**BIG DATA FUNDAMENTALS - DATA STORAGE NETWORKING**

**MACHINE LEARNING   
AND  
DATA ANALYSIS**

**(UNSUPERVISED)**

Gazi University Student Evaluation Questionnaire

**Submitted by**

Anju Thankachan C0746314

Arvind Damodaran C0753532

Sai Balakrishna Sai Prasad C0754209

Sree Mahalakshmi Sivaprakasam C0748070

Nittin Garg C0747736

**CONTENTS**

1. DATASET INFORMATION 3
2. ANALYSIS TOOLS 4
3. DATA CLEANSING 5
4. EXPLORATORY DATA ANALYSIS 6
5. FEATURE ENGINEERING 9
6. INSIGHTS 11

**DATASET INFORMATION:**

The dataset provides detailed information of evaluation scores given by the student from Gazi university in Ankara, Turkey. The main aim of the project is to analyze the various courses handled by instructors and the students’ feedback on the subject and the instructors’ way of subject handling. This dataset is multivariate which is obtained through the quantitative analysis using the questionnaire method, holds 28 course specific questions and 5 additional attributes which adds up to 33 variables and the descriptions are given below:

1.**instr**: Instructor’s identifier

values taken from {1,2,3}   
2.**class**: Course code (descriptor)

values taken from {1-13}   
3. **Repeat**: Number of times the student is taking this course

values taken from {0,1,2, 3,...}   
4.**attendance**: Code of the level of attendance

values from {0, 1, 2, 3, 4}   
5. **Difficulty**: Level of difficulty of the course as perceived by the student

values taken from {1,2,3,4,5}   
6. **Q1**: The semester course content, teaching method and evaluation system were provided at the start.   
7. **Q2**: The course aims, and objectives were clearly stated at the beginning of the period.   
8. **Q3**: The course was worth the amount of credit assigned to it.   
9. **Q4**: The course was taught according to the syllabus announced on the first day of class.   
10. **Q5**: The class discussions, homework assignments, applications and studies were satisfactory.   
11. **Q6**: The textbook and other courses resources were enough and up to date.   
12. **Q7**: The course allowed field work, applications, laboratory, discussion and other studies.   
13. **Q8**: The quizzes, assignments, projects and exams contributed to helping the learning.   
14. **Q9**: I greatly enjoyed the class and was eager to actively participate during the lectures.   
15. **Q10**: My initial expectations about the course were met at the end of the period or year.   
16. **Q11**: The course was relevant and beneficial to my professional development.   
17. **Q12**: The course helped me look at life and the world with a new perspective.   
18. **Q13**: The Instructor’s knowledge was relevant and up to date.   
19. **Q14**: The Instructor came prepared for classes.   
20. **Q15**: The Instructor taught in accordance with the announced lesson plan.   
21. **Q16**: The Instructor was committed to the course and was understandable.   
22. **Q17**: The Instructor arrived on time for classes.   
23. **Q18**: The Instructor has a smooth and easy to follow delivery/speech.   
24. **Q19**: The Instructor made effective use of class hours.   
25. **Q20**: The Instructor explained the course and was eager to be helpful to students.   
26. **Q21**: The Instructor demonstrated a positive approach to students.   
27. **Q22**: The Instructor was open and respectful of the views of students about the course.   
28. **Q23**: The Instructor encouraged participation in the course.   
29. **Q24**: The Instructor gave relevant homework assignments/projects and helped/guided students.   
30. **Q25**: The Instructor responded to questions about the course inside and outside of the course.   
31. **Q26**: The Instructor’s evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.   
32. **Q27**: The Instructor provided solutions to exams and discussed them with students.   
33. **Q28**: The Instructor treated all students in a right and objective manner.   
**Attribute information for 6-33:**  
Q1-Q28 are all Likert-type, meaning that the values are taken from {1,2,3,4,5}

**ANALYSIS TOOLS:**

The tools used to perform data analysis and manipulations are as follows:

* Python language:

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming. It uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution. Python’s design offers some support for functional programming in the Lisp tradition. It has filter, map, and reduce functions; list comprehensions, dictionaries, sets and generator expressions. The standard library has two modules (itertools and functools) that implement functional tools borrowed from Haskell and Standard ML.

