2013

Data Mining and Machine Learning Techniques for Extracting Patterns in Students' Evaluations of Instructors

Ernest Fokoue

Necla Gündüz

Data Mining and Machine Learning Techniques for Extracting Patterns in Students' Evaluations of Instructors

Necla Gündüz*

Department of Statistics
Gazi Üniversitesi, Fen Fakültesi, Istatistik Bölümü
06500 Teknikokullar, Ankara, Turkey
ngunduz@gazi.edu.tr

Ernest Fokoué
Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive, Rochester, NY 14623, USA
ernest.fokoue@rit.edu

Abstract

The evaluation of instructors by their students has been practiced at most universities for many decades, and there has always been a great interest in a variety of aspects of the evaluations. Are students matured and knowledgeable enough to provide useful and dependable feedback for the improvement of their instructors' teaching skills/abilities? Does the level of difficulty of the course have a strong relationship with the rating the student give an instructor? In this paper, we attempt to answer questions such as these using some state of the art statistical data mining techniques such support vector machines, classification and regression trees, boosting, random forest, factor analysis, kMeans clustering. hierarchical clustering. We explore various aspects of the data from both the supervised and unsupervised learning perspective. The data set analyzed in this paper was collected from a university in Turkey. The application of our techniques to this data reveals some very interesting patterns in the evaluations, like the strong association between the student's seriousness and dedication (measured by attendance) and the kind of scores they tend to assign to their instructors.

I. Introduction

The evaluation of instructors by their students has been practiced at most universities for many decades. Typically, these evaluations are administered in the form of long surveys answered by students at the end of the semester (quarter). Questions in the survey are related to such aspects as course organization, level and quality of delivery, clarity of course objectives, level of difficulty of the course, impact of the course on the student's overall university experience and goals, relevance of the course, preparedness and competency of the instructor, likeability and fairness of the instructor, overall satisfaction of the student, and overall rating of the instructor by the student, just to name a few. The overarching goal of students's evaluations of instructors is the extraction of feedback from students with the finality of inspiring their professors to teach better and help students learn more/better. Typically, most university administrators such as department heads,

*Corresponding Author

school directors, college deans, provosts and chancellors have tended to rely on a single grand average of the questionnaire scores as a measure of the quality of an instructor. Given the complex and multidimensional nature of the questionnaires administered, it is clearly misleading to summarize such evaluations with a single number. Besides, the averages usually relied upon are not valid, because of the non-numeric nature of the Likert-type of the evaluation responses/scores. Indeed, since the publication of the seminal Likert (1932) paper, Likert-type scores have been extensively used in a wide variety of fields ranging from Anthropology, Psychology, Education, Sociology, Sports just to name of a few. Unfortunately, with the astronomical number of applications of the Likert measurement system, there have also been innumerable abuses, especially the misuse of Likert-type scores as real-valued scores. Authors such as Sisson and Stocker (1989), Clason and Dormody (1994), Jamieson (2004) and Allen and Seaman (2007) provide pointers to the uses abuses of Likert-type data. Many authors have indeed cautioned experimenters on the meaninglessness of statements made based on analyses with inappropriate techniques. To quote Adams et al. (1965), *"Nothing is wrong per se in applying any statistical operation to measurements of given scale, but what may be wrong, depending on what is said about the results of these applications, is that the statement about them will not be empirically meaningful or else that it is not scientifically significant"*. Along the lines of Adams et al. (1965), many authors have written numerous articles providing guidelines as to which statistical techniques are most appropriate for Likert-type and the so-called Likert-scale datasets. Boone and Boone (2012) of instance provides a clear separation between Likert-type and Likert-scale, and strongly recommends nonparameteric techniques for Likert-type and parametric techniques for Likert-scale. To avoid such pitfalls of meaningless conclusions on our data, we strive to guarantee the validity of our analyses and summaries, by using mostly Likert-type specific (or at least Likert-type compatible) techniques and tools of exploratory data analysis, cluster analysis, dimensionality reduction and pattern recognition. The rest of this paper is organized as follows: in section 2 we present some general definitions and address important aspects of survey data such item reliability and respondent reliability. We also present empirical answers to most of the above questions using both appropriate exploratory data analysis tools and some straightforward tests of association. In section 3 we focus on the multivariate aspects of the data and answer most of the students' evaluation of instructors questions by using tools such as factor analytic and cluster analysis which both reveal very meaningful confirmation of some beliefs and perceptions about the rating of professors by their students. Section 4 uses some of the results from section 3 to perform predictive analytics on this data. We specifically apply state of the art pattern recognition techniques such as support vector machines, boosting, random forest and classification trees to predict the satisfaction level of a given student based on their answers to the 28 questions on the survey. Section 5 provides our conclusion and discussion, along with pointers to our future work.

II. Exploratory Data Analysis, Data Quality and Definitions

The dataset analyzed in this paper is made up of $n = 5820$ evaluations completed by students in recent years at a university in Turkey. The questionnaires have a total of 31 questions, 28 of which are course evaluation specific, and the remaining 3 represent scores on such items as *student's perceived difficulty level of the course*, *attendance*, *number of repetitions of the course*.

II.1 Definitions and data quality

The dataset is represented by an $n \times p$ matrix $\boldsymbol{X}$ whose ith row $\mathbf{x}_i^\top \equiv (\mathrm{x}_{i1}, \mathrm{x}_{i2}, \cdots, \mathrm{x}_{ip})$ denotes the p-tuple of characteristics, with each $\mathrm{x}_{ij} \in \{1, 2, 3, 4, 5\}$ representing the Likert-type level (order)

of preference of respondent i on item j. Recall that a Likert-type score is obtained by translating/mapping the response levels $\{\texttt{Strong Disagree}, \texttt{Disagree}, \texttt{Neutral}, \texttt{Agree}, \texttt{Strongly Agree}\}$ into pseudo-numbers $\{1,2,3,4,5\}$. With each $\mathtt{x}_{ij} \in \{1,2,3,4,5\}$, the data matrix $\mathbf{X}$ for a survey of $p = 7$ questions answered by $n = 10$ respondents looks like

$$\mathbf{X} = \begin{bmatrix} \mathtt{x}_{11} & \mathtt{x}_{12} & \cdots & \mathtt{x}_{1p} \\ \mathtt{x}_{21} & \mathtt{x}_{22} & \cdots & \mathtt{x}_{2p} \\ \mathtt{x}_{31} & \mathtt{x}_{32} & \cdots & \mathtt{x}_{3p} \\ \mathtt{x}_{41} & \mathtt{x}_{42} & \cdots & \mathtt{x}_{4p} \\ \mathtt{x}_{51} & \mathtt{x}_{52} & \cdots & \mathtt{x}_{5p} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathtt{x}_{n-1,1} & \mathtt{x}_{n-1,2} & \cdots & \mathtt{x}_{n-1,p} \\ \mathtt{x}_{n1} & \mathtt{x}_{n2} & \cdots & \mathtt{x}_{np} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 4 & 5 & 5 & 5 & 4 \\ 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ 5 & 4 & 4 & 3 & 3 & 1 & 5 \\ 2 & 1 & 1 & 1 & 5 & 5 & 3 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 4 & 4 & 3 & 5 & 5 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 & 5 & 5 \\ 5 & 5 & 5 & 1 & 5 & 3 & 4 \\ 1 & 1 & 1 & 1 & 1 & 5 & 1 \\ 2 & 5 & 1 & 3 & 5 & 3 & 3 \end{bmatrix} \tag{1}$$

A usually crucial part in the analysis of questionnaire data is the calculation of the Cronbach's alpha coefficient which measures the reliability/quality of the data. Let $X = (X_1, X_2, \cdots, X_p)^\top$ be a p-tuple representing the p items of a questionnaire. The Cronbach's alpha coefficient is a function of the ratio of the sum of the idiosyncratic item variances over the variance of the sum of the items, and is given by

$$\alpha = \left(\frac{p}{p-1}\right)\left[1 - \frac{\sum_{j=1}^{p} \mathbb{V}(X_j)}{\mathbb{V}\left(\sum_{\ell=1}^{p} X_\ell\right)}\right]$$

For our data, we found $\alpha = 0.992$, indicating a reliable (ie good quality) survey instrument from Cronbach's point of view. We must emphasize however, that this is just *item reliability*, which is of course of great importance, but herein contrasted with *respondent reliability* which we had to assess and address as a result of some patterns discovered in our data.

Definition 1. *Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a dataset with $\mathbf{x}_i^\top = (\mathtt{x}_{i1}, \mathtt{x}_{i2}, \cdots, \mathtt{x}_{ip})$. An observation vector $\mathbf{x}_i$ will be called a zero variation vector if $\mathtt{x}_{ij} = \texttt{constant}$, $j = 1, \cdots, p$. Respondents with zero variation response vectors will be referred to as* `single minded` *respondents/evaluators.*

Example 1. *Observations* 2 *and* 5 *in equation* (1) *are zero variation observations. In our data set of $n = 5820$ evaluations, we found a rather high prevalence of single minded evaluators, specifically, about half of the evaluations (*$2985/5820 \approx 51\%$*). In fact, zero variation responses essentially reduce a p items survey to a single item survey.*

Theorem 1. *Let $X = (X_1, X_2, \cdots, X_p)^\top$ be a p-tuple representing the p items of a questionnaire. If X is zero variation, then the Cronbach's alpha coefficient will be equal to* 1.

