Statistical Methods – COSC 6323 - HomeWork-12

By Dinesh Narlakanti (2083649)

INTRODUCTION

Students of the class COSC 6323 are asked to fill the Class Questionnaire before submission the assignment that released after every class. Questionnaire includes question about the usefulness of the class, difficulty of the assignment and feedback of the lecture.

This documents is focused on:
i) Checking whether there is correlation between sentiments about the usefulness of the class session and the difficulty of the associated assignment.
ii) Checking is there any significant differences in the students' sentiments at the beginning, in the middle, and towards the end of the semester.
iii) Constructing the word cloud of the free form student responses.

Steps Description

1. Installed packages to read excel file and stored data into a variable 'student_data'
2. Renamed column names as 'Survey Number', 'Q1', 'Q2', "Feedback".
3. Converting likert data to numberic for the Question 1 and 2. Strongly Disagree (1) to Strongly Agree (5) and Very Easy (1) to Very Difficult (5).
4. Performed spearman correlation test on both the scaled values to find the existence of correlation between them.
5. Subsetted session 1,6, and 11 to find the sentiment of students at the beginning, middle and end of the semester.
6. Checked normality of the numerically coverted scales of the sub setted data.
7. Performed Kruskal test to check whether there is difference in the sentiments of the students on the above subsetted data.
8. Installed packages related to text mining, test stemming, word cloud and color palettes.
9. Assigned feedback column to varialbe 'text'.
10. Loaded the text as a corpus.
11. Transformed all the data to lower case.
12. Removed the common engligh stopwords.
13. Removed punctuation.
14. Removed the words: "unanswered", "can", "get","null".
15. Stripped off the white spaces.
16. Building term documentation matrix by changing term document matrix into matrix, sorting and making it into a dataframe.
17. Generating word cloud.

SPEARMAN TEST RESULTS:

```
##
##  Spearman's rank correlation rho
##
## data:  student_data$Q1_scale and student_data$Q2_scale
## S = 363634, p-value = 0.02595
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##        rho
## -0.2016159
```

SPEARMAN TEST INTERPREATION:

The Spearman Correlation coefficient ranges from -1 to 1. Coefficient we obtained is -0.2016159.
Since the coefficient value is negative, we can say that there is negative correlation between the Questions: Usefulness and difficulty of assignment. In other words, as the the usefulness of the class increases, the difficulty level of the assignment decreases. We can also interpret that the students who were able to understand the class, felt the given assignment was easy.

The p-value here is 0.02595, which is less than 0.05 significance level. So, we can reject the null hypothesis and conclude that there is significant (negative) correlation between usefulness of the class and difficulty of the assignment.

SHAPIRO TEST RESULTS:

```
##
##  Shapiro-Wilk normality test
##
## data:  beginning_middle_end$Q1_scale
## W = 0.74859, p-value = 1.801e-06
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beginning_middle_end$Q2_scale
## W = 0.719, p-value = 5.702e-07
```

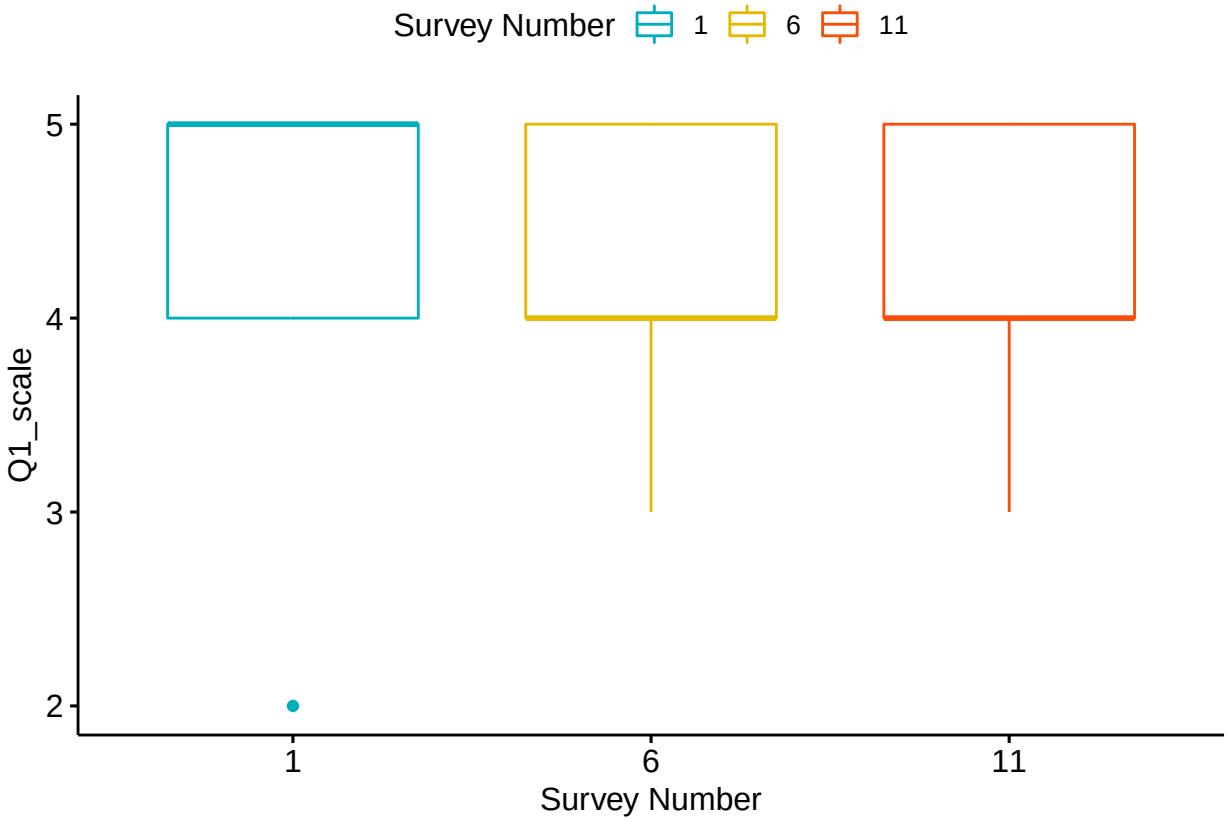
SHAPIRO TEST INTERPRETATION:

When observed, the p-values for both the scales after sub setting is less than 0.05. So, we can conclude that they are not normally distributed. As they are not normal, instead of anova test, Kruskal-Wallis test is more appropriate for analyzing the difference among the sessions.

KRUSHKAL-WALLIS TEST RESULTS:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  beginning_middle_end$`Survey Number` by beginning_middle_end$Q1_scale
## Kruskal-Wallis chi-squared = 4.184, df = 3, p-value = 0.2423
```

PLOTS TO SUPPORT KRUSKAL TEST:

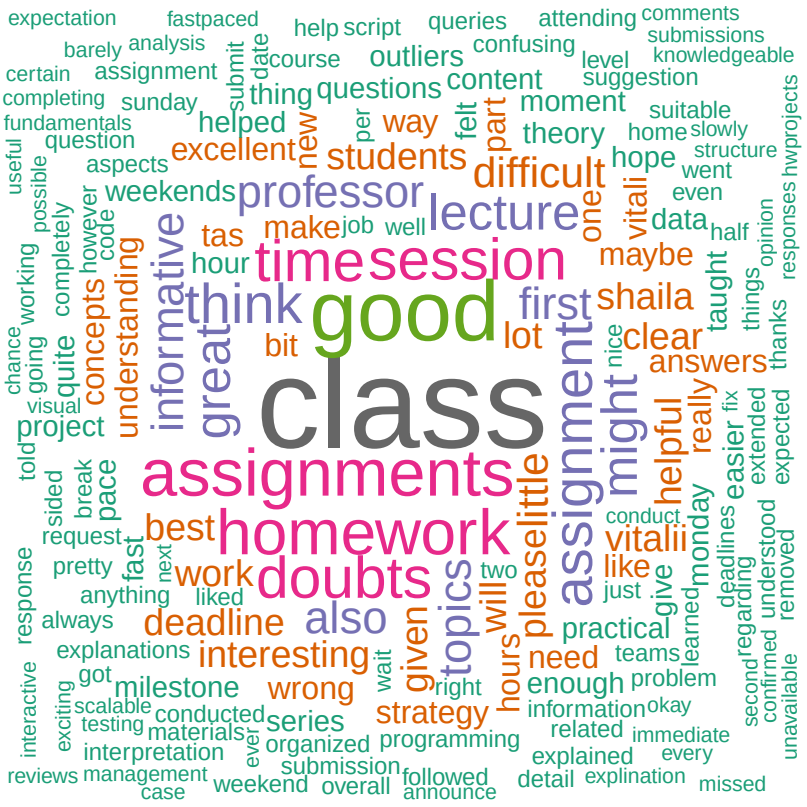


KRUSHKAL TEST and PLOT INTERPREATION:

As the p-value obtained here is 0.2423, which is greater than 0.05. So, we can accept the null hypothesis and conclude that there is no significant difference in the students' sentiments at beginning, middle and end of the semester.

Even when plotted, we can clearly observe that there is no difference among the ratings of sessions 1,6, and 11. Although, there is slight difference in the mean ratings but still it is not a significant difference.

WORD CLOUD OF THE STUDENT RESPONSES:



CONCLUSIONS DRAWN FROM THE WORD CLOUD:

1. The two most prominent words in the word cloud are 'class' and 'good'. We can interpret that most of the students reviewed that class was good.
2. From other highlighted words like 'assignments', 'homework', 'doubts', 'sessions', we can say that the theme of the questionnaire is about assignment/homework and feedback included not only about class but also about doubt clearing sessions.
3. Other words repeated throughout are 'vitalii' and 'shaila'. This indicated that students also gave feedback about TAs.
4. There are some appreciation words included in higher order and they are 'interesting', 'great', 'informative', 'understanding'. This tells that students appreciated the work of professors and TAs.
5. When checked the associated words with good at the correlation limit of 30% in the feedback, 'everything' and 'vitalii' matched with 47% and 31% respectively. We can interpret that Students almost liked everything about the course and also students gave good feedback about vitalii.
6. Frequent words that are repeated more than 8 time are:

```
##  [1] "class"       "good"        "great"       "session"     "lecture"
##  [6] "assignment"  "might"       "think"       "assignments" "time"
## [11] "doubts"      "homework"
```

7. Association with one of the most frequent words 'class' at 40% correlation limit is:'new'

```
## $class
```

```
##  new
## 0.47
```

This tells that some of the students feel new topics are introduced in the class.
8. Top 10 words repeated in the feedback are: