gapminder Analysis

Monica Buczynski

9/7/2020

Note: The purpose of this document is to showcase a sample of skills that I learned in R for Data Science (chapter: Many Models) by Garrett Grolemund and Hadley Wickham. Some scripts were taken from https://r4ds.had.co.nz/many-models.html. The code for each exercise was studied carefully for understanding and then was retyped manually into R to maximize the learning experience; however, many of the scripts were altered for further analysis and presentation aesthetics or I added my own code for further analysis.

Question: Examining gapminder data: "How does life expectancy (lifeExp) change over time (year) for each country (country)?"

```
# Examining data before analysis

dim(gapminder) # there are 1704 rows and 6 columns
```

```
## [1] 1704    6
```

```
# View first six lines of data to get an idea what kind of values we are working with
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

```
summary(gapminder) # 5 number summary of each variable
```

```
##         country        continent        year         lifeExp
##  Afghanistan:  12   Africa  :624   Min.   :1952   Min.   :23.60
##  Albania    :  12   Americas:300   1st Qu.:1966   1st Qu.:48.20
##  Algeria    :  12   Asia    :396   Median :1980   Median :60.71
##  Angola     :  12   Europe  :360   Mean   :1980   Mean   :59.47
##  Argentina  :  12   Oceania : 24   3rd Qu.:1993   3rd Qu.:70.85
##  Australia  :  12                  Max.   :2007   Max.   :82.60
##  (Other)    :1632
##       pop              gdpPercap
##  Min.   :6.001e+04   Min.   :   241.2
##  1st Qu.:2.794e+06   1st Qu.:  1202.1
##  Median :7.024e+06   Median :  3531.8
##  Mean   :2.960e+07   Mean   :  7215.3
##  3rd Qu.:1.959e+07   3rd Qu.:  9325.5
##  Max.   :1.319e+09   Max.   :113523.1
##
```

```
# it probably makes more sense to examine the 5 number summary of the variables *lifeExp*,
  # *pop* and *gdpPercap* though it could still be useful to examine the min and max values
  # for *year* to ensure that the description of the data matches (i.e. the years of the data)

(summary(gapminder1 <- gapminder %>%
select(-country:-year)))
```

```
##     lifeExp           pop              gdpPercap
##  Min.   :23.60   Min.   :6.001e+04   Min.   :   241.2
##  1st Qu.:48.20   1st Qu.:2.794e+06   1st Qu.:  1202.1
##  Median :60.71   Median :7.024e+06   Median :  3531.8
##  Mean   :59.47   Mean   :2.960e+07   Mean   :  7215.3
##  3rd Qu.:70.85   3rd Qu.:1.959e+07   3rd Qu.:  9325.5
##  Max.   :82.60   Max.   :1.319e+09   Max.   :113523.1
```

```
# I will add the standard deviations of these three variables:

sd(gapminder$lifeExp)
```

```
## [1] 12.91711
```

```
sd(gapminder$pop)
```

```
## [1] 106157897
```

```
sd(gapminder$gdpPercap)
```

```
## [1] 9857.455
```

```
# Make five number summary table with standard deviations:
```

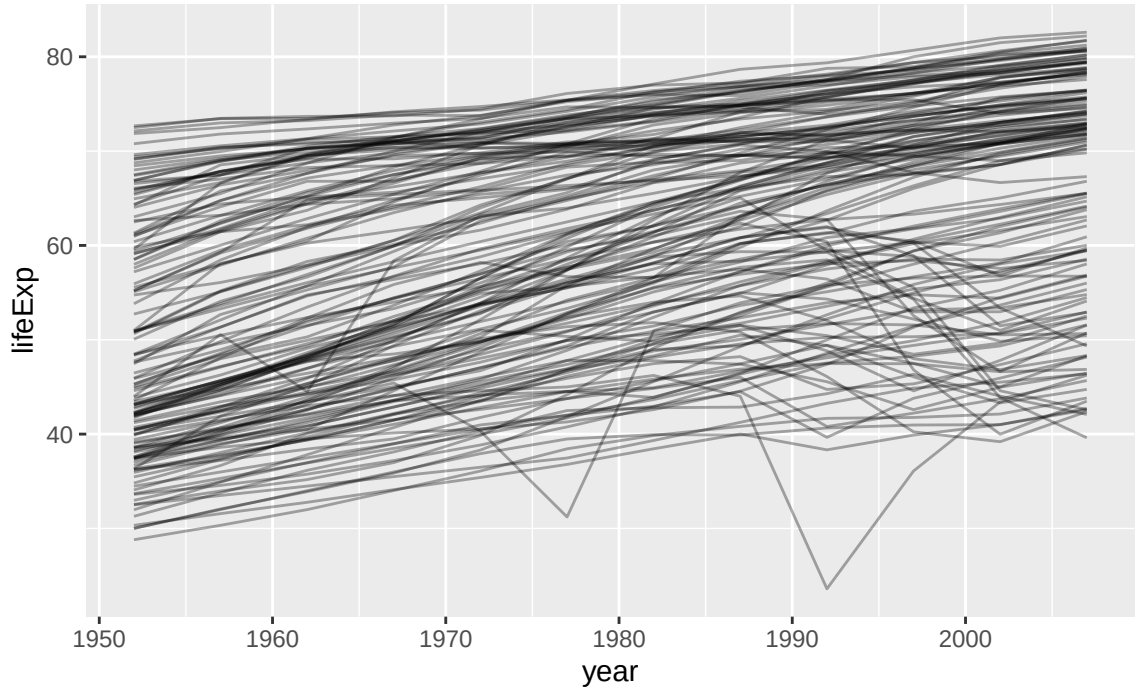
Table 1

Five-Number Summary and Standard Deviation Values

	lifeExp	pop	gdpPercap
Minimum	23.60	6.001e+04	241.2
1st Quartile	17.94	2.794e+06	1202.1
Median	60.71	7.024e+06	3531.8
Mean	59.47	2.960e+07	7215.3
3rd Quartile	70.85	1.959e+07	9325.5
Maximum	82.60	1.319e+09	113523.1
Standard Deviation	12.91711	106157897	9857.455

Note. Table 1 shows the five-number summaries and standard deviations of all variables used in my study

