EDA

Ruoyuan Qian

3/3/2020

Data cleaning

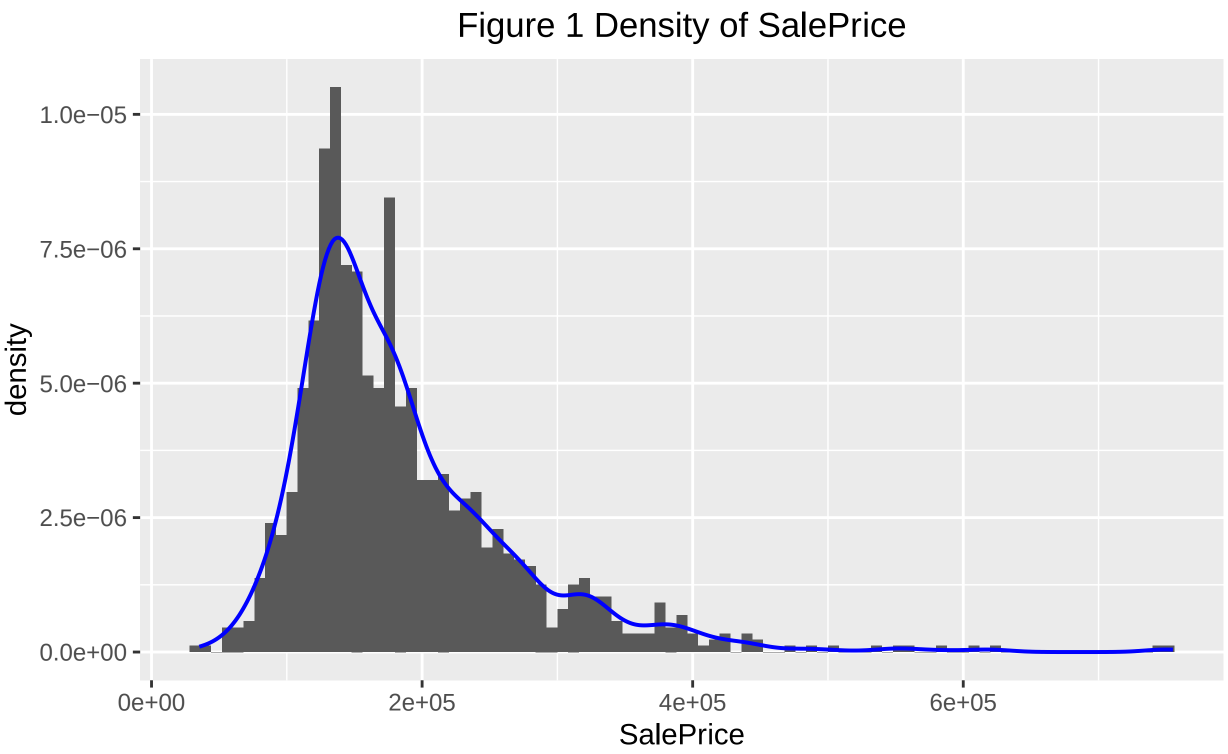
```
house = read.csv(file = "train.csv")
#skimr::skim(house)

sum_na = function(x){
  sum = sum(is.na(x))
  sum}

# names of predictor when its missing value larger than 500
missing_var = map(house,sum_na) %>%
  as.data.frame() %>%
  pivot_longer(
    Id : SalePrice,
    names_to = "variable",
    values_to = "value"
  ) %>%
  filter(value > 500 ) %>%
  pull(variable)

#house %>%
#select(-Alley,-FireplaceQu,-PoolQC,-Fence,-MiscFeature) %>%
#map(.,sum_na)
# names of variables when its value nears zero
near_0_var =
  house %>%
  nearZeroVar( names = TRUE)

final_house =
  house %>%
  #nearZeroVar( names = TRUE)
  select(-near_0_var,-missing_var,-Id) %>%
  #select(-Alley,-FireplaceQu,-PoolQC,-Fence,-MiscFeature) %>%
  drop_na()
```

