Statistics 5100 Spring 2020 Project 2

Notable Due Dates:

Report (First Draft):	March 11
Peer Review:	March 20
Proof of Writing Fellow Review:	March 24
Report (Final Draft):	March 25

This project is intended to give you experience designing and implementing regression models on a real dataset while formally communicating your results in writing. Unlike Project 1, you are not required to follow a specific paper template.

Data:

Tinder is a mobile dating app that allows users to locate potential dates in their local area. A 2017 publication, "Where Dating Meets Data: Investigating Social and Institutional Privacy Concerns in Tinder" in *Social Media + Society*, reported on results from a survey of nearly 500 US Tinder users. The authors of that paper, Lutz and Ranzini, have graciously made their data available for use in this STAT 5100 course (but you do *not* need to look at that paper to do this project). These data have been organized in the tinder.xlsx file on Canvas.

Responses to the authors' questionnaire have been combined within several categories of questions to give composite scores on some of the variables in the table on the following page. Unless coding levels are defined in the table, variables are coded on an increasing scale, so that higher values mean "more".

Of interest in this project is whether and how various demographic, psychological, motivational, and opinion characteristics affect (or at least predict) how genuine a person's self-presentation is on Tinder.

General Directions:

You will use the multiple linear regression methods discussed in this course to analyze the data, creating a valid linear model to predict a Tinder user's level of self-presentation genuineness. Randomly withhold 10% of the data as a test set for validation and report the requested MSPRs outlined later in this description.

Columns of File tinder.xlsx

Variable Name	Definition
ID	user identifier (arbitrary and meaningless other than to identify specific users)
Genuine	in terms of how they present themselves on Tinder, how genuine is the user (i.e., how honest and realistic, as opposed to fake or falsified or made-up)
SocPrivConc	how concerned is the user that other users will mis-use their private data
InstPrivConc	how concerned is the user that Tinder will mis-use their private data
Narcissism	how narcissistic is the user
SelfEsteem	how much self-esteem does the user have
Loneliness	how lonely is the user
Hookup	how interested is the user in using Tinder to hook up (especially for sex)
Friends	how interested is the user in using Tinder to build friendships
Partner	how interested is the user in using Tinder to develop a partnership
Travel	how interested is the user in using Tinder while traveling
SelfValidation	how interested is the user in using Tinder for self-validation
Entertainment	how interested is the user in using Tinder for entertainment
Orientation	user's sexual orientation (1=heterosexual, 2=homosexual, 3=bisexual, 4=other)
Gender	user's identified gender (1=male, 2=female, 3=other)
Education	user's education level (1=no schooling, 2=high school graduate, 3=some college, 4=undergrad degree, 5=masters degree, 6=doctoral degree, 7=other)
Income	user's estimated level of income (1=low, 2=medium, 3=high, 4=unknown)
Employment	user's current employment status (1=employed, 2=sef-employed, 3=out of work and looking, 4=out of work but not looking, 5=homemaker, 6=student, 7=military, 8=retired, 9=unable to work)
Age	user's age in years
ImpFitness	how important does the user think it is for members of their same gender to have physical fitness
ImpEnergy	how important does the user think it is for members of their same gender to have energy (or stamina)
ImpAttractive	how important does the user think it is for members of their same gender to have physical attractiveness

Specific Directions:

Your paper should include a discussion of the following elements:

- Describe the nature and background of the data and provide a motivation for the analysis (for a hypothetical audience).
- Discuss the approaches you used to build your regression model, including, but not limited to:
  - Variable transformations (must be attempted if model fails to meet assumptions)
  - Variable selection
  - Interaction terms
  - Satisfaction/Violation of assumptions regarding residuals.
- Report your final model equation and provide a few interpretations of some of the statistically significant coefficients in your model.
- Report the validation error (using the test set) and compare this to the validation error for a reduced model which contains only an intercept.
- Place the results of your study in context. Discuss the implications of your results for the general tinder user and identify potential future directions for research.
- Include the code you used in your analysis, with appropriate code comments, in the appendix.

The report should include clear graphical displays with appropriate explanations and interpretations, discussions of model assumptions, justifications of decisions made in the analysis (such as which variables to include in the model, which transformations to make, and how to deal with influential observations or outliers), and interpretations of model components. Figures and Tables should be included in the main body of the report and referenced in the text, per the homework style guide. Do not include any unnecessary computer output or code (i.e., the appendix should include only the code used to generate the results and plots referred to in the report). The length of the final report should be between 6-10 pages, not including the title page and appendix.

Please be sure to check out a copy of the assignment rubric in canvas before submitting either draft of your paper.