R Project Team 7

Arpit, Swathi, Arvind, Ritu

25 November 2016

Introduction

For the project we had to explore the posibility of AEDC variable % Developmentally vulnerable on two or more domains and use the following predictores :

- Internet Access at home
- Education

Methodology

We read the 3 excel sheets which had the required data and then selected only the variables which we thought are relevant for the analysis.

From AEDC sheet:

- % Children developmentally vulnerable on two or more domains We took percentage instead of the number of children as they are redundant data.

From Education sheet:

- % full-time participation at age 16
- ASR per 100

From Internet Access at home sheet:

- % dwellings with no Internet connection
- % dwellings with Internet connections
- % dwellings with Broadband Internet
- % dwellings with Dial-up Internet
- % dwellings with other Internet connections

We looked at number of NAs for each row and found that a row had 9 NAs and the total columns were 12. Keeping this row did not make sense and thus we deleted this row.

Issues

The issues we faced were with respect to what predictors to choose and how to deal with these variables. The other issue was to decide which variables to pick for the final model from the results we obtained from our correlation analysis.

Resolution

In order to keep the model relevant, We also attempted to take one variable from internet as well as education to make the fianl model more relevant to our problem statement.

Business Implications

This helped to understand the relation between internet access and education and what impact it has on % developmentally vulnerable on 2 or more domains. This helps to explain the how related these variables are and how varied levels of education or having internet access or not can impact our response variable.

Analysis

1. Data preparation and clean-up

```
# Reading data from excel file.
xl_workbook <- loadWorkbook("phidu_data_lga_sa.xls")
AEDC <- readWorksheet(xl_workbook, sheet = "Early_childhood_development",
                      header = TRUE, startRow = 5, endRow = 76)

str(AEDC)
```

```
## 'data.frame':    71 obs. of  69 variables:
##  $ Code                                                             : num  40070 40120
 40220 40250 40310 ...
##  $ Name                                                             : chr  "Adelaide
 (C)" "Adelaide Hills (DC)" "Alexandrina (DC)" "Anangu Pitjantjatjara (AC)" ...
##  $ Children.developmentally.vulnerable.on.one.or.more.domains           : chr  "28" "71" "2
7" "36" ...
##  $ Children.assessed.in.AEDC..first.year.of.school.                     : chr  "82" "410"
 "184" "45" ...
##  $ X..Children.developmentally.vulnerable.on.one.or.more.domains        : chr  "34.1" "17.
3" "14.7" "80.0" ...
##  $ Col6                                                             : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.vulnerable.on.two.or.more.domains           : chr  "12" "28" "1
1" "31" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..1                   : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.vulnerable.on.two.or.more.domains        : chr  "14.5" "6.8"
"5.9" "68.9" ...
##  $ Col10                                                            : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.vulnerable.in.physical.domain               : chr  "8" "24" "9"
"24" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..2                   : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.vulnerable.in.physical.domain            : chr  "9.6" "5.8"
 "4.9" "53.3" ...
##  $ Col14                                                            : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.at.risk.in.physical.domain                  : chr  "9" "56" "1
7" "8" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..3                   : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.at.risk.in.physical.domain               : chr  "10.8" "13.
6" "9.2" "17.8" ...
##  $ Col18                                                            : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.on.track.in.physical.domain                 : chr  "66" "331"
 "159" "13" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..4                   : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.on.track.in.physical.domain              : chr  "79.5" "80.
5" "85.9" "28.9" ...
##  $ Col22                                                            : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.vulnerable.in.social.domain                 : chr  "8" "26" "1
2" "21" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..5                   : chr  "83" "410"
 "185" "45" ...
##  $ X..Children.developmentally.vulnerable.in.social.domain              : chr  "9.6" "6.3"
 "6.5" "46.7" ...
##  $ Col26                                                            : logi  NA NA NA NA
NA NA ...
```

```
##  $ Children.developmentally.at.risk.in.social.domain                      : chr  "14" "62" "2
0" "11" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..6                     : chr  "83" "410"
 "185" "45" ...
##  $ X..Children.developmentally.at.risk.in.social.domain                   : chr  "16.9" "15.
1" "10.8" "24.4" ...
##  $ Col30                                                                  : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.on.track.in.social.domain                     : chr  "61" "322"
 "153" "13" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..7                     : chr  "83" "410"
 "185" "45" ...
##  $ X..Children.developmentally.on.track.in.social.domain                  : chr  "73.5" "78.
5" "82.7" "28.9" ...
##  $ Col34                                                                  : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.vulnerable.in.emotional.domain                : chr  "9" "28" "1
4" "24" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..8                     : chr  "82" "413"
 "185" "44" ...
##  $ X..Children.developmentally.vulnerable.in.emotional.domain             : chr  "11.0" "6.8"
"7.6" "54.5" ...
##  $ Col38                                                                  : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.at.risk.in.emotional.domain                   : chr  "17" "50" "1
5" "4" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..9                     : chr  "82" "413"
 "185" "44" ...
##  $ X..Children.developmentally.at.risk.in.emotional.domain                : chr  "20.7" "12.
1" "8.1" "9.1" ...
##  $ Col42                                                                  : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.on.track.in.emotional.domain                  : chr  "56" "335"
 "156" "16" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..10                    : chr  "82" "413"
 "185" "44" ...
##  $ X..Children.developmentally.on.track.in.emotional.domain               : chr  "68.3" "81.
1" "84.3" "36.4" ...
##  $ Col46                                                                  : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.vulnerable.in.language.and.cognitive.domain   : chr  "12" "16"
 "4" "28" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..11                    : chr  "83" "413"
 "184" "45" ...
##  $ X..Children.developmentally.vulnerable.in.language.and.cognitive.domain: chr  "14.5" "3.9"
"2.2" "62.2" ...
##  $ Col50                                                                  : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.at.risk.in.language.and.cognitive.domain      : chr  "5" "35" "1
9" "2" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..12                    : chr  "83" "413"
 "184" "45" ...
##  $ X..Children.developmentally.at.risk.in.language.and.cognitive.domain   : chr  "6.0" "8.5"
 "10.3" "4.4" ...
```

```
##  $ Col54                                                                    : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.on.track.in.language.and.cognitive.domain     : chr  "66" "362"
 "161" "15" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..13                    : chr  "83" "413"
 "184" "45" ...
##  $ X..Children.developmentally.on.track.in.language.and.cognitive.domain  : chr  "79.5" "87.
7" "87.5" "33.3" ...
##  $ Col58                                                                    : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.vulnerable.in.communication.domain            : chr  "13" "23"
 "4" "22" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..14                    : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.vulnerable.in.communication.domain         : chr  "15.7" "5.6"
"2.2" "48.9" ...
##  $ Col62                                                                    : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.at.risk.in.communication.domain               : chr  "13" "48" "1
5" "7" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..15                    : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.at.risk.in.communication.domain            : chr  "15.7" "11.
7" "8.1" "15.6" ...
##  $ Col66                                                                    : logi  NA NA NA NA
NA NA ...
##  $ Children.developmentally.on.track.in.communication.domain              : chr  "57" "340"
 "166" "16" ...
##  $ Children.assessed.in.AEDC..first.year.of.school..16                    : chr  "83" "411"
 "185" "45" ...
##  $ X..Children.developmentally.on.track.in.communication.domain           : chr  "68.7" "82.
7" "89.7" "35.6" ...
```

```
dim(AEDC)
```

```
## [1] 71 69
```

```
education <- readWorksheet(xl_workbook, sheet = "Education",
                           header = TRUE, startRow = 5, endRow = 76)

str(education)
```

```
## 'data.frame':    71 obs. of  10 variables:
##  $ Code                                : num  40070 40120 40220 40250 40310 ...
##  $ Name                                : chr  "Adelaide (C)" "Adelaide Hills (DC)" "Alexandri
na (DC)" "Anangu Pitjantjatjara (AC)" ...
##  $ Full.time.participation.at.age.16   : chr  "64" "551" "238" "42" ...
##  $ People.aged.16                      : chr  "77" "616" "301" "52" ...
##  $ X..full.time.participation.at.age.16: chr  "83.1" "89.4" "79.1" "80.8" ...
##  $ Col6                                : chr  NA NA NA NA ...
##  $ Number                              : chr  "1,818" "7,020" "7,191" "1,222" ...
##  $ ASR.per.100                         : chr  "11.7" "21.1" "31.5" "80.9" ...
##  $ SR                                  : chr  "34" "62" "92" "236" ...
##  $ Sig.                                : chr  "**" "**" "**" "**" ...
```

```
dim(education)
```

```
## [1] 71 10
```

```
internet_access <- readWorksheet(xl_workbook, sheet = "Internet_access",
                                 header = TRUE, startRow = 5, endRow = 76)
str(internet_access)
```

```
## 'data.frame':    71 obs. of  21 variables:
##  $ Code                                             : num  40070 40120 40220 40250 40310 ...
##  $ Name                                             : chr  "Adelaide (C)" "Adelaide Hills (D
C)" "Alexandrina (DC)" "Anangu Pitjantjatjara (AC)" ...
##  $ Private.dwellings.with.no.Internet.connection    : chr  "1,046" "1,879" "2,322" "374" ...
##  $ Total.private.dwellings                          : chr  "8,178" "13,614" "9,503" "527" ...
##  $ X..dwellings.with.no.Internet.connection         : chr  "12.8" "13.8" "24.4" "71.0" ...
##  $ Col6                                             : logi  NA NA NA NA NA NA ...
##  $ All.private.dwellings.with.Internet.connections  : chr  "6,890" "11,341" "6,942" "147" ...
##  $ Total.private.dwellings.1                        : chr  "8,178" "13,614" "9,503" "527" ...
##  $ X..dwellings.with.Internet.connections           : chr  "84.3" "83.3" "73.1" "27.9" ...
##  $ Col10                                            : logi  NA NA NA NA NA NA ...
##  $ Private.dwellings.with.Broadband.Internet        : chr  "6,115" "10,290" "6,076" "126" ...
##  $ Total.private.dwellings.2                        : chr  "8,178" "13,614" "9,503" "527" ...
##  $ X..dwellings.with.Broadband.Internet             : chr  "74.8" "75.6" "63.9" "23.9" ...
##  $ Col14                                            : logi  NA NA NA NA NA NA ...
##  $ Private.dwellings.with.Dial.up.Internet          : chr  "244" "660" "470" "17" ...
##  $ Total.private.dwellings.3                        : chr  "8,178" "13,614" "9,503" "527" ...
##  $ X..dwellings.with.Dial.up.Internet               : chr  "3.0" "4.8" "4.9" "3.2" ...
##  $ Col18                                            : logi  NA NA NA NA NA NA ...
##  $ Private.dwellings.with.other.Internet.connections: chr  "531" "391" "396" "4" ...
##  $ Total.private.dwellings.4                        : chr  "8,178" "13,614" "9,503" "527" ...
##  $ X..dwellings.with.other.Internet.connections     : chr  "6.5" "2.9" "4.2" "0.8" ...
```

```
dim(internet_access)
```

```
## [1] 71 21
```

```
# Keeping ony required columns and renaming them to something sensible.
AEDC <- AEDC %>%
  dplyr::select(Code, Name,
                vulnerable_on_2_domain_per =
                  X..Children.developmentally.vulnerable.on.two.or.more.domains)

str(AEDC)
```

```
## 'data.frame':    71 obs. of  3 variables:
##  $ Code                      : num  40070 40120 40220 40250 40310 ...
##  $ Name                      : chr  "Adelaide (C)" "Adelaide Hills (DC)" "Alexandrina (DC)"
 "Anangu Pitjantjatjara (AC)" ...
##  $ vulnerable_on_2_domain_per: chr  "14.5" "6.8" "5.9" "68.9" ...
```

```
dim(AEDC)
```

```
## [1] 71  3
```

```
education <- education %>%
  dplyr::select(Code,
                education_at_16_per = X..full.time.participation.at.age.16,
                left_school_at_10 = Number,
                left_school_asr_per_100 = ASR.per.100)
str(education)
```

```
## 'data.frame':    71 obs. of  4 variables:
##  $ Code                   : num  40070 40120 40220 40250 40310 ...
##  $ education_at_16_per    : chr  "83.1" "89.4" "79.1" "80.8" ...
##  $ left_school_at_10      : chr  "1,818" "7,020" "7,191" "1,222" ...
##  $ left_school_asr_per_100: chr  "11.7" "21.1" "31.5" "80.9" ...
```

```
dim(education)
```

```
## [1] 71  4
```

```
internet_access <- internet_access %>%
  select(Code,
         no_internet_per = X..dwellings.with.no.Internet.connection,
         total_internet_per = X..dwellings.with.Internet.connections,
         broadband_internet_per = X..dwellings.with.Broadband.Internet,
         dial_up_internet_per = X..dwellings.with.Dial.up.Internet,
         other_internet_per = X..dwellings.with.other.Internet.connections)

str(internet_access)
```

```
## 'data.frame':    71 obs. of  6 variables:
##  $ Code                  : num  40070 40120 40220 40250 40310 ...
##  $ no_internet_per       : chr  "12.8" "13.8" "24.4" "71.0" ...
##  $ total_internet_per    : chr  "84.3" "83.3" "73.1" "27.9" ...
##  $ broadband_internet_per: chr  "74.8" "75.6" "63.9" "23.9" ...
##  $ dial_up_internet_per  : chr  "3.0" "4.8" "4.9" "3.2" ...
##  $ other_internet_per    : chr  "6.5" "2.9" "4.2" "0.8" ...
```

```
dim(internet_access)
```

```
## [1] 71  6
```

```
# Joining all the data frames into 1 using Code variable.
data <- AEDC %>% left_join(education) %>% left_join(internet_access)
```

```
## Joining, by = "Code"
## Joining, by = "Code"
```

```
#Removing , from numerical column so that they can be converted to numerical.
data <- data %>% mutate(left_school_at_10 = gsub(",", "", left_school_at_10))
# Converting all numerical values to numeric as they were read as character by library.
data <- data %>% mutate_at(vars(-Name), as.numeric)

data <- data %>% mutate(LGA_type = gsub("[^\\(]*\\(", "", Name)) %>%
  mutate(LGA_type = gsub(")", "", LGA_type)) %>%
  mutate(Name = gsub(" \\(\\w+\\)", "", Name))

# Finding the number of NAs for each row.
rowSums(is.na(data))
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 0 0 0 0 9 0 0 0 0
## [36] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [71] 0
```

```
which(rowSums(is.na(data)) > 3)
```

```
## [1] 31
```

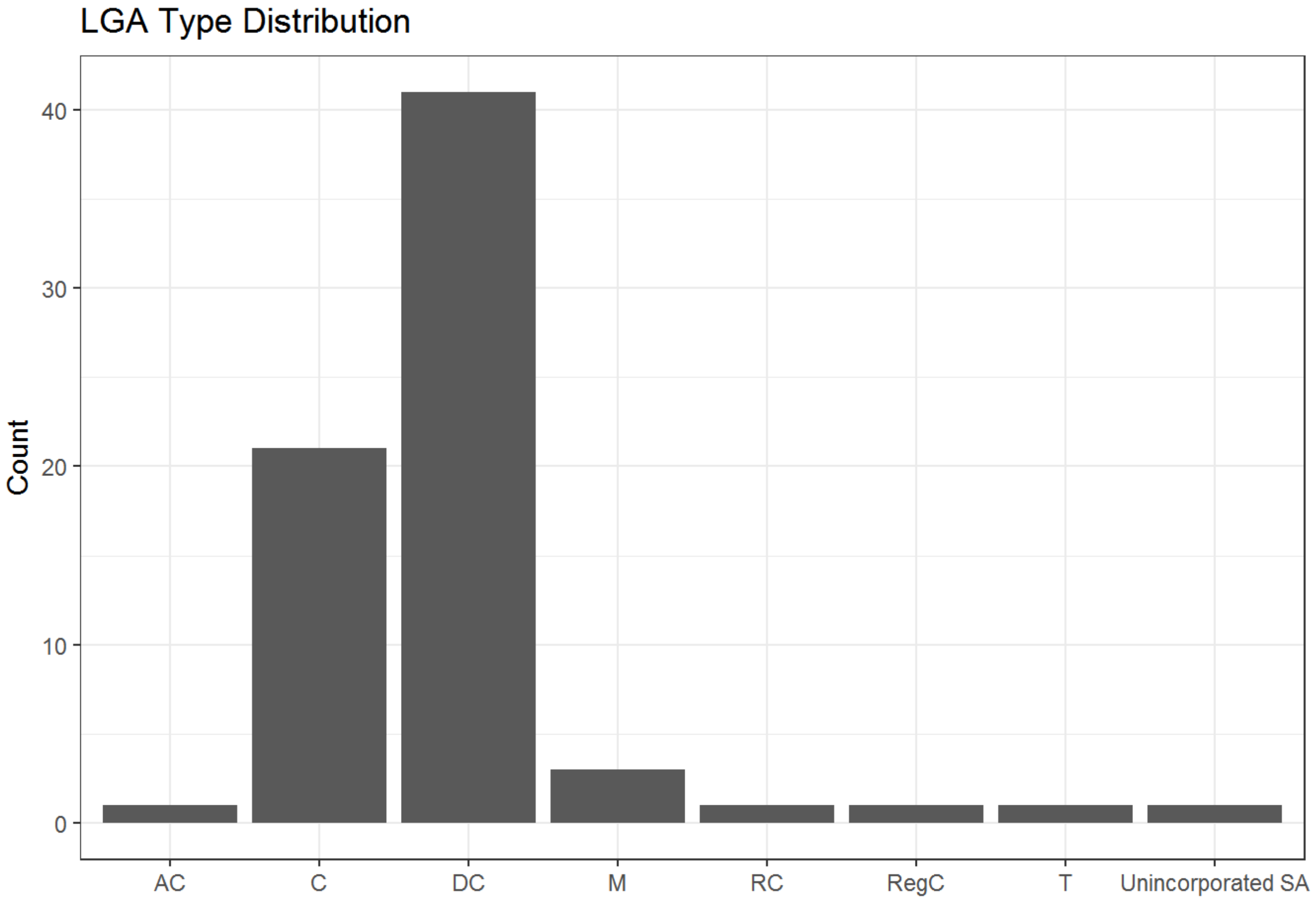
```
# This row has 9 NA's out of total 12 values and does not make sense to include it in our analys
is.
data <- data[rowSums(is.na(data)) < 3, ]

# Imputing all NAs. This will substitute all NAs with median.
data <- data %>% mutate_all(impute)
```

