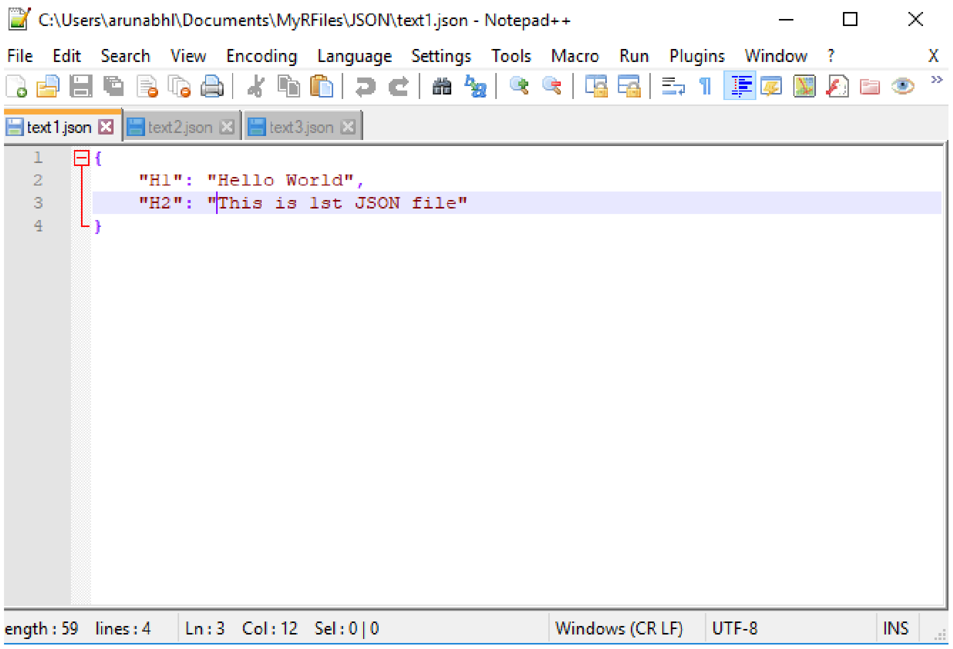
Problem Statement

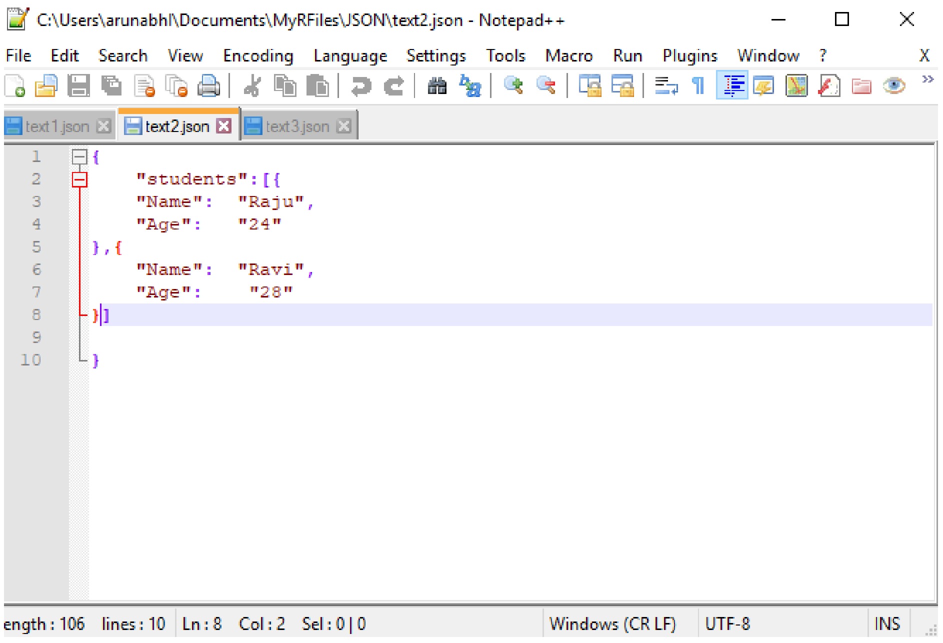
1. Read multiple json files into a working directory for further converting into a dataset. I have files text1, text2, text3 in the directory json.

