data analysis Report

Homework 1.1

José Manuel Pérez Ricote

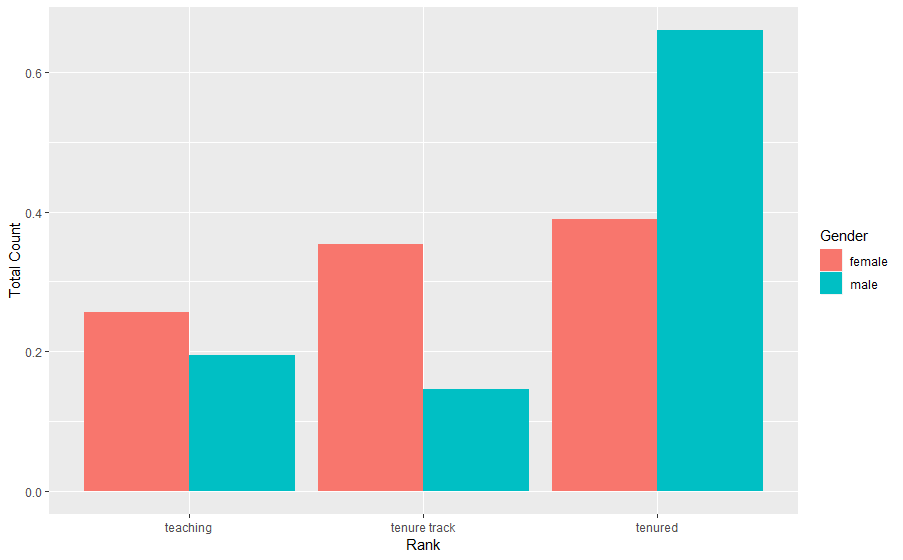
UPM (Universidad Politécnica de Madrid)

1. **RESEARCH QUESTIONS**

The dataset given (eval.R), includes different course evaluations of the professors on a given university, represented in attribute ‘eval’. There is a wide array of attributes that gives information about the different professors, each of them with a concrete evaluation. The goal of analysing this dataset is to explore the different possible relationships that there can be in data and describe them using plots and explanations.

Moreover, a group of six people, of different gender and groups inside the university, rated the beauty of the different professors. These six attributes and the beauty average of them all are also part of this dataset. Other important variables that we will use during this homework are the following:

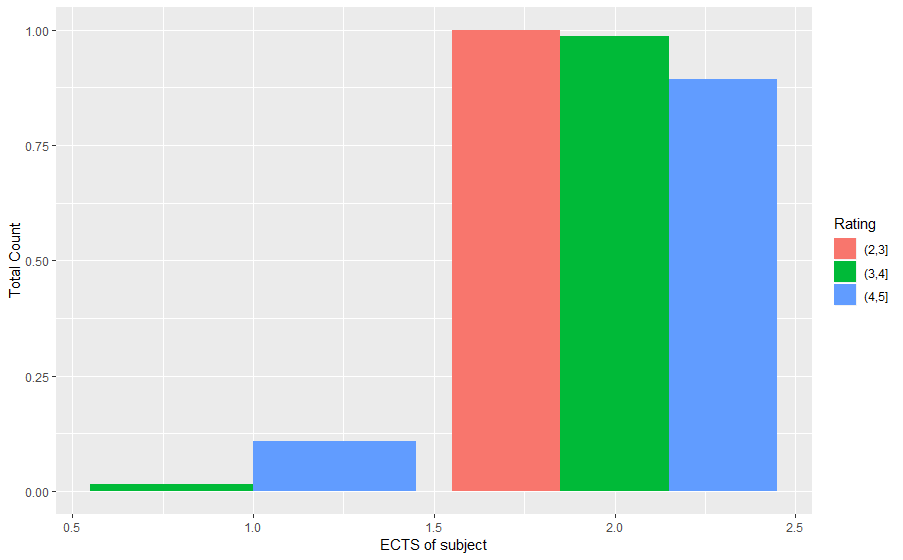
* Gender: gender of professor: female (1) and male (2).
* Rank: rank of professor: teaching (1), tenure track (2) and tenured (3).
* Cls\_credits: ECTS of the subject: one credit (1) and multi credit (2).
* Cls\_perc\_eval: Numerical variable, percentage of completed students that completed the evaluation.
* The first question that could be used to start exploring the data is if there is any evidence of independence between the rank of the professor and its gender. As both are categorical values, they must be discretized into factors with levels. Once this is done, they can be plotted using bar plots and analyse by the graphs the relationship between the variables.



