

Rを使った分析 スクリプト解説

この解説の見方の説明

- 灰色のボックスにはRのスク립トを記載しています。
- #以下はコメント（注釈）として、補足を記載しています。（#以下のコメントは入力してもコンピュータに無視され、影響がありません。入力不要です。）
- 文頭の「>」は、Consoleにコードを入力する際に自動で挿入される記号です。
「>」は入力する必要はありませんが、ここでは複数行にわたるスク립トを1つのスク립トとして分かりやすくするために記載しています。

```
> data <- read.csv("24nsnen.csv") # データの読み込み  
> rownames(data) <- data[,1] # 1列目を行名にする
```

- `data <- read.csv("24nsnen.csv")` :「24nsnen.csv」という名前のファイルを読み込んで「data」に格納します。
- `rownames(data) <- data[,1]` :dataの1列目(市区町村名)を行名(rownames)にします。

- 白いボックスにはRのスク립トごとの解説を記載しています。
- 該当のスク립トを記載した後、「:」以下には関数や引数（「,」で区切られている部分）の設定の解説をしています。

Rを使った分析＞使用したパッケージ

パッケージ名	概要
sf	地理空間データを扱うためのツールです。 sf は「 Simple Features 」の略で、地理空間データを効率的に操作、解析、可視化するための標準的な形式が提供されています。
maps	地図描画用の基本ライブラリ。
mapdata	より詳細な地図データを提供する補助ライブラリ。

`install.packages(“パッケージ名”)`でインストールできます。

Rを使った分析＞地価公示データの読み込みと確認

```
> library(sf) # パッケージの読み込み
> d <- read_sf("L01-25_26_GML") # 地価公示データの読み込み
> dim(d) # データ確認
```

- `library(sf)` : 地理情報が入ったGMLを扱うため、`sf`パッケージを読み込みます。
- `d <- read_sf("L01-25_26_GML")` : `read_sf()` は地理情報が入ったファイル(ShapefileやGML)を読み込む関数です。ここでは地価公示データの「L01-25_26_GML」を読み込んで、「d」に格納しています。
- `dim(d)` : `dim()` はデータの行数と列数を教えてくれる関数です。

```
> library(sf)
> d <- read_sf("L01-25_26_GML")
> dim(d)
[1] 622 147
> |
```

3行目まで実行したconsole画面です。

`dim`関数の結果は黒字で表示されています。

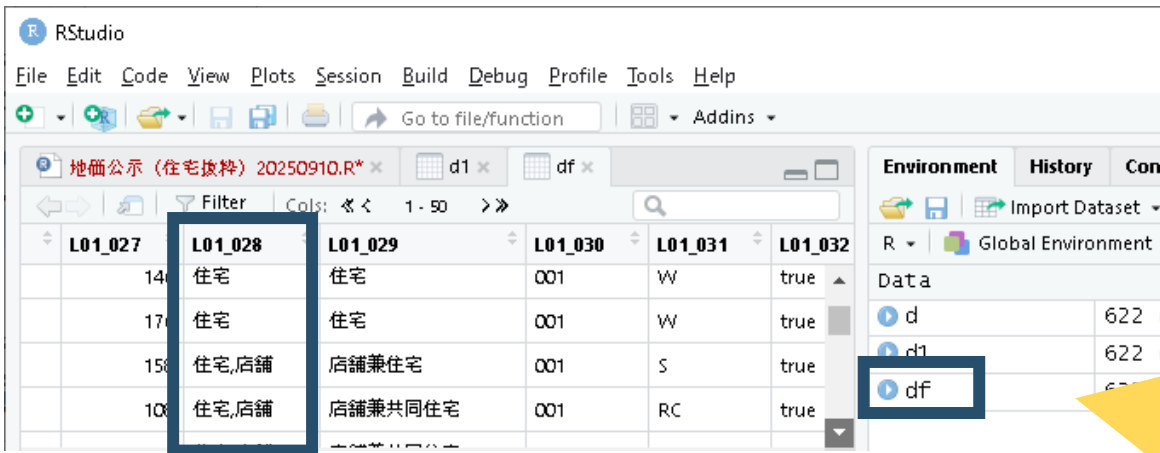
「d」は622行 147列のデータであることが分かります。

Rを使った分析＞地価公示データの整理①

```
> d1 <- st_drop_geometry(d) # 座標情報をなくしたデータフレームにする
> dim(d1)
> df <- as.data.frame(d1)
```

