Class 9: Structural Bioinformatics pt.1

Noel Lim (PID: A17652474)

The main database for structural data is called the PBD (Protein Data Bank). Let's see what it contains:

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

Read this into R

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

and answer the following questions:

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

```
pdbdb$Total
```

```
[1] "195,866" "12,328"  "13,746"  "4,532"   "213"     "22"
```

I need to remove the comma and convert to numeric to do math:

```
as.numeric( sub(",","", pdbdb$Total) )
```

```
[1] 195866  12328  13746   4532    213     22
```

I could turn this into a function to fix the whole table or any future table I read like this:

```
x <- pdbdb$Total
as.numeric( sub(",","",x))
```

```
[1] 195866  12328  13746   4532    213     22
```

```
comma2numeric <- function(x) {
  as.numeric( sub(",","", x))
}
```

Test it

```
comma2numeric(pdbdb$X.ray)
```

```
[1] 167317   9645   8735   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 167317 15698 12534              208      77    32 195866
[2,]             NA   9645  2639    34                8       2     0  12328
[3,]             NA   8735  4718   286                7       0     0  13746
[4,]             NA   2869   138  1507               14       3     1   4532
[5,]             NA    170    10    33                0       0     0    213
[6,]             NA     11     0     6                1       0     4     22
```

Or try a different read/import 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.
```