# Introduction to R Programming

Ozan Bakış[1]
[1] Bahcesehir University, Department of Economics and BETAM

# Outline

1. Data reshaping with dplyr/tidyverse

# Data reshaping I

```r
library(dplyr)
library(tidyr)
library(openxlsx)
f_url = "https://github.com/obakis/econ_data/raw/master/illere_gore_ihracat.xlsx"
download.file(url = f_url, destfile = "il_ihracat.xlsx", mode="wb")
dat = read.xlsx("il_ihracat.xlsx",  cols = 1:16, rows=5:1458, colNames = TRUE)
#head(dat)
dat = dat[,-c(3,4)]
names(dat)[1:2] = c("year","province")
names(dat)
```

```
##  [1] "year"      "province"  "January"   "February"  "March"
##  [6] "April"     "May"       "June"      "July"      "August"
## [11] "September" "October"   "November"  "December"
```

# Data reshaping II

```
dat = as_data_frame(dat)
str(dat)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':	1405 obs. of  14 variables:
##  $ year     : chr  "2018" NA NA NA ...
##  $ province : chr  NA "0" "1" "2" ...
##  $ January  : chr  "12456839.007999994" "124.199" "150321.90900000001" "12722.096" ...
##  $ February : chr  NA NA NA NA ...
##  $ March    : chr  NA NA NA NA ...
##  $ April    : chr  NA NA NA NA ...
##  $ May      : chr  NA NA NA NA ...
##  $ June     : chr  NA NA NA NA ...
##  $ July     : chr  NA NA NA NA ...
##  $ August   : chr  NA NA NA NA ...
##  $ September: chr  NA NA NA NA ...
##  $ October  : chr  NA NA NA NA ...
##  $ November : chr  NA NA NA NA ...
```

# Data reshaping III

```
##  $ December : chr  NA NA NA NA ...

# dat %>%
# mutate_each(funs(extract_numeric), year:december) -> dat1
dat %>%
  transmute_all(extract_numeric) -> dat1
#print(dat1[1350:1405,], n=10)

dat2 = fill(dat1, year, .direction = "down")
dat2 = dat2 %>%
  filter(! province %in% c(0,NA))
#print(dat2[,1:4], n=35, width=Inf)
dat_x1 = gather(data=dat2, key=month, value=export, -province, -year)
head(dat_x1)
```

# Data reshaping IV

```
## # A tibble: 6 x 4
##    year province month    export
##   <dbl>    <dbl> <chr>     <dbl>
## 1  2018        1 January 150322.
## 2  2018        2 January  12722.
## 3  2018        3 January  24786.
## 4  2018        4 January   2776.
## 5  2018        5 January   9008.
## 6  2018        6 January 529935.
```

```
dat_x1 %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(year,month, province) -> dat_x
print(dat_x,3)
```

# Data reshaping V

```
## # A tibble: 16,452 x 4
##     year province month     export
##    <dbl>    <dbl> <fct>      <dbl>
##  1  2002        1 January   35247.
##  2  2002        2 January     740.
##  3  2002        3 January    3163.
##  4  2002        4 January     190.
##  5  2002        5 January      19.3
##  6  2002        6 January 118803.
##  7  2002        7 January  13904.
##  8  2002        8 January    526.
##  9  2002        9 January   9959.
## 10  2002       10 January   6538.
## # ... with 16,442 more rows
```

```
saveRDS(dat_x, "tur_x.rds")
```

# Data reshaping VI

```r
f_url = "https://github.com/obakis/econ_data/raw/master/illere_gore_ithalat.xlsx"
download.file(url = f_url, destfile = "il_ithalat.xlsx", mode="wb")
dat = read.xlsx("il_ithalat.xlsx",
cols = 1:16, rows=5:1471, colNames = TRUE)
dat = dat[,-c(3,4)]
names(dat)[1:2] = c("year","province")
names(dat)
```

```
##  [1] "year"      "province"  "January"   "February"  "March"
##  [6] "April"     "May"       "June"      "July"      "August"
## [11] "September" "October"   "November"  "December"
```

# Data reshaping VII

```r
dat = as_data_frame(dat)
dat %>%
  transmute_all(extract_numeric) -> dat1

dat2 = fill(dat1, year, .direction = "down")
dat2 = dat2 %>%
  filter(! province %in% c(0,NA,99)) # imports

dat_m1 = gather(data=dat2, key=month, value=import, -c(province, year))
dat_m1 %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(year,month, province) -> dat_m
print(dat_m,3)
```

