HW04: Tidy data and joins

Carleena Ortega

02/10/2019

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.
2. Take advantage of this new data shape to scatterplot life expectancy for one country against that of another.
3. Re-lengthen the data.

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.

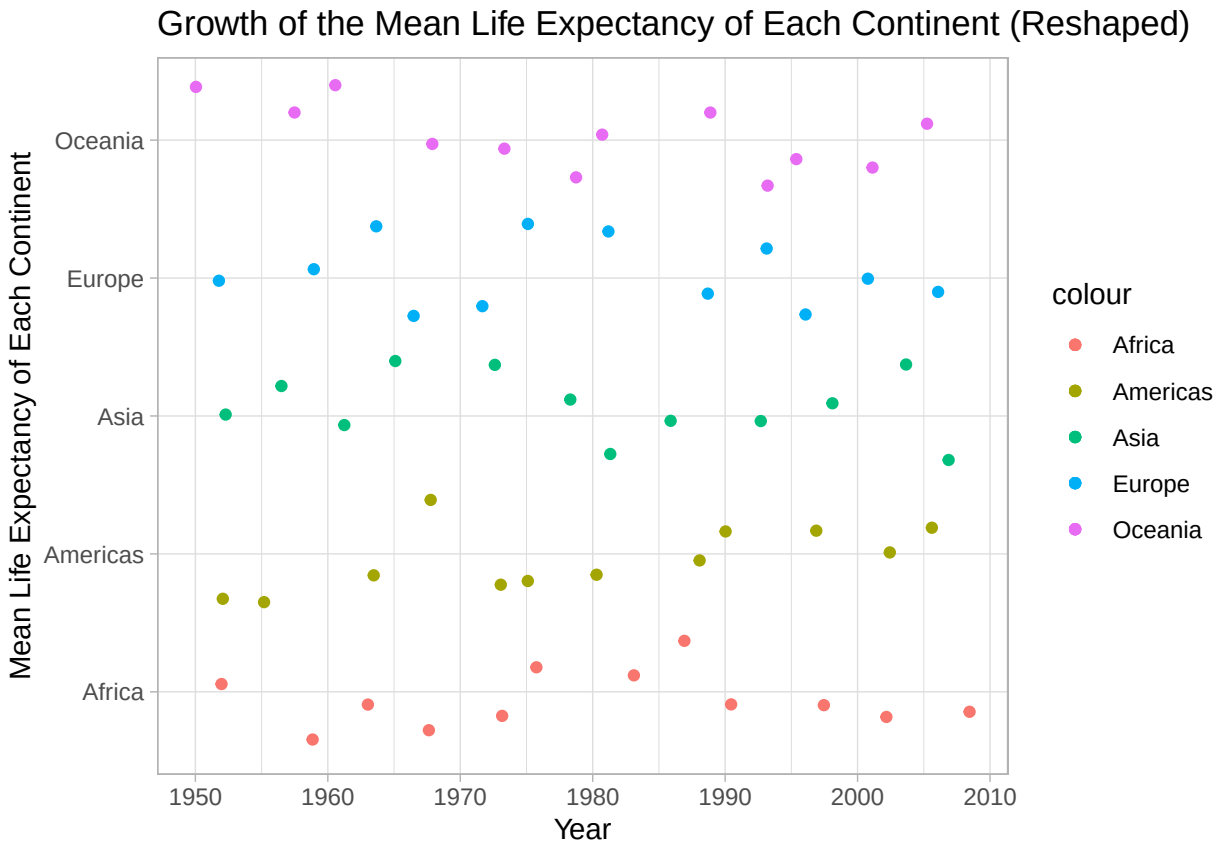
```
t1<-gapminder %>%
  group_by(continent,year) %>%
  select(continent,year,lifeExp) %>%
  arrange(continent,year) %>%
  summarize(mean_LE=mean(lifeExp))

t1_wide<-pivot_wider(t1,names_from="continent", values_from="mean_LE")
knitr::kable(t1_wide) %>%
  kable_styling("striped",latex_options="basic",
    full_width=FALSE,position="center")
```

