Tutorial Exercises Week 7

Question 1

Download the two datasets:

- cpb-house-prices.csv
- cpb-pop-growth.csv

Read in both datasets. When reading in the house price dataset you should use the following command:

```
read.csv("cpb-house-prices.csv", sep = ";", dec = ",")
```

This is because the dataset uses semicolons to separate the columns instead of commas, and uses commas for decimals.

Rename the 3 variables to: `"municipality"`, `"house_price_2022"`, `"house_price_2021"`.

The 2nd dataset is can be read in with the `read.csv()` function without any special options. Rename the 2 variables in that dataset to: `"municipality"`, `"pop_growth_2018_2023"`.

Merge the two datasets together by the variable `"municipality"`.

One municipality from the population growth dataset fails to merge with the house price dataset. Which municipality is this?

Question 2

How many municipalities from the house price dataset fail to merge with the population growth dataset?

Question 3

Create a scatter plot using ggplot of population growth on the horizontal axis and the house price in 2022 on the vertical axis.

Add the following layer to your plot to get a fitted line through the points:

```
geom_smooth(method = "lm")
```

Choose the answer below which best interprets what we can see in the plot.

- Municipalities with higher population growth on average have higher house prices.
- Municipalities with higher population growth on average have lower house prices.

Question 4

Reshape the original house price dataset from wide format to long format using the municipality as the ID variable. How many rows does the long format dataset have?

Question 5

If you correctly reshaped the dataset from the previous question the first 4 rows should look like:

```
  municipality         variable  value
1  Bloemendaal house_price_2022 1118.9
2     Blaricum house_price_2022 1099.1
3  Laren (NH.) house_price_2022 1030.1
4    Wassenaar house_price_2022  970.8
```

Suppose the long format dataset is called `df1_long`. Which of the following commands will return the dataset back to its original format (apart from the order of the observations)?

- `dcast(df1_long, municipality ~ variable)`
- `dcast(df1_long, variable ~ municipality)`
- `dcast(df1_long, value ~ municipality)`
- `dcast(df1_long, municipality ~ value)`

Question 6

Download the dataset municipality-province.csv.

This dataset contains two variables: the municipality and the province in which each municipality is located.

Read in the dataset and rename the variables to `"municipality", "province"`.

Merge the `municipality-province.csv` dataset with your previously-merged house price and population growth dataset.

Calculate the average of the variable `house_price_2022` by province.

Which province has the highest average?