209041841_GY7702_CW2

209041841

04/01/2021

GY7702-R-for-Data-Science

Datascience-Project

##*The University of Leicester* **Coursework 2** The link to the GitHub Repository

Option A

Option A.1

Exploratory Data Analysis

```
rm(list=ls())       # To clear environment
library(tidyverse)
library(knitr)
library(pastecs)
library(magrittr)
```

```
# Read OAC_2011 data
OAC_2011 <-
  readr::read_csv("Data/2011_OAC_Raw_kVariables.csv")
```

```
# Creating new table for assigned LAD - Wolverhampton

Wolverhampton_LAD <-
  readr::read_csv("Data/OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv") %>%
  dplyr::filter(LAD11CD == "E08000031") %>%
  dplyr::select(-LAD11NMW) %>%
  readr::write_csv("Data/Wolverhampton_LAD.csv")
```

```
# Read_LAD data

Wolverhampton_LAD <-
  readr::read_csv("Data/Wolverhampton_LAD.csv")
```

```r
# Tibble joining

OAC_2011 %>%

dplyr::inner_join(

  Wolverhampton_LAD,
  by = c("OA" = "OA11CD")
) %>%
  dplyr::select(OA, Total_Population, Total_Households, Total_Dwellings,
                Total_Household_Spaces, Total_Population_16_and_over,
                Total_Population_16_to_74, Total_Employment_16_to_74,
                Total_Pop_in_Housesholds_16_and_over,
                k004, k009, k010, k027, k031, k041, k046
  ) %>%
  readr::write_csv("Data/Wolverhampton_OAC2011.csv")
```

```r
# Read_Wolverhampton_data

Wolverhampton_2011OAC <-
  readr::read_csv("Data/Wolverhampton_OAC2011.csv")
```

Data Visualisation

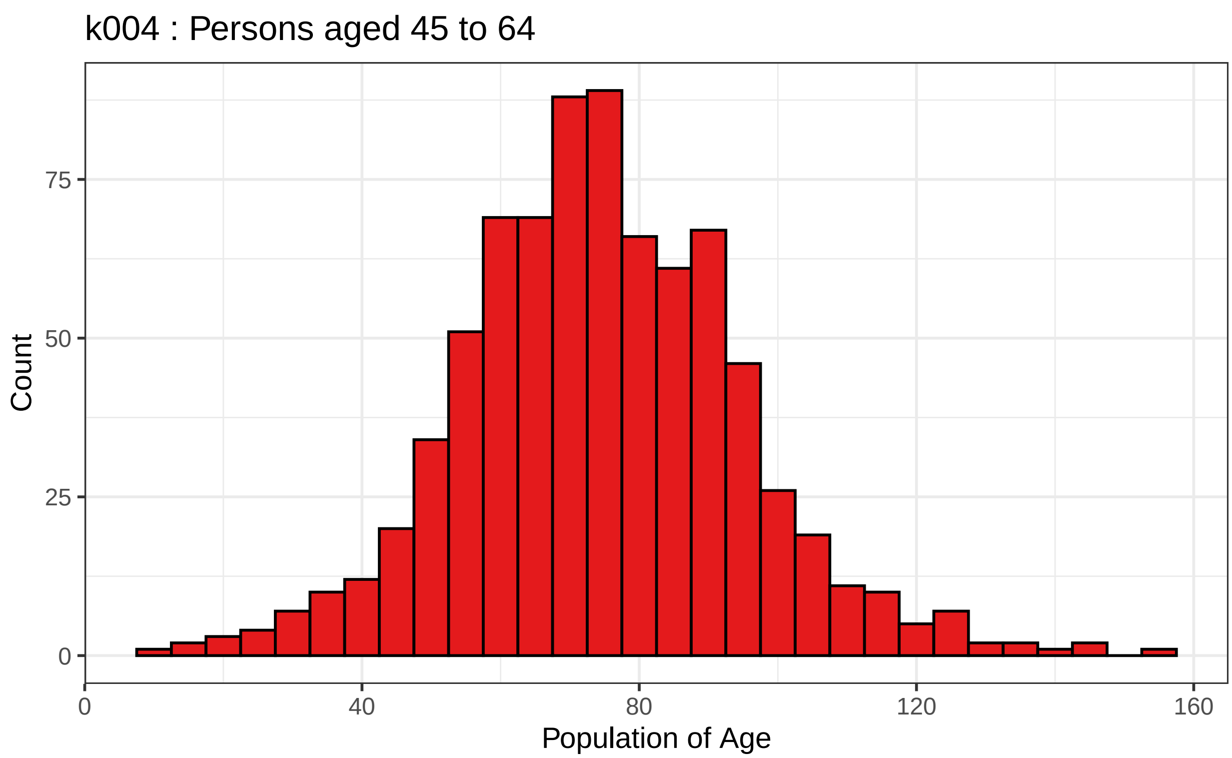
Distribution of variables

k004 - Persons aged 45 to 64

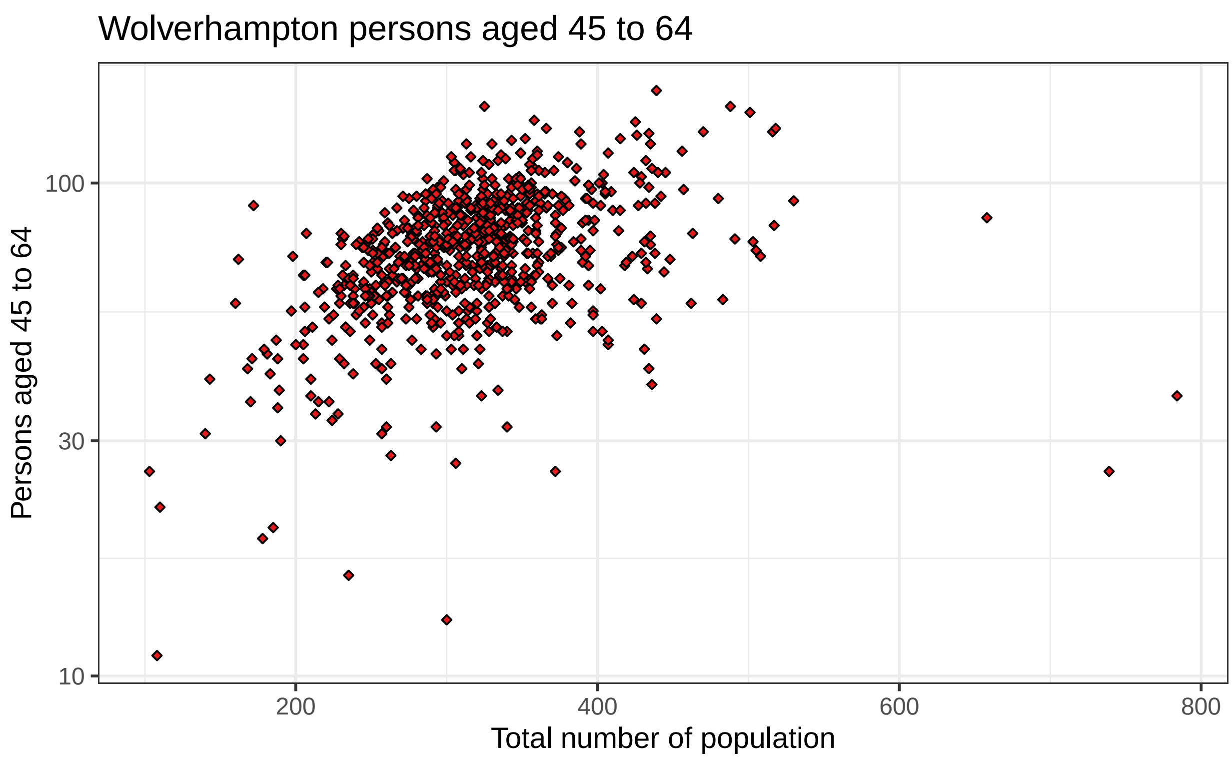
```r
summary(Wolverhampton_2011OAC$k004)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   62.00   74.00   74.59   88.00  154.00
```

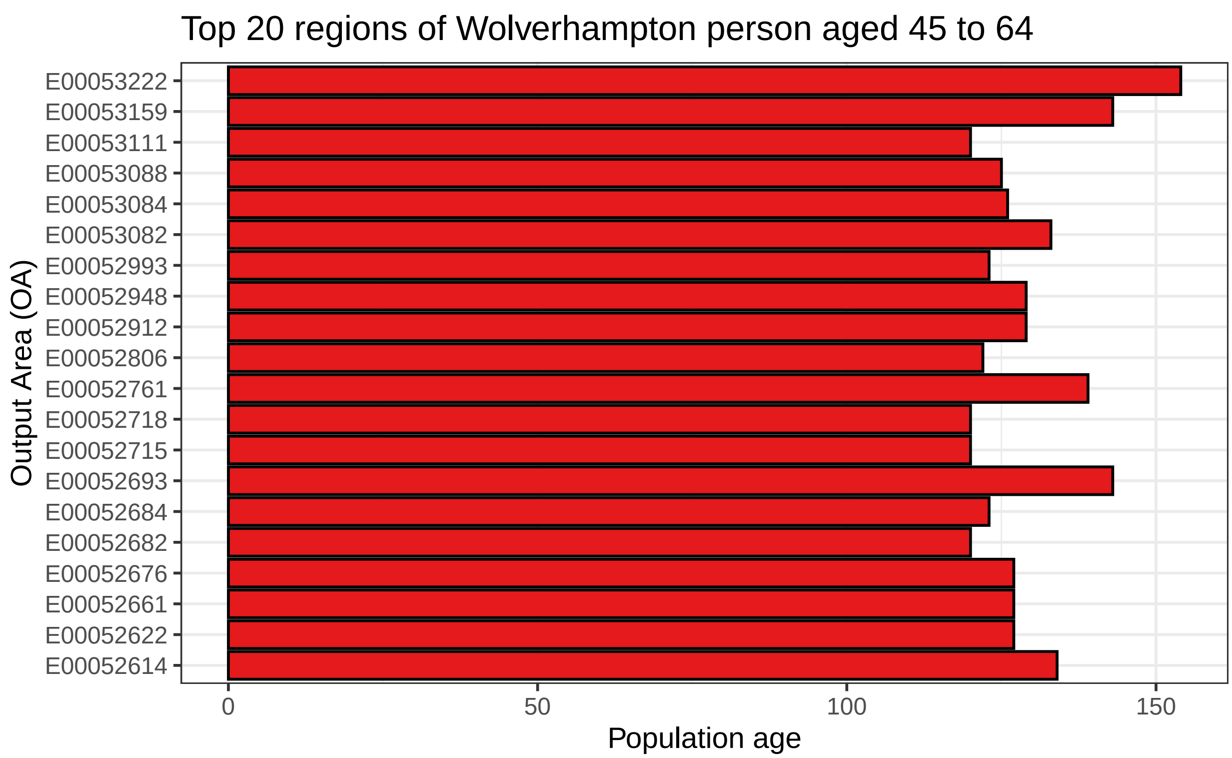
```r
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k004
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#e41a1c", colour="black") +
  ggplot2::ggtitle("k004 : Persons aged 45 to 64") +
  ggplot2::xlab("Population of Age") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```




```
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Population,
      y = k004
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#e41a1c") +
  ggplot2::ggtitle("Wolverhampton persons aged 45 to 64") +
  ggplot2::xlab("Total number of population") +
  ggplot2::ylab("Persons aged 45 to 64") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```
# Top 20 regions of Wolverhampton person aged 45 to 64

k004_max <-
Wolverhampton_2011OAC %>%
dplyr::filter(k004>20) %>%
  dplyr::select(OA, k004) %>%
  dplyr::slice_max(k004, n=20)

ggplot2::ggplot(k004_max,
       aes(
         x = k004,
         y = OA,
         )
       )+
ggplot2::geom_bar(position = "stack", stat = "identity", fill="#e41a1c", colour="black") +
ggplot2::ggtitle("Top 20 regions of Wolverhampton person aged 45 to 64")+
ggplot2::xlab("Population age")+
ggplot2::ylab("Output Area (OA)")+
ggplot2::theme_bw()
```



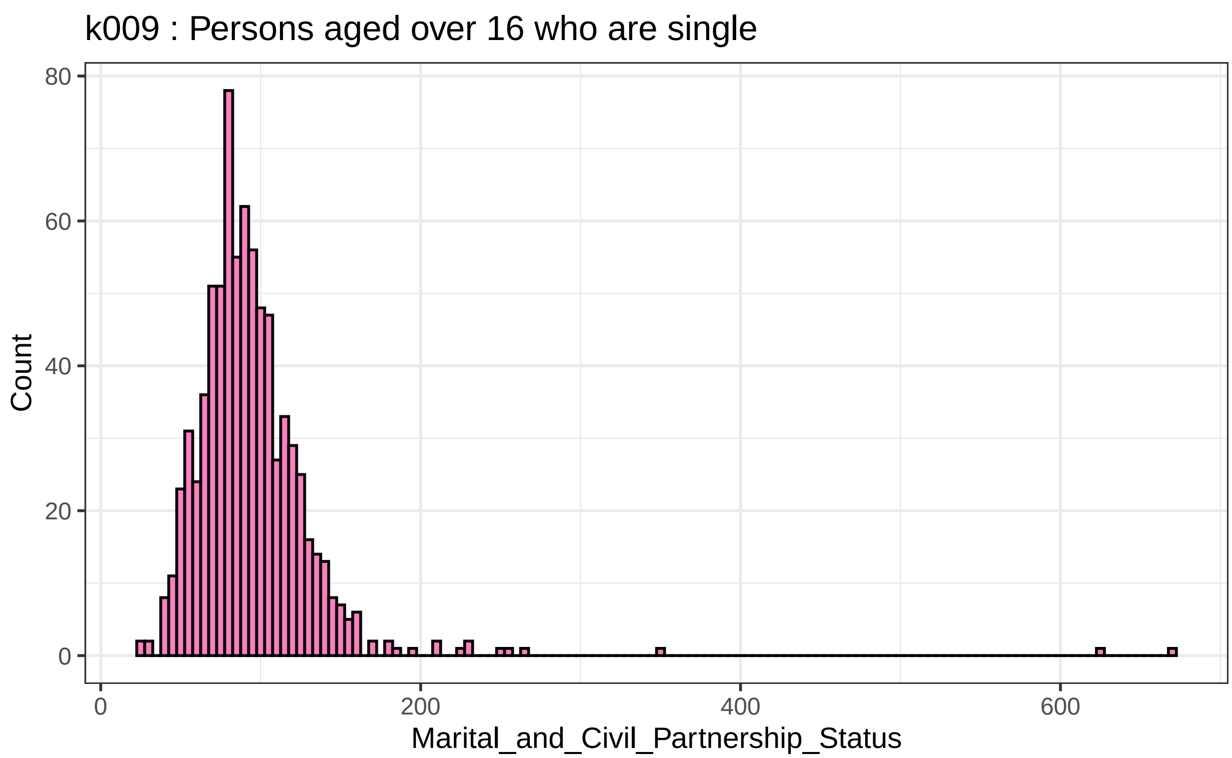

```r
# Bottom 20 regions of Wolverhampton person aged 45 to 64

k004_min <-
  Wolverhampton_2011OAC %>%
  dplyr::filter(k004>20) %>%
  dplyr::select(OA, k004) %>%
  dplyr::slice_min(k004, n=20)

ggplot2::ggplot(k004_min,
                aes(
                  x = k004,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#e41a1c", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton person aged 45 to 64")+
  ggplot2::xlab("Population age")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```

k009-Persons aged over 16 who are single

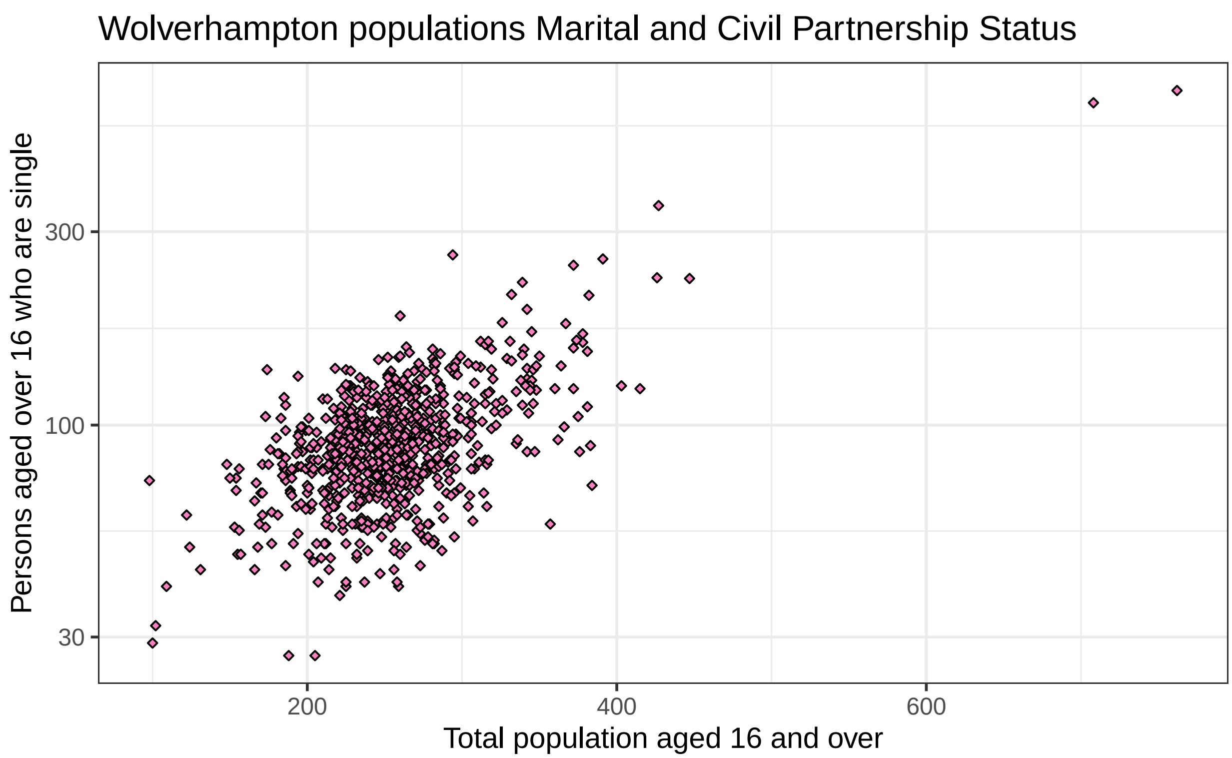
```
summary(Wolverhampton_2011OAC$k009)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   73.00   89.00   94.56  108.00  669.00
```

