Chapter 4: Data Management

Answers to Exercises

Brian Fogarty

Contents

Answers to Exercise 1

```
library(tidyverse)
library(readxl)

simd <- read_csv("simd2020.csv", na="*")

healthboard <- read_xlsx("scottish health boards by datazone.xlsx")

covid <- read_xlsx("covid total by health board.xlsx")
```

Answers to Exercise 2

Exercise 2.a

```
simd %>%
  count(Council_area, sort=TRUE)
```

```
# A tibble: 32 x 2
   Council_area          n
   <chr>             <int>
 1 Glasgow City        746
 2 City of Edinburgh   597
 3 Fife                494
 4 North Lanarkshire   447
 5 South Lanarkshire   431
 6 Aberdeenshire       340
 7 Highland            312
 8 Aberdeen City       283
 9 West Lothian        239
10 Renfrewshire        225
# ... with 22 more rows
```

Exercise 2.b

```
healthboard %>%
  count(health_board, sort=TRUE)
```

```
# A tibble: 14 x 2
   health_board                 n
   <chr>                    <int>
 1 Greater Glasgow and Clyde  1458
 2 Lothian                    1083
 3 Lanarkshire                 878
 4 Grampian                    749
 5 Tayside                     529
 6 Ayrshire and Arran          502
 7 Fife                        494
 8 Highland                    437
 9 Forth Valley                407
10 Dumfries and Galloway       201
11 Borders                     143
12 Western Isles                36
13 Shetland                     30
14 Orkney                       29
```

Exercise 2.c

```
covid %>%
  count(health_board, sort=TRUE)
```

```
# A tibble: 14 x 2
   health_board                 n
   <chr>                    <int>
 1 Ayrshire and Arran           1
 2 Borders                      1
 3 Dumfries and Galloway        1
 4 Fife                         1
 5 Forth Valley                 1
 6 Grampian                     1
 7 Greater Glasgow and Clyde    1
 8 Highland                     1
```

```
 9 Lanarkshire                  1
10 Lothian                      1
11 Orkney                       1
12 Shetland                     1
13 Tayside                      1
14 Western Isles                1
```

The reason for the different number of observations is simply because the `healthboard` dataset's unit of analysis (i.e., observations) is datazone and the `covid` dataset's unit of analysis is health board. So, there are 1,458 datazones in the Greater Glasgow and Clyde health board, but there is only 1 health board for Greater Glasgow and Clyde.

Answers to Exercise 3

We use the `filter()` and `is.na()` functions for this exercise.

Exercise 3.a

```
simd %>%
  filter(is.na(Broadband))
```

```
# A tibble: 2 x 38
  Data_Zone Intermediate_Zone Council_area Total_population Working_age_populat~
  <chr>     <chr>             <chr>                  <dbl>                <dbl>
1 S01010206 Petershill        Glasgow City               0                    0
2 S01010226 Sighthill         Glasgow City               0                    0
# ... with 33 more variables: Income_rate <dbl>, Income_count <dbl>,
#   Employment_rate <dbl>, Employment_count <dbl>, CIF <dbl>, ALCOHOL <dbl>,
#   DRUG <dbl>, SMR <dbl>, DEPRESS <dbl>, LBWT <dbl>, EMERG <dbl>,
#   Attendance <dbl>, Attainment <dbl>, no_qualifications <dbl>,
#   not_participating <dbl>, University <dbl>, drive_petrol <dbl>,
#   drive_GP <dbl>, drive_post <dbl>, drive_primary <dbl>, drive_retail <dbl>,
#   drive_secondary <dbl>, PT_GP <dbl>, PT_post <dbl>, PT_retail <dbl>, ...
```

```
simd %>%
  filter(!is.na(Broadband))
```

```
# A tibble: 6,974 x 38
   Data_Zone Intermediate_Zone   Council_area  Total_population Working_age_pop~
   <chr>     <chr>               <chr>                    <dbl>            <dbl>
 1 S01006506 Culter              Aberdeen City              894              580
 2 S01006507 Culter              Aberdeen City              793              470
 3 S01006508 Culter              Aberdeen City              624              461
 4 S01006509 Culter              Aberdeen City              537              307
 5 S01006510 Culter              Aberdeen City              663              415
 6 S01006511 Culter              Aberdeen City              759              453
 7 S01006512 Culter              Aberdeen City              539              345
 8 S01006513 Cults, Bieldside a~ Aberdeen City              788              406
 9 S01006514 Cults, Bieldside a~ Aberdeen City             1123              709
10 S01006515 Cults, Bieldside a~ Aberdeen City              816              529
# ... with 6,964 more rows, and 33 more variables: Income_rate <dbl>,
#   Income_count <dbl>, Employment_rate <dbl>, Employment_count <dbl>,
#   CIF <dbl>, ALCOHOL <dbl>, DRUG <dbl>, SMR <dbl>, DEPRESS <dbl>, LBWT <dbl>,
#   EMERG <dbl>, Attendance <dbl>, Attainment <dbl>, no_qualifications <dbl>,
```

```
#   not_participating <dbl>, University <dbl>, drive_petrol <dbl>,
#   drive_GP <dbl>, drive_post <dbl>, drive_primary <dbl>, drive_retail <dbl>,
#   drive_secondary <dbl>, PT_GP <dbl>, PT_post <dbl>, PT_retail <dbl>, ...
```

