UC data wrangling

Anupriya Kushwanshi

May 25, 2019

reading data

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------
------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.2.5
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.2.1       v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

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

```
house <- read.table("C:/Users/anupr/Documents/Flex2/DAM/Hours-to-Pay-Mortgage.csv",sep = ",", he
ader = T, col.names = c("City","State","Median Price","Mortgage Rate 30yr","Monthly Mortgage Pay
ment","Median Income","Hours/month needed","Periods","Present Value", "X", "bin"))

tail(house)
```

```
##             City         State Median.Price Mortgage.Rate.30yr
## 92       Oakland    California     5,99,000              3.55%
## 93        Boston Massachusetts     6,99,000              3.56%
## 94 San Francisco    California    11,50,000              3.55%
## 95         Miami       Florida     4,49,000              3.57%
## 96   Los Angeles    California     7,48,000              3.55%
## 97      New York      New York     7,98,000              3.60%
##    Monthly.Mortgage.Payment Median.Income Hours.month.needed Periods
## 92                    2,165        54,618               82.7     360
## 93                    2,530        55,777               94.7     360
## 94                    4,157        81,294              106.7     360
## 95                    1,627        31,051              109.4     360
## 96                    2,704        50,205              112.4     360
## 97                    2,902        53,373              113.5     360
##    Present.Value  X bin
## 92      4,79,200 NA  NA
## 93      5,59,200 NA  NA
## 94      9,20,000 NA  NA
## 95      3,59,200 NA  NA
## 96      5,98,400 NA  NA
## 97      6,38,400 NA  NA
```

```
house <- select(house, everything(), -c("X","bin"))
head(house)
```

```
##        City     State Median.Price Mortgage.Rate.30yr
## 1    Toledo      Ohio       74,900              3.61%
## 2   Memphis Tennessee       88,500              3.59%
## 3 Cleveland      Ohio       70,000              3.61%
## 4   Buffalo  New York       90,000              3.60%
## 5 Baltimore  Maryland     1,39,000              3.58%
## 6   Wichita    Kansas     1,53,900              3.57%
##   Monthly.Mortgage.Payment Median.Income Hours.month.needed Periods
## 1                      273        33,687               16.9     360
## 2                      321        36,445               18.4     360
## 3                      255        26,150               20.3     360
## 4                      327        31,918               21.4     360
## 5                      504        42,241               24.9     360
## 6                      558        45,947               25.3     360
##   Present.Value
## 1        59,920
## 2        70,800
## 3        56,000
## 4        72,000
## 5      1,11,200
## 6      1,23,120
```

Checking for null values

```
colSums(is.na(house))
```

```
##                     City                    State             Median.Price
##                        0                        0                        0
##       Mortgage.Rate.30yr Monthly.Mortgage.Payment            Median.Income
##                        0                        0                        0
##       Hours.month.needed                  Periods            Present.Value
##                        0                        0                        0
```

```
str(house)
```

```
## 'data.frame':    97 obs. of  9 variables:
##  $ City                   : Factor w/ 97 levels "Albuquerque",..: 91 52 20 13 9 96 28 55 39
22 ...
##  $ State                  : Factor w/ 32 levels "Alaska","Arizona",..: 24 28 24 22 14 12 11
32 11 24 ...
##  $ Median.Price           : Factor w/ 88 levels "1,24,800","1,39,000",..: 83 86 82 88 2 6 5
1 3 9 ...
##  $ Mortgage.Rate.30yr     : Factor w/ 12 levels "3.54%","3.55%",..: 8 6 8 7 5 4 4 4 9 8 ...
##  $ Monthly.Mortgage.Payment: Factor w/ 90 levels "1,000","1,015",..: 40 43 39 44 47 51 50 46
48 54 ...
##  $ Median.Income          : Factor w/ 96 levels "1,05,355","26,150",..: 7 10 2 4 23 37 28 9
22 34 ...
##  $ Hours.month.needed     : num  16.9 18.4 20.3 21.4 24.9 25.3 25.9 26.2 26.3 27.3 ...
##  $ Periods                : int  360 360 360 360 360 360 360 360 360 360 ...
##  $ Present.Value          : Factor w/ 88 levels "1,11,200","1,16,000",..: 80 85 79 86 1 5 4
88 2 8 ...
```

Changing the class of numerical variables

```
  house$Median.Price <- as.integer(house$Median.Price)
  house$Median.Income <- as.factor(house$Median.Income)
  house$Monthly.Mortgage.Payment <- as.numeric(house$Monthly.Mortgage.Payment)
  house$Hours.month.needed <- as.numeric(house$Hours.month.needed)
  house$Mortgage.Rate.30yr <- house$Mortgage.Rate.30yr
```

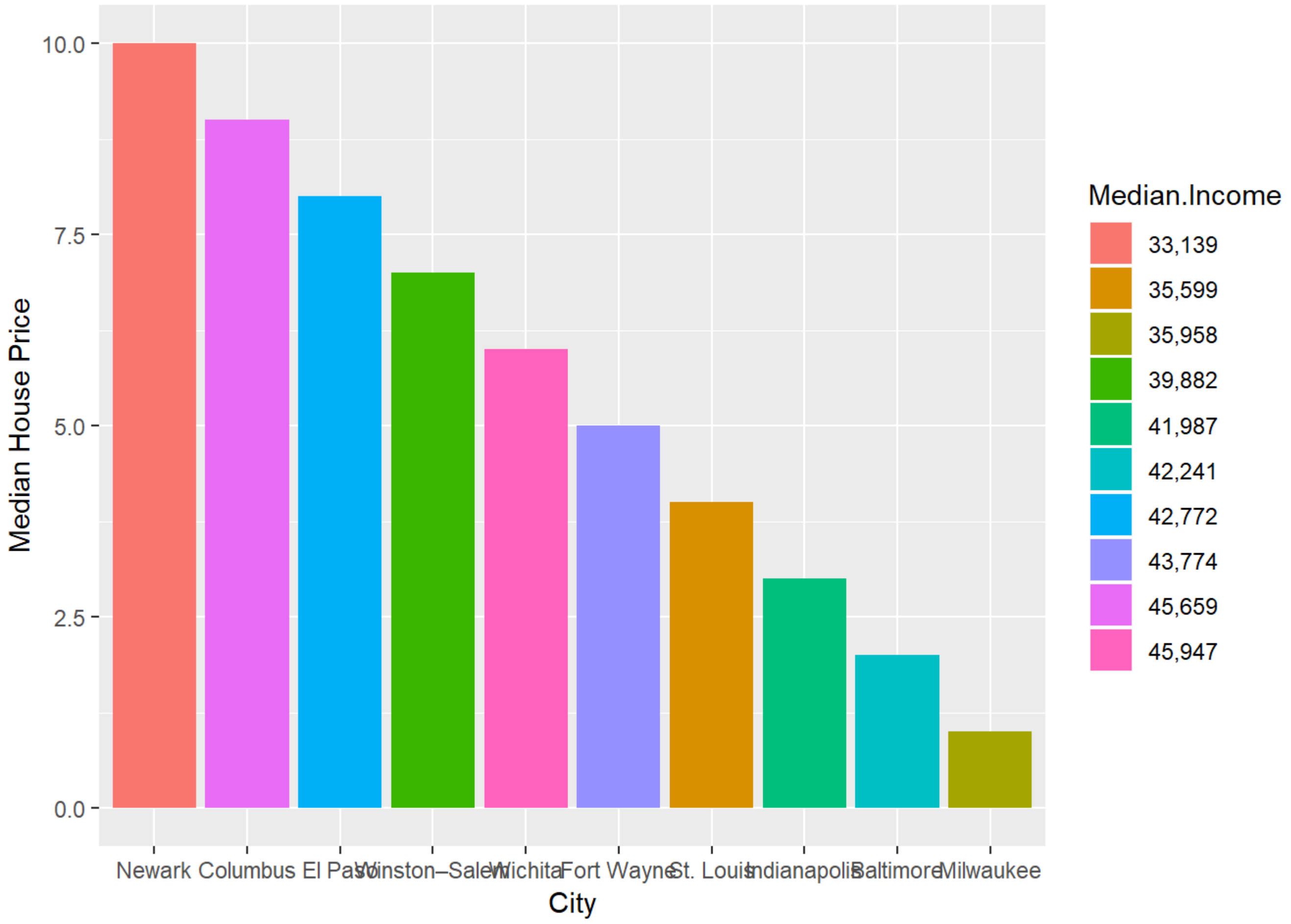
Plots using subsetted data

```
top_10_med_price <- head(arrange(house, by = Median.Price), 10)

str(top_10_med_price)
```

```
## 'data.frame':    10 obs. of  9 variables:
##  $ City                   : Factor w/ 97 levels "Albuquerque",..: 55 9 39 87 28 96 97 27 22
60
##  $ State                  : Factor w/ 32 levels "Alaska","Arizona",..: 32 14 11 17 11 12 23
29 24 20
##  $ Median.Price           : int  1 2 3 4 5 6 7 8 9 10
##  $ Mortgage.Rate.30yr     : Factor w/ 12 levels "3.54%","3.55%",..: 4 5 9 7 4 4 3 3 8 2
##  $ Monthly.Mortgage.Payment: num  46 47 48 49 50 51 52 53 54 55
##  $ Median.Income          : Factor w/ 96 levels "1,05,355","26,150",..: 9 23 22 8 28 37 17 2
7 34 5
##  $ Hours.month.needed     : num  26.2 24.9 26.3 31.6 25.9 25.3 30.3 28.5 27.3 38.6
##  $ Periods                : int  360 360 360 360 360 360 360 360 360 360
##  $ Present.Value          : Factor w/ 88 levels "1,11,200","1,16,000",..: 88 1 2 3 4 5 6 7 8
9
```

