Week 4 Exercises

Rebecca Hawthorne

July 17, 2023

Please complete all exercises below. You may use any library that we have covered in class. The data we will be using comes from the tidyr package, so you must use that.

1) Examine the who and population data sets that come with the tidyr library. the who data is not tidy, you will need to reshape the new_sp_m014 to newrel_f65 columns to long format retaining country, iso2, iso3, and year. The data in the columns you are reshaping contains patterns described in the details section below. You will need to assign three columns: diagnosis, gender, and age to the patterns described in the details.

Your tidy data should look like the following: country iso2 iso3 year diagnosis gender age count 1 Afghanistan AF AFG 1980 sp m 014 NA 2 Afghanistan AF AFG 1980 sp m 1524 NA 3 Afghanistan AF AFG 1980 sp m 2534 NA 4 Afghanistan AF AFG 1980 sp m 3544 NA 5 Afghanistan AF AFG 1980 sp m 4554 NA 6 Afghanistan AF AFG 1980 sp m 5564 NA

Details The data uses the original codes given by the World Health Organization. The column names for columns five through 60 are made by combining new_ to a code for method of diagnosis (rel = relapse, sn = negative pulmonary smear, sp = positive pulmonary smear, ep = extrapulmonary) to a code for gender (f = female, m = male) to a code for age group (014 = 0-14 yrs of age, 1524 = 15-24 years of age, 2534 = 25 to 34 years of age, 3544 = 35 to 44 years of age, 4554 = 45 to 54 years of age, 5564 = 55 to 64 years of age, 65 = 65 years of age or older).

Note: use data(who) and data(population) to load the data into your environment. Use the arguments cols, names_to, names_pattern, and values_to. Your regex should be = ("new_?(.)_(.)(.)")

https://tidyr.tidyverse.org/reference/who.html

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)

data(who)
data(population)

who_long <- who %>%
  pivot_longer(cols = 5:60,
```

```
               names_to = c('diagnosis', 'gender', 'age'),
               names_pattern = (regex= ("new_?(.*)_(.)(.*)")),
               values_to = "count")
head(who_long)
```

```
## # A tibble: 6 x 8
##   country     iso2  iso3   year diagnosis gender age   count
##   <chr>       <chr> <chr> <dbl> <chr>     <chr>  <chr> <dbl>
## 1 Afghanistan AF    AFG    1980 sp        m      014      NA
## 2 Afghanistan AF    AFG    1980 sp        m      1524     NA
## 3 Afghanistan AF    AFG    1980 sp        m      2534     NA
## 4 Afghanistan AF    AFG    1980 sp        m      3544     NA
## 5 Afghanistan AF    AFG    1980 sp        m      4554     NA
## 6 Afghanistan AF    AFG    1980 sp        m      5564     NA
```

2) There are two common keys between the data sets, with who as the left table, join the population data by country and year so that the population is available within the who dataset.

```
#df with left join to the original who data set

who_left_df <- who %>%
  left_join(population, by=c('country', 'year'))

head(who_left_df)
```

```
## # A tibble: 6 x 61
##   country   iso2  iso3   year new_sp_m014 new_sp_m1524 new_sp_m2534 new_sp_m3544
##   <chr>     <chr> <chr> <dbl>       <dbl>        <dbl>        <dbl>        <dbl>
## 1 Afghanis~ AF    AFG    1980          NA           NA           NA           NA
## 2 Afghanis~ AF    AFG    1981          NA           NA           NA           NA
## 3 Afghanis~ AF    AFG    1982          NA           NA           NA           NA
## 4 Afghanis~ AF    AFG    1983          NA           NA           NA           NA
## 5 Afghanis~ AF    AFG    1984          NA           NA           NA           NA
## 6 Afghanis~ AF    AFG    1985          NA           NA           NA           NA
## # i 53 more variables: new_sp_m4554 <dbl>, new_sp_m5564 <dbl>,
## #   new_sp_m65 <dbl>, new_sp_f014 <dbl>, new_sp_f1524 <dbl>,
## #   new_sp_f2534 <dbl>, new_sp_f3544 <dbl>, new_sp_f4554 <dbl>,
## #   new_sp_f5564 <dbl>, new_sp_f65 <dbl>, new_sn_m014 <dbl>,
## #   new_sn_m1524 <dbl>, new_sn_m2534 <dbl>, new_sn_m3544 <dbl>,
## #   new_sn_m4554 <dbl>, new_sn_m5564 <dbl>, new_sn_m65 <dbl>,
## #   new_sn_f014 <dbl>, new_sn_f1524 <dbl>, new_sn_f2534 <dbl>, ...
```

