

# data wrangling in R

Sadaf Zuhra

2024-09-10

## Activating packages

```
library(readxl)
library(tidyverse)
```

## Importing dataset from excel

```
library(readxl)
df <- read_excel("_Participatory Epidemiology of Ticks and Tick-Borne Piroplasmosis along with their Th",
  range = "A1:BS1236", col_types = c("date",
    "text", "text", "text", "numeric",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "numeric", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text", "text"))
```

## understanding the data structure

```
library(tidyverse)
as.tibble(df)
```

```
## # A tibble: 1,235 x 71
##   Timestamp      'Name of respondent' 'Mobile Number' Gender 'Age (Years)'
##   <dtm>          <chr>                <chr>          <chr>      <dbl>
```

```
## 1 2024-05-14 10:27:24 Arslan Muhammad Ali~ 03346640852 Male 34
## 2 2024-05-14 11:13:51 Faisal Hafeez 03071364100 Male 24
## 3 2024-05-15 20:19:13 Hafiz Muhammad Zoha~ 0301-8576149 Male 24
## 4 2024-05-15 20:35:03 Dr.Muhammad Shafeeq~ 0334-7916654 Male 32
## 5 2024-05-15 20:44:02 Amjad Shah 03022628543 Male 22
## 6 2024-05-15 21:34:38 Muhammad Usman Nazir 03030890765 Male 29
## 7 2024-05-15 23:28:17 Hizqeel Ahmed Muzaf~ 03211231783 Male 20
## 8 2024-05-15 23:48:35 USAMA IFTIKHAR 03035673115 Male 21
## 9 2024-05-16 07:53:42 Muhammad Ishfaq 03471620460 Male 26
## 10 2024-05-16 08:14:08 Mubashar Hanif 03260077937 Male 21
## # i 1,225 more rows
## # i 66 more variables: 'Education Status' <chr>, 'Respondent Occupation' <chr>,
## # 'Occupation type' <chr>, 'Socioeconomic Status' <chr>,
## # 'For how many years have you been farming or keeping animals?' <chr>,
## # 'What is your Annual Income (PKR)' <chr>, 'Name of District' <chr>,
## # 'Name of Tehsil' <chr>, 'Name of UC-Name/Village' <chr>,
## # 'Living Area' <chr>, ...
```

```
#head(df)
#tail(df)
#glimpse(df)
#View(df)
#summary(df)
```

## Subsetting of data on the basis of Variable

```
library(tidyverse)
#df <- df[,-c(1,2,3)]
df_1<- select(df,"Gender",
  "Education Status",
  "Occupation type","Socioeconomic Status",
  "Name of District","Name of Tehsil","Living Area",
  "Based on your observations, how would you rank the ticks in the matter of frequency",
  "In which host you have observed ticks?",
  "If it was animal host then select the host species in which ticks were observed",
  "Animal Gender",
  "Health Status of animal in which ticks are observed (select all possible)",
  "What is your level of knowledge about ticks?",
  "How frequently you observe the ticks?",
  "Did you find any impact of these ticks on animal health?",
  "What is the average number of ticks you have observed last year?",
  "Hygienic Measures of the area (select all possible)",
  "How frequently you have observed theileriosis in your animals?",
  "How frequently you have observed babesiosis in animals?")

head(df)
```

```
## # A tibble: 6 x 71
##   Timestamp      'Name of respondent' 'Mobile Number' Gender 'Age (Years)'
##   <dtm>          <chr>                <chr>          <chr>      <dbl>
## 1 2024-05-14 10:27:24 Arslan Muhammad Ali ~ 03346640852 Male 34
```

```
## 2 2024-05-14 11:13:51 Faisal Hafeez 03071364100 Male 24
## 3 2024-05-15 20:19:13 Hafiz Muhammad Zohaib 0301-8576149 Male 24
## 4 2024-05-15 20:35:03 Dr.Muhammad Shafeeq ~ 0334-7916654 Male 32
## 5 2024-05-15 20:44:02 Amjad Shah 03022628543 Male 22
## 6 2024-05-15 21:34:38 Muhammad Usman Nazir 03030890765 Male 29
## # i 66 more variables: 'Education Status' <chr>, 'Respondent Occupation' <chr>,
## # 'Occupation type' <chr>, 'Socioeconomic Status' <chr>,
## # 'For how many years have you been farming or keeping animals?' <chr>,
## # 'What is your Annual Income (PKR)' <chr>, 'Name of District' <chr>,
## # 'Name of Tehsil' <chr>, 'Name of UC-Name/Village' <chr>,
## # 'Living Area' <chr>,
## # 'Based on your observations, how would you rank the ticks in the matter of frequency' <dbl>, ...
```

