PROJECT 1: EDA WITH PANDAS NEW

AMES HOUSING DATASET

PROJECT GOALS

- Given the **AMES Housing Dataset**, the goals of the project were as follows:

I. **Data understanding** of the structure variables and types of data therein.

II. **Data cleaning** through identifying missing values, outliers and inconsistencies

III. **Data visualization** by creating visual representations such as histograms, scatter plots and boxplots.

IV. **Statistical summary** by calculating summary statistics such as mean, median, standard deviations and correlations.

V. **Feature engineering** by creating a new feature(Age) using existing features in the dataset.

VI. **Identifying key patterns and relationships** and explaining them

DATA

- **Data content**

The dataset contains a number of files having two most important ones namely: ***data/ames.csv*** and ***data/data_description.txt*** which were imported into the notebook for use.

- **Data structure**

The data contains 1460 rows and 80 columns. The main focus was on the 'Saleprice', 'totrmsabvgrd,' 'overallcond,' 'yrsold', 'yearbuilt' columns for the analysis needed.

- **Summary of descriptive statics**

	mean	median	Std deviation
SalePrice	180921.196	163000.0	79442.50288
TotRmsAbvGrd	6.5178082	6.0	1.625393290
OverallCond	5.5753424657	5.0	1.112799336712

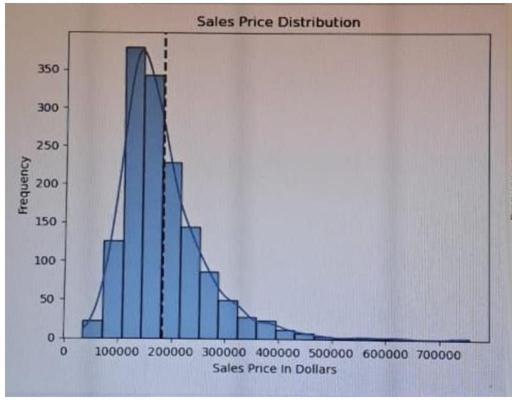
DATA

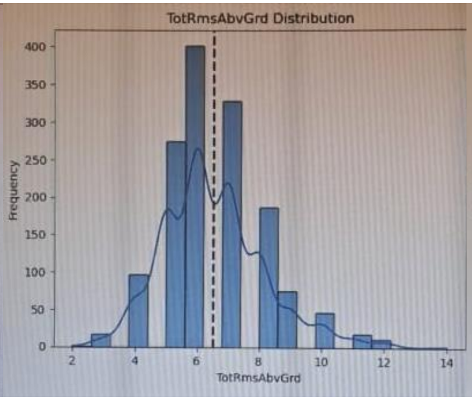
- **Check for correlations**

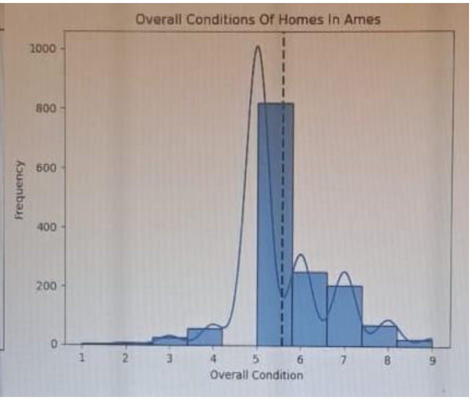
Pearson correlation was used to find the columns with the strongest positive correlation and most negative correlation.

The column with the strongest positive correlation was ***Overallquality*** with **0.790981600.** The column with the most negative correlation was **KitchAbvGrd** with **-0.13590737**.

