Week-6

Viraj Sapre, Sinchana Mysore Eshwar, Kartik Nagarajan

March 8, 2019

#Loading the libraries  
library(ggplot2)  
library(ggthemes)

## Warning: package 'ggthemes' was built under R version 3.5.2

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.5.2

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.5.2

## corrplot 0.84 loaded

library(GGally)

## Warning: package 'GGally' was built under R version 3.5.2

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(data.table)

## Warning: package 'data.table' was built under R version 3.5.2

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

library(scales)  
library(MVA)

## Warning: package 'MVA' was built under R version 3.5.2

## Loading required package: HSAUR2

## Warning: package 'HSAUR2' was built under R version 3.5.2

## Loading required package: tools

library(Rmisc)

## Warning: package 'Rmisc' was built under R version 3.5.2

## Loading required package: lattice

## Loading required package: plyr

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

# loading the dataset  
training <- read.csv("~/MS SEM 2/Multivariate Analysis - Raunak Parikh/MVA Grp Project/train.csv")  
View(training)

UNDERSTANDING THE DATA

dim(training) # checking the dimensions

## [1] 1460 81

str(training)# checking the structure of dataset

## '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 ...

summary(training)# checking the summary of dataset

## 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:1335 3rd Qu.:1391 3rd Qu.: 728 3rd Qu.: 0.000   
## 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

Checking for MISSING VALUES

#Missing data  
sum(is.na(training)/(nrow(training)\*nrow(training)))# printing percentage of missing data

## [1] 0.003267029

unique(nrow(training)) # printing all the unique values

## [1] 1460

colSums(sapply(training,is.na))# prinitng number of missing values in each column

## Id MSSubClass MSZoning LotFrontage LotArea   
## 0 0 0 259 0   
## Street Alley LotShape LandContour Utilities   
## 0 1369 0 0 0   
## LotConfig LandSlope Neighborhood Condition1 Condition2   
## 0 0 0 0 0   
## BldgType HouseStyle OverallQual OverallCond YearBuilt   
## 0 0 0 0 0   
## YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd   
## 0 0 0 0 0   
## MasVnrType MasVnrArea ExterQual ExterCond Foundation   
## 8 8 0 0 0   
## BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1   
## 37 37 38 37 0   
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating   
## 38 0 0 0 0   
## HeatingQC CentralAir Electrical X1stFlrSF X2ndFlrSF   
## 0 0 0 0 0   
## LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath   
## 0 0 0 0 0   
## HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd   
## 0 0 0 0 0   
## Functional Fireplaces FireplaceQu GarageType GarageYrBlt   
## 0 0 690 81 81   
## GarageFinish GarageCars GarageArea GarageQual GarageCond   
## 81 0 0 81 81   
## PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch   
## 0 0 0 0 0   
## ScreenPorch PoolArea PoolQC Fence MiscFeature   
## 0 0 1453 1179 1406   
## MiscVal MoSold YrSold SaleType SaleCondition   
## 0 0 0 0 0   
## SalePrice   
## 0

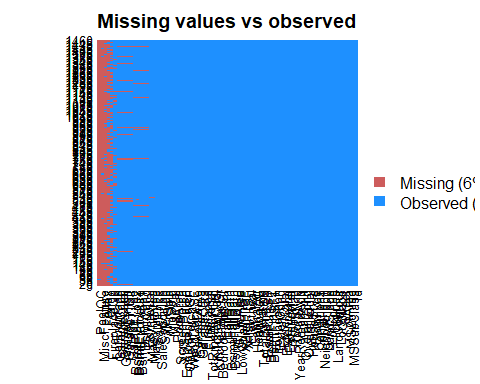
library(Amelia)

## Warning: package 'Amelia' was built under R version 3.5.2

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.5, built: 2018-05-07)  
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

missmap(training, main ="Missing values vs observed")



unique(nrow(training$SalePrice))

## NULL

Removing columns with NA values

training$Alley = NULL   
training$LotFrontage = NULL  
training$FireplaceQu = NULL  
training$Fence = NULL  
training$PoolQC = NULL   
training$MiscFeature = NULL   
training$BsmtQual = NULL   
training$BsmtCond = NULL   
training$BsmtExposure = NULL   
training$BsmtFinType1 = NULL   
training$BsmtFinType2 = NULL  
training$GarageType = NULL  
training$GarageYrBlt = NULL  
training$MasVnrType = NULL   
training$MasVnrArea = NULL   
training$GarageQual = NULL   
training$GarageFinish = NULL   
training$GarageCond = NULL  
training$Id=NULL  
training$BsmtFinSF1=NULL  
training$BsmtFinSF2=NULL  
training$X1stFlrSF=NULL  
training$X2stFlrSF

## NULL

training$Age=training$YrSold-training$YearBuilt

# creating dataframe of categorical and numerical variables  
catvar <- c('MSZoning','Street', 'Neighborhood', 'LandContour','BldgType', 'LandSlope', 'RoofStyle',  
 'HouseStyle','CentralAir','PavedDrive','SaleCondition','OverallCond' )  
numvar<-c('LotArea','TotalBsmtSF','GrLivArea','BedroomAbvGr','GarageCars','GarageArea','OpenPorchSF','EnclosedPorch','WoodDeckSF','PoolArea','Age')

training[!complete.cases(training),]

## [1] MSSubClass MSZoning LotArea Street LotShape   
## [6] LandContour Utilities LotConfig LandSlope Neighborhood   
## [11] Condition1 Condition2 BldgType HouseStyle OverallQual   
## [16] OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl   
## [21] Exterior1st Exterior2nd ExterQual ExterCond Foundation   
## [26] BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir   
## [31] Electrical X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath   
## [36] BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr   
## [41] KitchenQual TotRmsAbvGrd Functional Fireplaces GarageCars   
## [46] GarageArea PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch  
## [51] X3SsnPorch ScreenPorch PoolArea MiscVal MoSold   
## [56] YrSold SaleType SaleCondition SalePrice Age   
## <0 rows> (or 0-length row.names)

head(training)

## MSSubClass MSZoning LotArea Street LotShape LandContour Utilities  
## 1 60 RL 8450 Pave Reg Lvl AllPub  
## 2 20 RL 9600 Pave Reg Lvl AllPub  
## 3 60 RL 11250 Pave IR1 Lvl AllPub  
## 4 70 RL 9550 Pave IR1 Lvl AllPub  
## 5 60 RL 14260 Pave IR1 Lvl AllPub  
## 6 50 RL 14115 Pave IR1 Lvl AllPub  
## LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType  
## 1 Inside Gtl CollgCr Norm Norm 1Fam  
## 2 FR2 Gtl Veenker Feedr Norm 1Fam  
## 3 Inside Gtl CollgCr Norm Norm 1Fam  
## 4 Corner Gtl Crawfor Norm Norm 1Fam  
## 5 FR2 Gtl NoRidge Norm Norm 1Fam  
## 6 Inside Gtl Mitchel Norm Norm 1Fam  
## HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle  
## 1 2Story 7 5 2003 2003 Gable  
## 2 1Story 6 8 1976 1976 Gable  
## 3 2Story 7 5 2001 2002 Gable  
## 4 2Story 7 5 1915 1970 Gable  
## 5 2Story 8 5 2000 2000 Gable  
## 6 1.5Fin 5 5 1993 1995 Gable  
## RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation  
## 1 CompShg VinylSd VinylSd Gd TA PConc  
## 2 CompShg MetalSd MetalSd TA TA CBlock  
## 3 CompShg VinylSd VinylSd Gd TA PConc  
## 4 CompShg Wd Sdng Wd Shng TA TA BrkTil  
## 5 CompShg VinylSd VinylSd Gd TA PConc  
## 6 CompShg VinylSd VinylSd TA TA Wood  
## BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical X2ndFlrSF  
## 1 150 856 GasA Ex Y SBrkr 854  
## 2 284 1262 GasA Ex Y SBrkr 0  
## 3 434 920 GasA Ex Y SBrkr 866  
## 4 540 756 GasA Gd Y SBrkr 756  
## 5 490 1145 GasA Ex Y SBrkr 1053  
## 6 64 796 GasA Ex Y SBrkr 566  
## LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath  
## 1 0 1710 1 0 2 1  
## 2 0 1262 0 1 2 0  
## 3 0 1786 1 0 2 1  
## 4 0 1717 1 0 1 0  
## 5 0 2198 1 0 2 1  
## 6 0 1362 1 0 1 1  
## BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces  
## 1 3 1 Gd 8 Typ 0  
## 2 3 1 TA 6 Typ 1  
## 3 3 1 Gd 6 Typ 1  
## 4 3 1 Gd 7 Typ 1  
## 5 4 1 Gd 9 Typ 1  
## 6 1 1 TA 5 Typ 0  
## GarageCars GarageArea PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch  
## 1 2 548 Y 0 61 0  
## 2 2 460 Y 298 0 0  
## 3 2 608 Y 0 42 0  
## 4 3 642 Y 0 35 272  
## 5 3 836 Y 192 84 0  
## 6 2 480 Y 40 30 0  
## X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SaleType  
## 1 0 0 0 0 2 2008 WD  
## 2 0 0 0 0 5 2007 WD  
## 3 0 0 0 0 9 2008 WD  
## 4 0 0 0 0 2 2006 WD  
## 5 0 0 0 0 12 2008 WD  
## 6 320 0 0 700 10 2009 WD  
## SaleCondition SalePrice Age  
## 1 Normal 208500 5  
## 2 Normal 181500 31  
## 3 Normal 223500 7  
## 4 Abnorml 140000 91  
## 5 Normal 250000 8  
## 6 Normal 143000 16