There are 2 missing values and 6,974 non-missing values for `Broadband`. The missing and non-missing values for a given variable should always add up to the total number of observations in the dataset; unless there is some weird error in the dataset.

Exercise 3.b

```
simd %>%
  filter(is.na(Employment_rate))
```

```
# A tibble: 3 x 38
  Data_Zone Intermediate_Zone Council_area Total_population Working_age_populat~
  <chr>     <chr>             <chr>                  <dbl>                <dbl>
1 S01010206 Petershill        Glasgow City               0                    0
2 S01010226 Sighthill         Glasgow City               0                    0
3 S01010227 Sighthill         Glasgow City               0                    0
# ... with 33 more variables: Income_rate <dbl>, Income_count <dbl>,
#   Employment_rate <dbl>, Employment_count <dbl>, CIF <dbl>, ALCOHOL <dbl>,
#   DRUG <dbl>, SMR <dbl>, DEPRESS <dbl>, LBWT <dbl>, EMERG <dbl>,
#   Attendance <dbl>, Attainment <dbl>, no_qualifications <dbl>,
#   not_participating <dbl>, University <dbl>, drive_petrol <dbl>,
#   drive_GP <dbl>, drive_post <dbl>, drive_primary <dbl>, drive_retail <dbl>,
#   drive_secondary <dbl>, PT_GP <dbl>, PT_post <dbl>, PT_retail <dbl>, ...
```

```
simd %>%
  filter(!is.na(Employment_rate))
```

```
# A tibble: 6,973 x 38
   Data_Zone Intermediate_Zone   Council_area  Total_population Working_age_pop~
   <chr>     <chr>               <chr>                    <dbl>            <dbl>
 1 S01006506 Culter              Aberdeen City              894              580
 2 S01006507 Culter              Aberdeen City              793              470
 3 S01006508 Culter              Aberdeen City              624              461
 4 S01006509 Culter              Aberdeen City              537              307
 5 S01006510 Culter              Aberdeen City              663              415
 6 S01006511 Culter              Aberdeen City              759              453
 7 S01006512 Culter              Aberdeen City              539              345
 8 S01006513 Cults, Bieldside a~ Aberdeen City              788              406
 9 S01006514 Cults, Bieldside a~ Aberdeen City             1123              709
10 S01006515 Cults, Bieldside a~ Aberdeen City              816              529
# ... with 6,963 more rows, and 33 more variables: Income_rate <dbl>,
#   Income_count <dbl>, Employment_rate <dbl>, Employment_count <dbl>,
#   CIF <dbl>, ALCOHOL <dbl>, DRUG <dbl>, SMR <dbl>, DEPRESS <dbl>, LBWT <dbl>,
#   EMERG <dbl>, Attendance <dbl>, Attainment <dbl>, no_qualifications <dbl>,
#   not_participating <dbl>, University <dbl>, drive_petrol <dbl>,
#   drive_GP <dbl>, drive_post <dbl>, drive_primary <dbl>, drive_retail <dbl>,
#   drive_secondary <dbl>, PT_GP <dbl>, PT_post <dbl>, PT_retail <dbl>, ...
```

There are 3 missing values and 6,973 non-missing values for `Employment_rate`.

