NB Week 1 EDA

May 23, 2019

1 House Prices: Advanced Regression Techniques

1.1 Introduction:

This project and the data can be found in https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

We will start first with EDA to check the dataset, available rows, the distribution of the sale price (target).

1.2 EDA

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import os
        import sys
        from sklearn.preprocessing import LabelEncoder, OneHotEncoder
        %matplotlib inline
```

```
In [2]: # Read the data
        data = pd.read_csv('../train.csv', index_col=0)
        data.head()
```

```
Out[2]:     MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  \
        Id
        1           60       RL         65.0     8450   Pave   NaN      Reg
        2           20       RL         80.0     9600   Pave   NaN      Reg
        3           60       RL         68.0    11250   Pave   NaN      IR1
        4           70       RL         60.0     9550   Pave   NaN      IR1
        5           60       RL         84.0    14260   Pave   NaN      IR1
```

```
  LandContour Utilities LotConfig    ...     PoolArea PoolQC Fence  \
Id                                   ...
1          Lvl    AllPub    Inside   ...            0    NaN   NaN
2          Lvl    AllPub       FR2   ...            0    NaN   NaN
3          Lvl    AllPub    Inside   ...            0    NaN   NaN
4          Lvl    AllPub    Corner   ...            0    NaN   NaN
5          Lvl    AllPub       FR2   ...            0    NaN   NaN

  MiscFeature MiscVal MoSold  YrSold  SaleType  SaleCondition  SalePrice
Id
1          NaN       0      2    2008        WD         Normal     208500
2          NaN       0      5    2007        WD         Normal     181500
3          NaN       0      9    2008        WD         Normal     223500
4          NaN       0      2    2006        WD        Abnorml     140000
5          NaN       0     12    2008        WD         Normal     250000

[5 rows x 80 columns]
```

In [3]:
```python
# Read the description of the file
with open('../data_description.txt',  'r') as fi:
    print(fi.read())
```

```
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
       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 wi
       Family       Sale between family members
       Partial       Home was not completed when last assessed (associated with New Homes)
```

In [4]: `data.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1460 entries, 1 to 1460
Data columns (total 80 columns):
MSSubClass       1460 non-null int64
MSZoning         1460 non-null object
LotFrontage      1201 non-null float64
LotArea          1460 non-null int64
Street           1460 non-null object
Alley            91 non-null object
LotShape         1460 non-null object
LandContour      1460 non-null object
Utilities        1460 non-null object
LotConfig        1460 non-null object
LandSlope        1460 non-null object
Neighborhood     1460 non-null object
Condition1       1460 non-null object
Condition2       1460 non-null object
BldgType         1460 non-null object
HouseStyle       1460 non-null object
OverallQual      1460 non-null int64
OverallCond      1460 non-null int64
YearBuilt        1460 non-null int64
YearRemodAdd     1460 non-null int64
RoofStyle        1460 non-null object
RoofMatl         1460 non-null object
Exterior1st      1460 non-null object
```

```
Exterior2nd      1460 non-null object
MasVnrType       1452 non-null object
MasVnrArea       1452 non-null float64
ExterQual        1460 non-null object
ExterCond        1460 non-null object
Foundation       1460 non-null object
BsmtQual         1423 non-null object
BsmtCond         1423 non-null object
BsmtExposure     1422 non-null object
BsmtFinType1     1423 non-null object
BsmtFinSF1       1460 non-null int64
BsmtFinType2     1422 non-null object
BsmtFinSF2       1460 non-null int64
BsmtUnfSF        1460 non-null int64
TotalBsmtSF      1460 non-null int64
Heating          1460 non-null object
HeatingQC        1460 non-null object
CentralAir       1460 non-null object
Electrical       1459 non-null object
1stFlrSF         1460 non-null int64
2ndFlrSF         1460 non-null int64
LowQualFinSF     1460 non-null int64
GrLivArea        1460 non-null int64
BsmtFullBath     1460 non-null int64
BsmtHalfBath     1460 non-null int64
FullBath         1460 non-null int64
HalfBath         1460 non-null int64
BedroomAbvGr     1460 non-null int64
KitchenAbvGr     1460 non-null int64
KitchenQual      1460 non-null object
TotRmsAbvGrd     1460 non-null int64
Functional       1460 non-null object
Fireplaces       1460 non-null int64
FireplaceQu      770 non-null object
GarageType       1379 non-null object
GarageYrBlt      1379 non-null float64
GarageFinish     1379 non-null object
GarageCars       1460 non-null int64
GarageArea       1460 non-null int64
GarageQual       1379 non-null object
GarageCond       1379 non-null object
PavedDrive       1460 non-null object
WoodDeckSF       1460 non-null int64
OpenPorchSF      1460 non-null int64
EnclosedPorch    1460 non-null int64
3SsnPorch        1460 non-null int64
ScreenPorch      1460 non-null int64
PoolArea         1460 non-null int64
```

```
PoolQC           7 non-null object
Fence            281 non-null object
MiscFeature      54 non-null object
MiscVal          1460 non-null int64
MoSold           1460 non-null int64
YrSold           1460 non-null int64
SaleType         1460 non-null object
SaleCondition    1460 non-null object
SalePrice        1460 non-null int64
dtypes: float64(3), int64(34), object(43)
memory usage: 923.9+ KB
```

1.2.1 Notes on the feature columns:

The following columns have NA, but NA here indicate something: * Alley column => NA means "No alley access". * BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, BsmtFinSF1 columns => NA means "No Basement" * FireplaceQu column => NA means "No fireplace" * GarageType, GarageCond, GarageFinish columns => NA means "No Garage" * PoolQC column => No Pool * Fence column => No fence * MiscFeature column => None

So be carful with dropping NA values. Next cell, I will fillna as if I didn't pandas will ignore NA values.

```
In [5]: data.Alley = data.Alley.fillna(value = 'NoAlley')
        data.BsmtCond = data.BsmtCond.fillna(value = 'NoBsmt')
        data.BsmtQual = data.BsmtQual.fillna(value = 'NoBsmt')
        data.BsmtExposure = data.BsmtExposure.fillna(value= 'NoBsmt')
        data.BsmtFinType1 = data.BsmtFinType1.fillna(value= 'NoBsmt')
        data.BsmtFinType2 = data.BsmtFinType2.fillna(value= 'NoBsmt')
        data.LotFrontage = data.LotFrontage.fillna(value = 0)
        data.FireplaceQu = data.FireplaceQu.fillna(value = 'Nofireplace')
        data.GarageType = data.GarageType.fillna(value = 'NoGarage')
        data.GarageCond = data.GarageCond.fillna(value = 'NoGarage')
        data.GarageFinish = data.GarageFinish.fillna(value = 'NoGarage')
        data.GarageYrBlt = data.GarageYrBlt.fillna(value = 0)
        data.GarageQual = data.GarageQual.fillna(value = 'NoGarage')

        data.PoolQC = data.PoolQC.fillna(value = 'NoPool')
        data.Fence = data.Fence.fillna(value = 'NoFence')
        data.MiscFeature = data.MiscFeature.fillna(value = 'NoMisc')
        data.MasVnrType = data.MasVnrType.fillna(value = 'noMas')
        data.MasVnrArea = data.MasVnrArea.fillna(value = 'noMas')

        data.info()
```

