Mehrdad Baradaran
DataScience Assignment01
House_price dataset

Report 01 (House_price)
Prof. MR Kheradpishe
Shahid Beheshti University

Analysis of Housing Dataset

Author: Mehrdad Baradaran

Abstract

This report presents an analysis of a housing dataset, examining various statistical tests and hypotheses. The analysis seeks to understand how a given budget compares to the average house price, whether it is possible to purchase a house within that budget, and how the budget relates to factors like house size, zoning, and garage types.

1. Introduction

The housing dataset contains valuable information about house prices, zoning classifications, and garage types. This analysis aims to answer several key questions:

- How does a given budget compare to the average house price?
- Is it possible to purchase a house within the budget?
- How does the budget relate to the size of the house, zoning classification, and garage type?

2. Data Preprocessing

The dataset was preprocessed to handle missing values and log-transform the 'SalePrice' column to make it more suitable for analysis.

3. Exploratory Data Analysis (EDA)

Exploratory data analysis was performed, revealing insights into the dataset's statistics and distributions.

Take a First Look of our Data:

I created the function below to simplify the analysis of general characteristics of the data. Inspired on the str function of R, this function returns the types, counts, distinct, count nulls, missing ratio and uniques values of each field/feature.

If the study involve some supervised learning, this function can return the study of the correlation, for this we just need provide the dependent variable to the pred parameter.

Also, if its return is stored in a variable you can evaluate it in more detail, focus on specific field, or sort them from different perspectives.

Some Observations from the STR Details:

- The dependent variabel, **SalePrice**, are ***skewed*** and ***heavy-tailed distribution***. We need investigate its distribution with a plot and check if a **transformation by Log 1P** could correct it, withou drop most of the **outiliers**.

- Nulls: The data have 19 features with nulls, five of then area categorical and with more then 47% of missing ration. They are candidates to drop or use them to create another more interesting feature:

    - PoolQC
    - MiscFeature
    - Alley
    - Fence
    - FireplaceQu

- Features ***high skewed right***, ***heavy-tailed distribution***, and with ***high correlation*** to Sales Price. It is important to treat them (boxcox 1p transformation, Robustscaler, and drop some outliers):
    - TotalBsmtSF
    - 1stFlrSF
    - GrLivArea

- Features ***skewed***, ***heavy-tailed distribution***, and with ***good correlation*** to Sales Price. It is important to treat them (boxcox 1p transformation, Robustscaler, and drop some outliers):
    - LotArea
    - KitchenAbvGr
    - ScreenPorch
    - EnclosedPorch
    - MasVnrArea
    - OpenPorchSF
    - LotFrontage
    - BsmtFinSF1
    - WoodDeckSF
    - MSSubClass

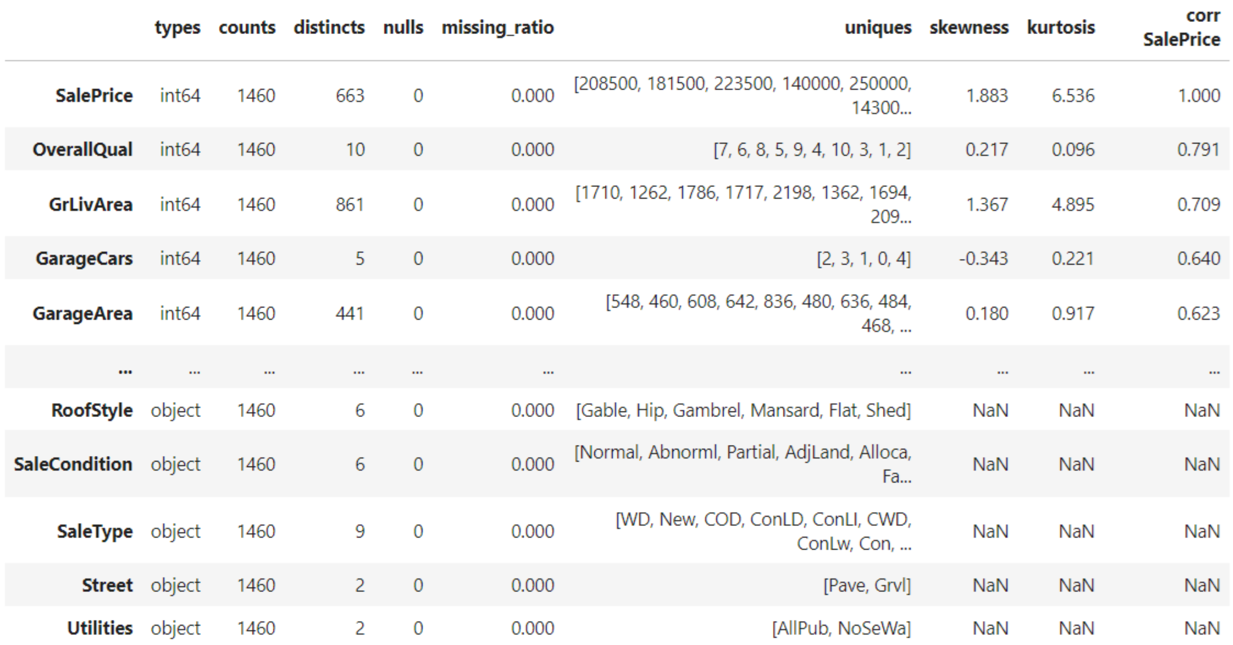
- Features ***high skewed***, ***heavy-tailed distribution***, and with ***low correlation*** to Sales Price. Maybe we can drop these features, or just use they with other to create a new more importants feature:
    - MiscVal
    - TSsnPorch

  - LowQualFinSF
  - BsmtFinSF2
  - BsmtHalfBa

- Features ***low skewed***, and with ***good to low correlation*** to Sales Price. Just use a Robustscaler probably reduce the few distorcions:
  - BsmtUnfSF
  - 2ndFlrSF
  - TotRmsAbvGrd
  - HalfBath
  - Fireplaces
  - BsmtFullBath
  - OverallQual
  - BedroomAbvGr
  - GarageArea
  - FullBath
  - GarageCars
  - OverallCond

- Transforme from Yaer Feature to Age, 2011 - Year feature, or YEAR(TODAY()) - Year Feature
  - YearRemodAdd:
  - YearBuilt
  - GarageYrBlt
  - YrSold

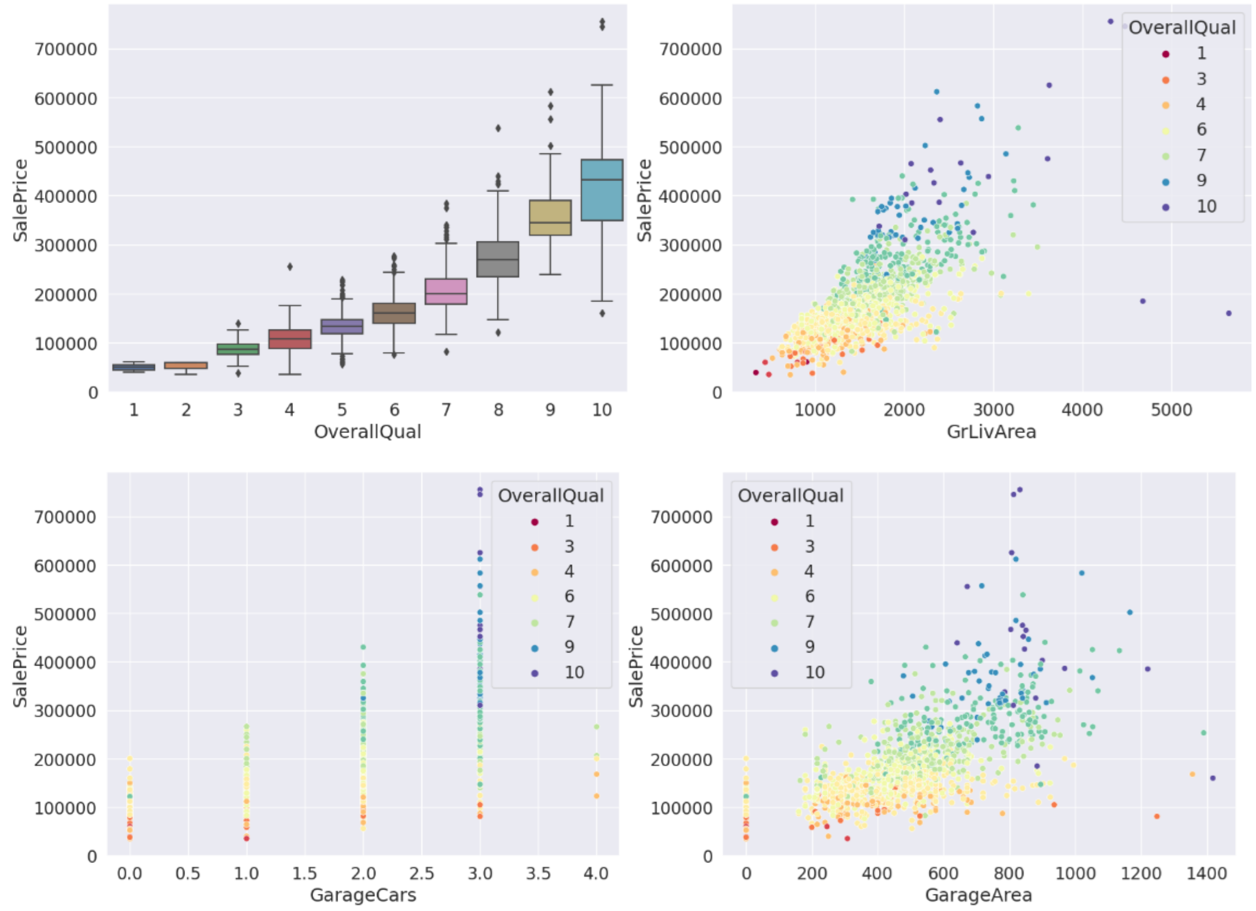
If we apply this data to a Keras, first we need to chnage the float64 and Int64 to float32 and Int32!

