Statistical Learning

HOME WORK#3 INFERENCIAL STATISITIC AND CLUSTER ANALYSIS

Julateh K. Mulbah
| DARLARNA UNIVERSITY **DATE: JUNE 12,2020**

1. Introduction

Linear regression is useful for finding a relationship between two variables. One is a predictor or independent variable, and the other is the response or dependent variable.

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable Y, based on the value of an independent variable X.

However, statistical inference is also the theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, normally on the basis of random sampling. The main task in this study is to investigate some inferential properties of the linear regression model.

The methodology pertaining to the questions are as follows:1.

I) x=rnorm(100)

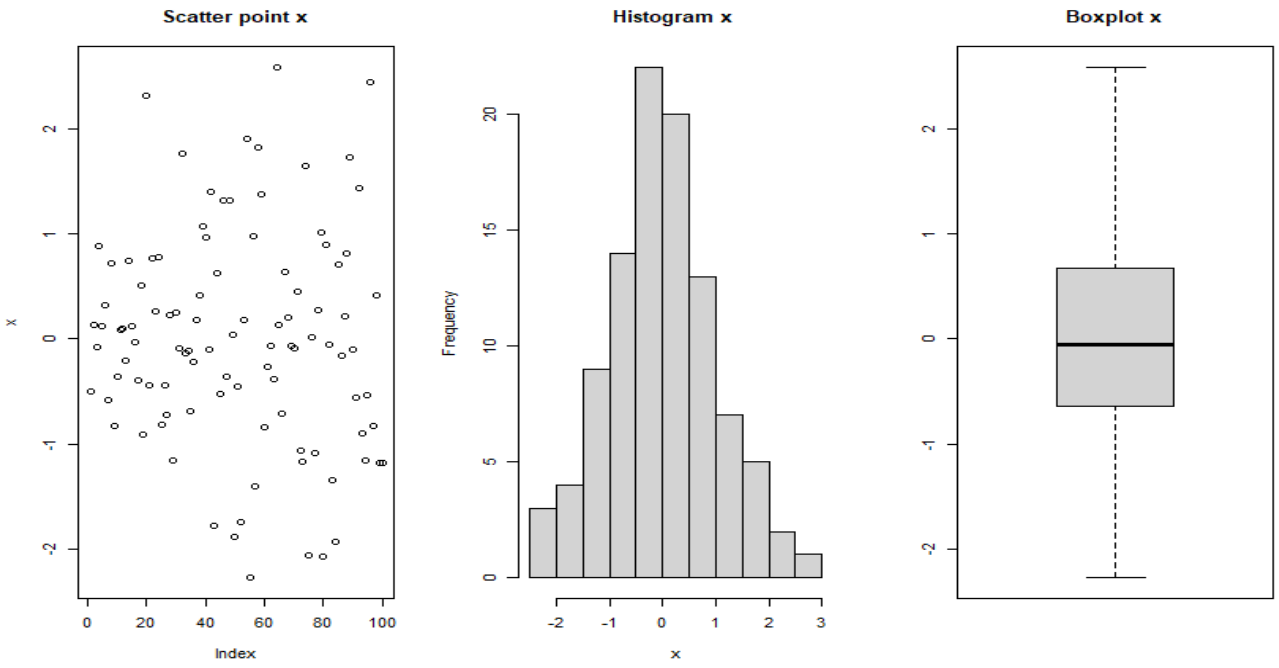


Fig 1. X distribution

X is normally distributed

II) y=-1+2x+rnorm(100)

