STAT 545A Assignment 04: Tidy data and joins

```
library(tidyverse)
library(ggplot2)
library(gapminder)
library(gridExtra)
library(grid)
library(spelling)
knitr::opts_chunk$set(echo = TRUE)
```

Exercise 1: Univariate Data Reshaping (30%)

Univariate Option 1: 1. Make a tibble with one row per year, and columns for life expectancy for two or more countries.

- Consider the Life expectancy data for North America:

```
subset <- gapminder %>%
  filter(country %in% c("Canada", "Mexico", "United States")) %>%
  select(country, year, lifeExp)

subset %>%
   knitr::kable(align = "c", caption = "Life Expectancy Per Year Selected Countries")
```

Table 1: Life Expectancy Per Year Selected Countries

country	year	lifeExp
Canada	1952	68.750
Canada	1957	69.960
Canada	1962	71.300
Canada	1967	72.130
Canada	1972	72.880
Canada	1977	74.210
Canada	1982	75.760
Canada	1987	76.860
Canada	1992	77.950
Canada	1997	78.610
Canada	2002	79.770
Canada	2007	80.653
Mexico	1952	50.789
Mexico	1957	55.190
Mexico	1962	58.299
Mexico	1967	60.110
Mexico	1972	62.361
Mexico	1977	65.032
Mexico	1982	67.405
Mexico	1987	69.498
Mexico	1992	71.455
Mexico	1997	73.670
Mexico	2002	74.902
Mexico	2007	76.195
United States	1952	68.440

country	year	lifeExp
United States	1957	69.490
United States	1962	70.210
United States	1967	70.760
United States	1972	71.340
United States	1977	73.380
United States	1982	74.650
United States	1987	75.020
United States	1992	76.090
United States	1997	76.810
United States	2002	77.310
United States	2007	78.242

- The following table contains the values of Life expectancy in a different column of each country by years in rows:

```
(subset_wide <- subset %>%
  pivot_wider(id_cols = year,
              names_from = country,
              values_from = lifeExp))
```

```
## # A tibble: 12 x 4
##     year Canada Mexico `United States`
##    <int>  <dbl>  <dbl>           <dbl>
##  1  1952   68.8   50.8            68.4
##  2  1957   70.0   55.2            69.5
##  3  1962   71.3   58.3            70.2
##  4  1967   72.1   60.1            70.8
##  5  1972   72.9   62.4            71.3
##  6  1977   74.2   65.0            73.4
##  7  1982   75.8   67.4            74.6
##  8  1987   76.9   69.5            75.0
##  9  1992   78.0   71.5            76.1
## 10  1997   78.6   73.7            76.8
## 11  2002   79.8   74.9            77.3
## 12  2007   80.7   76.2            78.2
```

```
subset_wide %>%
   knitr::kable(align = "c", caption = "Life Expectancy Per Year Selected Countries")
```

