library(readxl) library(ggplot2) library(reshape2)

```
library(readxl)
lab5df <- read_excel("/Users/donglinxiong/Downloads/Lab5/Lab5DataSet.xlsx")
```

#Inspect the data

```
lab5df
```

```
## # A tibble: 1,460 x 81
##       Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
##    <dbl>      <dbl> <chr>    <chr>         <dbl> <chr>  <chr> <chr>
##  1     1         60 RL       65             8450 Pave   NA    Reg
##  2     2         20 RL       80             9600 Pave   NA    Reg
##  3     3         60 RL       68            11250 Pave   NA    IR1
##  4     4         70 RL       60             9550 Pave   NA    IR1
##  5     5         60 RL       84            14260 Pave   NA    IR1
##  6     6         50 RL       85            14115 Pave   NA    IR1
##  7     7         20 RL       75            10084 Pave   NA    Reg
##  8     8         60 RL       NA            10382 Pave   NA    IR1
##  9     9         50 RM       51             6120 Pave   NA    Reg
## 10    10        190 RL       50             7420 Pave   NA    Reg
## # i 1,450 more rows
## # i 73 more variables: LandContour <chr>, Utilities <chr>, LotConfig <chr>,
## #   LandSlope <chr>, Neighborhood <chr>, Condition1 <chr>, Condition2 <chr>,
## #   BldgType <chr>, HouseStyle <chr>, OverallQual <dbl>, OverallCond <dbl>,
## #   YearBuilt <dbl>, YearRemodAdd <dbl>, RoofStyle <chr>, RoofMatl <chr>,
## #   Exterior1st <chr>, Exterior2nd <chr>, MasVnrType <chr>, MasVnrArea <chr>,
## #   ExterQual <chr>, ExterCond <chr>, Foundation <chr>, BsmtQual <chr>, ...
```

#remove row with missing value

```
#Data cleaning by remove missing value
lab5df <- na.omit(lab5df)
```

#Check the data type of each column

```
#over view the data type of each column
str(lab5df)
```

```
## tibble [1,460 x 81] (S3: tbl_df/tbl/data.frame)
##  $ Id           : num [1:1460] 1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : num [1:1460] 60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr [1:1460] "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : chr [1:1460] "65" "80" "68" "60" ...
##  $ LotArea      : num [1:1460] 8450 9600 11250 9550 14260 ...
##  $ Street       : chr [1:1460] "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr [1:1460] "NA" "NA" "NA" "NA" ...
##  $ LotShape     : chr [1:1460] "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr [1:1460] "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr [1:1460] "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr [1:1460] "Inside" "FR2" "Inside" "Corner" ...
```

```
##  $ LandSlope    : chr [1:1460] "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr [1:1460] "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr [1:1460] "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr [1:1460] "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr [1:1460] "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr [1:1460] "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : num [1:1460] 7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : num [1:1460] 5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : num [1:1460] 2003 1976 2001 1915 2000 ...
##  $ YearRemodAdd : num [1:1460] 2003 1976 2002 1970 2000 ...
##  $ RoofStyle    : chr [1:1460] "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr [1:1460] "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr [1:1460] "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr [1:1460] "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr [1:1460] "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : chr [1:1460] "196" "0" "162" "0" ...
##  $ ExterQual    : chr [1:1460] "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr [1:1460] "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr [1:1460] "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr [1:1460] "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr [1:1460] "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr [1:1460] "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr [1:1460] "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : num [1:1460] 706 978 486 216 655 ...
##  $ BsmtFinType2 : chr [1:1460] "Unf" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2   : num [1:1460] 0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : num [1:1460] 150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : num [1:1460] 856 1262 920 756 1145 ...
##  $ Heating      : chr [1:1460] "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr [1:1460] "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr [1:1460] "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr [1:1460] "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 1stFlrSF     : num [1:1460] 856 1262 920 961 1145 ...
##  $ 2ndFlrSF     : num [1:1460] 854 0 866 756 1053 ...
##  $ LowQualFinSF : num [1:1460] 0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : num [1:1460] 1710 1262 1786 1717 2198 ...
##  $ BsmtFullBath : num [1:1460] 1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : num [1:1460] 0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : num [1:1460] 2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : num [1:1460] 1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : num [1:1460] 3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : num [1:1460] 1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : chr [1:1460] "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : num [1:1460] 8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr [1:1460] "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : num [1:1460] 0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr [1:1460] "NA" "TA" "TA" "Gd" ...
##  $ GarageType   : chr [1:1460] "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : chr [1:1460] "2003" "1976" "2001" "1998" ...
##  $ GarageFinish : chr [1:1460] "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : num [1:1460] 2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : num [1:1460] 548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr [1:1460] "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr [1:1460] "TA" "TA" "TA" "TA" ...
```

```
##  $ PavedDrive   : chr [1:1460] "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF   : num [1:1460] 0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : num [1:1460] 61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: num [1:1460] 0 0 0 272 0 0 0 228 205 0 ...
##  $ 3SsnPorch    : num [1:1460] 0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : num [1:1460] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : num [1:1460] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : chr [1:1460] "NA" "NA" "NA" "NA" ...
##  $ Fence        : chr [1:1460] "NA" "NA" "NA" "NA" ...
##  $ MiscFeature  : chr [1:1460] "NA" "NA" "NA" "NA" ...
##  $ MiscVal      : num [1:1460] 0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : num [1:1460] 2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : num [1:1460] 2008 2007 2008 2006 2008 ...
##  $ SaleType     : chr [1:1460] "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr [1:1460] "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : num [1:1460] 208500 181500 223500 140000 250000 ...
```

