S2S Lab 4

Exercise Solutions

Exercise 1

The data set `VIT2005` from the `PASWR2` package stores information relating to 218 different apartments in Vitoria, Spain. For a full list of the variables included and the information they represent, use the code `help(VIT2005)`.

You can load `VIT2005` into your **Environment** tab using the following code.

```
library(PASWR2)
data("VIT2005")
```

a. The variable "out" details how much of an apartment is exposed to the elements. It has levels "E25", "E50", "E75" and "E100" corresponding to 25%, 50%, 75% and 100% exposure respectively.

   Create a frequency table, a pie chart and a barplot in base R which show the number of apartments from `VIT2005` within each category of the "out" variable.

   You may wish to relabel and reorder the categories of "out" first of all.

b. The variable "area" is a numeric variable detailing the size (in square metres) of each apartment.

   i. Create a new categorical variable, called "size", indicating whether an apartment is considered "Small", "Medium" or "Large". Apartments smaller than 100m^2 should be labelled as small, those between 100m^2 and 150m^2 should be labelled as medium and any apartments larger than 150m^2 should be labelled as large.

   The function `cut()` will be useful for this.

   ii. Using this new categorical variable, create a barplot in base R to investigate whether the proportion of apartments in each level of exposure is different for small, medium and large apartments.

   Make sure that apartment size is placed along the x-axis, there is a different bar for each level of exposure and that the y-axis shows proportion, not frequency.

c. "totalprice" gives the market price (in Euros) for each apartment.

   i. Create a histogram in base R showing the shape of the distribution of apartment prices. Superimpose a kernel density estimate of the distribution on the histogram.

   ii. Investigate whether the distribution of apartment price is normal using a QQ plot. Do you think this is a normal distribution?

d. Use three histograms to compare the distributions of apartment prices for small, medium and large apartments. Make sure you are using the same `breaks =` and range along the y-axes in all plots so that the distributions can be easily compared.

   What differences are there between these distributions?

   Remember to subset the data first so that the first plot is only for prices of small apartments etc.

Solution

a. Change the labels of "out" and reorder them to be "25%", "50%", "75%", "100%".

```
VIT2005$out <- factor(x = VIT2005$out, levels = c("E25", "E50", "E75", "E100"),
                      labels = c("25%", "50%", "75%", "100%"))
```

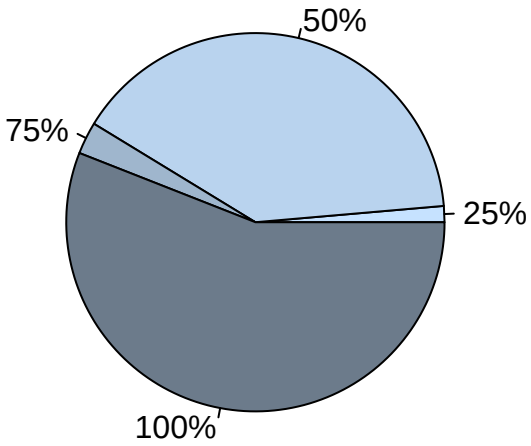
Frequency table:

```
expo_tab <- table(VIT2005$out)
expo_tab
```

```
 25%  50%  75% 100%
   3   87    6  122
```

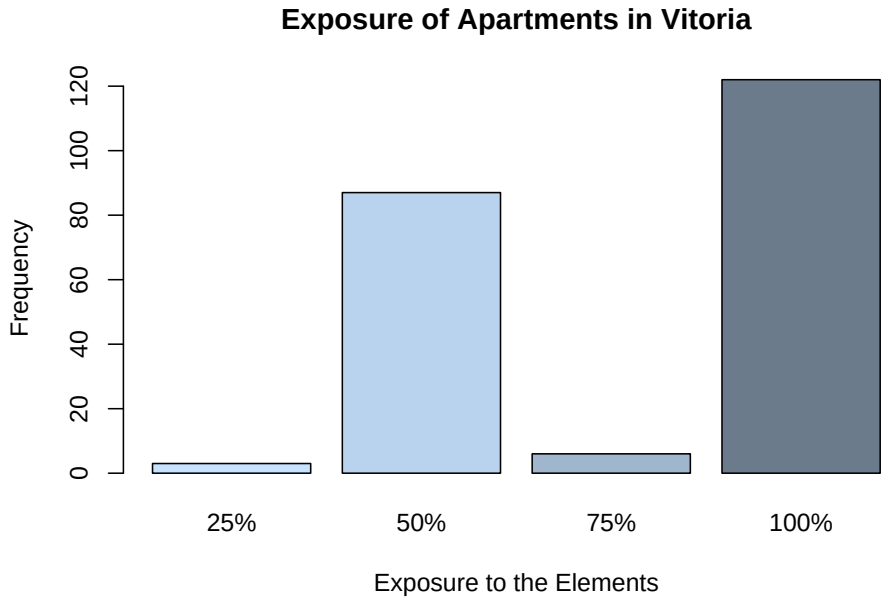
Pie chart:

```
pie(x = expo_tab,
    col = c("slategray1", "slategray2", "slategray3", "slategray4"),
    main = "Exposure of Apartments in Vitoria")
```



