supplement 2: Data management

Arho Toikka

2022-09-22

```
# Tupakan myyntipaikkatutkimus
# Salla-Maaria Patsi 04/2022


library(tidyverse)
library(readxl)
library(dplyr)
library(stringr)
library(snakecase)


#############
# Korjataan Rukan vanha postinumero ja postilaatikkoon merkityt myyntipaikat
# fix errors in post codes
Myyntilupien_listaus <- read_excel("Myyntilupien listaus.xlsx") %>%
  mutate(Postinumero = str_pad(subject_zip, width=5, side=c("left"), pad="0"))  %>%
  mutate(Postinumero = fct_recode(Postinumero, "93830"="93825"),
         Postinumero = str_replace(Postinumero, "1$", "0"))
dim(Myyntilupien_listaus)
# [1] 24554    28


# filtteraidaan aineisto niihin joilla lupa voimassa 1.1.2021
# filter to current licences (1.1.2021)
kaikki_myyntipaikat <- Myyntilupien_listaus %>%
  filter(licence_start_date < "2020-12-31 00:00:00.000000" &
             (is.na(licence_end_date) | licence_end_date > "2020-12-31 00:00:00.000000") &
                     (decision != "expired" | is.na(decision)) &
                       ((liikkeen_tyyppi != 21 & liikkeen_tyyppi != 23 &
                           liikkeen_tyyppi != 6) | is.na(liikkeen_tyyppi)))

dim(kaikki_myyntipaikat)
# [1] 5675   28

# remove duty free licences not accessible to general public
lentokenttienpostinumerot <- read_excel("tupakan_myyntipaikat_lentokenttien_postinumeroissa.xlsx") %>%
  filter(lentokenttamatkustajille == 1)

kaikki_myyntipaikat <- kaikki_myyntipaikat %>% filter(!case_id %in% lentokenttienpostinumerot$case_id)
# luokitellaan uudelleen liikkeen_tyyppi muuttuja
# classify numeric licence types as per Valvira information
kaikki_myyntipaikat <- kaikki_myyntipaikat %>%
```


