**6304 Project 3**

**Kansas City Housing Data Analysis Project**

**Objectives**

The goal of this project is to analyze the Kansas City housing data to understand various factors affecting house prices and build a predictive model for sale prices. Key objectives include:

1. Cleaning and preprocessing the dataset.

2. Performing exploratory data analysis.

3. Building and evaluating a multiple linear regression model.

4. Making predictions using the regression model.

5. Conducting an Analysis of Variance and interpreting the results.

**Preprocessing**

Data Import and Initial Inspection

1. Import the dataset using any appropriate method.

2. Determine and report the count of zero values in each column.

3. Replace zero values with NA. (Hint: command syntax follows data [data == 0] = NA.)

4. Determine and report the total number of missing values in the dataset.

5. Determine and report the number of missing values for each column.

**Data Cleaning**

1. Remove any columns you see as unnecessary to this assignment such as *id, date, waterfront, view,* and *yr\_renovated*.
2. Remove null values using na.omit().
3. Convert to factors any variables you believe should be factors to properly complete this assignment.

**Creating a Primary Data Set for Analysis**

Create your primary data set for analysis from the clean data created above. This primary data set will have:

1. 500 randomly sampled properties using the numerical portion of your U number.
2. Only properties built in 1950 or later and having 2, 3, or 4 bedrooms.

**Exploratory Data Analysis (EDA)**

Using your primary data set:

1. Display the structure of your primary dataset.
2. Show your primary data set meets each of the two conditions specified in the Creating a Primary Data Set for Analysis section. The unique() command may help with this.
3. Determine the 95% confidence interval for the *price* variable in primary data. Give a clear written interpretation for the confidence interval.
4. Use a hypothesis test determine if the population mean *price* is significantly greater than $650,000. Give a clear written interpretation of the results.

**Regression**

Using your primary data set:

1. Conduct a multiple linear regression with *price* as the dependent variable and *bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, condition, grade, sqft\_above, sqft\_basement, yr\_built, zipcode, lat, long~~, sqft\_living~~*~~, and~~ *~~sqft\_lot~~* as the independent variables. Show the results using the summary() command as well as confidence intervals on beta coefficients. Provide appropriate interpretations of your beta coefficients.
2. Remove variables with non-significant beta coefficients and re-run the model. Do the goodness of fit measures improve with the removal of these variables?
3. Determine whether this model meets the LINE assumptions of regression.
4. Determine whether there appears to be multicollinearity in your second model and which variables seem to be affected.
5. Abhinav needs to predict the *price* for a particular property. Use your second model to generate the prediction assuming the property has the following characteristics. If your model does not include some of these variables do not use them in your prediction.
   1. *bedrooms*: 4
   2. *bathrooms*: 3
   3. *sqft\_living*: 3600
   4. *sqft\_lot*: 250000
   5. *floors*: 2
   6. *condition*: 4
   7. *grade*: 5 (changed to 6)
   8. *sqft\_above*: 2600
   9. *sqft\_basement*: 1000
   10. *yr\_built*: 1989
   11. *zipcode*: 98133
   12. *latitude*: 47.3754
   13. *longitude*: -122.353

**Analysis of Variance**

Conduct three one-way ANOVAs using *price* as the dependent variable and *bedrooms, floors,* and*, condition* as the independent variables. For each independent variable:

1. Show and interpret results of an appropriate test for equality of variances.
2. Conduct the Analysis of Variance and show and interpret standard ANOVA output with the F test on overall results.
3. Conduct a Tukey HSD test, plotting the results of each test, and explaining whether/where there are differences in mean *price* among the different levels of each factor variable. Briefly explain the results shown in the plot. Make sure factor level names can be clearly and completely read on the appropriate axis of your plot.