Module 2: Data activities

Guilherme Amorim

2024-08-08

Unit 1

Loading required packages

R

```r
# module 2, unit 1
# data activity 1.2 - levels of antisocial behaviour

library(haven)
library(skimr)
library(tidyverse)
```

Python

```python
# first install packages (from terminal)
# pip3 install numpy

# pip3 install pandas

# pip3 install matplotlib

# pip3 install pyreadstat

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

Loading dataset

R

```r
dataset <- read_sav("C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_prac

str(dataset)
```

```
## tibble [8,843 x 32] (S3: tbl_df/tbl/data.frame)
##  $ rowlabel : num [1:8843] 1.37e+08 1.47e+08 1.37e+08 1.47e+08 1.37e+08 ...
##   ..- attr(*, "label")= chr "Case identifier (9 digits)"
##   ..- attr(*, "format.spss")= chr "F12.0"
##  $ split    : dbl+lbl [1:8843] 1, 3, 1, 3, 3, 3, 1, 2, 4, 1, 1, 2, 3, 4, 2, 3, 1, 4,...
##    ..@ label      : chr "Follow-up module split"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:4] 1 2 3 4
##    .. ..- attr(*, "names")= chr [1:4] "A (Experiences of the police)" "B (Attitudes to the CJS)" "C
##  $ sex      : dbl+lbl [1:8843] 2, 2, 2, 2, 2, 2, 1, 2, 1, 1, 1, 1, 2, 1, 2, 2, 1, 1,...
##    ..@ label      : chr "Adult number 1 (respondent): Sex"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:2] 1 2
##    .. ..- attr(*, "names")= chr [1:2] "Male" "Female"
##  $ yrsarea  : dbl+lbl [1:8843] 7, 6, 7, 7, 7, 7, 6, 5, 7, 7, 4, 5, 7, 7, 7, 7, 3, 7,...
##    ..@ label      : chr "How long lived in this area"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:9] 1 2 3 4 5 6 7 8 9
##    .. ..- attr(*, "names")= chr [1:9] "Less than 12 months" "12 months but less than 2 years" "2 yea
##  $ resyrago : dbl+lbl [1:8843] NA, NA,  2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
##    ..@ label      : chr "Living at this address 12 months ago or not?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:2] 1 2
##    .. ..- attr(*, "names")= chr [1:2] "Yes" "No"
##  $ work2    : dbl+lbl [1:8843] 1, 2, 2, 1, 2, 2, 1, 1, 2, 2, 1, 2, 2, 1, 2, 2, 1, 2,...
##    ..@ label      : chr "Any paid work in last week"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:4] 1 2 8 9
##    .. ..- attr(*, "names")= chr [1:4] "Yes" "No" "Refusal" "Don't know"
##  $ tenure1  : dbl+lbl [1:8843] 2, 1, 4, 2, 4, 1, 4, 1, 1, 1, 2, 1, 4, 1, 1, 1, 4, 1,...
##    ..@ label      : chr "In which way do you occupy this accommodation?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:8] 1 2 3 4 5 6 8 9
##    .. ..- attr(*, "names")= chr [1:8] "Own it outright" "Buying it with the help of a mortgage or lo
##  $ livharm1 : dbl+lbl [1:8843] 3, 1, 6, 1, 6, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, 6, 4, 6,...
##    ..@ label      : chr "ONS harmonised marital status"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] -1 1 2 3 4 5 6
##    .. ..- attr(*, "names")= chr [1:7] "Not classified" "Married/Civil Partnered" "Cohabiting" "Singl
##  $ agegrp7  : dbl+lbl [1:8843] 4, 5, 5, 5, 6, 6, 4, 5, 5, 7, 2, 7, 7, 4, 4, 7, 4, 6,...
##    ..@ label      : chr "Age group (7 bands)"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] 1 2 3 4 5 6 7
##    .. ..- attr(*, "names")= chr [1:7] "16-24" "25-34" "35-44" "45-54" ...
##  $ ethgrp2a : dbl+lbl [1:8843] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1,...
##    ..@ label      : chr "Ethnic Group (5 categories)"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] 1 2 3 4 5 98 99
##    .. ..- attr(*, "names")= chr [1:7] "White" "Mixed" "Asian or Asian British" "Black or Black Briti
##  $ educat3  : dbl+lbl [1:8843] 4, 4, 4, 2, 1, 2, 1, 4, 4, 3, 4, 3, 1, 2, 3, 1, 4, 3,...
##   ..@ label      : chr "Respondent education (5 categories)"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:5] 1 2 3 4 5
##   .. ..- attr(*, "names")= chr [1:5] "None" "O level/GCSE" "Apprenticeship or A/AS level" "Degree o
##  $ rural2   : dbl+lbl [1:8843] 1, 2, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 1, 1,...
##   ..@ label      : chr "Type of area 2004: urban/rural"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:2] 1 2
##   .. ..- attr(*, "names")= chr [1:2] "Urban" "Rural"
##  $ edeprivex: num [1:8843] 2 4 1 1 3 2 1 5 4 5 ...
##   ..- attr(*, "label")= chr "England: Index of multiple deprivation by quintile (1=20% most deprived
##   ..- attr(*, "format.spss")= chr "F8.0"
##  $ wdeprivex: num [1:8843] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "Wales: Index of multiple deprivation by quintile (1=20% most deprived wa
##   ..- attr(*, "format.spss")= chr "F8.0"
##  $ IndivWgtx: num [1:8843] 0.543 1.213 0.57 0.994 0.41 ...
##   ..- attr(*, "label")= chr "Individual-level weight (mean=1)"
##   ..- attr(*, "format.spss")= chr "F9.2"
##  $ cause2m  : dbl+lbl [1:8843] NA, NA, NA, NA, NA, NA, NA, NA,  7, NA, NA, NA, NA,  ...
##   ..@ label      : chr "One MAIN cause of crime in Britain today"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:15] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr [1:15] "A. Too lenient sentencing" "B. Poverty" "C. Lack of disciplin
##  $ walkdark : dbl+lbl [1:8843] NA, NA, NA, NA, NA, NA, NA, NA,  1, NA, NA, NA, NA,  ...
##   ..@ label      : chr "How safe do you feel walking alone after dark?"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:6] 1 2 3 4 8 9
##   .. ..- attr(*, "names")= chr [1:6] "Very safe" "Fairly safe" "A bit unsafe" "Very unsafe" ...
##  $ walkday  : dbl+lbl [1:8843] NA, NA, NA, NA, NA, NA, NA, NA,  1, NA, NA, NA, NA,  ...
##   ..@ label      : chr "How safe do you feel walking alone in this area during the day?"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:6] 1 2 3 4 8 9
##   .. ..- attr(*, "names")= chr [1:6] "Very safe" "Fairly safe" "A bit unsafe" "Very unsafe" ...
##  $ homealon : dbl+lbl [1:8843] NA, NA, NA, NA, NA, NA, NA, NA,  1, NA, NA, NA, NA,  ...
##   ..@ label      : chr "How safe do you feel when alone in home at night?"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:6] 1 2 3 4 8 9
##   .. ..- attr(*, "names")= chr [1:6] "Very safe" "Fairly safe" "A bit unsafe" "Very unsafe" ...
##  $ wburgl   : dbl+lbl [1:8843] NA,  3, NA,  2,  2,  3, NA, NA, NA, NA, NA, NA,  3, N...
##   ..@ label      : chr "How worried about having your home broken into?"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ wmugged  : dbl+lbl [1:8843] NA,  4, NA,  3,  2,  4, NA, NA, NA, NA, NA, NA,  3, N...
##   ..@ label      : chr "How worried about being mugged and robbed?"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ wcarstol : dbl+lbl [1:8843] NA, NA, NA, NA, NA, NA, NA,  3, NA, NA, NA,  4, NA, N...
##   ..@ label      : chr "How worried about having car stolen?"
##   ..@ format.spss: chr "F8.0"
##   ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ wfromcar : dbl+lbl [1:8843] NA, NA, NA, NA, NA, NA, NA,  3, NA, NA, NA,  3, NA, N...
##    ..@ label      : chr "How worried about having things stolen from your car?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##    .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ wraped   : dbl+lbl [1:8843] NA,  4, NA,  4,  2,  4, NA, NA, NA, NA, NA, NA,  3, N...
##    ..@ label      : chr "How worried about being raped?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##    .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ wattack  : dbl+lbl [1:8843] NA,  4, NA,  3,  2,  3, NA, NA, NA, NA, NA, NA,  3, N...
##    ..@ label      : chr "How worried about being physically attacked by strangers?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##    .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ wraceatt : dbl+lbl [1:8843] NA,  4, NA,  4,  3,  4, NA, NA, NA, NA, NA, NA,  4, N...
##    ..@ label      : chr "How worried about being attacked because of skin colour, ethnic origin or re
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:7] 1 2 3 4 5 8 9
##    .. ..- attr(*, "names")= chr [1:7] "Very worried" "Fairly worried" "Not very worried" "Not at all
##  $ worryx   : num [1:8843] NA -1.132 NA -0.258 1.184 ...
##   ..- attr(*, "label")= chr "Worry about being a victim of crime (high score = high level of worry)"
##   ..- attr(*, "format.spss")= chr "F9.2"
##  $ bcsvictim: dbl+lbl [1:8843] 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
##    ..@ label      : chr "Experience of any crime in the previous 12 months"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:2] 0 1
##    .. ..- attr(*, "names")= chr [1:2] "Not a victim of crime" "Victim of crime"
##  $ rubbcomm : dbl+lbl [1:8843] 3, 4, 3, 4, 3, 4, 3, 4, 4, 4, 3, 4, 3, 2, 3, 4, 3, 4,...
##    ..@ label      : chr "How common is litter or rubbish in immediate area?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:5] 1 2 3 4 5
##    .. ..- attr(*, "names")= chr [1:5] "Very common" "Fairly common" "Not very common" "Not at all co
##  $ vandcomm : dbl+lbl [1:8843] 3, 4, 4, 4, 3, 4, 3, 4, 4, 4, 3, 4, 3, 4, 4, 4, 3, 4,...
##    ..@ label      : chr "How common is vandalism or graffiti in immediate area?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:5] 1 2 3 4 5
##    .. ..- attr(*, "names")= chr [1:5] "Very common" "Fairly common" "Not very common" "Not at all co
##  $ poorhou  : dbl+lbl [1:8843] 3, 4, 3, 4, 3, 4, 3, 4, 4, 4, 3, 4, 3, 1, 4, 4, 3, 4,...
##    ..@ label      : chr "How common are homes in poor condition/run down?"
##    ..@ format.spss: chr "F8.0"
##    ..@ labels     : Named num [1:5] 1 2 3 4 5
##    .. ..- attr(*, "names")= chr [1:5] "Very common" "Fairly common" "Not very common" "Not at all co
##  $ antisocx : num [1:8843] 2.065 NA -0.236 NA NA ...
##   ..- attr(*, "label")= chr "Anti-social behaviour in their neighbourhood (high score =high levels of
##   ..- attr(*, "format.spss")= chr "F9.2"
```

Python

