HW04: Tidy data and joins

Carleena Ortega

11/10/2019

Contents

Exercise 1: Univariate Data Reshaping

Choose *EITHER* "Univariate Option 1" or "Univariate Option 2". Both of these problems have three components:

1. Putting data in wider format
2. Producing a plot of the wide data
3. Re-lengthening the wider data

Univariate Option 1

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

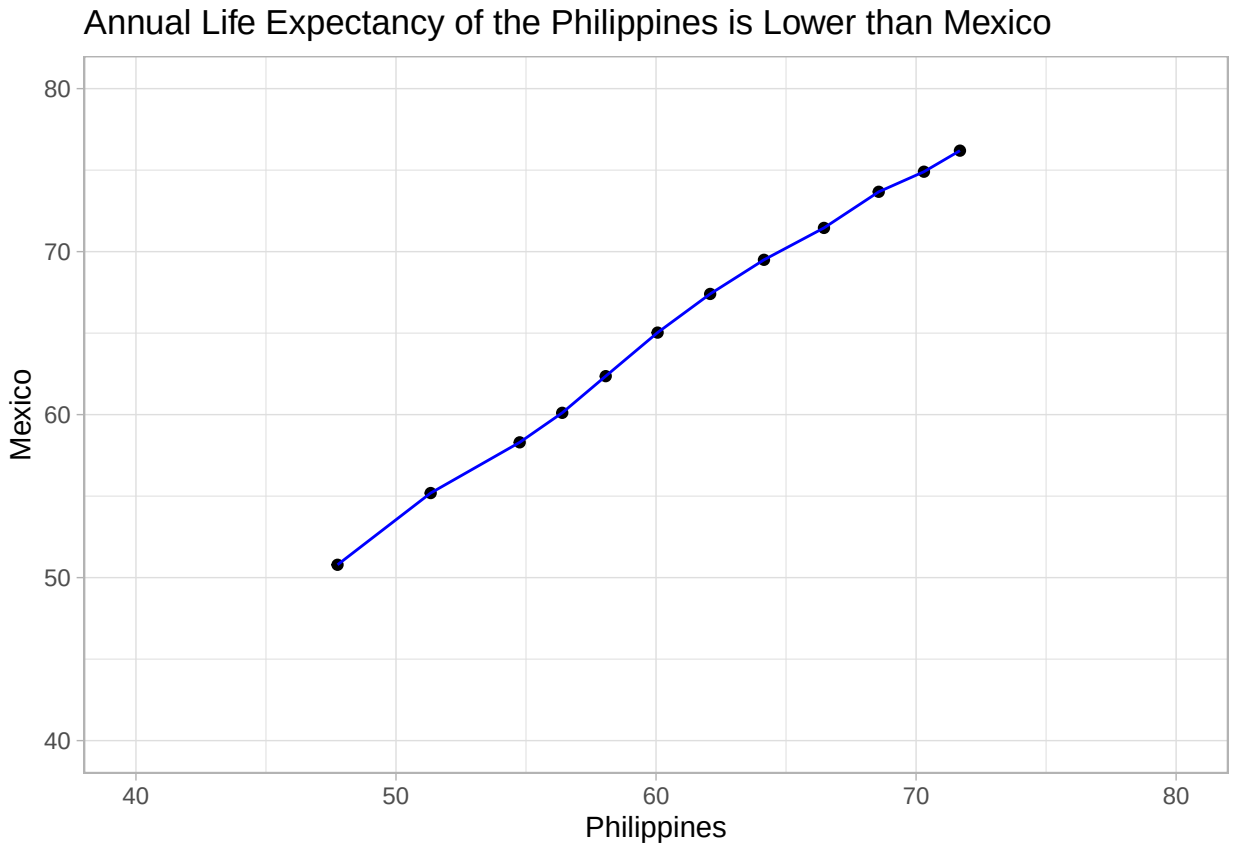
```
t1<-gapminder %>%
    select(year,country,lifeExp) %>%
    filter(country=="Philippines"|country=="Mexico") %>%
    group_by(year) %>%
  pivot_wider(names_from="country",values_from=c("lifeExp"))

knitr::kable(t1) %>%
  kable_styling("striped",latex_options="basic",
    full_width=FALSE,position="center")
```

