ISTE 782 Homework Instructions

Mick McQuaid

Fall 2018

Intro

There are six homework assignments due in this class.

i. Tableau
ii. Describe the 311 data
iii. Explore the 311 data
iv. Tidy the 311 data
v. Join other data to the 311 data
vi. Communicate your findings

Data sets

You will use a total of three data sets in this course, one of which will be provided by the instructor. One of your most important assignments is to identify two good datasets. The first one must be found very quickly since you must visualize it by the end of week 2.

The second data set is the New York 311 data set, a version of which will be provided on `serenity.ist.rit.edu/~mjmics`. Use the provided version rather than a version you may be able to obtain online. You can learn more about this famous data set in various articles, including one by Steven Berlin Johnson at wired.

The third data set must be linkable to the New York 311 data set so it must have a column of NYC boroughs. Otherwise, you are free to identify and use any data set that meets this minimum criteria.

General instructions

Part of your homework grade will depend on following these instructions carefully. These instructions serve two purposes. First, they make it easier for the instructor to concentrate on what you did right or wrong by removing variables like formatting. Second, they give you practice using a simple format for R reporting.

You may work in pairs on any assignment as long as you document your names and both turn in identical files. If you claim to work in a pair and turn in different files, I will question whether you are on the same page as a pair.

You will most likely use RStudio to do homeworks ii through vi but, if you wish, you may use the copy of a virtual machine located at serenity.ist.rit.edu/~mjmics. This virtual machine has a copy of R and all the packages used in this class preloaded. It will give you experience with working on a virtual machine which you may need in practice if you have to work with large data sets. RIT has a license for VMWare you can use as students to host this virtual machine. In practice, you would likely use a virtual machine hosted remotely, often via Amazon Web Services or a similar service, but the mode of interaction is the same. I am enthusiastic about the use of virtual machines and am happy to help you set this up.

Instructions for homework i

On the first day of class, we will do an exercise together with Tableau. To demonstrate your understanding of Tableau, you will do something similar to what we do in class, but with your own dataset.

First, obtain a dataset. You will turn this in. If it is too large to fit in myCourses, you will submit an explicit link. Be aware that I will deduct substantial points from your grade if you make it too difficult for me to find your data. The dataset should be at least as large and rich as the *Superstore* dataset we use in class.

Second, create some visualizations in Tableau, analogous to the visualizations we create in class with the *Superstore* data. Do not necessarily use the same selection of visualizations but rather identify those appropriate to your data.

Third, create an interactive dashboard like the one that we create in class, so that it is possible for the user to explore the data from different perspectives.

Fourth, create a series of storypoints like the ones we create in class, allowing the user to follow the narrative of discovery you've constructed through your visualizations and dashboard.

Fifth, answer the six discussion questions at the end of our Tableau video, also to be found in the Tableau workbook. Put these answers in a plain text file, not in a word processing document. Be aware that I will deduct substantial points if you turn in a word processing document that I can not open with a plain text editor.

Turn in a zip file called `i.zip` containing your Tableau workbook, named `i.twb`, your data file, named `data.csv`, and your answers to questions, named `i.txt`. Do not use any subdirectories in your .zip file. Be aware that I will deduct substantial points from your grade if I have to hunt through a directory structure to find your files. myCourses will take care of adding your name to your submission.

Instructions for homeworks ii through vi

R Markdown Files. For each homework assignment, you will turn in one or more files. You will always turn in an R markdown file and sometimes you will turn in one or more graphics files. Here's how to format the text files.

0. Zip any files you turn in. There should be a single zip file named with the number of the homework as a lowercase roman numeral. For example, homework ii should be enclosed in a zip file called `ii.zip`. MyCourses will add your name (plus a great many numbers) to the file name automatically.

1. Use a full-featured text editor that saves your files as UTF-8 or use R Studio.

2. Save the files with the extension `.Rmd` and make the file name match the roman numeral of the assignment but in lower case. For example, the second assignment will be called `ii.Rmd` and so on.

