**Practice on clean data: Starbuck drink**

**Step 1: Clean data**

1. Import the data file to R
2. The column names are not clean
3. Notice that some rows are totally missing. Try to delete all these rows in R (not Excel).
4. Some rows are duplicated:
   1. If they are totally the same, just keep a unique row
   2. Or keep the rows with more information
5. Some **size** has typo: instead of “venti”, we make typo to “vendi”
6. Any column has the wrong data type? If you can find them, let’s convert them into the correct data format.
7. The whip column should have only two values: 1 if we added whip cream and 0 otherwise. But we have some other values like 2 or 3. Please change all these typos to 1.
8. Some size is not quite popular, let’s skip them. Only keep these size: *short*, *tall*, *grande*, *venti*
9. Create a new column **milk\_type** as follow:

|  |  |
| --- | --- |
| Milk Type type of milk used | milk\_type |
| 0 | none |
| 1 | nonfat |
| 2 | 2% |
| 3 | soy |
| 4 | coconut |
| 5 | whole |

If you can find any other problems/errors in data, please share with us.

10. Save your final data into the “data/process” folder as the name: **starbucks\_clean.rds**

**Step 2: Summarize the data a bit**

1. How many drink we can order in Starbuck?
2. On average, a Starbuck drink has how much calories, sugar, and caffeine
3. Which drink has the largest and lowest calories, sugar, total fat, and caffeine?
4. Among all coffee drinks, which one has the largest sugar and caffeine? And the lowest?
5. After your analysis, which drink you want to recommend to your friends (e.g., girl or boyfriend)? And explain your choice.

**Appendix**

I borrowed and modified a bit data from this project: <https://github.com/PythonCoderUnicorn/starbucks>

Description of data can be found in the above link too.

Some visual intro about the data is here:

<https://www.behance.net/gallery/58743971/Starbucks-Menu-Infographic-Design>