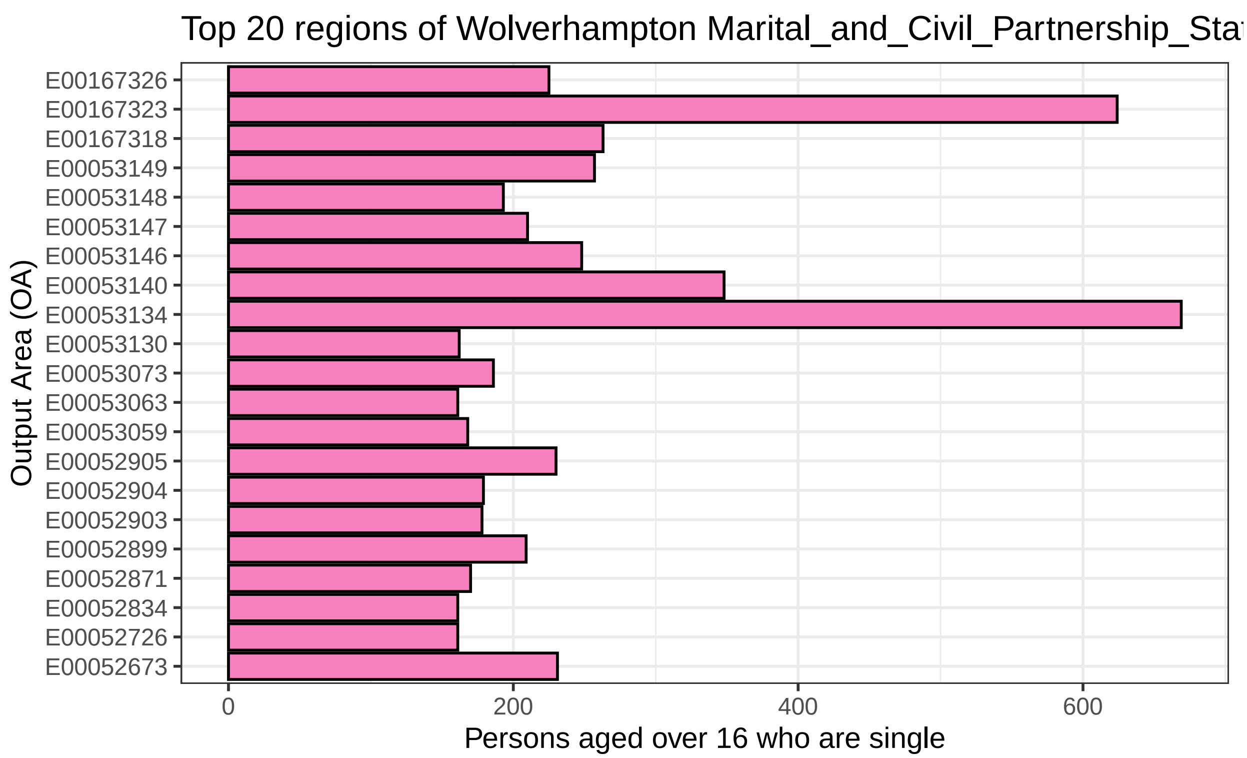
```
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k009
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#f781bf", colour="black") +
  ggplot2::ggtitle("k009 : Persons aged over 16 who are single") +
  ggplot2::xlab("Marital_and_Civil_Partnership_Status") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```



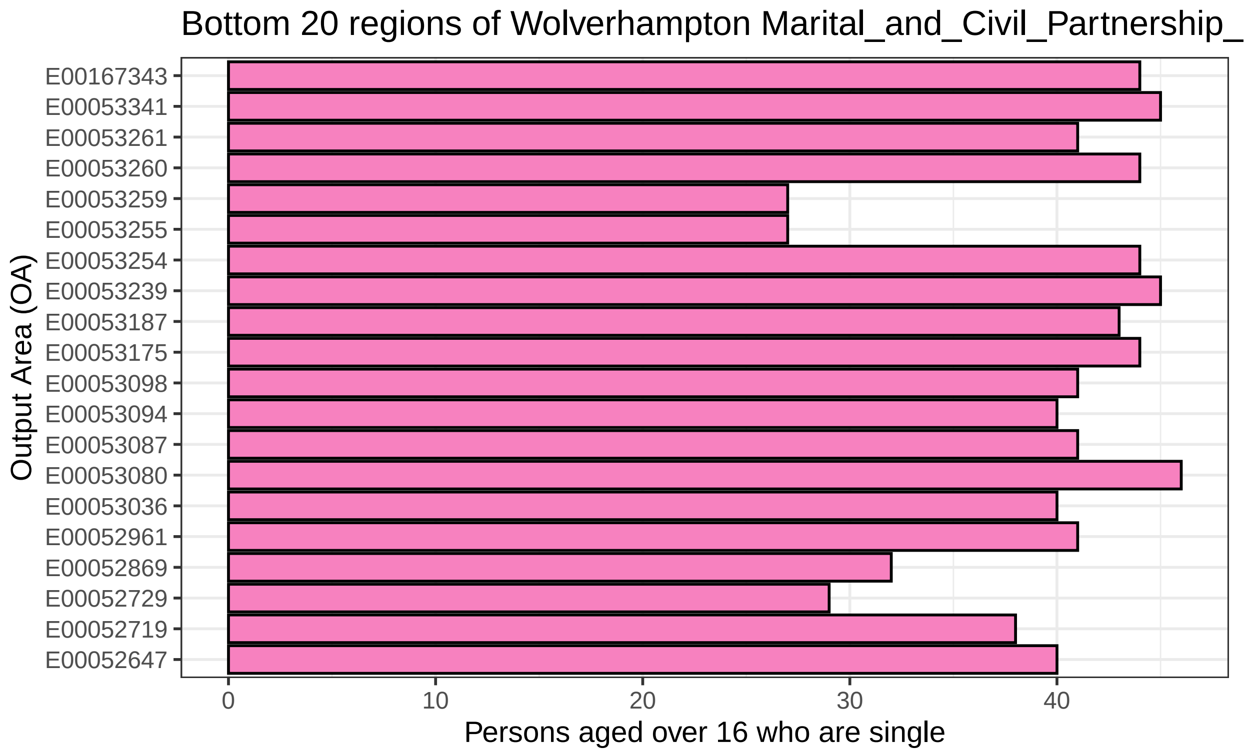

```r
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Population_16_and_over,
      y = k009
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#f781bf") +
  ggplot2::ggtitle("Wolverhampton populations Marital and Civil Partnership Status") +
  ggplot2::xlab("Total population aged 16 and over") +
  ggplot2::ylab("Persons aged over 16 who are single") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```
# Top 20 regions of Wolverhampton Marital_and_Civil_Partnership_Status

k009_max <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k009) %>%
  dplyr::slice_max(k009, n=20)

ggplot2::ggplot(k009_max,
                aes(
                  x = k009,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#f781bf", colour="black") +
  ggplot2::ggtitle("Top 20 regions of Wolverhampton Marital_and_Civil_Partnership_Status")+
  ggplot2::xlab("Persons aged over 16 who are single")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```




```r
# Bottom 20 regions of Wolverhampton Marital_and_Civil_Partnership_Status

k009_min <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k009) %>%
  dplyr::slice_min(k009, n=20)

ggplot2::ggplot(k009_min,
                aes(
                  x = k009,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#f781bf", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton Marital_and_Civil_Partnership_Status")+
  ggplot2::xlab("Persons aged over 16 who are single")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```

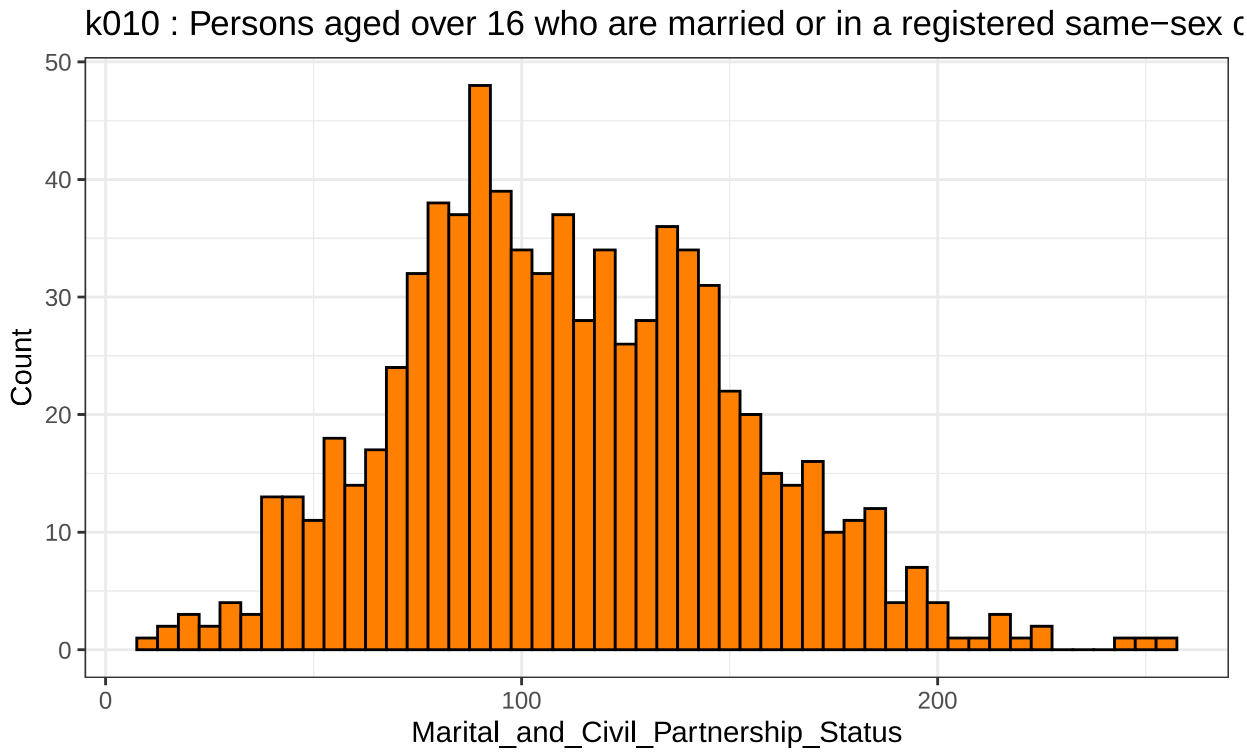


k010 - Persons aged over 16 who are married or in a registered same-sex civil partnership

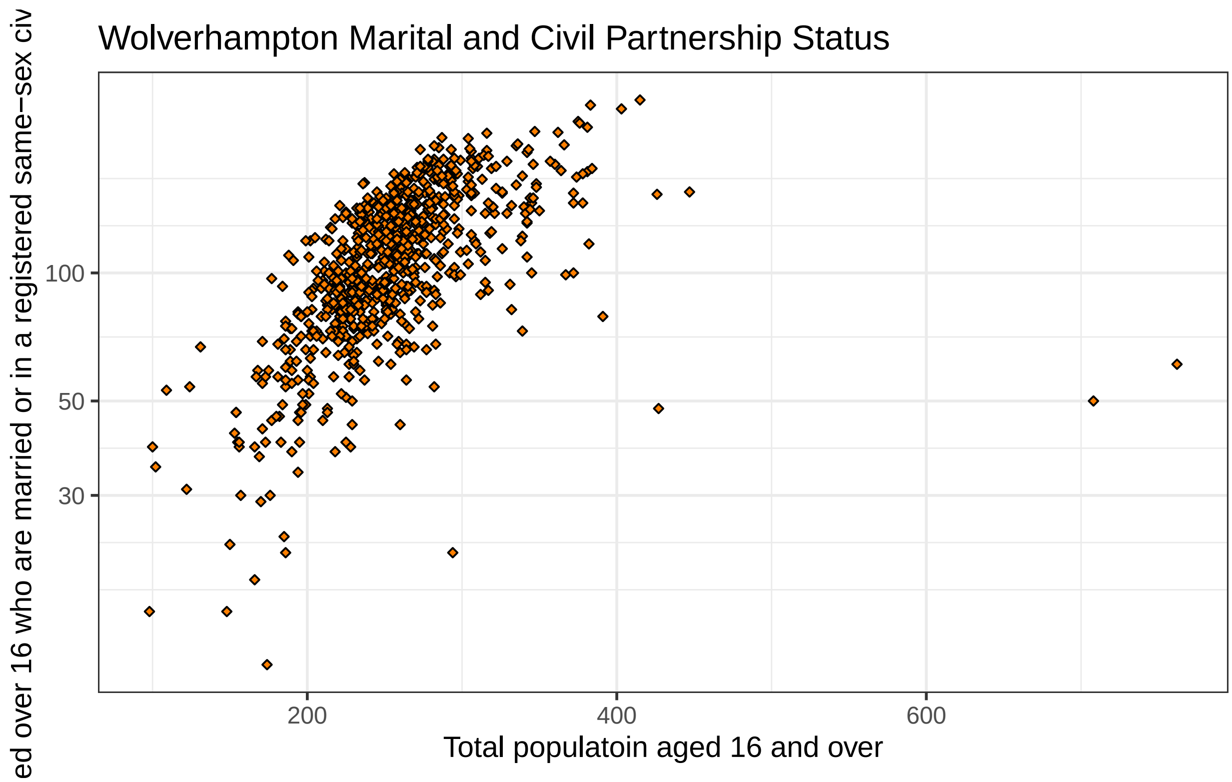
```
summary(Wolverhampton_2011OAC$k010)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0    83.0   109.0   111.6   139.0   255.0
```

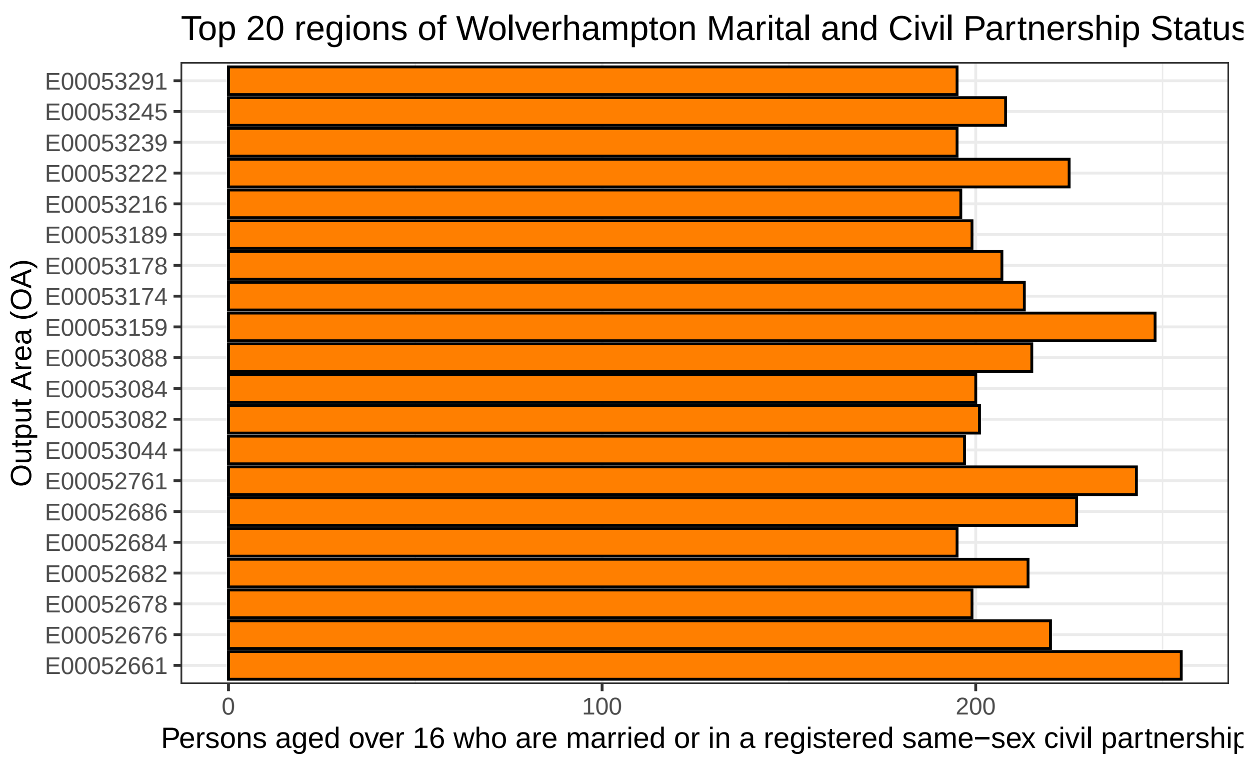
```
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k010
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#ff7f00", colour="black") +
  ggplot2::ggtitle("k010 : Persons aged over 16 who are married or in a registered same-sex civil partn
  ggplot2::xlab("Marital_and_Civil_Partnership_Status") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```



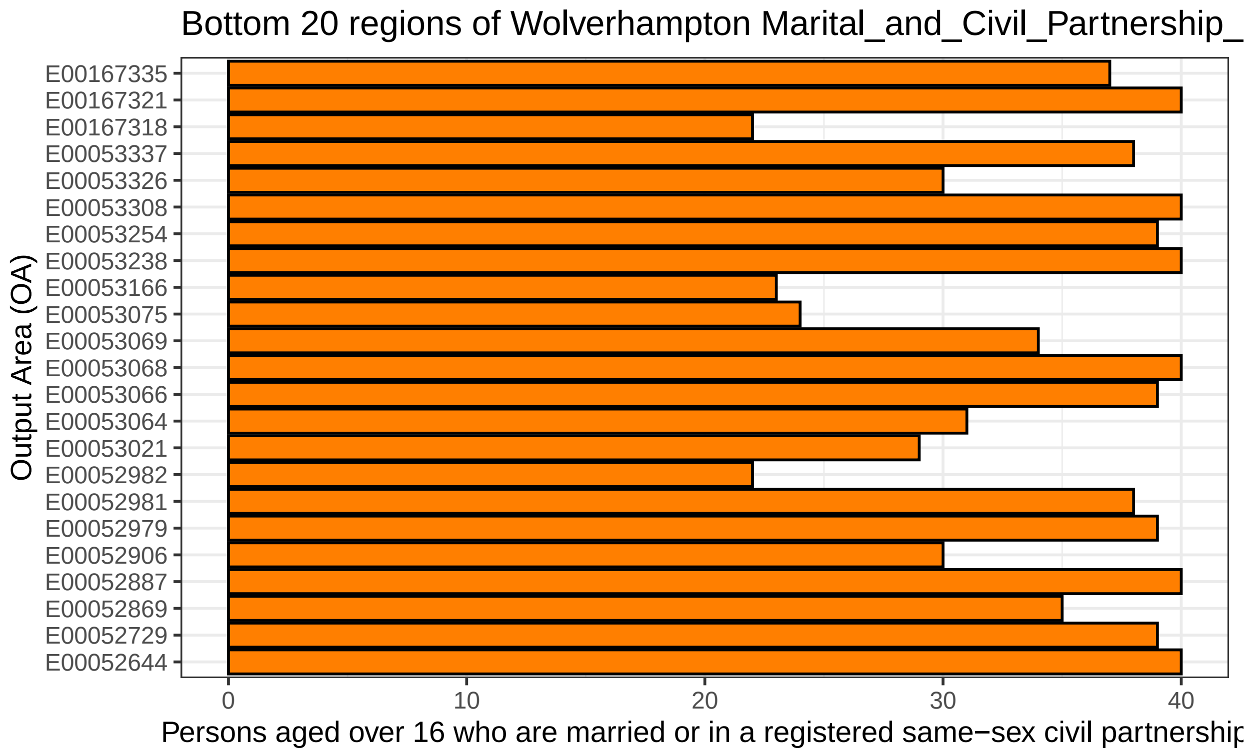

```
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Population_16_and_over,
      y = k010
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#ff7f00") +
  ggplot2::ggtitle("Wolverhampton Marital and Civil Partnership Status") +
  ggplot2::xlab("Total populatoin aged 16 and over") +
  ggplot2::ylab("Persons aged over 16 who are married or in a registered same-sex civil partnership") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```
# Top 20 regions of Wolverhampton Marital and Civil Partnership Status

k010_max <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k010) %>%
  dplyr::filter(k010>20) %>%
  dplyr::slice_max(k010, n=20)

ggplot2::ggplot(k010_max,
                aes(
                  x = k010,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#ff7f00", colour="black") +
  ggplot2::ggtitle("Top 20 regions of Wolverhampton Marital and Civil Partnership Status")+
  ggplot2::xlab("Persons aged over 16 who are married or in a registered same-sex civil partnership")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```