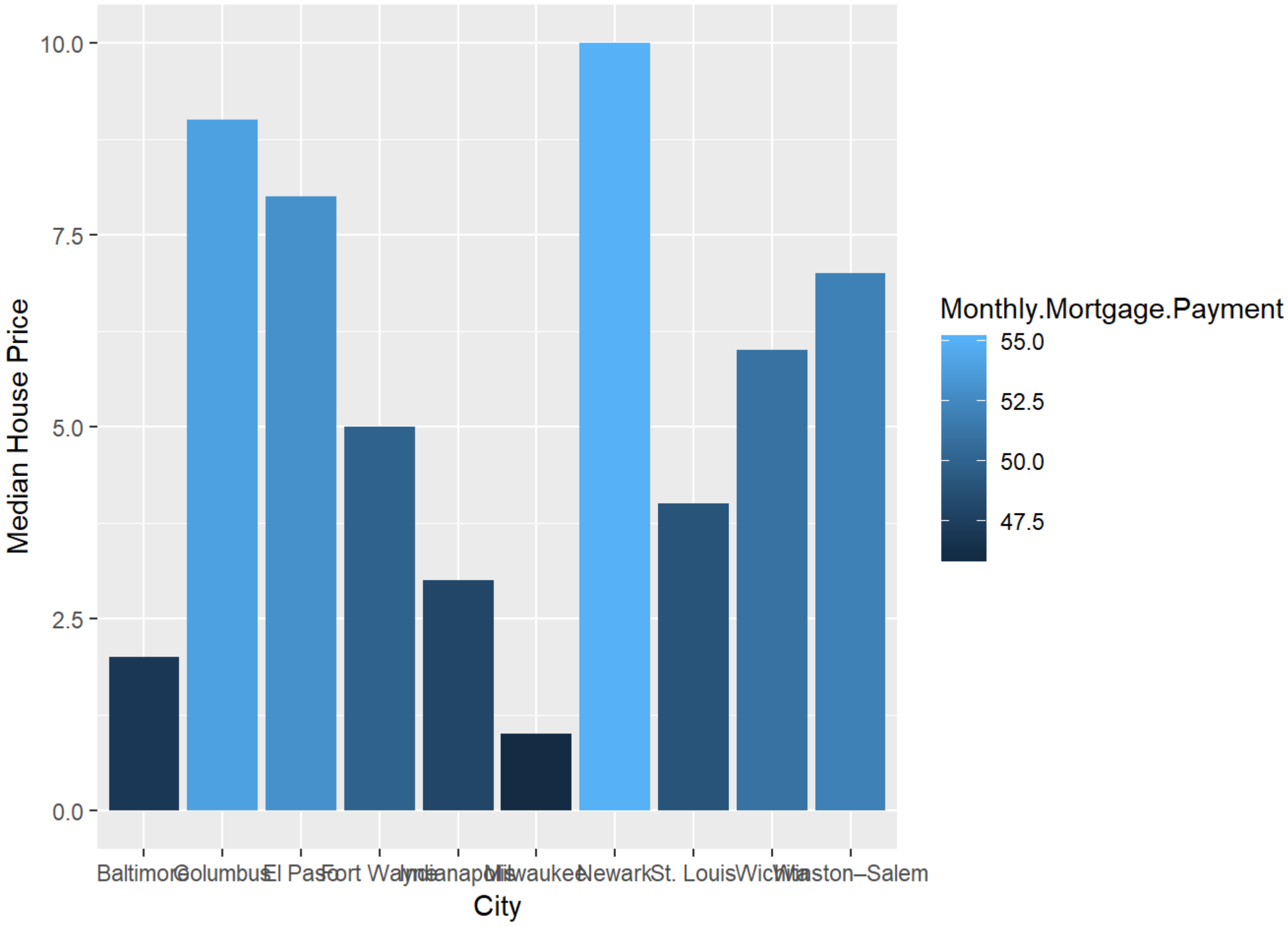
```
top_10_med_price%>%
  ggplot(aes(x=reorder(City,-Median.Price), y=Median.Price, fill =  Median.Income ))+geom_bar(st
at="identity")+xlab("City")+ylab("Median House Price")
```



```
top_10_med_price%>%
  ggplot(aes(x=reorder(City,-Present.Value), y=Median.Price, fill = Monthly.Mortgage.Payment  ))
+geom_bar(stat="identity")+xlab("City")+ylab("Median House Price")
```

```
## Warning in Ops.factor(Present.Value): '-' not meaningful for factors

## Warning in Ops.factor(Present.Value): '-' not meaningful for factors
```



Reading data in a faster way - good for big data

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.2
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
house.2 <- read.csv("C:/Users/anupr/Documents/Flex2/DAM/Hours-to-Pay-Mortgage.csv")
```

Read_csv and fread maintain white spaces. fread is fastest of all.

```
house.3 <- read_csv("C:/Users/anupr/Documents/Flex2/DAM/Hours-to-Pay-Mortgage.csv")
```

```
## Warning: Missing column names filled in: 'X10' [10]
```

```
## Parsed with column specification:
## cols(
##   City = col_character(),
##   State = col_character(),
##   `Median Home Listing Price` = col_number(),
##   `30-year Fixed Mortgage Rate` = col_character(),
##   `Monthly Mortgage Payment` = col_number(),
##   `Median Household Income` = col_number(),
##   `Hours per Month to Afford a Home` = col_double(),
##   `Number of Periods` = col_double(),
##   `Present Value` = col_number(),
##   X10 = col_logical(),
##   bin = col_double()
## )
```

```
system.time(house.1 <- fread("C:/Users/anupr/Documents/Flex2/DAM/Hours-to-Pay-Mortgage.csv"))
```

```
##    user  system elapsed 
##       0       0       0
```

exporting data in multiple sheets in excel

```
# install.packages("devtools")
'devtools::install_github("kassambara/r2excel")
library(r2excel)

multiple_df <- createWorkbook()
car_df <- createSheet(wb = multiple_df, sheetName = "Cars")
iris_df <- createSheet(wb = multiple_df, sheetName = "Iris")'
```

```
## [1] "devtools::install_github(\"kassambara/r2excel\")\nlibrary(r2excel)\n\nmultiple_df <- cre
ateWorkbook()\ncar_df <- createSheet(wb = multiple_df, sheetName = \"Cars\")\niris_df <- createS
heet(wb = multiple_df, sheetName = \"Iris\")"
```

read from database