Exercise 3.c

```
simd %>%
  filter(is.na(drive_GP))
```

```
# A tibble: 0 x 38
# ... with 38 variables: Data_Zone <chr>, Intermediate_Zone <chr>,
#   Council_area <chr>, Total_population <dbl>, Working_age_population <dbl>,
#   Income_rate <dbl>, Income_count <dbl>, Employment_rate <dbl>,
#   Employment_count <dbl>, CIF <dbl>, ALCOHOL <dbl>, DRUG <dbl>, SMR <dbl>,
#   DEPRESS <dbl>, LBWT <dbl>, EMERG <dbl>, Attendance <dbl>, Attainment <dbl>,
#   no_qualifications <dbl>, not_participating <dbl>, University <dbl>,
#   drive_petrol <dbl>, drive_GP <dbl>, drive_post <dbl>, ...
```

```
simd %>%
  filter(!is.na(drive_GP))
```

```
# A tibble: 6,976 x 38
   Data_Zone Intermediate_Zone   Council_area  Total_population Working_age_pop~
   <chr>     <chr>               <chr>                    <dbl>            <dbl>
 1 S01006506 Culter              Aberdeen City              894              580
 2 S01006507 Culter              Aberdeen City              793              470
 3 S01006508 Culter              Aberdeen City              624              461
 4 S01006509 Culter              Aberdeen City              537              307
 5 S01006510 Culter              Aberdeen City              663              415
 6 S01006511 Culter              Aberdeen City              759              453
 7 S01006512 Culter              Aberdeen City              539              345
 8 S01006513 Cults, Bieldside a~ Aberdeen City              788              406
 9 S01006514 Cults, Bieldside a~ Aberdeen City             1123              709
10 S01006515 Cults, Bieldside a~ Aberdeen City              816              529
# ... with 6,966 more rows, and 33 more variables: Income_rate <dbl>,
#   Income_count <dbl>, Employment_rate <dbl>, Employment_count <dbl>,
#   CIF <dbl>, ALCOHOL <dbl>, DRUG <dbl>, SMR <dbl>, DEPRESS <dbl>, LBWT <dbl>,
#   EMERG <dbl>, Attendance <dbl>, Attainment <dbl>, no_qualifications <dbl>,
#   not_participating <dbl>, University <dbl>, drive_petrol <dbl>,
#   drive_GP <dbl>, drive_post <dbl>, drive_primary <dbl>, drive_retail <dbl>,
#   drive_secondary <dbl>, PT_GP <dbl>, PT_post <dbl>, PT_retail <dbl>, ...
```

There are 0 missing values and 6,976 non-missing values for `drive_GP`.

Answers to Exercise 4

Exercise 4.a

```
simd_sub <- simd %>%
  select(Intermediate_Zone, Council_area, Broadband, Employment_rate, drive_GP)

glimpse(simd_sub)
```

```
Rows: 6,976
Columns: 5
$ Intermediate_Zone <chr> "Culter", "Culter", "Culter", "Culter", "Culter", "C~
$ Council_area      <chr> "Aberdeen City", "Aberdeen City", "Aberdeen City", "~
$ Broadband         <dbl> 0.105050505, 0.013586957, 0.005633803, 0.113074205, ~
```

```
$ Employment_rate   <dbl> 0.08, 0.05, 0.04, 0.08, 0.08, 0.04, 0.02, 0.03, 0.02~
$ drive_GP          <dbl> 3.074295, 4.309812, 3.784549, 2.778026, 2.358335, 1.~
```

Exercise 4.b

We'll do the rest of this exercise without saving the filtered versions as new objects.

```
simd_sub %>%
  filter(Council_area=="City of Edinburgh")
```

```
# A tibble: 597 x 5
   Intermediate_Zone          Council_area    Broadband Employment_rate drive_GP
   <chr>                      <chr>                <dbl>           <dbl>    <dbl>
 1 Balerno and Bonnington Vi~ City of Edinbu~   0.153              0.03     7.47
 2 Balerno and Bonnington Vi~ City of Edinbu~   0.0742             0.02     7.66
 3 Balerno and Bonnington Vi~ City of Edinbu~   0                  0.06     6.39
 4 Balerno and Bonnington Vi~ City of Edinbu~   0                  0.03     6.60
 5 Balerno and Bonnington Vi~ City of Edinbu~   0                  0.1      6.02
 6 Balerno and Bonnington Vi~ City of Edinbu~   0                  0.02     6.12
 7 Balerno and Bonnington Vi~ City of Edinbu~   0.133              0.02     5.51
 8 Balerno and Bonnington Vi~ City of Edinbu~   0.00885            0.03     3.76
 9 Currie West                City of Edinbu~   0.193              0.01     3.41
10 Currie West                City of Edinbu~   0                  0.08     1.57
# ... with 587 more rows
```

There are 597 datazones for Edinburgh.

