HW4

2019150445/Shin Baek Rok

2020 11 20

1.

```
x<-data.frame(ID=c(1,2,3,4),grp=c('A','A','B','B'),
              sex=c('F','M','F','M'),meanL=c(0.22,0.47,0.33,0.55),
              sdL=c(0.11,0.33,0.11,0.31),meanR=c(0.34,0.57,0.40,0.65),
              sdR=c(0.08,0.33,0.07,0.27))

x[-1] %>% gather(key=key, value=value, -c(1,2)) %>%
  unite(key,sex,key,sep='.') %>% spread(key,value) %>%
  rename(ID=grp) %>% mutate(ID=c(1,2)) %>% knitr::kable()
```

ID	F.meanL	F.meanR	F.sdL	F.sdR	M.meanL	M.meanR	M.sdL	M.sdR
1	0.22	0.34	0.11	0.08	0.47	0.57	0.33	0.33
2	0.33	0.40	0.11	0.07	0.55	0.65	0.31	0.27

2.

a)

```
Marriage %>% filter(year(dob)>2000) %>% select(dob) %>% arrange(dob) %>% head()
```

```
##          dob
## 1 2024-05-21
## 2 2025-10-29
## 3 2027-03-18
## 4 2028-05-26
## 5 2030-08-03
## 6 2031-07-19
```

Of the years larger than 2000, the closest year to 2000 is 2024 and we are in 2020. So all years after 2000 are all inappropriate values.

b)

```
safe.ifelse <- function(cond, yes, no) structure(ifelse(cond, yes, no), class = class(yes))
#ifelse function can make errors when we are dealing with date.

Marriage %>%
  mutate(dob=safe.ifelse(dob>'2020-01-01',ymd(paste(year(dob)-100,month(dob),day(dob))),dob)) %>%
  select(dob) %>% arrange(dob) %>% head(10)
```

```
##           dob
## 1  1924-05-21
## 2  1925-10-29
## 3  1927-03-18
## 4  1928-05-26
## 5  1930-08-03
## 6  1931-07-19
## 7  1941-05-28
## 8  1943-02-20
## 9  1943-02-26
## 10 1944-02-28
```

3.

```
library(readxl)
data <- read_excel("C:/Users/admin/Downloads/China-Global-Investment-Tracker-2020-Spring-FINAL.xlsx",sk

# Change column names
colnames(data) <- data %>% colnames() %>%
  str_replace_all(" ","_") %>% str_to_lower()

glimpse(data)
```

```
## Rows: 1,700
## Columns: 12
## $ year                 <dbl> 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005...
## $ month                <chr> "January", "January", "February", "March", "Ap...
## $ investor             <chr> "Minmetals", "China Academy of Sciences", "Min...
## $ quantity_in_millions <dbl> 500, 1740, 550, 670, 130, 120, 100, 4200, 1420...
## $ share_size           <chr> NA, NA, NA, "0.85", "0.17", "0.4", "1", "0.67"...
## $ transaction_party    <chr> "Cubapetroleo", "IBM", "Codelco", "Highlands P...
## $ sector               <chr> "Metals", "Technology", "Metals", "Metals", "E...
## $ subsector            <chr> NA, NA, "Copper", "Steel", "Oil", "Oil", "Auto...
## $ country              <chr> "Cuba", "USA", "Chile", "Papua New Guinea", "C...
## $ region               <chr> "North America", "USA", "South America", "East...
## $ bri                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ greenfield           <chr> "G", NA, "G", "G", NA, "G", NA, NA, NA, "G", N...
```

a)

```
data %>% select(country,region) %>% group_by(country) %>%
  summarize(region=n_distinct(region)) %>% head(5)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 5 x 2
##   country             region
##   <chr>                <int>
## 1 Afghanistan              1
## 2 Angola                   1
## 3 Antigua and Barbuda      1
## 4 Argentina                1
## 5 Australia                1
```

```
data %>% select(country,region) %>% group_by(country) %>%
  summarize(region=n_distinct(region)) %>%
  filter(region>=2|region==0)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 0 x 2
## # ... with 2 variables: country <chr>, region <int>
```

```
#0 rows(Every country is only listed as belonging to a single region)
```

