

Data Activities: Code and Output

by Maria Ingold

for UoEO MSc AI Numerical Analysis, 2024

CRIME SURVEY ACTIVITIES

DATA ACTIVITY 1.1: Download Crime Survey

Download the **Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset** from its catalogue page. It is an open access dataset which the data are available to download without any registration with the UK Data Service.

Ensure you download and save the SPSS.sav version of the dataset. Save the data into a folder that you would like to save all your data and R activities in as we will return to this dataset regularly during the module.

Set Working Directory

```
> setwd("~/Github/DataActivities")
```

Import packages

```
> library(tidyverse) # For data manipulation
— Attaching core tidyverse packages ————— tidyverse 2.0.0 —
✓ dplyr      1.1.4      ✓ readr      2.1.5
✓ forcats   1.0.0      ✓ stringr    1.5.1
✓ ggplot2   3.5.0      ✓ tibble     3.2.1
✓ lubridate 1.9.3      ✓ tidyr      1.3.1
✓ purrr     1.0.2
— Conflicts ————— tidyverse_conflicts() —
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
i Use the conflicted package to force all conflicts to become errors
```

```
> library(haven) # For reading SPSS files and as_factor
```

```
> library(psych) # For describe function
```

```
Attaching package: 'psych'
```

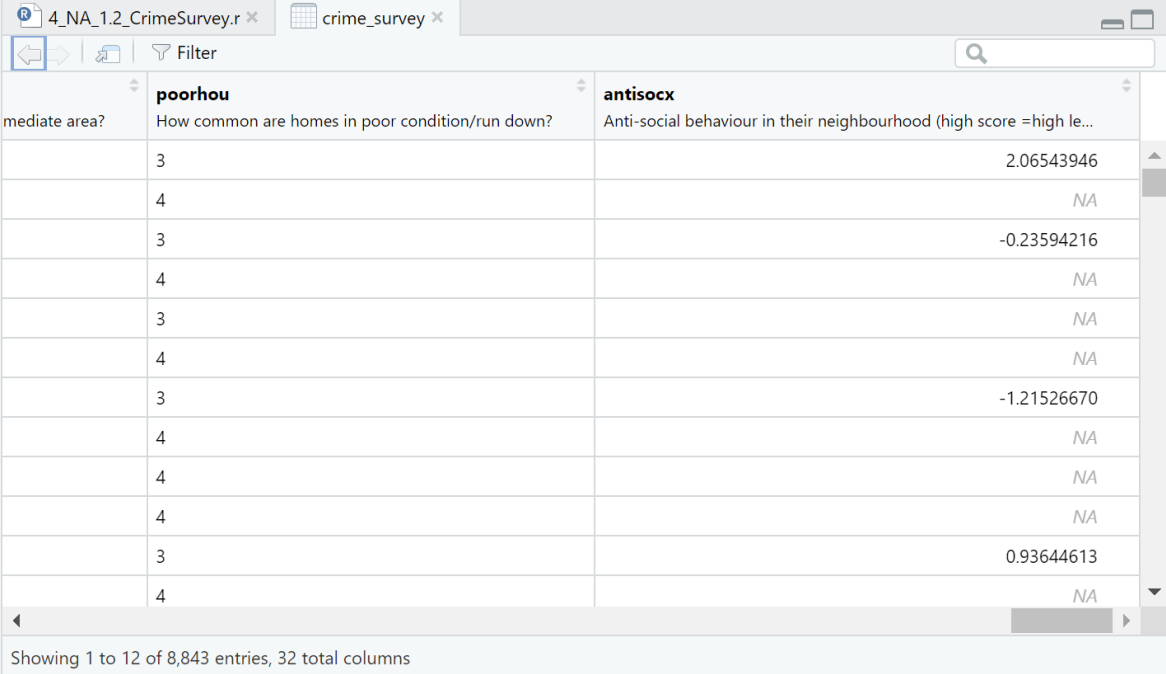
```
The following objects are masked from 'package:ggplot2':
```

```
  %+%, alpha
```

Load the data

```
> crime_survey <- haven::read_sav("csew1314teachingopen.sav")
```

```
> view(crime_survey)
```



