Department of Computer Science

CS5801 Quantitative Data Analysis
Assessment/Coursework for 2020/21

TABLE OF CONTENTS

Assessment Title	Quantitative Data Analysis
Module Leader	Isabel Sassoon, Martin Shepperd
Distribution Date	4/11/2020
Submission Deadline	11:00, Monday 11th of January 2021
Feedback by	8/02/2021
Contribution to overall assessment block	100 %
Indicative student time working on assessment	Hours
Word or Page Limit (if applicable)	n.a.
Assessment Type (individual or group)	Individual Coursework

MAIN OBJECTIVE OF THE ASSESSMENT

The coursework is based on undertaking an analysis of a real world data set[1]. It is a shared assessment block for CS5701 Quantitative Data Analysis and CS5702 Modern Data.

This assessment offers an opportunity to bring together your skills from CS5701 (Quantitative Data Analysis) and CS5702 (Modern Data). Note that there is also a second shared assessment block CS5802 which takes the form of a written examination.

DESCRIPTION OF THE ASSESSMENT

The requirements for the assessment block are as follows:

1. You will be provided with
   (i) a data set and its metadata, (note students will receive different data subsets but in the same data format)
   (ii) research questions to guide your analysis
   (ii) a pro forma .Rmd file to use to complete your coursework.

2. Using the pro forma please address the following:

   1. Organise and clean the data
      1.1 subset the data into the specific data subset allocated
      1.2 data quality analysis
      1.3 data cleaning
   2. Exploratory data analysis
   3. Modelling

[1] We have made some small modifications so please use our provided version from GitHub; do *not* use any original source version.

3.1 build a model for player salary
3.2 critique your model using relevant diagnostics
3.3 based on 3.2 suggest improvements to your model
4. Extension work:
4.1 Plan and build a model for the variable Hit indicator (hit.ind)

3. Generating your personal data sets

1. Each student should use data based on two baseball teams subsetted from the overall dataset baseball-2015.RDa which can be accessed from GitHub (https://raw.githubusercontent.com/mjshepperd/CS5702-Data/master/CS5801_data.rda).
2. The choice of teams depends on your student id.
3. To determine your particular two teams, extract the two least significant digits from your student id, e.g., if your id is 2012345 then the relevant two digits are '4' and '5'.
4. If your two digits are equal e.g., 2012344 then increment by one so in this example they would become '4' and '5'. If the digits are 2012399 then wrap around so they would become '9' and '0'.
5. Use the following table to map the digit to your team. Note that teams may have differing numbers of observations/cases/rows.
6. The code for subsetting is embedded in the RMarkdown template. You need to configure it for your particular teams.
7. If you are uncertain, please check!

Digit	teamID.x
0	BOS
1	TEX
2	LAN
3	NYN
4	CHN
5	PIT
6	LAA
7	TOR
8	HOU
9	KCA

General Guidance:

1. You are expected to use R and RMarkdown for your analysis.
2. Use the template RMarkdown as a starting point to structure your report but remember to remove our scaffolding and guidance comments before you submit. It is available from GitHub
3. Update the YAML to include your name and other identifier information.
4. Include all relevant R code chunks and provide explanation, comments and discussion as appropriate.
5. Follow the principles of 'literate programming' so choose meaningful variable and function names and add comments.
6. You can also submit supplementary files if you wish, but you *must* include a single report file that contains your entire report in .Rmd format. Make sure any supplementary files are cross-referenced from your main report.
7. Where appropriate cite external sources and add a bibliography at the end of your main report.
8. The report should be professionally presented with good structure, an absence of spelling errors and other typos and written in an appropriate style (i.e., simple to the point, unemotive language).
9. Whilst there are no formal page or word limits, we discourage excessive padding, so unnecessary words and waffle will militate against professional presentation.
10. Don't worry, sometimes even suitable models do not have good fit due to the nature of the data. In such circumstances you will not be penalised.
11. Whilst we encourage collaboration and sharing of ideas this is an individual report and so must be based on your own understanding, analysis and words. WiseFlow automatically cross-compares all submissions.

12. WiseFlow also has a plagiarism detector for external sources. We encourage you to use such sources including R packages, code, ideas for data analysis and other statistical sources, but you must acknowledge the sources. In other words, do *not* attempt to pass off the work of others as your own.
13. If you have questions, please check the assessment FAQs or ask one of us. Don't guess!