b)

```
table<-data %>% group_by(region, sector) %>% summarize(sum=sum(quantity_in_millions)) %>%
  mutate(prop=sum/sum(sum)) %>% select(-sum) %>% spread(region,prop)
```

```
## `summarise()` regrouping output by 'region' (override with `.groups` argument)
```

```
table[is.na(table)]<-0 #NA means zero

colnames(table)<-c('sec','NAf','Aus','EAs','Eur','NAm','SAm','SSAf','USA','WAs')
#change column names

table[,-1]<-round(table[,-1],2) #round
table
```

```
## # A tibble: 14 x 10
##    sec             NAf   Aus   EAs   Eur   NAm   SAm  SSAf   USA   WAs
##    <chr>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Agriculture    0     0.04  0.02  0.16  0.01 0.04   0.01  0.04  0.03
##  2 Chemicals      0.01  0     0     0.01  0    0.02   0.02  0.01  0
##  3 Energy         0.82  0.36  0.28  0.13  0.69 0.570  0.37  0.09  0.59
```

```
##  4 Entertainment  0     0.01  0.02  0.09  0    0      0     0.08  0
##  5 Finance        0     0.02  0.03  0.11  0    0.03   0.07  0.13  0.03
##  6 Health         0     0.06  0.01  0.02  0.02 0      0     0.04  0.01
##  7 Logistics      0.02  0     0.07  0.06  0    0.01   0     0.01  0.02
##  8 Metals         0.05  0.34  0.11  0.02  0.15 0.28   0.37  0.01  0.1
##  9 Other          0.04  0     0.07  0.03  0.01 0      0.02  0.08  0.05
## 10 Real estate    0.02  0.1   0.14  0.07  0.02 0.01   0.1   0.17  0.05
## 11 Technology     0     0     0.05  0.08  0.01 0.01   0.01  0.12  0.04
## 12 Tourism        0.01  0.01  0.04  0.04  0.04 0      0     0.11  0.01
## 13 Transport      0.03  0.05  0.15  0.17  0.04 0.04   0.04  0.12  0.07
## 14 Utilities      0     0     0.01  0.01  0    0      0     0     0
```

```
apply(table[,-1],1,mean)
```

```
##  [1] 0.038888889 0.007777778 0.433333333 0.022222222 0.046666667 0.017777778
##  [7] 0.021111111 0.158888889 0.033333333 0.075555556 0.035555556 0.028888889
## [13] 0.078888889 0.002222222
```

```
apply(table[,-1],1,mean) %>% max()
```

```
## [1] 0.4333333
```

```
#Energy sector commonly receives the great share of investment.
#But in USA, Real estate receives great share of investment.
```

c)

```
summary<-data %>% group_by(sector) %>%
  summarize(mean=mean(quantity_in_millions),sd=sd(quantity_in_millions)) %>%
  arrange(mean)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summary
```

```
## # A tibble: 14 x 3
##    sector         mean    sd
##    <chr>         <dbl> <dbl>
##  1 Other          357.  543.
##  2 Health         388.  375.
##  3 Real estate    447.  452.
##  4 Utilities      472.  452.
##  5 Transport      644. 1261.
##  6 Technology     669.  999.
##  7 Chemicals      701.  661.
##  8 Metals         725. 1160.
##  9 Tourism        746. 1143.
## 10 Entertainment  797. 1396.
## 11 Finance        804. 1303.
## 12 Energy         975. 1349.
## 13 Agriculture   1192. 5163.
## 14 Logistics     1278. 2946.
```

```r
data %>% group_by(sector) %>%
  summarize(mean=mean(log(quantity_in_millions)),sd=sd(log(quantity_in_millions))) %>%
  arrange(mean)
```

