Solutions to exercises

Listed below are the solutions to the exercises.

All solutions are found using RStudio, though **you should only do the exercises in RStudio if indicated in the list of exercises.** This may result in slight differences in numerical answers, which is due to rounding errors.

The solutions may often be computed in different ways and when two solutions are given it does not necessarily mean that more solutions does not exist. However, when two solutions are given we encourage you to think about why these two solutions are equivalent.

```
library(mosaic)
```

9.1:

a) - x: high scool - y: college

b) - x: education - y: children

c) - x: education - y: income

d) - x: income - y: housevalue

9.11:

a.i)

(20;90)

a.ii)

(37.5;40)

b)

```
pred <- -0.13 + 2.62 * 34.3
pred
```

```
## [1] 89.736
```

```
res <- 45.1 - pred
res
```

```
## [1] -44.636
```

The actual cell-phone usage is much lower than expected since the residual is negative.

c)

The correlation is positive since when the GDP increase the cellular usage seem to increase.

9.13:

The prediction equation is

$$\hat{y} = a + bx \qquad \Leftrightarrow \qquad a = \hat{y} - bx$$

and by the measure of correlation

$$r = b\left(\frac{s_x}{s_y}\right) \qquad \Leftrightarrow \qquad b = r\left(\frac{s_y}{s_x}\right)$$

Hence

```
b <- 0.6 * 120 / 80
b
```

```
## [1] 0.9
```

```
a <- 500 - b * 480
a
```

```
## [1] 68
```

and

$$\hat{y} = 68 + 0.9x$$

9.26:

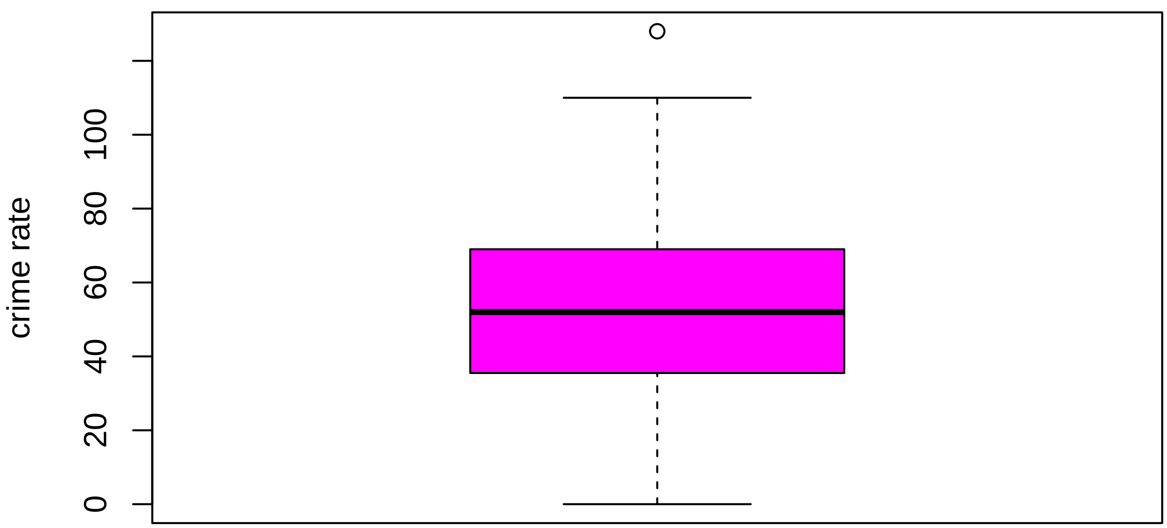
Additional sub-exercises:

Import data:

```
florida <- read.table("https://asta.math.aau.dk/datasets?file=fl-crime-extended.txt", header = TRUE)
```