#My data set has 81 columns and 1460 rows. The data type of each column is either integer or double and text.

#calcualte and interpret the following descriptive statistics for a numeric variable in the dataset

```
#mean
mean(lab5df$MoSold)
```

```
## [1] 6.321918
```

```
#median
median(lab5df$MoSold)
```

```
## [1] 6
```

```
#standard deviation
sd(lab5df$MoSold)
```

```
## [1] 2.703626
```

```
#Minimum
min(lab5df$MoSold)
```

```
## [1] 1
```

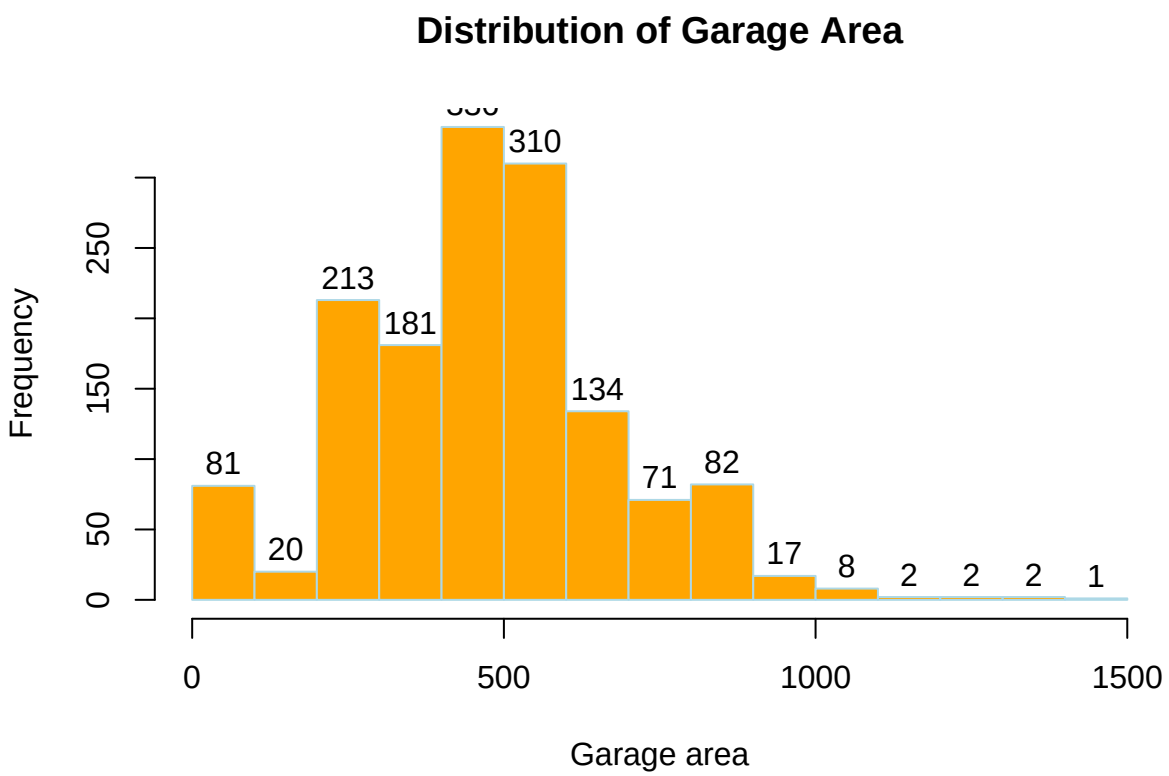
```
#Maximum
max(lab5df$MoSold)
```

```
## [1] 12
```