Proof. If $X = (X_1, X_2, \cdots, X_p)^\top$ is *zero variation*, then $X_j = W$ for $j = 1, \cdots, p$, and $\sum_{j=1}^{p} X_j = pW$. As a result, $\sum_{j=1}^{p} \mathbb{V}(X_j) = p\mathbb{V}(W)$ and $\mathbb{V}\left(\sum_{j=1}^{p} X_j\right) = \mathbb{V}(pW) = p^2\mathbb{V}(W)$. Therefore,

$$\alpha = \left(\frac{p}{p-1}\right)\left[1 - \frac{\sum_{j=1}^{p} \mathbb{V}(X_j)}{\mathbb{V}\left(\sum_{\ell=1}^{p} X_\ell\right)}\right] = \left(\frac{p}{p-1}\right)\left[1 - \frac{p\mathbb{V}(W)}{p^2\mathbb{V}(W)}\right] = \left(\frac{p}{p-1}\right)\left[1 - \frac{1}{p}\right] = 1$$

□

It is our view that *zero variation* responses are an indication that the respondent did not give deep thought to each of the questions/items of the survey. One could always argue that such evaluators came in with a single rating on all the items, ad that such responses are just fine. However, considering the sometimes drastically different foci of the questions, it is rather unlikely that a given instructor on a given course would perform exactly the same on all the items. On the other hand, such *zero variation* responses convey the impression that the respondent rushed the answering process. Finally, from a point of view of feedback to the instructor in order to help them improve the course, such answers provide very little if any feedback at all. For all the above reasons, we deem the *zero variation* responses unreliable. We use a straightforward adaptation of the Cronbach's alpha coefficient to measure and capture *respondent reliability*

Definition 2. *Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a dataset with $\mathbf{x}_i^\top = (\mathtt{x}_{i1}, \mathtt{x}_{i2}, \cdots, \mathtt{x}_{ip})$. Let the estimated variance of the ith respondent be $\tilde{S}_i^2 = \sum_{j=1}^{p} (\mathtt{x}_{ij} - \bar{\mathtt{x}}_i)^2/(p-1)$. Let $Z_j = \sum_{i=1}^{n} \mathtt{x}_{ij}$ represent the sum of the scores given by all the n respondents to item j. Our respondent reliability is estimated by*

$$\hat{\tilde{\alpha}} = \left(\frac{n}{n-1}\right)\left[1 - \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\mathtt{x}_{ij} - \frac{1}{p}\sum_{j=1}^{p}\mathtt{x}_{ij}\right)^2}{\displaystyle\sum_{j=1}^{p}\left(\sum_{i=1}^{n}\mathtt{x}_{ij} - \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{n}\mathtt{x}_{ij}\right)^2}\right]$$

Given a data matrix $\boldsymbol{X}$, respondent reliability can be computed in practice by simply taking the Cronbach's alpha coefficient of $\boldsymbol{X}^\top$, the transpose of the data matrix $\boldsymbol{X}$. Let m be the number of *nonzero variation*. If $m \ll p$ and m/n is very small, then respondent reliability will be very poor. Fortunately, for our data, respondent reliability is estimated at 0.996, which is very satisfactory. We think this large value is due to the fact that, despite having more than 50% *zero variation* respondents, we still a large enough sample. Despite this however, we will perform analyses taking into account the dichotomy between single minded respondents and their counterparts.

II.2 Exploratory Data Analysis and Basic Tests

As we said earlier, students' evaluations of instructors are administered with the goal of measuring the effectiveness (quality) of instructors and hopefully provide them (the instructors) with useful feedback to help them teach better. Clearly, such a goal is complex, and because of its complexity, there have always been heated and often very passionate debates about the validity and the appropriateness of such evaluations. As a matter of fact, many professors strongly believe and claim that students, especially undergraduate students, are neither mature enough nor knowledgeable enough nor objective enough to provide useful feedback to their instructors. To a certain degree, such anti-students' evaluations professors do have a valid point because even with the crucial issues of maturity, knowledge and objectivity, there are very important points of concerns with students' evaluations of instructors: (a) a complex multidimensional instrument like a 28 items questionnaire should never be summarized using a single number (as it is commonly practiced around the world), because such a simplistic summarization definitely fails to capture all the niceties inherent in the complex art of teaching (b) given the Likert type nature of the scores (responses), the often used grand average is at best misleading because averages computed on non-numeric variables are often meaningless as we will show later in very striking simple examples, see Adams et al. (1965). It makes sense that only a multidimensional summary or better yet a functional summary (density or mass function) can meaningfully capture the pattern underlying a multidimensional instrument like the students' evaluations of instructors. In

this section, we provide just such summarizations, bearing in mind the Likert-type nature of our data, and therefore providing only the kind of statistic analysis appropriate for such data types.

II.2.1 Beware of misleading and meaningless averages

It's a very common practice among people dealing with Likert type data to use averages and standard deviations as their measures of central tendency and measures of spread (variation) respectively. Universities use the grand mean as the overall `rating` of an instructor

$$\texttt{grandmean}(\boldsymbol{x}) = \frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\texttt{x}_{ij} = \texttt{mean}\left(\underset{j=1:p}{\texttt{mean}}(s(\boldsymbol{x}_j))\right),$$

where $s(\boldsymbol{x}_j) = \{\texttt{x}_{ij} : i = 1\cdots, n\}$. Typically, when an instructor opens the website containing her/his student evaluation data, there are 28 averages, one for each questions, and then there is the average of those averages which is the grand mean representing the overall rating of the instructor. With $\texttt{x}_{ij} \in \{1,2,3,4,5\}$, such a grand average is a best misleading and at worst just plain invalid. In the hierarchy of data types, Likert type scores are no more than ordinal, which prohibits the use of averages. By their very nature, Likert-type observations are inherently definitely not numerical in the usual sense of interval or ration data. Consider a sample of size 2 with $\texttt{x}_1 = 4 = \texttt{Agree}$ and $\texttt{x}_2 = 2 = \texttt{Disagree}$. When treated as a sample of numeric (ratio) observations $\bar{\texttt{x}} = (\texttt{x}_1 + \texttt{x}_2)/2 = (4+2)/2 = 3 = \texttt{Neutral} = (\texttt{Agree} + \texttt{Disagree})/2$, which means that if one `agrees` and the other `disagrees`, one should expect `neutrality`. It's somewhat weird that one should expect `neutrality` as a result of an `agreement` and a `disagreement`. Consider another sample of size 3 with $\texttt{x}_1 = 3 = \texttt{Neutral}$, $\texttt{x}_2 = 4 = \texttt{Agree}$ and $\texttt{x}_3 = 5 = \texttt{Strongly Agree}$. Then our corresponding average is $\bar{\texttt{x}} = (\texttt{x}_1 + \texttt{x}_2 + \texttt{x}_3)/3 = (3+4+5)/3 = 4 = \texttt{Agree} = (\texttt{Neutral} + \texttt{Agree} + \texttt{Strongly Agree})/3$. Such an average is meaningless and somewhat misleading. In fact, if consider even the following example where we have $\texttt{x}_1 = 1 = \texttt{Strongly Disagree}$, $\texttt{x}_2 = 1 = \texttt{Strongly Disagree}$, $\texttt{x}_3 = 5 = \texttt{Strongly Agree}$ and $\texttt{x}_4 = 5 = \texttt{Strongly Agree}$. Then, the average is $\bar{\texttt{x}} = (\texttt{x}_1 + \texttt{x}_2 + \texttt{x}_3 + \texttt{x}_4)/4 = (1+1+5+5)/4 = 3$. Written in orginal Likert, it is $\bar{\texttt{x}} = (\texttt{Strongly Disagree} + \texttt{Strongly Disagree} + \texttt{Strongly Agree} + \texttt{Strongly Agree})/4 = \texttt{Neutral}$. This means that two instances of `Strongly Disagree` combine with two instances of `Strongly Agree` to yield a `neutral` as the expected opinion. There is definitely no valid basis for such an average. In fact, if push the same kind of reasoning even further by considering a sample of 50 observations, with 25 instances of `Agree`, namely $\{\texttt{x}_i = 4, i = 1, \cdots, 25\}$, and 25 instances of `Disagree`, namely $\{\texttt{x}_i = 2, i = 26, \cdots, 50\}$, we get the average $\bar{\texttt{x}} = \sum_{i=1}^{50}\texttt{x}_i/50 = (\sum_{i=1}^{25}4/25 + \sum_{i=16}^{50}4/25)/2 = (4+2)/2 = 3 = \texttt{Neutral} = (\texttt{Agree} + \texttt{Disagree})/2$. It is definitely misleading to conclude that 25 agreements combined with 25 disagreements should yield neutrality. The use of averages for data collected/scored this way is at best misleading because the numbers assigned to the opinions do not have an objective numerical value in the sense of a measurement. Going back to our motivating example of the students' evaluation of instructors, the use of the grand mean as the overall rating of the instructor misses the subtle and important information reveal by appropriate frequencies (proportions) and the corresponding bar plots. When the grand mean is used, Instructor 1 scores an average of 3.4, which of course tells us nothing about the distribution of her scores. Figure (1) reveals that the distribution for this instructor is skewed to the left, with a pronounced/strong mode at 4 for most of the questions/items. Although we still do not advocate the use of a single number to summarize a complex instrument like a students' evaluations of instructor, we would recommend trusting the mode rather than the mean if a single number were to be used. This led us to defining a grand mode in place of

the invalid grand mean as follows: Let $s(\boldsymbol{x}_j) = \{\mathtt{x}_{ij} : i = 1 \cdots, n\}$ and let $\tilde{\mathtt{x}}_{ij} = \mathtt{unique}(\mathtt{x}_{ij})$. If m_j denotes the mode of variable X_j, then for $j = 1, \cdots, p$, we can readily compute the mode of the jth column as

$$m_j = \underset{\tilde{\mathtt{x}}_{ij} \in s(\boldsymbol{x}_j)}{\mathtt{arg\,max}} \{ frequency(\tilde{\mathtt{x}}_{ij}) \} = \mathtt{mode}(s(\boldsymbol{x}_j)).$$

The set $M = \{m_1, m_2, \cdots, m_p\}$ containing the modes for the p columns. Let $\tilde{m}_j = \mathtt{unique}(m_j)$. We can find the grand mode as

$$\mathtt{grandmode}(\boldsymbol{x}) = \underset{\tilde{m}_j \in M}{\mathtt{arg\,max}} \{ frequency(\tilde{m}_j) \} = \mathtt{mode} \left(\underset{j=1:p}{\mathtt{mode}}(s(\boldsymbol{x}_j)) \right),$$

The grand mode for instructor 1 is found to be 4, which, in light of the distribution of her scores, is a more accurate summarization of her effectiveness and teaching quality.

