Intro to Data Science - HW 6

```
# Enter your name here: Hrishikesh Telang
```


Attribution statement: (choose only one and delete the rest)

```
# 1. I did this homework by myself, with help from the book and the professor.
```

This module: Data visualization is important because many people can make sense of data more easily when it is presented in graphic form. As a data scientist, you will have to present complex data to decision makers in a form that makes the data interpretable for them. From your experience with Excel and other tools, you know that there are a variety of **common data visualizations** (e.g., pie charts). How many of them can you name?

The most powerful tool for data visualization in R is called **ggplot**. Written by computer/data scientist **Hadley Wickham**, this **"graphics grammar"** tool builds visualizations in layers. This method provides immense flexibility, but takes a bit of practice to master.

Step 1: Make a copy of the data

A. Read the **who** dataset from this URL: https://intro-datascience.s3.us-east-2.amazonaws.com/who.csv into a new dataframe called **tb**.

Your new dataframe, tb, contains a so-called **multivariate time series**: a sequence of measurements on 23 Tuberculosis-related (TB) variables captured repeatedly over time (1980-2013). Familiarize yourself with the nature of the 23 variables by consulting the dataset's codebook which can be found here: https://intro-datascience.s3.us-east-2.amazonaws.com/TB_data_dictionary_2021-02-06.csv.

```
tb <- read.csv("https://intro-datascience.s3.us-east-2.amazonaws.com/who.csv") #reading the data from c
head(tb) #Display tb on console.
```

```
##   iso2 year new_sp new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524 new_sp_m2534
## 1   AD 1989     NA         NA          NA          NA           NA           NA
## 2   AD 1990     NA         NA          NA          NA           NA           NA
## 3   AD 1991     NA         NA          NA          NA           NA           NA
## 4   AD 1992     NA         NA          NA          NA           NA           NA
## 5   AD 1993     15         NA          NA          NA           NA           NA
## 6   AD 1994     24         NA          NA          NA           NA           NA
##   new_sp_m3544 new_sp_m4554 new_sp_m5564 new_sp_m65 new_sp_mu new_sp_f04
## 1           NA           NA           NA         NA        NA         NA
## 2           NA           NA           NA         NA        NA         NA
## 3           NA           NA           NA         NA        NA         NA
## 4           NA           NA           NA         NA        NA         NA
## 5           NA           NA           NA         NA        NA         NA
## 6           NA           NA           NA         NA        NA         NA
##   new_sp_f514 new_sp_f014 new_sp_f1524 new_sp_f2534 new_sp_f3544 new_sp_f4554
```

```
## 1          NA          NA           NA           NA           NA           NA
## 2          NA          NA           NA           NA           NA           NA
## 3          NA          NA           NA           NA           NA           NA
## 4          NA          NA           NA           NA           NA           NA
## 5          NA          NA           NA           NA           NA           NA
## 6          NA          NA           NA           NA           NA           NA
##   new_sp_f5564 new_sp_f65 new_sp_fu
## 1           NA         NA        NA
## 2           NA         NA        NA
## 3           NA         NA        NA
## 4           NA         NA        NA
## 5           NA         NA        NA
## 6           NA         NA        NA
```

```
View(tb) #Viewing the contents of tb.
```

B. How often were these measurements taken (in other words, at what frequency were the variables measured)? Put your answer in a comment.

```
#The variables were measured as per the frequency of each year starting from 1980 to 2008. The iso2 att
```

Step 2: Clean-up the NAs and create a subset

A. Let's clean up the iso2 attribute in **tb**

Hint: use *is.na()* – well use *! is.na()*

```
tb<- tb[!is.na(tb$iso2),] #Filters out only those rows which are not NA values in iso2.
head(tb) #Display tb.
```

```
##   iso2 year new_sp new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524 new_sp_m2534
## 1   AD 1989     NA          NA           NA           NA           NA           NA
## 2   AD 1990     NA          NA           NA           NA           NA           NA
## 3   AD 1991     NA          NA           NA           NA           NA           NA
## 4   AD 1992     NA          NA           NA           NA           NA           NA
## 5   AD 1993     15          NA           NA           NA           NA           NA
## 6   AD 1994     24          NA           NA           NA           NA           NA
##   new_sp_m3544 new_sp_m4554 new_sp_m5564 new_sp_m65 new_sp_mu new_sp_f04
## 1           NA           NA           NA         NA        NA         NA
## 2           NA           NA           NA         NA        NA         NA
## 3           NA           NA           NA         NA        NA         NA
## 4           NA           NA           NA         NA        NA         NA
## 5           NA           NA           NA         NA        NA         NA
## 6           NA           NA           NA         NA        NA         NA
##   new_sp_f514 new_sp_f014 new_sp_f1524 new_sp_f2534 new_sp_f3544 new_sp_f4554
## 1          NA          NA           NA           NA           NA           NA
## 2          NA          NA           NA           NA           NA           NA
## 3          NA          NA           NA           NA           NA           NA
## 4          NA          NA           NA           NA           NA           NA
## 5          NA          NA           NA           NA           NA           NA
## 6          NA          NA           NA           NA           NA           NA
```

```
##   new_sp_f5564 new_sp_f65 new_sp_fu
## 1           NA         NA        NA
## 2           NA         NA        NA
## 3           NA         NA        NA
## 4           NA         NA        NA
## 5           NA         NA        NA
## 6           NA         NA        NA
```