```python
dataset_python = pd.read_spss("C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_modul

dataset_python.shape
```

```
## (8843, 32)
```

```
dataset_python.info()
```

```
## <class ’pandas.core.frame.DataFrame’>
## RangeIndex: 8843 entries, 0 to 8842
## Data columns (total 32 columns):
##  #   Column     Non-Null Count  Dtype
## ---  ------     --------------  -----
##  0   rowlabel   8843 non-null   float64
##  1   split      8843 non-null   category
##  2   sex        8843 non-null   category
##  3   yrsarea    8842 non-null   category
##  4   resyrago   1509 non-null   category
##  5   work2      8841 non-null   category
##  6   tenure1    8820 non-null   category
##  7   livharm1   8830 non-null   category
##  8   agegrp7    8843 non-null   category
##  9   ethgrp2a   8833 non-null   category
##  10  educat3    8822 non-null   category
##  11  rural2     8843 non-null   category
##  12  edeprivex  8140 non-null   float64
##  13  wdeprivex  703 non-null    float64
##  14  IndivWgtx  8843 non-null   float64
##  15  cause2m    2064 non-null   category
##  16  walkdark   2057 non-null   category
##  17  walkday    2071 non-null   category
##  18  homealon   2072 non-null   category
##  19  wburgl     2193 non-null   category
##  20  wmugged    2185 non-null   category
##  21  wcarstol   1763 non-null   category
##  22  wfromcar   1732 non-null   category
##  23  wraped     2183 non-null   category
##  24  wattack    2185 non-null   category
##  25  wraceatt   2184 non-null   category
##  26  worryx     2047 non-null   float64
##  27  bcsvictim  8843 non-null   category
##  28  rubbcomm   8843 non-null   category
##  29  vandcomm   8843 non-null   category
##  30  poorhou    8843 non-null   category
##  31  antisocx   2149 non-null   float64
## dtypes: category(26), float64(6)
## memory usage: 644.8 KB
```

Summary stats for antisocial behaviour (antisocx)

R

```
summary(dataset$antisocx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##  -1.215  -0.788  -0.185  -0.007   0.528   4.015    6694
```

Mean -0.01, median -0.18; data seems to be right skewed, and to take both positive and negative values; there are also a high proportion of missing values (76%)

Python

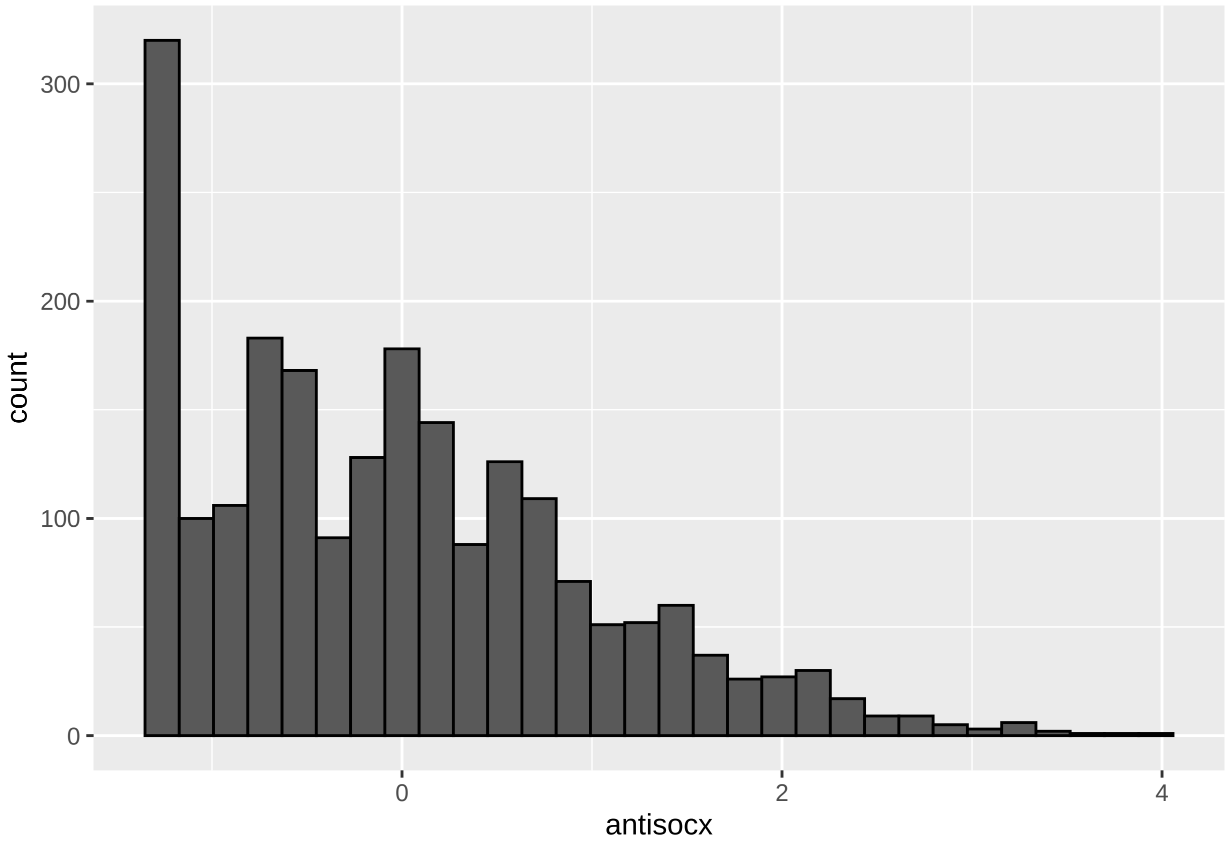
```
# convert data into a dataframe (not required here)
# df = pd.DataFrame(dataset_python)

dataset_python["antisocx"].describe()
```

```
## count    2149.000000
## mean       -0.007498
## std         0.991067
## min        -1.215267
## 25%        -0.788219
## 50%        -0.184597
## 75%         0.528008
## max         4.014557
## Name: antisocx, dtype: float64
```

Histogram

R

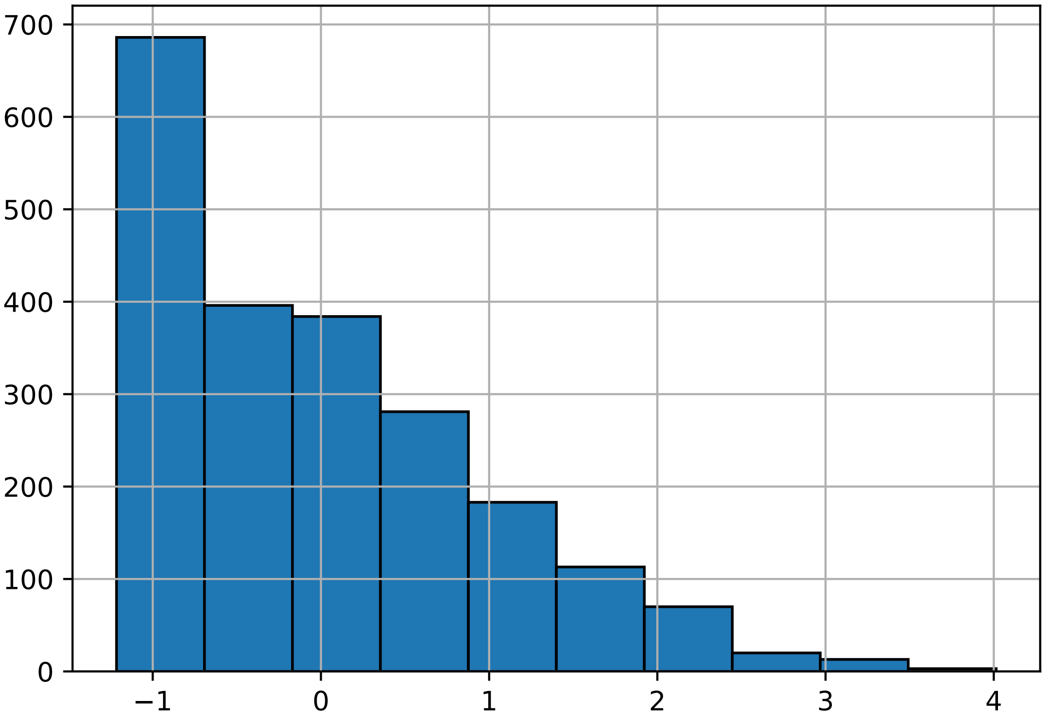
```
dataset %>%
  ggplot() +
geom_histogram(aes(antisocx),
               color = "black")
```



The histogram confirms the previous observations, and shows that the most common values are those in the bucket with the lowest values

Python

```
dataset_python["antisocx"].hist(edgecolor="black")
plt.show()
```



Unit 2

Experience of crime in the previous year

R

```
summary(dataset$bcsvictim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1564  0.0000  1.0000
```

```
str(dataset$bcsvictim)
```

```
##  dbl+lbl [1:8843] 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
##  @ label      : chr "Experience of any crime in the previous 12 months"
##  @ format.spss: chr "F8.0"
##  @ labels     : Named num [1:2] 0 1
##   ..- attr(*, "names")= chr [1:2] "Not a victim of crime" "Victim of crime"
```

```
# this variable is coded as numeric but represents a binary feature,
# therefore I will reassign it as a factor to allow counting using table()
```

```r
dataset$bcsvictim <- as.factor(dataset$bcsvictim)

str(dataset$bcsvictim)
```

```
##  Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
```

```r
table(dataset$bcsvictim)
```

```
## 
##    0    1 
## 7460 1383
```

Out of 8843 respondents, 1383 (15.6%) experienced crime in the previous year (note that the 15.6% value included above was inserted with in-line R code to perform that calculation within the document)

Python

```python
dataset_python['bcsvictim'].describe()
```

```
## count                      8843
## unique                        2
## top       Not a victim of crime
## freq                       7460
## Name: bcsvictim, dtype: object
```

```python
dataset_python['bcsvictim'].dtypes
```

```
## CategoricalDtype(categories=[’Not a victim of crime’, ’Victim of crime’], ordered=False, categories_
```

```python
# already coded as Categorical, so no need to recode (in opposition to R)

# code to convert to factor (categorical) if needed
# dataset_python['bcsvictim']=pd.Categorical(dataset_python['bcsvictim'])

# dataset_python.bcsvictim

dataset_python.bcsvictim.value_counts()
```

```
## bcsvictim
## Not a victim of crime    7460
## Victim of crime          1383
## Name: count, dtype: int64
```

Unit 3

“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’.

R

```r
head(dataset)
```

```
## # A tibble: 6 x 32
##    rowlabel split      sex     yrsarea resyrago work2   tenure1 livharm1 agegrp7
##       <dbl> <dbl+lbl>  <dbl+l> <dbl+l> <dbl+lb> <dbl+l> <dbl+l> <dbl+lb> <dbl+l>
## 1 137068050 1 [A (Exp~ 2 [Fem~ 7 [20 ~ NA       1 [Yes] 2 [Buy~ 3 [Sing~ 4 [45--~
## 2 147461190 3 [C (Cri~ 2 [Fem~ 6 [10 ~ NA       2 [No]  1 [Own~ 1 [Marr~ 5 [55--~
## 3 137116250 1 [A (Exp~ 2 [Fem~ 7 [20 ~  2 [No]  2 [No]  4 [Ren~ 6 [Wido~ 5 [55--~
## 4 147354190 3 [C (Cri~ 2 [Fem~ 7 [20 ~ NA       1 [Yes] 2 [Buy~ 1 [Marr~ 5 [55--~
## 5 137061230 3 [C (Cri~ 2 [Fem~ 7 [20 ~ NA       2 [No]  4 [Ren~ 6 [Wido~ 6 [65--~
## 6 136898230 3 [C (Cri~ 2 [Fem~ 7 [20 ~ NA       2 [No]  1 [Own~ 1 [Marr~ 6 [65--~
## # 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>, wraped <dbl+lbl>, wattack <dbl+lbl>,
## #   wraceatt <dbl+lbl>, worryx <dbl>, bcsvictim <fct>, rubbcomm <dbl+lbl>,
## #   vandcomm <dbl+lbl>, poorhou <dbl+lbl>, antisocx <dbl>
```