The screenshot shows the RStudio 'View' window for the 'crime_survey' dataset. The window title is '4_NA_1.2_CrimeSurvey.r'. The table has three columns: 'mediate area?', 'poorhou', and 'antisocx'. The 'poorhou' column is described as 'How common are homes in poor condition/run down?' and the 'antisocx' column is described as 'Anti-social behaviour in their neighbourhood (high score =high le...'. The table displays 12 rows of data. The 'poorhou' column values are 3, 4, 3, 4, 3, 4, 3, 4, 4, 4, 3, 4. The 'antisocx' column values are 2.06543946, NA, -0.23594216, NA, NA, NA, -1.21526670, NA, NA, NA, 0.93644613, NA. The status bar at the bottom indicates 'Showing 1 to 12 of 8,843 entries, 32 total columns'.

mediate area?	poorhou	antisocx
	3	2.06543946
	4	NA
	3	-0.23594216
	4	NA
	3	NA
	4	NA
	3	-1.21526670
	4	NA
	4	NA
	4	NA
	3	0.93644613
	4	NA

Display a bit about the whole dataset

```
> names(crime_survey) # Display the names of the variables in the dataset
[1] "rowlabel" "split" "sex" "yrsarea" "resyrago" "work2"
[7] "tenure1" "livharm1" "agegrp7" "ethgrp2a" "educat3" "rural2"
[13] "edeprivex" "wdeprivex" "IndivWgtx" "cause2m" "walkdark" "walkday"
[19] "homealon" "wburgl" "wmugged" "wcarstol" "wfromcar" "wrape"
[25] "wattack" "wraceatt" "worryx" "bcsvictim" "rubbcomm" "vandcomm"
[31] "poorhou" "antisocx"
```

```

> glimpse(crime_survey) # Display the structure of the dataset
Rows: 8,843
Columns: 32
$ rowlabel <dbl> 137068050, 147461190, 137116250, 147354190, 137061230, 1368982...
$ split <dbl+lbl> 1, 3, 1, 3, 3, 3, 1, 2, 4, 1, 1, 2, 3, 4, 2, 3, 1, 4, 3, 1...
$ sex <dbl+lbl> 2, 2, 2, 2, 2, 2, 1, 2, 1, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1, 2...
$ yrsarea <dbl+lbl> 7, 6, 7, 7, 7, 7, 6, 5, 7, 7, 4, 5, 7, 7, 7, 7, 3, 7, 2, 7...
$ resyrago <dbl+lbl> NA, NA, 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
$ work2 <dbl+lbl> 1, 2, 2, 1, 2, 2, 1, 1, 2, 2, 1, 2, 2, 1, 2, 2, 1, 2, 2, 1...
$ tenure1 <dbl+lbl> 2, 1, 4, 2, 4, 1, 4, 1, 1, 1, 2, 1, 4, 1, 1, 1, 4, 1, 1, 2...
$ livharm1 <dbl+lbl> 3, 1, 6, 1, 6, 1, 1, 1, 1, 3, 1, 1, 1, 3, 6, 4, 6, 3, 1...
$ agegrp7 <dbl+lbl> 4, 5, 5, 5, 6, 6, 4, 5, 5, 7, 2, 7, 7, 4, 4, 7, 4, 6, 5, 3...
$ ethgrp2a <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1...
$ educat3 <dbl+lbl> 4, 4, 4, 2, 1, 2, 1, 4, 4, 3, 4, 3, 1, 2, 3, 1, 4, 3, 3, 4...
$ rural2 <dbl+lbl> 1, 2, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1...
$ edeprivex <dbl> 2, 4, 1, 1, 3, 2, 1, 5, 4, 5, 1, NA, 2, 1, 3, 4, 5, 2, 1, 2...
$ wdeprivex <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 4, NA, NA, NA, NA,...
$ Indivwgtx <dbl> 0.5434830, 1.2128379, 0.5696113, 0.9942963, 0.4102766, 1.20282...
$ cause2m <dbl+lbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 7, NA, NA, NA, NA, 4, NA...
$ walkdark <dbl+lbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, 2, NA...
$ walkday <dbl+lbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, 1, NA...
$ homealon <dbl+lbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, 2, NA...
$ wburgl <dbl+lbl> NA, 3, NA, 2, 2, 3, NA, NA, NA, NA, NA, NA, 3, NA, NA...
$ wmugged <dbl+lbl> NA, 4, NA, 3, 2, 4, NA, NA, NA, NA, NA, NA, 3, NA, NA...
$ wcarstol <dbl+lbl> NA, NA, NA, NA, NA, NA, NA, NA, 3, NA, NA, NA, 4, NA, NA...
$ wfromcar <dbl+lbl> NA, NA, NA, NA, NA, NA, NA, NA, 3, NA, NA, NA, 3, NA, NA...
$ wrapped <dbl+lbl> NA, 4, NA, 4, 2, 4, NA, NA, NA, NA, NA, NA, 3, NA, NA...
$ watack <dbl+lbl> NA, 4, NA, 3, 2, 3, NA, NA, NA, NA, NA, NA, 3, NA, NA...
$ wraceatt <dbl+lbl> NA, 4, NA, 4, 3, 4, NA, NA, NA, NA, NA, NA, 4, NA, NA...
$ worryx <dbl> NA, -1.1319020, NA, -0.2575738, 1.1841520, -0.8226912, NA, NA,...
$ bcsvictim <dbl+lbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ rubbcomm <dbl+lbl> 3, 4, 3, 4, 3, 4, 3, 4, 4, 4, 3, 4, 3, 2, 3, 4, 3, 4, 3, 3...
$ vandcomm <dbl+lbl> 3, 4, 4, 4, 3, 4, 3, 4, 4, 4, 3, 4, 3, 4, 4, 4, 3, 4, 4, 3...
$ poorhou <dbl+lbl> 3, 4, 3, 4, 3, 4, 3, 4, 4, 4, 3, 4, 3, 1, 4, 4, 3, 4, 3, 3...
$ antisocx <dbl> 2.0654395, NA, -0.2359422, NA, NA, NA, -1.2152667, NA, NA, NA,...

