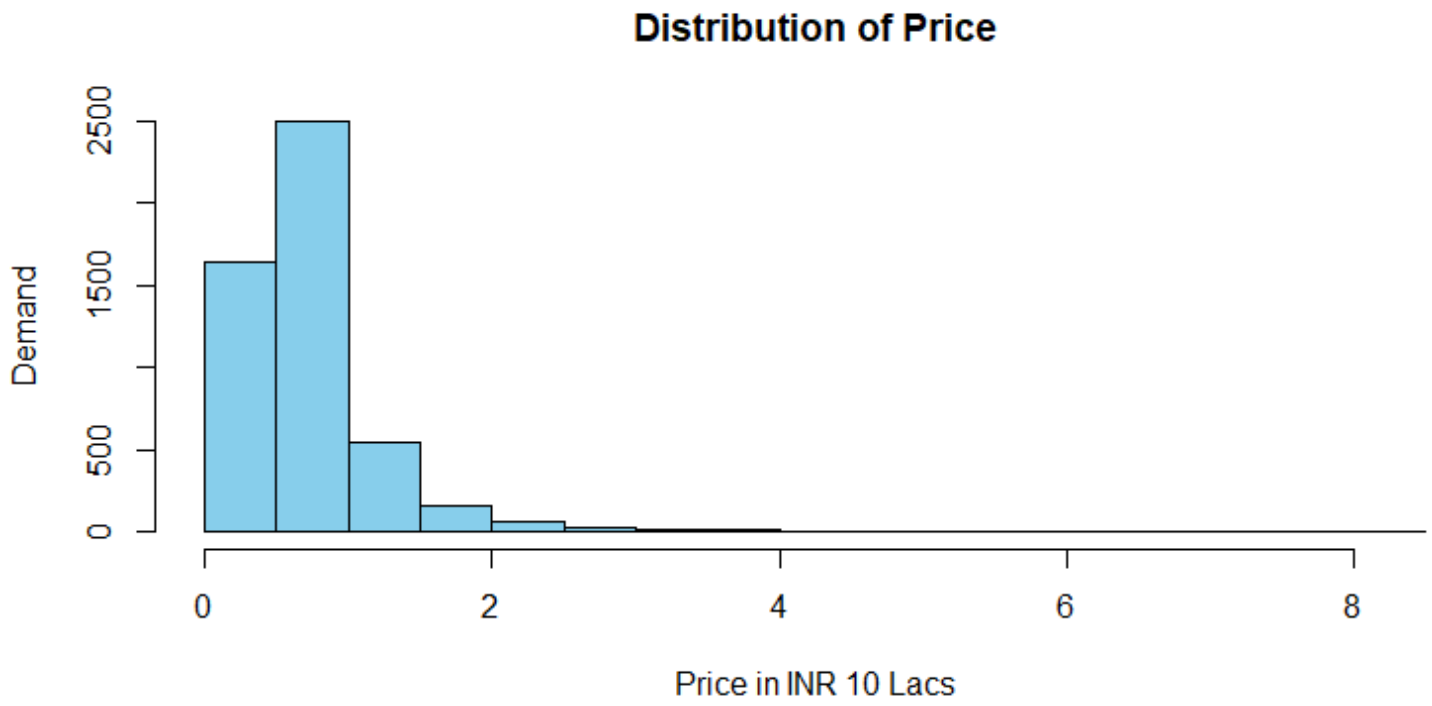
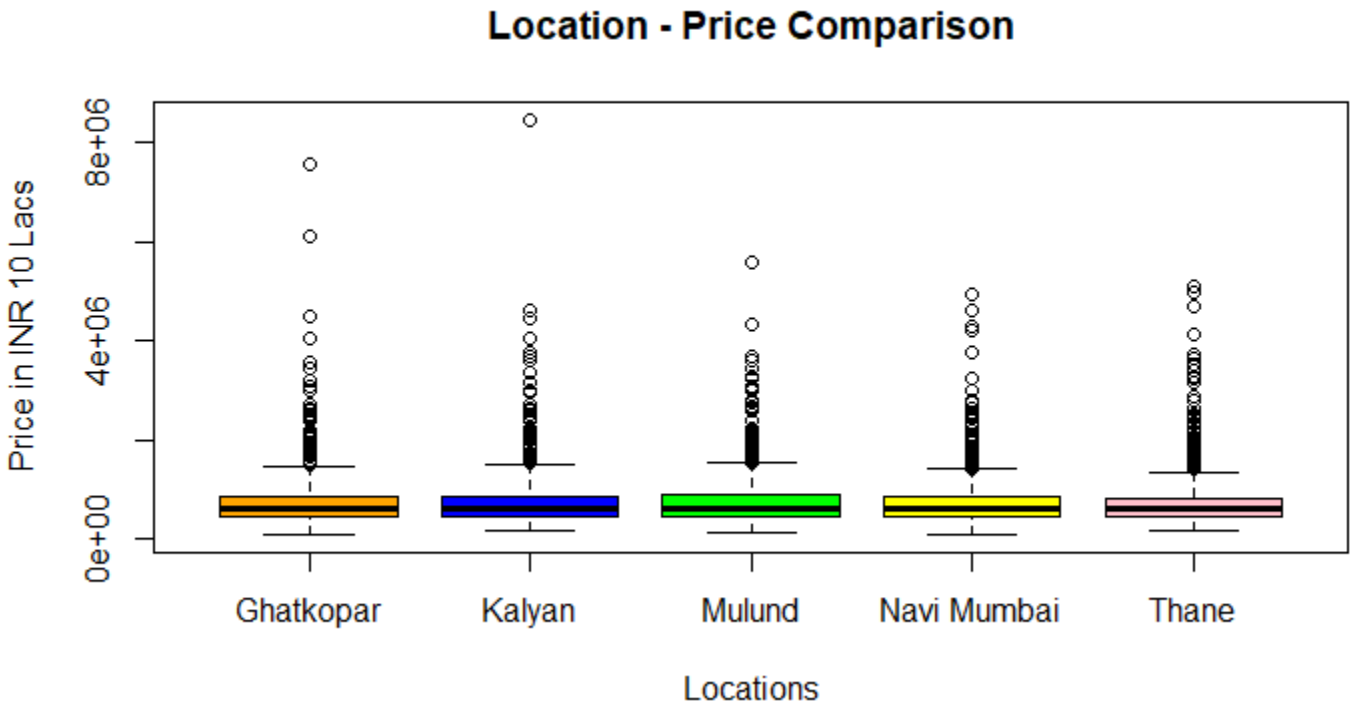
```
# Creating Real Estate Analysis Project in R
# Sachin Korgaonkar
# Data Science Project
# Dated 5-June-2020

# libraries import
library(ggplot2)
library(ggmap)
library(dplyr)

# Import Mumbai and Navi Mumbai real estate data
# This data is downloaded from kaggle
house.data <- read.csv('Mumbai_realestate_data.csv',header=TRUE)

# Data Cleanup Activity
# The date from this data frame is not required
# Create Data Frame from this data and store as house.data
house.data <- house.data[1:5000,]

attach(house.data)

# Data Exploration
# first view the structure of the data
# See all columns in data frame
glimpse(house.data)

# View a summary of the house data
summary(house.data)


# Put average price as benchmark
# In India INR 10 Lacs are minimum price for household.
pricesIn10Lacs <- house.data$price / 1000000