Answer: Created 3 json files under *("C:/Users/arunabhl/Documents/MyRFiles/JSON") folder*


C:\Users\arunabhl\Documents\MyRFiles\JSON\text1.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
text1.json text2.json text3.json

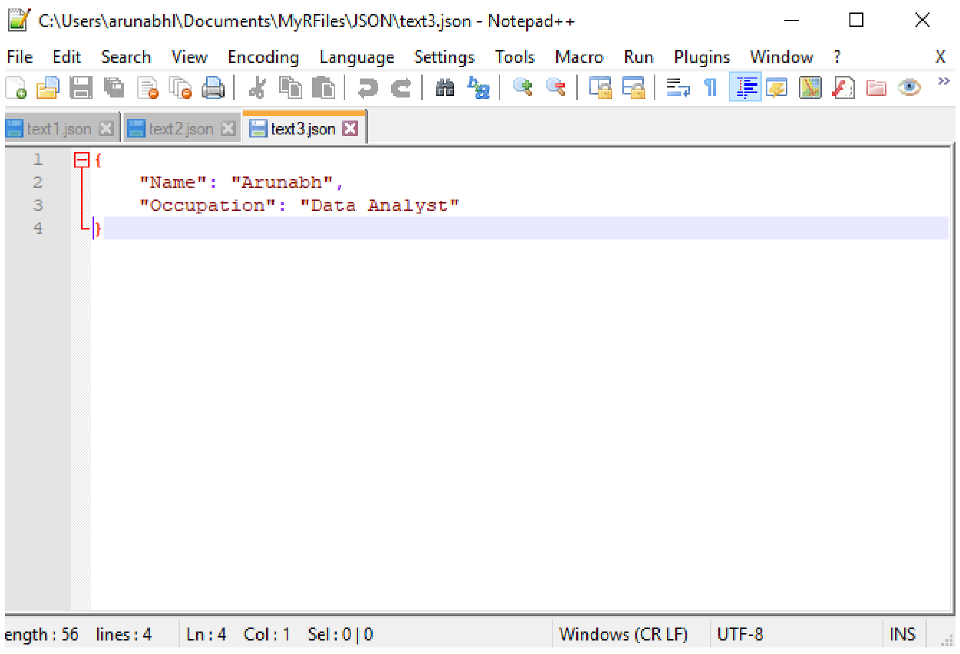
```
{
    "H1": "Hello World",
    "H2": "This is 1st JSON file"
}
```

ength : 59 lines : 4 Ln : 3 Col : 12 Sel : 0 | 0 Windows (CR LF) UTF-8 INS


C:\Users\arunabhl\Documents\MyRFiles\JSON\text2.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
text1.json text2.json text3.json

```
{
    "students":[{
    "Name":  "Raju",
    "Age":   "24"
},{
    "Name":  "Ravi",
    "Age":    "28"
}]

}
```

ength : 106 lines : 10 Ln : 8 Col : 2 Sel : 0 | 0 Windows (CR LF) UTF-8 INS


C:\Users\arunabhl\Documents\MyRFiles\JSON\text3.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
text1.json
text2.json
text3.json
ength : 56 lines : 4 Ln : 4 Col : 1 Sel : 0 | 0 Windows (CR LF) UTF-8 INS

```
{
    "Name": "Arunabh",
    "Occupation": "Data Analyst"
}
```

Step 1- Install ***rjson package*** and then ***library(rjson)***

> ***install.packages("rjson")***
>
> ***library(rjson)***

Step 2- set the working directory where the JSON files are located.

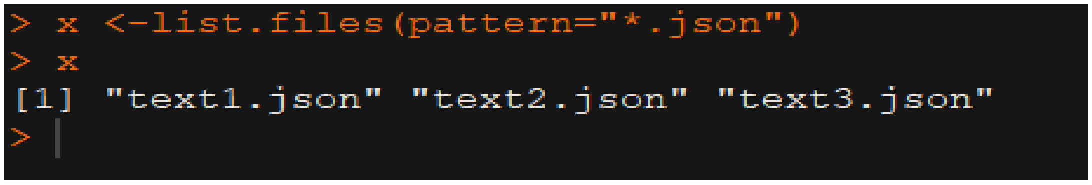
> ***setwd("C:/Users/arunabhl/Documents/MyRFiles/JSON")***

Step 3- Use the command-

> ***x <-list.files(pattern="*.json")***

Here the list of all files under the working directory with the pattern or extension as **.json** is declared on a variable 'x'.