```

```

> print(crime_survey) # Display the dataset
# A tibble: 8,843 x 32
  rowlabel split sex yrsarea resyrago work2 tenure1 livharm1 agegrp7
  <dbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>
1 137068050 1 [A (Expe... 2 [Fem... 7 [20 ... NA 1 [Yes] 2 [Buy... 3 [Sing... 4 [45-...
2 147461190 3 [C (Crim... 2 [Fem... 6 [10 ... NA 2 [No] 1 [Own... 1 [Marr... 5 [55-...
3 137116250 1 [A (Expe... 2 [Fem... 7 [20 ... 2 [No] 4 [Ren... 6 [Wido... 5 [55-...
4 147354190 3 [C (Crim... 2 [Fem... 7 [20 ... NA 1 [Yes] 2 [Buy... 1 [Marr... 5 [55-...
5 137061230 3 [C (Crim... 2 [Fem... 7 [20 ... NA 2 [No] 4 [Ren... 6 [Wido... 6 [65-...
6 136898230 3 [C (Crim... 2 [Fem... 7 [20 ... NA 2 [No] 1 [Own... 1 [Marr... 6 [65-...
7 135507330 1 [A (Expe... 1 [Mal... 6 [10 ... NA 1 [Yes] 4 [Ren... 1 [Marr... 4 [45-...
8 136450220 2 [B (Atti... 2 [Fem... 5 [5 y... NA 1 [Yes] 1 [Own... 1 [Marr... 5 [55-...
9 136111200 4 [D (Onli... 1 [Mal... 7 [20 ... NA 2 [No] 1 [Own... 1 [Marr... 5 [55-...
10 136599250 1 [A (Expe... 1 [Mal... 7 [20 ... NA 2 [No] 1 [Own... 1 [Marr... 7 [75+]

# i 8,833 more rows
# i 23 more variables: ethgrp2a <dbl+lbl>, educat3 <dbl+lbl>, rural2 <dbl+lbl>,
# edeprivex <dbl>, wdeprivex <dbl>, Indivwgtx <dbl>, cause2m <dbl+lbl>,
# walkdark <dbl+lbl>, walkday <dbl+lbl>, homealon <dbl+lbl>, wburgl <dbl+lbl>,
# wmugged <dbl+lbl>, wcarstol <dbl+lbl>, wfromcar <dbl+lbl>, wrapped <dbl+lbl>,
# watack <dbl+lbl>, wraceatt <dbl+lbl>, worryx <dbl>, bcsvictim <dbl+lbl>,
# rubbcomm <dbl+lbl>, vandcomm <dbl+lbl>, poorhou <dbl+lbl>, antisocx <dbl>
# i use `print(n = ...)` to see more rows

