forcats practical

Jumping Rivers

First we'll load the data and the relevant packages

```
library(tidyverse)
data(okcupid, package = "jrTidyverse")
```

This is a data set of dating profiles from the dating website OK Cupid. We're going to specifically focus on how drinking affects income in the land of OK Cupid.

Question 1

a) The column `drinks` corresponds to a persons answer about their drinking habits. What is the average income of each group of drinking habits? Save this as a data frame called `drinks_in` and have the column containing the average income called `av_in` Hint: use `group_by() then summarise()`

```
drinks_in = okcupid %>%
  group_by(drinks) %>%
  summarise(av_in = mean(income))
```

b) We can plot the average incomes using **ggplot2**

```
drinks_in %>%
  ggplot(aes(x = drinks, y = av_in)) +
  geom_point()
```

c) Previously we saw how to rename factors using `fct_recode()`. However, this will not work with missing values i.e. `NA`'s. A function that will is `fct_explicit_na()`. Try running

```
x = c(1,2,3,NA)
(y = factor(x))
## [1] 1    2    3    <NA>
## Levels: 1 2 3
```

Notice how the NA isn't included in the factors?

```
fct_explicit_na(y, "unknown")
## [1] 1       2       3       unknown
## Levels: 1 2 3 unknown
```

That will rename the NA factors as "Unknown". Use `mutate()` and `fct_explicit_na()` to rename the missing values to something more appropriate. Then plot the points using **ggplot2**.

```
drinks_in %>%
  mutate(drinks = fct_explicit_na(drinks, "Unknown")) %>%
  ggplot(aes(x = drinks, y = av_in)) +
  geom_point()
```

d) Reorder the points from lowest average income to highest, then plot using **ggplot2**
Hint: use `mutate()` and `fct_reorder()`

```
drinks_in %>%
  mutate(drinks = fct_explicit_na(drinks, "Unknown")) %>%
  mutate(drinks = fct_reorder(drinks, av_in)) %>%
```

```
  ggplot(aes(x = drinks, y = av_in)) +
  geom_point()
```

e) Reorder the points order from people who drink least to people who drink most and then plot using **ggplot2**. Put the category "Unknown" where you deem appropriate.
Hint: use `fct_relevel()`

```
drinks_in %>%
  mutate(drinks = fct_explicit_na(drinks, "Unknown")) %>%
  mutate(drinks = fct_relevel(drinks,
                              "Unknown", "not at all",
                              "rarely", "socially",
                              "often", "very often",
                              "desperately")) %>%
  ggplot(aes(x = drinks, y = av_in)) +
  geom_point()
```

f) Summarise the groups average income in the same way, but this time collapse "not at all" and "rarely" into "low", "socially" and "often" into "medium" and then "very often" and "desperately" into "high". Plot it using **ggplot2**.

```
okcupid %>%
  mutate(drinks = fct_explicit_na(drinks, "Unknown")) %>%
  mutate(drinks = fct_collapse(drinks,
                               Low = c("not at all", "rarely"),
                               Medium = c("socially", "often"),
                               High = c("very often", "desperately"))) %>%
  group_by(drinks) %>%
  summarise(av_in = mean(income)) %>%
  ggplot(aes(x = drinks, y = av_in)) +
  geom_point()
```