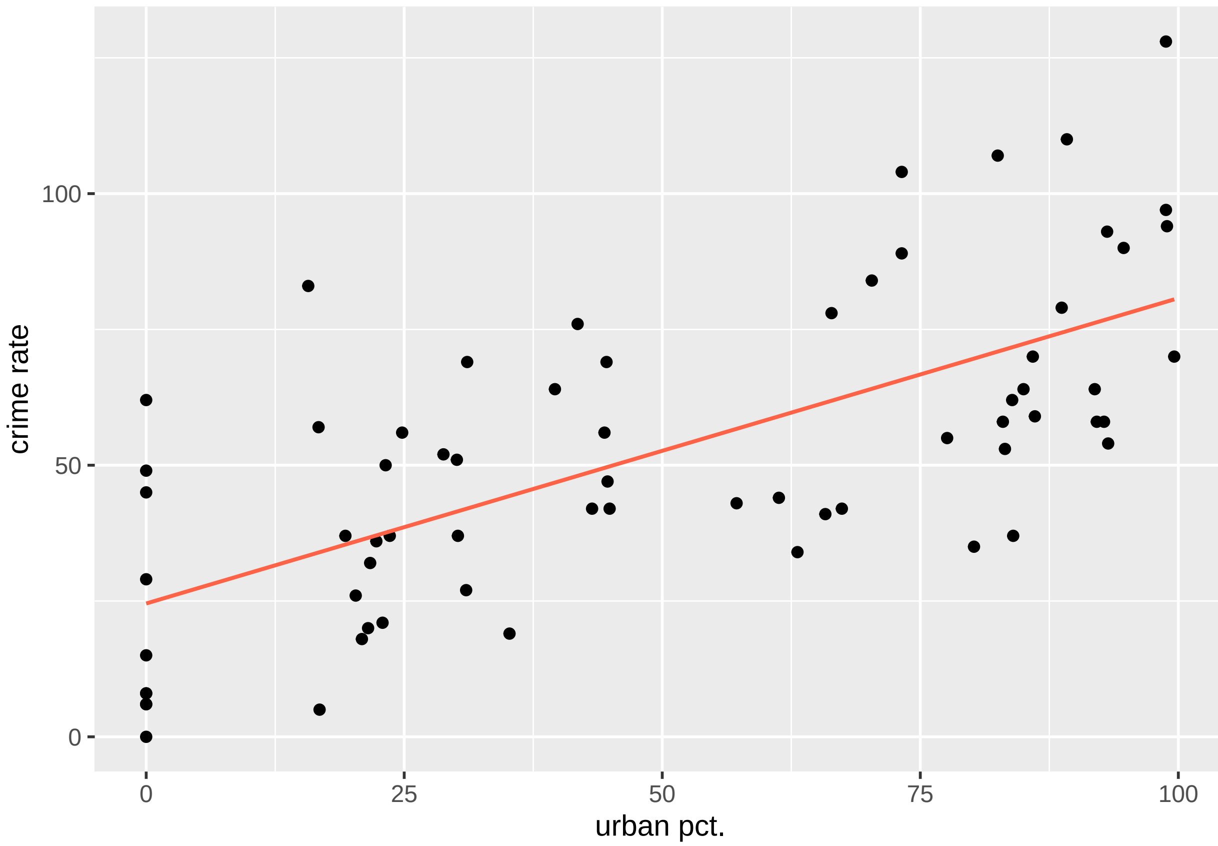
Create relevant figures:

```
boxplot(florida$C,ylab="crime rate",col="magenta")
```




```
gf_point(C ~ U, data = florida,
         xlab = "urban pct.", ylab = "crime rate") %>% gf_lm(color="tomato")
```

```
## Warning: Using the `size` aesthetic with geom_line was deprecated in ggplot2 3.4.0.
## i Please use the `linewidth` aesthetic instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Fit linear regression model:

```
fit <- lm(C ~ U, data = florida)
summary(fit)
```

```
##
## Call:
## lm(formula = C ~ U, data = florida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.766 -16.541  -4.741  16.521  49.632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.54125    4.53930   5.406 9.85e-07 ***
## U            0.56220    0.07573   7.424 3.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 65 degrees of freedom
## Multiple R-squared:  0.4588, Adjusted R-squared:  0.4505
## F-statistic: 55.11 on 1 and 65 DF,  p-value: 3.084e-10
```

b)

$$\hat{y} = 24.5 + 0.562x.$$

- Intercept: If urbanisation is zero, we expect a crime rate of 24.5
- Slope: If urbanisation increases by one, we expect a crime rate increase of 0.562

c)

```
100*0.562
```

```
## [1] 56.2
```

So the difference in crime rate between the two extremes is 56.2.

d)