year	Mexico	Philippines
1952	50.789	47.752
1957	55.190	51.334
1962	58.299	54.757
1967	60.110	56.393
1972	62.361	58.065
1977	65.032	60.060
1982	67.405	62.082
1987	69.498	64.151
1992	71.455	66.458
1997	73.670	68.564
2002	74.902	70.303
2007	76.195	71.688

This widened table makes it easier to compare the life expectancies of Mexico and the Philippines compared with the original untidy data. We can see that the life expectancy of Filipinos is lower than Mexicans.

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

```
t1 %>%
  ggplot(aes(Philippines,Mexico)) +
  geom_point()+ xlim(40,80)+ ylim(40,80)+
  geom_line(colour="blue")+
  labs(title="Annual Life Expectancy of the Philippines is Lower than Mexico")+
  theme_light()
```



The scatter plot comparing the life expectancy of the Philippines and Mexico show that both countries have an increasing life expectancy over the years. However, the life expectancy of the Philippines is lower.

3. Re-lengthen the data.

```
t1_relength<-t1 %>%
  pivot_longer(cols = c(-year),
               names_to  = "continent",
               values_to = "lifeExp") %>%
  arrange(continent)

  knitr::kable(t1_relength) %>%
    kable_styling("striped",latex_options="basic",font_size=8,
      full_width=FALSE, position="center")
```

year	continent	lifeExp
1952	Mexico	50.789
1957	Mexico	55.190
1962	Mexico	58.299
1967	Mexico	60.110
1972	Mexico	62.361
1977	Mexico	65.032
1982	Mexico	67.405
1987	Mexico	69.498
1992	Mexico	71.455
1997	Mexico	73.670
2002	Mexico	74.902
2007	Mexico	76.195
1952	Philippines	47.752
1957	Philippines	51.334
1962	Philippines	54.757
1967	Philippines	56.393
1972	Philippines	58.065
1977	Philippines	60.060
1982	Philippines	62.082
1987	Philippines	64.151
1992	Philippines	66.458
1997	Philippines	68.564
2002	Philippines	70.303
2007	Philippines	71.688

We've succesfully re-lengthened the previously widened data by using pivot_longer.

~~**Univariate Option 2**~~

1. Compute some measure of life expectancy (mean? median? min? max?) for all possible combinations of continent and year. Reshape that to have one row per year and one variable for each continent. Or the other way around: one row per continent and one variable per year.

2. Is there a plot that is easier to make with the data in this shape versis the usual form? Try making such a plot!

3. Re-lengthen the data.

Exercise 2: Multivariate Data Reshaping

Choose *EITHER* "Multivariate Option 1" or "Multivariate Option 2". All of these problems have two components:

1. Putting data in wider format
2. Re-lengthening the data

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.

```
t2<-gapminder %>%
    select(year,country,lifeExp,gdpPercap) %>%
    filter(country=="Canada"|country=="Japan") %>%
  pivot_wider(names_from="country",names_sep="_",
     values_from=c("lifeExp","gdpPercap"))

  knitr::kable(t2) %>%
  kable_styling("striped",latex_options = "basic",
    full_width = FALSE,position="center")
```

year	lifeExp_Canada	lifeExp_Japan	gdpPercap_Canada	gdpPercap_Japan
1952	68.750	63.030	11367.16	3216.956
1957	69.960	65.500	12489.95	4317.694
1962	71.300	68.730	13462.49	6576.649
1967	72.130	71.430	16076.59	9847.789
1972	72.880	73.420	18970.57	14778.786
1977	74.210	75.380	22090.88	16610.377
1982	75.760	77.110	22898.79	19384.106
1987	76.860	78.670	26626.52	22375.942
1992	77.950	79.360	26342.88	26824.895
1997	78.610	80.690	28954.93	28816.585
2002	79.770	82.000	33328.97	28604.592
2007	80.653	82.603	36319.24	31656.068

The widened table enables us to easily compare the life expectancy and the GDP per capita between Canada and Japan.

