Data Analysis using R

500666461

2023-11-17

Loading required libraries *dplyr*, *stringr* and *ggplot2*

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(ggplot2)
```

Task 1

Reading the dataset into R

```
cdrc_data <- read.csv("hh_churn_lsoa11_2023.csv")
str(cdrc_data) #extraneous
```

```
## 'data.frame':    42619 obs. of  28 variables:
##  $ area   : chr  "95AA01S1" "95AA01S2" "95AA01S3" "95AA02W1" ...
##  $ chn1997: num  1 0.663 0.7 0.591 0.602 0.633 0.658 0.685 0.675 0.539 ...
##  $ chn1998: num  1 0.651 0.688 0.576 0.591 0.618 0.64 0.663 0.644 0.522 ...
##  $ chn1999: num  1 0.636 0.671 0.56 0.586 0.601 0.61 0.638 0.607 0.498 ...
##  $ chn2000: num  1 0.629 0.663 0.549 0.584 0.593 0.591 0.624 0.584 0.483 ...
##  $ chn2001: num  1 0.618 0.653 0.542 0.581 0.583 0.577 0.615 0.569 0.473 ...
##  $ chn2002: num  1 0.603 0.633 0.537 0.569 0.56 0.559 0.596 0.55 0.458 ...
##  $ chn2003: num  1 0.59 0.611 0.534 0.554 0.535 0.537 0.566 0.523 0.443 ...
##  $ chn2004: num  1 0.569 0.591 0.529 0.541 0.514 0.512 0.538 0.485 0.427 ...
##  $ chn2005: num  1 0.541 0.57 0.524 0.524 0.494 0.492 0.521 0.444 0.408 ...
##  $ chn2006: num  1 0.517 0.555 0.518 0.51 0.478 0.48 0.506 0.412 0.391 ...
##  $ chn2007: num  1 0.506 0.548 0.517 0.506 0.467 0.475 0.489 0.395 0.378 ...
##  $ chn2008: num  1 0.495 0.538 0.506 0.502 0.46 0.465 0.47 0.374 0.366 ...
```

```
##  $ chn2009: num  1 0.48 0.528 0.49 0.494 0.439 0.445 0.44 0.343 0.35 ...
##  $ chn2010: num  1 0.47 0.522 0.48 0.487 0.415 0.43 0.417 0.322 0.332 ...
##  $ chn2011: num  1 0.462 0.515 0.47 0.478 0.402 0.42 0.406 0.311 0.321 ...
##  $ chn2012: num  1 0.446 0.495 0.448 0.454 0.387 0.402 0.385 0.292 0.307 ...
##  $ chn2013: num  1 0.429 0.469 0.411 0.419 0.369 0.379 0.362 0.274 0.292 ...
##  $ chn2014: num  1 0.416 0.453 0.387 0.398 0.353 0.364 0.351 0.268 0.284 ...
##  $ chn2015: num  0.969 0.397 0.435 0.362 0.371 0.335 0.354 0.333 0.256 0.276 ...
##  $ chn2016: num  0.928 0.373 0.412 0.331 0.337 0.314 0.344 0.308 0.239 0.267 ...
##  $ chn2017: num  0.914 0.336 0.367 0.297 0.304 0.281 0.317 0.273 0.217 0.25 ...
##  $ chn2018: num  0.879 0.287 0.31 0.262 0.27 0.239 0.282 0.232 0.188 0.227 ...
##  $ chn2019: num  0.799 0.256 0.279 0.243 0.252 0.216 0.264 0.207 0.169 0.21 ...
##  $ chn2020: num  0.723 0.244 0.266 0.233 0.244 0.205 0.257 0.192 0.16 0.199 ...
##  $ chn2021: num  0.537 0.186 0.182 0.172 0.177 0.145 0.189 0.132 0.11 0.145 ...
##  $ chn2022: num  0.29 0.102 0.072 0.087 0.081 0.066 0.09 0.055 0.044 0.071 ...
##  $ chn2023: int  0 0 0 0 0 0 0 0 0 0 ...
```

Task 2

Examining the LSOA code W01000092 Menai (Bangor), and identifying the residential mobility indices which are greater than 0.5 and less than 0.8.

