Contents

[1 – Introduction 2](#_Toc61251887)

[2 – Exploratory Data Analysis (EDA) 2](#_Toc61251888)

[2.1 – NEETs.1 2](#_Toc61251889)

[2.2 – UK.NEETs 3](#_Toc61251890)

[2.3 –Kaggle.NEETs 4](#_Toc61251891)

[3 – Analysis 6](#_Toc61251892)

[3.1 - Linear Regression 6](#_Toc61251893)

[3.2 – Cluster Analysis 8](#_Toc61251894)

[3.3 – Factor Analysis 10](#_Toc61251895)

[4 - Conclusion 12](#_Toc61251896)

[5 - References 13](#_Toc61251897)

[6 - Appendix 14](#_Toc61251898)

[6.1 – NEET.1 Unemployed Distribution 14](#_Toc61251899)

[6.2 – NEET.1 Economically\_Inactive Distribution 14](#_Toc61251900)

[6.3 – NEET.1 As\_Percentage Distribution 15](#_Toc61251901)

[6.4 – NEET.1 Fit Diagnostics for Unemployed 15](#_Toc61251902)

[6.5 – NEET.1 DW Code (does not work) 16](#_Toc61251903)

[6.6 – Kaggle.NEETs – Scatter pt\_neet 16](#_Toc61251904)

[6.7 – Kaggle.NEETs – Scatter t\_economically\_inactive 17](#_Toc61251905)

[6.8 – Kaggle.NEETs – Scatter t\_population 17](#_Toc61251906)

[6.7 – Code snippet for Linear Regression 18](#_Toc61251907)

[6.8 – Code snippet for Cluster Analysis 18](#_Toc61251908)

# 1 – Introduction

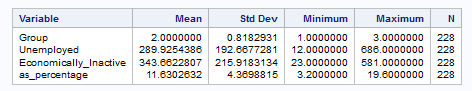
This report will look at 3 datasets around the subject of unemployment and specifically on people not in education or training (NEETs). These 3 datasets will be used to explore if there are any known patterns and any unknown patterns, with the aim to carry out a linear regression analysis along with a Cluster Analysis and a Factor Analysis.

# 2 – Exploratory Data Analysis (EDA)

Before carrying out any form of analysis on the datasets gathered, it is a good idea to explore the data to try and understand the datasets in more depth. The 3 datasets here are all around people not in education, employment or training (NEETs), this section will use exploratory data analysis on the 3 datasets to see if there is anything that stands out before carrying on with the analysis.

## 2.1 – NEETs.1

The dataset named NEETs.1 was a dataset gathered from the (Office for National Statistics, 2020) with a dataset named ‘*Young people not in education, employment or training (NEETs)’*, which looks at the number of NEETs from the UK within the years of 2001 to 2020. This dataset had a total of 228 observations with multiple columns, which can be seen in figure 1.



(Figure 1 – NEETs.1 Summary Analysis.)

After carrying out a summary analysis it is a good idea to run a distribution analysis to see where the data lies within the dataset, whilst checking normality which will be useful later. This can be done to test a hypothesis of:

|  |
| --- |
| H0: The data is normally distributed |
| H1: The data is not normally distributed |

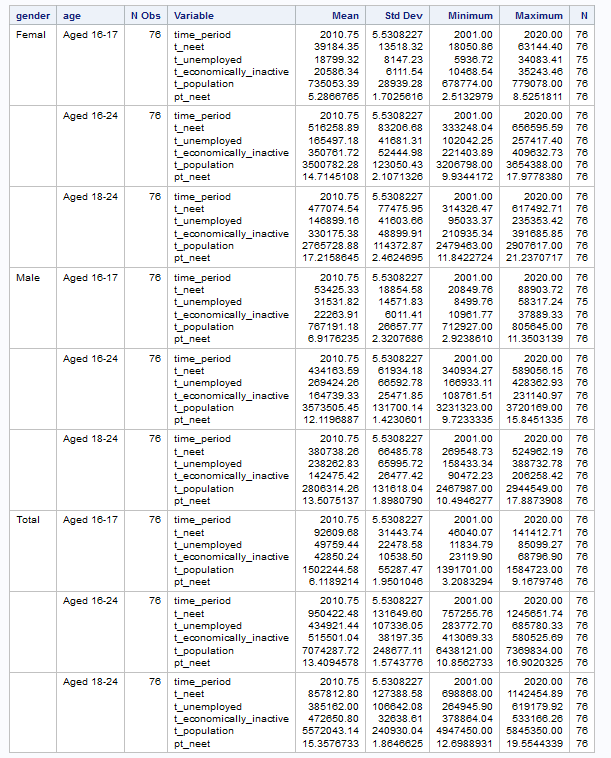
Below are the results for the test of normality for the different types of variables (excluding group as it is categorical).

