COMP1814 (2020/21)	**Statistical Techniques with R**	**Faculty Header ID: NA**	**Contribution: 50% of course**

Course Leader: Dr K Skindilias	**COMP1814 Coursework**	**Deadline Date:** 29 March 2021

This coursework should take an average student who is up-to-date with tutorial work approximately 25 hours Feedback and grades are normally made available within 15 working days of the coursework deadline

Learning Outcomes:

1. Identify challenges in data analytics; be able to critically evaluate and select appropriate solutions.
2. Demonstrate an understanding of the core methods and algorithms used in data analytic.
3. Analyse and manipulate data sets to extract statistics and features and provide analytic insights.
4. Critically evaluate, select and employ appropriate tools, technologies and data models to provide answers to analytic questions.

Plagiarism is presenting somebody else's work as your own. It includes: copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as your own work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by the University. Please see your student handbook for further details of what is / isn't plagiarism.

All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.

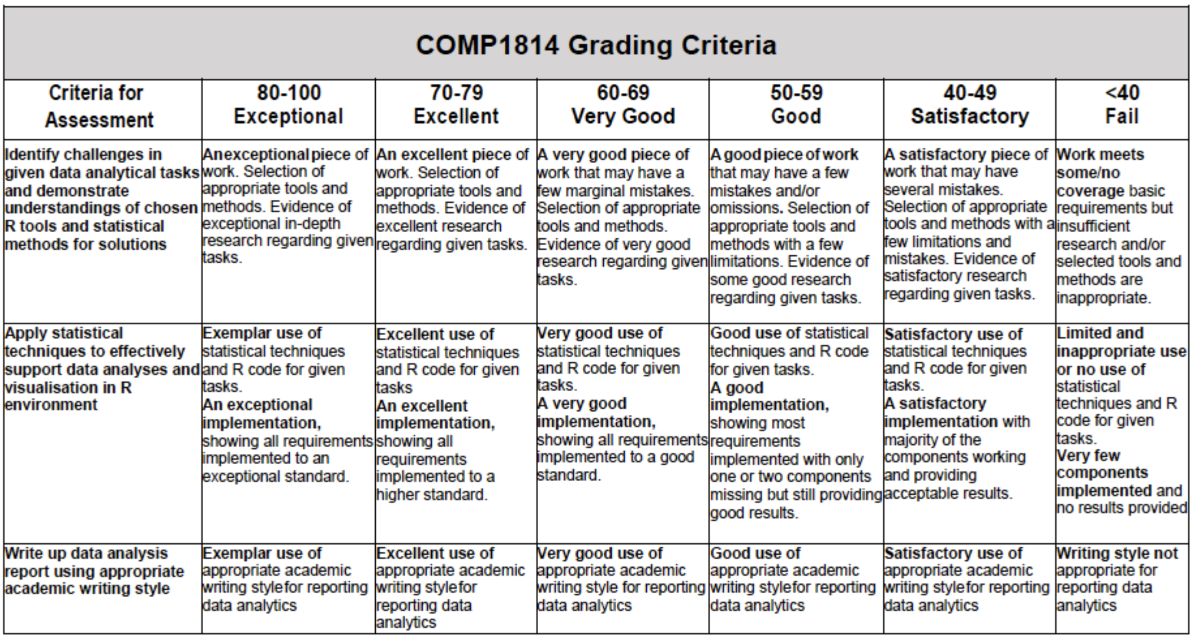
Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.

Coursework Submission Requirements

- An electronic copy of your work for this coursework must be fully uploaded on the Deadline Date of 29 March 2021 using the link on the coursework Moodle page for COMP1814.
- For this coursework you must **submit a single PDF document**. In general, any text in the document must not be an image (i.e. must not be scanned) and would normally be generated from other documents (e.g. MS Office using "Save As .. PDF"). An exception to this is hand written mathematical notation, but when scanning do ensure the file size is not excessive.
- There are limits on the file size (see the relevant course Moodle page). • _Make sure that any files you upload are virus-free and not protected by a password or corrupted otherwise they will be treated as null submissions.

