Homework Assignment 2

STA 141A

Due Tuesday, October 31 by 5:00 pm

Description

The U.S. Department of Transportation (DoT) records and publishes tables of information for public use. In this assignment, you'll analyze two tables from their Domestic Airline Consumer Airfare Report. Both tables contain directionless fare information.

The first of these, Table 1a, contains information for flights between pairs of airports. Only airports in cities with more than one airport are listed.

The second, Table 6, contains information about flights between pairs of cities. Specific airport details are omitted. Only cities that average at least 10 passengers per day are listed.

The datasets for this assignment are provided in a ZIP compressed file. The ZIP file contains:

- `airfare.csv`: The airfare dataset described above, in comma-separated values (CSV) format. Adapted from the U.S. Department of Transportation.
- `cpi_1996_2017.xls`: A consumer price index (CPI) dataset for transportation, in Microsoft Excel format. From the U.S. Bureau of Labor Statistics.

Questions

Use R to find answers to all of the following questions (that is, don't do any by hand or by point-and-click). Save your code in an R script. Try to complete at least one every day until the assignment is due.

1. Unzip and load the airfare dataset. Convert the columns to appropriate data types, then separate table 1a and 6 into different variables (to help you avoid double counting). You don't need to write an answer for this question, but please mark the code for this question in the appendix.
2. What timespan does the data cover? Do any quarters or years in that span have no data? Check separately for table 1a and table 6. In addition, check both tables for patterns in the missing values.
3. In 2017, which cities have the most connections to other cities? Which have the least? How do these results compare to 10 years earlier? 20 years earlier? Which cities have increased connectivity the most?
4. How has the approximate number of **total** passengers per quarter changed over the years? Create a graphic to show this and comment on patterns you see. Some quarters have a sharp decline in number of total passengers. What might explain these?
5. The average fares in the dataset are in nominal dollars (the actual price in dollars at the time). Inflation can confound conclusions based on nominal dollars over time. To deal with this, statisticians convert nominal dollars to real dollars. The conversion formula is explained at the end of this document.

   Load the CPI dataset. Create a new column `real17_fare` in the table 6 airfare dataset that has the average fare converted to real Q1 2017 dollars. You don't need to write an answer for this question, but please mark the code for this question in the appendix.
6. How have airfares changed over time? Use fares in real Q1 2017 dollars to investigate this graphically. Comment on patterns you see.

7. For 2015, what is the relationship between fare and distance? Use table 1a to investigate this visually **and** by using an appropriate statistical model or test. Comment on what you can infer from each and whether there is any disagreement. State the assumptions, use diagnostics to check whether they hold, and comment on how this affects your conclusions. Repeat your analysis with table 6. Comment on differences between your two results and why these occur.
8. Modify the model or test you used in the previous question to consider average daily passengers in addition to distance. Recheck the assumptions and then comment on what you can infer about the relationship between fare and average daily passengers. Is there any difference between the results for table 1a and table 6?
9. For 2015, identify city pairs where the carrier with the largest market share has fares below the average for that city pair. Investigate these using graphics, statistics, or models (as you see fit). Comment on patterns you find.
10. Use table 1a to compare Sacramento (SMF), Oakland (OAK), San Francisco (SFO), and San Jose (SJC). How do fares differ between these airports? Which airport has the most long-distance connections and how does this compare to the others? Do these results differ by year?

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted[1] tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 10 pages** including graphics, but excluding code and citations.

What To Submit

Submit a digital copy on Canvas. The digital copy must contain your report (as a PDF) and your code (as one or more R scripts).

Additionally, submit a printed copy to the box in the statistics department office[2]. The printed copy must contain your report and your code (in an appendix). Please print double-sided to save trees. It is your responsibility to make sure the graphics are legible in the printed copy!

Data Documentation

The airfare dataset contains the following features:

`year`	year
`quarter`	quarter
`city1`	descriptive label for `city_id1`
`city2`	descriptive label for `city_id2`
`fare`	mean fare in nominal dollars
`miles`	non-stop straight-line miles
`passengers`	mean passengers per day
`airport1`	IATA code for first airport
`airport2`	IATA code for second airport
`lg_fare`	mean fare in nominal dollars for carrier with largest market share
`lg_carrier`	carrier with largest market share
`lg_marketshare`	market share for carrier with largest market share
`low_fare`	mean fare in nominal dollars for carrier with lowest fare
`low_carrier`	carrier with lowest fare
`low_marketshare`	market share for carrier with lowest fare
`city_id1`	identification number for first city market
`city_id2`	identification number for second city market
`airport_id1`	identification number for first airport
`airport_id2`	identification number for second airport
`table`	name of table

[1] See the graphics checklist on Canvas.
[2] 4th floor of Mathematical Sciences Building

For more detailed information, see the original documentation provided by the Department of Transportation: `https://www.transportation.gov/sites/dot.gov/files/docs/FareReportMetaData.xlsx`

The CPI dataset is documented in its Excel file.

Relevant Functions

All of the functions from Assignment 1, as well as:

`sapply()`, `split()`, `tapply()`, `aggregate()`, `rep()`, `droplevels()`, `strsplit()`, `lm()`, `qqnorm()`, `qqline()`, `coef()`, `residuals()`, `fitted.values()`, `merge()`, `unzip()`, `read.csv()`, `read_excel()` in readxl

Nominal & Real Dollars

Consumer price index data can be used to convert nominal dollars to real dollars. The formula is

$$(\text{real } \$) = (\text{nominal } \$) \cdot \frac{\text{CPI}_b}{\text{CPI}_y}$$

where b is the base year and y is the nominal dollars year. For example, to convert nominal dollars from 1991 to real 2016 dollars, you would compute:

$$(\text{real } \$) = (\text{nominal } \$) \cdot \frac{\text{CPI}_{2016}}{\text{CPI}_{1991}}$$

This formula also works for other time scales such as quarters or months.

Based on: `http://itech.fgcu.edu/faculty/bhobbs/Nominal%20Real%20Price%20Index.htm`