Data 605 Final
Part II
Housing Price Prediction Data

Jun Pan

5/7/2019

Overview

- Train data: housing characteristics data. 81 variables and 1460 observations.
- Test data: using for prediction house price. Result will be submitted to Kaggle.com. 80 variables (without sales price) and 1459 observations.
- Using “glimps” function to dig deeper into the datasets. We can that part of the variables are numerical, part of the data are categorical. Categorical variables need to be convert to integer (levels) for further analysis. Also, there are some missing values in the dataset.

Definition of Variables (I)

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

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

Definition of Variables (II)

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

Definition of Variables (III)

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

Definition of Variables (IV)

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	Unfinished
NA	No Basement

Definition of Variables (V)

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

Definition of Variables (VI)

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

Definition of Variables (VII)

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

Overview of Train Data by “Glimps”

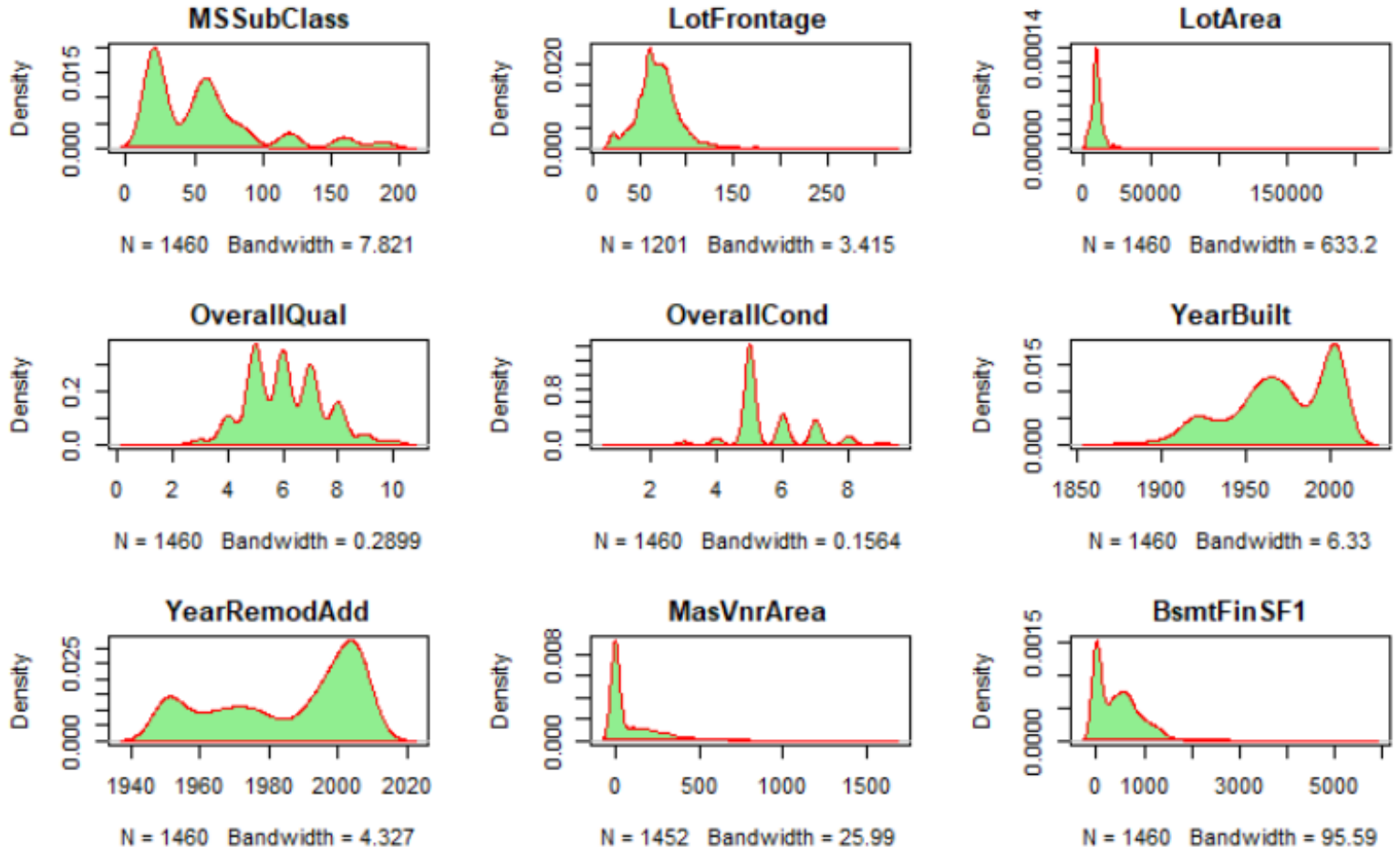
```
  {r}
glimpse(train)
```

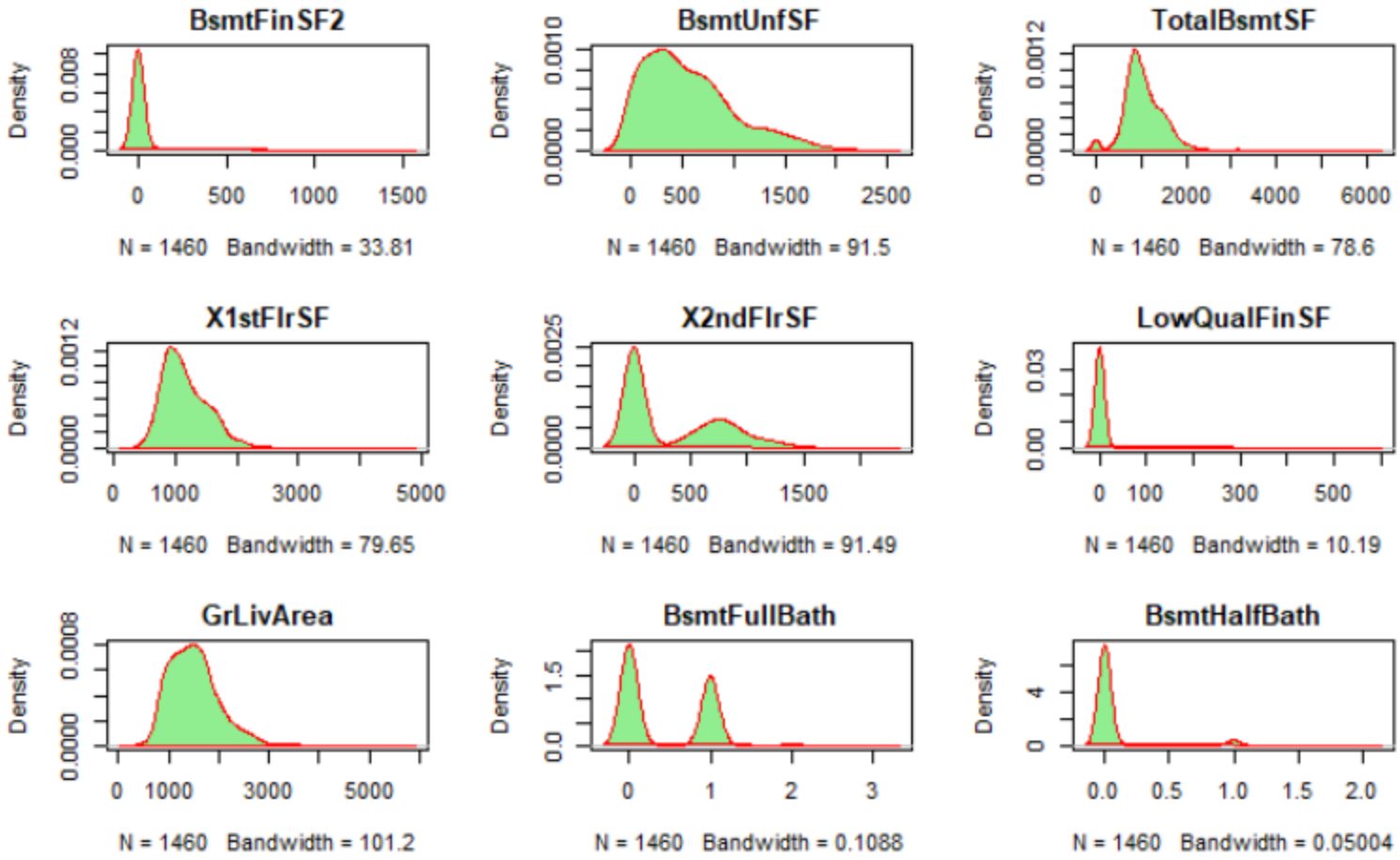
```
Observations: 1,460
Variables: 81
$ Id            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1...
$ MSSubClass    <int> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, 20, 20,...
$ MSZoning      <fct> RL, RL, RL, RL, RL, RL, RL, RL, RM, RL, RL, RL, RL, RL, RL, ...
$ LotFrontage   <int> 65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 85, NA, 91, NA, ...
$ LotArea       <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10382, 6120, 7...
$ Street        <fct> Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, ...
$ Alley         <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ LotShape      <fct> Reg, Reg, IR1, IR1, IR1, IR1, Reg, IR1, Reg, Reg, Reg, IR1, ...
$ LandContour   <fct> Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, ...
$ Utilities     <fct> AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, AllP...
$ LotConfig     <fct> Inside, FR2, Inside, Corner, FR2, Inside, Inside, Corner, In...
$ LandSlope     <fct> Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, ...
$ Neighborhood  <fct> CollgCr, Veenker, CollgCr, Crawfor, NoRidge, Mitchel, Somers...
$ Condition1    <fct> Norm, Feedr, Norm, Norm, Norm, Norm, Norm, Norm, PosN, Artery, Art...
$ Condition2    <fct> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Artery...
$ BldgType      <fct> 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 2fmCon...
$ HouseStyle    <fct> 2Story, 1Story, 2Story, 2Story, 2Story, 1.5Fin, 1Story, 2Sto...
$ OverallQual   <int> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4, 5, 5, ...
$ OverallCond   <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5, 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     <fct> Gable, Gable, Gable, Gable, Gable, Gable, Gable, Gable, Gabl...
$ RoofMatl      <fct> CompShg, CompShg, CompShg, CompShg, CompShg, CompShg, CompSh...
$ Exterior1st   <fct> VinylSd, MetalSd, VinylSd, Wd Sdng, VinylSd, VinylSd, VinylS...
$ Exterior2nd   <fct> VinylSd, MetalSd, VinylSd, Wd Shng, VinylSd, VinylSd, VinylS...
$ MasVnrType    <fct> BrkFace, None, BrkFace, None, BrkFace, None, Stone, Stone, N...
$ MasVnrArea    <int> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 306, 212,...
$ ExterQual     <fct> Gd, TA, Gd, TA, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, Gd, TA, ...
$ ExterCond     <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, ...
$ Foundation    <fct> PConc, CBlock, PConc, BrkTil, PConc, Wood, PConc, CBlock, Br...
```

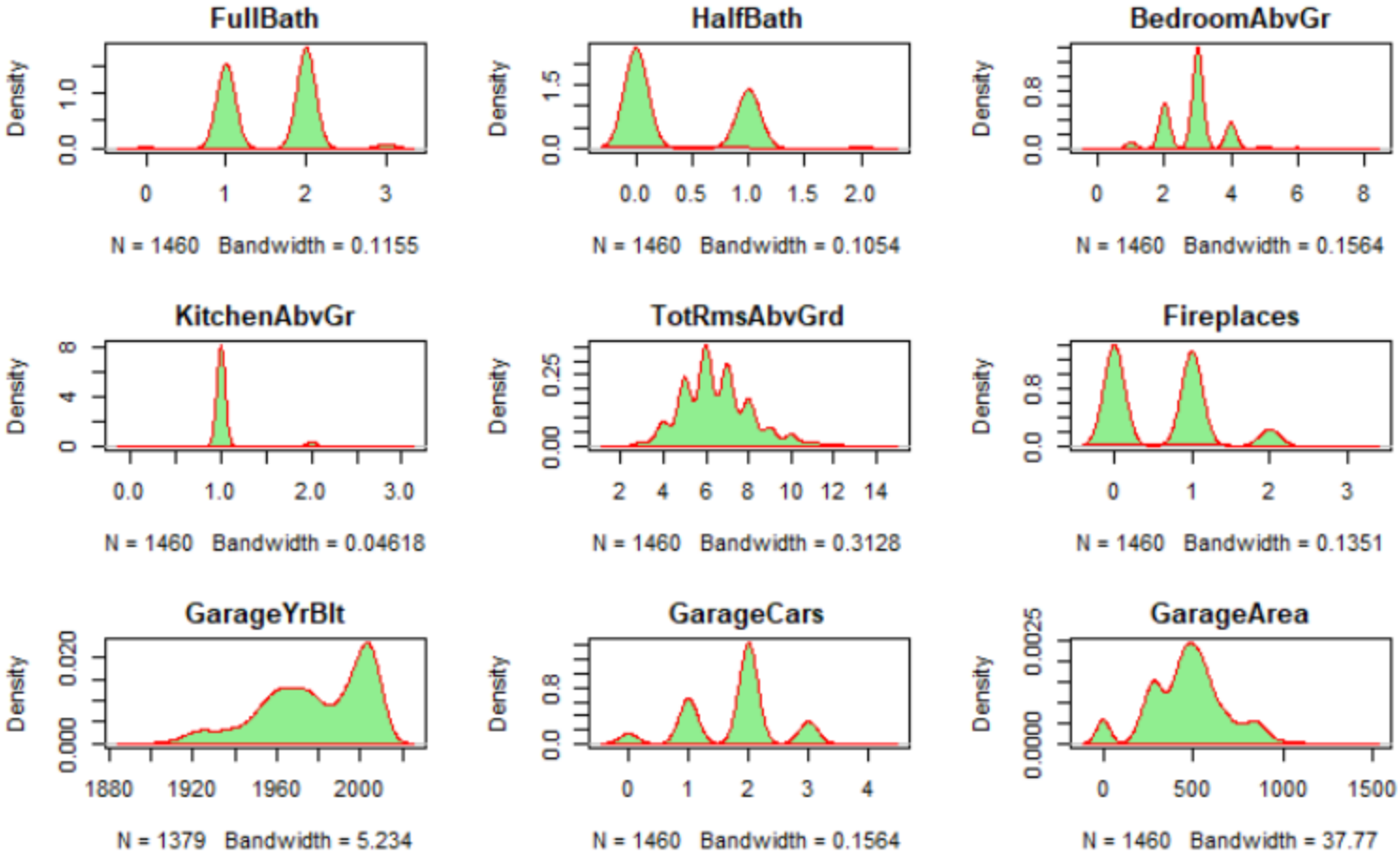
```
$ Foundation    <fct> PConc, CBlock, PConc, BrkTil, PConc, Wood, PConc, CBlock, Br...
$ BsmtQual      <fct> Gd, Gd, Gd, TA, Gd, Gd, Ex, Gd, TA, TA, TA, Ex, TA, Gd, TA, ...
$ BsmtCond      <fct> TA, TA, TA, Gd, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, ...
$ BsmtExposure  <fct> No, Gd, Mn, No, Av, No, Av, Mn, No, No, No, No, No, Av, No, ...
$ BsmtFinType1  <fct> GLQ, ALQ, GLQ, ALQ, GLQ, GLQ, GLQ, ALQ, Unf, GLQ, Rec, GLQ, ...
$ BsmtFinSF1    <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906, 998, 7...
$ BsmtFinType2  <fct> Unf, Unf, Unf, Unf, Unf, Unf, Unf, BLQ, Unf, Unf, Unf, Unf, ...
$ BsmtFinSF2    <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0,0,...
$ BsmtUnfSF     <int> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140, 134, 177, 1...
$ TotalBsmtSF   <int> 856, 1262, 920, 756, 1145, 796, 1686, 1107, 952, 991, 1040, ...
$ Heating       <fct> GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA, ...
$ HeatingQC     <fct> Ex, Ex, Ex, Gd, Ex, Ex, Ex, Gd, Ex, Ex, Ex, TA, Ex, TA, ...
$ CentralAir    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, ...
$ Electrical    <fct> SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, Fuse...
$ X1stFlrSF     <int> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 1077, 1040...
$ X2ndFlrSF     <int> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142, 0, 0, ...
$ LowQualFinSF  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, ...
$ BsmtHalfBath  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ FullBath      <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2, 1, 1, ...
$ HalfBath      <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, ...
$ BedroomAbvGr  <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2, 3, 3, ...
$ KitchenAbvGr  <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, ...
$ KitchenQual   <fct> Gd, TA, Gd, Gd, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, Gd, TA, ...
$ TotRmsAbvGrd  <int> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6, 6, 6,...
$ Functional    <fct> Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Min1, Typ, Typ, Typ,...
$ Fireplaces    <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0, 0, ...
$ FireplaceQu   <fct> NA, TA, TA, Gd, TA, NA, Gd, TA, TA, TA, NA, Gd, NA, Gd, Fa, ...
$ GarageType    <fct> Attchd, Attchd, Attchd, Detchd, Attchd, Attchd, Attchd, Attc...
$ GarageYrBlt   <int> 2003, 1976, 2001, 1998, 2000, 1993, 2004, 1973, 1931, 1939, ...
$ GarageFinish  <fct> RFn, RFn, RFn, Unf, RFn, Unf, RFn, RFn, Unf, RFn, Unf, Fin, ...
$ GarageCars    <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2, 1, ...
```

