MAT022 Foundations of Statistics and Data Science

Summative Assessment 2020/21

Summative assessment for the module is by means of a single **report** on your statistical analysis of data related to the **National Basketball Association** (NBA), a men's professional basketball league in the USA.

This form of assessment has been chosen because as professional statisticians and data scientists, you will often be asked to investigate a data set and report on whether it contains anything useful or interesting. The assessment will also help you to prepare for writing your MSc dissertation in the summer.

Assessment type	*Weight*	*Max. length*	*Format*	*Deadline*
Report	100%	10 pages	R Markdown and PDF	**Tue 09 Feb 2021**

Your report will be assessed according to how well you are able to

- **analyse** the data set, **40%**
- **interpret** the results of your analysis, and **30%**
- **present** the results of your analysis and your interpretation of the data set. **30%**

Your analysis should be performed using the `R` statistical software package, and your report prepared using the `R Markdown` typesetting system and the template provided.

1 The data

You are asked to write a report on data from the 2014–15 season of the **National Basketball Association** (NBA), a men's professional basketball league in the USA. The data set is a record of all shots taken by players in the NBA between October 2014 and March 2015, and consists of 128,069 observations on 23 variables as described in Table 1.

Please submit your report in PDF format, together with the `R Markdown` file used to generate the report, through Learning Central sometime before **12.00** on **Tuesday 09 February 2021**.

Variable	*Description*
GAME_ID	Unique id number of the game.
DATE	Date of the game.
HOME_TEAM	Team playing at home.
AWAY_TEAM	Team playing away from home.
PLAYER_NAME	Name of the shooting player.
PLAYER_ID	Unique id number of the shooting player.
LOCATION	Whether the player was on the home (H) or away (A) team.
W	Whether the player's team won (W) or lost (L) the game.
FINAL_MARGIN	The margin of victory for the player's team (negative means defeat).
SHOT_NUMBER	The number of the shot taken by the shooting player in that game.
PERIOD	The period of the game that the shot was taken.
GAME_CLOCK	The time remaining in the period when the shot was taken.
SHOT_CLOCK	The time remaining on the shot clock.
DRIBBLES	Number of dribbles by the player before the shot was taken.
TOUCH_TIME	The time that the ball was in the shooting player's hand.
SHOT_DIST	The distance of the shooting player from the basket.
PTS_TYPE	2 for shots from inside the arc, 3 for shots from outside the arc.
SHOT RESULT	Whether the shot was successful ('made') or unsuccessful ('missed')
CLOSEST_DEFENDER	Name of the nearest defender when the shot was taken.
CLOSEST_DEFENDER_ID	Unique id number of the nearest defender.
CLOSE_DEF_DIST	Distance of the nearest defender when the shot was taken.
FGM	Equal to 1 if the shot was made (scored) otherwise 0.
PTS	The number of points scored (0, 2 or 3)

Table 1: Description of the variables in the NBA shot logs data set.

ATL	*Atlanta Hawks*	MIA	*Miami Heat*
BKN	*Brooklyn Nets*	MIL	*Milwaukee Bucks*
BOS	*Boston Celts*	MIN	*Minnesota Timberwolves*
CHA	*Charlotte Hornets*	NOP	*New Orleans Pelicans*
CHI	*Chicago Bulls*	NYK	*New York Knicks*
CLE	*Cleveland Cavaliers*	OKC	*Oklahoma City Thunder*
DAL	*Dallas Mavericks*	ORL	*Orlando Magic*
DEN	*Denver Nuggets*	PHI	*Philadelphia 76ers*
DET	*Detroit Pistons*	PHX	*Phoenix Suns*
GSW	*Golden State Warriors*	POR	*Portland Trail Blazers*
HOU	*Houston Rockets*	SAC	*Sacramento Kings*
IND	*Indiana Pacers*	SAS	*San Antonio Spurs*
LAC	*Los Angeles Clippers*	TOR	*Toronto Raptors*
LAL	*Los Angeles Lakers*	UTA	*Utah Jazz*
MEM	*Memphis Grizzlies*	WAS	*Washington Wizards*

Table 2: Acronyms for the teams in the NBA.

2 The report

The ability to write clearly and concisely is an important professional competence. To encourage writing that is brief and to the point, your reports are limited to a **maximum of 10 pages**. It is often far more difficult to express yourself in 100 words than in 1000 words, especially when you have a lot to say, so be careful not to underestimate the challenge posed by this restriction. The modest page limit will also encourage you to be selective in the results you choose to present.

A suggested structure for your report is shown in Table 3. Note that the title page, abstract, table of contents, list of references and appendix do not contribute towards the page count.

Title	1 page
Abstract	100 words
Table of contents	–
1. Introduction	1/2 page
2. Background	1 page
3. (*descriptive analysis*)	2 pages
4. (*inferential analysis*)	2 – 3 pages
5. (*inferential analysis*)	2 – 3 pages
6. (*inferential analysis*)	2 – 3 pages
7. Conclusion	1/2 page
References	–
Appendices	2 pages max.

