Effect of COVID-19 on Mental Well-being: an Unsupervised Learning Analysis

Bachelor's Project Thesis

Jelmer van Lune, j.van.lune.1@student.rug.nl,
Supervisor: dr. M.K. van Vugt

Abstract: Just as seen during previous virus outbreaks, the COVID-19 pandemic can have a negative impact on the mental well-being of individuals. However, it is likely that not everyone responds mentally the same to the pandemic. In this thesis I investigated whether there are groups of people that respond different with respect to mental well-being to the pandemic, by performing unsupervised learning on a large-scale questionnaire study performed during the pandemic. Moreover, I examined what other factors differ between groups of individuals responding adaptively and maladaptively to the pandemic and how these groups evolved during the pandemic. Indeed, a K-Means clustering and to a lesser extent a Hierarchical Agglomerative clustering analysis indicated that there were two groups of people, one with a better average mental well-being, one with a worse average mental well-being. Other factors that differed between these group were age, gender, employment, financial worries, social contact, frequency of leaving the house, knowledge about the virus, confidence in government, being infected and knowing infected people. In both groups the mental well-being improved slightly as the pandemic progressed.

1 Introduction

The World Health Organization (WHO) declared in January 2020 the outbreak of a novel coronavirus, COVID-19. The virus was first detected in Wuhan, China. After this, the virus has been spreading rapidly all over the world. Later, in March 2020, WHO declared COVID-19 as a global pandemic (Bhattarai and Karki, 2020). According to the WHO COVID-19 dashboard at the time of writing (January 26th, 2021), there are approximately 100 million cases of the virus worldwide and over 2 million deaths caused by the virus (WHO, 2021). To control this pandemic, governments have taken certain measures. In the Netherlands for example, the first national government measures were a social distancing policy (keep a distance of 1.5m to each other), advice to often wash hands, and a request to stay home as much as possible, which were announced on March 9, 2020 (Antonides and van Leeuwen, 2020). On March 12, 2020, more measures were announced. Events with over 100 people were cancelled, visits to vulnerable people were limited, and there was advised to work from home as much as possible. Worldwide similar measures were implemented. The effect of the pandemic, and measures taken to prevent the spreading of the virus, have risen concern regarding their consequences to the mental health of the general population (Bhattarai and Karki, 2020). Mental health is an indicator of emotional, psychological and social well-being of an individual. It determines how an individual thinks, feels and handle situations (Srividya, Mohanavalli, and Bhalaji, 2018).

From previous experiences with coronaviruses, it can be derived that such an outbreak can have an effect on global mental health. For example, medical staff was mentally affected by the Korean MERS-CoV outbreak. Medical staff that performed MERS related tasks, showed symptoms of post-traumatic stress disorder (Torales, O'Higgins, Castaldelli-Maia, and Ventriglio, 2020). Futhermore, during the Ebola outbreaks the Democratic Republic of the Congo in 2018, and in Sierra Leone in 2014, high levels of anxiety and the impact of stigma were reported among medical staff (Torales et al., 2020). Not only the medical staff was affected mentally during the Ebola outbreak in Sierra Leone, also the patients and the general population were affected. During the outbreak an increase of people with mental health and psychological problems were reported among the general population. There was in increase of people having mild distress or depression, anxiety disorders and grief or social problems(Kamara, Walder, Duncan, Kabbedijk, Hughes, and Muana, 2017). From these examples, it becomes clear that mental health is affected by the outbreak of a virus.

It has also been shown that the current coronavirus outbreak (COVID-19) has an impact on the mental health of medical workers. Medical workers in Wuhan for example had to deal with isolation, lack of contact with family and friends, overwork, inadequate protection against contamination and a high risk of infection. This led to stress, anxiety and depressive symptoms (Torales et al., 2020). Besides the medical workers, non-medical workers can also be mentally affected by the outbreak. A study in Hong Kong, conducted using a questionnaire between April 24th to May 3rd 2020 during the COVID-19 pandemic, showed that the mental health of 25.4% of the participants (randomly sampled from the population) had deteriorated compared to before the pandemic (Choi, Hui, and Wan, 2020). Besides this, the study also showed that of the participants 19% had depression and 14% had anxiety. Factors that caused poorer mental health were being worried about being infected by COVID-19, bothered by mask shortage, bothered by not being able to work from home and not experiencing the SARS outbreak in 2003 (Choi et al., 2020). Another study, conducted using a questionnaire in Italy during the last 2 weeks of the initial lockdown during the first COVID-19 wave (April 19th till May 3rd 2020), showed a high prevalence of mental health issues among the general population. Depression and anxiety symptom prevalence was 24.7% and 23.2% respectively (Gualano, Lo Moro, Voglino, Bert, and Siliquini, 2020). The likelihood of a mental health issue outcome increased when more time was spent on the internet, being female and avoiding activities because of peer pressure. Besides this, younger people experienced higher anxiety levels because they are more likely to reach a greater amount of information through social media, which might influence stress. Also, media contributed to unwarranted public fear, distrust and intolerance towards "dangerous others" (Gualano et al., 2020). Increasing age, an absence of work-related troubles and being married reduced the likelihood of a mental health issue outcome (Gualano et al., 2020). Furthermore, a study in Austria found using a questionnaire that the depression and anxiety symptom levels are higher after four weeks of lockdown, compared to data before the lockdown. The lockdown seemed particularly stressful in Austria for people younger than 35 years old, women, people without work and people with a low income (Pieh, Budimir, and Probst, 2020). A different study in the Netherlands between April 1st and May 13th 2020, conducted on people with pre-existing mental health disorders (depression, anxiety or obsessive-compulsive disorders) before the COVID-19 pandemic, found that these people did not report greater increase in their symptoms during the pandemic. This suggests that for people already suffering mentally, the pandemic does not seem to have further increases symptom severity, compared to before the pandemic (Pan, Kok, Eikelenboom, Horsfall, Jörg, Luteijn, Rhebergen, Oppen, Giltay, and Penninx, 2020). However, people without mental health issues before the pandemic, showed a greater increase in mental health issue symptoms during the pandemic. Not only in these mentioned countries, but for many more countries there are reports of the alarming implications on emotional and social functioning or an increased vulnerability for mental health problems of the COVID-19 pandemic (Pieh et al., 2020). Pfefferbaum and North (2020) concluded that the COVID-19 pandemic has alarming implications for individuals and collective health and emotional and social functioning worldwide.

With unsupervised machine learning it is possible to identify structures in large datasets, for example questionnaire datasets. Unsupervised learning is a type of machine learning where is tried to directly infer properties and patterns in the data, without the help of a supervisor (Hastie, Tibshirani, and Friedman, 2009). Data can be clustered into groups to potentially find new meaningful information in the data. Unsupervised learning has been used before on questionnaires regarding mental health. In a study by Srividya et al. (2018), clustering is used on a mental health questionnaire. The questions collected information on higher levels of well-being: engagement, perseverance, optimism, connectedness, and happiness. Every question had to be answered with a score from 1 to 5. Three different clustering algorithms were used on the data, namely K-Means clustering, Hierarchical clustering and K-Medoids clustering. For every clustering algorithm, three clusters were found in the data. The three clusters were found to represent groups of people that were mentally distressed, neutral and happy. These labels were later used for supervised learning purposes. In a different study, conducted by Chattopadhyay, Kaur, Rabhi, and Acharya (2012), a different clustering technique, a Self-Organizing Map (SOM), is used to cluster data to find different grades of depression. The data for this study was gathered using a questionnaire about emotional, cognitive, motivational and vegetative constructs. The SOM was able to identify three different clusters in the data relatively well, mild cases of depression, moderate cases and severe cases.

Some of these above mentioned unsupervised learning techniques can potentially also be applied on the data of a questionnaire distributed during the COVID-19 pandemic. It has been shown that data from a survey can be clustered and participants can be divided in groups with better and worse mental well-being (Srividya et al., 2018 and Chattopadhyay et al., 2012). It has also been shown that the mental health of the general population can be negatively affected by COVID-19 and the measures taken to prevent the spreading of the virus (Choi et al., 2020, Gualano et al., 2020 and Pieh et al., 2020). Besides this however, it has been suggested that the mental health of people already suffering from mental health issues before the pandemic, did not worsen during the pandemic (Pan et al., 2020). It is likely however, that not everyone responds the same to the pandemic with respect to mental well-being. There might be people not affected mentally much by the pandemic. It might be interesting to know if such different groups of people actually exist, what differs between them and how they mentally progress during the pandemic. Therefore, the research question to be answered in this thesis is: Are there groups of people that differ in mental well-being during the COVID-19 pandemic, what are their differences, and how do these groups evolve as the pandemic progresses?

The hypothesis to the first part of the research question, whether there are groups of people that differ in mental well-being during the COVID-19 pandemic, is as follows. Multiple papers suggest that in the general population in different countries over the world there is an increase in mental health issues during the pandemic (Gualano et al., 2020, Choi et al., 2020 and Pieh et al., 2020). Meaning that generally there are people that do not react well to the pandemic, causing their mental well-being to worsen. However, it is likely there are also people not very much affected by the measures taken against the virus, with a good mental well-being before the pandemic and thus also during the pandemic. This means that there probably are groups of people with a good mental well-being and groups of people with worse mental well-being, during the pandemic.

Secondly I asked, what different factors between these groups could be. The hypothesis to this question is as follows. Younger people and females were more likely to have mental health issues during the pandemic, compared to older people and males (Pieh et al., 2020 and Gualano et al., 2020), which could be factors differentiating the groups. It also has been suggested that people without work and people having financial stress, are more likely to develop mental health issues during the pandemic (Pieh et al., 2020), which could be other factors differentiating the groups. Furthermore, avoiding activities was found to increase the likelihood of mental health issues (Gualano et al., 2020). Avoiding activities can result in leaving the house less frequently and less social in-person contact with friends or other people in general. Additionally, being married was mentioned as one of the factors of a decreased likelihood of a mental health issue, during the lockdown in Italy (Gualano et al., 2020). This could also suggest that in-person contact with someone close is an important factor for good mental well-being during the pandemic. Moreover, it is shown that the media contributed to poorer mental health. It caused unwarrented public fear (Gualano et al., 2020). Clear messages from the government in the media regarding the situation around COVID-19 could possibly prevent that. People would have confidence in the government to be able to fight the virus, because it is clear what the current situation is and how the government will respond. Finally, being worried about being infected by COVID-19 was also found to be factor that caused poorer mental health (Choi et al., 2020). This could mean that knowing infected people close to you, causes worse mental health. In summary I hypothesize that in the groups of people with a worse mental well-being, compared to the groups with better mental well-being: people are younger, there is a higher proportion of females, higher proportion of people with no work, people with more financial worries, people leaving the house less frequent, less in-person contact with friends relatives or other people in general, less confidence in the government able to fight the virus and knowing more infected people.

