DATA ANALYTICS INVOLVING THE USE OF R

DESCRIPTIVE ANALYSIS

Introduction

This Section requires one to use inferential statistics and dimension reduction approaches in order to extract components from the Turkiye Student Evaluation dataset and use the component scores in several analyses. Further, the section needs one to critically analyze the results, manage and manipulate the provided data, as well as illustrate the findings using various visualization. The process uncovers the properties of the dataset via data transformation and management procedures using R software to provide relevant recommendations.

Procedure

This analysis uses descriptive statistics, principal components, and dimensional reduction on survey data. This dataset contains 28 variables which are marked using a Likert scale. The descriptions of variables can be seen in the same link identified as column 6:33 in R. The study chose a multivariate analyses of variance (MANOVA) test to reveal the effect the independent variables have on the dependent variables Thus, several parametric assumption were tested to confirm whether the appropriate test was carried out. “QQplot”, “leveneTest”, and “boxplot” packages were adopted for conducting homogeneity of variance and normality checks which were used to quantitatively and qualitatively analyse the dataset. The dataset used is called Data1 (1).csv on R studio.

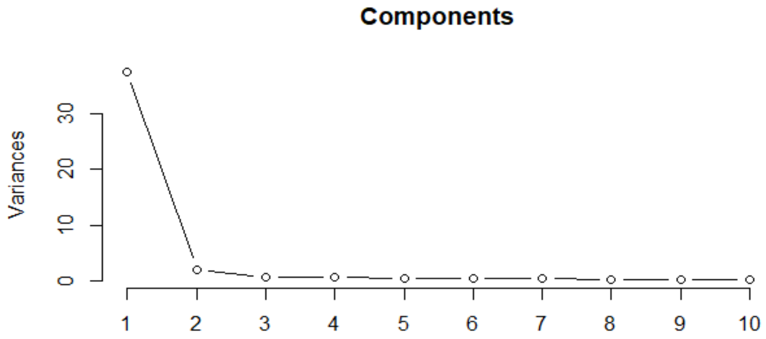
The dataset provided had over 5000 entries which can be regarded as large. As a result the initial stage, the Principal Component Analysis (PCA) and a Paran was conducted to reduce the dimensionality of the dataset. Principal component analysis for the variables was useful in identifying specific constructs or measures of the dataset. In this case, the approach identifies whether the constructs or variables are a subset of another construct in a stream of questions or if they stand on their own. When it is found that the attributes are components of other variables, they are integrated to form a single variable for analysis (Healy 2018). This will reveal what specific variables are critical for statistical analysis. The Para analysis as well as a screeplot was created for PCA. The Paran gives good results (Reilly & Eaves, 2000; Sarff, 1997; Velicer, Eaton, & Fava, 2000; Wang, 2001; Zwick & Velicer, 1986).

Dimension Reduction of Variables using Principle Component Analysis

Before conducting the PCA test, a test was carried out to determine if there is indeed a need to perform a dimensionality reduction on the dataset. The test conducted was the Keiser-Meyer-Olkin test. The test showed an overall Measure of Sampling adequacy (MSA) as 0.99 which means that there is enough reason for a dimensionality reduction to be performed on the dataset and a PCA can be used on the dataset.

Deciding the number of factors is a critical decsions is more important than other decisions {Çokluk, Ö. and Koçak, D. (2016)}. To check the number of components we should retain, several methods like (the qualitative assessment) scree plot, parallel analysis (Advanced statistical method) was conducted. The Parallel analysis by Horn (1965) is one of the several other methods which can be used to determine the number of factors which gives good results.

Figure 1: Principal Components Scree Plot



The scree plot in figure 1 retained to components 1 & 2 as being most significant to the 28 questions. The scree plot showed an 'elbow' shaped scatterplot from component 1 to component 2, whereas all other components had a similar variance. The parallel analysis method confirms the two components. On {library (Paran)} function, the parallel analysis suggests that two (2) components were retained.