Output -

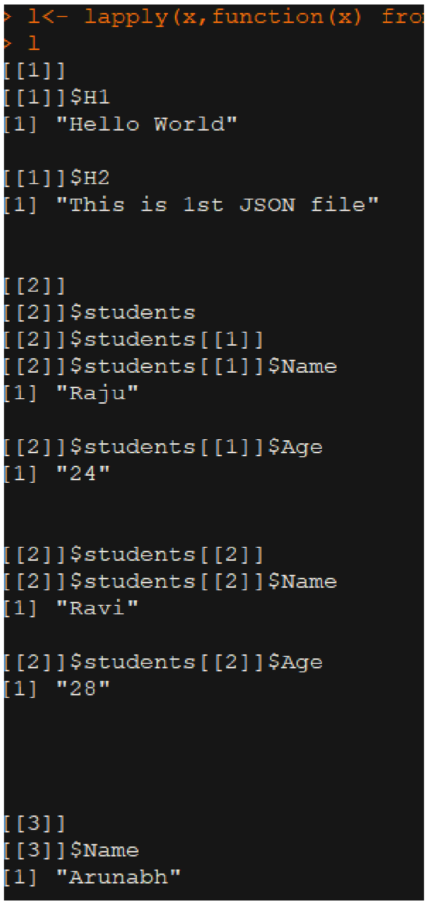


```
> x <-list.files(pattern="*.json")
> x
[1] "text1.json" "text2.json" "text3.json"
> 
```

Step 4 - Use lapply()

l<- lapply(x,function(x) fromJSON(file= x))

Output -



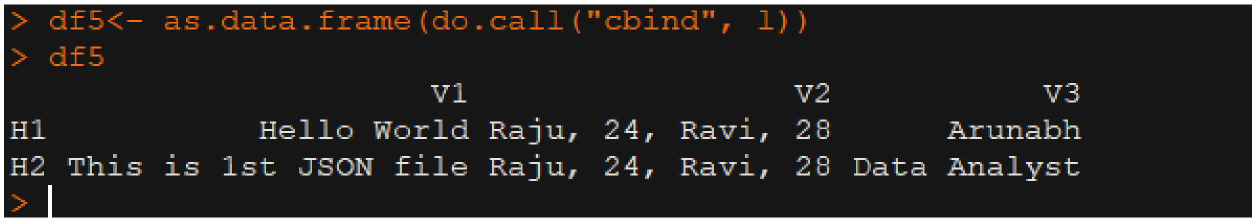
```
> l<- lapply(x,function(x) fro
> l
[[1]]
[[1]]$H1
[1] "Hello World"

[[1]]$H2
[1] "This is 1st JSON file"


[[2]]
[[2]]$students
[[2]]$students[[1]]
[[2]]$students[[1]]$Name
[1] "Raju"

[[2]]$students[[1]]$Age
[1] "24"


[[2]]$students[[2]]
[[2]]$students[[2]]$Name
[1] "Ravi"

[[2]]$students[[2]]$Age
[1] "28"




[[3]]
[[3]]$Name
[1] "Arunabh"
```

Here, the lapply functions is used on the variable 'x' and the output of all the jSoN files are as "list".

Step 5- Converting the list to dataframe in below script

df5<-as.data.frame(do.call("cbind", l))

Output-



```
> df5<- as.data.frame(do.call("cbind", l))
> df5
                          V1                 V2           V3
H1               Hello World Raju, 24, Ravi, 28      Arunabh
H2 This is 1st JSON file Raju, 24, Ravi, 28 Data Analyst
> 
```

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

}'

ANSWER:

We need to convert the JSON into a list and then to a data frame.