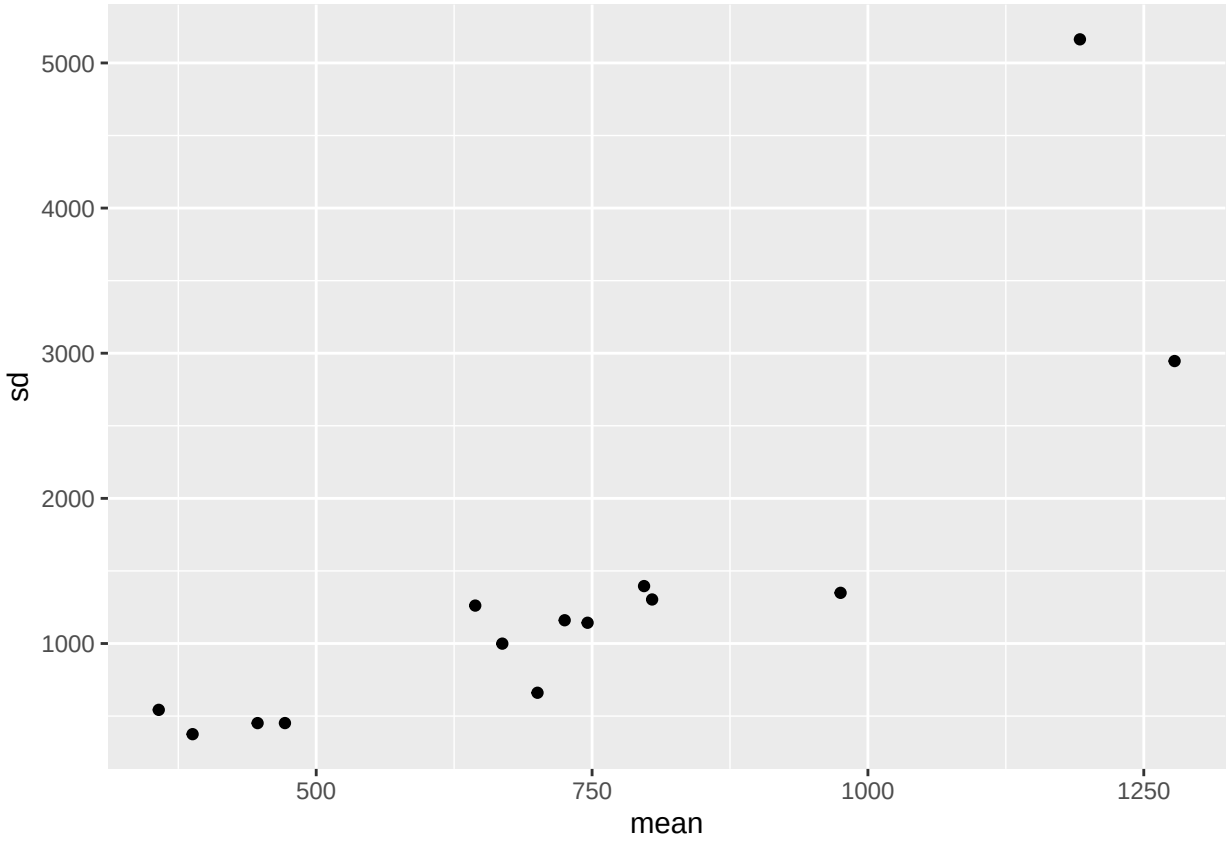
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 14 x 3
##    sector         mean    sd
##    <chr>         <dbl> <dbl>
##  1 Other          5.49 0.757
##  2 Health         5.63 0.776
##  3 Real estate    5.74 0.817
##  4 Utilities      5.76 0.919
##  5 Transport      5.81 0.990
##  6 Technology     5.86 1.04
##  7 Agriculture    5.90 1.08
##  8 Entertainment  5.91 1.12
##  9 Finance        5.98 1.13
## 10 Tourism        5.99 1.05
## 11 Logistics      6.02 1.29
## 12 Metals         6.03 0.997
## 13 Chemicals      6.07 1.05
## 14 Energy         6.25 1.12
```

```r
cor(summary$mean,summary$sd)
```

```
## [1] 0.8308919
```

```r
qplot(x=mean,y=sd,data=summary)
```




```
#There is high positive correlation between mean&sd(When mean increases, sd also increases.)
#Agriculture sector has high sd for its mean.
```

d)

