Title of the Project

Housing: Price Prediction

Submitted by

Name of the Candidate: Muskan Sureka

Internship Batch Number: 34

Supervised by

Name of the SME: Khushboo Garg

Month & Year of Submission: January,2023.

Student's Declaration

I hereby declare that the Project Work with the title (in block letters) “HOUSING: PRICE PREDICTION.”
submitted by me for the project allocated to Internship batch no.: 34 by FLIPROBO TECHNOLOGIES as a part of my internship phase of my PG DIPLOMA COURSE OF DATA SCIENCE AND NEURAL NETWORKS BY DATATRAINED INSITITUE is my original work and has not been submitted earlier to any other Institution for the fulfilment of the requirement for any course of study.

I also declare that no chapter of this manuscript in whole or in part has been incorporated in this report from any earlier work done by others or by me. However, extracts of any literature which has been used for this report has been duly acknowledged providing details of such literature in the references.

Signature: Muskan Sureka

Name: Muskan Sureka

Address: Alipore Residency,
3 Burdwan Road

Place: Kolkata

Internship batch No.: 34

Date: 21/01/2023

ACKNOWLEDGEMENT

The success of the project is the result of hard work and endeavor of not only one but rather numerous people. It is the outcome of the exceptional support of many individuals.

Firstly, I take immense pleasure in thanking DataTrained Institute and Flip Robo Technologies for furnishing me with a chance to conduct a research project by making it a part of the curriculum.

I also express my sincere indebtedness and profound gratitude to my teachers: Dr. Deepika Sharma, Mr. Ravikesh Pandey and my SME: Khushboo Garg.

I would also like to express my deep sense of gratitude to my project SME Khushboo Garg for her continuous guidance and support, which has helped me tremendously to complete this project with the best of my abilities. Without her help, completing this research would not have been possible.

Also, I would like to express my sincere gratitude towards my friends and family for providing me with valuable and significant ideas throughout the course of the project.

I hence express my sincere gratitude and appreciation to all the individuals who have helped in any possible way and contributed in any manner. Their constant support and assistance has been significant and of great value.

INDEX

1. INTRODUCTION

1.1 Introduction

Understanding the House Price Index (HPI) and How It Is Used:

What Is the House Price Index (HPI)?

The House Price Index (HPI) is a broad measure of the movement of single-family property prices in the United States. Aside from serving as an indicator of house price trends, it also functions as an analytical tool for estimating changes in the rates of mortgage defaults, prepayments, and housing affordability.1

KEY TAKEAWAYS

- The House Price Index (HPI) is a broad measure of the movement of single-family house prices in the United States.
- It is published by the Federal Housing Finance Agency (FHFA), using monthly and quarterly data supplied by Fannie Mae and Freddie Mac.
- The HPI is one of many economic indicators that investors use to keep a pulse on broader economic trends and potential shifts in the stock market.

Understanding the House Price Index (HPI)

The HPI is pieced together by the Federal Housing Finance Agency (FHFA), using data supplied by the Federal National Mortgage Association (FNMA), typically known as Fannie Mae, and the Federal Home Loan Mortgage Corp. (FHLMC), commonly known as Freddie Mac.

The HPI is based on transactions involving conventional and conforming mortgages on single-family properties. It is a weighted repeat sales index, measuring average price changes in repeat sales or refinancings on the same properties.

An HPI report is published every quarter, as well as a monthly report. Data is compiled by reviewing mortgages purchased or securitized by Fannie Mae and Freddie Mac.1

How the House Price Index (HPI) Is Used

The HPI is one of many economic indicators that investors use to keep a pulse on broader economic trends and potential shifts in the stock market.

The rise and fall of house prices can have big implications for the economy. Price increases generally create more jobs, stimulate confidence, and prompt higher consumer spending. This paves the way for greater aggregate demand, boosting gross domestic product (GDP) and overall economic growth.

When prices fall, the opposite tends to happen. Consumer confidence is eroded and the companies profiting from the demand for real estate lay off staff. This can sometimes trigger an economic recession.

The House Price Index (HPI) vs. the S&P CoreLogic Case-Shiller Home Price Indexes

The HPI is one of many trackers of home prices. Some of the most well-known alternatives are the S&P CoreLogic Case-Shiller Home Price indexes.

These indexes utilize different data and measuring techniques and, therefore, produce varying results. For example, the HPI weights all homes equally, while the S&P CoreLogic Case-Shiller Home Price indexes are value-weighted.

Moreover, while the Case-Shiller indexes only use purchase prices, the all-transactions HPI includes refinancing appraisals as well. The HPI also provides wider coverage.

Fannie Mae and Freddie Mac
As already mentioned, the HPI measures average price changes for homes that are sold or refinanced by looking at mortgages purchased or secured by Fannie Mae or Freddie Mac. That means loans and mortgages from other sources, such as the United States Department of Veterans Affairs and the Federal Housing Administration (FHA), do not feature in its data.

Fannie Mae
Fannie Mae is a government-sponsored enterprise (GSE) that is listed on the public market yet operates under a congressional charter. The company's goal is to keep mortgage markets liquid. It does this by purchasing and guaranteeing mortgages from the actual lenders, such as credit unions, and local and national banks—Fannie Mae cannot originate loans directly.6

The FNMA expands the liquidity of mortgage markets and facilitates homeownership for low-, moderate-, and middle-income Americans by creating a secondary market. Fannie Mae was created in 1938 during the Great Depression as part of the New Deal.

Freddie Mac
Like Fannie Mae, Freddie Mac, or the FHLMC, is also a GSE. It purchases, guarantees, and securitizes mortgages to form mortgage-backed securities (MBS). It then issues liquid MBS that generally carry a credit rating close to that of U.S. Treasuries.

Given its connection with the U.S. government, Freddie Mac can borrow money at interest rates that are generally lower than those available to other financial institutions.

How Do You Tell if a House Is a Good Price?
To determine if a house is a good price, you can check the sale prices of recently sold properties in the neighborhood, compare the price with other properties for sale in the market, speak with a real estate agent, and consider the appreciation value.

Should I Offer the Full Asking Price on a House?
Knowing whether or not you should offer the full asking price on a house will come down to a few factors. One of the main factors is if the property being sold is in a buyer's market or a seller's market. If it is a seller's market, you may have to offer the full asking price or above, whereas, in a buyer's market, you may be able to offer a lower price. If you need to offer the full asking price or more, it is generally recommended to offer 1% to 3% more.

What Brings Down the Value of a House?
Many factors bring down the value of a house, such as any new planned construction in the area that would be seen as less than desirable, such as a highway. Foreclosures in the neighborhood would bring down prices as well as the increased likelihood of natural disasters in the area or a greater impact due to climate change. Even rising interest rates can bring down the value of a house, as the increase in mortgage rates makes homes more expensive, which reduces the demand.

1.2 Housing Market Predictions for 2023: Will home prices drop?

High mortgage rates have put some much-needed pressure on the housing market in recent months after home prices hit record highs across the nation. But as mortgage rates have shown some decline lately, many economists are mixed about whether home prices will continue to decelerate through 2023—or crash.

The nation's overall housing supply remains limited, as those who purchased homes in recent years at extremely low mortgage rates are staying put. This tight inventory has kept prices from really dropping off, making homes still unaffordable for many, especially first-time homebuyers.

Even though home prices remain high year-over-year (YOY), they're not as eye-popping as they were in early 2022. How far home prices dip in 2023 will likely depend on where mortgage rates go.

Housing Market Forecast for 2023

As we enter 2023, housing experts maintain a watchful eye on the economy, which is still being pulled in all directions by high inflation, steep interest rates, ongoing geopolitical uncertainties and recession fears, to name a few.
After a couple of red-hot years for the housing market, there are indicators a correction is underway—but it's been slow-going. Mortgage rates are still hovering around double what they were a year ago. And nationwide home prices are still increasing on a monthly basis despite a decline in total sales. This continues to make it harder for many homebuyers to access affordable housing.

The median existing-home sales price was up 3.5% to $370,700 in November compared to a year ago, according to the National Association of Realtors (NAR). It was the 129th consecutive month of YOY price increases—a record streak—even though home prices have fallen from their record high of $413,800 in June.

Still, higher housing costs have taken a toll on home shoppers as mortgage applications are at their lowest level in over 25 years, according to the Mortgage Bankers Association (MBA).

And the total existing-home sales dropped 7.7% from October to November, marking the tenth consecutive month of declining sales, and down 35% from a year ago.

Because of this, some experts say the housing market has reached its bottom already.

“It seems we have already reached the bottom of the low home sales activity,” says Nadia Evangelou, senior economist and director of forecasting for the NAR. “And with mortgage rates stabilizing near 6%, we expect the housing market to turn around in 2023. . .and rebound in 2024.”

Housing Inventory Predictions for 2023

Low housing inventory has been a challenge since the 2008 housing crash, when the construction of new homes plummeted. And it hasn’t fully recovered.

Housing supply that remains near historic lows has held up demand compared to other downturns, consequently sustaining higher home prices.

“For most of this year, prospective home buyers have faced the dual challenges of elevated mortgage rates and limited housing inventory,” said NAR’s president Kenny Parcell, in a recent report.

At the current sales pace, inventory is at a 3.3-month supply, according to NAR.

“[This] is about half of what we’d like to see normally,” says Rick Sharga, executive vice president of market intelligence at ATTOM Data. “And we still have pent-up demands based on demographic trends.”

Housing inventory in November remained flat compared to October but was up from 2.1 months a year ago, according to NAR.

In the meantime, the ongoing slowdown in new construction continues to squeeze the already limited housing supply. Single-family construction starts and applications for building permits in November were down 4.1% and 7.1%, respectively, from the previous month, according to the U.S. Census Bureau and the U.S. Department of Housing and Urban Development.

Data from within the construction realm aligns with these figures, with builder confidence declining for the 12th consecutive month, marking the lowest confidence reading (outside the Spring 2020 Covid-19 pandemic) since mid-2012, according to the latest National Association of Home Builders (NAHB)/Wells Fargo Housing Market Index (HMI) report.

When Will the Housing Market Crash?

There are mixed signals from economists about if and when the housing market will crash, or if it will simply correct itself from the double-digit percentage jumps seen in home prices the past few years.

“We’re estimating about a 5% drop nationally,” says Sharga. “Some markets, believe it or not, will probably see prices continue to increase.”

Other experts point out that today’s homeowners stand on much more secure footing than those coming out of the 2008 financial crisis, so the likelihood of a housing market crash is low.

“Homeowner equity is at the highest level it’s been in the past several decades, so homeowners have a lot of value in their home,” says Nicole Bachaud, an economist at Zillow.

Bachaud also notes that mortgage products have become less risky.

“There are a lot more regulations and restrictions in the mortgage market that make it a lot stronger, and less volatile and less risky, than it was in the market after 2008,” she says.

In a housing market crash, you would typically see a 20% to 30% drop in home prices and a decline in home sales—far more than what’s currently happening. Another crash symptom that’s been missing is a jump in foreclosure activity.
“I think we’re more likely to see the market cool, rather than crash,” Sharga says.

Are a Lot of Foreclosures Coming?

Following a steady rise in foreclosures that resulted after the expiration of the Covid-19 foreclosure moratorium in September 2021, foreclosures may have hit their peak, according to ATTOM Data Solutions, a leading curator of real estate data. Though foreclosure starts are up 57% from a year ago, they were down roughly 5% between October and November 2022.

“Foreclosure starts in November nearly doubled from last year’s numbers, but are still just above 80 percent of pre-pandemic levels,” said Rick Sharga, executive vice president of market intelligence at ATTOM, in a report. “We may continue to see below-normal foreclosure activity, since unemployment rates are still very low, and mortgage delinquency rates are lower than historical averages.”

A key difference now compared to the last housing crisis is that many homeowners, and even those struggling to make payments, have had a large boost to their home values in recent years. That means they still have equity in their homes and are not underwater—when you owe more than the house is worth.

"(E)ven as the foreclosure moratorium was lifted…we didn't see a huge flood of foreclosures because people have so much equity," says Bachaud.

When Should I Buy a Home in 2023?

Buying a house—in any market—is a highly personal decision. Because homes represent the largest single purchase most people will make in their lifetime, it's crucial to be in a solid financial position before diving in.
Use a mortgage calculator to estimate your monthly housing costs based on your down payment and interest rate.
Trying to predict what might happen next year is not the best homebuying strategy. "Buyers sitting on the sidelines today in anticipation of lower prices tomorrow may end up disappointed," says Neda Navab, president of the U.S. region at Compass, a real estate tech company.

Navab expects home prices in the hotter markets during the past few years to decrease somewhat, but she doesn't expect a widespread, national price decline like what followed the 2008 financial crisis.

Instead of waiting for much lower prices, experts suggest buying a home based on your budget and needs. If you find a home you love in an area you love, and it also fits your budget, then chances are it might be right for you. However, if you make too many sacrifices just to get a house, you may end up with buyer's remorse, potentially forcing you to offload the house.

Tips for Buying in Today's Housing Market

Start with a budget and stick with it. Even with a slight uptick in the number of homes for sale, buyers are still facing elevated prices and mortgage rates nearing 7%.

"The biggest thing right now is the disconnect between buyers and sellers," says Rita

Tayenaka, owner of Orange County, California-based Coast to Canyon brokerage. "Buyers want to lowball, and sellers want last year's price."

While buyers are getting a bit more breathing room now, they should keep in mind that it's still a seller's market while they consider their options.

Tips for Selling in Today's Housing Market

The first step for a successful sale is to find a listing agent who knows the area and comes highly recommended. A good agent will work closely with you to price your home competitively while fielding questions and offers from prospective buyers.

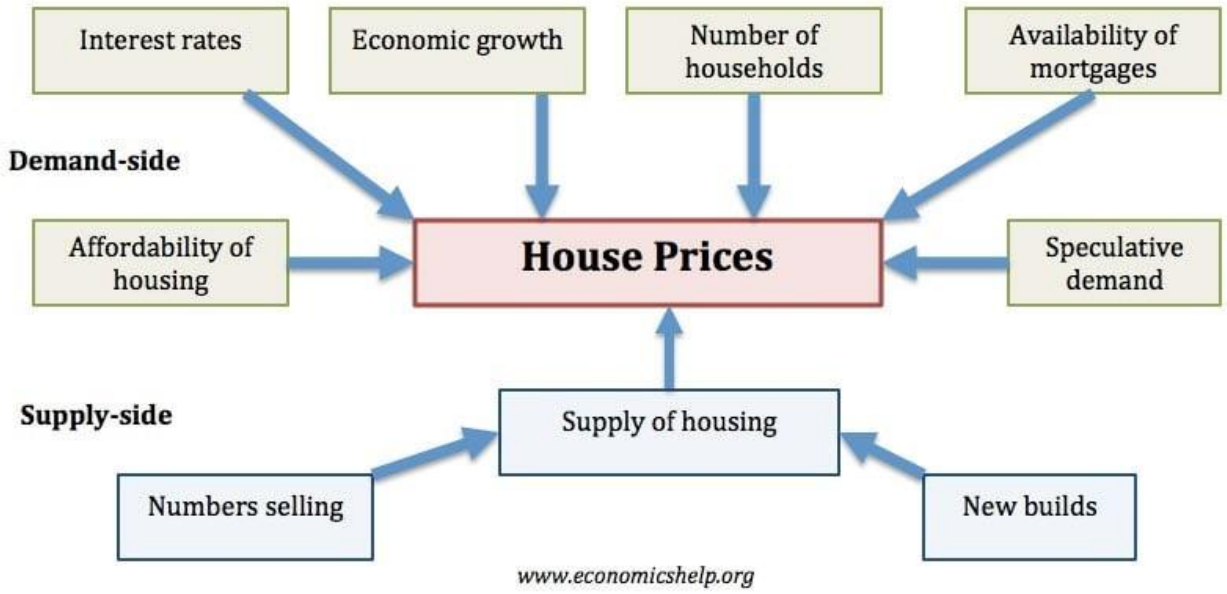
Tayenaka points to the outsize number of homes falling out of escrow recently as a cautionary tale for sellers who continue to demand 2021 prices. "Everyone thinks their house is special," she says.

Even though the market may still be tipped in your favor, it's in your best interest to present your home in the best possible light. Not everyone has cash dedicated to renovations and repairs, but a little sweat equity can go a long way. The first step is to declutter, organize and clean. Even if your home is outdated, a clean space gives buyers a chance to envision the house's potential.

1.3 Factos that affect the housing market

The housing market is influenced by the state of the economy, interest rates, real income and changes in the size of the population. As well as these demand-side factors, house prices will be determined by available supply. With periods of rising demand and limited supply, we will see rising house prices, rising rents and increased risk of homelessness.

Factors determining house prices

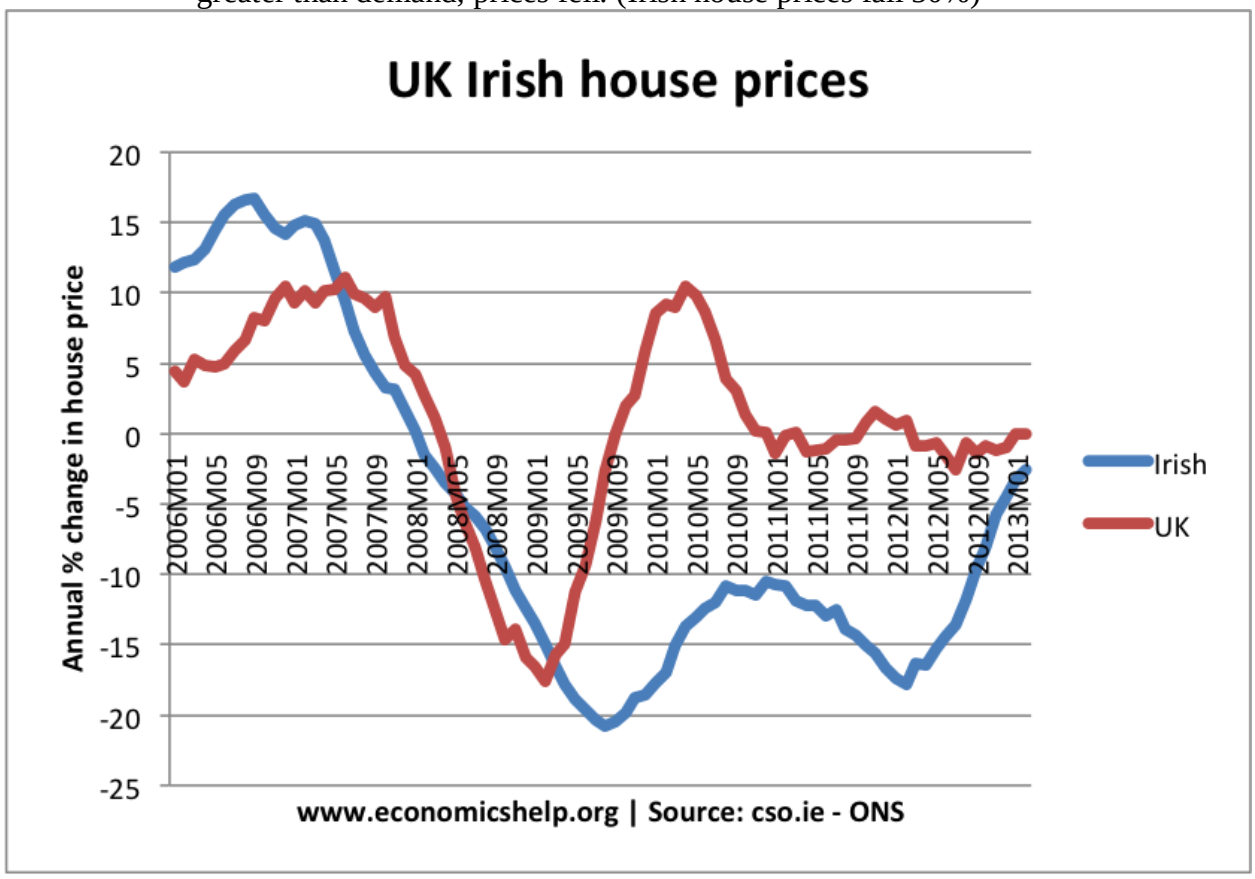


Main factors that affect the housing market

- **Economic growth.** Demand for housing is dependent upon income. With higher economic growth and rising incomes, people will be able to spend more on houses; this will increase demand and push up prices. In fact, demand for housing is often noted to be income elastic (luxury good); rising incomes leading to a bigger % of income being spent on houses. Similarly, in a recession, falling incomes will mean people can't afford to buy and those who lose their job may fall behind on their mortgage payments and end up with their home repossessed.
- **Unemployment.** Related to economic growth is unemployment. When unemployment is rising, fewer people will be able to afford a house. But, even the fear of unemployment may discourage people from entering the property market.
- **Interest rates.** Interest rates affect the cost of monthly mortgage payments. A period of high-interest rates will increase cost of mortgage payments and will cause lower demand for buying a house. High-interest rates make renting relatively more attractive compared to buying. Interest rates have a bigger effect if homeowners have large variable mortgages. For example, in 1990-92, the sharp rise in interest rates caused a very steep fall in UK house prices because many homeowners couldn't afford the rise in interest rates.
- **Consumer confidence**. Confidence is important for determining whether people want to take the risk of taking out a mortgage. In particular expectations towards

the housing market is important; if people fear house prices could fall, people will defer buying.

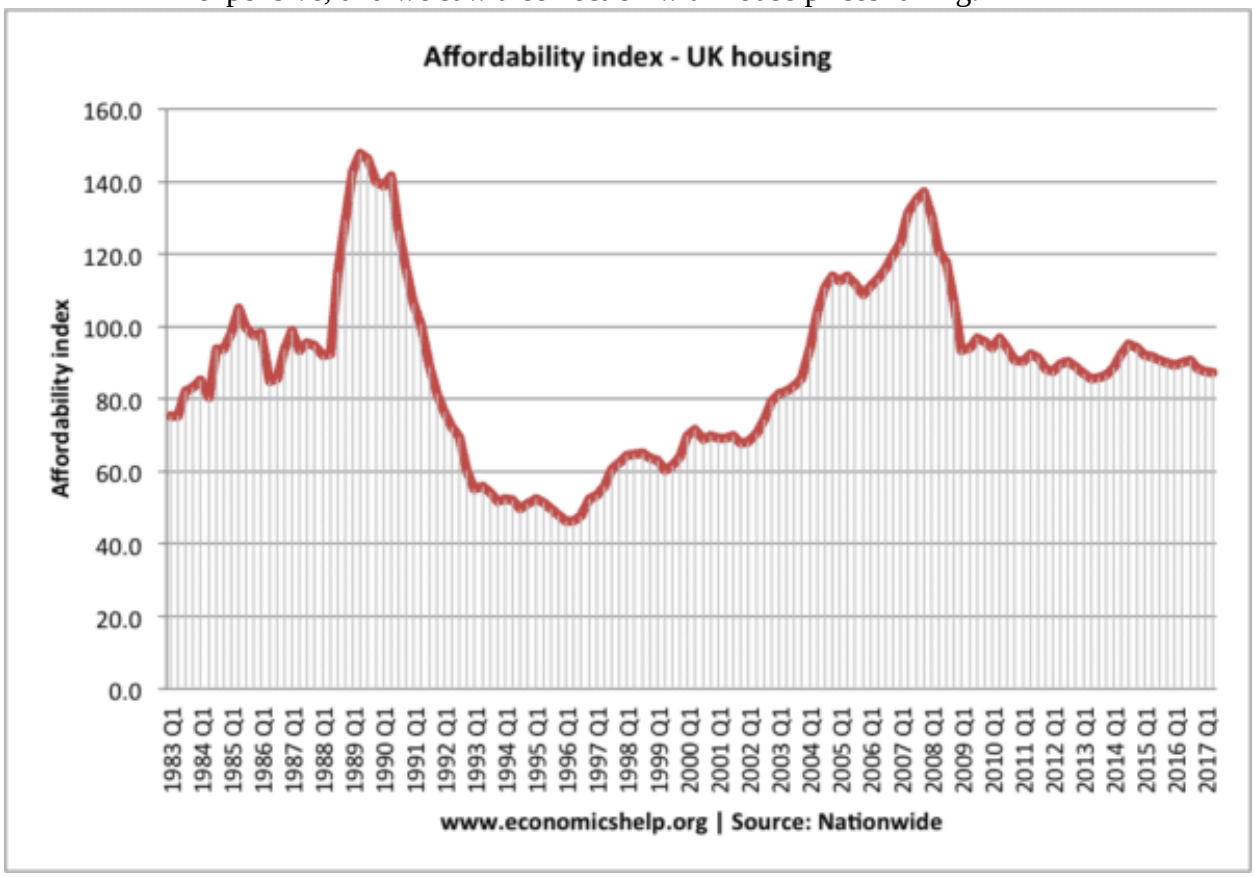
- **Mortgage availability**. In the boom years of 1996-2006, many banks were very keen to lend mortgages. They allowed people to borrow large income multiples (e.g. five times income). Also, banks required very low deposits (e.g. 100% mortgages). This ease of getting a mortgage meant that demand for housing increased as more people were now able to buy. However, since the credit crunch of 2007, banks and building societies struggled to raise funds for lending on the money markets. Therefore, they have tightened their lending criteria requiring a bigger deposit to buy a house. This has reduced the availability of mortgages and demand fell.
- **Supply**. A shortage of supply pushes up prices. Excess supply will cause prices to fall. For example, in the Irish property boom of 1996-2006, an estimated 700,000 new houses were built. When the property market collapsed, the market was left with a fundamental oversupply. Vacancy rates reached 15%, and with supply greater than demand, prices fell. (Irish house prices fall 50%)



By contrast, in the UK, housing supply fell behind demand. With a shortage, UK house prices didn't fall as much as in Ireland and soon recovered – despite the ongoing credit crunch.
The supply of housing depends on existing stock and new house builds. Supply of housing tends to be quite inelastic because to get planning permission and build houses is a time-consuming process. Periods of rising house prices may not cause an equivalent rise in supply, especially in countries like the UK, with limited land for home-building.

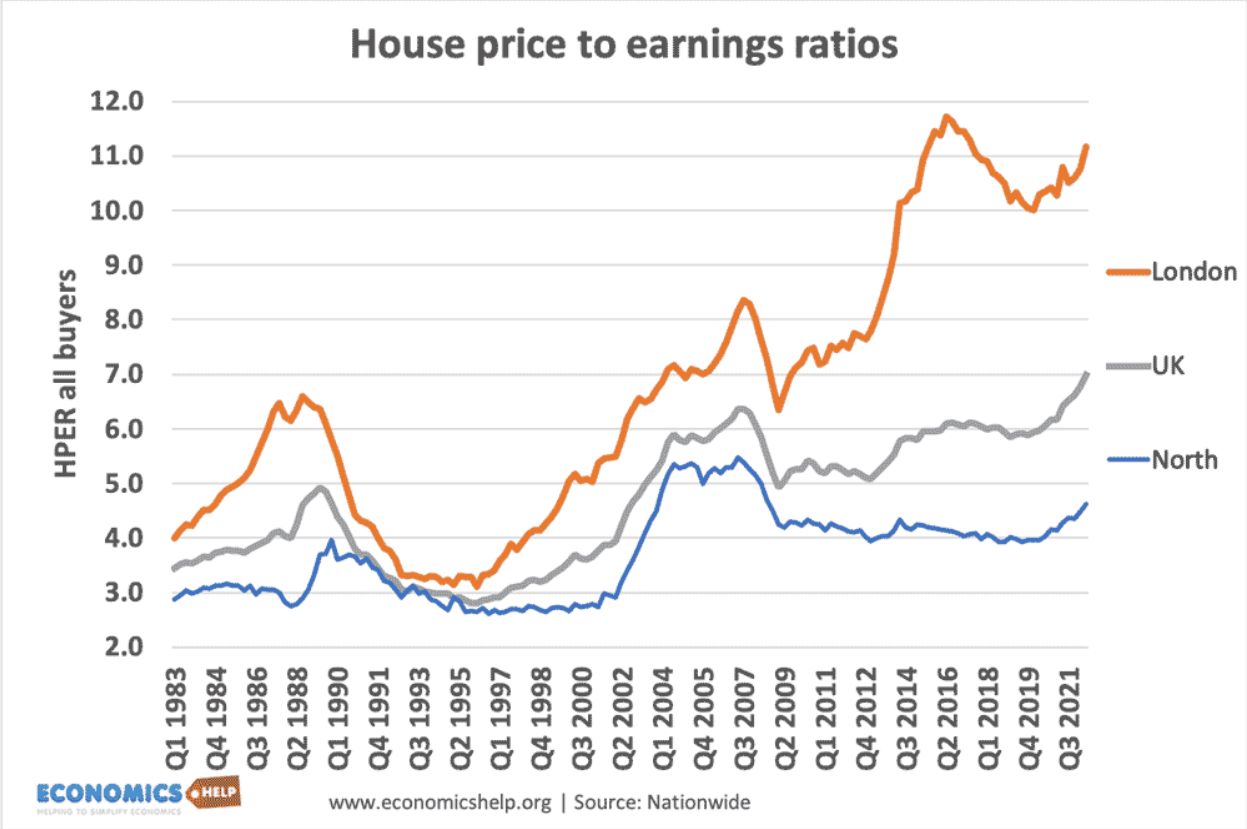
- **Affordability/house prices to earnings.** The ratio of house prices to earnings influences the demand. As house prices rise relative to income, you would expect fewer people to be able to afford. For example, in the 2007 boom, the ratio of

house prices to income rose to 5. At this level, house prices were relatively expensive, and we saw a correction with house prices falling.



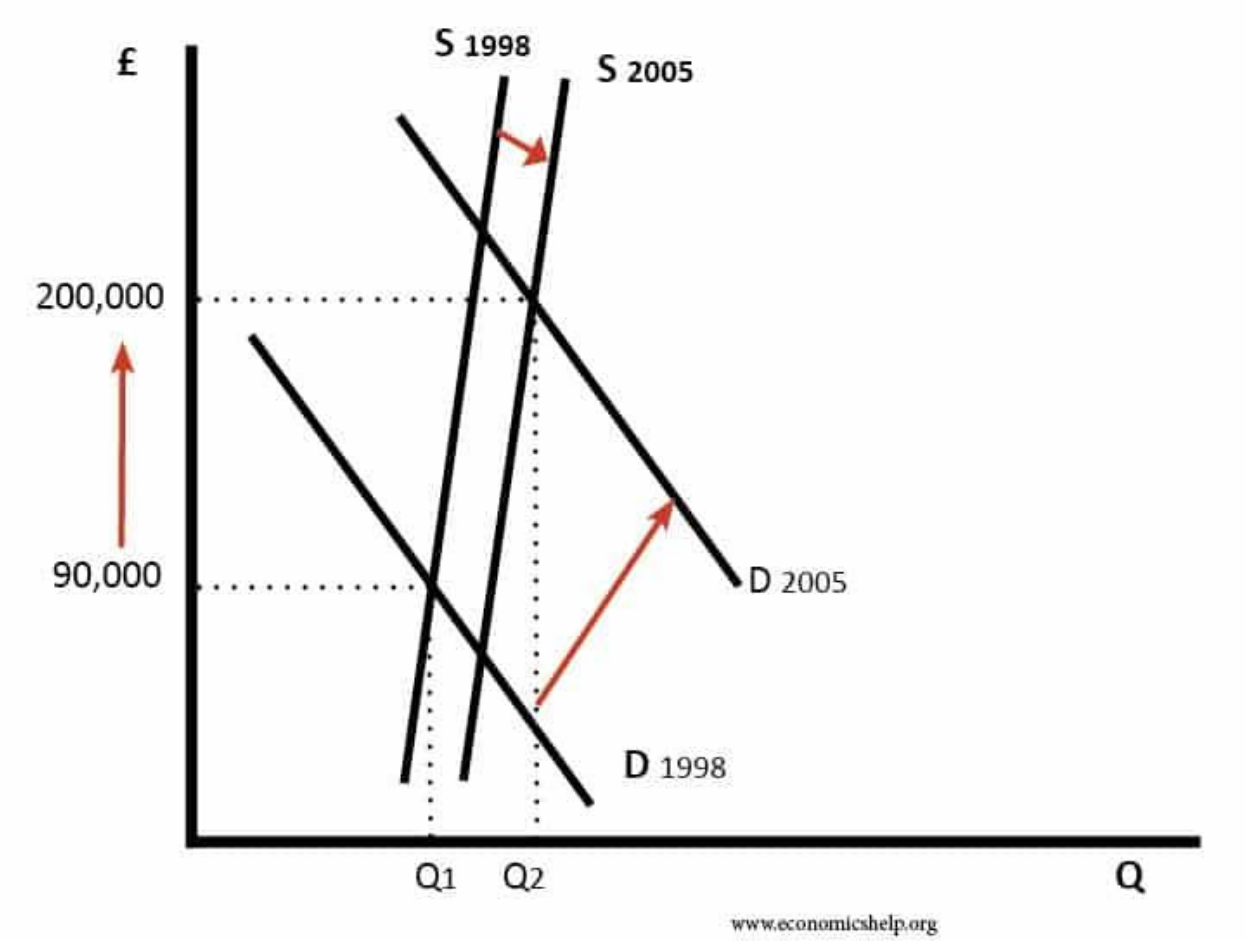
Another way of looking at the affordability of housing is to look at the percentage of take-home pay that is spent on mortgages. This takes into account both house prices, but mainly interest rates and the cost of monthly mortgage payments. In late 1989, we see housing become very unaffordable because of rising interest rates. This caused a sharp fall in prices in 1990-92.