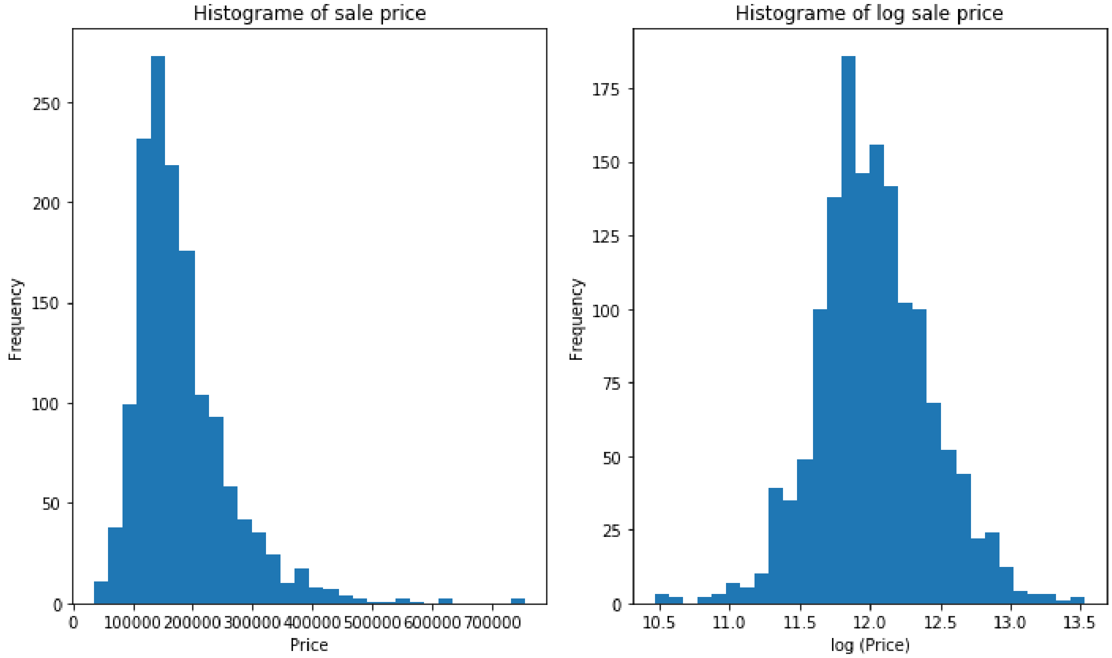
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1460 entries, 1 to 1460
```

```
Data columns (total 80 columns):
MSSubClass       1460 non-null int64
MSZoning         1460 non-null object
LotFrontage      1460 non-null float64
LotArea          1460 non-null int64
Street           1460 non-null object
Alley            1460 non-null object
LotShape         1460 non-null object
LandContour      1460 non-null object
Utilities        1460 non-null object
LotConfig        1460 non-null object
LandSlope        1460 non-null object
Neighborhood     1460 non-null object
Condition1       1460 non-null object
Condition2       1460 non-null object
BldgType         1460 non-null object
HouseStyle       1460 non-null object
OverallQual      1460 non-null int64
OverallCond      1460 non-null int64
YearBuilt        1460 non-null int64
YearRemodAdd     1460 non-null int64
RoofStyle        1460 non-null object
RoofMatl         1460 non-null object
Exterior1st      1460 non-null object
Exterior2nd      1460 non-null object
MasVnrType       1460 non-null object
MasVnrArea       1460 non-null object
ExterQual        1460 non-null object
ExterCond        1460 non-null object
Foundation       1460 non-null object
BsmtQual         1460 non-null object
BsmtCond         1460 non-null object
BsmtExposure     1460 non-null object
BsmtFinType1     1460 non-null object
BsmtFinSF1       1460 non-null int64
BsmtFinType2     1460 non-null object
BsmtFinSF2       1460 non-null int64
BsmtUnfSF        1460 non-null int64
TotalBsmtSF      1460 non-null int64
Heating          1460 non-null object
HeatingQC        1460 non-null object
CentralAir       1460 non-null object
Electrical       1459 non-null object
1stFlrSF         1460 non-null int64
2ndFlrSF         1460 non-null int64
LowQualFinSF     1460 non-null int64
GrLivArea        1460 non-null int64
BsmtFullBath     1460 non-null int64
```

```
BsmtHalfBath     1460 non-null int64
FullBath         1460 non-null int64
HalfBath         1460 non-null int64
BedroomAbvGr     1460 non-null int64
KitchenAbvGr     1460 non-null int64
KitchenQual      1460 non-null object
TotRmsAbvGrd     1460 non-null int64
Functional       1460 non-null object
Fireplaces       1460 non-null int64
FireplaceQu      1460 non-null object
GarageType       1460 non-null object
GarageYrBlt      1460 non-null float64
GarageFinish     1460 non-null object
GarageCars       1460 non-null int64
GarageArea       1460 non-null int64
GarageQual       1460 non-null object
GarageCond       1460 non-null object
PavedDrive       1460 non-null object
WoodDeckSF       1460 non-null int64
OpenPorchSF      1460 non-null int64
EnclosedPorch    1460 non-null int64
3SsnPorch        1460 non-null int64
ScreenPorch      1460 non-null int64
PoolArea         1460 non-null int64
PoolQC           1460 non-null object
Fence            1460 non-null object
MiscFeature      1460 non-null object
MiscVal          1460 non-null int64
MoSold           1460 non-null int64
YrSold           1460 non-null int64
SaleType         1460 non-null object
SaleCondition    1460 non-null object
SalePrice        1460 non-null int64
dtypes: float64(2), int64(34), object(44)
memory usage: 923.9+ KB
```