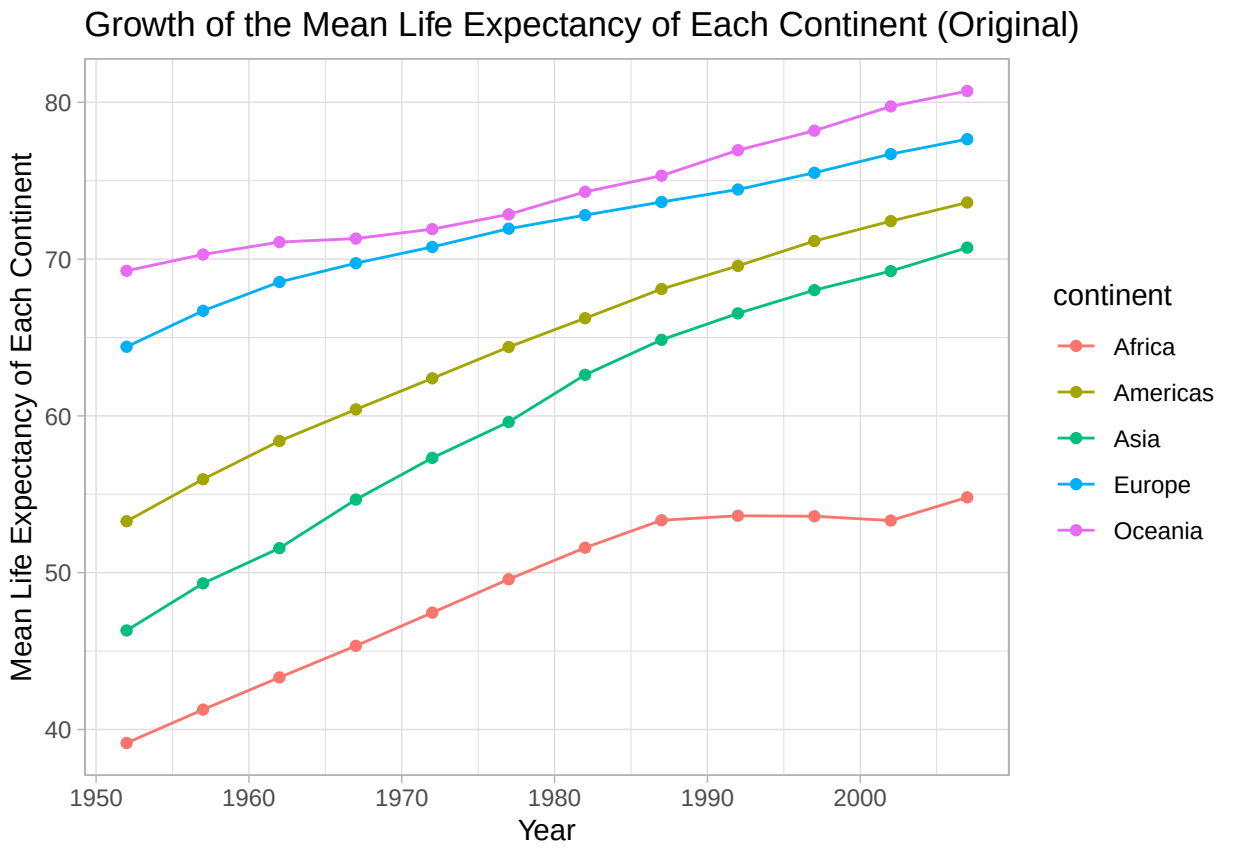
year	Africa	Americas	Asia	Europe	Oceania
1952	39.13550	53.27984	46.31439	64.40850	69.2550
1957	41.26635	55.96028	49.31854	66.70307	70.2950
1962	43.31944	58.39876	51.56322	68.53923	71.0850
1967	45.33454	60.41092	54.66364	69.73760	71.3100
1972	47.45094	62.39492	57.31927	70.77503	71.9100
1977	49.58042	64.39156	59.61056	71.93777	72.8550
1982	51.59287	66.22884	62.61794	72.80640	74.2900
1987	53.34479	68.09072	64.85118	73.64217	75.3200
1992	53.62958	69.56836	66.53721	74.44010	76.9450
1997	53.59827	71.15048	68.02052	75.50517	78.1900
2002	53.32523	72.42204	69.23388	76.70060	79.7400
2007	54.80604	73.60812	70.72848	77.64860	80.7195

