Introduction to Data Science (11372 & G 11516)

Semester 2 2020

UNIVERSITY OF CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 9

Dr. Ibrahim Radwan

DISTINCTIVE BY DESIGN

1

OUTLINE

UNIVERSITY OF CANBERRA

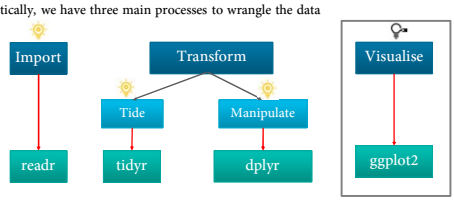
- Data Wrangling, a recap
- Exploratory Data Analysis
- Data Visualisation
- Grammar of Graphics
- EDA; Univariate Analysis

2

DATA WRANGLING

UNIVERSITY OF CANBERRA

- Practically, we have three main processes to wrangle the data



3

DATA IMPORT (RECAP)

UNIVERSITY OF CANBERRA

```
read_*(file, col_names = TRUE, col_types = NULL, locale = default_locale(), na = c("", "NA"),
  quoted_na = TRUE, comment = "", trim_ws = TRUE, skip = 0, n_max = Inf, guess_max = min(1000,
  n_max), progress = interactive())
```

Comma Delimited Files
read_csv("file.csv")

Tab Delimited Files
read_tsv("file.tsv")
Also **read_table()**

Semi-colon Delimited Files
read_csv2("file2.csv")

Files with Any Delimiter
read_delim("file.txt", delim = "|")

To save data into csv or txt file

Comma delimited file
write_csv(x, path, na = "NA", append = FALSE, col_names = !append**)**
File with arbitrary delimiter
write_delim(x, path, delim = " ", na = "NA", append = FALSE, col_names = !append**)**

4

DATA MANIPULATION (RECAP)

UNIVERSITY OF CANBERRA

- The `*dplyr*` package in `*tidyverse*` library presents five *verbs* for manipulating the data in data frames:
  1. filter() extracts a subset of the rows (i.e., observations) based on some criteria
  2. select() extracts a subset of the columns (i.e., features, variables) based on some criteria
  3. mutate() adds or modifies existing columns
  4. arrange() sorts the rows
  5. summarise() aggregates the data across rows (e.g., group them according to some criteria)
- Each of these functions takes a data frame as its first argument and returns a data frame.

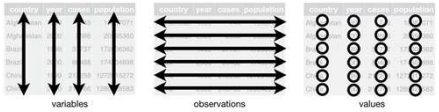
5

TIDY DATA (RECAP)

UNIVERSITY OF CANBERRA

- There are three interrelated rules which make a dataset tidy:
  1. Each variable must have its own column.
  2. Each observation must have its own row.
  3. Each value must have its own cell.

Having your data in a tidy format is crucial for data manipulation and exploring



Credit: R for Data Science

6

TIDY DATA (RECAP) – 2

UNIVERSITY OF CANBERRA

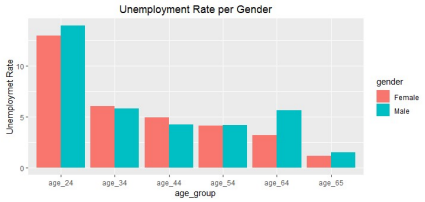
- The `*tidyr*` package presents four main *verbs/functions* to tide up the data:
  1. gather() collapses multiple columns into key-value pairs. It produces a "long" data format from a "wide" one.
  2. spread() takes two columns (key & value), and spreads into multiple columns: it makes "long" data wider. This is the reverse of gather.
  3. unite() unites multiple columns into one
  4. separate() takes a column and divides it into multiple columns
- Each of these functions takes a data frame as its first argument and returns a data frame.