|  |  |
| --- | --- |
| **Variable** | **P Value** |
| Unemployed | <0.010 |
| Economically\_inactive | <0.010 |
| As\_Percentage | <0.010 |

Seeing as the results all have a p-value of <0.010 we can state that we would reject the null hypothesis at a 5% level (even 1%) which means the data is not normally distributed, this would indicate there would need to be nonparametric tests going forward. With the summary analysis and the test for normality being done, the next step for this dataset would be the analysis.

## 2.2 – UK.NEETs

The dataset named UK.NEETs was a dataset gathered from (GOV.UK, 2020) which was named *‘Not in Education, Employment or Training (NEET) by gender’* which looked at the “percentage of 16-24-year olds not in education, employment or training” (GOV.UK, 2020). This dataset had a total of 685 observations with multiple columns, which can be seen in figure 2.



(Figure 2 – UK.NEETs Summary Analysis)

After carrying out a summary analysis, it could be a good idea to carry out a distribution analysis to see where the data lies within the whole dataset whilst checking for normality which will be useful later. This can be done to test a hypothesis of:

|  |
| --- |
| H0: The data is normally distributed |
| H1: The data is not normally distributed |

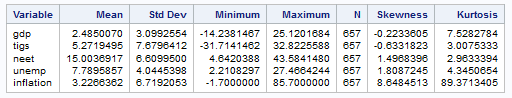
Below are the results for the tests for normality for the different types of variables (excluding group).

|  |  |
| --- | --- |
| **Variable** | **P Value** |
| t\_neet | <0.010 |
| t\_unemployed | <0.010 |
| t\_economically\_inactive | <0.010 |
| t\_population | <0.010 |
| pt\_neet | <0.010 |

Seeing as the results all have a p-value of <0.010 we can state that we would reject the null hypothesis at a 5% level which means the data is not normally distributed. This would indicate there would need to be nonparametric tests going forward. Next would be to look at the analysis for this dataset within section 3.

## 2.3 –Kaggle.NEETs

This dataset came from the website Kaggle where people can freely share datasets they have collected and manipulated. This dataset is made up of “economic indicators from OECD and IMF” (Tubi, 2010) from the dates 1997 to 2018 and was originally used to predict NEET rates. The dataset was made up of 7 variables (2 which were excluded being Date/Time and Location) with 657 observations, this can be seen below.



(Figure 3 – Kaggle.NEETs Summary)

And here is what the abbreviation stands for:

|  |
| --- |
| gdp: Annual Growth rate observations from OECD.  Covers the years between 1997 - 2018 |
| tigs: Trade in Goods and Services (Export) observations from OECD.  Covers the years between 1997 |
| unemp: Unemployment rate observations from IMF.  Covers the years between 1997 |
| inflation: Inflation, consumer prices observations from IMF.  Covers the years between 1997 |

Next will be a distribution analysis to get a better understanding of where the data is within the full dataset, whilst doing a distribution analysis it is a good idea to carry out a normality test which we can reference later.

|  |
| --- |
| H0: The data is normally distributed |
| H1: The data is not normally distributed |

Here are the results for the normality test:

|  |  |
| --- | --- |
| **Variable** | **P Value** |
| gdp | <0.010 |
| tigs | <0.010 |
| neet | <0.010 |
| unemp | <0.010 |
| inflation | <0.010 |

The results from the test for normality is significant at a 5% level, meaning the data is not normally distributed and we would assume there is not enough evidence to reject the alternative hypothesis. Here we can see that all the results have turned out the same results for normality, this could be a good thing as all the data is not normally distributed and we can keep in mind a nonparametric test, but it could also be a bad thing. This is something to be mindful of when moving forward with the rest of the analysis.

# 3 – Analysis

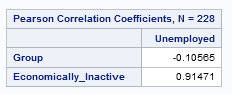
After carrying out some EDA above, this part of the report will focus on the main analysis of the datasets. This section will cover a linear regression, a cluster analysis and a PCA/Factor analysis.