Visualization

The response `SalePrice` is right skewed

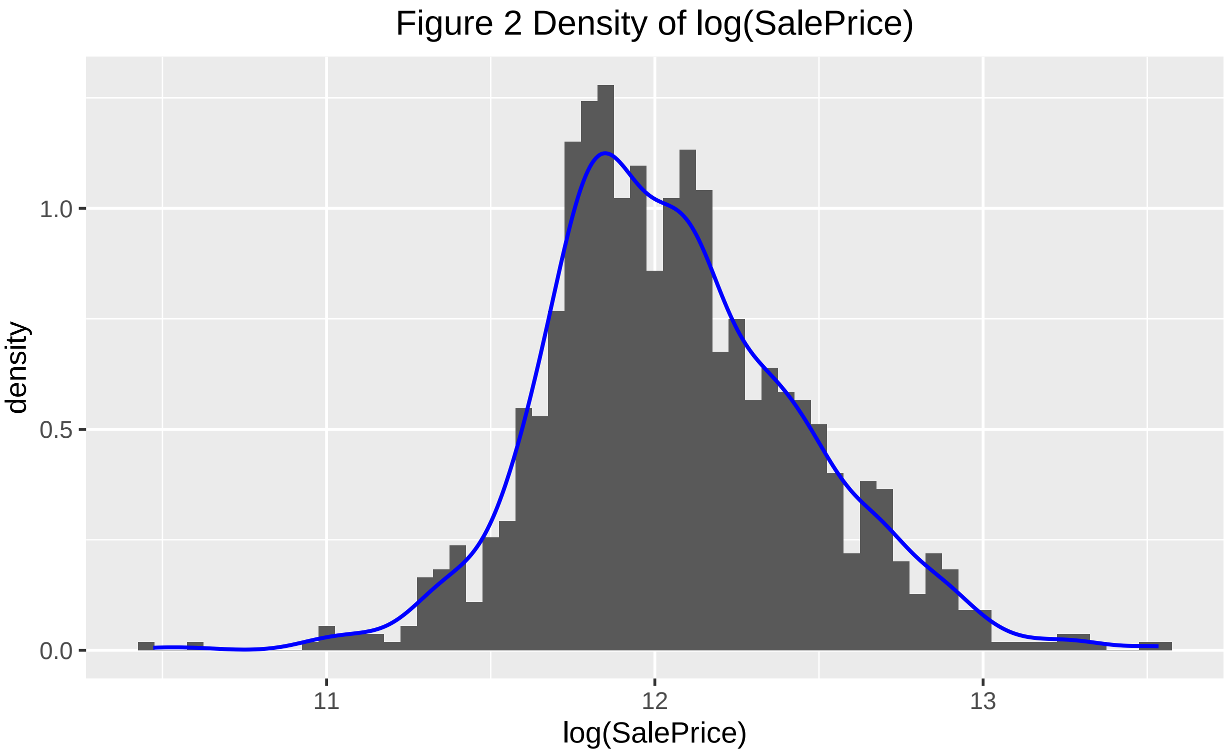
```
density_sale =
ggplot(final_house, aes(x = SalePrice, ..density..)) +
  geom_histogram(binwidth = 8000) +
  geom_line(stat = 'density',size = 0.7,color = "blue")+
  ggtitle("Figure 1 Density of SalePrice") +
  #ylab("Houses") +
  xlab("SalePrice") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
density_sale
```