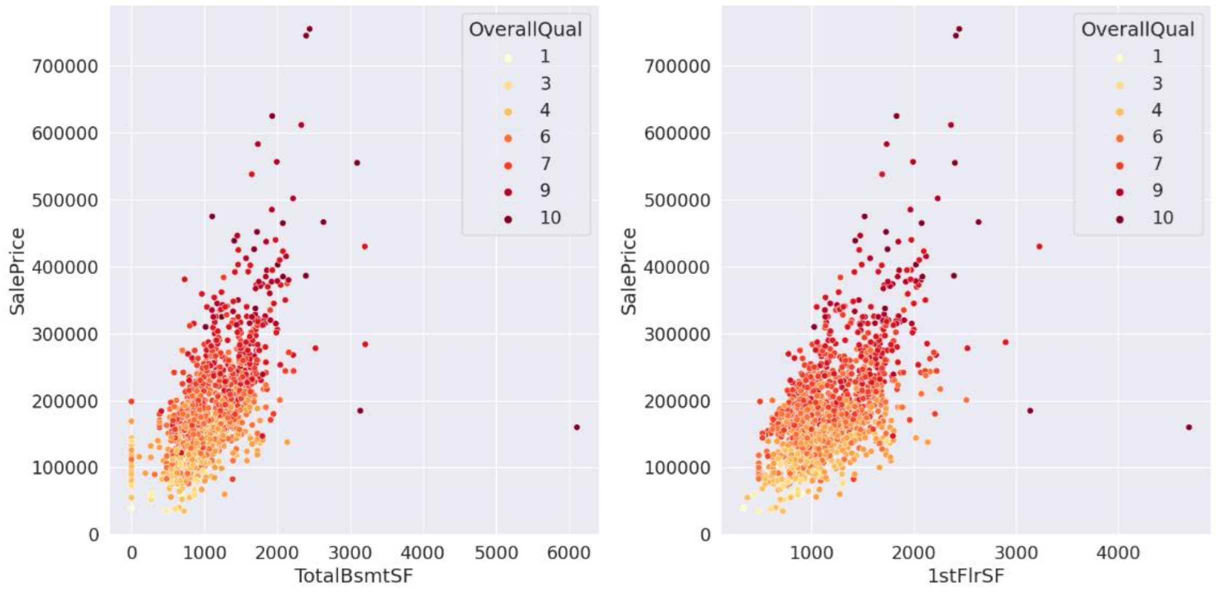


	types	counts	distincts	nulls	missing_ratio	uniques	skewness	kurtosis	corr SalePrice
SalePrice	int64	1460	663	0	0.000	[208500, 181500, 223500, 140000, 250000, 14300...	1.883	6.536	1.000
OverallQual	int64	1460	10	0	0.000	[7, 6, 8, 5, 9, 4, 10, 3, 1, 2]	0.217	0.096	0.791
GrLivArea	int64	1460	861	0	0.000	[1710, 1262, 1786, 1717, 2198, 1362, 1694, 209...	1.367	4.895	0.709
GarageCars	int64	1460	5	0	0.000	[2, 3, 1, 0, 4]	-0.343	0.221	0.640
GarageArea	int64	1460	441	0	0.000	[548, 460, 608, 642, 836, 480, 636, 484, 468, ...	0.180	0.917	0.623
...	...	...	...	...	...	...	...	...	...
RoofStyle	object	1460	6	0	0.000	[Gable, Hip, Gambrel, Mansard, Flat, Shed]	NaN	NaN	NaN
SaleCondition	object	1460	6	0	0.000	[Normal, Abnorml, Partial, AdjLand, Alloca, Fa...	NaN	NaN	NaN
SaleType	object	1460	9	0	0.000	[WD, New, COD, ConLD, ConLI, CWD, ConLw, Con, ...	NaN	NaN	NaN
Street	object	1460	2	0	0.000	[Pave, Grvl]	NaN	NaN	NaN
Utilities	object	1460	2	0	0.000	[AllPub, NoSeWa]	NaN	NaN	NaN

It is not surprise that overall quality has the highest correlation with SalePrice among the numeric variables (0.79). It rates the overall material and finish of the house on a scale from 1 (very poor) to 10 (very excellent). The positive correlation is certainly there indeed, and seems to be a slightly upward curve. Regarding outliers, I do not see any extreme values. If there is a candidate to take out as an outlier later on, it seems to be the expensive house with grade 4.

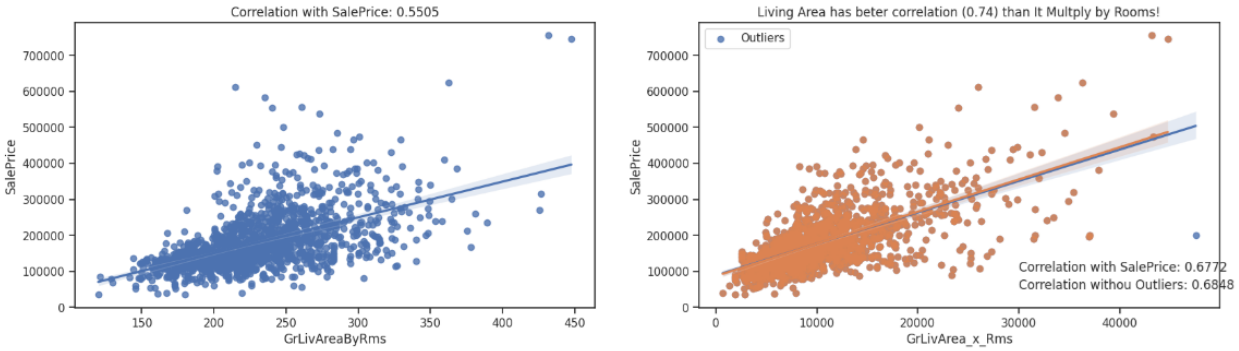
Especially the two houses with really big living areas and low SalePrices seem outliers. I will not take them out yet, as taking outliers can be dangerous. For instance, a low score on the Overall Quality could explain a low price. However, as you can see below, these two houses actually also score maximum points on Overall Quality. Therefore, I will keep theses houses in mind as prime candidates to take out as outliers.





Total Rooms above Ground and Living Area

From a previews experience with Boston data set, you probably main expect to much from the total rooms above ground, as its 'RM' feature (the average number of rooms per dwelling), but here is not the same scenario. Our common sense make to think that live area maybe has some correlation to it and probably we can combine this two features to produce a better predictor. Let's see.

