Homework assignment #4

2019150432 임효진

November 21, 2020

1

```
library(tidyverse)
library(tidyr)
df=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))

df %>% gather(measure, value, 4:7) %>%
    mutate(measure=paste(sex,".",measure),
           ID=ifelse(grp=="A", 1, 2)) %>%
    select(ID, measure, value) %>%
    spread(measure, value)
```

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

2

(a)

```
library(mosaicData)
library(lubridate)
packageVersion("mosaicData")
```

```
## [1] '0.20.1'
```

```
Marriage %>% select(dob) %>%
    filter(year(dob)>=2000)
```

```
##            dob
## 1  2064-04-11
## 2  2064-08-06
## 3  2062-02-20
## 4  2056-05-20
## 5  2066-12-14
## 6  2062-01-31
## 7  2051-07-02
## 8  2055-02-06
## 9  2067-11-15
## 10 2054-10-30
## 11 2059-11-28
## 12 2066-01-30
## 13 2043-02-20
## 14 2048-05-19
## 15 2045-04-10
## 16 2052-11-29
## 17 2058-10-20
## 18 2060-01-06
## 19 2057-04-06
## 20 2046-09-19
## 21 2064-06-05
## 22 2056-11-26
## 23 2024-05-21
## 24 2027-03-18
## 25 2041-05-28
## 26 2043-02-26
## 27 2055-02-18
## 28 2060-09-20
## 29 2031-07-19
## 30 2059-12-20
## 31 2068-02-25
## 32 2044-04-24
## 33 2057-05-18
## 34 2053-07-22
## 35 2063-04-13
## 36 2059-06-25
## 37 2058-03-02
## 38 2047-11-16
## 39 2053-06-23
## 40 2054-09-10
## 41 2052-10-01
## 42 2059-03-29
## 43 2062-09-27
## 44 2055-12-03
## 45 2055-04-08
## 46 2055-07-17
## 47 2058-08-21
## 48 2030-08-03
## 49 2025-10-29
## 50 2044-02-28
## 51 2048-09-17
## 52 2067-06-08
## 53 2061-06-24
## 54 2028-05-26
```

The variable dob corresponds to the date of the person. However there are values that don't make sense since the year should start with 19.

(b)

```
Marriage %>% select(dob) %>%
    mutate(year=ifelse(year(dob)>=2000,
                      year(dob)-100, year(dob)),
           month=month(dob),
           day=day(dob)) %>%
    unite(dob, year, month, day, sep = "-") %>%
    mutate(dob=as.Date(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

(a)

```
library(readxl)
data=read_excel("~/Desktop/data/China-Global-Investment-Tracker-2019-Spring-FINAL.xlsx",
                skip = 5)
colnames(data) <- data %>% colnames() %>%
    str_replace_all(" ","_") %>% str_to_lower()
glimpse(data)
```

```
## Rows: 1,571
## 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, "0.5", "0.85", "0.17", "0.4", "1", "0....
## $ 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...
```

```r
df1=data %>% group_by(region) %>%
    select(region, country) %>%
    summarise(count=n_distinct(country))

sum(df1$count)
```

```
## [1] 126
```

```r
n_distinct(data$country)
```

```
## [1] 125
```

```r
identical(sum(df1$count), n_distinct(data$country))
```

```
## [1] FALSE
```

```r
# There is one country beloning to 2 regions.

data %>% group_by(country) %>%
    summarise(count=n_distinct(region)) %>%
    filter(count==2)
```

```
## # A tibble: 1 x 2
##   country   count
##   <chr>     <int>
## 1 Indonesia     2
```

```r
data %>% select(country, region) %>%
    filter(country=="Indonesia") %>%
    distinct(region)
```

```
## # A tibble: 2 x 1
##   region
##   <chr>
## 1 East Asia
## 2 West Asia
```

```r
# Indonesia belongs to 2 regions which are East Asia and West Asia
```

(b)

```
tab=xtabs(quantity_in_millions~sector+region,
          data) %>%
    prop.table(2) %>% round(3)
colnames(tab)=c("Arb&N.Afrc", "Astrl",
                "E.Asia", "Erp", "N.Amrc",
                "S.Amrc", "SS Afrc",
                "USA", "W.Asia")
tab
```