Barplot:

```
barplot(height = expo_tab,
        col = c("slategray1", "slategray2", "slategray3", "slategray4"),
        main = "Exposure of Apartments in Vitoria",
        xlab = "Exposure to the Elements",
        ylab = "Frequency")
```

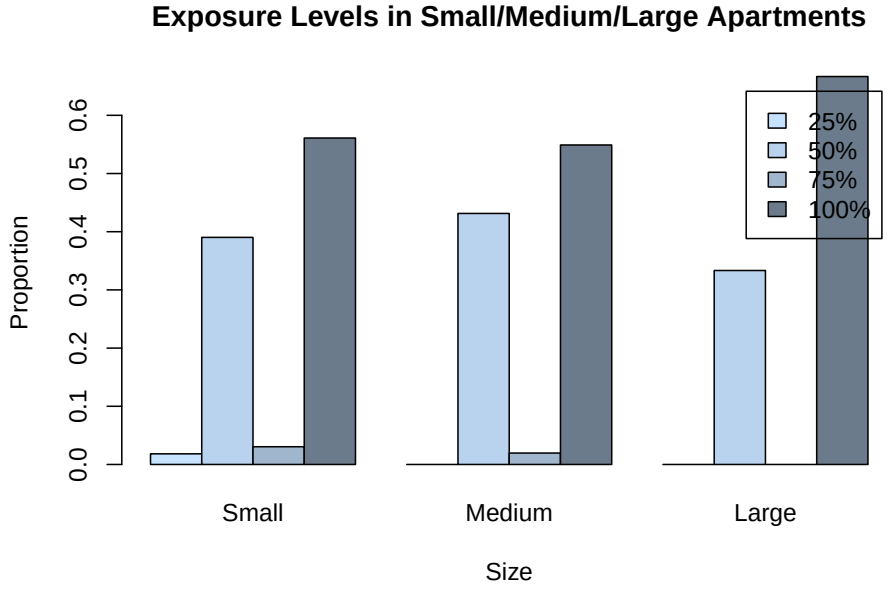


b. Create the new variable "size".

```
VIT2005$size <- cut(x = VIT2005$area,
                    breaks = c(50, 100, 150, 200),
                    labels = c("Small", "Medium", "Large"))
```

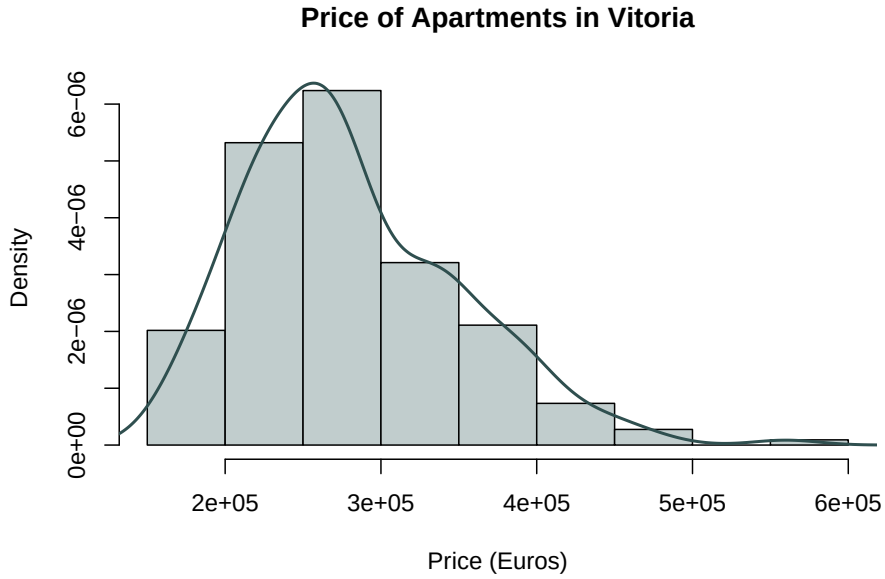
Barplot:

```
size_prop_tab <- prop.table(table(VIT2005$size, VIT2005$out), margin = 1)

barplot(height = t(size_prop_tab),
        beside = TRUE,
        legend.text = TRUE,
        col = c("slategray1", "slategray2", "slategray3", "slategray4"),
        main = "Exposure Levels in Small/Medium/Large Apartments",
        xlab = "Size",
        ylab = "Proportion")
```



