Class 9: Structural Bioinformatics (pt.1)

Sofia Lanaspa, A17105313

The main database for strucutal data on proteins is PDB(protein data book), let see what it contain:

Data from: https://www.rcsb.org/stats

```
pdbdb <- read.csv("Data Export Summary.csv")
```

```
pdbdb$total
```

```
NULL
```

```
#issue because numbers are in quotes due to the coma
as.numeric(sub(",","",pdbdb$Total)) #convert values to numeric to do math
```

```
[1] 195610  12318  13720   4531    213     22
```

Could turn the code above into a function

```
comma2numeric <- function(x){
  as.numeric(sub(",","",x))
}
comma2numeric((pdbdb$X.ray)) #test function
```

```
[1] 167192   9639   8730   2869    170     11
```

```
apply(pdbdb,2,comma2numeric)
```

```
Warning in FUN(newX[, i], ...): NAs introduced by coercion
```

```
     Molecular.Type  X.ray    EM   NMR Multiple.methods Neutron Other  Total
[1,]             NA 167192 15572 12529              208      77    32 195610
[2,]             NA   9639  2635    34                8       2     0  12318
[3,]             NA   8730  4697   286                7       0     0  13720
[4,]             NA   2869   137  1507               14       3     1   4531
[5,]             NA    170    10    33                0       0     0    213
[6,]             NA     11     0     6                1       0     4     22
```

OR TRY DIFFERENT FUNCTION

```
library(readr)
pdbdb <- read_csv("Data Export Summary.csv")
```

```
Rows: 6 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Molecular Type
dbl (3): Multiple methods, Neutron, Other
num (4): X-ray, EM, NMR, Total

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
#Sum of all x-rays divided the Total
(sum(pdbdb$"X-ray")/sum(pdbdb$Total)) * 100
```

```
[1] 83.30359
```

Q2: What proportion of structures in the PDB are protein?

```
#First value of column 'Total' is only protein, divided by all the values in 'Total'
(pdbdb$Total[1]/ sum(pdbdb$Total)) * 100
```

```
[1] 86.39483
```