* Jupyter notebook:

Jupyter Notebook is a web-based interactive computational environment for creating Jupyter notebook documents. The “notebook” term can colloquially refer to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, and containing an ordered list of input/output cells which can contain code, text, mathematics, plots and rich media, usually ending with the “. Ipynb” extension.

* CSV dataset:

It is the data file which is used to perform predictions on the data and obtain meaningful information. The file extension of the dataset is .csv.

**DATA CLEANSING:**

In this process, we are entrusted to find whether there are any missing values (NaN), outliers. There are no missing values in the dataset which is taken. When it comes to outliers, since the questionnaire is a quantitative analysis method and Likert scale is employed at the time of data collection, there is no possibility of outliers.

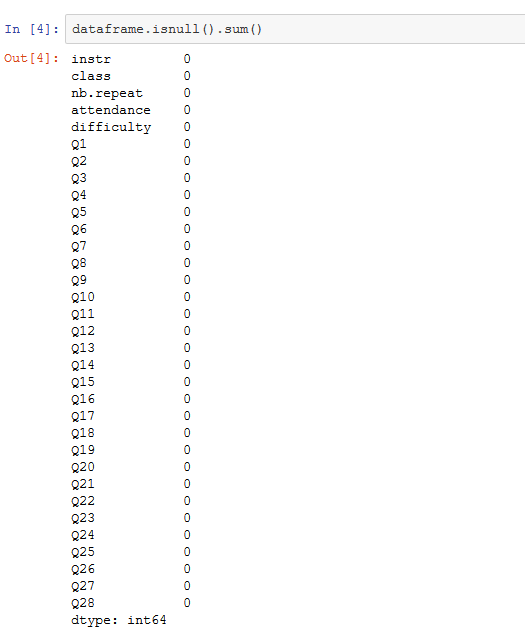


Fig 1. Checking missing values

**EXPLORATORY DATA ANALYSIS:**

In order to analyze the data, we have to know the entire statistics of the it and hence we are using the describe command.

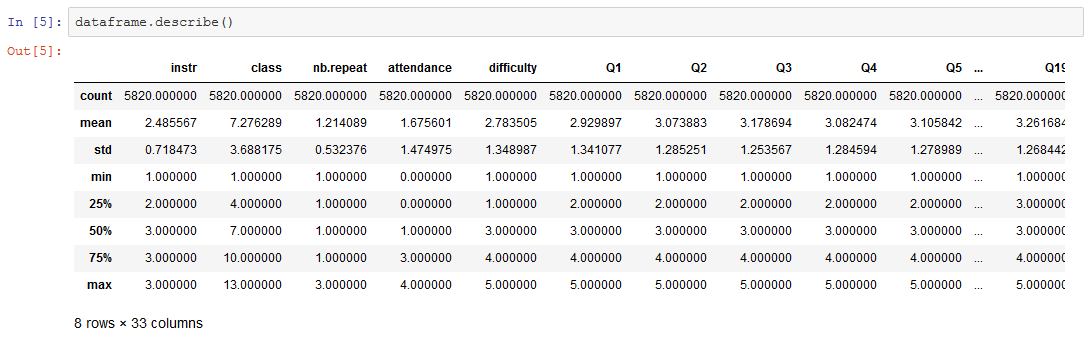


Fig 2. Describe command

We use plotting to gain better understanding of the dataset.

To understand for which course the students have given the most responses and it is found to be Course 3 from below graph.

