```
rm(list = ls())
```

```
# libraries
suppressMessages(library(tidyverse))
library(broom)
```

```
Data <- read.csv("/Users/mpaga/Downloads/train.csv",sep=",")
```

```
dim(Data)
```

```
[1] 1460   81
```

```
names(Data)
```

```
 [1] "Id"            "MSSubClass"    "MSZoning"      "LotFrontage"
 [5] "LotArea"       "Street"        "Alley"         "LotShape"
 [9] "LandContour"   "Utilities"     "LotConfig"     "LandSlope"
[13] "Neighborhood"  "Condition1"    "Condition2"    "BldgType"
[17] "HouseStyle"    "OverallQual"   "OverallCond"   "YearBuilt"
[21] "YearRemodAdd"  "RoofStyle"     "RoofMatl"      "Exterior1st"
[25] "Exterior2nd"   "MasVnrType"    "MasVnrArea"    "ExterQual"
[29] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"
[33] "BsmtExposure"  "BsmtFinType1"  "BsmtFinSF1"    "BsmtFinType2"
[37] "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"   "Heating"
[41] "HeatingQC"     "CentralAir"    "Electrical"    "X1stFlrSF"
[45] "X2ndFlrSF"     "LowQualFinSF"  "GrLivArea"     "BsmtFullBath"
[49] "BsmtHalfBath"  "FullBath"      "HalfBath"      "BedroomAbvGr"
[53] "KitchenAbvGr"  "KitchenQual"   "TotRmsAbvGrd"  "Functional"
[57] "Fireplaces"    "FireplaceQu"   "GarageType"    "GarageYrBlt"
[61] "GarageFinish"  "GarageCars"    "GarageArea"    "GarageQual"
[65] "GarageCond"    "PavedDrive"    "WoodDeckSF"    "OpenPorchSF"
[69] "EnclosedPorch" "X3SsnPorch"    "ScreenPorch"   "PoolArea"
[73] "PoolQC"        "Fence"         "MiscFeature"   "MiscVal"
[77] "MoSold"        "YrSold"        "SaleType"      "SaleCondition"
[81] "SalePrice"
```

```
#remove Id colum
Data["Id"] <- NULL
dim(Data)
```

```
[1] 1460   80
```

```
hist(Data$SalePrice)
```

```
summary(Data$SalePrice)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34900  129975  163000  180921  214000  755000
```

```
skimr::skim_without_charts(Data,where(is.numeric))-> num_skimData
```

```
(numVal_na <- num_skimData$skim_variable[num_skimData$n_missing !=0] )
```

```
[1] "LotFrontage" "MasVnrArea"  "GarageYrBlt"
```

```
mean(is.na(Data$LotFrontage))
```

```
[1] 0.1773973
```

```
mean(is.na(Data$MasVnrType))
```

```
[1] 0.005479452
```

```
mean(is.na(Data$GarageYrBlt))
```

```
[1] 0.05547945
```

```
skimr::skim_without_charts(Data,where(is.character))-> cat_skimData
```

```
summary(cat_skimData)
```

```
  Data Summary
                           Values
Name                       Data
Number of rows             1460
Number of columns          80
-----------------------
Column type frequency:
  character                43
-----------------------
Group variables            None
```

```
# missing values in cat features
cat_skimData$skim_variable[cat_skimData$n_missing !=0]
```

```
 [1] "Alley"        "MasVnrType"   "BsmtQual"     "BsmtCond"
 [5] "BsmtExposure" "BsmtFinType1" "BsmtFinType2" "Electrical"
 [9] "FireplaceQu"  "GarageType"   "GarageFinish" "GarageQual"
[13] "GarageCond"   "PoolQC"       "Fence"        "MiscFeature"
```

Let's impute some of these numerical features

```
numVal_na
```

```
[1] "LotFrontage" "MasVnrArea"  "GarageYrBlt"
```

```
colMeans(is.na(Data[numVal_na]))
```

```
LotFrontage  MasVnrArea GarageYrBlt
0.177397260 0.005479452 0.055479452
```

```
# list of num featute to impute
imputeVal_list <- apply(Data[numVal_na],2, FUN = "median",na.rm = T,simplify = list)
```

```
#impute numerical features
Data[numVal_na] <- replace_na(Data[numVal_na] ,
                              replace = imputeVal_list
)
```

```
# check  na
colMeans(is.na(Data[numVal_na]))
```

```
LotFrontage  MasVnrArea GarageYrBlt
          0           0           0
```

```
Data |>
  select_if(is.numeric) |>
  unique() |>
  dim()
```

```
[1] 1460   37
```

```
Data |>
  select_if(is.numeric)  ->numData
  lm(SalePrice~.,numData) |>
    summary() |>
    tidy() |>
    filter(p.value <=0.5) |>
    nrow()
```

```
[1] 28
```

28 numerical features have predictive effect on target feature.

```
# correlation
for (feature in names(numData)[-1]){
  if (cor(numData[names(numData)][1],numData[feature]) >= 0.8 ) print(feature)
}
```

There is no carrelated features in numeric features

cat var feat engineering

```
# n cat features presenting missing values
cat_skimData$skim_variable[cat_skimData$n_missing  !=0] |> length()
```

```
[1] 16
```

```
#duplicated rows in numData
nrow(unique(numData)) != dim(numData)[1]
```

```
[1] FALSE
```

16 features have missing values or NA

let's use random forest to predict missing values

to be continued !