CBA – Statistical Analysis 1: Estimation & Testing

Assignment 1: Solutions ***Part A***

1. Prepare a brief report summarizing the home values (prices) in this area. Use both graphical and numerical summaries. Your report should briefly describe what those summaries tell you, and anything of particular note/interest.

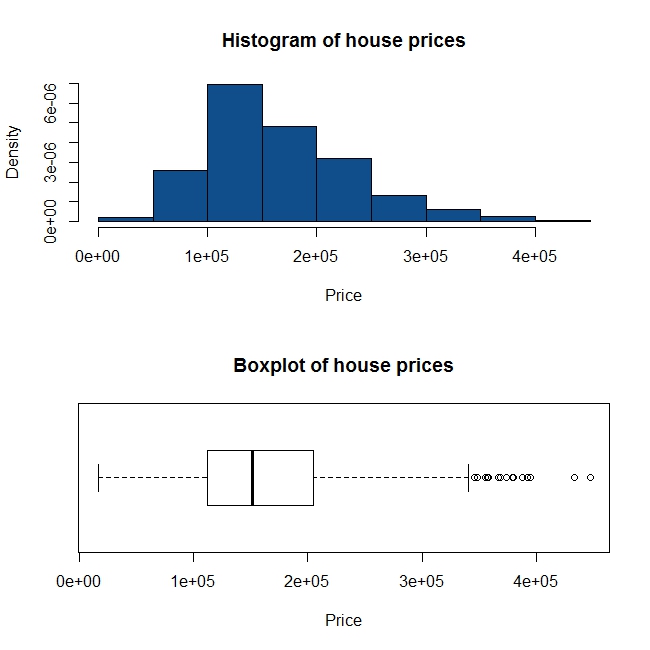
***Solution:***

* To get a preliminary understanding of the *House Prices* in our data, let us first take a look at the basic summary.



(R rounds-off the *summary* values)

* The key observations from this table can be made about the range of values taken by the variable. The range is $429578. We are told that location does not have a bearing on the house values. Therefore this large value tells us that either the other factors affecting the house prices have large variation in their magnitude or that they are affecting the price variable unevenly.
* The closeness of the median and the 1st & 3rd quartiles to the minimum value is indicating that the prices are more concentrated in the lower half of the data than the upper half. But from a home buyer’s perspective, it is good to note that the inter-quartile range in not very large ($93221). 50% of the house prices lie in an interval of this length.
* The value of Standard Deviation (67651.56) indicates that there is high deviation in the values from their mean.
* To try and observe each of these two phenomena, let us take a look at some graphical measures.
* Let us take a look at the histogram and the boxplot of the *Price* variable. This will help us get a better idea about the distribution of the values.



* The right-skewed histogram is supporting our initial proposal that there are more houses with prices closer to the minimum value. Or in other words, there are more values below the mean than those above the mean. There must be few houses with very high prices that are stretching the histogram in this way.
* The boxplot tells us the same story! In addition to the skewness, the boxplot is indicating that approximately 10 values are potential outliers i.e. values higher than (3rd quartile + 1.5\*IQR). We must investigate them further, to understand the factors which are contributing to such sharp increase in their prices as compared to other houses.

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

***Solution:***

We saw from the histogram and the boxplot that this variable is skewed. But to validate this, let us plot a Normal Quantile plot of the *Price* variable.



Based on the Q-Q plot we see that normal model would not describe the distribution of values of this variable very well. There are not many values that are running on the red line or lie between the 95% CI dotted-red lines. To support this, let us take a look at the skewness and kurtosis of the variable.

skewness(Price)

0.8749042

kurtosis(Price)

3.750459

The skewness and kurtosis values support our belief that this variable not very close to being normally distributed and it is right-skewed.

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

Given this:

1. Calculate the following probabilities – P(Price > 92.8K), P(Price < 255.5K). Do these numbers agree with what you see in the data?
2. Once again, assuming the above normal distribution, what percentage of houses should have a value less than 232K? Does that agree with the data?
3. 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?

***Solution:***

A(i)

1-pnorm(92.8,164,68)

0.8524638

quantile(Price,pnorm(92.8,164,68))

14.75362%

101204 = 101.2K

Assuming that Price ~ N(164K, (68K)^2), P(Price > 92.8K) = 0.85

But in our data, P(Price > 101.2K) = 0.85

The difference is almost $10K. Therefore we cannot say that the data does not agree well to the assumed distribution.

A(ii)

pnorm(255.5,164,68)

0.9107823

quantile(Price,pnorm(255.5,164,68))

91.07823%

261214.1 = 261.2K

Assuming that Price ~ N(164K, (68K)^2), P(Price < 255.5K) = 0.91

But in our data, P(Price < 261.2K) = 0.91

The difference is almost $6K. Therefore we cannot say that the data does not agree well to the assumed distribution.

B.

pnorm(232,164,68)

0.8413447

quantile(Price,pnorm(232,164,68))

84.13447%

232936.6 = 232.9K

Assuming that Price ~ N(164K, (68K)^2), approximately 84% houses should have a value less than $232K. And in our data, values of 84% of the houses lie below 232.9K

Alternatively, 878 out of 1047 i.e. approximately 83.86% houses have values lower than 232K.

Therefore we might be able to say that this approximation using the assumed distribution agrees well with our data.

