**House Price**

**Ali Bakhshesh**

**Analysis**

**1.Data Exploration**

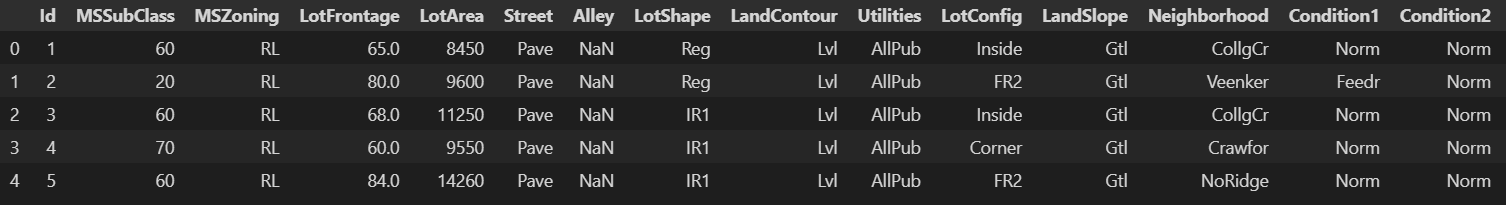
The house price dataset provides comprehensive information on residential properties, including key features such as square footage, number of bedrooms and bathrooms, location, and sale prices. This dataset is a valuable resource for real estate market analysis, helping to understand housing market trends and factors influencing property values.

The dataset consists of following attributes:

1. **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
2. **MSSubClass**: The building class
3. **MSZoning**: The general zoning classification
4. **LotFrontage**: Linear feet of street connected to property
5. **LotArea**: Lot size in square feet
6. **Street**: Type of road access
7. **Alley**: Type of alley access
8. **LotShape**: General shape of property
9. **LandContour**: Flatness of the property
10. **Utilities**: Type of utilities available
11. **LotConfig**: Lot configuration
12. **LandSlope**: Slope of property
13. **Neighborhood**: Physical locations within Ames city limits
14. **Condition1**: Proximity to main road or railroad
15. **Condition2**: Proximity to main road or railroad (if a second is present)
16. **BldgType**: Type of dwelling
17. **HouseStyle**: Style of dwelling
18. **OverallQual**: Overall material and finish quality
19. **OverallCond**: Overall condition rating
20. **YearBuilt**: Original construction date
21. **YearRemodAdd**: Remodel date
22. **RoofStyle**: Type of roof
23. **RoofMatl**: Roof material
24. **Exterior1st**: Exterior covering on house
25. **Exterior2nd**: Exterior covering on house (if more than one material)
26. **MasVnrType**: Masonry veneer type
27. **MasVnrArea**: Masonry veneer area in square feet
28. **ExterQual**: Exterior material quality
29. **ExterCond**: Present condition of the material on the exterior
30. **Foundation**: Type of foundation
31. **BsmtQual**: Height of the basement
32. **BsmtCond**: General condition of the basement
33. **BsmtExposure**: Walkout or garden level basement walls
34. **BsmtFinType1**: Quality of basement finished area
35. **BsmtFinSF1**: Type 1 finished square feet
36. **BsmtFinType2**: Quality of second finished area (if present)
37. **BsmtFinSF2**: Type 2 finished square feet
38. **BsmtUnfSF**: Unfinished square feet of basement area
39. **TotalBsmtSF**: Total square feet of basement area
40. **Heating**: Type of heating
41. **HeatingQC**: Heating quality and condition
42. **CentralAir**: Central air conditioning
43. **Electrical**: Electrical system
44. **1stFlrSF**: First Floor square feet
45. **2ndFlrSF**: Second floor square feet
46. **LowQualFinSF**: Low quality finished square feet (all floors)
47. **GrLivArea**: Above grade (ground) living area square feet
48. **BsmtFullBath**: Basement full bathrooms
49. **BsmtHalfBath**: Basement half bathrooms
50. **FullBath**: Full bathrooms above grade
51. **HalfBath**: Half baths above grade
52. **Bedroom**: Number of bedrooms above basement level
53. **Kitchen**: Number of kitchens
54. **KitchenQual**: Kitchen quality
55. **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
56. **Functional**: Home functionality rating
57. **Fireplaces**: Number of fireplaces
58. **FireplaceQu**: Fireplace quality
59. **GarageType**: Garage location
60. **GarageYrBlt**: Year garage was built
61. **GarageFinish**: Interior finish of the garage
62. **GarageCars**: Size of garage in car capacity
63. **GarageArea**: Size of garage in square feet
64. **GarageQual**: Garage quality
65. **GarageCond**: Garage condition
66. **PavedDrive**: Paved driveway
67. **WoodDeckSF**: Wood deck area in square feet
68. **OpenPorchSF**: Open porch area in square feet
69. **EnclosedPorch**: Enclosed porch area in square feet
70. **3SsnPorch**: Three season porch area in square feet
71. **ScreenPorch**: Screen porch area in square feet
72. **PoolArea**: Pool area in square feet
73. **PoolQC**: Pool quality
74. **Fence**: Fence quality
75. **MiscFeature**: Miscellaneous feature not covered in other categories
76. **MiscVal**: $Value of miscellaneous feature
77. **MoSold**: Month Sold
78. **YrSold**: Year Sold
79. **SaleType**: Type of sale
80. **SaleCondition**: Condition of sale

