The following report is about analyzing the dataset which is realised by the Economic and Social Research Council (ESRC).

We take into consideration two datasets, one representing 4682 English words and the other one specifying 379 ambiguous words from the first dataset which have several meanings. In total we are using a set of 5553 english words, we want to analyze the following psycholinguistic dimensions: length of the words, arousal, valence, dominance, concreteness, imegeability, familiarity, age of acquisition, semsize, gender and web corpus frequency.

**1.1 DATA SEMANTICS**

In this section we represent two tables for both datasets – (1) Words Glasgow & (2) Words Polysemy.

Through the variable “Name” we indicate the name of the attribute/variable we are considering. In total we have 13 attributes (columns in our dataset), with the following table we demonstrate each feature with its description, type, and domain.

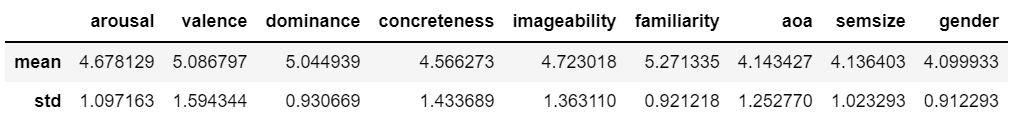
The first three dimensions are usually used to measure the emotional and psycological empact of a word. All the other variables indicate the grade of knowing the word and the use of it.

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **Description** | **TYPE** | **DOMAIN** |
| Word | English words | Categorical - Nominal (string) | 4683 (number of records) |
| Length | Word length | Numerical - Discrete (integer) | [2; 16] |
| Arousal | Measure of excitement (excitement, calmness) | Numerical - Contininuous (float) | [2.057, 8.177] |
| Valence | Measure of value or worth (positive, negative) | Numerical - Contininuous (float) | [1.03, 8.647] |
| Dominance | Measure of the degree of control  (dominant, controlled) | Numerical - Contininuous (float) | [1.941, 8.371] |
| Concreteness | Measure of how concrete or abstract something is (concrete, abstract) | Numerical - Contininuous (float) | [1.636, 6.938] |
| Imageability | Measure of generating a mental image of something (imageable, unimageable) | Numerical - Contininuous (float) | [1.737, 6.941] |
| Familiarity | Measure of how familiar a word is  (familiar, unfamiliar) | Numerical - Contininuous (float) | [1.647, 6.939] |
| Age of acquisition (aoa) | Measure of the age at which a word was initially acquired | Numerical - Contininuous (float) | [1.219, 6.971] |
| Semsize | Measure of magnitude  (big, small) | Numerical - Contininuous (float) | [1.375, 6.912] |
| Gender | Measure of a word considered to be associated with male or female behavior  (masculine, feminine) | Numerical - Contininuous (float) | [1.0, 6.971] |
| Polysemy | Measure of semantically ambiguous words which convey multiple meanings  (homographs) | Categorical - Binary (integer) | {0,1} |
| Web corpus frequency | Measure of frequency of a word in Google Newspapers Corpus | Numerical - Contininuous (float) | [12770.0, 2022459848.0] |

Out of all attributes we distinguish 9 variables as 9 dimensions of each record in our dataset, which are the following:

Table

Description automatically generated



The following is the table of the second dataset – ‘Words Polysemy’

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **Description** | **TYPE** | **DOMAIN** |
| Word | English words | Categorical - Nominal (string) | 872 (number of records) |
| Length | Word length | Numerical - Discrete (integer) | [2; 16] |
| Arousal | Measure of excitement (excitement, calmness) | Numerical - Contininuous (float) | [2.057, 8.177] |
| Valence | Measure of value or worth (positive, negative) | Numerical - Contininuous (float) | [1.03, 8.647] |
| Dominance | Measure of the degree of control  (dominant, controlled) | Numerical - Contininuous (float) | [1.941, 8.371] |
| Concreteness | Measure of how concrete or abstract something is (concrete, abstract) | Numerical - Contininuous (float) | [1.636, 6.938] |
| Imageability | Measure of generating a mental image of something (imageable, unimageable) | Numerical - Contininuous (float) | [1.737, 6.941] |
| Familiarity | Measure of how familiar a word is  (familiar, unfamiliar) | Numerical - Contininuous (float) | [1.647, 6.939] |
| Age of acquisition (aoa) | Measure of the age at which a word was initially acquired | Numerical - Contininuous (float) | [1.219, 6.971] |
| Semsize | Measure of magnitude  (big, small) | Numerical - Contininuous (float) | [1.375, 6.912] |
| Gender | Measure of a word considered to be associated with male or female behavior  (masculine, feminine) | Numerical - Contininuous (float) | [1.0, 6.971] |

**1.2 DISTRIBUTION OF THE VARIABLES AND STATISTICS**

In this section we represent the distribution of variables with the help of histograms and charts.