```
# Overview: Life expectancy seems to be increasing on average;
#   however, there are some countries which do not follow this pattern

gapminder %>%
  ggplot(aes(year, lifeExp, group = country)) +
  geom_line(alpha = 1/3)
```

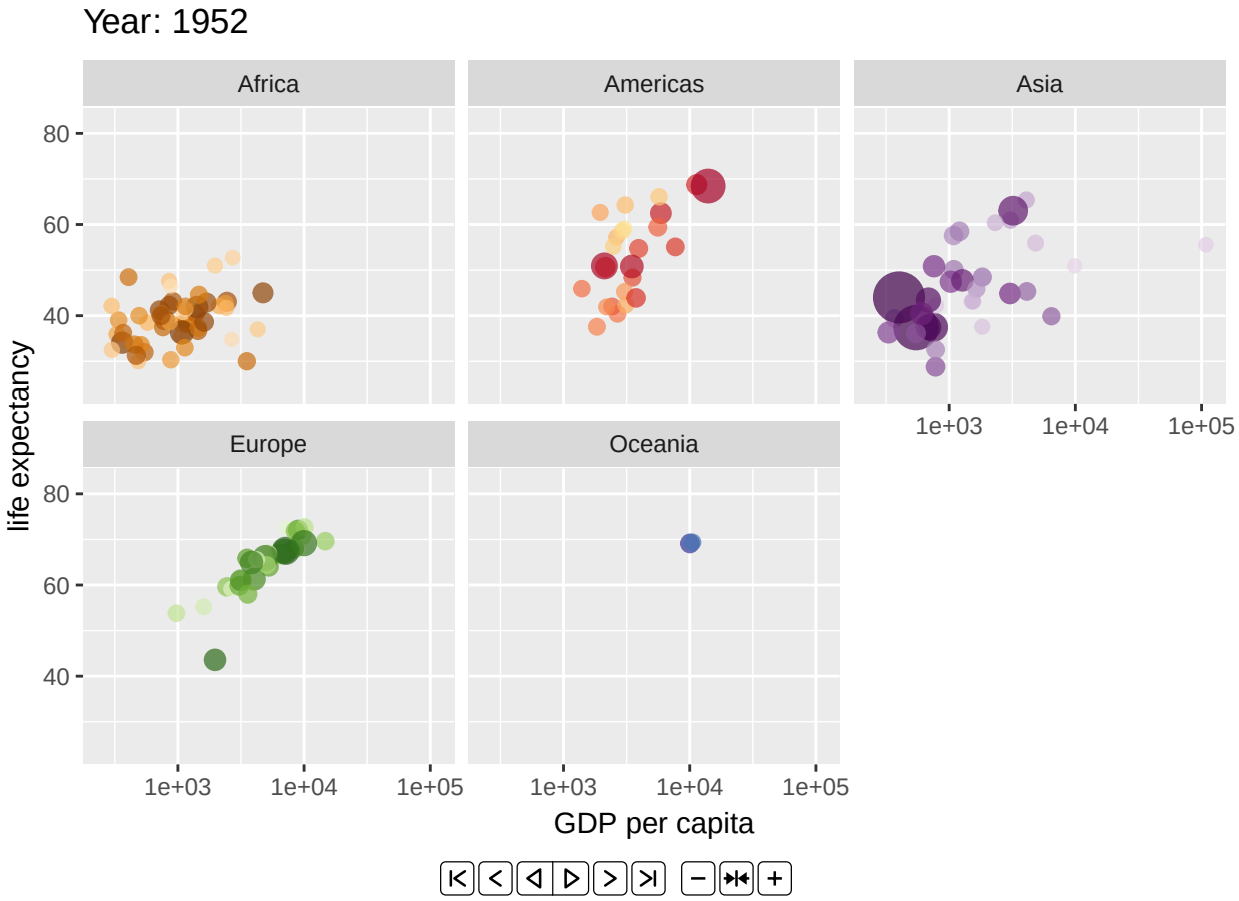


```
# There is a strong positive linear relationship between
#   the variables gdpPercap and lifeExp with a Pearson
#   correlation coefficient of 0.584