```
'install.packages("RODBC")
require(RODBC)'
```

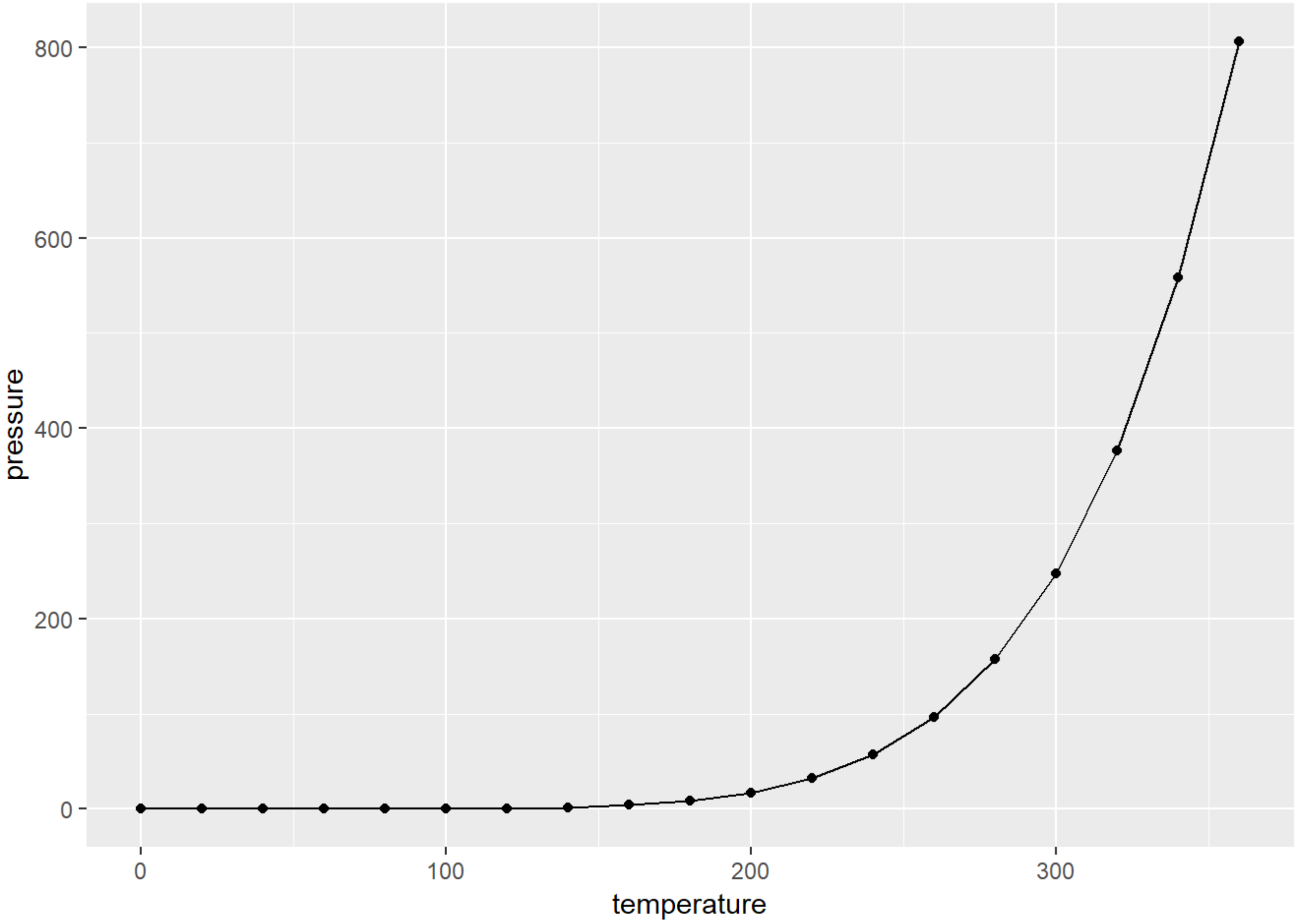
```
## [1] "install.packages(\"RODBC\")\nrequire(RODBC)"
```

Qplot - plotting 2 types at once

```
data("mpg")
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(~ f        18    29 p     comp~
## 2 audi         a4      1.8  1999     4 manua~ f        21    29 p     comp~
## 3 audi         a4      2    2008     4 manua~ f        20    31 p     comp~
## 4 audi         a4      2    2008     4 auto(~ f        21    30 p     comp~
## 5 audi         a4      2.8  1999     6 auto(~ f        16    26 p     comp~
## 6 audi         a4      2.8  1999     6 manua~ f        18    26 p     comp~
```

```
qplot(temperature, pressure, data = pressure, geom = c("line", "point"))
```



Barplot

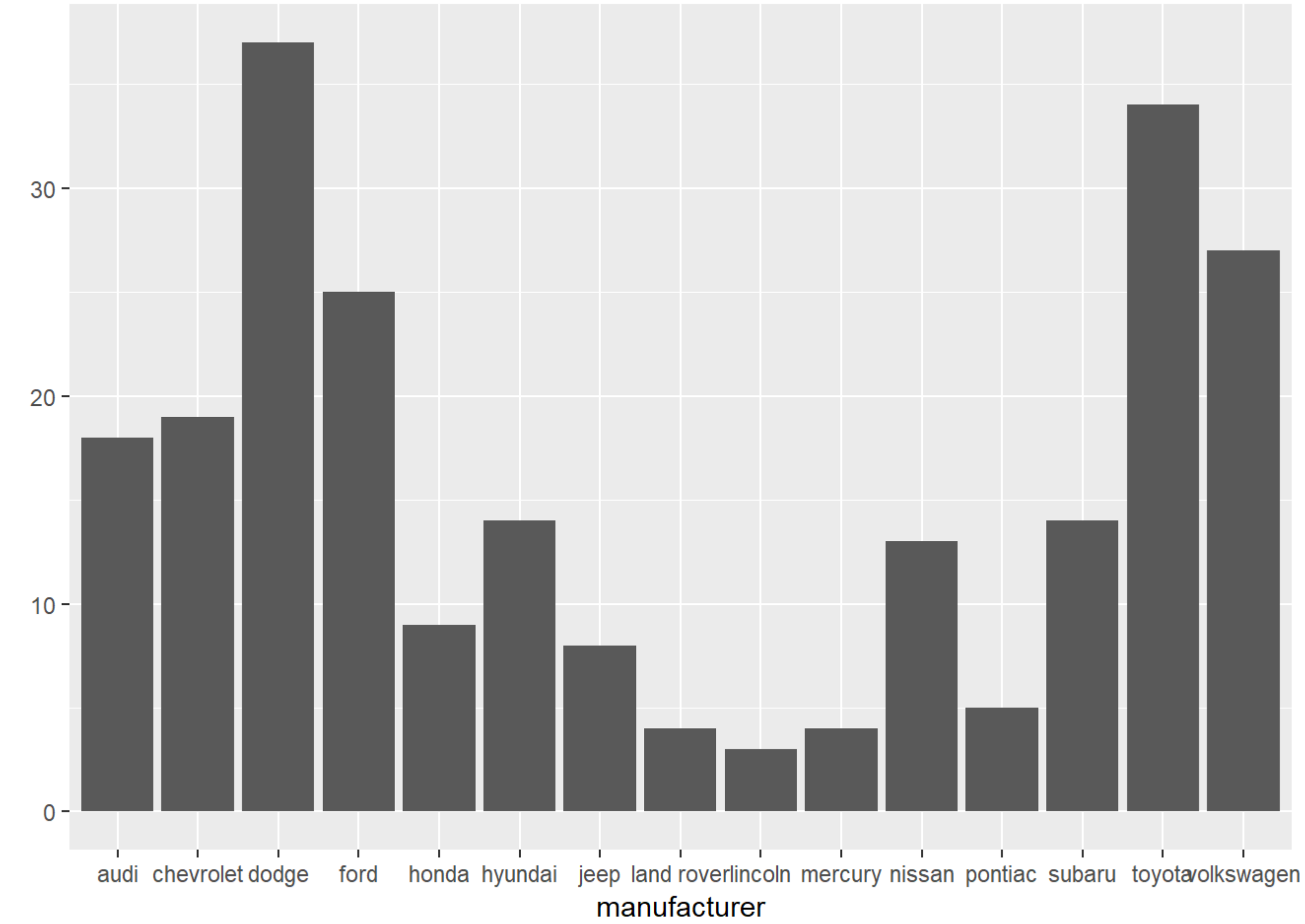
```
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(~ f        18    29 p     comp~
## 2 audi         a4      1.8  1999     4 manua~ f        21    29 p     comp~
## 3 audi         a4      2    2008     4 manua~ f        20    31 p     comp~
## 4 audi         a4      2    2008     4 auto(~ f        21    30 p     comp~
## 5 audi         a4      2.8  1999     6 auto(~ f        16    26 p     comp~
## 6 audi         a4      2.8  1999     6 manua~ f        18    26 p     comp~
```

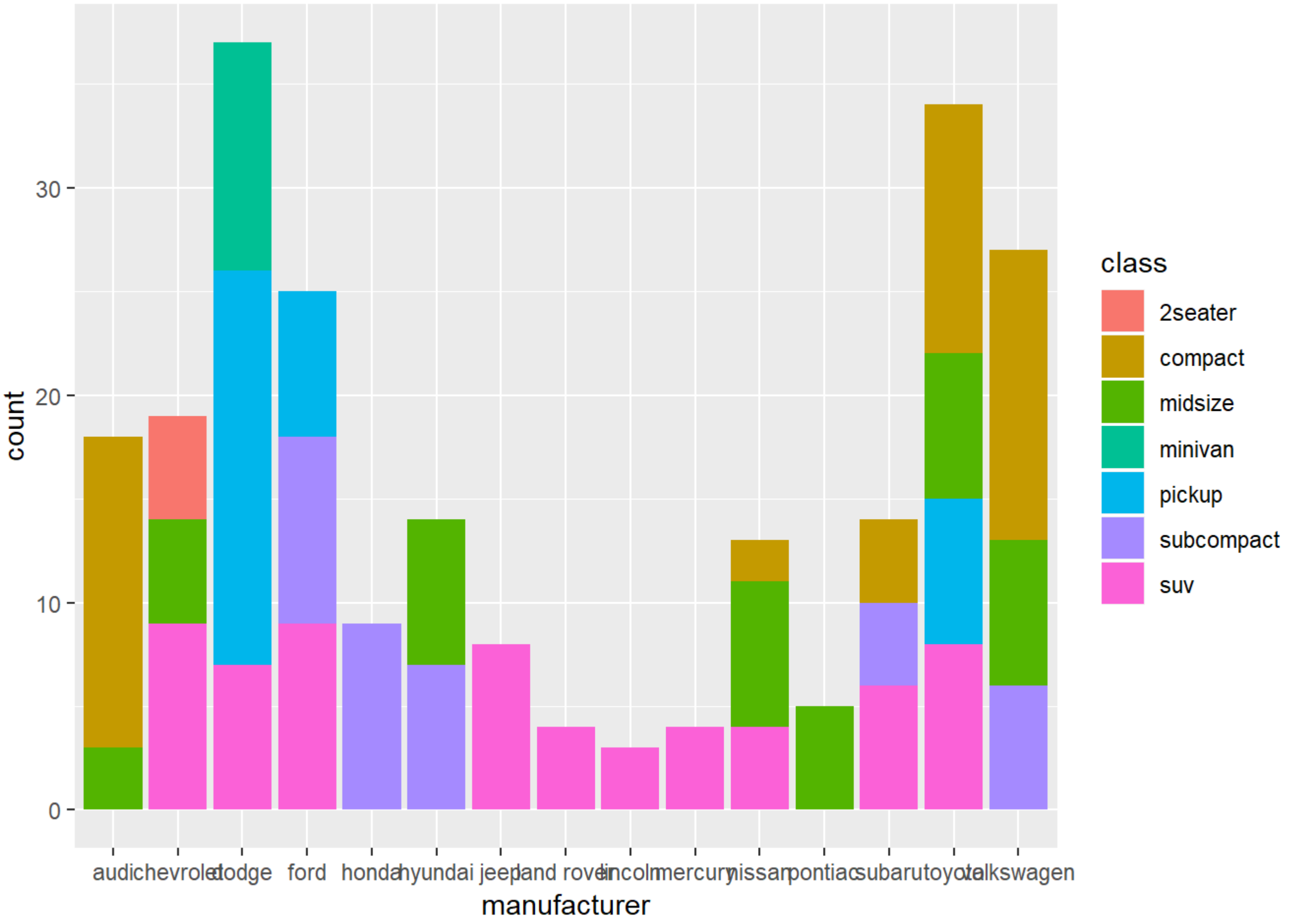
```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
qplot(data = mpg, manufacturer)
```



