Health Data Exploration

Required packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

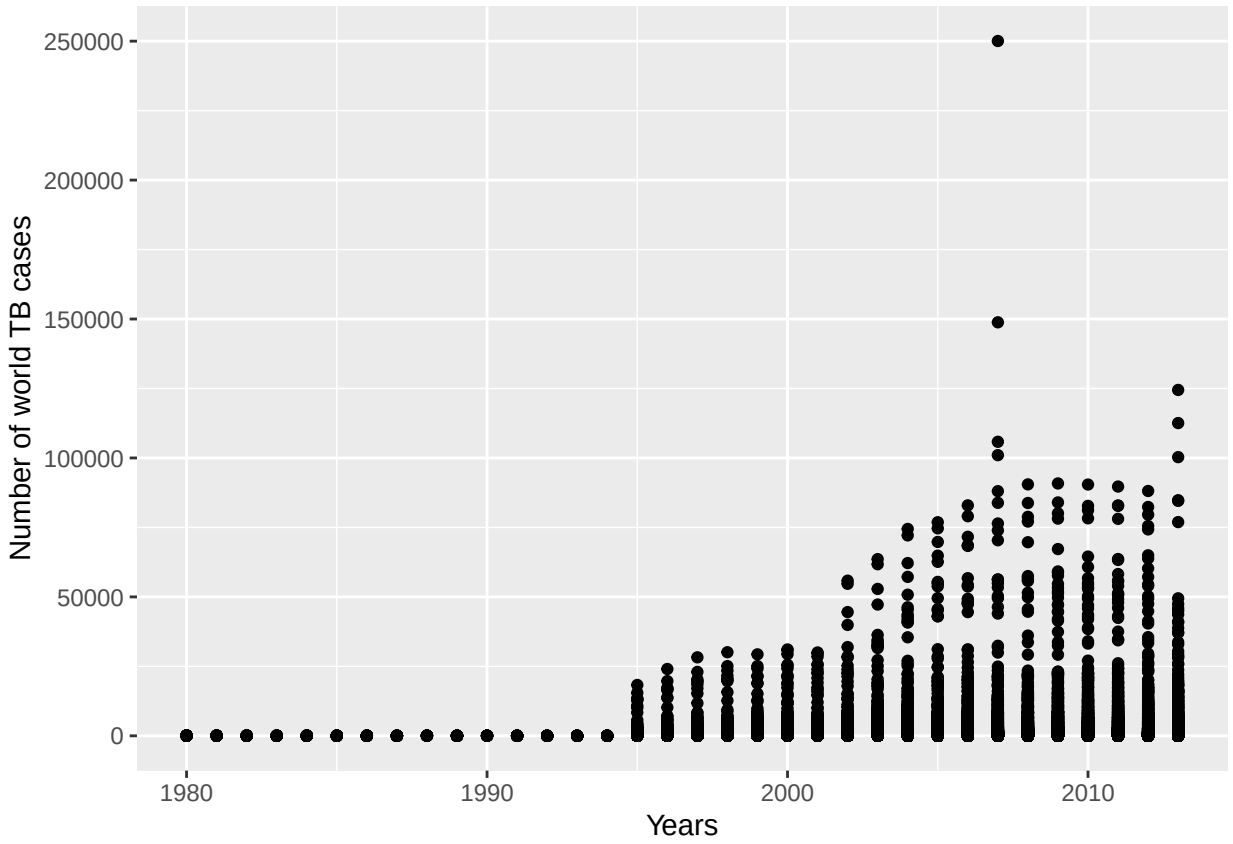
Load the R data file `who.rds` into the current session.

```
df_who <-readRDS("who.rds")
#df_who
#df_who %>%
#  count(country)
#df_who %>%
#  count(year)
```

