GY7702_Assignment_2

Student ID: 209047191

Date of Submission: 04/01/2021

Contents

1 GY7702 Assignment 1

This assignment entails developing skills in exploratory data analysis and geodemographic classification. With the data focused in the 2011 Output Area Classification for the assigned LAD and this document was created in RMarkdown. **Option B** was selected from the assignment and the code, output and analysis are presented in this document.

This assignment was created using R, Rstudio, RMarkdown and GitHub.

For all P-values for normality test > 0.01 is used as the threshold for significant normality.

Github Link: space-uni

Bitly Link created from the above Github link: https://bit.ly/3889N1a

1.1 References

This assignment would like to acknowledge that this document includes teaching materials from Dr Stefano De Sabbata for the module GY7702 R for Data Science. And the associated teaching materials can be found here

R for Data Science by Garrett Grolemund and Hadley Wickham, O'Reilly Media, 2016. See online book

This repository / document contains public sector information licensed under the Open Government Licence v3.0: 2011_OAC_Raw_kVariables.csv, 2011_OAC_Raw_kVariables_Lookup.csv, OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv and covid_cases_total_MSOA_20201123.csv. See also **Gale et al (2016)**, **Chris Gale's GitHub pages**, **UK Census geography**, **Coronavirus (COVID-19) in the UK website** and **Office for National Statistics**.

1.2 Introduction

Main Datasets:

- 2011_OAC_Raw_kVariables.csv: a dataset containing the 60 final variables used by **Gale et al (2016)** to create the 2011 Output Area Classification (2011OAC, see also **Chris Gale's GitHub pages**);
- 2011_OAC_Raw_kVariables_Lookup.csv: a table containing all the information (metadata) about the 60 variables included in 2011_OAC_Raw_kVariables.csv;
- OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv: a dataset indicating for each census Output Area (OA), the Lower-Super Output Area (LSOA), Middle-Super Output Area (MSOA) and Local Authority District (LAD) that contain it (see also **UK Census geography**)
- covid_cases_total_MSOA_20201123.csv: a dataset listing the total number of covid-19 cases for each MSOA in England (see also **Coronavirus (COVID-19) in the UK website**)

1.3 Prerequisites

```
# load libraries

library(tidyverse)
library(magrittr)
library(lubridate)
library(knitr)
library(ggplot2)
```

```r
library(psych)
library(GGally)
library(fmsb)

# load datasets

OAC_2011 <- read_csv("Data/2011_OAC_Raw_kVariables.csv")
OAC_2011_lookup <- read_csv("Data/2011_OAC_Raw_kVariables_Lookup.csv")
OA11 <- read_csv("Data/OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv")
```

1.4 LAD Dataset

Joining the 2011_OAC_Raw_kVariables.csv and OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv datasets, and creating a dataset containing only the data regarding the LAD of "Cotswold".

```r
# creating a new dataset pertaining to the LAD
OAC_2011_LAD <- OAC_2011 %>%
  # joining the 2011_OAC_Raw_kVariables and OA11_LSOA11_MSOA11_LAD11_EW_LUv2 datasets
  dplyr::full_join(
    OA11, by = c("OA" = "OA11CD"
  )) %>%
  # filtering for LAD of Cotswold
  dplyr::filter(
    LAD11NM == "Cotswold"
  )
```

1.5 Option B

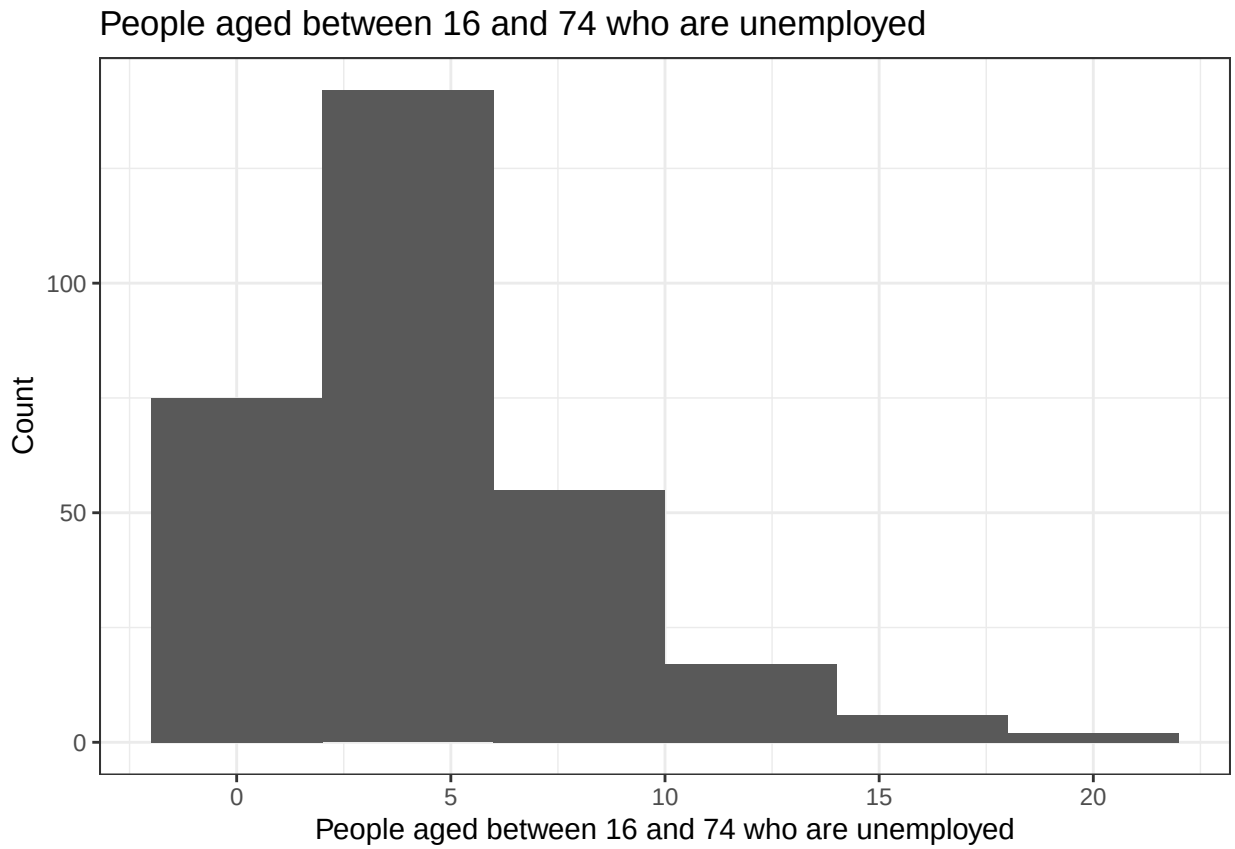
1.5.1 Question B.1

An exploratory analysis of the variables listed below, from the 2011_OAC_Raw_kVariables.csv dataset, for the OAs in Cotswold.

- k045 (People aged between 16 and 74 who are unemployed)
- k046 (Employed persons aged between 16 and 74 who work part-time)
- k047 (Employed persons aged between 16 and 74 who work part-time)
- k048 (Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries)
- k049 (Employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries)
- k050 (Employed persons aged between 16 and 74 who work in the manufacturing industry)
- k059 (Employed persons aged between 16 and 74 who work in the education sector)

1.5.1.1 k045 (People aged between 16 and 74 who are unemployed)

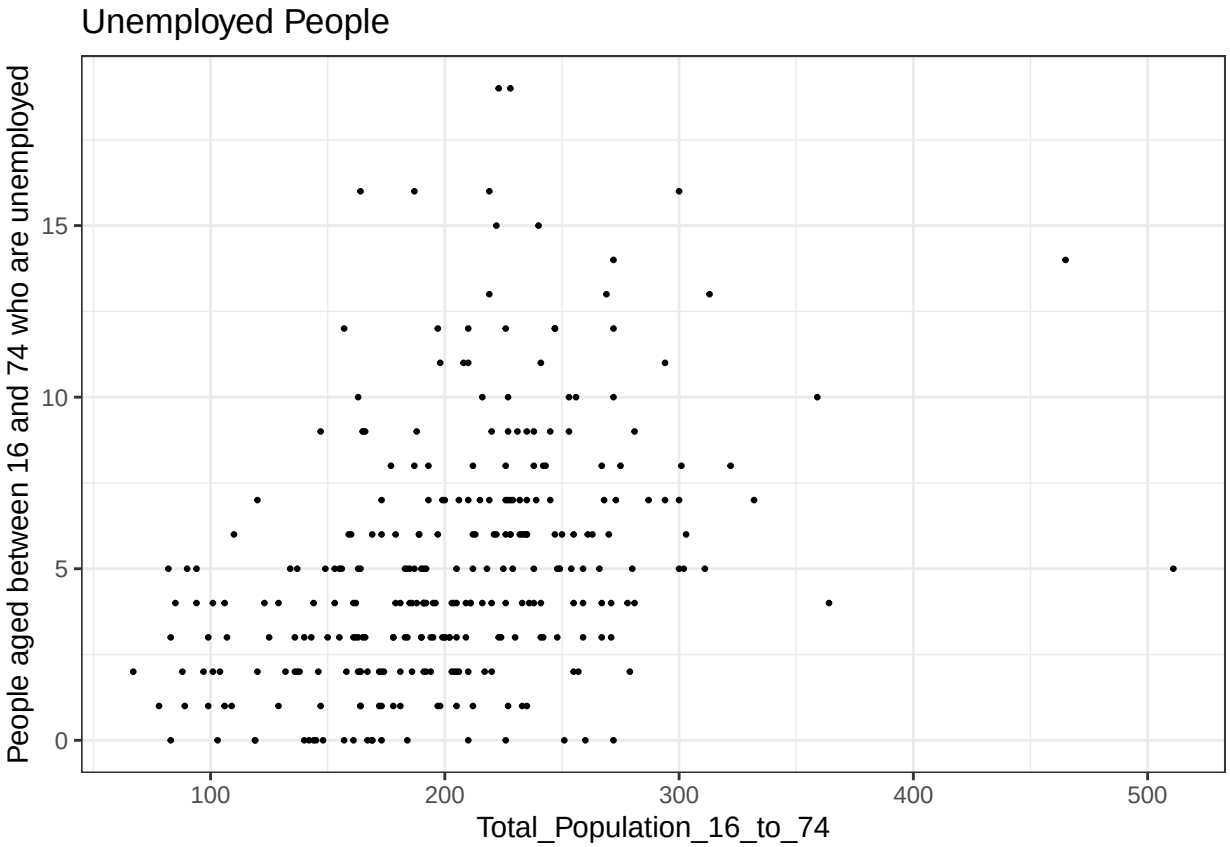
```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who are unemployed
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k045
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("People aged between 16 and 74 who are unemployed") +
  ggplot2::xlab("People aged between 16 and 74 who are unemployed") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```