Exercise 4.c

```
simd_sub %>%
  filter(Council_area=="City of Edinburgh" & Employment_rate >= .25)
```

```
# A tibble: 12 x 5
   Intermediate_Zone           Council_area   Broadband Employment_rate drive_GP
   <chr>                       <chr>               <dbl>           <dbl>    <dbl>
 1 Clovenstone and Wester Hai~ City of Edinb~   0                  0.26    0.868
 2 The Calders                 City of Edinb~   0                  0.26    3.00
 3 Murrayburn and Wester Hail~ City of Edinb~   0                  0.3     3.58
 4 Murrayburn and Wester Hail~ City of Edinb~   0                  0.25    3.05
 5 Moredun and Craigour        City of Edinb~   0                  0.36    3.25
 6 Niddrie                     City of Edinb~   0                  0.25    2.37
 7 Niddrie                     City of Edinb~   0.0220             0.28    2.59
 8 Bingham, Magdalene and The~ City of Edinb~   0                  0.32    1.80
 9 Restalrig and Lochend       City of Edinb~   0                  0.27    2.88
10 Restalrig and Lochend       City of Edinb~   0                  0.25    1.22
11 Great Junction Street       City of Edinb~   0                  0.4     1.62
12 Muirhouse                   City of Edinb~   0.00651            0.36    2.43
```

There are 12 datazones in Edinburgh where 25% or more of its residents are employment deprived.

Exercise 4.d

```
simd_sub %>%
  filter(Council_area=="City of Edinburgh" & Broadband >= .25)
```

```
# A tibble: 5 x 5
  Intermediate_Zone            Council_area   Broadband Employment_rate drive_GP
  <chr>                        <chr>               <dbl>           <dbl>    <dbl>
1 Newington and Dalkeith Road  City of Edinb~     0.352            0.01     2.61
2 Old Town, Princes Street an~ City of Edinb~     0.253            0.16     1.64
3 Craigmillar                  City of Edinb~     0.292            0.13     2.23
4 Ratho, Ingliston and Gogar   City of Edinb~     0.272            0.04     5.23
5 Queensferry West             City of Edinb~     0.352            0.02     3.26
```

There are 5 datazones in Edinburgh where 25% or more of its premises do not have access to superfast broadband.

Exercise 4.e

```
simd_sub %>%
  filter(Council_area=="City of Edinburgh" & Employment_rate >= .25 & Broadband >= .25)
```

```
# A tibble: 0 x 5
# ... with 5 variables: Intermediate_Zone <chr>, Council_area <chr>,
#   Broadband <dbl>, Employment_rate <dbl>, drive_GP <dbl>
```

There are 0 datazones in Edinburgh where 25% or more of its residents are employment deprived AND 25% or more of its premises do not have access to superfast broadband.

