**­­­­3.2.2     Exploratory Data Analysis and Data Preparation (24%)**

Summarise the main characteristics of the dataset, using tables and statistical graphics, and/or other data visualisation methods.

Describe how you split the dataset and present stats such as count, mean, etc.

Describe how you constructed and/or transformed the data/feature.

**Data Analysis**

Table I. Raw Data Overview

Fig I. Histogram

Fig II. Violin Plot

Fig III. Pair Plot

Fig IV. Correlation Matrix

**Data Preparation**

1. Data Cleaning

* Corrupted Data
* Missing Data
* Statistical Outliers
* Imbalanced Data – Down Sampling and Up Weighting? Or just delete data?

1. Feature Scaling
2. Transforming categorial data

* One Hot Encoding

Order of writing:

Raw Data Overview

Data Cleaning

* Corrupted Data
* Missing Data
* Deal with Imbalanced samples

Data Presentation

Data Processing

1. Raw Data Overview

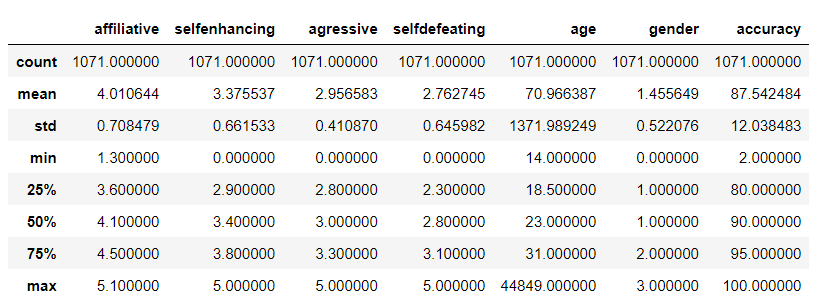


Table I. Feature used to predict gender and gender

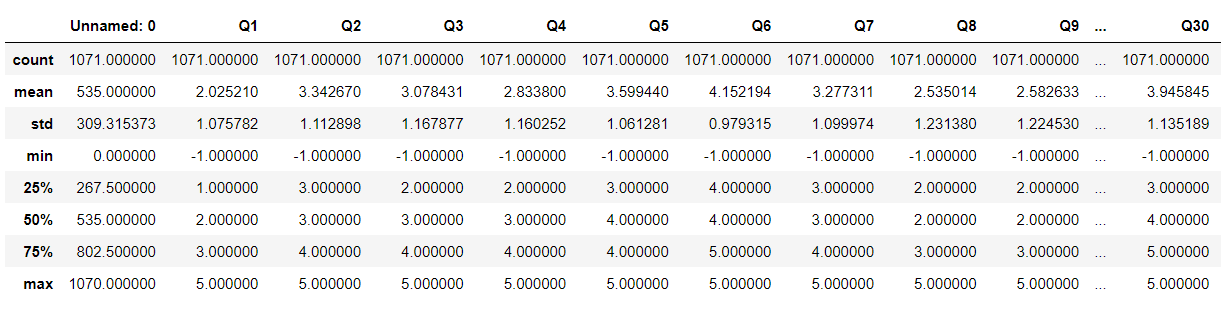


Table II. Questions used to calculate the four scale scores

The data used to predict gender are sourced from the study by Martin et al, titled Individual differences in uses of humour and their relation to psychological well-being: Development of the Humour Styles Questionnaire.

From Table I we can see that the data require some cleaning before being sent for any analysis and further processing.

The data frame contains corrupted data. There are

1. A maximum age of 44849, with some cases greater than the oldest natural human age of 120
2. Gender of 0, which is not in any of the three permitted category.
3. A maximum affiliative score of 5.1, calculation from missing question responses
4. Some participants’ test has a test accuracy of 0

To counter three scenarios mentioned above, following strategies are used:

1. Use median/mean = 22 to replace any age filled above age
2. Delete all gender filled with 0
3. For the affiliative score calculated to be 5.1, the samples are regarded to be erroneous and will be deleted.
4. Delete all participants’ test accuracy of 0

This should ensure that all the data are good for both visual presentation and model training in the next stage.

There are no missing data/uncalculated data from features’ table as well as gender’s column. Although it is worth noting that from Table II, there are clearly missing data filled with -1. They are the questions which participants did not answer. They are used to calculate the four scales score as well as accuracy. This could result in outliers for the features we use to predict gender, which then affect the accuracy of models trained.