cor(gapminder$gdpPercap, gapminder$lifeExp, method = "pearson")
```

```
## [1] 0.5837062
```

Animation of Life Expectancy Trends by country with respect to GDP per capita

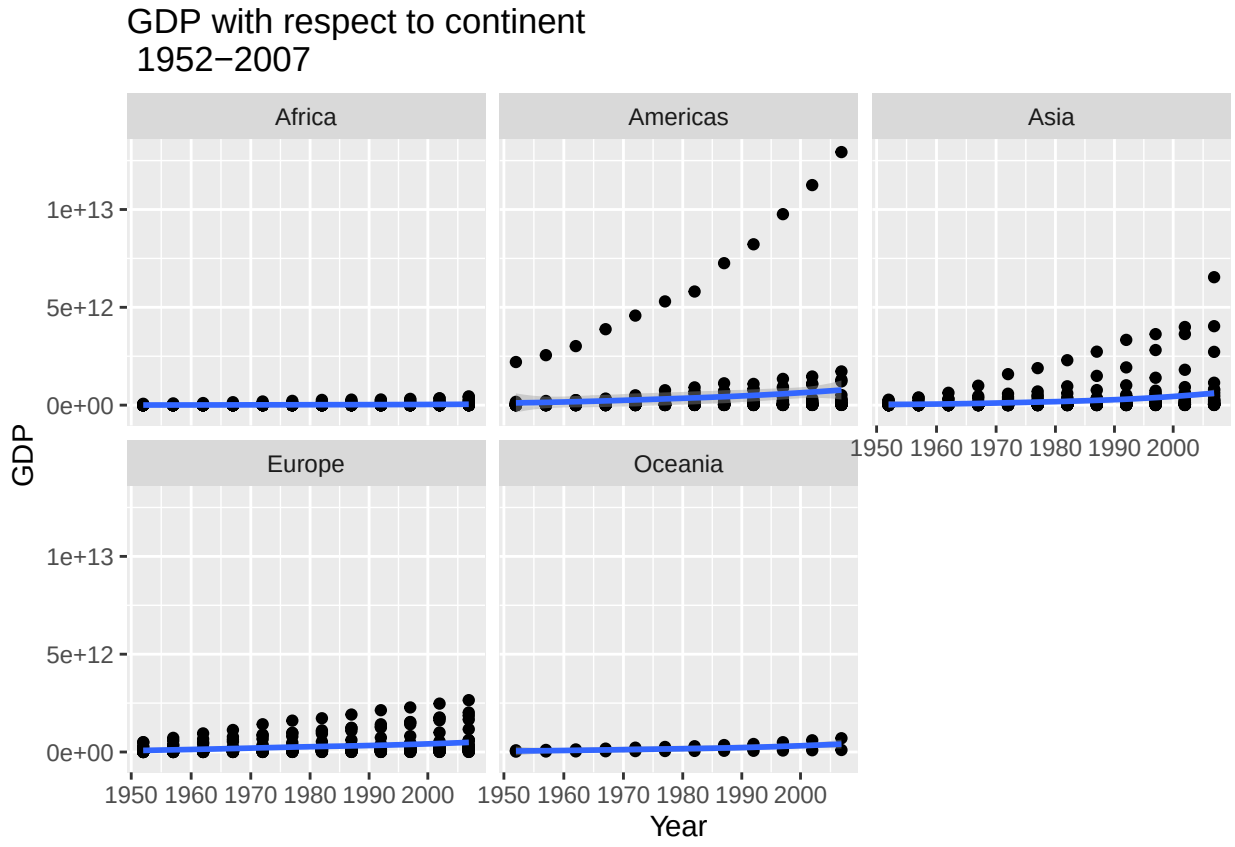
```
ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, colour = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  facet_wrap(~continent) +
  labs(title = 'Year: {frame_time}', x = 'GDP per capita', y = 'life expectancy') +
  transition_time(year) +
  ease_aes('linear')
```



```r
# Creating a new variable *totdgp* to represent the total GDP and identifying the top 10 entries
  #of countries with the highest total GDP.
# Note:  To the best of my knowledge it is unknown whether or not GDP has been adjusted for inflation.

head(gapminder %>%
  mutate(totgdp = gdpPercap * pop) %>%
  select(totgdp, country, year, gdpPercap, pop, continent) %>%
  arrange(desc(totgdp)), 10)
```

```
## # A tibble: 10 x 6
##     totgdp country        year gdpPercap        pop continent
##      <dbl> <fct>         <int>     <dbl>      <int> <fct>
##  1 1.29e13 United States  2007    42952.  301139947 Americas
##  2 1.12e13 United States  2002    39097.  287675526 Americas
##  3 9.76e12 United States  1997    35767.  272911760 Americas
##  4 8.22e12 United States  1992    32004.  256894189 Americas
##  5 7.26e12 United States  1987    29884.  242803533 Americas
##  6 6.54e12 China          2007     4959. 1318683096 Asia
##  7 5.81e12 United States  1982    25010.  232187835 Americas
##  8 5.30e12 United States  1977    24073.  220239000 Americas
##  9 4.58e12 United States  1972    21806.  209896000 Americas
## 10 4.04e12 Japan          2007    31656.  127467972 Asia
```

```
# Examine how total GDP in each continent has changed with respect to time.

# unique(gapminder$continent) # used to find how the countries are named in the data set


gapminder2 <- gapminder %>%
   mutate(gdp = gdpPercap * pop)

ggplot(gapminder2, aes(year, gdp)) +
  geom_point() +
  geom_smooth() +
  scale_x_log10() +
  facet_wrap(.~continent) +
  labs(title = 'GDP with respect to continent\n 1952-2007  ' , y = 'GDP', x = 'Year')
```




```
summary(gapminder2$year) # to find min and max years for title
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1952    1966    1980    1980    1993    2007
```

```
# The blue line shows the trend line of each continent.  In particular, there seems to be some
  # "outliers" in the Americas and in Asia where some countries are richer than others.
  # Let's check that out and find out which countries in each continent have the
  # highest totgdp! But first, lets divide the "Americas" into NORTH and SOUTH America.
```

