Homework # 7

MATH 110

In this homework, we will work with a baseball player dataset composed of the files Master.csv, Salaries.csv, Pitching.csv, Inflation.csv.

The file contents are as follows:

- Master.csv – contains information (such as date of birth, place of birth, height, weight, batting hand, throwing hand, full name, etc.) about each player, each player is given a unique playerID that is used to reference them in other data files; each row in this data file represents a single player and each player appears exactly once in this data file. This data.frame is included just for orientation, you don't actually need it to do the homework.

- Pitching.csv – contains information on the pitching records of individual players during individual season; each row in this data file represents a player in a particular season and each player may appear several times if they played during multiple seasons (if a player switched teams during a season, that player will have multiple rows for that season, a separate record for their performance with each team). Pitchers are a particular position in baseball, so not all players are pitchers. We will concentrate on the statistic ERA, provided for each pitcher. ERA can be used as a measure of how good a particular pitcher is, with a lower ERA being better.

- Salaries.csv – contains salary information; players will appear multiple times if they played in multiple seasons. **Salaries are only provided for the years 1985-2014**.

- Inflation.csv - contains two variables, year and inflation2015. The second variable gives the purchasing power of one dollar in year (for year between 1980 and 2015) in terms of 2015 dollars. This data was generated using the Bureau of Labor Statistics' Consumer Price Index Inflation Calculator.

In this analysis you will consider player salaries and ERA over time. Restrict your attention to the years 1985-2014, since there is no data for previous years. Salaries should be adjusted to 2015 dollars (i.e. map each salary to an adjusted salary). Since ERAs over all players can vary from season to season for various reasons, map each players ERA to a quantile for the particular year considered. (For example, an ERA of 3.00 might be in the .7 quantile for the year 1990, meaning that 70% of pitchers had an ERA less than 3.00). You may find the `order` function useful.

After completing the mappings above, you will have data samples that include a year, an adjusted salary, and an adjusted ERA. Each such sample corresponds to a particular pitcher in a particular year. (Hint: pre-process the data to form a data.frame containing these three variables over all samples.) You goal is now to investigate the relationship between these three variables using ggplot. To start, consider how salaries have changed over time, without considering ERA. Then consider how salary depends on ERA in a particular year. Then go further and try to consider all variables at once. Be sure to split your workflow into different directories, as we have done before, and create your analysis using Rmarkdown.