```
R=cor(C~U,data=florida)
R
```

```
## [1] 0.6773678
```

```
R^2
```

```
## [1] 0.4588271
```

R is the correlation between predictions and observations which for simple linear regression is equivalent to the correlation between the response and the explanatory variable. Thus an R value of 0.677 indicate a moderate linear dependence between crime rate and urbanisation. In addition, the correlation is positive indicating a positive relationship.

R^2 is a similar measure giving how much of the variation in the response variable is explained by the explanatory variables relative to the total variation of the response variable; thus the interpretation is elegant. In this case with $R^2 = 0.459$, 45.9% of the total variation of crime rate is explained by the urbanisation.

In conclusion the urbanisation seems to be important in relation to modelling the crime rate, however it seems reasonable to include more variables in the model that may explain crime rate(we may look forward to this in the next lecture).

9.33:

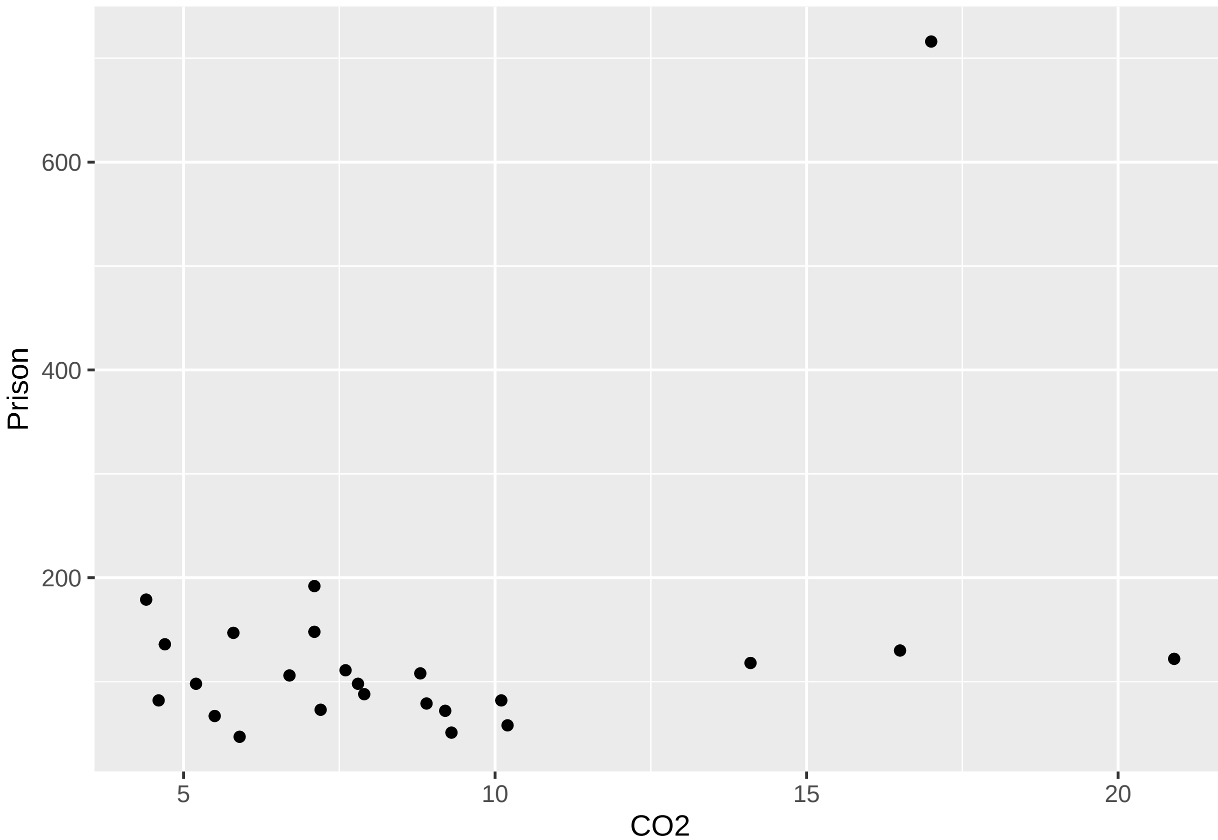
Additional sub-exercises:

Import data:

```
oecd <- read.table("https://asta.math.aau.dk/datasets?file=OECD_Agresti_ed5.dat", header = TRUE)
```

Create relevant figures:

```
gf_point(Prison ~ CO2, data = oecd,
         ylab = "Prison", xlab = "CO2")
```



Compute correlation:

