Statistical Learning: Home Exercise 3

Solutions by Saumya Gupta

h20saugu@du.se

May 6, 2021

Introduction

This report aims to describe the author's solutions to the third home exercise of the Statistical Learning course at Dalarna University during the second period of the Masters in the Data Science programme. The report will explain the solutions to two problem statements as part of the exercise. The first problem includes a mathematical derivation and testing certain expectations related to linear regression analysis. This report will explain the provided dataset, data modifications done, methods and models used, establish the results, and discuss the findings for the second problem. The author will also list the limitations of the approach.

Problem 1

Solution for part 1: X is a feature that follows a normal distribution with different mean and variance depending on which class it comes from, that is, $X_k \sim N(\mu_k, \sigma_k^2)$ where $k = 1, 2, 3, \dots, K$ denotes the classes. Say, Y represents these classes for X. For the Bayesian classifier, the Bayes' rule is the equation (1). Here, $P(Y = k|X = x)$ represents the probability that Y is class k, given that the data point in consideration is x. $f_k(X = x|Y = k)$ represents the density function or the distribution of X within the class k evaluated at x. It is assumed to be a normal distribution. $p(Y = k)$ represents the marginal probability of class k. In the denominator, we have the marginal probability of x, given by the sum of the entity calculated in the numerator for all classes.

$$P(Y = k|X = x) = \frac{f_k(X = x|Y = k)\, p(Y = k)}{\sum_k^K f_k(X = x|Y = k)\, p(Y = k)} \quad (1)$$

Here, classifier classifies $X = x$ to class k, if $P(Y = k|X = x) > P(Y = m|X = x)\ \forall\, m \neq k$. Let us represent $P(Y = k|X = x)$ as L_k (L for likelihood), $p(Y = k)$ as π_k for simplicity. For our case, we can then modify equation (1) into equation (2). Here, $\frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2}$ represents the function for normal distribution.

$$L_k = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2}}{\sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2}} \quad (2)$$

Here, the classifier will classify an observation i to class k if $L_{k,i} \geq L_{m,i}, \forall\, m \neq k; k = 1, \dots, K, m = 1, \dots, K$. There are two things to look at now:

1. In the operands of the inequality, the denominators will be the same for all likelihood calculations and will not add any information to the question - "likelihood of belonging to which class is the greatest?". Hence only the numerator must be compared for all classes.
2. Using a log transformation (log() being a monotonic function); $\log(L_{k,i}) \geq \log(L_{m,i})$ can simplify our computations for comparison.

Keeping these in mind, it boils down to using the function (1) for likelihood comparisons for all classes. The function is further simplified to function (2) using the product, quotient, and power rule for log(), and with the logarithm of 1 and logarithm of base rule.

$$\log\left(\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}\right) \qquad (1)$$

$$= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}\right) + \log\left(e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}\right)$$

$$= \log(\pi_k) + \log(1) - \log\left(\sqrt{2\pi\sigma_k^2}\right) - \frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2 \log(e)$$

$$= \log(\pi_k) + (0) - \log(\sqrt{2\pi}) - \log(\sigma_k) - \left[\frac{x^2 + \mu_k^2 - 2x\mu_k}{2\sigma_k^2}\right](1)$$

$$= \log(\pi_k) - \log(\sqrt{2\pi}) - \log(\sigma_k) - \frac{x^2}{2\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma_k^2}$$

$$= \log(\pi_k) - \log(\sigma_k) - \frac{x^2}{2\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma_k^2} - \log(\sqrt{2\pi})$$

Again, $-\log(\sqrt{2\pi})$ here, is a constant and will not add to the comparison information.

$$= \left[\frac{-1}{2\sigma_k^2}\right]x^2 + \left[\frac{\mu_k}{\sigma_k^2}\right]x + \left[\log\left(\frac{\pi_k}{\sigma_k}\right) - \frac{\mu_k^2}{2\sigma_k^2}\right] \qquad (2)$$

Hence, we proved that the Bayes' classifier for our case is quadratic, as seen from function (2). Therefore, we assign the class with the highest value of this function (log-likelihood) to x.

