INFO370 Lab 02: Explore and clean data

Deadline: Apr 14, 11:59pm

Instructions

This lab is largely about understanding data and simple manipulations of data. In particular, you need to work with missings, strings, and plotting.
You are using AirBnB data from Seattle. Your task is to address the question:

```
How does the listing price depend on the size of the apartment?
```

The data is downloaded from AirBnB, http://insideairbnb.com/get-the-data.html, "as-is". There is no documentation regarding the variables and what do they mean. So we cannot provide much guidance.
Working together is fun and useful but you have to submit your own work. Discussing the solutions and problems with your classmates is all right but do not copy-paste their solution! First understand, and thereafter create your own solution. Please list all your collaborators on the solution.

1 Describe the data

1. Load "airbnb-seattle-listings" data from canvas.
   Hint: this is a large file. If your computer gets too slow and sluggish, use only subset of this data (check out nrows option for `pd.read_csv`). If this is the case, state clearly that you are only working on a subset.

2. Below, we only need variables *price*, *bedrooms*, *name* and *square feet*. For clarity, drop all other variables from data. Print a few lines of data and comment if this looks good.
   Note: you may also achieve this when loading data (in the read csv step above), check out its documentation.

3. Now use the selected columns and perform the basic data description:
   (a) How many observations, columns do you have?
   (b) How many missing values do you have in each variable
   (c) What are the data types of each column?

4. Based on your results in the previous question: what do you think, what variables can you use for the analysis below? Do the data types look what you expect?

---

2 Prepare and analyze variables

Let's now take a closer look at the following variables: *price* and *bedrooms*.

1. Convert price to numeric. Ensure you do not introduce any missings into price. If you see missings in price, demonstrate that this was due to invalid number in the original data.
   Hint: check out `pd.Series.str.replace` method and `pd.to_numeric` function.
   Hint 2: you may create a new variable, the numeric price, and thereafter print a sub-data frame that only contains the original and the numeric price. These two should match!

2. For all these variables, find min, max and mean value. Discuss if these values are reasonable.
   Hint: check out the corresponding methods like `pd.Series.min()` etc.

---

3 How is price related to size?

Finally, let's answer the question: how is price related to *bedrooms*?

1. Compute mean price by different values for bedrooms. What do you see? Does it make sense? Hint: use `.groupby`

2. Plot the mean price as a function of *bedrooms*. Ensure you label the plot appropriately. Comment the result.
   You may do barplot using `plt.bar`, but you may also consider different plots.
   Hint: you may use `.groupby(..., as_index=False)`. This preserves the grouping variable as variable (otherwise it will be index).

---

4 Add-on bonus task (not graded)

If you have time and interest, here is an extra task for you:

1. You may have seen a curious result: the cheapest price is 0 and the smallest listing has 0 bedrooms.
   Let's take another look at these listings:
   (a) How many 0-price cases there are?
   (b) How many 0-BR cases there are?
   Analyze a few 0-price and 0-BR listings (check description, name and perhaps also picture url). Can you see what is the reason they have 0-price and 0-bedrooms?

2. Replicate the table in question 3.1 with all missings and otherwise suspicious cases removed.
