HW01

Jackson Dial

2022-09-07

# Shell Scripting

### Download from the Sakai HW1 folder the file named “heart-disease.csv.gz”. Upload this file to your DCC home directory.

# sftp> put Data/heart-disease.csv.gz .

### Look at the first 5 lines and last 5 lines of the data file heart-disease.csv.gz without unzipping it. While this is not a very large dataset, we use it here to simulate working with a very large dataset that would be difficult or even impossible to open due to its size. Find out what “gzip” does by running the command “man gzip”. You may use “gzip” to complete this task.

#zcat heart-disease.csv.gz | head -n 6  
  
#zcat heart-disease.csv.gz | tail -n 5

### How many lines are there in this file? Find out without unzipping the file.

#zcat heart-disease.csv.gz | wc -l  
  
#There are 303 lines in the file

### Unzip heart-disease.csv.gz without deleting the original zipped file. You may use gunzip to accomplish this.

# gunzip -c heart-disease.csv.gz > heart-disease.csv

### Find out how many patients are diagnosed with heart disease. To accomplish this, use the “grep” to find the lines that contain the word “TRUE” and count the lines. Since there is only one column containing TRUE/FALSE, hence we don’t need to worry about searching by column here, but if there were other columns in the same dataset that also contained true/false, you would need to search by column.

# grep -c "TRUE" heart-disease.csv  
# 139

### How many columns and how many rows are there in this dataset? Is the row count the same as the previous row count in question 3? If it is different, why do you think so? Show your commands that lead you to your answer.

# awk '{print NF}' heart-disease.csv | sort -nu | tail -n 1  
# there are 16 columns including the index column, 15 not including the index column  
  
# cat heart-disease.csv | wc -l  
# there are 303 rows  
  
# the row count is the same as above.

# Complete a Tutorial for “dplyr”

# install.packages("tidyverse")  
# install.packages("devtools")

library(tidyverse)  
# library(devtools)  
# devtools::install\_github("tidyverse/dplyr")

starwars %>%   
 filter(species == "Droid")

## # A tibble: 6 x 14  
## name height mass hair\_color skin\_color eye\_color birth\_year sex gender   
## <chr> <int> <dbl> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 C-3PO 167 75 <NA> gold yellow 112 none masculi~  
## 2 R2-D2 96 32 <NA> white, blue red 33 none masculi~  
## 3 R5-D4 97 32 <NA> white, red red NA none masculi~  
## 4 IG-88 200 140 none metal red 15 none masculi~  
## 5 R4-P17 96 NA none silver, red red, blue NA none feminine  
## 6 BB8 NA NA none none black NA none masculi~  
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,  
## # vehicles <list>, starships <list>

starwars %>%   
 select(name, ends\_with("color"))

## # A tibble: 87 x 4  
## name hair\_color skin\_color eye\_color  
## <chr> <chr> <chr> <chr>   
## 1 Luke Skywalker blond fair blue   
## 2 C-3PO <NA> gold yellow   
## 3 R2-D2 <NA> white, blue red   
## 4 Darth Vader none white yellow   
## 5 Leia Organa brown light brown   
## 6 Owen Lars brown, grey light blue   
## 7 Beru Whitesun lars brown light blue   
## 8 R5-D4 <NA> white, red red   
## 9 Biggs Darklighter black light brown   
## 10 Obi-Wan Kenobi auburn, white fair blue-gray  
## # ... with 77 more rows

starwars %>%   
 mutate(name, bmi = mass / ((height / 100) ^ 2)) %>%  
 select(name:mass, bmi)

## # A tibble: 87 x 4  
## name height mass bmi  
## <chr> <int> <dbl> <dbl>  
## 1 Luke Skywalker 172 77 26.0  
## 2 C-3PO 167 75 26.9  
## 3 R2-D2 96 32 34.7  
## 4 Darth Vader 202 136 33.3  
## 5 Leia Organa 150 49 21.8  
## 6 Owen Lars 178 120 37.9  
## 7 Beru Whitesun lars 165 75 27.5  
## 8 R5-D4 97 32 34.0  
## 9 Biggs Darklighter 183 84 25.1  
## 10 Obi-Wan Kenobi 182 77 23.2  
## # ... with 77 more rows

starwars %>%   
 arrange(desc(mass))

## # A tibble: 87 x 14  
## name height mass hair\_color skin\_color eye\_color birth\_year sex gender  
## <chr> <int> <dbl> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 Jabba ~ 175 1358 <NA> green-tan~ orange 600 herm~ mascu~  
## 2 Grievo~ 216 159 none brown, wh~ green, y~ NA male mascu~  
## 3 IG-88 200 140 none metal red 15 none mascu~  
## 4 Darth ~ 202 136 none white yellow 41.9 male mascu~  
## 5 Tarfful 234 136 brown brown blue NA male mascu~  
## 6 Owen L~ 178 120 brown, grey light blue 52 male mascu~  
## 7 Bossk 190 113 none green red 53 male mascu~  
## 8 Chewba~ 228 112 brown unknown blue 200 male mascu~  
## 9 Jek To~ 180 110 brown fair blue NA male mascu~  
## 10 Dexter~ 198 102 none brown yellow NA male mascu~  
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,  
## # films <list>, vehicles <list>, starships <list>

starwars %>%  
 group\_by(species) %>%  
 summarise(  
 n = n(),  
 mass = mean(mass, na.rm = TRUE)  
 ) %>%  
 filter(  
 n > 1,  
 mass > 50  
 )

