Anmol More (11915043)

Please use table of contents (PDF Bookmarks) for easy navigation

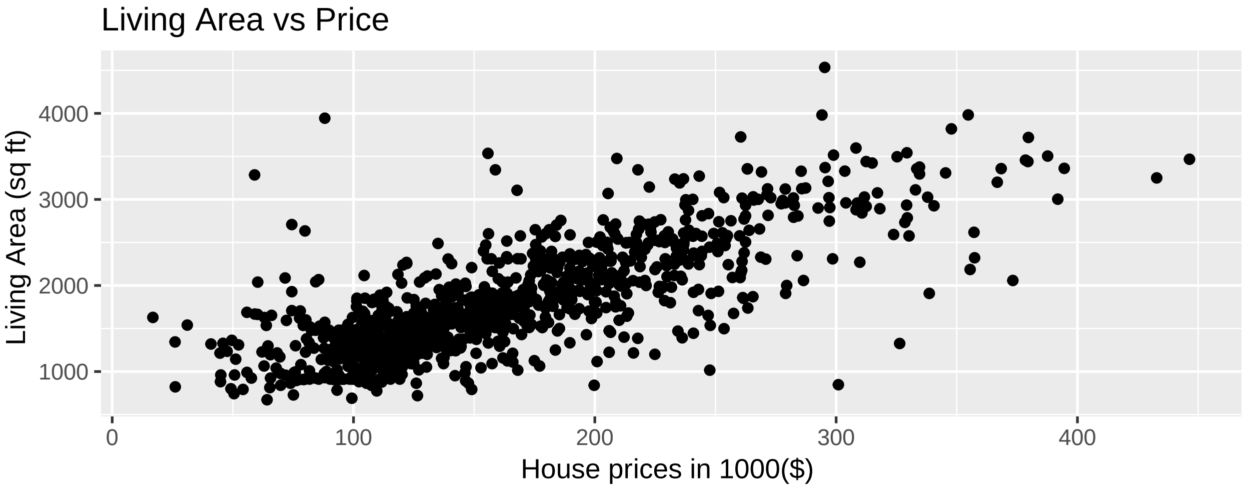
Part (A)

1. Report summarizing home prices

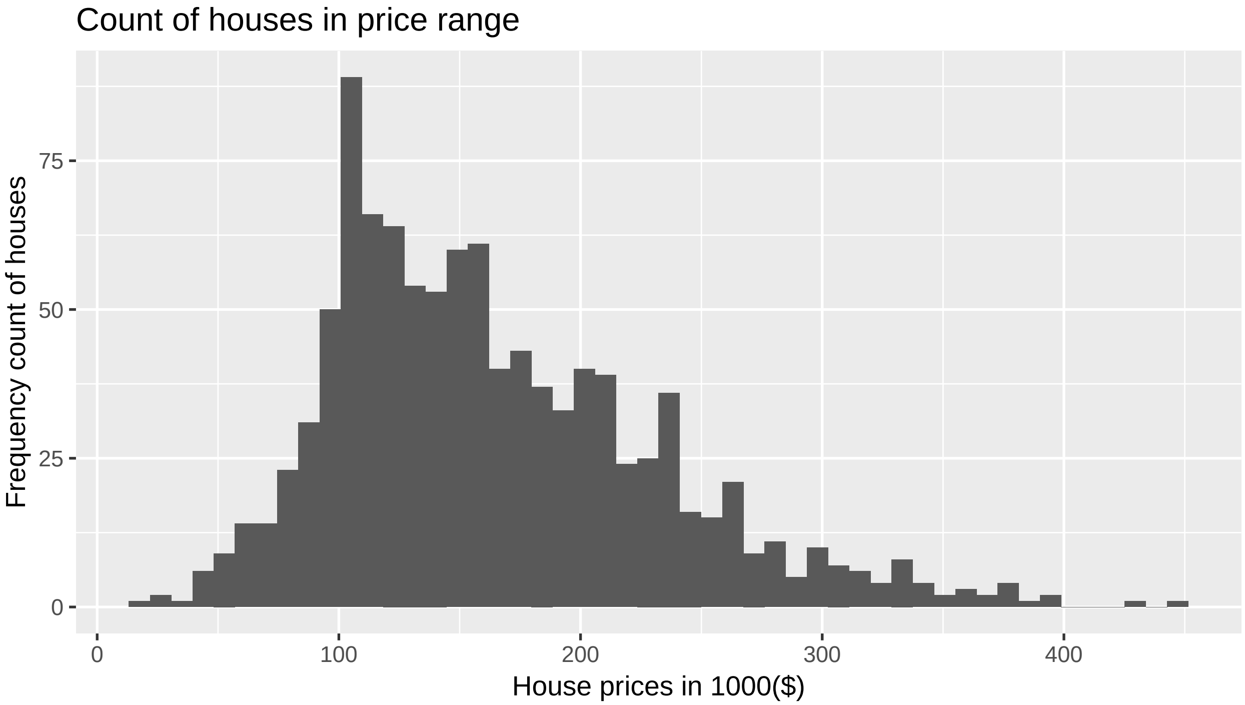
```
##      Price         Living.Area     Bathrooms        Bedroooms
##  Min.   : 16858   Min.   : 672   Min.   :1.000   Min.   :1.000
##  1st Qu.:112014   1st Qu.:1336   1st Qu.:1.500   1st Qu.:3.000
##  Median :151917   Median :1672   Median :2.000   Median :3.000
##  Mean   :163862   Mean   :1807   Mean   :1.918   Mean   :3.183
##  3rd Qu.:205235   3rd Qu.:2206   3rd Qu.:2.500   3rd Qu.:4.000
##  Max.   :446436   Max.   :4534   Max.   :4.500   Max.   :6.000
##     Lot.Size           Age           Fireplace
##  Min.   :0.0000   Min.   :  0.00   Min.   :0.0000
##  1st Qu.:0.2100   1st Qu.:  6.00   1st Qu.:0.0000
##  Median :0.3900   Median : 18.00   Median :1.0000
##  Mean   :0.5696   Mean   : 28.06   Mean   :0.5931
##  3rd Qu.:0.6000   3rd Qu.: 34.00   3rd Qu.:1.0000
##  Max.   :9.0000   Max.   :247.00   Max.   :1.0000
```

