**Dataset Description**

# **JOKES DATASET**

# 

## **TASK**

One dataset will be provided to each team. The dataset can be found in the folder corresponding to the team number.

The team’s task is to understand the dataset completely, and perform exploratory data analysis on it. This would entail using knowledge acquired from the class sessions to plot visualizations which would help understand the dataset better, and derive insights from it. These visualizations and basic insights are a must for the project. However, the focus of this project is to augment theoretical learning of the class with invaluable hands-on knowledge.

This project is also meant to increase awareness and encourage exploration of fields and topics which are overlapping with the field of Data Science. Thus, a portion of the project is meant to test the team’s willingness to learn and explore about new topics and techniques. Take note, however, that these might be topics which are not part of the Data Science course, but nonetheless will impart a great deal of knowledge.

To aid the thinking process, there is a section titled ‘Possible tracks for insights’ which is meant to provide an example of a possible direction to begin thinking in. The questions provided in that section are NOT a checklist of things to be done, as the basic visualizations need to be implemented, even though they are not mentioned in the section. It could also be the case that the question provided in this section does not provide any insight. In such cases, simply consider it as a direction to be thinking in, and feel free to skip the results of the question in the report. Marks will be awarded for any and all techniques not part of the course which have been implemented.

# 

# **DESCRIPTION**

A dataset is a collection of data. In a dataset, each row is a particular instance or record of an event, and each column corresponds to one variable or feature.

4.1 Million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users: collected between April 1999 - May 2003. Each row corresponds to one user. The first column is the number of jokes the user has rated. If a user has not rated a joke, it is shown as 99.

The data are contained in the files `links.csv`, `movies.csv`, `ratings.csv` and `tags.csv`.

**Take note**, however, that the data may have to be cleaned, with a number of erroneous or blank values, with a number of outliers. The dataset will be a CSV file or a number of them. Your task is to clean the dataset, and perform analysis on it, plot relevant visualizations and to gain some meaningful insight into the data using the Data Science skills you have acquired. If you deem any columns to be irrelevant to your analysis, you can discard them. However, you will have to report clearly why this column was discarded.

## **COLUMN EXPLANATION**

In this section, a short description of the meaning of each column can be found. Along with this, the type of data in the column is indicated.

1. The first column is the number of jokes, out of 100 the user has rated. If a user has not rated a joke, it is shown as 99.
2. The remaining columns are the ratings of the corresponding jokes

## **POSSIBLE TRACKS FOR INSIGHTS**

In this section, a small number of possible insights by manipulating and analysing the data are presented. This is certainly **NOT** a checklist of things to do in your project, but rather aims to provide a generic example of the direction to begin thinking in. It could certainly be the case that more valuable insights may be possible, which are not listed here. Some suggestions may be beyond the syllabus and are included only for learning, so please feel free to go out and learn how to do them.

Which joke was the highest rated?

For each joke, show a boxplot for the ratings. What insights can you derive from this?

If you had to recommend a joke to a user that they have not rated before, how would you go about doing this, in theory? [Hint : Collaborative Filtering] Include an explanation of how this algorithm would function, in your report (don’t have to implement)

Prove or disprove the hypothesis : Those who rate more jokes give higher ratings.

## **POTENTIALLY USEFUL LINKS**

In this section, you will find a number of links, which can be used to brush up your knowledge or help resolve issues.

### Python Tutorial

### <https://www.learnpython.org/>

### Conda Cheatsheet

<https://docs.conda.io/projects/conda/en/4.6.0/_downloads/52a95608c49671267e40c689e0bc00ca/conda-cheatsheet.pdf>

### Pandas Documentation

<https://pandas.pydata.org/pandas-docs/stable/>

### Pandas Quickstart (Short Tutorial)

<https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html>

### NumPy Quickstart (Short Tutorial)

<https://docs.scipy.org/doc/numpy/user/quickstart.html>

### Seaborn Documentation

<https://seaborn.pydata.org/introduction.html>

### Seaborn In-Depth Tutorial

<https://seaborn.pydata.org/tutorial.html#tutorial>

### Matplotlib Tutorial

<https://matplotlib.org/3.1.1/tutorials/index.html>

### Scikit-Learn Tutorial

<https://scikit-learn.org/stable/tutorial/index.html>