- Your work will not be printed in colour. Please ensure that any pages with colour are acceptable when printed in Black and White.
- You must NOT submit a paper copy of this coursework.
- All coursework must be submitted as above. Under no circumstances can they be accepted by academic staff

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances, and penalties for Assessment Offences. See http://www2.gre.ac.uk/current-students/regs



COMP1814 Grading Criteria

Criteria for Assessment	80-100 Exceptional	70-79 Excellent	60-69 Very Good	50-59 Good	40-49 Satisfactory	<40 Fail
Identify challenges in given data analytical tasks and demonstrate understandings of chosen R tools and statistical methods for solutions	**Anexceptional piece of** work. Selection of appropriate tools and methods. Evidence of exceptional in-depth research regarding given tasks.	**An excellent piece of** work. Selection of appropriate tools and methods. Evidence of excellent research regarding given tasks.	**A very good piece of** work that may have a few marginal mistakes. Selection of appropriate tools and methods. Evidence of very good research regarding given tasks.	**A good piece of work** that may have a few mistakes and/or omissions. Selection of appropriate tools and methods with a few limitations. Evidence of some good research regarding given tasks.	**A satisfactory piece of** work that may have several mistakes. Selection of appropriate tools and methods with a few limitations and mistakes. Evidence of satisfactory research regarding given tasks.	**Work meets some/no coverage** basic requirements but insufficient research and/or selected tools and methods are inappropriate.
Apply statistical techniques to effectively support data analyses and visualisation in R environment	**Exemplar use of** statistical techniques and R code for given tasks. **An exceptional implementation,** showing all requirements implemented to an exceptional standard.	**Excellent use of** statistical techniques and R code for given tasks **An excellent implementation,** showing all requirements implemented to a higher standard.	**Very good use of** statistical techniques and R code for given tasks. **A very good implementation,** showing all requirements implemented to a good standard.	**Good use of** statistical techniques and R code for given tasks. **A good implementation,** showing most requirements implemented with only one or two components missing but still providing good results.	**Satisfactory use of** statistical techniques and R code for given tasks. **A satisfactory implementation** with majority of the components working and providing acceptable results.	**Limited and inappropriate use or no use of** statistical techniques and R code for given tasks. **Very few components implemented** and no results provided
Write up data analysis report using appropriate academic writing style	**Exemplar use of** appropriate academic writing stylefor reporting data analytics	**Excellent use of** appropriate academic writing stylefor reporting data analytics	**Very good use of** appropriate academic writing style for reporting data analytics	**Good use of** appropriate academic writing stylefor reporting data analytics	**Satisfactory use of** appropriate academic writing stylefor reporting data analytics	**Writing style not** appropriate for reporting data analytics

Detailed Coursework specification: See next page

Hypothesis testing in R.
Comparing means and fitting distributions

200 mice received a treatment “Nutritional Supplement” during 6 months. We want to know whether the treatment has an impact on the weight of the mice.

Another 200 rats received the same treatment but results seem to differ.

To answer to this question, the weight of the mice has been measured before and after the treatment. This gives us 200 sets of values before treatment and 200 sets of values after treatment from measuring twice the weight of the **same mice**, and another 200 sets of values before treatment and 200 sets of values after treatment from measuring twice the weight of the **same rats.**

For this test you are required to create two datasets for each set of mice and rats. The data will be created from artificial data.

Create the following:

Task 1: Data Generation

a. The weights of **mice** as “before” and “after” the treatment, coming from a normal distribution with mean = 20, and variance = 2.
   For the “after” treatment add to the mean 1 unit and to the variance 0.5 units, that is mean = 21 and variance = 2.5. [5 marks]
b. The weights of **rats** as “before” and “after” the treatment, coming from a Weibull distribution with shape = 10, and scale = 20.
   For the “after” treatment remove from the shape 1 unit and add to the scale 1 unit, that is shape = 9 and scale = 21. [5 marks]
c. Using the function ‘qplot’ with ‘geom = density’ (from your lecture notes) compare for each of your data sets mice(before, after) and rats(before, after) [10 marks]
d. Perform the same operation using ‘geom = boxplot’. [5 marks]

Task 2: Appropriateness for Hypothesis t-testing

a. For your **mice** data set (combined “before” + “after”) examine whether the data passes normality qualitatively (QQ plot) and quantitatively (Shapiro-Wilk test). [5 marks]
b. For your **rats** data set (combined “before” + “after”) examine whether the data passes normality qualitatively (QQ plot) and quantitatively (Shapiro-Wilk test). [5 marks]
c. Explain the output of your analysis for each of the dataset and discuss the appropriate test to test your hypothesis. [15 marks]

Task 3: Hypothesis testing

a. For the normal data set (**mice**) examine perform a paired t-test and explain your findings. You need to extract and comment on all output of the t-test:
   a. T-test statistic
   b. Degrees of freedom
   c. P-value
   d. Confidence Interval
   e. Sample estimates

[20 marks]

b. For the **rats** dataset perform a non-parametric -test and comment on your finding.
[10 marks]

Task 4: Fitting distributions

a. Lastly, for the **rats** datasets use the function 'fitdist' (from the 'fitdistrplus' R package) and examine the best-fit distribution (even if we know the TRUE distribution). Fit a Weibull, a lognormal and a Gamma distributions and discuss your findings making use of the package comparison tool for a Density, CDF, QQ, and PP. [20 marks]

--------- End of Coursework --