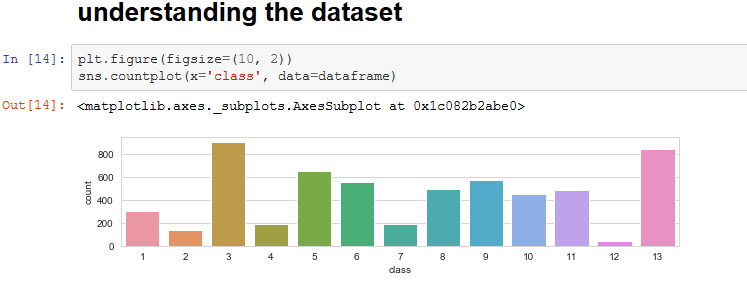


Fig 3. Understanding the dataset by plotting

Based on the following graph, students have given good rating for Q14, Q15, Q17, Q19, Q20, Q21, Q22, Q25, Q28.

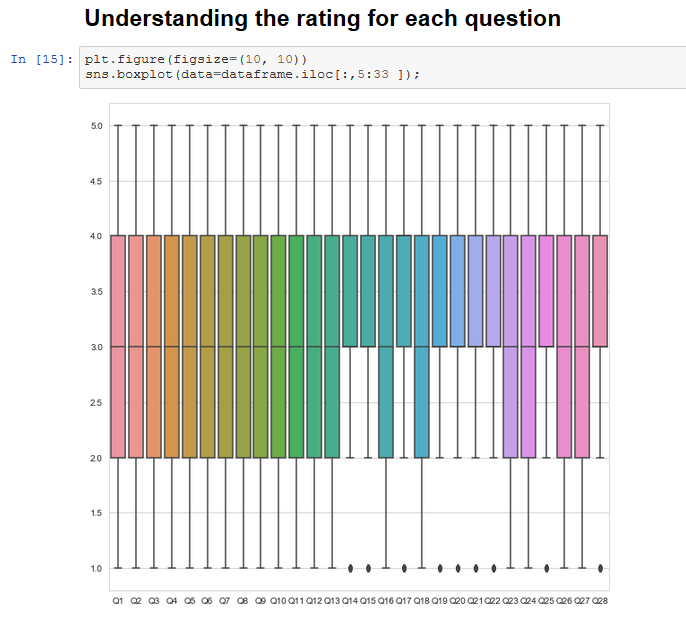


Fig 4. Understanding the rating for each question by plotting

To see the rating based on class

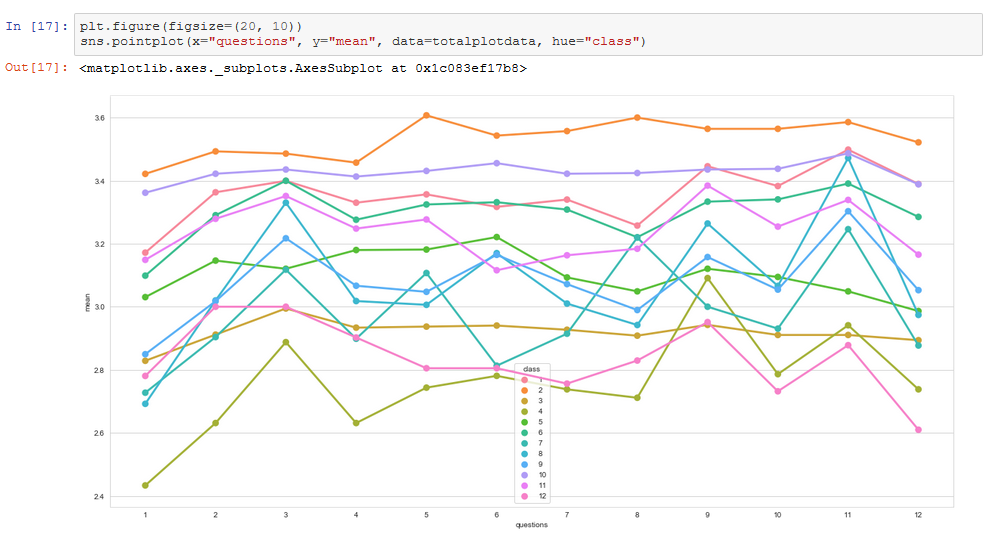
Based on the below graph, class 2 has the best rating and class 4 has the worst rating.