- `d1 <- st_drop_geometry(d)` : `st_drop_geometry()`は地図の形「geometry」を取り除く関数です。「d」は座標情報などを含む地図データなので、座標情報を取り除いて普通の表として扱えるようにし、「d1」に格納します。
- `dim(d1)` : 「d1」の行数、列数を教えてくれます。「d」よりも1列減っていることが確認できます。
- `df <- as.data.frame(d1)` : `as.data.frame()`はデータを普通のデータフレーム(行と列だけの表)として扱うための関数です。
「d1」は`st_drop_geometry`で座標情報を取り除きましたが、ここで内部的に残っている地理空間属性を完全に消して「df」に格納しています。

Rを使った分析＞地価公示データの整理②



RStudio interface showing the 'df' data frame. The 'L01_028' column is highlighted, showing values like '住宅', '住宅', '住宅,店舗', and '住宅,店舗'. The 'Environment' pane on the right shows 'df' as a data frame with 622 rows.

	L01_027	L01_028	L01_029	L01_030	L01_031	L01_032
14		住宅	住宅	001	W	true
17		住宅	住宅	001	W	true
15		住宅,店舗	店舗兼住宅	001	S	true
10		住宅,店舗	店舗兼共同住宅	001	RC	true

右上の画面で、dfのデータを確認します。

地点ごとの住所や概況、建物がある場合は建物の概況などが表形式でまとまっていることが確認できます。

今回は住宅に関する分析をしたいので、土地の利用状況の情報がある列を確認すると、「L01_028」の列に「空地、住宅」などの情報があることが確認できました。

この列の住宅、併用住宅(住宅,店舗など)のデータを抜き出していきます。

データ列の情報はホームページでも確認できます。

[国土数値情報 | 地価公示データ](#)

```
> df_house <- subset(df, grepl("住宅", L01_028)) # dfから住宅を含む行だけ抜き出す
```

- `df_house <- subset(df, grepl("住宅", L01_028))`: `subset()`は条件に合う行だけを取り出す関数です。
`grepl("住宅", L01_028)` は、「L01_028 という列に「住宅」という文字が含まれているかどうか」を調べています。
つまり「L01_028」の列から「住宅」を含む行だけ取り出し、「df_house」に格納しています。

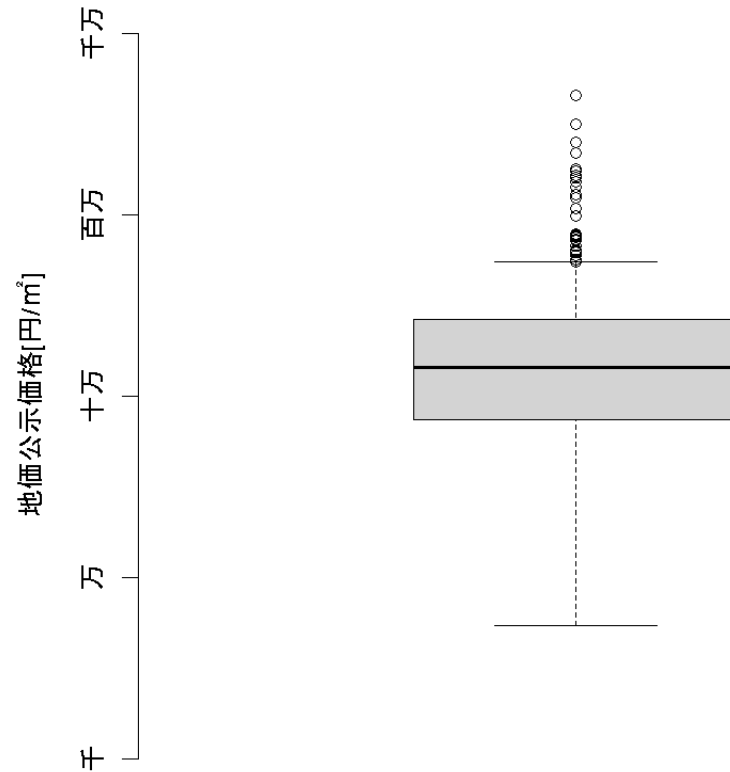
Rを使った分析＞地価公示データの確認（箱ひげ図）

```
> png("地価公示価格(住宅)_箱ひげ図.png",width=1000,height=1000,pointsize=24)
> boxplot(df_house$L01_008,log="y",ylab="地価公示価格[円/㎡]",axes=F,ylim=c(10^3,10^7)) # 箱ひげ図作図
> axis(side=2,at=10^(3:7),labels=c("千","万","十万","百万","千万")) # 軸の設定
> dev.off()
```