```
sec_year<-data %>% group_by(sector, year) %>%
  summarize(total=sum(quantity_in_millions)) %>% spread(sector, total)
```

```
## `summarise()` regrouping output by 'sector' (override with `.groups` argument)
```

```
sec_year<-sec_year %>% replace(is.na(sec_year),0)
sec_year
```

```
## # A tibble: 16 x 15
##     year Agriculture Chemicals Energy Entertainment Finance Health Logistics
##    <dbl>       <dbl>     <dbl>  <dbl>         <dbl>   <dbl>  <dbl>     <dbl>
##  1  2005           0       180   6360             0       0      0         0
##  2  2006         480         0  10140             0     100      0         0
##  3  2007           0       700   2610             0   19370      0       150
##  4  2008         340         0  22100             0    4650    360         0
##  5  2009         370         0  34480             0    3100      0         0
##  6  2010        1580      1200  36510           100    3030    270       100
```

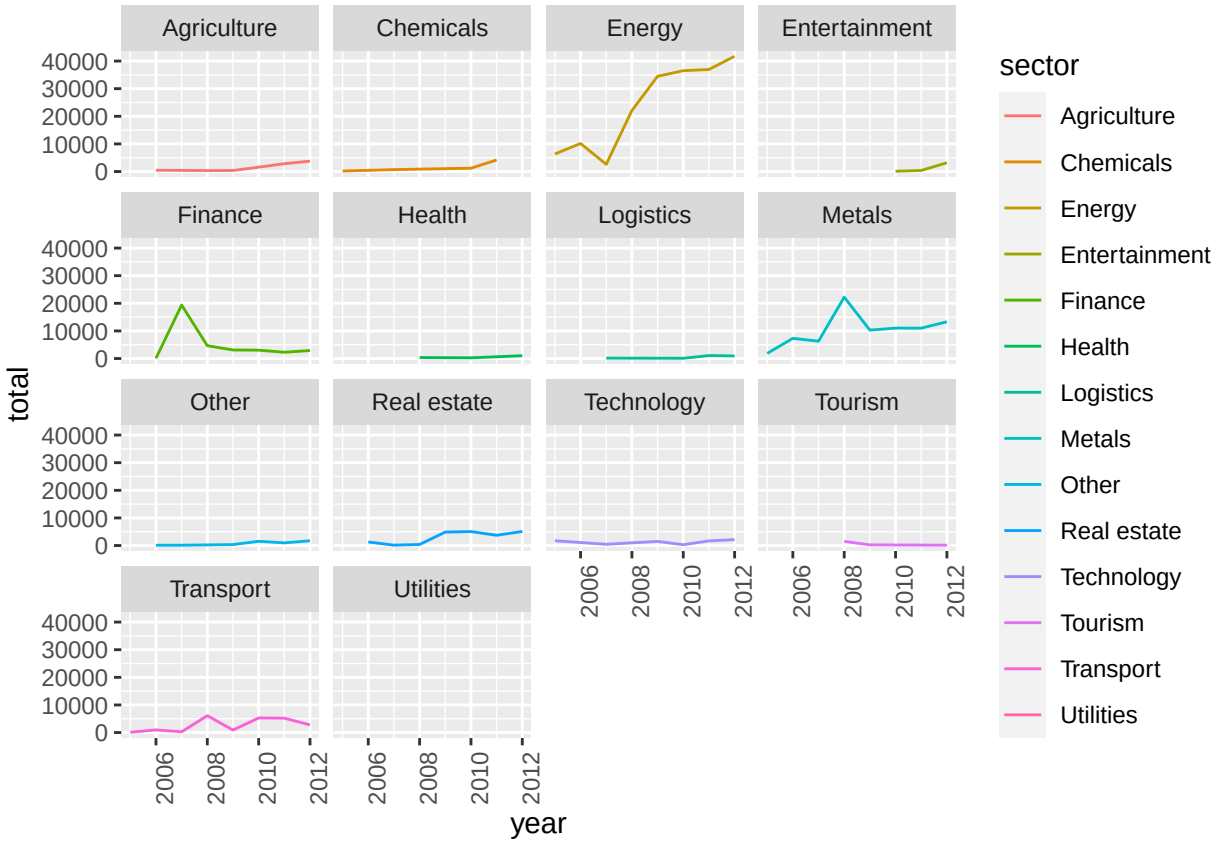
```
##  7  2011         2830       4190  36950          400     2280       0       1080
##  8  2012         3750          0  41740         3170     2900    1020        940
##  9  2013         9640       1260  35230          350     1020     980        470
## 10  2014         7030        620  29190         2110     6390     590        760
## 11  2015         1090          0  32110         3740    13360    2410       5230
## 12  2016         5610       1610  34720        22030     3080    2820       3100
## 13  2017        45870          0  20490         5950    15990    5830      23880
## 14  2018         2440       2580  28200         4710     4400    5890        250
## 15  2019         2420        270  24130         9740     1900    2270       1100
## 16  2020            0          0      0          310      480     450          0
## # ... with 7 more variables: Metals <dbl>, Other <dbl>, `Real estate` <dbl>,
## #   Technology <dbl>, Tourism <dbl>, Transport <dbl>, Utilities <dbl>
```