Interpresation: The mean of MoSold is 6.32, the median is 6, the standard deviation is 2.7, the minimum is 1, and the maximum is 12.

#Create a histogram to visualize the distribution of a numeric variable

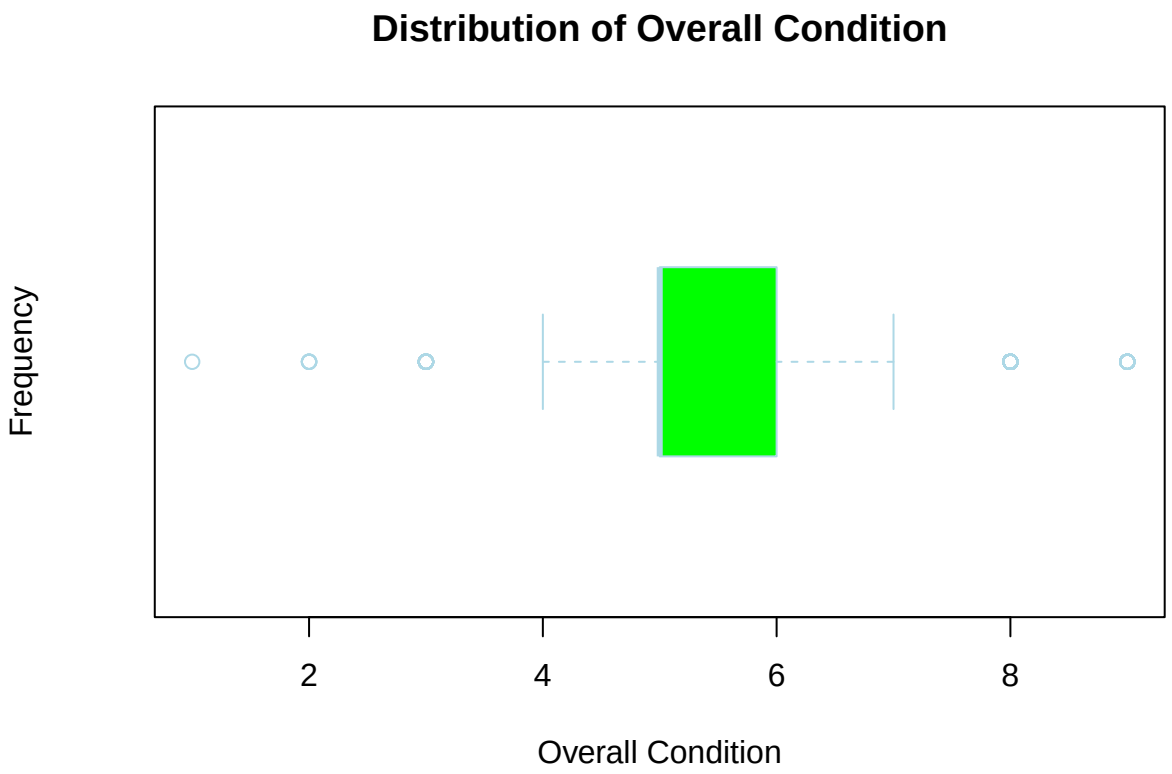
```r
#choose an appropriate varaible and customize the histogram
hist(lab5df$GarageArea,
     main = "Distribution of Garage Area",
     xlab = "Garage area",
     ylab = "Frequency",
     col = "Orange",
     border = "lightblue",
     labels = TRUE
     )
```



interpretation: The histogram shows the distribution of garage area. The majority of the garage area is between 0 and 1000.

