EDA

Archit

Data loading

```
house <- read.csv("D:/SEM-6/Statistics/Project/Housing Prices/train.csv", header = TRUE)
head(house, n=3)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0        <NA>     Attchd        2003          RFn          2
## 2          1          TA     Attchd        1976          RFn          2
## 3          1          TA     Attchd        2001          RFn          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
```

```
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
```

Missing values

```
missing_counts <- colSums(is.na(house))
missing_cols <- names(which(missing_counts > 1))
print("Missing Values:")
```

```
## [1] "Missing Values:"
```

```
missing_counts
```

```
##            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             1             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
```

```
sort(missing_cols)
```

```
##  [1] "Alley"        "BsmtCond"     "BsmtExposure" "BsmtFinType1" "BsmtFinType2"
##  [6] "BsmtQual"     "Fence"        "FireplaceQu"  "GarageCond"   "GarageFinish"
## [11] "GarageQual"   "GarageType"   "GarageYrBlt"  "LotFrontage"  "MasVnrArea"
## [16] "MasVnrType"   "MiscFeature"  "PoolQC"
```

Filling missing values and encoding categorical variables

Analyzing the values of the variable PoolQC first as it had highest number of NA values

1. PoolQC

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

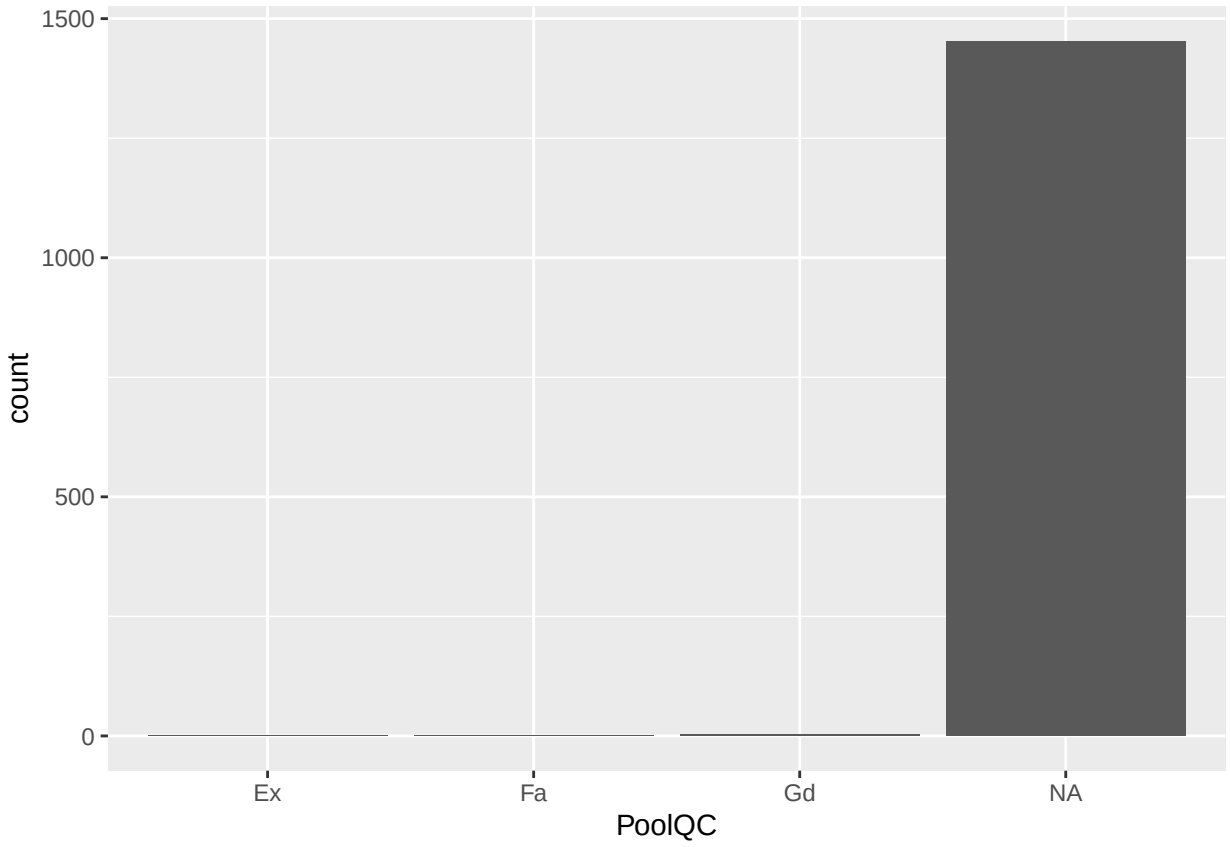
```
summary(house$PoolQC)
```

```
##    Length     Class      Mode
##      1460 character character
```

```
head(house$PoolQC)
```

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

```
ggplot(data = house, aes(x = PoolQC)) +  geom_bar()
```




```
#
```

Clearly the variable has no values at all in any field so a good option would be to just drop it and not analyze the target variable on the basis of it

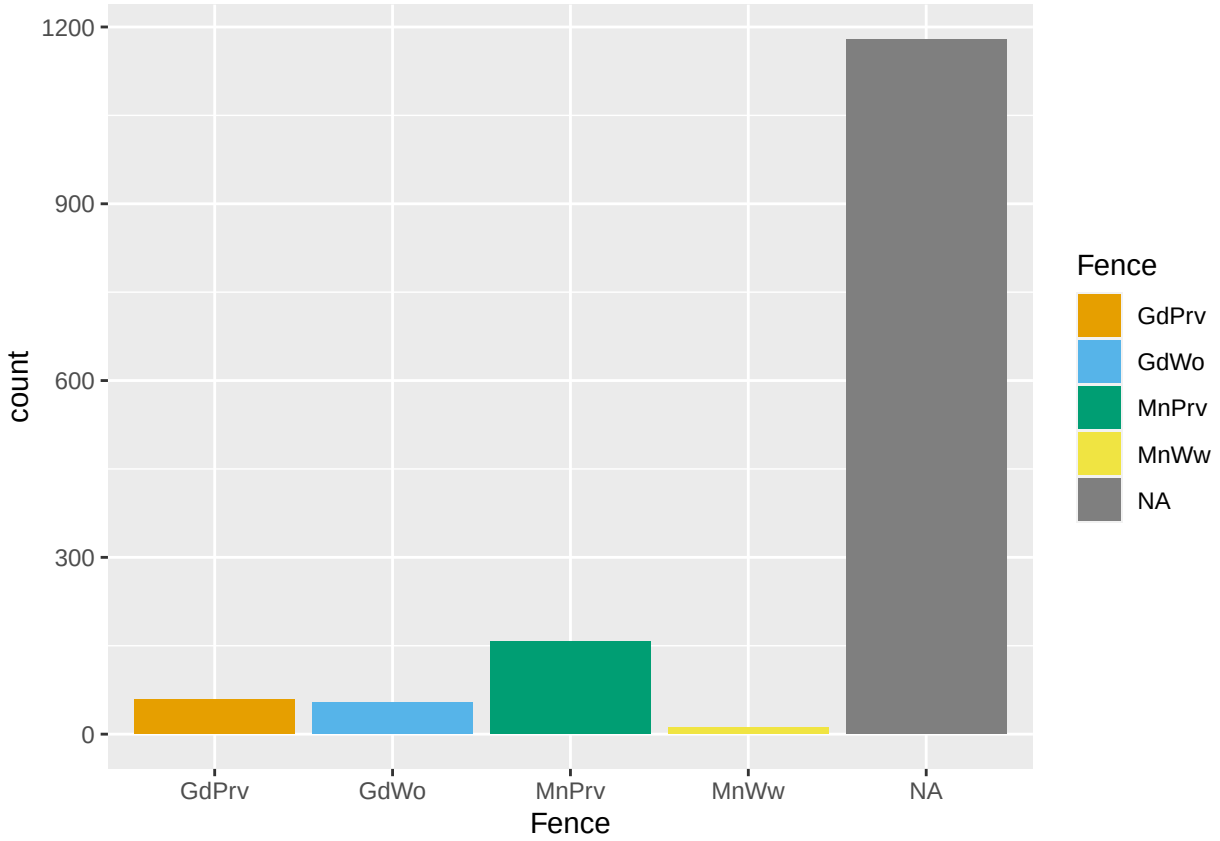
```
house$PoolQC <- NULL
```

2. Fence GdPrv Good Privacy MnPrv Minimum Privacy GdWo Good Wood MnWw Minimum Wood/Wire NA No Fence

```
summary(house$Fence)
```

```
##    Length     Class      Mode
##      1460 character character
```

```
ggplot(data = house, aes(x = Fence, fill = Fence)) +
  geom_bar() +
  scale_fill_manual(values = c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79
```



This NA is indicative of the number of houses which have no fence and these need not be filled with values as they would then lead to wrong impression of the houses while predicting the data so these will get converted into a category while encoding

```
factor_var <- factor(house$Fence, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Fence = numeric_labels
print(house$Fence[1:10])
```

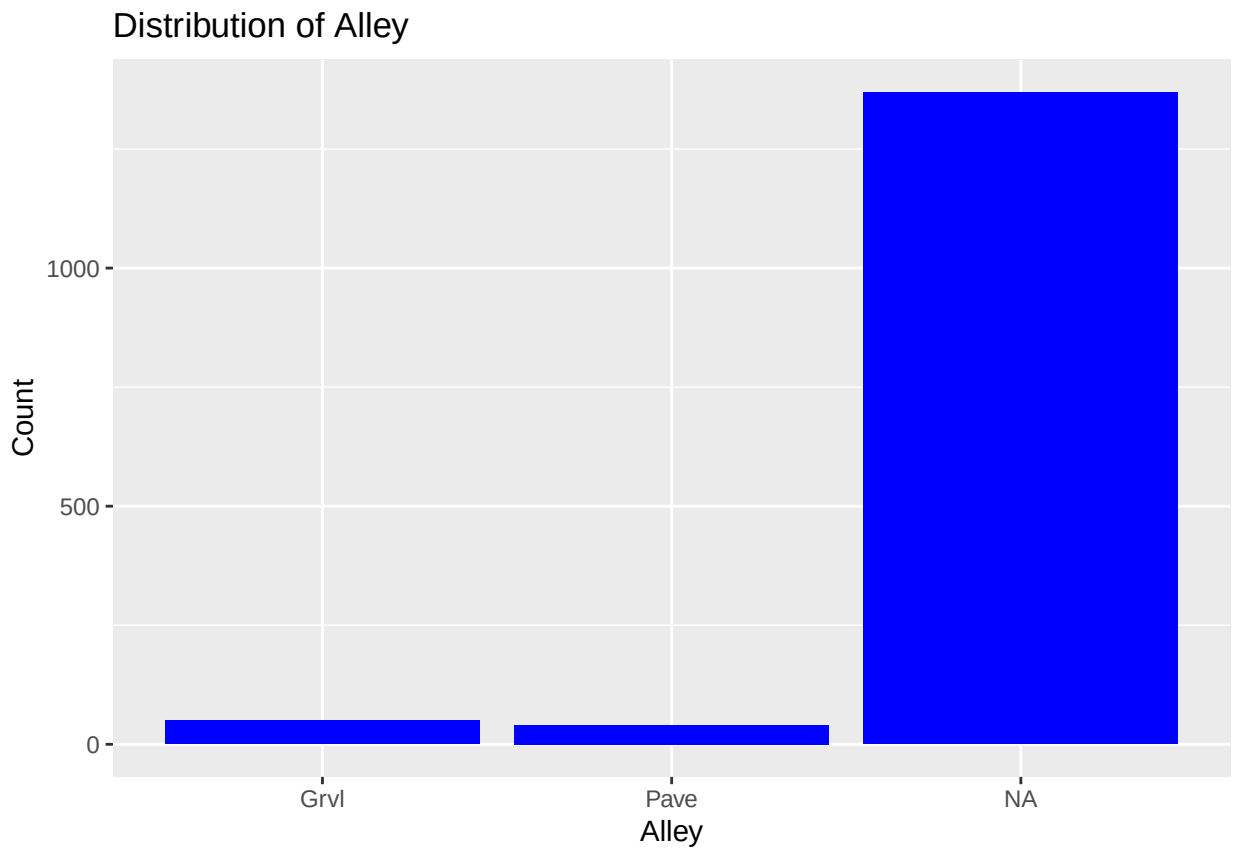
```
##  [1] 5 5 5 5 5 3 5 5 5 5
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "GdPrv" "GdWo"  "MnPrv" "MnWw"  NA
```

3. Alley

```
library(ggplot2)
ggplot(house, aes(x = Alley)) +
  geom_bar(fill = "blue") +
  labs(title = "Distribution of Alley", x = "Alley", y = "Count")
```



In this case also the NA indicates absence of any pavement for the houses which would mean that the houses dont have an access to alley and the NA values are important

```
factor_var <- factor(house$Alley, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Alley = numeric_labels
print(house$Alley[1:10])
```

```
##  [1] 3 3 3 3 3 3 3 3 3 3
```

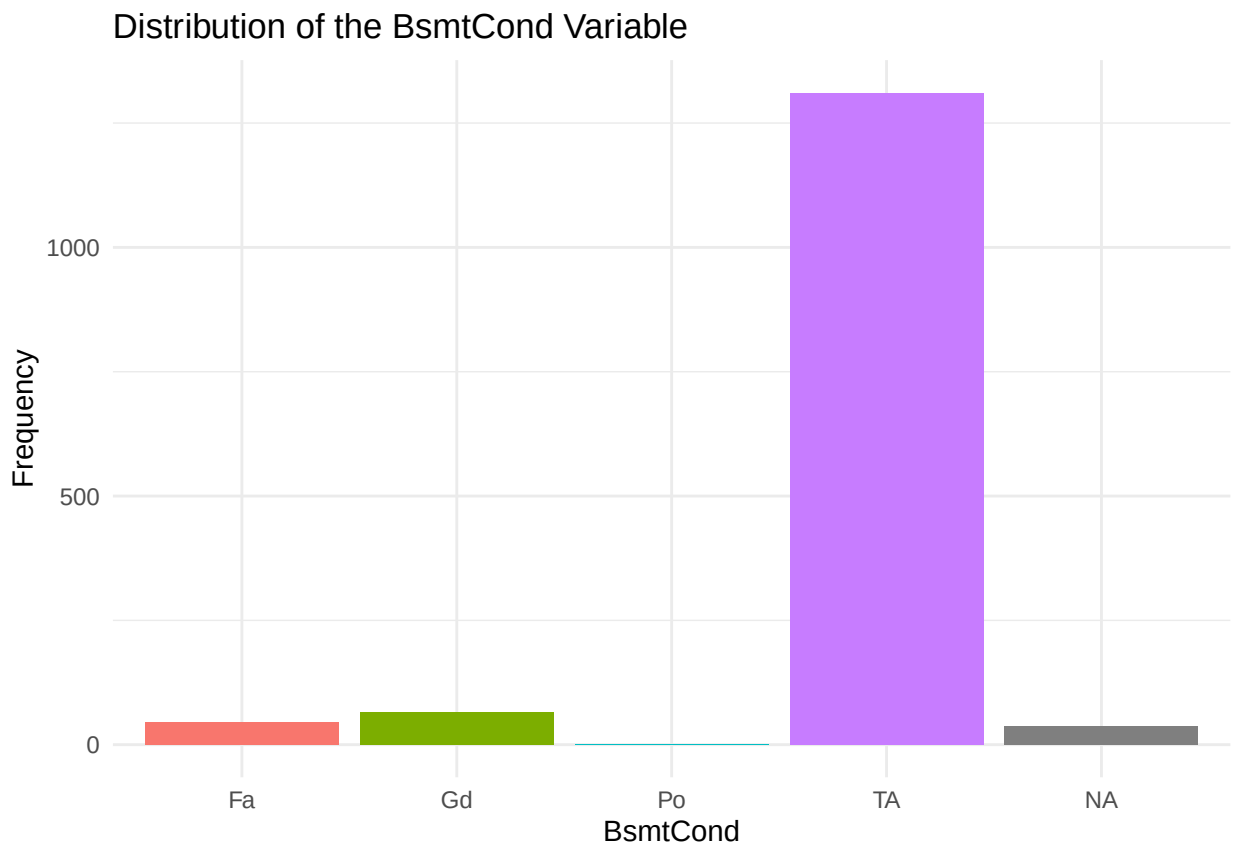
```
categories <- levels(factor_var)
categories
```

```
## [1] "Grvl" "Pave" NA
```

4. BsmtCond Evaluates the general condition of the basement

```
library(ggplot2)

my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999")

ggplot(house, aes(x = BsmtCond, fill = BsmtCond)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
```

```
  labs(title = "Distribution of the BsmtCond Variable",
       x = "BsmtCond",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$BsmtCond, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$BsmtCond = numeric_labels
print(house$BsmtCond[1:10])
```

```
##  [1] 4 4 4 2 4 4 4 4 4 4
```

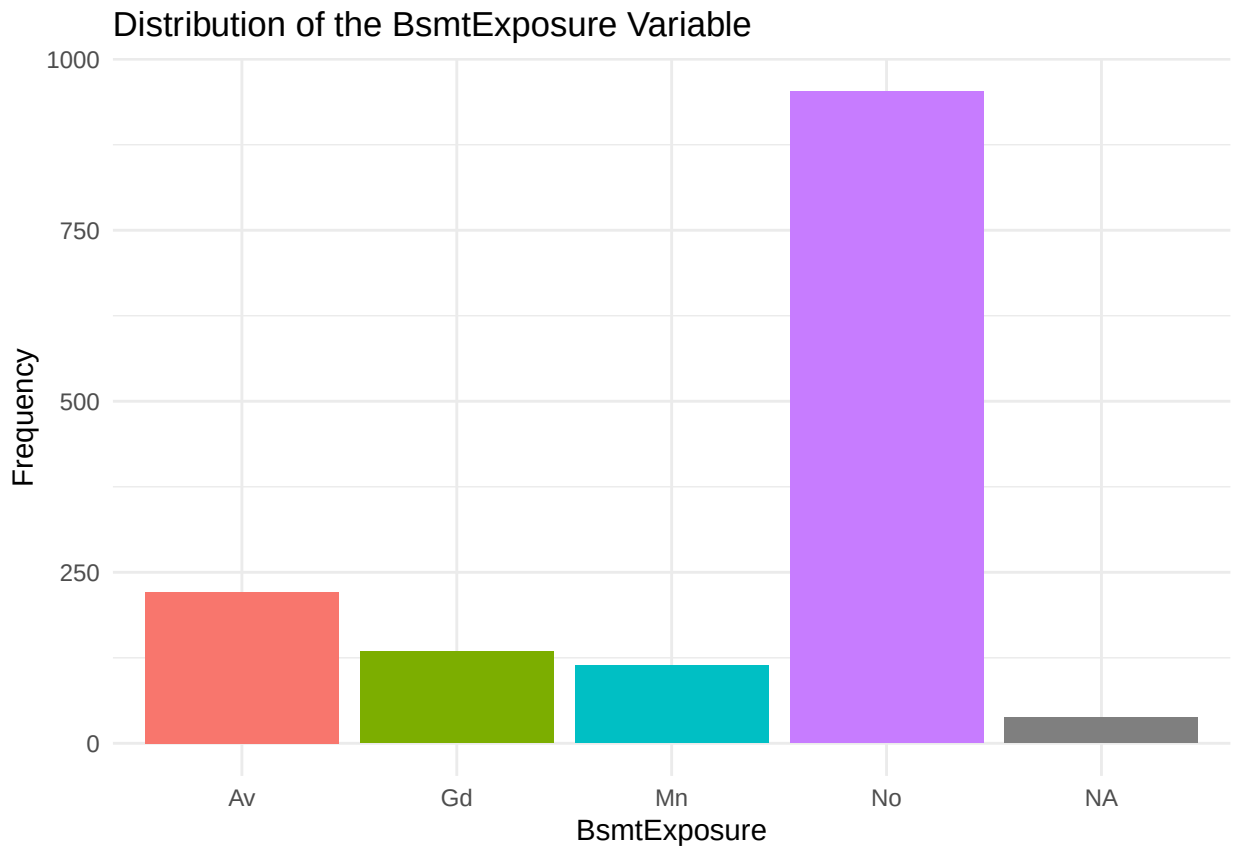
```
categories <- levels(factor_var)
categories
```

```
## [1] "Fa" "Gd" "Po" "TA" NA
```

5. BsmtExposure Refers to walkout or garden level walls

```
library(ggplot2)

my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999")
```

```
ggplot(house, aes(x = BsmtExposure, fill = BsmtExposure)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the BsmtExposure Variable",
       x = "BsmtExposure",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$BsmtExposure, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$BsmtExposure = numeric_labels
print(house$BsmtExposure[1:10])
```

```
##  [1] 4 2 3 4 1 4 1 3 4 4
```

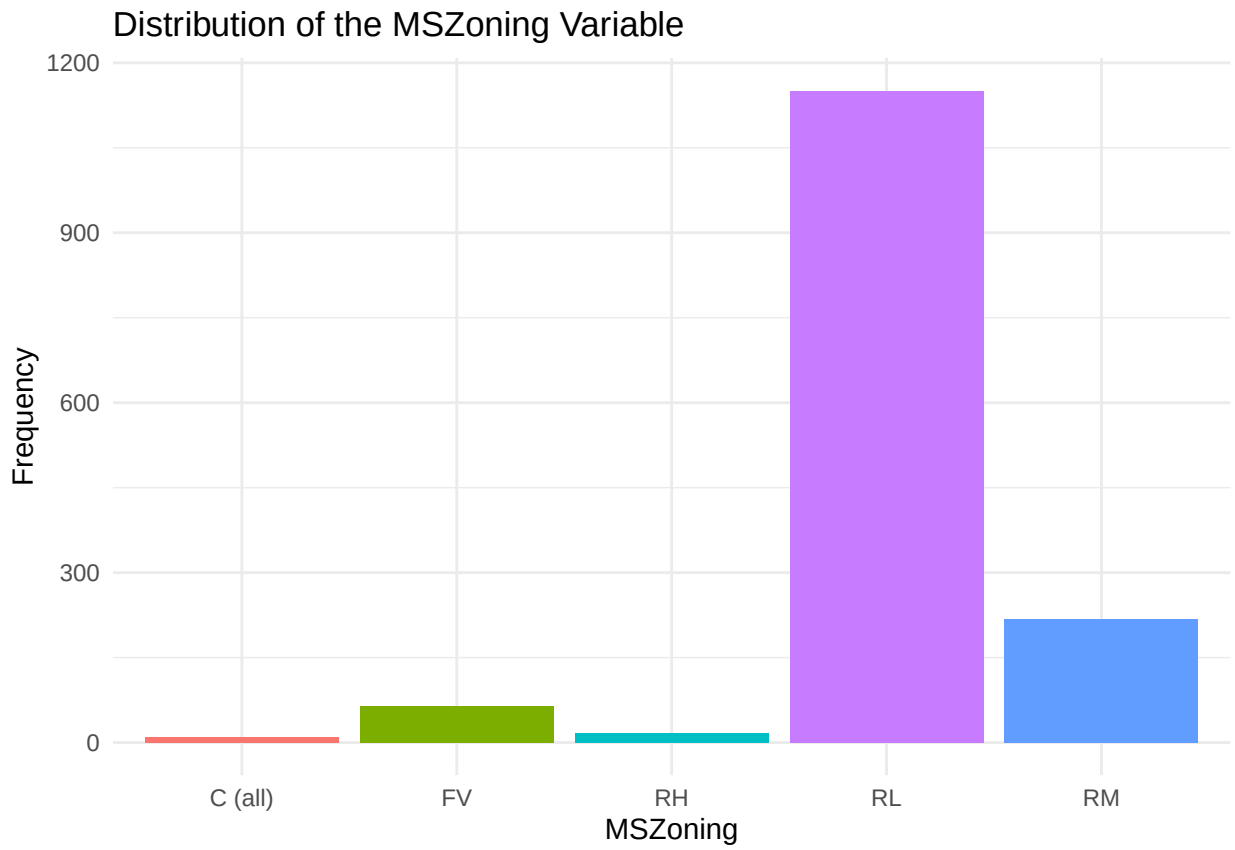
```
categories <- levels(factor_var)
categories
```

```
## [1] "Av" "Gd" "Mn" "No" NA
```

6. MSZoning

Identifies the general zoning classification of the sale

```
library(ggplot2)

my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","blue")

ggplot(house, aes(x = MSZoning, fill = MSZoning)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the MSZoning Variable",
       x = "MSZoning",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
print("Number of NA's:")
```

```
## [1] "Number of NA's:"
```

```
sum(is.na(house$MSZoning))
```

```
## [1] 0
```

```
house$MSZoning[is.na(house$MSZoning)] <- names(sort(table(house$MSZoning), decreasing = TRUE))[1]
sum(is.na(house$MSZoning))
```

```
## [1] 0
```

Filling the NA values with mode of the variable i.e. RL category and then encoding it

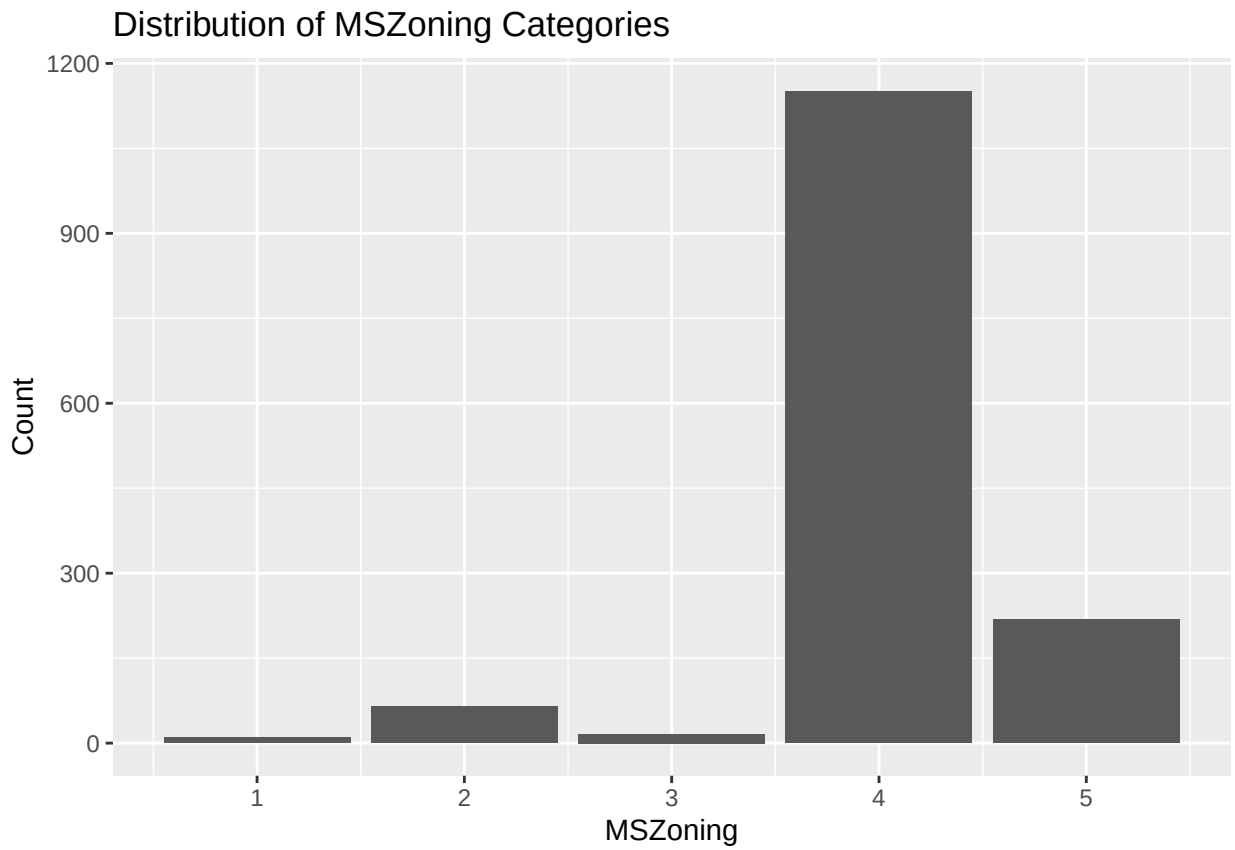
```
factor_var <- factor(house$MSZoning, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$MSZoning = numeric_labels
print(house$MSZoning[1:10])
```

```
##  [1] 4 4 4 4 4 4 4 4 5 4
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "C (all)" "FV"      "RH"      "RL"      "RM"
```

```
library(ggplot2)

# bar plot after encoding
ggplot(house, aes(x = MSZoning)) +
  geom_bar() +
  labs(x = "MSZoning", y = "Count", title = "Distribution of MSZoning Categories")
```

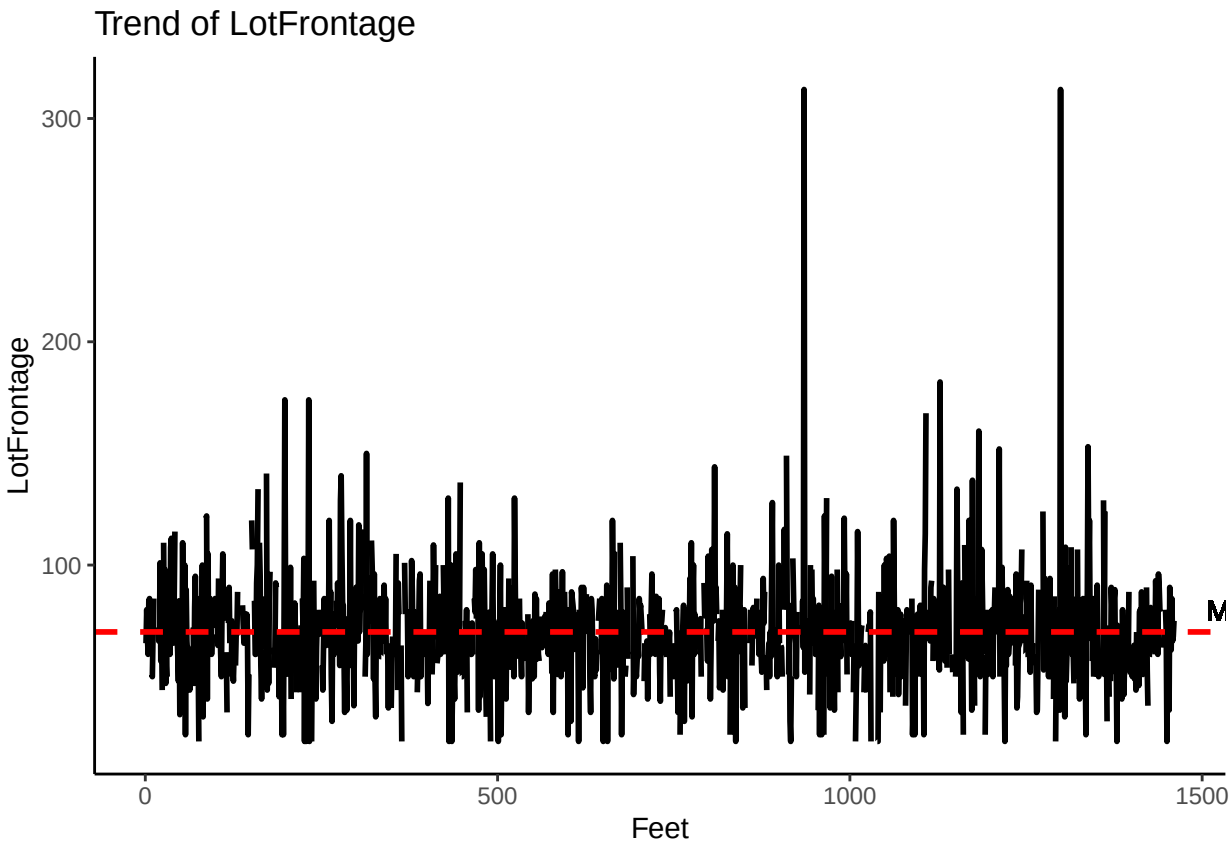


7. LotFrontage Linear feet of street connected to property

```
ggplot(house, aes(x = seq_along(LotFrontage), y = LotFrontage)) +
  geom_line(color = "black", size = 1) +
  geom_hline(yintercept = mean(house$LotFrontage, na.rm = TRUE),
             color = "red", linetype = "dashed", size = 1) +
  geom_text(aes(x = max(seq_along(house$LotFrontage)),
                y = mean(house$LotFrontage, na.rm = TRUE),
                label = paste("Mean:", round(mean(house$LotFrontage, na.rm = TRUE), 2))),
            color = "black", size = 4, hjust = -0.2, vjust = -0.5) +
  labs(x = "Feet", y = "LotFrontage", title = "Trend of LotFrontage") +
  theme_classic()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

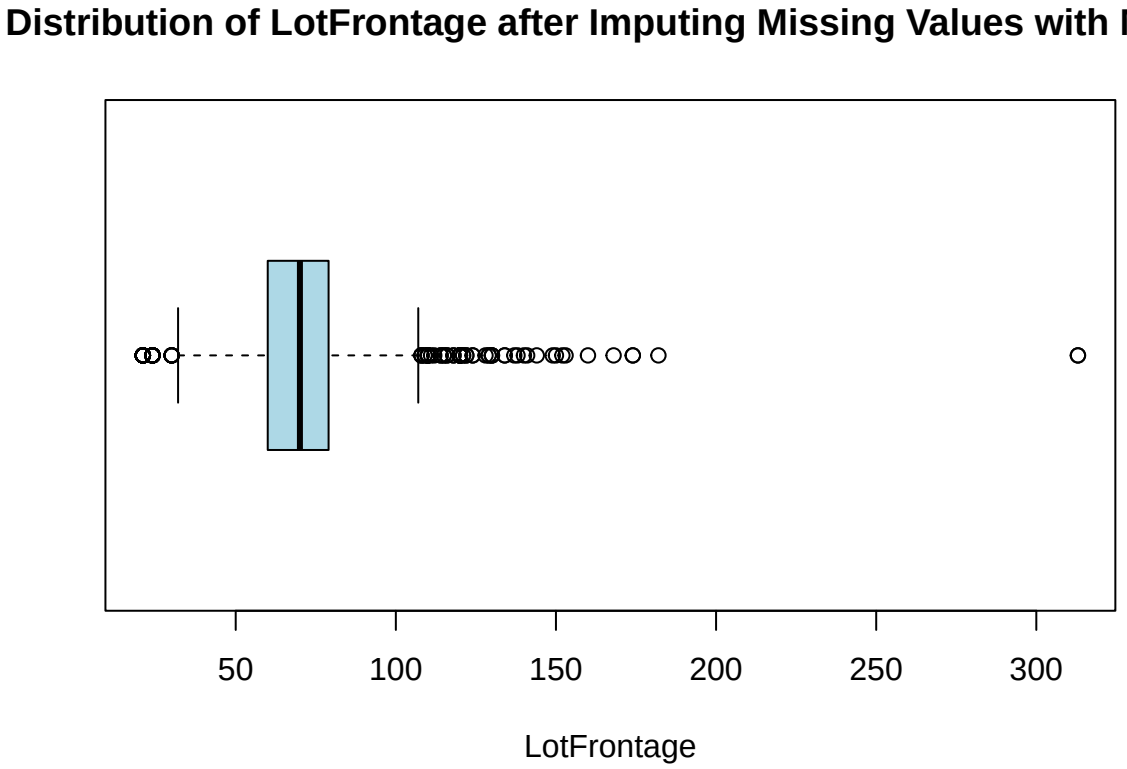
```
## Warning: Use of ‘house$LotFrontage‘ is discouraged.
## i Use ‘LotFrontage‘ instead.
## Use of ‘house$LotFrontage‘ is discouraged.
## i Use ‘LotFrontage‘ instead.
## Use of ‘house$LotFrontage‘ is discouraged.
## i Use ‘LotFrontage‘ instead.
```



```
mean_LotFrontage <- mean(house$LotFrontage, na.rm = TRUE)
house$LotFrontage[is.na(house$LotFrontage)] <- mean_LotFrontage
sum(is.na(house$MSZoning))
```

```
## [1] 0
```

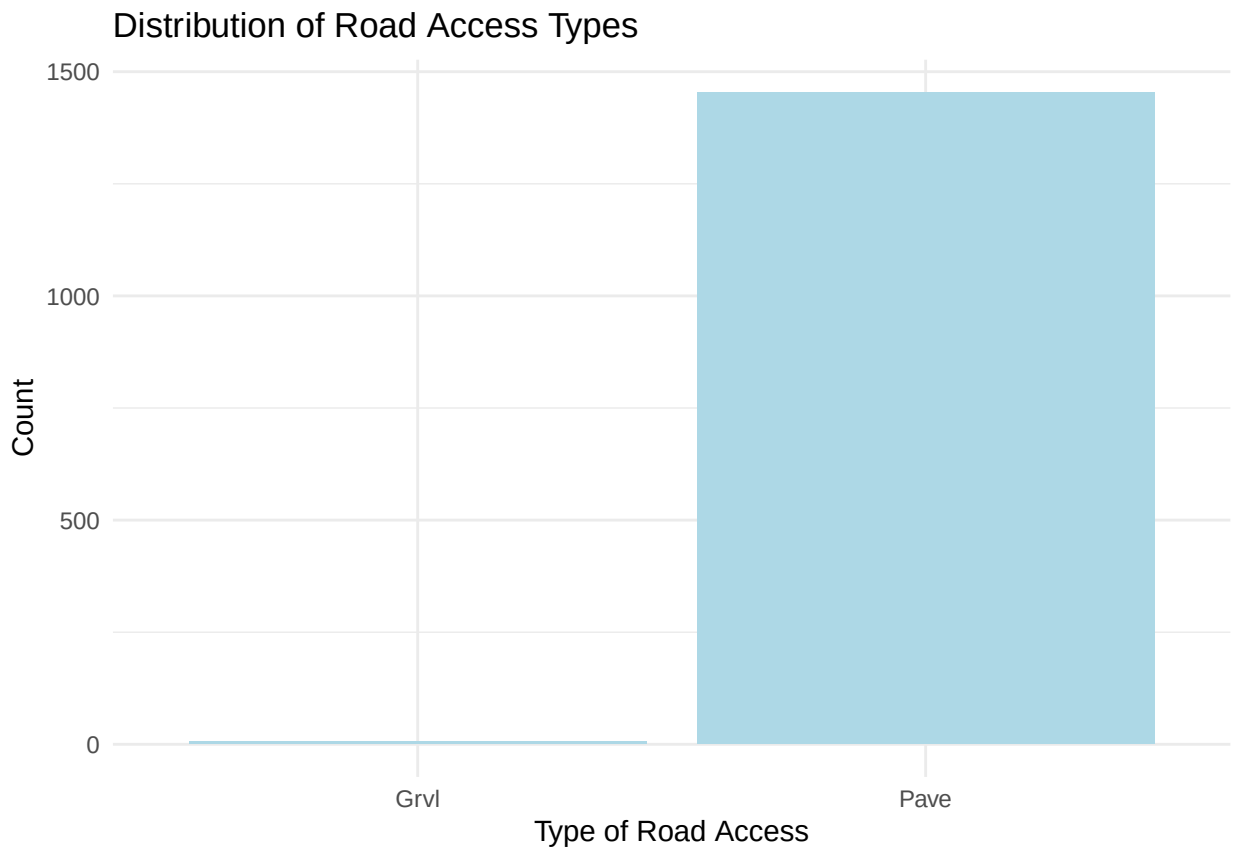
```
boxplot(house$LotFrontage, main = "Distribution of LotFrontage after Imputing Missing Values with Mean"
        xlab = "LotFrontage", col = "lightblue", border = "black", horizontal = TRUE)
```



Imputing Values with the mean of the variable

8. Street Type of road access to property

```
ggplot(house, aes(x = Street)) +
  geom_bar(fill = "lightblue") +
  labs(x = "Type of Road Access", y = "Count",
       title = "Distribution of Road Access Types") +
  theme_minimal()
```



Encoding categories:

```
factor_var <- factor(house$Street, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Street = numeric_labels
print(house$Street[1:10])
```

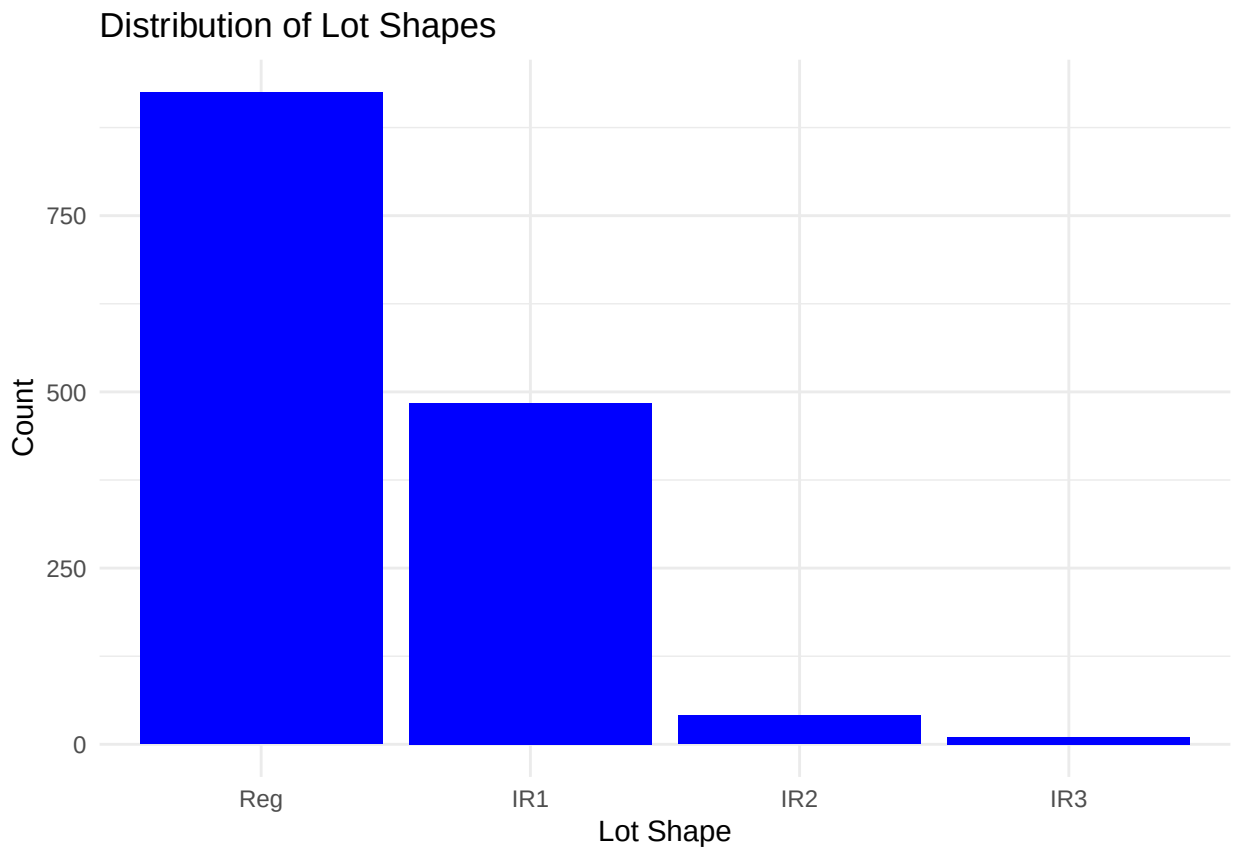
```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Grvl" "Pave"
```

9. LotShape General shape of property

```
ggplot(house, aes(x = LotShape)) +
  geom_bar(fill = "blue") +
  scale_x_discrete(limits = c("Reg", "IR1", "IR2", "IR3")) +
  labs(x = "Lot Shape", y = "Count",
       title = "Distribution of Lot Shapes") +
  theme_minimal()
```



Encoding categories:

```
factor_var <- factor(house$LotShape, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$LotShape = numeric_labels
print(house$LotShape[1:10])
```

```
##  [1] 4 4 1 1 1 1 4 1 4 4
```

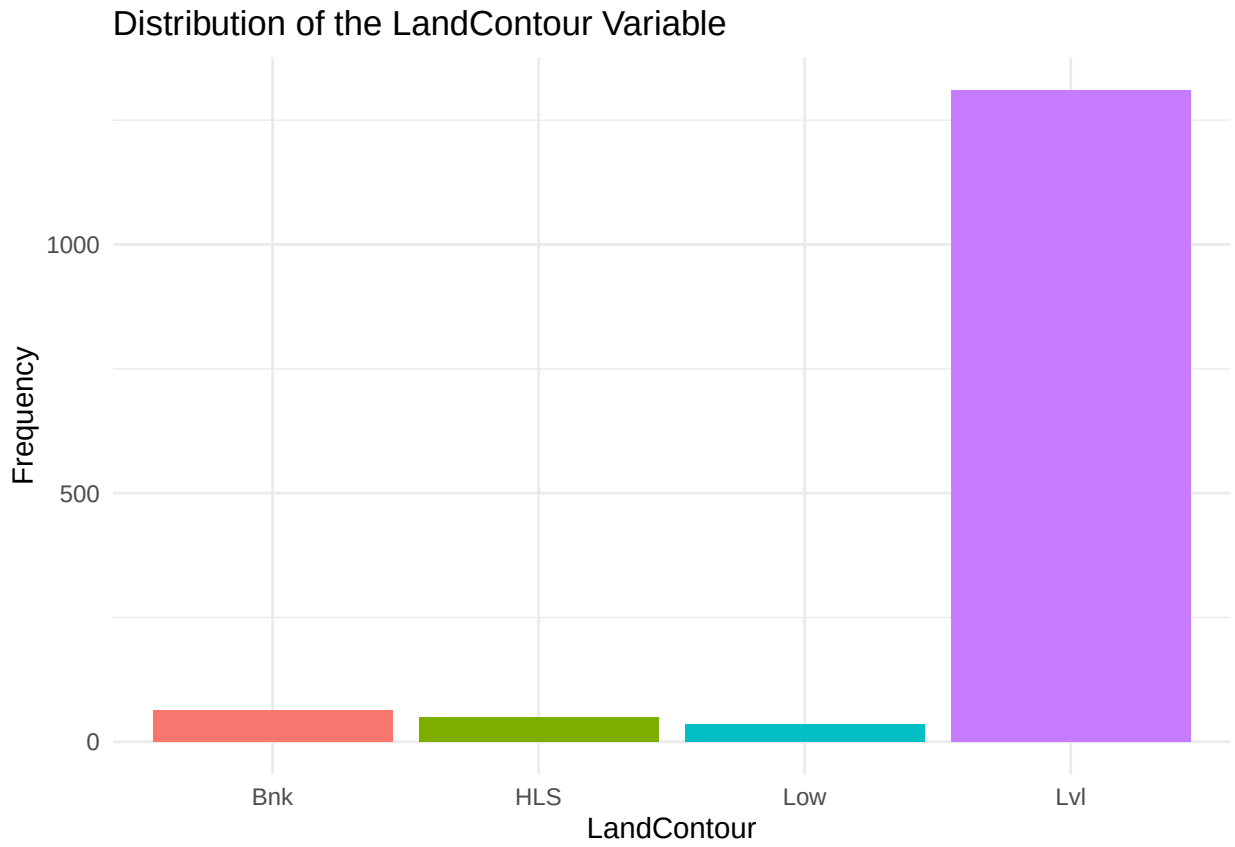
```
categories <- levels(factor_var)
categories
```

```
## [1] "IR1" "IR2" "IR3" "Reg"
```

10. LandContour Flatness of the property Lvl Near Flat/Level Bnk Banked - Quick and significant rise from street grade to building HLS Hillside - Significant slope from side to side Low Depression

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","blue")

ggplot(house, aes(x = LandContour, fill = LandContour)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the LandContour Variable",
       x = "LandContour",
```

```
      y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Encoding categories:

```
factor_var <- factor(house$LandContour, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$LandContour = numeric_labels
print(house$LandContour[1:10])
```

```
##  [1] 4 4 4 4 4 4 4 4 4 4
```

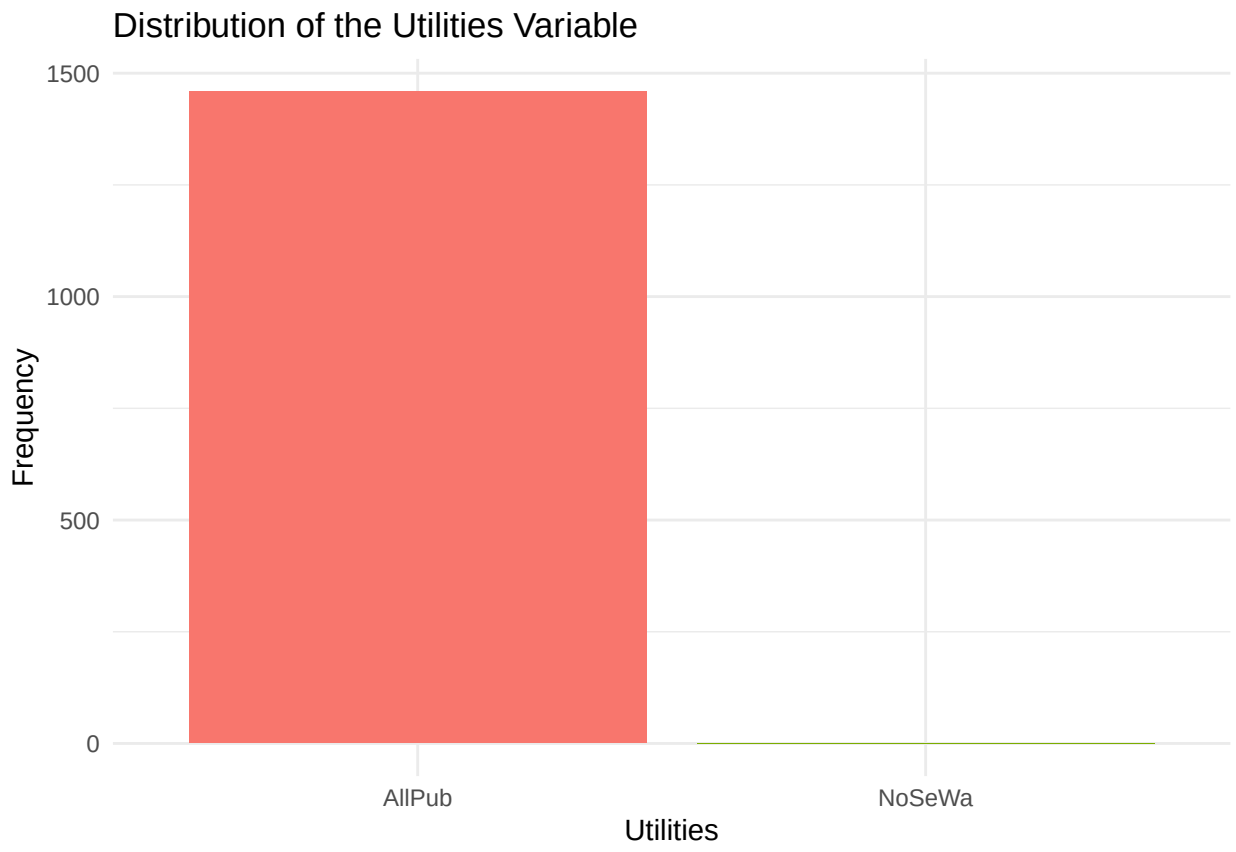
```
categories <- levels(factor_var)
categories
```

```
## [1] "Bnk" "HLS" "Low" "Lvl"
```

11. Utilities Type of utilities available

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","blue")

ggplot(house, aes(x = Utilities, fill = Utilities)) +
  geom_bar() +
```

```
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Utilities Variable",
       x = "Utilities",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
table(house$Utilities)
```

```
##
## AllPub NoSeWa
##   1459      1
```

Encoding categories:

```
factor_var <- factor(house$Utilities, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Utilities = numeric_labels
print(house$Utilities[1:10])
```

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

```
categories <- levels(factor_var)
categories
```

```
## [1] "AllPub" "NoSeWa"
```

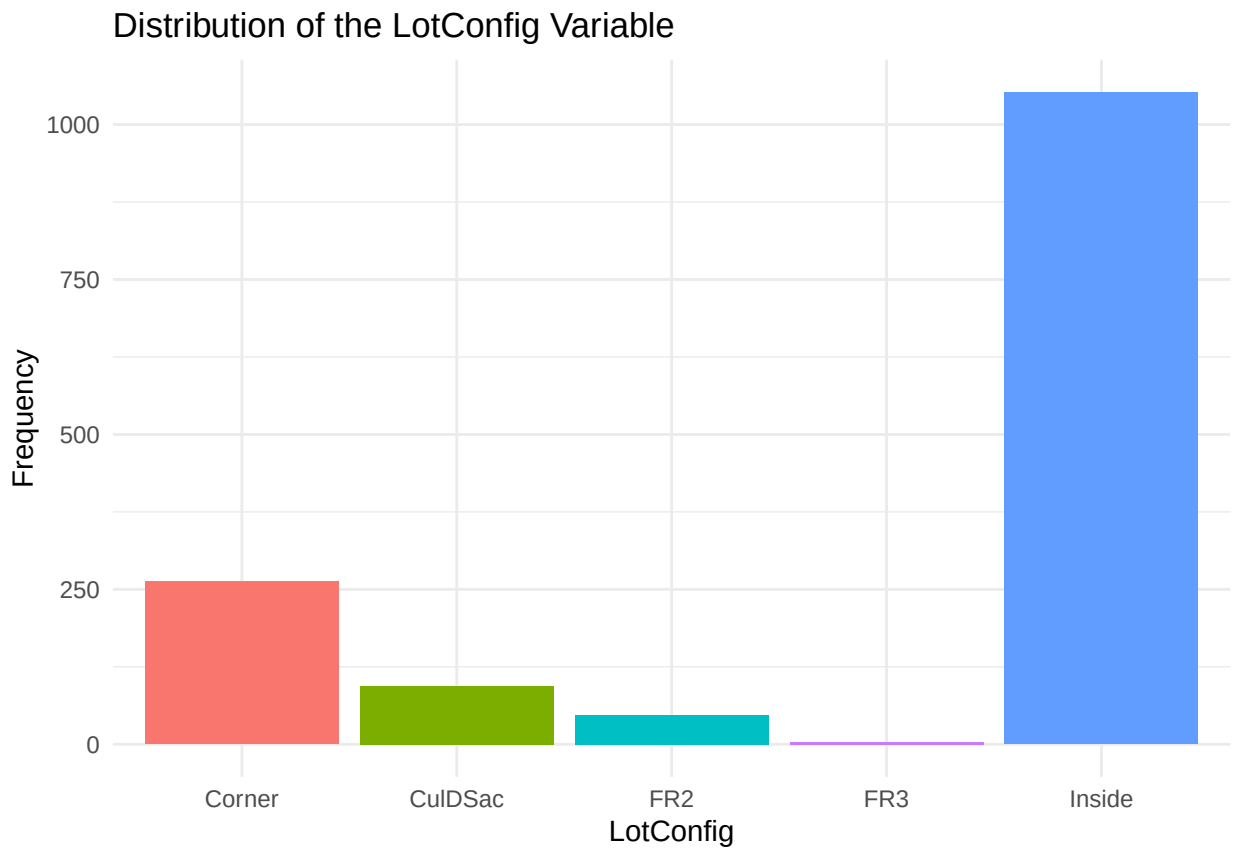
```
mode_value <- as.character(table(house$Utilities))[which.max(table(house$Utilities))]
house$Utilities[is.na(house$Utilities)] <- mode_value

sum(is.na(house$Utilities))
```

```
## [1] 0
```

Removed 2 rows which had NA values for this variable

12.LotConfig: Lot configuration

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","blue")

ggplot(house, aes(x = LotConfig, fill = LotConfig)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the LotConfig Variable",
       x = "LotConfig",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Encoding categories:

```
factor_var <- factor(house$LotConfig, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$LotConfig = numeric_labels
print(house$LotConfig[1:20])
```

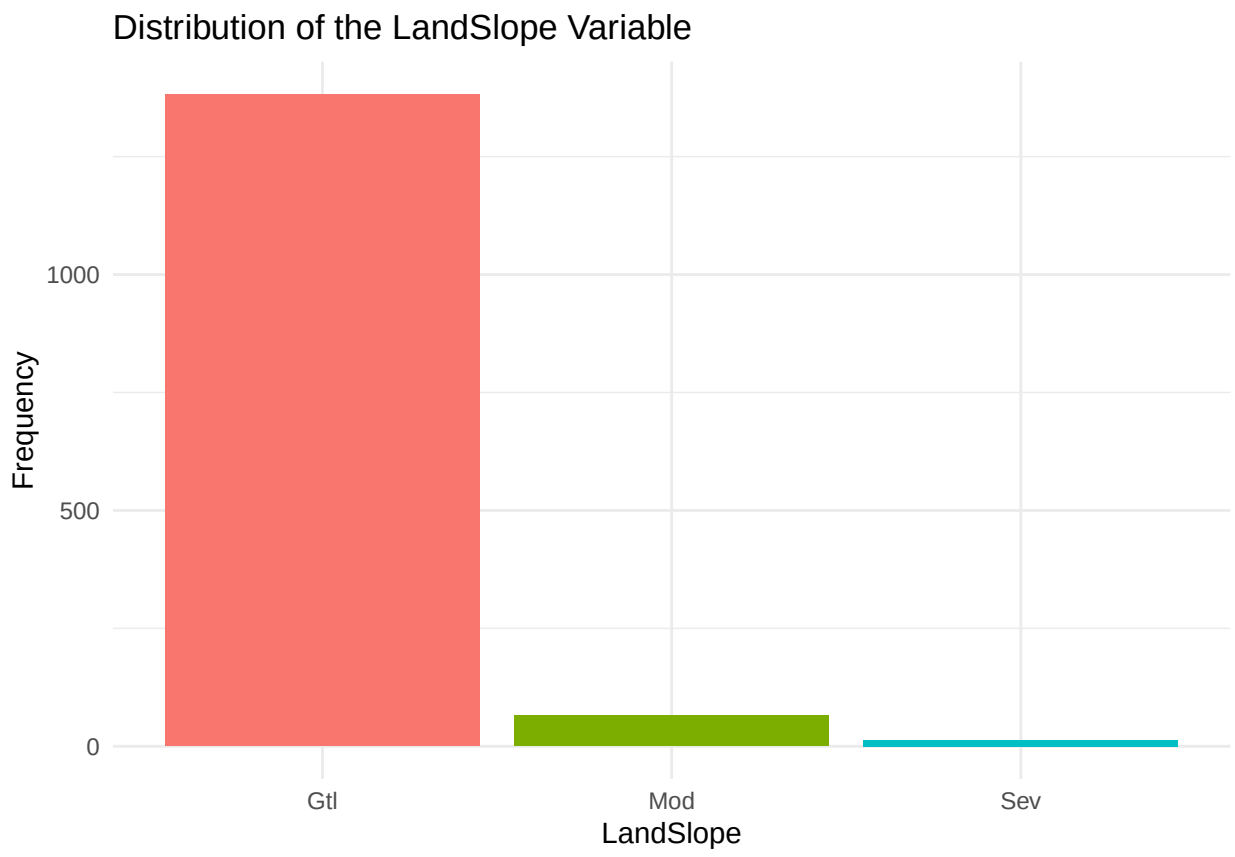
```
##  [1] 5 3 5 1 3 5 5 1 5 1 5 5 5 5 1 1 2 5 5 5
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Corner"  "CulDSac" "FR2"     "FR3"     "Inside"
```

13.LandSlope Slope of property Gtl Gentle slope Mod Moderate Slope Sev Severe Slope

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue")

ggplot(house, aes(x = LandSlope, fill = LandSlope)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the LandSlope Variable",
       x = "LandSlope",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



```r
factor_var <- factor(house$LandSlope, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$LandSlope = numeric_labels
print(house$LandSlope[1:20])
```

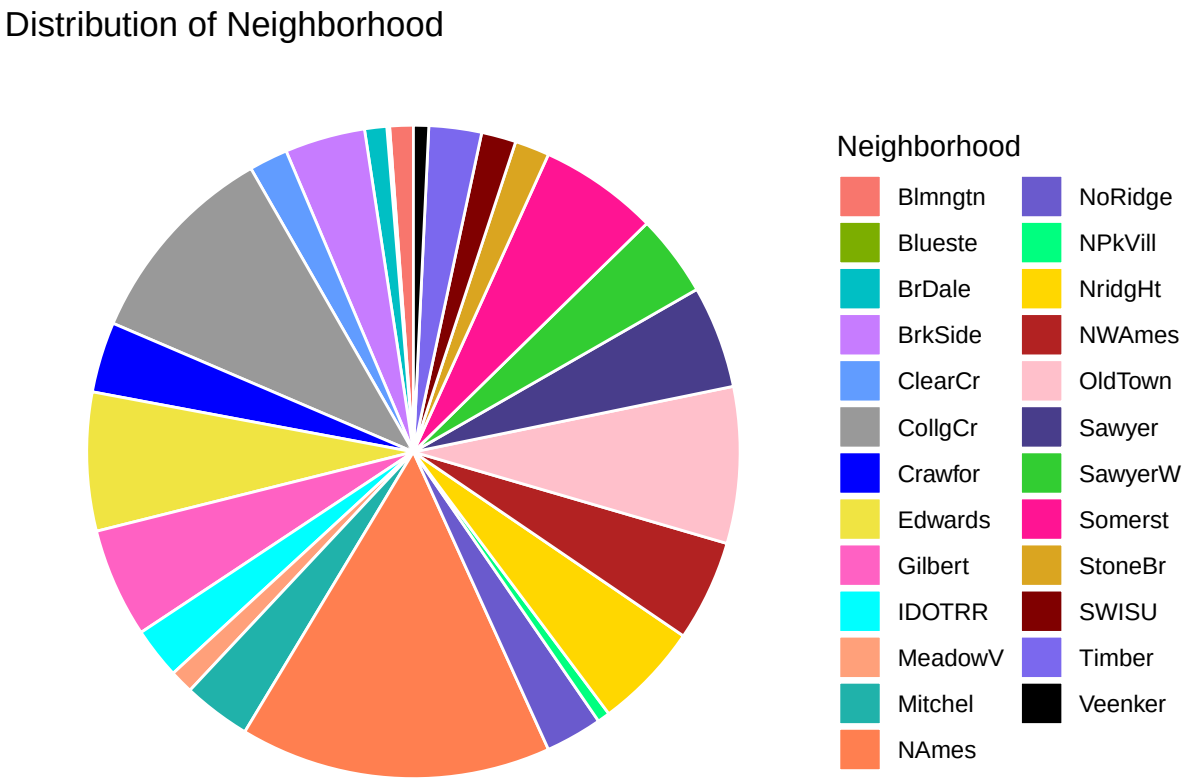
```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```r
categories <- levels(factor_var)
categories
```

```
## [1] "Gtl" "Mod" "Sev"
```

14. Neighborhood

```r
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F

ggplot(house, aes(x = "", fill = Neighborhood)) +
  geom_bar(width = 1, color = "white") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of Neighborhood",
       fill = "Neighborhood") +
  theme_void()
```



Encoding labels

```
factor_var <- factor(house$Neighborhood, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Neighborhood = numeric_labels
print(house$Neighborhood[1:10])
```

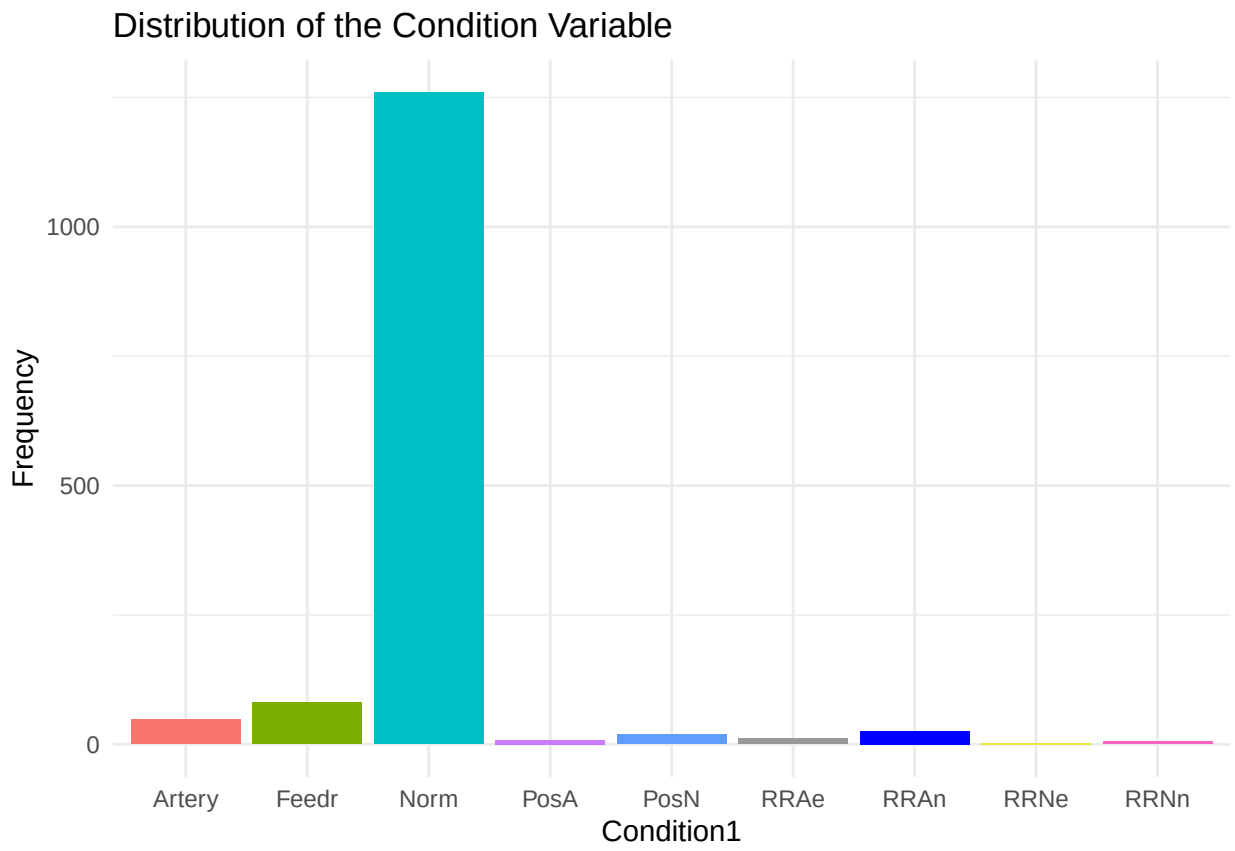
```
##  [1]  6 25  6  7 14 12 21 17 18  4
```

```
categories <- levels(factor_var)
categories
```

```
##  [1] "Blmngtn" "Blueste" "BrDale"  "BrkSide" "ClearCr" "CollgCr" "Crawfor"
##  [8] "Edwards" "Gilbert" "IDOTRR"  "MeadowV" "Mitchel" "NAmes"   "NoRidge"
## [15] "NPkVill" "NridgHt" "NWAmes"  "OldTown" "Sawyer"  "SawyerW" "Somerst"
## [22] "StoneBr" "SWISU"   "Timber"  "Veenker"
```