```
In [9]: fig = plt.figure(figsize=(10,6))
        plt.subplot(121)
        plt.hist(data.SalePrice, bins=30)
        plt.xlabel('Price')
        plt.ylabel('Frequency')
        plt.title('Histograme of sale price');
        plt.subplot(122)
        plt.hist(np.log(data.SalePrice), bins=30)
        plt.xlabel('log (Price)')
        plt.ylabel('Frequency')
        plt.title('Histograme of log sale price')
```

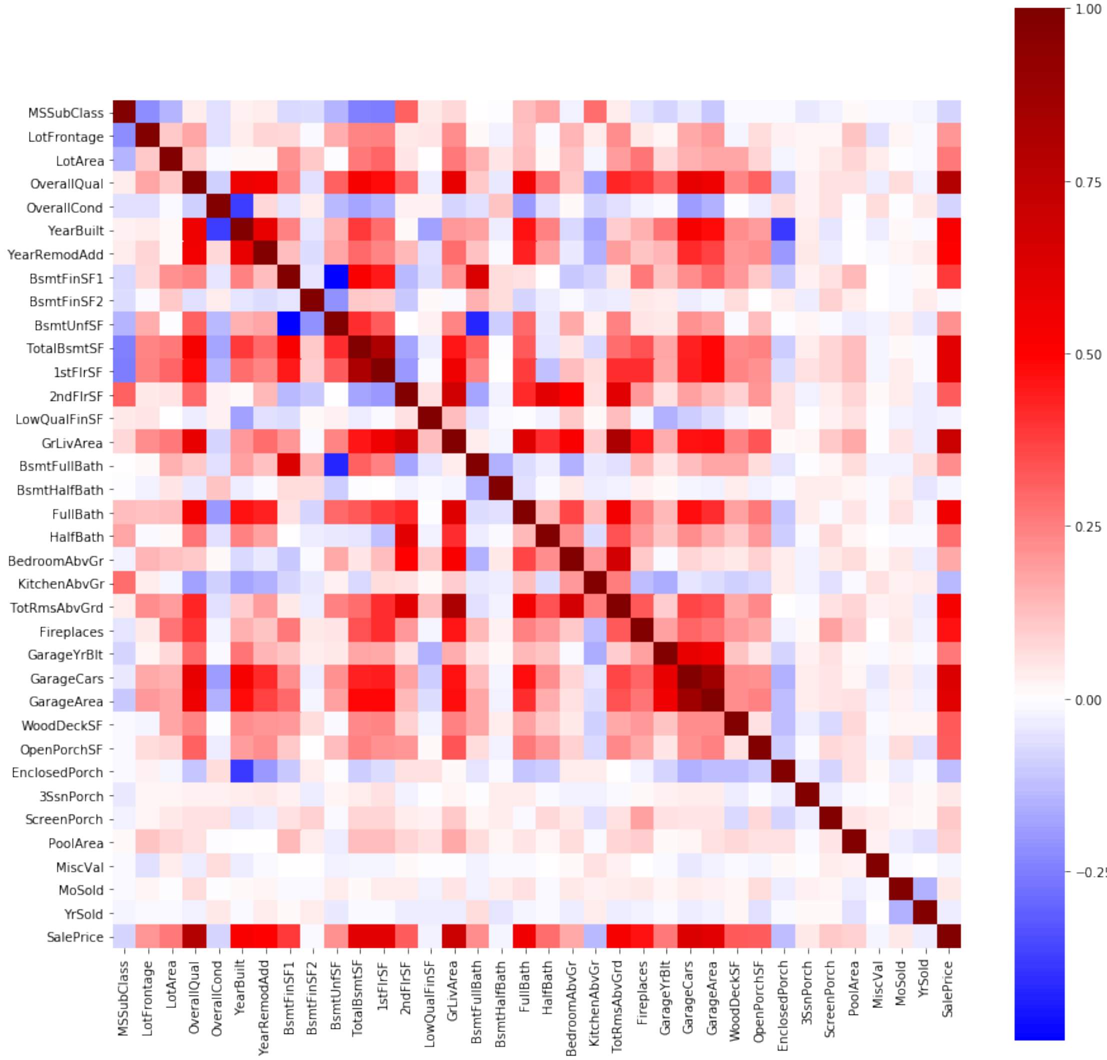
```
plt.tight_layout()
```



The plot of log(sale price) looks normal without any outliers.

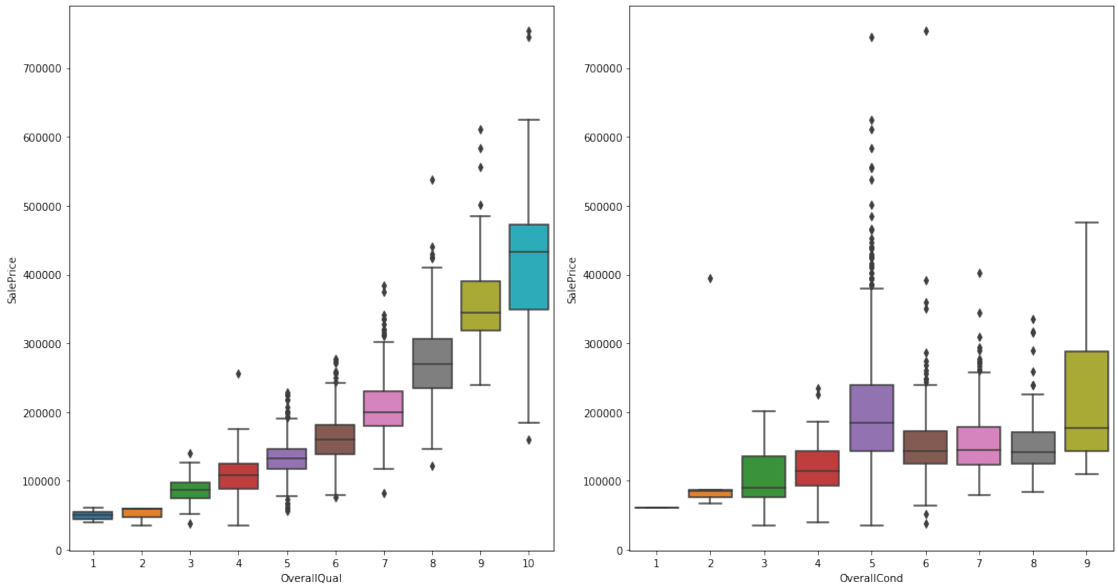
1.3 EDA for Numerical Columns:

```
In [7]: # heatmap of the Sale price, with the numerical columns
        corr = data.corr()
        fig, ax = plt.subplots(figsize=(15,15))
        sns.heatmap(corr, square=True, ax=ax, cmap='seismic', center= 0.0)
        plt.xticks(fontsize=10);
        plt.yticks(fontsize=10);
```



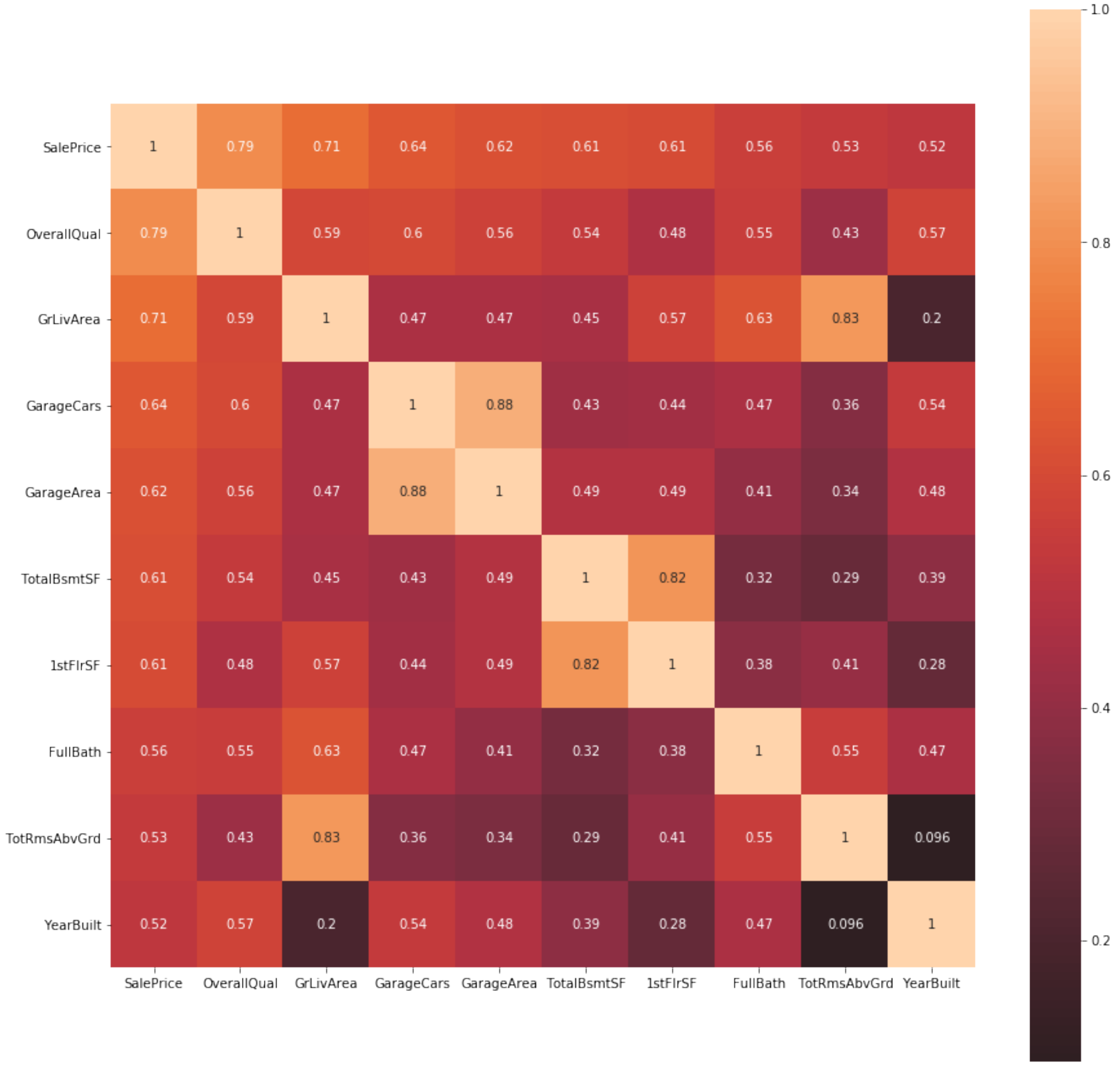
- From the above figure, there are some features which have high corrolation with the "Sale price" column, most of them with positive corrolation.
- It is interesting to find that OverallQual has high corrolation with the Sale price, on the otherhand Overallcond has a small corrolation factor with sale price.

In [8]:
```
# Check OverallQual and OverallCond columns
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(121)
g = sns.catplot(x="OverallQual", y="SalePrice", kind="box", data=data, ax = ax)
plt.close(g.fig)
ax = fig.add_subplot(122)
g = sns.catplot(x="OverallCond", y="SalePrice", kind="box", data=data, ax = ax)
plt.close(g.fig)