The affected data will be treated as outliers.

1. Exploratory Visualization

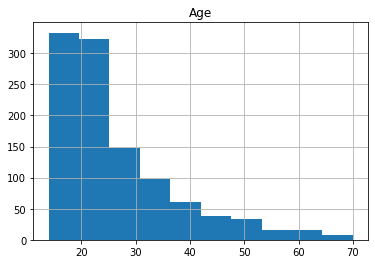
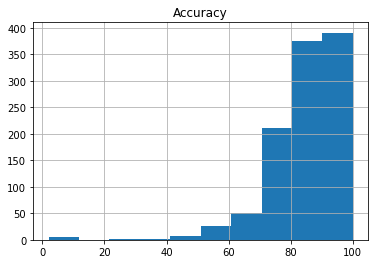
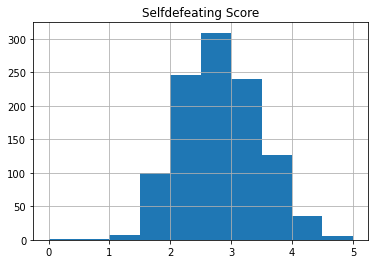
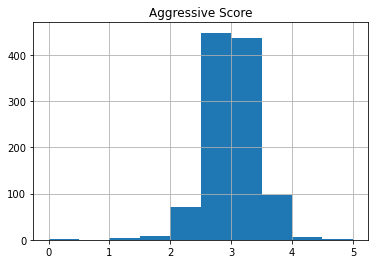
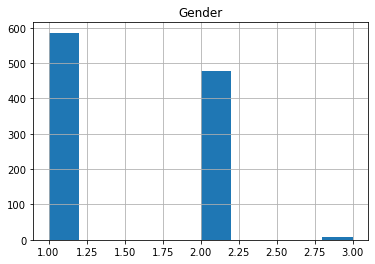
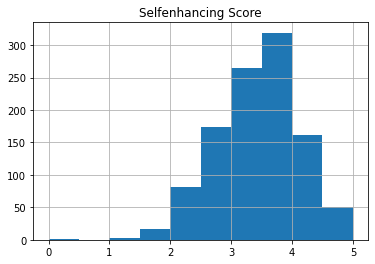
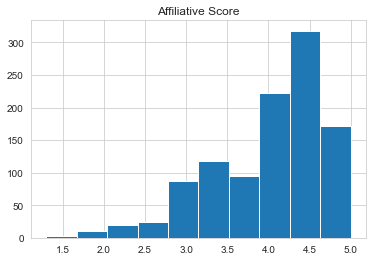


FIG. 1. Histogram

FIG. 2. Violin Plot

From histogram and violin plot, we can see that Self-defeating score has a Gaussian distribution with an average score of 2.8. Affiliative and accuracy are highly negatively skewed, i.e., they concentrate on the right part of the graph. Meaning affiliative type of humour are most prevalent type among participants, with an average score of 4.

There are more male participants than female, however there are significant less participants who filled their gender as others. The ratio is 585:477:8 (male: female: others).­