```r
gapminder2 %>%
  filter(continent == "Americas")
```

```
## # A tibble: 300 x 7
##    country   continent  year lifeExp      pop gdpPercap          gdp
##    <fct>     <fct>     <int>   <dbl>    <int>     <dbl>        <dbl>
##  1 Argentina Americas   1952    62.5 17876956     5911. 105676319105.
##  2 Argentina Americas   1957    64.4 19610538     6857. 134466639306.
##  3 Argentina Americas   1962    65.1 21283783     7133. 151820757737.
##  4 Argentina Americas   1967    65.6 22934225     8053. 184688236498.
##  5 Argentina Americas   1972    67.1 24779799     9443. 233996596624.
##  6 Argentina Americas   1977    68.5 26983828    10079. 271970723960.
##  7 Argentina Americas   1982    69.9 29341374     8998. 264010673179.
##  8 Argentina Americas   1987    70.8 31620918     9140. 289004799539.
##  9 Argentina Americas   1992    71.9 33958947     9308. 316104097627.
## 10 Argentina Americas   1997    73.3 36203463    10967. 397053586287.
## # ... with 290 more rows
```

```r
# Renamed "Americas" to "North America."
  # This isn't exactly what we want because we need to assign "South America" as well.

gapminder3 <- gapminder2
gapminder3$continent <- gsub("Americas", "North America", gapminder3$continent)

# Now I can see all the countries with their respective continents
(by_countcont <- gapminder3 %>%
  group_by(country, continent) %>%
  nest())
```

```
## # A tibble: 142 x 3
## # Groups:   country, continent [142]
##    country     continent     data
##    <fct>       <chr>         <list>
##  1 Afghanistan Asia          <tibble [12 x 5]>
##  2 Albania     Europe        <tibble [12 x 5]>
##  3 Algeria     Africa        <tibble [12 x 5]>
##  4 Angola      Africa        <tibble [12 x 5]>
##  5 Argentina   North America <tibble [12 x 5]>
##  6 Australia   Oceania       <tibble [12 x 5]>
##  7 Austria     Europe        <tibble [12 x 5]>
##  8 Bahrain     Asia          <tibble [12 x 5]>
##  9 Bangladesh  Asia          <tibble [12 x 5]>
## 10 Belgium     Europe        <tibble [12 x 5]>
## # ... with 132 more rows
```

```r
# Filter to see which countries are in the Americas so we know which countries need to have
  # their continent value changed.

by_countcont2 <- by_countcont %>%
  filter(continent == "North America") %>%
  select(-data)

# To view all of the countries in the by_countcont2 df:

view(by_countcont2$country)
```

```
# Here is a table regarding the countries in the "Americas" and their respective
  # continent according to the Nations Online Project:
```

Table 2
North American and South American countries

North America	South America
Canada	Argentina
Costa Rica	Bolivia
Cuba	Brazil
Dominican Republic	Chile
El Salvador	Colombia
Haiti	Ecuador
Honduras	Paraguay
Jamaica	Peru
Mexico	Uruguay
Nicaragua	Venezuela
Panama	
Puerto Rico	
Trinidad and Tobago	
United States	

Note. South America: French Guiana (FR), Guyana and Suriname were omitted from the gapminder2 dataset. North America: Greenland (DK), Anguilla (UK), Antigua and Barbuda, Aruba (NL), Bahamas, Barbados, Bermuda (UK), British Virgin Islands (UK), Cayman Islands (UK), Curaçao (NL), Dominica, Grenada, Saint George's, Guadeloupe (FR), Martinique (FR), Montserrat (UK), Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Trinidad and Tobago and US Virgin Islands (USA).

```
# add "South America" to the continent column

gapminder3$continent[gapminder3$country == "Argentina" |
                       gapminder3$country == "Bolivia" |
                       gapminder3$country == "Brazil" |
                       gapminder3$country == "Chile" |
                       gapminder3$country ==  "Colombia" |
                       gapminder3$country ==  "Ecuador" |
                       gapminder3$country ==  "Paraguay" |
                       gapminder3$country ==  "Peru" |
                       gapminder3$country ==  "Uruguay" |
                       gapminder3$country ==  "Venezuela"] <- "South America"

# to check if South America and North America are listed as values in the continent
unique(gapminder3$continent)
```

```
## [1] "Asia"          "Europe"        "Africa"        "South America"
## [5] "Oceania"       "North America"
```

```
# check if we caught all the "Americas" countries from the original data set by seeing
  # if the summation of the countries in the new df that are in North America or South America
  # is equal to the number of countries in the original df that are in the "Americas"

# original data frame with "Americas"

head(gapminder_count_test <- gapminder %>%
  group_by(country,continent) %>%
    nest(), 6)
```

```
## # A tibble: 6 x 3
## # Groups:   country, continent [6]
##   country     continent data
##   <fct>       <fct>     <list>
## 1 Afghanistan Asia      <tibble [12 x 4]>
## 2 Albania     Europe    <tibble [12 x 4]>
## 3 Algeria     Africa    <tibble [12 x 4]>
## 4 Angola      Africa    <tibble [12 x 4]>
## 5 Argentina   Americas  <tibble [12 x 4]>
## 6 Australia   Oceania   <tibble [12 x 4]>
```

```
 sum(gapminder_count_test$continent == 'Americas')
```

```
## [1] 25
```

```
# new df with "North America" and "South America"

# list of all of the countries and their respective continents

(gapminder4 <- gapminder3 %>%
  group_by(country,continent) %>%
    nest() %>%
   filter(continent == "South America" | continent == "North America") %>%
select(-data))
```

```
## # A tibble: 25 x 2
## # Groups:   country, continent [25]
##    country            continent
##    <fct>              <chr>
##  1 Argentina          South America
##  2 Bolivia            South America
##  3 Brazil             South America
##  4 Canada             North America
##  5 Chile              South America
##  6 Colombia           South America
##  7 Costa Rica         North America
##  8 Cuba               North America
##  9 Dominican Republic North America
## 10 Ecuador            South America
## # ... with 15 more rows
```