Following this step, we derived the correlation of each question with the identified components, in order to that the library (psych) package was used. Based on the items loading scores, RC2 had a higher correlation matrix to Q1 – Q12 while RC1 had a higher correlation matrix to Q13 – Q28. Based on the question, it is quite clear that Component 2 (RC2) focuses more on the course content and Component 1 (RC1) focuses on the instructors. Therefore, we labelled RC1 as instructors' performance and RC2 as course satisfaction. Thereafter, we generated the components scores of each of the 5820 participant and added it to the data set for analysis.

Furthermore, most items loading for the data set were above 0.4 which means that all items are relevant that there is no need to remove any on the items loaded (Steven, 2002), see the appendix 2.0 for more information.

Given that the dataset has several independent and dependant variables a MANOVA test was suggested. For the analysis the following variables were selected:

- Independent variables – Instructors, difficulty and Number of Repeats
- RC1 {Instructors Performance} – Dependent variable (numeric)
- RC2 { Course satisfaction}– Dependent variable(numeric)

The following question will be the hypothesis and focal point of this analysis; is there an effect of instructors and difficulty on student satisfaction?

H1: There is an effect of instructor, difficulty and number of repeats on RC1 and RC2

RC1, RC2 ~ Instr*Difficulty*Nb.Repeats

Parametric Assumptions for MANOVA

In order to know if a Manova is the appropriate test to carry out, certain parametric assumption should be met.

Normality Check

Firstly, Shapiro Wilik Test does not respond well with large data set therefore QQnorm was used for checking for normality. Due to the large data size, we use a qualitative approach to check for normality which is based on the QQplot. Based on Figure 2.0 & 2.1, the graph shows that the correlation between the instructors performance are fairly normal distribution, also figure 2.1 shows that the correlation between the course satisfaction and all independent variables are fairly normal.