Finally, I asked how the groups evolve mentally as the pandemic progresses. The hypothesis for the question is as follows. People with mental health issues before the pandemic didn't suffer more during the pandemic (Pan et al., 2020). This could suggest that once the mental well-being of people not responding well to the pandemic has deteriorated, it will not deteriorate further as the pandemic progresses. This means that people found in a group with generally bad mental well-being during the pandemic, probably also will not suffer more than they already did during the rest of the pandemic. Furthermore, people with on average a better mental well-being during the pandemic will probably maintain this mental well-being, because it is likely they are not affected much by the measures.

To test these hypotheses, data gathered by a large-scale worldwide questionnaire conducted during the pandemic (March 19th till July 13th) will be clustered using unsupervised learning techniques. The clustering algorithms used are K-Means clustering and Hierarchical Agglomerative clustering, because these are two of the most practical and most commonly used clustering algorithms and are used on questionnaire data in previous studies by for example Srividya et al. (2018).

2 Methods

First I asked whether there exist groups of people that differ in mental well-being during the COVID-19 pandemic, and what other factors differ between these groups. To answer these two questions, unsupervised clustering is performed on data of a large-scaled questionnaire study conducted during the pandemic. K-Means clustering and Hierarchical Agglomerative clustering are performed to see if participants of the questionnaire can be divided into groups that differ in mental well-being. If that is the case, other factors differing between the groups are evaluated. Secondly, I also asked how the mental well-being of the observed groups evolves as the pandemic progresses. To answer this question, the data of follow-up studies of the questionnaire study are evaluated from March to July 2020.

2.1 Questionnaire data

Data about mental well-being were derived from the PsyCorona project, which is a large-scaled questionnaire project that aims to identify psychological and cultural factors that affect mental health during the COVID-19 pandemic. This project started on March 19th 2020 and is still ongoing. In this thesis the data from this project from a period between March 19th and July 13th are used. During this period participants of this study were asked about their experiences, feelings and circumstances. Examples of information gathered by this survey are ratings on the presence on different emotions, employment status, social contacts and personal traits. In appendix A all the variables of the survey included in this study are shown and explained.

Mental well-being can be assessed to some extent on the basis of the presence of positive and negative affect. In this study, the positive affect variables are calm, energetic, inspired and relaxed. The negative affect variables are anxious, bored, depressed, exhausted and nervous. These positive and negative affect variables are used to examine whether it is possible to divide the data into groups that differ on this dimension, and whether there are other variables that predict whether an individual has good or poor mental health.

In total 62,902 people participated in the PsyCorona survey during this period. Approximately half participated as volunteers and half were recruited via paid panels. Important to note here is that the participants recruited via paid panels were carefully sampled to ensure representativeness. This was not the case for the volunteers. Of these participants 61.46% is female, 38.06% is male and 0.48% filled in "other", while the average age is between 35 and 44 years old. The data are collected from all over world, with participants living in 115 different countries. Approximately half of the participants were from the following countries: United States of America, The Netherlands, Greece, Romania, Indonesia, Republic of Serbia, Italy and the United Kingdom.

Every participant completed the baseline of the survey. This is the first time they complete they survey and can be anywhere between March 19th and July 13th. For this baseline survey they fill out a number of personality questionnaires and other self-report measures that characterize them as an individual. After they had completed the baseline survey, the participants had the possibility to participate in additional follow-up surveys, which were deployed from March 27th to July 13th. These follow-up surveys focused more on how people responded to and acted during the pandemic. These different waves of the survey could have a period of a week, two weeks or a month between them. It is possible to skip waves, or not complete any waves at all except for the baseline. In table 2.1 the dates are shown of when each wave of the follow-survey was deployed, and how many people participated in each wave.

The K-Means and Hierarchical Agglomerative clustering are only performed on the baseline data. This means clusters will be formed based on how the participants felt the first time they completed the survey anywhere between March 19th and July 13th. The follow-up survey wave data are used in this study to investigate how the people in the clusters evolved over time as the pandemic progressed.

Table 2.1: Survey dates in 2020 and number of participants for each wave of the follow-up surveys.

wave	date	participants
1	27/03	1,511
2	11/04	6,268
3	18/04	5,561
4	25/04	8,030
5	02/05	7,366
6	09/05	6,563
7	16/05	5,318
8	23/05	5,357
9	30/05	4,858
10	06/06	4,151
11	13/06	4,952
12	13/07	4,360

2.2 Data preparation

Before the K-Means clustering and Hierarchical clustering can be performed, the baseline data is cleaned up and prepared for clustering. For these clustering algorithms to work on the baseline data, first of all there should be no NA values for the baseline variables of the participants in the dataset. To accomplish this, ordinal baseline variables from the original dataset were removed if 10% or more of the participants did not complete that question. As mentioned before, the complete list of the ordinal and binomial variables that remained and thus used in this study are shown in appendix A. The variables `employstatus 1, 2 and 3` were originally binomial variables, but are merged to be an ordinal variable. These variables indicated if an individual worked 1-23 per week, 24-39 hours per week or 40 or more hours per week. These variables are merged to one ordinal variable, representing the amount of hours an individual worked per week. I chose for this approach because the three different variables represented the same construct and could easily be merged to decrease the number of NA values in the dataset. Furthermore, if there still are NA-values for participants in the dataset, imputation is performed. For ordinal variables, the general median of that variable is imputed, which is a common practice. For all the ordinal variables 1.18% of the values are imputed values. Besides this, for the binomial variables a zero is imputed. I chose for this method, because the questions for these variables were tick box questions. If a participant did not tick the box for that question, the question did not apply to them, meaning a zero would be a meaningful imputation. For all the binomial variables 80.69% of the values are imputed zeros. There is one exception, for the gender variable the most frequent answer is imputed. For this variable 0.27% of the values are imputed values.

Secondly, only numerical variables are suitable for clustering. For this reason, variables containing text, such as the country of residence or their response ID are removed.

The final step of the data preparation is to scale the baseline data. This is done because the variables have different ranges. Because of those different ranges, variables with higher ranges could have a bigger influence on the clustering results compared to variables with a smaller range. This should not be the case. Every variables should have the same influence on the clustering. To scale the data, the data is transformed such that each variable has a mean of 0 and a standard deviation of 1. After the clusters are formed and the clusters get evaluated the scaling and imputations are removed, as to analyze the original data.

Only the baseline data is cleaned up and prepared. This is not done on the follow-up survey data, because this data is not clustered, only evaluated to see how mental well-being evolves as the pandemic progresses.

2.3 Working of the clustering algorithms

2.3.1 K-Means clustering

The K-Means clustering algorithm is an iterative algorithm, and one of the simplest unsupervised learning techniques. It works by first defining a target number k, which represents the number of clusters that should be formed. For all k clusters, a centroid (cluster center) is defined. After this, each data point (in this case each participant) is assigned to the nearest centroid. The nearest centroid is calculated by taking the least squared Euclidean distance. When every datapoint is assigned to a cluster, the centroids of those clusters are updated by calculating the means of the features of the datapoints in each cluster. This assigning of datapoints to a cluster and updating the centroids is repeated until the centroids no longer change, and therefore the data is not reassigned. The clusters have been formed.

Two important parameters that need to be chosen when the clustering is performed is the number of clusters, and the initialization method used to initialize the centroids. The initialization method determines how the first centroids are chosen. In this study I chose the k-means++ initialization, which is the best initialization method in terms of speed and accuracy (Arthur and Vassilvitskii, 2007).

Common methods to determine the optimal number of clusters in the data for K-Means clustering are the elbow method and the silhouette score method, as for example used by Srividya et al. (2018). The elbow method works by calculating the within-cluster-sum-of-squares (WCSS) for K-Means clusterings with different numbers of clusters. The WCSS is defined by equation 2.1.

$$WCSS = \sum_{i=1}^{n}(X_i - Y_i)^2 \tag{2.1}$$

Here Y_i is the centroid corresponding to datapoint X_i, and n is the number of features (variables) of the datapoint. This method is based on the principle that clustering performance increases (WCSS decreases), when the number of clusters increases. However the rate of this increase is usually decreasing. Plotting the WCSS against increasing number of clusters can show an ‘elbow’ which indicates significant drop in rate of performance increase. The optimal number of clusters is the number corresponding to the elbow point.

The silhouette score method works by calculating the average silhouette coefficient (SC) for K-Means clusterings with a varying number of clusters. The SC is calculated by taking into account the mean intra-cluster distance (mean distance to the other instances in the same cluster) and the mean nearest-cluster distance (mean distance to the instances of the next closest cluster) for each data point. For each datapoint the SC is calculated using the equation 2.2.

$$SC = \frac{x - y}{max(x, y)} \tag{2.2}$$

Here y is the mean intra-cluster distance and x depicts mean nearest-cluster distance for a single datapoint. For every datapoint in the clustering this value is calculated and the average for all datapoints is taken. This average is taken for every K-Means clustering with a varying number of clusters. These average silhouette scores are plotted against the number of clusters. The silhouette score gives information about how well how well samples are clustered with other samples that are similar to each other. An average silhouette score with a value near 1 means the datapoints are mostly far away from neighbouring clusters. An average silhouette score with a value near 0 means the datapoints are mostly on or very close to the decision boundary between two neighboring clusters. An average silhouette score with a value near -1 means the datapoints are mostly in the wrong cluster. The optimal number of clusters is the number of clusters that scored the highest average SC for its K-Means clustering.

2.3.2 Hierarchical Agglomerative clustering

Besides K-Means clustering, Hierarchical Agglomerative clustering will also be performed to investigate whether the choice of clustering algorithm has an influence on the results. The Hierarchical Agglomerative clustering algorithm works on the basis of a bottom up approach. Initially, each datapoint is considered as a single-element cluster. At each step of the algorithm, two clusters that are the most similar are merged together into a new bigger cluster. This continues until a stopping criterion is satisfied, which in this case will be that a specific number of clusters is reached.

The same number of clusters will be used for this clustering algorithm as was found for K-Means clustering using the elbow and silhouette score method, so the two algorithms can be easily compared. Another important parameter is the linkage criterion, which specifies how exactly the most similar clusters are measured. I chose to use Ward's linkage, because this method picks the two clusters to merge such that the variance within all clusters increases the least. This often leads to clusters that are relatively equally sized. Besides this, this method is the most used and works on most datasets (Müller, Guido, et al., 2016).

2.4 Evaluation of the clusterings

In total three clusterings are performed, two K-Means clusterings and one Hierarchical Agglomerative clustering. The main technique used in this study is K-Means clustering. Hierarchical Agglomerative clustering is only performed to compare the results to the results of K-Means clustering, and to investigate whether the choice of clustering algorithm has an influence on the results.