```

```

> summary(crime_survey) # Summary statistics for all variables
  rowlabel      split      sex      yrsarea
Min.   :135230190  Min.   :1.000  Min.   :1.000  Min.   :1.000
1st Qu.:136324135  1st Qu.:1.000  1st Qu.:1.000  1st Qu.:5.000
Median :136731120  Median :2.000  Median :2.000  Median :6.000
Mean   :138726222  Mean   :2.456  Mean   :1.543  Mean   :5.511
3rd Qu.:137148275  3rd Qu.:3.000  3rd Qu.:2.000  3rd Qu.:7.000
Max.   :147639290  Max.   :4.000  Max.   :2.000  Max.   :7.000
      NA's :1

  resyrago      work2      tenure1      livharm1      agegrp7
Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :1.000
1st Qu.:1.000  1st Qu.:1.000  1st Qu.:1.000  1st Qu.:1.000  1st Qu.:3.000
Median :2.000  Median :1.000  Median :2.000  Median :2.000  Median :4.000
Mean   :1.568  Mean   :1.468  Mean   :2.416  Mean   :2.532  Mean   :4.129
3rd Qu.:2.000  3rd Qu.:2.000  3rd Qu.:4.000  3rd Qu.:3.000  3rd Qu.:6.000
Max.   :2.000  Max.   :2.000  Max.   :6.000  Max.   :6.000  Max.   :7.000
NA's   :7334  NA's   :2      NA's   :23     NA's   :13
      ethgrp2a      educat3      rural2      edeprivex      wdeprivex
Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :1.000
1st Qu.:1.000  1st Qu.:2.000  1st Qu.:1.000  1st Qu.:2.000  1st Qu.:2.000
Median :1.000  Median :3.000  Median :1.000  Median :3.000  Median :3.000
Mean   :1.244  Mean   :2.847  Mean   :1.236  Mean   :3.045  Mean   :3.057
3rd Qu.:1.000  3rd Qu.:4.000  3rd Qu.:1.000  3rd Qu.:4.000  3rd Qu.:4.000
Max.   :5.000  Max.   :5.000  Max.   :2.000  Max.   :5.000  Max.   :5.000
NA's   :10     NA's   :21     NA's   :703   NA's   :8140
      Indivwgtx      cause2m      walkdark      walkday
Min.   :0.2192  Min.   :1.000  Min.   :1.000  Min.   :1.000
1st Qu.:0.5725  1st Qu.:4.000  1st Qu.:1.000  1st Qu.:1.000
Median :0.8211  Median :5.000  Median :2.000  Median :1.000
Mean   :0.9957  Mean   :5.457  Mean   :2.089  Mean   :1.274
3rd Qu.:1.2357  3rd Qu.:7.000  3rd Qu.:3.000  3rd Qu.:1.000
Max.   :5.1740  Max.   :13.000  Max.   :4.000  Max.   :4.000
NA's   :6779  NA's   :6779  NA's   :6786  NA's   :6772
      homealon      wburgl      wmugged      wcarstol      wfromcar
Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :1.00  Min.   :1.000
1st Qu.:1.000  1st Qu.:2.000  1st Qu.:2.000  1st Qu.:3.00  1st Qu.:2.000
Median :1.000  Median :3.000  Median :3.000  Median :3.00  Median :3.000
Mean   :1.423  Mean   :2.676  Mean   :2.897  Mean   :3.02  Mean   :2.934
3rd Qu.:2.000  3rd Qu.:3.000  3rd Qu.:3.000  3rd Qu.:4.00  3rd Qu.:3.000
Max.   :4.000  Max.   :5.000  Max.   :5.000  Max.   :5.00  Max.   :4.000
NA's   :6771  NA's   :6650  NA's   :6658  NA's   :7080  NA's   :7111
      wrapped      wattack      wraceatt      worryx
Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :-1.389
1st Qu.:3.000  1st Qu.:2.000  1st Qu.:3.000  1st Qu.:-0.771
Median :4.000  Median :3.000  Median :4.000  Median :-0.226
Mean   :3.346  Mean   :2.927  Mean   :3.523  Mean   :-0.024
3rd Qu.:4.000  3rd Qu.:4.000  3rd Qu.:4.000  3rd Qu.:0.360
Max.   :5.000  Max.   :5.000  Max.   :5.000  Max.   :2.902
NA's   :6660  NA's   :6658  NA's   :6659  NA's   :6796
      bcsvictim      rubbcomm      vandcomm      poorhou      antisocx
Min.   :0.0000  Min.   :1.00  Min.   :1.000  Min.   :1.000  Min.   :-1.215
1st Qu.:0.0000  1st Qu.:3.00  1st Qu.:3.000  1st Qu.:3.000  1st Qu.:-0.788
Median :0.0000  Median :4.00  Median :4.000  Median :4.000  Median :-0.185
Mean   :0.1564  Mean   :3.42  Mean   :3.669  Mean   :3.548  Mean   :-0.007
3rd Qu.:0.0000  3rd Qu.:4.00  3rd Qu.:4.000  3rd Qu.:4.000  3rd Qu.:0.528
Max.   :1.0000  Max.   :5.00  Max.   :5.000  Max.   :5.000  Max.   :4.015
      NA's :6694