```
ggplot(mpg, aes(manufacturer, fill=class))+geom_bar()
```



Histogram

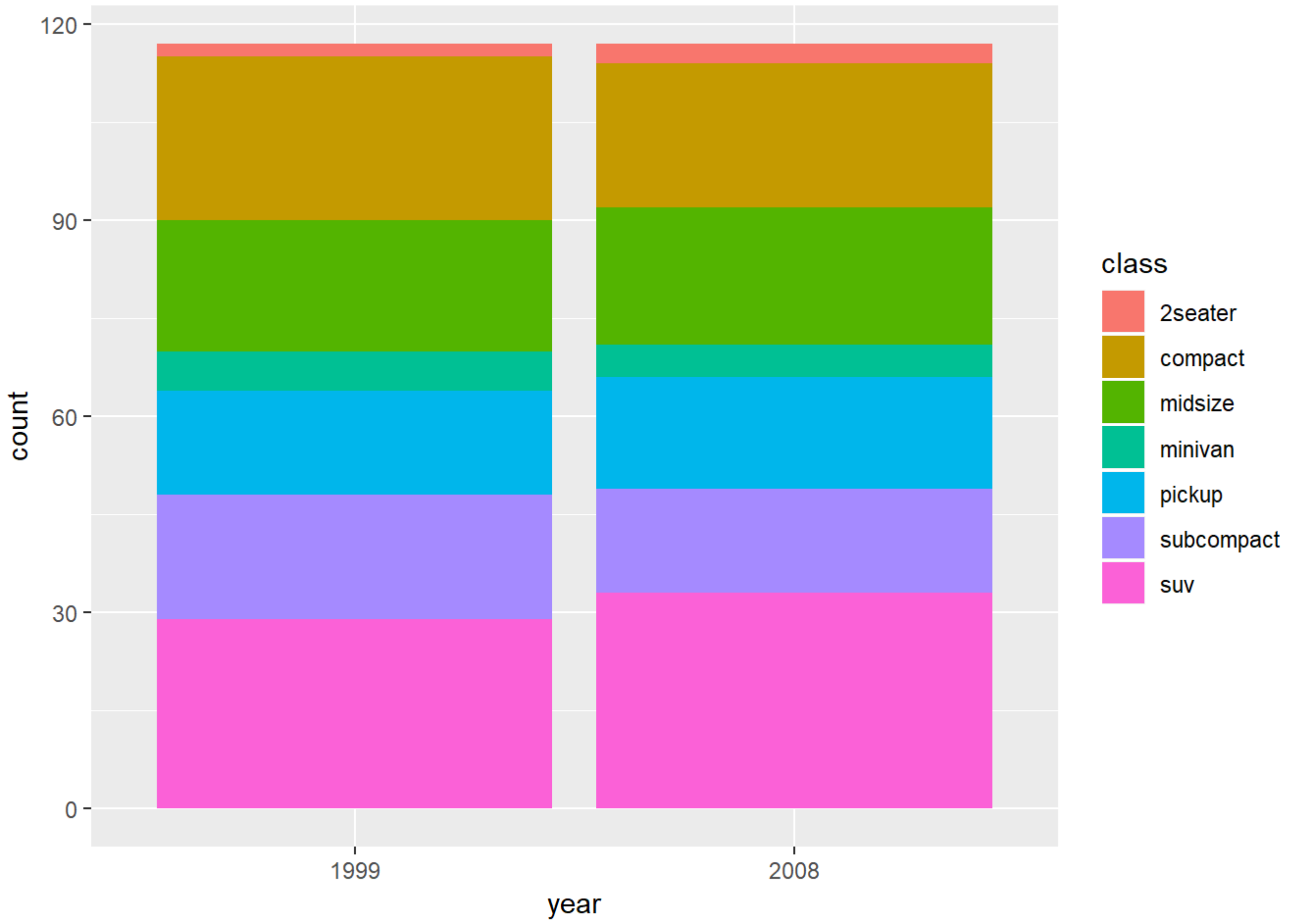
```
mpg$year <- as.factor(mpg$year)
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : Factor w/ 2 levels "1999","2008": 1 1 2 2 1 1 2 1 1 2 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
unique(mpg$year)
```

```
## [1] 1999 2008
## Levels: 1999 2008
```

```
ggplot(mpg, aes(x=year, fill=class))+geom_bar()
```



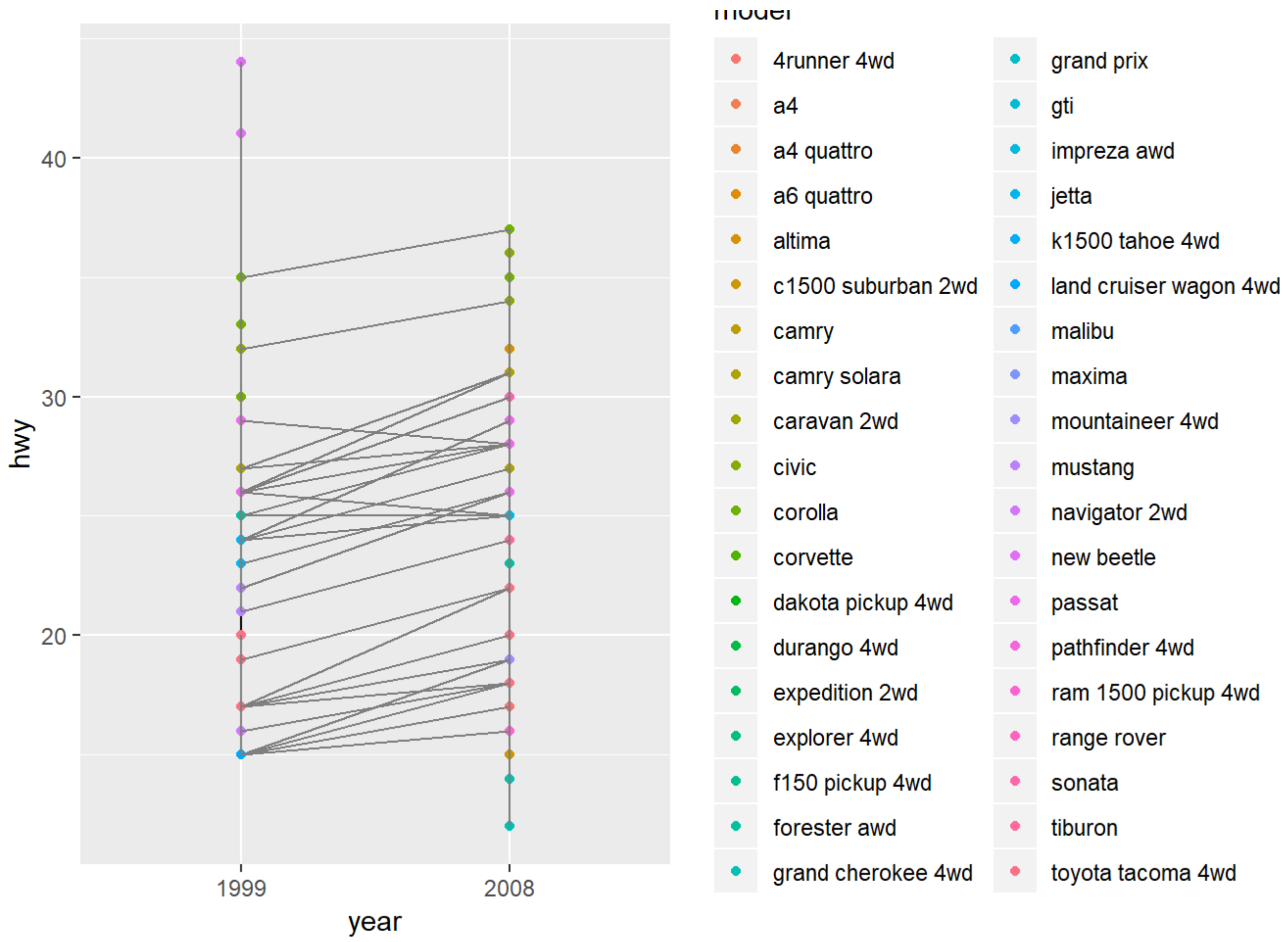
Stem and leaf plot

