structural_bioinformaticspt1

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
PDB <- "PDB_Data_bioinformatics.csv"
PDBstats = read.csv(PDB)


#Q1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy
PDBstats
```

```
           Molecular.Type   X.ray     EM    NMR Multiple.methods Neutron Other
1          Protein (only) 167,317 15,698 12,534              208      77    32
2 Protein/Oligosaccharide   9,645  2,639     34                8       2     0
3              Protein/NA   8,735  4,718    286                7       0     0
4     Nucleic acid (only)   2,869    138  1,507               14       3     1
5                   Other     170     10     33                0       0     0
6  Oligosaccharide (only)      11      0      6                1       0     4
    Total
1 195,866
2  12,328
```

```
3  13,746
4   4,532
5     213
6      22
```

```r
#removed commas
totalSUM<-sum(as.numeric(gsub(",","",PDBstats$Total)))
EMSUM<-sum(as.numeric(gsub(",","",PDBstats$EM)))
XraySUM<-sum(as.numeric(gsub(",","",PDBstats$X.ray)))
Q1<-((EMSUM+XraySUM)/totalSUM)*100
(EMSUM/totalSUM)*100
```

```
[1] 10.2348
```

```r
(XraySUM/totalSUM)*100
```

```
[1] 83.25592
```

```r
#answer is
#EM:10.2348
#Xray:83.25592
#Both:93.49072
#could also create a function and then do colSums(apply(PDBstats, 2, covert_comma_numbers))
#convert comma numbers is your function, I didn't include this in my code but it is basically
#could also do this in the tidyverse using the readr library and also read_csv, this will au

#Q2: What proportion of structures in the PDB are protein?
totalnocommas<-as.numeric(gsub(",","",PDBstats$Total))
sum(totalnocommas[1:3])
```

```
[1] 221940
```

```r
#protiens is 221940

#Q3: Type HIV in the PDB website search box on the home page and determine how many HIV-1 pro
#4563 structures.
```