1. Suppose that the observational units in a study are the 21 purchases that I made on amazon.com in the year 2011. Identify each of the following as a categorical variable, a quantitative variable, or not a variable. (Circle your answer for each; do not bother to explain.)

a) How much did I spend on the purchase?
   - Categorical
   - Quantitative
   - Not a variable

b) Was the purchase shipped to me or to someone else?
   - Categorical
   - Quantitative
   - Not a variable

c) Do I tend to spend more on purchases sent to others than on purchases sent to me?
   - Categorical
   - Quantitative
   - Not a variable

d) Did the purchase include free shipping?
   - Categorical
   - Quantitative
   - Not a variable

e) What was the average price of these purchases?
   - Categorical
   - Quantitative
   - Not a variable

2. A few years ago, I asked students in my classes whether they drink coffee every day, sometimes, or (almost) never. I also asked each student to report his/her gender. Results are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Every day</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>Sometimes</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>(Almost) never</td>
<td>19</td>
<td>7</td>
</tr>
</tbody>
</table>

a) What kind of graphical display would be appropriate to display these data? (Circle your choice. Do not bother to explain. Do not bother to create the graph.)
   - histogram
   - boxplot
   - segmented bar graph

b) What proportion of the female students drink coffee every day?
   \[
   \frac{15}{32} \approx 0.469
   \]

c) What proportion of the students who drink coffee every day are female?
   \[
   \frac{15}{22} \approx 0.682
   \]

3. The National Retail Federation conducted a national survey of 8526 consumers on September 1-9, 2009. Among the findings reported were that:
• 29.6% of those surveyed said that the state of the U.S. economy would affect their Halloween spending plans;
• the average amount that the respondents said they expect to spend on Halloween is $56.31.

a) Identify the observational units in this study.

The observational units are the 8526 consumers who were surveyed.

b) Identify a quantitative variable in this study.

A quantitative variable is the amount that the person expects to spend on Halloween.

c) The information above does not say whether randomness was used in this study. Which kind of randomness would have been appropriate? (Circle your answer; do not bother to explain.)

random sampling random assignment

4. Create a hypothetical example of 10 exam scores (integers between 0 and 100, inclusive, with repeats allowed) with the property that only one of the exam scores is greater than the mean.

One example that works is: 80, 80, 80, 80, 80, 80, 80, 80, 80, 100

5. In a study published in a 2007 issue of the journal *Preventive Medicine*, researchers found that smokers were more likely to have used candy cigarettes as children than non-smokers were.

a) What type of study is this? (Circle your answer; do not bother to explain.)

observational experimental

b) When hearing about this study, John responded: “But isn’t the smoking status of the person’s parents a confounding variable here?” When Karen asked what he meant, John said: “Children whose parents smoke are more likely to become smokers themselves when they become adults.” What else does John need to say in order to explain how the parents’ smoking status can be a confounding variable in this study?

In order for the parents’ smoking status to be a confounding variable, it must be the case that children of parents who smoke are more likely to use candy cigarettes than children whose parents do not smoke.

6. Consider the following four distributions of quiz scores:
Arrange these in order from smallest standard deviation to largest standard deviation. (Just write in the letters A, B, C, D. Do not bother to calculate any standard deviations. Do not bother to explain.)

Smallest: B   Next-to-smallest: C   Next-to-largest: D   Largest: A

7. The book *Day Hikes in San Luis Obispo County* by Robert Stone gives information on 72 different hikes that one can take in the county. The following stemplot displays the distribution of the lengths (in miles) of these hikes (for instance, the shortest hike is 0.6 miles and the longest is 9.5 miles):

```
0 | 68
1 | 00000555555588
2 | 0000000012555556668
3 | 00002224458
4 | 0000566
5 | 0055668
6 | 0000
7 | 004
8 |
9 | 5
```

a) Determine the five-number summary of this distribution. Also indicate how you calculate these values.

The minimum hike length is 0.6 miles. The maximum hike length is 9.5 miles.

Because \((72 + 1)/2 = 36.5\), the median is the average of the 36th and 37th values, which is \((2.8 + 3.0) / 2 = 2.9\) miles.

