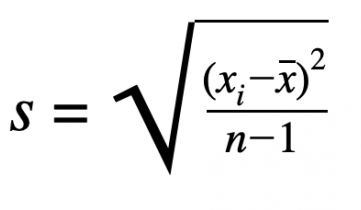
**Assignment 2**

**Explanation:**

* *You have a dataset that has houses for sale. Each row is a house for sale. You will analyze this data. The data is* [*here*](https://docs.google.com/spreadsheets/d/1YUTZTTGexqfWDmBYhY7cX67V0_4LK7e_/edit?usp=sharing&ouid=105769116295230040317&rtpof=true&sd=true)*.*
* *You will submit your jupyter codes through online.yildiz.edu.tr. You will need to* ***zip*** them*. You can do it by clicking the* ***right mouse button****, then click* ***send*** *and finally click* ***zipped folder***
* *Submit a Jupyter file which has a cell that has the question and in the following cell there must be the answer. An example jupyter file can be found* [*here*](https://drive.google.com/file/d/1iqYrbbispm5U2uez4VPt9c_GLpLbKXFH/view?usp=sharing)*:*
* *You can change your groups or keep the same group. If your group members decided not to work somehow or withdraw the lecture or any other unexpected thing happened , you can continue alone. But send me an email regarding your situation. Groups are 2-3 students, again.*
* *If you miss the first due date, you will lose 1 points for each hour. You can be at most 2 days late which means you can get at most 52 points if you submit it at 24.01.2024 - 23.45.*
* *This sheet is* ***3 pages.***

**TASKS**:

1. Import the data as a dataframe.
2. Prices of some houses are missing. Find them and replace them with the average prices of the districts they belong to. For example if the house is from Şişli, then assign the **average** price of the houses in Şişli.
3. There is some other missing data as well in the given dataset. Fill in the missing values using the following rules:
   1. If it’s a continuous number like the size of the house, then put the **median** of the given dataset.
   2. If it’s a categorical variable like Heating type, assign the **most frequent** value.
4. There are some repetitive rows, i.e. the same houses are given in different lines. If everything except the price is the same, then they belong to the same house which means these rows are repetitive. Find and clear them by erasing the data which has the larger price since it means that the house is not sold and the owner has reduced the price in a new ad to sell the house.
5. Find the average price of houses for each district (district) and create a new dataframe that only has the districts with these average prices. You will use this dataframe in the following questions as well.
6. Find the standard deviation of the prices of each district. You can find the definition of the standard deviation from wikipedia or use the following formula where Xbar is the average of the data. Add the standard deviation of each district to the district dataframe.



1. Let’s denote the average and the standard deviation with greek letters μ and σ, respectively. Then a **reasonable price range** can be defined as: [μ - 2σ, μ + 2σ], i.e., the price of a house is reasonable if it’s at most 2 standard deviations away from the mean. We have found the mean and the standard deviation of each district. You will use them to find whether a house has a reasonable price or not. If it’s within the range defined by the mean and standard deviation of the district that it belongs to, then it has a reasonable price. Add a new column to the house dataframe called “Price range” and write “IN” or “OUT” if they are within the range or out of the range, respectively.
2. Find the number of houses whose price is less than 250K TL, between 250K(exclusive) and 500k (inclusive) and larger than 500k(exclusive). (K means thousand.) The output should be dictionary like this:

{ "0-250":20,

"250-500":40,

"500 and over"50}

1. Write a function which generates the above lists automatically. That is, write a function which calculates the number of houses for a given list of prices. For example, for the above question, the input will be [250,500]. If you are given [250, 500, 750, 1000], then the function will return a dictionary like this:

{ "0-250":20,

"250-500":30,

"500-750":40,

"750-1000":40,

"1000 and over": 50}

1. Find the average price with respect to “room count” and return it as a dictionary like above.

1. Find the average price of 3+1 houses in each district and return it as a dictionary like above.
2. Print the house dataframe which is ordered in terms of the prices of the houses.
3. Print the dataframe which is ordered in terms of first **district** and then **street lexicographically.** .
4. Make a scatter plot of size of the houses (X axis) and the price of the houses (Y axis.)
5. Divide the price of the houses by its net m2 size and add it as a new column to the house datafame.
6. Find the average m2 price of each district and add it to the district dataframe.
7. Make a bar plot for the average house price of the districts from the largest to the smallest.
8. In this question you will predict the price of a house and add a new column which has the price predictions. How to predict the price of a house? Well, a very easy way of doing this is to find the average of **n nearest neighbors** (i.e., houses). For this question, let’s pick 3 neighbors. How to find whether two houses are close to each other or not? You have the coordinates!