

|  |  |  |
| --- | --- | --- |
|  | **PES UNIVERSITY, Bangalore**  (Established under Karnataka Act No. 16 of 2013) | **UE18CS203** |
| **B.Tech, Sem III**  **Session : Aug-Dec, 2019**  **UE18CS203 – INTRODUCTION TO DATA SCIENCE** | | |

**REPORT**

**ON**

**EXPLORATORY ANALYSIS ON**

**JOKES RATING SET**

**SECTION : A**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **SRN** | **Name** | **Contact No.** | **Email ID** | **Sign** |
| 1) | PES1201801475 | SUDARSHAN.R | 8792607638 | sudarshanmurthy333@gmail.com | Sudarshan |
| 2) | PES1201801502 | ASHISH.D | 9740471350 | ashish742001@gmail.com | ashish |
|  |  |  |  |  |  |

**ABOUT THE DATA SET**

Data set: Jokes rating set

* Dataset has the ratings of the jokes from the user in the range of -10 to +10
* The first column indicates the number of jokes rated and the remaining 100 columns are the jokes given
* Each row is an user rating og the jokes , if the joke is not rated it is given as 99
* Size:24982 \*101

**ABSTRACT**

Data analysis is the process of systematically examining data with the purpose of spotlighting useful information. Data analysis is the foundation of scientific research.

The purpose of the assignment is to get hands-on random dataset and dealing with its various aspects .The dataset might be in correct format or it might not be ,in case where data cleansing is done .

Descriptive and visualization analysis can be done and will be shown below.

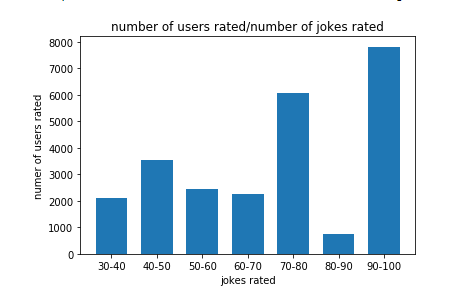
Kind of questions which could be asked

* Range of jokes for which maximum number of users have rated
* Those who rate more jokes give higher rating ? etc..

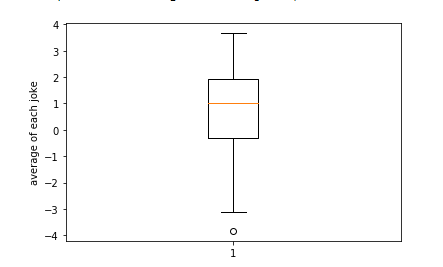
**EXPLORATORY ANALYSIS**

1) Maximum (8k) users have rated 90-100 jokes and it is the maximum for any range

Around 32% of users have rated maximum number of jokes



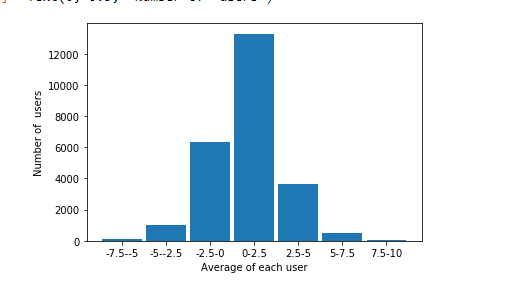
2) Median of the jokes is 1



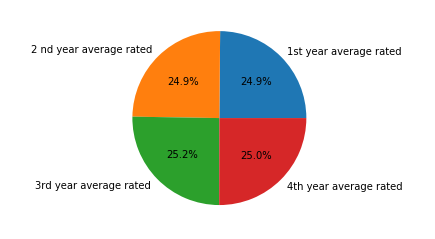
And there is an outlier which indicates that the joke is least rated and it is joke number 48

3) More number of users has an average of 0 to 2.5

And as the graph is almost symmetric, therefore we can conclude that Mean~Median



4) On dividing the set uniformly based on the year of rating ,it can be said that average number of jokes rated is uniform (25%)

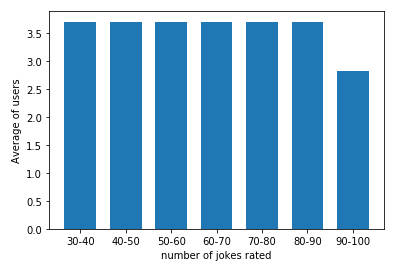


**Data cleaning**:

If the user hasn’t rated a joke , the default value was 99 which did not belong in the range (-10,10) . Therefore there was a need to do datacleaning

We first found the average of each column by skipping those jokes which had 99 as there rating ,later the 99 rated jokes were replaced by its corresponding column average.

* **Hypothesis testing: Those who rate more jokes give higher rating ?**



* From the plot we can say that the users who rate more jokes give a lesser rating .
* Hence we disprove the hypothesis and conclude that “**those who rate more number of jokes has lesser ratings**”

CONCLUSION

* We could find the range of number of jokes for which more number of users rated,which was between 90-100
* Median of entire joke set was 1.And with Boxplot plotted against average of each joke it was found that there was an outlier indicating that the joke was underrated (which was found out to be joke number 48) and the maximum rated joke was found out to be 50th joke
* By plotting average of each user against number of user ,it was found out that the maximum number of users average rating was between 0 to 2.5
* From a pie chart ,it was found out that ,on dividing the set uniformly based on the year of rating ,it can be said that average number of jokes rated is uniform
* Hypothesis testing :
  + - Hypothesis: Those who rate more jokes give higher rating?
      * We reject the hypothesis