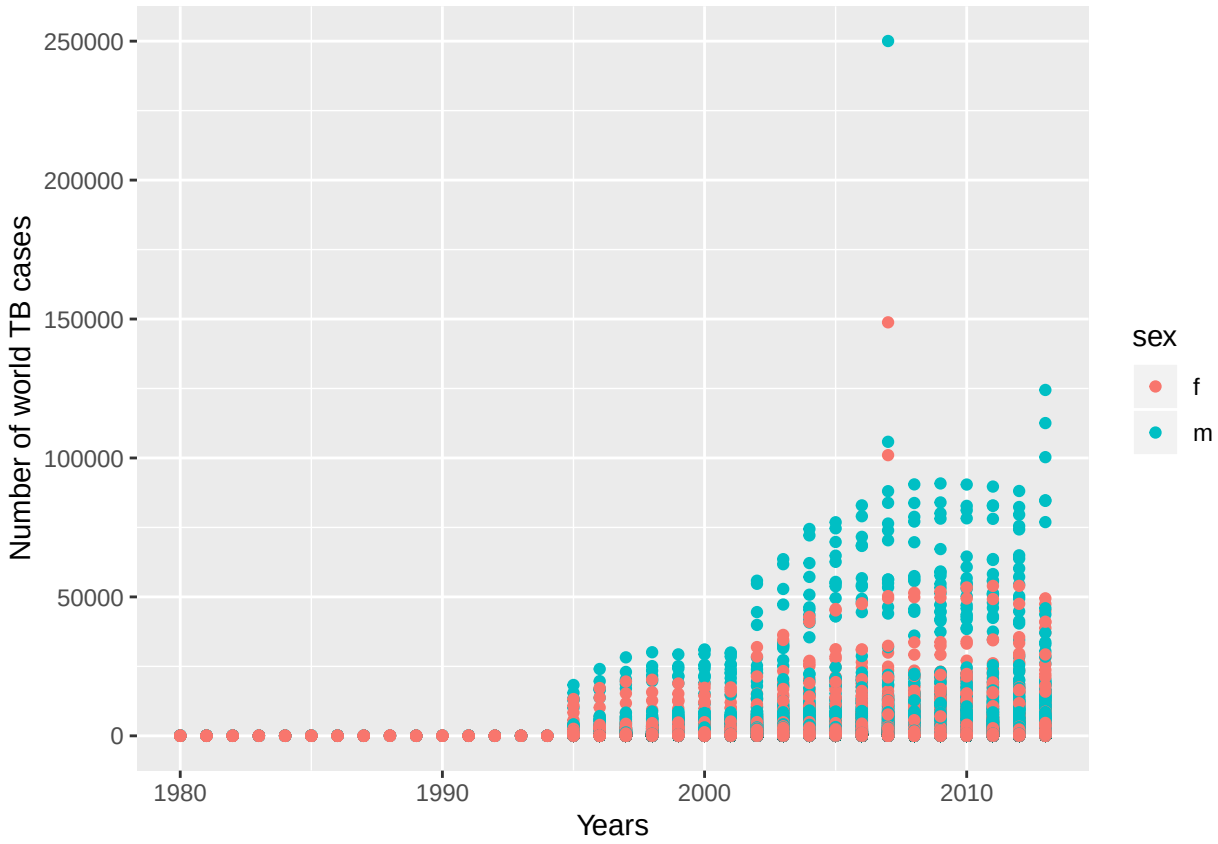
Note the dataset `who` has tuberculosis cases for 219 countries over the years of 1980 to 2013. However, data are not available for each year for every country.

You can look at the overall trend in the number of tuberculosis cases for these countries over three decades of given time period.

```
ggplot(data=df_who, aes(x=year,y=cases))+
  geom_point()+
  xlab("Years")+
  ylab("Number of world TB cases")
```




```
ggplot(data=df_who, aes(x=year,y=cases, color = sex))+
  geom_point()+
  xlab("Years")+
  ylab("Number of world TB cases")
```