B. Create a subset of **tb** containing **only the records for Canada ("CA" in the iso2 variable)**. Save it in a new dataframe called **tbCan**. Make sure this new df has **29 observations and 23 variables**.

```
tbCan <- subset(tb, tb$iso2 == 'CA') #subsets records only for iso2 variables which correspond to Canad
tbCan #Display tbCan on console.
```

```
##     iso2 year new_sp new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524
## 872   CA 1980    951         NA          NA          12           54
## 873   CA 1981    803         NA          NA           8           49
## 874   CA 1982    812         NA          NA           6           52
## 875   CA 1983    771         NA          NA           9           47
## 876   CA 1984    811         NA          NA           3           44
## 877   CA 1985    791         NA          NA          11           42
## 878   CA 1986    752         NA          NA           9           58
## 879   CA 1987    668         NA          NA           9           40
## 880   CA 1988    682         NA          NA           4           43
## 881   CA 1989    652         NA          NA          10           45
## 882   CA 1990    549         NA          NA           3           35
## 883   CA 1991    543         NA          NA           7           37
## 884   CA 1992    506         NA          NA           6           42
## 885   CA 1993    488         NA          NA           8           33
## 886   CA 1994    483         NA          NA           2           42
## 887   CA 1995    436         NA          NA           1           28
## 888   CA 1996    430         NA          NA           3           28
## 889   CA 1997    473         NA          NA           0           21
## 890   CA 1998    438         NA          NA           4           33
## 891   CA 1999    455         NA          NA           0           23
## 892   CA 2000    492         NA          NA           5           34
## 893   CA 2001    458         NA          NA           6           24
## 894   CA 2002    408         NA          NA           0           25
## 895   CA 2003    332         NA          NA           1           26
## 896   CA 2004    438         NA          NA           2           25
## 897   CA 2005    433         NA          NA           3           37
## 898   CA 2006    407          1           1           2           34
## 899   CA 2007    463          4           1           5           31
## 900   CA 2008    488          0           2           2           39
##     new_sp_m2534 new_sp_m3544 new_sp_m4554 new_sp_m5564 new_sp_m65 new_sp_mu
## 872           75           83          100          108        186        NA
## 873           61           64           87          103        141        NA
## 874           66           69           90           91        150        NA
## 875           63           62           90           92        123        NA
## 876           75           58           68           83        169        NA
## 877           70           59           77           81        168        NA
## 878           73           62           59           73        147        NA
## 879           71           60           49           64        129        NA
```

```
## 880          73          62          52          68         131          NA
## 881          56          60          54          62         122          NA
## 882          70          55          40          42         100          NA
## 883          79          53          37          36         110          NA
## 884          47          58          41          51          79          NA
## 885          47          53          43          33          74          NA
## 886          54          42          43          34          87          NA
## 887          31          60          34          41          70          NA
## 888          49          48          31          34          70          NA
## 889          55          44          30          44          90          NA
## 890          43          51          31          26          80          NA
## 891          47          51          36          33          94          NA
## 892          45          46          41          32          79          NA
## 893          49          56          40          22          76          NA
## 894          34          50          34          27          64          NA
## 895          36          37          32          21          42          NA
## 896          34          38          32          31          64          NA
## 897          45          44          40          20          68          NA
## 898          34          33          42          26          64          NA
## 899          41          51          50          35          75          NA
## 900          36          49          53          38          62           0
##     new_sp_f04 new_sp_f514 new_sp_f014 new_sp_f1524 new_sp_f2534 new_sp_f3544
## 872         NA          NA          18           62          51          34
## 873         NA          NA           6           46          57          26
## 874         NA          NA           7           51          57          30
## 875         NA          NA          11           50          50          29
## 876         NA          NA           9           51          59          28
## 877         NA          NA           5           30          56          19
## 878         NA          NA          10           33          54          33
## 879         NA          NA           8           39          48          29
## 880         NA          NA           6           38          56          27
## 881         NA          NA           6           37          51          23
## 882         NA          NA           1           30          38          26
## 883         NA          NA           4           23          37          31
## 884         NA          NA           2           27          28          21
## 885         NA          NA           6           22          50          22
## 886         NA          NA           3           37          37          19
## 887         NA          NA           7           33          28          22
## 888         NA          NA           2           23          34          28
## 889         NA          NA           1           36          44          26
## 890         NA          NA           1           26          31          26
## 891         NA          NA           4           33          31          28
## 892         NA          NA           4           33          40          30
## 893         NA          NA           5           23          41          33
## 894         NA          NA           6           32          31          26
## 895         NA          NA           3           21          28          25
## 896         NA          NA           0           34          55          34
## 897         NA          NA           6           28          40          27
## 898          0           4           4           39          30          25
## 899          0           2           2           32          33          33
## 900          0           3           3           36          39          39
##     new_sp_f4554 new_sp_f5564 new_sp_f65 new_sp_fu
## 872           31           33        104        NA
## 873           28           35         92        NA
```

```
## 874          25          38          80          NA
## 875          24          35          86          NA
## 876          28          36         100          NA
## 877          28          48          97          NA
## 878          20          26          95          NA
## 879          17          26          79          NA
## 880          16          26          80          NA
## 881          24          21          81          NA
## 882          17          20          72          NA
## 883           9          20          60          NA
## 884          11          15          78          NA
## 885          21          21          55          NA
## 886          11          13          59          NA
## 887          12          18          51          NA
## 888          14          16          50          NA
## 889          13          16          53          NA
## 890          14          18          54          NA
## 891          13          11          51          NA
## 892          25          12          66          NA
## 893          16          14          53          NA
## 894          17          17          45          NA
## 895          15           9          36          NA
## 896          19          22          48          NA
## 897          24          13          37          NA
## 898          16           6          52          NA
## 899          11          13          51          NA
## 900          27          20          45           0
```

```
View(tbCan) #Viewing the contents of tbCan.
```

C. A simple method for dealing with small amounts of **missing data** in a numeric variable is to **substitute the mean of the variable in place of each missing datum**. This expression locates (and reports to the console) all the missing data elements in the variable measuring the **number of positive pulmonary smear tests for male children 0-4 years old** (there are 26 data points missing)

```
tbCan$new_sp_m04[is.na(tbCan$new_sp_m04)]
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [26] NA

Error in eval(expr, envir, enclos): object 'tbCan' not found
Traceback:
```

D. Write a comment describing how that statement works.

```
#The expression catches all the NA expressions with is.na() and returns a TRUE/FALSE vector. The outer
```

E. Write 4 more statements to check if there is missing data for the number of positive pulmonary smear tests for: **male and female** children 0-14 years old (**new_sp_m014** and **new_sp_f014**), and **male and female citizens 65 years of age and older**, respectively. What does empty output suggest about the number of missing observations?

```
tbCan$new_sp_m014[is.na(tbCan$new_sp_m014)]
```

```
## integer(0)
```

```
tbCan$new_sp_f014[is.na(tbCan$new_sp_f014)]
```

```
## integer(0)
```

```
tbCan$new_sp_m65[is.na(tbCan$new_sp_m65)]
```

```
## integer(0)
```

```
tbCan$new_sp_f65[is.na(tbCan$new_sp_f65)]
```

```
## integer(0)
```

```
#The empty output suggests that there are no NA values present in all four of the attributes ie: new_sp
```

There is an R package called **imputeTS** specifically designed to repair missing values in time series data. We will use this instead of mean substitution. The **na_interpolation()** function in this package takes advantage of a unique characteristic of time series data: **neighboring points in time can be used to "guess" about a missing value in between**.

