SESSION 2-ASSIGNMENT 2

Read multiple JSON files into a directory to convert into a dataset.I have files text1, text2, text3 in the directory JSON.

```
library(rjson)
filenames <- list.files("Users/Desktop/json", pattern="*.json", full.names=TRUE) # this should give you a character vector, with each file name represented by an entry
myJSON <- lapply(filenames, function(x) fromJSON(file=x)) # a list in which each element is one of your original JSON files
```

2. Parse the following JSON into a data frame.

```
js<-'{
"name": null, "release_date_local": null, "title": "3 (2011)",
"opening_weekend_take": 1234, "year": 2011,
"release_date_wide": "2011-09-16", "gross": 59954
}'
```

```
require(RJSONIO)
js<-'[{"name": null, "release_date_local": null, "title": "3 (2011)",
"opening_weekend_take": 1234, "year": 2011,
"release_date_wide": "2011-09-16", "gross": 59954}]'

js <- fromJSON(js)
js <- lapply(js, function(x) {
  x[sapply(x, is.null)] <- NA
  unlist(x)
})
```

Then finally use do.call method

```
asDataFrame <- do.call("rbind", lapply(js, as.data.frame))
```

Output:

```
name release_date_local title      opening_weekend_take year
[1,] NA   NA            "3 (2011)" "1234"              "2011"
     release_date_wide gross
[1,] "2011-09-16"      "59954"
```

Write a script for Variable Binning using R.

Datasets like iris or GermanCredit are not applicable due to not having NAs, strings or zeros, so I wrote some code below to replicate my data.

Raw data to be binned.

```
OVERDUEAMOUNT_numbers <- rnorm(10000, mean = 9000, sd = 3000)
OVERDUEAMOUNT_zeros <- rep(0, 3000)
OVERDUEAMOUNT_NAs <- rep(NA, 4000)
OVERDUEAMOUNT <- c(OVERDUEAMOUNT_numbers, OVERDUEAMOUNT_zeros,
OVERDUEAMOUNT_NAs)

PROFESSION_f1 <- rep("438", 3000)
PROFESSION_f2 <- rep("000", 4000)
PROFESSION_f3 <- rep("selfemployed", 5000)
PROFESSION_f4 <- rep(NA, 5000)
PROFESSION <- c(PROFESSION_f1, PROFESSION_f2, PROFESSION_f3, PROFESSION_f4)

ID <- sample(123456789:987654321, 17000, replace = TRUE); n_distinct(ID)

df_Raw <- cbind.data.frame(ID, OVERDUEAMOUNT, PROFESSION)
colnames(df_Raw) <- c("ID", "OVERDUEAMOUNT", "PROFESSION")
```

Convert PROFESSION to factor to replicate this variable is processed & prepared for further import into R. Reshuffle the dataframe row-wise to make it look like real data.

```
df_Raw$PROFESSION <- as.factor(df_Raw$PROFESSION)
df_Raw <- df_Raw[sample(nrow(df_Raw)), ]
```

Dataframe with bins.

```
variable <- c(rep("OVERDUEAMOUNT", 7), rep("PROFESSION", 4))
min <- c(0, c(-Inf, 1500, 4000, 8000, 12000), "", c("438", "000", "selfemployed", ""))
max <- c(0, c(1500, 4000, 8000, 12000, Inf), "", c("438", "000", "selfemployed", ""))
bin <- c(c(1, 2, 3, 4, 5, 6, 7), c(1, 2, 3, 4))

binsDF <- cbind.data.frame(variable, min, max, bin)
colnames(binsDF) <- c("variable", "min", "max", "bin")
```