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/
```

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)?

```
# some starting points
library(ggplot2)
attach(mpg)

# Q2.a
# Plot variables to get a sense of their relationship.
# Both variables are continous so scatter plot can be a good option.
# It seems like a strong linear relationship.
plot(cty, hwy, main = "Highway and City Fuel Efficiency")

# Lets see correlation coefficient to quantify the strength of linear
relationship
cor(cty, hwy)

# Let's fit a regression line to be able to mathematically describe the
relationship
# By looking at the summary of the model,
# we can express the relationship between hwy and city with the following
expression
# E[hwy] = 0.89204 + 1.33746 * cty
fitted = lm(hwy ~ cty)
summary(fitted)
abline(fitted)

# Q2.b
```

```
# This time we want to see the relationship between 1 categorical variable (i.e
manufacturer)
# and a continous variable (i.e cty or hwy). Hence, we can draw a seperate
boxplot for each catogry.
# las: for orienting the x-axis labels
# -1* hwy is to reverse the order
# cex: for
group_order = with(mpg, reorder(manufacturer, -1*hwy, median, ))
boxplot(hwy ~  group_order, las = 3, xlab="", cex=0.8)
# According to plot Honda has the best fuel efficiency whereas landrover has the
highest consumption

#Q2.c
# We are asked to investigate the relationship between
# 2 continuous (i.e hwy/cty and displ) and 1 categorial variable (i.e cyl)
# There can be different options to visualise this information;
# - Facet of Scatter plots
# - 1 Scatter plot with categorical variable used for colouring points
# - 1 Scatter plot with shapes of the points are based on categorical variable's
value
# ...

# Given that there are only 4 different cylinder,
# using colors wouldn't be confusing.
plot(displ, hwy, col=cyl, cex=1, main="Cylinders - Displacement - Highway Fuel
Efficiency")
legend("topright" , legend=sort(unique(cyl)), pch=rep(1,4)
,col=sort(unique(cyl)), cex=1, title = "Cylinders")

# Looking at the plot above, we can clearly see the relationship between hwy-
cyl-displ.
# Higher number of cylinders and higher displacement are signals of low hwy.
# We can also see that higher number of cylinders tend to have a higher
displacement value.
# Adding another variable to existing plot by labeling the points
# We can also observe that drv="f" cars tends to have higher hwy and lower displ
and cylinders
# Whereas drv="r" or "4" have opposite relationships.

plot(displ, hwy, col=cyl, cex=2, main="Drive Type - Cylinders - Displacement -
Highway Fuel Efficiency")
text(displ, hwy, labels=drv, cex=0.8)
legend("topright" , legend=sort(unique(cyl)), pch=rep(1,4)
,col=sort(unique(cyl)), cex=1, title = "Cylinders")

# Q2.d

# Investigate relationship between categorical variables using table function
# We can see dodge, forde, toyota and volkswagen are among the common
manufacturers.
# Also can't see a big change between the years.
table(manufacturer, year)

# We can see that dodge and chevrolet mostly producec automatic trans. cars.
# Whereas Toyota and Volkswagen have both automatic and manual cars more
balanced.
table(manufacturer, trans)
# You can turn those counts into percentage by dividing the total number of cars
of each brand as follows;
round(table(manufacturer, trans) / as.vector(table(manufacturer)),2)

# Investigate relationship between 2 numerical variables using scatterplot
matrix and correlation matrix.
# We can see a strong relationship between displ & cty/hwy
```

```
# Also displ and cyl looks associated with each other.
plot(mpg[,c(3,4,5,8,9)])
cor(mpg[,c(3,4,5,8,9)])

# Investigating relationship between a categorical and a numeric variable
unique(drv)

# par function lets us set graphical parameters
# using mfrow argument, we create a grid then several graphs can be visualised
# as elements of that grid. This makes it easier to compare several graphs.
par(mfrow=c(3,1))

# hist function requires a vector as input so mpg[drv=="f","hwy"] will not work
because this type
# of selection returns a list. Therefore, we first select the "hwy" column as a
vector then filter
# it based on the value of "drv".
hist((hwy[drv=="f"]), main = "Drive f", xlab="hwy")
hist((hwy[drv=="r"]),main = "Drive r", xlab="hwy")
hist((hwy[drv=="4"]),main = "Drive 4", xlab="hwy")

# We can see that full drive cars have higher hwy values -- mostly around 25-30
-- whereas Rear and 4 wheel drives have hwy values
# clustered around 15-25.