#Missing data  
sum(is.na(training)/(nrow(training)\*nrow(training)))# printing percentage of missing data

## [1] 0

unique(nrow(training)) # printing all the unique values

## [1] 1460

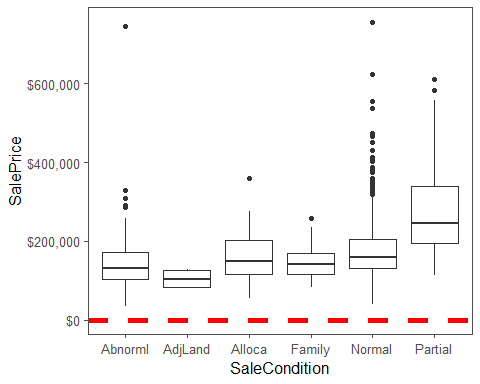
colSums(sapply(training,is.na))# prinitng number of missing values in each column

## MSSubClass MSZoning LotArea Street LotShape   
## 0 0 0 0 0   
## LandContour Utilities LotConfig LandSlope Neighborhood   
## 0 0 0 0 0   
## Condition1 Condition2 BldgType HouseStyle OverallQual   
## 0 0 0 0 0   
## OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl   
## 0 0 0 0 0   
## Exterior1st Exterior2nd ExterQual ExterCond Foundation   
## 0 0 0 0 0   
## BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir   
## 0 0 0 0 0   
## Electrical X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath   
## 0 0 0 0 0   
## BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr   
## 0 0 0 0 0   
## KitchenQual TotRmsAbvGrd Functional Fireplaces GarageCars   
## 0 0 0 0 0   
## GarageArea PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch   
## 0 0 0 0 0   
## X3SsnPorch ScreenPorch PoolArea MiscVal MoSold   
## 0 0 0 0 0   
## YrSold SaleType SaleCondition SalePrice Age   
## 0 0 0 0 0

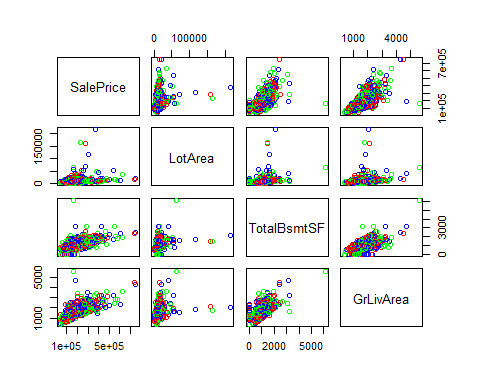
attach(training)   
catdf<-training[,catvar]  
numdf<-training[,numvar]

VISUALIZING THE DATA

ggplot(training, aes(x = SaleCondition, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



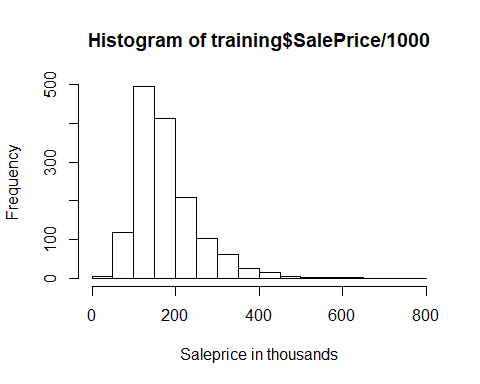
pairs(~SalePrice+LotArea+TotalBsmtSF+GrLivArea, data=training,col=c('red','blue','green'))



as.factor(training$SalePrice)