2. Re-lengthen the data.

```
t2_longer<-t2 %>%
pivot_longer(cols=c(-year),names_to=c(".value","Country"),names_sep="_")

knitr::kable(t2_longer) %>%
  kable_styling("striped",latex_options = "basic",
    full_width = FALSE,position="center")
```

year	Country	lifeExp	gdpPercap
1952	Canada	68.750	11367.161
1952	Japan	63.030	3216.956
1957	Canada	69.960	12489.950
1957	Japan	65.500	4317.694
1962	Canada	71.300	13462.486
1962	Japan	68.730	6576.649
1967	Canada	72.130	16076.588
1967	Japan	71.430	9847.789
1972	Canada	72.880	18970.571
1972	Japan	73.420	14778.786
1977	Canada	74.210	22090.883
1977	Japan	75.380	16610.377
1982	Canada	75.760	22898.792
1982	Japan	77.110	19384.106
1987	Canada	76.860	26626.515
1987	Japan	78.670	22375.942
1992	Canada	77.950	26342.884
1992	Japan	79.360	26824.895
1997	Canada	78.610	28954.926
1997	Japan	80.690	28816.585
2002	Canada	79.770	33328.965
2002	Japan	82.000	28604.592
2007	Canada	80.653	36319.235
2007	Japan	82.603	31656.068

We have successfully returned the widened data to its original format.

~~**Multivariate Option 2**~~

1. Compute some measure of life expectancy and GDP per capita (or two other numeric variables) (mean? median? min? max?) for all possible combinations of continent and year. Reshape that to have one row per year and one variable for each continent-measurement combination. Or the other way around: one row per continent and one variable for each year-measurement combination. 2.Re-lengthen the data.

Exercise 3: Table Joins

Do *ALL* of the activities in this section.

Read in the made-up wedding guestlist and email addresses using the following lines (go ahead and copy-paste these):

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

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

The guestlist is as follows:

```
guest %>%
  knitr::kable() %>%
  kable_styling("striped",latex_options = "basic",
    full_width = FALSE,position="center")
```

party	name	meal_wedding	meal_brunch	attendance_wedding	attendance_brunch	attendance_
1	Sommer Medrano	PENDING	PENDING	PENDING	PENDING	PENDING
1	Phillip Medrano	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
1	Blanka Medrano	chicken	Menu A	CONFIRMED	CONFIRMED	CONFIRMI
1	Emaan Medrano	PENDING	PENDING	PENDING	PENDING	PENDING
2	Blair Park	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
2	Nigel Webb	NA	NA	CANCELLED	CANCELLED	CANCELLI
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	CONFIRMI
5	Chanelle Shah	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
6	Jolene Welsh	NA	NA	CANCELLED	CANCELLED	CANCELLI
6	Hayley Booker	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
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	CONFIRMI
10	Cosmo Dunkley	PENDING	PENDING	PENDING	PENDING	PENDING
11	Cai Mcdaniel	fish	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
12	Daisy-May Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMI
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	CONFIRMI
13	Rosanna Bird	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
13	Kurtis Frost	PENDING	PENDING	PENDING	PENDING	PENDING
14	Huma Stokes	NA	NA	CANCELLED	CANCELLED	CANCELLI
14	Samuel Rutledge	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMI
15	Eddison Collier	PENDING	PENDING	PENDING	PENDING	PENDING
15	Stewart Nicholls	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMI

The email addresses are as displayed:

```
email %>%
  knitr::kable() %>%
  kable_styling("striped",latex_options = "basic",
    full_width = FALSE,position="center")
```

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

Then, complete the following tasks using the tidyverse (tidyr, dplyr, ...). No need to do any pivoting – feel free to leave guest in its current format.