# Q2.e
# We can see that year column have only 2 entries. Hence, we can divide the data
into 2 groups
# and apply hypothesis test to see if the difference between fuel efficiency of
those groups
# are statistically significant or not.
unique(year)

# Looking at the results, p-value is too large to reject null hypothesis.
# Hence, we can say this data doesn't provide sufficient evidence of difference
between the efficiency of
# the cars that are produced at 1999 and 2008.
a = as.data.frame(mpg[year==1999,"hwy"])
b = as.data.frame(mpg[year==2008,"hwy"])
t.test(a$hwy,b$hwy,"two.sided")
```

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

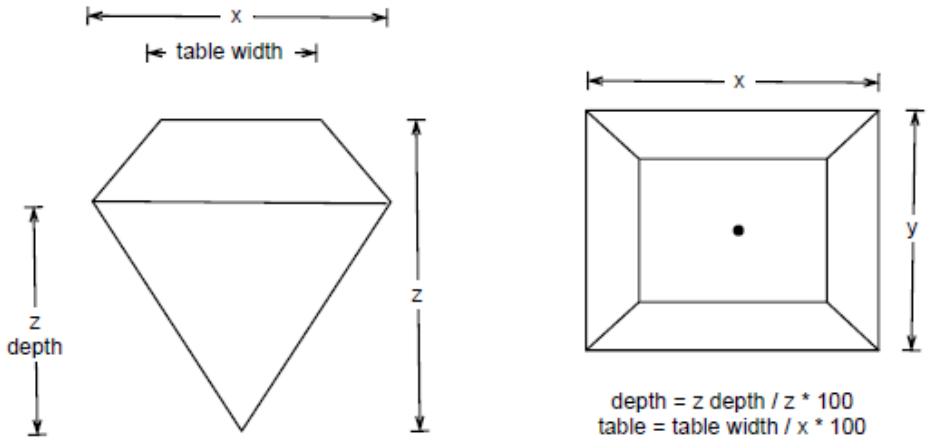


Fig. 2.1: How the variables x, y, z, table and depth are measured.

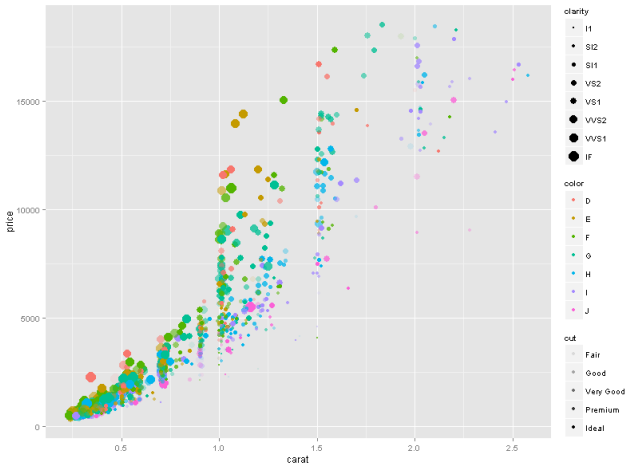
(a) Taking a random sample using the code below, create a subset of the diamonds data set: ‘dsmall’ to use in the following analysis.