#create a box plot to show the distribution of a another numeric variable

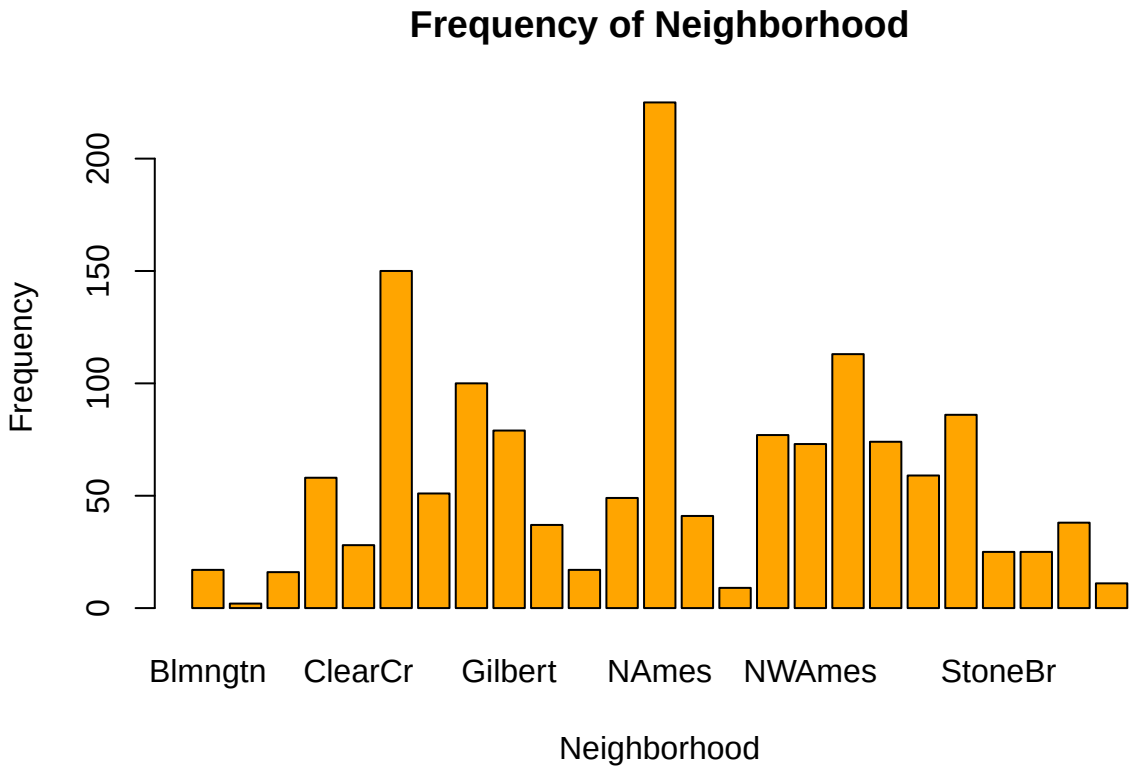
```r
#choose an appropriate varaible and customize the box plot
boxplot(lab5df$OverallCond,
        main = "Distribution of Overall Condition",
        xlab = "Overall Condition",
        ylab = "Frequency",
        col = "GREEN",
        border = "LIGHTBLUE",
        horizontal = TRUE
        )
```



Interpretation: The box plot shows the distribution of overall condition. The majority of the overall condition is between 5 and 7.