```r
str(dataset$agegrp7)
```

```
##  dbl+lbl [1:8843] 4, 5, 5, 5, 6, 6, 4, 5, 5, 7, 2, 7, 7, 4, 4, 7, 4, 6, 5, ...
##  @ label      : chr "Age group (7 bands)"
##  @ format.spss: chr "F8.0"
##  @ labels     : Named num [1:7] 1 2 3 4 5 6 7
##   ..- attr(*, "names")= chr [1:7] "16-24" "25-34" "35-44" "45-54" ...
```

```r
output_directory<-"C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_practic

dataset %>%
  filter(agegrp7 == 7,
         bcsvictim == 1)%>%
  write_csv(file = paste0(output_directory, "Datasets/R/crime_75victim.csv"))
```

Python

```python
# general exploration of the data frame
dataset_python.head()
```

```
##       rowlabel                              split  ...            poorhou  antisocx
## 0  137068050.0      A (Experiences of the police)  ...    Not very common  2.065439
## 1  147461190.0  C (Crime prevention and security)  ...  Not at all common       NaN
## 2  137116250.0      A (Experiences of the police)  ...    Not very common -0.235942
## 3  147354190.0  C (Crime prevention and security)  ...  Not at all common       NaN
## 4  137061230.0  C (Crime prevention and security)  ...    Not very common       NaN
##
## [5 rows x 32 columns]
```

```
dataset_python.dtypes
```

```
## rowlabel      float64
## split        category
## sex          category
## yrsarea      category
## resyrago     category
## work2        category
## tenure1      category
## livharm1     category
## agegrp7      category
## ethgrp2a     category
## educat3      category
## rural2       category
## edeprivex     float64
## wdeprivex     float64
## IndivWgtx     float64
## cause2m      category
## walkdark     category
## walkday      category
## homealon     category
## wburgl       category
## wmugged      category
## wcarstol     category
## wfromcar     category
## wraped       category
## wattack      category
## wraceatt     category
## worryx        float64
## bcsvictim    category
## rubbcomm     category
## vandcomm     category
## poorhou      category
## antisocx      float64
## dtype: object
```

```
dataset_python.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 8843 entries, 0 to 8842
## Data columns (total 32 columns):
##  #   Column     Non-Null Count  Dtype
## ---  ------     --------------  -----
##  0   rowlabel   8843 non-null   float64
##  1   split      8843 non-null   category
##  2   sex        8843 non-null   category
##  3   yrsarea    8842 non-null   category
##  4   resyrago   1509 non-null   category
##  5   work2      8841 non-null   category
##  6   tenure1    8820 non-null   category
##  7   livharm1   8830 non-null   category
##  8   agegrp7    8843 non-null   category
##  9   ethgrp2a   8833 non-null   category
##  10  educat3    8822 non-null   category
##  11  rural2     8843 non-null   category
##  12  edeprivex  8140 non-null   float64
##  13  wdeprivex  703 non-null    float64
##  14  IndivWgtx  8843 non-null   float64
##  15  cause2m    2064 non-null   category
##  16  walkdark   2057 non-null   category
##  17  walkday    2071 non-null   category
##  18  homealon   2072 non-null   category
##  19  wburgl     2193 non-null   category
##  20  wmugged    2185 non-null   category
##  21  wcarstol   1763 non-null   category
##  22  wfromcar   1732 non-null   category
##  23  wraped     2183 non-null   category
##  24  wattack    2185 non-null   category
##  25  wraceatt   2184 non-null   category
##  26  worryx     2047 non-null   float64
##  27  bcsvictim  8843 non-null   category
##  28  rubbcomm   8843 non-null   category
##  29  vandcomm   8843 non-null   category
##  30  poorhou    8843 non-null   category
##  31  antisocx   2149 non-null   float64
## dtypes: category(26), float64(6)
## memory usage: 644.8 KB
```

```
# exploration of the two columns to be used
dataset_python.agegrp7.head()
```

```
## 0    45-54
## 1    55-64
## 2    55-64
## 3    55-64
## 4    65-74
## Name: agegrp7, dtype: category
## Categories (7, object): ['16-24', '25-34', '35-44', '45-54', '55-64', '65-74', '75+']
```

```
dataset_python.bcsvictim.head()
```

```
## 0    Not a victim of crime
## 1    Not a victim of crime
## 2    Not a victim of crime
## 3          Victim of crime
## 4    Not a victim of crime
## Name: bcsvictim, dtype: category
## Categories (2, object): ['Not a victim of crime', 'Victim of crime']
```

```
# filter dataset and inspect results

## method 1: filter based on dataframe column names
dataset_python_crime75victim = dataset_python[(dataset_python.agegrp7 == "75+") & (dataset_python.bcsvi

dataset_python_crime75victim.shape
```

```
## (67, 32)
```

```
## method 2: 'query' method
dataset_python_crime75victim = dataset_python.query("agegrp7 == '75+' & bcsvictim == 'Victim of crime'"

dataset_python_crime75victim.shape
```

```
## (67, 32)
```

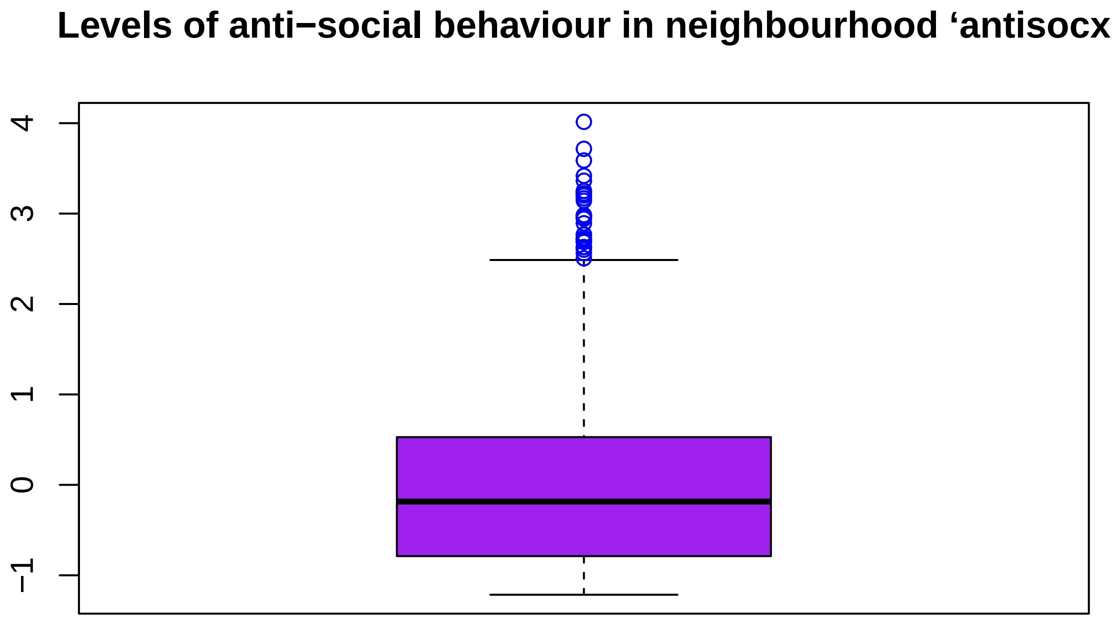
```
# save dataset

output_directory="C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_practic

dataset_python_crime75victim.to_csv(output_directory + "Datasets/Python/crime_75victim.csv")
```

Unit 5 - data activities

1. Create a boxplot for the variable 'antisocx'

R (base)

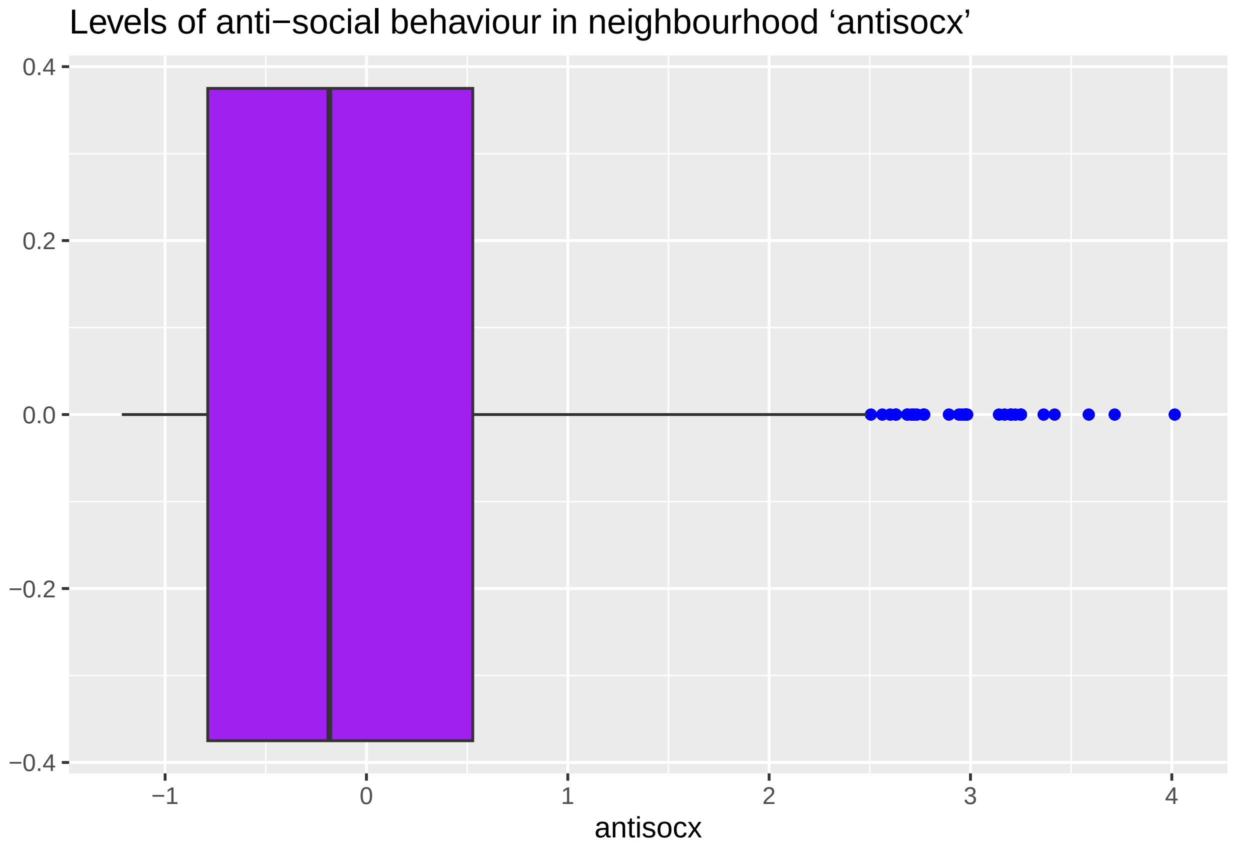
```
boxplot(dataset$antisocx,
        main = "Levels of anti-social behaviour in neighbourhood ‘antisocx",
        col = "purple",
        outcol = "blue")
```



R (ggplot)