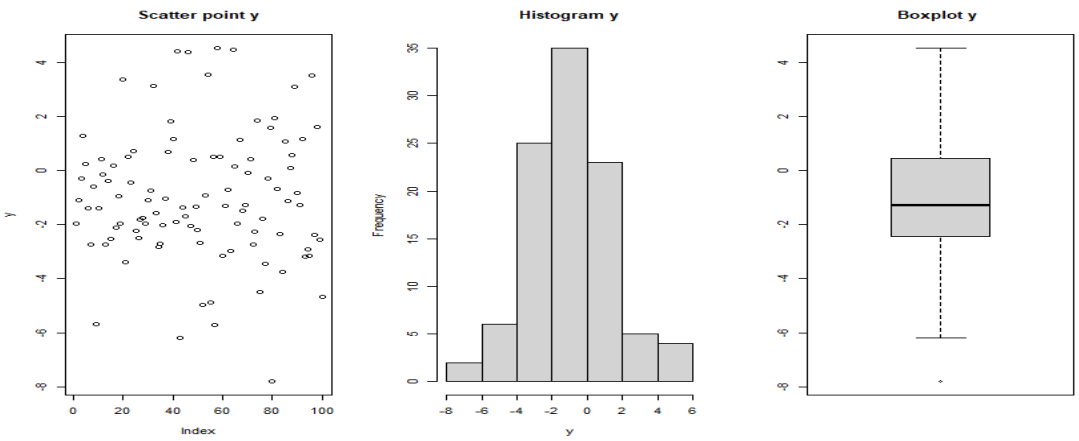


Fig 2 y distribution.

y is normally distributed

III) lm (y~x)

A significant regression equation was found (F (1, 98) = 588.9, p < 2.2e-16), with an R2 0.8573. y predicted is equal to -0.9885+ 1.8946 (x). y increased by 1.8946 for each x. Furthermore, x is statistically significant because the p-value has a significance of 0.05.

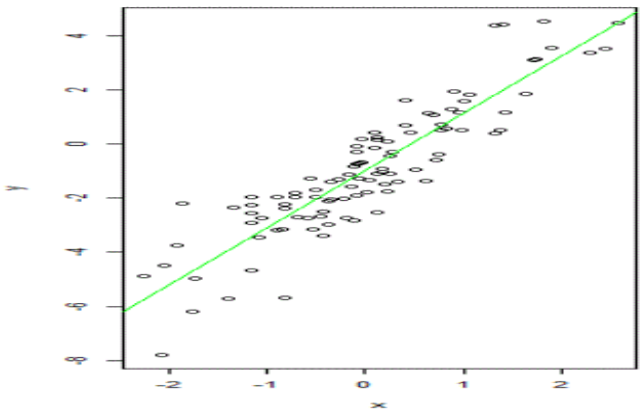


Fig 3. The Linear Regression Line a significant linear relationship.

IV) The variables in deviation form are:u=x-mean(x) and v= y-mean(y) model parameter using lm(v~u) is given as

Linear regression was modelled to predict the value of v based on u. A significant regression equation was found (F (1, 98) = 588.9, p < 2.2e-16), with an R2 of 0.8573. v predicted is equal to 4.207e-17 + 1.895 (u). v increased by 1.895 for each u.

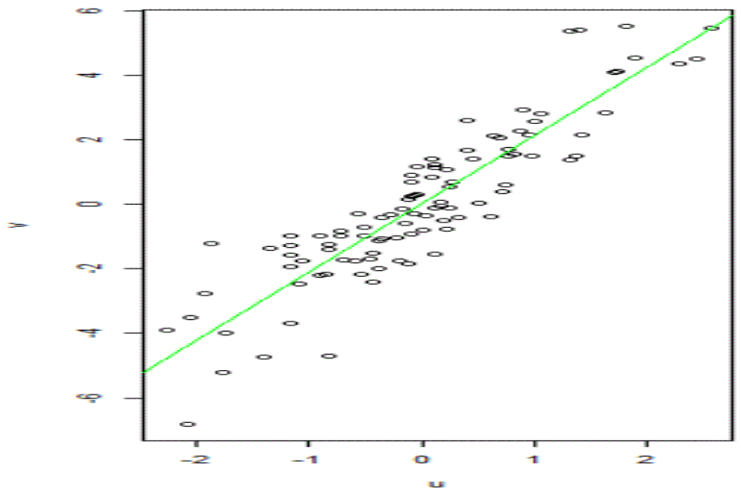


Fig:4. The linear regression equation –lm (v~u)

IV) lm (v~u)

v predicted is equal to 0+ 1.895 (x). v increased by 1.8946 for each u.

