SYS 5581 Project - Extract, Transform, and Load Data

Nick Coronato

Version of 2021-02-16 | Due 2021-02-22

Step 1: Identify a Time Series data set that you want to work with

For this project, I will be analyzing a set of exercise data for 186 patients.

Step 2: Acquire the data from its source location, reproducibly.

For this project, my data is stored #on my local machine

Note: Ideally the data will be stored at and read from a Github repository. (Note: permission was granted to use this data, and no identifiable patient information is included in the raw data.)

```
url = 'https://github.com/uva-eng-time-series-sp21/coronato-nicholas/blob/main/CPET_raw_data.csv'

(CPET_raw  <-  read_delim("CPET_raw_data2.csv",",",
                         col_types = cols(.default = col_character(),
                                          "HR" = col_double(),
                                          "VO2" = col_double(),
                                          "VO2/kg" = col_double(),
                                          "VCO2" = col_double(),
                                          "RQ" = col_double(),
                                          "VE" = col_double(),
                                          "VE/VO2" = col_double(),
                                          "VE/VCO2" = col_double(),
                                          "Work" = col_double(),
                                          "PetO2" = col_double(),
                                          "PetCO2" = col_double(),
                                          "VEO22"= col_double()
                         )))
```

```
## # A tibble: 16,564 x 30
##    PatientId SessionId Time  LocalTime TestLevel    HR SpO2    VO2 ‘VO2/kg‘
##    <chr>     <chr>     <chr> <chr>     <chr>     <dbl> <chr> <dbl>    <dbl>
##  1 1         1         0:00~ 0:00:20   Baseline     74 <NA>  0.601      6.4
##  2 1         1         0:00~ 0:00:40   Baseline     74 <NA>  0.492      5.2
##  3 1         1         0:01~ 0:01:00   Baseline     73 <NA>  0.476      5
##  4 1         1         0:01~ 0:01:20   Baseline     74 <NA>  0.44       4.7
##  5 1         1         0:01~ 0:01:40   Baseline     75 <NA>  0.452      4.8
##  6 1         1         0:02~ 0:02:00   Baseline     74 <NA>  0.467      4.9
##  7 1         1         0:02~ 0:02:20   Baseline     78 <NA>  0.536      5.7
##  8 1         1         0:02~ 0:00:20   Exercise     86 <NA>  0.808      8.6
##  9 1         1         0:03~ 0:00:40   Exercise     86 <NA>  0.696      7.4
## 10 1         1         0:03~ 0:01:00   Exercise     86 <NA>  0.796      8.4
## # ... with 16,554 more rows, and 21 more variables: VCO2 <dbl>, RQ <dbl>,
## #   VE <dbl>, ‘VE/VO2‘ <dbl>, ‘VE/VCO2‘ <dbl>, Work <dbl>, PetO2 <dbl>,
## #   PetCO2 <dbl>, VEO22 <dbl>, FEO2 <chr>, FECO2 <chr>, RER <chr>, RR <chr>,
## #   METS <chr>, TMSPD <chr>, TMELV <chr>, Vtex <chr>, Vtin <chr>, Source <chr>,
## #   TypeUser <chr>, Summary <chr>
```

Step 3: Organize your data into a *tidy* data frame.

Organize by taking out the non-useful variables.

```
CPET_raw <- select(CPET_raw, -LocalTime, -FEO2, -FECO2, -RER,   -RR,    -METS,  -Vtex,  -Vtin, -Source,
```

Make a new variable called Index so that each observation is individually identifiable, i.e. Session 1, Obs 1

```
#load package
require(data.table)

#  Turn data.frame into a data.table
CPET_ts2 <- data.table( CPET_raw )

#  Get running count by SessionId
CPET_ts2[ , Index := 1:.N , by = c("SessionId") ]
```

Make Index variable to be a two digit readout (i.e. 01, 02, . . .)

Convert time column into a more usable value (seconds instead of HH:MM:SS)

This can be used to create a dataframe of HR over time, per patient session.