```
# Bottom 20 regions of Wolverhampton Marital_and_Civil_Partnership_Status

k010_min <-
  Wolverhampton_2011OAC %>%
  dplyr::filter(k010>20) %>%
  dplyr::select(OA, k010) %>%
  dplyr::slice_min(k010, n=20)

ggplot2::ggplot(k010_min,
                aes(
                  x = k010,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#ff7f00", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton Marital_and_Civil_Partnership_Status")+
  ggplot2::xlab("Persons aged over 16 who are married or in a registered same-sex civil partnership")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```

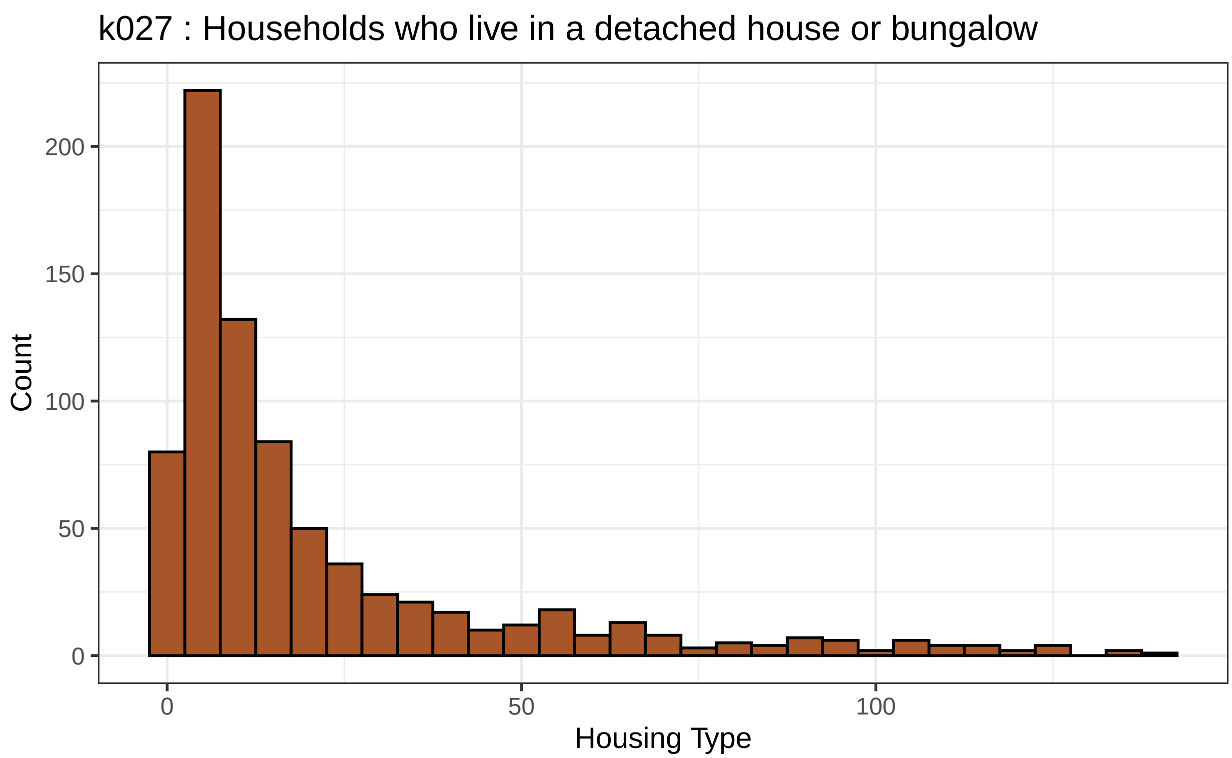


k027-Households who live in a detached house or bungalow

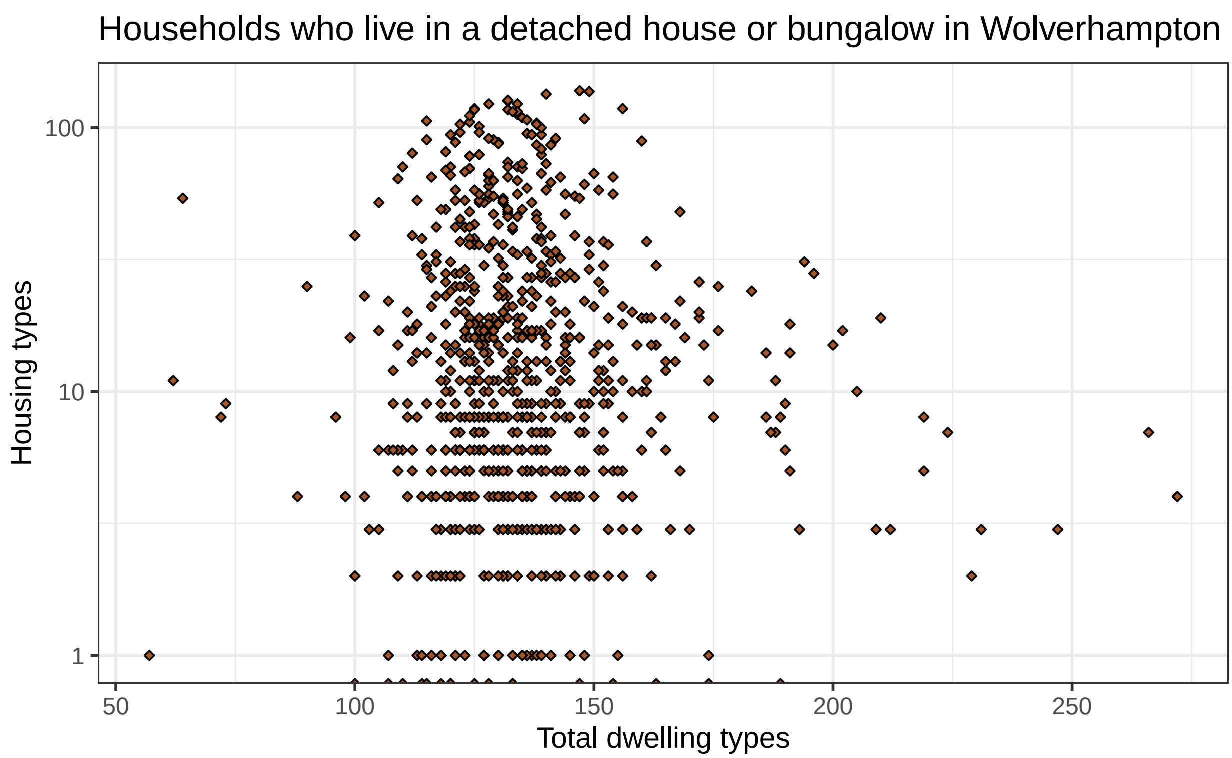
```r
summary(Wolverhampton_2011OAC$k027)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   10.00   21.22   25.00  138.00
```

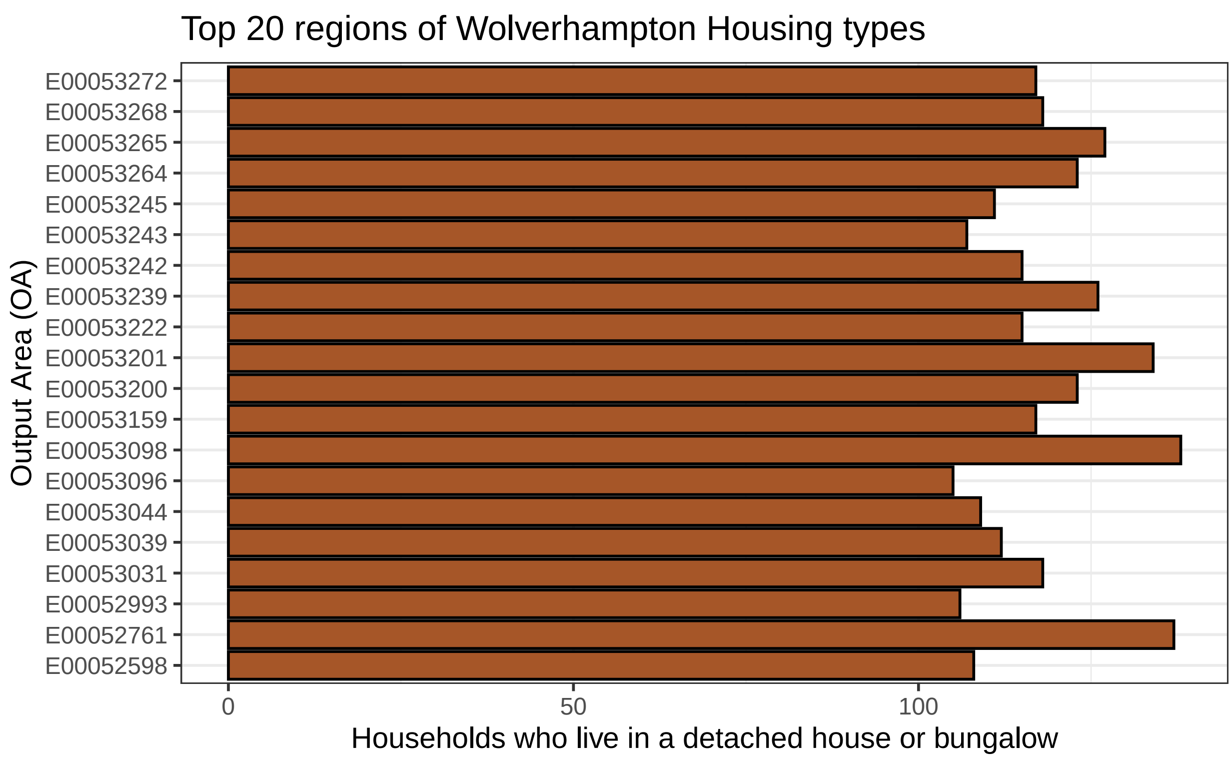
```r
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k027
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#a65628", colour="black") +
  ggplot2::ggtitle("k027 : Households who live in a detached house or bungalow") +
  ggplot2::xlab("Housing Type") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```



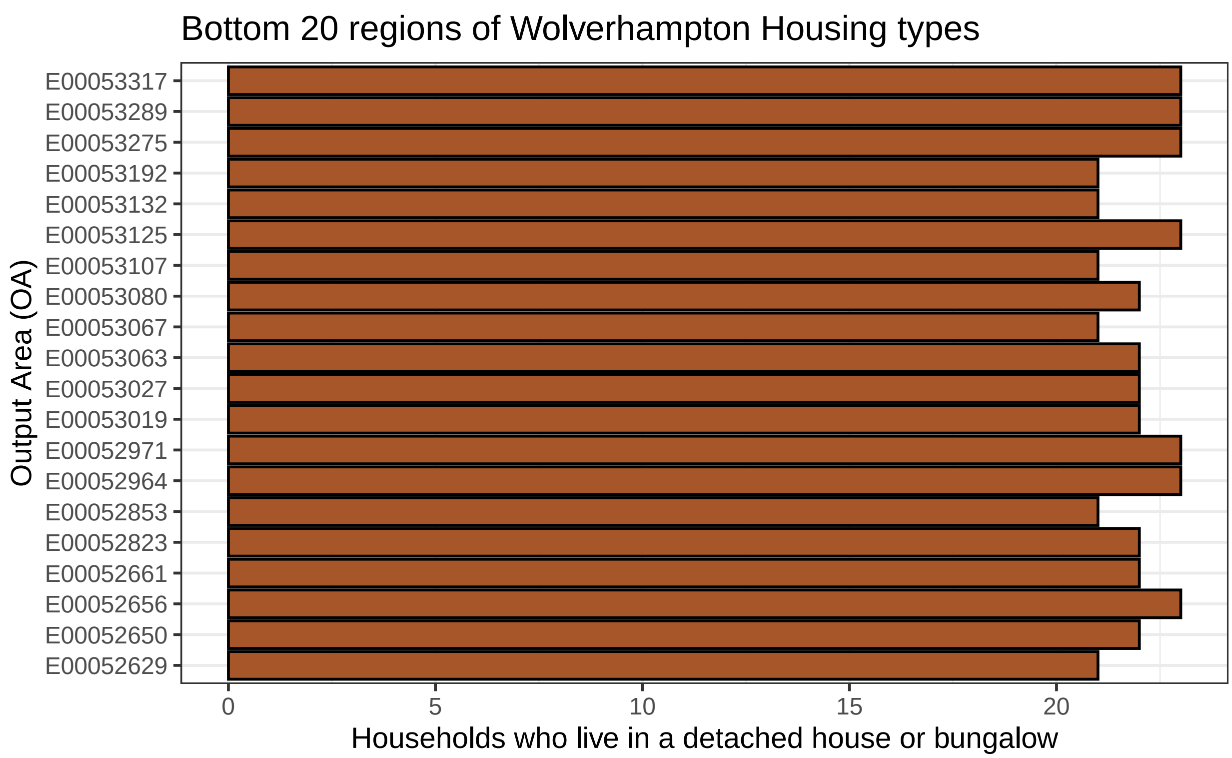

```r
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Household_Spaces,
      y = k027
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#a65628") +
  ggplot2::ggtitle("Households who live in a detached house or bungalow in Wolverhampton") +
  ggplot2::xlab("Total dwelling types") +
  ggplot2::ylab("Housing types") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```r
# Top 20 regions of Wolverhampton Housing types

k027_max <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k027) %>%
  dplyr::slice_max(k027, n=20)

ggplot2::ggplot(k027_max,
                aes(
                  x = k027,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#a65628", colour="black") +
  ggplot2::ggtitle("Top 20 regions of Wolverhampton Housing types")+
  ggplot2::xlab("Households who live in a detached house or bungalow")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```




```
# Bottom 20 regions of Wolverhampton Housing types

k027_min <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k027) %>%
  dplyr::filter(k027>20) %>%
  dplyr::slice_min(k027, n=20)

ggplot2::ggplot(k027_min,
                aes(
                  x = k027,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#a65628", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton Housing types")+
  ggplot2::xlab("Households who live in a detached house or bungalow")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```

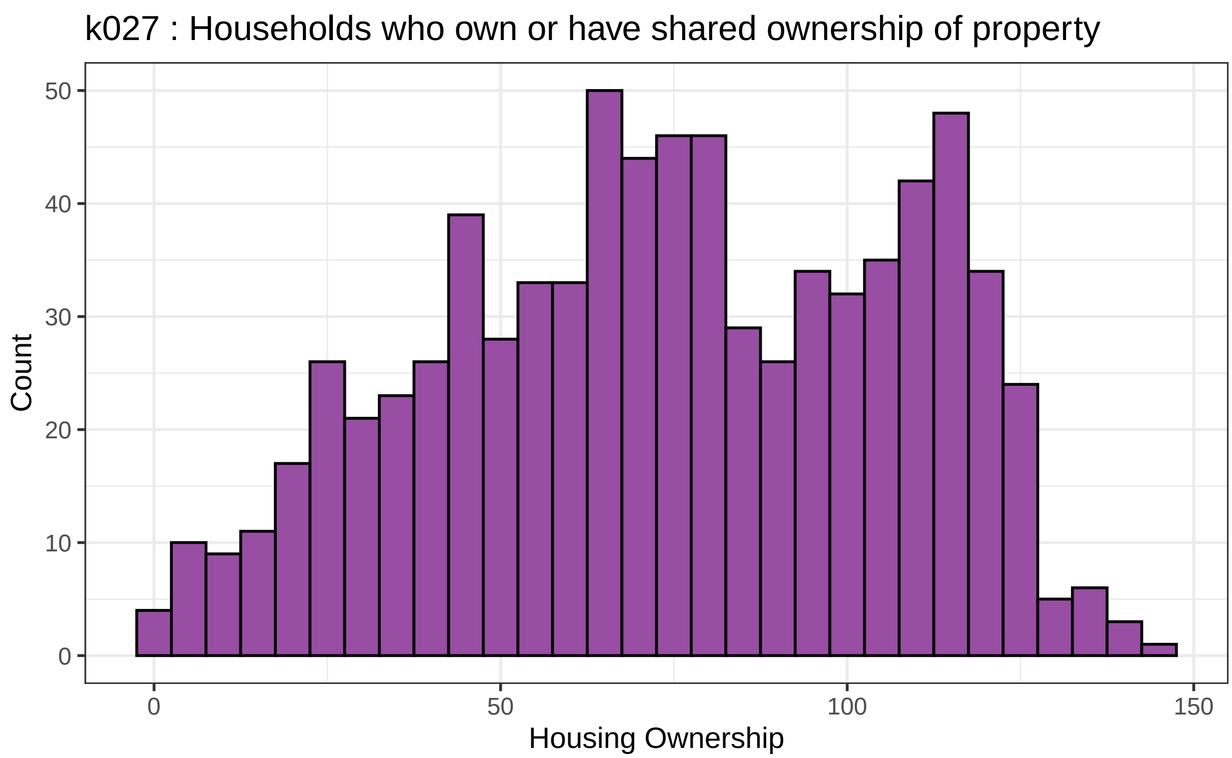


k031-Households who own or have shared ownership of property

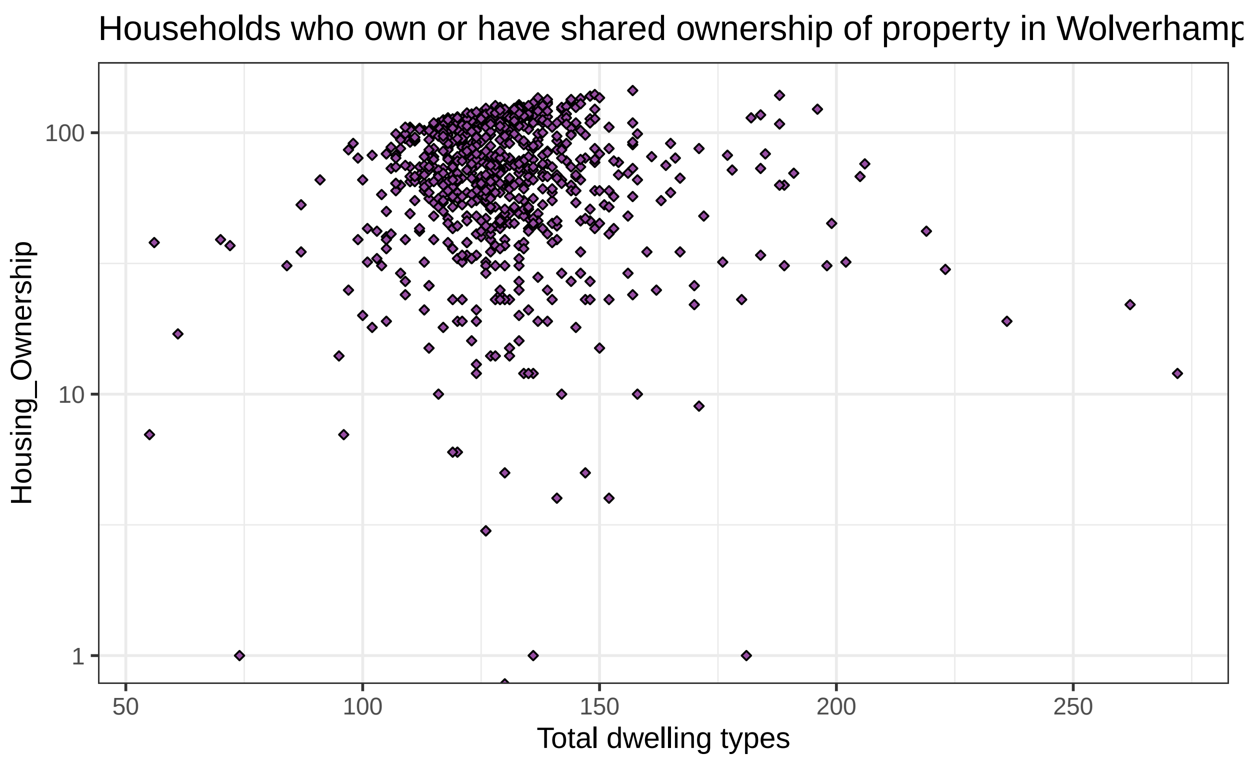
```
summary(Wolverhampton_2011OAC$k031)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   49.00   74.00   74.18  103.00  145.00
```

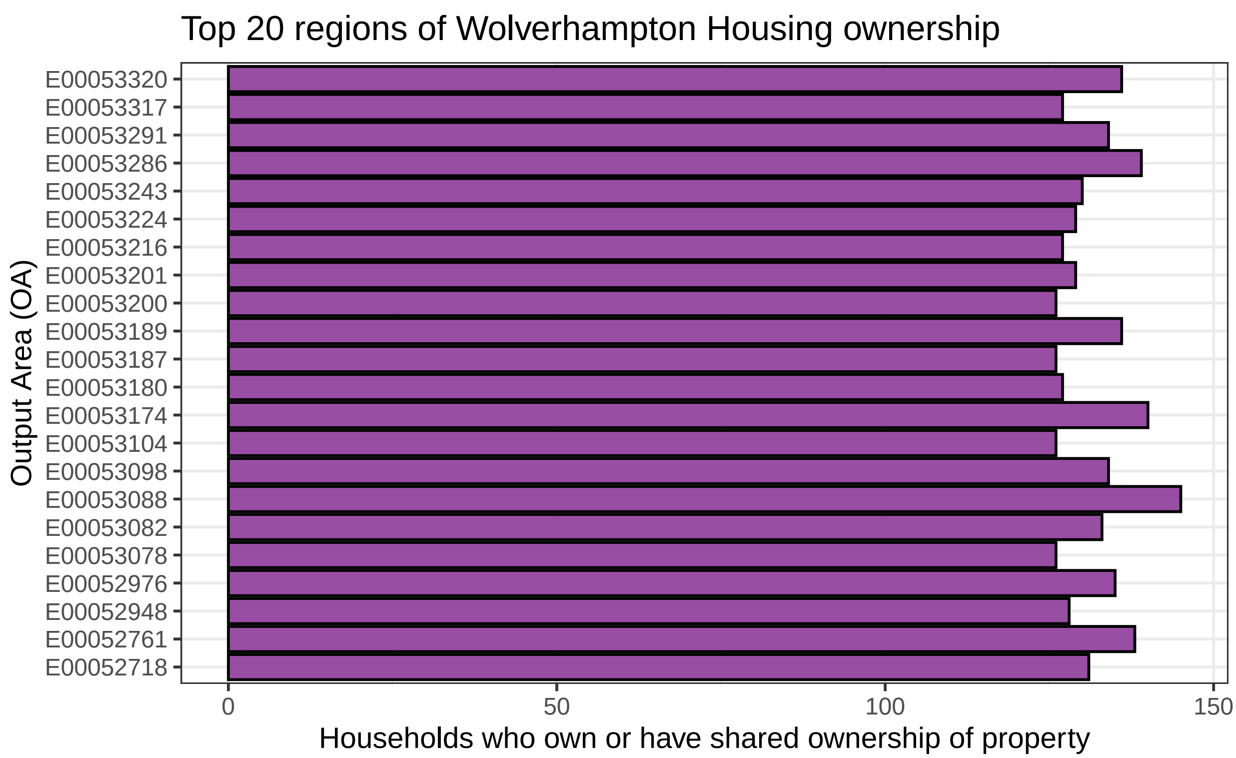
```
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k031
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#984ea3", colour="black") +
  ggplot2::ggtitle("k027 : Households who own or have shared ownership of property") +
  ggplot2::xlab("Housing Ownership") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```