7

DATA VISUALISATION

UNIVERSITY OF CANBERRA

- *"The simple graph has brought more information to the data analyst's mind than any other device."* — John Tukey



8

DATA VISUALISATION (2)

UNIVERSITY OF CANBERRA

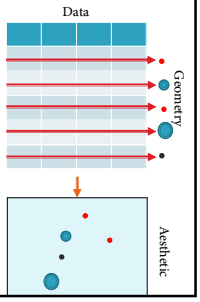
- Presenting the data variables into a pictorial or a graphical format
- Visualising data provides a guide to:
  - Check changes in variables
  - Inspect the differences or relations between variables
  - Find patterns in the data
  - Grasp new concepts or insights from the data
- Data visualisation should be easy to the stakeholders
- Data visualisation aids data modelling processes

9

GRAPH COMPONENTS

UNIVERSITY OF CANBERRA

- To build a graph in R , you will need to specify three components:
  1. **Data**: the set of records/variables that we need to represent with a graph
  2. **Geometry**: the type of the plot, which will be generated, usually it is a function such as (scatterplot, boxplot, barplot, histogram, smooth density, etc.)
  3. **Aesthetic mapping**: the coordinate map and the other visual cues, such as size, scale and color.



10

BUILD GRAPHS IN R

UNIVERSITY OF CANBERRA

- There are two functions in ggplot2 library to build a graph in R:

qplot()

Quick **plot** that encapsulates the three graph components all in the call of the function.

ggplot()

Build the graph **layer by layer**. Start by defining an empty frame and then add the subsequent operations.

We will be focusing on using this method to build the graphs

11

BUILD GRAPHS IN R (2)

UNIVERSITY OF CANBERRA

- *Graph grammar* is an elegant way to use few functions to be able to build many graphs and plots layer by layer by combining these functions together.
- The template for building a ggplot graph:

```
ggplot(data = <DATA>) +
    <GEOM_FUNCTION>(mapping=aes(<MAPPINGS>))
```

- The ggplot2 is one of the core members of the `*tidyverse*` library, so you will be able to use its functions when loading the `tidyverse` library.
  - *library(tidyverse)*

12

GRAPHS IN R - EXAMPLE

UNIVERSITY OF CANBERRA

- Let us start by looking at this dataset (mpg), which provides information about "Fuel economy data from 1999 to 2008 for 38 popular models of car"
- Then, we need to build a plot to answer some questions such as:
  - Do cars with big engines use more fuel than the car with small engines?
- To answer this question, let us have a look to the variables of this dataset:

```
> mpg
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compa~
 2 audi         a4           1.8  1999     4 manual(m~  f        21    29 p     compa~
 3 audi         a4           2    2008     4 manual(m~  f        20    31 p     compa~
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compa~
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compa~
 6 audi         a4           2.8  1999     6 manual(m~  f        18    26 p     compa~
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compa~
 8 audi         a4 quattro   1.8  1999     4 manual(m~  4        18    26 p     compa~
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compa~
10 audi         a4 quattro   2    2008     4 manual(m~  4        20    28 p     compa~
# ... with 224 more rows
```


13

GRAPHS IN R – EXAMPLE (2)

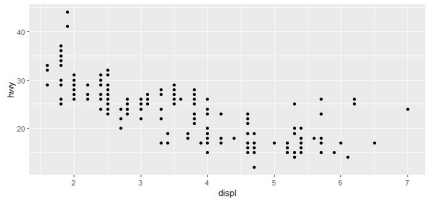
UNIVERSITY OF CANBERRA

Create an empty coordinate system

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

Add a layer of scatter plot to represent the relationship between two variables

You may add as many layers as needed

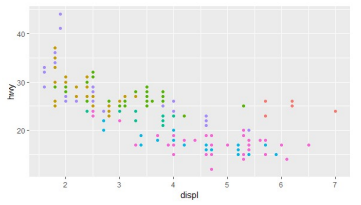


14

GRAPHS IN R – EXAMPLE (3)

UNIVERSITY OF CANBERRA

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = class))
```



Three variables are represented in the 2D graph

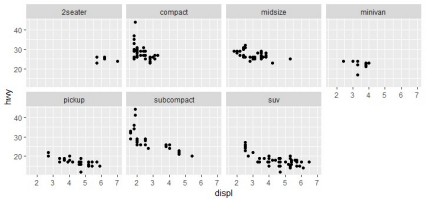
15

GRAPHS IN R – EXAMPLE (4)

UNIVERSITY OF CANBERRA

You may add additional variables by using facets (e.g. subplots)

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~class, nrow = 2)
```