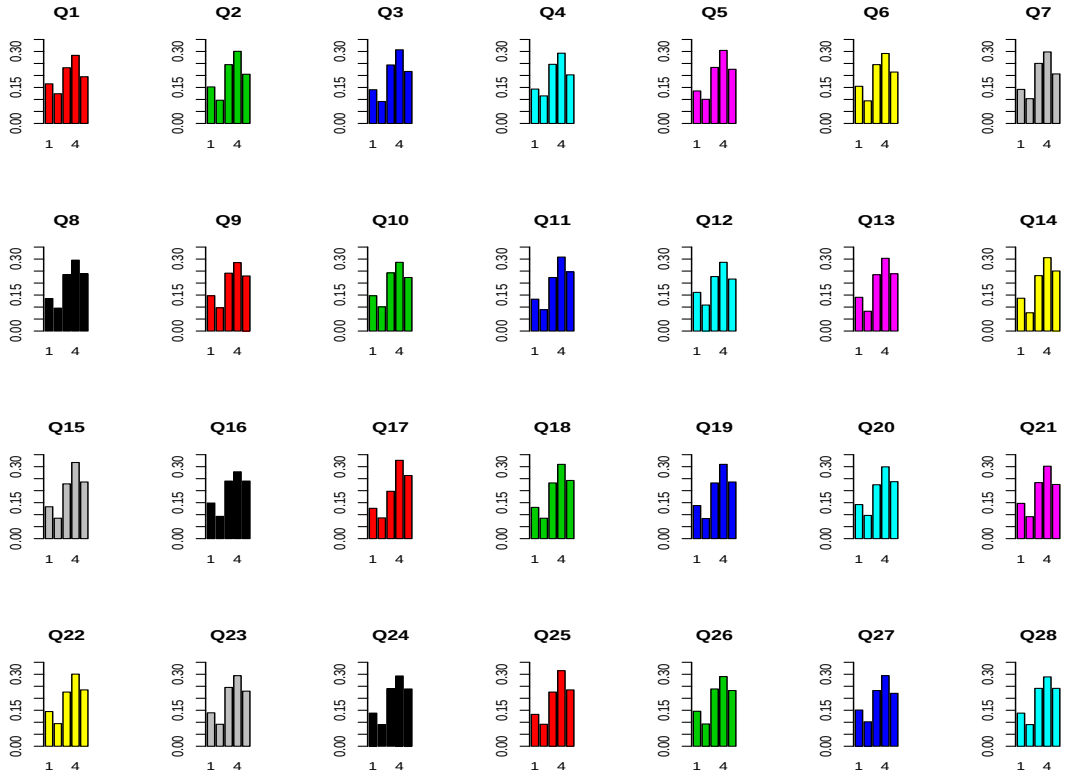

Figure 1: *Barplot of the scores received by Instructor 1 from the $n_1 = 775$ who evaluated her courses.*

It might be tempting, given the ordinal nature of Likert-type data, to use the grand median

$$\mathtt{grandmedian}(\boldsymbol{x}) = \mathtt{median} \left(\underset{j=1:p}{\mathtt{median}}(s(\boldsymbol{x}_j)) \right),$$

in place of the grand mean. From our experience, such a summarization is not as accurate as the grand mode, partly due to the floor and ceiling effect. See Clason and Dormody (1994)

If instead of considering only instructor 1 we use the entirety of the data with all the $n = 5820$ evaluations, the distribution of all the 28 course specific questions is given in Figure (2). Clearly, most questions attain their mode at 3, and we find the grand mode to be 3.

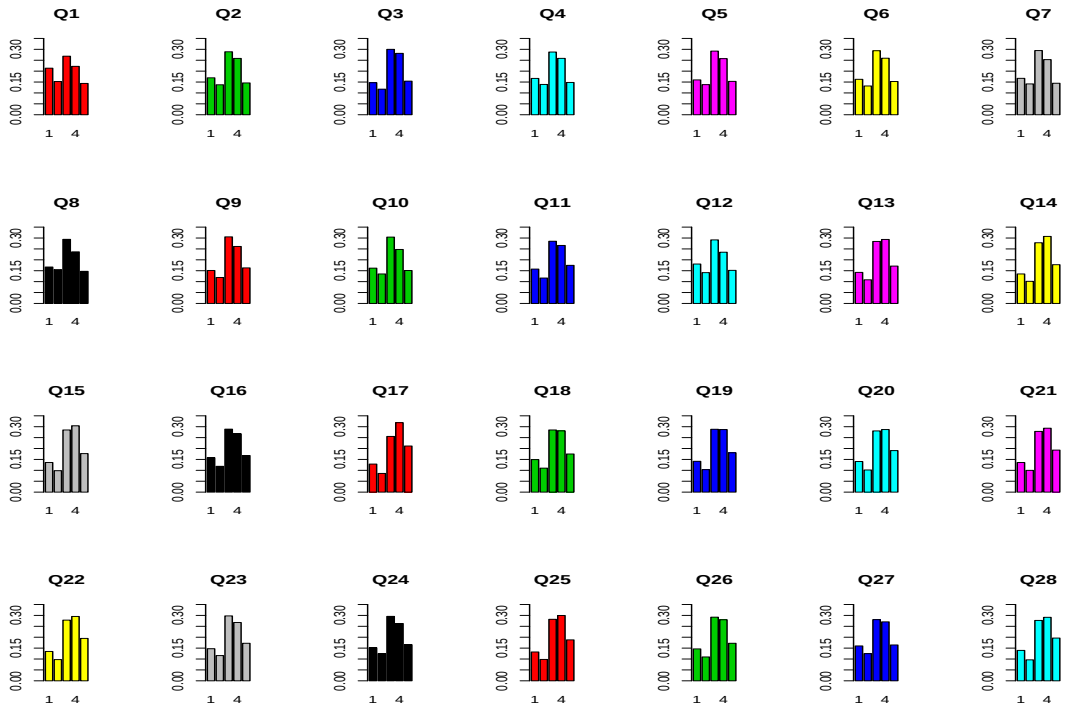

Figure 2: *Barplots for all the $n = 5820$ student evaluation regardless of the professor.*

Thanks to the distributional features of the scores of the instructors in this dataset, namely the skewness to the left, we able to comment in a more complete manner on the effectiveness (or at least the students' perception thereof). With the highest frequencies being between 3 and 5, it is fair to say that the instructors evaluated here are **not** negatively perceived by their students.

	Zero Variation	Nonzero Variation	Total
Instructor 1	0.0789	0.0543	0.1332
Instructor 2	0.1309	0.1172	0.2481
Instructor 3	0.3031	0.3156	0.6187
Total	0.5129	0.4871	**1**

Table 1: *Distribution of response variation among instructors*

II.2.2 Examining the Effect of Response Variation

Despite the ability to provide a more meaningful single summarization of the whole evaluation through the grand mode along with distributional qualifications, we still need to answer relational questions like the association between student maturity and their rating, student seriousness/dedication/objectivity and their rating. We now propose to focus on *zero variation*

responses, as we believe that the reflect the reliability of the respondent. In a sense, we claim that a student who gives a *zero variation* response is providing a less objective and less mature answer to the survey. We then try to find out if there is an association between *zero variation* and the answers to the questions. First and foremost, it is interesting to assess the association between response variation and instructors. See Table 1.

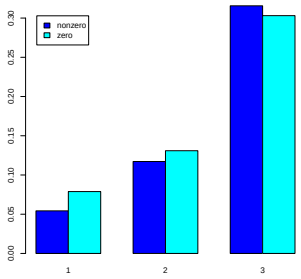

Figure 3: *Barplot of the distribution of response variation as a function of instructor.*

As expected based on the above barplot, the chi-squared test of association between *Instructor* and *Response Variation* is significant, namely with $\chi^2_{\texttt{obs}} = 28.45$, $\texttt{df} = 2$, *p-value* $= 0.000$.

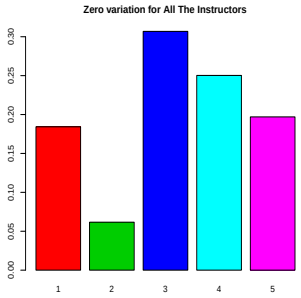

Figure 4: *Barplot of the distribution of Zero Variation Responses for all the instructors.*

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Instructor 1	0.1678	0.0545	0.2288	0.2767	0.2723
Instructor 2	0.1378	0.0407	0.3018	0.3084	0.2113
Instructor 3	0.2086	0.0726	0.3294	0.2183	0.1712

Table 2: *Proportion of each response category for answers with zero variation*

The lack of richness of the *zero variation* responses in this study is less concerning because most of such responses are either neutral or positive. In a sense, those who were single minded about their rating of the courses, were so mostly not because of dissatisfaction, which somewhat means that their feedback was really not needed. For that reason, we can proceed with the remaining aspects of the analysis of this data, secured that both the item reliability (measured by Cronbach's alpha) and the respondent reliability are satisfactory.

II.2.3 Examining Various Important Associations

We now examine a variety of association between different important variables. Taking the view that *attendance* is a measure of dedication/seriousness, and therefore a decent and plausible indicator of the ability/authority of the student to correctly assess their instructor, we will now test the association of various variables with *attendance*. In other words, if a student is not dedicated ie not serious (as measured by attendance), their assessment should probably not be taken seriously. We also consider the variable *difficulty*, a self reported variable provided to allow the student to indicate their perception of the level of difficulty of the course. This variable is particularly important because some instructors strongly believe that students tend to give a negative feedback when they perceive the course to be difficult.

Association between Attendance Level and Response Variation for Instructor 3.

	Poor	Minimal	Good	Good	Excellent
nonzero	0.14912524	0.08747570	0.07220217	0.12663149	0.07470147
zero	0.22299361	0.07497917	0.05942794	0.07275757	0.05970564

Table 3: *Cross tabulation of Attendance level vs Response Variation for Instructor 3.*

The corresponding chi-squared test of association between *Attendance level* and *Response Variation* is significant, specifically with $\chi^2_{\texttt{obs}} = 117.7398$, $\nu = \texttt{df} = 4$, *p-value* $< 2.2e-16$.

