Spitogatos Data Analysis Assignment

Context

SpaN is a company that provides an online portal for real estate services. The main functionality of the portal is that property listings are published by real estate agents and visitors can search for properties based on a set of search criteria.

As a memeber of the Data Analysis team of SpaN, you closely cooperate with various departments to assist in decision making and strategy execution, based on actionable insights that you get out of data.

The following assignment consists of 3 parts. Each part has specific requirements and deliverables as described in the corresponding sections. After you complete the assignment, you need to provide a link to a public git repository that includes your code (in one file or several folders) and any other corresponding deliverables.

The dataset

The given dataset is this month's snapshot of all the listings of residential properties (houses) for sale listed on the portal of SpaN for 4 specific areas of Athens (**geography_name**). Each listing has a unique **id** but it can be the case that the same actual property is uploaded by multiple real estate agents so multiple different listings (with different **id**s) for the same property are created. Each agent is identified by a unique **agent_id** .

The rank that listings are ordered by in a specific area when a user makes a search depends on the type of the listing (**ad_type**) and their listing score (**ranking_score**). There are four different listing types: simple listings, that appear last, "UP" listings that rank above simple, "PREMIUM" listings that rank above "UP", and "STAR" listings that appear at the top of the list. Within each listing type group properties are ranked based on the listing score.

The size of the property (**sq_meters**), its **price** and the area (**geography_name**) are the main search filters that visitors use in their initial search. The rest of the columns of the dataset are all further attributes of the properties listed and can be used as filters on a search. The **year_of_construction** column represents the year that the house was built and the *value 2155* means that the house is under construction.

Assignment Part #1

The marketing department wants to issue a press release regarding house prices in different areas of Athens. They ask if you could help them by providing some useful statistics that would show the price levels of houses for sale this month that real estate journalists would find useful.

For this purpose, you will need to calculate tables that show some metrics, namely the mean, median and standard deviation of *property prices_per_sq_meter per house type (**subtype**)* and *per area (**geography_name**)*. Before you calculate the final metrics, keep in mind that you should clean the dataset from any entries that you think should not be included in the analysis, because they will corrupt the right image of the price_per_sq_meter levels of each area.

Assignment Part #2

The Sales Manager of SpaN, after conducting qualitative research, by asking different agents in each area in Athens, wants to examine the possibility of offering special discounts for some listing types, based on the competitiveness of each area. To decide what type of discount should be given to agents in each area she would need to see an analysis of the competitiveness of each area.

A highly competitive area would mean that it would be hard for a simple listing to rank high in the search results of this area just by having a high **ranking score**.

To help the sales manager decide the level of discount to be given to agents in each area, you would need to:

- Identify and calculate some *competitiveness metrics* that would show the level of difficulty for a listing to rank high in a specific area
- Plot those metrics in graphs that you believe would convey the right information to the sales manager, to be able to make the right decisions

Assignment Part #3

The product team of SpaN wants to launch a new page on the portal that would help agents decide the correct price they should set for a property for sale in Athens. The agent would need to input certain attributes of a property and an algorithm would value the property, based on historical data. The team is building an MVP (minimum viable product) that would be launched in beta, in order to measure the willingness of agents to use the new page and get feedback on the accuracy of the predictions, based on the experience of agents in the market.

They ask if you could help them identify what are the most important attributes that an agent would have to input to get a valid prediction of a property's price valuation and also build the model that would predict the value (**price**).

Using the data from the given dataset:

- Identify the most important attributes in predicting the price of a property.
- Build a model that valuates each residential property