# Price Distribution
# Find out how Price Distribution is aligned.
# X will show from minimum to maximum price.
# Y will show the approx total numbers (frequency) of prices appear in the list.
# Graph shows which price has maximum demand.
hist(pricesIn10Lacs,
     data = house.data,
     main = 'Distribution of Price',
     xlab = 'Price in INR 10 Lacs',
     ylab = 'Demand',
     col = 'Sky blue',
     bins = 5
)
```

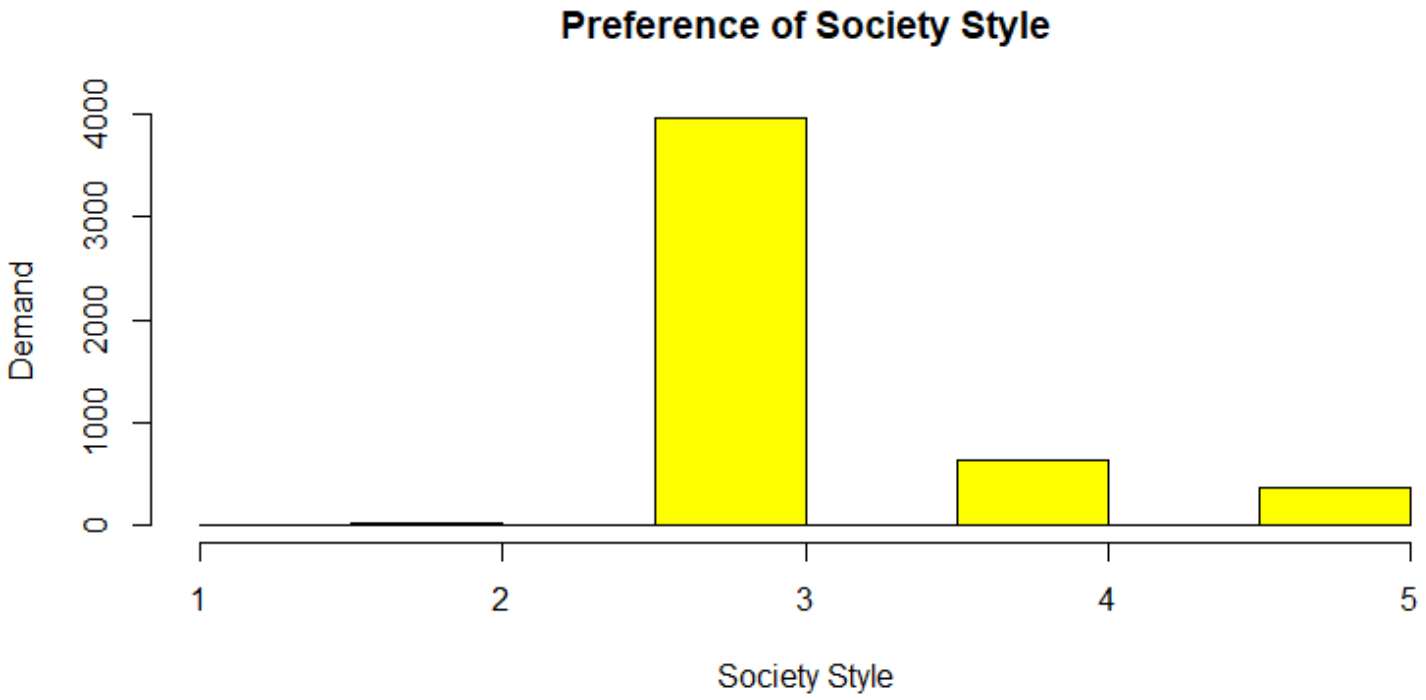