2. Exploratory Data Analysis

```
#library(plyr)
count(data,'LGA_type')  # frequency tables for categorical variable LGA type
```

```
## # A tibble: 1 × 2
##   `"LGA_type"`     n
##          <chr> <int>
## 1     LGA_type    70
```

```
#barplots for categorical variables
ggplot(data=data, aes(x=LGA_type)) + geom_bar() + ggtitle("LGA Type Distribution") + theme_bw()
+ xlab("") + ylab("Count")
```



```
#summary of numerical predictors % full-time participation at age 16,Number that left school at
 age 10 or below, ASRper100 of people who left school at year 10 or below, % dwellings with no i
nternet access, % dwellings with total internet access, % dwellings with broadband internet acce
ss, % dwellings with dial-up internet access and % other internet access.

summary(data$education_at_16_per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61.90   79.12   84.05   84.18   87.48  160.00
```

```
summary(data$left_school_at_10 )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     310    1017    2782    5741    6192   40240
```

```
summary(data$left_school_asr_per_100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.70   30.82   35.15   33.59   37.78   80.90
```

```
summary(data$no_internet_per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.60   23.08   27.60   27.33   31.45   71.00
```

```
summary(data$total_internet_per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.90   65.40   69.10   69.28   73.82   91.00
```

```
summary(data$broadband_internet_per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.90   56.23   60.80   61.20   65.47   83.20
```

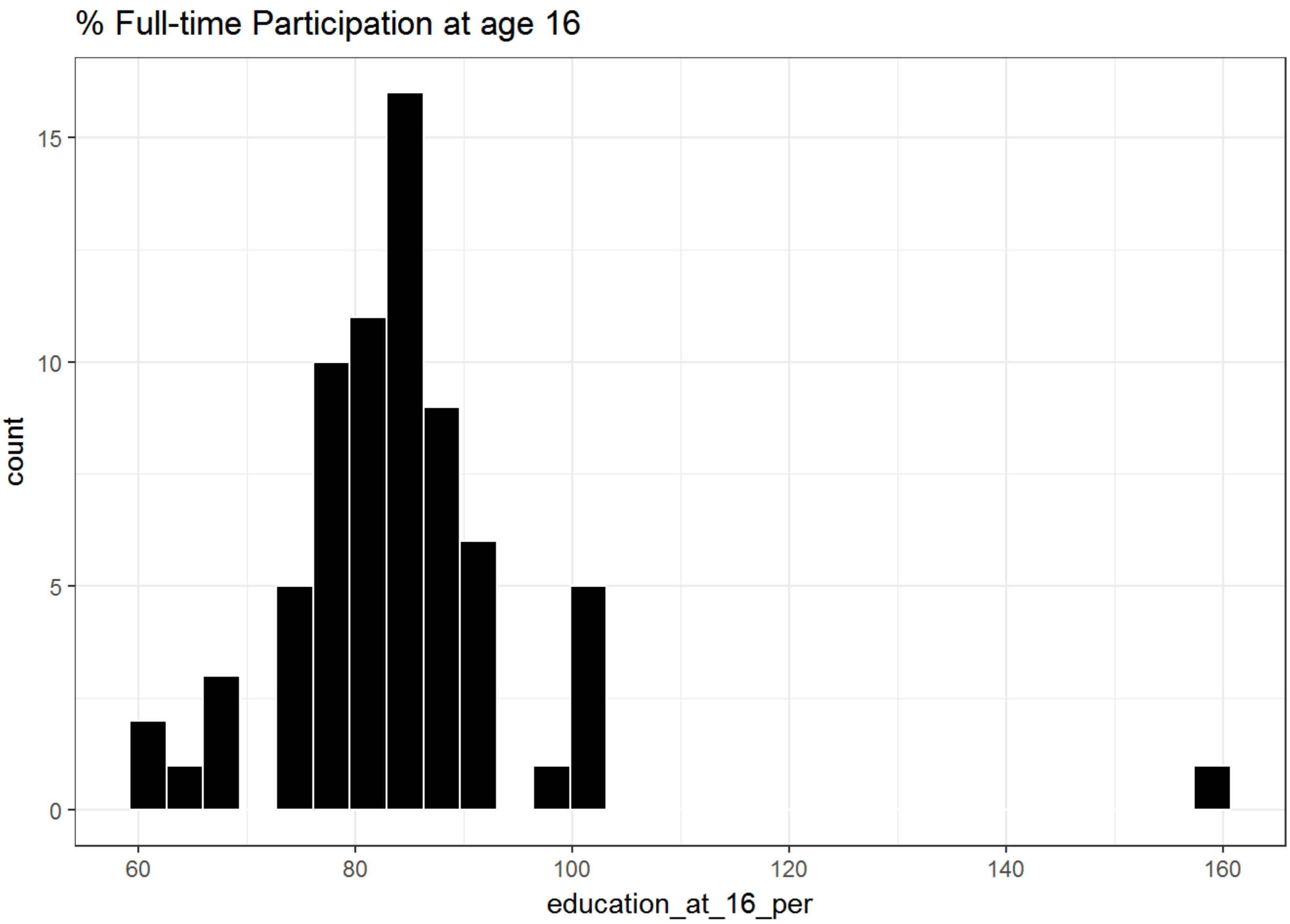
```
summary(data$dial_up_internet_per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.400   3.325   3.950   4.039   4.600   6.300
```

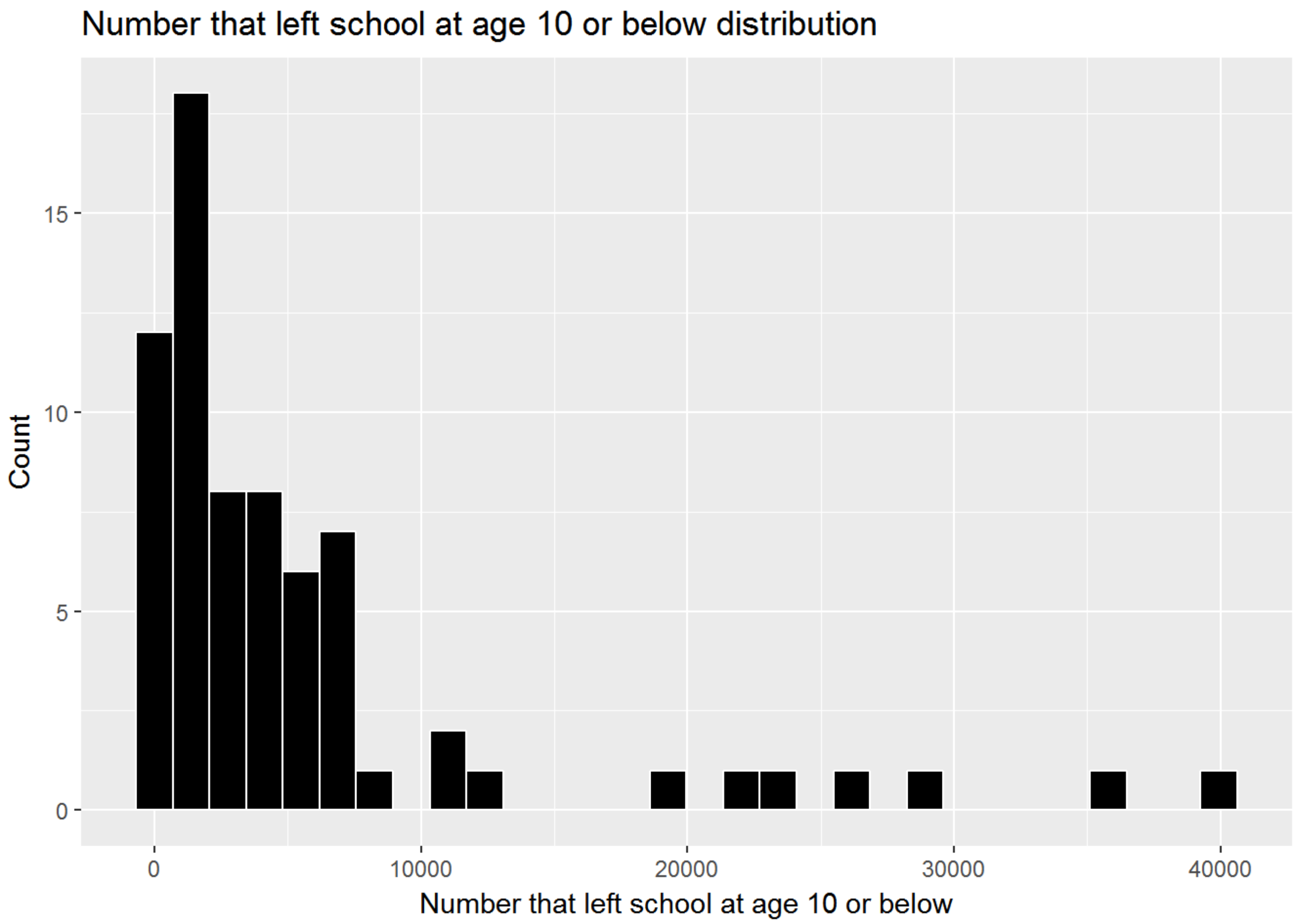
```
summary(data$other_internet_per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   3.500   4.100   4.051   4.500   6.500
```

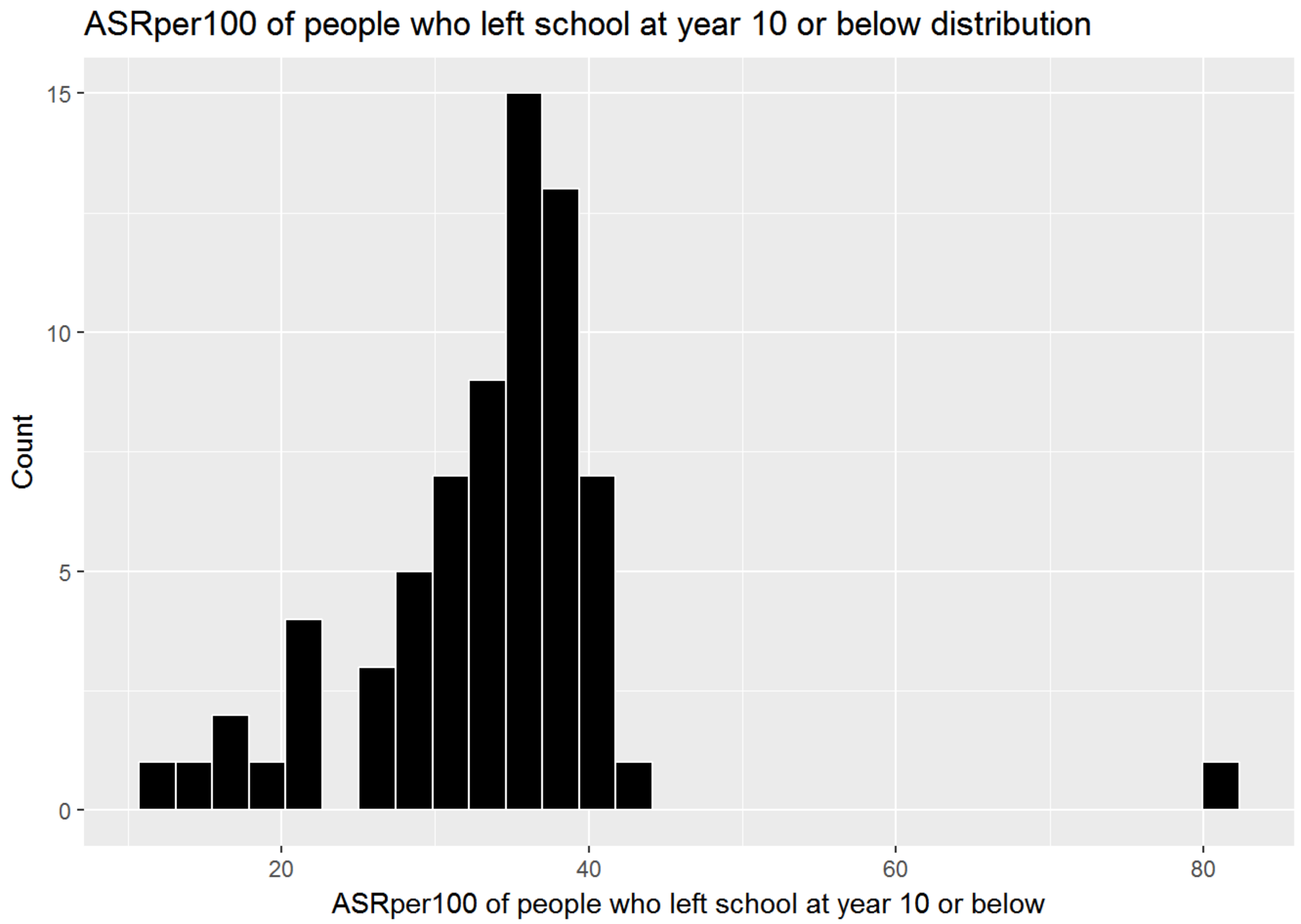
```
#Histograms of Numeric predictors % full-time participation at age 16,Number that left school at
 age 10 or below, ASRper100 of people who left school at year 10 or below, % dwellings with no i
nternet access, % dwellings with total internet access, % dwellings with broadband internet acce
ss, % dwellings with dial-up internet access and % other internet access.

ggplot(data, aes(x=education_at_16_per)) +
  geom_histogram(bins=30, color = "white", fill = "black") +
  ggtitle("% Full-time Participation at age 16") + theme_bw()  #Histogram for % Full-time Partic
ipation at age 16
```



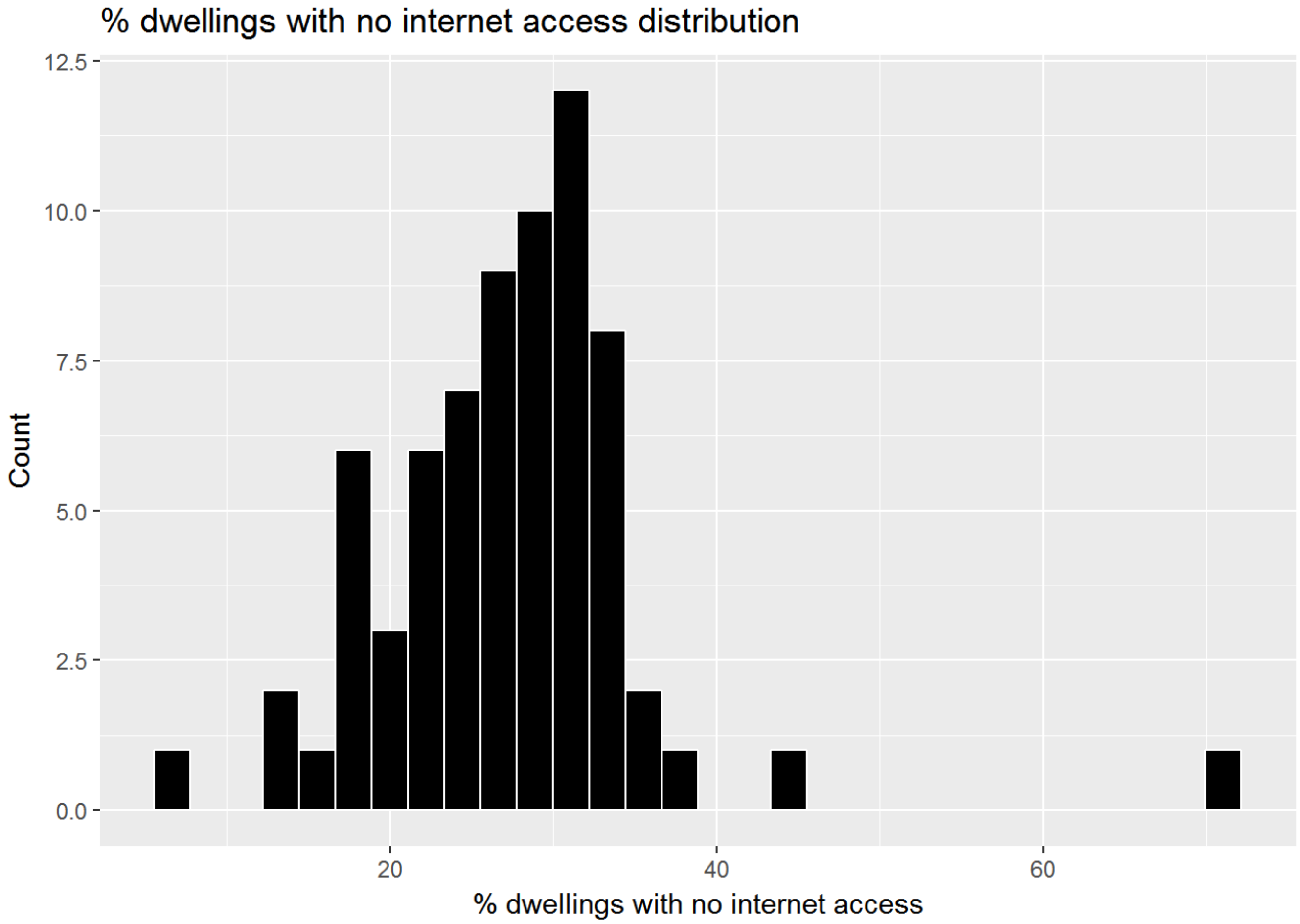
```
ggplot(data, aes(x=left_school_at_10 )) +geom_histogram(bins=30, color = "white", fill =
"black") +labs(
       x = "Number that left school at age 10 or below",
       y = "Count", title="Number that left school at age 10 or below distribution")
```