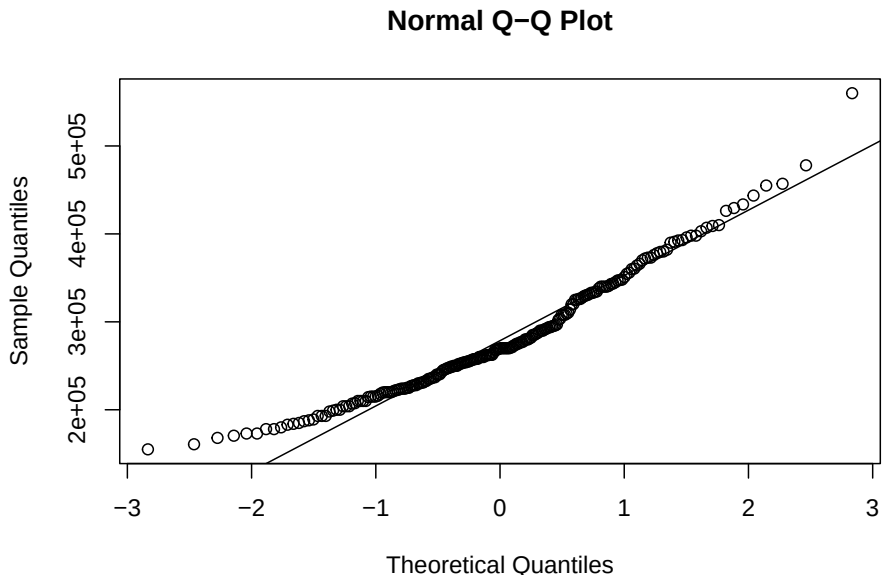
c. Distribution of apartment price:

```
hist(x = VIT2005$totalprice,
     breaks = 10,
     freq = FALSE,
     col = "azure3",
     main = "Price of Apartments in Vitoria",
     xlab = "Price (Euros)")

price_dens <- density(VIT2005$totalprice)

lines(price_dens, lwd = 2, col = "darkslategray")
```



Normal QQ plot for apartment price:

```
qqnorm(y = VIT2005$totalprice)
qqline(y = VIT2005$totalprice)
```

