GY7702_CW2_179024704

Jess Campbell

22/12/2020

This document was written in RMarkdown by opening a new R Markdown saved under GY7702_Coursework file and then Knit to a pdf file.

A GitHub was also created for this coursework.

GitHub link: https://github.com/jesslc54/GY7702_CW2

Option A

Question A.1

Conduct an exploratory analysis of the variables listed below, from the 2011_OAC_Raw_kVariables.csv dataset, for the OAs in the LAD assigned to you in the table in the Appendix. Include the code, the output (can include graphics) and a description of the findings. The latter should be up to 500 words and it can be written as a final discussion after the analysis, or as a description of each step of the analysis, or a combination of the two.

VariableCode= VariableName

k004= Persons aged 45 to 64

k009= Persons aged over 16 who are single

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

k027= Households who live in a detached house or bungalow

k031= Households who own or have shared ownership of property

k041= Households with two or more cars or vans

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

Variable codes used throughout instead of variable name as more concise.

Load libraries

```
# load tidyverse and knitr libraries
library(tidyverse)
library(knitr)
```

Load data

```
# upload .cvs files
# read and name the .cvs files
raw_2011OAC<- read_csv("2011_OAC_Raw_kVariables.csv")
census_OA <- read_csv("OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv")
```

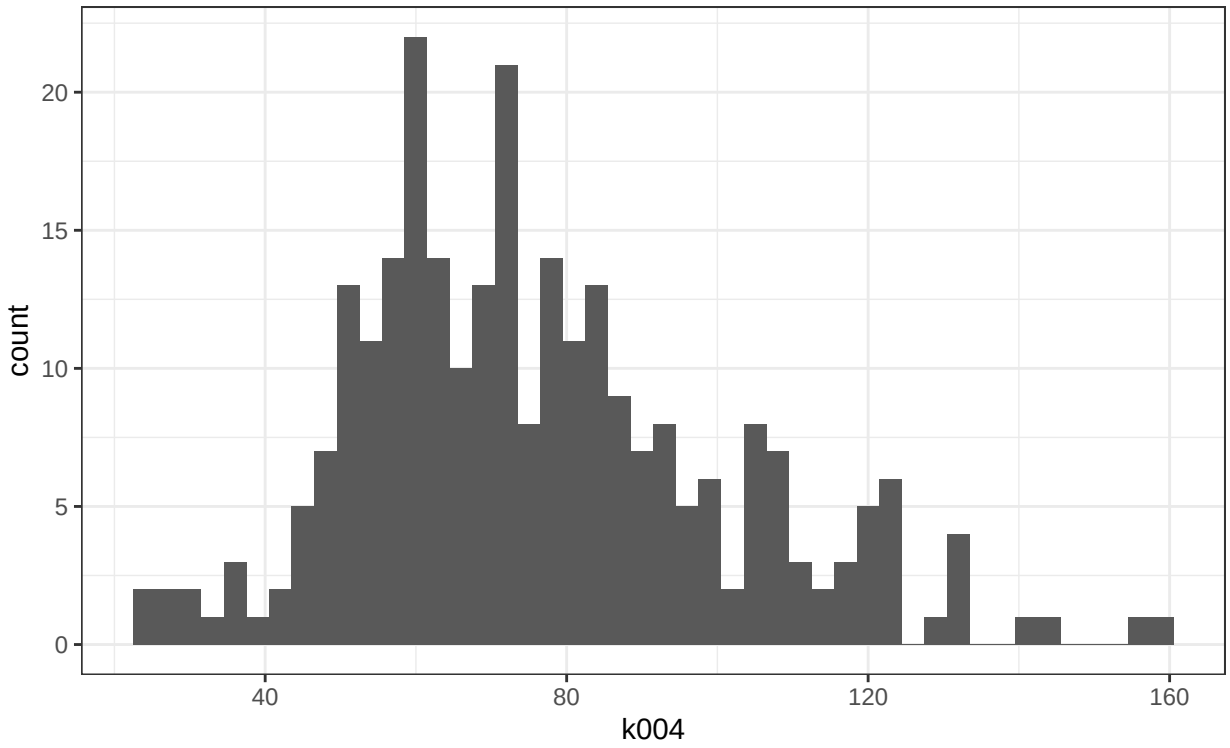
Join data and filter/ rename for Hyndburn

```
# rename resulting table
hyndburn_2011OAC <-
# full join tables by output area
  dplyr::full_join(
   census_OA, raw_2011OAC,
  by = c("OA11CD" = "OA")
)%>%
#filter to only show Hyndburn data
  dplyr::filter(
    LAD11NM == "Hyndburn"
  ) %>%
# select only relevent columns in table
  dplyr::select(
    "LAD11NM",
    "Total_Population",
    "Total_Households",
    "Total_Employment_16_to_74",
    "k004", "k009", "k010", "k027", "k031", "k041", "k046"
    )
```

