

# p8106\_mtp\_rq2166

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()
```

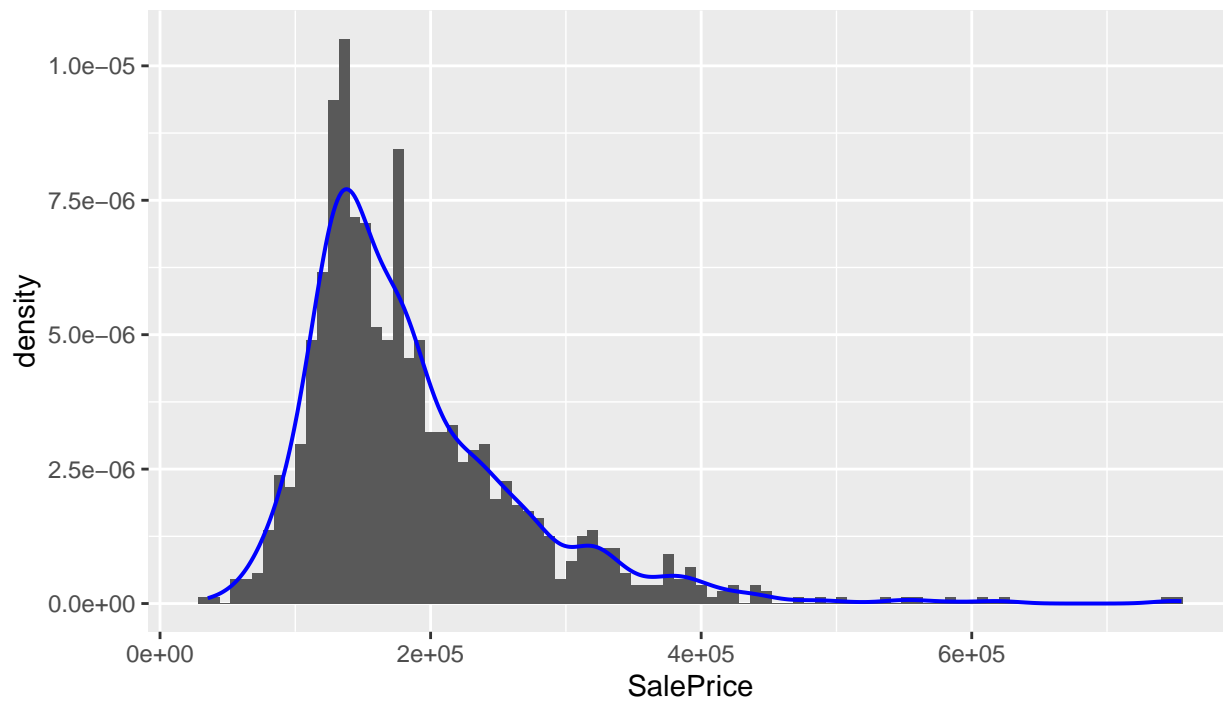
## Visualization

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
```