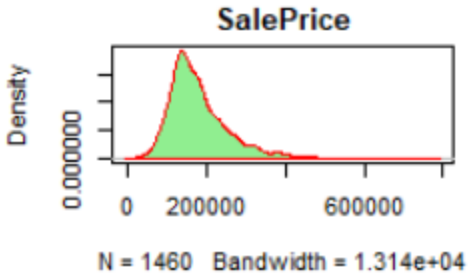
```
$ GarageCars    <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2, 1, ...
$ GarageArea    <int> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384, 736, ...
$ GarageQual    <fct> TA, TA, TA, TA, TA, TA, TA, TA, Fa, Gd, TA, TA, TA, TA, TA, ...
$ GarageCond    <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, ...
$ PavedDrive    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, ...
$ WoodDeckSF    <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, 160, 0,...
$ OpenPorchSF   <int> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, 33, 213, 112...
$ EnclosedPorch <int> 0, 0, 0, 272, 0, 0, 0, 228, 205, 0, 0, 0, 0, 0, 176, 0, 0, 0...
$ X3SsnPorch    <int> 0, 0, 0, 0, 0, 320, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ ScreenPorch   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 176, 0, 0, 0, 0, 0, 0...
$ PoolArea      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ PoolQC        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ Fence         <fct> NA, NA, NA, NA, NA, MnPrv, NA, NA, NA, NA, NA, NA, NA, NA, G...
$ MiscFeature   <fct> NA, NA, NA, NA, NA, Shed, NA, Shed, NA, NA, NA, NA, NA, NA, ...
$ MiscVal       <int> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 700, 500...
$ MoSold        <int> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3, 10, 6,...
$ YrSold        <int> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008, 2008, ...
$ SaleType      <fct> WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, New, WD, New, WD...
$ SaleCondition <fct> Normal, Normal, Normal, Abnorml, Normal, Normal, Normal, Nor...
$ SalePrice     <int> 208500, 181500, 223500, 140000, 250000, 143000, 307000, 2000...
```

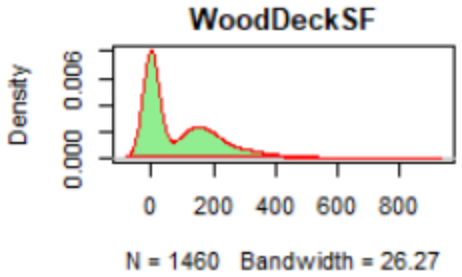
Density Plot of Numeric Variables in Train Data

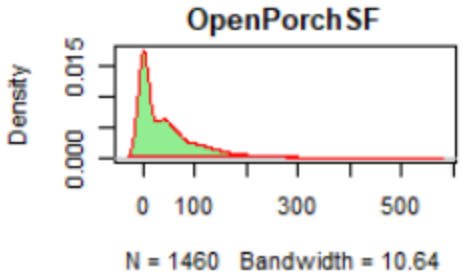


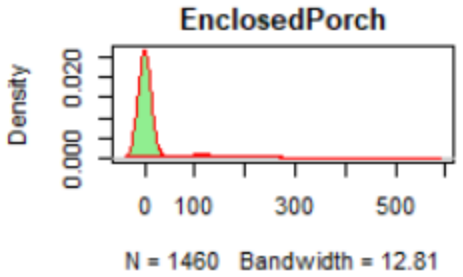

BsmtFinSF2
Density
0.000
0.008
0
500
1000
1500
N = 1460 Bandwidth = 33.81
BsmtUnfSF
Density
0.0000
0.0010
0
500
1500
2500
N = 1460 Bandwidth = 91.5
TotalBsmtSF
Density
0.0000
0.0012
0
2000
4000
6000
N = 1460 Bandwidth = 78.6
X1stFlrSF
Density
0.0000
0.0012
0
1000
3000
5000
N = 1460 Bandwidth = 79.65
X2ndFlrSF
Density
0.0000
0.0025
0
500
1500
N = 1460 Bandwidth = 91.49
LowQualFinSF
Density
0.00
0.03
0
100
300
500
N = 1460 Bandwidth = 10.19
GrLivArea
Density
0.0000
0.0008
0
1000
3000
5000
N = 1460 Bandwidth = 101.2
BsmtFullBath
Density
0.0
1.5
0
1
2
3
N = 1460 Bandwidth = 0.1088
BsmtHalfBath
Density
0
4
0.0
0.5
1.0
1.5
2.0
N = 1460 Bandwidth = 0.05004

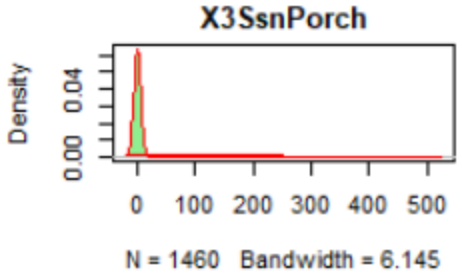

FullBath
Density
0.0
1.0
0
1
2
3
N = 1460 Bandwidth = 0.1155
HalfBath
Density
0.0
1.5
0.0
0.5
1.0
1.5
2.0
N = 1460 Bandwidth = 0.1054
BedroomAbvGr
Density
0.0
0.8
0
2
4
6
8
N = 1460 Bandwidth = 0.1564
KitchenAbvGr
Density
0
4
8
0.0
1.0
2.0
3.0
N = 1460 Bandwidth = 0.04618
TotRmsAbvGrd
Density
0.00
0.25
2
4
6
8
10
12
14
N = 1460 Bandwidth = 0.3128
Fireplaces
Density
0.0
0.8
0
1
2
3
N = 1460 Bandwidth = 0.1351
GarageYrBlt
Density
0.000
0.020
1880
1920
1960
2000
N = 1379 Bandwidth = 5.234
GarageCars
Density
0.0
0.8
0
1
2
3
4
N = 1460 Bandwidth = 0.1564
GarageArea
Density
0.0000
0.0025
0
500
1000
1500
N = 1460 Bandwidth = 37.77

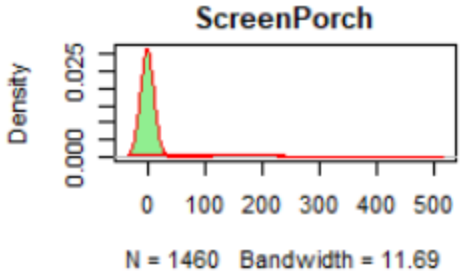

SalePrice
Density
0.000000
0
200000
600000
N = 1460 Bandwidth = 1.314e+04

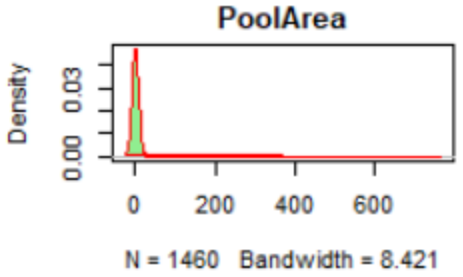

WoodDeckSF
Density
0.000
0.006
0
200
400
600
800
N = 1460 Bandwidth = 26.27

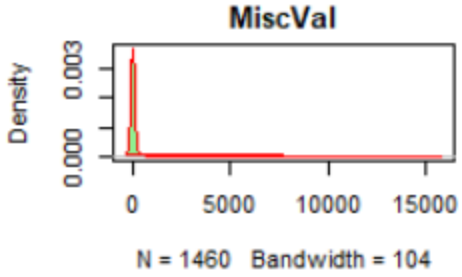

OpenPorchSF
Density
0.000
0.015
0
100
300
500
N = 1460 Bandwidth = 10.64

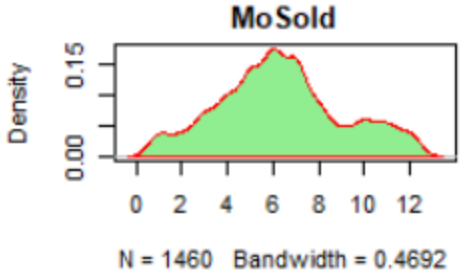

EnclosedPorch
Density
0.000
0.020
0
100
300
500
N = 1460 Bandwidth = 12.81

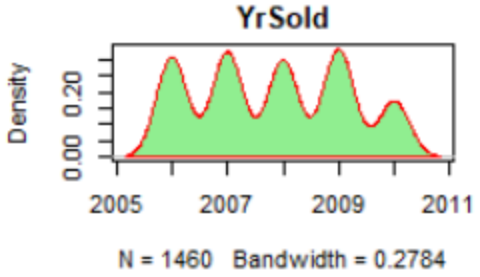

X3SsnPorch
Density
0.00
0.04
0
100
200
300
400
500
N = 1460 Bandwidth = 6.145


ScreenPorch
Density
0.000
0.025
0
100
200
300
400
500
N = 1460 Bandwidth = 11.69


PoolArea
Density
0.00
0.03
0
200
400
600
N = 1460 Bandwidth = 8.421


MiscVal
Density
0.000
0.003
0
5000
10000
15000
N = 1460 Bandwidth = 104


MoSold
Density
0.00
0.15
0
2
4
6
8
10
12
N = 1460 Bandwidth = 0.4692


YrSold
Density
0.00
0.20
2005
2007
2009
2011
N = 1460 Bandwidth = 0.2784

Univariate Descriptive Stat

Sample of Univariate Descriptive Stat.

GrLivArea	SalePrice
864	129900
2630	315000
816	110000
2353	260000
1646	248900
1524	260000
1928	219500
1372	250580
1394	167500
1487	113000
2414	160000
1694	136500

```
   GrLivArea      SalePrice
Min.   : 334   Min.   : 34900
1st Qu.:1130   1st Qu.:129975
Median :1464   Median :163000
Mean   :1515   Mean   :180921
3rd Qu.:1777   3rd Qu.:214000
Max.   :5642   Max.   :755000
```

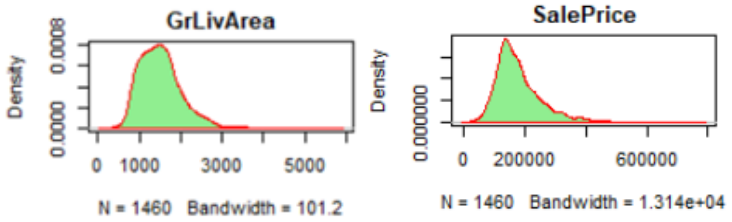
Univariate Descriptive Stat

Univariate Descriptive Statistics

	GrLivArea	SalePrice
Number of Observations	1460	1460
Non-missing values	1460	1460
Minimum	334 sq. ft.	$34900
Maximum	5642 sq. ft.	$755000
Median	1464 sq. ft.	$163000
1st quartile	1129.5 sq. ft.	$129975
3rd quartile	1776.75 sq. ft.	$214000
Average(mean)	1515.46 sq. ft.	$180921.2
Standard deviation	525.48	79442.5
Mode	864 sq. ft.	$140000
Interquartile range(IQR)	647.25 sq. ft.	$84025

GrLivArea & SalePrice

GrLivArea	SalePrice
Min. : 334	Min. : 34900
1st Qu.:1130	1st Qu.:129975
Median :1464	Median :163000
Mean :1515	Mean :180921
3rd Qu.:1777	3rd Qu.:214000
Max. :5642	Max. :755000



Correlation Matrixes (GrLivArea & SalePrice)

Correlation Matrix of Observations

	(Y>y)	(Y<=y)	Total
(X>x)	720	374	1094
(X<=x)	8	358	366
Total	728	732	1460

Correlation Matrix of Joint Probabilities

	(Y>y)	(Y<=y)	Total
(X>x)	0.4932	0.2562	0.75
(X<=x)	0.0055	0.2452	0.25
Total	0.5000	0.5000	1.00

a. $P(X>x \mid Y>y) = 99\%$
b. $P(X>x \ \& \ Y>y) = 49.32\%$
c. $P(X<x \mid Y>y) = 1.1\%$

If GrLivArea is independent to SalePrice, it should satisfy the following condition.
$P(X>x \mid Y>y)P(Y>y)=P(X>x \cap Y>y)=P(Y>y \mid X>x)P(X>x)$

However, according to the results of (a), (b) and (c), we reject the Null hypothesis and accept the Alternative Hypothesis.

