Overview

This assessment task allows you to consolidate and apply the concepts and skills you've learnt throughout the semester. This assessment requires you to develop a **reproducible data analysis project.**

Your reproducible data analysis project will be hosted as a repository on GitHub and you are required to submit the URL to your GitHub repository.

Scenario and aim

You are a data analyst with the Chicago Bulls competing in the NBA (National Basketball Association). In the most recent NBA season (pretend last season was 2018-19), your team placed 27th out of 30 (based on win-loss record). Your team's budget for player contracts next season is $118 million, ranked 26th out of 30 (for the purpose of this assignment, next season is 2019-20). For context, the team with the highest payroll budget is Portland with $148 million, while the best performing team was Milwaukee Bucks (who clinched the best league record in 2018-19) with $131 million.

You have been tasked by the general manager of Chicago Bulls to find the best five starting players (one from each position) your team can afford. (Make sure you don't use up all of your money on just these five players, you still need to fill a full team roster, but are just focused on finding five starting players here). You can choose players that were already playing for Chicago Bulls in 2018-19, you just need to prove that they are worth it.

To complete the task assigned to you by the general manager, you will need to develop a reproducible data analysis project and generate a *knitr* report using R Markdown describing your analysis and findings.

The data

To help you complete this task, you have been provided with the following data sets:

1. `2018-19_nba_player-statistics.csv` : sourced from basketball-reference.com
2. `2018-19_nba_player-salaries.csv` : sourced from hoopshype.com/salaries
3. `2019-20_nba_team-payroll.csv` : sourced from hoopshype.com/salaries
4. `2018-19_nba_team-statistics_1.csv` : sourced from basketball-reference.com
5. `2018-19_nba_team-statistics_2.csv` : sourced from basketball-reference.com

Check the downloaded zip file `project-data.zip` to access these data sets.

Note: You can use some or all of these data sets.

You can find a description of each of the variables in these data sets (Check Variable Description file attached).

What you need to do

To complete the task assigned to you by the general manager, you will need to develop a reproducible data analysis project and generate a *knitr* report using R Markdown describing your analysis and findings. You will need to host your reproducible data analysis project on GitHub in its own repository.

Your GitHub repository must include the following parts:

- a `README.md` file adequately describing the GitHub repository and it's contents. You may add in variable descriptions for each dataset in the README.md file, or include a separate appropriately named document which describes there
- a `data` folder containing the raw data and any tidy/processed data files (formatted appropriately in separate `raw` and `processed` directories)
- a knitr report (e.g. as a html file) detailing your analysis and findings (see below for more details)
- a raw R Markdown file (.rmd) used to generate the knitr report

You GitHub repository may also include the following parts if relevant:

- an `R` or `funcs` folder containing .R scripts used to create your own functions (or other code sourced in the R Markdown file)
- an RStudio project file (.Rproj)
- a `figs` folder containing any figures that were generated by the R Markdown file
- a 'images' folder containing any images that are sourced by the R Markdown file
- a Markdown file (.md) generated from the R Markdown file so the rendered report can be viewed in GitHub

The Knitr Report

Your knitr report should demonstrate best practice **literate programming** and show the code used and subsequent output generated from your analysis, as well as

written descriptions and interpretations of each step. The report should include the following sections:

1. Introduction

This section should provide relevant background information and justification for the project, including:

a) relevant background information of basketball, including key metrics, position requirements etc

b) description of the scenario

c) the aim of the project

d) justification and importance

Note that you may choose a different order to present each of the elements listed above.

2. Reading and cleaning the raw data

This section should document the process used to read and clean the raw data. It should also include a description of the data sets used and variables in each. For brevity, you could provide a link to the specific variable descriptions, rather than writing these out in full within your report.

3. Exploratory analysis

This section should document your exploratory data analysis and may include but is not limited to:

a) checking for errors and missing values within the datasets

b) checking the distribution of variables

c) checking for relationships between variables, or differences between groups

d) justification for decisions made about data modelling

Note that this section and the data cleaning section may be an iterative process, as you might find things about the data that need to be 'cleaned up' once you have explored the data further.

4. Data modelling and results

This section may include but is not limited to:

a) data modelling (e.g. creating a linear regression)

b) assumption checking

c) model output and interpretation of your model

5. Player recommendations

This section will be the key part that is presented to the general manager. Here you should present your recommendations for the best five starting players, but also think about what other important information they would want to know, and how it is best to present that information to them.

6. Summary

Provide a brief summary which describes the key points and findings from your project. It will also be important to acknowledge any limitations of your model and overall approach to answering the question asked of you by the general manager.

7. Reference List

Provide a reference list of any sources you used in the development of your report and justification of your arguments. Please use the Vancouver reference style (Links to an external site.) for the reference list and in-text references.

Submission

You are required to submit the URL to your GitHub repository.

Tips and resources

- It is highly recommended you have completed the Moneyball case studies (part 1 and part 2) to help you with this project.
- Players that were traded during the season may appear more than once (on more than one row) in the `2018-19_nba_player-statistics.csv` data set, so it is important to handle these duplicates appropriately. These players will also have a row where the team is TOT. This row is their "total".
- Think about: how will you determine which metrics are more important to each position?
- You can find information about each position in basketball here.

- Refer to the marking rubric for a guide on how this assignment will be assessed