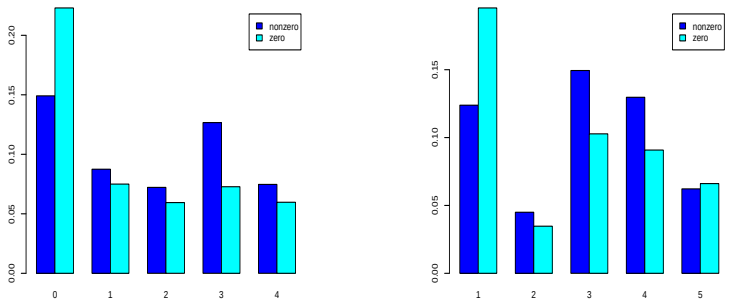

Figure 5: *Instructor 3: (left) Response variation vs attendance level; (right) Response variation vs difficulty level.*

Association between Difficulty Level and Response Variation for Instructor 3.

	Too Easy	Easy	Normal	Difficult	Too Difficult
nonzero	0.12385448	0.04498750	0.14940294	0.12968620	0.06220494
zero	0.19550125	0.03471258	0.10274924	0.09080811	0.06609275

Table 4: *Cross tabulation of Difficulty level vs Response Variation given by students for Instructor 3.*

The corresponding chi-squared test of association between *Difficulty level* and *Response variation* is significant, specifically with $\chi^2_{\texttt{obs}} = 117.7398$, $\nu = \texttt{df} = 4$, *p-value* $< 2.2e-16$.

Various Tests of Association using the whole dataset

It appears that for all the evaluations provided for Instructor 3, both *Attendance Level Difficulty Level* are strongly associated with *Response Variation*. The question then arises as to whether that association holds when all the 5820 evaluations are considered together.

	Poor	Minimal	Good	Good	Excellent
nonzero	0.1273	0.0880	0.0718	0.1218	0.0782
zero	0.1995	0.0887	0.0643	0.0933	0.0672

Table 5: *Cross tabulation of Attendance level vs Response Variation for all the Instructors.*

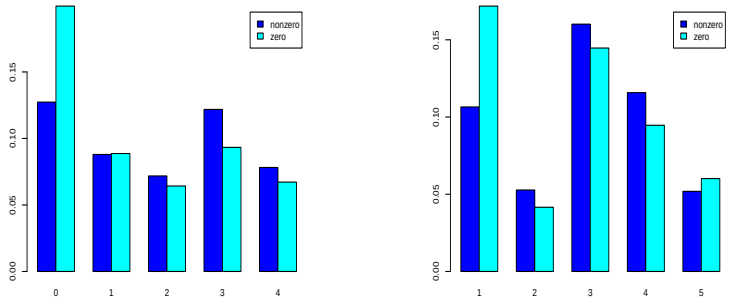

Figure 6: *All instructors: (left) Response variation vs attendance level; (right) Response variation vs Difficulty level.*

The corresponding chi-squared test of association between *Attendance level* and *Response variation* is significant, namely with $\chi^2_{\texttt{obs}} = 118.3$, $\nu = \texttt{df} = 4$, *p-value* $< 2.2e-16$

	Too Easy	Easy	Normal	Difficult	Too Difficult
nonzero	0.1065	0.0527	0.1601	0.1158	0.0519
zero	0.1718	0.0416	0.1447	0.0947	0.0601

Table 6: *Cross tabulation of Difficulty level vs Response Variation for all the Instructors.*

The corresponding chi-squared test of association between *Difficulty level* and *Response variation* is significant, namely with $\chi^2_{\texttt{obs}} = 113.5$, $\nu = \texttt{df} = 4$, *p-value* $< 2.2e-16$.
As the barplots in Figure 6 clearly shows, students with poor attendance give an overwhelmingly large number of *zero variation* answers whereas students with reasonable to excellent attendance level tend to give *nonzero variation* answers. This somewhat confirms or at least supports the strongly held belief that only those answers provided by dedicated/serious students should be taken into account. On the other hand, students who perceive a course as too easy and therefore boring or at least uninteresting also tend to give an overwhelming proportion of *zero variation* answers. Interestingly, students who think the course has a normal difficulty level tend to take time to provide varied answers to different questionnaire items.
There was a total of 13 courses included in the 5820 evaluations considered here. With the interesting patterns discovered between response variation and both attendance and difficulty

level, it is interesting to examine if there might be a similar type of strong association between the courses and the response variation.

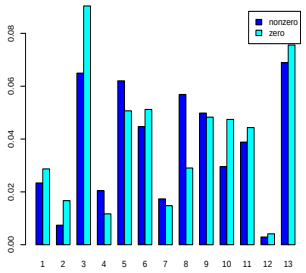

Figure 7: *Cross-tabulation of response variation versus course indicator for all the instructors.*

The above barplot does indeed show that there is an association, and the corresponding chi-squared test of association between *Course* and *Response variation* is significant, with $\chi^2_{\texttt{obs}} = 150.7$, $\nu = \texttt{df} = 12$, *p-value* $< 2.2e-16$.

	Too Easy	Easy	Normal	Difficult	Too Difficult
Poor	0.2263	0.0170	0.0380	0.0249	0.0206
Minimum	0.0137	0.0311	0.0653	0.0431	0.0234
Reasonable	0.0093	0.0131	0.0591	0.0393	0.0153
Good	0.0158	0.0187	0.0859	0.0679	0.0268
Excellent	0.0132	0.0144	0.0565	0.0352	0.0259

Table 7: *Cross tabulation of Difficulty level vs Attendance Level for all the Instructors.*

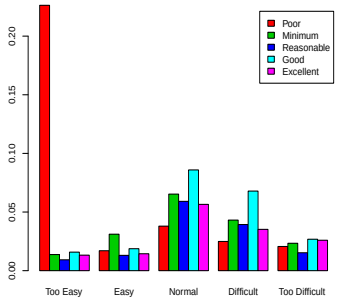

Figure 8: *Cross-tabulation of attendance level versus difficulty level for all the instructors.*

The corresponding chi-squared test of association between *Difficulty level* and *Attendance Level* is significant, with $\chi^2_{\texttt{obs}} = 2528.06$, $\nu = \texttt{df} = 16$, *p-value* $< 2.2e-16$. The most obvious feature of this association is the astronomically high proportion of poor attendance in courses deemed too easy. No surprise here, just plain common sense. Sadly however, there is no category in which excellent attendance dominates.

Although all the 28 items in the questionnaire were carefully selected by the designers of the students' evaluation, one could make a strong case that some questions are better indicators of overall assessment than others. We have selected 4 such questions, and we know examine the association between those questions and both *attendance level* and *difficulty level*.

Q10: My initial expectations about the course were met at the end of the period or year.

Q14: The Instructor came prepared for classes.

Q20: The Instructor explained the course and was eager to be helpful to students.

Q24: The Instructor gave relevant homework assignments/projects, and helped students.

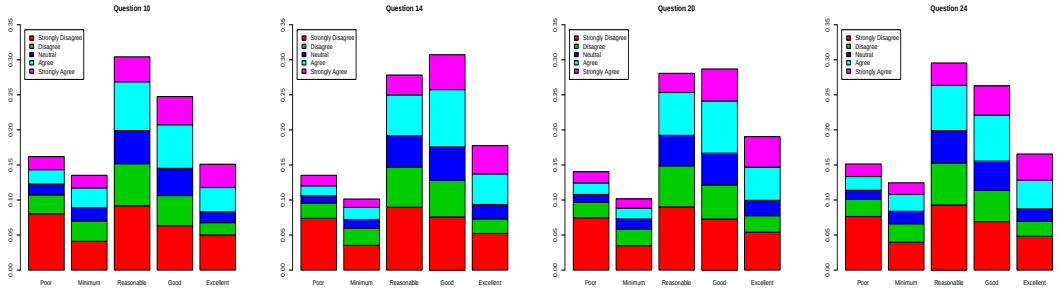

Figure 9: *Cross-tabulation of attendance level versus scores on Questions 10,14,20,24*

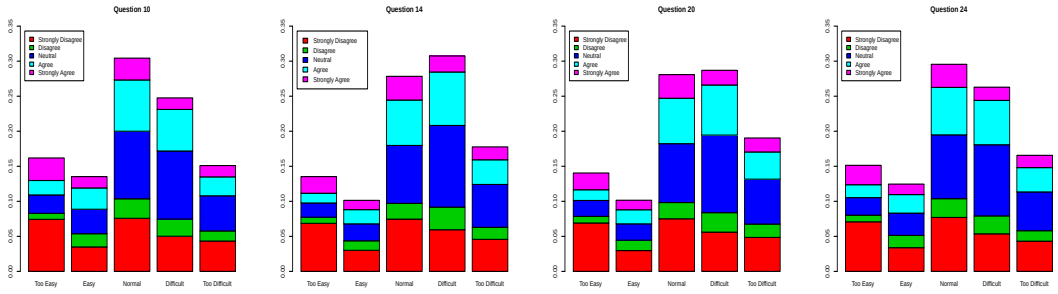

Figure 10: *Cross-tabulation of difficulty level versus scores on Questions 10,14,20,24*

One of the most striking remarks in the distribution of *difficulty level* and *scores* is the fact that `disagree` is the smallest percentage across the board. Also, when students deem the class `too easy`, they tend to express their dissatisfaction in a forceful and really angry tone with the `strong disagree` answers dominating the subgroup. Based on suggestions and speculations by various instructors, we hypothesized 3 groups of students overall, namely `Satisfied`, `Neutral` and `Dissatisfied`. In some cases, we just considered a binary categorization, focusing on dissatisfaction as our *positive*, leading to `Not Dissatisfied Dissatisfied` as our class labels. Using such classifications of students, we can say that Figure 10 and 9 reveal rather clear separations among the groups. On Figure 9 for instance, Question 14 does reveal a clear difference of percentages when our three hypothesized groups are considered. In the next section, we use cluster analysis to extract possible groups in the data, and we then use various techniques of pattern recognition to assess how well one can predict the group of a student based on the pattern of their answers.