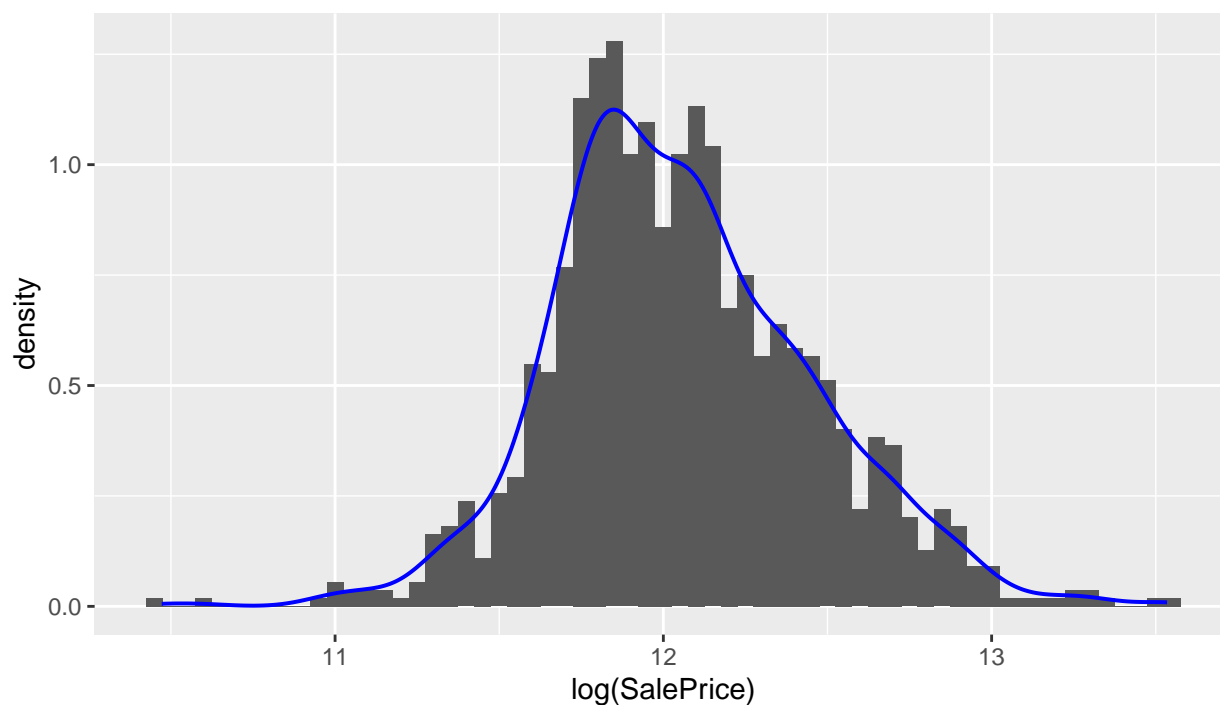
Figure 1 Density of SalePrice



The response  $\log(\text{SalePrice})$  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
```

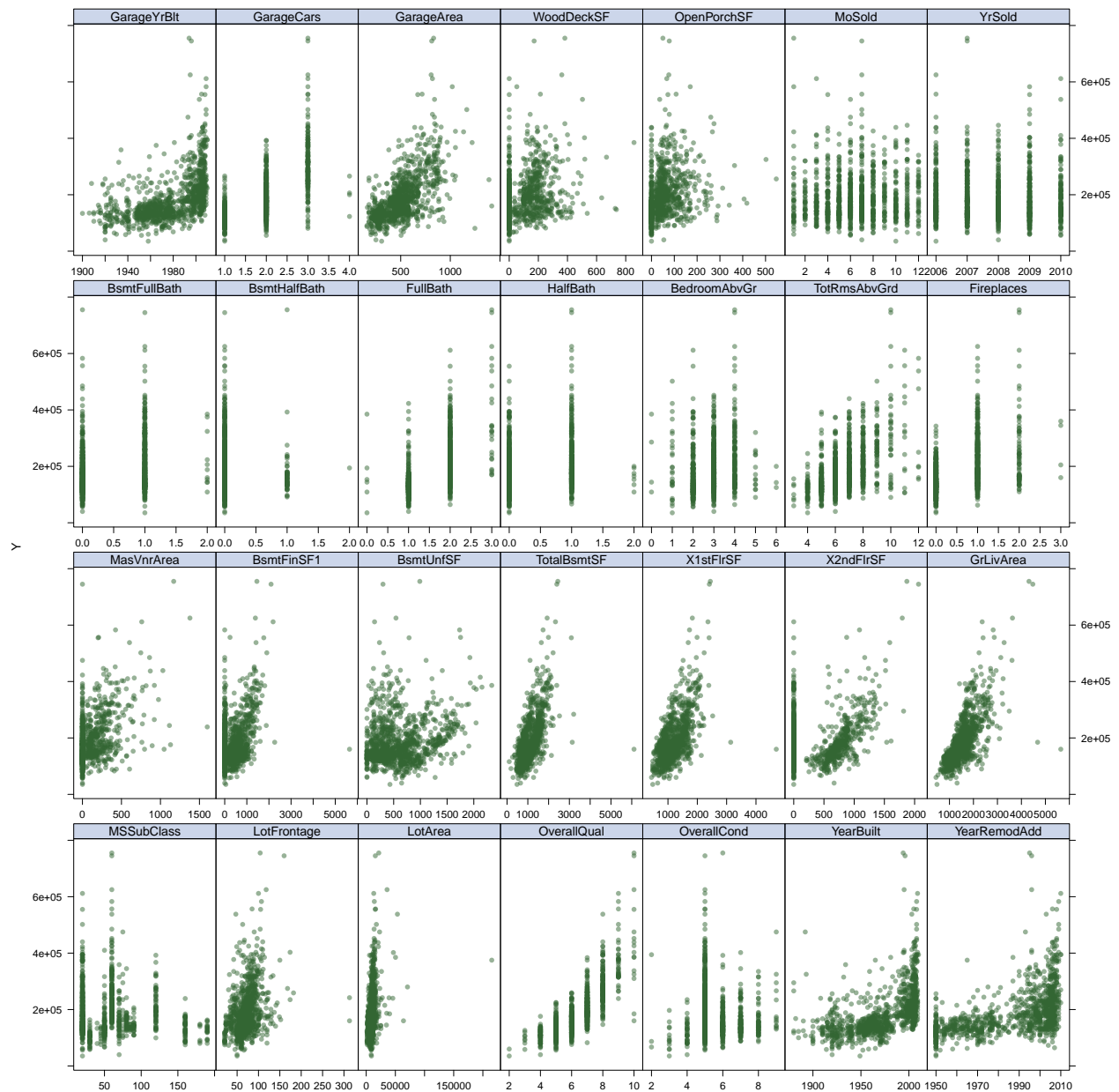
Figure 2 Density of log(SalePrice)



```
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))
```



```

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

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

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