So GrLiv Area and SalePrice are not independent to each other.

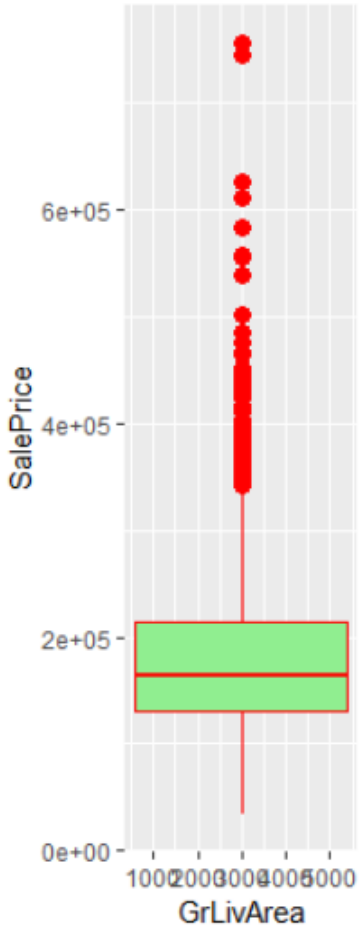
GrLivArea & SalePrice, Pearson Chi Square Test

```
        Pearson's Chi-squared test with Yates' continuity correction

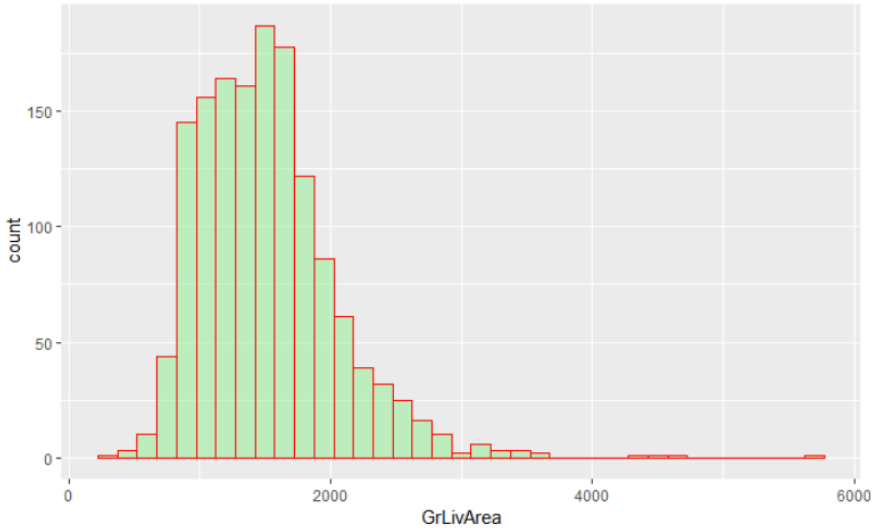
data:  house.data
X-squared = 441.58, df = 1, p-value < 2.2e-16
```

Because we have only 2 variables `GrLivArea` and `SalePrice`, degrees of freedom(df) = 1. p-value is almost "0", which is far smaller compared to 0.05 significance level. So we reject _Null Hypothesis(H0), and accept Alternative Hypothesis (HA) that`GrLivArea` has significant influence on sale price of the house.

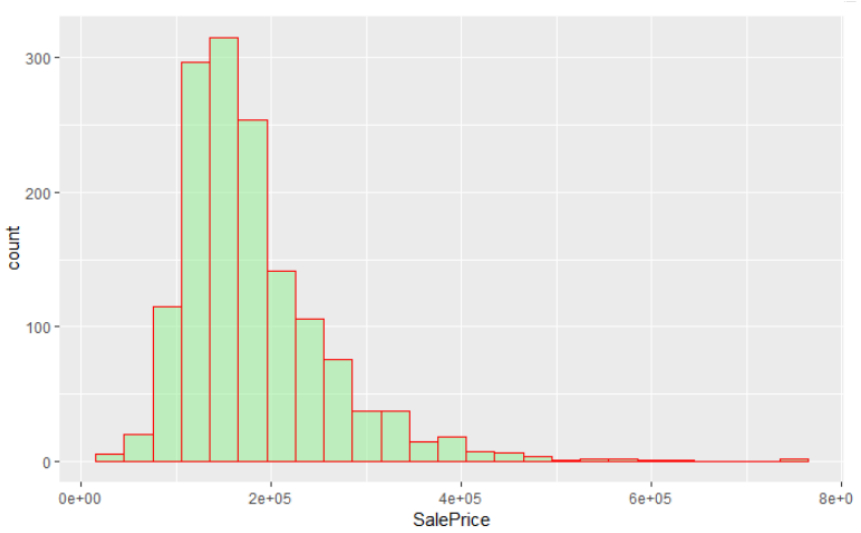
Boxplot (GrLivArea & SalePrice)



Histogram (GrLivArea & SalePrice)

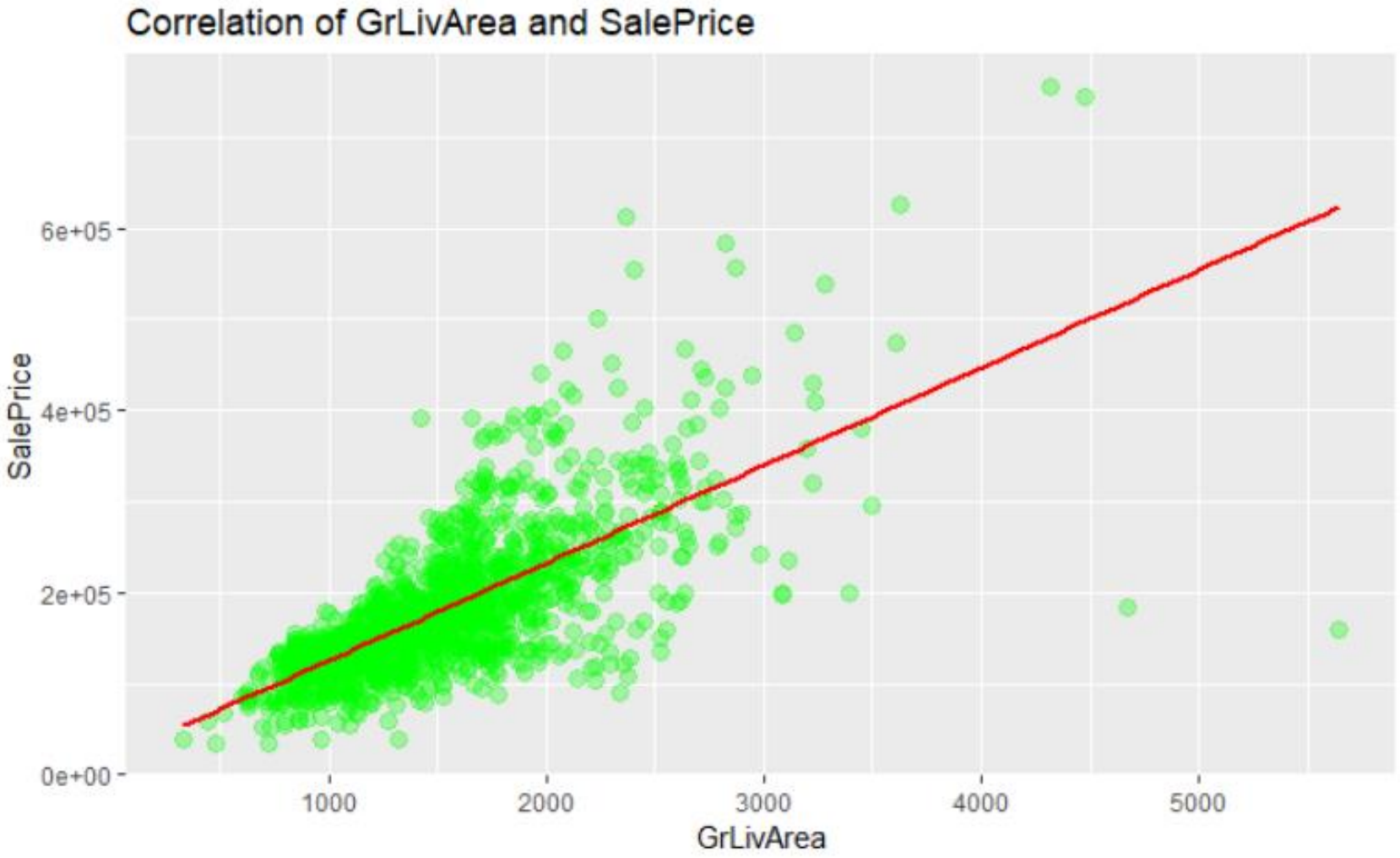


Histogram shows distribution of "GrLivArea" . Average area is 1515.46 sq.ft. with standard deviation as 525.48. It also shows right tail, suggesting existence of outliers to the right of the average.



Histogram shows distribution of sale price of houses. Average sale price is $180921.2, with sandard deviation of $79442.5. It also shows right tail, suggesting existence of outliers to the right of the average.

Scatter Plot(GrLivArea, SalePrice)



From the above boxplot, histograms and scatterplot , we can notice there are some outliers and the variation among “GrLivArea” and “SalePrice” is not constant. This causes a longer tail on the right side.

