FIT3152 Data analytics. Tutorial 01: Introduction to R. Review of basic statistics

Note: Hints and tips will be added to this document during Weeks 1 and 2.

Group Questions:

1. Tutorial Activity as a class or in groups: Compare the two figures in Lecture 1 showing Internet activity over 2018/2019, and 2020 (Slides 5 and 6). Answer the following:
   - What are the trends – that is, what types of online activities are increasing in prevalence?
   - Looking at a particular activity, what types of data could be collected?
   - What could that data be used to study?
   - What changes might be due to COVID-19?

Tips

For example, think about Instagram.
The data collected is: time spent scrolling, number of likes, number of followers, number following, number of posts made, frequency of posting, amount of private messaging, ...
Time spent scrolling could be used as an indicator of how much free time that person has.
The number of followers might be a guide to the size of their friendship circle.
...

2. Tutorial Activity as a class or in groups: Using the examples of applications of data science in the real world as inspiration, find a recent application of data science from the media. Answer the following:
   - What is the problem to be solved?
   - What type of data is collected?
   - What type of analysis is performed?
   - What is the outcome?
   - How might you use this data to investigate another aspect of (human) activity?

Present your findings in Tutorial 1 as a 2 – 3-minute verbal report as part of the group discussion.

Tips

Sources of news articles you might find useful are:
https://www.theage.com.au/
https://www.nature.com/
https://www.news.com.au/

Individual/Small Group Questions:

Note, much of the data for the following questions has been sourced from http://www.statsci.org/datasets.html and links within.

1. Using the data sets provided as csv files and the lecture notes, try and reproduce all of the statistics and graphics from Lecture 1.

   Files are: {InvestA, InvestB, Toothbrush, Workers, Food Retail 2014-2020 time series}.csv

Tips

```
If you are having trouble reading files into R, put your data file on to
the desktop and set desktop as your working directory.
```

2. The following data records the length of rivers in the South Island of New Zealand. The lengths are given in kilometres. Data is grouped depending on where it flows into. Source: http://www.statsci.org/data/oz/nzrivers.html

   Pacific Ocean:
   209, 48, 169, 138, 64, 97, 161, 95, 145, 90, 121, 80, 56, 64, 209, 64, 72, 288, 322.

   Tasman Sea:
   76, 64, 68, 64, 37, 32, 32, 51, 56, 40, 64, 56, 80, 121, 177, 56, 80, 35, 72, 72, 108, 48.

   (a) Calculate the summary stats for each group of rivers. Draw a boxplot.
   (b) Test the hypothesis that rivers flowing into the Tasman Sea are shorter on average than those flowing into the Pacific Ocean. Use a significance of 1%

Tips

```
Will be added here.
```

3. When anthropologists analyze human skeletal remains, an important piece of information is living stature. Since skeletons are commonly based on statistical methods that utilize measurements on small bones, the following data was presented in a paper in the American Journal of Physical Anthropology to validate one such method. Variables are: MetaCarp – Metacarpal bone length in cm, Stature (Height of skeleton) in cm. Source: http://www.statsci.org/data/general/stature.html

MetaCarp	Stature
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

   Draw a scatterplot of the data with Stature as the vertical axis. Calculate the regression equation predicting Stature from MetaCarp. Comment on the accuracy of the model. Superimpose the line of best fit on your scatterplot.

Tips

```
If you are having trouble adding an extra line on scatter plot, check out
the abline function.
```

4. The ocean swell produces spectacular eruptions of water through a hole in the cliff at Kiama, about 120km south of Sydney, known as the Blowhole. The times at which 65 successive eruptions occurred from 1340 hours on 12 July 1998 were observed using a digital watch. Source: http://www.statsci.org/data/oz/kiama.html

   Challenge: download the data into R directly from: http://www.statsci.org/data/oz/kiama.txt (See ATHR page 18) or alternatively use the file: Data: kiama.txt

   Read these data into R, creating a vector named 'kiama'. Calculate the mean, standard deviation. Draw the default histogram. Using help, try and draw an improved histogram of your own design by changing range, class intervals and colour etc.