- **Geographical factors.** Many housing markets are highly geographical. For example, national house prices may be falling, but some areas (e.g. London, Oxford) may still see rising prices. Desirable areas can buck market trends as demand is high, and supply limited. For example, houses near good schools or a good rail link may have a significant premium to other areas.



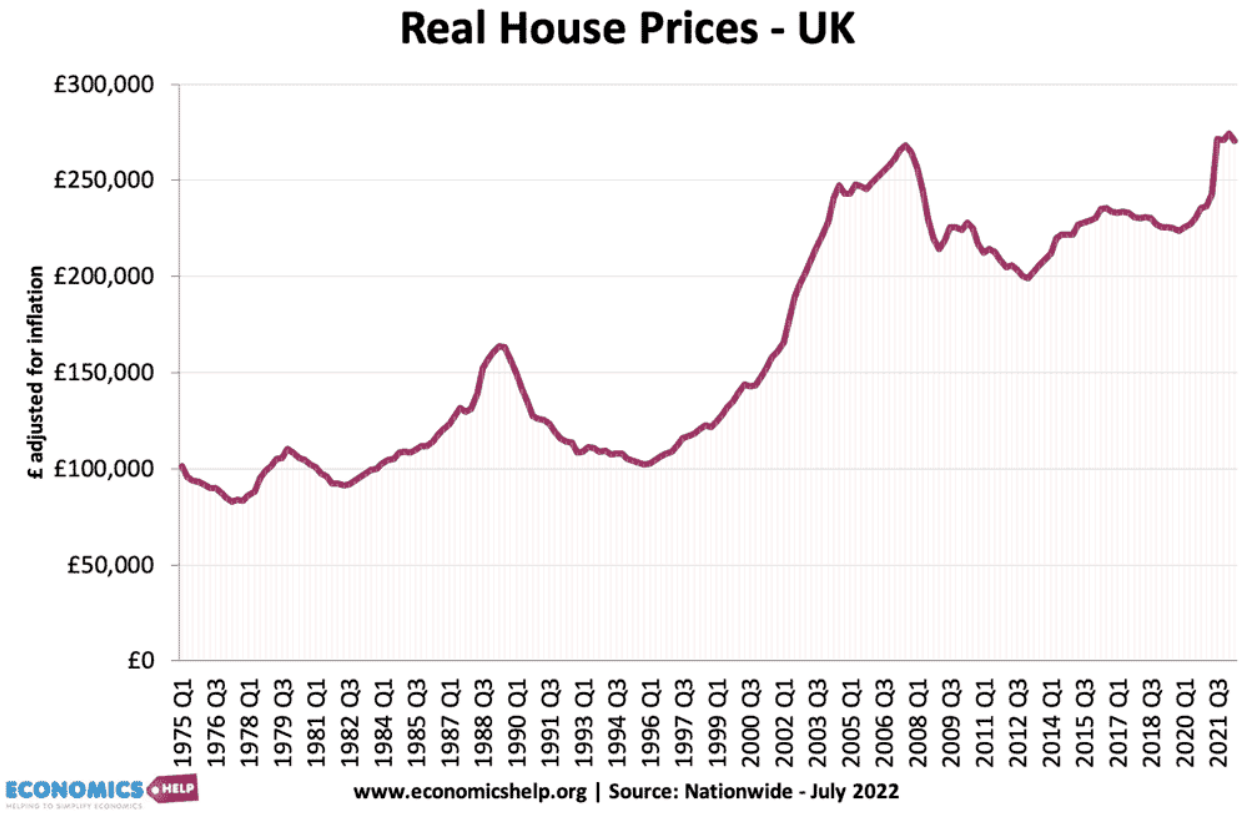
This graph shows that first time buyers in London face much more expensive house prices – over 1.0 times earnings compared to the north, where house prices are only 4.5 times earnings.

House prices using supply and demand



This diagram shows the period 1998 to 2005 where house prices more than doubled in the UK. During these years, there was only a limited increase in supply. By contrast, there was a significant rise in demand due to low-interest rates, positive economic growth, a rising population, high mortgage availability and confidence the housing market was a good investment.

UK Housing Market since 1976



Between 1993-2007 house prices rose sharply. Between 2007-2012 house prices fell as a result of:

- The credit crunch and a decline in bank lending (mortgage lending much stricter criteria.)
- House prices became over-valued in boom years meaning few first-time buyers could afford.
- Recession and rise in unemployment discouraged many from buying.

Since 2012, house prices have risen faster than inflation – despite a relatively weak economy. This is partly due to an ongoing shortage of supply and a rising population.

Why Boom and Busts are common in Housing Markets

If you look at a graph of UK or American house prices - It tends to remind you of one thing - A Rollercoaster. (BTW: since end of graph, UK house prices have fallen £20,000)

Why are House prices so Volatile? Why don't house prices increase at a constant rate, or even remain close to the inflation rate.

On first glance, you might expect house prices to be steady:

- Houses are difficult to buy and sell. There are both financial costs and costs in terms of time. It's not like a stock commodity which can easily be traded.
- The majority of people who buy, do it to live in - not as a speculative investment.

These are some of the reason to explain house price volatility.

Interest rates.

The last boom in house prices in the UK, was burst by a rapid increase in interest rates. If interest rates doubled from 6% to 12% alot homeowners suddenly start defaulting; it is a major disincentive to buy. Interest rates are used to control inflation and the economic cycle - not to stabilise house prices.

In the US, many housing problems were exacerbated when the Federal reserve increased rates from 2% to 4% in 2006. 4% interest rates are still relatively low, but, many had borrowed up to the hilt. This small rise in interest rates stretched their affordability.

Time Delays in Building House

When house prices are rising, builders want to increase supply. However, from planning to completion can take up to 2 years. Therefore, if builders start building at peak of boom, when

prices are falling, new houses are still coming onto the market. Therefore, the increase in supply, magnifies the falling prices (particularly a problem in the US at the moment). At the start of the boom, an inelastic supply squeezes prices upwards, even with relatively moderate demand. (this is the case in the UK)

Confidence Factor

When prices are falling, people want to delay their purchase - after all it could save you tens of thousands of dollars. Therefore, falling house prices deter buyers, causing even lower prices. This is exacerbated by media headlines which highlight the 'housing crisis'

When prices are rising, the opposite happens. People see it as an opportunity to make equity gains or consolidate debt. It becomes easy for people to slip into notion that 'house prices always rise' Housing has generally been seen as the safest investment you can make - I mean it's not like buying some obscure dot com firm.

Mortgage Industry is Cyclical.

When prices are rising, mortgage firms become more willing to lend 100% mortgages. Homeowners need less deposit, because the hope is for rising house prices to effectively create the deposit. When prices are falling, the need for a deposit rises to insure mortgage firms against negative equity. This is the reason why mortgages are often more diffficult to secure during falling prices.

- Of course, during the present credit crunch the cyclical nature of the mortgage industry has reached unprecedented levels.

The big question is whether we will ever be able to restore normalcy to housing markets. Will governments and financial authorities be able to prevent future housing booms and busts?

1.4 RESEARCH OBJECTIVES

This research paper has the following objectives:

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

Business Goal:

You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

1.5 RESEARCH METHODOLOGY

❖ RESEARCH DESIGN

The research design is the conceptual framework around which the survey is undertaken. Here a part of the research undertaken is a **Exploratory Research** as it is describing the perception of the respondents.

❖ SAMPLE SIZE

The Sample Size is **1460**. Data has been collected and analysed on the basis of the responses. The research was conducted with an aim of getting respondents from all across USA.

❖ PERIOD OF STUDY

The period of study was more than **a week.**

❖ DATA COLLECTION

1. **Primary Data-** Dataset and Data description provided by Flip Robo technologies for project completion.
2. **Secondary Data-**The data includes reference from previously published research papers, journals, books and articles.

❖ METHOD OF ANALYSIS

In this research, tables and various types of graphs such as bar graph and count plot have been used. This helped to present the data in a meaningful way and making it easily understandable. We will use LINEAR REGRESSION MODEL TO PREDICT THE HOUSE PRICE FOR TEST.CSV DATA.

1.6 RESEARCH LIMITATIONS

- **Non Representative Sample**: The research project was based on the survey conducted of only 1460 respondents. Hence, such sample size cannot be said to be genuine and actual representative of the people.

- **Shortage of Time**: The time frame of the study was restricted and limited. It therefore becomes difficult to have detailed study on project work. The period was not enough for the proper study and investigation on the project.

- **Use of only Virtual Research Methods**: The survey was based on the data collected by questionnaire which was circulated virtually having limited questions which in turn resulted in collection of insufficient data due to which there was inadequacy in the research.

- **Lack of Scientific Method**: The absence of use of scientific and logical training in research approach turned out to be a great hindrance in the exploration programme.

2. STEPS USED IN PREDICTING HOUSING PRICE:

2.1 Choosing the model

We will use Linear Regression model for this dataset as, we need to predict the customer retention which is a continuous data. Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between one response variable (denoted by y) and one or more explanatory variables (denoted by X1, X2, X3…. Xn). The response variable will dependent on how the explanatory variables changes and not the other way round. Response variable is also known as target or dependent variable while the explanatory variable is known as independent or predictor variables.

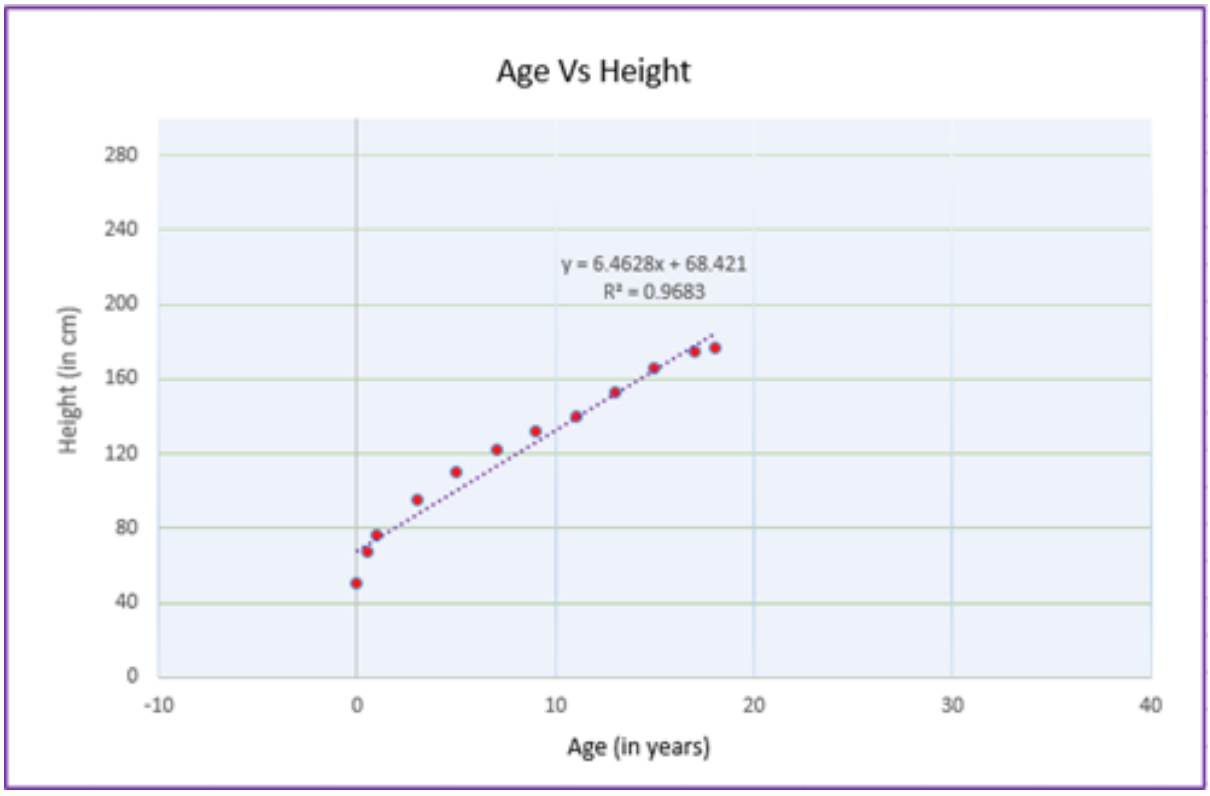


Fig: Simple Linear Regression between Height and age

There are 2 types of linear regression:

1. Simple Linear Regression

2. Multiple Linear Regression

Simple Linear Regression*: It is a type of linear regression model where there is only independent or explanatory variable. For e.g., the above scatter plot follows a simple linear regression with age being an independent variable is responsible for any change in height (dependent variable).*

Multiple Linear Regression*: It is similar to simple linear regression but here we have more than one independent or explanatory variable.*

Linear Regression can be written mathematically as follows:

$$\mathbf{Y = \beta 0 + \beta 1.X1 + \beta 2.\ X2 + \beta 3.\ X3 + \beta 4.\ X4 + \beta 5.\ X5 + \beta 5.\ X6 + \epsilon}$$

$$\mathbf{charges = \beta 0 + \beta 1.bmi + \beta 2.age + \beta 3.sex + \beta 4\ .children + \beta 5.region + \beta 5.smoker + \epsilon}$$

charges= response variable, generally denoted by Y

bmi, age, sex, children, region, smoker=Predictor variables, denoted by X1, X2, X3 and X4 respectively

β0 = Y-intercept (always a constant)

β1, β2, β3, β4, β5 = regression coefficients

ϵ = Error terms (Residuals)

C**omponents of Linear Regression:**

1. Regression Coefficient (or β1):

The Regression Coefficient in the above equation talks about the change in the value of dependent variable corresponding to the unit change in the independent variable. So, for e.g. if X1 increases or decreases by one unit, then Y will increase or decrease by β1 units. An important assumption followed by an ideal linear regression is that any increase or decrease in

one independent variable will not have any corresponding changes in other independent variables.

2. Intercept (or β0):

Intercept is a constant value which tells us at what point in the x-y coordinate graph, should the regression line must start if it follows a linear regression. Since it is a constant value, hence it is not dependent on any change in independent variables. Even if the values of X=0, intercept will have a constant value. If the value of intercept is 0, that means, the line will start at the origin point (0,0).

3. Error Terms or Residuals (ϵ):

It is the difference between the actual and the predicted data point in the x-y coordinate graph

Objective of Linear Regression:

The goal of linear regression is to perform predictive analytics and it is done by making the machine learn the science of generating a trained (best fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets.

S**teps to be followed in Linear Regression Algorithm:**

1. Reading and understanding the data

a. Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization

b. Cleaning and manipulating data to make it up to the standards that exploratory data analysis can be performed by treating null values if any, updating to necessary formats, changing data types if needed, removing unwanted rows or columns etc. The raw data in whatever condition you get must be squeaky cleaned of any muck before assessing it for visualization.

2. Visualizing the data (Exploratory Data Analysis)

a. Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.

b. Visualizing categorical variables using barplots or boxplots in order to interpret business/domain inferences.

3. Data Preparation

a. Converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be represented during model building in order to contribute to the best fitted line for the purpose of better prediction.

4. Splitting the data into training and test sets

a. Splitting the data into two sections in order to train a subset of dataset to generate a trained (fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets. Generally, the train-test split ratio is 70:30 or 80:20.

b. Rescaling the trained model: It is a method used to normalize the range of numerical variables with varying degrees of magnitude. For e.g. height or bmi or age are of different magnitude and units or some feature may have values in 10000s while feature may contain values in the magnitude of 10s or 100s, then the contribution of each feature for the dependent variable will be different

5. Building a linear model

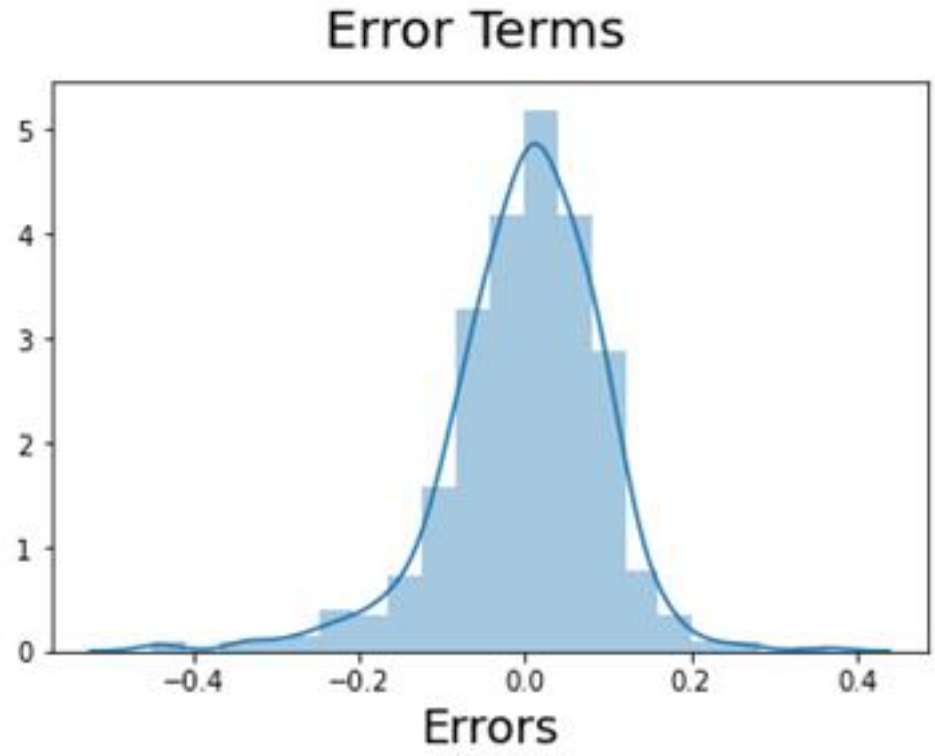
a. Forward Selection: We start with null model and add variables one by one. These variables are selection on the basis of high correlation with target variable. First we select the one, which has highest correlation and then we move on to the second highest and so on.

b. Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p- values).

c. RFE or Recursive Feature Elimination is more like an automated version of feature selection technique where we select that we need “m” variables out of “n” variables and then machine provides a list of features with importance level given in terms of rankings. A rank 1 means that feature is important for the model, while a rank 4 implies that we are better off, if we don’t consider the feature.

6. Residual analysis of the train data:

a. It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0.



The residual errors are centered around 0

7. Making predictions using the final model and evaluation:

a. We will predict the test dataset by transforming it onto the trained dataset

b. Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score. A difference of 2–3% between r2_score of train and test score is acceptable as per the standards.

2.2 Steps in EDA and Preprocessing:

1. Identification of variables and data types
2. Analyzing the basic metrics
3. Non-Graphical Univariate Analysis
4. Missing value treatment
5. Graphical Univariate Analysis
6. Bivariate Analysis
7. Encoding the categorical Data
8. Outlier treatment
9. Variable transformations
10. Correlation Analysis
11. Dimensionality Reduction
12. Scaling of Independent features

2.3 Steps followed in project

1. We need to apply feature engineering/ EDA on the dataset. After that you can split the dataset into train_model and test_model; use this train_model to train your model and test_model to validate your model.

2. After splitting the dataset you need to train at least 4-5 models.

3. Check performance of each model

For regression problem- create regression model and check the r2 score and metrics of each model .

4. Check the cross validation score for each model(for classification as well as for regression model).

5. Choose the model as the best model.

For regression problem- model with least difference between performance parameter and cross validation computed on same performance parameter is the best model. Example- Difference between r2 score and cross validation computed on r2 scoring parameter.

6. Perform hyper parameter tuning on the best model and check performance of the best model.

7. Save the best model.

*3. **ANALYSIS**: DATA FINDINGS AND INTERPRETATION*

3.1 Data shape and Data Types

The shape property returns a tuple containing the shape of the DataFrame.

The shape is the number of rows and columns of the DataFrame

When doing data analysis, it is important to make sure you are using the correct data types; otherwise you may get unexpected results or errors. In the case of pandas, it will correctly infer data types in many cases and you can move on with your analysis without any further thought on the topic.

Despite how well pandas works, at some point in your data analysis processes, you will likely need to explicitly convert data from one type to another. This article will discuss the basic pandas data types (aka dtypes), how they map to python and numpy data types and the options for converting from one pandas type to another.

Pandas Data Types

A data type is essentially an internal construct that a programming language uses to understand how to store and manipulate data. For instance, a program needs to understand that you can add two numbers together like 5 + 10 to get 15. Or, if you have two strings such as “cat” and “hat” you could concatenate (add) them together to get “cathat.”

A possible confusing point about pandas data types is that there is some overlap between pandas, python and numpy. This table summarizes the key points:

Pandas dtype mapping

Pandas dtype	**Python type**	**NumPy type**	**Usage**
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	Int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	Float	float_, float16, float32, float64	Floating point numbers
bool	Bool	bool_	True/False values

Pandas dtype mapping

Pandas dtype	Python type	NumPy type	Usage
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

For the most part, there is no need to worry about determining if you should try to explicitly force the pandas type to a corresponding to NumPy type. Most of the time, using pandas default int64 and float64 types will work. The only reason I included in this table is that sometimes you may see the numpy types pop up on-line or in your own analysis.

For this article, I will focus on the follow pandas types:

- object
- int64
- float64
- datetime64
- bool

The category and timedelta types are better served in an article of their own if there is interest. However, the basic approaches outlined in this article apply to these types as well.

One other item I want to highlight is that the object data type can actually contain multiple different types. For instance, the a column could include integers, floats and strings which collectively are labeled as an object . Therefore, you may need some additional techniques to handle mixed data types in object columns.

Columns and datatypes of it on our dataset:

There are total 81 columns in our dataset and they are:

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

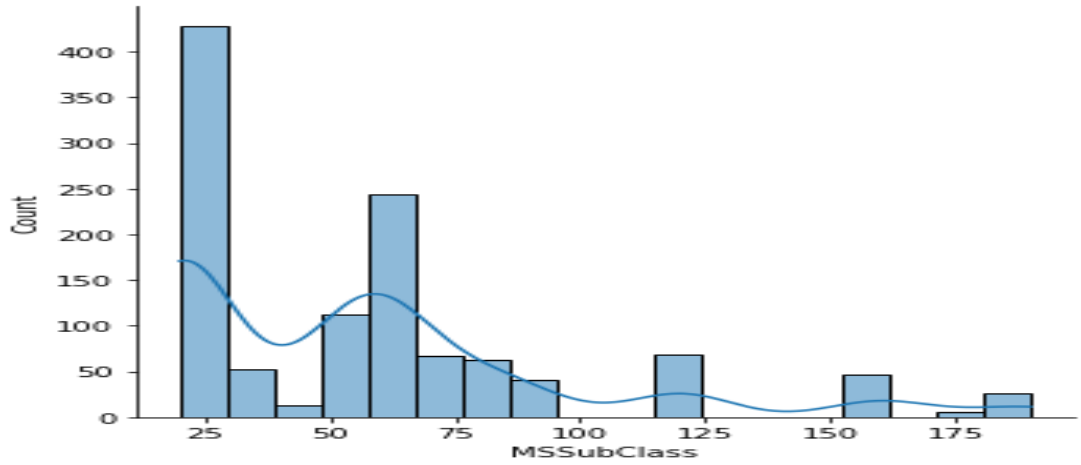
Family Sale between family members

Partial Home was not completed when last assessed (associated with New Hom es)

3.2 DATA VISUALIZATION

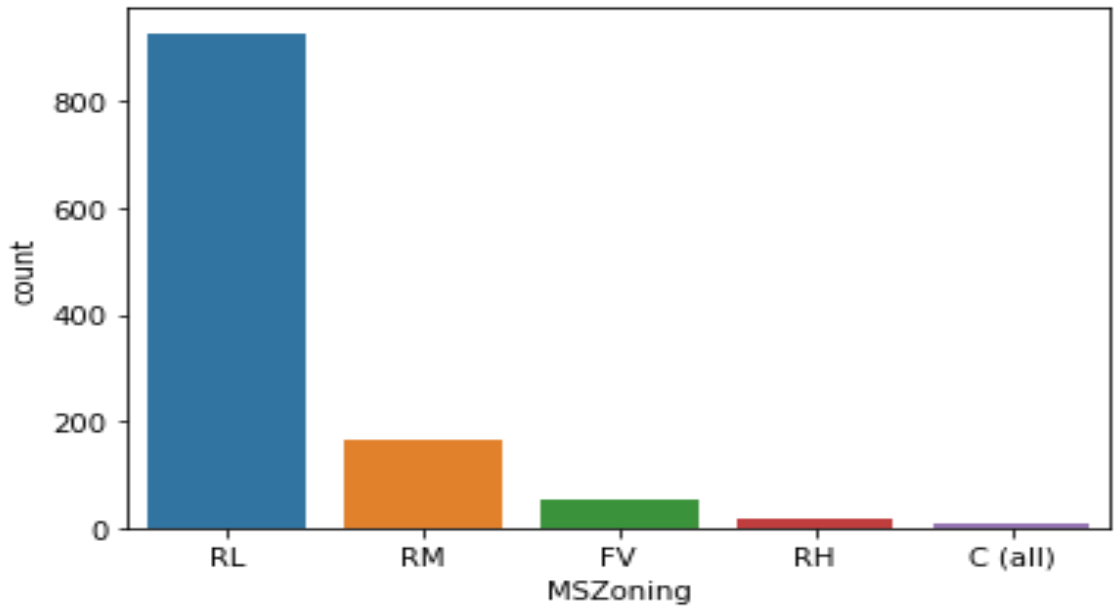
We make the dataframe for the nominal categorial data under our dataset train.csv.
We see that,some columns are spread over categories while some are in integer format. For example: MSSubClass is integer while Street is in categorial format.
We use countplot for data visualization of the categorial data as it gives the frequency and value counts of specific categories and Distplot for integer form columns.

1.

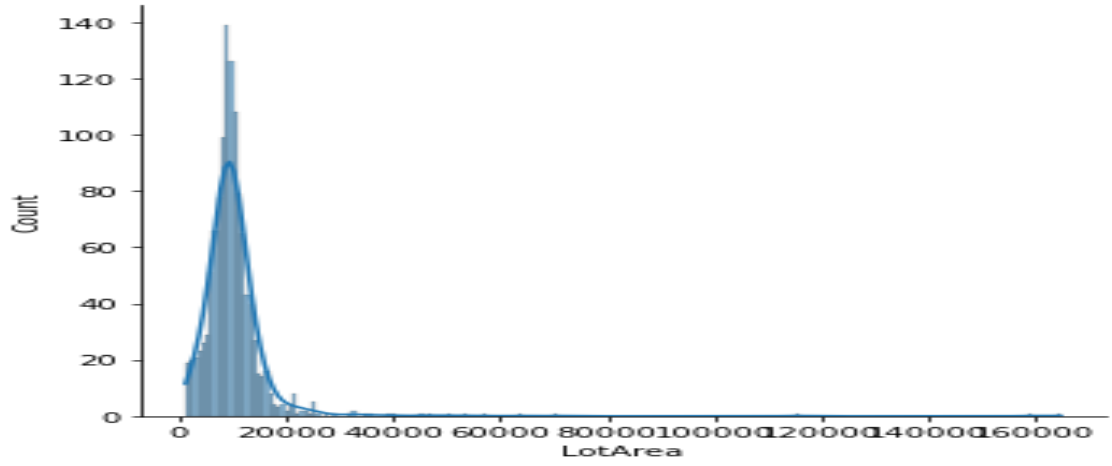


2.

```
RL          928
RM          163
FV           52
RH           16
C (all)       9
```

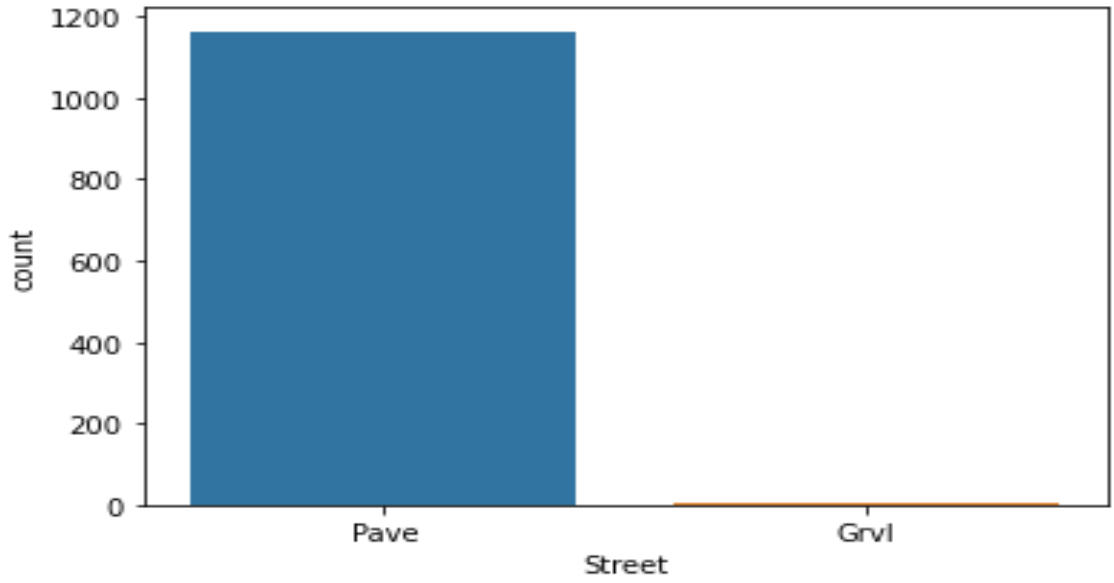


3.