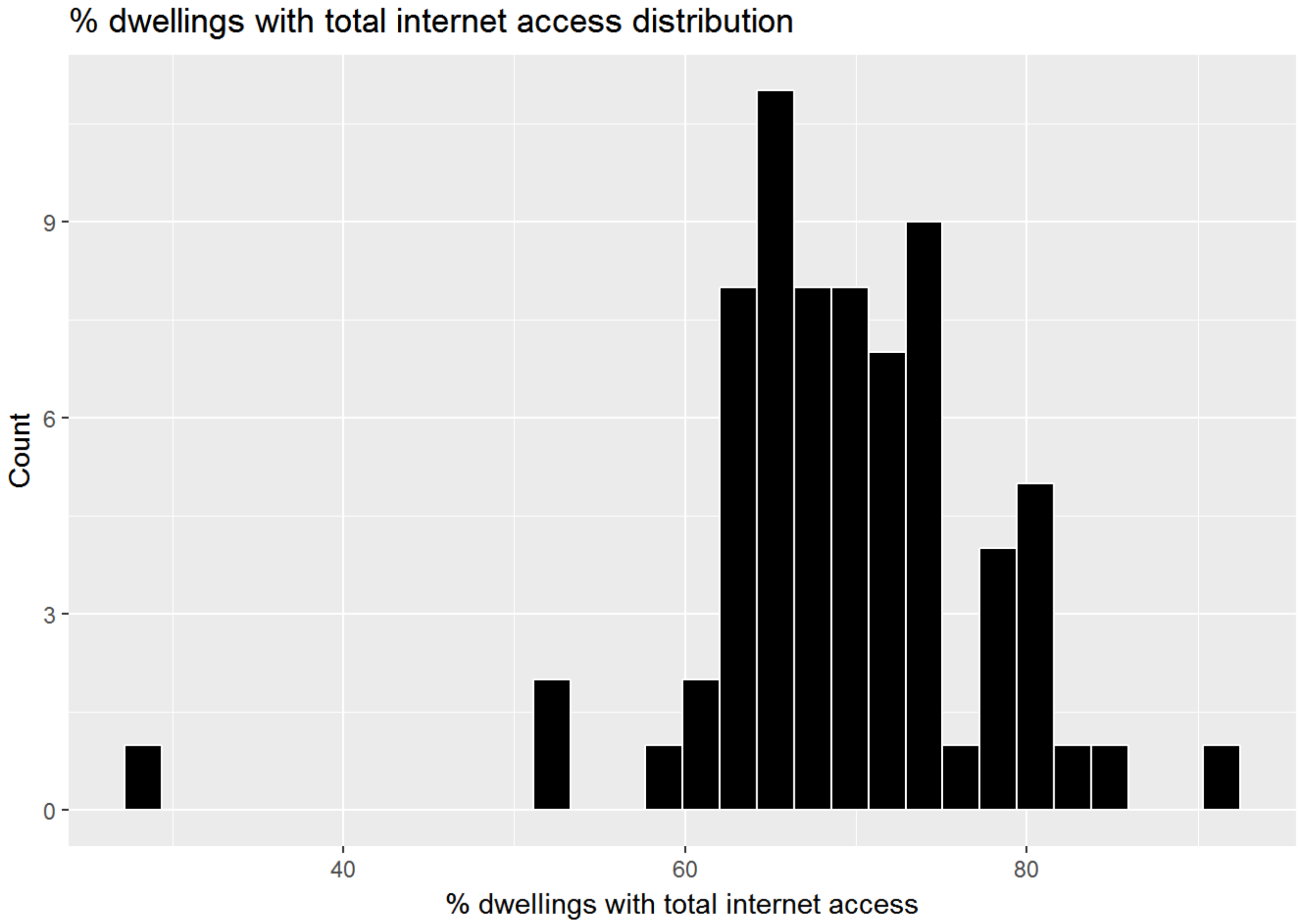
```
ggplot(data, aes(x=left_school_asr_per_100)) +
  geom_histogram(bins=30, color = "white", fill = "black") +labs(
       x = "ASRper100 of people who left school at year 10 or below",
       y = "Count", title="ASRper100 of people who left school at year 10 or below
distribution")
```



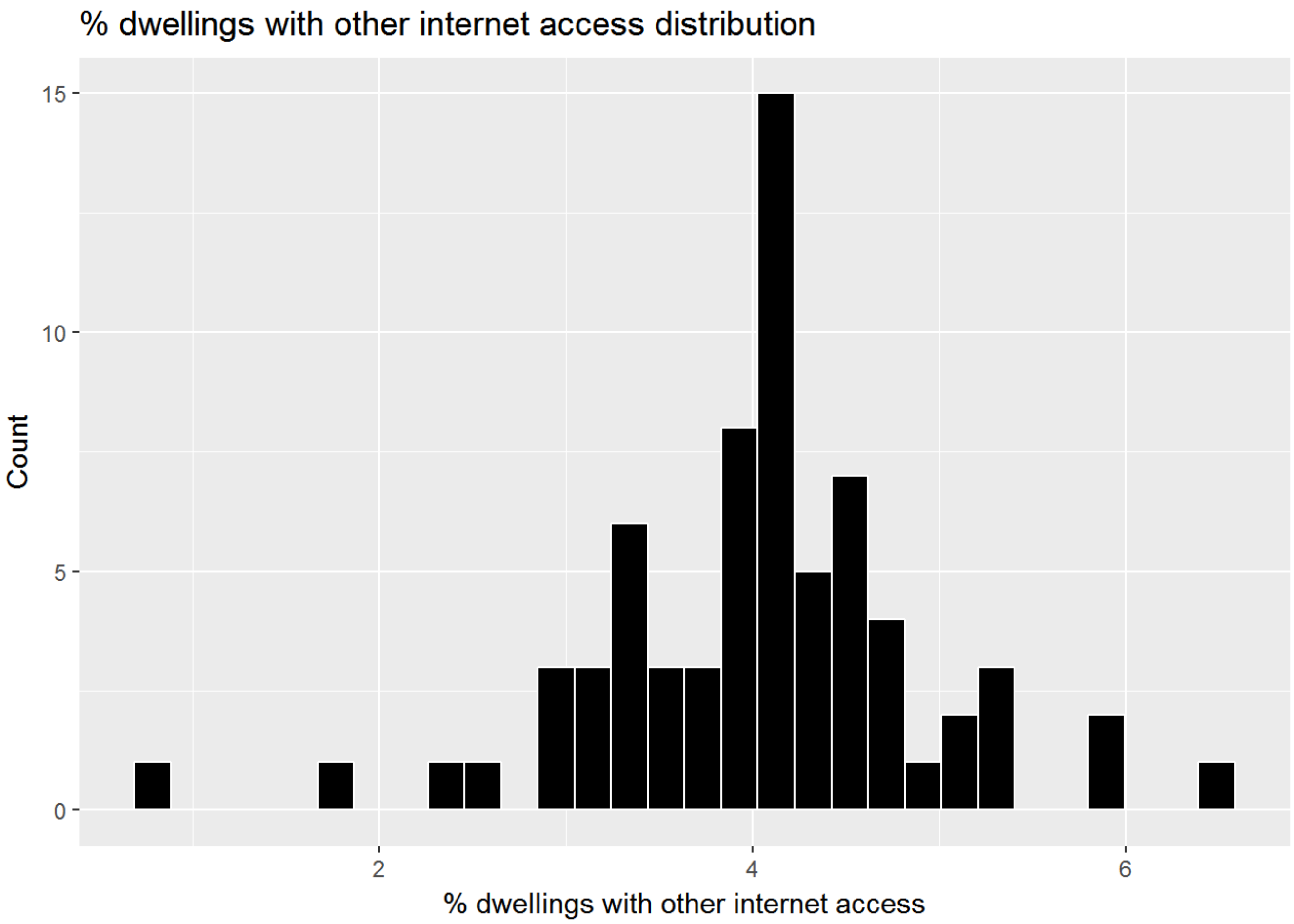
```
ggplot(data, aes(x=no_internet_per)) +
  geom_histogram(bins=30, color = "white", fill = "black") +labs(
      x = "% dwellings with no internet access",
      y = "Count", title="% dwellings with no internet access distribution")
```



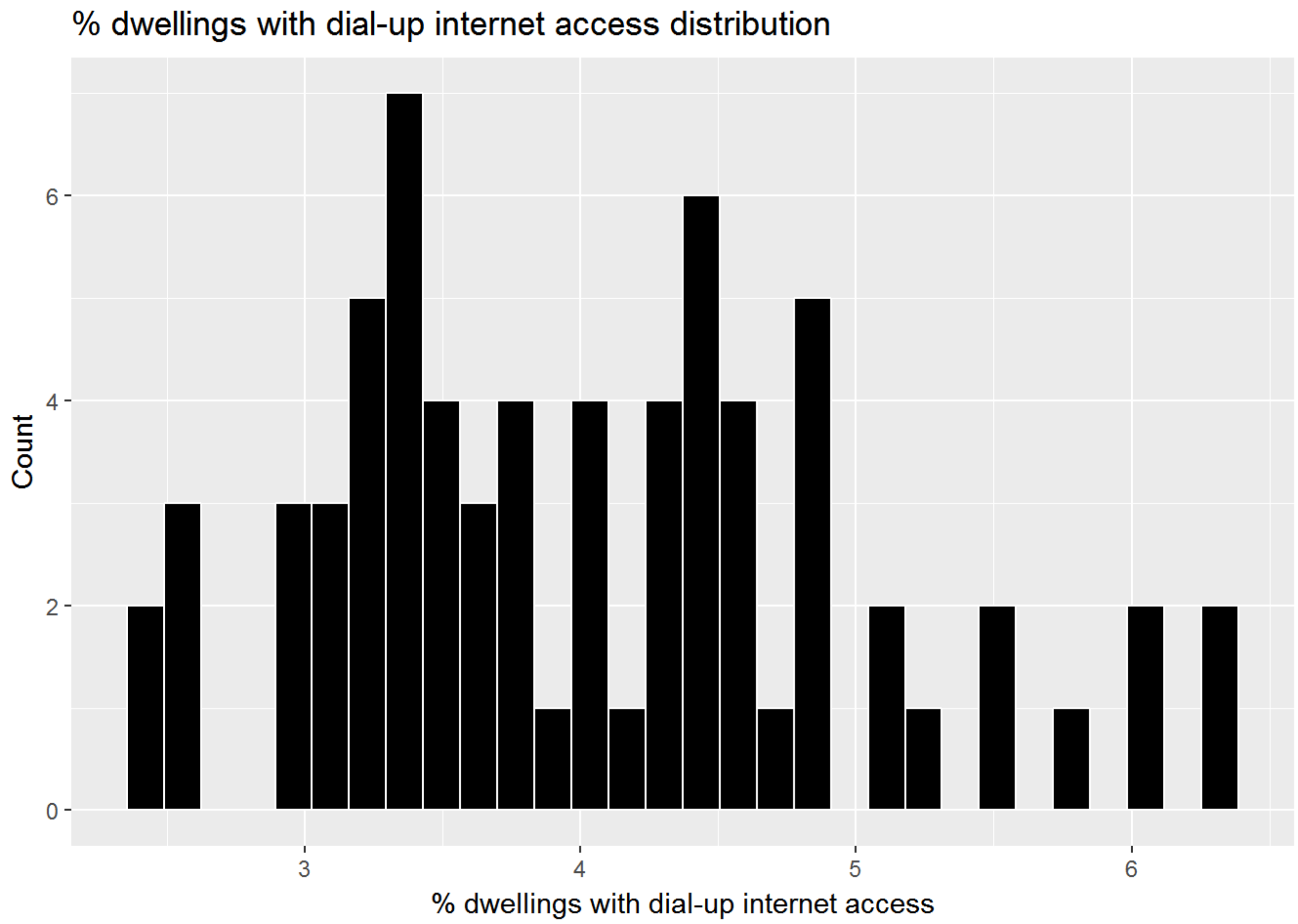
```
ggplot(data, aes(x=total_internet_per)) +
  geom_histogram(bins=30, color = "white", fill = "black") +labs(
       x = "% dwellings with total internet access",
       y = "Count", title="% dwellings with total internet access distribution")
```



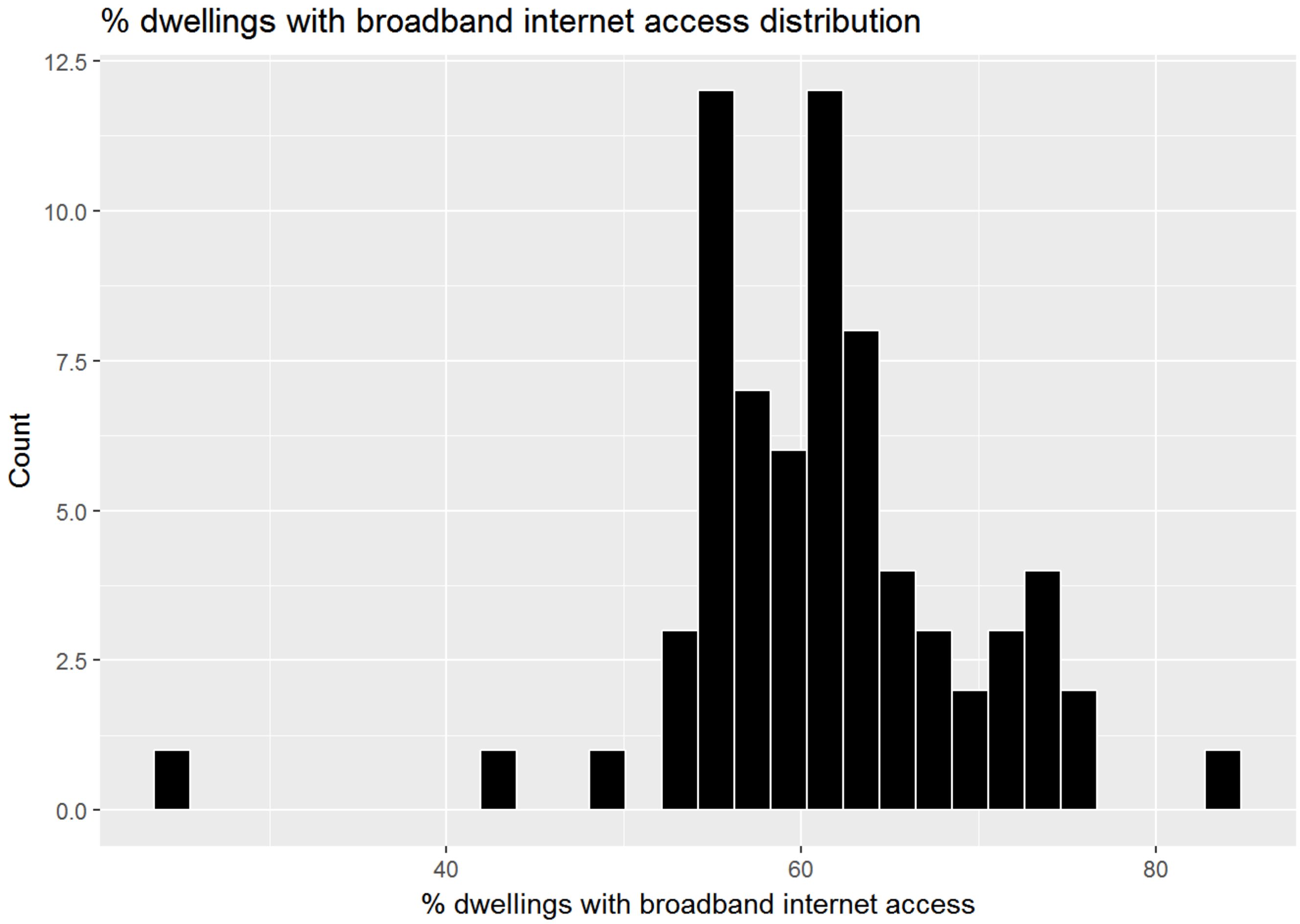
```
ggplot(data, aes(x=other_internet_per)) +
  geom_histogram(bins=30, color = "white", fill = "black") +labs(
      x = "% dwellings with other internet access",
      y = "Count", title="% dwellings with other internet access distribution")
```



```
ggplot(data, aes(x=dial_up_internet_per)) +
  geom_histogram(bins=30, color = "white", fill = "black") +labs(
      x = "% dwellings with dial-up internet access",
      y = "Count", title="% dwellings with dial-up internet access distribution")
```



```
ggplot(data, aes(x=broadband_internet_per)) +
  geom_histogram(bins=30, color = "white", fill = "black") +labs(
      x = "% dwellings with broadband internet access",
      y = "Count", title="% dwellings with broadband internet access distribution")
```



```
#Summary tables of predictors and response by LGA type


data %>% group_by(LGA_type)%>%
  dplyr::summarize(mean_edu_at_16_per =mean(education_at_16_per),
                   max_edu_at_16_per=max(education_at_16_per),
                   min_edu_at_16_per=min(education_at_16_per)) %>% kable
```

LGA_type	mean_edu_at_16_per	max_edu_at_16_per	min_edu_at_16_per
AC	80.80000	80.8	80.8
C	83.75714	92.4	74.7
DC	85.57805	160.0	62.5
M	79.13333	88.9	69.2
RC	76.90000	76.9	76.9
RegC	86.00000	86.0	86.0
T	82.10000	82.1	82.1
Unincorporated SA	61.90000	61.9	61.9

