Final_Exam

Keith Lee

2022-10-26

```
library(dslabs)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.6      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1.

(a). How many observations and variable are there? What are the types of each variable?

```
full_data <- us_contagious_diseases
dim(full_data) #to find number of variables and observations
```

```
## [1] 16065     6
```

```
str(full_data) #str() shows the types of each variable
```

```
## 'data.frame':    16065 obs. of  6 variables:
##  $ disease        : Factor w/ 7 levels "Hepatitis A",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ state          : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year           : num  1966 1967 1968 1969 1970 ...
##  $ weeks_reporting: num  50 49 52 49 51 51 45 45 45 46 ...
##  $ count          : num  321 291 314 380 413 378 342 467 244 286 ...
##  $ population     : num  3345787 3364130 3386068 3412450 3444165 ...
```

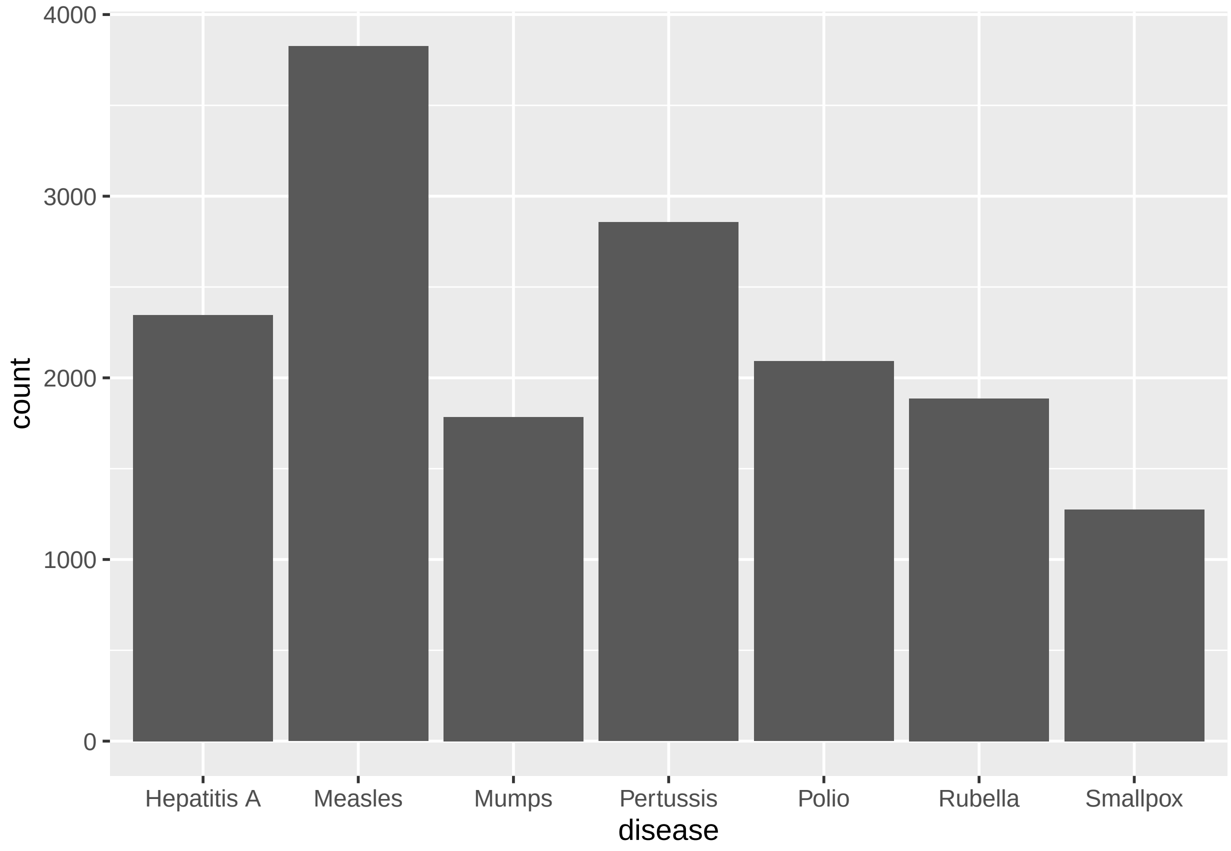
Disease and state are factor variables which is also another term for categorical variables. Weeks_report, year, count, and population are numeric variables. However, year can also be considered as categorical variable.

(b). Compute the frequency of each type of disease and visualize the proportion using barchart.