Yes I expect the slope of the two regression models to be same. The reasons are that looking at the above simulated data, it is realized that the data is normally distributed. Moreover, the centered values are also normally distributed. The values of x and y were only transformed using centering technique as new observations for u and v respectively. And so it does not make the values lose their coefficient values or slope parameter. In other words, the unit value does not change but the intercept is expected to change in value any time the observed value is cantered at its mean.

1b) II) Regression of u onto v without intercept

Linear regression was modelled to predict the value of u based on v. A significant regression equation was found (F (1, 98) = 588.9, p < 2.2e-16), with an R2 of 0.8573. v predicted is equal to 0 + 1.895 (u). v increased by 1.895 for each u.

From the regression the t-statistic is 24.39

The parameter is 0.45251

It can be observed that the two regression models' parameters are different but t –statistic is the same.

The proof is given below.

Mathematical proof that the equality of the t-statistic is expected or disprove by showing a counter example.

Null hypothesis: $U_v = U_u$ ---------- - (1)

Alternative hypothesis: $U_v = ! U_u$---------- -(2)

Using the conventional p-value of 0.05.

Using t-statistic= as per attached in the Appendex-Page#13 and page#14

R code for both Null and Alternative hypothesis.

```
> ## v onto u
> n <-length(u)
> t <- sqrt(n - 1)*(u %*% v)/sqrt(sum(u^2) * sum(v^2)-(u %*% v)^2)
> as.numeric(t)
[1] 24.3909


> ## u onto v
> n <-length(v)
> t <- sqrt(n - 1)*(v %*% u)/sqrt(sum(v^2) * sum(u^2)-(v %*% u)^2)
> as.numeric(t)

[1] 24.3909
```

Now we see that the t above is exactly the t-statistic given in the summary of both models.

Hence proved.

Part:2……

Introduction

Today, many researchers believe that they are five core personality traits. The "big five" are broad categories of personality traits. While there is a significant body of literature supporting this five-factor model of personality, researchers don't always agree on each dimension's exact labels. These personality traits are openness, conscientiousness, extraversion, agreeableness, and neuroticism.

However, it was the model to comprehend the relationship between personality and academic behaviors.[2]This model was defined by several independent sets of researchers who used factor analysis of verbal descriptors of human behaviour.[1]

This study intends to use clustering analysis to cluster the numerical response base of research conducted about the five personality traits. The response was based on the agreement with the statement on a scale of 1-5. The rest of the study is as follows in section2 and 3, methodology. The results will be presented in section 4 and the conclusion in section 5.

2. Methodology

2.1 Data Preparation.

The data is from an online survey of more than 1000000 participants test consisting of 50 statements. For every statement, the participants could give a numerical response based on an agreement with the statement. A subset consisting of 100,000 were taken to build the model. Another subset-50,000 were chosen so that the observations will be assigned with new clusters centers. However, data preparation was carried out to give an insight into the nature of the data, which allows for better questions. In other to ensure that the data is complete and insight is accurate, an analysis was made to detect missing values. No missing values were observed. Furthermore, it was observed that each variable has five levels and are numeric in nature of range 1-5. The values were standardized to apply PCA on the dataset. Some of the distributions of the dataset are shown in Figure 1.

