week4_assignment

Kohei Nishitani

2024-02-06

Section 1: Description of the data

This dataset originates from King County, WA, USA (https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data), contains a wide range of real estate sales data which is comprehensive details on homes sold within the county, featuring over 20 columns that captures various aspects of the propertes listed. The primary objective of utilizing this dataset is to analyze the real estate market trends in King County, focusing on factors that influence property prices. By examining attributes such as the size of the living space, number of bedrooms and bathrooms, and additional features like waterfront views and grade of the house, we can identify patterns and insights that are crucial for buyers, sellers, and investors in the housing market.

Section 2: Reading the data into R.

```
# set working directory
setwd("D:/Workspace/McDaniel-Repository/515/week4")

# read csv and assign this to data variable, and specify the header
data <- read.csv('kc_house_data.csv',header=TRUE)
```

Section 3: Clean the data

```
cln_data <- data %>%
  mutate(
  # these variables should be factor not number data type
  id = as.factor(id),
  waterfront = as.factor(waterfront),
  condition = as.factor(condition),
  view = as.factor(view),
  zipcode = as.factor(zipcode),
  yr_renovated = if_else(yr_renovated == 0, NA_integer_, yr_renovated),
  bedrooms = as.numeric(bedrooms),
  sqft_living = as.numeric(sqft_living),
  sqft_living15 = as.numeric(sqft_living15),
  sqft_lot = as.numeric(sqft_lot),
  sqft_lot15 = as.numeric(sqft_lot15),
  sqft_above = as.numeric(sqft_above),
  sqft_basement = as.numeric(sqft_basement),
  yr_built = as.numeric(yr_built),

  ) %>%
  # at least less than 100 years old houses
  filter(yr_built > 1924) %>%
  # select few relevant columns for convenience
  select(price, bedrooms, bathrooms, waterfront, grade)
```

Section 4: Characteristics of the data

This real estate sales data contains rows and variables like (id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long, sqft_living15, sqft_lot15).
This dataframe has 21613 rows and 21 columns. The names of the columns and a brief description of each are in the table below:

Column ID, Names and Descriptions

Column Number	Column Name	Description
1	price	Price of each home sold
2	bedrooms	# of bedrooms
3	bathrooms	# of bathrooms
4	waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
5	grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design

Section 5 Subset and Summary

```
##   Avg_Price Min_Price Max_Price Avg_Bedrooms Min_Bedrooms Max_Bedrooms
## 1  534435.3     75000   7062500     3.393391            0           33
##   Avg_Grade Min_Grade Max_Grade
## 1  7.720437         1        13
```