III. Pattern Recognition and Association Analysis

In this section, we turn our attention to the multivariate aspects of our data set. We perform Factor Analysis to extract meaningful latent structure and cluster analysis to identify potential groups in the way students rate their professors. As we said clearly in section 1 and 2, Likert-type scores are inherently non-numeric, and applying techniques designed for numeric data on Likert-type will yield answers that are potentially meaningless or at best very difficult to interpret. When it comes to correlation analysis for instance, the default choice is the Pearson correlation measure. With Likert type data however, one wonders if Pearson correlation should ever be used. Based on recommendations by Adams et al. (1965), Boone and Boone (2012), and Clason and Dormody (1994), the correct type of correlation for Likert-type data should be Kendall-tau-B correlation or the Spearman correlation, as these are designed for (ordinal) ranked data. There have been many recent interesting contributions to the multivariate analysis of Likert-type: in her doctoral thesis, Javaras (2004) provides a wide variety of univariate and multivariate tools for analyzing Likert-type data. Narli (2010) proposes the use of rough sets in the analysis of Likert-scale data. We start off by checking how different the Pearson correlation matrix would be from the Kendall-tau B correlation matrix on our data. Recall, that given two random variables X_i and X_j for which observed (realized) values $\mathtt{x}_{1i}, \mathtt{x}_{2i}, \cdots, \mathtt{x}_{ni}$ and $\mathtt{x}_{1j}, \mathtt{x}_{2j}, \cdots, \mathtt{x}_{nj}$ have been respectively gathered, the so-called Pearson sample correlation matrix is given by

$$r_{ij} = \texttt{correlation}(\mathtt{x}_i, \mathtt{x}_j) = r(\mathtt{x}_i, \mathtt{x}_j) = \frac{1}{n-1} \sum_{\ell=1}^{n} \left(\frac{\mathtt{x}_{\ell i} - \bar{\mathtt{x}}_i}{s_{\mathtt{x}_i}} \right) \left(\frac{\mathtt{x}_{\ell j} - \bar{\mathtt{x}}_j}{s_{\mathtt{x}_j}} \right)$$

For a random p-tuple $X = (X_1, X_2, \cdots, X_p)^\top$ and the corresponding data matrix $\boldsymbol{X} = (\mathtt{x}_{ij}), i = 1, \cdots, n, j = 1, \cdots, p$, the Pearson sample correlation matrix is given by

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

As we have been stressing all along, the Likert-type nature of our data makes the matrix $\boldsymbol{R}$ meaningless, in the sense that the averages on which it is based may not have an interpretable meaning. If consider two Likert type (ordered categorical) variables X and Y once again, their Kendall τ-B correlation coefficient $\tau_B(X, Y)$ is given by

$$\tau_B(X, Y) = \frac{n_c(X, Y) - n_d(X, Y)}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

where $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i - 1)/2$, $n_2 = \sum_j u_j(u_j - 1)/2$, t_i=*number of tied values in the i-th group of ties for the first quantity*, u_j=*number of tied values in the j-th group of ties for the second quantity*, n_c = *number of concordant pairs*, n_d=*number of discordant pairs*. For a p-tuple $X = (X_1, \cdots, X_p)$ of p Likert type variables, the Kendall Tau-B correlation matrix is $\boldsymbol{K}$ where

$$\boldsymbol{K} = \begin{bmatrix} \tau_B(X_1, X_1) & \tau_B(X_1, X_2) & \cdots & \tau_B(X_1, X_p) \\ \tau_B(X_2, X_1) & \tau_B(X_2, X_2) & \cdots & \tau_B(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \tau_B(X_p, X_1) & \tau_B(X_p, X_2) & \cdots & \tau_B(X_p, X_p) \end{bmatrix}$$

Figure 11 depicts the graphical representation of both the Pearson and the Kendall τ-B correlation matrices for our data. Surprisingly, the two matrices are very similar, to the point of being almost indistinguishable.

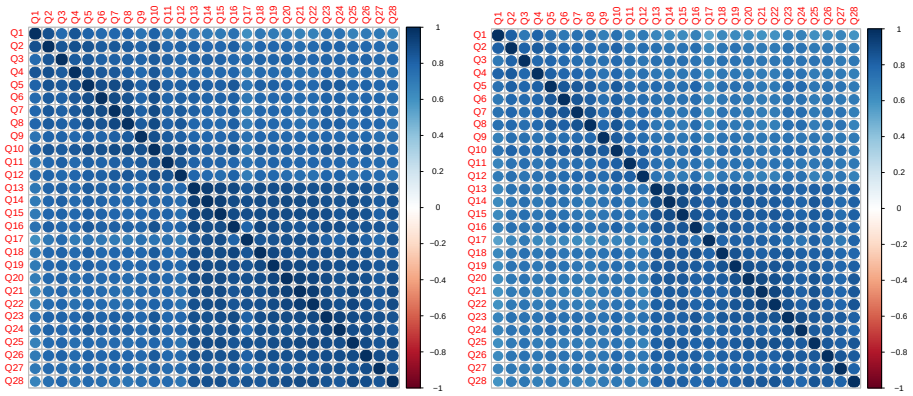

Figure 11: *(left) Pearson correlation matrix (right) Kendall τB correlation matrix for our whole dataset.*

III.1 Are there distinguishable groups among students?

A natural question that arises in the presence of data like the one we have, is whether it can be clustered. In other words, is there such a thing as different groups of students as far as their patterns of feedback to instructors are concerned? Can the patterns of students's evaluations of their instructors be grouped into distinct and clearly describable categories? Now, one of the most celebrated approaches to cluster analysis is the ubiquitous kMeans clustering algorithm. Obviously, as the name suggests, it is based the computation of averages that represent the centers of potential underlying groups. With Likert-type data, it has been stressed all along that averages are potentially meaningless because for the inherently categorical non numeric nature of such data. With correlation matrices computed earlier looking very similar, one might conjecture that for this data, it might be worth using numeric techniques on our data. The kMeans clustering algorithm in this case would proceed by partitioning the data into k clusters such that $\mathcal{P} = C_1 \cup C_2 \cup \cdots \cup C_k$ such that the best cluster minimizes the within-cluster sum of squares (WCSS). In other words, if $\mathcal{P}^*$ denotes the best partitioning (clustering) of the data, we must have

$$\mathcal{P}^* = \underset{\mathcal{P}}{\texttt{argmin}} \left\{ \sum_{j=1}^{k} \sum_{i=1}^{n} I(\boldsymbol{x}_i \in C_j) \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2 \right\}$$

where $\boldsymbol{\mu}_j$ is the mean vector (center) for cluster C_j. Figure (12) seems to clearly suggest that one should retain three distinct clusters. Indeed, two clusters would capture a very low percentage of the variation in the data, while four clusters do not substantially improve the amount of variation captured by three clusters. We therefore retained three clusters and carefully examined both the percentage of observations in each one of them and the values of the centers. As Table 8 shows, one could venture to say that almost 60% of the students have a neutral opinion of the courses they took, and this seems to apply to almost all the 28 questions of the survey. The cluster analytic result also suggests that 17% expressed maximum satisfaction with the courses they took. Finally

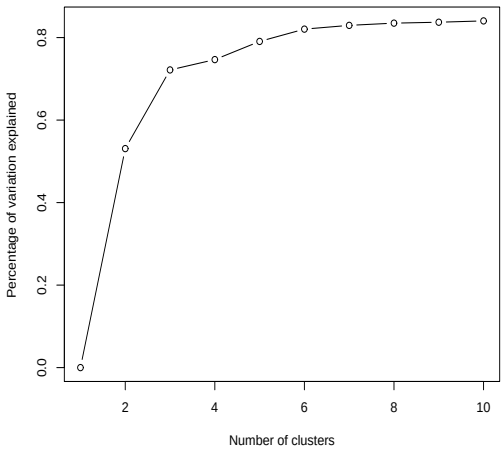

Figure 12: *Percentage of variation explained by cluster size*

a third group of the students seems to be the group of very dissatisfied students, with our data showing roughly 23% of such students. These numbers apply to all the 5820 evaluations analyzed. It certainly would be more interesting, in the context of instructor's improvement, to extract such clustering for each course in order to help the instructor identify areas of improvement.

	Cluster 1	Center 2	Center 3
Average of Center	4.80	1.52	3.37
Number of Observations	1010	1364	3446
Percentage of observations	17.35%	23.44%	59.21%
Suggested class label	Satisfied	Dissatisfied	Neutral

Table 8: *Clusters extracted using kMeans clustering.*

Besides *k*Means clustering, we thought it useful to also perform hierarchical clustering. We use the Euclidean distance coupled with the Ward linkage to generate three clusters. We extracted the corresponding labels, $y \in \{\texttt{Dissatisfied}, \texttt{Neutral}, \texttt{Satisfied}\}$ and performed classification with various pattern recognition techniques. We also merged the `Neutral` and `Satisfied` into one group call `Not Dissatisfied` and performed binary classification, again using many different pattern recognition techniques.