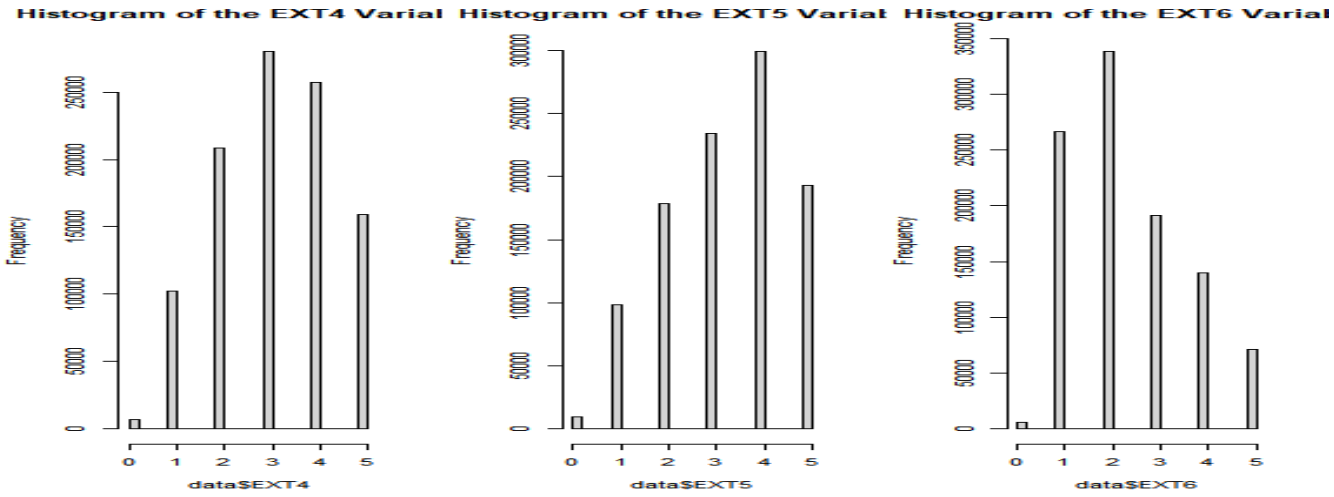


Figure 1. Variable distribution.

2.2 Correlation analysis

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

The correlation heatmap in Fig 3 shows the correlation among variables in the dataset.

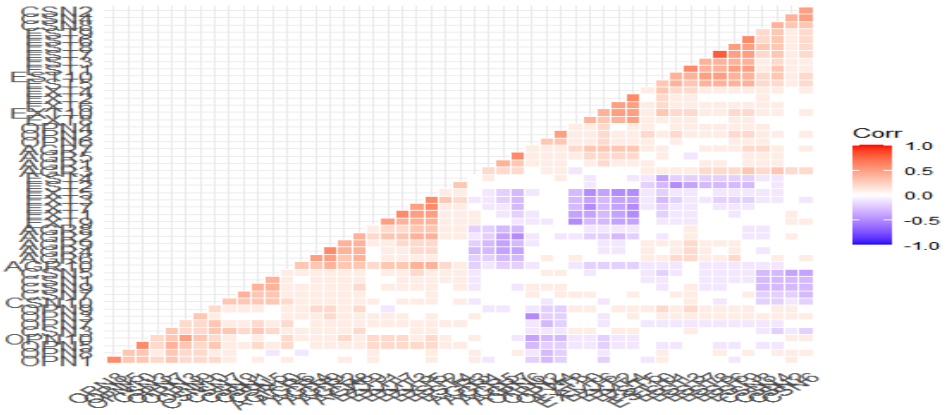


Figure 2. Correlation map

We can see that there is high correlation among variables where the colour is relatively dark.

2.3 Dimension reduction

Dimension reduction technique was used in this analysis to reduce the number of features in the dataset, while creating new combinations of attributes. I employed PCA .it allows to reduce a “complex” data set like the one at hand to a lower dimension in order to reveal the structures or the dominant types of variations in both the observations and the variables. The PCA analysis results show that 37 PCs explains more than 90% of the total variation in the dataset. Precisely 90.32%.

The first two principal components account for 24.31% of the total variation as shown in fig 5

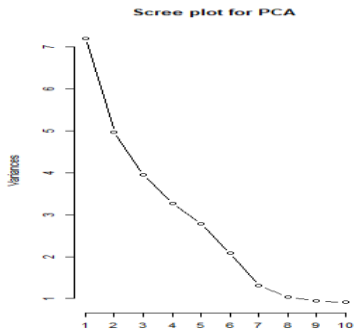


Figure 3. PCA suggesting the number of PCs and each contribution to the total variation.

3. Method

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group. In order words, clustering is the grouping of specific objects based on their characteristics and their similarities.

Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters.

Hierarchical and k-means methods have been used in this analysis. Hierarchical clustering can be agglomerative or Divisive approach. Divisive Approach was used where is also known as the Top-Down Approach. It begins with all the objects in the same cluster

Moreover, an optimization technique was employed to select the best linkage to form the clusters.