```
for_plot<-data %>% group_by(sector, year) %>%
  summarize(total=sum(quantity_in_millions))
```

```
## `summarise()` regrouping output by 'sector' (override with `.groups` argument)
```

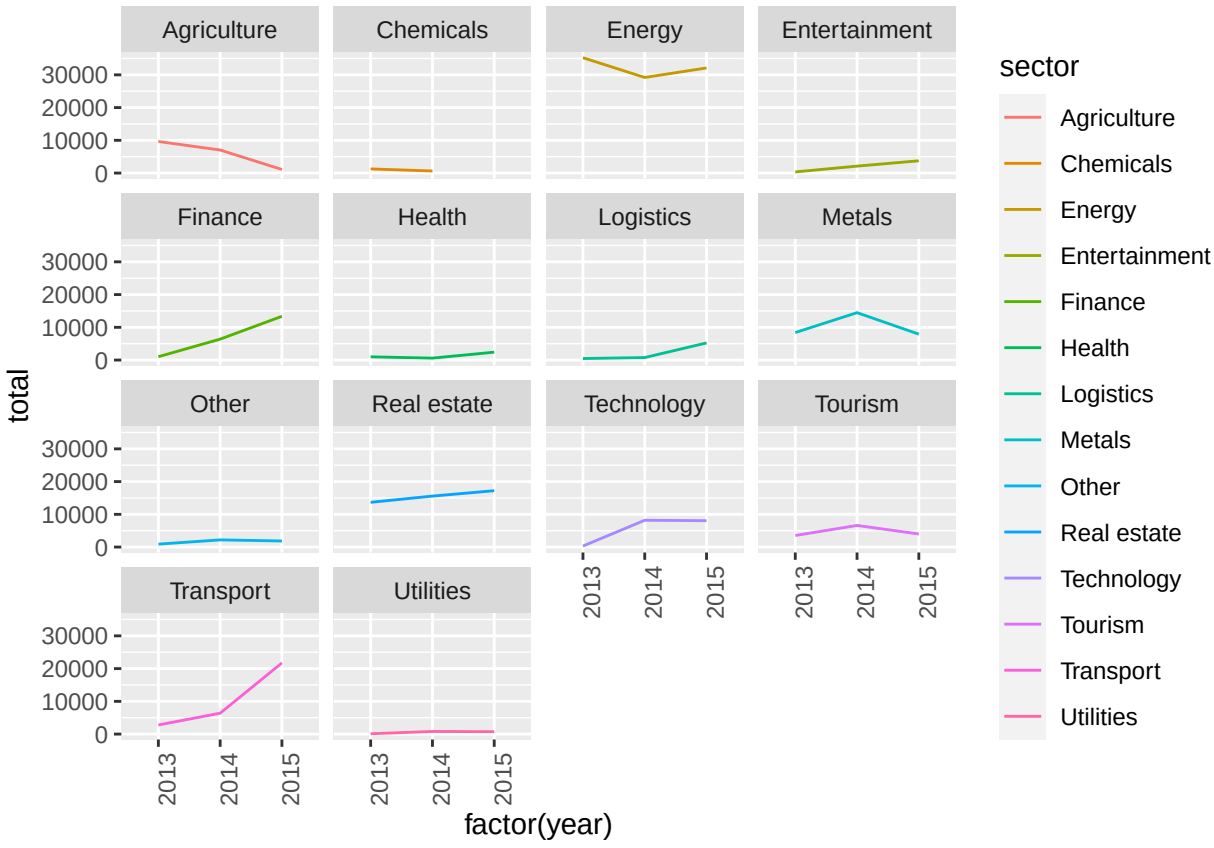
```
for_plot<-for_plot %>% replace(is.na(for_plot),0)

