PREDICT 401 Introduction to Statistical Analysis Data Analysis Assignment #1

Introduction and Objectives:

Exploratory data analysis (EDA) is an essential first step in data analysis and model building. EDA has two components: description and exploration. Statistical summaries and visual displays are used to better understand the data. It has certain characteristics:

- displays properties of the variables,
- reveals associations between variables,
- provides insights into potential analytical problems; and,
- sets the stage for further analysis, data collection or model building.

This assignment will demonstrate methods used during EDA; and, require application of these methods to a data set. A quiz will follow.

Overview:

The first part of this assignment will introduce and characterize the standard normal density function using summary statistics and graphical methods. The second part will use data from the Case Study "Coca Cola Develops the African Market" to demonstrate aspects of an exploratory data analysis (EDA). The third part requires application of these methods to data from the Case Study "Soap Companies Do Battle". Instructions are provided to guide the investigations. No report is required. A quiz must be completed. The quiz will have questions on aspects of EDA, and on computational results obtained from the analysis of the Case Study "Soap Companies Do Battle". To prepare for the quiz it is essential to follow the instructions in this assignment.

Preliminaries:

The language R will be required throughout the assignment. An R package will also be required: "moments". It is available from the Comprehensive R Archive Network (CRAN). In addition, download the two data sets "Coke.csv" and "soap_sales.csv" from the course shell.

References:

1) Chihara and Hesterberg, *Mathematical Statistics with Resampling an d R.* pages 13-30.
2) http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm

Part 1 Assessing Normality

This part of the assignment introduces methods which can be used to characterize an empirical distribution of data. Both summary statistics and visual methods will be demonstrated using simulated data from a standard normal density function. The reason for this is that statistical methods based on the normal distribution are used frequently. The valid application of these methods depends on the degree to which a normal distribution models the data. The work in this section gives a baseline for evaluating data in the future.

Check the references. Use the R code from Appendix A of this assignment. The code is documented with numerous comments. The output gives a baseline for comparing the results which will be obtained in Part 2 and Part 3 of this Assignment. Execute the code several times to see the degree of variation in the results obtained from sampling a normal distribution.

Part 2 "Coca Cola Develops the African Market"

Refer to Black *Business Statistics* page 96 problem 1. The data in "Coke.csv" will be used to characterize the production process so that a performance assessment can be made. Use the R code from Appendix B of this assignment and execute. Complete the assigned readings and consider the following questions as you review the results of this analysis.

1) Is one of the four: summary(), hist(), stem() and boxplot() more effective and preferable for identifying outliers?
2) What is the advantage of using a stem-and-leaf display relative to a histogram?
3) Why calculate the percent coefficient of variation and 20% trimmed mean? What is their usefulness for characterizing these data?
4) How does the distribution of the Coke.csv data compare to a normal distribution based on skewness, kurtosis, the QQ plot and the empirical distribution display?
5) Assuming the data are collected in serial order, what conclusions might be reached from the index plot of Fill? How is the bottling process working?

Part 3 "Soap Companies Do Battle"

Refer to Black *Business Statistics* page 48 problem 2. The data in "soap_sales.csv" will be used. Using the code for Part 2 as an example, adapt the code to the sales data. Complete the following steps and be prepared to answer questions on a quiz.

1) Execute summary(), stem(), histogram() and boxplot(). Observe the results.
2) Calculate and compare the 20% trimmed mean to the mean from summary().
3) Calculate the skewness and kurtosis, plot qqnorm(), qqline() and plot.ecdf(). Compare to the results of Parts 1 and 2.
4) Construct a six-cell relative frequency table with lower cell boundary starting at 10 and cell width of 5. Calculate the mean and standard deviation from the grouped data.
5) Using the results from (4), plot a histogram of the relative frequencies with overlay as shown in Part 2. Compare this plot with the stem-and-leaf plot.
6) Plot sales as a function on week similarly to the index plot shown in Part 2.

