Minilab 4a Worksheet

Exploratory Data Analysis

In this minilab, we will look at visualisation and summary statistics for two quantitative variables through scatterplots and correlation.

1. Scatterplots and Correlation

In the lectures for this module this week, we have looked at the idea of *correlation* and calculating Pearson's correlation coefficient (denoted by r). This is a single number summary statistic of the relationship between two quantitative variables. We always have $-1 \leq r \leq 1$. *Make sure you check out the lectures on Aula if you have not done so already.* Below are six examples of scatterplots, where each has a different value of the correlation coefficient.

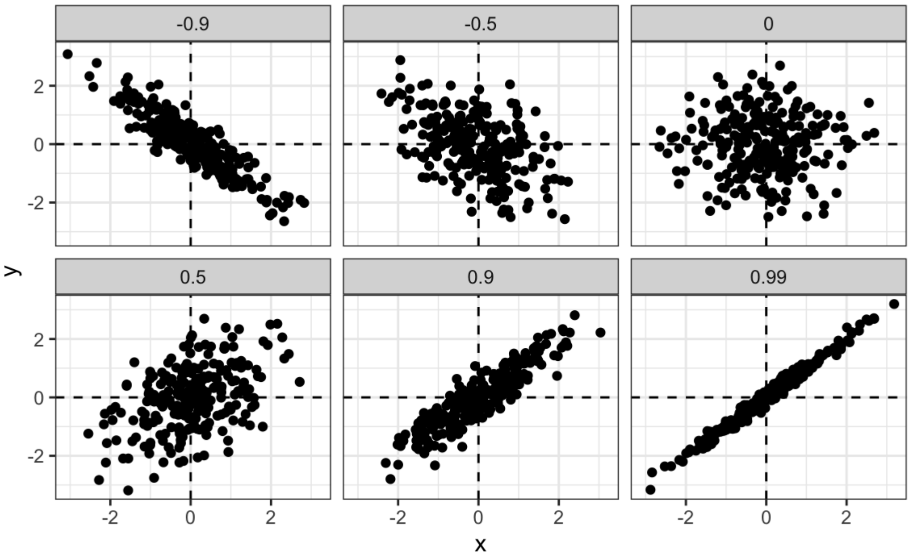


Image from: https://rafalab.github.io/dsbook/regression.html

In this minilab, we now look at how to produce scatterplots and calculate Pearson's correlation coefficient using R and the tidyverse.

As an example dataset, we will use *Anscombe's quartet* which is actually four datasets with exactly the same summary statistics. This data is built into R, so we can take a look for ourselves.

(1) Given a dataset with two quantitative variables, we can calculate summary statistics for each variable (including the mean, median, quartiles, max and min) using summary(). The standard deviation of each variable is calculated using sd() and the correlation coefficient between two variables is calculated using cor().

```
library(tidyverse)
# Anscombe's quartet
View(anscombe)
# first Anscombe dataset given by columns 1 and 5
x = anscombe[,1]
y = anscombe[,5]
summary(x)
sd(x)
summary(y)
sd(y)
cor(x,y)
```

An important next step in Exploratory Data Analysis (EDA) is to construct a scatterplot to visualise the relationship between the two variables.

```
library(tidyverse)
ggplot(NULL,aes(x=x,y=y)) +
  geom_point()
```

Our eyes detect a relationship and the correlation coefficient supports this since we have calculated above (using the `cor()` function) $r = 0.8164205$ which indicates a strong positive relationship.

(2) Repeat these calculations for the second dataset from Anscombe's quartet.

```
# Anscombe's quartet
# second Anscombe dataset given by columns 2 and 6
x = anscombe[,2]
y = anscombe[,6]
```

What do you notice about the summary statistics?

What do your eyes notice about the relationship between the two variables?

We can easily calculate a correlation coefficient. But, because there is a clear *nonlinear* relationship, there is no conclusion we can draw from the correlation coefficient.

(3) Repeat these calculations for the third dataset from Anscombe's quartet (columns 3 and 7).

Notice that there is one very clear outlier, which is the third data point in the dataset (corresponding to $x = 13$). We can omit this data point (use the R code below) and repeat the analysis.

```
x = anscombe[-3,3]
y = anscombe[-3,7]
```

What can you conclude from the corresponding scatterplot and value of the correlation coefficient?

(4) For the fourth dataset from Anscombe's quartet (columns 4 and 8), there is one clear outlier.

If that outlier is omitted from the dataset, what is the corresponding value of the correlation coefficient?

