Explore the Data

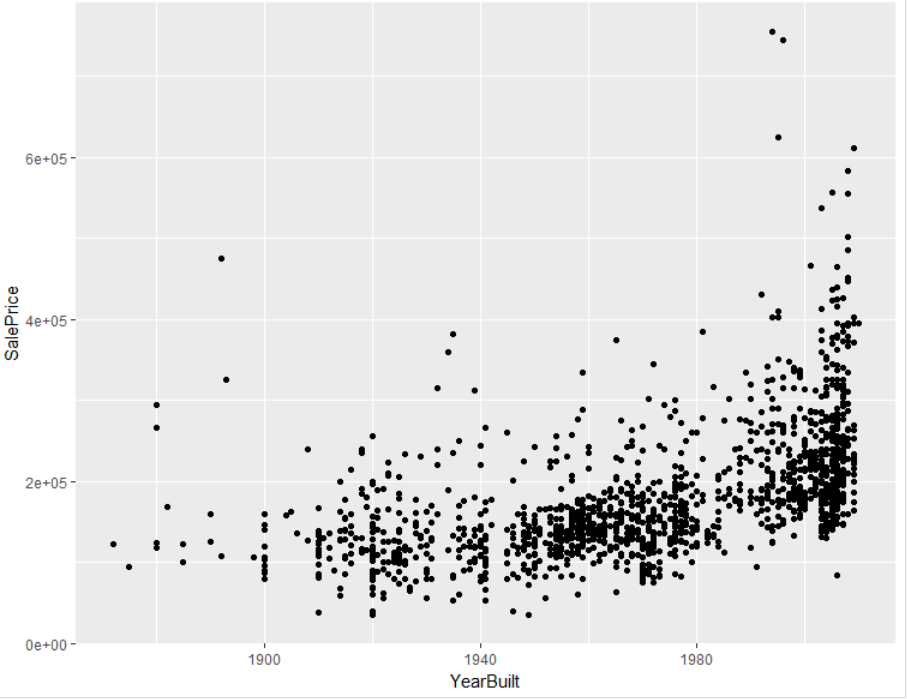
|  |
| --- |
| 'data.frame': 1460 obs. of 81 variables:  $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  $ MSSubClass : int 60 20 60 70 60 50 20 60 50 190 ...  $ MSZoning : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...  $ LotFrontage : int 65 80 68 60 84 85 75 NA 51 50 ...  $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...  $ Street : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...  $ Alley : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...  $ LotShape : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...  $ LandContour : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...  $ Utilities : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...  $ LotConfig : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...  $ LandSlope : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...  $ Condition1 : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...  $ Condition2 : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...  $ BldgType : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...  $ HouseStyle : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...  $ OverallQual : int 7 6 7 7 8 5 8 7 7 5 ...  $ OverallCond : int 5 8 5 5 5 5 5 6 5 6 ...  $ YearBuilt : int 2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...  $ YearRemodAdd : int 2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...  $ RoofStyle : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...  $ RoofMatl : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...  $ Exterior1st : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...  $ Exterior2nd : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...  $ MasVnrType : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...  $ MasVnrArea : int 196 0 162 0 350 0 186 240 0 0 ...  $ ExterQual : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...  $ ExterCond : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...  $ Foundation : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...  $ BsmtQual : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...  $ BsmtCond : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...  $ BsmtFinSF1 : int 706 978 486 216 655 732 1369 859 0 851 ...  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...  $ BsmtFinSF2 : int 0 0 0 0 0 0 0 32 0 0 ...  $ BsmtUnfSF : int 150 284 434 540 490 64 317 216 952 140 ...  $ TotalBsmtSF : int 856 1262 920 756 1145 796 1686 1107 952 991 ...  $ Heating : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...  $ HeatingQC : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...  $ CentralAir : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...  $ Electrical : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...  $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...  $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...  $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...  $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...  $ BsmtFullBath : int 1 0 1 1 1 1 1 1 0 1 ...  $ BsmtHalfBath : int 0 1 0 0 0 0 0 0 0 0 ...  $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...  $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...  $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...  $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...  $ KitchenQual : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...  $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...  $ Functional : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...  $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  $ FireplaceQu : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...  $ GarageType : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...  $ GarageYrBlt : int 2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...  $ GarageCars : int 2 2 2 3 3 2 2 2 2 1 ...  $ GarageArea : int 548 460 608 642 836 480 636 484 468 205 ...  $ GarageQual : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...  $ GarageCond : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...  $ PavedDrive : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...  $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...  $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...  $ EnclosedPorch: int 0 0 0 272 0 0 0 228 205 0 ...  $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...  $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...  $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...  $ PoolQC : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...  $ Fence : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...  $ MiscFeature : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...  $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...  $ MoSold : int 2 5 9 2 12 10 8 11 4 1 ...  $ YrSold : int 2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...  $ SaleType : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...  $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ... |
| The dataset contain 1460 observation and 81 variables. Sale Price is the response variable .Id variables is used to keep count so it not useful here  The other 79 predictors variable seem like can be used to build the model.   |  | | --- | |  | |  | | The table above show some basic statistic of the attribute. And we notice something that for some of the factor attribute such  As garageType, BsmtFinType2 ..etc… NA in this variables does not mean missing value but it mean another factor levels  We will need to rename this NA factor level to None factor level later so our process of building the model will work.  We plot the SalePrice response variable to see what charterisctic it have | |
| |  | | --- | |  | |