Exploratory Data Analysis

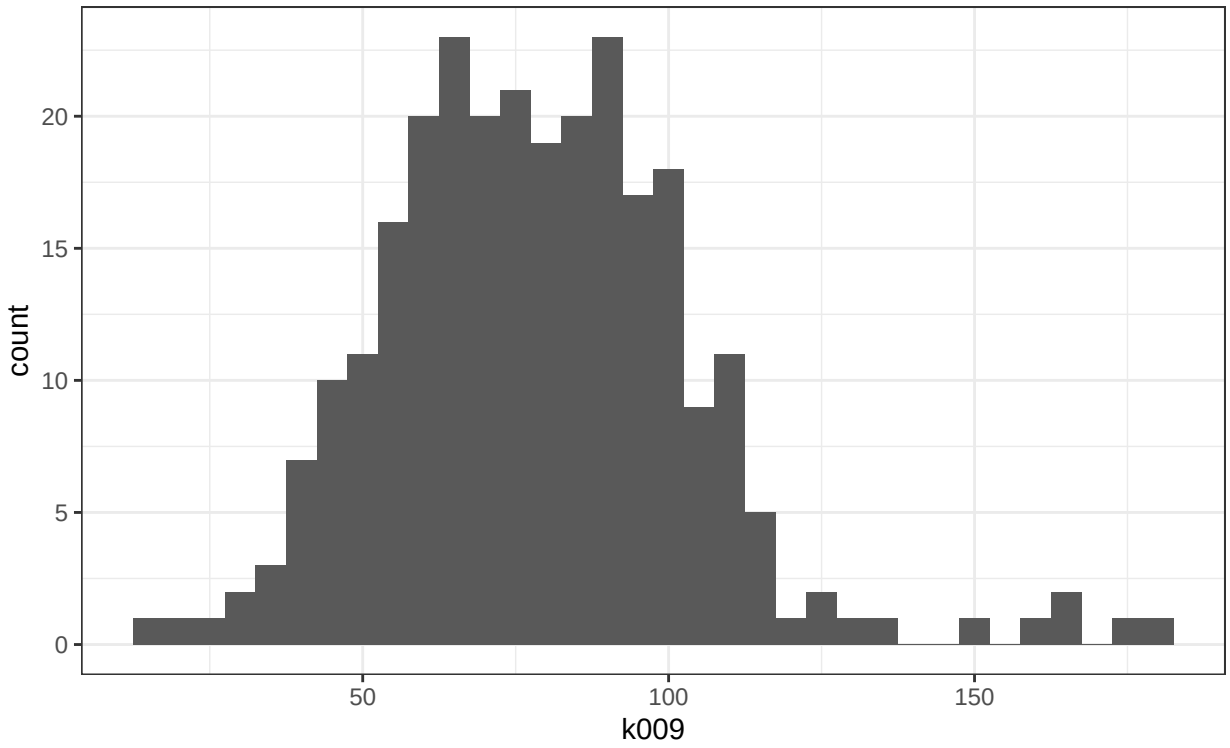
Plot for persons aged 45 to 64

```
#
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k004
    )
  ) +
  ggplot2::geom_histogram(binwidth = 3) +
  ggplot2::theme_bw()
```



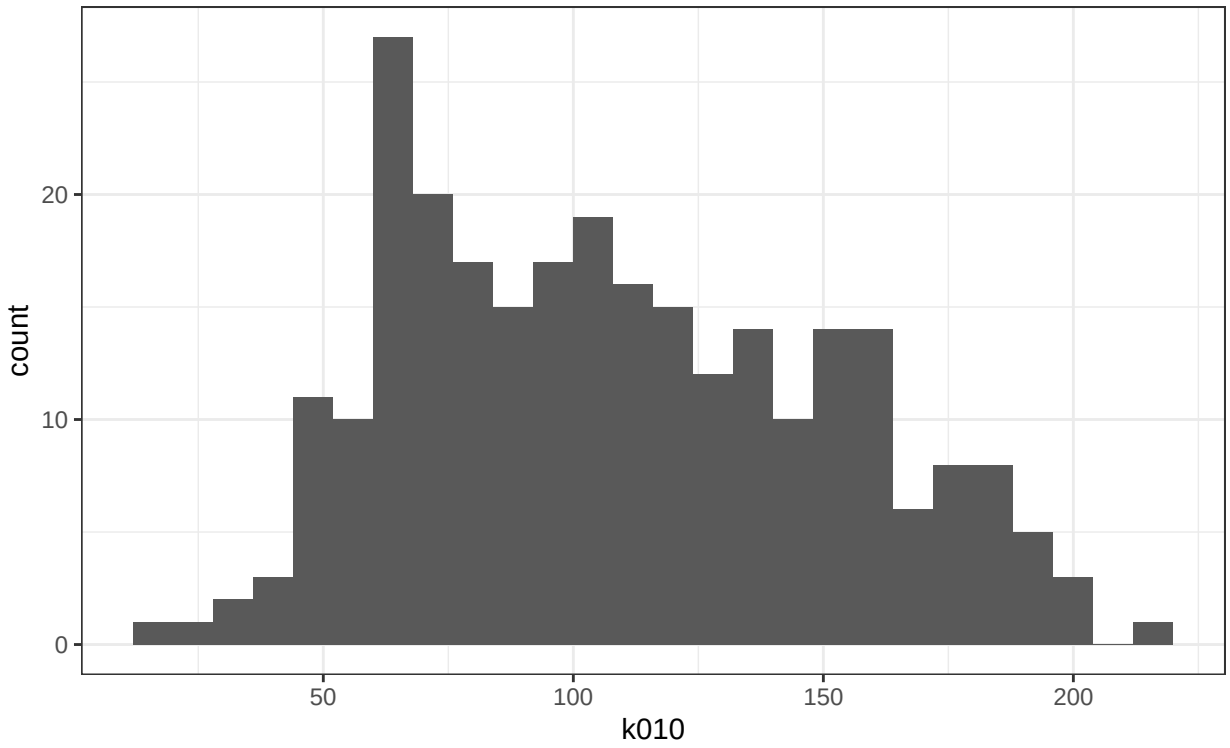
Plot for persons aged over 16 who are single

```
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k009
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::theme_bw()
```