In this plot, we can see that there is a clear dependency between both attributes: rank and gender. This is explained with the plot because for tenured positions male have much more opportunities of being in such interval than women. However, for less important ranks of professors, the probability of being a woman increases whereas the probability of choosing a man decreases. We can analyse this conclusion as a coherent one, because one of nowadays real-life problems is the inequality between genders. Men continue having more rights when accessing to high rewarded positions in companies, research groups, universities, political parties, and so on.

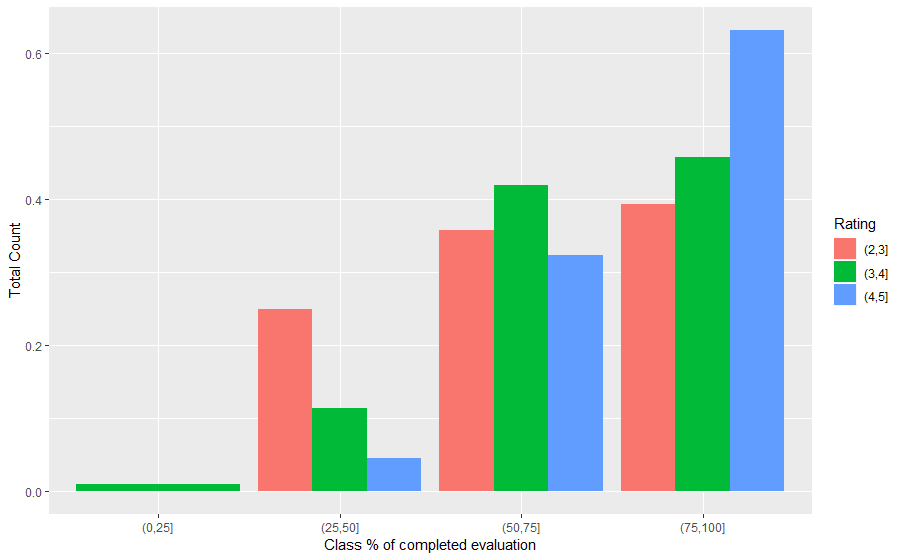
* The second research question is more open to different analysis, as it comes in a wider perspective. We should choose two different variables that can have influence on course evaluations and focus on them. To do so, we set all values to numeric ones by creating a new dataset. Afterwards, we generate a matrix of correlation for the whole dataset and detect which attributes have a higher absolute value in correlation with course evaluations attribute.

All this analysis concluded in two variables, cls\_credits (with a correlation of -0.23) and cls\_perc\_eval (with a correlation of 0.22). First one, cls\_credits, divides the subjects in two types: single credit subjects and multiple credit ones. The other expresses the percentage of students that passed evaluation. As cls\_perc\_eval and eval are both numerical attributes, they should be converted into discrete values. One this is done, we proceed to plot them.

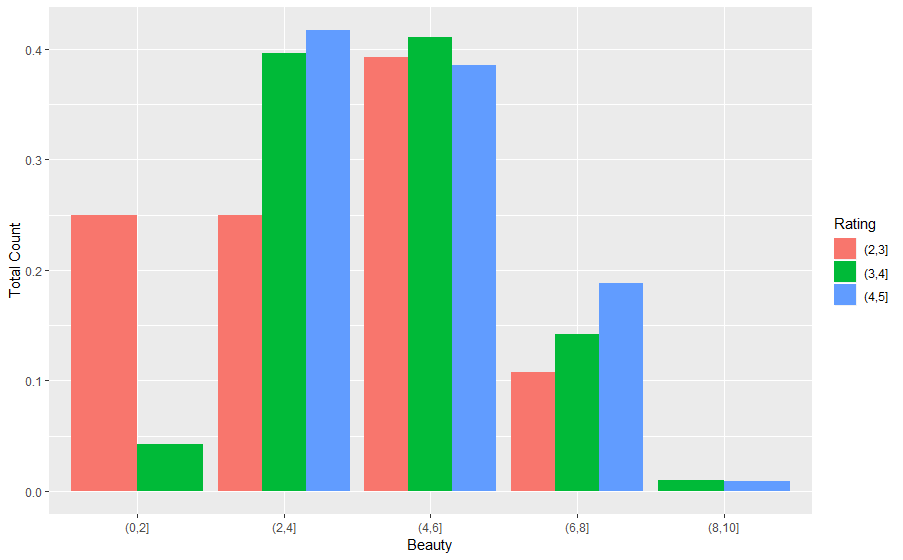


Single ECTS subjects are more likely to be better rated, explaining the negative correlation between the variables. However, as there are very few observations for single ECTS subjects, it can’t be considered as a key attribute to analyse whether the rating for the course is going to be good or not.