2.4.1 Determine optimal number of clusters

Before the clusterings can be performed, the optimal number of clusters in the data is examined. As explained earlier, this is done using the elbow method and the silhouette score method. For these methods K-Means clustering with a varying number of clusters is used on the prepared scaled baseline data with all the variables, as shown in appendix A, included. When the optimal number of clusters in the data is validated, the actual K-Means clusterings can be performed for that amount of clusters. Hierarchical Agglomerative is performed for the same number of clusters as the K-Means clusterings. Each participant gets labeled with what cluster they belong to for the different clustering methods.

2.4.2 Clustering methods

The first performed clustering is K-Means clustering on the prepared scaled baseline data with all the variables included. To answer the first part of the research question, whether there exist different groups of people that differ in mental well-being during the pandemic, the positive and negative affect variables are evaluated, in each group observed by the K-Means clustering. This is done using the unscaled data without imputations.

To answer the second part of the research question, what other factors differ between the groups, the other variables in the dataset are evaluated for each cluster using the unscaled data without imputations. This is done to see if these variables differ between the groups, and if that is the case, how they differ.

To answer the last part of the research question, how the mental well-being of the observed groups evolves as the pandemic progresses, the data of the positive and negative affect variables in the follow-up surveys are used. These data can provide insight in how mental well-being of the observed clusters evolves as the pandemic progresses. The first wave is recorded on March 27th and the last wave is recorded on July 13th.

To validate the performed K-Means clustering, a second K-Means clustering is performed. This time not every baseline variable is included in the clustering, namely the affect variables are not included. This is done to see whether the clusters will be similar without the influence of the dependent affect variables, and whether the clustering is driven by the affect emotions or not. Only the other factors that do not say something directly about mental well-being influence the clustering. The same number of clusters are used for this second clustering as for the first clustering, so they can be easily compared.

Besides the K-Means clusterings, also Hierarchical Agglomerative clustering is performed, to investigate whether the choice of clustering algorithm has an influence on the results. This clustering is performed on the same scaled prepared baseline data as the first K-Means clustering. For this clustering also the same number of clusters are used as the K-Means clustering, so they can be easily compared. The observed clusters from the Hierarchical Agglomerative clustering are compared to the observed clusters from the first K-Means clustering, to see if they are similar.

2.4.3 Statistics

To examine the differences between the observed clusters for the three different clustering methods, statistics are used. These statistics are performed on the unscaled baseline data without the imputations. For the ordinal variables, among which the affect variables, first the mean values of each variable in the different clusters are calculated together with their standard error. To test whether there is a significant difference between the observed clusters for the ordinal variables in the dataset, multiple Mann-Whitney U tests are used, one for each variable. I chose for this statistical test, because the variables are ordinal. Since multiple statistical test are performed simultaneously, the p-values obtained by these tests need to be corrected. This is done using the Benjamini–Hochberg procedure with a family wise error rate of 0.05, to control the false discovery rate. For each clustering method, there are also boxplots created of the affect variables for each cluster, to better examine the values for these variables.

For the binomial variables in the dataset a different statistical test is used to examine if they difference between the observed clusters, for each of the three clustering methods. For the binomial variables, the percentage of successes for each variable for each cluster is calculated. To determine if the relative amount of successes is significantly different between the clusters, a χ^2 test is performed for each binomial variable. This statistical test is chosen, because it is suitable to test if the differences between proportions of groups are significant. Again, since multiple comparisons are done simultaneously, the p-values obtained by these tests need to be corrected. This is done the same way as described earlier, using the Benjamini–Hochberg procedure with a family wise error rate of 0.05.

With these statistics it is possible to see if and how the average mental well-being differs between clusters, and what other factors differ between the clusters in what way, for each clustering method.

Furthermore, for the first K-Means clustering with the affect variables included, the follow-up survey data is analyzed to examine how the mental well-being of the participants in the observed clusters evolve as the pandemic progresses. This is only done for this clustering, since this is the main clustering method used in this study. For each affect variable, for each wave of the follow-up survey, for each observed cluster of participants the mean values are calculated. Besides this, for each calculated mean, a standard error is also calculated. These means and standard errors are plotted, against the waves. With these plots it is possible to see how positive and negative affect on average changes over time, which is used to asses the mental well-being of the clusters as the pandemic progresses. The first wave is recorded on March 27th and the last wave is recorded on July 13th.

Finally, The percentages of people classified to the same clusters is calculated for the second K-Means clustering and the Hierarchical Agglomerative clustering, compared to first K-Means clustering. This gives information on how much the clusters observed by these clustering methods overlap with the first K-Means clustering clusters, and thus how similar they are.

2.5 Implementation

The data preparation, clustering and evaluation of the clustering is done using the programming language Python. The K-Means and Hierarchical Agglomerative clusterings are implemented with the use of the Scikit-learn library.

3 Results

First I asked whether there are groups of people that differ in mental well-being during the COVID-

19 pandemic. To answer this question, the baseline data of the PsyCorona survey with all countries (n=62,902) is clustered using the K-Means clustering and Hierarchical Agglomerative clustering algorithm, and the mental well-being of the observed clusters is evaluated using the affect variables. Secondly, I asked what other differences there are between these groups. To answer this question the values of the other variables in the data are evaluated to see if and how they differ between the observed clusters. Finally, I asked how these observed groups evolve as the pandemic progresses. To answer this question, the average mental well-being of the clusters is evaluated during the pandemic in a period between March 27th and July 13th 2020.

In total three clusterings are performed. First of all, K-Means clustering is performed on the data with the affect variables included. The results for this clustering will be presented first. Secondly, K-Means clustering is performed on the data without the affect variables included. This is done to see if the clustering will be similar without the influence of the dependent affect variables, and whether the clustering is driven by the affect emotions or not. Finally, Hierarchical Agglomerative Clustering is performed on the data with the affect variables included, to investigate if the choice of clustering algorithm has an effect on the results.

3.1 K-Means clustering

3.1.1 Determine number of clusters

To determine the optimal number of clusters in the data for K-Means clustering, the elbow method and silhouette score method are performed. Figure 3.1 (elbow method) shows that no large significant improvements in the fit of the K-means clustering model is shown after 2 clusters, meaning the optimal number of clusters seems to be 2. This result is reinforced by the the silhouette scores plotted in Figure 3.2. The silhouette score gives information about how well how well samples are clustered with other samples that are similar to each other. The silhouette score is the highest for 2 clusters, meaning the optimal number of clusters is 2.

3.1.2 Mental well-being in clusters

The optimal amount of clusters in the data is 2, so the K-Means clustering on the baseline data is

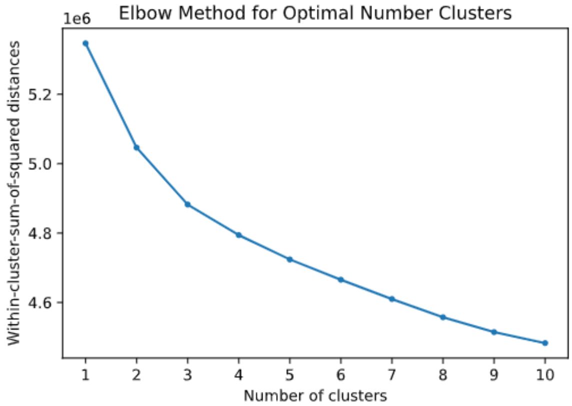

Figure 3.1: Elbow method for determining the optimal number of clusters for K-Means clustering on baseline data.

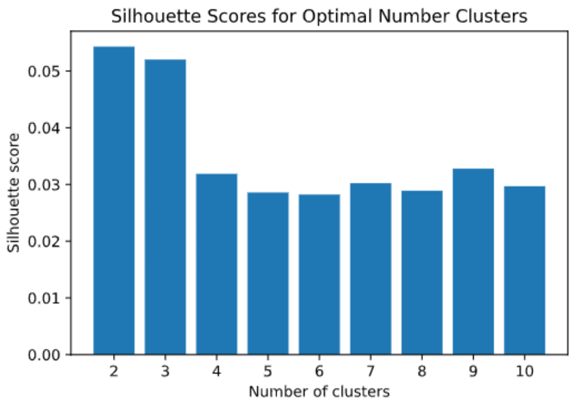

Figure 3.2: Silhouette score method for determining the optimal number of clusters for K-Means clustering on baseline data. A higher silhouette score is better.

performed with 2 clusters. Each participants is labelled with what cluster they belong to (cluster 1 or cluster 2), as to evaluate the clusters with the original unscaled data. The clustering divided the participants in two groups of roughly the same size. After this, I examined if these two clusters differed in average level of mental well-being. This is done using the positive and negative affect variables.

Figure 3.3 shows the average differences of the affect variables between the two clusters. The negative affect variables are: anxious, bored, depressed, exhausted, and nervous. The positive affect variables are: calm, energetic, inspired and relaxed.

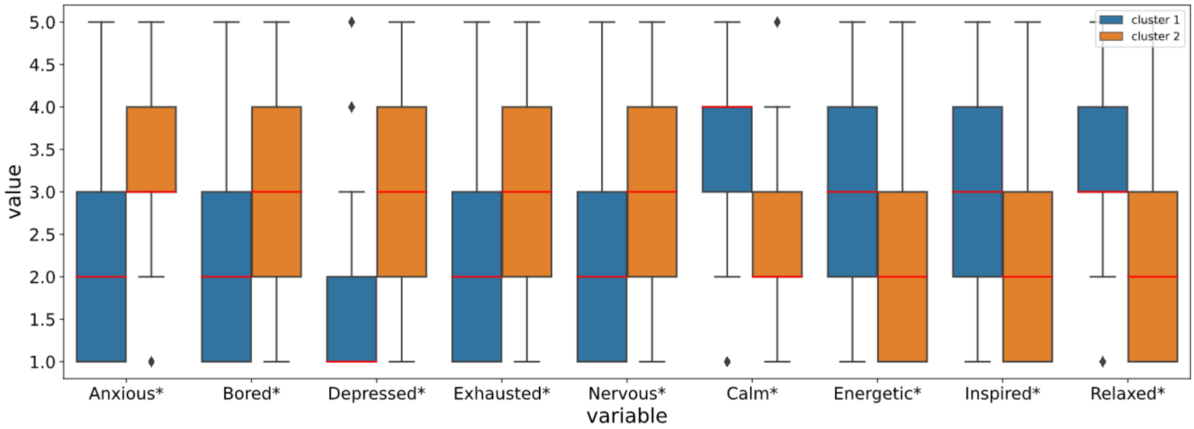

Figure 3.3: Boxplots of negative and positive affect variables for cluster 1 and cluster 2 for K-Means clustering with inclusion of affect variables. Blue represents cluster 1 and orange represents cluster 2. Variables that differ significantly between the two clusters are marked with a star. The red lines indicate the median.