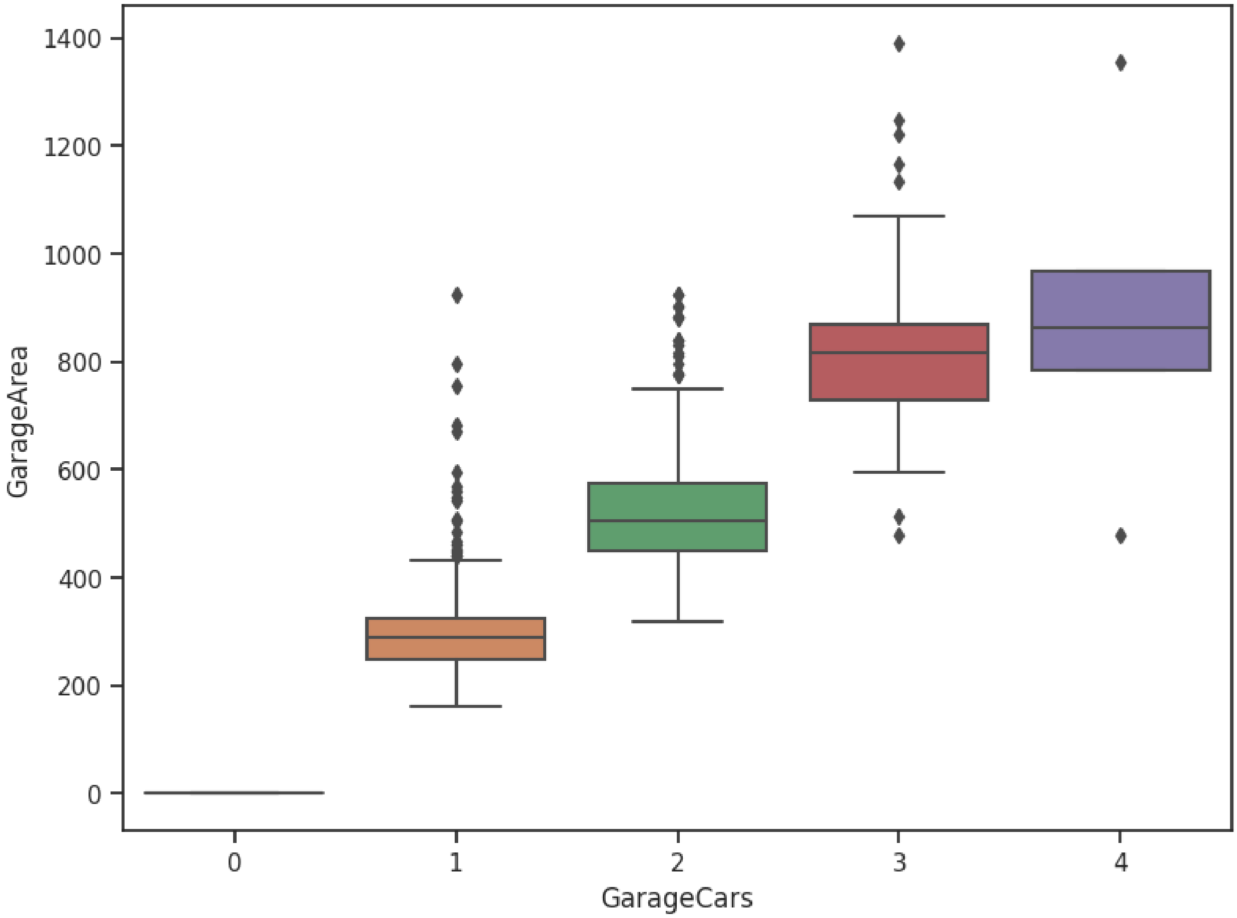


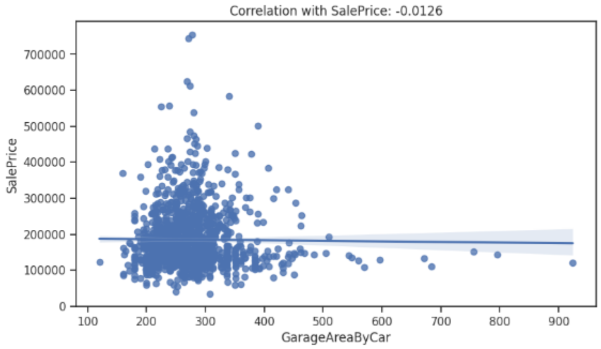
As we can see, the interaction between the two features did not present a better correlation than that already seen in the living area, include it improves to 0.74 with the cut of the outliers.

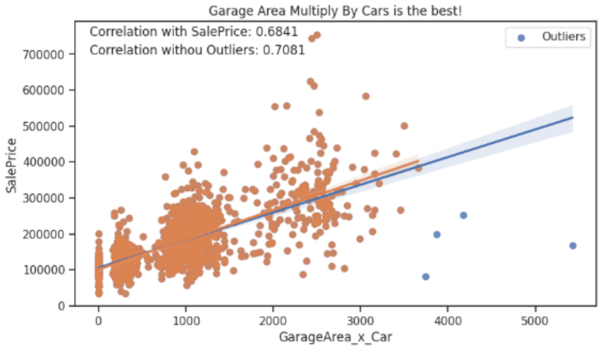
On the other hand, the ***multiplication*** not only demonstrated the living area **outliers** already identified, but it still **emphasized another**. If the strategy is to ***drop the TotRmsAbvGrd***, we should also ***exclude this additional outlier***.

Garage areas and parking

From the boxplot below, we can note that more than 3 parking cars and more than 900 of area are outliers, since a few number of their observations. Although there is a relationship between them, most likely with a smaller number of parking spaces, there may be more garage area for other purposes, reason why the correlation between them is 0.88 and not 1.







As can be seen the area by car is little useful, but contrary to common sense the multiplication of the area by the number of vacancies yes is. In the division we lose the magnitude and we have to maintain one or another functionality to recover it. With the multiplication we solve the problem of 1 parking space of 10 square feet against another of 10 with 1 square feet each. We could still ***improve the correlation*** by **0.06**, already considering the exclusion of only 4 outliers.

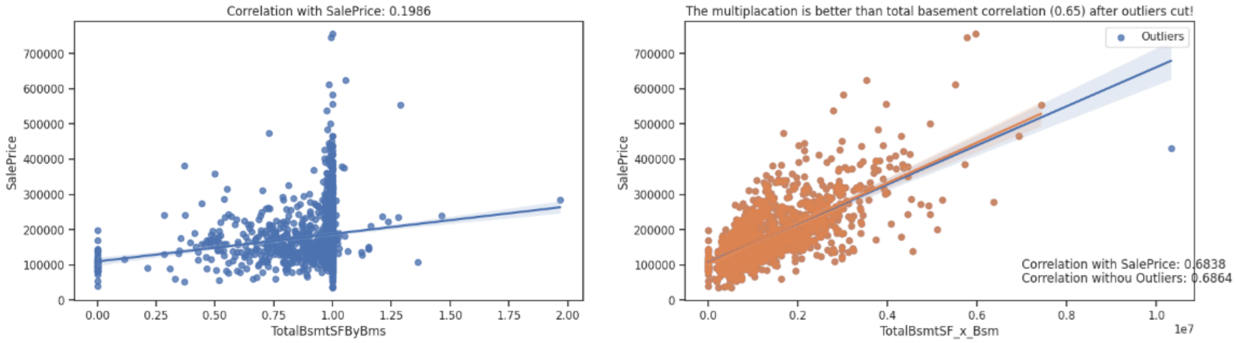
The identification of the outliers was facilitated, note that before we would have a greater number of outliers, since the respective of each features alone are not coincident.

So let's continue with the multiplication strategy, remove the two original metrics that have high correlation with each other, and exclude the 4 outliers from the training base.

Total Basement Area Vs 1st Flor Area

In our country it is not common to have Basement, I think we thought it was a little spooky. So I looked a bit more "carefully" at this variable...

I noticed that in Ames has a lot of variation, but the predictive effect is very small, so I decided to study its composition with the first floor.



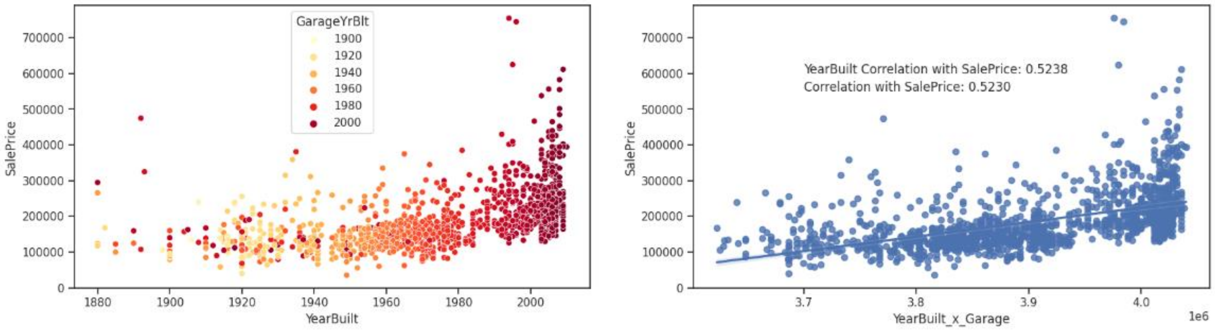
Similar to what we saw in the garage analysis, we again have a better correlation by multiplying the variables, but now we don't have a significant gain with outliers exclusion. So let's continue with the multiplication strategy and remove only the two original metrics that have high correlation with each other.

Year Built Vs Garage Year Built

Of course when we buy a property the date of its construction makes a lot of difference as it can be a source of great headaches. Depending on the age and conditions there will be need for renovations and very old houses there may be cases where the garage has been built or refit after the house itself.

Well, I'd be more worried about the plumbing, the electricity, ... the garage is only for car and trunk, or is it not? Is that so? it will be?

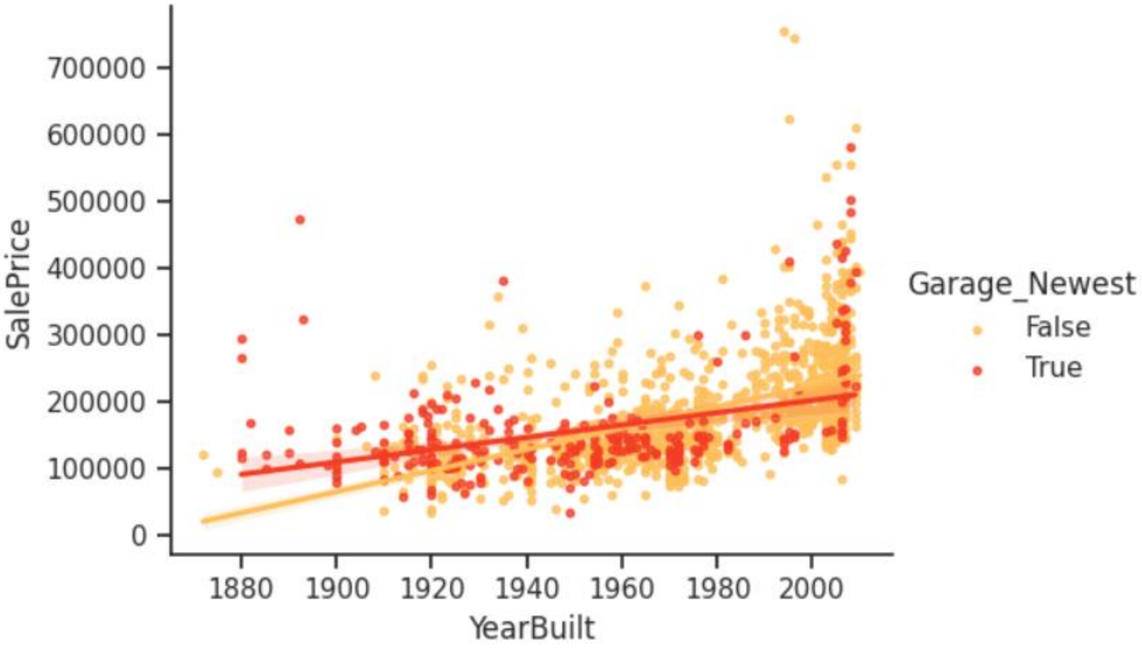
So, let's see the graphs below, and confirm that this two features are highly correlated, but as expect is not easy to find a good substitute by iteration.