Upon completion of Parts 1-3, answer the questions in the quiz posted on the course shell.

```
# Predict 401 Data Analysis Project 1
#-------------------------------------------------------------------------
# Part 1 Assessing normality
# Appendix A
#-------------------------------------------------------------------------
require(moments)

# Set the seed for the random number generator so that results can be compared.
set.seed(123)

# Generate standard normal random variables using the function rnorm().
normal_x <- rnorm(10000, mean=0, sd=1)

# Check summary statistics.  Skewness is particularly important. Data which
# are skewed present estimation and inference problems.  For a random sample
# from a normal distribution, the values for skewness should be close to zero
# and the kurtosis values produced by R should close to 3.
summary(normal_x)
skewness(normal_x)
kurtosis(normal_x)

# Plot a histogram with density function overlay.
hist(normal_x, prob=T, ylim=c(0.0,0.5))
lines(density(normal_x),lwd=2, col="darkred")

# Demonstrate QQ Plot by comparing two standard normal variables.  A QQ plot
# is a scatterplot of two sets of data.  The values of the quantiles for the
# two sets are plotted against each other.  If the distributions are the same,
# the resulting plot is a straight line.

# normal_x is one set of data.  Generate a second vector normal_w.
normal_w <- rnorm(10000, mean=0, sd=1)

# Sort and match the ordered sets of data to form the plot.
normal_x <- sort(normal_x)
normal_w <- sort(normal_w)
plot(normal_x, normal_w, main = "Scatterplot of two normal random variables")

# This can be done for any set of data with supplied functions.
# The unknown distribution is plotted against the standard normal distribution.
# The closer to a straight line, the better the normal approximation.
# qqnorm() and qqline() provide the capability to make this comparison.
qqnorm(normal_x)
qqline(normal_x)
# The normal QQ plot illustrates desirable agreement.  Due to sampling
# variability there will always be some departures. Note the next section.
```

```
# Another way to compare an empirical distribution to the standard normal is shown
# by using the plot.ecdf() function.  Since this assignment uses smaller sample
# sizes, a sample of size 50 will be used and compared to normal_x. Execute this
# portion of the code several times to see the degree of variabilty that occurs.
# The degree of sampling variability this simulation reveals is important to
# remember.  This is why we need statistical tests to determine when a departure
# is extreme enough to be declared statistically significant.
normal_w <- rnorm(50, mean=0, sd=1)
plot.ecdf(normal_x,xlab = "Standard Normal Variable x", main = "Comparison to Standard Normal")
plot.ecdf(normal_w, col = "blue", pch =2, add=TRUE)
abline(v = 0.0, lty = 2, col = "red")
legend ("topleft", legend = c("normal", "sample"), col = c('black', "blue"), pch = c(19, 2))

#---------------------------------------------------------------------------
# Predict 401 Data Analysis Project 1
# Part 2
# Appendix B
#---------------------------------------------------------------------------
# Case Study "Coca Cola Develops the African Market"
# EDA using data shown in problem 1 page 96 of Business Statistics.

require(moments)
coke <- read.csv(file.path("c:/RBlack/","Coke.csv"), sep=" ")
str(coke)  # Check structure of the dataset.

# Generate summary statistics and visual displays
summary(coke$Fill)
sd(coke$Fill)
stem(coke$Fill)
boxplot(coke$Fill)
100*sd(coke$Fill)/mean(coke$Fill)  # Compute percent CV.

# Compare summary statistics with 20 percent trimmed mean.
mean(coke$Fill, trim=.2)

# Evaluate the distribution of the data.
hist(coke$Fill, prob=T, ylim=c(0.0,1.5))
lines(density(coke$Fill),lwd=2, col="darkred")

# Determine skewness and kurtosis of data.
skewness(coke$Fill)
kurtosis(coke$Fill)

# Evaluate QQ plot of filled coke cans.
qqnorm(coke$Fill)
qqline(coke$Fill)
```

```
# Comparison of Fill volume versus standard normal using empirical distribution functions.
# To do this, we standardize the data to a mean of zero and standard deviation of one.

mu <- mean(coke$Fill)
std <- sd(coke$Fill)
Fill <- (coke$Fill - mu)/std

normal <- rnorm(1000, mean = 0, sd = 1)

plot.ecdf(normal, xlab = "Standard Normal Variable x", main = "Comparison to Standard Normal")
plot.ecdf(Fill, col = "blue", pch =2, add=TRUE)
abline(v = 0.0, lty = 2, col = "red")
legend ("topleft", legend = c("normal", "sample"), col = c('black', "blue"), pch = c(19, 2))

# Prepare a relative frequency table.
# First define cell boundaries. Second, define cell midpoints.
cells <- seq(from = 339, to = 341.2, by = 0.2)
center <- seq(from = 339.1, to = 341.1, by = 0.2)

Fill_Volume <- coke$Fill
# Cut() places each fill volume into its associated cell.
Fill_Volume <- cut(Fill_Volume, cells, include.lowest=TRUE, right = FALSE)
# table() followed by prop.table() calculates proportions in each cell.
Fill_Volume <- prop.table(table(Fill_Volume))
# Include the cell center in the data frame.
Fill_Volume <- data.frame(Fill_Volume, center)
# Print out the data frame and compare to the stem-and-leaf plot.
Fill_Volume

# Superimpose on histogram using established breaks from cells.
# First establish the count in each cell.
count <- Fill_Volume$Freq*length(coke$Fill)
Fill_Volume <- data.frame(Fill_Volume, count)

# Plot the frequency (count) for each cell with overlay.
hist(coke$Fill, breaks=cells, main = "Frequency in Each Cell", right = FALSE)
lines(Fill_Volume$center, Fill_Volume$count, type = "b", col = "red")

# Calculate the mean and standard deviation from the grouped data and compare.
mean <- sum(Fill_Volume$Freq*Fill_Volume$center)
mean
delta2 <- (Fill_Volume$center - mean)**2
std <- sqrt(sum(delta2*Fill_Volume$Freq))
std

# Add an index variable to the data frame so that a scatter plot can be made.
```

```
index <- seq(1,50)
sample <- data.frame(coke, index)
plot (sample$index, sample$Fill, ylim = c(335, 345), main = "Fill versus Index")
abline(h = mean(sample$Fill))

coke$Fill <- sort(coke$Fill)
coke$Fill[48]  # 95th percentile value

#  Determine if the data can be approximated using a normal distribution.
#------------------------------------------------------------------------
```