(5) In a previous minilab we looked at the gapminder dataset.

```
# install.packages("gapminder")    # run only once
library(gapminder)
ggplot(gapminder, aes(x=gdpPercap, y=lifeExp)) +
  geom_point(aes(colour=continent))
```

We can use a dplyr pipe and a summary function to calculate a *correlation coefficient*. For example, for the data from 2007, we can calculate a correlation coefficient between the two variables gdpPercap and lifeExp for each continent. In this case, note that in this dataset Oceania only has two countries (Australia and New Zealand) so the correlation is $r = 1$ (there is ALWAYS a straight line through exactly two points).

```
gapminder %>%
  filter(year==2007) %>%
  group_by(continent) %>%
  summarise(num_countries=n_distinct(country),
            r=cor(gdpPercap,lifeExp))
```

Exercise. For the “diamonds” dataset, write a dplyr pipe that produces a summary table showing for each *cut* of diamond the number of diamonds of that cut and the correlation coefficient of carat and price. Which cut of diamond has the highest correlation coefficient between carat and price?

2. Multivariate Data

In a typical dataset, we often have many variables, some of which are quantitative and some of which are categorical.

(1) We will look at a dataset which gives credit card balance ($), income (in thousands), credit card limit, credit rating, age, and years of education for 400 people.

```
install.packages("ISLR")
library(tidyverse)
library(ISLR)
credit = as_tibble(Credit)
credit = select(credit,Balance,Limit,Income)
summary(credit)
```

The summary() function provides a useful summary of each variable separately.

(2) Let us take a look at the correlation between the variables (take them in pairs). First we could look at the scatterplot between Balance and Limit. We can also calculate the correlation coefficient (which is $r = 0.8616973$).

```
ggplot(credit, aes(x=Balance, y=Limit)) +
  geom_point()
cor(credit$Balance,credit$Limit)
```

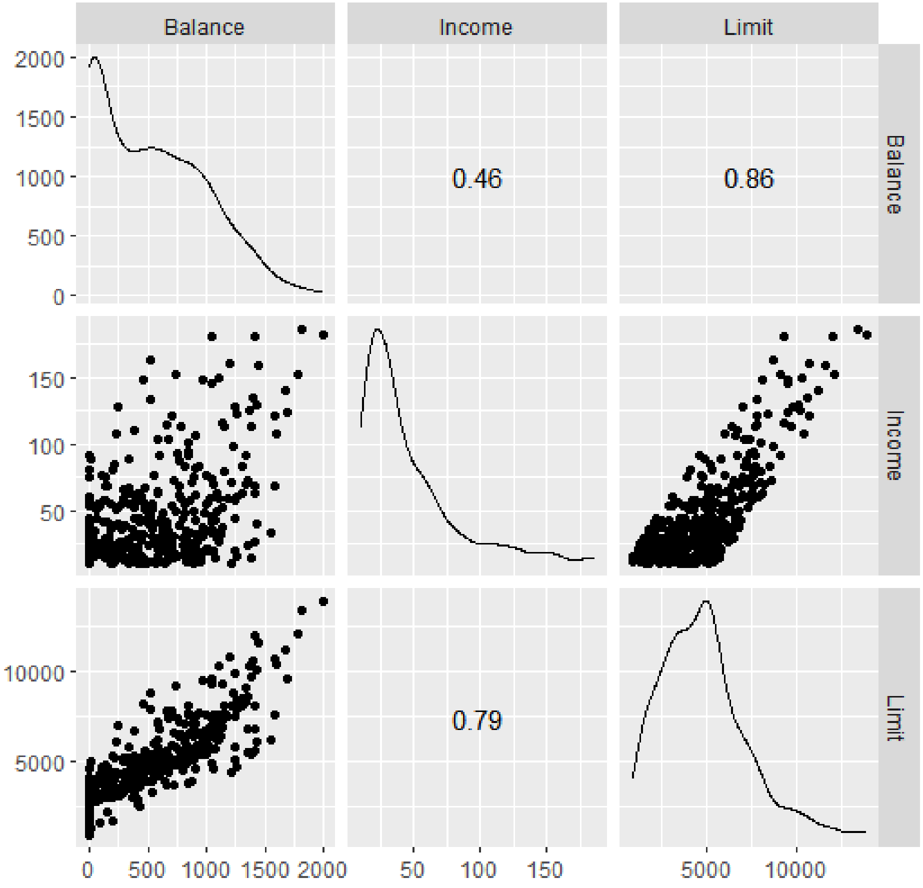
Notice the clear way in which a scatterplot is build using ggplot by specifying the dataset (in this case the *credit* tibble), the aesthetic mapping (saying that the x-axis will have the *Balance* variable and the y-axis will have the *Limit* variable), and then what type of geometric object will be plotted (in this case *points*).