However, by making the year of construction of the garage an indicator of whether it is newer, it becomes easiest to identify a pattern of separation.

And more, note that we have a rising price due to the lower age. Maybe the old cars had the garage would only be for themselves...

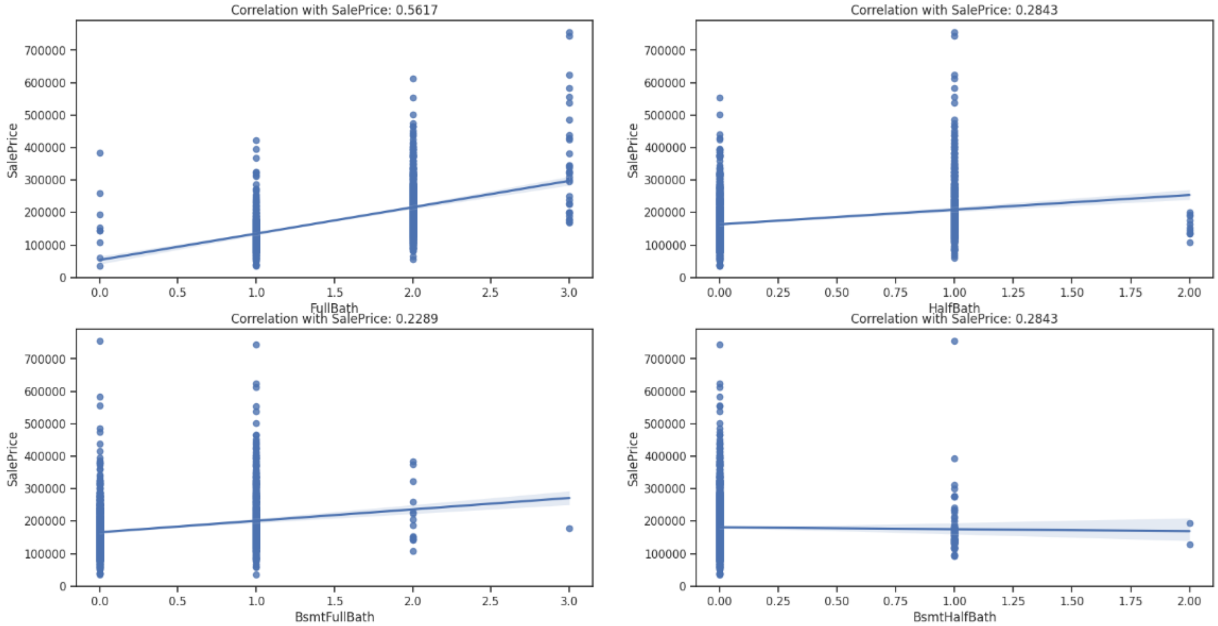
..., or put it in the barn. Today we must have other more usable uses for garage, right...?



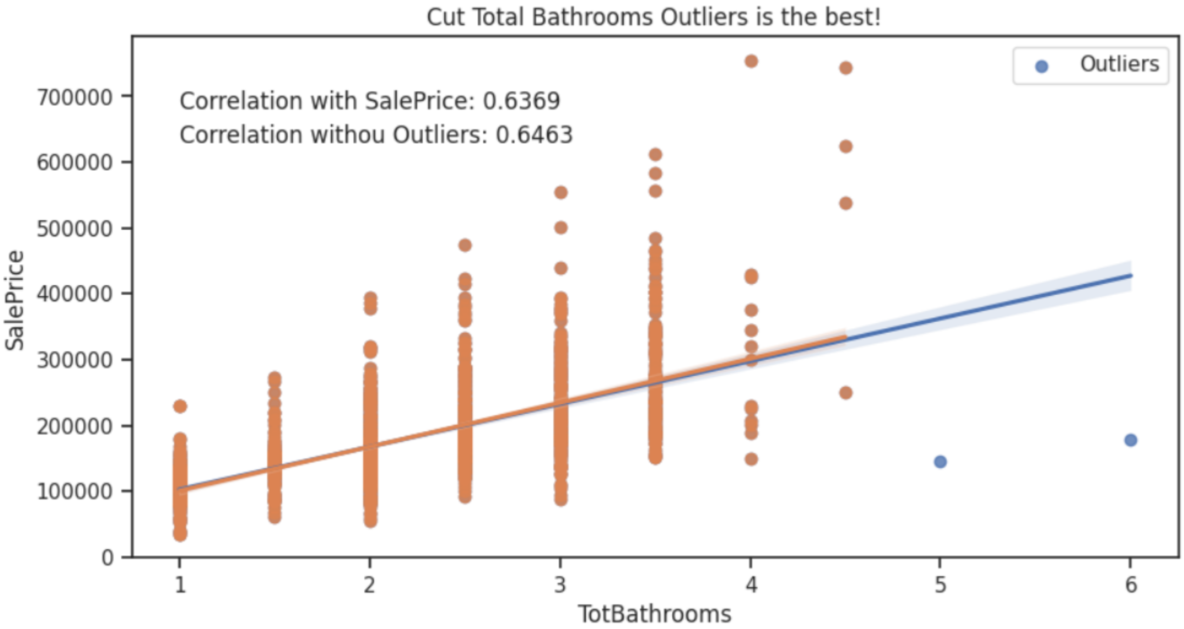
But note that although we have a rising price the newer the house, the growth rate is very smooth, even with the rate gain with a newer garage. This makes sense, given that the prices of these regressors are meeting with the mean price of each year.

Bathrooms Features

It's time to take a break and go to the toilet, to our luck there are 4 bathroom variables in our data set. FullBath has the largest correlation with SalePrice between than. The others individually, these features are not very important.



However, I assume that I if I add them up into one predictor, this predictor is likely to become a strong one. A half-bath, also known as a powder room or guest bath, has only two of the four main bathroom components-typically a toilet and sink. Consequently, I will also count the half bathrooms as half.

