Thursday Exam 3 - STAT 252

### Description of the Data & Transformation

The code provided transforms the Sacramento housing dataset using tidyverse tools to prepare it for analysis or modeling. Specifically, it uses the mutate() function to add new variables that capture both geographic and numerical context. The first transformation creates a new variable, RegionGroup, which classifies each city in the dataset into one of five geographic clusters based on its location in or around the Sacramento area. These groups—such as “Sacramento Core,” “Roseville Area,” “Auburn & Foothills”, “South & Rural”, and “El Dorado Corridor”—are based on regional proximity and serve to simplify the diversity of neighborhoods into meaningful clusters for analysis. This kind of grouping is especially helpful in modeling and exploratory analysis when regional differences may influence housing prices or property characteristics.

The Sacramento dataset contains real estate transaction information for 932 homes in the Sacramento, California area. Originally compiled by the SpatialKey software team, the dataset includes housing sales reported over a five-day period by the *Sacramento Bee*. It was later cleaned and supplemented using Google to fill in missing or inaccurate location data. The dataset is structured as a tibble (a modern type of data frame in R) and includes key variables such as city, zip code, number of beds and baths, sqft (square footage), type of property (e.g., Residential or Condo), and price. It also contains geographic coordinates (latitude and longitude) that allow for spatial analysis.

## Transformation of Data

## Models

model\_1 <- lm(price ~ sqft, data = sacramento\_df)  
model\_2 <- lm(price ~ sqft + beds, data = sacramento\_df)  
model\_3 <- lm(price ~ sqft + beds + baths, data = sacramento\_df)  
model\_4 <- lm(price ~ sqft + RegionGroup, data = sacramento\_df)  
model\_5 <- lm(price ~ sqft \* RegionGroup, data = sacramento\_df)

## Model 1

summary(model\_1)

Call:  
lm(formula = price ~ sqft, data = sacramento\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-231889 -54717 -11822 38993 600141   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 13859.393 6948.714 1.995 0.0464 \*   
sqft 138.546 3.796 36.495 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 84130 on 930 degrees of freedom  
Multiple R-squared: 0.5888, Adjusted R-squared: 0.5884   
F-statistic: 1332 on 1 and 930 DF, p-value: < 2.2e-16

## Model 2

summary(model\_2)

Call:  
lm(formula = price ~ sqft + beds, data = sacramento\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-200938 -52196 -9148 37268 620281   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 60850.544 10424.530 5.837 7.33e-09 \*\*\*  
sqft 161.336 5.339 30.218 < 2e-16 \*\*\*  
beds -26035.145 4366.612 -5.962 3.53e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 82610 on 929 degrees of freedom  
Multiple R-squared: 0.604, Adjusted R-squared: 0.6031   
F-statistic: 708.5 on 2 and 929 DF, p-value: < 2.2e-16

## Model 3

summary(model\_3)

Call:  
lm(formula = price ~ sqft + beds + baths, data = sacramento\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-202306 -52808 -9260 37341 610425   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 58481.747 10755.143 5.438 6.91e-08 \*\*\*  
sqft 158.167 6.403 24.702 < 2e-16 \*\*\*  
beds -27035.285 4507.272 -5.998 2.86e-09 \*\*\*  
baths 5342.591 5958.426 0.897 0.37   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 82610 on 928 degrees of freedom  
Multiple R-squared: 0.6043, Adjusted R-squared: 0.6031   
F-statistic: 472.5 on 3 and 928 DF, p-value: < 2.2e-16

## Model 4

summary(model\_4)  
Call:  
lm(formula = price ~ sqft + RegionGroup, data = sacramento\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-194463 -48130 -11590 35790 569866   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 1.891e+04 6.593e+03 2.869 0.00421\*\*  
sqft 1.268e+02 3.776e+00 33.575 < 2e-16\*\*\*  
RegionGroupGroup 2 - Roseville Area 4.432e+04 7.911e+03 5.602 2.79e-08\*\*\*  
RegionGroupGroup 3 - El Dorado Corridor 9.647e+04 1.055e+04 9.147 < 2e-16\*\*\*  
RegionGroupGroup 4 - Auburn & Foothills 1.070e+05 2.541e+04 4.210 2.80e-05\*\*\*  
RegionGroupGroup 5 - South & Rural 2.150e+04 1.406e+04 1.529 0.12661  
  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 79560 on 926 degrees of freedom  
Multiple R-squared: 0.6338, Adjusted R-squared: 0.6319   
F-statistic: 320.6 on 5 and 926 DF, p-value: < 2.2e-16

## Model 5

summary(model\_5)

Call:  
lm(formula = price ~ sqft \* RegionGroup, data = sacramento\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-205411 -48032 -11800 34921 569552   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) 30127.922 7940.240 3.794  
sqft 119.564 4.736 25.248  
RegionGroupGroup 2 - Roseville Area 34770.167 21341.692 1.629  
RegionGroupGroup 3 - El Dorado Corridor 38261.435 27074.270 1.413  
RegionGroupGroup 4 - Auburn & Foothills 154963.526 59190.862 2.618  
RegionGroupGroup 5 - South & Rural -70973.072 32761.265 -2.166  
sqft:RegionGroupGroup 2 - Roseville Area 6.390 10.388 0.615  
sqft:RegionGroupGroup 3 - El Dorado Corridor 28.444 11.821 2.406  
sqft:RegionGroupGroup 4 - Auburn & Foothills -21.796 26.429 -0.825  
sqft:RegionGroupGroup 5 - South & Rural 48.461 15.338 3.159  
 Pr(>|t|)   
(Intercept) 0.000158 \*\*\*  
sqft < 2e-16 \*\*\*  
RegionGroupGroup 2 - Roseville Area 0.103609   
RegionGroupGroup 3 - El Dorado Corridor 0.157934   
RegionGroupGroup 4 - Auburn & Foothills 0.008989 \*\*   
RegionGroupGroup 5 - South & Rural 0.030538 \*   
sqft:RegionGroupGroup 2 - Roseville Area 0.538611   
sqft:RegionGroupGroup 3 - El Dorado Corridor 0.016316 \*   
sqft:RegionGroupGroup 4 - Auburn & Foothills 0.409753   
sqft:RegionGroupGroup 5 - South & Rural 0.001632 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 79070 on 922 degrees of freedom  
Multiple R-squared: 0.6399, Adjusted R-squared: 0.6363   
F-statistic: 182 on 9 and 922 DF, p-value: < 2.2e-16