## 3.1 - Linear Regression

For the first data analysis model within this report, a linear regression will be performed on the dataset NEETs.1. This will be carried out after doing the EDA above and seeing the summary and distribution of the dataset. Seeing as we covered the test for normality within the EDA, the next step would be to check if there is a linear relationship between the variables, this can be done with a correlation analysis.

|  |
| --- |
| H0: There is no linear relationship between the variables. |
| H1: There is a linear relationship between the variables. |

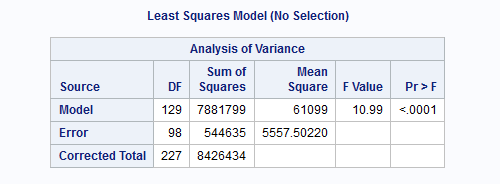
After carrying out a correlation analysis, the results are shown below:



(Figure 4 – Correlation Analysis NEETs.1)

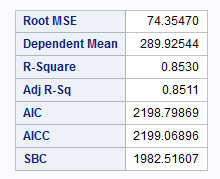
With the results shown, here we can see that the unemployed variable has little or no association to the type of group a person is in, with the unemployed variable having a strong positive association with the economically\_inactive variable. This would mean there is only 1 linear relationship between the 3 variables, which means there could be issues whilst carrying out a linear regression analysis.

The next step would be to carry out a linear regression analysis, which will be carried out by using the dependent variable *unemployed* and using the independent variables *group* as a classification variable & *economically\_inactive* as a continuous variable. This will be carried out starting with 2 intercepts, the two independent variables above with the results shown below.



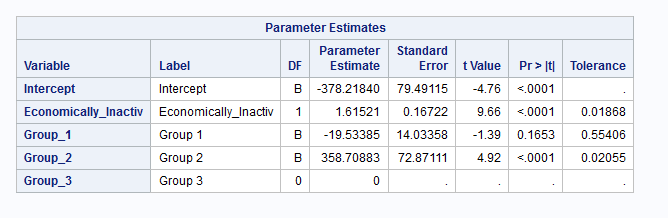
(Figure 5 – Analysis of Variance NEETs.1)

With the results shown with a p value of <.0001, we can reject the null hypothesis at the 5% level and even a 1% level. This means we accept that a linear model is appropriate for this dataset, and therefor will need to carry out a few assumptions to validate this. It is also interesting to see that the overall average of the data was. This can be seen under dependent mean which shows 289.92 with an R-Square value of 0.85 which means there is an 85% of variability within the unemployed.



(Figure 6 – Additional information about NEETs.1)

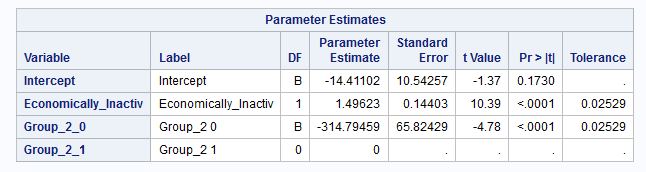
Next would be to look at a few more assumptions, here we will check the tolerance between all the variables.



(Figure 7 – Tolerance of NEETs.1)

Here the tolerance values are ranging from 0.01 being the lowest to 0.55 being the highest, this would suggest that the variable economically\_inactive does not seem to be independent which we would need to be mindful of when going forward with the next few steps, and group 1 (and group 3 which seems to be omitted) looks like the only ones to not be significant.

Here are the results with just the significant variables.



(Figure 8 – Tolerance of significant variables in NEETs.1)

The Bi value for economically\_inactive as the predictor value tells us that for each increase of 1 in unemployed cause the economically\_inactive to increase by 1.49.