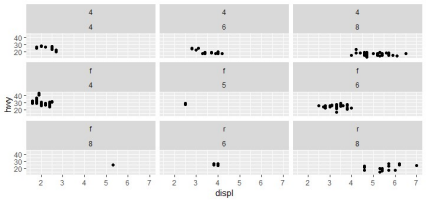
16

GRAPHS IN R – EXAMPLE (5)

UNIVERSITY OF CANBERRA

The variables () should be discrete when using facets

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ cyl)
```



17

EXPLORATORY DATA ANALYSIS (EDA)

UNIVERSITY OF CANBERRA

- EDA: *is the process of exploring the data variables toward discovering some trends or patterns from the data. This leads the modelling step toward fixing issues or guiding the decision making.*
- To understand the variables in a dataset, we may transform these variables into other format or extract their summaries (e.g. mean, variance, etc.) or to get insights about the distribution of these variables.
- The most elegant way to understand relationships with-in a variable or between variables is by ***visualising*** these relationships.

18

EXPLORATORY DATA ANALYSIS (EDA) – 2

UNIVERSITY OF CANBERRA

- To extract the relationships between variables or to discover the patterns/distributions of the variables, we need to check on the type of these variables.
- To conduct the data analysis on variables for sake of understanding their relationships, this analysis can be either:
  - Uni-variate analysis
    - Discover the variations of the data into **one** variable
  - Multi-variate analysis
    - Discover the co-variation of **multiple** variables
    - Bi-variate analysis is a special type of this analysis with only **two** variables.

19

EXPLORATORY DATA ANALYSIS (EDA) – 3

UNIVERSITY OF CANBERRA

- EDA
  - Uni-variate analysis
    - Discrete
    - Continuous
  - Bi-variate analysis (can be extended to multi-variate analysis)
    - Discrete
    - Continuous

We will start by using the visualisation to do the EDA for both of the univariate and bi-variate analysis.

20

EDA – UNIVARIATE

UNIVERSITY OF CANBERRA

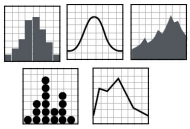
- There are two types of visualization-based univariate analysis:
  - Visualising variation of continuous variable
  - Visualising variation of discrete variable
- Examples of univariate continuous :
  - Histograms, etc.
- Examples of univariate discrete:
  - bar plots, etc.

21

EDA – UNIVARIATE, CONTINUOUS

UNIVERSITY OF CANBERRA

- The analysis is done based on just one variable, where it is a 'numerical' continuous variable.
- ggplot2 provides many functions to plot the variation of the univariate continuous variable such as:
  - *geom_histogram()*, for a histogram plot
  - *geom_density()*, for a density plot
  - *geom_area()*, for an area plot
  - *geom_dotplot()*, for a dot plot
  - *geom_freqpoly()*, for a frequency polygon

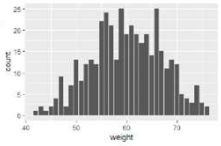


22

HISTOGRAMS

UNIVERSITY OF CANBERRA

- Summarise the variable by extracting a 1-D distribution of the variation in the data within this variable.
- The data are summarised using suitable binwidths.
- A histogram plot shows properties, such as:
  - center (i.e., the location) of the data;
  - spread (i.e., the scale) of the data;
  - skewness (i.e., left or right) of the data and
  - presence of outliers



23

HISTOGRAMS (2)

UNIVERSITY OF CANBERRA

- A Histogram plot of a variable can answer some questions on the data such as:
  - What kind of distribution do the data come from?
  - How spread out are the data?
  - Are the data symmetric or skewed?
  - Are there outliers in the data?

24

BUILDING A HISTOGRAM

UNIVERSITY OF CANBERRA

1- a) First layer contains the data specification
b) Also specifies the variable

2- a) Second layer specifies the type of plot; the geometric function

3- a) Third and following layers specify axis labels and further visual cues

```
ggplot(data,aes(x = variable_name))+
geom_histogram() +
xlab("x axis label") +
ylab("y axis label")+
ggtitle("plot title")
```

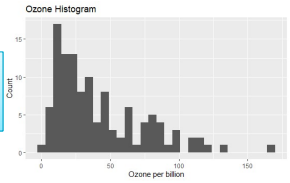

25

EXAMPLE

UNIVERSITY OF CANBERRA

Basic histogram visualization