```
# Applying a shapiro test to the k045 variable
OAC_2011_LAD  %>%
  dplyr::pull(k045) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.91701, p-value = 9.176e-12
```

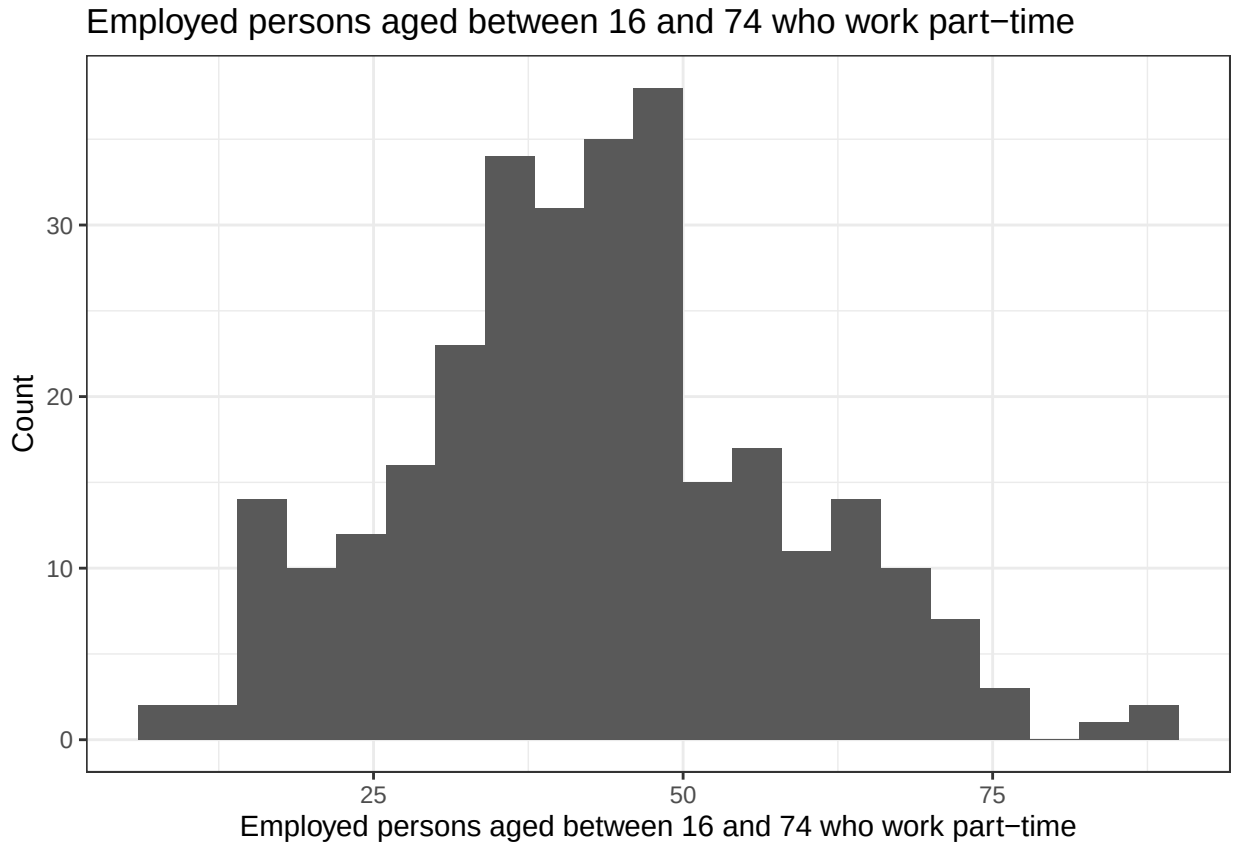
```
# Plotting total employment counts against the variable k045
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Population_16_to_74,
      y = k045
    )
  ) +
  ggplot2::geom_point(size = 0.5) +
  ggplot2::ggtitle("Unemployed People") +
  ggplot2::xlab("Total_Population_16_to_74") +
  ggplot2::ylab("People aged between 16 and 74 who are unemployed") +
  ggplot2::theme_bw()
```



It is clear from the initial plot and the Shapiro-Wilk test that the counts of unemployed people (aged 16-74) are skewed and does not have significant normality. The unemployed people graph against the total population counts (aged 16-74) appear to have a spread out positive trend.

1.5.1.2 k046 (Employed people aged between 16 and 74 who work part-time)

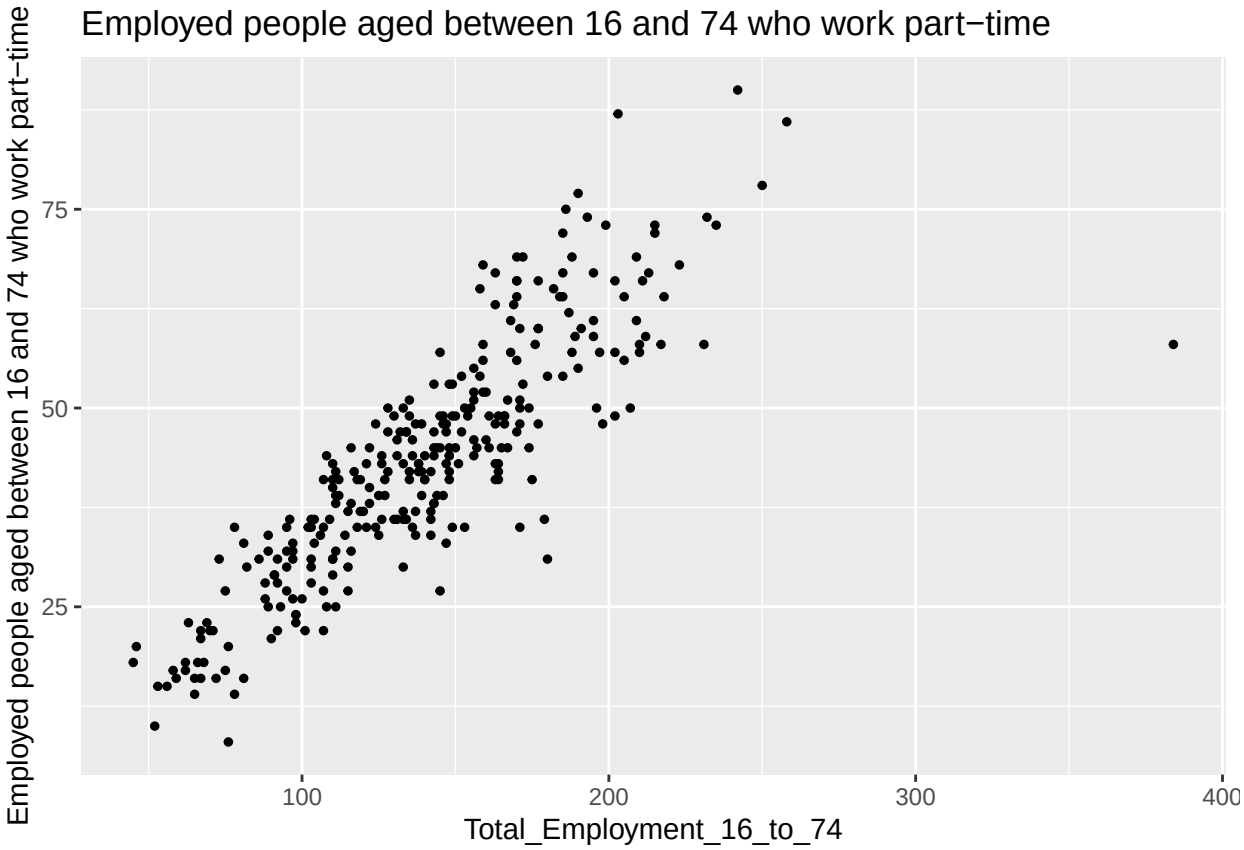
```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who work part-time
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k046
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("Employed persons aged between 16 and 74 who work part-time") +
  ggplot2::xlab("Employed persons aged between 16 and 74 who work part-time") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```




```
# Applying a Shapiro-Wilk test to the k046 variable
OAC_2011_LAD  %>%
  dplyr::pull(k046) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.99081, p-value = 0.05991
```

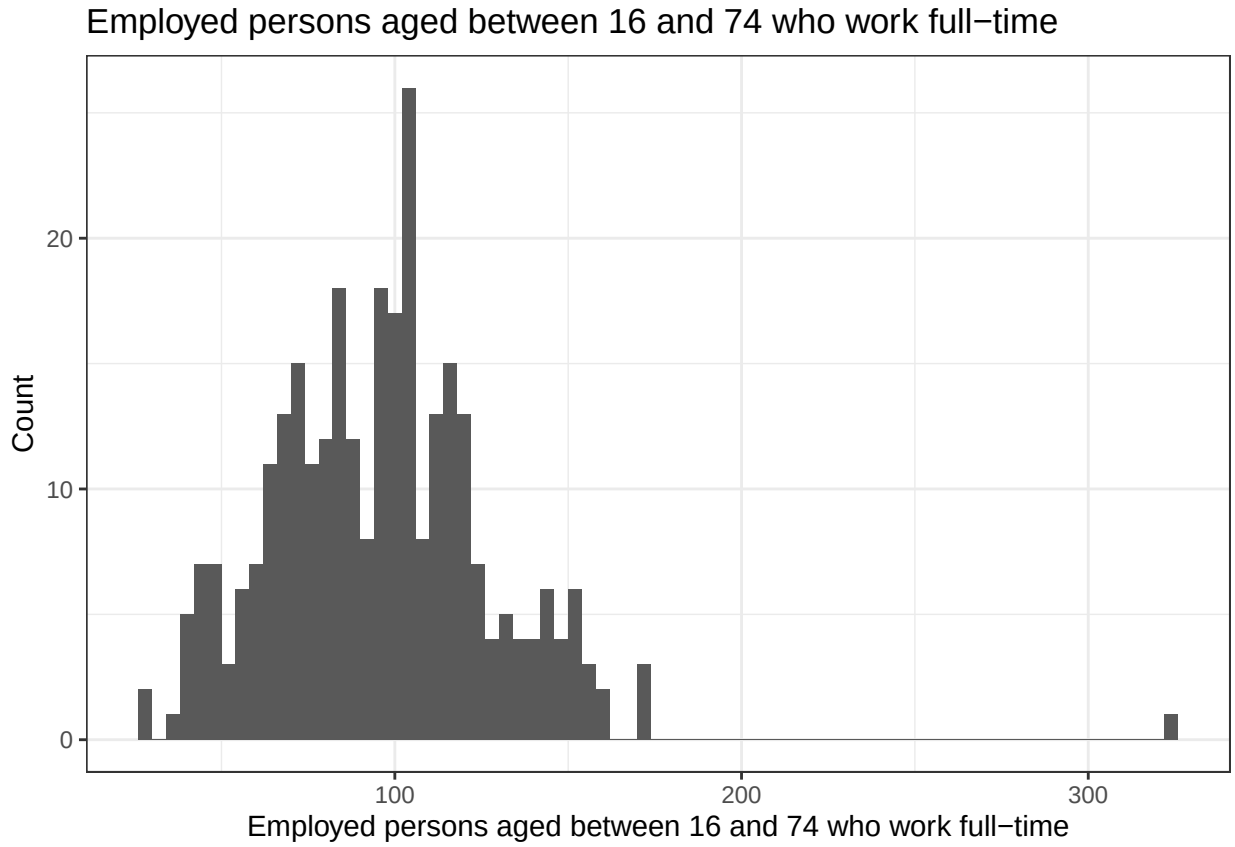
```r
# Plotting total employment counts against the variable k046
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Employment_16_to_74,
      y = k046
    )
  ) +
  ggplot2::geom_point(size = 1) +
  ggplot2::ggtitle("Employed people aged between 16 and 74 who work part-time") +
  ggplot2::xlab("Total_Employment_16_to_74") +
  ggplot2::ylab("Employed people aged between 16 and 74 who work part-time")
```



From the initial graph of counts of people who work part time (aged 16-74), the data may appear to have a normal distribution shape. Utilising the Shapiro-Wilk test proves that the p-value $= 0.05991 > 0.01$ which confirms the data having significant normality. Including the fact that the final graph of part time vs total counts of employed people has a much tighter concentration of a positive trend.

1.5.1.3 k047 (Employed persons aged between 16 and 74 who work full-time)

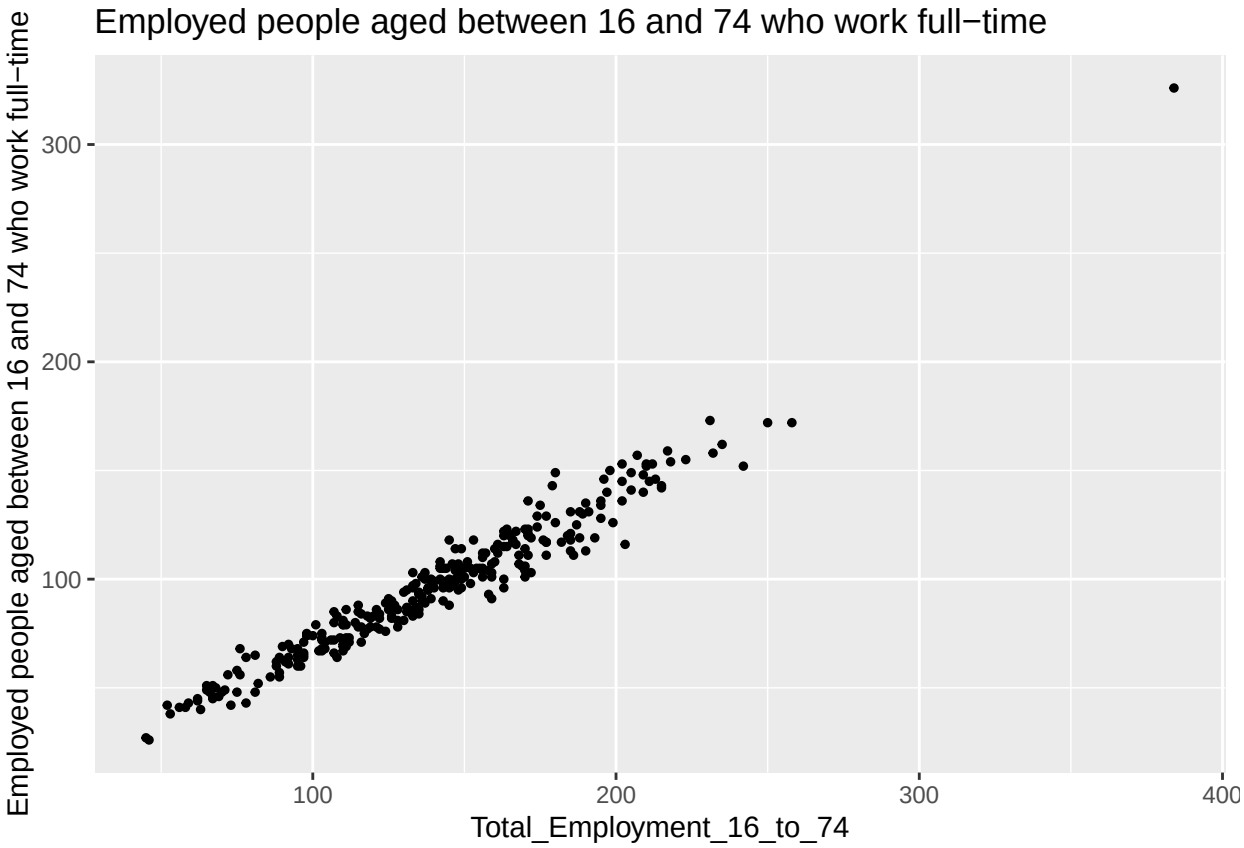
```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who work full-time
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k047
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("Employed persons aged between 16 and 74 who work full-time") +
  ggplot2::xlab("Employed persons aged between 16 and 74 who work full-time") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```