This figure is used to determine what the differences between the clusters are regarding affect. The p-values for these variables can be found in appendix B in table B.1. All the affect variables were significantly different between the two clusters. People in cluster 1 on average were less anxious, bored, depressed, nervous and exhausted compared to the people in cluster 2. On the other hand, people in cluster 1 on average were more calm, energetic, inspired and relaxed compared to cluster 2. These results suggest that generally the individuals in cluster 1 have a better mental well-being compared to the individuals in cluster 2, at the baseline of the survey.

3.1.3 Other different factors between clusters

To find what the differences are between the people in each cluster besides their mental well-being, the other baseline variables in the survey are evaluated. These results are presented in appendix B, in table B.1 and table B.2. Note that previously in section 3.1.2, cluster 1 is defined as the cluster with on average a better mental well-being, compared to the cluster 2.

First of all, on average the age of the participants is significant different between the two clusters. People in cluster 1 are generally older compared to people in cluster 2.

Secondly, there is a significant smaller proportion of females present in cluster 1 compared to cluster 2.

Thirdly, the proportion of unemployed people is lower in cluster 1 compared to cluster 2. However, the average amount of hours worked that the people worked is not significantly different between the two clusters. Following this, evaluating the financial situation of the people in the clusters, on average the people cluster 2 think their personal situation will get even worse due to economic consequences of coronavirus, compared to the people in cluster 1. Besides this, people in cluster 2 are on average also more financially strained compared to cluster 1.

The amount of social contact with friends or relatives is also evaluated. On average, the people in cluster 1 had less in-person contact with friends or relatives compared to the people in cluster 2, although the difference is small. On the other hand, people in cluster 1 had more often online contact with friends or relatives compared to cluster 1.

Besides this, the social contact of the people in clusters with other people in general was evaluated. No significant difference between the two clusters was found for how often people had social in-person contact with other people in general. However, people in cluster 1 had more online contact with other people in general compared to people in cluster 2. Also, the people in cluster 1 felt on average less lonely, compared to cluster 2.

Besides this, I looked at the how often people left their house on average in the clusters. On average, the people in cluster 1 left the house more often compared to cluster 2. Moreover, I found that a greater proportion of people in cluster 1 left the house for leisure purposes alone, compared to cluster 2. On the other hand, a smaller proportion of people in cluster 1 left the house for leisure purposes with others, compare to cluster 2.

Moreover, people in cluster 1 were more confident that their country would be able to fight the virus, compared to cluster 2.

Finally, the proportion of people in cluster 1 knowing any infected people is smaller compared to the proportion in cluster 2. In addition, there is a smaller proportion of people in cluster 1 infected with the virus themselves, compared to cluster 2.

3.1.4 Mental well-being over time

To evaluate how the mental well-being of each cluster develops over time, the positive and negative affect variables are explored using the follow-up survey data. The first wave of the follow-up survey (wave 1) was recorded on March 27th 2020. The last wave of this dataset (wave 12) was recorded on July 13th 2020.

Figure 3.4 shows the mean values and standard errors of all the positive affect variables for each cluster for each wave. The distance between the waves in the plot is related to the actual time period between the recording of the waves. For all positive affect variables it is shown that over time the the two lines for the clusters follow approximately the same trajectory, although the values of cluster 1 are for every wave higher compared to cluster 2. Besides this, calmness, energeticness, inspiration, and relaxation for both clusters all increased slightly over time from wave 1 to wave 12.

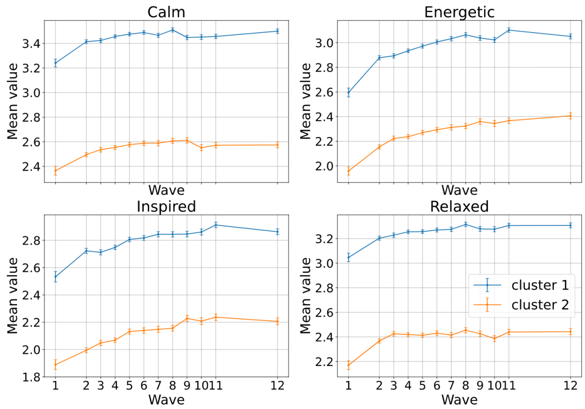

Figure 3.4: Mean values of the positive affect variables for each wave for each cluster, including the standard error. The blue line represents cluster 1. The orange line represents cluster 2.

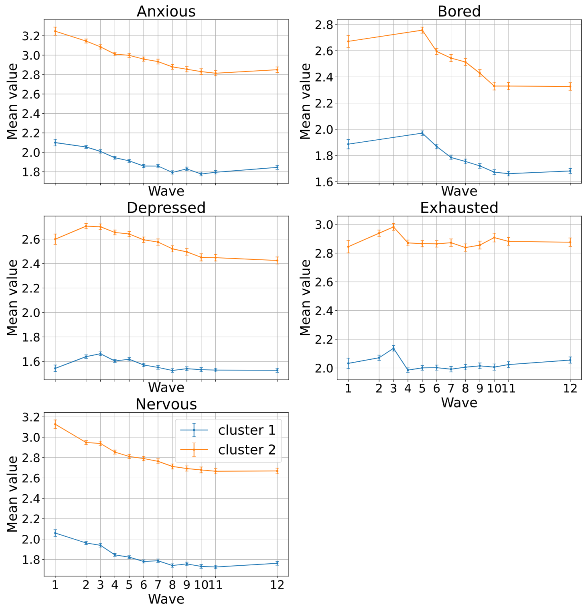

Figure 3.5: Mean values of the negative affect variables for each wave for each cluster, including the standard error. The blue line represents cluster 1. The orange line represents cluster 2.

Figure 3.5 shows the mean values and standard errors of all the negative affect variables for each cluster for each wave. This figure shows that for all negative affect variables the lines for the clusters follow approximately the same trajectory, while the values of cluster 2 are always higher compared to cluster 1. Besides this, anxiety, boredom and nervousness all decrease over time for both clusters from wave 1 to wave 12. This is not the case for the other two affect variables. Depression seems to stay the same on average for cluster 1, when comparing wave 1 to wave 12. For cluster 2, depression decreases slightly over time. Besides this, Exhaustion seems to stay the same for both clusters from wave 1 to wave 12.

Both Figure 3.4 and Figure 3.5 show that generally the mental well-being of both clusters increases

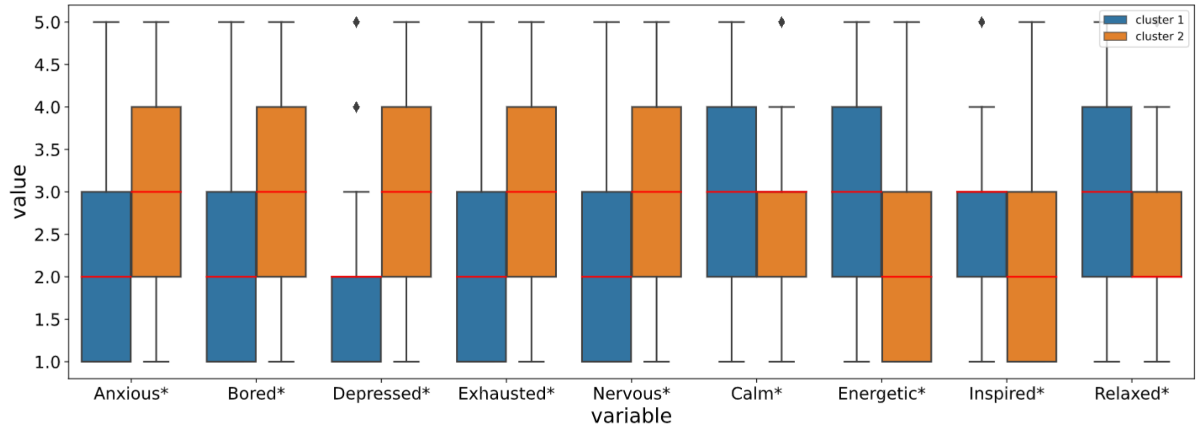

Figure 3.6: Boxplots of negative and postive affect variables for cluster 1 and cluster 2 for K-Means clustering without the inclusion of affect variables. Blue represents cluster 1 and orange represents cluster 2. Variables that differ significantly between the two clusters are marked with a star. The red lines indicate the median.

from March 27th to July 13th, while still maintaining a difference between the clusters. This means that the mental well-being of both the groups of people with initially better (cluster 1) and initially worse (cluster 2) mental well-being increases from March 27th to July 13th.

3.1.5 Similarity clusters when affect emotions removed from the mix

One may wonder whether the affect variables are what drives the clustering, or whether the two groups groups are still visible in the data when affect is removed from the mix. A second K-Means clustering is performed on the baseline data without the affect variables included. The percentage of people classified to the same cluster in both clusterings is calculated, to compare the two clusterings.

The proportion of people classified to the same cluster in the second clustering compared to the first clustering is 80.81%. This means that 80.81% percent of the people classified to cluster 1 or 2 in the first clustering, are classified to the same cluster (1 or 2) in the second clustering. This also means that 19.19% of the people classified to cluster 1 in the first clustering are now classified to cluster 2 or vice versa.

Moreover, Figure 3.6 shows the average differences of the affect variables between the clusters for this new clustering. The p-values for these variables can be found in in appendix C in table C.1. These boxplots will be compared to the boxplots of the first clustering shown earlier in Figure 3.3.

Figure 3.6 shows that people in cluster 1 were less anxious, bored, depressed, nervous and exhausted compared to the people in cluster 2. Besides this, people in cluster 1 were more calm, energetic, inspired and relaxed compared to cluster 2. This is similar to what was discussed for the first clustering with the inclusion of the affect variables. However, when looking at the boxplots in Figure 3.3 and 3.6 and the means in tables B.1 and C.1, the differences for all the affect variables between the clusters seem to be smaller when comparing the second clustering to the first.

In summary, the results are relatively similar when comparing the second clustering without the inclusion of the affect variables to the first clustering with the inclusion of the affect variables. The second clustering was also able to find two groups of people that responded mentally different to the COVID-19 pandemic. Similar clusters can be found when the affect emotions are removed from the mix. The observed clusters in the first clustering are not only driven by affect, but also by the other variables in the dataset. This means that the other variables included in the clustering, such as the factors differing between the clusters described earlier, influence the mental well-being of individuals in the clusters.