Although we use some of these features for analysis. In the blow you can see the first five rows of the dataset and their attributes.

A screen shot of a computer

Description automatically generated

A screen shot of a black screen

Description automatically generated

A screenshot of a computer

Description automatically generated

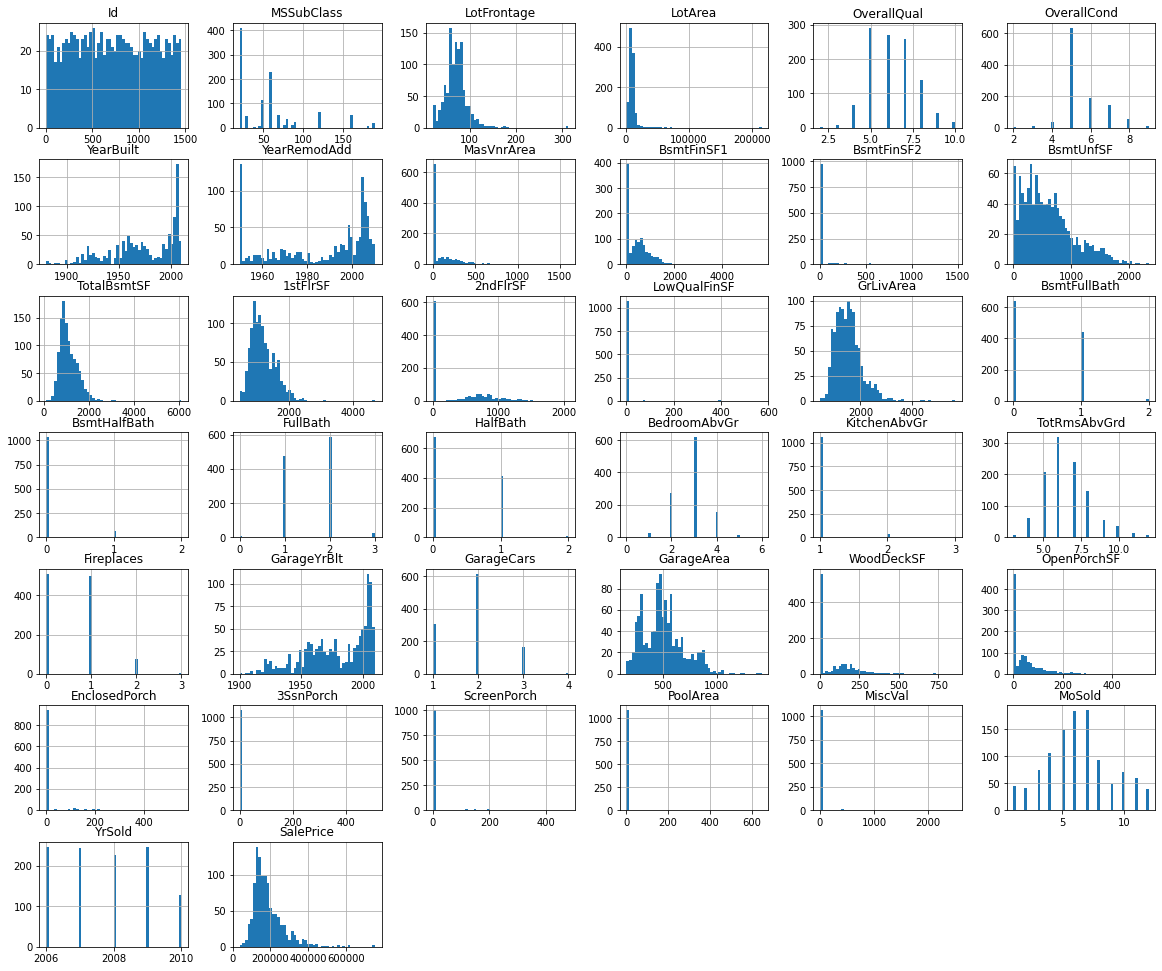
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

And this is the histogram of numerical data distribution.



