Project description

The data you work with will not always feel familiar —sometimes, you'll encounter data from peculiar sources, operating with peculiar measurements. Let's work with something exotic to keep you on your toes: Yandex.Realty has an archive of property ads for St. Petersburg, Russia, and the surrounding areas covering the past few years. You'll need to learn how to determine the market value of real estate. Your task is to define parameters. This will make it possible to create an automated system that is capable of detecting anomalies and fraud.
There are two different types of data available for every apartment for sale. The first type is user input. The second type was calculated automatically based upon the map data. This includes distance from the city center, the airport, and the nearest park or body of water.

Instructions on completing the project

Step 1. Open the data file and study the general information
File path: */datasets/real_estate_data_us.csv*. Download dataset
Step 2. Data preprocessing

- Determine and study the missing values:
    - A practical replacement can be presumed for some missing values. For example, if the user doesn't enter the number of balconies, then there probably aren't any. The correct course of action here is to replace these missing values with 0. There's no suitable replacement value for other data types. In this case, leave these values blank. A missing value is also a key indicator that mustn't be hidden.
    - Fill in the missing values where appropriate. Explain why you've chosen to fill the missing values in these particular columns and how you selected the values.
    - Describe the factors that may have led up to the missing values.
- Convert the data to the required types:

    - Indicate the columns where the data types have to be changed and explain why.

Step 3. Make calculations and add the following entries to the table:

- the price per square meter
- the day of the week, month, and year that the ad was published
- which floor the apartment is on (first, last, or other)
- the ratio between the living space and the total area, as well as between the kitchen space and the total area.

Step 4. Conduct some exploratory data analysis and follow the instructions below:

- Carefully investigate the following parameters: square area, price, number of rooms, and ceiling height. Plot a histogram for each parameter.
- Examine the time it's taken to sell the apartment and plot a histogram. Calculate the mean and median and explain the average time it usually takes to complete a sale. When can a sale be considered to have happened rather quickly or taken an extra long time?
- Remove rare and outlying values and describe the patterns you've discovered.
- Which factors have had the biggest influence on an apartment’s price? Examine whether the value depends on the total square area, number of rooms, floor (top or bottom), or the proximity to the city center area. Also check whether the publication date has any effect on the price: specifically, day of the week, month, and year. Note that using scatter plot is preferable to hexbin. If you do decide to use hexbin, please use scatter plot too, and then compare the results. It is also recommended to check the hexbin documentation and carefully study its parameters.
- Select the 10 localities with the largest number of ads then calculate the average price per square meter in these localities. Determine which ones have the highest and lowest housing prices. You can find this data by name in the ’*locality_name’* column.

- Thoroughly look at apartment offers: Each apartment has information about the distance to the city center. Select apartments in Saint Petersburg (*‘locality_name’*). Your task is to pinpoint which area is considered to be in the city center. In order to do that, create a column with the distance to the city center in km and round to the nearest whole number. Next, calculate the average price for each kilometer and plot a graph to display how prices are affected by the distance to the city center. Find a place on the graph where it shifts significantly. That's the city center border.
- Select all the apartments in the city center and examine correlations between the following parameters: total area, price, number of rooms, ceiling height. Also identify the factors that affect an apartment’s price: number of rooms, floor, distance to the city center, and ad publication date. Draw your conclusions. Are they different from the overall deductions about the entire city?

Step 5. Write an overall conclusion

v