```
# Bedrooms Distribution
# How many bedroom people demand?
# Graph shows 2.5 to 3 bedroom apartments are more popular
hist(house.data$bedrooms,
     main = 'Distribution of Bedrooms',
     xlab = 'Number of Bedrooms',
     ylab = 'Demand',
     col = 'Green'
)
```

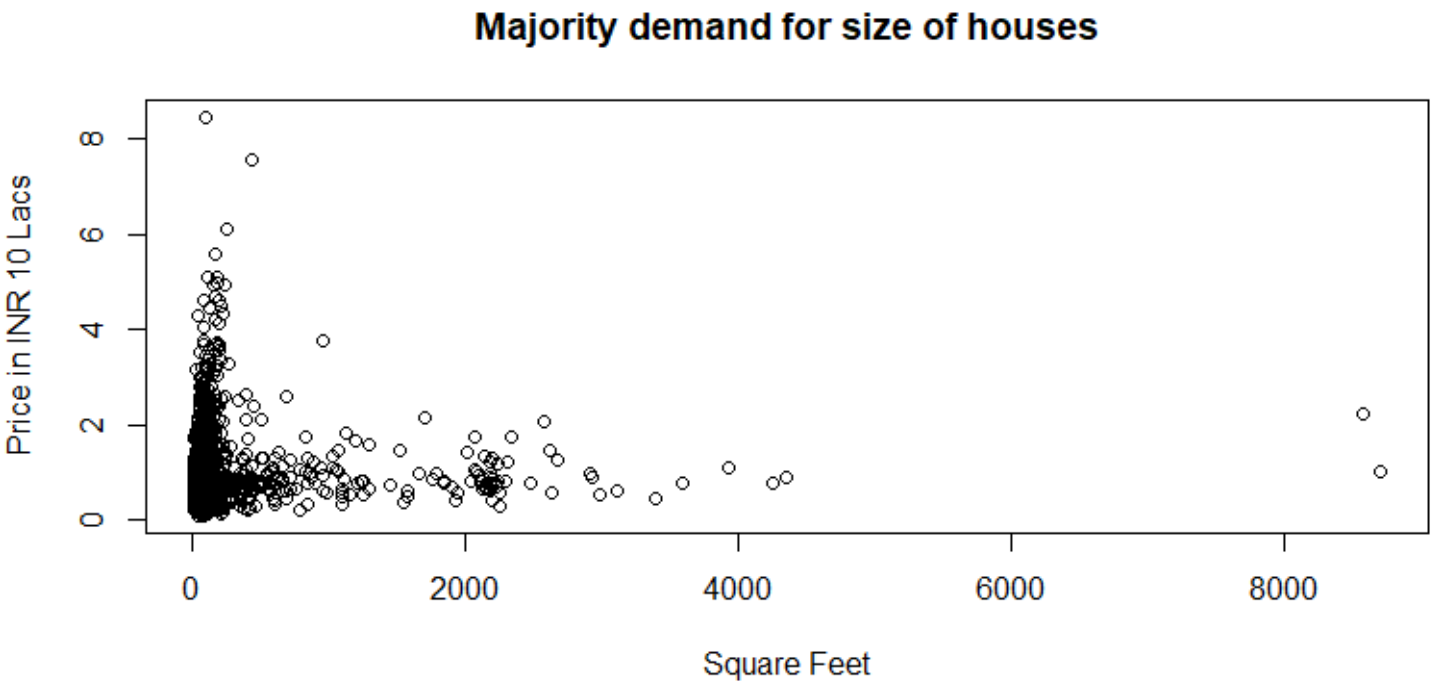
```
# Identify which location is in our budget
plot(price ~ location,
     main = 'Location - Price Comparison',
     xlab = 'Locations',
     ylab = 'Price in INR 10 Lacs',
     col = c('Orange', 'Blue','Green','Yellow','Pink')
)
```



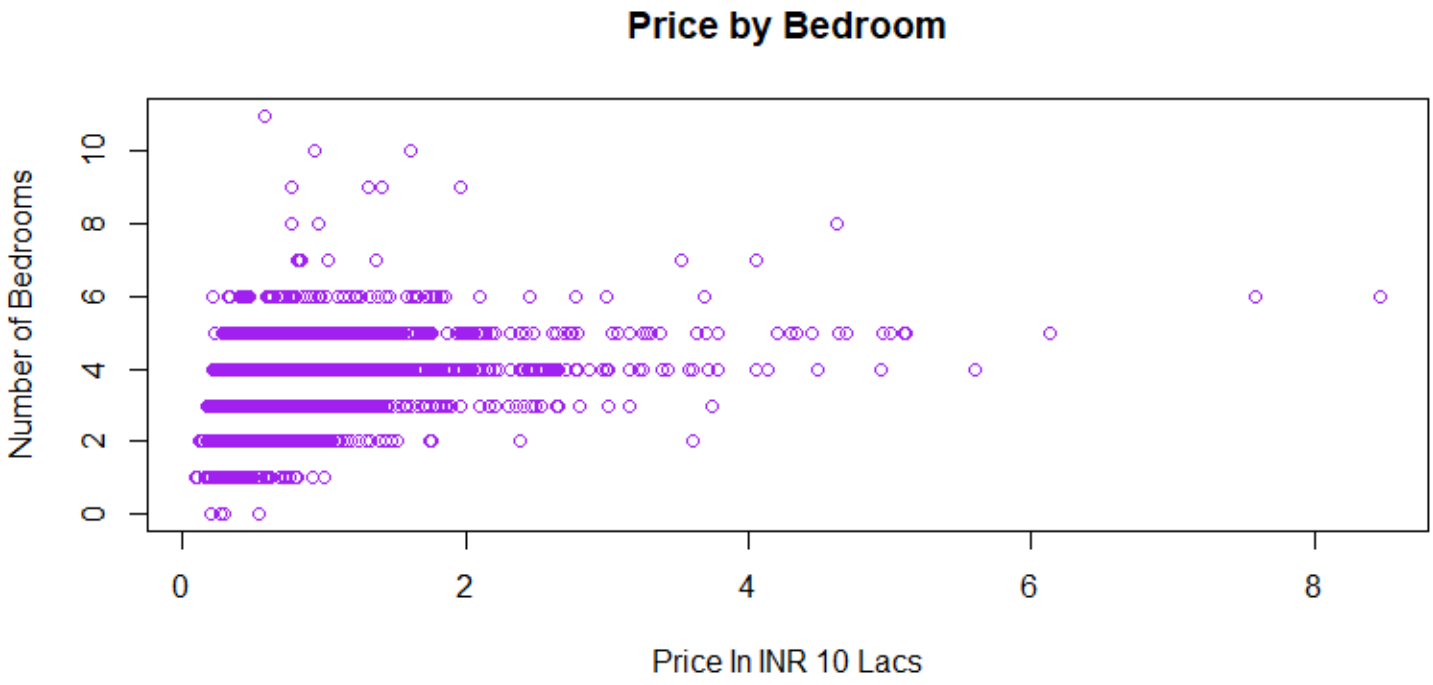

```
# How popular is society style. Whether people prefer Single tower or township
# Graph Shows people prefer township
hist(house.data$towers,
