Statistics and Data Analysis, Fall 2015
Homework 1: Descriptive Statistics

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

This homework is due ***6:35 pm, October 12, 2015***. Each student should submit her/his own ***hard copy*** to the instructor at the beginning of the class. All the data for this homework are contained in the file "SDA-Fa15_hw01_data.xlsx". Discuss with your classmates but NEVER copy one's work.

1. (15 points; 3 points each) Consider the 300 numbers contained in the sheet "P1". Let them be x_1, x_2, ..., and x_{300}.
   (a) Divide each of them by 5 and obtain the remainders. Let the 300 remainders be r_1, r_2, ..., and r_{300}. What is the sum of the 300 remainders, $\sum_{i=1}^{300} r_i$?
   (b) Find the average of those numbers in x_1, x_2, ..., and x_{300} that are no less than 500.
   (c) Find the number of even values in x_1, x_2, ..., and x_{300} (a value is an even value if it can be divided by 2).
   (d) For r_1, r_2, ..., and r_{300}, draw a bar chart with five classes, one for each distinct value.
   (e) For x_1, x_2, ..., and x_{300}, draw a histogram with ten classes $[0, 100)$, $[100, 200)$, ..., and $[900, 1000)$. Label the frequency of each class on your histogram.

2. (20 points; 5 points each) Consider the wholesale data contained in the sheet "Wholesale".
   (a) For each channel-region pair, find the amount of Frozen food sales to the buyer of the highest sales amount.
   (b) Among all buyers, find the 50 buyer who bought the most Frozen food. In other words, the Frozen food sales amounts with respect to these buyers are ranked at the top 50. Then draw two pie charts, one according to channel and one according to region, for the 50 buyers.
   Hint. Though not covered in lectures, the MS Excel operation "Sort" can be helpful.
   (c) For buyer i, $i = 1, ..., 440$, let y_i be the total amount of money it spent in the six categories. For example, for buyer 1 (in row 2), we have $y_1 = 65080$. Find the number of buyers whose y_i is at least 40000.
   (d) Continue from Part (c). For each channel-region pair, find the number of buyers whose y_i is at least 40000.

3. (20 points; 5 points each) Consider the daily bike rental data contained in the sheet "Bike".
   (a) Draw a histogram for the daily total rentals by choosing the number of classes by yourself. Then report something about the distribution you observe.
   (b) Draw a line chart for the 31 casual, registered, and total daily rentals in January, 2011. You need to put the three curves on one single chart.
   (c) For each of the 31 days in January, 2011, find the ratio of casual rentals to total rentals. For example, for January 1, 2011, the ratio is $\frac{331}{985} \approx 0.336$. Report any pattern of this curve that you observe.
   (d) Use all the data you have, try to find (at least) one factor that will help you predict the ratio of casual rentals to total rentals. Do some data processing and analysis to give us some numbers or figures to support the use of that factor (i.e., to show us an evidence that your factor indeed has a significant impact on that ratio). Then explain why the factor may affect that ratio in words.

4. (15 points; 5 points each) Consider the daily bike rental data contained in the sheet "Bike".

   (a) Find the mode and median of the "weathersit" column. Explain why its mean is meaningless.
   Hint. Are the data nominal, ordinal, or quantitative?

   (b) Use $[0, 10)$, $[10, 20)$, ..., and $[90, 100)$ as classes to construct a frequency distribution for the "humidity" variable.

   (c) Find the mode, median, and mean of the "humidity" variable. For the mode, use the frequency distribution you prepare in Part (b). Among the three measurements, which is the largest and which is the smallest? Comment on their order and the shape of the distribution.

5. (15 points; 5 points each) Consider the daily bike rental data contained in the sheet "Bike".

   (a) For the data contained in the "casual" column, calculate the variance and standard deviation. Should you use the formula for population or sample? Briefly explain why.

   (b) Continue from Part (a). Find all numbers whose z-scores are greater than 3 or less than -3. For them, list the corresponding dates and numbers of daily casual rentals.

   (c) Recall that the data points you find in Part (b) are potential outliers and should be double checked. Investigate those days and try to determine whether they are outliers. If you find anything that strongly suggests them as outliers, report it; otherwise, try to explain why these numbers are so large or small.

6. (15 points; 5 points each) Consider the daily bike rental data contained in the sheet "Bike".

   (a) For the three variables "temp", "humidity", and "cnt", calculate the three correlation coefficients, one for a pair of variables.

   (b) Based on your findings in Part (a), comment on how temperature and humidity tend to affect the number of daily rentals. Describe the relationship and how strong the relationship is.

   (c) Which variable, temperature or humidity, seems to have a stronger impact on the number of daily rentals? Does that fit your intuition? Briefly explain.