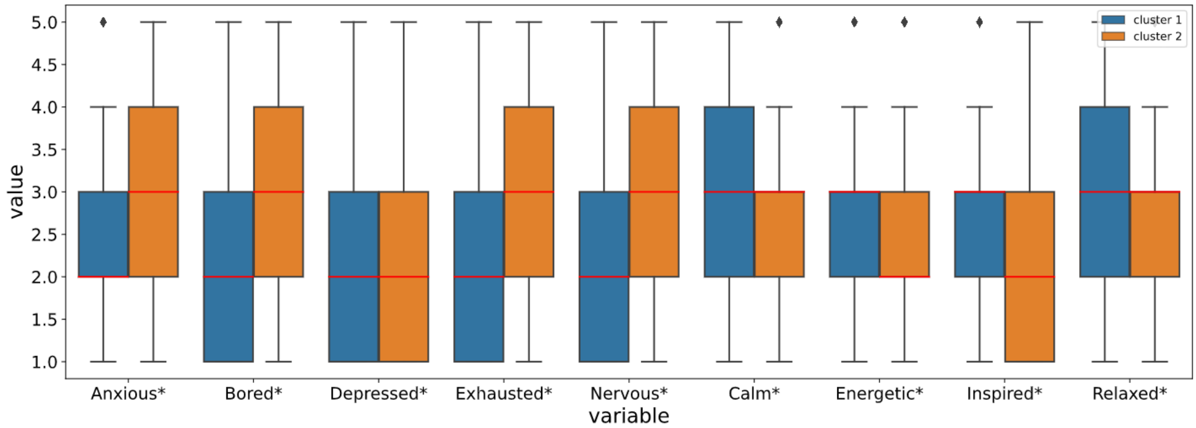

Figure 3.7: Boxplots of negative and positive affect variables for cluster 1 and cluster 2 for Hierarchical Agglomerative clustering. Blue represents cluster 1 and orange represents cluster 2. Variables that differ significantly between the two clusters are marked with a star. The red line indicates the median.

3.2 Hierarchical Agglomerative clustering

Finally, to investigate if the results would differ for a different clustering algorithm than K-Means clustering, Hierarchical Agglomerative clustering is performed. This is done on the same baseline data as the first K-Means clustering. As before, 2 clusters are used for the clustering. When the clustering is performed each participant is labelled with what cluster they belong to.

The proportion of people classified to the same cluster in the Hierarchical Agglomerative clustering compared to the first K-Means clustering is 68.18%. This means that 68.18% percent of the people classified to cluster 1 or 2 in the K-Means clustering are classified to the same cluster (1 or 2) in the Hierarchical Agglomerative clustering.

Besides this, the average mental well-being of these new clusters are evaluated. Figure 3.7 shows the average differences of the affect variables between the clusters for this new clustering. The p-values for these variables can be found in in appendix D in table D.1. Although all affect variables are significant different between the two clusters, Figure 3.7 shows that depression and energeticness have generally the same values for both clusters. Besides this, the differences between the cluster for the other affect variables also seem smaller compared to the differences for the first K-Means clustering shown in Figure 3.3. The differences between the clusters are also even smaller compared to the second K-Means clustering in Figure 3.6.

The Hierarchical Agglomerative clustering algorithm was not able the divide the data in two groups differing in mental well-being as clearly as the K-Means clustering algorithm. This means the choice of clustering algorithm had an influence on the results of the clustering.

4 Discussion

First I asked whether there are groups of people that differ in mental well-being during the COVID-19 pandemic. I hypothesized that there probably are groups of people with better mental well-being and groups of people with worse mental well-being during the COVID-19 pandemic, because it is likely that not everyone responds mentally the same to the pandemic. I found that K-Means clustering is able to distinguish two groups of people in the PsyCorona data that clearly differ in mental well-being. The people in the first group were on average more calm, energetic, inspired and relaxed, compared to the second group. Besides this, people in the second group were more anxious, bored, depressed, exhausted and nervous compared to the first group. This result is reinforced by the K-Means clustering without the inclusion of these affect variables.

Approximately the same result were found when affect had no influence on the clustering. Even when these emotions were removed from the mix, two groups of people could be distinguished that differ clearly in mental well-being in the same way, although the differences in mental well-being between the two groups were slightly smaller. In contrast to the K-Means clustering algorithm, the Hierarchical Agglomerative clustering algorithm could not as clearly find two groups differing in mental well-being in the data. The people in the first group were on average more calm, energetic, inspired and relaxed, compared to the second group, while people in the second group were more anxious, bored, depressed, exhausted and nervous compared to the first group. However, the differences between the groups were much smaller compared to the groups formed by K-Means clustering. This means K-Means clustering was the most effective in dividing the data in groups of people that differ in mental well-being. The hypothesis is confirmed by the results in this thesis. This study concludes that their are two groups of people that respond differently during the pandemic. One group has a better mental well-being, while the other group has a worse mental well-being during the pandemic. It is however not certain if the COVID-19 pandemic caused the formation of these groups, or if these groups already existed, because there is no data from the participants from before the start of the pandemic.

For future research these labels generated by K-Means clustering on the PsyCorona data can be used as labels for research with supervised machine learning about this topic. This is for example also done in the study by Srividya et al. (2018). First labels about mental health were obtained by using clustering algorithms. After this, supervised learning techniques, such as a random forest classifier, were deployed to predict the mental health of an individual. The labels generated by this study can also be used for that purpose, without inventing labels yourself for this dataset. With supervised learning the mental well-being of an individual can be predicted, and possibly any further mental health issues can be prevented.

Secondly I asked, what other differences there could be between the observed groups. I hypothesized that in the group of people with a worse mental well-being, compared to the groups with better mental well-being: people are younger, there is a higher proportion of females, higher proportion of people with no work, people have more financial worries, people leave the house less frequent, have less in-person contact with friends, relatives or other people in general, less confidence in the government able to fight the virus and knowing more infected people.

I found that it was indeed the case that a higher proportion of females was present in the worse mental well-being group, and people were on average younger, compared to the better mental well-being group. This means that generally females and young people suffer more mentally during the pandemic compared to males and older people.

Secondly, I found that there was indeed a higher proportion of unemployed people in the worse mental well-being group, compared to the better mental well-being group. However, the average amount of hours worked per group was found to be same. This means that people without work suffer mentally more compared to people with work. People also had more financial worries in the worse mental well-being cluster compared to the better mental well-being cluster, which matches the hypothesis. Having financial worries during the pandemic could have a negative effect on mental well-being of an individual.

Thirdly, this study found that the worse mental well-being group had more in-person contact with friends and relatives, compared to the better mental well-being group. No difference between the two groups was found for how often people had social in-person contact with other people in general. This does not match the hypothesis. People with a worse mental well-being had more face to face contact with people, compared to people with a better mental well-being. The difference between the two groups is nevertheless relatively small. It is a possibility that the mental well-being of these people was worse because of a stigma on meeting with friend and relatives when this is not allowed by government rules. Future research on this matter may provide more insight into this matter. In contrast to this, I however found that the better mental well-being cluster had more online contact with friends, relatives or other people in general. Staying in touch with friends and relatives via voice or video chat, or other online meetings in general could be good for the mental well-being of individuals during the pandemic.

This study also found that people in the better mental well-being group left the house more often compared to the worse mental well-being group, which matches the hypothesis. A greater proportion of people in the better mental well-being group left the house for leisure purposes alone, compared to the worse mental well-being group. However, a greater proportion of the worse mental well-being group left the house for leisure purposes with others. This means people with a better mental well-being go more often running or walking for example. People with a worse mental well-being met up more often with their friends and families for activities. This is consistent with the finding that people with a worse mental well-being had more in person contact with friends and relatives. It is possible that the people in this group did not take the initiative to go outside, but their friends and relatives did.

Furthermore, people in the better mental well-being group were more confident in their government to be able to fight the virus, compared to the worse mental well-being group. This finding matches the hypothesis and suggests that clear messages and information about the virus provided by the government via media during the pandemic could lead to better mental well-being among the people. It is possible there would be less unwarrented public fear, described as a factor by Gualano et al. (2020), if the government is transparent and does the right things.

Finally, people in the better mental well-being group knew generally less people infected with the coronavirus compared to the worse mental well-being group. Additionally, a smaller proportion of people in the better mental well-being group had been infected with the virus themselves compared to the worse mental well-being group. This suggests that generally knowing infected people or being infected has a negative influence on the mental well-being of an individual, which matches the hypothesis. When you know more people infected by the virus, you might be more worried about getting infected yourself, which was described as a factor that had a negative influence on mental well-being by Choi et al. (2020).

As mentioned by Pfefferbaum and North (2020) the COVID-19 pandemic has alarming implications for the mental health of people. The mental health of especially females, young people, unemployed people and people with financial worries should be closely monitored during this pandemic in all countries over the world. Furthermore, voice and video chatting with friends and relative, leaving the house for runs/walks or other activities alone and clear government information could have a positive effect on mental well-being.

Finally I asked, how the mental well-being of these two groups evolves as the pandemic progresses. I hypothesized that the average mental well-being of the people in both groups (better and worse mental well-being groups) would probably stay the same as the pandemic progressed. To test this hypothesis the average mental well-being of the earlier defined groups are evaluated during the pandemic from March 27th to July 13th 2020. In this study I found that generally speaking the average mental well-being increased in both the better and worse mental well-being groups worldwide from March 27th to July 13th 2020, while the difference in mental well-being between the groups over time is maintained. These results do not match the hypothesis. However, these results should be interpreted with caution, because an important factor is not kept constant across this time period. Because the severity of the coronavirus situation was different for each country in the world during this time period, the stringency of the measures taken by the governments could also have changed over time. It could be the case that the mental health of the two groups increased because the stringency of measures decreased in some countries. A decrease in the stringency of the measures could cause people to for example leave the house more often, which could increase mental well-being. However, this is not sure, because the measures differ for each country and this study focused on data from countries all over the world.

For future research, the data of the PsyCorona survey could be clustered using K-Means clustering for individual countries. This way, the clusters could be evaluated over time, which could be related to the stringency of the measures taken in a specific country. It is likely that the mental well-being over time is directly affected by the stringency of the measure take in a country. Groups in different countries then should respond differently, according to the measures taken in the country.

In summary, I found two clear clusters of people that differed in positive and negative affect during the COVID-19 pandemic. Other factors that differed between these clusters were age, gender, employment, financial worries, social contact, frequency of leaving the house, knowledge about the virus, confidence in government, being infected and knowing infected people. As the pandemic progressed, the mental well-being of both groups slightly increased. For future research, the data of individual countries could be clustered to investigate the impact of specific COVID-19 measures. Furthermore, the obtained cluster labels can be used for supervised learning purposes, to predict mental well-being of individuals.