for_plot %>% filter(year %in% 2005:2012) %>%
  ggplot() +geom_line(aes(x=year, y=total, col=sector))+
  facet_wrap(.~sector)+theme(axis.text.x = element_text(angle = 90))
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



```
#Energy sector contributed the most to investment growth from 2005-2012
```

```
for_plot %>% filter(year %in% 2013:2015) %>%
  ggplot() +geom_line(aes(x=factor(year), y=total, col=sector,group=sector))+
  facet_wrap(.~sector)+theme(axis.text.x = element_text(angle = 90))
```




```
#Energy contributed the most to investment.Also Finance&Transport&Entertainment has growth.
```

4.

a)

```
mort<-read_csv("C:/indicatordeadkids35.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   X1 = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
mort<-mort %>% rename(country=X1)
year<-colnames(mort)[-1] %>% as.integer()
year %>% head()
```

```
## [1] 1760 1761 1762 1763 1764 1765
```

```
year %>% class()
```

```
## [1] "integer"
```

b)

```
long<-mort %>% gather(key='year',value='mortality',-1)
long$year<-as.numeric(long$year)
long %>% glimpse()
```

```
## Rows: 50,038
## Columns: 3
## $ country   <chr> "Afghanistan", "Albania", "Algeria", "Angola", "Argentina...
## $ year      <dbl> 1760, 1760, 1760, 1760, 1760, 1760, 1760, 1760, 1760, 176...
## $ mortality <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

c)

```
pop<-read_tsv('http://johnmuschelli.com/intro_to_r/data/country_pop.txt')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Rank = col_double(),
##   `Country (or dependent territory)` = col_character(),
##   Population = col_number(),
##   Date = col_character(),
##   `% of world population` = col_character(),
##   Source = col_character()
## )
```

```
pop<-pop %>% rename(country=2, percent=5)
pop %>% glimpse()
```

```
## Rows: 242
## Columns: 6
## $ Rank       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ country    <chr> "China", "India", "United States", "Indonesia", "Brazil"...
## $ Population <dbl> 1347350000, 1210193422, 315091000, 237641326, 193946886,...
## $ Date       <chr> "31-Dec-11", "1-Mar-11", "10-Jan-13", "1-May-10", "1-Jul...
## $ percent    <chr> "19.07%", "17.13%", "4.46%", "3.36%", "2.75%", "2.57%", ...
## $ Source     <chr> "Official estimate", "2011 census", "Official population...
```

d)

```
pop_levels<-pop %>% arrange(-Population) %>% select(country)
long<-long %>% mutate(sorted=factor(country, levels=pop_levels$country))
long %>% head()
```