The first variable represented is ‘Polysemy’ that we consider as a target variable. It is represented as a boolean variable giving two values 0 and 1 (word has one meaning or several meanings respectively).

We observed that 4303 words has polymesy equal to 0, and 379 equal to 1.

A picture containing bar chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

We visualised density plots, bar charts and boxplots, similarly as previous, in order to learn the statistics and distribution of all the variables with respect to our target variable ‘Polysemy’.

Table

Description automatically generated

**1.3 ASSESING DATA QUALITY**

This following section studies and evaluates data quality. In order to asses the quality of the dataset it explores and handles missing values, outliers, and any other semantic errors or inconsistencies.

This part evaluates the quality of data. It investigates and handles missing values and outliers. Firstly, it has been decided to investigate the dataset for the none values, and to simplify the next work those were replaced using a specific method (see chap. 1.3.2). Finally, Z-score method helped identifying outliers in the data.

First we investigated for dublicate records and we found out that in our data we do not face the issues regarding to it.

Another thing that we checked was the actual length of the words’ strings. We compared string lengths to our attribute ‘length’ in order to make sure that the data represented with this attribute does not contain any semantic errors. finally we found zero errors in these variables.

After that, we checked the number of ambiguous (polysemous) words. For this part we used another dataset representing only polysemous words, we separated the description part from the actual word and counted the unique values. Finally, we compared the words from first dataset with the polysemy value equal to 1 and unique words from the second dataset. On the first phase we faced the difference between the total number of words. However, we found out that the difference was cause only because of the word ‘apple’ which was represented in the second dataset in two differenc ways (with capital letter representing the brand - Apple).

With the help of these several checks we made some assessments of our dataset.

**Missing values**

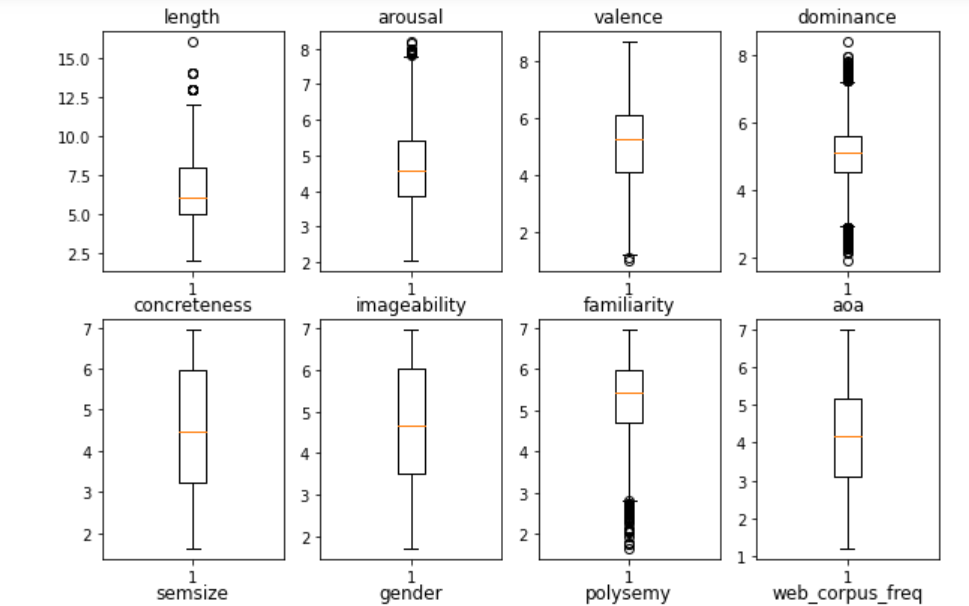
We found some missing values in only one attribute which represents ‘web corpus frequency’. As the number of those missing values (which is 14) was not significant w.r.t the number of all records.

So we decide to fix it through the average of their values.

**Outliers**

The outliers are data that can have different characteristics from all the others in the data set. They have an unusal value of an attribut from the usual values of that attribute. So, the outlier are defined as anomal object, which are different from the “rumors” that can be created beacause we are interested in.

In the analisys, we apply the boxplot method to define the number of outliers in all the cathegories.



**1.4 VARIABLE TRANSFORMATIONS**

For this part of the project first transformation that we decided to make was logarithm transform of the attribute ‘web corpus frequency’. As this variable is represented as a high continuous numbers, it was not convinient to analyze it with other attributes.

We also tried square root and z-score transformation. Z-score transformation helped us identifying outliers of some attributes.

Finally, we decided to only make log and z-score transformation for the future analysis.