References

Gerrit Antonides and Eveline van Leeuwen. Covid-19 crisis in the Netherlands: "Only together we can control Corona". *Mind & Society*, September 2020. ISSN 1860-1839.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

Aabishkar Bhattarai and Bijaya Karki. Covid-19 pandemic and mental health issues. *Journal of Lumbini Medical College*, 8(1):181–182, Jun. 2020.

Subhagata Chattopadhyay, Preetisha Kaur, Fethi Rabhi, and U. Rajendra Acharya. Neural Network Approaches to Grade Adult Depression. *Journal of Medical Systems*, 36(5):2803–2815, October 2012. ISSN 1573-689X.

Edmond Pui Hang Choi, Bryant Pui Hung Hui, and Eric Yuk Fai Wan. Depression and Anxiety in Hong Kong during COVID-19. *International Journal of Environmental Research and Public Health*, 17(10):3740, May 2020. ISSN 1660-4601.

Maria Rosaria Gualano, Giuseppina Lo Moro, Gianluca Voglino, Fabrizio Bert, and Roberta Siliquini. Effects of covid-19 lockdown on mental health and sleep disturbances in italy. *International Journal of Environmental Research and Public Health*, 17(13), 2020. ISSN 1660-4601.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Unsupervised Learning*, pages 485–585. Springer New York, New York, NY, 2009. ISBN 978-0-387-84858-7.

Stania Kamara, Anna Walder, Jennifer Duncan, Antoinet Kabbedijk, Peter Hughes, and Andrew Muana. Mental health care during the ebola virus disease outbreak in sierra leone. *Bulletin of the World Health Organization*, 95:842–847, 12 2017.

Andreas C Müller, Sarah Guido, et al. *Introduction to machine learning with Python: a guide for data scientists.* " O'Reilly Media, Inc.", 2016.

Kuan-Yu Pan, Almar A L Kok, Merijn Eikelenboom, Melany Horsfall, Frederike Jörg, Rob A Luteijn, Didi Rhebergen, Patricia van Oppen, Erik J Giltay, and Brenda W J H Penninx. The mental health impact of the covid-19 pandemic on people with and without depressive, anxiety, or obsessive-compulsive disorders: a longitudinal study of three dutch case-control cohorts. *The lancet. Psychiatry*, December 2020. ISSN 2215-0366.

Betty Pfefferbaum and Carol S. North. Mental health and the covid-19 pandemic. *New England Journal of Medicine*, 383(6):510–512, 2020.

Christoph Pieh, Sanja Budimir, and Thomas Probst. The effect of age, gender, income, work, and physical activity on mental health during coronavirus disease (covid-19) lockdown in austria. *Journal of Psychosomatic Research*, 136: 110186, 2020. ISSN 0022-3999.

M. Srividya, S. Mohanavalli, and N. Bhalaji. Behavioral Modeling for Mental Health using Machine Learning Algorithms. *Journal of Medical Systems*, 42(5):88, April 2018. ISSN 1573-689X.

Julio Torales, Marcelo O'Higgins, João Mauricio Castaldelli-Maia, and Antonio Ventriglio. The outbreak of covid-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, 66(4):317–320, 2020.

WHO. WHO Coronavirus Disease (COVID-19) Dashboard, 2021. URL https://covid19.who.int/. Accessed: 26-01-2021.

A Appendix: meaning of variables in the dataset

Table A.1: Ordinal variables in the dataset, their meaning and their minimal and maximal possible values.

Variable name	Question asked	Min value	Max value
affAnx	How did you feel over the last week? - Anxious	1	5
affBor	How did you feel over the last week? - Bored	1	5
affCalm	How did you feel over the last week? - Calm	1	5
affDepr	How did you feel over the last week? - Depressed	1	5
affEnerg	How did you feel over the last week? - Energetic	1	5
affExh	How did you feel over the last week? - Exhausted	1	5
affInsp	How did you feel over the last week? - Inspired	1	5
affNerv	How did you feel over the last week? - Nervous	1	5
affRel	How did you feel over the last week? - Relaxed	1	5
age	What is your age?	1	8
bor02	Indicate your agreement or disagreement with the following statements. - Time is moving very slowly.	-3	3
c19Eff	Agree or disagree: - I think that the country that I'm living in is able to fight the Coronavirus.	-3	3
c19Hope	Agree or disagree: - I have high hopes that the situation regarding coronavirus will improve.	-3	3
C19Know	How knowledgeable are you about the recent outbreak of Covid-19, commonly referred to as the Coronavirus, in the country I'm living in?	1	5
c19perBeh01	To minimize my chances of getting coronavirus, I... - ...wash my hands more often.	-3	3
c19perBeh02	To minimize my chances of getting coronavirus, I... - ...avoid crowded spaces.	-3	3
c19perBeh03	To minimize my chances of getting coronavirus, I... - ...put myself in quarantine.	-3	3
c19ProSo01	I am willing to... - ...help others who suffer from coronavirus.	-3	3
c19ProSo03	I am willing to... - ...protect vulnerable groups from coronavirus even at my own expense.	-3	3
c19RCA01	I would sign a petition that supports... - ...mandatory vaccination once a vaccine has been developed for coronavirus.	-3	3
c19RCA02	I would sign a petition that supports... - ...mandatory quarantine for those that have coronavirus and those that have been exposed to the virus.	-3	3

consp01	I think that...... - . . . many very important things happen in the world, which the public is never informed about.	0	10
consp02	I think that...... - . . . politicians usually do not tell us the true motives for their decisions.	0	10
disc01	Agree or disagree: - I fear that things will go wrong in society.	-2	2
disc02	Agree or disagree: - I feel concerned when I think about the future of society.	-2	2
disc03	Agree or disagree: - I am satisfied with society.	-2	2
discPers	Do you have anyone with whom you can discuss very personal matters?	-1	1
ecoHope	Agree or disagree: - I have high hopes that the situation regarding the economic and financial consequences of coronavirus will improve.	-3	3
ecoProSo01	To help with the economic and financial consequences of coronavirus, I am willing to... - ...help others who suffer from such consequences.	-3	3
ecoProSo03	To help with the economic and financial consequences of coronavirus, I am willing to... - ...protect vulnerable groups from such consequences, even at my own expense.	-3	3
ecoRCA02	If it would alleviate the economic and financial consequences of coronavirus, I would sign a petition that supports... - ...giving the government more authority over people.	-3	3
ecoRCA03	If it would alleviate the economic and financial consequences of coronavirus, I would sign a petition that supports... - ...increased government spending.	-3	3
edu	What is your highest level of education?	1	7
employstatus_123	How much hours did you work during the last week?	0	3
fail01	Agree or disagree: - Not a lot is done for people like me in the country I'm living in.	-2	2
happy	In general, how happy would you say you are?	1	10
houseLeaveAmount	In the past week, how often did you leave your home?	1	4
isoFriends_inPerson	In the past 7 days, how much social contact have you had with people who live outside your househ... - In the past 7 days, how many days did you have in-person (face-to-face) contact with ... - ...friends or relatives	0	7

isoFriends_online	In the past 7 days, how much social contact have you had with people who live outside your househ... - In the past 7 days, how many days did you have online (video or voice) contact with ... - ...friends or relatives	0	7
isoOthPpl_inPerson	In the past 7 days, how much social contact have you had with people who live outside your househ... - In the past 7 days, how many days did you have in-person (face-to-face) contact with ... - ...other people in general	0	7
isoOthPpl_online	In the past 7 days, how much social contact have you had with people who live outside your househ... - In the past 7 days, how many days did you have online (video or voice) contact with ... - ...other people in general	0	7
lifeSat	In general, how satisfied are you with your life?	1	6
lone01	During the past week, did you... - ...feel lonely?	1	5
MLQ	My life has a clear sense of purpose.	-3	3
neuro01	I see myself as someone who... - ...is very concerned.	-3	3
neuro02	I see myself as someone who... - ...easily gets nervous.	-3	3
neuro03	I see myself as someone who... - ...is relaxed, can easily deal with stress.	-3	3
para01	I need to be on my guard against others	0	10
para02	People are trying to make me upset	0	10
para03	Strangers and friends look at me critically	0	10
PFS01	Agree or disagree: - I am financially strained.	-2	2
PLRAC19	How likely is it that the following will happen to you in the next few months? - You will get infected with coronavirus.	1	8
PLRAEco	How likely is it that the following will happen to you in the next few months? - Your personal situation will get worse due to economic consequences of coronavirus.	1	8
posrefocus01	When dealing with stressful situations, what do you usually do? - I distract myself to avoid thinking about the subject.	1	5
posrefocus02	When dealing with stressful situations, what do you usually do? - I do things to distract myself from my thoughts and feelings.	1	5
posrefocus03	When dealing with stressful situations, what do you usually do? - I force myself to think about something else.	1	5
probSolving01	When dealing with stressful situations, what do you usually do? - I try to come up with a strategy about what to do.	1	5

probSolving02	When dealing with stressful situations, what do you usually do? - I make a plan of action.	1	5
probSolving03	When dealing with stressful situations, what do you usually do? - I think hard about what steps to take.	1	5
tempFocFut	Agree or disagree: - I think about what my future has in store.	-3	3
tempFocPast	Agree or disagree: - I replay memories of the past in my mind.	-3	3
tempFocPres	Agree or disagree: - I focus on what is currently happening in my life.	-3	3
tightLoose	To what extent do you think that the country you currently live in should have the following characteristics right now? - 1\: Be loose:9\: Be tight	1	9
tightNorms	To what extent do you think that the country you currently live in should have the following characteristics right now? - 1\: Have flexible social norms:9\: Have rigid social norms	1	9
tightTreat	To what extent do you think that the country you currently live in should have the following characteristics right now? - 1\: Treat people who don't conform to norms kindly:9\: Treat people who don't conform to norms harshly	1	9

Table A.2: Binomial variables in the dataset and their meaning.

Variable name	**Question asked**
coronaClose_1	Do you personally know anyone who currently has coronavirus? - Yes, myself
coronaClose_2	Do you personally know anyone who currently has coronavirus? - Yes, a member of my family
coronaClose_3	Do you personally know anyone who currently has coronavirus? - Yes, a close friend
coronaClose_4	Do you personally know anyone who currently has coronavirus? - Yes, someone I know
coronaClose_5	Do you personally know anyone who currently has coronavirus? - Yes, someone else
coronaClose_6	Do you personally know anyone who currently has coronavirus? - No, I do not know anyone
employstatus_10	Which of the following categories best describes your employment status during the last week? - Volunteering
employstatus_4	Which of the following categories best describes your employment status during the last week? - Not employed, looking for work
employstatus_5	Which of the following categories best describes your employment status during the last week? - Not employed, not looking for work
employstatus_6	Which of the following categories best describes your employment status during the last week? - Homemaker

employstatus_7	Which of the following categories best describes your employment status during the last week? - Retired
employstatus_8	Which of the following categories best describes your employment status during the last week? - Disabled, not able to work
employstatus_9	Which of the following categories best describes your employment status during the last week? - Student
gender	What is your gender?
houseLeaveWhy_1	I had to go to work.
houseLeaveWhy_2	I had errands to run.
houseLeaveWhy_4	For leisure purposes with others (e.g., meeting up with friends, seeing family, going to the cinema, etc.)
houseLeaveWhy_6	Selected Choice Other, please specify:
houseLeaveWhy_7	For leisure purposes alone (e.g., running, going for a walk, etc.)
relYesNo	Are you religious?