Step 1: Install rjson package and then pass the command library(rjson)

install.packages("rjson")

library(rjson)

Step 2: Assign the json into a variable

js_ans<- '{

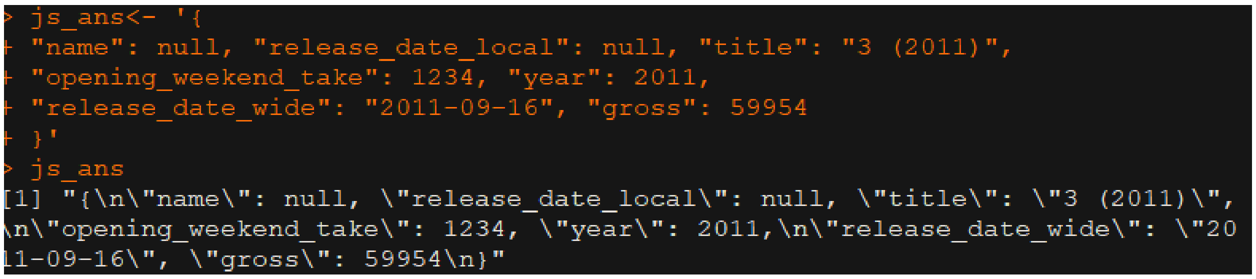
"name": null, "release_date_local": null, "title": "3 (2011)",

"opening_weekend_take": 1234, "year": 2011,

"release_date_wide": "2011-09-16", "gross": 59954

}'

Output:

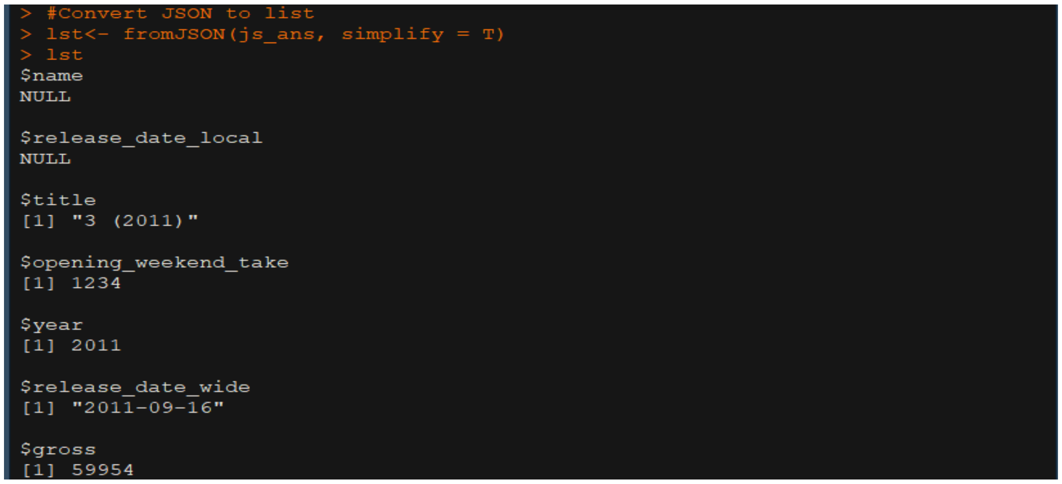


```
> js_ans<- '{
+ "name": null, "release_date_local": null, "title": "3 (2011)",
+ "opening_weekend_take": 1234, "year": 2011,
+ "release_date_wide": "2011-09-16", "gross": 59954
+ }'
> js_ans
[1] "{\n\"name\": null, \"release_date_local\": null, \"title\": \"3 (2011)\",
\n\"opening_weekend_take\": 1234, \"year\": 2011,\n\"release_date_wide\": \"20
11-09-16\", \"gross\": 59954\n}"
```