```
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.*

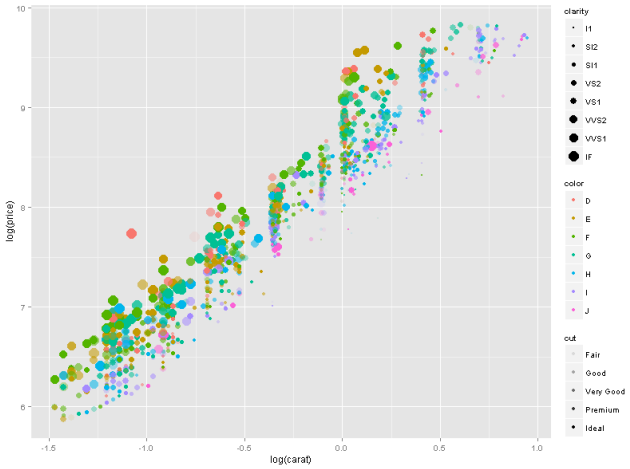
```
> qplot(carat, price, data = dsmall, color = color, size = clarity, alpha
= cut)
```

```
# note that I have not assigned the categorical variables (color, cut,
clarity) to any particular symbol setting - notwithstanding, you can still
see a systematic effect by the patterning in the plot.
```




```
> # taking logs of both numeric variables gives a linear model.

> qplot(log(carat), log(price), data = dsmall, color = color, size =
clarity, alpha = cut)
```



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.

(a) predicting height

```
> rm(list = ls())
> body.dat <-
read_csv("body.dat.csv")
> by(body.dat, body.dat[25],
function(df)
round(cor(df[1:23], df[,24]),
digits = 2))
```

```
Gender: Female
              Height
ShoulderWidth   0.47
Pelvis          0.42
Hips            0.37
ChestDepth      0.11
ChestDiam       0.24
ElbowDiam       0.41
WristDiam       0.40
KneeDiam        0.35
AnkleDiam       0.40
ShoulderGirth   0.28
Chest           0.20
Waist           0.14
Abdomen         0.23
HipGirth        0.30
ThighGirth      0.23
Bicep           0.12
Forearm         0.24
KneeGirth       0.40
CalfGirth       0.32
AnkleGirth      0.31
WristGirth      0.36
Age            -0.06
Weight          0.43
--------------------------------
Gender: Male
              Height
ShoulderWidth   0.48
Pelvis          0.42
Hips            0.50
ChestDepth      0.33
ChestDiam       0.31
ElbowDiam       0.49
WristDiam       0.33
KneeDiam        0.35
AnkleDiam       0.42
ShoulderGirth   0.30
Chest           0.25
Waist           0.21
Abdomen         0.26
HipGirth        0.35
ThighGirth      0.24
Bicep           0.20
Forearm         0.27
KneeGirth       0.41
CalfGirth       0.24
AnkleGirth      0.36
WristGirth      0.34
Age            -0.04
Weight          0.53
```

(b) predicting weight

```
> # swap columns
> body.dat =
body.dat[,c(1:22,24,23,25)]
> by(body.dat, body.dat[25],
function(df)
round(cor(df[1:23], df[,24]),
digits = 2))
```

