other?

(b) Using the same approach, which variables are best for predicting weight in each gender?

(c) Which variables are most highly correlated? Are the same variables most highly correlated for men and women?

(d) Which measure is the best means of distinguishing between men and women? Show your results and analysis graphically.

4 The data file "Dunhumby1-20.csv" is a cut down and modified set of test data from the Kaggle competition to predict when consumers would next visit a Dunnhumby supermarket and how much they would spend. See: http://www.kaggle.com/c/dunnhumbychallenge for more information. The current modified data set contains the customer ID, Date of visit, Date since last visit, and Spend for 20 customers from the test set.

Tell me as much as you can about those customers using descriptive statistics. Using one or more graphics – such as histograms, boxplots, scatterplots or anything else you can think of make a visual display to show the differences and similarities between the customers. Are there particular customers whose next visit, and spend, would be easier or harder to predict than the cohort in general? *Use ggplot2 for your graphics.*