15. Condition1 Proximity to various conditions

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F

ggplot(house, aes(x = Condition1, fill = Condition1)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Condition Variable",
       x = "Condition1",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Encoding labels

```
factor_var <- factor(house$Condition1, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Condition1 = numeric_labels
print(house$Condition1[1:10])
```

```
##  [1] 3 2 3 3 3 3 3 5 1 1
```

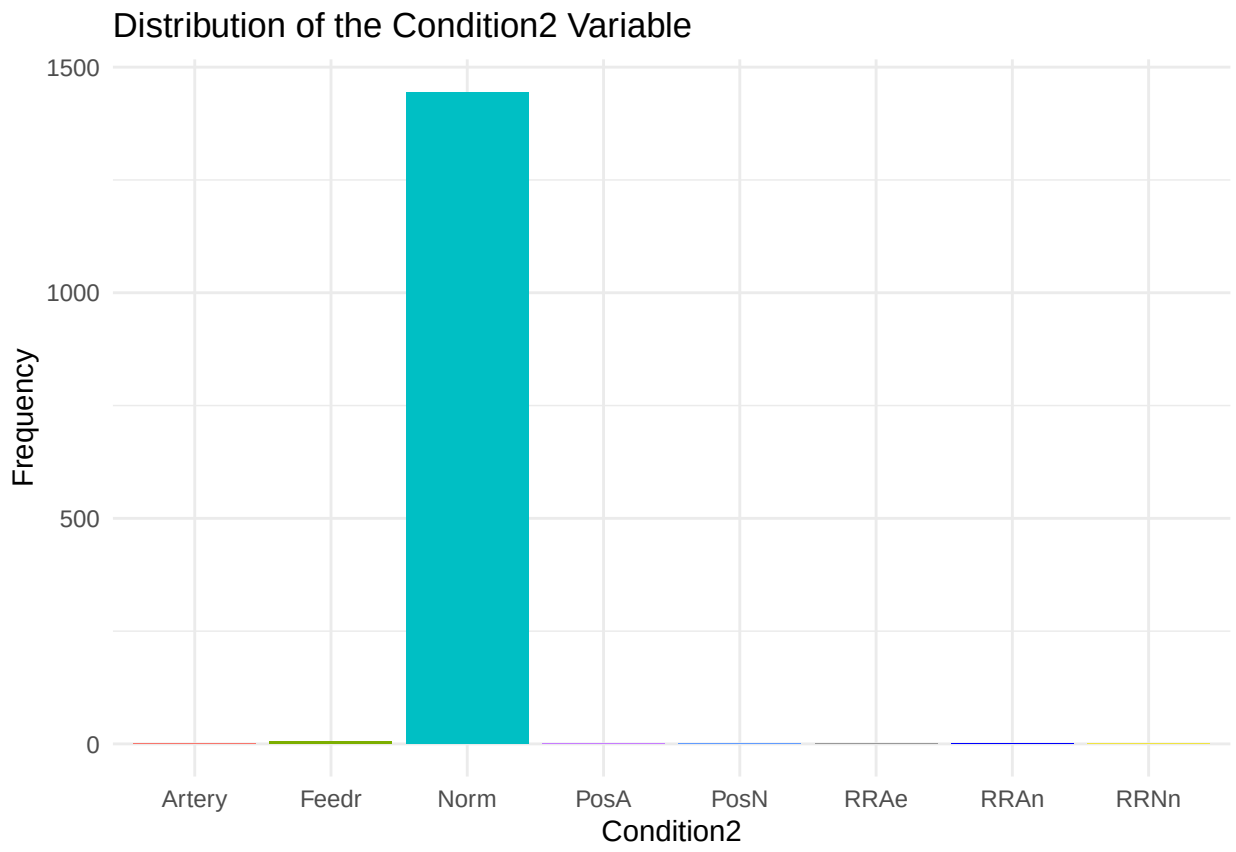
```
categories <- levels(factor_var)
categories
```

```
## [1] "Artery" "Feedr"  "Norm"   "PosA"   "PosN"   "RRAe"   "RRAn"   "RRNe"
## [9] "RRNn"
```

16. Condition2 Proximity to various conditions (if more than one is present)

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F

ggplot(house, aes(x = Condition2, fill = Condition2)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Condition2 Variable",
       x = "Condition2",
```

```
      y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$Condition2, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Condition2 = numeric_labels
print(house$Condition2[1:10])
```

```
##  [1] 3 3 3 3 3 3 3 3 3 1
```

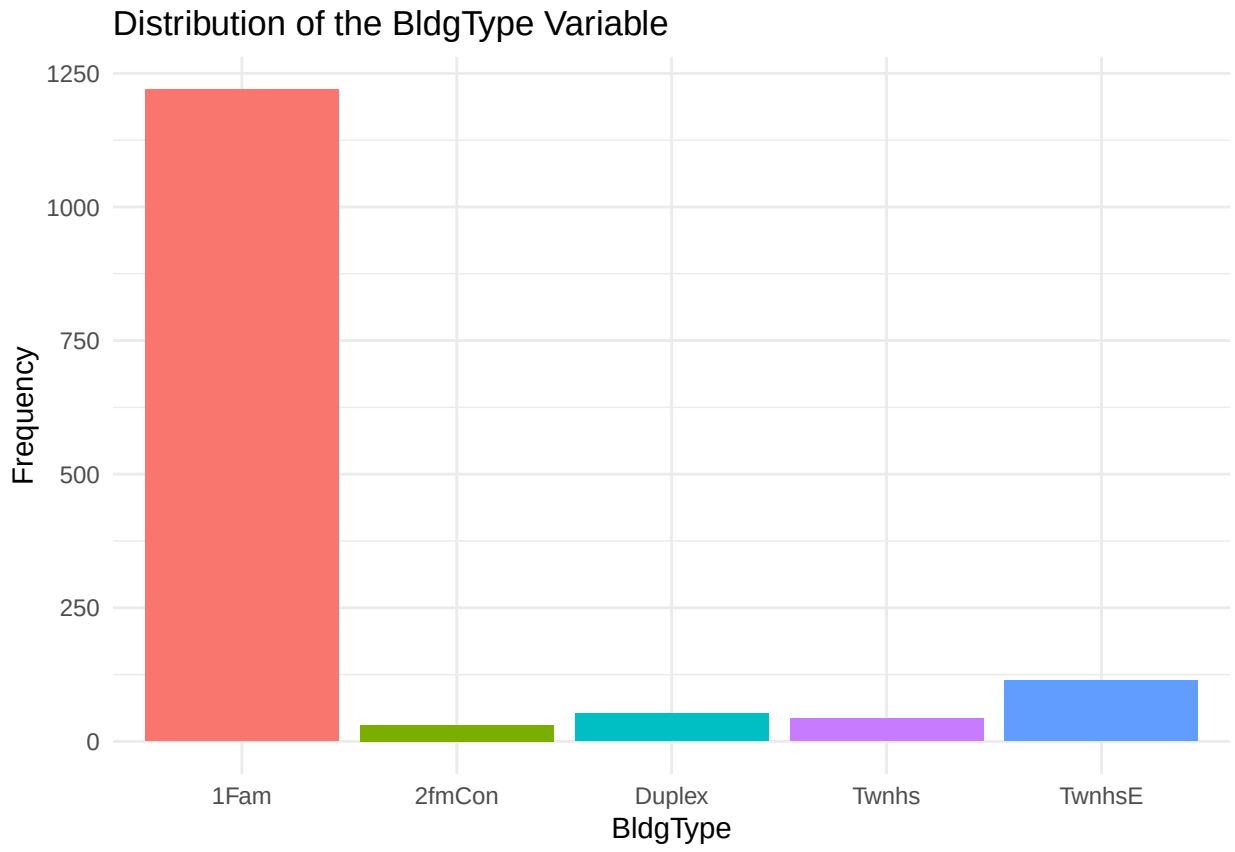
```
categories <- levels(factor_var)
categories
```

```
## [1] "Artery" "Feedr"  "Norm"   "PosA"   "PosN"   "RRAe"   "RRAn"   "RRNn"
```

17. BldgType Type of dwelling

```
library(ggplot2)

my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999")

ggplot(house, aes(x = BldgType, fill = BldgType)) +
```

```
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the BldgType Variable",
       x = "BldgType",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$BldgType, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$BldgType = numeric_labels
print(house$BldgType[1:10])
```

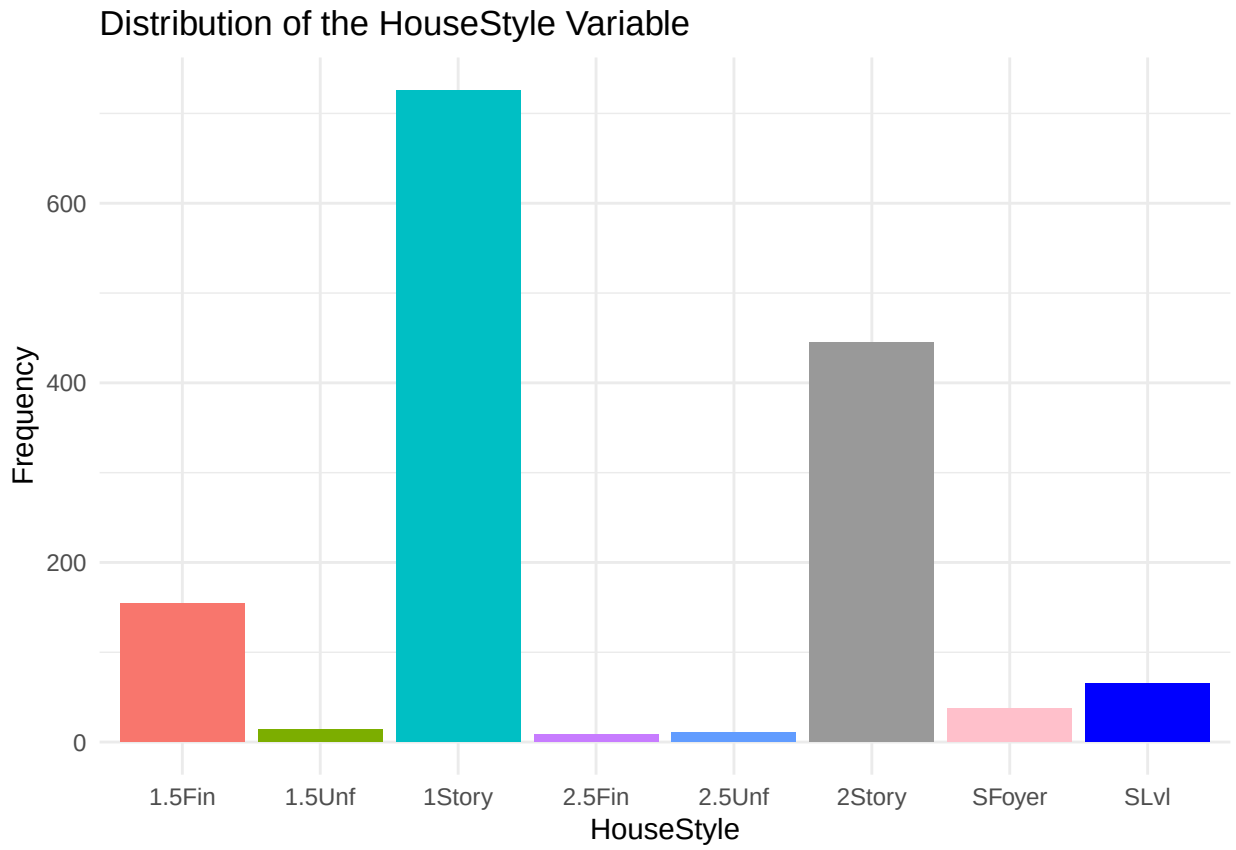
```
##  [1] 1 1 1 1 1 1 1 1 1 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "1Fam"   "2fmCon" "Duplex" "Twnhs"  "TwnhsE"
```

18 HouseStyle Style of dwelling

```
library(ggplot2)

my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink","blue")

ggplot(house, aes(x = HouseStyle, fill = HouseStyle)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the HouseStyle Variable",
       x = "HouseStyle",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$HouseStyle, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$HouseStyle = numeric_labels
print(house$HouseStyle[1:10])
```

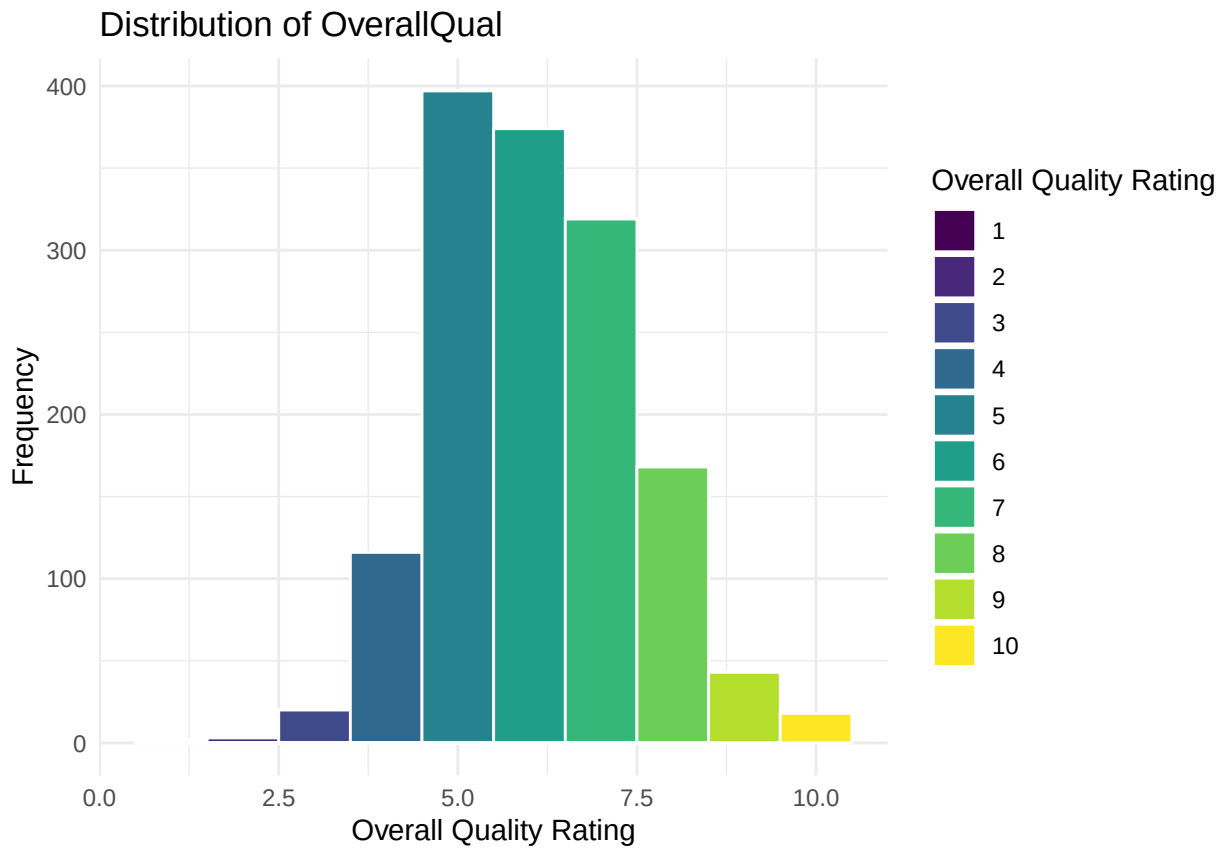
```
##  [1] 6 3 6 6 6 1 3 6 1 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "1.5Fin" "1.5Unf" "1Story" "2.5Fin" "2.5Unf" "2Story" "SFoyer" "SLvl"
```

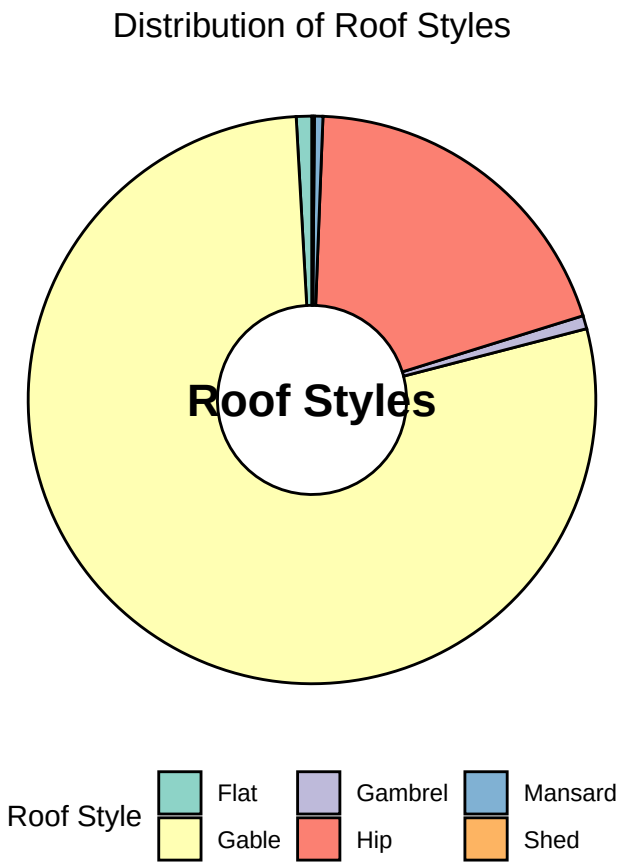
19. OverallQual Rates the overall material and finish of the house

```
ggplot(house, aes(x = OverallQual, fill = factor(OverallQual))) +
  geom_histogram(color = "white", binwidth = 1) +
  scale_fill_viridis_d() +
  labs(title = "Distribution of OverallQual",
       x = "Overall Quality Rating",
       y = "Frequency",
       fill = "Overall Quality Rating") +
  theme_minimal()
```



20. RoofStyle Type of roof

```
ggplot(house, aes(x = "", fill = RoofStyle)) +
  geom_bar(width = 1, color = "black") +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Set3") +
  labs(title = "Distribution of Roof Styles",
       fill = "Roof Style") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "bottom") +
  annotate("text", x = 0, y = 0, label = "Roof Styles", size = 6, fontface = "bold")
```




```
factor_var <- factor(house$RoofStyle, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$RoofStyle = numeric_labels
print(house$RoofStyle[1:10])
```

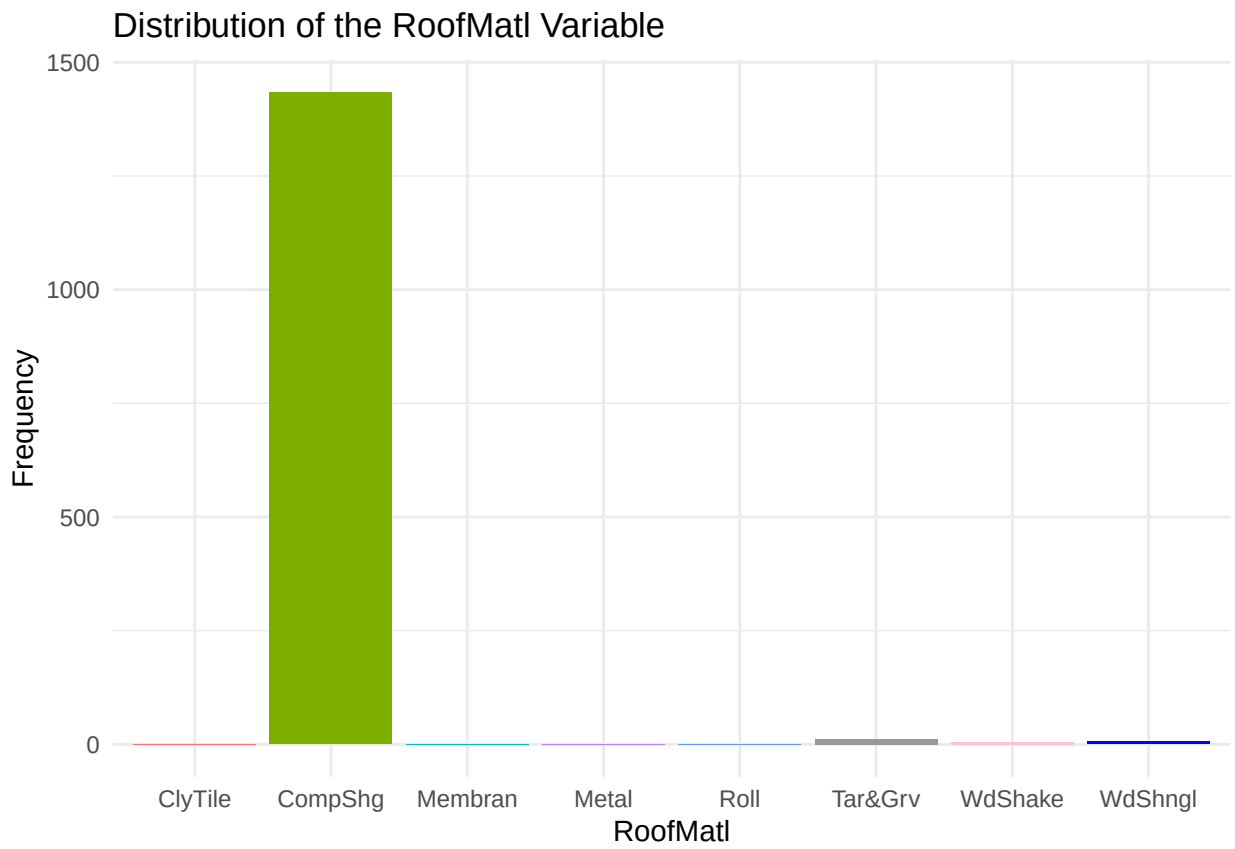
```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Flat"    "Gable"   "Gambrel" "Hip"     "Mansard" "Shed"
```

21. RoofMatl

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink","blue")

ggplot(house, aes(x = RoofMatl, fill = RoofMatl)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the RoofMatl Variable",
       x = "RoofMatl",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$RoofMatl, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$RoofMatl = numeric_labels
print(house$RoofMatl[1:10])
```

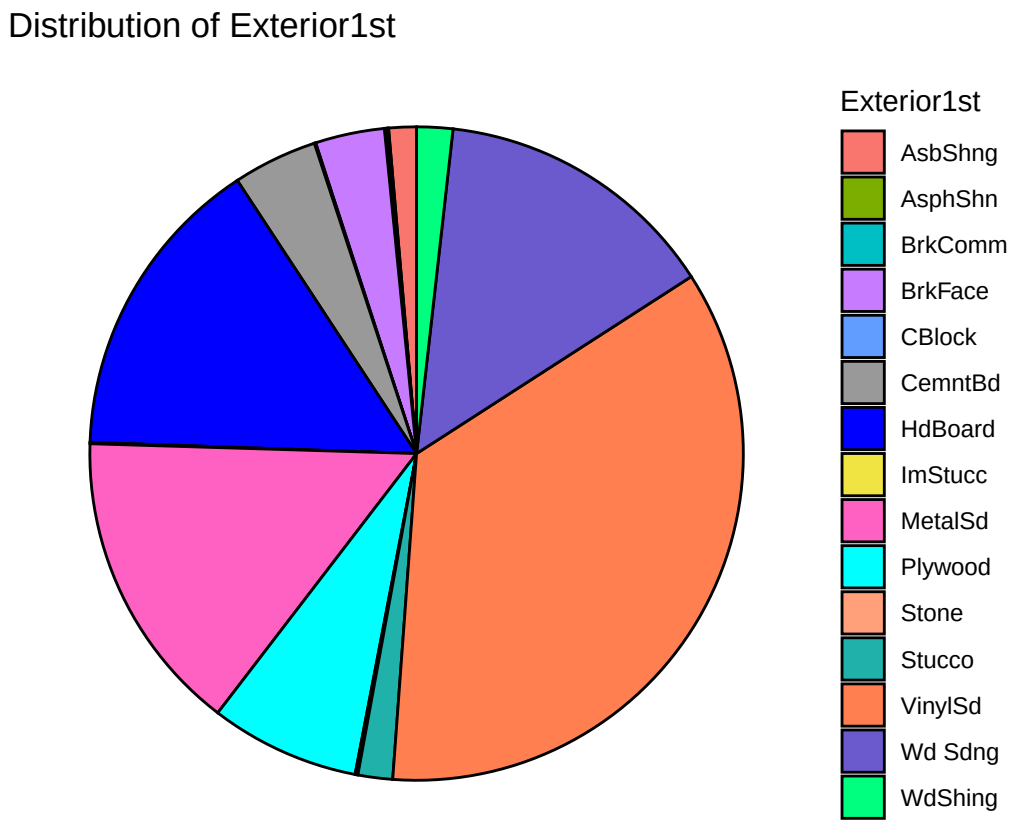
```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "ClyTile" "CompShg" "Membran" "Metal"   "Roll"    "Tar&Grv" "WdShake"
## [8] "WdShngl"
```

22. Exterior1st Exterior covering on house

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = "", fill = Exterior1st)) +
  geom_bar(width = 1, color = "black") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of Exterior1st",
       fill = "Exterior1st") +
  theme_void()
```




```
factor_var <- factor(house$Exterior1st, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Exterior1st = numeric_labels
print(house$Exterior1st[1:10])
```

```
##  [1] 13  9 13 14 13 13 13  7  4  9
```

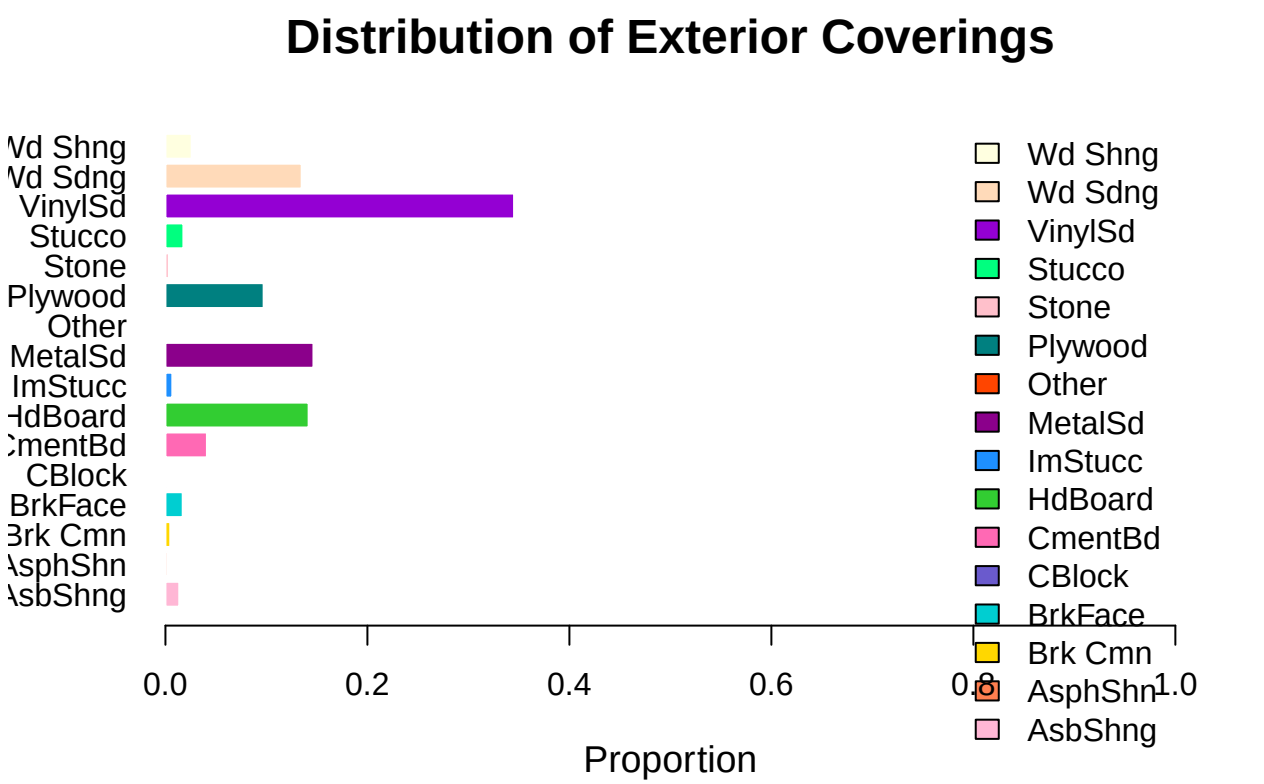
```
categories <- levels(factor_var)
categories
```

```
##  [1] "AsbShng" "AsphShn" "BrkComm" "BrkFace" "CBlock"  "CemntBd" "HdBoard"
##  [8] "ImStucc" "MetalSd" "Plywood" "Stone"   "Stucco"  "VinylSd" "Wd Sdng"
## [15] "WdShing"
```

23. Exterior2nd Exterior covering on house (if more than one material)

```
ext_counts <- table(house$Exterior2nd)
ext_props <- prop.table(ext_counts)
ext_names <- names(ext_counts)
ext_colors <- c("#FFB7D5", "#FF7F50", "#FFD700", "#00CED1", "#6A5ACD", "#FF69B4",
                "#32CD32", "#1E90FF", "#8B008B", "#FF4500", "#008080", "#FFC0CB",
                "#00FF7F", "#9400D3", "#FFDAB9", "#FFFFE0", "#F0E68C")

barplot(ext_props, horiz = TRUE, col = ext_colors, border = "white",
```

```
        main = "Distribution of Exterior Coverings", xlab = "Proportion",
        xlim = c(0, 1), las = 1, cex.main = 1.5, cex.lab = 1.2,
        legend.text = ext_names, args.legend = list(x = "topright", bty = "n"))
```




```
factor_var <- factor(house$Exterior2nd, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Exterior2nd = numeric_labels
print(house$Exterior2nd[1:10])
```

```
##  [1] 14  9 14 16 14 14 14  7 16  9
```

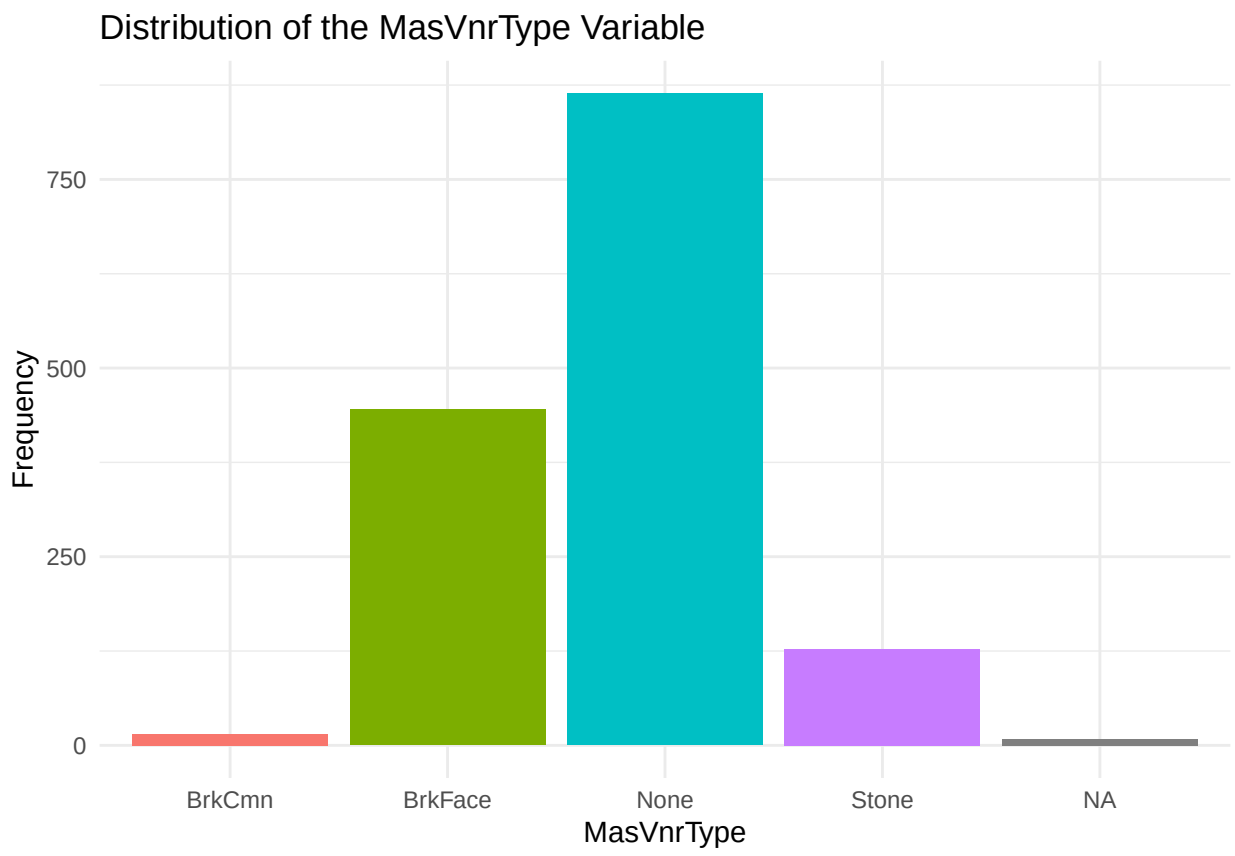
```
categories <- levels(factor_var)
categories
```

```
##  [1] "AsbShng" "AsphShn" "Brk Cmn" "BrkFace" "CBlock"  "CmentBd" "HdBoard"
##  [8] "ImStucc" "MetalSd" "Other"   "Plywood" "Stone"   "Stucco"  "VinylSd"
## [15] "Wd Sdng" "Wd Shng"
```

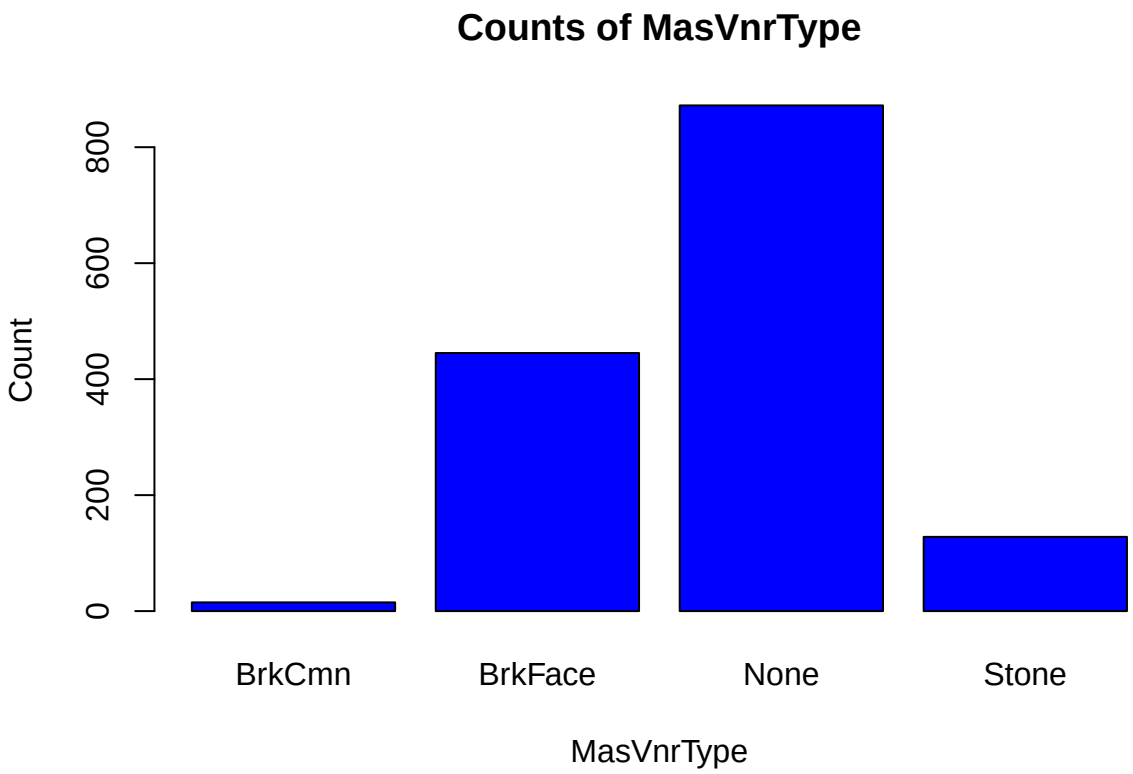
24. MasVnrType Masonry veneer type

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")

ggplot(house, aes(x = MasVnrType, fill = MasVnrType)) +
```

```
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the MasVnrType Variable",
       x = "MasVnrType",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
house$MasVnrType <- ifelse(is.na(house$MasVnrType), "None", house$MasVnrType)
mvt_counts <- table(house$MasVnrType)

barplot(mvt_counts, col = "blue", main = "Counts of MasVnrType",
        xlab = "MasVnrType", ylab = "Count")
```



Changed the NA to already existing None category in the variable

```
factor_var <- factor(house$MasVnrType, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$MasVnrType = numeric_labels
print(house$MasVnrType[1:10])
```

```
##  [1] 2 3 2 3 2 3 4 4 3 3
```

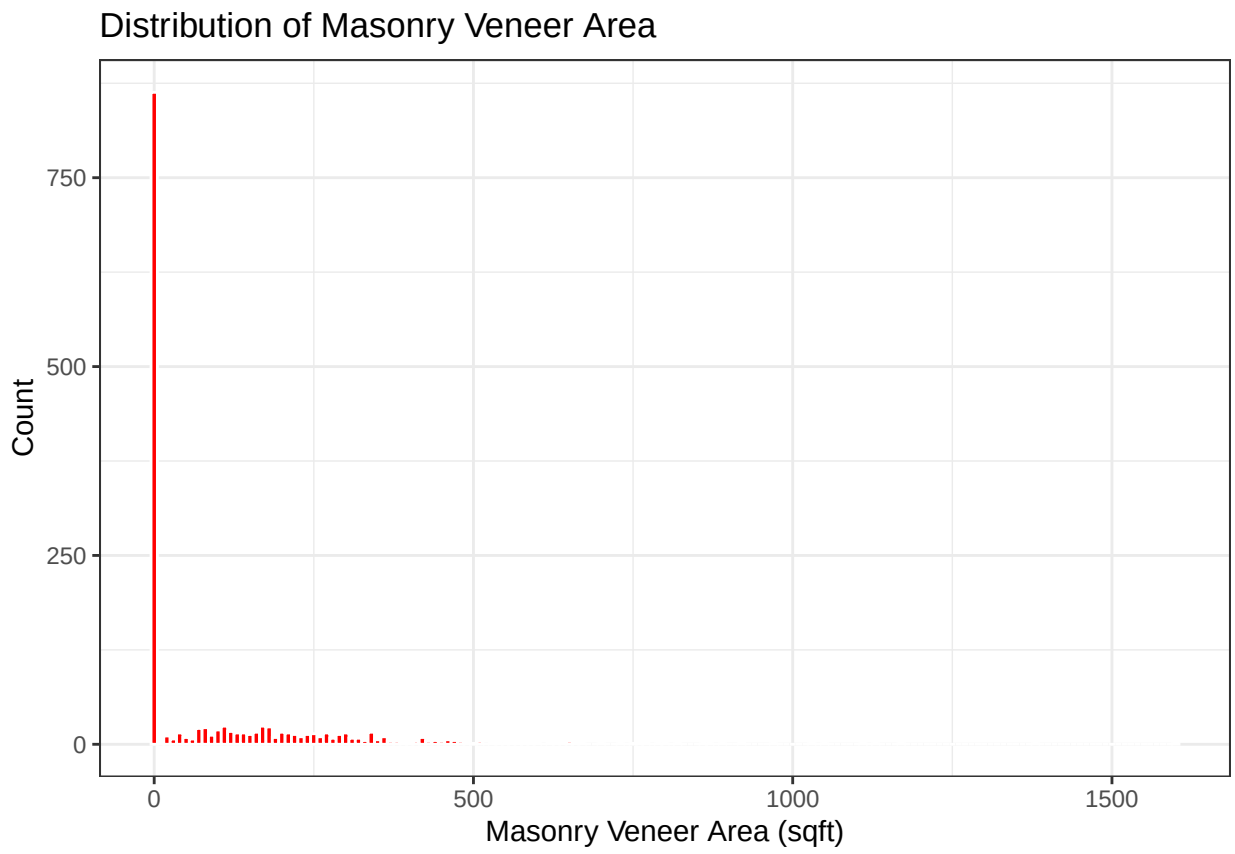
```
categories <- levels(factor_var)
categories
```

```
## [1] "BrkCmn"  "BrkFace" "None"    "Stone"
```

25. MasVnrArea Masonry veneer area in square feet

```
ggplot(house, aes(x = MasVnrArea)) +
  geom_histogram(binwidth = 10, fill = "red", color = "white") +
  labs(title = "Distribution of Masonry Veneer Area",
       x = "Masonry Veneer Area (sqft)", y = "Count") +
  theme_bw()
```

```
## Warning: Removed 8 rows containing non-finite values (‘stat_bin()‘).
```

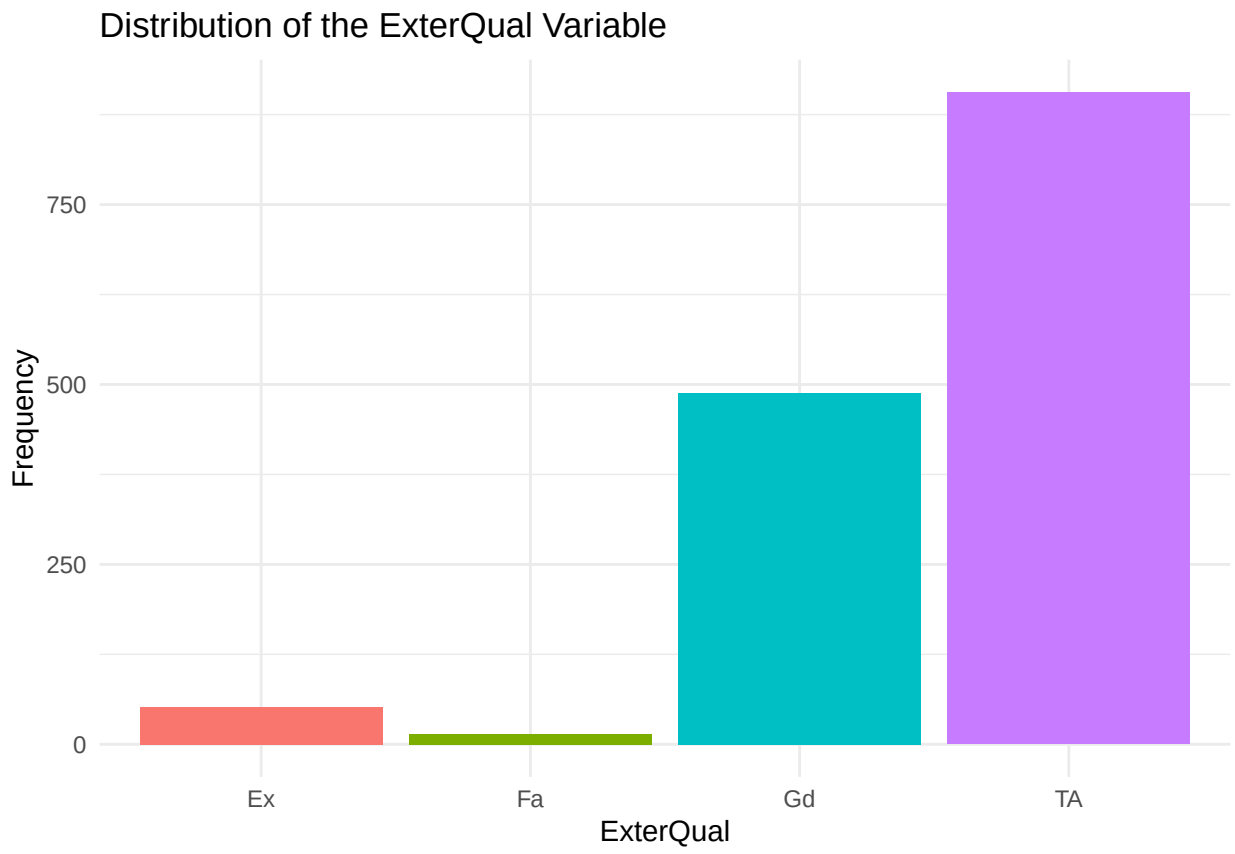



```
house$MasVnrArea <- ifelse(is.na(house$MasVnrArea), 0, house$MasVnrArea)
```

Imputed the mode that is 0 for NA values

26. ExterQual Evaluates the quality of the material on the exterior Ex Excellent Gd Good TA Average/Typical Fa Fair Po Poor

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")

ggplot(house, aes(x = ExterQual, fill = ExterQual)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the ExterQual Variable",
       x = "ExterQual",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$ExterQual, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$ExterQual = numeric_labels
print(house$ExterQual[1:10])
```

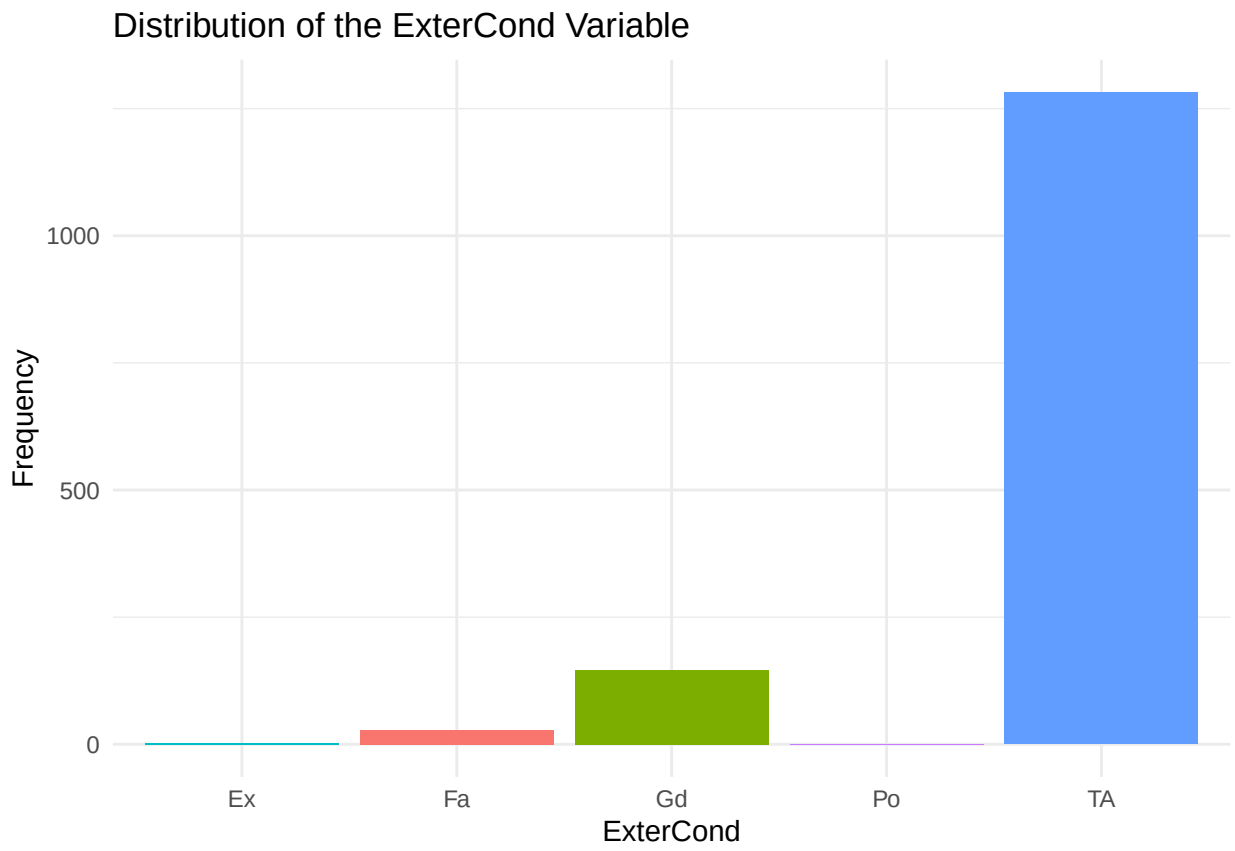
```
##  [1] 3 4 3 4 3 4 3 4 4 4
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "TA"
```

27. ExterCond Evaluates the present condition of the material on the exterior

```
my_colors <- c("#00BFC4","#F8766D", "#7CAE00",  "#C77CFF", "#619CFF")

ggplot(house, aes(x = ExterCond, fill = ExterCond)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the ExterCond Variable",
       x = "ExterCond",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$ExterCond, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$ExterCond = numeric_labels
print(house$ExterCond[1:10])
```

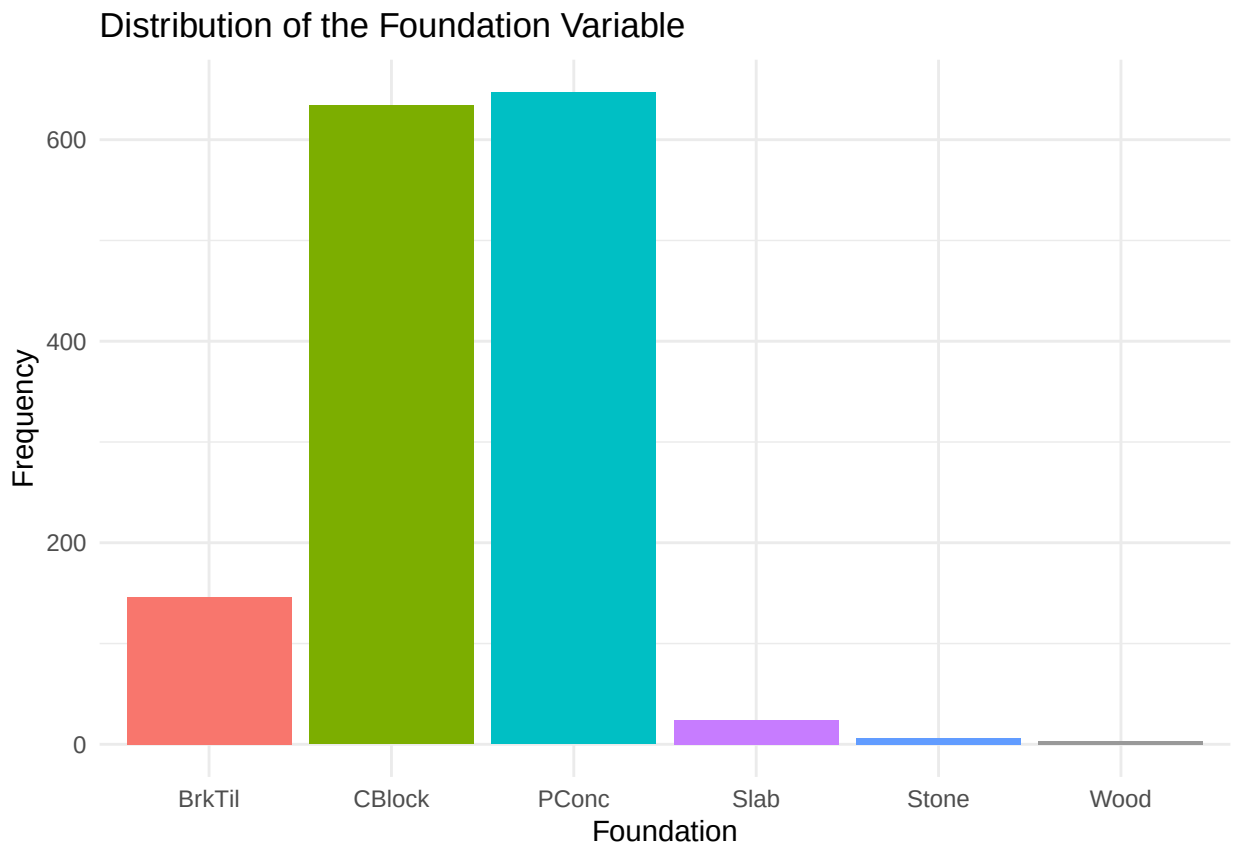
```
##  [1] 5 5 5 5 5 5 5 5 5 5
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "Po" "TA"
```

28. Foundation Type of foundation

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")

ggplot(house, aes(x = Foundation, fill = Foundation)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Foundation Variable",
       x = "Foundation",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$Foundation, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Foundation = numeric_labels
print(house$Foundation[1:10])
```

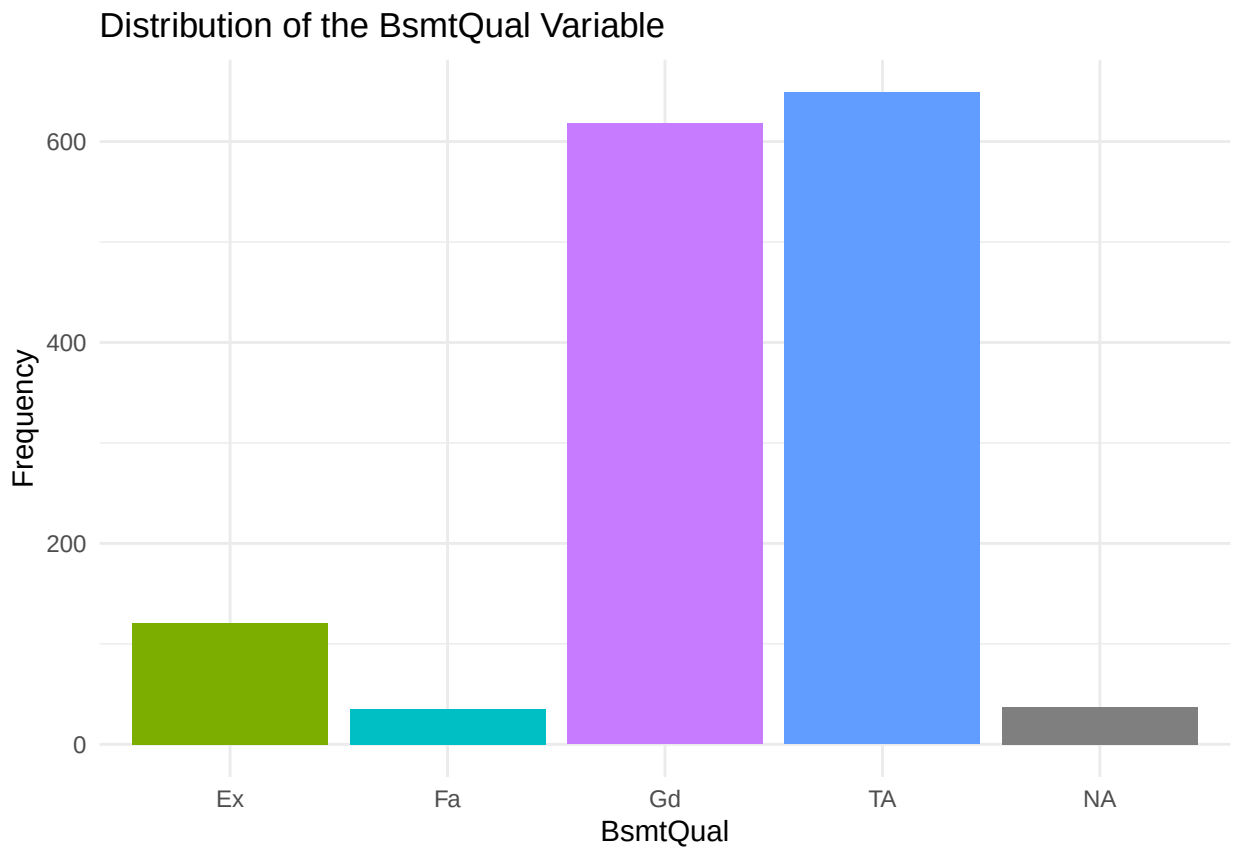
```
##  [1] 3 2 3 1 3 6 3 2 1 1
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "BrkTil" "CBlock" "PConc"  "Slab"   "Stone"  "Wood"
```

29. BsmtQual Evaluates the height of the basement

```
my_colors <- c( "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF","pink")

ggplot(house, aes(x = BsmtQual, fill = BsmtQual)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the BsmtQual Variable",
       x = "BsmtQual",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Should not convert the NA values as they represent no basement

```
factor_var <- factor(house$BsmtQual, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$BsmtQual = numeric_labels
print(house$BsmtQual[1:10])
```

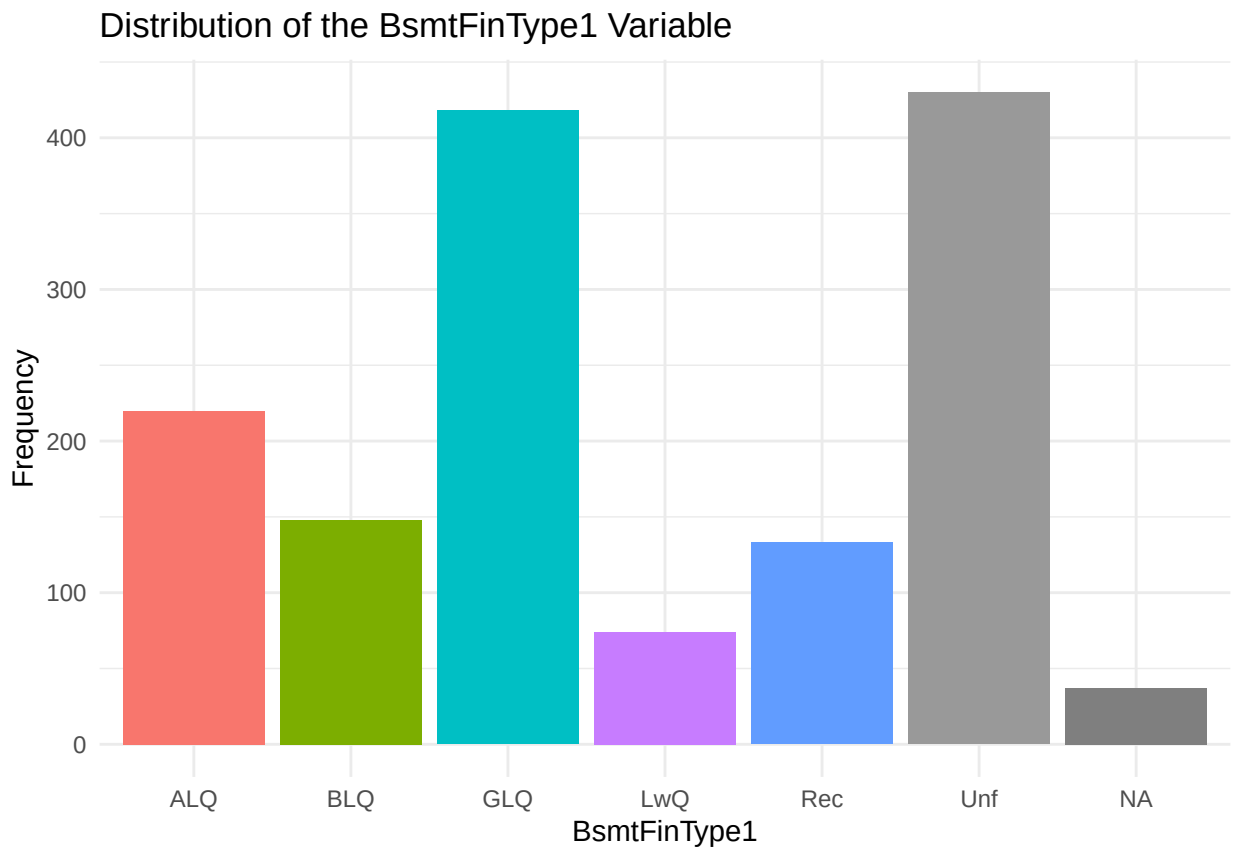
```
##  [1] 3 3 3 4 3 3 1 3 4 4
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "TA" NA
```

30. BsmtFinType1 Rating of basement finished area

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")
ggplot(house, aes(x = BsmtFinType1, fill = BsmtFinType1)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the BsmtFinType1 Variable",
       x = "BsmtFinType1",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Should not convert the NA values as they represent no basement

```
factor_var <- factor(house$BsmtFinType1, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$BsmtFinType1 = numeric_labels
print(house$BsmtFinType1[1:10])
```

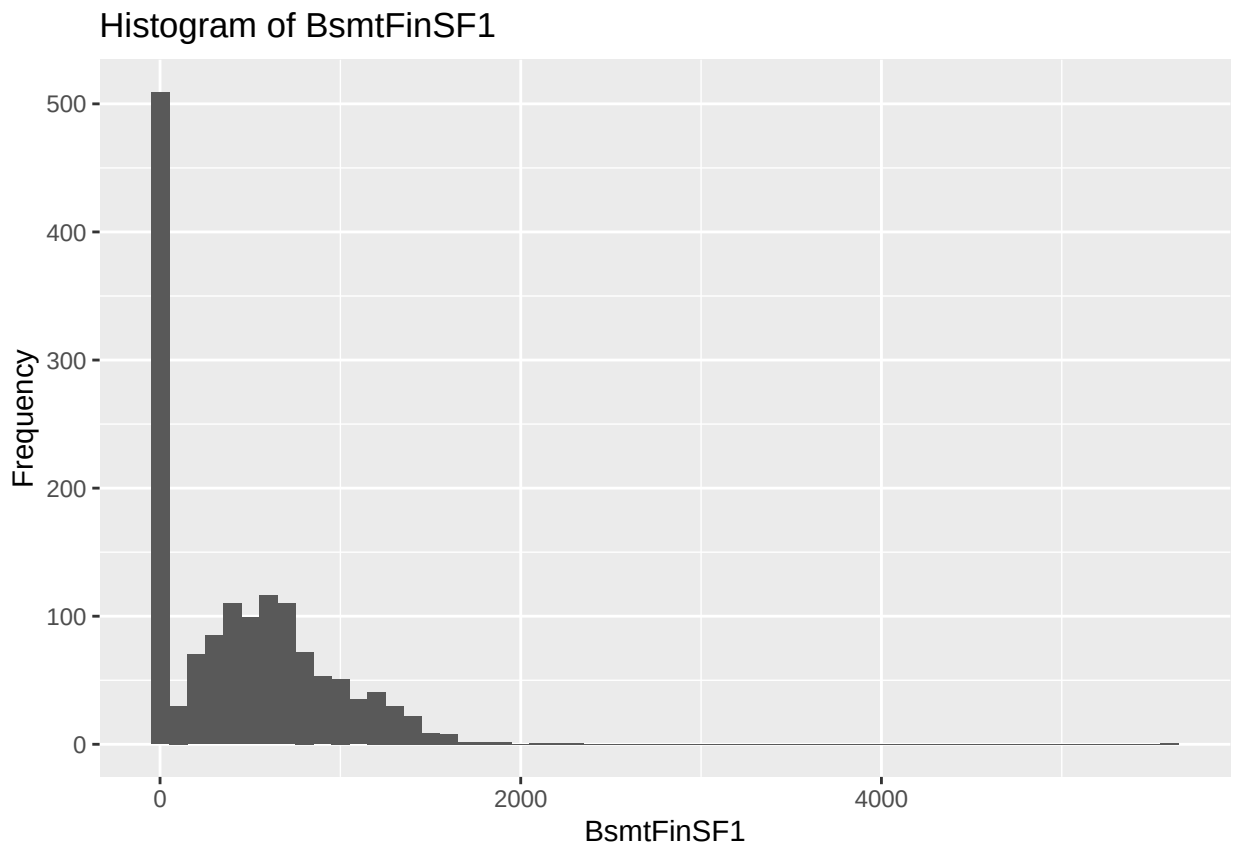
```
##  [1] 3 1 3 1 3 3 3 1 6 3
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf" NA
```

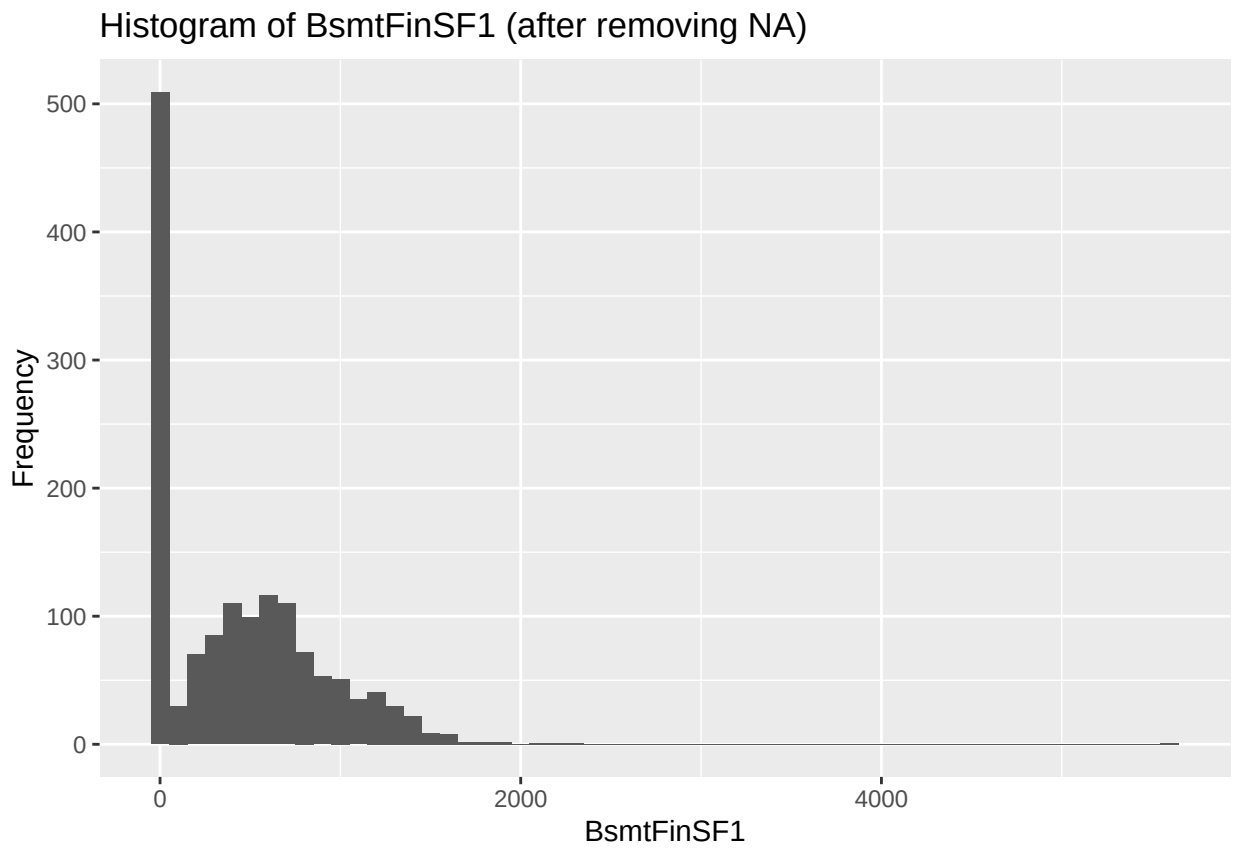
31. BsmtFinSF1 Type 1 finished square feet