Step 3: Converting json string into list

lst<- fromJSON(js_ans, simplify = T)

lst

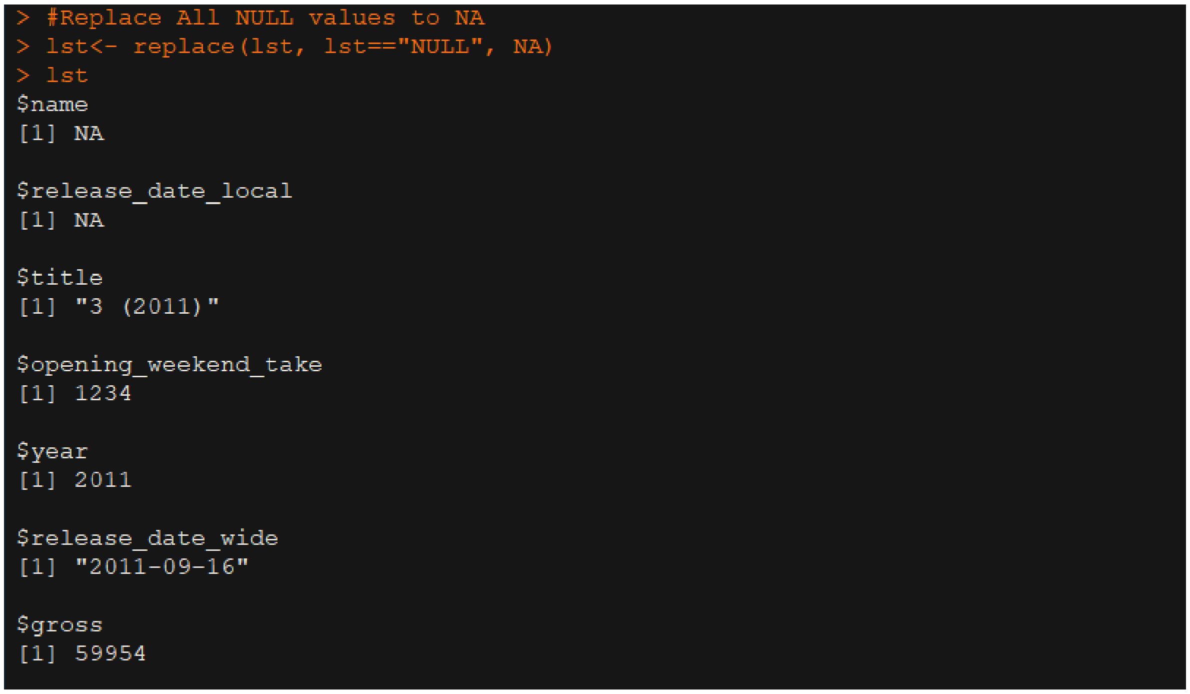


```
> #Convert JSON to list
> lst<- fromJSON(js_ans, simplify = T)
> lst
$name
NULL

$release_date_local
NULL

$title
[1] "3 (2011)"

$opening_weekend_take
[1] 1234

$year
[1] 2011

$release_date_wide
[1] "2011-09-16"

$gross
[1] 59954
```

Step4: Replace all NULL values to NA

lst<- replace(lst, lst="NULL", NA)

lst

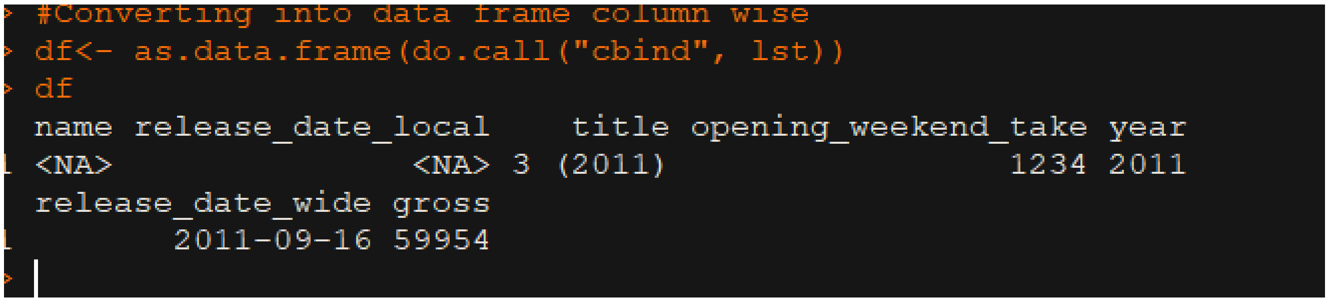


```
> #Replace All NULL values to NA
> lst<- replace(lst, lst=="NULL", NA)
> lst
$name
[1] NA

$release_date_local
[1] NA

$title
[1] "3 (2011)"

$opening_weekend_take
[1] 1234

$year
[1] 2011

$release_date_wide
[1] "2011-09-16"

$gross
[1] 59954
```

