ISE 5103 Intelligent Data Analytics
Homework #5

Instructor: Charles Nicholson

See course website for due date

Learning objective: Data wrangling, ordinary least squares regression (OLS) modeling and analysis.

Submission notes:

1. Teams of 1 or 2 (optional) – make sure to set this up correctly in Canvas and on the Kaggle.com competition site. Ask if you don't know!
2. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader may view your R code, but should never have to in order to find your solutions.
   (a) I expect high-quality, clear, concise yet complete, easy to read PDFs.
   (b) 10 page max – 10% penalty per page over the allowance.
3. In the PDF, clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.) Also, note that only relevant and informative computer output should be provided. For example, I do not want to see "warning" messages, or the results of "library" commands, etc.
4. Make sure to provide comments on what your R code is doing. Keep it clean and clear!
5. You will submit your complete R script. Note: include library commands to load all packages that are used in the completion of the assignment. Place these statements at the top of your script/code.
6. Do not zip your files for submission. Submit exactly two files. Name the files LastName-HW1 with the appropriate file extension (that is, .pdf for the write-up and .R or .Rmd for the code)

Kaggle competition notes:

- In order to join the competition, you need to create a Kaggle account. Only one account per student is allowed.
- Join the competition here: `https://www.kaggle.com/t/e5aa719be3584ef091d9f61fc931fb36`
- Once you join Kaggle and the competition, to create a team:
  - Have one person click on "Team"
  - Then, request a merge by searching for one of the other team members user name and "Request Merge"
  - Create a team name as stated in the instructions. Even if you want to work alone, you need to form a team so that we can recognize if you are online or on-campus student.
- Grades will, in part, be based on the quality of your predictions as compared to the other teams in the class. *It is your responsibility to read the rules and information on the competition website.*

1 Sales Prediction

In many businesses, identifying which customers will make a purchase (and when), is a critical exercise. This is true for both brick-and-mortar outlets and online stores. The data provided in this assignment is website traffic data acquired from an online retailer. You will be predicting customer sales.

The data and many details about the problem can be found here:
`https://www.kaggle.com/t/e5aa719be3584ef091d9f61fc931fb36`.

The data provides information on customer's website site visit behavior. Customers may visit the store multiple times, on multiple days, with or without making a purchase.

Your goal is to predict how much sales revenue can be expected from each customer. The variable `revenue` lists the amount of money that a customer spends on a given visit. Your goal is to predict how much money a customer will spend, in total, across all visits to the website, during the allotted one-year time frame (August 2016 to August 2017).

More specifically, you will need to predict a transformation of the aggregrate customer-level sales value based on the natural log. That is, if customer i has k_i revenue transactions, then you should compute:

$$custRevenue_i = \sum_{j}^{k_i} revenue_{ij} \quad \forall i \in customers$$

And then transform this variable as follows:

$$targetRevenue_i = \ln(custRevenue_i + 1) \quad \forall i \in customers$$

You will be evaluated on how well you can predict the target revenue on a test data set available at the Kaggle.com website.

(a) (25 points) Conduct exploratory data analysis (EDA), using any techniques you prefer, to help understand the data and prepare you for the *data wrangling* step next.

While you may do as much you like, make sure to choose *at least* 5 EDA analyses of value and provide commentary as to what you have found. Please note: Each visualization, table of statistics, etc., that you include must be described thoughtfully. Your "insight" represents half of the points for this problem – make it good.

(b) (30 points) Prepare the data for modeling. This requires you to consider missing values, outliers, transformations, aggregations, and/or any other data preparation technique you find useful. You may wish to consider using the `dplyr`, `lubridate`, and/or `forcats` packages, among others, to help you manage this process. This process will likely be iterative with the previous step (EDA) and with the next step (building and evaluating a model).

The deliverable for this step is a concise summary of the choices you have made and an explanation for your choices.

(c) (45 points) Build the best possible linear model using `lm` to predict the target value. Note: you may use any methods you like to anaylze the data, but the model used for predictions must be `lm`.

i. (15 points) For your best model, report the variables, coefficient estimates, and p-values. Additionally, report the re-sampled RMSE and R^2 values. What are your thoughts on the quality of the model? Did you have any problems during the modeling process? If so, how did you overcome those?

ii. (30 points) Submit your model predictions to the Kaggle.com competition website and outperform your peers in high quality predictions on the test data. You can submit multiple times each day to get feedback on the "public leaderboard". The final competition placement will be based on the "private leaderboard" standings. See the competition website for more details.