```
ggplot(house, aes(x=BsmtFinSF1)) +
  geom_histogram(binwidth = 100) +
  labs(x = "BsmtFinSF1", y = "Frequency") +
  ggtitle("Histogram of BsmtFinSF1")
```



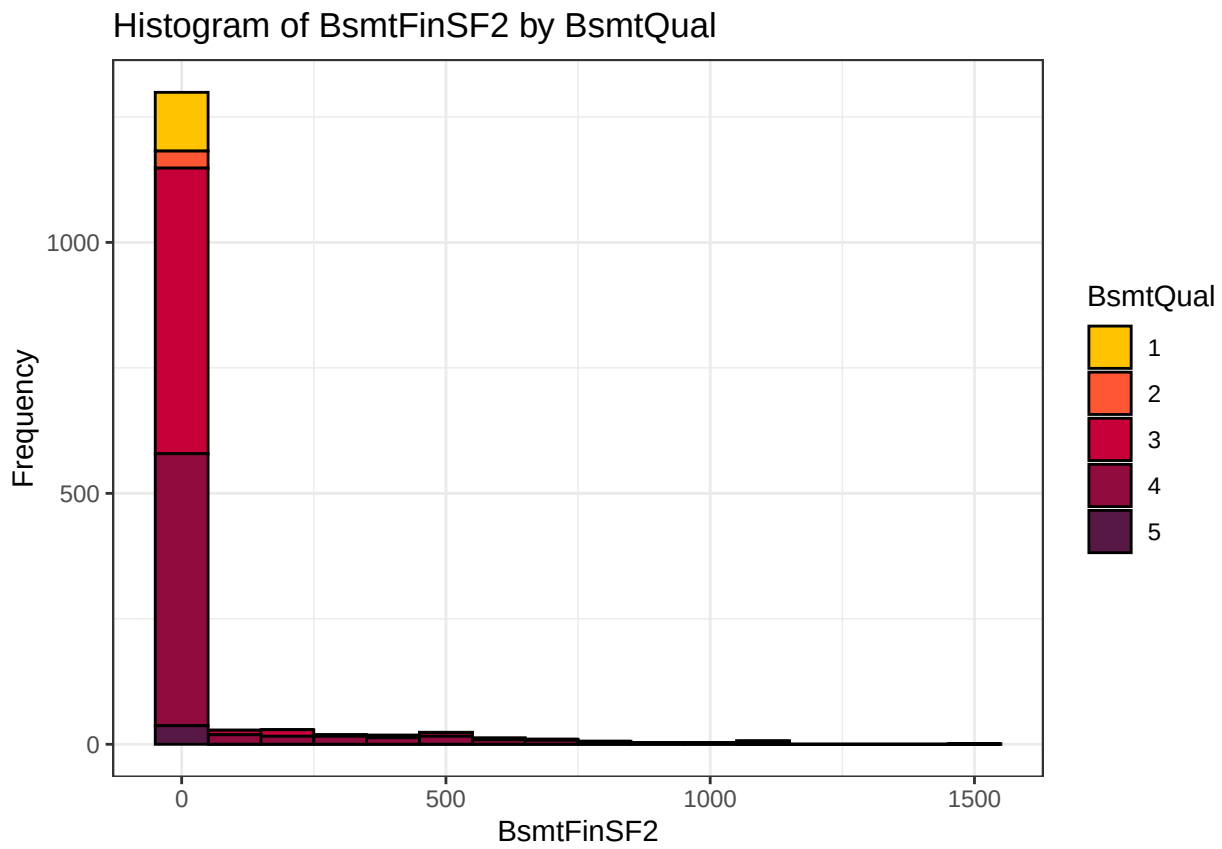
Removed the only row with single NA value

```
house$BsmtFinSF1[is.na(house$BsmtFinSF1)] <- mean(house$BsmtFinSF1, na.rm = TRUE)
ggplot(house, aes(x=BsmtFinSF1)) +
  geom_histogram(binwidth = 100) +
  labs(x = "BsmtFinSF1", y = "Frequency") +
  ggtitle("Histogram of BsmtFinSF1 (after removing NA)")
```



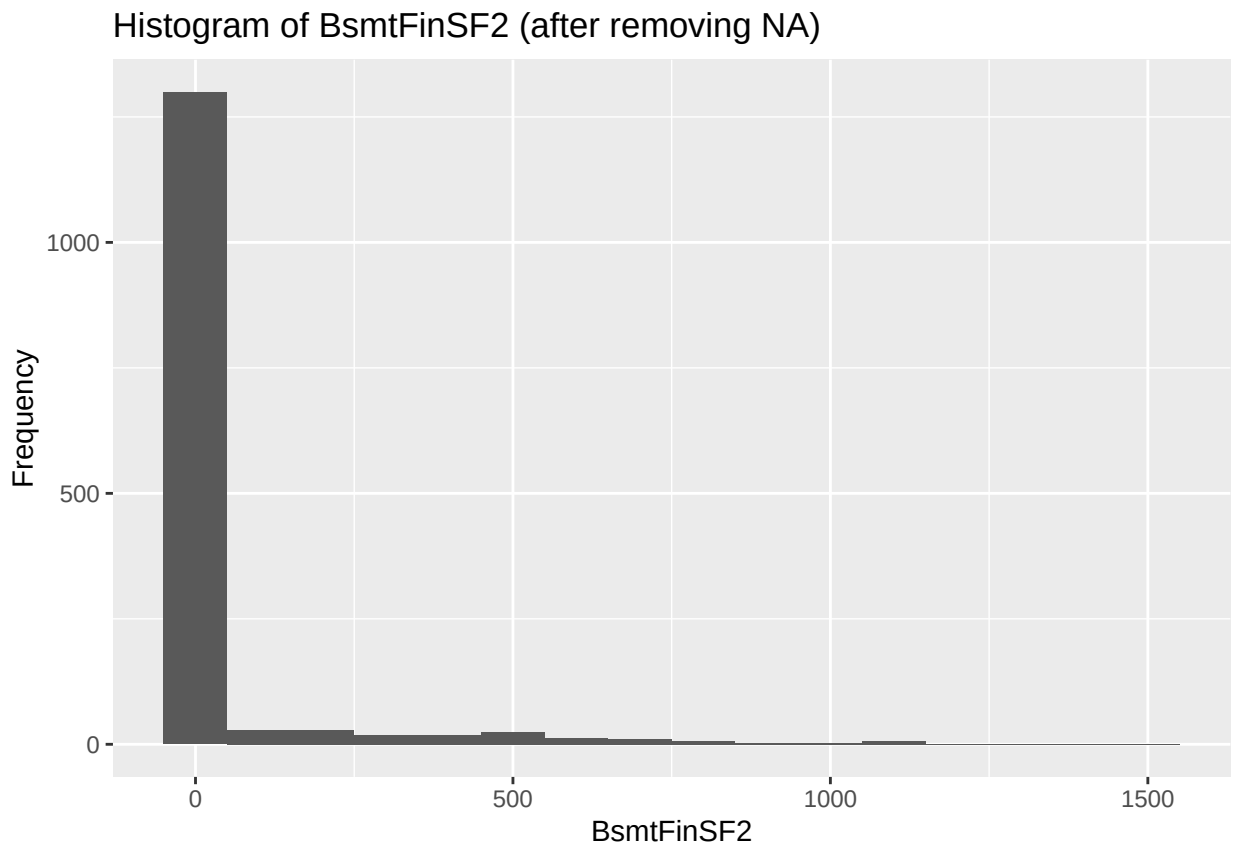
32. BsmtFinSF2 Type 2 finished square feet

```
ggplot(house, aes(x=BsmtFinSF2, fill=factor(BsmtQual))) +
  geom_histogram(binwidth = 100, color="black") +
  scale_fill_manual(values=c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845")) +
  labs(x = "BsmtFinSF2", y = "Frequency", fill = "BsmtQual") +
  ggtitle("Histogram of BsmtFinSF2 by BsmtQual") +
  theme_bw()
```



Removed the only row with single NA value

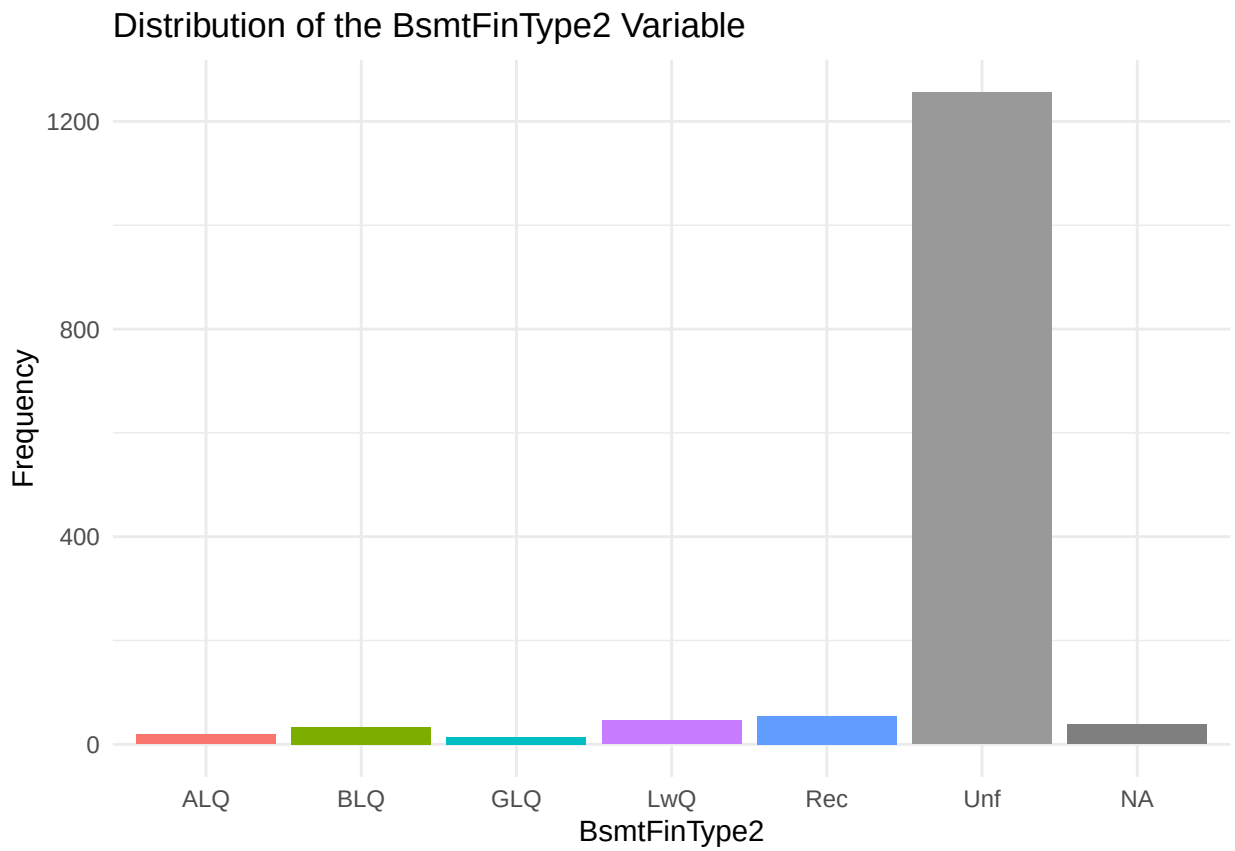
```
house$BsmtFinSF2[is.na(house$BsmtFinSF2)] <- mean(house$BsmtFinSF2, na.rm = TRUE)

ggplot(house, aes(x=BsmtFinSF2)) +
  geom_histogram(binwidth = 100) +
  labs(x = "BsmtFinSF2", y = "Frequency") +
  ggtitle("Histogram of BsmtFinSF2 (after removing NA)")
```



33. BsmtFinType2

Rating of basement finished area (if multiple types)

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")
ggplot(house, aes(x = BsmtFinType2, fill = BsmtFinType2)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the BsmtFinType2 Variable",
       x = "BsmtFinType2",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



NA represent the no basement categories and need not be removed

```
factor_var <- factor(house$BsmtFinType2, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$BsmtFinType2 = numeric_labels
print(house$BsmtFinType2[1:10])
```

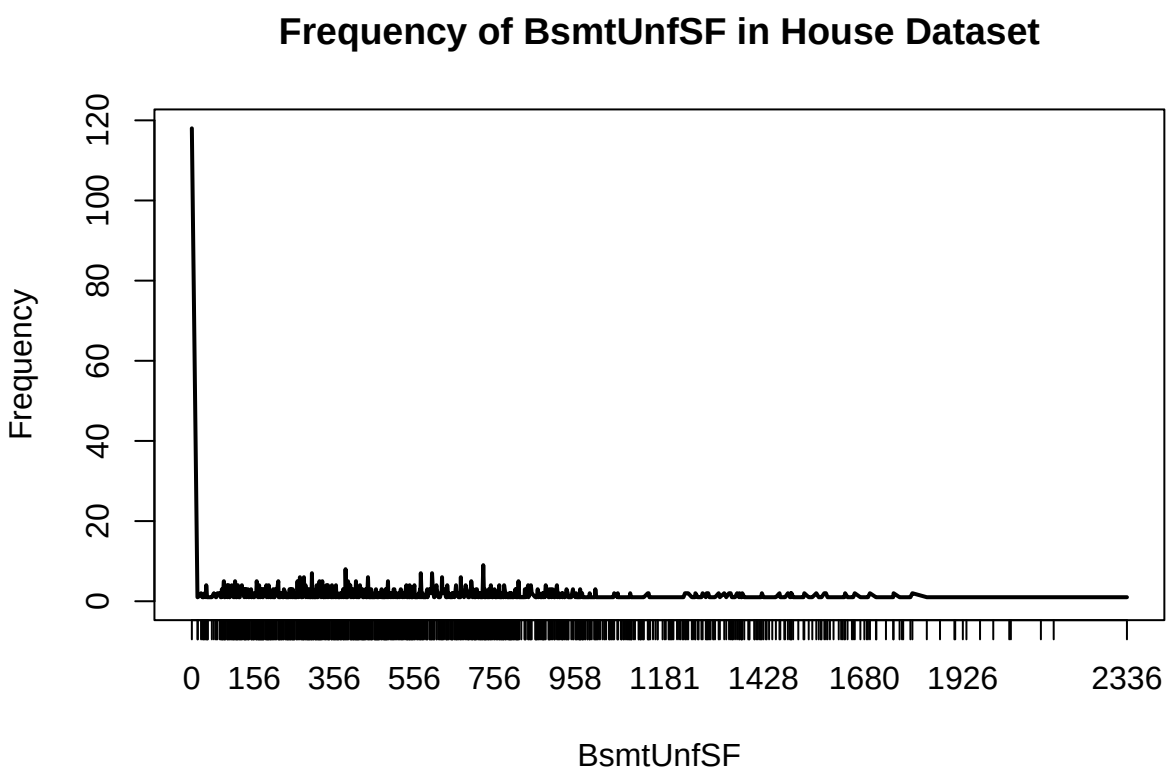
```
##  [1] 6 6 6 6 6 6 6 2 6 6
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf" NA
```

34. BsmtUnfSF Unfinished square feet of basement area

```
plot(table(house$BsmtUnfSF), type = "l",
     xlab = "BsmtUnfSF", ylab = "Frequency",
     main = "Frequency of BsmtUnfSF in House Dataset")
```




```
mean_BsmtUnfSF <- mean(house$BsmtUnfSF, na.rm = TRUE)

# Replace NA values with mean value
house$BsmtUnfSF[is.na(house$BsmtUnfSF)] <- mean_BsmtUnfSF
```

35. TotalBsmtSF Total square feet of basement area

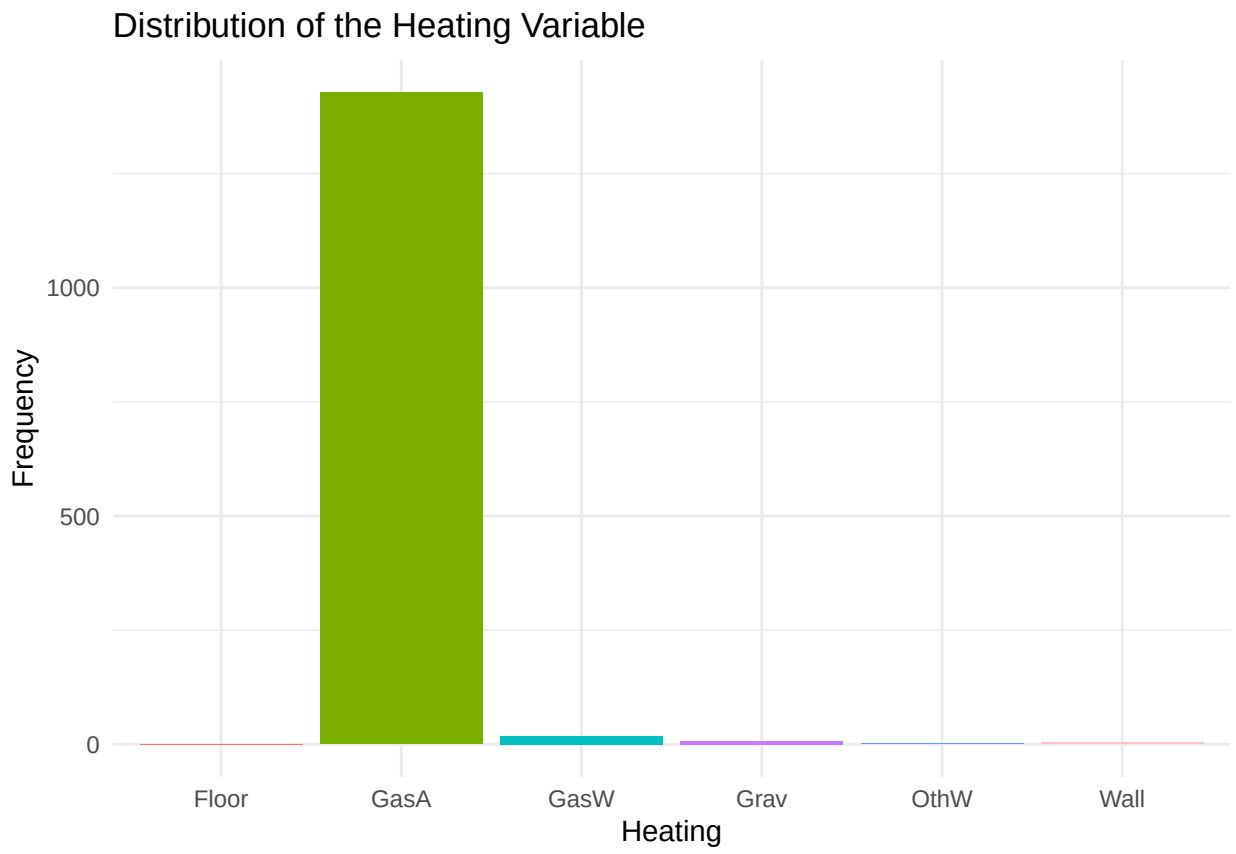
```
sum(is.na(house$TotalBsmtSF))
```

```
## [1] 0
```

36. Heating Type of heating

Floor Floor Furnace GasA Gas forced warm air furnace GasW Gas hot water or steam heat Grav Gravity furnace OthW Hot water or steam heat other than gas Wall Wall furnace

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "pink")
ggplot(house, aes(x = Heating, fill = Heating)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Heating Variable",
       x = "Heating",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$Heating, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Heating = numeric_labels
print(house$Heating[1:10])
```

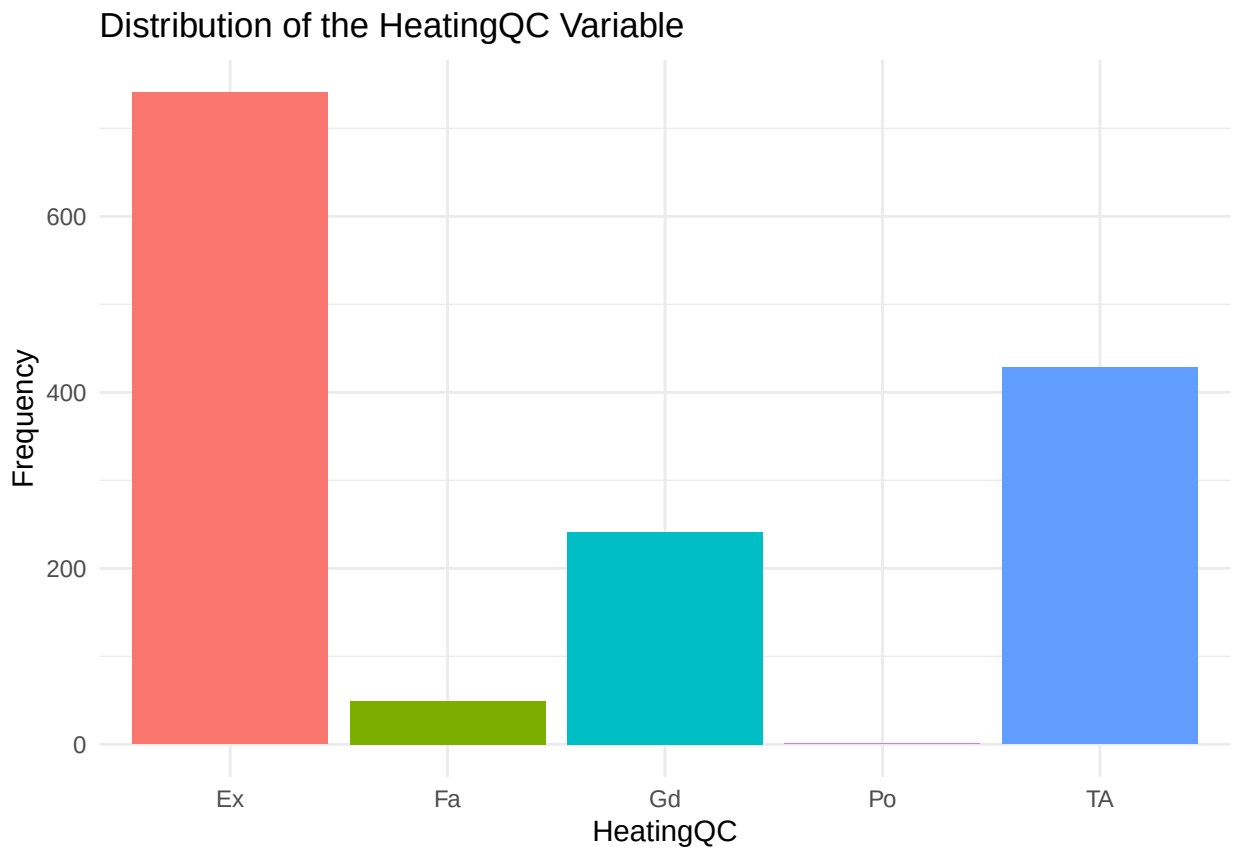
```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Floor" "GasA"  "GasW"  "Grav"  "OthW"  "Wall"
```

37. HeatingQC Heating quality and condition

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "pink")
ggplot(house, aes(x = HeatingQC, fill = HeatingQC)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the HeatingQC Variable",
       x = "HeatingQC",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$HeatingQC, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$HeatingQC = numeric_labels
print(house$HeatingQC[1:10])
```

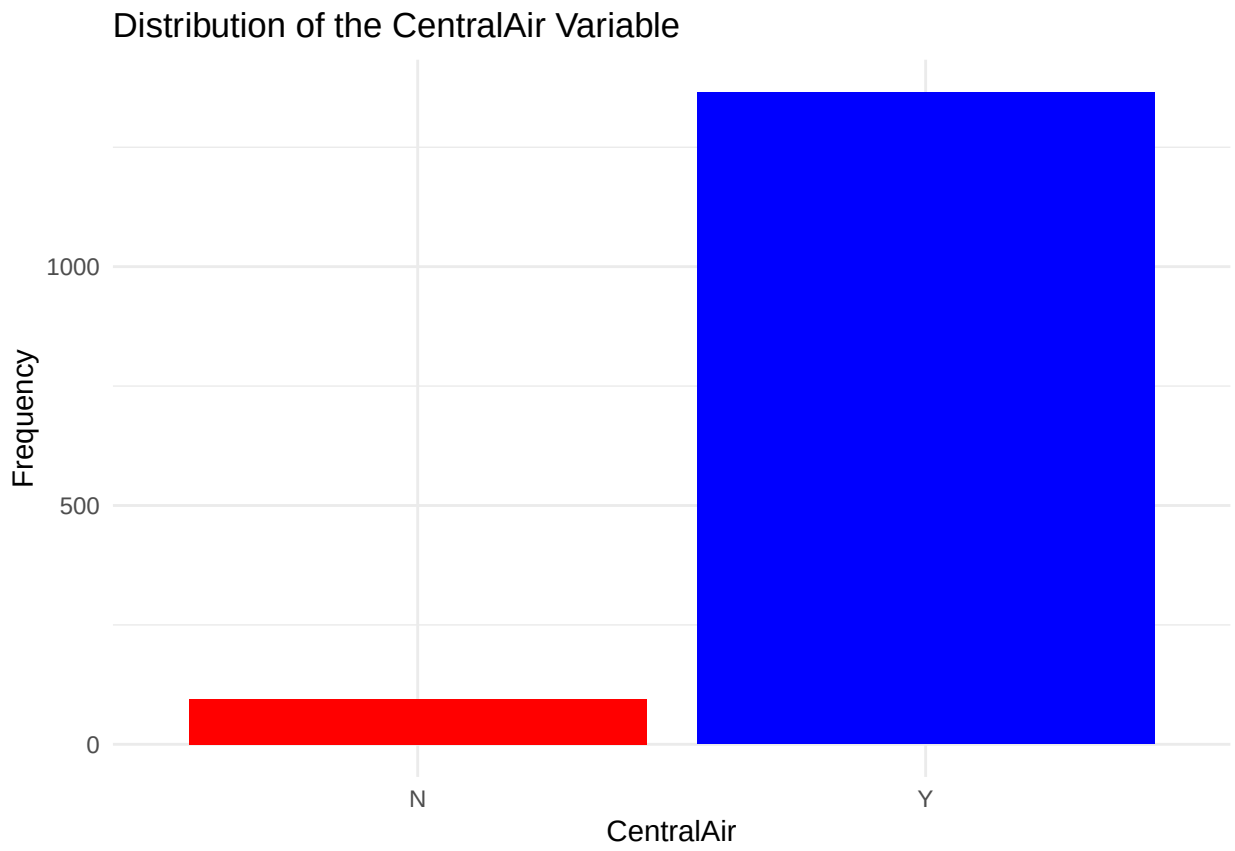
```
##  [1] 1 1 1 3 1 1 1 1 3 1
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "Po" "TA"
```

38. CentralAir and Central Central air conditioning

```
my_colors <- c("red", "blue", "#00BFC4", "#C77CFF", "#619CFF", "pink")
ggplot(house, aes(x = CentralAir, fill = CentralAir)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the CentralAir Variable",
       x = "CentralAir",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$CentralAir, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$CentralAir = numeric_labels
print(house$CentralAir[1:10])
```

```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "N" "Y"
```

```
factor_var <- factor(house$Central, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Central = numeric_labels
print(house$Central[1:10])
```

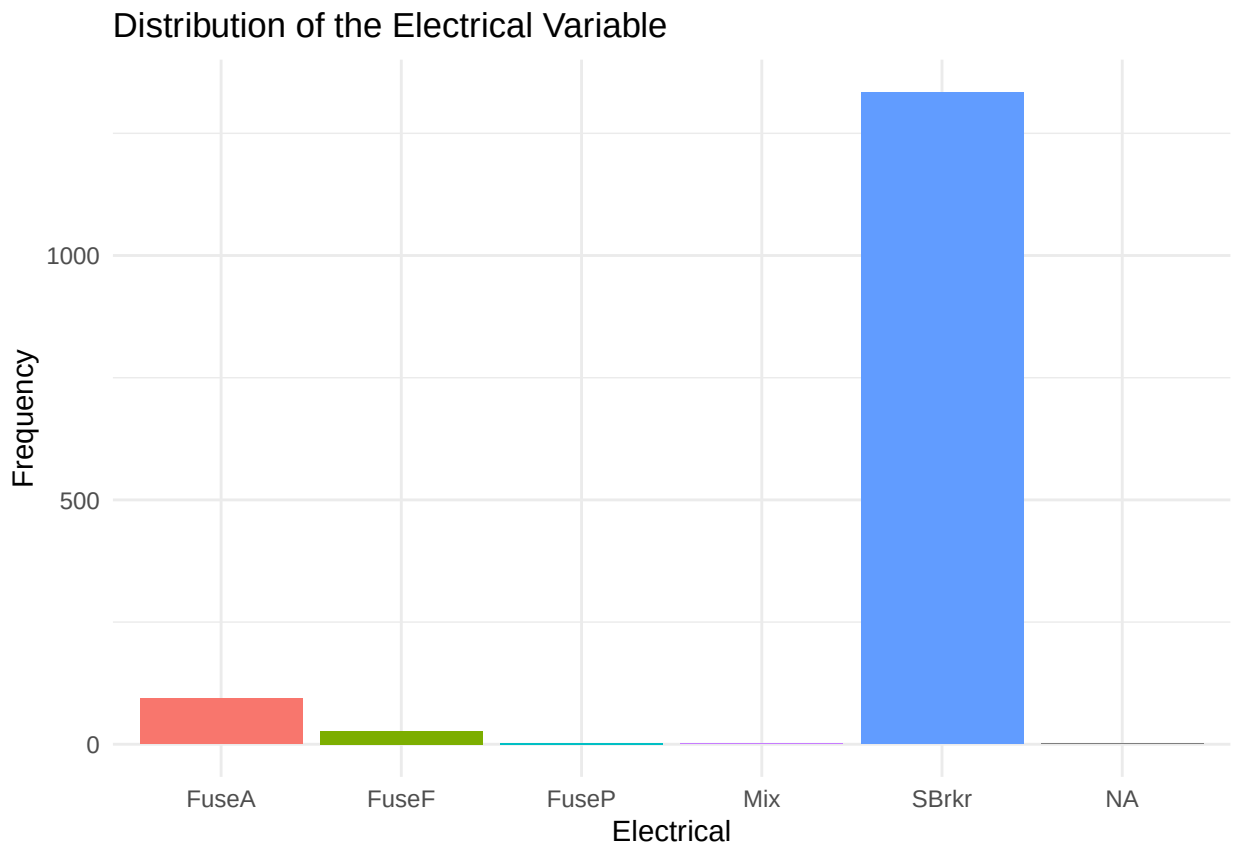
```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "1" "2"
```

39. Electrical Electrical system

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")
ggplot(house, aes(x = Electrical, fill = Electrical)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Electrical Variable",
       x = "Electrical",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$Electrical, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Electrical = numeric_labels
print(house$Electrical[1:10])
```

```
##  [1] 5 5 5 5 5 5 5 5 2 5
```

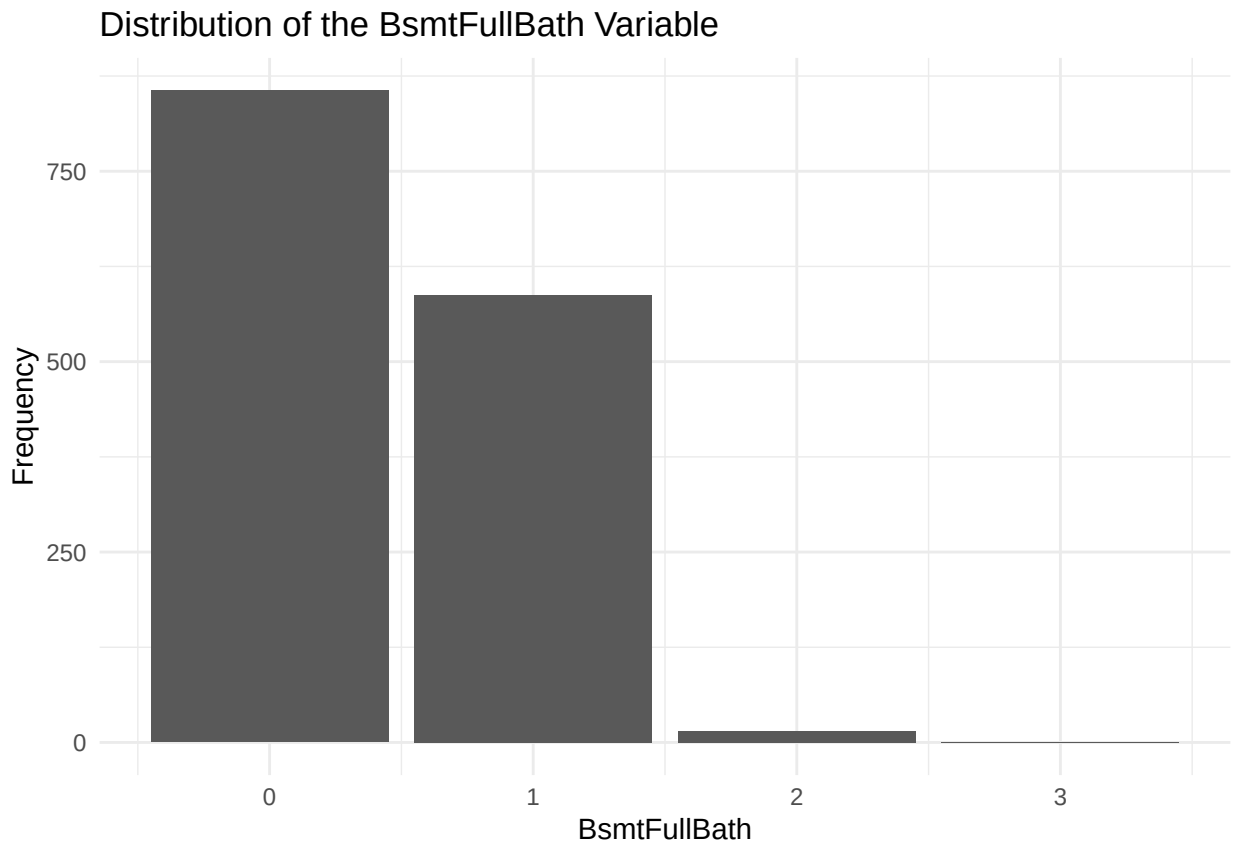
```
categories <- levels(factor_var)
categories
```

```
## [1] "FuseA" "FuseF" "FuseP" "Mix"   "SBrkr" NA
```

40. BsmtFullBath Basement full bathrooms

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")
ggplot(house, aes(x = BsmtFullBath, fill = BsmtFullBath)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the BsmtFullBath Variable",
       x = "BsmtFullBath",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```



Removes the only row with NA

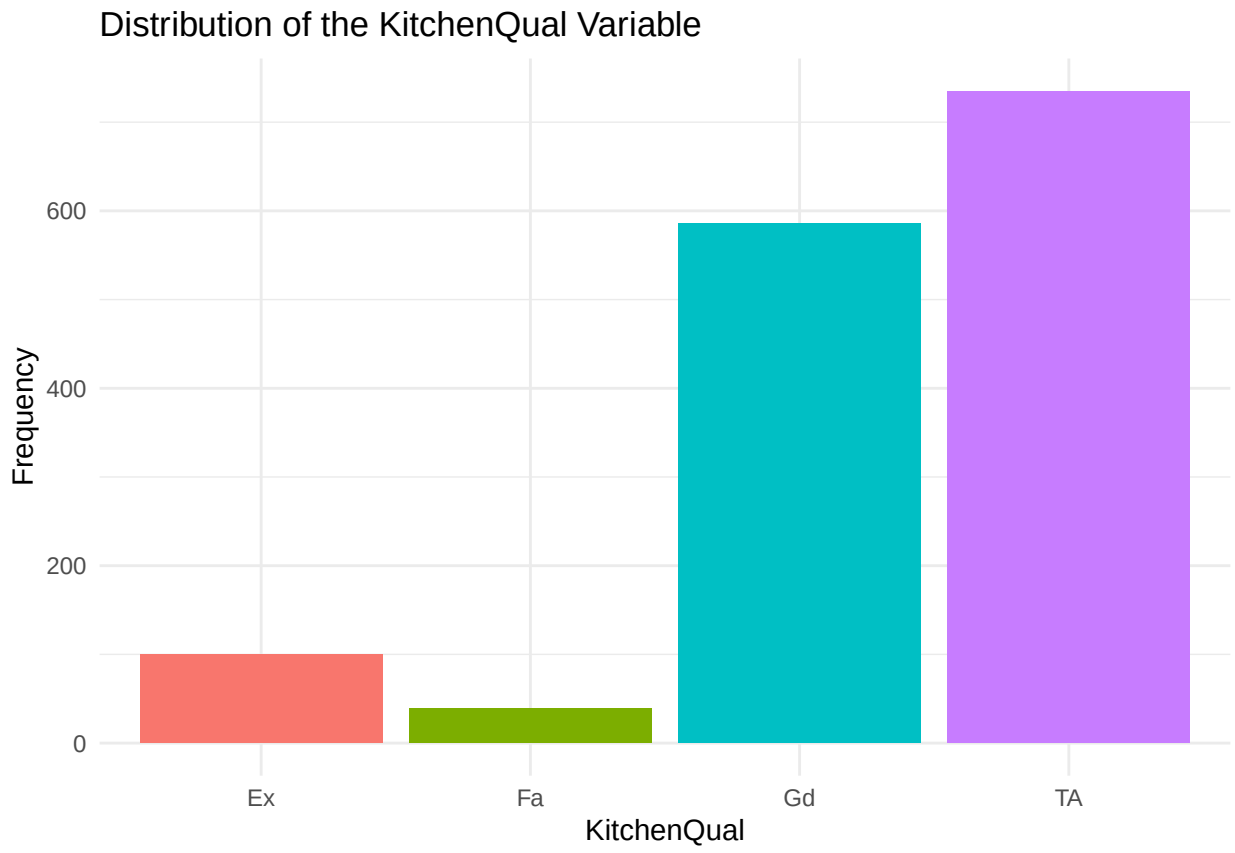
```
sum(is.na(house$BsmtFullBath))
```

```
## [1] 0
```

```
median_value <- median(house$BsmtFullBath, na.rm = TRUE)
house$BsmtFullBath[is.na(house$BsmtFullBath)] <- median_value
```

41. KitchenQual Kitchen quality Ex Excellent Gd Good TA Typical/Average Fa Fair Po Poor

```
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999","pink")
ggplot(house, aes(x = KitchenQual, fill = KitchenQual)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the KitchenQual Variable",
       x = "KitchenQual",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



removed single row with mode value

```
sum(is.na(house$KitchenQual))
```

```
## [1] 0
```

```
mode_value <- as.character(table(house$KitchenQual))[which.max(table(house$KitchenQual))]
house$KitchenQual[is.na(house$KitchenQual)] <- mode_value
```

```
factor_var <- factor(house$KitchenQual, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$KitchenQual = numeric_labels
print(house$KitchenQual[1:10])
```

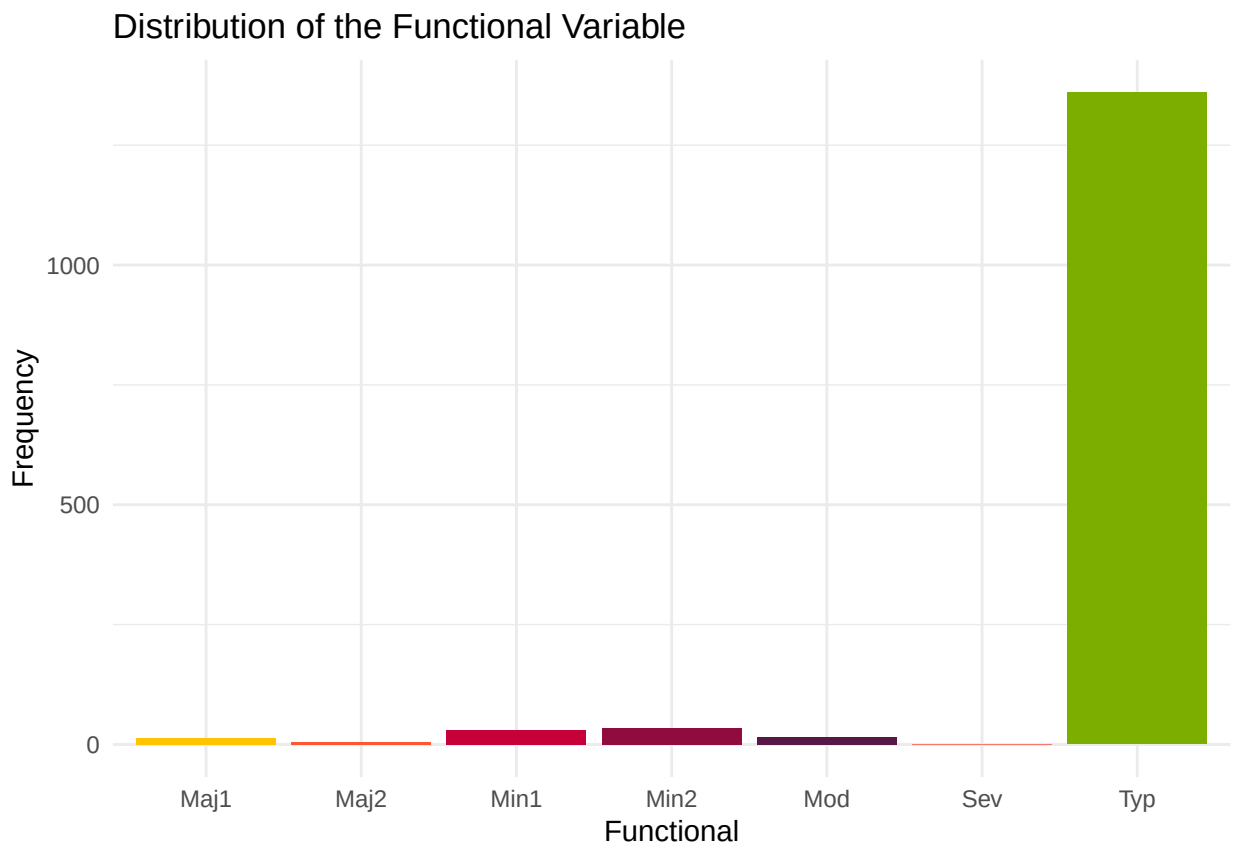
```
##  [1] 3 4 3 3 3 4 3 4 4 4
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "TA"
```

42. Functional Home functionality (Assume typical unless deductions are warranted)

```
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#F8766D", "#7CAE00", "#00BFC4",
ggplot(house, aes(x = Functional, fill = Functional)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the Functional Variable",
       x = "Functional",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



```
mode_Functional <- as.character(names(sort(table(house$Functional),
                                           decreasing = TRUE)[1]))

house$Functional[is.na(house$Functional)] <- mode_Functional
```

```
sum(is.na(house$Functional))
```

```
## [1] 0
```

```
factor_var <- factor(house$Functional, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$Functional = numeric_labels
print(house$Functional[1:10])
```

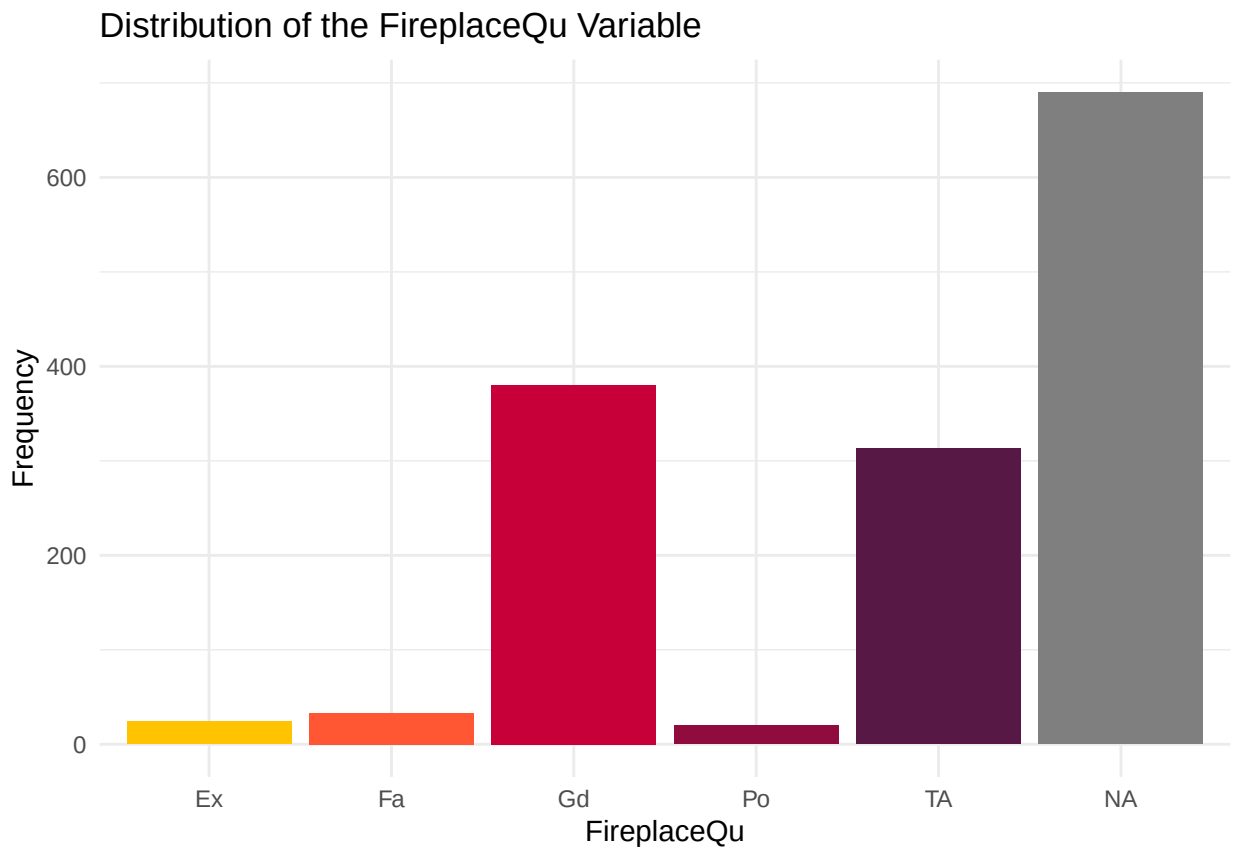
```
##  [1] 7 7 7 7 7 7 7 7 3 7
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Maj1" "Maj2" "Min1" "Min2" "Mod"  "Sev"  "Typ"
```

43. FireplaceQu Fireplace quality

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#F8766D", "#7CAE00", "#00BFC4",
ggplot(house, aes(x = FireplaceQu, fill = FireplaceQu)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the FireplaceQu Variable",
       x = "FireplaceQu",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Again the NA is a no fireplace category which cannot be removed or imputed

```
factor_var <- factor(house$FireplaceQu, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$FireplaceQu = numeric_labels
print(house$FireplaceQu[1:10])
```

```
##  [1] 6 5 5 3 5 6 3 5 5 5
```

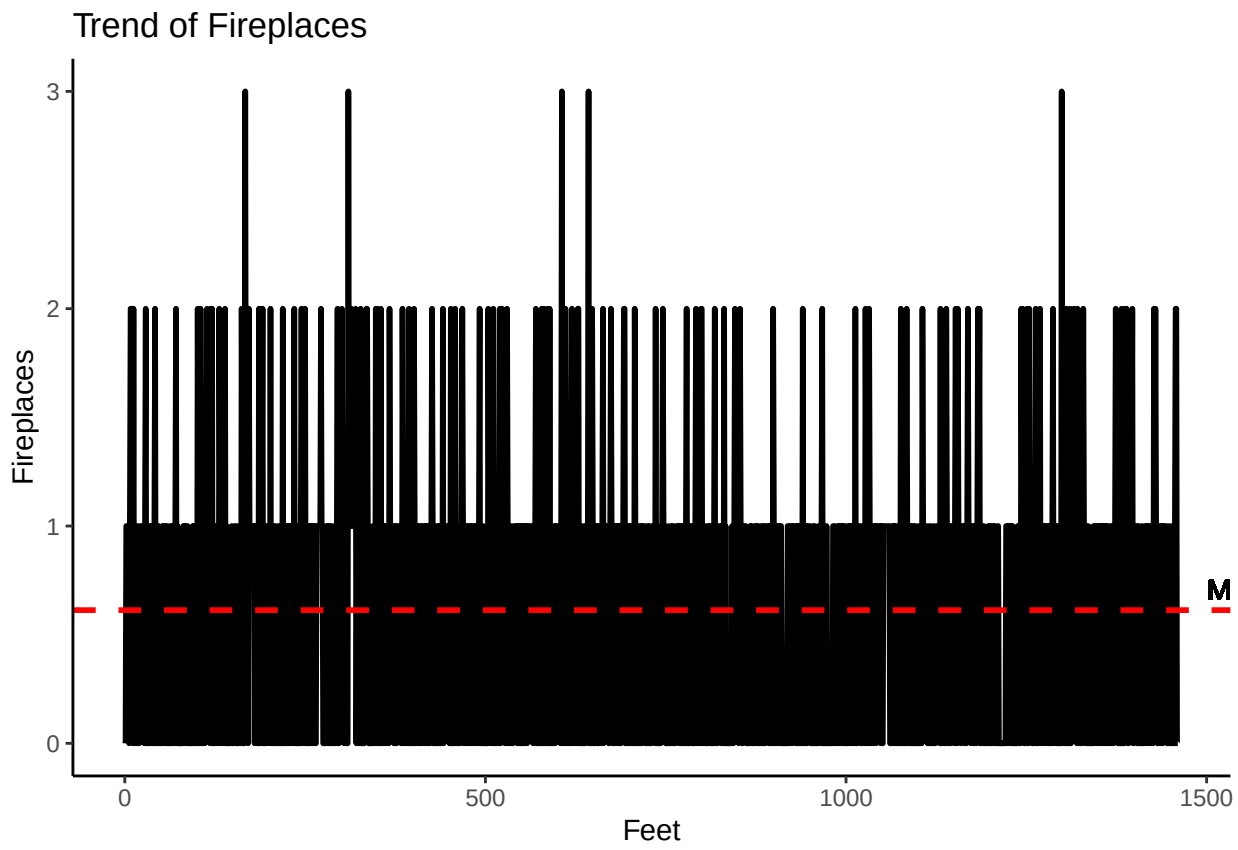
```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "Po" "TA" NA
```

*44. Fireplaces** Number of fireplaces

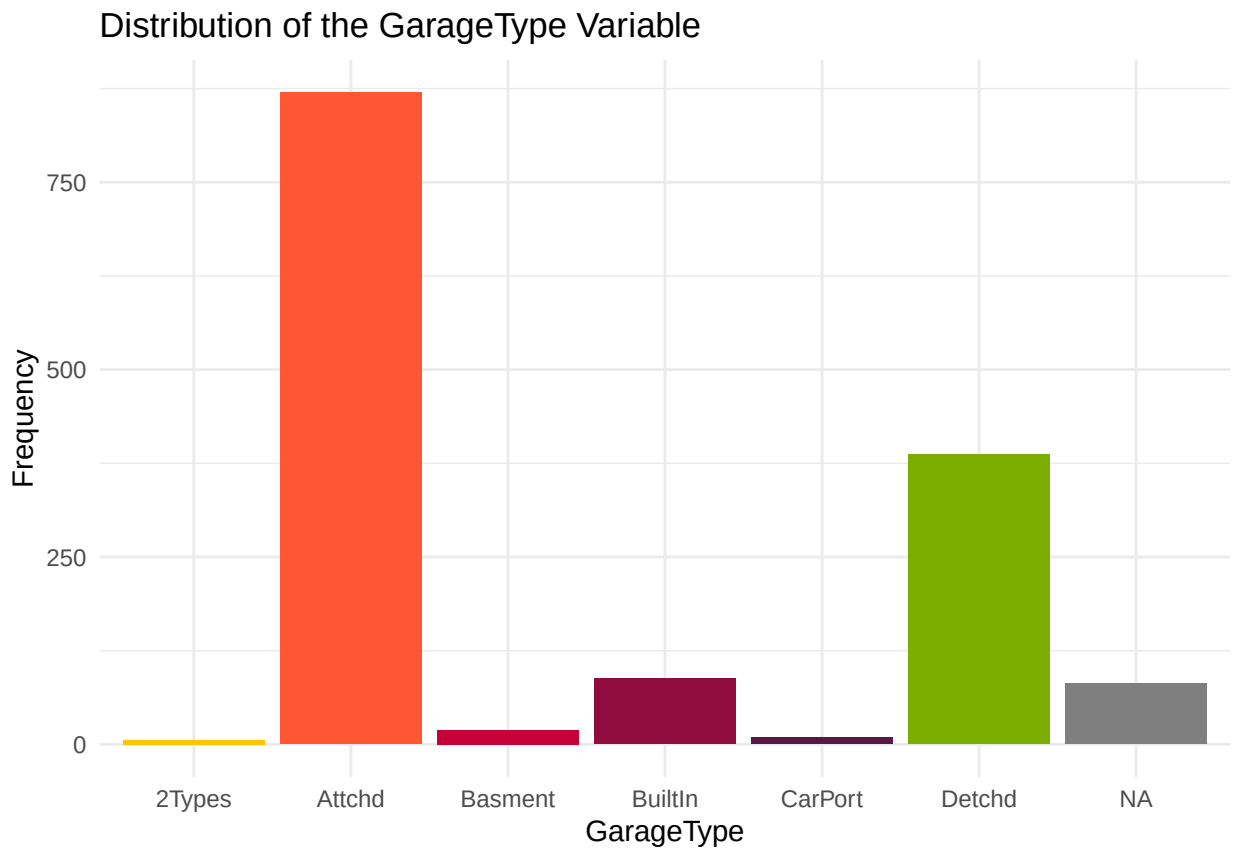
```
ggplot(house, aes(x = seq_along(Fireplaces), y = Fireplaces)) +
  geom_line(color = "black", size = 1) +
  geom_hline(yintercept = mean(house$Fireplaces, na.rm = TRUE),
             color = "red", linetype = "dashed", size = 1) +
  geom_text(aes(x = max(seq_along(house$Fireplaces)),
                y = mean(house$Fireplaces, na.rm = TRUE),
                label = paste("Mean:", round(mean(house$Fireplaces, na.rm = TRUE), 2))),
            color = "black", size = 4, hjust = -0.2, vjust = -0.5) +
  labs(x = "Feet", y = "Fireplaces", title = "Trend of Fireplaces") +
  theme_classic()
```

```
## Warning: Use of ‘house$Fireplaces‘ is discouraged.
## i Use ‘Fireplaces‘ instead.
## Use of ‘house$Fireplaces‘ is discouraged.
## i Use ‘Fireplaces‘ instead.
## Use of ‘house$Fireplaces‘ is discouraged.
## i Use ‘Fireplaces‘ instead.
```



45. GarageType Garage location

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = GarageType, fill = GarageType)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the GarageType Variable",
       x = "GarageType",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$GarageType, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$GarageType = numeric_labels
print(house$GarageType[1:10])
```

```
##  [1] 2 2 2 6 2 2 2 2 6 2
```

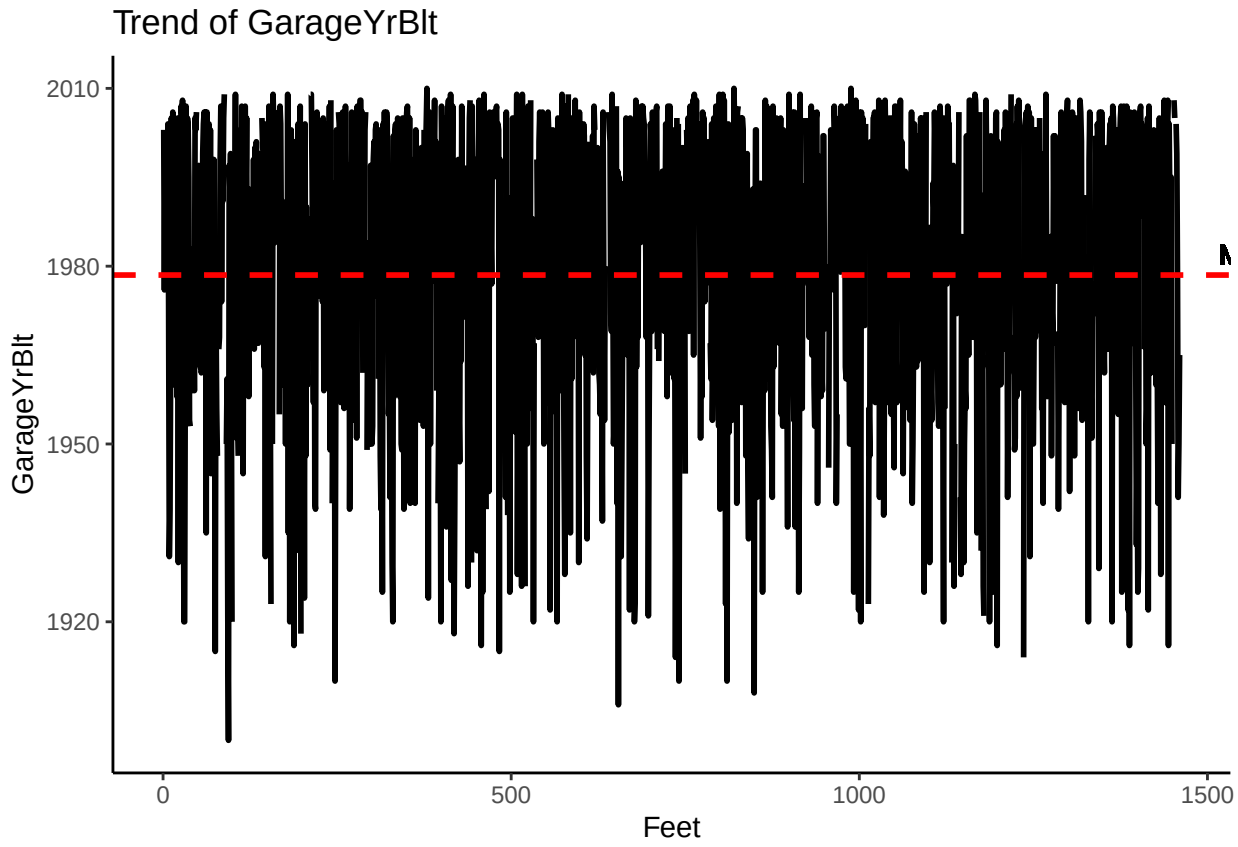
```
categories <- levels(factor_var)
categories
```

```
## [1] "2Types"  "Attchd"  "Basment" "BuiltIn" "CarPort" "Detchd"  NA
```

46. GarageYrBlt Year garage was built

```
ggplot(house, aes(x = seq_along(GarageYrBlt), y = GarageYrBlt)) +
  geom_line(color = "black", size = 1) +
  geom_hline(yintercept = mean(house$GarageYrBlt, na.rm = TRUE),
             color = "red", linetype = "dashed", size = 1) +
  geom_text(aes(x = max(seq_along(house$GarageYrBlt)),
                y = mean(house$GarageYrBlt, na.rm = TRUE),
                label = paste("Mean:", round(mean(house$GarageYrBlt, na.rm = TRUE), 2))),
            color = "black", size = 4, hjust = -0.2, vjust = -0.5) +
  labs(x = "Feet", y = "GarageYrBlt", title = "Trend of GarageYrBlt") +
  theme_classic()
```

```
## Warning: Use of ‘house$GarageYrBlt‘ is discouraged.
## i Use ‘GarageYrBlt‘ instead.
## Use of ‘house$GarageYrBlt‘ is discouraged.
## i Use ‘GarageYrBlt‘ instead.
## Use of ‘house$GarageYrBlt‘ is discouraged.
## i Use ‘GarageYrBlt‘ instead.
```




```
sum(is.na(house$GarageYrBlt))
```

```
## [1] 81
```

```
mode_GarageYrBlt <- as.numeric(names(sort(table(house$GarageYrBlt), decreasing = TRUE)[1]))
mode_GarageYrBlt
```

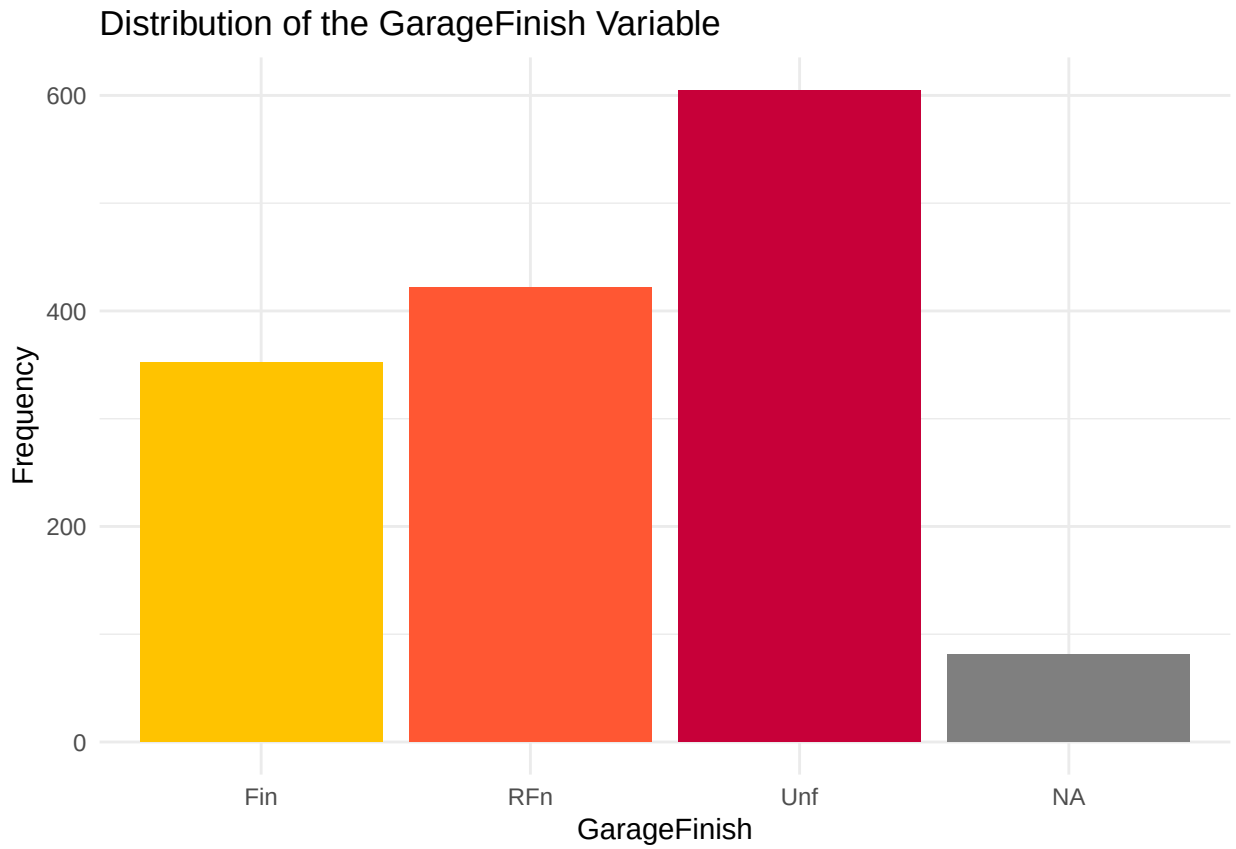
```
## [1] 2005
```

```
house$GarageYrBlt[is.na(house$GarageYrBlt)] <- mode_GarageYrBlt
```

Filled NA values with the mode of the values

47. GarageFinish Interior finish of the garage

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = GarageFinish, fill = GarageFinish)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the GarageFinish Variable",
       x = "GarageFinish",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



NA is No garage here

```
factor_var <- factor(house$GarageFinish, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$GarageFinish = numeric_labels
print(house$GarageFinish[1:10])
```

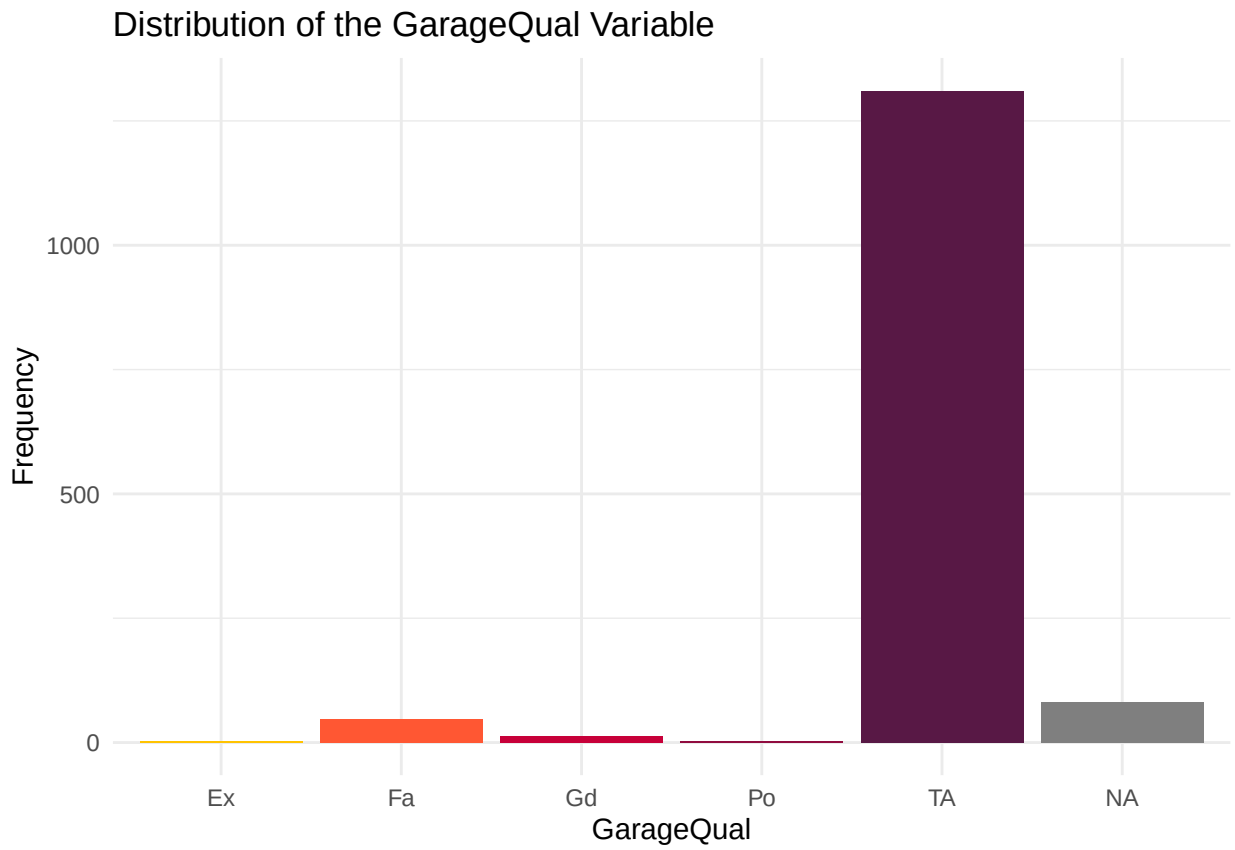
```
##  [1] 2 2 2 3 2 3 2 2 3 2
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Fin" "RFn" "Unf" NA
```

48. GarageQual Garage quality

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = GarageQual, fill = GarageQual)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the GarageQual Variable",
       x = "GarageQual",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



NA is no garage

