Chapter 10: A Brief Note on Feature Engineering

Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

Icahn
School of
Medicine at
**Mount
Sinai**

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 15 min, max) while following along in the notes

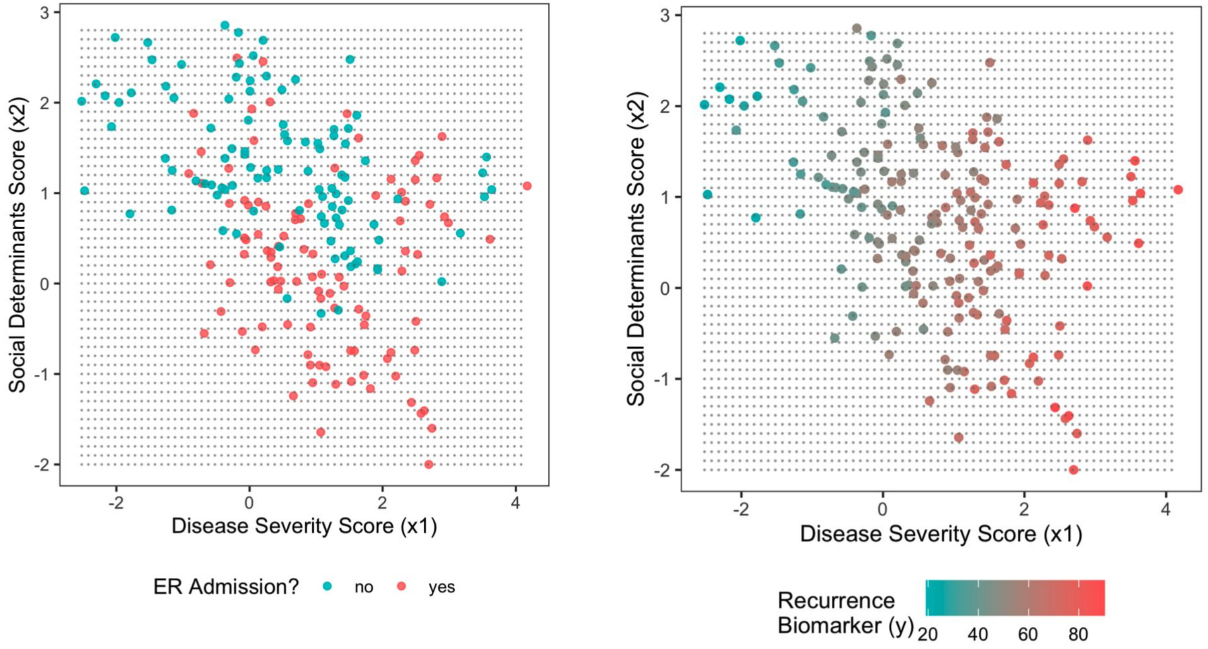
From the Ch. 10 notes:

“In many datasets, the features are chosen at the **study design** stage. The analyst (statistician, data scientist, etc.) has no say in what the features look like or which features are included.

This paradigm is changing as data science increasingly focuses on large, observational datasets, like those from electronic medical records (EMRs). In these types of studies, the raw data were not collected for the study itself, but to fulfill some other purpose. The analyst must choose how to build features from the raw data and use them in models.”

