STAT 526 — HW 1–KEY

1. Part of this class involves working in groups where each group must submit consensus answers on the final project and all homework assignments. Your first task for homework one is to brainstorm a well documented strategy addressing the details of how your group plans to complete and submit each assignment. All due dates are posted on the course schedule so you should reference this while writing your answer. You should set up some expectations related to what effort each group member should bring before collaborating and comparing as a group online. Also your group should outline a schedule of meeting times for each assignment and possibly assign a different group leader for each meeting. Although your group will certainly adjust as the semester progresses this will give you a starting place and guarantee that everyone has the same expectations from the start of the semester and all online meetings are on your calendar.

2. *USA Today* reported on a study that suggests "frequently heading the ball in soccer lowers players' IQs." A psychologist tested 60 male soccer players ages 14-29 who played up to five times a week. Players who averaged 10 or more headers a game had an average IQ of 103 while players who headed one or fewer times per game had an average IQ of 112.

   (a) Describe the study population of interest to the psychologist.
   All male soccer players ages 14-29 that play up to five times a week.

   (b) Describe the sample used by the psychologist.
   The 60 male soccer players tested.

   (c) Identify the two different variables.
   Number of headers and IQ.

   (d) Classify these variables as either qualitative or quantitative.
   IQ is a quantitative variable, while headers is a categorical (2 categories) or qualitative variable.

   If groups provide a good explanation along with an answer of numerical for number of headers we would not take points off. Please Note: The problem summarizes the data in such as way as to assume categorical data types for the two header groups. If we have the mean response reported for 2 or 3 separate groups this implies that one variable is numerical and one variable (based on the group assignments) is categorical.

   A complete answer addressing any uncertainty in the variable type would read something such as the following: The actual number of headers is a numerical variable; however, the data has been grouped into two categories before summarizing the results. This means a variable that originally had a numerical representation has been converted to a categorical representation.

   (e) Describe the statistics reported in this study.
   Players who averaged 10 or more headers a game had an average IQ of 103. Players who headed one or fewer times per game had an average IQ of 112.

   (f) Describe the corresponding parameters related to this study. Remember that even though we often do not know the actual numerical values of parameters we may still describe what they should represent using words.
   The true population mean of IQ for soccer players that average more than 10 headers a game, and the true population mean for players who headed one or fewer times per game.

   (g) Provided details of a graphical display that would be appropriate to visualize the differences in IQ based on the two groups of soccer players.
   Side by side box plots are a good way to compare quantitative variables across multiple groups or categories. Another answer we allowed was to report stacked histograms (one for each category for headers) in this case you must specify that the axes on both histogram plots should be identical to provide easy visual comparisons.

3. **How Many Cleaning Crews?** High-end, "boutique" hotels can be found in all major cities. Because these hotels tend to be small (less than 100 rooms, typically) and promise outstanding customer service, they have attracted the type of business clientele that can afford to pay a premium for their room. One perk that is offered hotel guests is that their room will be cleaned within the hour from the time when the guest calls for service. In order to meet this promise, hotel managers must make sure that they have enough cleaning crews on site so that at least one is available for re-directing at any time. In principle, the manager could hire more cleaning crews than he or she thinks the hotel will need. But on the other hand, hiring more staff than strictly needed only adds to the cost of each room. Therefore, being able to predict the number of cleaning crews that might be required to meet the needs of a boutique hotel with a given number of occupied rooms is an important component of a sound business model. To this effect, the American Association of Hotels and Restaurants (AAHR) conducted a study to collect information from a nationwide random sample of boutique hotels considered to be effectively managed. A total of 53 hotels were included in the study. The number of occupied rooms and the number of cleaning crews employed by the hotel were recorded for several different hotels in 2014.

(a) Describe the study population of interest.
   Boutique Hotel effectively managed according to the AAHR with less than 100 room all in major cities. (Although not stated we would probably assume major US cities but this would need to be further verified.)

(b) Describe the sample used in this example.
   The total of 53 hotels sampled.

(c) Identify the two different variables.
   Number of occupied rooms (X) and number of cleaning crews (Y).

(d) Classify these variables as either qualitative or quantitative.
   Both of these are quantitative variables.

(e) Describe one or more statistics that might be of interest related to this dataset.
   We might be interested in the sample correlation between the number of occupied rooms and the number of cleaning crews. There are many possible answers here.

(f) Import the dataset CleaningOriginal.csv into R and use this to create an appropriate graphical display for each variable separately. Below each graphical display put a detailed description of the variable and distribution. Be sure to put this in the context of the example.

Distribution of Crew Members   **Distribution of Rooms**
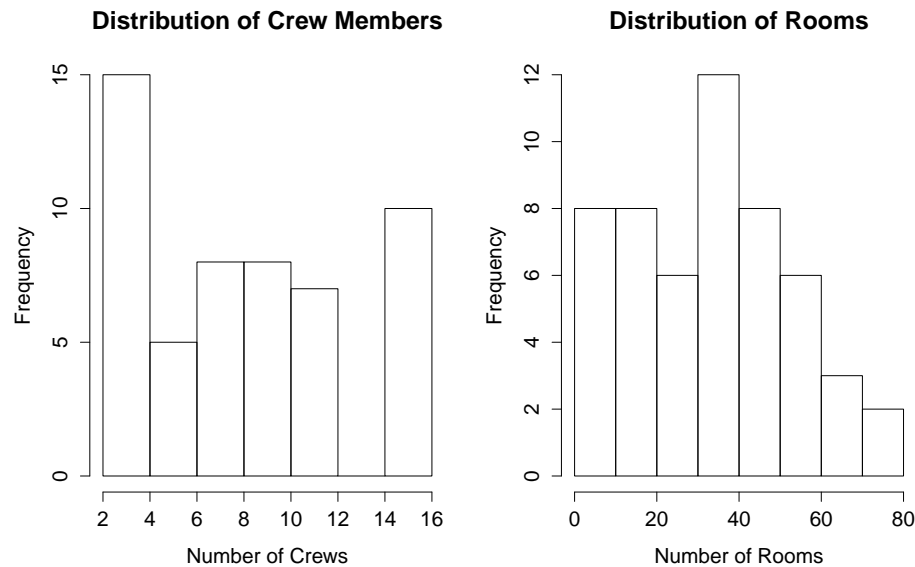
The answers will vary. Ideally you should mention the shape, center and spread of these distributions.