```
#df with left join to the long data set

who_long_left_df <- who_long %>%
  left_join(population, by=c('country', 'year'))

head(who_long_left_df)
```

```
## # A tibble: 6 x 9
##   country     iso2  iso3   year diagnosis gender age   count population
##   <chr>       <chr> <chr> <dbl> <chr>     <chr>  <chr> <dbl>      <dbl>
## 1 Afghanistan AF    AFG    1980 sp        m      014      NA         NA
## 2 Afghanistan AF    AFG    1980 sp        m      1524     NA         NA
## 3 Afghanistan AF    AFG    1980 sp        m      2534     NA         NA
## 4 Afghanistan AF    AFG    1980 sp        m      3544     NA         NA
```

```
## 5 Afghanistan AF    AFG    1980 sp        m      4554     NA         NA
## 6 Afghanistan AF    AFG    1980 sp        m      5564     NA         NA
```

3) Split the age column into two columns, min age and max age. Notice that there is no character separator. Check the documentation with ?separate to understand other ways to separate the age column. Keep in mind that 0 to 14 is coded as 014 (3 characters) and the other age groups are coded with 4 characters. 65 only has two characters, but we will ignore that until the next problem.

```
who_long_lef_age_split_df <- who_long_left_df %>%
  separate('age', c('min_age', 'max_age'), sep = -2)

head(who_long_lef_age_split_df)
```

```
## # A tibble: 6 x 10
##   country    iso2  iso3   year diagnosis gender min_age max_age count population
##   <chr>      <chr> <chr> <dbl> <chr>     <chr>  <chr>   <chr>   <dbl>      <dbl>
## 1 Afghanist~ AF    AFG    1980 sp        m      0       14         NA         NA
## 2 Afghanist~ AF    AFG    1980 sp        m      15      24         NA         NA
## 3 Afghanist~ AF    AFG    1980 sp        m      25      34         NA         NA
## 4 Afghanist~ AF    AFG    1980 sp        m      35      44         NA         NA
## 5 Afghanist~ AF    AFG    1980 sp        m      45      54         NA         NA
## 6 Afghanist~ AF    AFG    1980 sp        m      55      64         NA         NA
```

4) Since we ignored the 65+ group in the previous problem we will fix it here. If you examine the data you will notice that 65 was placed into the max_age column and there is no value for min_age for those records. To fix this use mutate() in order to replace the blank value in the min_age column with the value from the max_age column and another mutate to replace the 65 in the max column with an Inf. Be sure to keep the variables as character vectors.

```
who_long_lef_age_split_fixed_df <- who_long_lef_age_split_df %>%
 mutate(min_age = replace(min_age, min_age == '', '65')) %>%
  mutate(max_age = replace(max_age, min_age == '65', 'Inf'))
```

5) Find the count per diagnosis for males and females.

See ?sum for a hint on resolving NA values.

```
sum_per_sex_of_cases_diagnosis_df <- who_long_lef_age_split_fixed_df %>%
 group_by(diagnosis, gender) %>%
 summarize(sum_per_sex_of_cases_diagnosis = sum(count, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'diagnosis'. You can override using the
## `.groups` argument.
```

