MGSC 310-01 Updated Final Project: Outline

(a) Chosen data set:

- "Goodreads-books" from Kaggle https://www.kaggle.com/jealousleopard/goodreadsbooks

(b) Chosen outcome & the variables that will be used to predict it:

- Outcome variable: "average_rating" variable
- Possible predictors:
  - bookID
    - id number for each book
  - title
    - the name of the book
  - authors
    - names of the authors of the book
  - isbn
    - International Standard Book Number (ISBN) for identification
  - Isbn13
    - A thirteen-digit ISBN for identification
  - Language_code
    - The language the book is written in
  - Num_pages
    - Number of pages in the book
  - Ratings_count
    - Total number of ratings of the book
  - Text_reviews_count
    - Total number of written text reviews of the book
  - Publication_date
    - Month/day/year
  - Publisher
    - Publisher's name

(c) Motivation to project – the business use case of the prediction:

We are using the Goodreads book dataset to build a predictive model for book ratings. The application could be used to provide quantitative insights to consumers, publishers, and authors, who are looking for more than the qualitative feedback from book reviews. Linear and decision tree regression analyses will help us predict our outcome variable, "average_rating", based on the other variables in the data set. We can also report the most important predictors of our outcome for marketing purposes.

(d) Methods that will be used to analyze question of interest:

1. One technique we will use for our analysis is a linear regression. We will take our predictor variables and use them to fit a trend line to predict our continuous

outcome variable. We are operating under the assumption that we expect to find a linear relationship between our variables. We will interpret the results of our model, using the regression coefficients.

2. A second method we plan to use is lasso. We want to apply regularization to our linear regression because we do not expect all of our predictor variables will have a strong relationship with our outcome. Lasso is appropriate for this since it acts as a variable selector. We want to identify which variables are most important for our predictive model and to overcome any limitations from the previous method of a linear regression, such as multicollinearity.
3. We will then use the elastic net method because it performs both ridge and lasso regularization techniques on our linear regression, using both penalties simultaneously. This allows us to consider prediction accuracy as well as the variable selection feature of lasso.
4. A random forest for a regression allows us to handle the qualitative predictors we have in our book data set.

(e) Group:

1. Adam Gonzalez
2. Jon Le
3. Erin Lee
4. Debbie Lu
5. Corinne Smith