Table 3: Suggested report structure

- The **title page** should contain the title of your report, your name and student number, and the date on which your report was completed.
- The **abstract** should contain a short summary of the report and its main conclusions.
- The **table of contents** should list the number and title of each section against the number of the page on which the section begins.
- The **introduction** should consist of a few short paragraphs, describing the purpose of the report and providing a brief outline of its contents.
- The **background** section should include a brief review of any relevant literature, and provide a context for the work presented in the report.
- The report should contain a relatively short section on a **descriptive analysis** of the data set, with a title chosen to reflect what the section contains.
- The main part of the report should consist of two or three sections on different **inferential analyses** of the data set. Here you should formulate hypotheses, conduct statistical tests, then present and discuss the results of these tests. The titles of these sections should reflect what the sections contain.
- The **conclusion** should consist of a few short paragraphs, providing a summary of the report and a brief outline of some ideas for future work.
- The report may contain a single **appendix** for large figures and tables, limited to a maximum of two pages.

3 Assessment criteria

Detailed assessment criteria are shown in Table 4.

Level	**Analysis** (40%)	**Discussion** (30%)	**Presentation** (30%)
Distinction (70–100)	Hypotheses are interesting and original. Methods are appropriate and applied carefully and precisely. An interesting descriptive analysis is included and reported correctly.	Inferences are valid and supported by evidence. Original and interesting conclusions are articulated. There is some shrewd speculation about possible causal factors.	A high standard of writing is maintained throughout. The narrative is clear, coherent, eloquent and refined. Figures and tables are used creatively.
Merit (60–69)	Hypotheses are formulated correctly. Methods are appropriate and applied correctly. A moderately interesting descriptive analysis is included and reported correctly.	Inferences are valid and supported by evidence. Interesting conclusions are articulated. There is some speculation about possible causal factors.	A good standard or writing is maintained throughout. The narrative is clear and coherent. Figures and tables are used to illustrate the narrative.
Pass (50–59)	Hypotheses are formulated correctly. Methods are applied correctly for the most part. A descriptive analysis is included and reported correctly.	Inferences are mostly valid and supported by some evidence. Some relatively interesting conclusions are articulated.	An acceptable standard of writing is maintained throughout. The narrative is lacklusture and sometimes unclear. Figures and tables do not always illustrate the narrative.
Fail (0–49)	The analysis is bland and almost entirely descriptive.	Inferences are invalid or not supported by evidence. There is little of any interest.	The report is poorly written. The narrative is disjointed and hard to follow.

Table 4: Assessment criteria

Plagiarism

You may find existing studies of the NBA data set online. Plagiarism is to present other people's work or ideas as your own, by incorporating it into your work without full acknowledgement. The need to acknowledge others' work applies not only to text, but also to computer code, figures, tables etc. You must also attribute text, data, or other resources downloaded from websites. Following submission your report will be analysed by the *TurnitIn* software, and any report in which plagiarism is detected will receive a mark of zero.

4 Guidelines for writing reports

The golden rule when writing is to always **think of the reader**. For scientific reports, readers will typically want to read something interesting and learn something in the process.

What do we mean by interesting?	
Not interesting	The average exam mark of statistics and data science students.
Quite interesting	The average mark of male students, the average mark of female students, and the results of a test of whether any difference is statistically significant.
Very interesting	The average mark of male students, the average mark of female students, a statistical test of whether any difference is significant, and some speculation about why there is a significant difference, or alternatively why there is not.

Audience. The target audience for your report is this year's cohort students on the *Foundations of Statistics and Data Science* module, so you can assume that your readers are familiar with the methods and terminology established within the lectures and notebooks. If you choose to use methods that have not been covered in lectures, you must ensure that any new terms are properly defined and references to the relevant literature included.

Analysis. The reader shoud be satisfied that you have performed your analysis correctly, and in particular that you have verified the conditions that are necessary to apply the various methods. Your methods should be introduced with a brief summary of their main features, but technical details should not be discussed at length although you might consider providing the interested reader with references to the relevant literature.

Navigation. Do not assume that the reader will read the report from start to finish, as one might read a novel. Reports should be made easy to navigate using numbered sections and subsections together with cross-referencing. Once you have written a first draft, it will need careful editing before it becomes a coherent and polished report. This stage always takes longer than you think!

Scientific writing. For scientific reports we aim for a style of writing that is *clear* and *concise*. Make sure that sentences are unambiguous and that a good standard of writing is maintained throughout the report.

- Sections should not start abruptly with the subject matter, but rather with an introductory sentence or short paragraph. Sections should also end with concluding sentence or short paragraph.
- *All* figures and tables must be numbered and have captions. Figures or tables that are not mentioned at least once in the text should not be included.
- A *qualified statement* is one that express some level of uncertainty about its own accuracy, and should always be used when drawing conclusions from the results of a statistical analysis, and especially when speculating about possible causal factors. Common phrases that indicate qualified statements include "*This suggests that ...*", "*It appears that ...*", "*We might conclude that ...*", "*There is some evidence to indicate ...*" and so on.