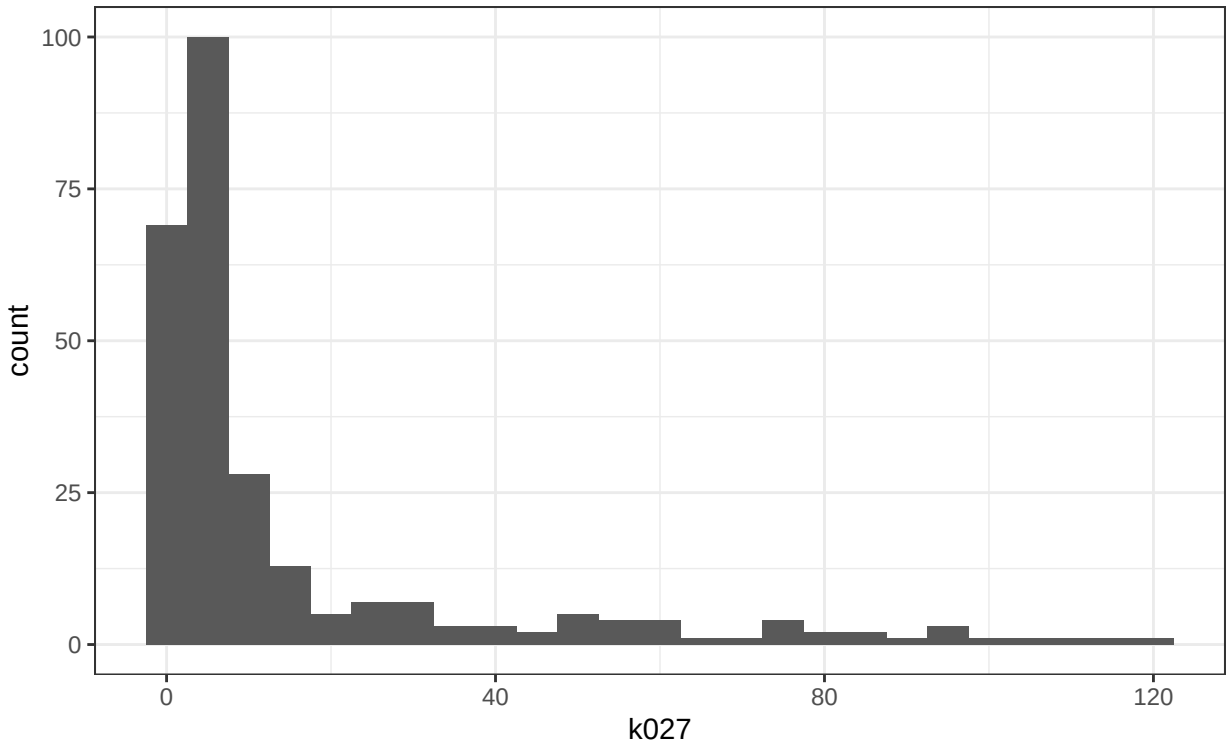
Plot for persons aged over 16 who are married or in a registered same-sex civil partnership

```
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k010
    )
  ) +
  ggplot2::geom_histogram(binwidth = 8) +
  ggplot2::theme_bw()
```



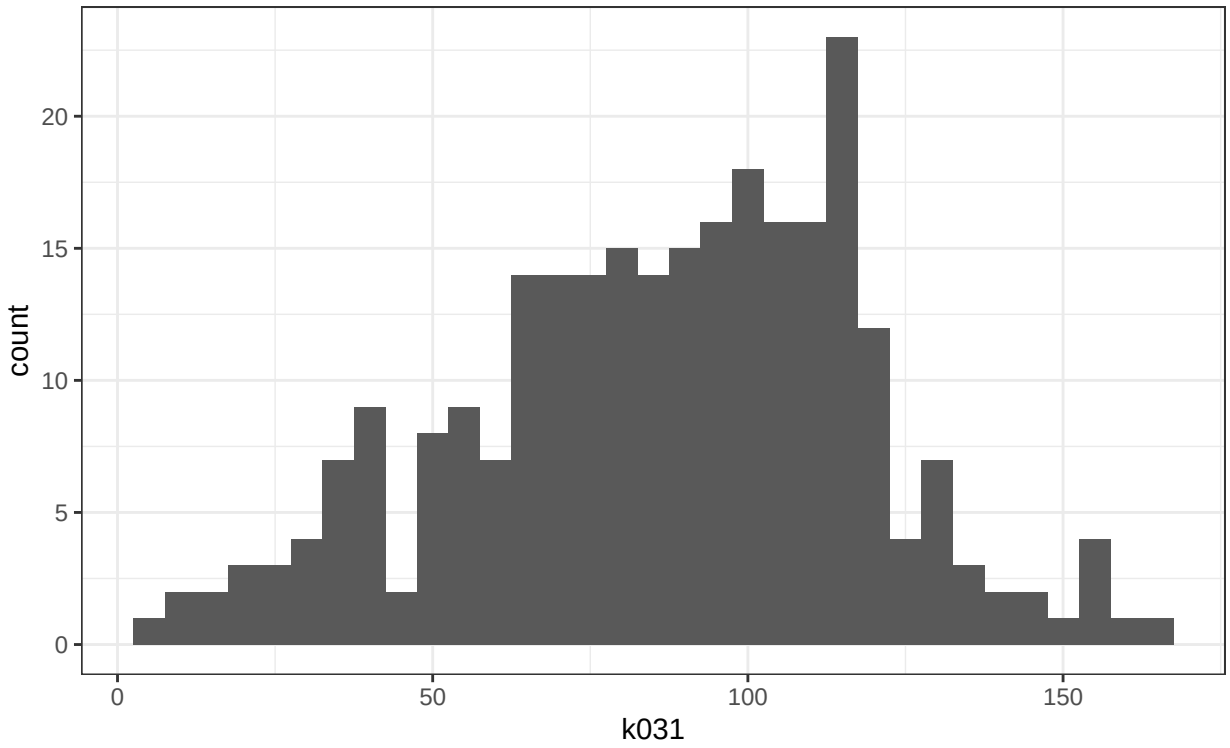
Plot for households who live in a detached house or bungalow

```
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k027
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::theme_bw()
```