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(mean_left_school_at_10 =mean(left_school_at_10),
                   max_left_school_at_10=max(left_school_at_10),
                   min_left_school_at_10=min(left_school_at_10)) %>% kable
```

LGA_type	mean_left_school_at_10	max_left_school_at_10	min_left_school_at_10
AC	1222.000	1222	1222
C	13696.381	40245	1818
DC	2131.415	7191	310
M	2664.667	6116	800
RC	6761.000	6761	6761
RegC	3604.000	3604	3604
T	6003.000	6003	6003
Unincorporated SA	1306.000	1306	1306

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(left_school_asr_per_100 =mean(left_school_asr_per_100),
                   max_left_school_asr_per_100=max(left_school_asr_per_100),
                   min_left_school_asr_per_100=min(left_school_asr_per_100)) %>% kable
```

LGA_type	left_school_asr_per_100	max_left_school_asr_per_100	min_left_school_asr_per_100
AC	80.90000	80.90000	80.90000
C	27.41905	27.41905	27.41905
DC	35.69512	35.69512	35.69512
M	28.83333	28.83333	28.83333
RC	40.10000	40.10000	40.10000
RegC	33.20000	33.20000	33.20000
T	33.70000	33.70000	33.70000
Unincorporated SA	37.50000	37.50000	37.50000

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(no_internet_per =mean(no_internet_per),
                   max_no_internet_per=max(no_internet_per),
                   min_no_internet_per=min(no_internet_per)) %>% kable
```

LGA_type	no_internet_per	max_no_internet_per	min_no_internet_per
AC	71.00000	71.00000	71.00000

LGA_type	no_internet_per	max_no_internet_per	min_no_internet_per
C	23.26190	23.26190	23.26190
DC	29.00732	29.00732	29.00732
M	20.46667	20.46667	20.46667
RC	31.50000	31.50000	31.50000
RegC	18.70000	18.70000	18.70000
T	25.20000	25.20000	25.20000
Unincorporated SA	27.40000	27.40000	27.40000

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(total_internet_per =mean(total_internet_per),
                   max_total_internet_per=max(total_internet_per),
                   min_total_internet_per=min(total_internet_per)) %>% kable
```

LGA_type	total_internet_per	max_total_internet_per	min_total_internet_per
AC	27.90000	27.90000	27.90000
C	73.34762	73.34762	73.34762
DC	67.62927	67.62927	67.62927
M	76.26667	76.26667	76.26667
RC	64.40000	64.40000	64.40000
RegC	78.90000	78.90000	78.90000
T	71.30000	71.30000	71.30000
Unincorporated SA	65.40000	65.40000	65.40000

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(broadband_internet_per =mean(broadband_internet_per),
                   max_broadband_internet_per=max(broadband_internet_per),
                   min_broadband_internet_per=min(broadband_internet_per)) %>% kable
```

LGA_type	broadband_internet_per	max_broadband_internet_per	min_broadband_internet_per
AC	23.90000	23.90000	23.90000
C	65.77143	65.77143	65.77143
DC	59.17561	59.17561	59.17561
M	69.20000	69.20000	69.20000
RC	55.20000	55.20000	55.20000

LGA_type	broadband_internet_per	max_broadband_internet_per	min_broadband_internet_per
RegC	70.80000	70.80000	70.80000
T	63.80000	63.80000	63.80000
Unincorporated SA	55.30000	55.30000	55.30000

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(other_internet_per =mean(other_internet_per),
                   max_other_internet_per=max(other_internet_per),
                   min_other_internet_per=min(other_internet_per)) %>% kable
```

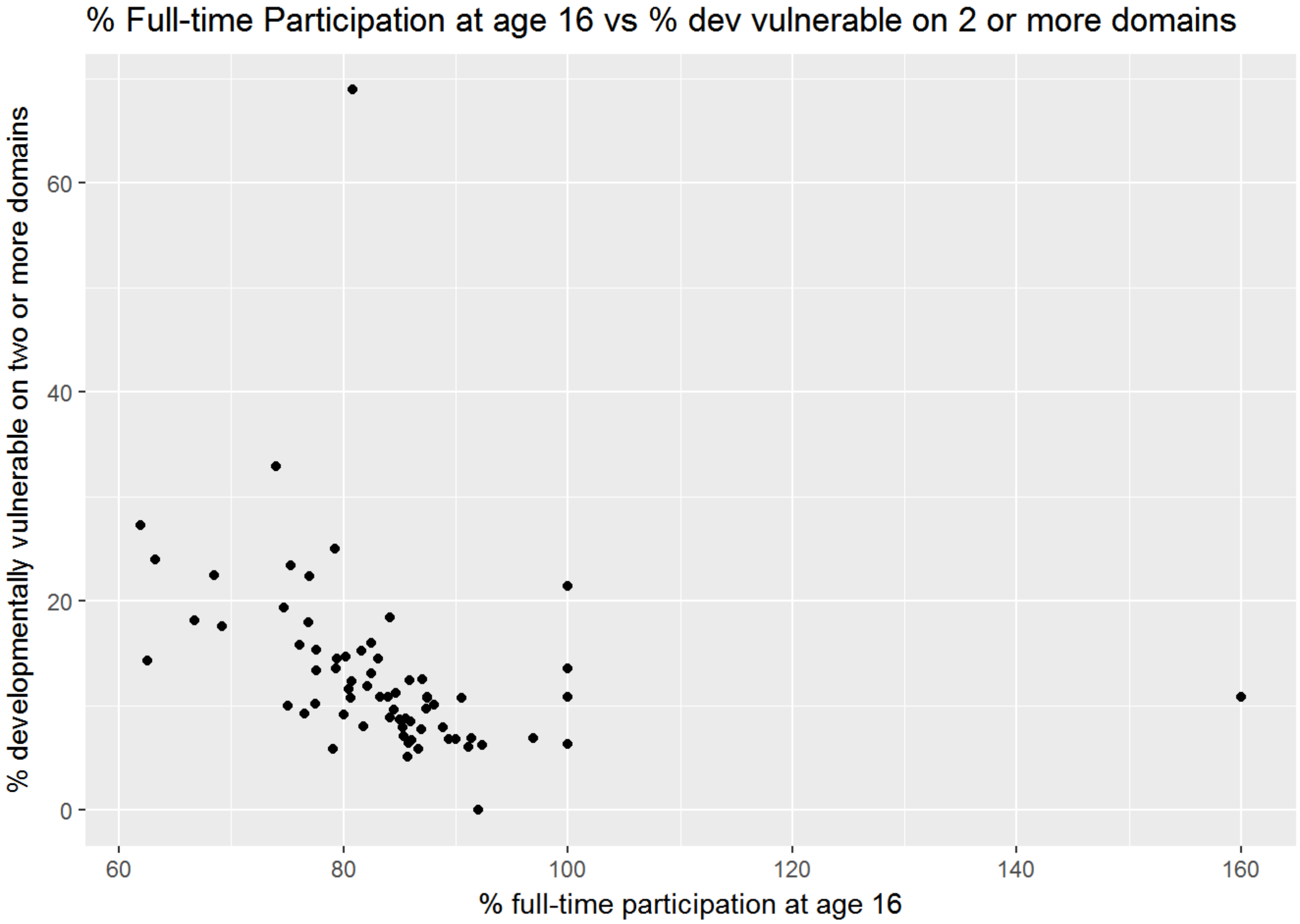
LGA_type	other_internet_per	max_other_internet_per	min_other_internet_per
AC	0.800000	0.800000	0.800000
C	4.285714	4.285714	4.285714
DC	3.931707	3.931707	3.931707
M	4.233333	4.233333	4.233333
RC	5.300000	5.300000	5.300000
RegC	3.700000	3.700000	3.700000
T	4.000000	4.000000	4.000000
Unincorporated SA	5.900000	5.900000	5.900000

```
data %>% group_by(LGA_type)%>%
  dplyr::summarize(mean_dial_up_internet_per =mean(dial_up_internet_per),
                   max_dial_up_internet_per=max(dial_up_internet_per),
                   min_dial_up_internet_per=min(dial_up_internet_per)) %>% kable
```

LGA_type	mean_dial_up_internet_per	max_dial_up_internet_per	min_dial_up_internet_per
AC	3.200000	3.2	3.2
C	3.300000	4.6	2.4
DC	4.519512	6.3	2.4
M	2.866667	3.2	2.5
RC	4.000000	4.0	4.0
RegC	4.400000	4.4	4.4
T	3.600000	3.6	3.6
Unincorporated SA	4.300000	4.3	4.3

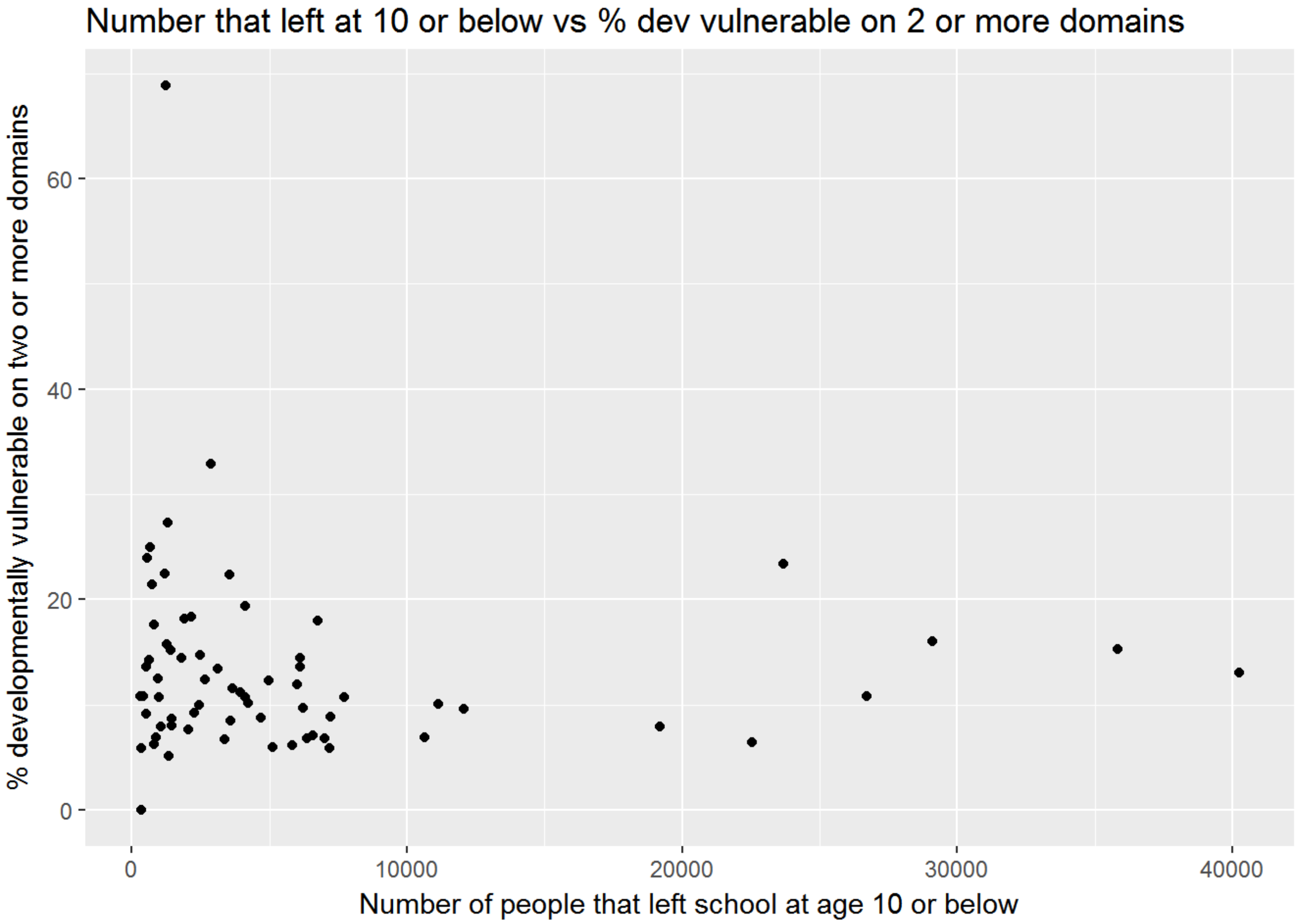
```
#Scatterplots of response variable AEDC % developmentally vulnerable on two or more domains versus all numeric predictors,  % full-time participation at age 16,Number that left school at age 10 or below, ASRper100 of people who left school at year 10 or below, % dwellings with no internet access, % dwellings with total internet access, % dwellings with broadband internet access, % dwellings with dial-up internet access and % other internet access.

ggplot(data, aes(x=education_at_16_per, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "% full-time participation at age 16",
       y = "% developmentally vulnerable on two or more domains", title="% Full-time Participation at age 16 vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuous.
```



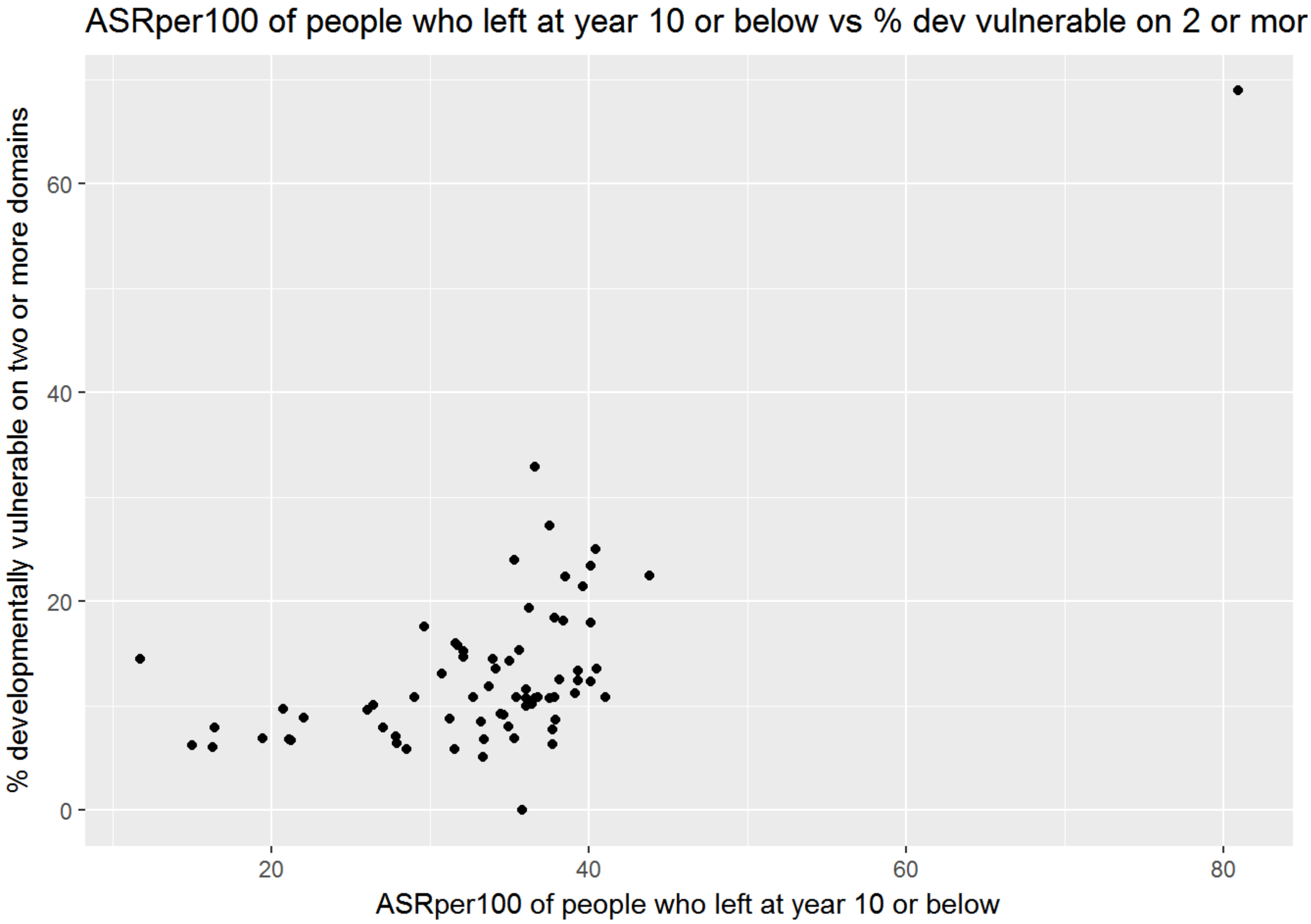
```
ggplot(data, aes(x=left_school_at_10, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "Number of people that left school at age 10 or below",
       y = "% developmentally vulnerable on two or more domains", title="Number that left at 10 or below vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuous.
```