```
head(sum_per_sex_of_cases_diagnosis_df )
```

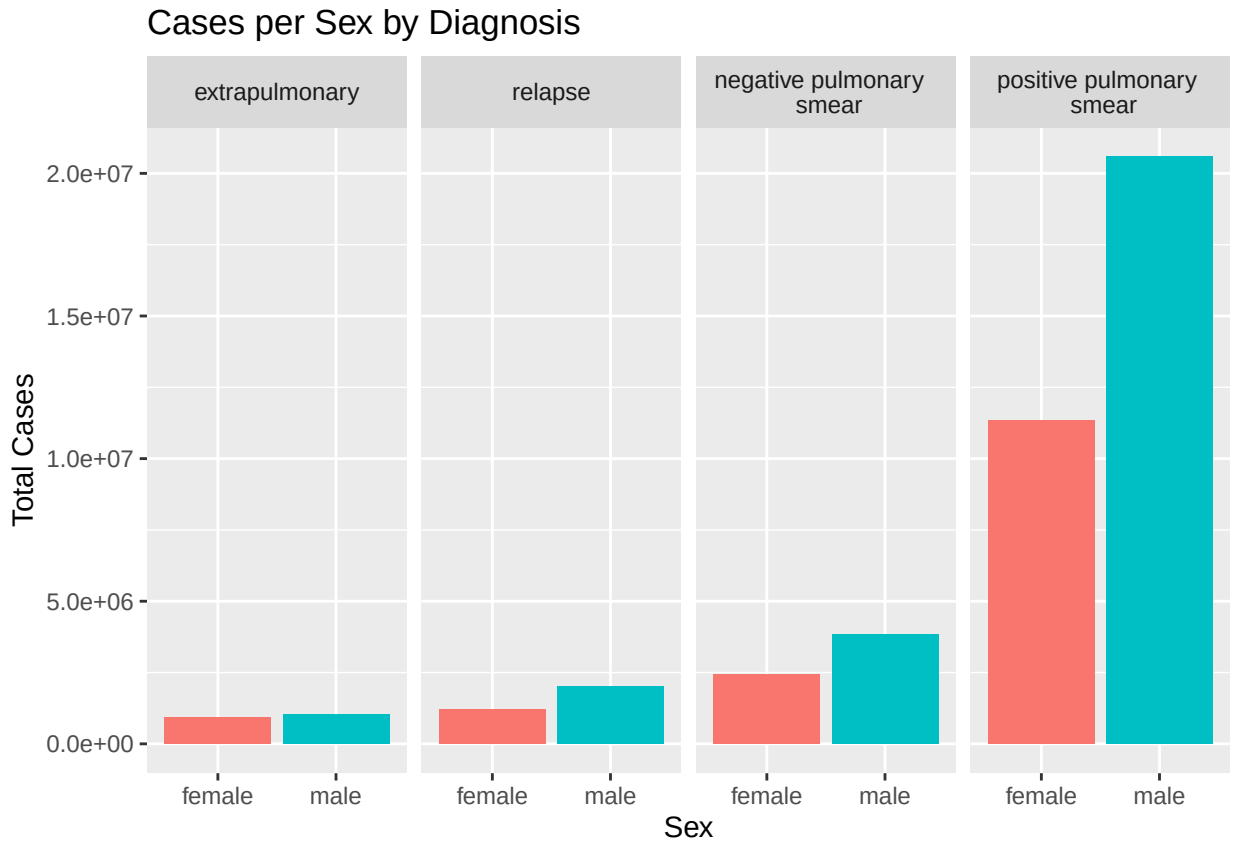
```
## # A tibble: 6 x 3
## # Groups:   diagnosis [3]
##   diagnosis gender sum_per_sex_of_cases_diagnosis
##   <chr>     <chr>                           <dbl>
## 1 ep        f                              941880
## 2 ep        m                             1044299
## 3 rel       f                             1201596
## 4 rel       m                             2018976
## 5 sn        f                             2439139
## 6 sn        m                             3840388
```

6) Now create a plot using ggplot and geom_col where your x axis is gender, your y axis represents the counts, and facet by diagnosis. Be sure to give your plot a title and resolve the axis labels.

```r
#create new labels for the facets
diagnosis.labs <- c("relapse", "negative pulmonary  \n\ smear",
                    "positive pulmonary \n\ smear", "extrapulmonary")
names(diagnosis.labs) <- c("rel", "sn", "sp", "ep")



ggplot(sum_per_sex_of_cases_diagnosis_df) +
  geom_col(aes(x=gender,y=sum_per_sex_of_cases_diagnosis, fill = gender),
           show.legend = FALSE) +
  facet_grid(.~diagnosis, labeller = labeller(diagnosis = diagnosis.labs)) +
  scale_x_discrete(labels=c("f" = "female", "m" = "male"))+
  labs(x='Sex',
       y='Total Cases',
       title='Cases per Sex by Diagnosis')
```



7) Find the percentage of population by year, gender, and diagnosis. Be sure to remove rows containing NA values.

```r
percentage_of_pop <- who_long_lef_age_split_fixed_df %>%
 drop_na() %>%
 group_by(year,diagnosis, gender) %>%
 summarize(percentage = (100*(sum(count))/(sum(population))))
```

```
## `summarise()` has grouped output by 'year', 'diagnosis'. You can override using
## the `.groups` argument.
```