```
# Applying a Shapiro-Wilk test to the k047 variable
OAC_2011_LAD  %>%
  dplyr::pull(k047) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.93416, p-value = 3.3e-10
```

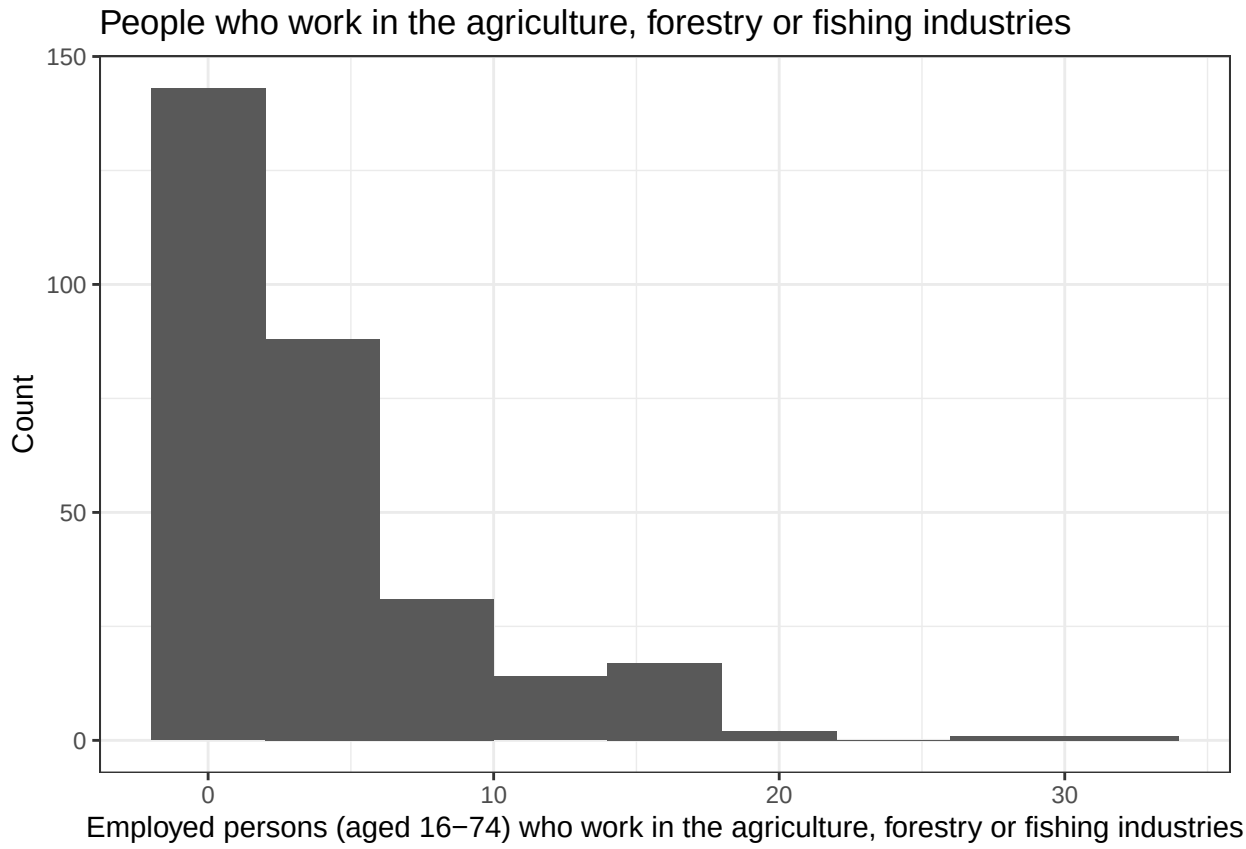
```r
# Plotting total employment counts against the variable k047
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Employment_16_to_74,
      y = k047
    )
  ) +
  ggplot2::geom_point(size = 1) +
  ggplot2::ggtitle("Employed people aged between 16 and 74 who work full-time") +
  ggplot2::xlab("Total_Employment_16_to_74") +
  ggplot2::ylab("Employed people aged between 16 and 74 who work full-time")
```



From the initial graph of counts of people who work full time (aged 16-74), the data may appear initially appear to have normality but upon close inspection it is clearly skewed which also correlates with the the Shapiro-Wilk test. Also the final graph of full time vs total counts of employed people has a much tighter concentration of a positive trend.

1.5.1.4 k048 (Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries)

```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who work in the agriculture, forestry or fishing industries
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k048
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("People who work in the agriculture, forestry or fishing industries") +
  ggplot2::xlab(
    "Employed persons (aged 16-74) who work in the agriculture, forestry or fishing industries"
  ) +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```

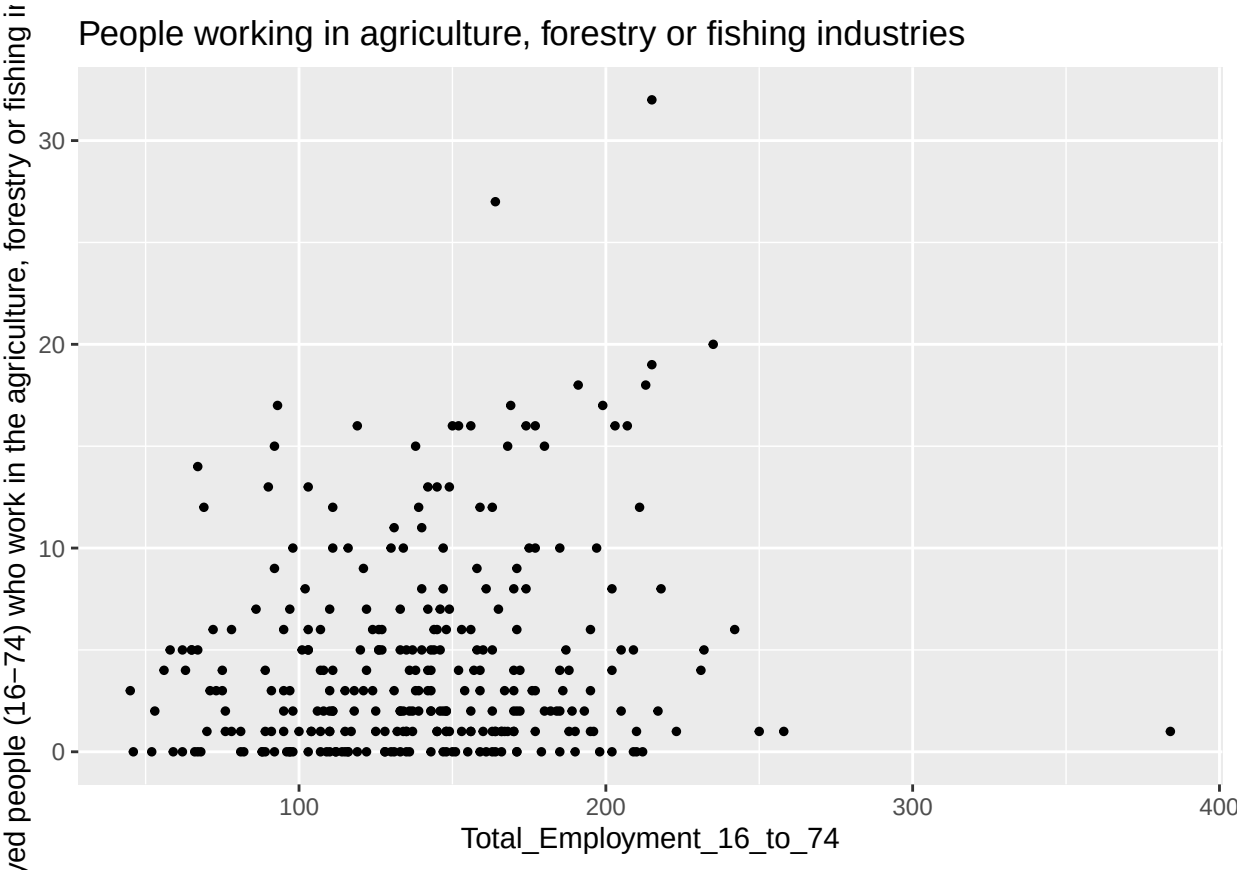



```
# Applying a Shapiro-Wilk test to the k048 variable
OAC_2011_LAD  %>%
  dplyr::pull(k048) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  .
## W = 0.79925, p-value < 2.2e-16
```

```
# Plotting total employment counts against the variable k048
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Employment_16_to_74,
      y = k048
    )
  ) +
  ggplot2::geom_point(size = 1) +
  ggplot2::ggtitle("People working in agriculture, forestry or fishing industries") +
  ggplot2::xlab("Total_Employment_16_to_74") +
  ggplot2::ylab(
    "Employed people (16-74) who work in the agriculture, forestry or fishing industries"
  )
```



From the first plot of counts of people who work in the agriculture, forestry or fishing industries and Shapiro-Wilk test, the data is skewed and does not have significant normality. And the final graph against total employment counts (aged 16-74), the data does not have a clear trend and distinctive relationship. Therefore indicating that people who worked in this industry are constant needed and employed regardless of the variation of employed people.

1.5.1.5 k049 (Employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries)

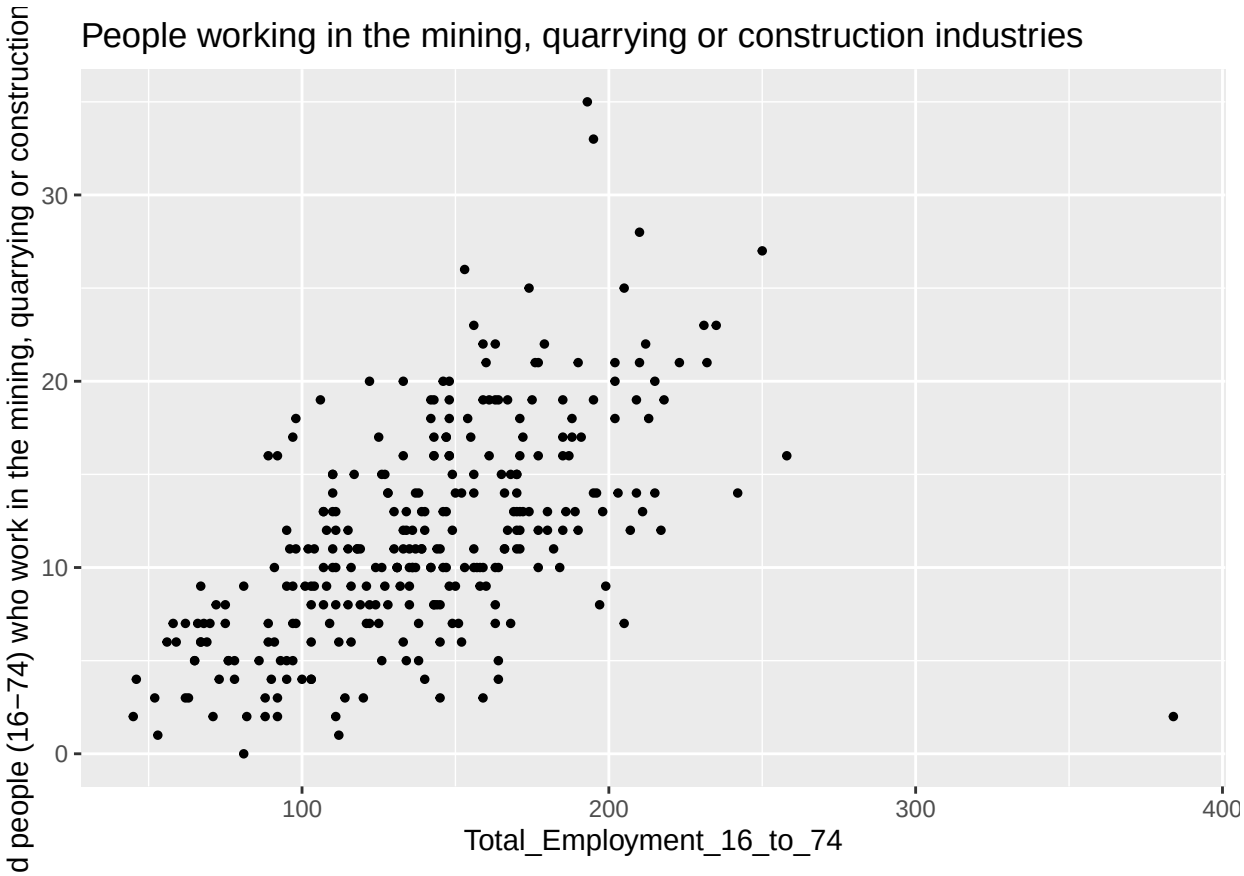
```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who work in the mining, quarrying or construction industries
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k049
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("People who work in the mining, quarrying or construction industries") +
  ggplot2::xlab(
    "Employed persons (aged 16-74)who work in the mining, quarrying or construction industries"
  ) +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```