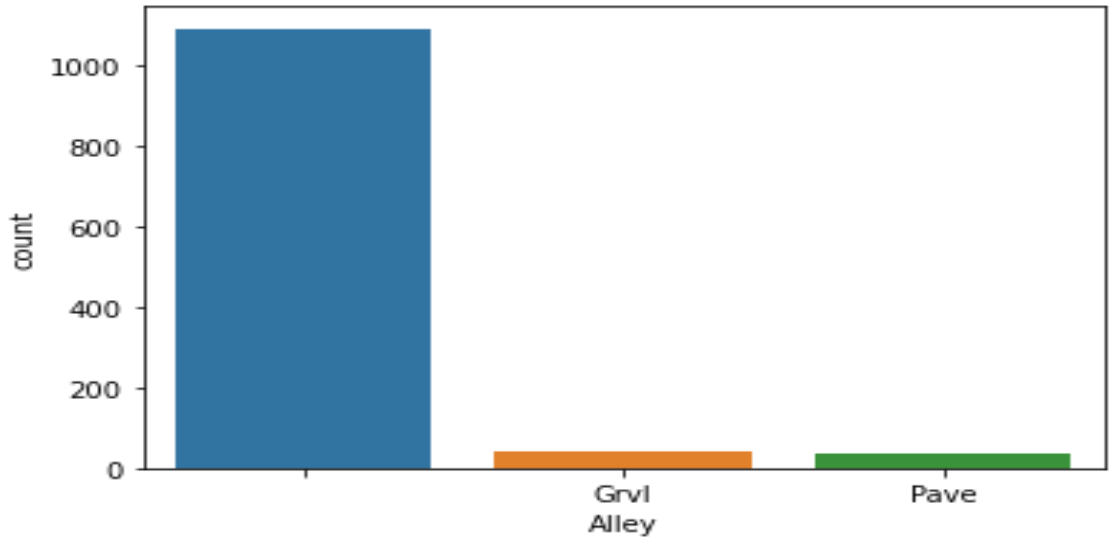
4.

```
Pave    1164
Grvl       4
```



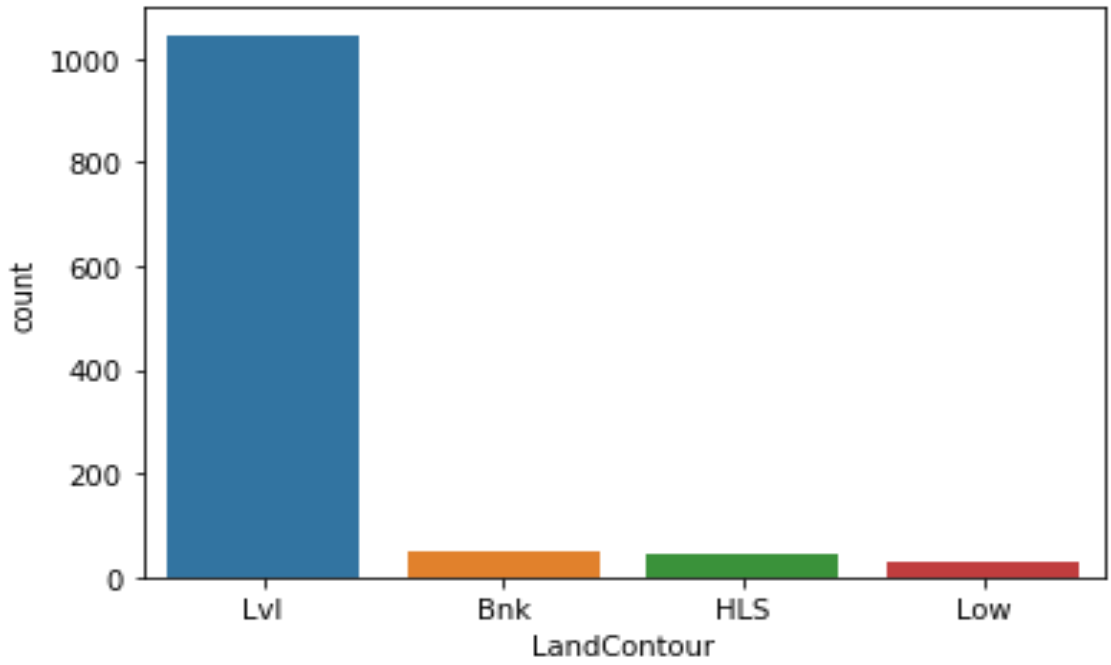
5.

```
        1091
Grvl      41
Pave      36
```



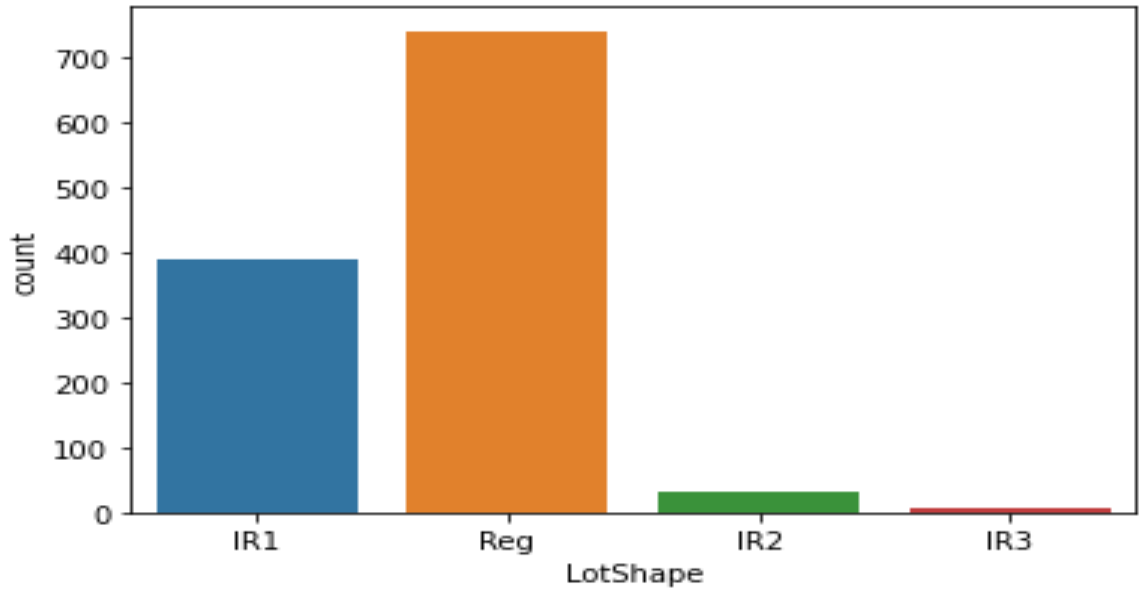
6.

```
Lvl     1046
Bnk       50
HLS       42
Low       30
```



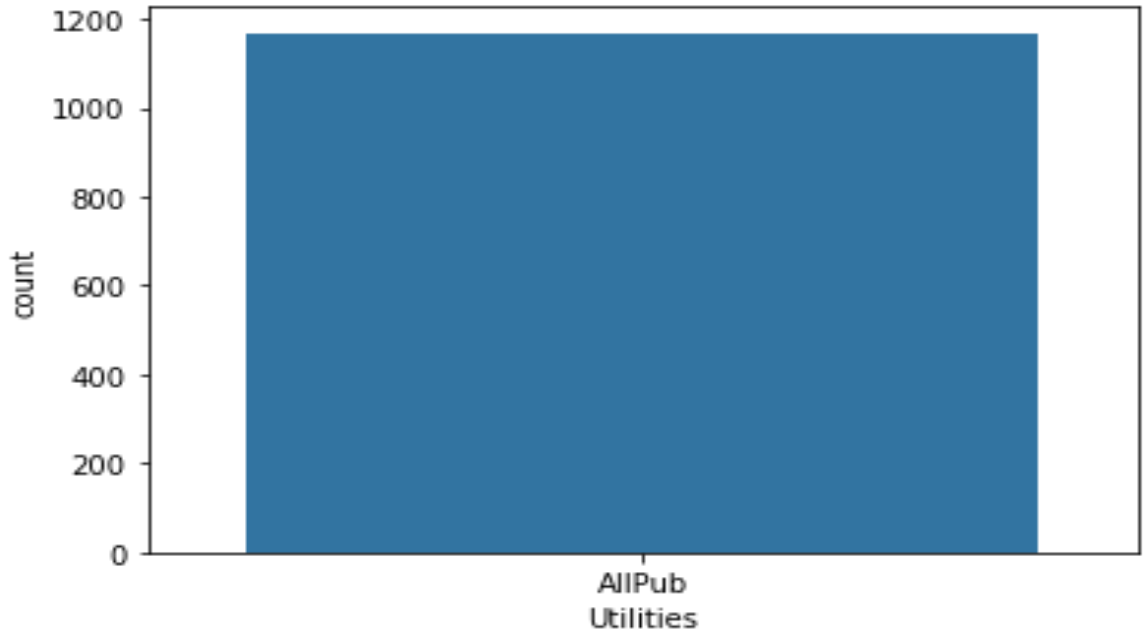
7.

```
Reg     740
IR1     390
IR2      32
IR3       6
```



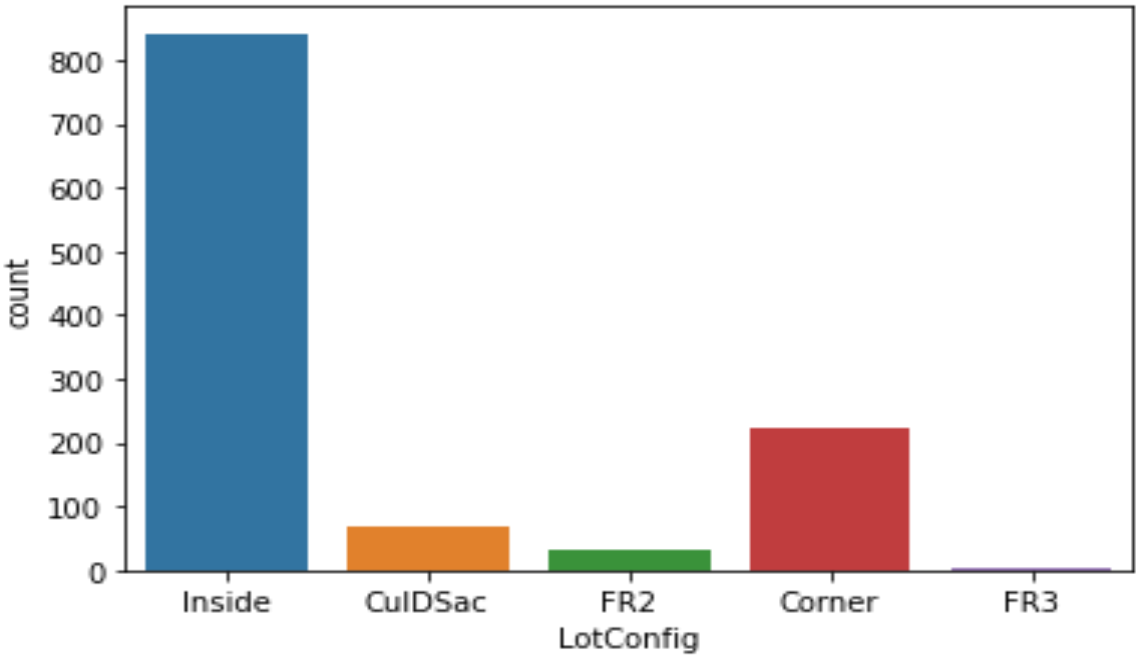
8.

```
AllPub    1168
```



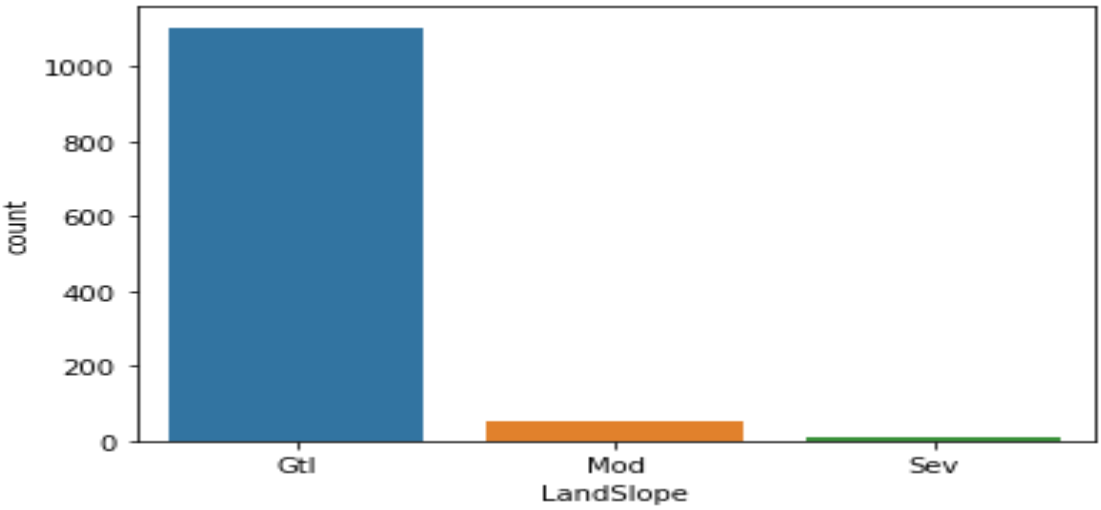
9.

```
Inside     842
Corner     222
CulDSac     69
FR2         33
FR3          2
```



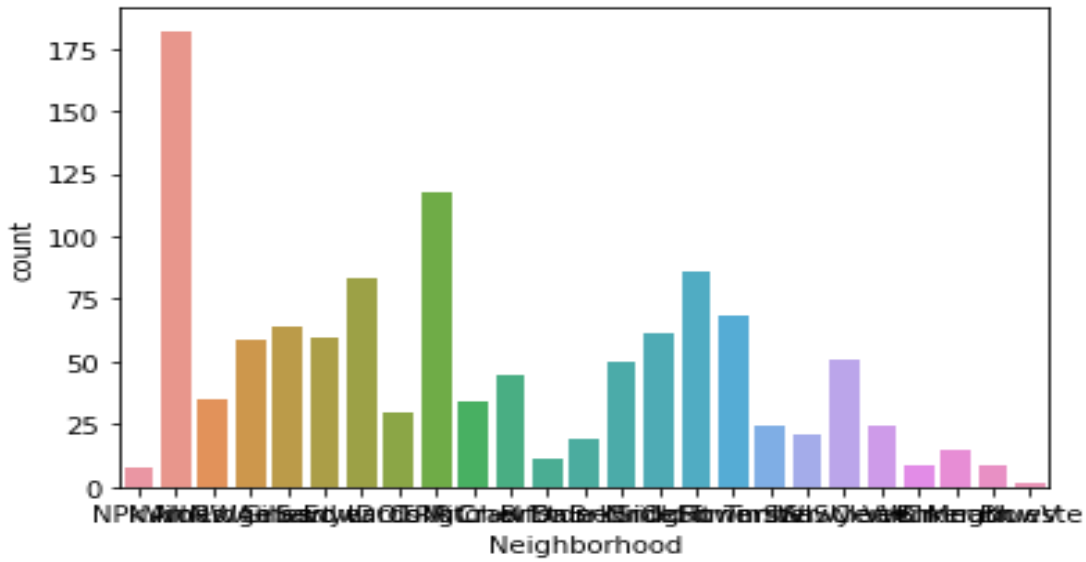
10.

```
Gtl    1105
Mod      51
Sev      12
```



11.

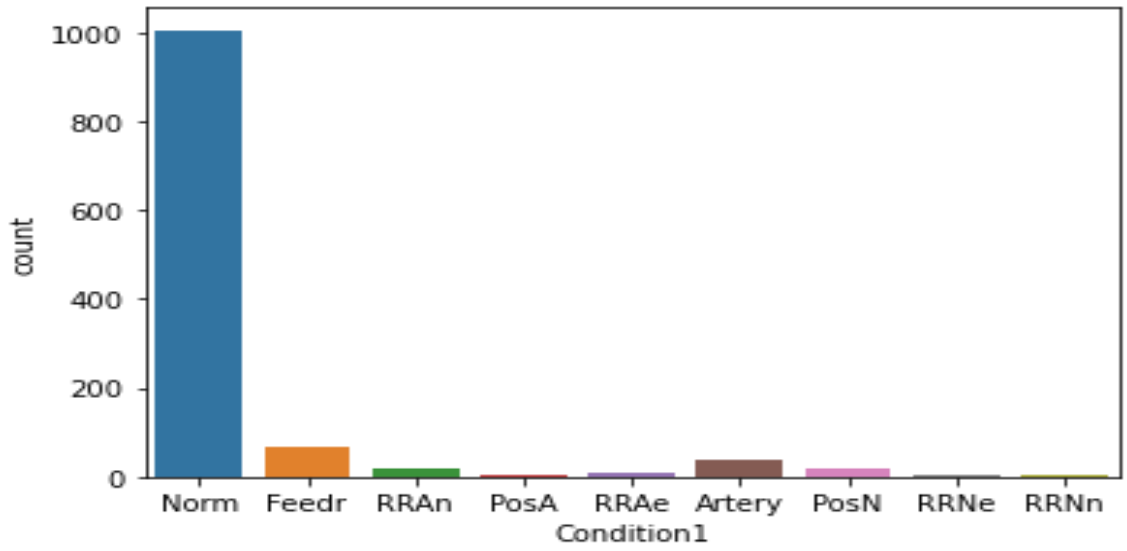
```
NAmes      182
CollgCr    118
OldTown     86
Edwards     83
Somerst     68
Gilbert     64
NridgHt     61
Sawyer      60
NWAmes      59
SawyerW     51
BrkSide     50
Crawfor     45
NoRidge     35
Mitchel     34
IDOTRR      30
Timber      24
ClearCr     24
SWISU       21
StoneBr     19
Blmngtn     15
BrDale      11
MeadowV      9
Veenker      9
NPkVill      8
Blueste      2
```



12.

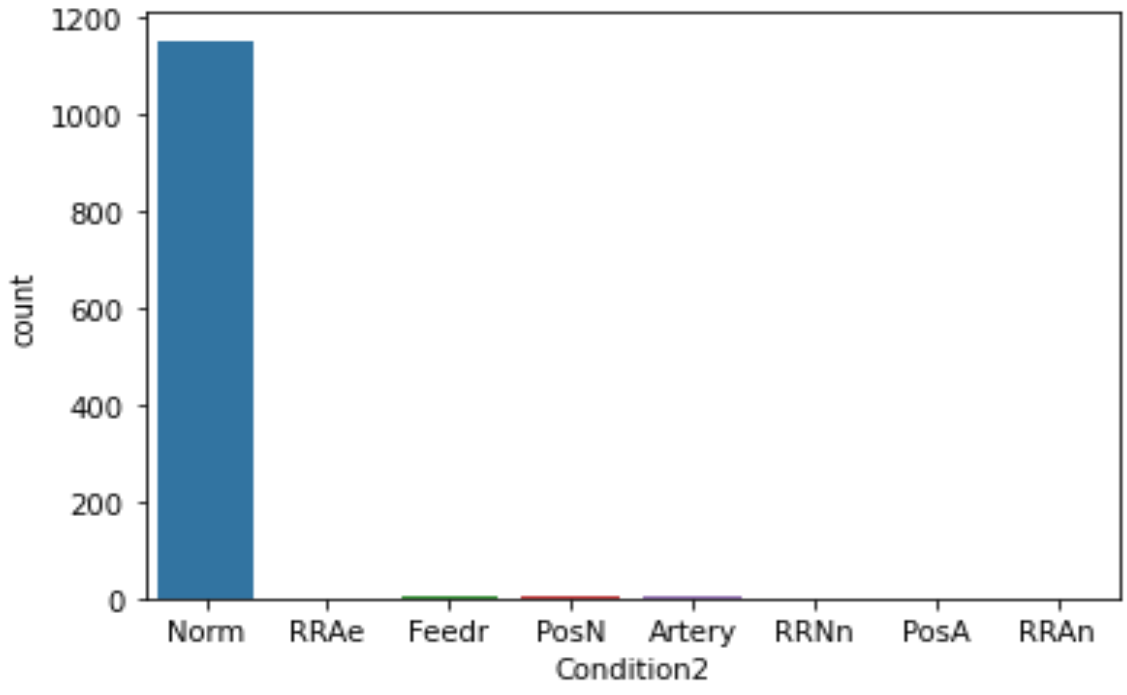
```
Norm       1005
Feedr        67
Artery       38
RRAn         20
PosN         17
RRAe          9
PosA          6
```

```
RRNn          4
RRNe          2
```



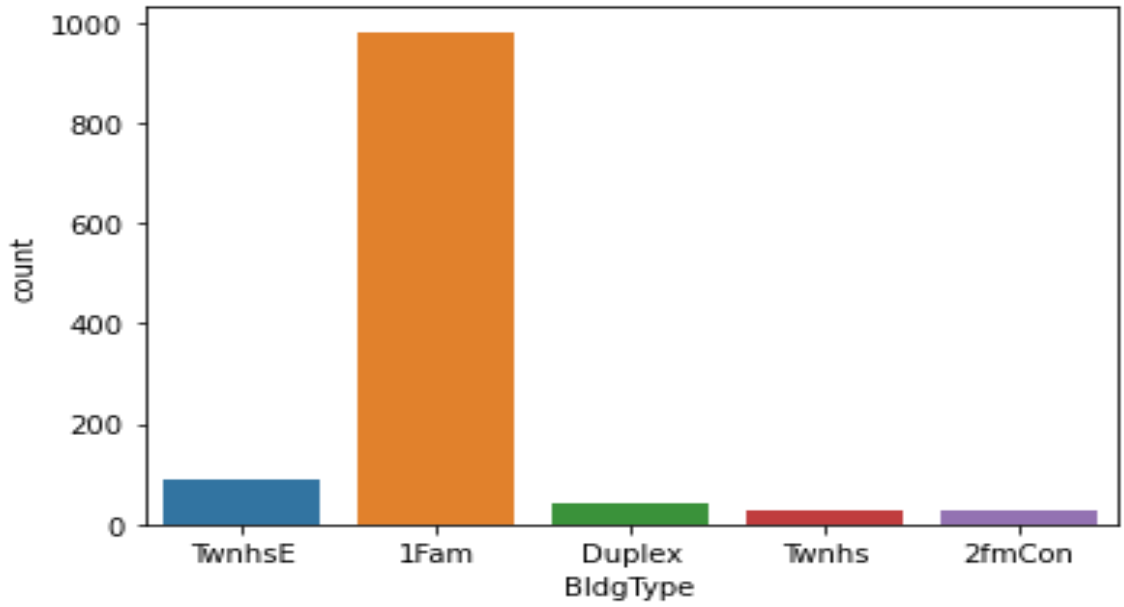
13.

```
Norm       1154
Feedr         6
PosN          2
Artery        2
RRAe          1
RRNn          1
PosA          1
RRAn          1
```



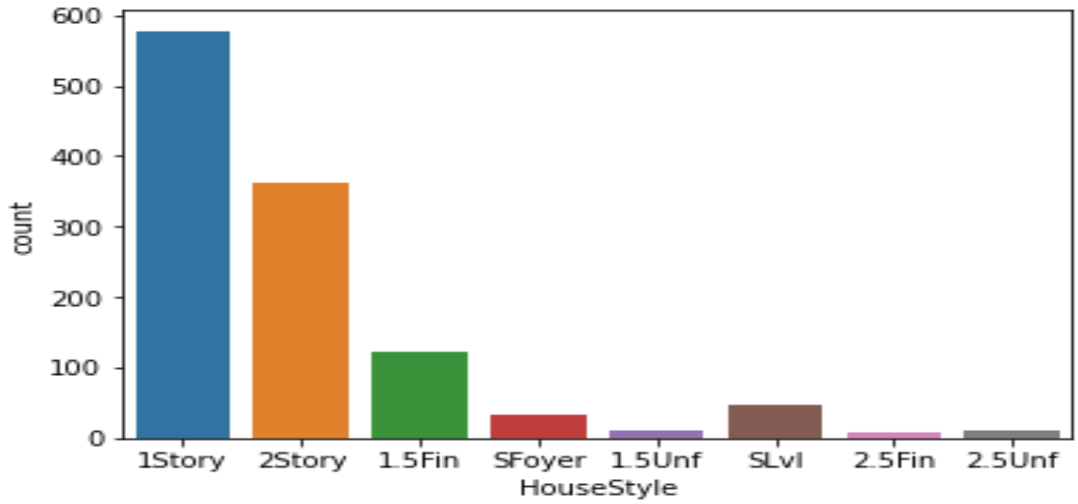
14.

```
1Fam       981
TwnhsE      90
Duplex      41
Twnhs       29
2fmCon      27
```



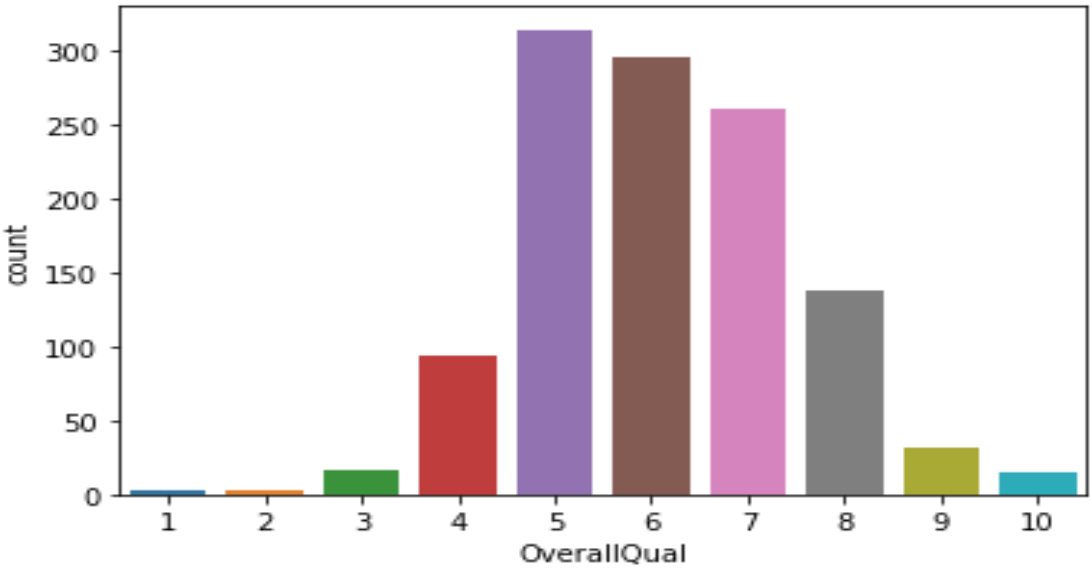
15.

```
1Story     578
2Story     361
1.5Fin     121
SLvl        47
SFoyer      32
1.5Unf      12
2.5Unf      10
2.5Fin       7
```



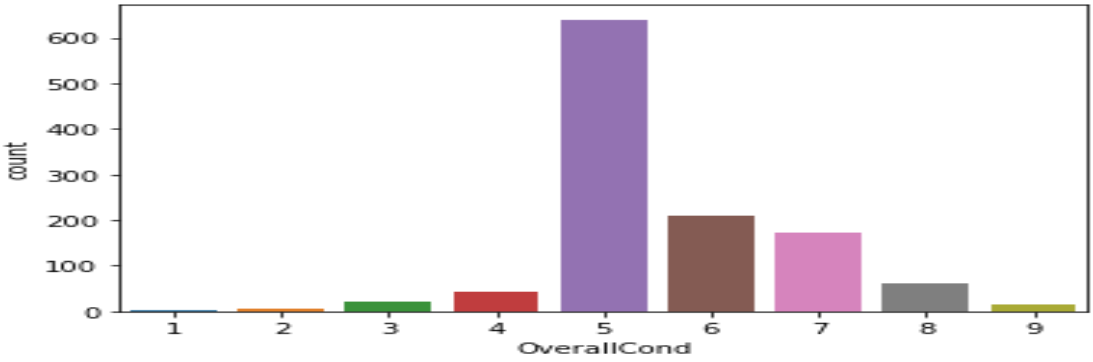
16.

```
5       314
6       295
7       260
8       138
4        93
9        32
3        16
10       15
2         3
1         2
```

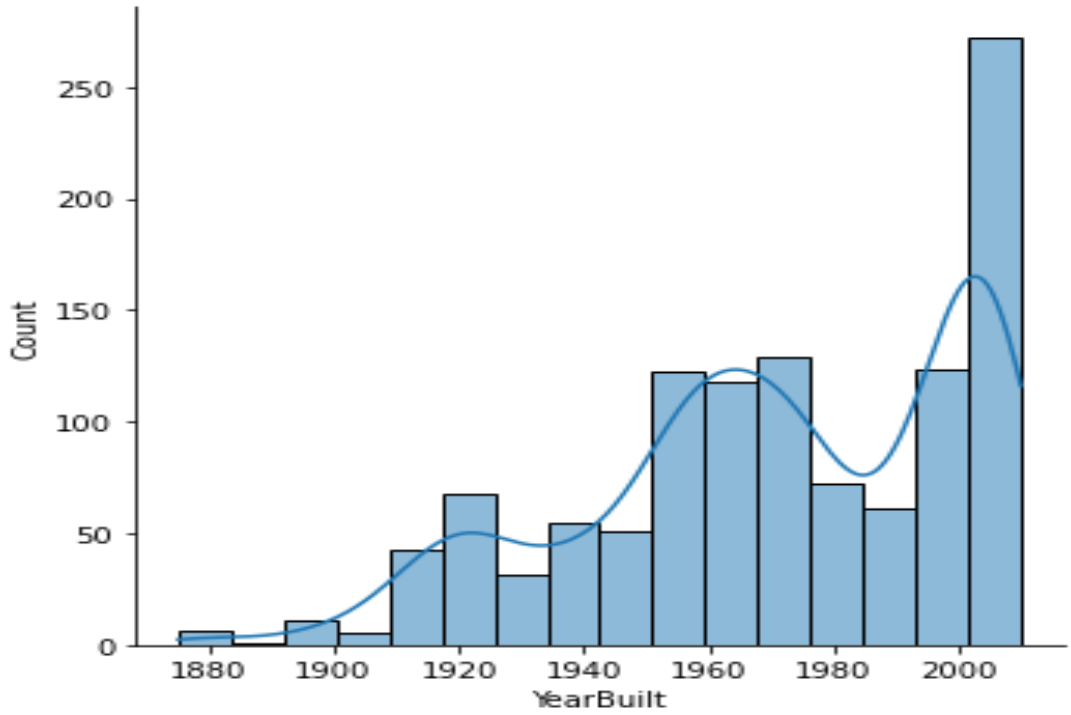


17.

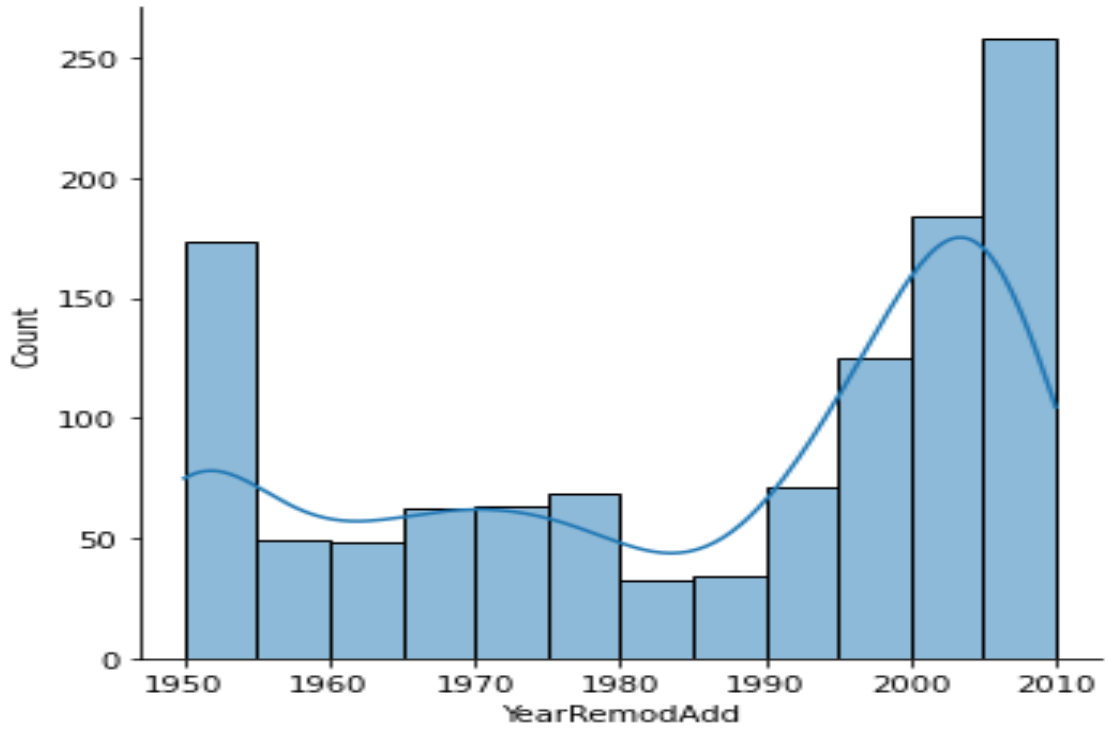
```
5     640
6     209
7     172
8      61
4      43
3      21
9      16
2       5
1       1
```



18.

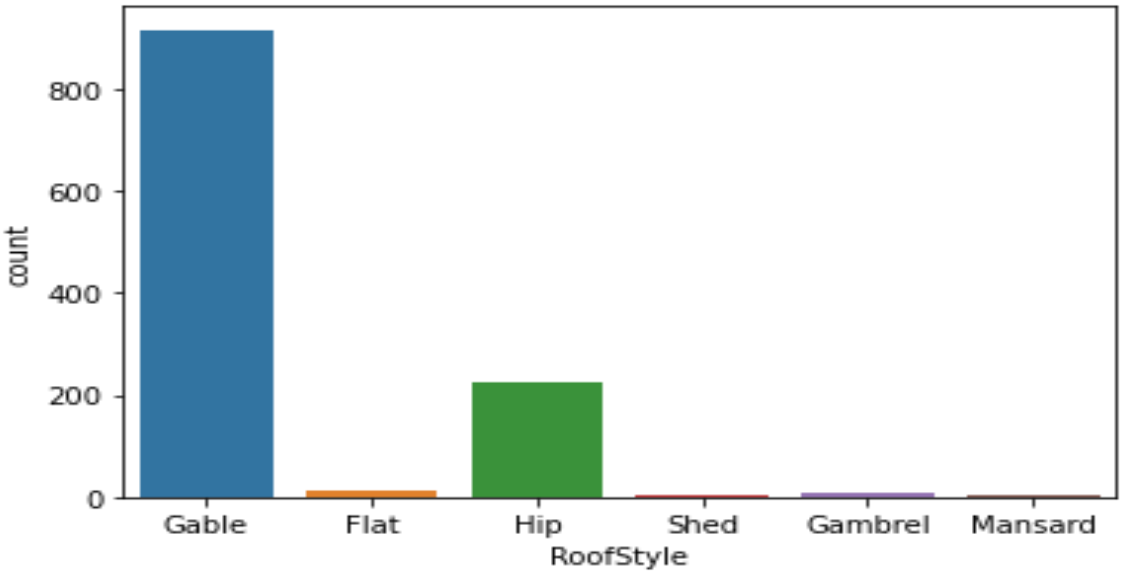

Count
250
200
150
100
50
0
1880
1900
1920
1940
1960
1980
2000
YearBuilt

19.