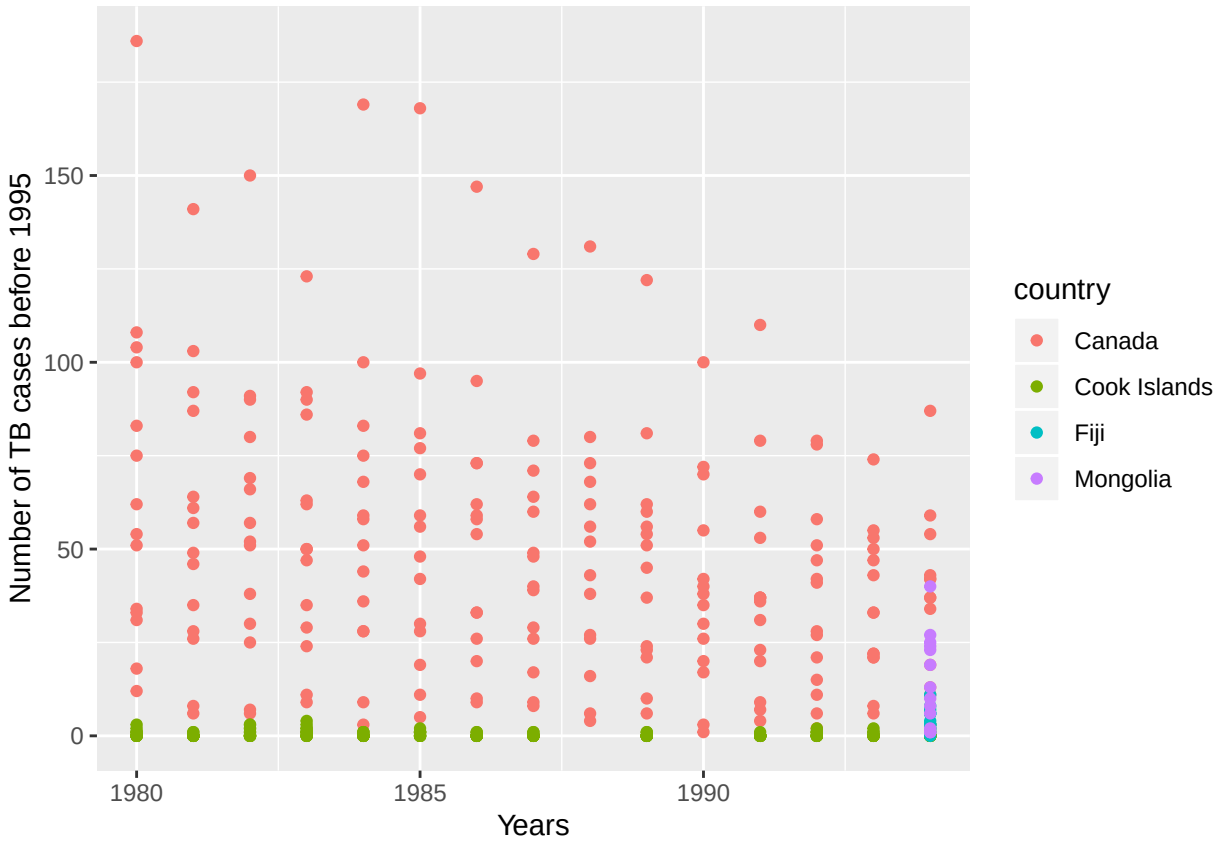
Overall an increasing number of TB cases over time. Note that there aren't many cases recorded until year 1995. You can explore which countries had TB cases before 1995

```
df_80_94 <- filter(df_who, year %in% 1980:1994)

df_80_94 %>%
count(country)
```

```
## # A tibble: 4 x 2
##   country          n
##   <chr>        <int>
## 1 Canada         210
## 2 Cook Islands   182
## 3 Fiji            14
## 4 Mongolia        14
```

```
ggplot(data=df_80_94, aes(x=year,y=cases, color = country))+
  geom_point()+
  xlab("Years")+
  ylab("Number of TB cases before 1995")
```



