# QAC 305 – Exploratory Data Analysis and Pattern Recognition

## Assignment 2 - Due March 9th

## Working with Missing Data

Consider the following dataset:

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

The data is contained in a data frame with 397 observations on the following 6 variables:

- `Rank` - a factor with levels `AssocProf` `AsstProf` `Prof`
- `Discipline` - a factor with levels `A` ("theoretical" departments) or `B` ("applied" departments).
- `yrs.since.phd` - years since PhD.
- `yrs.service`- years of service.
- `Sex` - a factor with levels `Female` `Male`
- `Salary` - nine-month salary, in dollars.

1. Load the dataset from the file **prof_salary.Rdata**. Since the file is already in R format, use the **load( )** function.

2. What is the percentage of missing data in each variable?

3. What are the patterns of missing data? Include both a table and graph.

4. Is there a relationship between missing values on salary and any of the other variables?

5. Perform a linear regression predicting salary from the other five variables. Use list-wise deletion.

6. Perform a linear regression predicting salary from the five other variables. Use **kNN** to impute missing values. Use the **imp_var=FALSE** option to avoid creating additional variables.

7. Perform a linear regression predicting salary form the other five variables. Use **missForest** to impute missing values. Use a random number seed of 1234.

8. It turns out in speaking with the researchers, that men with many years of experience were less likely to answer the salary question. What missing data mechanism does this describe?

9. Researchers went back to the original respondents and had them complete any missing questions. An analysis on a complete dataset is given below.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
  rankAssocProf  12907.6     4145.3   3.114  0.00198 **
  rankProf       45066.0     4237.5  10.635  < 2e-16 ***
  disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
  yrs.since.phd    535.1      241.0   2.220  0.02698 *
  yrs.service     -489.5      211.9  -2.310  0.02143 *
  sexMale         4783.5     3858.7   1.240  0.21584
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  Residual standard error: 22540 on 390 degrees of freedom
  Multiple R-squared:  0.4547,         Adjusted R-squared:  0.4463
  F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

Which approach yielded the closest results to the complete dataset?

## Matrix Algebra

Consider the matrices and vectors defined below.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 2 & 2 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & -1/2 \\ -1 & 1 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Perform the following operation, indicating any that are impossible to perform.

10. What is the order of
    a. **A**
    b. **C**
    ***c. X***

11. Perform the following operations, indicating any that are impossible.
    **a. A + B**
    **b. B + A**
    **c. A * x**
    **d. B * C**
    **e. C * B**

12. Is **D** the inverse of **A**? Prove your answer.

13. What is the transpose of **C**?

14. Consider matrix **B**
    a. Is **B** symmetric? Prove your answer.
    b. What is the principal diagonal of **B?**
    c. What is the trace of **B**?

15. **A** * **E** = **B**. Solve for **E**.