```
(AN <- sum(gapminder4$continent == 'North America'))
```

```
## [1] 15
```

```
(AS <- sum(gapminder4$continent == 'South America'))
```

```
## [1] 10
```

```
AN + AS
```

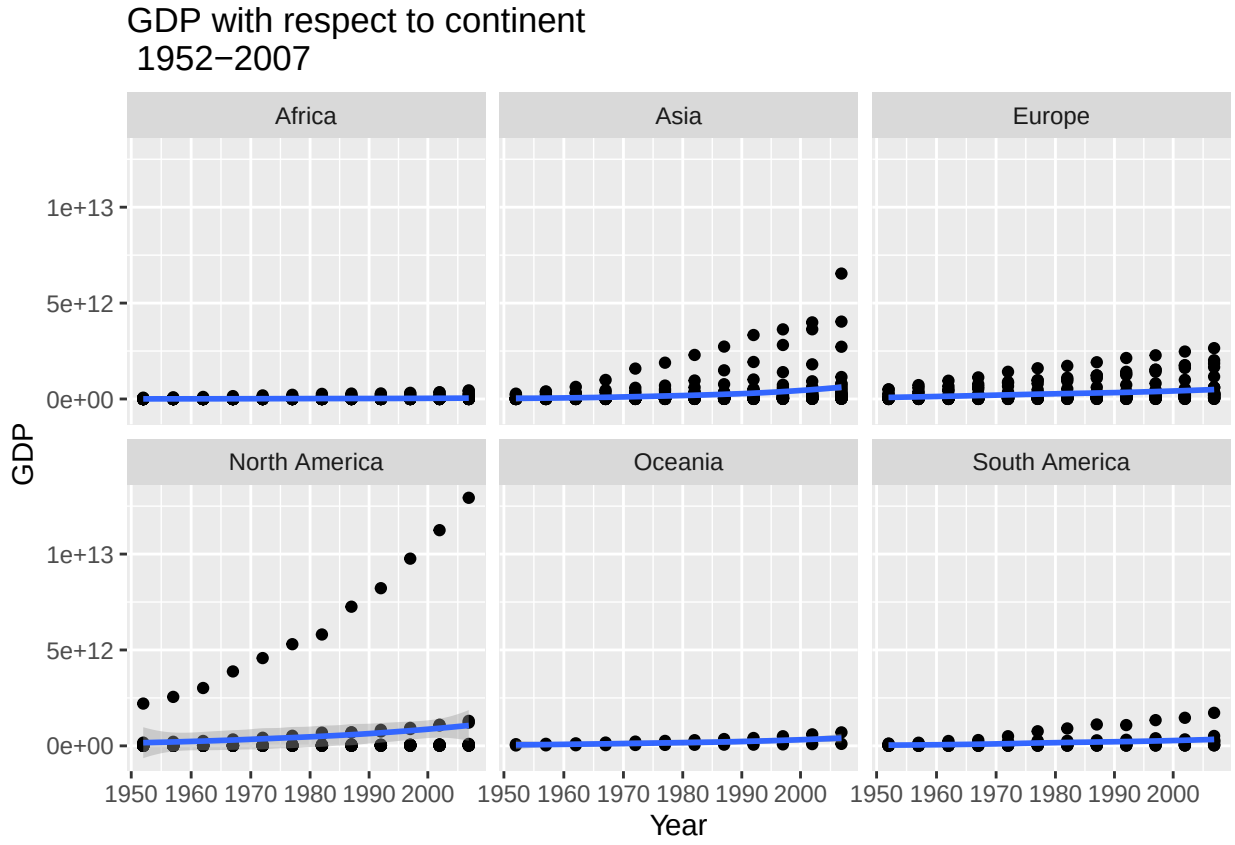
```
## [1] 25
```

```
# both have 25.  Lets set this up as a logical statement

AN + AS ==  sum(gapminder_count_test$continent == 'Americas')
```

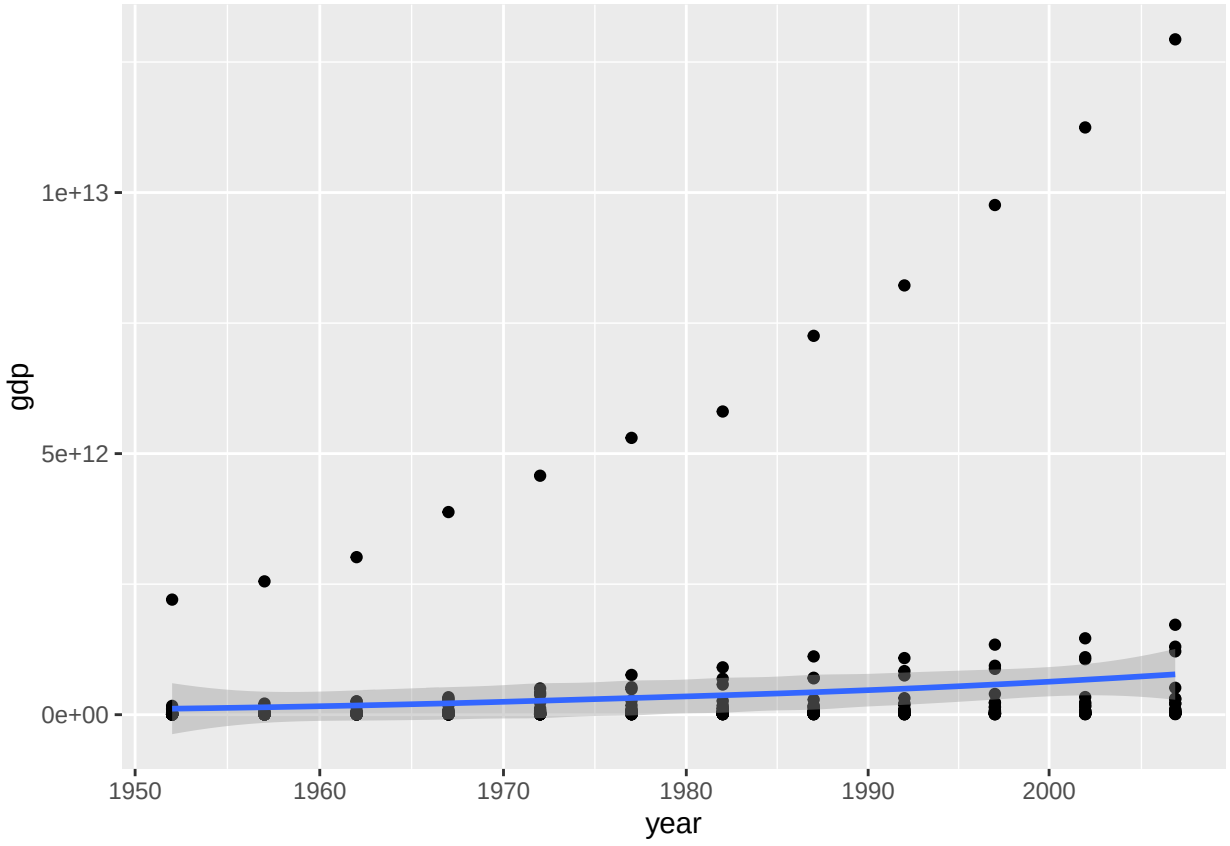
```
## [1] TRUE
```