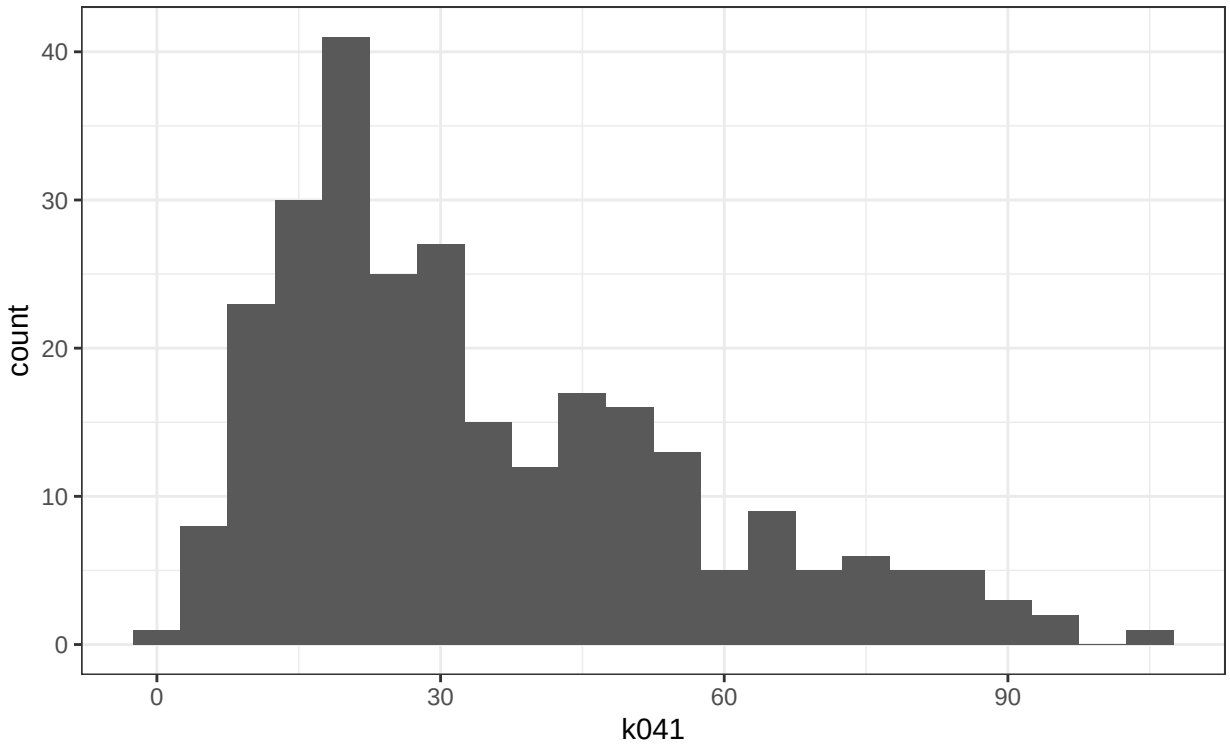
Plot for households who own or have shared ownership of property

```
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k031
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::theme_bw()
```