Linear Regression Model (GrLivArea vs. SalePrice)

```
```{r, echo=T, warning=F, message=F}
lm_model_price_area <- lm(train$SalePrice ~ train$GrLivArea)
summary(lm_model_price_area)

```
```

```
Call:
lm(formula = train$SalePrice ~ train$GrLivArea)

Residuals:
    Min      1Q  Median      3Q     Max
-462999  -29800   -1124   21957  339832

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     18569.026   4480.755   4.144 3.61e-05 ***
train$GrLivArea   107.130      2.794  38.348  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56070 on 1458 degrees of freedom
Multiple R-squared:  0.5021,    Adjusted R-squared:  0.5018
F-statistic:  1471 on 1 and 1458 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.5021 means that regression model can explain 50.21% of the variation in data.

Residual standard error: 56070` suggests that the average distance of the data points from the fitted line is about 56070. And 95% of times sale price should fall between 2*56070.

Box-Cox Transformation

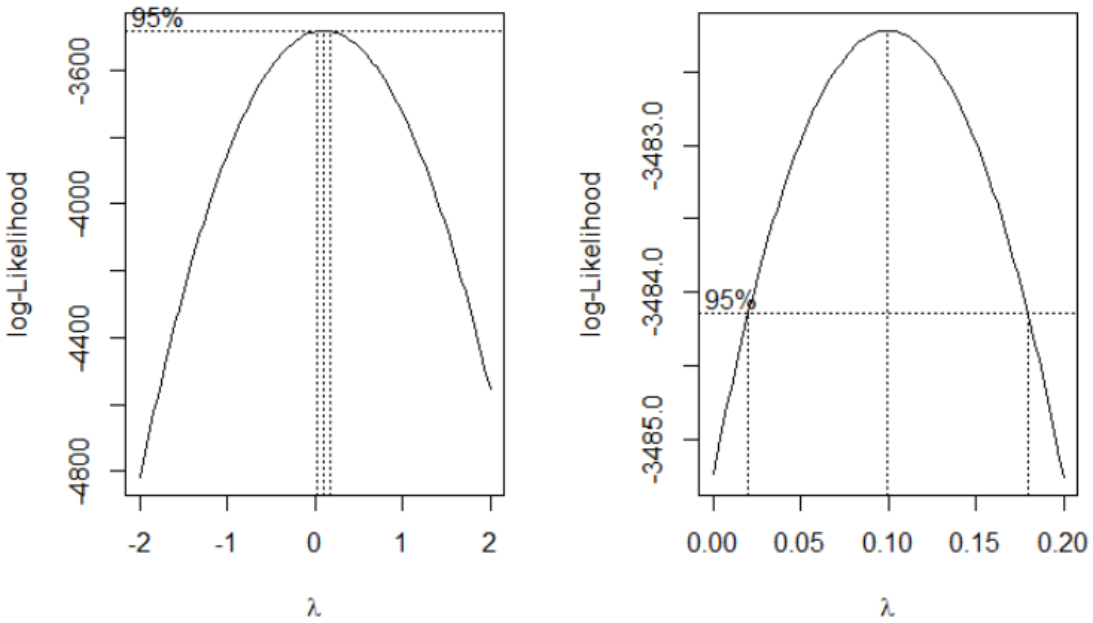
Transform non-normal dependent variables into a normal shape

y > 0

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ log\, y, & \text{if } \lambda = 0 \end{cases}$$

y < 0

$$y(\lambda) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0 \\ log\,(y+\lambda_2), & \text{if } \lambda_1 = 0 \end{cases}$$

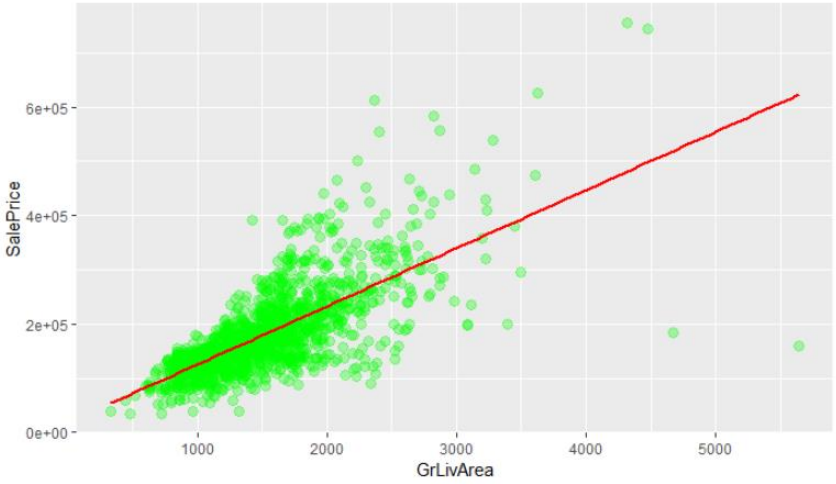


From above `boxcox` plot, optimal λ is 0.1010101 with confidence interval (0.020.02 ~ 0.180.18)

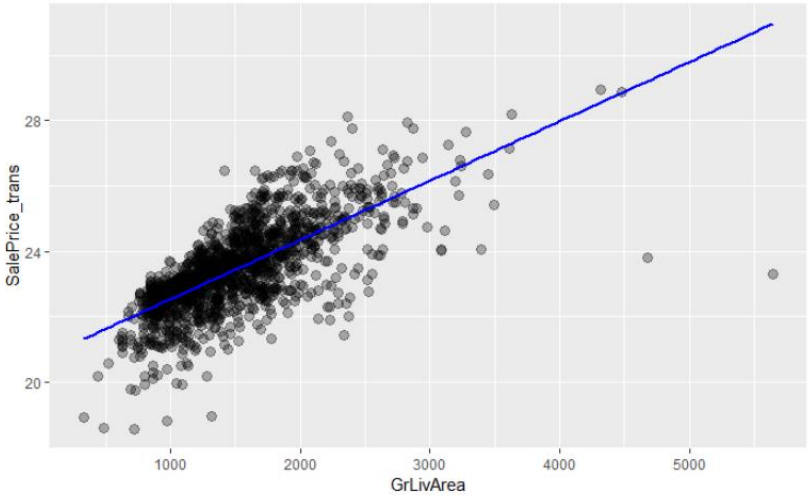
Because λ is < 0.50, there is no need to transform data.

Scatter plot using transformed “SalePrice”

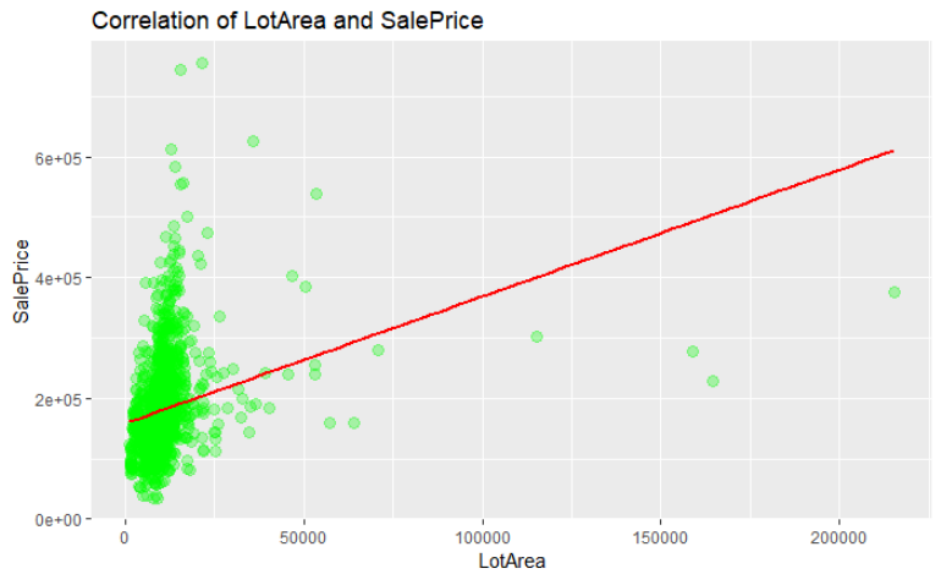
Before transformation



After transformation

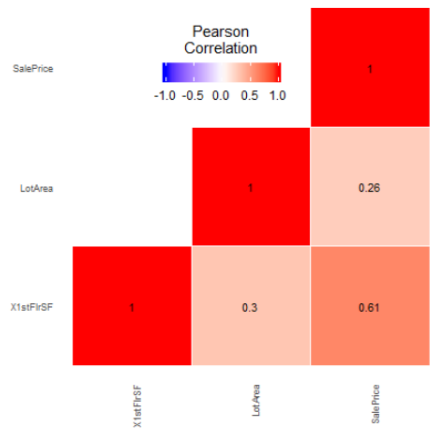


There are not too much difference using Box-Cox transformation


Correlation of LotArea and SalePrice
SalePrice
6e+05
4e+05
2e+05
0e+00
0
50000
100000
150000
200000
LotArea


Correlation of X1stFlrSF and SalePrice
SalePrice
6e+05
4e+05
2e+05
0e+00
1000
2000
3000
4000
X1stFlrSF

Pearson Correlation



For every two variables, we have generated an 80 percent of confidence interval. All the p values are < 0.001. Hence, for the three iterations of testing, we can reject the the null hypothesis and conclude that the true correlation is not 0 for the selected variables.

```
        Pearson's product-moment correlation

data:  corr_data$X1stFlrSF and corr_data$SalePrice
t = 29.078, df = 1458, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
80 percent confidence interval:
 0.5841687 0.6266715
sample estimates:
      cor
0.6058522
```

```
        Pearson's product-moment correlation

data:  corr_data$LotArea and corr_data$SalePrice
t = 10.445, df = 1458, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
80 percent confidence interval:
 0.2323391 0.2947946
sample estimates:
      cor
0.2638434
```

```
        Pearson's product-moment correlation

data:  corr_data$X1stFlrSF and corr_data$LotArea
t = 11.985, df = 1458, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
80 percent confidence interval:
 0.2686127 0.3297222
sample estimates:
      cor
0.2994746
```

Family-wise Error

Discuss the meaning of your analysis. Would you be worried about familywise error? Why or why not?

family wise error is a measurment of error when it comes to performing several iterations of estimates. This might cause results to be interpreted as being more independent then they really are. Our three tests of correlation had low p values, hence we can use that to derive the familywise error rate.

```{r}
n=3

alpha=(0.5)/n

print(paste0("Familywise error rate is ", 1-alpha))
```

```
[1] "Familywise error rate is 0.833333333333333"
```

Linear Regression Model After Transformation

```
Call:
lm(formula = train$SalePrice_trans ~ train$GrLivArea)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6488 -0.4945  0.0843  0.5314  3.1560

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.073e+01  7.656e-02  270.75   <2e-16 ***
train$GrLivArea 1.813e-03  4.774e-05   37.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9581 on 1458 degrees of freedom
Multiple R-squared:  0.4975,    Adjusted R-squared:  0.4971
F-statistic:  1443 on 1 and 1458 DF,  p-value: < 2.2e-16
```

As we see, Multiple R-squared value is smaller than the non-transformation model . The transformation is worthless in this case.

Pearson Correlation Matrix
'LotArea','TotalBsmtSF','GrLivArea', 'SalePrice'

	LotArea	TotalBsmtSF	GrLivArea	SalePrice
LotArea	1.0000000	0.2608331	0.2631162	0.2638434
TotalBsmtSF	0.2608331	1.0000000	0.4548682	0.6135806
GrLivArea	0.2631162	0.4548682	1.0000000	0.7086245
SalePrice	0.2638434	0.6135806	0.7086245	1.0000000

Correlation between `TotalBsmtSF` and `SalePrice` is 0.610.61. So bigger basement area will predict the better sale price. Square value of the coefficient is 0.3721. It means 37.21% of the variance in the sale price of a house can be explained by the total area of the basement.
Correlation between `GrLivArea` and `SalePrice` is 0.710.71. Bigger living area will predict the better sale price. Square value of the coefficient is 0.5041. It means 50.41% percent of the variance in the sale price of a house can be explained by the total living area.

Precision Matrix (Inverse Matrix)

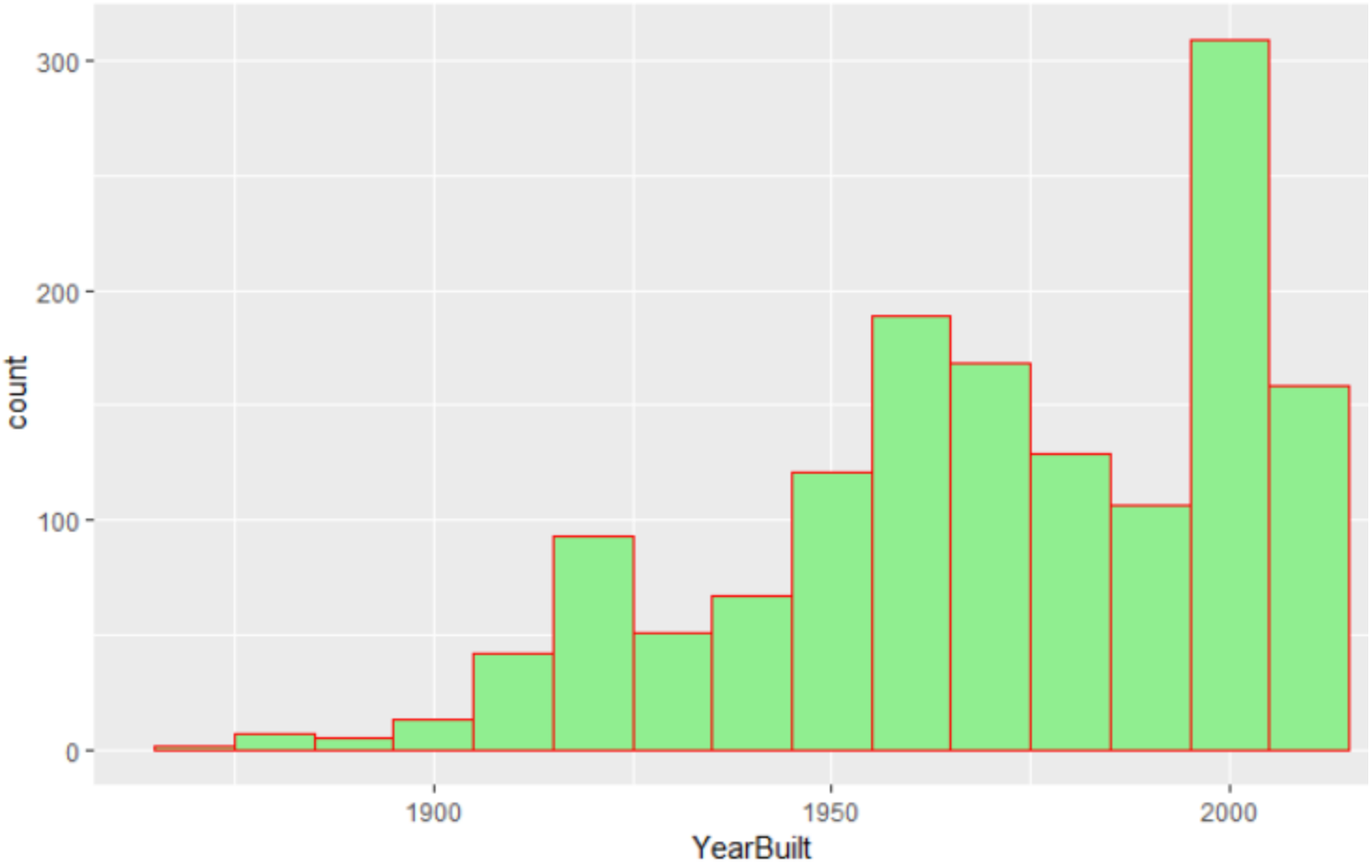
	LotArea	TotalBsmtSF	GrLivArea	SalePrice
LotArea	1.10622180	-0.1703170	-0.1623394	-0.07232846
TotalBsmtSF	-0.17031695	1.6321069	-0.0397442	-0.92832834
GrLivArea	-0.16233936	-0.0397442	2.0350650	-1.37487844
SalePrice	-0.07232846	-0.9283283	-1.3748784	2.56296011

Pearson Correlation Matrix x Precision Matrix

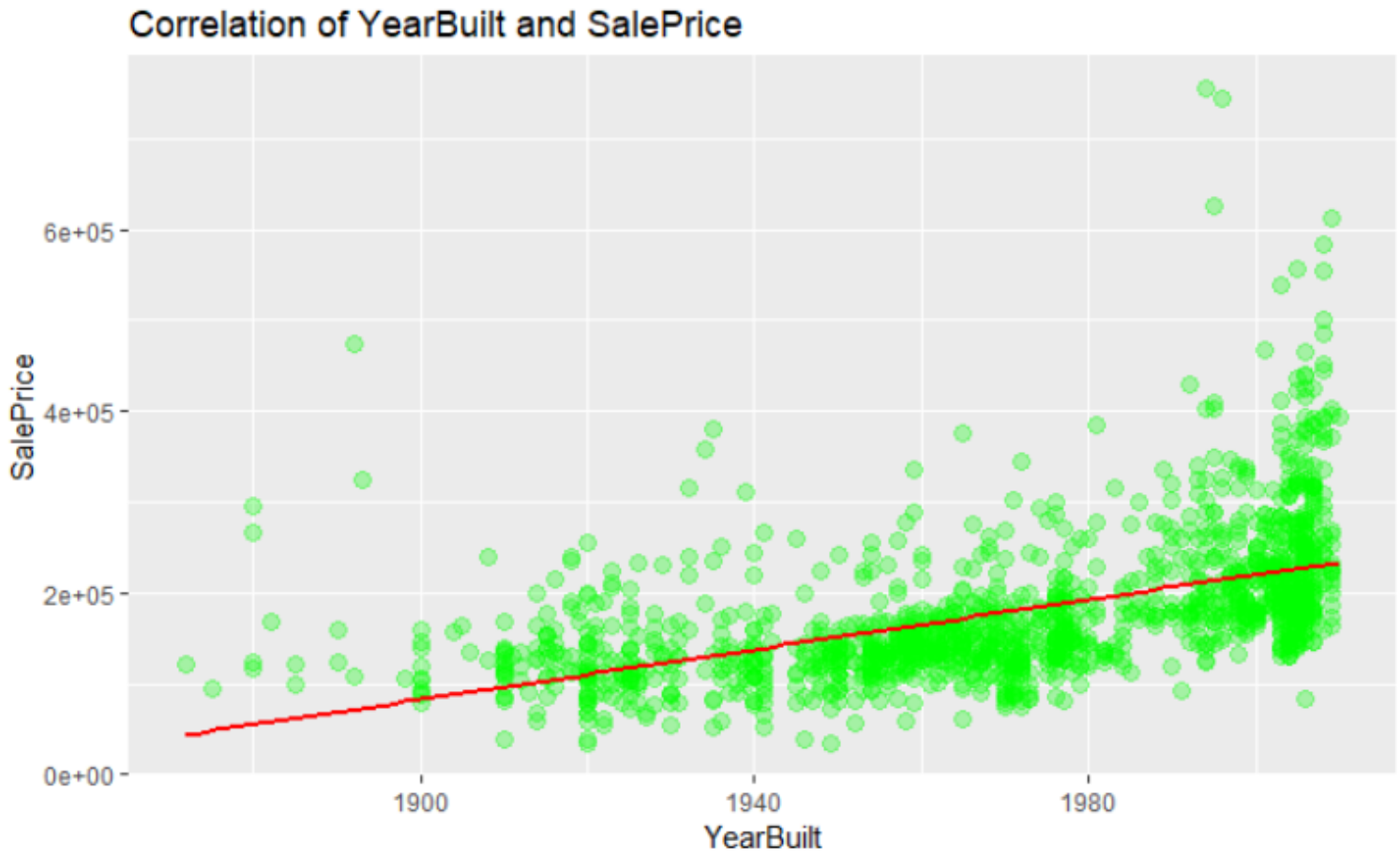
	LotArea	TotalBsmtSF	GrLivArea	SalePrice
LotArea	1	0	0	0
TotalBsmtSF	0	1	0	0
GrLivArea	0	0	1	0
SalePrice	0	0	0	1

Correlation Matrix multiplied by Precision Matrix and Precision Matrix multiplied by Correlation Matrix results in `identity matrix`.

Calculus-Based Probability & Statistics



Firstly, find a right skewed variable for study.


Correlation of YearBuilt and SalePrice
SalePrice
6e+05
4e+05
2e+05
0e+00
1900
1940
1980
YearBuilt

Results are same using ‘fitdistr’ and ‘optim’ functions

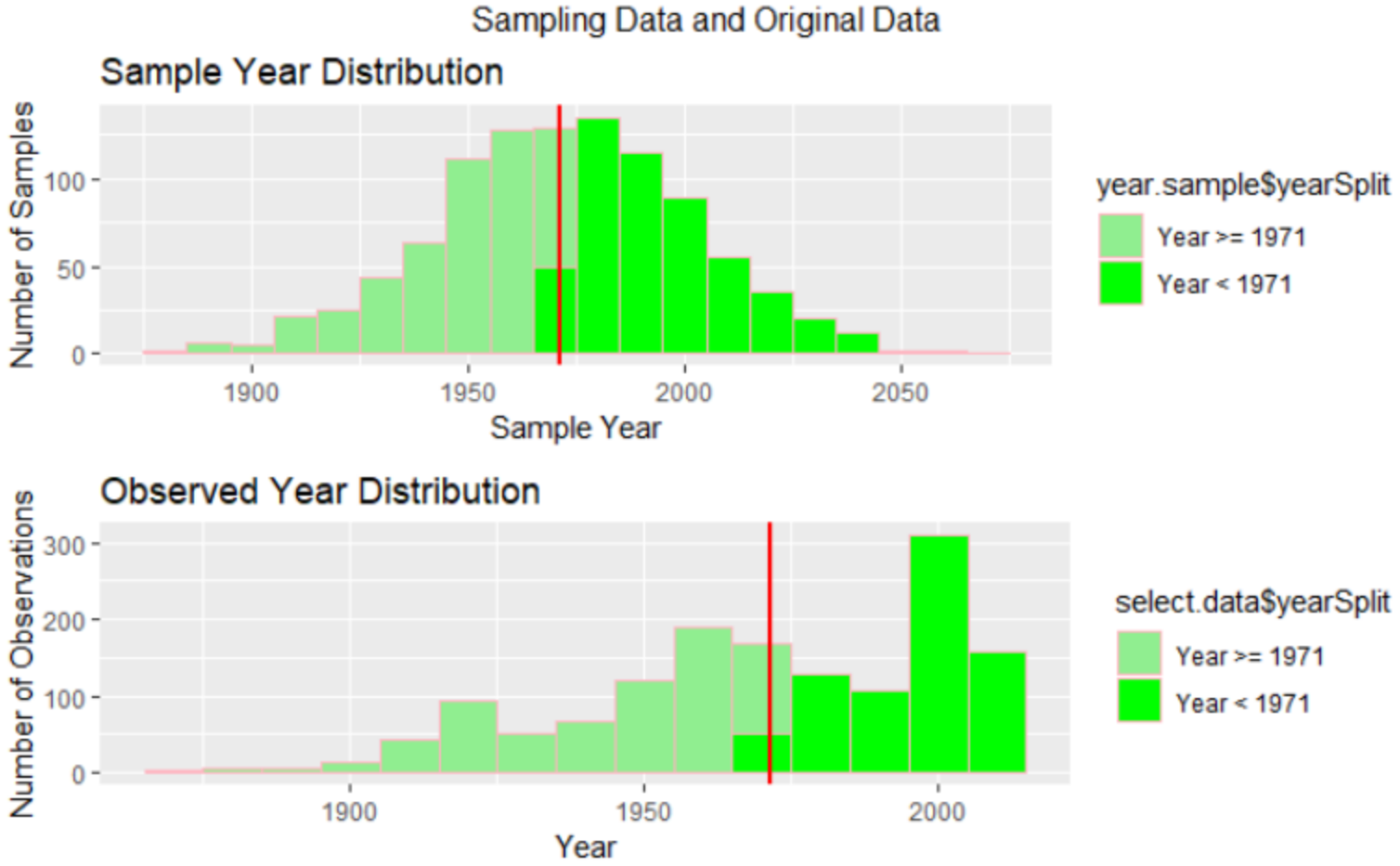
'fitdistr‘ function