Describing 5 point summary from above

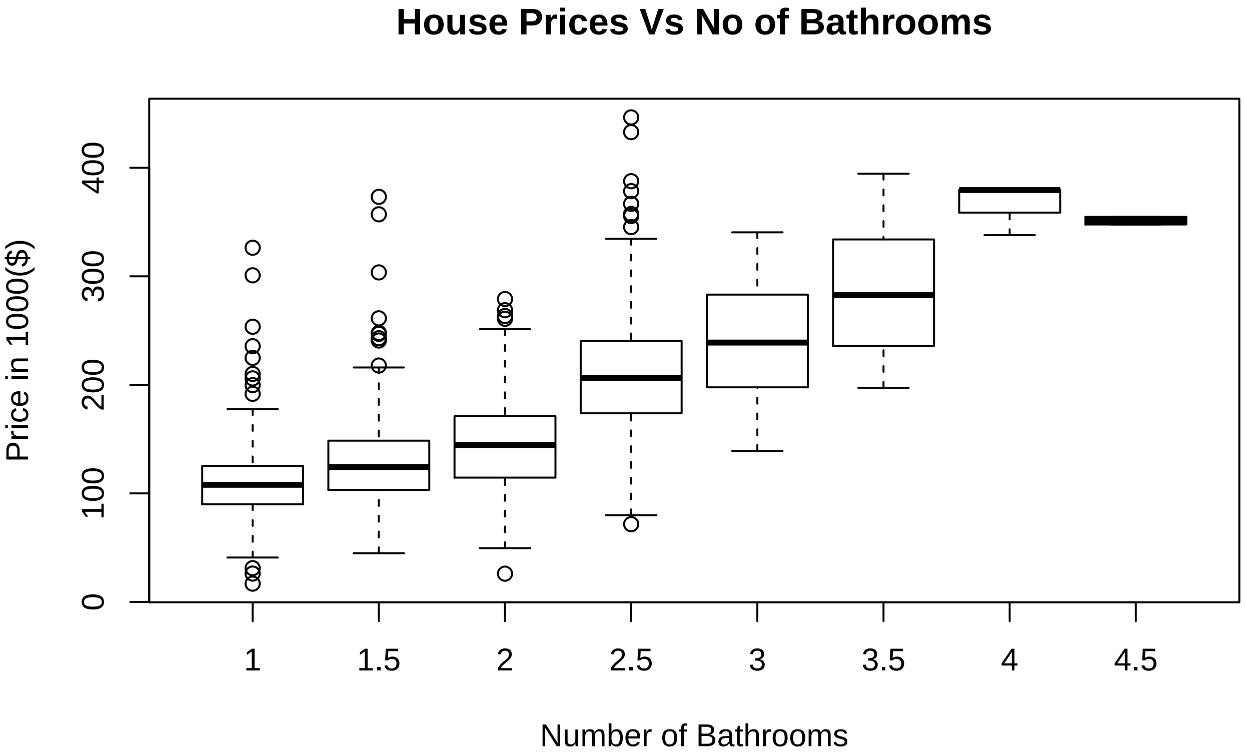
1. House Prices have mean of 16K but median is around 15K, so positively skewed data.
2. No of bedrooms vary from 1 to 6 and no of bathrooms are varying from 1 to 4.5
3. Age of houses is highly skewed towards right, as mean and median are far away. Also we have outliers where age is 247 or even 0
4. Around 40% of houses doesn't have Fireplace
5. Lot size and living area doesn't directly relates. With twice or thrice increase in lot size, Living Area tends to be increase by small margin. This to some extent points that bigger homes might have more open area
6. As seen in scatter plot below, Prices tend to increase with increase in Living Area



