CPSC 4800 - ASSIGNMENT 3.3

EXPLORATORY DATA ANALYSIS ON HOUSE PRICE DATA SET

Thi Thu Thuy Tran
Student ID: 100420937

Instructor: Nasim Tabatabaei

SUMMARY

Data was collected on almost every aspect of residential homes in Ames, Iowa to determine house selling price. Data set is called HousePrice with:

- 1460 observations and 81 variables: 1 dependent variable (*SalePrice*), 1 index column (PassesgerID) and 79 independent variables in which 36 vriables are quantitative & 43 variables are categorical
- Some variables have significant missing values (more than 30%): *LotFrontage, FireplaceQu, MasVnType, Fence, Alley, Miscfeature,* and *PoolQC* (refer to figure '*Count of missing values'*)

EDA

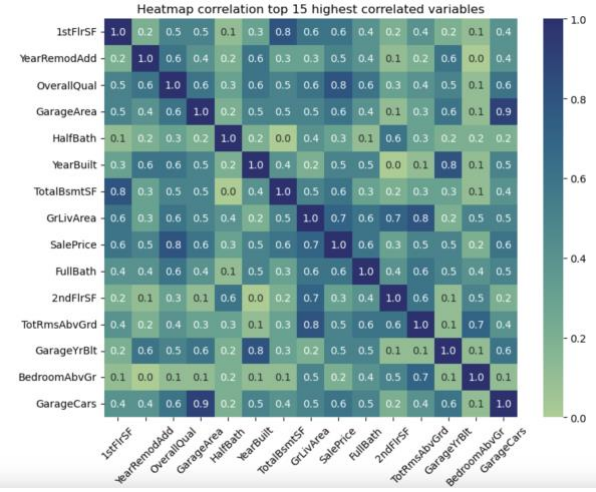
	SalePrice	YearBuilt	LotArea	GrLivArea	BedroomAbvGr	FullBath	OverallQual	GarageArea	GarageCars
min	34900.0	1872.0	1300.0	334.0	0.0	0.0	1.0	0.0	0.0
max	755000.0	2010.0	215245.0	5642.0	8.0	3.0	10.0	1418.0	4.0
mean	180921.0	1971.0	10517.0	1515.0	3.0	2.0	6.0	473.0	2.0
std	79443.0	30.0	9981.0	525.0	1.0	1.0	1.0	214.0	1.0
25%	129975.0	1954.0	7554.0	1130.0	2.0	1.0	5.0	334.0	1.0
50%	163000.0	1973.0	9478.0	1464.0	3.0	2.0	6.0	480.0	2.0
75%	214000.0	2000.0	11602.0	1777.0	3.0	2.0	7.0	576.0	2.0

Summary stats of a few numberical factors that most people tend to start with when searching for a house:

- Sellling price range from 755k, average price is $180k
- The oldest house was built in 1872 and newest one was 2010, majority are built from the 1950s to 2000
- Smallest land is 1300 squared feet while the biggest one is 215k squared feet, average around 10k squared feet
- Living area ranging from 334 to 5642 sqt with average of 1515 sqt
- Each house has 0 to 8 bedrooms above ground level with 0 to 3 full baths
- House quality are ranked on a scale from 1 to 10. The average quality of all houses is 6 with most houses are ranked from 5 to 7.
- Garage area ranges from 0 to 1418 sqt with average of 473 sqt meaning that some houses does not have garage (GarageArea = 0) and some has very big garage (~ 3 times average)
- Garage area as car capacity range from 0 to 4 cars with average of 2 cars. We can expect that it is likely to be in line with garage area

House price distribution: the distribution of house price is skewed to the right with most of selling price between $110k to $200k (refer to figure '*Distribution of Selling Price'*)

Correlation among quantitative variables



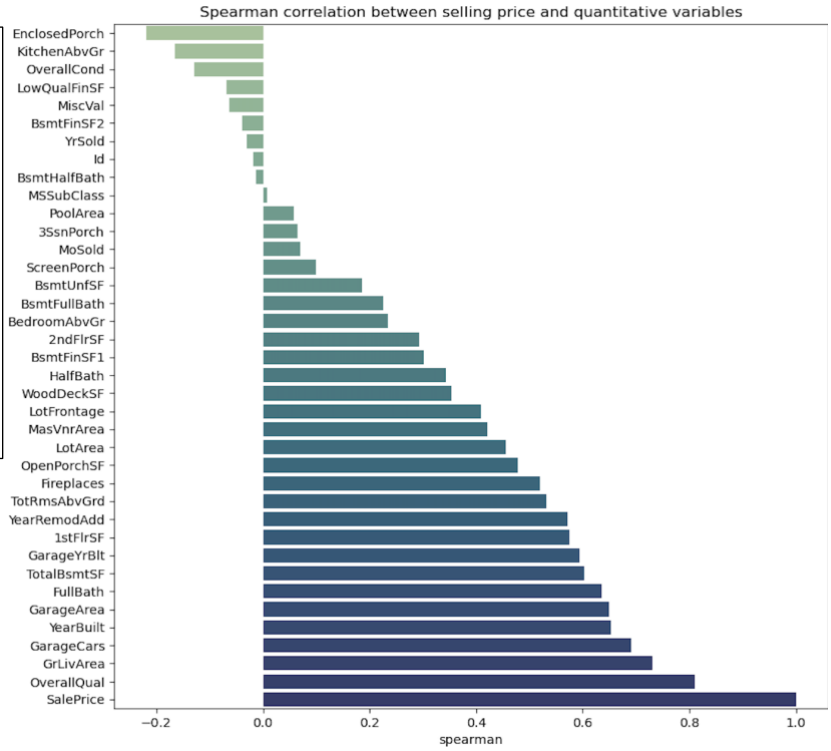
The heatmap on the left aims to illustrate the relationship between quantitative variables of dataset. We pick only 15 highest ones for better visualisation