```
      mean             sd
 1971.2678082     30.1925588
(   0.7901754) (   0.5587384)
```

‘Optim’function

```
     mean
1971.2765   30.1827

$value
[1] 7046.74

$counts
function gradient
      57       NA

$convergence
[1] 0

$message
NULL
```

Histogram Comparing of Original and Sampling Data



To generate 1000 samples, 'rnorm' function with the optimal parameters generated by 'optim' function will be used.
Mean and SD of samples and observed data is same, 1971, 30 respectively. Red line represents average of the data.

Good of Fit Tests

- Chi-Square test was used to see if the sample generated represents a normal distribution. We guess that there should be 50% cases where year is greater than or equal to `average` and 50% cases less than `average`.

- Hypothesis,
- H_0 : Sample data follow a specified distribution.
- H_A : Sample data do not follow the specified distribution.

```
        Chi-squared test for given probabilities

data:  sample.rows
X-squared = 1.024, df = 1, p-value = 0.3116
```

- Because p-value is 0.31 which is greater than 0.05, we accept null hypothesis(H0). In conclusion, that sample data represents normal distribution.

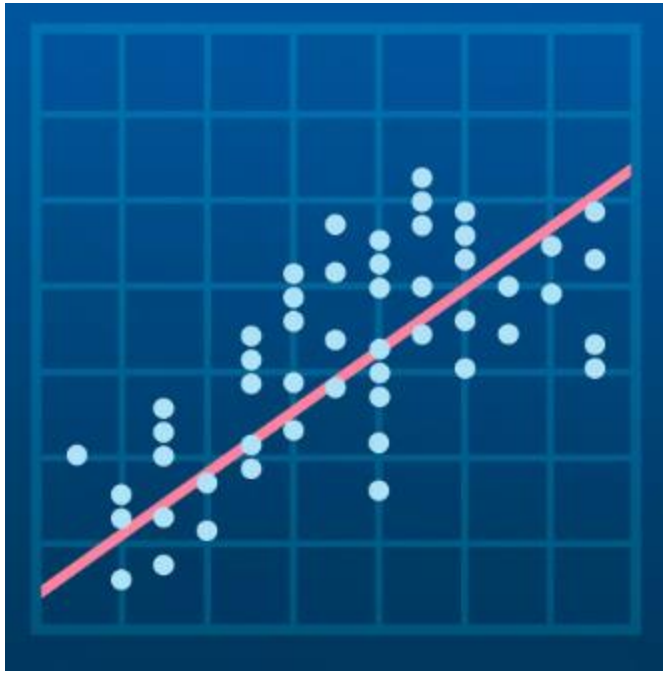
Second Chi Square Test

- Following is a test to see whether sample represents actual observed data.
- Hypothesis:
- H_0: Sample data represents actual observed data.
- H_A: Sample data do not represent actual observed data.

```
        Chi-squared test for given probabilities

data:  sample.rows
X-squared = 0.064103, df = 1, p-value = 0.8001
```

- Because p-value is 0.8 which is greater than 0.05, we accept null hypothesis(H0). In conclusion, that sample data represents actual observed data.

Build Model

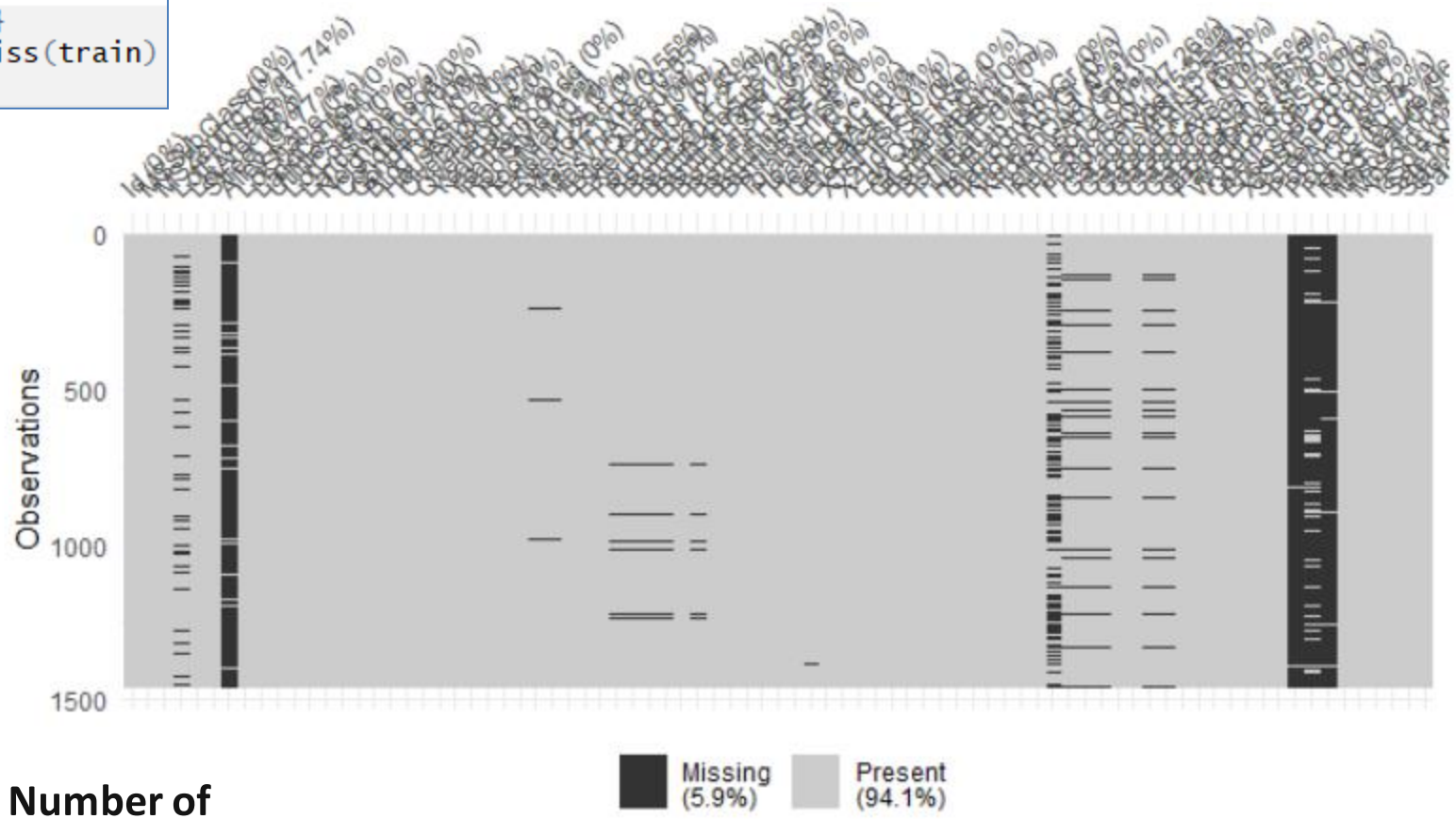


Plan to Deal with Data

- 1. For some variables with more than 50% of missing information such as “Alley”, “PoolQC”, “Fence”, “Miss Feature”, I will drop it.
- 2. For numerical variables, I will try to keep as many as possible. If there is missing information, I can impute it with mean.
- 3. For categorical variables, it is hard to analysis using “as is” condition. Too drop all categorical data is not wise, because it has lots of information. For this kind of situation, I like to keep some categorical variables for analysis by transforming from a factor in character into an ordinal variables coded with a serials of numbers. For some categorical variables which can not each to give a ordinal code, I am going to drop it.

Check Missing Data (train data)