B Appendix: statistics variables K-Means clustering with affect variables included

Table B.1: Mean values in each cluster for each ordinal variable, together with the standard error of the mean and the corrected p-value calculated using Mann-Whitney U test. The values are rounded to 4 decimal places. The variables highlighted in italic are the factors discussed in the results section.

	Mean: 1	SEM: 1	Mean: 2	SEM: 2	p-value
affAnx	2.1086	0.0058	3.3446	0.0064	<0.0001
affBor	2.2441	0.0067	3.1952	0.0072	<0.0001
affCalm	3.4193	0.0054	2.4241	0.0056	<0.0001
affDepr	1.5732	0.0045	2.9081	0.0066	<0.0001
affEnerg	2.9780	0.0059	2.1615	0.0056	<0.0001
affExh	1.9993	0.0059	3.0065	0.0067	<0.0001
affInsp	2.7928	0.0064	2.0636	0.0058	<0.0001
affNerv	1.9780	0.0054	3.2088	0.0064	<0.0001
affRel	3.2033	0.0057	2.2623	0.0057	<0.0001
age	3.2058	0.0092	2.6036	0.0084	<0.0001
bor02	-0.4163	0.0100	0.5194	0.0104	<0.0001
c19Eff	1.4072	0.0079	0.3516	0.0093	<0.0001
c19Hope	1.6950	0.0071	0.8024	0.0091	<0.0001
C19Know	3.8290	0.0045	3.6317	0.0049	<0.0001
c19perBeh01	2.4713	0.0051	2.1733	0.0070	<0.0001
c19perBeh02	2.5781	0.0046	2.2981	0.0066	<0.0001
c19perBeh03	1.9274	0.0081	1.7776	0.0085	<0.0001
c19ProSo01	1.2132	0.0078	0.7134	0.0086	<0.0001
c19ProSo03	0.7501	0.0090	0.3239	0.0096	<0.0001
c19RCA01	1.2950	0.0104	1.2628	0.0104	<0.01
c19RCA02	2.1463	0.0070	1.9861	0.0078	<0.0001
consp01	6.6225	0.0155	7.0960	0.0144	<0.0001
consp02	6.9098	0.0149	7.4276	0.0140	<0.0001
disc01	0.3558	0.0057	0.9289	0.0050	<0.0001
disc02	0.6099	0.0056	1.0679	0.0049	<0.0001
disc03	-0.1286	0.0056	-0.6799	0.0055	<0.0001
discPers	0.8333	0.0031	0.6383	0.0045	<0.0001
ecoHope	0.9835	0.0091	0.1945	0.0100	<0.0001
ecoProSo01	1.0318	0.0076	0.5221	0.0085	<0.0001
ecoProSo03	0.6926	0.0088	0.2645	0.0093	<0.0001
ecoRCA02	-0.1924	0.0107	-0.3941	0.0105	<0.0001
ecoRCA03	0.8139	0.0097	0.6572	0.0100	<0.0001
edu	4.5268	0.0081	4.2563	0.0079	<0.0001
employstatus123_merge	2.2745	0.0057	2.2153	0.0064	0.1792
fail01	-0.3791	0.0063	0.2655	0.0062	<0.0001
happy	7.3518	0.0088	5.3092	0.0109	<0.0001
houseLeaveAmount	2.4644	0.0060	2.2982	0.0058	<0.0001
isoFriends_inPerson	2.0296	0.0136	2.0670	0.0136	<0.05
isoFriends_online	4.6377	0.0137	4.1750	0.0142	<0.0001

isoOthPpl_inPerson	1.9474	0.0124	1.9235	0.0125	0.0717
isoOthPpl_online	3.0451	0.0152	2.6535	0.0148	<0.0001
lifeSat	4.7416	0.0049	3.5278	0.0069	<0.0001
lone01	1.8733	0.0052	2.9672	0.0062	<0.0001
MLQ	1.5065	0.0069	0.1760	0.0090	<0.0001
neuro01	-0.0833	0.0093	1.0987	0.0079	<0.0001
neuro02	-0.6556	0.0090	0.8322	0.0087	<0.0001
neuro03	1.0549	0.0073	-0.3325	0.0085	<0.0001
para01	5.3886	0.0167	6.2659	0.0151	<0.0001
para02	2.2818	0.0137	3.9129	0.0158	<0.0001
para03	1.9617	0.0131	3.5713	0.0160	<0.0001
PFS01	-0.4086	0.0065	0.3726	0.0065	<0.0001
PLRAC19	3.3135	0.0079	3.8126	0.0082	<0.0001
PLRAEco	3.9179	0.0095	4.9241	0.0098	<0.0001
posrefocus01	2.9367	0.0059	3.2125	0.0058	<0.0001
posrefocus02	3.1513	0.0058	3.3357	0.0056	<0.0001
posrefocus03	2.8475	0.0059	3.1027	0.0057	<0.0001
probSolving01	4.0281	0.0048	3.5477	0.0055	<0.0001
probSolving02	3.7737	0.0055	3.2745	0.0058	<0.0001
probSolving03	3.9684	0.0049	3.5922	0.0054	<0.0001
tempFocFut	1.4372	0.0069	1.4142	0.0076	0.0887
tempFocPast	0.3705	0.0095	0.9913	0.0089	<0.0001
tempFocPres	1.5972	0.0058	1.0288	0.0073	<0.0001
tightLoose	6.0820	0.0138	5.9499	0.0140	<0.0001
tightNorms	5.6396	0.0140	5.4964	0.0143	<0.0001
tightTreat	5.6750	0.0144	5.5600	0.0145	<0.0001

Table B.2: Percentage of successes in each cluster for each binomial variable, together with the corrected p-value calculated using χ^2 test. The values are rounded to 4 decimal places. The variables highlighted in italic are the factors discussed in the results section.

	Percentage: 1	**Percentage: 2**	**p-value**
coronaClose_1	0.5779	2.1193	<0.0001
coronaClose_2	2.4822	3.6208	<0.0001
coronaClose_3	3.3190	4.3636	<0.0001
coronaClose_4	11.2392	12.7897	<0.0001
coronaClose_5	10.7339	11.7749	<0.0001
coronaClose_6	75.9932	71.6257	<0.0001
employstatus_10	2.4916	2.0489	<0.0001
employstatus_4	5.5138	12.0118	<0.0001
employstatus_5	4.1622	5.9547	<0.0001
employstatus_6	7.3423	8.0260	<0.0001
employstatus_7	12.3192	5.7626	<0.0001
employstatus_8	1.0927	2.5099	<0.0001
employstatus_9	16.3551	24.0780	<0.0001
gender (female)	57.2000	65.4100	<0.0001
houseLeaveWhy_1	22.4689	19.4263	<0.0001

houseLeaveWhy_2	41.3977	39.6722	<0.0001
houseLeaveWhy_4	4.8917	7.2384	<0.0001
houseLeaveWhy_6	23.3847	20.1754	<0.0001
houseLeaveWhy_7	20.6120	16.8011	<0.0001
relYesNo	51.4116	46.8818	<0.0001

C Appendix: statistics variables K-Means clustering without affect variables included

Table C.1: Mean values in each cluster for each ordinal variable, together with the standard error of the mean and the corrected p-value calculated using Mann-Whitney U test. The values are rounded to 4 decimal places.

	Mean: 1	SEM: 1	Mean: 2	SEM: 2	p-value
affAnx	2.4869	0.0064	3.0237	0.0075	<0.0001
affBor	2.4609	0.0068	3.0423	0.0078	<0.0001
affCalm	3.1731	0.0056	2.6101	0.0064	<0.0001
affDepr	1.8863	0.0055	2.6838	0.0075	<0.0001
affEnerg	2.8113	0.0058	2.2684	0.0062	<0.0001
affExh	2.2837	0.0063	2.7757	0.0074	<0.0001
affInsp	2.6666	0.0062	2.1299	0.0064	<0.0001
affNerv	2.3440	0.0062	2.9036	0.0073	<0.0001
affRel	2.9587	0.0059	2.4525	0.0065	<0.0001
age	3.0874	0.0086	2.6765	0.0092	<0.0001
bor02	-0.1520	0.0099	0.3035	0.0110	<0.0001
c19Eff	1.4063	0.0075	0.2167	0.0098	<0.0001
c19Hope	1.6900	0.0068	0.6939	0.0097	<0.0001
C19Know	3.8949	0.0041	3.5225	0.0052	<0.0001
c19perBeh01	2.6061	0.0039	1.9633	0.0081	<0.0001
c19perBeh02	2.7016	0.0034	2.1048	0.0076	<0.0001
c19perBeh03	2.1230	0.0069	1.5092	0.0097	<0.0001
c19ProSo01	1.3989	0.0069	0.4127	0.0091	<0.0001
c19ProSo03	0.9894	0.0081	-0.0356	0.0100	<0.0001
c19RCA01	1.5421	0.0092	0.9441	0.0115	<0.0001
c19RCA02	2.3144	0.0058	1.7514	0.0091	<0.0001
consp01	6.7332	0.0145	7.0193	0.0155	<0.0001
consp02	7.0107	0.0139	7.3693	0.0151	<0.0001
disc01	0.4895	0.0054	0.8324	0.0056	<0.0001
disc02	0.7563	0.0052	0.9405	0.0056	<0.0001
disc03	-0.1693	0.0054	-0.6991	0.0057	<0.0001
discPers	0.8649	0.0027	0.5684	0.0051	<0.0001
ecoHope	1.0216	0.0086	0.0444	0.0104	<0.0001
ecoProSo01	1.2563	0.0066	0.1709	0.0089	<0.0001
ecoProSo03	0.9493	0.0078	-0.1172	0.0097	<0.0001
ecoRCA02	-0.0260	0.0101	-0.6319	0.0108	<0.0001
ecoRCA03	0.9830	0.0089	0.4218	0.0107	<0.0001
edu	4.6020	0.0076	4.1255	0.0084	<0.0001
employstatus123_merge	2.2798	0.0054	2.1966	0.0069	0.4225
fail01	-0.3570	0.0060	0.3203	0.0065	<0.0001
happy	7.2231	0.0086	5.2069	0.0116	<0.0001
houseLeaveAmount	2.3812	0.0056	2.3827	0.0063	0.3104
isoFriends_inPerson	1.9999	0.0129	2.1097	0.0144	<0.0001
isoFriends_online	4.8771	0.0124	3.8049	0.0153	<0.0001
isoOthPpl_inPerson	1.8718	0.0117	2.0170	0.0135	<0.0001