データの散らばりを確認するため、箱ひげ図を用いて標準地の地価公示価格(円/㎡)を確認します。
ホームページで属性名を確認すると、「L01_008」の列が地価公示価格であることが分かります。

- `png("地価公示価格(住宅)_箱ひげ図.png",width=1000,height=1000,pointsize=24)`: `png`関数でPNGファイルに保存するためのデバイスを開きます。""でファイル名を、`width`で横幅(ピクセル単位)を、`height`で高さを、`pointsize`で文字の大きさを指定しています。ファイル名の指定だけでも出力できます。
- `boxplot(df_house$L01_008,log="y",ylab="地価公示価格[円/㎡]",axes=F)`: `boxplot()`は箱ひげ図を描く関数です。「`df_house`」の「`L01_008`」列(地価公示価格)で箱ひげ図を描くよう指示しています。`log="y"`ではY軸を対数目盛にし、`ylab`ではY軸のラベルを設定しています。`axes=F`は軸を自動で描かないよう、`ylim`でY軸の範囲を設定しています。
- `axis(side=2,at=10^(3:7),labels=c("千","万","十万","百万","千万"))`: `axis()`でY軸を手動で描いています。`side=2`はY軸を描くよう指示を、`at=10^(3:7)`は10の3乗から10の7乗の位置に目盛を描くよう指示を、`labels=c("千","万","十万","百万","千万")`でそれぞれの目盛のラベルを指示しています。
- `dev.off()`: `dev.off`関数でデバイスを閉じてファイルを保存します。作業ディレクトリにファイルが保存されました

Rを使った分析＞箱ひげ図の確認



✓ 箱ひげ図のPNGファイルが作成されました。

地価公示価格は、数千円台から数十万円台に分布し、四分位範囲は数万円から数十万円であることが分かります。

また外れ値が多くあり、㎡単価が数百万円台となっている地点が数か所あることがうかがえます。

Rを使った分析＞地価公示データの確認（CSVファイル）

平均値(mean)や最大値(max)、最小値(min)でも集計できます。

箱ひげ図から外れ値が何件かあることが確認できたので、今回は中央値を選択しました。

```
> d2 <- aggregate(L01_008~L01_001,data=df_house,FUN=median) # 行政区コードごとに中央値を集計
> dim(d2)
> write.csv(d2,"行政区域コード別集計(住宅).csv",row.names=F) # CSVファイルで保存
```

- `d2 <- aggregate(L01_008~L01_001,data=df_house,FUN=median)`: `aggregate()` は「グループごとに計算する」関数です。L01_008（地価公示価格）の中央値(median)をL01_001(行政区域コード)ごとに計算して「d2」に格納するよう指示しています。
- `dim(d2)`: d2の行数、列数を確認します。「[1] 32 2」と返され、36市区町村のうち、4つの行政区はデータがないことが分かります。
- `write.csv(d2,"行政区域コード別集計(住宅).csv",row.names=F)`: `write.csv()` はデータをcsvファイルとして保存する関数です。""はファイル名を指定し、`row.names=F`は「行番号をつけない」という指定です。

