DSCI401 - Homework 5

Due: November 3th, 2024

Homework should be submitted as an R Markdown file with links to Google colab notes where necessary. Homework should be turned in on Sakai.

1. Generate the code to convert the following data frame to wide format.

```
dat <- data.frame(grp = c("A","A","B","B"),
           sex = c("F","M","F","M"),
           meanL = c(0.225,0.47,0.325,0.547),
           sdL = c(0.106,.325,.106,.308),
           meanR = c(.34,.57,.4,.647),
           sdR = c(0.0849, 0.325, 0.0707, 0.274)
)
```

   The output should look like this:

```
## # A tibble: 2 x 9
##   grp   F.meanL F.sdL F.meanR  F.sdR M.meanL M.sdL M.meanR M.sdR
##   <chr>   <dbl> <dbl>   <dbl>  <dbl>   <dbl> <dbl>   <dbl> <dbl>
## 1 A       0.225 0.106    0.34 0.0849   0.47  0.325   0.57  0.325
## 2 B       0.325 0.106    0.4  0.0707   0.547 0.308   0.647 0.274
```

2. Consider the pccc_icd10_dataset.

   (a) Remove all the columns labeled with "g" and a number.

   (b) Convert this to a long data set with three columns: id, type (pc or dx), and code.

```
library(pccc)
head(pccc_icd10_dataset)

##   id     dx1     dx2     dx3     dx4     dx5     dx6     dx7     dx8     dx9
## 1  1 S9410XS  I67841  E70339    <NA> S14121A  M66229 S92065G   O0973    <NA>
## 2  2    <NA> S53422D S92244B  M66342    <NA> S32442A T1582XD S72325C S52131B
## 3  3    <NA> S91225S    <NA> W6119XD   C8397 M80819K S72114R    <NA> Y382X3D
## 4  4 S7226XK   Y93G2   L0592  K08530    <NA> S62637D T84612A    <NA>    <NA>
## 5  5 S92246A   O4212   D2920 S42434S  F15980    <NA> S52572R M8080XA X731XXD
## 6  6    <NA> S52291C    <NA>    <NA>   E7140  H05222 S60549S    <NA> S32616G
##      dx10     pc1     pc2     pc3     pc4     pc5     pc6     pc7     pc8
## 1    <NA> 0PSH3CZ 0JPT3XZ 037906Z 0JHD3HZ 0KQ54ZZ 0WPK3YZ 01B04ZX 0DWV07Z
## 2   O1400 0DVM7DZ 0NRJ47Z DWY48ZZ 0HRWX7Z BP091ZZ 0Y0H4JZ    <NA> 0B9880Z
## 3  I70519 0PBV4ZX 0XM20ZZ 0DWD4UZ 2W07XYZ F0636ZZ 0RUP37Z    <NA> 0WCP8ZZ
## 4    <NA> DDY37ZZ 07LL0CZ 0Y9930Z 037M3GZ 04100Z4    <NA> 0SPG33Z 0TRC07Z
## 5 S42471K 02UL4KZ 03VD0ZZ 02110K8 3E050HZ 3E0U0GB    <NA> 0SPQ30Z 0WWBXYZ
## 6    <NA> 0D740DZ 0V1Q4JJ 10A07Z6 03150AK 047J47Z 0NQHXZZ 08BY3ZZ 047B376
##       pc9    pc10     g1     g2     g3     g4     g5     g6     g7     g8
## 1 09513ZZ 0V554ZZ 239196 672832 683784 757546     NA 168052 104625     NA
## 2    <NA>    <NA> 931331 404900 912213     NA 964580 371556 778488 115827
## 3 0DUM4KZ BN02ZZZ 627455 638100 745829 843799 322975     NA     NA 932106
## 4 041M0KQ DB10B8Z 809782 153243 413723 130995 211708 610135     NA 471383
## 5    <NA> 0SWN38Z     NA 636794     NA 928572 930823 168586 133292 699936
## 6 0SRQ07Z 0GPR00Z 281891 318962 542326 705580 700647 929863 338026 525937
##       g9    g10
## 1 850974     NA
## 2 440619 955264
## 3 289004 242699
## 4 191245 135116
## 5 500743     NA
## 6 412691     NA
```

3. In the ‘Lahman‘ package using the ‘People‘ dataset answer the following questions.

   (a) Make birthDate and deathDate date variables and if deathDate is NA put in the date November 3rd 2024.

   (b) Using those variables find the median age of how long players lived or currently are. Does this change drasticly by birth country?

   (c) Look at the age of a player at their debut versus their final game, what are the median ages for both of those? Visualize in a side by side boxplot.

   (d) Using Debut and finale game how long does players usually play. Give the 5 number summary of this.