Student Name _____ **Section #** _____
STAT 202 BASIC STATISTICS **Lab: Exploratory Data Analysis**

Preparing Your Solution:

The lab report should include a well written response to the following questions and all relevant supporting graphs and analyses performed using *StatCrunch*.

Part I: (FAST FOOD DATA) Load the data set "STAT202-Lab-data-FastFoodDiet" from the class' Blackboard account into *StatCrunch.* The data set contains information obtained through a nutrition study of food items offered by popular fast food chains.

1. How many cases are in the data set? How many variables are in the data set? Which variables are categorical variables? Which variables are quantitative variables?

2a. Construct a histogram of the variable **Carbs**. Describe the shape of the distribution of **Carb** values in terms of mode and symmetry, including any skewness that you observe and the direction of the skew.

2b. Construct a boxplot of **Carb** values.

"*Graph*"→ "*Boxplot*" Select column **Carb**. Under other options: Select "*Use fences to identify outliers*". Select "*Compute*".

Does the variable contain any outliers? If so, list the specific fast food chains and items that contain outlier **Carb** values. (On the boxplot, click the outlier points to select, turning them pink, then scroll through the dataset to investigate. Is there some variable which is common to most or all of the selected rows? If so, what is it? When you are done, click "clear" to remove the selections and the pink highlighting.)

2c. Calculate the appropriate set of summary statistics to describe the center and spread of the distribution of **Carb** values.

"Stat" -> "Summary Stats" -> "Columns" and then select the appropriate set of summary statistics.

3. Construct a histogram of the **Serving Size** values. Investigate the observations that fall in the lowest 3 bins by clicking on each of the bins. The individual observations that fall in those bins should be highlighted. **For the rest of the Lab, we want to remove the fast food chain that appears to offer the greatest number of small-serving-size items as highlighted in the lowest three bins of your histogram** (**Hint:** Scroll down the data set and see which brand has the largest number of highlighted cases in the data set). To eliminate this restaurant chain from the data set,

Data → Row Selection → Select Where under *Expression* enter **"Fast Food Restaurant" = "Name of Restaurant"**. (**Note:** You must include the quotations as shown here) Select *Compute.*

Next go to menu "*Edit*" → "*Rows*" → select *Delete* and select *Compute*.

4. Construct a histogram for **Sodium** (mg).

a. Calculate the mean and the median of the **Sodium** variable. What do they tell us about the shape of the distribution?

b. Report the appropriate measures of center and spread for the distribution.

c. Click on the bins with values higher than the mean to highlight items with an above-average sodium count. Examine the selected rows, and list which types of food have an above-average sodium count.

d. What percentage of cases in this data set offer food items with less than 240 mg of sodium? Explain how you obtained your answer.

5. Construct side-by-side box plots to compare the variation in item **Calories** sold by the various fast food chains. You will do this comparison in two separate ways – first, **without** transforming (re-scaling) **Calorie** data, and then after transforming **Calorie** data to adjust for unequal **Serving Sizes**.

A. With *Untransformed* Calorie data:

"*Graph*"→ "*Boxplot*" Select column **Calories.** *Group by:* select **Fast Food Restaurant**. Under other options: Select "*Use fences to identify outliers*". Click "*Compute*".

Compare the median Calories across the different fast food brands. Identify the two fast food brands that have the lowest and highest median calories on offer.

B. With *Transformed* Calorie data:

You may have noticed that the service size (in grams) is not the same across the different items. For example, the McDonald's 4-Piece Chicken Nuggets have a service size of 65 grams, while the 4-Piece Chicken Nuggets offered by Chick-fil-A has a serving size of only 57 grams. One way to account for the different serving sizes is to transform the **Calorie** count into a per-gram Calorie Count. Do this by dividing all of the **Calorie** values by their corresponding service size values as follows:

"Data" → "Compute" → "Expression". In Expression type the following: **"Calories"/"Serving Size (g)"** Make sure you include the quotation marks. In Column Label, type the new variable name: **CalPerGram**. Select *Compute*.

Now construct the side-by-side boxplots using the **CalPerGram** column variable and identify the two fast food brands that have the lowest and highest median **CalPerGram** on offer. What do you observe?

Part II The **US Bureau of Justice Statistics** (https://www.bjs.gov/index.cfm?ty=tp&tid=31) defines Violent Crime as follows: "Violent crime includes murder, rape and sexual assault, robbery, and assault". Download the Dataset "**STAT202-Lab-data-USViolentCrime1960-2012**" from the class Blackboard page and upload on to StatCrunch. The Data has both the total violent crimes committed in the US – including murders, rapes, robbery, and aggravated assaults -- for each year from 1960 to 2012. It also has the corresponding rates of violent crimes per-one-hundred thousand of population.

1. Construct a simple Time Plot in StatCrunch that shows the trend in all five rates – **Aggravated Assault, Robbery, Forcible Rape, Murder, Violent Crimes**– from 1960 to 2012. To construct the time plot, do the following:

"Graph" → "Index/Time Plot" → under Select Column, select the relevant columns → under X axis format, select "Time" and select the "Year" as type, with 1960 as the Start year, and select Increment of 1 → for Display, select Points, size 2 and Line, size 1. Then Select *Compute*.

2. Copy and paste the plot in your document and describe the common trend across all the crime variables in the plot.

3. Identify the year the crime rate peaked for each of the five crime variables in your plot.

4. Do you have a hypothesis about the common trend in crime rates that you observe in your plot, post-peak?