

Project Instructions: Data Mining with R AND SQL!!!

The purpose of the course project is to demonstrate your “hacking skills” and ability to work through the analytics pipeline from beginning to end by *integrating both software tools* introduced in the course. To do this, we will have you utilize a special R package called ‘*sqldf*’ that will allow you to run SQL queries from directly inside R/RStudio. You will also familiarize yourself with the ‘*ggplot2*’ package to aid in your analysis of visualizations.

The first part of the project entails finding an appropriate dataset, cleaning that dataset, and mining the dataset using your barrage of descriptive tools. You will need to find your own business / economics relevant dataset either by gathering your own data or sourcing data from publicly available information. You should conduct a thorough exploratory analysis using the programming skills developed thus far by leveraging R, RStudio, SQL, and the ggplot2 system (package). The end-product should result in a TIDY dataset, a report explaining your preprocessing, cleaning, exploratory (descriptive) analysis (including your SQL queries), and a detailed description of the steps taken along the way.

- Groups – you may self-select into a group with no more than 4 team members total.
- Data Collection: Each student group will collect a business or economics relevant dataset - *you must get approval from the instructor before proceeding to the steps below.*
 1. The data can be observational data found online (recommended) or experimental data generated or collected by you and your team (the latter is more difficult, time consuming, and is not recommended, but will be rewarded on the back end with more leniency in grading).
 2. It is highly recommended that you choose a cross-sectional dataset (or collapse a panel/longitudinal dataset into a cross-sectional dataset) but other data structures should also work.
 3. Your data should have *at least 5 variables (fields)* total in addition to primary/foreign key fields. At least three of these should be numeric (discrete or continuous) and at least two of these should be categorical. It is to your advantage to use more variables so that you can easily meet the criteria below regarding the number of visualizations and queries.
 4. You should have *at least 100 records* (observations/records) total.
 5. You should have *at least two separate tables* (that could in theory be merged using primary/foreign keys later) so you can demonstrate your ability to join (either by merging the datasets or running a join query). (hint: if you downloaded a single flat file dataset and parsed/separated it into multiple tables that are TIDY, that would also do the trick!)
 6. You need to get instructor approval of your dataset(s) to make sure you are setting yourself up for success (I will provide time in class to do this).
- You’ll need to install the ‘*sqldf*’ and ‘*ggplot2*’ packages (make sure you also install the dependencies) once and will need to call the library each time you start the RStudio session. Running queries (that don’t require you to create or insert records) is as easy as creating a

string variable with the query information inside of it and then passing the variable to the `sqldf()` function to run the query (resulting output will show up in the console window below). Using the `ggplot2` package will require some more “hacking.” The instructor may provide a brief video segment showing an example of using `ggplot2` (with a code chunk that will be made available to you).

- Here are some places to find datasets (there are *many many* more):
 1. Government Open Data: <https://www.data.gov>
 2. Gapminder: <https://www.gapminder.org/data/>
 3. Federal Reserve of St Louis: <https://fred.stlouisfed.org/>
 4. Penn World Tables: <https://cid.econ.ucdavis.edu/pwt.html>
 5. Yahoo Finance: <https://finance.yahoo.com>
 6. UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets.php>
 7. Kaggle: <https://www.kaggle.com/datasets>
- In case you would like a longer, more comprehensive list of data sources, here is one: <https://www.kdnuggets.com/2017/12/big-data-free-sources.html>.
- **Deliverables** - will submit four items (submission details below):
 1. Your group’s R code used in your analysis (which should include your visualizations build in the `ggplot2` system along with SQL queries using the `sqldf` package).
 2. A written report summarizing your process, all of the steps taken in cleaning the data, and findings from your exploratory analysis that integrate the statistical output from R directly into your report (you can screen shot and paste or export diagrams and paste).
 3. The TIDY data that you cleaned up (the end product of cleaning).
 4. The raw dataset that you started with (the starting product).
 5. The items above should be emailed per the submission instructions below (provided they are not excessively large – specifically items 3 and 4).
- The report should contain R code, output, as well as written narrative to describe and document the steps taken in your analysis from start (cleaning / preprocessing) to finish. Here is a rough outline for how to structure the report (should not exceed 10 pages):
 1. The first section should include an executive summary walking through the data collection process, describing the variables of interest, and discussing the overall structure of the dataset.
 2. The second section should include a carefully documented set of instructions of steps taken in preprocess and cleaning the data. Did you have to deal with missing or null values? Did you have to rename column headers or recode data? Did you have to rescale or transform any variables? Include any cleaning related exploratory visualizations here.
 3. The third section should detail your exploratory analysis on the cleaned data. You should leverage the `ggplot2` package to develop several (at least 6 unique) visualizations that tell provide a visualized narrative regarding “what happened” in your data. How are your data (variables) distributed? Are there any obvious relationships between variables in the data? How strong are those relationships? Did you use a smoother to show the relationship more clearly? By unique visualizations,

I mean to not only use the same plot device. For example, a scatter diagram and a bar histogram would be considered two unique visualizations, but using 6 different scatter plots would not suffice (would count as one unique visualization). Did you use facet wrappers? Did you color your plots appropriately? You should also be able to run at least 12 queries to demonstrate your ability to use SQL in the context of invoking the *sqldf* package. The queries should be designed to help tell the narrative of “what is going on in your data” alongside with the visualization you create (hint: associating the SQL query results with the visualizations is a good idea). You should explicitly include the syntax for the queries that you run (we will also see it in your code, but it should also be included explicitly in the report).

4. The final section should wrap up your results with a conclusion and discussion of lessons learned along the way.
- Your grade will be based on how well you accomplish the tasks above as well as the visibility of individual group member contributions. Your grade will be a weighted average based on the quality of your report and presentation performance.
 - You should email your **deliverables** to slevkoff@sandiego.edu and CC carlr@sandiego.edu. In the subject heading of the email, include “BUAN 370 DATA MINING PROJECT F21” and include the associated deliverables as email attachments (one submission per group). The deadline to submit the report for the course project will be Friday, 12/10/21. Your group will also make a presentation reviewing your analysis to be presented during the final exam period for your section and should not exceed 20 minutes.