**2. Data preprocessing:**

The data preprocessing and cleaning include following steps:

* **Omitting null values**: All the rows with null values could be eliminated or we could fill them with some proper values, but we did the first one because the number null values was not so huge compared to rest of the data.
* **Removing some columns**: In this project we confront many features but we just used a few of them specially numerical columns.

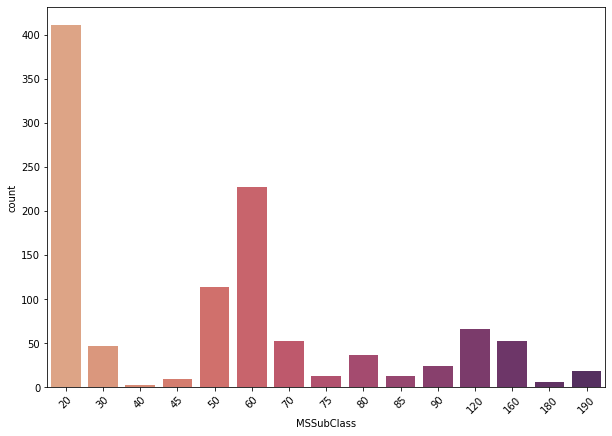
**3. Statistical Tests and Analysis**

**Question1:** What are the building classes and why it is important?

In the first question we just need to visualize some features to answer the question and its not necessary to do any statistical tests.

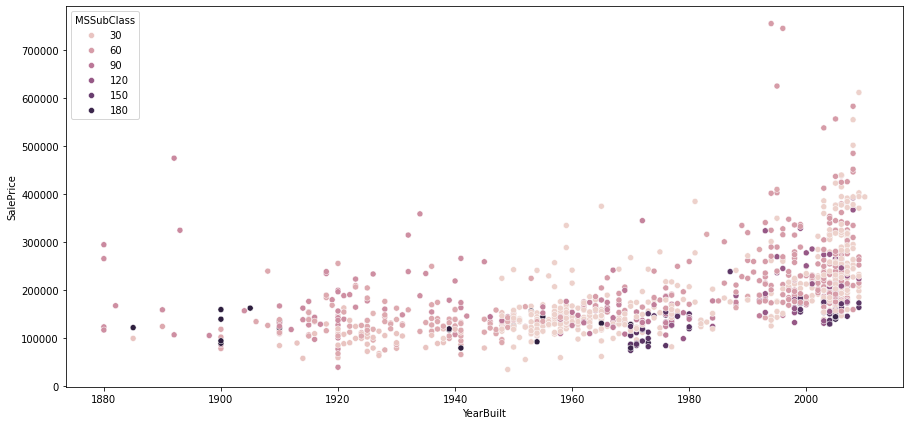
A pie chart with different colored circles

Description automatically generatedWe have drawn some plots and according to them we will answer the question.



As it is obvious in the above plots, we can see that the most frequent class of buildings is class 20.

In the following plot we have a scatter plot which demonstrates the relation among MSclass, Year Built and Sale price.

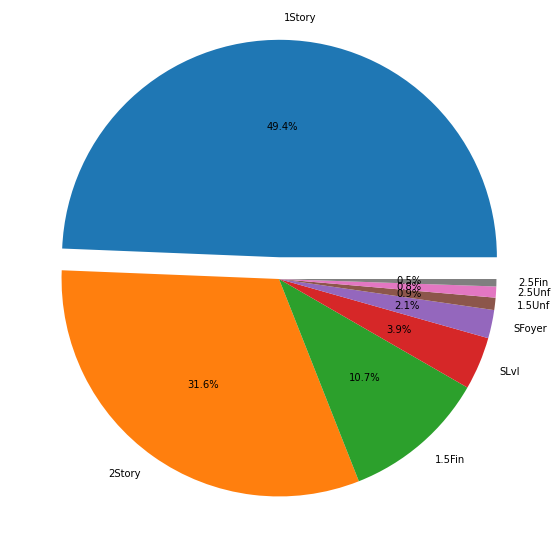


We cannot decide definitely about the effect of being in which class on price but as it was expected it is clear that newer houses are more expensive than older ones.

We also did some exploration about the style of building (1story, 2story, etc).