(g) Comment on what is missing in the graphical displays given in part (3f) related to visualizing a comparison of the two variables and their corresponding simultaneous relationship.
We can see what the distribution of each variable is quite well. However, it's difficult to see if there is a relationship between the two variables (e.g., a linear relationship).

(h) Create an appropriate graphical display that will allow you to visualize the relationship between the two variables simultaneously. Fully describe this graphical display within the context of this example.

Rooms vs. Cleaning Crews Scatterplot
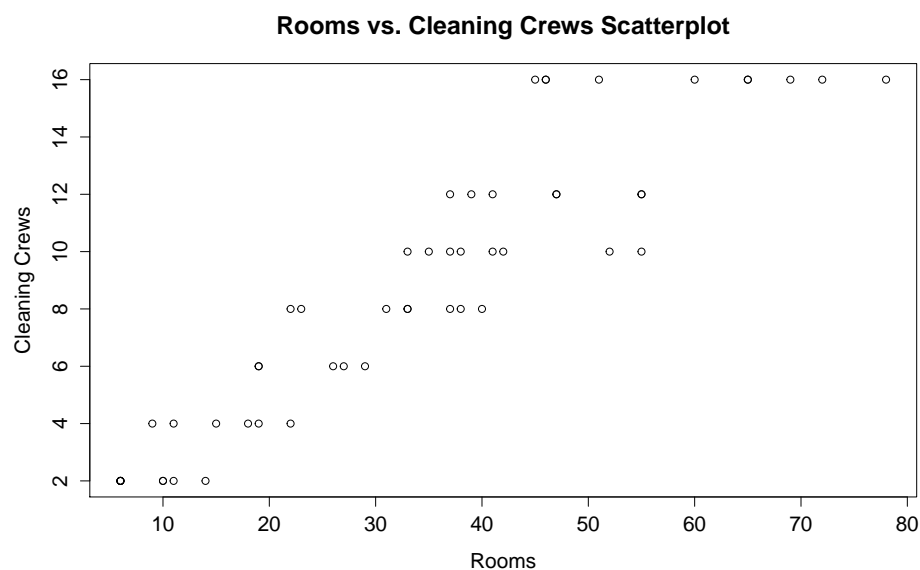
It looks like as the number of rooms increases, so does the number of cleaning crews on staff. The relationship starts out linear, but appears to flatten off at around 50 rooms.

4. In an experiment to investigate the effect of color of paper (blue, green, orange) on response rates for questionnaires distributed by the "windshield method" in supermarket parking lots. 15 similar demographic supermarket parking lots were chosen in a metropolitan area and each color was assigned at random to five of the lots. The response rates (in percent) are given in the file Response.csv.

(a) Use R to create an appropriate graphical display. Your graphical display should help you visualize the differences/similarities based on the color of paper and the relation to the response rate.

Response Rate by Color of Flier
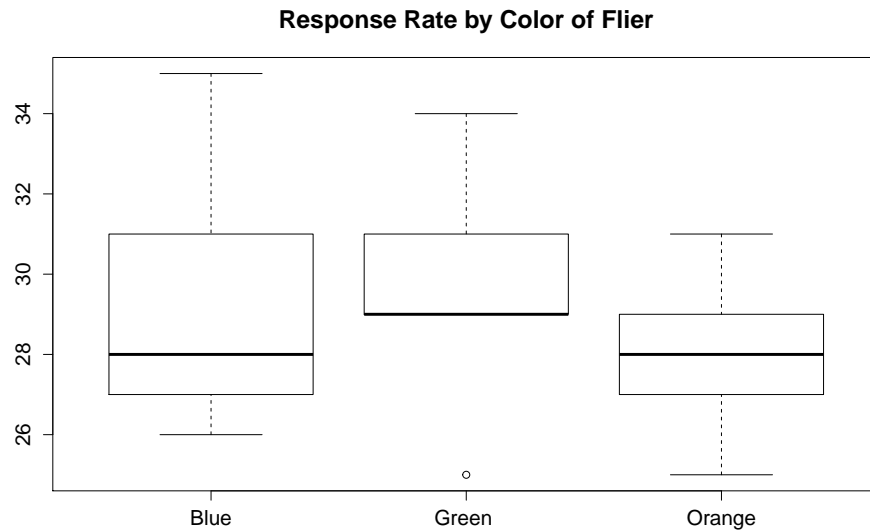
(b) Provide a detailed description of the graphical display in the context of this example.
It doesn't look like any color performs significantly better than the others. All of the IQRs overlap and the medians seem to be between 28-30 for all three groups. Color does not appear to have an effect on the response rate.