F. Install the **imputeTS** package (if needed) and use **na_interpolation( )** on the variable from part C. Don't forget that you need to save the results back to the **tbCan** dataframe. Also update any attribute discussed in part E (if needed).

```
#install.packages('imputeTS') #This installs imputeTS
library(imputeTS) #We are calling the library of imputeTS
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
tbCan$new_sp_m04 <- na_interpolation(tbCan$new_sp_m04) #guesses the missing values in between
head(tbCan)
```

```
##     iso2 year new_sp new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524
## 872   CA 1980    951          1          NA          12           54
## 873   CA 1981    803          1          NA           8           49
## 874   CA 1982    812          1          NA           6           52
## 875   CA 1983    771          1          NA           9           47
## 876   CA 1984    811          1          NA           3           44
## 877   CA 1985    791          1          NA          11           42
##     new_sp_m2534 new_sp_m3544 new_sp_m4554 new_sp_m5564 new_sp_m65 new_sp_mu
## 872           75           83          100          108        186        NA
## 873           61           64           87          103        141        NA
## 874           66           69           90           91        150        NA
```

```
## 875           63           62           90           92        123        NA
## 876           75           58           68           83        169        NA
## 877           70           59           77           81        168        NA
##     new_sp_f04 new_sp_f514 new_sp_f014 new_sp_f1524 new_sp_f2534 new_sp_f3544
## 872         NA          NA          18           62           51           34
## 873         NA          NA           6           46           57           26
## 874         NA          NA           7           51           57           30
## 875         NA          NA          11           50           50           29
## 876         NA          NA           9           51           59           28
## 877         NA          NA           5           30           56           19
##     new_sp_f4554 new_sp_f5564 new_sp_f65 new_sp_fu
## 872           31           33        104        NA
## 873           28           35         92        NA
## 874           25           38         80        NA
## 875           24           35         86        NA
## 876           28           36        100        NA
## 877           28           48         97        NA
```

G. Rerun the code from C and E above to check that all missing data have been fixed.

```
tbCan$new_sp_m014 <- na_interpolation(tbCan$new_sp_m014)
tbCan$new_sp_f014 <- na_interpolation(tbCan$new_sp_f014)
tbCan$new_sp_m65 <- na_interpolation(tbCan$new_sp_m65)
tbCan$new_sp_f65 <- na_interpolation(tbCan$new_sp_f65)

head(tbCan) #Displays tbCan
```

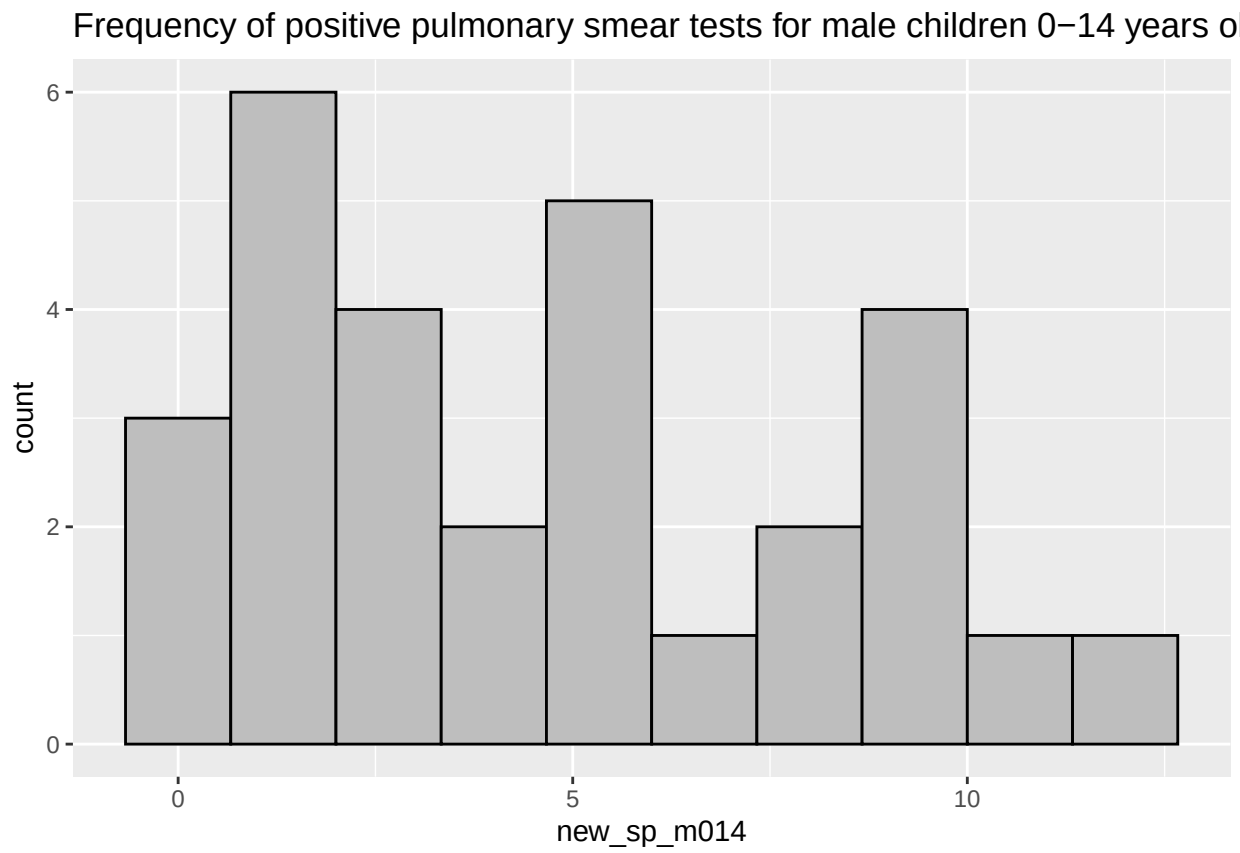
```
##     iso2 year new_sp new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524
## 872   CA 1980    951          1          NA          12           54
## 873   CA 1981    803          1          NA           8           49
## 874   CA 1982    812          1          NA           6           52
## 875   CA 1983    771          1          NA           9           47
## 876   CA 1984    811          1          NA           3           44
## 877   CA 1985    791          1          NA          11           42
##     new_sp_m2534 new_sp_m3544 new_sp_m4554 new_sp_m5564 new_sp_m65 new_sp_mu
## 872           75           83          100          108        186        NA
## 873           61           64           87          103        141        NA
## 874           66           69           90           91        150        NA
## 875           63           62           90           92        123        NA
## 876           75           58           68           83        169        NA
## 877           70           59           77           81        168        NA
##     new_sp_f04 new_sp_f514 new_sp_f014 new_sp_f1524 new_sp_f2534 new_sp_f3544
## 872         NA          NA          18           62           51           34
## 873         NA          NA           6           46           57           26
## 874         NA          NA           7           51           57           30
## 875         NA          NA          11           50           50           29
## 876         NA          NA           9           51           59           28
## 877         NA          NA           5           30           56           19
##     new_sp_f4554 new_sp_f5564 new_sp_f65 new_sp_fu
## 872           31           33        104        NA
## 873           28           35         92        NA
## 874           25           38         80        NA
## 875           24           35         86        NA
```

```
## 876          28          36         100         NA
## 877          28          48          97         NA
```

Step 3: Use ggplot to explore the distribution of each variable

Don't forget to install and library the ggplot2 package. Then: H. Create a histogram for **new_sp_m014**. Be sure to add a title and briefly describe what the histogram means in a comment.