3. 1 Hierarchical clustering.

Table 1. Hierarchical Method

Method	Accuracy
Average	0.6862563
Single	0.5518285
Ward.D2	0.9768917
Complete	0.7984202

It is realized from Table 1 that the ward method is the best method to solve the problem at hand. Moreover, since we know beforehand that the clusters are five, the dendrogram was cut horizontally at height 5.

Furthermore, k-means clustering is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. Most often the number of k must be defined or determined before the algorithm can be fitted.

Fig 4 below shows an elbow technique used to determine the number of k for the clustering.

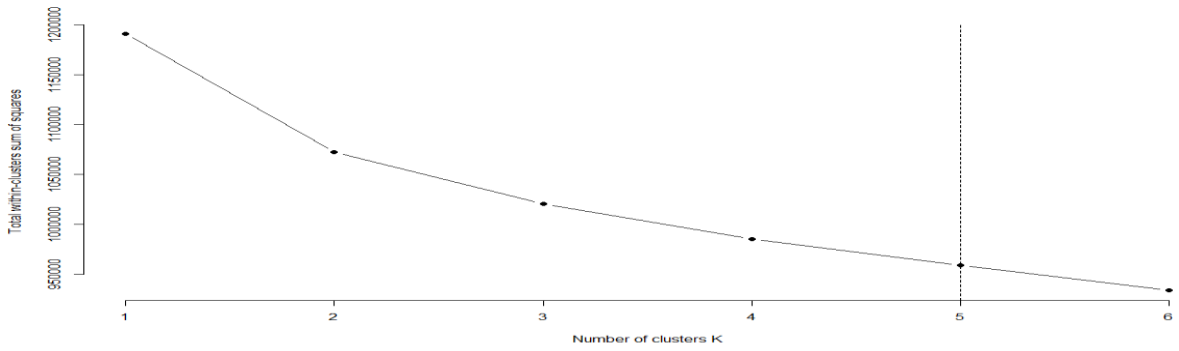


Figure 5: Showing the result of k-means for 'N' = 100,000 and 'K = 5

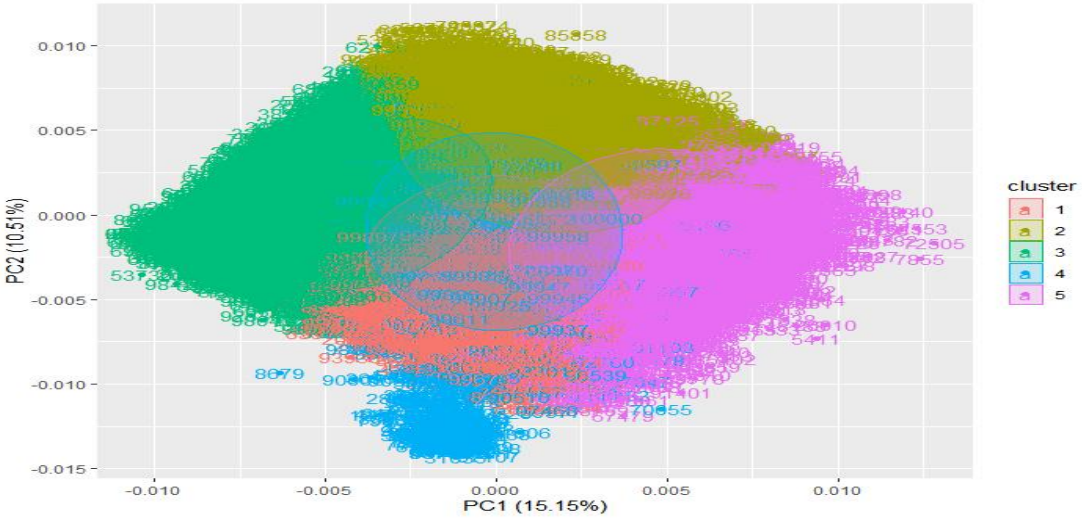
4.2 Determine the distance between clusters

Table 2. Distance between clusters centers

cluster	1	2	3	4
1				
2	3.621179			
3	7.040810	4.374418		
4	5.103402	4.075270	4.229122	
5	4.611820	4.749355	4.942298	3.842672

As you can see, Clusters 1 and 3 are relatively well-separated from each other, while cluster 2, 4, 5 not as much.