Step 5: converts list to dataframe in column wise **(FINAL OUTPUT)**

df<-as.data.frame(do.call("cbind", lst))

df



```
> #Converting into data frame column wise
> df<- as.data.frame(do.call("cbind", lst))
> df
  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
> 
```

3. Write a script for variable binning using R.

Answer- Variable Binning in R is the method of converting a numerical variable of a dataset into a categorical variable. To do so we use the cut() command, which divides the range of a variable (say x) into intervals.

Example - I have imported a csv file called "LungCapData" consisting of different variables like "Age", "Height", "Smoke", "Gender" and "Caesarean". Considering the "Height" variable we are going to divide this variable into different ranges.

Step1: Importing the File

LungCapData<- read.csv(file.choose(), header = T, sep = "\t")

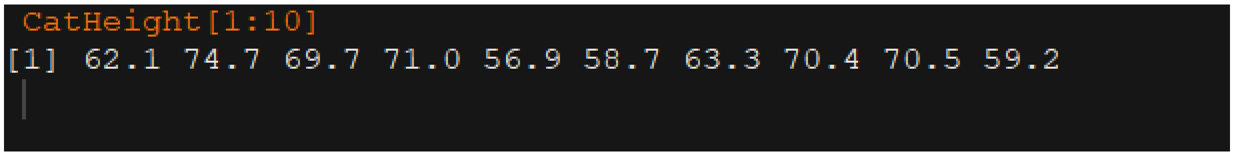
LungCapData

Step2: #Creating a new Variable to save Height

CatHeight<- LungCapData$Height

CatHeight[1:10]

Output:



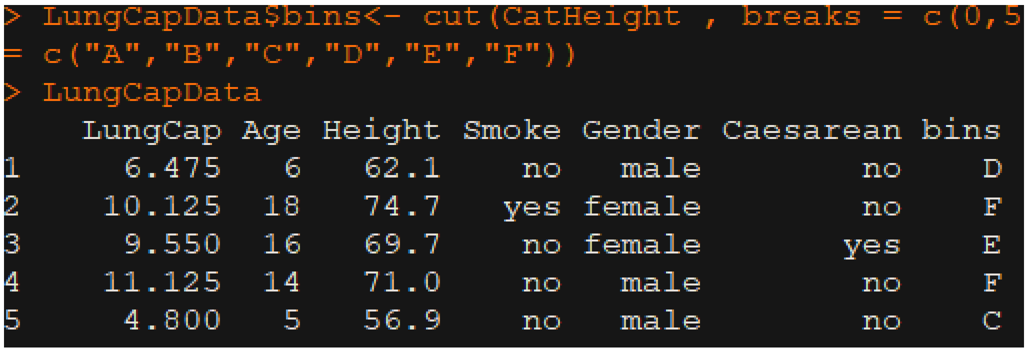
```
 CatHeight[1:10]
[1] 62.1 74.7 69.7 71.0 56.9 58.7 63.3 70.4 70.5 59.2
```

Step3: #Dividing the height categories A<=50, B=50-55, C=55-60, D=60-65, E=65-70, F=70+

#Also adding a column to the dataset

LungCapData$bins<- cut(CatHeight , breaks = c(0,50,55,60,65,70,100), labels = c("A","B","C","D","E","F"))

LungCapData



```
> LungCapData$bins<- cut(CatHeight , breaks = c(0,5
= c("A","B","C","D","E","F"))
> LungCapData
   LungCap Age Height Smoke Gender Caesarean bins
1    6.475   6   62.1    no   male        no    D
2   10.125  18   74.7   yes female       no    F
3    9.550  16   69.7    no female      yes    E
4   11.125  14   71.0    no   male        no    F
5    4.800   5   56.9    no   male        no    C
```

Explanations:

breaks() is used to break as per as the ranges required

labels =c() gives label names to the CatHeight bins.

LungCapData$bins is creating a new column consisting of the bins labels information.

Please find the **R-script** attached to see the **Output** in R named "R Bins By Variable".