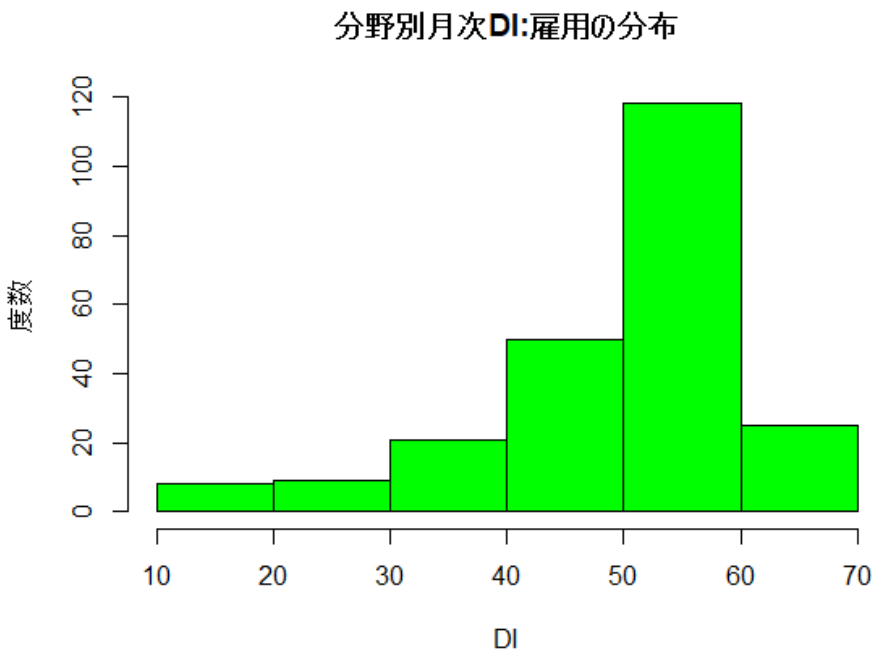
課題 1

コード

```
#-----課題1--------------------------------
#カレントディレクトリをデータがあるところに変更
setwd("C:/Users/81802/OneDrive/東工大/授業の資料/2021 1Q/確率･統計学/中間レポート")

#カンマ区切り(CSV)形式のファイルをデータフレームとして読み込み
di <- read.table("DI4domain.csv", sep=",", header=T)

#di のヒストグラム作成
hist(di$Koyou, col="green", main="分野別月次 DI:雇用の分布
", xlab="DI", ylab="度数",breaks=seq(10, 70, 10))
```

結果



課題 2

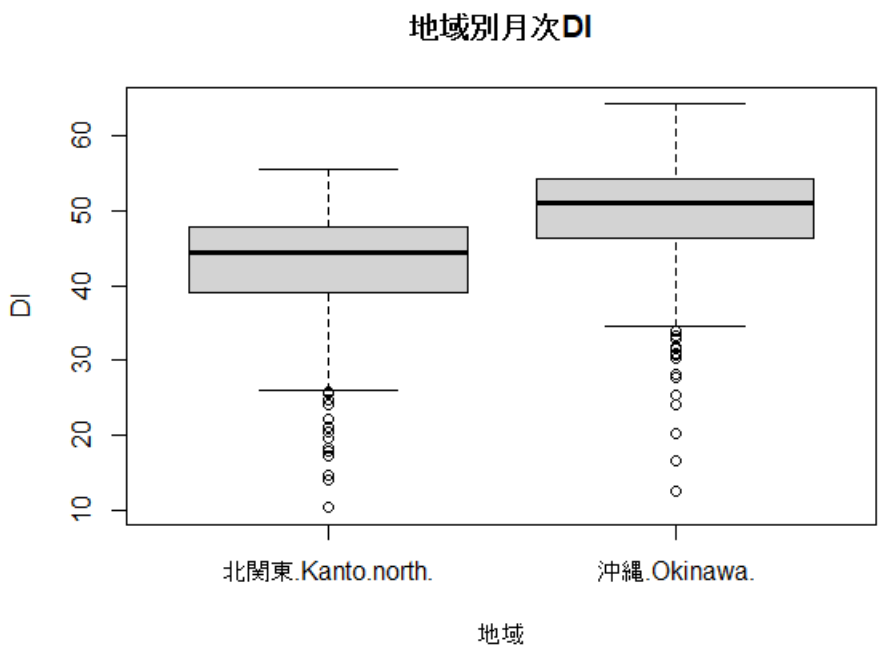
コード

```
#-----課題2-------------------------------
#カレントディレクトリをデータがあるところに変更
setwd("C:/Users/81802/OneDrive/東工大/授業の資料/2021 1Q/確率・統計学/中間レポート")

#カンマ区切り(CSV)形式のファイルをデータフレームとして読み込み
di_area <- read.table("DI4area.csv",sep=",", header=T)

#初期値
max_average <- 0 #第2列から第(x-1)列までの DI の平均値の最大値(暫定の最大値)
min_average <- 100#第2列から第(x-1)列までの DI の平均値の最小値(暫定の最小値)
max_row <- 0       #暫定の最大値がある列
min_row <- 0       #暫定の最小値がある列

#第2列から第16列まで一列ずつ処理していく
for(x in 2:16){
```

```
  #第 x 列の DI の平均値が暫定の最大値より大きいとき
  if(mean(di_area[,x])>max_average){
    max_average <- mean(di_area[,x])#暫定の最大値を書き換え
    max_row <- names(di_area)[x]    #暫定の最大値がある列の書き換え
  }

  #第 x 列の DI の平均値が暫定の最小値より小さいとき
  if(mean(di_area[,x])<min_average){
    min_average <- mean(di_area[,x])#暫定の最小値を書き換え
    min_row <- names(di_area)[x]#暫定の最小値がある列の書き換え
  }
}

#期間内の DI の平均値が最も高い地域と、最も低い地域を表示
print(max_row)
print(min_row)

#箱ひげ図を作成
boxplot(di_area[[min_row]], di_area[[max_row]],names=c(min_row,max_row),
main="地域別月次 DI", xlab="地域", ylab="DI")
```