```
cor(Prison ~ CO2, data = oecd)
```

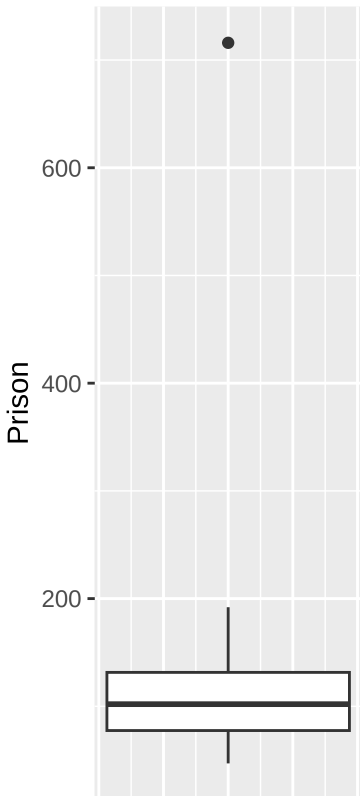
```
## [1] 0.390552
```

There is a weak positive linear relationship between carbon dioxide emissions and prison populations. The positivity means that increasing carbon dioxide emissions increases the prison population.

a)

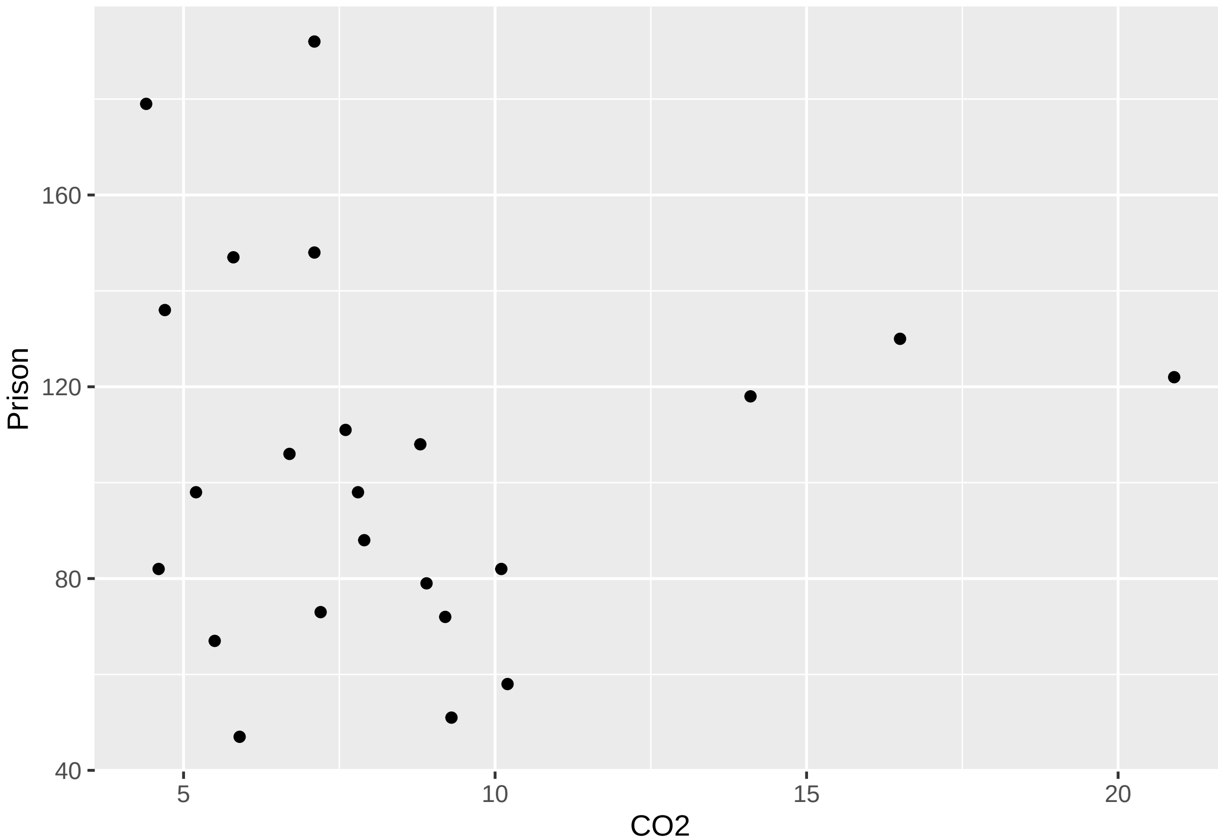
There is a nation with a very high prison population (over 700). When looking at a boxplot of prison population we also see that this is an outlier.

```
gf_boxplot(Prison ~ 1, data = oecd,
           xlab = "") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```



Removing the point and computing correlation yields the following:

```
oecd2 <- subset(oecd, Prison != max(Prison))
gf_point(Prison ~ CO2, data = oecd2,
         ylab = "Prison", xlab = "CO2")
```




```r
cor(Prison ~ CO2, data = oecd2)
```

```
## [1] 0.0004736938
```

We see no correlation what so ever, showing us that one data point may have a huge influence on summary statistics (such as the mean, variance, correlation, and so on), thus it is important to always look at plots of the data.