```{r}
vis_miss(train)
```



Number of missing values

```{r}
n_miss(train)
```

```
[1] 6965
```

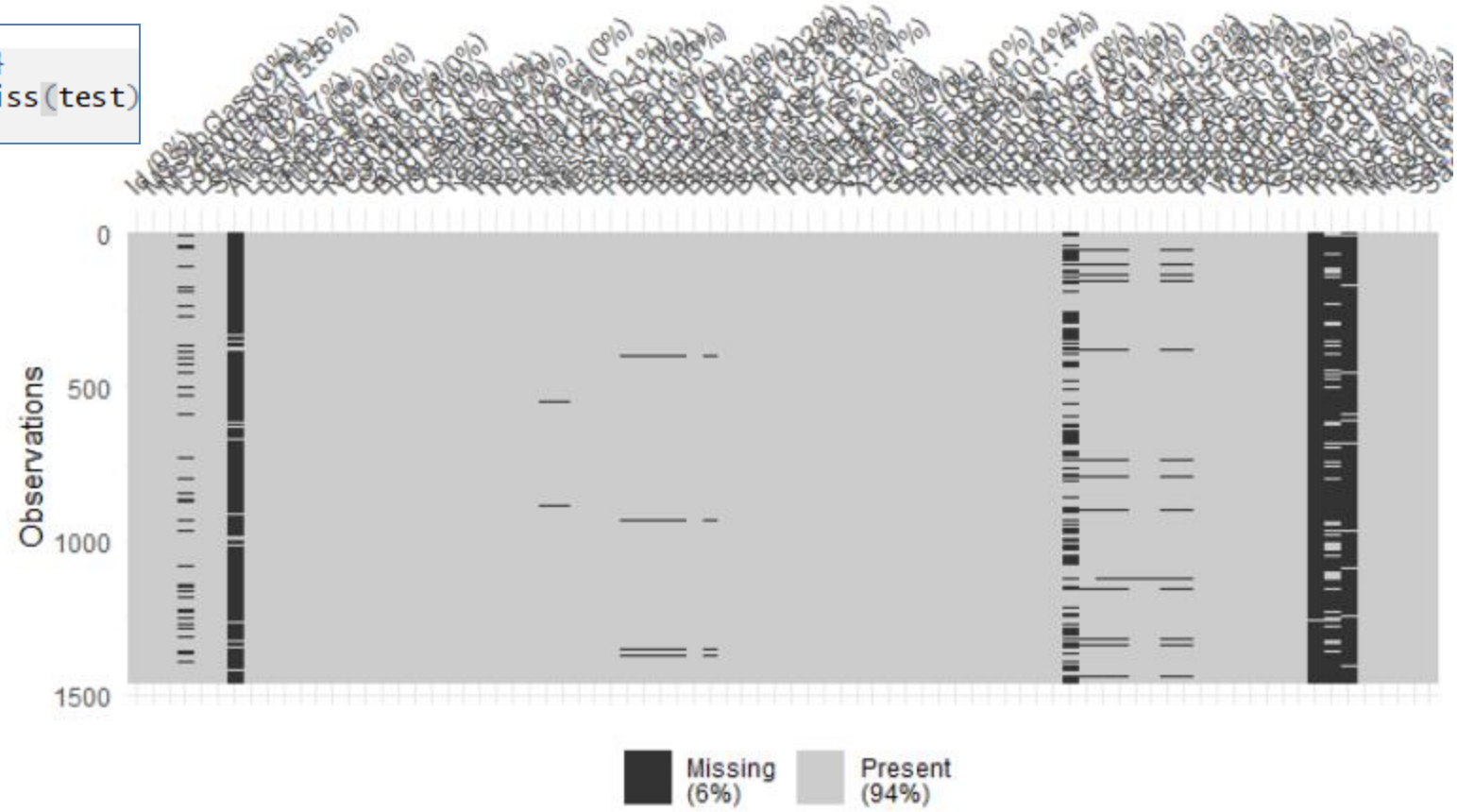
% of Missing

```{r}
prop_miss(train)
```

```
[1] 0.05889565
```

Check Missing Data (test data)

```r
vis_miss(test)
```



Number of missing values

```r
n_miss(test)
```

```
[1] 7000
```

% of Missing

```r
prop_miss(test)
```

```
[1] 0.05997258
```

Detail Missing Information in Train

LotFrontage	259
Alley	1369
MasVnrType	8
MasVnrArea	8
BsmtQual	37
BsmtCond	37
BsmtExposure	38
BsmtFinType1	37
BsmtFinType2	38
Electrical	1
FireplaceQu	690
GarageType	81
GarageYrBlt	81
GarageFinish	81
PoolQC	1453
Fence	1179
MiscFeature	1406

Detail Missing Information in Test (I)

MSZoning	4
LotFrontage	227
Alley	1352
Utilities	2
Exterior1st	1
Exterior2nd	1
MasVnrType	16
MasVnrArea	15
BsmtQual	44
BsmtCond	45
BsmtExposure	44
BsmtFinType1	42
BsmtFinSF1	1
BsmtFinType2	42
BsmtFinSF2	1
BsmtUnfSF	1
TotalBsmtSF	1

Detail Missing Information in Test (II)

BsmtFullBath	2
BsmtHalfBath	2
KitchenQual	1
Functional	2
FireplaceQu	730
GarageType	76
GarageYrBlt	78
GarageFinish	78
GarageCars	1
GarageArea	1
GarageQual	78
GarageCond	78
PoolQC	1456
Fence	1169
MiscFeature	1408
SaleType	1

“dplyr::select” to build subset data

```{r}
train1 <- dplyr::select(train,MSSubClass,Neighborhood,LotFrontage,LotArea,BldgType,OverallQual,OverallCond,YearBuilt,YearRemodAdd,MasVnrArea,BsmtFinSF1,BsmtFinSF2,BsmtUnfSF,TotalBsmtSF,CentralAir,X1stFlrSF,X2ndFlrSF,LowQualFinSF,GrLivArea,TotRmsAbvGrd,GarageCars,GarageArea,WoodDeckSF,OpenPorchSF,EnclosedPorch,ScreenPorch,X3SsnPorch,PoolArea,MiscVal,MoSold,YrSold,SaleType,SaleCondition,SalePrice)
```

Transform charater vector to numeric

```{r}

train1$CentralAir <- ifelse(train1$CentralAir=="Yes", 1, 0)
train1$SaleType <- ifelse(train1$SaleType==c("WD","New","VMD"), 1, 0)
train1$SaleCondition <- ifelse(train1$SaleCondition=="Normal", 1, 0)

```

```{r}
train1$BldgType <- as.character(train1$BldgType)
train1$BldgType[which(train1$BldgType == "1Fam")] <- "5"
train1$BldgType[which(train1$BldgType == "2fmCon")] <- "4"
train1$BldgType[which(train1$BldgType == "Duplex")] <- "3"
train1$BldgType[which(train1$BldgType == "Twnhs")] <- "2"
train1$BldgType[which(train1$BldgType == "TwnhsE")] <- "1"
train1$BldgType <- as.numeric(train1$BldgType)
```

Impute missing data with mean

```{r}
train1$LotFrontage[is.na(train1$LotFrontage)] <- mean(train1$LotFrontage, na.rm=TRUE)
train1$MasVnrArea[is.na(train1$MasVnrArea)] <- mean(train1$MasVnrArea, na.rm=TRUE)
```

Neighborhood Variable Transformation

- Neighborhood is an important factor for per square foot price. Because I am not familiar with the neighborhood in dataset, I will get the median sale price of each neighborhood first.

```
```{r}
df <- train %>%

  group_by(Neighborhood) %>%
  summarize(medianSalePrice = median(SalePrice)) %>% arrange(desc(medianSalePrice))
df
```

Neighborhood <fctr>	medianSalePrice <dbl>
NridgHt	315000
NoRidge	301500
StoneBr	278000
Timber	228475
Somerst	225500
Veenker	218000
Crawfor	200624
ClearCr	200250
CollgCr	197200
Blmngtn	191000

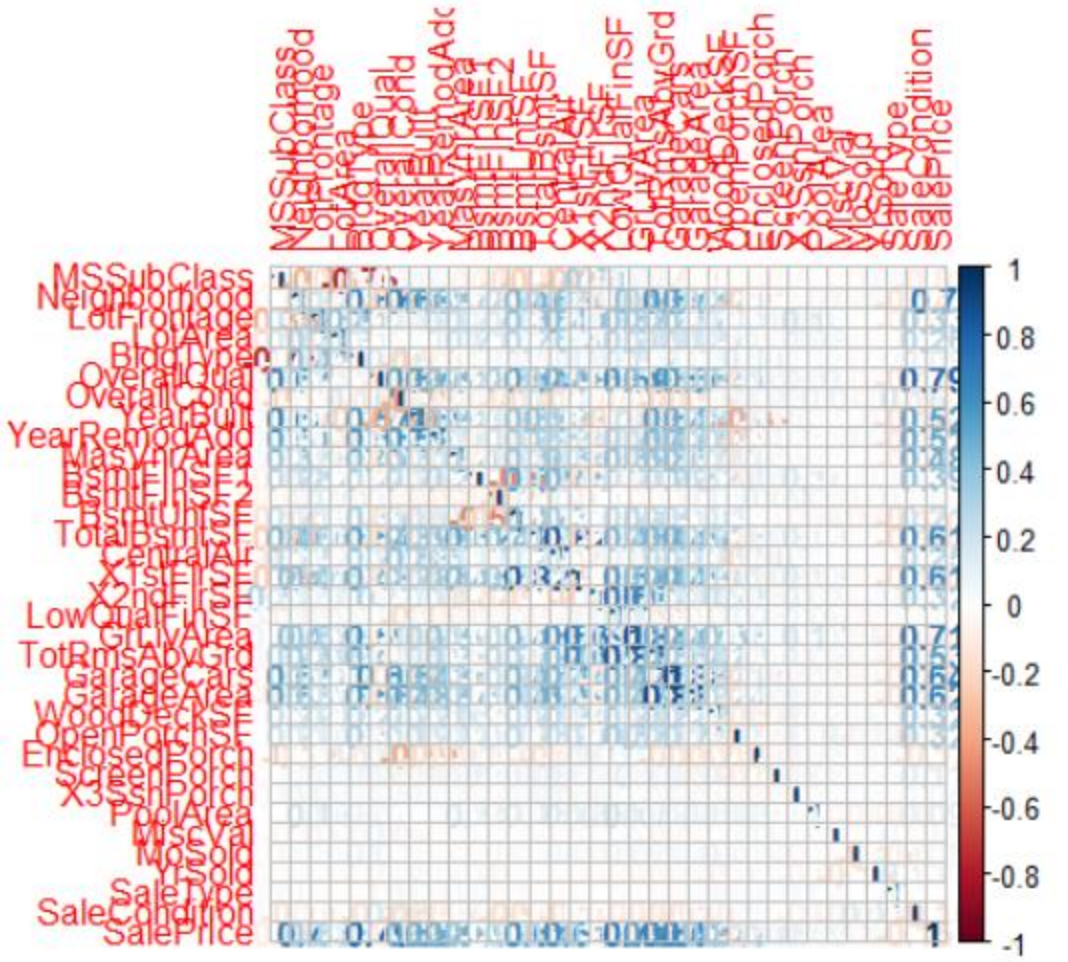
Neighborhood <fctr>	medianSalePrice <dbl>
NWAmes	182900
Gilbert	181000
SawyerW	179900
Mitchel	153500
NPkVill	146000
NAmes	140000
SWISU	139500
Blueste	137500
Sawyer	135000
BrkSide	124300

Neighborhood <fctr>	medianSalePrice <dbl>
Edwards	121750
OldTown	119000
BrDale	106000
IDOTRR	103000
MeadowV	88000

For each neighborhood, I will impute as a score from 25 to 1 according the medianSalePrice from Highest (NridgHt) to Lowest (MeadowV)
```

No Missing Data after Imputation

Correlation Plot



Summary (I)

```
  MSSubClass     Neighborhood    LotFrontage        LotArea          BldgType
Min.   : 20.0   Min.   : 1.00   Min.   : 21.00   Min.   :  1300   Min.   :1.000
1st Qu.: 20.0   1st Qu.: 7.00   1st Qu.: 60.00   1st Qu.:  7554   1st Qu.:5.000
Median : 50.0   Median :13.00   Median : 70.05   Median :  9478   Median :5.000
Mean   : 56.9   Mean   :12.84   Mean   : 70.05   Mean   : 10517   Mean   :4.507
3rd Qu.: 70.0   3rd Qu.:17.00   3rd Qu.: 79.00   3rd Qu.: 11602   3rd Qu.:5.000
Max.   :190.0   Max.   :25.00   Max.   :313.00   Max.   :215245   Max.   :5.000
 OverallQual      OverallCond      YearBuilt     YearRemodAdd    MasVnrArea
Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950   Min.   :   0.0
1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   1st Qu.:   0.0
Median : 6.000   Median :5.000   Median :1973   Median :1994   Median :   0.0
Mean   : 6.099   Mean   :5.575   Mean   :1971   Mean   :1985   Mean   : 103.7
3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   3rd Qu.: 164.2
Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010   Max.   :1600.0
  BsmtFinSF1      BsmtFinSF2       BsmtUnfSF       TotalBsmtSF      CentralAir
Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Min.   :0.0000
1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   1st Qu.:1.0000
Median : 383.5   Median :   0.00   Median : 477.5   Median : 991.5   Median :1.0000
Mean   : 443.6   Mean   :  46.55   Mean   : 567.2   Mean   :1057.4   Mean   :0.9349
3rd Qu.: 712.2   3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2   3rd Qu.:1.0000
Max.   :5644.0   Max.   :1474.00   Max.   :2336.0   Max.   :6110.0   Max.   :1.0000
  X1stFlrSF      X2ndFlrSF     LowQualFinSF       GrLivArea     TotRmsAbvGrd
Min.   : 334   Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   : 2.000
1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.: 5.000
Median :1087   Median :   0   Median :  0.000   Median :1464   Median : 6.000
Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   : 6.518
3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.: 7.000
Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :14.000
```

Summary (II)

```
   GarageCars      GarageArea      WoodDeckSF      OpenPorchSF     EnclosedPorch
Min.   :0.000   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
1st Qu.:1.000   1st Qu.: 334.5   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
Median :2.000   Median : 480.0   Median :  0.00   Median : 25.00   Median :  0.00
Mean   :1.767   Mean   : 473.0   Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
3rd Qu.:2.000   3rd Qu.: 576.0   3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
Max.   :4.000   Max.   :1418.0   Max.   :857.00   Max.   :547.00   Max.   :552.00
  ScreenPorch      X3SsnPorch        PoolArea          MiscVal            MoSold
Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :    0.00   Min.   : 1.000
1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:    0.00   1st Qu.: 5.000
Median :  0.00   Median :  0.00   Median :  0.000   Median :    0.00   Median : 6.000
Mean   : 15.06   Mean   :  3.41   Mean   :  2.759   Mean   :   43.49   Mean   : 6.322
3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:    0.00   3rd Qu.: 8.000
Max.   :480.00   Max.   :508.00   Max.   :738.000   Max.   :15500.00   Max.   :12.000
    YrSold        SaleType      SaleCondition      SalePrice
Min.   :2006   Min.   :0.0000   Min.   :0.0000   Min.   : 34900
1st Qu.:2007   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:129975
Median :2008   Median :0.0000   Median :1.0000   Median :163000
Mean   :2008   Mean   :0.3151   Mean   :0.8205   Mean   :180921
3rd Qu.:2009   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:214000
Max.   :2010   Max.   :1.0000   Max.   :1.0000   Max.   :755000
```

Fix Colinearity and Replace Outliers

#Colinearity
From the correlation matrix and plot, we find that 'TotalBsmtSF' is highly associated with 'GrLivArea (0.825)' and 'BsmtFinSF1'. So we will DROP 'TotalBsmtSF'.

'GrLivArea (0.825)' is also highly with 'TotRmsAbvGrd'(0.825), 'X1stFlrSF'(0.566) and X2ndFlrSF(0.688). So we will also DROP the above three variables.