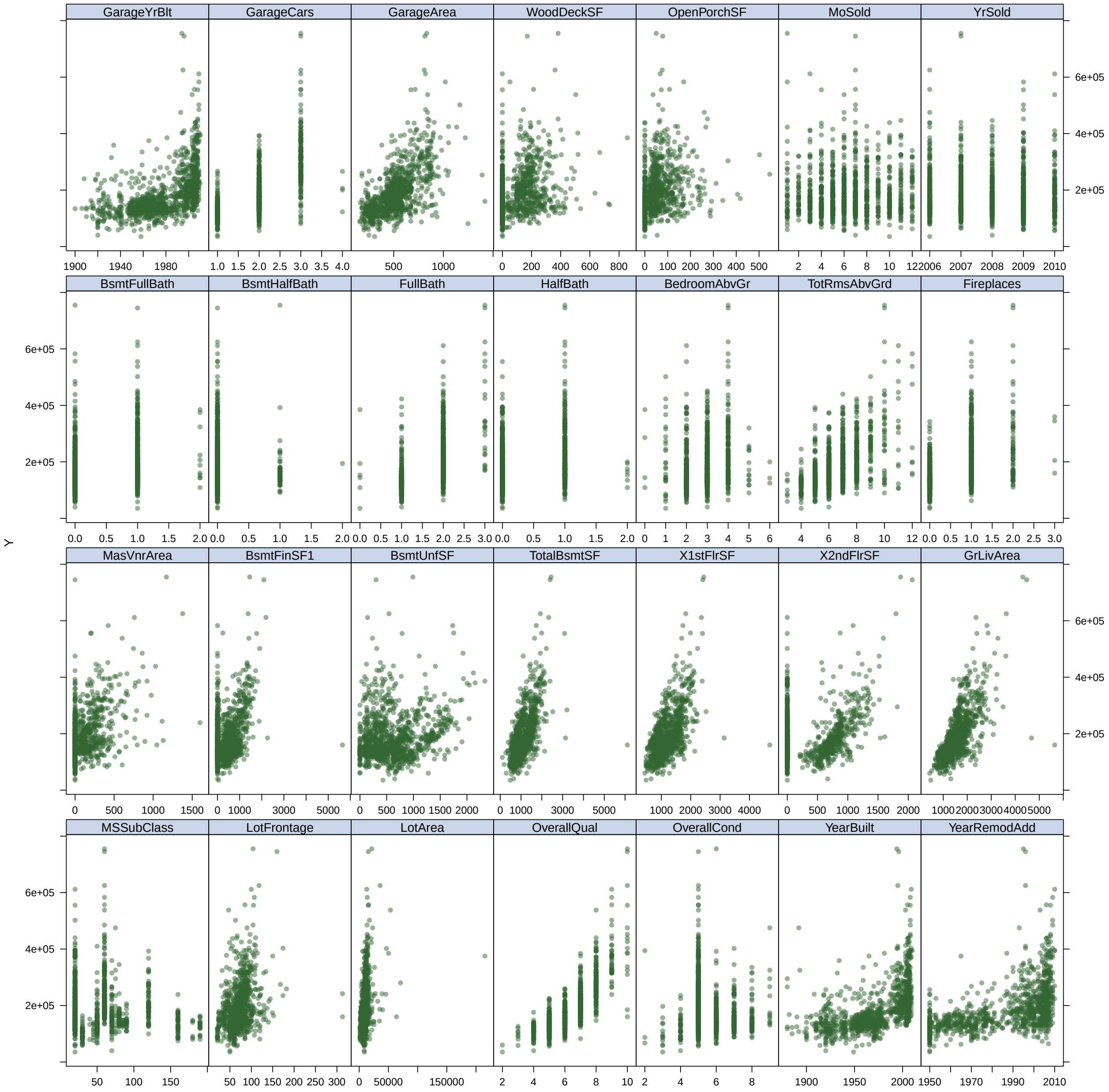


The response `log(SalePrice)` is bell-shaped

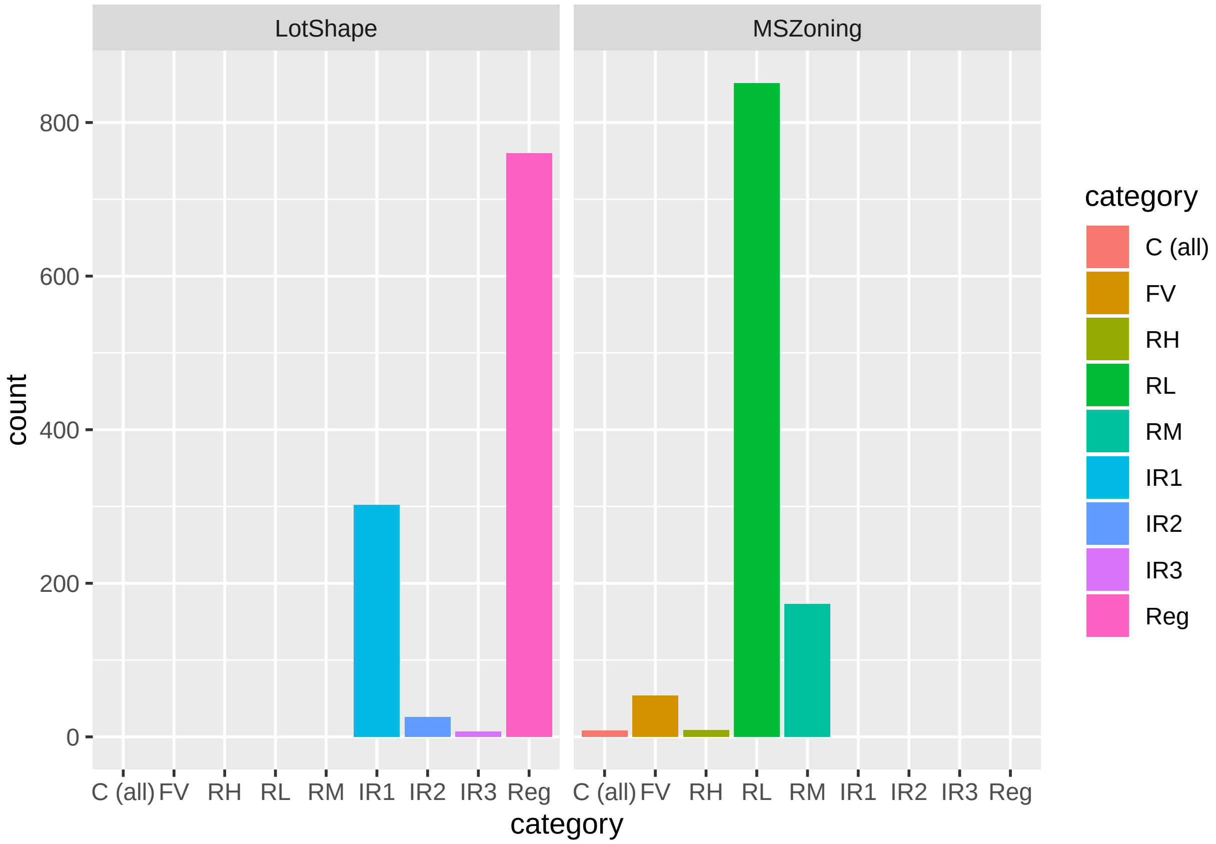
```
density_log_sale =
final_house %>%
  mutate(log_SalePrice = log(SalePrice)) %>%
  ggplot(aes(x = log_SalePrice, ..density..)) +
  geom_histogram(binwidth = 0.05) +
  geom_line(stat = 'density',size = 0.7,color = "blue")+
  ggtitle("Figure 2 Density of log(SalePrice)") +
  #ylab("Houses") +
  xlab("log(SalePrice)") +
  theme(plot.title = element_text(hjust = 0.5))
density_log_sale
```



```
numeric_var_index =
 final_house%>%
  map(.,is.numeric) %>%
  unlist() %>%
  as.vector()

x <- model.matrix(SalePrice~.,
                  final_house[,which(numeric_var_index == TRUE)])[,-1]
y <- final_house$SalePrice


theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <- rgb(.8, .1, .1, 1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(.0, .2, .6, .2)
trellis.par.set(theme1)
featurePlot(x, y, plot = "scatter", labels = c("","Y"),
            type = c("p"), layout = c(7, 4))
```



```
final_house %>%
  select(which(numeric_var_index == FALSE)) %>%
  pivot_longer(
   MSZoning : SaleCondition,
   names_to = "variable",
   values_to = "category"