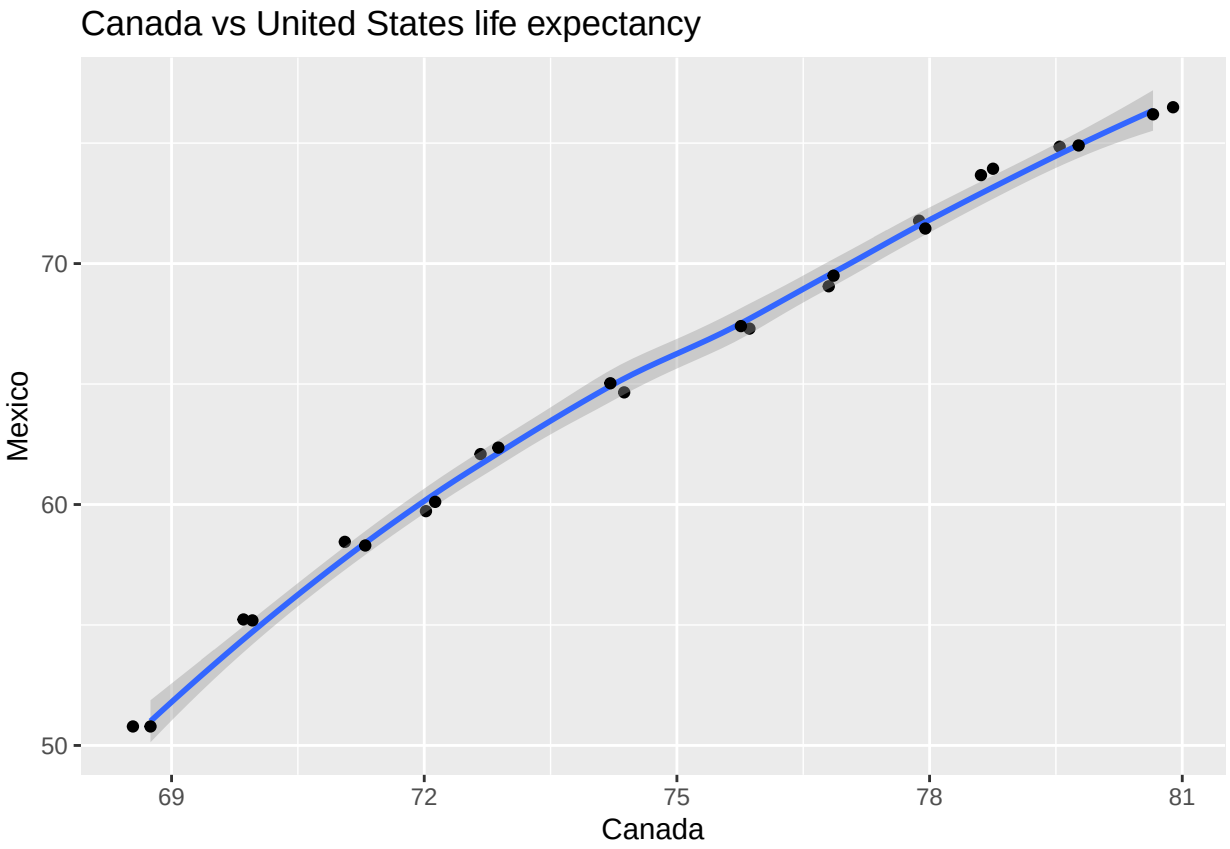
Table 2: Life Expectancy Per Year Selected Countries

year	Canada	Mexico	United States
1952	68.750	50.789	68.440
1957	69.960	55.190	69.490
1962	71.300	58.299	70.210
1967	72.130	60.110	70.760
1972	72.880	62.361	71.340
1977	74.210	65.032	73.380
1982	75.760	67.405	74.650
1987	76.860	69.498	75.020
1992	77.950	71.455	76.090
1997	78.610	73.670	76.810
2002	79.770	74.902	77.310

year	Canada	Mexico	United States
2007	80.653	76.195	78.242

2. Take advantage of this new data shape to scatterplot life expectancy for one country against that of another.

```
(subset_wide %>%
  ggplot(aes(Canada, Mexico)) +
  geom_jitter() +
  geom_smooth(method = "loess") +
  geom_point() +
  ggtitle("Canada vs United States life expectancy"))
```



3. Re-lengthen the data.

```
subset_wide %>%
  pivot_longer(cols = c(-year),
               names_to  = "country",
               values_to = "LifeExp") %>%
   knitr::kable(align = "c", caption = "Life Expectancy Per Year Selected Countries")
```