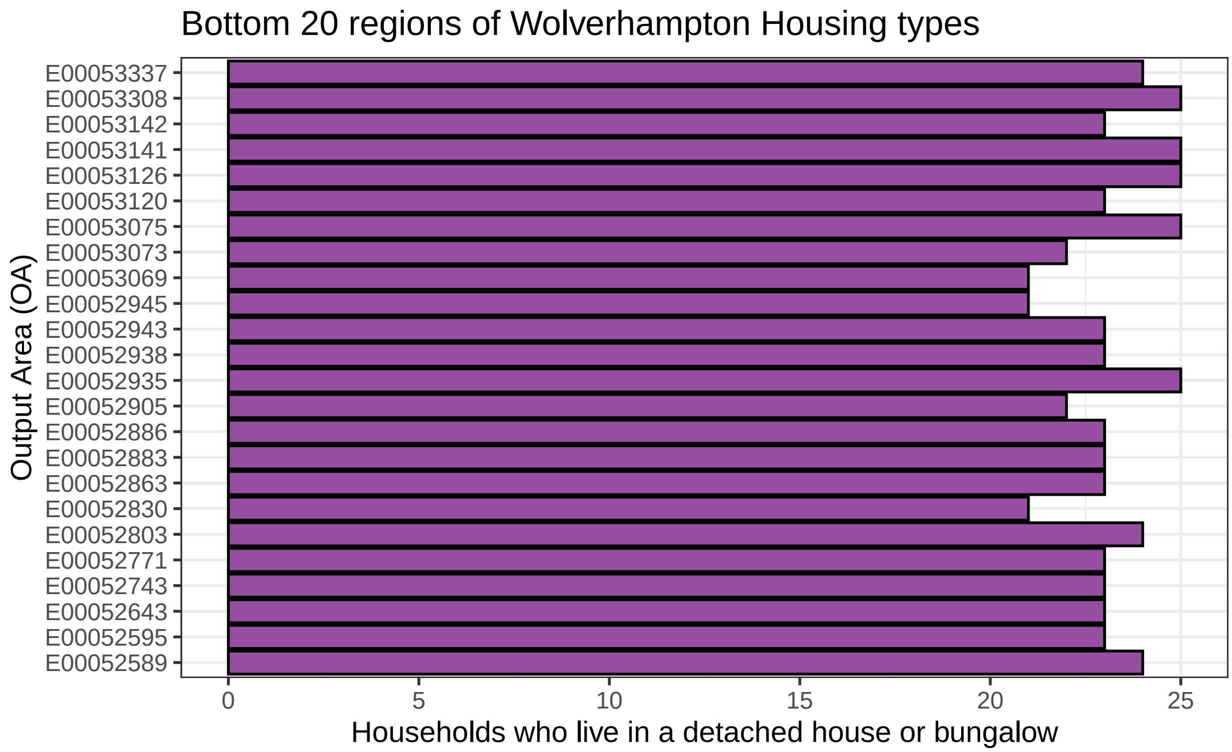

```r
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Households,
      y = k031
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#984ea3") +
  ggplot2::ggtitle("Households who own or have shared ownership of property in Wolverhampton") +
  ggplot2::xlab("Total dwelling types") +
  ggplot2::ylab("Housing_Ownership") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```r
# Top 20 regions of Wolverhampton Housing ownership

k031_max <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k031) %>%
  dplyr::filter(k031>20) %>%
  dplyr::slice_max(k031, n=20)

ggplot2::ggplot(k031_max,
                aes(
                  x = k031,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#984ea3", colour="black") +
  ggplot2::ggtitle("Top 20 regions of Wolverhampton Housing ownership")+
  ggplot2::xlab("Households who own or have shared ownership of property")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```




```
# Bottom 20 regions of Wolverhampton Housing types

k031_min <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k031) %>%
  dplyr::filter(k031>20) %>%
  dplyr::slice_min(k031, n=20)

ggplot2::ggplot(k031_min,
                aes(
                  x = k031,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#984ea3", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton Housing types")+
  ggplot2::xlab("Households who live in a detached house or bungalow")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```

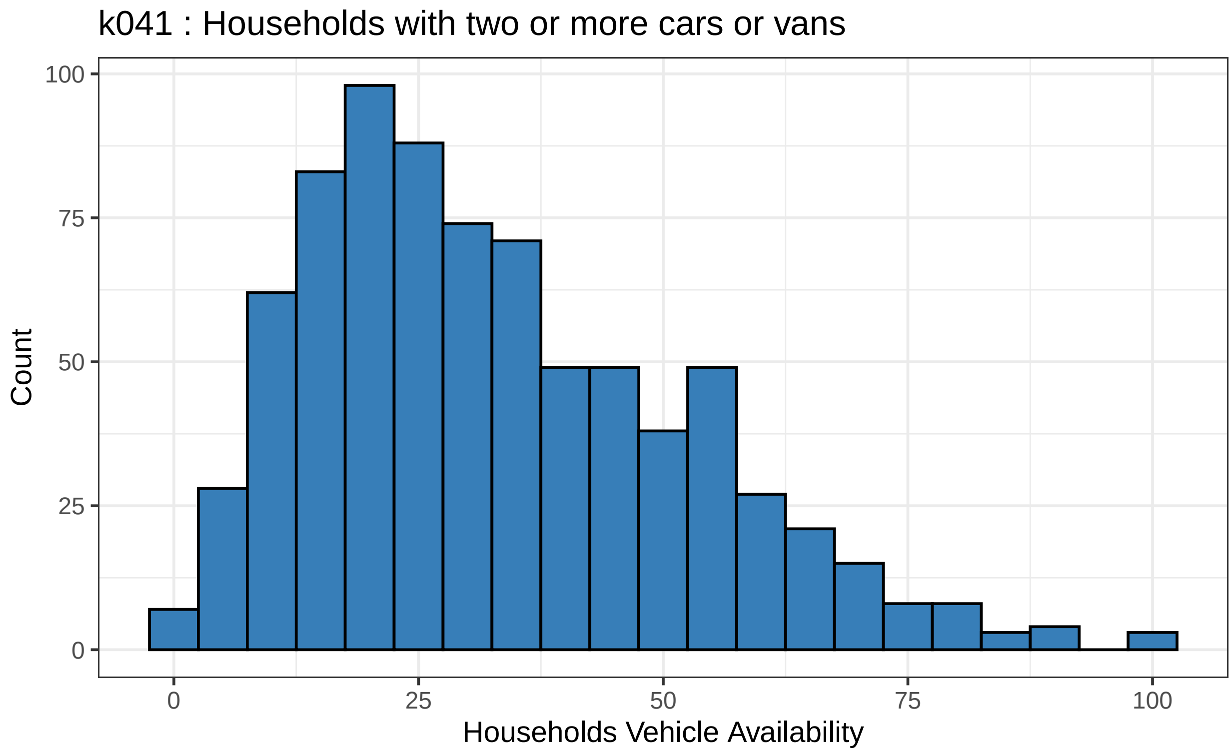


k041-Households with two or more cars or vans

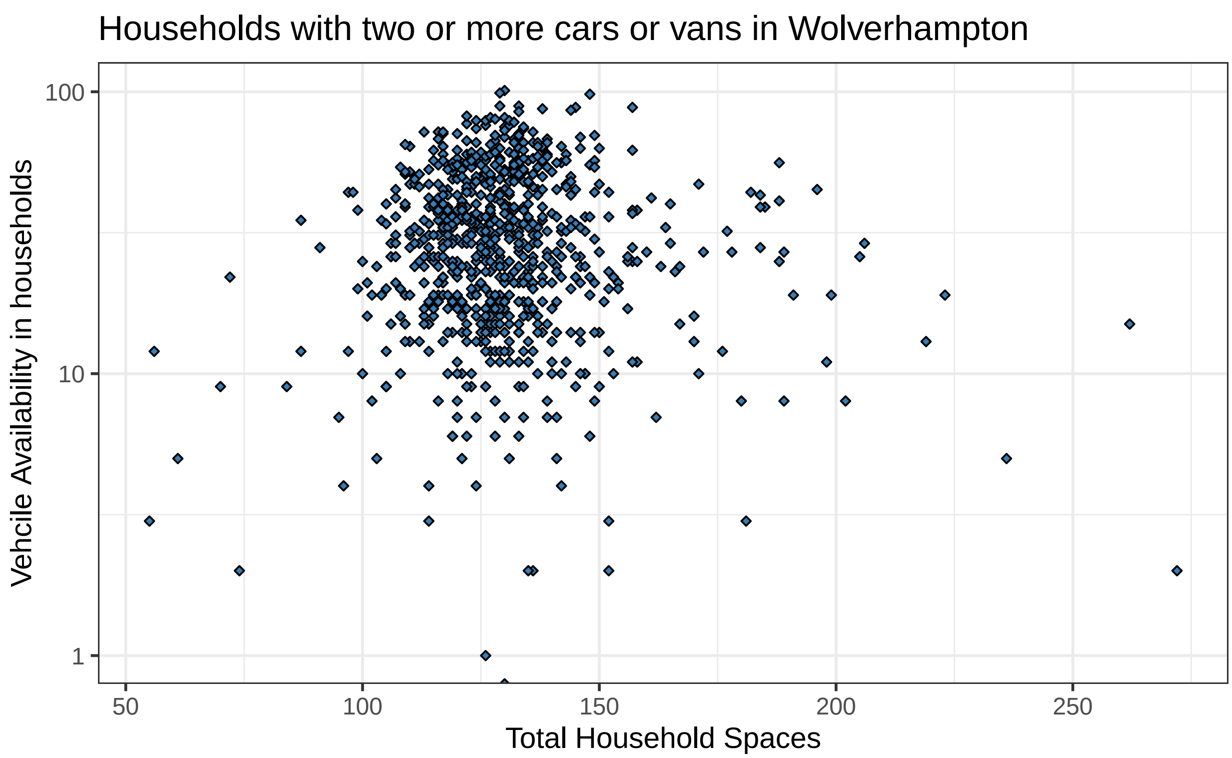
```
summary(Wolverhampton_2011OAC$k041)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   29.00   32.87   45.00  101.00
```

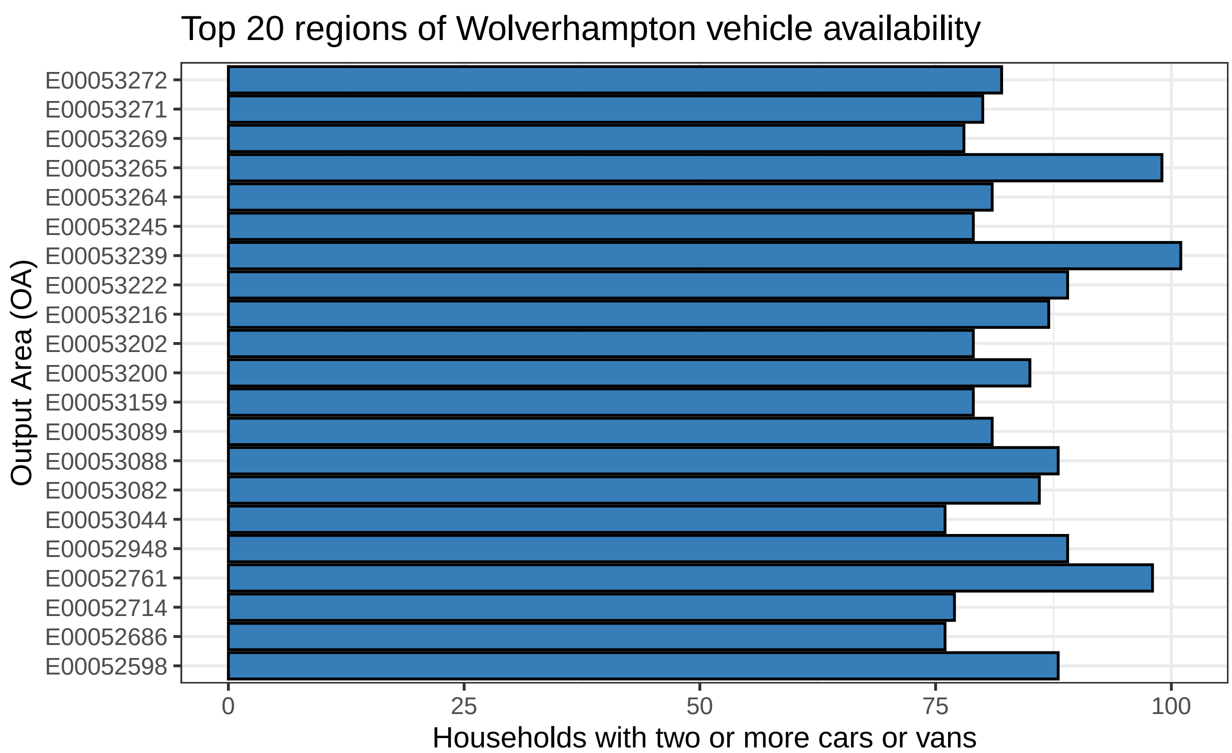
```
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k041
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#377eb8", colour="black") +
  ggplot2::ggtitle("k041 : Households with two or more cars or vans") +
  ggplot2::xlab("Households Vehicle Availability") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```



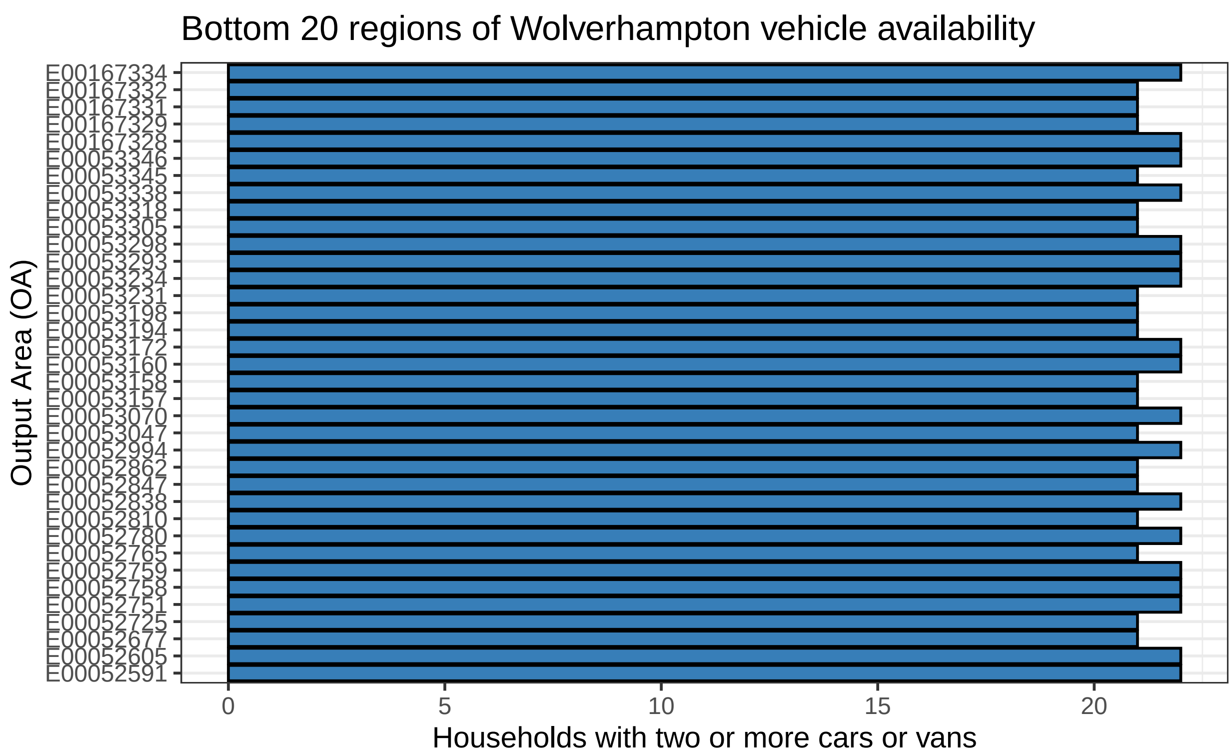