     main = 'Preference of Society Style',
     xlab = 'Society Style',
     ylab = 'Demand',
     col = 'yellow'
)
```



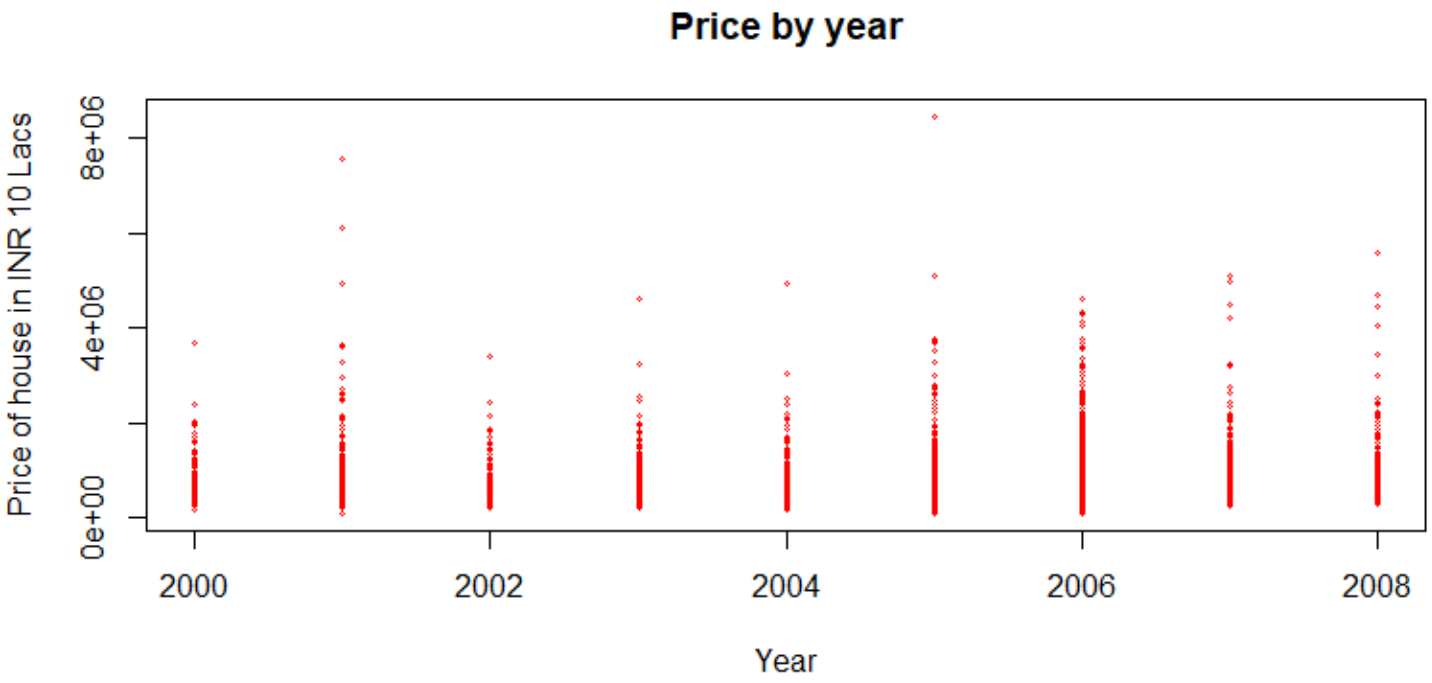

```
# Put SQFT by 100
squareFt <- house.data$sqft_lot15 /100
# Price by Sqft
plot(y = pricesIn10Lacs,
     main = 'Majority demand for size of houses',
     x = squareFt,
     xlab = 'Square Feet',
     ylab = 'Price in INR 10 Lacs'
)
```



```
# Prices by bedrooms
plot(pricesIn10Lacs,
     bedrooms,
     data = house.data,
     main = 'Price by Bedroom',
     col = 'purple',
     xlab = 'Price In INR 10 Lacs',
     ylab = 'Number of Bedrooms'
)
```




```
# 2. MLR Model
# Drop Date from model,
# create baseline model
house.model <- lm(price ~ ., data = house.data)

