Lab 7

A) Kaggle Data Exploration

Please download the data for **KKBox's Churn Prediction Challenge** competition from the kaggle website:

https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data

For the training data, train.csv

1) How many features are numerical data? And how many are categorical data
2) Make histogram for the numerical features, to see how they distribute
3) Make table for the categorical features, to see how they distribute

B)

2.4 Load and attach the data set central .park (UsingR). The WX variable contains a list of numbers representing bad weather (e.g., 1 for fog, 3 for thunder, 8 for smoke or haze). NA is used when none of the types occurred. Make a table of the data, then make a table with the extra argument exclude=FALSE. Why is the second table better?

2.8 The data set npdb (UsingR) contains information on malpractice awards in the United States. Attach the data set and make a table of the state variable. Which state had the most awards? (Using sort () on your table is useful here.)

2.9 For the malpractice-award data set npdb (UsingR), the variable ID is an identification number unique to a doctor but not traceable back to the doctor. It allows a look at a doctor's malpractice record without sacrificing anonymity.
The commands

```
> table(npdb$ID)
```

create a table of malpractice awards for each of the 6,369 doctors. What does the command table (table (ID)) do, and why is this interesting?

2.10 The data set MLBattend (UsingR) contains attendance information for major league baseball between 1969 and 2000. The following commands will extract just the wins for the New York Yankees, in chronological order.

```
> attach(MLBattend)
> wins[franchise == "NYA"]
[1] 80 93 82 79 80 89 83 97 100 100 89 103
59 79 91
…
> detach(MLBattend) # tidy up
```

Add the names 1969:2000 to your variable. Then make a barplot and dot chart showing this data in chronological order.

2.16 The data set rivers contains the lengths (in miles) of 141 major rivers in North America.
1. What proportion are less than 500 miles long?
2. What proportion are less than the mean length?
3. What is the 0.75 quantile?

2.23 The data set npdb (UsingR) contains malpractice-award information. The variable amount is the size of malpractice awards in dollars. Find the mean and median award amount. What percentile is the mean? Can you explain why this might be the case?

2.30 For the data sets bumpers (UsingR), firstchi (UsingR), and math (UsingR), make histograms. Try to predict the mean, median, and standard deviation. Check your guesses with the appropriate R commands.

2.32 Fit a density estimate to the data set pi2000 (UsingR). Compare with the appropriate histogram. Why might you want to add an argument like breaks =0:10−.5 to hist()?

2.34 The data set DDT (MASS) contains independent measurements of the pesticide DDT on kale. Make a histogram and a boxplot of the data. From these, estimate the mean and standard deviation. Check your answers with the appropriate functions.

2.35 There are several built-in data sets on the 50 United States. For instance, state.area (,) showing the area of each U.S. state, and state.abb (,) showing a common abbreviation. First, use state.abb to give names to the state.area variable, then find the percent of states with area less than New Jersey (NJ). What percent have area less than New York (NY)? Make a histogram of all the data. Can you identify the outlier?

2.36 The time variable of the nym. 2002 (UsingR) data set contains the time to finish the 2002 New York City marathon for a random sample of runners. Make a histogram and describe the shape. Can you explain why the shape is as it is?

2.39 The data set hall.fame (UsingR) contains baseball statistics for several baseball players. Make histograms of the following variables and describe their shapes: HR, BA, and OBP.

2.41 Why are the boxplot whiskers chosen with the factor of 1.5? Why not some other factor? You can see the results of other choices by setting the range=argument. Use x=rnorm(1000) for your data. Try values of 0.5, 1, 1.5, and 2 to see which shows the tails of the distribution best. (This random sample should not have a heavy tail or a light tail, meaning it will usually have a handful of points beyond the whiskers in a sample of this size.)

2.42 The data set cf b (UsingR) contains a sampling of the data from a survey of consumer finances. For the variables AGE, EDUC, NETWORTH, and log (SAVING +1), describe their distribution using the concepts of modes, symmetry, and tails. Can you convince yourself that these distributions should have the shape they do? Why?

2.43 The brightness (UsingR) data set contains the brightness for 966 stars in a sector of the sky. It comes from the Hipparcos catalog. Make a histogram of the data. Describe the shape of the distribution.

2.44 It can be illuminating to view two different graphics of the same data set at once.
A simple way to stack graphics is to specify that a figure will contain two graphics by using the command

```
> par(mfrow=c(2,1) # 2 rows, 1 column for
graphic figures
```

Then, if x is the data set, the commands

```
> hist(x)
> boxplot(x, horizontal=TRUE)
```

will produce stacked graphics. (The graphics device will remain divided until you change it back with a command such as par (mfrow=c(1, 1)) or close the device.) For the data set lawsuits (UsingR), make stacked graphics of lawsuits and log (lawsuits). Could you have guessed where the middle 50% of the data would have been without the help of the boxplot?

2.45 Sometimes a data set is so skewed that it can help if we transform the data prior to looking at it. A common transformation for long-tailed data sets is to take the logarithm of the data. For example, the exec.pay (UsingR) data set is highly skewed. Look at histograms before and after taking a logarithmic transform. Which is better at showing the data and why? (You can transform with the command log (1+exec.pay, 10).) Find the median and the mean for the transformed data. How do they correspond to the median and mean of the untransformed data?