```
stem(mpg$displ)
```

```
##
##   The decimal point is at the |
##
##   1 | 6666688888888888888999
##   2 | 0000000000000000000002222224444444444444
##   2 | 55555555555555555555777777778888888888
##   3 | 000000001111113333333334444
##   3 | 555556677788888888999
##   4 | 00000000000000022224
##   4 | 6666666666677777777777777777
##   5 | 002222233333344444444
##   5 | 67777777799
##   6 | 0122
##   6 | 5
##   7 | 0
```

Line Chart - changes over time

```
mpg%>%
ggplot(aes(x=year, y=hwy))+ geom_line() +geom_point(aes(colour = model))+  geom_line(aes(group =
model), colour = "grey50")
```



case study - data cleaning

```
data("who")
head(who)
```

```
## # A tibble: 6 x 60
##   country iso2  iso3   year new_sp_m014 new_sp_m1524 new_sp_m2534
##   <chr>   <chr> <chr> <int>       <int>        <int>        <int>
## 1 Afghan~ AF    AFG    1980          NA           NA           NA
## 2 Afghan~ AF    AFG    1981          NA           NA           NA
## 3 Afghan~ AF    AFG    1982          NA           NA           NA
## 4 Afghan~ AF    AFG    1983          NA           NA           NA
## 5 Afghan~ AF    AFG    1984          NA           NA           NA
## 6 Afghan~ AF    AFG    1985          NA           NA           NA
## # ... with 53 more variables: new_sp_m3544 <int>, new_sp_m4554 <int>,
## #   new_sp_m5564 <int>, new_sp_m65 <int>, new_sp_f014 <int>,
## #   new_sp_f1524 <int>, new_sp_f2534 <int>, new_sp_f3544 <int>,
## #   new_sp_f4554 <int>, new_sp_f5564 <int>, new_sp_f65 <int>,
## #   new_sn_m014 <int>, new_sn_m1524 <int>, new_sn_m2534 <int>,
## #   new_sn_m3544 <int>, new_sn_m4554 <int>, new_sn_m5564 <int>,
## #   new_sn_m65 <int>, new_sn_f014 <int>, new_sn_f1524 <int>,
## #   new_sn_f2534 <int>, new_sn_f3544 <int>, new_sn_f4554 <int>,
## #   new_sn_f5564 <int>, new_sn_f65 <int>, new_ep_m014 <int>,
## #   new_ep_m1524 <int>, new_ep_m2534 <int>, new_ep_m3544 <int>,
## #   new_ep_m4554 <int>, new_ep_m5564 <int>, new_ep_m65 <int>,
## #   new_ep_f014 <int>, new_ep_f1524 <int>, new_ep_f2534 <int>,
## #   new_ep_f3544 <int>, new_ep_f4554 <int>, new_ep_f5564 <int>,
## #   new_ep_f65 <int>, newrel_m014 <int>, newrel_m1524 <int>,
## #   newrel_m2534 <int>, newrel_m3544 <int>, newrel_m4554 <int>,
## #   newrel_m5564 <int>, newrel_m65 <int>, newrel_f014 <int>,
## #   newrel_f1524 <int>, newrel_f2534 <int>, newrel_f3544 <int>,
## #   newrel_f4554 <int>, newrel_f5564 <int>, newrel_f65 <int>
```

```
dim(who)
```

```
## [1] 7240   60
```

```
str(who)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7240 obs. of  60 variables:
##  $ country     : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ iso2        : chr  "AF" "AF" "AF" "AF" ...
##  $ iso3        : chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ year        : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ new_sp_m014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_m1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_m2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_m3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_m4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_m5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_m65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sp_f65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_m65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_sn_f65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_m65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_ep_f65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_m65  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f014 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f1524: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f2534: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f3544: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f4554: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f5564: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ newrel_f65  : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
unique(who$newrel_f2534)
```

```
##   [1]    NA    34   707     0  2480   808    81   165    37   302     1
##  [12]    33 15912   280    96     7   360   134   664    46   898  5673
##  [23]    24   102   251   658    12  1079   161   626  1075   147 44985
##  [34]   346   907     2   475    35  1827    18    31     9    19  7094
##  [45]  8352    27   113   240   387   803   122    20   374     3   213
##  [56]   298   323  1170    22   208  1056   275    50  2241   246 28125
##  [67]   827   781    32    30   217     4   627    49  2005  9717   164
##  [78]   604   243    58   158  1293   512   124    76  2219  1798     8
##  [89]   351    17   153  1315   541     6  2569  1137  1007    90   487
## [100] 11994    60    42   103   141  5798   131  2554   310   940  8705
## [111]   569   388    63   785   289    11   588 41071   490   500    13
## [122]  1087    65   320   564    15   300    14   349  1136  2830  2906
## [133]   839   693    71  1617   537  2487  5157  4649
```

```
colnames(who)
```

```
##  [1] "country"      "iso2"         "iso3"         "year"        
##  [5] "new_sp_m014"  "new_sp_m1524" "new_sp_m2534" "new_sp_m3544"
##  [9] "new_sp_m4554" "new_sp_m5564" "new_sp_m65"   "new_sp_f014" 
## [13] "new_sp_f1524" "new_sp_f2534" "new_sp_f3544" "new_sp_f4554"
## [17] "new_sp_f5564" "new_sp_f65"   "new_sn_m014"  "new_sn_m1524"
## [21] "new_sn_m2534" "new_sn_m3544" "new_sn_m4554" "new_sn_m5564"
## [25] "new_sn_m65"   "new_sn_f014"  "new_sn_f1524" "new_sn_f2534"
## [29] "new_sn_f3544" "new_sn_f4554" "new_sn_f5564" "new_sn_f65"  
## [33] "new_ep_m014"  "new_ep_m1524" "new_ep_m2534" "new_ep_m3544"
## [37] "new_ep_m4554" "new_ep_m5564" "new_ep_m65"   "new_ep_f014" 
## [41] "new_ep_f1524" "new_ep_f2534" "new_ep_f3544" "new_ep_f4554"
## [45] "new_ep_f5564" "new_ep_f65"   "newrel_m014"  "newrel_m1524"
## [49] "newrel_m2534" "newrel_m3544" "newrel_m4554" "newrel_m5564"
## [53] "newrel_m65"   "newrel_f014"  "newrel_f1524" "newrel_f2534"
## [57] "newrel_f3544" "newrel_f4554" "newrel_f5564" "newrel_f65"
```

There are so many columns with similar data. We see there are some numbers in each such column. This can be count.

```
who_gather <- gather(who,new_sp_m014:newrel_f65 ,key="key",value = "cases" , na.rm = TRUE)
head(arrange(who_gather, by=desc(cases)))
```

```
## # A tibble: 6 x 6
##   country iso2  iso3   year key           cases
##   <chr>   <chr> <chr> <int> <chr>         <int>
## 1 India   IN    IND    2007 new_sn_m3544 250051
## 2 India   IN    IND    2007 new_sn_f3544 148811
## 3 China   CN    CHN    2013 newrel_m65   124476
## 4 China   CN    CHN    2013 newrel_m5564 112558
## 5 India   IN    IND    2007 new_ep_m3544 105825
## 6 India   IN    IND    2007 new_ep_f3544 101015
```

```
count(who_gather,key)
```

```
## # A tibble: 56 x 2
##    key              n
##    <chr>        <int>
##  1 new_ep_f014   1032
##  2 new_ep_f1524  1021
##  3 new_ep_f2534  1021
##  4 new_ep_f3544  1021
##  5 new_ep_f4554  1017
##  6 new_ep_f5564  1017
##  7 new_ep_f65    1014
##  8 new_ep_m014   1038
##  9 new_ep_m1524  1026
## 10 new_ep_m2534  1020
## # ... with 46 more rows
```

```
 who_sep <- separate(who_gather, key, c("new or old", "type", "sex_age"), sep='_')
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 2580 rows
## [73467, 73468, 73469, 73470, 73471, 73472, 73473, 73474, 73475, 73476,
## 73477, 73478, 73479, 73480, 73481, 73482, 73483, 73484, 73485, 73486, ...].
```