```
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Households,
      y = k041
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#377eb8") +
  ggplot2::ggtitle("Households with two or more cars or vans in Wolverhampton") +
  ggplot2::xlab("Total Household Spaces") +
  ggplot2::ylab("Vehcile Availability in households") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```r
# Top 20 regions of Wolverhampton vehicle availability

k041_max <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k041) %>%
  dplyr::filter(k041>20) %>%
  dplyr::slice_max(k041, n=20)

ggplot2::ggplot(k041_max,
                aes(
                  x = k041,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#377eb8", colour="black") +
  ggplot2::ggtitle("Top 20 regions of Wolverhampton vehicle availability")+
  ggplot2::xlab("Households with two or more cars or vans")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```




```r
# Bottom 20 regions of Wolverhampton vehicle availability

k041_min <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k041) %>%
  dplyr::filter(k041>20) %>%
  dplyr::slice_min(k041, n=20)

ggplot2::ggplot(k041_min,
                aes(
                  x = k041,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#377eb8", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton vehicle availability")+
  ggplot2::xlab("Households with two or more cars or vans")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```

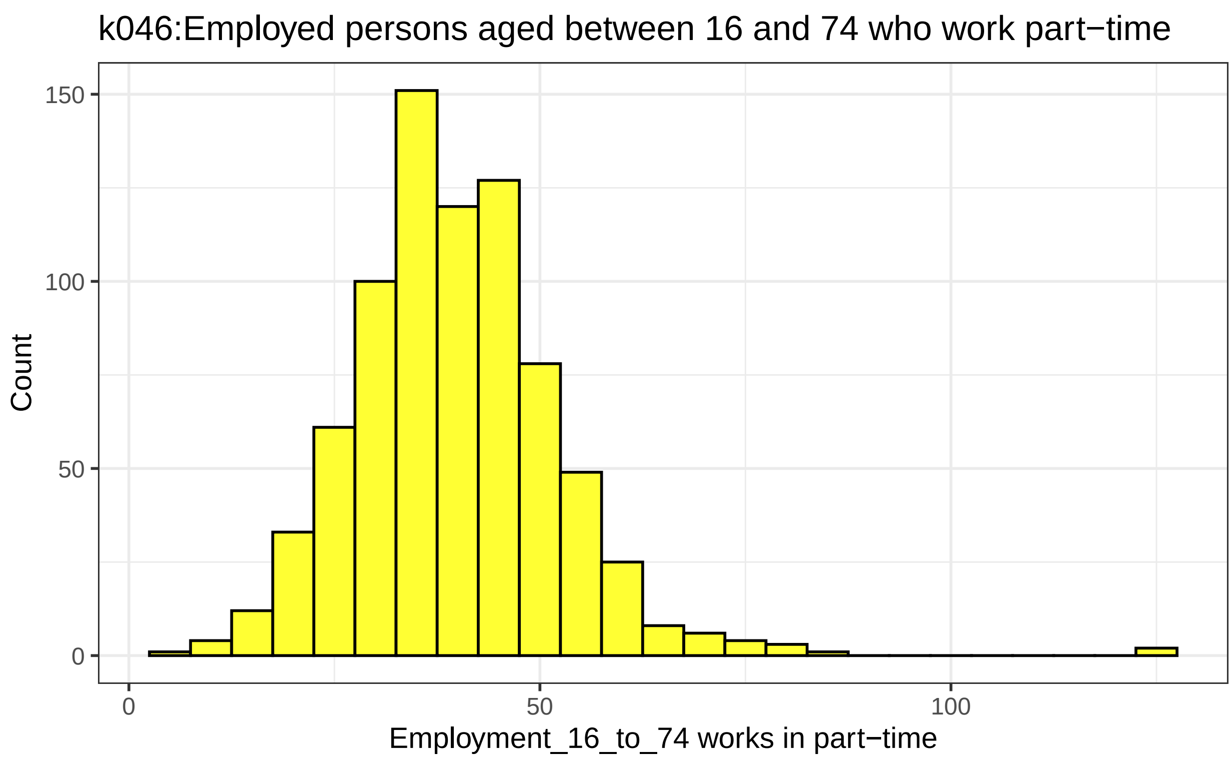


k046-Employed persons aged between 16 and 74 who work part-time

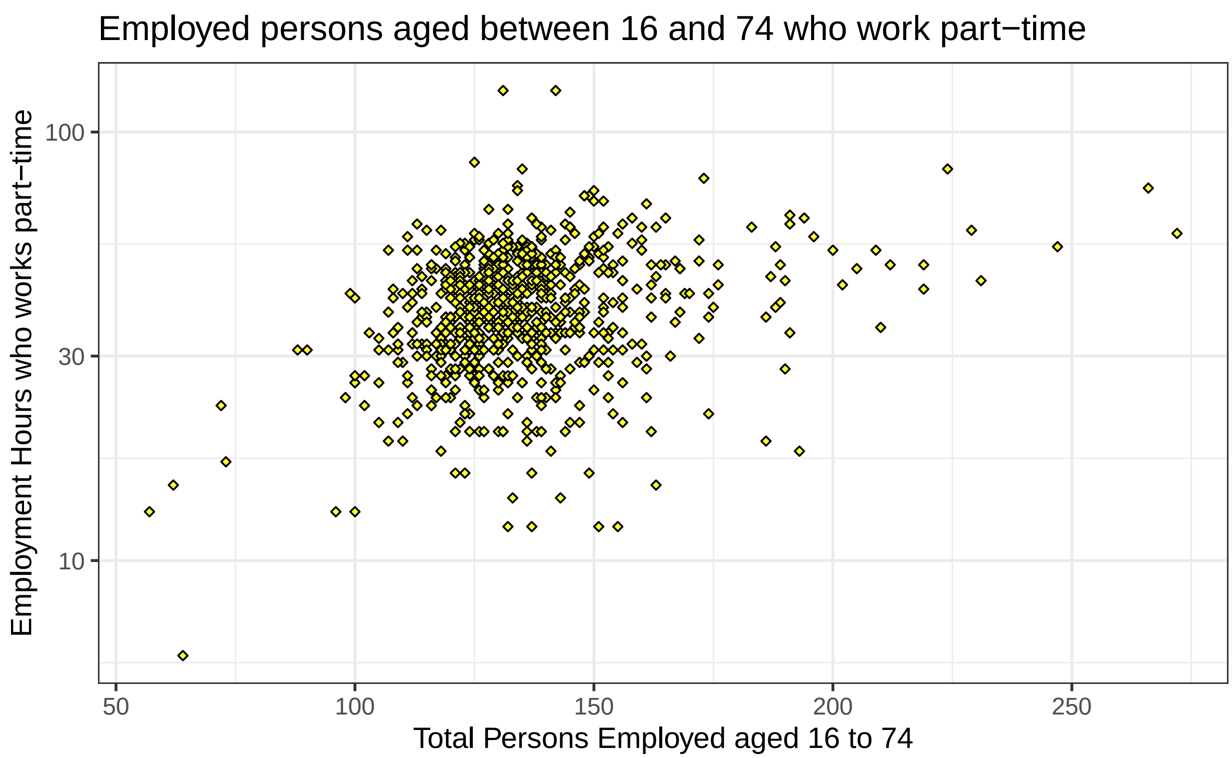
```r
summary(Wolverhampton_2011OAC$k046)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   32.00   39.00   39.62   47.00  125.00
```

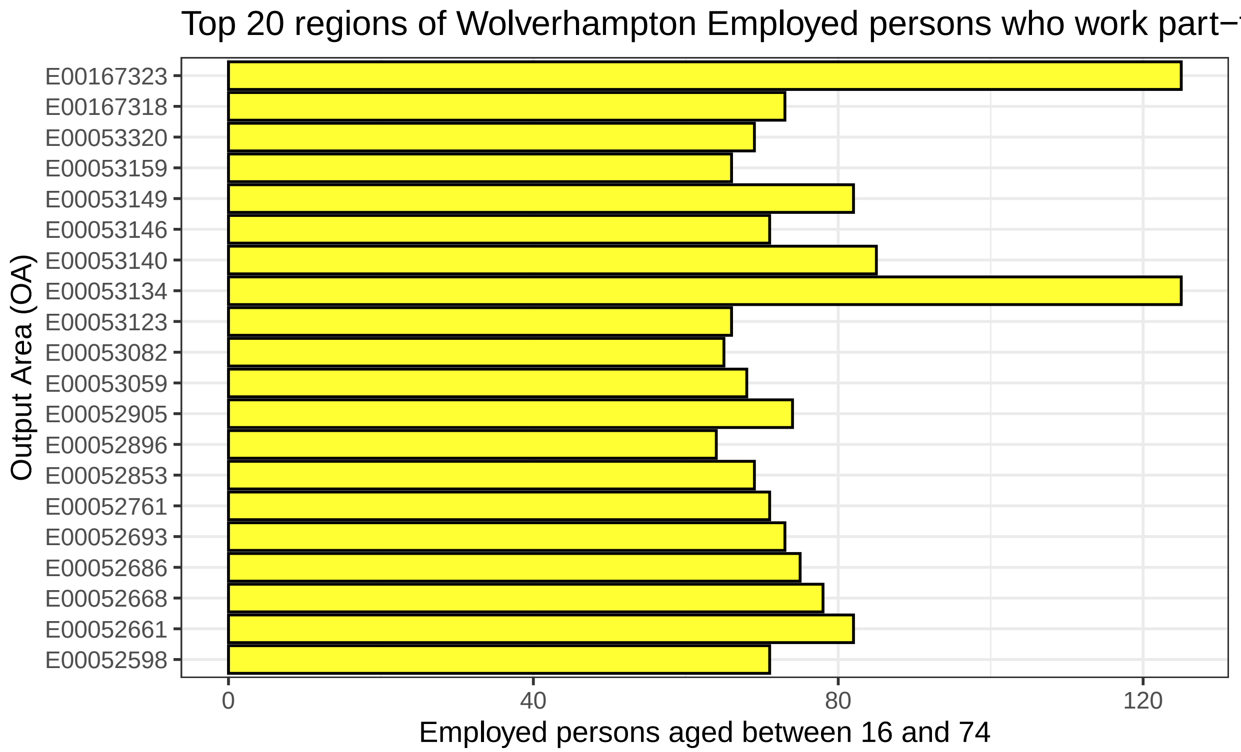
```r
# Histogram

Wolverhampton_2011OAC %>%
  ggplot2::ggplot (
    aes(
      x = k046
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5, fill="#ffff33", colour="black") +
  ggplot2::ggtitle("k046:Employed persons aged between 16 and 74 who work part-time") +
  ggplot2::xlab("Employment_16_to_74 works in part-time") +
  ggplot2::ylab("Count") +
  ggplot2::theme_bw()
```



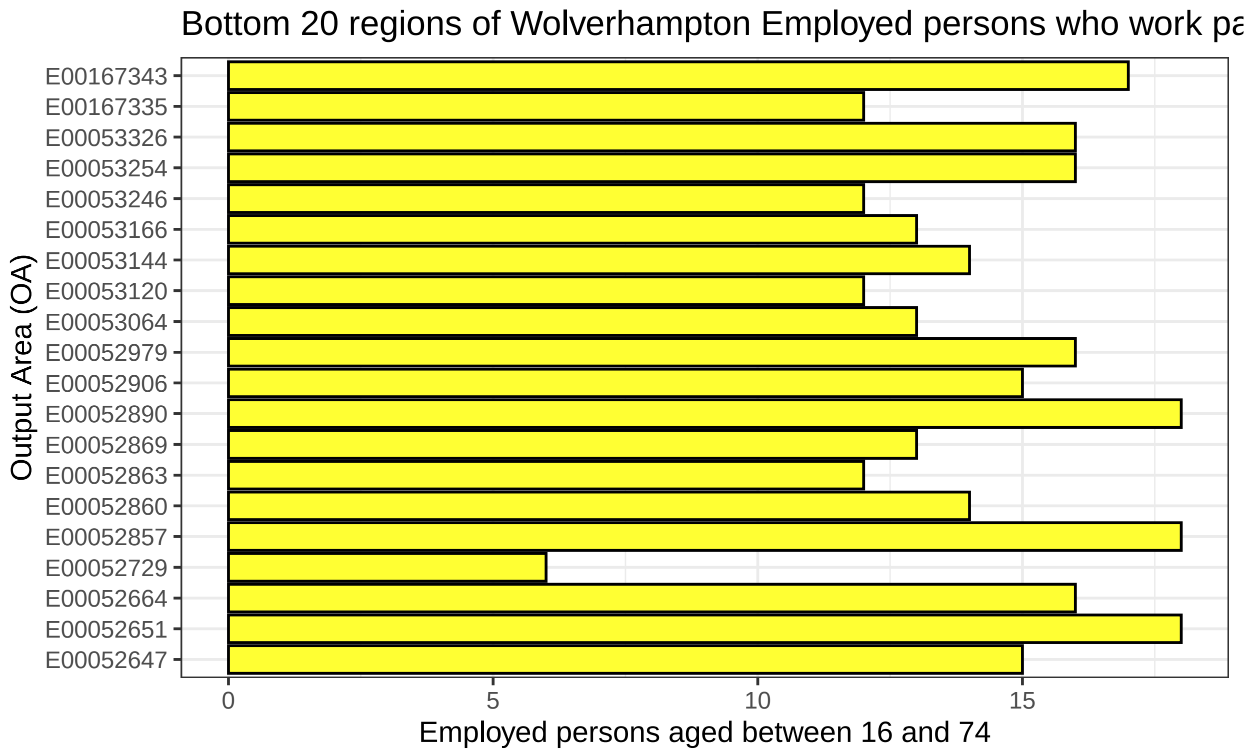

```
# Scatterplot

Wolverhampton_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Household_Spaces,
      y = k046
    )
  )+
  ggplot2::geom_point(color= "black", shape = 23, size = 1, fill = "#ffff33") +
  ggplot2::ggtitle("Employed persons aged between 16 and 74 who work part-time") +
  ggplot2::xlab("Total Persons Employed aged 16 to 74") +
  ggplot2::ylab("Employment Hours who works part-time") +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```




```
# Top 20 regions of Wolverhampton Employed persons who work part-time

k046_max <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k046) %>%
  dplyr::slice_max(k046, n=20)

ggplot2::ggplot(k046_max,
                aes(
                  x = k046,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#ffff33", colour="black") +
  ggplot2::ggtitle("Top 20 regions of Wolverhampton Employed persons who work part-time")+
  ggplot2::xlab("Employed persons aged between 16 and 74")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```




```r
# Bottom 20 regions of Wolverhampton Employed persons who work part-time

k046_min <-
  Wolverhampton_2011OAC %>%
  dplyr::select(OA, k046) %>%
  dplyr::slice_min(k046, n=20)

ggplot2::ggplot(k046_min,
                aes(
                  x = k046,
                  y = OA,
                )
)+
  ggplot2::geom_bar(position = "stack", stat = "identity", fill="#ffff33", colour="black") +
  ggplot2::ggtitle("Bottom 20 regions of Wolverhampton Employed persons who work part-time")+
  ggplot2::xlab("Employed persons aged between 16 and 74")+
  ggplot2::ylab("Output Area (OA)")+
  ggplot2::theme_bw()
```



Exploratory statistics

The graphics above provide preliminary evidence that the distribution of people aged 20 to 24 might, in fact, be different in different 2011 supergroups. In the remainder of the practical session, we are going to explore that hypothesis further. First, load the necessary statistical libraries.

The code below calculates the percentage of people aged 20 to 24 (i.e., u011) over total population per OA, but it also recodes (see recode) the names of the 2011OAC supergroups to a shorter 2-letter version, which is useful for the tables presented further below.

Only the OA code, the recoded 2011OAC supergroup name, and the newly created perc_age_20_to_24 are retained in the new table leic_2011OAC_20to24. Such a step is sometimes useful as stepping stone for further analysis and can make the code easier to read further down the line. Sometimes it is also a necessary step when interacting with certain libraries, which are not fully compatible with Tidyverse libraries, such as leveneTest

Calculate percentage for the each variables

```
Percentage <-
  Wolverhampton_2011OAC %>%
  dplyr::mutate(
    Perc_k004 = (k004 / Total_Population) * 100,
    Perc_k009 = (k009 / Total_Population_16_and_over) * 100,
    Perc_k010 = (k010 / Total_Population_16_and_over) * 100,
    Perc_k027 = (k027 / Total_Household_Spaces) * 100,
    Perc_k031 = (k031 / Total_Households) * 100,
```

```
    Perc_k041 = (k041 / Total_Households) * 100,
    Perc_k046 = (k046 / Total_Employment_16_to_74) * 100
  ) %>%
  dplyr::select(OA, Perc_k004, Perc_k009, Perc_k010,
                Perc_k027, Perc_k031, Perc_k041, Perc_k046
                )
```

Descriptive statistics

