Problem Set 1

This problem set is designed to get you more comfortable analyzing data with R. You will take a raw data set of airline tickets sold in the United States from the Bureau of Transportation Statistics, clean it, and perform some simple analyses.

You are allowed to work in groups of up to four students, but you must disclose the members of your group. Individual submissions are required. The code you submit may be identical to the one of the other group members, but we expect the comments and answers to the questions to be your own.

Your submission should consist of: (i) a R markdown document with code, figures and comments, and (ii) a zip folder containing all files needed for replication. You may upload these materials via the course's Canvas website (please do not email us with your homework submission).

Each of the four questions below summarize the required tasks, with the individual bullet points detailing the steps needed to complete them.

0) Download Rstudio. You can find the latest version here:

`https://www.rstudio.com/products/rstudio/download/#download`

1) Download the DB1BMarket data table from the following URL:

`https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=125`

a) Take only data from the first quarter of 2015.

b) Take the following variables: ItinID, MktID, OriginAirportID, DestAirportID, TkCarrierChange, TicketCarrier, Passengers, MarketFare, and MarketDistance.

c) Download the data and bring it into R.

2) Remove tickets that can't be assigned to a unique carrier, remove markets (a unidirectional origin-destination pair) with less than 20 passengers per day, and remove tickets with extreme prices.

a) Notice that the data are at the one-way ticket level. Start by removing any tickets that have a ticket carrier change.

b) Remove tickets with prices less than $25 or more than $2,500.

c) Calculate the total number of passengers in each market and remove tickets that belong to a market with less than 20 passengers per day, on average. Following the above, an example of a market is LAX to JFK, which is a different market than JFK to LAX.

3) You will create two datasets: one at the market-carrier level and another at the market level. The former dataset should have average prices and total passengers for each market-airline and the latter should have each market's average price, number of firms, market size, and HHI, or degree of concentration.

a) For each market-airline,

i) Calculate the average price.

ii) Calculate the total number of passengers.

iii) Calculate the average distance.

b) For each market,

i) Calculate the average price.

ii) Calculate the average distance.

iii) Calculate the total number of firms.

iv) Calculate the HHI, defined as the sum of squared market squares. E.g., if a market has two firms, each with 50% market share, the HHI is 50^2 + 50^2 or 5,000).

v) Download "populations.R". This dataset contains, for each market, the populations of the MSAs that contain the origin and destination airports. Merge this dataset with your market-level dataset. Calculate the size of each market, defined as the geometric mean of the populations of the endpoints.

4) Generate tables with summary statistics for each of your datasets and generate plots characterizing the distributions of market level prices and HHI as well as the relationship between them.

a) Report summary statistics for your tables (hint: use the kable function in the knitr package).

b) Using ggplot, generate a scatter plot of HHI versus prices at the market level. Also, using ggplot, generate density plots of market level prices and HHI. Be sure to label each of your three plots.