Learning Outcomes and Marking Criteria

LO1: Design and implement methods and protocols for data preparation and exploration using advanced statistical techniques.
LO2: Apply these methods on real data to generate novel insight, critically evaluate its value and design a framework for data management and sharing.

Below we give the detailed marking scheme.

Task	Total marks available	Marking criteria (for full marks)
0. Understandability of the analysis (10)		
0.1 Quality of report	5	Clearly and professionally presented. Appropriate use of cited external sources.
0.2 Quality of code including comments, clear layout and structure, meaningful identifiers.	5	Easy to understand, well documented code following principles of literate programming.
1. Organise and clean the data (25)		
1.1 Subset the data into the specific dataset allocated	5	Uses R code to correctly select the subset of data allocated.
1.2 Data quality analysis	10	Provides a description of a comprehensive plan to assess the quality of the data, and documents its findings. Includes all columns/variables (5), and full implementation (5).
1.3 Data cleaning	10	Explains data quality issues found in 1.2, (5) justifies and documents the responses made (if any) (5). NB even if no data quality issues are identified you should still check and report.
2. Exploratory Data Analysis (20)		
2.1 EDA plan	5	Outlines a suitable plan to explore the data.
2.2 EDA and summary of results	10	Undertakes and summarises the findings of the data exploration, particularly with respect to the research questions. Use appropriate summary statistics (univariate and multivariate) and visualisations.
2.3 Additional insights and issues	5	Highlights potential further issues or insights uncovered in 2.2.
3. Modelling (25)		

3.1 Build a model for player salary	10	Given the research question (i.e. target attribute) outlines an analysis plan that incorporates/references any findings from the data cleaning (1.3) and EDA (2.2) (5). Uses R to build a suitable model (10).
3.2 Critique model using relevant diagnostics	10	Offers an interpretation of the model characteristics, goodness of fit and graphical diagnostics (5) for the model built in 3.1. Explains any potential weaknesses (5). (When multiple models are considered a concise explanation for the selection of one model to present in answer to 3.1.)
3.3 Suggest improvements to your model	5	Based on the findings in 3.2 articulates possible alternative approaches to address them (5).
4. Extension work	20	
4.1 Plan and build a model for the variable Hit indicator (hit.ind)	20	Given the second research question (i.e. ,involving the binary target attribute) a plan of analysis is outlined based on relevant EDA on this attribute (10). A model is included, explained, critiqued and documented (10).

FORMAT OF THE ASSESSMENT

[provide guidelines on expected format and length of submission]

SUBMISSION INSTRUCTIONS

You must submit your coursework as an RMarkdown file on WiseFlow by 11th of January 2021 at 11am. You can follow the link to Wiseflow through the module's section on Blackboard Learn or login in directly at https://uk.wiseflow.net/brunel. The name of your file should follow the normal convention set out in the student handbook and must therefore include your student ID number (e.g., 0612345.Rmd). It can also include the module code (e.g., CS2001_0612345.Rmd). You may, if you wish, submit ancillary files (e.g., pdf) as Appendices, however you must cross reference them from your main file and make it clear what is their purpose.

Your main submission *must* follow the pro forma we have provided.

AVOIDING PLAGIARISM

Please ensure that you understand the meaning of plagiarism and the seriousness of the offence. Information on plagiarism can be found on the College's Student Handbook.

LATE COURSEWORK

The clear expectation is that you will submit your coursework by the submission deadline stated in the study guide. In line with the University's policy on the late submission of coursework (revised in July 2016), coursework submitted up to 48 hours late will be accepted, but capped at a threshold pass (D- for undergraduate or C- for postgraduate). Work submitted over 48 hours after the stated deadline will automatically be given a fail grade (F).

Please refer to the Computer Science Student Handbook, available on Blackboard Learn, for information on submitting late work, penalties applied and procedures in the case of mitigating circumstances.

Appendix 1 : The metadata

The baseball-2015.RDa data set includes the following attributes:

Column Name	Column Description
playerID	Player ID code
teamID.x	Team ID
G	Games: number of games in which a player played
R	Runs
H	Hits: times reached base because of a batted, fair ball without error by the defense
RBI	Runs Batted In
weight	Player's weight in pounds
height	Player's height in inches
salary	The salary of the player
birthdate	The date of birth of the player
career.length	The career length of the player in years
bats	Whether they bat with their Left (L) or Right (R) hand
age	The age of the player
hit.ind	This will be 1 if the player has made at least one hit in the 2015 season and 0 if they have not.

The data has one row per player, and some players can play for more than one team in a season. We are focusing on the data from the 2015 season.