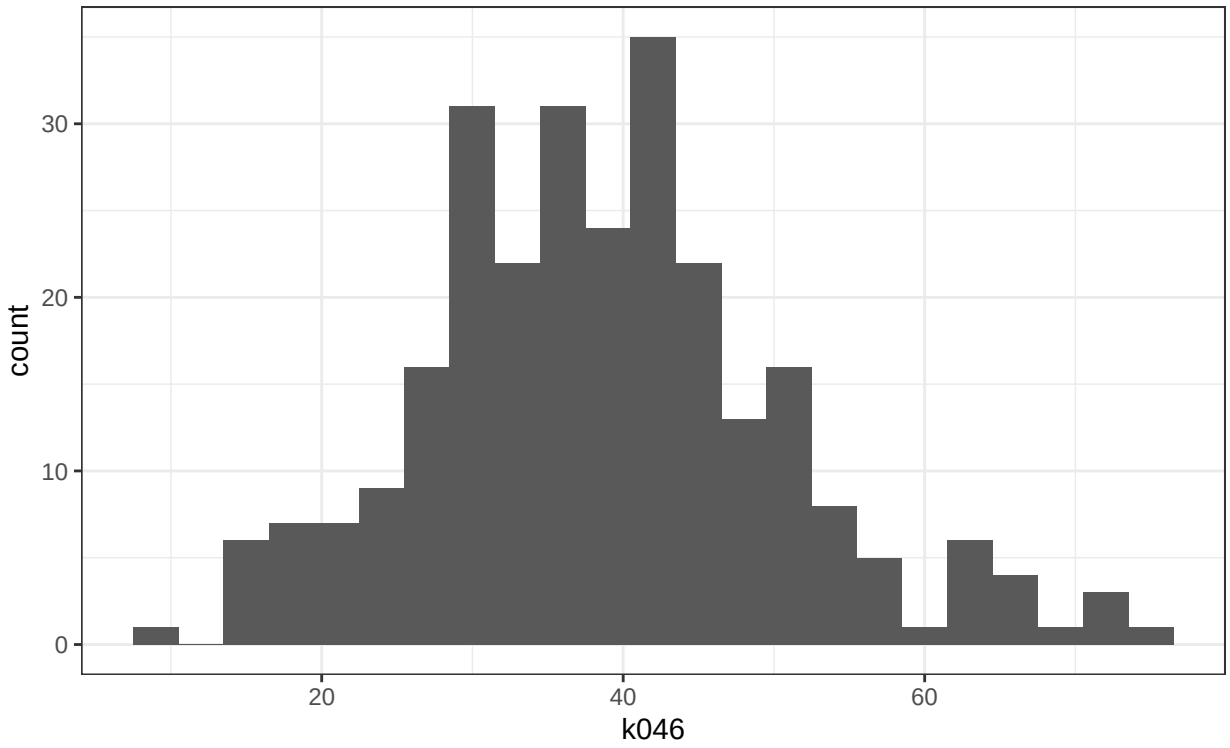
plot for households with two or more cars or vans

```
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k041
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::theme_bw()
```



Plot for employed persons aged between 16 and 74 who work part-time

```
hyndburn_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = k046
    )
  ) +
  ggplot2::geom_histogram(binwidth = 3) +
  ggplot2::theme_bw()
```



Desciptive statistic of each variable

```
hyndburn_2011OAC_stat_desc <- hyndburn_2011OAC %>%
  dplyr::select(k004, k009, k010, k027, k031, k041, k046) %>%
  pastecs::stat.desc(norm = TRUE)

hyndburn_2011OAC_stat_desc %>%
  knitr::kable(digits = 3)
```

	k004	k009	k010	k027	k031	k041	k046
nbr.val	269.000	269.000	269.000	269.000	269.000	269.000	269.000
nbr.null	0.000	0.000	0.000	7.000	0.000	0.000	0.000
nbr.na	0.000	0.000	0.000	0.000	0.000	0.000	0.000
min	24.000	15.000	19.000	0.000	6.000	2.000	9.000
max	160.000	181.000	219.000	120.000	163.000	104.000	76.000
range	136.000	166.000	200.000	120.000	157.000	102.000	67.000
sum	20502.000	21126.000	29368.000	4225.000	23322.000	9358.000	10394.000
median	72.000	77.000	103.000	5.000	90.000	28.000	38.000
mean	76.216	78.535	109.175	15.706	86.699	34.788	38.639
SE.mean	1.500	1.541	2.593	1.490	1.918	1.320	0.722
CI.mean.0.95	2.953	3.034	5.105	2.934	3.777	2.599	1.421
var	605.282	638.847	1808.137	597.447	989.816	468.750	140.209
std.dev	24.602	25.275	42.522	24.443	31.461	21.651	11.841
coef.var	0.323	0.322	0.389	1.556	0.363	0.622	0.306
skewness	0.654	0.753	0.313	2.293	-0.247	0.896	0.394
skew.2SE	2.201	2.534	1.054	7.719	-0.832	3.015	1.326
kurtosis	0.351	1.892	-0.826	4.664	-0.296	0.072	0.372

	k004	k009	k010	k027	k031	k041	k046
kurt.2SE	0.593	3.197	-1.396	7.880	-0.500	0.122	0.628
normtest.W	0.968	0.963	0.972	0.628	0.988	0.923	0.985
normtest.p	0.000	0.000	0.000	0.000	0.020	0.000	0.006

Question A.2

Use the variables explored in Question A.1 to create a robust, multiple linear regression model having as outcome (dependent) variable the presence (per OA in the LAD assigned to you in the table in the Appendix) of households who own or have shared ownership of property. Present the model that achieves the best fit and the process through which it has been identified. Include the code, the output (can include graphics), a final model. The latter two should be up to 500 words and it can be written as a final discussion after the analysis, or as a description of each step of the analysis, or a combination of the two.

property ownershipi = (model) + errori

Alternatively, if no robust model or no significant model can be created for the LAD assigned to you, include the code and the output (can include graphics) that illustrate that finding, and a related discussion (still, up to 500 words). The latter could be written as a final discussion after the analysis, or as a description of each step of the analysis, or a combination of the two

```
# load magrittr and stargazer library
library(magrittr)
library(stargazer)
library(lmtest)
library(lm.beta)
```