## [1] 208500 181500 223500 140000 250000 143000 307000 200000 129900  
## [10] 118000 129500 345000 144000 279500 157000 132000 149000 90000   
## [19] 159000 139000 325300 139400 230000 129900 154000 256300 134800  
## [28] 306000 207500 68500 40000 149350 179900 165500 277500 309000  
## [37] 145000 153000 109000 82000 160000 170000 144000 130250 141000  
## [46] 319900 239686 249700 113000 127000 177000 114500 110000 385000  
## [55] 130000 180500 172500 196500 438780 124900 158000 101000 202500  
## [64] 140000 219500 317000 180000 226000 80000 225000 244000 129500  
## [73] 185000 144900 107400 91000 135750 127000 136500 110000 193500  
## [82] 153500 245000 126500 168500 260000 174000 164500 85000 123600  
## [91] 109900 98600 163500 133900 204750 185000 214000 94750 83000   
## [100] 128950 205000 178000 118964 198900 169500 250000 100000 115000  
## [109] 115000 190000 136900 180000 383970 217000 259500 176000 139000  
## [118] 155000 320000 163990 180000 100000 136000 153900 181000 84500   
## [127] 128000 87000 155000 150000 226000 244000 150750 220000 180000  
## [136] 174000 143000 171000 230000 231500 115000 260000 166000 204000  
## [145] 125000 130000 105000 222500 141000 115000 122000 372402 190000  
## [154] 235000 125000 79000 109500 269500 254900 320000 162500 412500  
## [163] 220000 103200 152000 127500 190000 325624 183500 228000 128500  
## [172] 215000 239000 163000 184000 243000 211000 172500 501837 100000  
## [181] 177000 200100 120000 200000 127000 475000 173000 135000 153337  
## [190] 286000 315000 184000 192000 130000 127000 148500 311872 235000  
## [199] 104000 274900 140000 171500 112000 149000 110000 180500 143900  
## [208] 141000 277000 145000 98000 186000 252678 156000 161750 134450  
## [217] 210000 107000 311500 167240 204900 200000 179900 97000 386250  
## [226] 112000 290000 106000 125000 192500 148000 403000 94500 128200  
## [235] 216500 89500 185500 194500 318000 113000 262500 110500 79000   
## [244] 120000 205000 241500 137000 140000 180000 277000 76500 235000  
## [253] 173000 158000 145000 230000 207500 220000 231500 97000 176000  
## [262] 276000 151000 130000 73000 175500 185000 179500 120500 148000  
## [271] 266000 241500 290000 139000 124500 205000 201000 141000 415298  
## [280] 192000 228500 185000 207500 244600 179200 164700 159000 88000   
## [289] 122000 153575 233230 135900 131000 235000 167000 142500 152000  
## [298] 239000 175000 158500 157000 267000 205000 149900 295000 305900  
## [307] 225000 89500 82500 360000 165600 132000 119900 375000 178000  
## [316] 188500 260000 270000 260000 187500 342643 354000 301000 126175  
## [325] 242000 87000 324000 145250 214500 78000 119000 139000 284000  
## [334] 207000 192000 228950 377426 214000 202500 155000 202900 82000   
## [343] 87500 266000 85000 140200 151500 157500 154000 437154 318061  
## [352] 190000 95000 105900 140000 177500 173000 134000 130000 280000  
## [361] 156000 145000 198500 118000 190000 147000 159000 165000 132000  
## [370] 162000 172400 134432 125000 123000 219500 61000 148000 340000  
## [379] 394432 179000 127000 187750 213500 76000 240000 192000 81000   
## [388] 125000 191000 426000 119000 215000 106500 100000 109000 129000  
## [397] 123000 169500 67000 241000 245500 164990 108000 258000 168000  
## [406] 150000 115000 177000 280000 339750 60000 145000 222000 115000  
## [415] 228000 181134 149500 239000 126000 142000 206300 215000 113000  
## [424] 315000 139000 135000 275000 109008 195400 175000 85400 79900   
## [433] 122500 181000 81000 212000 116000 119000 90350 110000 555000  
## [442] 118000 162900 172500 210000 127500 190000 199900 119500 120000  
## [451] 110000 280000 204000 210000 188000 175500 98000 256000 161000  
## [460] 110000 263435 155000 62383 188700 124000 178740 167000 146500  
## [469] 250000 187000 212000 190000 148000 440000 251000 132500 208900  
## [478] 380000 297000 89471 326000 374000 155000 164000 132500 147000  
## [487] 156000 175000 160000 86000 115000 133000 172785 155000 91300   
## [496] 34900 430000 184000 130000 120000 113000 226700 140000 289000  
## [505] 147000 124500 215000 208300 161000 124500 164900 202665 129900  
## [514] 134000 96500 402861 158000 265000 211000 234000 106250 150000  
## [523] 159000 184750 315750 176000 132000 446261 86000 200624 175000  
## [532] 128000 107500 39300 178000 107500 188000 111250 158000 272000  
## [541] 315000 248000 213250 133000 179665 229000 210000 129500 125000  
## [550] 263000 140000 112500 255500 108000 284000 113000 141000 108000  
## [559] 175000 234000 121500 170000 108000 185000 268000 128000 325000  
## [568] 214000 316600 135960 142600 120000 224500 170000 139000 118500  
## [577] 145000 164500 146000 131500 181900 253293 118500 325000 133000  
## [586] 369900 130000 137000 143000 79500 185900 451950 138000 140000  
## [595] 110000 319000 114504 194201 217500 151000 275000 141000 220000  
## [604] 151000 221000 205000 152000 225000 359100 118500 313000 148000  
## [613] 261500 147000 75500 137500 183200 105500 314813 305000 67000   
## [622] 240000 135000 168500 165150 160000 139900 153000 135000 168500  
## [631] 124000 209500 82500 139400 144000 200000 60000 93000 85000   
## [640] 264561 274000 226000 345000 152000 370878 143250 98300 155000  
## [649] 155000 84500 205950 108000 191000 135000 350000 88000 145500  
## [658] 149000 97500 167000 197900 402000 110000 137500 423000 230500  
## [667] 129000 193500 168000 137500 173500 103600 165000 257500 140000  
## [676] 148500 87000 109500 372500 128500 143000 159434 173000 285000  
## [685] 221000 207500 227875 148800 392000 194700 141000 755000 335000  
## [694] 108480 141500 176000 89000 123500 138500 196000 312500 140000  
## [703] 361919 140000 213000 55000 302000 254000 179540 109900 52000   
## [712] 102776 189000 129000 130500 165000 159500 157000 341000 128500  
## [721] 275000 143000 124500 135000 320000 120500 222000 194500 110000  
## [730] 103000 236500 187500 222500 131400 108000 163000 93500 239900  
## [739] 179000 190000 132000 142000 179000 175000 180000 299800 236000  
## [748] 265979 260400 98000 96500 162000 217000 275500 156000 172500  
## [757] 212000 158900 179400 290000 127500 100000 215200 337000 270000  
## [766] 264132 196500 160000 216837 538000 134900 102000 107000 114500  
## [775] 395000 162000 221500 142500 144000 135000 176000 175900 187100  
## [784] 165500 128000 161500 139000 233000 107900 187500 160200 146800  
## [793] 269790 225000 194500 171000 143500 110000 485000 175000 200000  
## [802] 109900 189000 582933 118000 227680 135500 223500 159950 106000  
## [811] 181000 144500 55993 157900 116000 224900 137000 271000 155000  
## [820] 224000 183000 93000 225000 139500 232600 385000 109500 189000  
## [829] 185000 147400 166000 151000 237000 167000 139950 128000 153500  
## [838] 100000 144000 130500 140000 157500 174900 141000 153900 171000  
## [847] 213000 133500 240000 187000 131500 215000 164000 158000 170000  
## [856] 127000 147000 174000 152000 250000 189950 131500 152000 132500  
## [865] 250580 148500 248900 129000 169000 236000 109500 200500 116000  
## [874] 133000 66500 303477 132250 350000 148000 136500 157000 187500  
## [883] 178000 118500 100000 328900 145000 135500 268000 149500 122900  
## [892] 172500 154500 165000 118858 140000 106500 142953 611657 135000  
## [901] 110000 153000 180000 240000 125500 128000 255000 250000 131000  
## [910] 174000 154300 143500 88000 145000 173733 75000 35311 135000  
## [919] 238000 176500 201000 145900 169990 193000 207500 175000 285000  
## [928] 176000 236500 222000 201000 117500 320000 190000 242000 79900   
## [937] 184900 253000 239799 244400 150900 214000 150000 143000 137500  
## [946] 124900 143000 270000 192500 197500 129000 119900 133900 172000  
## [955] 127500 145000 124000 132000 185000 155000 116500 272000 155000  
## [964] 239000 214900 178900 160000 135000 37900 140000 135000 173000  
## [973] 99500 182000 167500 165000 85500 199900 110000 139000 178400  
## [982] 336000 159895 255900 126000 125000 117000 395192 195000 197000  
## [991] 348000 168000 187000 173900 337500 121600 136500 185000 91000   
## [1000] 206000 82000 86000 232000 136905 181000 149900 163500 88000   
## [1009] 240000 102000 135000 100000 165000 85000 119200 227000 203000  
## [1018] 187500 160000 213490 176000 194000 87000 191000 287000 112500  
## [1027] 167500 293077 105000 118000 160000 197000 310000 230000 119750  
## [1036] 84000 315500 287000 97000 80000 155000 173000 196000 262280  
## [1045] 278000 139600 556581 145000 115000 84900 176485 200141 165000  
## [1054] 144500 255000 180000 185850 248000 335000 220000 213500 81000   
## [1063] 90000 110500 154000 328000 178000 167900 151400 135000 135000  
## [1072] 154000 91500 159500 194000 219500 170000 138800 155900 126000  
## [1081] 145000 133000 192000 160000 187500 147000 83500 252000 137500  
## [1090] 197000 92900 160000 136500 146000 129000 176432 127000 170000  
## [1099] 128000 157000 60000 119500 135000 159500 106000 325000 179900  
## [1108] 274725 181000 280000 188000 205000 129900 134500 117000 318000  
## [1117] 184100 130000 140000 133700 118400 212900 112000 118000 163900  
## [1126] 115000 174000 259000 215000 140000 135000 93500 117500 239500  
## [1135] 169000 102000 119000 94000 196000 144000 139000 197500 424870  
## [1144] 80000 80000 149000 180000 174500 116900 143000 124000 149900  
## [1153] 230000 120500 201800 218000 179900 230000 235128 185000 146000  
## [1162] 224000 129000 108959 194000 233170 245350 173000 235000 625000  
## [1171] 171000 163000 171900 200500 239000 285000 119500 115000 154900  
## [1180] 93000 250000 392500 745000 120000 186700 104900 95000 262000  
## [1189] 195000 189000 168000 174000 125000 165000 158000 176000 219210  
## [1198] 144000 178000 148000 116050 197900 117000 213000 153500 271900  
## [1207] 107000 200000 140000 290000 189000 164000 113000 145000 134500  
## [1216] 125000 112000 229456 80500 91500 115000 134000 143000 137900  
## [1225] 184000 145000 214000 147000 367294 127000 190000 132500 101800  
## [1234] 142000 130000 138887 175500 195000 142500 265900 224900 248328  
## [1243] 170000 465000 230000 178000 186500 169900 129500 119000 244000  
## [1252] 171750 130000 294000 165400 127500 301500 99900 190000 151000  
## [1261] 181000 128900 161500 180500 181000 183900 122000 378500 381000  
## [1270] 144000 260000 185750 137000 177000 139000 137000 162000 197900  
## [1279] 237000 68400 227000 180000 150500 139000 169000 132500 143000  
## [1288] 190000 278000 281000 180500 119500 107500 162900 115000 138500  
## [1297] 155000 140000 160000 154000 225000 177500 290000 232000 130000  
## [1306] 325000 202500 138000 147000 179200 335000 203000 302000 333168  
## [1315] 119000 206900 295493 208900 275000 111000 156500 72500 190000  
## [1324] 82500 147000 55000 79000 130500 256000 176500 227000 132500  
## [1333] 100000 125500 125000 167900 135000 52500 200000 128500 123000  
## [1342] 155000 228500 177000 155835 108500 262500 283463 215000 122000  
## [1351] 200000 171000 134900 410000 235000 170000 110000 149900 177500  
## [1360] 315000 189000 260000 104900 156932 144152 216000 193000 127000  
## [1369] 144000 232000 105000 165500 274300 466500 250000 239000 91000   
## [1378] 117000 83000 167500 58500 237500 157000 112000 105000 125500  
## [1387] 250000 136000 377500 131000 235000 124000 123000 163000 246578  
## [1396] 281213 160000 137500 138000 137450 120000 193000 193879 282922  
## [1405] 105000 275000 133000 112000 125500 215000 230000 140000 90000   
## [1414] 257000 207000 175900 122500 340000 124000 223000 179900 127500  
## [1423] 136500 274970 144000 142000 271000 140000 119000 182900 192140  
## [1432] 143750 64500 186500 160000 174000 120500 394617 149700 197000  
## [1441] 191000 149300 310000 121000 179600 129000 157900 240000 112000  
## [1450] 92000 136000 287090 145000 84500 185000 175000 210000 266500  
## [1459] 142125 147500  
## 663 Levels: 34900 35311 37900 39300 40000 52000 52500 55000 55993 ... 755000

hist(training$SalePrice / 1000, xlab = "Saleprice in thousands")