結果

[1] "沖縄.Okinawa."
[1]”北関東.Kanto.north."



課題 3

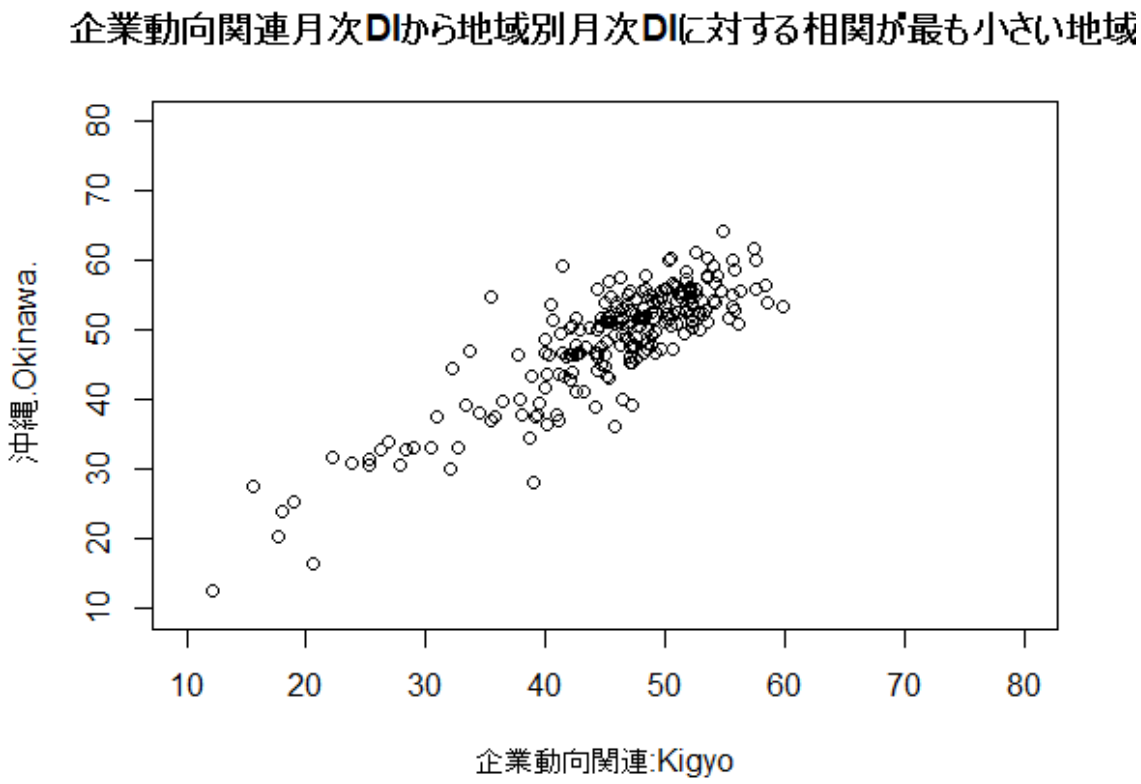
コード

```
#-----課題3-----------------------------------------------
#カレントディレクトリをデータがあるところに変更
setwd("C:/Users/81802/OneDrive/東工大/授業の資料/2021 1Q/確率･統計学/中間レポート")

#カンマ区切り(CSV)形式のファイルをデータフレームとして読み込み
di_area <- read.table("DI4area.csv",sep=",", header=T)
di <- read.table("DI4domain.csv", sep=",", header=T)
```

```
#初期値
min_cor <- 1 #企業動向関連月次DIから第2列から第(x-1)列までのそれぞれの月次DIに対するそれぞれの相関係数の最小値(暫定の相関係数の最小値)
min_cor_di_area <- "error"#暫定の相関係数の最小値がある列の名前

#第2列から最後の列まで1行ずつ処理を行う
for(x in 2:ncol(di_area)){

  #第x列との相関係数が暫定の相関係数の最小値より小さいとき
  if(cor(di$Kigyo,di_area[,x]) < min_cor){
    min_cor <- cor(di$Kigyo,di_area[,x])#暫定の相関係数の最小値を書き換え
    min_cor_di_area <- names(di_area)[x]#暫定の相関係数の最小値がある列の書き換え
  }
}

#最も相関が低いものを表示
print(min_cor_di_area)

#散布図を描画
plot(di$Kigyo, di_area[[min_cor_di_area]], xlim=c(10,80), ylim=c(10,80), main="企業動向関連月次DIから地域別月次DIに対する相関が最も小さい地域", xlab="企業動向関連:Kigyo", ylab=min_cor_di_area)
```