Figure 2.0

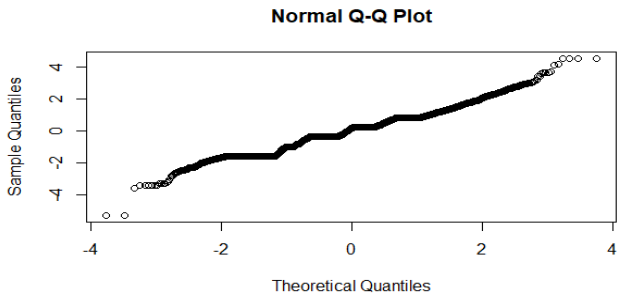


Fig 2.0

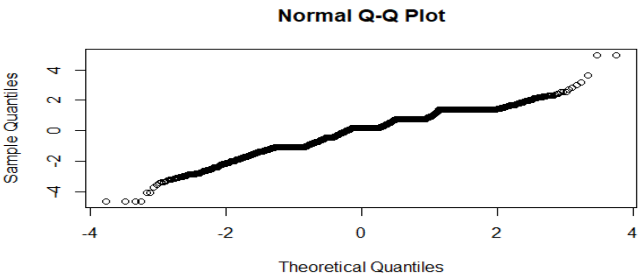


Fig 2.1

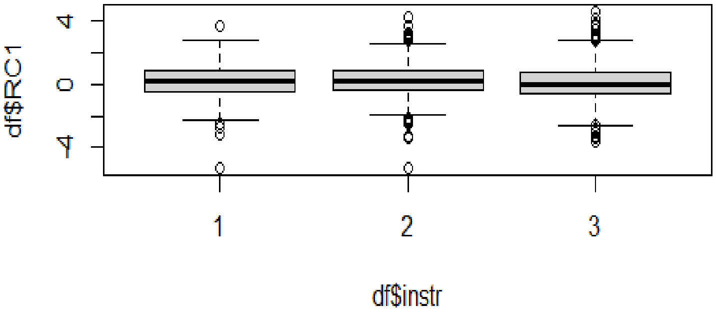
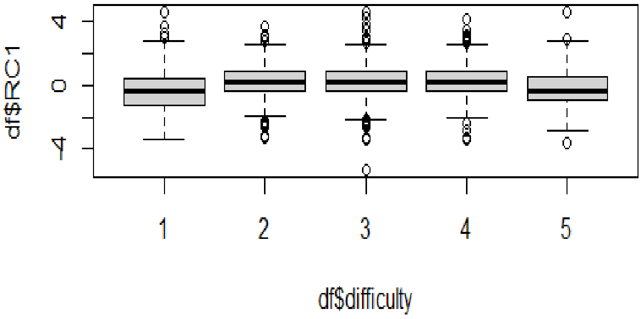
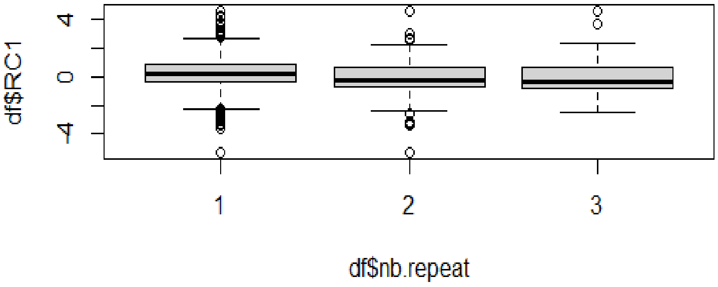
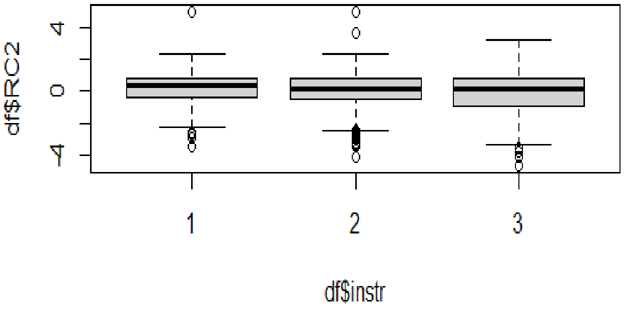
Homogeneity of variance

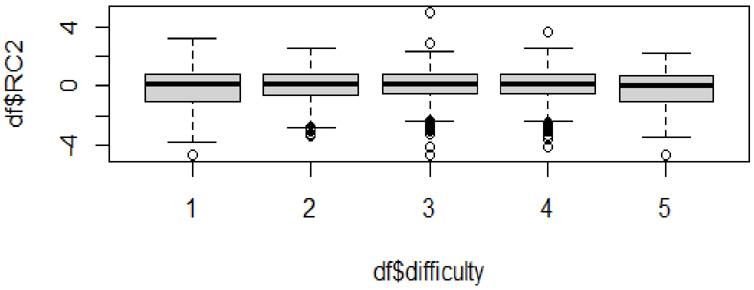
The next assumption to test is the homogeneity of variance which means that mean of the dataset are the same. To check for homogeneity the LeveneTest was carried out because it is ideal to check the quantitative analysis for homogeneity. Usually when Levene test is carried out on large dataset such as this one, the parametric are violated as seen in this dataset. It should be viewed with caution. Based on the result the LeveneTest for homogeneity of variance was significant for instructors performance on instructors [F (2, 5817) = 7.8481, P<0.01. Also, for Instructors performance on difficulty the parametric assumption was violated [F (4, 5815) = 13.915, P<0.01.However, the effect on instructors performance on number of repeats was not significant [F (2, 5817) = 0.0914, P>0.01.

Regarding R2 which focuses on course satisfaction, the homogeneity of variance was not significant as it approximately equal to 0.01. [F (2, 5817) = 4.571, P=0.01, therefore the variance is not different from each other and the parametric was not violated. Conversely, the effect on course satisfaction on difficulty was significant [F (4, 5815) = 11.244 P<0.01. Finally, the homogeneity of variance of course satisfaction on number of student repeats is not significant [F (2, 5817) = 0.3571 P>0.01, indicating that the parametric assumption was violated.