```
# Applying a Shapiro-Wilk test to the k049 variable
OAC_2011_LAD  %>%
  dplyr::pull(k049) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  .
## W = 0.97295, p-value = 2.174e-05
```

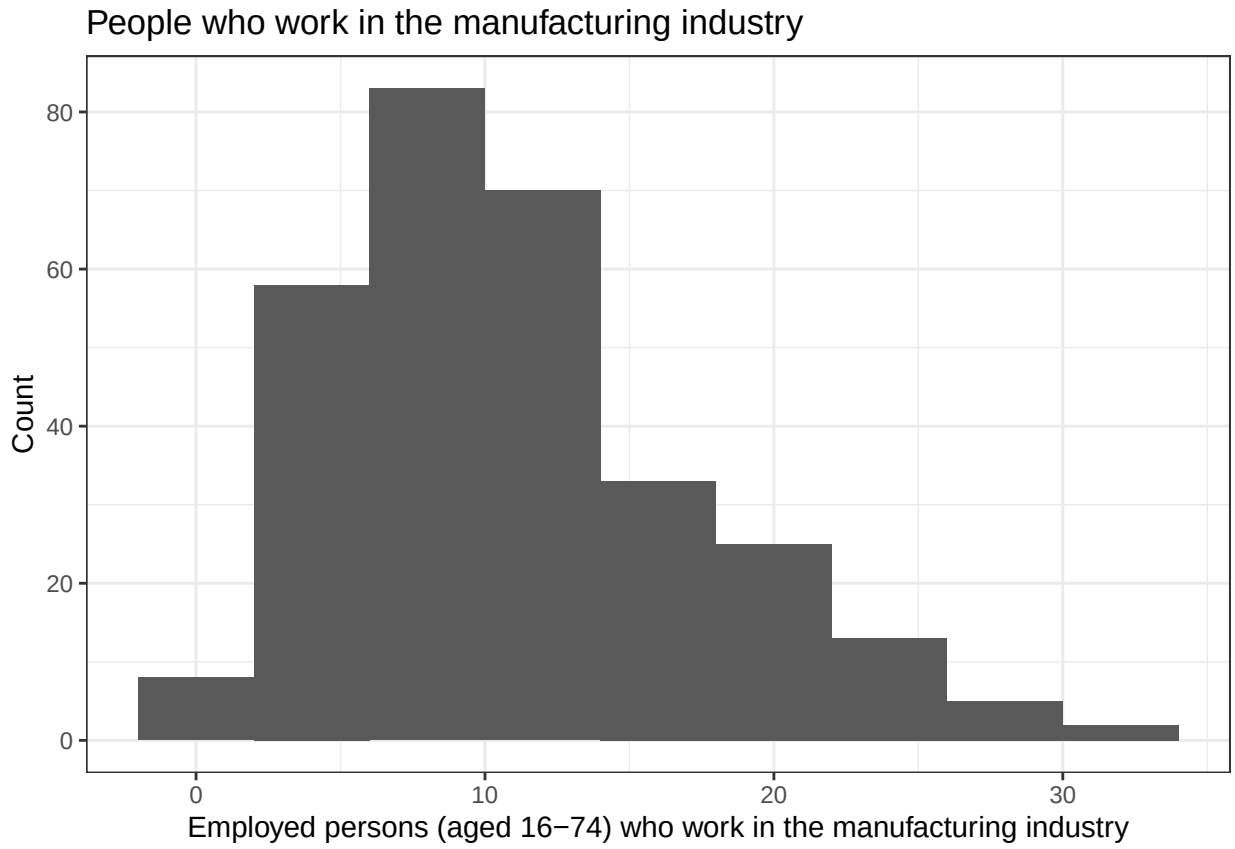
```
# Plotting total employment counts against the variable k049
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Employment_16_to_74,
      y = k049
    )
  ) +
  ggplot2::geom_point(size = 1) +
  ggplot2::ggtitle("People working in the mining, quarrying or construction industries") +
  ggplot2::xlab("Total_Employment_16_to_74") +
  ggplot2::ylab(
    "Employed people (16-74) who work in the mining, quarrying or construction industries"
  )
```



This data on employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries does not have significant normality and are skewed towards the left. The counts against the total employed people appears to have a positive trend.

1.5.1.6 k050 (Employed persons aged between 16 and 74 who work in the manufacturing industry)

```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who work in the manufacturing industry
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k050
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("People who work in the manufacturing industry") +
  ggplot2::xlab(
    "Employed persons (aged 16-74) who work in the manufacturing industry"
  ) +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```

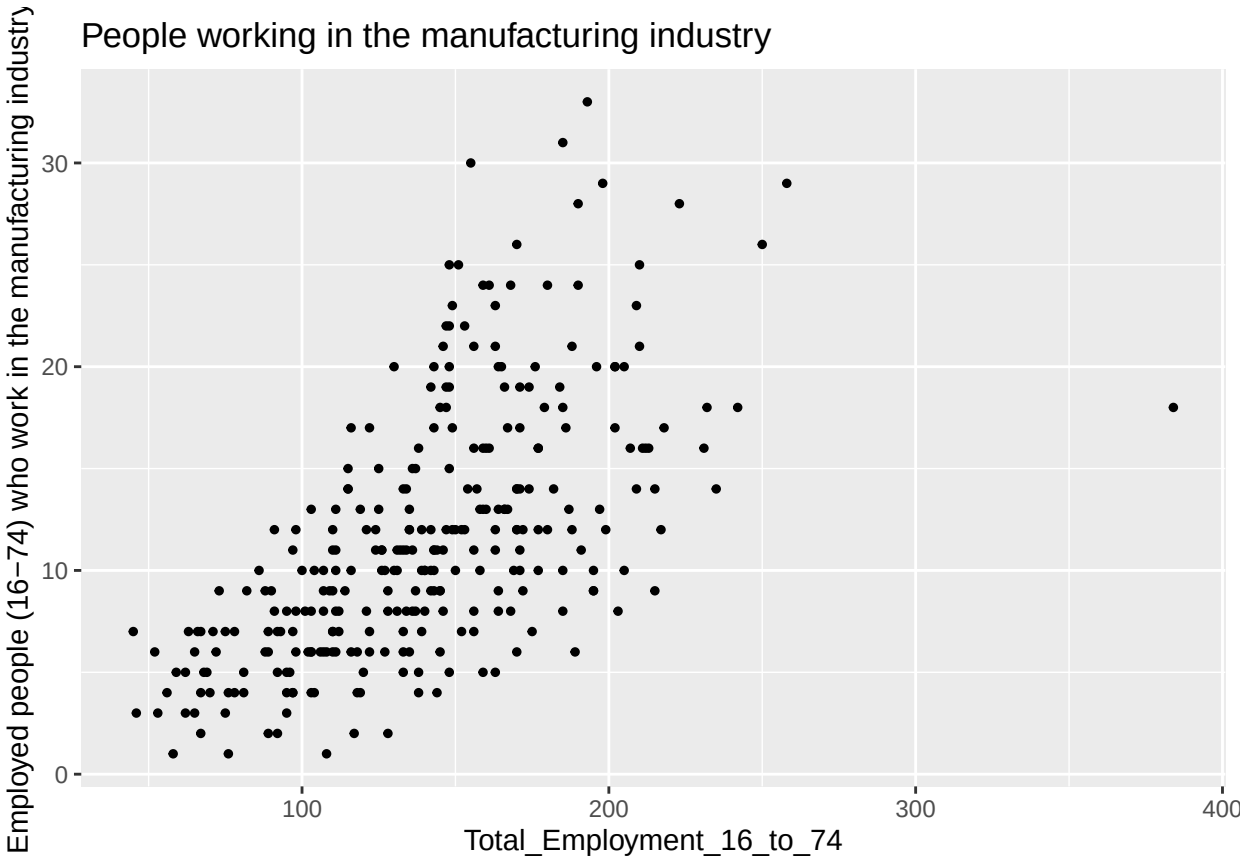



```
# Applying a Shapiro-Wilk test to the k050 variable
OAC_2011_LAD  %>%
  dplyr::pull(k050) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  .
## W = 0.94702, p-value = 7.325e-09
```

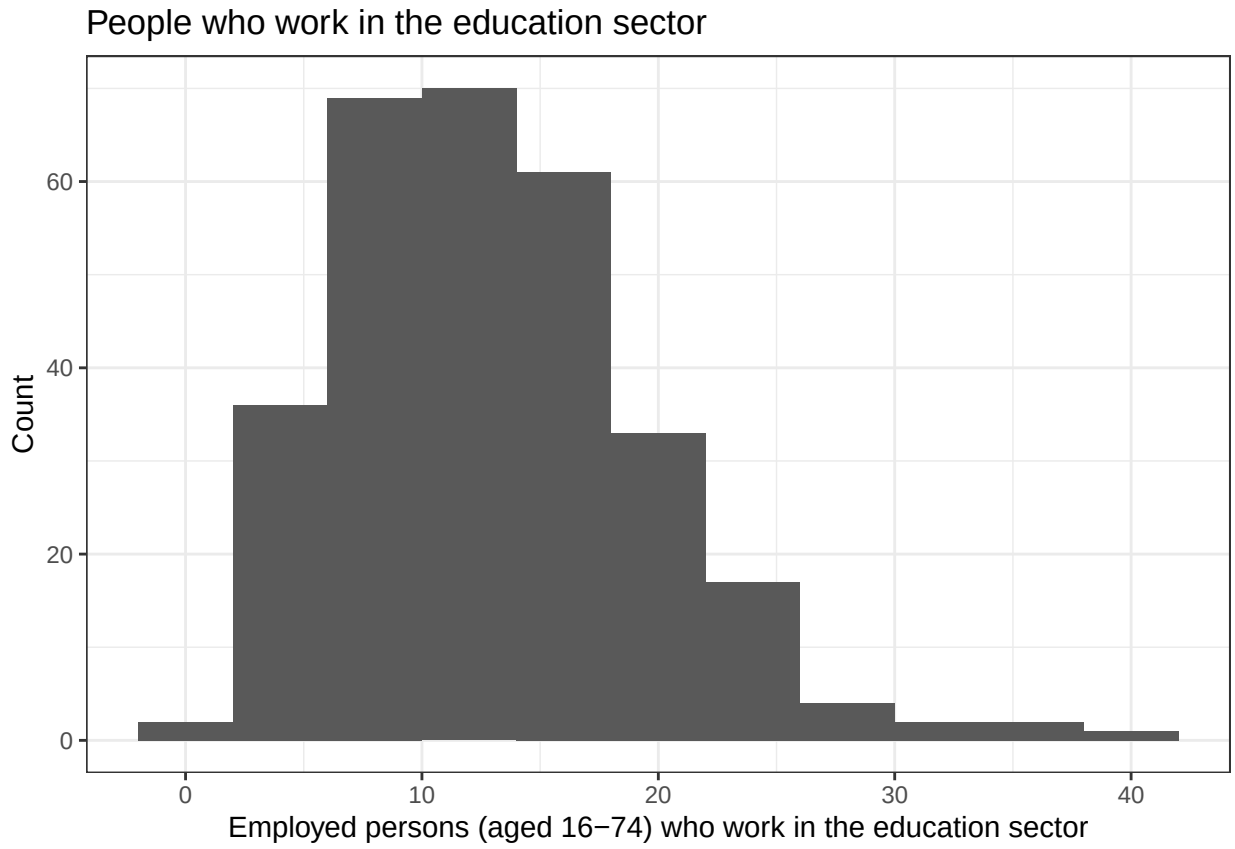
```
# Plotting total employment counts against the variable k050
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Employment_16_to_74,
      y = k050
    )
  ) +
  ggplot2::geom_point(size = 1) +
  ggplot2::ggtitle("People working in the manufacturing industry") +
  ggplot2::xlab("Total_Employment_16_to_74") +
  ggplot2::ylab(
    "Employed people (16-74) who work in the manufacturing industry"
  )
```



This data on employed persons aged between 16 and 74 who work in the manufacturing industry does not have significant normality and are skewed towards the left. The counts against the total employed people appears to have a spreaded out positive trend.

