**STAT4610 Project 1**

**Descriptive Statistics and Data Analysis**

For this project you will use data from home sales in Delta County, Colorado from 1995 through 2012. You are interested in analyzing the residential real estate market across this time period for reasons unknown to you. You may access the Excel data file under “Files” on Canvas.

1. The first thing you notice is that the data set is a mess. You are only interested in homes that have sold, so start by removing all records where the sale price or the sale date are null. [HINT: the sort or filter functions might come in handy here.] How many observations does this leave you?

2. Even with the non-sales removed, there are still some problems with the data. Are there any “homes” that sold for “prices” that seem unreasonable? Are there any “homes” with unusual numbers of bedrooms or bathrooms? [HINT: you are interested in the market as it pertains to “usual” residential homes. However, just because the number of beds and baths is zero, it doesn’t imply that these are zero-bed, zero-bath homes.] Continue “cleaning” the data as you see fit. How many observations does this leave you?

3. Calculate the following descriptive statistics for the sale price across the entire range of the cleaned data: mean, median, mode, min, max, standard deviation, variance, range, sum and count. Show your results here. [HINT: You may find the “data analysis/descriptive statistics” function under the data tab to be useful here. In that case, you can just paste an Excel table here to complete this requirement.] Calculate the correlation between the number of bedrooms and bathrooms for all the observations where these are not void (or zero). Briefly describe what this means.

4. Draw a histogram of the sale prices, using bins in increments of $50,000. [HINT: You can do this with a pivot table, a lot of “countif” statements and a bar chart, or the histogram tool under data analysis/descriptive statistics (to perhaps guide you towards the simple solution).] Show your histogram here. Briefly describe the “distribution” (shape) of this variable based on the histogram and the descriptive statistics above. How many outliers are in the data (mild outliers are >2σ from µ, strong outliers are >3σ from µ)? Quantify the outliers in terms of their distance from the mean.

5. Draw a scatterplot of sale prices versus the year sold. [HINT: the “year” command in Excel might be useful here, when applied to the sale\_date column.] Insert a trendline on this chart. [HINT: right-click on the data to get started.] Is there a trend? If so, describe it briefly.

6. Starting with January 1995 as month 1, calculate the months of sale for all of the homes that sold in this timeframe. [HINT: the “month” command works the same way as the year—you just have to do some math.] Draw time series charts (line charts) of the following across all 211 months: Monthly count of homes sold, Average monthly price of homes sold, and Monthly standard deviation of price of homes sold. Include those charts below, and briefly describe what you observe in each one. You might want to note what happened in 2008 to the US housing market. [HINT: once you’ve made your monthly calculations, you might want to use a pivot table to create the data for the line charts.]

7. You are interested in the volume and average price of home sales by city across the 18 years of data. Unfortunately, you aren’t sure if this can be easily captured in a single graph, so you decide to build a dynamic graph that allows you to animate this process.

a. Build a table where the rows are the cities and the columns are the years, and where the entries in the table are the numbers of homes sold by city for each year. Build a second table where the entries are the average home prices for each city for each year. [HINT: a pivot table might come in handy here. When you extract the values from the pivot table, paste them as “values” so you don’t keep the pivot table formatting.] You can most likely remove two cities at this point (Cory and Lazear) by putting a filter on your pivot table (or by deleting them from the completed table).

b. Move these tables to a clean worksheet that will be turned in with this word doc. (This way you won’t be submitting a 5M file for this assignment.)

c. Since we know that all these cities are in Colorado, remove the state from the city name. [HINT: there are many ways to do this—“text to columns” under data analysis might be the easiest, but leave yourself a column for the state so you don’t overwrite any useful data.]

d. Set up two columns of data as the basis for a scatterplot, where the row labels are the city names, the first column is the count and the second column is the average price. Set up a counter cell (that you can change manually) that can designate the appropriate column in your count and average price tables. Fill in these columns from the two tables using the “vlookup” command. [HINT: don’t forget the “$” for cell references when you need them.] [HINT: if you’re not familiar with the “f4” function key, now might be a good time to get to know it.] Experiment with manually changing the “counter” cell to see the two data columns change.

e. Draw a scatterplot based on the two lookup columns of data. Fix the axes so that the graph can accommodate the entire range of both variables without any adjustment. Label each point with its city name. [HINT: right-click again.] Create a cell for the year, and have that cell display in the chart title. Again, manually change the counter cell to make this work.

f. From the “Developer” tab (you might have to install this through File/Options) insert a scrollbar (from the Form Controls.) Set the scrollbar to step through the values in your control cell (corresponding to column numbers in your table) and point the scrollbar output to the control cell. Your animation should now run by clicking on the scrollbar. Format the graph as you see fit, but ensure you have axis labels and a title. Submit this Excel file in addition to this completed word doc through Canvas prior to lesson 3.