Project McNulty MVP for Gray Davidson **5/4/2018**

Description:
For McNulty I will be working with a social psychometric dataset combining the "Big 5" personality traits, also called the "five-factor model", with Angela Duckworth's Grit scale and considerable demographic data such as Age, Gender, Religious Beliefs, Native Language, etc.

My preference for this dataset springs from the domain and from the structure of the data. Not only are there a variety of interesting questions to ask with the data but the component parts are well stated (one measure might be "E7 I talk to a lot of different people at parties." This is in contrast to "feature_042" which was the extent of the labeling present in many online datasets I investigated. My background also touches tangentially on social psychology and I have read Angela Duckworth's book about her investigations of "grit."

Data:
A summary of the data columns and types:

RangeIndex: 4270 entries, 0 to 4269
Data columns (total 98 columns):
dtypes: int64(95), object(3)

To fit this in a reasonable space I have removed about 40 columns of the form "GS5 4270 non-null int64" but as you cas see above there are 98 total column.

country 4226 non-null object	married 4270 non-null int64
surveyelapse 4270 non-null int64	familysize 4270 non-null int64
GS1 4270 non-null int64	E1 4270 non-null int64
GS2 4270 non-null int64	E2 4270 non-null int64
GS3 4270 non-null int64	E3 4270 non-null int64
GS4 4270 non-null int64	E4 4270 non-null int64
… … ...	… … …
VCL13 4270 non-null int64	O4 4270 non-null int64
VCL14 4270 non-null int64	O5 4270 non-null int64
VCL15 4270 non-null int64	O6 4270 non-null int64
VCL16 4270 non-null int64	O7 4270 non-null int64
education 4270 non-null int64	O8 4270 non-null int64
urban 4270 non-null int64	O9 4270 non-null int64
gender 4270 non-null int64	O10 4270 non-null int64
engnat 4270 non-null int64	operatingsystem 4270 non-null object
age 4270 non-null int64	browser 4270 non-null object
hand 4270 non-null int64	screenw 4270 non-null int64
religion 4270 non-null int64	screenh 4270 non-null int64

orientation 4270 non-null int64 race 4270 non-null int64 voted 4270 non-null int64	introelapse 4270 non-null int64 testelapse 4270 non-null int64

Note that the int64 data for the majority of columns was submitted on a Likert Scale which ranges from 0 (strongly disagree) to 5 (strongly agree).

Classifier:
I would like to get a model running and then play with the outcome variables (as I have several demographic classifiers available) to decide which to use as my classifier. (Or perhaps there will be interesting stories to tell from the data for several classifiers).

Unknowns:

A) One known unknown is the scaling of the Likert scale features.
B) Another is whether I will need to collapse the individual big-5 features and grit-scale features into aggregate test scores *before* entering them into the model.
C) One big unknown is what model to make use of and I've been very thankful for the discussion on Thurs./Fri. of how to motivate model selection. It is one of my personal goals for this project to make choices with good reasons rather than simply because they are the available methods or code blocks. (i.e. what should K be in my k-fold cross-validation?).

Intermediate Moonshot:
I intend to explore the extra tools available to us in this section of the course so even though I do not need to use RDS/SQL for this work I intend to do so anyway to practice these resources.

Moonshots:
Use Flask.
Use a more complex visualization tool (D3 or other) to augment my presentation.
Use a more complex presentation tool to create an interactive presentation.