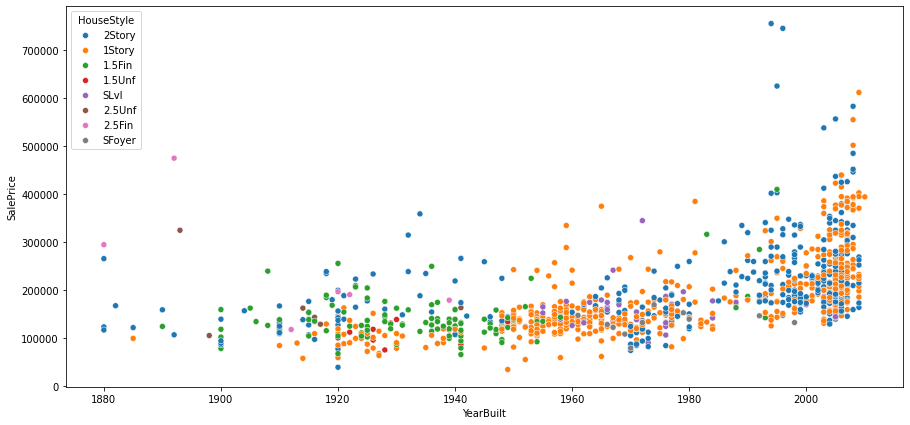
A graph of different colored squares

Description automatically generatedThe results have been visualized as you can see in the blow:



The buildings mostly include 1 and 2 story houses.

We have another plot which this time shows the relation among House style, Year Built and Sale price.



From my point of view the most important result we can get from this plot is that in recent years mostly 1 and 2 story houses have been built and other type of houses are gradually exticting.

**Question2:** How does the overall quality (OverallQual) of a house relate to its sale price?

Here we start our statistical tests and please NOTE that the value of α for all tests and in all questions is 0.05.

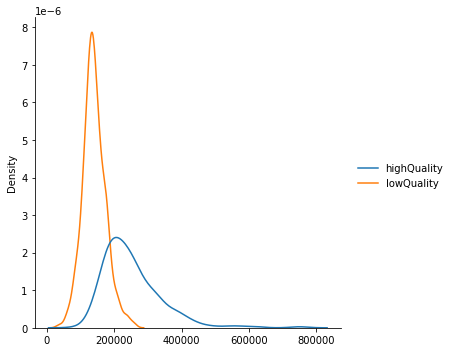
Quality of a house is specified by a discrete number between 1 and 10.

We consider houses with quality more than 6 as high-quality house and vice versa.

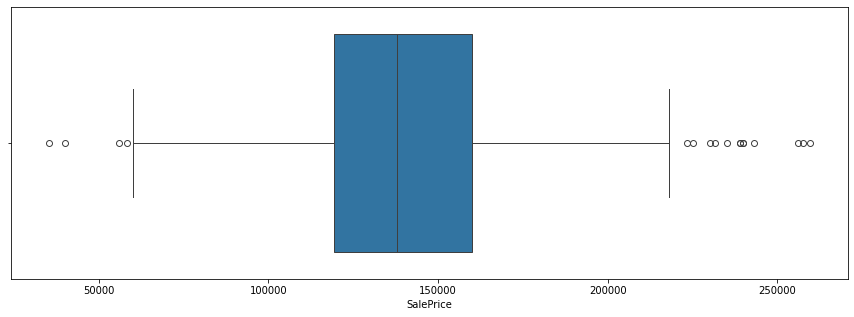
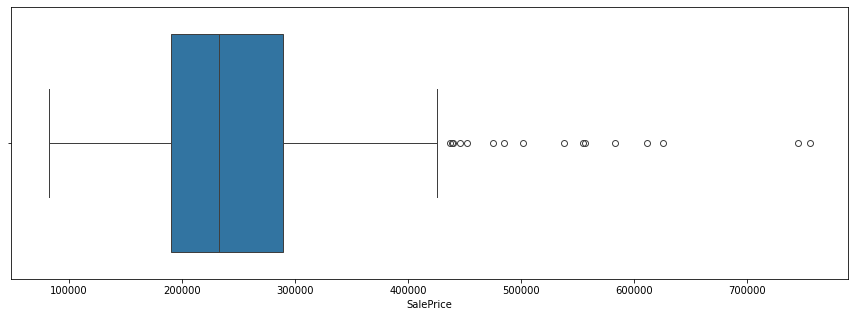
We have used ***two sample t-test*** to check whether quality have effect on sale price or not.

The null hypothesis is NOT.

But the ***P\_VALUE*** near 0 means that our null hypothesis has been rejected and according to following plot houses with higher quality are more expensive.



