Intro to R for Epidemiologists

Lab 6 (2/19/15)

Many of these questions go beyond the information provided in the lecture. Therefore, you may need to use R help files and the internet to search for answers. Feel free to ask questions of the instructor, the TAs, or your classmates, but try to work through as much as you can independently.

For the lab, you are expected to create an R script (.R file in the R editor) with your code corresponding to each question. Begin each question with a commented line of code indicating the question. As an example:

```r
# Jenna Krall

# Question 1.
head(iris)
```

Part 1. Manipulating Datasets with `dplyr`

We will use the `hflights` dataset in R. This dataset contains information on flights departing from Houston's airports in 2011 (Source: Bureau of Transporation, Research and Innovation Technology Administration). In this section, you should use functions from the `dplyr` package.

1. Install the `hflights` package and load the `hflights` dataset (Hint: After installing the package, use the command `data(hflights)` to load the dataset)
2. Create a data frame tbl from the `hflights` dataset (Hint: `?tbl_df`) and call this data frame tbl `flights`
3. Create a new dataset that only includes observations from August where the taxi in time was less than 15 minutes.
4. Using the dataset in (3.), create a new dataset with only the following variables: DayofMonth, DepTime, ArrTime, UniqueCarrier, ActualElapsedTime, DepDelay, ArrDelay, Origin, Dest, and Distance.
5. Sort the new dataset created in (4.) by descending delay in departure time.
6. Group your dataset in (5.) dataset by airline carrier.
   (a) How many unique carriers are included in this dataset?
   (b) How many observations are in each airline carrier group?
7. Calculate the mean arrival delay to each destination using your `flights` dataset from (2.).
8. For each day of the year, count the total number of flights in `flights` and sort in descending order.
9. Create a new variable named "late" that takes the value of 1 if departure delay $\geq$ 15 minutes and 0 if departure delay $<$ 15 minutes. What percentage of flights depart over 15 minutes late?

Part 2. Manipulating Datasets with `tidyr`

1. Download the file "artificialLongData.csv" from the course website.
2. Using the `tidyr` package, convert the dataset from wide-form to long-form data