BZAN 535 **Homework 6** **Due: October 5, 2016**

1. The file *Kroger_trips_grouped_by_date_hw6_2016.jmp* contains spending at Kroger over a 15-month period (April 2010 – June 2011) for 100,000 households, **GROUP BY date**.
    a. Present a graph of side-by-side box plots for the number of households shopping by day of the week. Write a paragraph explaining the insights from this graph.
    b. In what way do the results differ from what you would expect, based on your experience of grocery shopping? Do you think Kroger has modified the data – or might this be accurate for some subset of Kroger shoppers? (This question necessarily requires some speculation. Day 1 corresponds to Sunday.)

2. This question (and #3) concerns the mean dollars spent per visit for each of the 445 days.
    a. Averages tend to be normally distributed. For Day of week = 1, produce a normal quantile plot for mean spend. This will reveal 7 outliers. Report the p-value for the Goodness of Fit test with and without these 7 values. State the null and alternative hypotheses, and your conclusion about this distribution.
    b. Do the same for each of the remaining days of the week, excluding outliers as indicated by each box plot and then checking for normality. Which day(s) of the week is (are) not normally distributed, even with the removal of outliers?
    c. Examine all the outlier points and comment on what the unusually large mean spend dates have in common, and on what the unusually small spend dates have in common.

3. After excluding outliers in 2a and 2b, you will have about 400 days left.
    a. Present a histogram for mean spend from all of the days of week together. Describe the resulting distribution. What is the standard deviation for this sample of about 400 values and why is it larger than the standard deviations for individual days?
    b. Construct an ANOVA using X = Day of Week, Y = mean spend. Report the F statistic and p-value. What hypotheses is this testing? Summarize the conclusion for this test in the language of the problem.
    c. This same ANOVA can be obtained by using Fit Model and putting Day of Week as the only term in the model, and mean spend as Y. As we did in HW5, check for autocorrelation using the Durbin-Watson test. Report the autocorrelation value. To what do you attribute this positive autocorrelation? To help you answer this question, create a plot using Graph Builder taking X = Date and overlay using Day of Week. Interpret this graph and the autocorrelation.

4. Are days that have more shoppers than usual for that day of the week prone to have higher dollars per basket (than is typical for that day of the week)?

a. Produce a graph using Graph Builder with mean dollar spend on the y-axis, number of trips on the x-axis, and use Overlay for Day of Week. Display this graph and use it to answer the question above.
b. Discuss the graph further, explain which days are similar for this pair of variables and which are unique.

5. Given that there are slightly more than 12 million trips, the individual trip data are a bit overwhelming. So I post just a sample of 120,000 trips (a 1% sample).
   a. Using JMP's Table->Subset, select a random sample of just 120 rows from this file of daily trips. Using Distribution, obtain a histogram for visit_spend with this sample. Report the mean and standard deviation for this distribution. Save this sample for question 5b.
   b. The mean and standard deviation are equal for the exponential distribution. Use "Continuous Fit", under Analyze -> Distribution in JMP, to see whether the exponential distribution is a suitable model for the amount spent per day by this population of families. Show and interpret the goodness of fit test and the diagnostic (quantile) plot.
   c. Now return to the original data file of 120,000 rows. Show the histogram, diagnostic quantile plot, and goodness of fit test for exponential distribution fit for the same variable in 5b.
   d. Explain what you learn from the difference in the goodness of fit tests in 5b (where n=120) and 5c (where n = 120,000).

6. Using the 120,000 rows, report an 80% prediction interval for sales at Kroger. Explain why your chosen method (t-based or quantiles-based) was the most appropriate.

7. I took a simple random sample of 120,000 rows – a 1% sample. Suppose instead that I had sampled all trips for 1000 randomly selected households (1% of households). The resulting sample would be about the same size. Why is it better to sample 1% of households and have all their trips? What more could we have learned about shopping behavior that way?