It is easy to see how this code could be repeated to look at Balance and Income as a pair of variables and Income and Limit as a pair of variables.

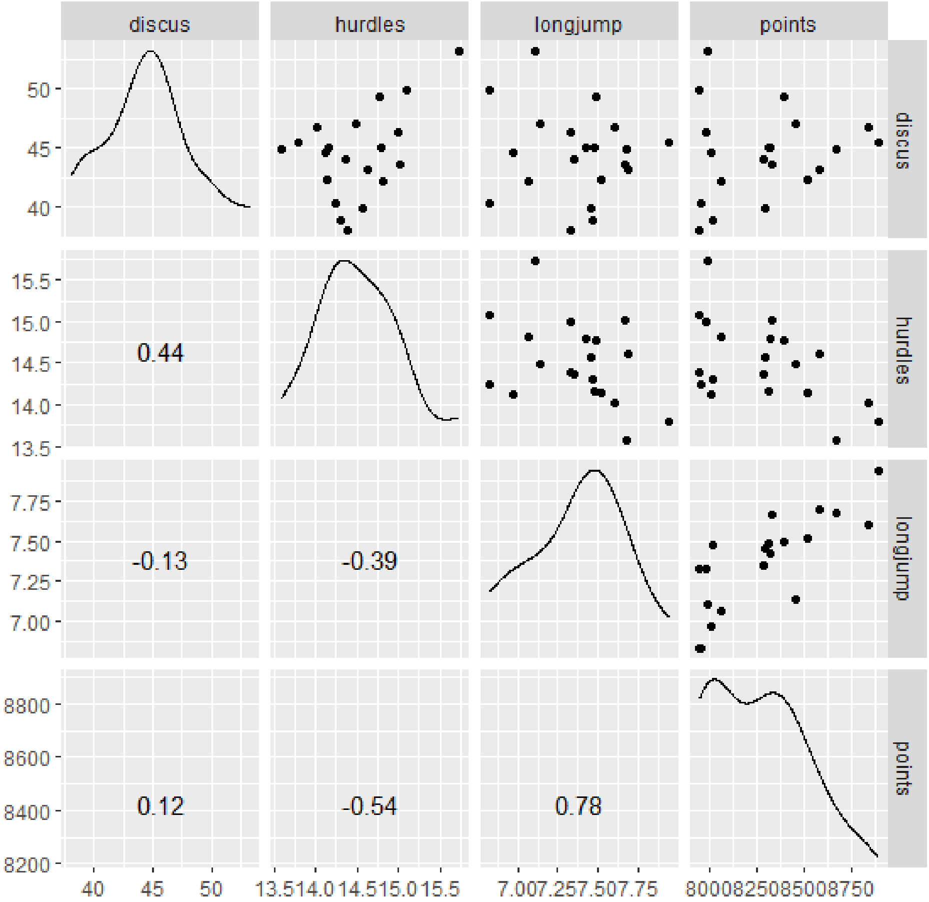
(3) There is a quick and easy way to look at all the pairs of quantitative variables at the same time in a *scatter matrix* (a matrix or grid of scatterplots).

```
install.packages("GGally")    # run only once
library(GGally)
ggscatmat(select(credit,Balance,Limit,Income))
```

For each pair of variables, this lovely plot gives a scatterplot and a correlation coefficient, together with a smoothed histogram for each variable (along the diagonal). We can see that the variables are clearly all positively correlated (from the correlation coefficients or through the scatterplots).



Exercise. In the Decathlon, athletes compete in ten track and field events, gaining points in each event depending on their performance. The scatter matrix below shows partial results of the top 20 finishers in the Decathlon at the 2016 Rio Olympics. Performances for the *discus* throw (units of metres), 110m *hurdles* (units of seconds) and *longjump* (units of metres) events are given, along with the overall *points* gained over all ten events. The winner is the athlete with the most overall points (Ashton Eaton with 8893 points).



(a) Briefly explain the <u>implications</u> arising from the negative correlations shown.

(b) Consider the performance of the overall winner (Ashton Eaton with 8893 points). From the scatterplots shown, what do you notice about this athlete's performance in each of the three events?

Summary

In this minilab, we have seen how to plot a scatterplot and calculate Pearson's correlation coefficient. For multivariate data, use a scatter matrix to quickly explore the relationships between any quantitative variables. This should be an automatic "first step" when starting to explore a given dataset.