結果

[1] "沖縄.Okinawa."



課題 4

コード

```
#----------課題 4-----------------------------------------------------
target_file <- "第三次産業.csv"
#---------------------------------------------
#カレントディレクトリをデータがあるところに変更
setwd("C:/Users/81802/OneDrive/東工大/授業の資料/2021 1Q/確率・統計学/中間レポート")

#カンマ区切り(CSV)形式のファイルをデータフレームとして読み込み
di <- read.table("DI4domain.csv", sep=",", header=T)
target_data <- read.table(target_file, sep=",", header=T)  #60*element_count
```

```r
#両方のデータにデータがある月のみ残す
merged_data <- merge(di, target_data)
di_improved <- merged_data[,c(1:ncol(di))]
target_data <- merged_data[,c(1, c((ncol(di)+1):ncol(merged_data)))]

#target_data の列の数
element_count <- ncol(target_data)
#----------------------------------------------
#1～element_count-1 まで
rank <- 1
#----------------------------------------------
#要素がある列の数ぶんの零ベクトル
correlation_vector <- numeric(element_count-1)

#第2列から target_data の最後の列まで1行ずつ処理を行う
for(x in 2:element_count){

  #target_data の第x列の標準偏差が 0 でないとき
  if(var(target_data[,x]) != 0){

    #DI_Total と correlation_vector の x 列目の相関係数
    correlation_vector[x-1] <- cor(di_improved[,2],target_data[,x])
  }
}

#相関係数の低い列から順に並べて、第何列かを correlation_list に記述していく
correlation_list <- sort.list(correlation_vector) + 1  #target_data の第2列から最後の列までの並べ替えであるため、すべての列が 1 小さいことに注意

#相関係数の高い列から順に並べてその列を correlation_list に記述していく
correlation_list <- rev(correlation_list)

#相関係数を降順に並べ替え[正→負]
correlation_vector <- rev(sort(correlation_vector))
```

```
#rank 番目の相関係数とその項目を代入
correlation_rank <- correlation_vector[rank]
correlation_rank_target_data <- names(target_data)[correlation_list[rank]
]

#相関係数が最も高い順位の相関係数とその項目を表示
print(correlation_rank)
print(correlation_rank_target_data)

#折れ線グラフ作成
merged_data <- merge(di_improved[,c(1,2)], target_data[,c(1,correlation_l
ist[rank])], all=T)  #DI_Total と target_data の相関係数一位の列をくっつける
default_mai <-par()   #デフォルトの値を保存
mai <- par()$mai        #グラフパラメータの設定
mai[4] <- mai[1]        #余白サイズの設定(上下と左右の幅を揃える)
par(mai = mai)          #指定した余白サイズの適用

#target_data の相関関係が高い列の時系列の折れ線グラフ表示
plot(merged_data[,3], type="l", xlab=names(merged_data)[1],ylab=names(mer
ged_data)[3], main="全国 DI に最も相関のあるデータ")

#二つのグラフを同時に表示
par(new=T)
plot(merged_data[,2], type="l", xlab="", ylab="", col="red", axes=FALSE,
ylim= c(10,70))   #DI_Total の時系列の折れ線グラフ
#2 軸目の表示
axis(side=4)    #目盛りを右に表示
# 2 軸目のラベル設定
mtext(names(merged_data)[2],side=4,line=2)   #(ラベル名，ラベルの置く位置，グ
ラフの枠からの距離)
par(mai = default_mai$mai)  #デフォルトの値に戻す
```

