Practical Exam: House sales

RealAgents is a real estate company that focuses on selling houses.

RealAgents sells a variety of types of house in one metropolitan area.

Some houses sell slowly and sometimes require lowering the price in order to find a buyer.

In order to stay competitive, RealAgents would like to optimize the listing prices of the houses it is trying to sell.

They want to do this by predicting the sale price of a house given its characteristics.

If they can predict the sale price in advance, they can decrease the time to sale.

Data

The dataset contains records of previous houses sold in the area.

Column Name	Criteria
house_id	Nominal. Unique identifier for houses. Missing values not possible.
city	Nominal. The city in which the house is located. One of 'Silvertown', 'Riverford', 'Teasdale' and 'Poppleton'. Replace missing values with "Unknown".
sale_price	Discrete. The sale price of the house in whole dollars. Values can be any positive number greater than or equal to zero. Remove missing entries.
sale_date	Discrete. The date of the last sale of the house. Replace missing values with 2023-01-01.
months_listed	Continuous. The number of months the house was listed on the market prior to its last sale, rounded to one decimal place. Replace missing values with mean number of months listed, to one decimal place.
bedrooms	Discrete. The number of bedrooms in the house. Any positive values greater than or equal to zero. Replace missing values with the mean number of bedrooms, rounded to the nearest integer.

Column Name	Criteria
house_type	Ordinal. One of "Terraced" (two shared walls), "Semi-detached" (one shared wall), or "Detached" (no shared walls). Replace missing values with the most common house type.
area	Continuous. The area of the house in square meters, rounded to one decimal place. Replace missing values with the mean, to one decimal place.

Task 1

The team at RealAgents knows that the city that a property is located in makes a difference to the sale price.

Unfortuntately they believe that this isn't always recorded in the data.

Calculate the number of missing values of the `city`.

- You should use the data in the file "house_sales.csv".
- Your output should be an object `missing_city`, that contains the number of missing values in this column.

Task 2

Before you fit any models, you will need to make sure the data is clean.

The table below shows what the data should look like.

Create a cleaned version of the dataframe.

- You should start with the data in the file "house_sales.csv".
- Your output should be a dataframe named `clean_data`.
- All column names and values should match the table below.

Column Name	Criteria
house_id	Nominal. Unique identifier for houses. Missing values not possible.
city	Nominal. The city in which the house is located. One of 'Silvertown', 'Riverford', 'Teasdale' and 'Poppleton' Replace missing values with "Unknown".
sale_price	Discrete. The sale price of the house in whole dollars. Values can be any positive number greater

Column Name	Criteria
	than or equal to zero. Remove missing entries.
sale_date	Discrete. The date of the last sale of the house. Replace missing values with 2023-01-01.
months_listed	Continuous. The number of months the house was listed on the market prior to its last sale, rounded to one decimal place. Replace missing values with mean number of months listed, to one decimal place.
bedrooms	Discrete. The number of bedrooms in the house. Any positive values greater than or equal to zero. Replace missing values with the mean number of bedrooms, rounded to the nearest integer.
house_type	Ordinal. One of "Terraced", "Semi-detached", or "Detached". Replace missing values with the most common house type.
area	Continuous. The area of the house in square meters, rounded to one decimal place. Replace missing values with the mean, to one decimal place.

Task 3

The team at RealAgents have told you that they have always believed that the number of bedrooms is the biggest driver of house price.

Producing a table showing the difference in the average sale price by number of bedrooms along with the variance to investigate this question for the team.

- You should start with the data in the file 'house_sales.csv'.
- Your output should be a data frame named `price_by_rooms`.
- It should include the three columns `bedrooms`, `avg_price`, `var_price`.
- Your answers should be rounded to 1 decimal place.

Task 4

Fit a baseline model to predict the sale price of a house.

1. Fit your model using the data contained in “train.csv”

2. Use “validation.csv” to predict new values based on your model. You must return a dataframe named `base_result`, that includes `house_id` and `price`. The price column must be your predicted values.