```
factor_var <- factor(house$GarageQual, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$GarageQual = numeric_labels
print(house$GarageQual[1:10])
```

```
##  [1] 5 5 5 5 5 5 5 5 2 3
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Ex" "Fa" "Gd" "Po" "TA" NA
```

49. GarageCars Drop the only row with NA

```
sum(is.na(house$GarageCars))
```

```
## [1] 0
```

```
house$GarageCars[is.na(house$GarageCars)] <- mean(house$GarageCars, na.rm = TRUE)
sum(is.na(house$GarageCars))
```

```
## [1] 0
```

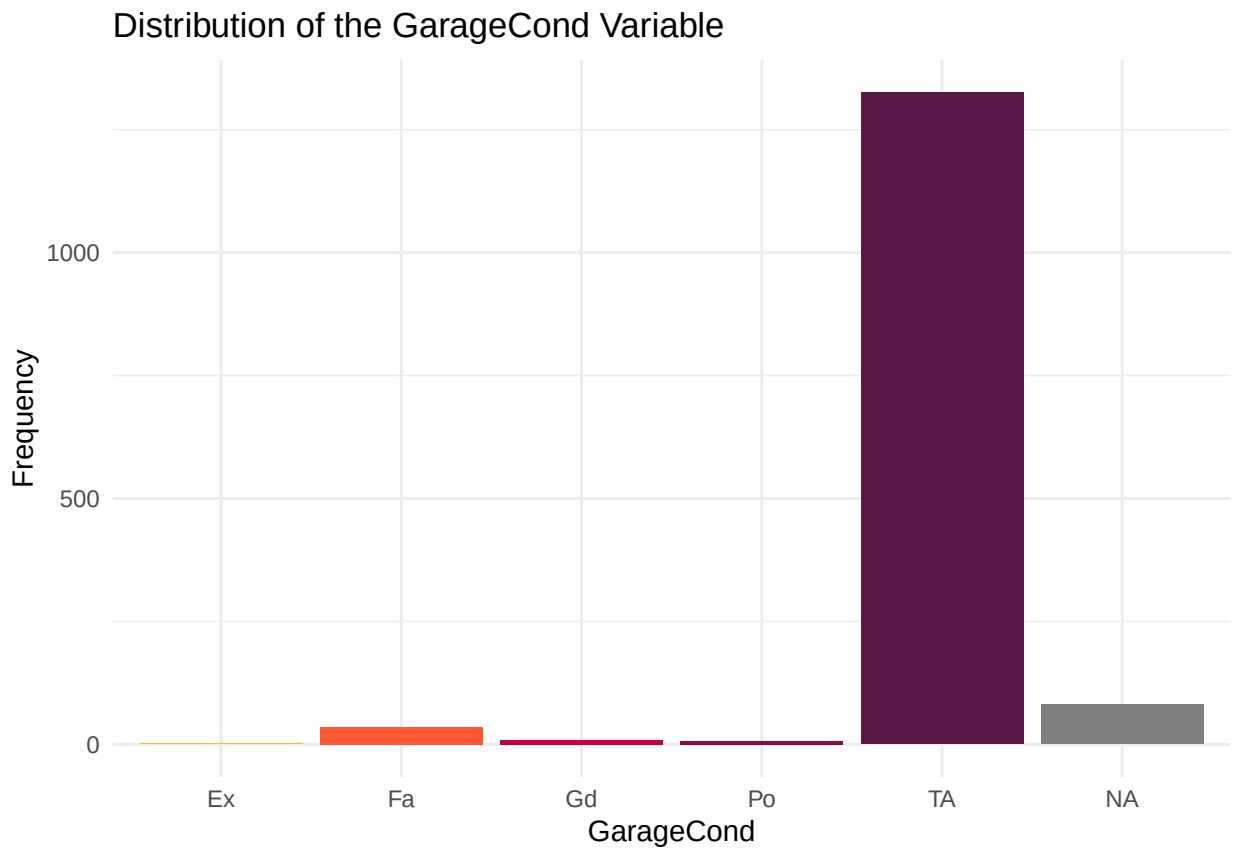
50. GarageArea

```
sum(is.na(house$GarageArea))
```

```
## [1] 0
```

51. GarageCond Garage condition

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = GarageCond, fill = GarageCond)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the GarageCond Variable",
       x = "GarageCond",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Na means no garage

```
factor_var <- factor(house$GarageCond, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$GarageCond = numeric_labels
print(house$GarageCond[1:10])
```

```
##  [1] 5 5 5 5 5 5 5 5 5 5
```

```
categories <- levels(factor_var)
categories
```

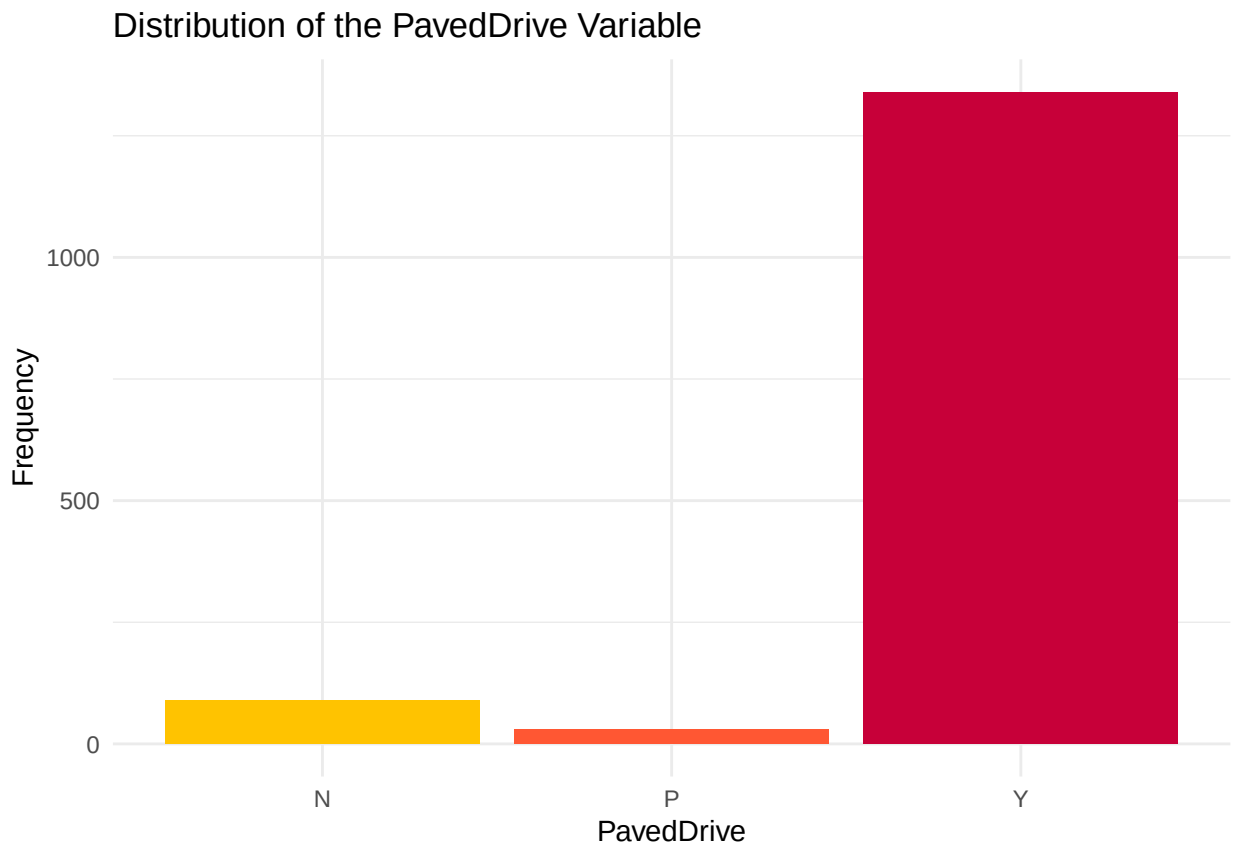
```
## [1] "Ex" "Fa" "Gd" "Po" "TA" NA
```

52. PavedDrive Paved driveway

Y Paved P Partial Pavement N Dirt/Gravel

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = PavedDrive, fill = PavedDrive)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the PavedDrive Variable",
       x = "PavedDrive",
```

```
      y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$PavedDrive, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$PavedDrive = numeric_labels
print(house$PavedDrive[1:10])
```

```
##  [1] 3 3 3 3 3 3 3 3 3 3
```

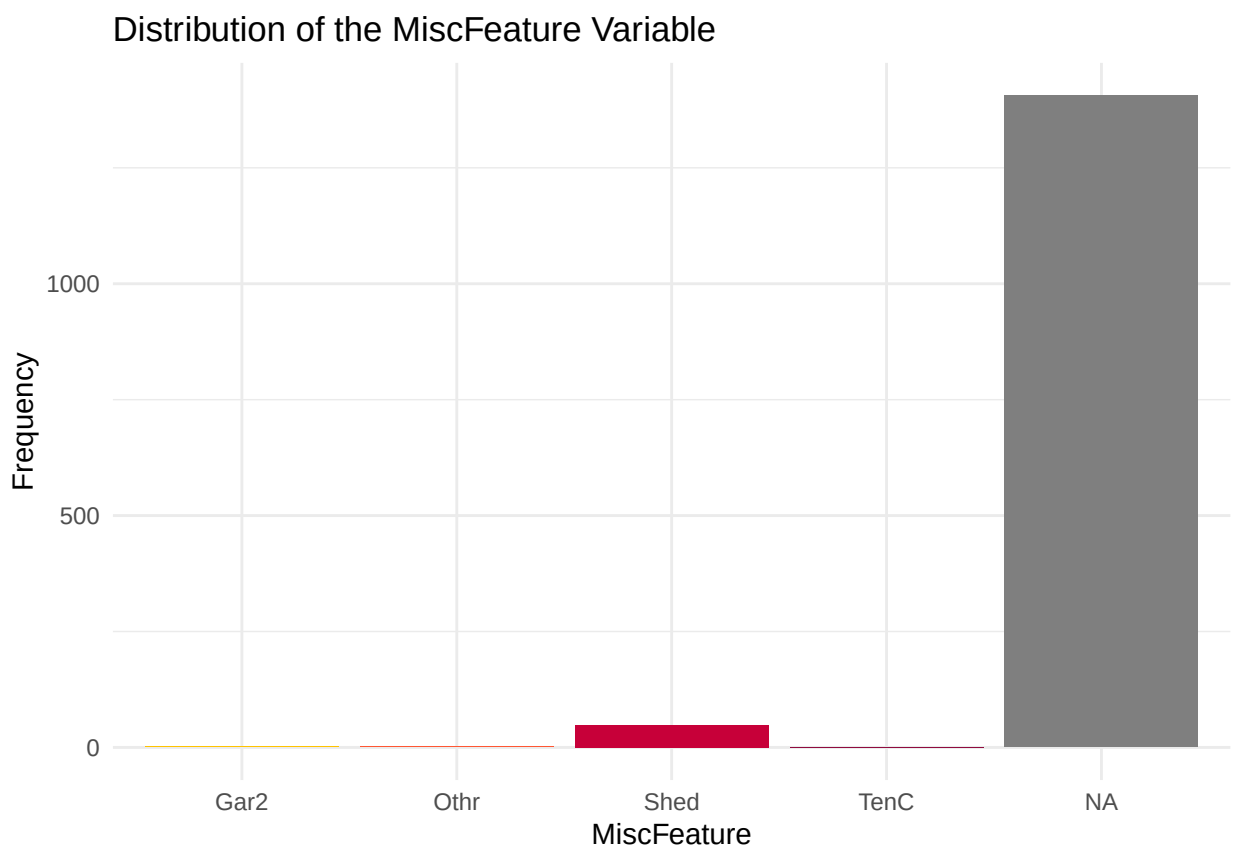
```
categories <- levels(factor_var)
categories
```

```
## [1] "N" "P" "Y"
```

53. MiscFeature Miscellaneous feature not covered in other categories

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = MiscFeature, fill = MiscFeature)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
```

```
  labs(title = "Distribution of the MiscFeature Variable",
       x = "MiscFeature",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



NA is no feature so cannot impute it

```
factor_var <- factor(house$MiscFeature, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$MiscFeature = numeric_labels
print(house$MiscFeature[1:10])
```

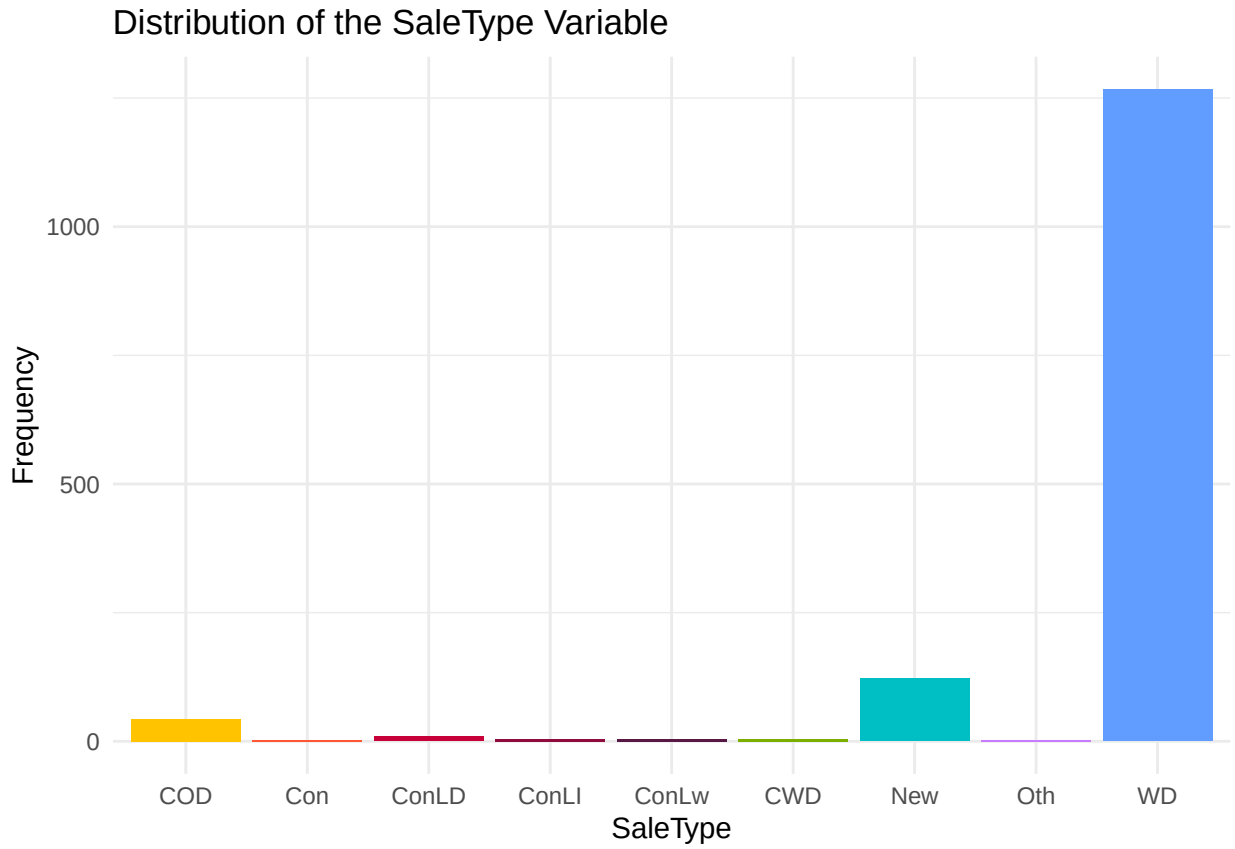
```
##  [1] 5 5 5 5 5 3 5 3 5 5
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Gar2" "Othr" "Shed" "TenC" NA
```

54. SaleType Type of sale WD Warranty Deed - Conventional CWD Warranty Deed - Cash VWD Warranty Deed - VA Loan New Home just constructed and sold COD Court Officer Deed/Estate Con Contract 15% Down payment regular terms ConLw Contract Low Down payment and low interest ConLI Contract Low Interest ConLD Contract Low Down Oth Other

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = SaleType, fill = SaleType)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the SaleType Variable",
       x = "SaleType",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Removed single NA value row

```
mode_value <- as.character(table(house$SaleType))[which.max(table(house$SaleType))]
house$SaleType[is.na(house$SaleType)] <- mode_value
sum(is.na(house$SaleType))
```

```
## [1] 0
```

```
factor_var <- factor(house$SaleType, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$SaleType = numeric_labels
print(house$SaleType[1:10])
```

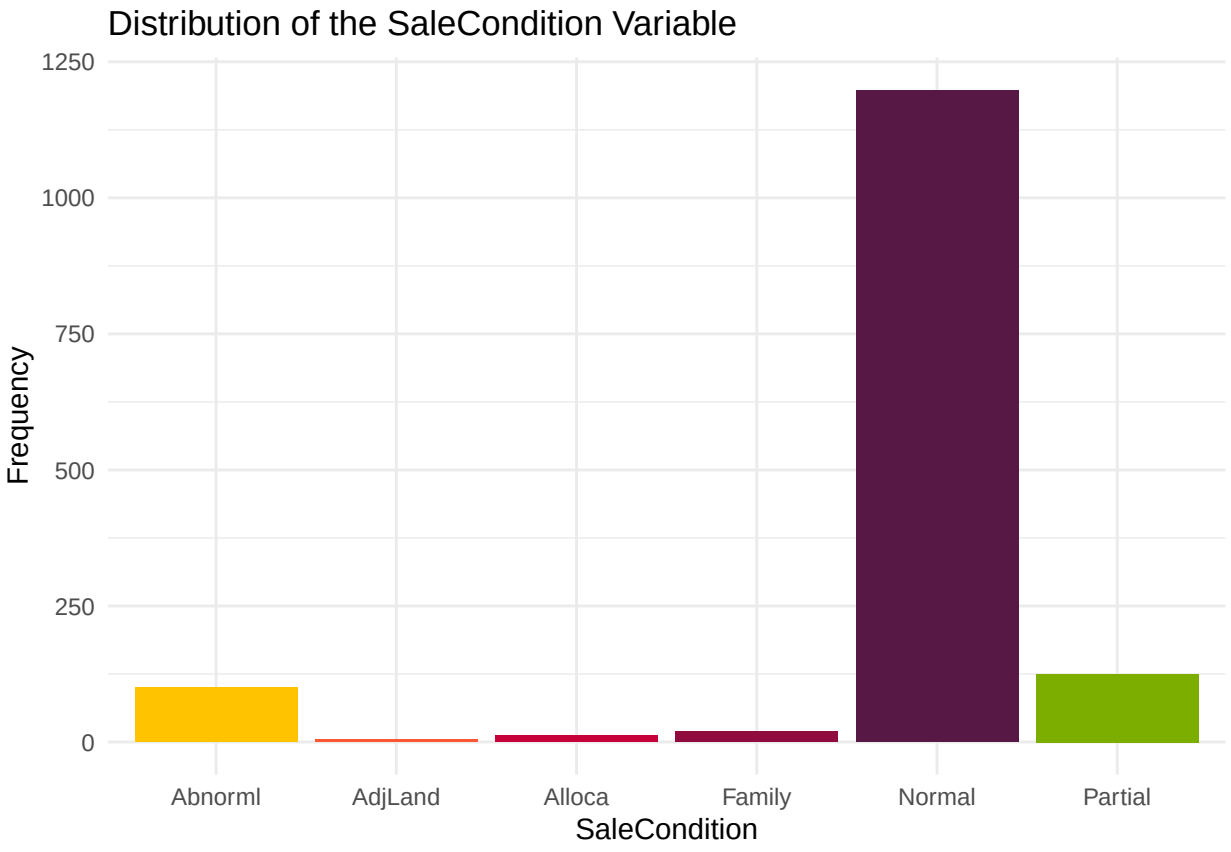
```
##  [1] 9 9 9 9 9 9 9 9 9 9
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "COD"   "Con"   "ConLD" "ConLI" "ConLw" "CWD"   "New"   "Oth"   "WD"
```

55. SaleCondition

```
library(ggplot2)
my_colors <- c("#FFC300", "#FF5733", "#C70039", "#900C3F", "#581845", "#7CAE00", "#00BFC4", "#C77CFF",
ggplot(house, aes(x = SaleCondition, fill = SaleCondition)) +
  geom_bar() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of the SaleCondition Variable",
       x = "SaleCondition",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```




```
factor_var <- factor(house$SaleCondition, exclude = NULL)
numeric_labels <- as.integer(factor_var)
house$SaleCondition = numeric_labels
print(house$SaleCondition[1:10])
```

```
##  [1] 5 5 5 1 5 5 5 5 1 5
```

```
categories <- levels(factor_var)
categories
```

```
## [1] "Abnorml" "AdjLand" "Alloca"  "Family"  "Normal"  "Partial"
```

```
write.csv(house,file="train.csv", row.names=FALSE, col.names = TRUE)
```

```
## Warning in write.csv(house, file = "train.csv", row.names = FALSE, col.names =
## TRUE): attempt to set 'col.names' ignored
```

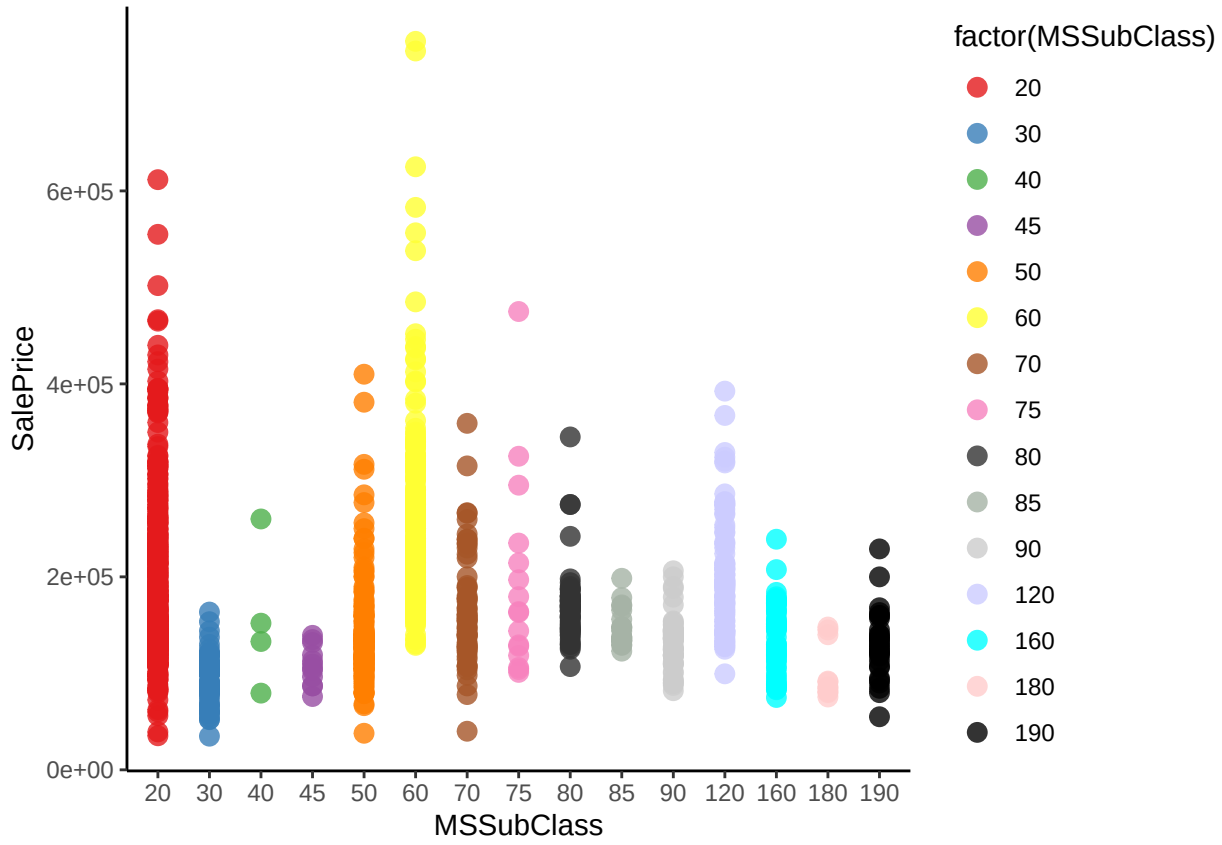
Analysis with Target variable

```
sum(is.na(house))
```

```
## [1] 0
```

1. MSSubClass

```
colors <- c(rgb(0.894, 0.102, 0.110),
            rgb(0.216, 0.494, 0.722),
            rgb(0.302, 0.686, 0.290),
            rgb(0.596, 0.306, 0.639),
            rgb(1.000, 0.498, 0.000),
            rgb(1.000, 1.000, 0.200),
            rgb(0.651, 0.337, 0.157),
            rgb(0.969, 0.510, 0.745),
            rgb(0.200, 0.200, 0.200),
            rgb(0.650, 0.700, 0.650),
            rgb(0.800, 0.800, 0.800),
            rgb(0.800, 0.800, 1.000),
            rgb(0.000, 1.000, 1.000),
            rgb(1.000, 0.800, 0.800),
            rgb(0.000, 0.000, 0.000))

# create a scatter plot with 15 different colors for MSSubClass
ggplot(house, aes(x = factor(MSSubClass), y = SalePrice, color = factor(MSSubClass))) +
  geom_point(size = 3, alpha = 0.8) +
  scale_color_manual(values = colors) +
  labs(x = "MSSubClass", y = "SalePrice") +
  theme_classic()
```




```
cor(house$MSSubClass, house$SalePrice)
```

```
## [1] -0.08428414
```

```
res <- aov(SalePrice ~ MSSubClass, data = house)
summary(res)
```

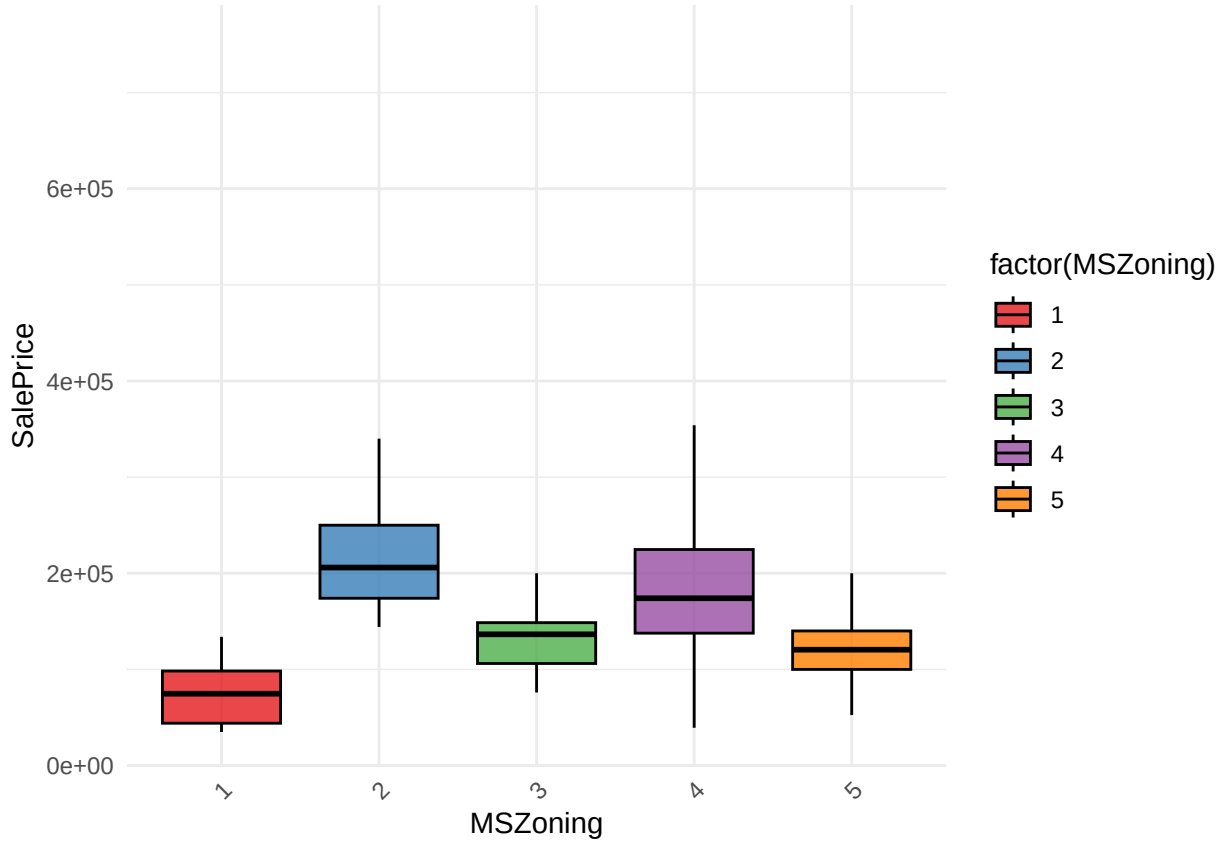
```
##               Df    Sum Sq   Mean Sq F value  Pr(>F)
## MSSubClass     1 6.541e+10 6.541e+10   10.43 0.00127 **
## Residuals   1458 9.143e+12 6.271e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A low p value 0.00127 indicates that the variable MSSubClass is related to the target variable

2.MSZoning

```
library(ggplot2)
colors <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

# Create a box plot of SalePrice by MSZoning
ggplot(house, aes(x = factor(MSZoning), y = SalePrice, fill = factor(MSZoning))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
```

```
  labs(x = "MSZoning", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the graph we can infer that the category 2(Commercial) houses have a higher price comparatively and the Agriculture sector houses are cheaper

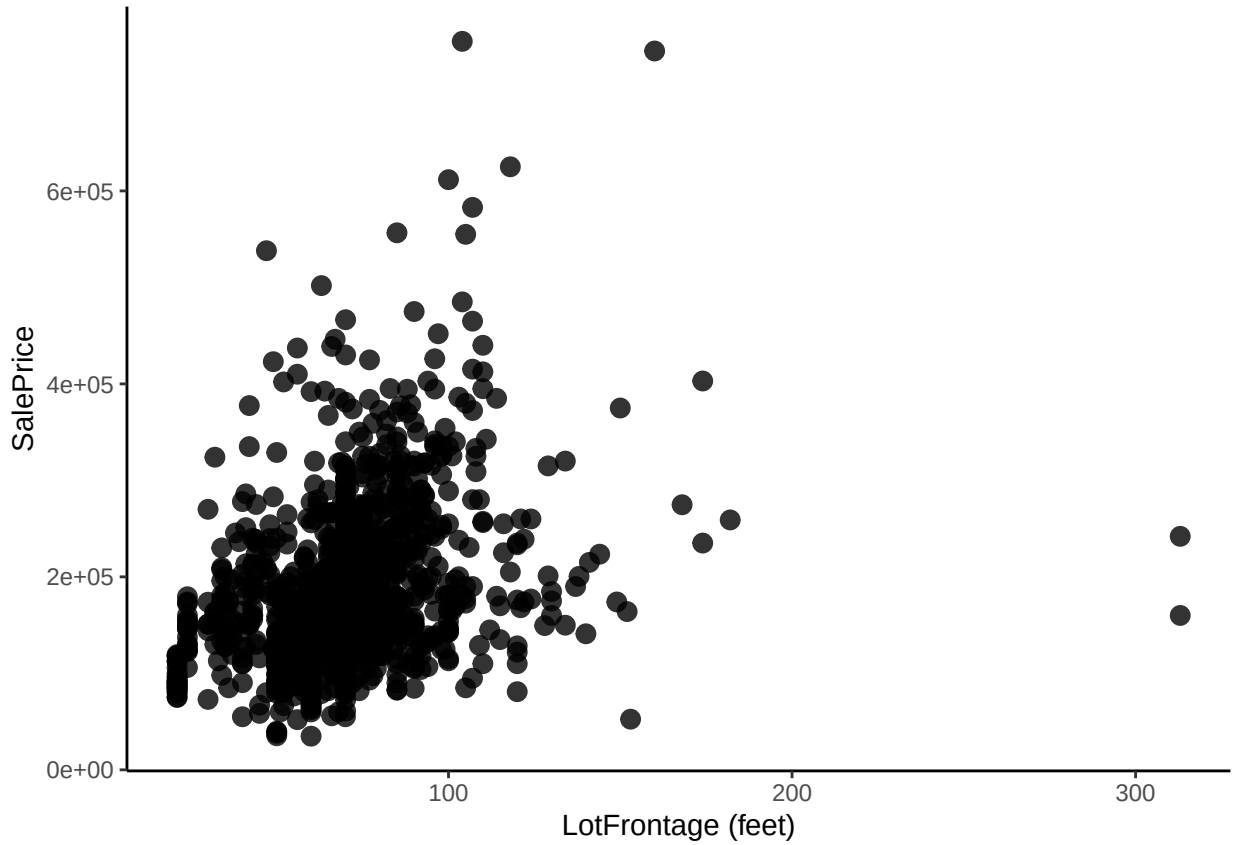
```
res <- aov(SalePrice ~ MSZoning, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value  Pr(>F)
## MSZoning       1 2.564e+11 2.564e+11   41.76 1.4e-10 ***
## Residuals   1458 8.952e+12 6.140e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A large F value here means that the difference in mean SalePrice for different MSZoning categories is large compared to the in category variability

3. LotFrontage

```
ggplot(house, aes(x = LotFrontage, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "LotFrontage (feet)", y = "SalePrice") +
  theme_classic()
```




```
cor(house$LotFrontage, house$SalePrice)
```

```
## [1] 0.3349009
```

We cannot analyze much from the graph however most of the data seems to be near the 1-400000$ range when the distance is less than 150 feet

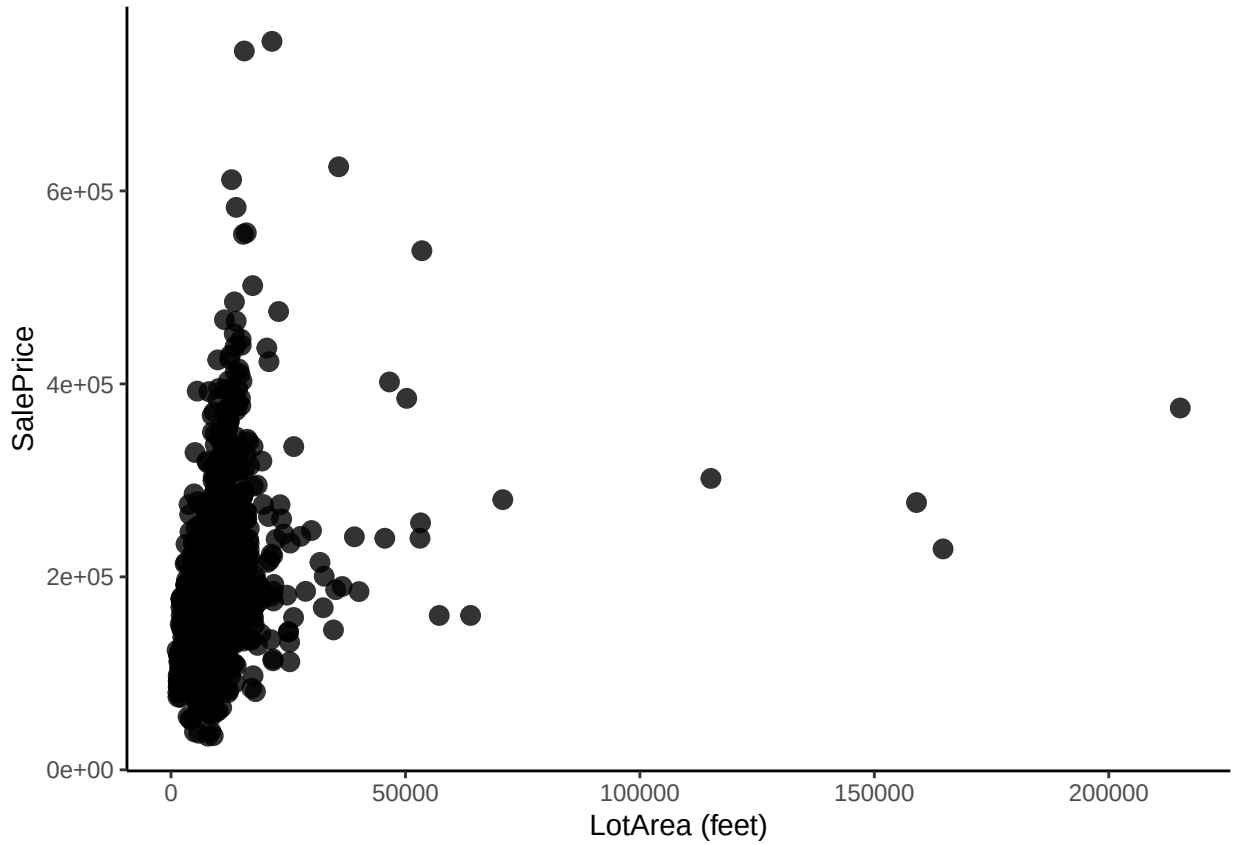
```
res <- aov(SalePrice ~ LotFrontage, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## LotFrontage    1 1.033e+12 1.033e+12   184.2 <2e-16 ***
## Residuals   1458 8.175e+12 5.607e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again a very high F value and low p value indicate the rejection of null hypothesis i.e. the variable is strongly related to the target variable SalePrice

4. LotArea

```
ggplot(house, aes(x = LotArea, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "LotArea (feet)", y = "SalePrice") +
  theme_classic()
```




```
cor(house$LotArea, house$SalePrice)
```

```
## [1] 0.2638434
```

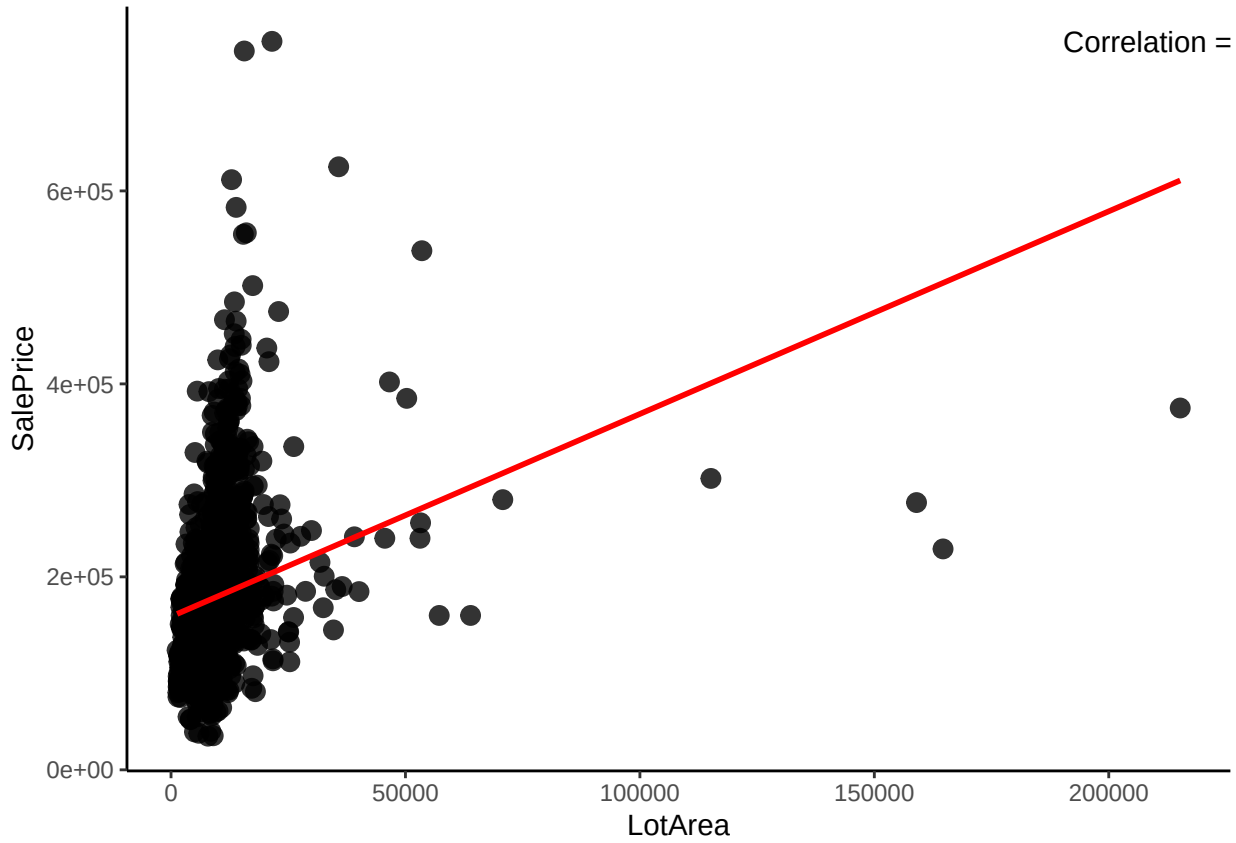
From the graph we can see that the values vary over a big range for lot area between 0 to 30000 feet

```
correlation <- cor(house$LotArea, house$SalePrice)
correlation
```

```
## [1] 0.2638434
```

```
ggplot(house, aes(x = LotArea, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "LotArea", y = "SalePrice") +
  annotate("text", x = max(house$LotArea), y = max(house$SalePrice),
           label = paste("Correlation =", round(correlation, 2))) +
  theme_classic()
```

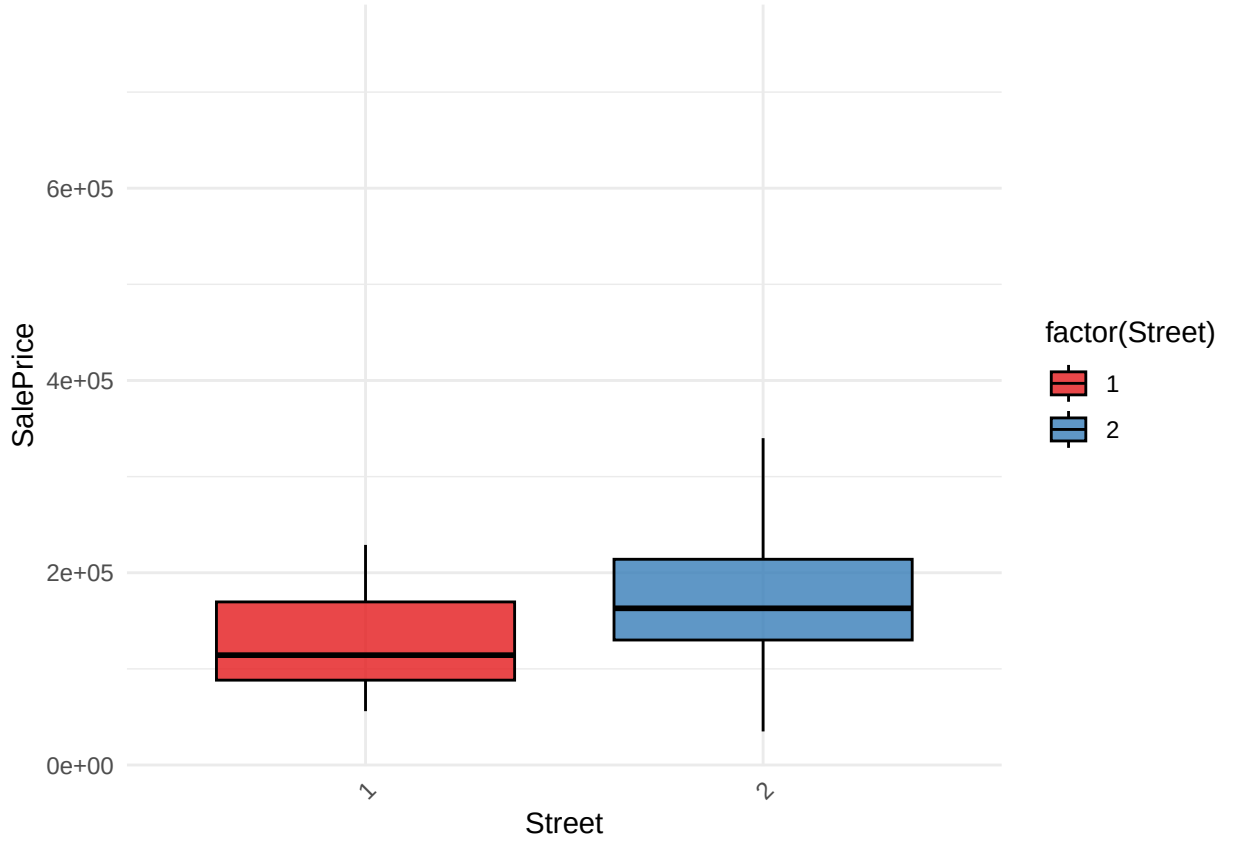
```
## `geom_smooth()` using formula = 'y ~ x'
```



A Correlation value of 0.26 indicates some positive relation i.e. an increase here leads to increase in the SalePrice

5. Street

```
library(ggplot2)
colors <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

# Create a box plot of SalePrice by Street
ggplot(house, aes(x = factor(Street), y = SalePrice, fill = factor(Street))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "Street", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Paved Streets have a higher price for the house compared to Gravel streets

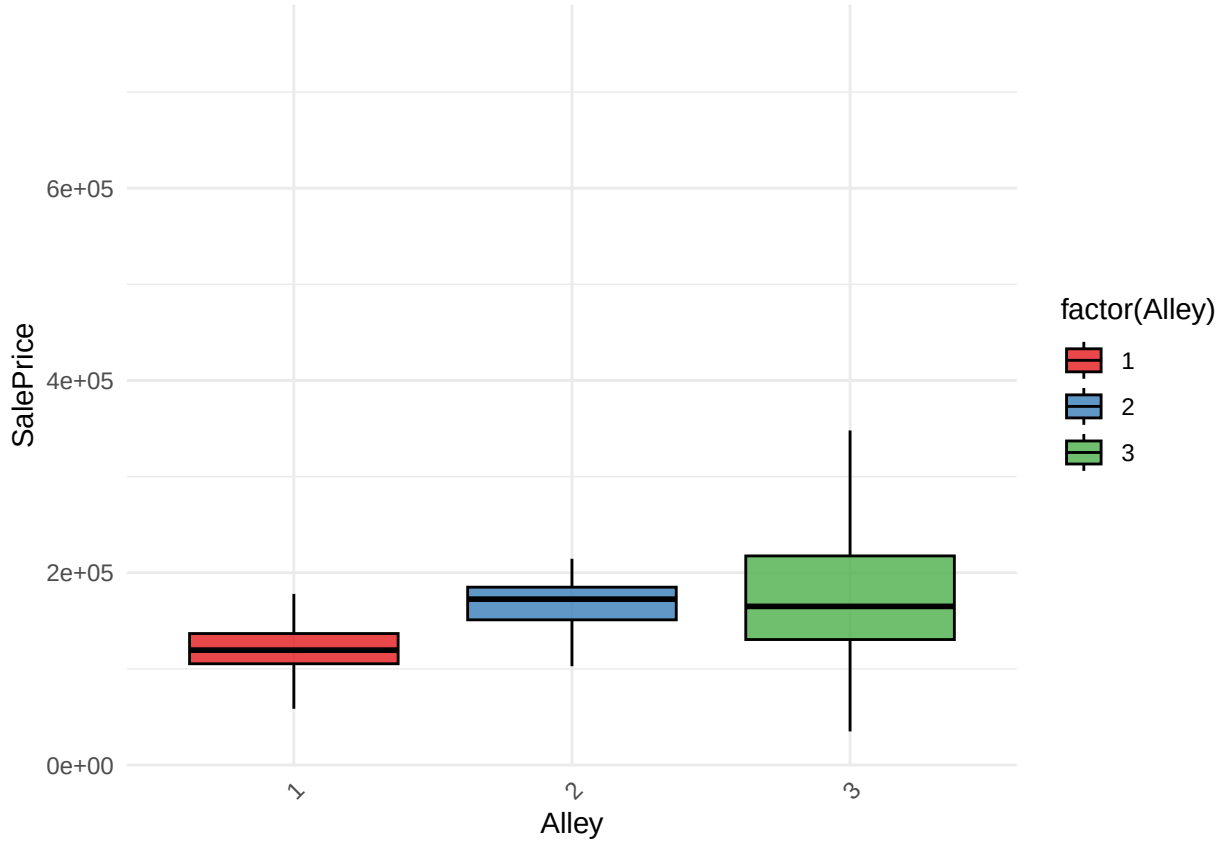
```
res <- aov(SalePrice ~ Street, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Street         1 1.551e+10 1.551e+10   2.459  0.117
## Residuals   1458 9.192e+12 6.305e+09
```

The F and p value suggest that the mean is not very different for the categories and it is difficult to identify a relation with the category and target variable(unable to reject null hypothesis)

6. Alley

```
library(ggplot2)
colors <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

ggplot(house, aes(x = factor(Alley), y = SalePrice, fill = factor(Alley))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "Alley", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The paved Alleys have higher prices compared to no alley access and gravel alley

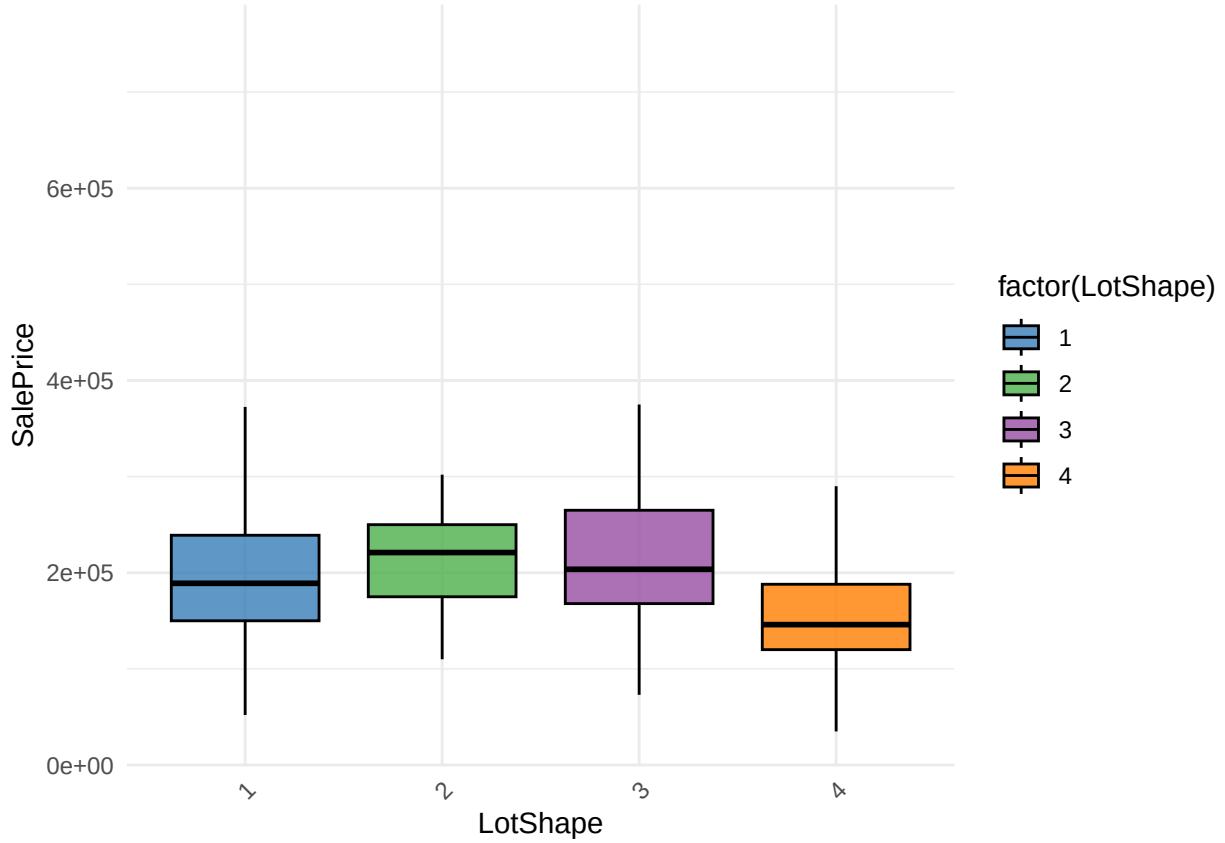
```
res <- aov(SalePrice ~ Alley, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## Alley          1 1.801e+11 1.801e+11   29.09 8.04e-08 ***
## Residuals   1458 9.028e+12 6.192e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A positive correlation and high F value indicate positive relation between the two variables

7. LotShape

```
library(ggplot2)
colors <- c("#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

ggplot(house, aes(x = factor(LotShape), y = SalePrice, fill = factor(LotShape))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "LotShape", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the graph we can see that the 4th category which is irregular lot shape has the lowest SalePrices and the rest have close means

```
res <- aov(SalePrice ~ LotShape, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## LotShape       1 6.015e+11 6.015e+11   101.9 <2e-16 ***
## Residuals   1458 8.606e+12 5.903e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

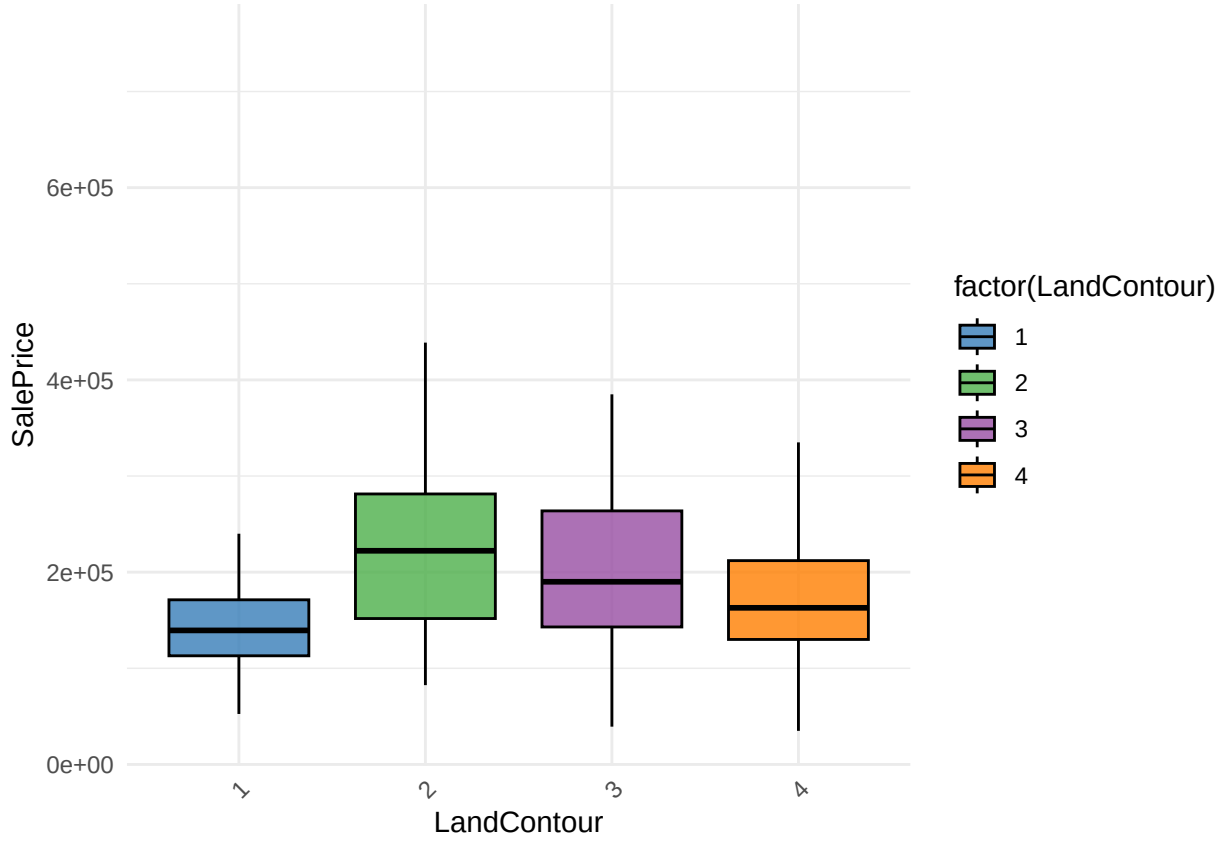
```
kruskal.test(house$LotShape ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$LotShape by house$SalePrice
## Kruskal-Wallis chi-squared = 787.89, df = 662, p-value = 0.0005189
```

8. LandContour

```
library(ggplot2)
colors <- c("#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")
```

```
ggplot(house, aes(x = factor(LandContour), y = SalePrice, fill = factor(LandContour))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "LandContour", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the graph clearly the banked and hillside houses have a higher sale Price

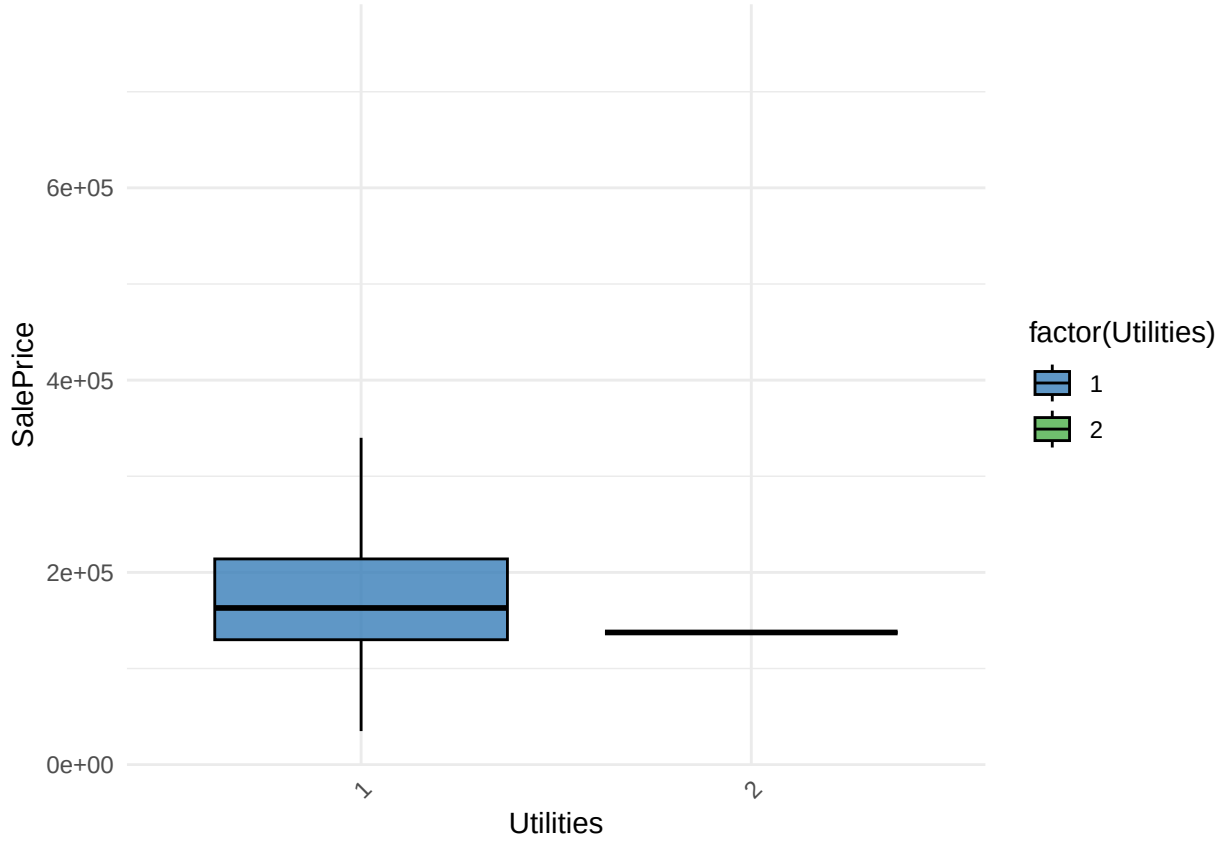
```
res <- aov(SalePrice ~ LandContour, data = house)
summary(res)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## LandContour     1 2.199e+09 2.199e+09   0.348  0.555
## Residuals    1458 9.206e+12 6.314e+09
```

9. Utilities

```
library(ggplot2)
colors <- c("#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

ggplot(house, aes(x = factor(Utilities), y = SalePrice, fill = factor(Utilities))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "Utilities", y = "SalePrice") +
```

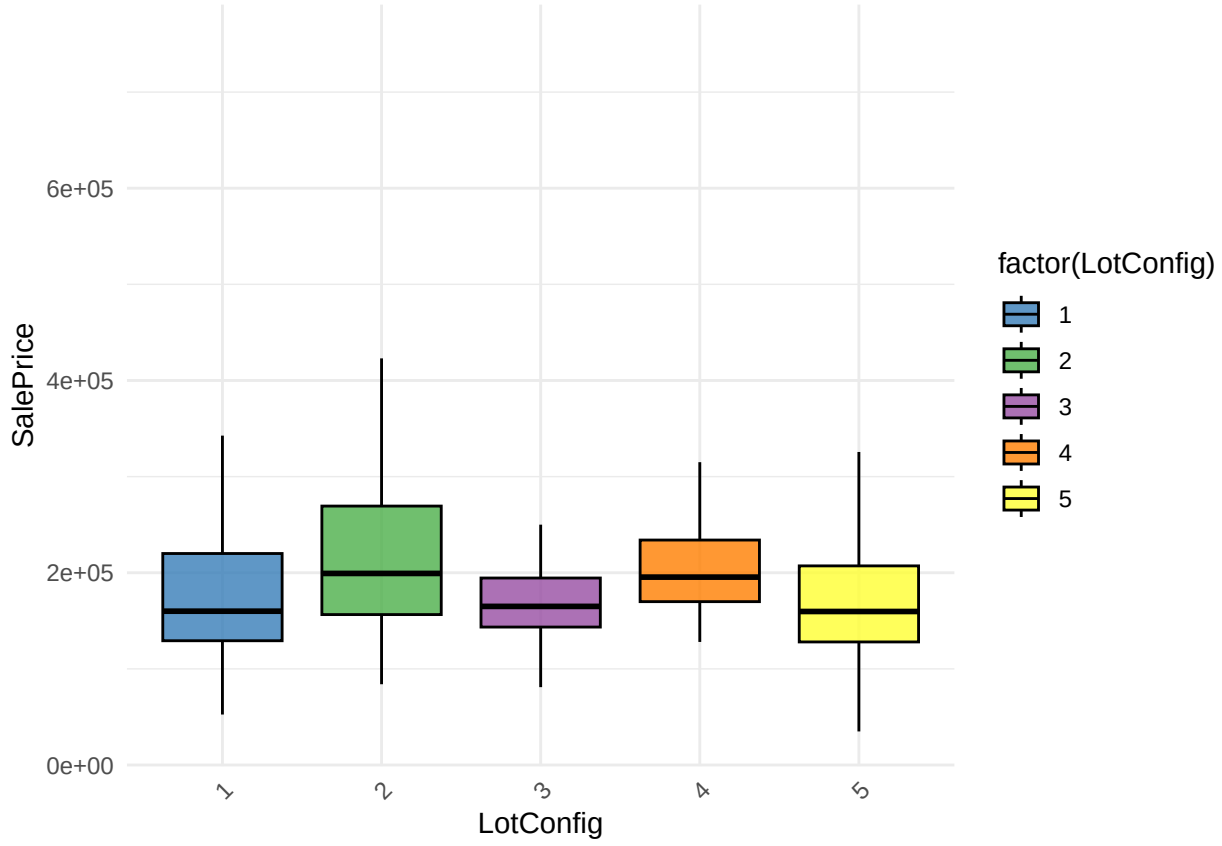
```
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Everyone has the same utility and this can be seen from the graph