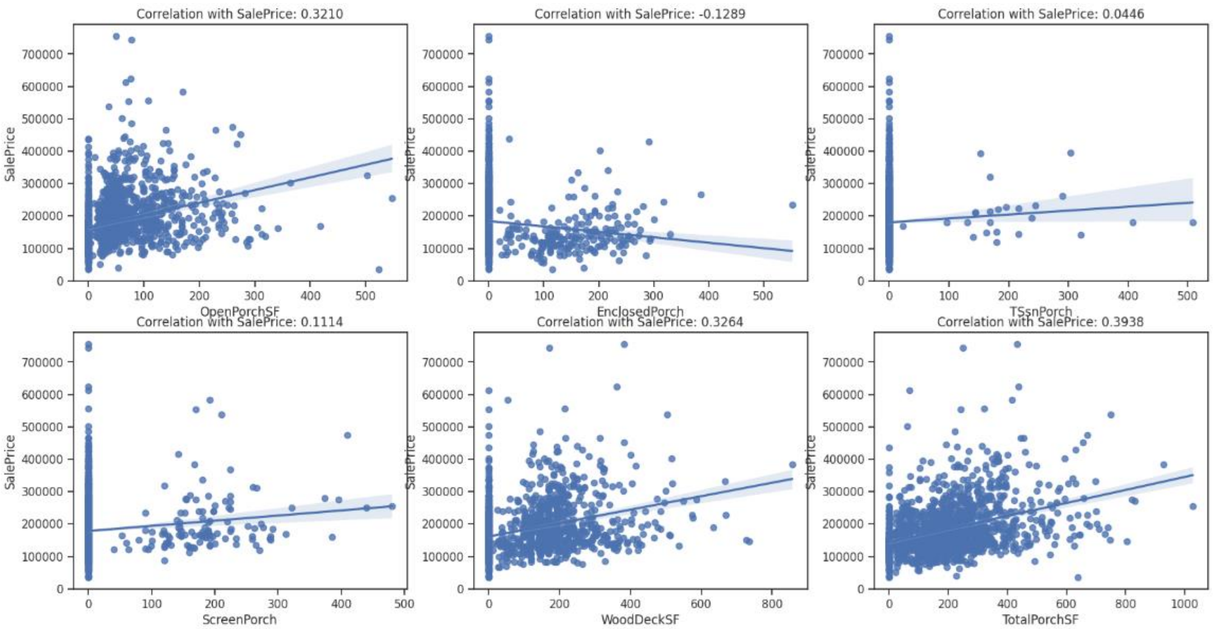


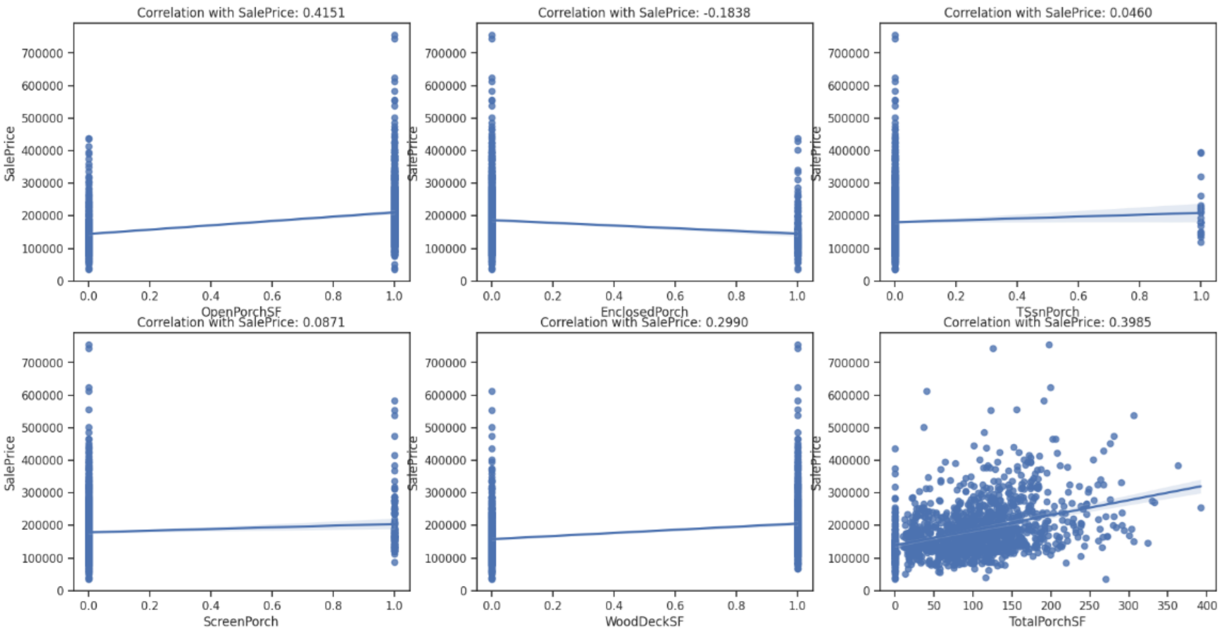
So, with our best predictor, we can cut only two outliers, use it and substitute all others bath features with a existence indicator.

Reviwe Porch Features:

The porch is where many people feel more comfortable to watch life go by, or you prefer the sofa in front of the TV, I think there are people that solved this to the family don't can fighting about this...

... this idea should make a house worth more, should not it?

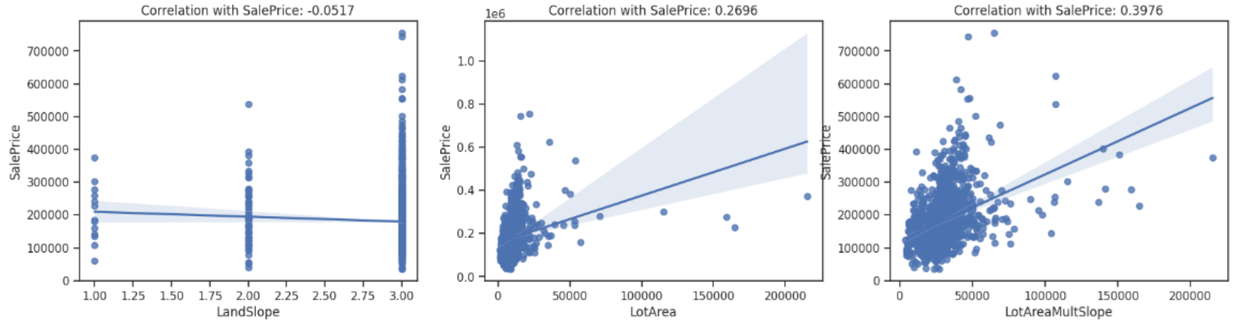




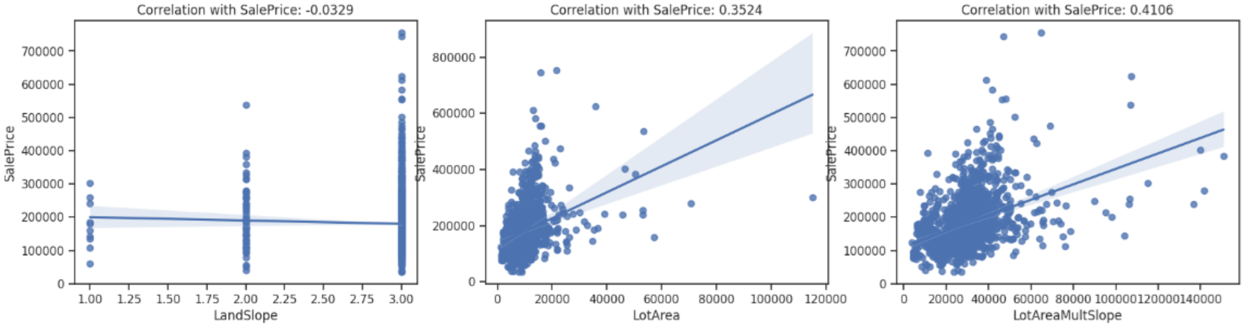
As we have seen, porch features have low correlation with price, and by the graphics we see all most has low bas and high variance, being a high risk to end complex models and fall into ouverfit.

Slope of property and Lot area

Everyone knows that the size of the lot matters, but has anyone seen any ad talking about the slope?

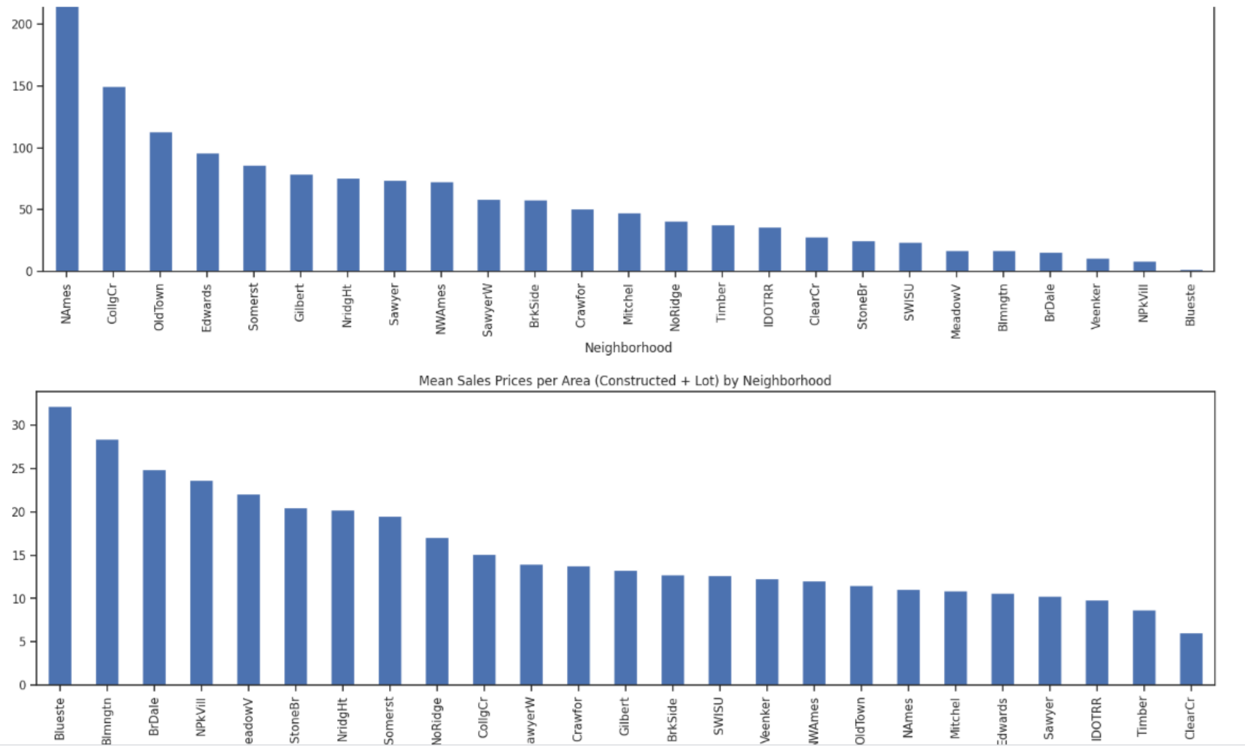


It is interesting to note that the slope has a low correlation, but as an expected negative. On the other hand, the lot size does not present such a significant correlation, contrary to the interaction between these two characteristics, which is better and also allow us to identify some outliers. Let's take a look at the effect of removing the outliers.