	Cluster 1	Center 2	Center 3
Average of Center	2.93	4.25	1.40
Number of Observations	2176	2471	1173
Percentage of observations	37.38%	42.45%	20.15%
Suggested class label	Neutral	Satisfied	Dissatisfied

Table 9: *Clusters extracted using hierarchical clustering with the Euclidean distance and the Ward linkage.*

We now use the Jaccard distance combined with the Ward linkage. Prior to using the Jaccard distance, the data must be binarized. Consider two subjects with observation vectors $\mathbf{x}_i^\top =$

	Cluster 1	Center 2	Center 3
Average of Center	3.00	4.44	1.25
Number of Observations	916	1335	734
Percentage of observations	30.68%	44.72%	24.60%
Suggested class label	Neutral	Satisfied	Dissatisfied

Table 10: *Clusters extracted using hierarchical clustering on evaluations with zero variation.*

$(\mathtt{x}_{i1}, \mathtt{x}_{i2}, \cdots, \mathtt{x}_{ip})$ and $\mathbf{x}_j^\top = (\mathtt{x}_{j1}, \mathtt{x}_{j2}, \cdots, \mathtt{x}_{jp})$ respectively, where each item $\mathtt{x}_{\kappa\ell} \in \{1,2,3,4,5\}$ is gathered on a Likert-scale. How does one go about measuring the similarity between the two subjects? For instance, if $p = 2$ and we have $\mathbf{x}_1^\top = (1,2), \mathbf{x}_2^\top = (2,3), \mathbf{x}_3^\top = (4,5), \mathbf{x}_4^\top = (3,3), \mathbf{x}_5^\top = (4,4), \mathbf{x}_6^\top = (3,5), \mathbf{x}_7^\top = (1,5)$ How does equidistance (or the lack of it) affect the interpretation of distances between two Likert-scale vectors. For instance, if we compute some of the City block (Manhattan) distances among the above subjects (vectors): $\|\mathbf{x}_1 - \mathbf{x}_2\|_1 = 2, \|\mathbf{x}_2 - \mathbf{x}_3\|_1 = 2, \|\mathbf{x}_4 - \mathbf{x}_5\|_1 = 2, \|\mathbf{x}_4 - \mathbf{x}_6\|_1 = 2, \|\mathbf{x}_5 - \mathbf{x}_6\|_1 = 2, \|\mathbf{x}_6 - \mathbf{x}_7\|_1 = 2$. What does one do with the fact that all the above distances are 2? Due to the very nature of Likert-type data, those distances cannot be interpreted in the same way that Manhattan distances are interpreted!

Many researchers have argued that a suitable transformation of Likert-type should suffice to allow the use of general parametric statistical techniques. Wu (2007) provides an empirical study of the kind of transformations that make it possible to derive numerical scores from Likert-type data. For our data however, we simply binarize the scores and use the Jaccard distance on them. Indeed, once binarized, the data can be analyzed using other distances and tools dedicated to indicator variables. Consider two subjects with responses $\mathbf{x}_i^\top = (\mathtt{x}_{i1}, \mathtt{x}_{i2}, \cdots, \mathtt{x}_{ip})$ and $\mathbf{x}_j^\top = (\mathtt{x}_{j1}, \mathtt{x}_{j2}, \cdots, \mathtt{x}_{jp})$ where $\mathtt{x}_{\kappa\ell} \in \{1,2,3,4,5\}$. We binarize the Likert-type in such a way that

$$\mathbf{x} \in \{1,2,3,4,5\}^p \mapsto \tilde{\mathbf{x}} \in \{0,1\}^{5p},$$

For instance, four subjects in three dimensions

$$\boldsymbol{X} = \begin{bmatrix} 3 & 4 & 1 \\ 3 & 5 & 5 \\ 5 & 1 & 3 \\ 5 & 1 & 1 \end{bmatrix} \mapsto \tilde{\boldsymbol{X}} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

After the data is binarized, we can then calculate $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, the similarity between the two subjects using the Jaccard similarity index defined by

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = J(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \frac{\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j}{|\tilde{\mathbf{x}}_i|^2 + |\tilde{\mathbf{x}}_j|^2 - \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j}.$$

When performing cluster analysis, we use the Jaccard distance defined as

$$d_J(\mathbf{x}_i, \mathbf{x}_j) = 1 - J(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = 1 - \frac{\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j}{|\tilde{\mathbf{x}}_i|^2 + |\tilde{\mathbf{x}}_j|^2 - \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j}.$$

This distance is also known as the *asymmetric binary distance* because matching 1's counts toward similarity, whereas matching 0's do not. Another very popular similarity measure is the Sorensen

similarity given by

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = S(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = 2\frac{\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j}{|\tilde{\mathbf{x}}_i|^2 + |\tilde{\mathbf{x}}_j|^2}.$$

Now, for the example given above, the corresponding Manhattan and Jaccard distances are respectively given below:

	1	**2**	**3**
2	5		
3	7	8	
4	5	10	2

$\Longrightarrow$

	1	**2**	**3**
2	0.8		
3	1.0	1.0	
4	0.8	1.0	0.5

Noted that the dissimilarity between subject 1 and subject 3 is perfectly captured by the Jaccard distance, whereas the Manhattan distance does not capture the fact that subjects 1 and 3 mismatch on every question. Using the above definition of the Jaccard distance, we performed hierarchical clustering on the whole data. Also, the Jaccard distance is particularly appealing because it has been shown that it has probabilistic basis. See Real and Vargas (1996) and Jousselme et al. (2001). Table 11 provides the details of the clustering results.

	`Cluster 1`	`Center 2`	`Center 3`
`Average of Center`	1.03	3.33	3.95
`Median of Median`	1.00	3.00	4.00
`Number of Observations`	659	4100	1061
`Percentage of observations`	11.32%	70.45%	18.23%
`Suggested class label`	`Dissatisfied`	`Neutral`	`Satisfied`

Table 11: *Clusters extracted using hierarchical clustering on binarized version of the data. For this binarized version of the data, we used the Jaccard distance along with Ward linkage.*

III.2 Are there meaning concepts underlying the items of the evaluation?

It goes without saying that 28 questions for a single respondent can be quite overwhelming. Besides, it's indeed very likely that many of the questions end measuring the same aspect of the perception of the student. Recall for instance that the correlation matrices calculated earlier revealed extremely large correlation values. We should therefore expect the 28 dimensional questionnaire given to students to boil down to a much lower number of latent concepts. From a factor analytic perspective, this means that the student evaluation vector $X^\top = (X_1, \cdots, X_{28})$ does have a representation of the form

$$X = \mathbf{\Lambda} Z + \boldsymbol{\epsilon}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{28 \times q}$ and $Z^\top = (Z_1, \cdots, Z_q)$ for some $q \ll 28$. Factor Analysis typically assumes that the factor scores vector Z has a multivariate Gaussian (normal) distribution. Such an assumption is bound to be violated here because of the non-normality of the vector X. Many authors have performed factor analysis on Likert-type data despite this non-normality. Muthen and Kaplan (1992) and Lubke and Muthen (2004) provide a detail account of the pitfalls resulting from the misuses of factor analysis on Likert-type data. It turns out that part of the problem with the use of factor analysis on Likert-type data stems from the fact that some analysts use the Pearson covariance matrix as their main ingredients. To somehow avoid the pitfalls and hope for meaningful factor analytic results, we use the Kendall τ-B correlation matrix as the basis of our factor

analysis. Based of Table 12 our factor analytic seem to reveal the following facts: Questions 13 to 28 have estimated factor loadings that are all higher on factor 1 than they are on Factor 2. These 18 questions are all related to how the student rate the competence of the instructor teaching the course. We therefore name the first factor score Z_1 the "**instructor rating score**". Questions 1 to 12 have estimated factor loadings that are all higher on factor 2 than they are on Factor 1. These 12 questions are all related to how satisfied the student was about the course. We therefore name the second factor score Z_2 the "**student satisfaction score**".

	Factor 1	Factor 2
Q1	0.376	0.781
Q2	0.495	0.767
Q3	0.567	0.689
Q4	0.475	0.770
Q5	0.505	0.793
Q6	0.497	0.776
Q7	0.465	0.819
Q8	0.456	0.815
Q9	0.545	0.699
Q10	0.524	0.791
Q11	0.564	0.680
Q12	0.486	0.751
Q13	0.753	0.558
Q14	0.794	0.517
Q15	0.791	0.514
Q16	0.705	0.611
Q17	0.827	0.391
Q18	0.762	0.541
Q19	0.790	0.517
Q20	0.825	0.475
Q21	0.844	0.447
Q22	0.846	0.446
Q23	0.756	0.564
Q24	0.713	0.593
Q25	0.826	0.463
Q26	0.749	0.543
Q27	0.695	0.567
Q28	0.811	0.452

Table 12: *Two factor model from the $p = 28$ questions on the student evaluation. There was a total of $n = 5820$ evaluations submitted by the students and used to estimate these factor loadings.*

The two factors described above captured 85% of the variation, and any attempt to generate/derive more factors resulted in very little gain along with the loss of interpretability inherent in these two factors. From a practical perspective, it seems to make sense that a student's answers would be summarized into their overall satisfaction along with some rating of the instructor who led the whole experience on the course. Clearly one could hypothesize more factors, but these two tend to intuitively capture what one would expect.

IV. Supervized Learning Techniques

Our final analysis of this data consists of performing classification using the labels generated by the clustering algorithms. We specifically used classification trees, and then we boosted the trees and finally check the performance of random forests. The great appeal of trees was triggered by our interest in finding out if there were some questions that drove the classification and that could therefore considered somewhat key questions in the evaluation. Classification trees are usually highly preferred by analysts who desire an interpretable learning machine. Understanding trees is indeed straightforward as they are intuitively appealing piecewise functions operating on a partitioning of the input space. Given $\mathcal{D} = \Big\{(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_n, Y_n)\Big\}$, with $\boldsymbol{x}_i \in \mathcal{X}^p$, $Y_i \in \{1, \cdots, g\}$. If T denotes the tree represented by the partitioning of $\mathcal{X}$ into q regions $R_1, R_2, \cdots, R_q$ such that $T = \cup_{\ell=1}^{q} R_\ell$, then, given a new point $\boldsymbol{x}^*$ in node ℓ, its predicted response $\hat{Y}^*$ is

$$\hat{Y}^*_{\texttt{Tree}} = \hat{f}_{\texttt{Tree}}(\boldsymbol{x}^*) = arg \max_{j \in \{1, \cdots, g\}} \Big\{p_{j\ell}\Big\}$$

where

$$p_{j\ell} = \frac{1}{|R_\ell|} \sum_{\boldsymbol{x}_i \in R_\ell} I(Y_i = j)$$

estimates the proportion of node ℓ observations that belongs to class j. Using the three labels obtained from kMeans clustering, namely $\{\texttt{Satisfied}, \texttt{Dissatisfied}, \texttt{Neutral}\}$, we obtained the tree of Figure 13.