```
table(full_data$disease) #table() shows how many observations exist in each variable
```

```
##
## Hepatitis A     Measles       Mumps   Pertussis       Polio     Rubella
##        2346        3825        1785        2856        2091        1887
##    Smallpox
##        1275
```

```
ggplot(data=full_data) + geom_bar(mapping=aes(x=disease)) #used bar chart to easily observe the overall
```



(c). Compute the 0.1, 0.5, 0.9 quantiles of the population for each type of diseases. Write a paragraph to compare the quantiles of the population across the disease.

```
full_data %>%
  group_by(disease) %>%
  summarize(quantile_0.1 = quantile(count, probs=0.1, na.rm=TRUE))
```

```
## # A tibble: 7 x 2
##   disease     quantile_0.1
##   <fct>              <dbl>
## 1 Hepatitis A           10
## 2 Measles                0
## 3 Mumps                  0
## 4 Pertussis              4
## 5 Polio                   0
## 6 Rubella                 0
## 7 Smallpox                0
```

```
full_data %>%
  group_by(disease) %>%
  summarize(quantile_0.5 = quantile(count, probs=0.5, na.rm=TRUE))
```

```
## # A tibble: 7 x 2
##   disease     quantile_0.5
##   <fct>              <dbl>
## 1 Hepatitis A         138.
## 2 Measles             577
## 3 Mumps                29
## 4 Pertussis            81
## 5 Polio                57
## 6 Rubella               5
## 7 Smallpox              8
```

```
full_data %>%
  group_by(disease) %>%
  summarize(quantile_0.9 = quantile(count, probs=0.9, na.rm=TRUE))
```

```
## # A tibble: 7 x 2
##   disease     quantile_0.9
##   <fct>              <dbl>
## 1 Hepatitis A        1014.
## 2 Measles           13697
## 3 Mumps              1163.
## 4 Pertussis          1953
## 5 Polio               634
## 6 Rubella             641.
## 7 Smallpox            538.
```

For 10% of counts, we had proportion of 10 cases of Hepatitis A, 4 cases of Pertussis, and the rest 0. For 90% of counts, we had proportion of 1013.5 cases of Hepatitis A, 13697 cases for Measles, 1163 cases for Mumps, 1953 cases for Pertussis, 634 cases for Polio, 631.4 cases for Rubella, and 537.6 cases for Smallpox. For median values, we have 138.5 for Hepatitis A, 577 for Measles, 29 for Mumps, 81 for Pertussis, 57 for Polio, 5 for Rubella, and 8 for Smallpox.

2.

Find the top 5 states with the most "Mumps" cases over the 10 years from 1991 to 2000 (both years inclusive). Find the bottom 5 states with the least Hepatitis A cases over the 5 years from 1994 to 1998 (both years inclusive). Most

```
full_data %>% #used pipe operator for chaining commands
  filter(year >= 1991, year <= 2000, disease == "Mumps") %>% #filer() for getting the range of year and
  arrange(desc(count)) %>% #arrange() gives in order from lowest to highest, but also used desc() to ha
  head(5) #used head() function to get only first 5 observations
```

```
##   disease          state year weeks_reporting count population
## 1   Mumps     California 1991              49   389   30311890
## 2   Mumps South Carolina 1991              44   384    3527239
## 3   Mumps          Texas 1994              46   378   18376501
## 4   Mumps        Florida 1991              48   359   13246692
## 5   Mumps          Texas 1991              48   340   17305041
```

Least