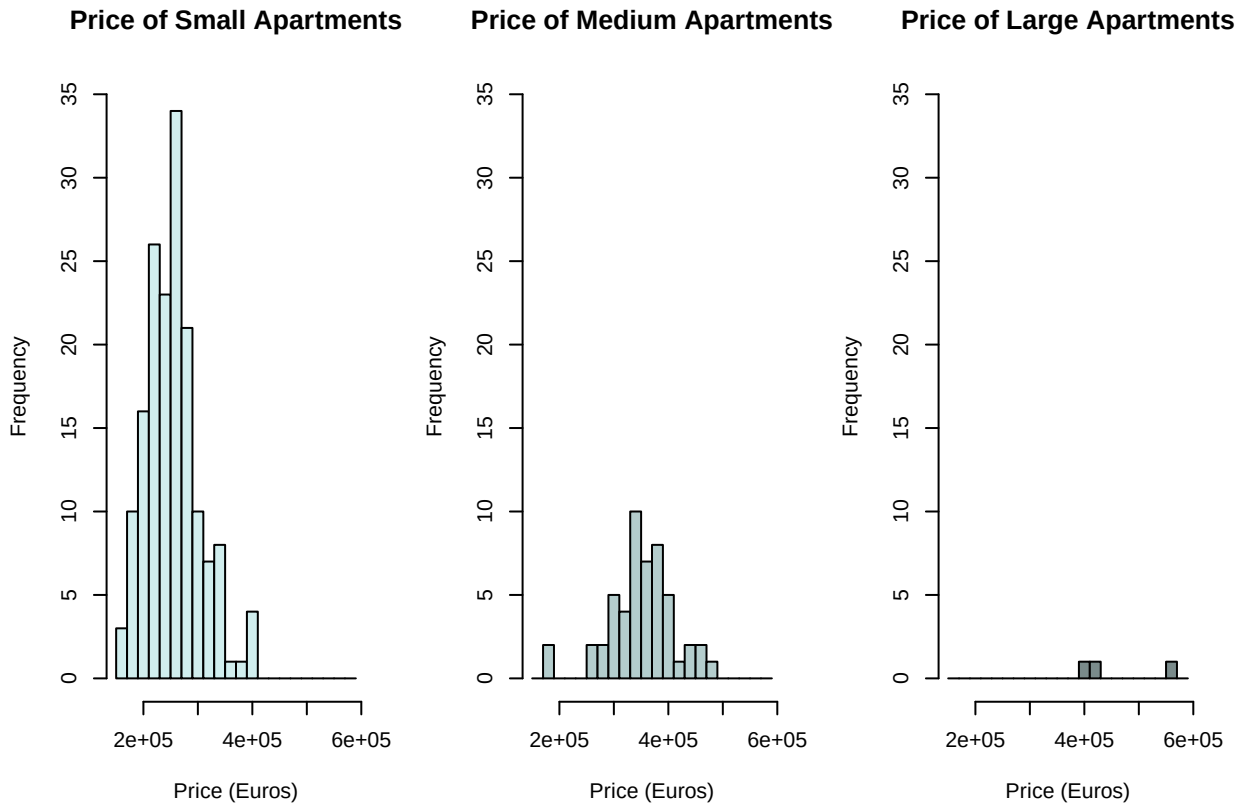


d. Subset the data to show the price of apartments for small, medium and large apartments separately.

```
price_small <- subset(x = VIT2005,
                      subset = (size == "Small"),
                      select = totalprice)

price_medium <- subset(x = VIT2005,
                       subset = (size == "Medium"),
                       select = totalprice)

price_large <- subset(x = VIT2005,
                      subset = (size == "Large"),
                      select = totalprice)
```

Plot three histograms, on the same scale, showing the distributions of apartment prices.

```
par(mfrow = c(1, 3))

hist(x = price_small$totalprice,
     breaks = seq(from = 150000, to = 600000, by = 20000),
     col = "lightcyan2",
     ylim = c(0, 35),
     main = "Price of Small Apartments",
     xlab = "Price (Euros)")

hist(x = price_medium$totalprice,
     breaks = seq(from = 150000, to = 600000, by = 20000),
     col = "lightcyan3",
     ylim = c(0, 35),
     main = "Price of Medium Apartments",
     xlab = "Price (Euros)")

hist(x = price_large$totalprice,
     breaks = seq(from = 150000, to = 600000, by = 20000),
     col = "lightcyan4",
     ylim = c(0, 35),
     main = "Price of Large Apartments",
     xlab = "Price (Euros)")

par(mfrow = c(1, 1))
```

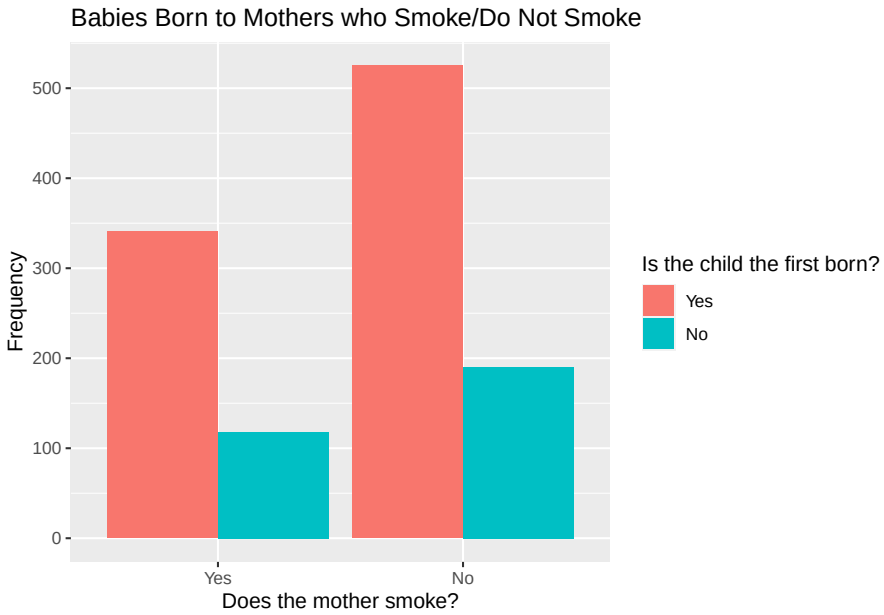


Exercise 2

The file "*babies.csv*" contains data on the birth weight of babies born to different mothers and whether each mother currently smokes or does not smoke (a full list of the variables included in this data set can be found here).