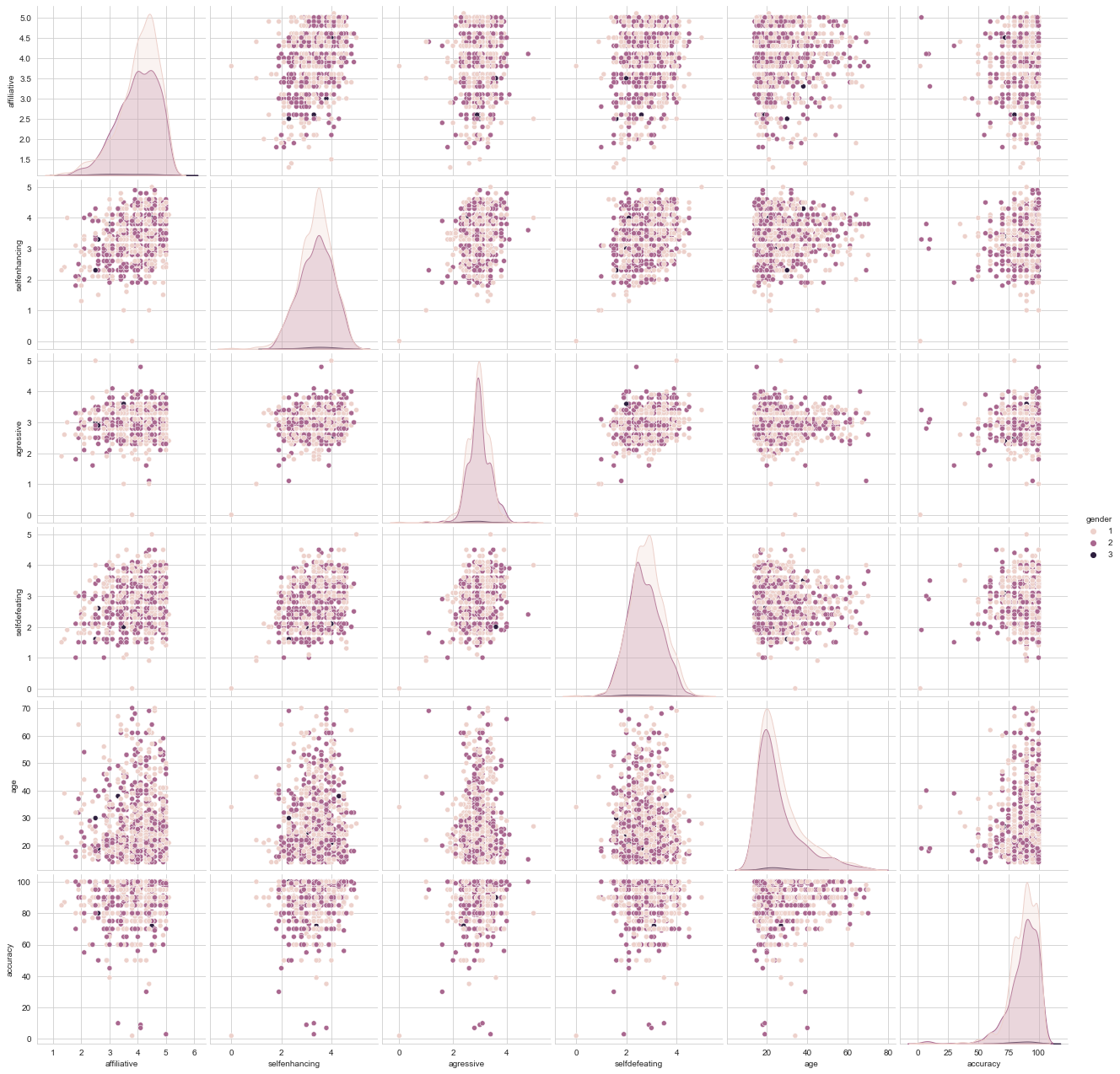


FIG. 3. Pair Plot

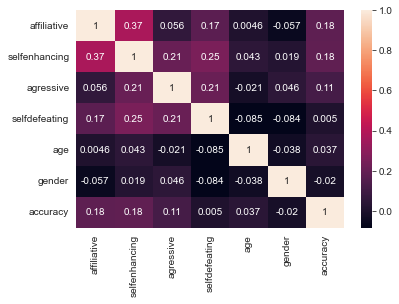


FIG. 4. Correlation Matrix

From the pair plot, we can see the three types of gender all have spread-out distribution on all the features relationships. However, there are quite a few outliers on each category.

It is noticeable that from the correlation matrix, gender does not have any strong correlation with any other categories. 6 correlations are negatively scored, meaning they have a weak negative correlation with each other.

Affiliative and self-enhancing has the highest positive correlation. Age and self-defeating has the highest negative correlation.

1. Data Processing

Before sending the data into the model, there are still some processing needed such as dealing with outliers.

~~A validation set may be put into action with a 70% for training, 10% for test set and 10% validation. It will be used to counter overfitting problem if occurs and used again to evaluate the models.~~

Transforming categorical data process for gender is necessary with one hot encoding, as the gender column contains discrete data which has no relationship with each other.

As noted earlier, due to the imbalanced data for gender, dealing with these samples is also necessary. With not enough data for 3-others to make accurate prediction, the following methods are used to counter this:

1. oversampling the minority class, done using synthetic minority sampling oversampling technique (SMOTE). It works by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. [1]
2. Using different models’ performance evaluation metrices, for example avoid using accuracy, using some others metrices such as F1 score and ROC AUC scores.