## # A tibble: 8 x 3  
## species n mass  
## <chr> <int> <dbl>  
## 1 Droid 6 69.8  
## 2 Gungan 3 74   
## 3 Human 35 82.8  
## 4 Kaminoan 2 88   
## 5 Mirialan 2 53.1  
## 6 Twi'lek 2 55   
## 7 Wookiee 2 124   
## 8 Zabrak 2 80

# Dataset Summary and Plotting

### Read the dataset into R by using a code chunk in your Rmd file.

# setwd("C:/Users/jacks/OneDrive/Desktop/DUKE\_FALL2022/707/BIOSTAT707")  
dat <- read\_csv("Data/heart\_failure.csv")

### Use the summary function to show the summary statistics for the dataset. Print the summary. How many rows are there in this dataframe?

library(pander)  
summary(dat) %>%   
 pander()

Table continues below

| age | anaemia | creatinine\_phosphokinase | diabetes |
| --- | --- | --- | --- |
| Min. :40.00 | Min. :0.0000 | Min. : 23.0 | Min. :0.0000 |
| 1st Qu.:51.00 | 1st Qu.:0.0000 | 1st Qu.: 116.5 | 1st Qu.:0.0000 |
| Median :60.00 | Median :0.0000 | Median : 250.0 | Median :0.0000 |
| Mean :60.83 | Mean :0.4314 | Mean : 581.8 | Mean :0.4181 |
| 3rd Qu.:70.00 | 3rd Qu.:1.0000 | 3rd Qu.: 582.0 | 3rd Qu.:1.0000 |
| Max. :95.00 | Max. :1.0000 | Max. :7861.0 | Max. :1.0000 |

Table continues below

| ejection\_fraction | high\_blood\_pressure | platelets | serum\_creatinine |
| --- | --- | --- | --- |
| Min. :14.00 | Min. :0.0000 | Min. : 25100 | Min. :0.500 |
| 1st Qu.:30.00 | 1st Qu.:0.0000 | 1st Qu.:212500 | 1st Qu.:0.900 |
| Median :38.00 | Median :0.0000 | Median :262000 | Median :1.100 |
| Mean :38.08 | Mean :0.3512 | Mean :263358 | Mean :1.394 |
| 3rd Qu.:45.00 | 3rd Qu.:1.0000 | 3rd Qu.:303500 | 3rd Qu.:1.400 |
| Max. :80.00 | Max. :1.0000 | Max. :850000 | Max. :9.400 |

| serum\_sodium | sex | smoking | time | DEATH\_EVENT |
| --- | --- | --- | --- | --- |
| Min. :113.0 | Min. :0.0000 | Min. :0.0000 | Min. : 4.0 | Min. :0.0000 |
| 1st Qu.:134.0 | 1st Qu.:0.0000 | 1st Qu.:0.0000 | 1st Qu.: 73.0 | 1st Qu.:0.0000 |
| Median :137.0 | Median :1.0000 | Median :0.0000 | Median :115.0 | Median :0.0000 |
| Mean :136.6 | Mean :0.6488 | Mean :0.3211 | Mean :130.3 | Mean :0.3211 |
| 3rd Qu.:140.0 | 3rd Qu.:1.0000 | 3rd Qu.:1.0000 | 3rd Qu.:203.0 | 3rd Qu.:1.0000 |
| Max. :148.0 | Max. :1.0000 | Max. :1.0000 | Max. :285.0 | Max. :1.0000 |

nrow(dat) #there are 299 rows in this df

## [1] 299

### How many people have anaemia? (Hint: Use functions in the dplyr package to help you get to this number)

table(dat$anaemia)

##   
## 0 1   
## 170 129

dat %>%   
 filter(anaemia == 1) %>%   
 nrow()

## [1] 129

#129 people have anaemia

### How many death events occur in people who smoke?

dat %>%   
 filter(smoking == 1,  
 DEATH\_EVENT == 1) %>%   
 nrow()

## [1] 30

# There are 30 death events in people who smoke

### Plot a histogram of the number of deaths for each diabetes group using ggplot2. You may have to install the package using the function install.packages(“ggplot2”) and read through the ggplot tutorial to learn how to use ggplot2 commands. [Hint: remember that a death event is not continuous and should be designated as a Factor data type in R]

ggplot(dat, aes(x = as.factor(diabetes), y = DEATH\_EVENT))+  
 geom\_bar(stat = "identity")+  
 theme(panel.grid.minor = element\_blank(),  
 panel.grid.major.x = element\_blank())+  
 labs(x = "Diabetes? (0 = No, 1 = Yes)",  
 y = "Number of Death Events")