```
ggplot(dataset, aes(antisocx))+
  geom_boxplot(fill="purple",
               outlier.colour = "blue")+
  labs(title="Levels of anti-social behaviour in neighbourhood ‘antisocx’")
```

```
## Warning: Removed 6694 rows containing non-finite values (‘stat_boxplot()‘).
```

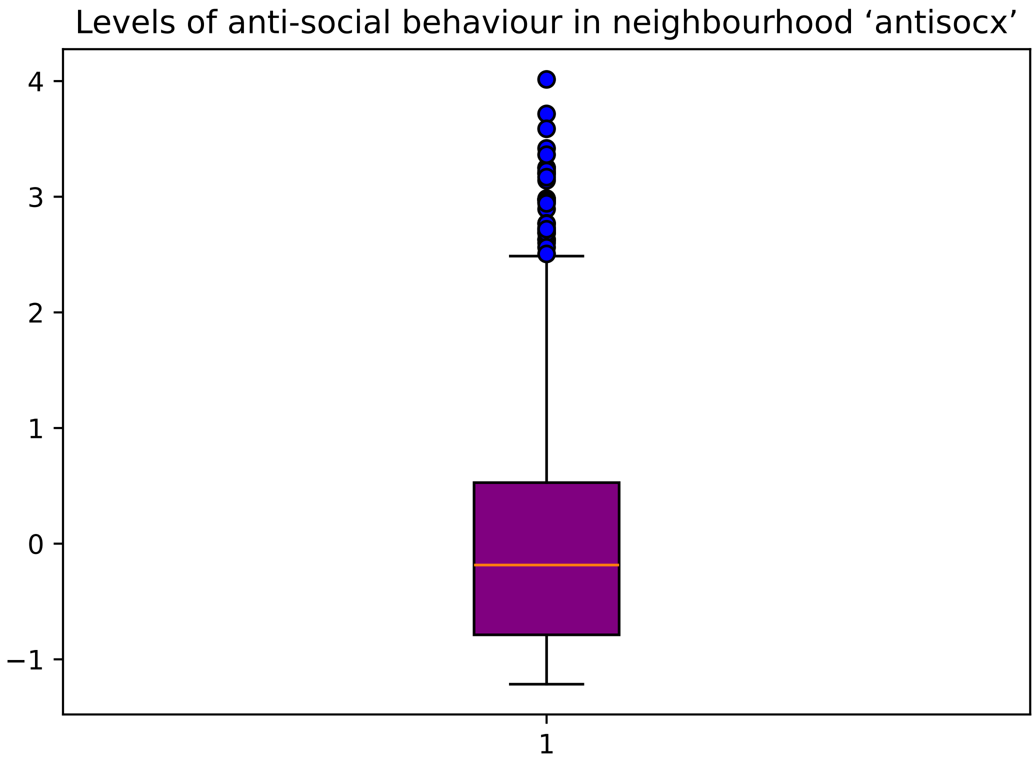


Python (pyplot)

```
plt.boxplot(dataset_python.antisocx.dropna(),
  patch_artist=True,
  boxprops=dict(facecolor="purple"),
  flierprops=dict(markerfacecolor='blue')) # notice I had to drop NAs, otherwise this wouldn't plot
```

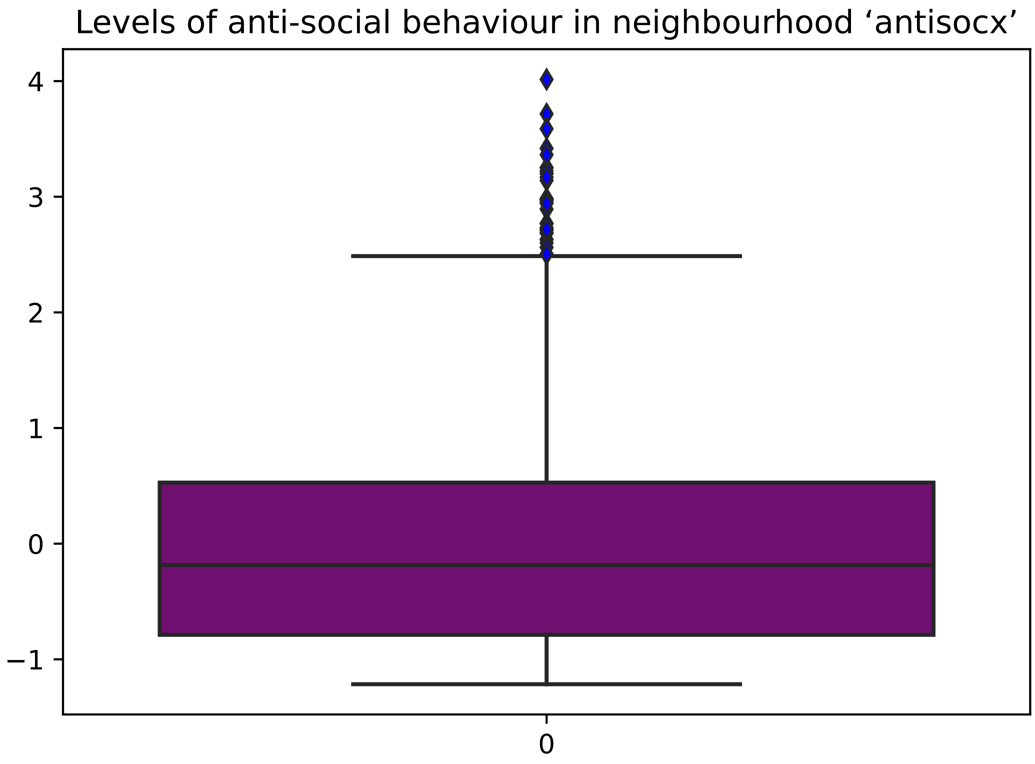
```
## {’whiskers’: [<matplotlib.lines.Line2D object at 0x000001B1CF023710>, <matplotlib.lines.Line2D objec
```

```
plt.title("Levels of anti-social behaviour in neighbourhood ‘antisocx’")

plt.show()
```



Python (Seaborn)

```
import seaborn as sb

sb.boxplot(dataset_python.antisocx,
color="purple",
flierprops={"markerfacecolor": "blue"})

plt.title("Levels of anti-social behaviour in neighbourhood ‘antisocx’")

plt.show()
```



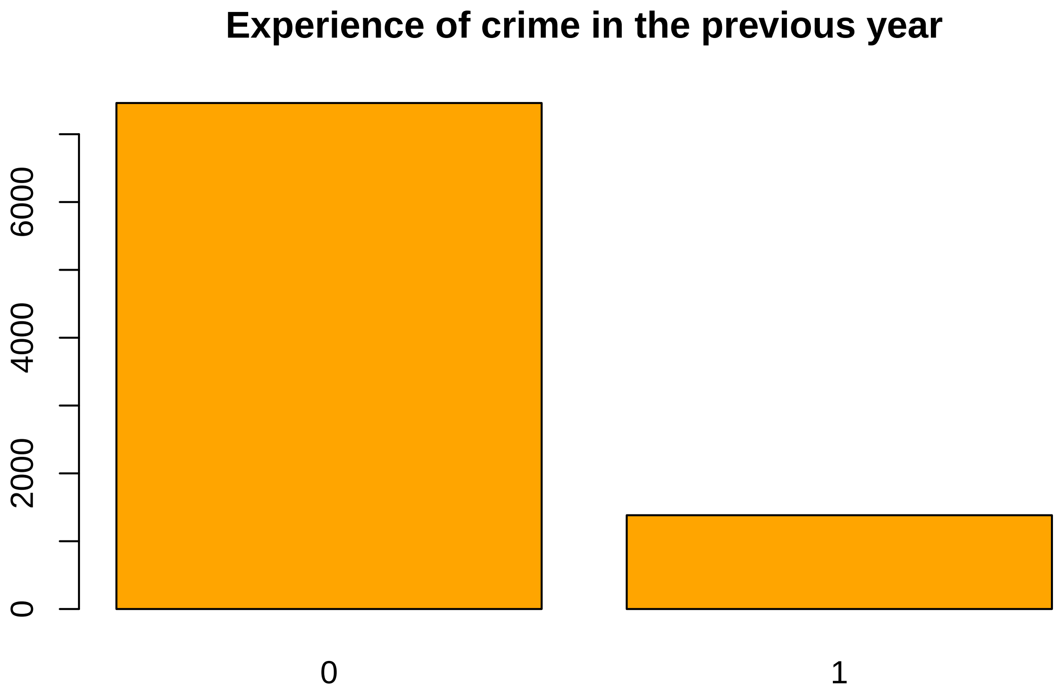
Python (plotly)

```
import plotly.express as px

px.box(dataset_python,y="antisocx")
```

2. 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).

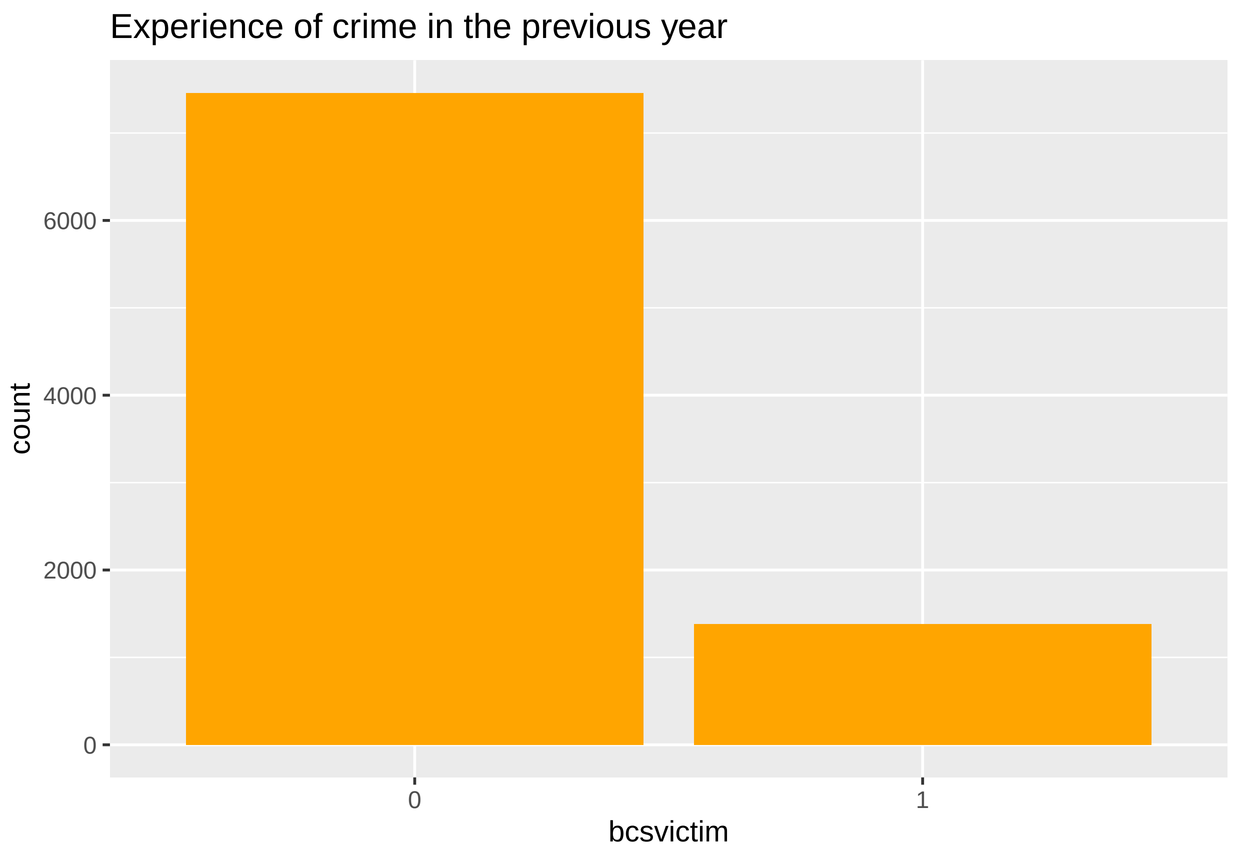
R (base)