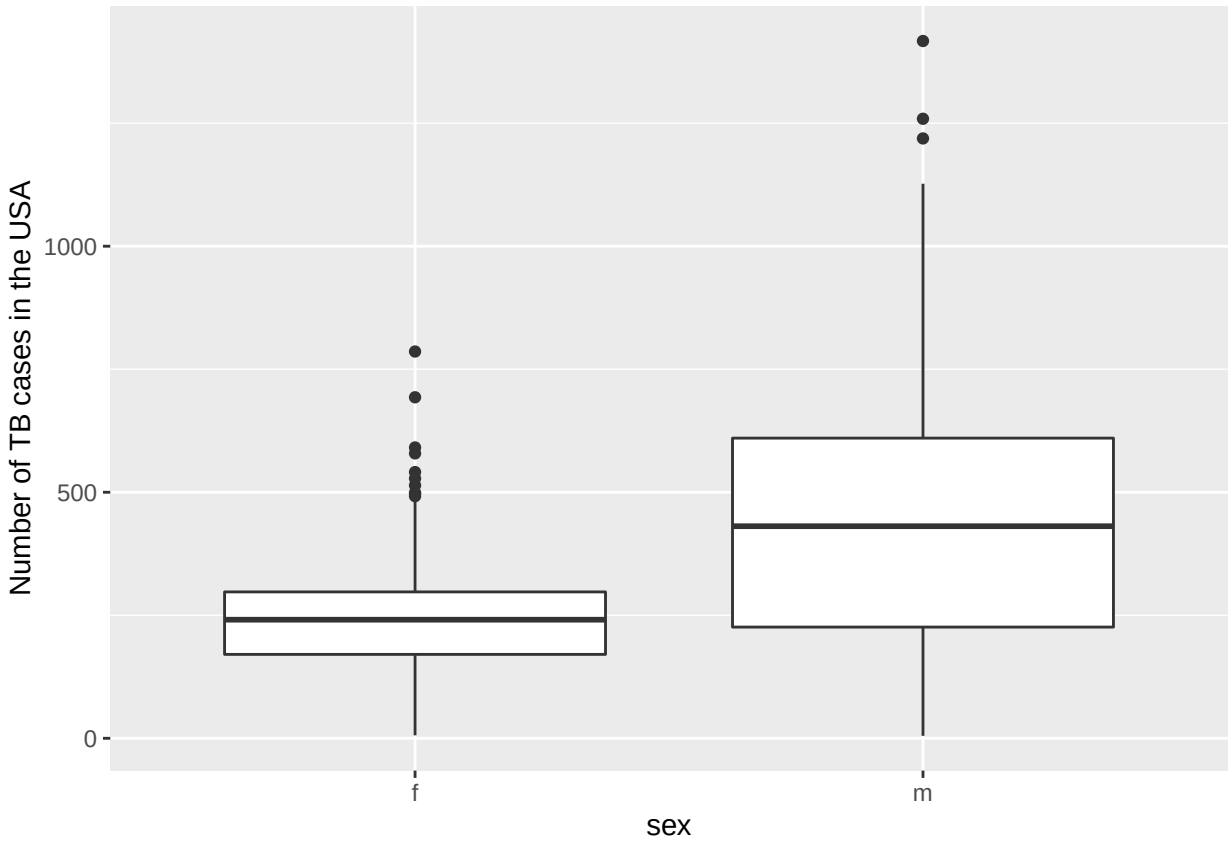
Note only four countries had records of TB cases before year 1995. Majority of these cases were reported in Canada and Cooks Island with a downward trend over time.