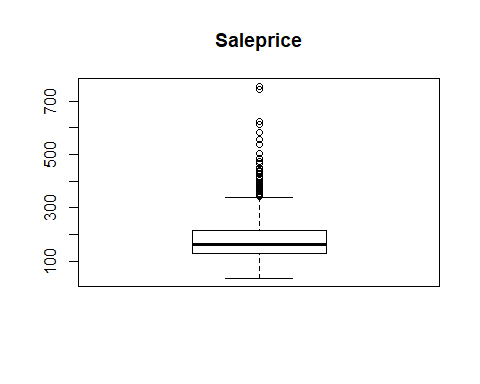
library(moments)

## Warning: package 'moments' was built under R version 3.5.2

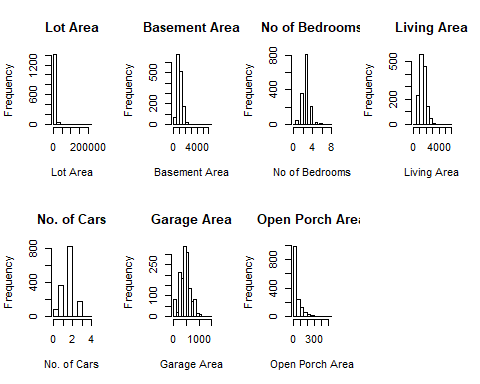
skewness(SalePrice)

## [1] 1.880941

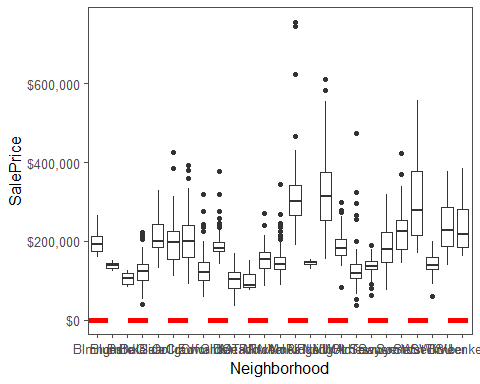
boxplot(training$SalePrice/ 1000, main = "Saleprice")



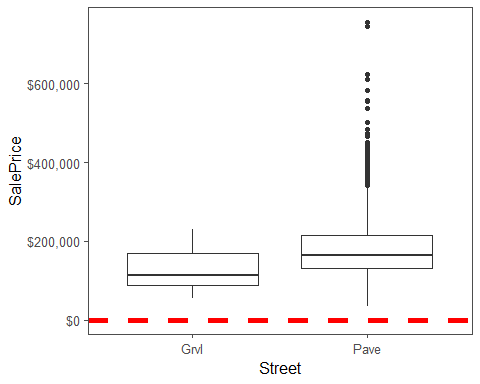
par(mfrow=c(2,4))  
hist(training$LotArea,xlab="Lot Area", main="Lot Area")  
hist(training$TotalBsmtSF, xlab="Basement Area", main="Basement Area")  
hist(training$BedroomAbvGr, xlab="No of Bedrooms", main="No of Bedrooms")  
hist(training$GrLivArea, xlab="Living Area",main="Living Area")  
hist(training$GarageCars, xlab="No. of Cars",main="No. of Cars")  
hist(training$GarageArea, xlab="Garage Area",main="Garage Area")  
hist(training$OpenPorchSF, xlab="Open Porch Area",main="Open Porch Area")



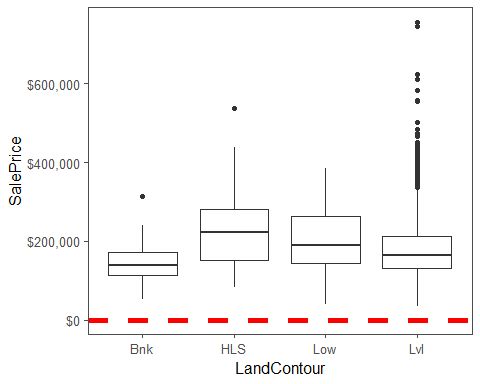
ggplot(training, aes(x = Neighborhood, y = SalePrice)) +  
 geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



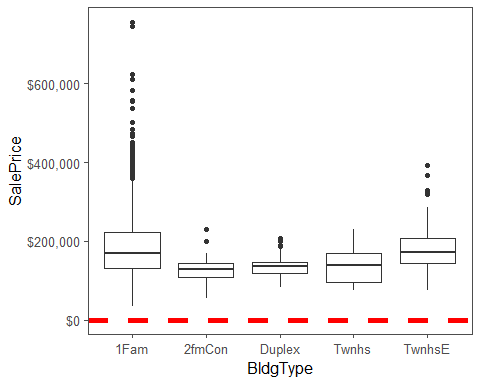
ggplot(training, aes(x = Street, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



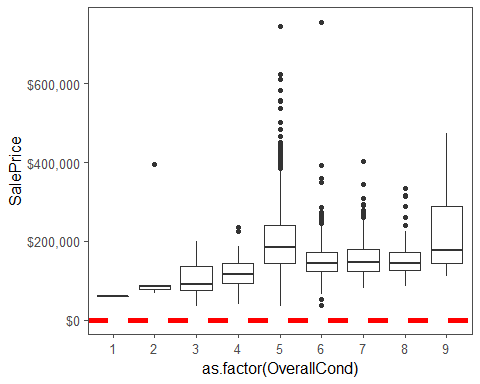
ggplot(training, aes(x = LandContour, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



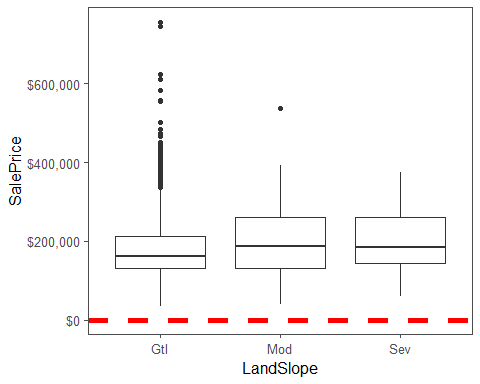
ggplot(training, aes(x = BldgType, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



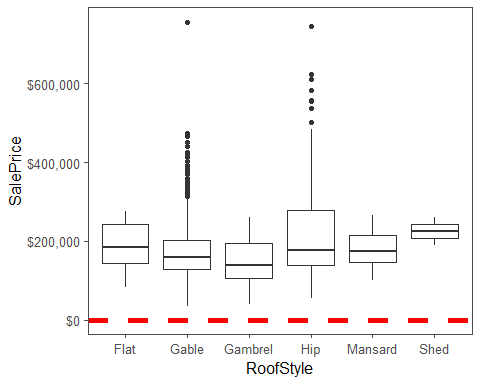
ggplot(training, aes(x = as.factor(OverallCond), y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



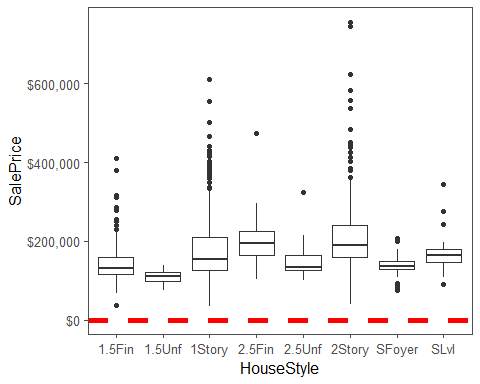
ggplot(training, aes(x = LandSlope, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



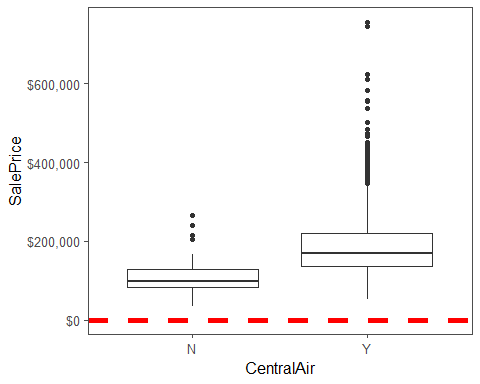
ggplot(training, aes(x = RoofStyle, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



ggplot(training, aes(x = HouseStyle, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



ggplot(training, aes(x = CentralAir, y = SalePrice)) +geom\_boxplot() +  
 geom\_hline(aes(yintercept=80),   
 colour='red', linetype='dashed', lwd=2) +  
 scale\_y\_continuous(labels=dollar\_format()) +  
 theme\_few()



library(PerformanceAnalytics)