- **Data visualizations**

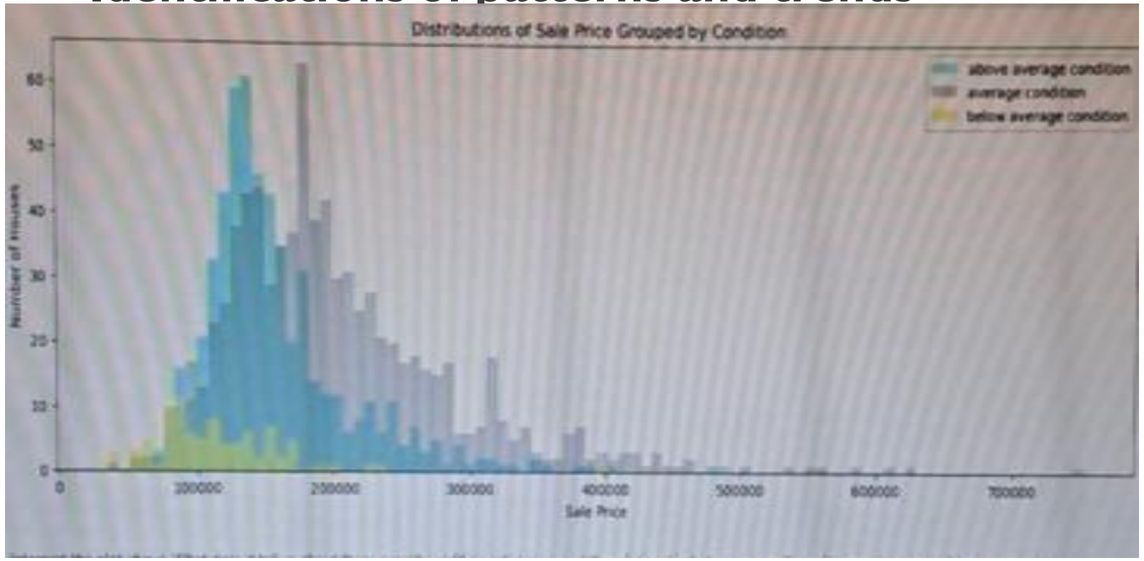


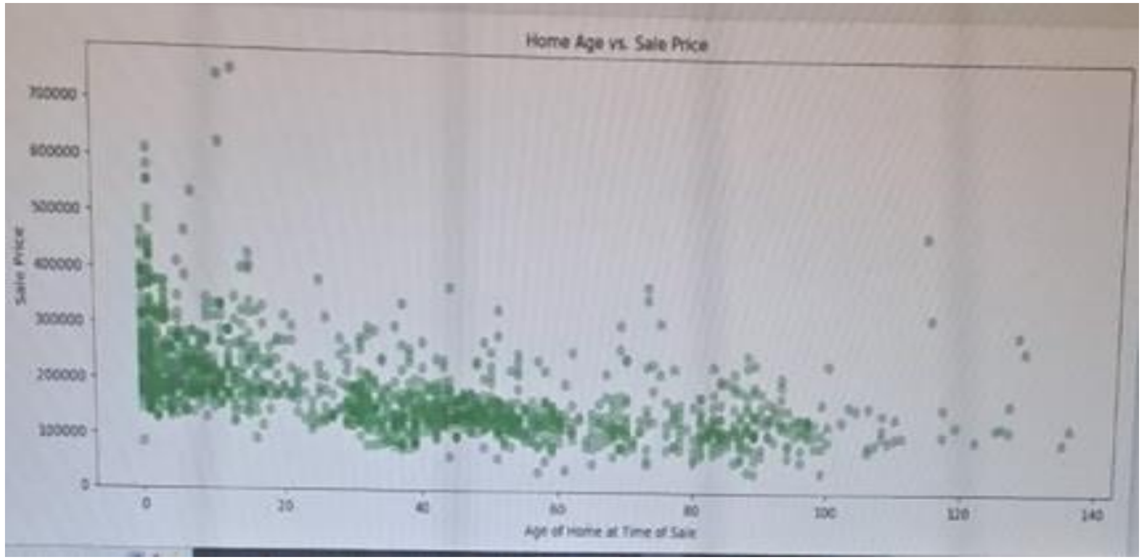




DATA

- **Identifications of patterns and trends**





- **<u>Summary insights</u>**
- The SalePrice is greatly influenced by a number of factors.

METHODS USED

- **Data loading:** The data was loaded into a pandas DataFrame in Python.
- **Data transformation:** The feature engineering was used to create new feature (Age) using existing ones(yearbuilt) and (yearsold) columns.
- **Exploratory visualizations:** Histograms, boxplots and scatter plots were used to explore distributions of several columns, correlations and trends.
- **Statistical summaries:** Descriptive statistics were computed for some numerical columns including mean, median and standard deviation using pandas. Pearson's correlation coefficient was calculated to assess linear relationships between 'SalePrice' and the other columns.

RESULTS

- Key findings from descriptive statistics

The average price of a home in Ames is 163,000 dollars, has an average of 6 rooms above grade and its overall condition is 5.0 which is good.

The prices have a standard deviation of 79442.50 dollars from the mean, rooms above grade by 1.625 and overall condition 1.112.

- Insights from visualizations

The histograms developed showed that; the saleprices are not evenly distributed, having a high concentration at 100,000 to 200,000 ; the number of rooms are mostly around 6,: the overall condition is mainly concentrated past the 5 mark.

- **Key patterns and trends**

the analysis revealed that customers and sales people value overall quality.

RESULTS

- **Relationships between variables**

There is a strong correlation between ‘SalePrice’ and ‘OverallQuality’ but little correlation with ‘KitchAbvGrd’.

The strong correlation suggests customers value OverallQuality over the grade of kitchens in a home as suggested by SalePrices.

- **Insights for further analysis**

Given the strong correlation between the ‘SalePrice’ and the ‘overall quality’ , the owners should focus on improving the overall quality of homes in order to attract big prices and customers.