Solution for part 2: The data for the three variables, x_1, x_2, y have been simulated using the following three lines of R code.

```
x1 = runif(100)
x2 = 0.5 * x1 + rnorm(100) / 10
y = 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

We wish to model the following three linear relationships using the variables created:

a) $y = \alpha_1 + \beta_1 \, x1 + \epsilon_1$
b) $y = \alpha_2 + \beta_2 \, x2 + \epsilon_2$
c) $y = \alpha_3 + \beta_3 \, x1 + \gamma_3 x2 + \epsilon_3$

Now, from the code, we know that:

$$x_1 \sim U(0,1)$$

Furthermore, say, we consider $\epsilon \sim \mathrm{N}(0,1)$ then:

$$x_2 = 0.5x_1 + \frac{\epsilon}{10} \quad (3)$$

$$y = 2 + 2x_1 + 0.3x_2 + \epsilon \quad (4)$$

Equation (4) tells us about the actual relationship between the 3 variables. That is, $\alpha_3 = 2$, $\beta_3 = 2$, $\gamma_3 = 0.3$ and $\epsilon_3 = \epsilon$. From the Equation (3), we can also deduce the actual relationships between y & x_1 and y & x_2 given in Equation (5) and (6). That is, $\alpha_1 = 2$, $\beta_1 = 2.15$, $\epsilon_1 = 1.03\epsilon$, $\alpha_2 = 2$, $\beta_2 = 4.3$, $\epsilon_2 = 0.6\epsilon$.

$$x_1 = 2x_2 - \frac{\epsilon}{5} \quad (\text{from } (3))$$

$$y = 2 + 2.15x_1 + 1.03\epsilon \quad (5)$$

$$y = 2 + 4.3x_2 + 0.6\epsilon \quad (6)$$

If there is nothing wrong with the model, we should expect the model estimates (alphas and betas values) to be somewhere close to those mentioned in the previous paragraph. Nevertheless, let us look at the correlation relationships in Table 1, calculated using the programmatically simulated variables.

Table 1. Correlations between x_1, x_2 and y

	x1	x2	y
x1	1.00	0.84	0.45
X2	-	1.00	0.42
y	-	-	1.00

There is a high correlation between x_1 and x_2. Using these variables together in multivariate regression analysis could lead to high variance in the estimated parameters and phenomenon where other predictors can linearly predict one predictor with considerable accuracy (multicollinearity). It means the estimates are too sensitive to a minor change in model to be considered reliable.

The correlation of these variables with y is moderate (well, < 0.5). Still, if we are to model the relationships for y & x_1 or y & x_2 separately, we could expect estimates close to the real ones. Hence, the following are the expectation for the mentioned linear models:

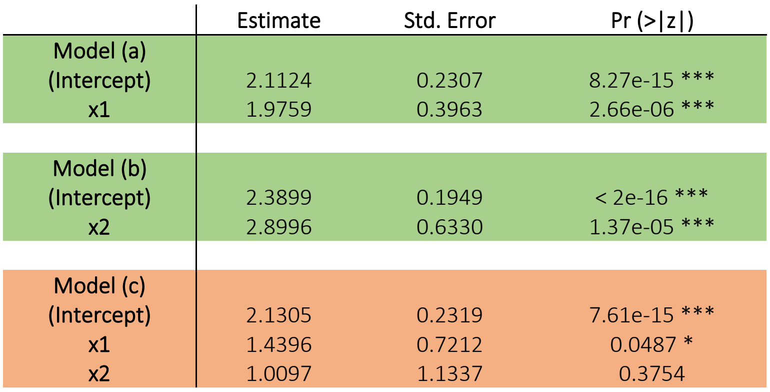
i. We expect α_1 and α_2 to be very close to 2 because in a) and b) the variables are being used separately to model y. We would expect α_1 to be nearer to 2 compared to α_2, because of higher correlation of x_1 to y. We will not expect the same with α_3 because of the non-reliability of the parameter estimates (expected to have high standard errors).

ii. For similar reasons mentioned in part (i), we will expect β_1 to be close to its actual values, which is 2.15. So yes, it could be close to 2. But we will not expect β_3 to be close to its actual value for the reasons mentioned in (i).

iii. We will not expect any of β_2 and γ_3 to be close to 0.3. For the same reason, β_2 is expected to be close to 4.3. γ_3 whose actual value is 0.3 is expected to show estimates with high standard errors, hence unreliable.