```
barplot(table(as.factor(dataset$bcsvictim)),
        col="orange",
        main="Experience of crime in the previous year")
```



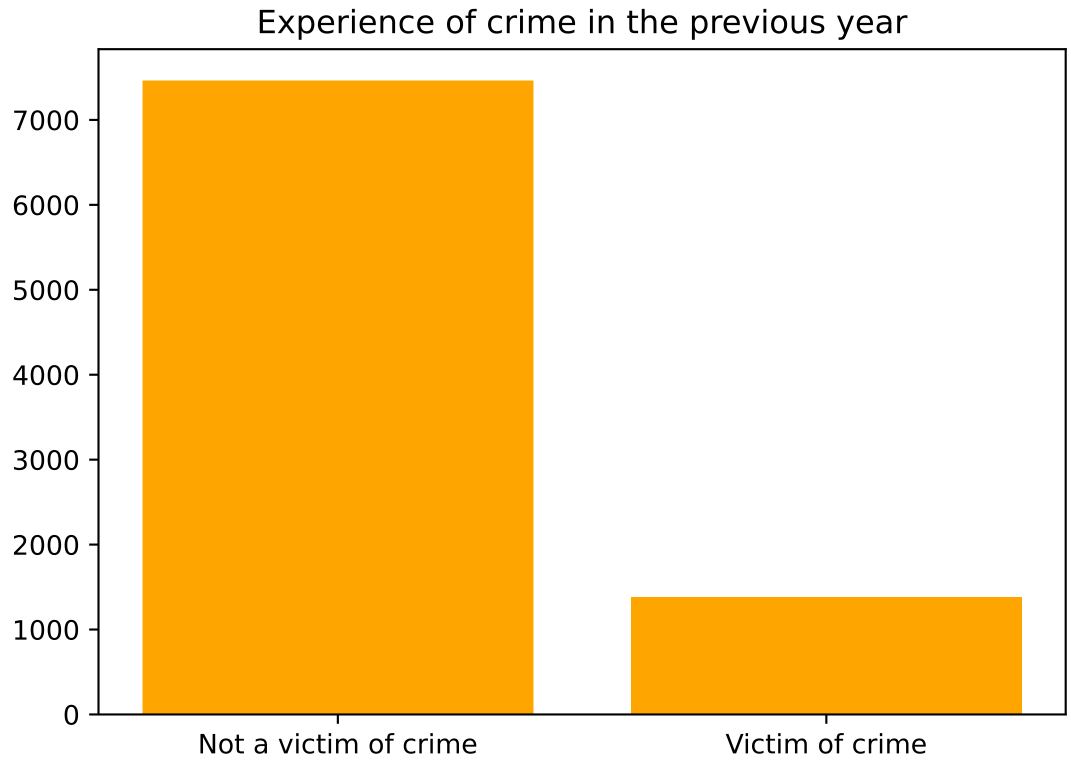
R (ggplot)

```
ggplot(dataset, aes(as.factor(bcsvictim)))+
  geom_bar(fill="orange")+
  labs(title="Experience of crime in the previous year",
       x="bcsvictim")
```



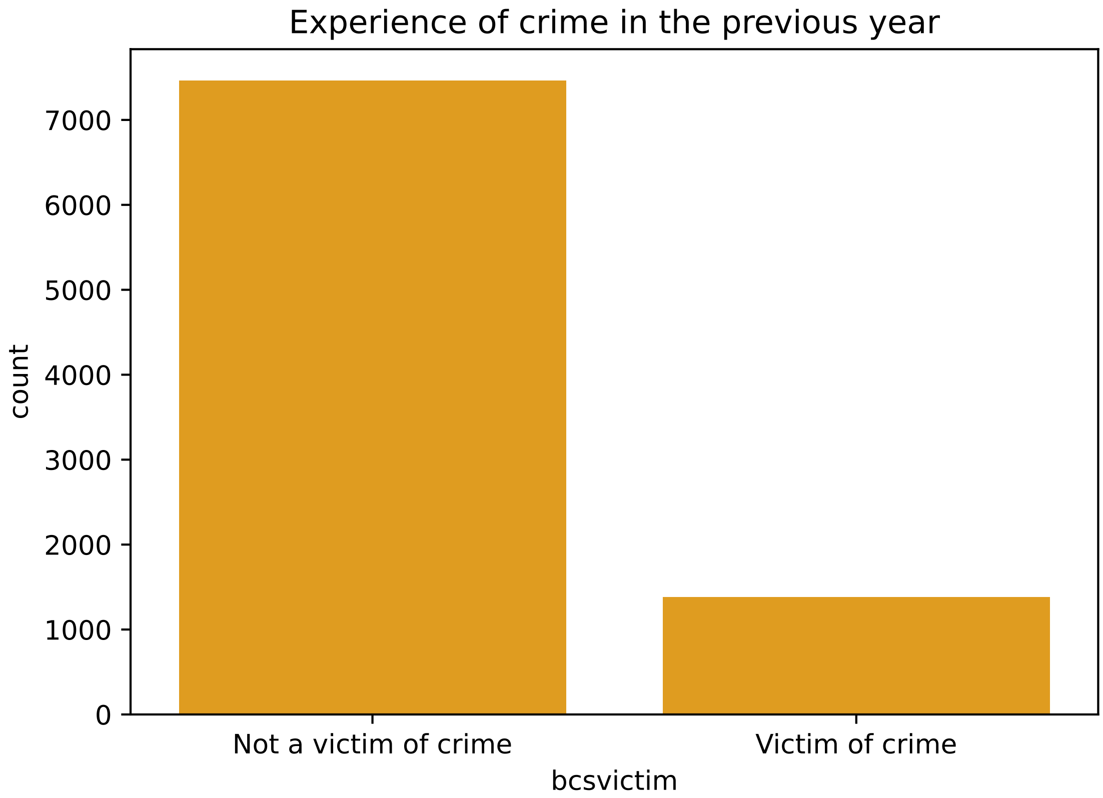
Python (pyplot)

```
plt.bar(dataset_python.bcsvictim.value_counts().index,
dataset_python.bcsvictim.value_counts().values,
color="orange"
)

plt.title("Experience of crime in the previous year")


plt.show()
```



Python (Seaborn)

```
sb.countplot(dataset_python,
  x="bcsvictim",
  color="orange")

plt.title("Experience of crime in the previous year")

plt.show()
```



Python (plotly)

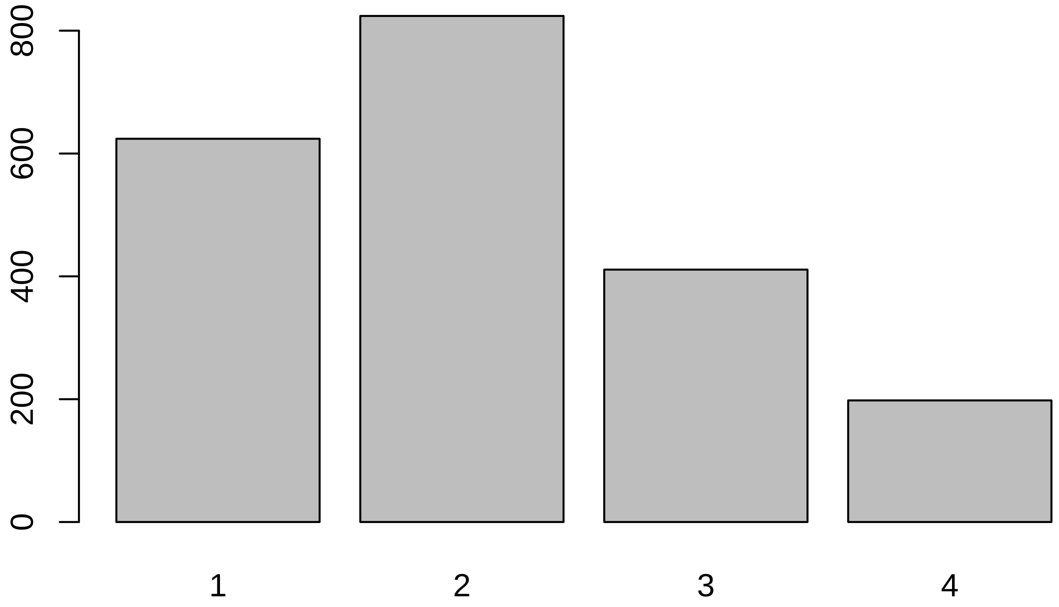
```
# px.bar(data=dataset_python.groupby('bcsvictim').count().reset_index(),x="bcsvictim",y="rowlabel",colo
```

Unit 5 - notes activities

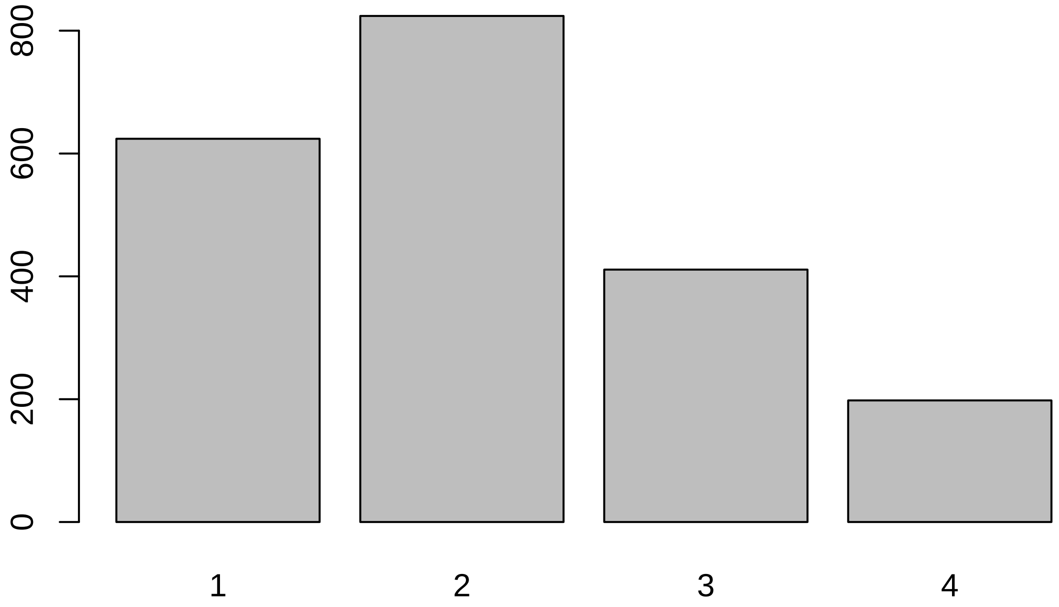
Barcharts

R

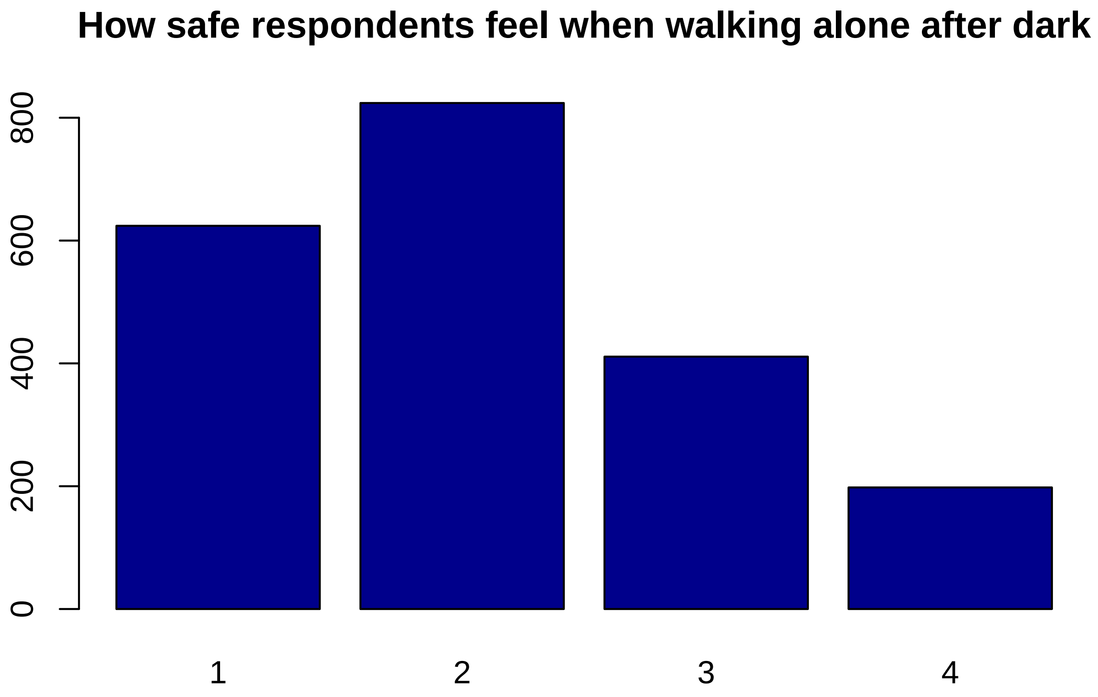
```
barplot(
  table(dataset$walkdark)
)
```