- The relationship amongs variables looks reasonable e.g number of full bath is positively correlated with living area (about 0.6) or garage area is strongly correlated with garage cars (~0.9)
- Interesting relationship was spotted is that garage area is positively correlated with the year that garage was built (~0.6) meaning newer house tends to have bigger garage, though the correlation does not imply causation.

Correlation between selling price & other quantitative variables

- **Bar chart of spearman correlation** illustrates the correltion of selling price with each of the quantitative variables

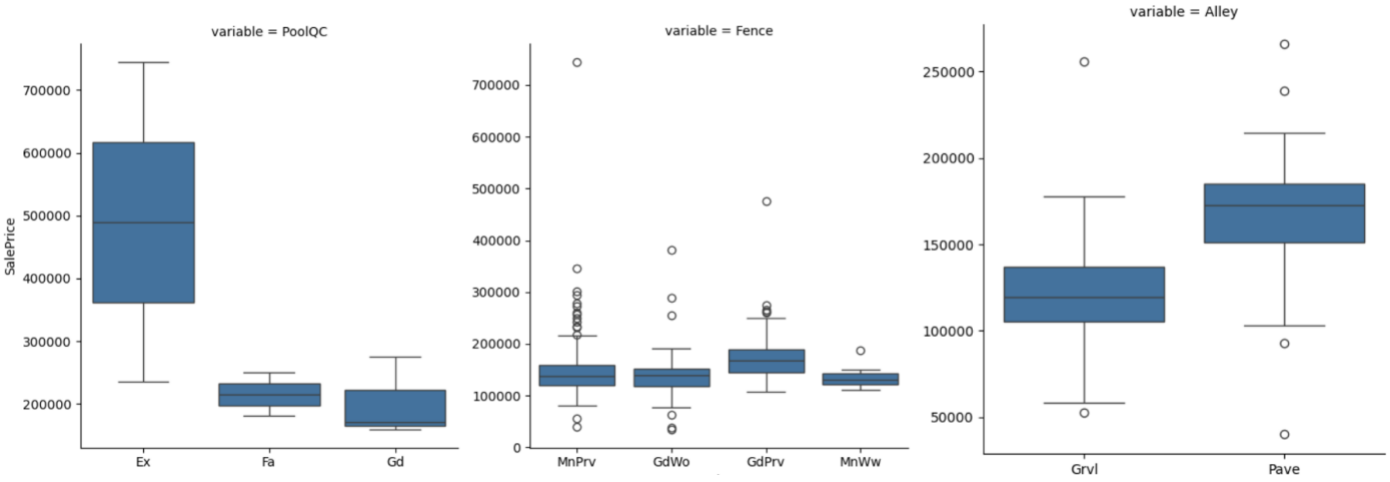
- Variables that are positively & strongly correlated with selling price (> 0.6) are: number of full baths, garage area, year built, size of garage in car capacity, ground living area and Overall material and finish quality.