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!

```
ggplot(t1_wide) +
  geom_jitter(aes(year,"Africa",color="Africa"))+
  geom_jitter(aes(year,"Americas", color="Americas"))+
  geom_jitter(aes(year,"Asia",color="Asia"))+
  geom_jitter(aes(year,"Europe",color="Europe"))+
  geom_jitter(aes(year,"Oceania",color="Oceania"))+
  labs(x="Year",y="Mean Life Expectancy of Each Continent",
  title="Growth of the Mean Life Expectancy of Each Continent (Reshaped)")+
  theme_light()
```



```
ggplot(t1) +
  geom_point(aes(year,mean_LE,color=continent))+
  geom_line(aes(year,mean_LE,color=continent))+
  labs(x="Year",y="Mean Life Expectancy of Each Continent",
  title="Growth of the Mean Life Expectancy of Each Continent (Original)")+
  theme_light()
```



The x axis in the reshaped data is the continents and it was difficult to add scales and label the axis properly. Coding for the plots is much easier for the original data , it is more visually appealing, and the trends are more apparent. Hence, comparing the original and the reshaped plots suggests that plotting the original data is simpler, neater and more understandable.

3. Re-lengthen the data.

```
t1_wide %>%
  pivot_longer(cols = c(-year),
               names_to  = "Continent",
               values_to = "Mean Life Expectancy") %>%
  knitr::kable() %>%
    kable_styling("striped",latex_options="basic",font_size=8,
      full_width=FALSE, position="center")
```

year	Continent	Mean Life Expectancy
1952	Africa	39.13550
1952	Americas	53.27984
1952	Asia	46.31439
1952	Europe	64.40850
1952	Oceania	69.25500
1957	Africa	41.26635
1957	Americas	55.96028
1957	Asia	49.31854
1957	Europe	66.70307
1957	Oceania	70.29500
1962	Africa	43.31944
1962	Americas	58.39876
1962	Asia	51.56322
1962	Europe	68.53923
1962	Oceania	71.08500
1967	Africa	45.33454
1967	Americas	60.41092
1967	Asia	54.66364
1967	Europe	69.73760
1967	Oceania	71.31000
1972	Africa	47.45094
1972	Americas	62.39492
1972	Asia	57.31927
1972	Europe	70.77503
1972	Oceania	71.91000
1977	Africa	49.58042
1977	Americas	64.39156
1977	Asia	59.61056
1977	Europe	71.93777
1977	Oceania	72.85500
1982	Africa	51.59287
1982	Americas	66.22884
1982	Asia	62.61794
1982	Europe	72.80640
1982	Oceania	74.29000
1987	Africa	53.34479
1987	Americas	68.09072
1987	Asia	64.85118
1987	Europe	73.64217
1987	Oceania	75.32000
1992	Africa	53.62958
1992	Americas	69.56836
1992	Asia	66.53721
1992	Europe	74.44010
1992	Oceania	76.94500
1997	Africa	53.59827
1997	Americas	71.15048
1997	Asia	68.02052
1997	Europe	75.50517
1997	Oceania	78.19000
2002	Africa	53.32523
2002	Americas	72.42204
2002	Asia	69.23388
2002	Europe	76.70060
2002	Oceania	79.74000
2007	Africa	54.80604
2007	Americas	73.60812
2007	Asia	70.72848
2007	Europe	77.64860
2007	Oceania	80.71950