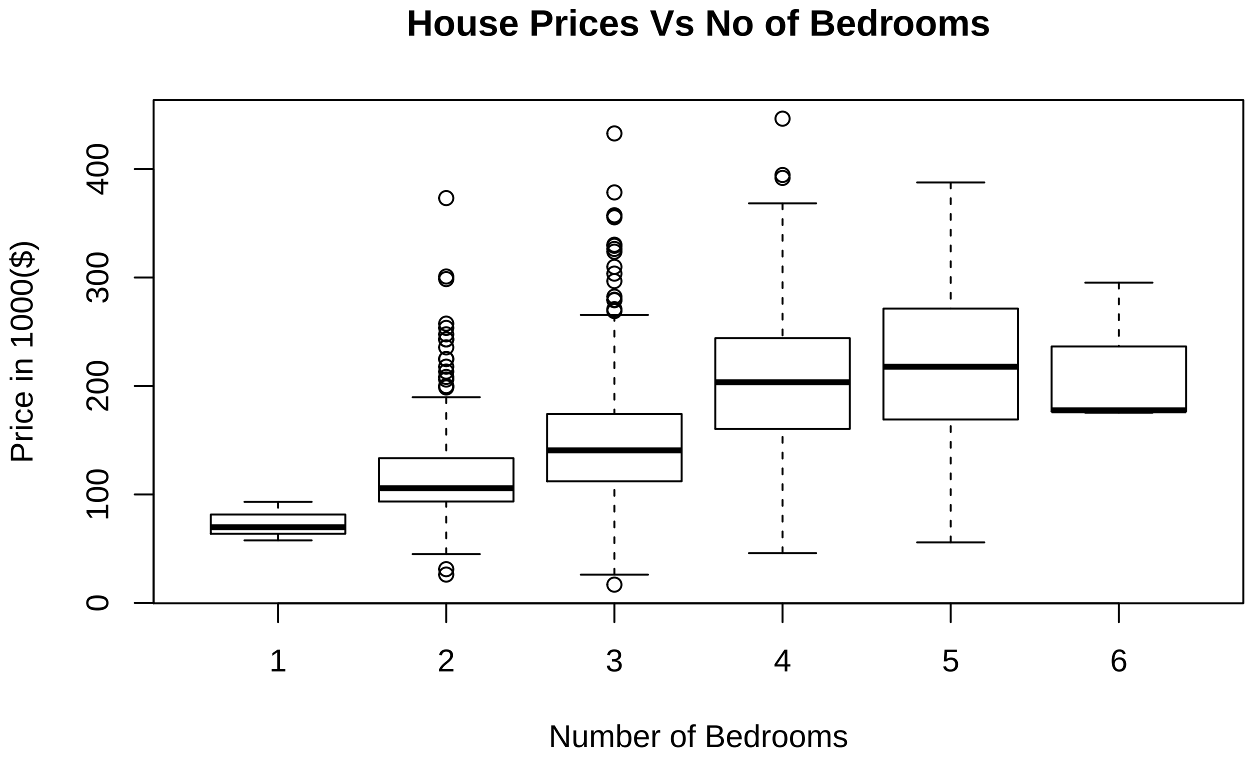
Histogram shows house prices are normally distributed in our sample, with bit rightly skewed



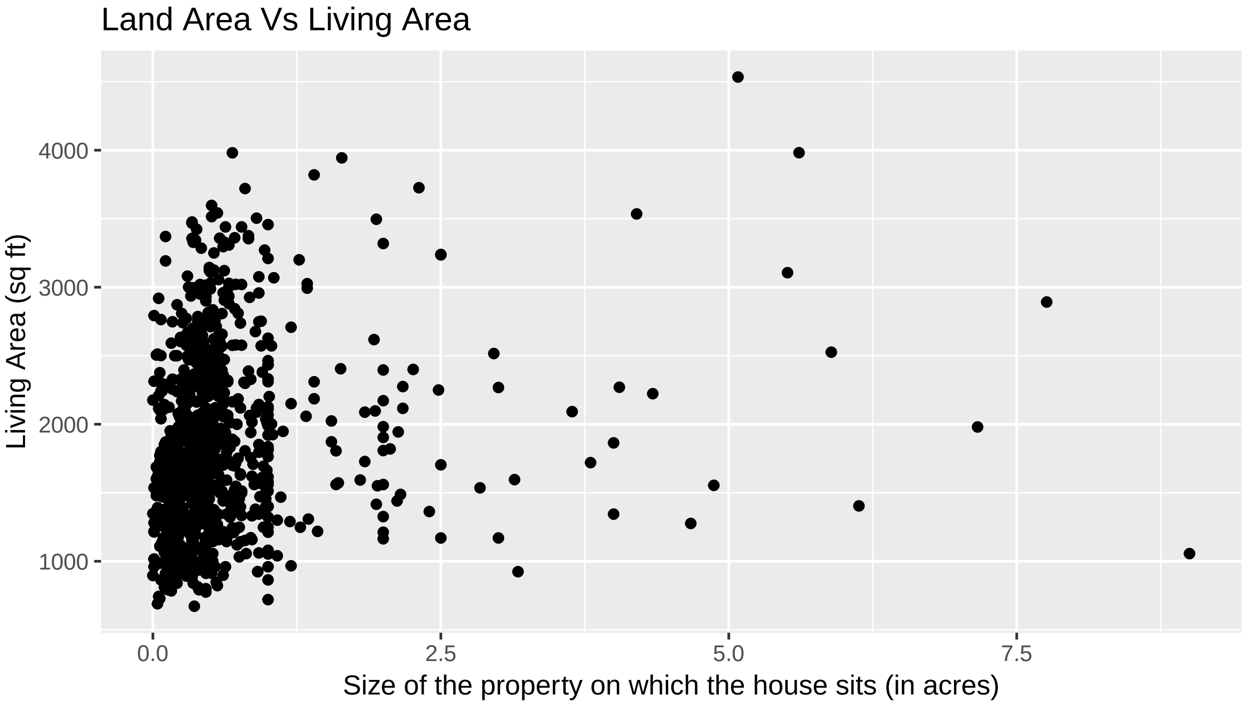
No of Bathrooms have direct affect on Prices, except for case of houses with 4.5 bathrooms