Table 3: Life Expectancy Per Year Selected Countries

year	country	LifeExp
1952	Canada	68.750
1952	Mexico	50.789
1952	United States	68.440

year	country	LifeExp
1957	Canada	69.960
1957	Mexico	55.190
1957	United States	69.490
1962	Canada	71.300
1962	Mexico	58.299
1962	United States	70.210
1967	Canada	72.130
1967	Mexico	60.110
1967	United States	70.760
1972	Canada	72.880
1972	Mexico	62.361
1972	United States	71.340
1977	Canada	74.210
1977	Mexico	65.032
1977	United States	73.380
1982	Canada	75.760
1982	Mexico	67.405
1982	United States	74.650
1987	Canada	76.860
1987	Mexico	69.498
1987	United States	75.020
1992	Canada	77.950
1992	Mexico	71.455
1992	United States	76.090
1997	Canada	78.610
1997	Mexico	73.670
1997	United States	76.810
2002	Canada	79.770
2002	Mexico	74.902
2002	United States	77.310
2007	Canada	80.653
2007	Mexico	76.195
2007	United States	78.242

Exercise 2: Multivariate Data Reshaping (30%)

Multivariate Option 1 1. Make a tibble with one row per year, and columns for life expectancy and GDP per capita (or two other numeric variables) for two or more countries.

```
subset2 <- gapminder %>%
  filter(country %in% c("Chile", "Argentina", "Peru")) %>%
  select(country, year, lifeExp, gdpPercap)

subset2_wide <- subset2  %>%
   pivot_wider(id_cols     = year,
               names_from  = country,
               names_sep   = "_",
               values_from = c(lifeExp, gdpPercap))

subset2_wide %>%
knitr::kable(align = "c", caption = "Life Expectancy & GDP percapita by Year Selected Countries")
```

Table 4: Life Expectancy & GDP percapita by Year Selected Countries

year	lifeExp_Argentina	lifeExp_Chile	lifeExp_Peru	gdpPercap_Argentina	gdpPercap_Chile	gdpPercap_Peru
1952	62.485	54.745	43.902	5911.315	3939.979	3758.523
1957	64.399	56.074	46.263	6856.856	4315.623	4245.257
1962	65.142	57.924	49.096	7133.166	4519.094	4957.038
1967	65.634	60.523	51.445	8052.953	5106.654	5788.093
1972	67.065	63.441	55.448	9443.039	5494.024	5937.827
1977	68.481	67.052	58.447	10079.027	4756.764	6281.291
1982	69.942	70.565	61.406	8997.897	5095.666	6434.502
1987	70.774	72.492	64.134	9139.671	5547.064	6360.943
1992	71.868	74.126	66.458	9308.419	7596.126	4446.381
1997	73.275	75.816	68.386	10967.282	10118.053	5838.348
2002	74.340	77.860	69.906	8797.641	10778.784	5909.020
2007	75.320	78.553	71.421	12779.380	13171.639	7408.906

2. Re-lengthen the data.

```
(subset2_wide %>%
  pivot_longer(cols      = c(-year),
               names_to  = c(".value", "country"),
               names_sep = "_")) %>%
knitr::kable(align = "c", caption = "Life Expectancy & GDP percapita by Year Selected Countries")
```

Table 5: Life Expectancy & GDP percapita by Year Selected Countries

year	country	lifeExp	gdpPercap
1952	Argentina	62.485	5911.315
1952	Chile	54.745	3939.979
1952	Peru	43.902	3758.523
1957	Argentina	64.399	6856.856
1957	Chile	56.074	4315.623
1957	Peru	46.263	4245.257
1962	Argentina	65.142	7133.166
1962	Chile	57.924	4519.094
1962	Peru	49.096	4957.038
1967	Argentina	65.634	8052.953
1967	Chile	60.523	5106.654
1967	Peru	51.445	5788.093
1972	Argentina	67.065	9443.039
1972	Chile	63.441	5494.024
1972	Peru	55.448	5937.827
1977	Argentina	68.481	10079.027
1977	Chile	67.052	4756.764
1977	Peru	58.447	6281.291
1982	Argentina	69.942	8997.897
1982	Chile	70.565	5095.666
1982	Peru	61.406	6434.502
1987	Argentina	70.774	9139.671
1987	Chile	72.492	5547.064
1987	Peru	64.134	6360.943
1992	Argentina	71.868	9308.419

year	country	lifeExp	gdpPercap
1992	Chile	74.126	7596.126
1992	Peru	66.458	4446.381
1997	Argentina	73.275	10967.282
1997	Chile	75.816	10118.053
1997	Peru	68.386	5838.348
2002	Argentina	74.340	8797.641
2002	Chile	77.860	10778.784
2002	Peru	69.906	5909.020
2007	Argentina	75.320	12779.380
2007	Chile	78.553	13171.639
2007	Peru	71.421	7408.906