1.5.1.7 k059 (Employed persons aged between 16 and 74 who work in the education sector)

```
# Plotting a histogram displaying the counts of people aged
# between 16 and 74 who work in the education sector
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = k059
    )
  ) +
  ggplot2::geom_histogram(binwidth = 4) +
  ggplot2::ggtitle("People who work in the education sector") +
  ggplot2::xlab(
    "Employed persons (aged 16-74) who work in the education sector"
  ) +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```

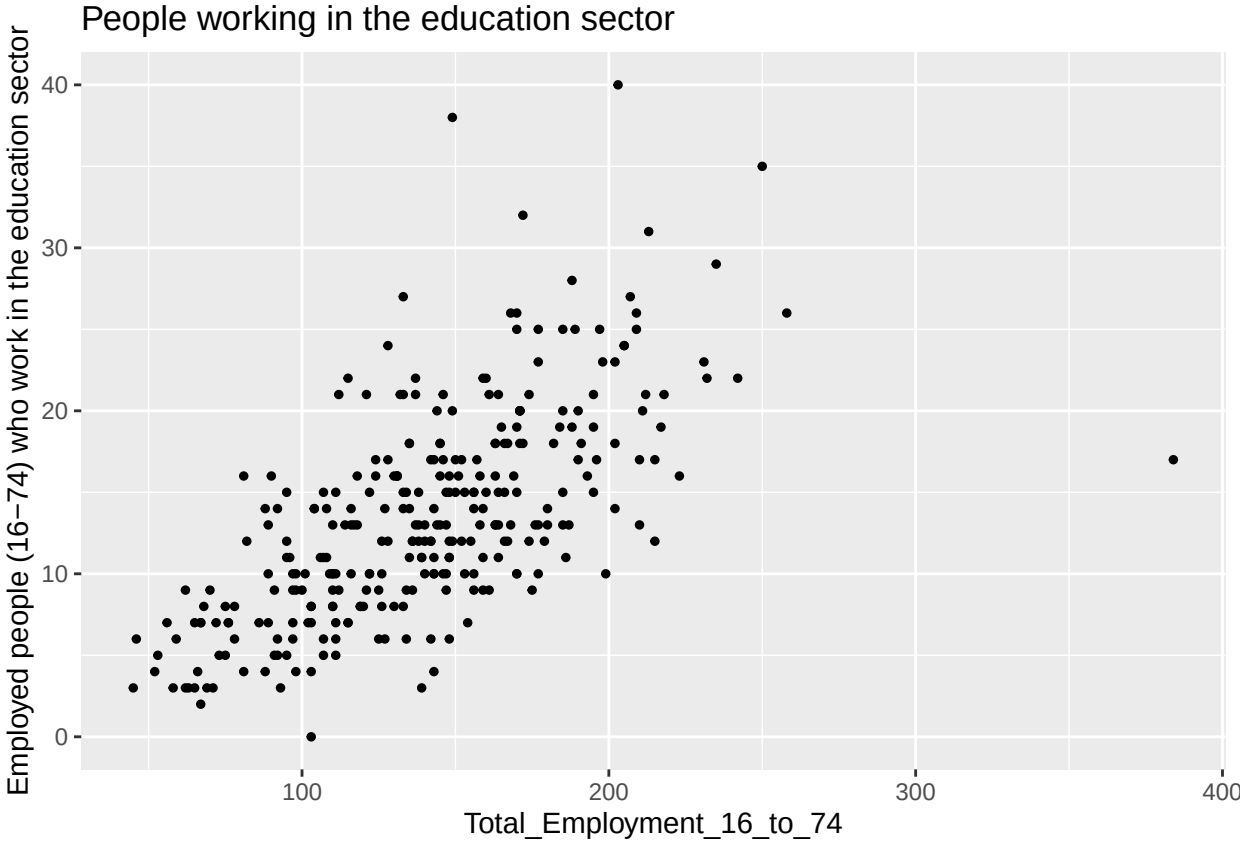



```
# Applying a Shapiro-Wilk test to the k059 variable
OAC_2011_LAD  %>%
  dplyr::pull(k059) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  .
## W = 0.965, p-value = 1.347e-06
```

```
# Plotting total employment counts against the variable k059
OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = Total_Employment_16_to_74,
      y = k059
    )
  ) +
  ggplot2::geom_point(size = 1) +
  ggplot2::ggtitle("People working in the education sector") +
  ggplot2::xlab("Total_Employment_16_to_74") +
  ggplot2::ylab(
    "Employed people (16-74) who work in the education sector"
  )
```



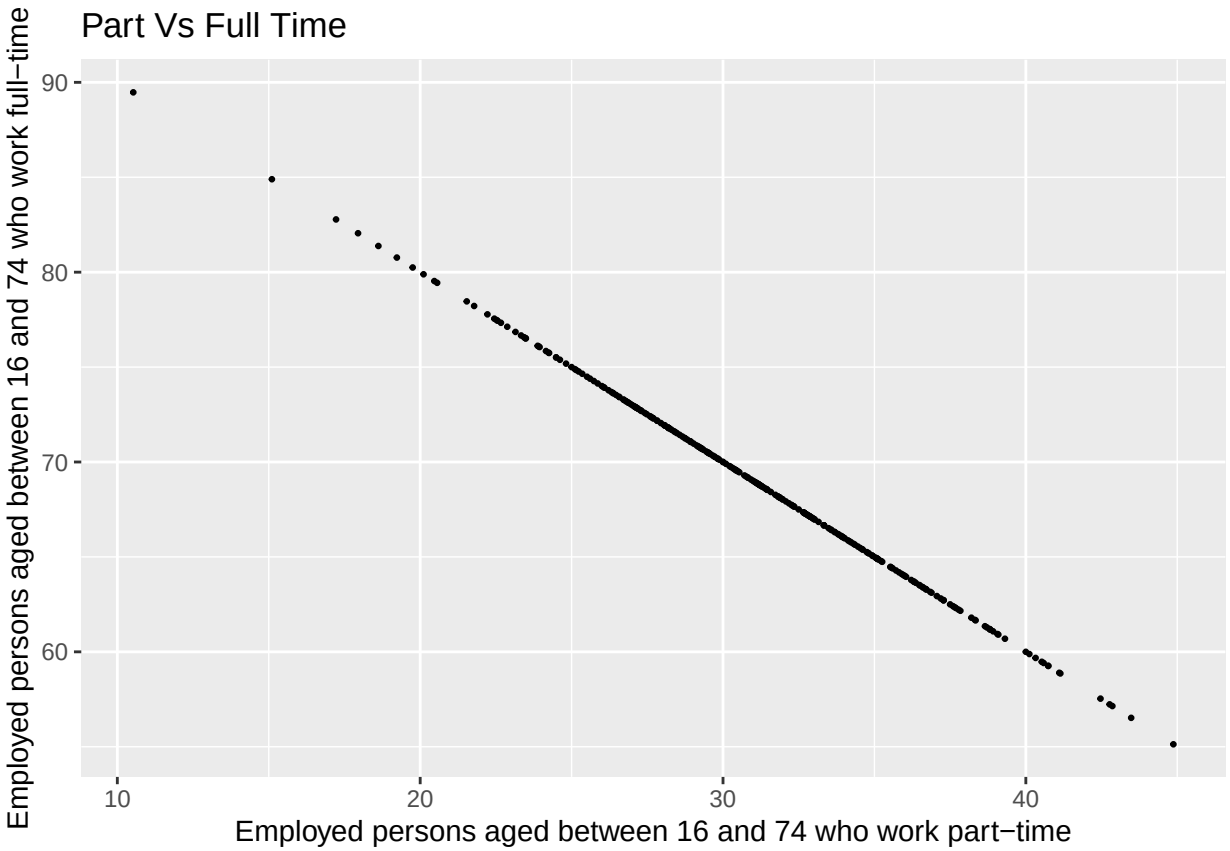
This data on employed persons aged between 16 and 74 who work in the education sector does not have significant normality from the Shapiro-Wilk test and skewed data. The counts of employed people who work in the education sector against the total employed people appears to have a positive trend.

1.5.2 Data Analysis

To perform data analysis with the selected variables, normalisation is applied to ensure all ranges are the same. The umemployed variable (k045) are treated differently as its variable statistical unit is the total count of populations (aged 16 to 74)

```
# normalising the selected variables
perc_OAC_2011_LAD <-
  # dataset focused in Cotswold
  OAC_2011_LAD %>%
  dplyr::mutate(
    # unemployed
    perc_unemployed = (k045 / Total_Population_16_to_74) * 100,
    # part time
    perc_part = (k046 / Total_Employment_16_to_74) * 100,
    # full time
    perc_full = (k047 / Total_Employment_16_to_74) * 100,
    # agriculture, forestry or fishing industries
    perc_agri_ind = ((k048)/ Total_Employment_16_to_74) * 100,
    # mining, quarrying or construction industries
    perc_mini_ind = ((k049)/ Total_Employment_16_to_74) * 100,
    # manufacturing industry
    perc_manu_ind = ((k050)/ Total_Employment_16_to_74) * 100,
    # education sector
    perc_edu = (k059 / Total_Employment_16_to_74) * 100
  ) %>%
  dplyr::select(
    OA, perc_unemployed, perc_part, perc_full, perc_agri_ind,perc_mini_ind,
    perc_manu_ind, perc_edu, Total_Population_16_to_74, Total_Employment_16_to_74
  )
```

```
# plotting normalised full time vs part time
perc_OAC_2011_LAD %>%
  ggplot2::ggplot(
    aes(
      x = perc_part,
      y = perc_full
    )
  ) +
  ggplot2::geom_point(size = 0.5) +
  ggplot2::ggtitle("Part Vs Full Time") +
  ggplot2::xlab("Employed persons aged between 16 and 74 who work part-time") +
  ggplot2::ylab("Employed persons aged between 16 and 74 who work full-time")
```



The above plot displays the relationship between employed people who work full time vs part time having a negative correlation.

```
# generating a descriptive statistics for all variables involved
perc_OAC_2011_LAD %>%
  dplyr::select(
    perc_unemployed,
    perc_part,
    perc_full,
    perc_agri_ind,
    perc_mini_ind,
    perc_manu_ind,
    perc_edu
    ) %>%
  pastecs::stat.desc(basic = FALSE, desc = FALSE, norm = TRUE) %>%
  knitr::kable()
```

	perc_unemployed	perc_part	perc_full	perc_agri_ind	perc_mini_ind	perc_manu_ind	perc_edu
skewness	1.1216162	-0.2510575	0.2510575	1.6563111	0.3749887	0.5257389	0.6989383
skew.2SE	3.9654670	-0.8876122	0.8876122	5.8558774	1.3257703	1.8587467	2.4710918
kurtosis	1.8521475	0.4633599	0.4633599	3.1983027	0.3949318	0.0172893	1.3236897
kurt.2SE	3.2848746	0.8217915	0.8217915	5.6723469	0.7004309	0.0306635	2.3476287
normtest.W	0.9295332	0.9942294	0.9942294	0.8198092	0.9861126	0.9787220	0.9735634
normtest.p	0.0000000	0.3214299	0.3214299	0.0000000	0.0057991	0.0002115	0.0000274

The results from this stats test for the normalised transformed variables confirms the previously mentioned the perc_full and perc_part are the only variables that passes the threshold value, therefore having significant normality. The other variable perc_mini_ind is the other variable that comes close to significance of 0.0057991 and the rest of the variables are clearly skewed.

