Fraud Detection
Rita P. Ribeiro

2022/2023
DCC - FCUP

1 Hands On: Data Understanding

1.1 Data Import

1. The Echocardiogram data set in the UCI Machine Learning repository contains information on a set of patients that suffered heart attacks at some point in the past.

(a) Import from the file echocardiogram.csv the Echocardiogram data set to a tibble. Inspect the type of each attribute, for that you can use function `spec()` or `str()`.

```
ec <- read_csv("echocardiogram.csv")
ec
spec(ec)
```

(b) Read the information on the data set and find out how missing values are represented and make sure that they are properly represented in R. Import the data set and inspect the type of each attribute again.

```
ec <- read_csv("echocardiogram.csv",na = "?")
ec
spec(ec)
```

(c) According to the information on the webpage, is there any redundant or irrelevant attribute that you can remove? Remove them.

```
ec <- ec %>% select(-c(mult,name,group))
```

(d) Run the function `summary()` on your tibble.

```
summary(ec)
```

(e) Is there any data type change that you find necessary? You can use the functions `as.numeric()` or `as.factor()` inside `mutate()` for coercing an attribute to a numeric or categorical one, respectively. Obtain the summary of the new tibble.

```
ec <- ec %>% mutate(survival=as.numeric(survival),
                    still_alive=as.factor(still_alive),
                    pericardial_effusion=as.factor(pericardial_effusion),
                    alive_at=as.factor(alive_at))

summary(ec)
```

1.2 Summarization

2. Load the data set carIns_noNAs.Rdata about the insurance risk rating of cars based on several characteristics of each car. Detailed information on this can be found in here. Using the package `dplyr`, answer the following questions:

```
load("carIns_noNAs.Rdata")
```

(a) Obtain the number of cars by `bodyStyle`.

```
carIns_noNAs %>% group_by(bodyStyle) %>% count()
```

(b) Obtain the number of cars by `bodyStyle` and `fuelType`.

```
carIns_noNAs %>% group_by(bodyStyle,fuelType) %>% count()
```

(c) Obtain the mean and the standard deviation of the attribute `cityMpg` by `bodyStyle` in ascending order.

```
carIns_noNAs %>%  group_by(bodyStyle) %>%
  summarize(cityMpg.mean = mean(cityMpg), cityMpg.sd = sd(cityMpg)) %>% arrange(cityMpg.mean)
```

1.3 Visualization

3. Using the package `ggplot2`, create graphs that you find adequate to answer the following questions on the data carIns_noNAs.

```
library(ggplot2)
```

(a) Show the relationship between the attributes `cityMpg` and `highwayMpg`

```
ggplot(carIns_noNAs,aes(x=cityMpg,y=highwayMpg)) + geom_point() +
     ggtitle("Relationship between cityMpg and highwayMpg")
```

(b) Show the distribution of cars by `bodyStyle`.

```
ggplot(carIns_noNAs,aes(x=bodyStyle)) + geom_bar() +
  ggtitle("Distribution of cars across bodyStyle")
```

(c) Show the distribution of cars by `price`. Suggestion: create bins of width equal to 5000.

```
ggplot(carIns_noNAs,aes(x=price)) + geom_histogram(binwidth = 5000) +
  ggtitle("Histogram of price")
```

(d) Add the information of the density estimation to the previous graph.

```
ggplot(carIns_noNAs,aes(x=price))  + geom_histogram(binwidth = 5000, aes(y=..density..)) +
  geom_density(color="blue") + geom_rug() +
  ggtitle("Histogram of price")
```

(e) Show the distribution of `price` by `make` attribute. Suggestion: use boxplots and the function `coord_flip()`.

```
ggplot(carIns_noNAs,aes(x=make,y=price)) + geom_boxplot() + coord_flip()
```

(f) Show the distribution of `price` by `nDoors` attribute. Suggestion: use histograms.

```
ggplot(carIns_noNAs,aes(x=price)) + geom_histogram(binwidth = 5000) + facet_wrap(~nDoors) +
  ggtitle("Histogram of price by nDoors")
```

(g) Show the distribution of `price` by `bodyStyle` and `nDoors` attributes. Suggestion: use histograms.

```
ggplot(carIns_noNAs,aes(x=price)) + geom_histogram(binwidth = 5000) + facet_grid(fuelType ~ aspiration) +
  ggtitle("Histogram of price by aspiration and fuel type")
```

(h) Add the parameter `scales="free_y"` to the facet function in the previous graph.

```
ggplot(carIns_noNAs,aes(x=price)) + geom_histogram(binwidth = 5000) + facet_grid(fuelType ~ aspiration,scales="free_y") +
  ggtitle("Histogram of price by aspiration and fuel type")
```

Please note:

- "%>%" is the chaining operator or pipe: `x %>% f(y)` becomes `f(x,y)`
- "." represents the previous value in the chain, i.e. `x %>% f(.)` becomes `f(x)`
- "~" is used for anonymous functions. i.e. `function(x) x + 2` can be written as
  - `~ .x + 2`, where `.x` represents the first argument of the function
  - `~ . + 2`, in case the first argument is the previous value in the chain