Assignment 3 Part-3

Name: P.N.Bhavani

Id: 100437867

Introduction:

An outline of the House Price Dataset exploratory data analysis (EDA) is given in this report. Understanding the structure of the dataset, dealing with missing values, displaying data distributions, and examining correlations between variables are all steps in the EDA process.

Data Structure: The Dataset consists of 81variables(columns) and 1460 entries/observations, from which the primary columns are as follows:

- OverallQual, GrLivArea, TotalBsmtSF, GarageCars, GarageArea, YearBuilt, YearRemodAdd, FullBath, KitchenQual, TotRmsAbvGrd, Neighborhood, LotArea, Fireplaces and Saleprice.

By focusing on these variables in further exploratory data analysis, we can gain insights into the factors that most significantly impact house prices. This will help us understand the relationships and correlations within the data, which is essential for building effective predictive models.

Summary Statistics: For first few columns-

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	\
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	
std	421.610009	42.300571	24.284752	9981.264932	1.382997	
min	1.000000	20.000000	21.000000	1300.000000	1.000000	
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	

	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	\
count	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	...	
mean	5.575342	1971.267808	1984.865753	103.685262	443.639726	...	
std	1.112799	30.202904	20.645407	181.066207	456.098091	...	
min	1.000000	1872.000000	1950.000000	0.000000	0.000000	...	
25%	5.000000	1954.000000	1967.000000	0.000000	0.000000	...	
50%	5.000000	1973.000000	1994.000000	0.000000	383.500000	...	
75%	6.000000	2000.000000	2004.000000	166.000000	712.250000	...	
max	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	...	

	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	\
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	
mean	94.244521	46.660274	21.954110	3.409589	15.060959	
std	125.338794	66.256028	61.119149	29.317331	55.757415	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	25.000000	0.000000	0.000000	0.000000	
75%	168.000000	68.000000	0.000000	0.000000	0.000000	
max	857.000000	547.000000	552.000000	508.000000	480.000000	

	PoolArea	MiscVal	MoSold	YrSold	SalePrice
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	2.758904	43.489041	6.321918	2007.815753	180921.195890
std	40.177307	496.123024	2.703626	1.328095	79442.502883
min	0.000000	0.000000	1.000000	2006.000000	34900.000000
25%	0.000000	0.000000	5.000000	2007.000000	129975.000000
50%	0.000000	0.000000	6.000000	2008.000000	163000.000000
75%	0.000000	0.000000	8.000000	2009.000000	214000.000000
max	738.000000	15500.000000	12.000000	2010.000000	755000.000000

[8 rows x 38 columns]

An overview of the dataset's distribution of different parameters, including Id, MSSUBClass, LotFrontage, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MassVnrArea etc.. is given by these statistics.

Categorical and Numerical variables distribution:

Based on the inputs/values and their usage we can divide columns to Categorical and Numerical variables

Categorical variables: MSSubClass, MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, Electrical, KitchenQual, Functional, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PavedDrive, PoolQC, Fence, GarageCars, MiscFeature, SaleType, SaleCondition, OverallQual, OverallCond, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces

Numerical variables: LotFrontage, LotArea, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, GarageYrBlt, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold, SalePrice

Univariate Analysis:

Plotting bar charts for categorical variables analysis: As there are many variables that are categorical, let us just focus on the primary columns we defined before to focus on for analysis.

Interpretations:

OverallQal: The most common ratings are quality 5 and 6, which show that many homes are either ordinary or marginally better than average.
Rare Ratings: It is rare to find extremely high (10) and low (1, 2) quality ratings.
Market insight: The majority of homes are of a moderate caliber.

GarageCars: According to the chart, the majority of homes (approximately 850) have two-car garages. One-car (400) and three-car (200) garages are also typical in homes. Four-car garages and no garage are uncommon. This suggests that two-car garages are preferred.

FullBath: The "Count Plot of FullBath" bar chart displays the distribution of homes according to the quantity of complete bathrooms. With about 800 occurrences, two complete bathrooms are found in the majority of homes. With slightly more than 600 instances, homes with one complete bathroom come in second. Few homes have three complete bathrooms or none at all. This suggests that most of the houses in the dataset have one or two complete bathrooms.