Let us explore the trend in the number of tuberculosis cases in the United States.

```
usa<- filter(df_who,country == "United States of America")
#usa %>%
 # count(year)
```

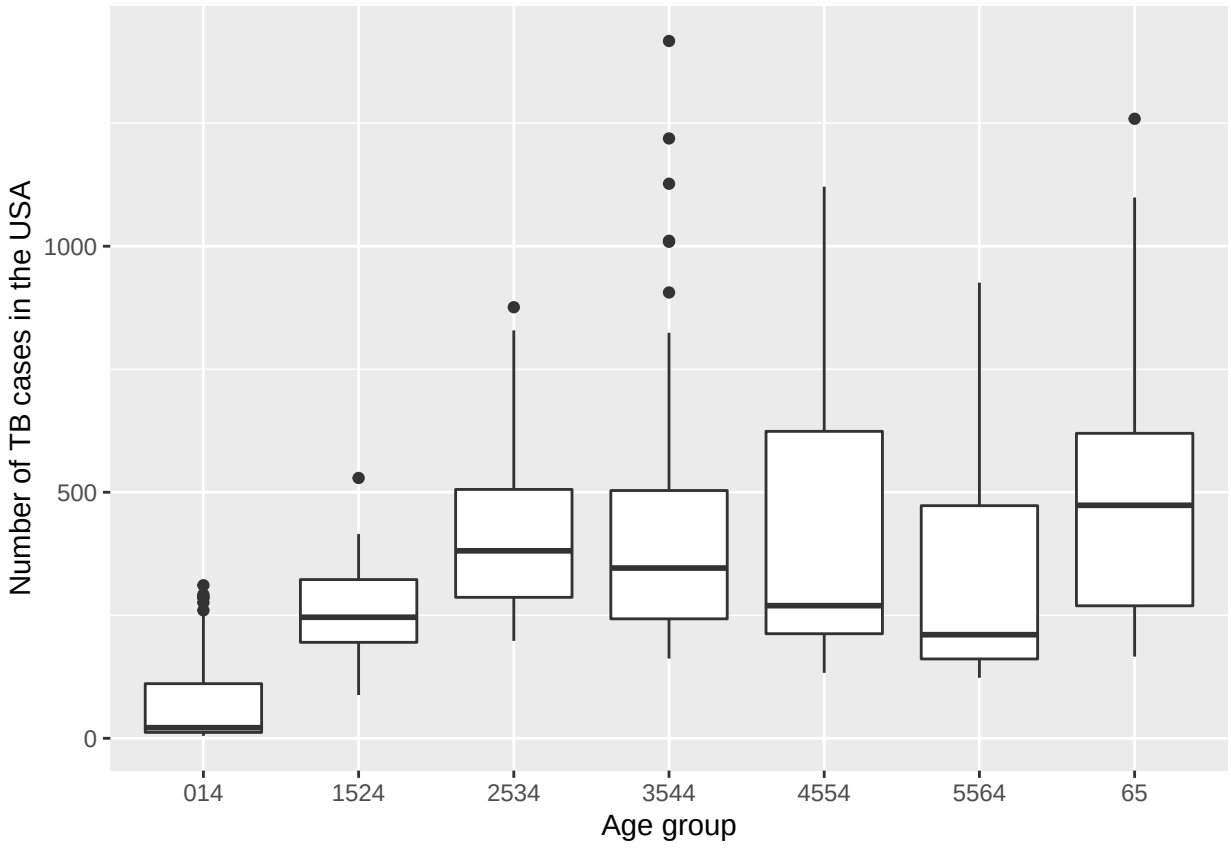
The following boxplot presents the distribution of the number of TB cases by gender in the United States.

```
ggplot(data = usa, mapping = aes(x = sex, y = cases)) + geom_boxplot()+ylab("Number of TB cases in the U
```