10. LotConfig

```
library(ggplot2)
colors <- c("#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

ggplot(house, aes(x = factor(LotConfig), y = SalePrice, fill = factor(LotConfig))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "LotConfig", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Corner lots have a higher price compared to others

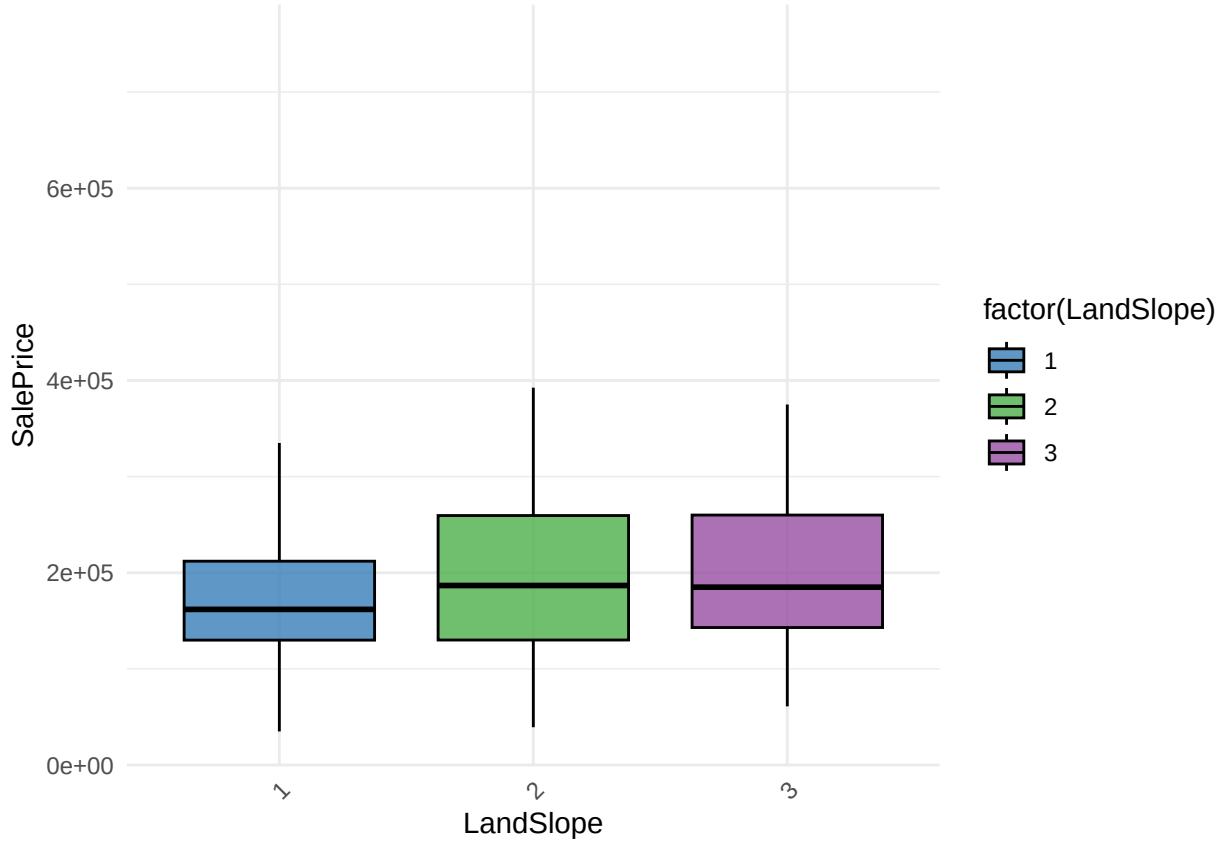
```
res <- aov(SalePrice ~ LotConfig, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## LotConfig      1 4.182e+10 4.182e+10   6.653   0.01 **
## Residuals   1458 9.166e+12 6.287e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test suggests that there is a weak relation between the two variables

11. LandSlope

```
library(ggplot2)
colors <- c("#377EB8", "#4DAF4A", "#984EA3", "#FF7F00", "#FFFF33")

ggplot(house, aes(x = factor(LandSlope), y = SalePrice, fill = factor(LandSlope))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = colors) +
  labs(x = "LandSlope", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

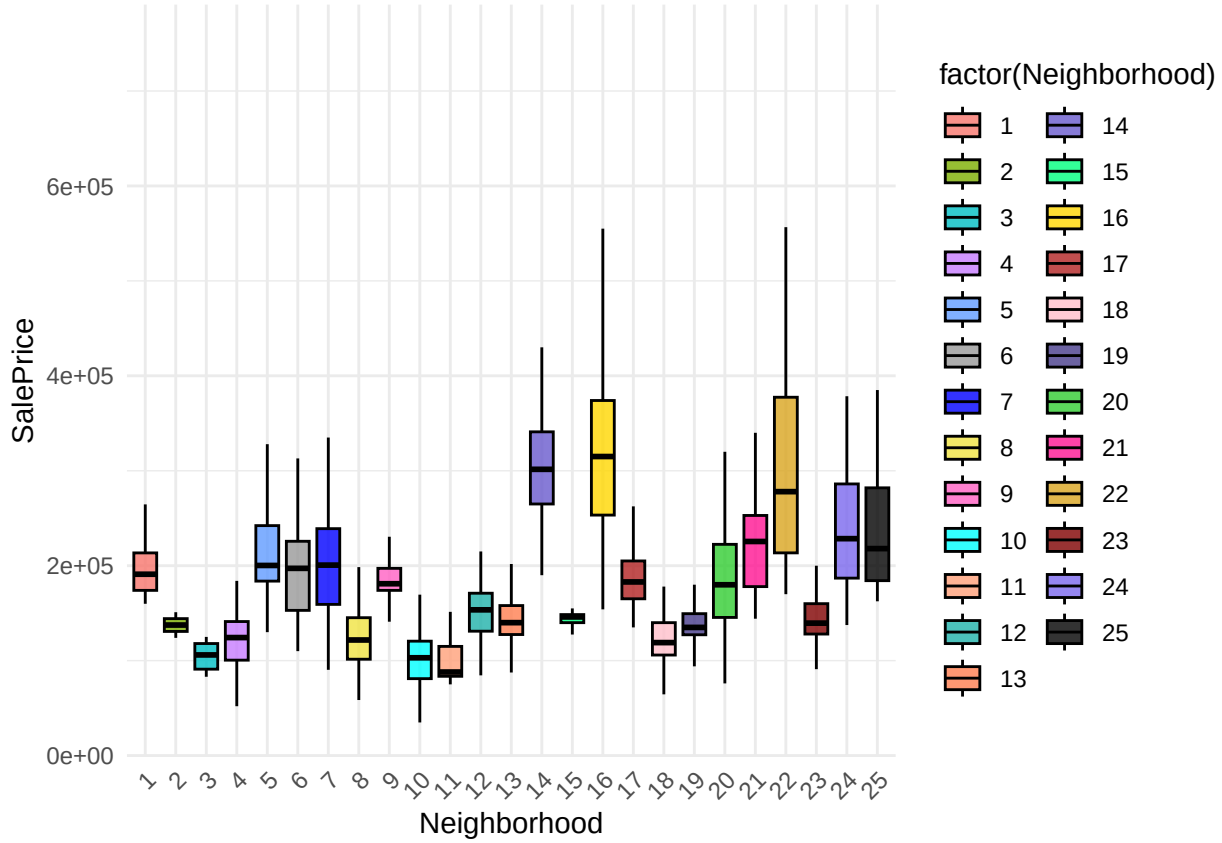



```
res <- aov(SalePrice ~ LandSlope, data = house)
summary(res)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## LandSlope      1 2.409e+10 2.409e+10   3.825 0.0507 .
## Residuals   1458 9.184e+12 6.299e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12. Neighborhood

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = factor(Neighborhood), y = SalePrice, fill = factor(Neighborhood))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Neighborhood", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



It is difficult to directly analyze something from the graph and thus we perform some tests further

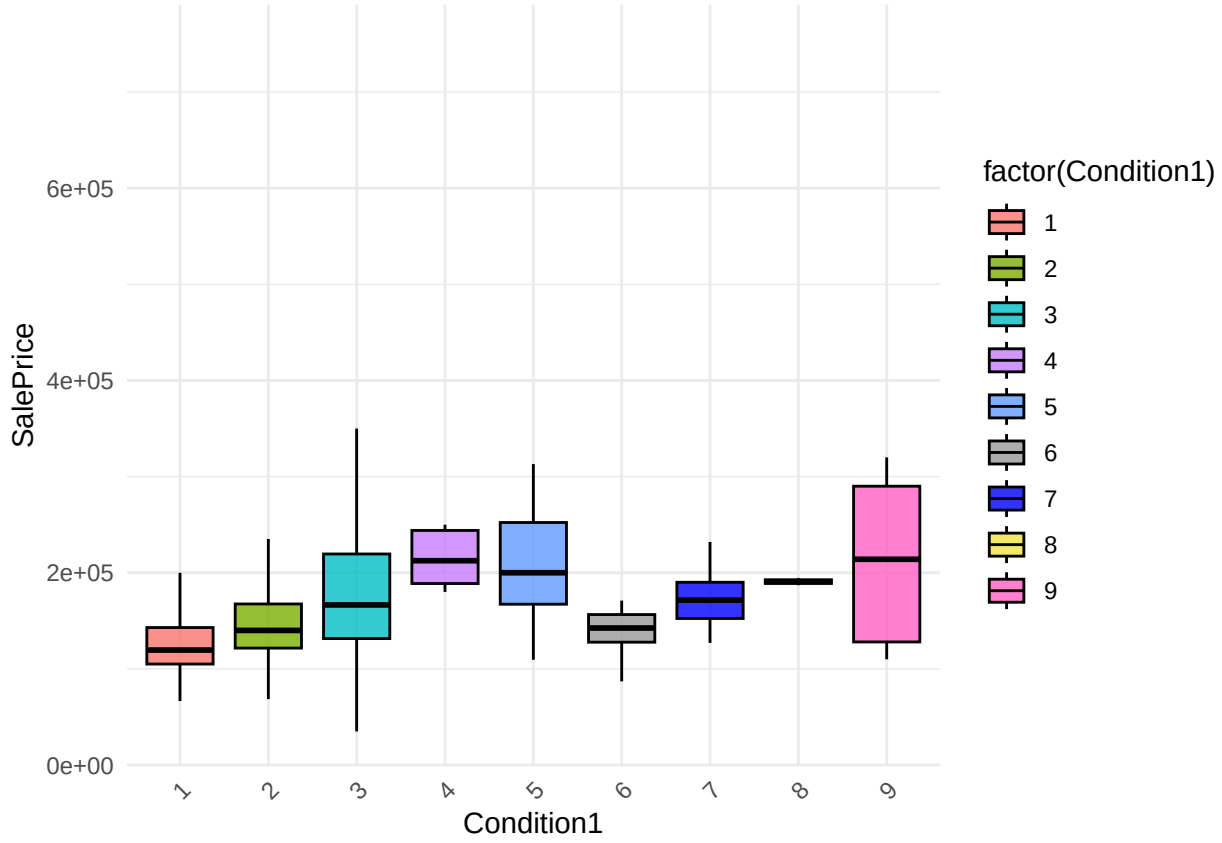
```
kruskal.test(house$Neighborhood ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Neighborhood by house$SalePrice
## Kruskal-Wallis chi-squared = 711.23, df = 662, p-value = 0.09026
```

This suggests that the target variable change with the Neighborhood ,there are certainly some relevant categories which are higher than average

13. Condition1

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = factor(Condition1), y = SalePrice, fill = factor(Condition1))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Condition1", y = "SalePrice") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



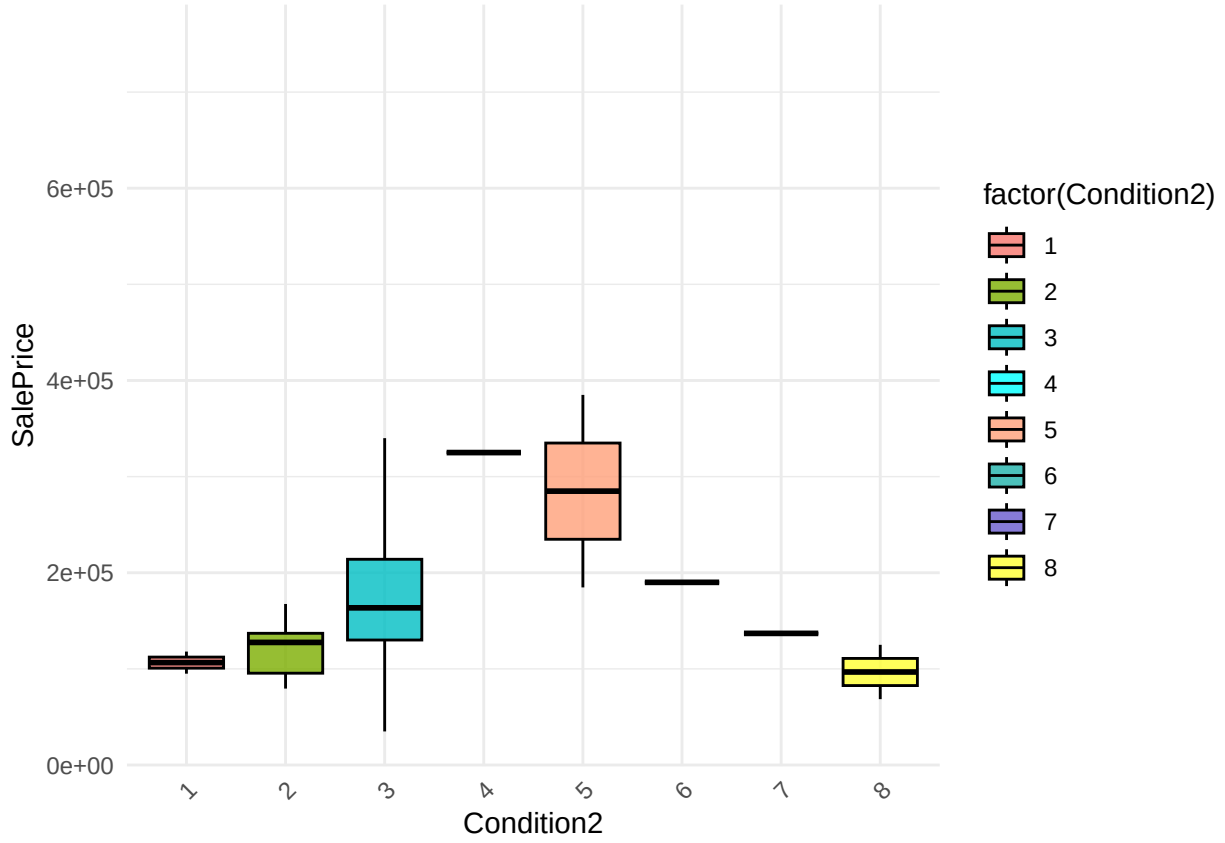

```
kruskal.test(house$Condition1 ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Condition1 by house$SalePrice
## Kruskal-Wallis chi-squared = 596.45, df = 662, p-value = 0.9676
```

the test suggests that both variables dont have a strong relation and can be used for prediction further

14. Condition2

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4","#00FFFF", "#FFA07A", "#20B2AA",  "#6A5ACD", "#FFFF33")
ggplot(house, aes(x = factor(Condition2), y = SalePrice, fill = factor(Condition2))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Condition2", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Clearly category 55 has a high SalePrice as compared to any other category

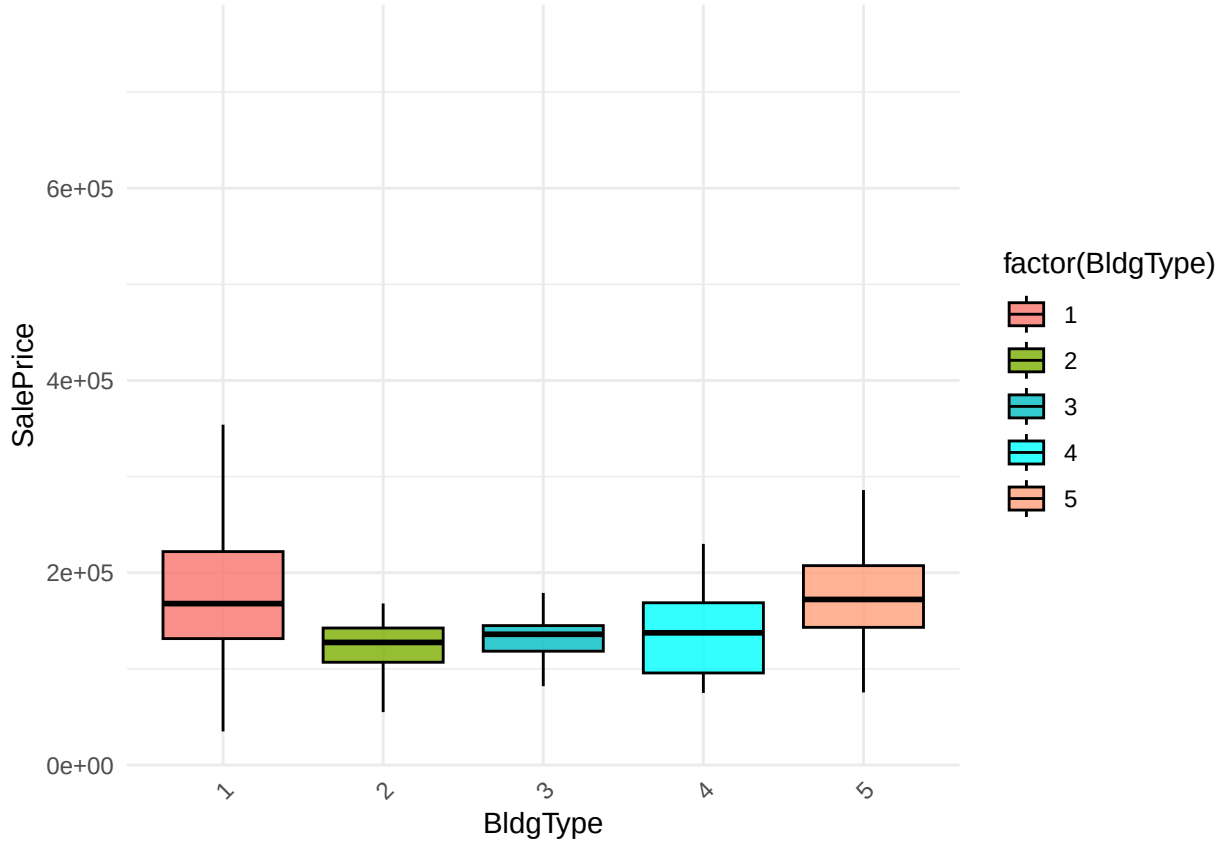
```
kruskal.test(house$Condition2 ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Condition2 by house$SalePrice
## Kruskal-Wallis chi-squared = 630.45, df = 662, p-value = 0.8059
```

This indicates that there is significant relation between the variables

*15. BldgType**

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4","#00FFFF", "#FFA07A", "#20B2AA",  "#6A5ACD", "#FFFF33")
ggplot(house, aes(x = factor(BldgType), y = SalePrice, fill = factor(BldgType))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BldgType", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



There is not much visible difference from the graph for the changes in Price

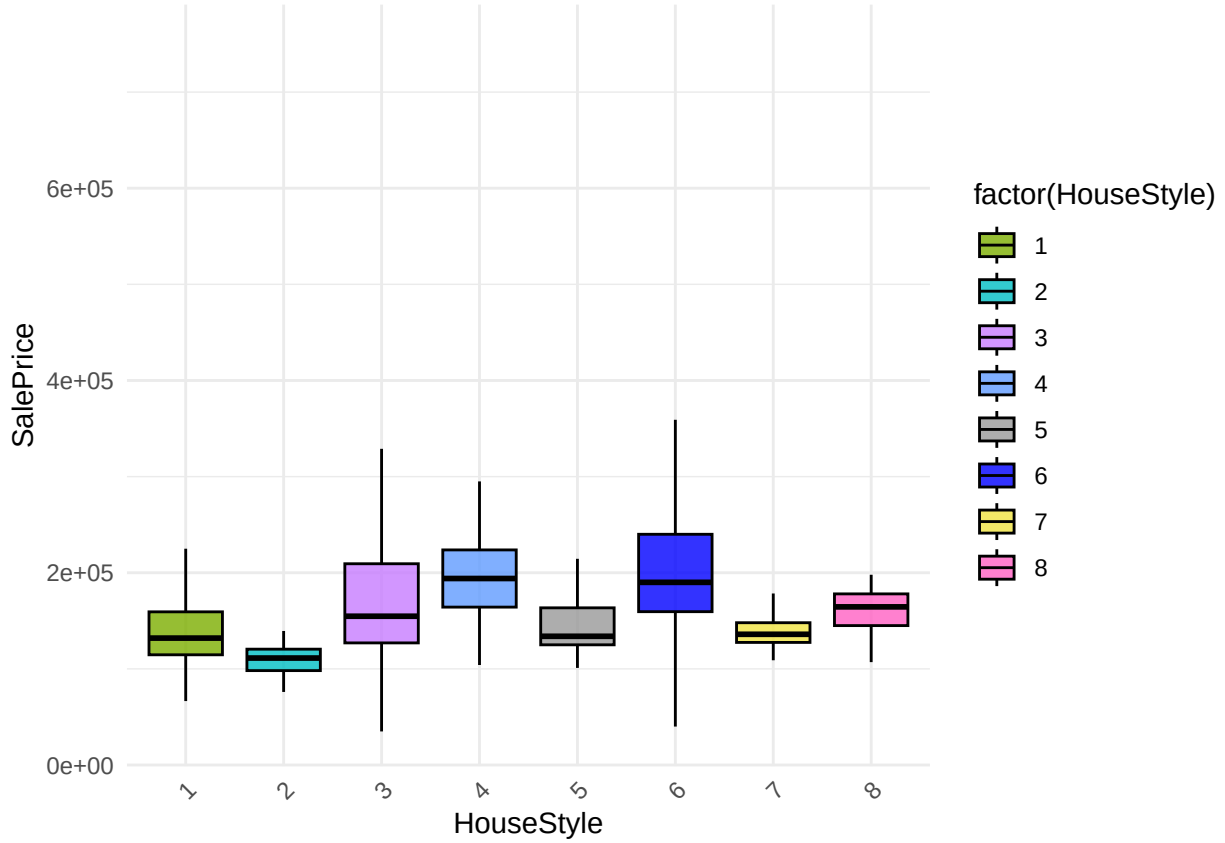
```
kruskal.test(house$BldgType ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BldgType by house$SalePrice
## Kruskal-Wallis chi-squared = 655.9, df = 662, p-value = 0.5595
```

The big p value indicates that there is not quite a difference between various categories for the target variable

16__1. HouseStyle

```
library(ggplot2)
my_colors <- c( "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#FF61C3", "#
ggplot(house, aes(x = factor(HouseStyle), y = SalePrice, fill = factor(HouseStyle))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "HouseStyle", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Clearly various categories have quite different means and values for Price

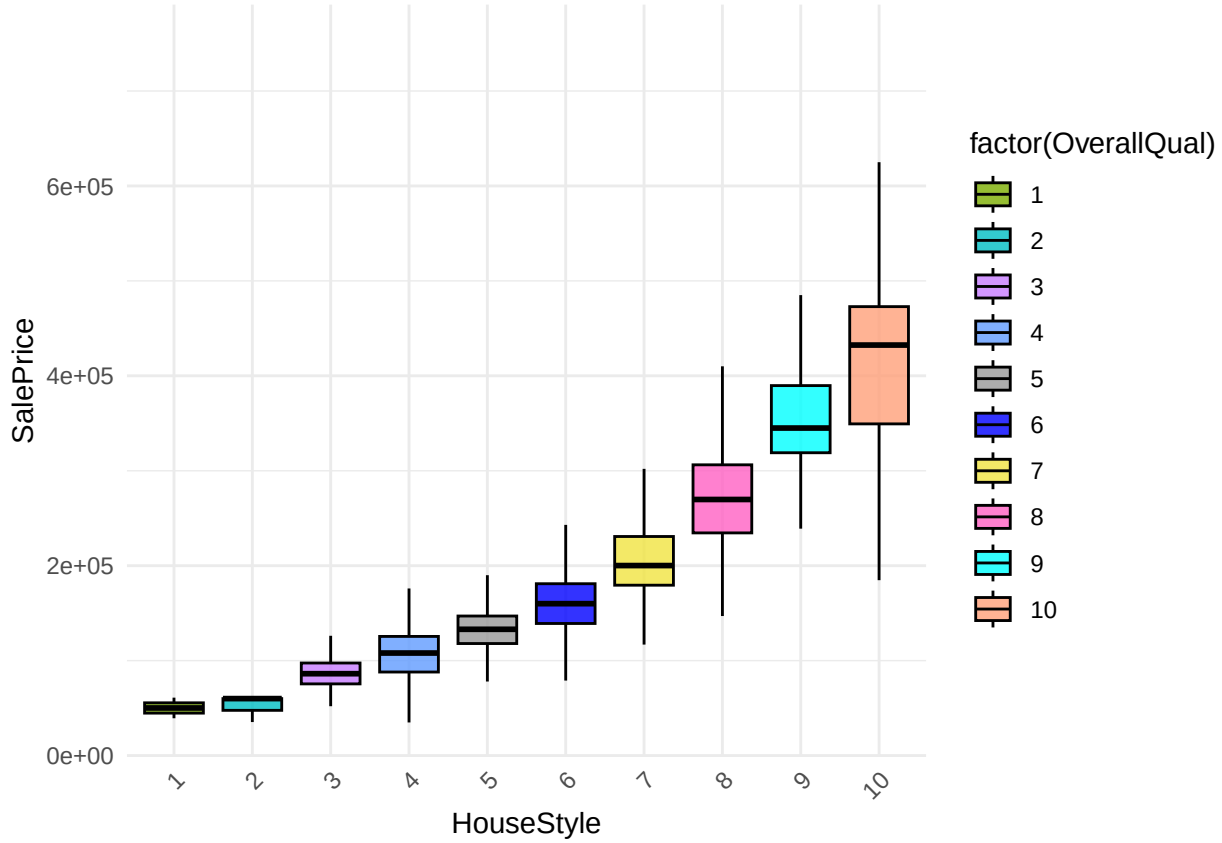
```
kruskal.test(house$HouseStyle ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$HouseStyle by house$SalePrice
## Kruskal-Wallis chi-squared = 686.23, df = 662, p-value = 0.2494
```

From the test also we see the same

17. OverallQual

```
library(ggplot2)
my_colors <- c( "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#FF61C3", "#
ggplot(house, aes(x = factor(OverallQual), y = SalePrice, fill = factor(OverallQual))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "HouseStyle", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A clear and huge difference as the rating increase the prices of the house have risen, a clear positive correlation

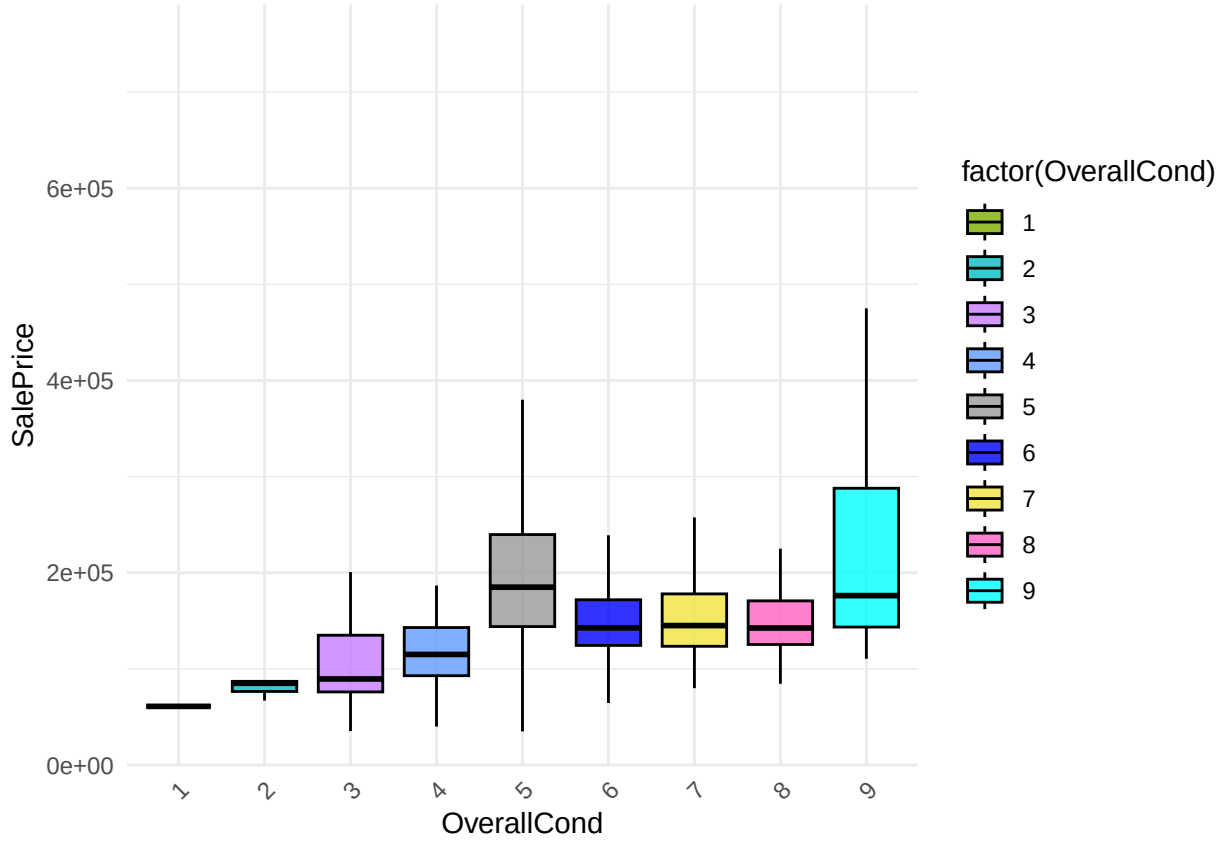
```
kruskal.test(house$OverallQual ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$OverallQual by house$SalePrice
## Kruskal-Wallis chi-squared = 1164.7, df = 662, p-value < 2.2e-16
```

The test also points out the same suggesting relation between the two variables

17. OverallCond

```
library(ggplot2)
my_colors <- c( "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#FF61C3", "#
ggplot(house, aes(x = factor(OverallCond), y = SalePrice, fill = factor(OverallCond))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "OverallCond", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



An almost increasing rating leads to higher prices except the fifth category which is the average condition category of the house

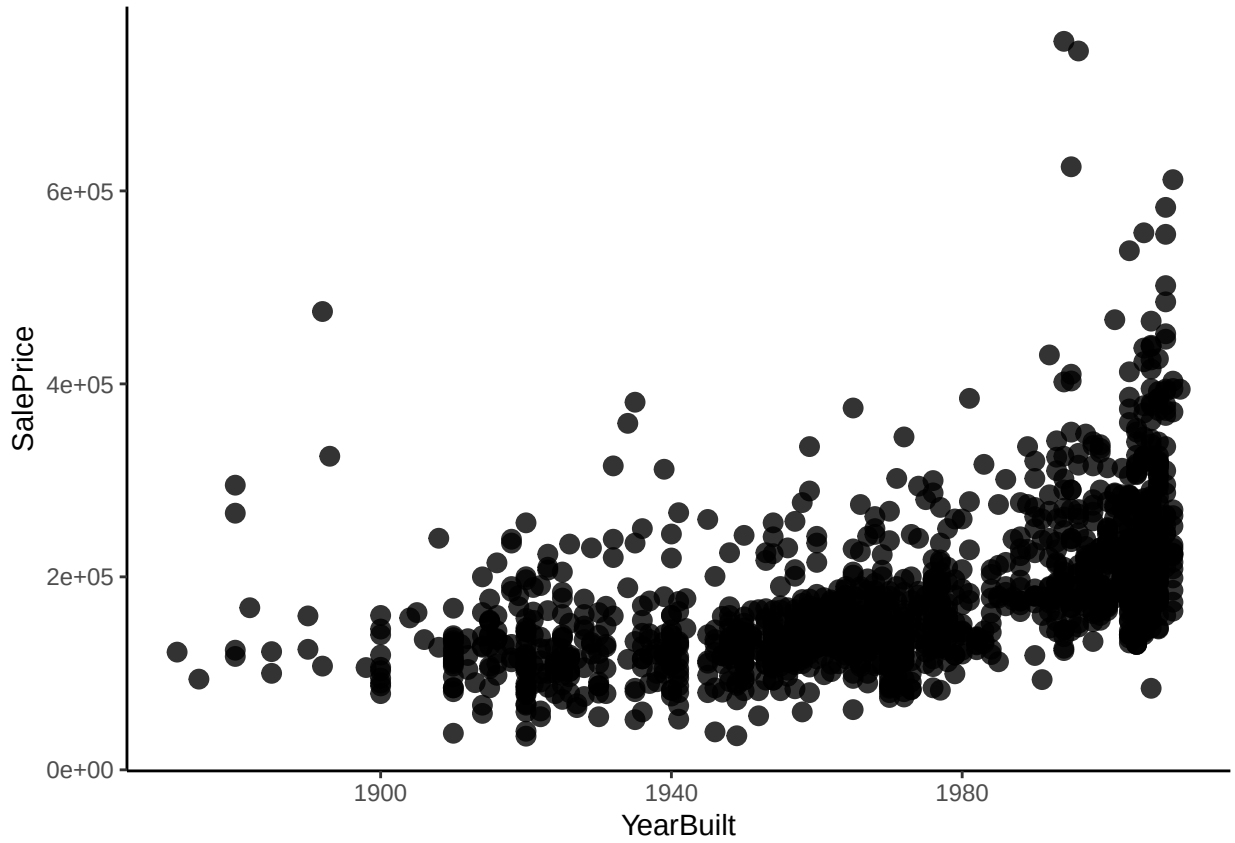
```
kruskal.test(house$OverallCond ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$OverallCond by house$SalePrice
## Kruskal-Wallis chi-squared = 679.32, df = 662, p-value = 0.312
```

The test suggests that the overall condition is not leading to a huge difference in between its categories

18. YearBuilt

```
ggplot(house, aes(x = YearBuilt, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "YearBuilt", y = "SalePrice") +
  theme_classic()
```



We can clearly see as the year increases the price have a steady increase

```
kruskal.test(house$YearBuilt ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$YearBuilt by house$SalePrice
## Kruskal-Wallis chi-squared = 1019.7, df = 662, p-value < 2.2e-16
```

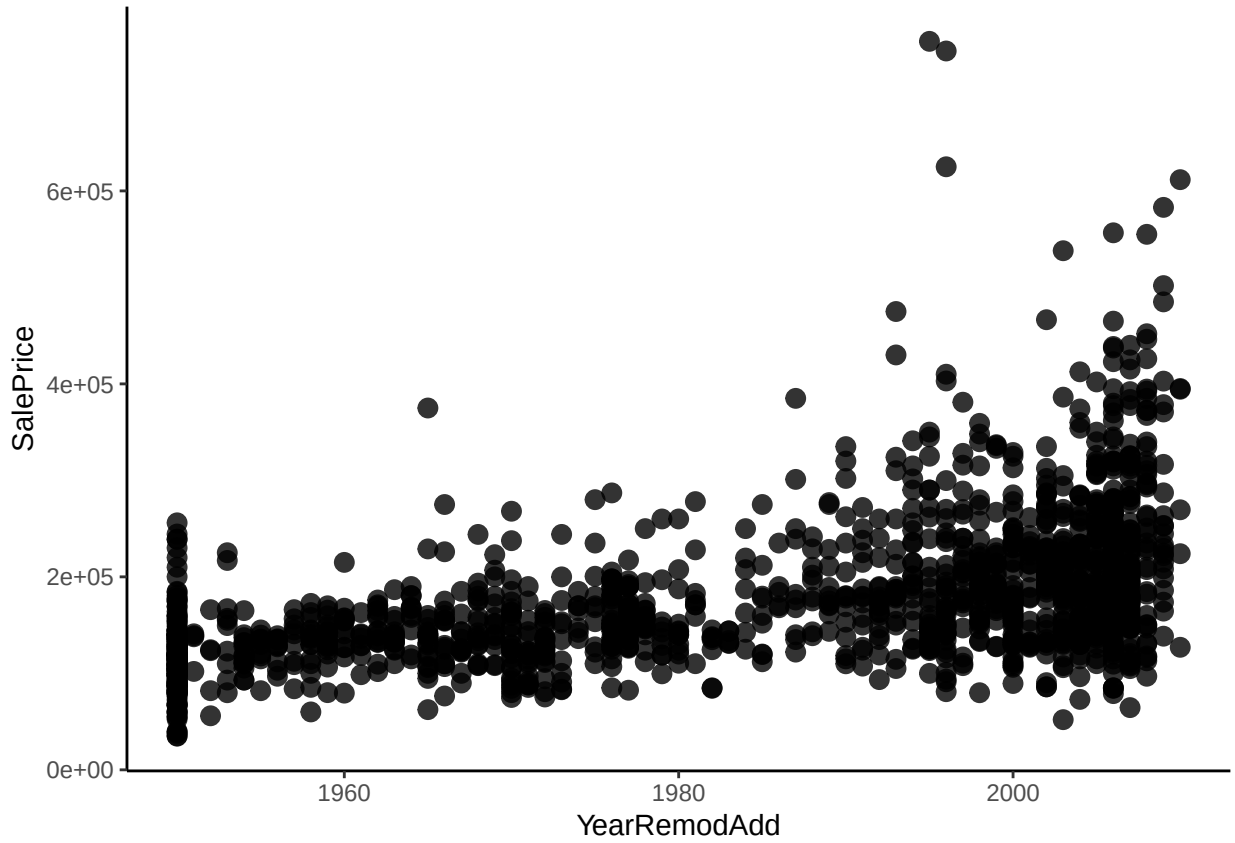
```
cor(house$YearBuilt, house$SalePrice)
```

```
## [1] 0.5228973
```

A very good correlation indicating linear increase with year and also a very small p value indicating the strong relation with the target variable

19. YearRemodAdd

```
ggplot(house, aes(x = YearRemodAdd, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "YearRemodAdd", y = "SalePrice") +
  theme_classic()
```



again a similar anayais to year built

```
kruskal.test(house$YearRemodAdd ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$YearRemodAdd by house$SalePrice
## Kruskal-Wallis chi-squared = 909.23, df = 662, p-value = 4.875e-10
```

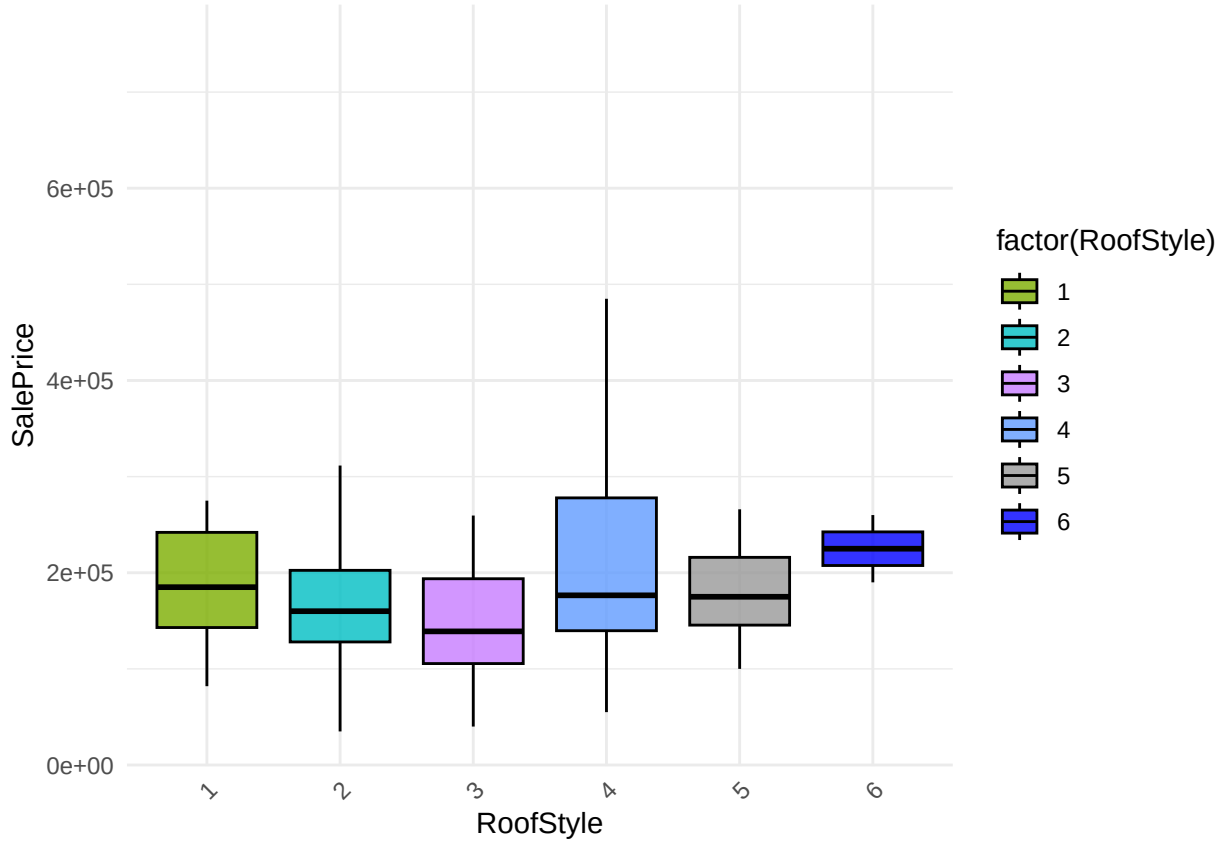
```
cor(house$YearRemodAdd, house$SalePrice)
```

```
## [1] 0.507101
```

Similar to year built

20. RoofStyle

```
library(ggplot2)
my_colors <- c( "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#FF61C3", "#
ggplot(house, aes(x = factor(RoofStyle), y = SalePrice, fill = factor(RoofStyle))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "RoofStyle", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



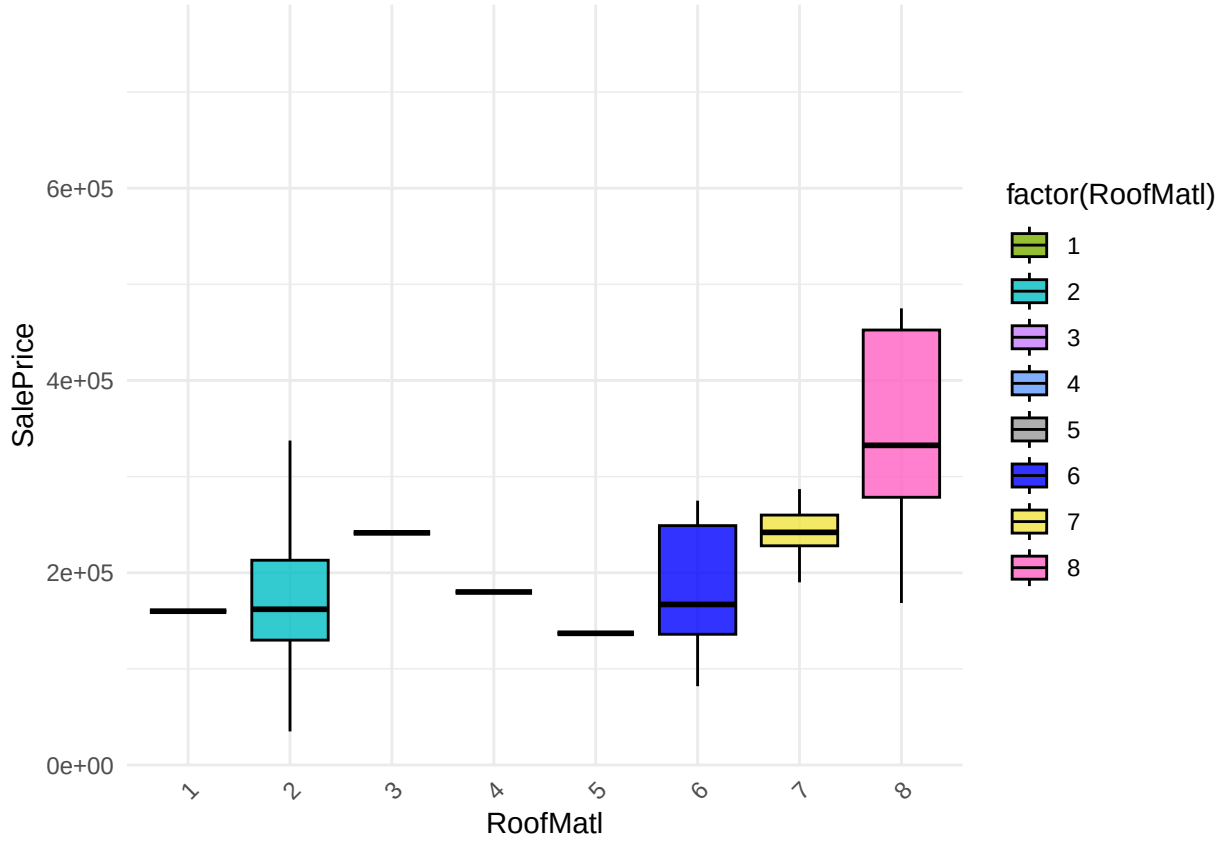
Not much difference across categories

```
kruskal.test(house$RoofStyle ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$RoofStyle by house$SalePrice
## Kruskal-Wallis chi-squared = 728.88, df = 662, p-value = 0.0361
```

21. RoofMatl

```
library(ggplot2)
my_colors <- c( "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#FF61C3", "#
ggplot(house, aes(x = factor(RoofMatl), y = SalePrice, fill = factor(RoofMatl))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "RoofMatl", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Very vivid output having different prices for 8th category(high price)

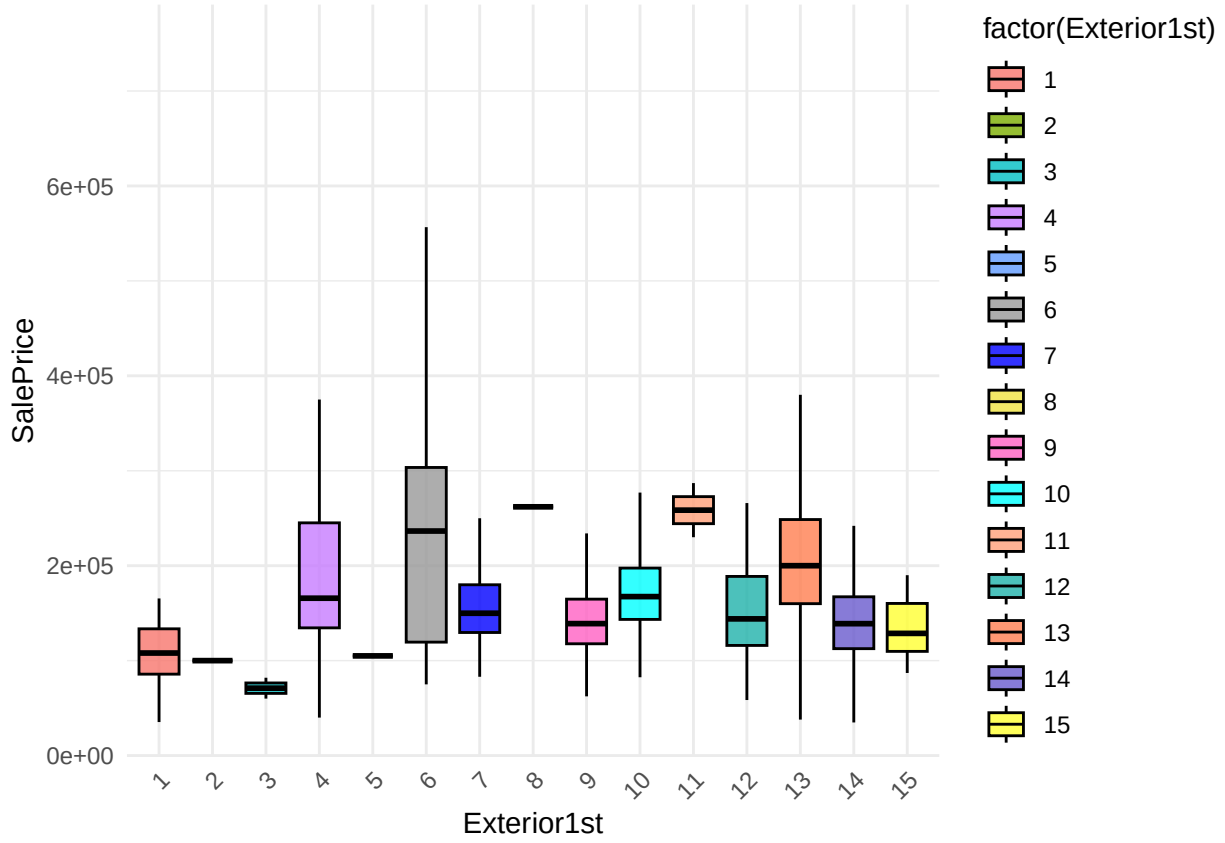
```
kruskal.test(house$RoofMatl ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$RoofMatl by house$SalePrice
## Kruskal-Wallis chi-squared = 602.46, df = 662, p-value = 0.9525
```

Clearly the values suggest no relation between the two variables

21__1. Exterior1st

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = factor(Exterior1st), y = SalePrice, fill = factor(Exterior1st))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Exterior1st", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



QUite some variation among categories for the prices of house

```
kruskal.test(house$Exterior1st ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Exterior1st by house$SalePrice
## Kruskal-Wallis chi-squared = 630.6, df = 662, p-value = 0.8047
```

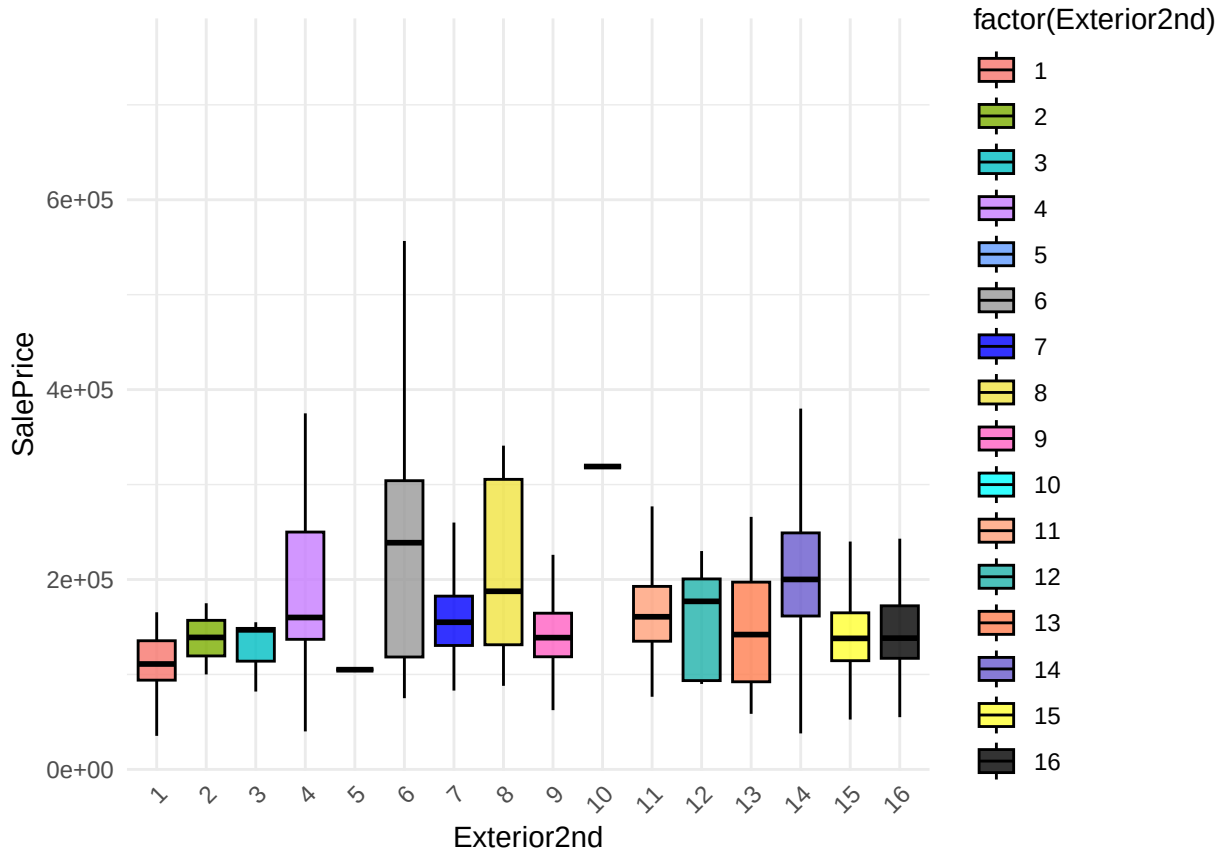
```
summary(aov(Exterior1st~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1    160  159.97    15.8 7.37e-05 ***
## Residuals   1458  14758   10.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

the ANOVA test suggests a relation with the target variable because of mean differences between various categories but the Kruskal wallis test looks on the medians of the categories which would not be quite different

22. Exterior2nd

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = factor(Exterior2nd), y = SalePrice, fill = factor(Exterior2nd))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Exterior2nd", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Similar analysis like the previous variable

```
kruskal.test(house$Exterior2nd ~ house$SalePrice)
```

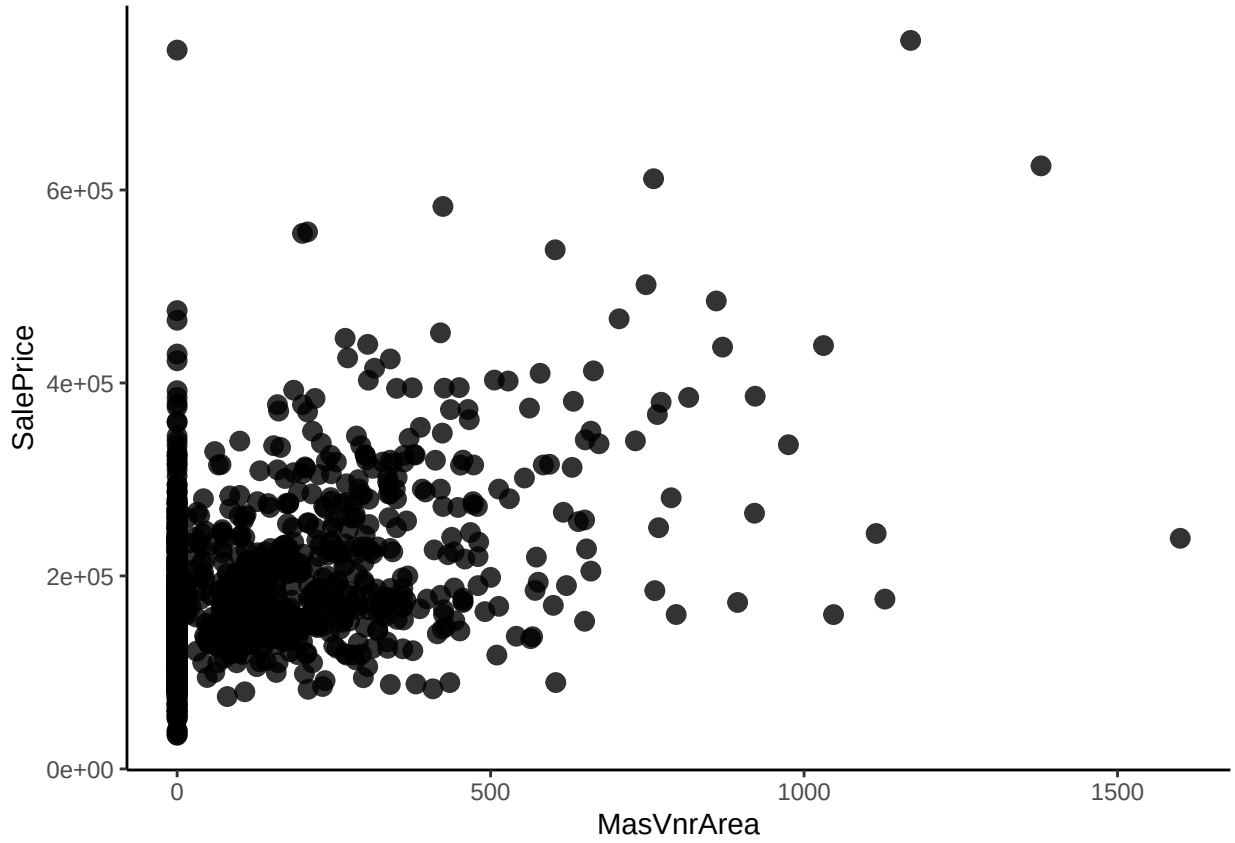
```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Exterior2nd by house$SalePrice
## Kruskal-Wallis chi-squared = 648.27, df = 662, p-value = 0.6412
```

```
summary(aov(Exterior2nd~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1    197  196.93   15.87 7.12e-05 ***
## Residuals   1458  18093   12.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

23. MasVnrArea

```
ggplot(house, aes(x = MasVnrArea, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "MasVnrArea", y = "SalePrice") +
  theme_classic()
```



This suggests that the prices increase gradually over the Area of veneer

```
kruskal.test(house$MasVnrArea ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$MasVnrArea by house$SalePrice
## Kruskal-Wallis chi-squared = 815.17, df = 662, p-value = 4.061e-05
```

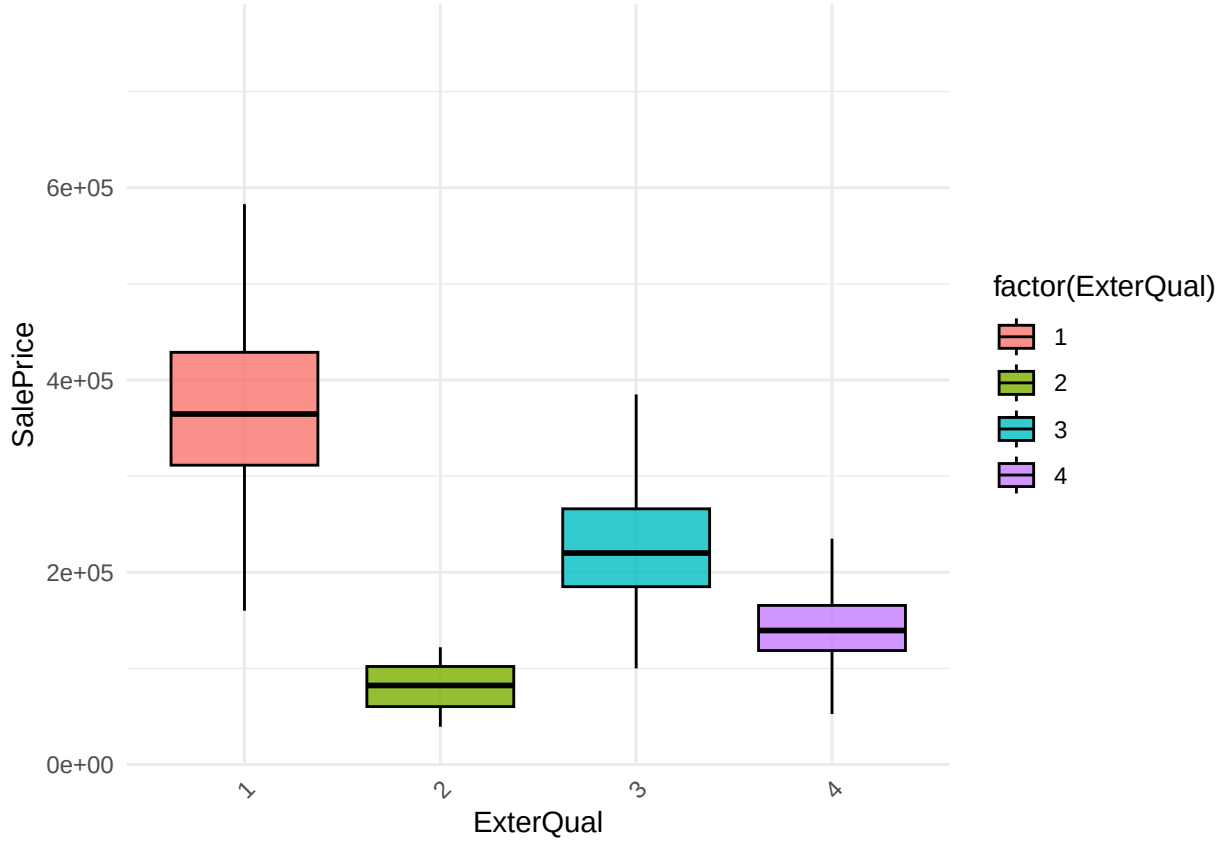
```
cor(house$MasVnrArea, house$SalePrice)
```

```
## [1] 0.4726145
```

The results also suggest a strong relation between them

24. ExterQual

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = factor(ExterQual), y = SalePrice, fill = factor(ExterQual))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "ExterQual", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A big difference between various categories

```
kruskal.test(house$ExterQual ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$ExterQual by house$SalePrice
## Kruskal-Wallis chi-squared = 1035.2, df = 662, p-value < 2.2e-16
```

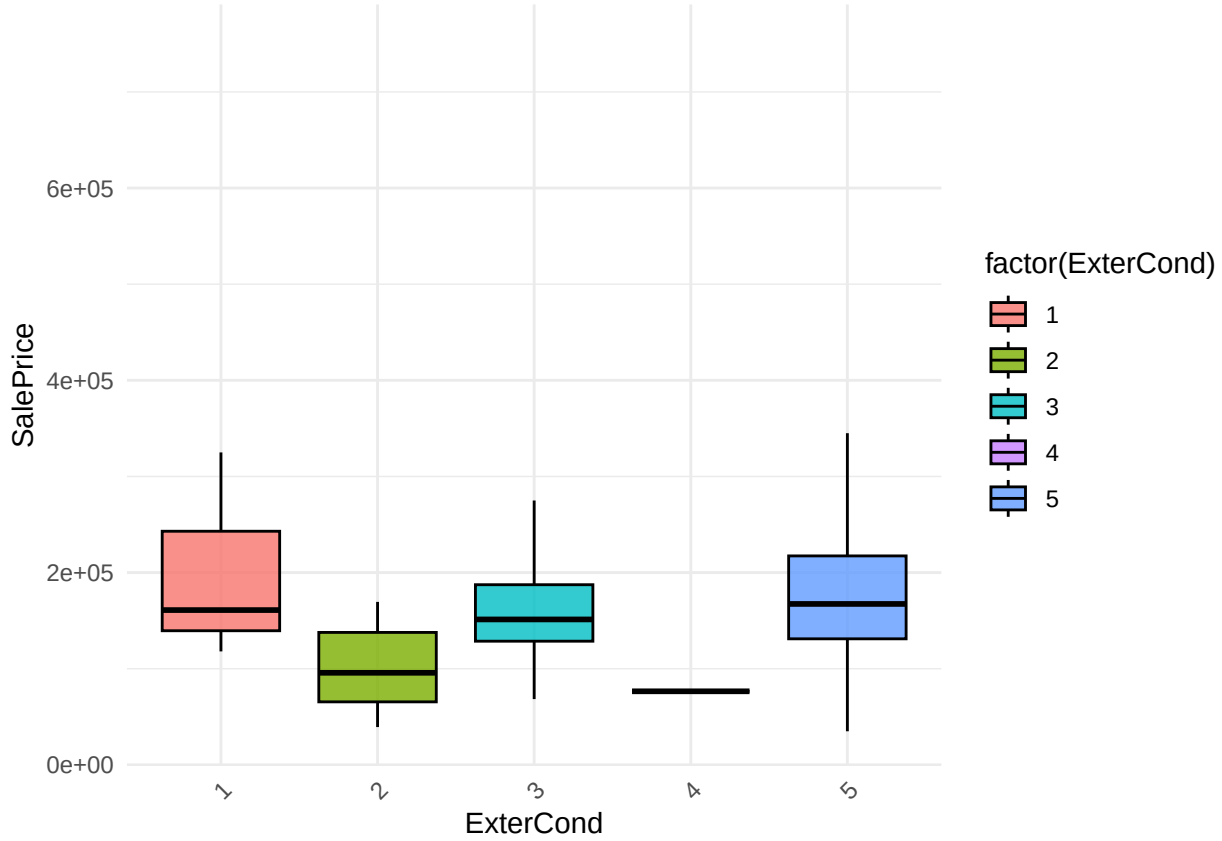
```
summary(aov(ExterQual~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  285.0  285.03     995 <2e-16 ***
## Residuals   1458  417.7    0.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both tests indicate a very strong relation between the two variables

25. ExterCond

```
library(ggplot2)
my_colors <- c("#F8766D", "#7CAE00", "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442", "#F
ggplot(house, aes(x = factor(ExterCond), y = SalePrice, fill = factor(ExterCond))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "ExterCond", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Different categories have different prices cannot infer much from it

```
kruskal.test(house$ExterCond ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$ExterCond by house$SalePrice
## Kruskal-Wallis chi-squared = 605.05, df = 662, p-value = 0.9445
```

```
summary(aov(ExterCond~SalePrice,data=house))
```

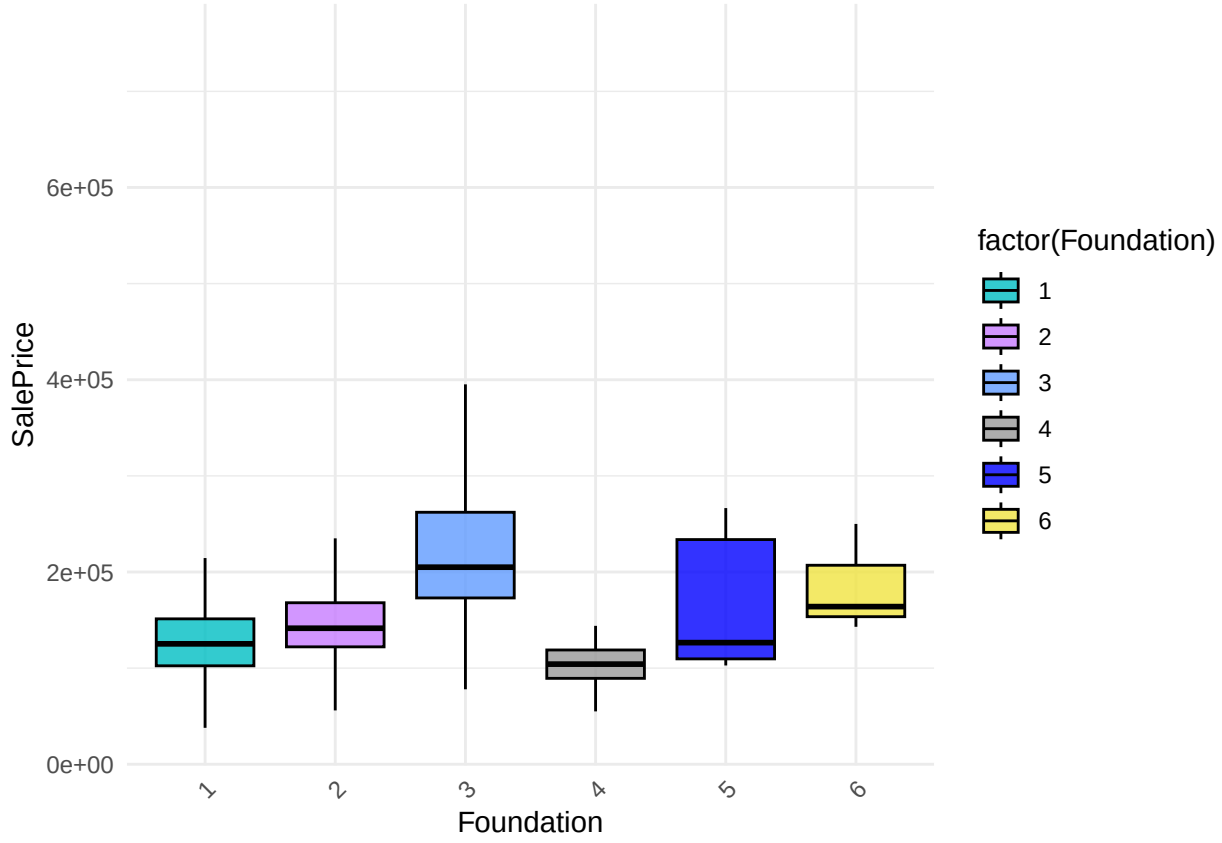
```
##               Df Sum Sq Mean Sq F value Pr(>F)
```

