Course Project

First draft due online through Canvas on Mar 28
Second draft due online through Canvas on Apr 25
Video presentation due online through Canvas on May 9
Final version due online through Canvas on May 16

1 Introduction

There are three main goals this project aims to achieve:

1. The primary goal is to have you practice data management, visualization and analysis techniques your learned in class using real life data.

2. The secondary goal is to practice writing full-scale data analysis reports, similar to the ones you will be preparing in our other classes or as part of your future job as a data analyst.

3. The final goal is to have you prepare and deliver a short presentation of your findings, as well as participate in the discussion of findings delivered by your classmates.

All students will be divided into groups (number of groups and size of each group will depend on the number of students in the class), and each group will be working on its own version of the project.

Important: most of the assignment description below is written with regards to analyzing merged dataset that consists of the data on sales of liquors in Iowa with Iowa's demographic and economic data available through American Community Survey. However, you are also allowed to analyze any other dataset of your choice, provided that it has at least 1000 observations across at least 10 variables. If your group does decide to go with an alternative dataset, make sure to confirm it with me before doing any serious analysis.

2 Data

All data files for the project are located inside a single archive "project.data.zip" available through Canvas. The table below summarizes the content inside each file:

File		Description
1	project.sales.zipcodes	Average annual liquor sales per zipcode
2	project.sales.cities	Average annual liquor sales per city
3	project.sales.counties	Average annual liquor sales per county
4	project.acs.counties	ACS data on Iowa counties
5	project.acs.cities	ACS data on Iowa cities
6	project.acs.zipcodes	ACS data on Iowa zipcodes
7	iowa.geographies	Full list of counties, cities and zipcodes in Iowa state

The first three files contain the same data on liquor sales in Iowa over 2012-2016 that you have already worked with in other assignments, except this time it contains full range of liquor categories. The variables in those files are as follows:

File		Variable	Description
1	project.sales.zipcodes	`zipcode`	Zipcode
		`category`	Liquor category
		`sale.dollars`	Average annual cost of liquor sold in dollars
		`sale.volume`	Average annual volume of liquor sold in liters
1	project.sales.cities	`city`	City name
		`category`	Liquor category
		`sale.dollars`	Average annual cost of liquor sold in dollars
		`sale.volume`	Average annual volume of liquor sold in liters
1	project.sales.counties	`county`	County name
		`category`	Liquor category
		`sale.dollars`	Average annual cost of liquor sold in dollars
		`sale.volume`	Average annual volume of liquor sold in liters

Files 4-6 contain ACS data on various economic and demographic variables across Iowa geographies. Each file contains a unique variable that defines the geography (`zipcode`, `city` or `county`) and then the following variables common across all 3 files:

Variable	Description
high.school	Percent of population, high school graduate or higher
bachelor	Percent of population, bachelor's degree or higher
unemployment	Unemployment rate, population 16 years and over
income	Median earnings, dollars
population	Total population
pop.white	Total population, white
pop.black	Total population, black
pop.indian	Total populationm, American Indian and Alaska Native
pop.asian	Total population, Asian
pop.hawai	Total population, Native Hawaiian and Other Pacific Islander
pop.other	Total population, other single race
pop.multi	Total population, two or more races

Last file shows a mapping between Iowa counties, cities and zipcodes, in case you would like to pinpoint certain locations without using Tableau's mapping features.

3 Research agenda

There are lots of potential issues one can look into using merged data. Below is a list of questions that can be analyzed using appropriate visualizations of the merged data across all 3 geography levels (counties, cities, zipcodes):

1. What is the distribution of per capita sales across geographies? What are the ranks of top 10 geographies for per capita consumption across every liquor category? Are there any outliers, i.e. locations that have high per capita sales in only one specific liquor category?

2. What about pairwise patterns in total/per capita sales across geographies? E.g. what if you plot average sales per zipcode vs average sales in corresponding city? Does the pattern differ across liquor categories?

3. Do geographies with higher median income consume more alcohol in total? What about per capita? Does it depend on liquor category?

4. Does employment affect what liquor categories are sold most? Is the pattern the same as the one for median income?

5. Are there any preference patterns for liquor consumption among different races? What are most popular liquor categories among geographies with highest share of minorities?

