**Assignment 8: Develop a Data Analysis Project**

**Shristi Dhakal**

**Question 1:**

**Develop a question of your choice that can be addressed by identifying, collecting, and analyzing relevant data. You need to find relevant data by yourself, and describe the data such as the source, attributes, size, how the data were collected, is the dataset sample data or population data? etc. The dataset should have at least six distinct variables (i.e. columns) and a sample size (i.e. rows) of 500 or more. (3 points)**

I chose the Housing dataset on [Kaggle](https://www.kaggle.com/c/house-prices-advanced-regression-techniques) that has two data files - train.csv and test.csv - used for training and testing the models. I will use the data from train.csv in this project as I am only interested to look at the first 1460 houses in Ames, Iowa. I will split this dataset into training and testing to predict the final sale price of a home. This data set has 81 attributes and 1460 rows, and includes attributes such as Lot Frontage, Lot Area, Neighborhood, Condition, House Style, Year Built, Year Remodeled, Foundation, Total Rooms Above Ground, Garage Type, etc. of residential homes in Ames. For this project; however, I am only interested to look at the relationship between the living area above ground represented by the variable "GrLivArea" and the final sale price of a home represented by the variable "SalePrice". Both of these variables have numeric data types, so it is relatively easier to manipulate the data. This is a population data set. The goal here is to see if a linear relationship exixts between these two variables and if it does, to predict the sale price based on the living area above ground.

**Project Objective:** Can we predict sale price based on living area above ground?

**Questions 2 & 3:**

**Perform exploratory data analysis (EDA). Describe the EDA process and result with at least four data visualizations. Explain whether the data is sufficient to answer the question you developed based on EDA result. If it is not sufficient, how did you address the issue? (3 points)**

**Describe any data cleaning or transformations that you perform and why they are motivated by your EDA? (2 point)**

A screenshot of a cell phone

Description automatically generated



A screenshot of a cell phone

Description automatically generated

One of the keys to effective data analysis is selecting variables that're relevant to your project objective. There are a lot of variables in this data set that I will not need to achieve my project goal, so I will only select the variables that I need.

Since I'm interested to know the relationship between living area above ground and the sale price, I select the variables "GrLivArea" and "SalePrice" manually, plus the variable "YearBuilt" just to see if this provides any interesting insight in addition to the other two variables. This ensures no redundancy in the data since now there are only three variables and they are very distinct from each other. The data narrowed down after variables selection will be stored in a new variable "all\_data", which will be used in rest of the project.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The mean of YearBuilt is 1971.27, which means that we have more data for the recent years; the mean is much closer to the maximum value of 2010 than to the minimum value of 1872. Also, we see that sale price of homes range from 34900 dollars to 755000 dollars and the average is 180921.196 dollars, which is much closer to the min value than the max value.

A screenshot of a cell phone

Description automatically generated

Next, we'll see if a linear relationship exists between GrLivArea and SalePrice or not and check if there are any outliers that are skewing some of the statistics.

A close up of a logo

Description automatically generatedA screenshot of a social media post

Description automatically generated

The scatter plot looks more or less linear; however, it appears that there are a couple outliers present towards the bottom right of the plot. If the general pattern is that the sale price of a home increases with the increase in the living area above ground, it is difficult to make sense about what is going on with the two data points at the bottom right because their sale prices are very low given that their respective values for GrLivArea are high.



A close up of text on a white background

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Both GrLivArea and SalePrice have right-tailed histograms with majority values falling towards the left.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The heatmap shows the same pattern as the scatterplot i.e. sale price seems to increase as the living area above ground increases.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The new statistics after removing the outliers is slightly different than the statistics including those data points. This may not make a lot of difference in the analysis, but it is still better to remove exceptions like this for more interpretable results. So, I'll use the new data without the outliers.

The data seems to be sufficient to answer the question "Can we predict sale price of a home based on the living area of the home above ground?" because we can already see in the visualizations above a linear relationship between GrLivArea and SalePrice. Also, there are 1460 rows of data - this appears to be enough to split into training and validation and get at least satisfactory results.

**Question 4:**

**Apply relevant inference or predication methods such as linear regression or K-nearest neighborhood (KNN)to analyze your processed data, and validate the analysis results using cross-validation. Explain the training process, and the loss functions used in the analysis. Using examples (i.e. the values of the loss functions) to explain how the minimal value(s) of the loss function is/are found. (7 points)**

I'll use the linear regression prediction method since all data is numeric. However, the data needs to be normalized before using it in modeling so that values in one variable do not overweight the values in other variable(s) leading to inaccurate results. In this case, the unit in which sale prices are measured (U.S. dollars) is different than the unit in which the living area above ground is measured (square ft.), so we do not want sale prices to outweigh the living area values.

**Normalization:**

Min-max normalization subtracts the min. value in the feature and then divides by the range for each feature. The range is the difference between the original maximum and original minimum. This is a good method for preserving the shape of the original distribution.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA picture containing knife

Description automatically generatedA screenshot of a cell phone

Description automatically generated

0.65 is not a very strong correlation, but it is not bad either. This means that these variables are more related to each other than not.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA picture containing photo, table, kitchen

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

This looks like a good enough model. But we don't know how the data was split, specifically we don't know if the split was random enough. I'll use k-fold cross-validation method to validate this.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

This scatterplot looks a little different than the earlier scatterplot with the 80/20 data split.

A screenshot of a cell phone

Description automatically generated

The accuracy score of the earlier model with the 80/20 split (0.577) is more than the accuracy of this model (0.537), so I'll use the original model.

**Calculating slope of the original model**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA picture containing screenshot

Description automatically generated

The regression slope falls inside the interval of the bootstrapped slopes and sits firmly in the middle of the interval. Thus, this looks like a good enough model to be used for prediction.

**Loss Functions:**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA close up of a white wall

Description automatically generated

The sale prices look very low because they are normalized to be in the range 0-20. In the normalized form, the theta should be around 2-3.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The value of theta that minimizes loss seems to be 4. I had guessed the theta to be somewhere between 2-3 (in the peak of the distribution), but it turned out to be different, which is fine.

**Question 5:**

**Summarize and interpret your results including at least four data visualizations. Provide an evaluation of your approach and discuss any limitations of the methods you used. (2 points)**

I think that I took a pretty reasonable approach to predict the sale prices of homes in Ames based on the living area above ground. I used EDA, 80/20 splits for training and testing, and k-folds cross validation to come up with a good enough Linear Regression Model. The Mean Squared Error (MSE) used to calculate loss function is also one of the simplest and interpretable functions. Plus, I also did the additional Minimum Squared Error loss function to calculate ideal theta value.