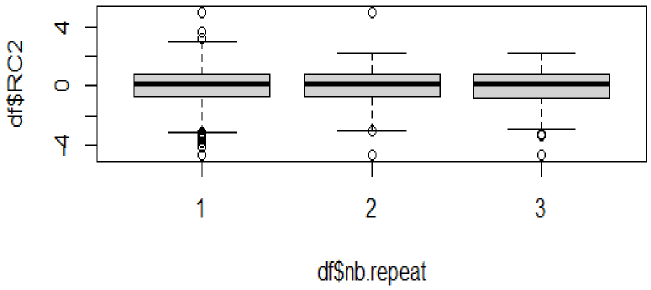
Visual Representation of Homogeneity of Variance.

The LeveneTest a quantitative result but to visualize this data Boxplot was plotted. The boxplot shows the spread of the data between the dependent variables and independent variables. The boxplots below show similarities between the dependent variables and independent variables.

 BoxPlot 1	 Box plot 2
 Box Plot 3	 Box Plot 4



Box Plot 5



Box Plot 6

Boxplot 1, the spread especially similar between instructors performance with instructors, this was a quite obvious similarity. Boxplot 2 shows that the spread between the instructor's performance and the levels of difficultly are similar. The outliers (observation distant from the rest of the data) were more prominent in level 2, 3 & 4 compared to level 1 & 5. Boxplot 3 shows a similar spread between the range numbers. Majority of the outliers were on number 1. Boxplot 4 shows a similar spread was similar between course satisfaction and instructors. Instructors 1 & 2 are similar compared to 3, most of the outliers are in instructor 2. Boxplot 5 shows an also similar spread between course satisfaction and difficulty, most outliers where in levels 3 & 4. Box plot 6 shows a similar spread between the range numbers. Majority of the outliers were on number 1.

MANOVA: Omnibus Test

```
                Df   Pillai    approx F  num Df den Df   Pr(>F)
df$instr         2   0.027219  40.088     4      11622 < 2.2e-16 ***
df$difficulty    4   0.052011  38.788     8      11622 < 2.2e-16 ***
df$nb.repeat     2   0.005587   8.139     4      11622 1.502e-06 ***
Residuals     5811
```

For the omnibus test MANOVA showed a highly significant main effect of instructors [F(2,5811) = 40.1, p < 0.001,] , difficulty [F(4,5811) = 38.8, p < 0.001,] and [F(2,5811) = 8.1, p < 0.001]

Furthermore individual ANOVAS are carried out to test the significance of each independent variable on each of the dependent variables.

Individual ANOVA
Further Investigation based on the individual ANOVAS indicates that:

Response 1 with Eta-Squared η2:

- [F(2,5811) = 22.9, p < 0.001, **η2 =** 0.003] RC1 on Instructors
- [F(4,5811) = 67.3, p < 0.001, **η2 =** 0.047] RC1 on Difficulty
- [F(2,5811) = 15.6, p < 0.001, **η2 =** 0.005] RC1 on Nb. Repeats

Response 2:

- [F(2,5811) = 56.9, p < 0.001, **η2 =** 0.016] RC2 on Instructors
- [F(4,5811) = 11.1, p < 0.001, **η2 =** 0.007] RC2 on Difficulty
- [F(2,5811) = 0.8, p < 0.001, **η2 =** 0.00] RC2 on Nb. Repeats

All the independent variables had a significant effect of the dependent variables listed however, that course satisfaction had an almost insignificant effect on the number of repeats. In addition to this, the effect size is added to the ANOVAS using the Eta-squared to measure the effect. The effect size shows the number of variances explained, the lower the variance the model captures, the lesser the effect.