  ) %>%
  filter(variable %in% c("MSZoning","LotShape")) %>%
  ggplot( mapping = aes(x = category,
                        fill = category))+
  geom_bar() +
  facet_grid(~variable)
```



```
variable_name = names(final_house %>%
  select(which(numeric_var_index == FALSE)))

dataframe =
  final_house %>%
  select(which(numeric_var_index == FALSE))

summary =
final_house %>%
  select(which(numeric_var_index == FALSE)) %>%
  pivot_longer(
   everything(),
   names_to = "variable",
   values_to = "category"
  ) %>%
  group_by(variable,category)%>%
  count() %>%
  mutate(n = freq) %>%
  select(-freq)


plotss = NULL

  for(i in 1:length(variable_name)){
plot_i =
  summary %>%
  filter(variable == variable_name[i]) %>%
  ggplot(mapping = aes(x = category,
                       y = n,fill = category)) +
   geom_bar(stat = 'identity',position = 'dodge') +
 scale_fill_hue(c = 80)+
 ggtitle(paste("Figure", i+2 ,"Distribution of",variable_name[i]))+
  labs(x = variable_name[i]) +
 theme(plot.title = element_text(hjust = 0.5),
       legend.position="right")
    #plots = paste(plots,plot_tem,"+")
#paste("plot",i) <- plot_i
#plotss = c(plotss,print(plot_i))

  }
```