After running the regression model programmatically, we get the results recorded in Table 2. We found that contrary to our expectations, α_3 is also close to 2 with high significance and low standard error. Just as we expected, β_3 and γ_3 estimates have high standard errors and are nowhere close to 2 and 0.3, respectively. Both models (a) and (b) show significant relationships with y. But contradictory to what we thought, β_1 is close to 2.15, β_2 is not close to 4.3.

Table 2. Summary of Linear Models in (a), (b) and (c)



	Estimate	Std. Error	Pr (>|z|)
Model (a)			
(Intercept)	2.1124	0.2307	8.27e-15 ***
x1	1.9759	0.3963	2.66e-06 ***
Model (b)			
(Intercept)	2.3899	0.1949	< 2e-16 ***
x2	2.8996	0.6330	1.37e-05 ***
Model (c)			
(Intercept)	2.1305	0.2319	7.61e-15 ***
x1	1.4396	0.7212	0.0487 *
x2	1.0097	1.1337	0.3754

Problem 2

Background: An interactive online personality test was conducted (2016-2018), based on the "Big-Five Factor Markers" from the International Personality Item Pool (IPIP). "Big-Five Factor Markers" represent the five main personality traits suggested for overall grouping for several personality traits - extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience. In the test, they recorded the answers of the test participants for research use with approval. Results of statistical analyses on such data could reveal plenty of insights on applied psychology. Problem statement 2 provides the same dataset for the clustering analysis that we will perform to group the participants based on their responses to the test. The data set is vast, so we need to do a sampling of an appropriate subset and build clusters on that, validate by assigning cluster labels and visualising results. Lastly, we should use the already built clusters and the assigned labels to make predictions for unseen observations.

Data and Related Methods

Dataset: As mentioned previously, the data comes from an online test where over a million (exactly 1013558) people gave numerical responses, based on their agreement with each of the 50 statements, on a scale of 1-5 (1:Disagree, 3:Neutral, 5:Agree). The 50 statements contain 10 sample statements each for the five personality traits labelled as EXT1, EXT2…, EXT10 for extraversion, EST1, EST2…, EST10 for neuroticism, AGR1, AGR2…, AGR10 for agreeableness, CSN1, CSN2…, CSN10 for conscientiousness and OPN1, OPN2…, OPN10 for openness to experience. One interesting observation regarding the statements used in the test is that questions are present in both positive and negative/ reversed forms. For example, for openness to experience, we have both "I have a vivid imagination." (positive) and "I do not have a good imagination." (reversed). Therefore, while describing the personality using the responses, we cannot treat each statement equally for labelling.

Dealing Missing Data: The dataset consists of missing values in 1780 rows. Investigation shows that all the data (the answers to all 50 questions) for these 1780 participants are missing. Therefore, we remove these rows altogether, given that we still are left with sufficient data.