Statistical outliers for the four scale scores will be removed. Some of them are results of unfilled/missing data from the questions column. Interquartile range method will be used since most of the features are either positively or negatively skewed from histogram and violin plot. Following formulas are applied:

IQR = Q3 – Q1

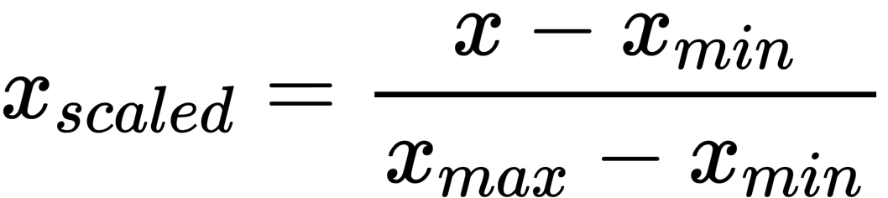
Outlier < Q1 – 1.5 x IQR

Outlier > Q3 + 1.5 x IQR

Q1 is the first quartile and Q3 is the third quartile.

Feature scaling is done using min-max normalization, it rescales the dataset such that all feature values are in the range [0,1]. As the cleaned data set does not have any outliers, therefore min-max scaler will not be affected.

Formula:



This is picked from comparing with other two different type of scalers: 1. Standard Scaler and 2. Robust Scaler.

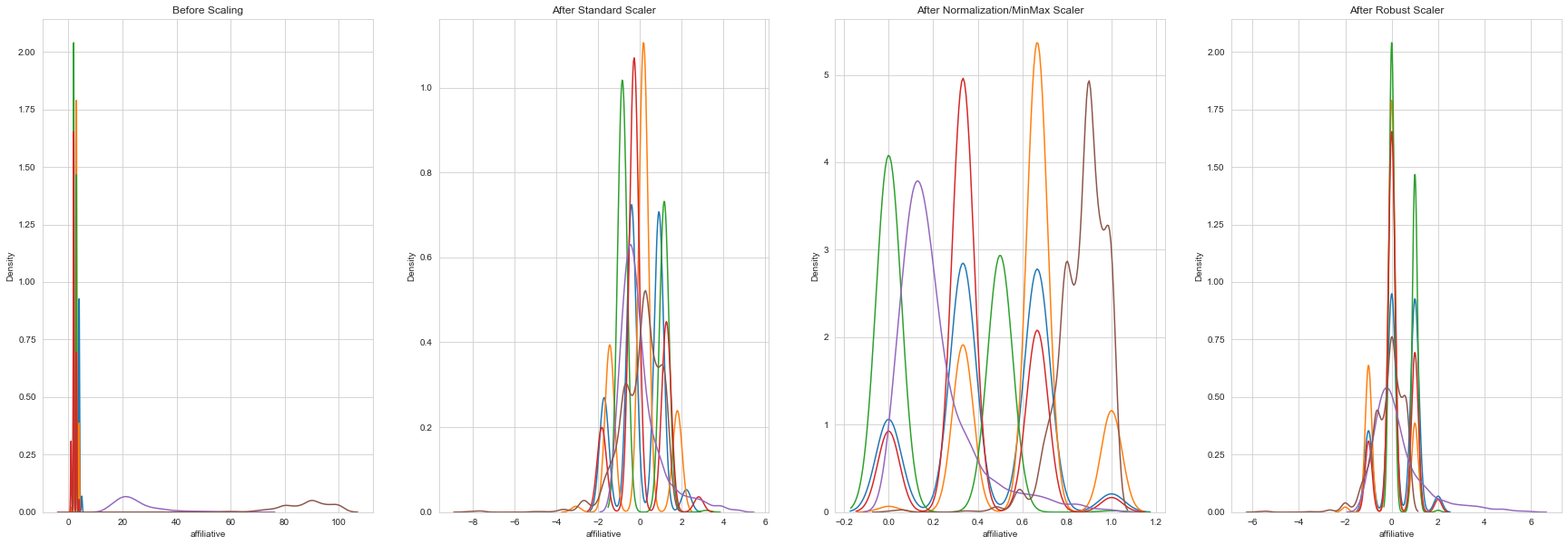


FIG.5. Scaler Comparison

The data will be split with 80% ( samples after SMOTE)for training set and 20% ( samples after SMOTE)for test set.

[1]: https://arxiv.org/abs/1106.1813