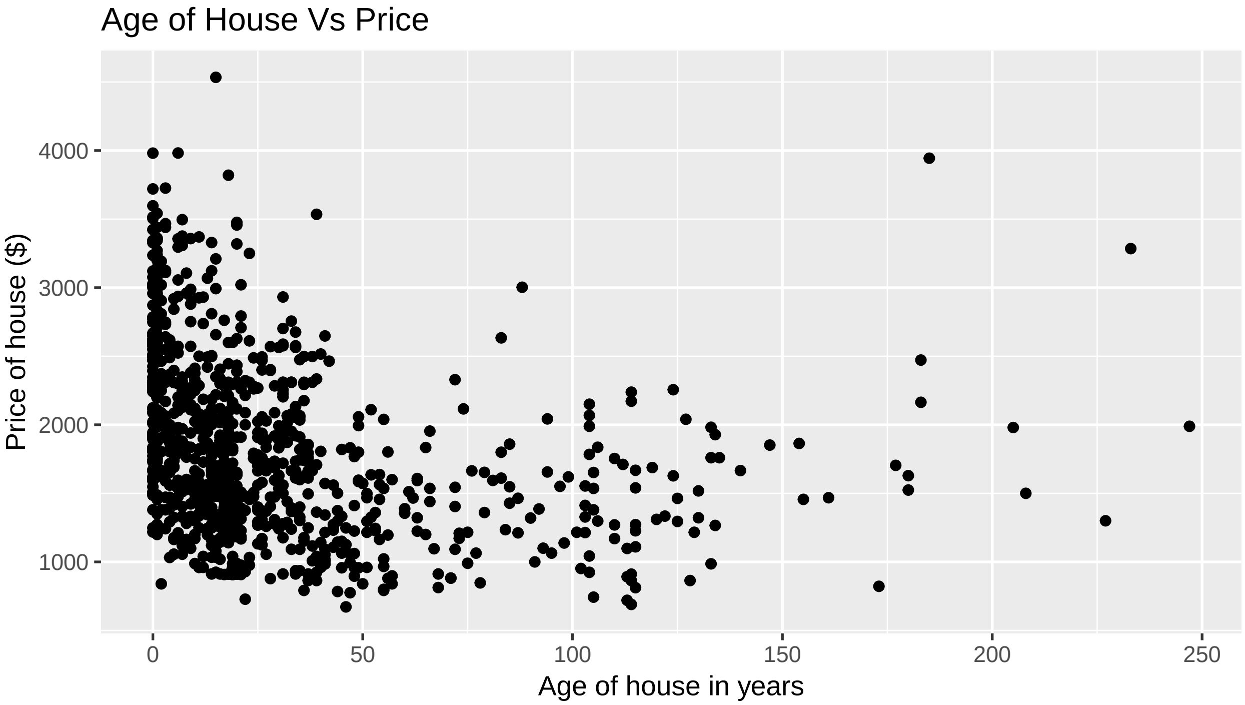
However increase in No of Bedrooms beyond 3 doesn't directly fetch more price



People tend to prefer more living area. But, lot area doesn't have linear relation with living area as seen in plot below.