Fig 5. Class based rating

Calculate mean for each question response for all the classes.

Based on below graph we can see that Instructor 1 and 2 are performing well and got similar ratings Instructor 3 got less ratings. So, we can further explore which course instructor 3 teaches and find out the which course got least ratings.

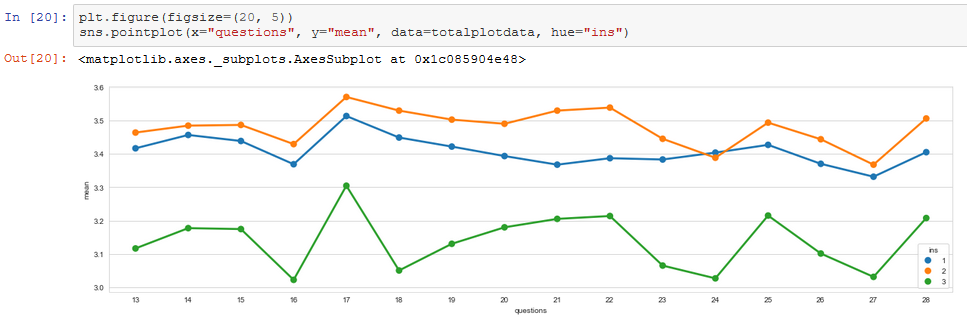


Fig 6. Instructor based rating

Calculate mean for each question response for all the classes for Instructor 3

By below analysis we can recommend the instructor 3 for check on course 4 and 13.

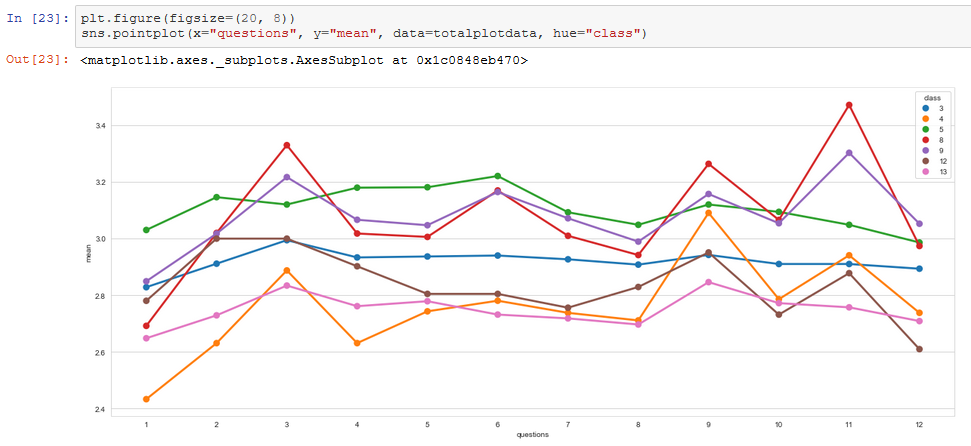


Fig 7. Instructor 3 mean response for courses 4 and 13

**FEATURE ENGINEERING:**

Scaling has not been performed as the dataset is already scaled because the data is recorded based on Likert scale. Then dimensionality reduction has been performed using PCA (Principal Component Analysis).

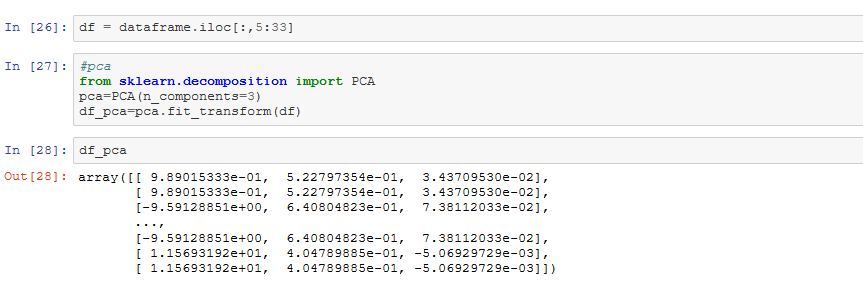


Fig 8. PCA

Elbow method is employed to find the number of clusters to be used in K-means.