```
##                 region
## sector          Arb&N.Afrc Astrl E.Asia   Erp N.Amrc S.Amrc SS Afrc   USA
##   Agriculture        0.000 0.030  0.024 0.167  0.015  0.049   0.009 0.042
##   Chemicals          0.013 0.002  0.001 0.015  0.004  0.022   0.020 0.012
##   Energy             0.809 0.366  0.304 0.141  0.731  0.541   0.377 0.092
##   Entertainment      0.000 0.011  0.017 0.082  0.005  0.000   0.003 0.084
##   Finance            0.000 0.020  0.035 0.113  0.003  0.028   0.067 0.128
##   Health             0.000 0.064  0.002 0.018  0.018  0.000   0.000 0.036
##   Logistics          0.004 0.000  0.089 0.052  0.000  0.006   0.002 0.006
##   Metals             0.057 0.340  0.112 0.014  0.116  0.304   0.368 0.009
##   Other              0.040 0.000  0.072 0.034  0.009  0.000   0.014 0.079
##   Real estate        0.021 0.103  0.130 0.071  0.025  0.008   0.102 0.166
##   Technology         0.000 0.001  0.046 0.078  0.010  0.007   0.008 0.119
##   Tourism            0.012 0.009  0.032 0.040  0.036  0.000   0.000 0.107
##   Transport          0.044 0.051  0.125 0.165  0.028  0.034   0.030 0.122
##   Utilities          0.000 0.002  0.012 0.010  0.000  0.002   0.000 0.000
##                 region
## sector          W.Asia
##   Agriculture    0.028
##   Chemicals      0.000
##   Energy         0.622
##   Entertainment  0.002
##   Finance        0.020
##   Health         0.010
##   Logistics      0.006
##   Metals         0.102
##   Other          0.036
##   Real estate    0.053
##   Technology     0.043
##   Tourism        0.012
##   Transport      0.066
##   Utilities      0.000
```

```
apply(tab, 2, which.max)
```

```
## Arb&N.Afrc      Astrl     E.Asia        Erp     N.Amrc     S.Amrc    SS Afrc
##          3          3          3          1          3          3          3
##        USA     W.Asia
##          10           3
```

```
#Energy sector recieves the most investent.
```

Although sector energy which corresponds to the index 3 most commonly recieves the shares there are exceptions in Europe and USA.

(c)

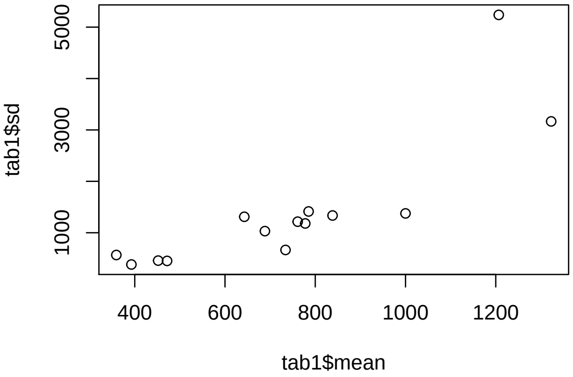
```
tab1=data %>% group_by(sector) %>%
    summarise(mean=mean(quantity_in_millions),
              sd=sd(quantity_in_millions))
tab1 %>% arrange(mean)
```

```
## # A tibble: 14 x 3
##    sector         mean    sd
##    <chr>         <dbl> <dbl>
##  1 Other          359.  567.
##  2 Health         393.  382.
##  3 Real estate    452.  458.
##  4 Utilities      472.  452.
##  5 Transport      643. 1310.
##  6 Technology     688. 1031.
##  7 Chemicals      734.  665.
##  8 Metals         761. 1215.
##  9 Tourism        778. 1180.
## 10 Entertainment  785. 1414.
## 11 Finance        838. 1336.
## 12 Energy        1000  1376.
## 13 Agriculture   1207. 5238.
## 14 Logistics     1323. 3166.
```

```
tab2=data %>% group_by(sector) %>%
    summarise(mean=mean(log10(quantity_in_millions)),
              sd=sd(log10(quantity_in_millions)))
tab2 %>% arrange(mean)
```