Variables' Usage: The variables in the dataset containing numerical input from 1-5 for agreement are ordinal categorical. To preserve the information in order, we treat these variables as numeric. Grace-Martin (2018) states that doing this requires the assumption that the numerical distance between each set of following categories is equal, which, if true, the analyses based on these numbers will render results that are very close to reality. The assumption holds for our case. Hence, we move forward with 50 numeric variables. Since these have similar ranges, we do not do standardisation.

Sampling Appropriate Subset: As stated previously, the dataset is extensive. We often refer to the whole data as "population". Clustering a large dataset is a challenging problem that requires advanced solutions, such as MapReduce. However, for this analysis, we will build clusters on a sample of this population, representing the whole population. To do that, we calculate the mean vector (centroid) and variance-covariance matrix (dispersion) first to understand the distribution of this multivariate data at hand. These two statistics for analysing multivariate data are analogous to the mean and the variance statistics for univariate data, respectively.

We then use the dmvnorm function from mvtnorm (Genz et al., 2009) to provide the density function for the multivariate normal distribution with a mean equal to the calculated centroid and covariance matrix equal to the calculated dispersion. To be precise, it does not return the density function; instead, it evaluates the density heights at the values we pass to it, using the statistics. We pass the whole dataset to the dmvnorm function and get the density heights for each point. In one dimension, the density height for an observation is analogous to the probability of the observation in the dataset. Notice, here we assume the data distribution to be multivariate normal and proceed further with this assumption. We found that checking it through tests (say, energy test) requires much memory.

Once we get the probabilities for each of these observations, we pass these probabilities to the sample function of base (R Core Team, R Foundation for Statistical Computing, 2020). Given the number of items to be chosen, this function will randomly pick observations from the population according to the probabilities mentioned. Hence it will somewhat replicate the distribution in the sample result. We now proceed with this method of random, unbiased sample picking.

Sample Size: Bullen (2021), in her article about sample sizes for survey research, states that a good maximum sample size with a margin of error of ∓ 3 % is usually around 10% of the population if this does not exceed 1000. Siddiqui (2013) concludes in his study that there is no rule of thumb for the minimum sample size for cluster analysis. However, we start by checking how the total with-cluster sum of squares for the K-Means clustering with the number of clusters (k) equals four changes for different sample sizes. Fig 1 shows that the metric increases with sample size. The selection of 4 for k-value is motivated in a later section. At one point, the expectations were that increase in the sample size would not have much impact on the metric, which we cannot observe in the figure. Since we do not want to increase our sample size beyond 2000, we go with 1000 as the sample size and proceed towards building clusters.