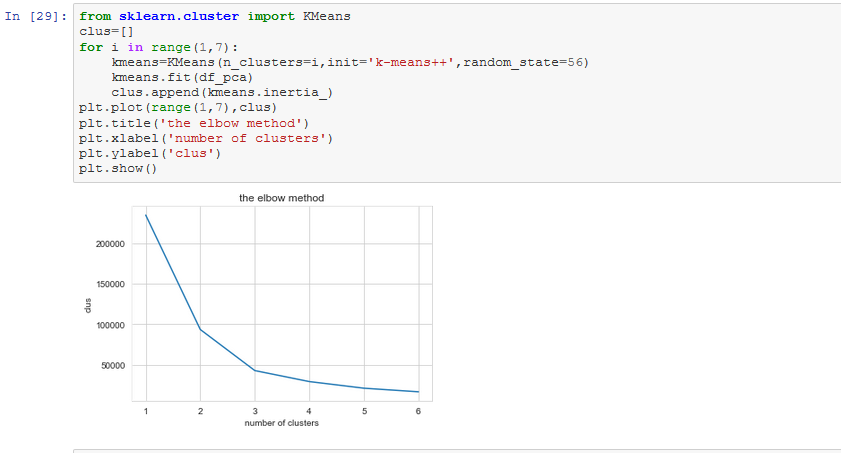
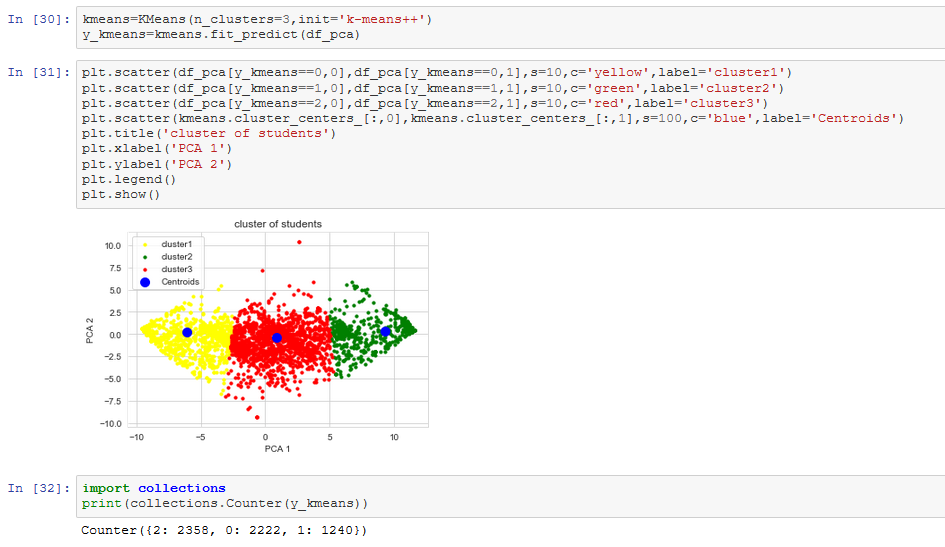


Fig 9. Elbow method modelling

K-means clustering is used to find a centroid point based on the Euclidean distance for every fields. The point changes for every cluster and the centroid point moves. In our project, k-means ++ is used which is more optimal than ordinary k-means based on time complexity where the former has O (log n), the centroid point changes.

Fig 10. K-means clustering and graph plotting

**INSIGHTS:**

Negative results: 38%

Neutral: 40.5%

Positive results: 21.5%

Based on the insights obtained from the K-means clustering, very few majority of the students are happy with the university in general. It is observed that the university has to perform better in terms of imparting knowledge to the students by recruiting competent instructors and adding value-added courses which might be useful for the students’ future.

In addition to this, the model has been deployed on cloud to ensure portability.