## renaming data columns

```
library(tidyverse)
df_1 <- df_1 %>% rename("gender" = "Gender",
  "educational_status" = "Education Status" ,
  "ocupation_types" ="Occupation type",
  "sec"="Socioeconomic Status",
  "dist_name"="Name of District",
  "tehsil_name"="Name of Tehsil",
  "living_area"="Living Area",
  "tick_freq"= "Based on your observations, how would you rank the ticks in the matter of frequency",
  "host_animal"= "In which host you have observed ticks?" ,
  "host_specie"="If it was animal host then select the host species in which ticks were observed",
  "animal_gender"="Animal Gender",
  "host_health" = "Health Status of animal in which ticks are observed (select all possible)",
  "knowledge" ="What is your level of knowledge about ticks?",
  "obs_freq"= "How frequently you observe the ticks?",
  "impact_anl_health"="Did you find any impact of these ticks on animal health?" ,
  "ave_tick_ly"= "What is the average number of ticks you have observed last year?",
  "hm"="Hygienic Measures of the area (select all possible)" ,
  "theileriosis_freq" = "How frequently you have observed theileriosis in your animals?",
  "babesiosis_freq" = "How frequently you have observed babesiosis in animals?") %>% print
```

```
## # A tibble: 1,235 x 19
##   gender educational_status occupation_types sec dist_name tehsil_name
##   <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Male Postgraduated Government Employ Middle C~ faisalab~ Faisalabad
## 2 Male Postgraduated Student Middle C~ Fsd Fsd sadar
## 3 Male Postgraduated Government Employ Middle C~ Narowal Shakergarh
## 4 Male Graduated Government Employ Middle C~ Rajan pur Jampur
## 5 Male Graduated Student Middle C~ Kasur Kasur
## 6 Male Postgraduated Private Employ Middle C~ Bahawalp~ Bahawalpur
## 7 Male Intermediate (11-12) Student Middle C~ Narowal Narowal
## 8 Male Intermediate (11-12) Student Middle C~ Gujranwa~ Wazirabad
## 9 Male Graduated Self Business Middle C~ Muzaffar~ Kotadu
## 10 Male Intermediate (11-12) Student Lower Cl~ Vehari Vehari
## # i 1,225 more rows
## # i 13 more variables: living_area <chr>, tick_freq <dbl>, host_animal <chr>,
```

```
## # host_specie <chr>, animal_gender <chr>, host_health <chr>, knowledge <chr>,
## # obs_freq <chr>, impact_animal_health <chr>, ave_tick_ly <chr>, hm <chr>,
## # theileriosis_freq <chr>, babesiosis_freq <chr>
```

## recoding data columns

```
df_1 %>%
  mutate(dist_name = recode(dist_name, "Rajan pur"="Rajanpur", "Rajan 6" = "Rajanpur", "faisalabad"="Faisalabad",
                             "Fsd"="Faisalabad")) %>%
  drop_na(dist_name) %>%
  summarise(unique(dist_name))
```

```
## # A tibble: 13 x 1
##   'unique(dist_name)'
##   <chr>
## 1 Faisalabad
## 2 Narowal
## 3 Rajanpur
## 4 Kasur
## 5 Bahawalpur
## 6 Gujranwala
## 7 Muzaffargarh
## 8 Vehari
## 9 Toba Tek Singh
## 10 Sheikupura
## 11 Lodhran
## 12 Hafizabad
## 13 Burj
```

```
df_1 <- df_1 %>%
  mutate(dist_name = recode(dist_name, "Rajan pur"="Rajanpur", "Rajan 6" = "Rajanpur", "faisalabad"="Faisalabad",
                             "Fsd"="Faisalabad")) %>%
  drop_na(dist_name)
```

