Practice Problems: Exam 1

1. Consider the diamonds data set, available on the class website.
   (a) Is the data set structured or unstructured? Why? The data set is structured because it fits into one or more spreadsheet-like tables, each with rows and columns, and no additional dimensions.
   (b) What are the cases? Each diamond is a case of the diamonds data set.
   (c) What are the variables? The variables of the diamonds data set follow: price, carat, cut, color, clarity, x, y, z, depth, and table, as described in the code book.
   (d) What kind of variable is each variable? Be specific! price, carat, x, y, z, depth, and table are quantitative variables. Cut, color, and clarity are ordinal categorical variables, as indicated by the order specified in the code book.
   (e) What kind of graph(s) would you use to visualize each kind of variable? Histograms would be appropriate for the quantitative variables. (There are too many cases to make a stemplot.) Bar plots and pie charts would be appropriate for the ordinal categorical variables, although a bar chart may more effectively convey the order.
   (f) If you wanted to use StatCrunch to visualize the distribution of the cut variable with a pie chart, would you choose "with data," or "with summary?" How would the data need to be structured to compel the other choice? You would select "with data," because you have the categories in a column for each of the tens of thousands of diamonds, in rows, in the data set. You would instead select "with summary" if your data set only had a single row for each category (in this case, five rows: Fair, Good, Very Good, Premium, Ideal), together with these five category names in one column and the respective counts of the number of diamonds in each category, in a second column.
   (g) Are there any identifier variables in the diamonds data set? There is no label in the diamonds data set. However, StatCrunch numbers the rows so the row labels could be used as an identifier variable.

2. What feature of a QQ plot (normal probability plot) indicates that the distribution of the variable in question is approximately normal. If the QQ plot shows a straight line, that indicates the distribution of the variable in question is approximately normal. Note even in the ideal case (data are simulated from a normal model), the QQ plot will show deviations from a straight line especially near the tails. The more samples, the smaller the deviations, however.

3. Boxplots.
   (a) What feature of a boxplot indicates median? The line in the center of the "box" indicates the median.
   (b) What feature of a boxplot indicates Q1? The bottom of the box indicates the first quartile (Q1).
   (c) What feature of a boxplot indicates Q3? The top of the box indicates the third quartile (Q3).
   (d) What feature of a boxplot indicates the interquartile range (IQR)? The vertical extent of the box (Q1 to Q3) indicates the IQR.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

(e) What feature of a boxplot indicates the maximum? The highest identified outlier, or if there are no identified outliers, the top of the whiskers in the box and whisker plot. In other words the highest feature of the boxplot.

(f) How are the fences determined? Fences are determined by the 1.5 IQR rule. Basically the IQR defines a ruler. Stretch the rule by a factor of 1.5. Put the stretched ruler on top of Q3. That is the top fence. Put the stretched ruler beneath Q1. That is the bottom fence. Data points that lie outside the fences are identified as (possible) outliers.

4. Histograms.

   (a) What features of a histogram indicate that the distribution of the variable in question is symmetric? The histogram can be folded about the center and the tails will approximately line up. Randomness present in all distributions will make the "line up" only approximate, even in the ideal situation where data are simulated from a perfectly symmetric distribution.

   (b) What features of a histogram indicate that the distribution of the variable in question is skewed? Skewed left or right? Skewed means one tail is longer than the other. For left skewed the left tail is longer, and vice versa for right skewed. A more precise measure of skewness is the relative position of mean and median. If the mean is greater than the median, the distribution is right skewed, and vice versa for left skewed.

   (c) What features of a histogram indicate the distribution of the variable is unimodal? Bimodal? Multimodal? Uniform? Respectively, one peak, two peak, three or more peaks, no peaks (for uniform). No peaks mean that all bars of a histogram have roughly the same height, however randomness will prevent this property from being exact.

   (d) How can standardizing a variable (converting to z-scores) change the shape of a histogram? It can't. Only differences the the width or positions of the bins will alter the appearance of a histogram of a variable that has been standardized. All properties of the shape of a histogram, modes, gaps, outliers, etc, will stay the same.

5. Explain the difference between resistant and not resistant and give example statistics for each type. Resistant refers to a measure that is resistant to outliers, meaning large outliers don't change the measure much. Examples: Median, Q1, Q3, IQR, Percentiles. Not resistant is the opposite: examples: Mean, Standard Deviation, Min, Max, Range.

6. Give the name for explain the condition listed in the book for using a normal model to describe what values to expect from a variable. The Nearly Normal Condition. The distribution of the variable should be unimodal and symmetric and lack substantial outliers.

7. A standardized test is given to a large population of students across the USA. After an analysis of the data, researchers determine that a Normal model fits the scores well. The scores have a mean of 1300 points and a standard deviation of 50 points.

   (a) What percentile does a score of 1400 lie on? 97.7^{th}.

   (b) What z-score does a score of 1200 have? -2.