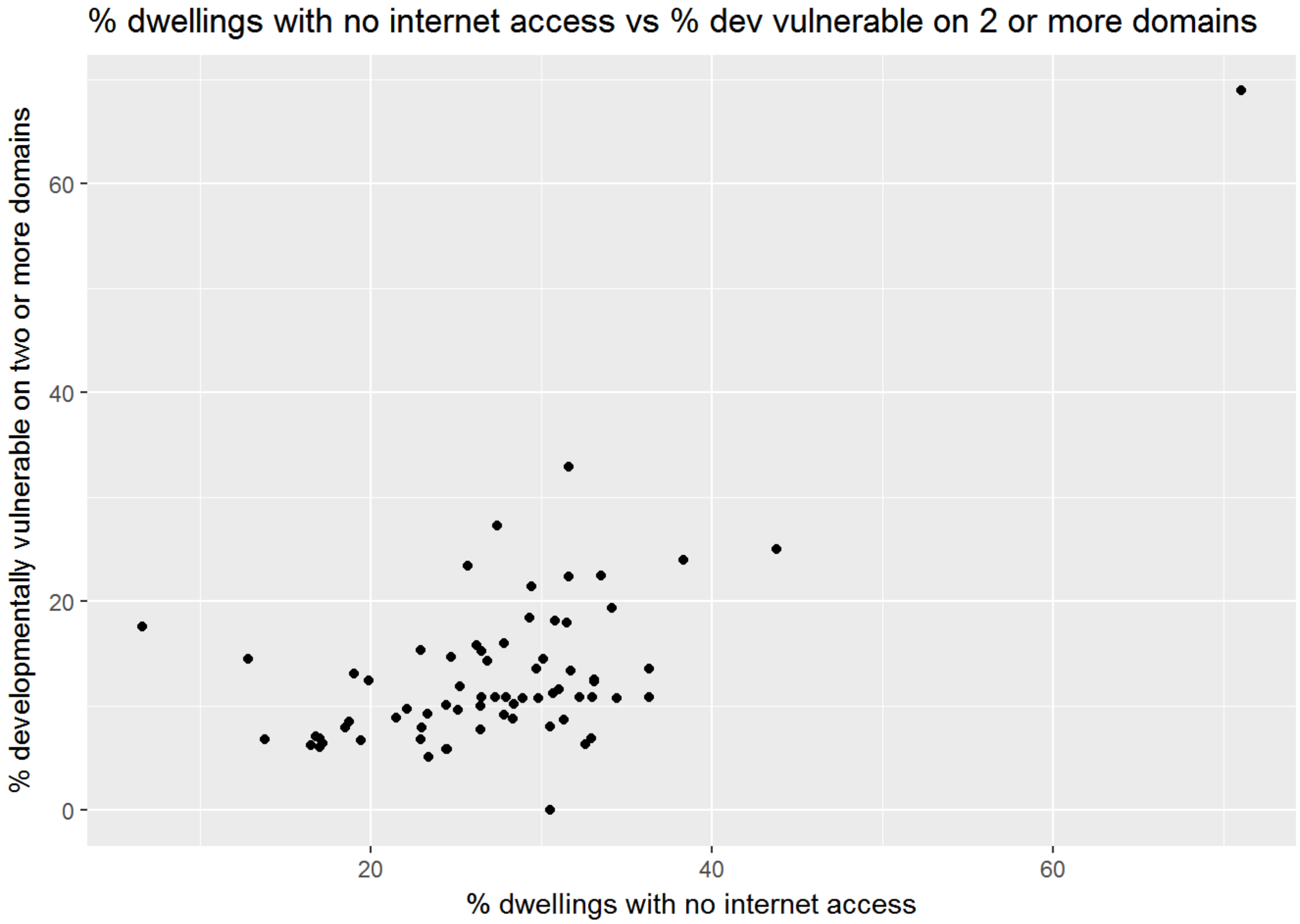
```
ggplot(data, aes(x=left_school_asr_per_100, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "ASRper100 of people who left at year 10 or below",
       y = "% developmentally vulnerable on two or more domains", title="ASRper100 of people who
 left at year 10 or below vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuou
s.
```



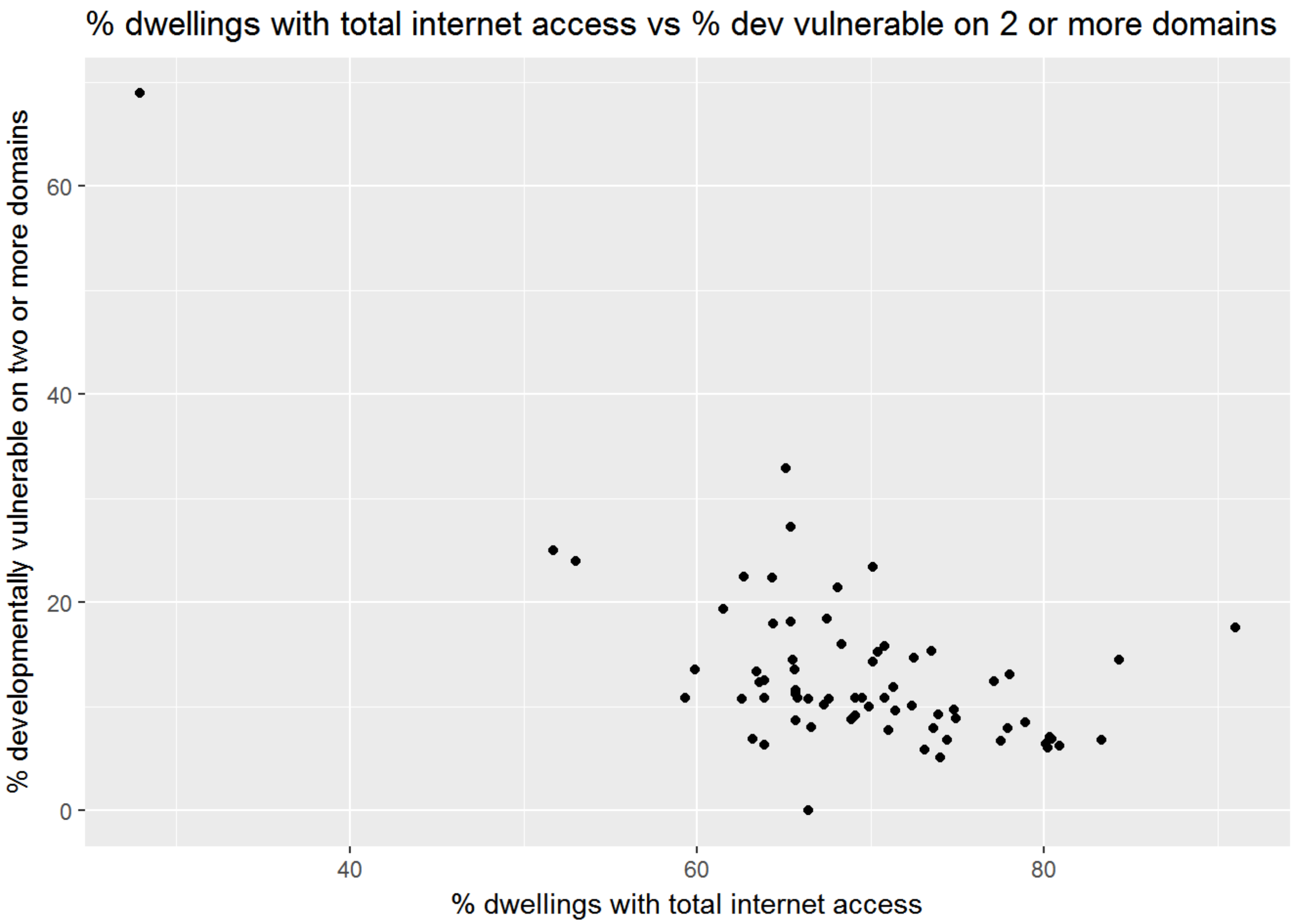
```
ggplot(data, aes(x=no_internet_per, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "% dwellings with no internet access",
       y = "% developmentally vulnerable on two or more domains", title="% dwellings with no int
ernet access vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuou
s.
```



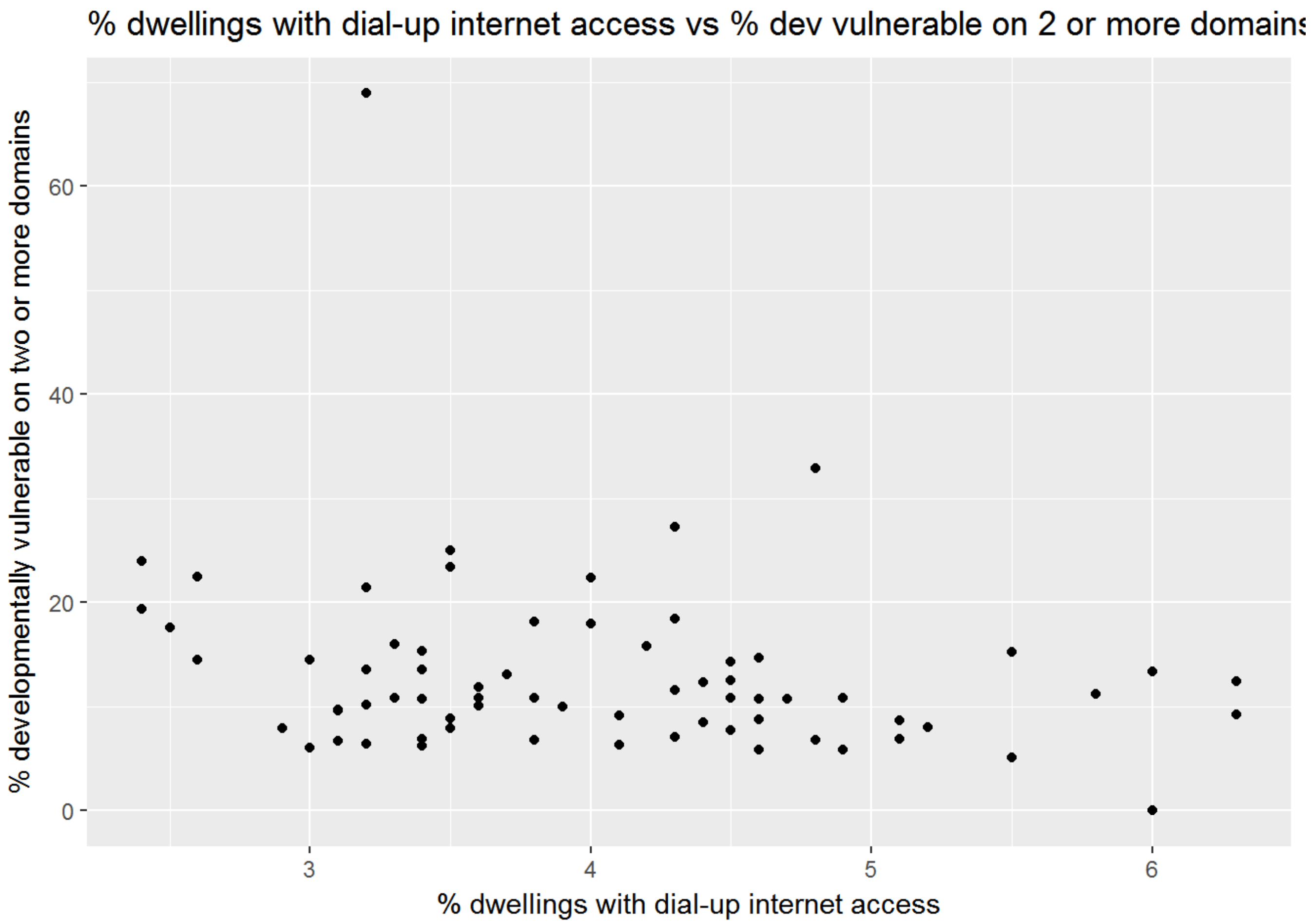
```
ggplot(data, aes(x=total_internet_per, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "% dwellings with total internet access",
       y = "% developmentally vulnerable on two or more domains", title="% dwellings with total
 internet access vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuou
s.
```



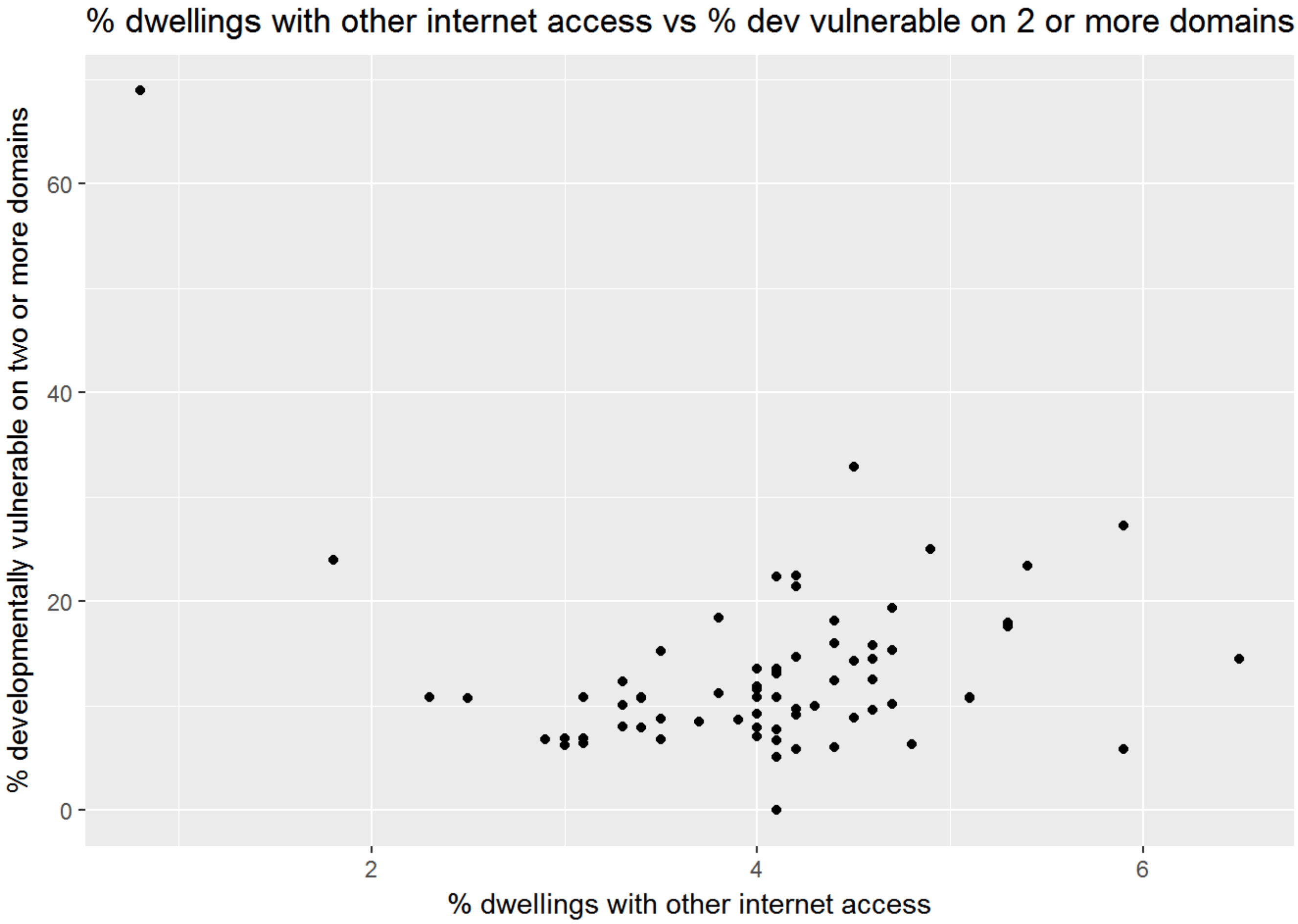
```
ggplot(data, aes(x=dial_up_internet_per, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "% dwellings with dial-up internet access",
       y = "% developmentally vulnerable on two or more domains", title="% dwellings with dial-u
p internet access vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuou
s.
```



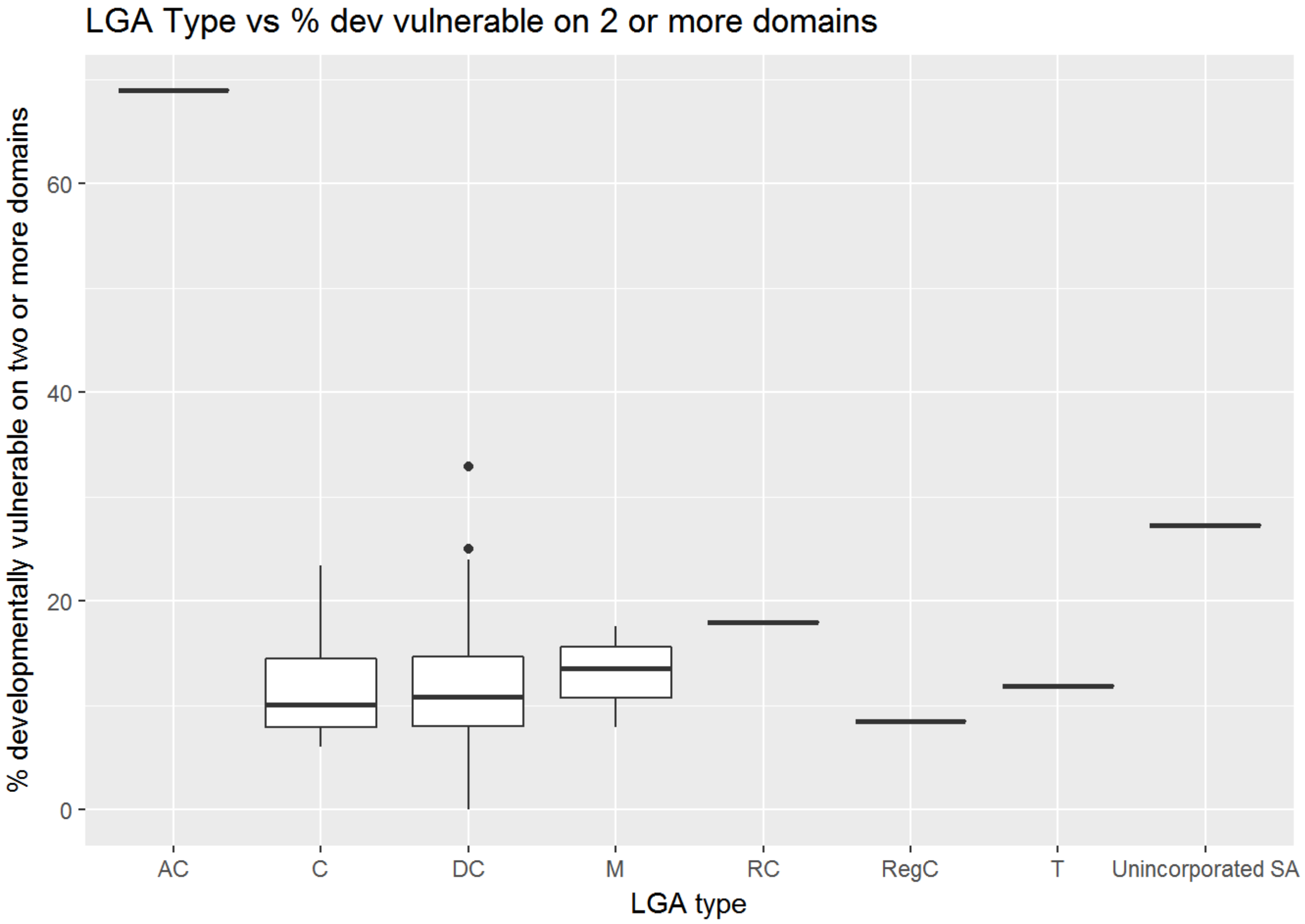
```
ggplot(data, aes(x=other_internet_per, y=vulnerable_on_2_domain_per)) + geom_point() +labs(
       x = "% dwellings with other internet access",
       y = "% developmentally vulnerable on two or more domains", title="% dwellings with other
 internet access vs % dev vulnerable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuou
s.
```



```
#boxplot of response variable % developmentally vulnerable on 2 or more domains for categorical
 predictor LGA_type

ggplot(data, aes(x = LGA_type, y = vulnerable_on_2_domain_per)) + geom_boxplot() +labs(
       x = "LGA type",
       y = "% developmentally vulnerable on two or more domains", title="LGA Type vs % dev vulne
rable on 2 or more domains")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to continuou
s.
```



Comments

From the frequency table of LGA_Type we can see that the LGA_type DC is the most frequent with 41 occurrences followed by C with 21.

