MSIS 2607 - Winter 2020 - Project 2

Logistics:

Assigned: Thursday, February 11, 2020
Due: Thursday, February 29, 2020

Objective:

Perform a data analysis on a data.gov dataset using Jupyter Notebooks.

The requirements for this project are:

- Select two related datasets from data.gov, Kaggle, or other open source datasets.
- Choose datasets which no one else has (do your best to avoid using someone else's dataset).
- Merge the two datasets using an appropriate Pandas merge or join. For example, the Happiness Index (happiness scores by country) could be inner joined with a dataset of GDP by country (or other country level data). The analysis could them look at happiness by GDP or happiness by country population.
- Use Jupyter Notebooks and submit the .ipynb file
- Clean the data to deal with missing entries after the join, and explain your approach. If dropping rows or columns, explain your reasoning. If imputing values, explain why you chose the imputation values.
- Find 3 interesting facts/patterns about the dataset and present your findings with the use of graphs. Each finding should have a concise description, and a chart to show or justify the conclusion.
- One of your findings must be done with the use of a decision tree, using the decision branches to guide your analysis. You may use a different machine learning model instead of a decision tree.
- Tell a story with the data. Make sure your findings fit into an overall narrative.

Guidelines for judging 'interesting':

There are multiple ways things get to be 'interesting'. Here's two of the best heuristics we know:

- This fact/pattern is so interesting you would go to a party and say: "Guess what I found out about xyz!"
- This fact/pattern is crucial to understanding the topic: e.g. For gerrymandering, that would be something like 'There are x many districts, that are up for debate every y years, and z are the decision-makers. If *a* many districts shift to red/blue, then the odds of the election swaying one way is *b%* higher.' By the way, this would be one of the three sections - not all three in one.

Storytelling & visualizations:

- Present the data in a manner which draws people in and keeps them engaged
- Be concise, clear, concrete, correct, coherent, complete, and courteous (7 C's of communication)
- Use comments for code, and Jupyter elements for storytelling
- Pictures are worth a thousand words. Use them to distill complicated data into an easily graspable chart or table.

Resources:

- https://catalog.data.gov/dataset
- http://jupyter.org/
- https://matplotlib.org/
- https://seaborn.pydata.org/
- https://pandas.pydata.org/
- https://www.mindtools.com/pages/article/newCS_85.htm
- https://datavizblog.com/2013/05/26/dataviz-history-charles-minards-flow-map-of-napoleo

  ns-russian-campaign-of-1812-part-5/

Collaboration:

You will work individually on the assignment. You are allowed and encouraged to use Google extensively.

Submission:

- Name your final file <your_username>_project2_winter2020.ipynb (mine would look like dvrdoljak_ project2_winter2020.ipynb).
- Make sure it runs completely and correctly on your computer
- Submit it via Camino
- (We will run your program on our computer to test your answers)
- Include a link to your datasets. If you performed significant data prep, you should upload your prepared data to a Google Drive folder and share the link in your submission.

Grading Rubric:

Section	Grade	Criteria
Interesting Fact 1	15%	Interestingness, factfullness, analysis, presentation, data preparation
Interesting Fact 2	15%	Interestingness, factfullness, analysis, presentation, data preparation
Interesting Fact 3	15%	Interestingness, factfullness, analysis, presentation, data preparation

Data Preparation and Merging	20%	Correct and appropriate merge of two datasets. Missing values correctly address, and reasoning explained clearly in notebook.	
Use of Machine Learning		20%	Correct use of a decision tree (or other ML model) for analysis
Use of comments & Readability; General & Submission		15%	Documentation of author & dates; Explanation of steps Use of whitespace; Use of new lines; Naming convention of variables; Sequencing of code and outputs Directions followed correctly, code is correct and error free