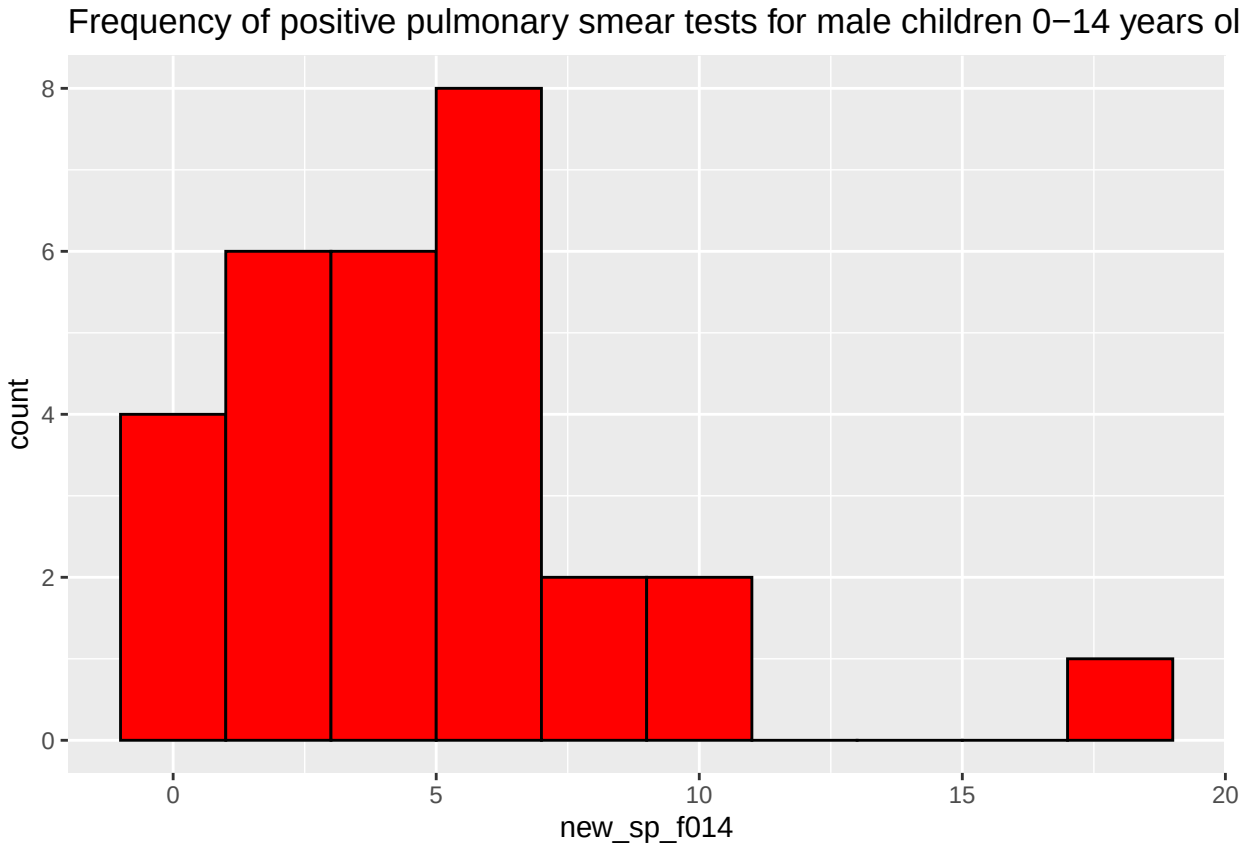
```
library(ggplot2) #Calls the ggplot2 library function
ggplot(tbCan) + ggtitle('Frequency of positive pulmonary smear tests for male children 0-14 years old')
```



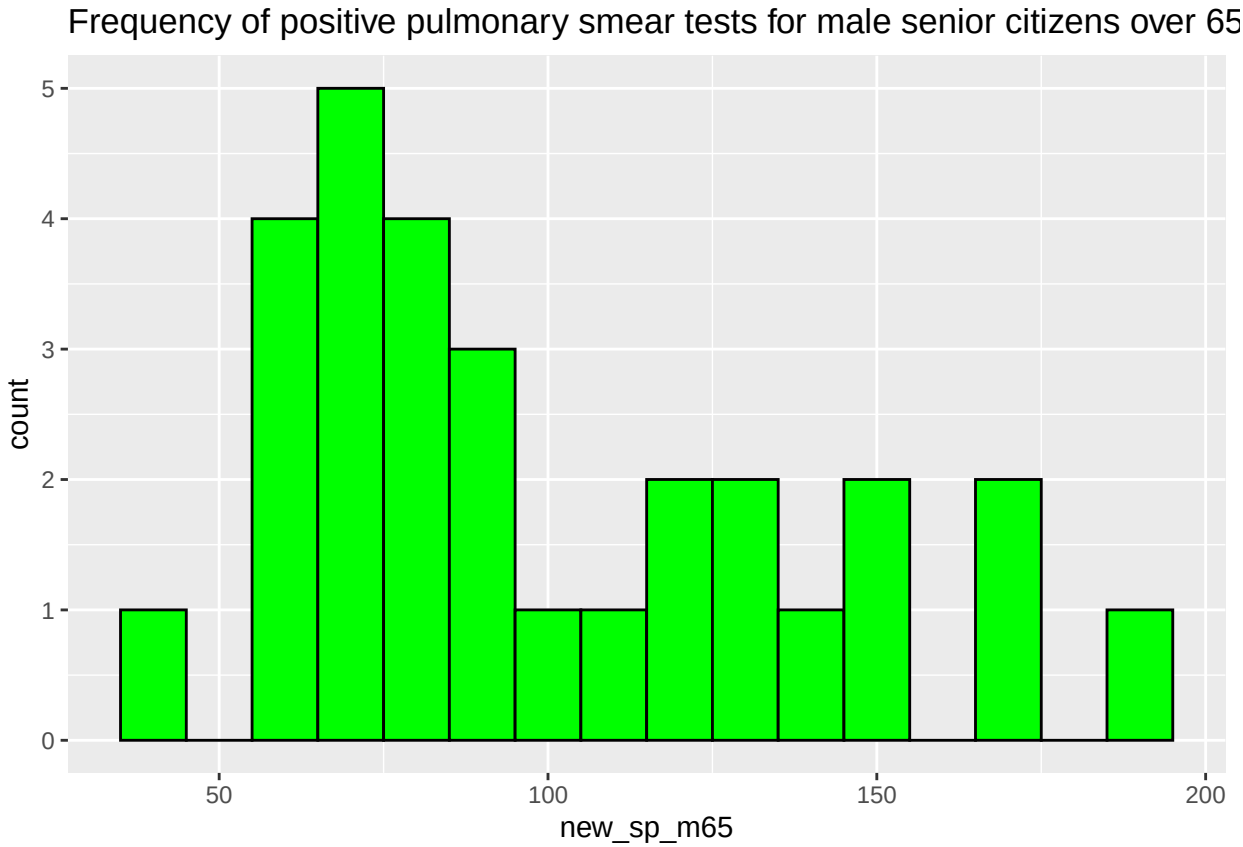

```
#The following histogram function here illustrates the positive pulmonary smear tests for male children
```