```
# generating a descriptive statistics for all variables involved
# but applying an inverse hyberbolic sine to check for significances and
# normaility
perc_OAC_2011_LAD %>%
  dplyr::select(
    perc_unemployed,
    perc_part,
    perc_full,
    perc_agri_ind,
    perc_mini_ind,
    perc_manu_ind,
    perc_edu
    ) %>%
  # using transmute to only include the transformed variables
  dplyr::transmute(
    ihs_perc_unemployed = asinh(perc_unemployed),
    ihs_perc_agri_ind = asinh(perc_agri_ind),
    ihs_perc_mini_ind = asinh(perc_mini_ind),
    ihs_perc_manu_ind = asinh(perc_manu_ind),
    ihs_perc_edu = asinh(perc_edu)
    )%>%
  pastecs::stat.desc(basic = FALSE, desc = FALSE, norm = TRUE) %>%
  knitr::kable()
```

	ihs_perc_unemployed	ihs_perc_agri_ind	ihs_perc_mini_ind	ihs_perc_manu_ind	ihs_perc_edu
skewness	-0.4680663	0.0827402	-1.581264	-0.8159525	-1.4819676
skew.2SE	-1.6548455	0.2925273	-5.590549	-2.8847948	-5.2394872
kurtosis	-0.0173013	-1.1579338	5.026327	1.4288624	7.9576899
kurt.2SE	-0.0306847	-2.0536524	8.914438	2.5341576	14.1133539
normtest.W	0.9686727	0.9408481	0.901114	0.9639849	0.9251219
normtest.p	0.0000047	0.0000000	0.000000	0.0000010	0.0000000

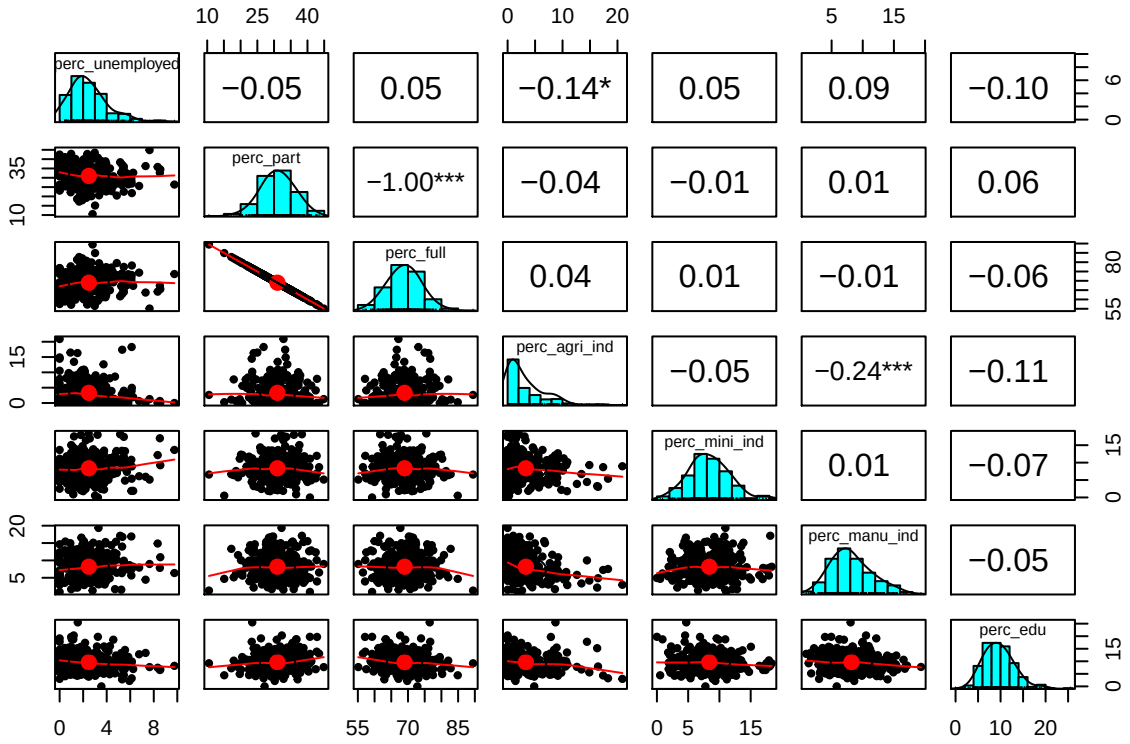
From this table above transforming the variables using the inverse hyperbolic sine still does not result in normally distributed variables. Thus, we should eliminate Pearson's r as an test to explore any correlations between the selected variables.

```
# this code chunk checks the number of ties for each
# normalised variables
ties_perc_part <-
  perc_OAC_2011_LAD %>%
  dplyr::count(perc_part) %>%
  dplyr::filter(n > 1) %>%
  # Specify wt = n() to count rows
  # otherwise n is taken as weight
  dplyr::count(wt = n) %>%
  dplyr::pull(n)

ties_perc_full <-
  perc_OAC_2011_LAD %>%
```

```
  dplyr::count(perc_full) %>%
  dplyr::filter(n > 1) %>%
  dplyr::count(wt = n) %>%
  dplyr::pull(n)

ties_perc_agri_ind <-
  perc_OAC_2011_LAD %>%
  dplyr::count(perc_agri_ind) %>%
  dplyr::filter(n > 1) %>%
  dplyr::count(wt = n) %>%
  dplyr::pull(n)

ties_perc_mini_ind <-
  perc_OAC_2011_LAD %>%
  dplyr::count(perc_mini_ind) %>%
  dplyr::filter(n > 1) %>%
  dplyr::count(wt = n) %>%
  dplyr::pull(n)

ties_perc_manu_ind <-
  perc_OAC_2011_LAD %>%
  dplyr::count(perc_manu_ind) %>%
  dplyr::filter(n > 1) %>%
  dplyr::count(wt = n) %>%
  dplyr::pull(n)

ties_perc_edu <-
  perc_OAC_2011_LAD %>%
  dplyr::count(perc_edu) %>%
  dplyr::filter(n > 1) %>%
  dplyr::count(wt = n) %>%
  dplyr::pull(n)
```

Concluded that all variables have many ties, with the ties ranging from 44 to 117, therefore it would not be appropriate to use Spearman's rho and Kendall's tau is advised.

```
# applying kendall test in a pairs.panels format
perc_OAC_2011_LAD %>%
  dplyr::select(perc_unemployed, perc_part, perc_full, perc_agri_ind,
                perc_mini_ind, perc_manu_ind, perc_edu
              ) %>%
  psych::pairs.panels(
    method = "kendall",
    stars = TRUE
  )
```



From this overall kendall tau test, it can be seen that perc_manu_ind and perc_umployed has the highest positive relationship yet still weak. As it has a variance of 1.8% and the relationship is significant. From inspection of the histogram plots, perc_mini_ind has the best indication for significant normality. The "highest" negative relationship is with the perc_mini_ind and perc_manu_ind with the variance of 5.76 %, but the relationship is not the strongest.

1.5.3 Question B.2

In this section a geodemographic classification are created for the OAs in Cotswold using the variables listed below from the 2011_OAC_Raw_kVariables.csv dataset. The variables are:

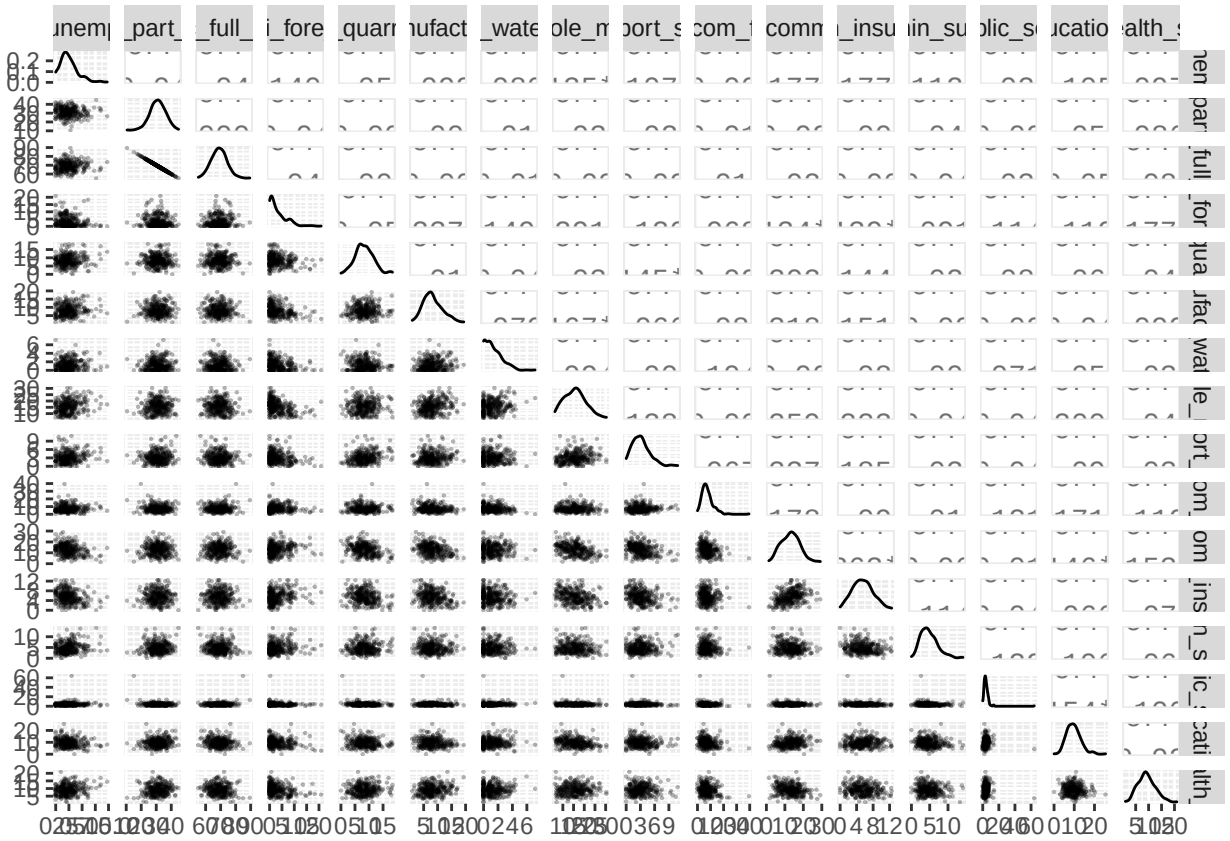
- k045 (People aged between 16 and 74 who are unemployed)
- k046 (Employed persons aged between 16 and 74 who work part-time)
- k047 (Employed persons aged between 16 and 74 who work part-time)
- k048 (Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries)
- k049 (Employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries)
- k050 (Employed persons aged between 16 and 74 who work in the manufacturing industry)
- k051 (Employed persons aged between 16 and 74 who work in the energy, water or air conditioning supply industries)
- k052 (Employed persons aged between 16 and 74 who work in the wholesale and retail trade; repair of motor vehicles and motor cycles industries)
- k053 (Employed persons aged between 16 and 74 who work in the transport or storage industries)
- k054 (Employed persons aged between 16 and 74 who work in the accommodation or food service activities industries)
- k055 (Employed persons aged between 16 and 74 who work in the information and communication or professional, scientific and technical activities industries)
- k056 (Employed persons aged between 16 and 74 who work in the financial, insurance or real estate industries)
- k057 (Employed persons aged between 16 and 74 who work in the administrative or support service activities industries)
- k058 (Employed persons aged between 16 and 74 who work in the in public administration or defence; compulsory social security industries)
- k059 (Employed persons aged between 16 and 74 who work in the education sector)
- k060 (Employed persons aged between 16 and 74 who work in the human health and social work activities industries)

```
# Normalising all the variables and assigning it to the
# employed_people variable
employed_people <-
  OAC_2011_LAD %>%
  dplyr::select(
    OA, Total_Population_16_to_74,
    Total_Employment_16_to_74,
    k045:k060
  ) %>%
  # normalising variables
  # k045 is treated separately
  dplyr::mutate(
    k045 = (k045 / Total_Population_16_to_74) * 100,
    # scale across
    dplyr::across(
      k046:k060,
      #scale
      function(x){ (x / Total_Employment_16_to_74) * 100 }
    )
  ) %>%
  # renaming the variables for better interpretation
  dplyr::rename(
    unemployed = k045,
```

```
    part_time = k046,
    full_time = k047,
    # agriculture, forestry or fishing industries
    agri_fore_fish_ind = k048,
    # mining, quarrying or construction industries
    mini_quarr_cons_ind = k049,
    # manufacturing industry
    manufacturing_ind = k050,
    # energy, water or air conditioning supply industries
    ener_water_airc_ind = k051,
    # wholesale and retail trade; repair of motor vehicles and motor cycles industries
    whole_motor_ind = k052,
    # transport or storage industries
    transport_storage_ind = k053,
    # accommodation or food service activities industries
    accom_food_ind = k054,
    # information and communication or professional,
    # scientific and technical activities industries
    info_comm_tech_ind = k055,
    # financial, insurance or real estate industries
    finan_insur_real_ind = k056,
    # administrative or support service activities industries
    admin_support_ind = k057,
    # public administration or defence; compulsory social security industries
    public_social_ind = k058,
    education_sector = k059,
    # human health and social work activities industries
    health_swa_ind = k060
  ) %>%
  # rename columns
  # function appends perc_ to each variable
  dplyr::rename_with(
    function(x){
      paste0("perc_", x)
      },
    unemployed:health_swa_ind
  )
```

To perform the analysis with the aforementioned variables, they all needed to be all normalised so all the variables are within the same ranges for optimum analysis. Also they are renamed for easier interpretation with the later statistical analysis.