Because \((36 + 1)/2 = 18.5\), the quartiles are the average of the 18th and 19th values. The lower quartile is \((2.0 + 2.0) / 2 = 2.0\) miles, and the upper quartile is \((4.5 + 4.6) / 2 = 4.55\) miles.

b) Is the 9.5-mile hike an outlier according to the \(1.5\times\text{IQR}\) rule? Justify your answer with appropriate calculations.

\[
\text{IQR} = (4.55 - 2.0) = 2.55\text{ miles.}\]
\[
1.5\times\text{IQR} = 1.5\times2.55 = 3.825.\]

The mileage needed to be an outlier on the high end is \(4.55 + 3.825 = 8.375\), so the 9.5 mile hike is an outlier.
8. In a recent study, researchers purchased 40 food items in New York City and determined the actual calorie content of each through a laboratory analysis. They then calculated the percentage difference between the actual calorie content and the calorie count listed on the item’s label. (A positive percentage difference corresponds to a food item whose actual calorie content was higher than what the label claimed.) Each food item was also classified according to whether it was marketed locally, nationally, or regionally. The boxplots below were constructed to compare the distributions:

Write a paragraph summarizing what these boxplots reveal about the percentage differences between actual and advertised calorie content among the three marketing groups of food items.

The locally marketed items have the highest discrepancy between actual and advertised calorie counts. All of these items had a higher actual than advertised calorie count, and the median is about a 70% discrepancy. Moreover, there is large variability in these discrepancy amounts, from about 5% to about 250%. The regionally marketed items also tended to understate their calorie amounts, but much less so and with much less variability. The median for the regional items is only about a 30% difference. The nationally marketed items list the calorie amounts very close to the actual amounts; the median is about 0 and there is very little variability. The national group even has a low outlier at about -20%, indicating that one item had about 20% fewer calories than listed. The distribution of percentage differences is skewed to the right for the local items and fairly symmetric for the regional and national items.

9. Between the months of September 1990 and May 1997, a statistics teacher gathered data on the average temperature for that month (in degrees Fahrenheit) and the amount of gas usage in his home for that month (in units called therms). Summary statistics for these variables follow:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg temp</td>
<td>71</td>
<td>46.35</td>
<td>1.80</td>
<td>15.16</td>
<td>45.00</td>
<td>26.00</td>
</tr>
<tr>
<td>gas usage per day</td>
<td>71</td>
<td>5.311</td>
<td>0.420</td>
<td>3.538</td>
<td>5.000</td>
<td>6.600</td>
</tr>
<tr>
<td>Pearson correlation of avg temp and gas usage per day = -0.930</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Determine and report the equation of the least squares line for predicting a month’s gas usage per day based on its average temperature.

The slope coefficient is: 
\[ b_1 = r \frac{s_y}{s_x} = (-0.930) \times (3.538)/(15.16) \approx -0.217 \]

The intercept coefficient is: 
\[ b_0 = \bar{y} - b_1 \bar{x} = 5.311 - (-0.217) \times 46.35 \approx 15.369 \]

The equation of the least squares line is therefore:
predicted gas usage per day = 15.369 – 0.217 \times \text{average temperature}.

b) Interpret the value of the slope coefficient.

The predicted decrease in gas usage for each additional degree (F) of temperature is 0.217 therms.

c) In the following scatterplot with the least squares line superimposed, circle the point corresponding to the month with the largest positive residual:

The largest positive residual corresponds to the point farthest above the least squares line.

d) Calculate the value of $r^2$, and write a sentence interpreting what this value means.

The value of $r^2 = (-0.930)^2 \approx 0.865$, which means that 86.5% of the variability in gas usage per day across these months is explained by knowing the average temperature for the month.

e) Predict the gas usage per day for a month in which the average temperature is 50 degrees Fahrenheit.

This prediction is: predicted gas usage per day = 15.369 – 0.217 \times 50 \approx 4.519 \text{ therms per day}.

t) The following scatterplots display the age (in months) at which a child first speaks and the child’s score on a cognitive aptitude test taken later in childhood.

Graph A (on the left) displays all of the data in a sample, and Graph B (on the right) excludes the two children who took longer than 24 months to speak.

In one of these graphs the correlation coefficient is $-0.034$, and in the other graph the correlation coefficient is $-0.640$. 
a) Identify which correlation coefficient goes with Graph A (on the left). Briefly explain your choice.

The correlation coefficient for the graph on the left is $-0.640$. The two unusual values (children who took a long time to speak and had very low aptitude scores) create a fairly strong negative association. The graph on the right shows almost no association between the variables, so its correlation coefficient is $-0.034$.

b) Write a paragraph, as if to the parent of an infant, summarizing what these scatterplots reveal about whether there is a relationship between the age at which a child first speaks and his/her cognitive aptitude.

These data reveal that for most children, there is hardly any association between age of first speaking and aptitude score. But there is some evidence that children who take very long to speak tend to have quite low aptitude scores. Unless a child takes very long to speak, though, there is no connection between speaking age and aptitude.