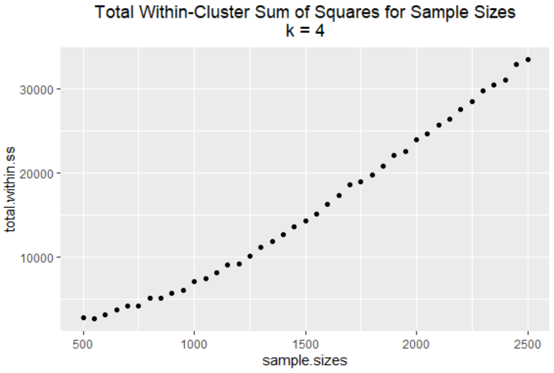


Figure 1. The total within-cluster sum of squares for four clusters build using K-Means clustering on different sample sizes. Here, we extract all samples using the densities calculated with population centroid and dispersion.

Clustering: For K-Means clustering, we need to specify the k-value, which we will also provide while hierarchical clustering for cutting the dendrogram. We find this using the elbow method and the K-Means algorithm. In Figure 2, we see the total within-cluster sum of squares for clustering with different k-values for two different sample sizes. We notice that irrespective of the sample size, we see the elbow at k = 4. Hence, we choose four as the number of clusters for both K-Means and the hierarchical clustering approach. We now know the reason behind the use of k = 4 in Fig 1.

While performing K-Means, we use 50 random sets for centroids (nstart) to end up with stable clusters.

For hierarchical clustering, we explore different agglomeration methods: complete, average, and single linkage.

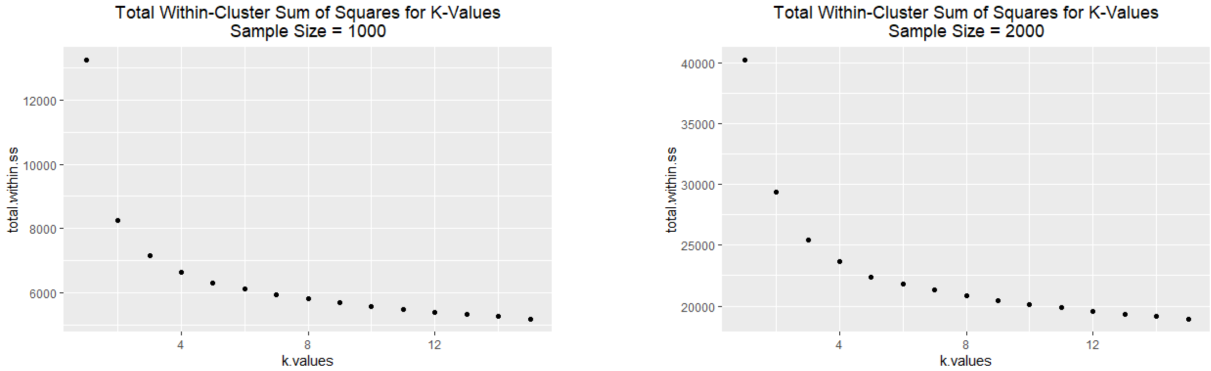


Figure 2. The total within-cluster sum of squares for different k-values on the sample size of (left) 1000 and (right) 2000. Here, we extract all samples using the densities calculated with population centroid and dispersion.

Clustering Results

Table 3. shows cluster sizes for clusters produced from K-Means and Hierarchical clustering with complete, average, and single linkage methods. We see clusters with single observations for average and single linkage methods, which is not likely for our case, and we should see more than one observations per cluster. On the other hand, complete linkage method clusters show some similarity to K-Means clusters in terms of sizes. Therefore, from this point on, we exclude average and single linkage clusters from our analysis.

Table 4 shows the agreement between K-Means clustering and complete linkage hierarchical clustering. We can say that 98.9% classified in cluster 1 by hierarchical was also classified in cluster 1 by K-Means. Similarly, for cluster 2, 3 and 4, the percentages are 70.8%, 86.4% and 14.3%. There is less agreement for cluster 4. 116 observations in the sample assigned to cluster 4 by hierarchical were assigned cluster 1 by K-Means. Hence, there is disagreement here.

Table 3. Cluster Sizes of the 4 Clusters Produced from K-Means and Hierarchical Clustering (Complete, Average and Single Linkage)

	1	2	3	4
K-Means	637	113	103	147
Hierarchical (Complete)	766	94	119	21
Hierarchical (Average)	994	1	3	2
Hierarchical (Single)	997	1	1	1

Table 4. Agreement Between K-Means and Hierarchical Clustering with Complete Linkage

		K-Means Clustering			
		1	2	3	4
Hierarchical Clustering	1	630	7	0	0
with	2	13	80	20	0
Complete Linkage	3	7	7	89	0
Method	4	116	0	10	21

To visualise the clusters from both algorithms, we use the first two principal components, which explain 52.3 % of the variance in the 50 variables. Fig. 3 shows the clusters from both algorithms. Here, the overlap between the 1st and the 4th cluster across both clustering algorithms is visible.