3.1 Add emails

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

```
email<-separate_rows(email,guest,sep=",")
# this separates the guests' names within the same party into its own row (unlike the table above)

email[,1] <- trimws(email[,1])
#we trim the white spaces between the guests since it was hindering the proper joining of the tables

left_join(guest,email,by=c("name"="guest")) %>%
knitr::kable() %>%
kable_styling("striped", latex_options = "scale_down",full_width = FALSE)
```

party	name	meal_wedding	meal_brunch	attendance_wedding	attendance_brunch	attendance_golf	email
1	Sommer Medrano	PENDING	PENDING	PENDING	PENDING	PENDING	sommm@gmail.com
1	Phillip Medrano	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	sommm@gmail.com
1	Blanka Medrano	chicken	Menu A	CONFIRMED	CONFIRMED	CONFIRMED	sommm@gmail.com
1	Emaan Medrano	PENDING	PENDING	PENDING	PENDING	PENDING	sommm@gmail.com
2	Blair Park	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	bpark@gmail.com
2	Nigel Webb	NA	NA	CANCELLED	CANCELLED	CANCELLED	bpark@gmail.com
3	Sinead English	PENDING	PENDING	PENDING	PENDING	PENDING	singlish@hotmail.ca
4	Ayra Marks	vegetarian	Menu B	PENDING	PENDING	PENDING	marksa42@gmail.com
5	Atlanta Connolly	PENDING	PENDING	PENDING	PENDING	PENDING	NA
5	Denzel Connolly	fish	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	NA
5	Chanelle Shah	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	NA
6	Jolene Welsh	NA	NA	CANCELLED	CANCELLED	CANCELLED	jw1987@hotmail.com
6	Hayley Booker	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	jw1987@hotmail.com
7	Amayah Sanford	NA	PENDING	CANCELLED	PENDING	PENDING	erikaaaaaaa@gmail.com
7	Erika Foley	PENDING	PENDING	PENDING	PENDING	PENDING	erikaaaaaaa@gmail.com
8	Ciaron Acosta	PENDING	Menu A	PENDING	PENDING	PENDING	shining_ciaron@gmail.com
9	Diana Stuart	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	doodledianastu@gmail.com
10	Cosmo Dunkley	PENDING	PENDING	PENDING	PENDING	PENDING	NA
11	Cai Mcdaniel	fish	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	NA
12	Daisy-May Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	caldwellfamily5212@gmail.com
12	Martin Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING	caldwellfamily5212@gmail.com
12	Violet Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING	caldwellfamily5212@gmail.com
12	Nazifa Caldwell	chicken	PENDING	PENDING	PENDING	PENDING	caldwellfamily5212@gmail.com
12	Eric Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	caldwellfamily5212@gmail.com
13	Rosanna Bird	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	rosy1987b@gmail.com
13	Kurtis Frost	PENDING	PENDING	PENDING	PENDING	PENDING	rosy1987b@gmail.com
14	Huma Stokes	NA	NA	CANCELLED	CANCELLED	CANCELLED	humastokes@gmail.com
14	Samuel Rutledge	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	humastokes@gmail.com
15	Eddison Collier	PENDING	PENDING	PENDING	PENDING	PENDING	eddison.collier@gmail.com
15	Stewart Nicholls	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	eddison.collier@gmail.com

Now we can associate each e-mail address with the respective guests in the guest list.