I. Create histograms (using ggplot) of each of the other three variables from E with ggplot( ). Which parameter do you need to adjust to make the other histograms look right?

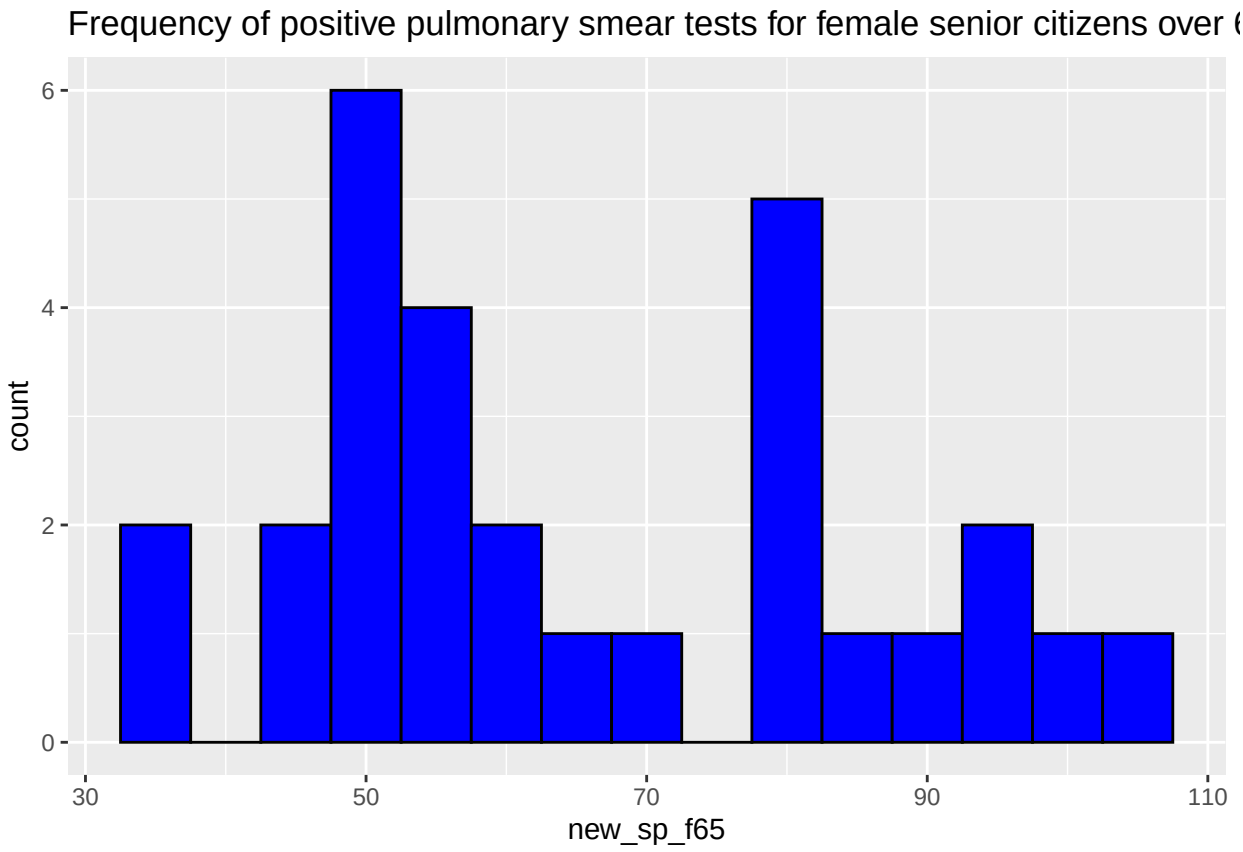
```
ggplot(tbCan) + ggtitle('Frequency of positive pulmonary smear tests for male children 0-14 years old')
```




```
ggplot(tbCan) + ggtitle('Frequency of positive pulmonary smear tests for male senior citizens over 65 y
```




```
ggplot(tbCan) + ggtitle('Frequency of positive pulmonary smear tests for female senior citizens over 65
```