KitchenQual: The distribution of kitchen quality ratings is displayed in the "Count Plot of KitchenQual" bar chart. With over 700 occurrences, the majority of kitchens receive a typical/average (TA) rating, while good (Gd) receives a rating of about 600. Fair (Fa) kitchens are the least common, occurring less than 50 times, whereas excellent (Ex) kitchens are less frequent, occurring approximately 100 times. This distribution shows that the majority of the dataset's kitchens are standard and of high quality.

TotRmsAbvGrd: The "Count Plot of TotRmsAbvGrd" bar chart displays the distribution of all the rooms above ground. With about 400 instances, six rooms are the most often used number. The dataset's typical dwelling size is indicated by the fact that most homes have five to eight rooms. Fewer homes exhibit a fairly bell-shaped distribution, with either extremely few or numerous rooms.

Neighborhood: The "Count Plot of Neighborhood" bar chart illustrates how homes are distributed among different communities. The neighborhoods "NAmes" and "CollgCr," which have the highest counts, are those with the greatest number of homes in the dataset. While some neighborhoods have fewer occurrences, other neighborhoods, such as "OldTown," "Edwards," and "Somerst," also have noteworthy counts. This aids in our comprehension of the neighborhoods that are most represented in the dataset.

Fireplaces: The "Count Plot of Fireplaces" bar chart displays the distribution of homes according to the quantity of fireplaces. With counts between 700 and 650, respectively, the majority of homes have either one or no fireplace. Very few homes have three fireplaces, while homes with two fireplaces are far less common—about 100 instances. This suggests that the dataset's average fireplace size is either 0 or 1.

Plotting bar charts for numerical variables analysis: As there are many variables that are numerical, let us just focus on the primary columns we defined before to focus on for analysis.

GrLivArea: The histogram displays a distribution that is skewed to the right. The majority of residences are between 1000 and 2500 square feet in size, with the most common being between 1500 and 2000 square feet. Homes larger than three thousand square feet are less prevalent. This suggests that the vast majority of the dataset's residences have moderate living spaces.

TotalBsmtSF: The histogram of the numerical variable "TotalBsmtSF" (Total Basement Square Footage) shows a distribution that is skewed to the right. According to the research, the majority of basements are between 0 and 2000 square feet, with 1000 square feet being the most common size. The exception is for basements larger than 3,000 square feet. This suggests that there are several basements of a respectable size in the dataset.

GarageArea: With a peak of about 500 square feet, the histogram of the numerical variable "GarageArea" reveals that the majority of garage spaces in the sample fall between 200 and 700 square feet. Numerous homes

without garages are indicated by the obvious bump at 0 square feet. Because of the right-skewed distribution, fewer homes have garages larger than 700 square feet. This suggests that the majority of garages in the sample are of a moderate size.

YearBuilt: The distribution of building construction years is displayed by the histogram of the numerical variable "YearBuilt". Over time, the number of buildings being built rises, reaching notable peaks in the years 1950, 1975, and 2000. This points to times when building activity was higher, with the year 2000 seeing the most structures constructed.

YearRemodAdd: The numerical variable "YearRemodAdd" histogram displays the frequency of additions or remodeling over time. Two significant peaks, one around 1950 and the other around 2010, show periods of strong activity. The frequency fluctuates between the 1950s and the 1980s before beginning to rise once more in the 1990s. This implies that there was comparatively less activity in the years between these two periods, with the majority of expansions or remodels taking place during those times.

LotArea: The distribution of the numerical variable "LotArea" is heavily skewed to the right, according to its histogram. Few homes are larger than 50,000 square feet, and the majority have lot areas ranging from 0 to 20,000 square feet. This suggests that while some outliers have noticeably larger lots, lower lot sizes are more prevalent in the sample.

Saleprice: The majority of homes sold were between $100,000 and $200,000, according to the histogram, which displays a distribution that is skewed to the right. As the sale price rises, the frequency falls, suggesting that there are fewer homes in the higher price levels. This implies that there are fewer expensive homes in the dataset and that lower-priced homes are more prevalent.

Bivariate Analysis:

Considering few useful pairs of numerical columns to study the effect on Saleprice(target variable):

In Between numerical vs numerical variables(Scatterplot):

OverallQual, GrLivArea, TotalBsmtSF, GarageCars, GarageArea, YearBuilt, YearRemodAdd, FullBath, KitchenQual, TotRmsAbvGrd, Neighborhood, LotArea, Fireplaces and Saleprice.