Multiple Regression Model and Analysis

```
hyndburn_2011OAC_norm <-
  hyndburn_2011OAC %>%
  dplyr::select(
Total_Population, Total_Households, Total_Employment_16_to_74,
k031, k004, k009, k010, k027, k041, k046
  ) %>%
  # percentage number of households
  dplyr::mutate(
    k027 = (k027 / Total_Households) * 100
  ) %>%
  # percentage across household columns
  dplyr::mutate(
    dplyr::across(
      k031:k041,
```

```
      function(x){ (x / Total_Households) * 100 }
    )
  ) %>%
# percentage number of population
  dplyr::mutate(
    k004 = (k004 / Total_Population) * 100
  ) %>%
  # percentage across population columns
  dplyr::mutate(
    dplyr::across(
      k009:k010,
      function(x){ (x / Total_Population) * 100 }
    )
  ) %>%
# percentage number of employment
  dplyr::mutate(
    k046 = (k046/ Total_Employment_16_to_74) * 100
  ) %>%
  # rename columns
  dplyr::rename_with(
    function(x){ paste0("perc_", x) },
    c(k027, k031, k041, k004, k009, k010, k046)
  )
```

```
hyndburn_household_model <-
  hyndburn_2011OAC_norm %$%
  lm(
    perc_k031 ~
    perc_k004 + perc_k009 + perc_k010 + perc_k027 + perc_k041 + perc_k046
  )
hyndburn_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
## -34.341  -7.651   1.047   7.726  27.813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.80473    6.37141   6.561 2.84e-10 ***
## perc_k004   -0.45974    0.15929  -2.886 0.004226 **
## perc_k009   -0.03426    0.13339  -0.257 0.797483
## perc_k010    0.85316    0.10020   8.515 1.32e-15 ***
## perc_k027   -0.39933    0.07820  -5.107 6.32e-07 ***
## perc_k041    1.04552    0.08873  11.783  < 2e-16 ***
## perc_k046   -0.43627    0.12224  -3.569 0.000426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11.45 on 262 degrees of freedom
## Multiple R-squared:  0.7456, Adjusted R-squared:  0.7397
## F-statistic:   128 on 6 and 262 DF,  p-value: < 2.2e-16
```

```
# shapiro-wilks test
hyndburn_household_model %>%
  stats::rstandard() %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.98771, p-value = 0.02157
```

```
# breusch-pagan test
hyndburn_household_model %>%
  lmtest::bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 27.524, df = 6, p-value = 0.0001155
```

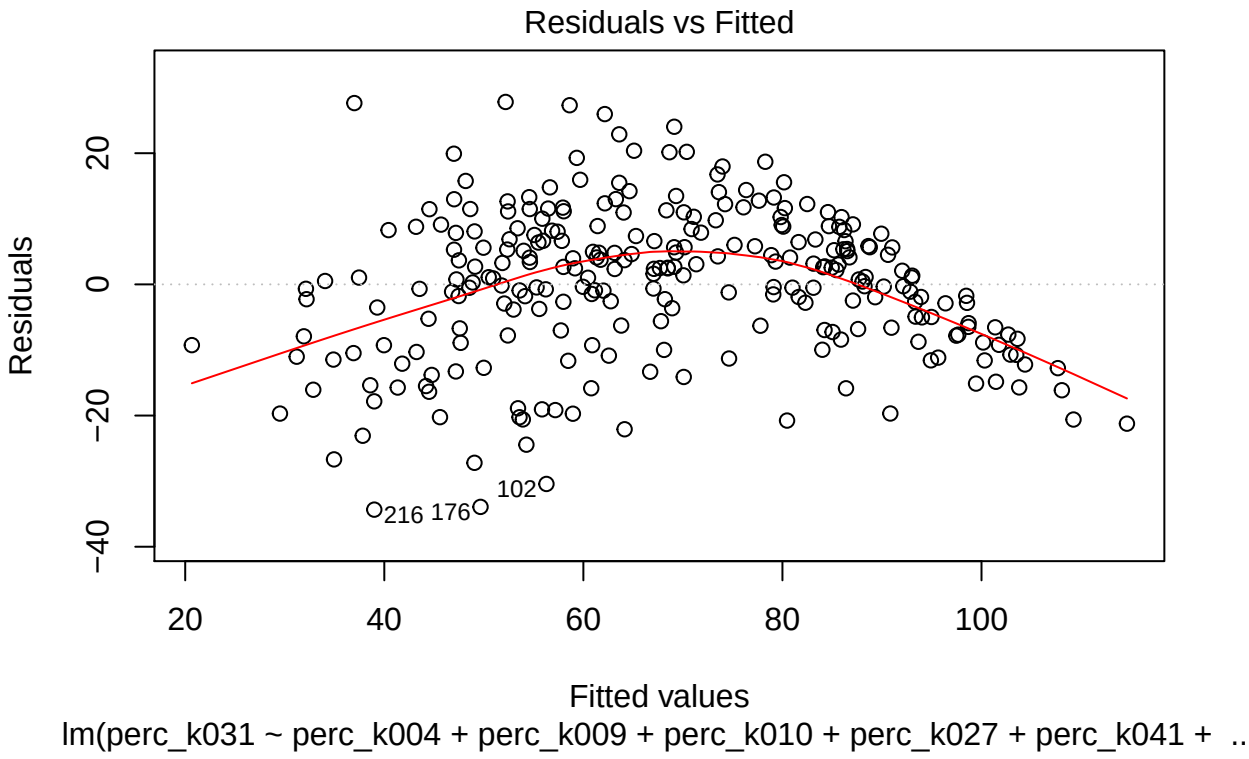
```
# durbin-watson test
hyndburn_household_model %>%
  lmtest::dwtest()
```