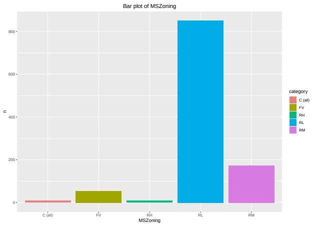
```
plots = function(dataframe){
variable_name = names(dataframe)

summary =
final_house %>%
  select(which(numeric_var_index == FALSE)) %>%
  pivot_longer(
   everything(),
   names_to = "variable",
   values_to = "category"
  ) %>%
  group_by(variable,category)%>%
  count() %>%
  mutate(n = freq) %>%
  select(-freq)

plot_tem =
  summary %>%
  filter(variable == variable_name) %>%
  ggplot(mapping = aes(x = category,
                       y = n,fill = category)) +
   geom_bar(stat = 'identity',position = 'dodge') +
 scale_fill_hue(c = 80)+
 ggtitle(paste("Bar plot of",variable_name))+
  labs(x = variable_name) +
 theme(plot.title = element_text(hjust = 0.5),
       legend.position="right")
    #plots = paste(plots,plot_tem,"+")
plot_tem
  }


plot_name = NULL
for(i in 1: length(dataframe)){
  plot_name_tem = paste("plots(dataframe %>% select(",i,"))",",")
  plot_name = c(plot_name,plot_name_tem)
}
```

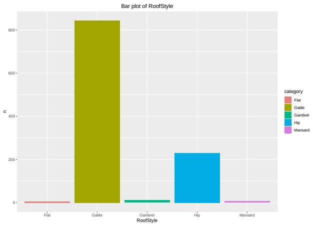
```
as.factor(plot_name)
```

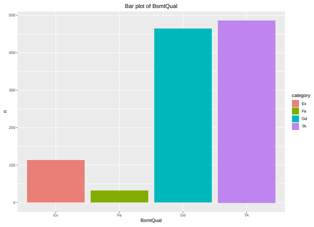
```
##  [1] plots(dataframe %>% select( 1 )) ,
##  [2] plots(dataframe %>% select( 2 )) ,
##  [3] plots(dataframe %>% select( 3 )) ,
##  [4] plots(dataframe %>% select( 4 )) ,
##  [5] plots(dataframe %>% select( 5 )) ,
##  [6] plots(dataframe %>% select( 6 )) ,
##  [7] plots(dataframe %>% select( 7 )) ,
##  [8] plots(dataframe %>% select( 8 )) ,
##  [9] plots(dataframe %>% select( 9 )) ,
## [10] plots(dataframe %>% select( 10 )) ,
## [11] plots(dataframe %>% select( 11 )) ,
## [12] plots(dataframe %>% select( 12 )) ,
## [13] plots(dataframe %>% select( 13 )) ,
## [14] plots(dataframe %>% select( 14 )) ,
## [15] plots(dataframe %>% select( 15 )) ,
## [16] plots(dataframe %>% select( 16 )) ,
## [17] plots(dataframe %>% select( 17 )) ,
## [18] plots(dataframe %>% select( 18 )) ,
## [19] plots(dataframe %>% select( 19 )) ,
## [20] plots(dataframe %>% select( 20 )) ,
## [21] plots(dataframe %>% select( 21 )) ,
## [22] plots(dataframe %>% select( 22 )) ,
## [23] plots(dataframe %>% select( 23 )) ,
## [24] plots(dataframe %>% select( 24 )) ,
## [25] plots(dataframe %>% select( 25 )) ,
## [26] plots(dataframe %>% select( 26 )) ,
## 26 Levels: plots(dataframe %>% select( 1 )) , ...
```

