Homework 7

MTH 3270 Data Science

Due Sat., Apr. 9

Read These Chapters of the Book	**Then Do These Exercises**
Appendix E	Problem 5* (**App E**)
10	Problem 3** (**Ch 10**)
11	Problem 4 (**Ch 11**), Problem 6*** (parts ***a*** and ***c*** only, and just do **decision tree**, **random forest**, and ***k*-NN**) (**Ch 11**)

* For **Problem 5** (**App E**):

- The `HELPrct` data is in the `"mosaicData"` package. You can view its help page by typing:

```
library(mosaicData)           # Contains the HELPrct data set
? HELPrct
```

- The response variable (Y), `homeless`, is **dichotomous**, so a ***logistic regression*** analysis is appropriate. It's best to create a **recoded** (**0** or **1**) version of the response variable for use in the model fitting step:

```
library(dplyr)      # For mutate() and case_when()

HELPrct <- mutate(HELPrct,
                  homeless01 = case_when(homeless == "housed" ~ 0,
                                         homeless == "homeless" ~ 1))
```

- **Don't** use any of the **categorical** variables as explanatory variables (X's) in the model. To see which variables are **numeric** (or **integer**) and which are **categorical** (**factors**), type:

```
str(HELPrct)
```

** For **Problem 3** (**Ch 10**):

- The `HELPrct` data is in the `"mosaicData"` package.
- For **Part *a***, the **"null model"** is one that *doesn't* contain *any* explanatory variables (X's). After creating the **recoded** (**0** or **1**) version of `homeless` (see above), it can be fitted using:

```
my.logreg <- glm(homeless01 ~ 1, data = HELPrct, family = "binomial")
```

*** For **Problem 6** (**Ch 11**):

- The `NHANES` data set is in the `"NHANES"` package. The help page has a description of the data set:

```
library(NHANES)
? NHANES
```

- There are *many* `NA`s in the data set. In fact, *every row* has at least one `NA`:

```
any(complete.cases(NHANES))
```

  One way to deal with the `NA`s is to *first* use `select()` (from `"dplyr"`) to create a new data frame containing only the variables (columns) from `NHANES` that you want to use in your classification models, *then* use `na.omit()` (or `complete.cases()`) to create a new version of that data frame which contains only the observations (rows) that don't have any `NA`s.

- **Don't** use any of the **categorical** variables as explanatory variables (X's) in the classification model. To see which variables are **numeric** (or **integer**) and which are **categorical** (**factors**), type:

```
str(NHANES)
```

- It's *possible* that your **decision tree** might end up having only one (root) node. If this happens, try changing the value of the **complexity parameter** (or **tuning parameter**), i.e. the `cp` value – see Exercise 6 in Class Notes 6.