Confidence Intervals

The confidence intervals shows the how confident of the population parameter. A common confidence level is 95% which means that there is a 95% confidence of the true mean in the population. Based on the Figures below, the following have been deduced;

In fig 4.1The Confidence interval between each instructors do not overlap significantly therefore, we cannot be confidence that they do not share the same population mean. However, instructors two (2) had the most effect on instructors on instructor’s performance (RC1).

In Fig 4.2 Between difficulty level 2, 3 and 4 the means overlap, so there is not a significant difference between the groups. The lowest mean is level 1 while, the level 3 which had the most effect on instructor performance.

In Fig 4.3 The confidence interval between the repeat times 1 and 2 on Instructor performance do not overlap, so we can be 95% confidence that they do not share the same population mean however, the number two and three do overlap therefore is not much confidence.

In Fig 4.4 Based on the plot none of the instructors overlap. Instructor one had the most effect on course satisfaction having the highest mean. There is a 95% confidence that none of the instructors share the same population mean with regards to course satisfaction.

In Fig 4.5 The confidence interval between levels 1 and 2 overlap, between level 3 and 4 there is a slight shift below, however, level 4 and 5 do not overlap. Level 3 of difficulty had the most effect on course satisfaction.

In Fig 4.6 the confidence intervals do not overlap, though it is a slight fluctuation between the mean groups. The assumption is that the means of all group is significant. The third time student repeat the course has the most effect on how satisfied they are with it.

Fig 4.1

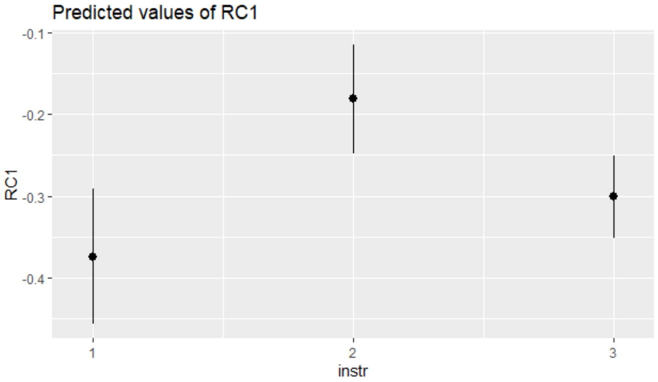
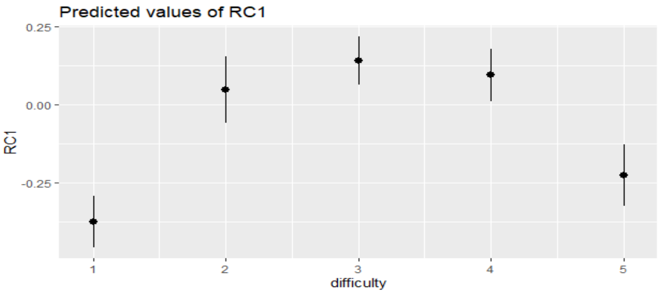


