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 https://ggplot2.tidyverse.org/reference/
```

Tips

```
When you are installing packages, you might get the following message:

  There are binary versions available but the source versions are later:
      binary source needs_compilation
XXXXXX  0.2.3  0.2.4              TRUE
XXXXXX  1.2.2  1.2.4              TRUE

Do you want to install from sources the packages which need compilation?
 (Yes/no/cancel)

My advice is to choose "no", which means you might not get the absolute
latest version of the package – but you will still get a good enough
recent working version.

Major Tip
If you can't write a script in R, ask your tutor how to now!

A good resource to use is the ggplot2 cheat sheet. See link below, and on
Moodle

https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
```

2. Create some simple graphs to gain a better understanding of the mpg data, which comes as part of the ggplot2 package. For information on the data set use `? mpg`. Use this simple, data set to create the best looking graphs you can using base graphics and ggplot2. Review the elements of a figure (Slide 54) as design factors you should consider. For example, think about the weight of lines, colours used, size of typefaces, position of elements to create simple stylish figures.

   Some motivating questions to investigate:

(a) What is the relationship between city (cty) fuel consumption and highway (hwy) consumption?

(b) How is fuel consumption (cty/hwy) related to manufacturer, transmission, class etc.? Are there manufacturers or car types with particularly high or low fuel consumption?

(c) How is fuel consumption related to the number of cylinders (cyl), or engine displacement (displ)?

(d) Are there any other interesting relationships you can find in the data?

(e) Did cars become more or less fuel efficient over time? How strong is your evidence (perhaps use a non-graphical justification for this last part)?

3. 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

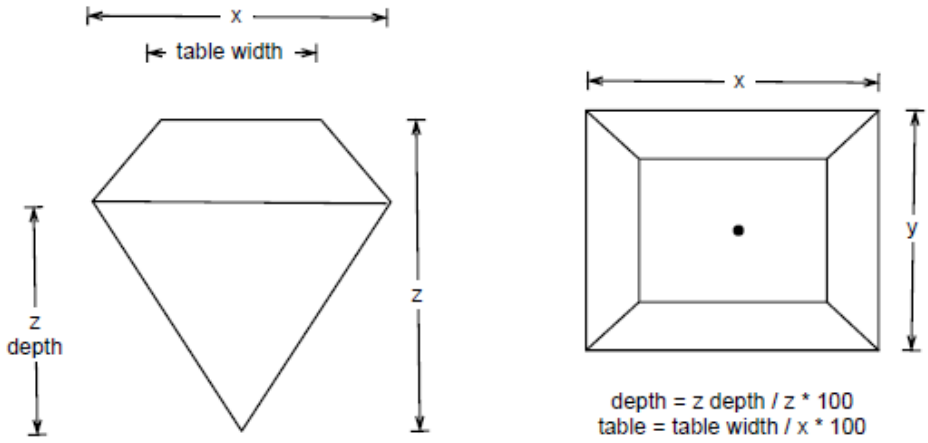

Fig. 2.1: How the variables x, y, z, table and depth are measured.

(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.*

Tips

```
Try and plot price as a function of each of the variables.

You can add extra dimensions to plot by varying size or colour of the
plotted points.
```

4. 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 pairs of 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.

Tips

Consider the correlation between height and other variables.

To examine males or females separately you may need to subset part of your data set. Some ideas: https://www.statmethods.net/management/subset.html

5. The data file "Dunnhumby1-20.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, facets and 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.*

Tips

```
You may need to use a grouping function, such as "by" to calculate stats
for each customer. Open the file and make sure you understand the
contents of each column.

For plotting, start with something easy. Try and plot a histogram of
amount spent during visit for all customers. Once you can plot that use
faceting to create individual histograms for each customer.
```

Extension: (a former sample exam question given without solution)

6. A World Health study is examining how life expectancy varies between men and women in different countries and at different times in history. The table below shows a sample of the data that has been recorded. There are approximately 15,000 records in all.

Country	Year of Birth	Gender	Age at Death
Australia	1818	M	9
Afghanistan	1944	F	40
USA	1846	F	12
India	1926	F	6
China	1860	F	32
India	1868	M	54
Australia	1900	F	37
China	1875	F	75
England	1807	M	15
France	1933	M	52
Egypt	1836	M	19
USA	1906	M	58

Using one of the graphic types from the Visualization Zoo (see formulae and references for a list of types) suggest a suitable graphic to help the researcher display as many variables as clearly as possible.

Explain your decision. Which graph elements correspond to the variables you want to display? You may want do a brief sketch to show how your graphic would be constructed.

Tips

```
Think about the number of dimensions in the data, and how each attrbitue
would be best shown.
```