#generate a bar chart to display the frequency of a categorical variable

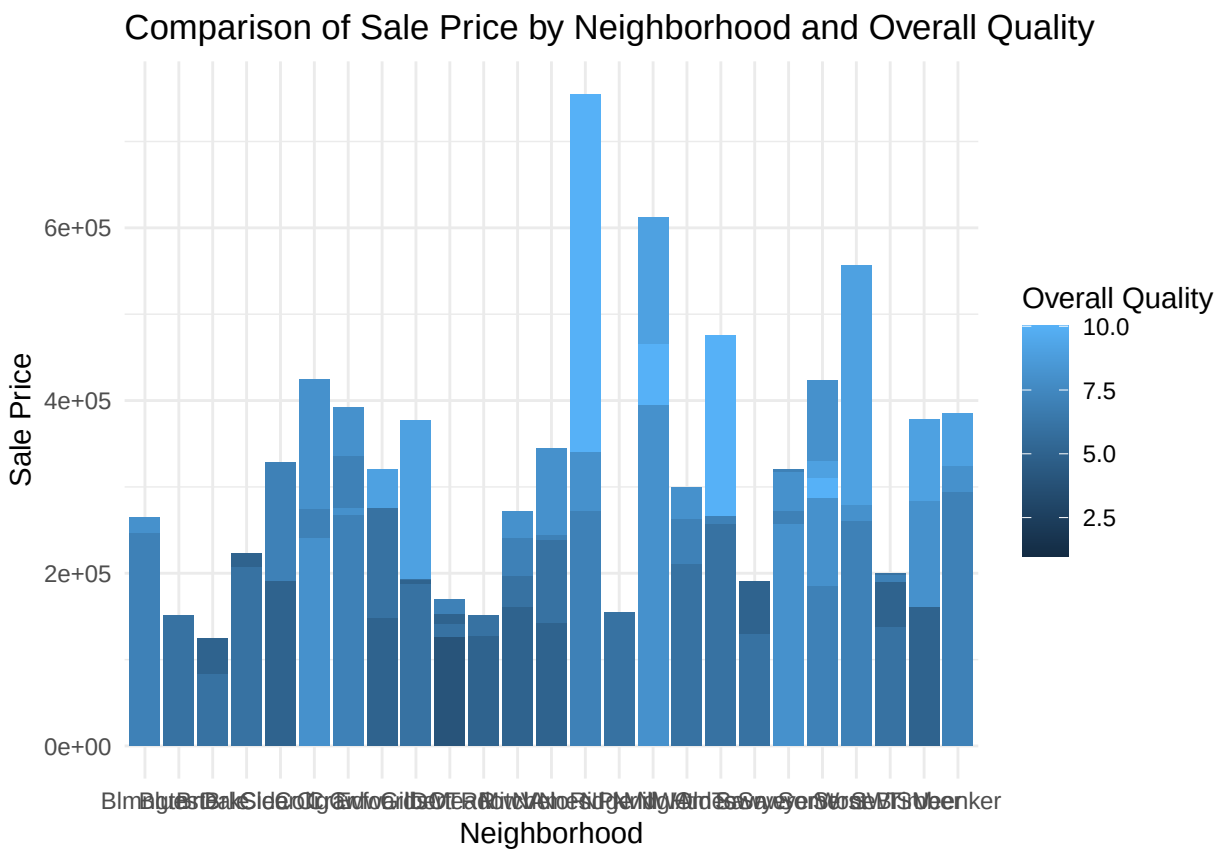
```
#lable the x-axis and y-axis
barplot(table(lab5df$Neighborhood),
        main = "Frequency of Neighborhood",
        xlab = "Neighborhood",
        ylab = "Frequency",
        col = "orange",
        border = "black"
        )
```



Interpretation: The bar chart shows the frequency of neighborhood. The majority of the neighborhood is between 0 and 150.

#Create a grouped bar chart showing the comparison of at least two categorical variables against one numberica variable