Fig 4.2



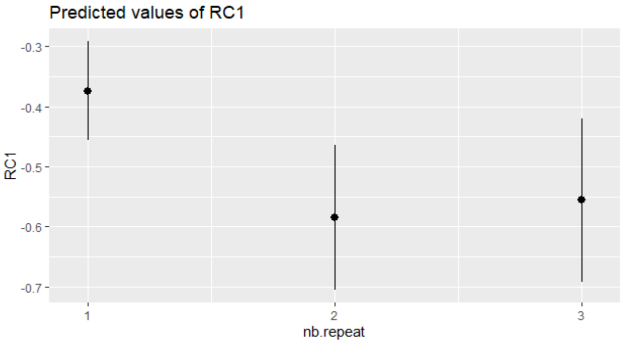


Fig 4.3

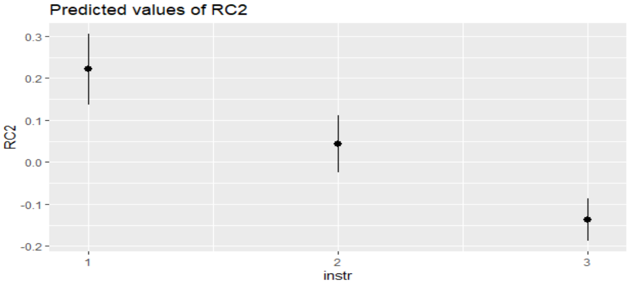


Fig 4.4

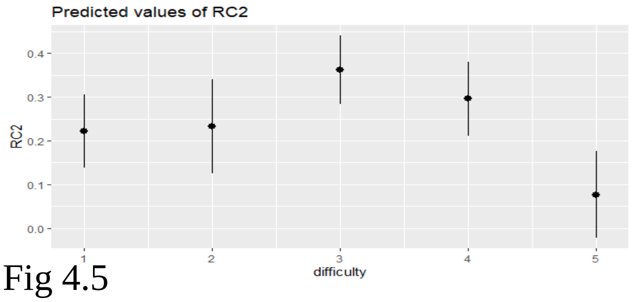


Fig 4.5

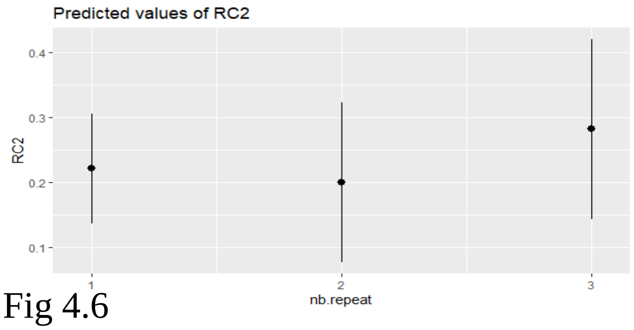


Fig 4.6

Post Hoc Test & Tukey Interpretation

Following this all individual ANOVAS were significant so the post-hoc is necessary to be tested. Tukey's HSD procedure provides the preferred method post-hoc analysis for allowing all possible pairwise comparison, it also to show where there is significance (Zach (2020) How to perform Tukey's test in R). For this reason the post hoc test using Tukey was conducted.

TUKEY Interpretation

(RC1 ~ instr + difficulty + nb.repeat) – Significant Comparision

Paired Levels	RC1 ~ Instructors Interpretation	Significance Level
2--1	0.183(0.084, 0.285), P<0.05	Significant
3--2	-0.203(-0.274, -0.132), P<0.05)	Significant
	RC1 ~ Difficulty	
2--1	0.363(0.232, 0.494), P<0.05	Significant
3--1	0.477(0.386, 0.568), P<0.05	Significant
4--1	0.443(0.343, 0.544), P<0.05	Significant
5--2	-0.243(-0.397, -0.090) P<0.05	Significant
5--3	-0.357(-0.479, -0.236) P<0.05	Significant
5--4	-0.323(-0.452, -0.195) P<0.05	Significant
	RC1 ~ Nb.repeat	
2--1	-0.200(-0.301, -0.100) P<0.05	Significant
3--1	-0.172(-0.301, -0.044) P<0.05	Significant

RC2 ~ instr + difficulty + nb.repeat, data = df)

	RC2 ~ Instructors	
	Interpretation	**Significance level**
2--1	-0.171(-0.274, -0.068) P<0.05	Significant
3--1	-0.376(-0.467, -0.284)	Significant
3--2	-0.205(-0.277, -0.133) P<0.05)	Significant
	RC2 ~difficulty	
3--1	0.136(0.044, 0.229) P<0.05	Significant
5--1	-0.141(-0.266, -0.016) P<0.05	Significant
3--2	0.131(-0.001, 0.263) P<0.05	Significant
5--3	-0.278(-0.401, -0.154) P<0.05	Significant
5--4	-0.215(-0.346, -0.085) P<0.05	Significant
	RC2 ~ nb.repeat	None significant