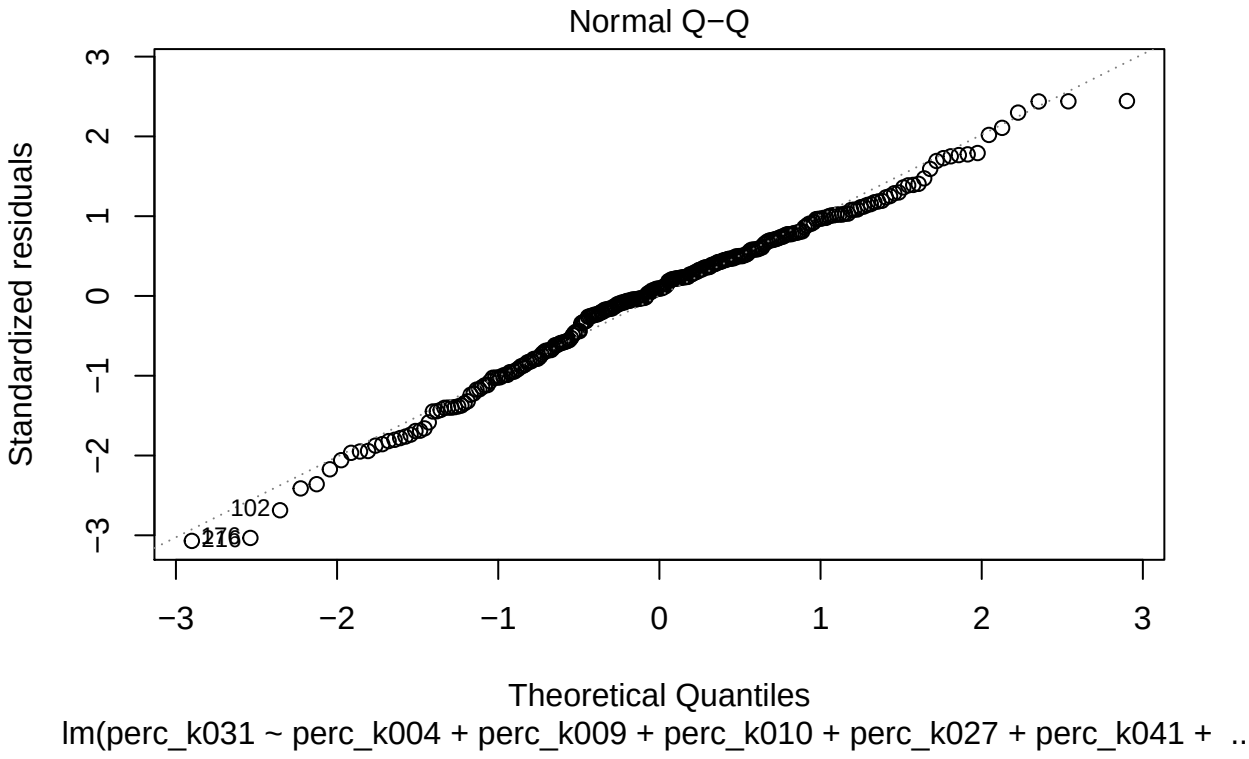
```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.8661, p-value = 0.1206
## alternative hypothesis: true autocorrelation is greater than 0
```

