**12 January 2020**

**SDM Homework A: Spring 2020**

**Part 1: Descriptive statistics**

This assignment uses part of data from the Baldrige scoring process collected by the National Institute of Standards and Technology (NIST) for years 1990 to 2006. The Malcolm Baldrige National Quality Award (MBNQA) is considered as one of the most powerful catalysts of quality and organizational performance excellence in the United States, and it has had significant influence throughout the world. This data consists of performance evaluation scores of organizations from different sectors (variable “sector”) like (1) **manufacturing, (2) services, (3) small business, (4) education, (5) healthcare, and (6) nonprofit** from 1990 to 2006. The dataset has scores on 7 categories that include **(1) leadership; (2) strategic planning; (3) customer focus; (4) measurement, information and analysis; (5) workforce focus; (6) process management; and (7) results.** In turn, these categories may have subcategories and the Criteria has sometimes added, discontinued or modified subcategories over time.

There is a variable “slnoskm17mar11” that simply provides a unique rowname to each observation. The variable “applicant” is applicant number within a particular year. There is another variable called “permanentid” that has an identifier for organizations to identify them uniquely over time if they appear in the data more than once.

Note that the scores are assigned by multiple volunteer examiners for each subcategory. We have individual (median) scores and consensus scores from the examiners. **Prefix “i”** in the column names tells us that this score is the median value of the all individual scores given by the individual examiners for the specified category and subcategories. **Prefix “c”** means that the score is a consensus scores received by an organization. In early years, the consensus score were provided only if the organization scored above a certain cut-off value.

For example,

i1.1score : Individual median score for 1st subcategory in Leadership,

i1.max : It is the maximum possible score for 1st subcategory in Leadership

icat1total: Total score for category 1, which is the summation of all individual subcategory scores.

icat1max: Max possible score for category 1, which is the summation of all individual max scores

iirtotal: Total of all category scores from 1 to 7

iirmax: Max possible total score across categories 1 to 7

The same naming scheme is followed for consensus score as well.

NOTE: The data may require you to do some pre-processing before it can be used for analysis—you should develop necessary R code for all such preprocessing and data cleaning.

1. Download the dataset “baldrige2011.xlsx” posted on Canvas. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please use “stargazer” library for showing all the output tables and “ggplot2” for all graphs.
2. What are the number of observations, mean, median, standard deviation, min, maximum, and mode of iirtotal and ccrtotal in this data? Does it make more sense to use mean, median, or mode as a measure of central tendency for these two variables?
3. List the mean, median, standard deviation, min and max of iirtotal and ccrtotal by Sector (make sure that you label sectors such that 1=mfg, 2=service, 3=small, 4=education, 5=health, 6=nonprofit and the output shows sector names and not numerals that denote the sector). Which sector has the highest variation in ccrtotal?
4. Identify the outliers in this data set in terms of ccrtotal using a box plot. How does the mean and standard deviation of ccrtotal change if the outliers are included versus excluded from the data set?
5. Draw a graph to represent the relationship between icat4total and ccat7total. Make a comment on the graph and how will you interpret it. Also compute the correlation coefficient between these two variables. What does this coefficient tell you about the relationship between the variables? (Hint: correlation can’t be calculated for 2 variables with different number of observations)
6. Provide a count of organizations that have appeared more than once in the data and identify them using their permanentid. Make a table that shows counts of organizations appearing more than once (i.e., twice, thrice, four times, etc).
7. Some companies have been evaluated for 6 or more times in the data. Identify these companies and plot a line graph for any one of these companies that you found most interesting by looking over the trends in icat4total and ccat7total scores during the years they were evaluated. Explain why you found that company interesting and make some conjectures about the relationship between icat4total and ccat7total based on what you observe.

**Part 2: Data Visualization**

2. Download the dataset “baldrige2011.xlsx” posted on Canvas. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please use “stargazer” library for showing all the output tables and “ggplot2” for all graphs.

1. Investigate the distribution of icat4total and icat7total scores for healthcare sector with the help of a histogram plot. Attach the resulting graph, copy and paste the accompanying R code for computing this histogram, and listthree key observations.
2. Now, investigate the relationship between icat4total and ccat7total scores for those companies that have both the scores available and belong to manufacturing sector using a scatterplot. Attach the resulting graph, copy and paste the accompanying R code that you used to draw this scatterplot, and list two key observations about the relationship between the scores that you see from this plot
3. Plot a line graph to understand the trends in average iirtotal scores by sector. Also, draw vertical lines at years 1995 and 1998 to separate out and better visualize the trends in 3 different time periods (before 1995, 1995-1998, after 1998).
4. Plot a line graph to understand the trends in number of applicants by sector. Also, draw vertical lines at years 1995 and 1998 to separate out and better visualize the trends in 3 different time periods (before 1995, 1995-1998, after 1998).
5. Draw a box plot for all individual category totals (icat1total,icat2total….icat7total) in manufacturing sector and identify the scores that have outliers.
6. Finally, refine your analysis in part 2b by conditioning the scatterplot between icat4total and ccat7total score during 3 different time periods (upto 1994, 1995 to 1998, and 1999 onwards). To do this, install the R library “lattice” by typing *install.packages(“lattice”)*, then load it for your current session by typing *library(lattice)*. Since we want to condition the scatterplot for different time periods, create a new variable, say, *time\_period* for which

time\_period = 0 if year <= 1994 and

time\_period= 1 if 1995 <= year <= 1998

time\_period= 2 if year >= 1999

and compute a scatterplot between icat4total and icat7total score, conditional on time\_period. Attach the resulting graph, copy and paste the accompanying R code for computing this scatterplot, and list three key observations about the relationships between advertising, income and sales that you see from this plot.

1. Create a table of summary stats (N, mean, sd, min, max) for all individual category totals (icat1total, icat2total, …, icat7total). Note: The output should be neatly formatted in a Table and the values should be rounded to 2 decimal places. Please use “stargazer” library for creating the stats table.
2. Create a table of pairwise Pearson correlation coefficients for all consensus category totals (ccat1total,ccat2total, …, ccat7total) also showing p-value for correlations. Note: The output should be neatly formatted in a Table and the values should be rounded to 2 decimal places.

**Part 3: Confidence Intervals/ Prediction**

1. Download the dataset “baldrige2011.xlsx” posted on Canvas. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please use “stargazer” library for showing all the output tables and “ggplot2” for all graphs.
   1. Construct a 95% confidence interval for the average icat7total. Based on this confidence interval, what is the maximum icat7total that a company can score with 95% confidence. Write the R code that you used to arrive at this answer.
   2. Is icat7total score statistically different for companies in healthcare and education sector? If so, which set of firms have higher score? To answer these questions, split the data into two subsets for healthcare and education firms and construct a 95% confidence intervals for each subset. Based on these two confidence intervals, is it likely that the average score of healthcare firms is same as that of education firms? Write the R code that you used to derive this inference.