

ISE 5103 Intelligent Data Analytics

Homework #3

Instructor: Charles Nicholson

See course website for due date

Learning objective: Data understanding and preparation with an emphasis on transformations, missing value imputation, and feature engineering.

Submission notes:

1. Clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.)
2. All *relevant* computer output should be provided unless noted otherwise.
3. The R code itself is part of your solution – make sure to *provide comments* on what your code is doing. Keep it clean and clear!
4. 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.
5. You may use “R Markdown” to *help* with your submission. However, please edit the final submission to clearly and concisely respond to the questions. The goal is to limit your complete homework submission to under 10 pages.
6. Do not zip your files for submission. Submit exactly two files. Name the files “LastName-HW1” with the appropriate file extension (that is, .R, .pdf, .docx, or .html)

1 Glass Identification

The study of classification of types of glass is motivated by criminological investigations. At the scene of the crime, the glass left can be used as evidence... if it is correctly identified.

The data set we consider consists of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na (Sodium), Mg (Magnesium), Al (Aluminum), Si (Silicon), K (Potassium), Ca (Calcium), Ba (Barium), and Fe (Iron).

The data is available here: <http://archive.ics.uci.edu/ml/datasets/Glass+Identification> and is also available in the `mlbench` package as the dataset `Glass`.

- (a) Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors. Do there appear to be any outliers in the data? Are any predictors skewed?
- (b) Identify three attributes that you think could benefit from a skew transformation (there might be more than three that could benefit, but three is enough for this problem). For these attributes,
 - i. Use the `symbox` function from package `car` to consider possible power transformations
 - ii. Use the `boxcox` method from the `EnvStats` package to determine an optimal value for Box-Cox value of λ .
- (c) Use PCA to help evaluate the data. Does this provide any insight? If so, what?

- (d) Perform a linear discriminant analysis (LDA) using the `lda` function from the MASS package. Compare and contrast the results of LDA with PCA.

2 Missing Data

The `freetrade` data frame in the `Amelia` package has economic and political data on nine developing countries in Asia from 1980 to 1999. These 9 variables include year, country, average tariff rates, Polity IV score, total population, gross domestic product per capita, gross international reserves, a dummy variable for if the country had signed an IMF agreement in that year, a measure of financial openness, and a measure of US hegemony. Unfortunately, this data has missing values.

You will perform the following linear regression to predict the value of `tariff`:

```
lm(data=freetrade, tariff ~ year + country + polity + pop + gdp.pc + intresmi + signed + fiveop + usheg)
```

- (a) Perform the regression using listwise deletion.
- (b) Perform the regression using mean imputation.
- (c) Perform the regression using multiple imputation – in particular, *multivariate imputation by chained equations*. Note: you will have to find an imputation method that will work with this data!
- (d) Compare the coefficients for each of the regression models.

3 House prices data

The `housingData.csv` file in the course website is real data associated with 1,000 residential homes sold in Ames, Iowa between 2006 and 2010. The data set includes over 70 explanatory variables – many of which are factors with several levels. The file `housingVariables.pdf` provides a concise explanation of the variables and the factor levels in the data.

In a later assignment you will be challenged to create predictive models for the final price of each home. In preparation of this task I am asking you to perform exploratory data analysis, visualization, and feature engineering on this data and the wide range of variables.

- (a) Explore and visualize the data. (Now, that is an open-ended question!)
- (b) Construct new features, and/or collapse (reconstruct) factor levels as you see fit.
- (c) Justify the choices you’ve made in feature engineering.

Note: w.r.t. grading, I am looking for a logical and thoughtful process which should be supported visualizations and/or tables.

4 Kaggle.com – a little more data understanding

To complete this problem you need to join Kaggle.com – this site is where “the world’s largest community of data scientists compete” to solve real-world and sponsored analytics problems – often for significant cash awards.

Your course project will be based on data from sites like Kaggle.com (there are several), so spend a little time to see if there are some competitions that you find personally interesting.

- (a) Explore the site, competitions, and data – choose one data set to download from a competition to download. Provide the url and a *brief* description of the data (one or two sentences is fine!).
- (b) Perform an initial basic exploratory analysis of the data which includes at minimum: the number of rows, number of variables, descriptive statistics, a selection of visualizations, information on missing value counts, and some form of outlier labeling/detection.