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") %>%
    group_by(year) %>%
  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

2. Re-lengthen the data.

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

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

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

~~**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.csv
```

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.

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

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	NA
1	Blanka Medrano	chicken	Menu A	CONFIRMED	CONFIRMED	CONFIRMED	NA
1	Emaan Medrano	PENDING	PENDING	PENDING	PENDING	PENDING	NA
2	Blair Park	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	bpark@gmail.com
2	Nigel Webb	NA	NA	CANCELLED	CANCELLED	CANCELLED	NA
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	NA
7	Amayah Sanford	NA	PENDING	CANCELLED	PENDING	PENDING	erikaaaaaa@gmail.com
7	Erika Foley	PENDING	PENDING	PENDING	PENDING	PENDING	NA
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	NA
12	Violet Caldwell	PENDING	PENDING	PENDING	PENDING	PENDING	NA
12	Nazifa Caldwell	chicken	PENDING	PENDING	PENDING	PENDING	NA
12	Eric Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	NA
13	Rosanna Bird	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	rosy1987b@gmail.com
13	Kurtis Frost	PENDING	PENDING	PENDING	PENDING	PENDING	NA
14	Huma Stokes	NA	NA	CANCELLED	CANCELLED	CANCELLED	humastokes@gmail.com
14	Samuel Rutledge	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	NA
15	Eddison Collier	PENDING	PENDING	PENDING	PENDING	PENDING	eddison.collier@gmail.com
15	Stewart Nicholls	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	NA

3.2 Filter emails

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

```
emails_nolist<-right_join(guest,email2,by=c("name"="guest"))
```

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

```
knitr::kable(emails_nolist) %>%
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
NA	Phillip Medrano	NA	NA	NA	NA	NA	sommm@gmail.com
NA	Blanka Medrano	NA	NA	NA	NA	NA	sommm@gmail.com
NA	Emaan Medrano	NA	NA	NA	NA	NA	sommm@gmail.com
2	Blair Park	chicken	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	bpark@gmail.com
NA	Nigel Webb	NA	NA	NA	NA	NA	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
6	Jolene Welsh	NA	NA	CANCELLED	CANCELLED	CANCELLED	jw1987@hotmail.com
NA	Hayley Booker	NA	NA	NA	NA	NA	jw1987@hotmail.com
7	Amayah Sanford	NA	PENDING	CANCELLED	PENDING	PENDING	erikaaaaaa@gmail.com
NA	Erika Foley	NA	NA	NA	NA	NA	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
12	Daisy-May Caldwell	chicken	Menu B	CONFIRMED	CONFIRMED	CONFIRMED	caldwellfamily5212@gmail.com
NA	Martin Caldwell	NA	NA	NA	NA	NA	caldwellfamily5212@gmail.com
NA	Violet Caldwell	NA	NA	NA	NA	NA	caldwellfamily5212@gmail.com
NA	Nazifa Caldwell	NA	NA	NA	NA	NA	caldwellfamily5212@gmail.com
NA	Eric Caldwell	NA	NA	NA	NA	NA	caldwellfamily5212@gmail.com
13	Rosanna Bird	vegetarian	Menu C	CONFIRMED	CONFIRMED	CONFIRMED	rosy1987b@gmail.com
NA	Kurtis Frost	NA	NA	NA	NA	NA	rosy1987b@gmail.com
14	Huma Stokes	NA	NA	CANCELLED	CANCELLED	CANCELLED	humastokes@gmail.com
NA	Samuel Rutledge	NA	NA	NA	NA	NA	humastokes@gmail.com
15	Eddison Collier	PENDING	PENDING	PENDING	PENDING	PENDING	eddison.collier@gmail.com
NA	Stewart Nicholls	NA	NA	NA	NA	NA	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

3.3 Make a guestlist

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