Answers to Exercise 5

Let's do the merge as a two-step process. We'll first merge `simd` and `healthboard`, and then `covid`.

```
merged <- simd %>%
  inner_join(healthboard, by=c("Data_Zone"="data_zone"))

glimpse(merged)
```

```
Rows: 6,976
Columns: 41
$ Data_Zone              <chr> "S01006506", "S01006507", "S01006508", "S010065~
$ Intermediate_Zone      <chr> "Culter", "Culter", "Culter", "Culter", "Culter~
$ Council_area           <chr> "Aberdeen City", "Aberdeen City", "Aberdeen Cit~
$ Total_population       <dbl> 894, 793, 624, 537, 663, 759, 539, 788, 1123, 8~
$ Working_age_population <dbl> 580, 470, 461, 307, 415, 453, 345, 406, 709, 52~
$ Income_rate            <dbl> 0.08, 0.05, 0.06, 0.10, 0.10, 0.04, 0.02, 0.02,~
$ Income_count           <dbl> 71, 43, 40, 52, 68, 30, 13, 14, 17, 5, 14, 24, ~
$ Employment_rate        <dbl> 0.08, 0.05, 0.04, 0.08, 0.08, 0.04, 0.02, 0.03,~
$ Employment_count       <dbl> 49, 25, 19, 26, 32, 17, 8, 13, 12, 7, 14, 24, 4~
$ CIF                    <dbl> 65, 45, 45, 80, 95, 50, 40, 40, 25, 25, 35, 40,~
$ ALCOHOL                <dbl> 28.728183, 129.921017, 71.021154, 80.473293, 89~
$ DRUG                   <dbl> 30.36573, 126.43368, 18.26983, 28.48559, 44.290~
$ SMR                    <dbl> 69.55405, 80.57479, 41.14113, 103.48468, 138.64~
$ DEPRESS                <dbl> 0.13154961, 0.14250310, 0.12812500, 0.16396396,~
```

```
$ LBWT                  <dbl> 0.00000000, 0.00000000, 0.03703704, 0.04761905,~
$ EMERG                 <dbl> 74.21743, 86.08168, 69.31582, 88.17561, 88.7019~
$ Attendance            <dbl> 0.85207, 0.84746, 0.90476, 0.94268, 0.79739, 0.~
$ Attainment            <dbl> 5.882353, 5.961538, 5.750000, 6.200000, 5.86666~
$ no_qualifications     <dbl> 52.758631, 95.854081, 38.559683, 80.060071, 77.~
$ not_participating     <dbl> 0.000000000, 0.017699115, 0.014925373, 0.000000~
$ University            <dbl> 0.297297, 0.117188, 0.185185, 0.250000, 0.16494~
$ drive_petrol          <dbl> 2.540103, 3.915072, 3.323025, 2.622991, 2.11500~
$ drive_GP              <dbl> 3.074295, 4.309812, 3.784549, 2.778026, 2.35833~
$ drive_post            <dbl> 1.616239, 2.555858, 1.440991, 2.620681, 2.40841~
$ drive_primary         <dbl> 2.615747, 3.646697, 3.247325, 1.936908, 1.84567~
$ drive_retail          <dbl> 1.544260, 2.849656, 2.062255, 2.160142, 1.78463~
$ drive_secondary       <dbl> 9.930833, 11.042816, 10.616768, 10.036471, 9.65~
$ PT_GP                 <dbl> 8.863589, 9.978272, 8.620700, 7.935112, 5.56896~
$ PT_post               <dbl> 5.856135, 7.515000, 4.321493, 8.433328, 6.96642~
$ PT_retail             <dbl> 6.023406, 7.926029, 5.770910, 8.329819, 6.63260~
$ Broadband             <dbl> 0.105050505, 0.013586957, 0.005633803, 0.113074~
$ crime_count           <dbl> 11.139188, 10.126535, 8.101228, 4.050614, 11.13~
$ crime_rate            <dbl> 124.59942, 127.69905, 129.82737, 75.43043, 168.~
$ overcrowded_count     <dbl> 87, 85, 31, 42, 50, 27, 27, 15, 10, 29, 12, 39,~
$ nocentralheat_count   <dbl> 10, 4, 8, 6, 7, 8, 9, 4, 3, 1, 1, 9, 0, 0, 0, 0~
$ overcrowded_rate      <dbl> 0.102112676, 0.101674641, 0.048211509, 0.072413~
$ nocentralheat_rate    <dbl> 0.011737089, 0.004784689, 0.012441680, 0.010344~
$ urban                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,~
$ intermediate_zone     <chr> "Culter", "Culter", "Culter", "Culter", "Culter~
$ council_area          <chr> "Aberdeen City", "Aberdeen City", "Aberdeen Cit~
$ health_board          <chr> "Grampian", "Grampian", "Grampian", "Grampian",~
```

```
merged1 <- merged %>%
