MIDTERM: DATA VISUALIZATION ANALYTICS SECTION POINTS: 50

1. Proportion of Tax Underpayment. A sample of 10,001 Federal income tax returns with errors has been collected, and the amount of payment error for each is provided in Worksheet 1. A positive value indicates the taxpayer is underpaid, and a negative value indicates the taxpayer is overpaid. (2.5 x4 =10)

a. What is the sample proportion of erroneous income tax returns corresponding to tax underpayments?
b. What is the margin of error for a 95% confidence interval on the proportion of erroneous tax returns corresponding to tax underpayments?
c. Construct a column chart displaying the sample proportion of tax underpayments with error bars depicting a 95% confidence interval. Construct the chart to explicitly compare the sample proportion and confidence interval to a 50% proportion benchmark.
d. Can we conclude with 95% confidence that more than 50% of tax payment errors correspond to tax underpayments?

2. Based on the total passenger traffic, the airports in the file Airports are among the busiest in North America. Use the Pivot Table function and,
(2.5x4 =10)

   a. Create a frequency distribution of the total passenger traffic using a bin width of 10, with the first bin starting at 30 million.
   b. Construct a histogram of the frequency counts.
   c. What is the most common passenger traffic range based on this histogram?
   d. Describe the shape of the histogram.

3. Courtney Boyce is auditing a sample of the travel expense reports submitted by company employees over the past year. (5+5=10)

   a. Using the data in Worksheet 3, create a relative frequency distribution of each report's first digit of the expense reported.

   b. Assessing this distribution with Benford's Law, does Courtney have any reason to suspect reporting error or fraud?

4. In a recent report, the top five most-visited English-language websites were google.com (GOOG), facebook.com (FB), youtube.com (YT), yahoo.com (YAH), and wikipedia.com (WIKI). The file WebSites contains a sample of the favorite websites of 50 Internet users. (4+3+3=10)
   a. Create a frequency distribution of the website data. Add the relative frequency column and the percent frequency column, too.

b. Using a column chart, visualize the frequency distribution for these data.
c. Based on the sample, which website is listed most frequently as the favorite website for Internet users? Which is second?

5. The file MarathonRecords contains marathon world records for ages from 6 to 90 for women and men (records for ages 10 and 11 are missing).

(2.5x4=10)

a. Create a scatter chart with age on the horizontal axis and the women's marathon record on the vertical axis. Use "Female Marathon Records (in minutes)" as the chart title and "Age (years)" as the horizontal axis title. Edit the chart to improve interpretation.

b. Create a scatter chart with age on the horizontal axis and the men's marathon record on the vertical axis. Use "Male Marathon Records (in minutes)" as the chart title and "Age (years)" as the horizontal axis title. Edit the chart to improve interpretation.

c. Create a scatter chart that plots both the women's record versus age and the men's record versus age. Select Scatter with Straight Lines. Use "Marathon Records (in minutes)" as the chart title and "Age (years)" as the horizontal axis title. Edit the chart to improve interpretation.

d. Based on the charts in parts a, b, and c, what observations can you make regarding the women's and men's marathon records?