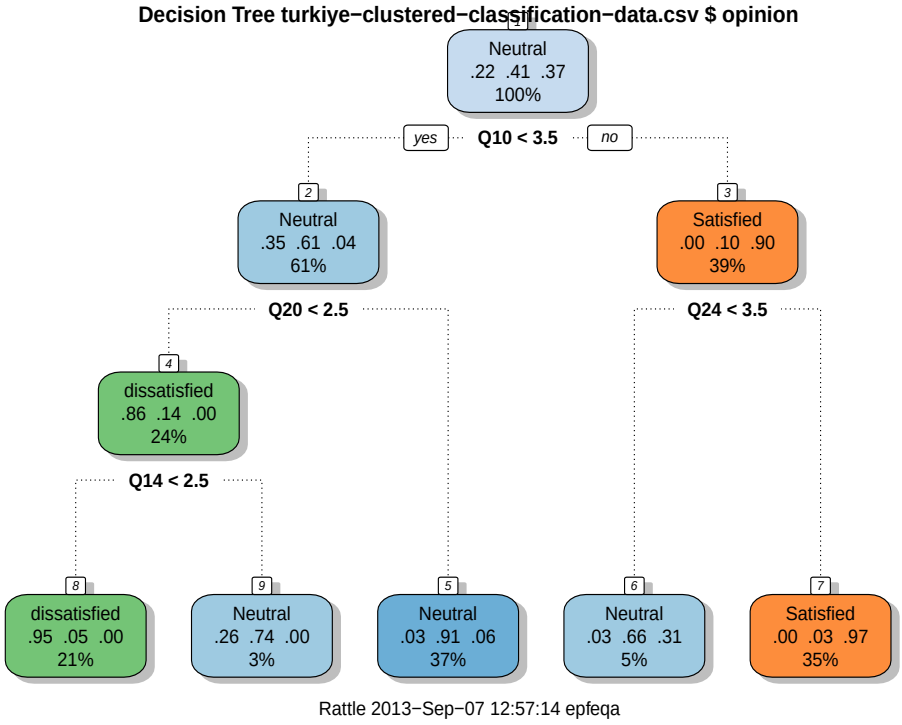

Figure 13: *Tree yielded for the* 3 *class classification of the whole dataset.*

Before we delve into details about what the tree of Figure 13 reveals, it is important to mention that the threshold values are somewhat weird. The root for instance checks `Q10 < 3.5`. Clearly, no student has any mechanism for answering 3.5 to any question. The appearance of 3.5 reveals that the tree method is treating Likert-type scores as if they were real values. We should bear in mind that abuse of variable type while drawing far reaching conclusions from this tree result.
Now, the root of the tree being `Q10` seems to suggest that `Question 10` was found by the tree building algorithm to be one of the variables with the greatest importance. This dominance of `Question 10` on the tree somewhat makes sense, because it is the question that seeks to know whether the student's objectives and goals were satisfied at the end of the school term (period). It is worth mentioning that the remaining variables actually used by the tree are `Q14`, `Q20` and `Q24` which we believe are very important questions when it comes to feedback from students. From a predictive analytics perspective we also generated the confusion matrix on the test set shown in Table 13.

	Dissatisfied	Neutral	Satisfied
Dissatisfied	0.19	0.01	0.00
Neutral	0.01	0.39	0.01
Satisfied	0.00	0.04	0.35

Table 13: *Confusion matrix based on the test set for the three class recognition task.*

Table 13 shows that the test error is 7%, which seems to suggest that the classification is very good at predicting the satisfaction of a student based on the way they answered the survey.
Our next and final task is binary classification, with labels `Dissatisfied, Not Dissatisfied`. Here, the second label, `Not Dissatisfied`, is simply created by merging `Neutral` and `Satisfied`. We did this primarily to generate comparative ROC curves for five different pattern recognition methods, namely *classification trees, Random Forest, Boosting (ada), Support Vector Machines and Neural Networks*. For brevity, we show only the binary tree, but we present the ROC curve revealing the predictive performance of the other methods.
When we had the three class recognition task, the root of the tree was `Question 10`. In this binary classification task, the root is now `Question 20`. Unsurprisingly, the two questions driving this new tree are questions that were in the previous tree. It is re-assuring also to see that `Question 20` acts here in the same manner it did in the previous tree. The huge dimensionality reduction in both trees, namely from 28 variables to 4 and now to 2 is somewhat to be expected giving the extremely high correlations encountered earlier.
The ROC Curve reveals that all the other four methods produce a better predictive performance than the single classification tree. It suffices to note that the satisfaction level of a student on a course can be predict quite accurately based on their answers.

V. Conclusion and Discussion

We have provided a comprehensive statistical analysis of a relatively large dataset containing students' evaluations of various courses at a university in Turkey.
Factor Analytic results appear to reveal a very plausible two factor model suggesting that students' evaluations inherently reveal the overall satisfaction of the student at the end of the course along with impact the instructor had on their overall satisfaction. With the instructor's factor coming out as the most dominant one, it is fair to say that the instructor does play a central role in the over experience of the student.

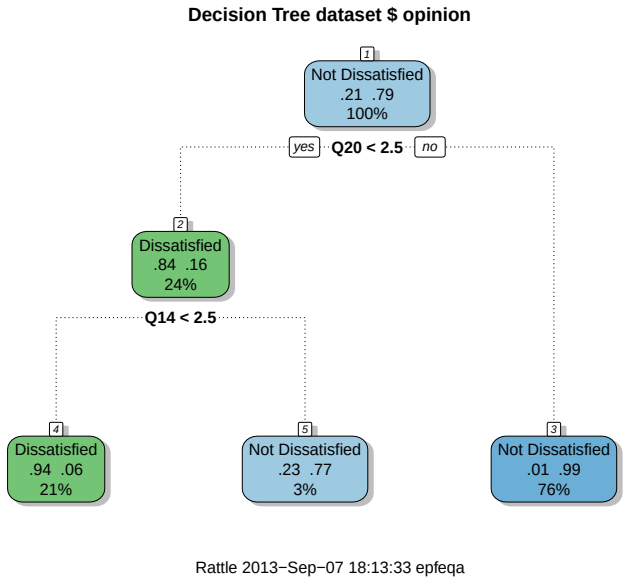

Figure 14: *Tree yielded on the binary classification of our students into* `Dissatisfied` *and* `Not Dissatisfied`

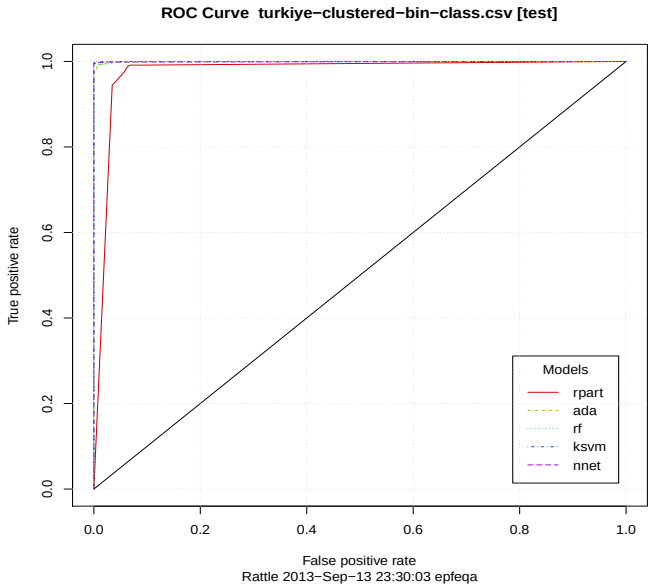

Figure 15: *ROC curve comparing the five classification methods used, namely Classification trees, Random Forest, Boosting (ada), Support Vector Machines and Neural Networks.*

Anyone analyzing students' evaluations should be careful to consider the number of *zero variation* responses and examining their association with the pattern of answers provided by the students. We strongly believe that these *zero variation* responses somewhat determine the quality of the survey and reliability of the answers provided.

We have shown evidence to support the fact dedicated students (attendance) will tend to reveal a more satisfactorily learning experience than those students who do not take their course seriously. We combined unsupervised and supervised learning techniques and were able, not only to find meaningful and interpretable groups in the data, but also identify the items in the questionnaire that appeared to be driving the students' assessment of their learning experience.

The dominance of `Question 10` on the three class recognition tree confirmed our intuition in the sense that it is the question that seems to measure the overall satisfaction of a student on a course, and it is re-assuring to have the tree model reveal it.

From a questionnaire design perspective, it is our view that 28 questions is a bit too much for the students, and this usually large survey might be the reason why some students ended up giving *zero variation* responses. We would also like to suggest the use of two questions that have been found to be very revealing of the experience of student, namely (1) *What is your overall rating of this instructor?* (2) *Would you recommend this course to any other student?*. Although these two questions are inherently correlated

Part of our future work on this data will consist of adapting traditional classification trees to Likert-type data. This essentially boils down to using Likert-type specific loss functions for splitting the nodes of the tree. We specifically plan on deriving adaptations of the Jaccard distance as loss function or using the cross entropy measure on the tendencies of respondents.

We are also planning to include the final grades of the respondents. The motivation for this is the fact that many instructors around the world have repeatedly argued that students who know (based on quiz scores, homework assignment scores, and midterm exam scores) that they will be receiving a good grade on the course tend to rate their professors very highly. We plan on finding out if there is evidence to support such a belief.

References

Adams, E. W., R. F. Fagot, and R. E. Robinson (1965, June). A theory of appropriate statistics. *Psychometrica 30*(2), 99–127.

Allen, I. E. and C. A. Seaman (2007). Likert scales and data analyses. *47*, 419–442.

Boone, H. and D. A. Boone (2012). Analyzing Likert Data. *Journal of Extension 50*(2).

Clason, D. L. and T. J. Dormody (1994). Analyzing Data Measured by Individual Likert type items. *Journal of Agricultural Education 35*(4), 31–35.

Jamieson, S. (2004). Likert Scales: How to (ab)use Them. *Medical Education 38*(38), 1212–1218.

Javaras, K. N. (2004, Hilary Term). *Statistical Analysis of Likert Data on Attitudes*. Ph. D. thesis, Balliol College, University of Oxford.

Jousselme, A. L., D. Granier, and P. E. DorÃl'c (2001). A new distance between two bodies of evidence. *Information Fusion 2*, 91–101.

Likert, R. (1932, June). A Technique for the Measurement of Attitudes. *Archives of Psychology 22*(140), 5–55.