```
# we get TRUE so this means that we have accounted for all of the countries that we needed to
```

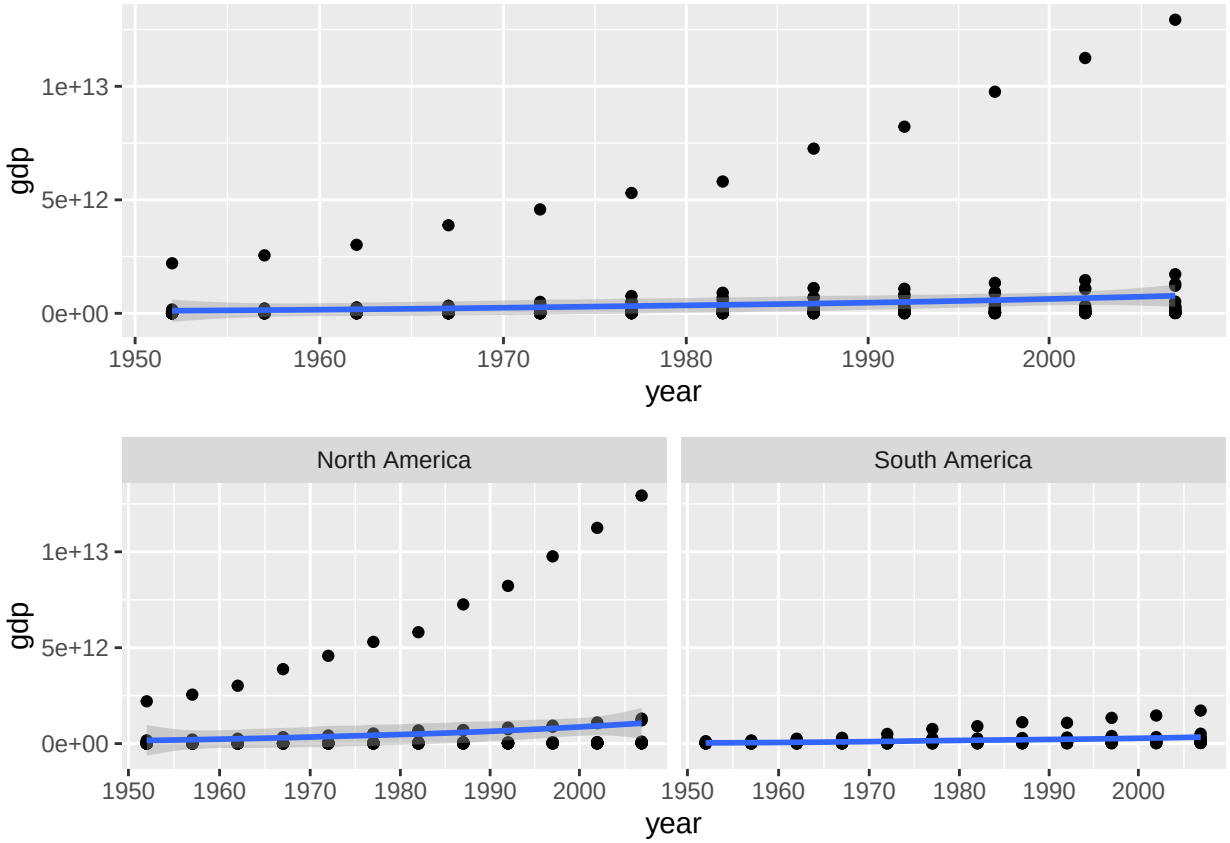
```
# From our new visual, we have now matched each country with its respective continent.
ggplot(gapminder3, aes(year, gdp)) +
  geom_point() +
  geom_smooth() +
  scale_x_log10() +
  facet_wrap(.~continent) +
  labs(title = 'GDP with respect to continent\n 1952-2007' , y = 'GDP', x = 'Year')
```




```
p2 <- gapminder3 %>%
  filter(continent == "North America" | continent == "South America") %>%
  ggplot(aes(year, gdp)) +
  geom_point() +
  geom_smooth() +
  scale_x_log10() +
  facet_wrap(.~continent)