```
head(who_sep)
```

```
## # A tibble: 6 x 8
##   country     iso2  iso3   year `new or old` type  sex_age cases
##   <chr>       <chr> <chr> <int> <chr>        <chr> <chr>   <int>
## 1 Afghanistan AF    AFG    1997 new          sp    m014        0
## 2 Afghanistan AF    AFG    1998 new          sp    m014       30
## 3 Afghanistan AF    AFG    1999 new          sp    m014        8
## 4 Afghanistan AF    AFG    2000 new          sp    m014       52
## 5 Afghanistan AF    AFG    2001 new          sp    m014      129
## 6 Afghanistan AF    AFG    2002 new          sp    m014       90
```

split sex and age after the first character

```
who_sep_1 <- separate(who_sep, sex_age, c("sex","age"), sep=1)
head(who_sep_1)
```

```
## # A tibble: 6 x 9
##   country     iso2  iso3   year `new or old` type  sex   age   cases
##   <chr>       <chr> <chr> <int> <chr>        <chr> <chr> <chr> <int>
## 1 Afghanistan AF    AFG    1997 new          sp    m     014       0
## 2 Afghanistan AF    AFG    1998 new          sp    m     014      30
## 3 Afghanistan AF    AFG    1999 new          sp    m     014       8
## 4 Afghanistan AF    AFG    2000 new          sp    m     014      52
## 5 Afghanistan AF    AFG    2001 new          sp    m     014     129
## 6 Afghanistan AF    AFG    2002 new          sp    m     014      90
```

```
final_who <- select(who_sep_1, everything(), -c("iso2","iso3"))
head(final_who)
```

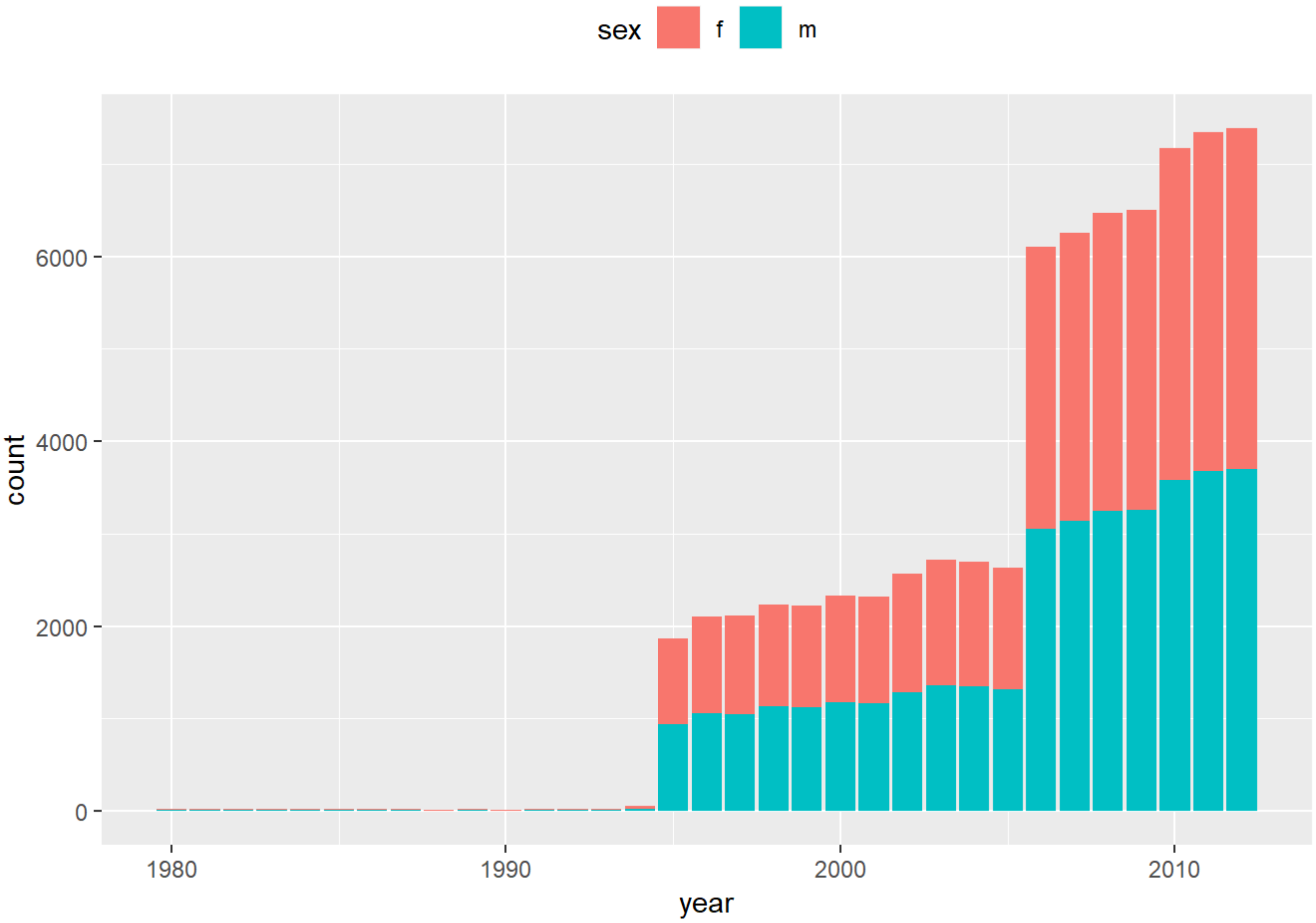
```
## # A tibble: 6 x 7
##   country      year `new or old` type  sex   age   cases
##   <chr>       <int> <chr>        <chr> <chr> <chr> <int>
## 1 Afghanistan  1997 new          sp    m     014       0
## 2 Afghanistan  1998 new          sp    m     014      30
## 3 Afghanistan  1999 new          sp    m     014       8
## 4 Afghanistan  2000 new          sp    m     014      52
## 5 Afghanistan  2001 new          sp    m     014     129
## 6 Afghanistan  2002 new          sp    m     014      90
```

```
final_who <- na.omit(final_who)
str(final_who)
```

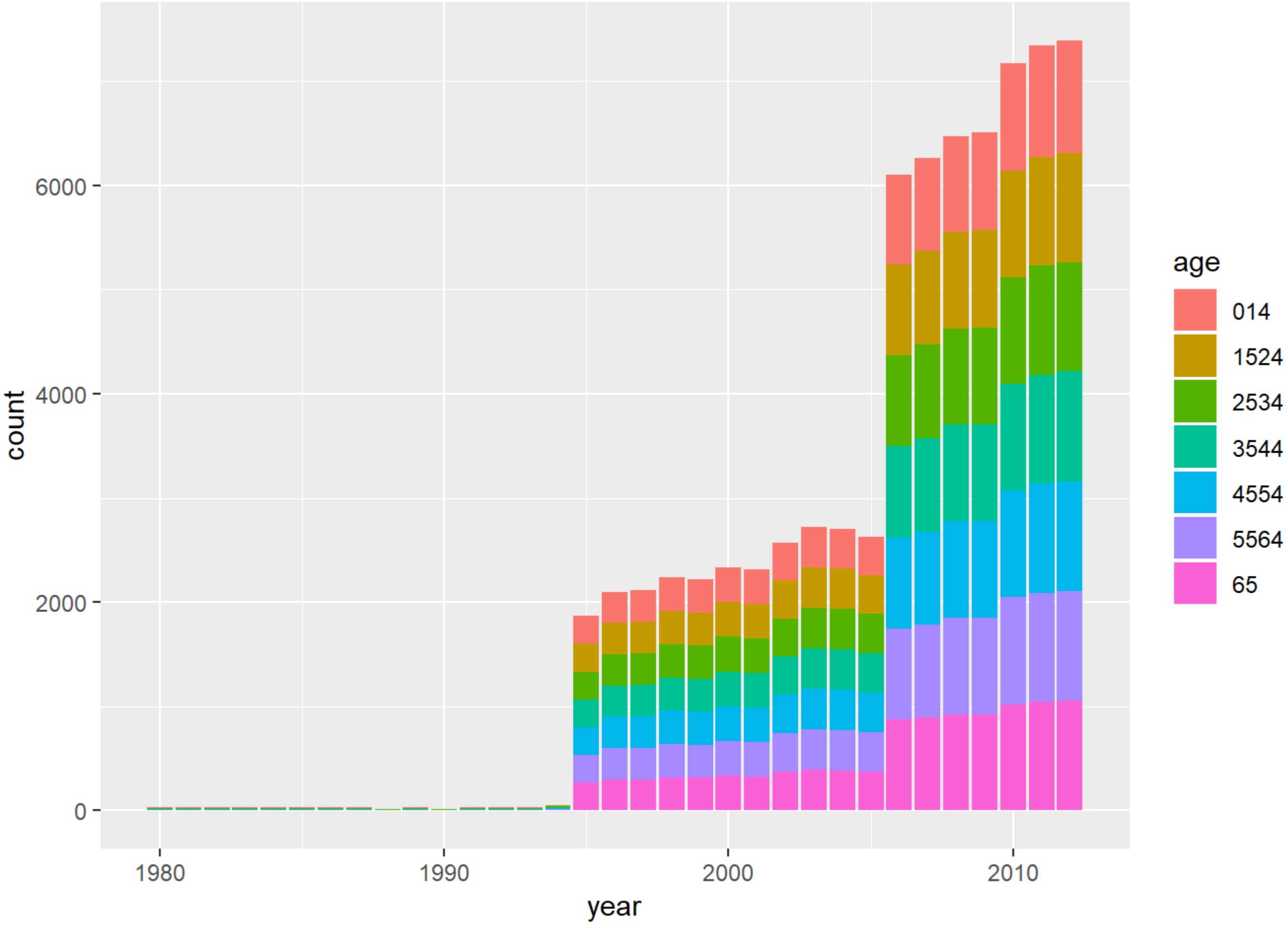
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    73466 obs. of  7 variables:
##  $ country   : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ year      : int  1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 ...
##  $ new or old: chr  "new" "new" "new" "new" ...
##  $ type      : chr  "sp" "sp" "sp" "sp" ...
##  $ sex       : chr  "m" "m" "m" "m" ...
##  $ age       : chr  "014" "014" "014" "014" ...
##  $ cases     : int  0 30 8 52 129 90 127 139 151 193 ...
##  - attr(*, "na.action")= 'omit' Named int  73467 73468 73469 73470 73471 73472 73473 73474 73
475 73476 ...
##   ..- attr(*, "names")= chr  "73467" "73468" "73469" "73470" ...
```