```

```

> n_count <- nrow(crime_survey) # Number of rows in the dataset

```

DATA ACTIVITY 1.2: Create summary statistic for antisocx

Using the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset, assess the level of anti-social behaviour that the survey respondents experience in their neighbourhood by **creating a summary statistic**, using the 'antisocx' variable.

```
> print(attr(crime_survey$antisocx, "label")) # Display antisocx label  
[1] "Anti-social behaviour in their neighbourhood (high score =high levels of anti-social behaviour)"
```

```
> summary(crime_survey$antisocx) # Summary statistics for antisocx  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
-1.215 -0.788  -0.185  -0.007  0.528   4.015  6694
```

This is a 5-Figure Summary:

- Minimum
- 1st Quartile
- Median
- Mean
- 3rd Quartile
- Maximum
- Number of Not Applicables

```
> describe(crime_survey$antisocx) # Descriptive statistics for antisocx  
vars  n  mean  sd median trimmed  mad  min  max range skew kurtosis  se  
x1    1 2149 -0.01 0.99  -0.18  -0.11 1.06 -1.22 4.01  5.23  0.8    0.23 0.02
```

The column names mean:

- item name
- item number
- number of valid cases
- mean
- standard deviation
- trimmed mean (with trim defaulting to .1)
- median (standard or interpolated)
- mad: median absolute deviation (from the median)
- minimum
- maximum
- skew
- kurtosis
- standard error

DATA ACTIVITY 2.1: Crime in 12 months prior to survey for `bcsvictim`

Explore whether survey respondents experienced any crime in the 12 months prior to the survey using the variable `bcsvictim`.

```
> print(attr(crime_survey$bcsvictim, "label")) # Display bcsvictim label
[1] "Experience of any crime in the previous 12 months"

> print(attr(crime_survey$bcsvictim, "labels")) # Display bcsvictim format
Not a victim of crime      Victim of crime
              0              1

> summary(crime_survey$bcsvictim) # Summary statistics for bcsvictim
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000 0.0000  0.0000  0.1564 0.0000  1.0000

> describe(crime_survey$bcsvictim) # Descriptive statistics for bcsvictim
  vars   n mean  sd median trimmed mad min max range skew kurtosis se
x1     1 8843 0.16 0.36     0    0.16  0  0  1     1 1.89     1.58  0

> victims <- sum(crime_survey$bcsvictim == 1) # Number victims in 12 months prior
> print(paste("Victims : ", victims, " out of: ", n_count)) # Display number
[1] "Victims : 1383 out of: 8843"
```

DATA ACTIVITY 2.2: Create `bcsvictim` frequency table

Create a frequency table to count if the survey respondents experienced any crime in the previous 12 months. Use the `table()` command.

```
> table(crime_survey$bcsvictim) # Frequency table for bcsvictim variable

  0     1
7460 1383
```

DATA ACTIVITY 2.3: Use `bcsvictim` labels using `as_factor`

Assess the results and decide if you need to convert this variable into a factor variable. Use `as_factor`.

