## SI 618 Homework 7

**Assignment Instructions**

This homework has two parts. In part 1, you'll prepare data using Python.  In part 2, you'll load the data into R to do some aggregation and visualization.

For Part 1, you'll need to download a Yelp dataset, which is available for class use only in an Mbox account (link below).  The dataset is over 100Mb compressed, and 300Mb+ uncompressed, so it will take a few minutes to download. The Mbox link is here:

<https://umich.box.com/s/1lc3o4ib3mgybo86r04ginn90rm0pukx>

To decompress the file yelp\_academic\_dataset.json.gz, you should be able to use standard decompress tools, like gzip (Mac/Linux) or the default Windows compressed file utility.  The decompressed file is in JSON format.

Next, download and unzip the attached homework zip file. Your ultimate goal is to generate a report that looks like 'si618hw7\_sample\_report.html'. The 'businessdata\_desired\_output.tsv' is the desired output of your data preparation step.

**Part 1. Data Preparation (35 points)**

Write Python code to extract data about all businesses from the uncompressed file yelp\_academic\_dataset.json. Note that "business categories" are actually a list in the JSON file, so we'll just use the first element of this list as the main\_category for a business. The output of this step should be a TSV file named businessdata.tsv that contains the columns:

name, city, state, stars, review\_count, main\_category

for each business. If there are missing values, output 'NA' instead. For unicode issues, refer to [http://docs.python.org/2/howto/unicode.html](http://docs.python.org/2/howto/unicode.html" \t "_blank)

We've provided the desired output in the file businessdata\_desired\_output.tsv

**Part 2. Data Aggregation and Visualization in R (65 points)**

Now that you have your extracted data file, you should write an R Markdown document that will generate a report looking like the provided si618hw7\_report\_sample\_output.html. You should use the qplot() or ggplot() functions in the ggplot2 package and either data.table syntax or functions in the plyr package to get these results. The steps involved and points for each are described in the provided si618hw7\_report\_sample\_output.html.

**What to submit:**

A zip file named 'si618hw7\_youruniquename.zip' containing:

* Your Python code for Part 1 named 'si618hw7part1\_youruniquename.py'
* The output file of your Python code for Part 1 named 'businessdata.tsv'
* The R Markdown file you wrote for Part 2 named 'si618hw7report\_youruniquename.Rmd'