```
head(percentage_of_pop)
```

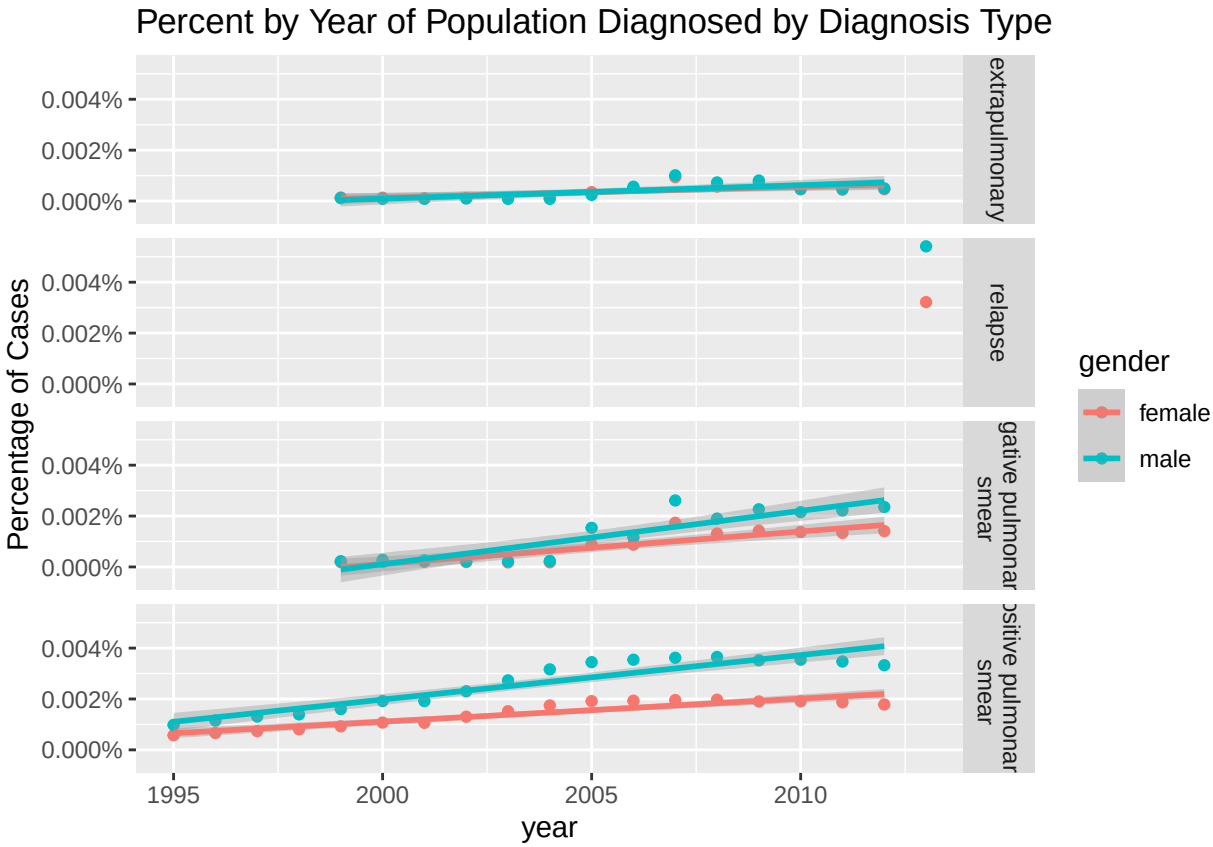
```
## # A tibble: 6 x 4
## # Groups:   year, diagnosis [3]
##    year diagnosis gender percentage
##   <dbl> <chr>     <chr>       <dbl>
## 1  1995 sp        f        0.000574
## 2  1995 sp        m        0.000982
## 3  1996 sp        f        0.000663
## 4  1996 sp        m        0.00115
## 5  1997 sp        f        0.000737
## 6  1997 sp        m        0.00131
```

8) Create a line plot in ggplot where your x axis contains the year and y axis contains the percent of world population. Facet this plot by diagnosis with each plot stacked vertically. You should have a line for each gender within each facet. Be sure to format your y axis and give your plot a title.

```
diagnosis.labs <- c("relapse", "negative pulmonary  \n\ smear",
                    "positive pulmonary \n\ smear", "extrapulmonary")
names(diagnosis.labs) <- c("rel", "sn", "sp", "ep")

ggplot(percentage_of_pop) +
  geom_point(aes(x=year,y=(percentage/100), color=gender)) +
  scale_color_discrete(labels=c("f" = "female", "m" = "male")) +
  geom_smooth(method='lm', aes(y=(percentage/100),x=year, color = gender)) +
  facet_grid(rows = vars(diagnosis), labeller = labeller(diagnosis = diagnosis.labs)) +
  scale_y_continuous(labels=scales::percent_format())+
  labs(title='Percent by Year of Population Diagnosed by Diagnosis Type',
       y='Percentage of Cases',)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