Count
250
200
150
100
50
0
1950
1960
1970
1980
1990
2000
2010
YearRemodAdd

20.

```
Gable      915
Hip        225
Flat        12
Gambrel      9
Mansard      5
Shed         2
```

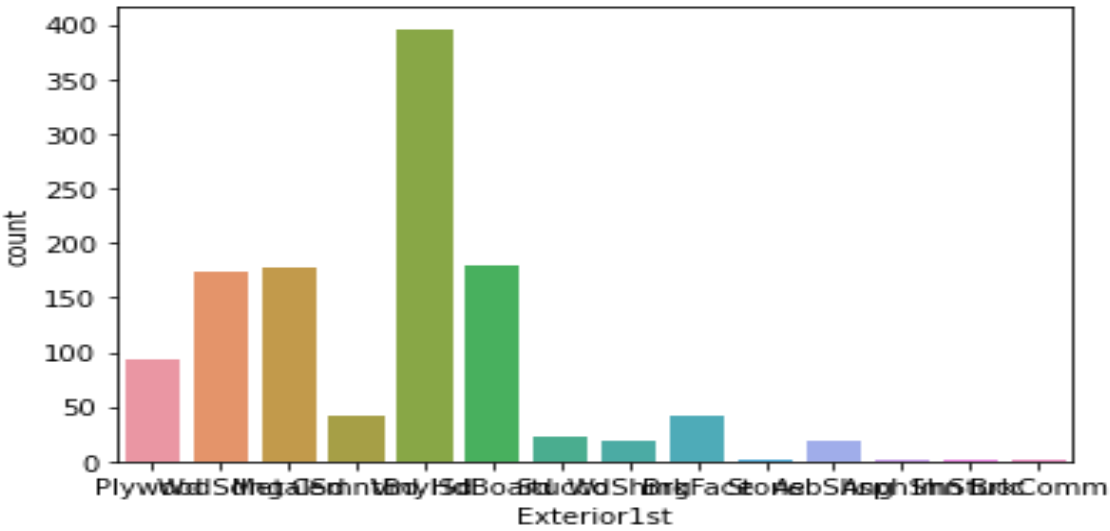


21.

```
CompShg    1144
Tar&Grv      10
WdShngl       6
WdShake       4
Roll          1
ClyTile       1
Metal         1
Membran       1
```

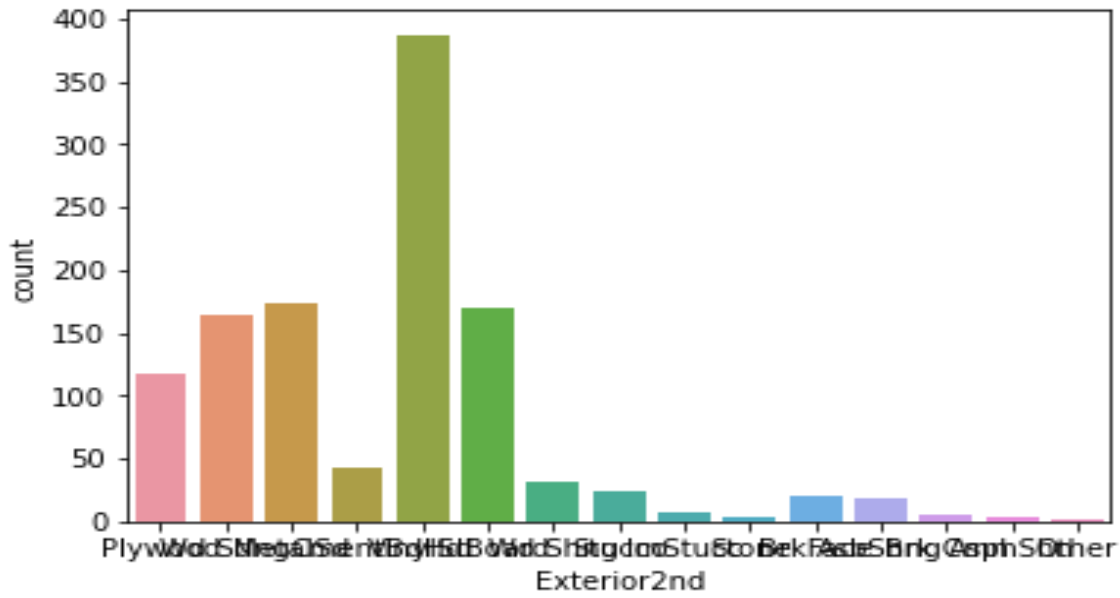
22.

```
VinylSd    396
HdBoard    179
MetalSd    178
Wd Sdng    174
Plywood     93
CemntBd     42
BrkFace     41
Stucco      22
WdShing     19
AsbShng     19
Stone        2
AsphShn      1
ImStucc      1
BrkComm      1
```



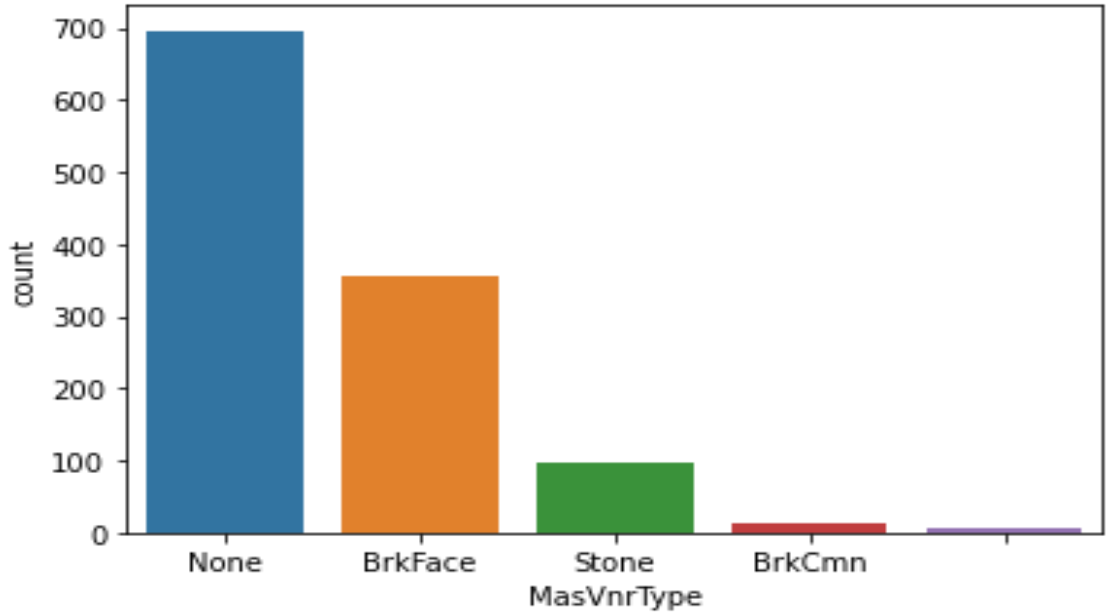
23.

```
VinylSd    387
MetalSd    173
HdBoard    170
Wd Sdng    165
Plywood    118
CmentBd     42
Wd Shng     31
Stucco      23
BrkFace     20
AsbShng     18
ImStucc      8
Brk Cmn      5
Stone        4
AsphShn      3
Other        1
```



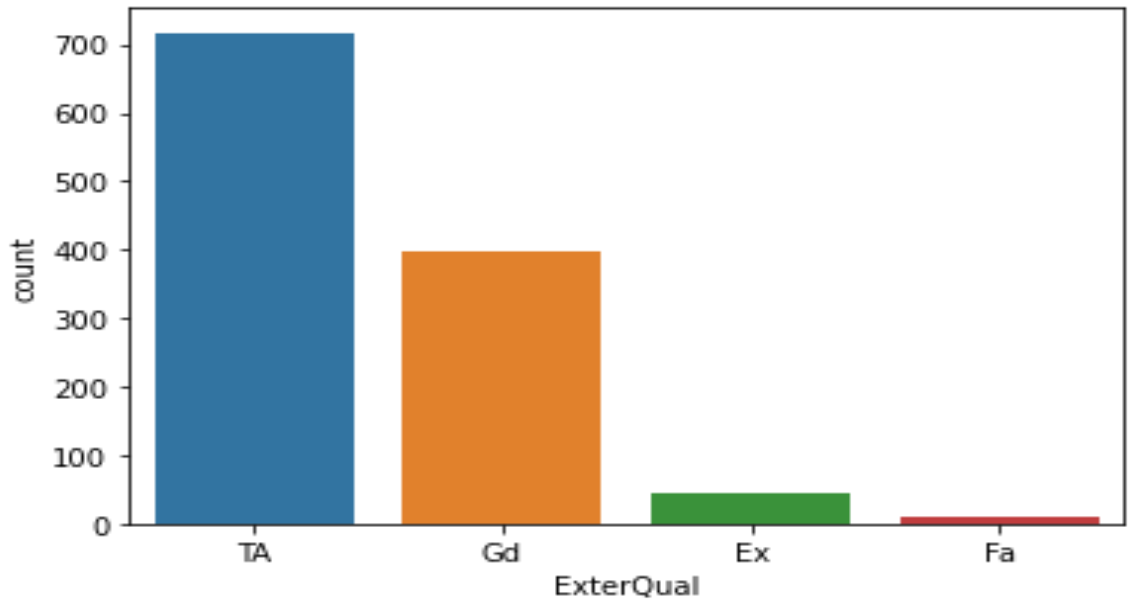
24.

```
None        696
BrkFace     354
Stone        98
BrkCmn       13
              7
```



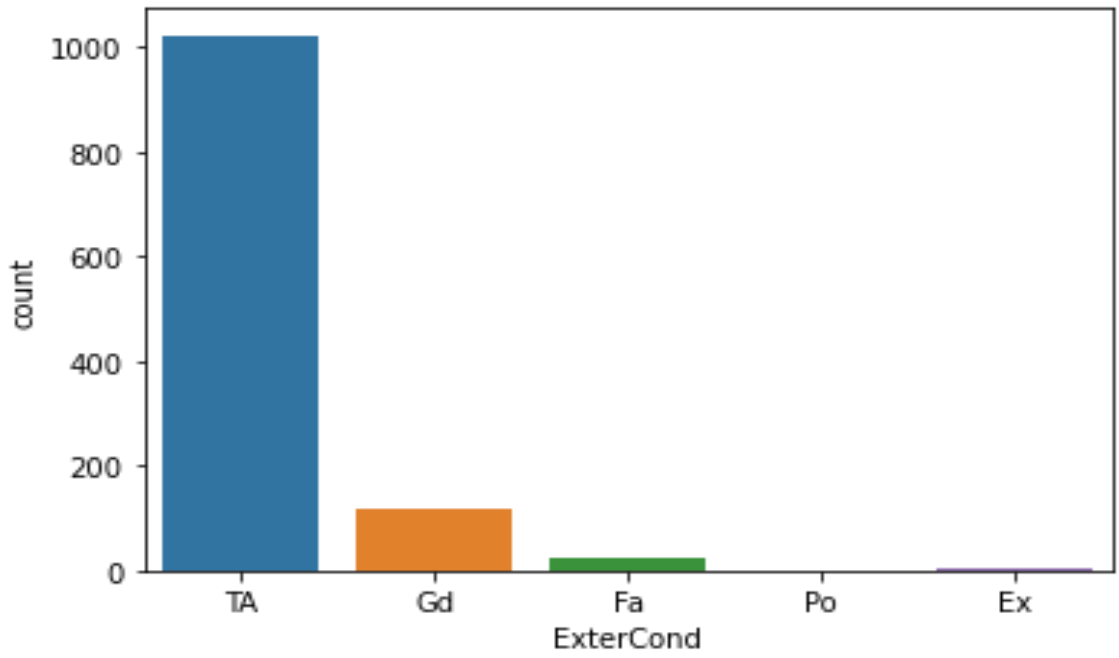
25.

```
TA     717
Gd     397
Ex      43
Fa      11
```



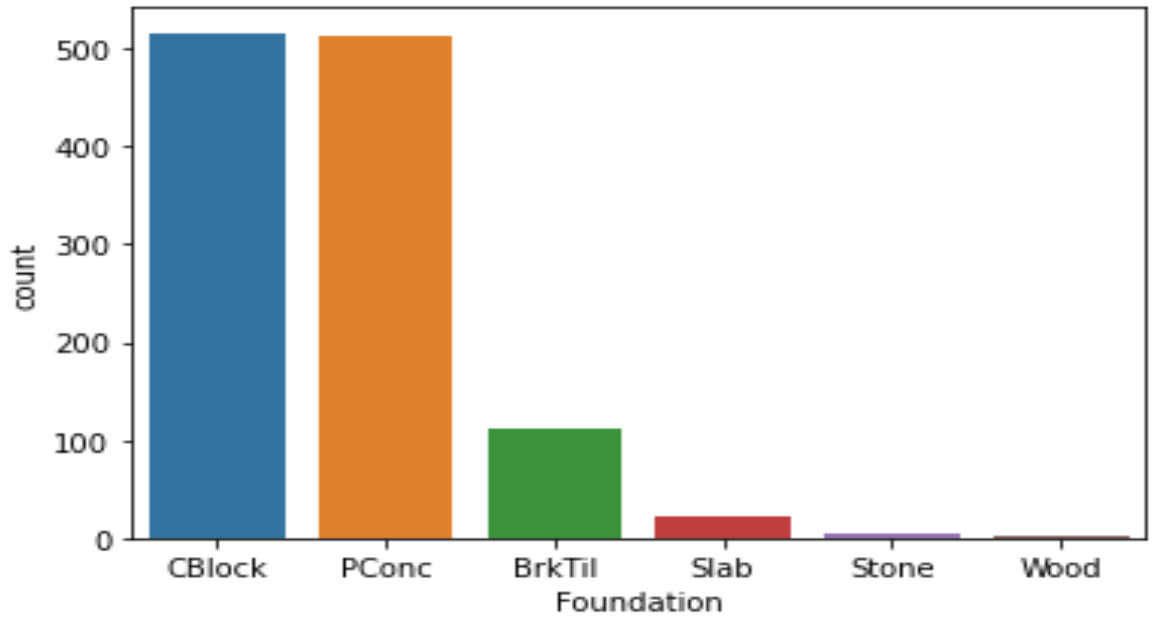
26.

```
TA    1022
Gd     117
Fa      26
Ex       2
Po       1
```



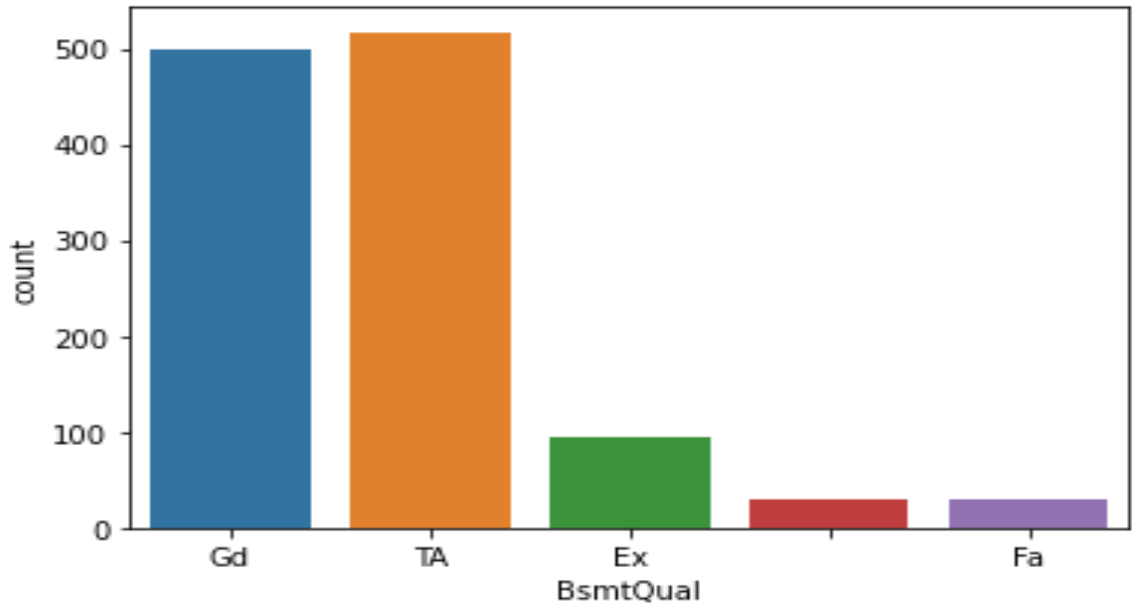
27.

```
5CBlock    516
PConc      513
BrkTil     112
Slab        21
Stone        5
Wood         1
```



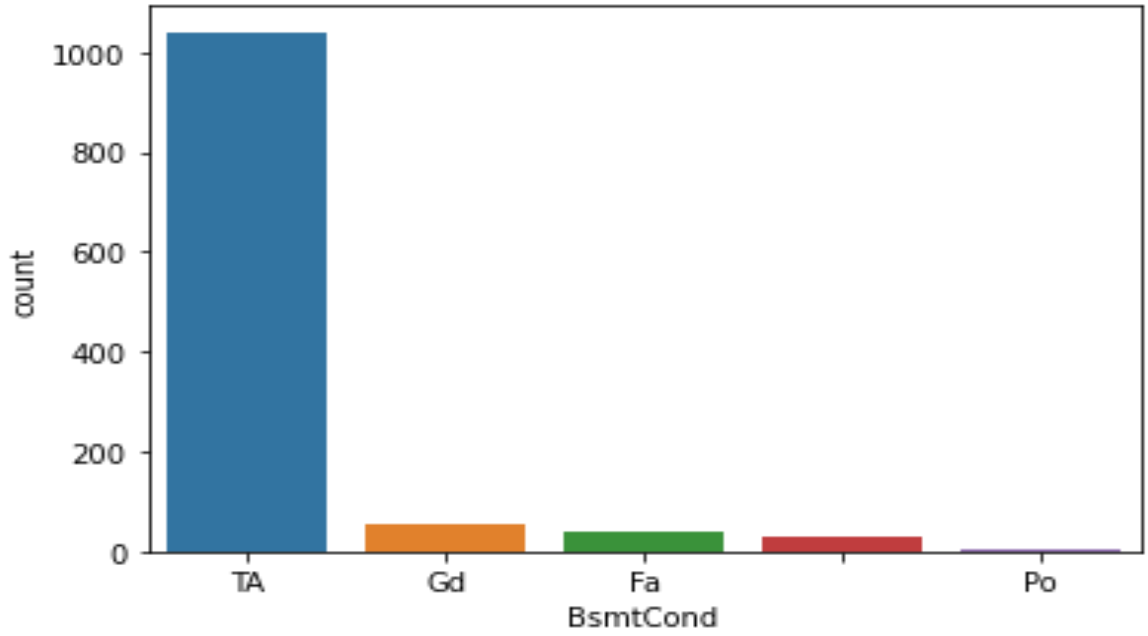
28.

```
TA     517
Gd     498
Ex      94
        30
Fa      29
```



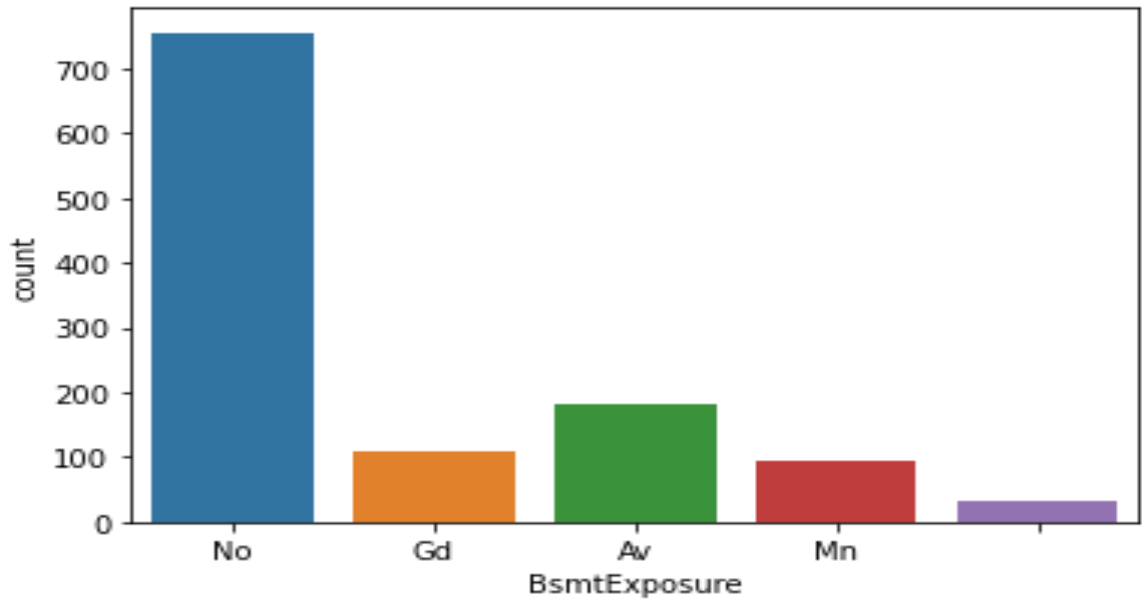
29.

```
TA     1041
Gd       56
Fa       39
         30
Po        2
```



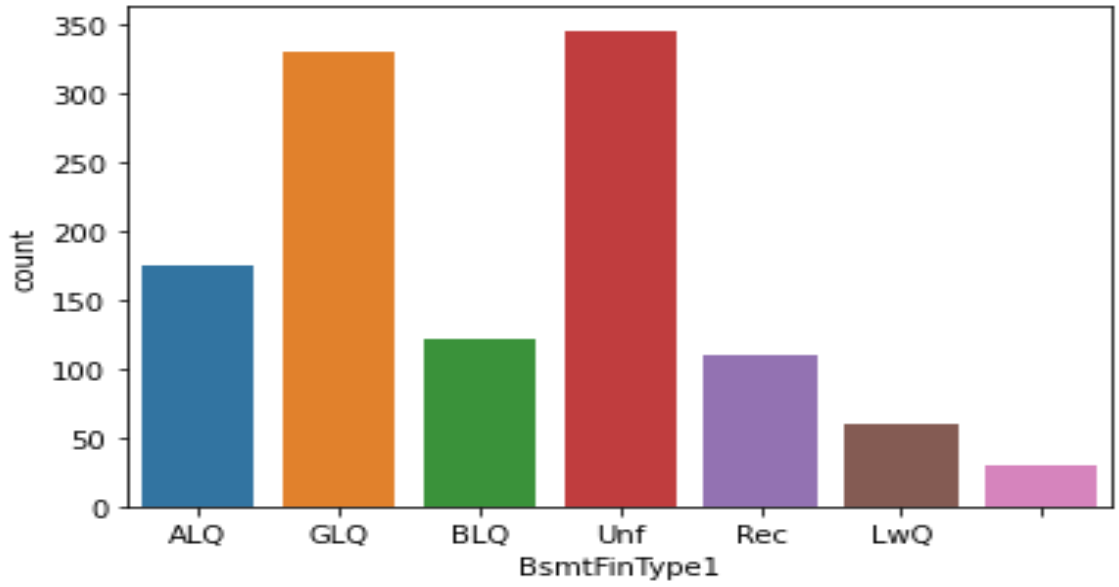
30.

```
No     756
Av     180
Gd     108
Mn      93
        31
```

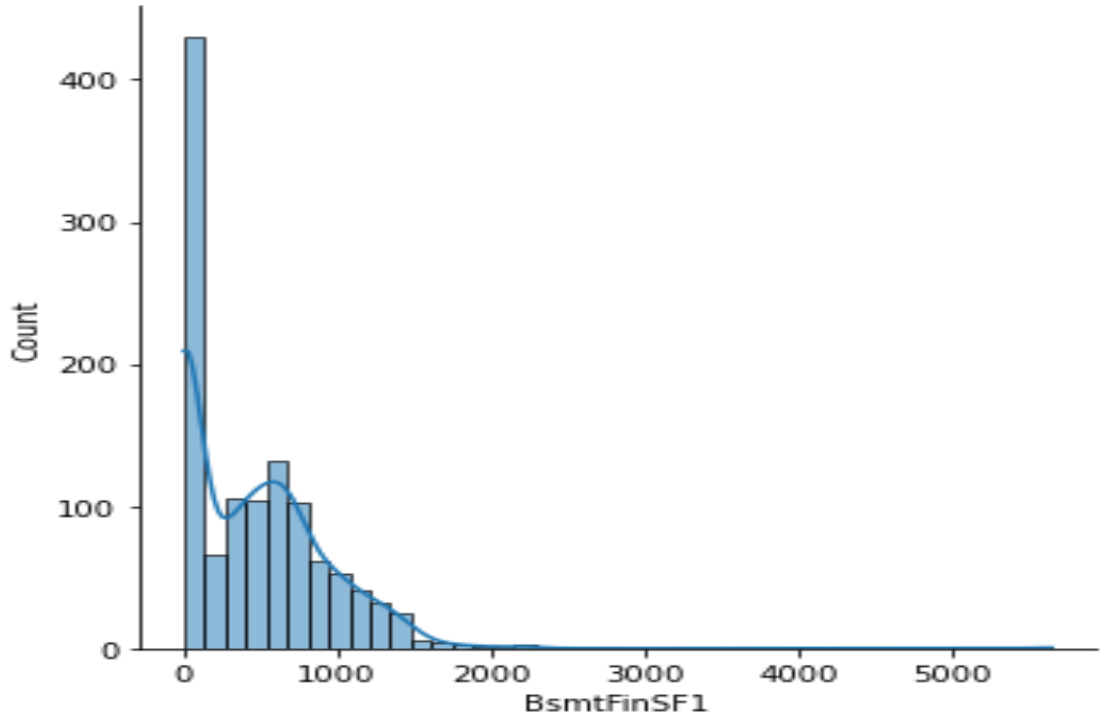


31.

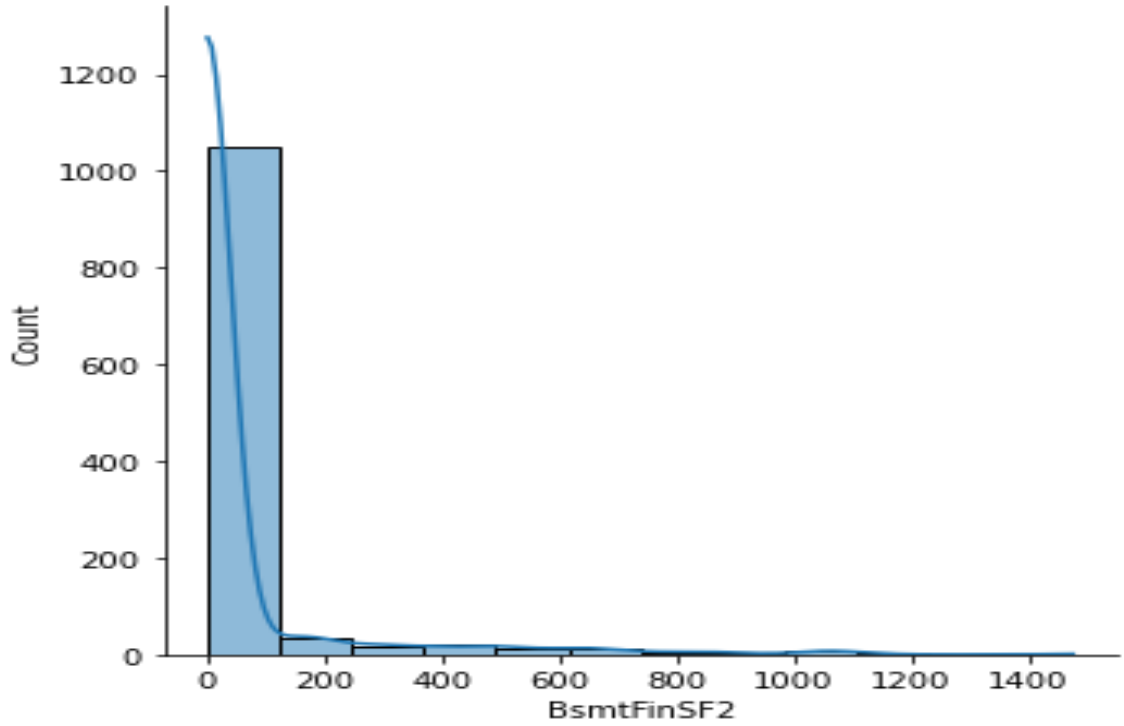
```
Unf     345
GLQ     330
ALQ     174
BLQ     121
Rec     109
LwQ      59
         30
```



32.

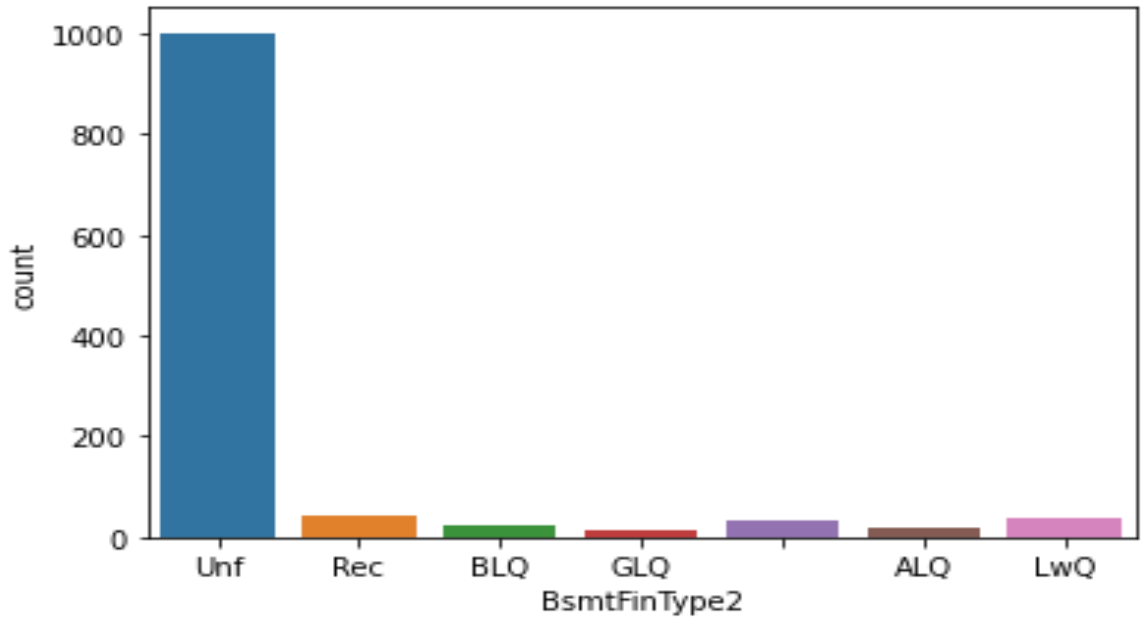


33.

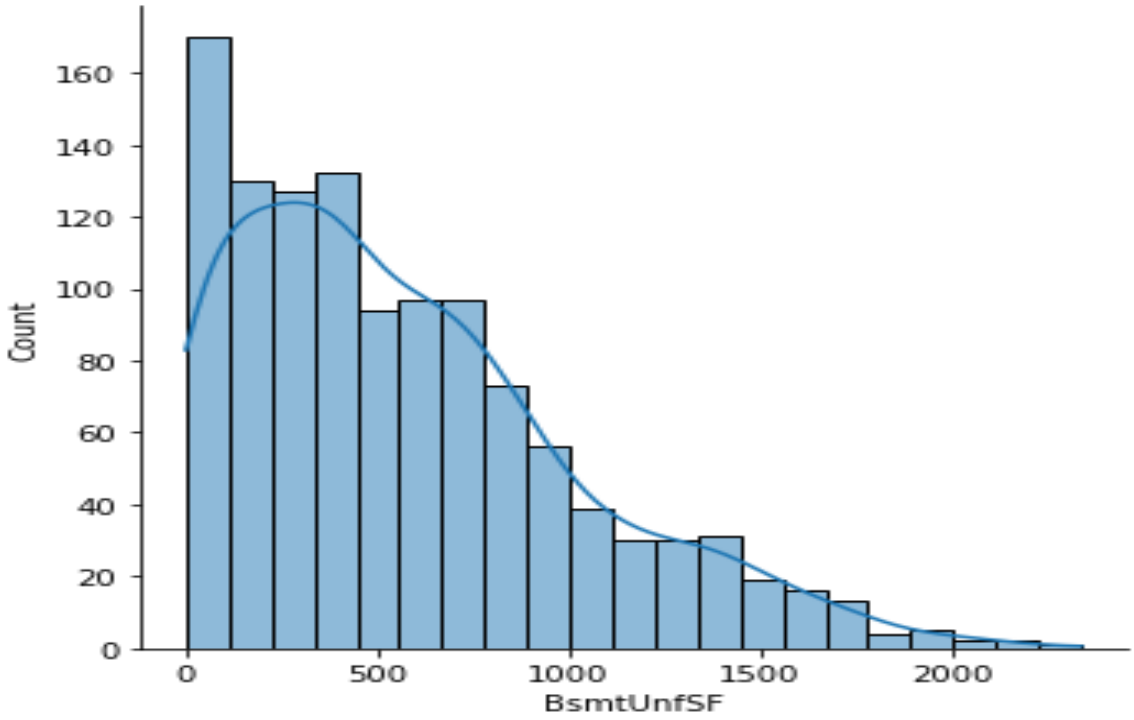


34.