Visualising new who dataset

```
final_who%>%
  ggplot(aes(year, fill=sex))+geom_bar()+theme(legend.position = "top")
```

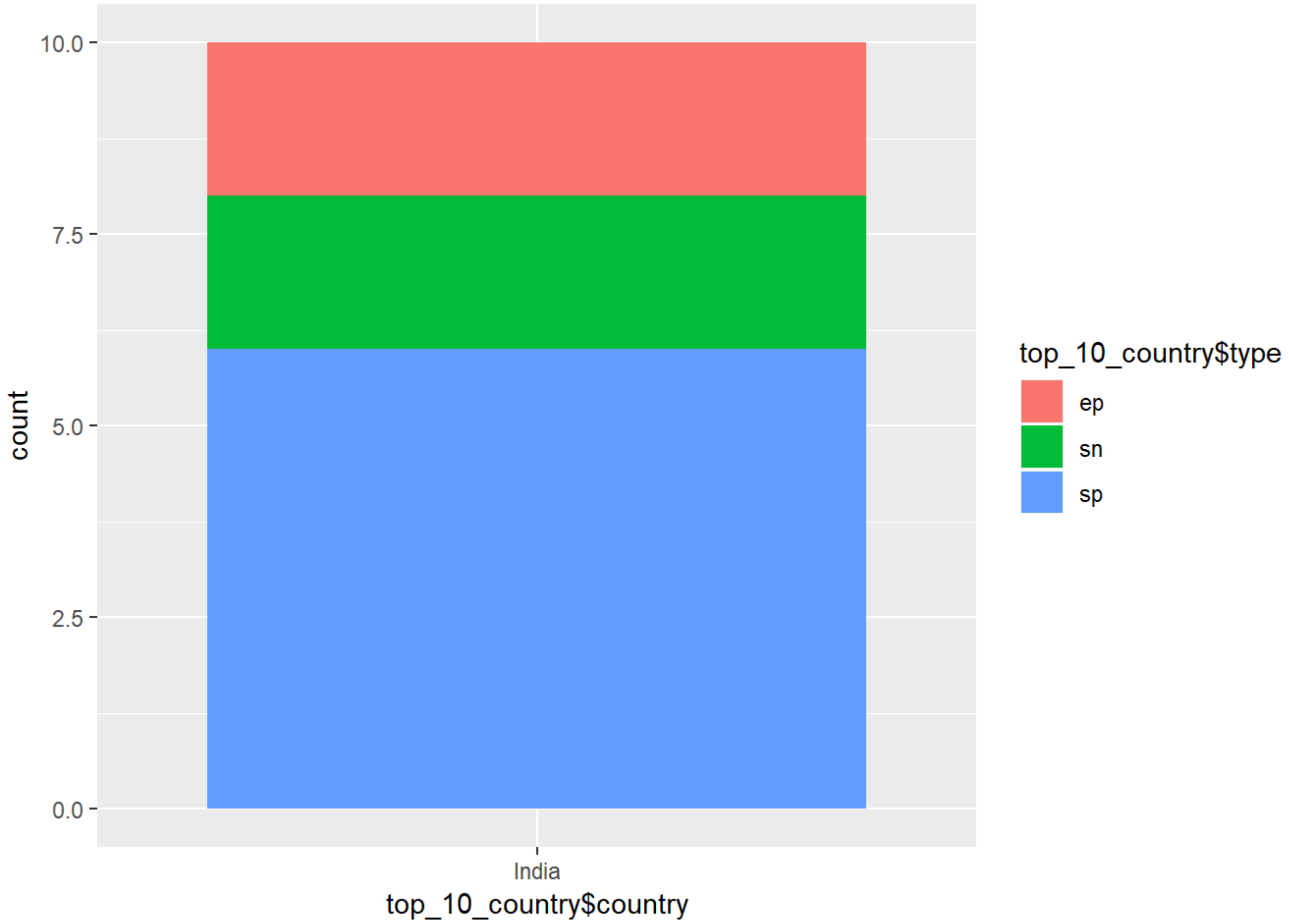


```
final_who%>%
  ggplot(aes(year, fill=age))+geom_bar()
```



Top 10 cases are from India

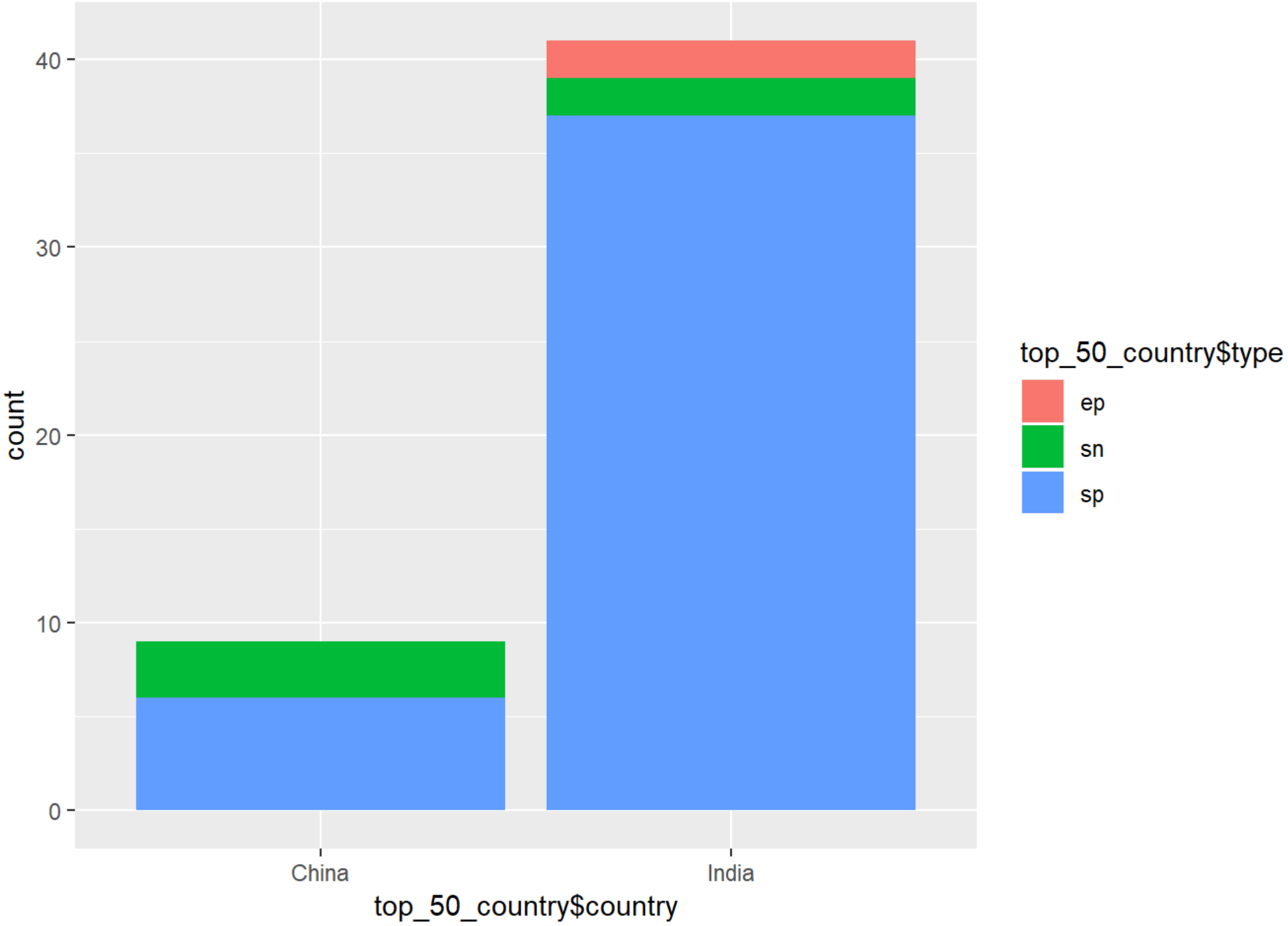
```
top_10_country <- head(arrange(final_who, desc(cases)), 10)
top_10_country%>%
  ggplot(aes(x=top_10_country$country, fill=top_10_country$type))+geom_bar()
```