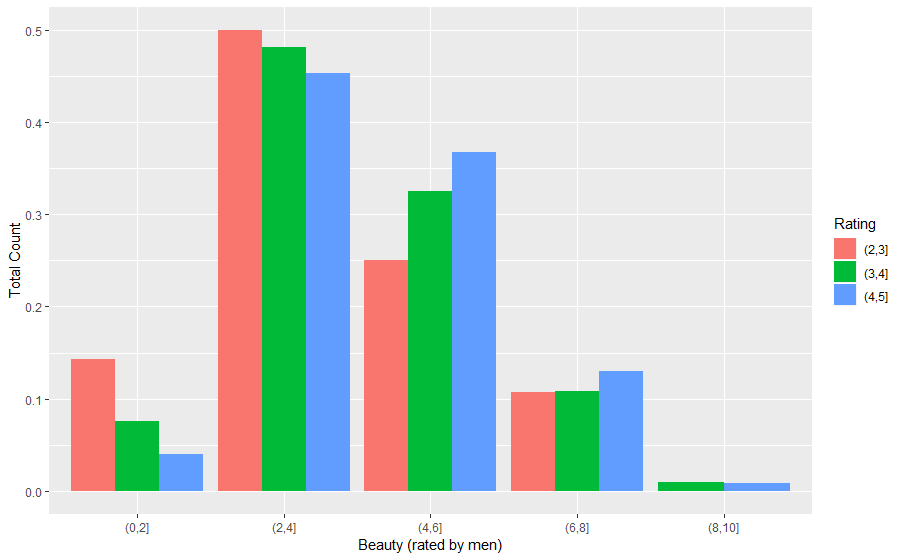
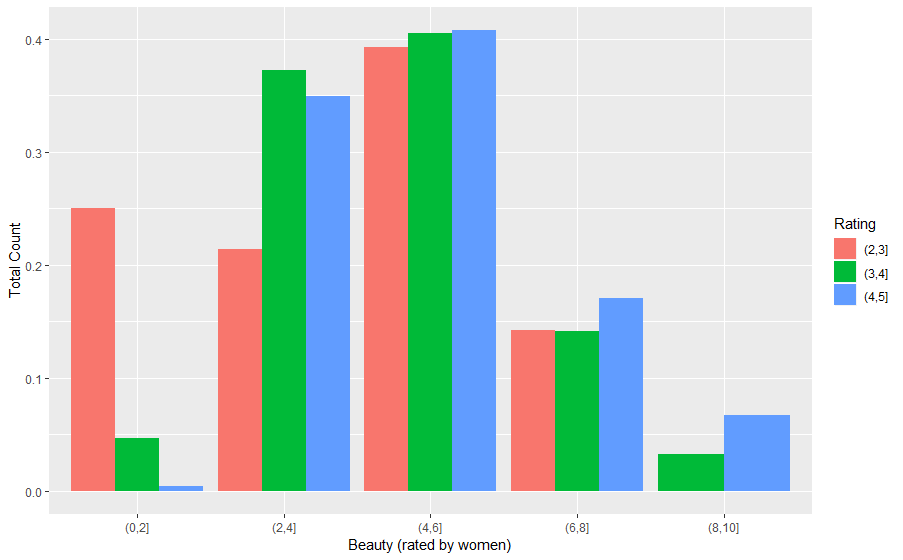
In the second plot, we can observe that correlation is positive, as better rated subjects are those with higher percentage of passed students in the evaluation. As before, it is not the best variable to predict the rating of a subject because for subjects with less than 50% of passed students there are almost no observations.



* In third place, we analyse the dependency that exists between beauty of professor’s (from the rating of six students) and course ratings. We should be aware of the six different attributes, so we divide the exploration in three parts: compare the beauty average with course evaluations, compare men’s average rating of professor’s beauty with course evaluations and compare women’s average rating of professor’s beauty with course evaluations. The plot corresponding to the first comparison is the following:

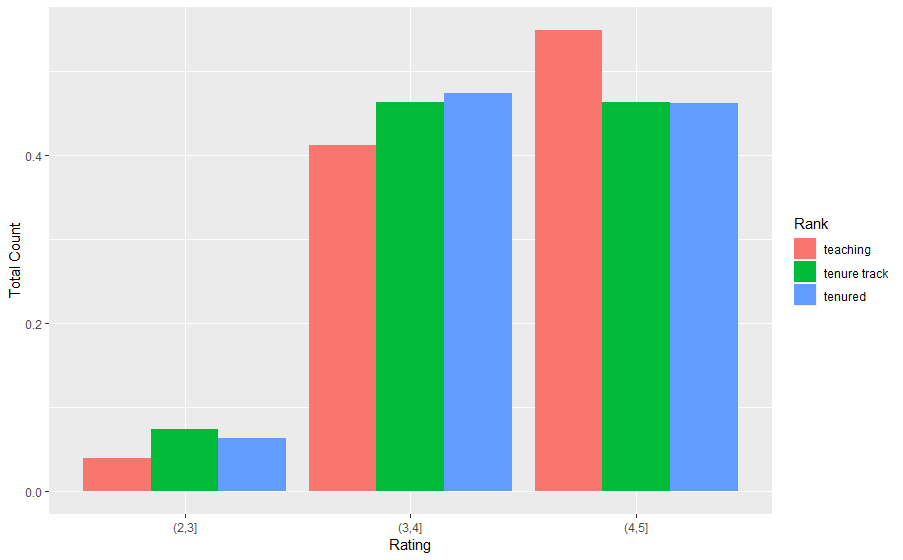


There is not a strong dependency between variables, as intermediate values dominate the plot. In extreme interval, likes (0,2] or (6,8] we can observe the positive correlation, as uglier professors are the worst rated, and most handsome professors are rated with high marks. After this, we continue plotting the comparison between men’s and women’s average rating of professor’s beauty, and course ratings.



In the left, we see the plot when women rated, whereas in the right side we can observe the plot when men rated. If we focus on the centred intervals, which have more observations than others, we can say that women tend to overrate professor’s beauty, and men tend to underrate it. From this analysis, we can conclude that beauty is another variable that depends linearly with course evaluations, but in a weak way. Previous conclusions, such as uglier professors are the ones with worst ratings are verified in these two new plots.

* Finally, we are going to explore the relationship between professor’s rank and course evaluations. If we look the R-matrix variable that contained the correlated values for all the variables of ‘eval’ dataset, we will see that correlation between rank and ‘eval’ is negative with -0.1 value.



We can verify this negative and weak dependency between variables because teaching bar plots (the lowest ranked ones) have better ratings than in other ranks (tenure track or tenured). However, as the difference between those bar plots depending on the rank is so small, we can conclude that such linear dependency is very weak.

1. **DATA ANALYSIS PLAN**

The dataset containing the course evaluations for every subject in the university does not have a labelled class which could be used as an output to target by doing a prediction. Therefore, the goal of the dataset cannot be to predict a concrete attribute by terms of others. We decide to choose as goal the exploration one, rather than the descriptive. In the first case, we analyse the relationships that exist beyond the data given. In the second one, data is described (range, values, mean, median…) using descriptive statistics.

The benefits of applying exploration rather than description are obvious. It is possible to compare two variables that we consider important in the analysis. Also, plots are very common for inferential statistics, showing the data in graphs. In this way, we extract knowledge that can be used later to express the conclusion on every test.

As we set exploration as the main objective of the data analysis, we should divide our planning into three different sections. These sections answer questions such as what is the level of measurement, the study design or the research questions.

* **Research Question**

Does it exist any association between course evaluations and any of the other variables?

In our case of study, we want to determine if there exists any association between the variables of the dataset, and ‘eval’ variable. There are three types of associations (agreement, correlation and regression), and in the study of design will be decided which to choose.

Although there are different research questions that were solved using descriptive plots that could help us doing it, the main goal of the work is to see which variable is highly associated with ‘eval’. Therefore, we include another research question that expresses the basic target of the task.

* **Study Design**

As we want to measure any type of relationship between ‘eval’ and other variables, we can simply deal with the analysis using a regression method. Doing it, we measure the association between the outcome variable ‘eval’ and the whole set of variables, except ‘eval’.

However, it can be useful to measure the correlation between two variables, being one of them the outcome one. In this case, we can compare directly the dependency of variables in our outcome attribute.

Finally, we choose regression model as it gives a more precise knowledge of the data, by terms of computing the contribution of each predictor variable.

* **Level of Measurement**

The type of data that is contained in the attributes is also called the level of measurement of the attribute. For regression models, it depends on the data type of the outcome variable whether the regression is linear or not. For our study, the outcome variable is measured in numerical values within an interval from 0 to 5. Therefore, we should apply a linear regression to measure quantitatively the data.