TEB2164 Introduction to Data Science
Lab 7a- Managing data frame

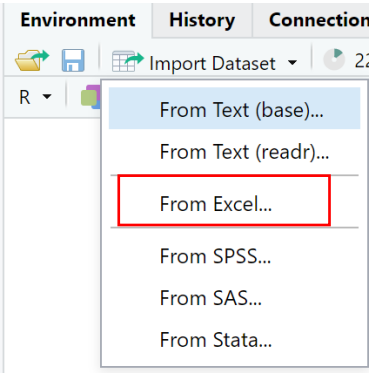
Learning Outcomes

At the end of the session, you will be able to:

- Importing data (from excel/csv to data frame)
- Exporting data (save into excel/csv)
- Managing NA/empty cells in data frame
- Filtering, searching, and arranging data using data frame

Activity

1. Importing data
   1.1. Installing library
   - Click Import Dataset>From Excel, then allow the installation of 'readxl' library:



- Or type in console:

```
install.packages("readxl")
```

- Verify package installation:

```
any(grepl("readxl",installed.packages()))
```

- To use the library:

```
library(readxl)
```

1.2. Identify and setting the folder path

- Write and run the following in R. Make your conclusion about the code:

```
# Get the working directory then store 'uforeport.xls' and
'titanic.csv' in the same folder of your R Script
getwd()

# You can also change your working directory as you wish using
setwd('C:/Users/User/Desktop') #file path may differ for everyone

# To ensure you store in the same working director, 'uforeport.xls'
and 'titanic.csv' should be listed in one of many files that you
have
list.files()
```

1.3. Importing excel file

- Write and run the following in R. Make your conclusion about the code:

```
uforeport <- read_excel("uforeport.xls")
View(uforeport)
```

1.4. Importing csv file (no library required)

- Write and run the following in R. Make your conclusion about the code:

```
titanic <- read.csv("titanic.csv")
View(titanic)
```

2. Identifying NA/empty cells in data frame
   - Write and run the following in R. Make your conclusion about the code:

```
# To count the total NA values
sum(is.na(uforeport))
sum(is.na(titanic))

# To get the positions where NA values are
which(is.na(uforeport))
which(is.na(titanic))

# To see how it work in smaller dataset, run code below
demo = c(1, 2, NA, 4, NA, 6, 7)
sum(is.na(demo))
which(is.na(demo))

# Get count of NA in each column
print(sapply(uforeport, function(x) sum(is.na(x))))
print(sapply(titanic, function(x) sum(is.na(x))))
```

3. Managing NA/empty cells in data frame
   - Write and run the following in R. Make your conclusion about the code:

```
dim(uforeport)
uforeport_cleaned = na.omit(uforeport)
dim(uforeport_cleaned)
dim(titanic)
titanic_cleaned = na.omit(titanic)
dim(titanic_cleaned)
```

4. Filtering values in data frame

4.1. Import necessary library

- Write and run the following in R. Make your conclusion about the code:

```
install.packages("dplyr")
library(dplyr)
```

4.2. Get the columns name

- Write and run the following in R. Make your conclusion about the code:

```
colnames(uforeport_cleaned)
colnames(titanic_cleaned)
```

or

```
names(uforeport_cleaned)
names(titanic_cleaned)
```

4.3. Renaming unconventional variable names

- Write and run the following in R. Make your conclusion about the code:

```
names(uforeport_cleaned)
# replace blank with underscore
names(uforeport_cleaned) = gsub(" ", "_",
colnames(uforeport_cleaned))
names(uforeport_cleaned)
```

4.4. Filtering values in data frame

- Write and run the following in R. Make your conclusion about the code:

```
print(filter(uforeport_cleaned,City == 'Belton'))
print(filter(uforeport_cleaned,Colors_Reported == 'GREEN'))

print(filter(titanic_cleaned,sex == 'female'))
print(filter(titanic_cleaned,fare > 500))
```

5. Searching in data frame

5.1. Filtering data by multiple conditions

- Write and run the following in R. Make your conclusion about the code:

```
print(filter(titanic_cleaned,sex == 'female' & fare > 500))
```

or

```
titanic_cleaned %>% filter(sex == 'female', fare > 500)
ufo_green = uforeport_cleaned %>%  filter(Colors_Reported ==
'GREEN')
```

6. Arranging values in data frame

6.1. Sorting by values in columns

- Write and run the following in R. Make your conclusion about the code:

```
# Sort by ascending order
titanic_sortbyfare = arrange(titanic_cleaned, fare)
# Sort by descending order
titanic_sortbyfare = arrange(titanic_cleaned, desc(fare))
```

7. Exporting data frame to excel and csv

- Write and run the following in R. Make your conclusion about the code:

```
# Save as EXCEL file
install.packages("xlsx")
library(xlsx)
write.xlsx(ufo_green, "ufo_green.xlsx")
# Save as CSV file
write.csv(titanic_sortbyfare, "titanic_sortbyfare.csv")
```