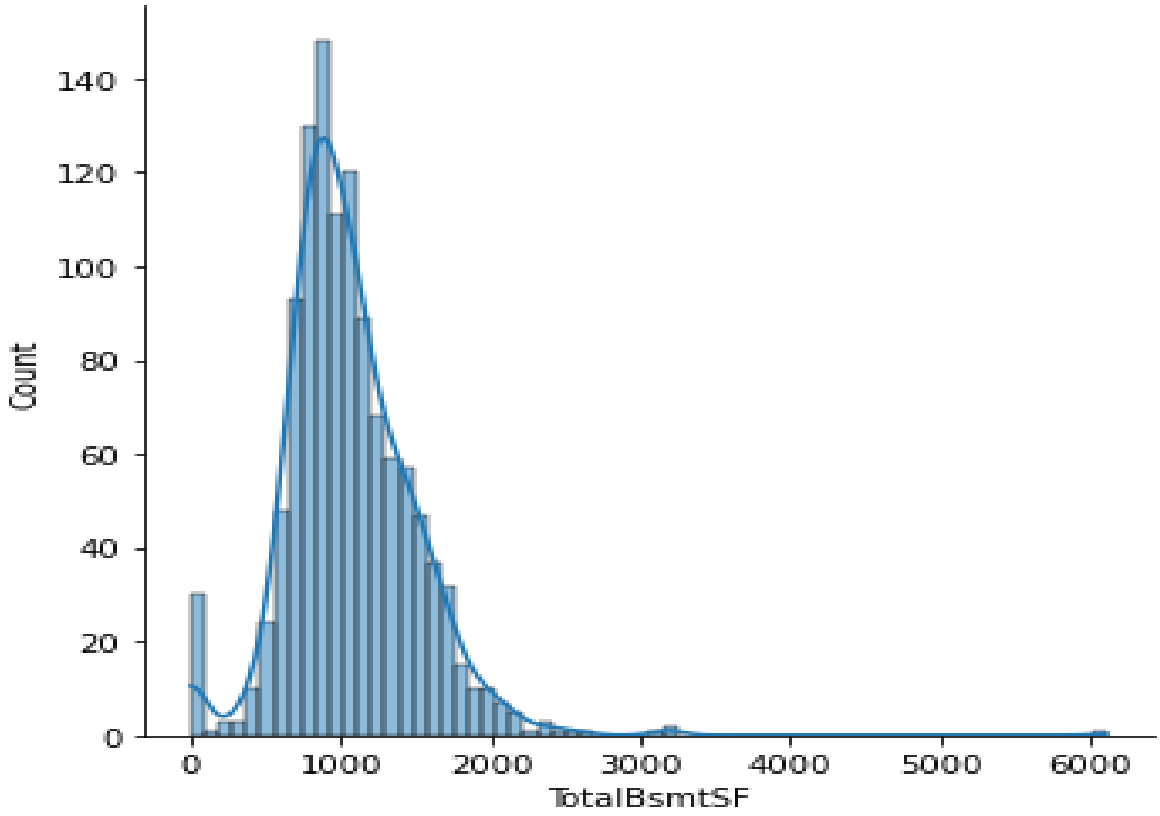
```
Unf     1002
Rec       43
LwQ       40
          31
BLQ       24
ALQ       16
GLQ       12
```



35.

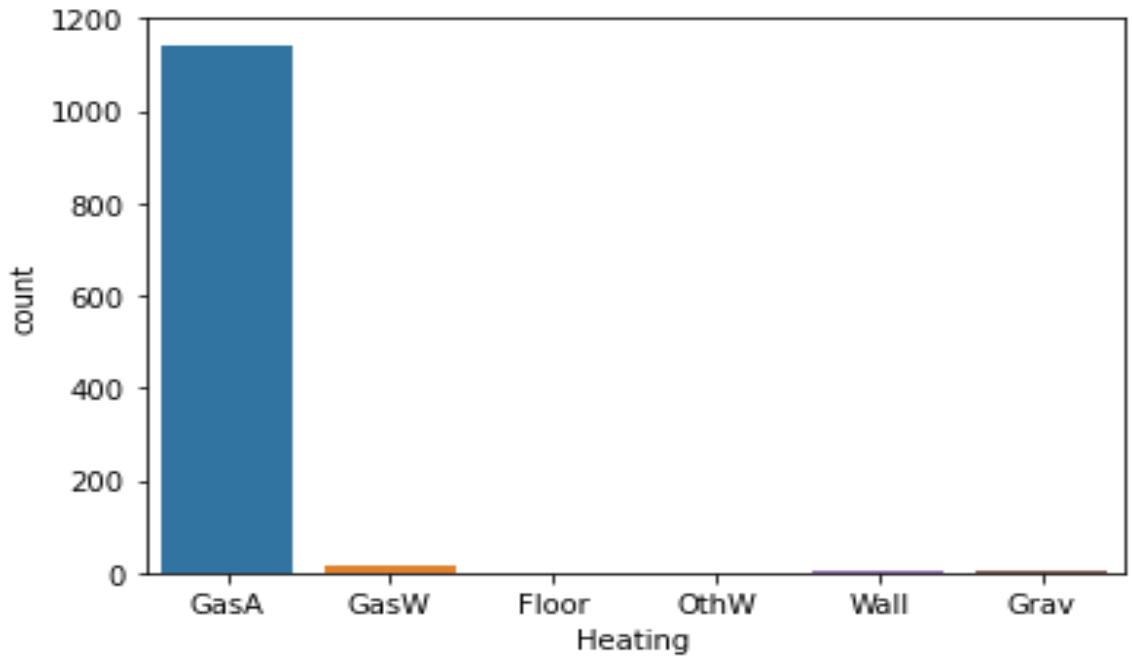

160
140
120
100
80
60
40
20
0
Count
0
500
1000
1500
2000
BsmtUnfSF

36.


140
120
100
80
60
40
20
0
Count
0
1000
2000
3000
4000
5000
6000
TotalBsmtSF

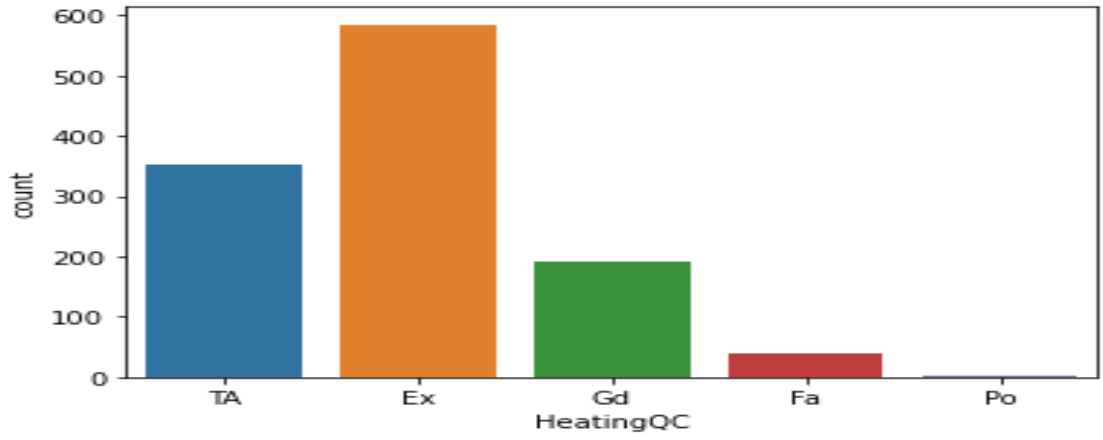
37.

```
GasA      1143
GasW        14
Grav         5
Wall         4
Floor        1
OthW         1
```



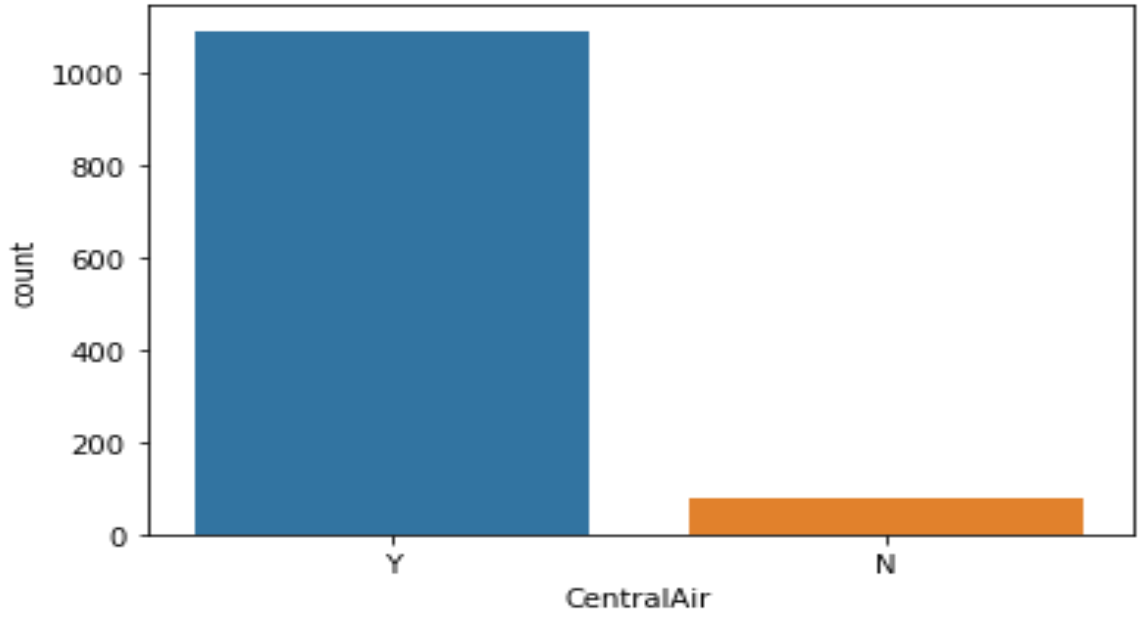
38.

```
Ex     585
TA     352
Gd     192
Fa      38
Po       1
```



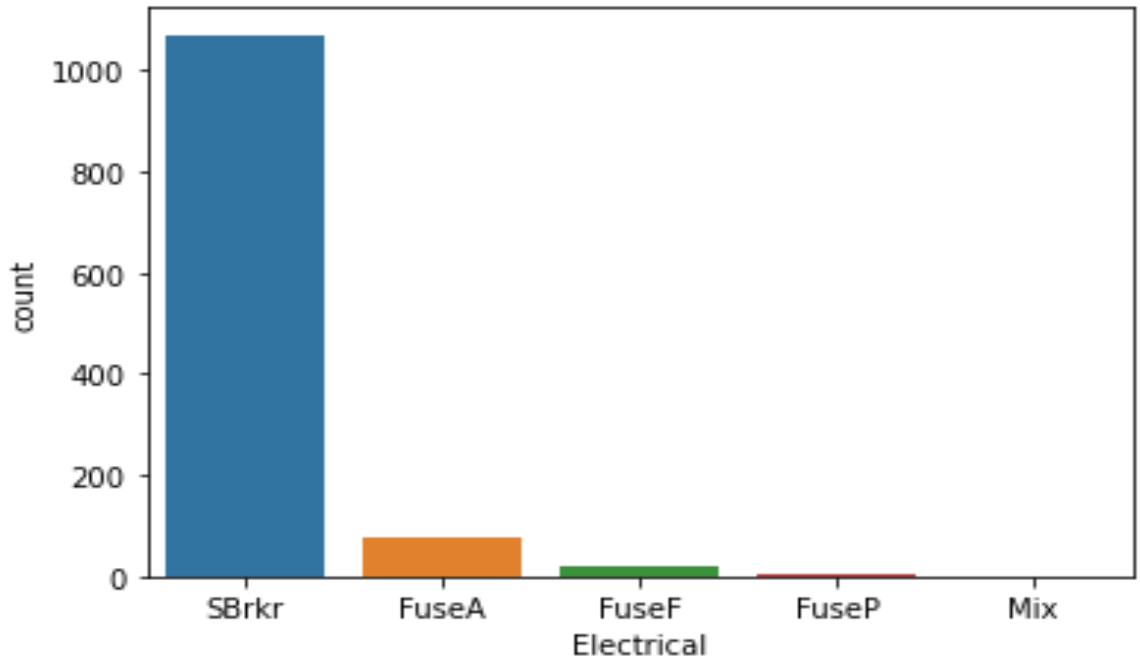
39.

```
Y    1090
N      78
```

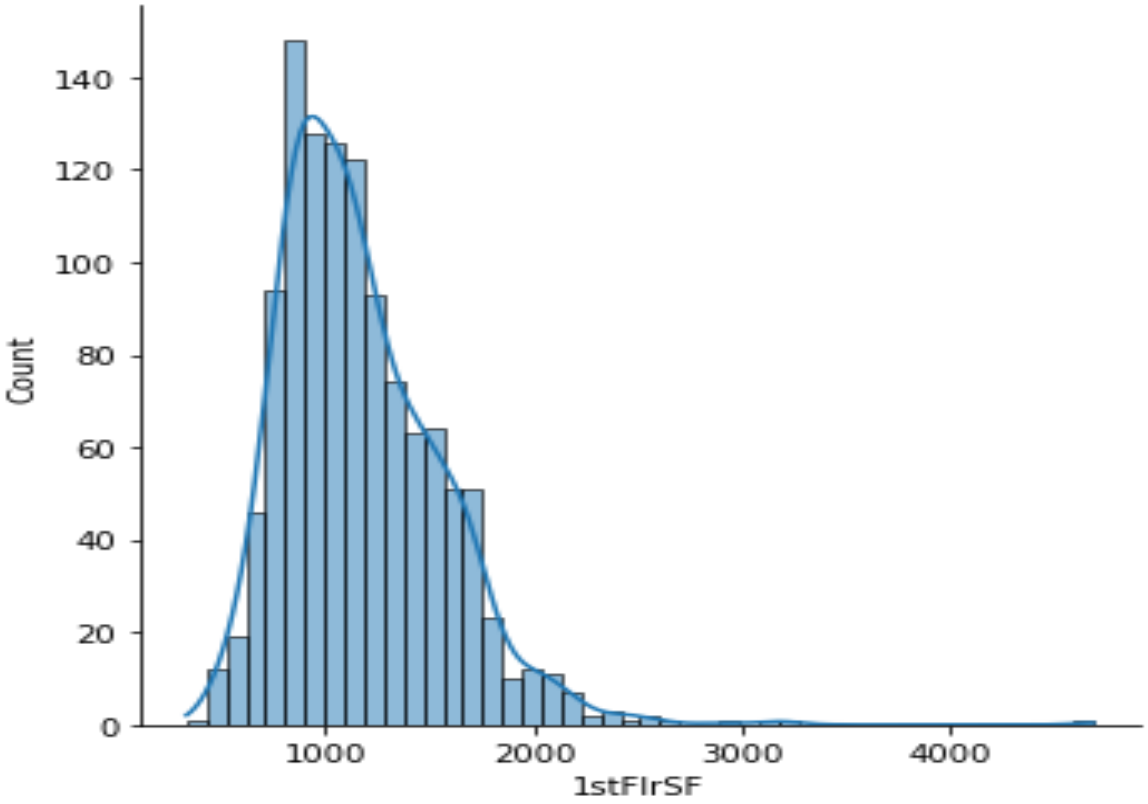


40. Column41-

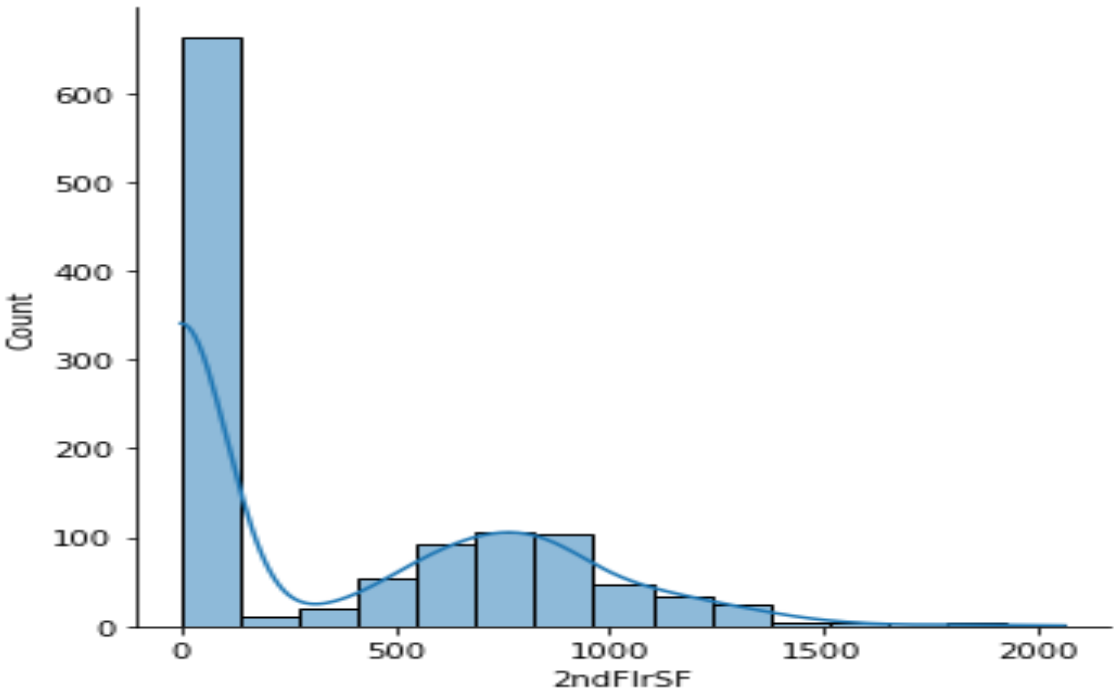
```
SBrkr    1070
FuseA      74
FuseF      21
FuseP       2
Mix         1
```



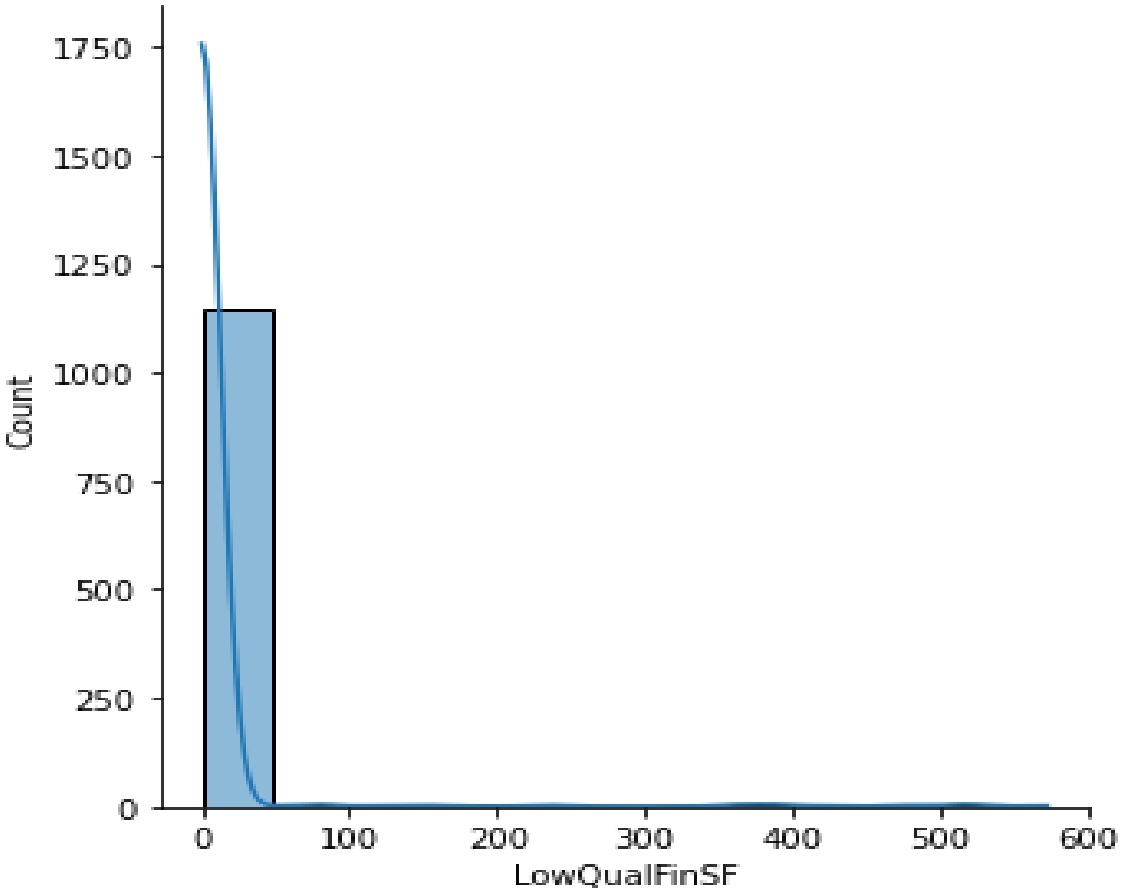
41.


140
120
100
80
60
40
20
0
Count
1000
2000
3000
4000
1stFlrSF

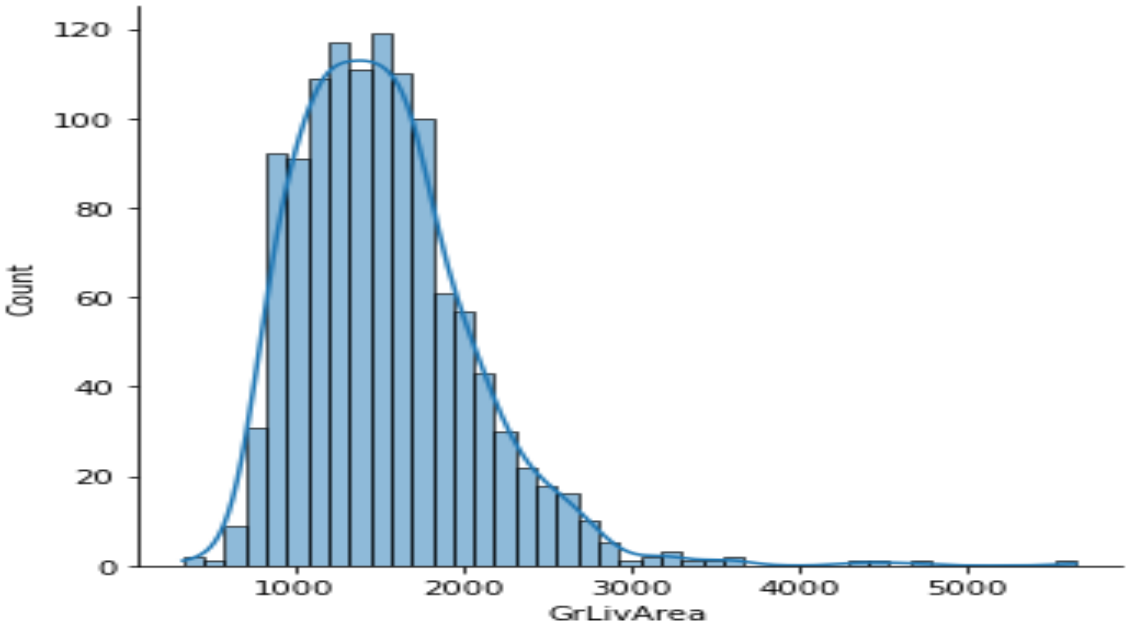
42.


600
500
400
300
200
100
0
Count
0
500
1000
1500
2000
2ndFlrSF

43.

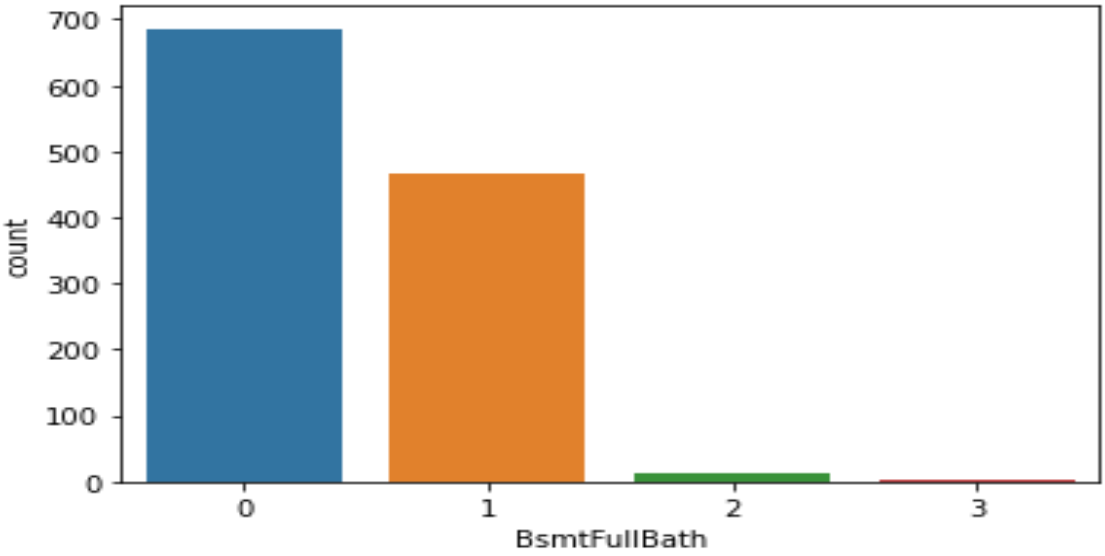

1750
1500
1250
1000
750
500
250
0
Count
0
100
200
300
400
500
600
LowQualFinSF

44.


120
100
80
60
40
20
0
Count
1000
2000
3000
4000
5000
GrLivArea

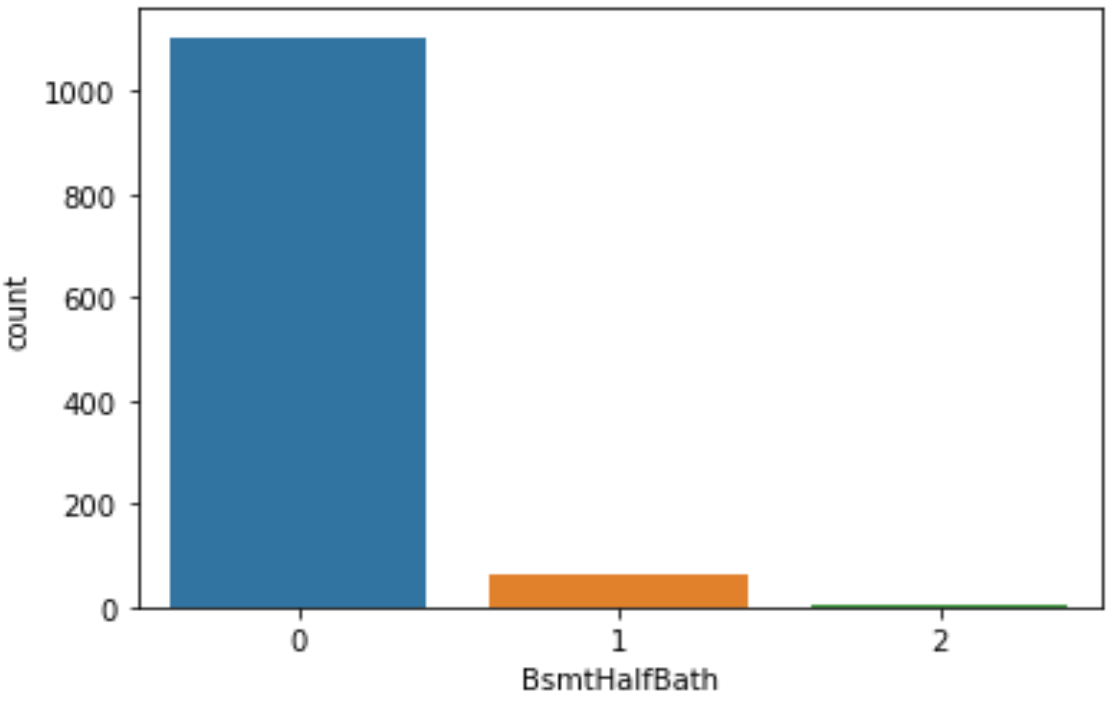
45.

```
0     686
1     468
2      13
3       1
```



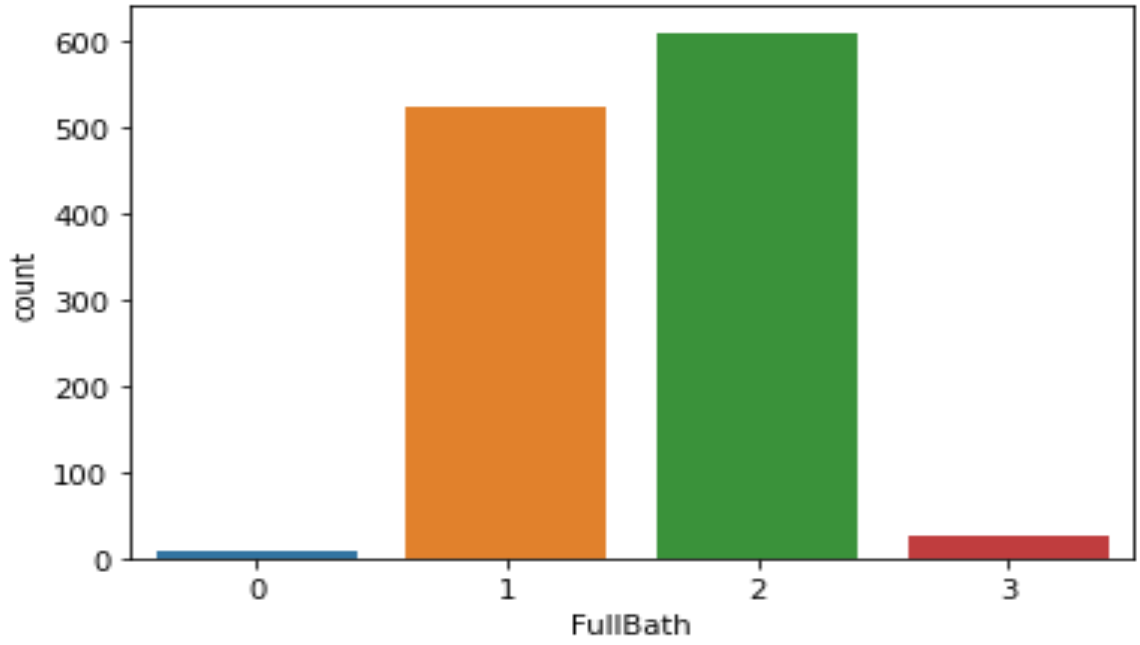
46.

```
0     1105
1       61
2        2
```



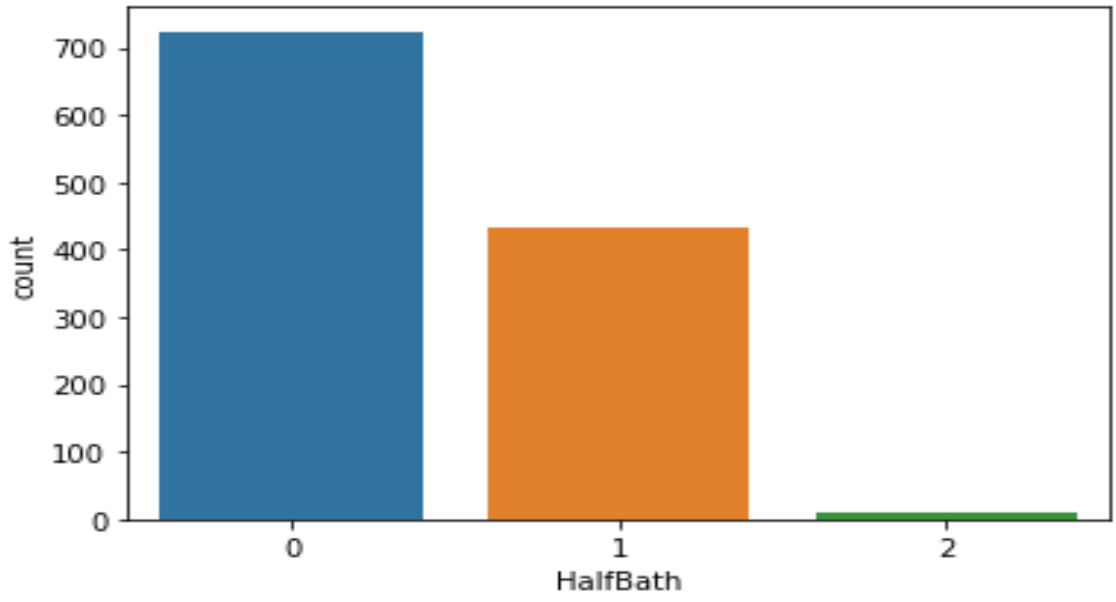
47.

```
2    610
1    524
3     27
0      7
```