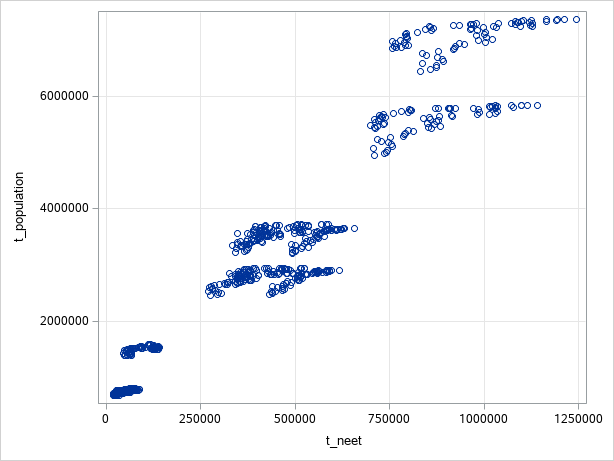
It is also important to highlight the Residual Predicted value plot ([top left](#_6.4_–_Neet.1)) which almost looks to have a pattern which again, we would need to be mindful of going forward. The next step would be to carry out a Durbin-Watson test by adding DW into the code window within SAS. At this step, [DW](#_6.5_–_Neet.1) would not work so this step will be ignored and will move onto the next step.

This would conclude the linear regression model for this dataset, but there are still a few things to take away. A lot of the variables within this dataset could have been more accurate and could have had a longer time frame which could have improved this type of model. There were also a few steps which could not have be carried out which could have affected the output.

## 3.2 – Cluster Analysis

This section will focus on looking at a cluster analysis for the dataset UK.NEETs, by doing a cluster analysis on this dataset could bring to light some insightful groups within the dataset. Whilst some datasets could already be grouped, for example, group by age, gender, location etc, cluster analysis can be used to find hidden patterns within the data.

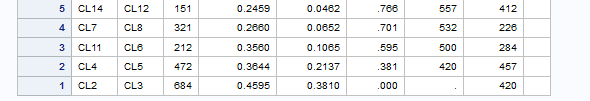
Since the EDA has already been completed above which stated the data was not normally distributed, it is clear to see the known groups so let us start with looking at the data as a scatter plot to see if anything clearly stands out about the data.



(Figure 9 – Scatter graph of total neets v total population)

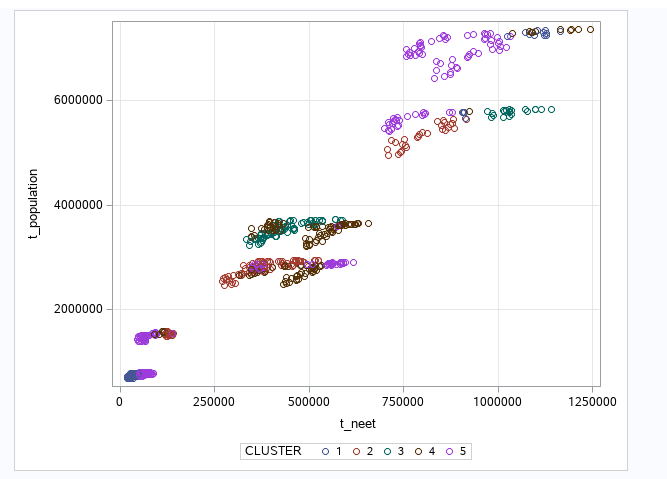
From first look, the data looks like it could be split into 6 small groups, or 3 decent sized groups. Whilst not been tested statistically at all, this is something which quickly stands out from first look, that is the data looks to be within groups as the data t\_population grows, so does the t\_neets. Within the Appendix [6.6](#_6.6_–_Kaggle.NEETs), [6.7](#_6.7_–_Kaggle.NEETs), [6.8,](#_6.8_–_Kaggle.NEETs) there will be scatters for the rest of the other variables against t\_neet.

The next step would be to start carrying out a cluster observation to see how many clusters we would need to use. After carrying out this analysis, there were a total of 683 clusters as it does a cluster for the number of observations. The cluster history table shows a lot of information from Semipartial R-square, R-square, Pseudo F Statistic and Pseudo t-Squared. With this information, here we want to find a pattern which goes from a high value to a low value with the least number of clusters.



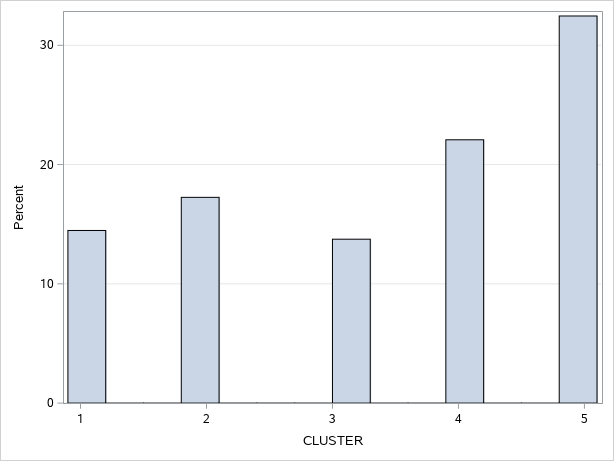
(Figure 10 – First 5 clusters from Cluster History)

Here we can see whilst the jump between 2 to 3 is big, the jump between 4 to 5 is also quite big. So, let us start with saying there are 5 clusters and using the output data to visualise this.



