Assessment Part 2: Data Import

In this part of the assessment, you will import real datasets and learn more about useful arguments to **readr** functions. You will encounter common issues that arise when importing raw data. This part of the assessment will require you to program in R.

Use the **readr** package in the **tidyverse** library:

```
library(tidyverse)
```

Question 14

1.0/1.0 point (graded)
Inspect the file at the following URL:

http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data

Which **readr** function should be used to import this file?

○ `read_table`

◉ `read_csv` ✔

○ `read_csv2`

○ `read_tsv`

○ None of the above

Answer
Correct: Correct - this is a comma-separated value file.

Submit	You have used 1 of 2 attempts

ⓘ   Answers are displayed within the problem

Question 15

1.0/1.0 point (graded)

Check the documentation for the **readr** function you chose in the previous question to learn about its arguments. Determine which arguments you need to the file from the previous question:

```
url <- "http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data"
```

Does this file have a header row? Does the **readr** function you chose need any additional arguments to import the data correctly?

○ Yes, there is a header. No arguments are needed.

○ Yes, there is a header. The `header=TRUE` argument is necessary.

○ Yes, there is a header. The `col_names=TRUE` argument is necessary.

○ No, there is no header. No arguments are needed.

○ No, there is no header. The `header=FALSE` argument is necessary.

● No, there is no header. The `col_names=FALSE` argument is necessary. ✔

Answer
Correct:
Correct! There are no variable names in the first row, and the correct argument to skip the header in **readr** is `col_names=FALSE` .

Submit   You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 16

2.0/2.0 points (graded)

Inspect the imported data from the previous question.

How many rows are in the dataset?

569   ✔ **Answer:** 569

569

Answer code

```
url <- "http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data"
         df <- read_csv(url, col_names = FALSE)
         nrow(df)
```

How many columns are in the dataset?

| 32 | ✔ **Answer:** 32 |

32

Answer code

```
ncol(df)
```

Submit   You have used 1 of 10 attempts

ⓘ Answers are displayed within the problem

Question 17

3.0/3.0 points (graded)
Read in the table from the following URL using a function from the **readr** library and save it as
`co2_mauna_loa` :

```
url <- "ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_annmean_mlo.txt"
```

Use the `skip` argument to skip all of the documentation rows so that the column names are
`c("year", "mean", "unc")` or `c("#", "year", "mean", "unc")` . (The "#" column could be removed
after import with, for example, `select` .

How many rows must be skipped?
This may require some experimentation - this is normal for data wrangling.

| 56 | ✔ **Answer:** 56 |

56

Answer code

```
co2_mauna_loa <- read_table(url, skip = 56) # skip documentation
co2_mauna_loa
```

Which **readr** function correctly imports the table as a data frame after skipping the documentation rows?

- ○ `read_csv`

- ○ `read_csv2`

- ○ `read_tsv`

- ● `read_table` ✔

- ○ `read_lines`

Answer code

```
        !between(0.5, x_hat - qnorm(.975)*se_hat, x_hat + qnorm(.975)*se_hat)    # predicts winner
        between(0.481, x_hat - qnorm(.975)*se_hat, x_hat + qnorm(.975)*se_hat)    # does not cover p
```

How many rows are in `co2_mauna_loa`?

60

✔ **Answer:** 60

60

Answer code

```
nrow(co2_mauna_loa)
```

Submit   You have used 1 of 10 attempts

ℹ Answers are displayed within the problem