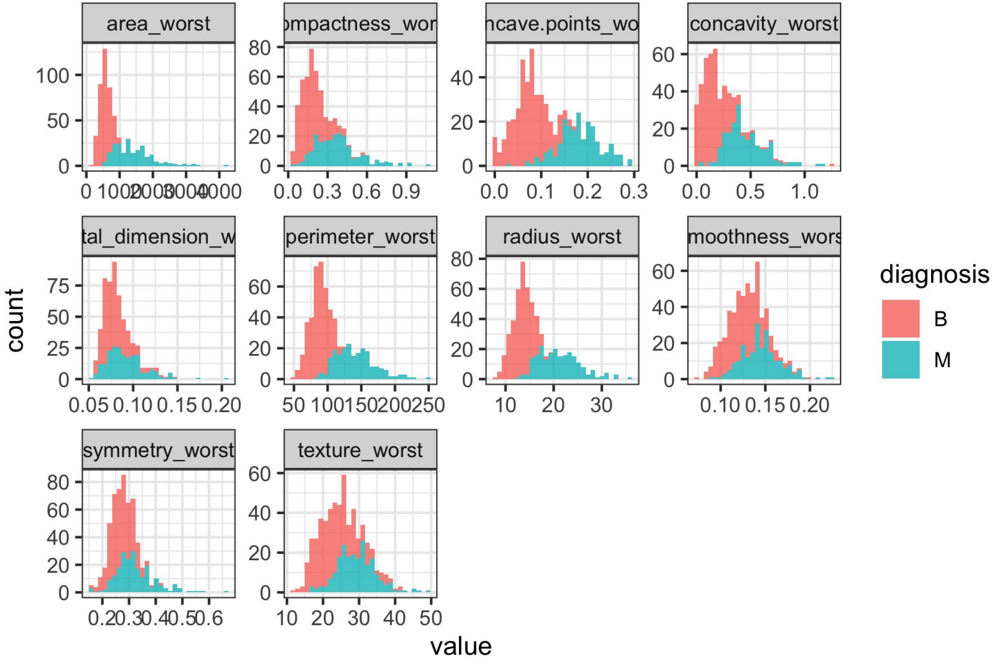
Question 10.1

The examples in Chapters 2 and 3 used the same two features. What were these features? How were they represented? What are some alternatives to this choice of features?



Question 10.2

In Chapter 7, we looked at the Wisconsin Breast Cancer Dataset, which includes 30 different imaging features relevant to predicting whether a tumor is benign or malignant. How were these features represented? What are some alternatives to this choice?



Question 10.3

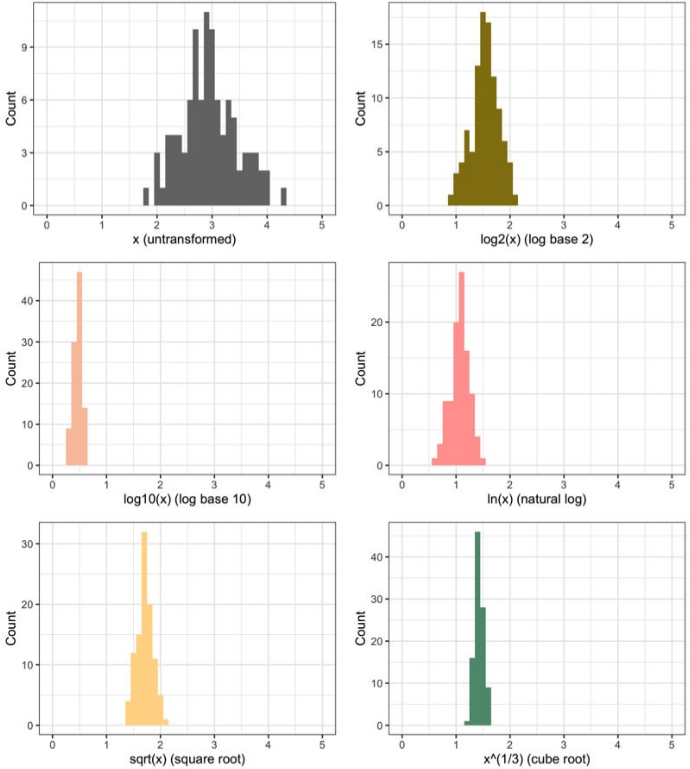
In Chapters 8 and 9, we looked at two datasets that were collected for the purposes of answering particular questions. Do you agree with these study designers' choice of features? What other features could potentially have been relevant to answering each research question?

MORT	Total age-adjusted mortality from all causes, in deaths per 100,000 population
PRECIP	Mean annual precipitation (in inches)
EDUC	Median number of school years completed for persons of age 25 years or older
NONWHITE	Percentage of the 1960 population that is nonwhite
NOX	Relative pollution potential of oxides of nitrogen
SO2	Relative pollution potential of sulfur dioxide

LOW	Low birth weight (0 = birth weight $\geq$ 2500 g; 1 = birth weight $<$ 2500 g)
AGE	Age of mother in years
LWT	Mother's weight in pounds at last menstrual period
RACE	Race (1 = white, 2 = black, 3 = other)
SMOKE	Smoking status during pregnancy (1 = yes, 0 = no)
PTL	History of premature labor (0 = none, 1 = one, etc.)
HT	History of hypertension (0 = no, 1 = yes)
UI	Presence of uterine irritability (0 = no, 1 = yes)
FTV	Number of physician visits during the first trimester
BWT	Birth weight in grams