Exercise 3: Table Joins (30%)

```
guest <- read_csv("https://raw.githubusercontent.com/STAT545-UBC/Classroom/master/data/wedding/attend.cs
```

```
## Parsed with column specification:
## cols(
##   party = col_double(),
##   name = col_character(),
##   meal_wedding = col_character(),
##   meal_brunch = col_character(),
##   attendance_wedding = col_character(),
##   attendance_brunch = col_character(),
##   attendance_golf = col_character()
## )
```

```
email <- read_csv("https://raw.githubusercontent.com/STAT545-UBC/Classroom/master/data/wedding/emails.cs
```

```
## Parsed with column specification:
## cols(
##   guest = col_character(),
##   email = col_character()
## )
```

```
guest %>%
knitr::kable(align = "c", caption = "Guest List")
```

Table 6: Guest List

party	name	meal_wedding	meal_brunch	attendance_wedding	attendance_brunch	attendance_g
1	Sommer Medrano	PENDING	PENDING	PENDING	PENDING	PENDING
1	Phillip Medrano	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRME
1	Blanka Medrano	chicken	Menu A	CONFIRMED	CONFIRMED	CONFIRME
1	Emaan Medrano	PENDING	PENDING	PENDING	PENDING	PENDING
2	Blair Park	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRME
2	Nigel Webb	NA	NA	CANCELLED	CANCELLED	CANCELLE
3	Sinead English	PENDING	PENDING	PENDING	PENDING	PENDING
4	Ayra Marks	vegetarian	Menu B	PENDING	PENDING	PENDING
5	Atlanta Connolly	PENDING	PENDING	PENDING	PENDING	PENDING
5	Denzel Connolly	fish	Menu B	CONFIRMED	CONFIRMED	CONFIRME
5	Chanelle Shah	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRME

party	name	meal_wedding	meal_brunch	attendance_wedding	attendance_brunch	attendance_g
6	Jolene Welsh	NA	NA	CANCELLED	CANCELLED	CANCELLE
6	Hayley Booker	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRME
7	Amayah Sanford	NA	PENDING	CANCELLED	PENDING	PENDING
7	Erika Foley	PENDING	PENDING	PENDING	PENDING	PENDING
8	Ciaron Acosta	PENDING	Menu A	PENDING	PENDING	PENDING
9	Diana Stuart	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRME
10	Cosmo Dunkley	PENDING	PENDING	PENDING	PENDING	PENDING
11	Cai Mcdaniel	fish	Menu C	CONFIRMED	CONFIRMED	CONFIRME
12	Daisy-May Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRME
12	Martin Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING
12	Violet Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING
12	Nazifa Caldwell	chicken	PENDING	PENDING	PENDING	PENDING
12	Eric Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRME
13	Rosanna Bird	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRME
13	Kurtis Frost	PENDING	PENDING	PENDING	PENDING	PENDING
14	Huma Stokes	NA	NA	CANCELLED	CANCELLED	CANCELLE
14	Samuel Rutledge	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRME
15	Eddison Collier	PENDING	PENDING	PENDING	PENDING	PENDING
15	Stewart Nicholls	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRME

```
email %>%
knitr::kable(align = "c", caption = "Email List")
```

