Homework 1

Costa Stavrianidis

2022-09-11

Shell Scripting

Question 1

```
sftp> put heart-disease.csv.gz .
Uploading heart-disease.csv.gz to /hpc/home/cs621/./heart-disease.csv.gz
heart-disease.csv.gz                                  100% 4314   111.2KB/s   00:00
```

Question 2

```
(base) cs621@dcc-login-01  ~ $ zcat heart-disease.csv.gz | head -5
"age" "sex" "cp" "trestbps" "chol" "fbs" "restecg" "thalach" "exang" "oldpeak" "slope
" "ca" "thal" "num" "diagnosed"
"1" 67 1 4 160 286 0 2 108 1 1.5 2 "3.0" "3.0" 2 TRUE
"2" 67 1 4 120 229 0 2 129 1 2.6 2 "2.0" "7.0" 1 TRUE
"3" 37 1 3 130 250 0 0 187 0 3.5 3 "0.0" "3.0" 0 FALSE
"4" 41 0 2 130 204 0 2 172 0 1.4 1 "0.0" "3.0" 0 FALSE

(base) cs621@dcc-login-01  ~ $ zcat heart-disease.csv.gz | tail -5
"298" 45 1 1 110 264 0 0 132 0 1.2 2 "0.0" "7.0" 1 TRUE
"299" 68 1 4 144 193 1 0 141 0 3.4 2 "2.0" "7.0" 2 TRUE
"300" 57 1 4 130 131 0 0 115 1 1.2 2 "1.0" "7.0" 3 TRUE
"301" 57 0 2 130 236 0 2 174 0 0 2 "1.0" "3.0" 1 TRUE
"302" 38 1 3 138 175 0 0 173 0 0 1 "?" "3.0" 0 FALSE
```

Question 3

```
(base) cs621@dcc-login-01  ~ $ zcat heart-disease.csv.gz | wc -l
303
```

Question 4

```
(base) cs621@dcc-login-01  ~ $ gunzip -k heart-disease.csv.gz
(base) cs621@dcc-login-01  ~ $ ls
heart-disease.csv  heart-disease.csv.gz  R  testdirectory
```

Question 5

```
(base) cs621@dcc-login-01  ~ $ grep -c "TRUE" heart-disease.csv
139
```

Question 6

```
# Columns
(base) cs621@dcc-login-01  ~ $ awk '{print NF; exit}' heart-disease.csv
15

# Rows
(base) cs621@dcc-login-01  ~ $ wc -l heart-disease.csv
303 heart-disease.csv

# Row count is the same as in question 3.
```

Complete a Tutorial for “dplyr”

Question 8

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

```
## # A tibble: 6 × 14
##   name   height  mass hair_color skin_color eye_c…¹ birth…² sex   gender homew…³
##   <chr>   <int> <dbl> <chr>      <chr>      <chr>     <dbl> <chr> <chr>  <chr>
## 1 C-3PO     167    75 <NA>       gold       yellow      112 none  mascu… Tatooi…
## 2 R2-D2      96    32 <NA>       white, bl… red          33 none  mascu… Naboo
## 3 R5-D4      97    32 <NA>       white, red red          NA none  mascu… Tatooi…
## 4 IG-88     200   140 none       metal      red          15 none  mascu… <NA>
## 5 R4-P17     96    NA none       silver, r… red, b…      NA none  femin… <NA>
## 6 BB8        NA    NA none       none       black        NA none  mascu… <NA>
## # … with 4 more variables: species <chr>, films <list>, vehicles <list>,
## #   starships <list>, and abbreviated variable names ¹eye_color, ²birth_year,
## #   ³homeworld
```

```
starwars %>%
  select(name, ends_with("color"))
```

```
## # A tibble: 87 × 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 × 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 × 14
##    name        height  mass hair_…¹ skin_…² eye_c…³ birth…⁴ sex   gender homew…⁵
##    <chr>        <int> <dbl> <chr>   <chr>   <chr>     <dbl> <chr> <chr>  <chr>
##  1 Jabba Desi…    175  1358 <NA>    green-… orange    600   herm… mascu… Nal Hu…
##  2 Grievous       216   159 none    brown,… green,…    NA   male  mascu… Kalee
##  3 IG-88          200   140 none    metal   red        15   none  mascu… <NA>
##  4 Darth Vader    202   136 none    white   yellow     41.9 male  mascu… Tatooi…
##  5 Tarfful        234   136 brown   brown   blue       NA   male  mascu… Kashyy…
##  6 Owen Lars      178   120 brown,… light   blue       52   male  mascu… Tatooi…
##  7 Bossk          190   113 none    green   red        53   male  mascu… Trando…
##  8 Chewbacca      228   112 brown   unknown blue      200   male  mascu… Kashyy…
##  9 Jek Tono P…    180   110 brown   fair    blue       NA   male  mascu… Bestin…
## 10 Dexter Jet…    198   102 none    brown   yellow     NA   male  mascu… Ojom
## # … with 77 more rows, 4 more variables: species <chr>, films <list>,
## #   vehicles <list>, starships <list>, and abbreviated variable names
## #   ¹hair_color, ²skin_color, ³eye_color, ⁴birth_year, ⁵homeworld
```

```
starwars %>%
  group_by(species) %>%
  summarise(
    n = n(),
    mass = mean(mass, na.rm = TRUE)
  ) %>%
  filter(
    n > 1,
    mass > 50
  )
```

```
## # A tibble: 8 × 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

Question 9

```
heart <- read_csv("heart_failure.csv", show_col_types = FALSE)
```

Question 10