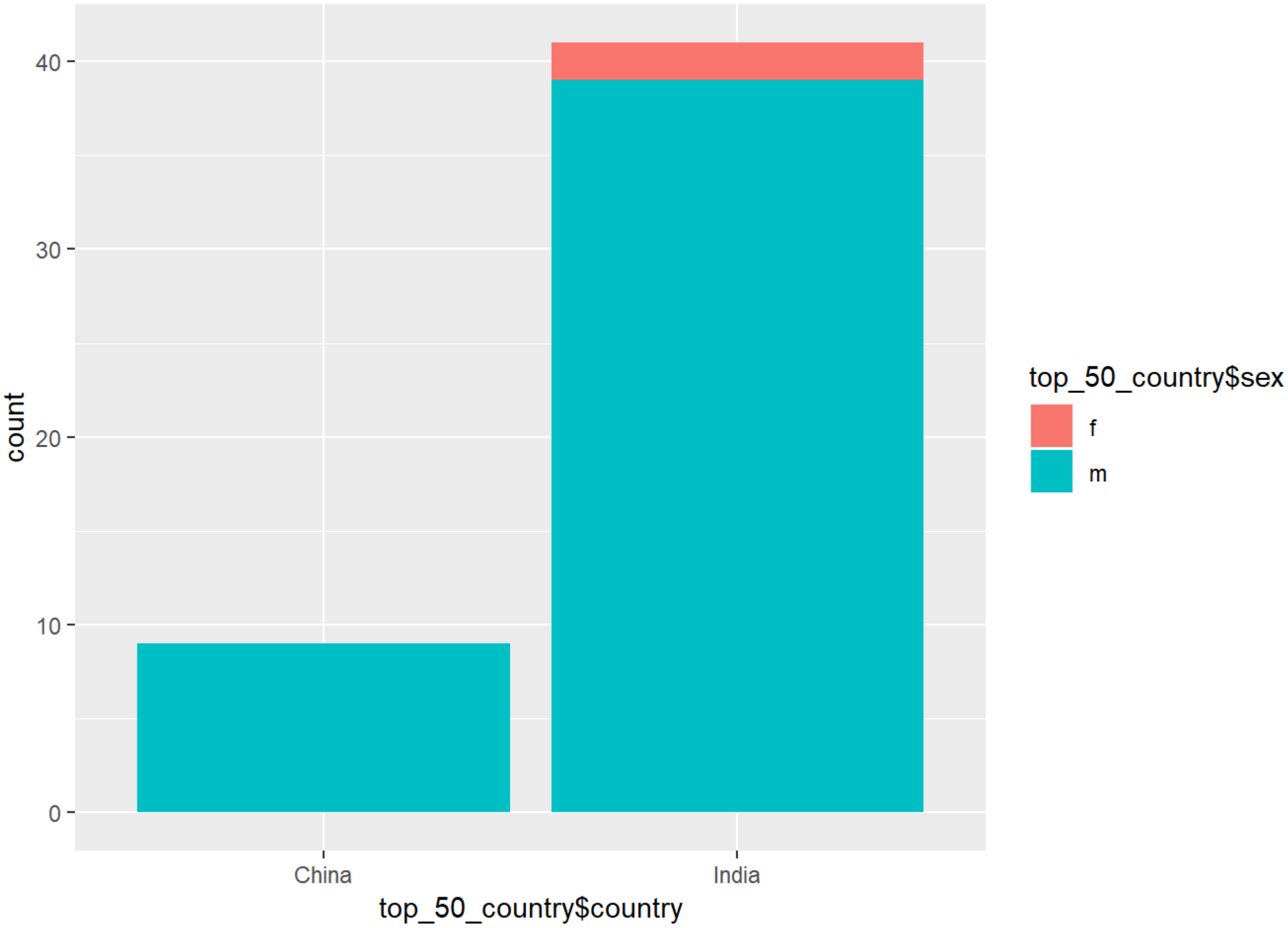
Top

50 cases are from India and China

```
top_50_country <- head(arrange(final_who, desc(cases)), 50)
top_50_country%>%
  ggplot(aes(x=top_50_country$country, fill=top_50_country$type))+geom_bar()
```

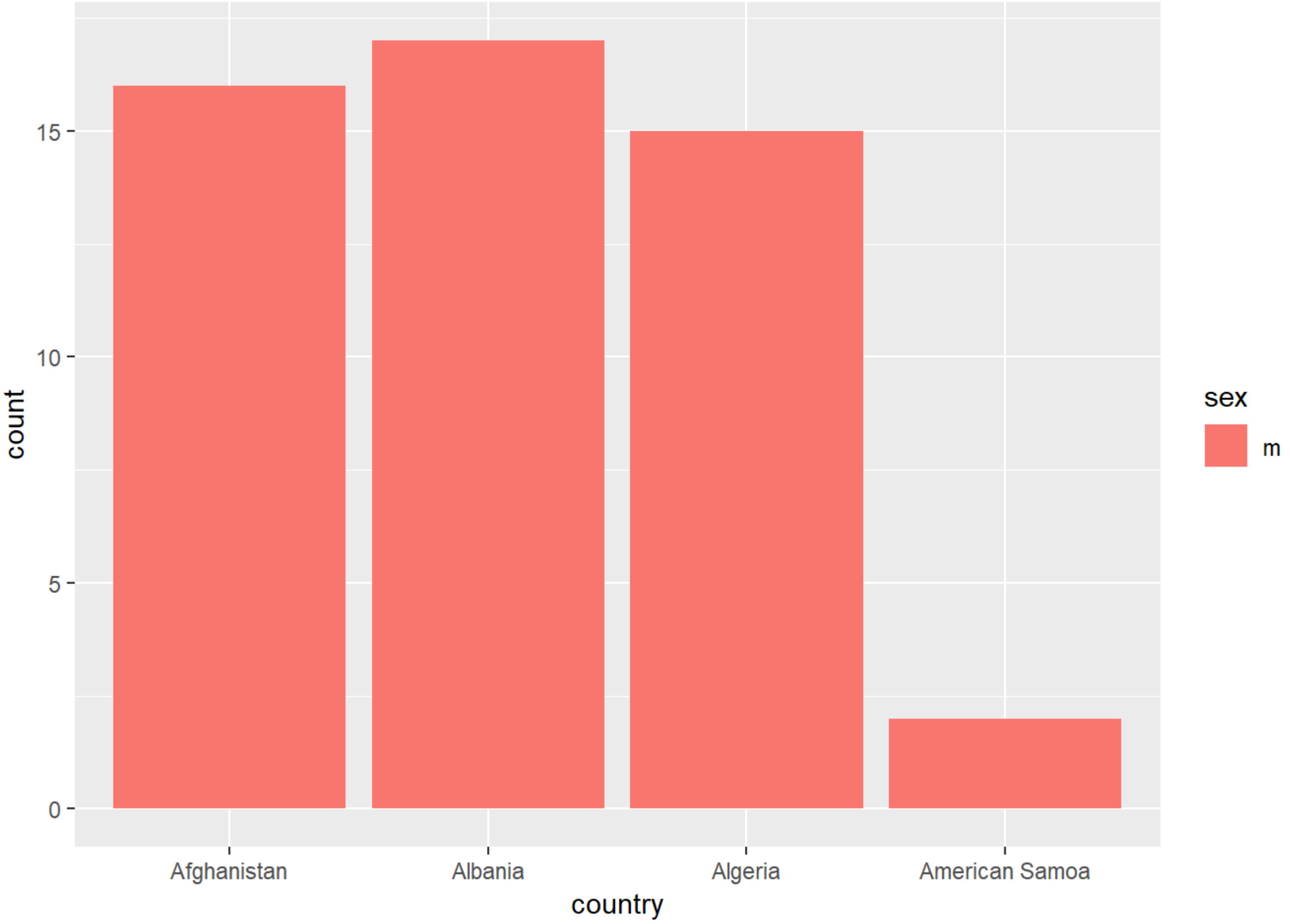


```
top_50_country <- head(arrange(final_who, desc(cases)), 50)
top_50_country%>%
  ggplot(aes(x=top_50_country$country, fill=top_50_country$sex))+geom_bar()
```



Bottom 50

```
bottom_50_country <- head(arrange(final_who), 50)
bottom_50_country%>%
  ggplot(aes(x=country, fill=sex))+geom_bar()
```


count
15
10
5
0
Afghanistan
Albania
Algeria
American Samoa
country
sex
m