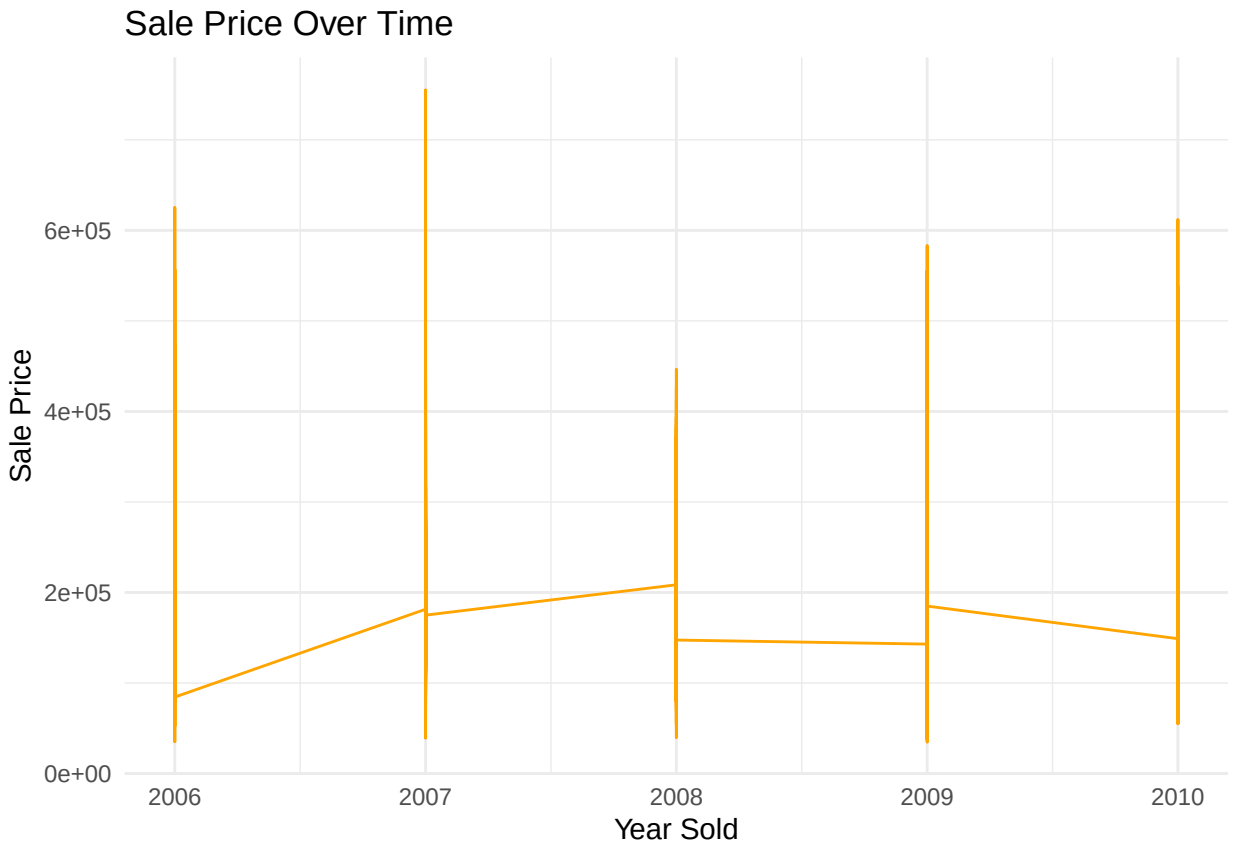
```
library(ggplot2)
#customize the grouped bar chart to make it visually appealing
ggplot(lab5df, aes(x = Neighborhood, y = SalePrice, fill = OverallQual)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Comparison of Sale Price by Neighborhood and Overall Quality",
       x = "Neighborhood",
       y = "Sale Price",
       fill = "Overall Quality") +
  theme_minimal()
```



Interpretation: The grouped bar chart shows the comparison of sale price by neighborhood and overall quality. The majority of the sale price is between 0 and 200000.

#Create a line graph to visualize a time-series dataset within the dataset. Label the x and y axis and provide a title for your graph

```
#customize the line graph to make it visually appealing
ggplot(lab5df, aes(x = YrSold, y = SalePrice, group = 1)) +
  geom_line(color = "Orange") +
  labs(title = "Sale Price Over Time",
       x = "Year Sold",
       y = "Sale Price") +
  theme_minimal()
```



Interpretation: The line graph shows the sale price over time. The majority of the sale price is between 0 and 800000.