9) Now unite the min and max age variables into a new variable named age_range. Use a '-' as the separator.

```
who_long_left_age_together_df <- who_long_lef_age_split_fixed_df %>%
  unite(col = 'age_range', min_age:max_age,sep='-')

head(who_long_left_age_together_df)
```

```
## # A tibble: 6 x 9
##   country     iso2  iso3   year diagnosis gender age_range count population
##   <chr>       <chr> <chr> <dbl> <chr>     <chr>  <chr>     <dbl>      <dbl>
## 1 Afghanistan AF    AFG    1980 sp        m      0-14         NA         NA
## 2 Afghanistan AF    AFG    1980 sp        m      15-24        NA         NA
## 3 Afghanistan AF    AFG    1980 sp        m      25-34        NA         NA
## 4 Afghanistan AF    AFG    1980 sp        m      35-44        NA         NA
## 5 Afghanistan AF    AFG    1980 sp        m      45-54        NA         NA
## 6 Afghanistan AF    AFG    1980 sp        m      55-64        NA         NA
```

10) Find the percentage contribution of each age group by diagnosis. You will first need to find the count of all diagnoses then find the count of all diagnoses by age group. Join the former to the later and calculate the percent of each age group. Plot these as a geom_col where the x axis is the diagnosis, y axis is the percent of total, and faceted by age group.

```
count_diagnosis_df <- who_long_left_age_together_df %>%
 group_by(diagnosis) %>%
 summarize(total_count = sum(count, na.rm = TRUE))

head(count_diagnosis_df )
```

```
## # A tibble: 4 x 2
##   diagnosis total_count
##   <chr>           <dbl>
## 1 ep            1986179
## 2 rel           3220572
## 3 sn            6279527
## 4 sp           31911240
```

```
count_diagnosis_by_age_df <- who_long_left_age_together_df %>%
 group_by(diagnosis, age_range) %>%
 summarize(count_by_age = sum(count, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'diagnosis'. You can override using the
## `.groups` argument.
```

```
head(count_diagnosis_by_age_df )
```

```
## # A tibble: 6 x 3
## # Groups:   diagnosis [1]
##   diagnosis age_range count_by_age
##   <chr>     <chr>            <dbl>
## 1 ep        0-14            249998
## 2 ep        15-24           314716
## 3 ep        25-34           398758
## 4 ep        35-44           526041
## 5 ep        45-54           205633
## 6 ep        55-64           137356
```

```
count_diagnosis_join_df <- count_diagnosis_by_age_df %>%
  left_join(count_diagnosis_df , by=c('diagnosis'))


head(count_diagnosis_join_df)
```

```
## # A tibble: 6 x 4
## # Groups:   diagnosis [1]
##   diagnosis age_range count_by_age total_count
##   <chr>     <chr>            <dbl>       <dbl>
## 1 ep        0-14            249998     1986179
## 2 ep        15-24           314716     1986179
## 3 ep        25-34           398758     1986179
## 4 ep        35-44           526041     1986179
## 5 ep        45-54           205633     1986179
## 6 ep        55-64           137356     1986179
```

```
count_diagnosis_join_percentage_df <- count_diagnosis_join_df %>%
  group_by(diagnosis, age_range) %>%
  summarize(percentage_of_cases=(100*count_by_age/total_count) )
```

```
## `summarise()` has grouped output by 'diagnosis'. You can override using the
## `.groups` argument.
```

```
head(count_diagnosis_join_percentage_df)
```

```
## # A tibble: 6 x 3
## # Groups:   diagnosis [1]
##   diagnosis age_range percentage_of_cases
##   <chr>     <chr>                  <dbl>
```

```
## 1 ep        0-14                  12.6
## 2 ep        15-24                 15.8
## 3 ep        25-34                 20.1
## 4 ep        35-44                 26.5
## 5 ep        45-54                 10.4
## 6 ep        55-64                  6.92
```

```r
ggplot(count_diagnosis_join_percentage_df) +
  geom_col(aes(x=diagnosis,y=(percentage_of_cases/100), fill = diagnosis)) +
  facet_grid(.~age_range) +
  scale_fill_discrete (name = "Diagnosis type", labels = diagnosis.labs) +
  scale_y_continuous(labels=scales::percent_format())+
  theme(axis.text.x = element_text(angle = 45)) +
  labs(x='Diagnosis',
       y='Percentage of Cases',
       title='Percentage of each Diagnosis Type by Age')
```