```
## SalePrice      1   10.8  10.751   20.34  7e-06 ***
## Residuals   1458  770.6   0.529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This shows that the mean of categories varies largely with the price but no the medians hence a realtion with the Saleprice is evident

26. Foundation

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue", "#F0E442",  "#00FFFF", "#FFA07A", "#
ggplot(house, aes(x = factor(Foundation), y = SalePrice, fill = factor(Foundation))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Foundation", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$Foundation ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Foundation by house$SalePrice
## Kruskal-Wallis chi-squared = 853.34, df = 662, p-value = 6.458e-07
```

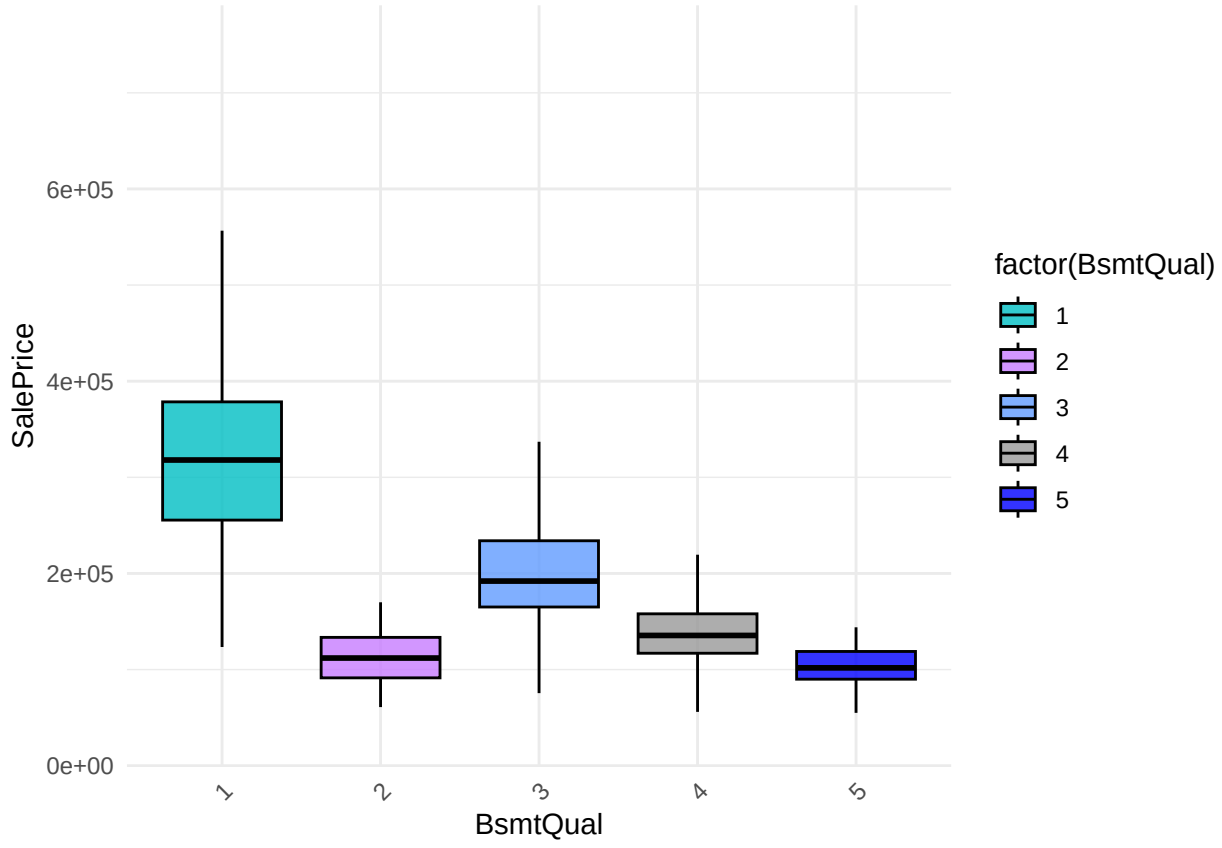
```
summary(aov(Foundation~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  111.4  111.38   249.8 <2e-16 ***
## Residuals   1458  650.0    0.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

both the tests suggest a strong relation with the Saleprice

27. BsmtQual

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtQual), y = SalePrice, fill = factor(BsmtQual))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtQual", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$BsmtQual ~ house$SalePrice)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtQual by house$SalePrice
## Kruskal-Wallis chi-squared = 963.18, df = 662, p-value = 1.503e-13
```

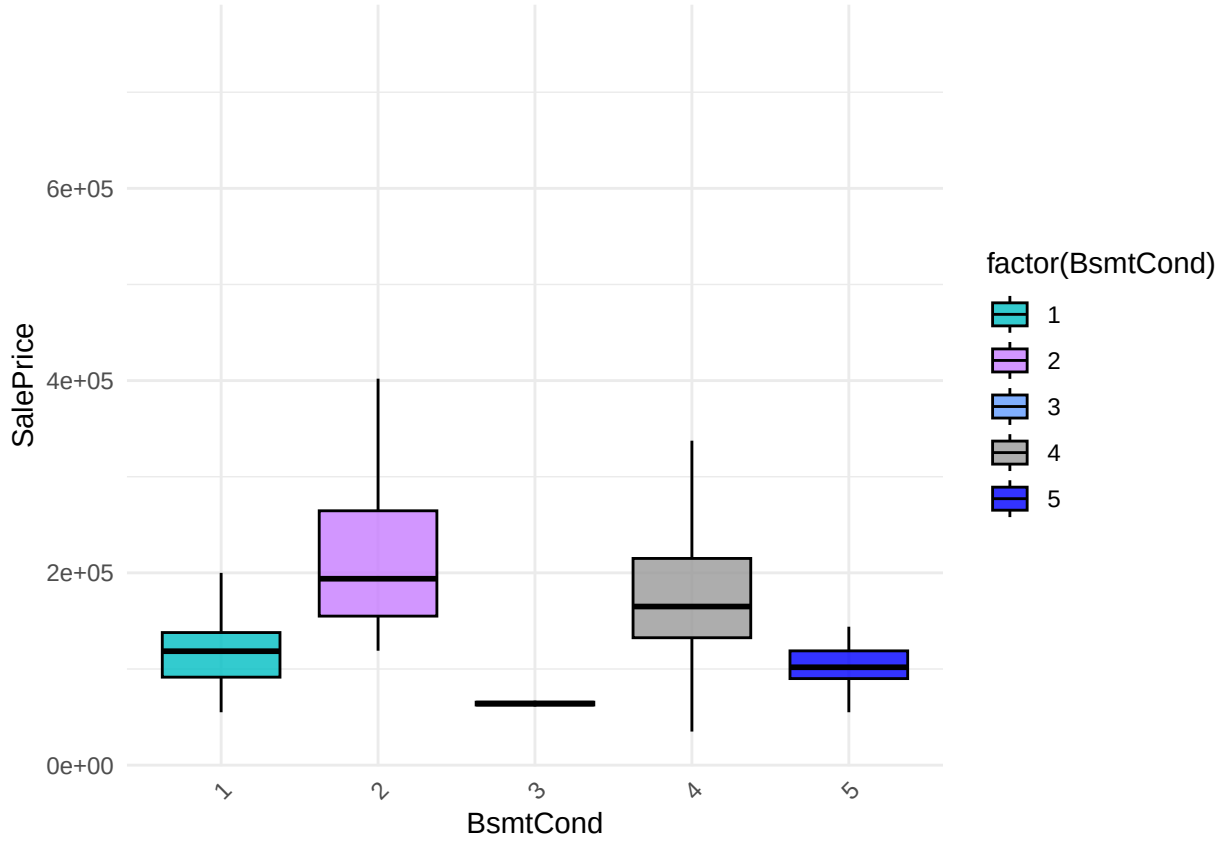
```
summary(aov(BsmtQual~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  454.8   454.8   914.7 <2e-16 ***
## Residuals   1458  725.0     0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A small p value for both clearly indicates the dtrong relation with the prediction variable

28. BsmtCond

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtCond), y = SalePrice, fill = factor(BsmtCond))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtCond", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A difference is very visible in the various categories

```
kruskal.test(house$BsmtCond ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtCond by house$SalePrice
## Kruskal-Wallis chi-squared = 737.64, df = 662, p-value = 0.02155
```

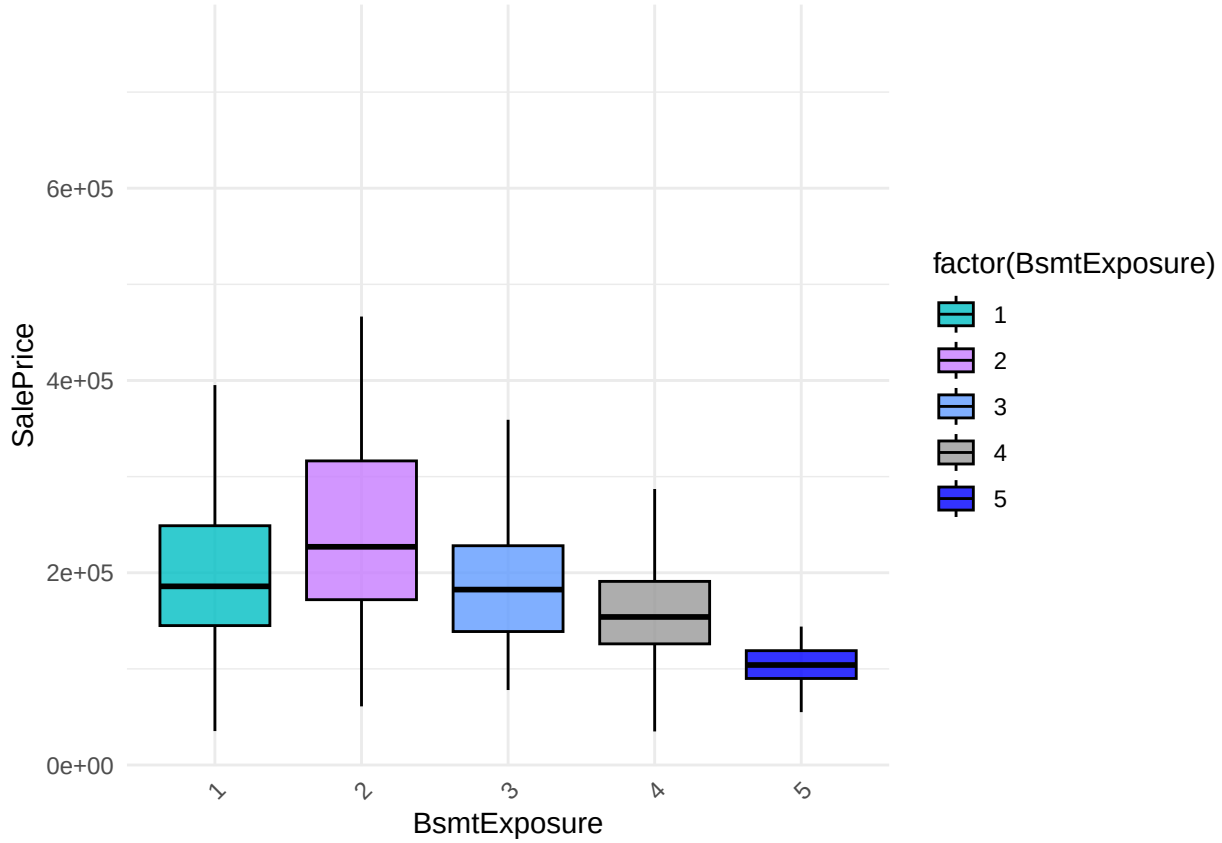
```
summary(aov(BsmtCond~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1    0.2  0.1514   0.331  0.565
## Residuals   1458  667.6  0.4579
```

A small p value with regards to median suggests that there is a chance of outliers also

29. BsmtExposure

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtExposure), y = SalePrice, fill = factor(BsmtExposure))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtExposure", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Clearly a decreasing trend in the geraph is visible as the exposure decrease price decreases

```
kruskal.test(house$BsmtExposure ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtExposure by house$SalePrice
## Kruskal-Wallis chi-squared = 798.84, df = 662, p-value = 0.0001949
```

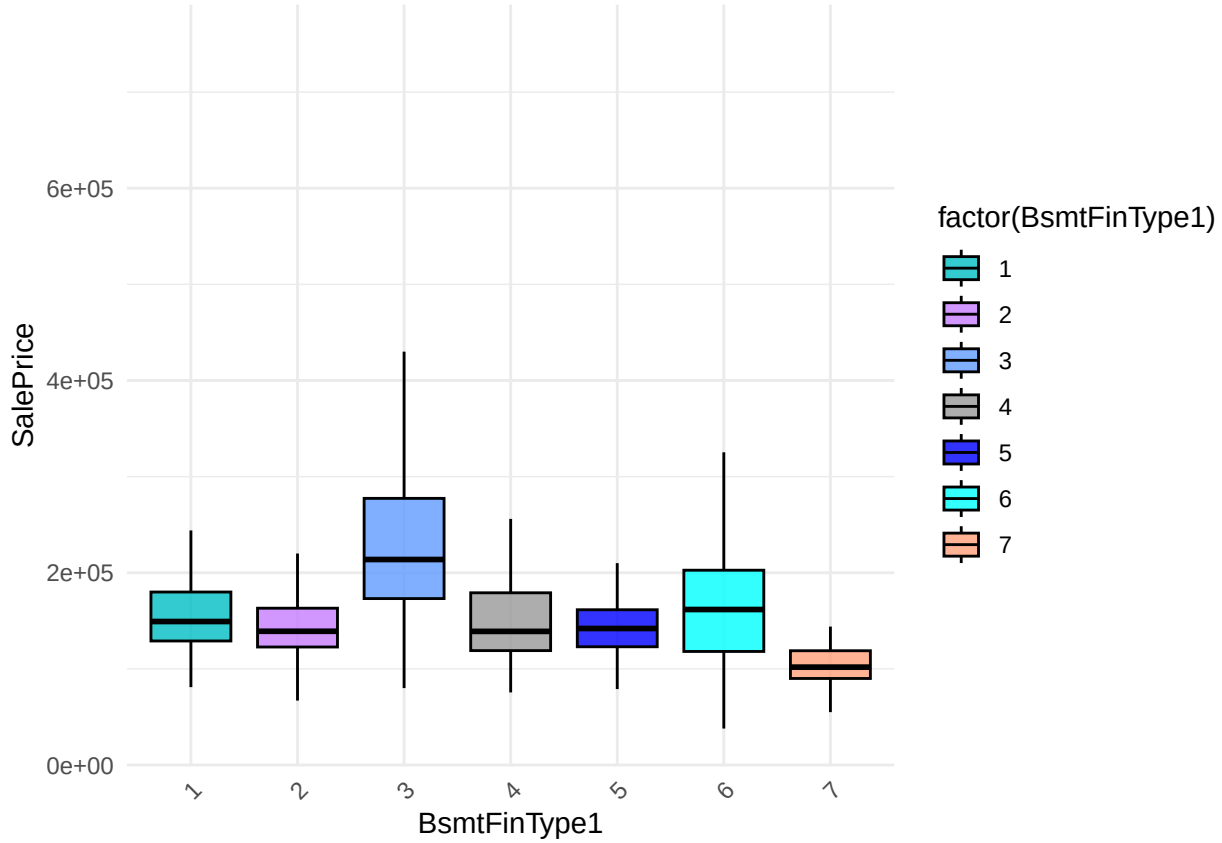
```
summary(aov(BsmtExposure~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  189.3  189.34     154 <2e-16 ***
## Residuals   1458 1793.1    1.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Very small p values also sugest the same

30. BsmtFinType1

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtFinType1), y = SalePrice, fill = factor(BsmtFinType1))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtFinType1", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Something cannot be deduced directly from the graph

```
kruskal.test(house$BsmtFinType1 ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtFinType1 by house$SalePrice
## Kruskal-Wallis chi-squared = 668.38, df = 662, p-value = 0.4235
```

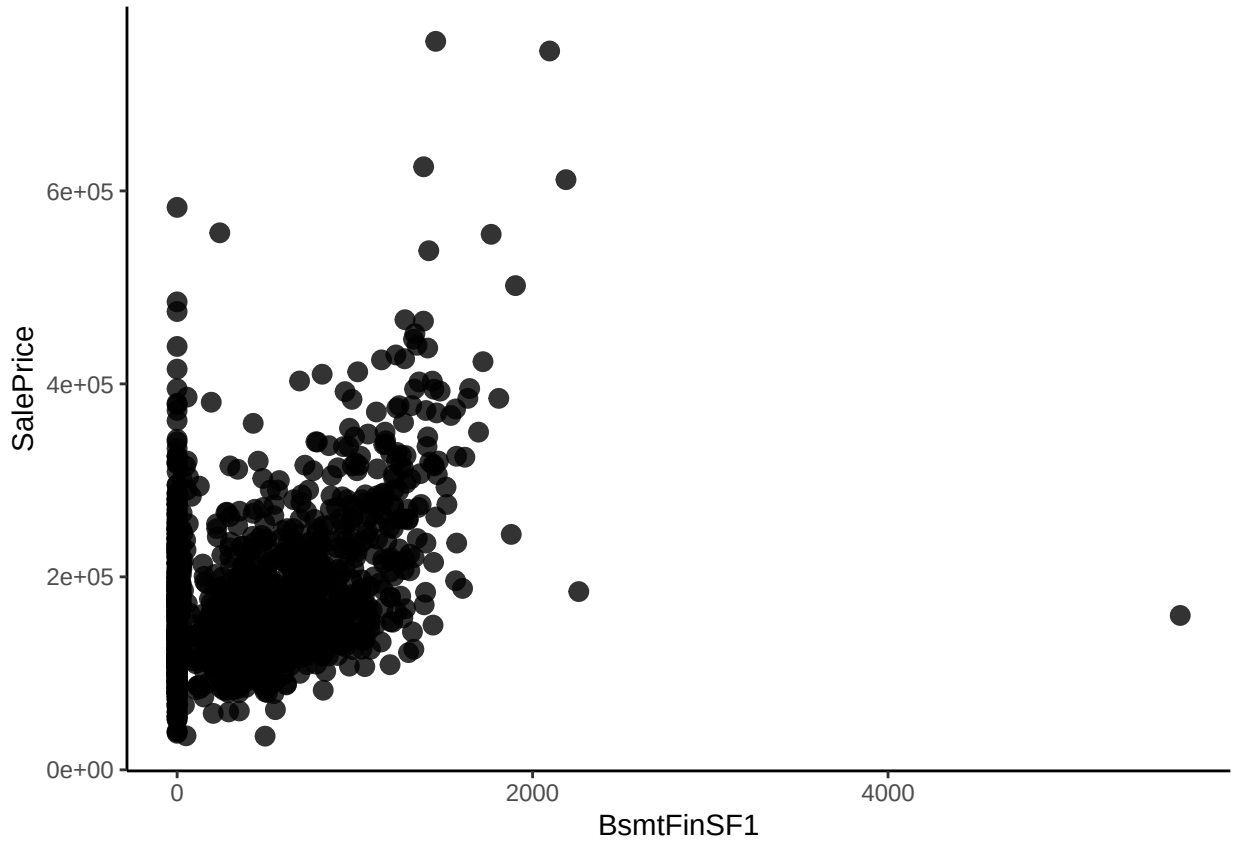
```
summary(aov(BsmtFinType1~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1     55   54.50   15.67 7.91e-05 ***
## Residuals   1458   5072    3.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

with respect to mean there is a difference and no presence of maybe outliers due to lack of differentiation in median

31. BsmtFinSF1

```
ggplot(house, aes(x = BsmtFinSF1, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "BsmtFinSF1", y = "SalePrice") +
  theme_classic()
```



This suggests that the prices increase gradually over the finished square feet

```
kruskal.test(house$BsmtFinSF1 ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtFinSF1 by house$SalePrice
## Kruskal-Wallis chi-squared = 835.52, df = 662, p-value = 4.84e-06
```

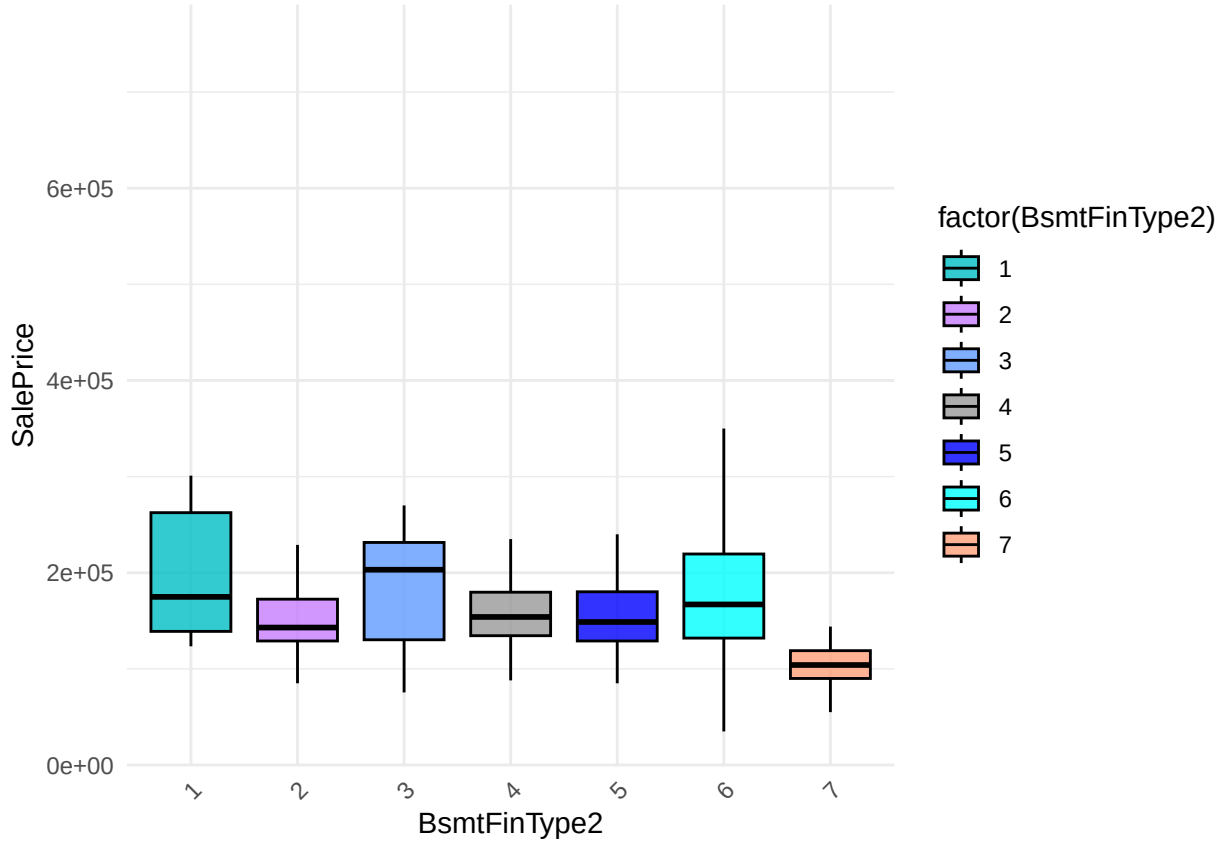
```
cor(house$BsmtFinSF1, house$SalePrice)
```

```
## [1] 0.3864198
```

A positive correlation also suggests a linear positive relation with the target variable along with a small p value

32. BsmtFinType2

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtFinType2), y = SalePrice, fill = factor(BsmtFinType2))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtFinType2", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A very less difference can be seen from the graph

```
kruskal.test(house$BsmtFinType1 ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtFinType1 by house$SalePrice
## Kruskal-Wallis chi-squared = 668.38, df = 662, p-value = 0.4235
```

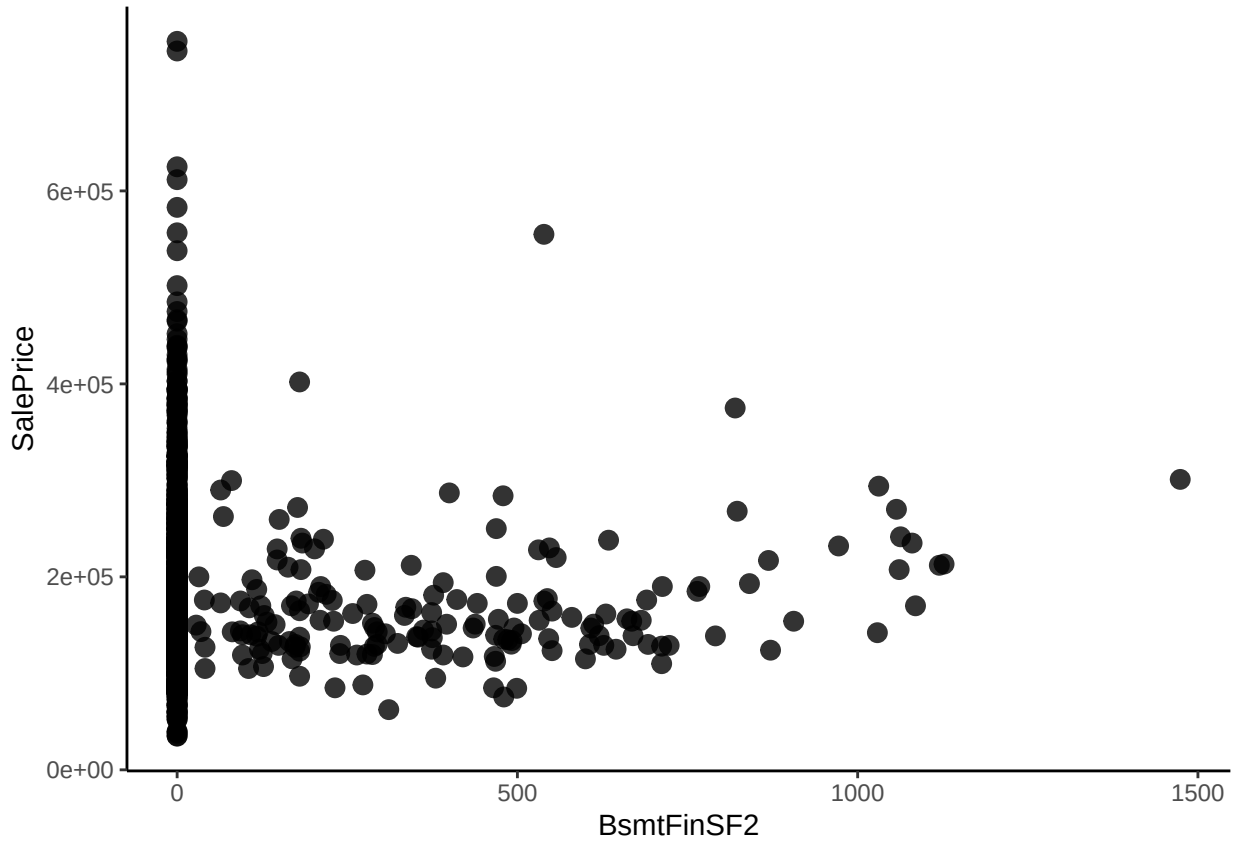
```
summary(aov(BsmtFinType1~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1     55   54.50   15.67 7.91e-05 ***
## Residuals   1458   5072    3.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a small p value for mean also indicates linear type relation with the target variable

33. BsmtFinSF2

```
ggplot(house, aes(x = BsmtFinSF2, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "BsmtFinSF2", y = "SalePrice") +
  theme_classic()
```



This suggests that the prices decrease once and then similar over the finished square feet

```
kruskal.test(house$BsmtFinSF2 ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtFinSF2 by house$SalePrice
## Kruskal-Wallis chi-squared = 578.74, df = 662, p-value = 0.9912
```

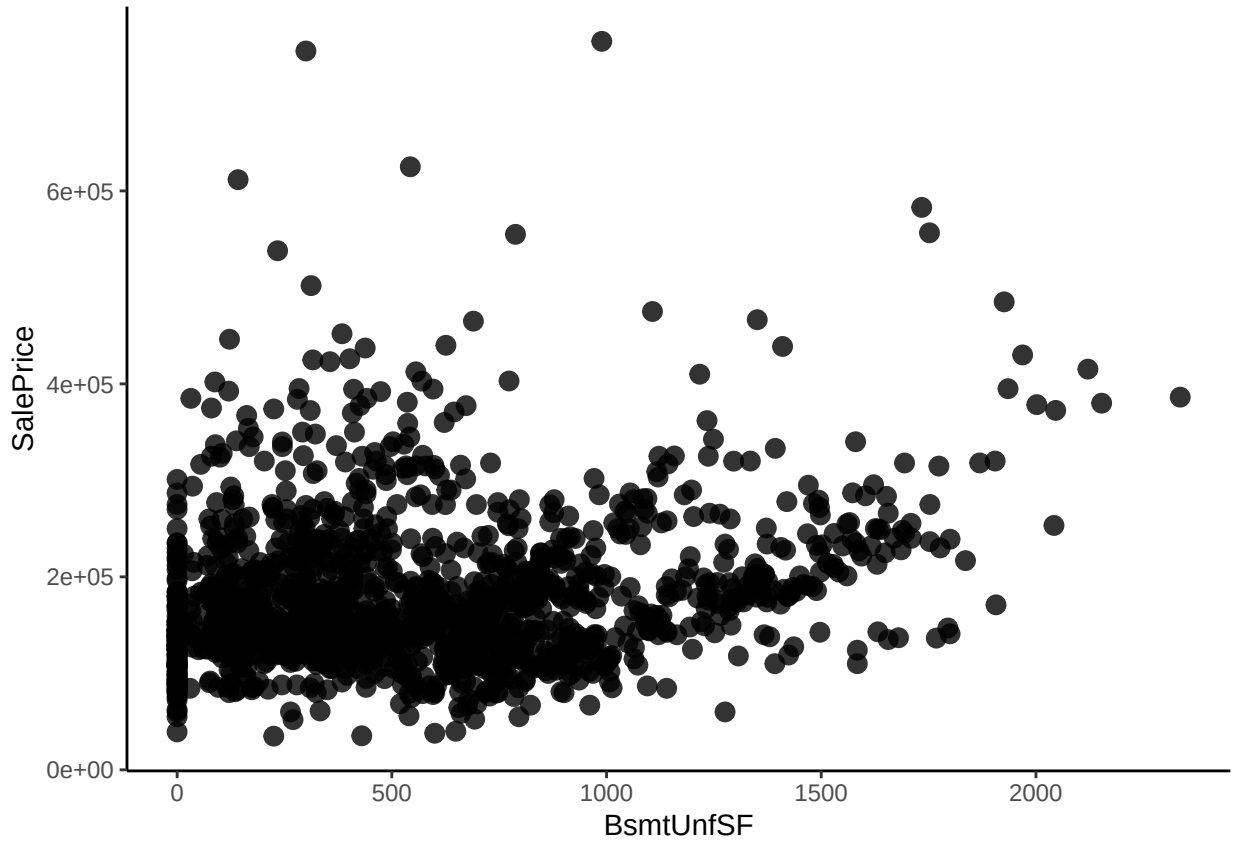
```
cor(house$BsmtFinSF2, house$SalePrice)
```

```
## [1] -0.01137812
```

Clearly there is a negative relation with the output because of sudden decrease

34. BsmtUnfSF

```
ggplot(house, aes(x = BsmtUnfSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "BsmtUnfSF", y = "SalePrice") +
  theme_classic()
```



A very gradual increase with the uninfished square feet area(maybe because poeple want to build their own house)

```
kruskal.test(house$BsmtUnfSF ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtUnfSF by house$SalePrice
## Kruskal-Wallis chi-squared = 701.04, df = 662, p-value = 0.1422
```

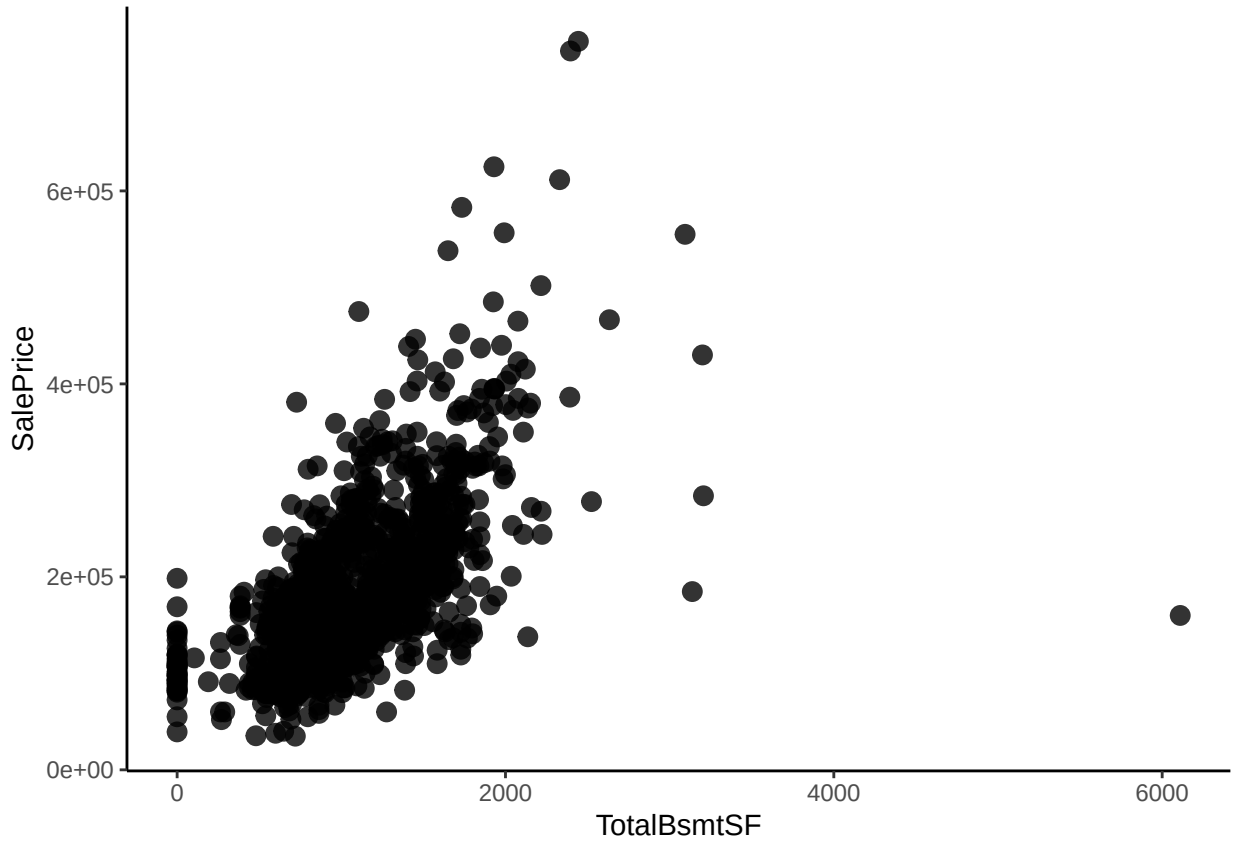
```
cor(house$BsmtUnfSF, house$SalePrice)
```

```
## [1] 0.2144791
```

A positive correlation but not a very strong relation with the target variable

35. TotalBsmtSF

```
ggplot(house, aes(x = TotalBsmtSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "TotalBsmtSF", y = "SalePrice") +
  theme_classic()
```



a very big increase in the price as the total square feet increases

```
kruskal.test(house$TotalBsmtSF ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$TotalBsmtSF by house$SalePrice
## Kruskal-Wallis chi-squared = 945.67, df = 662, p-value = 2.329e-12
```

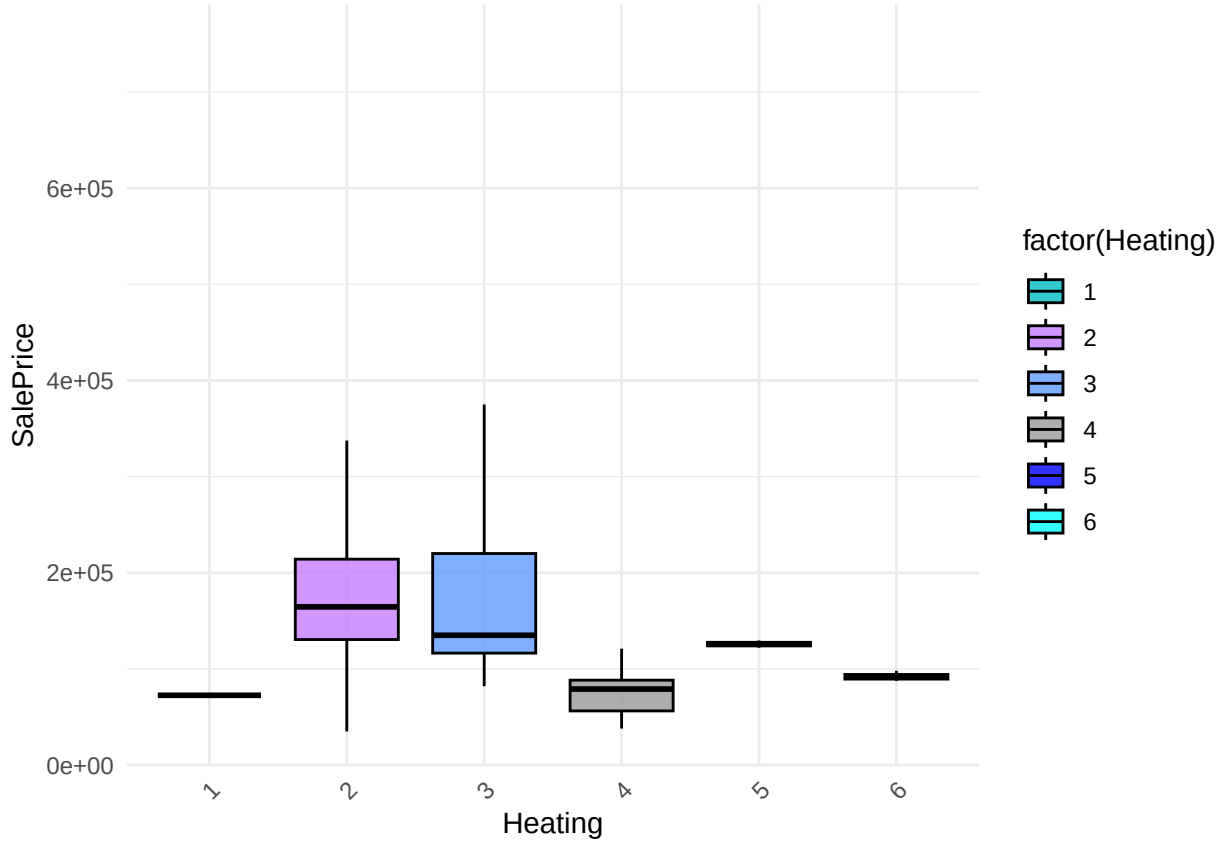
```
cor(house$TotalBsmtSF, house$SalePrice)
```

```
## [1] 0.6135806
```

A very high postiive correlation with the output is clearly visible along with a very small p value

36. Heating

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(Heating), y = SalePrice, fill = factor(Heating))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Heating", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Two categories majorly have prices in the same range and the rest have few values

```
kruskal.test(house$Heating ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Heating by house$SalePrice
## Kruskal-Wallis chi-squared = 659.15, df = 662, p-value = 0.5239
```

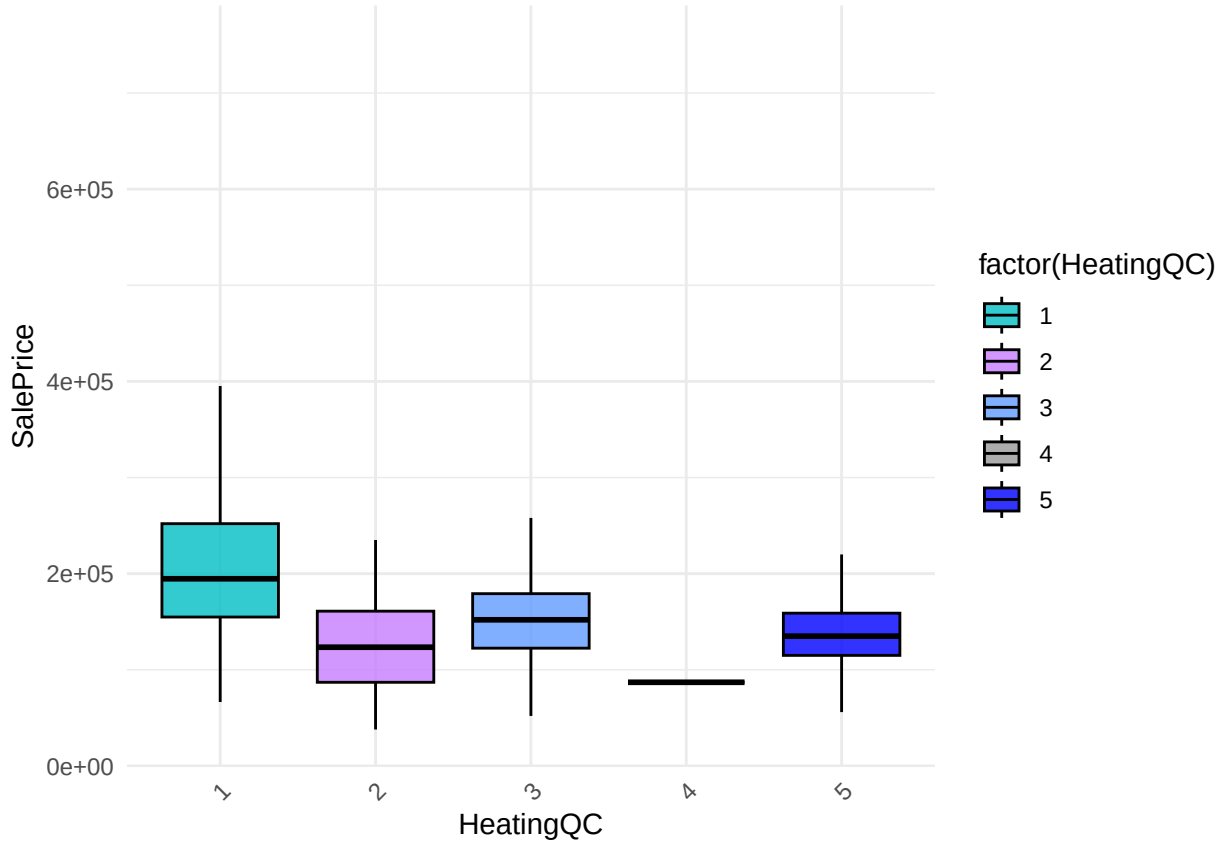
```
summary(aov(Heating~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1   1.24  1.2407   14.38 0.000156 ***
## Residuals   1458 125.84  0.0863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly the first test suggests that there is similarity based on medians which is clearly visible and the means show a change because of the posiiblity that these just may be outliers

37. HeatingQC

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(HeatingQC), y = SalePrice, fill = factor(HeatingQC))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "HeatingQC", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Clearly the excellent range have a higher price

```
kruskal.test(house$HeatingQC ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$HeatingQC by house$SalePrice
## Kruskal-Wallis chi-squared = 766.85, df = 662, p-value = 0.002889
```

```
summary(aov(HeatingQC~SalePrice,data=house))
```

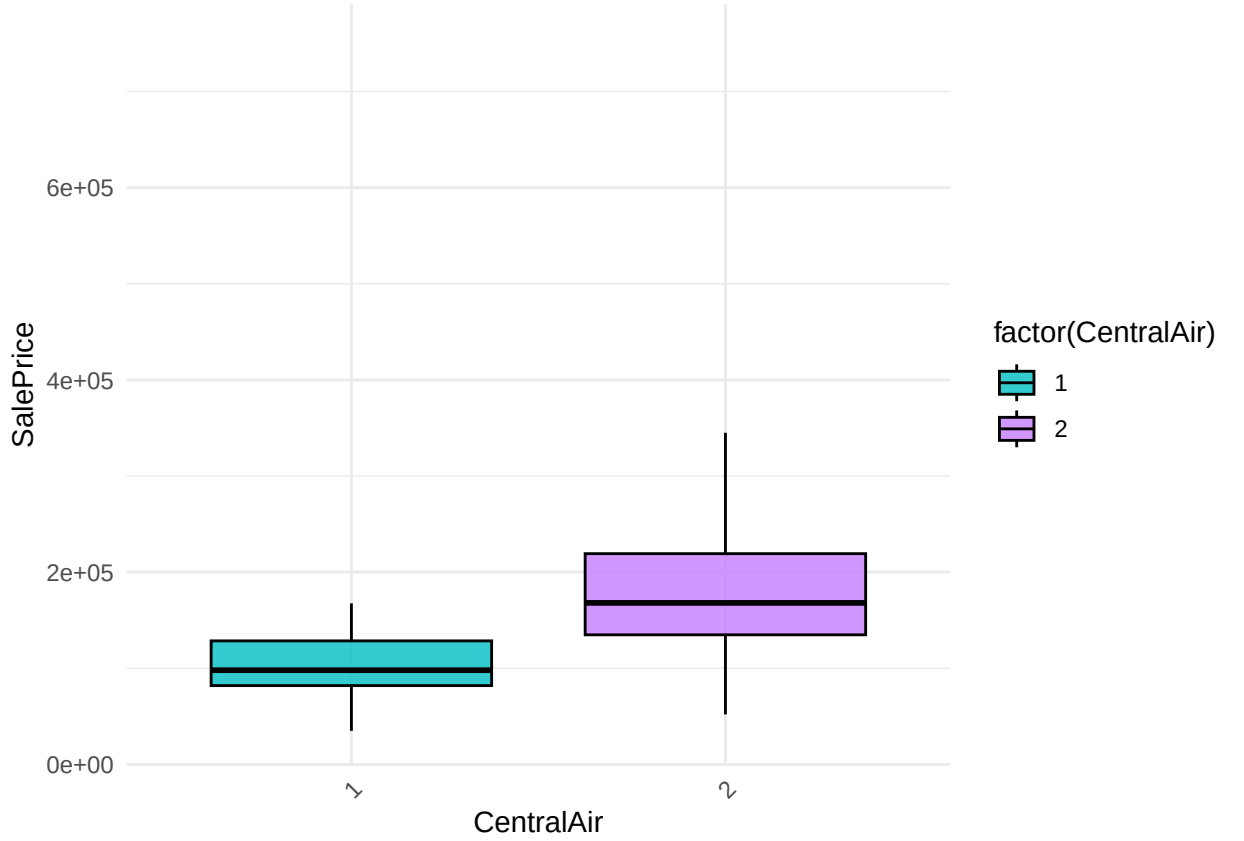
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1    707   707.0     278 <2e-16 ***
## Residuals   1458   3708     2.5
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

both small p values indicate that the variable has a strong relation with the target variable

38. CentralAir

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(CentralAir), y = SalePrice, fill = factor(CentralAir))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "CentralAir", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Houses having central air conditioning have a higher rate clearly

```
kruskal.test(house$CentralAir ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$CentralAir by house$SalePrice
## Kruskal-Wallis chi-squared = 826.29, df = 662, p-value = 1.301e-05
```

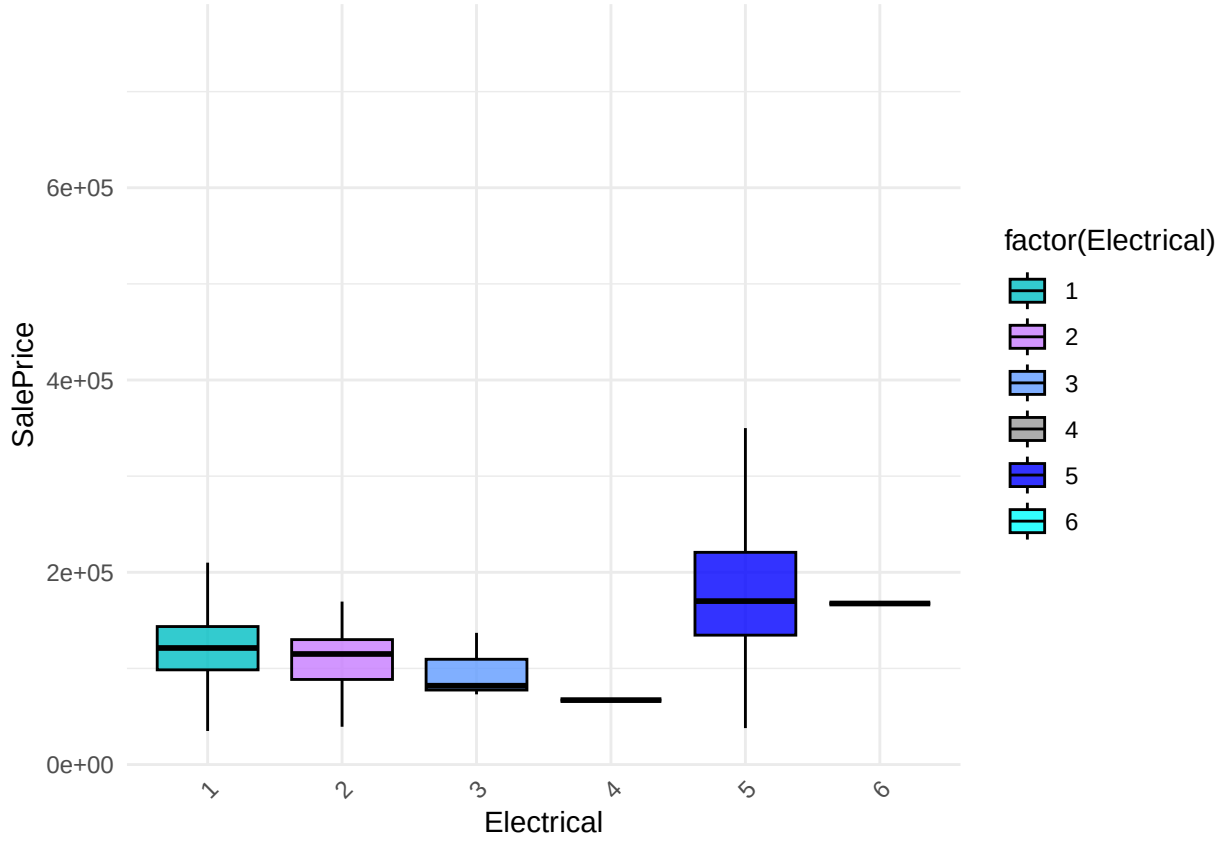
```
summary(aov(CentralAir~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   5.61   5.610   98.31 <2e-16 ***
## Residuals   1458  83.21   0.057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small p values also indicate its strong relation with the target variable

39. Electrical

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(Electrical), y = SalePrice, fill = factor(Electrical))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Electrical", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Clearly the mix type has higher price range, rest of them follow a decreasing trend due to decrease in quality

```
kruskal.test(house$Electrical ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Electrical by house$SalePrice
## Kruskal-Wallis chi-squared = 721.71, df = 662, p-value = 0.05345
```

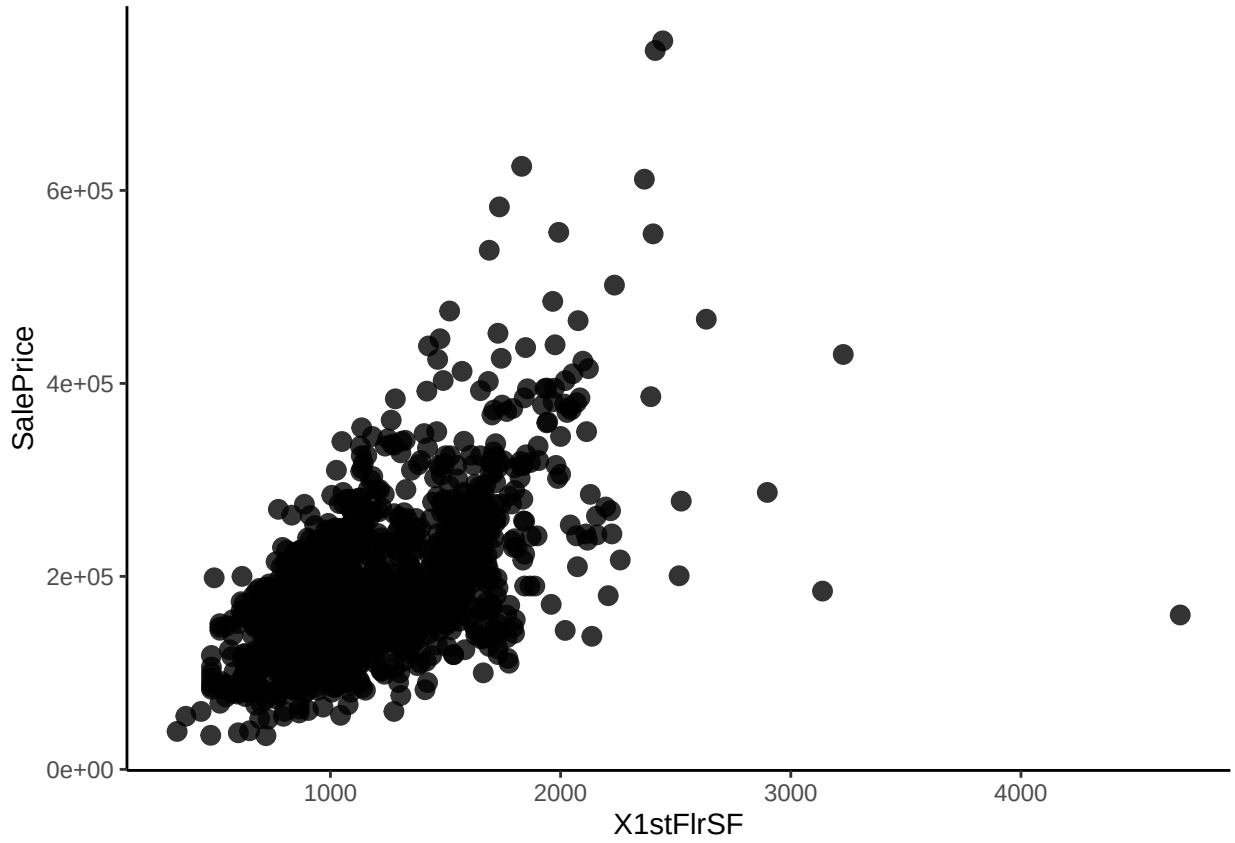
```
summary(aov(Electrical~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   88.9   88.93   85.01 <2e-16 ***
## Residuals   1458 1525.2    1.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p values for mean also suggest the same inference from the graph

40. X1stFlrSF

```
ggplot(house, aes(x = X1stFlrSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "X1stFlrSF", y = "SalePrice") +
  theme_classic()
```



a almost linear increase in the price as the first floor square feet increases

```
kruskal.test(house$X1stFlrSF ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$X1stFlrSF by house$SalePrice
## Kruskal-Wallis chi-squared = 918.61, df = 662, p-value = 1.292e-10
```

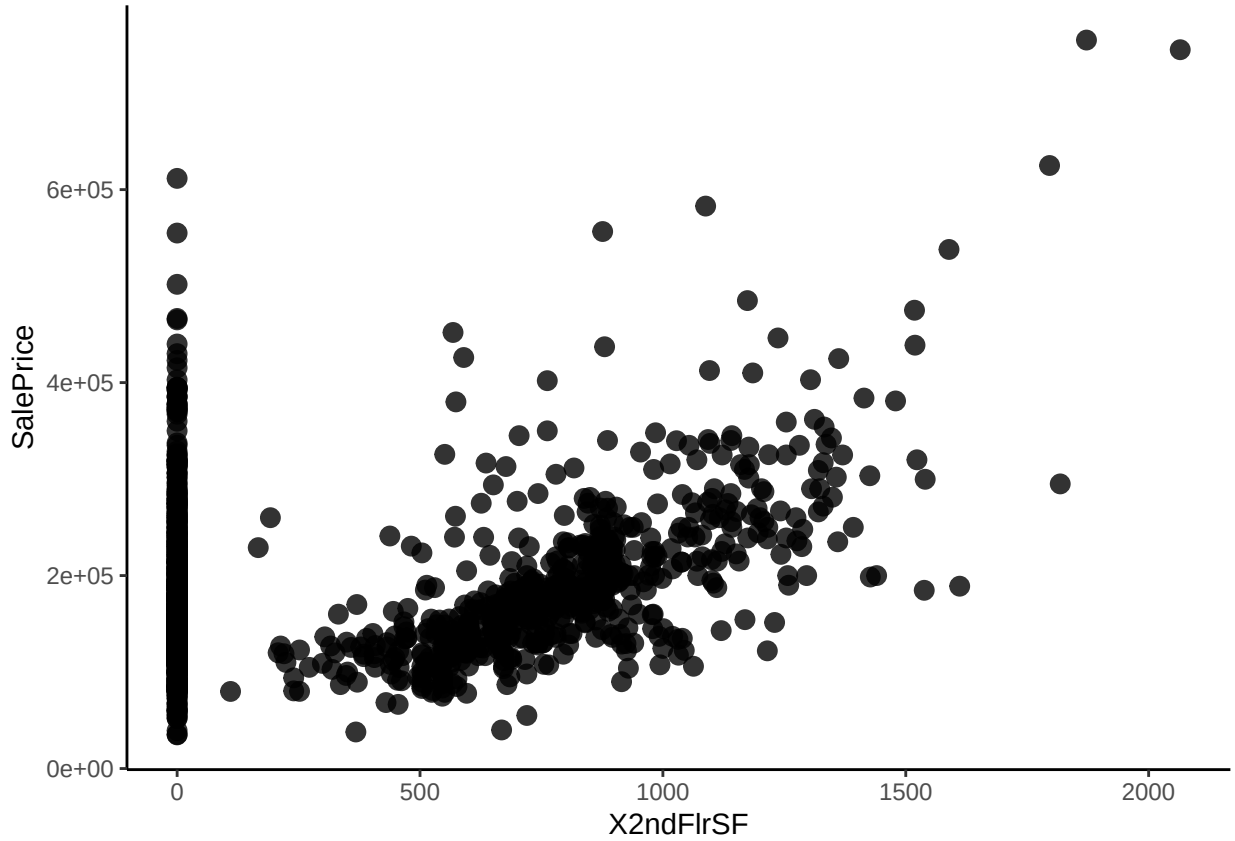
```
cor(house$X1stFlrSF, house$SalePrice)
```

```
## [1] 0.6058522
```

a small p value lesser than confidence level also indicates the same and the positive correlation also points out the same fact

41. X2ndFlrSF

```
ggplot(house, aes(x = X2ndFlrSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "X2ndFlrSF", y = "SalePrice") +
  theme_classic()
```



a almost linear increase in the price as the second floor square feet increases similar to above variable

```
kruskal.test(house$X2ndFlrSF ~ house$SalePrice)
```

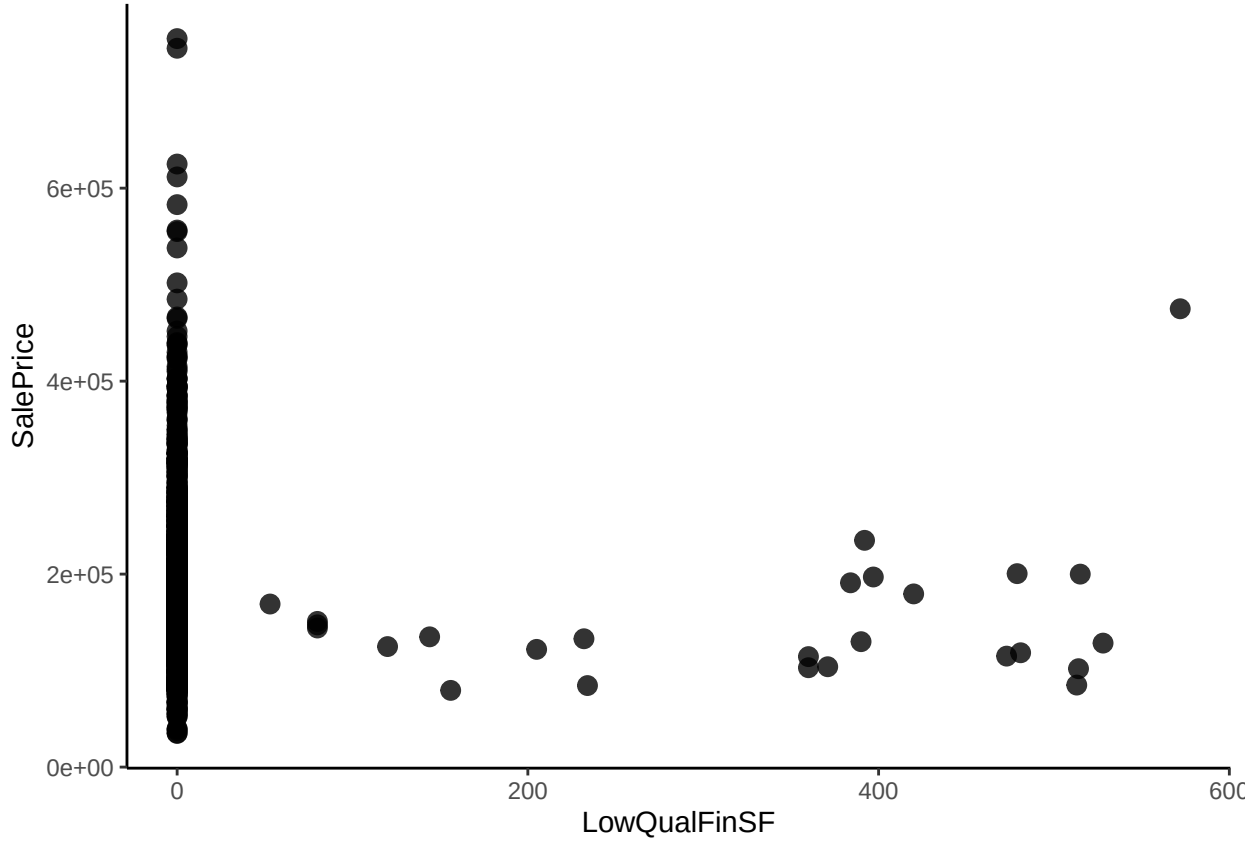
```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$X2ndFlrSF by house$SalePrice
## Kruskal-Wallis chi-squared = 793.81, df = 662, p-value = 0.0003078
```

```
cor(house$X2ndFlrSF, house$SalePrice)
```

```
## [1] 0.3193338
```

42. LowQualFinSF

```
ggplot(house, aes(x = LowQualFinSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "LowQualFinSF", y = "SalePrice") +
  theme_classic()
```



a very scattered chart with probable outliers

```
kruskal.test(house$LowQualFinSF ~ house$SalePrice)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  house$LowQualFinSF by house$SalePrice
## Kruskal-Wallis chi-squared = 649.75, df = 662, p-value = 0.6257
```

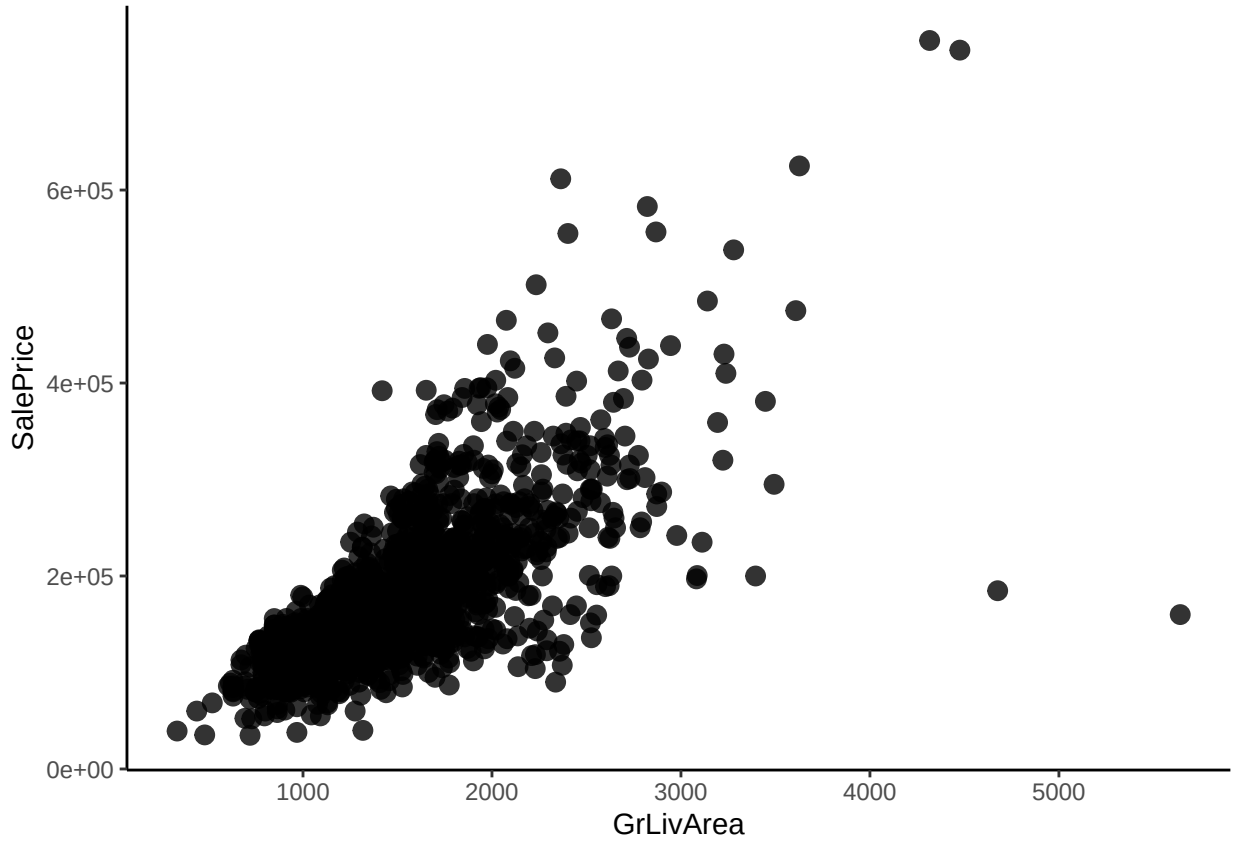
```
cor(house$LowQualFinSF, house$SalePrice)
```

```
## [1] -0.02560613
```

a high p value suggests difficult inference from the variable for the target variable

43. GrLivArea

```
ggplot(house, aes(x = GrLivArea, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "GrLivArea", y = "SalePrice") +
  theme_classic()
```



a almost linear increase in the price as the square feet increases

```
kruskal.test(house$GrLivArea ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GrLivArea by house$SalePrice
## Kruskal-Wallis chi-squared = 1047.6, df = 662, p-value < 2.2e-16
```