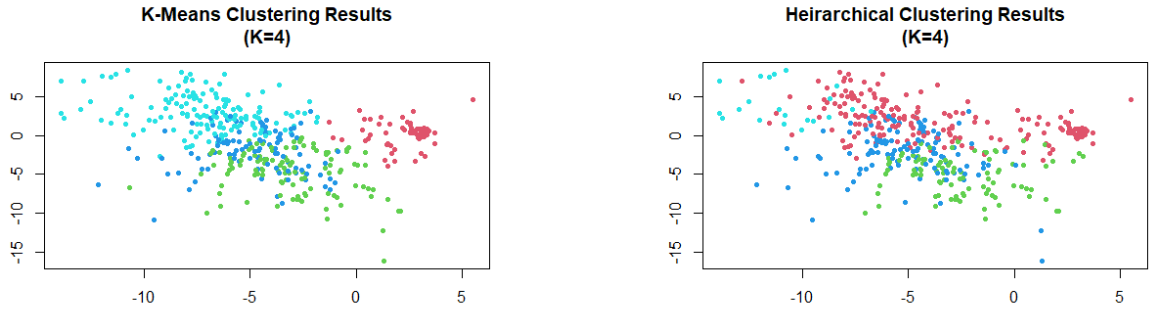


Figure 3. Clusters build using (left) K-Means and (right) hierarchical clustering with the complete linkage method. Again, clusters are in different colours.

Labelling

We proceed towards labelling the clusters build using K-Means clustering because of better cluster separation than that build using Hierarchical clustering. A thing to note here is that Fig 3. only captures ~ 50% of the variance in the variables. Finally, we bind the cluster numbers to the actual data, resulting in 51 variables, 51st denoting the assigned cluster.

Handling Reversed Statements: As stated previously, the test obtains answers for many statements in a reversed form. For labelling, we want to aggregate values within clusters, and for these aggregation results to make sense, we multiply answers to each of these negative statements with -1. To understand this, consider an extrovert person who will rate “I start conversations.” as 5, “I have little to say.” as 1. It makes sense. To conclude about this person, we want to take the sum and say higher the value more the person is an extrovert. If we take both the questions equally, then 10 is the highest answer (5 + 5). For this case, taking a sum gives us 6, which is less than 10, and the person is not a strong extrovert. Now consider if we do not treat the two questions equally, and we multiply the answer to the reversed question with -1 and then take a sum; $5 + (-1)1$, we get 4, which tells us that the person is a strong extrovert because 4 is the highest answer possible in this case.

Hence, we do two things for labelling in the specified order and Table 4. shows the aggregate values obtained as a result:

1. Take column-wise means within each cluster. We get four rows (1 for each cluster) and 50 columns (1 for each statement). Each cell represents the average answer for that statement for that cluster.
2. Next, we take row sums but trait wise. Row mean of all “EXT” statements are taken into one column, “sum_EXT”, all “EST” statements into “sum_EST” and so on. So we get four rows (1

for each cluster) but only five columns (1 for each personality trait). Each cell represents the sum of the average answer for that personality trait for that cluster.

Table 4. Sum of Average Answers for Big Five Personality Traits for Clusters Build Using K-Means Clustering

	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness
1	-0.14	-23.90	6.10	6.06	12.06
2	-5.27	-25.82	14.13	10.41	19.35
3	4.17	-26.11	15.74	7.90	21.23
4	4.51	-18.25	15.63	14.04	21.07

Based on the number of negative and positive statements for each personality trait, we find the highest and lowest scores for each personality trait and decide the labels based on this continuous scale of lowest to highest. Table 5 shows the scale with buckets. The labels are given based on the value (1-5) to which the score for a particular trait for a particular cluster is close.

Table 4. Minimum and Maximum Scores for the Big Five

	1 (Min)	2	3	4	5 (Max)
Extraversion	-20	-10	0	10	20
Neuroticism	-44	-32	-20	-8	4
Agreeableness	-14	-4	6	16	26
