Data Science Decal, Fall 2017
Project 1: Data Cleaning & Visualization

Machine Learning at Berkeley

1 Introduction

In this project, you will explore a new dataset and build a complete end-to-end pipeline for data visualization, from organizing and cleaning the data to visualizing its various aspects. Typically in any real-world use case, you would have the problem presented to you (such as "How are chronic diseases in the U.S. tied to socioeconomic mobility?"); you would then collect data related to answering this question and start exploring it to find answers. However, in order to ensure quality datasets, I have made a short list from which you may decide and pick (next section).

You may use any python libraries you wish for this project. You will be given raw, real-world data and be asked to follow the checkpoints for cleaning/visualization/creating regressions on the data. We will not ask you to do anything specific - it is up to YOU to figure out what methods and avenues best fit.

There will be 2 components to this project: the jupyter notebook that contains all the code you used to manipulate the data, and a 2-page typed writeup explaining the major decisions you made and your biggest findings. The writeup will be the main component of this project, with the notebook simply to verify what you state. Don't freak out about the writeup; I anticipate it being a relatively simple task once you have the notebook. YOU NEED TO WORK IN GROUPS OF 3-4.

2 The Datasets

Your first task is to choose and download one of the datasets listed below. These are all datasets available on the public domain, and they all have a certain degree of complexity to them - some of them have data spread across multiple files, others have documentation that needs to be read to fully understand the data, and yet others have data that is really oddly formatted and stored. I have chosen these datasets on purpose to allow you to illustrate your data cleaning and organizing skills.

- Food Environment Atlas - contains data on how local food choices affect diet in the US. Here is the page for the dataset. Download the dataset here. Here is the accompanying documentation.
- School System Finances - a survey of the finances of school systems in the US. Here is the page for the dataset. Download the dataset here. Here is the accompanying documentation.
- Word Development Indicators - contains country level information on development. Here is the page for the dataset. Download the dataset here. Here is the accompanying documentation.

3 Setup

First download your dataset. If it's not in .csv format, you may have to save it as a .csv file. The dataset may be an excel doc with multiple sheets, in which case you would have to save multiple .csv files. Then create a new jupyter notebook, and load all your dataset files. At the top of your notebook, please note which dataset you have chosen! Also document in your writeup any important extra steps you needed to do before you could read your data into a dataframe in python.

4 Understanding

Once you have loaded your dataset, you first need to get somewhere so that it's easy to work with (like a jupyter notebook!). As you do so, you should explore your dataset. Create overviews, make some

summary statistics, and explore any documentation that may have come with your dataset to really understand it and its structure. Make sure to note down in your writeup any important steps/realizations here.

5 Cleaning and Organizing

Now, based on what you have understood about your data, you should clean and organize it (so that it's easier to create visualizations/fit models later). This may be the most confusing part of the project, but at the end of this process, you should have one perfectly clean dataframe that you can easily create plots from.

First organize it. Don't be afraid to throw out large parts of your dataset if you feel it is not relevant information. Combine the multiple frames you may have into one standardized complete view of your data. Basically, do all the restructuring you need here. Note important parts of this in the writeup.

Now you should have all data you think is relevant. Explore if this data is usable; if it has too many missing values, you may need to drop those parts. If there are not too many missing values in some parts, you may want to explore how to impute the missing values with workable data; some options we talked about in homework was using a simple average, using a linear regression to estimate values based on other factors, or dropping the data entirely. Note any steps you make in your writeup.

6 Visualization and Regression

This is the most fun part of the project - all the other steps were leading up to this! Now you can go ahead and create your graphs, models, and other visualizations. Feel free to use any libraries to create interesting graphs. Note that each graph should be properly labeled with x and y units, proper x and y scales, a title, and a legend if necessary.

Make sure you demonstrate your ability to create rich graphs by creating various different types of graphs, both for numerical as well as categorical data. One way to create an interesting story from your plots is to compare one aspect - for example survival rate from the titanic dataset - with all other features, such as age, embarking town, class, etc and see how the breakdown differs. You can also create stories by breaking down plots into smaller and smaller categories - for example when plotting embark town vs survival rate, you can further split each embark town into categories for gender.

Make sure you also demonstrate your ability to create regressions on your data. Include at least one linear regression and one logistic regression. Make sure you document why you thought a certain type of regression would fit your data better and what use your model could have for predictions. How accurate were your regressions? Why do you suspect they performed so well/ineffectively on that certain data? Is there a way you could make the regression better (optional: try adding regularization to see if you could get better results). Dedicate a non-trivial portion of the writeup to this subsection.

Also remember to document any work you do here in your writeup! Include important graphs and plots you have created, as well as a thorough discussion to what you have discovered about the dataset through these visualizations.

7 Submission

Remember your write-up must be at least 2 pages in length. Please include the pictures of your visualizations/regressions in it to illustrate points you talk about! However, these pictures are not counted in the 2-page requirement: you must have at least 2 pages of text. There is no upper limit on how long the report can be. You must submit your both your writeup and jupyter notebook. Save them both as PDFs and attach them together, with the jupyter notebook at the end. Make sure to note your group member's names at the top of your writeup. Only one team member should upload the PDF to Gradescope.