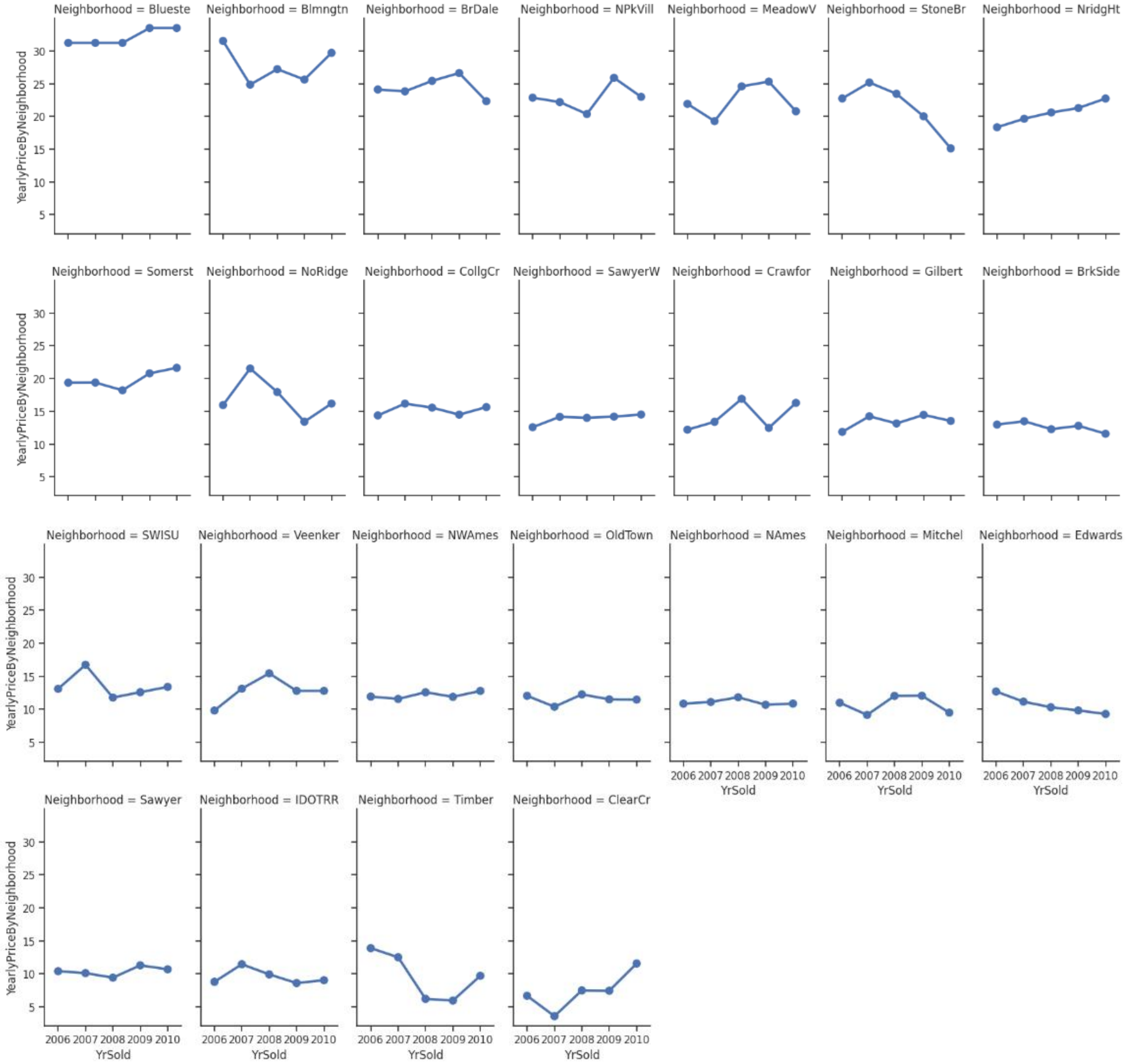


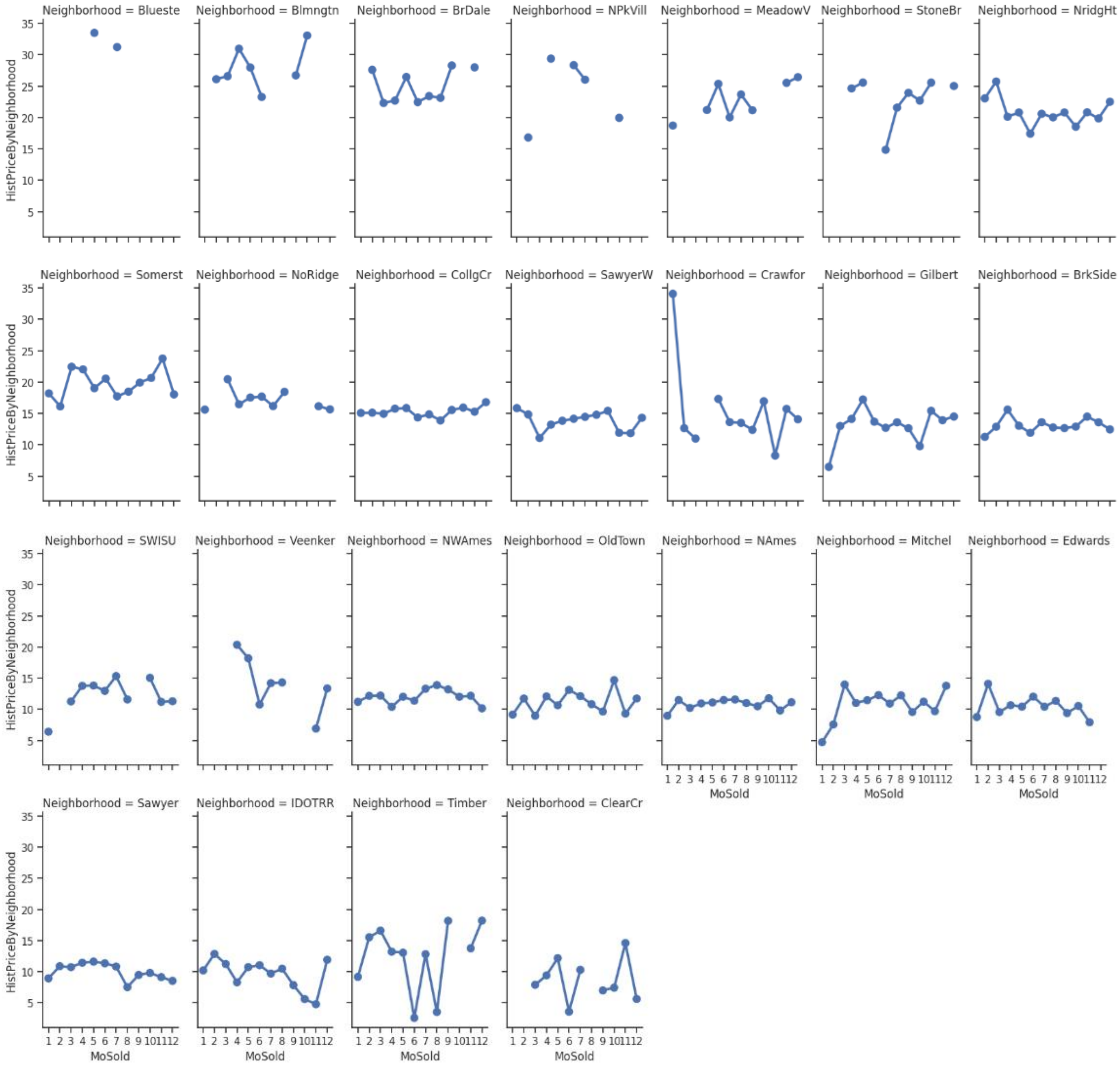
Neighborhood

Let's watch how much the neighborhood may be influencing the price.



As we can see prices are affected by the neighborhood, yes, if more similar more they attract. But we will delve a little and see how the year and month of the sale also has great influence on the price variation and confirm the seasonality.


Neighborhood = Blueste
Neighborhood = Blmngtn
Neighborhood = BrDale
Neighborhood = NPkVill
Neighborhood = MeadowV
Neighborhood = StoneBr
Neighborhood = NridgHt
Neighborhood = Somerst
Neighborhood = NoRidge
Neighborhood = CollgCr
Neighborhood = SawyerW
Neighborhood = Crawfor
Neighborhood = Gilbert
Neighborhood = BrkSide
Neighborhood = SWISU
Neighborhood = Veenker
Neighborhood = NWAmes
Neighborhood = OldTown
Neighborhood = NAmes
Neighborhood = Mitchel
Neighborhood = Edwards
Neighborhood = Sawyer
Neighborhood = IDOTRR
Neighborhood = Timber
Neighborhood = ClearCr
YearlyPriceByNeighborhood
30
25
20
15
10
5
2006
2007
2008
2009
2010
YrSold

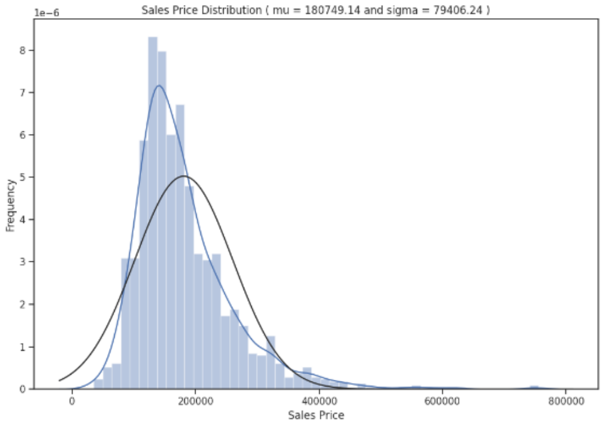


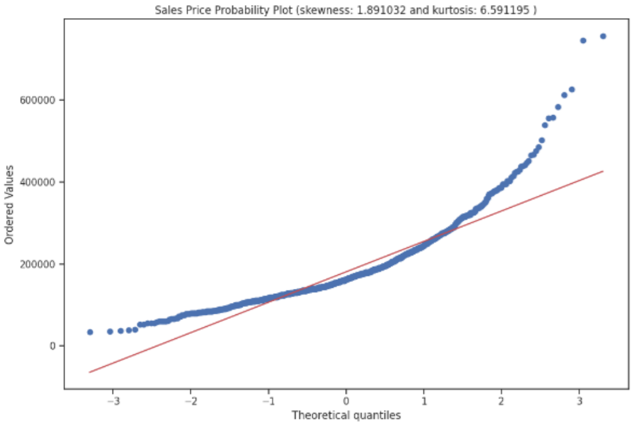
As we expected, the seasonality does have some effect, but of course we draw this conclusion based only on the above graphs is precipitated if not erroneous, given that even having restricted the views still exist houses with different characteristics in the same neighborhood.