# Data reshaping VIII

```
## # A tibble: 16,488 x 4
##     year province month    import
##    <dbl>    <dbl> <fct>     <dbl>
##  1  2002        1 January  44761.
##  2  2002        2 January   1868.
##  3  2002        3 January   1295.
##  4  2002        4 January    680.
##  5  2002        5 January    271.
##  6  2002        6 January 358921.
##  7  2002        7 January   4623.
##  8  2002        8 January   1687.
##  9  2002        9 January   5557.
## 10  2002       10 January   3799.
## # ... with 16,478 more rows
```

```
saveRDS(dat_m, "tur_m.rds")
```

# Data reshaping IX

```r
f_url = "https://github.com/obakis/econ_data/raw/master/illere_gore_gsyh.xlsx"
download.file(url = f_url, destfile = "il_gsyh.xlsx", mode="wb")
dat = read.xlsx("il_gsyh.xlsx", rows=9:89, colNames = FALSE)
#head(dat)
nms1=as.vector(outer(c("agr","ind","ser","sectot", "tax","gdp"),2004:2014,paste, sep="_"))
nms = c("nuts3","province",nms1)
nms
```

```
##  [1] "nuts3"       "province"    "agr_2004"    "ind_2004"   
##  [5] "ser_2004"    "sectot_2004" "tax_2004"    "gdp_2004"   
##  [9] "agr_2005"    "ind_2005"    "ser_2005"    "sectot_2005"
## [13] "tax_2005"    "gdp_2005"    "agr_2006"    "ind_2006"   
## [17] "ser_2006"    "sectot_2006" "tax_2006"    "gdp_2006"   
## [21] "agr_2007"    "ind_2007"    "ser_2007"    "sectot_2007"
## [25] "tax_2007"    "gdp_2007"    "agr_2008"    "ind_2008"   
## [29] "ser_2008"    "sectot_2008" "tax_2008"    "gdp_2008"   
## [33] "agr_2009"    "ind_2009"    "ser_2009"    "sectot_2009"
```

# Data reshaping X

```
## [37] "tax_2009"    "gdp_2009"    "agr_2010"    "ind_2010"
## [41] "ser_2010"    "sectot_2010" "tax_2010"    "gdp_2010"
## [45] "agr_2011"    "ind_2011"    "ser_2011"    "sectot_2011"
## [49] "tax_2011"    "gdp_2011"    "agr_2012"    "ind_2012"
## [53] "ser_2012"    "sectot_2012" "tax_2012"    "gdp_2012"
## [57] "agr_2013"    "ind_2013"    "ser_2013"    "sectot_2013"
## [61] "tax_2013"    "gdp_2013"    "agr_2014"    "ind_2014"
## [65] "ser_2014"    "sectot_2014" "tax_2014"    "gdp_2014"
```

```
names(dat)=nms
dat$province=NULL
dat = as_tibble(dat)
dat1 = gather(data=dat, key=output, value=TL, -nuts3)
head(dat1)
```

# Data reshaping XI

```
## # A tibble: 6 x 3
##   nuts3 output         TL
##   <chr> <chr>       <dbl>
## 1 TR100 agr_2004  530330.
## 2 TR211 agr_2004  833109.
## 3 TR212 agr_2004  847308.
## 4 TR213 agr_2004  526600.
## 5 TR221 agr_2004 1544359.
## 6 TR222 agr_2004  965057.
```

```
dat2 = dat1 %>% separate(output, c("out", "year"))
dat3 = dat2 %>%
  spread(key = out, value = TL)
print(dat3,3)
```