```
cdrc_data_menai_df <- filter(cdrc_data, area == "W01000092")
cdrc_data_menai <- unlist(cdrc_data_menai_df[1, -1])

cdrc_data_menai_sub <- cdrc_data_menai[cdrc_data_menai > 0.5 & cdrc_data_menai < 0.8]
# printing the named vector
print(cdrc_data_menai_sub)
```

```
## chn2002 chn2003 chn2004 chn2005 chn2006 chn2007 chn2008 chn2009 chn2010 chn2011
##   0.791   0.783   0.776   0.771   0.765   0.758   0.752   0.747   0.743   0.729
## chn2012 chn2013 chn2014 chn2015 chn2016 chn2017
##   0.701   0.664   0.634   0.609   0.579   0.529
```

```
# printing the names of the vector
print(names(cdrc_data_menai_sub))
```

```
##  [1] "chn2002" "chn2003" "chn2004" "chn2005" "chn2006" "chn2007" "chn2008"
##  [8] "chn2009" "chn2010" "chn2011" "chn2012" "chn2013" "chn2014" "chn2015"
## [15] "chn2016" "chn2017"
```

```
# printing the actual years
x <- str_replace(names(cdrc_data_menai_sub), "chn", "")
print(x)
```

```
##  [1] "2002" "2003" "2004" "2005" "2006" "2007" "2008" "2009" "2010" "2011"
## [11] "2012" "2013" "2014" "2015" "2016" "2017"
```

Task 3

Appending additional columns to the data

```r
# adds column one (averaged residential mobility for each region) as ARMI
cdrc_data$ARMI <- round(rowMeans(cdrc_data[, -c(1, 28)]), 3)
# adds column two (grouping regions into low, medium and high) as ARMIgrpd
cdrc_data$ARMIgrpd <- cut(cdrc_data$ARMI, breaks = c(0, 0.2, 0.5, 1), labels = c("Low", "Medium", "High
str(cdrc_data)
```