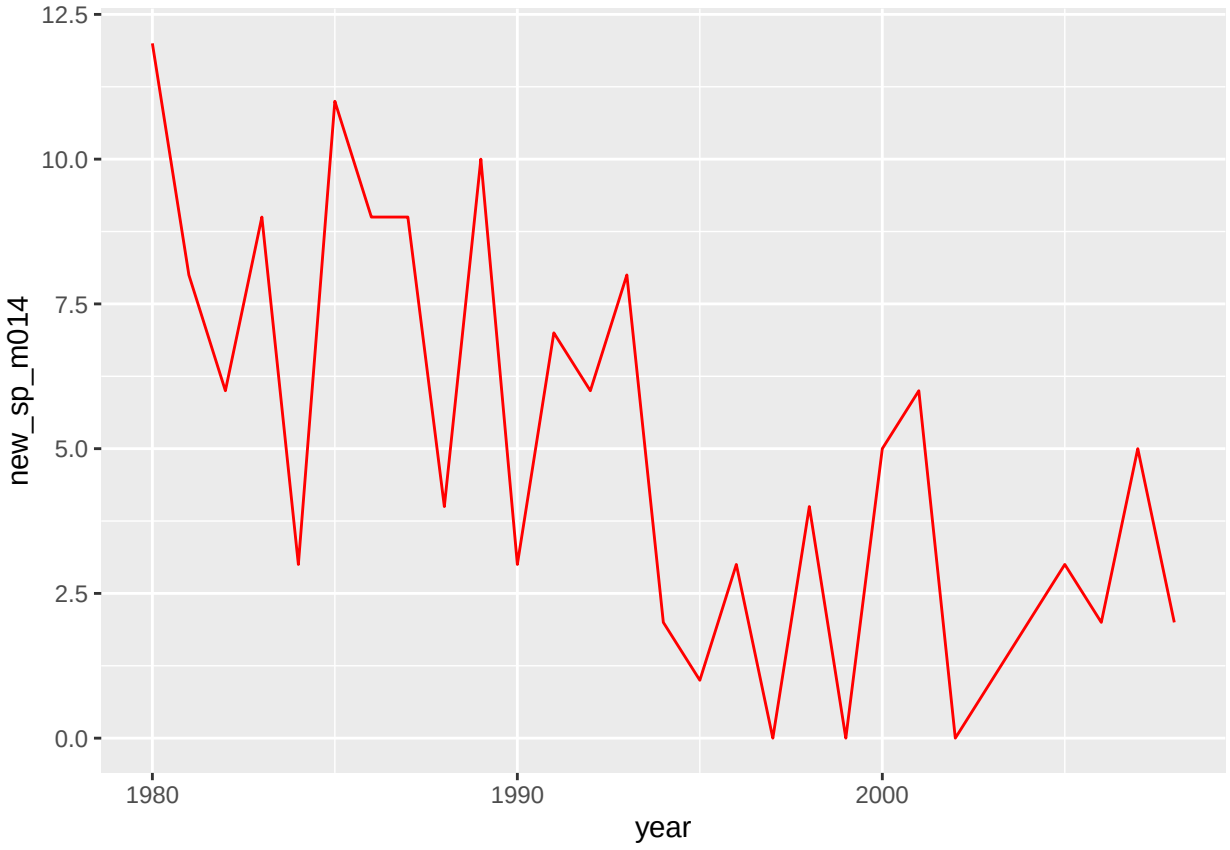



```
#I used the 'binwidth' parameter to make the necessary adjustments so as to make each of these histogra
```

Step 4: Explore how the data changes over time

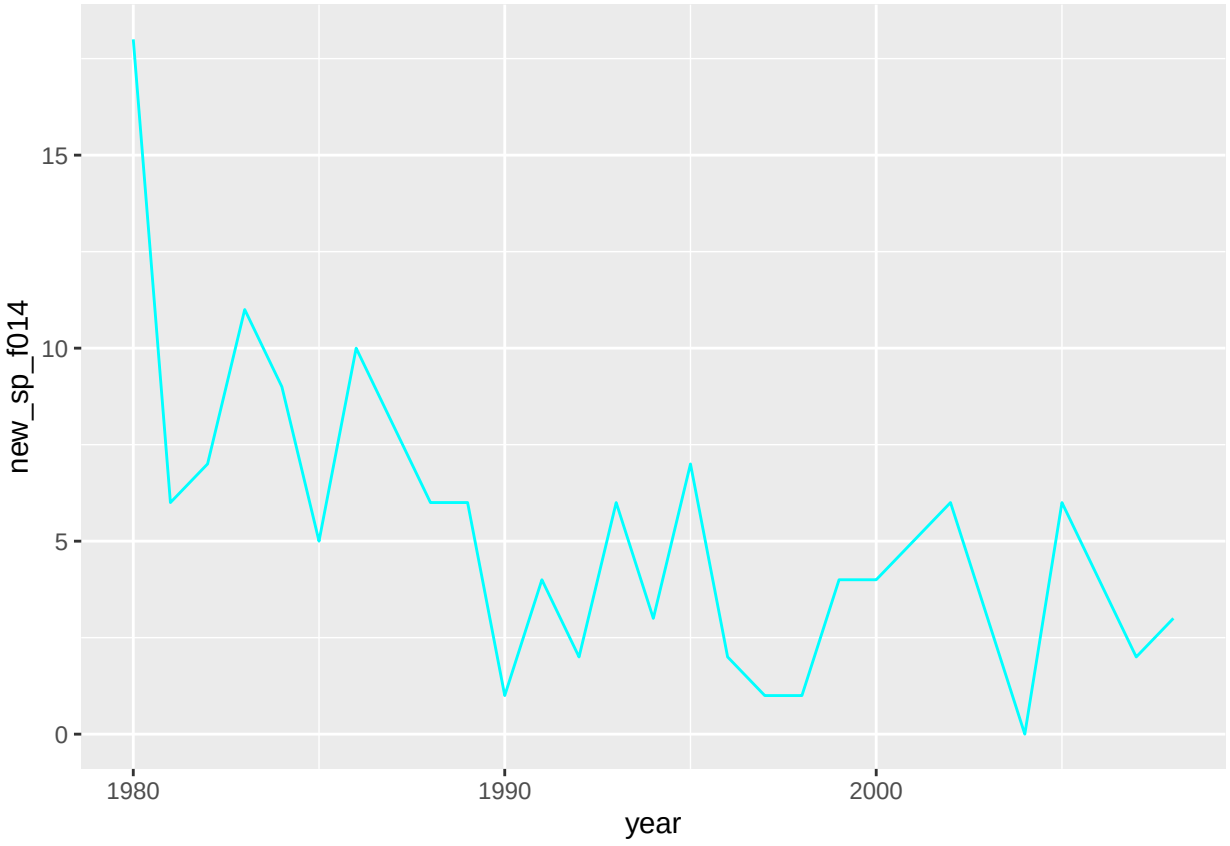
J. These data were collected in a period of several decades (1980-2013). You can thus observe changes over time with the help of a line chart. Create a **line chart**, with **year** on the X-axis and **new_sp_m014** on the Y-axis.

```
ggplot(tbCan) + geom_line(aes(x=year, y=new_sp_m014), colour='red')
```