```
# Calculating descriptive statistics

wolverhampton_stat_desc <-
  Percentage %>%
  dplyr::select(Perc_k004, Perc_k009, Perc_k010,
                Perc_k027, Perc_k031, Perc_k041, Perc_k046) %>%
  pastecs::stat.desc(norm =TRUE)

wolverhampton_stat_desc %>%
  knitr::kable(digits = 5)
```

	Perc_k004	Perc_k009	Perc_k010	Perc_k027	Perc_k031	Perc_k041	Perc_k046
nbr.val	785.00000	785.00000	785.00000	785.00000	785.00000	785.00000	785.00000
nbr.null	0.00000	0.00000	0.00000	17.00000	1.00000	1.00000	0.00000
nbr.na	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
min	3.51827	13.17073	6.89655	0.00000	0.00000	0.00000	12.37113
max	52.32558	89.45578	72.47387	96.21212	99.27007	77.69231	64.10256
range	48.80731	76.28505	65.57732	96.21212	99.27007	77.69231	51.73143
sum	18691.32819	29049.88200	33988.64147	12658.03389	45435.69382	20146.86762	23688.50896
median	24.13793	36.28319	42.91498	7.62712	58.20896	23.00885	29.50820
mean	23.81061	37.00622	43.29763	16.12488	57.87986	25.66480	30.17644
SE.mean	0.20505	0.37399	0.42974	0.72517	0.91407	0.53335	0.21622
CI.mean.0.95	0.40251	0.73414	0.84358	1.42351	1.79432	1.04696	0.42444
var	33.00465	109.79665	144.97175	412.81289	655.88755	223.30052	36.70061
std.dev	5.74497	10.47839	12.04042	20.31780	25.61030	14.94324	6.05810
coef.var	0.24128	0.28315	0.27808	1.26003	0.44247	0.58225	0.20076
skewness	-0.05498	0.83897	-0.16098	2.06691	-0.24200	0.68742	0.99092
skew.2SE	-0.31502	4.80731	-0.92245	11.84346	-1.38664	3.93891	5.67801
kurtosis	1.02426	2.32522	-0.32131	3.86446	-0.97436	-0.11994	3.00249
kurt.2SE	2.93822	6.67018	-0.92172	11.08572	-2.79509	-0.34407	8.61303
normtest.W	0.99115	0.96405	0.99216	0.71265	0.95823	0.95469	0.95548
normtest.p	0.00012	0.00000	0.00037	0.00000	0.00000	0.00000	0.00000

Shapiro test, Density histogram and QQ plot

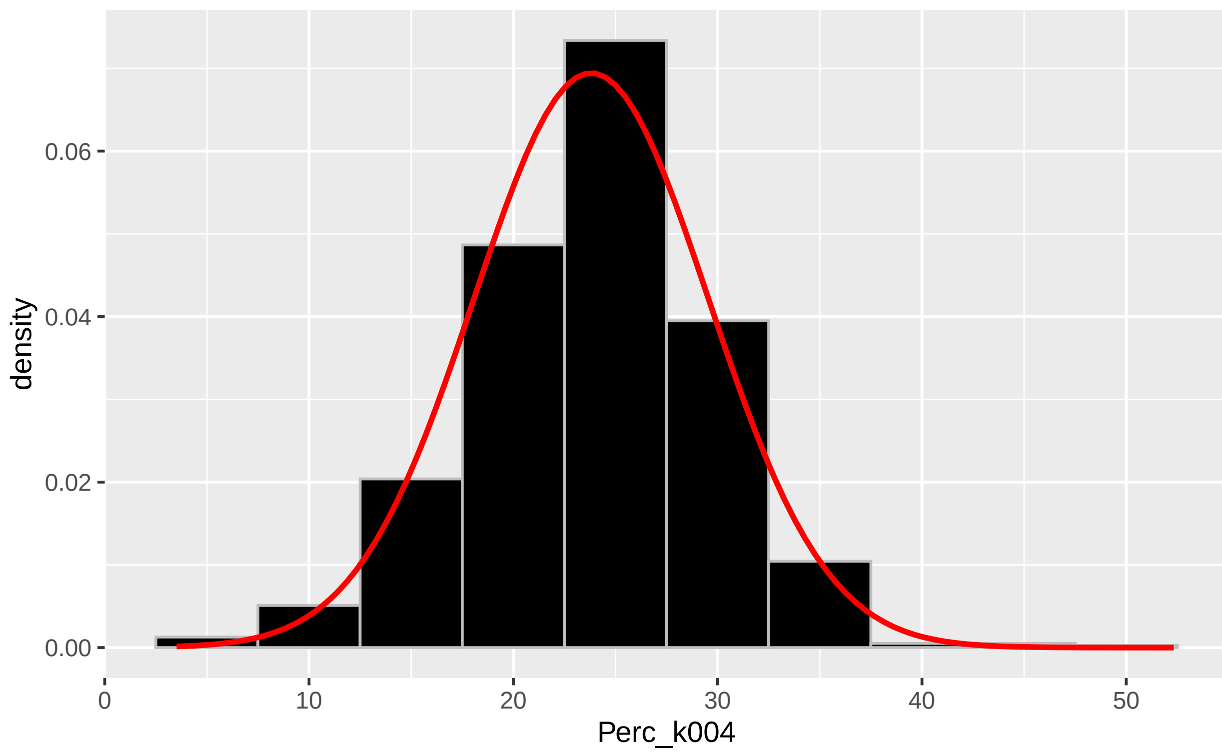
k004 - Persons aged 45 to 64

```
# Shapiro_Test
```

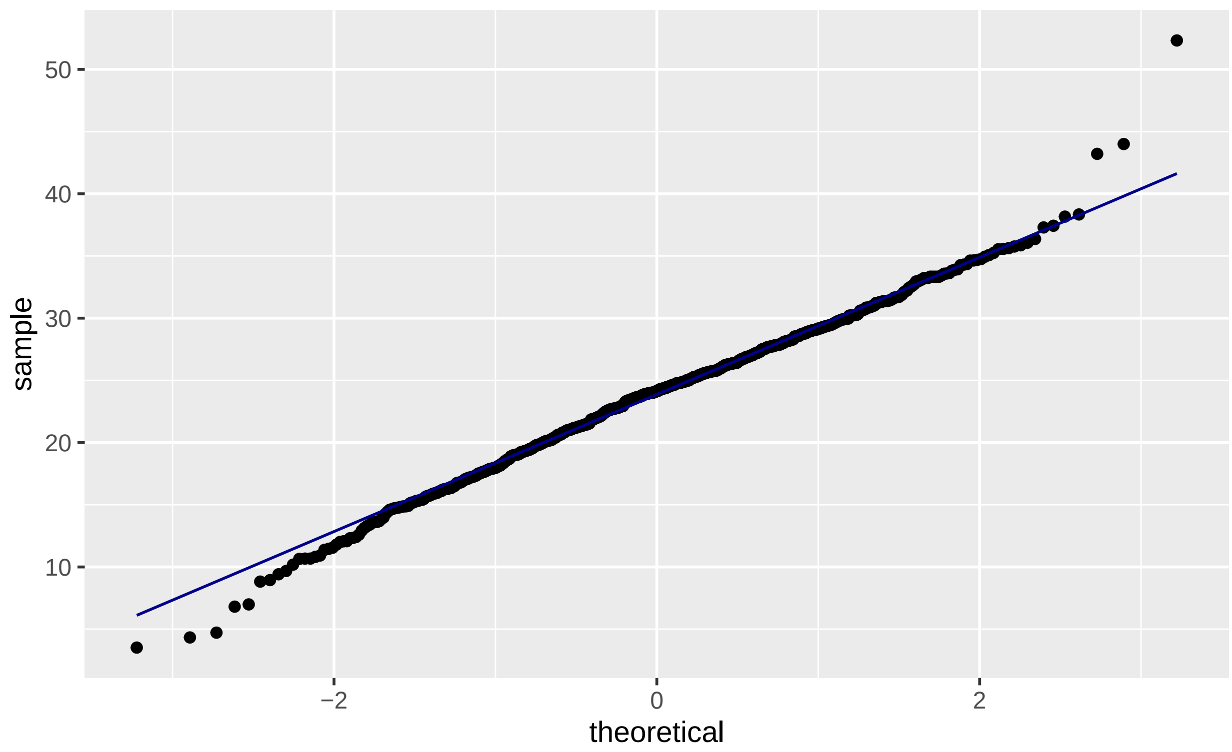
```r
Percentage %>%
  dplyr::pull(Perc_k004) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.99115, p-value = 0.0001186
```

```r
# Density_Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k004
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k004) %>% mean(),
      sd = Percentage %>% pull(Perc_k004) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```
# QQ-Plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k004
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



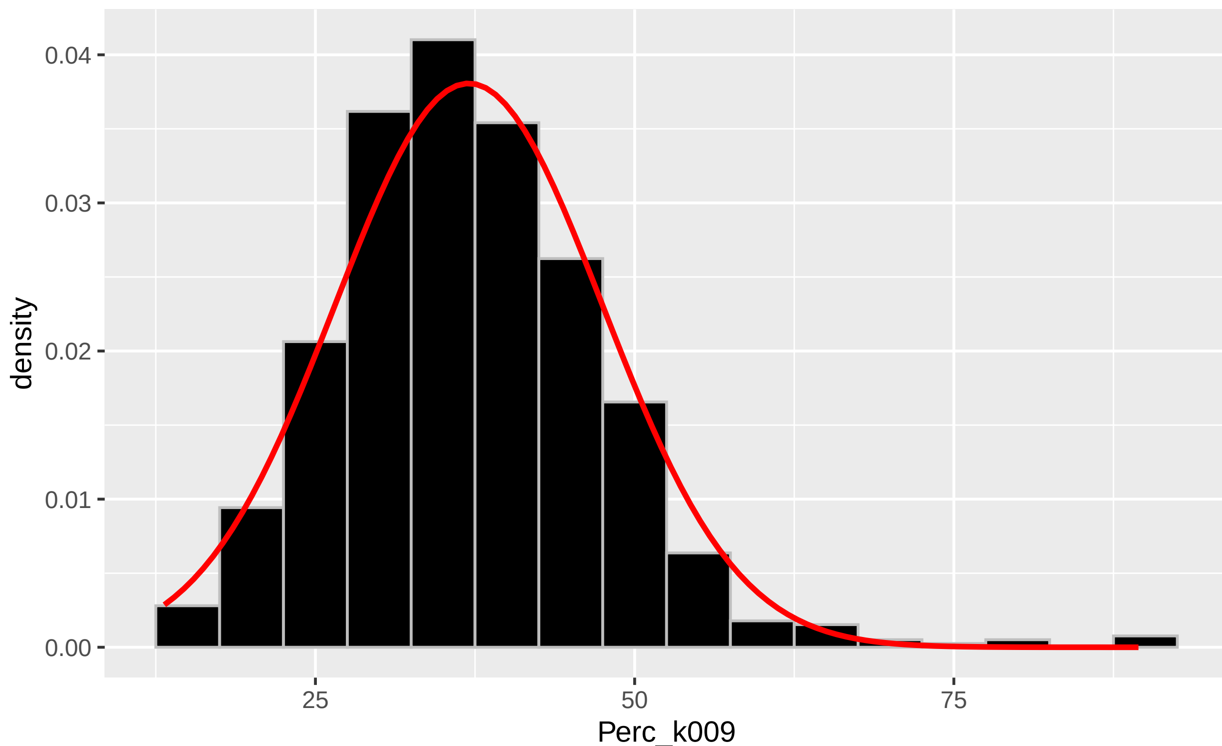
k009 - Persons aged over 16 who are single

```
# Shapiro-Test

Percentage %>%
  dplyr::pull(Perc_k009) %>%
  stats::shapiro.test()
```

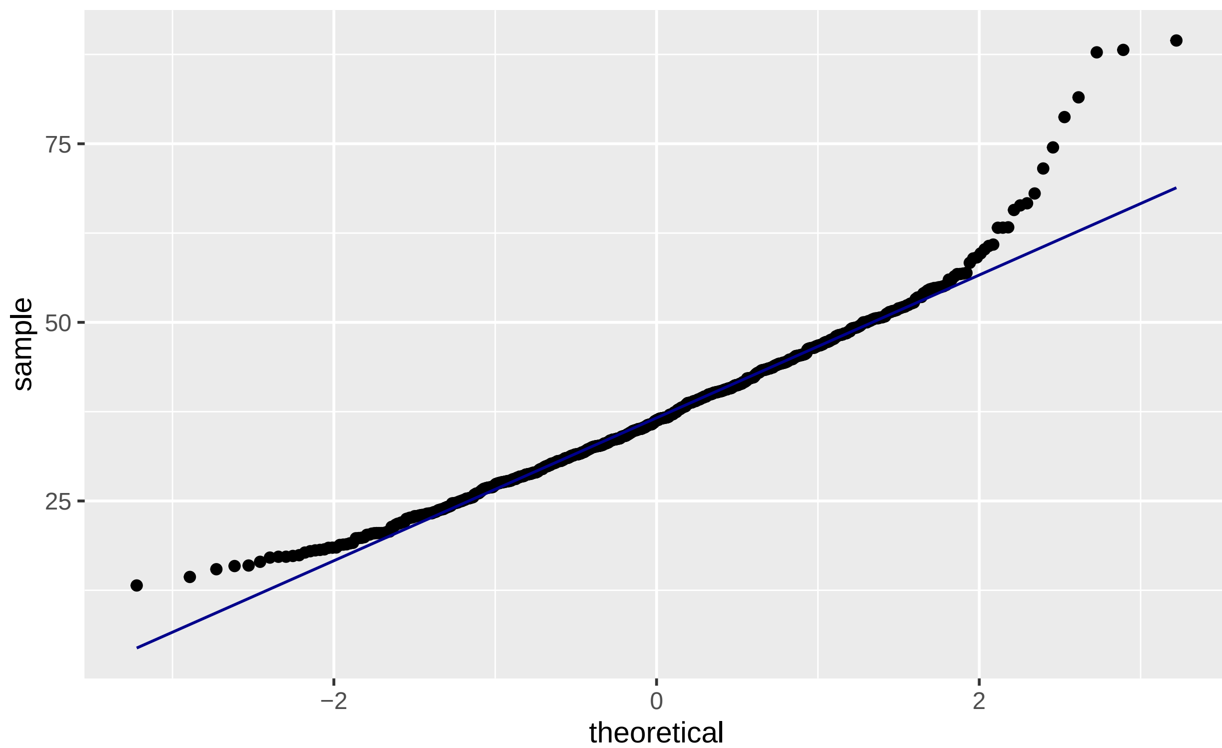
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.96405, p-value = 5.709e-13
```

```
# Density-Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k009
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
```

```
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k009) %>% mean(),
      sd = Percentage %>% pull(Perc_k009) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```
# QQ-plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k009
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



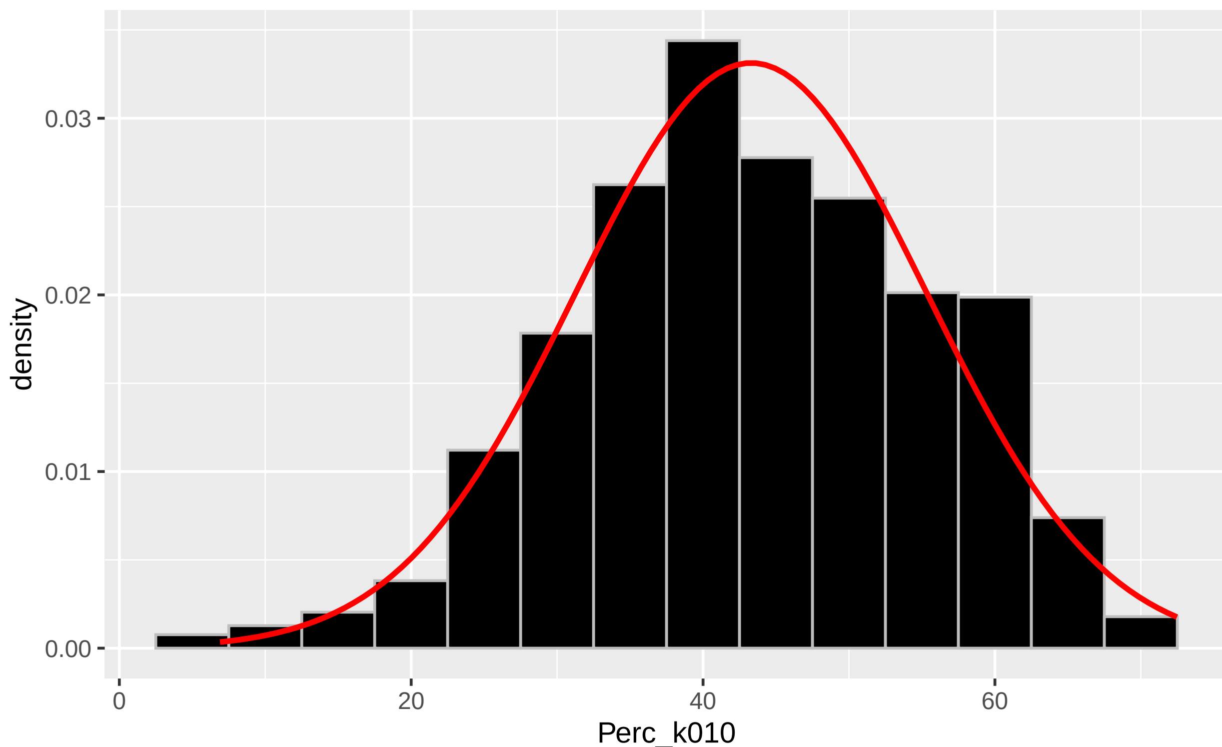
k010 - Persons aged over 16 who are married or in a registered same-sex civil partnership

```
# Shapiro-test

Percentage %>%
  dplyr::pull(Perc_k010) %>%
  stats::shapiro.test()
```

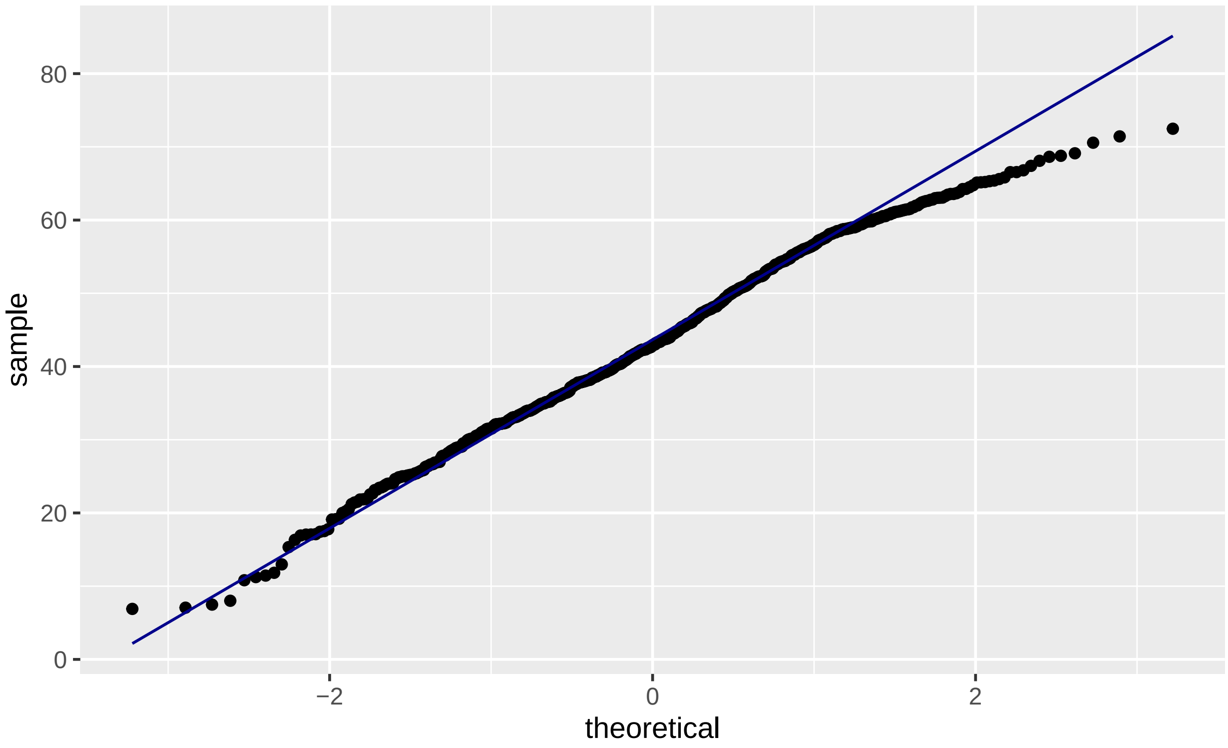
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.99216, p-value = 0.0003657
```