```
## 'data.frame':    42619 obs. of  30 variables:
##  $ area    : chr  "95AA01S1" "95AA01S2" "95AA01S3" "95AA02W1" ...
##  $ chn1997 : num  1 0.663 0.7 0.591 0.602 0.633 0.658 0.685 0.675 0.539 ...
##  $ chn1998 : num  1 0.651 0.688 0.576 0.591 0.618 0.64 0.663 0.644 0.522 ...
##  $ chn1999 : num  1 0.636 0.671 0.56 0.586 0.601 0.61 0.638 0.607 0.498 ...
##  $ chn2000 : num  1 0.629 0.663 0.549 0.584 0.593 0.591 0.624 0.584 0.483 ...
##  $ chn2001 : num  1 0.618 0.653 0.542 0.581 0.583 0.577 0.615 0.569 0.473 ...
##  $ chn2002 : num  1 0.603 0.633 0.537 0.569 0.56 0.559 0.596 0.55 0.458 ...
##  $ chn2003 : num  1 0.59 0.611 0.534 0.554 0.535 0.537 0.566 0.523 0.443 ...
##  $ chn2004 : num  1 0.569 0.591 0.529 0.541 0.514 0.512 0.538 0.485 0.427 ...
##  $ chn2005 : num  1 0.541 0.57 0.524 0.524 0.494 0.492 0.521 0.444 0.408 ...
##  $ chn2006 : num  1 0.517 0.555 0.518 0.51 0.478 0.48 0.506 0.412 0.391 ...
##  $ chn2007 : num  1 0.506 0.548 0.517 0.506 0.467 0.475 0.489 0.395 0.378 ...
##  $ chn2008 : num  1 0.495 0.538 0.506 0.502 0.46 0.465 0.47 0.374 0.366 ...
##  $ chn2009 : num  1 0.48 0.528 0.49 0.494 0.439 0.445 0.44 0.343 0.35 ...
##  $ chn2010 : num  1 0.47 0.522 0.48 0.487 0.415 0.43 0.417 0.322 0.332 ...
##  $ chn2011 : num  1 0.462 0.515 0.47 0.478 0.402 0.42 0.406 0.311 0.321 ...
##  $ chn2012 : num  1 0.446 0.495 0.448 0.454 0.387 0.402 0.385 0.292 0.307 ...
##  $ chn2013 : num  1 0.429 0.469 0.411 0.419 0.369 0.379 0.362 0.274 0.292 ...
##  $ chn2014 : num  1 0.416 0.453 0.387 0.398 0.353 0.364 0.351 0.268 0.284 ...
##  $ chn2015 : num  0.969 0.397 0.435 0.362 0.371 0.335 0.354 0.333 0.256 0.276 ...
##  $ chn2016 : num  0.928 0.373 0.412 0.331 0.337 0.314 0.344 0.308 0.239 0.267 ...
##  $ chn2017 : num  0.914 0.336 0.367 0.297 0.304 0.281 0.317 0.273 0.217 0.25 ...
##  $ chn2018 : num  0.879 0.287 0.31 0.262 0.27 0.239 0.282 0.232 0.188 0.227 ...
##  $ chn2019 : num  0.799 0.256 0.279 0.243 0.252 0.216 0.264 0.207 0.169 0.21 ...
##  $ chn2020 : num  0.723 0.244 0.266 0.233 0.244 0.205 0.257 0.192 0.16 0.199 ...
##  $ chn2021 : num  0.537 0.186 0.182 0.172 0.177 0.145 0.189 0.132 0.11 0.145 ...
##  $ chn2022 : num  0.29 0.102 0.072 0.087 0.081 0.066 0.09 0.055 0.044 0.071 ...
##  $ chn2023 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ARMI    : num  0.925 0.458 0.489 0.429 0.439 0.412 0.428 0.423 0.364 0.343 ...
##  $ ARMIgrpd: Factor w/ 3 levels "Low","Medium",..: 3 2 2 2 2 2 2 2 2 2 ...
```

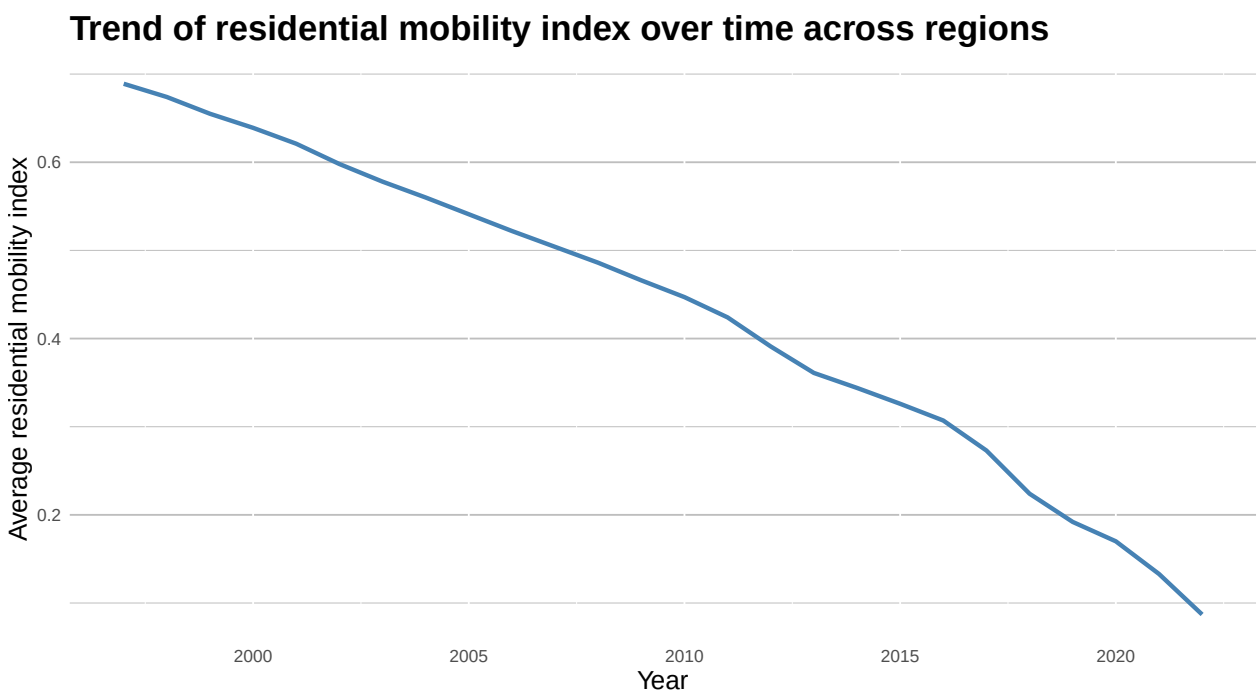
Task 4

First insight (with chart)

```r
avg_armi <- round(colMeans(cdrc_data[, 2:27]), 3)
years <- as.numeric(str_replace(names(avg_armi), "chn", ""))
ts_df <- data.frame(year = years, avg_rmi = avg_armi)

ggplot(ts_df, aes(x = year, y = avg_rmi)) +
  geom_line(color = "steelblue", linewidth = 1.2) +
  labs(x = "Year",
       y = "Average residential mobility index",
       title = "Trend of residential mobility index over time across regions") +
```