Download the data from Moodle and load the data set into your **Enivronment** tab using the code below and use `babies` to recreate the following two plots using **ggplot2**.

```
babies <- read.csv(file = "babies.csv")
```

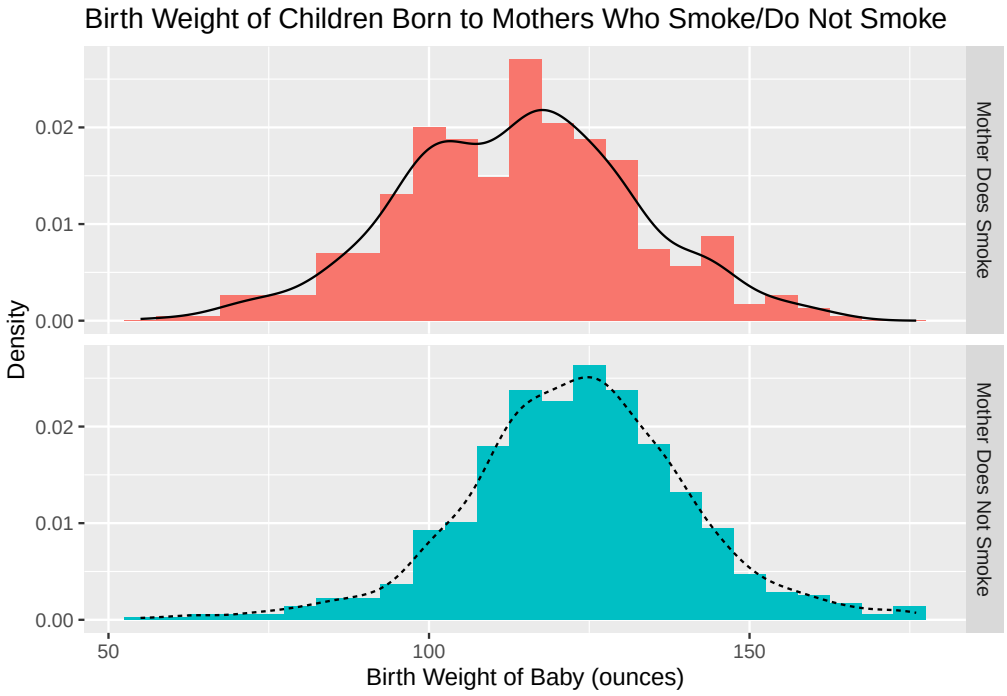


Solution

To make this first plot, the labels of the variables "smoke" and "parity" first have to be changed so they show as "Yes" and "No" in the plot.

```
babies$smoke <- factor(x = babies$smoke, labels = c("No", "Yes"))
babies$smoke <- factor(x = babies$smoke, levels = c("Yes", "No"))
babies$parity <- factor(x = babies$parity, labels = c("Yes", "No"))
```

```
ggplot(data = babies) +
  geom_bar(aes(x = smoke, fill = parity), position = "dodge") +
  labs(title = "Babies Born to Mothers who Smoke/Do Not Smoke",
       x = "Does the mother smoke?", y = "Frequency") +
  guides(fill = guide_legend("Is the child the first born?"))
```



Solution

To make the second plot, the labels for the "smoke" variable again have to be changed so they show as "Mother Does Smoke" and "Mother Does Not Smoke".

```
babies$smoke <- factor(x = babies$smoke, labels = c("Mother Does Smoke", "Mother Does Not Smoke"))
```

```
ggplot(data = babies) +
  geom_histogram(aes(x = bwt, y = after_stat(density), fill = smoke), binwidth = 5) +
  geom_density(aes(x = bwt, linetype = smoke)) +
  facet_grid(smoke ~ .)+
  labs(title = "Birth Weight of Children Born to Mothers Who Smoke/Do Not Smoke",
       x = "Birth Weight of Baby (ounces)", y = "Density") +
  guides(fill = "none", linetype = "none")
```