行政区域ごとの地価公示価格の中央値が抜き出されたcsvファイルが保存されました。

Rを使った分析＞地図グラフの作成①

```
> d3 <- aggregate(cbind(X,Y)~L01_001,  
                  data=cbind(d1,st_coordinates(d)),FUN=mean) # 座標情報のみを取り出し行政区域コード別に集計する(平均値)  
> head(d3)
```

地図グラフにプロットする点の位置を設定します。

- `d3 <- aggregate(cbind(X,Y)~L01_001, data=cbind(d1,st_coordinates(d)),FUN=mean)` :
このスクリプトでは、住宅の地価地点の平均位置を市町村ごとに求めてd3に格納しています。
`st_coordinates(d)`で座標情報をなくす前のデータ「d」から座標(x:経度、y:緯度)を取り出し、
`cbind(d1,st_coordinates(d))`で座標をなくした「d1」と結合しています。`aggregate(cbind(X,Y) ~ L01_001, ...)` は、行政区域コードごとに座標の平均を計算しています。
- `head(d3)`: d3の表頭から6行目までを確認します。
行政区域コード、x座標(経度)の平均、y座標(緯度)の平均が抜き出されていることが確認できます。

Rを使った分析＞地図グラフの作成②

```
> d23 <- merge(x=d2,y=d3,by="L01_001")
> dim(d23)
> write.csv(d23,"座標付公示価格(住宅).csv",row.names=F)
> library(maps)
> library(mapdata)
> png("地価公示価格分布.png",width=1000,height=1000,pointsize=24)
> map("japan","kyoto")
```

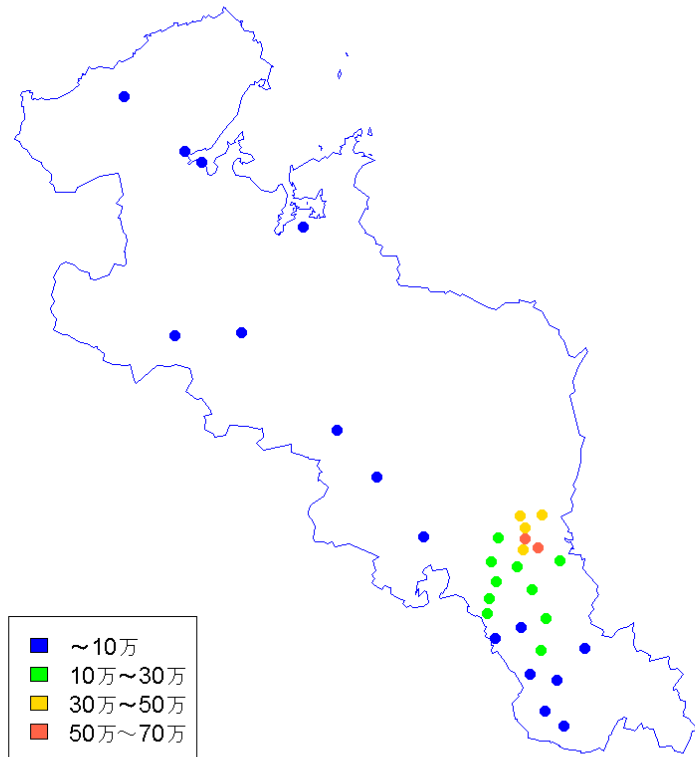
- `d23 <- merge(x=d2,y=d3,by="L01_001")` :`merge()`はデータを結合する関数です。行政区域コードによってd2(行政区域コードごとの地価公示価格の中央値)とd3(行政区域コードごとの座標の平均値)を結合し、d23に格納しています。
- `dim(d23)`:d23の行数、列数を確認します。「[1] 32 4」と返され、32行4列のデータになっていることが分かります。
- `write.csv(d23,"座標付公示価格(住宅).csv",row.names=F)`:d23をCSVファイルとして保存しています。
- `library(maps),library(mapdata)`:mapsとmapdataライブラリを読み込みます。
- `png("地価公地理分布.png",width=1000,height=1000,pointsize=24)`: png関数でPNGファイルに保存するためのデバイスを開きます。
- `map("japan","kyoto")`:日本地図の中から京都府の地図を描画します。