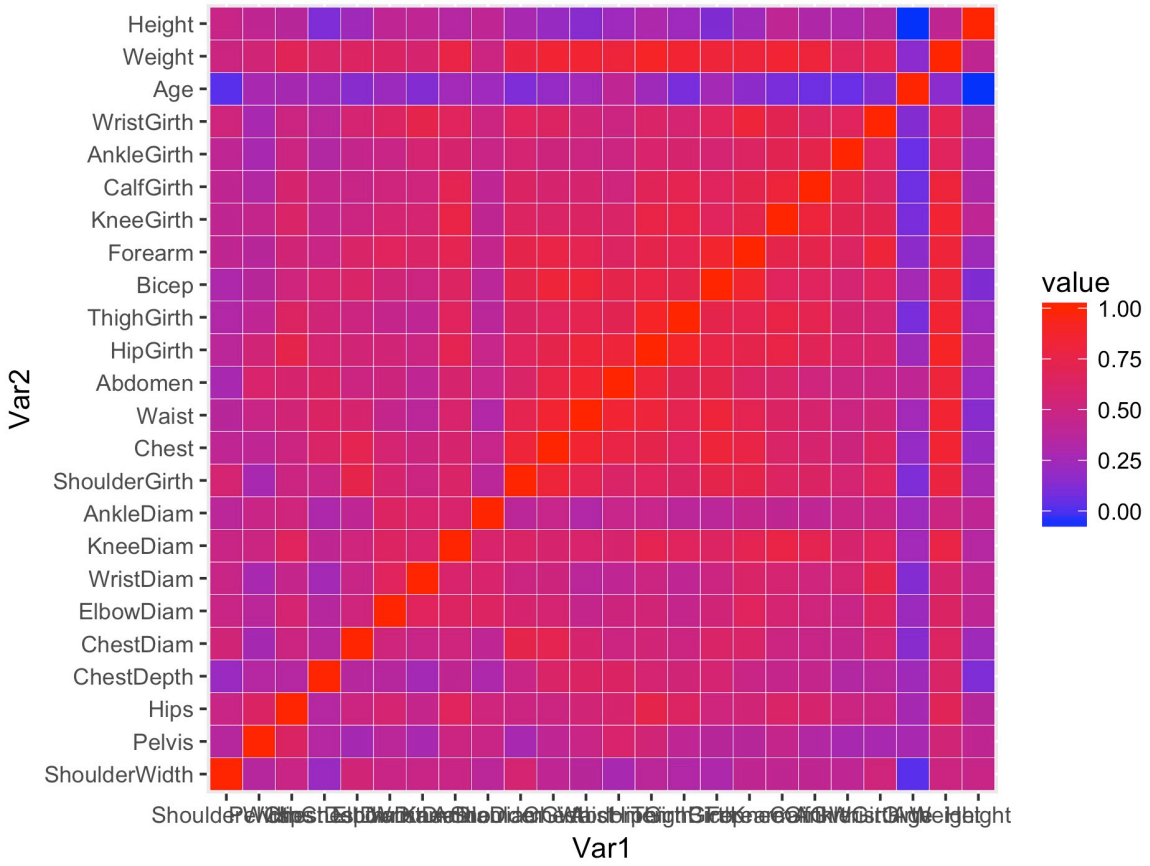
```
Gender: Female
              Weight
ShoulderWidth   0.50
Pelvis          0.55
Hips            0.69
ChestDepth      0.61
ChestDiam       0.64
ElbowDiam       0.63
WristDiam       0.58
KneeDiam        0.78
AnkleDiam       0.49
ShoulderGirth   0.79
Chest           0.84
Waist           0.86
Abdomen         0.80
HipGirth        0.90
ThighGirth      0.86
Bicep           0.82
Forearm         0.83
KneeGirth       0.84
CalfGirth       0.80
AnkleGirth      0.67
WristGirth      0.72
Age             0.15
Height          0.43
--------------------------------
Gender: Male
              Weight
ShoulderWidth   0.42
Pelvis          0.58
Hips            0.67
ChestDepth      0.72
ChestDiam       0.73
ElbowDiam       0.60
WristDiam       0.53
KneeDiam        0.51
AnkleDiam       0.51
ShoulderGirth   0.76
Chest           0.79
Waist           0.81
Abdomen         0.78
HipGirth        0.88
ThighGirth      0.77
Bicep           0.69
Forearm         0.69
KneeGirth       0.74
CalfGirth       0.69
AnkleGirth      0.64
WristGirth      0.58
Age             0.14
Height          0.53
```

(c) Looking at all the cross correlations is probably overkill but you can use the function below (which uses a technique introduced in Lecture 3):

```
> by(body.dat, body.dat[25], function(df) round(cor(df[1:24]), digits =
2))
```

Alternatively:

```
# Code by Abishek to draw heat map of correlations
body = read.csv("body.dat.csv")
split = by(body, body$Gender, function(df) round(cor(df[c(1:24)]),digits =
2))
library(reshape2)
female =  melt(split$Female) #you could adapt for males
g = ggplot(data = female, aes(x=Var1, y=Var2, fill=value))
g = g + geom_tile(color = "white")+  scale_fill_gradient(low = "blue",
high = "red")
g
ggsave("bodyfheat.jpg", g, width = 16, height = 12, units = "cm")
```



(d) if you had to use a single measure to distinguish men and women. What would it be? Histogram? Boxplot? ...