```
# now after removing missing values

barplot(
  table(dataset$walkdark, useNA = "no")
)
```

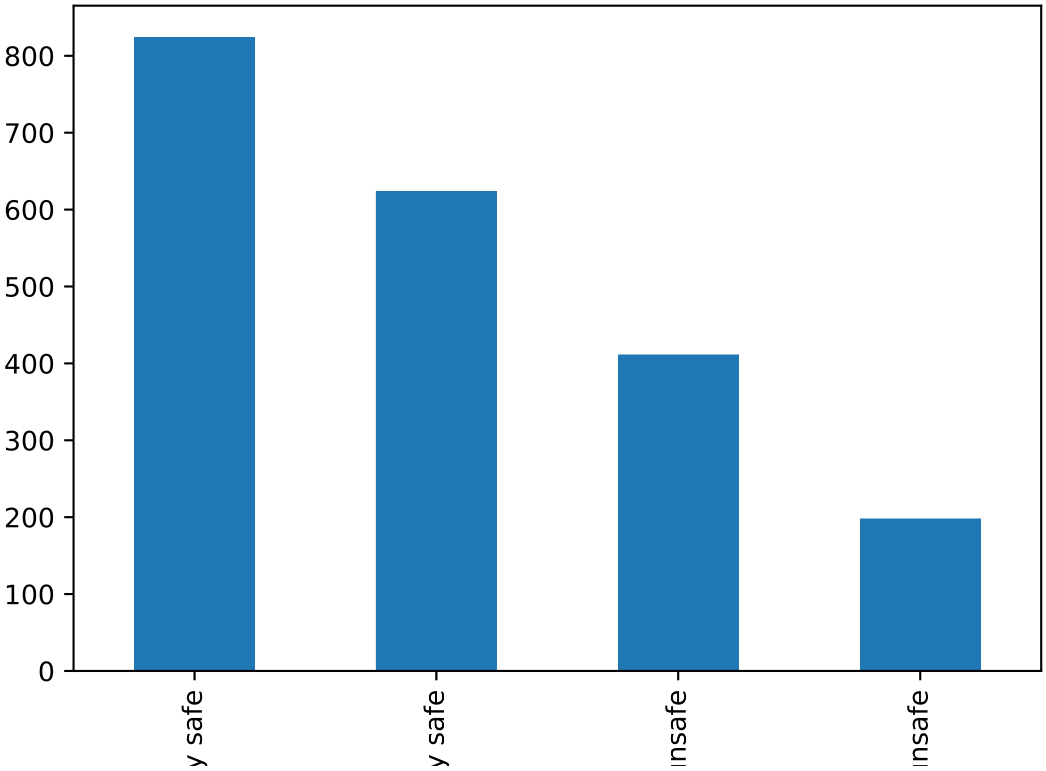


```r
# some customisation

barplot(
  table(dataset$walkdark, useNA = "no"),
  main = "How safe respondents feel when walking alone after dark",
  col = "darkblue"
)
```

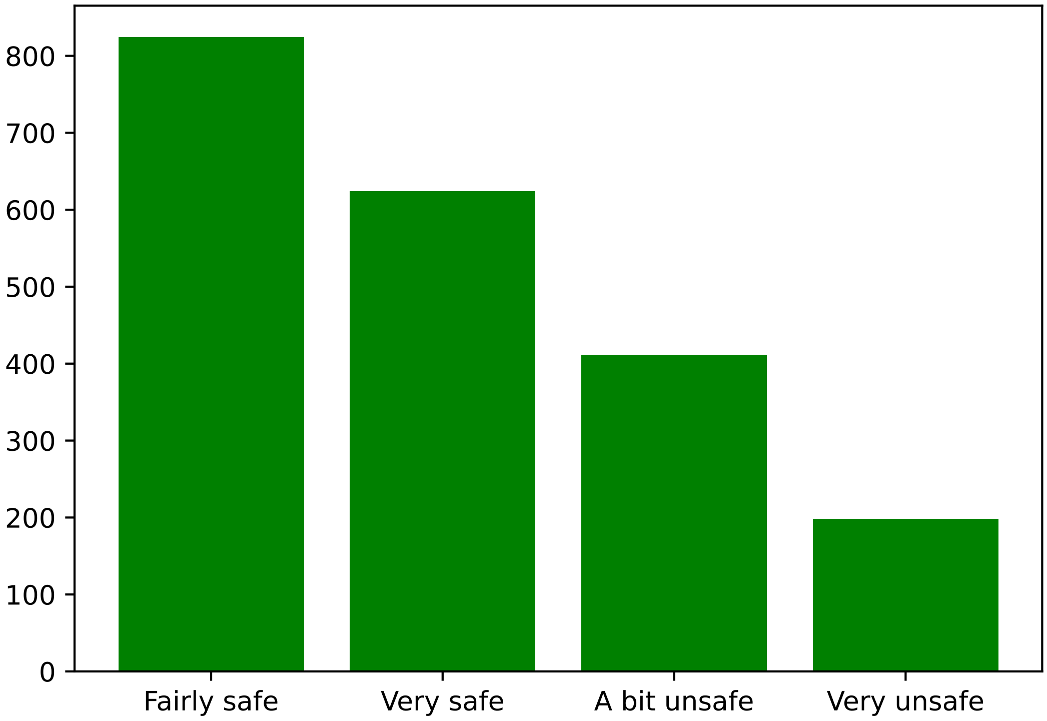


Python

```
dataset_python.walkdark.value_counts().plot(kind="bar")

plt.show()
```



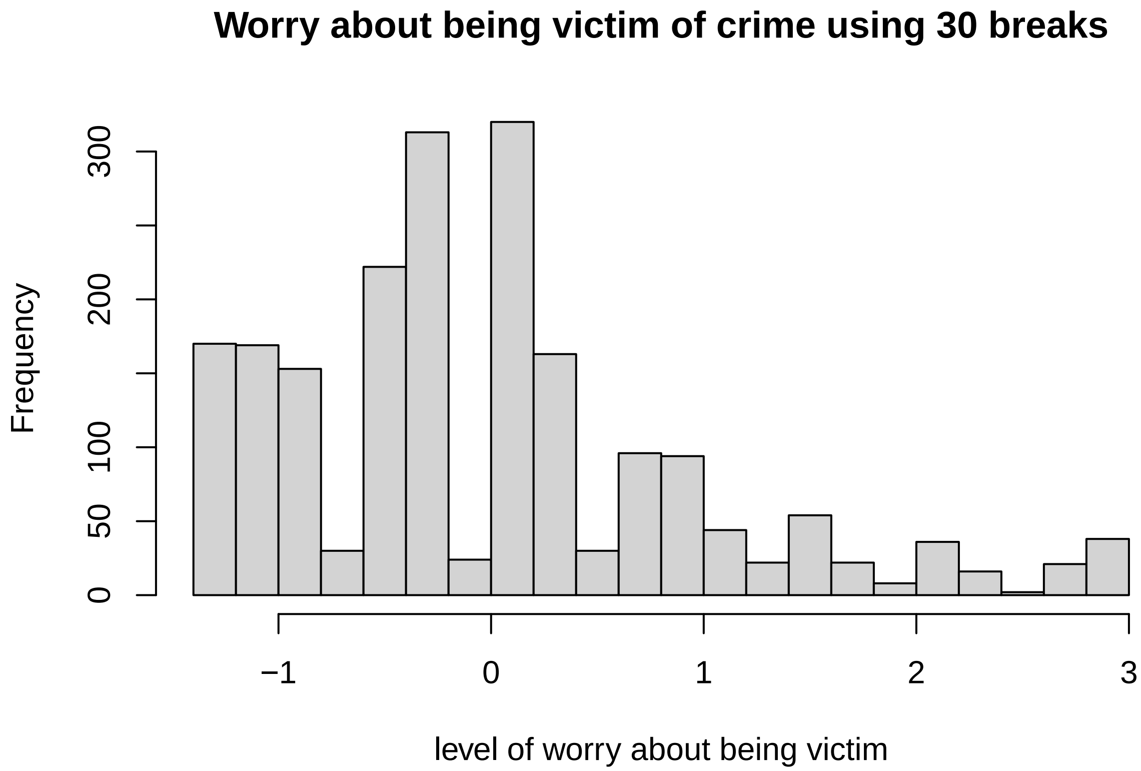
```
# alternative method

plt.bar(list(dataset_python.walkdark.value_counts().index),dataset_python.walkdark.value_counts().values