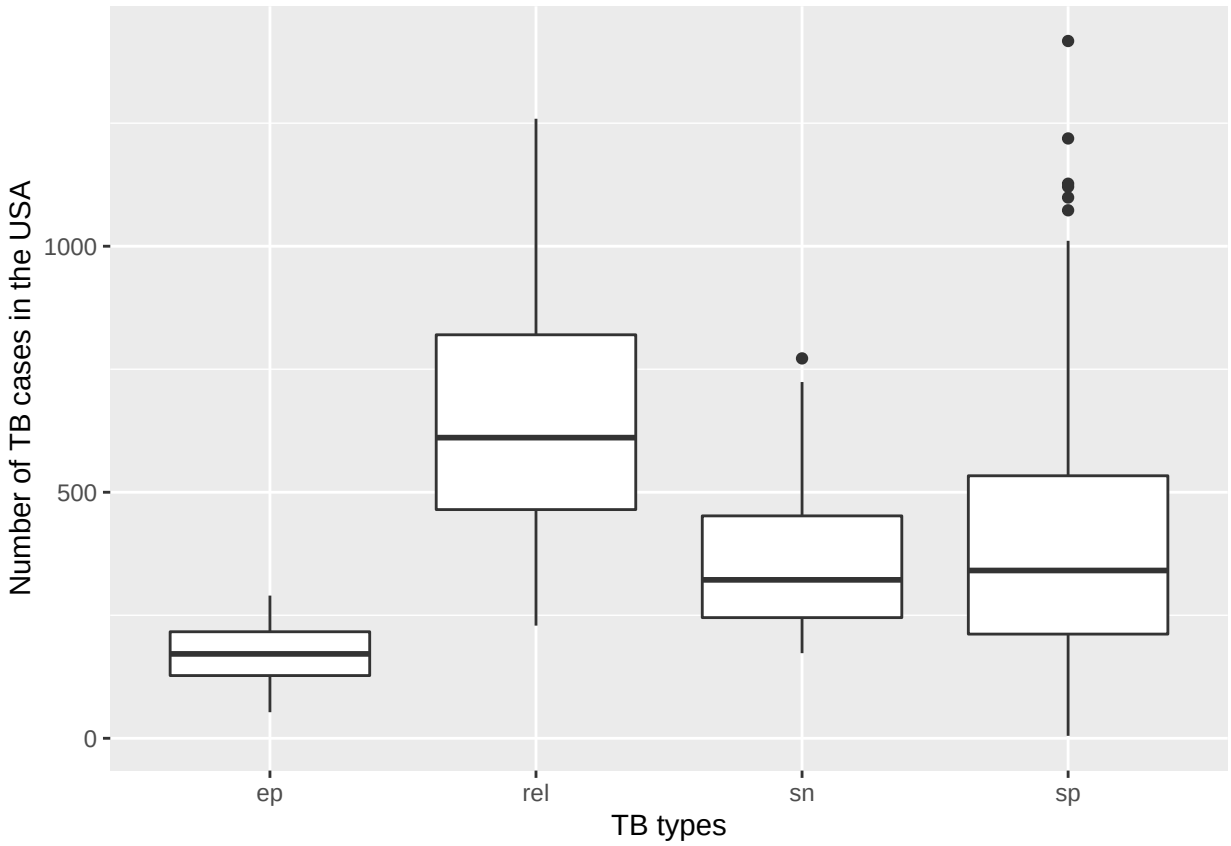
Note that males had higher number TB cases. There is also higher variability in the number of TB cases among males.

```
ggplot(data = usa, mapping = aes(x = age, y = cases)) + geom_boxplot()+xlab("Age group")+ylab("Number o
```



There is wide range of variability in the number of TB cases among different age groups with highest number of cases in the age group 35-44.

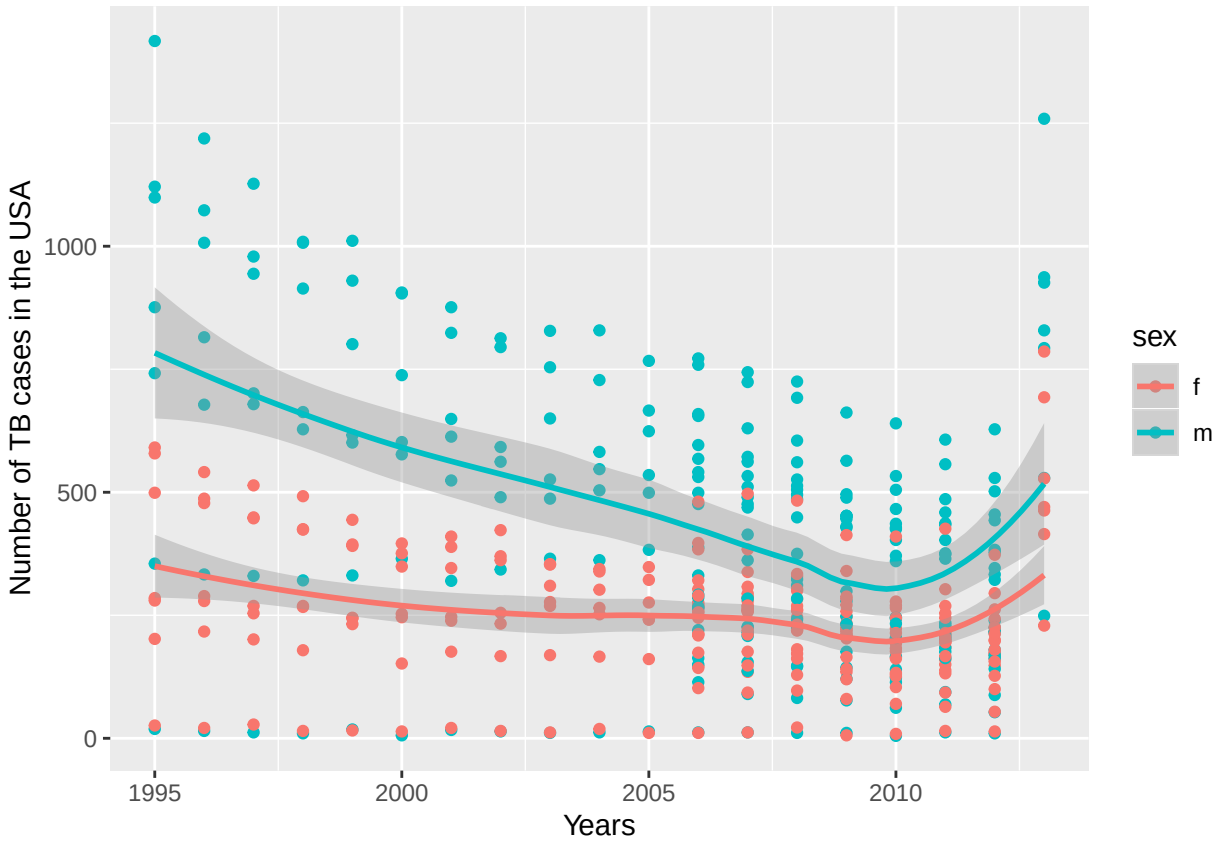
```
ggplot(data = usa, mapping = aes(x = type, y = cases)) + geom_boxplot() + xlab("TB types") +ylab("Numbe
```



In the USA, expulmonary (ep) and smear neagtive (sn) TB types are less common compared to the relapse (rel) of TB and smear positive (sp) type.

```
ggplot(data=usa, aes(x=year,y=cases, color = sex))+
  geom_point()+
  geom_smooth()+
  xlab("Years")+
  ylab("Number of TB cases in the USA")
```