New constructions fetch comparateively more price and prices tend of decrease with Age of House, however property area & Price doesn't have a linear relationship

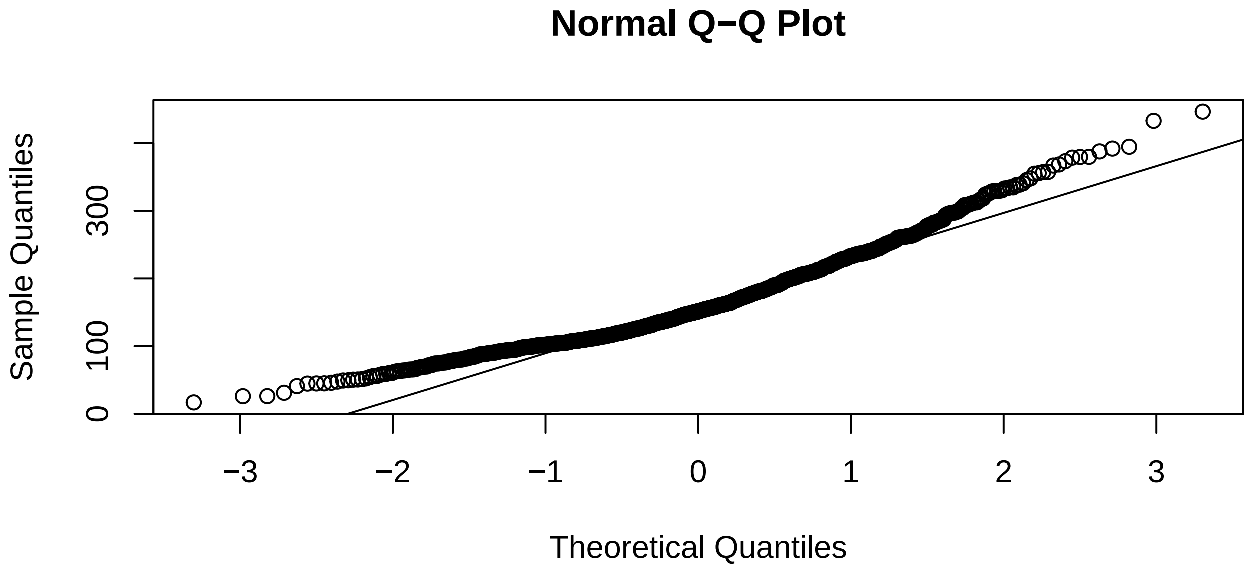


2. Does the normal model provide a good description of the prices? Use a Normal Quantile plot to frame your response.

Yes, normal model provides good description, as prices are normally distributed with bit right skewness. As seen in Normal quantile plot below, we can say - prices are normally distributed, with homes in range 100-300K $ doesn't show much deviation from diagonal line. Houses with prices below 100k or above 300K $ tend of vary on many other factors and that's why we see lower and upper quartiles are deviating from normal

Normal quantile plot for house prices

```r
qqnorm(house_prices$Price/1000);qqline(house_prices$Price/1000, main="Fig : Normal Quantile Plot")
```



3. Irrespective of your response to Q2, assume that Price ~ N(164K, (68K)^2). Given this

A. Calculate the following probabilities – P(Price > 92.8K), P(Price < 255.5K). Do these numbers agree with what you see in the data?

P(Price > 92.8K) = 1 - P(Z < (92.8 - 164)/68) = 1 - P(Z < -1.05) = 1 - .147 = .853

P(Price < 255.5K) = P(Z < (255.5 - 164)/68) = P(Z < 1.35) = .911

```
print(quantile(house_prices$Price, (1-.853)))
```

```
##    14.7%
## 101139.2
```

```
print(quantile(house_prices$Price, .911))
```

```
##    91.1%
## 261275.6
```

Theoritical assumption of normal distribution, doesn't completely matches data.

85.3% of houses have price > 101K and 91.1% have price < 261K in our sample data, as calculated below

B. Once again, assuming the above normal distribution, what percentage of houses should have a value less than 232K? Does that agree with the data?

P(Price < 232K) = P(Z < (232 - 164)/68) = P(Z < 1) = 84.13 % Calculating quantile from sample data:

```
quantile(house_prices$Price, .8413)
```

```
## 84.13%
## 232934
```

Yes, it matches as data approximately. We can see sample data quantile calculation have roughly 84.13% houses less than 232K.

C. Based on the theoretical model, what do you expect should be the price of a house that is exactly on the 3rd quartile (75th percentile).How does that compare to the actual?

Assuming normal distribution given in question, Let x be 3rd quartile range P(Price < x) = P(Z < (x-164)/68) = 0.75 (x - 164)/68 = 0.6744 x = 209.86K

```
quantile(house_prices$Price, .75)
```

```
##    75%
## 205235
```