(Figure 11 – Scatter chart of clustered UK-NEETs)

Whilst not visually appealing, the cluster analysis has worked and here we can see there are multiple clusters within the total neets and total population throughout the chart. With cluster 5 having a higher percent of observations.



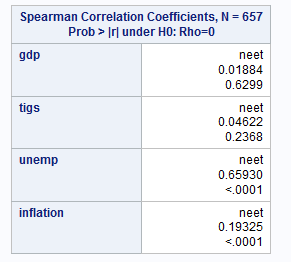
(Figure 12 – Number of observations per cluster)

To conclude on this section, we can see that cluster analysis can work with how ever many clusters you want to get, but this can not be statistically correct. However, cluster analysis with k-means could have also been run to check this cluster against that one, which could have brought to light more information about how many clusters would have been correct.

## 3.3 – Factor Analysis

This section of the report will go over a Factor analysis for the dataset Kaggle.NEETs, where the results will be explored and interpreted to see if there is any useful information which reflects the number of NEETs.

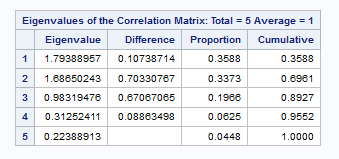
Before starting with any analysis, lets first see if there is any correlation between the variables within the dataset. This can be done with a correlation analysis and looking at the nonparametric results (Spearman). The results are as follows:



(Figure 13 – Spearman correlation for Kaggle.NEETs)

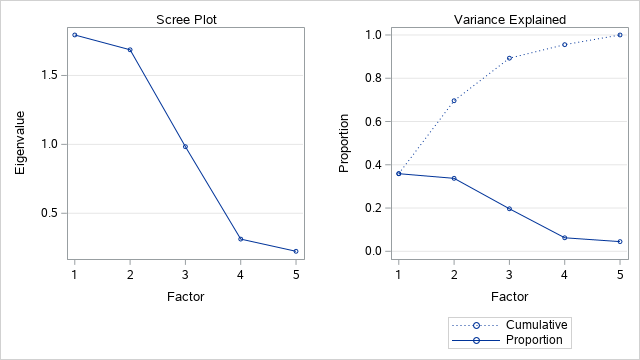
From the correlation analysis we can see that there are 2 variables which have a strong correlation between the variable NEET, these variables are inflation and unemployment. This would suggest that these 2 variables affect the numbers of NEETs which makes sense.

Next would be to carry out a factor analysis now that this assumption has been met, we will start by following the Kaiser rule which was introduced by Kaiser (1959) which recommends only including components in the analysis with an Eigenvalue greater than one. Whilst looking at Factor Analysis for this dataset, only 2 factors had Eigenvalue greater than 1 and it made up a total of 69% of the proportion of variance. If we include the third factor which had an Eigenvalue of 0.983, this will take the total proportion of variance up to 89% which could improve the model.



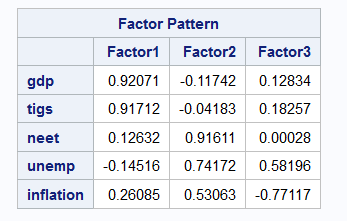
(Figure 14 – Eigenvalues for Kaggle.NEETs)

Even though Cattell (1966) proposed the graphical approach to looking at the factors, stating to look for the factors above the ‘elbow’, this Scree Plot would suggest either taking 2 or 3 factors, which could be fine, but they have much lower Eigenvalue.