plt.tight_layout()
```



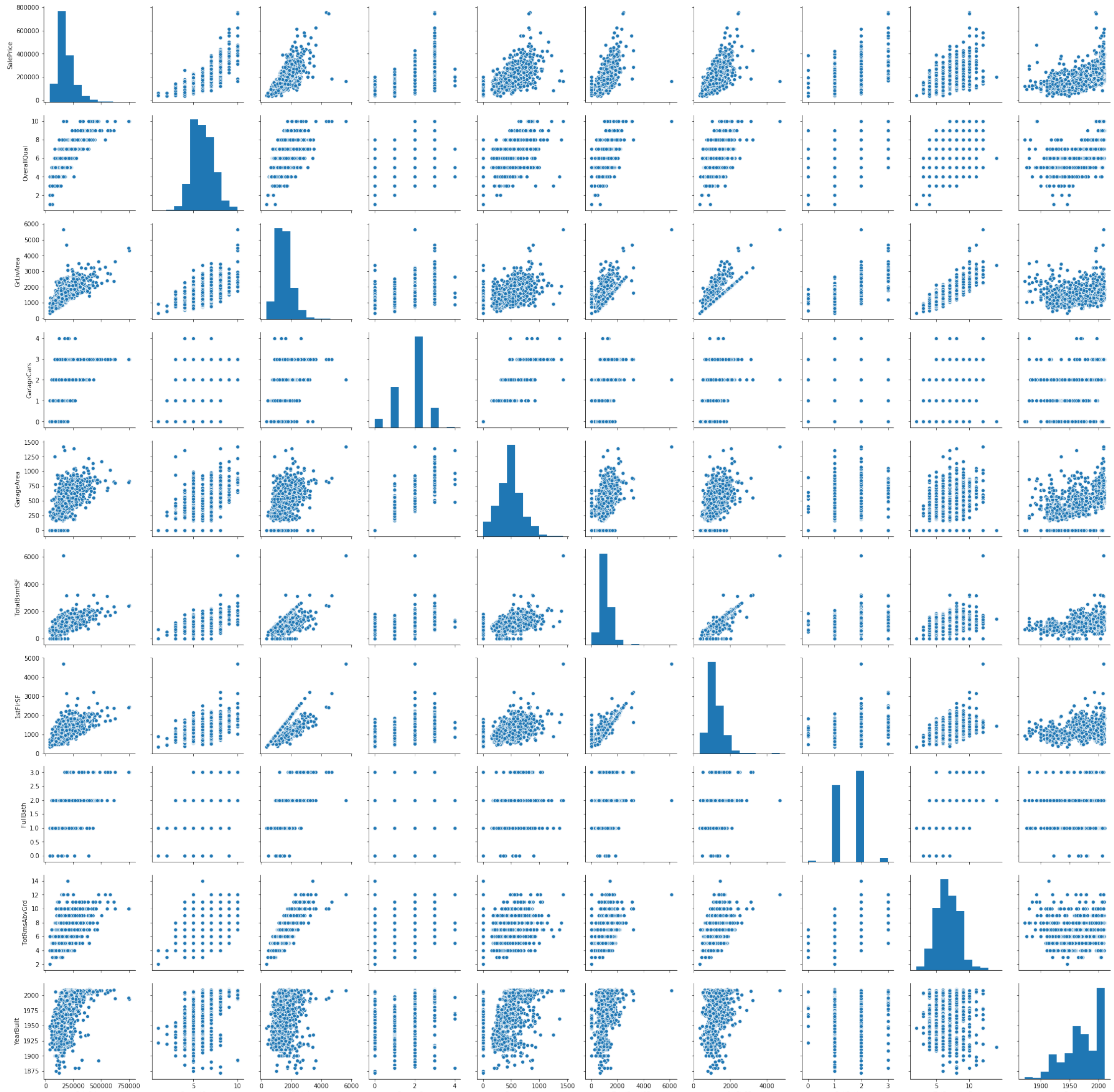
- It make sense now why OverallQual gives high corrolation with saleprice than OverallCond

```
In [9]: corr = data.corr()
        cols = corr.nlargest(10, 'SalePrice')['SalePrice'].index    # Take the max 10
        corr = data[cols].corr()
        fig, ax = plt.subplots(figsize=(15,15))
        sns.heatmap(corr, square=True, ax=ax, center= 0.0, xticklabels=cols, yticklabels=cols,
        plt.xticks(fontsize=10);
        plt.yticks(fontsize=10);
```



- Numerical columns with high corrolation with Sale price are: ['SalePrice', 'OverallQual', 'GrLivArea', 'GarageCars', 'GarageArea','TotalBsmtSF', '1stFlrSF', 'FullBath', 'TotRmsAbvGrd', 'YearBuilt']

```
In [10]: sns.pairplot(data[cols]);
```

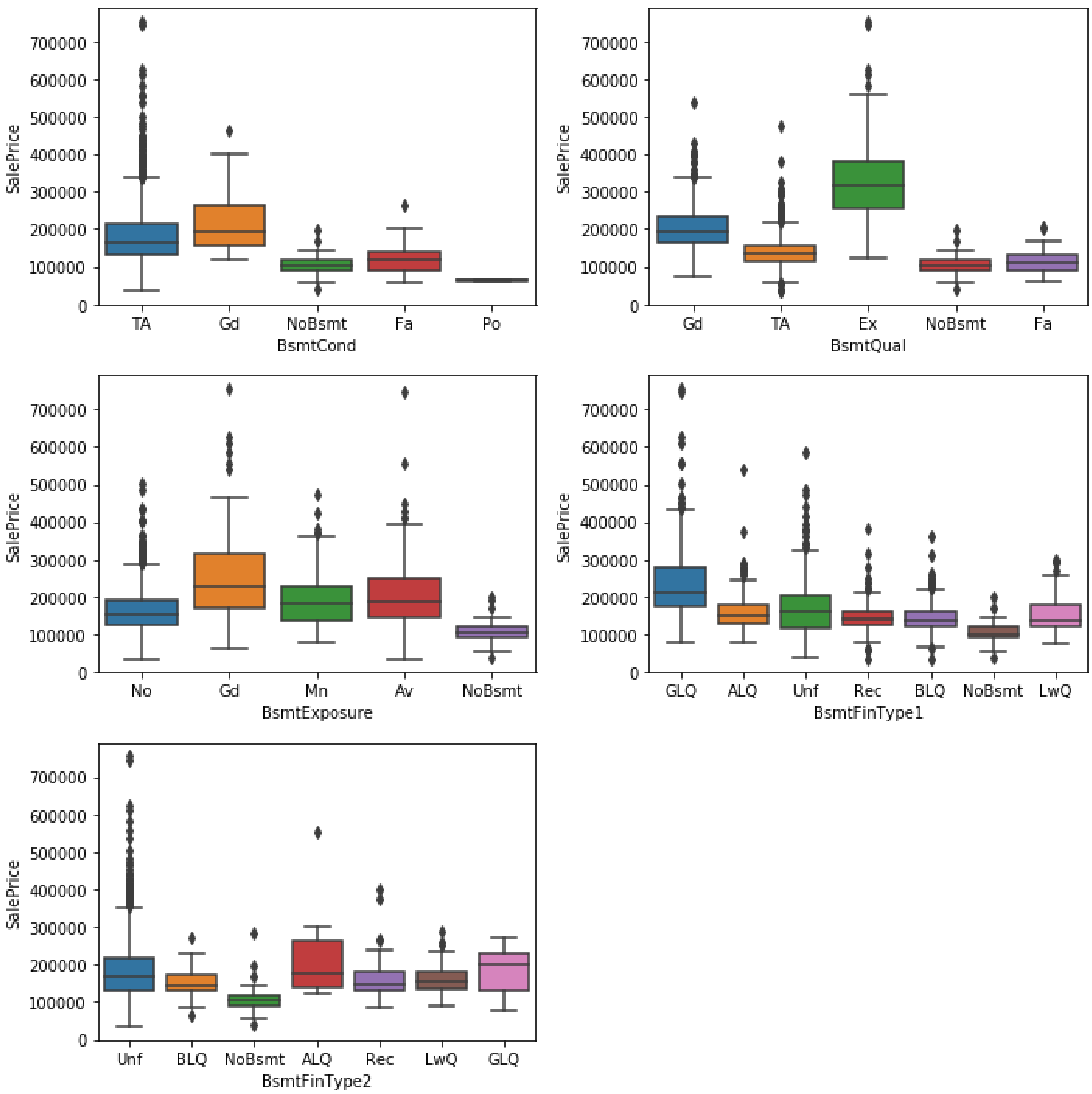


The following columns are integer and they give different corrolation with sale price: * Garage-Cars: Size of garage in car capacity, * FullBath: Full bathrooms above grade, and * TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

1.4 EDA for Categorical columns