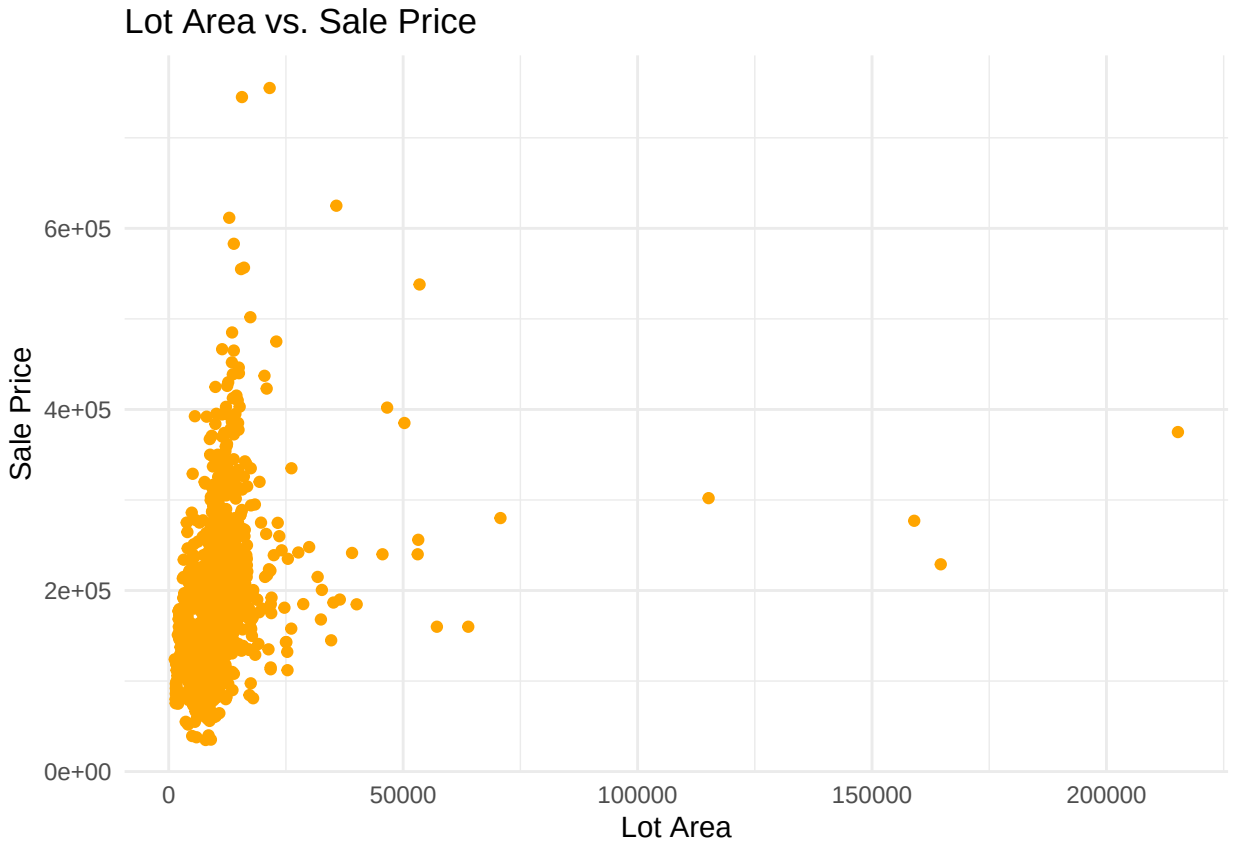
#generate an area chart to represent the cumulative sum of a numeric variable over time.

```
#Customize the area chart to make it visually appealing
ggplot(lab5df, aes(x = YrSold, y = cumsum(SalePrice))) +
  geom_area(fill = "lightblue") +
  labs(title = "Cumulative Sale Price Over Time",
       x = "Year Sold",
       y = "Cumulative Sale Price") +
  theme_minimal()
```

Interpretation: The area chart shows the cumulative sale price over time. The majority of the cumulative sale price is between 0 and 10000000.the shape of the area chart is similar to the line graph.

#Create a scatter plot to explor the relationship between two numeric variables

```
#Customize the scatter plot to make it visually appealing
ggplot(lab5df, aes(x = LotArea, y = SalePrice)) +
  geom_point(color = "orange") +
  labs(title = "Lot Area vs. Sale Price",
       x = "Lot Area",
       y = "Sale Price") +
  theme_minimal()
```

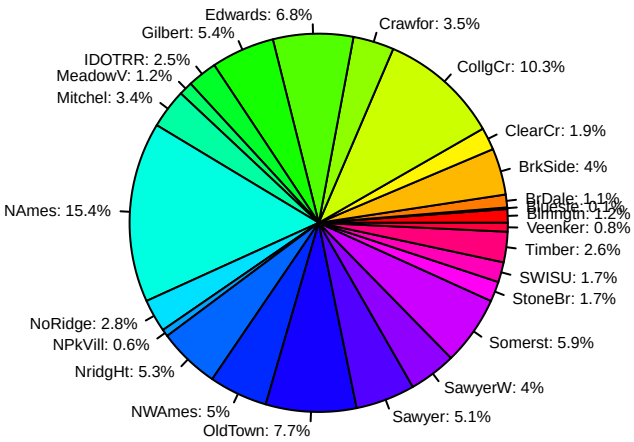


Interpretation: The scatter plot shows the relationship between lot area and sale price. The majority of the lot area is between 0 and 200000.And 0-50000 lot area has the most sale price.

