Practice Problems: Week 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) How many tables does the data set have? The data set has one table.
   (c) What are the cases? Each diamond is a case of the diamonds data set.
   (d) 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.
   (e) 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.
   (f) 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.
   (g) 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.
   (h) Are there any labels in the diamonds data set? There is no label in the diamonds data set. However, StatCrunch numbers the rows so the row numbers could be used as a label. This choice might cause complications if you ever decided to reorder the rows. (But you won't have a reason to reorder rows in this class.)

2. In the diamonds data set, you create a new variable called "ideal." You assign the value 1 to this new variable for every diamond in the data set with the ideal cut and you assign 0 to this new variable for every other diamond.
   (a) How could you more simply describe the sum of the values of the ideal variable in terms of statistics that are important for categorical variables? The sum of the values is the *count* of ideal diamonds in the data set.
   (b) How could you more simply describe the average (mean) of the values of the ideal variable in terms of statistics that are important for categorical variables? The mean of the values is the *proportion* of ideal diamonds in the data set.

3. What does the distribution of a variable tell us? Your answer should be the loose definition given in class. The distribution of a variable tells us (1) what values the variable takes and (2) how often it takes these values.

4. What graphs help us visualize the distribution of each kind of variable we studied (different answers for different kinds of variables)? When would you choose one type of graph over the

other? Stemplots and histograms help us visualize the distribution of quantitative variables. Stemplots are only appropriate when there are few cases. Histograms are appropriate for both few and many cases, although stemplots provide more information when there are few cases. Pie charts and bar plots are appropriate for categorical variables (both nominal and ordinal) although bar plots may be used to better convey the order of an ordinal variable (if that order is used for the bars). Pie charts are only appropriate for graphs that include all cases and categories within a distribution because the pie conveys the idea of a whole. On the other hand, categories with too few elements can be deleted from a bar plot.

5. Describe a bar plot with Pareto ordering. Why would such an ordering be desirable? A Pareto ordering of a bar plot places bars in order of descending height. Such an ordering would be desirable to draw attention to the relative number of elements in each category. Unfortunately, this ordering is usually different from the ordering of an ordinal variable, so a choice has to be made as to what features to highlight.

6. Name two things you can do with StatCrunch (or other statistical software) to better visualize a distribution of a quantitative variable with a histogram? (1) Eliminate outliers with the "where" selection, and (2) change the bin width to adjust the resolution of the bins.

7. Put the following data into stemplot. Trim and split stems, and include all data points. You may use StatCrunch, but you do not have to.

$$236, 375, 383, 412, 413, 426, 440, 458, 459, 465, 491, 507$$

2	3
2	
3	
3	78
4	1124
4	5569
5	0