```
full_data %>% #used pipe operator for chaining commands
  filter(year >= 1994, year <= 1998, disease == "Hepatitis A") %>% #filer() for getting the range of ye
  arrange((count)) %>% #arrange() gives in order from lowest to highest
  head(5) #used head() function to get only first 5 observations
```

```
##       disease        state year weeks_reporting count population
## 1 Hepatitis A  Mississippi 1998              16     0    2787267
## 2 Hepatitis A North Dakota 1998              31     3     638665
## 3 Hepatitis A      Vermont 1995              26     3     587845
## 4 Hepatitis A     Delaware 1998              25     4     758939
## 5 Hepatitis A      Vermont 1998              34     4     601400
```

3.

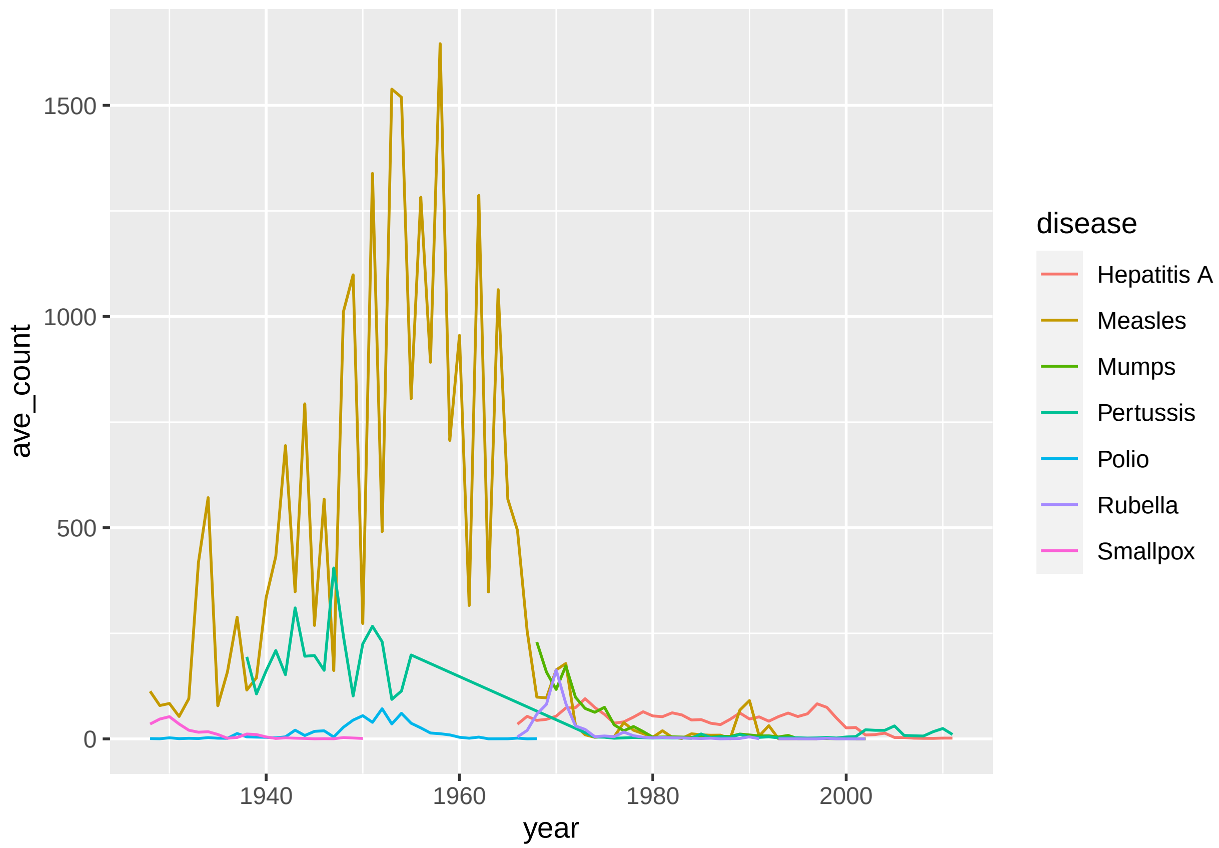
For the state of Texas, (a) Add a variable ave_count, representing the average reported case count per weeks_reporting for each year.

```
full_data %>%
  filter(state == "Texas") %>% #filter to get only state of Texas
  group_by(year) %>% #grouped by each year
  mutate(ave_count = count/weeks_reporting) #added ave_count variable by mutate()
```

```
## # A tibble: 315 x 7
## # Groups:   year [84]
##    disease     state  year weeks_reporting count population ave_count
##    <fct>       <fct> <dbl>           <dbl> <dbl>      <dbl>     <dbl>
##  1 Hepatitis A Texas  1966              52  1808   10470937      34.8
##  2 Hepatitis A Texas  1967              51  2727   10628322      53.5
##  3 Hepatitis A Texas  1968              50  2190   10798697      43.8
##  4 Hepatitis A Texas  1969              50  2312   10986554      46.2
##  5 Hepatitis A Texas  1970              51  2741   11196730      53.7
##  6 Hepatitis A Texas  1971              51  3731   11433080      73.2
##  7 Hepatitis A Texas  1972              46  3407   11694123      74.1
##  8 Hepatitis A Texas  1973              48  4569   11976810      95.2
##  9 Hepatitis A Texas  1974              43  3200   12277800      74.4
## 10 Hepatitis A Texas  1975              49  2845   12593389      58.1
## # ... with 305 more rows
```

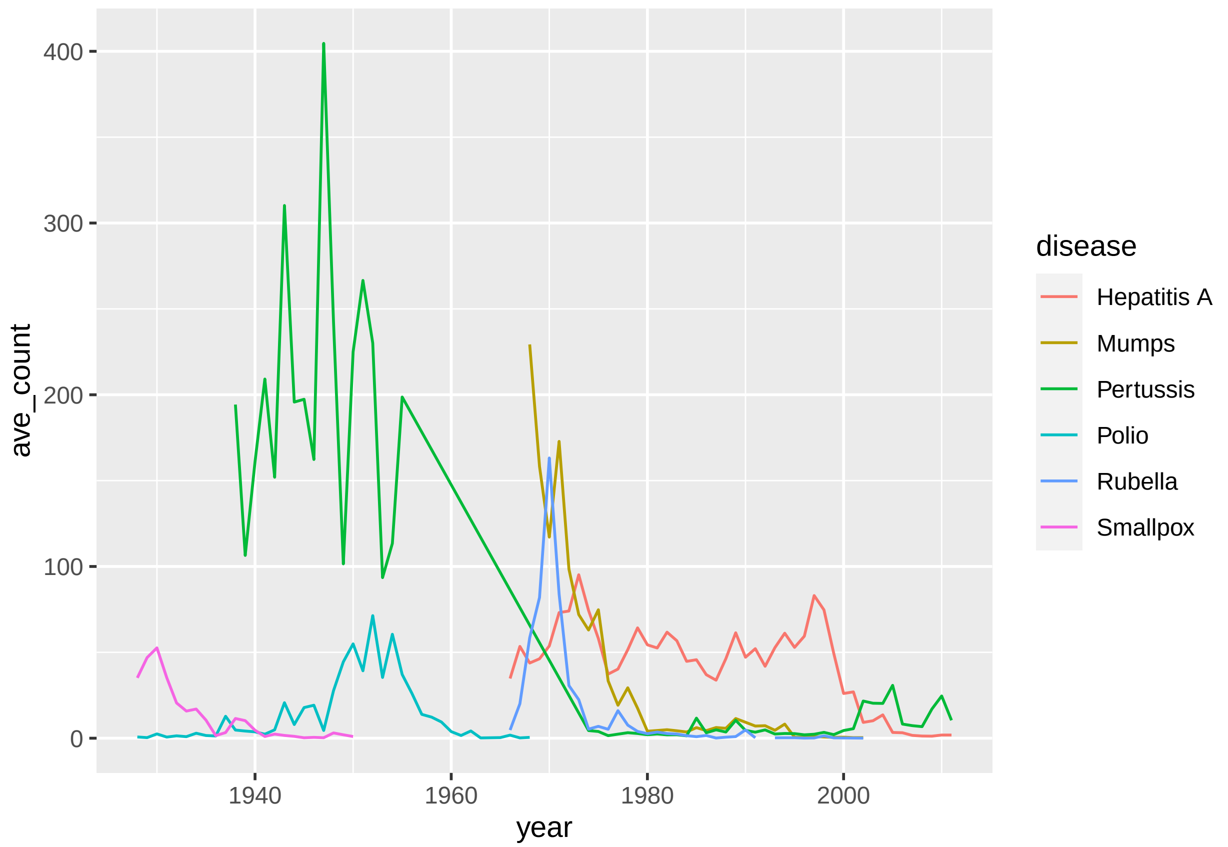
(b). Plot ave_count against the year while using different colors for different diseases.

```
full_data %>%
  filter(state == "Texas") %>% #used filter to get state of Texas
  group_by(year) %>% #grouped by each year
  mutate(ave_count = count/weeks_reporting) %>% #add ave_count variable to the data
  ggplot(mapping = aes(x=year, y=ave_count, color=disease)) + geom_line() #ggplot + geom_line to plot y
```



(c). Remove all the observations for disease "Measles" and redo the plot in (b).

```
full_data %>%
  subset(disease != "Measles") %>% #removed Measles by using subset() and !=
  filter(state == "Texas") %>% #to get state of Texas
  group_by(year) %>% #grouped by each year
  mutate(ave_count = count/weeks_reporting) %>% #add ave_count variable to the data
  ggplot(mapping = aes(x=year, y=ave_count, color=disease)) + geom_line() #line graph to see year again
```



4.

Redo the problem 3 for the state of New York. Write a paragraph to compare the results of the two states. (a)

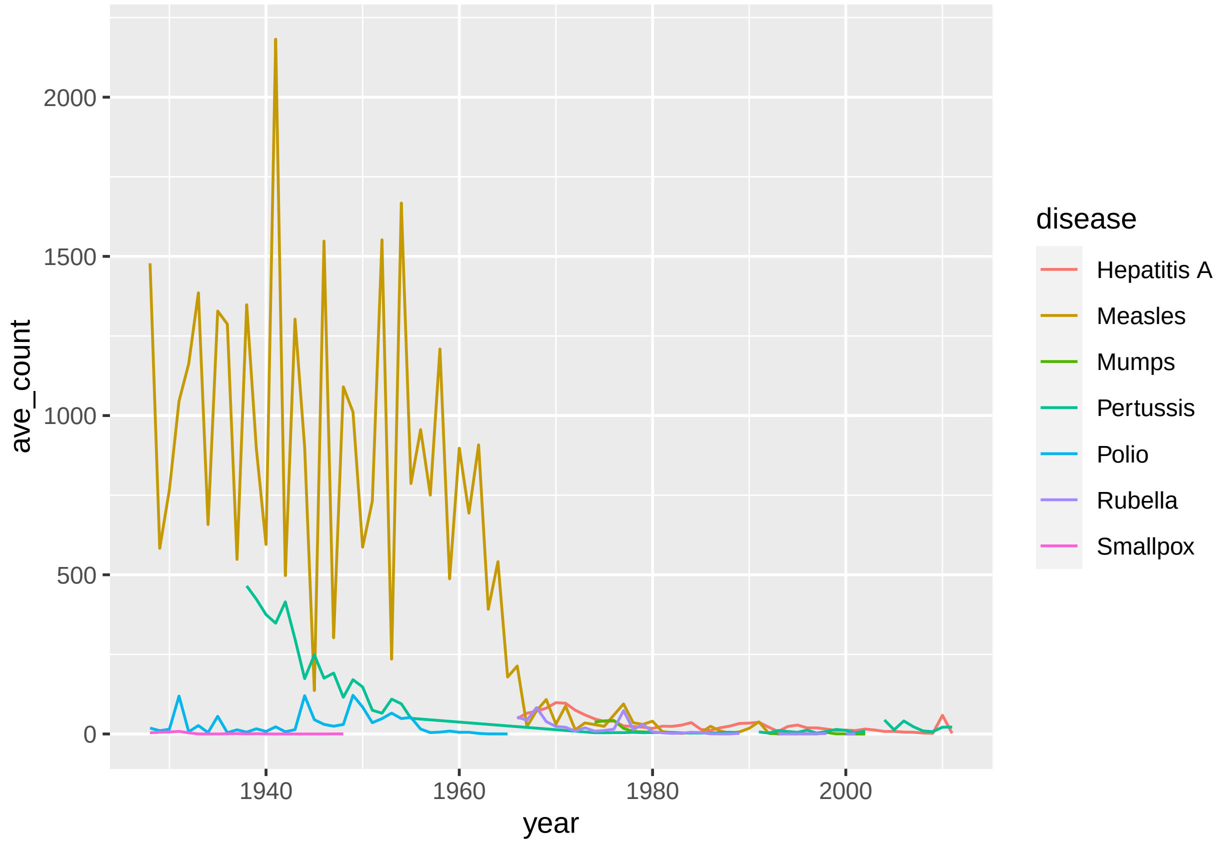
```
full_data %>%
  filter(state == "New York") %>%
  mutate(ave_count = count/weeks_reporting) %>%
  head(20)
```

```
##         disease    state year weeks_reporting count population ave_count
## 1  Hepatitis A New York 1966              50  2435   17895985  48.70000
## 2  Hepatitis A New York 1967              52  3394   18025684  65.26923
## 3  Hepatitis A New York 1968              52  3728   18128492  71.69231
## 4  Hepatitis A New York 1969              49  3976   18200269  81.14286
## 5  Hepatitis A New York 1970              51  5024   18236967  98.50980
## 6  Hepatitis A New York 1971              50  4825   18236388  96.50000
## 7  Hepatitis A New York 1972              46  3438   18203314  74.73913
## 8  Hepatitis A New York 1973              47  2821   18144367  60.02128
## 9  Hepatitis A New York 1974              46  2193   18066218  47.67391
## 10 Hepatitis A New York 1975              49  1932   17975503  39.42857
## 11 Hepatitis A New York 1976              51  2072   17878766  40.62745
## 12 Hepatitis A New York 1977              50  1266   17782428  25.32000
## 13 Hepatitis A New York 1978              49  1155   17692772  23.57143
## 14 Hepatitis A New York 1979             50  1065   17615962  21.30000
## 15 Hepatitis A New York 1980             38   666   17558072  17.52632
## 16 Hepatitis A New York 1981             45  1103   17523755  24.51111
## 17 Hepatitis A New York 1982             48  1145   17512164  23.85417
## 18 Hepatitis A New York 1983             47  1305   17521154  27.76596
## 19 Hepatitis A New York 1984             47  1683   17548657  35.80851
## 20 Hepatitis A New York 1985             30   434   17592652  14.46667
```

(b)

```
full_data %>%
  filter(state == "New York") %>%
  mutate(ave_count = count/weeks_reporting) %>%
  ggplot(mapping = aes(x=year, y=ave_count, color=disease)) + geom_line()
```

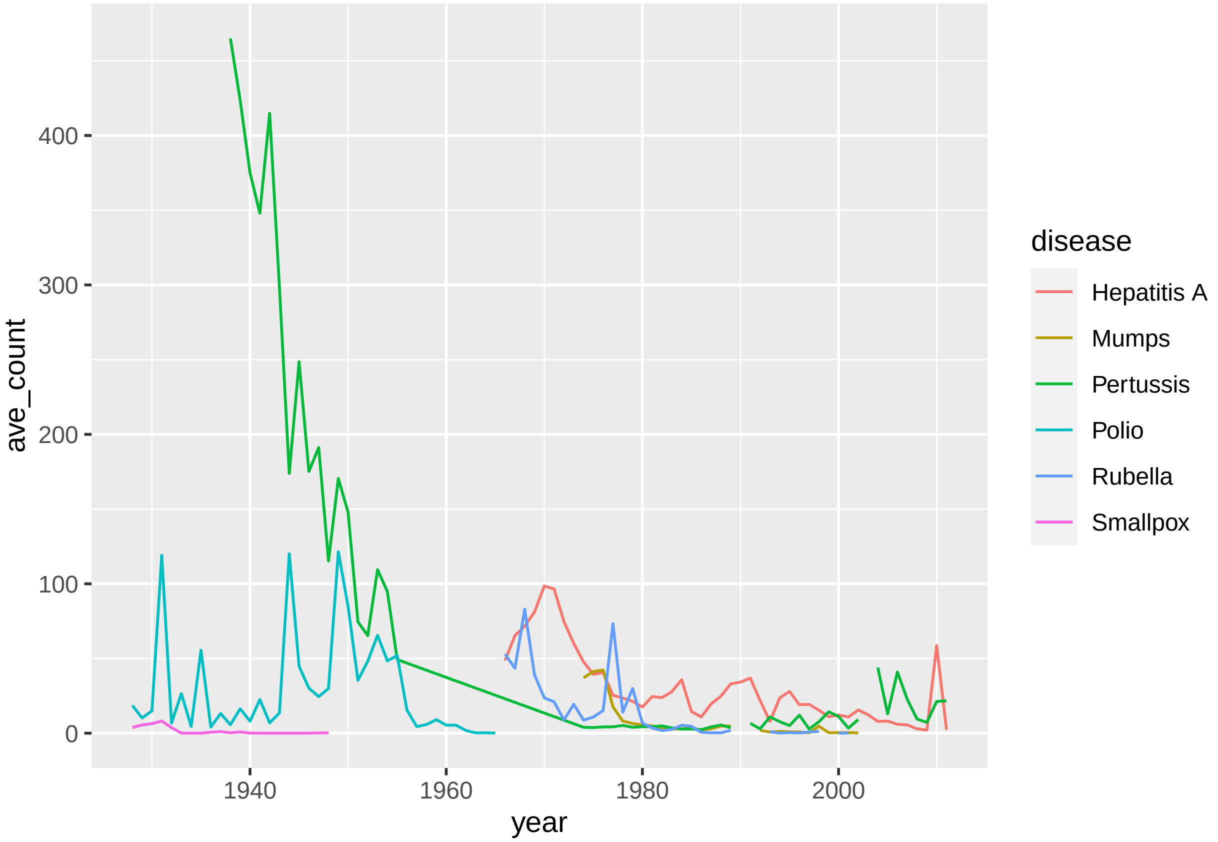
```
## Warning: Removed 8 row(s) containing missing values (geom_path).
```



(c)

```
full_data %>%
  subset(disease != "Measles") %>%
  filter(state == "New York") %>%
  mutate(ave_count = count/weeks_reporting) %>%
  ggplot(mapping = aes(x=year, y=ave_count, color=disease)) + geom_line()
```

```
## Warning: Removed 8 row(s) containing missing values (geom_path).
```



From the line graph that we created previously for Texas and New York, we can first clearly see that Measles had far more counts than other diseases. Also, measles prevalence peaked in the 1940s in New York and 1950s in Texas. After removing the measles, we can see that pertussis had most prevalence rate in the 1940s in New York. Also, in Texas, pertussis had the most counts from the 1940s to 1960s and started to decrease after the 1960s. Polio had more cases in New York than in Texas from the 1930s to the 1960s. We had very few counts for smallpox in both New York and Texas. For rubella, Texas peaked in the 1970s, while New York had relatively low cases. New York had Hepatitis cases in the mid-1960s, peaked in 1970, then decreased afterward. Texas also had a Hepatitis case in the 1960s, but it didn't decrease and kept the pace until 2000. We had few instances of mumps in New York, but Texas had relatively high cases.

5.

(a) For each state and year, find the total count of all diseases for the given state and year.