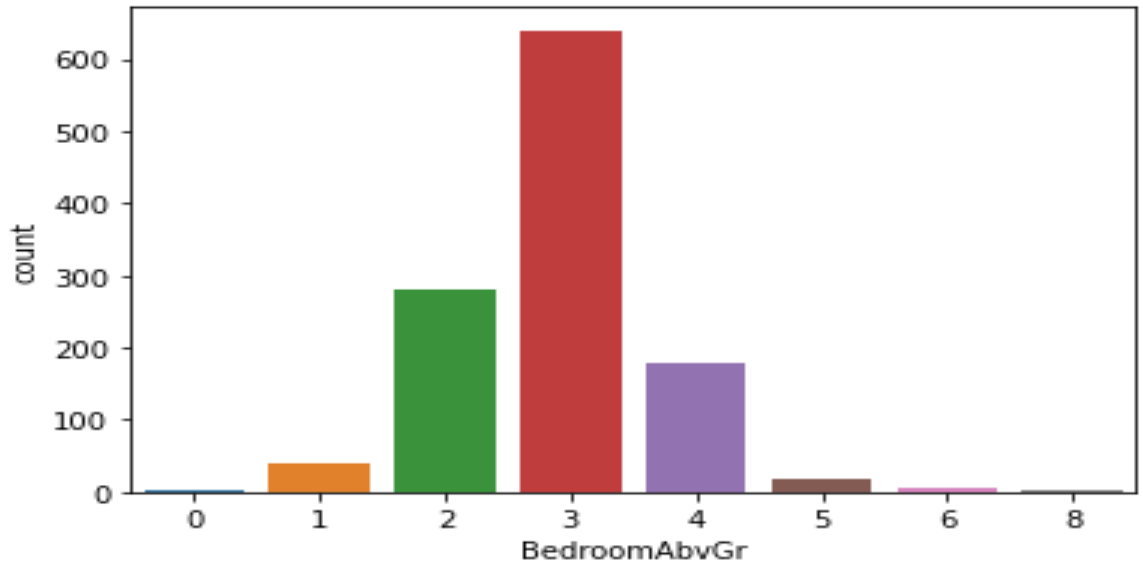
48.

```
0    724
1    434
2     10
```



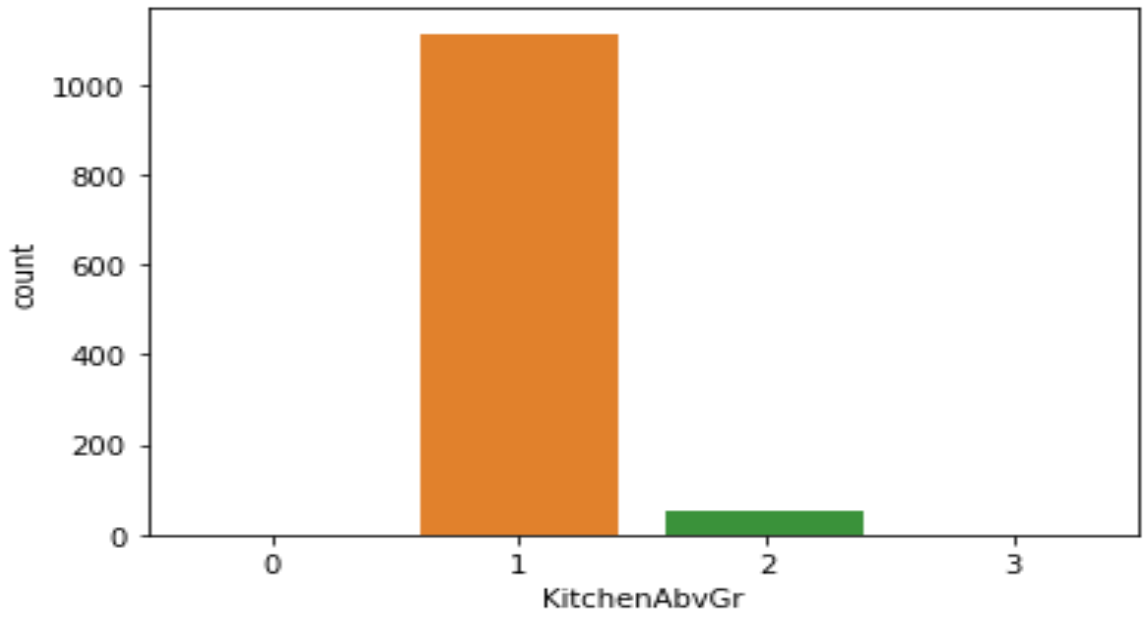
49.

```
3     640
2     281
4     180
1      39
5      18
6       5
0       4
8       1
```



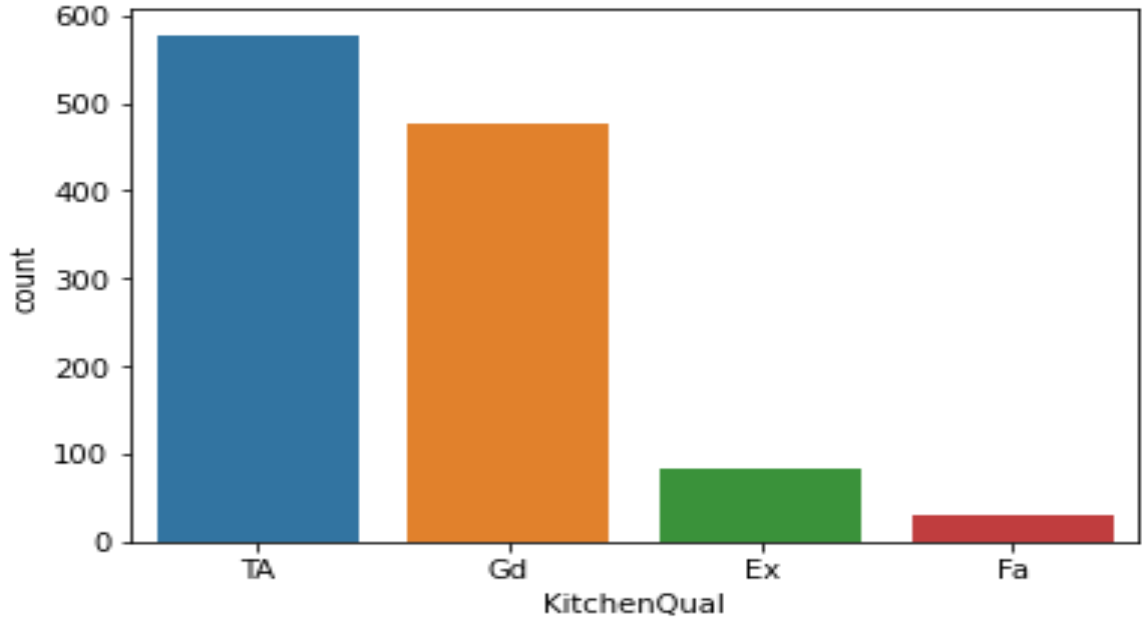
50.

```
A1     1114
2        52
3         1
0         1
```



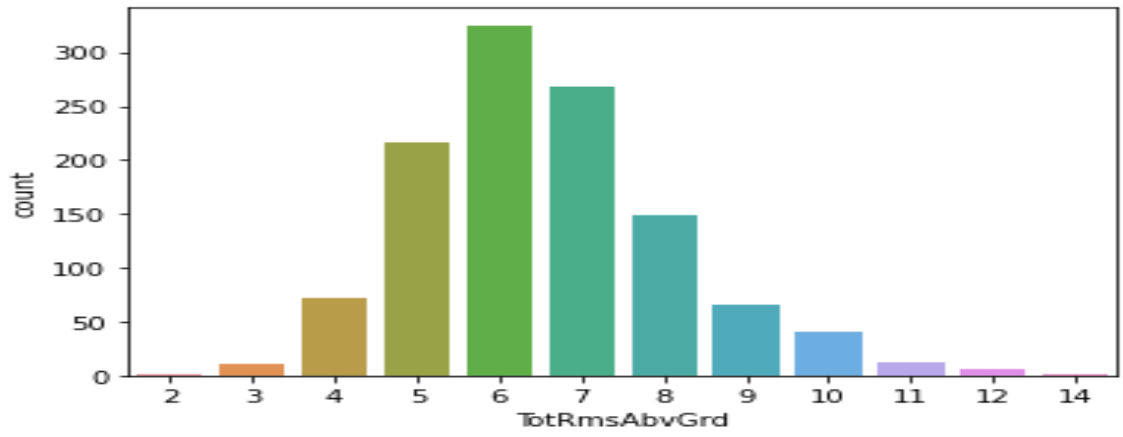
51.

```
TA     578
Gd     478
Ex      82
Fa      30
```



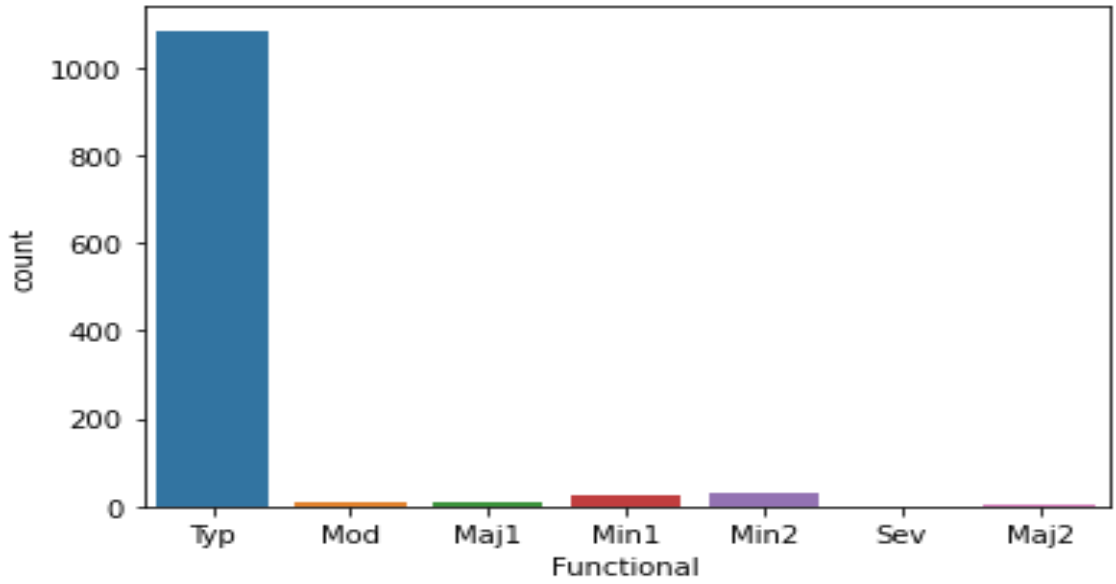
52.

```
6      325
7      268
5      217
8      148
4       72
9       65
10      41
11      13
3       11
12       6
2        1
14       1
```



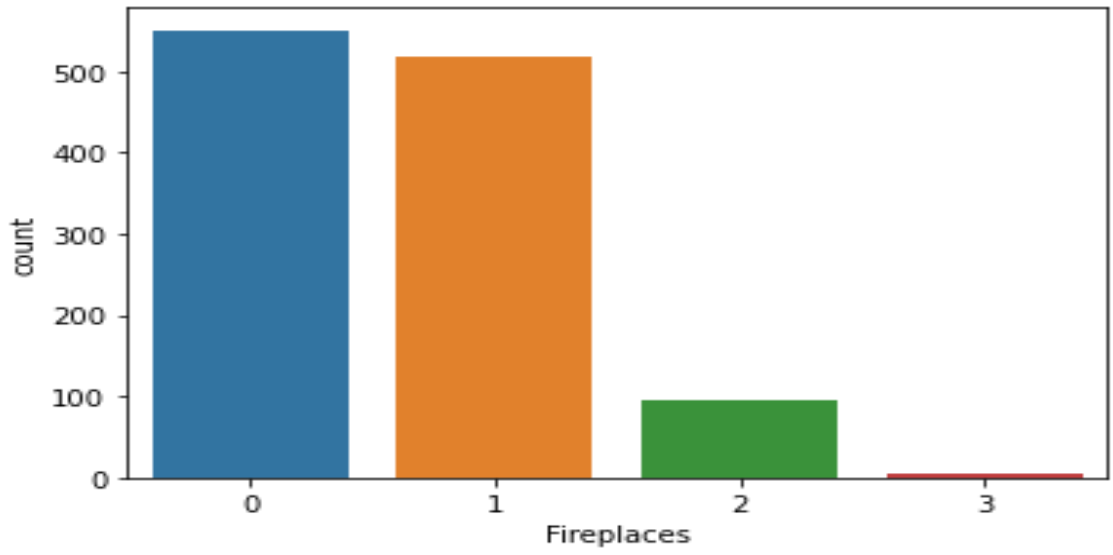
53.

```
Typ      1085
Min2       30
Min1       25
Mod        12
Maj1       11
Maj2        4
Sev         1
```



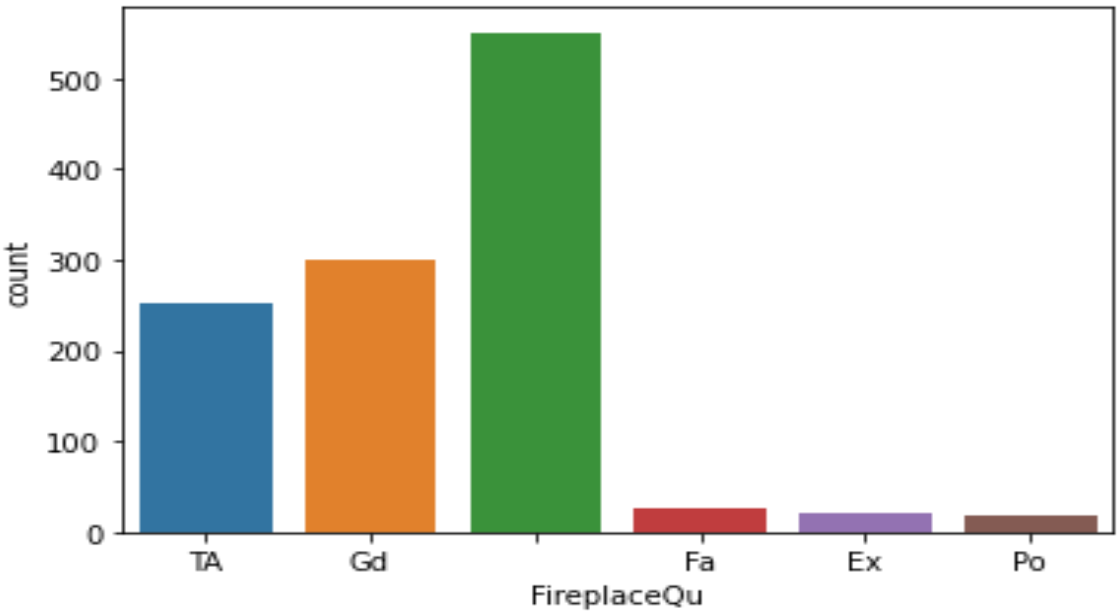
54.

```
0     551
1     518
2      94
3       5
```



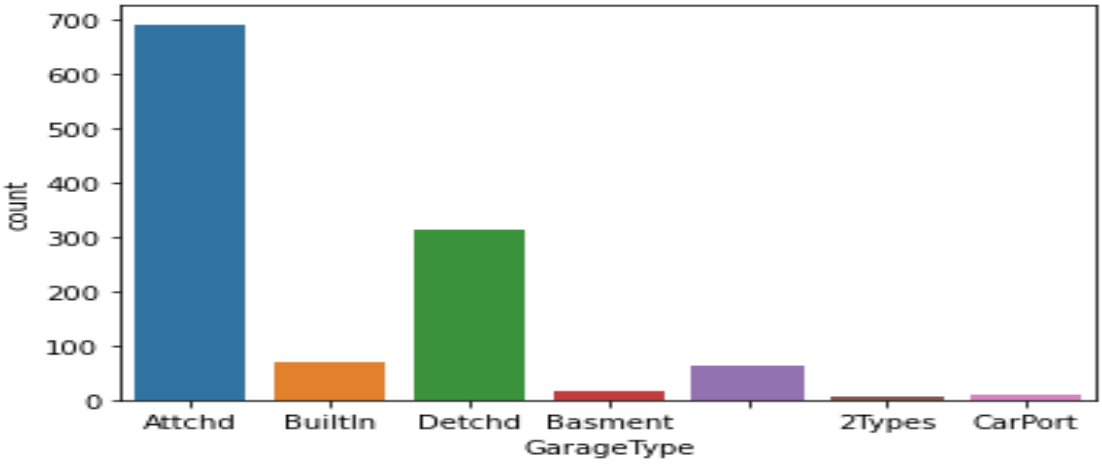
55.

```
      551
Gd    301
TA    252
Fa     25
Ex     21
Po     18
```



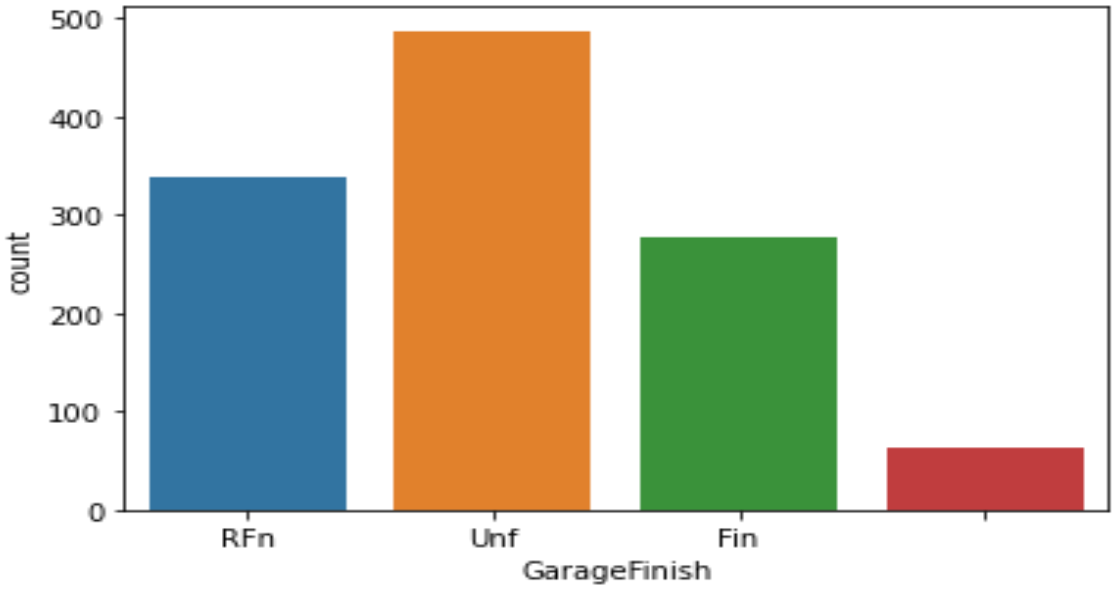
56.

```
Attchd     691
Detchd     314
BuiltIn     70
            64
Basment     16
CarPort      8
2Types       5
```



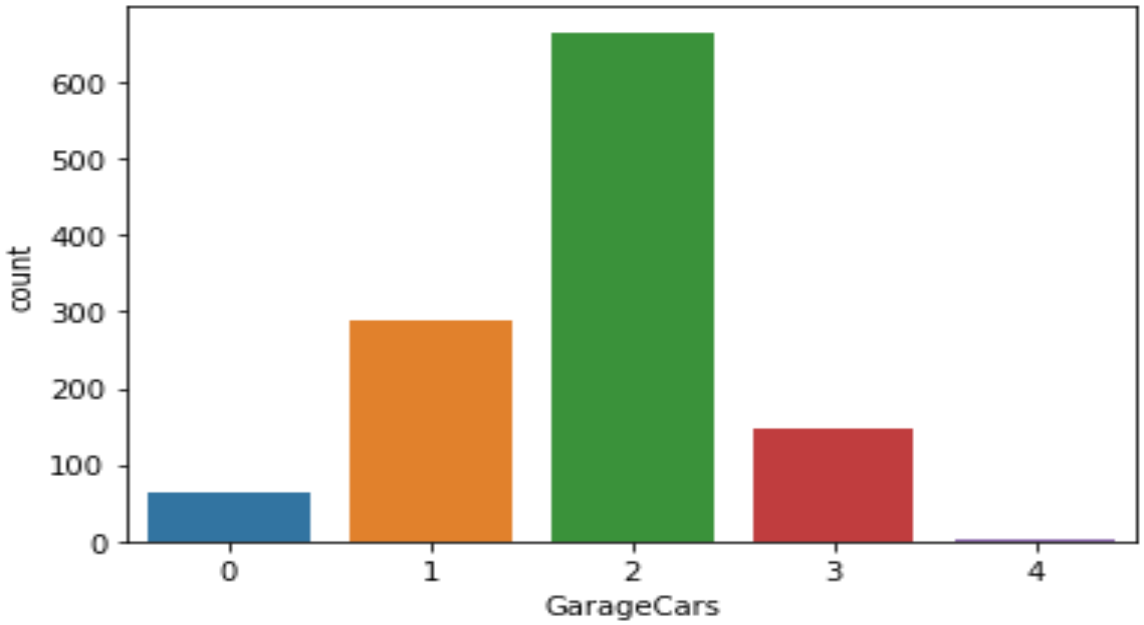
57.

```
Unf     487
RFn     339
Fin     278
         64
```

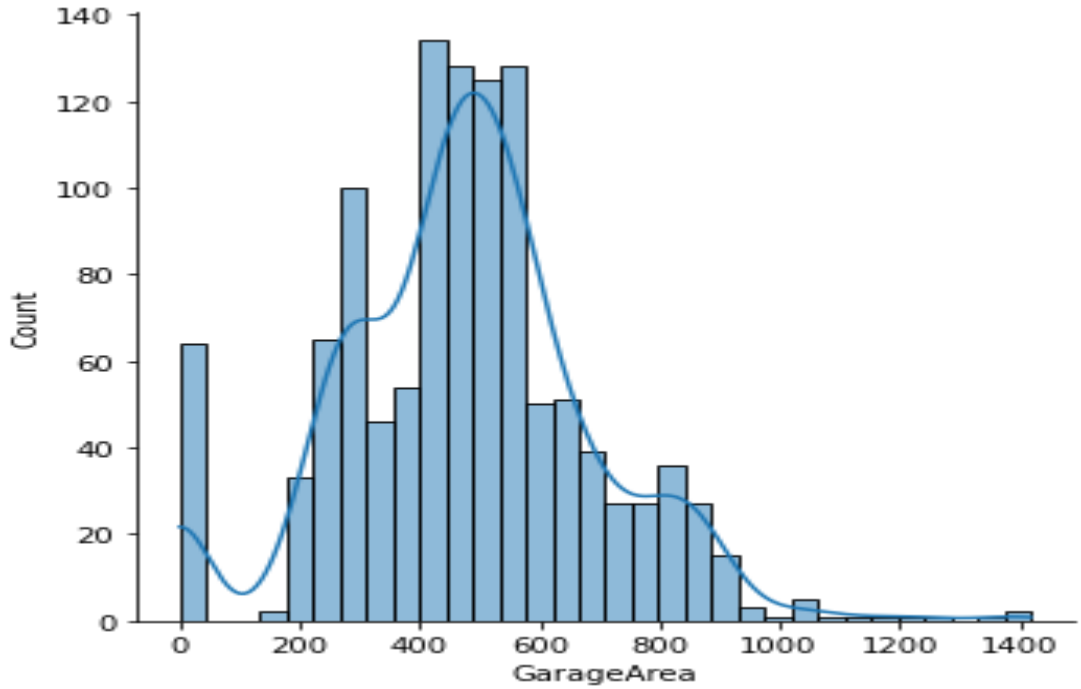


58.

```
2     665
1     288
3     147
0      64
4       4
```

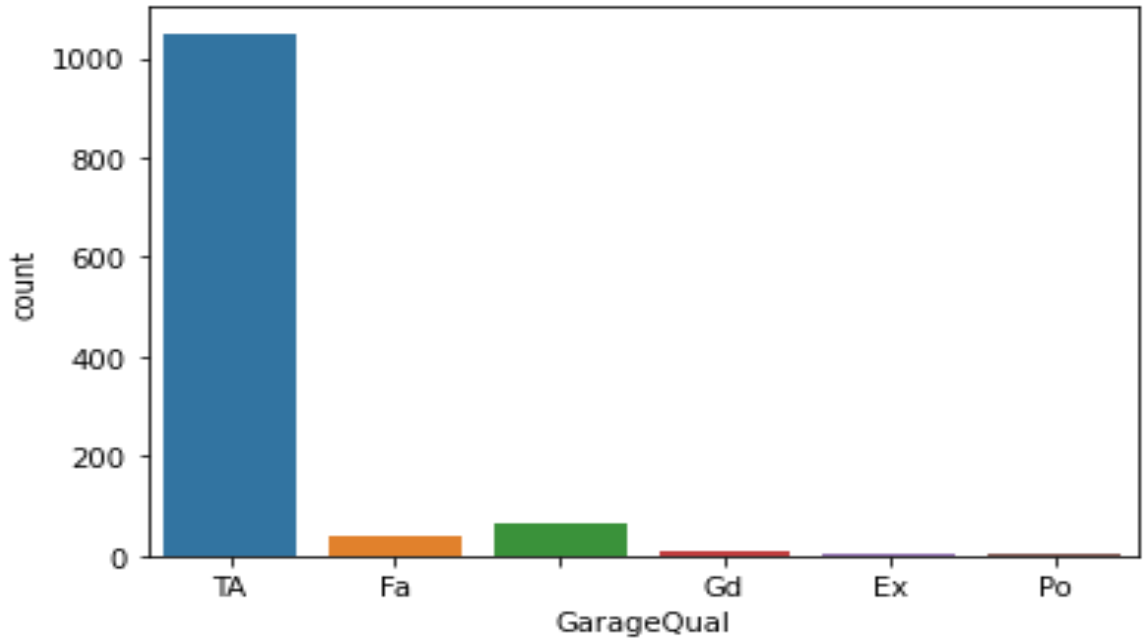


59.



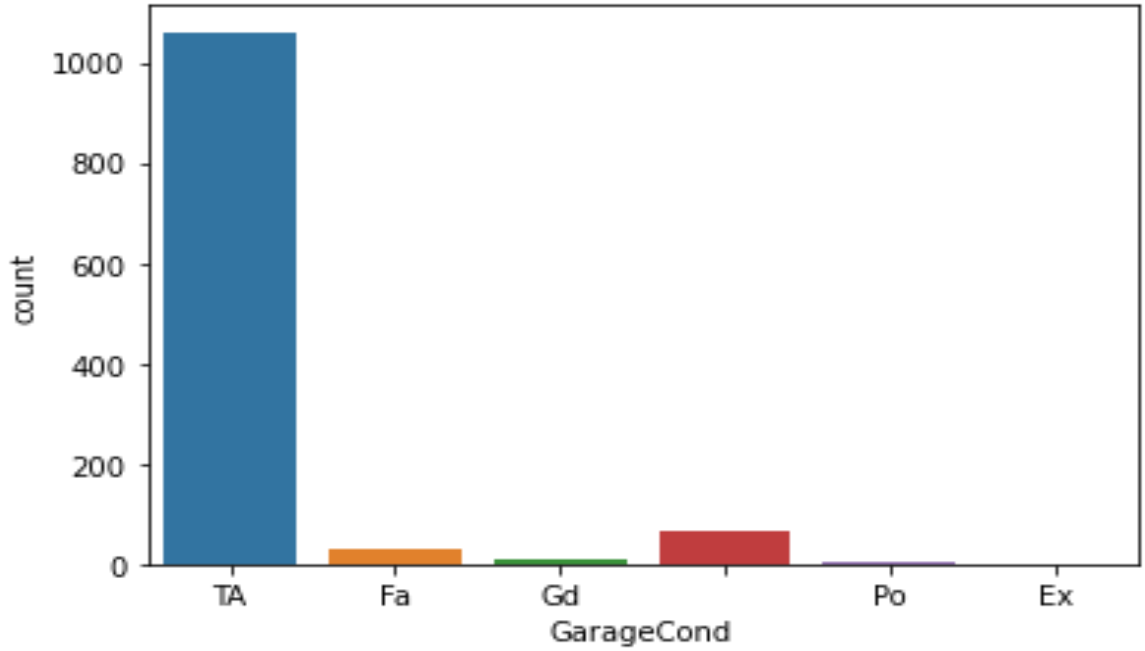
60.

```
TA      1050
         64
Fa       39
Gd       11
Ex        2
Po        2
```



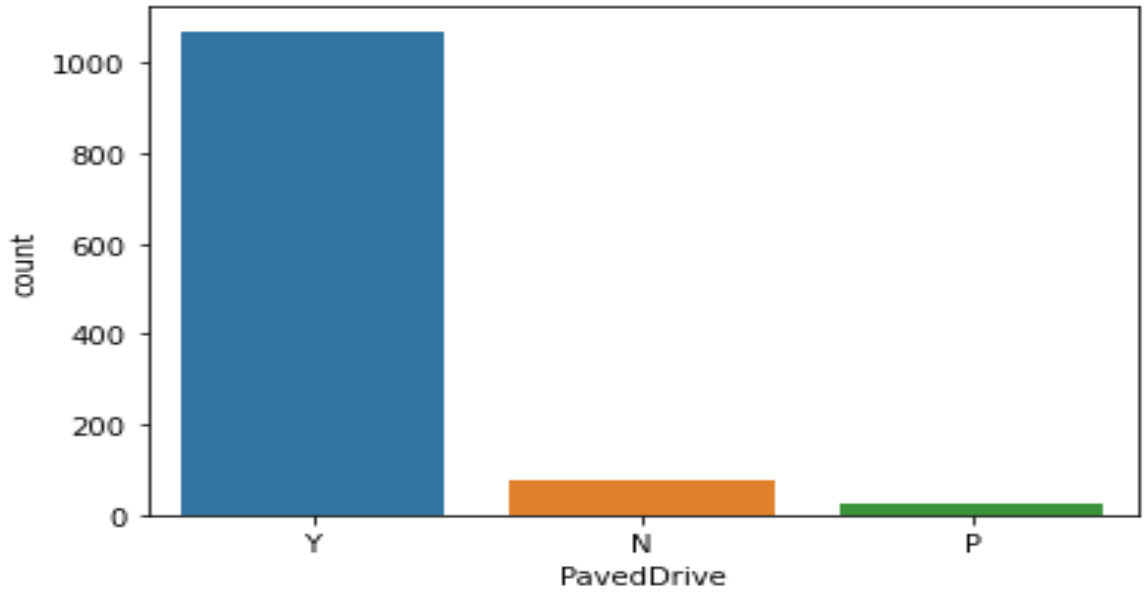
61.

```
TA     1061
        64
Fa       28
Gd        8
Po        6
Ex        1
```

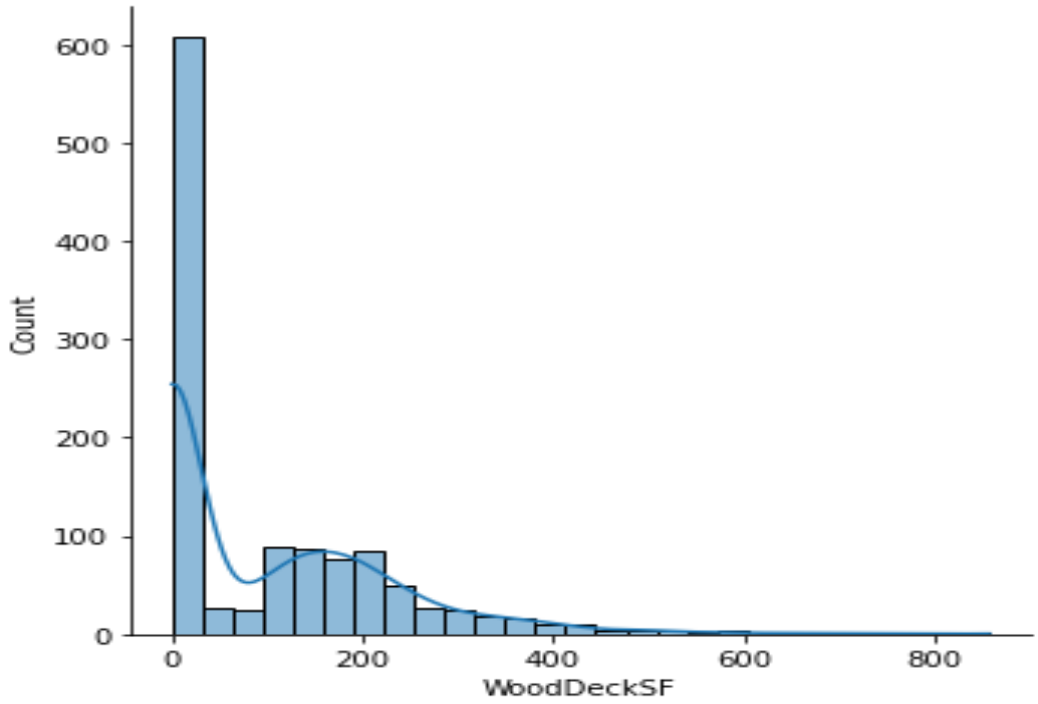


62.