#Design a pie chart to illustrate the composition of a categorical variable

```
# Create the pie chart
pie(table(lab5df$Neighborhood),
    main = "Composition of Neighborhood",
    col = rainbow(length(table(lab5df$Neighborhood))),
    labels = paste(names(table(lab5df$Neighborhood)), ": ", round(prop.table(table(lab5df$Neighborhood)
                   sep = ""),
    cex = 0.5
    )
```

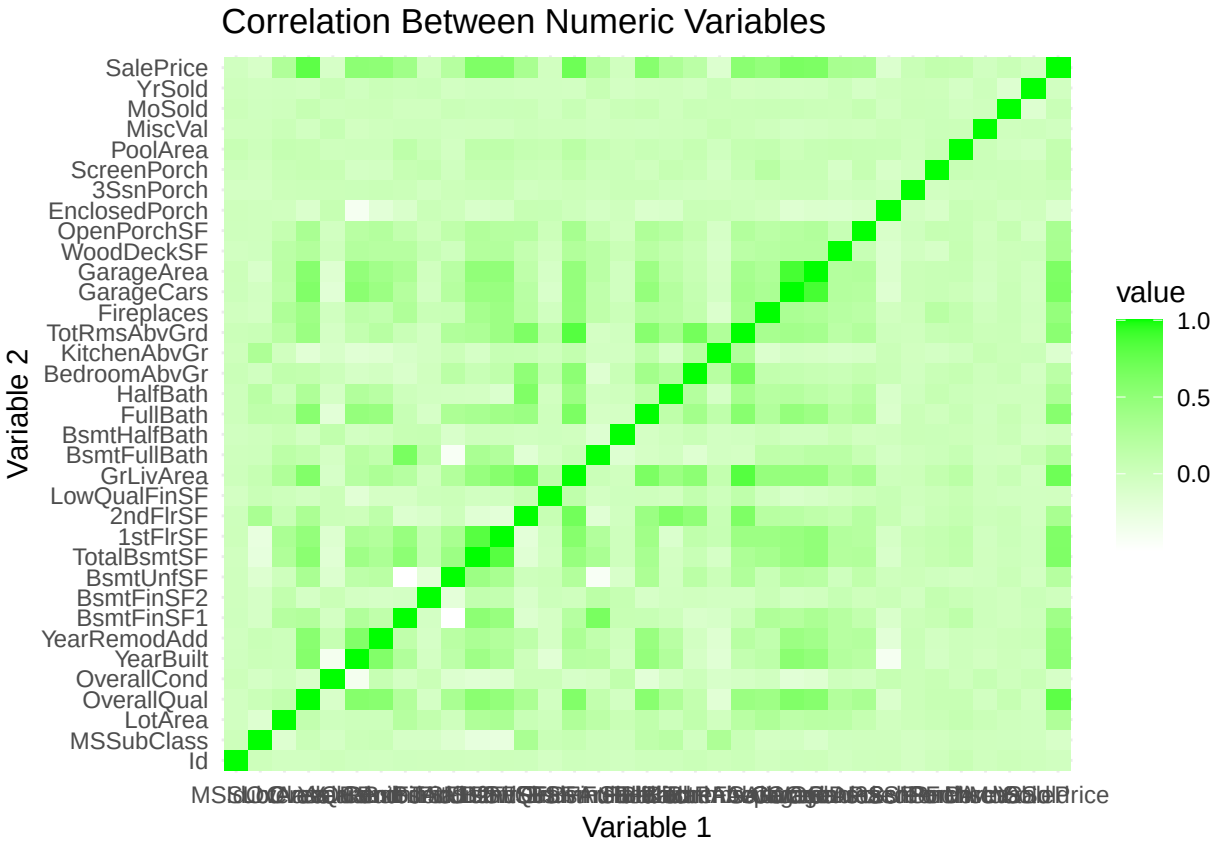
Composition of Neighborhood



Interpretation: The pie chart shows the composition of neighborhood. The majority of the neighborhood is between 0 and 20%.

#Build a heatmap to visualize the correlation between numeric variables

```
library(reshape2)
numeric_data <- lab5df[, sapply(lab5df, is.numeric)]
correlation_matrix <- cor(numeric_data)
melted_correlation_matrix <- melt(correlation_matrix)
ggplot(melted_correlation_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "green") +
  labs(title = "Correlation Between Numeric Variables",
       x = "Variable 1",
       y = "Variable 2") +
  theme_minimal()
```



interpretation: The heatmap shows the correlation between numeric variables. The majority of the correlation is between 0 and 1. The darker the color, the higher the correlation.it indicates that the correlation between numeric variables is strong.