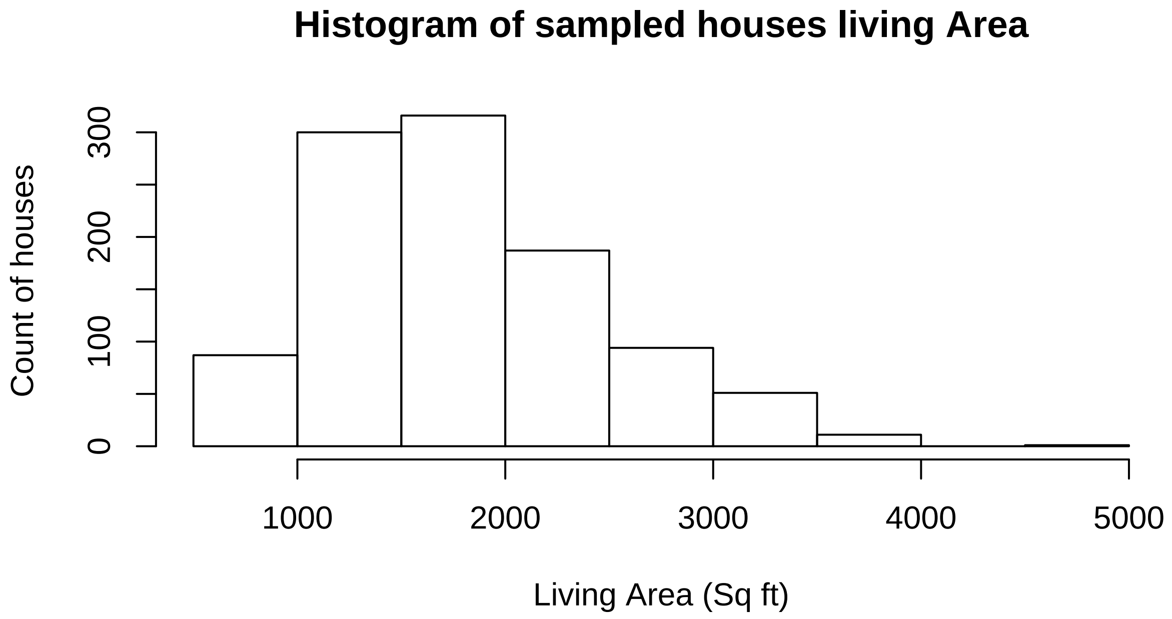
Against the actual data the third quartile varies by approximately 4K $

4. Create a histogram and boxplot for the Living Area variable. What does the histogram tell you that the boxplot does not, and vice-versa? Is the distribution symmetric? Check the skewness measure to see if it is consistent with your observation.

Histogram gives the frequency count for houses having living areas in each of the bins, however it doesn't tells anything about mean, but looking at skewness of histogram that can be imputed. It does not says what percentage of the houses lie in outlier range.

Boxplot clearly tells, which houses can be considered in outlier range and what's the median or min and max values

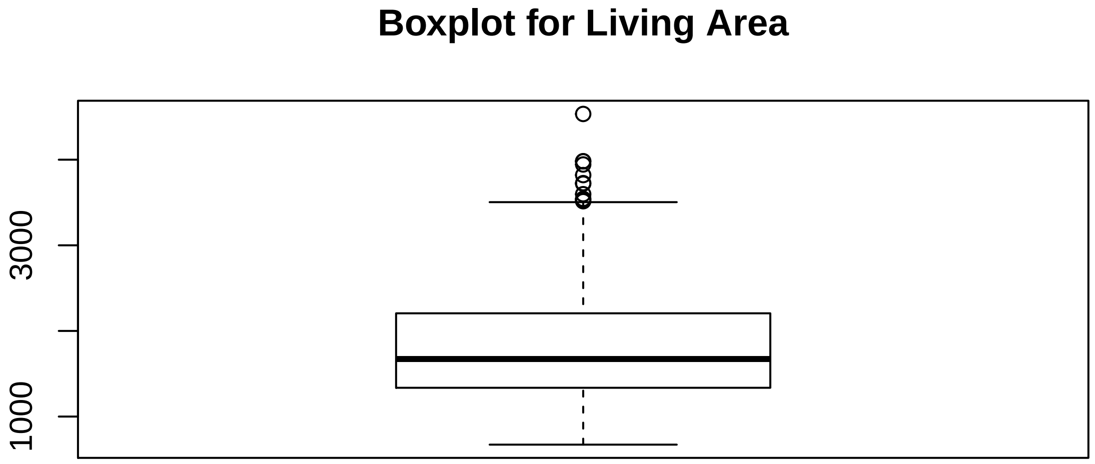
```
hist(house_prices$Living.Area, xlab = "Living Area (Sq ft)", ylab = "Count of houses",
     main = "Histogram of sampled houses living Area")
```



```
print(skewness(house_prices$Living.Area))
```

```
## [1] 0.8066481
```