4.0 Results



4.1 k-means Clustering

Figure 6. K-means clustering plot

All Clusters are relatively tight.

4.2 Number of observations/sums of square in each cluster for the five cluster solution

Table 3. Clusters features

features	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Case	23314	20582	16089	21418	18597
Within Sum of Square	800951.4	835109.4	733615.8	884790.4	822259.4

It is realized that cluster 4 has more number of observations within than all clusters. Also, cluster 3 has the lowest observations. Moreover, the sum of square within cluster 4 is the largest .This means the records are loose .Cluster 3 has the smallest sum of squares.

4.3 Characteristics of clusters

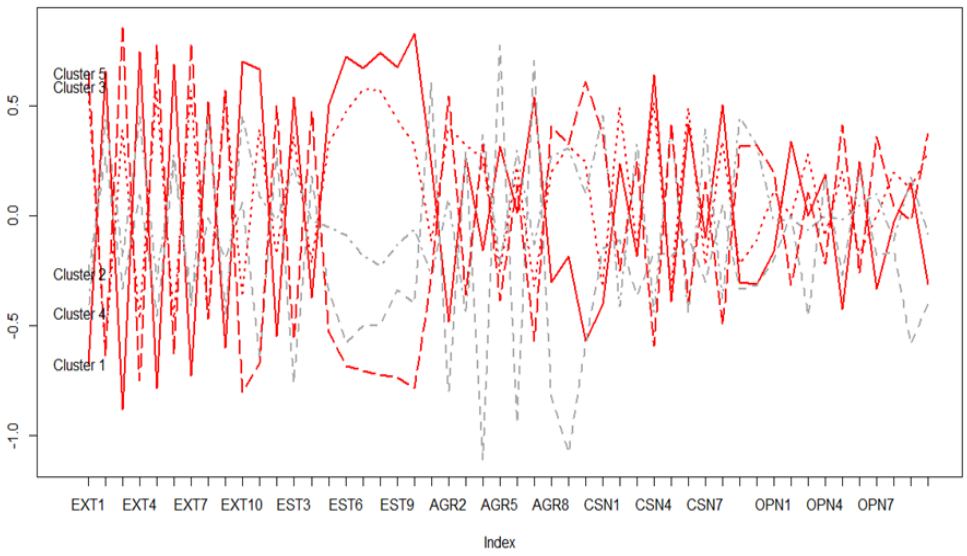


Fig 7. Clusters centroid.

Cluster 1, 2& 5 are characterized with people who feel more comfortable around people but very quiet around strangers. Moreover, cluster 3 are people who most often leave their belonging around to the extreme but are very quiet around strangers. On the other hand, cluster 4 people are characterized with an attitude of not interested in other people problems also low in sympathizing with others' s feelings.

4.4 Testing datasetTable 5. Clusters features

Test set	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Number of case	5733	8937	93712	2833	48392

From Table 5, more cases were assigned to cluster 5 and least being cluster 1.

4. 5 Hierarchical clustering

Choosing the linkage to merge the form the clusters.