- **LotArea vs Saleprice:** For a set of properties, the scatter plot displays the association between LotArea and SalePrice. The majority of data points are concentrated in the lower LotArea (0–50,000) and SalePrice (0–600,000) ranges. This implies that sale prices are typically lower in places with smaller lots. Though they are less frequent, there are a few outliers with bigger lot sizes (up to 200,000) and different sale prices. Overall, the plot shows that higher sale prices are not always associated with greater lot areas.
- **YearBuilt vs Saleprice:** The scatter plot shows the correlation between a home's sale price and the year it was constructed. There is a noticeable rise in sale prices as the years go by, especially after the 2000s. This implies that newer homes typically fetch higher values than older ones.
- **YearRemodAdd vs Saleprice:** The scatter plot displays the correlation between a home's sale price (SalePrice) and the year it was added to or renovated (YearRemodAdd). It shows that homes that have been added to or renovated in more recent years typically sell for more money, with prices for homes added or renovated after 2000 rising significantly. This suggests that the year of additions or remodeling and the sale price of homes are positively correlated.
- **TotalBsmtSF vs Saleprice:** The scatter plot displays the correlation between a home's sale price (SalePrice) and its total basement square footage (TotalBsmtSF). According to the research, there is a positive association between the size of the basement and the house's sale price. With a few exceptions, the majority of data points fall into the ranges of 0 to 2000 square feet for TotalBsmtSF and $0 to $400,000 for SalePrice. This implies that the market value of a home can be greatly impacted by a larger basement.
- **GarageArea vs Saleprice:** The correlation between garage area (measured in square feet) and sale price (measured in dollars) is displayed in the scatter plot. It shows a positive correlation: the sale price tends to increase as the garage area grows. The dispersed data points, however, suggest some variation

in the sale prices for a particular garage location. There are outliers where the sale prices deviate from the average, especially for extremely small and very big garage areas.

In Between numerical vs categorical variables(boxplot):

- **OverallQual vs Saleprice:** The box plot illustrates the correlation between a home's sale price (SalePrice) and overall quality (OverallQual). The median sale price climbs dramatically as the quality grade rises from 1 to 10. Furthermore, there is more variety in the range of sale prices for higher quality ratings. Higher quality ratings are more likely to contain outliers, indicating that some excellent homes fetch remarkably high prices. Overall, the plot shows that a home's quality and sale price are positively correlated.
- **Fireplaces vs Saleprice:** It demonstrates that homes with more fireplaces typically sell for a higher median price. Houses with one fireplace, for instance, have a typical sale price of about $200,000, but those without one have a median sale price of about $130,000. Even higher median prices are found for homes with two or more fireplaces. Furthermore, the dispersion of sale prices increases as the number of fireplaces increases, suggesting that homes with more fireplaces have more price variability. This implies that greater property values might be linked to the existence of fireplaces.
- **GarageCars vs Saleprice:** It shows that the median sale price of homes with additional garage spaces is typically higher. In particular, homes with three or more garage spaces have the highest median sale prices, while those without a garage have the lowest median sale prices. More garage spaces are linked to higher and more variable sale prices, as seen by the dispersion and unpredictability of sale prices increasing with the number of garage spaces. This implies that garage space and home value are positively correlated.

Correlation Heatmap for numerical variables:

For housing data, the correlation coefficients between numerical variable pairs are graphically represented by the correlation heatmap. With values between -1 and 1, the color intensity shows the direction and strength of the correlations, ranging from red (positive correlation) to blue (negative correlation).

- The largest positive correlations are seen between SalePrice and GrLivArea (0.71), GarageArea(0.62), 1stFirSF(0.61) and TotalBsmtSF (0.61). This suggests that higher sale prices are linked to larger living areas, more garage spaces, and larger basements.
- Additionally, there are somewhat favorable associations between SalePrice and YearBuilt (0.52), and YearRemodAdd (0.51), indicating that these criteria also impact higher sale prices.
- The variables MasVnrArea (0.47), GarageYrBlt (0.47), GarageArea (0.47), WoodDeckSF (0.32), OpenPorchSF (0.32), and 2ndFlrSF (0.32) all have smaller positive associations with SalePrice.
- These characteristics are less relevant or negatively connected to sale prices, as seen by the negative correlations found with SalePrice for EnclosedPorch (-0.13), BsmtUnfSF (-0.21), BsmtFinSF2(-0.01), LowQualFinSF(-0.03), MiscVal(-0.02) and Yrsold(-0.03)