(Figure 15 – Screen Plot for Kaggle.NEETs)

Factor analysis will now be run with only 3 factors to account for the 89% of variance, with a factor pattern with these results below:

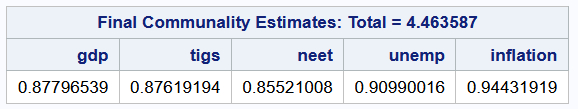


(Figure 16 – Factor Pattern for Kaggle.NEETs with 3 factors)

From these results we can see that gdp and tigs both have a strong loading for factor 1, with neet, unemp being strongly loaded for factor 2 with inflation having a moderate loading. With factor 3 not having a strong loading for any of the variables, other than the strong negative loading for inflation. This could be because factor 3 did not have an Eigenvalue >1. From this we can now explore the variance by each factor which can give us a total of 4.46 dispersed by variable.

|  |
| --- |
| Factor 1 is highly loaded with gdp and tigs which could represent the global |
| Factor 2 is highly loaded with neet, which represents the number of NEETs for the country. |

With the Final Communality Estimates were as follows:



(Figure 17 – Final Communality Estimates used)

# 4 - Conclusion

Throughout this report there have been 3 different statistical tests processed which looked at the number of NEETs within the datasets, which came from 3 different sources. Whilst this report has carried out some more in-depth statistical analysis compared to the first report (CW1), there was a lot more to learn from this report.

The total number of NEETs per country is highly affected by its population, which makes sense when you think about it as it could be very difficult to get everyone within a country a job. Another take away was how much more difficult these statistical tests were, this could have been due to my understanding of the tests or could also have required more pre-processing.

Another take away is how this data could have changed (I am assuming significantly) with the rising COVID-19 cases since the end of 2020 here within the UK (GOV.UK, 2021) with many local shops closing here within the valleys. This could also be interesting to see how this has affected the other countries with their shops and businesses trying to move to online only, which could be carried out within further research.

Throughout this report there have also been a few limitations, December 21st to January 13th there was no internet within the house due to moving several days before Christmas. I have carried out as much as I could with the data I acquired before the internet being off (along with a sas studio virtual machine) and very limited mobile data and signal within the house.

# 5 - References

GOV.UK. (2020, 12 13). *Education and training statistics for the UK*. Retrieved from GOV.UK: https://explore-education-statistics.service.gov.uk/find-statistics/education-and-training-statistics-for-the-uk/2020

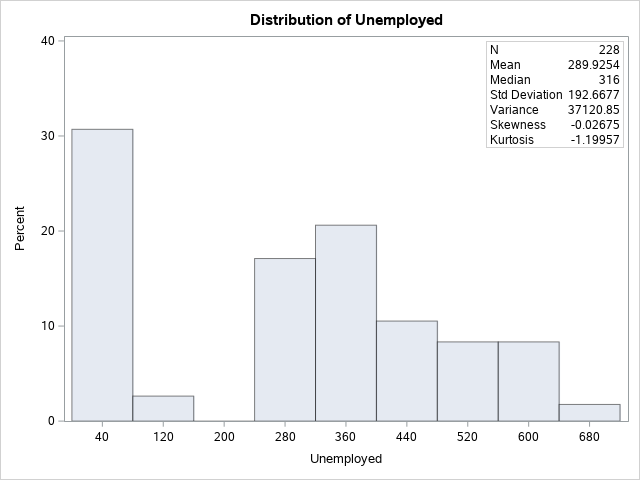
GOV.UK. (2021, 01 10). *Coronavirus (COVID-19) in the UK*. Retrieved from GOV.UK: https://coronavirus.data.gov.uk/

Office for National Statistics. (2020, 12 14). *Young people not in education, employment or training (NEET)* . Retrieved from Office for National Statistics: https://cy.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/youngpeoplenotineducationemploymentortrainingneettable1

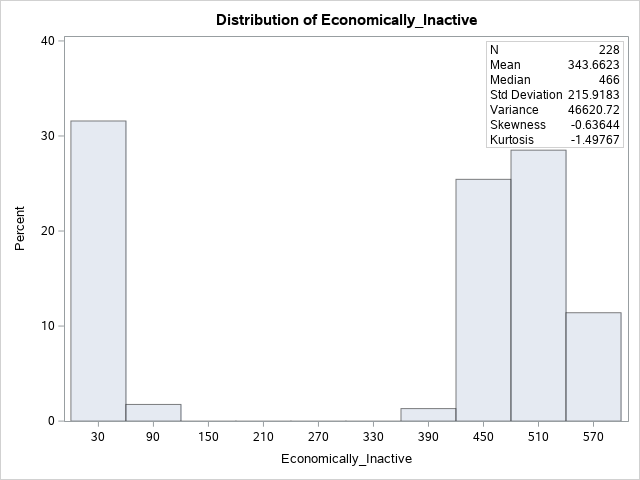
Tubi. (2010, 08 01). *Youth Not in Employment Education or Training*. Retrieved 12 05, 2010, from Kaggle: https://www.kaggle.com/keremtugberk/youth-not-in-employment-education-or-training/version/1