The barplot for LGA type shows frequencies of each LGA Type. We can see that DC has the highest frequency of 41 followed by C which has a frequency of 21.

Summary of Numerical predictors

% education at 16 is in the range 61.90 to 160.00 Number that have left school at 10 or belo is in the range 310 to 40240 ASR per 100 that have left school at 10 or below is in range 11.70 to 80.90 % dwellings with no internet access is in the range 6.60 to 71.00 % dwellings with total internet access is in the range 27.90 to 91.00 % dwellings with broadband internet access is in the range 23.90 to 83.20 % dwellings with dial up internet access is in the range 2.4 to 6.3 % dwellings with other internet access is in the range 0.80 to 6.50

Analysing the histograms

The most frequent % of full time participation at 16 is around 85%. There are over 18 instances where around 2500-3500 students have left school below at age 10 or below The most frequent occurrence of ASR per 100 that have left school at 10 or below is 34-36%. The most frequent occurrence of % dwellings with no internet access is 30-32%. The most frequent occurrence of % dwellings with total internet access is 64-66% The most frequent occurrence of % dwellings with dial up internet access is around 3.3-3.4% The most frequent occurrence of % dwellings with broadband internet access is 60-63%. The most frequent occurrence of % dwellings with other internet access is 60-63%.

Summary (grouped by LGA type)

The mean % of full time participation at 16 is highest in RegC and lowest in RC. The mean number of students that left school at 10 or below is highest in C and lowest in Unincorporated part of SA. The mean ASR per 100 of students that left school at 10 or below is highest in AC and lowest in C. The mean % of dwellings with total internet access is highest in RegC and lowest in RC. The mean % of dwellings with broadband is highest in RegC and lowest in RC and unincoprporated SA. The mean % of dwellings with no internet access is highest in AC and lowest in RegC. The mean % of dwellings with other internect access is highest in unincorporated SA and lowest in AC. The mean % of dwellings with dial up is highest in DC and lowest in M.

Analysis of Scatterplot

The most % of full time participation at 16 is concentrated between 80-90% that corresponds to between 15-20% being developmentally vulnerable on two or more domains Number that have left school at 10 or below are mostly concentrated between 1000 and 10000. The total internet access,no internet access, broadband, dial up and other do not show any specific relation or pattern.

Analysis of boxplot

The boxplot b/w LGA Type and % developmentally vulnerable on two or more domains shows that the lga type AC has the highest mean, min and max values for % developmentally vulnerable on two or more domains

3. Creating Geographic maps of AEDC variable per LGA

```
# I extracted data from the zip file and the folder is copied parallel to the Rmd file.
shp <- readOGR(dsn = "ASGC_LGA2011"           # folder with the .shp file
               , layer = "LGA11aAust")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "ASGC_LGA2011", layer: "LGA11aAust"
## with 565 features
## It has 3 fields
```

```
# Exploring the data and summary.
head(shp@data)
```

```
##   STATE_CODE LGA_CODE11            LGA_NAME11
## 0          1      10050            Albury (C)
## 1          1      10110 Armidale Dumaresq (A)
## 2          1      10150          Ashfield (A)
## 3          1      10200            Auburn (C)
## 4          1      10250           Ballina (A)
## 5          1      10300         Balranald (A)
```

```
str(shp@data)
```

```
## 'data.frame':    565 obs. of  3 variables:
##  $ STATE_CODE: Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ LGA_CODE11: Factor w/ 565 levels "10050","10110",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ LGA_NAME11: Factor w/ 564 levels "Adelaide (C)",..: 4 11 13 14 18 20 22 31 35 36 ...
```

```
summary(shp@data)
```

```
##    STATE_CODE    LGA_CODE11                LGA_NAME11
##  1      :153   10050  :  1   Campbelltown (C)   :  2
##  5      :139   10110  :  1   Adelaide (C)       :  1
##  2      : 80   10150  :  1   Adelaide Hills (DC):  1
##  3      : 74   10200  :  1   Albany (C)         :  1
##  4      : 71   10250  :  1   Albury (C)         :  1
##  6      : 29   10300  :  1   Alexandrina (DC)   :  1
##  (Other): 19   (Other):559   (Other)            :558
```

```
# Subsetting the data for South Australia. state code for South Australia is 4.
shp_sa<-subset(shp,shp@data$STATE_CODE=="4")
head(shp_sa@data)
```

```
##     STATE_CODE LGA_CODE11                 LGA_NAME11
## 307          4      40070               Adelaide (C)
## 308          4      40120        Adelaide Hills (DC)
## 309          4      40220           Alexandrina (DC)
## 310          4      40250 Anangu Pitjantjatjara (AC)
## 311          4      40310               Barossa (DC)
## 312          4      40430          Barunga West (DC)
```

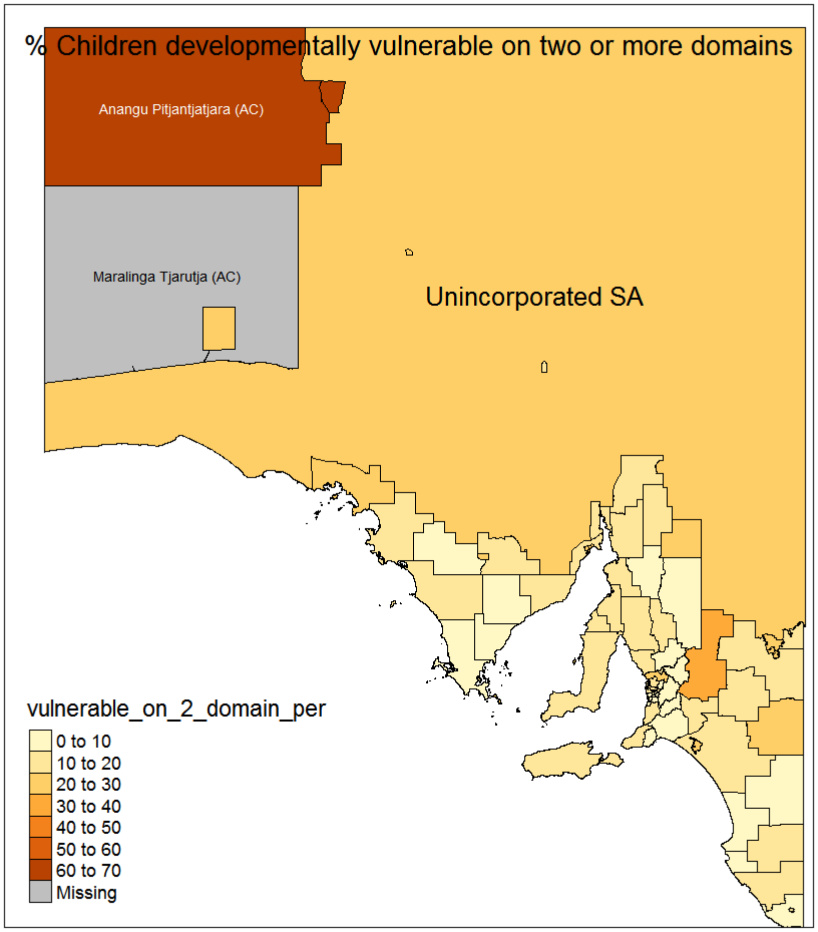
```
str(shp_sa@data)
```

```
## 'data.frame':    71 obs. of  3 variables:
##  $ STATE_CODE: Factor w/ 9 levels "1","2","3","4",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ LGA_CODE11: Factor w/ 565 levels "10050","10110",..: 308 309 310 311 312 313 314 315 316 3
17 ...
##  $ LGA_NAME11: Factor w/ 564 levels "Adelaide (C)",..: 1 2 5 8 27 28 40 75 84 96 ...
```

```
summary(shp_sa)
```

```
## Object of class SpatialPolygonsDataFrame
## Coordinates:
##        min       max
## x 129.0013 141.00296
## y -38.0626 -25.99615
## Is projected: FALSE
## proj4string : [+proj=longlat +ellps=GRS80 +no_defs]
## Data attributes:
##    STATE_CODE   LGA_CODE11                      LGA_NAME11
##  4      :71   40070  : 1   Adelaide (C)              : 1
##  1      : 0   40120  : 1   Adelaide Hills (DC)       : 1
##  2      : 0   40220  : 1   Alexandrina (DC)          : 1
##  3      : 0   40250  : 1   Anangu Pitjantjatjara (AC): 1
##  5      : 0   40310  : 1   Barossa (DC)              : 1
##  6      : 0   40430  : 1   Barunga West (DC)         : 1
##  (Other): 0   (Other):65   (Other)                   :65
```

```
data <- data %>% mutate(Code = as.character(Code))
# Adding a column named code of type character to be able to join the 2 data frames.
shp_sa@data <- shp_sa@data %>% mutate(Code = as.character(LGA_CODE11))
shp_sa@data <- left_join(shp_sa@data, data)

# Plotting the required map.
qtm(shp_sa, "vulnerable_on_2_domain_per",format = "World"
    ,text="LGA_NAME11",text.size="AREA"
    ,borders="black", title = "% Children developmentally vulnerable on two or more domains"
    )
```


% Children developmentally vulnerable on two or more domains
Anangu Pitjantjatjara (AC)
Maralinga Tjarutja (AC)
Unincorporated SA
vulnerable_on_2_domain_per
0 to 10
10 to 20
20 to 30
30 to 40
40 to 50
50 to 60
60 to 70
Missing

```
#plotting required ggmap
address <- data$Name #the row that contains all the location is selected
lonlat <- geocode(address) #this is setting the longitude and latitude
```

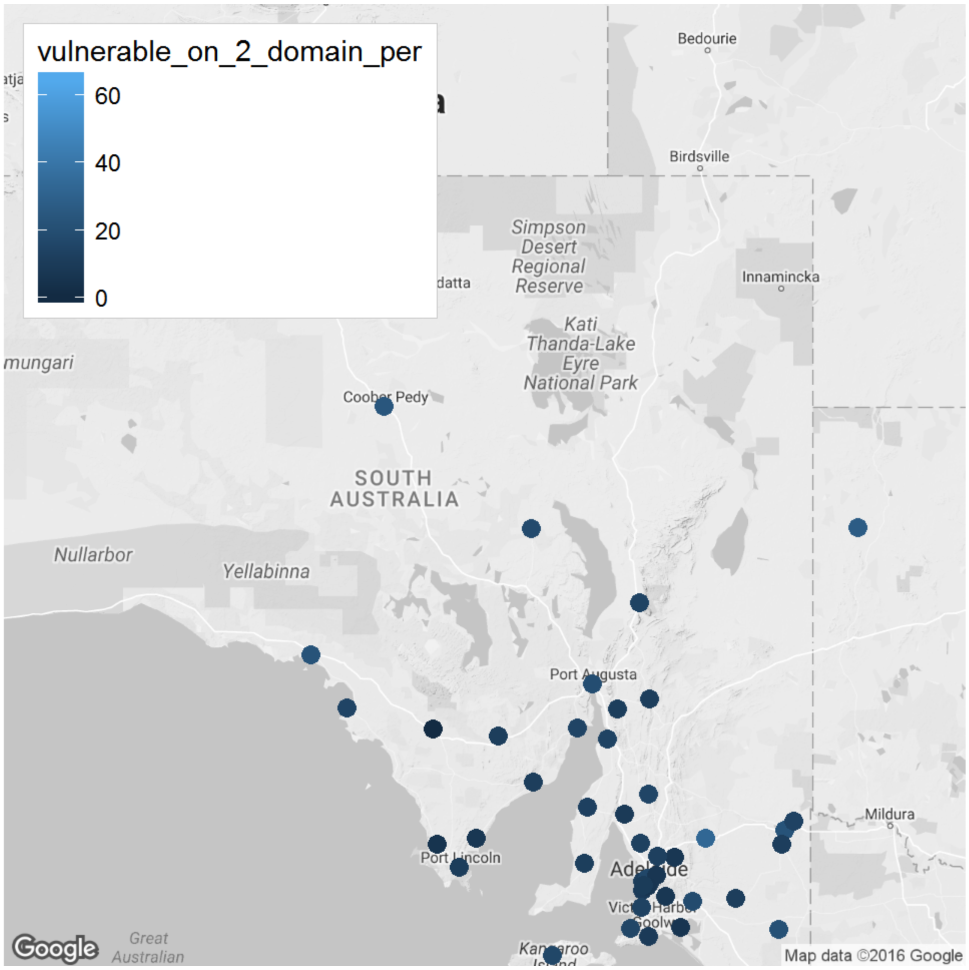
```
## Warning: geocode failed with status ZERO_RESULTS, location = "Anangu
## Pitjantjatjara"
```

```
data_new <- data
data_new <- cbind(data_new,lonlat) #the latitude and longitude is added to the data
SAMap <- qmap(location = "South Australia", zoom = 6, color="bw", legend= "topleft")
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

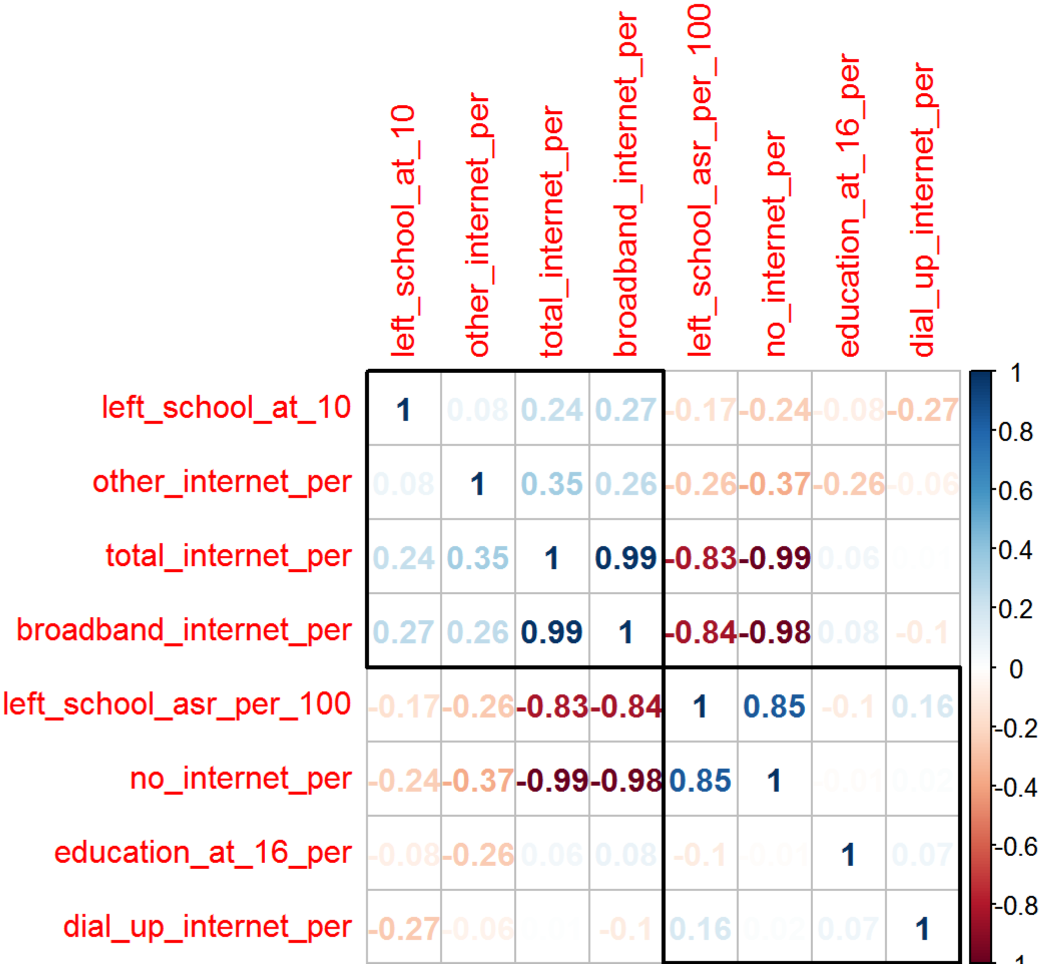
```
SAMap + geom_point(aes(x = lon, y = lat, colour = vulnerable_on_2_domain_per), data = data_new,
size = 3)
```

