Lab Four: Exploratory Data Analysis with One and Two Variables

STA 111 (Summer Session I)

Lab Objective

To explore data with histograms and scatter plots.

Introduction

We will continue to use the RStudio pre-installed on the lab computers.

Before continuing, download the lab report template by typing:

```
download.file("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2016/Labs/Lab.Rmd", destfile ="Lab.Rmd")
```

Note: remember to use the right double quotation marks when you copy commands into R!

Lab Procedures

What are the characteristics of U.S. movies that make the most money? Let's explore this question with the data set *"Movies2012.txt"*. It comprises data on the 250 top domestic grossing movies of all time as of November 2012. The variables are:

Variable	Description
Ranking	Ranking on Domestic gross sales
Title	Movie title
Year	Release year
Domestic	Domestic gross sales
Foreign	Foreign gross sales
Worldwide	Worldwide gross sales
Budget	Budget
Rating	MPAA rating
Best_Picture	Academy Awards Best Picture (nominated or won)
Genre	Main/first genre of the movie
All_Genres	List of all genres the movie falls into
Director	Name of the director

There are missing data in this file. We'll ignore them for simplicity. In general, when confronted with missing data, it is best to get the advice of a professional statistician before doing analyses. First, read in the data. Type:

```
Movies2012 = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2016/Labs/Movies2012.txt",header=T)
```

Note: again, remember to use the right quotation marks when you copy the command into R!

Check that you have the right file by typing `head(Movies2012)` to see the first few rows.

Data Analysis Tip: The unit of measurement for the monetary variables is not stated. That's bad practice. Always include a description of the units somewhere on the file. Based on knowledge of movie revenues, it is clear that that the unit of measurement is $1,000,000.

Questions

1. After reading in the data, describe the distributions of foreign and domestic grosses. That is, say where most values are, note any outliers (*an outlier is an observation point that is distant from other observations*), and say whether the distribution is tightly packed around its mean or is spread out. Also, report the mean and standard deviation.

   You should try the `summary` command. For example, you would type `summary(Data$Variable1)` to see a summary of a variable called *"Variable1"* in a data frame named *"Data"*. You should also try the `mean` and `sd` commands. In addition to summarizing the data, you can use histograms to get a visual representation of the distribution of the data. The command is `hist(Data$Variable1,col="orange")`. Note that I specified color just for fun; I didn't need to.

   Further, if you want to compare 2 histograms in one window you can combine them by first typing `par(mfrow=c(1,2))`, then `hist(Data$Variable1,col="orange")`. Add another histogram by simply typing `hist(Data$Variable2,col="darkblue")` next, where *"Variable2"* is another variable in *"Data"*. By the way, `par(mfrow=c(1,2)` specifies that you want all plots henceforth arranged into 1 row and 2 columns. If that's no longer what you want, simply type `par(mfrow=c(1,1))`.

2. Which sentence best describes the distributions of domestic and foreign grosses? You can just write the letter of your choice on the lab report.

   (a) Domestic and foreign grosses are very similar.

   (b) Domestic and foreign grosses have similar distributional shapes, but foreign grosses tend to be larger than domestic grosses.

   (c) Domestic and foreign grosses have similar distributional shapes, but domestic grosses tend to be larger than foreign grosses.

   (d) The two distributions look nothing like each other, because one has a long left tail and the other has a long right tail.

3. What are the names of the two movies that are the largest outliers on all three gross sales variables?

Hint: It turns out that the movies are already arranged by gross sales so that you only need to look at the first few rows of the data to answer this question.

4. We can examine the relationship between world-wide gross and movie genre using a box plot. Use the variable *"Genre"* for this analysis. The command for a box plot is `boxplot(Contvar~Catvar,data=Data)` where Contvar represents the continuous variable that you are trying to graph, Catvar allows you to break the box plot down by different values of a categorical variable and Data is still your data frame. If you don't want to break the continuous variable by a categorical variable, the command is `boxplot(Contvar,data=Data)` instead.

   Answer the three questions below:

   (a) Out of Comedy and Animated movies, which one has a distribution of world-wide grosses that is most similar to the distribution of world-wide grosses for Action movies? Justify your choice in at most two sentences.

   (b) Compare the distributions for Drama movies and Adventure movies. Do they have reasonably similar medians? Is one more spread out than the other (if so, say which one)?

   (c) If you directed a movie and wanted to make lots of money worldwide, which type appears to give you the best chance of doing so? Base your answer on the results of the box plot.

5. Describe the relationship between domestic gross and foreign gross. To make a scatter plot between Variable1 and Variable2, type `plot(Data$Variable1,Data$Variable2)`.

   Items to include in your description are the general trend of the relationship (e.g., positive and linear, negative and linear, some other pattern, no clear pattern) and whether there are any outliers or points that do not fit the pattern.

6. Report the three pairwise correlations between Foreign, Domestic, and World-wide gross. To find correlations, type `cor(Movies2012[,c("Variable1","Variable2","Variable3",...)])` which will show a matrix of the correlations between all the variables used as input (you can use as many as you'd like). Note that the diagonal is always 1. *Make sure you know why that is.*

   Do the correlations suggest strongly positive linear relationships, weakly positive linear relationships, no linear relationships, weakly negative linear relationships, or strongly negative linear relationships?

7. Why are the correlations between Domestic and Worldwide, and Foreign and Worldwide, stronger than than the correlation between Domestic and Foreign? The answer has to do with the definitions of the

variables.

8. Outliers can have a strong effect on correlations. Let's check to see if excluding Avatar and Titanic changes the correlations substantially. To exclude Avatar and Titanic, type

```
Movies2012 = Movies2012[-which(Movies2012$Title=="Avatar" | Movies2012$Title=="Titanic"),]
```

Data Analysis Tip: Note that != is defined as "not equal to" and | is the "OR" operator.

Now, re-calculate the correlations in (6). Did the correlations get stronger or weaker? Does the substance of your conclusions in (6) change very much when excluding Avatar and Titanic?

Data Analysis Tip: It is not acceptable to exclude outliers from analyses unless you have a scientific reason to do so (e.g., a data entry error, or maybe the outlying unit is not part of your target population). Hiding outliers is fudging data to get results you want. That is dishonest and unethical. When you see outliers, do analyses with and without them. When the results do not change much, report the results based on the full data set, and tell your audience that the results were not sensitive to the outliers. When the results do change substantially, report both sets of analyses: one with and one without the outliers. This honestly informs people that your conclusions are not on very solid ground, because particular data points affect the results greatly.

Feel free to explore relationships between sales and other characteristics of movies, like rating, best picture nominations/wins, and director.

This ends the lab. Remember to turn in your lab reports on Sakai.