用いたデータ:

サービス産業動向調査:2013 年 1 月:月次調査:月次:事業活動の産業（中分類）別売上高（月次）【2013 年 1 月～】

第三次産業.csv は、このデータベースを DI4domain.csv と merge できるように、その csv

の日付を 1 行目につけて整理したもの

相関係数の高いデータ：

X80 娯楽業（相関係数：0.7339278）

直接相関か、因果関係か？

近年娯楽と観光に関連する業種はインバウンド需要もあり、産業の規模は年々大きくなっているが、しかし、日本は基本的には自動車産業が主軸であり、娯楽と観光に関連する業種は国の景気に大きな影響を与えるほどの規模はなく、娯楽と観光に関連する業種の売り上げを原因として DI が影響を受ける可能性は低く、景気から娯楽業の売り上げに対する因果があるかについて考える。

まず、約 0.73 という大きな相関は因果関係を示すものの一つとなる。

また、折れ線グラフを見たときに、一番右側のコロナによる大きな落ち込みや、およそ４０か月から４５か月のあたりで顕著に表れるが、時系列でみたときに、景気が落ちると娯楽業の売り上げが落ち、景気が良くなるとそれに付随して娯楽業の売り上げが上がるという少しの時間のラグを見つけられ、この時間の差は因果関係を示すものの一つとなる。

データの存在する２０１３年１月からの８年にわたる情報を用いたことで、一時的なトレンドではなく、ある程度普遍的にこの相関が成り立つということがわかる。また、補助のデータを見ればわかるように、サービス産業の中で、景気と相関の高いデータの上位５つは、娯楽・観光業とそれに関係する旅客運送業や宿泊業が占めている。これは、景気との相関が娯楽業に特異的に成り立つのではなく、娯楽・観光関連業全体に成り立つということを示しており、娯楽・観光関連業の売り上げは、景気が大きな要因となっていることが読み取れる。

最後に、家計に余裕がなくなったときに、どの部分のお金を削ることが出来るかと考えたときに、一番削りやすいのは娯楽費であると考えることに無理はなく、「景気が悪くなったので、各家庭の家計に余裕はなくなり、各家庭の娯楽に使われるお金が減り、結果として娯楽業の売り上げが減る」と考えることは、合理的である。また、逆に、「景気が良くなったので、各家庭の家計に余裕ができ、各家庭の娯楽に使われるお金が増え、結果として娯楽業の売り上げが上がる」と考えることにも無理はない。