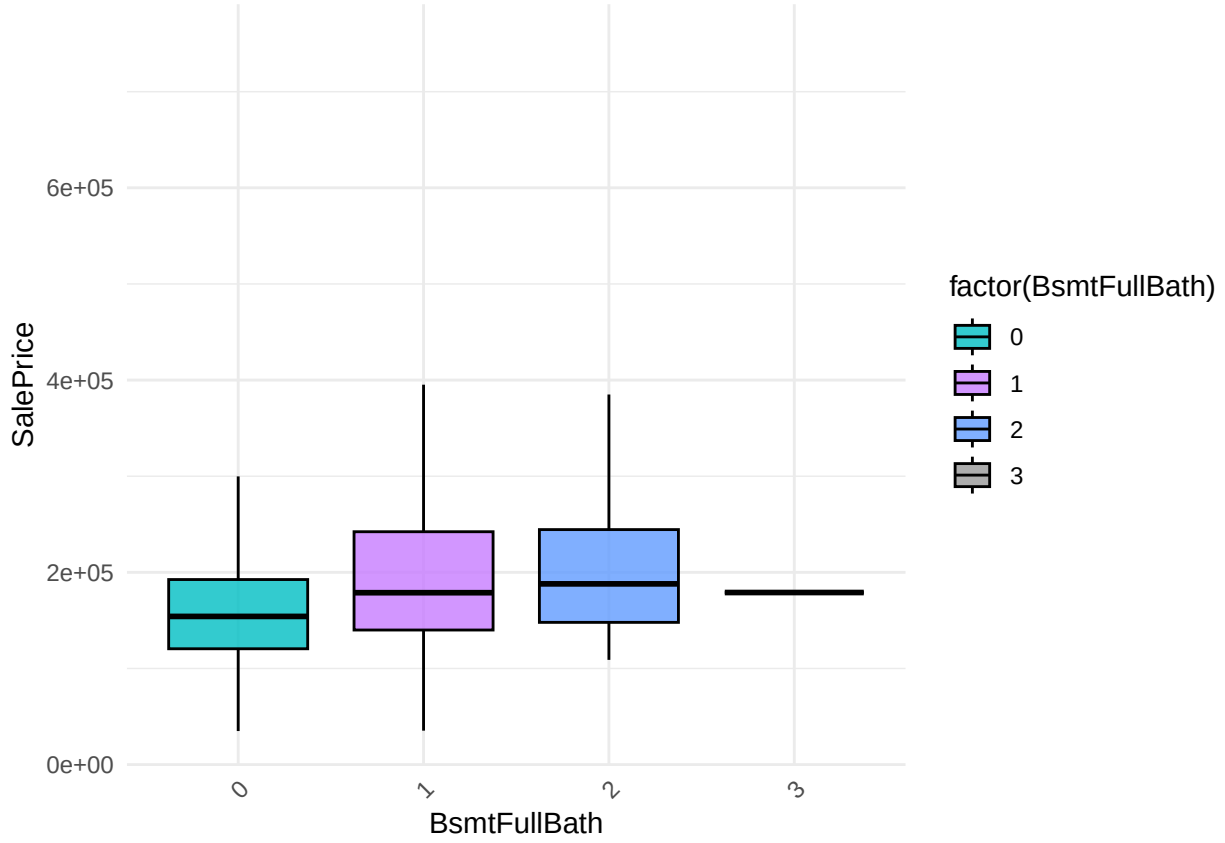
```
cor(house$GrLivArea, house$SalePrice)
```

```
## [1] 0.7086245
```

A very strong linear relationship is visible

43. BsmtFullBath

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtFullBath), y = SalePrice, fill = factor(BsmtFullBath))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtFullBath", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Almost equal distribution

```
kruskal.test(house$BsmtFullBath ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtFullBath by house$SalePrice
## Kruskal-Wallis chi-squared = 699.82, df = 662, p-value = 0.1495
```

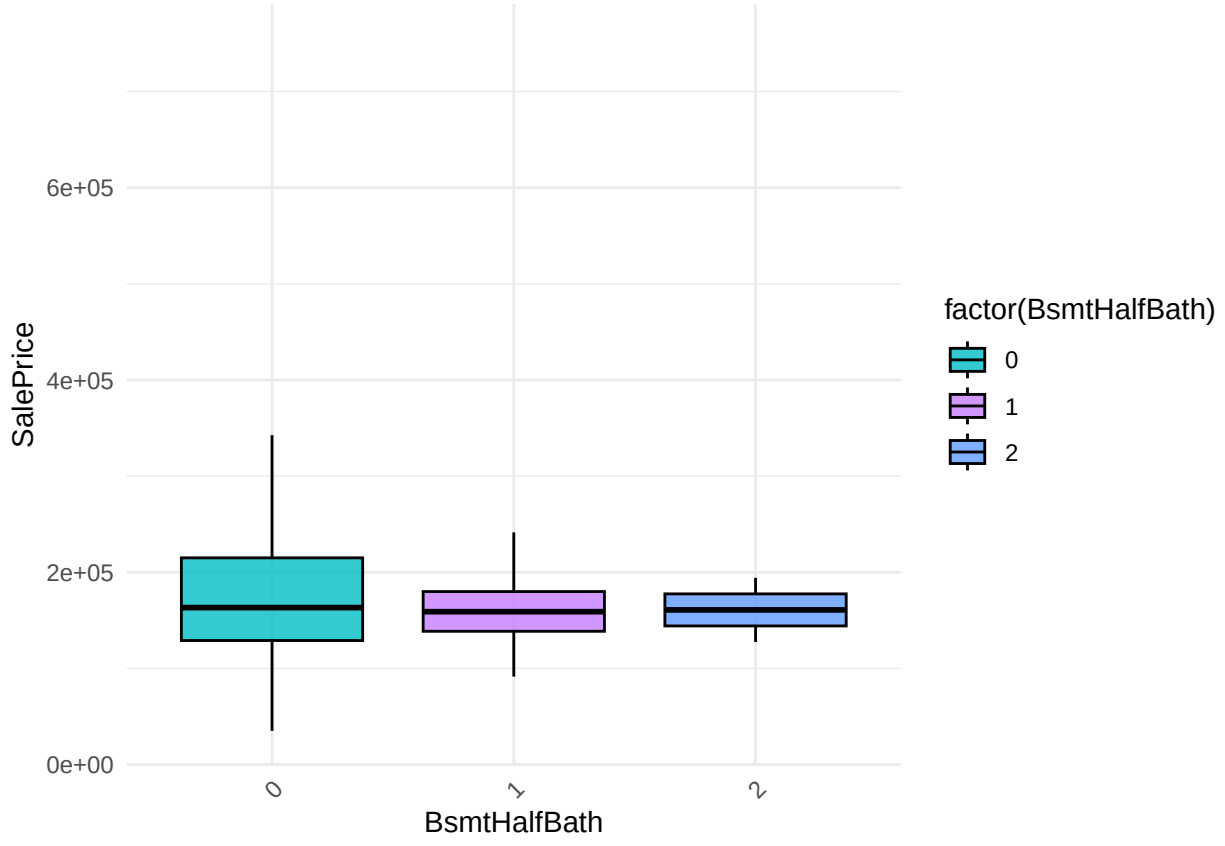
```
summary(aov(BsmtFullBath~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   20.3  20.266    79.3 <2e-16 ***
## Residuals   1458  372.6   0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A high p value suggests less difference between categories

44. BsmtHalfBath

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BsmtHalfBath), y = SalePrice, fill = factor(BsmtHalfBath))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BsmtHalfBath", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Difficult to infer from the graph

```
kruskal.test(house$BsmtHalfBath ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BsmtHalfBath by house$SalePrice
## Kruskal-Wallis chi-squared = 586.71, df = 662, p-value = 0.9836
```

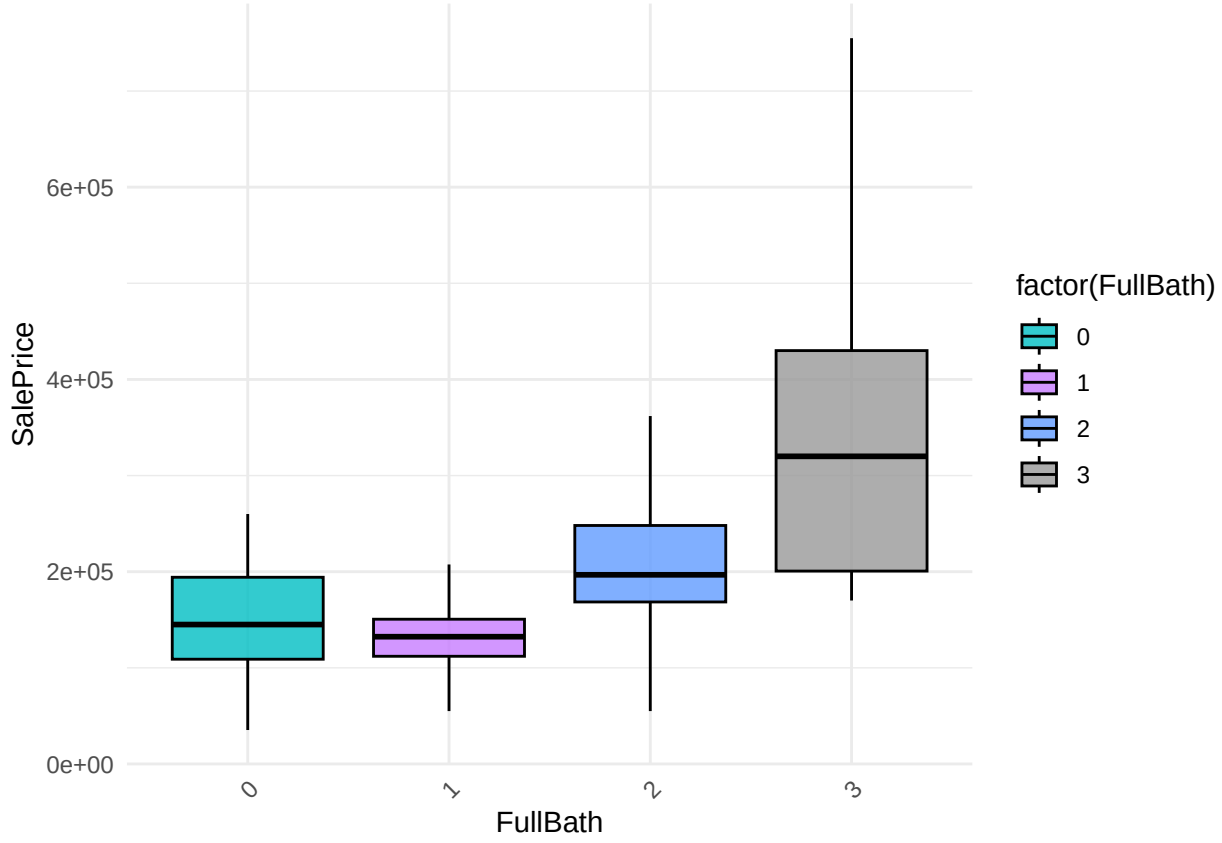
```
summary(aov(BsmtHalfBath~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   0.02 0.02360   0.414   0.52
## Residuals   1458  83.14 0.05703
```

Very less difference in the mean as well as median indicating no relation at all with the target variable

45. FullBath

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(FullBath), y = SalePrice, fill = factor(FullBath))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "FullBath", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A high price for 3rd category

```r
kruskal.test(house$FullBath ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$FullBath by house$SalePrice
## Kruskal-Wallis chi-squared = 996.26, df = 662, p-value = 6.369e-16
```

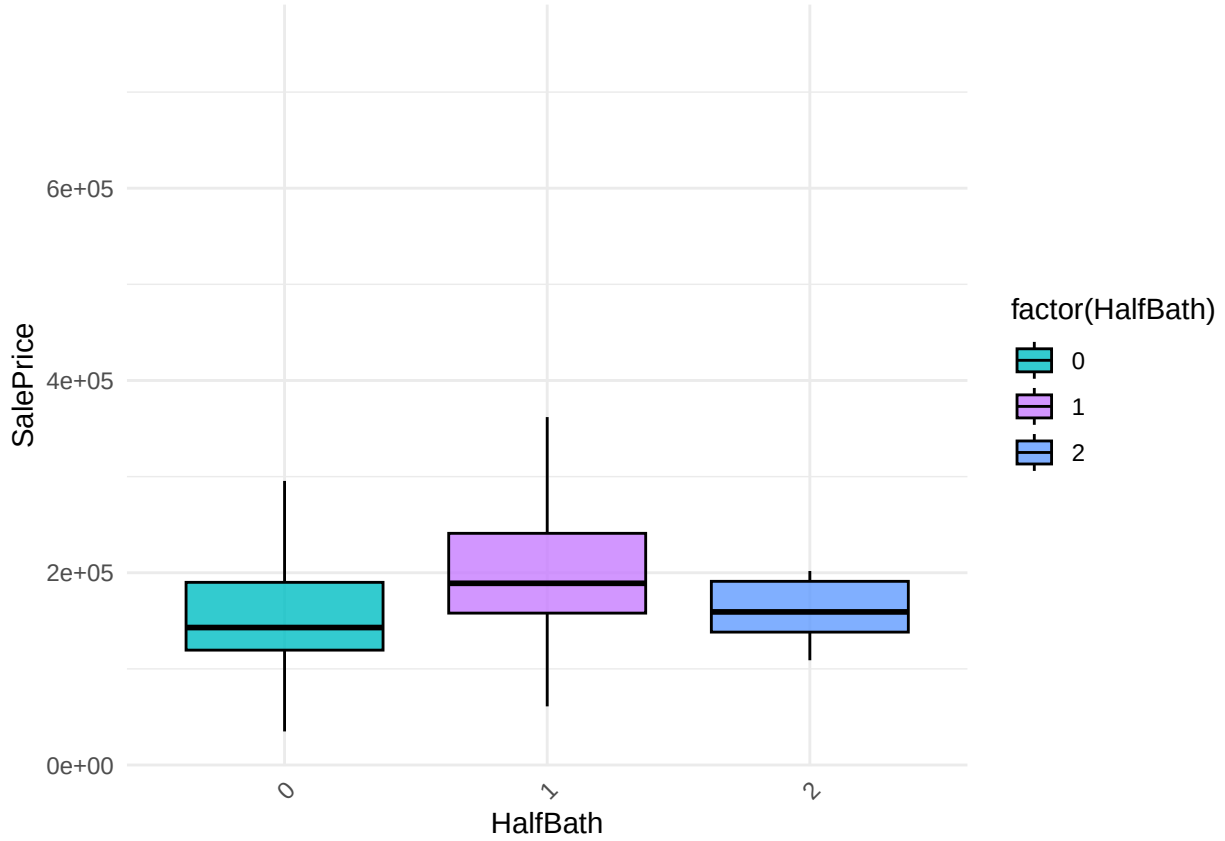
```r
summary(aov(FullBath~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  139.2  139.20   668.4 <2e-16 ***
## Residuals   1458  303.6    0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly small p values suggest that the price changes are present within categories

46. HalfBath

```r
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(HalfBath), y = SalePrice, fill = factor(HalfBath))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "HalfBath", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Less difference between various categories

```
kruskal.test(house$HalfBath ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$HalfBath by house$SalePrice
## Kruskal-Wallis chi-squared = 787.8, df = 662, p-value = 0.0005229
```

```
summary(aov(HalfBath~SalePrice,data=house))
```

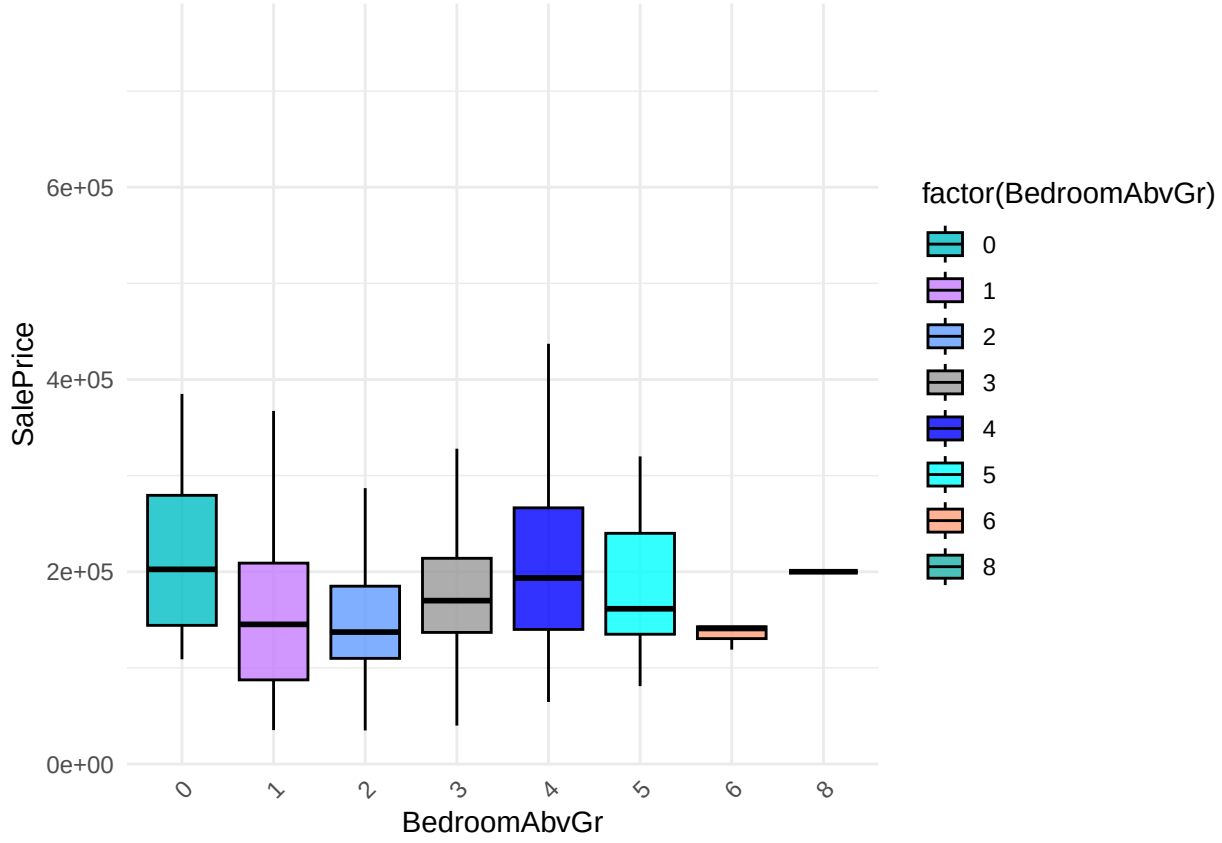
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   29.8  29.782     128 <2e-16 ***
## Residuals   1458  339.2   0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Less difference on the grounds of median but means separate the categories

47. BedroomAbvGr

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(BedroomAbvGr), y = SalePrice, fill = factor(BedroomAbvGr))) +
```

```
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "BedroomAbvGr", y = "SalePrice") +
  theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Difference between various categories is evident

```
kruskal.test(house$BedroomAbvGr ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$BedroomAbvGr by house$SalePrice
## Kruskal-Wallis chi-squared = 749.84, df = 662, p-value = 0.009826
```

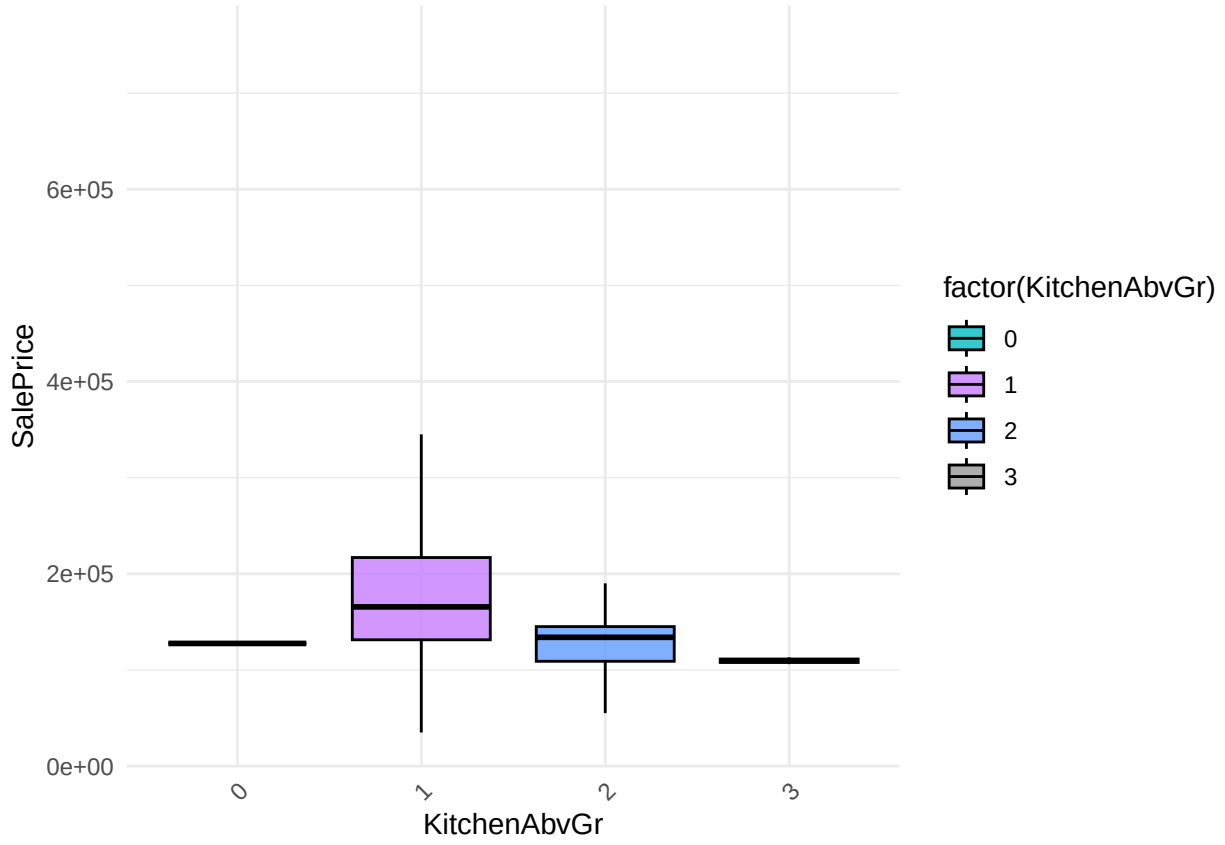
```
summary(aov(BedroomAbvGr~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1   27.5  27.474   42.46 9.93e-11 ***
## Residuals   1458  943.5   0.647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p values also suggest the same

48. KitchenAbvGr

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(KitchenAbvGr), y = SalePrice, fill = factor(KitchenAbvGr))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "KitchenAbvGr", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$KitchenAbvGr ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$KitchenAbvGr by house$SalePrice
## Kruskal-Wallis chi-squared = 639.49, df = 662, p-value = 0.7282
```

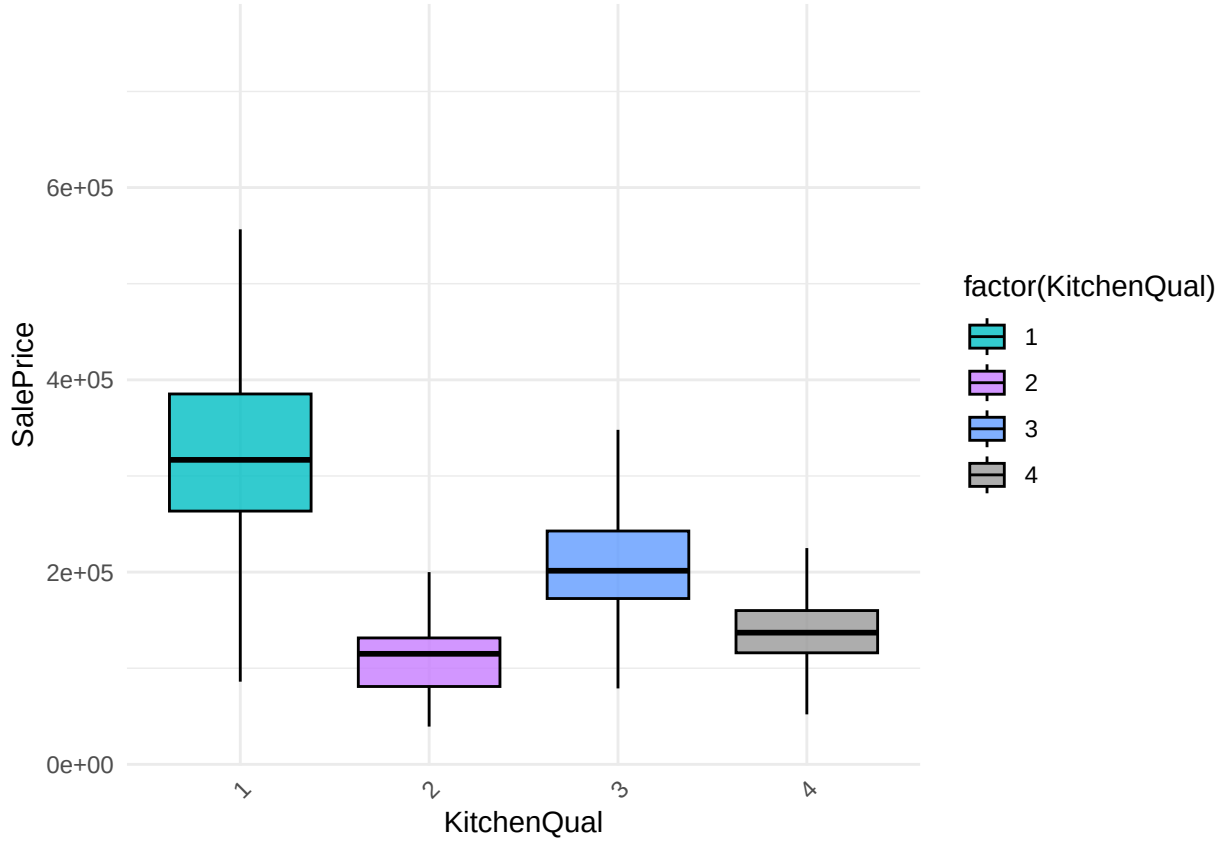
```
summary(aov(KitchenAbvGr~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1   1.31  1.3083   27.44 1.86e-07 ***
## Residuals   1458  69.52  0.0477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small p value for anova test suggest different means within categories

49. KitchenQual

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(KitchenQual), y = SalePrice, fill = factor(KitchenQual))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "KitchenQual", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A clear difference on excellent kitchen qualities

```
kruskal.test(house$KitchenQual ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$KitchenQual by house$SalePrice
## Kruskal-Wallis chi-squared = 957.94, df = 662, p-value = 3.452e-13
```

```
summary(aov(KitchenQual~SalePrice,data=house))
```

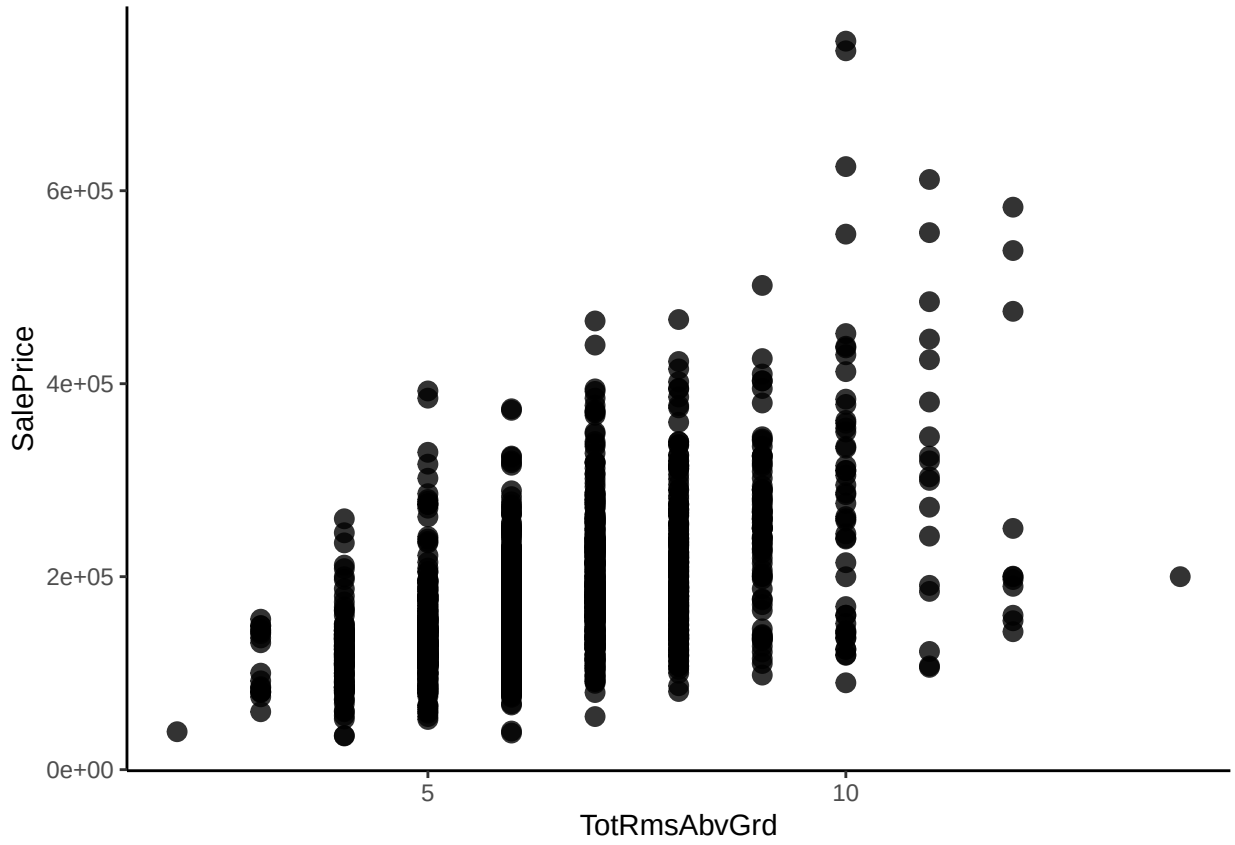
```
##               Df Sum Sq Mean Sq F value Pr(>F)
```

```
## SalePrice      1  349.1   349.1   775.3 <2e-16 ***
## Residuals   1458  656.4     0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A very strong relation with the target variable is indicative

50.TotRmsAbvGrd

```
library(ggplot2)
ggplot(house, aes(x = TotRmsAbvGrd, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "TotRmsAbvGrd", y = "SalePrice") +
  theme_classic()
```



a increase in price as rooms increase

```
kruskal.test(house$TotRmsAbvGrd ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$TotRmsAbvGrd by house$SalePrice
## Kruskal-Wallis chi-squared = 854.04, df = 662, p-value = 5.95e-07
```

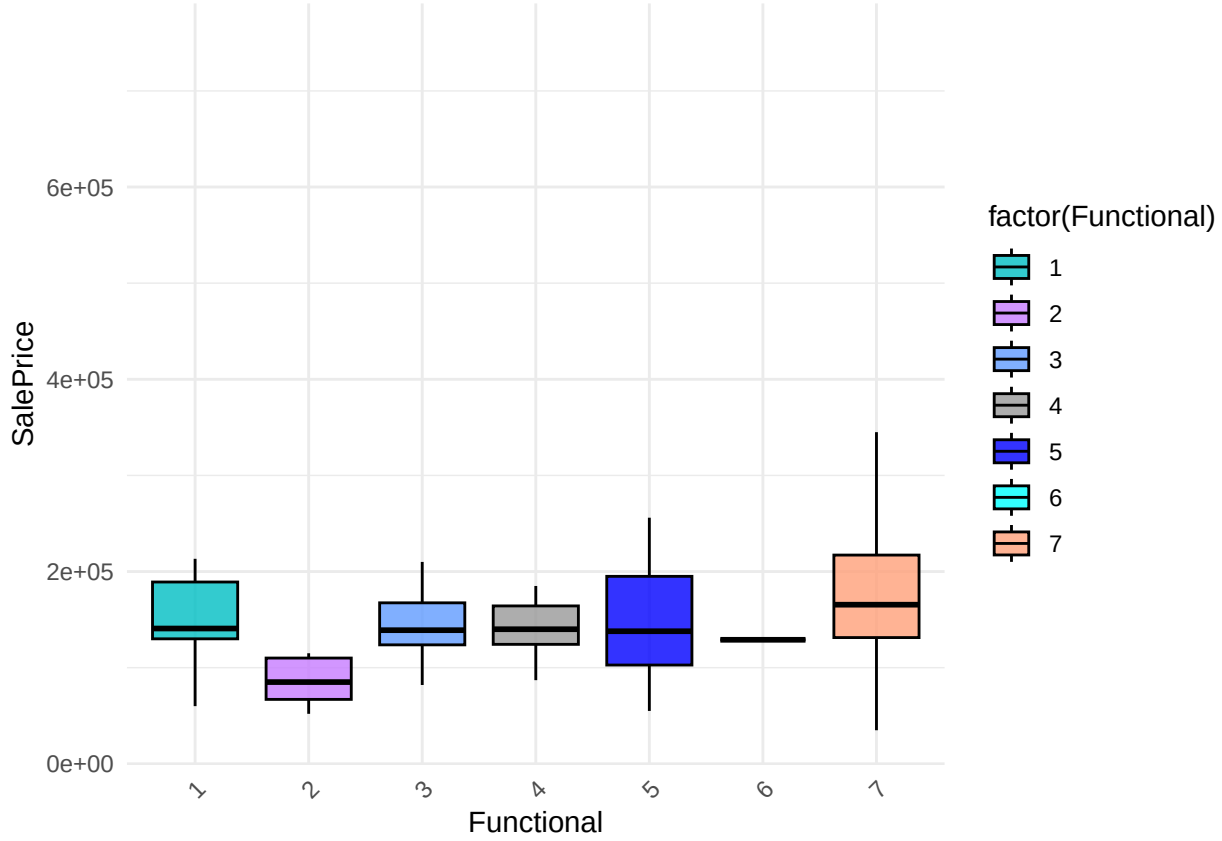
```
cor(house$TotRmsAbvGrd, house$SalePrice)
```

```
## [1] 0.5337232
```

Very strong relation with the target variable

51. Functional

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(Functional), y = SalePrice, fill = factor(Functional))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Functional", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$Functional ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Functional by house$SalePrice
## Kruskal-Wallis chi-squared = 581.34, df = 662, p-value = 0.9891
```

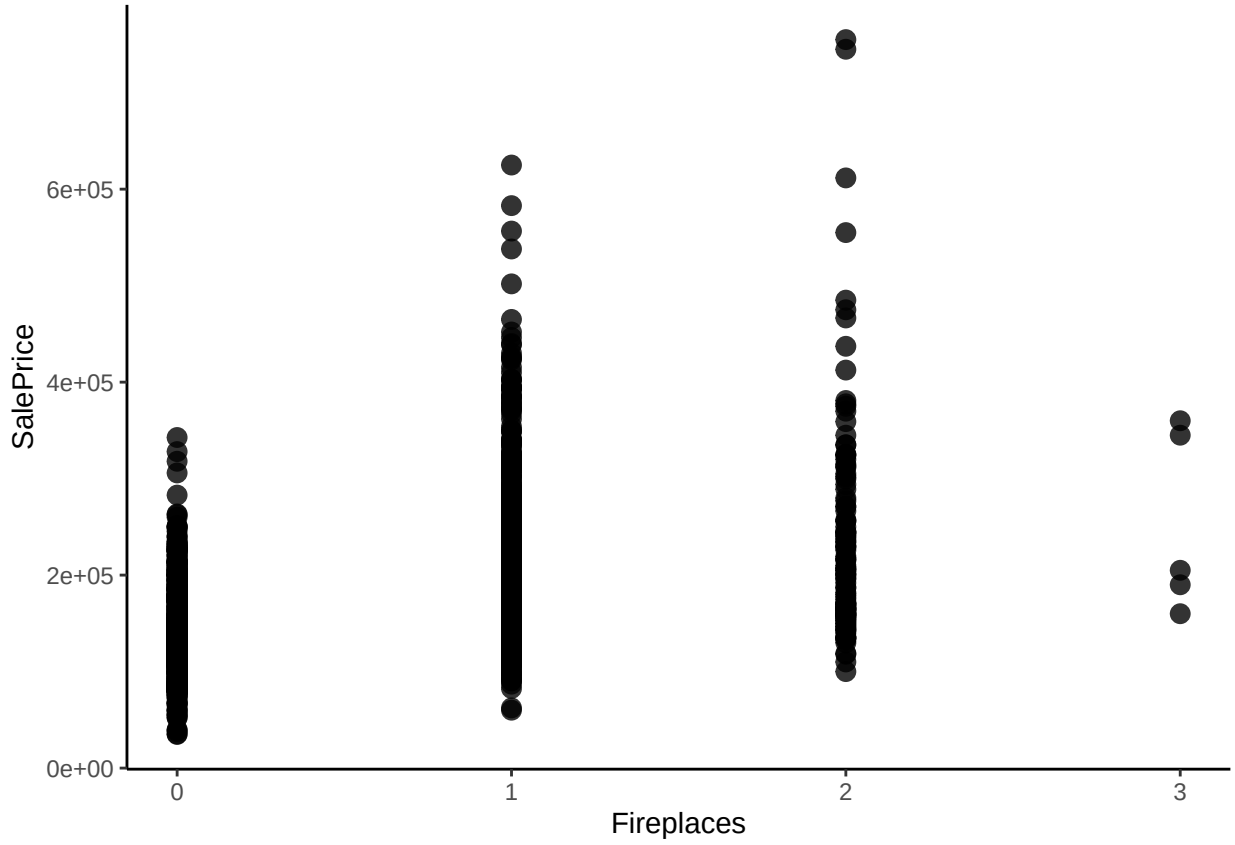
```
summary(aov(Functional~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1   18.6  18.624   19.65 9.98e-06 ***
## Residuals   1458 1381.6   0.948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A significant difference with respect to mean value

52. Fireplaces

```
library(ggplot2)
ggplot(house, aes(x = Fireplaces, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "Fireplaces", y = "SalePrice") +
  theme_classic()
```



1 to 2 fireplaces have higher rates

```
kruskal.test(house$Fireplaces ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Fireplaces by house$SalePrice
## Kruskal-Wallis chi-squared = 848.1, df = 662, p-value = 1.184e-06
```

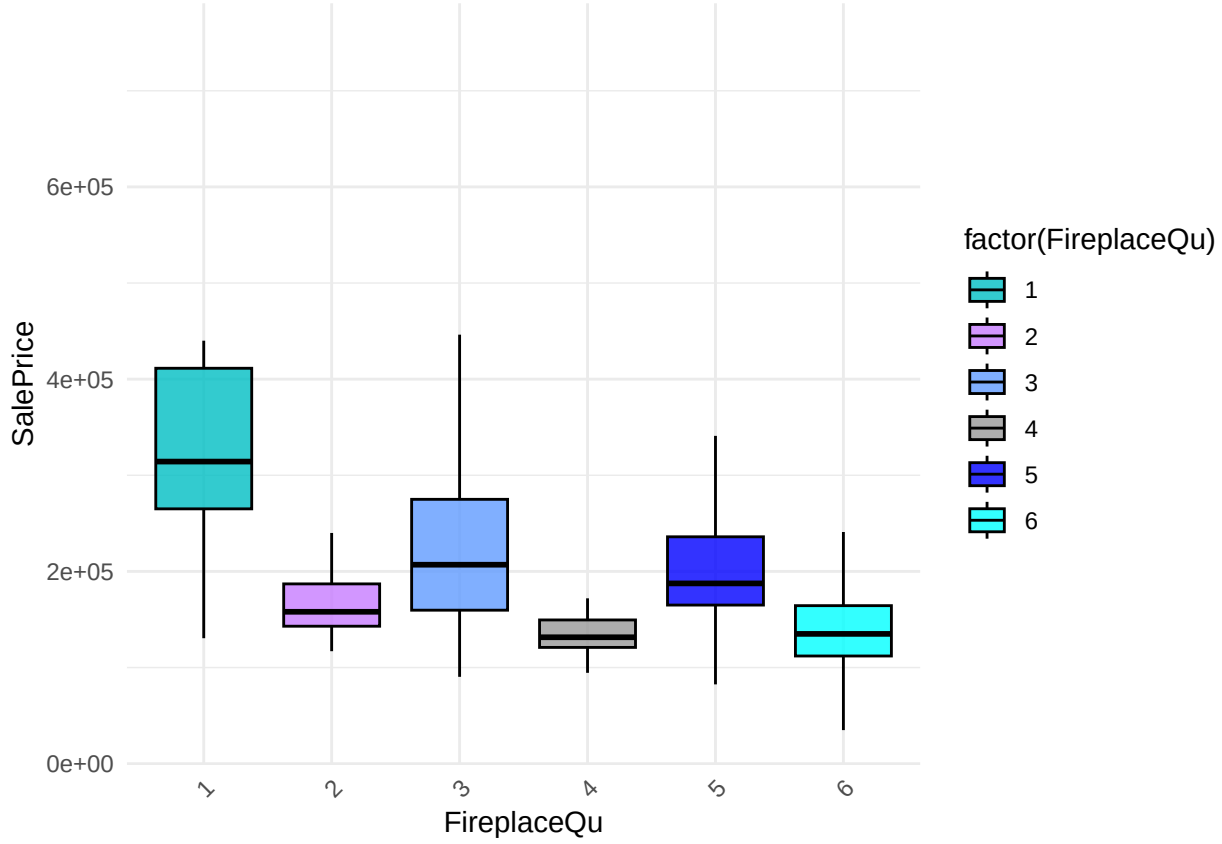
```
cor(house$Fireplaces, house$SalePrice)
```

```
## [1] 0.4669288
```

strong positive correlation and a small p value indicate the same

53. FireplaceQu

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(FireplaceQu), y = SalePrice, fill = factor(FireplaceQu))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "FireplaceQu", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$FireplaceQu ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$FireplaceQu by house$SalePrice
## Kruskal-Wallis chi-squared = 847.12, df = 662, p-value = 1.326e-06
```

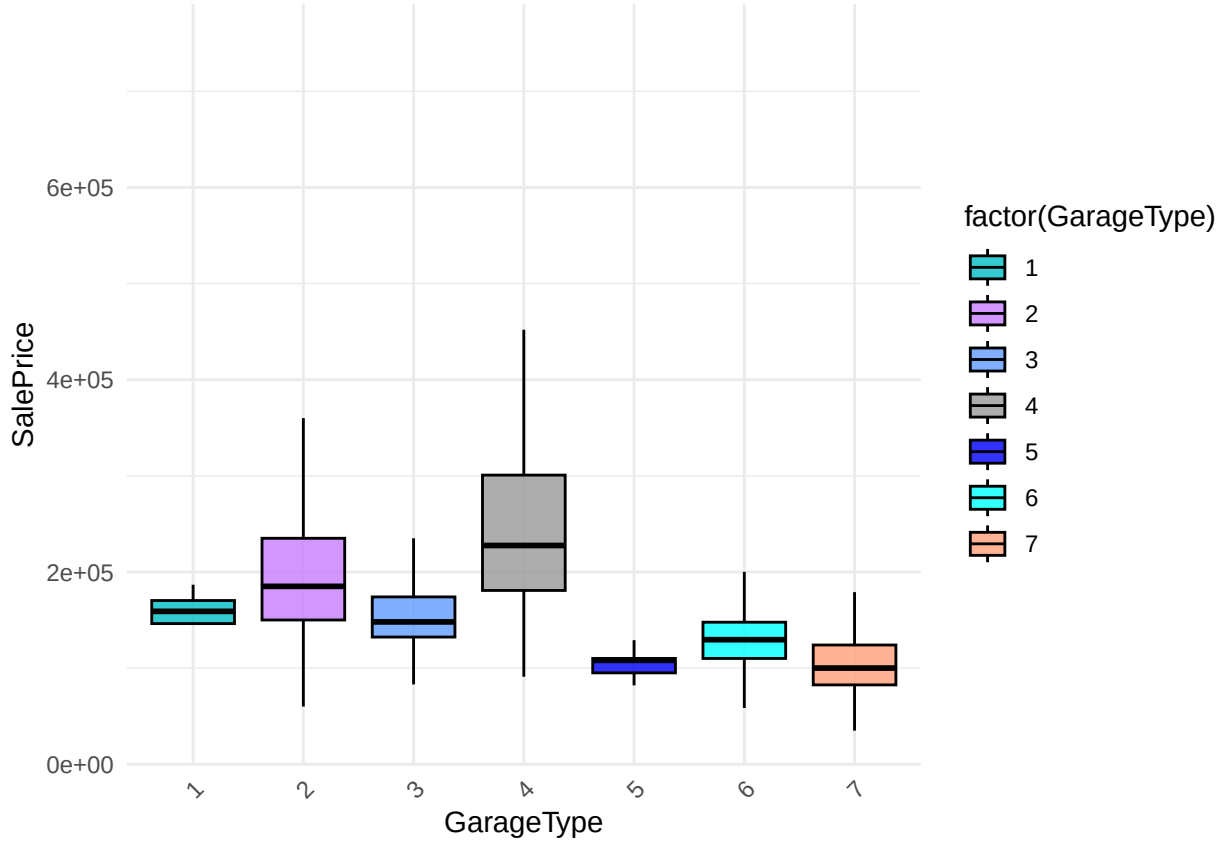
```
summary(aov(FireplaceQu~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  603.2   603.2   390.5 <2e-16 ***
## Residuals   1458 2252.2     1.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A very strong relation both from the graph as well as the p values

54. GarageType

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(GarageType), y = SalePrice, fill = factor(GarageType))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "GarageType", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$GarageType ~ house$SalePrice)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageType by house$SalePrice
## Kruskal-Wallis chi-squared = 831.9, df = 662, p-value = 7.171e-06
```

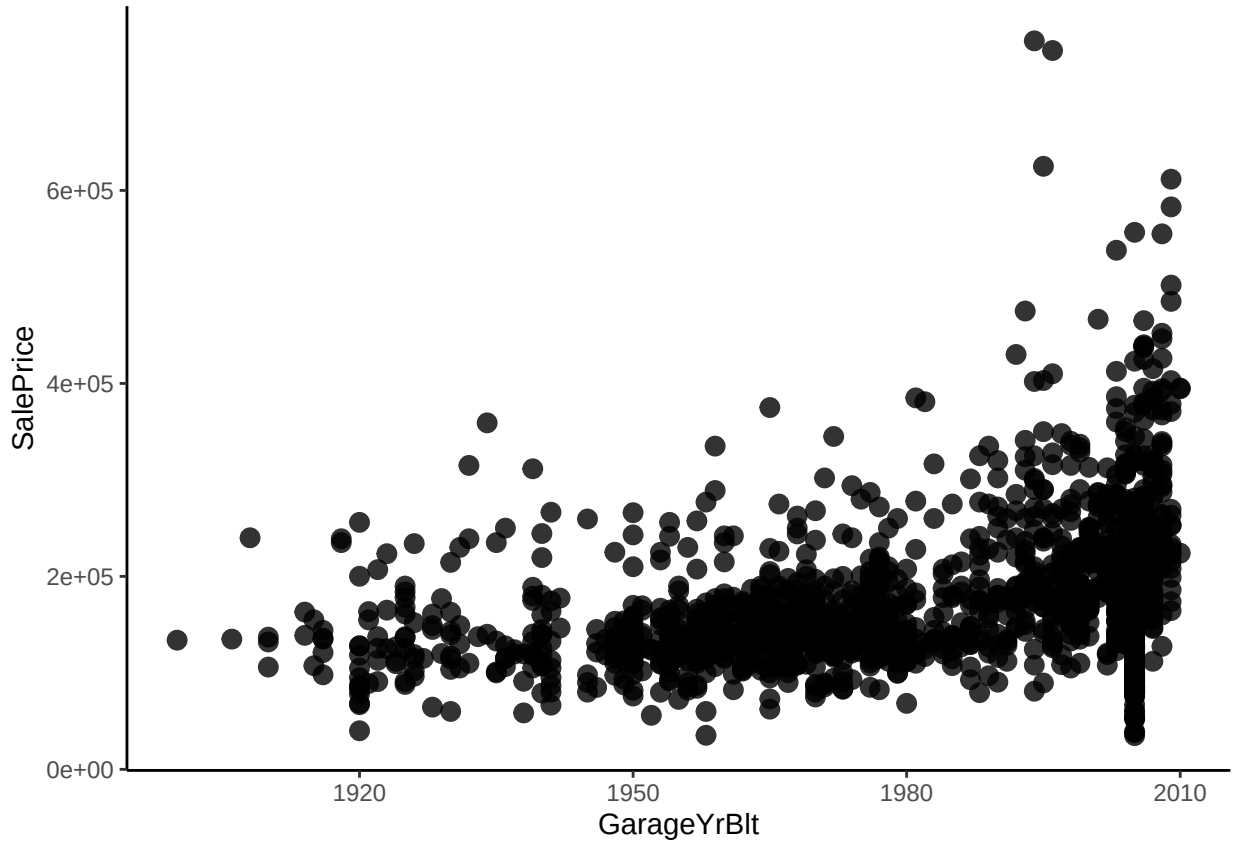
```
summary(aov(GarageType~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1    940   940.4   303.8 <2e-16 ***
## Residuals   1458   4512     3.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again a strong variation with the target variable

55. GarageYrBlt

```
library(ggplot2)
ggplot(house, aes(x = GarageYrBlt, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "GarageYrBlt", y = "SalePrice") +
  theme_classic()
```



1 Rates increase gradually as the year goes ahead

```
kruskal.test(house$GarageYrBlt ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageYrBlt by house$SalePrice
## Kruskal-Wallis chi-squared = 880.72, df = 662, p-value = 2.237e-08
```

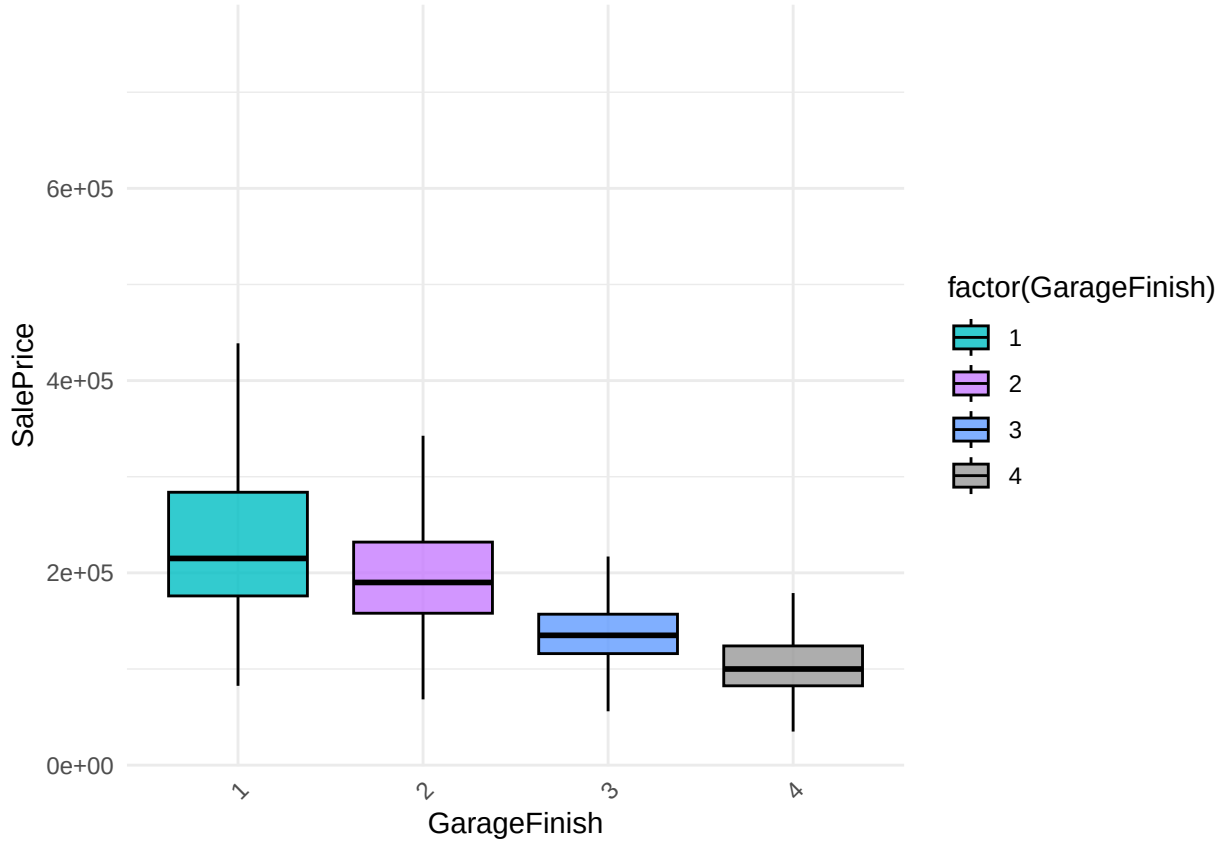
```
cor(house$GarageYrBlt, house$SalePrice)
```

```
## [1] 0.397778
```

A positive correlation also suggests the same

56. GarageFinish

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(GarageFinish), y = SalePrice, fill = factor(GarageFinish))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "GarageFinish", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A steady decrease as the quality goes from top to no garage at all in the prices

```
kruskal.test(house$GarageFinish ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageFinish by house$SalePrice
## Kruskal-Wallis chi-squared = 965.92, df = 662, p-value = 9.686e-14
```

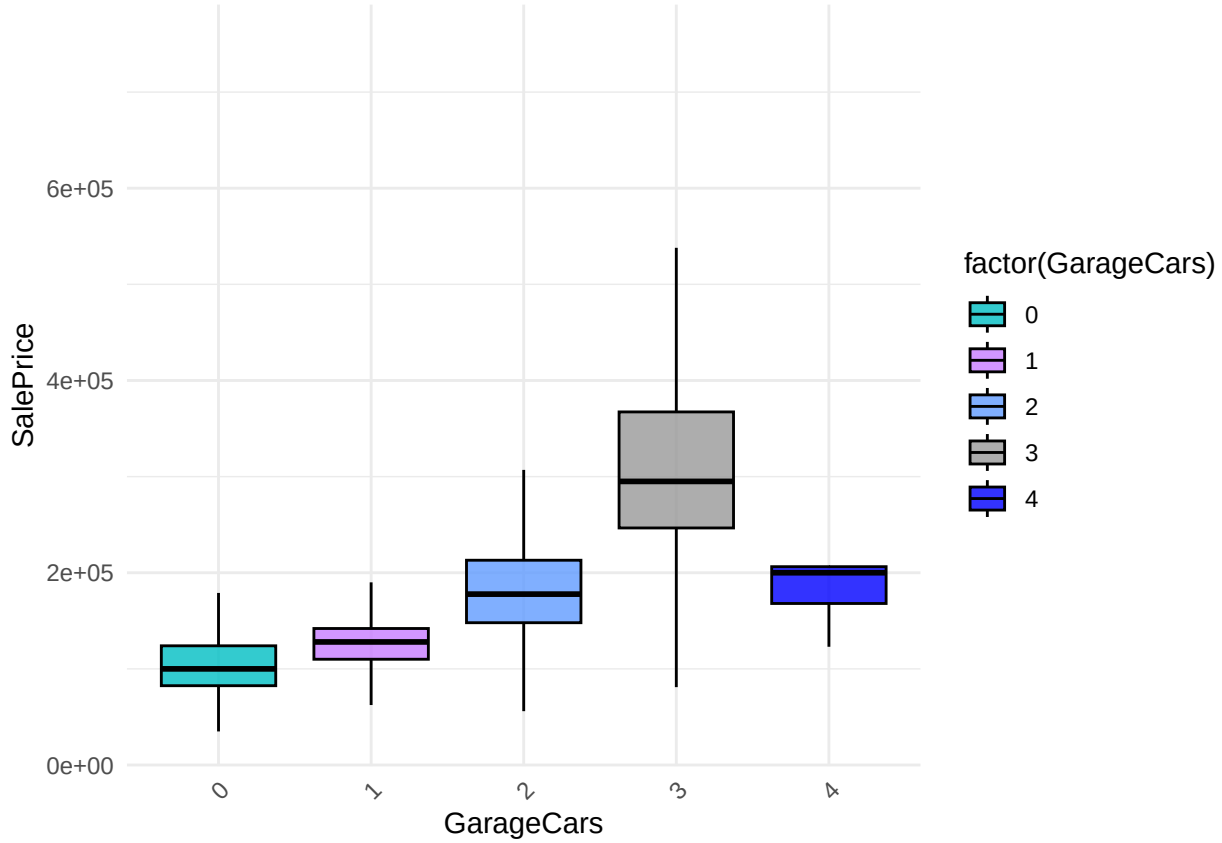
```
summary(aov(GarageFinish~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  350.9   350.9   629.8 <2e-16 ***
## Residuals   1458  812.2     0.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Same indication from the p values also

57. GarageCars

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(GarageCars), y = SalePrice, fill = factor(GarageCars))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "GarageCars", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



garage with capacity of 3 cars have the highest rates

```
kruskal.test(house$GarageCars ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageCars by house$SalePrice
## Kruskal-Wallis chi-squared = 1039, df = 662, p-value < 2.2e-16
```

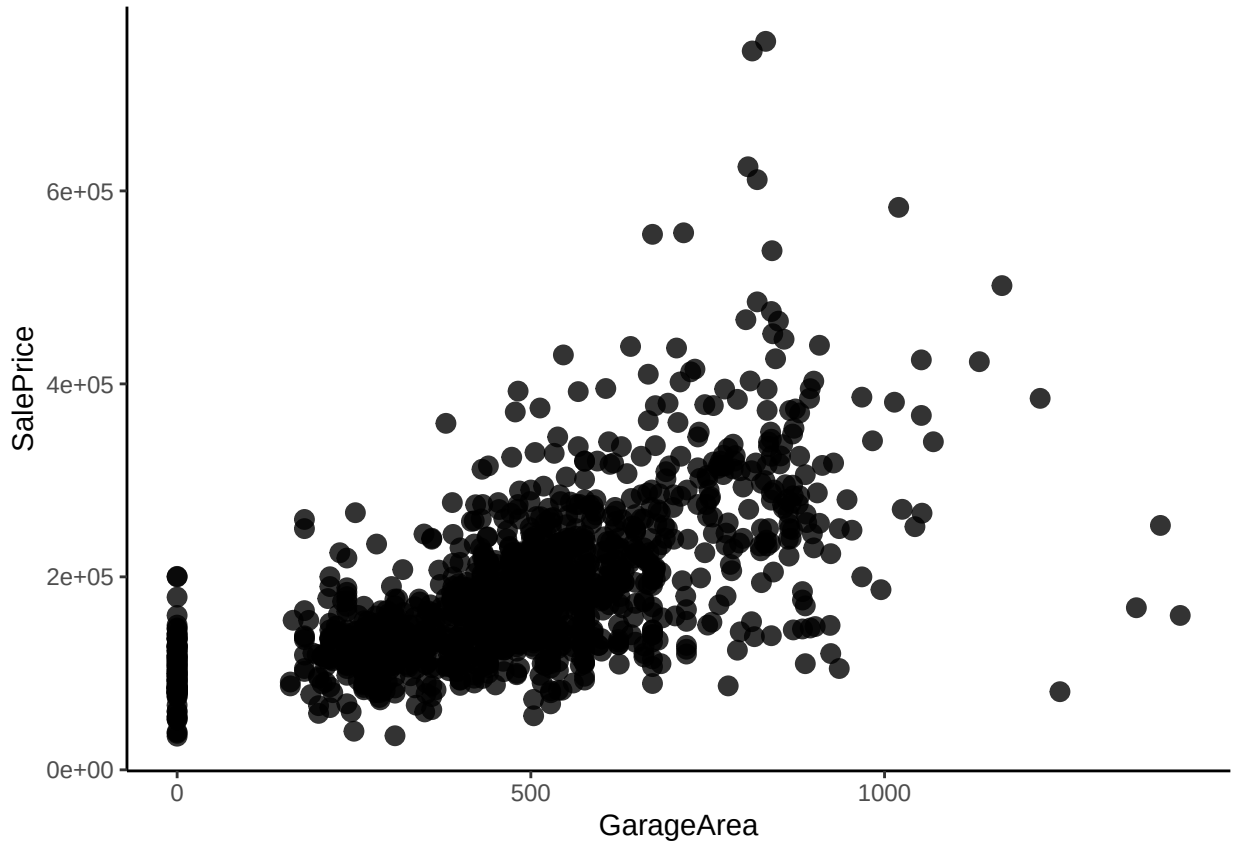
```
summary(aov(GarageCars~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1  334.2   334.2    1014 <2e-16 ***
## Residuals   1458  480.6     0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small p values suggest the similar analysis for a strong relation between the two variables

58. GarageArea

```
library(ggplot2)
ggplot(house, aes(x = GarageArea, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "GarageArea", y = "SalePrice") +
  theme_classic()
```



Rates increase gradually as the area increases

```
kruskal.test(house$GarageArea ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageArea by house$SalePrice
## Kruskal-Wallis chi-squared = 979.39, df = 662, p-value = 1.081e-14
```

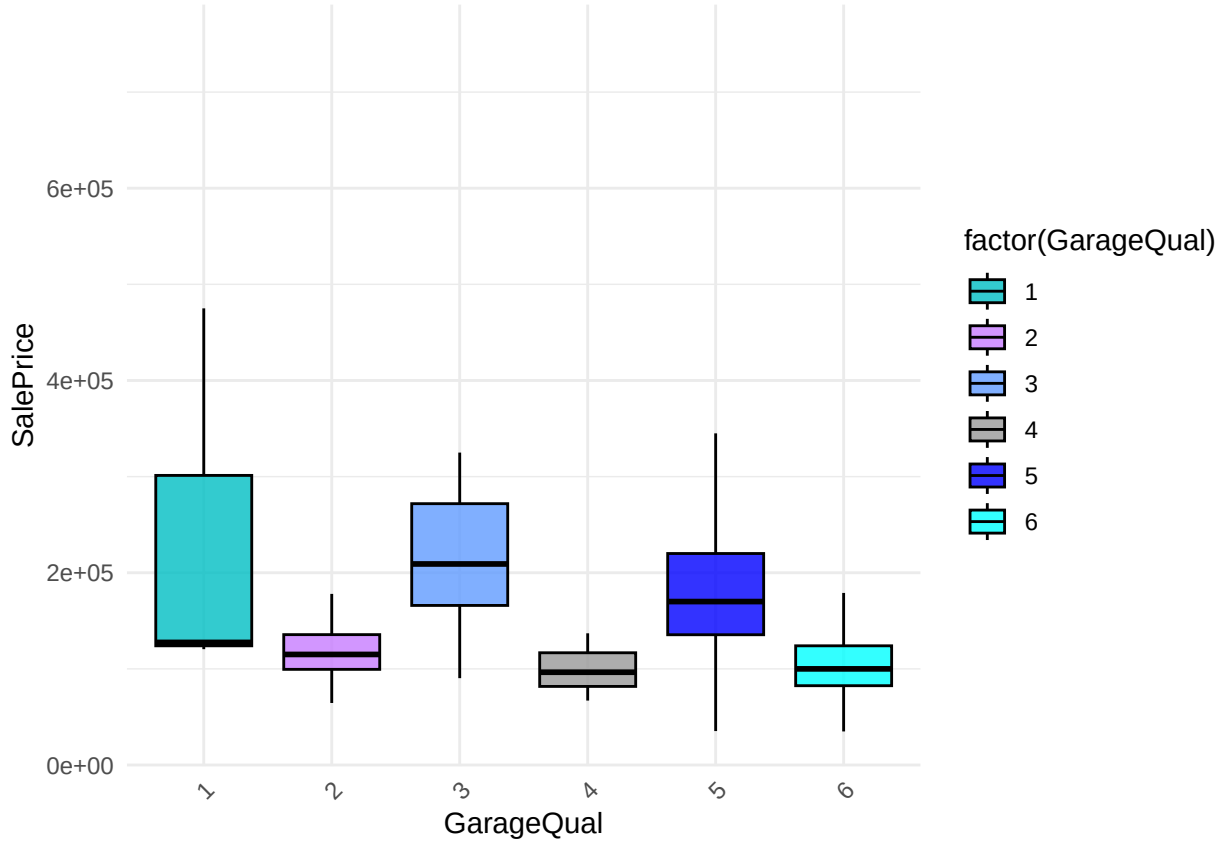
```
cor(house$GarageArea, house$SalePrice)
```

```
## [1] 0.6234314
```

A strong positive correlation also gives the same insight

59. GarageQual

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(GarageQual), y = SalePrice, fill = factor(GarageQual))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "GarageQual", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Excellent quality garages have a high price for the houses compared to the other qualities

```
kruskal.test(house$GarageQual ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageQual by house$SalePrice
## Kruskal-Wallis chi-squared = 636.85, df = 662, p-value = 0.7523
```