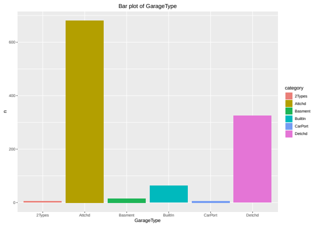
```
multiplot(
 plots(dataframe %>% select( 1 )) ,
plots(dataframe %>% select( 2 )) ,
plots(dataframe %>% select( 3 )) ,
plots(dataframe %>% select( 4 )) ,
plots(dataframe %>% select( 5 )) ,
plots(dataframe %>% select( 6 )) ,
plots(dataframe %>% select( 7 )) ,
plots(dataframe %>% select( 8 )) ,
plots(dataframe %>% select( 9 )) ,
plots(dataframe %>% select( 10 )) ,
plots(dataframe %>% select( 11 )) ,
plots(dataframe %>% select( 12 )) ,
plots(dataframe %>% select( 13 )) ,
plots(dataframe %>% select( 14 )) ,
plots(dataframe %>% select( 15 )) ,
plots(dataframe %>% select( 16 )) ,
plots(dataframe %>% select( 17 )) ,
plots(dataframe %>% select( 18 )) ,
plots(dataframe %>% select( 19 )) ,
plots(dataframe %>% select( 20 )) ,
plots(dataframe %>% select( 21 )) ,
plots(dataframe %>% select( 22 )) ,
```

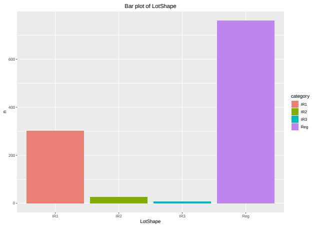
```r
plots(dataframe %>% select( 23 )) ,
plots(dataframe %>% select( 24 )) ,
plots(dataframe %>% select( 25 )) ,
plots(dataframe %>% select( 26 )) ,
          cols=4)
```

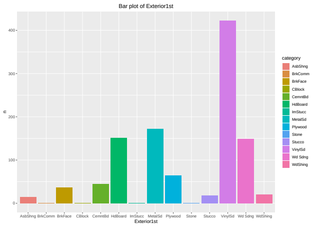

Bar plot of MSZoning
n
0
200
400
600
800
C (all)
FV
RH
RL
RM
MSZoning
category

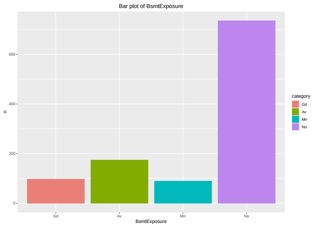

Bar plot of RoofStyle
n
0
200
400
600
800
Flat
Gable
Gambrel
Hip
Mansard
RoofStyle
category

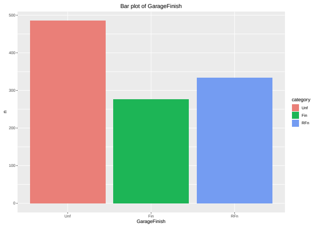

Bar plot of BsmtQual
n
0
100
200
300
400
500
Ex
Fa
Gd
TA
BsmtQual
category

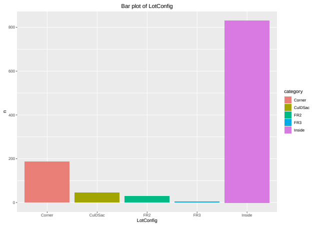

Bar plot of GarageType
n
0
200
400
600
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd
GarageType
category