```
In [11]: # Columns related to Basement
         fig = plt.figure(figsize=(10,10))
         ax = fig.add_subplot(321)
         g = sns.catplot(x="BsmtCond", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         ax = fig.add_subplot(322)
         g = sns.catplot(x="BsmtQual", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
```

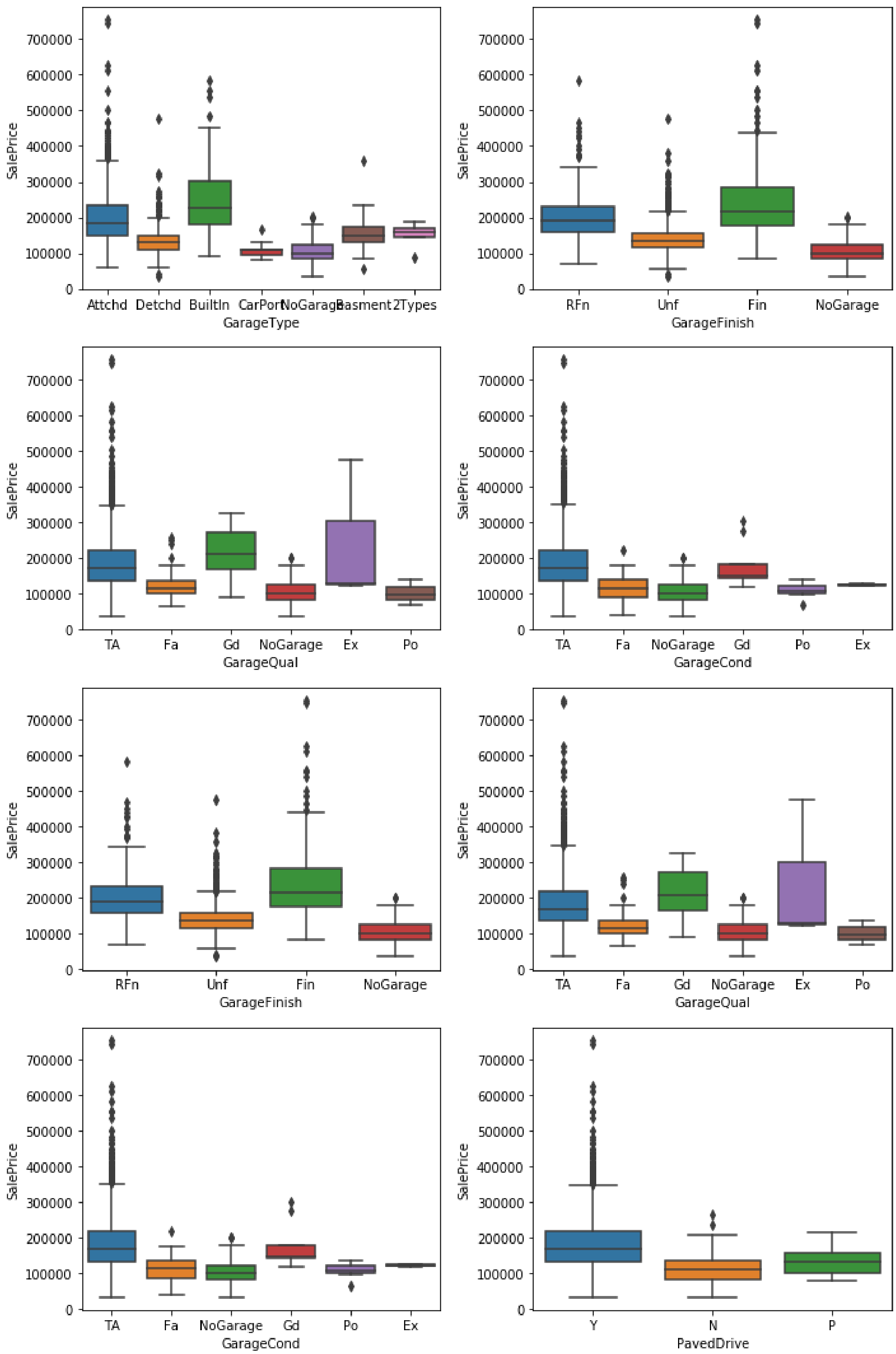
```
ax = fig.add_subplot(323)
g = sns.catplot(x="BsmtExposure", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(324)
g = sns.catplot(x="BsmtFinType1", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(325)
g = sns.catplot(x="BsmtFinType2", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
plt.tight_layout()
```



In [12]:
```
# I will use only BsmtCond and BsmtQual and drop the rest
# It is better to use lable encoder for these columns than one-hot code:
```

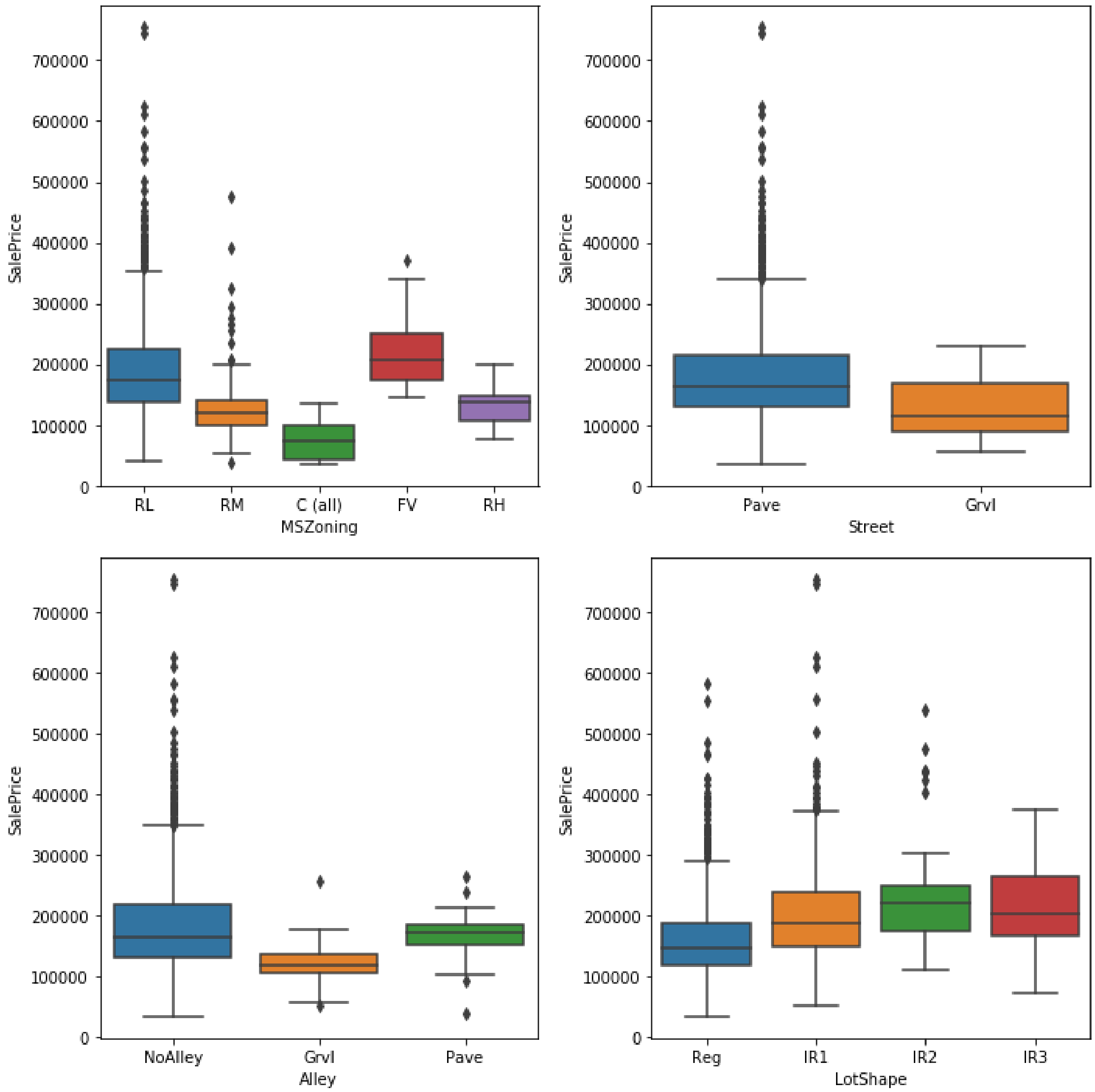
```
data.BsmtCond = data.BsmtCond.map({'Ex':5 ,'Gd':4 , 'TA':3 ,'Fa':2 ,'Po':1 , 'NoBsmt'
data.BsmtQual = data.BsmtQual.map({'Ex':5 ,'Gd':4 , 'TA':3 ,'Fa':2 ,'Po':1 , 'NoBsmt'
data.BsmtExposure = data.BsmtExposure.map({'Gd':4, 'Av':3, 'Mn':2, 'No':1, 'NoBsmt':0
data.BsmtFinType1 = data.BsmtFinType1.map({'GLQ':6,'ALQ':5,'BLQ':4,'Rec':3,'LwQ':2,'U
data.BsmtFinType2 = data.BsmtFinType2.map({'GLQ':6,'ALQ':5,'BLQ':4,'Rec':3,'LwQ':2,'U
```

In [13]:
```
# Columns related to Garage
fig = plt.figure(figsize=(10,15))
ax = fig.add_subplot(421)
g = sns.catplot(x="GarageType", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(422)
g = sns.catplot(x="GarageFinish", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(423)
g = sns.catplot(x="GarageQual", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(424)
g = sns.catplot(x="GarageCond", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(425)
g = sns.catplot(x="GarageFinish", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(426)
g = sns.catplot(x="GarageQual", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(427)
g = sns.catplot(x="GarageCond", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(428)
g = sns.catplot(x="PavedDrive", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)