While the following worked, it rewrote the table with this text. That wasn't what I wanted.

```
> table(as_factor(crime_survey$bcsvictim))

Not a victim of crime      Victim of crime
              7460              1383
```

DATA ACTIVITY 3.0: Subset of 75+ who were bcsvictim

Create a subset of individuals who belong to the '75+' age group and who were a 'victim of crime' that occurred in the previous 12 months. Save this dataset under a new name 'crime_75victim'.

```
> print(attr(crime_survey$agegrp7, "label")) # Display agegrp7 label
[1] "Age group (7 bands)"

> print(attr(crime_survey$agegrp7, "labels")) # Display agegrp7 label
16-24 25-34 35-44 45-54 55-64 65-74 75+
  1     2     3     4     5     6     7

> crime_75victim <- crime_survey %>%
+   filter(agegrp7 == 7 & bcsvictim == 1)

> print(paste("Victims 75+ : ", nrow(crime_75victim), " out of: ", n_count))
[1] "Victims 75+ : 67 out of: 8843"
```

DATA ACTIVITY 4.1: Boxplot for antisocx

Create a boxplot for the variable 'antisocx'

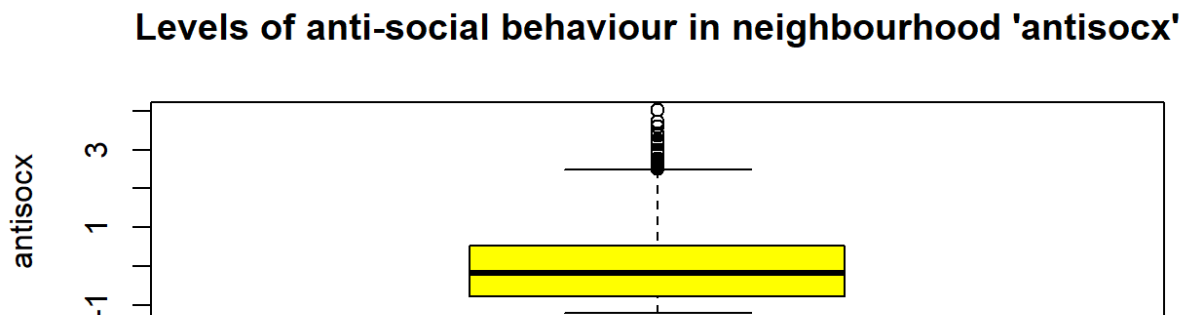
Follow the instructions below to create a boxplot for assessing levels of anti-social behaviour that the survey respondents experience in their neighbourhood (use the variable: antisocx).

If you're using 'graphics': Add "Levels of anti-social behaviour in neighbourhood 'antisocx'" as a title and colour the plot in purple and colour the outliers in blue.

If you're using 'ggplot2': Add "Levels of anti-social behaviour in neighbourhood 'antisocx' as a title, colour the plot in yellow and colour the outliers in red.

Boxplot gave me:

```
> boxplot(crime_survey$antisocx, main = "Levels of anti-social behaviour i
n neighbourhood 'antisocx'", ylab = "antisocx", col="yellow")
```



It didn't seem flexible enough to change outliers.

So, I changed the code to:

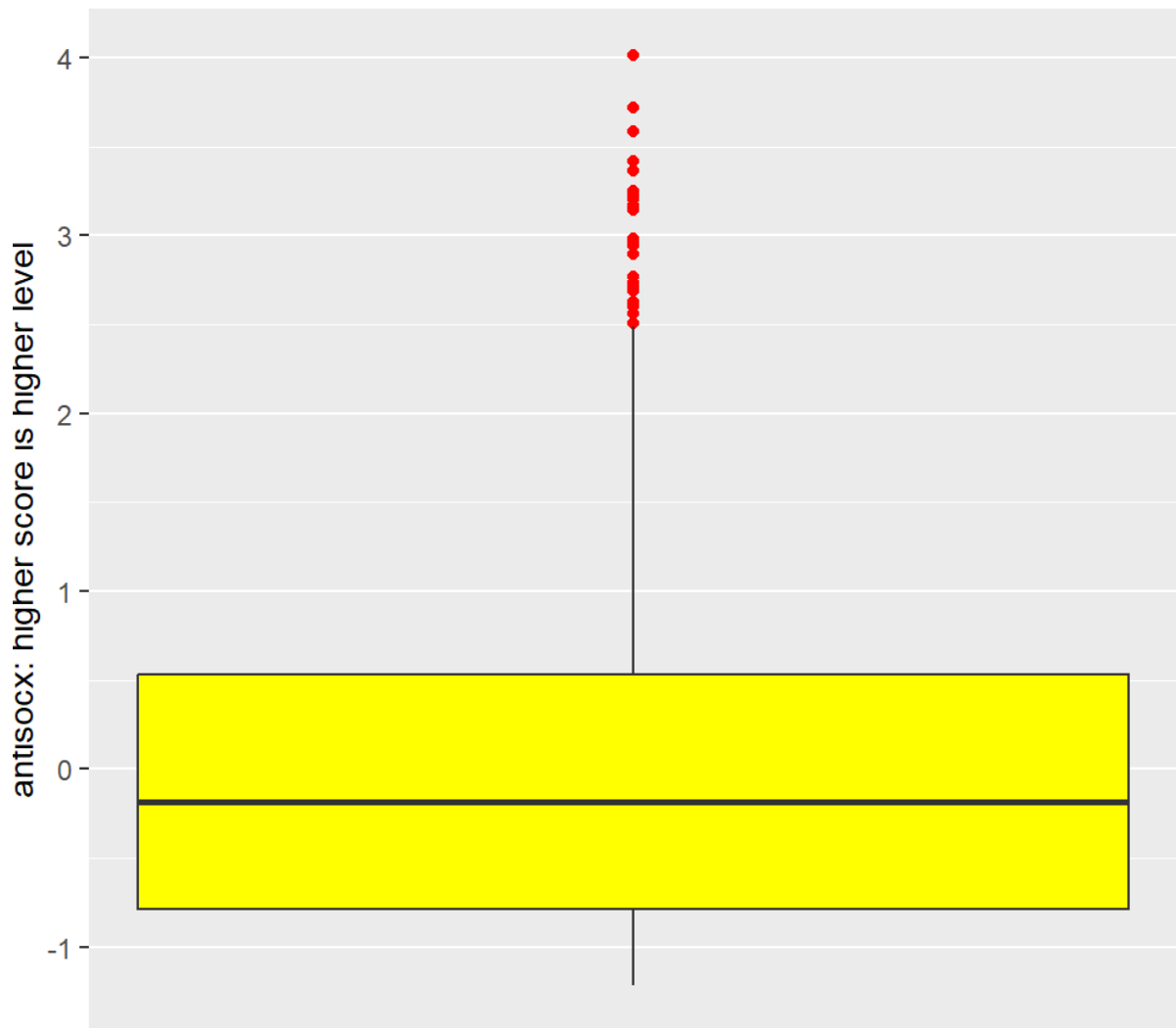
```
> ggplot(crime_survey, aes(y = antisocx)) +  
+   geom_boxplot(fill = "yellow", outlier.color = "red", na.rm = TRUE) +  
+   labs(title = "Levels of anti-social behaviour in neighbourhood 'antisocx'",  
+        y = "antisocx: higher score is higher level") +  
+   scale_x_continuous(breaks = NULL) # Remove x-axis labels
```

Because `geom_boxplot` includes:

```
geom_boxplot(  
  mapping = NULL,  
  data = NULL,  
  stat = "boxplot",  
  position = "dodge2",  
  ...,  
  outliers = TRUE,  
  outlier.colour = NULL,  
  outlier.color = NULL,  
  outlier.fill = NULL,  
  outlier.shape = 19,  
  outlier.size = 1.5,  
  outlier.stroke = 0.5,  
  outlier.alpha = NULL,  
  notch = FALSE,  
  notchwidth = 0.5,  
  staplewidth = 0,  
  varwidth = FALSE,  
  na.rm = FALSE,  
  orientation = NA,  
  show.legend = NA,  
  inherit.aes = TRUE  
)
```

Which gave me:

Levels of anti-social behaviour in neighbourhood 'antisocx'