isoOthPpl_online	3.2232	0.0145	2.3724	0.0152	<0.0001
lifeSat	4.7259	0.0047	3.3915	0.0072	<0.0001
lone01	2.0615	0.0055	2.8683	0.0069	<0.0001
MLQ	1.5605	0.0062	-0.0641	0.0092	<0.0001
neuro01	0.2373	0.0091	0.8427	0.0090	<0.0001
neuro02	-0.3653	0.0091	0.6542	0.0096	<0.0001
neuro03	0.9113	0.0074	-0.3283	0.0090	<0.0001
para01	5.6072	0.0160	6.1007	0.0159	<0.0001
para02	2.4016	0.0136	3.9703	0.0164	<0.0001
para03	2.0587	0.0131	3.6553	0.0166	<0.0001
PFS01	-0.3388	0.0063	0.3844	0.0068	<0.0001
PLRAC19	3.4890	0.0077	3.6534	0.0087	<0.0001
PLRAEco	4.1104	0.0092	4.8089	0.0107	<0.0001
posrefocus01	3.0240	0.0057	3.1369	0.0061	<0.0001
posrefocus02	3.2416	0.0055	3.2445	0.0060	0.3370
posrefocus03	2.9405	0.0056	3.0171	0.0061	<0.0001
probSolving01	4.1086	0.0042	3.3835	0.0059	<0.0001
probSolving02	3.8708	0.0049	3.0867	0.0061	<0.0001
probSolving03	4.0621	0.0043	3.4244	0.0058	<0.0001
tempFocFut	1.6086	0.0062	1.1929	0.0084	<0.0001
tempFocPast	0.5774	0.0090	0.8079	0.0097	<0.0001
tempFocPres	1.6734	0.0053	0.8586	0.0078	<0.0001
tightLoose	6.2865	0.0131	5.6728	0.0148	<0.0001
tightNorms	5.8157	0.0134	5.2538	0.0148	<0.0001
tightTreat	5.8020	0.0138	5.3838	0.0152	<0.0001

Table C.2: Percentage of successes in each cluster for each binomial variable, together with the corrected p-value calculated using χ^2 test. The values are rounded to 4 decimal places.

	Percentage: 1	**Percentage: 2**	**p-value**
coronaClose_1	1.0588	1.7056	<0.001
coronaClose_2	2.9550	3.1654	<0.001
coronaClose_3	3.8237	3.8556	<0.0001
coronaClose_4	10.7393	13.0067	<0.0001
coronaClose_5	10.1612	12.1068	<0.0001
coronaClose_6	75.3921	72.5928	<0.0001
employstatus_10	2.5463	1.9224	<0.0001
employstatus_4	5.1749	13.2796	<0.0001
employstatus_5	3.8804	6.5440	<0.0001
employstatus_6	7.3862	8.0581	<0.0001
employstatus_7	10.8550	6.7825	<0.0001
employstatus_8	0.9339	2.8944	<0.0001
employstatus_9	18.2667	22.6386	0.1314
gender (female)	60.5695	61.8372	<0.001
houseLeaveWhy_1	21.6192	20.1164	<0.0001
houseLeaveWhy_2	40.0903	41.1144	<0.0001
houseLeaveWhy_4	4.2722	8.3291	<0.0001

houseLeaveWhy_6	23.0953	20.1308	<0.0001
houseLeaveWhy_7	20.1204	16.9365	<0.0001
relYesNo	53.3524	43.8281	<0.0001

D Appendix: statistics variables Hierarchical Agglomerative clustering with affect variables included

Table D.1: Mean values in each cluster for each ordinal variable, together with the standard error of the mean and the corrected p-value calculated using Mann-Whitney U test. The values are rounded to 4 decimal places.

	Mean: 1	SEM: 1	Mean: 2	SEM: 2	p-value
affAnx	2.4745	0.0070	2.9437	0.0069	<0.0001
affBor	2.4507	0.0074	2.9531	0.0072	<0.0001
affCalm	3.1393	0.0062	2.7416	0.0060	<0.0001
affDepr	1.9247	0.0062	2.5038	0.0068	<0.0001
affEnerg	2.7794	0.0064	2.4035	0.0058	<0.0001
affExh	2.2549	0.0068	2.7016	0.0068	<0.0001
affInsp	2.5989	0.0068	2.2742	0.0061	<0.0001
affNerv	2.3232	0.0067	2.8108	0.0067	<0.0001
affRel	2.9276	0.0064	2.5757	0.0061	<0.0001
age	3.3742	0.0096	2.5006	0.0078	<0.0001
bor02	-0.1134	0.0108	0.1889	0.0102	<0.0001
c19Eff	1.1650	0.0089	0.6381	0.0091	<0.0001
c19Hope	1.4538	0.0081	1.0765	0.0087	<0.0001
C19Know	3.8313	0.0047	3.6440	0.0047	<0.0001
c19perBeh01	2.4963	0.0050	2.1735	0.0069	<0.0001
c19perBeh02	2.5993	0.0044	2.3003	0.0064	<0.0001
c19perBeh03	1.9501	0.0082	1.7688	0.0084	<0.0001
c19ProSo01	1.1864	0.0081	0.7726	0.0083	<0.0001
c19ProSo03	0.7285	0.0095	0.3739	0.0092	<0.0001
c19RCA01	1.4028	0.0105	1.1720	0.0102	<0.0001
c19RCA02	2.1741	0.0071	1.9741	0.0077	<0.0001
consp01	6.7905	0.0158	6.9283	0.0144	<0.0001
consp02	7.1379	0.0151	7.2492	0.0139	<0.0001
disc01	0.5358	0.0058	0.7317	0.0053	<0.0001
disc02	0.7767	0.0057	0.8903	0.0052	<0.0001
disc03	-0.2445	0.0060	-0.5389	0.0054	<0.0001
discPers	0.8237	0.0034	0.6888	0.0042	<0.0001
ecoHope	0.7795	0.0099	0.4288	0.0096	<0.0001
ecoProSo01	1.0015	0.0080	0.5855	0.0082	<0.0001
ecoProSo03	0.6889	0.0092	0.3002	0.0089	<0.0001
ecoRCA02	-0.1613	0.0112	-0.4057	0.0101	<0.0001
ecoRCA03	0.8791	0.0101	0.6126	0.0096	<0.0001
edu	4.6101	0.0083	4.2074	0.0077	<0.0001
employstatus123_merge	1.6698	0.0054	0.9539	0.0069	<0.0001
fail01	-0.2211	0.0067	0.0819	0.0062	<0.0001
happy	6.8912	0.0107	5.8590	0.0113	<0.0001
houseLeaveAmount	2.4598	0.0063	2.3141	0.0056	<0.0001
isoFriends_inPerson	1.9409	0.0140	2.1305	0.0133	<0.0001
isoFriends_online	4.6437	0.0141	4.2264	0.0138	<0.0001

isoOthPpl_inPerson	2.0201	0.0133	1.8455	0.0118	<0.0001
isoOthPpl_online	3.1004	0.0158	2.5962	0.0143	<0.0001
lifeSat	4.4975	0.0061	3.8272	0.0070	<0.0001
lone01	2.1396	0.0063	2.6568	0.0063	<0.0001
MLQ	1.2470	0.0080	0.4974	0.0089	<0.0001
neuro01	0.2876	0.0098	0.6921	0.0087	<0.0001
neuro02	-0.3019	0.0100	0.4175	0.0092	<0.0001
neuro03	0.7672	0.0085	0.0184	0.0086	<0.0001
para01	5.6433	0.0171	5.9820	0.0152	<0.0001
para02	2.7054	0.0155	3.4273	0.0151	<0.0001
para03	2.3678	0.0151	3.1003	0.0151	<0.0001
PFS01	-0.2853	0.0069	0.2092	0.0065	<0.0001
PLRAC19	3.5512	0.0083	3.5713	0.0080	0.1100
PLRAEco	4.1555	0.0101	4.6444	0.0099	<0.0001
posrefocus01	3.0566	0.0061	3.0884	0.0057	<0.0001
posrefocus02	3.2415	0.0060	3.2438	0.0055	0.3870
posrefocus03	2.9677	0.0061	2.9799	0.0056	0.1132
probSolving01	3.9581	0.0051	3.6433	0.0054	<0.0001
probSolving02	3.7040	0.0058	3.3715	0.0057	<0.0001
probSolving03	3.9195	0.0051	3.6619	0.0053	<0.0001
tempFocFut	1.4763	0.0071	1.3825	0.0074	<0.0001
tempFocPast	0.5412	0.0098	0.7986	0.0089	<0.0001
tempFocPres	1.5201	0.0063	1.1372	0.0070	<0.0001
tightLoose	6.1349	0.0145	5.9133	0.0135	<0.0001
tightNorms	5.7092	0.0147	5.4497	0.0136	<0.0001
tightTreat	5.7235	0.0151	5.5295	0.0139	<0.0001

Table D.2: Percentage of successes in each cluster for each binomial variable, together with the corrected p-value calculated using χ^2 test. The values are rounded to 4 decimal places.

	Percentage: 1	**Percentage: 2**	**p-value**
coronaClose_1	0.0000	2.5108	<0.0001
coronaClose_2	5.4534	0.9568	<0.0001
coronaClose_3	5.9628	1.9908	<0.0001
coronaClose_4	16.1379	8.4210	<0.0001
coronaClose_5	16.4251	6.7540	<0.0001
coronaClose_6	62.8419	83.3690	<0.0001
employstatus_10	0.0239	4.2254	<0.0001
employstatus_4	1.7335	14.8303	<0.0001
employstatus_5	0.6804	8.8518	<0.0001
employstatus_6	1.6035	12.9643	<0.0001
employstatus_7	17.2354	1.9611	<0.0001
employstatus_8	0.0034	3.3547	<0.0001
employstatus_9	10.4759	28.6326	<0.0001
gender (female)	57.0979	65.4484	<0.0001
houseLeaveWhy_1	26.3095	16.3071	<0.0001
houseLeaveWhy_2	41.4558	39.7456	<0.0001

houseLeaveWhy_4	1.0291	10.4267	<0.0001
houseLeaveWhy_6	21.3211	22.1994	<0.0001
houseLeaveWhy_7	19.4783	18.0603	<0.001
relYesNo	50.9813	47.5813	<0.0001