Lubke, G. and B. Muthen (2004). Factor-analyzing Likert scale data under the assumption of multivariate normality complicates a meaningful comparison of observed groups or latent classes. *Structural Equation Modeling 11*, 514–534.

Muthen, B. and D. Kaplan (1992). A comparison of some methodologies for the factor analysis of non-normal likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology 45*, 19–30.

Narli, S. (2010, March). An alternative evaluation method for likert type attitude scales: Rough set data analysis. *Scientific Research and Essays 5*(6), 519–528.

Real, R. and J. M. Vargas (1996). The probabilistic basis of jaccard's index of similarity. *Systematic Biology 45*(3), 380–385.

Sisson, D. and H. Stocker (1989). Analyzing and interpreting likert-type survey data. *The Delta Pi Epsilon Journal 31*(2), 81–85.

Wu, C.-H. (2007). An Empirical Study on the Transformation of Likert-scale Data to Numerical scores. *Applied Mathematical Sciences 1*(58), 2851 – 2862.

- Q1: The semester course content, teaching method and evaluation system were provided at the start.
- Q2: The course aims and objectives were clearly stated at the beginning of the period.
- Q3: The course was worth the amount of credit assigned to it.
- Q4: The course was taught according to the syllabus announced on the first day of class.
- Q5: The class discussions, homework assignments, applications and studies were satisfactory.
- Q6: The textbook and other courses resources were sufficient and up to date.
- Q7: The course allowed field work, applications, laboratory, discussion and other studies.
- Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
- Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
- Q10: My initial expectations about the course were met at the end of the period or year.
- Q11: The course was relevant and beneficial to my professional development.
- Q12: The course helped me look at life and the world with a new perspective.
- Q13: The Instructor's knowledge was relevant and up to date.
- Q14: The Instructor came prepared for classes.
- Q15: The Instructor taught in accordance with the announced lesson plan.
- Q16: The Instructor was committed to the course and was understandable.
- Q17: The Instructor arrived on time for classes.
- Q18: The Instructor has a smooth and easy to follow delivery/speech.
- Q19: The Instructor made effective use of class hours.
- Q20: The Instructor explained the course and was eager to be helpful to students.
- Q21: The Instructor demonstrated a positive approach to students.
- Q22: The Instructor was open and respectful of the views of students about the course.
- Q23: The Instructor encouraged participation in the course.
- Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students.
- Q25: The Instructor responded to questions about the course inside and outside of the course.
- Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.
- Q27: The Instructor provided solutions to exams and discussed them with students.
- Q28: The Instructor treated all students in a right and objective manner.

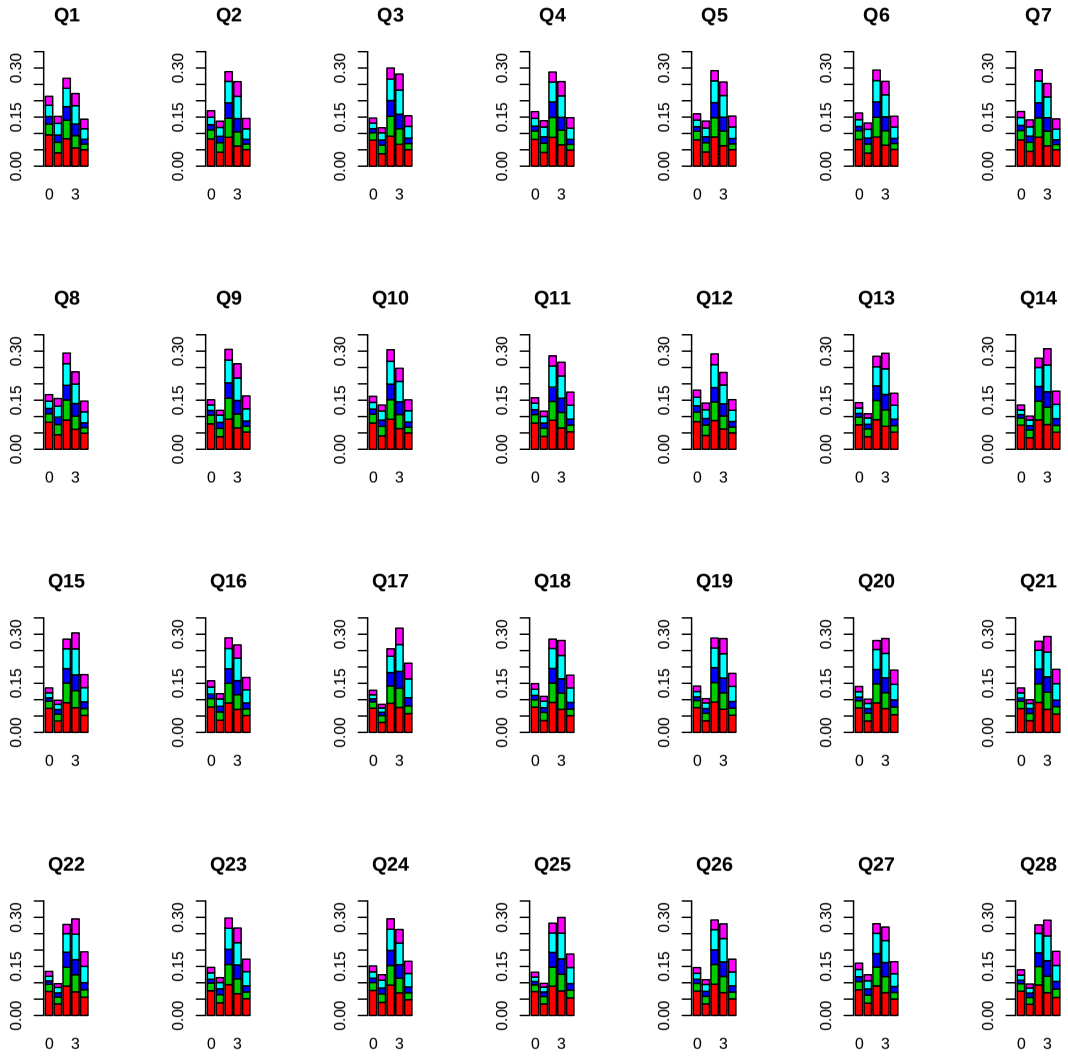

Figure 16: *Barplot of the cross-tabulation of attendance level versus scores on all the questions.*

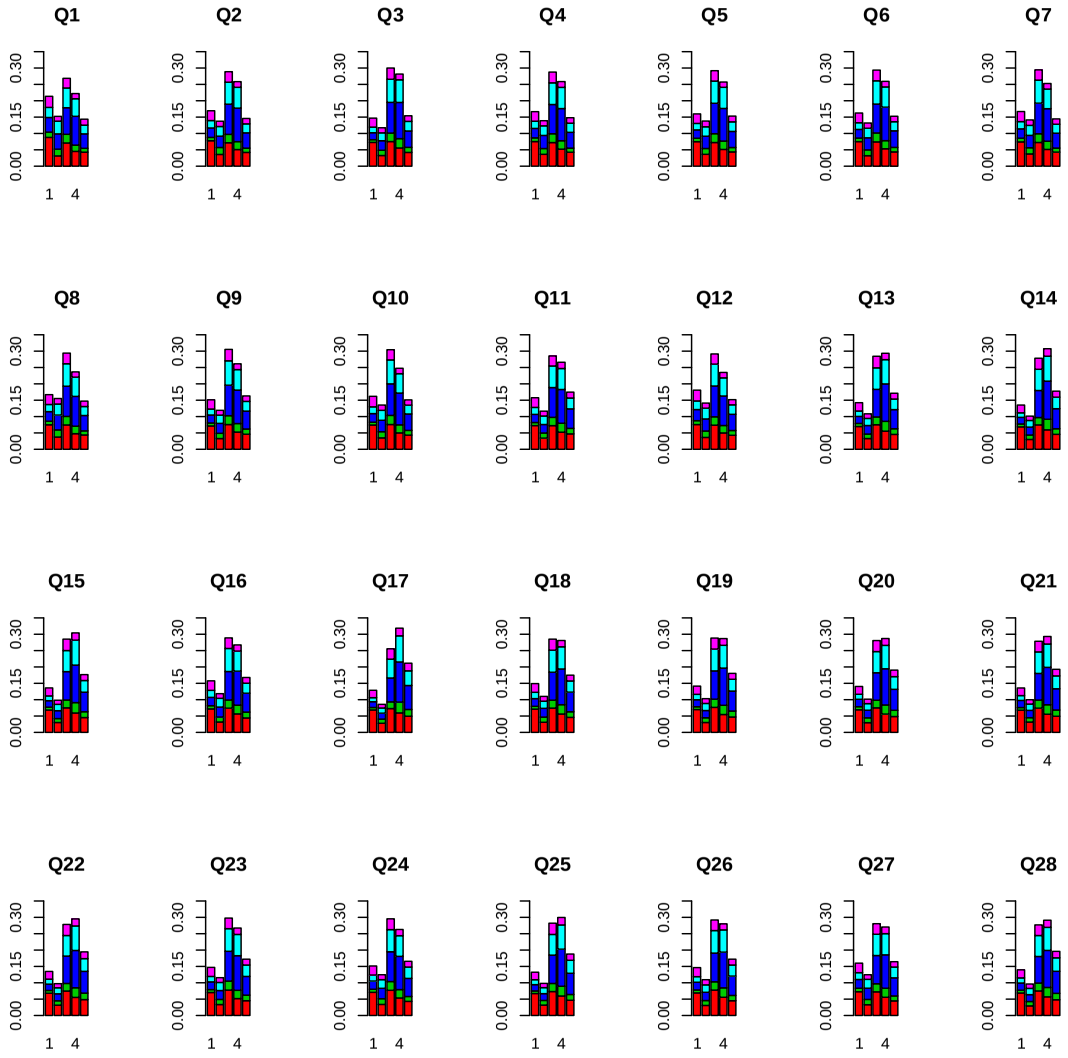

Figure 17: *Barplot of the cross-tabulation of difficulty level versus scores on all the questions.*