  inner_join(covid, by=c("health_board"="health_board"))

glimpse(merged1)
```

```
Rows: 6,976
Columns: 43
$ Data_Zone             <chr> "S01006506", "S01006507", "S01006508", "S010065~
$ Intermediate_Zone     <chr> "Culter", "Culter", "Culter", "Culter", "Culter~
$ Council_area          <chr> "Aberdeen City", "Aberdeen City", "Aberdeen Cit~
$ Total_population      <dbl> 894, 793, 624, 537, 663, 759, 539, 788, 1123, 8~
$ Working_age_population <dbl> 580, 470, 461, 307, 415, 453, 345, 406, 709, 52~
$ Income_rate           <dbl> 0.08, 0.05, 0.06, 0.10, 0.10, 0.04, 0.02, 0.02,~
$ Income_count          <dbl> 71, 43, 40, 52, 68, 30, 13, 14, 17, 5, 14, 24, ~
$ Employment_rate       <dbl> 0.08, 0.05, 0.04, 0.08, 0.08, 0.04, 0.02, 0.03,~
$ Employment_count      <dbl> 49, 25, 19, 26, 32, 17, 8, 13, 12, 7, 14, 24, 4~
$ CIF                   <dbl> 65, 45, 45, 80, 95, 50, 40, 40, 25, 25, 35, 40,~
$ ALCOHOL               <dbl> 28.728183, 129.921017, 71.021154, 80.473293, 89~
$ DRUG                  <dbl> 30.36573, 126.43368, 18.26983, 28.48559, 44.290~
$ SMR                   <dbl> 69.55405, 80.57479, 41.14113, 103.48468, 138.64~
$ DEPRESS               <dbl> 0.13154961, 0.14250310, 0.12812500, 0.16396396,~
$ LBWT                  <dbl> 0.00000000, 0.00000000, 0.03703704, 0.04761905,~
$ EMERG                 <dbl> 74.21743, 86.08168, 69.31582, 88.17561, 88.7019~
$ Attendance            <dbl> 0.85207, 0.84746, 0.90476, 0.94268, 0.79739, 0.~
$ Attainment            <dbl> 5.882353, 5.961538, 5.750000, 6.200000, 5.86666~
$ no_qualifications     <dbl> 52.758631, 95.854081, 38.559683, 80.060071, 77.~
$ not_participating     <dbl> 0.000000000, 0.017699115, 0.014925373, 0.000000~
```

```
$ University             <dbl> 0.297297, 0.117188, 0.185185, 0.250000, 0.16494~
$ drive_petrol           <dbl> 2.540103, 3.915072, 3.323025, 2.622991, 2.11500~
$ drive_GP               <dbl> 3.074295, 4.309812, 3.784549, 2.778026, 2.35833~
$ drive_post             <dbl> 1.616239, 2.555858, 1.440991, 2.620681, 2.40841~
$ drive_primary          <dbl> 2.615747, 3.646697, 3.247325, 1.936908, 1.84567~
$ drive_retail           <dbl> 1.544260, 2.849656, 2.062255, 2.160142, 1.78463~
$ drive_secondary        <dbl> 9.930833, 11.042816, 10.616768, 10.036471, 9.65~
$ PT_GP                  <dbl> 8.863589, 9.978272, 8.620700, 7.935112, 5.56896~
$ PT_post                <dbl> 5.856135, 7.515000, 4.321493, 8.433328, 6.96642~
$ PT_retail              <dbl> 6.023406, 7.926029, 5.770910, 8.329819, 6.63260~
$ Broadband              <dbl> 0.105050505, 0.013586957, 0.005633803, 0.113074~
$ crime_count            <dbl> 11.139188, 10.126535, 8.101228, 4.050614, 11.13~
$ crime_rate             <dbl> 124.59942, 127.69905, 129.82737, 75.43043, 168.~
$ overcrowded_count      <dbl> 87, 85, 31, 42, 50, 27, 27, 15, 10, 29, 12, 39,~
$ nocentralheat_count    <dbl> 10, 4, 8, 6, 7, 8, 9, 4, 3, 1, 1, 9, 0, 0, 0, 0~
$ overcrowded_rate       <dbl> 0.102112676, 0.101674641, 0.048211509, 0.072413~
$ nocentralheat_rate     <dbl> 0.011737089, 0.004784689, 0.012441680, 0.010344~
$ urban                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,~
$ intermediate_zone      <chr> "Culter", "Culter", "Culter", "Culter", "Culter~
$ council_area           <chr> "Aberdeen City", "Aberdeen City", "Aberdeen Cit~
$ health_board           <chr> "Grampian", "Grampian", "Grampian", "Grampian",~
$ `2020`                 <dbl> 7870, 7870, 7870, 7870, 7870, 7870, 7870, 7870,~
$ `2021`                 <dbl> 39166, 39166, 39166, 39166, 39166, 39166, 39166~
```