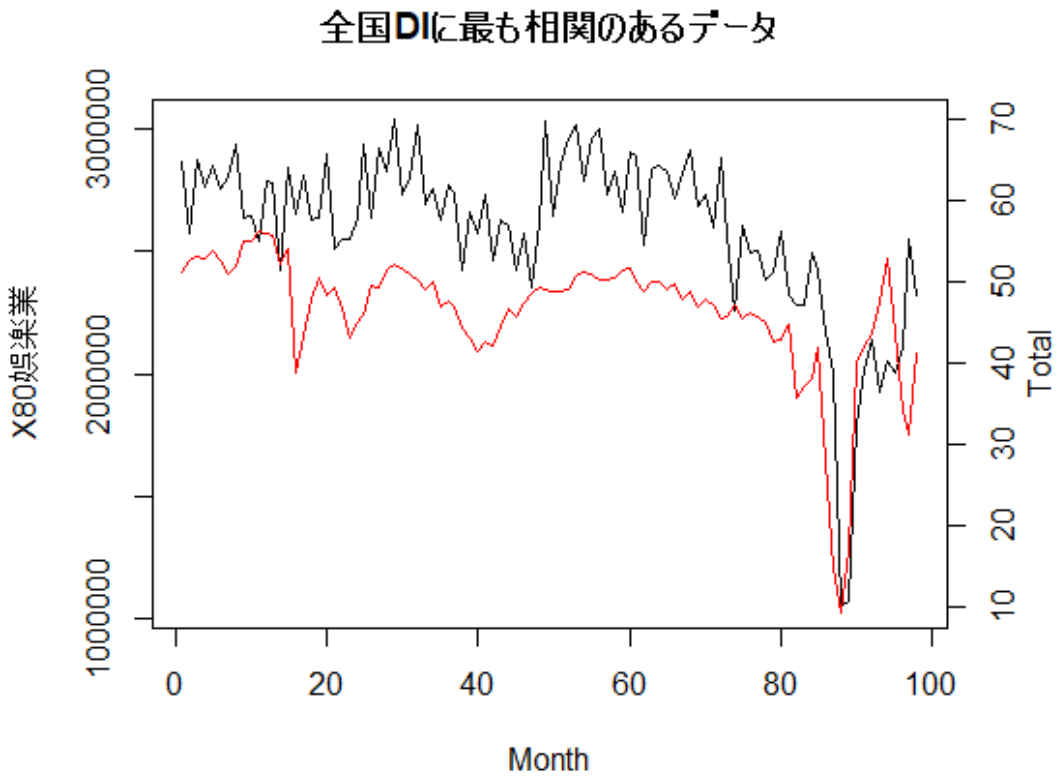
これらにより、景気から娯楽業に対して、因果関係がある可能性が高い。

補助のデータ：

第三次産業.csv:N生活関連サービス業，娯楽業　（※「家事サービス業」を除く）
(相関係数: 0.7230791)
同調査：43 道路旅客運送業（相関係数：0.6049286）
同調査：X75 宿泊業(相関係数：0.5552554)
同調査：M宿泊業.飲食サービス業（相関係数：0.5451438）

家計調査: 家計収支編：二人以上の世帯：月次：品目分類（2020 年改訂）（総数：金額）
134 パスタ【円】（相関係数：-0.58723）（外食の代わりに自炊で質素な食事が増える）
同調査：328 ソース【円】（相関係数：-0.4681847）
同調査：139 他の麺類【円】（相関係数：-0.4116774）
商業動態統計調査：確報：月次：百貨店・スーパー販売　経済産業局別販売額　月次：百貨店：北海道（相関係数：0.546774）（不景気になると、多くの人は質素倹約に努めようとするため、一般的に贅沢とされる百貨店での買い物は控えられる）
機械受注統計調査：月次：産業機械_合成樹脂加工機械（相関係数：0.4594177）
同調査：工作機械（相関係数：0.4564218）