plt.tight_layout()
```

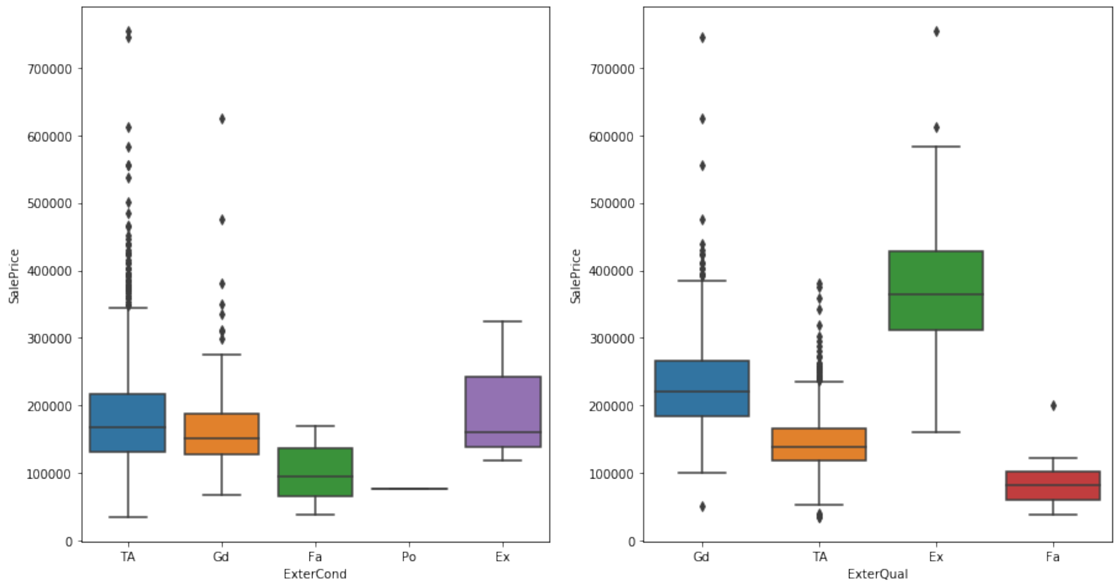

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Attchd
Detchd
BuiltIn
CarPort
NoGarage
Basment
2Types
GarageType
RFn
Unf
Fin
NoGarage
GarageFinish
TA
Fa
Gd
NoGarage
Ex
Po
GarageQual
TA
Fa
NoGarage
Gd
Po
Ex
GarageCond
Y
N
P
PavedDrive

In [14]:
```
data.GarageType = data.GarageType.map({'2Types':4 , 'Attchd': 5, 'Basment':3 ,'BuiltI
                                       'CarPort' :1, 'Detchd':2 , 'NoGarage': 0})

data.GarageCond = data.GarageCond.map({'NoGarage':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4,
data.GarageQual = data.GarageQual.map({'NoGarage':0, 'Po':1, 'Fa':2, 'TA':3, 'Gd':4,
data.GarageFinish = data.GarageFinish.map({'Fin':3, 'RFn':2, 'Unf':1, 'NoGarage':0})
data.PavedDrive = data.PavedDrive.map({'Y':2,'P':1, 'N':0 })
```

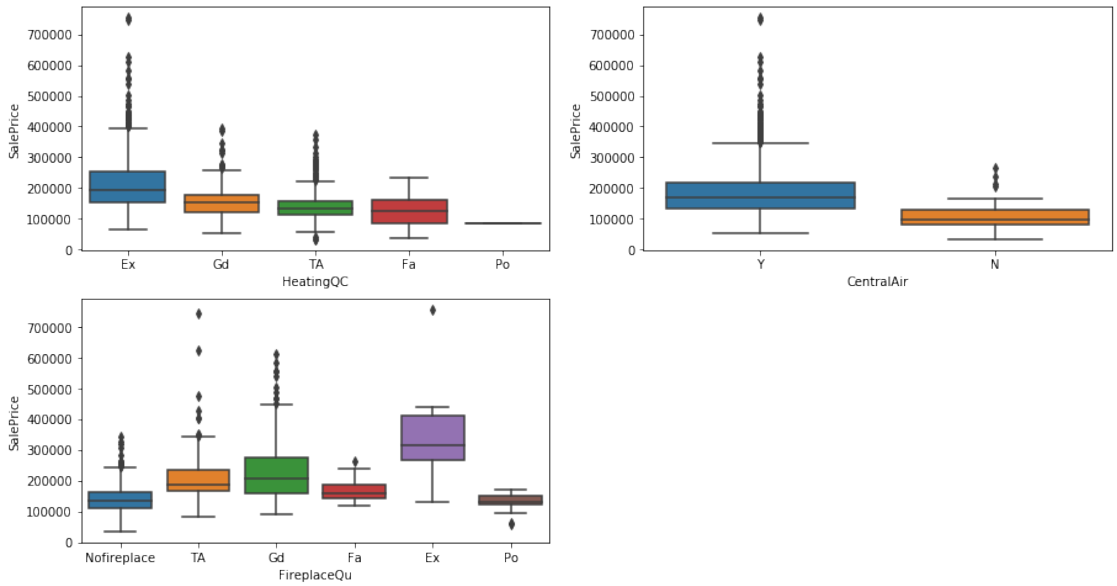
In [15]:
```
# Columns related to surrounding condition
fig = plt.figure(figsize=(10,10))
ax = fig.add_subplot(221)
g = sns.catplot(x="MSZoning", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(222)
g = sns.catplot(x="Street", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(223)
g = sns.catplot(x="Alley", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)
ax = fig.add_subplot(224)
g = sns.catplot(x="LotShape", y="SalePrice", kind="box", data=data, ax=ax)
plt.close(g.fig)