## Warning: package 'PerformanceAnalytics' was built under R version 3.5.2

## Loading required package: xts

## Warning: package 'xts' was built under R version 3.5.2

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':  
##   
## first, last

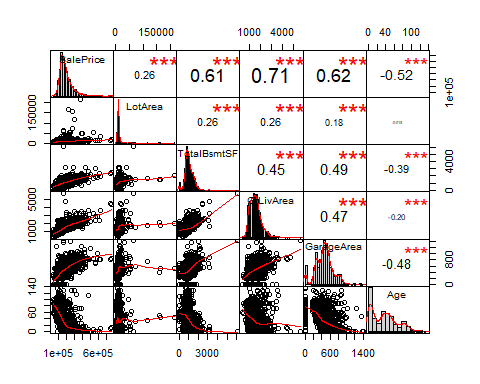
## The following objects are masked from 'package:dplyr':  
##   
## first, last

##   
## Attaching package: 'PerformanceAnalytics'

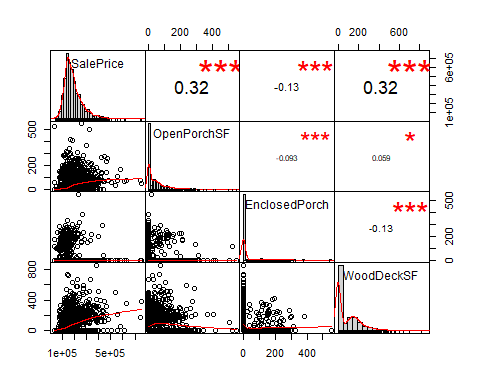
## The following objects are masked from 'package:moments':  
##   
## kurtosis, skewness

## The following object is masked from 'package:graphics':  
##   
## legend

my\_data <- training[, c('SalePrice','LotArea','TotalBsmtSF','GrLivArea','GarageArea','Age')]  
  
chart.Correlation(my\_data, histogram=TRUE, pch=19)



my\_data <- training[, c('SalePrice','OpenPorchSF','EnclosedPorch','WoodDeckSF')]  
  
chart.Correlation(my\_data, histogram=TRUE, pch=19)



library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.5.2

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:plyr':  
##   
## is.discrete, summarize

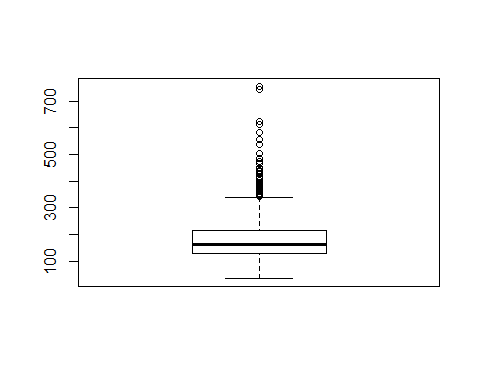
## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

describe(training)

