Lab 1 (40 points): Introduction to Statistical Packages
Objectives: Reading files, cleaning and manipulating the data.

A. (40 points) Airline Dataset 2008. This semester, we are going to be exploring airline on-time performance of domestic flights operated by large air carriers. The information was compiled from the Bureau of Transportation Statistics. We will only be analyzing the data from randomly selected flights from November 2008 which is in the data set airline2008NovS.txt. The variable names and definitions are listed in the file airline2008_dataset_definition.docx. In this lab, we are going to explore what is included in the data set, load it into the software package and do some basic manipulations.

1. (8 points) How many variables does this data set contain? Which are categorical or qualitative variables and which are quantitative or numeric variables? Besides looking at the documentation file provided, you might want to look at the data file itself in a spreadsheet, notepad or the software package (R only).

2. (6 pts.) Write two questions that can be asked from this data set. If the question is asked in this lab, you will not receive credit for that question. In the project due at the end of the semester, your group will have to pose one general question that can be answered by two different inferential methods that are discussed this semester. It is not required that you use one of the questions in this part in your project.

3. (5 points) Read the data into your software package. This part is code only. No output is required.

4. (6 points) Are there missing values (NA) in the dataset? If so, please remove the rows that contain NAs in the dataset. Please save this data file and use the new file for the rest of the semester.
    a. (3 pts.) Code
    b. (2 pts.) How many observations are there after removing the incomplete data?
    c. (1 pt.) In which directory did you save your cleaned data set?

5. (6 points) For readability, we want to transform airport codes to their full name in the variable "Dest". Please only change the name of the four airports listed below using the following abbreviated names. The full names are in parenthesis after the abbreviations to be used in the lab. You can see why the three letter code of abbreviations was started!

    ATL → Atlanta (Hartsfield Jackson Atlanta International Airport),
    CHS →CharlestonAFB (Charleston Air Force Base International Airport),
    DFW → DallasFtWorth (Dallas Fort Worth International Airport),
    MSP → MinneapolisStPaul (Minneapolis-St Paul International/Wold-Chamberlain Airport)

    a. (3 pts.) Code
    b. (3 pts.) Print out the first 6 rows of the data and highlight which airports were changed.

6. (9 points) We are going to see if the variable "ActualElapsedTime" can be calculated from other variables in the data set.
    a. (3 pts.) Write down a mathematical equation to calculate "ActualElapsedTime" from “AirTime”, “TaxiIn” and “Taxiout”.
    b. (3 pts.) Code for part a).
    c. (3 pts.) Show that your code as correct by displaying the original variable "ActualElapsedTime" and the variable that you calculated. Please only print out the first 6 rows.