For example

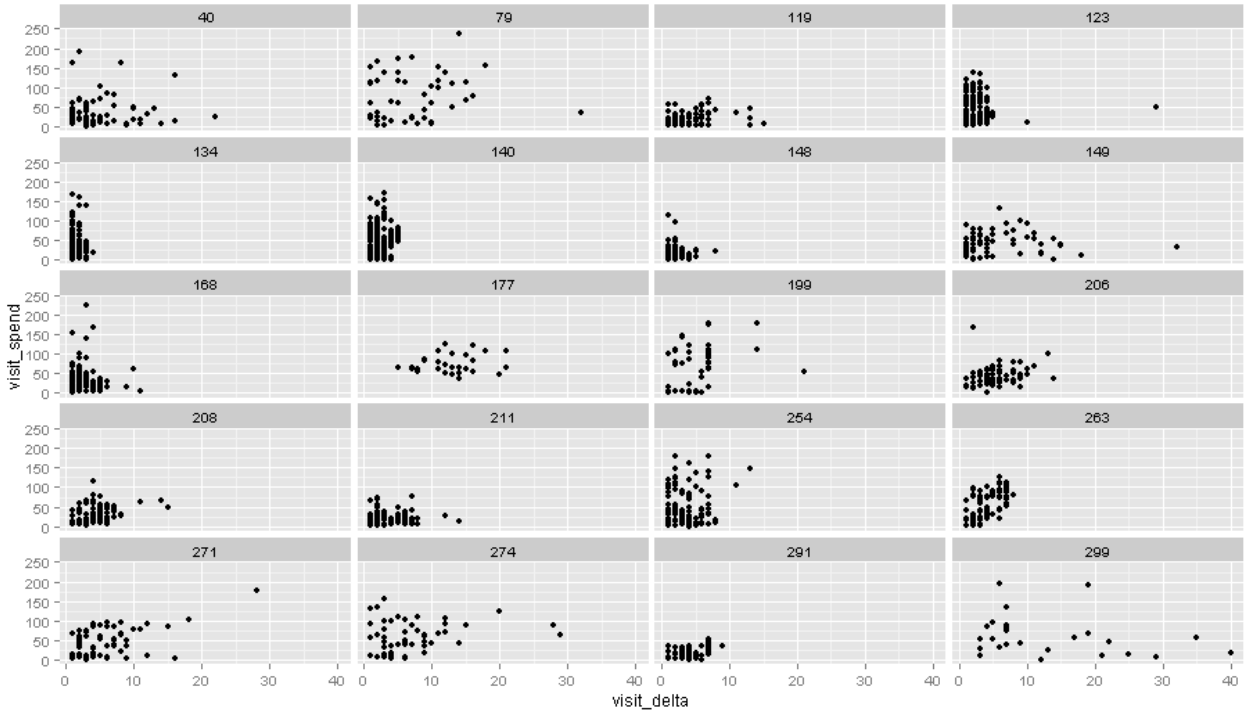
```
> qplot(ShoulderWidth, data = body.dat, geom = "histogram", facets =
Gender ~ .)
```

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.*

   Below is one example of how to investigate the 2 different customers looking at scatterplots showing time between visits and amount spent.

```
> qplot(visit_delta, visit_spend, data = Dunnhumby1.20, geom = "point",
facets = customer_id ~ .) + facet_wrap(~ customer_id, ncol = 4)
```



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.