Tips

```
Will be added here.
```

5. The timber data are for specimens of 50 varieties of timber, for modulus of rigidity, modulus of elasticity and air dried density, arranged in increasing order of magnitude of the density. Source: http://www.statsci.org/data/oz/timber.html

   Read these data into R, creating a data frame named 'timber'. You can use the data file: timber.txt or load directly from: http://www.statsci.org/data/oz/timber.txt

   (a) which variable: elasticity or density is a better predictor of rigidity?

   (b) using your choice of variable calculate the regression equation predicting rigidity, draw a scatterplot of the data, showing the fitted model.

   (c) challenge: calculate the regression equation predicting rigidity as a function of both elasticity and density. Comment on the quality of your model vs the single predictor in (b).

Tips

```
If you want to load the data from an online data source, you can use the
path to the file on web (such as
"http://www.statsci.org/data/oz/timber.txt")  together with the same
functions that are used to load a local file into R.

Correlation between a predictor and the response variable can be a good
measure to assess quality of the predictor.

Also, to assess quality of a model, one option could be considering R-
squared value to see how good a fit the model is to the data.
```

6. Challenge: Using the data: InvestA.csv draw a boxplot. You will need to use the help file to work out the syntax – try ?boxplot as a starting point...

   Using the data: InvestA.csv, now use the 'aggregate' function to calculate the mean of each group. This is similar to the 'tapply' function but returns a data frame. Use help to work out the syntax...

Tips

```
If you are having trouble finding out how to group data inside the
boxplot, have a closer look at the "formula" argument of boxplot
function. Also, it is possible to use formula argument in aggregate
function. in the same way with boxplot function.
```

7. Analyse Victorian Retail Turnover: for Food retailing; Household goods retailing; Clothing, footwear and personal accessory retailing; Department stores; Other retailing; Cafes, restaurants and takeaway food service for the period Jan 2010 – Dec 2020 using Australian Bureau of Statistics data. You will need to copy the data from the Excel file: 8501.0 Retail Trade, Australia.xls.

   Draw a time series plot of the data and plot the time series decomposition. Comment on the main elements in the time series.

   Can you see any COVID-19 effects during 2020?

Tips

```
It is an .xls file with more than one pages, if you are having trouble
loading the data located at the second page, you might try to copy the
data on to the clipboard and load into R from clipboard. This operation
works differently for Mac, Windows and Linux. Quick Google search can be
the way to go!
```

Additional Statistics Notes:

Q2) Hypothesis Testing and p-value

In a statistical hypothesis test, a hypothesis about a certain population parameter is tested using sample data usually consisting of one or two groups.
The hypothesis is evaluated against a null hypothesis that assumes equality or no difference between the group(s) and the value being tested.
p-values evaluate how well the used sample data support the argument of the null hypothesis.
High p-values indicate that the sample data is likely with a true null hypothesis, while a low p-value indicates that the data is unlikely with a true null.
Therefore, a low p-value indicates that there is more evidence to conclude that the tested hypothesis is true against the null hypothesis.
Compare p-value against the desired significance to determine the likelihood that the hypothesis is true.

Q3)Regression

Provides a model for the relationship between different variables with the primary aim of prediction.
Simple linear regression models the relationship between one dependent variable and an associated independent variable.
Can plot expected value of dependent variable for all possible values of the independent variable using the derived formula.
Minimize sum of squared error (squared residuals) to come up with best fit.

Q4) Histogram

A histogram graphically displays the shape and spread of data.
Groups data into a number of class intervals and counts the frequency in each.

Q5) Correlation and Prediction

Correlation indicates the degree of association between two variables.
Usually quantifies the strength of the linear relationship between the variables (from -1 to +1).
Positive correlation – value of one variable increases with the increase in the second variable.
Negative Correlation – value of one variable decreases with the increase in the second variable
Higher association between two variables (positive or negative correlation) indicates a better predictive capability.