Table 7: Email List

guest	email
Sommer Medrano, Phillip Medrano, Blanka Medrano, Emaan Medrano	sommm@gmail.com
Blair Park, Nigel Webb	bpark@gmail.com
Sinead English	singlish@hotmail.ca
Ayra Marks	marksa42@gmail.com
Jolene Welsh, Hayley Booker	jw1987@hotmail.com
Amayah Sanford, Erika Foley	erikaaaaaa@gmail.com
Ciaron Acosta	shining_ciaron@gmail.com
Diana Stuart	doodledianastu@gmail.com
Daisy-May Caldwell, Martin Caldwell, Violet Caldwell, Nazifa Caldwell, Eric Caldwell	caldwellfamily5212@gmail.com
Rosanna Bird, Kurtis Frost	rosy1987b@gmail.com
Huma Stokes, Samuel Rutledge	humastokes@gmail.com
Eddison Collier, Stewart Nicholls	eddison.collier@gmail.com
Turner Jones	tjjones12@hotmail.ca
Albert Marshall, Vivian Marshall	themarshallfamily1234@gmail.com

3.1 (10%)

1. For each guest in the guestlist (guest tibble), add a column for email address, which can be found in the email tibble.

- First I have to fix the guess names in the email file and rename the variable guest to name:

```
email_list <- email %>%
  separate_rows(guest, sep = ", ") %>%
   rename(name = guest)
```

- Then I can merge the two files: guest and emails, with guest names as identifier

```
guest %>%
  left_join(email_list, by = "name")
```

```
## # A tibble: 30 x 8
##    party name  meal_wedding meal_brunch attendance_wedd~ attendance_brun~
##    <dbl> <chr> <chr>        <chr>       <chr>            <chr>
##  1     1 Somm~ PENDING      PENDING     PENDING          PENDING
##  2     1 Phil~ vegetarian   Menu C      CONFIRMED        CONFIRMED
##  3     1 Blan~ chicken      Menu A      CONFIRMED        CONFIRMED
##  4     1 Emaa~ PENDING      PENDING     PENDING          PENDING
##  5     2 Blai~ chicken      Menu C      CONFIRMED        CONFIRMED
##  6     2 Nige~ <NA>         <NA>        CANCELLED        CANCELLED
##  7     3 Sine~ PENDING      PENDING     PENDING          PENDING
##  8     4 Ayra~ vegetarian   Menu B      PENDING          PENDING
##  9     5 Atla~ PENDING      PENDING     PENDING          PENDING
## 10     5 Denz~ fish         Menu B      CONFIRMED        CONFIRMED
## # ... with 20 more rows, and 2 more variables: attendance_golf <chr>,
## #   email <chr>
```

3.2 (10%)

Who do we have emails for, yet are not on the guestlist?

```
email_list %>%
  anti_join(guest, by = "name")
```

```
## # A tibble: 3 x 2
##   name            email
##   <chr>           <chr>
## 1 Turner Jones    tjjones12@hotmail.ca
## 2 Albert Marshall themarshallfamily1234@gmail.com
## 3 Vivian Marshall themarshallfamily1234@gmail.com
```

3.3 (10%)

Make a guestlist that includes everyone we have emails for (in addition to those on the original guestlist).

```
guest %>%
  right_join(email_list, by = "name")
```

```
## # A tibble: 28 x 8
##    party name  meal_wedding meal_brunch attendance_wedd~ attendance_brun~
##    <dbl> <chr> <chr>        <chr>       <chr>            <chr>
##  1     1 Somm~ PENDING      PENDING     PENDING          PENDING
##  2     1 Phil~ vegetarian   Menu C      CONFIRMED        CONFIRMED
##  3     1 Blan~ chicken      Menu A      CONFIRMED        CONFIRMED
##  4     1 Emaa~ PENDING      PENDING     PENDING          PENDING
##  5     2 Blai~ chicken      Menu C      CONFIRMED        CONFIRMED
##  6     2 Nige~ <NA>         <NA>        CANCELLED        CANCELLED
##  7     3 Sine~ PENDING      PENDING     PENDING          PENDING
##  8     4 Ayra~ vegetarian   Menu B      PENDING          PENDING
```

```
##  9     6 Jole~ <NA>         <NA>        CANCELLED        CANCELLED
## 10     6 Hayl~ vegetarian   Menu C      CONFIRMED        CONFIRMED
## # ... with 18 more rows, and 2 more variables: attendance_golf <chr>,
## #   email <chr>
```