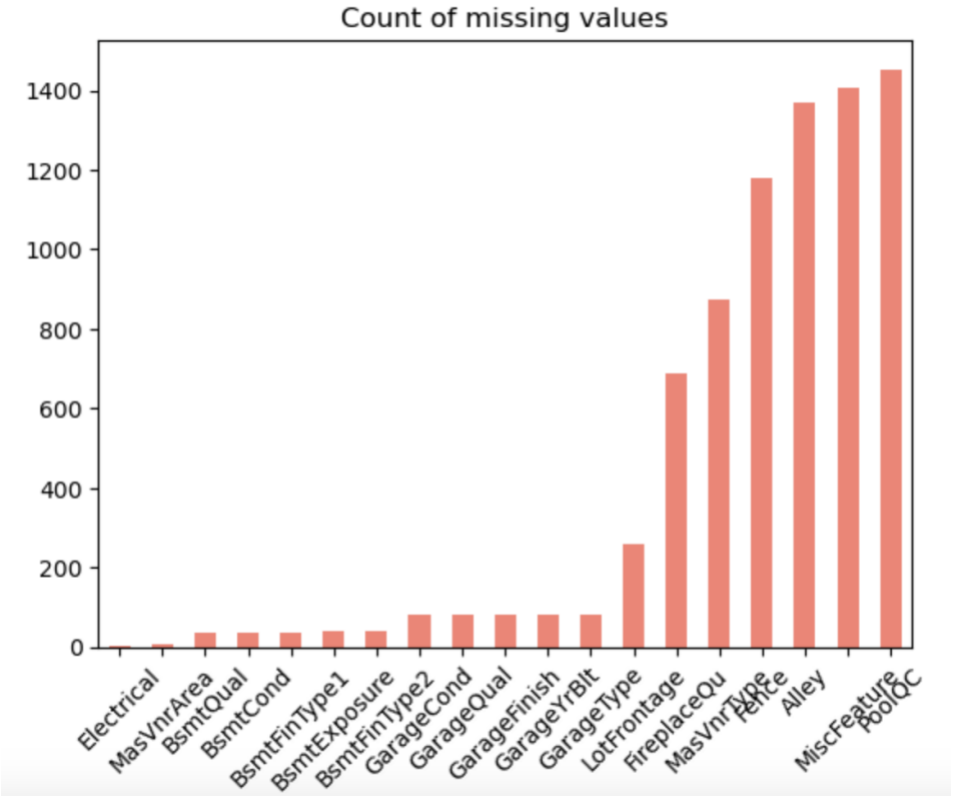


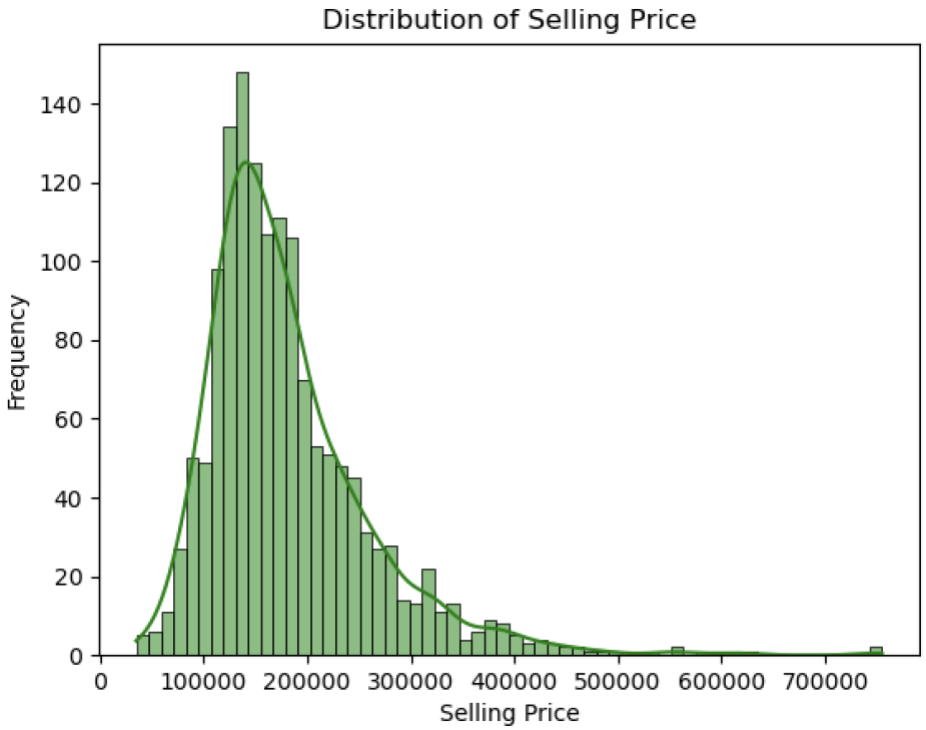
Mean selling price versus different quantitative variables

Box plot helps to illustrates weather or not as well as magnituge that the mean selling price varies among values of each categorical variables

- All box plots indicates that selling price is somewhat depends on buyer's preferences about these categorical variables because the mean selling price varies among value of each categorical variable. For instance:
  - House will pools generally has higher selling price already but those in excellent condition are siginificantly high.
  - differences in the ability to provide priviacy as well as material of fence have influces of selling price but quite minimal
  - Houses with paved alley has signiicantly higher average selling price than those with gravel alley
- Refer to jupyternotebook for box plots of a range of qualitative variables




Count of missing values
1400
1200
1000
800
600
400
200
0
Electrical
MasVnrArea
BsmtQual
BsmtCond
BsmtFinType1
BsmtExposure
BsmtFinType2
GarageCond
GarageQual
GarageFinish
GarageYrBlt
GarageType
LotFrontage
FireplaceQu
MasVnrType
Fence
Alley
MiscFeature
PoolQC


Distribution of Selling Price
Frequency
140
120
100
80
60
40
20
0
0
100000
200000
300000
400000
500000
600000
700000
Selling Price