```
## # A tibble: 14 x 3
##    sector         mean    sd
##    <chr>         <dbl> <dbl>
##  1 Other          2.38 0.330
##  2 Health         2.45 0.337
##  3 Real estate    2.50 0.354
##  4 Utilities      2.50 0.399
##  5 Transport      2.51 0.435
##  6 Technology     2.55 0.457
##  7 Agriculture    2.56 0.475
##  8 Entertainment  2.56 0.483
##  9 Logistics      2.59 0.565
## 10 Finance          2.61 0.495
## 11 Tourism          2.62 0.464
## 12 Metals           2.63 0.446
## 13 Chemicals        2.67 0.450
## 14 Energy           2.72 0.490
```

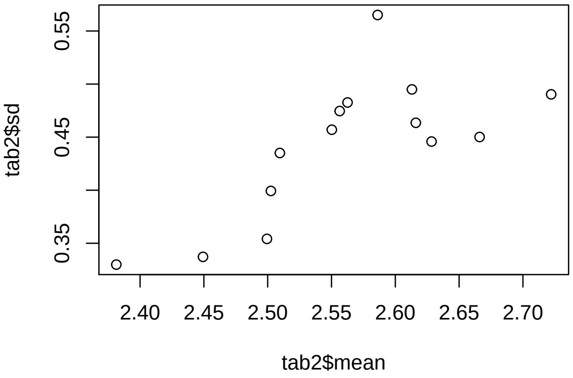
```
plot(tab1$mean, tab1$sd)
```



```
cor(tab1$mean, tab1$sd)
```

```
## [1] 0.8263623
```

```
plot(tab2$mean, tab2$sd)
```



```
cor(tab2$mean, tab2$sd)
```

```
## [1] 0.7298429
```

There seems to be a positive relation between the mean and the standard deviation. However thr chemical sector tends to hve a lower standard deviation compared to its mean.

(d)

```
tab3=xtabs(quantity_in_millions~year+sector, data)
```

Except for 2007, energy sector contributed most in both 2005~2012 and 2013~2015. However in 2013~2015 the contribution proportion of energy decreased noticeably and the amount of investment in sectors of finance, transport, and technology increased.

4

(a)

```
mort=read_csv("http://johnmuschelli.com/intro_to_r/data/indicatordeadkids35.csv")
mort %>% rename("country"=X1)
```

```
## # A tibble: 197 x 255
##    country `1760` `1761` `1762` `1763` `1764` `1765` `1766` `1767` `1768` `1769`
##    <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Afghan~     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  2 Albania     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  3 Algeria     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  4 Angola      NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  5 Argent~     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  6 Armenia     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  7 Aruba       NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  8 Austra~     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  9 Austria     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 10 Azerba~     NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
## # ... with 187 more rows, and 244 more variables: `1770` <dbl>, `1771` <dbl>,
## #   `1772` <dbl>, `1773` <dbl>, `1774` <dbl>, `1775` <dbl>, `1776` <dbl>,
## #   `1777` <dbl>, `1778` <dbl>, `1779` <dbl>, `1780` <dbl>, `1781` <dbl>,
## #   `1782` <dbl>, `1783` <dbl>, `1784` <dbl>, `1785` <dbl>, `1786` <dbl>,
## #   `1787` <dbl>, `1788` <dbl>, `1789` <dbl>, `1790` <dbl>, `1791` <dbl>,
## #   `1792` <dbl>, `1793` <dbl>, `1794` <dbl>, `1795` <dbl>, `1796` <dbl>,
## #   `1797` <dbl>, `1798` <dbl>, `1799` <dbl>, `1800` <dbl>, `1801` <dbl>,
## #   `1802` <dbl>, `1803` <dbl>, `1804` <dbl>, `1805` <dbl>, `1806` <dbl>,
## #   `1807` <dbl>, `1808` <dbl>, `1809` <dbl>, `1810` <dbl>, `1811` <dbl>,
## #   `1812` <dbl>, `1813` <dbl>, `1814` <dbl>, `1815` <dbl>, `1816` <dbl>,
## #   `1817` <dbl>, `1818` <dbl>, `1819` <dbl>, `1820` <dbl>, `1821` <dbl>,
## #   `1822` <dbl>, `1823` <dbl>, `1824` <dbl>, `1825` <dbl>, `1826` <dbl>,
## #   `1827` <dbl>, `1828` <dbl>, `1829` <dbl>, `1830` <dbl>, `1831` <dbl>,
## #   `1832` <dbl>, `1833` <dbl>, `1834` <dbl>, `1835` <dbl>, `1836` <dbl>,
## #   `1837` <dbl>, `1838` <dbl>, `1839` <dbl>, `1840` <dbl>, `1841` <dbl>,
## #   `1842` <dbl>, `1843` <dbl>, `1844` <dbl>, `1845` <dbl>, `1846` <dbl>,
## #   `1847` <dbl>, `1848` <dbl>, `1849` <dbl>, `1850` <dbl>, `1851` <dbl>,
## #   `1852` <dbl>, `1853` <dbl>, `1854` <dbl>, `1855` <dbl>, `1856` <dbl>,
## #   `1857` <dbl>, `1858` <dbl>, `1859` <dbl>, `1860` <dbl>, `1861` <dbl>,
## #   `1862` <dbl>, `1863` <dbl>, `1864` <dbl>, `1865` <dbl>, `1866` <dbl>,
## #   `1867` <dbl>, `1868` <dbl>, `1869` <dbl>, ...
```