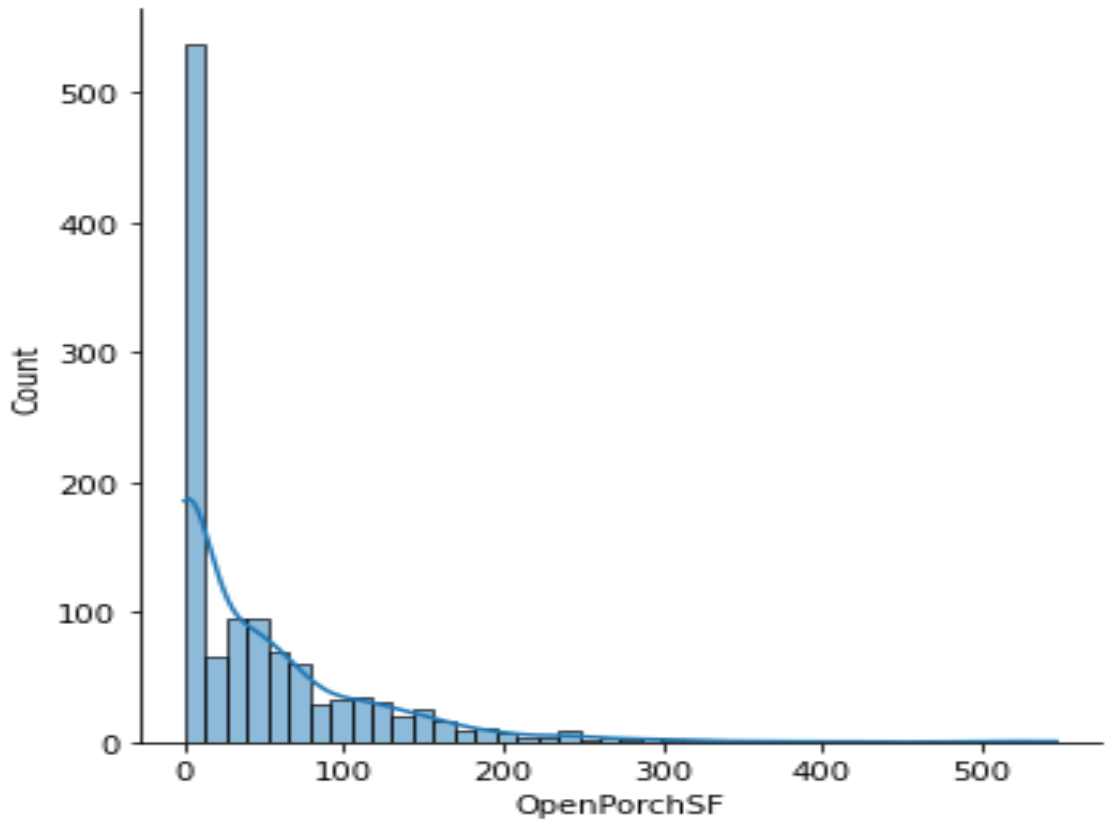
```
Y     1071
N       74
P       23
```



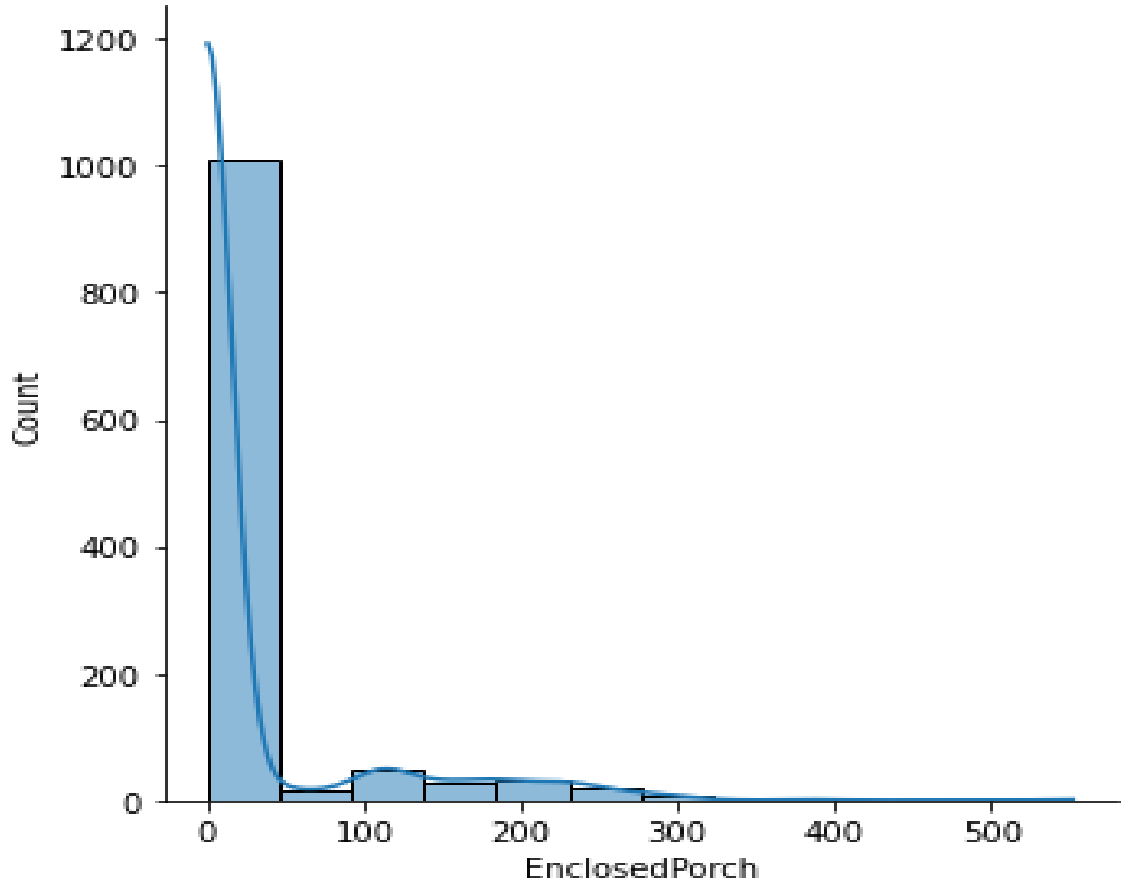
63.


600
500
400
300
200
100
0
Count
0
200
400
600
800
WoodDeckSF

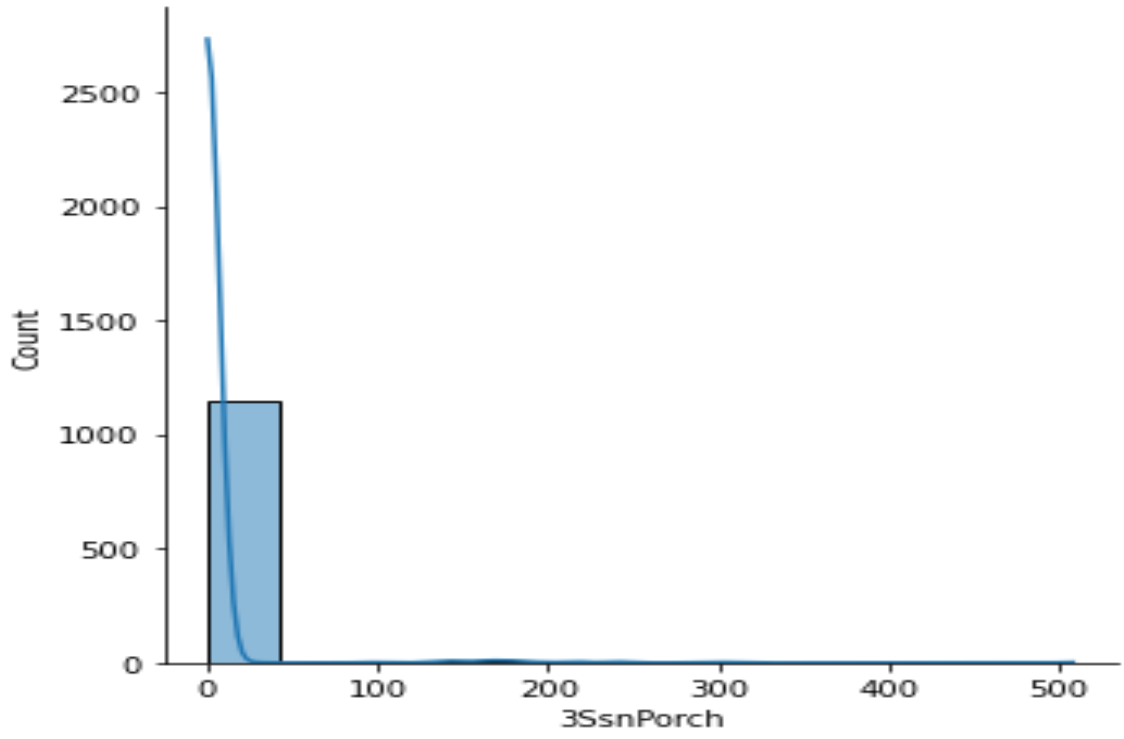
64.


500
400
300
200
100
0
Count
0
100
200
300
400
500
OpenPorchSF

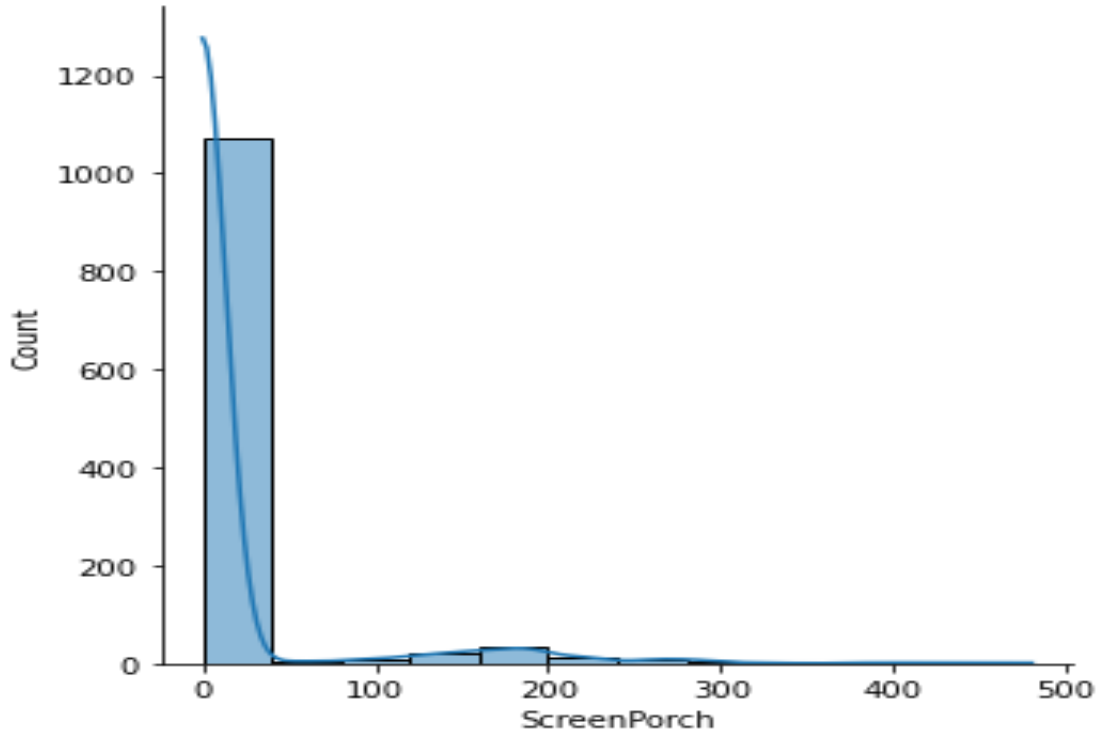
65.


1200
1000
800
600
400
200
0
Count
0
100
200
300
400
500
EnclosedPorch

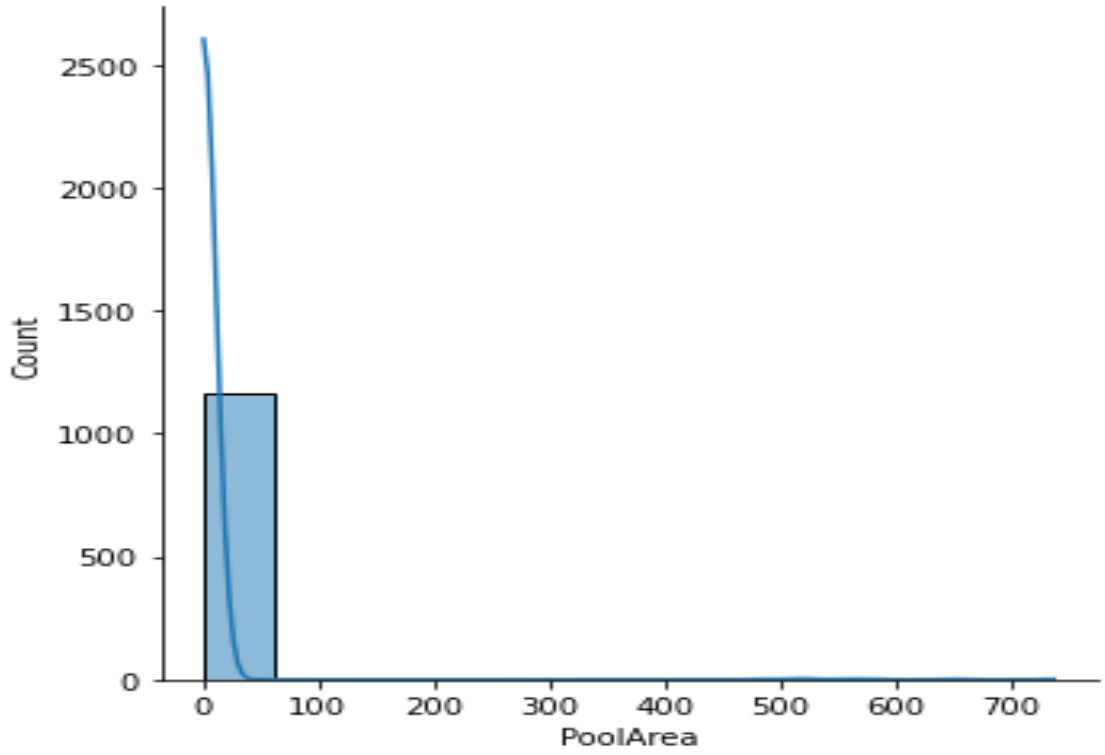
66.


2500
2000
1500
1000
500
0
Count
0
100
200
300
400
500
3SsnPorch

67.

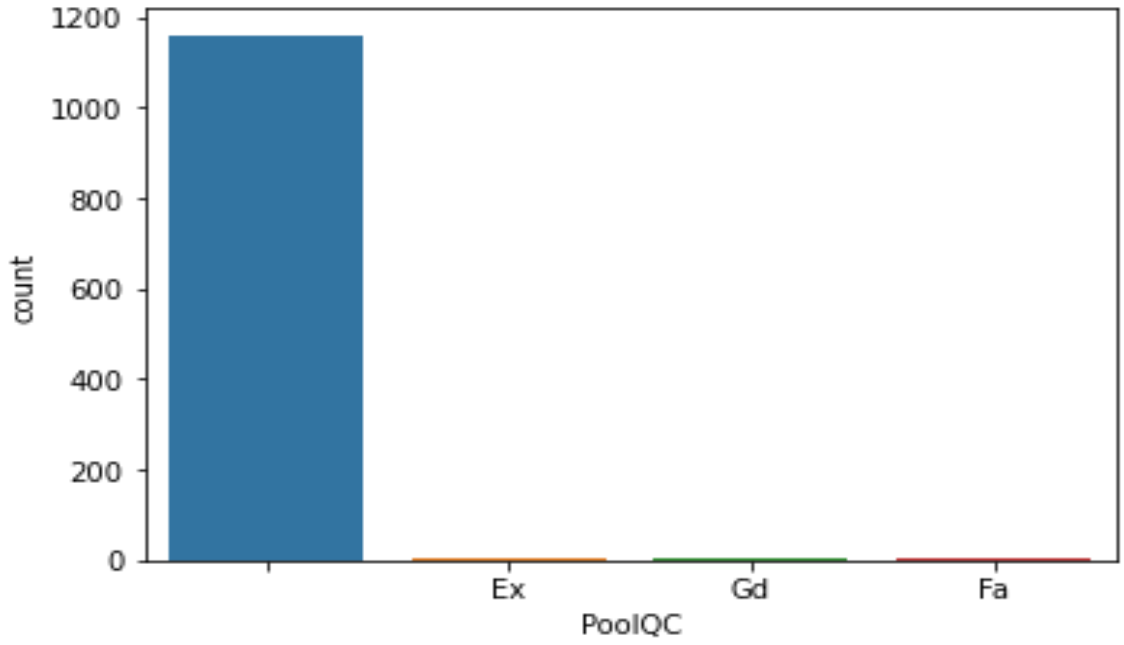

1200
1000
800
600
400
200
0
Count
0
100
200
300
400
500
ScreenPorch

68.


2500
2000
1500
1000
500
0
Count
0
100
200
300
400
500
600
700
PoolArea

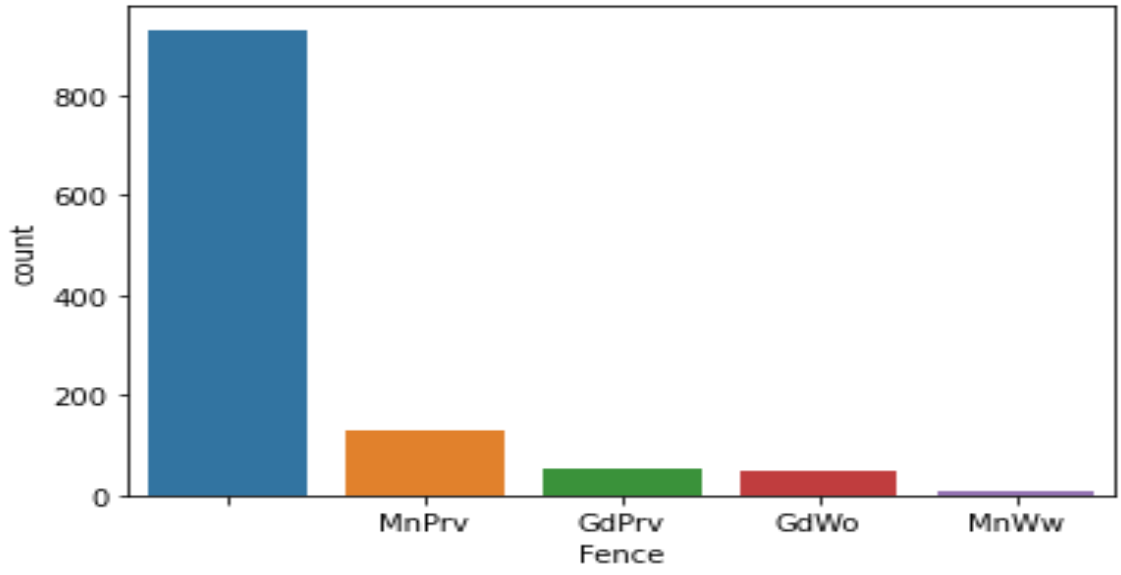
69.

```
      1161
Gd       3
Ex       2
Fa       2
```



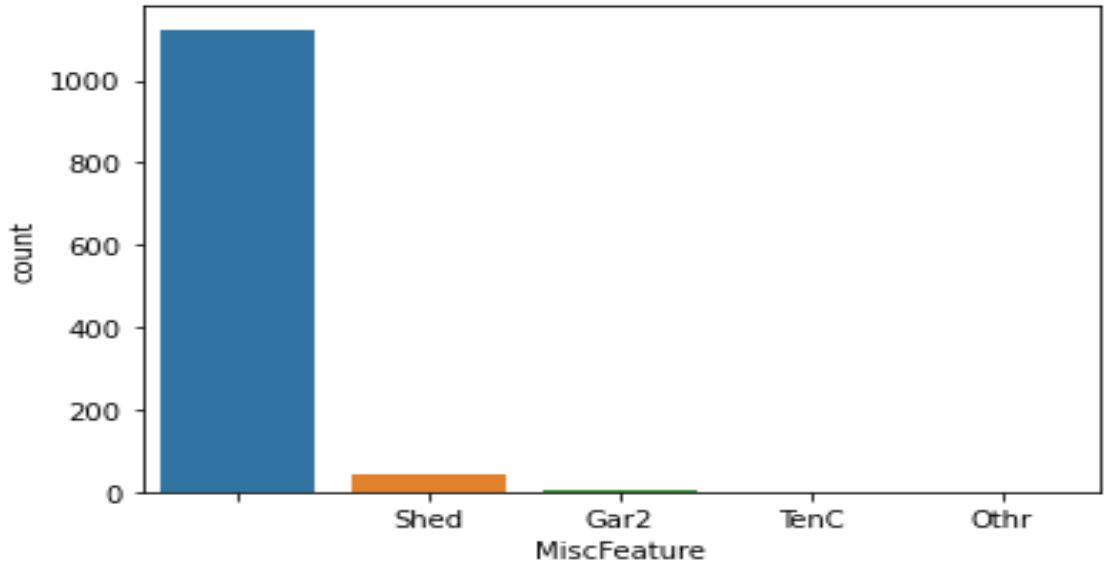
70.

```
        931
MnPrv   129
GdPrv    51
GdWo     47
MnWw     10
```

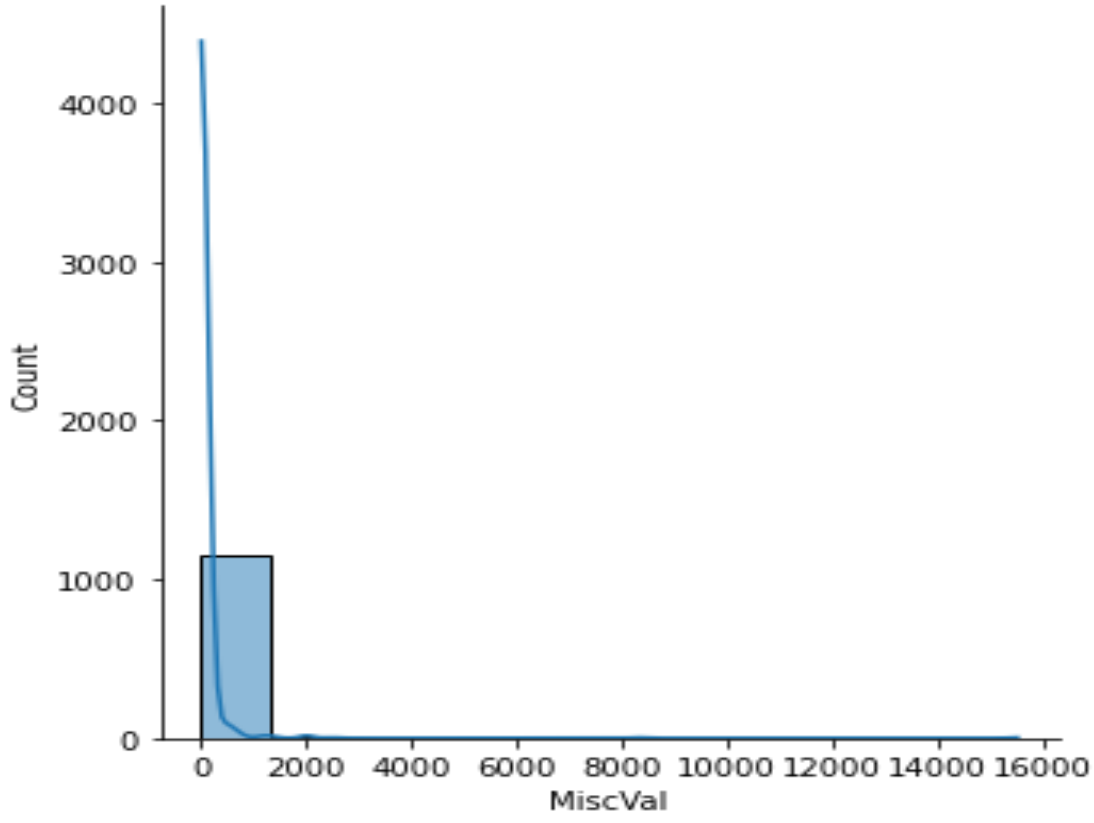


71.

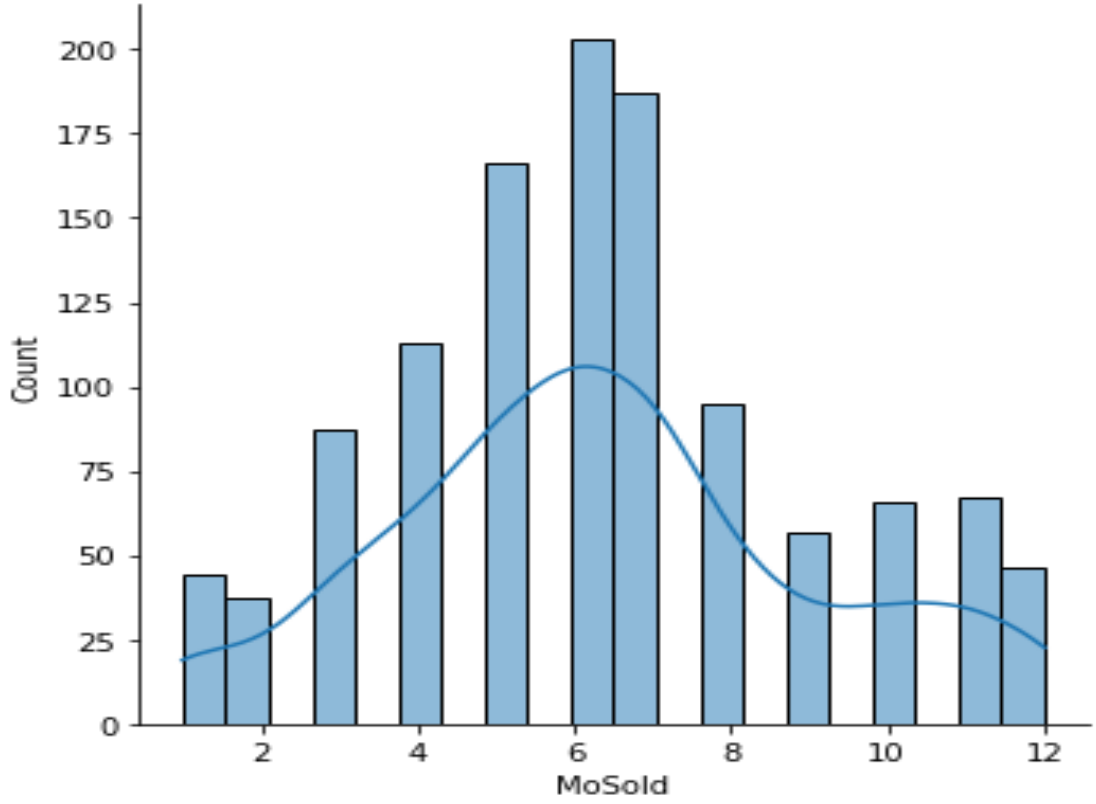
```
       1124
Shed     40
Gar2      2
TenC      1
Othr      1
```



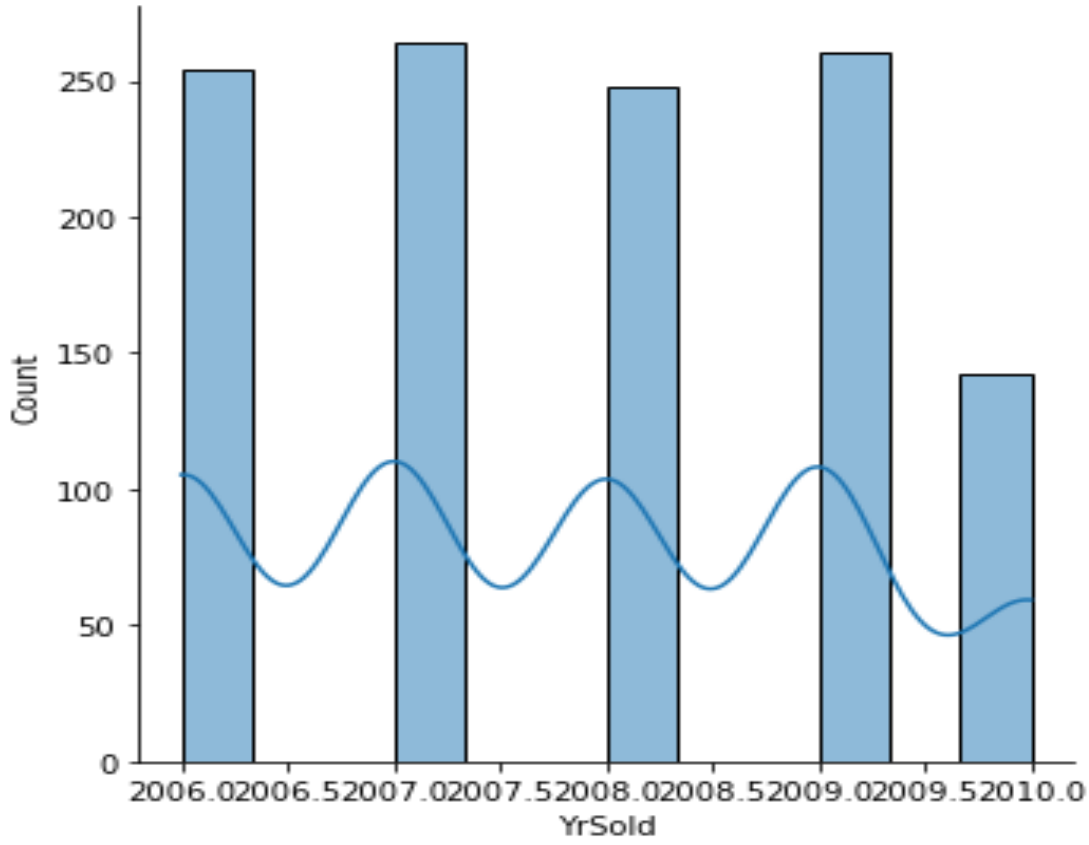
72.



73.

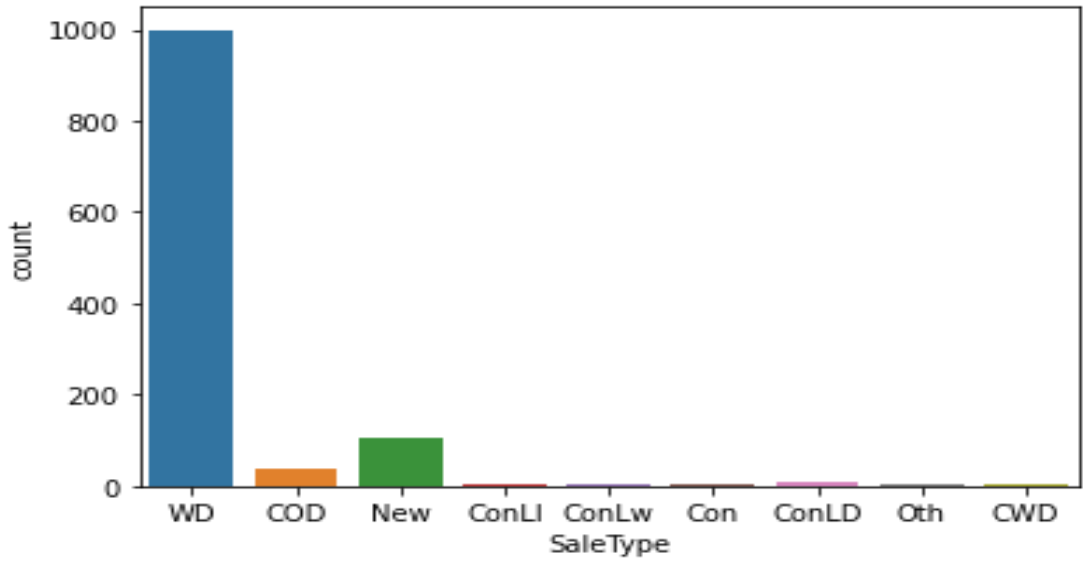

Count
MoSold
0
25
50
75
100
125
150
175
200
2
4
6
8
10
12

74.