C.

qnorm(0.75,164000,68000)

209865.3

quantile(Price,0.75)

75%

205235

If we assume the proposed theoretical model which says Price ~ N(164K, (68K)^2), the value of the house exactly on the 3rd quartile is 209.87K. In our actual data, the value of the house on the 3rd quartile is 205.24K

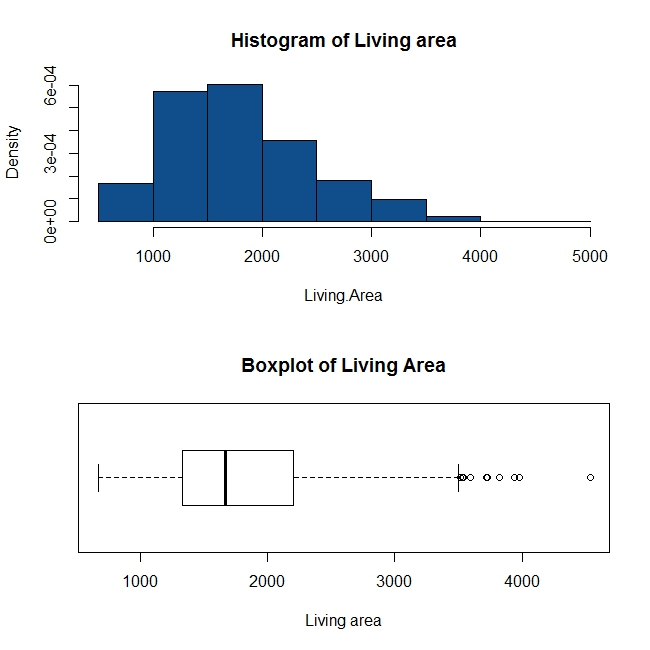
1. Create a histogram and boxplot for the Living Area variable. What does the histogram tell you that the boxplot does not, and vice-versa? Is the distribution symmetric? Check the skewness measure to see if it is consistent with your observation.

***Solution:***

* Looking at the histogram and the boxplot below, we can observe the distribution of values of the *Living Area* variable. The histogram tells us that the distribution is almost bimodal; which the boxplot cannot portray in any way. On the other hand, the boxplot is giving us information about the potential outliers (values beyond the right whisker). The histogram is indicative of the fact that there are extreme values on the right side but it cannot tell us what these values would exactly be.
* Both the plots together tell us that the distribution of *Living Area* is not symmetric.
* The skewness value of 0.8 supports this claim.

skewness(Living.Area)

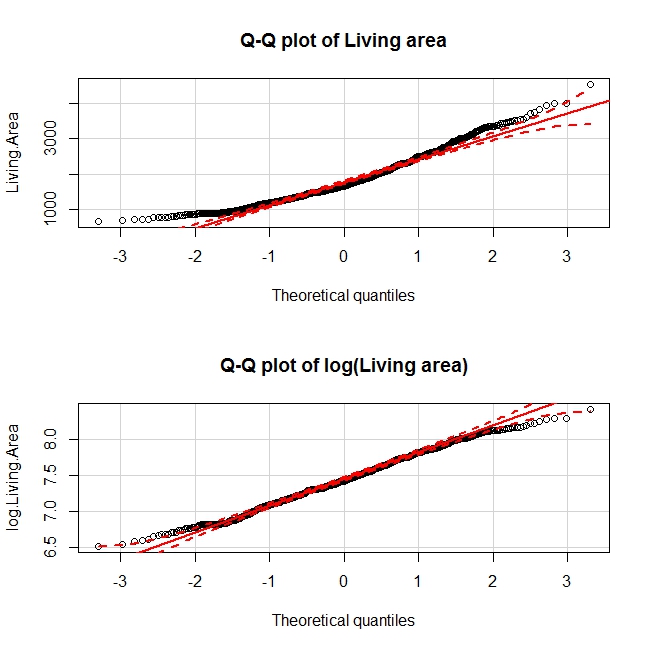
0.8066481



1. 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?

***Solution:***

log.Living.Area<-log(Living.Area)



The *log(Living.Area)* is certainly closer to a normal distribution than the same variable without taking a log. This is so because taking a log changes the scale on X-axis in such a way that the data gets squeezed. The very large (or very small) values become less extreme, bringing the shape of the distribution closer to a normal distribution. The skewness values below support our observation.

|  |  |
| --- | --- |
| skewness(log.Living.Area) | skewness(Living.Area) |
| 0.005214516 | 0.8066481 |
| kurtosis(log.Living.Area) | kurtosis(Living.Area) |
| 2.522292 | 3.385202 |

Assignment 1: Solutions ***Part B***

1. 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.

***Solution:***

The assumption we make before creating confidence intervals is that of a Normal population distribution or that of a large enough sample size to apply the CLT.

To check if the sample size is sufficient to apply the CLT:

Since a DMOE is not mentioned we can use thumb rules.

. The sample size is big enough to apply the CLT.

Since the population variance is known we use the z-distribution to obtain CIs for the population mean. The CI is as follows:

1. **90% CI is (160618.6,167105.6)**, where

Sample mean=163862.1

This means that we have 90% confidence that the population mean will lie in the range **(160618.6, 167105.6) for a random sample**.

Similarly,

1. **95% CI is (159985.7,167738.5),** where

Sample mean=163862.1

1. **99% CI is (158759.5,168964.7),** where

Sample mean=163862.1

We see that if % of C.I increases then margin of error increases and MOE is largest for the 99% CI and smallest for the 90% CI. This makes sense; as if we want a higher level of confidence we must settle for a larger margin of error.

1. 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.

***Solution:***

We have an estimated interval for the population mean, not for the entire data. There is no way to get an estimate of 95th percentile using these intervals. We can use a point estimate from the data.

1. 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, over time. Clearly, this is not a great assumption but make it anyway.)

***Solution:***

Averages follow Normal Distribution. We can use the CLT to say this. We have seen that the sample size is sufficient to apply the CLT.

Sampling distribution of the mean is normal. The parameters of this distribution are the (mean =population mean) and (standard deviation=population standard deviation/√n)