DATA ACTIVITY 4.2: barplot / ggplot for bcsvictim

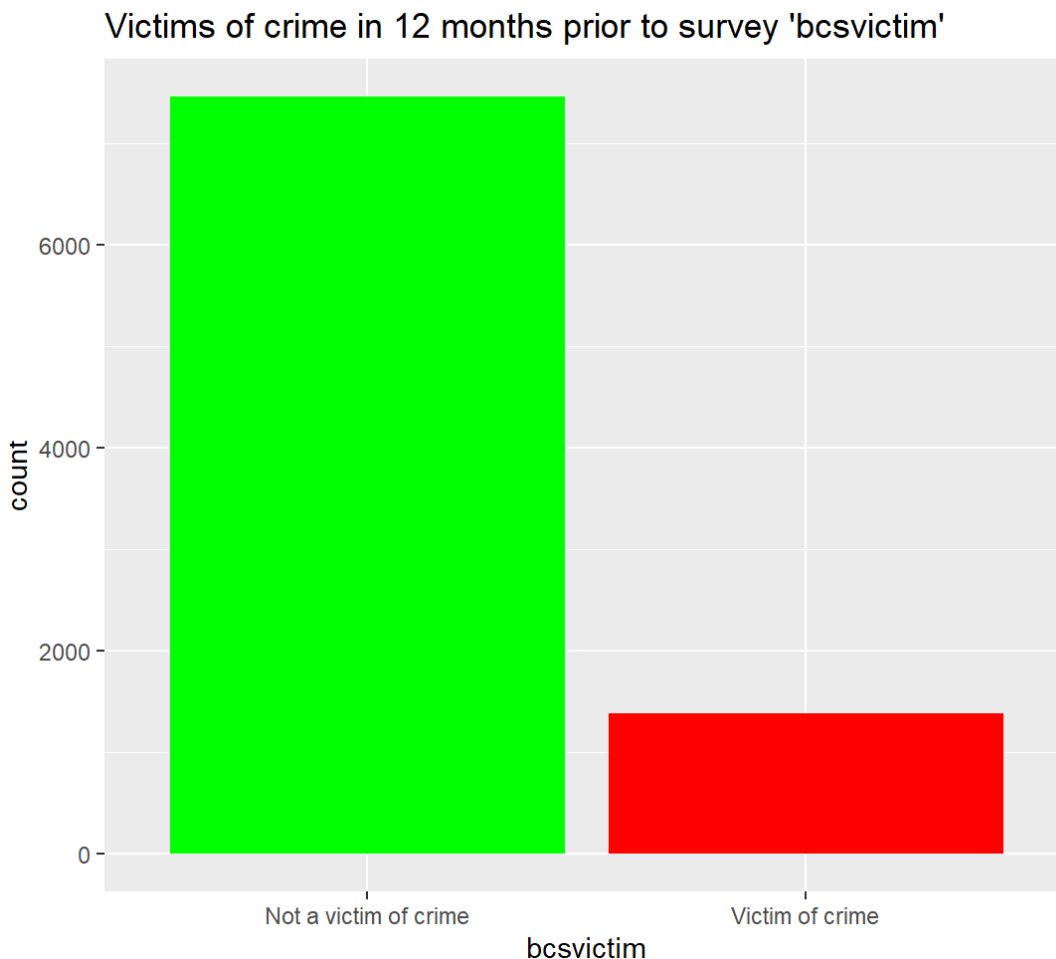
Create a bar plot using either the **barplot()** function or the **ggplot()** function to assess whether or not the survey respondents experienced crime in the 12 months prior to the survey (use the variable '**bcsvictim**'). Give the graph a suitable title and choose a colour for the bars (e.g., orange).

To just do a colour it's the following in the below.

```
> ggplot(crime_survey, aes(x = as_factor(bcsvictim))) +  
+   geom_bar(fill = "blue") +  
+   labs(title = "Victims of crime in 12 months prior to survey 'bcsvictim'",  
+         x = "bcsvictim")
```

I went further and set the bars to different colours depending on victim or not. That required using fill in aes to get the values, scale_fill_manual to set them and theme to remove the legend.

```
> ggplot(crime_survey, aes(x = as_factor(bcsvictim),  
+                           fill = as_factor(bcsvictim))) +  
+   geom_bar() +  
+   scale_fill_manual(values = c("Not a victim of crime" = "green",  
+                                "Victim of crime" = "red")) +  
+   labs(title = "Victims of crime in 12 months prior to survey 'bcsvictim'",  
+         x = "bcsvictim") +  
+   theme(legend.position = "none")
```



DATA ACTIVITY 7.1: Crosstab for bcsvictim and agegrp7

Create a crosstab to assess how individuals' experience of any crime in the previous 12 months **bcsvictim** vary by age group **agegrp7**. Create the crosstab with **bcsvictim** in the rows and **agegrp7** in the columns, and produce row percentages, rounded to 2 decimal places.

DATA ACTIVITY 7.2: Analyse crime likelihood

Looking at the crosstab you have produced, which age groups were the most likely, and least likely, to be victims of crime?