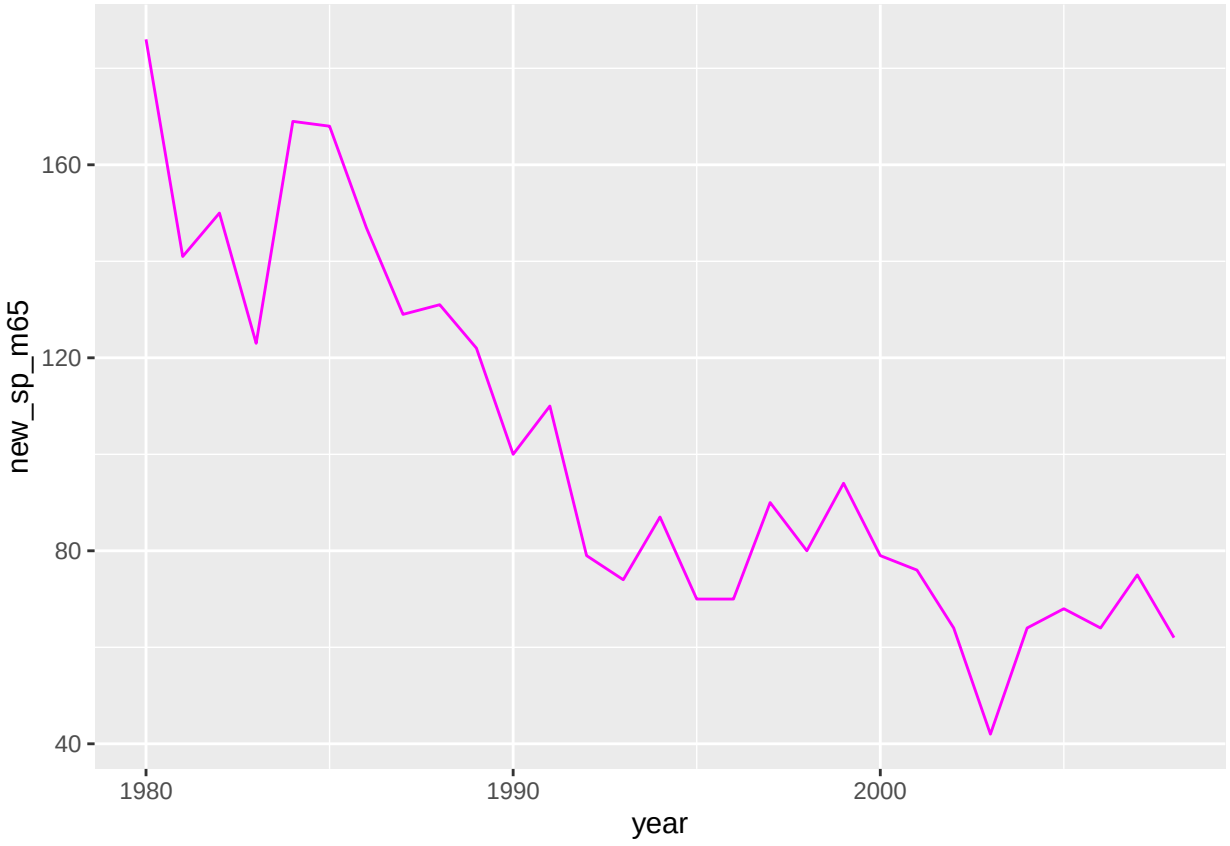


K. Next, create similar graphs for each of the other three variables. Change the **color** of the line plots (any color you want).

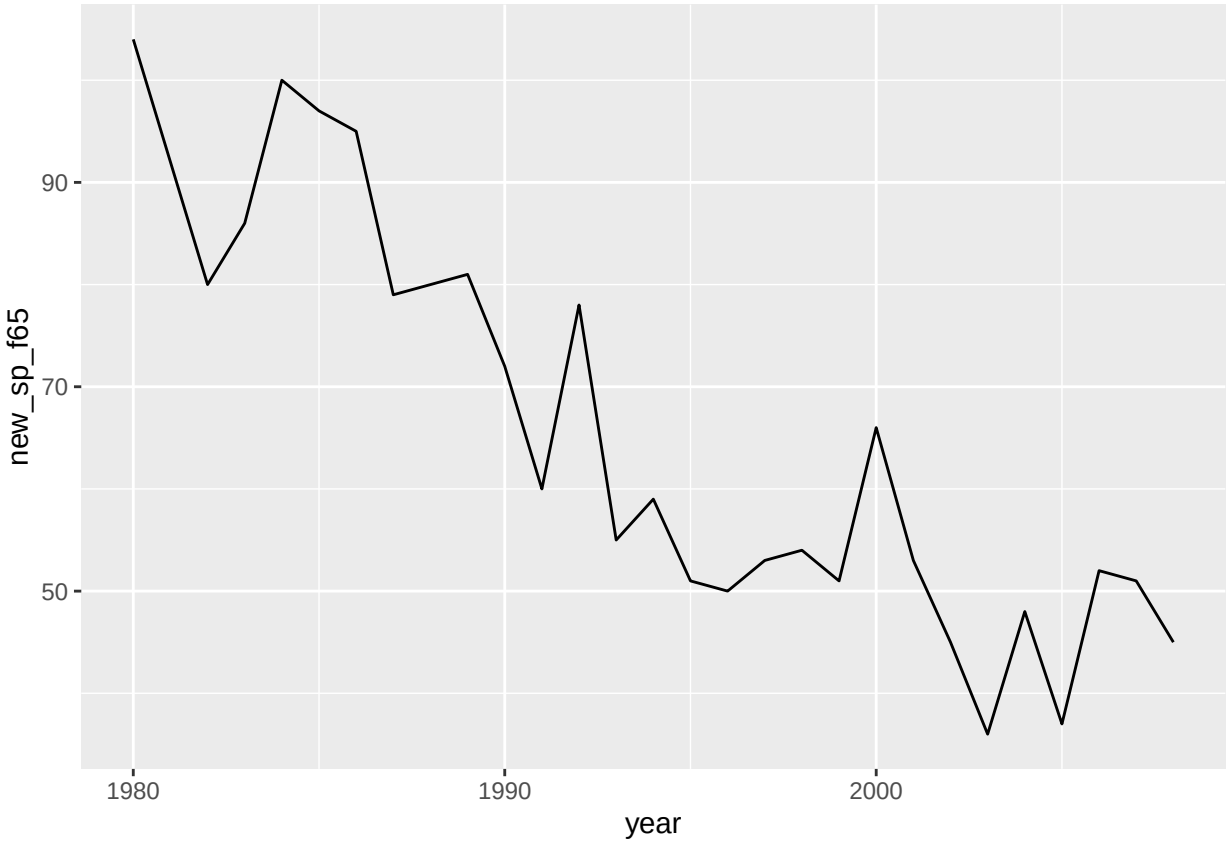
```
ggplot(tbCan) + geom_line(aes(x=year, y=new_sp_f014), colour='cyan')
```




```
ggplot(tbCan) + geom_line(aes(x=year, y=new_sp_m65), colour='magenta')
```




```
ggplot(tbCan) + geom_line(aes(x=year, y=new_sp_f65), colour='black')
```



L. Using vector math, create a new variable by combining the numbers from **new_sp_m014** and **new_sp_f014**. Save the resulting vector as a new variable in the **tbCan** df called **new_sp_combined014**. This new variable represents the number of positive pulmonary smear tests for male AND female children between the ages of 0 and 14 years of age. Do the same for SP **tests among citizens 65 years of age and older** and save the resulting vector in the tbCan variable called **new_sp_combined65**.