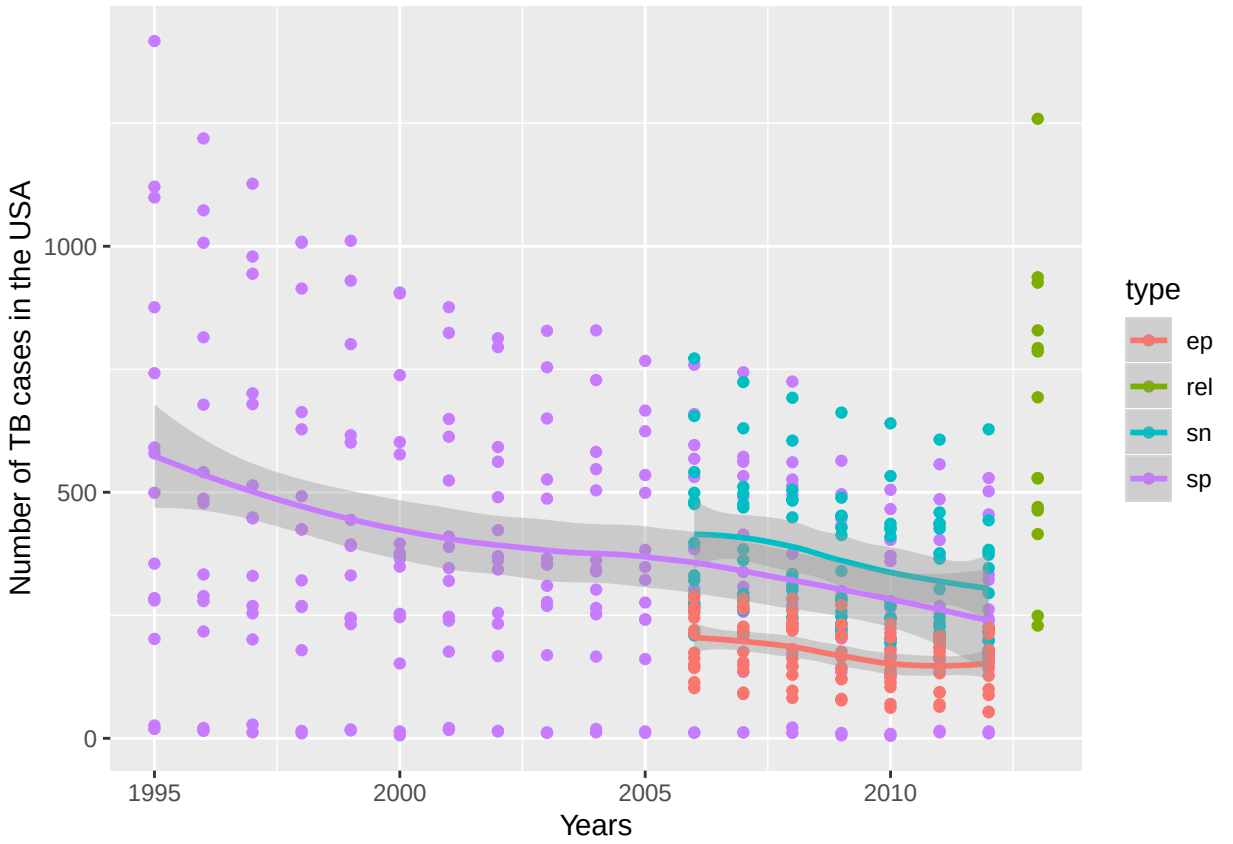
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Overall there is a downward trend in the number of TB cases. Among males a decreasing trend is more visible compared to that among females.

```
ggplot(data=usa, aes(x=year,y=cases, color = type))+
  geom_point()+
  geom_smooth()+
  xlab("Years")+
  ylab("Number of TB cases in the USA")
```

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



Until 2005, the TB cases were all smear positive. There is also a decreasing trend in all types of TB cases with an exception to higher relapse in the year 2013.

You can report descriptive statistics on the number of TB cases by gender and TB types as follows:

```
dftbl1 <- usa %>%
  group_by(sex) %>%
  summarise(mean = mean(cases),
            stdev = sd(cases),
            N = n())

kable(dftbl1)
```

sex	mean	stdev	N
f	245.6320	128.8939	231
m	442.7186	286.0144	231

```
dftbl2 <- usa %>%
  group_by(sex,type) %>%
  summarise(mean = mean(cases),
            stdev = sd(cases),
            N = n())

kable(dftbl2)
```

sex	type	mean	stdev	N
f	ep	170.7959	60.57983	49
f	rel	512.0000	183.37393	7
f	sn	282.8367	82.43061	49
f	sp	245.4683	135.00096	126
m	ep	173.6735	62.34193	49
m	rel	788.8571	322.07371	7
m	sn	441.6327	145.71686	49
m	sp	528.5397	305.59041	126