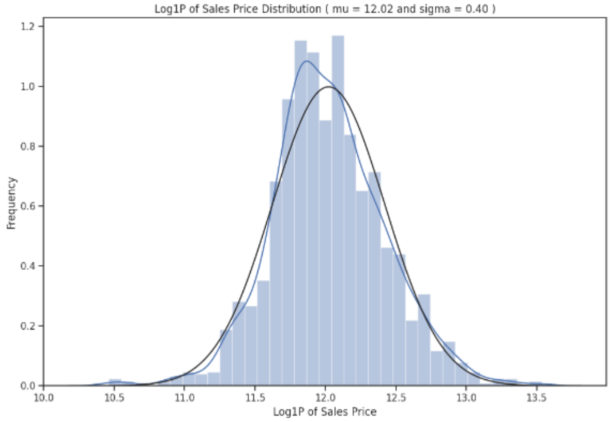
However, this is sufficient to understand that the timing of the sale matters, so the model will probably have to take this into account, or this will be part of the residual errors.

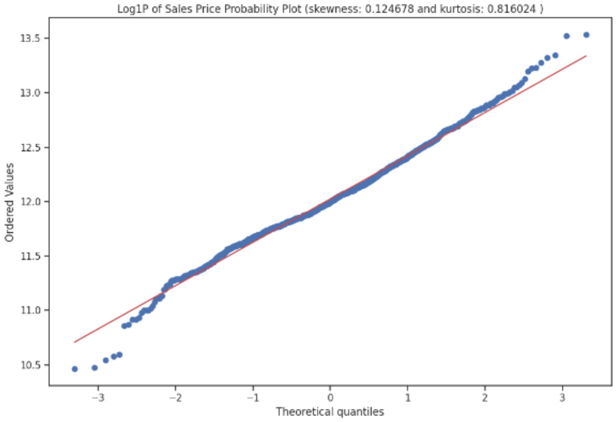
Check the Dependent Variable - SalePrice:

Since most of the machine learning algorithms start from the principle that our data has a normal distribution, we first take a look at the distribution of our **dependent variable**. For this, I create a procedure to plot the **Sales Distribution** and **QQ-plot** to identify substantive departures from normality, likes ***outliers***, ***skewness*** and ***kurtosis***.









From the first graph above we can see that Sales Price distribution is ***skewed***, has a **peak**, it **deviates from normal distribution** and is **positively biased**. From the **Probability Plot**, we could see that **Sales Price** also does **not align with the diagonal red line** which represent normal distribution. The form of its distribution confirm that is a skewed right.

With ***skewness positive of 1.9***, we confirm the **lack of symmetry** and indicate that Sales Price are **skewed right**, as we can see too at the Sales Distribution plot, skewed right means that the right tail is **long relative to the left tail**. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers, and **positive** kurtosis indicates a **heavy-tailed distribution** and **negative** kurtosis indicates a **light tailed distribution**. So, with 6.5 of positive kurtosis **Sales Price** are definitely heavy-tailed and has some **outliers** that we need take care.

4. Hypothesis Testing

Let's say our name is Mehrdad Baradaran and we are planning to move to Ames, Iowa, with a $120 000 budget to buy a house.
We have no idea about the real estate market in the city.
However the City Hall owns a precious piece of information : the House Price Dataset.
It contains about 1500 lines of data about houses in the city, with attributes like Sale Price, Living Area, Garage Type, etc...

The bad news is we can not access the entire database, it is too expensive.
The good news is the City Hall proposes some samples : free for up to 25 observations, with a small fee for up to 100 observations.
So we'll make use of this great offer to know a bit more about the real estate market, and understand what we can get for our money.

What we will be trying to do in this tutorial is make assumptions on the whole population of houses based only on the samples at our disposal.
This is what statistical tests do, but one must know a few principles before using them.

The process

The basic process of statistical tests is the following :

- Stating a Null Hypothesis (most often : "the two values are not different")
- Stating an Alternative Hypothesis (most often : "the two values are different")
- Defining an alpha value, which is a confidence level (most often : 95%). The higher it is, the harder it will be to validate the Alternative Hypothesis, but the more confident we will be if we do validate it.
- Depending on data at disposal, we choose the relevant test (Z-test, T-test, etc... More on that later)
- The test computes a score, which corresponds to a p-value.
- If p-value is below 1-alpha (0.05 if alpha is 95%), we can accept the Alternative Hypothesis (or "reject the Null Hypothesis"). If it is over, we'll have to stick with the Null Hypothesis (or "fail to reject the Null Hypothesis").

There's a built-in function for most statistical tests out there.
Let's also build our own function to summarize all the information.
All tests we will conduct from now on are based on alpha = 95%.

Two-tailed and One-tailed

Two-tails tests are used to show two values are just "different".
One-tail tests are used to show one value is either "larger" or "lower" than another one.

This has an influence on the p-value : in case of one-tail tests, p-value has to be divided by 2.

Most of the functions we'll use (those from the statweights modules) do that by themselves if we input the right information in the parameters.
We'll have to do it on our own with functions from the scipy module.

Types of tests

There are different types of tests, here are the ones we will cover :

- T-tests. Used for small sample sizes (n<30), and when population's standard deviation is unknown.
- Z-tests. Used for large sample sizes (n=>30), and when population's standard deviation is known.
- F-tests. Used for comparing values of more than two variables.
- Chi-square. Used for comparing categorical data.

Normal distribution

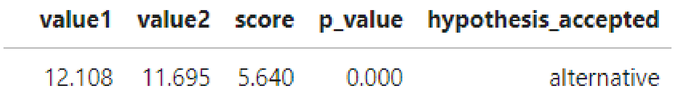
Also, most tests - parametric tests - require a population that is normally distributed.
It it not the case for SalePrice - which we'll use for most tests - but we can fix this by log-transforming the variable.
Note that to go back to our original scale and understand values vs. our $120 000, we'll to exponantiate values back.

So let's say we are ready to dive into the data, but not ready to pay the small fee for the large sample size. We'll be starting with the free samples of 25 observations.

One sample T-test | Two-tailed | Means

So first question we want to ask is : How are our $120 000 situated vs. the average Ames house SalePrice?
In other words, is 120 000 (11.7 logged) any different from the mean SalePrice of the population?
To know that from a 25 observations sample, we need to use a One Sample T-Test.

Null Hypothesis : Mean SalePrice = 11.695
Alternative Hypothesis : Mean SalePrice ≠ 11.695



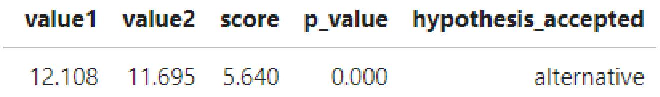
value1	value2	score	p_value	hypothesis_accepted
12.108	11.695	5.640	0.000	alternative

So we know our initial budget is significantely different from the mean SalePrice.
From the table above, it unfortunately seems lower.

One sample T-test | One-tailed | Means

Let's make sure our budget is lower by running a one-tailed test.
Question now is : is 120 000 (11.695 logged) lower than the mean SalePrice of the population?

Null Hypothesis : Mean SalePrice <= 11.695
Alternative Hypothesis : Mean SalePrice > 11.695



value1	value2	score	p_value	hypothesis_accepted
12.108	11.695	5.640	0.000	alternative

Unfortunately it is!
We have 95% chance of believing that our starting budget won't let us buy a house at the average Ames price.

Two sample T-test | Two-tailed | Means

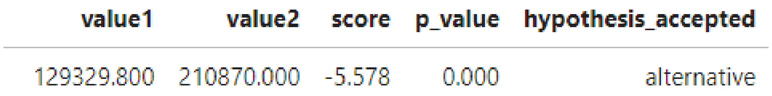
Now that our expectations are lowered, we realize something important :
The entire dataset probably contains some big houses fitted for entire families as well as small houses for fewer inhabitants.
Prices are probably really different in-between the two types.
And we are moving in alone, so we probably don't need that big of a house.

What if we could ask the City Hall to give us a sample for big houses, and a sample for smaller houses?
We first could see if there is a significant difference in prices.
And then see how our $120 000 are doing against the small houses average SalePrice.