# 6 - Appendix

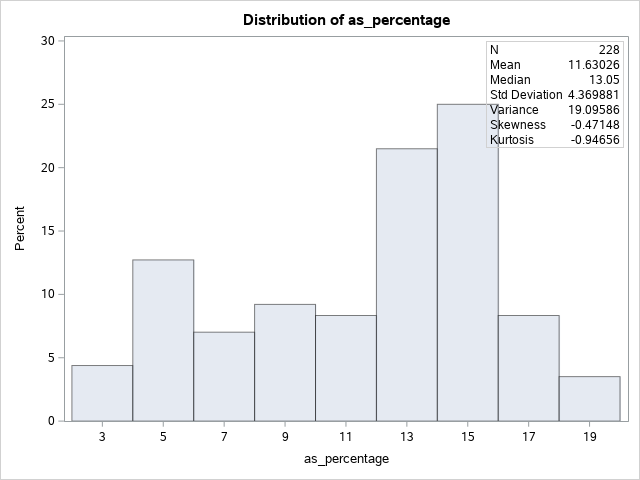
## 6.1 – NEET.1 Unemployed Distribution



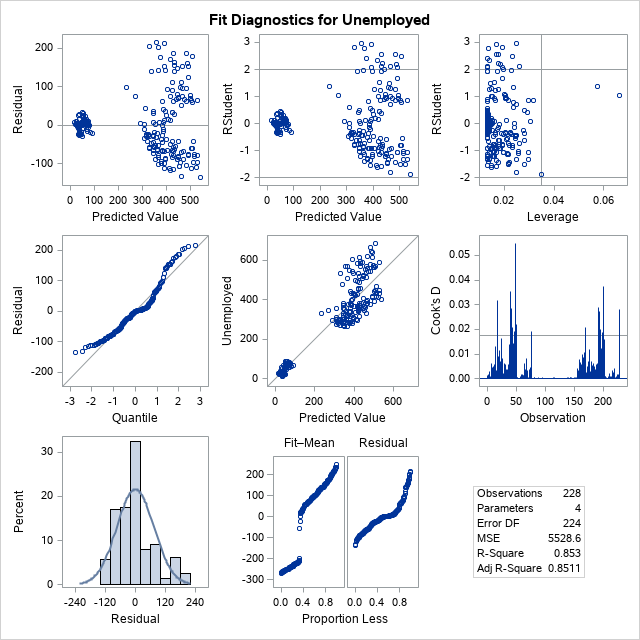
## 6.2 – NEET.1 Economically\_Inactive Distribution



## 6.3 – NEET.1 As\_Percentage Distribution



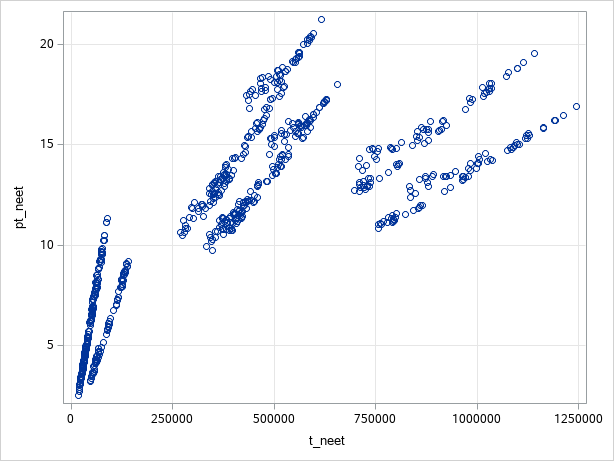
## 6.4 – NEET.1 Fit Diagnostics for Unemployed