Conscientiousness	-14	-4	6	16	26
Openness	-8	2	12	22	32

Lastly, the labels assigned to clusters are as follows:

1. Cluster "**1**": "**Neutral**"
2. Cluster "**2**": "**ReservedCompassionateCurious**"
3. Cluster "**3**": "**StableCompassionateCurious**"
4. Cluster "**4**": "**CompassionateCuriousOrganized**"

Cluster "1" people are ordinary in all personality traits (they get 3 on the scale of 1-5 for all traits). On the other hand, cluster "2" people are a bit reserved for scoring close to the second bucket in extraversion, compassionate for closing to the fourth bucket in agreeableness, and curious for closing to the fourth bucket in openness to experience and same for the other two clusters. Fig. 4 shows the cluster sizes from clustering representing the population.

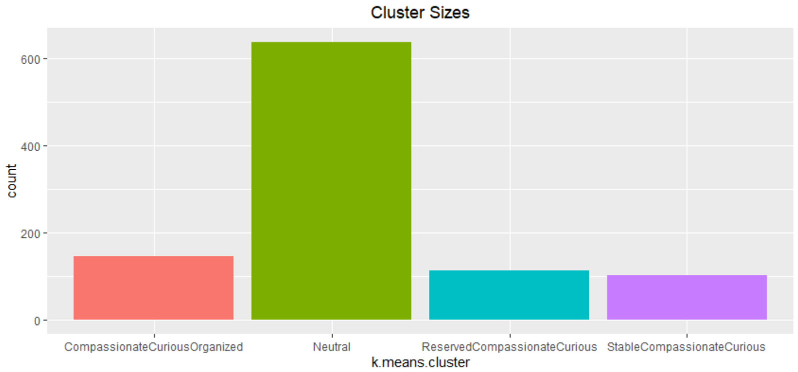


Figure 4. Cluster sizes for clusters build using K-Means clustering. Clusters are labelled.

Prediction

We take another random sample of 100 observations from the population for predictions on a new subset, which has zero intersection with the previous sample used for building the K-Means clusters. For a particular observation in the sample, we assign the cluster label closest to, in other words, whose centroid has the lowest Euclidean distance from it. We do it with all the observations in the sample. Fig. 5 shows the number of assignments for each cluster to the observations in the new sample. We see only 9% neutral personalities and 48% reserved, compassionate, and curious personalities in this subset.

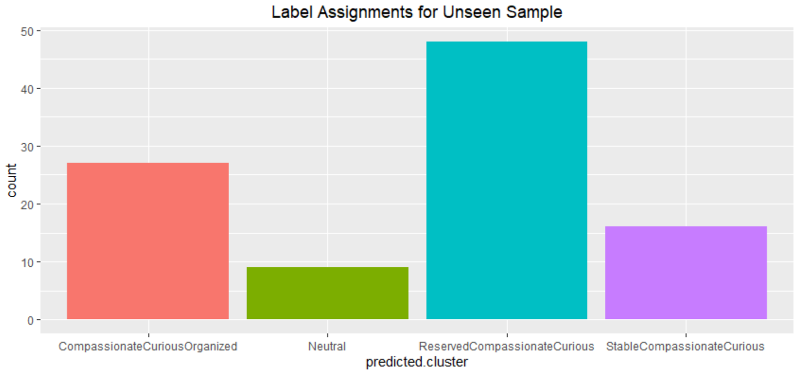


Figure 4. The number of assignments of each cluster to the observations from the unseen subset.

Limitations

The author identifies the following limitations to the analysis:

1. We assume the distribution of population data to be multivariate normal during the sampling task. However, we can do further investigation into more appropriate sampling techniques.
2. In this analysis, we use PCA after clustering to validate the clusters by visualisation. Unfortunately, the first 2 PCs here explain only 52% of the data.
3. PCA could also be done before the clustering to reduce dimensions to make more sense of the observation to feature ratio. However, this may or may not help in revealing clusters.

References

Bullen, P. B. (2021). How to choose a sample size (for the statistically challenged). Retrieved from Practical tools for international development: https://www.tools4dev.org/resources/how-to-choose-a-sample-size/

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2009). Computation of Multivariate Normal and t Probabilities. Heidelberg: Springer-Verlag.

Grace-Martin, K. (2018). Pros and Cons of Treating Ordinal Variables as Nominal or Continuous. Retrieved April 26, 2021, from The Analysis Factor: https://www.theanalysisfactor.com/pros-and-cons-of-treating-ordinal-variables-as-nominal-or-continuous/

R Core Team, R Foundation for Statistical Computing. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Siddiqui, K. (2013). Heuristics for Sample Size Determination in Multivariate Statistical Techniques. World Applied Sciences Journal, 27(2), 285-287. Retrieved from https://ssrn.com/abstract=2447286