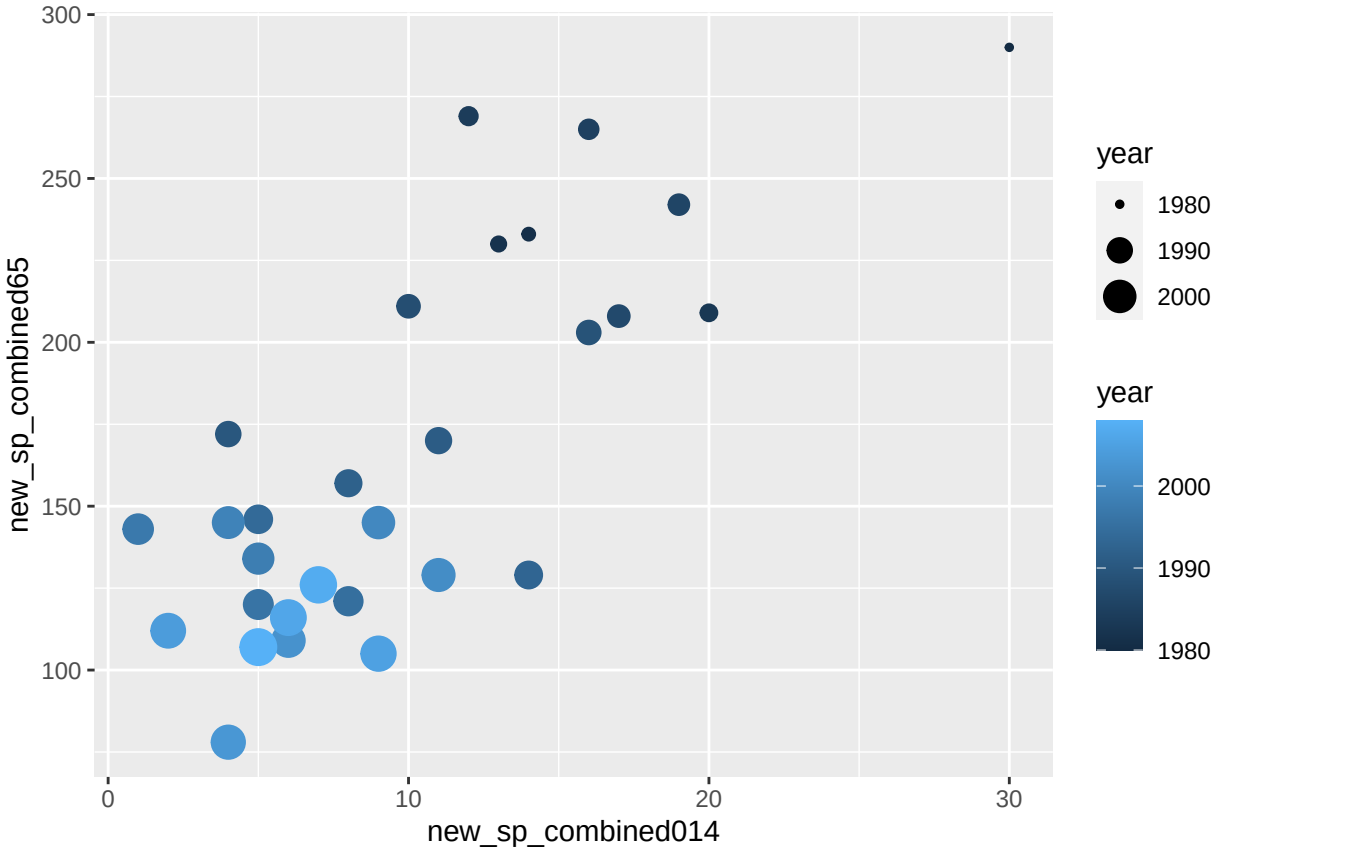
```
tbCan$new_sp_combined014 <- tbCan$new_sp_m014 + tbCan$new_sp_f014 #adds the values and stores in new_sp
tbCan$new_sp_combined65 <- tbCan$new_sp_m65 + tbCan$new_sp_f65 #adds the values and stores in new_sp_co
head(tbCan)
```

```
##     iso2 year new_sp new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524
## 872   CA 1980    951          1          NA          12           54
## 873   CA 1981    803          1          NA           8           49
## 874   CA 1982    812          1          NA           6           52
## 875   CA 1983    771          1          NA           9           47
## 876   CA 1984    811          1          NA           3           44
## 877   CA 1985    791          1          NA          11           42
##     new_sp_m2534 new_sp_m3544 new_sp_m4554 new_sp_m5564 new_sp_m65 new_sp_mu
## 872           75           83          100          108        186        NA
## 873           61           64           87          103        141        NA
## 874           66           69           90           91        150        NA
## 875           63           62           90           92        123        NA
## 876           75           58           68           83        169        NA
## 877           70           59           77           81        168        NA
```

```
##     new_sp_f04 new_sp_f514 new_sp_f014 new_sp_f1524 new_sp_f2534 new_sp_f3544
## 872         NA          NA          18           62           51           34
## 873         NA          NA           6           46           57           26
## 874         NA          NA           7           51           57           30
## 875         NA          NA          11           50           50           29
## 876         NA          NA           9           51           59           28
## 877         NA          NA           5           30           56           19
##     new_sp_f4554 new_sp_f5564 new_sp_f65 new_sp_fu new_sp_combined014
## 872           31           33        104        NA                30
## 873           28           35         92        NA                14
## 874           25           38         80        NA                13
## 875           24           35         86        NA                20
## 876           28           36        100        NA                12
## 877           28           48         97        NA                16
##     new_sp_combined65
## 872               290
## 873               233
## 874               230
## 875               209
## 876               269
## 877               265
```

M. Finally, create a **scatter plot**, showing **new_sp_combined014** on the x axis, **new_sp_combined65** on the y axis, and having the **color and size** of the point represent **year**.

```
ggplot(tbCan) + geom_point(aes(x=new_sp_combined014, y=new_sp_combined65, colour=year, size=year))
```



N. Interpret this visualization – what insight does it provide?

```
#It can be perceived that the number of Tuberculosis cases have significantly reduced over the past two
```