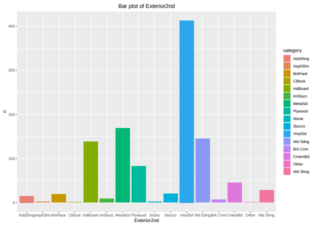

Bar plot of LotShape
n
0
200
400
600
IR1
IR2
IR3
Reg
LotShape
category

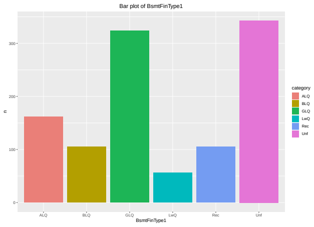

Bar plot of Exterior1st
n
0
100
200
300
400
AsbShng
BrkComm
BrkFace
CBlock
CemntBd
HdBoard
ImStucc
MetalSd
Plywood
Stone
Stucco
VinylSd
Wd Sdng
WdShing
Exterior1st
category

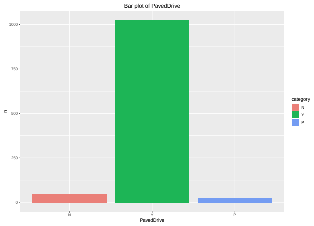

Bar plot of BsmtExposure
n
0
200
400
600
Gd
Av
Mn
No
BsmtExposure
category

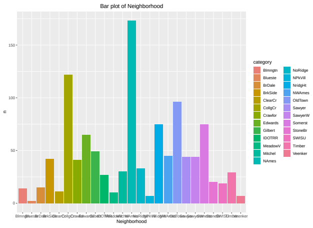

Bar plot of GarageFinish
n
0
100
200
300
400
500
Unf
Fin
RFn
GarageFinish
category

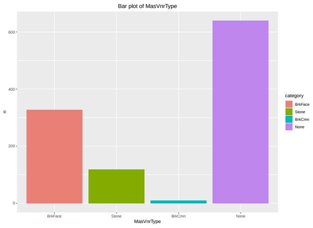

Bar plot of LotConfig
n
0
200
400
600
800
Corner
CulDSac
FR2
FR3
Inside
LotConfig
category

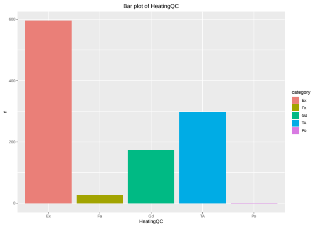

Bar plot of Exterior2nd
n
0
100
200
300
400
AsbShng
AsphShn
BrkFace
CBlock
HdBoard
ImStucc
MetalSd
Plywood
Stone
Stucco
VinylSd
Wd Sdng
Brk Cmn
CmentBd
Other
Wd Shng
Exterior2nd
category

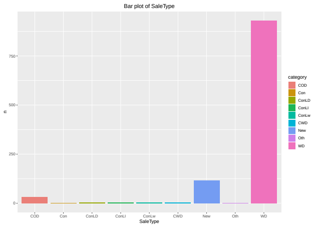

Bar plot of BsmtFinType1
n
0
100
200
300
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinType1
category

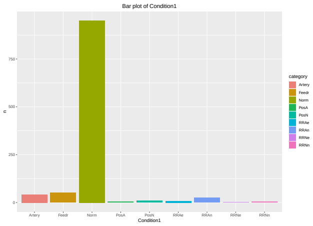

Bar plot of PavedDrive
n
0
250
500
750
1000
N
Y
P
PavedDrive
category

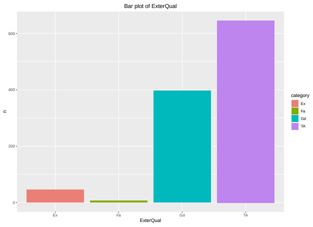

Bar plot of Neighborhood
n
0
50
100
150
Blmngtn
Blueste
BrDale
BrkSide
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
MeadowV
Mitchel
NAmes
NoRidge
NPkVill
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU
Timber
Veenker
Neighborhood
category