3.2 Filter emails

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

```
anti_join(email,guest,by=c("guest"="name")) %>%
knitr::kable() %>%
kable_styling("striped", latex_options = "scale_down",full_width = FALSE)
```

```
## Warning: Column `guest`/`name` joining character vector and factor,
## coercing into character vector
```

guest	email
Turner Jones	tjjones12@hotmail.ca
Albert Marshall	themarshallfamily1234@gmail.com
Vivian Marshall	themarshallfamily1234@gmail.com

We have three guests who are in the e-mail list but have not yet been included in the guest list! We should contact them soon and ask for their meal preferences and attendance to the events.

3.3 Make a guestlist

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

```
full_join(guest,email,by=c("name"="guest")) %>%
knitr::kable() %>%
kable_styling("striped", latex_options = "scale_down",full_width = FALSE)
```

party	name	meal_wedding	meal_brunch	attendance_wedding	attendance_brunch	attendance_golf	email
1	Sommer Medrano	PENDING	PENDING	PENDING	PENDING	PENDING	sommm@gmail.com
1	Phillip Medrano	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	sommm@gmail.com
1	Blanka Medrano	chicken	Menu A	CONFIRMED	CONFIRMED	CONFIRMED	sommm@gmail.com
1	Emaan Medrano	PENDING	PENDING	PENDING	PENDING	PENDING	sommm@gmail.com
2	Blair Park	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	bpark@gmail.com
2	Nigel Webb	NA	NA	CANCELLED	CANCELLED	CANCELLED	bpark@gmail.com
3	Sinead English	PENDING	PENDING	PENDING	PENDING	PENDING	singlish@hotmail.ca
4	Ayra Marks	vegetarian	Menu B	PENDING	PENDING	PENDING	marksa42@gmail.com
5	Atlanta Connolly	PENDING	PENDING	PENDING	PENDING	PENDING	NA
5	Denzel Connolly	fish	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	NA
5	Chanelle Shah	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	NA
6	Jolene Welsh	NA	NA	CANCELLED	CANCELLED	CANCELLED	jw1987@hotmail.com
6	Hayley Booker	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	jw1987@hotmail.com
7	Amayah Sanford	NA	PENDING	CANCELLED	PENDING	PENDING	erikaaaaaa@gmail.com
7	Erika Foley	PENDING	PENDING	PENDING	PENDING	PENDING	erikaaaaaa@gmail.com
8	Ciaron Acosta	PENDING	Menu A	PENDING	PENDING	PENDING	shining_ciaron@gmail.com
9	Diana Stuart	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	doodledianastu@gmail.com
10	Cosmo Dunkley	PENDING	PENDING	PENDING	PENDING	PENDING	NA
11	Cai Mcdaniel	fish	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	NA
12	Daisy-May Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	caldwellfamily5212@gmail.com
12	Martin Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING	caldwellfamily5212@gmail.com
12	Violet Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING	caldwellfamily5212@gmail.com
12	Nazifa Caldwell	chicken	PENDING	PENDING	PENDING	PENDING	caldwellfamily5212@gmail.com
12	Eric Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	caldwellfamily5212@gmail.com
13	Rosanna Bird	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	rosy1987b@gmail.com
13	Kurtis Frost	PENDING	PENDING	PENDING	PENDING	PENDING	rosy1987b@gmail.com
14	Huma Stokes	NA	NA	CANCELLED	CANCELLED	CANCELLED	humastokes@gmail.com
14	Samuel Rutledge	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	humastokes@gmail.com
15	Eddison Collier	PENDING	PENDING	PENDING	PENDING	PENDING	eddison.collier@gmail.com
15	Stewart Nicholls	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	eddison.collier@gmail.com
NA	Turner Jones	NA	NA	NA	NA	NA	tjjones12@hotmail.ca
NA	Albert Marshall	NA	NA	NA	NA	NA	themarshallfamily1234@gmail.com
NA	Vivian Marshall	NA	NA	NA	NA	NA	themarshallfamily1234@gmail.com

Now we have one table to refer to for the guests, their meal preferences, attendance, and e-mail. This makes it easier to plan for the wedding!