There are two duplicate variables we should remove - `intermediate_zone` and `council_area`. We can remove them using the `select()` and `-c()` functions.

```
merged2 <- merged1 %>%
  select(-c(intermediate_zone,council_area))

glimpse(merged2)
```

```
Rows: 6,976
Columns: 41
$ Data_Zone              <chr> "S01006506", "S01006507", "S01006508", "S010065~
$ Intermediate_Zone      <chr> "Culter", "Culter", "Culter", "Culter", "Culter~
$ Council_area           <chr> "Aberdeen City", "Aberdeen City", "Aberdeen Cit~
$ Total_population       <dbl> 894, 793, 624, 537, 663, 759, 539, 788, 1123, 8~
$ Working_age_population <dbl> 580, 470, 461, 307, 415, 453, 345, 406, 709, 52~
$ Income_rate            <dbl> 0.08, 0.05, 0.06, 0.10, 0.10, 0.04, 0.02, 0.02,~
$ Income_count           <dbl> 71, 43, 40, 52, 68, 30, 13, 14, 17, 5, 14, 24, ~
$ Employment_rate        <dbl> 0.08, 0.05, 0.04, 0.08, 0.08, 0.04, 0.02, 0.03,~
$ Employment_count       <dbl> 49, 25, 19, 26, 32, 17, 8, 13, 12, 7, 14, 24, 4~
$ CIF                    <dbl> 65, 45, 45, 80, 95, 50, 40, 40, 25, 25, 35, 40,~
$ ALCOHOL                <dbl> 28.728183, 129.921017, 71.021154, 80.473293, 89~
$ DRUG                   <dbl> 30.36573, 126.43368, 18.26983, 28.48559, 44.290~
$ SMR                    <dbl> 69.55405, 80.57479, 41.14113, 103.48468, 138.64~
$ DEPRESS                <dbl> 0.13154961, 0.14250310, 0.12812500, 0.16396396,~
$ LBWT                   <dbl> 0.00000000, 0.00000000, 0.03703704, 0.04761905,~
$ EMERG                  <dbl> 74.21743, 86.08168, 69.31582, 88.17561, 88.7019~
$ Attendance             <dbl> 0.85207, 0.84746, 0.90476, 0.94268, 0.79739, 0.~
$ Attainment             <dbl> 5.882353, 5.961538, 5.750000, 6.200000, 5.86666~
$ no_qualifications      <dbl> 52.758631, 95.854081, 38.559683, 80.060071, 77.~
$ not_participating      <dbl> 0.000000000, 0.017699115, 0.014925373, 0.000000~
$ University             <dbl> 0.297297, 0.117188, 0.185185, 0.250000, 0.16494~
```

```
$ drive_petrol           <dbl> 2.540103, 3.915072, 3.323025, 2.622991, 2.11500~
$ drive_GP               <dbl> 3.074295, 4.309812, 3.784549, 2.778026, 2.35833~
$ drive_post             <dbl> 1.616239, 2.555858, 1.440991, 2.620681, 2.40841~
$ drive_primary          <dbl> 2.615747, 3.646697, 3.247325, 1.936908, 1.84567~
$ drive_retail           <dbl> 1.544260, 2.849656, 2.062255, 2.160142, 1.78463~
$ drive_secondary        <dbl> 9.930833, 11.042816, 10.616768, 10.036471, 9.65~
$ PT_GP                  <dbl> 8.863589, 9.978272, 8.620700, 7.935112, 5.56896~
$ PT_post                <dbl> 5.856135, 7.515000, 4.321493, 8.433328, 6.96642~
$ PT_retail              <dbl> 6.023406, 7.926029, 5.770910, 8.329819, 6.63260~
$ Broadband              <dbl> 0.105050505, 0.013586957, 0.005633803, 0.113074~
$ crime_count            <dbl> 11.139188, 10.126535, 8.101228, 4.050614, 11.13~
$ crime_rate             <dbl> 124.59942, 127.69905, 129.82737, 75.43043, 168.~
$ overcrowded_count      <dbl> 87, 85, 31, 42, 50, 27, 27, 15, 10, 29, 12, 39,~
$ nocentralheat_count    <dbl> 10, 4, 8, 6, 7, 8, 9, 4, 3, 1, 1, 9, 0, 0, 0, 0~
$ overcrowded_rate       <dbl> 0.102112676, 0.101674641, 0.048211509, 0.072413~
$ nocentralheat_rate     <dbl> 0.011737089, 0.004784689, 0.012441680, 0.010344~
$ urban                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,~
$ health_board           <chr> "Grampian", "Grampian", "Grampian", "Grampian",~
$ `2020`                 <dbl> 7870, 7870, 7870, 7870, 7870, 7870, 7870, 7870,~
$ `2021`                 <dbl> 39166, 39166, 39166, 39166, 39166, 39166, 39166~
```