## training   
##   
## 60 Variables 1460 Observations  
## ---------------------------------------------------------------------------  
## MSSubClass   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 15 0.94 56.9 43.19 20 20   
## .25 .50 .75 .90 .95   
## 20 50 70 120 160   
##   
## Value 20 30 40 45 50 60 70 75 80 85  
## Frequency 536 69 4 12 144 299 60 16 58 20  
## Proportion 0.367 0.047 0.003 0.008 0.099 0.205 0.041 0.011 0.040 0.014  
##   
## Value 90 120 160 180 190  
## Frequency 52 87 63 10 30  
## Proportion 0.036 0.060 0.043 0.007 0.021  
## ---------------------------------------------------------------------------  
## MSZoning   
## n missing distinct   
## 1460 0 5   
##   
## Value C (all) FV RH RL RM  
## Frequency 10 65 16 1151 218  
## Proportion 0.007 0.045 0.011 0.788 0.149  
## ---------------------------------------------------------------------------  
## LotArea   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 1073 1 10517 5718 3312 5000   
## .25 .50 .75 .90 .95   
## 7554 9478 11602 14382 17401   
##   
## lowest : 1300 1477 1491 1526 1533, highest: 70761 115149 159000 164660 215245  
## ---------------------------------------------------------------------------  
## Street   
## n missing distinct   
## 1460 0 2   
##   
## Value Grvl Pave  
## Frequency 6 1454  
## Proportion 0.004 0.996  
## ---------------------------------------------------------------------------  
## LotShape   
## n missing distinct   
## 1460 0 4   
##   
## Value IR1 IR2 IR3 Reg  
## Frequency 484 41 10 925  
## Proportion 0.332 0.028 0.007 0.634  
## ---------------------------------------------------------------------------  
## LandContour   
## n missing distinct   
## 1460 0 4   
##   
## Value Bnk HLS Low Lvl  
## Frequency 63 50 36 1311  
## Proportion 0.043 0.034 0.025 0.898  
## ---------------------------------------------------------------------------  
## Utilities   
## n missing distinct   
## 1460 0 2   
##   
## Value AllPub NoSeWa  
## Frequency 1459 1  
## Proportion 0.999 0.001  
## ---------------------------------------------------------------------------  
## LotConfig   
## n missing distinct   
## 1460 0 5   
##   
## Value Corner CulDSac FR2 FR3 Inside  
## Frequency 263 94 47 4 1052  
## Proportion 0.180 0.064 0.032 0.003 0.721  
## ---------------------------------------------------------------------------  
## LandSlope   
## n missing distinct   
## 1460 0 3   
##   
## Value Gtl Mod Sev  
## Frequency 1382 65 13  
## Proportion 0.947 0.045 0.009  
## ---------------------------------------------------------------------------  
## Neighborhood   
## n missing distinct   
## 1460 0 25   
##   
## lowest : Blmngtn Blueste BrDale BrkSide ClearCr  
## highest: Somerst StoneBr SWISU Timber Veenker  
## ---------------------------------------------------------------------------  
## Condition1   
## n missing distinct   
## 1460 0 9   
##   
## Value Artery Feedr Norm PosA PosN RRAe RRAn RRNe RRNn  
## Frequency 48 81 1260 8 19 11 26 2 5  
## Proportion 0.033 0.055 0.863 0.005 0.013 0.008 0.018 0.001 0.003  
## ---------------------------------------------------------------------------  
## Condition2   
## n missing distinct   
## 1460 0 8   
##   
## Value Artery Feedr Norm PosA PosN RRAe RRAn RRNn  
## Frequency 2 6 1445 1 2 1 1 2  
## Proportion 0.001 0.004 0.990 0.001 0.001 0.001 0.001 0.001  
## ---------------------------------------------------------------------------  
## BldgType   
## n missing distinct   
## 1460 0 5   
##   
## Value 1Fam 2fmCon Duplex Twnhs TwnhsE  
## Frequency 1220 31 52 43 114  
## Proportion 0.836 0.021 0.036 0.029 0.078  
## ---------------------------------------------------------------------------  
## HouseStyle   
## n missing distinct   
## 1460 0 8   
##   
## Value 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer SLvl  
## Frequency 154 14 726 8 11 445 37 65  
## Proportion 0.105 0.010 0.497 0.005 0.008 0.305 0.025 0.045  
## ---------------------------------------------------------------------------  
## OverallQual   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 10 0.951 6.099 1.522 4 5   
## .25 .50 .75 .90 .95   
## 5 6 7 8 8   
##   
## Value 1 2 3 4 5 6 7 8 9 10  
## Frequency 2 3 20 116 397 374 319 168 43 18  
## Proportion 0.001 0.002 0.014 0.079 0.272 0.256 0.218 0.115 0.029 0.012  
## ---------------------------------------------------------------------------  
## OverallCond   
## n missing distinct Info Mean Gmd   
## 1460 0 9 0.814 5.575 1.111   
##   
## Value 1 2 3 4 5 6 7 8 9  
## Frequency 1 5 25 57 821 252 205 72 22  
## Proportion 0.001 0.003 0.017 0.039 0.562 0.173 0.140 0.049 0.015  
## ---------------------------------------------------------------------------  
## YearBuilt   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 112 1 1971 33.88 1916 1925   
## .25 .50 .75 .90 .95   
## 1954 1973 2000 2006 2007   
##   
## lowest : 1872 1875 1880 1882 1885, highest: 2006 2007 2008 2009 2010  
## ---------------------------------------------------------------------------  
## YearRemodAdd   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 61 0.997 1985 23.05 1950 1950   
## .25 .50 .75 .90 .95   
## 1967 1994 2004 2006 2007   
##   
## lowest : 1950 1951 1952 1953 1954, highest: 2006 2007 2008 2009 2010  
## ---------------------------------------------------------------------------  
## RoofStyle   
## n missing distinct   
## 1460 0 6   
##   
## Value Flat Gable Gambrel Hip Mansard Shed  
## Frequency 13 1141 11 286 7 2  
## Proportion 0.009 0.782 0.008 0.196 0.005 0.001  
## ---------------------------------------------------------------------------  
## RoofMatl   
## n missing distinct   
## 1460 0 8   
##   
## Value ClyTile CompShg Membran Metal Roll Tar&Grv WdShake WdShngl  
## Frequency 1 1434 1 1 1 11 5 6  
## Proportion 0.001 0.982 0.001 0.001 0.001 0.008 0.003 0.004  
## ---------------------------------------------------------------------------  
## Exterior1st   
## n missing distinct   
## 1460 0 15   
##   
## Value AsbShng AsphShn BrkComm BrkFace CBlock CemntBd HdBoard ImStucc  
## Frequency 20 1 2 50 1 61 222 1  
## Proportion 0.014 0.001 0.001 0.034 0.001 0.042 0.152 0.001  
##   
## Value MetalSd Plywood Stone Stucco VinylSd Wd Sdng WdShing  
## Frequency 220 108 2 25 515 206 26  
## Proportion 0.151 0.074 0.001 0.017 0.353 0.141 0.018  
## ---------------------------------------------------------------------------  
## Exterior2nd   
## n missing distinct   
## 1460 0 16   
##   
## Value AsbShng AsphShn Brk Cmn BrkFace CBlock CmentBd HdBoard ImStucc  
## Frequency 20 3 7 25 1 60 207 10  
## Proportion 0.014 0.002 0.005 0.017 0.001 0.041 0.142 0.007  
##   
## Value MetalSd Other Plywood Stone Stucco VinylSd Wd Sdng Wd Shng  
## Frequency 214 1 142 5 26 504 197 38  
## Proportion 0.147 0.001 0.097 0.003 0.018 0.345 0.135 0.026  
## ---------------------------------------------------------------------------  
## ExterQual   
## n missing distinct   
## 1460 0 4   
##   
## Value Ex Fa Gd TA  
## Frequency 52 14 488 906  
## Proportion 0.036 0.010 0.334 0.621  
## ---------------------------------------------------------------------------  
## ExterCond   
## n missing distinct   
## 1460 0 5   
##   
## Value Ex Fa Gd Po TA  
## Frequency 3 28 146 1 1282  
## Proportion 0.002 0.019 0.100 0.001 0.878  
## ---------------------------------------------------------------------------  
## Foundation   
## n missing distinct   
## 1460 0 6   
##   
## Value BrkTil CBlock PConc Slab Stone Wood  
## Frequency 146 634 647 24 6 3  
## Proportion 0.100 0.434 0.443 0.016 0.004 0.002  
## ---------------------------------------------------------------------------  
## BsmtUnfSF   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 780 0.999 567.2 486.6 0.0 74.9   
## .25 .50 .75 .90 .95   
## 223.0 477.5 808.0 1232.0 1468.0   
##   
## lowest : 0 14 15 23 26, highest: 2042 2046 2121 2153 2336  
## ---------------------------------------------------------------------------  
## TotalBsmtSF   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 721 1 1057 459.5 519.3 636.9   
## .25 .50 .75 .90 .95   
## 795.8 991.5 1298.2 1602.2 1753.0   
##   
## lowest : 0 105 190 264 270, highest: 3094 3138 3200 3206 6110  
## ---------------------------------------------------------------------------  
## Heating   
## n missing distinct   
## 1460 0 6   
##   
## Value Floor GasA GasW Grav OthW Wall  
## Frequency 1 1428 18 7 2 4  
## Proportion 0.001 0.978 0.012 0.005 0.001 0.003  
## ---------------------------------------------------------------------------  
## HeatingQC   
## n missing distinct   
## 1460 0 5   
##   
## Value Ex Fa Gd Po TA  
## Frequency 741 49 241 1 428  
## Proportion 0.508 0.034 0.165 0.001 0.293  
## ---------------------------------------------------------------------------  
## CentralAir   
## n missing distinct   
## 1460 0 2   
##   
## Value N Y  
## Frequency 95 1365  
## Proportion 0.065 0.935  
## ---------------------------------------------------------------------------  
## Electrical   
## n missing distinct   
## 1460 0 5   
##   
## Value FuseA FuseF FuseP Mix SBrkr  
## Frequency 94 27 3 1 1335  
## Proportion 0.064 0.018 0.002 0.001 0.914  
## ---------------------------------------------------------------------------  
## X2ndFlrSF   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 417 0.817 347 450.2 0.0 0.0   
## .25 .50 .75 .90 .95   
## 0.0 0.0 728.0 954.2 1141.0   
##   
## lowest : 0 110 167 192 208, highest: 1611 1796 1818 1872 2065  
## ---------------------------------------------------------------------------  
## LowQualFinSF   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 24 0.052 5.845 11.55 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 0   
##   
## lowest : 0 53 80 120 144, highest: 513 514 515 528 572  
## ---------------------------------------------------------------------------  
## GrLivArea   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 861 1 1515 563.1 848 912   
## .25 .50 .75 .90 .95   
## 1130 1464 1777 2158 2466   
##   
## lowest : 334 438 480 520 605, highest: 3627 4316 4476 4676 5642  
## ---------------------------------------------------------------------------  
## BsmtFullBath   
## n missing distinct Info Mean Gmd   
## 1460 0 4 0.733 0.4253 0.5085   
##   
## Value 0 1 2 3  
## Frequency 856 588 15 1  
## Proportion 0.586 0.403 0.010 0.001  
## ---------------------------------------------------------------------------  
## BsmtHalfBath   
## n missing distinct Info Mean Gmd   
## 1460 0 3 0.159 0.05753 0.1088   
##   
## Value 0 1 2  
## Frequency 1378 80 2  
## Proportion 0.944 0.055 0.001  
## ---------------------------------------------------------------------------  
## FullBath   
## n missing distinct Info Mean Gmd   
## 1460 0 4 0.766 1.565 0.5521   
##   
## Value 0 1 2 3  
## Frequency 9 650 768 33  
## Proportion 0.006 0.445 0.526 0.023  
## ---------------------------------------------------------------------------  
## HalfBath   
## n missing distinct Info Mean Gmd   
## 1460 0 3 0.706 0.3829 0.4852   
##   
## Value 0 1 2  
## Frequency 913 535 12  
## Proportion 0.625 0.366 0.008  
## ---------------------------------------------------------------------------  
## BedroomAbvGr   
## n missing distinct Info Mean Gmd   
## 1460 0 8 0.815 2.866 0.818   
##   
## Value 0 1 2 3 4 5 6 8  
## Frequency 6 50 358 804 213 21 7 1  
## Proportion 0.004 0.034 0.245 0.551 0.146 0.014 0.005 0.001  
## ---------------------------------------------------------------------------  
## KitchenAbvGr   
## n missing distinct Info Mean Gmd   
## 1460 0 4 0.133 1.047 0.09174   
##   
## Value 0 1 2 3  
## Frequency 1 1392 65 2  
## Proportion 0.001 0.953 0.045 0.001  
## ---------------------------------------------------------------------------  
## KitchenQual   
## n missing distinct   
## 1460 0 4   
##   
## Value Ex Fa Gd TA  
## Frequency 100 39 586 735  
## Proportion 0.068 0.027 0.401 0.503  
## ---------------------------------------------------------------------------  
## TotRmsAbvGrd   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 12 0.958 6.518 1.762 4 5   
## .25 .50 .75 .90 .95   
## 5 6 7 9 10   
##   
## Value 2 3 4 5 6 7 8 9 10 11  
## Frequency 1 17 97 275 402 329 187 75 47 18  
## Proportion 0.001 0.012 0.066 0.188 0.275 0.225 0.128 0.051 0.032 0.012  
##   
## Value 12 14  
## Frequency 11 1  
## Proportion 0.008 0.001  
## ---------------------------------------------------------------------------  
## Functional   
## n missing distinct   
## 1460 0 7   
##   
## Value Maj1 Maj2 Min1 Min2 Mod Sev Typ  
## Frequency 14 5 31 34 15 1 1360  
## Proportion 0.010 0.003 0.021 0.023 0.010 0.001 0.932  
## ---------------------------------------------------------------------------  
## Fireplaces   
## n missing distinct Info Mean Gmd   
## 1460 0 4 0.806 0.613 0.6566   
##   
## Value 0 1 2 3  
## Frequency 690 650 115 5  
## Proportion 0.473 0.445 0.079 0.003  
## ---------------------------------------------------------------------------  
## GarageCars   
## n missing distinct Info Mean Gmd   
## 1460 0 5 0.802 1.767 0.7609   
##   
## Value 0 1 2 3 4  
## Frequency 81 369 824 181 5  
## Proportion 0.055 0.253 0.564 0.124 0.003  
## ---------------------------------------------------------------------------  
## GarageArea   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 441 1 473 234.9 0.0 240.0   
## .25 .50 .75 .90 .95   
## 334.5 480.0 576.0 757.1 850.1   
##   
## lowest : 0 160 164 180 186, highest: 1220 1248 1356 1390 1418  
## ---------------------------------------------------------------------------  
## PavedDrive   
## n missing distinct   
## 1460 0 3   
##   
## Value N P Y  
## Frequency 90 30 1340  
## Proportion 0.062 0.021 0.918  
## ---------------------------------------------------------------------------  
## WoodDeckSF   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 274 0.858 94.24 125 0 0   
## .25 .50 .75 .90 .95   
## 0 0 168 262 335   
##   
## lowest : 0 12 24 26 28, highest: 668 670 728 736 857  
## ---------------------------------------------------------------------------  
## OpenPorchSF   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 202 0.909 46.66 62.43 0 0   
## .25 .50 .75 .90 .95   
## 0 25 68 130 175   
##   
## lowest : 0 4 8 10 11, highest: 406 418 502 523 547  
## ---------------------------------------------------------------------------  
## EnclosedPorch   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 120 0.369 21.95 39.39 0.0 0.0   
## .25 .50 .75 .90 .95   
## 0.0 0.0 0.0 112.0 180.1   
##   
## lowest : 0 19 20 24 30, highest: 301 318 330 386 552  
## ---------------------------------------------------------------------------  
## X3SsnPorch   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 20 0.049 3.41 6.739 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 0   
##   
## Value 0 23 96 130 140 144 153 162 168 180  
## Frequency 1436 1 1 1 1 2 1 1 3 2  
## Proportion 0.984 0.001 0.001 0.001 0.001 0.001 0.001 0.001 0.002 0.001  
##   
## Value 182 196 216 238 245 290 304 320 407 508  
## Frequency 1 1 2 1 1 1 1 1 1 1  
## Proportion 0.001 0.001 0.001 0.001 0.001 0.001 0.001 0.001 0.001 0.001  
## ---------------------------------------------------------------------------  
## ScreenPorch   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 76 0.22 15.06 28.27 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 160   
##   
## lowest : 0 40 53 60 63, highest: 385 396 410 440 480  
## ---------------------------------------------------------------------------  
## PoolArea   
## n missing distinct Info Mean Gmd   
## 1460 0 8 0.014 2.759 5.497   
##   
## Value 0 480 512 519 555 576 648 738  
## Frequency 1453 1 1 1 1 1 1 1  
## Proportion 0.995 0.001 0.001 0.001 0.001 0.001 0.001 0.001  
## ---------------------------------------------------------------------------  
## MiscVal   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 21 0.103 43.49 85.67 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 0   
##   
## Value 0 50 350 400 450 500 550 600 700 800  
## Frequency 1408 1 1 11 4 10 1 5 5 1  
## Proportion 0.964 0.001 0.001 0.008 0.003 0.007 0.001 0.003 0.003 0.001  
##   
## Value 1150 1200 1300 1400 2000 2500 3500 8300 15500  
## Frequency 1 2 1 1 4 1 1 1 1  
## Proportion 0.001 0.001 0.001 0.001 0.003 0.001 0.001 0.001 0.001  
## ---------------------------------------------------------------------------  
## MoSold   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 12 0.985 6.322 3.041 2 3   
## .25 .50 .75 .90 .95   
## 5 6 8 10 11   
##   
## Value 1 2 3 4 5 6 7 8 9 10  
## Frequency 58 52 106 141 204 253 234 122 63 89  
## Proportion 0.040 0.036 0.073 0.097 0.140 0.173 0.160 0.084 0.043 0.061  
##   
## Value 11 12  
## Frequency 79 59  
## Proportion 0.054 0.040  
## ---------------------------------------------------------------------------  
## YrSold   
## n missing distinct Info Mean Gmd   
## 1460 0 5 0.955 2008 1.498   
##   
## Value 2006 2007 2008 2009 2010  
## Frequency 314 329 304 338 175  
## Proportion 0.215 0.225 0.208 0.232 0.120  
## ---------------------------------------------------------------------------  
## SaleType   
## n missing distinct   
## 1460 0 9   
##   
## Value COD Con ConLD ConLI ConLw CWD New Oth WD  
## Frequency 43 2 9 5 5 4 122 3 1267  
## Proportion 0.029 0.001 0.006 0.003 0.003 0.003 0.084 0.002 0.868  
## ---------------------------------------------------------------------------  
## SaleCondition   
## n missing distinct   
## 1460 0 6   
##   
## Value Abnorml AdjLand Alloca Family Normal Partial  
## Frequency 101 4 12 20 1198 125  
## Proportion 0.069 0.003 0.008 0.014 0.821 0.086  
## ---------------------------------------------------------------------------  
## SalePrice   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 663 1 180921 81086 88000 106475   
## .25 .50 .75 .90 .95   
## 129975 163000 214000 278000 326100   
##   
## lowest : 34900 35311 37900 39300 40000, highest: 582933 611657 625000 745000 755000  
## ---------------------------------------------------------------------------  
## Age   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 122 0.999 36.55 33.96 1 1   
## .25 .50 .75 .90 .95   
## 8 35 54 84 91   
##   
## lowest : 0 1 2 3 4, highest: 127 128 129 135 136  
## ---------------------------------------------------------------------------