```
full_data %>%
  group_by(year, state) %>% #groupd by each year and state
  summarise(total_count = sum(count)) #used to summarise to create a new data frame which gives total c
```

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

```
## # A tibble: 4,284 x 3
## # Groups:   year [84]
##     year state                  total_count
##    <dbl> <fct>                        <dbl>
##  1  1928 Alabama                       9246
##  2  1928 Alaska                           0
##  3  1928 Arizona                       1268
##  4  1928 Arkansas                      9157
##  5  1928 California                    4960
##  6  1928 Colorado                      2510
##  7  1928 Connecticut                  10247
##  8  1928 Delaware                       607
##  9  1928 District Of Columbia          2609
## 10  1928 Florida                       1892
## # ... with 4,274 more rows
```

(b) For each state and year, find the disease count density, which is defined by the total count of all diseases divided by the population for the given state and year.

```
full_data %>%
  group_by(year, state) %>% #grouped by each year and state
  summarize(disease_count_density = sum(count)/mean(population)) #used summarise to create a disease_co
```

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

```
## # A tibble: 4,284 x 3
## # Groups:   year [84]
##     year state                  disease_count_density
##    <dbl> <fct>                                  <dbl>
##  1  1928 Alabama                           0.00357
##  2  1928 Alaska                           NA
##  3  1928 Arizona                           0.00303
##  4  1928 Arkansas                          0.00499
##  5  1928 California                        0.000948
##  6  1928 Colorado                          0.00247
##  7  1928 Connecticut                       0.00651
##  8  1928 Delaware                          0.00259
##  9  1928 District Of Columbia              0.00552
## 10  1928 Florida                           0.00139
## # ... with 4,274 more rows
```

6.

(a). Find the 3 state and year pairs that have the largest total count of all diseases.

```
full_data %>%
  group_by(state, year) %>% #to get pair of each year and state
  summarize(total_count = sum(count)) %>% #to make total count of all disease in each year and state
  arrange(desc(total_count)) #arranged hihgest to lowest to get the largest count
```

```
## `summarise()` has grouped output by 'state'. You can override using the
## `.groups` argument.
## # A tibble: 4,284 x 3
## # Groups:   state [51]
##    state         year total_count
##    <fct>        <dbl>       <dbl>
##  1 Pennsylvania  1938      146097
##  2 New York      1941      123598
##  3 Pennsylvania  1941      116071
##  4 California    1942      106847
##  5 Illinois      1938      104641
##  6 Ohio          1941       94161
##  7 New York      1938       94131
##  8 New York      1954       94116
##  9 New York      1946       91117
## 10 Pennsylvania  1935       89890
## # ... with 4,274 more rows
```

Pennsylvania 1938, New York 1941, and Pennsylvania 1941 had the largest count of all diseases.

(b). Fint the 3 state and year paris that have the largest disease count density.

```
full_data %>%
  group_by(state, year) %>% #paired each year and state
  summarize(disease_count_density = sum(count)/mean(population)) %>% #to get count density of each coun
  arrange(desc(disease_count_density))  #arranged highest to lowest to get all diseases
```

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

```
## # A tibble: 4,284 x 3
## # Groups:   state [51]
##    state      year disease_count_density
##    <fct>     <dbl>                 <dbl>
##  1 Vermont    1936                0.0297
##  2 Utah       1942                0.0289
##  3 Wisconsin  1938                0.0277
##  4 Vermont    1943                0.0243
##  5 Vermont    1938                0.0234
##  6 Utah       1938                0.0228
##  7 Utah       1934                0.0220
##  8 Montana    1939                0.0214
##  9 Utah       1940                0.0210
## 10 Vermont    1955                0.0195
## # ... with 4,274 more rows
```

Vermont 1936, Utah 1942, Wisconsin 1938 had the largest disease count density.