plt.tight_layout()
```



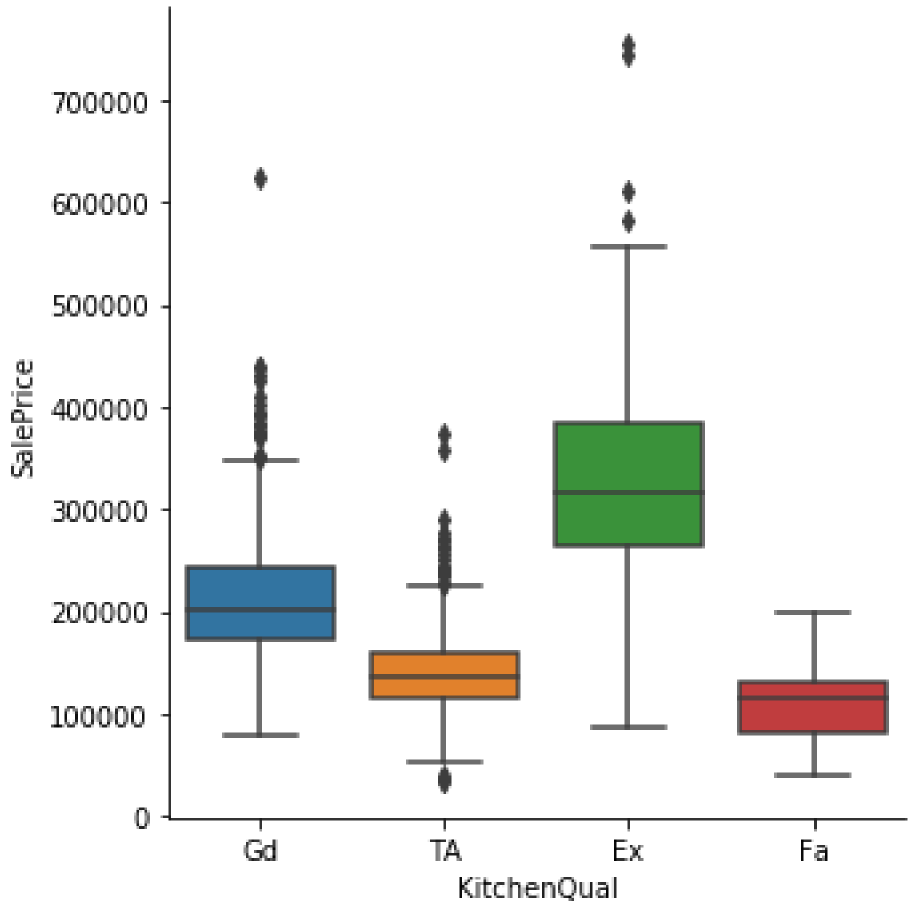

```
In [16]: # Columns related to surrounding condition
         fig = plt.figure(figsize=(15,8))
         ax = fig.add_subplot(121)
         g = sns.catplot(x="ExterCond", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         ax = fig.add_subplot(122)
         g = sns.catplot(x="ExterQual", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         data.ExterCond = data.ExterCond.map({"Ex":4,'Gd':3,'TA':2,'Fa':1,'Po':0})
         data.ExterQual = data.ExterQual.map({"Ex":4,'Gd':3,'TA':2,'Fa':1,'Po':0})
```



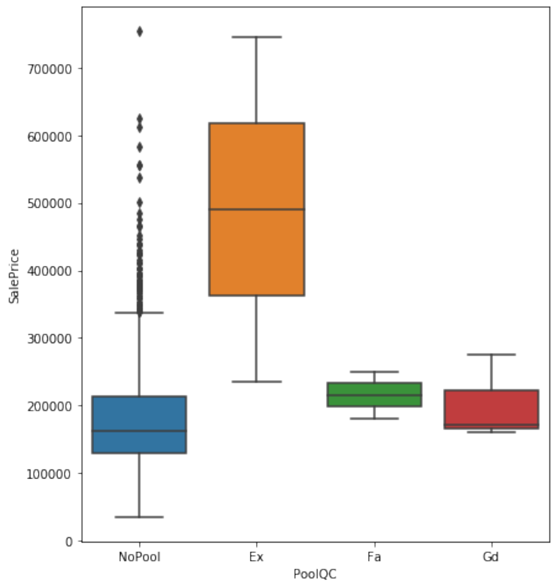

```
In [17]: fig = plt.figure(figsize=(15,8))
         ax = fig.add_subplot(221)
         g = sns.catplot(x="HeatingQC", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         ax = fig.add_subplot(222)
         g = sns.catplot(x="CentralAir", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         ax = fig.add_subplot(223)
         g = sns.catplot(x="FireplaceQu", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         data.CentralAir = data.CentralAir.map({'Y':1, 'N':0})
         data.HeatingQC = data.HeatingQC.map({"Ex":4,'Gd':3,'TA':2,'Fa':1,'Po':0})
         data.FireplaceQu = data.FireplaceQu.map({"Ex":5,'Gd':4,'TA':3,'Fa':2,'Po':1, 'Nofirep
```

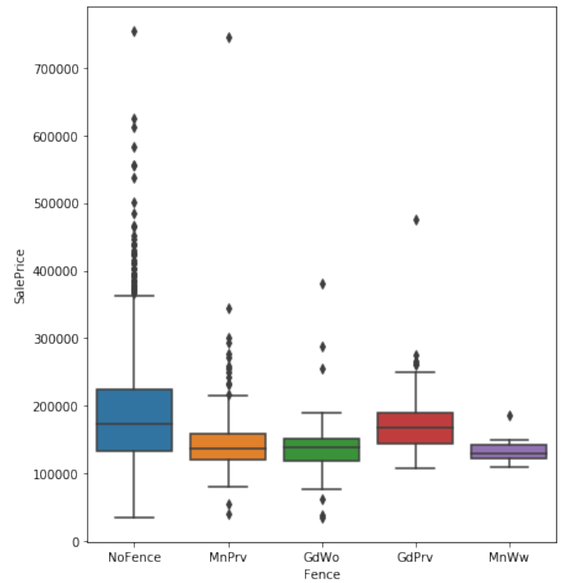



```
In [18]: sns.catplot(x="KitchenQual", y="SalePrice", kind="box", data=data)
         data.KitchenQual = data.KitchenQual.map({"Ex":4,'Gd':3,'TA':2,'Fa':1,'Po':0})
```