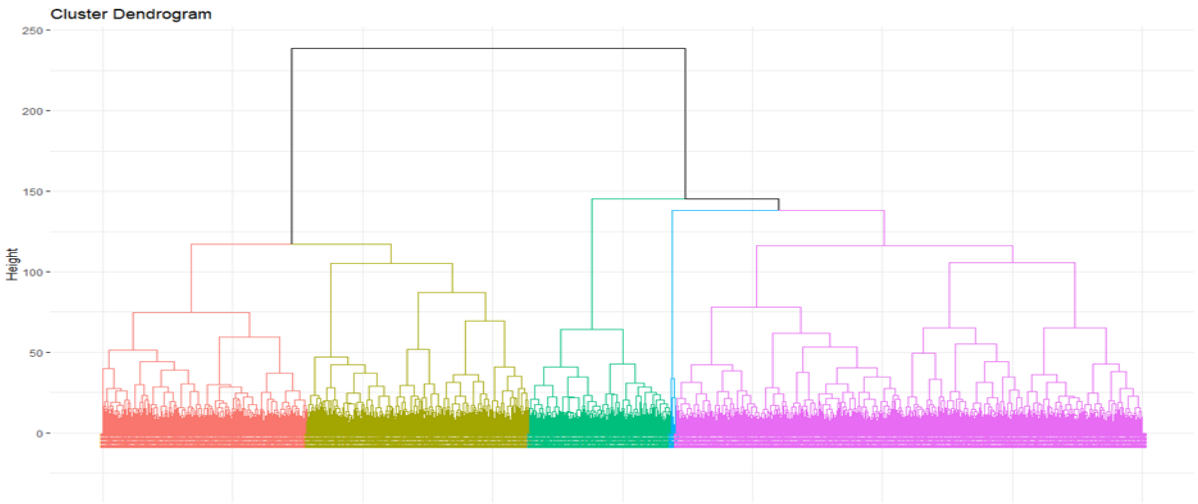


Fig 8. Hierarchical clustering using ward.D2 method.

From Fig 8, all the five clusters are well separated as shown on the dendrogram.

Conclusion.

This study's objective is to use an unsupervised machine learning algorithm to cluster the five personality traits dataset into a group of common characteristics. Both k-means and hierarchical clustering were used. The dataset had observations of over 1000000 numerical responses of a survey concerning the big five personality traits. A subset was used to build both the k-means and hierarchical models. Moreover, the k means model was used to predict new cases.

However, the dimension of the dataset was relatively high and, therefore, a dimension technique. Principal Component Analysis was employed to reduce the dimension without losing many variations in the dataset. The resulting Principal Component Analysis shows that out of the total number of 50 PCs, 37 PCs accounted for 90.32% variation in the dataset. Since it is difficult to plot on more than two-dimension plane figures, only two PCs, which explained 24.31% variation was used to plot the clustering results.

Furthermore, Model optimization techniques were employed to optimize the clustering model to solve the problem at hand. An elbow technique was used in the k-means clustering to ascertain the number of k for the clustering. The results show that k= five, as shown in Fig 5. In the hierarchical clustering, the ward.D2 method scored high accuracy in terms of building up the dendrogram. The result of both clustering algorithms is relatively similar.

The results of the clustering labels have been explicitly shown in Fig 7 in section 4.