```{r}
train1$TotalBsmtSF <- NULL
train1$TotRmsAbvGrd<- NULL
train1$X1stFlrSF<- NULL
train1$X2ndFlrSF<- NULL
```

#Outlinear
From the density plots and summary. We feel that the following varibles ("LotFrontage","LotArea","MasVnrArea","BsmtFinSF1","BsmtFinSF2","BsmtUnfSF","GrLivArea","Sale Price","WoodDeckSF","OpenPorchSF","EnclosedPorch","ScreenPorch","X3SsnPorch","PoolArea","Misc Val") may have outliers. We will replace the outliers.

Full Model: All Variables

```r
full.model <- lm(SalePrice~MSSubClass+Neighborhood+LotFrontage+LotArea+BldgType+OverallQual+OverallCond+YearBuilt+YearRemodAdd+MasVnrArea+BsmtFinSF1+BsmtFinSF2+BsmtUnfSF+CentralAir+LowQualFinSF+GrLivArea+GarageCars+GarageArea+WoodDeckSF+OpenPorchSF+EnclosedPorch+ScreenPorch+X3SsnPorch+PoolArea+MiscVal+MoSold+YrSold+SaleCondition, data=train1)
summary(full.model)
```

```
Call:
lm(formula = SalePrice ~ MSSubClass + Neighborhood + LotFrontage +
    LotArea + BldgType + OverallQual + OverallCond + YearBuilt +
    YearRemodAdd + MasVnrArea + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
    CentralAir + LowQualFinSF + GrLivArea + GarageCars + GarageArea +
    WoodDeckSF + OpenPorchSF + EnclosedPorch + ScreenPorch +
    X3SsnPorch + PoolArea + MiscVal + MoSold + YrSold + SaleCondition,
    data = train1)

Residuals:
    Min      1Q  Median      3Q     Max
-166234  -14552   -1481   13165   90418

Coefficients: (3 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.039e+06  9.529e+05  -1.091 0.275636
MSSubClass    -1.263e+02  2.528e+01  -4.998 6.52e-07 ***
Neighborhood   1.257e+03  1.482e+02   8.485  < 2e-16 ***
LotFrontage    1.273e+02  7.180e+01   1.773 0.076414 .
LotArea        8.313e-01  3.262e-01   2.549 0.010909 *
BldgType      -1.281e+03  9.046e+02  -1.416 0.157132
OverallQual    1.036e+04  7.542e+02  13.739  < 2e-16 ***
OverallCond    5.887e+02  7.076e+02   0.832 0.405551
YearBuilt     -2.058e+00  4.159e+01  -0.049 0.960545
YearRemodAdd   1.617e+02  4.327e+01   3.737 0.000193 ***
MasVnrArea     1.332e+01  5.158e+00   2.583 0.009907 **
BsmtFinSF1     3.308e+01  3.103e+00  10.659  < 2e-16 ***
BsmtFinSF2     2.136e+00  6.601e+00   0.324 0.746343
BsmtUnfSF      8.206e+00  2.812e+00   2.919 0.003571 **
CentralAir    -1.282e+04  2.890e+03  -4.436 9.86e-06 ***
LowQualFinSF  -2.113e+01  1.317e+01  -1.604 0.108987
GrLivArea      5.761e+01  2.873e+00  20.048  < 2e-16 ***
GarageCars    -5.591e+02  1.878e+03  -0.298 0.765994
GarageArea     1.199e+01  6.365e+00   1.883 0.059900 .
OpenPorchSF    3.094e+01  1.486e+01   2.083 0.037423 *
EnclosedPorch  9.079e+00  1.374e+01   0.661 0.508722
ScreenPorch    1.300e+01  1.581e+01   0.822 0.411132
X3SsnPorch            NA         NA      NA       NA
PoolArea              NA         NA      NA       NA
MiscVal               NA         NA      NA       NA
MoSold         1.384e+02  2.315e+02   0.598 0.549904
YrSold         3.556e+02  4.747e+02   0.749 0.453875
SaleCondition -5.027e+03  1.673e+03  -3.005 0.002699 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23420 on 1434 degrees of freedom
Multiple R-squared:  0.7773,    Adjusted R-squared:  0.7734
F-statistic: 200.2 on 25 and 1434 DF,  p-value: < 2.2e-16
```

Reduced Model: With All Significant Variables in Full.model

```{r}
reduced.model <- lm(SalePrice~MSSubClass+Neighborhood+LotArea+OverallQual+YearBuilt+YearRemodAdd+MasVnrArea+BsmtFinSF1+BsmtUnfSF+CentralAir+GrLivArea+WoodDeckSF+OpenPorchSF+SaleCondition, data=train1)
summary(reduced.model)
```

```
Call:
lm(formula = SalePrice ~ MSSubClass + Neighborhood + LotArea +
    OverallQual + YearBuilt + YearRemodAdd + MasVnrArea + BsmtFinSF1 +
    BsmtUnfSF + CentralAir + GrLivArea + WoodDeckSF + OpenPorchSF +
    SaleCondition, data = train1)

Residuals:
    Min      1Q  Median      3Q     Max
-158644  -14844   -1347   13089   89419

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.476e+05  8.461e+04  -4.108 4.21e-05 ***
MSSubClass    -1.103e+02  1.533e+01  -7.193 1.02e-12 ***
Neighborhood   1.295e+03  1.477e+02   8.771  < 2e-16 ***
LotArea        1.017e+00  3.114e-01   3.266 0.001116 **
OverallQual    1.083e+04  7.302e+02  14.830  < 2e-16 ***
YearBuilt      1.730e+00  3.313e+01   0.052 0.958350
YearRemodAdd   1.713e+02  3.992e+01   4.291 1.90e-05 ***
MasVnrArea     1.681e+01  5.082e+00   3.307 0.000965 ***
BsmtFinSF1     3.474e+01  2.897e+00  11.994  < 2e-16 ***
BsmtUnfSF      9.076e+00  2.595e+00   3.498 0.000484 ***
CentralAir    -1.183e+04  2.778e+03  -4.257 2.21e-05 ***
GrLivArea      5.751e+01  2.699e+00  21.310  < 2e-16 ***
WoodDeckSF     1.950e+01  6.077e+00   3.210 0.001358 **
OpenPorchSF    3.155e+01  1.480e+01   2.132 0.033161 *
SaleCondition -4.968e+03  1.649e+03  -3.012 0.002639 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23490 on 1445 degrees of freedom
Multiple R-squared:  0.7744,    Adjusted R-squared:  0.7722
F-statistic: 354.3 on 14 and 1445 DF,  p-value: < 2.2e-16
```

Backward Elimination from Full.Model

```{r}
backward.model<- step (full.model, direction = "backward")
summary(backward.model)
```

```
Call:
lm(formula = SalePrice ~ MSSubClass + Neighborhood + LotFrontage +
    LotArea + OverallQual + YearRemodAdd + MasVnrArea + BsmtFinSF1 +
    BsmtUnfSF + CentralAir + LowQualFinSF + GrLivArea + GarageArea +
    WoodDeckSF + OpenPorchSF + SaleCondition, data = train1)

Residuals:
    Min      1Q  Median      3Q     Max
-166192  -14817   -1306   13389   90623

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.390e+05  7.421e+04  -4.568 5.33e-06 ***
MSSubClass    -1.022e+02  1.539e+01  -6.643 4.34e-11 ***
Neighborhood   1.228e+03  1.339e+02   9.173  < 2e-16 ***
LotFrontage    1.197e+02  7.116e+01   1.682  0.09269 .
LotArea        8.100e-01  3.202e-01   2.529  0.01153 *
OverallQual    1.043e+04  7.345e+02  14.203  < 2e-16 ***
YearRemodAdd   1.660e+02  3.805e+01   4.362 1.38e-05 ***
MasVnrArea     1.350e+01  5.098e+00   2.648  0.00820 **
BsmtFinSF1     3.338e+01  2.901e+00  11.507  < 2e-16 ***
BsmtUnfSF      8.195e+00  2.592e+00   3.161  0.00160 **
CentralAir    -1.219e+04  2.713e+03  -4.494 7.56e-06 ***
LowQualFinSF  -2.145e+01  1.293e+01  -1.658  0.09748 .
GrLivArea      5.677e+01  2.675e+00  21.225  < 2e-16 ***
GarageArea     1.002e+01  3.841e+00   2.609  0.00918 **
WoodDeckSF     1.960e+01  6.046e+00   3.242  0.00121 **
OpenPorchSF    2.996e+01  1.474e+01   2.033  0.04221 *
SaleCondition -4.786e+03  1.642e+03  -2.915  0.00361 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23390 on 1443 degrees of freedom
Multiple R-squared:  0.7765,    Adjusted R-squared:  0.774
F-statistic: 313.4 on 16 and 1443 DF,  p-value: < 2.2e-16
```

Model with Top 5

```r
cors <- sapply(train1, cor, y=train1$SalePrice)
mask <- (rank(-abs(cors)) <= 6 )
best5.pred <- train1[, mask]

best5.pred <- subset(best5.pred, select = c(-SalePrice) )
summary(best5.pred)
```

```r
model.best5 <- lm (SalePrice ~ Neighborhood + OverallQual + GrLivArea + GarageCars + GarageArea, data=train1)

model.best5<- step (model.best5, direction = "backward")
```

```r
summary(model.best5)
```

```
Call:
lm(formula = SalePrice ~ Neighborhood + OverallQual + GrLivArea +
    GarageArea, data = train1)

Residuals:
    Min      1Q  Median      3Q     Max
-140933  -15612   -1271   13633  100098

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -14472.334   4212.272  -3.436 0.000608 ***
Neighborhood   1723.765    141.016  12.224  < 2e-16 ***
OverallQual   11896.404    735.461  16.175  < 2e-16 ***
GrLivArea        61.218      2.660  23.017  < 2e-16 ***
GarageArea       28.109      4.024   6.986 4.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25900 on 1455 degrees of freedom
Multiple R-squared:  0.7237,    Adjusted R-squared:  0.7229
F-statistic: 952.6 on 4 and 1455 DF,  p-value: < 2.2e-16
```

Compare Model using ANOVA

```r
anova(full.model,reduced.model,backward.model, model.best5, test="Chisq")
```

```
Analysis of Variance Table

Model 1: SalePrice ~ MSSubClass + Neighborhood + LotFrontage + LotArea +
    BldgType + OverallQual + OverallCond + YearBuilt + YearRemodAdd +
    MasVnrArea + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF + CentralAir +
    LowQualFinSF + GrLivArea + GarageCars + GarageArea + WoodDeckSF +
    OpenPorchSF + EnclosedPorch + ScreenPorch + X3SsnPorch +
    PoolArea + MiscVal + MoSold + YrSold + SaleCondition
Model 2: SalePrice ~ MSSubClass + Neighborhood + LotArea + OverallQual +
    YearBuilt + YearRemodAdd + MasVnrArea + BsmtFinSF1 + BsmtUnfSF +
    CentralAir + GrLivArea + WoodDeckSF + OpenPorchSF + SaleCondition
Model 3: SalePrice ~ MSSubClass + Neighborhood + LotFrontage + LotArea +
    OverallQual + YearRemodAdd + MasVnrArea + BsmtFinSF1 + BsmtUnfSF +
    CentralAir + LowQualFinSF + GrLivArea + GarageArea + WoodDeckSF +
    OpenPorchSF + SaleCondition
Model 4: SalePrice ~ Neighborhood + OverallQual + GrLivArea + GarageArea
  Res.Df          RSS  Df     Sum of Sq  Pr(>Chi)
1   1434 786766781016
2   1445 797081291276 -11  -10314510260  0.064784 .
3   1443 789652291334   2    7428999941  0.001147 **
4   1455 976387611229 -12 -186735319894 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Prepare Test Data

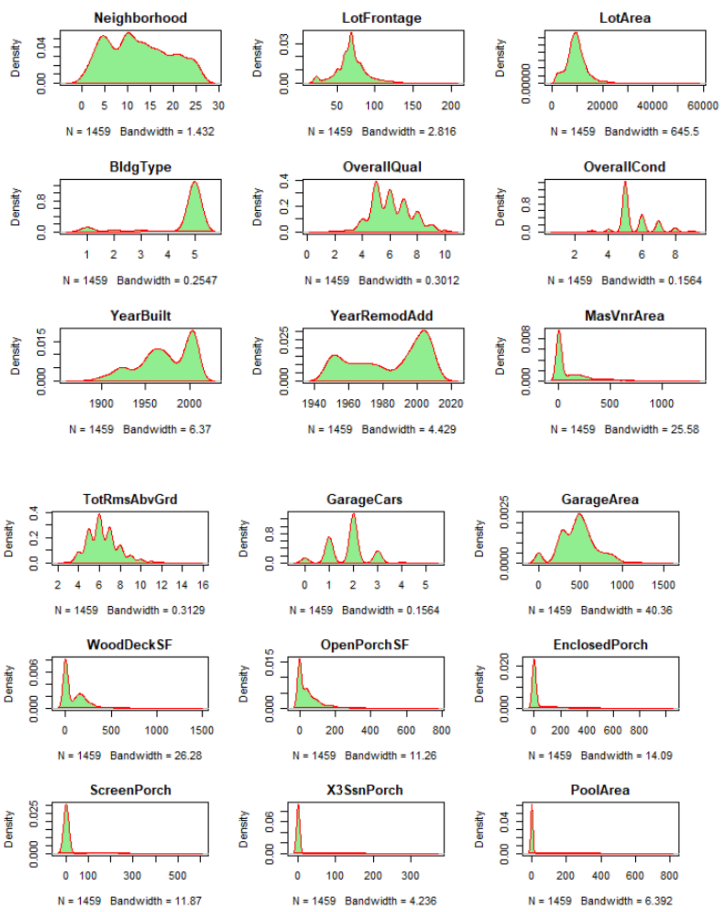
- Select interesting columns
- Change the categorical variables to numeric
- Impute missing data
- Drop a few columns of colinearity
- Fix the outliers