```
# Density-Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k010
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
```

```
    ),
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k010) %>% mean(),
      sd = Percentage %>% pull(Perc_k010) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```
# QQ-plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k010
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



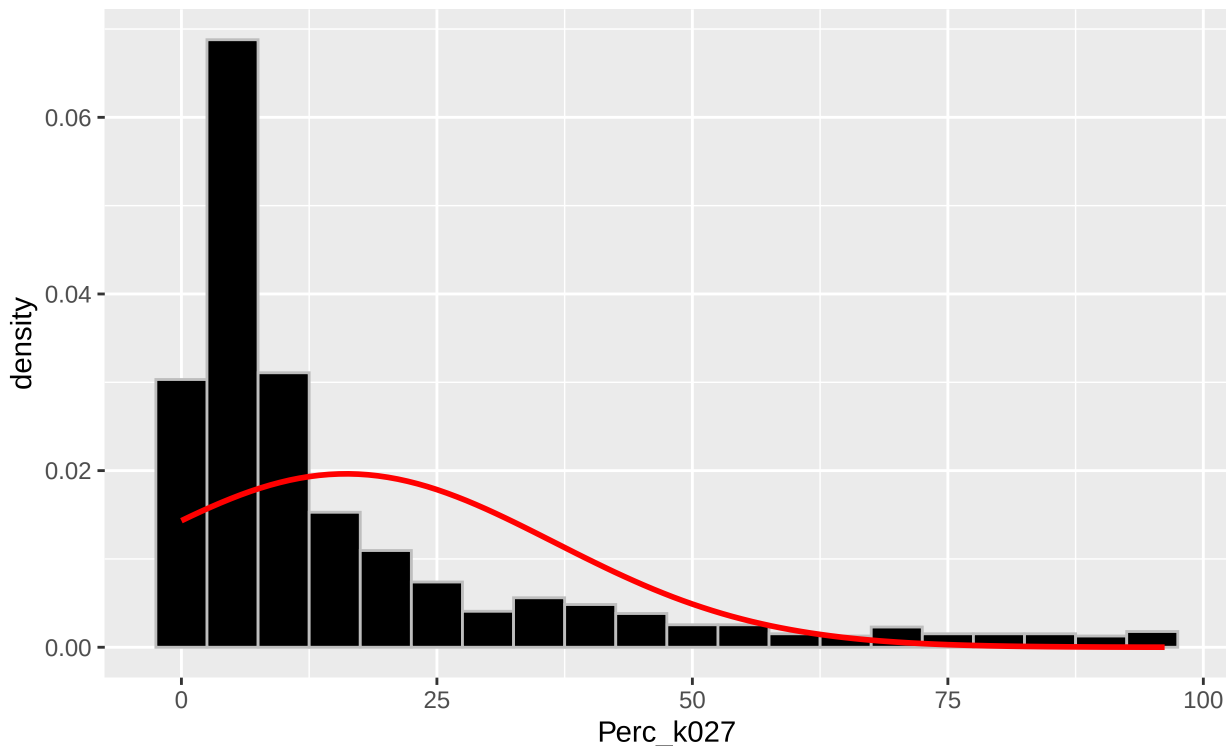
k027 - Households who live in a detached house or bungalow

```
# Shapiro-Test

Percentage %>%
  dplyr::pull(Perc_k027) %>%
  stats::shapiro.test()
```

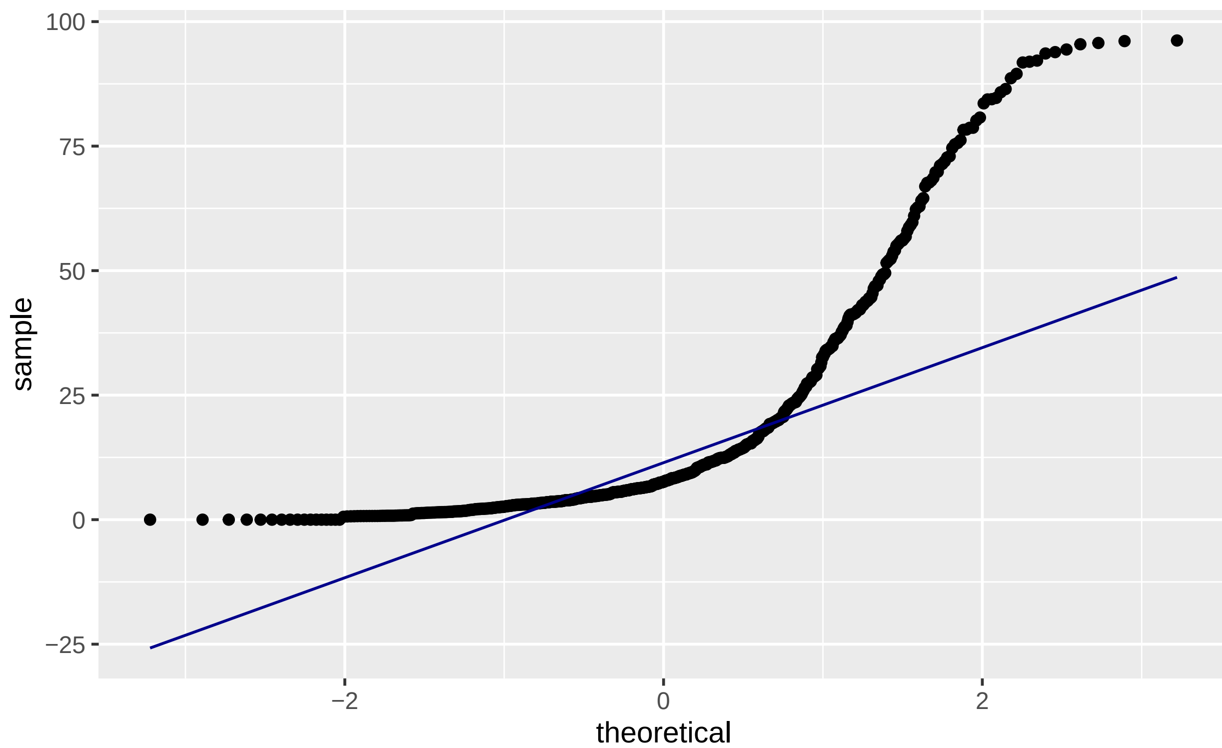
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.71265, p-value < 2.2e-16
```

```
# Density-Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k027
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
```

```
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k027) %>% mean(),
      sd = Percentage %>% pull(Perc_k027) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```
# QQ-plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k027
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



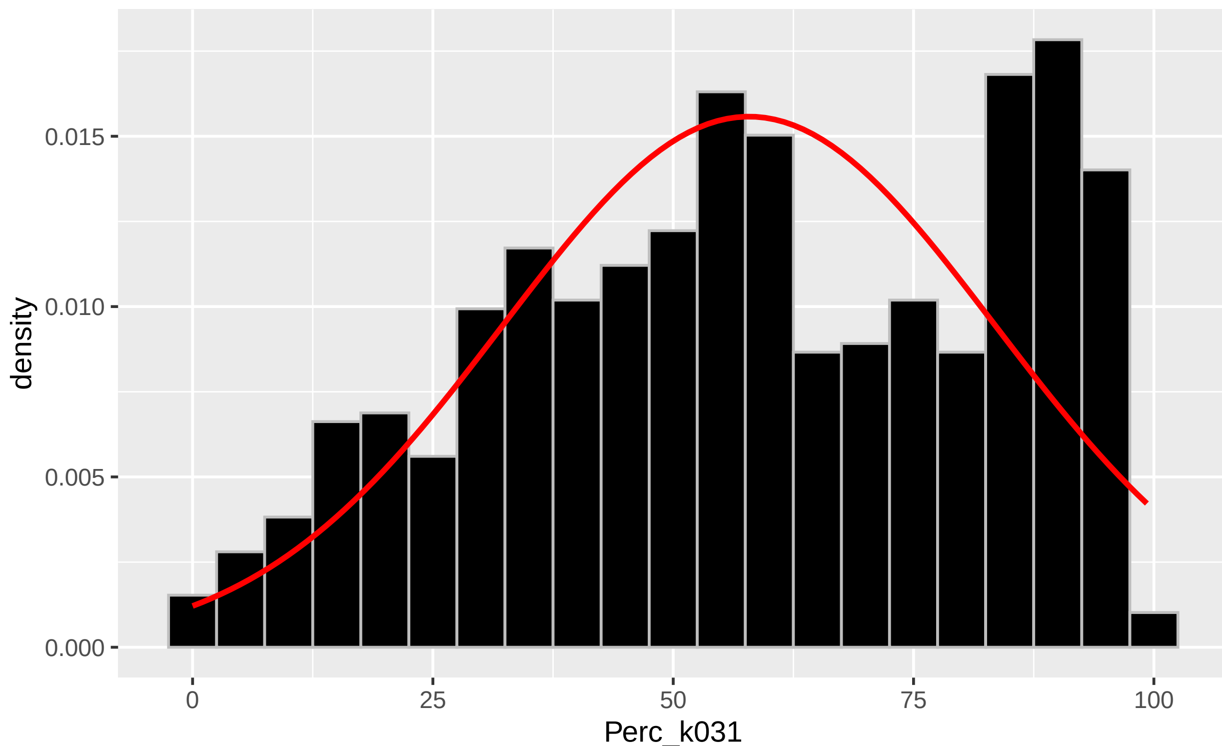
k031 - Households who own or have shared ownership of property

```
# Shapiro-Test

Percentage %>%
  dplyr::pull(Perc_k031) %>%
  stats::shapiro.test()
```

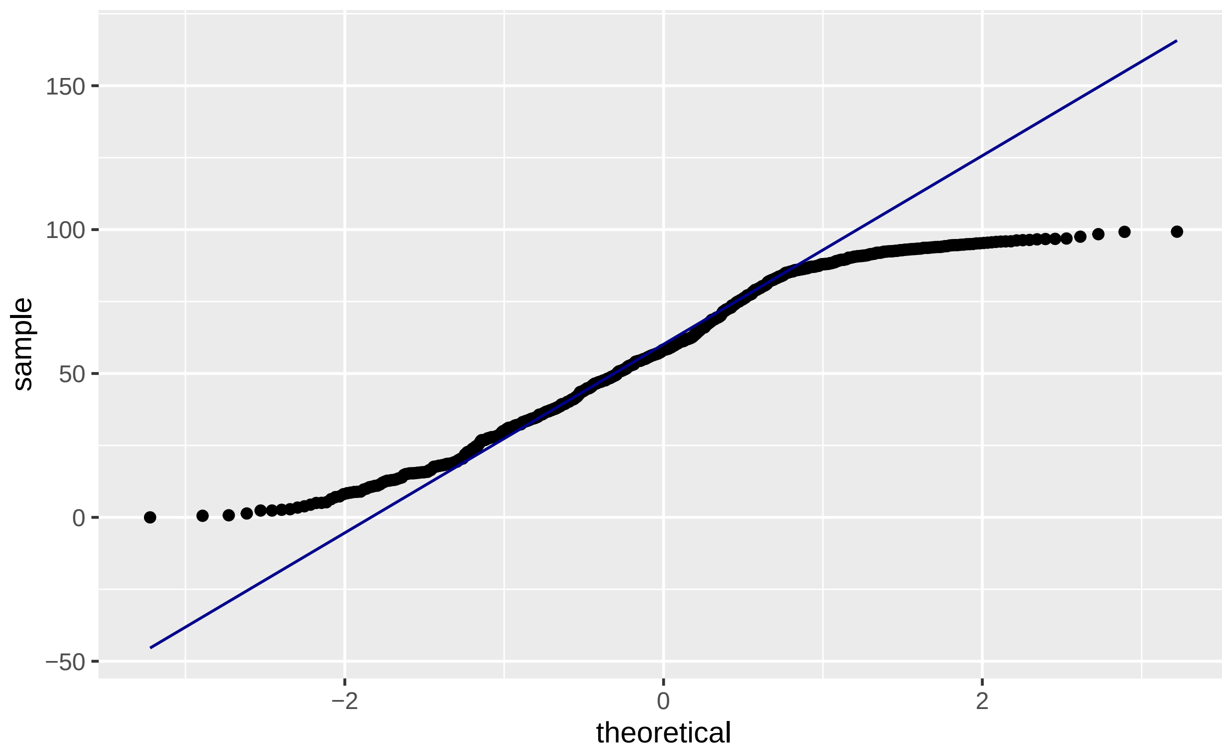
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95823, p-value = 3.72e-14
```

```
# Density-Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k031
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
```

```
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k031) %>% mean(),
      sd = Percentage %>% pull(Perc_k031) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```
# QQ-plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k031
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



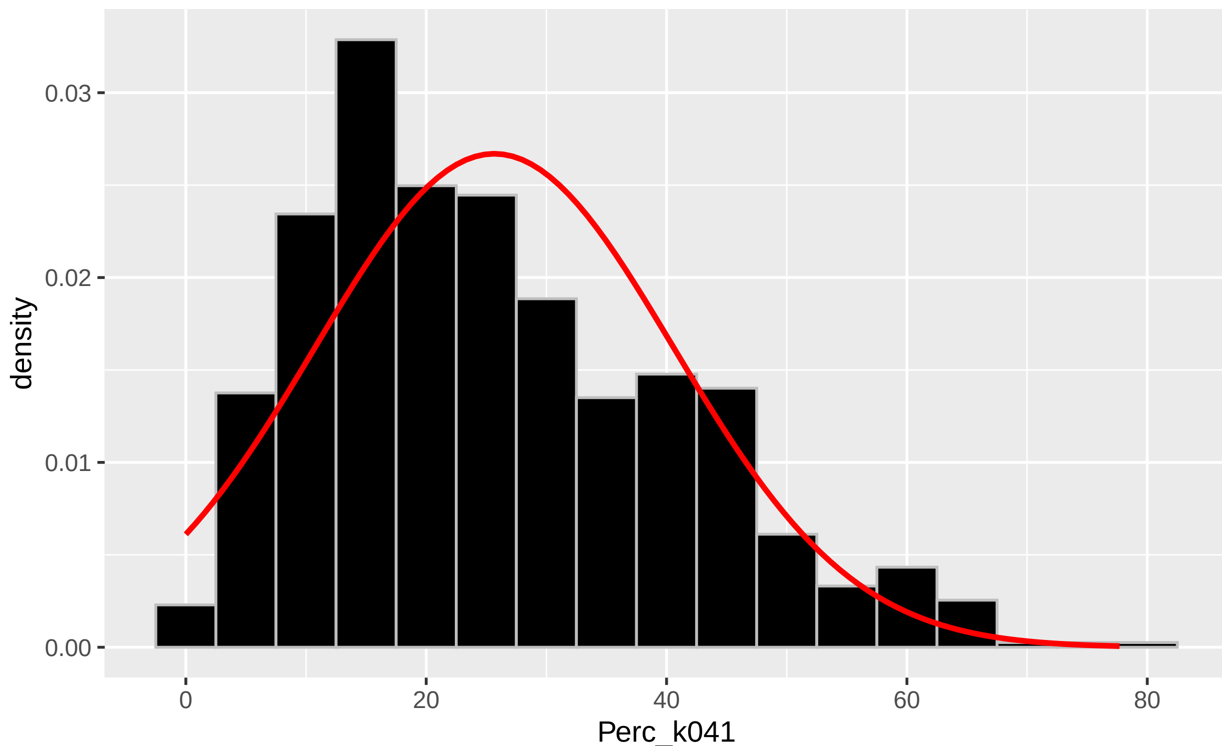
k041 - Households with two or more cars or vans

```
# Shapiro-Test

Percentage %>%
  dplyr::pull(Perc_k041) %>%
  stats::shapiro.test()
```

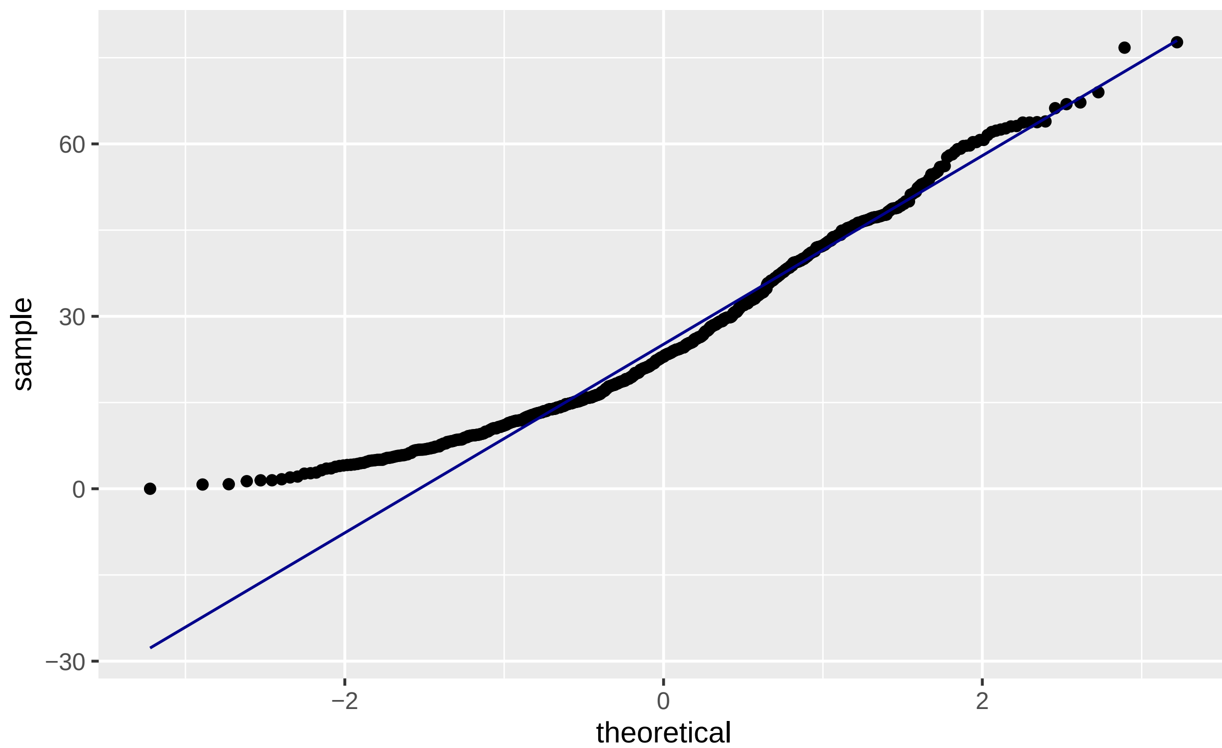
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95469, p-value = 7.991e-15
```

```
# Density-Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k041
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
```

```
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k041) %>% mean(),
      sd = Percentage %>% pull(Perc_k041) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```
# QQ-plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k041
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