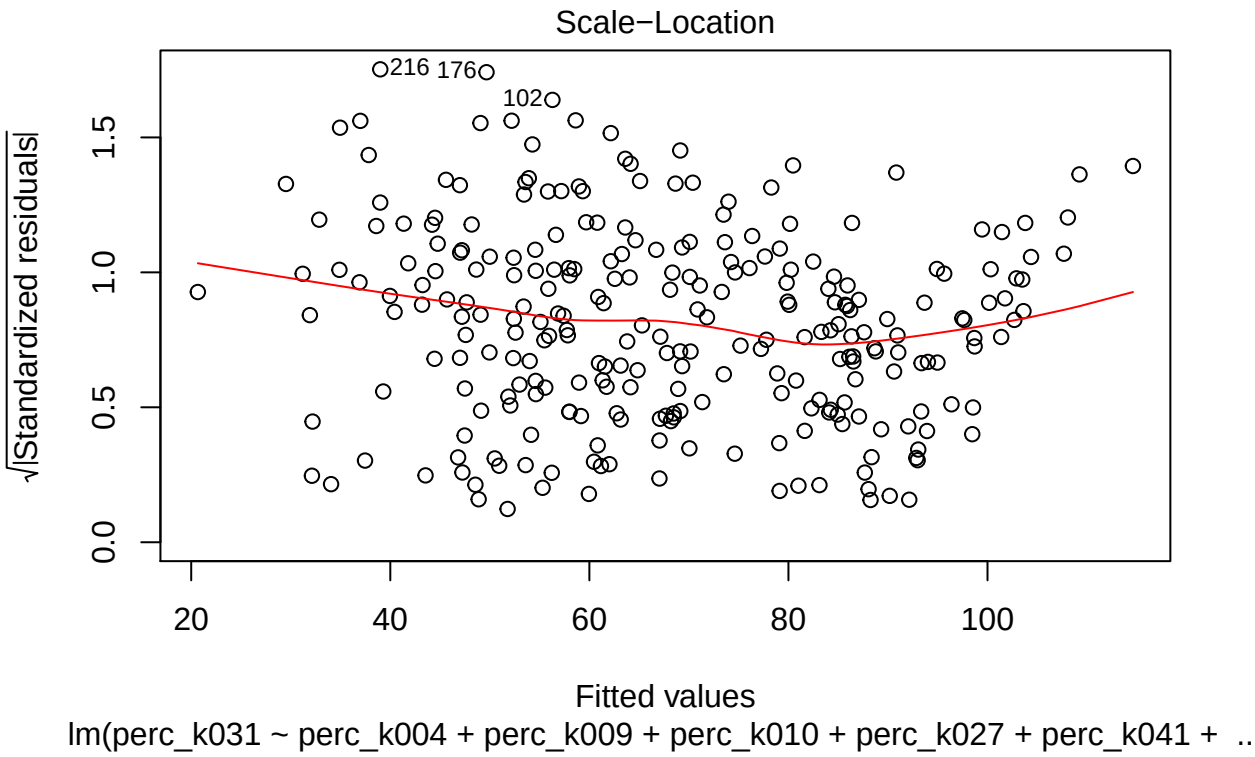
Discussion

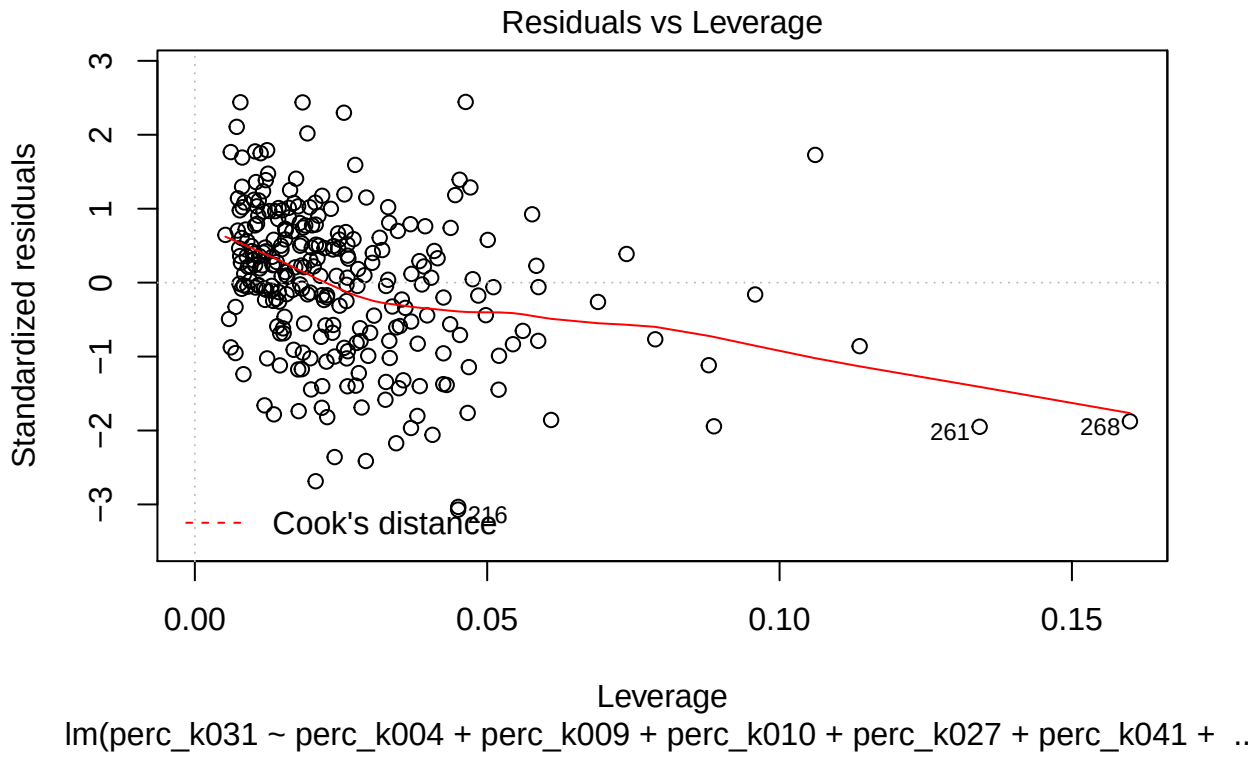
The output above suggests that the model is fit (F(6,262)= 128, p<.001) where the p value is significant, indicating that a model based on households, population and emplyment variables can account for 73.97% of households who own or have shared ownership of property. However the model is only partially robust. The residuals are normally distributed (Shapiro-Wilk test, W= 0.98, p= 0.02) but the residuals don't satisfy the homoscedasticity assumption (Breusch-Pagan test, BP=27.5, p= <.001), nor the independence assumption (Durbin-Watson test, DW=1.86, p <.001)

```
hyndburn_household_model %>%
  plot()
```


Residuals vs Fitted
Residuals
20
0
−20
−40
20
40
60
80
100
216
176
102
Fitted values
lm(perc_k031 ~ perc_k004 + perc_k009 + perc_k010 + perc_k027 + perc_k041 + ..


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
102
176
216
Theoretical Quantiles
lm(perc_k031 ~ perc_k004 + perc_k009 + perc_k010 + perc_k027 + perc_k041 + ..


Scale−Location
216
176
102
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
40
60
80
100
Fitted values
lm(perc_k031 ~ perc_k004 + perc_k009 + perc_k010 + perc_k027 + perc_k041 + ..



normal Q-Q plot shows high normality of the residuals

```
lm.beta(hyndburn_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
##  0.000000000 -0.121120894 -0.009619977  0.392309107 -0.250214222
##    perc_k041    perc_k046
##  0.742177820 -0.134997572
```

perc_k010 and perc_041 standardized coefficients shows that if persons aged over 16 who are married or in a registered same-sex civil partnership; or Households with two or more cars or vans increased then so would the households who own or have shared ownership of property, whereas the other variables would have a negative impact.

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
```

```
##  Max.   :25.0   Max.   :120.00
```