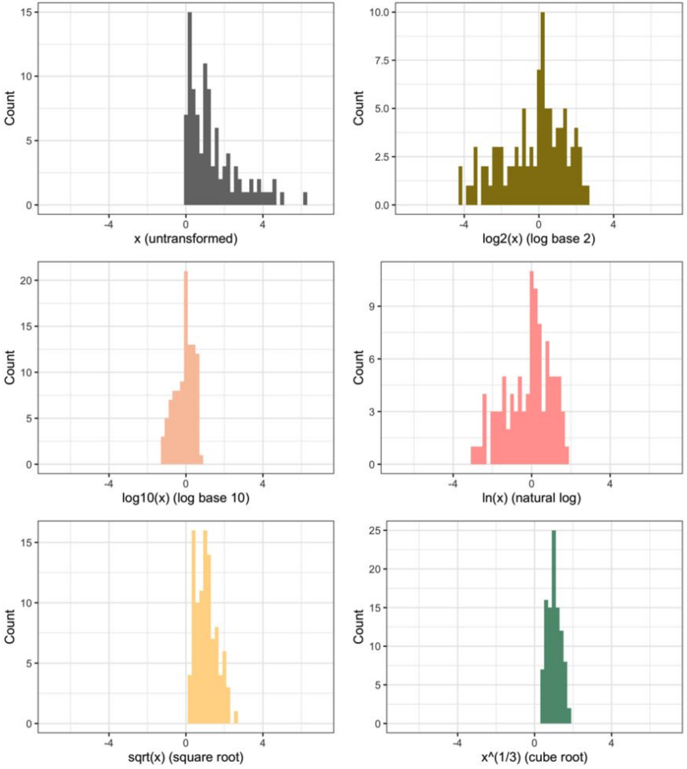
Question 10.4

Here are 100 random samples from a normal distribution with $\mu = 3.0$ and $\sigma = 0.5$ and five different transformations of those samples. What do you notice about the shape and position of the data under the different transformations?



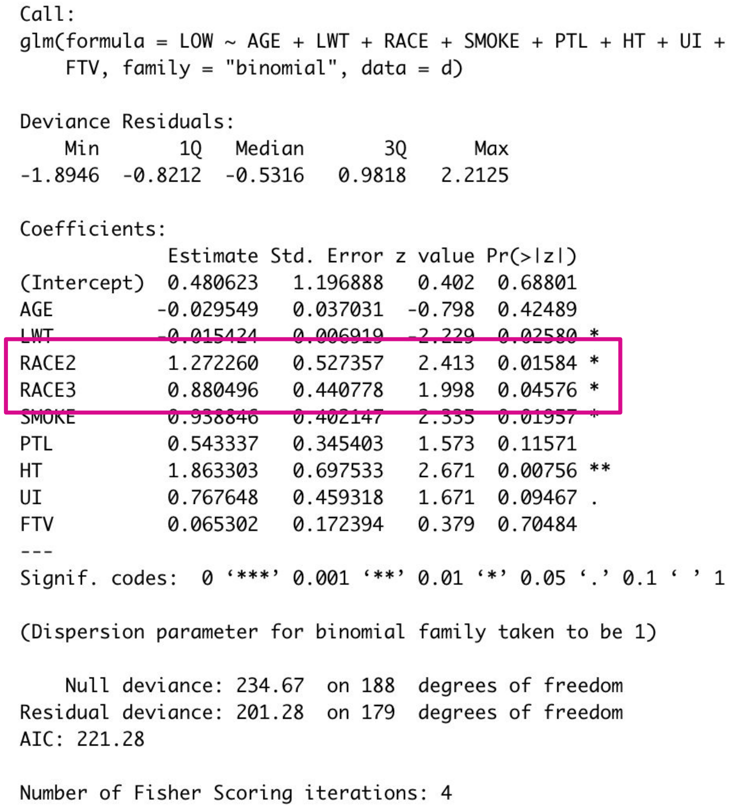
Question 10.5

Here are 100 random samples from an exponential (see Section 4.7) distribution with $\lambda = 0.8$ and the same five transformations of those samples. What do you notice about the shape and position of the data under the different transformations?



Question 10.6

In Section 9.3, we saw an example of a model that predicts whether or not a mother will give birth to a low birthweight baby. One of the factors considered in that model is the mother's race, which was coded (crudely and probably inaccurately, I might add) as `1 = white, 2 = Black, 3 = other`. You can tell how the feature `RACE` was coded by examining the model output. How many indicator variables were used? Which level of the feature was used as the reference category?



```
Call:
glm(formula = LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI +
    FTV, family = "binomial", data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8946  -0.8212  -0.5316   0.9818   2.2125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.480623   1.196888   0.402  0.68801
AGE         -0.029549   0.037031  -0.798  0.42489
LWT         -0.015424   0.006919  -2.229  0.02580 *
RACE2        1.272260   0.527357   2.413  0.01584 *
RACE3        0.880496   0.440778   1.998  0.04576 *
SMOKE        0.938846   0.402147   2.335  0.01957 *
PTL          0.543337   0.345403   1.573  0.11571
HT           1.863303   0.697533   2.671  0.00756 **
UI           0.767648   0.459318   1.671  0.09467 .
FTV          0.065302   0.172394   0.379  0.70484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.28  on 179  degrees of freedom
AIC: 221.28

Number of Fisher Scoring iterations: 4
```