**1.5 PAIRWISE CORRELATIONS AND EVENTUAL ELIMINATION OF VARIABLES**

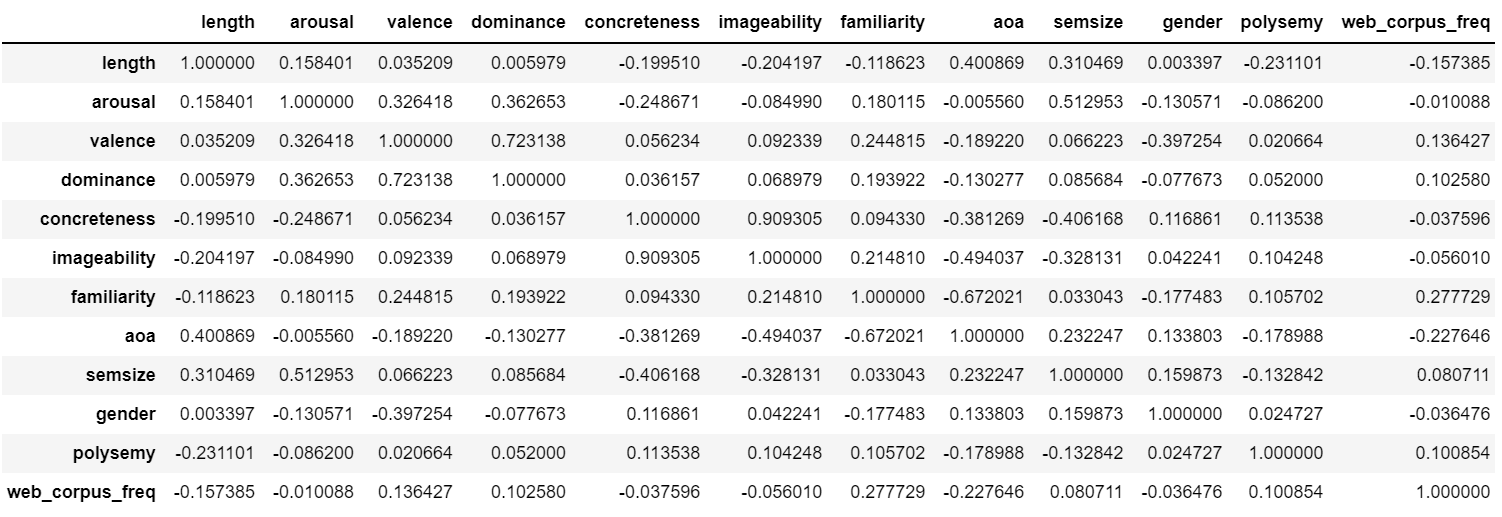
With correlation we mean statistic, random or not, between two random variables casuali o dati bivariati.

In the following table we can see the grade of correlation between the words. The correlation can be done dividing in couple all the columns in the data frame.

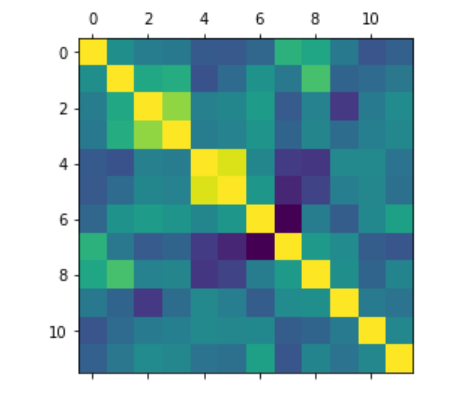
We observe that the correlation of a variables with itself is 1. For this reason, all the diagonal values are 1.

Table

Description automatically generated

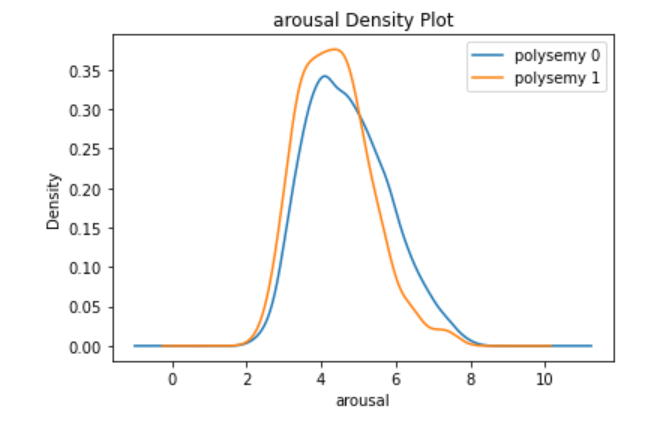


When we apply a graphic rappresentation to “scatter plots”, we can have the following graphic:



After that, we confront the target value “*polysemy*” with the other variables, so we can understand how it can influence the others.

We confront for example the *target value* with the variable “*arousal*”, and we can observe that if the polymey grows, so also the grade of arousal grows for a word.



**CLUSTERING**

**…**