Rを使った分析＞地図グラフの作成③

```
> k <- cut(d23$L01_008, breaks = c(0, 100000, 300000, 500000, 700000), labels = FALSE)
> palette(c("blue", "green", "gold", "tomato"))
> points(d23$X, d23$Y, pch=19, col=k)
> legend("bottomleft", legend = c("～10万円", "10万～30万円", "30万～50万円", "50万～70万円"), fill =
c(1,2,3,4))
> dev.off()
```

- `k <- cut(d23$L01_008, breaks = c(0,100000,300000,500000,700000), labels = FALSE)`:
`cut`関数で「d23」の地価公示価格を4つの区間(～10万円、10万～30万円、30万円～50万円、50万円～70万円)に分類して「k」に格納します。`labels = FALSE`で分類ラベルを設定しないため、分類結果は1～4の数値で返されます。
- `palette(c("blue", "green", "gold", "tomato"))`:座標をプロットする際に1～4の分類によって色分けをする、色を指定します。1～4に対してそれぞれ、青色、緑色、黄色、赤色になるように設定しました。
- `points(d23$X, d23$Y, pch=19, col=k)`:`points`関数で座標を地図上に描画します。`pch=19`で点の形を「塗りつぶしの円」に指定し、`col=k`で点の色を上記で分類した4分類に設定しています。
- `legend("bottomleft", legend=c("～10万", "10万～30万", "30万～50万", "50万～70万"), fill=c(1,2,3,4))`:
凡例として4分類のラベル表示を指定し、凡例の表示位置を左下に指定しています。

Rを使った分析＞地図グラフの確認



✓ 地図グラフのPNGファイルが作成されました。

✓ 作成したCSVファイルと並べて見ると、理解が深まります。

北部地域と南部地域では、 m^2 単価（中央値）が10万円以下となっている一方で、京都市域、乙訓地域、山城北部地域などで10万円から30万円以上となっていることが分かります。

また、相楽東部地域の2町1村及び伊根町にはデータがないことが分かります。

Rを使った分析＞人口データと合わせる

エラーが起こるときは、以下のように入力してみてください。
`read.csv("2504ssnen.csv",fileEncoding = "UTF-8-BOM")`

```
> d4 <- read.csv("2504ssnen.csv")  
> d234 <- merge(x=d23,y=d4,by.x="L01_001",by.y="団体コード")  
> dim(d234)
```