```
year=as.integer(colnames(mort)[-1])
```

(b)

```
long=mort %>% rename("country"=X1) %>%
    gather(year, mortality, -country) %>%
    mutate(year=as.numeric(year))
```

(c)

```
pop=read_tsv("http://johnmuschelli.com/intro_to_r/data/country_pop.txt")
pop=pop %>% rename("country"=colnames(pop[2]),
           "percent"=colnames(pop[5]))
```

(d)

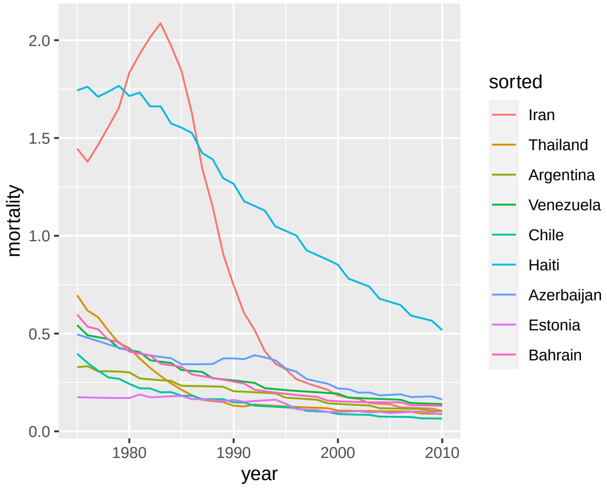
```
c1=pop %>% arrange(desc(Population))
pop_levels=c1$country
long=long %>% mutate(sorted=factor(country,
                             levels = pop_levels))
```

(e)

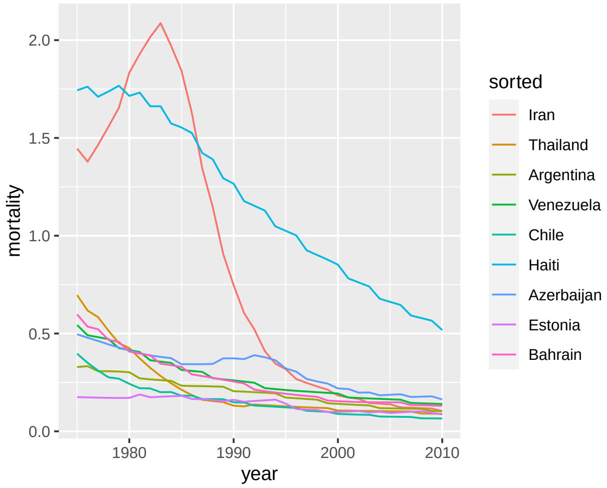
```
long_sub=long %>%
    filter(between(year, 1975, 2010),
           sorted %in% c("Venezuela", "Bahrain",
                         "Estonia", "Iran",
                         "Thailand", "Chile",
                         "Western Sahara",
                         "Azerbaijan",
                   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

(f)

```
qplot(x=year, y=mortality, data=long_sub,
      color=sorted, geom = "line")
```



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


mortality
year
sorted
Iran
Thailand
Argentina
Venezuela
Chile
Haiti
Azerbaijan
Estonia
Bahrain