**Statistical Analysis (I): Estimation & Testing Homework Assignment 1Instructions**

1. This assignment uses the dataset *houseprices.csv,* which is available on the LMS under the folder *Datasets*.

2. The assignment is an individual submission. Please remember to include your name and ID at the top of the assignment.

3. Your report should not be more than 4 pages (sides) long. Please ensure that all figures and tables are included in this space.

4. **The submissions should be submitted 12 midnight on the due date via LMS** Instructions for uploading on Turnitin will be announced by your AAs separately. Late submissions will be penalized.

5. The homework submissions will not be graded for correctness, but rather for effort. The scoring will be binary for each question– you will get full credit if your effort is deemed satisfactory (all questions are responded and the thought process is clearly explained), otherwise you will get no credit.

6. You can refer to any resource within and outside of the course material but only with appropriate citations. You can also discuss with other students but the submitted write-up should be entirely your own work. Significant overlaps with other submissions and lack of citations maybe considered as possible instances of violation of the honor code.

7. The solutions for the homework assignments will be made available to you after

the due date.

**8.** Have fun!

**Description of the dataset:**

A real estate agent is trying to understand the nature of housing stock and home prices in and around a medium sized town in upstate New York. She has collected data from a random sample of 1047 homes sold in the last 12 months. Data was collected on the following variables, and is available in the attached *houseprices.csv* file.

*Price* – the sale price of the house in *$*

*Living Area* – in *Sq. ft*.

*Bathrooms* – number of bathrooms in the house (powder rooms with no tub or shower area are considered 0.5 baths)

*Bedrooms* – the number of bedrooms

*Lot Size* – size of the property on which the house sits (in acres).

*Age* – of the house in years

*Fireplace* – whether or not the house has a fireplace (Yes = 1, No = 0)

Your task in this assignment is to analyze this dataset in order to gain some understanding of this particular real estate market – the values of homes, their characteristics in terms of size and other features, and relationships between these. This understanding will prove immensely helpful to the real estate agent in advising her clients. Since all of the homes are from the same

geographical area, location (which usually has a huge bearing on home values) is not a major concern here.

Most of the analysis will be done in response to the specific questions posed on the homework assignments. But feel free to explore and play around with the data set to enhance your own understanding of how to make sense of data.

**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.

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

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?

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?

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?

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.

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?

**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.

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.

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.)