summary(house.model)

# Round Coefficient Table
coeffs <- summary(house.model)$coefficients
coeffs <- round(coeffs,4)
coeffs

# Create a scatter plot
plot(price ~ yr_built,
     data = house.data,
     cex = .4,
     col = 'red',
     main = 'Price by year',
     xlab = 'Year',
     ylab = 'Price of house in INR 10 Lacs'
     )
```



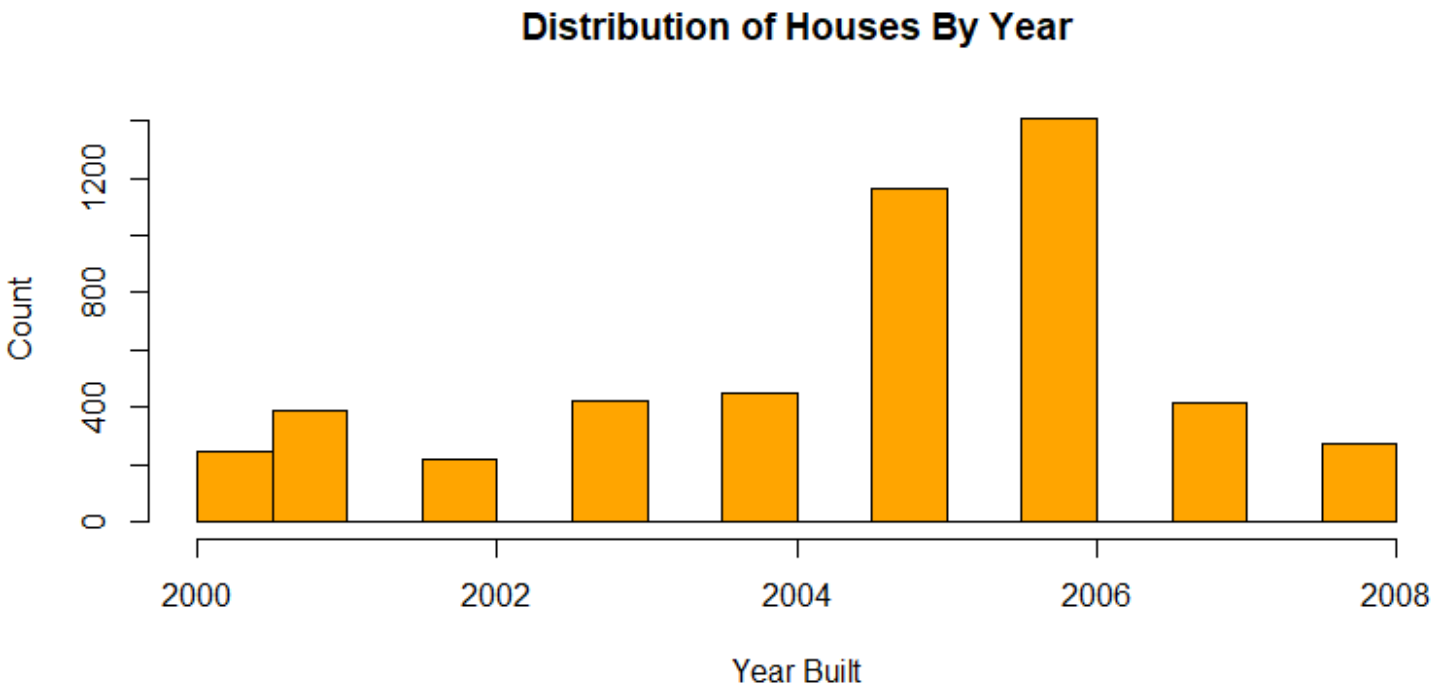

```
# I want to see if house prices on average vary by quarter centuries
# Grab the price and year to convert year into decade factor
priceByDecade <- data.frame(Price = house.data$price, Decade = house.data$yr_built)

# Find the earliest Year Built
min(priceByDecade$Decade)
```