Pätsi S-M, *et al. Tob Control* 2023;0:1–7. doi: 10.1136/tc-2022-057798

```
  mutate(liikkeen_tyyppi_ryhmiteltyna = liikkeen_tyyppi %>%
           as_factor() %>%
           fct_recode("Kaupat" = "1",
                      "Kaupat" = "4",
                      "Kioski_huoltoasema" = "2",
                      "Kioski_huoltoasema" = "3",
                      "Ravintolat" = "5",
                      "Erikoisliikkeet" = "7",
                      "Tupakkakauppa" = "8") %>% fct_explicit_na("Ei tiedossa"))

table(kaikki_myyntipaikat$liikkeen_tyyppi_ryhmiteltyna)
# Kaupat Kioski_huoltoasema         Ravintolat
#Erikoisliikkeet      Tupakkakauppa        Ei tiedossa
#   1502                738               1021
#119                 49               2246


kaikki_myyntipaikat %>% filter(Postinumero=="01530") -> q
dim(kaikki_myyntipaikat)
# [1] 5675   29
# -> 1 muuttuja lisaa


# erotellaan aineisto myynnissa olevien tuotteiden mukaan
# remove e-cigarettes
tupakan_myyntipisteet <- kaikki_myyntipaikat %>% filter(myynnissa_olevat_tuotteet != 2)
dim(tupakan_myyntipisteet)
# [1] 5639   29



# group_by zipcode and count types
tupakan_myyntipaikat_postinumeroittain <- tupakan_myyntipisteet %>%
  group_by(Postinumero) %>%
  add_count(liikkeen_tyyppi_ryhmiteltyna, name="liiketyypit") %>%
  mutate(n_kaupat = sum(liikkeen_tyyppi_ryhmiteltyna=="Kaupat"),
         n_kioski = sum(liikkeen_tyyppi_ryhmiteltyna=="Kioski_huoltoasema"),
         n_ravintolat = sum(liikkeen_tyyppi_ryhmiteltyna=="Ravintolat"),
         n_erikoisliikkeet = sum(liikkeen_tyyppi_ryhmiteltyna=="Erikoisliikkeet"),
         n_tupakkakauppa = sum(liikkeen_tyyppi_ryhmiteltyna=="Tupakkakauppa"),
         n_ei_tiedossa = sum(liikkeen_tyyppi_ryhmiteltyna=="Ei tiedossa")) %>%
  summarise(myyntipaikat = n(), kaupat = first(n_kaupat), kioskit = first(n_kioski),
            ravintola = first(n_ravintolat), erikoisliikkeet = first(n_erikoisliikkeet),
            tupakkakaupat = first(n_tupakkakauppa), ei_tiedossa = first(n_ei_tiedossa)) %>%
  mutate(tarkistussumma = kaupat + kioskit + ravintola + erikoisliikkeet + tupakkakaupat + ei_tiedossa)




##########
# read Statistics finland data and fix variable names
pxweb_data <- read.csv2("data2021/pxweb_data_2022.csv") %>%
  mutate(Postinumero = str_pad(Postinumero, width=5, side=c("left"), pad="0"))
```

```
dim(pxweb_data)
#[1] 3027  124

duplicates_when_cut <- str_remove(names(pxweb_data), "[\\.]+[\\.].*") %>% duplicated()
pxweb_data <- pxweb_data %>% select(-which(duplicates_when_cut)) %>%
  rename_with(~str_extract(., "[^,]+")) %>%
  rename_with(~to_snake_case(., unique_sep="toistuu"))



#######
# combine licence data with Statistics Finland
tupakan_myyntipaikat_postinumeroittain <- tupakan_myyntipaikat_postinumeroittain %>% left_join(pxweb_data, by=c
dim(tupakan_myyntipaikat_postinumeroittain)
# [1] 1379  112


# Add those locations that have no sales but do have 500+ inhabitants and replace missing prevalence with 0
ei_myyntia <- pxweb_data %>%
  anti_join(tupakan_myyntipaikat_postinumeroittain,
            by=c("postinumero"="Postinumero")) %>%
  filter(asukkaat_yhteensä_2020_he > 500)

tupakan_myyntipaikat_postinumeroittain <- tupakan_myyntipaikat_postinumeroittain %>%
  full_join(ei_myyntia %>% rename(Postinumero=postinumero))

tupakan_myyntipaikat_postinumeroittain <- tupakan_myyntipaikat_postinumeroittain %>%
  mutate(myyntipaikat = myyntipaikat %>% replace_na(0),
         kaupat = kaupat %>% replace_na(0),
         kioskit = kioskit %>% replace_na(0),
         ravintola = ravintola %>% replace_na(0),
         erikoisliikkeet = erikoisliikkeet %>% replace_na(0),
         tupakkakaupat = tupakkakaupat %>% replace_na(0),
         ei_tiedossa = ei_tiedossa %>% replace_na(0))

# Make variables from raw data

tupakan_myyntipaikat_postinumeroittain <- tupakan_myyntipaikat_postinumeroittain %>%
  mutate(# sales locations / 1000 inhabitants, also by type
    tupakan_myyntipaikkoja_per_1000_asukasta = myyntipaikat / asukkaat_yhteensä_2020_he * 1000,
         kauppoja_per_1000_asukasta =  kaupat / asukkaat_yhteensä_2020_he * 1000,
         kioskit_per_1000_asukasta =  kioskit / asukkaat_yhteensä_2020_he * 1000,
         ravintolat_per_1000_asukasta =  ravintola / asukkaat_yhteensä_2020_he * 1000,
         erikoisliikkeet_per_1000_asukasta =  erikoisliikkeet / asukkaat_yhteensä_2020_he * 1000,
         tupakkakaupat_per_1000_asukasta = tupakkakaupat / asukkaat_yhteensä_2020_he * 1000,
    # raw counts of socio-demographic variables into proportions
         alin_tuloluokka_osuus = alimpaan_tuloluokkaan_kuuluvat_asukkaat_2020_hr /
      x_18_vuotta_täyttäneet_yhteensä_2020_ko,
         tyottomienosuus = työttömät_2019_pt/
      (työlliset_2019_pt+työttömät_2019_pt),
         tyollistenosuus = työttömät_2019_pt /
      (työlliset_2019_pt+työttömät_2019_pt),
         perusasteen_osuus = perusasteen_suorittaneet_2020_ko /
```

```
      x_18_vuotta_täyttäneet_yhteensä_2020_ko,
         keskiasteen_osuus = (ylioppilastutkinnon_suorittaneet_2020_ko+
                                ammatillisen_tutkinnon_suorittaneet_2020_ko) /
      x_18_vuotta_täyttäneet_yhteensä_2020_ko,
         korkeakoulutettujen_osuus = (alemman_korkeakoulututkinnon_suorittaneet_2020_ko+
                                        ylemmän_korkeakoulututkinnon_suorittaneet_2020_ko) /
      x_18_vuotta_täyttäneet_yhteensä_2020_ko,
    # population density
    asukastiheys = asukkaat_yhteensä_2020_he / (pinta_ala/1000),
         tyopaikkaomavaraisuus = työpaikat_yhteensä_2019_tp / työlliset_2019_pt  * 100)

dim(tupakan_myyntipaikat_postinumeroittain)
# [1] 1379  126
# -> 14 muuttujaa lisaa



write.csv2(tupakan_myyntipaikat_postinumeroittain,
           "tupakan_myyntipaikat_postinumeroittain_valmis_anayysiaineisto.csv",
           row.names = F)
```