```
ggplot(data= airquality, aes(x = Ozone)) +
  geom_histogram() +
  xlab("Ozone per billion") +
  ylab("Count") +
  ggtitle("Ozone Histogram")
```

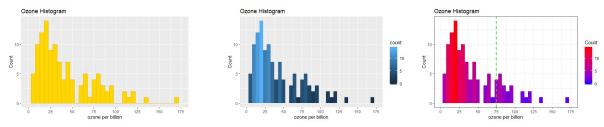


26

EXAMPLE (2)

UNIVERSITY OF CANBERRA

Adding extra visual cues and more options

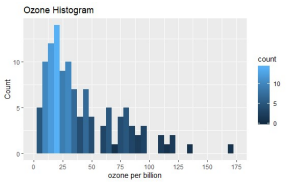


The full code is shared under week 10 on Canvas

27

HISTOGRAM INTERPRETATION

UNIVERSITY OF CANBERRA



Back to our questions...

- What kind of distribution do the data come from?
- How spread out are the data?
- Are the data symmetric or skewed?
- Are there outliers in the data?

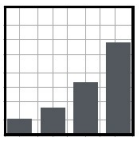
The data is right skewed and there are clearly outliers as the graph stretches on the scale to175 Ozone per billion. It seems the median is around the 20 Ozone per billion.

28

EDA – UNIVARIATE, DISCRETE

UNIVERSITY OF CANBERRA

- The analysis is done based on just one variable, where it is a discrete (i.e. categorical) variable.
- ggplot2 provides one functions to plot the variation of the univariate discrete, which is:
  - *geom_bar()*, for a bar plot



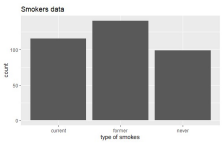
29

UNIVARIATE BAR PLOT

UNIVERSITY OF CANBERRA

- Summarise the variable by extracting a frequency of each level/category of the univariate variable.
- The x-axis represents the levels/categories of the data
- The y-axis becomes the counts of each category.

- Bar plots helps answering the following questions:
  - What is the spread of observations between data levels?
  - Which level has the most observations or least observations etc.



30

BUILDING A BAR PLOT

UNIVERSITY OF CANBERRA

1- a) First layer contains the data specification
b) Also specifies the variable

2- a) Second layer specifies the type of plot; the geometric function

3- a) Third and following layers specify axis labels and further visual cues

```
ggplot(data,aes(x = variable_name))+
geom_bar() +
xlab("x axis label") +
ylab("y axis label")+
ggtitle("plot title")
```

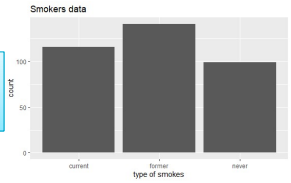

31

EXAMPLE

UNIVERSITY OF CANBERRA

Basic bar plot visualization

```
ggplot(data, aes(x = Smoke)) +
  geom_bar()+
  xlab("type of smokes") +
  ylab("count") +
  ggtitle("Smokers data")
```

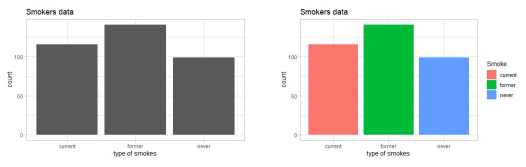


32

EXAMPLE (2)

UNIVERSITY OF CANBERRA

Adding extra visual cues and more options

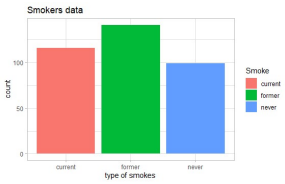


The full code is shared under week 10 on Canvas

33

BAR PLOT INTERPRETATION

UNIVERSITY OF CANBERRA



Back to our questions...
What is the spread of observations between levels of the data? Which level has the most observations or least observations etc. ?

Former smokers are more than the current, etc.

34

RECOMMENDED READING

UNIVERSITY OF CANBERRA

- You are recommended to read chapters 3 from the *"R for Data Science"* book:
  - https://r4ds.had.co.nz/data-visualisation.html

35

ANNOUNCEMENT

UNIVERSITY OF CANBERRA

- This week online test has been released since yesterday and is **due this Sunday**.
  - One attempt
  - 60 minutes long
  - Weights 20%

36