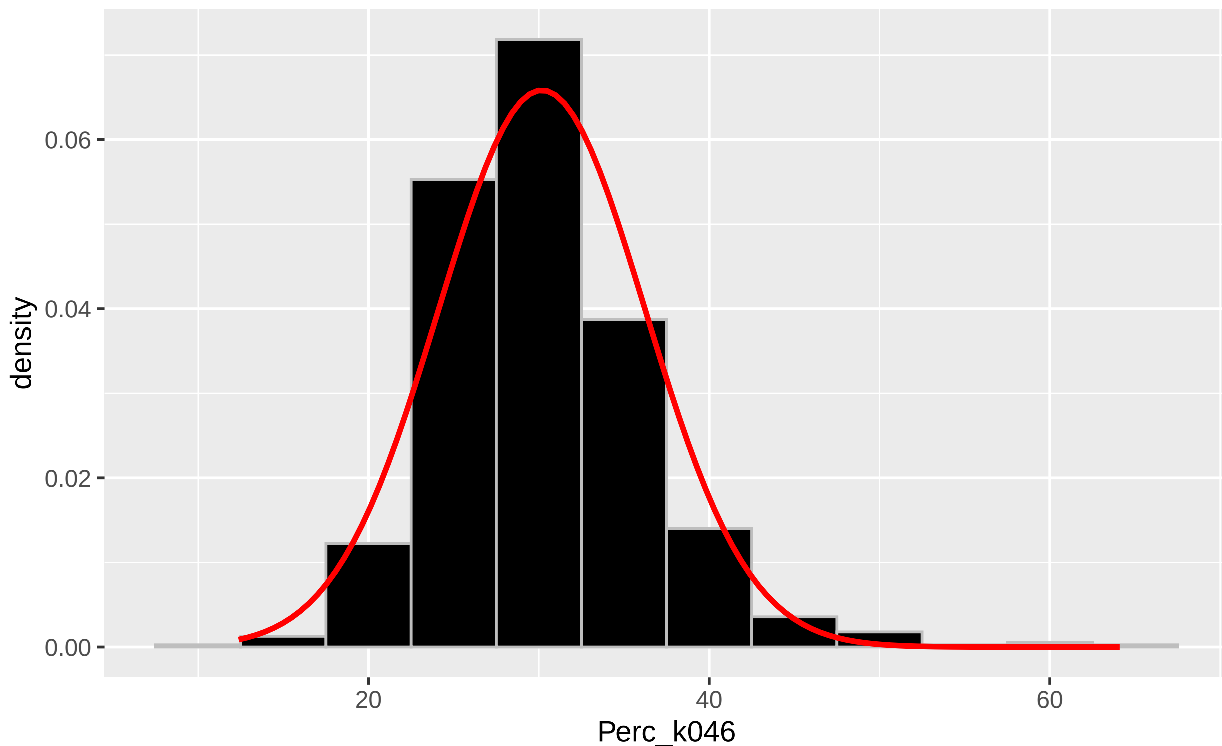
k046 - Employed persons aged between 16 and 74 who work part-time

```
# Shapiro-Test

Percentage %>%
  dplyr::pull(Perc_k046) %>%
  stats::shapiro.test()
```

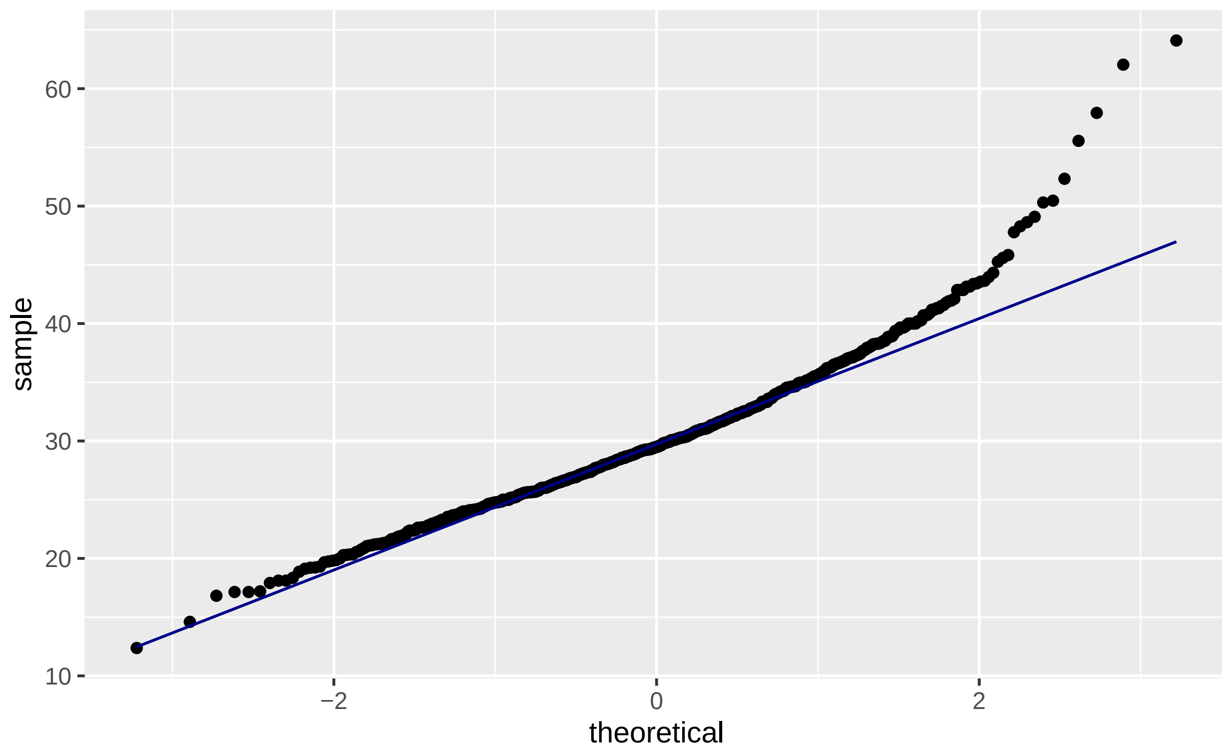
```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95548, p-value = 1.117e-14
```

```
# Density-Histogram

Percentage %>%
  ggplot2::ggplot(
    aes(
      x = Perc_k046
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
```

```r
    binwidth = 5,
    fill = "black",
    colour = "grey"
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = Percentage %>% pull(Perc_k046) %>% mean(),
      sd = Percentage %>% pull(Perc_k046) %>% sd()
    ),
    colour = "red", size = 1
  )
```




```r
# QQ-plot

Percentage %>%
  ggplot2::ggplot(
    aes(
      sample = Perc_k046
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(col = "darkblue")
```



Results and Discussion

Option A.2

Multiple Linear regression

###Select and normalize variables

```
library(stargazer)
library(lmtest)
library(car)
library(lm.beta)
```

```
# Selecting the dependent and independent variables

Wolverhampton_Household <-
  Wolverhampton_2011OAC %>%
  dplyr::select(
    OA, Total_Population, Total_Population_16_and_over, Total_Household_Spaces,
    Total_Households, Total_Employment_16_to_74,
    k004, k009, k010, k027, k031, k041, k046
  ) %>%

# percentage of dependent and independent variables

dplyr::mutate (
```

```r
  k004 = ( k004 / Total_Population) * 100,
  k009 = ( k009 / Total_Population_16_and_over) * 100,
  k010 = ( k010 / Total_Population_16_and_over) * 100,
  k027 = ( k027 / Total_Household_Spaces) * 100,
  k031 = ( k031 / Total_Households) * 100,
  k041 = ( k041 / Total_Households) * 100,
  k046 = ( k046 / Total_Employment_16_to_74) * 100
  ) %>%

#  rename columns

  dplyr::rename_with(
    function(x) {(paste0("Perc_", x))},
    c(k004, k009, k010, k027, k031, k041, k046)
  )
```

```r
# Selected variables

#Perc_k004 : Persons aged 45 to 64
#Perc_k009 : Persons aged over 16 who are single
#Perc_k010 : Persons aged over 16 who are married or in a registered same-sex civil partnership
#Perc_k027 : Households who live in a detached house or bungalow
#Perc_k041 : Households with two or more cars or vans
#Perc_k046 : Employed persons aged between 16 and 74 who work part-time

# create household model

Household_model <-
  Wolverhampton_Household %$%
  lm(
    Perc_k031 ~
      Perc_k004 + Perc_k009 + Perc_k010 + Perc_k027 + Perc_k041 + Perc_k046
    )
```

```r
#print summary
Household_model %>%
  summary()
```

```
##
## Call:
## lm(formula = Perc_k031 ~ Perc_k004 + Perc_k009 + Perc_k010 +
##     Perc_k027 + Perc_k041 + Perc_k046)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.758  -6.096   0.545   6.160  46.007
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.62323    6.02163   1.432 0.152533
## Perc_k004    0.30026    0.08081   3.716 0.000217 ***
## Perc_k009   -0.16804    0.07011  -2.397 0.016771 *
## Perc_k010    0.85996    0.07447  11.548  < 2e-16 ***
```

```
## Perc_k027   -0.15791    0.02447  -6.453 1.93e-10 ***
## Perc_k041    0.89511    0.05118  17.491  < 2e-16 ***
## Perc_k046   -0.30935    0.06346  -4.875 1.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.805 on 778 degrees of freedom
## Multiple R-squared:  0.8545, Adjusted R-squared:  0.8534
## F-statistic: 761.7 on 6 and 778 DF,  p-value: < 2.2e-16
```

```r
# Not rendered in bookdown
stargazer(Household_model, header=FALSE)
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & Perc\_k031 \\
## \hline \\[-1.8ex]
##  Perc\_k004 & 0.300$^{***}$ \\
##   & (0.081) \\
##   & \\
##  Perc\_k009 & $-$0.168$^{**}$ \\
##   & (0.070) \\
##   & \\
##  Perc\_k010 & 0.860$^{***}$ \\
##   & (0.074) \\
##   & \\
##  Perc\_k027 & $-$0.158$^{***}$ \\
##   & (0.024) \\
##   & \\
##  Perc\_k041 & 0.895$^{***}$ \\
##   & (0.051) \\
##   & \\
##  Perc\_k046 & $-$0.309$^{***}$ \\
##   & (0.063) \\
##   & \\
##  Constant & 8.623 \\
##   & (6.022) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 785 \\
## R$^{2}$ & 0.855 \\
## Adjusted R$^{2}$ & 0.853 \\
## Residual Std. Error & 9.805 (df = 778) \\
## F Statistic & 761.719$^{***}$ (df = 6; 778) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
```

```
## \end{tabular}
## \end{table}
```

```r
# Conduct shapiro-test for Households.
# Normality

Household_model %>%
  rstandard() %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.99156, p-value = 0.0001853
```

```r
# Homoscedasticity
# Breusch-Pagan test

Household_model %>%
  bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 32.155, df = 6, p-value = 1.524e-05
```

```r
# Independence
# Durbin-Watson test

Household_model %>%
  dwtest()
```

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.7526, p-value = 0.0002195
## alternative hypothesis: true autocorrelation is greater than 0
```

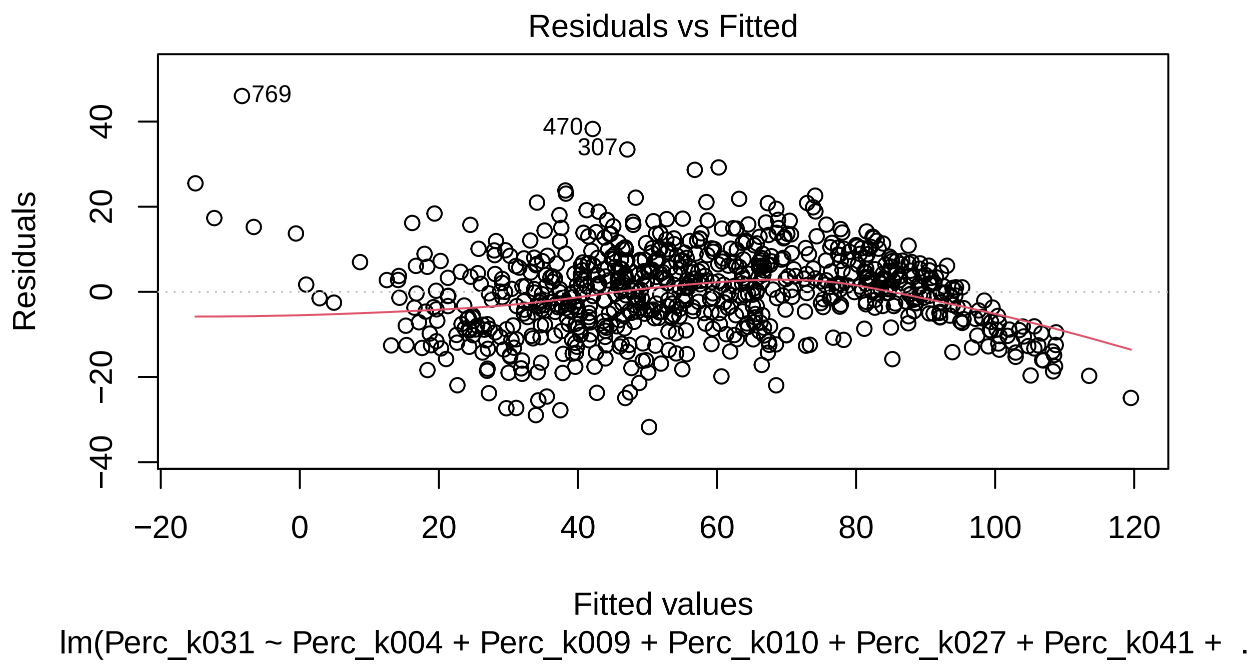
```r
# Conduct vif model

Household_model %>%
  vif()
```

```
## Perc_k004 Perc_k009 Perc_k010 Perc_k027 Perc_k041 Perc_k046
##  1.757381  4.400610  6.555691  2.016023  4.768866  1.205221
```

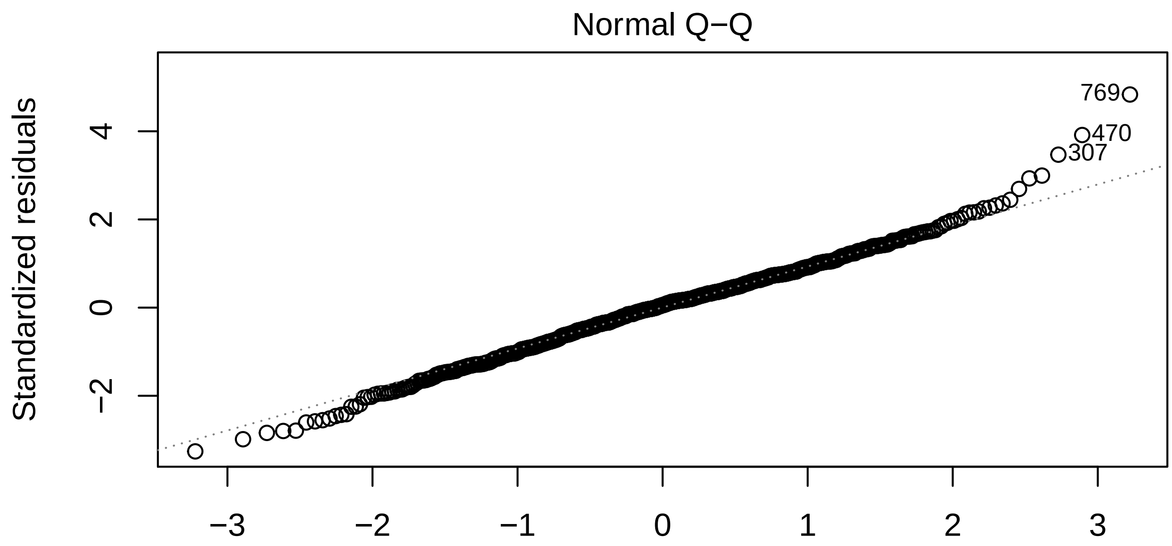
```r
# Conduct lm.beta

lm.beta(Household_model)
```

```
##
## Call:
## lm(formula = Perc_k031 ~ Perc_k004 + Perc_k009 + Perc_k010 +
##     Perc_k027 + Perc_k041 + Perc_k046)
##
## Standardized Coefficients::
## (Intercept)   Perc_k004   Perc_k009   Perc_k010   Perc_k027   Perc_k041
##  0.00000000  0.06735401 -0.06875274  0.40430088 -0.12527696  0.52228662
##   Perc_k046
## -0.07317581
```

```r
# Plotting residual to better understanding the variables
# Explore the residuals visually

# cook's distance c = 1

Household_model %>%
  plot(which = c(1))
```

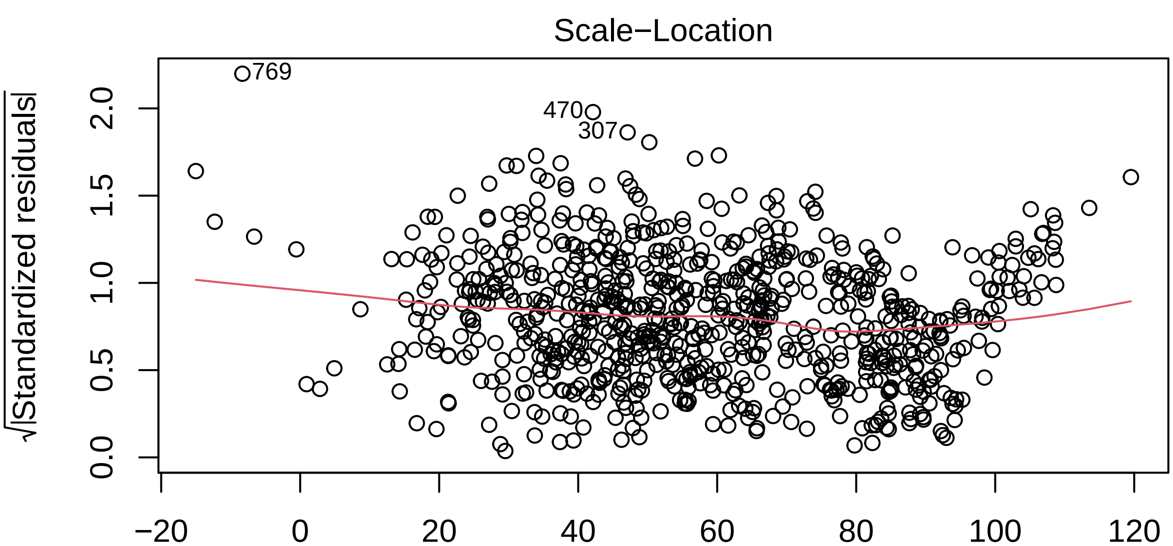



```r
# cook's distance c = 2

Household_model %>%
  plot(which = c(2))
```



lm(Perc_k031 ~ Perc_k004 + Perc_k009 + Perc_k010 + Perc_k027 + Perc_k041 + .

```r
# cook's distance c = 3

Household_model %>%
  plot(which = c(3))
```



lm(Perc_k031 ~ Perc_k004 + Perc_k009 + Perc_k010 + Perc_k027 + Perc_k041 + .

```r
# cook's distance c = 5

Household_model %>%
  plot(which = c(5))
```