(p1 <- gapminder2 %>%
  filter(continent == "Americas") %>%
  ggplot(aes(year, gdp)) +
  geom_point() +
  geom_smooth() +
  scale_x_log10())
```




```
ggarrange(p1, p2, ncol = 1)
```




```
# A noticeable visual difference is that the GDP of North America
  # increases over time versus the GDP of South America.
```

```r
# From 1952 to 2007, China (3 times), Japan (5 times)
  # and the United States (10 times) were in the top 1% of total GDP.

quantile(gapminder3$gdp, probs =  0.99)
```

```
##          99%
## 2.731639e+12
```

```r
top_counts_world_1 <- filter(gapminder3, gdp > 2.731639e+12)

view(top_counts_world_1 %>%
  arrange(desc(gdp)))

view(unique(top_counts_world_1$country))

table(top_counts_world_1$country == "United States")
```

```
##
## FALSE  TRUE
##     8    10
```

```r
table(top_counts_world_1$country == "Japan")
```

```
##
## FALSE  TRUE
##    13     5
```

```r
table(top_counts_world_1$country == "China")
```

```
##
## FALSE  TRUE
##    15     3
```

```r
# 0%, 25%, 50%, 75%, 100% for GDP of countries for any given year
  # between 1952-2007 within North America

do.call("rbind",
        tapply(gapminder3$gdp,   # Specify numeric column
               gapminder3$continent,  # Specify group variable
               quantile))
```

```
##                         0%         25%          50%          75%         100%
## Africa           52784691  2075656710   5327685830  16310607775 4.479709e+11
## Asia            629774980 12085020512  39646366753 153393367013 6.539501e+12
## Europe         1075341715 42410474179 112434821368 235979751201 2.650871e+12
## North America  2003975797  9210229784  18183688019  60194411071 1.293446e+13
## Oceania       21058193787 55303147125  97141165914 267463697959 7.036584e+11
## South America  3037550252 22586520566  67147057973 177135142324 1.722599e+12
```

```r
# Top 25% of GDP from North America from

# View top 25% countries with all other columns
 (NA_25 <- gapminder3 %>%
  filter(continent == "North America" & gdp > 60194411071) %>%
  arrange(desc(gdp)))
```

```
## # A tibble: 45 x 7
```

```
##    country       continent      year lifeExp       pop gdpPercap     gdp
##    <fct>         <chr>         <int>   <dbl>     <int>     <dbl>   <dbl>
##  1 United States North America  2007    78.2 301139947    42952. 1.29e13
##  2 United States North America  2002    77.3 287675526    39097. 1.12e13
##  3 United States North America  1997    76.8 272911760    35767. 9.76e12
##  4 United States North America  1992    76.1 256894189    32004. 8.22e12
##  5 United States North America  1987    75.0 242803533    29884. 7.26e12
##  6 United States North America  1982    74.6 232187835    25010. 5.81e12
##  7 United States North America  1977    73.4 220239000    24073. 5.30e12
##  8 United States North America  1972    71.3 209896000    21806. 4.58e12
##  9 United States North America  1967    70.8 198712000    19530. 3.88e12
## 10 United States North America  1962    70.2 186538000    16173. 3.02e12
## # ... with 35 more rows
```

```
# View which countries are "unique" --> United States (12 times), Mexico (12 times),
  # Canada (12 times), Cuba (5 times), Puerto Rico (3 times), Guatemala (1 time)

view(unique(NA_25$country))

# How many: view as list or search with code via TRUE
as.data.frame(table(NA_25$country))
```

```
##                          Var1 Freq
## 1                 Afghanistan    0
## 2                     Albania    0
## 3                     Algeria    0
## 4                      Angola    0
## 5                   Argentina    0
## 6                   Australia    0
## 7                     Austria    0
## 8                     Bahrain    0
## 9                  Bangladesh    0
## 10                    Belgium    0
## 11                      Benin    0
## 12                    Bolivia    0
## 13    Bosnia and Herzegovina    0
## 14                   Botswana    0
## 15                     Brazil    0
## 16                   Bulgaria    0
## 17               Burkina Faso    0
## 18                    Burundi    0
## 19                   Cambodia    0
## 20                   Cameroon    0
## 21                     Canada   12
## 22  Central African Republic    0
## 23                       Chad    0
## 24                      Chile    0
## 25                      China    0
## 26                   Colombia    0
## 27                    Comoros    0
## 28          Congo, Dem. Rep.    0
## 29               Congo, Rep.    0
## 30                 Costa Rica    0
## 31              Cote d'Ivoire    0
## 32                    Croatia    0
```

```
## 33                    Cuba    5
## 34          Czech Republic    0
## 35                 Denmark    0
## 36                Djibouti    0
## 37      Dominican Republic    0
## 38                 Ecuador    0
## 39                   Egypt    0
## 40             El Salvador    0
## 41       Equatorial Guinea    0
## 42                 Eritrea    0
## 43                Ethiopia    0
## 44                 Finland    0
## 45                  France    0
## 46                   Gabon    0
## 47                  Gambia    0
## 48                 Germany    0
## 49                   Ghana    0
## 50                  Greece    0
## 51               Guatemala    1
## 52                  Guinea    0
## 53           Guinea-Bissau    0
## 54                   Haiti    0
## 55                Honduras    0
## 56        Hong Kong, China    0
## 57                 Hungary    0
## 58                 Iceland    0
## 59                   India    0
## 60               Indonesia    0
## 61                    Iran    0
## 62                    Iraq    0
## 63                 Ireland    0
## 64                  Israel    0
## 65                   Italy    0
## 66                 Jamaica    0
## 67                   Japan    0
## 68                  Jordan    0
## 69                   Kenya    0
## 70         Korea, Dem. Rep.    0
## 71              Korea, Rep.    0
## 72                  Kuwait    0
## 73                 Lebanon    0
## 74                 Lesotho    0
## 75                 Liberia    0
## 76                   Libya    0
## 77              Madagascar    0
## 78                  Malawi    0
## 79                Malaysia    0
## 80                    Mali    0
## 81              Mauritania    0
## 82               Mauritius    0
## 83                  Mexico   12
## 84                Mongolia    0
## 85              Montenegro    0
## 86                 Morocco    0
```

