ECON 494

Dr. Levkoff

Project I: Descriptive Analytics Data Mining Project

The purpose of the two course projects is to demonstrate your ability to work through the analytics pipeline from beginning to end. The first project entails finding an appropriate dataset, cleaning that dataset, and mining the dataset using your barrage of descriptive tools, and then pushing the deliverables to your cloud repository to demonstrate your working knowledge of repository management. You will need to find your own business relevant dataset either by gathering your own data or sourcing data from publicly available information. The first project for the course involves cleaning the data and conducting a thorough exploratory analysis using the programming skills developed thus far by leveraging R and the ggplot2 package. It is highly recommended that you choose a cross-sectional dataset for this exercise with the prediction (project II) portion of the project sequence in mind. The end-product should result in a TIDY dataset and a report explaining your preprocessing, cleaning, exploratory (descriptive) analysis, and a detailed description of the steps taken along the way.

- Data Collection: Each student will collect a business or economics relevant dataset -you should clear this with the instructor before proceeding to the steps below – you will use the same dataset to "piggyback" the second project off of your results of the first, so successful choices early on set you up for a successful project (II) later. 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 and time consuming, but will be rewarded on the back end with more leniency in grading). It is highly recommended (with project II in mind) that you choose a cross-sectional dataset (or collapse a panel/longitudinal dataset into a cross-sectional dataset).
- 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 R code used in your analysis.
    2. A written report summarizing your process, all of the steps taken in cleaning the data, and findings that integrates the statistical output from R.
    3. The TIDY data that you cleaned.
    4. The raw dataset that you started with.
    5. The above items should be posted to a *private* GitHub repository. You will need to change the repository settings to add myself as a collaborator on the repository so that I can access your work. You can find info. on how to make a GitHub repository *private* here: https://docs.github.com/en/enterprise/2.13/user/articles/setting-repository-visibility. Here is a link with instructions on how to invite a collaborator: https://docs.github.com/en/free-pro-team@latest/github/setting-up-and-managing-your-github-user-account/inviting-collaborators-to-a-personal-repository. You will need my GitHub user name to add me as a collaborator: slevkoff.

- 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:

    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 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?
    4. The final section should wrap up your results with a conclusion and discussion of lessons learned along the way.

- You should upload your **deliverables** to a GitHub repository and provide the link to that repository to slevkoff@sandiego.edu. Make sure you've added slevkoff as a collaborator on the *private* GitHub repository per the instructions above in step 5) of the deliverables section. In the subject heading of the email, include *"PROJECT I ECON 494 F20*". The deadline to submit the project is Sunday, 10/25/20, before midnight.