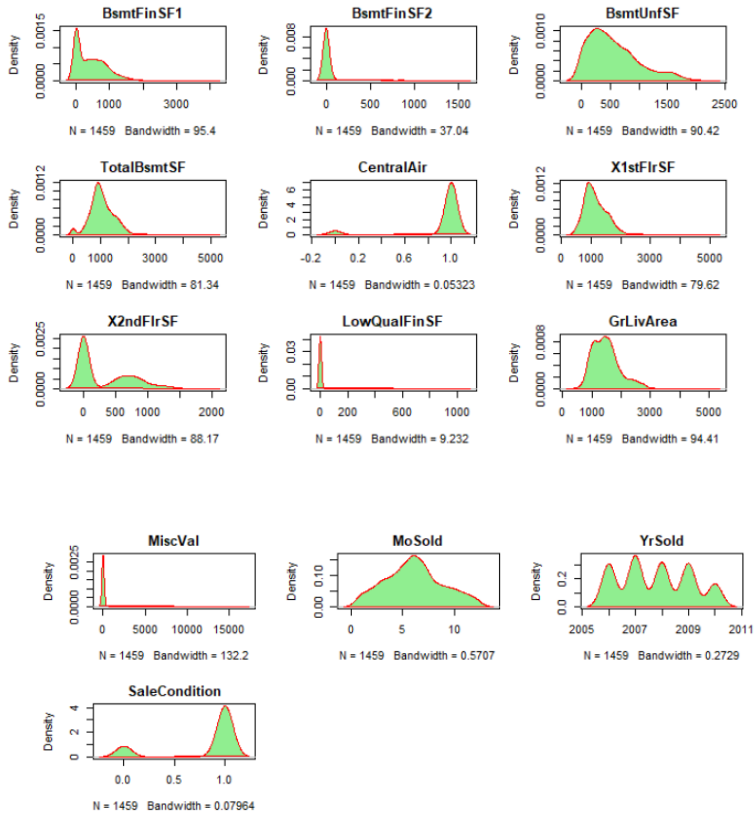
Summary of Test Data

```
   MSSubClass      Neighborhood    LotFrontage        LotArea          BldgType
Min.   : 20.00   Min.   : 1.00   Min.   : 21.00   Min.   : 1470   Min.   :1.000
1st Qu.: 20.00   1st Qu.: 7.00   1st Qu.: 60.00   1st Qu.: 7391   1st Qu.:5.000
Median : 50.00   Median :12.00   Median : 68.58   Median : 9399   Median :5.000
Mean   : 57.38   Mean   :12.55   Mean   : 68.58   Mean   : 9819   Mean   :4.482
3rd Qu.: 70.00   3rd Qu.:17.00   3rd Qu.: 78.00   3rd Qu.:11518   3rd Qu.:5.000
Max.   :190.00   Max.   :25.00   Max.   :200.00   Max.   :56600   Max.   :5.000
 OverallQual      OverallCond      YearBuilt     YearRemodAdd    MasVnrArea
Min.   : 1.000   Min.   :1.000   Min.   :1879   Min.   :1950   Min.   :   0.0
1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1953   1st Qu.:1963   1st Qu.:   0.0
Median : 6.000   Median :5.000   Median :1973   Median :1992   Median :   0.0
Mean   : 6.079   Mean   :5.554   Mean   :1971   Mean   :1984   Mean   : 101.8
3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004   3rd Qu.: 163.5
Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010   Max.   :1290.0
   BsmtFinSF1       BsmtFinSF2        BsmtUnfSF       TotalBsmtSF     CentralAir
Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0   Min.   :0.0000
1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 219.5   1st Qu.: 784   1st Qu.:1.0000
Median : 351.0   Median :   0.00   Median : 460.0   Median : 988   Median :1.0000
Mean   : 439.2   Mean   :  52.62   Mean   : 554.3   Mean   :1046   Mean   :0.9308
3rd Qu.: 752.0   3rd Qu.:   0.00   3rd Qu.: 797.5   3rd Qu.:1304   3rd Qu.:1.0000
Max.   :4010.0   Max.   :1526.00   Max.   :2140.0   Max.   :5095   Max.   :1.0000
   X1stFlrSF       X2ndFlrSF      LowQualFinSF       GrLivArea     TotRmsAbvGrd
Min.   : 407.0   Min.   :   0   Min.   :   0.000   Min.   : 407   Min.   : 3.000
1st Qu.: 873.5   1st Qu.:   0   1st Qu.:   0.000   1st Qu.:1118   1st Qu.: 5.000
Median :1079.0   Median :   0   Median :   0.000   Median :1432   Median : 6.000
Mean   :1156.5   Mean   : 326   Mean   :   3.543   Mean   :1486   Mean   : 6.385
3rd Qu.:1382.5   3rd Qu.: 676   3rd Qu.:   0.000   3rd Qu.:1721   3rd Qu.: 7.000
Max.   :5095.0   Max.   :1862   Max.   :1064.000   Max.   :5095   Max.   :15.000
```

```
  GarageCars      GarageArea       WoodDeckSF      OpenPorchSF     EnclosedPorch
Min.   :0.000   Min.   :   0.0   Min.   :   0.00   Min.   :  0.00   Min.   :   0.00
1st Qu.:1.000   1st Qu.: 318.0   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00
Median :2.000   Median : 480.0   Median :   0.00   Median : 28.00   Median :   0.00
Mean   :1.766   Mean   : 472.8   Mean   :  93.17   Mean   : 48.31   Mean   :  24.24
3rd Qu.:2.000   3rd Qu.: 576.0   3rd Qu.: 168.00   3rd Qu.: 72.00   3rd Qu.:   0.00
Max.   :5.000   Max.   :1488.0   Max.   :1424.00   Max.   :742.00   Max.   :1012.00
 ScreenPorch      X3SsnPorch        PoolArea          MiscVal
Min.   :  0.00   Min.   :  0.000   Min.   :  0.000   Min.   :    0.00
1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:    0.00
Median :  0.00   Median :  0.000   Median :  0.000   Median :    0.00
Mean   : 17.06   Mean   :  1.794   Mean   :  1.744   Mean   :   58.17
3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:    0.00
Max.   :576.00   Max.   :360.000   Max.   :800.000   Max.   :17000.00
    MoSold          YrSold     SaleCondition
Min.   : 1.000   Min.   :2006   Min.   :0.0000
1st Qu.: 4.000   1st Qu.:2007   1st Qu.:1.0000
Median : 6.000   Median :2008   Median :1.0000
Mean   : 6.104   Mean   :2008   Mean   :0.8252
3rd Qu.: 8.000   3rd Qu.:2009   3rd Qu.:1.0000
Max.   :12.000   Max.   :2010   Max.   :1.0000
```

Density Plot of Test Data





Predict SalePrice by 4 Models

#Full Model Prediction

```{r, echo=T, warning=F, message=F}
full.model.pred <- cbind(test1, s<-predict(full.model, test1))
names(full.model.pred)[ncol(full.model.pred)] <- "SalePrice"

full.model.submission <- dplyr::select(full.model.pred,Id,SalePrice)

write.csv(full.model.submission, file="full.model.submission.csv")
```

#Reduced Model Prediction

```{r, echo=T, warning=F, message=F}
reduced.model.pred <- cbind(test1, s<-predict(reduced.model, test1))
names(reduced.model.pred)[ncol(reduced.model.pred)] <- "SalePrice"

reduced.model.submission <- dplyr::select(reduced.model.pred,Id,SalePrice)

write.csv(reduced.model.submission, file="reduced.model.submission.csv")
```

#Backward Model Prediction

```{r, echo=T, warning=F, message=F}
backward.model.pred <- cbind(test1, s<-predict(backward.model, test1))
names(backward.model.pred)[ncol(backward.model.pred)] <- "SalePrice"

backward.model.submission <- dplyr::select(backward.model.pred,Id,SalePrice)

write.csv(backward.model.submission, file="backward.model.submission.csv")
```

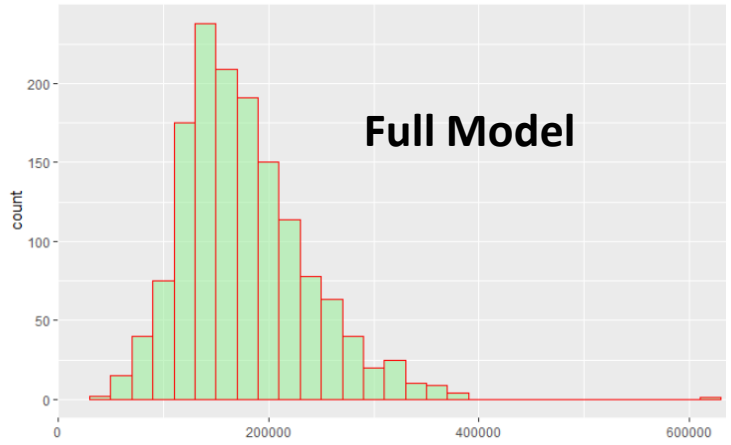
#Model Best 5

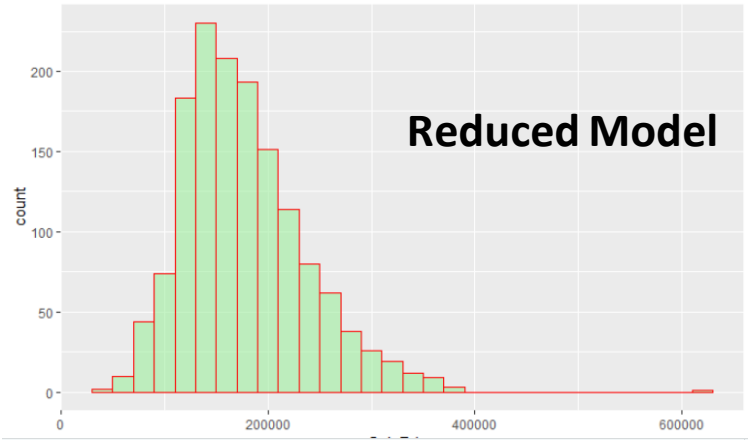
```{r, echo=T, warning=F, message=F}
model.best5.pred <- cbind(test1, s<-predict(model.best5, test1))
names(model.best5.pred)[ncol(model.best5.pred)] <- "SalePrice"

model.best5.submission <- dplyr::select(model.best5.pred,Id,SalePrice)

write.csv(model.best5.submission, file="model.best5.submission.csv")
```

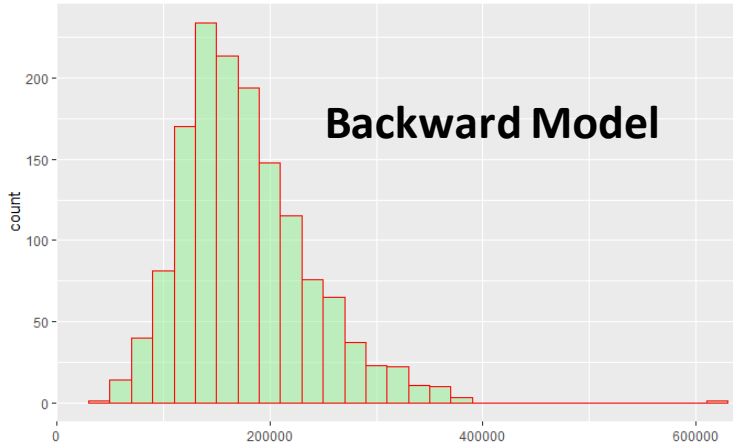
Summary of SalePrice by 4 Medels

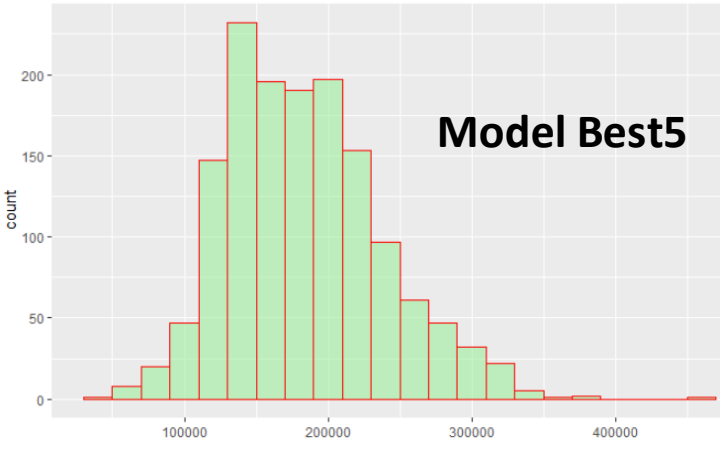
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Full Model	46543	135286	168060	176848	209841	626329
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Reduced Model	49574	135729	168059	177033	209958	629382
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Backward Model	49577	135547	167838	176919	209709	623086
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Model Best5	49480	143550	178091	183741	216539	457456

Histograms of SalePrice in 4 Models









Score of Full Model

Name	Submitted	Wait time	Execution time	Score
submission_11.csv	just now	1 seconds	0 seconds	0.18519

Complete

Jump to your position on the leaderboard

Make a submission for junpan43

Reduced Model

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,557 teams · Ongoing

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission_22.csv	just now	0 seconds	0 seconds	0.18486

Complete

Jump to your position on the leaderboard

Make a submission for junpan43

You have 8 submissions remaining today. This resets an hour from now (00: 00 UTC).

Backward Model

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,557 teams · Ongoing

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission_33.csv	just now	1 seconds	0 seconds	0.18576

Complete

Jump to your position on the leaderboard

Best 5 Model

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,557 teams · Ongoing

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission_44.csv	just now	0 seconds	0 seconds	0.21266

Complete

Jump to your position on the leaderboard