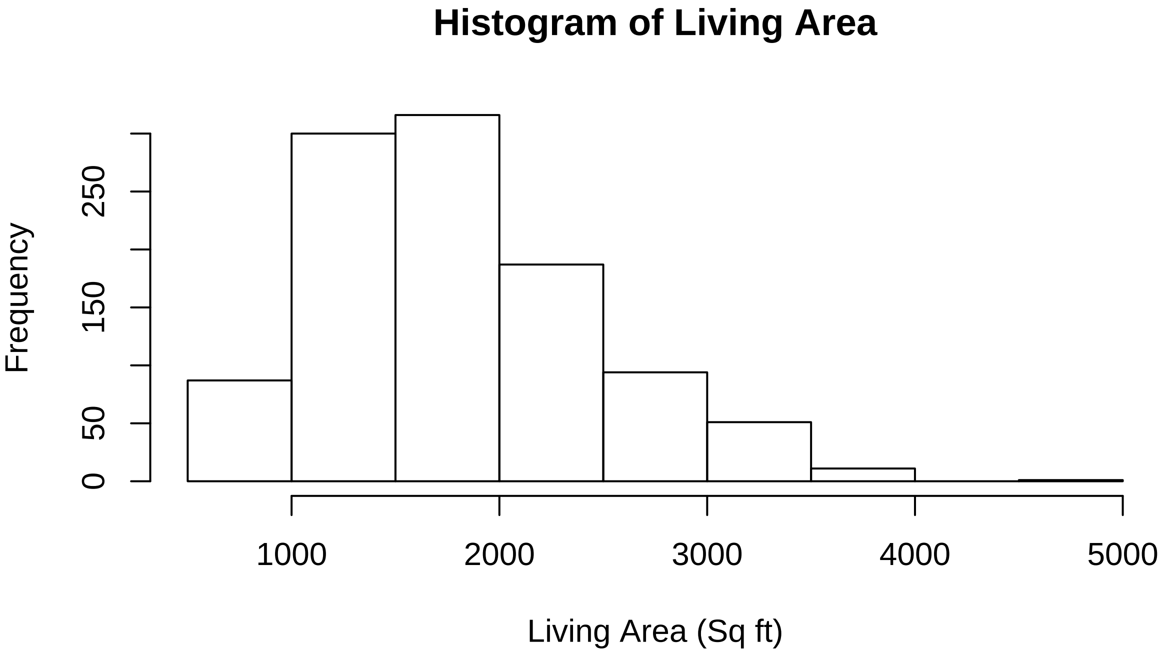
```
boxplot(house_prices$Living.Area, main= "Boxplot for Living Area")
```



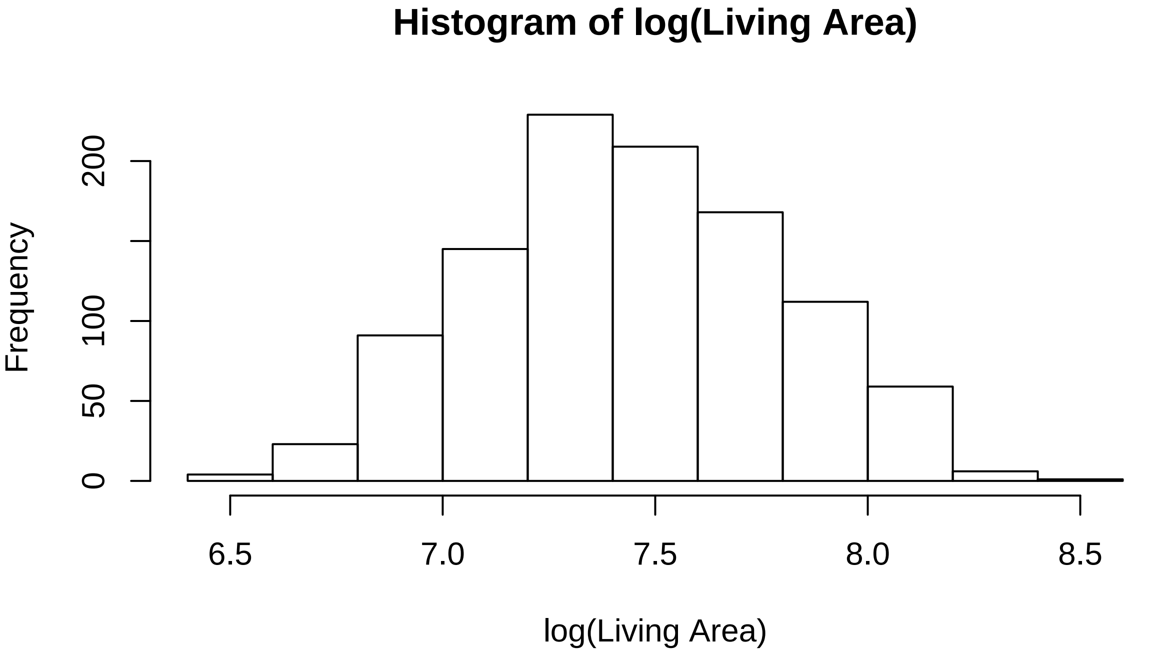
5. Create a new column in the dataset by taking the logarithm of the Living Area variable. Is the normal distribution a better fit for this variable or the original (Living Area) variable? Why do you think this is the case?

Normal distribution is better fit for log(Living Area) variable as seen in graphs below. A natural log graph for y = log(x) flattens out for higher values of x. In this case, lot of values of living area are on higher side, so log(Living Area) flattens it. So, log(Living Area) takes care of skewdness on right side

```
house_prices$Log.Living.Area = log(house_prices$Living.Area)
hist(house_prices$Living.Area, main="Histogram of Living Area", xlab = "Living Area (Sq ft)")
```



```
hist(house_prices$Log.Living.Area, main="Histogram of log(Living Area)", xlab = "log(Living Area)")
```



Part (B)

6. Create the 90%, 95%, and 99% confidence intervals for the average home price and explain what these mean. How do the margins of error for these three confidence intervals compare? Does that make sense? Before creating the confidence intervals, be sure to check the conditions necessary to create confidence intervals (and briefly describe this in your submission). Assume that the population standard deviation of the home prices is 64,000.