```
# initial kendall test on all the variables
employed_people %>%
  # all the selected variables
  dplyr::select(perc_unemployed:perc_health_swa_ind) %>%
  GGally::ggpairs(
    upper = list(continuous = wrap(ggally_cor, method = "kendall")),
    lower = list(continuous = wrap("points", alpha = 0.3, size=0.1))
  )
```



From the above plot, due to many variables in the analysis it is incredibly difficult to interpret the data and any discernable relationships. Upon closer inspection in a new window, it is possible to interpret the data. public_social_ind is highly skewed compared to all the other variables, which explains that other factors do not affect the relationship of people working in the public administration or defence; compulsory social security industries.

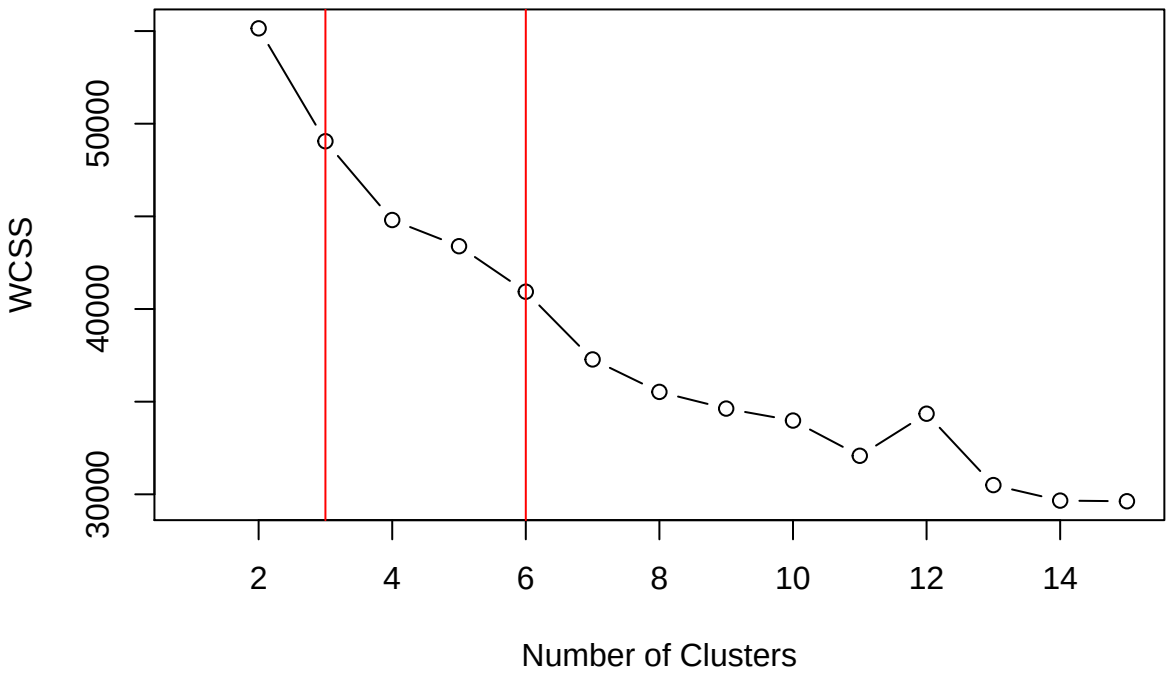
1.5.3.1 Elbow Method

```
# Data for elbow method
data_for_testing <-
  employed_people %>%
  dplyr::select(perc_unemployed:perc_health_swa_ind)

# Calculate WCSS and silhouette
# for k = 2 to 15
# Set up two vectors where to store
# the calculated WCSS and silhouette value
testing_wcss <- rep(NA, 15)
testing_silhouette <- rep(NA, 15)

# for k = 2 to 15
for (testing_k in 2:15){
  # Calculate kmeans
  kmeans_result <-
    stats::kmeans(data_for_testing, centers = testing_k, iter.max = 50)
```

```
  # Extract WCSS
  # and save it in the vector
  testing_wcss[testing_k] <- kmeans_result %$% tot.withinss

  # Calculate average silhouette
  # and save it in the vector
  testing_silhouette[testing_k] <-
    kmeans_result %$% cluster %>%
    cluster::silhouette(
      data_for_testing %>% dist()
    ) %>%
    magrittr::extract(, 3) %>% mean()
}

# Calculate the gap statistic using bootstrapping
testing_gap <-
  cluster::clusGap(data_for_testing, FUN = kmeans,
    K.max = 15, B = 50
  )
```

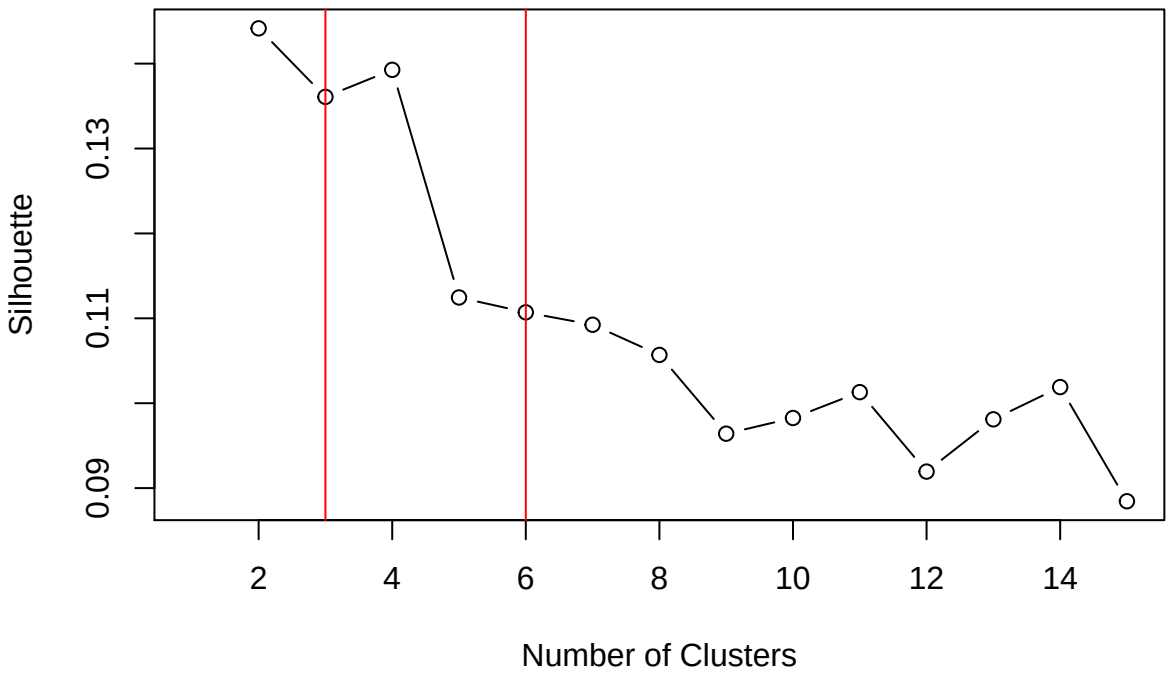
1.5.3.2 Plots

```
# Plots
# Checking the number of clusters associating to WCSS
plot(2:15, testing_wcss[2:15], type="b", xlab="Number of Clusters",
     ylab="WCSS", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```



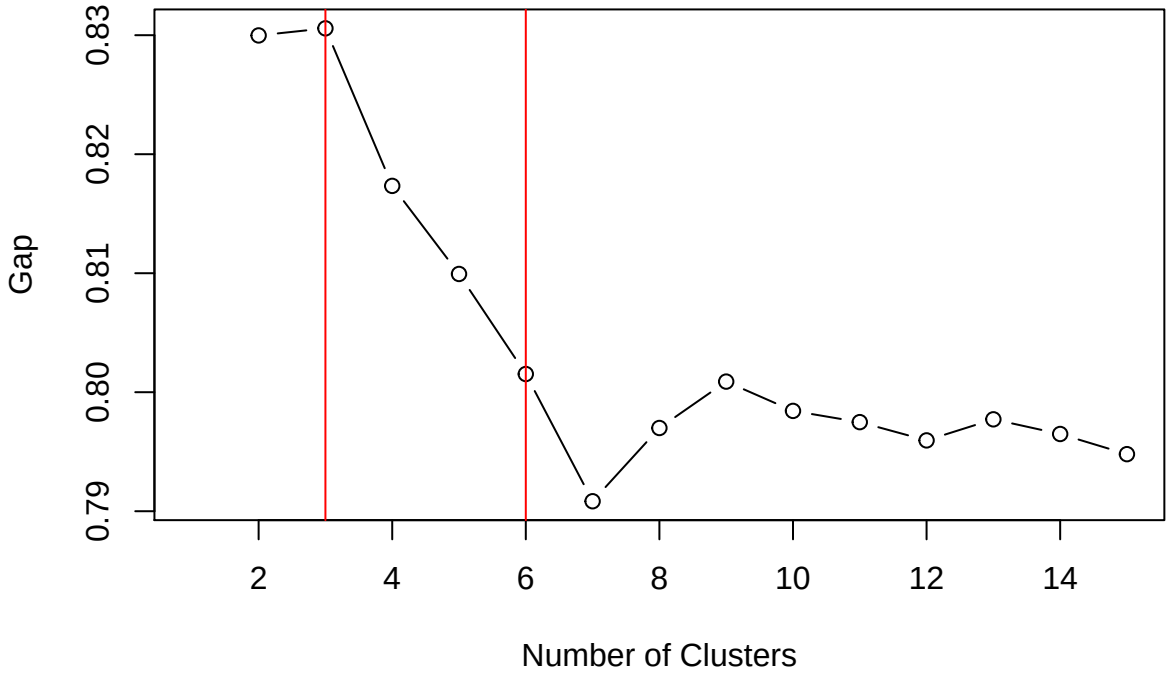

```
## integer(0)
```

```
# Plots
# Checking the number of clusters associating to silhouette measure
plot(2:15, testing_silhouette[2:15], type="b", xlab="Number of Clusters",
     ylab="Silhouette", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```




```
## integer(0)
```

```
# Plots
# Checking the number of clusters associating to gap statistics
plot(2:15, testing_gap[["Tab"]][2:15, "gap"], type="b", xlab="Number of Clusters",
     ylab="Gap", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```



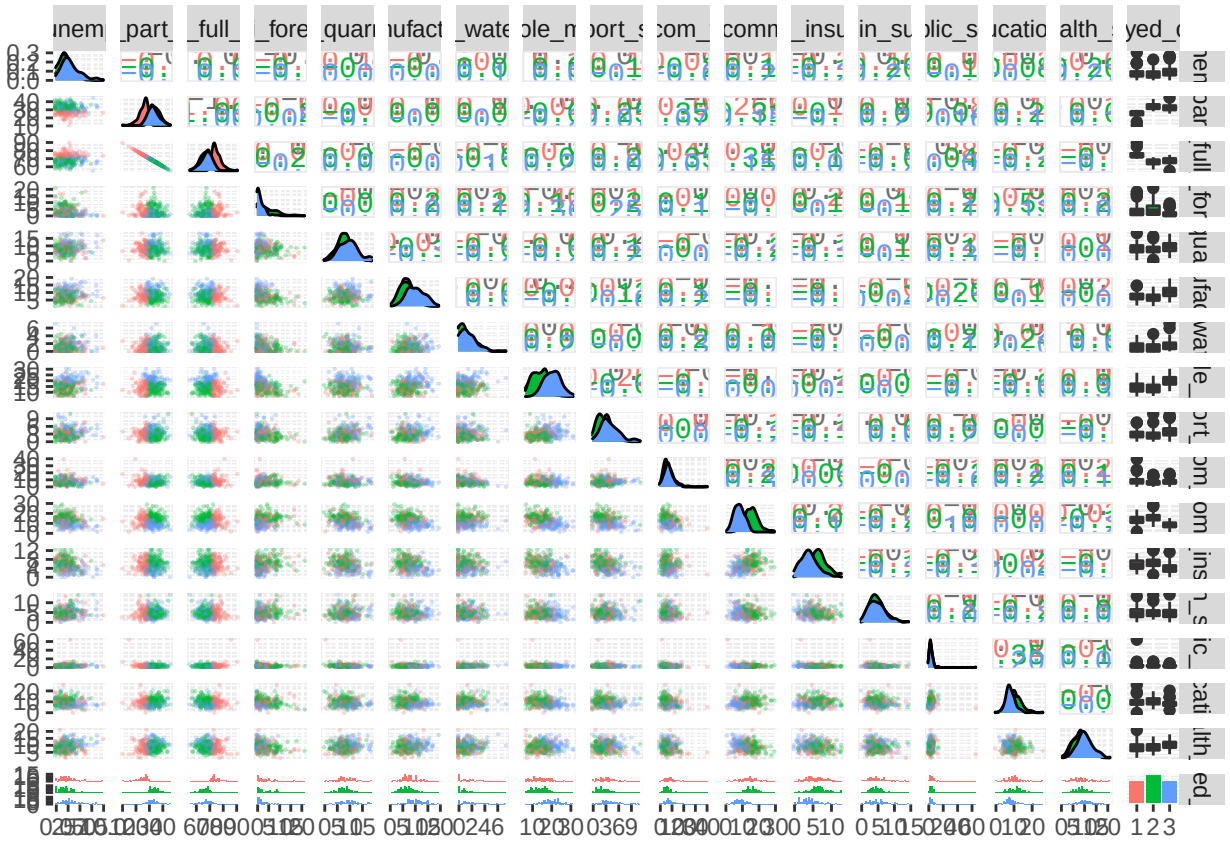

```
## integer(0)
```

From the above plots, the WCSS plot displays that 6 clusters may be optimum. But the Silhouette and Gap plot of 6 clusters is not appropriate as the statistical result are too low. Therefore to counteract the differences, 3 clusters is determined for the kmeans calculations.

```
# calculating the kmeans with 3 clusters
employed_kmeans <- employed_people %>%
  dplyr::select(perc_unemployed:perc_health_swa_ind) %>%
  stats::kmeans(
    centers = 3,
    iter.max = 50
  )

employed_people <-
  employed_people %>%
  tibble::add_column(
    employed_cluster =
      employed_kmeans %$%
        cluster %>%
          as.character()
  )
```

```
# analysing the computed results through a visual analysis
employed_people %>%
  dplyr::select(perc_unemployed:perc_health_swa_ind, employed_cluster) %>%
  GGally::ggpairs(
    mapping = aes(color = employed_cluster),
```

```
    lower = list(continuous = wrap("points", alpha = 0.3, size=0.1))
  )
```