plt.show()
```



## Histograms

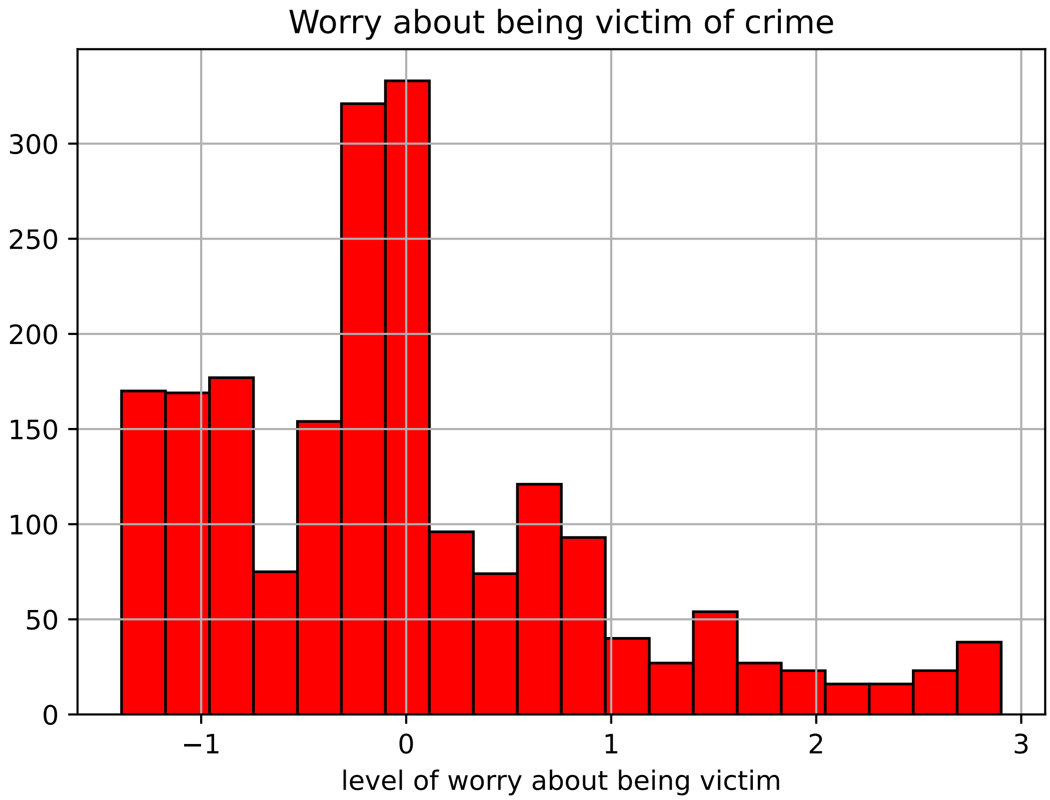
R

```r
hist(dataset$worryx,
     breaks=20,
     main="Worry about being victim of crime using 30 breaks",
     xlab = "level of worry about being victim")
```



Python

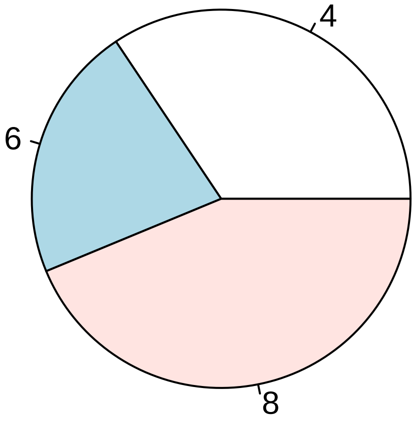
```
dataset_python.worryx.hist(edgecolor="black", color="red", bins=20)
plt.xlabel("level of worry about being victim")
plt.title("Worry about being victim of crime")

plt.show()
```



## Pie charts

R

```
data("mtcars")

propcyl<-table(mtcars$cyl)

pie(propcyl)
```



```
write_csv(mtcars, "C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_practic
```

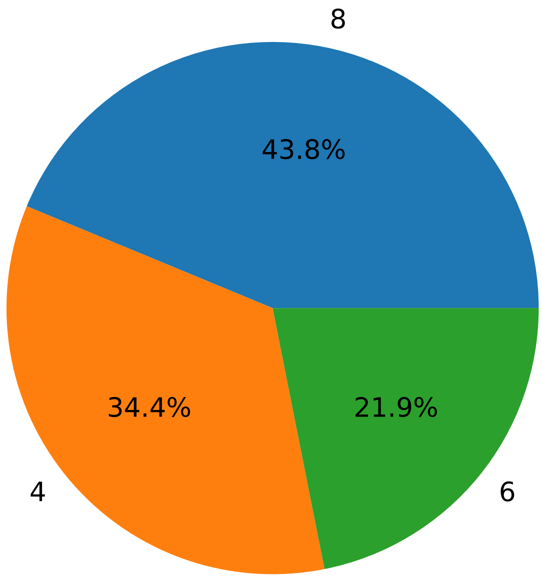
Python

```
mtcars=pd.read_csv("C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_practi

mtcars.cyl.value_counts()
```

```
## cyl
## 8    14
## 4    11
## 6     7
## Name: count, dtype: int64
```

```
sizes=mtcars.cyl.value_counts().values

labels=mtcars.cyl.value_counts().index

fig, ax = plt.subplots()
ax.pie(sizes, labels=labels, autopct='%1.1f%%')
```

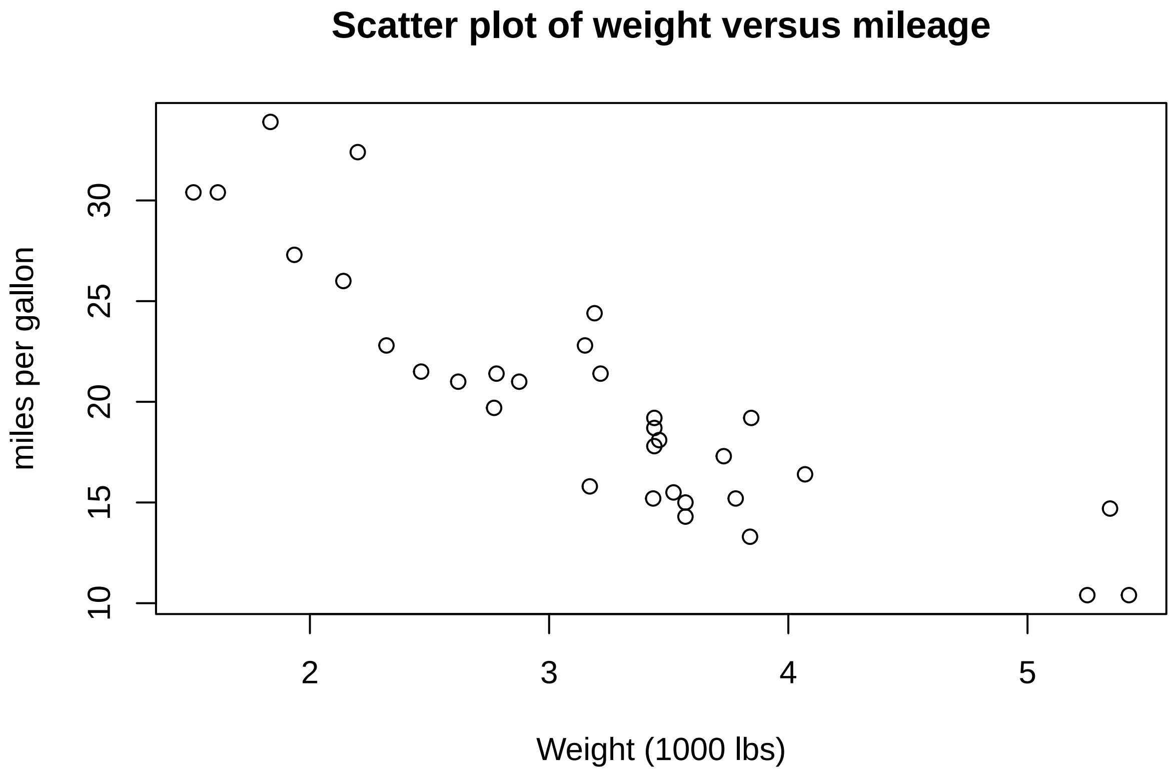
```
## ([<matplotlib.patches.Wedge object at 0x000001B1D6510990>, <matplotlib.patches.Wedge object at 0x0000
plt.show()
```



Scatterplots

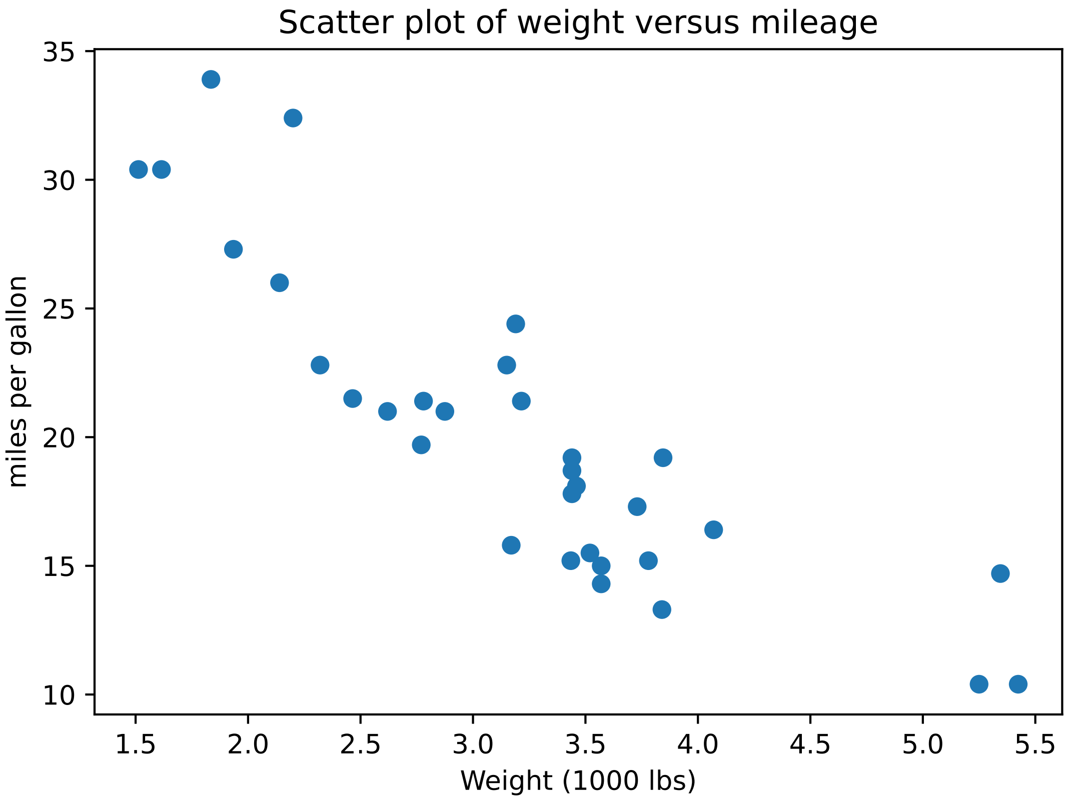
R