3. The beginning of the file contain the following info: the homework number, your name, the date, and the output format. For example, for the second homework, my file would begin as follows (note that each line begins flush left).

```
---
title: 'Homework ii'
author: 'Mick McQuaid'
date: '2018-09-14'
output: pdf_document
---
```

7. Whenever your answer includes R code, write the code flush left surrounded by code fences. A code fence consists of a line with three backticks flush left and no other text except the letter r in curly braces (you can put certain other code inside the curly braces as well and you will learn that later) for the opening code fence and none at all for the closing code fence. Here is an example.

~~~
```{r}
pairs(Auto)
plot(mpg,cylinders)
sapply(Auto[,3:7],mean)
```
~~~

8. If you want to include mathematical expressions, you may write them in LaTeX and surround them with dollar signs. This will allow you to say things like $\widehat{\mu}_0$. If you cannot understand LaTeX (it will be demonstrated for you and examples will be given), you can write the expressions out phonetically. For example, the above expression is pronounced *mu hat nought* and is written in LaTeX as `$\widehat\mu_0$`.

9. R markdown is documented in Chapters 21, 23, and 24 of our textbook as well as in other publications we will discuss. You will need to familiarize yourself with this format to some extent.

Graphics files. The homework specification often calls upon you to produce graphics. I may even add more graphics requirements to some of the assignments. For really basic graphics such as those in the first assignment, I will be content with the R code that produces the graphics. In some cases, I will require you to submit graphics files with your output.

In such cases, each graphic should be saved in a pdf file produced by R. It should be possible to refer to these files within your `.Rmd` document by using the following syntax. In this example, `description` is the caption for the picture and `iva.pdf` is the name of the graphics file. You should discuss any graphics files you create in your `.Rmd` file, whether you include a copy of the separate file or not. If you fail to mention a graphic file, I will have to assume that you do not understand what it represents.

```
![description](iva.pdf)
```

hw ii. Describe the 311 data

For this assignment, you will give a preliminary description of the 311 data. It should include pictures and tables and a data dictionary and be presented in an r markdown document called `ii.md`. It should be included as the only file in a .zip file called `ii.zip` and there should be no subdirectories in the .zip file. You will need to investigate the nature of the data using whatever means you can think of, such as googling. (A data dictionary gives definitions of columns and any specifications of limitations of what can be entered in the column. For example, a column containing zip codes consists of either exactly five digits or exactly nine digits with a dash between the fifth and sixth digit. A column containing borough names consists only of the following entries: Bronx, Brooklyn, Manhattan, Queens, Staten Island.)

HW III. Explore the 311 data

For this assignment, you will conduct an exploratory data analysis of the 311 data. This is a much deeper dive than the previous assignment. Here you will look for connections between columns. These questions could include, for instance, what complaints are most associated with which agencies or which boroughs generate the most complaints of each major category. You will be graded on the questions you raise as well as the answers you discover. Turn in an R markdown file recording your exploration.

HW IV. Tidy the 311 data

The 311 data contains many infelicities, some of which can be corrected by `tidyr`. For this assignment, you will improve the 311 data and introduce another related data set, which may also require the use of `tidyr`. The related data set should have a column of NY boroughs or be connectable by some other means to the 311 data set. You will turn in an R markdown document showing what you did to prepare the data. What you did should be reproducible using your file.

HW V. Join other data to the 311 data

For this assignment, you will connect your other data set to the 311 data set, using `dplyr`. As usual, you will turn in an R markdown file showing what you did to connect the files, as well as a short table or tables consisting of an extract of the data, and a data dictionary for all the data. You will continue to explore connections between columns but you need not report on the new connections until HW vi.

For this assignment, you will communicate your findings through a polished R markdown report. These findings include the connections you made in previous assignments as well as new ideas you incorporate as a result of additional exploration. Unlike your previous assignments, this assignment requires you to develop polished titles, legends, axes, and scales for your plots and to comment on them in a superb narrative, free of spelling and grammatical errors. This should be a complete report on the 311 data (and your other data set) that you would be proud to show in a portfolio or to a potential employer. Include the data dictionary from HW v as an appendix. Be aware that you are communicating findings not exploring in this report, so omit any dead ends you may have included in HW iii.