CPTR 420 Test 3.

December 12, 2023 1:30 – 3:30

THE SCENARIO

Sinclair Realtors is a real estate company based in the town of Brereton. Sinclair obviously keeps data on their home sales. They have extracted a subset of the data representing sales for January to November 2023. They want to use this data to gather insights and make plans. The data is in a csv file called HOUSING.csv

THE DATA

The data file HOUSING.csv contains data that represents homes for sale on Brereton. The columns are described below:

UnitId – unique identifier for the home
LotFrontage – distance in feet from house to street
LotArea – land area in sq. ft.
BldgType – describes building e.g. duplex, I-family, 2-fanily, etc.
HouseStyle – describes the style e.g. 1-storey, 2-storey, etc.
RoofMatl - roofing material
Exterior – house exterior material
Foundation – type of foundation
Heating - type of heating
CentralAir yes or no
Gr:Bsmt – area above ground:basement area
FullBath – number of full baths
HalfBath – number of half baths
BedroomAbvGr -number of bedrooms above grade/ground
TotRmsAbvGrd - total #number of rooms above grade/ground
Fireplaces – number of fireplaces

GarageType – attached, detached, etc.
GarageArea – square footage of garage
WoodDeck – square footage of deck
OpenPorchSF – square footage of porch
SaleDate – date the house was sold
SalePrice – price obtained
PriceCategory – price categorized as A, B, C, D, E. see below for the ranges

Price Category Ranges

>0 and <200000	E
>=200000 and < 300000	D
>=300000 and <400000	C
>= 400000 and < 550000	B
>= 550000	A

INSTRUCTIONS :

These exercises must be done using r. Your script must be well documented. You may use the bullet points for documentation

Note that the data must be preprocessed before use for example: **(30 points)**

- The UnitId, Fireplace, and Heating columns are not necessary, remove them
- Some observations are missing SalePrice. Remove these observations
- Frontage values should be numeric. Remove the ft. and make the result numeric (beware of space)
- Some observations are missing frontage values. This should be replaced with the minimum legal lot frontage i.e. 20 ft

- All other areas are in square feet, some of the columns have sq. ft. added to the data. The sq. ft. must be removed so that only the numeric value is left. (Beware of spaces)
- Above ground area and basement area are put in one column called Gr_Bsmt, these values should be separated into 2 columns
- The date sold must be a date in r format i.e. YYYY-MM-DD
- You may do any other preprocessing you deem necessary. Comment accordingly.

List any assumptions you make.

Sinclair Management wants to do some exploratory analysis on the data and get answers to the following questions (some in chart format): (30 points)

- How many of each style of house were sold? (sort by most to least)
- What are the different types of building and average Lot Area for each type?
- How does price vary with total area i.e. above ground + basement + woodDeck + OpenPorch square footage? Prepare a chart. The chart should also use color to indicate the SalesCategory of the homes. Add appropriate titles labels and legends.
- What is the sale trend throughout the year (in terms of number of homes sold)? Prepare a well labeled line or bar chart

Price Prediction **(30 points)**

Jason Trent, a resident of Brereton, is moving to Japan and has asked Sinclair Realtors to manage the sale of his home. He already has an official evaluation from the city but Sinclair wants to see how it compares to prices received for other homes that Sinclair sold this year.

He owns a house that has the following properties:

LotFrontage	98 ft.
LotArea	11478 sq. ft.
Above ground sq ft	1704 sq. ft.
Basement sq. ft.	1704 sq. ft.
TotalBaths	2
BedroomAbvGrd	3
TotRmsAbvGrd	7

GarageArea	772 sq. ft.
OpenPorchSF	50 sq. ft.

Based on these attributes, Jeremy wants to determine the price category for his house.

- Using KNN classification method to prepare a model
  - Note that some columns will have to be removed and data might have to be normalized
  - Use a random 75/25 train/test split and k that is square root of the training set size.
- Evaluate your model i.e.
  - How many predictions are correct
  - How many predictions are incorrect
  - what percentage of the predictions are correct.
- Use the model to help predict the **price category** for Jason’s home.