Data correlation analysis

In order to analyse correlation among user demographic features (D01-D07) and personality features (A-E), for this purpose it was decided to use heat map, from Seaborn library, because it describes features relations well, even in case of multidimensionality.

To answer the first question, I will be analysing the following features of heat map.

Heat map allows to identify correlation between two variables in range (-1,1), with strong positive dependency for values close to 1 (both variables increases or both variables decrease accordingly),

strong negative dependency for values close to -1 (if first variable increases second will decreases and vice versa). For values close to zero, there is no strong trend (correlation) between variables.

The correlation details will be described from psychological features perspective, it will allow to describe dependency of one psychological on a few demographic features, so that it makes not only logical but also conceptual sense.

D01 Gender

D02 Year of birth

D03 Education - basic, vocational, secondary, higher education

D04 Professional status

D05 Size of the city of residence - village, up to 20k, up to 100k, up to 500k, above

D06 Financial situation

D07 Household size

A - D04, D06, D07 have strongest positive impact on A, D05, D02 have strongest negative impact on A.

Education and gender have least impact on A, while professional status and size of the city have biggest impact on A among all demographical features.

B - D02, D03, D05 have strongest positive impact on B, moreover most demographic features have strongest positive correlation with B (in comparison with their impact on other phycological features). All demographic feature has negative correlation with B.

Year of birth, education and size of the city have biggest impact on B among all demographical features.

C - D01, D07 have strongest positive impact on C, D02 have strongest negative impact on C.

Most demographic features have relatively weak correlation with C (in comparison with their impact on other phycological features).

Education, financial situation and size of the city have least impact on C.

D - D01, D04 have strongest negative impact on D. In case of strongly positive correlated features, none of them has significant impact on D, the strongest dependency is with Year of birth (D02) with minor 0.06 value.

Professional status and gender have biggest impact on D among all demographical features.

E - D01 and D07 have strongest positive impact on E, D02 have strongest negative impact on E.

Most demographic features have relatively weak correlation with E (in comparison with their impact on other phycological features), except year of birth (D02) with negative 0.3 value.

Education, size of the city and financial situation have least impact on E.

In conclusion, all demographic features are correlated with phycological features to some extent, although none of those connections are very strong, values range from negative 0.3 to positive 0.3. There will be also provided a list of each demographic feature with its strongest correlation with one of phycological feature.

D01 correlates the most with D, with a value of negative 0.2.

D02 correlates the most with B and E, with a value of positive 0.3 and negative 0.3 respectively.

D03 correlates the most with B, with a value of negative 0.2.

D04 correlates the most with A, with a value of negative 0.2.

D05 correlates the most with A and B, with a value of negative 0.3 and positive 0.2 respectively.

D06 correlates the most with A, with a value of positive 0.1.

D07 correlates the most with C and E, with a value of positive 0.1 for both features.

Groups division analysis (preprocessing)

To answer the question of dividing users into groups and identifying user similarities within those groups, two data sets must be analysed.

First set, personality.csv, contains users(respondents) phycological information, it includes 5 features(A-E).

Second set, users.csv, contains users(respondents) demographic information, it includes 7 features(D01-D07).

Both sets are unlabelled, meaning the grouping task cannot be solved with classification model. This means clustering model will be an optimal option to work with.

Due to high dimensionality of dataset, it might be difficult to pick specific clustering model, such as DBSCAN vs mean-shift clustering.

Because of this, most widely used clustering model, K-Means, was used to be a model in order identify user groups.

K-Means requires specification of number of clusters. To identify the number of clusters the elbow method was used.

To identify number of clusters it was decided to use merged version of users and personality datasets. Sets were merged on user identifier feature, present in both sets, to avoid further redundancy.

In this case elbow methods function graph representation does not have an optimal option for number of clusters. To make result more obvious, it was decided to try KneeLocator method, after performing KneeLocator on dataset, optimal number of clusters turned out to be 4.

Because of not straight forward result, additional techniques were performed. Davies-Bouldin and Silhouette Coefficient scores. In both cases optimal number of clusters was 4, although it worth to be mentioned, that in case of Silhouette Coefficient for clusters in range (2-10), score was pretty high, in range 0.7 up to 1.0.

Groups division analysis (model)

Number of groups must be determined from the point of view of phycological profiles, so only (A-E) feature will be used to fit into model.

While demographic features, of users within each cluster, will be used for feature variation analysis. 50 turned out to be optimal number

of epochs for training stage.

Such a decision was made because in case of choosing both, 4 and 5 clusters, number of users in each cluster was on acceptable range,

meaning there was no cluster with less than 100 users and more than 600 users in another one.

Groups division analysis (clusters analysis)

Clusters will be analysed from both psychological and demographic perspective, in order to provide enough specification for of them.

All analysis will be performed with respect to feature normalization step.

Features which do not have specific tendencies will be described first, then features which uniquely describe each cluster will be mentioned separately.

Common phycological features analysis –

A - Most users in cluster 0 have value of A close to -0.75, this trend is also visible in clusters 1,3 and 4.

B - Values of B range from negative 1.5 to positive 1.5 in clusters 0-3, cluster 4 is the only exception.

C, D - For all clusters values of C and D are mostly unique, they will be observed independently for each cluster.

E - Value of E has the same value in most cases, in cluster 0, 1, 2 and 4, those values are approximately (-1.25, -0.75, 0, 0.75, 1.5), the only exception is cluster 3, the most common value there is negative 1.25.

Common demographic features analysis –

D01 - For D01, in cluster 0, 2 and 3, most of the values belong to range (-0.5, 2), exception is cluster 1 with significant number of users with value D01 equals negative 2 and 1. D01 in, cluster 4, has the same tendency as D01 in cluster 4.

D02 - Most of the values of D02 belong to range from (-2, 1), cluster 4 might look like an exception, most probably it varies a little due small size of cluster.

D03 - For D03 varies from -2.5 up to 2.5, values within each cluster has a normal distribution, except cluster 4 and 5, in those cases the fact of normal distribution is not too straight forward.

D04 - For D04 values in all clusters varies from -2.5 up to 2.5, the maximal values close to zero, values of D04 within each cluster have a normal distribution.

D05 - For D05 values in all clusters varies from -2.5 up to 2.5, the distribution is very similar to D04.

D06, D07 - For D06 and D07 values in all clusters varies from -2.5 up to 2.5, and it has very similar distribution to D04 and D05, the only interesting exception is D07 for cluster 4, in this case values mostly ranging from negative 2 up to positive 0.5.

Unique phycological and demographic

features analysis –

Cluster 0 -

Phycological features analysis -

For cluster 0 values of C, in most cases, equal to negative 0.5 and 1. I worth to be mention that in all clusters, C which is equal to negative 0.5, is common case.

For cluster 0 value of D equals to negative 0.75, this trend also appears in cluster 0, 2 and 3.

Cluster 1 -

Phycological features analysis -

For cluster 1 values of C, in most cases, equal to negative 1.5 and negative 0.5.

For cluster 1 values of D, in most cases, equal to positive 1.5, 2 and 3, this trend is unique for this cluster.

Cluster 2 -

Phycological features analysis -

Feature A equals to 1.2 in most cases, this trend is not observed in any other cluster.

Cluster 4 -

Phycological features analysis -

Feature A not only has high value of negative 0.75, as is in most of the clusters, but also value of 1.2 at the same frequency rate.

Only in cluster 4, most values of B are less than negative 1.5, while in all other clusters value of B ranges from negative 1.5 to positive 1.5.

The value of D in most cases equal to positive 0.75, this trend is unique for this cluster.