Homework 6

Mariia Nikitash
Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. You may work with students in both sections (375-01, -02). Only one person in the group should submit to Canvas.
Due: check on Canvas.

1. Consider the `billboard` dataset that is supplied with the tidyverse which shows the Billboard top 100 song rankings in the year 2000. Apply the tidyverse's data wrangling verbs to answer these questions. For each question, **give only the code**.

I make it tidy first:

```
table <- billboard %>% pivot_longer(cols=4:79, names_to = "Week", values_to = "Rank", values_drop_na = TRUE)
```

a) Show for each track, how many weeks it spent on the chart

```
table %>% group_by(track) %>% summarize(totalWeek= n()) %>% View()
```

b) List tracks in decreasing order of number of weeks spent on the chart

```
table %>% group_by(track) %>% summarize(totalWeek= n()) %>% arrange(-totalWeek) %>% View()
```

c) Show for each track, its top rank

```
table %>% group_by(track) %>% summarize(TopRank= max(Rank)) %>% View()
```

d) List tracks in increasing order of its top rank

```
table %>% group_by(track) %>% summarize(TopRank= max(Rank)) %>% arrange(TopRank) %>% View()
```

e) Show for each artist, their top rank

```
table %>% group_by(artist) %>% summarize(TopRank= max(Rank)) %>% View()
```

f) List artists in increasing order of their top rank

```
table %>% group_by(artist) %>% summarize(TopRank= max(Rank)) %>% arrange(TopRank) %>% View()
```

g) List tracks that spent more than 35 weeks in the charts

```
table %>% group_by(track) %>% summarize(totalWeek= n()) %>% filter(totalWeek > 35) %>% View()
```

h) List tracks that spent more than 35 weeks in the charts along with their artists
table %>% group_by(artist, track) %>% summarize(totalWeek= n()) %>%
filter(totalWeek > 35) %>% View()

Hint: ***First***, **convert to a tidy table**. **Show code first for this step**. All the above questions can then be answered with a single data pipeline.

2. The `demographics.csv`[1] file (available in the Datasets module on Canvas) gives the proportion of a country's population in different age groups and some other demographic data such as mortality rates and expected lifetime.

(a) The data is not "tidy". In 2-3 sentences, explain why.
This data is not "tidy" since not every variable forms a column. Series Code variable consists of other variables, which should form another columns.

(b) Transform the table to tidy data with one country per row. [Give code]
demographics %>% select(-`Series Name`) %>% pivot_wider(names_from = "Series Code", values_from = "YR2015") %>% View()

(c) Add the male/female population numbers together (i.e., ignore sex-related differences). [Hint: You will have to mutate for every pair of columns, e.g., mutate(SP.POP.0014.IN=SP.POP.0014.MA.IN+SP.POP.0014.FE.IN] [Give code]

demographics %>% select(-`Series Name`) %>% pivot_wider(names_from = "Series Code", values_from = "YR2015") %>%
mutate(SP.POP.80UP=SP.POP.80UP.MA+SP.POP.80UP.FE, SP.POP.65UP.IN= SP.POP.65UP.FE.IN+SP.POP.65UP.MA.IN,
SP.DYN.AMRT=SP.DYN.AMRT.MA+SP.DYN.AMRT.FE,SP.POP.0014.IN=SP.POP.0014.MA.IN+ SP.POP.0014.FE.IN, SP.POP.1564.IN=SP.POP.1564.FE.IN+SP.POP.1564.MA.IN,
SP.POP.TOTL.IN=SP.POP.TOTL.MA.IN+SP.POP.TOTL.MA.IN) %>%
select(-'SP.POP.65UP.FE.IN', -'SP.POP.65UP.MA.IN', -'SP.POP.80UP.MA', -'SP.POP.80UP.FE', -'SP.POP.1564.FE.IN', -'SP.POP.1564.MA.IN', -'SP.POP.0014.MA.IN',
-'SP.POP.0014.FE.IN',-'SP.DYN.AMRT.MA', -'SP.DYN.AMRT.FE',
-'SP.POP.TOTL.MA.IN',-'SP.POP.TOTL.FE.IN')%>% View()

Note: the Series Name column is included only to show the meaning of the Series Code values. This column can be dropped. At the end, the data should look as below (only part of the data is shown).

[1] Original dataset:
https://databank.worldbank.org/source/population-estimates-and-projections/Type/TABLE/preview/on#

Country Name	Country Code	SP.DYN.LE00.IN	SP.URB.TOTL	SP.POP.TOTL	SP.POP.80UP	SP.POP.1564.IN	SP.POP.0014.IN	SP.DYN.AMRT	SP.POP.TOTL.IN	SP.POP.65UP.IN
Afghanistan	AFG	63.37700	8535606	34413603	85552	18116800	15443807	455.4700	34413603	852996
Albania	ALB	78.02500	1654503	2880703	66965	1979175	537788	150.4100	2880703	363740
Algeria	DZA	76.09000	28146511	39728025	453741	25993589	11404930	191.6310	39728025	2329506
American Samoa	ASM	NA	48689	55812	NA	NA	NA	NA	NA	NA
Andorra	AND	NA	68919	78011	NA	NA	NA	NA	NA	NA
Angola	AGO	59.39800	17691524	27884381	69363	14113726	13136043	485.9310	27884381	634612

(d) Write code to show the top 5 countries with the lowest proportion of the population below 14 years old (i.e., SP.POP.0014.IN/SP.POP.TOTL) [Code, and list of 5 countries]

```
demo %>% select(-'SP.DYN.LE00.IN', -'SP.URB.TOTL',
-'SP.POP.80UP', -'SP.POP.1564.IN', -'SP.DYN.AMRT',
-'SP.POP.TOTL.IN', -'SP.POP.65UP.IN') %>%
mutate(Lowest.Proportion.Under.14 =
SP.POP.0014.IN/SP.POP.TOTL) %>% select(-'SP.POP.0014.IN',
-'SP.POP.TOTL', -'Country Code') %>%
arrange(Lowest.Proportion.Under.14) %>% slice_head(n=5)%>%
View()
> demo %>% select(-'SP.DYN.LE00.IN', -'SP.URB.TOTL',
-'SP.POP.80UP', -'SP.POP.1564.IN', -'SP.DYN.AMRT',
-'SP.POP.TOTL.IN', -'SP.POP.65UP.IN') %>%
mutate(Lowest.Proportion.Under.14 =
SP.POP.0014.IN/SP.POP.TOTL) %>% select(-'SP.POP.0014.IN',
-'SP.POP.TOTL', -'Country Code') %>%
arrange(Lowest.Proportion.Under.14) %>% slice(1:5)%>% View()
```