# Data reshaping XII

```
## # A tibble: 891 x 8
##    nuts3 year     agr     gdp     ind  sectot     ser     tax
##    <chr> <chr>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 TR100 2004  530330.  1.73e8  4.71e7  1.51e8  1.04e8  2.18e7
##  2 TR100 2005  601554.  2.01e8  5.42e7  1.76e8  1.21e8  2.55e7
##  3 TR100 2006  566282.  2.37e8  6.55e7  2.07e8  1.41e8  2.97e7
##  4 TR100 2007  526371.  2.67e8  7.25e7  2.36e8  1.63e8  3.03e7
##  5 TR100 2008  512739.  3.01e8  7.93e7  2.68e8  1.89e8  3.31e7
##  6 TR100 2009  542941.  3.00e8  7.25e7  2.68e8  1.95e8  3.25e7
##  7 TR100 2010  572084.  3.44e8  8.35e7  3.02e8  2.18e8  4.15e7
##  8 TR100 2011  643161.  4.19e8  1.12e8  3.68e8  2.56e8  5.04e7
##  9 TR100 2012  782857.  4.76e8  1.24e8  4.20e8  2.95e8  5.59e7
## 10 TR100 2013  733029.  5.53e8  1.46e8  4.84e8  3.37e8  6.85e7
## # ... with 881 more rows
```

```
saveRDS(dat3,"tur_gdp.rds")
```

# Data reshaping XIII

```r
f_url = "https://github.com/obakis/econ_data/raw/master/illere_gore_isgucu.xlsx"
download.file(url = f_url, destfile = "il_isgucu.xlsx", mode="wb")

dat = read.xlsx("il_isgucu.xlsx", colNames = TRUE)
head(dat)
```

```
##   pr_no        pr_name lfp_rate un_rate emp_rate year nuts3
## 1     1          Adana     49.0    26.5     36.0 2008 TR621
## 2     2       Adıyaman     38.0    17.9     31.2 2008 TRC12
## 3     3 Afyonkarahisar     44.7    10.8     39.9 2008 TR332
## 4     4           Ağrı     48.0    10.1     43.2 2008 TRA21
## 5     5         Amasya     56.2     6.9     52.4 2008 TR834
## 6     6         Ankara     44.9    13.6     38.8 2008 TR510
```

```r
dat=as_tibble(dat)
saveRDS(dat,"tur_labor.rds")
saveRDS(dat[1:81,c("pr_no","nuts3")],"province-nuts3.rds")
```

# Joining data frames I

```
##See http://dplyr.tidyverse.org/reference/join.htmlfor more on joining
tur_m = readRDS("tur_m.rds")
tur_x = readRDS("tur_x.rds")
tur_xm = full_join(tur_m, tur_x, by=c("year","province","month"))
tur_xm %>%
  arrange(year,month, province) -> tur_xm
print(tur_xm,3)
```

```
## # A tibble: 16,512 x 5
##     year province month    import   export
##    <dbl>    <dbl> <fct>     <dbl>    <dbl>
##   1  2002        1 January  44761.  35247.
##   2  2002        2 January   1868.    740.
##   3  2002        3 January   1295.   3163.
##   4  2002        4 January    680.    190.
##   5  2002        5 January    271.     19.3
```

# Joining data frames II

```
##  6  2002        6 January 358921. 118803.
##  7  2002        7 January   4623.  13904.
##  8  2002        8 January   1687.    526.
##  9  2002        9 January   5557.   9959.
## 10  2002       10 January   3799.   6538.
## # ... with 16,502 more rows
```

```
saveRDS(tur_xm, "tur_xm.rds")
```

# Joining data frames III

```r
# f_url = "https://github.com/obakis/econ_data/raw/master/tur_xm.rds"
# download.file(url = f_url, destfile = "tur_xm.rds", mode="wb")
# f_url = "https://github.com/obakis/econ_data/raw/master/tur_labor.rds"
# download.file(url = f_url, destfile = "tur_labor.rds", mode="wb")
xm = readRDS("tur_xm.rds")
lab = readRDS("tur_labor.rds")

ihs <- function(x){
  log(x + sqrt(x**2 + 1))
}

library(dplyr)
xm %>%
  group_by(province, year) %>%
  summarise(
    export = sum(export, na.rm=TRUE),
    import = sum(import, na.rm=TRUE)
```

# Joining data frames IV

```
    )  %>%
  group_by(province) %>%
  arrange(province, year) %>%
  mutate(
    ihs_x = ihs(export),
    ihs_m = ihs(import)
    ) %>%
  mutate(
    gr_x =  100*(ihs_x - dplyr::lag(ihs_x))/dplyr::lag(ihs_x),
    gr_m =  100*(ihs_m - dplyr::lag(ihs_m))/dplyr::lag(ihs_m)
    )  %>%
  rename(pr_no = province)  %>%
  mutate(
    gr_x = ifelse(is.na(gr_x) | is.infinite(gr_x), NA,gr_x),
    gr_m = ifelse(is.na(gr_m) | is.infinite(gr_m), NA,gr_m)
    )  -> xm_y
dat1 = inner_join(lab, xm_y, by=c("year","pr_no"))
```