```
## # A tibble: 6 x 4
##   country      year mortality sorted
##   <chr>       <dbl>     <dbl> <fct>
## 1 Afghanistan  1760        NA Afghanistan
## 2 Albania      1760        NA Albania
## 3 Algeria      1760        NA Algeria
## 4 Angola       1760        NA Angola
## 5 Argentina    1760        NA Argentina
## 6 Armenia      1760        NA Armenia
```

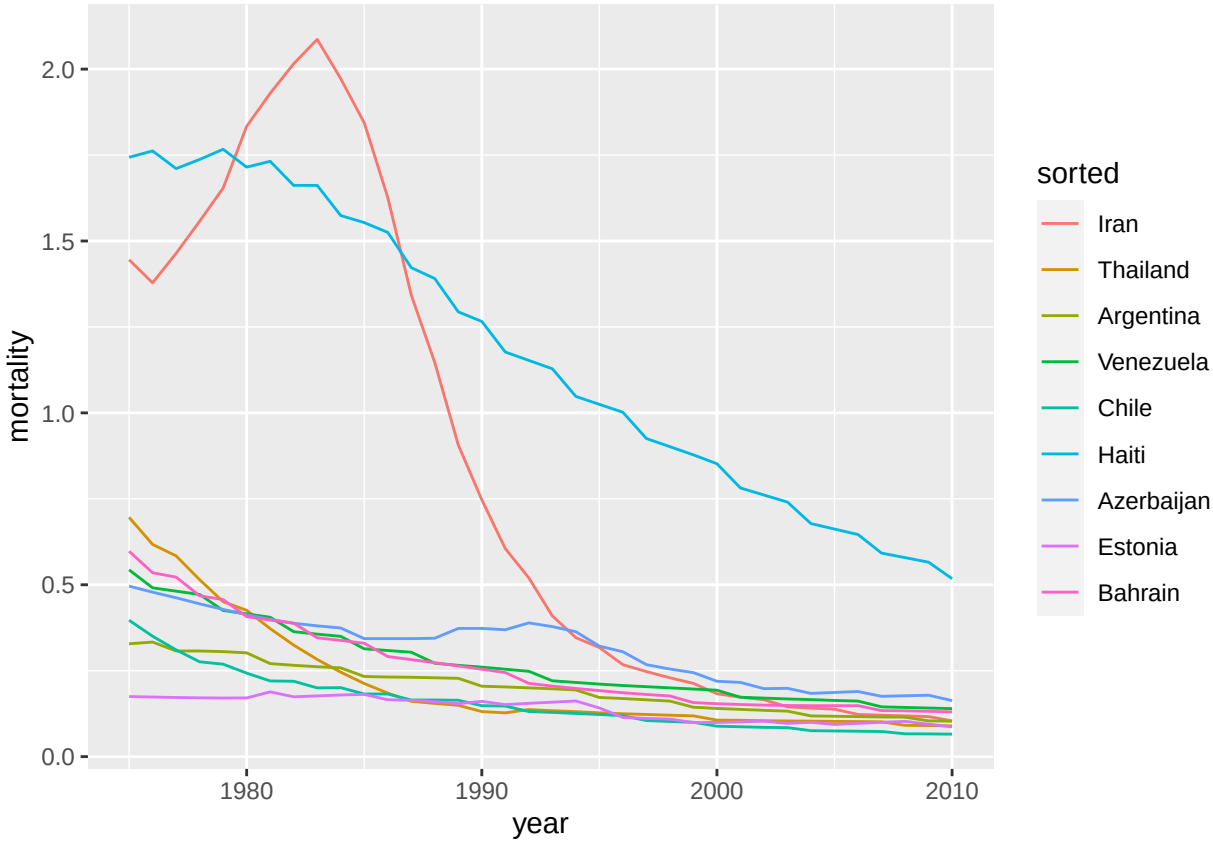
e)