```
  theme(axis.ticks.y = element_blank(),
        panel.background = element_blank(),
        axis.text.y = element_text(size = 10),
        axis.ticks.x = element_blank(),
        panel.grid.minor.y = element_line(colour = "grey"),
        axis.text.x = element_text(size = 10),
        panel.grid.major.y = element_line(colour = "grey"),
        axis.title = element_text(size = 15, margin = margin(t = 15)),
        plot.margin = margin(20, 0, 20, 0),
        plot.title = element_text(size = 20, face = "bold"))
```

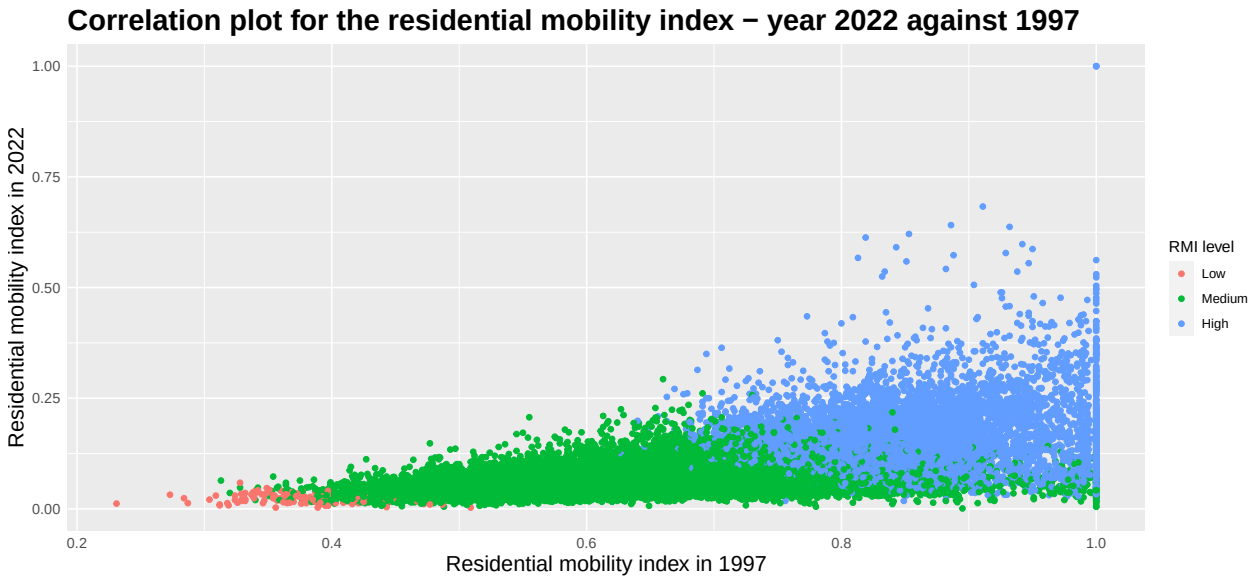


The chart is a time series plot showing the change in residential mobility index (averaged across all the regions) over time from 1997 to 2022. The plot clearly shows a sharp downward trend in the residential mobility index over time. The downward trend is also observed to be uniform, with the line approximately straight. The inference from the chart is that, on average, as years pass by, the residential mobility decreases; with this trend applying to all regions.

Second insight (with insight)

```
ggplot(cdrc_data, aes(x = chn1997, y = chn2022, color = ARMIgrpd)) +
  geom_point() +
  labs(x = "Residential mobility index in 1997",
       y = "Residential mobility index in 2022",
       color = "RMI level",
       title = "Correlation plot for the residential mobility index - year 2022 against 1997") +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_text(size = 10),
```

```
        axis.ticks.x = element_blank(),
        axis.text.x = element_text(size = 10),
        axis.title = element_text(size = 15, margin = margin(t = 15)),
        plot.margin = margin(20, 0, 20, 0),
        plot.title = element_text(size = 20, face = "bold"))
```



The residential mobility indices in 2022 are plotted against the residential mobility indices for 1997 to detect correlation in the indices across the extreme years. The plot clearly shows a positive and strong correlation. The points are classified in the plot according to the attribute Average Residential Mobility Index (ARMI) in the dataset. The interpretation from the plot could be that regions with high RMI in 1997 would also very likely have high RMI in 2022. Correspondingly, regions with low RMI in 1997 would likely have low RMI in 2022.