We could also do all of the above in one step.

```
merged3 <- simd %>%
  inner_join(healthboard, by=c("Data_Zone"="data_zone")) %>%
  inner_join(covid, by=c("health_board"="health_board")) %>%
  select(-c(intermediate_zone,council_area))

glimpse(merged3)
```

```
Rows: 6,976
Columns: 41
$ Data_Zone              <chr> "S01006506", "S01006507", "S01006508", "S010065~
$ Intermediate_Zone      <chr> "Culter", "Culter", "Culter", "Culter", "Culter~
$ Council_area           <chr> "Aberdeen City", "Aberdeen City", "Aberdeen Cit~
$ Total_population       <dbl> 894, 793, 624, 537, 663, 759, 539, 788, 1123, 8~
$ Working_age_population <dbl> 580, 470, 461, 307, 415, 453, 345, 406, 709, 52~
$ Income_rate            <dbl> 0.08, 0.05, 0.06, 0.10, 0.10, 0.04, 0.02, 0.02,~
$ Income_count           <dbl> 71, 43, 40, 52, 68, 30, 13, 14, 17, 5, 14, 24, ~
$ Employment_rate        <dbl> 0.08, 0.05, 0.04, 0.08, 0.08, 0.04, 0.02, 0.03,~
$ Employment_count       <dbl> 49, 25, 19, 26, 32, 17, 8, 13, 12, 7, 14, 24, 4~
$ CIF                    <dbl> 65, 45, 45, 80, 95, 50, 40, 40, 25, 25, 35, 40,~
$ ALCOHOL                <dbl> 28.728183, 129.921017, 71.021154, 80.473293, 89~
$ DRUG                   <dbl> 30.36573, 126.43368, 18.26983, 28.48559, 44.290~
$ SMR                    <dbl> 69.55405, 80.57479, 41.14113, 103.48468, 138.64~
$ DEPRESS                <dbl> 0.13154961, 0.14250310, 0.12812500, 0.16396396,~
$ LBWT                   <dbl> 0.00000000, 0.00000000, 0.03703704, 0.04761905,~
$ EMERG                  <dbl> 74.21743, 86.08168, 69.31582, 88.17561, 88.7019~
$ Attendance             <dbl> 0.85207, 0.84746, 0.90476, 0.94268, 0.79739, 0.~
$ Attainment             <dbl> 5.882353, 5.961538, 5.750000, 6.200000, 5.86666~
$ no_qualifications      <dbl> 52.758631, 95.854081, 38.559683, 80.060071, 77.~
$ not_participating      <dbl> 0.000000000, 0.017699115, 0.014925373, 0.000000~
$ University             <dbl> 0.297297, 0.117188, 0.185185, 0.250000, 0.16494~
$ drive_petrol           <dbl> 2.540103, 3.915072, 3.323025, 2.622991, 2.11500~
```

```
$ drive_GP              <dbl> 3.074295, 4.309812, 3.784549, 2.778026, 2.35833~
$ drive_post            <dbl> 1.616239, 2.555858, 1.440991, 2.620681, 2.40841~
$ drive_primary         <dbl> 2.615747, 3.646697, 3.247325, 1.936908, 1.84567~
$ drive_retail          <dbl> 1.544260, 2.849656, 2.062255, 2.160142, 1.78463~
$ drive_secondary       <dbl> 9.930833, 11.042816, 10.616768, 10.036471, 9.65~
$ PT_GP                 <dbl> 8.863589, 9.978272, 8.620700, 7.935112, 5.56896~
$ PT_post               <dbl> 5.856135, 7.515000, 4.321493, 8.433328, 6.96642~
$ PT_retail             <dbl> 6.023406, 7.926029, 5.770910, 8.329819, 6.63260~
$ Broadband             <dbl> 0.105050505, 0.013586957, 0.005633803, 0.113074~
$ crime_count           <dbl> 11.139188, 10.126535, 8.101228, 4.050614, 11.13~
$ crime_rate            <dbl> 124.59942, 127.69905, 129.82737, 75.43043, 168.~
$ overcrowded_count     <dbl> 87, 85, 31, 42, 50, 27, 27, 15, 10, 29, 12, 39,~
$ nocentralheat_count   <dbl> 10, 4, 8, 6, 7, 8, 9, 4, 3, 1, 1, 9, 0, 0, 0, 0~
$ overcrowded_rate      <dbl> 0.102112676, 0.101674641, 0.048211509, 0.072413~
$ nocentralheat_rate    <dbl> 0.011737089, 0.004784689, 0.012441680, 0.010344~
$ urban                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,~
$ health_board          <chr> "Grampian", "Grampian", "Grampian", "Grampian",~
$ `2020`                <dbl> 7870, 7870, 7870, 7870, 7870, 7870, 7870, 7870,~
$ `2021`                <dbl> 39166, 39166, 39166, 39166, 39166, 39166, 39166~
```