```
summary(heart)
```

```
##       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
##  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
##  Min.   :113.0   Min.   :0.0000   Min.   :0.0000   Min.   :  4.0
##  1st Qu.:134.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 73.0
##  Median :137.0   Median :1.0000   Median :0.0000   Median :115.0
##  Mean   :136.6   Mean   :0.6488   Mean   :0.3211   Mean   :130.3
##  3rd Qu.:140.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:203.0
##  Max.   :148.0   Max.   :1.0000   Max.   :1.0000   Max.   :285.0
##   DEATH_EVENT
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3211
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
nrow(heart)
```

```
## [1] 299
```

```
# There are 299 rows in this dataset.
```

Question 11

```
heart %>% count(anaemia)
```

```
## # A tibble: 2 × 2
##   anaemia     n
##     <dbl> <int>
## 1       0   170
## 2       1   129
```

```
# There are 129 people with anaemia.
```

Question 12

```
heart %>% filter(smoking == 1) %>% count(DEATH_EVENT)
```

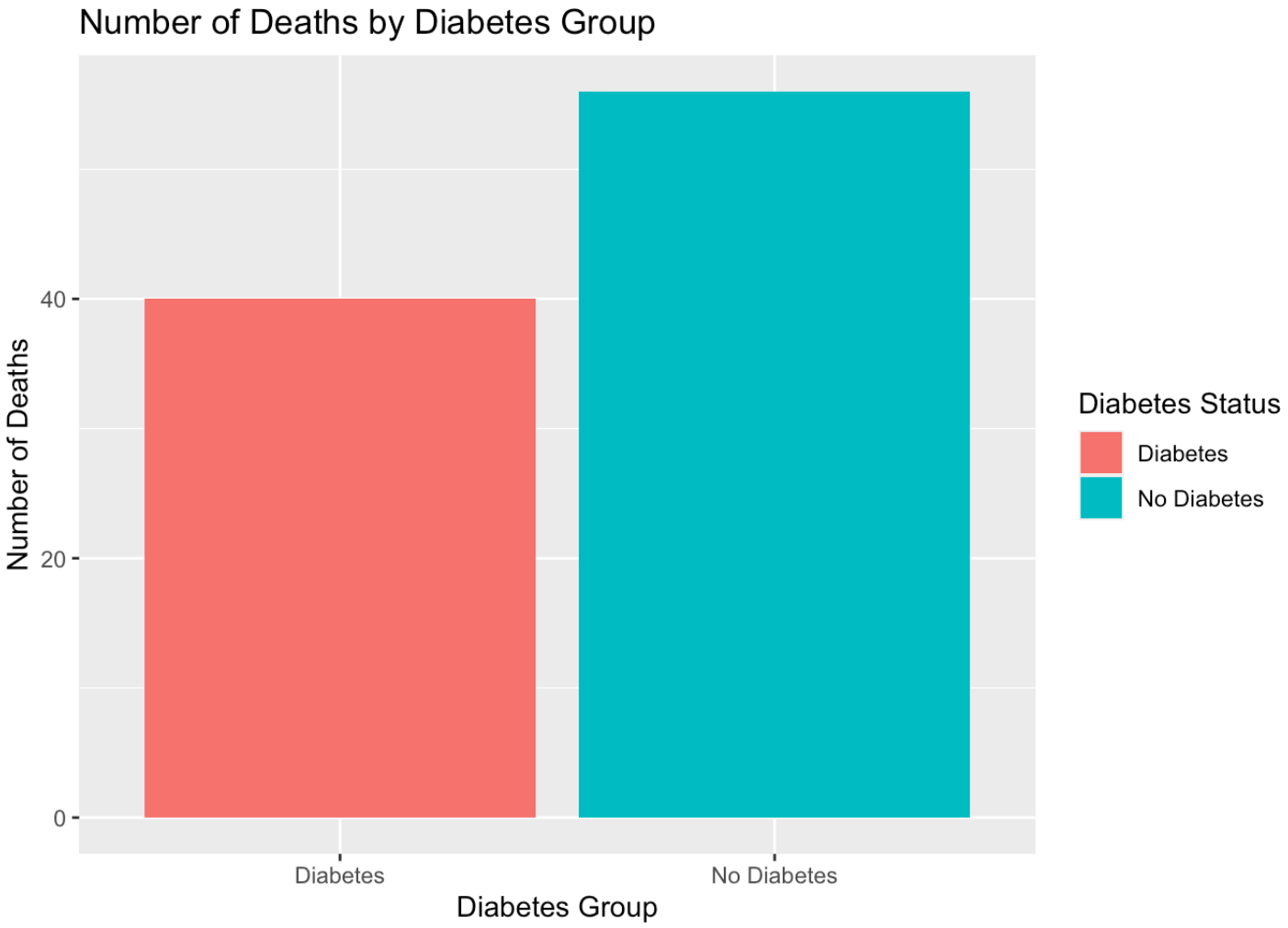
```
## # A tibble: 2 × 2
##   DEATH_EVENT     n
##         <dbl> <int>
## 1           0    66
## 2           1    30
```

```
# There were 30 death events in people who smoked.
```

Question 13

```
heart_death <- heart %>% filter(DEATH_EVENT == 1) %>% mutate(diabetes1 = ifelse(diabe
tes == 0, "No Diabetes", "Diabetes"))

ggplot(heart_death, aes(x=diabetes1, fill=diabetes1)) + geom_bar() +
  ggtitle("Number of Deaths by Diabetes Group") + xlab("Diabetes Group") +
  ylab("Number of Deaths") + labs(fill="Diabetes Status")
```


Number of Deaths by Diabetes Group
Number of Deaths
40
20
0
Diabetes
No Diabetes
Diabetes Group
Diabetes Status
Diabetes
No Diabetes