High and low-quality house price distribution plot

low-quality house price box plot

high-quality house price box plot

**Question3:** How do the different types of heating (Heating) affect the sale prices?

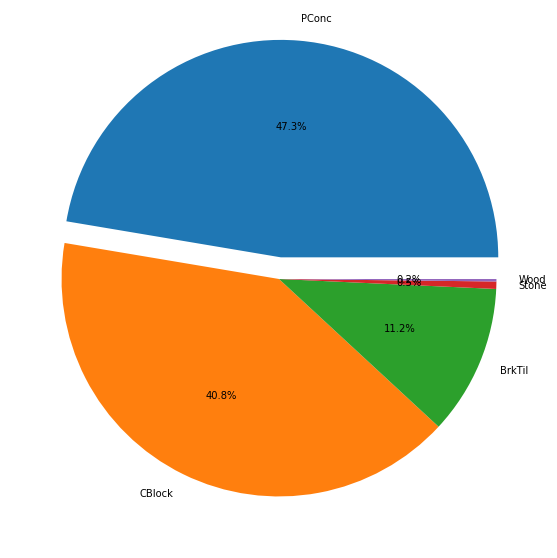
A blue circle with a orange line

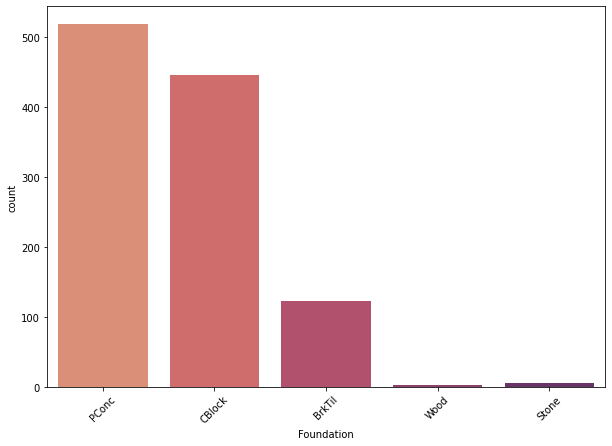
Description automatically generatedAccording to the following plot we do not have enough information and data to answer this question because approximately 99 percent of our houses in this dataset use GasA for heating.

A white and pink rectangle with black border

Description automatically generated

**Question4:** How do the different types of Foundation relate to sale prices?

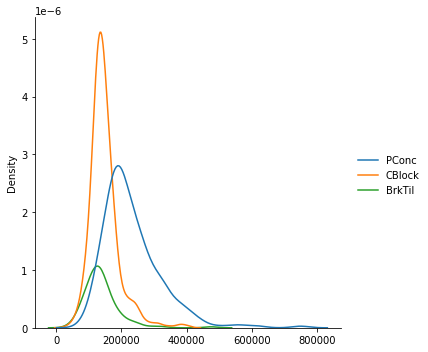
After plotting all categories of foundations, we see that two categories are not very frequent so we did our statistic test on other three remaining categories.



So, we will categorize houses according to their foundation into three groups: PConc, CBlock and BrkTil.

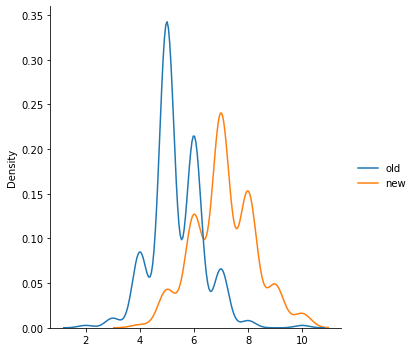
Using ***ANOVA*** test we will check if the average amount of price of all categories is equal or not.

***P\_VALUE*** for this test has been acquired about 0 again so it means that sale price will be change significantly according to the type of foundation. And by looking at the plot below it is clear that houses with PConc foundation have more price compared to other types.



**Question5:** How does Overall quality relate to year built?

To answer this question, we have considered houses which have been built before 1975 as old houses and vice versa.

In this question we just relied on plotting and as you can see and as we expected newer houses are of better quality.

Old and new houses quality distribution plot

**4.Result**

Generally according to the visualizations and tests we did we can mention following results:

* It is expectable but we admit it by done tests and visualizations that

houses with more quality are more expensive to buy.

* New houses have three specialties 1. They are mostly 1 and 2 story

houses, 2. They are more expensive, 3. They have better quality to live

in.

* Foundation is very important and is very defining feature for sale price

of house.

* Also, we couldn’t do any proper test but we understood that most of

the houses use GasA for heating.

* The other important thing we found out about this data is that most of

the houses are of the class 20.

* And at the end this is just an assumption, but it seems that houses of

lower classes e.g 20 and 30 are more expensive and newer than

others.