折れ線図：



黒：x80 娯楽業の売り上げ　　赤：分野別 DI:合計

課題 5

コード

```
#-----課題5-----------------------------
```

```r
#変更部
Variant_name <- "Medium" #人口の Variant を決める
Covid19_file <- "Covid19.csv" #処理済み(国名の表記を国連に合わせる)
population_file <- "WPP2019_TotalPopulationBySex.csv" #文字化け修正済み
target_file <- "doctor_count.csv" #1列目が、ISO3 の国名コードで、2 列目以降が各年の調べたいデータ
excluded_population <- 100000 #この値以下の人口の国を除く
rank <- 1  #target_file の中で相関係数が何番目に強い列を表示するか
worst_rank <- FALSE #TRUE:target_file の中で相関係数が最も高い列、FALSE:相関係数が最も低い列
#--------
#カレントディレクトリをデータがあるところに変更
setwd("C:/Users/81802/OneDrive/東工大/授業の資料/2021 1Q/確率・統計学/中間レポート")
#---------ファイル読み込み-------------------------
#df_covid19 <- read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv")
#カンマ区切り(CSV)形式のファイルをデータフレームとして読み込み
df_covid19 <- read.csv(Covid19_file, sep=",",header=T)
df_population <- read.csv(population_file, sep=",", header=T)
#相関を調べたいデータを読み込み
target_data <- read.table(target_file, sep=",", header=T)

#----------データをきれいにする----------------------
#2021 年のデータのみを残す
df_population2021 <- subset(df_population, Time=="2021")
#Variant_name のみ残す
df_population2021_Medium <- subset(df_population2021, Variant==Variant_name)
#アルファベット順に並べなおす
df_population2021_Medium <- df_population2021_Medium[order(df_population2021_Medium[,2]),]
#列名を Location にする
colnames(df_covid19)[1] <- "Location"

#国連のデータの国と地域以外の大きいエリア(アフリカ全体の人口など)を除く
```

```r
#1 行目から 189 行目まで 1 行ずつ処理する
for(x in 1:189){
  #国名番号が同じ値になるまで消し続ける(両データがアルファベット順に並んでいるので、かぶらない番号の時にその行を抜くことで揃えられる)
  while(df_population2021_Medium[x,1]!=df_covid19[x,13]){
    #人口のデータの x 行目を取り除く
    df_population2021_Medium <- df_population2021_Medium[-x,]
  }
}
#Location を基準にして df_population2021_Medium,と df_covid19 の二つを一つにする
merged_data <- merge(df_population2021_Medium, df_covid19, by="Location")
#100 万人当たりの死者数を計算し、それぞれベクトルに入れる
Death_per_1M <- round((merged_data$Deaths/merged_data$PopTotal) * 1000, 3)
#100 万人当たりの死者数という新しい行を作る
merged_data <- cbind(merged_data, Death_per_1M)
#NA が残ると欠損値の処理がしにくくなるので、必要な列のみを残す
merged_data <- merged_data[,c(1,2,9,10,12,13,15,21:24)]

#-----極端に人口の少ない国を除く-----------------------
#次処理のループで取り消したことでずれる分をカウントする
counter <- 0
#1行目から最後の行まで一行ずつ処理する
for (x in 1:nrow(merged_data)) {
  #人口が基準値以下の場合
  if(merged_data[x-counter,3] <= excluded_population / 1000){
    #第(x-counter)行を取り除く
    merged_data <- merged_data[-(x-counter),]
    #次から一行ずれるので、ずらす分を1増やす
    counter = counter + 1
  }
}

#-----順位を下から何番目かに変える-----
#相関を調べたいデータの列数
element_count = ncol(target_data)
```

```
#worst_rank が true の場合
if(worst_rank){
  #下からの順位に変える
  rank = element_count - rank
}

#-----相関を調べたいデータの各列で相関を調べる-----
#各列の相関をベクトルの各要素に入れるときに必要なベクトルを作成
correlation_vector <- numeric(element_count-1)
data_count <- numeric(element_count-1)
#第2列から target_data の最後の列まで1列ずつ処理を行う
for(x in 2:element_count){
  #merged_data を merged_data_1 に代入
  merged_data_1 <- merged_data
  #merged_data_1 の最後の列に調べたいデータを ISO3(JPN など)を基準にくっつける
  merged_data_1 <- merge(merged_data_1, target_data[,c(1, x)], by="ISO3",
 all=T)
  #データが欠損した行を除く
  merged_data_1 <- na.omit(merged_data_1)
  data_count[x-1] <- nrow(merged_data_1)
  #merged_data の第 12 列の標準偏差が 0 でないとき
  if(var(merged_data_1[,12]) != 0){
    #DI_Total と correlation_vector の x 列目の相関係数
    correlation_vector[x-
1] <- cor(merged_data_1[,11],merged_data_1[,12])
  }
}

#-----相関係数が最も高い列を求める------
#相関係数の低い列から順に並べて、第何列かを correlation_list に記述していく
correlation_list <- sort.list(correlation_vector) + 1  #target_data の第2
列から最後の列までの並べ替えであるため、すべての列が 1 小さいことに注意
#相関係数の高い列から順に並べてその列を correlation_list に記述していく
correlation_list <- rev(correlation_list)
saved_correlation_vector <- correlation_vector
#相関係数を降順に並べ替え[正→負]
```

```
correlation_vector <- rev(sort(correlation_vector))
#rank 番目の相関係数とその項目を代入
correlation_rank <- correlation_vector[rank]
correlation_rank_target_data <- names(target_data)[correlation_list[rank]
]
#相関係数が rank 番目の列の各値を表示
print(correlation_rank)                    #相関係数
print(correlation_rank_target_data)         #年
print(data_count[correlation_list[rank]-1]) #データ数

#最も新しい年の各値を表示
print(saved_correlation_vector[element_count-1])#相関係数
print(names(target_data)[element_count-1])       #年
print(data_count[element_count-1])               #データ数

#最も新しい年のデータ数が少ない場合に使用
print(saved_correlation_vector[element_count-2])#相関係数
print(names(target_data)[element_count-2])       #年
print(data_count[element_count-2])               #データ数

#相関係数の推移
plot(saved_correlation_vector,type="l")
```