```
plot(mtcars$wt,
     mtcars$mpg,
     main = "Scatter plot of weight versus mileage",
     xlab = "Weight (1000 lbs)",
     ylab = "miles per gallon")
```



Python

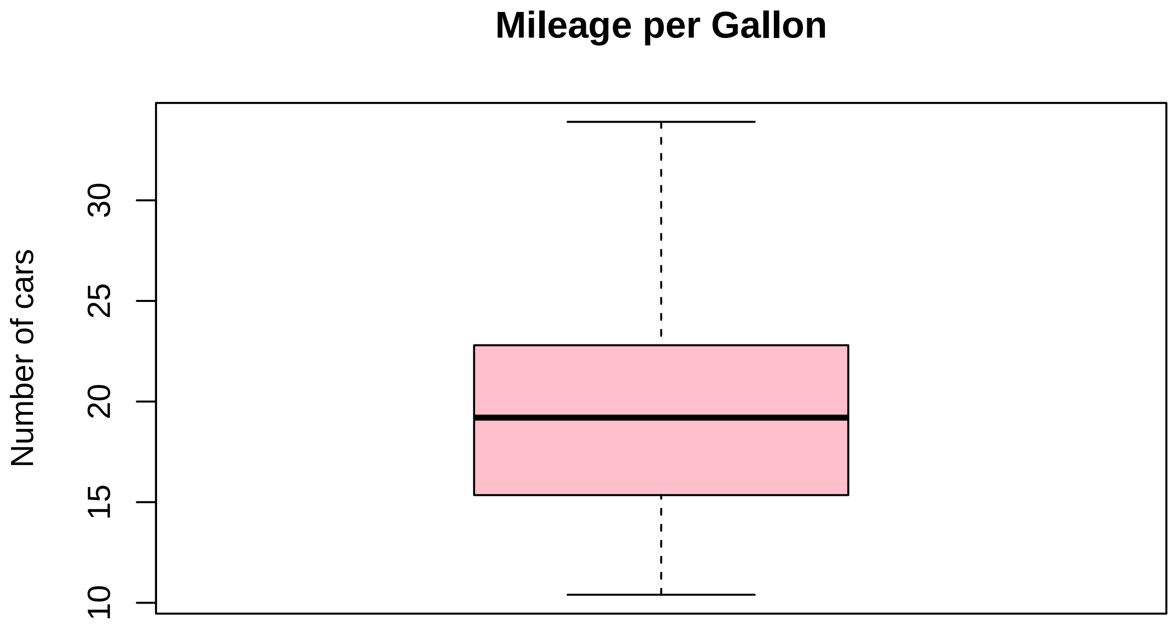
```
plt.scatter(mtcars.wt, mtcars.mpg)
plt.title("Scatter plot of weight versus mileage")
plt.xlabel("Weight (1000 lbs)")
plt.ylabel("miles per gallon")
plt.show()
```



Boxplots

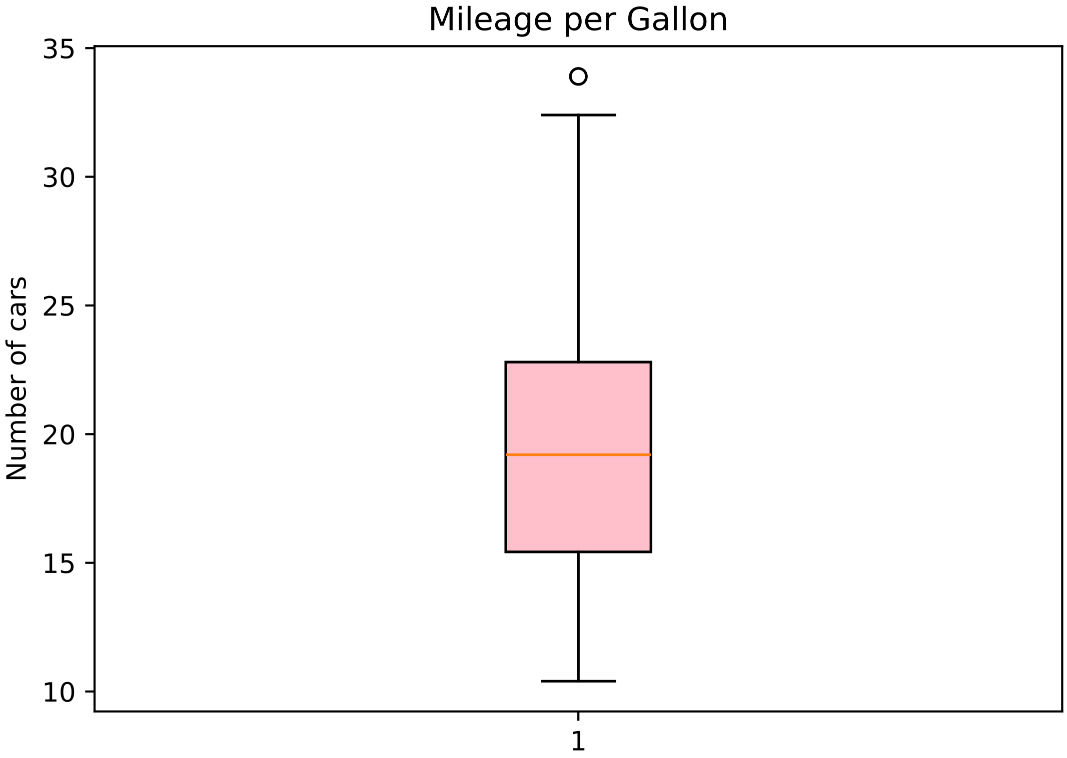
R

```
boxplot(mtcars$mpg,
        main= "Mileage per Gallon",
        ylab="Number of cars",
        col = "pink")
```



Python

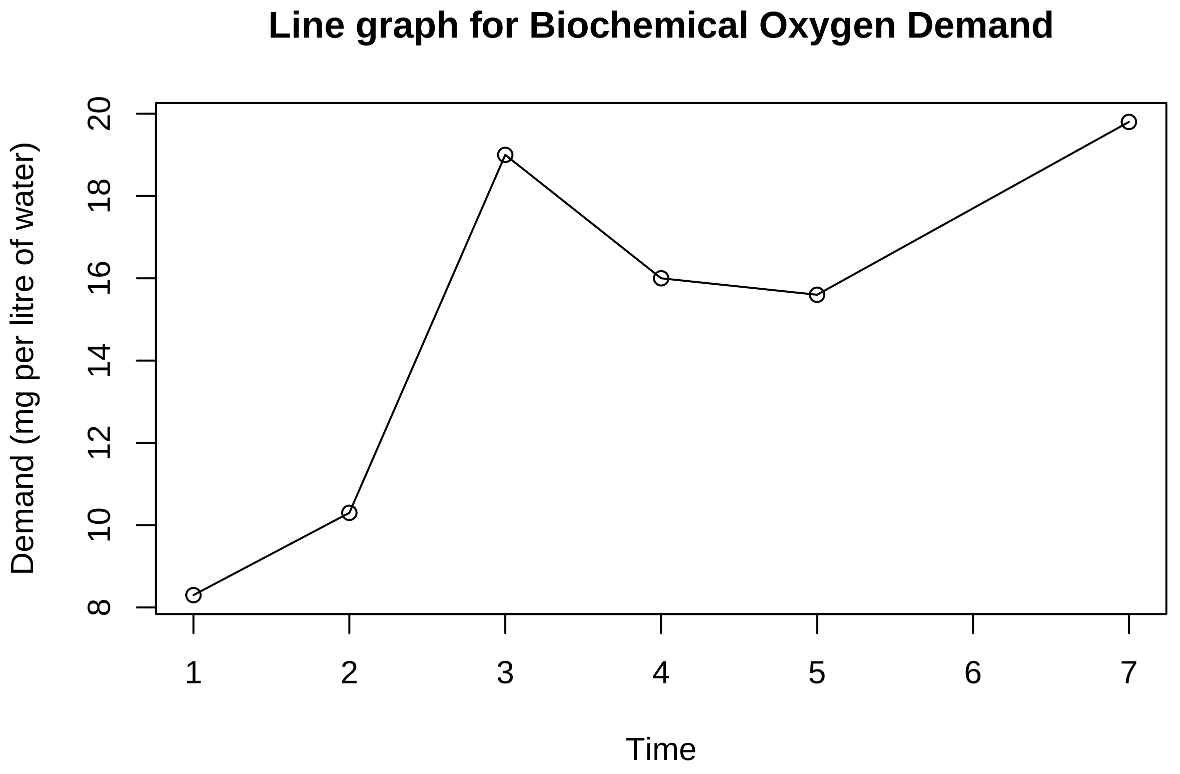
```
fig = plt.boxplot(mtcars.mpg, patch_artist=True, boxprops=dict(facecolor="pink"))
plt.title("Mileage per Gallon")
plt.ylabel("Number of cars")
plt.show()
```



Line charts

R

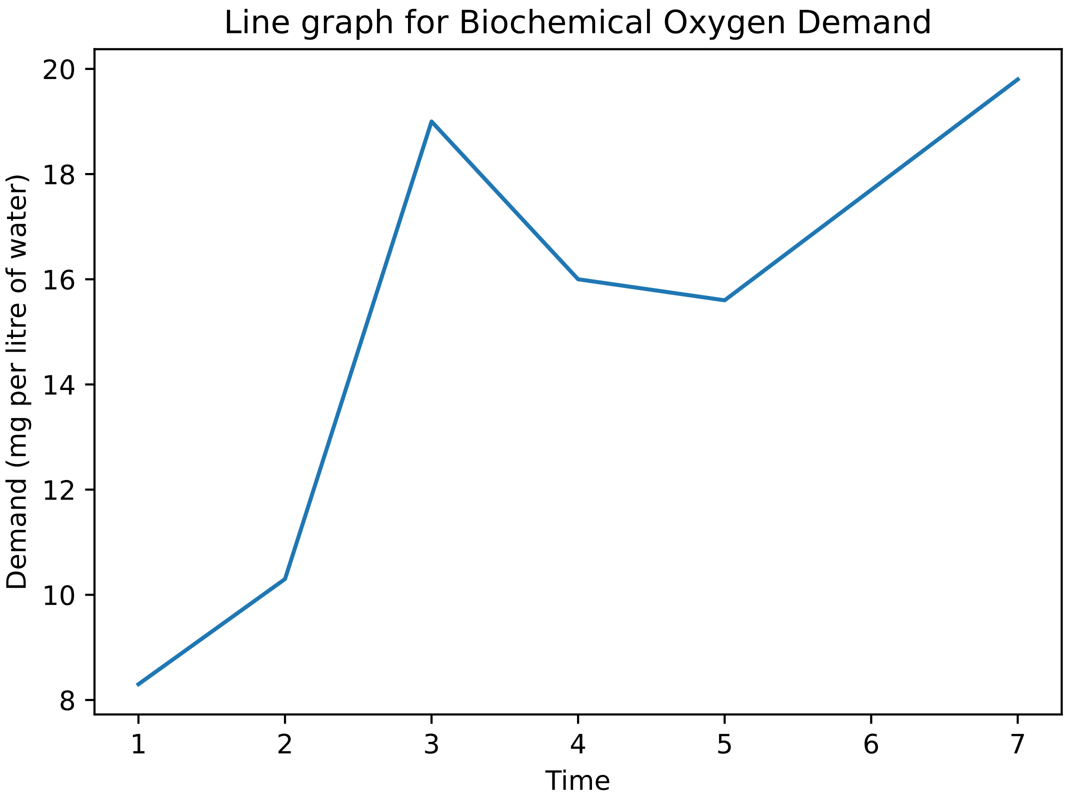
```
data(BOD)

plot(BOD$Time,
     BOD$demand,
     type = "o",
     main = "Line graph for Biochemical Oxygen Demand",
     xlab = "Time",
     ylab ="Demand (mg per litre of water)")
```



```
write_csv(BOD, "C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_practice/B
```

Python

```
BOD=pd.read_csv("C:/Users/guilhermep/OneDrive - Nexus365/PgDip/Coding/Module 2/pgdip_module2_r_practice/

plt.plot(BOD.Time,
BOD.demand)

plt.title("Line graph for Biochemical Oxygen Demand")
plt.xlabel("Time")
plt.ylabel("Demand (mg per litre of water)")
plt.show()
```


Line graph for Biochemical Oxygen Demand
Demand (mg per litre of water)
20
18
16
14
12
10
8
1
2
3
4
5
6
7
Time