boxplot(training$SalePrice / 1000 )



cat\_var <- names(training)[which(sapply(training, is.factor))]  
cat\_var

## [1] "MSZoning" "Street" "LotShape" "LandContour"   
## [5] "Utilities" "LotConfig" "LandSlope" "Neighborhood"   
## [9] "Condition1" "Condition2" "BldgType" "HouseStyle"   
## [13] "RoofStyle" "RoofMatl" "Exterior1st" "Exterior2nd"   
## [17] "ExterQual" "ExterCond" "Foundation" "Heating"   
## [21] "HeatingQC" "CentralAir" "Electrical" "KitchenQual"   
## [25] "Functional" "PavedDrive" "SaleType" "SaleCondition"

num\_var <- c('SalePrice','LotArea','TotalBsmtSF','GrLivArea','GarageArea','Age','WoodDeckSF','OpenPorchSF','PoolArea')  
training\_pca<-training[,num\_var]  
training\_pca<-training\_pca[,-1]  
head(training\_pca)

## LotArea TotalBsmtSF GrLivArea GarageArea Age WoodDeckSF OpenPorchSF  
## 1 8450 856 1710 548 5 0 61  
## 2 9600 1262 1262 460 31 298 0  
## 3 11250 920 1786 608 7 0 42  
## 4 9550 756 1717 642 91 0 35  
## 5 14260 1145 2198 836 8 192 84  
## 6 14115 796 1362 480 16 40 30  
## PoolArea  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

library(stats)  
library(factoextra)

## Warning: package 'factoextra' was built under R version 3.5.2

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

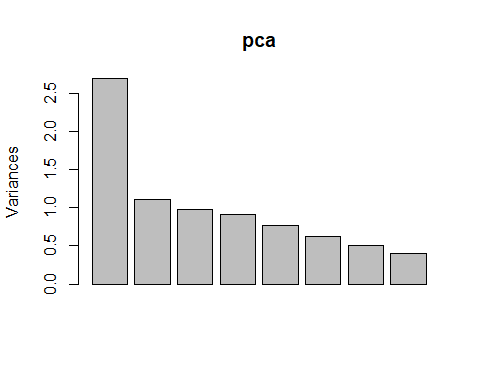
pca <- prcomp(training\_pca, scale. = T, center = T)  
pca

## Standard deviations (1, .., p=8):  
## [1] 1.6418166 1.0507294 0.9868895 0.9566831 0.8790785 0.7914566 0.7142146  
## [8] 0.6339774  
##   
## Rotation (n x k) = (8 x 8):  
## PC1 PC2 PC3 PC4 PC5  
## LotArea 0.2418914 -0.53667137 0.33118509 -0.50019393 0.25022784  
## TotalBsmtSF 0.4643873 0.02001449 0.01588152 -0.01512856 0.27800523  
## GrLivArea 0.4442693 -0.20795746 -0.15580977 -0.13480990 -0.09162792  
## GarageArea 0.4673895 0.22979967 0.01392001 0.02602217 0.25292157  
## Age -0.3637131 -0.52270812 -0.08856449 -0.27421668 -0.16430841  
## WoodDeckSF 0.2783972 -0.09067257 0.59834342 0.25762648 -0.68773801  
## OpenPorchSF 0.2881744 0.05557345 -0.62974978 -0.30531710 -0.52339677  
## PoolArea 0.1310271 -0.57532197 -0.32117044 0.70423120 0.11751223  
## PC6 PC7 PC8  
## LotArea -0.441528308 -0.17986725 -0.06067467  
## TotalBsmtSF 0.059861391 0.82388749 0.15427861  
## GrLivArea 0.619087116 -0.15618376 -0.54851639  
## GarageArea 0.213838887 -0.48550484 0.61843098  
## Age 0.460170910 0.14230075 0.50244056  
## WoodDeckSF -0.006370446 0.04844563 0.12048271  
## OpenPorchSF -0.360004424 0.01893306 0.14197874  
## PoolArea -0.176273669 -0.07611783 0.04548668

eigenvalues <- get\_eigenvalue(pca)  
eigenvalues <- pca$sdev^2  
sum(eigenvalues)