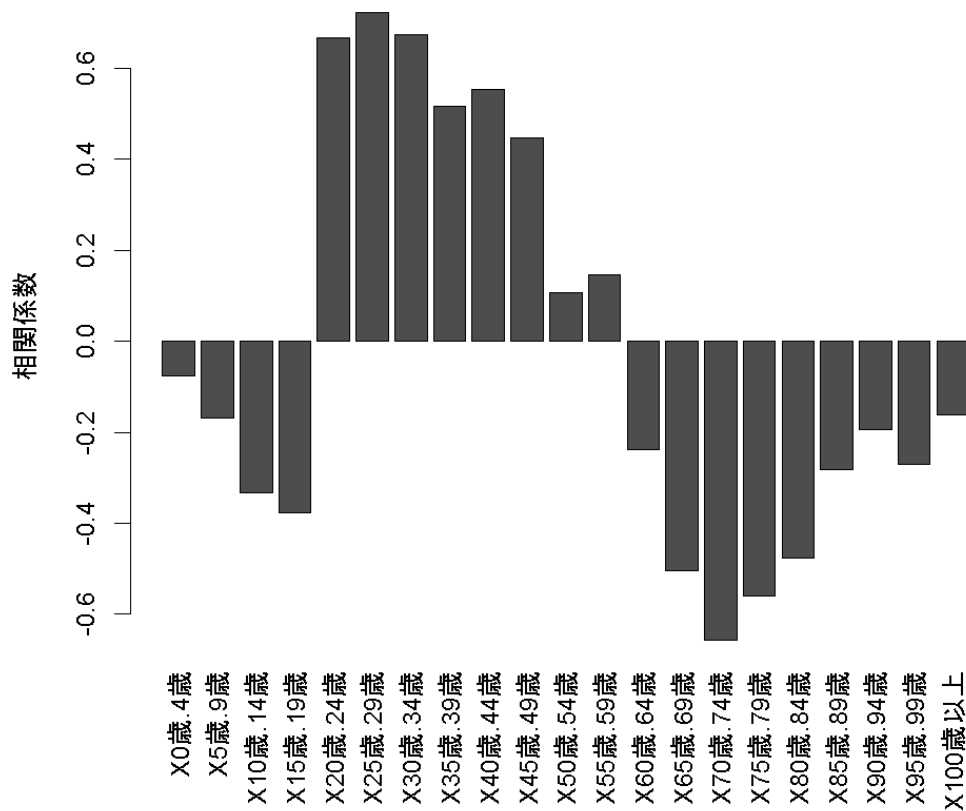
- `d4 <- read.csv("2504ssnen.csv")`:年齢階級別人口データ「2504ssnen.csv」を読み込み、d4に格納しています。
- `d234 <- merge(x=d23,y=d4,by.x="L01_001",by.y="団体コード")`:d23(地価公示価格データ)とd4(人口データ)を「団体コード」で結合しています。
- `dim(d234)`:d234の行数、列数を確認すると[1] 32 29と返され、32行29列のデータとなっています。

Rを使った分析＞地価公示価格と年齢階級別人口割合との相関係数

```
> prop <- d234[,9:29]/d234[,8] #年齢階級別人口割合を算出する
> png("地価公示価格_相関係数(住宅).png",width=1000,height=1000,pointsize=24)
> par(mar=c(10,4,4,2))
> barplot(cor(d234[, "L01_008"],prop),las=3,ylab="相関係数")
> dev.off()
```

- `prop <- d234[,9:29]/d234[,8]`: `d234`のすべての行の9列から29列目(各年齢階級別人口)を`d234`のすべての行の8列目(人口総数)で割って、`prop`に格納しています。
- `png("地価公示価格_相関係数(住宅).png",width=1000,height=1000,pointsize=24)`: `PNG`関数でデバイスを開いてファイル名、横幅、高さ、フォントサイズを設定しています。
- `par(mar=c(10,4,4,2))`: グラフの余白(margins)を指定しています。(下、左、上、右)の順に行数で指定しています。
- `barplot(cor(d234[, "L01_008"],prop),las=3,ylab="相関係数")`: `barplot()`は棒グラフを描画する関数です。`cor(d234[, "L01_008"],prop)`は市区町ごとの地価公示価格(中央値)と各年齢階級別人口割合の相関係数を計算しています。`las=3`はx軸のラベルを90度回転させて重ならないようにし、`ylab`ではy軸のラベルを設定しています。
- `dev.off()`: デバイスを閉じてPNGファイルを保存します。

Rを使った分析＞地価公示価格と年齢階級別人口割合との相関係数



✓ 地価公示価格と年齢階級別人口割合の相関係数グラフのPNGファイルが作成されました。

地価公示価格（中央値）と20～24歳から55～59歳の年齢階級割合は相関係数が正の値に、他の年齢階級では相関係数が負の値となっていることが分かります。

20～24歳から30～34歳までの年齢階級では相関係数が約0.7と強い正の相関がみられ、反対に70～74歳の年齢階級では相関係数が約-0.7と強い負の相関がみられます。

このことは、地価公示価格（中央値）が高い地域では20～24歳から30～34歳の年齢階級の割合が高く、反対に地価公示価格（中央値）が低い地域では70～74歳の年齢階級割合が高い傾向があることを示しています。

19歳未満の人口割合の傾向と、その親世代にあたる年齢階級の割合とで別の傾向があることから、同じ年齢階級の中でも子どもの有無によって異なる傾向があることも推測できます。