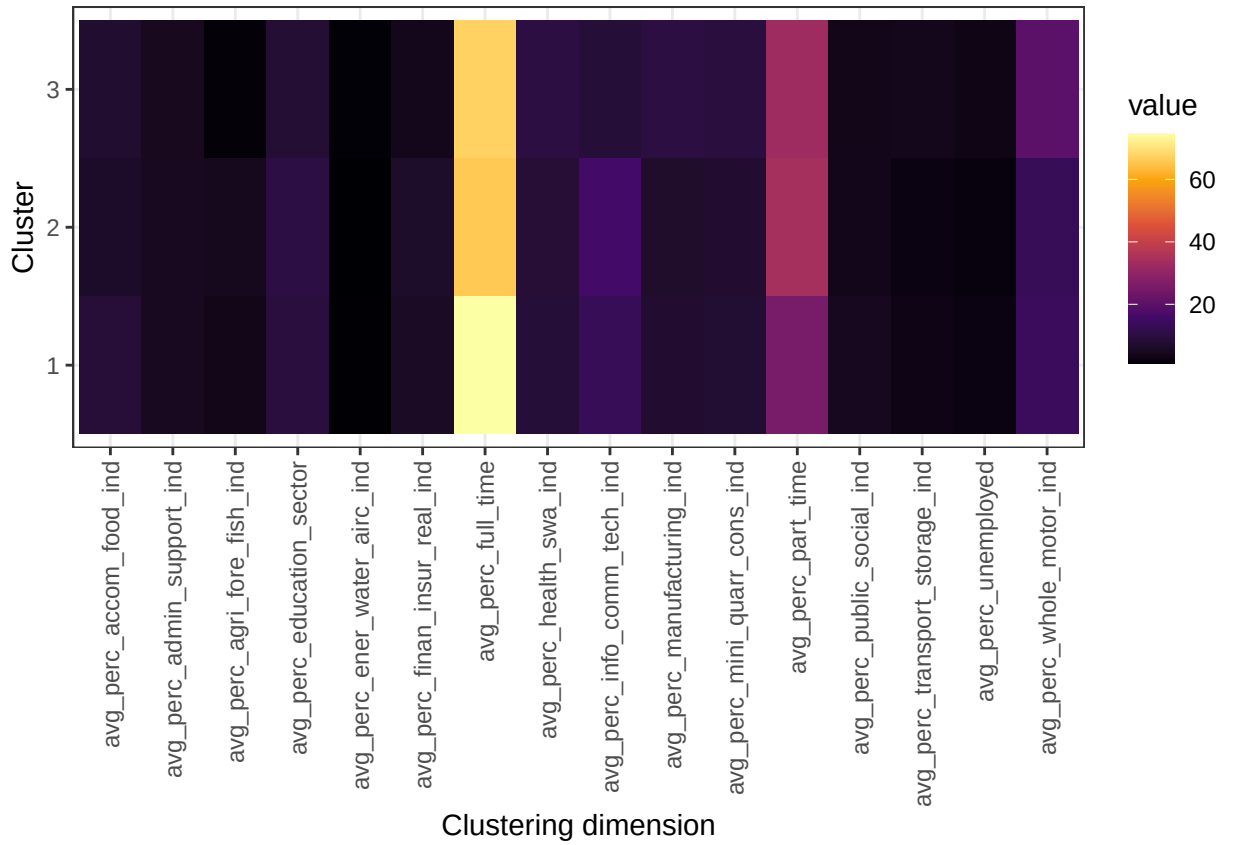


The above plot is near discernable. When opened in a new window, it can be better interpreted and the chosen number of 3 clusters appear to closely classify the data. But some data do appear to overlap with each other in many of the variables plotted against each other.

1.5.3.3 Heatmap

```
# Creating the heatmaps for the average values of the variables used
# in the clustering process for each cluster.
employed_cluster_avgs <-
  employed_people %>%
  group_by(employed_cluster) %>%
  dplyr::summarise(
    dplyr::across(
      perc_unemployed:perc_health_swa_ind,
      mean
    )
  ) %>%
  # rename columns
  dplyr::rename_with(
    function(x){ paste0("avg_", x) },
    perc_unemployed:perc_health_swa_ind
  )
```

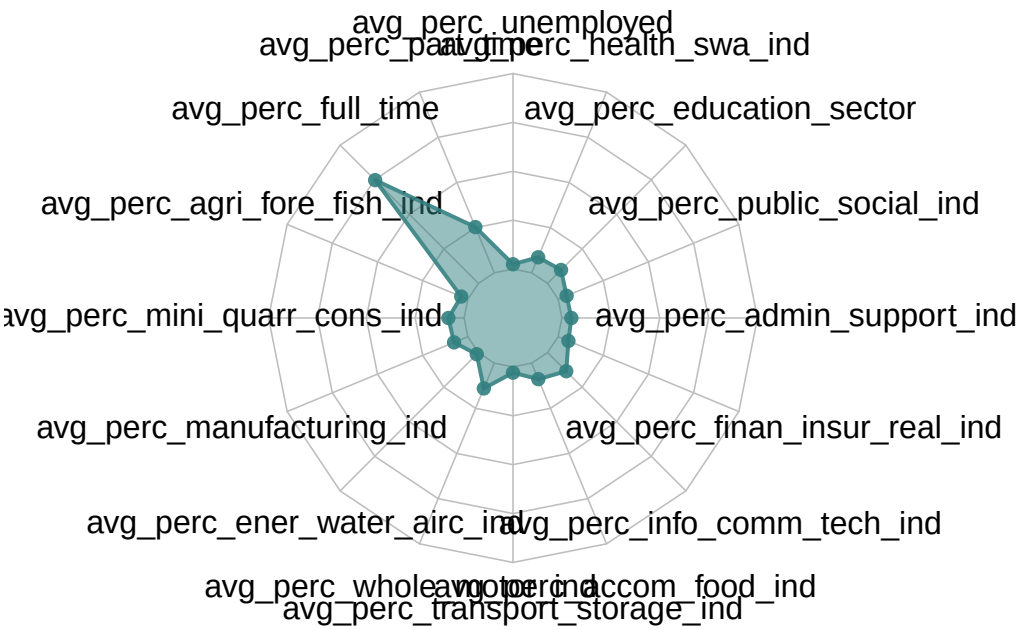
```r
employed_cluster_avgs %>%
  tidyr::pivot_longer(
    cols = -employed_cluster,
    names_to = "clustering_dimension",
    values_to = "value"
  )  %>%
  ggplot2::ggplot(
    aes(
      x = clustering_dimension,
      y = employed_cluster
    )
  ) +
  ggplot2::geom_tile(
    aes(
      fill = value
    )
  ) +
  ggplot2::xlab("Clustering dimension") +
  ggplot2::ylab("Cluster") +
  ggplot2::scale_fill_viridis_c(option = "inferno") +
  ggplot2::theme_bw() +
  ggplot2::theme(
    axis.text.x =
      element_text(
        angle = 90,
        vjust = 0.5,
        hjust=1
      )
    )
```



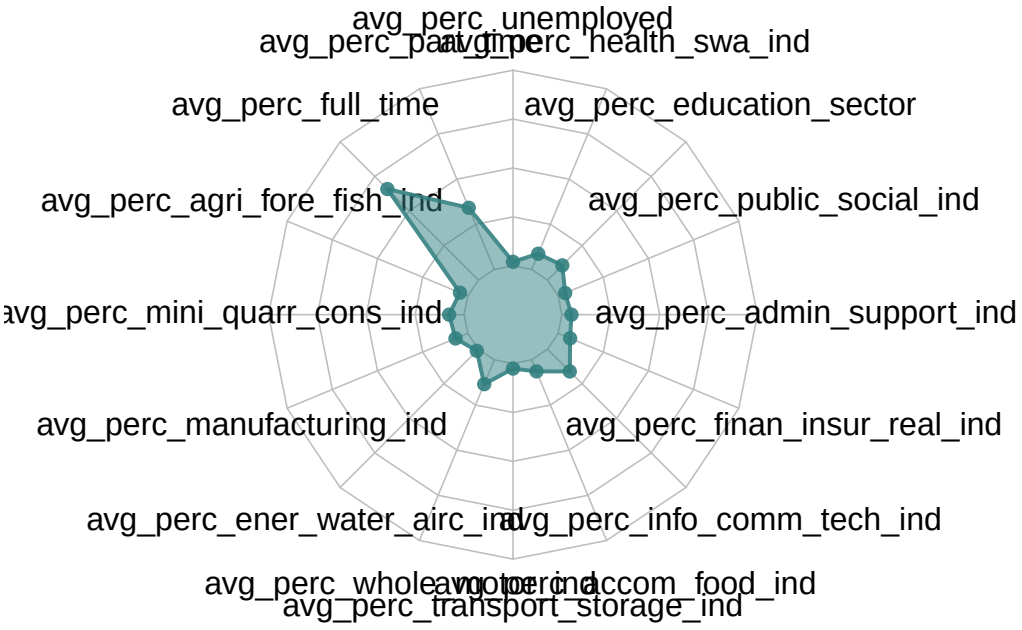
From the above heatmap the avg_perc_full_time has high values for 2 clusters and oppositely for avg_perc_part_time with the first and third clusters. Interestingly the avg_perc_whole_motor_ind has the highest avg value relationship with the third cluster and all other variables have low avg values associating to each cluster.

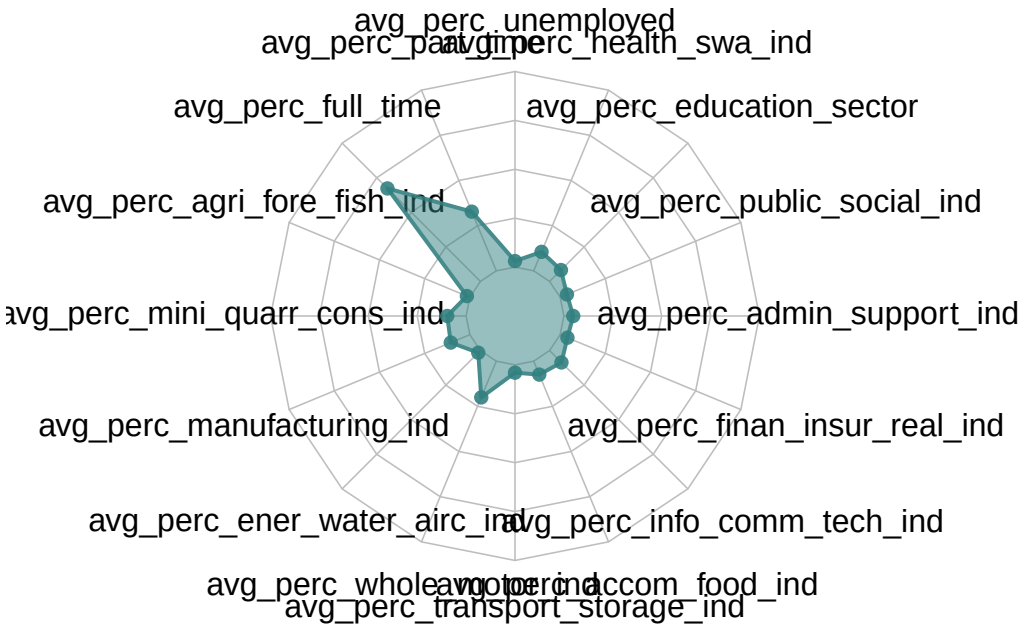
```
par(mar=rep(3,4))
par(mfrow=c(1,1))

for(cluster_number in 1:3){
  rbind (
    # The radar chart requires a maximum and a minimum row
    # before the actual data
    rep(100, 16), # max 100% for 16 variables
    rep(0, 16),   # min 0% for 16 variables
    employed_cluster_avgs %>%
      dplyr::filter(employed_cluster == cluster_number) %>%
      dplyr::select(-employed_cluster) %>%
      as.data.frame()
    ) %>%
    fmsb::radarchart(
      title = paste("Cluster", cluster_number)
      # affecting the polygon
      ,cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,20,5), cglwd=0.8,
      # affecting the grid
      pcol=rgb(0.2,0.5,0.5,0.9) , pfcol=rgb(0.2,0.5,0.5,0.5) , plwd=2
      )
}
```

Cluster 1



Cluster 2





Above are the radarplots for the 3 calculated kmeans clusters and it is clear that the full time and part time variable are distinctive having higher values compared to the rest in all 3 clusters. This type of visualisation allowed us to detect the financial variable (avg_perc_finan_insur_read_ind) having noticeable changes in clusters 1 and 2. Similarly for the manufacturing variable (avg_perc_manufacturing_ind) having higher values in all three clusters.