

Homework 8

Andrew Shao (NetID: as13381)

You are required to process the data in the files `ProjectTycho_Level1_v1.0.0.csv` and `us_state_populations_ext.rds` via the 5 sequential steps given in the questions below.

Question 1 (1 pt): Load the data from the file `ProjectTycho_Level1_v1.0.0.csv` and remove duplicate rows. Name the resulting data frame as `ProjectTycho_Level1`. Output the dimension of the data frame.

```
ProjectTycho_Level1 <- distinct(read_csv('ProjectTycho_Level1_v1.0.0.csv', show_col_types
↪   = F))
dim(ProjectTycho_Level1)
```

Answer:

```
## [1] 759465      7
```

Question 2 (1 pt): For data frame `ProjectTycho_Level1`, drop its rows that have `disease = "DIPHTHERIA"`. After that, output the distinct values of `ProjectTycho_Level1$disease`.

```
ProjectTycho_Level1 <- filter(ProjectTycho_Level1, disease != 'DIPHTHERIA')
unique(pull(ProjectTycho_Level1, disease))
```

Answer:

```
## [1] "HEPATITIS A" "MEASLES"      "MUMPS"          "PERTUSSIS"      "POLIO"
## [6] "RUBELLA"     "SMALLPOX"
```

Question 3 (1 pt): Separate the column `epi_week` of `ProjectTycho_Level1` into two new columns named as `year` and `week` which are in the integer type. After that, provide the output of `head(ProjectTycho_Level1)` and `dim(ProjectTycho_Level1)`.

```
ProjectTycho_Level1 <- ProjectTycho_Level1 %>%
  mutate(year = epi_week %/% 100,
         week = epi_week %% 100) %>%
  select(-c('epi_week'))
head(ProjectTycho_Level1)
```

Answer:

```
## # A tibble: 6 x 8
##   state loc      loc_type disease    cases incidence_per_100000  year  week
##   <chr> <chr>    <chr>    <chr>    <dbl>          <dbl> <dbl> <dbl>
## 1 MN    MINNESOTA STATE    HEPATITIS A      3          0.08  1966    1
## 2 CO    COLORADO  STATE    HEPATITIS A      1          0.05  1966    1
## 3 AZ    ARIZONA   STATE    HEPATITIS A      6          0.37  1966    1
## 4 MT    MONTANA   STATE    HEPATITIS A      2          0.28  1966    1
## 5 LA    LOUISIANA STATE    HEPATITIS A      1          0.03  1966    1
## 6 WA    WASHINGTON STATE    HEPATITIS A      5          0.16  1966    1
```

```
dim(ProjectTycho_Level1)
```

```
## [1] 600482      8
```

Question 4 (1 pt): From ProjectTycho_Level1, create a new data frame, named as ProjectTycho_count, that contains the count of cases of each disease for each state at each year, with column names disease, state, year, weeks_reporting and count. Note that you first need to drop the rows with cases = NA. Use ungroup() if group_by() is used in your processing. You may see the data frame us_contagious_diseases of package dslabs as an example for the resulting data frame. Provide the output of head(ProjectTycho_count) and dim(ProjectTycho_count).

```
ProjectTycho_count <- ProjectTycho_Level1 %>%
  drop_na(cases) %>%
  group_by(disease, loc, year) %>%
  summarise(weeks_reporting = n(),
            count = sum(cases)) %>%
  ungroup() %>%
  rename(state = loc)
```

Answer:

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

```
head(ProjectTycho_count)
```

```
## # A tibble: 6 x 5
##   disease    state    year weeks_reporting count
##   <chr>    <chr>    <dbl>          <int> <dbl>
## 1 HEPATITIS A ALABAMA  1966          50  321
## 2 HEPATITIS A ALABAMA  1967          49  291
## 3 HEPATITIS A ALABAMA  1968          52  314
## 4 HEPATITIS A ALABAMA  1969          49  380
## 5 HEPATITIS A ALABAMA  1970          51  413
## 6 HEPATITIS A ALABAMA  1971          51  378
```

```
dim(ProjectTycho_count)
```

```
## [1] 14265      5
```

Question 5 (1 pt): Load the data from the file `us_state_populations.rds`. Add the population information as a column to the data frame `ProjectTycho_count`. Note that the function `str_to_upper()` may be useful here. After that, provide the output of `head(ProjectTycho_count)` and `dim(ProjectTycho_count)`.

```
us_state_populations <- readRDS('us_state_populations.rds') %>% mutate(state =  
  ↪ str_to_upper(state))
```

```
ProjectTycho_count <- ProjectTycho_count %>%  
  left_join(us_state_populations, by = c('state', 'year'))  
head(ProjectTycho_count)
```

Answer:

```
## # A tibble: 6 x 6  
##   disease      state  year weeks_reporting count population  
##   <chr>      <chr>  <dbl>          <int> <dbl>      <dbl>  
## 1 HEPATITIS A ALABAMA 1966           50    321    3345787  
## 2 HEPATITIS A ALABAMA 1967           49    291    3364130  
## 3 HEPATITIS A ALABAMA 1968           52    314    3386068  
## 4 HEPATITIS A ALABAMA 1969           49    380    3412450  
## 5 HEPATITIS A ALABAMA 1970           51    413    3444165  
## 6 HEPATITIS A ALABAMA 1971           51    378    3481798
```

```
dim(ProjectTycho_count)
```

```
## [1] 14265      6
```