6. Is there any notable difference in liquor sales across communities with varying levels of education?

You are also encouraged to analyze any other questions and look for any other interesting patterns in merged data.

4 Instructions

1. Your project should include the following major pieces:

   - Analyzing and visualizing ACS data using Tableau and/or R.
   - Analyzing and visualizing aggregated sales data using Tableau and/or R.
   - Merging aggregated liquor sales data with ACS data per each geography (zipcodes, cities, counties) using R and/or Tableau.
   - Visualizing and identifying patterns in liquor sales across geographies and ACS metrics using R and/or Tableau.
   - Summarizing your findings in a short video presentation and participating in discussion of your findings and findings of other students.
   - Writing a detailed report in a form of a data analysis paper.

2. While it remains up to you to decide on what exactly you are going to submit in your draft versions, it is recommended to do it as follows:

   - Your first draft should contain a short Word/PDF document with data summary for two datasets separately (ACS and aggregated sales), as well as some key visualizations for each dataset (e.g. map of Iowa counties with median income from ACS data or distribution of average annual sales per county from liquor sales data). You should include a few paragraphs of text explaining how these elements align with your assigned research questions.
   - Your second draft should contain preliminary visualizations for merged data (e.g. one key chart per each of the required questions from research agenda) and a few paragraphs of text describing them.

3. When merging ACS and annual sales data, make sure to use full outer merge, as it is possible that not all ACS geographies have recorded liquor sales and/or not all sales geographies have corresponding matches among ACS geographies.

4. **Important:** you should have 3 merged datasets as the result, one for each geo level of ACS data, e.g. 3 Tableau workbooks or 3 data objects in R. **DO NOT** merge all data into one Tableau

workbook — even though it may seem like it is working correctly, it will produce completely wrong results.

5. Canvas submissions for both draft versions should include only your Word/PDF paper with whatever findings you have. Do not submit any data files and/or scripts for your draft versions.

6. Canvas submissions for the final version should include three things:

   - R scripts for merging and aggregating data, if you use R to do it.
   - R scripts and/or Tableau workbook with data visualizations.
   - PDF or Word document with charts, tables, data summaries and your analysis.

   Do not upload any data to Canvas for your submissions. I will generate the data from your R scripts and/or Tableau workbooks.

5 Project Paper Structure

The final version of the project paper should contain the following parts:

1. **Introduction.** Here you should explain the goals of the project and briefly outline any key expectations you might have (e.g. "Expecting higher sales of alcohol in areas with higher income per capita").

2. **Data summary.** This section should contain a short description of the nature of two datasets you are using, key numerical characteristics and some visualizations of each of the two datasets. At the minimum, it should include:

   - Name, type and units of measurement for all variables
   - Basic numerical stats for each variable (min, max, median, mean)
   - Relevant/interesting visualizations for some variables (histogram and/or bar chart and/or box plot), if there are some noteworthy patterns.

3. **Research agenda.** Here you should describe the research question(s) that were assigned to you, and explain how you are going to approach them using the available data.

4. **Data analytics.** Here you will present data analytics pieces that show your answers to the questions outlines above. These primarily should take the form of charts (scatter plots, histograms, bar charts, etc.), but you may need some tables as well. There are no specific requirements for this part, as you are free to choose any form of data analytics to support your findings. However, when putting everything together, make sure to follow these guidelines:

- All charts must have clearly outlined titles, units of measurement on both axis, legend codes (if any). All variables should have proper names such as "Sales per capita" instead of "sales.p.c".
- Avoid putting several similar charts next to each other (i.e. patter of sales for each liquor category). Either make a single chart with multiple elements (i.e. a box plot for each category) or show only charts with most notable patterns, putting the rest into Appendix.
- Avoid having long white spaces before/after charts. Either make them small enough to fit on the page along with text description, or rearrange your paragraphs of text so that they fit nicely above and below your chart.

5. **Conclusion.** Here you should outline your key findings, as well as whether your expectations described in introduction were fulfilled.

6. **Appendix.** Here you should put all visualizations and tables that are relevant to your analysis, but are not important enough to be put into main section. For example, in the main body of your paper you may show two charts for most interesting liquor categories, while the same charts for all other categories should be put into Appendix.