```
long_sub<-long %>% filter(year>=1975 & year<=2010) %>%
  filter(sorted %in% c("Venezuela", "Bahrain", "Estonia",
                       "Iran", "Thailand", "Chile",
                       "Western Sahara", "Azerbaijan", "Argentina", "Haiti")) %>%
  filter(!is.na(mortality))
head(long_sub)
```

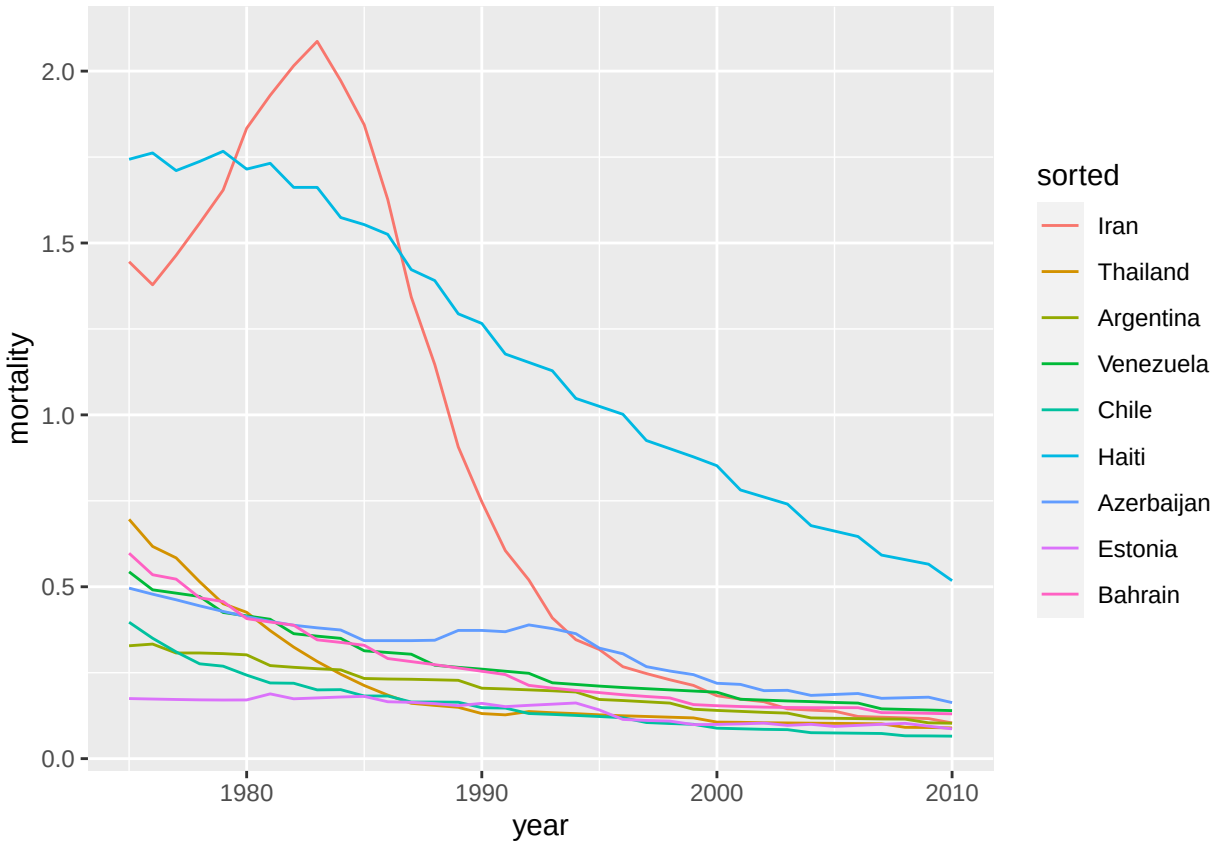
```
## # A tibble: 6 x 4
##   country     year mortality sorted
##   <chr>      <dbl>     <dbl> <fct>
## 1 Argentina   1975     0.328 Argentina
## 2 Azerbaijan  1975     0.496 Azerbaijan
## 3 Bahrain     1975     0.597 Bahrain
## 4 Chile       1975     0.397 Chile
## 5 Estonia     1975     0.175 Estonia
## 6 Haiti       1975     1.74  Haiti
```

f)

```
qplot(x=year,y=mortality,data=long_sub,col=sorted,geom='line')
```




```
long_sub %>% ggplot()+geom_line(aes(year,mortality,col=sorted))
```


2.0
1.5
1.0
0.5
0.0
mortality
1980
1990
2000
2010
year
sorted
Iran
Thailand
Argentina
Venezuela
Chile
Haiti
Azerbaijan
Estonia
Bahrain