## [1] 8

plot(pca)



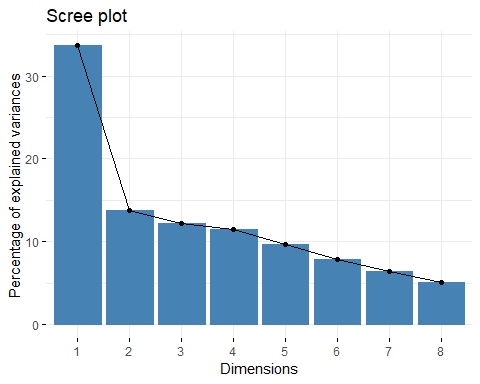
summary(pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.642 1.0507 0.9869 0.9567 0.8791 0.7915 0.71421  
## Proportion of Variance 0.337 0.1380 0.1217 0.1144 0.0966 0.0783 0.06376  
## Cumulative Proportion 0.337 0.4749 0.5967 0.7111 0.8077 0.8860 0.94976  
## PC8  
## Standard deviation 0.63398  
## Proportion of Variance 0.05024  
## Cumulative Proportion 1.00000

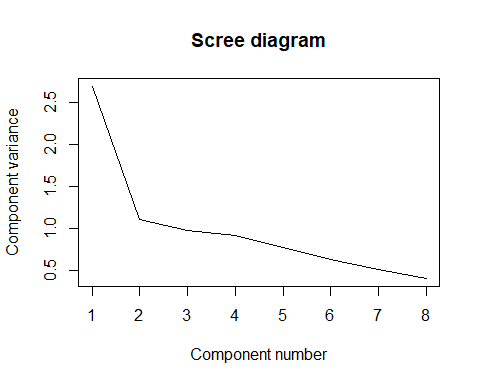
head(pca$x)

## PC1 PC2 PC3 PC4 PC5 PC6  
## [1,] 0.2885164 0.7704301 -0.6004514 0.04757420 0.4424969 -0.172769650  
## [2,] 0.2589971 0.0938194 1.5057769 0.73809412 -0.5918576 -0.072244372  
## [3,] 0.5128518 0.6067160 -0.3491214 -0.03772016 0.7502032 -0.004687438  
## [4,] -0.7263789 -0.7028615 -0.5681891 -0.65423267 0.2549489 1.316707774  
## [5,] 2.2678310 0.4162006 0.1660223 -0.08222949 -0.2246076 0.383537683  
## [6,] -0.2588210 0.2827174 0.1374989 -0.02789137 0.4923935 -0.575813356  
## PC7 PC8  
## [1,] -0.7447214 -0.63132618  
## [2,] 0.5500647 0.30514579  
## [3,] -0.8298439 -0.53911747  
## [4,] -0.7707521 0.96411258  
## [5,] -1.0108129 0.04231143  
## [6,] -0.6432933 -0.36557181

library(factoextra)  
fviz\_screeplot(pca, ncp = 35)



#plot(pca, type = "l", main = "Scree diagram")  
plot(eigenvalues, xlab = "Component number", ylab = "Component variance", type = "l", main ="Scree diagram")

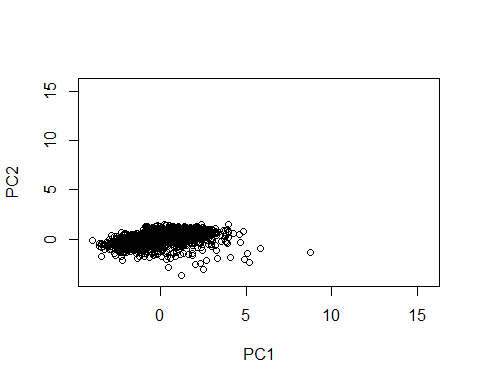


diag(cov(pca$x))

## PC1 PC2 PC3 PC4 PC5 PC6 PC7   
## 2.6955617 1.1040324 0.9739509 0.9152426 0.7727790 0.6264036 0.5101026   
## PC8   
## 0.4019273

xlim <- range(pca$x[,1])

plot(pca$x,xlim=xlim,ylim=xlim)



pca$rotation[,1]

## LotArea TotalBsmtSF GrLivArea GarageArea Age WoodDeckSF   
## 0.2418914 0.4643873 0.4442693 0.4673895 -0.3637131 0.2783972   
## OpenPorchSF PoolArea   
## 0.2881744 0.1310271

pca$rotation[,2]

## LotArea TotalBsmtSF GrLivArea GarageArea Age WoodDeckSF   
## -0.53667137 0.02001449 -0.20795746 0.22979967 -0.52270812 -0.09067257   
## OpenPorchSF PoolArea   
## 0.05557345 -0.57532197

pca$rotation[,3]

## LotArea TotalBsmtSF GrLivArea GarageArea Age WoodDeckSF   
## 0.33118509 0.01588152 -0.15580977 0.01392001 -0.08856449 0.59834342   
## OpenPorchSF PoolArea   
## -0.62974978 -0.32117044

pca$rotation[,4]

## LotArea TotalBsmtSF GrLivArea GarageArea Age WoodDeckSF   
## -0.50019393 -0.01512856 -0.13480990 0.02602217 -0.27421668 0.25762648   
## OpenPorchSF PoolArea   
## -0.30531710 0.70423120

pca$rotation[,5]

## LotArea TotalBsmtSF GrLivArea GarageArea Age WoodDeckSF   
## 0.25022784 0.27800523 -0.09162792 0.25292157 -0.16430841 -0.68773801   
## OpenPorchSF PoolArea   
## -0.52339677 0.11751223

training\_fca<-training[,num\_var]  
training\_fca<-training\_fca[,-1]  
library(psych)

## Warning: package 'psych' was built under R version 3.5.3

##   
## Attaching package: 'psych'

## The following object is masked from 'package:Hmisc':  
##   
## describe

## The following objects are masked from 'package:scales':  
##   
## alpha, rescale

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

fit.pc <- principal(training\_fca, nfactors=3, rotate="varimax")  
fit.pc

## Principal Components Analysis  
## Call: principal(r = training\_fca, nfactors = 3, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC1 RC2 RC3 h2 u2 com  
## LotArea 0.06 0.40 0.65 0.58 0.42 1.7  
## TotalBsmtSF 0.68 0.22 0.26 0.58 0.42 1.5  
## GrLivArea 0.57 0.49 0.21 0.60 0.40 2.2  
## GarageArea 0.78 0.05 0.17 0.65 0.35 1.1  
## Age -0.76 0.29 -0.05 0.67 0.33 1.3  
## WoodDeckSF 0.28 -0.09 0.69 0.57 0.43 1.4  
## OpenPorchSF 0.53 0.43 -0.39 0.61 0.39 2.8  
## PoolArea -0.04 0.71 0.06 0.51 0.49 1.0  
##   
## RC1 RC2 RC3  
## SS loadings 2.35 1.23 1.20  
## Proportion Var 0.29 0.15 0.15  
## Cumulative Var 0.29 0.45 0.60  
## Proportion Explained 0.49 0.26 0.25  
## Cumulative Proportion 0.49 0.75 1.00  
##   
## Mean item complexity = 1.6  
## Test of the hypothesis that 3 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.13   
## with the empirical chi square 1360.13 with prob < 1.6e-289   
##   
## Fit based upon off diagonal values = 0.75

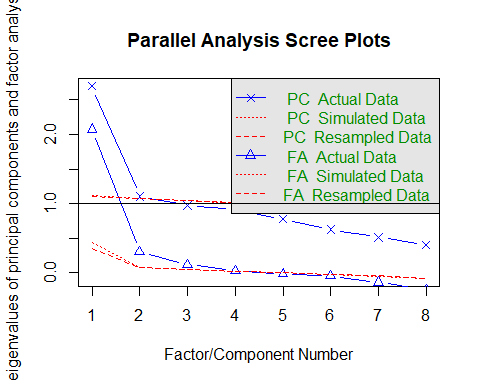
round(fit.pc$values, 3)

## [1] 2.696 1.104 0.974 0.915 0.773 0.626 0.510 0.402

fit.pc$loadings

##   
## Loadings:  
## RC1 RC2 RC3   
## LotArea 0.402 0.646  
## TotalBsmtSF 0.682 0.218 0.264  
## GrLivArea 0.569 0.485 0.210  
## GarageArea 0.785 0.172  
## Age -0.761 0.290   
## WoodDeckSF 0.282 0.693  
## OpenPorchSF 0.529 0.429 -0.387  
## PoolArea 0.712   
##   
## RC1 RC2 RC3  
## SS loadings 2.348 1.230 1.196  
## Proportion Var 0.293 0.154 0.150  
## Cumulative Var 0.293 0.447 0.597

fa.parallel(training\_fca)



## Parallel analysis suggests that the number of factors = 3 and the number of components = 2

fa.diagram(fit.pc) # Visualize the relationship