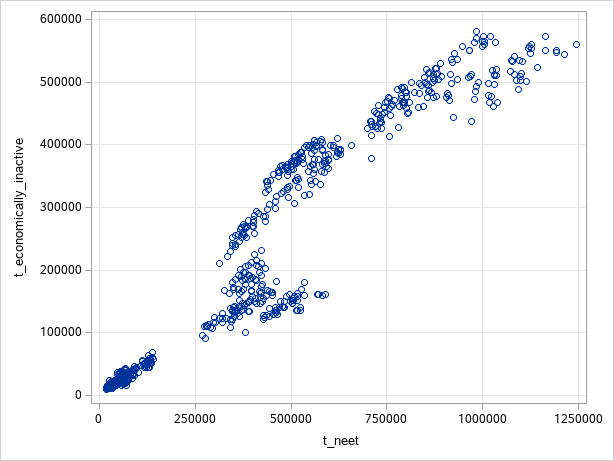
## 6.5 – NEET.1 DW Code (does not work)

|  |
| --- |
| ods noproctitle;  ods graphics / imagemap=on;  proc glmselect data=MYWORK.'NEETS.1'n outdesign(addinputvars)=Work.reg\_design;  class Group / param=glm;  model Unemployed=Economically\_Inactive Group / showpvalues selection=none;  run;  proc reg data=Work.reg\_design alpha=0.05 plots(only)=(diagnostics residuals  observedbypredicted);  where Group is not missing;  ods select ParameterEstimates DiagnosticsPanel ResidualPlot  ObservedByPredicted;  model Unemployed=&\_GLSMOD / tol dw;  run;  quit;  proc delete data=Work.reg\_design;  run; |

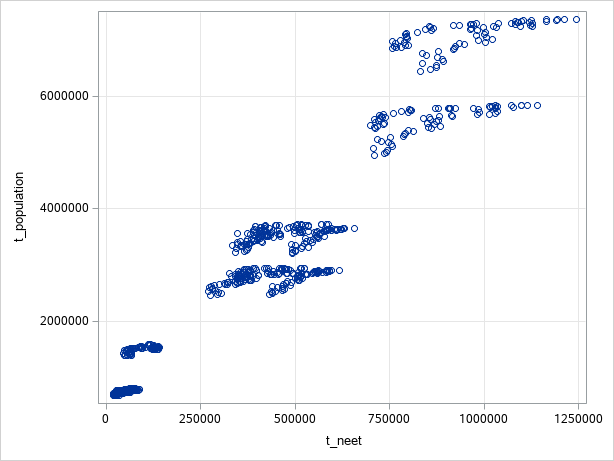
## 6.6 – Kaggle.NEETs – Scatter pt\_neet



## 6.7 – Kaggle.NEETs – Scatter t\_economically\_inactive



## 6.8 – Kaggle.NEETs – Scatter t\_population



## 6.7 – Code snippet for Linear Regression

|  |
| --- |
| ods noproctitle;  ods graphics / imagemap=on;  proc glmselect data=MYWORK.'NEETS.1'n outdesign(addinputvars)=Work.reg\_design;  class Group / param=glm;  model Unemployed=Economically\_Inactive Group / showpvalues selection=none;  run;  proc reg data=Work.reg\_design alpha=0.05 plots(only)=(diagnostics residuals  observedbypredicted);  where Group is not missing;  ods select ParameterEstimates DiagnosticsPanel ResidualPlot  ObservedByPredicted;  model Unemployed=&\_GLSMOD /dw;  run;  quit;  proc delete data=Work.reg\_design;  run; |

## 6.8 – Code snippet for Cluster Analysis

|  |
| --- |
| ods noproctitle;  /\*\*\* Standardize variables and create distances \*\*\*/  proc distance data=MYWORK.'UK.NEETS'n method=dsqcorr out=Work.\_tmp\_distances;  var interval(t\_neet t\_unemployed t\_economically\_inactive t\_population pt\_neet  / std=std);  copy t\_neet t\_unemployed t\_economically\_inactive t\_population pt\_neet;  run;  proc cluster data=Work.\_tmp\_distances method=ward pseudo rmsstd  plots(only)=(dendrogram) outtree=WORK.clustertree;  var Dist:;  copy t\_neet t\_unemployed t\_economically\_inactive t\_population pt\_neet;  run;  /\* do the same but only with 5 clusters \*/  proc tree data=Work.clustertree noprint n=5 out=treedata;  copy t\_neet t\_unemployed t\_economically\_inactive t\_population pt\_neet;  run;  proc delete data=Work.\_tmp\_distances;  run; |