```
In [19]: fig = plt.figure(figsize=(15,8))
         ax = fig.add_subplot(121)
         g = sns.catplot(x="PoolQC", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         ax = fig.add_subplot(122)
         g = sns.catplot(x="Fence", y="SalePrice", kind="box", data=data, ax=ax)
         plt.close(g.fig)
         data.PoolQC = data.PoolQC.map({"Ex":4,'Gd':3,'TA':2,'Fa':1,'NoPool':0})
         data.Fence = data.Fence.map({'GdPrv':4 , 'MnPrv':3 , 'GdWo':2 , 'MnWw':1 , 'NoFence':
```





1.5 Converting categorical columns

- The task no is to convert the rest of the categorical columns into one-hot code.

In [20]: data.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1460 entries, 1 to 1460
Data columns (total 80 columns):
MSSubClass        1460 non-null int64
MSZoning          1460 non-null object
LotFrontage       1460 non-null float64
LotArea           1460 non-null int64
Street            1460 non-null object
Alley             1460 non-null object
LotShape          1460 non-null object
LandContour       1460 non-null object
Utilities         1460 non-null object
LotConfig         1460 non-null object
LandSlope         1460 non-null object
Neighborhood      1460 non-null object
Condition1        1460 non-null object
Condition2        1460 non-null object
BldgType          1460 non-null object
HouseStyle        1460 non-null object
OverallQual       1460 non-null int64
OverallCond       1460 non-null int64
YearBuilt         1460 non-null int64
```

```
YearRemodAdd      1460 non-null int64
RoofStyle         1460 non-null object
RoofMatl          1460 non-null object
Exterior1st       1460 non-null object
Exterior2nd       1460 non-null object
MasVnrType        1460 non-null object
MasVnrArea        1460 non-null object
ExterQual         1460 non-null int64
ExterCond         1460 non-null int64
Foundation        1460 non-null object
BsmtQual          1460 non-null int64
BsmtCond          1460 non-null int64
BsmtExposure      1460 non-null int64
BsmtFinType1      1460 non-null int64
BsmtFinSF1        1460 non-null int64
BsmtFinType2      1460 non-null int64
BsmtFinSF2        1460 non-null int64
BsmtUnfSF         1460 non-null int64
TotalBsmtSF       1460 non-null int64
Heating           1460 non-null object
HeatingQC         1460 non-null int64
CentralAir        1460 non-null int64
Electrical        1459 non-null object
1stFlrSF          1460 non-null int64
2ndFlrSF          1460 non-null int64
LowQualFinSF      1460 non-null int64
GrLivArea         1460 non-null int64
BsmtFullBath      1460 non-null int64
BsmtHalfBath      1460 non-null int64
FullBath          1460 non-null int64
HalfBath          1460 non-null int64
BedroomAbvGr      1460 non-null int64
KitchenAbvGr      1460 non-null int64
KitchenQual       1460 non-null int64
TotRmsAbvGrd      1460 non-null int64
Functional        1460 non-null object
Fireplaces        1460 non-null int64
FireplaceQu       1460 non-null int64
GarageType        1460 non-null int64
GarageYrBlt       1460 non-null float64
GarageFinish      1460 non-null int64
GarageCars        1460 non-null int64
GarageArea        1460 non-null int64
GarageQual        1460 non-null int64
GarageCond        1460 non-null int64
PavedDrive        1460 non-null int64
WoodDeckSF        1460 non-null int64
OpenPorchSF       1460 non-null int64
```

```
EnclosedPorch    1460 non-null int64
3SsnPorch        1460 non-null int64
ScreenPorch      1460 non-null int64
PoolArea         1460 non-null int64
PoolQC           1460 non-null int64
Fence            1460 non-null int64
MiscFeature      1460 non-null object
MiscVal          1460 non-null int64
MoSold           1460 non-null int64
YrSold           1460 non-null int64
SaleType         1460 non-null object
SaleCondition    1460 non-null object
SalePrice        1460 non-null int64
dtypes: float64(2), int64(52), object(26)
memory usage: 963.9+ KB
```

```
In [21]: data.Electrical.unique()  # One missing value in this column
```

```
Out[21]: array(['SBrkr', 'FuseF', 'FuseA', 'FuseP', 'Mix', nan], dtype=object)
```

We have only one missing observation, I will drop it.

```
In [22]: data.dropna(inplace=True)
```

```
In [27]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1459 entries, 1 to 1460
Data columns (total 80 columns):
MSSubClass       1459 non-null int64
MSZoning         1459 non-null object
LotFrontage      1459 non-null float64
LotArea          1459 non-null int64
Street           1459 non-null object
Alley            1459 non-null object
LotShape         1459 non-null object
LandContour      1459 non-null object
Utilities        1459 non-null object
LotConfig        1459 non-null object
LandSlope        1459 non-null object
Neighborhood     1459 non-null object
Condition1       1459 non-null object
Condition2       1459 non-null object
BldgType         1459 non-null object
HouseStyle       1459 non-null object
OverallQual      1459 non-null int64
OverallCond      1459 non-null int64
YearBuilt        1459 non-null int64
```

```
YearRemodAdd     1459 non-null int64
RoofStyle        1459 non-null object
RoofMatl         1459 non-null object
Exterior1st      1459 non-null object
Exterior2nd      1459 non-null object
MasVnrType       1459 non-null object
MasVnrArea       1459 non-null object
ExterQual        1459 non-null int64
ExterCond        1459 non-null int64
Foundation       1459 non-null object
BsmtQual         1459 non-null int64
BsmtCond         1459 non-null int64
BsmtExposure     1459 non-null int64
BsmtFinType1     1459 non-null int64
BsmtFinSF1       1459 non-null int64
BsmtFinType2     1459 non-null int64
BsmtFinSF2       1459 non-null int64
BsmtUnfSF        1459 non-null int64
TotalBsmtSF      1459 non-null int64
Heating          1459 non-null object
HeatingQC        1459 non-null int64
CentralAir       1459 non-null int64
Electrical       1459 non-null object
1stFlrSF         1459 non-null int64
2ndFlrSF         1459 non-null int64
LowQualFinSF     1459 non-null int64
GrLivArea        1459 non-null int64
BsmtFullBath     1459 non-null int64
BsmtHalfBath     1459 non-null int64
FullBath         1459 non-null int64
HalfBath         1459 non-null int64
BedroomAbvGr     1459 non-null int64
KitchenAbvGr     1459 non-null int64
KitchenQual      1459 non-null int64
TotRmsAbvGrd     1459 non-null int64
Functional       1459 non-null object
Fireplaces       1459 non-null int64
FireplaceQu      1459 non-null int64
GarageType       1459 non-null int64
GarageYrBlt      1459 non-null float64
GarageFinish     1459 non-null int64
GarageCars       1459 non-null int64
GarageArea       1459 non-null int64
GarageQual       1459 non-null int64
GarageCond       1459 non-null int64
PavedDrive       1459 non-null int64
WoodDeckSF       1459 non-null int64
OpenPorchSF      1459 non-null int64
```

```
EnclosedPorch    1459 non-null int64
3SsnPorch        1459 non-null int64
ScreenPorch      1459 non-null int64
PoolArea         1459 non-null int64
PoolQC           1459 non-null int64
Fence            1459 non-null int64
MiscFeature      1459 non-null object
MiscVal          1459 non-null int64
MoSold           1459 non-null int64
YrSold           1459 non-null int64
SaleType         1459 non-null object
SaleCondition    1459 non-null object
SalePrice        1459 non-null int64
dtypes: float64(2), int64(52), object(26)
memory usage: 923.3+ KB
```

```
In [23]: n_data = pd.get_dummies(data, drop_first= True)
```

```
In [24]: n_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1459 entries, 1 to 1460
Columns: 533 entries, MSSubClass to SaleCondition_Partial
dtypes: float64(2), int64(52), uint8(479)
memory usage: 1.3 MB
```

```
In [25]: n_data.head()
```

```
Out[25]:     MSSubClass  LotFrontage  LotArea  OverallQual  OverallCond  YearBuilt  \
         Id
         1           60         65.0     8450            7            5       2003
         2           20         80.0     9600            6            8       1976
         3           60         68.0    11250            7            5       2001
         4           70         60.0     9550            7            5       1915
         5           60         84.0    14260            8            5       2000

             YearRemodAdd  ExterQual  ExterCond  BsmtQual          ...             \
         Id                                                       ...
         1           2003          3          2         4         ...
         2           1976          2          2         4         ...
         3           2002          3          2         4         ...
         4           1970          2          2         3         ...
         5           2000          3          2         4         ...

             SaleType_ConLI  SaleType_ConLw  SaleType_New  SaleType_Oth  SaleType_WD  \
         Id
         1                0               0             0             0            1
```

```
2                0                0                0                0                1
3                0                0                0                0                1
4                0                0                0                0                1
5                0                0                0                0                1

    SaleCondition_AdjLand  SaleCondition_Alloca  SaleCondition_Family  \
Id
1                       0                     0                     0
2                       0                     0                     0
3                       0                     0                     0
4                       0                     0                     0
5                       0                     0                     0

    SaleCondition_Normal  SaleCondition_Partial
Id
1                      1                      0
2                      1                      0
3                      1                      0
4                      0                      0
5                      1                      0

[5 rows x 533 columns]
```

In [28]:
```
n_data.to_csv('../clean_data.csv')
```

In [29]:
```
data.to_csv('../semi_clean_data.csv')
```