Exercise-7

Matthew McAvoy

September 14, 2016

To prepare the .arff file, I followed directions provided in class on how to coerce it into the right csv format using emacs. It is now in a readable format for R.

```
ckddata <- read.csv("C:/Users/homur/OneDrive/New College/EDA/Week 4/ckd.csv")
```

I want to begin by looking at what might be in the blood pressure, blood urea, creatinine, and potassium fields. From the codebook, I see blood pressure = bp, blood urea = bu, creatinine = sc, and potassium = pot. Following are normal ranges for each field so that I will know what to expect and what might look suspicious. Ranges gathered from quick internet search.

Blood Pressure: 80 - 12 mm/Hg Blood Urea: 7 - 20 mg/dL Creatinine: 0.8 - 1.2 mg/dL Potassium: 3.7 - 5.2 mEq/L

```
str(ckddata$bp)
```

```
##  Factor w/ 12 levels "","?","100","110",..: 11 8 11 10 11 12 10 2 3 12 ...
```

```
str(ckddata$bu)
```

```
##  Factor w/ 120 levels "","?","1.5","10",..: 66 35 84 87 54 53 85 60 90 7 ...
```

```
str(ckddata$sc)
```

```
##  Factor w/ 88 levels "","?","0.4","0.5",..: 12 7 18 58 14 11 51 11 19 79 ...
```

```
str(ckddata$pot)
```

```
##  Factor w/ 44 levels "","?","2.5","2.7",..: 2 2 2 3 2 8 18 2 2 13 ...
```

It looks like there arn't as many different blood pressures as we might have expected. And with creatinine being such a small range, there is a high degree of variability. Additionally, each field has a ? in the data. This should be removed when we want to put it into numeric form.

```
bp = ckddata$bp
bu = ckddata$bu
sc = ckddata$sc
pot = ckddata$pot
x <- data.frame(bp, bu, sc, pot)
```

We can use the transform function to coerce ?'s into NA's.

```
x_num <- transform(x, bp = as.numeric(as.character(bp)), bu = as.numeric(as.character(bu)), sc = as.num
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

Now we can see how many NA's in total and then in summary we can see how many there are in each field.

```
sum(is.na(x_num))
```

```
## [1] 140
```

```
summary(x_num)
```

```
##        bp               bu               sc              pot
##  Min.   : 50.00   Min.   :  1.50   Min.   : 0.400   Min.   : 2.500
##  1st Qu.: 70.00   1st Qu.: 27.00   1st Qu.: 0.900   1st Qu.: 3.800
##  Median : 80.00   Median : 42.00   Median : 1.300   Median : 4.400
##  Mean   : 76.47   Mean   : 57.43   Mean   : 3.072   Mean   : 4.627
##  3rd Qu.: 80.00   3rd Qu.: 66.00   3rd Qu.: 2.800   3rd Qu.: 4.900
##  Max.   :180.00   Max.   :391.00   Max.   :76.000   Max.   :47.000
##  NA's   :13       NA's   :20       NA's   :18       NA's   :89
```

It looks like there aren't that many missing entries for each field except potassium which has quite a bit. All together there are 140 missing entries.

Looking at ranges, blood pressure seems mostly in line, the max of 180 seems startling. Depending on how high blood urea can go, even the mean is outside our expected range. Either this means the data was input wrong multiple times or there are alot of people with unnaturally high blood urea. The max of 391 I'm guessing is an input error. For creatinine, the max of 76 again seems to be an input error, along with potassium's 47. Let's remove these that we guess were input error.

```
rm_bu <- grep("[0-9]{3}", x_num$bu)
x_num$bu[rm_bu] <- NA
rm_sc <- grep("[0-9]{2}", x_num$sc)
x_num$sc[rm_sc] <- NA
rm_pot <- grep("[0-9]{2}", x_num$pot)
x_num$pot[rm_pot] <- NA
summary(x_num)
```

```
##        bp               bu              sc             pot
##  Min.   : 50.00   Min.   : 1.50   Min.   :0.40   Min.   :2.50
##  1st Qu.: 70.00   1st Qu.:25.00   1st Qu.:0.90   1st Qu.:3.80
##  Median : 80.00   Median :38.00   Median :1.20   Median :4.40
##  Mean   : 76.47   Mean   :41.11   Mean   :2.01   Mean   :4.38
##  3rd Qu.: 80.00   3rd Qu.:50.00   3rd Qu.:2.40   3rd Qu.:4.90
##  Max.   :180.00   Max.   :98.60   Max.   :9.70   Max.   :7.60
##  NA's   :13       NA's   :71      NA's   :43     NA's   :91
```

We have more NA's now, but our ranges are better.