[1] 2000

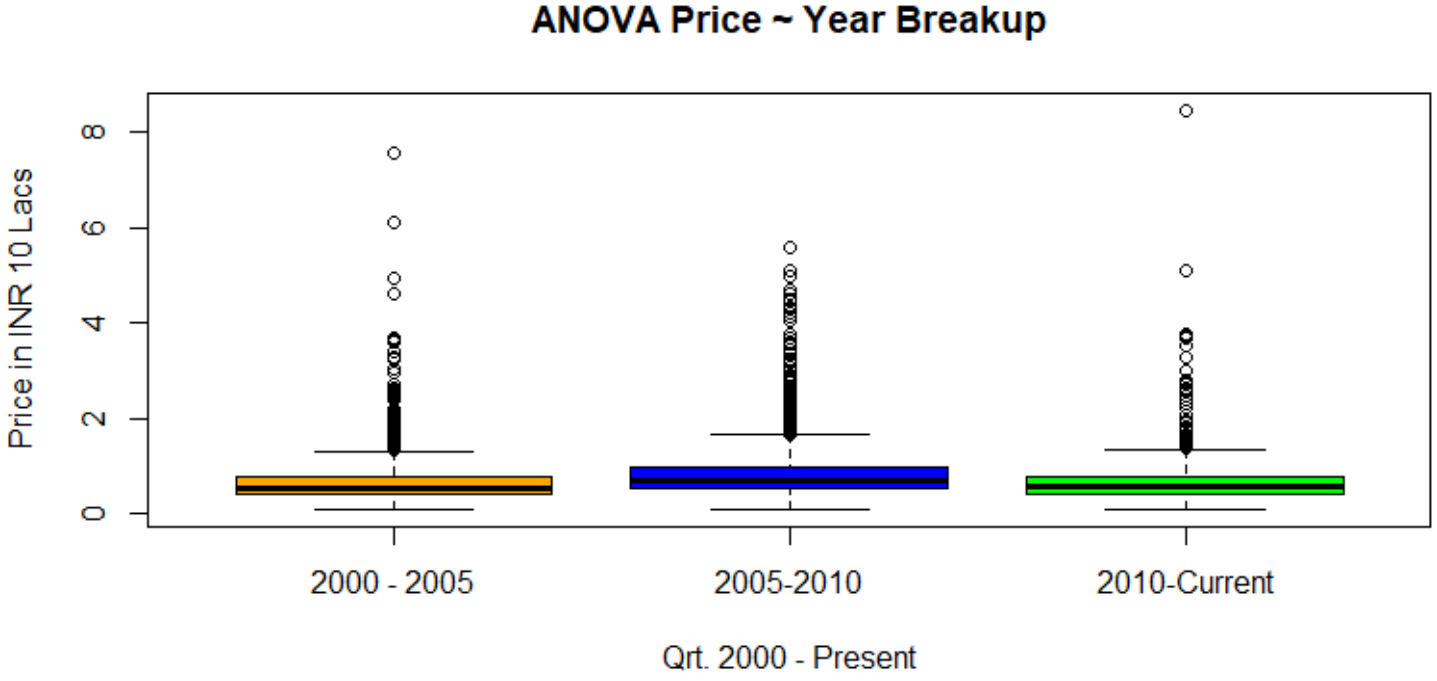
```
# Find the latest Year built
max(priceByDecade$Decade)
```

[1] 2008

```
# Find the Distribution by Year
hist(priceByDecade$Decade,
     bins = 10,
     main = 'Distribution of Houses By Year',
     xlab = 'Year Built',
     ylab = 'Count',
     col = 'orange'
     )
```




```
# Create a Break every 25 years
for(i in 1:5000){
        if (priceByDecade$Decade[i] < 2005){
                priceByDecade$Decade[i] <- '2000 - 2005'
        }
        else if (priceByDecade$Decade[i] >2005 && priceByDecade$Decade[i] < 2010){
                priceByDecade$Decade[i] <- '2005-2010'
        }
        else{
                priceByDecade$Decade[i] <- '2010-Current'
        }
}

# Most of the projects are old than 2010
# Make Sure each year is
priceByDecade$Decade <- as.factor(priceByDecade$Decade)

# Creating Analysis of variance (ANOVA)
# ANOVA is a collection of statistical models and their associated estimation procedures

anova <- aov(Price ~ Decade, data = priceByDecade)

summary(anova)

TukeyHSD(anova)

plot(pricesIn10Lacs ~ Decade,
     data = priceByDecade,
     main = 'ANOVA Price ~ Year Breakup',
     xlab = 'Qrt. 2000 - Present',
     ylab = 'Price in INR 10 Lacs',
     col = c('Orange', 'Blue','Green','Yellow','Pink')
     )
```


ANOVA Price ~ Year Breakup
Price in INR 10 Lacs
0
2
4
6
8
2000 - 2005
2005-2010
2010-Current
Qrt. 2000 - Present