We do ask the City Hall, and because they understand it is also for the sake of this tutorial, they accept.
They say they'll split the dataset in two, based on the surface area of the houses.
They will give us a sample from the top 50% houses in terms of surface, and another sample from the bottom 50%.

Now we first want to know if the two samples, extracted from two different populations, have significant differences in their average SalePrice.

Null Hypothesis : SalePrice of smaller houses = SalePrice of larger houses
Alternative Hypothesis : SalePrice of smaller houses ≠ SalePrice of larger houses



value1	value2	score	p_value	hypothesis_accepted
129329.800	210870.000	-5.578	0.000	alternative

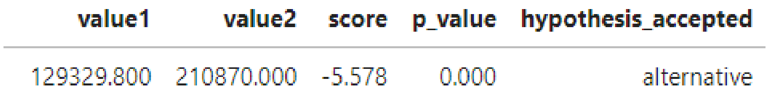
As expected, the two samples show some significant differences in SalePrice.

add Codeadd Markdown

Two sample T-test | One-tailed | Means

Obviously, larger houses have a higher SalePrice.
Let's prove it this with one-tailed test.

Null Hypothesis : SalePrice of smaller houses >= SalePrice of larger houses
Alternative Hypothesis : SalePrice of smaller houses < SalePrice of larger houses



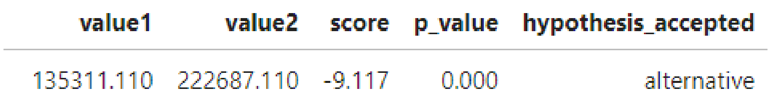
value1	value2	score	p_value	hypothesis_accepted
129329.800	210870.000	-5.578	0.000	alternative

Still as expected, SalePrice is significantly higher for larger houses.

Two sample Z-test | One-tailed | Means

Now that the City Hall has already splitted the population in two, why not ask them for larger samples? We'll pay a fee but that's all right, this is fake money.

Null Hypothesis : SalePrice of smaller houses >= SalePrice of larger houses
Alternative Hypothesis : SalePrice of smaller houses < SalePrice of larger houses



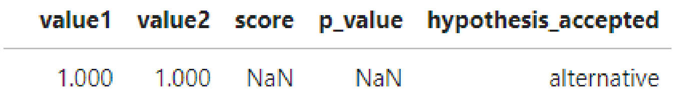
value1	value2	score	p_value	hypothesis_accepted
135311.110	222687.110	-9.117	0.000	alternative

Higher sample sizes show the same results : SalePrice is significantely higher for larger houses.

Two sample Z-test | One-tailed | Proportions

Instead of means, we can also run tests on proportions.
Is the proportion of houses over $120 000 higher in the larger houses populations than in smaller houses population?

Null Hypothesis : Proportion of smaller houses with SalePrice over 11.695 >= Proportion of larger houses with SalePrice over 11.695
Alternative Hypothesis : Proportion of smaller houses with SalePrice over 11.695 < Proportion of larger houses with SalePrice over 11.695



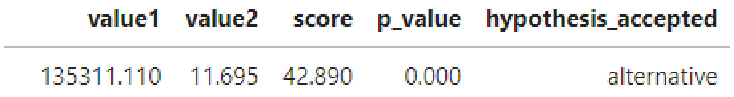
value1	value2	score	p_value	hypothesis_accepted
1.000	1.000	NaN	NaN	alternative

Logically, the test shows that the larger houses population has a higher ratio of houses sold over $120 000 vs. the smaller houses population.

One sample Z-test | One-tailed | Means

So now let's see how our $120 000 (11.7 logged) are doing against smaller houses only, based on the 100 observations sample.

Null Hypothesis : Mean SalePrice of smaller houses => 11.695
Alternative Hypothesis : Mean SalePrice of smaller houses < 11.695



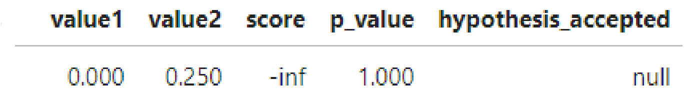
value1	value2	score	p_value	hypothesis_accepted
135311.110	11.695	42.890	0.000	alternative

That's quite depressing : our $120 000 do not even beat the average price of smaller houses.

One sample Z-test | One-tailed | Proportions

Our $120 000 do not seem too far from the average SalePrice of small houses though.
Let's see if at least 25% of houses have a SalePrice in our budget.

Null Hypothesis : Proportion of smaller houses with SalePrice under 11.695 <= 25%
Alternative Hypothesis : Proportion of smaller houses with SalePrice under 11.695 > 25%



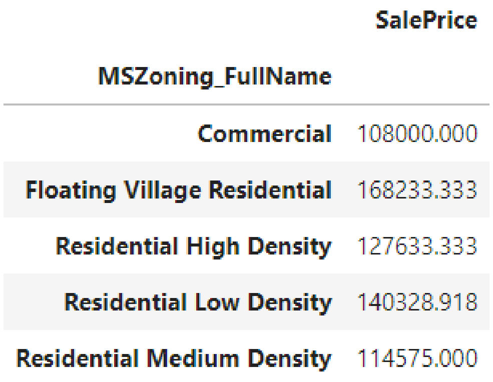
value1	value2	score	p_value	hypothesis_accepted
0.000	0.250	-inf	1.000	null

So at least, now we know we can buy a house among at least 25% of the smaller houses.

F-test (ANOVA)

The House Price Dataset has a MSZoning variable, which identifies the general zoning classification of the house.
For instance, it lets you know if the house is situated in a residential or a commerical zone.

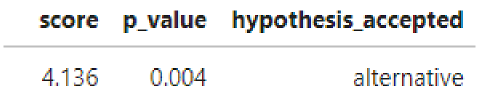
We'll therefore try to know if there is a significant difference in SalePrice based on the zoning.
And then know where we will be more likely to live with our budget.
Based on the 100 observations samples of smaller houses, let's first have an overview of mean SalePrice by zone.



	SalePrice
MSZoning_FullName	
Commercial	108000.000
Floating Village Residential	168233.333
Residential High Density	127633.333
Residential Low Density	140328.918
Residential Medium Density	114575.000

To know if there is a significant difference between these values, we run an ANOVA test. (because there a more than 2 values to compare)
The test won't not able to tell us what attributes are different from the others, but at least we'll know if there is a difference or not.

Null Hypothesis : No difference between SalePrice means
Alternative Hypothesis : Difference between SalePrice means

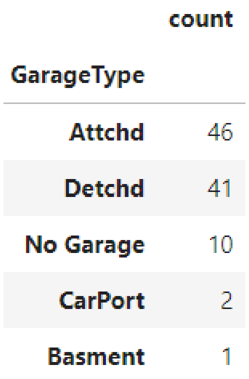


score	p_value	hypothesis_accepted
4.136	0.004	alternative

There is a difference between SalePrices based on where the house is located.
Looking at the Average SalePrice by zone, Commerical Zones and Residential High Density zones seem to be the most affordable for our budget.

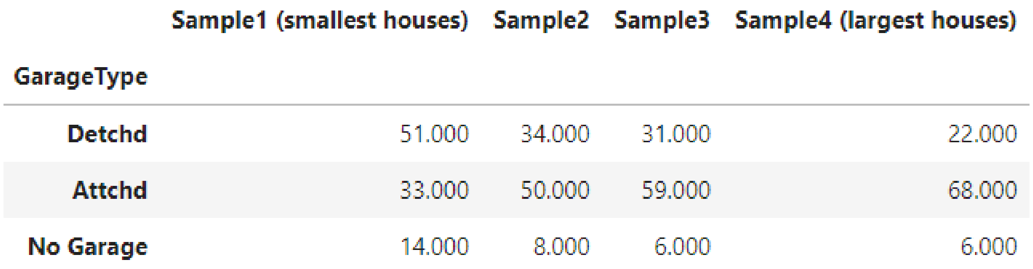
Chi-square test

One last question we'll address : can we get a garage? If yes, what type of garage?
If not, then we won't bother saving up for a car, and we'll try to get a house next to Public Transportion.
The dataset contains a categorical variable, GarageType, that will help us answer the question.



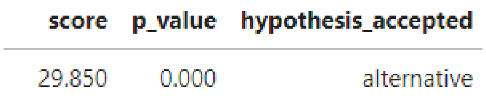
	count
GarageType	
Attchd	46
Detchd	41
No Garage	10
CarPort	2
Basment	1

We know we can get a house in at least the bottom 25% of smaller houses.
We would ideally like to know if distribution of Garage Types among these 25% is different than in the three other quarters
We are now friends with the City Hall, so we can ask them one last favor :
Split the smaller houses population in 4 based on surface, and give us a sample of each quarter.
Because we working here with categorical data, we'll run a Chi-Square test.



	Sample1 (smallest houses)	Sample2	Sample3	Sample4 (largest houses)
GarageType				
Detchd	51.000	34.000	31.000	22.000
Attchd	33.000	50.000	59.000	68.000
No Garage	14.000	8.000	6.000	6.000

Null Hypothesis : No difference between GarageType distribution
Alternative Hypothesis : Difference between GarageType distribution



score	p_value	hypothesis_accepted
29.850	0.000	alternative

Clearly there's a difference in GarageType distribution according to size of houses.
The sample that concerns us, Sample1, has the highest proportion of "No Garage" and "Detached Garage". We'll probably have to stick with Public Transportation.