# Joining data frames V

```
dat1 %>% select(-pr_name) -> dat
head(dat,3)
```

```
## # A tibble: 3 x 12
##   pr_no lfp_rate un_rate emp_rate  year nuts3 export import ihs_x
##   <dbl>    <dbl>   <dbl>    <dbl> <dbl> <chr>  <dbl>  <dbl> <dbl>
## 1     1     49      26.5     36    2008 TR621 1.30e6 2.15e6  14.8
## 2     2     38      17.9     31.2  2008 TRC12 5.91e4 3.63e4  11.7
## 3     3     44.7    10.8     39.9  2008 TR332 2.38e5 3.44e4  13.1
## # ... with 3 more variables: ihs_m <dbl>, gr_x <dbl>, gr_m <dbl>
```

```
saveRDS(dat, "tur_xmlab.rds")
```

```
xm %>%
  filter(year %in% c(2009,2010)) %>%
  group_by(province, year) %>%
  summarise(
```

# Joining data frames VI

```
  export = sum(export, na.rm=TRUE),
  import = sum(import, na.rm=TRUE)
  )  %>%
group_by(province) %>%
arrange(province, year) %>%
mutate(
  ihs_x = ihs(export),
  ihs_m = ihs(import)
  ) %>%
mutate(
  gr_x =  100*(ihs_x - dplyr::lag(ihs_x))/dplyr::lag(ihs_x),
  gr_m =  100*(ihs_m - dplyr::lag(ihs_m))/dplyr::lag(ihs_m)
  )  %>%
rename(pr_no = province)  %>%
mutate(
  gr_x = ifelse(is.na(gr_x) | is.infinite(gr_x), NA,gr_x),
  gr_m = ifelse(is.na(gr_m) | is.infinite(gr_m), NA,gr_m)
```

# Joining data frames VII

```
    )  -> xm_2y
xm_2y
```

```
## # A tibble: 162 x 8
## # Groups:   pr_no [81]
##    pr_no  year  export   import ihs_x ihs_m   gr_x  gr_m
##    <dbl> <dbl>   <dbl>    <dbl> <dbl> <dbl>  <dbl> <dbl>
##  1     1  2009 1135887. 1692782.  14.6  15.0 NA     NA
##  2     1  2010 1352306. 2229404.  14.8  15.3  1.19   1.83
##  3     2  2009   58091.   33336.  11.7  11.1 NA     NA
##  4     2  2010   71639.   85425.  11.9  12.0  1.80   8.47
##  5     3  2009  208636.   40512.  12.9  11.3 NA     NA
##  6     3  2010  217496.   72668.  13.0  11.9  0.321  5.17
##  7     4  2009   44339.   45227.  11.4  11.4 NA     NA
##  8     4  2010   76904.   58973.  11.9  11.7  4.83   2.33
##  9     5  2009   21629.   13072.  10.7  10.2 NA     NA
## 10     5  2010   53018.   41629.  11.6  11.3  8.40  11.4
```

# Joining data frames VIII

```
## # ... with 152 more rows
```

```
dat1 = inner_join(lab, xm_2y, by=c("year","pr_no"))
dat1 %>% select(-pr_name) -> dat2y
head(dat2y,3)
```

```
## # A tibble: 3 x 12
##   pr_no lfp_rate un_rate emp_rate  year nuts3 export import ihs_x
##   <dbl>    <dbl>   <dbl>    <dbl> <dbl> <chr>  <dbl>  <dbl> <dbl>
## 1     1     45.6    20.5     36.2  2009 TR621 1.14e6 1.69e6  14.6
## 2     2     42.1    16.5     35.1  2009 TRC12 5.81e4 3.33e4  11.7
## 3     3     44       7.7     40.6  2009 TR332 2.09e5 4.05e4  12.9
## # ... with 3 more variables: ihs_m <dbl>, gr_x <dbl>, gr_m <dbl>
```

```
saveRDS(dat2y, "tur_xmlab2y.rds")
```