HW11

Tural Ismayilov & Polad Mahmudov & Mansur Alizada

December 8, 2017

Team named "GARABAGH" members:
Tural Ismayilov (Software Enigeenring) Mansur Alizada (Computer Science) Polad Mahmudov (Software Engineering)

We are working on the project named House Prices: Advanced Regression Techniques on the Kaggle

link:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Bitbucket link:
https://turalismayilov@bitbucket.org/garabagh/dm.git

EX1

Business understanding

Identifying your business goals

Background

Our customers, want to sell a house and they do not know the price which they can take - it can't be too low or too high. To find house price they usually try to find similar properties in their neighborhood and based on gathered data they try to assess their house price. But this is not always good. (Guessing the price by looking at similar house may be useful, but there can be certain details that have big influence on the price, but are not considered therefore may make difficult the selling proccess.) Although that method can work, but there is better way for predicting house price.

Bussiness goals:

Increase the gain of seller.

Bussiness success criteria:

Sales increasing, both the seller and buyer stay satisfied with the price of house.

Assessing your situation

Inventory of resources

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, with its description data is available.

Requirements, assumptions, and constraints

Achieving top 20% result with RMSE in Kaggle is required.

Risks and contingencies

If an Internet outage in our dormitory could pose a problem, perhaps our contingency could be to work at university until the outage has ended. If the prediction is not well enough, our contingency is to approach our instructor for help.

Terminology

Bias - Positive values of bias indicate the model tends to overestimate the price (on average) while negative values indicate the model tends to underestimate price.
Maimum Deviation - It identifies the worst prediction tey made in the validation data set.
Training set - this is a set of examples used to fit the parameters of the mode.
Validation set - The fitted model is used to predict the responses fot the observations in a second dataset called validation dataset.
Test set - is a dataset used to provude an unbiased evaluation of final model fit on the training dataset.
Root Mean Square Error - used to obtain the coefficient estimates from the original dataset.
Mean Absolute Deviation - Average error regardless of sign.

Costs and benefits

The cost is the loss in gain when the house is sold with underestimated price, and gain happens when it's sold with enough gain.

Defining your data-mining goals

Data-mining goals

Data mining goal is to predict house price for its indicators.

Data-mining success criteria

Prediction with RMSE less than 0.11979.

EX2 #Data understanding ##Gathering data ###Outline data requirements We need previous sales of house and its details to fit regression model on them for prediction.

Verify data availability

Approximately 6 percent of elements are NA, so in that case, we will use mice for handling missing data.

```
train <- read.csv('train.csv')
(sum(is.na(train)) / (nrow(train)*ncol(train))) * 100
```

```
## [1] 5.889565
```

Define selection criteria

The data will be taken from the link provided above.

```
train <- read.csv('train.csv')
str(train)
```

```
## '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 ...
```

Describing data

```
cat(readLines('data description.txt'), sep = '\n')
```

```
## Warning in readLines("data description.txt"): incomplete final line found
## on 'data description.txt'
```