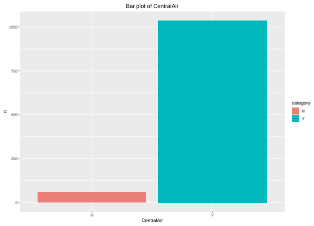

Bar plot of MasVnrType
n
0
200
400
600
BrkFace
Stone
BrkCmn
None
MasVnrType
category

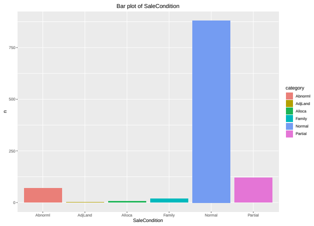

Bar plot of HeatingQC
n
0
200
400
600
Ex
Fa
Gd
TA
Po
HeatingQC
category

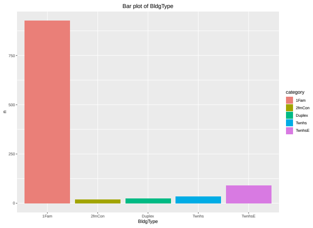

Bar plot of SaleType
n
0
250
500
750
COD
Con
ConLD
ConLI
ConLw
CWD
New
Oth
WD
SaleType
category

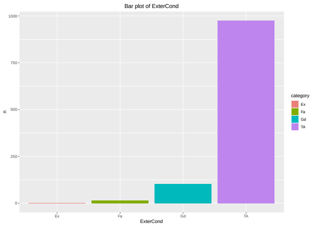

Bar plot of Condition1
n
0
250
500
750
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNe
RRNn
Condition1
category

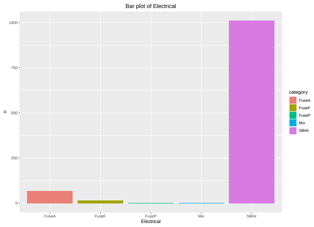

Bar plot of ExterQual
n
0
200
400
600
Ex
Fa
Gd
TA
ExterQual
category

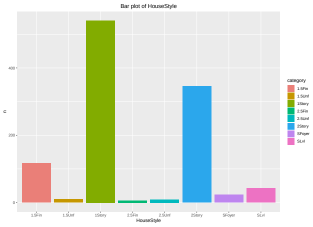

Bar plot of CentralAir
n
0
250
500
750
1000
N
Y
CentralAir
category

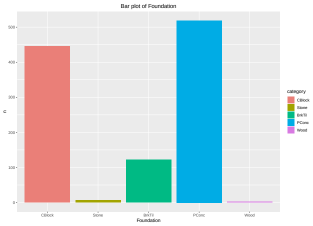

Bar plot of SaleCondition
n
0
250
500
750
Abnorml
AdjLand
Alloca
Family
Normal
Partial
SaleCondition
category

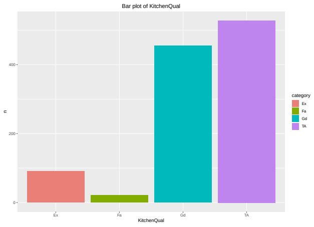

Bar plot of BldgType
n
0
250
500
750
1Fam
2fmCon
Duplex
Twnhs
TwnhsE
BldgType
category


Bar plot of ExterCond
n
0
250
500
750
1000
Ex
Fa
Gd
TA
ExterCond
category


Bar plot of Electrical
n
0
250
500
750
1000
FuseA
FuseF
FuseP
Mix
SBrkr
Electrical
category


Bar plot of HouseStyle
n
0
200
400
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle
category


Bar plot of Foundation
n
0
100
200
300
400
500
CBlock
Stone
BrkTil
PConc
Wood
Foundation
category


Bar plot of KitchenQual
n
0
200
400
Ex
Fa
Gd
TA
KitchenQual
category