Count
YrSold
0
50
100
150
200
250
2006.0
2006.5
2007.0
2007.5
2008.0
2008.5
2009.0
2009.5
2010.0

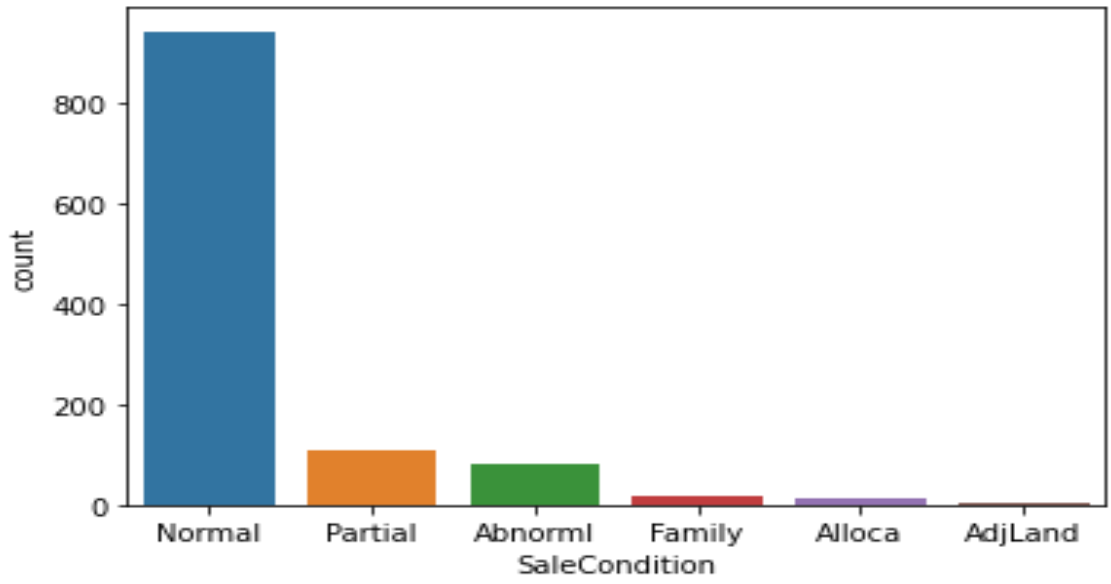
75.

```
WD        999
New       106
COD        38
ConLD       8
ConLI       5
ConLw       4
Oth         3
CWD         3
Con         2
```

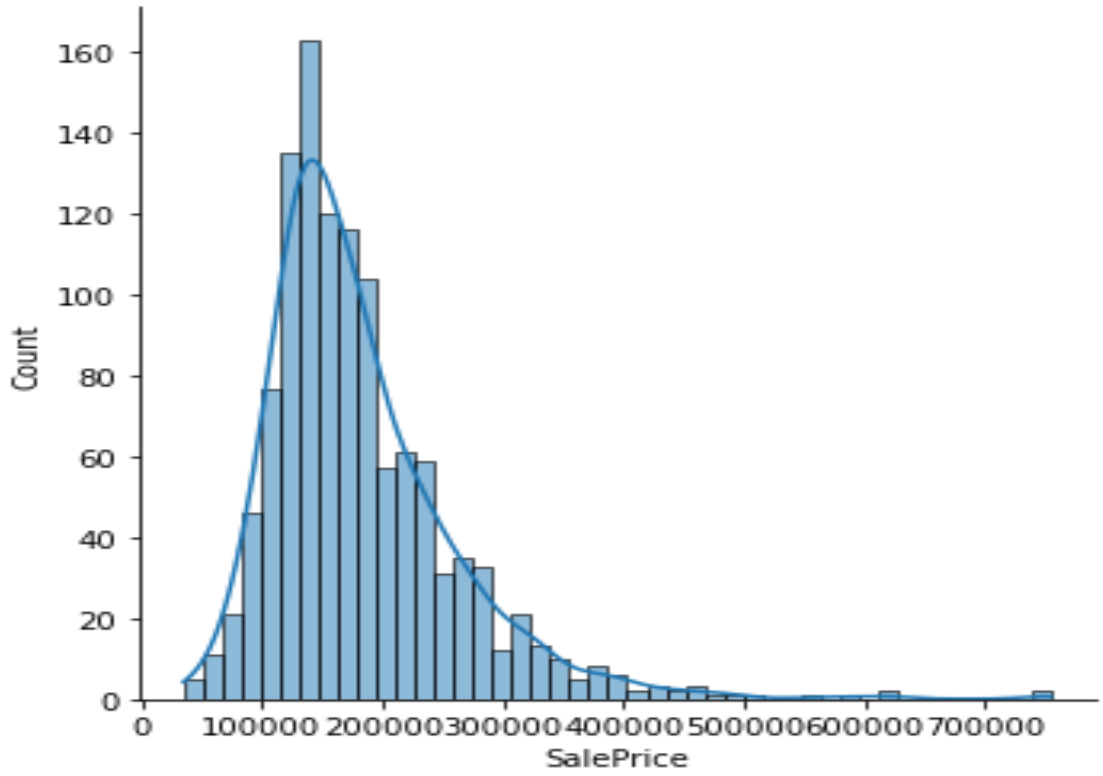


76.

```
Normal     945
Partial    108
Abnorml     81
Family      18
Alloca      12
AdjLand      4
```

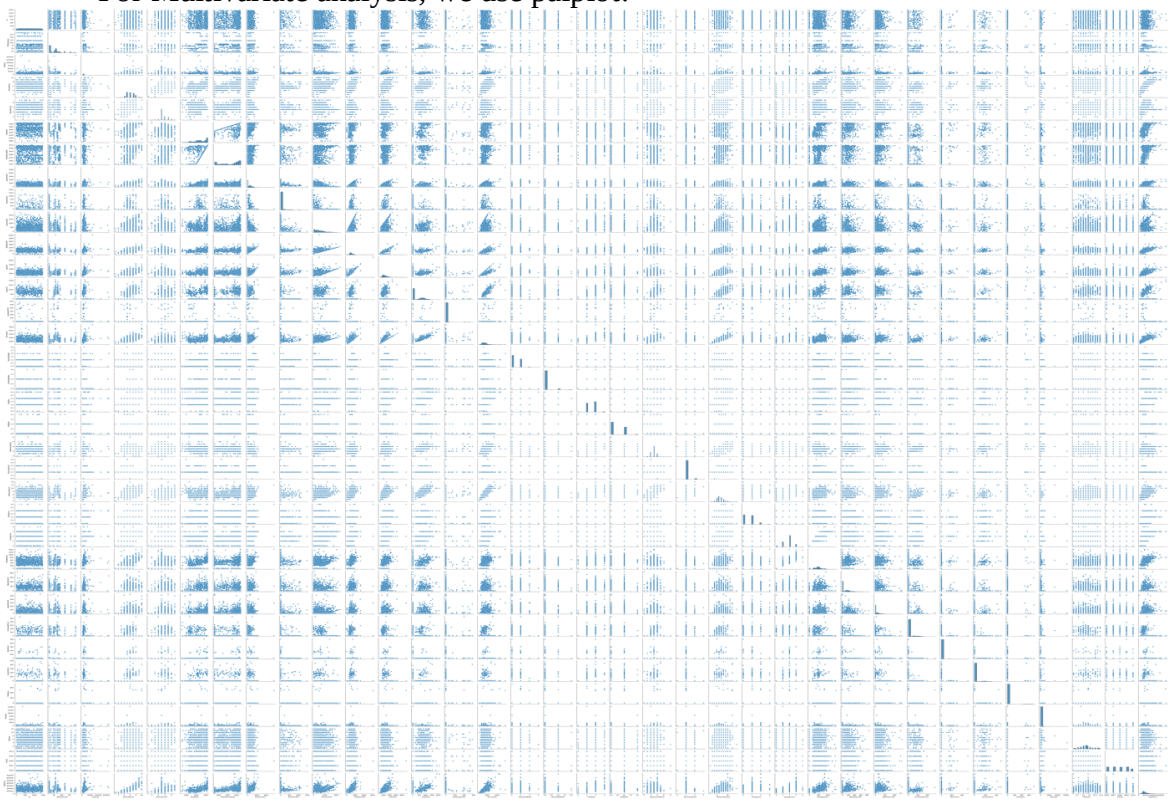


77.



We see that house sale price is mainly between 1.5L to 3L.

For Multivariate analysis, we use paiplot:



3.3 ENCODING OF DATAFRAME

Encoding a dataframe means changing the data type of a particular column to the required type as the dataset demands. There are various types on Encoding techniques:

I. Classic Encoders

We started with the most basic techniques, classic encoders. As the name suggests, these encoders are well known and widely used. Their concept is also pretty straight-forward.

1) Ordinal Encoding

The ordinal features are features that have an order. This type of data is also called **ordinal data**. Let's look at the Height column in the data frame. The categories are: *very short*, *short*, *normal*, *tall*, *very tal*l and it makes sense to put them in increasing/decreasing order. By encoding the columns manually, we can significantly boost the model performance.

2) One-hot encoding

Let's look at the column Type. This is **nominal** data which is the opposite to **ordinal data** in the Height column. The easiest way to turn this column into numerical is to use **one-hot encoding** by following the 2 steps

- Split all the categories in one column to different columns
- Put the checkmark *1* for the appropriate location

The get_dummiesfunction in pandas can achieve this goal

3) Binary Encoding

Imagine that you have 200 different categories. One-hot encoding will create 200 different columns. That a lot of columns will takes up a lot of memory. It the meantime, **binary encoding** only need 8 columns. It takes advantage of the binary system and so there might be multiple ones in a row. The logical explanation behind binary encoding is:

- Going down the column, every time it sees a new category, it gives a number, starting from 1 (and the next one is 2)
- Convert these number into binary
- Place each digit in this binary in a separate column.

4) Frequency Encoding

Give each category the **probability** (occurrence/total event). This means that if there are two categories in a column with the same probability (3 `fire` and 3 `bugs`), you cannot really tell the difference between them after being frequency encoded. The trade-off is no new column will be introduced.

5) Hashing Encoding

Hashing converts categorical variables to a higher dimensional space of integers. I won't comment on the methodology much here since scikit-learn explains it very well.

The `n_feature` is the number of columns you want to add. These new columns distinguish the corresponding category. However, you can adjust to any number. This is like binary encoding on steroids!

Advantage

- Deal with large scale categorical features
- High speed and reduced memory usage

Disadvantage

- No inverse-transformation method

Note: What is a good number of returning features? If m are distinct features, and n_feature=k , then m<2^k

II. Contrast encoders

Contrast encoding allows for recentering of categorical variables such that the intercept of a model is not the mean of one level of a category, but instead, the mean of all data points in the data set.

Many people argue that these encodings are not very effective. However, I will leave them here as references.

6) Helmert (reverse) Encoding

Helmert encoding compares each level of a categorical variable to the mean of the subsequent levels.

7) Backward Difference Encoding

In **backward difference encoding**, the mean of the dependent variable for a level is compared with the mean of the dependent variable for the prior level.

III. Bayesian Target Encoders

The general idea of this method is to take the target into account.

Advantage:

- Require minimal effort, only create one column for any number of categories in that feature
- Most favorite encoding scheme in Kaggle competition

Disadvantage:

- Only work for supervised learning (thus, inherently leaky). This means that when dealing with unsupervised data, it gets worse!

- Need regularization for the previous reason

8) Target Encoding

Target-based encoding basically means using the target to encode categorical features. The formula for it is:

$$TE_i = \frac{\text{total true}(y_i)}{\text{total}(y_i)}$$

where y_i is a category and λ is a smoothing function

In the table below, Legendary is our target. Since there are only 1 out of 3 Fire pokemon that are legendary, its value is 1/3. Think about target encoding as frequency encoding on the target!

Note: There are some different version that multiplies the output by a (Laplace)smoothing value. This is to avoid data leakage.

9) Leave One Out Encoding

Leave One Out Encoding (LOOE) is very similar to target encoding but excludes the current row's target when calculating the mean target for a level to reduce the effect of outliers.

Additionally, you can add some (Gaussian) noise to the data to prevent overfitting by changing the sigma value between 0 and 1.

10) Weight of Evidence Encoding

Weight of Evidence Encoding (WoE) is a measure of how much the evidence supports or undermines a hypothesis.

$$WoE = \left[\ln \left(\frac{\text{Distribution of goods} + adj}{\text{Distribution of Bads} + adj} \right) \right]$$

where adj is the adjacent factor is a function that avoids division by 0.

Advantage:

- Work well with logistic regression since WoE transformation has the same logistic scale.
- Can use WoE to compare across feature since their values are standardized.

Disadvantages:

- May lose information due to some category may have the same WoE
- Does not take into account features correlation
- Overfitting

Note: We can adjust the adj factor by changing regularization. (By default it is 1). When setting it equal to 0. You come back to the original WOE and may encounter division by 0

11) James-Stein Encoding (JSE)

This is target encoding but is more robust. **James-Stein (JS)** is it works best for the feature that has a normal distribution. JS is defined by the formula:

$$JS_i = (1 - B) \cdot \text{mean}(y_i) + B \cdot \text{mean}(y)$$

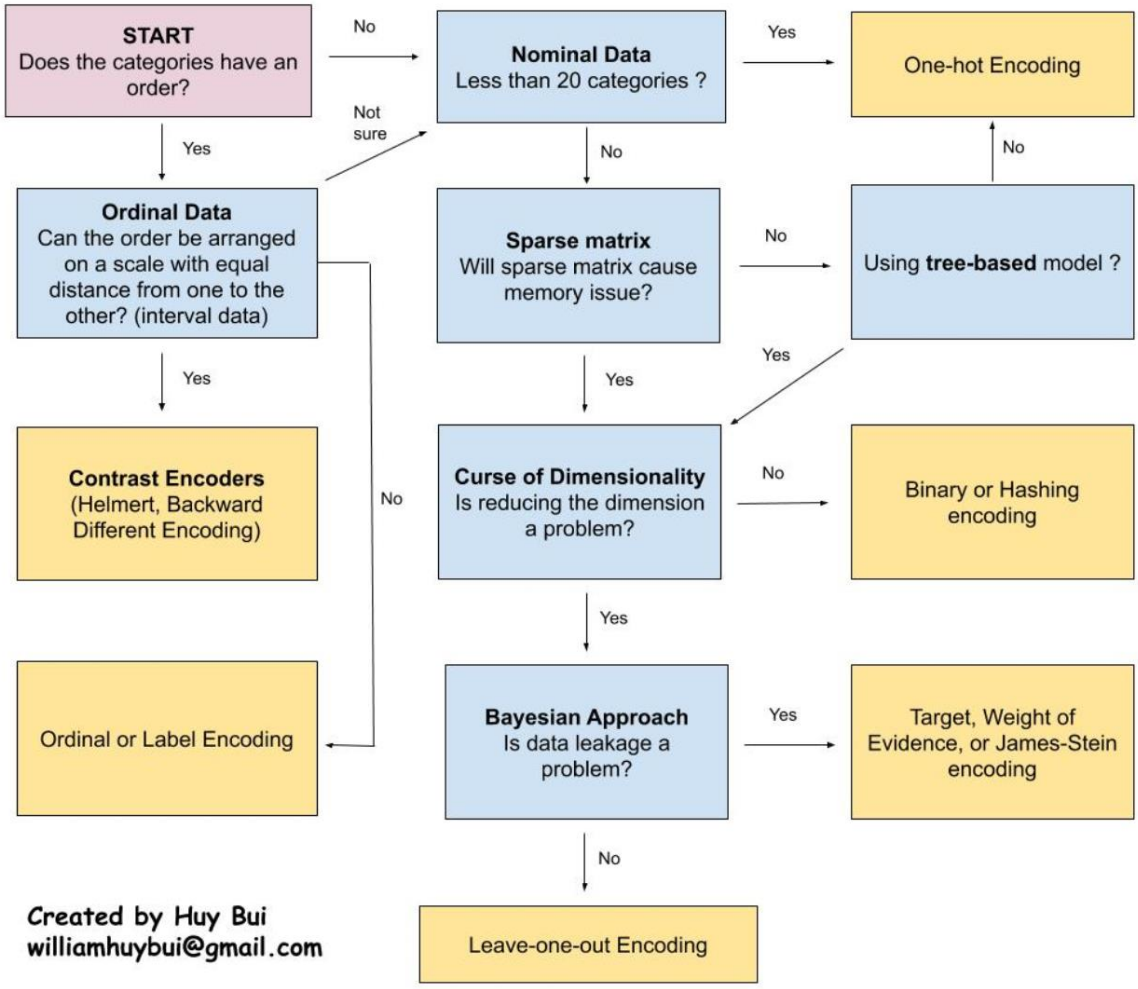
where $\text{mean}(y)$ is the global mean of the target, $\text{mean}(y_i)$ is the mean of the category, and B is the weight.

The weight B depends on the variances *σ(y)* and *σ(yi)*.

12) M-estimator Encoding

M-Estimate encoder is a simplified version of Target Encoder. The stands for maximum likelihood-type. It has only one hyper-parameter m, which represents the power of regularization. The higher the value of m results into stronger shrinking. Recommended values m are in the range of 1 to 100.

There is no single formula for encoding a feature. However, if you understand the 12 encoding techniques I introduced above, you would be able to move fast. Moreover, it always worth tries all the techniques that apply to the feature and decides which one works best. Try to input different regularization coefficient values and see if they increase your score. The cheat-sheet below will help us make some initial decisions.



Categorical encoding cheat sheet

For our dataframe, we use the Label Encoder to change the object data type columns to integers. This will make describing, correlating and model testing much easier and accurate.

3.4 DESCRIBING THE DATASET

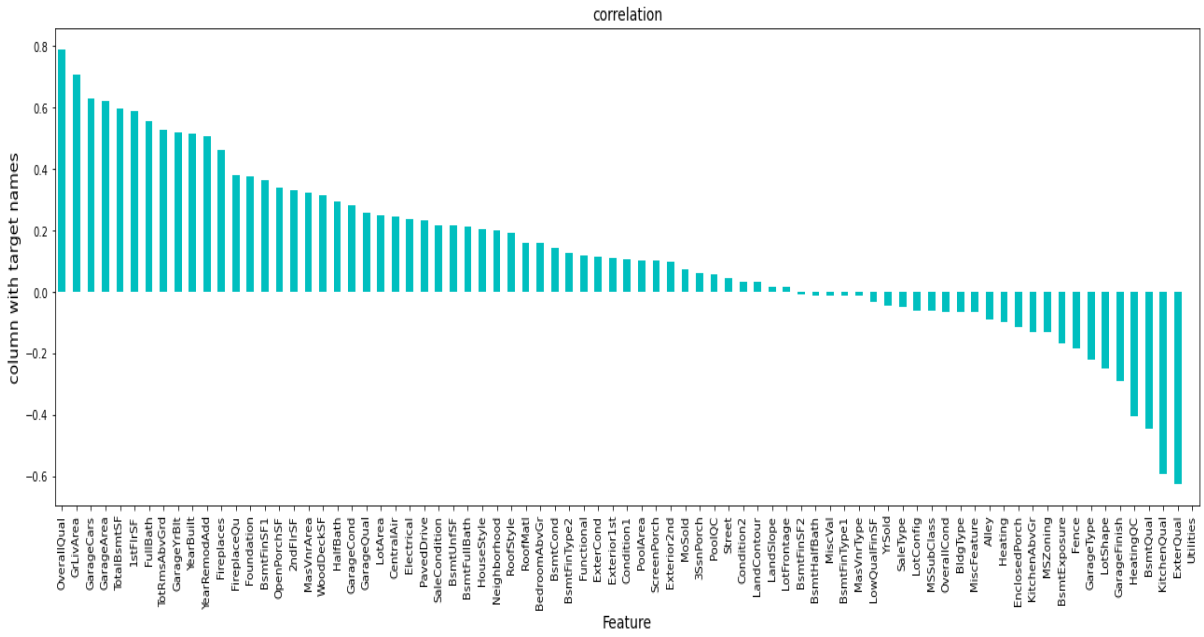
Describing the dataset gives us an understanding of various values like count of the attribute,mean of that attribute, standard deviation, minimum, 25th percentile, median/50th percentile, 75th percentile and the maximum value of that attribute.

We also use heatmap visualization for checking the relation between the described data.

After this, we check the correlation with the target column which is column80- Sale Price

We also form a heatmap for visualising the correlation between each column with each other.

For checking whether the columns are positively or negatively correlated with the target column, we see the following chart:



Keeping +/-1.5 as the range for skewness, here are the columns which do not lie within the range- MSZoning, LotArea, Street etc. Since no column has skewness, we will not treat them.

3.5 CHECKING FOR OUTLIERS

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Why outlier analysis?
Most data mining methods discard outliers noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring one and hence, the outlier analysis becomes important in such case.

Detecting Outlier:
Clustering based outlier detection using distance to the closest cluster:
In the K-Means clustering technique, each cluster has a mean value. Objects belong to the cluster whose mean value is closest to it. In order to identify the Outlier, firstly we need to initialize the threshold value such that any distance of any data point greater than it from its nearest cluster identifies it as an outlier for our purpose. Then we need to find the distance of the test data to each cluster mean. Now, if the distance between the test data and the closest cluster to it is greater than the threshold value then we will classify the test data as an outlier.

Algorithm:

1. Calculate the mean of each cluster
2. Initialize the Threshold value
3. Calculate the distance of the test data from each cluster mean
4. Find the nearest cluster to the test data
5. If (Distance > Threshold) then, Outlier

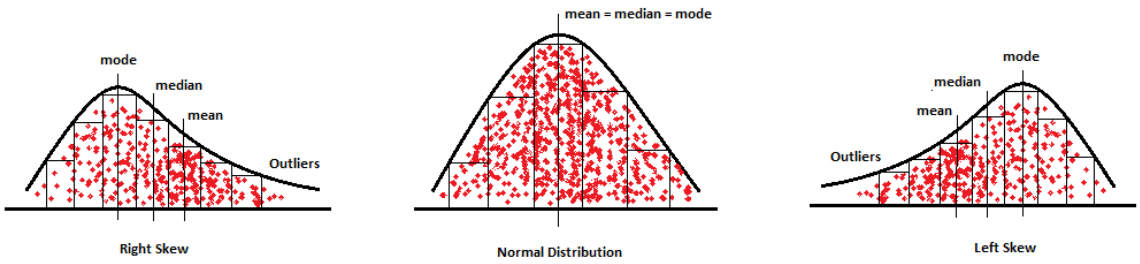
We use boxplot visualization for checking the outliers for the columns.

For removal of outliers we use Z-score.

The Z-score is a way to standardize the data to standard scale i.e. how far the data point is from the mean. The z-score can come positive or negative based on the help of mean and standard deviation values.

The data point away from the mean with some standard deviation is called a z-score.

The z-score can be perfectly found in a normal distribution curve with no left skew and right skew. The below image shows these curves.



Normal Distribution: The normal distribution is a curve in which the data is spread symmetrically on both sides of the mean.

Right Skew: The data is mostly skewed on the right side because most of the data is on the right side. If we talk about outliers they are mostly on the right side too.

Left Skew: The data is mostly skewed on the left side because most of the data is on the left side. If we talk about outliers they are mostly on the left side too.

Mean and Standard Deviation

Mean: The mean is an average value of the data that tells about the center value of the data.

Standard deviation: It is a spread of the data around the mean with one standard deviation.

Application of z-score in Machine learning

- To standardize the data as a part of data pre-processing.
- To compare the z-score values of different standard distributions for better results. Standard scaling is a crucial process in the data pre-processing of the machine learning algorithm.

3.6 SEPARATING COLUMNS INTO FEATURES AND TARGET

In order to make a prediction (in this case, whether a customer will recommend the e-commerce site to a friend or not), one needs to separate the dataset into two components:

- the **dependent** variable or **target** which needs to be predicted
- the **independent** variables or **features** that will be used to make a prediction

In machine learning, the concept of dependent and independent variables is important to understand. In the above dataset, if you look closely, the first 80 columns determine the outcome of the 81st , or last, column (Recommended). Intuitively, it means that the decision to buy a product of a given category is determined by the location, pathway, fireplace and many other factors affecting the price of the house.. So, we can say that Recommended is the dependent variable, the value of which is determined by the other four variables.

With this in mind, we need to split our dataset into the matrix of independent variables and the vector or dependent variable. Mathematically, Vector is defined as a matrix that has just one column.

For dataset splitting we give the following command:

```
features=df.drop('SalePrice',axis=1)
target=df['SalePrice']
```

3.7 SCALING THE DATASET

Scaling is a method of standardization that's most useful when working with a dataset that contains continuous features that are on different scales, and you're using a model that operates in some sort of linear space (like linear regression or K-nearest neighbors)

Feature scaling transforms the features in your dataset so they have a mean of zero and a variance of one This will make it easier to linearly compare features. Also, this a requirement for many models in scikit-learn.

Feature Scaling is one of the most important transformation we need to apply to our data. Machine Learning algorithms (Mostly Regression algorithms) don't perform well when the inputs are numerical with different scales.

when different features are in different scales, after applying scaling all the features will be converted to the same scale. Let's take we have two features where one feature is measured on a scale from 1 to 10 and the second feature is measured on a scale from 1 to 100,00, respectively. If we calculate the mean squared error, algorithm will mostly be busy in optimizing the weights corresponding to second feature instead of both the features. Same will be applicable when the algorithm uses distance calculations like Euclidian or Manhattan distances, second feature will dominate the result. So, if we scale the features, algorithm will give equal priority for both the features.

There are two common ways to get all attributes to have the same scale: *min-max scaling* and *standardization.*

We will use Min-Max scaling technique for our dataset.

Min-Max scaling, We have to subtract min value from actual value and divide it with max minus min. Scikit-Learn provides a transformer called MinMaxScaler. It has a feature_range hyperparameter that lets you change the range if you don't want 0 to1 for any reason.

class sklearn.preprocessing.MinMaxScaler(feature_range=0,1,*, copy=True, clip=False).

After this, we split the data into train and test data. We now check the training and testing accuracy using random state for loop for the range (0,100).

We now check the r2 score and handle overfitting and underfitting of the data.

Cross validation gives us the training and testing score, model accuracy and we do this using the CV mean using the for loop function.

We now draw the Best Fit line to represent the number of datapoints which shows good fit of our model. Equation of this line is : y=mx+c

3.8 REGULARIZATION OF THE DATASET

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "*In regularization technique, we reduce the magnitude of the features by keeping the same number of features.*"

Techniques of Regularization

There are mainly two types of regularization techniques, which are given below:

- **Ridge Regression**
- **Lasso Regression**

Ridge Regression

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
- In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

Lasso Regression:

- Lasso regression is another regularization technique to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator.**
- It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.

- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as **L1 regularization.**

Key Difference between Ridge Regression and Lasso Regression

- **Ridge regression** is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.
- **Lasso regression** helps to reduce the overfitting in the model as well as feature selection.

For our dataset, we use the Lasso Regression for regularization giving the following command:

```
from sklearn.model_selection import GridSearchCV

from sklearn.linear_model import Lasso

parameters={'alpha': [.0001, .001, .01, .1, 1, 10],

        'random_state':list(range(0,10))}

ls=Lasso()

clf= GridSearchCV(ls,parameters)

clf.fit(features_train,target_train)

print(clf.best_params_)
```

We get an output stating: {'alpha': 0.0001, 'random_state': 0}

We now do the final model training using the above best parameters and get the test score as 93.16%.

3.9 ENSEMBLE TECHNIQUE

Ensemble method in Machine Learning is defined as the multimodal system in which different classifier and techniques are strategically combined into a predictive model (grouped as Sequential Model, Parallel Model, Homogeneous and Heterogeneous methods etc.) Ensemble method also helps to reduce the variance in the predicted data, minimize the biasness in the predictive model and to classify and predict the statistics from the complex problems with better accuracy.

Ensemble Methods help to create multiple models and then combine them to produce improved results, some ensemble methods are categorized into the following groups:

1. Sequential Methods

In this kind of Ensemble method, there are sequentially generated base learners in which data dependency resides. Every other data in the base learner is having some dependency on previous data. So, the previous mislabeled data are tuned based on its weight to get the performance of the overall system improved.

Example: Boosting

2. Parallel Method

In this kind of Ensemble method, the base learner is generated in parallel order in which data dependency is not there. Every data in the base learner is generated independently.

Example: Stacking

3. Homogeneous Ensemble

Such an ensemble method is a combination of the same types of classifiers. But the dataset is different for each classifier. This will make the combined model work more precisely after the aggregation of results from each model. This type of ensemble method works with a large number of datasets. In the homogeneous method, the feature selection method is the same for different training data. It is computationally expensive.

Example: Popular methods like bagging and boosting comes into the homogeneous ensemble.

4. Heterogeneous Ensemble

Such an ensemble method is the combination of different types of classifiers or machine learning models in which each classifier built upon the same data. Such a method works for small datasets. In heterogeneous, the feature selection method is different for the same training data. The overall result of this ensemble method is carried out by averaging all the results of each combined model.

Below are the technical classification of Ensemble Methods:

1. Bagging

This ensemble method combines two machine learning models i.e. Bootstrapping and Aggregation into a single ensemble model. The objective of the bagging method is to reduce the high variance of the model. The decision trees have variance and low bias. The large dataset is (say 1000 samples) sub-sampled (say 10 sub-samples each carries 100 samples of data). The multiple decision trees are built on each sub-sample training data. While banging the sub-sampled data on the different decision trees, the concern of over-fitting of training data on each decision tree is reduced. For the efficiency of the model, each of the individual decision trees is grown deep containing sub-sampled training data. The results of each decision tree are aggregated to understand the final prediction. The variance of the aggregated data comes to reduce. The accuracy of the prediction of the model in the bagging method depends on the number of decision-tree used. The various sub-sample of a sample data is chosen randomly with replacement. The output of each tree has a high correlation.

2. Boosting

The boosting ensemble also combines different same type of classifier. Boosting is one of the sequential ensemble methods in which each model or classifier run based on features that will utilize by the next model. In this way, the boosting method makes out a stronger learner model from weak learner models by averaging their weights. In other words, a stronger trained model depends on the multiple weak trained models. A weak learner or a wear trained model is one that is very less correlated with true classification. But the next weak learner is slightly more correlated with true classification. The combination of such different weak learners gives a strong learner which is well-correlated with the true

3. Stacking

This method also combines multiple classifications or regression techniques using a meta-classifier or meta-model. The lower levels models are trained with the complete training dataset and then the combined model is trained with the outcomes of lower-level models. Unlike boosting, each lower-level model is undergone into parallel training. The prediction from the lower level models is used as input for the next model as the training dataset and form a stack in which the top layer of the model is more trained than the bottom layer of the model. The top layer model has good prediction accuracy and they built based on lower-level models. The stack goes on increasing until the best prediction is carried out with a minimum error. The prediction of the combined model or meta-model is based on the prediction of the different weak models or lower layer models. It focuses to produce less bias model.

4. Random Forest

The random forest is slightly different from bagging as it uses deep trees that are fitted on bootstrap samples. The output of each tress is combined to reduce variance. While growing each tree, rather than generating a bootstrap sample based on observation in the dataset, we also sample the dataset based on features and use only a random subset of such a sample to

build the tree. In other words, sampling of the dataset is done based on features that reduce the correlation of different outputs. The random forest is good for deciding for missing data. Random forest means random selection of a subset of a sample which reduces the chances of getting related prediction values. Each tree has a different structure. Random forest results in an increase in the bias of the forest slightly, but due to the averaging all the less related prediction from different trees the resultant variance decreases and give overall better performance.

For our dataset, we use the Random Forest Regressor and get the output as:

{'criterion': 'mae', 'max_features': 'log2'}.

Now, we put the output and check the r2 and cross val score of the model. We get an output:

R2 Score: 87.63086012762095
Cross val score: 84.77104845915167

We are getting model accuracy and cross validation both above 99% which shows our model is performing very good.

4. CONCLUSION

4.1 SAVING THE MODEL

Pickle is a useful Python tool that allows you to save your ML models, to minimise lengthy re-training and allow you to share, commit, and re-load pre-trained machine learning models. Most data scientists working in ML will use Pickle or Joblib to save their ML model for future use.

Pickle is a generic object serialization module that can be used for serializing and deserializing objects. While it's most commonly associated with saving and reloading trained machine learning models, it can actually be used on any kind of object. Here's how you can use Pickle to save a trained model to a file and reload it to obtain predictions.

To save the ML model using Pickle all we need to do is pass the model object into the dump() function of Pickle. This will serialize the object and convert it into a "byte stream" that we can save as a file called model.pkl. You can then store, or commit to Git, this model and run it on unseen test data without the need to re-train the model again from scratch.

We use the pickle.dump method to save the final model.

4.2 CONCLUSION

To load a saved model from a Pickle file, all you need to do is pass the "pickled" model into the Pickle load() function and it will be deserialized. By assigning this back to a model object, you can then run your original model's predict() function, pass in some test data and get back an array of predictions.

After this, for the final step we reload the saved model and test it on the test.csv data for predicting the house price of the test dataset. We see that, predicted house price values are much higher than the original ones because, real estate properties are appreciated with more facilities and features being added to and around the house property.

We see that our model is performing very good with the accuracy score of above 87%.

Final findings:

1. For predicting the price of the house, the following variables are important:

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

Fireplaces: Number of fireplaces

PoolQC: Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Hom es)

2. All these variables affect the price of the house. The location of the house matters a lot, if it is closer to a railway station/ metro/ subway so that trav elling is easier and hassle free. Fire places and pool quality matters due to the weather conditions in USA. During winters, fireplace keeps the whole house warm and during summers, a pool day relaxes you out. The foundat ion, fencing and pathways need to be strong and well conditioned which e nsures safety. All these variables are correlated and impact house pricing.

The business goal will be satisfied if the company uses this LINEAR REGR ESSION MACHINE LEARNING MODEL to predict house prices. We see, for our test data the predicted prices are quite high than the original prices.

5. BIBLIOGRAPHY

References:

- https://pbpython.com/pandas_dtypes.html
- https://www.datasciencemadesimple.com/encode-decode-column-dataframe-python/#:~:text=encode%20%28%29%20function%20with%20codec%20%E2%80%98base64%E2%80%99%20and%20error,be%20Decode%20a%20column%20of%20dataframe%20in%20python%3A
- https://www.geeksforgeeks.org/machine-learning-outlier/
- https://medium.com/pythoneers/z-distribution-or-z-score-application-in-machine-learning-fbba081cd9fe
- https://www.investopedia.com/terms/h/house-price-index-hpi.asp
- https://towardsdatascience.com/data-scaling-for-machine-learning-the-essential-guide-d6cfda3e3d6b
- https://blog.finxter.com/the-complete-guide-to-min-max-scaler-in-machine-learning-with-ease/#:~:text=Min-Max%20scaling%20is%20a%20normalization%20technique%20that%20enables,range%20using%20each%20feature%E2%80%99s%20minimum%20and%20maximum%20value.
- https://www.javatpoint.com/regularization-in-machine-learning
- https://www.educba.com/ensemble-methods-in-machine-learning/
- https://practicaldatascience.co.uk/machine-learning/how-to-save-and-load-machine-learning-models-using-pickle
- https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/

THE END