```
## Warning: Removed 24 rows containing missing values (geom_point).
```

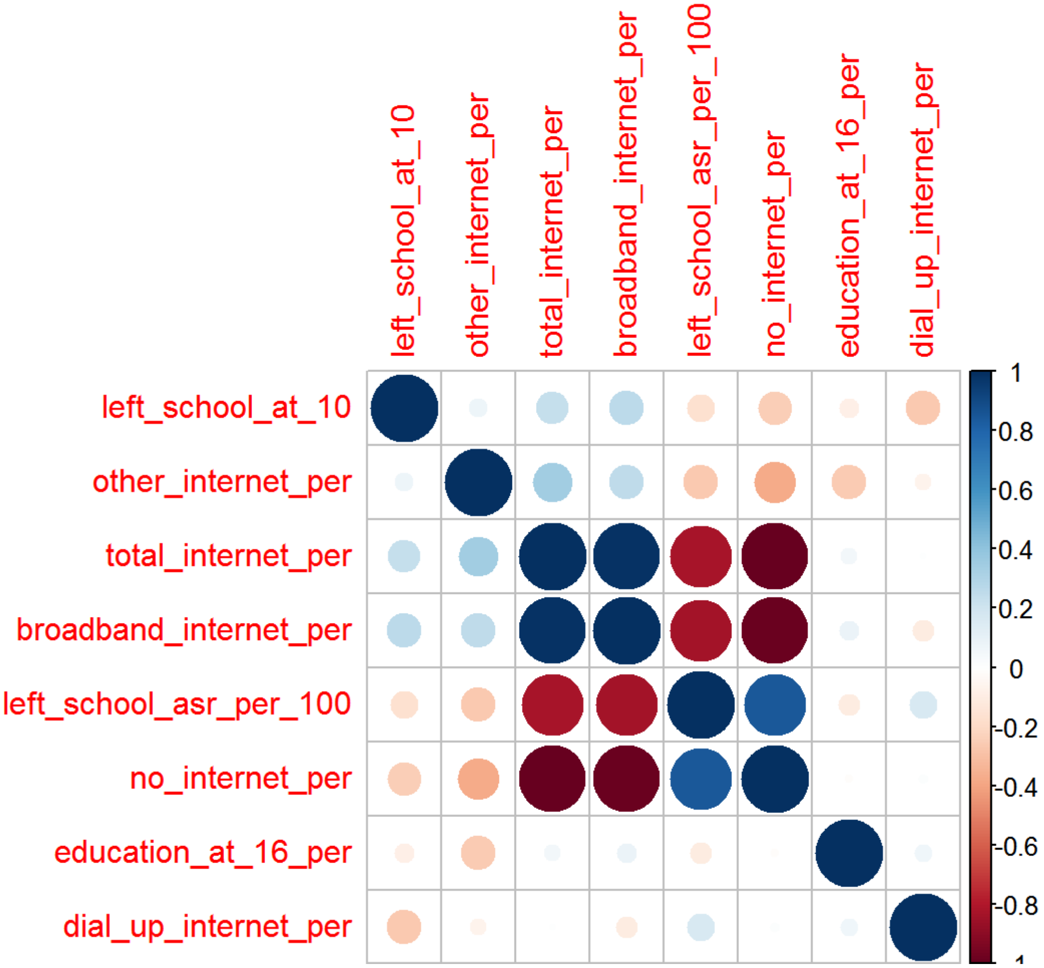


4. Performing Correlation Analysis

```
predictors <- data %>%
  dplyr::select(-c(Code,Name, vulnerable_on_2_domain_per, LGA_type))

corr <- cor(predictors)
corrplot(corr, diag=T, method="number", order='hclust', addrect = 2, tl.cex=1)
```



```
corrplot(corr, diag=T, order='hclust', tl.cex=1)
```



Explaination of Correlation plot

The correlation plot clearly shows that these variable pairs are highly correlated:

- (Positive) Total internet connection percentage and broadband internet percentage
- (Negative) Total internet connection percentage and Student who left school at 10 (ASR per 100)
- (Negative) Total internet connection percentage and no internet connection percentage
- (Negative) Broadband internet percentage and Student who left school at 10 (ASR per 100)
- (Negative) Broadband internet percentage and no internet connection percentage
- (Positive) No internet connection percentage and Student who left school at 10 (ASR per 100)

Also, No internet connection percentage is not correlated to percentage of children going to school full time and thus they can be used to as a pair of predictors.

5. Regression model specification and refining:

```
#running regression
reg_fit <- lm( vulnerable_on_2_domain_per ~ education_at_16_per +
                 no_internet_per +
                 LGA_type, data = data)
#Reviewing regression summary for model1
summary(reg_fit)
```

```
##
## Call:
## lm(formula = vulnerable_on_2_domain_per ~ education_at_16_per +
##     no_internet_per + LGA_type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.914  -2.802  -1.167   2.872  17.602
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                57.30632   10.11545   5.665 4.42e-07 ***
## education_at_16_per        -0.16648    0.05086  -3.273  0.00177 **
## no_internet_per             0.35275    0.10703   3.296  0.00165 **
## LGA_typeC                 -40.37263    7.28698  -5.540 7.09e-07 ***
## LGA_typeDC                -40.83534    6.82732  -5.981 1.32e-07 ***
## LGA_typeM                 -38.31829    7.97964  -4.802 1.09e-05 ***
## LGA_typeRC                -37.61549    8.33937  -4.511 3.06e-05 ***
## LGA_typeRegC              -41.08524    9.10466  -4.513 3.04e-05 ***
## LGA_typeT                 -40.62743    8.69455  -4.673 1.73e-05 ***
## LGA_typeUnincorporated SA -29.36646    8.63448  -3.401  0.00120 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.078 on 60 degrees of freedom
## Multiple R-squared:  0.7225, Adjusted R-squared:  0.6809
## F-statistic: 17.36 on 9 and 60 DF,  p-value: 1.088e-13
```

Comments on Fstatistic and p-value for model-1

The overall p-value for F-statistic is 1.088e-13 which is less that 1% hence the model is very significant.

Comments on variability for model-1

The R-squared value(0.7225) for the model shows a variaility of about 72.25%. The variability is hence fairly high.

The Adjusted R-squared value which accounts for the number of predictor variables is 0.6809. This value gives us a more effective/realistic assesment of the variance. The Adjusted R-squared value also shows that the The model is good for regression as the variability is fairly high.

Analysis of Predictor: education_at_16_per

Estimate of the education_at_16_per co-effeicient is -0.16648 whose p-value = 0.00177. The p-value is less than 1%. Hence the coefficient is very significant.

The p-value is associated with 3 stars. This coeffeciet is hence very significant as the p-value is tending to 0

Further we interpret that if education_at_16_per increases by 1 unit, vulnerable_on_2_domain_per decreases by 0.16648

Analysis of Predictor: Predictor: no_internet_per

Estimate of the no_internet_per co-effeicient is 0.35275 whose p-value = 0.00165. The p-value is less than 1%. Hence the no_internet_per coefficient is also very significant.

The p-value is associated with 2 stars. This coeffecient is hence substantially significant as the p-value is tending to 0.001

Further we interpret that if no_internet_per increases by 1 unit, vulnerable_on_2_domain_per increases by 0.35275 units

Analysis of Predictor: LGA_type

LGA_typeC

Estimate of the LGA_typeC co-effeicient is -40.37263 whose p-value = 7.09e-07 . The p-value is less than 1%. Hence the LGA_typeC coefficient is very significant.

The p-value is associated with 2 stars. This coeffecient is hence substantially significant as the p-value is tending to 0.001

Further we interpret that if LGA_typeC increases by 1 unit, vulnerable_on_2_domain_per decreases by -40.37263 units

LGA_typeDC

Estimate of the LGA_typeDC co-effeicient is -40.83534 whose p-value = 1.32e-07. The p-value is less than 1%. Hence the LGA_typeDC coefficient is very significant

The p-value is associated with 3 stars. This coeffecient is hence very significant as the p-value is tending to 0

Further we interpret that if LGA_typeDC increases by 1 unit, vulnerable_on_2_domain_per decreases by -40.83534 units

LGA_typeM

Estimate of the LGA_typeM co-effeicient is -38.31829 whose p-value = 1.09e-05. The p-value is less than 1%. Hence the LGA_typeM coefficient is very significant

The p-value is associated with 3 stars. This coeffecient is hence very significant as the p-value is tending to 0

Further we interpret that if LGA_typeM increases by 1 unit, vulnerable_on_2_domain_per decreases by -38.31829 units

LGA_typeRC

Estimate of the LGA_typeRC co-effeicient is -37.61549 whose p-value = 3.06e-05. The p-value is less than 1%. Hence the LGA_typeRC coefficient is very significant

The p-value is associated with 3 stars. This coeffecient is hence very significant as the p-value is tending to 0

Further we interpret that if LGA_typeRC increases by 1 unit, vulnerable_on_2_domain_per decreases by -37.61549 units

LGA_typeRegC

Estimate of the LGA_typeRegC co-effeicient is -41.08524 whose p-value = 3.04e-05. The p-value is less than 1%. Hence the LGA_typeRC coefficient is very significant

The p-value is associated with 3 stars. This coeffecient is hence very significant as the p-value is tending to 0

Further we interpret that if LGA_typeRegC increases by 1 unit, vulnerable_on_2_domain_per decreases by -41.08524 units

LGA_typeT

Estimate of the LGA_typeT co-effeicient is -40.62743 whose p-value = 1.73e-05. The p-value is less than 1%. Hence the LGA_typeT coefficient is very significant

The p-value is associated with 3 stars. This coeffecient is hence very significant as the p-value is tending to 0

Further we interpret that if LGA_typeT increases by 1 unit, vulnerable_on_2_domain_per decreases by -40.62743 units

LGA_typeUnincorporated SA

Estimate of the LGA_typeUnincorporated SA co-effeicient is -29.36646 whose p-value = 0.00120. The p-value is less than 1%. Hence the LGA_typeUnincorporated SA coefficient is very significant

The p-value is associated with 2 stars. This coeffecient is hence substantially significant as the p-value is tending to 0.001

Further we interpret that if LGA_typeUnincorporated SA increases by 1 unit, vulnerable_on_2_domain_per decreases by -29.36646 units

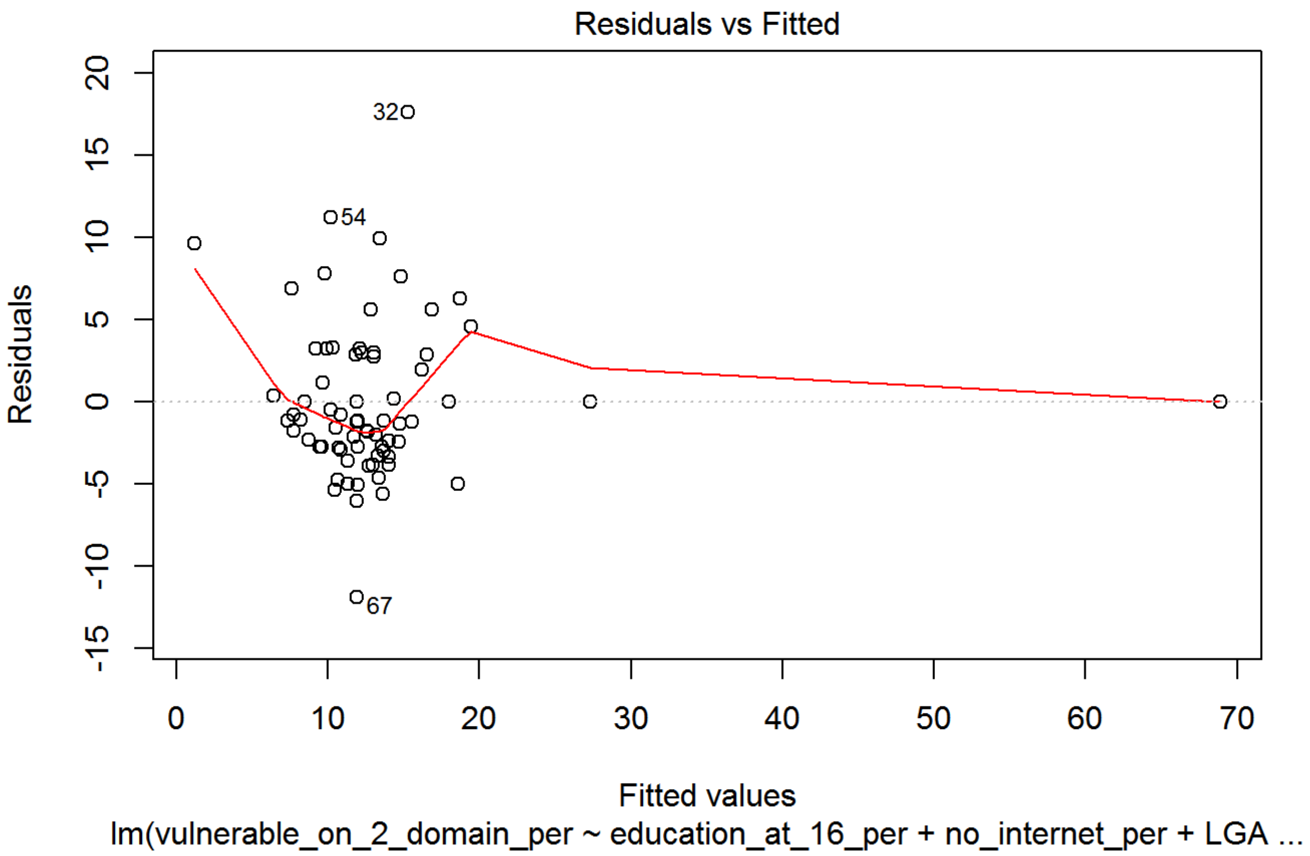
Comments on Residual standard error for model1

We find that the residual standard error is only 5.078 on 60 degrees of freedom. This tells us that the vulnerable_on_2_domain_per value predicted by the model is fairly close to the actual value.

6. Running residual Diagnostics

RESIDUALS VS FITTED PLOT

```
#Running residual analysis

# RESIDUALS VS FITTED PLOT
plot(reg_fit, which=1  ) # plot regression  diagnostics plot for ref_fit
```



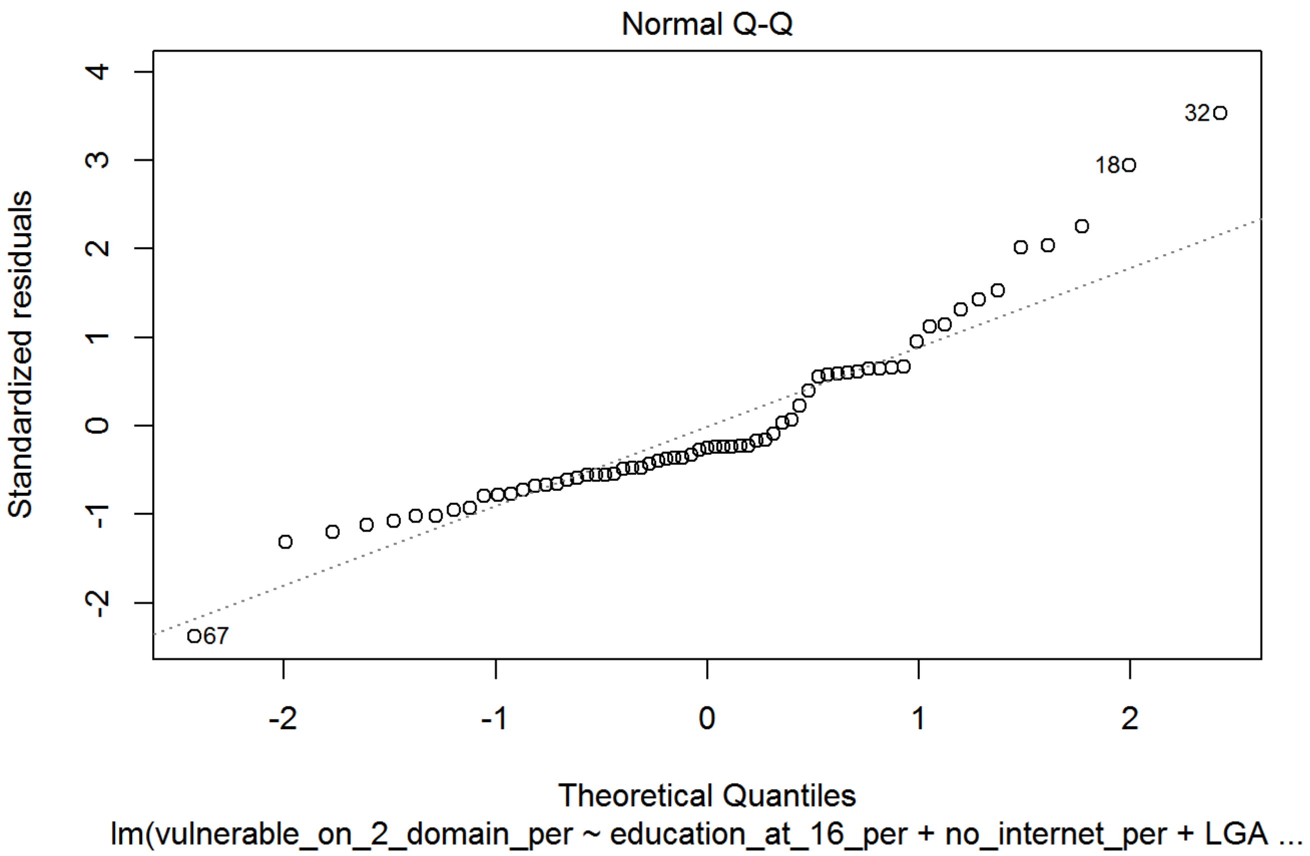
Analysis - RESIDUALS VS FITTED PLOT :

The RESIDUALS VS FITTED PLOT depicts that the residual points are randomly scattered on either side of regression line. They do not form any particular pattern. Hence, the plot indicates that the model does not have a non-linear relationship and is therefore valid.

Check for Normaility