データの収集方法及び整理

The world bank のデータを用いて調べた。また、人口が１０万人よりも少ない国を除くことにした。１０万人以下の国では、１人が無くなると、１００万人当たりの死者数が１０人分以上加算されるため、正しいデータを取ることが出来なくなるため、これを行う。

結果

結論としては、新型コロナウイルスに影響する要因は、その国の社会の高齢化の進行度だ。65 歳以上の人口(全体の割合)のデータの 2017 年のデータを用い、その相関係数は、0.6017298 となった。この相関を調べるときには、１７７か国の国のデータを用いており、データ数の信頼性は高いと考えられる。

次に、疑似相関の可能性について考える。

今回調べたデータのうち、絶対値が 0.5 を超える相関を示した年があったデータは、年少

人口（相関係数：-0.5729729：2018 年、データ数：１７７か国)、１０００人当たりの医師数（相関係数：0.6853606：１９７５年、データ数：８２か国）（相関係数：0.5345396:2016 年、データ数：102 か国)、平均余命（相関係数：-0.5632511：1961 年、データ数：147 か国)、（相関係数：-0.3663833:2016 年、データ数：162 か国）の３つのデータとなった。このうち、１９７５年の１０００人当たりの医師数や１９６１年の平均余命のデータと相関が高いのは、４０年前から５０年前に人口に対して医師が多い国や平均余命が高い国は、その時から医療が発達している国であり、昔から医療が発達していることで高齢になるまで生きることが出来る割合が高くなり、高齢化が進んだことで、新型コロナウイルス感染症による死亡率が上がったことが考えられる。

１０００人あたりの医師数との相関が２０１６年でも大きいのは、高齢化と関連があると思う。人口に対して医師が多いことは、医療が発達しているということを表し、それは高齢化が進むことと関連があると考えられる。つまり、医師が多いことで新型コロナウイルスからより多くの人を助けられるという要因よりも、新型コロナウイルスにかかると重篤になりやすい高齢者が多いという要因の方が、新型コロナの死者数に大きな影響を与えているということを示している。

年少人口との相関は、高齢化が進んだ国は、比較的先進的な国であり、先進国では、一人当たりにかける教育費が上がることで、産む子供の数が減り、少子化が進む傾向にあり、高齢化と少子化は表裏一体であることを示している。先進国であることが要因となるならば糖尿病や衛生サービスなどと相関がみられると考えられるので、先進国であることと新型コロナウイルスの死者数の間に直接的な関係があるとは考えにくく、先進国かどうかは第三の変数にならないと考える。

今回調べたデータに、第三の変数を示すものはなく、高齢化が進んだ国では、新型コロナウイルスにかかると重篤になりやすいといわれる高齢者が多いので、新型コロナウイルスによる死者が増えると考えることに無理はないため、新型コロナウイルスの死者数に影響を与える要因として国内の高齢化の度合いを出すことは妥当である。

調べたデータ

病院のベッド(1,000 人あたり)、医師(1,000 人あたり)、看護師・助産師(1,000 人あたり)、現在の医療費(GDP に占める割合)、一人当たりの現在の健康支出(現在の米ドル)、一人当たりの現在の健康支出：PPP(現在の国際$)、自己負担支出(現在の医療費の割合)、糖尿病の罹患率(人口 20~79 歳の割合)、喫煙率：合計：15 歳以上、出生時の平均余命：合計(年)、安全に管理された衛生サービスを利用している人(人口の割合)、65 歳以上の人口(全体の割合)、人口年齢 0~14(総人口の割合)