Seem like most of the house is in 100k-250k range. The history graph is right skewed and

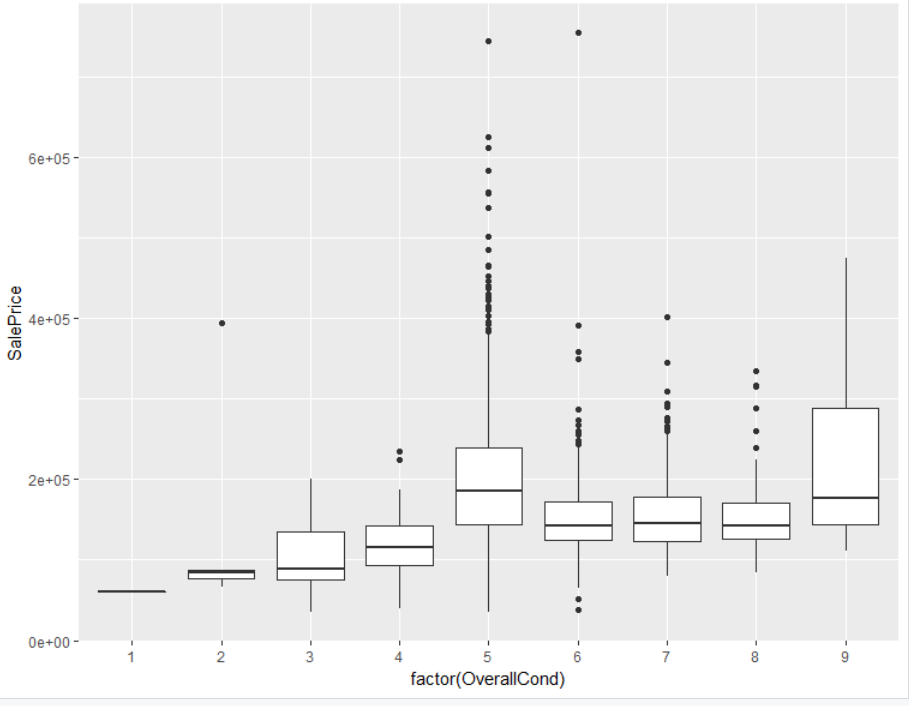
The box plot show there are house that is expensive this maybe outlier that we need to watch for

When building the model.

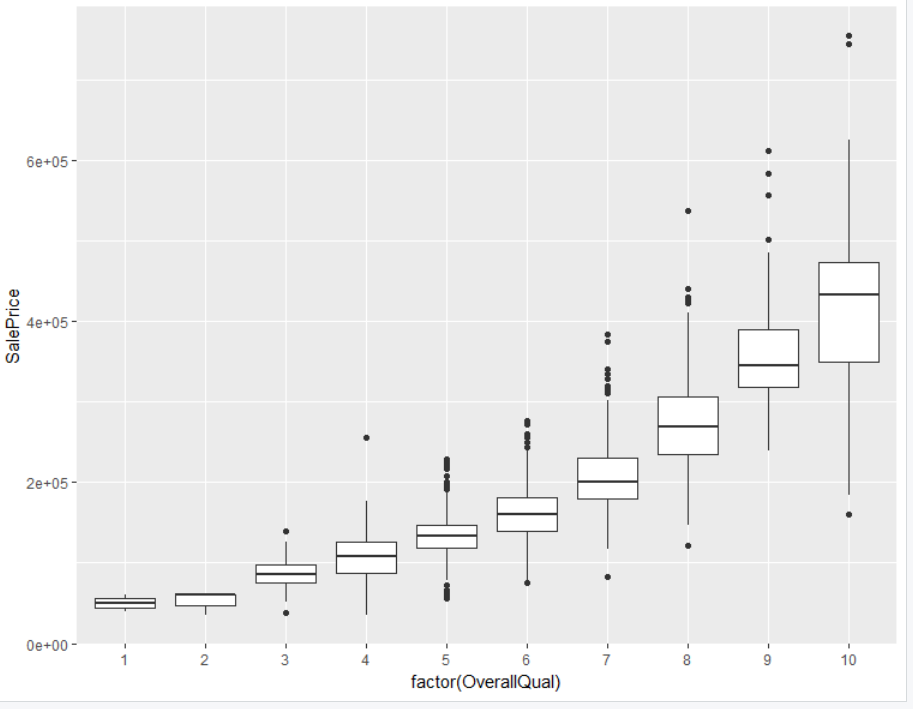
There are some exploratory variables that especally interested to me like the year built, over all condition and over quality rating of the house.



Seem like we have a positive relationship between year built and sale price here . this can be used as a good predictor



The sale price tend to increase as the overall condition rating go from 1 to 5 but does not change much or might even decrease as the over condition rating go from 5 to 9. This might not be a good predictor

wewe

we see as strong positvie relationship here between SalePrice and overall quality rating. Seem to be a good predictors.

Colncusion: the pupose of this initail explonatory analyis is to get the feel of the data, see if the expnotary varibles have missing values or need to be clean up and covert to the right type later.

The response varlbe is right skewed and have outilier that are in the upper price range this maybe

Used to improve our model later. We also did some predictors varibles to see if there is any realationship between those predictors and our response variable.