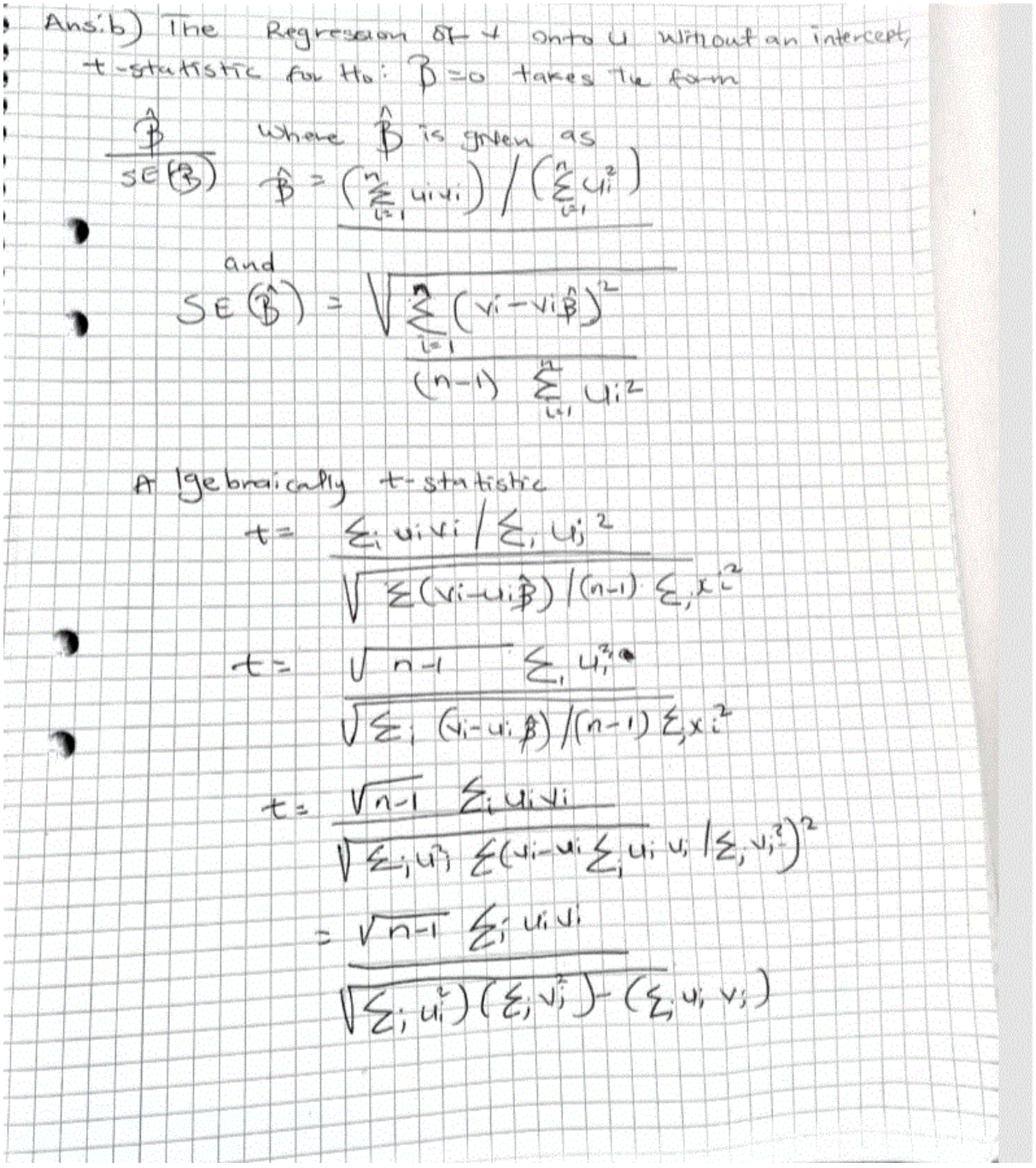
The only challenge in this analysis is that the hierarchical clustering is computationally expensive.

References:

[1] Antoneko.D.P, Toy.S., & Niederhauser, D.S. (2012). Using Cluster analysis for data mining in Educational technology research. Online Publication,.

[2] https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

Appendex:1: The regression of V onto u without an intercept.



Ans:b) The Regression of v onto u without an intercept, t-statistic for $H_0: \beta = 0$ takes the form

$$\frac{\hat{\beta}}{SE(\hat{\beta})}$$

where $\hat{\beta}$ is given as

$$\hat{\beta} = \left(\sum_{i=1}^{n} u_i v_i\right) \Big/ \left(\sum_{i=1}^{n} u_i^2\right)$$

and

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n} (v_i - v_i \hat{\beta})^2}{(n-1)\sum_{i=1}^{n} u_i^2}}$$

Algebraically t-statistic

$$t = \frac{\sum_i u_i v_i / \sum_i u_i^2}{\sqrt{\sum (v_i - u_i \hat{\beta}) / (n-1) \sum_i x_i^2}}$$

$$t = \frac{\sqrt{n-1} \sum_i u_i^2 \bullet}{\sqrt{\sum_i (v_i - u_i \beta) / (n-1) \sum_i x_i^2}}$$

$$t = \frac{\sqrt{n-1} \sum_i u_i v_i}{\sqrt{\sum_i u_i^2 \sum (v_i - u_i \sum_i u_i v_i / \sum_i v_i^2)^2}}$$

$$= \frac{\sqrt{n-1} \sum_i u_i v_i}{\sqrt{\left(\sum_i u_i^2\right)\left(\sum_i v_i^2\right) - \left(\sum_i u_i v_i\right)}}$$

Appendix#2: Regression of U onto V without an intercept.