We know sample mean = mean(house_prices$Price) = 163.8621K

Also given, population SD = 64K

With these two values, we can derive CI for average (mean) home prices by Central Limit theorem

Calculating margin of error for each of intervals -

- 90% CI - [160618.4, 167105.9]

We have 90% confidence in method that mean of house prices in newyork will lie in range of 160.6184K to 167.1059K

- 95% CI - [159985.4, 167738.8]

We have 95% confidence in method that mean of house prices in newyork will lie in range of 159.9854K to 167.7388K

- 99% CI - [158759.1, 168965.1]

We have 99% confidence in method that mean of house prices in newyork will lie in range of 158.7591K to 168.9651K

With increase in CI, our range for population mean estimate keeps on increasing. More confident we want to be about average of home prices in a range, bigger range we will provide

```r
sample_mean <- mean(house_prices$Price)
population_sd <- 64000
sample_size <- nrow(house_prices)

calculate.ci <- function(ci_percentage) {
  z_value = abs(round(qnorm((1- (ci_percentage/100))/2),2))
  margin_of_error <- z_value*population_sd/sqrt(sample_size)
  range_of_mean <- list((sample_mean - margin_of_error), (sample_mean + margin_of_error))
  return(range_of_mean)
}
print(calculate.ci(90))
```

```
## [[1]]
## [1] 160618.4
##
## [[2]]
## [1] 167105.9
```

```r
print(calculate.ci(95))
```

```
## [[1]]
## [1] 159985.4
##
## [[2]]
## [1] 167738.8
```

```
print(calculate.ci(99))
```

```
## [[1]]
## [1] 158759.1
##
## [[2]]
## [1] 168965.1
```

7. Your friend has asked you to provide an estimate for the 95th percentile of home prices in this market. Which (if any) of the above confidence intervals can you use to give an answer? Describe briefly.

Approach 1

CLT only talks about calculating population mean from sample mean. From CIs calculated above, we cannot infer range of 95th percentile of home prices in market

Also, another intuition which invalidates CLT for percentile calculation is -

If we just use min and max of mean range calculated above and get 95th percentile assuming sample is normally distributed, it comes as - [265K, 273K]. This is no where close to sample 95th percentile 295K.

So, this method looks incorrect !

```
print(qnorm(.95)*64000+159985)
```

```
## [1] 265255.6
```

```
print(qnorm(.95)*64000+167738)
```

```
## [1] 273008.6
```

```
print(quantile(house_prices$Price, .95))
```

```
##      95%
## 295378.6
```

Approach 2

Going by references -

https://www.r-bloggers.com/r-and-tolerance-intervals/

https://stats.stackexchange.com/questions/44743/tolerance-bound-for-a-normalized-variable

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Intervals_for_a_Percentile_of_a_Normal_Distribution.pdf

It looks like tolerance bound could be a way with an assumption that house prices of population is normally distributed. Using normtol.int from tolerance package, This value comes in range of [26K, 301K] for 95%CI that 95th percentile of home will lie in this range. **However range being so wide, usability of this approach is also little**

```
normtol.int(x = house_prices$Price, alpha = 0.05, P = 0.95, side = 2)
```

```
##   alpha    P    x.bar 2-sided.lower 2-sided.upper
## 1  0.05 0.95 163862.1      26239.24        301485
```

8. The sample data given to you all come from home sales within the past 12 months. Suppose you had sample data of the same size each year going back several years, and calculated the average sale price for each year. What kind of distribution do you expect to see for these averages and why? (Include the parameters of the distribution in your response, assuming that the house prices don't change i.e. go up or down, overtime. Clearly, this is not a great assumption but make it anyway.)

This question is talking about taking samples over each year and calculating average sale price, additionally ignoring time value of house prices. Which is basically "sample mean of each of samples" more correctly named as - "sampling distribution of sample mean"

This is idea of derivation of CLT itself, so we can say -

1. Mean of all such samples over the years will follow a normal distribution with mean (meu) and standard deviation ((sigma/ sqrt(n))). n <- sample size (1047 in our case)
2. Even if our whole population is not normally distributed, our sample means will be still normally distributed. As our sample size is large, 1047 > 30
3. Even if population SD is not known, we should be able to calculate population mean with our sample because sample size is large enough (>1000) for consideration of t distributions as well

Why -

CLT tells us "When sampling is done from a population with mean (meu) and finite standard deivation (sigma), the sampling distribution of sample mean X(bar) will tend to a normal distribution with mean (meu) and standard deviation (sigma/ sqrt(n)) as sample size n becomes large" - Ref 5-5 "Complete Business Statistics"

Ref - https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-mean/v/sampling-distribution-of-the-sample-mean