## filtering rows

```
library(tidyverse)
glimpse(df_1)
```

```
## Rows: 1,232
## Columns: 19
## $ gender          <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Ma~
## $ educational_status <chr> "Postgraduated", "Postgraduated", "Postgraduated", ~
## $ occupation_types  <chr> "Government Employ", "Student", "Government Employ"~
## $ sec              <chr> "Middle Class", "Middle Class", "Middle Class", "Mi~
## $ dist_name         <chr> "Faisalabad", "Faisalabad", "Narowal", "Rajanpur", ~
## $ tehsil_name       <chr> "Faisalabad", "Fsd sadar", "Shakergarh", "Jampur", ~
```

```
## $ living_area      <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Peri ~
## $ tick_freq        <dbl> 10, 6, 10, 3, 6, 7, 6, 6, 7, 5, 8, 5, 7, 5, 7, 4~
## $ host_animal      <chr> "Domestic Animal", "Domestic Animal", "Domestic Ani~
## $ host_specie      <chr> "Cattle, Goat, Sheep, Cat, Dog, Horse, Buffalo, Don~
## $ animal_gender    <chr> "Female", "Male", "Female", "Both", "Both", "Both",~
## $ host_health      <chr> "Healthy", "Debilitated", "Healthy", "Healthy, Dise~
## $ knowledge        <chr> "High", "High", "High", "Moderate", "Moderate", "Mo~
## $ obs_freq         <chr> "In specific season only", "Frequently", "Seldom", ~
## $ impact_ani_health <chr> "Seldom become diseased", "Seldom become diseased",~
## $ ave_tick_ly      <chr> "Above 50", "up to 10", "up to 10", "11-20", "21-30~
## $ hm               <chr> "Fair", "Good", "Fair", "Fair", "Fair", "Fair", "Po~
## $ theileriosis_freq <chr> "Less frequent", "Seldom", "Less frequent", "Seldom~
## $ babesiosis_freq  <chr> "Less frequent", "Seldom", "Less frequent", "Less f~
```

```
filter(df_1, dist_name=="Rajanpur"| dist_name=="Rajan pur") %>%
  print()
```

```
## # A tibble: 437 x 19
##   gender educational_status occupation_types sec dist_name tehsil_name
##   <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Male Graduated Government Employ Middle C~ Rajanpur Jampur
## 2 Male Graduated Government Employ Middle C~ Rajanpur Jampur
## 3 Male Under Matric Private Employ Middle C~ Rajanpur Rajan pur
## 4 Male Uneducated Private Employ Middle C~ Rajanpur Rajan pur
## 5 Male Matric Self Business Middle C~ Rajanpur Rajan pur
## 6 Male Matric Self Business Middle C~ Rajanpur Rajan pur
## 7 Male Under Matric Self Business Middle C~ Rajanpur Rajan pur
## 8 Male Under Matric Private Employ Middle C~ Rajanpur Rajan pur
## 9 Male Intermediate (11-12) Private Employ Middle C~ Rajanpur Rajan pur
## 10 Male Matric Self Business Middle C~ Rajanpur Rajan pur
## # i 427 more rows
## # i 13 more variables: living_area <chr>, tick_freq <dbl>, host_animal <chr>,
## # host_specie <chr>, animal_gender <chr>, host_health <chr>, knowledge <chr>,
## # obs_freq <chr>, impact_ani_health <chr>, ave_tick_ly <chr>, hm <chr>,
## # theileriosis_freq <chr>, babesiosis_freq <chr>
```

```
filter(df_1, dist_name %in% c("Rajanpur","Rajan pur","Rajan 6")) %>%
  print()
```

```
## # A tibble: 437 x 19
##   gender educational_status occupation_types sec dist_name tehsil_name
##   <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Male Graduated Government Employ Middle C~ Rajanpur Jampur
## 2 Male Graduated Government Employ Middle C~ Rajanpur Jampur
## 3 Male Under Matric Private Employ Middle C~ Rajanpur Rajan pur
## 4 Male Uneducated Private Employ Middle C~ Rajanpur Rajan pur
## 5 Male Matric Self Business Middle C~ Rajanpur Rajan pur
## 6 Male Matric Self Business Middle C~ Rajanpur Rajan pur
## 7 Male Under Matric Self Business Middle C~ Rajanpur Rajan pur
## 8 Male Under Matric Private Employ Middle C~ Rajanpur Rajan pur
## 9 Male Intermediate (11-12) Private Employ Middle C~ Rajanpur Rajan pur
## 10 Male Matric Self Business Middle C~ Rajanpur Rajan pur
## # i 427 more rows
```

```
## # i 13 more variables: living_area <chr>, tick_freq <dbl>, host_animal <chr>,  
## #   host_specie <chr>, animal_gender <chr>, host_health <chr>, knowledge <chr>,  
## #   obs_freq <chr>, impact_animal_health <chr>, ave_tick_ly <chr>, hm <chr>,  
## #   theileriosis_freq <chr>, babesiosis_freq <chr>
```