```
summary(aov(GarageQual~SalePrice,data=house))
```

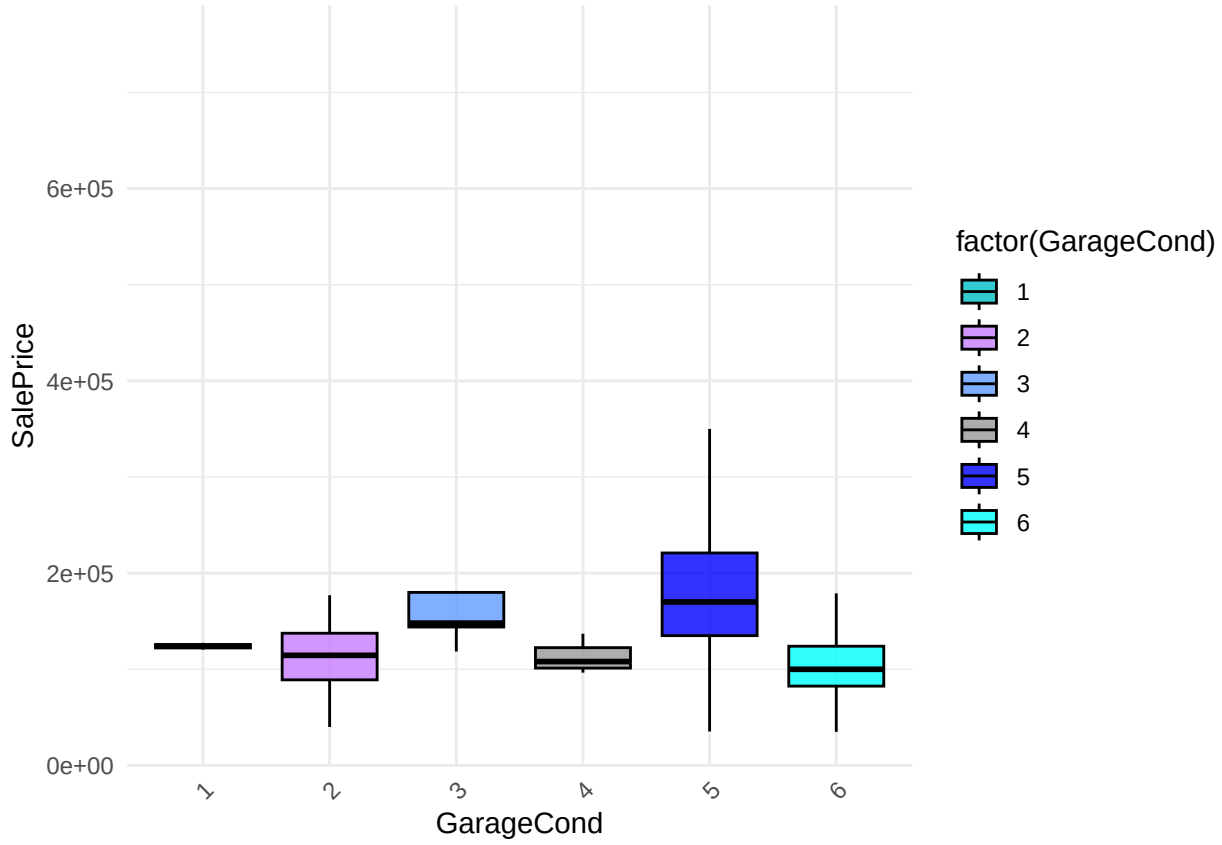
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1    0.0  0.0288   0.069  0.793
## Residuals   1458  612.3  0.4199
```

P values suggest that the variable doesnt hold a strong relation and we have evidence to reject the null hypothesis

60. GarageCond

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(GarageCond), y = SalePrice, fill = factor(GarageCond))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
```

```
  scale_fill_manual(values = my_colors) +
  labs(x = "GarageCond", y = "SalePrice") +
  theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$GarageCond ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$GarageCond by house$SalePrice
## Kruskal-Wallis chi-squared = 620.07, df = 662, p-value = 0.8768
```

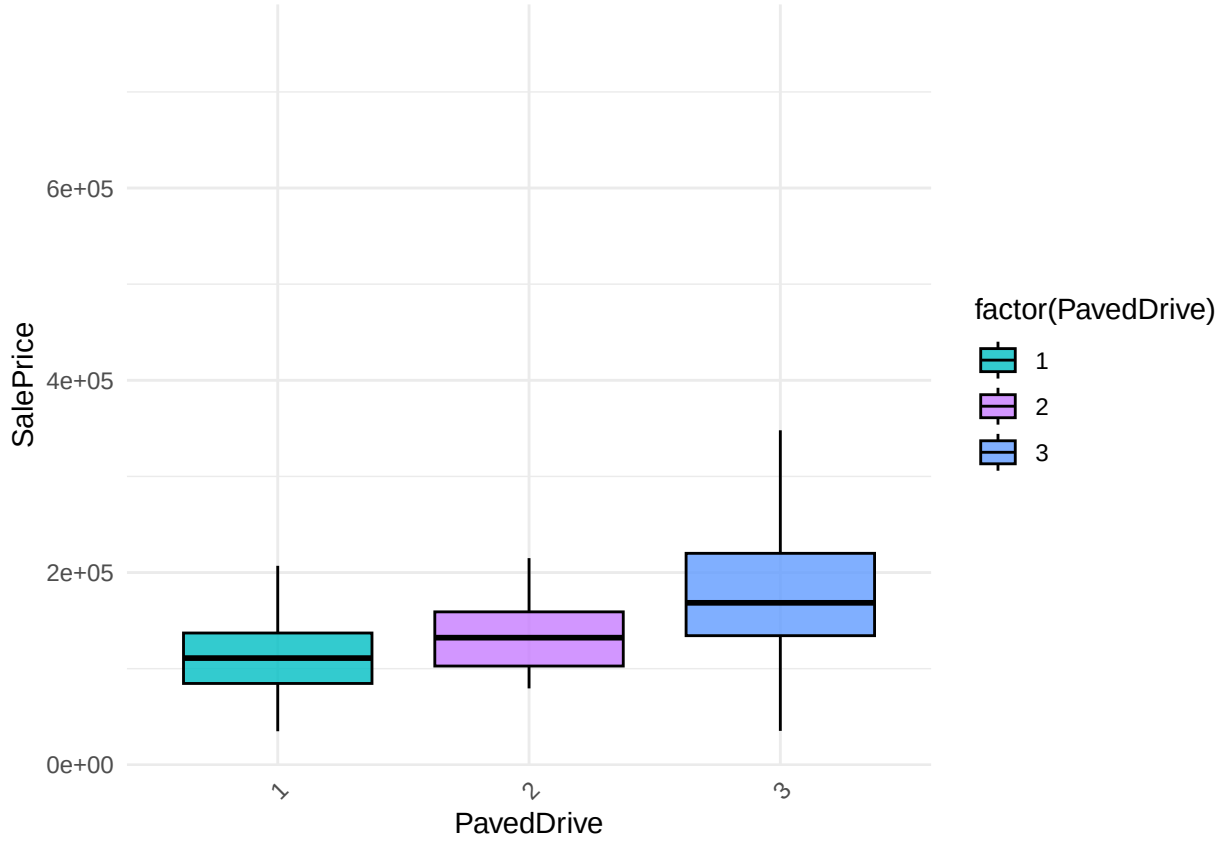
```
summary(aov(GarageCond~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1    0.3  0.2965   0.923  0.337
## Residuals   1458  468.5  0.3213
```

The values dont suggest a relation among the variables

61. PavedDrive

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(PavedDrive), y = SalePrice, fill = factor(PavedDrive))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "PavedDrive", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Category 3 have comparatively higher prices to other

```
kruskal.test(house$PavedDrive ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$PavedDrive by house$SalePrice
## Kruskal-Wallis chi-squared = 676.04, df = 662, p-value = 0.344
```

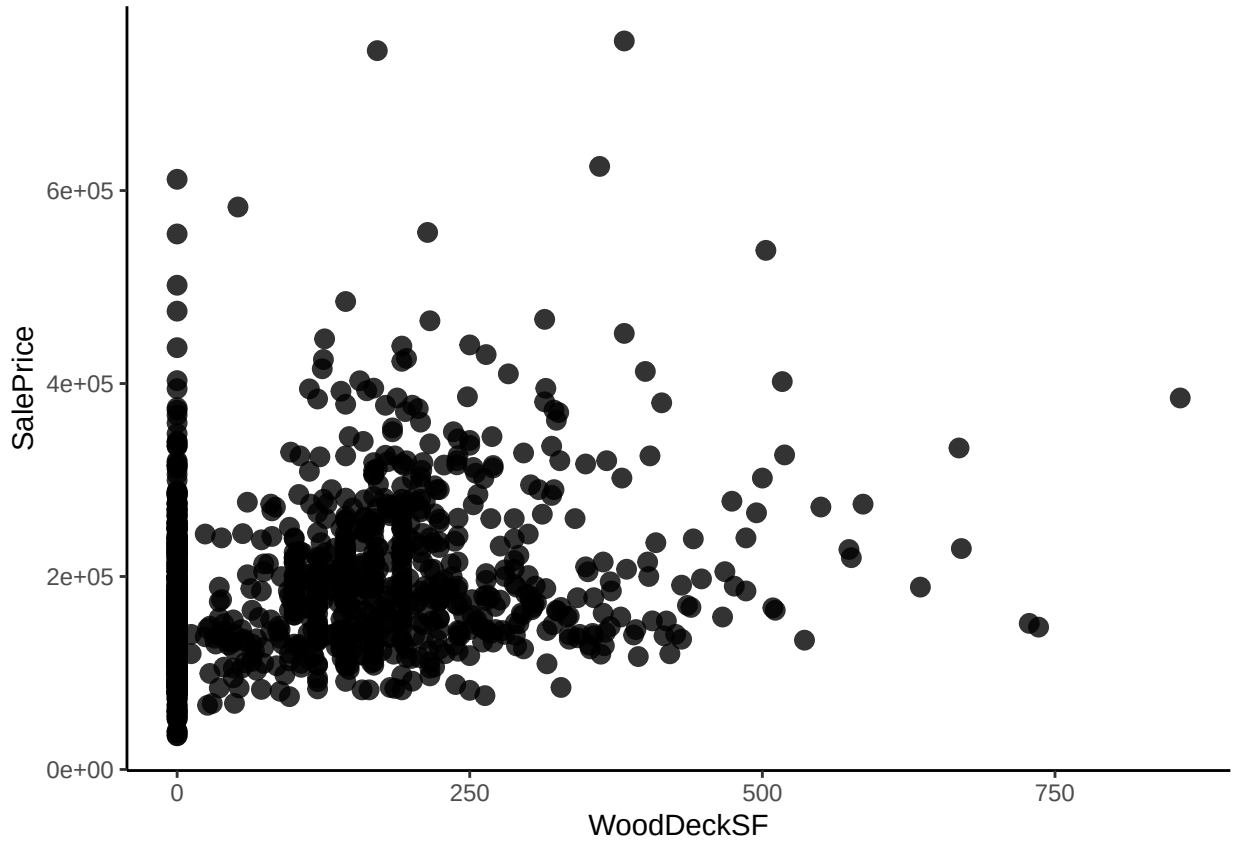
```
summary(aov(PavedDrive~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   19.3  19.258   82.45 <2e-16 ***
## Residuals   1458  340.5   0.234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mean of categories is different however median values are similar(maybe outliers)

62. WoodDeckSF

```
library(ggplot2)
ggplot(house, aes(x = WoodDeckSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "WoodDeckSF", y = "SalePrice") +
  theme_classic()
```



Rates increase gradually as the year goes ahead

```
kruskal.test(house$WoodDeckSF ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$WoodDeckSF by house$SalePrice
## Kruskal-Wallis chi-squared = 758.05, df = 662, p-value = 0.005543
```

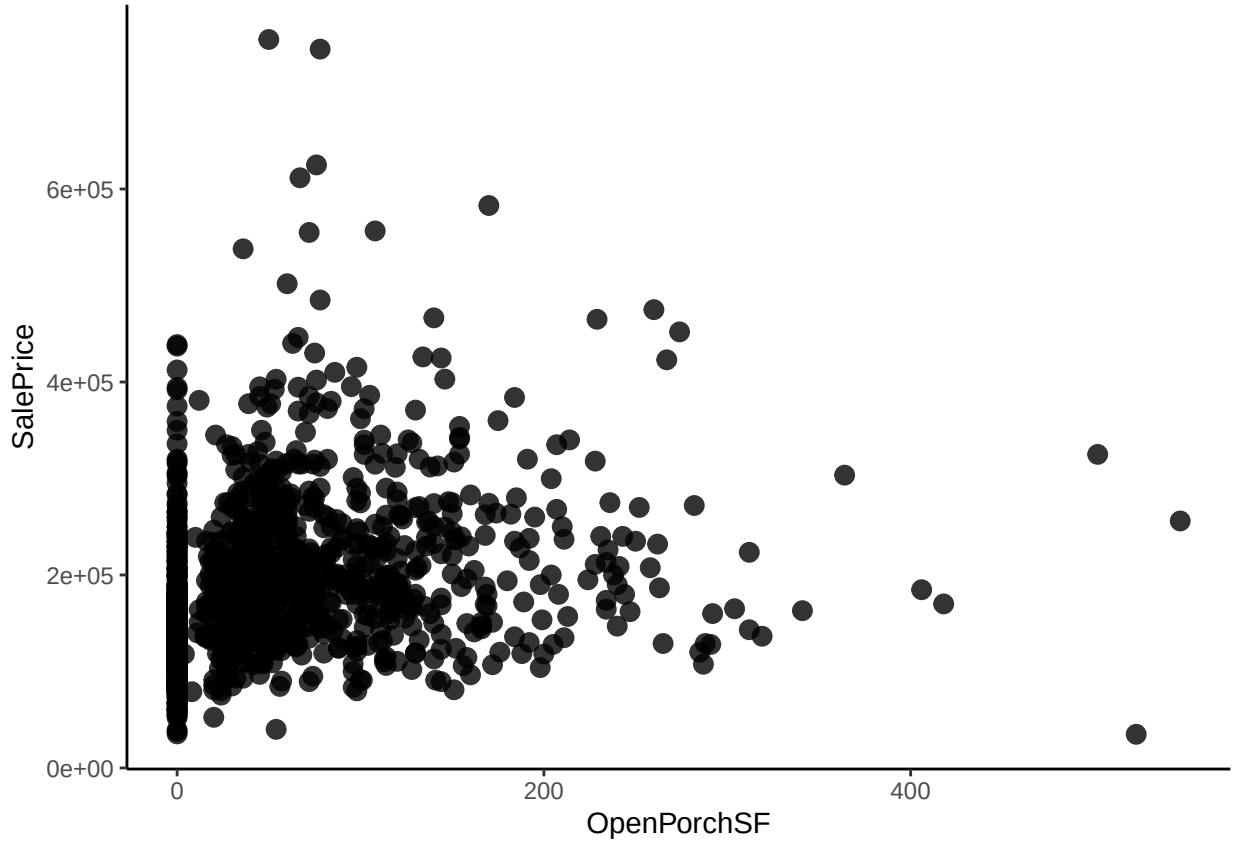
```
cor(house$WoodDeckSF, house$SalePrice)
```

```
## [1] 0.3244134
```

A positive correlation with the target

63. OpenPorchSF

```
library(ggplot2)
ggplot(house, aes(x = OpenPorchSF, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "OpenPorchSF", y = "SalePrice") +
  theme_classic()
```



Similar trend like previous variable

```
kruskal.test(house$OpenPorchSF ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$OpenPorchSF by house$SalePrice
## Kruskal-Wallis chi-squared = 844.73, df = 662, p-value = 1.74e-06
```

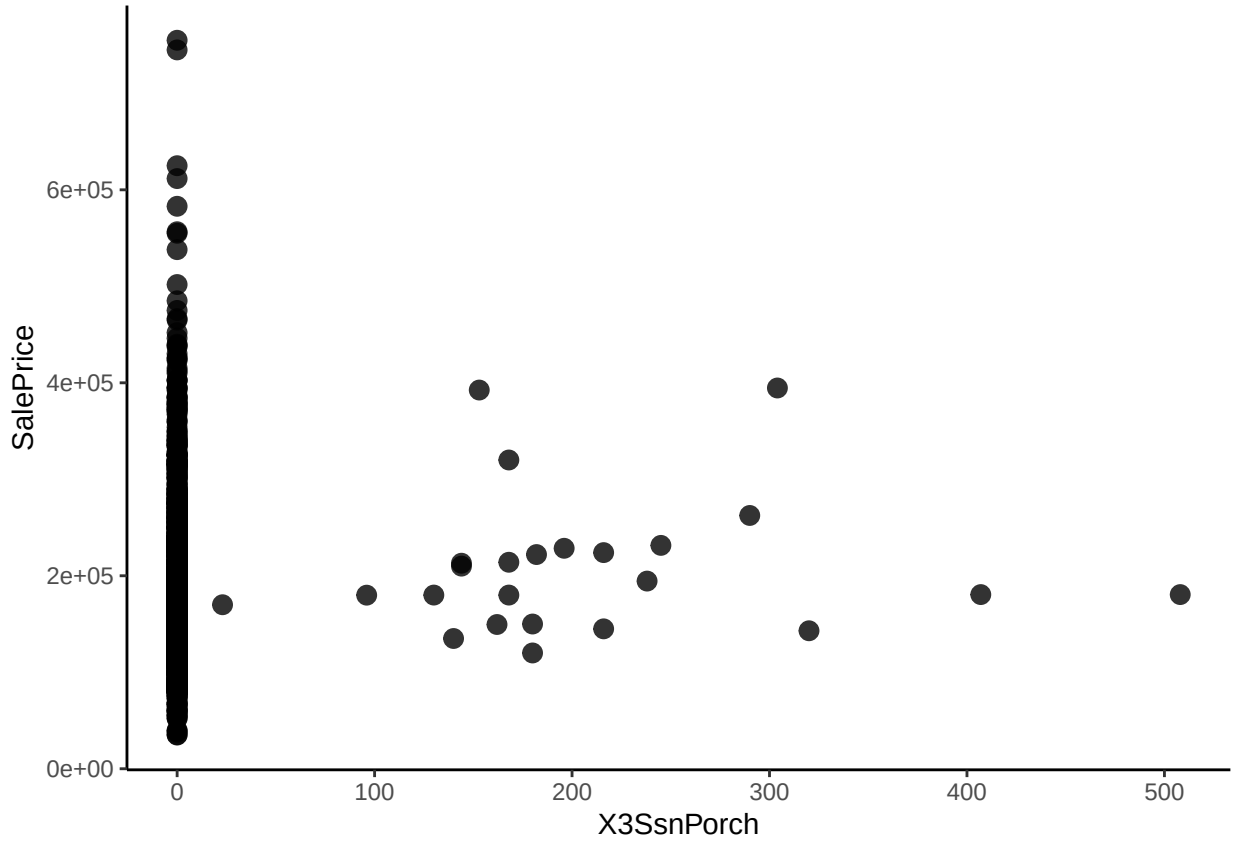
```
cor(house$OpenPorchSF, house$SalePrice)
```

```
## [1] 0.3158562
```

64. X3SsnPorch

```
library(ggplot2)
ggplot(house, aes(x = X3SsnPorch, y = SalePrice)) +
```

```
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "X3SsnPorch", y = "SalePrice") +
  theme_classic()
```




```
kruskal.test(house$X3SsnPorch ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$X3SsnPorch by house$SalePrice
## Kruskal-Wallis chi-squared = 555.67, df = 662, p-value = 0.999
```

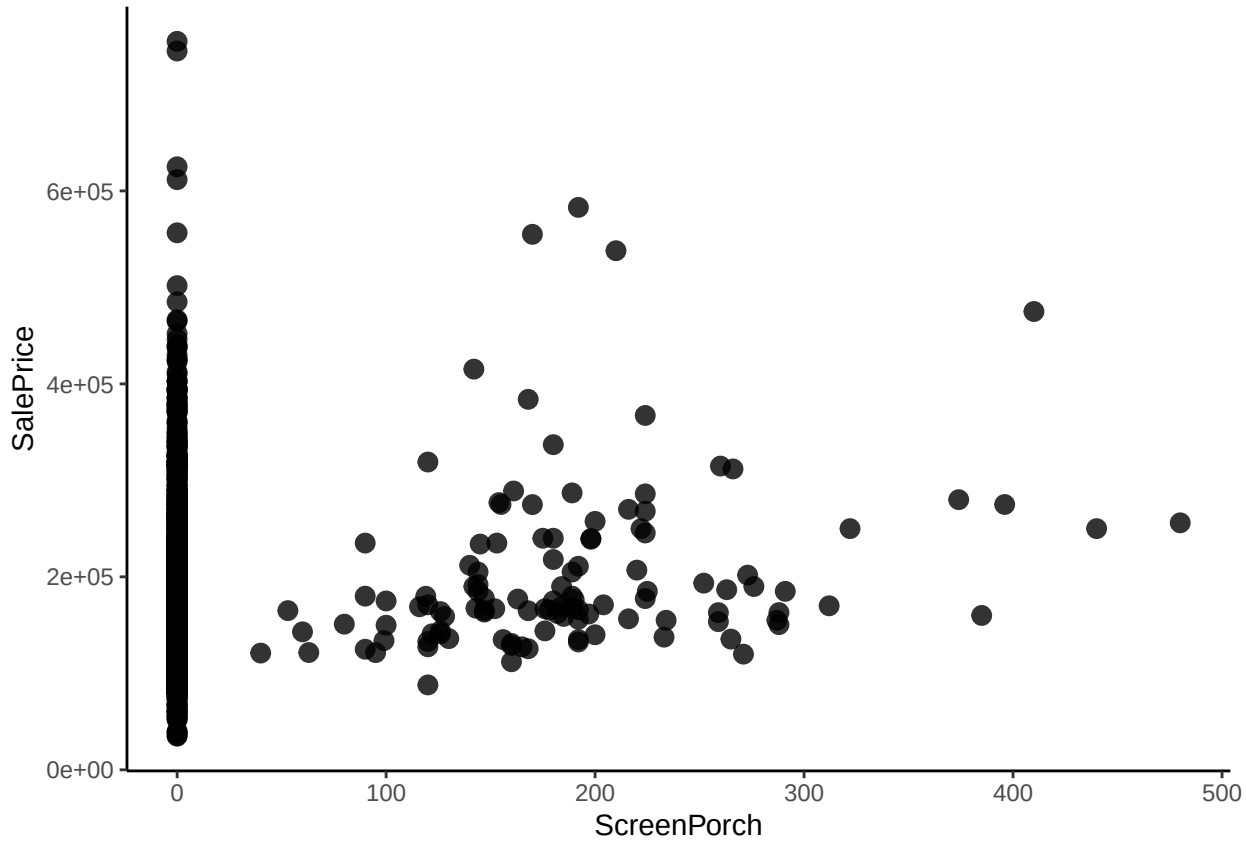
```
cor(house$X3SsnPorch, house$SalePrice)
```

```
## [1] 0.04458367
```

Almost no inference can be drawn from this variable

65. ScreenPorch

```
library(ggplot2)
ggplot(house, aes(x = ScreenPorch, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "ScreenPorch", y = "SalePrice") +
  theme_classic()
```




```
kruskal.test(house$ScreenPorch ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$ScreenPorch by house$SalePrice
## Kruskal-Wallis chi-squared = 638.19, df = 662, p-value = 0.7402
```

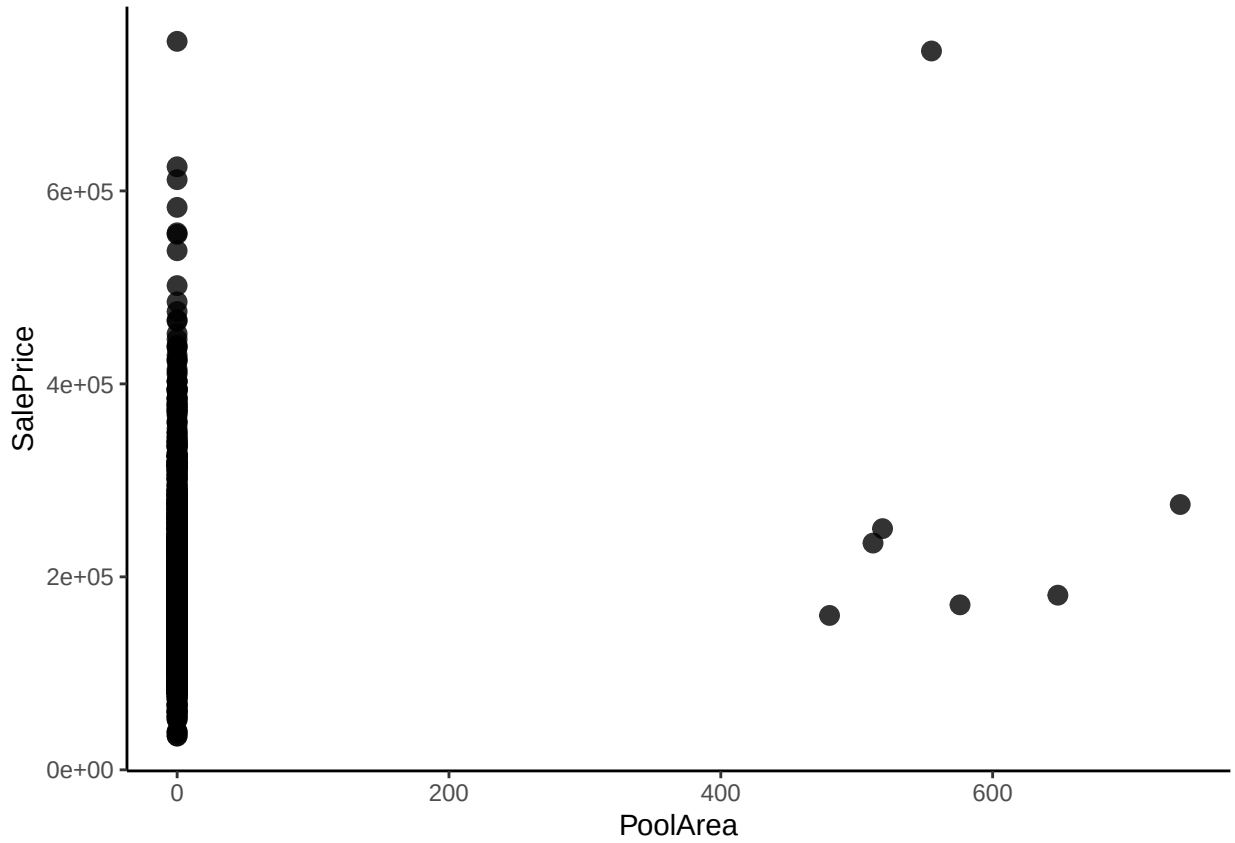
```
cor(house$ScreenPorch, house$SalePrice)
```

```
## [1] 0.1114466
```

Again very less inference can be drawn from it

66. PoolArea

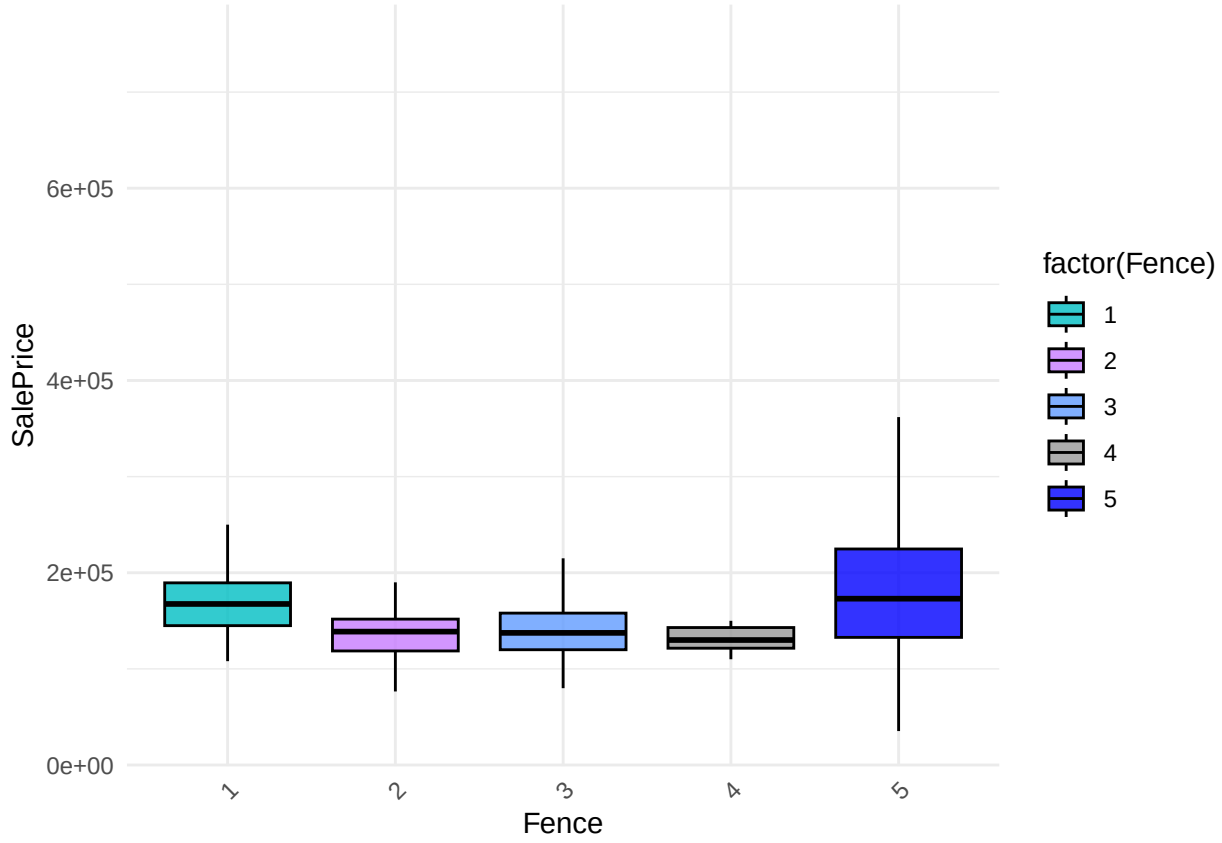
```
library(ggplot2)
ggplot(house, aes(x = PoolArea, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "PoolArea", y = "SalePrice") +
  theme_classic()
```



Almost no pool in most of the houses make this variable redundant

67. Fence

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(Fence), y = SalePrice, fill = factor(Fence))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "Fence", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$Fence ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$Fence by house$SalePrice
## Kruskal-Wallis chi-squared = 606.36, df = 662, p-value = 0.9401
```

```
summary(aov(Fence~SalePrice,data=house))
```

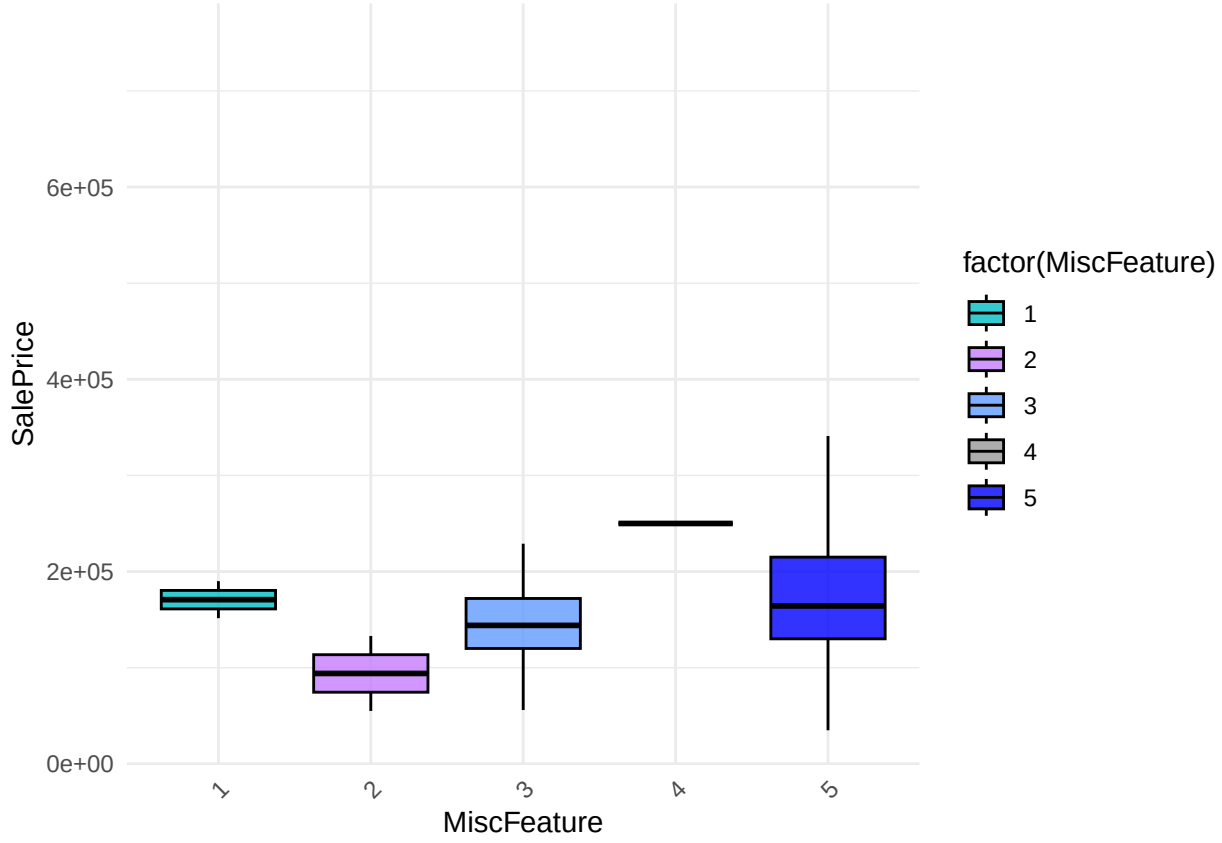
```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SalePrice      1   33.8   33.84   29.42 6.81e-08 ***
## Residuals   1458 1677.1    1.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Difference in mean can be seen clearly but the medians tend to be similar

68. MiscFeature

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(MiscFeature), y = SalePrice, fill = factor(MiscFeature))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
```

```
  scale_fill_manual(values = my_colors) +
  labs(x = "MiscFeature", y = "SalePrice") +
  theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



quite a difference between categories

```
kruskal.test(house$MiscFeature ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$MiscFeature by house$SalePrice
## Kruskal-Wallis chi-squared = 628.67, df = 662, p-value = 0.8195
```

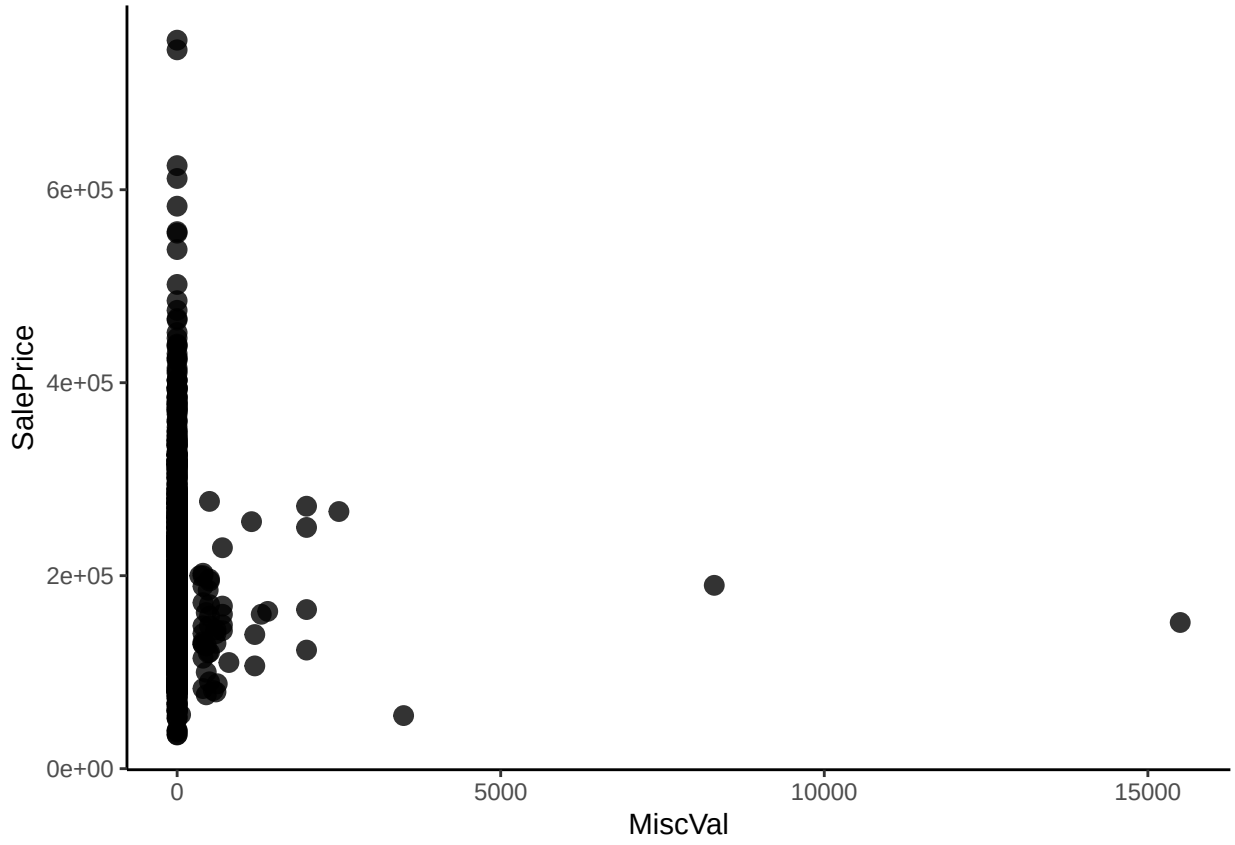
```
summary(aov(MiscFeature~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## SalePrice      1   1.29  1.2909   7.943 0.00489 **
## Residuals   1458 236.96  0.1625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again a difference in the mean values can be seen with the tests

69. MiscVal

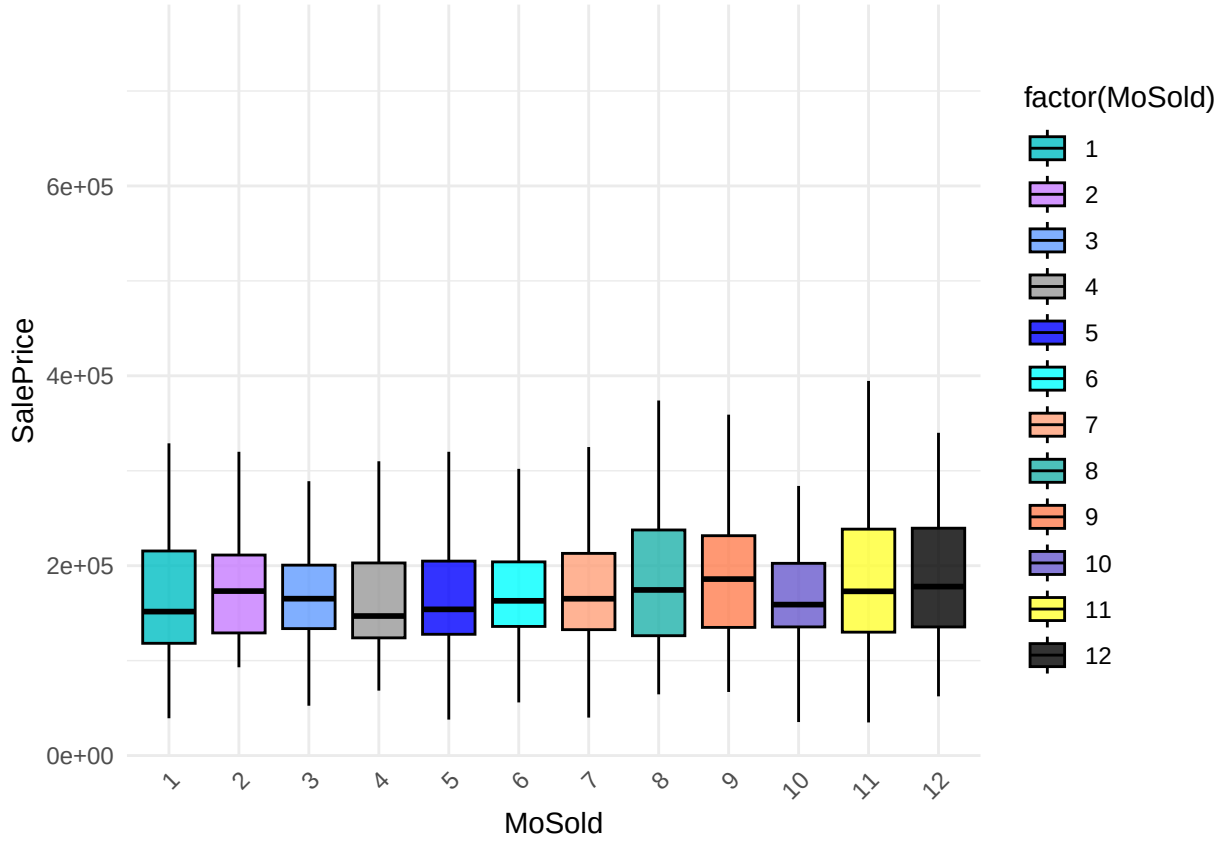
```
library(ggplot2)
ggplot(house, aes(x = MiscVal, y = SalePrice)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "MiscVal", y = "SalePrice") +
  theme_classic()
```



Almost all houses don't have a miscellaneous feature making it redundant

70. MoSold

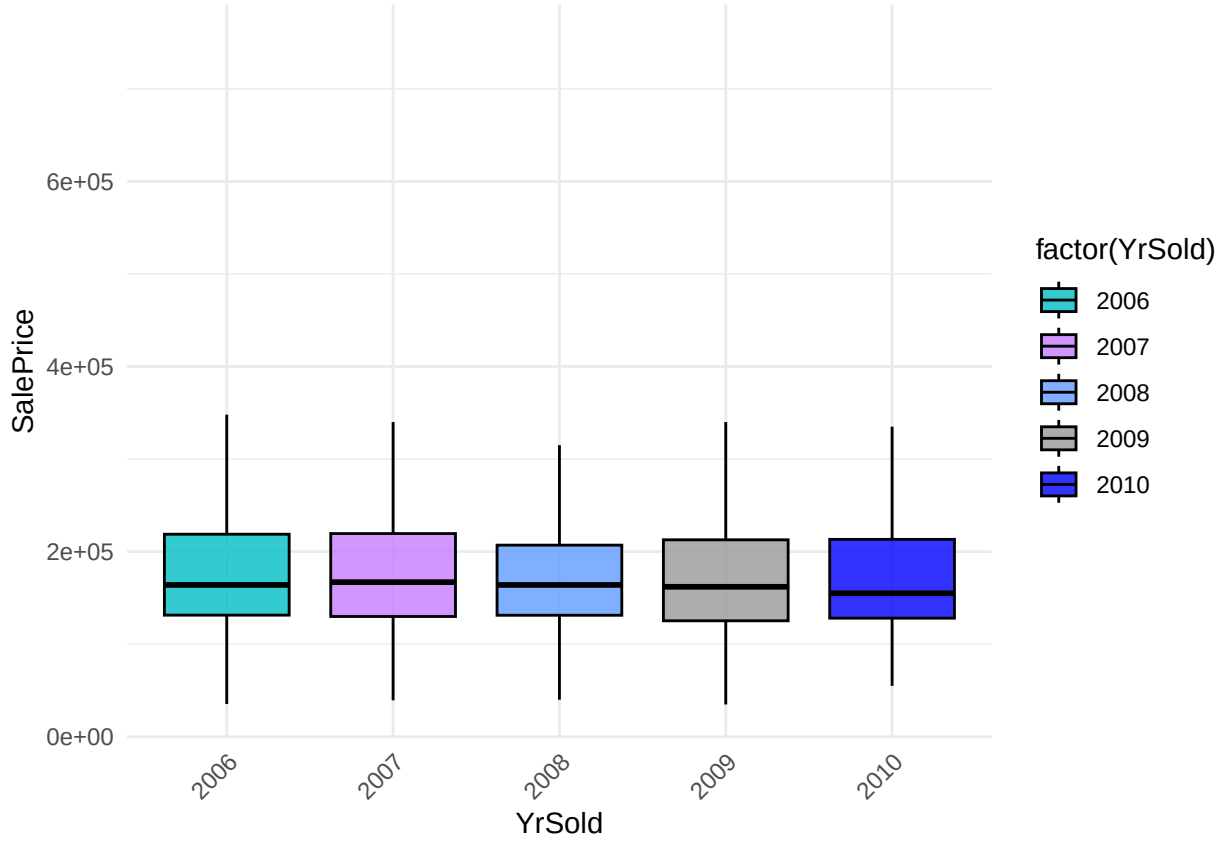
```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(MoSold), y = SalePrice, fill = factor(MoSold))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "MoSold", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Almost no difference between months as far as prices are concerned

71. YrSold

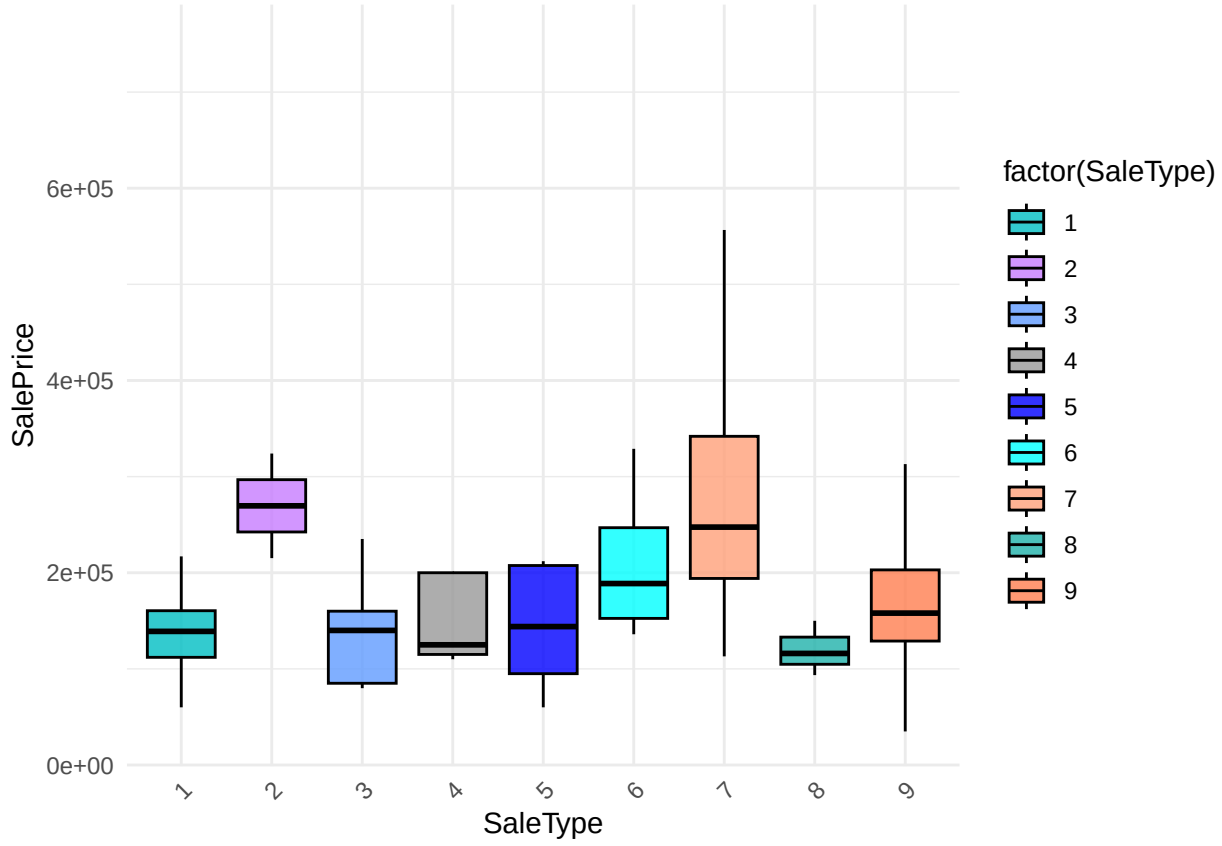
```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(YrSold), y = SalePrice, fill = factor(YrSold))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "YrSold", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Again almost no difference between the years of the selling of house

72. SaleType

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(SaleType), y = SalePrice, fill = factor(SaleType))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
  scale_fill_manual(values = my_colors) +
  labs(x = "SaleType", y = "SalePrice") +
  theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$SaleType ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$SaleType by house$SalePrice
## Kruskal-Wallis chi-squared = 883.21, df = 662, p-value = 1.623e-08
```

```
summary(aov(SaleType~SalePrice,data=house))
```

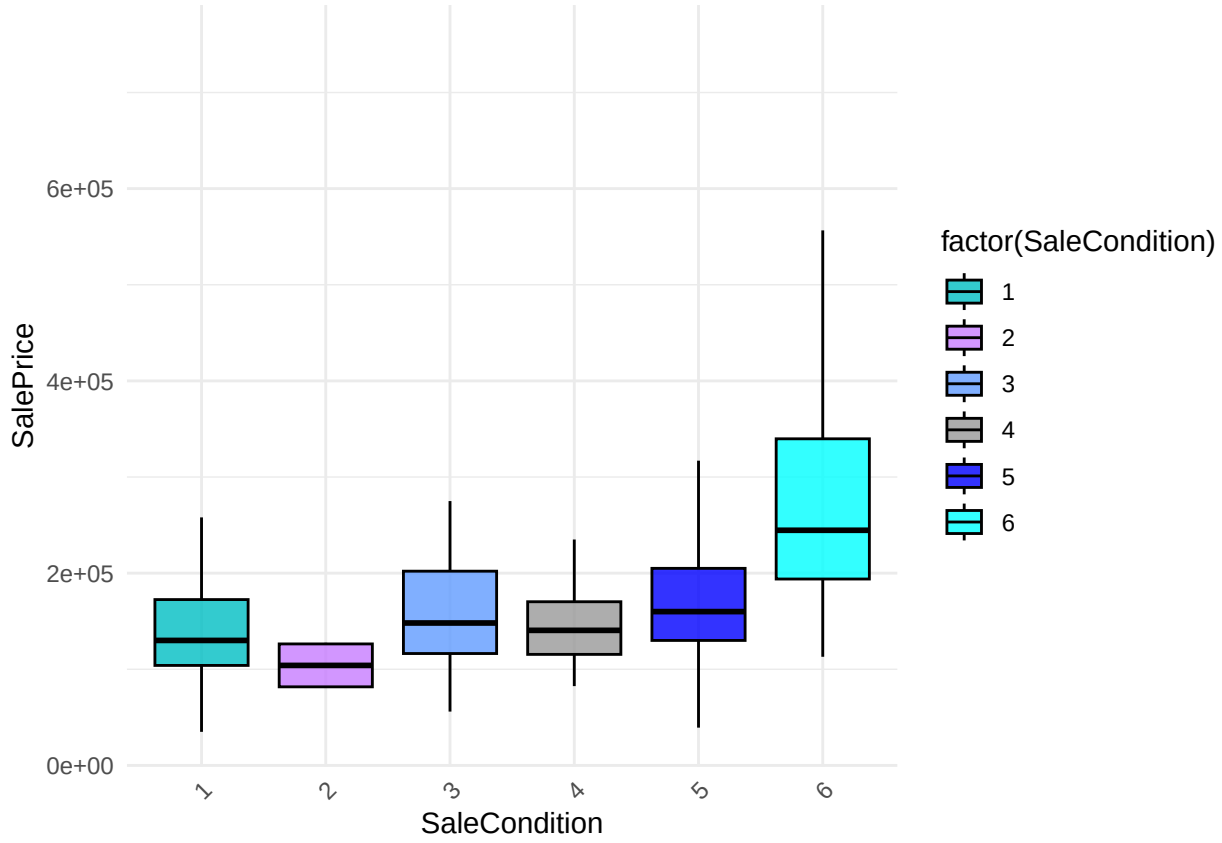
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1      9   9.019   3.708 0.0543 .
## Residuals   1458   3546   2.432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A strong relation with the target variable is visible

73. SaleCondition

```
library(ggplot2)
my_colors <- c( "#00BFC4", "#C77CFF", "#619CFF", "#999999", "blue",   "#00FFFF", "#FFA07A", "#20B2AA",
ggplot(house, aes(x = factor(SaleCondition), y = SalePrice, fill = factor(SaleCondition))) +
  geom_boxplot(alpha = 0.8, color = "black", outlier.shape = NA) +
```

```
  scale_fill_manual(values = my_colors) +
  labs(x = "SaleCondition", y = "SalePrice") +
  theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
kruskal.test(house$SaleCondition ~ house$SalePrice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  house$SaleCondition by house$SalePrice
## Kruskal-Wallis chi-squared = 909.33, df = 662, p-value = 4.808e-10
```

```
summary(aov(SaleCondition~SalePrice,data=house))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SalePrice      1   80.3   80.29   69.35 <2e-16 ***
## Residuals   1458 1687.8    1.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A very strong relation with the target variable is visible

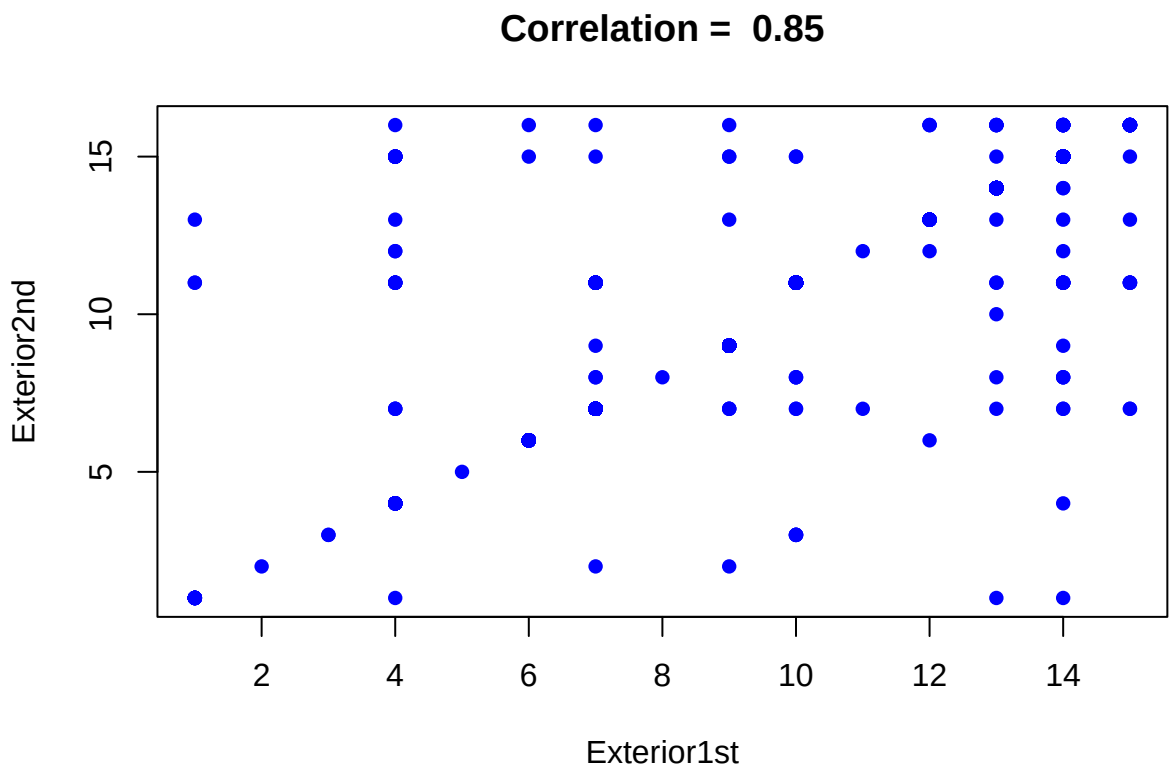
Dimension Reduction

We check the correlation of variables to check the relation between them

We remove those variables which had not much relation with the Sale Price derived from graphical analysis and tests

```
reduced = subset(house, select = c(-YrSold,-MoSold,-MiscVal,         -PoolArea,-ScreenPorch,-X3SsnPorch
```

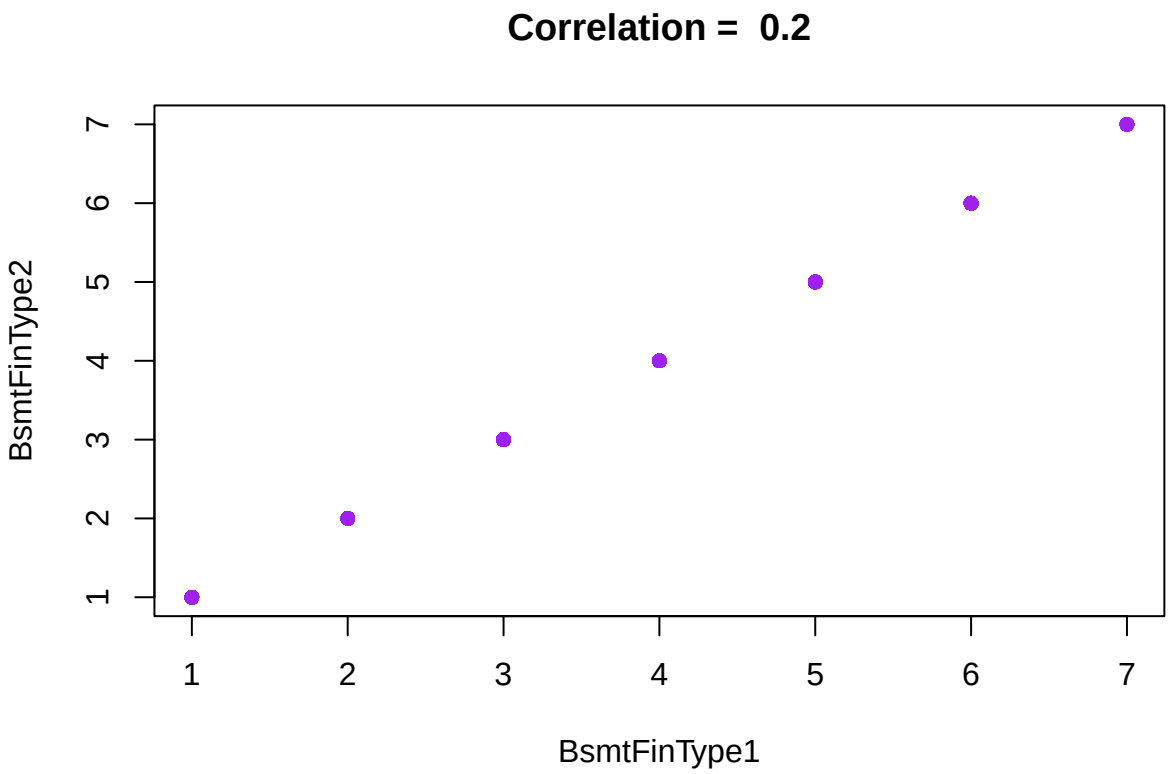
```
correlation <- cor(house$Exterior1st, house$Exterior2nd)

plot(house$Exterior1st, house$Exterior2nd, pch = 16, col = "blue", xlab = "Exterior1st", ylab = "Exteri
     main = paste("Correlation = ", round(correlation, 2)))
```



Clearly a very high overlap between the two fields with +0.85 correlation

```
reduced = subset(reduced, select = -Exterior2nd)
```

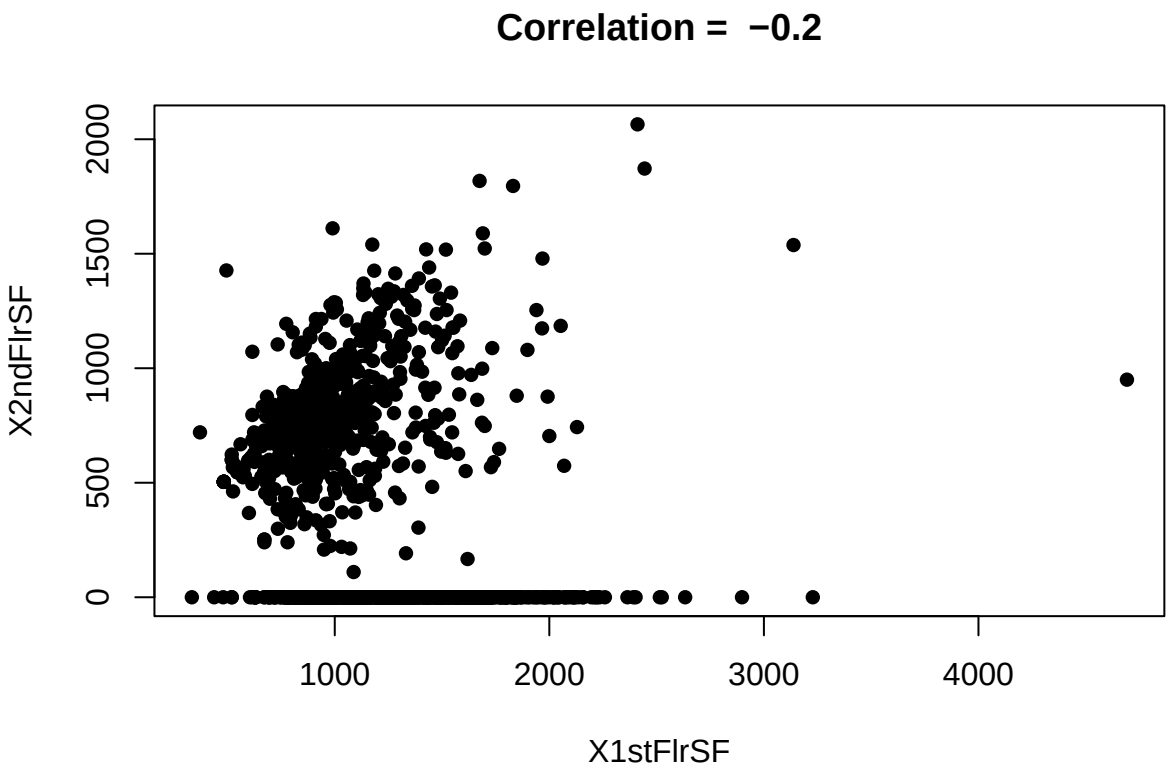
```
correlation <- cor(house$BsmtFinType1, house$BsmtFinType2)

plot(house$BsmtFinType1, house$BsmtFinType1, pch = 16, col = "purple", xlab = "BsmtFinType1", ylab = "Bs
     main = paste("Correlation = ", round(correlation, 2)))
```



From the plots done while analyzing variables with the target variable we can see that the Type 1 and Type 2 of BsmtFin had quite similar plots and the exact same categories and hence we can remove type 2 because it measures more than 1 which is not required

```
reduced = subset(reduced, select = -BsmtFinType1)
```

```
correlation <- cor(house$X1stFlrSF, house$X2ndFlrSF)

plot(house$X1stFlrSF, house$X2ndFlrSF, pch = 16, col = "black", xlab = "X1stFlrSF", ylab = "X2ndFlrSF",
     main = paste("Correlation = ", round(correlation, 2)))
```



We can drop one of the variables which has less relation with the target variable as both follow a linear relation with the target variable, we remove the 2nd because it had a correlation of 0.3 while the first variable has a correlation of 0.6 with the Sale Price

```
reduced = subset(reduced, select = -X2ndFlrSF)
```

We remove utilities also as it consists of all values in a signle class. Also central is a almost similar variable of centralAir so we drop it.

```
reduced = subset(reduced, select = c(-Utilities,-Central))
```

BldgType also had a high p value on performing the tests and can be removed. Also LotConfig had a smaller p value so we drop it.

```
reduced = subset(reduced, select = c(-BldgType,-LotConfig,-BsmtUnfSF))
```

Various other variables which had either a low p value or very less relation with the target variable can also be removed to reduce dimensions further

```
reduced = subset(house, select = c(-LowQualFinSF,-MiscFeature,-Foundation,-Electrical,-Fence,-HalfBath,
```

```
house = reduced
write.csv(house,"train.csv")
```

##Modelling

###Validation and train split

```r
#set.seed(123)
train_data = read.csv("D:/SEM-6/Statistics/Project/Housing Prices/train.csv", header = TRUE)

train_indices <- sample(nrow(train_data), round(0.8 * nrow(train_data)), replace = FALSE)
train <- train_data[train_indices, ]
validation <- train_data[-train_indices, ]
temp = validation
#validation <- validation[, -which(names(validation) == "SalePrice")]
#write.csv(validation,"validation.csv")
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```r
lm_model = train(SalePrice ~ ., data=train,method="lm")
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
```

```
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
predictions <- predict(lm_model, newdata = validation)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52640  125148  171013  187422  232513  548571
```

```
summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52640  125148  171013  187422  232513  548571
```

```
mse <- mean((predictions - validation$SalePrice))
mse
```

```
## [1] -1472.815
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.2.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
rf_model <- randomForest(SalePrice ~ ., data = train)

predictions <- predict(rf_model, newdata = validation)
validation$SalePrice = predictions
write.csv(validation,"validation.csv")
summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   85105  133450  163762  188452  215171  567165
```

```
summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   85105  133450  163762  188452  215171  567165
```

```
mse <- mean((temp$SalePrice - validation$SalePrice)^2)
mse
```

```
## [1] 1102466148
```

```
test_data <- read.csv("test.csv")
sum(is.na(test_data))
```

```
## [1] 0
```

```
missing_cols <- colSums(is.na(test_data)) > 0
missing_cols
```

```
##           X.1             X            Id      MSZoning       LotArea
##         FALSE         FALSE         FALSE         FALSE         FALSE
##         Alley     LotShape   LandContour     Utilities     LotConfig
##         FALSE         FALSE         FALSE         FALSE         FALSE
##  Neighborhood    Condition2      BldgType   OverallQual   OverallCond
##         FALSE         FALSE         FALSE         FALSE         FALSE
##     YearBuilt  YearRemodAdd   Exterior1st   Exterior2nd     ExterQual
##         FALSE         FALSE         FALSE         FALSE         FALSE
##      BsmtQual  BsmtExposure  BsmtFinType1    BsmtFinSF1     BsmtUnfSF
##         FALSE         FALSE         FALSE         FALSE         FALSE
##   TotalBsmtSF       Heating    CentralAir     X1stFlrSF     X2ndFlrSF
##         FALSE         FALSE         FALSE         FALSE         FALSE
##     GrLivArea  BsmtFullBath      FullBath  KitchenAbvGr   KitchenQual
##         FALSE         FALSE         FALSE         FALSE         FALSE
##  TotRmsAbvGrd    Functional    Fireplaces    GarageType  GarageFinish
##         FALSE         FALSE         FALSE         FALSE         FALSE
##    GarageCars    GarageArea    GarageQual    PavedDrive    X3SsnPorch
##         FALSE         FALSE         FALSE         FALSE         FALSE
##   ScreenPorch      PoolArea       MiscVal        MoSold      SaleType
##         FALSE         FALSE         FALSE         FALSE         FALSE
## SaleCondition       Central     SalePrice
##         FALSE         FALSE         FALSE
```

```
for (col in names(test_data)[missing_cols]) {
  col_mean <- mean(test_data[[col]], na.rm = TRUE)
  test_data[[col]][is.na(test_data[[col]])] <- col_mean
}
sum(is.na(test_data))
```

```
## [1] 0
```

```
write.csv(test_data,"test.csv")
```

```
library(caret)
train_data = read.csv("D:/SEM-6/Statistics/Project/Housing Prices/train.csv", header = TRUE)

lm_model = train(SalePrice ~ ., data=train_data,method="lm")
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
```

```
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
```

```
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
predictions <- predict(lm_model, newdata = test_data)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
test_data$SalePrice = predictions
write.csv(test_data,"test.csv")
```

```
library(randomForest)

train_data <- read.csv("train.csv")
rf_model <- randomForest(SalePrice ~ ., data = train_data)

test_data <- read.csv("test.csv")
predictions <- predict(rf_model, newdata = test_data)

test_data$SalePrice <- predictions
write.csv(test_data, "test.csv", row.names = FALSE)
```