```
#CHECK FOR NORMAITY -> Q-Q PLOT
plot(reg_fit, which=2)
```

```
## Warning: not plotting observations with leverage one:
##   4, 19, 27, 37, 70
```



Analysis - Normality Check:

The plot shows that the residuals are normally distributed. The residuals do not deviate severely from the median.

This further confirms that this is a good model

Outlier Test

```
#RUNNING CHECK FOR OUTLIERS -OUTLIER TEST
outlierTest(reg_fit)
```

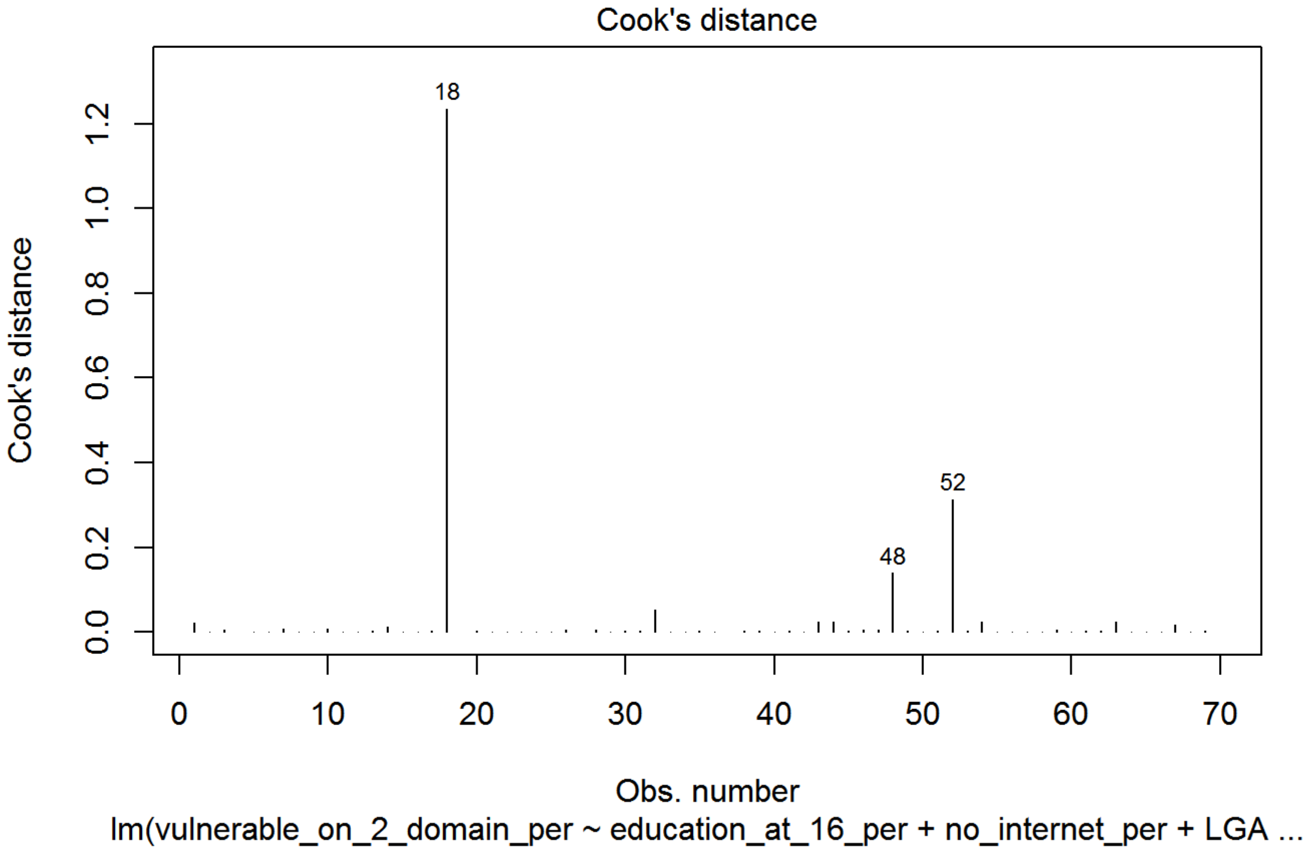
```
##    rstudent unadjusted p-value Bonferonni p
## 32 3.944125          0.0002153     0.013994
```

Analysis - Outlier Test:

The test indicates that the row# 18 could be an outlier for the predicted response.

Influential Observations : Cooks’D Plot

```
#CHECKING FOR INFLUENTIAL OBSERVATIONS USING COOK'S D PLOT

plot(reg_fit, which=4)   # plot regression (cook's d-plot)  diagnostics for fit1
```



Analysis - Cooks’s D Plot:

The Cook’s d plot once again indicates that row#18 has the highest cook’s distance and is hence a outlier. row#18 could possibly be an infuential variable.

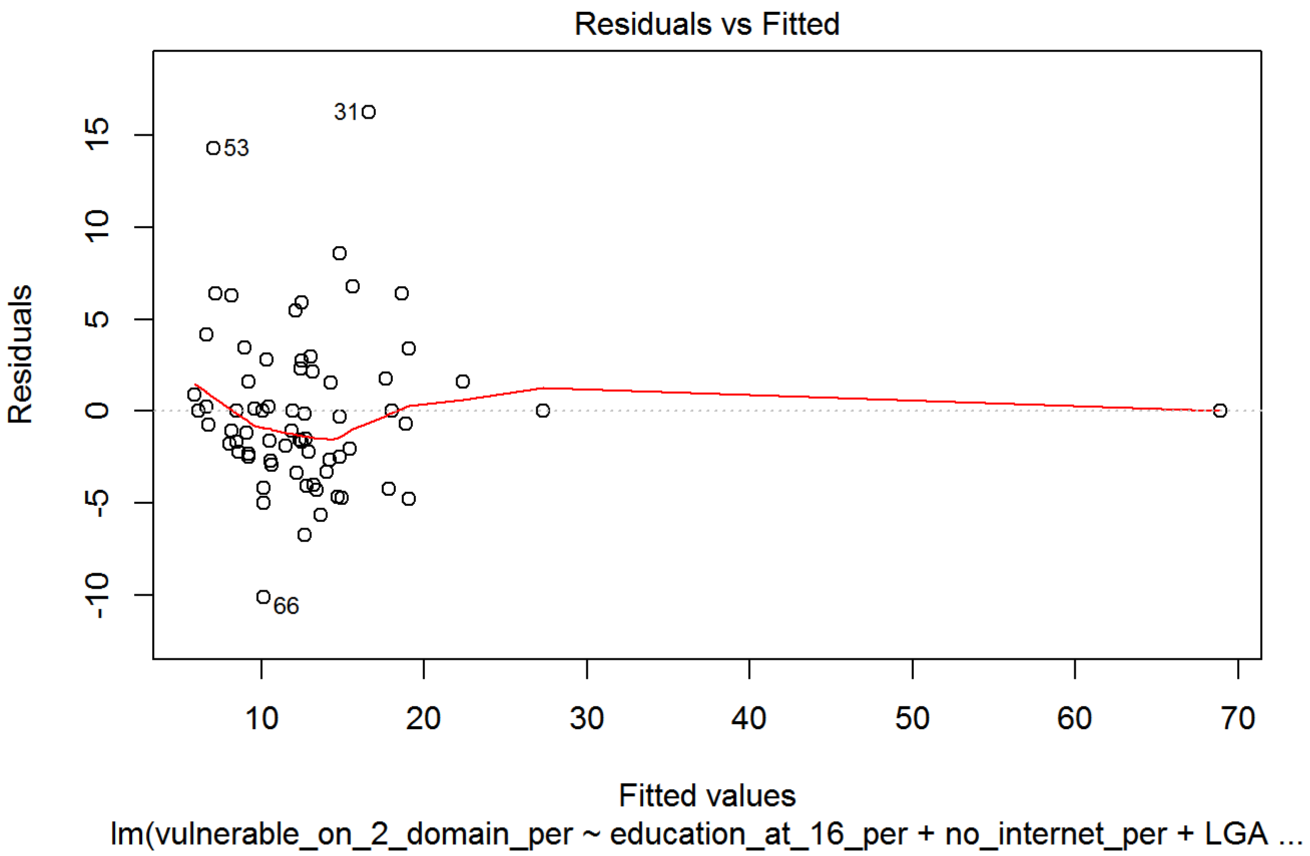
Let us validate this by excluding row#18 from our analysis as follows:

Influential Observations - On removing row number 18

```
data_new <- data

data_new <- data_new %>% filter(row.names(data_new)!= 18)  # exclude outlier from the modelling

#Therefore -  running regression on iris_new2
reg_fit2 <- lm( vulnerable_on_2_domain_per ~ education_at_16_per +
                 no_internet_per  +
                 LGA_type,
               data = data_new)


#Reviewing regression summary - regression analysis
summary(reg_fit2)
```

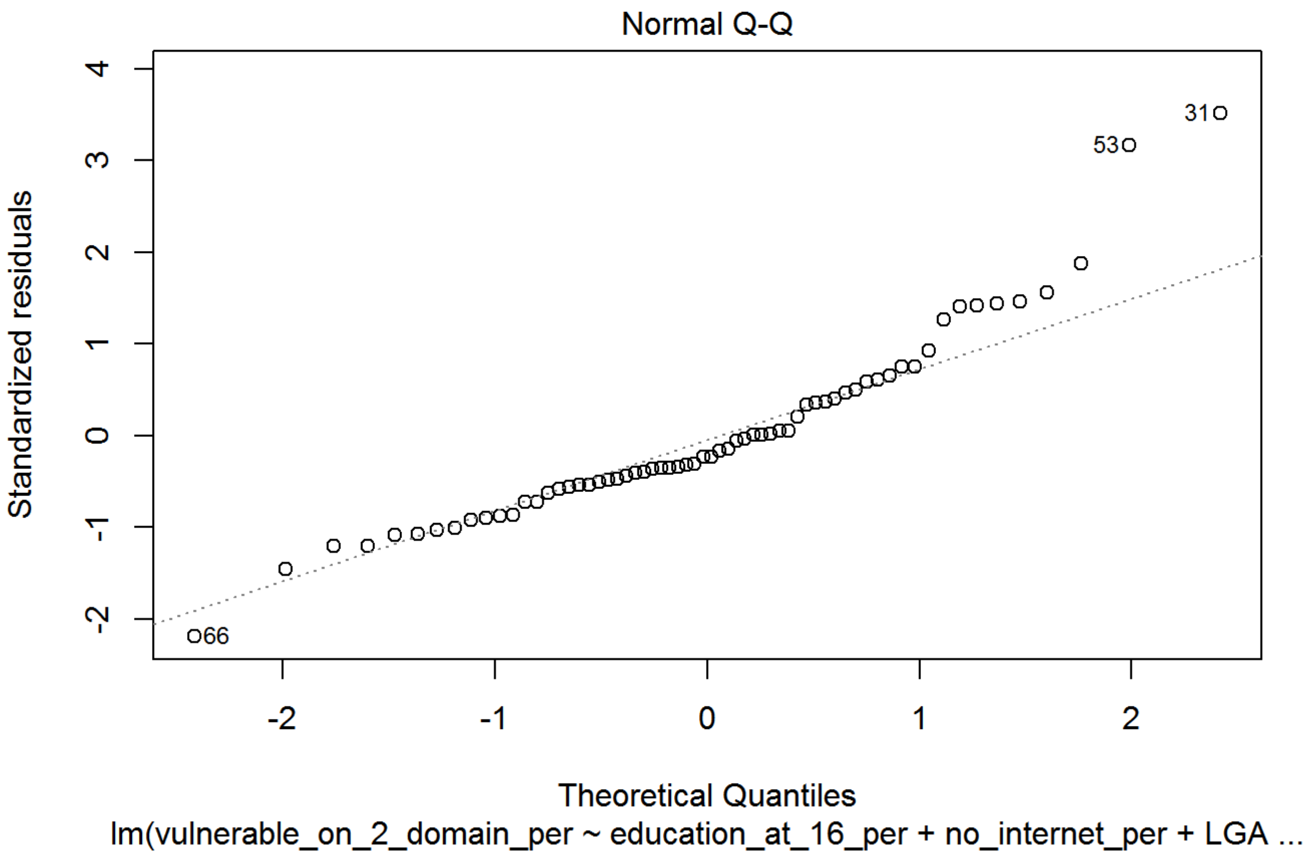
```
## 
## Call:
## lm(formula = vulnerable_on_2_domain_per ~ education_at_16_per + 
##     no_internet_per + LGA_type, data = data_new)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.1598  -2.5136  -0.6755   1.7730  16.2657 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                74.52790   10.89689   6.839 5.06e-09 ***
## education_at_16_per        -0.34083    0.07279  -4.683 1.71e-05 ***
## no_internet_per             0.30861    0.10080   3.062  0.00331 ** 
## LGA_typeC                 -41.96438    6.81506  -6.158 7.07e-08 ***
## LGA_typeDC                -42.42409    6.38752  -6.642 1.09e-08 ***
## LGA_typeM                 -40.83959    7.48512  -5.456 1.01e-06 ***
## LGA_typeRC                -40.03912    7.81570  -5.123 3.48e-06 ***
## LGA_typeRegC              -42.48733    8.50329  -4.997 5.52e-06 ***
## LGA_typeT                 -42.42254    8.12910  -5.219 2.45e-06 ***
## LGA_typeUnincorporated SA -34.58630    8.22102  -4.207 8.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.737 on 59 degrees of freedom
## Multiple R-squared:  0.7624, Adjusted R-squared:  0.7261 
## F-statistic: 21.03 on 9 and 59 DF,  p-value: 2.404e-15
```

```
# RESIDUALS VS FITTED PLOT
plot(reg_fit2, which=1  ) # plot regression  diagnostics plot for ref_fit2
```

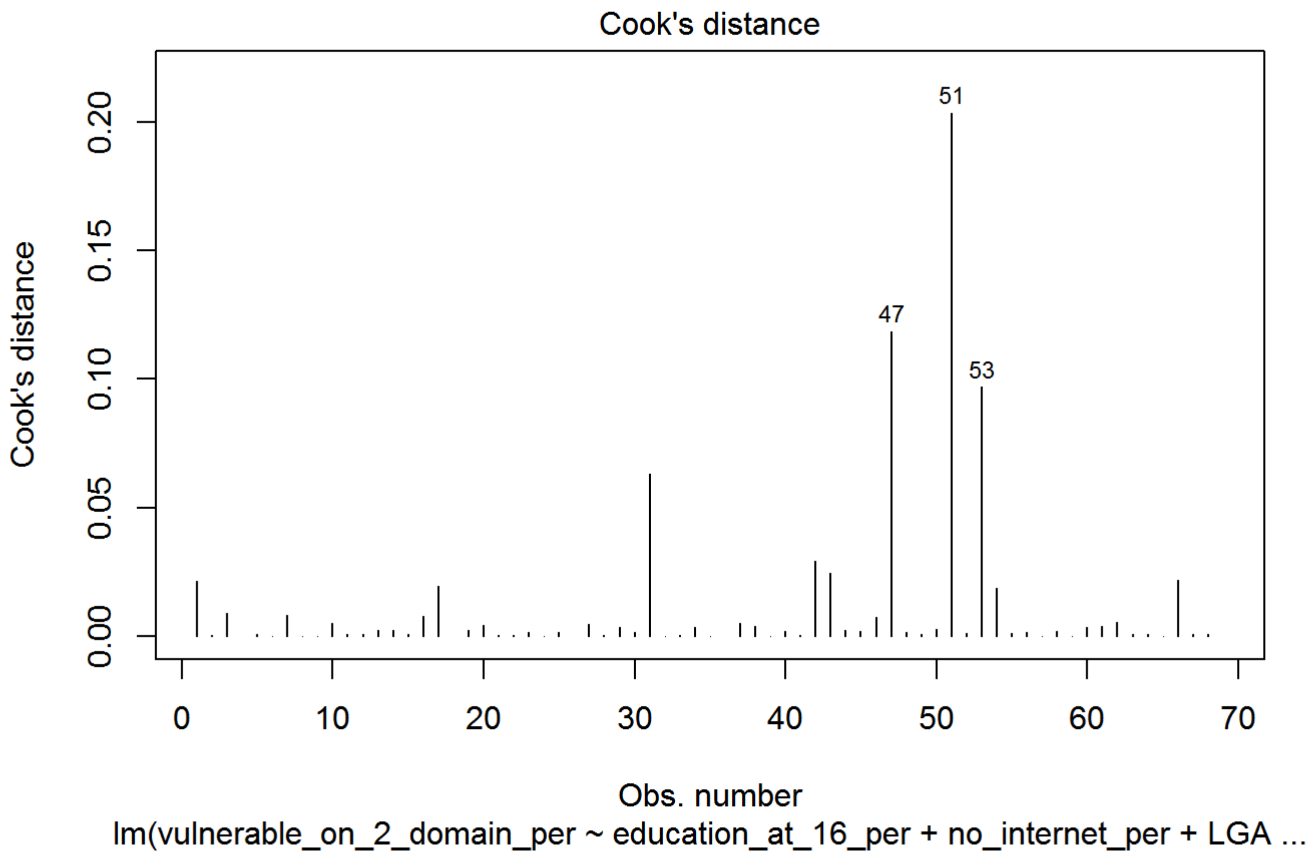


```
#CHECK FOR NORMAITY -> Q-Q PLOT
plot(reg_fit2, which = 2)
```

```
## Warning: not plotting observations with leverage one:
##   4, 18, 26, 36, 69
```



```
#CHECKING FOR INFLUENTIAL OBSERVATIONS USING COOK'S D PLOT
plot(reg_fit2, which=4)   # plot regression (cook's d-plot)  diagnostics for ref_fit2
```



###

Analysis:

On excluding row#18 and performing regression analysis/residual analysis, we find that the new model has 3 outliers viz 47 51 53. Further, we observe that the residuals deviate from the edian to a greater extent in this case.

Row#18 is hence an influencial variable to the linear regression model and we must therefore not exclude it.