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

Exploring data

```
summary(train)
```

```
##        Id           MSSubClass       MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                   Max.   :313.00
##                                                  NA's   :259
##     LotArea        Street      Alley      LotShape  LandContour
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50
##  Median :  9478               NA's:1369   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1
##  AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260
##  NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81
##                FR2    :  47   Sev:  13   OldTown:113   Artery :  48
##                FR3    :   4              Edwards:100   RRAn   :  26
##                Inside :1052              Somerst: 86   PosN   :  19
```

```
##                                          Gilbert: 79   RRAe   :  11
##                                          (Other):707   (Other):  15
##    Condition2     BldgType      HouseStyle   OverallQual
##  Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000
##  Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000
##  Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000
##  PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099
##  RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000
##  PosA   :   1                 1.5Unf : 14   Max.   :10.000
##  (Other):   2                 (Other): 19
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141
##  Median :5.000   Median :1973   Median :1994   Gambrel:  11
##  Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2
##
##     RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##  CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0
##  WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.7
##  ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8   3rd Qu.: 166.0
##  Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0
##  (Other):   2   (Other):128   (Other):136                 NA's   :8
##  ExterQual ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure
##  Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
##  Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
##  Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
##  TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
##            TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
##                      Wood  :  3
##
##  BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  ALQ :220     Min.   :   0.0   ALQ :  19    Min.   :   0.00
##  BLQ :148     1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00
##  GLQ :418     Median : 383.5   GLQ :  14    Median :   0.00
##  LwQ : 74     Mean   : 443.6   LwQ :  46    Mean   :  46.55
##  Rec :133     3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00
##  Unf :430     Max.   :5644.0   Unf :1256    Max.   :1474.00
##  NA's: 37                      NA's:  38
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseP:   3   Median :1087   Median :   0   Median :  0.000
```

```
##  Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##  SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu   GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870
##  Median : 6.000   Min1:  31   Median :1.000   Gd  :380    Basment: 19
##  Mean   : 6.518   Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9
##  Max.   :14.000   Sev :   1   Max.   :3.000   NA's:690    Detchd :387
##                   Typ :1360                               NA's   : 81
##   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
##  Min.   :1900   Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3
##  1st Qu.:1961   RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48
##  Median :1980   Unf :605     Median :2.000   Median : 480.0   Gd  :  14
##  Mean   :1979   NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3
##  3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   NA's:  81
##  NA's   :81
##  GarageCond  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  Ex  :   2   N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  Fa  :  35   P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Gd  :   9   Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
##  Po  :   7              Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  TA  :1326              3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  NA's:  81              Max.   :857.00   Max.   :547.00   Max.   :552.00
##
##    X3SsnPorch      ScreenPorch        PoolArea        PoolQC
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Ex  :   2
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2
##  Median :  0.00   Median :  0.00   Median :  0.000   Gd  :   3
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NA's:1453
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##    Fence      MiscFeature    MiscVal             MoSold
##  GdPrv:  59   Gar2:   2   Min.   :    0.00   Min.   : 1.000
```

```
##  GdWo :  54   Othr:   2   1st Qu.:    0.00   1st Qu.: 5.000
##  MnPrv: 157   Shed:  49   Median :    0.00   Median : 6.000
##  MnWw :  11   TenC:   1   Mean   :   43.49   Mean   : 6.322
##  NA's :1179   NA's:1406   3rd Qu.:    0.00   3rd Qu.: 8.000
##                           Max.   :15500.00   Max.   :12.000
##
##      YrSold        SaleType    SaleCondition    SalePrice
##  Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
##                 (Other):   9
```

Verifying data quality

There are some missing data, we have decided to solve this problem by using mice. This is the initial plan, if the mice method does not work properly we will narrow data size.

EX3

Setting up and planning your project

Create a project repository either in GitHub or Bitbucket.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Register you project by adding a new entry into the List of projects

New slide was added also to the project repository and the link is:
https://docs.google.com/presentation/d/1veA_WQcfRRx7hQnE8qklmsLYceWzQGrPHrieSEWqcaI/edit#slide=id.g2a4c3a4a7d_6_0

Make a detailed plan of your project with a list of tasks. Specify how many hours each team member is going to contribute to each task

Acquiring the data and create our environment
- Downloading the data
- Doing some plotting
Exploring the data and engineer Features
- Working with Numeric Features
- Displaying the correlation between columns
- Creating a pivot table to further investigate the relationships
- Handling null values
- Wrangling the non-numeric Features
- Transforming features
Building a linear model
Begining Linear regression modelling
Evaluating the performance and visualize results

Trying to improve the model
Making a submission
- Creating csv file with ids and predictions according to the sample format
Submitting our results

For the initial calculations each team member will spend approximately 18 hours.

Add the results from business understanding, data understanding and planning to your project repository. Report the links to where these results are listed.

It was added to project repository.

Prepare to pitch your project at the practice session