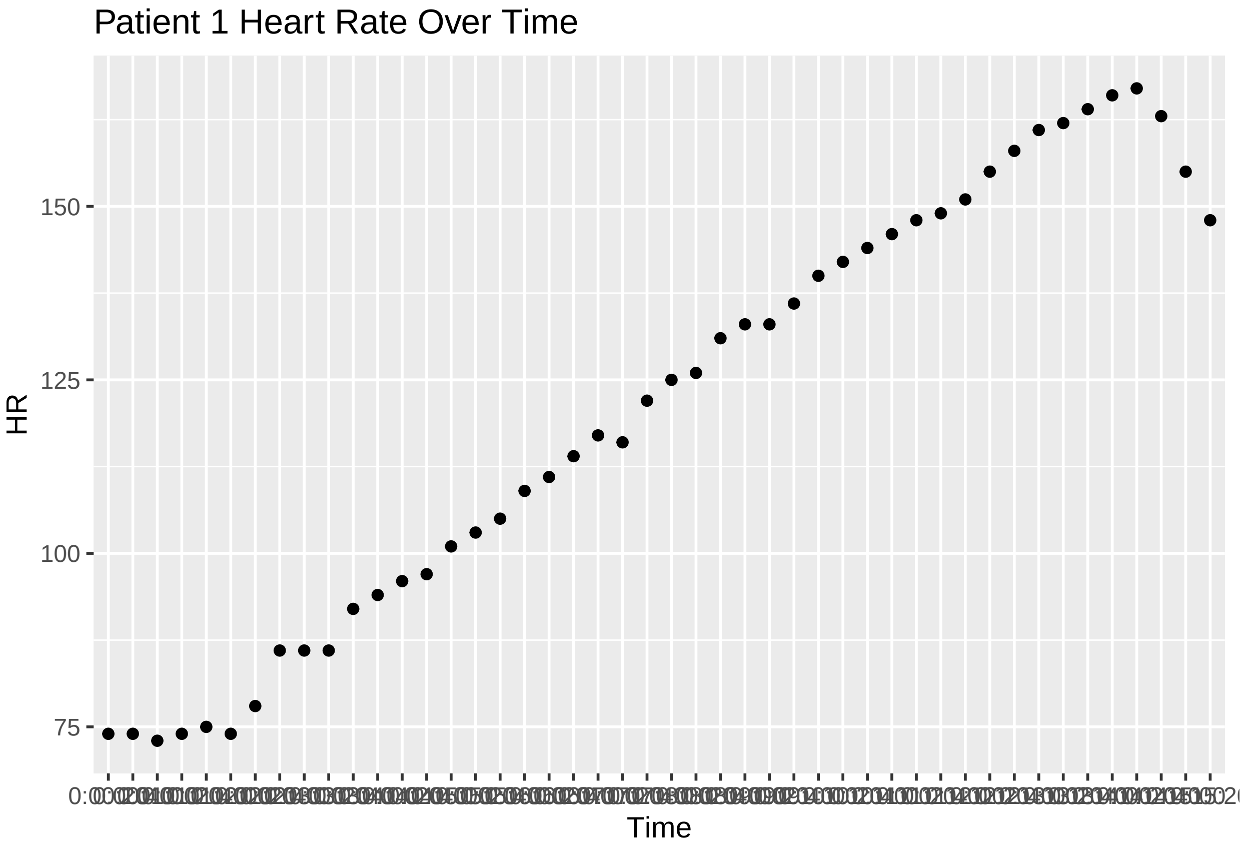
```
#This can be used to create a df of HR over time, per patient session

(CPET_ts %>%
  group_by(SessionId, NewTime) %>%
  summarise(HR) -> HR_by_patient)
```

```
## ‘summarise()‘ has grouped output by ’SessionId’, ’NewTime’. You can override using the ‘.groups‘ argu
```

```
## # A tibble: 16,564 x 3
## # Groups:   SessionId, NewTime [16,562]
##    SessionId NewTime     HR
##    <chr>     <Period> <dbl>
##  1 1         20S         74
##  2 1         40S         74
##  3 1         1M 0S       73
##  4 1         1M 20S      74
##  5 1         1M 40S      75
##  6 1         2M 0S       74
##  7 1         2M 20S      78
##  8 1         2M 40S      86
##  9 1         3M 0S       86
## 10 1         3M 20S      86
## # ... with 16,554 more rows
```

This chunk is for example purposes; ggplot of Patient 1’s heart rate over time.



Generate and print the tsibble.

```
## # A tsibble: 16,564 x 22 [1m 1s]
## # Key:       identifier [16,564]
##    PatientId SessionId identifier NewTime Time  TestLevel    HR SpO2    VO2
##    <chr>     <chr>     <chr>      <Perio> <chr> <chr>     <dbl> <chr> <dbl>
##  1 1         1         1.01       20S     0:00~ Baseline     74 <NA>  0.601
##  2 1         1         1.02       40S     0:00~ Baseline     74 <NA>  0.492
##  3 1         1         1.03       1M 0S   0:01~ Baseline     73 <NA>  0.476
##  4 1         1         1.04       1M 20S  0:01~ Baseline     74 <NA>  0.44
##  5 1         1         1.05       1M 40S  0:01~ Baseline     75 <NA>  0.452
##  6 1         1         1.06       2M 0S   0:02~ Baseline     74 <NA>  0.467
##  7 1         1         1.07       2M 20S  0:02~ Baseline     78 <NA>  0.536
##  8 1         1         1.08       2M 40S  0:02~ Exercise     86 <NA>  0.808
##  9 1         1         1.09       3M 0S   0:03~ Exercise     86 <NA>  0.696
## 10 1         1         1.10       3M 20S  0:03~ Exercise     86 <NA>  0.796
## # ... with 16,554 more rows, and 13 more variables: ‘VO2/kg‘ <dbl>, VCO2 <dbl>,
## #   RQ <dbl>, VE <dbl>, ‘VE/VO2‘ <dbl>, ‘VE/VCO2‘ <dbl>, Work <dbl>,
## #   PetO2 <dbl>, PetCO2 <dbl>, VEO22 <dbl>, TMSPD <chr>, TMELV <chr>,
## #   Index <chr>
```