**Problem3**

In this analysis, I used multiple datasets from the Great British Bake Off, including bakers.csv, bakers.comsv, and results.comsv, and cleaned, organized, and merged these data to generate a complete and easy to analyze dataset.

Firstly, by reading each dataset, we cleaned the data. For example, some rows in the dataset do not contain specific information related to baking competitions (such as blank lines or comments), so we filtered them. In addition, we standardized the column names in the dataset to make them more concise and easy to read (for example, renaming name to baker name). During the cleaning process, we also rounded and converted the "Sports Balls" data from the competition to integers to ensure data consistency.

After cleaning each dataset, we merge the bakers and bakers datasets using the leftjoining function, and then merge them with the results dataset to form a complete dataset that includes participants, their works, and their competition results. The variables in this dataset include key information such as season, episode number, baker name, and result (competition result, such as "Star Baker"). The final dataset consists of several rows and columns, sorted appropriately to analyze the season, number of sets, and performance of participants.

By analyzing this dataset, we extracted Star Bakers from seasons 5 to 10. In each episode of these seasons, the best performer was selected as Star Baker, and the championship was chosen in the final episode. In these six seasons of the competition, we can see some contestants winning frequently and performing very steadily, but there are also some unexpected results, such as the final champion not winning multiple times in the early stages of the competition, demonstrating the uncertainty of the competition.

Next, we imported viewers.csv to analyze the number of viewers per season. We cleaned up the data and calculated the average audience for seasons 1 and 5. Overall, the final generated aggregated dataset provides us with information on various aspects such as the performance of participants, the winners of each episode, and the number of viewers. This provides a solid foundation for subsequent analysis, including exploring the predictability of performance and changes in program popularity.