```
## 87                Mozambique    0
## 88                   Myanmar    0
## 89                   Namibia    0
## 90                     Nepal    0
## 91               Netherlands    0
## 92               New Zealand    0
## 93                 Nicaragua    0
## 94                     Niger    0
## 95                   Nigeria    0
## 96                    Norway    0
## 97                      Oman    0
## 98                  Pakistan    0
## 99                    Panama    0
## 100                 Paraguay    0
## 101                     Peru    0
## 102              Philippines    0
## 103                   Poland    0
## 104                 Portugal    0
## 105              Puerto Rico    3
## 106                  Reunion    0
## 107                  Romania    0
## 108                   Rwanda    0
## 109    Sao Tome and Principe    0
## 110             Saudi Arabia    0
## 111                  Senegal    0
## 112                   Serbia    0
## 113             Sierra Leone    0
## 114                Singapore    0
## 115          Slovak Republic    0
## 116                 Slovenia    0
## 117                  Somalia    0
## 118             South Africa    0
## 119                    Spain    0
## 120                Sri Lanka    0
## 121                    Sudan    0
## 122                Swaziland    0
## 123                   Sweden    0
## 124              Switzerland    0
## 125                    Syria    0
## 126                   Taiwan    0
## 127                 Tanzania    0
## 128                 Thailand    0
## 129                     Togo    0
## 130      Trinidad and Tobago    0
## 131                  Tunisia    0
## 132                   Turkey    0
## 133                   Uganda    0
## 134           United Kingdom    0
## 135            United States   12
## 136                  Uruguay    0
## 137                Venezuela    0
## 138                  Vietnam    0
## 139       West Bank and Gaza    0
## 140              Yemen, Rep.    0
```

```
## 141                   Zambia    0
## 142                 Zimbabwe    0
```

```
table(NA_25$country == "United States")
```

```
##
## FALSE  TRUE
##    33    12
```

```
table(NA_25$country == "Mexico")
```

```
##
## FALSE  TRUE
##    33    12
```

```
table(NA_25$country == "Canada")
```

```
##
## FALSE  TRUE
##    33    12
```

```
table(NA_25$country == "Cuba")
```

```
##
## FALSE  TRUE
##    40     5
```

```
table(NA_25$country == "Puerto Rico")
```

```
##
## FALSE  TRUE
##    42     3
```

```
table(NA_25$country == "Guatemala")
```

```
##
## FALSE  TRUE
##    44     1
```

Build a nested data frame to implement a common function for multiple countries

```
by_country <-  gapminder %>%
  group_by(country, continent) %>%
  nest()

by_country
```

```
## # A tibble: 142 x 3
## # Groups:   country, continent [142]
##    country     continent data
##    <fct>       <fct>     <list>
##  1 Afghanistan Asia      <tibble [12 x 4]>
##  2 Albania     Europe    <tibble [12 x 4]>
##  3 Algeria     Africa    <tibble [12 x 4]>
##  4 Angola      Africa    <tibble [12 x 4]>
##  5 Argentina   Americas  <tibble [12 x 4]>
##  6 Australia   Oceania   <tibble [12 x 4]>
##  7 Austria     Europe    <tibble [12 x 4]>
##  8 Bahrain     Asia      <tibble [12 x 4]>
##  9 Bangladesh  Asia      <tibble [12 x 4]>
## 10 Belgium     Europe    <tibble [12 x 4]>
## # ... with 132 more rows
```

```
# How to view all data for Afghanistan, for example.

which(by_country == "Afghanistan", arr.ind=TRUE)
```

```
##      row col
## [1,]   1   1
```

```
# Use arr.ind=TRUE to show column and row.
# Afghanistan is in the first row. Therefore..

by_country$data[[1]]
```

```
## # A tibble: 12 x 4
##     year lifeExp      pop gdpPercap
##    <int>   <dbl>    <int>     <dbl>
##  1  1952    28.8  8425333      779.
##  2  1957    30.3  9240934      821.
##  3  1962    32.0 10267083      853.
##  4  1967    34.0 11537966      836.
##  5  1972    36.1 13079460      740.
##  6  1977    38.4 14880372      786.
##  7  1982    39.9 12881816      978.
##  8  1987    40.8 13867957      852.
##  9  1992    41.7 16317921      649.
## 10  1997    41.8 22227415      635.
## 11  2002    42.1 25268405      727.
## 12  2007    43.8 31889923      975.
```

```
# Write function for model

country_model <- function(df) {
  lm(lifeExp ~ year, data = df)
```

```
}


# Apply our function to every data frame using map

models <- map(by_country$data, country_model)

# instead of creating a new object in the global environment,
  # we're going to create a new variable in the by_country data frame.
  # Use dplyr::mutate():

by_country <- by_country %>%
  mutate(model = map(data, country_model))

# Mutating the column into the data frame will prevent one from forgetting
  # to reorder/subset one df without doing the same to the other

# View new data frame:
by_country
```

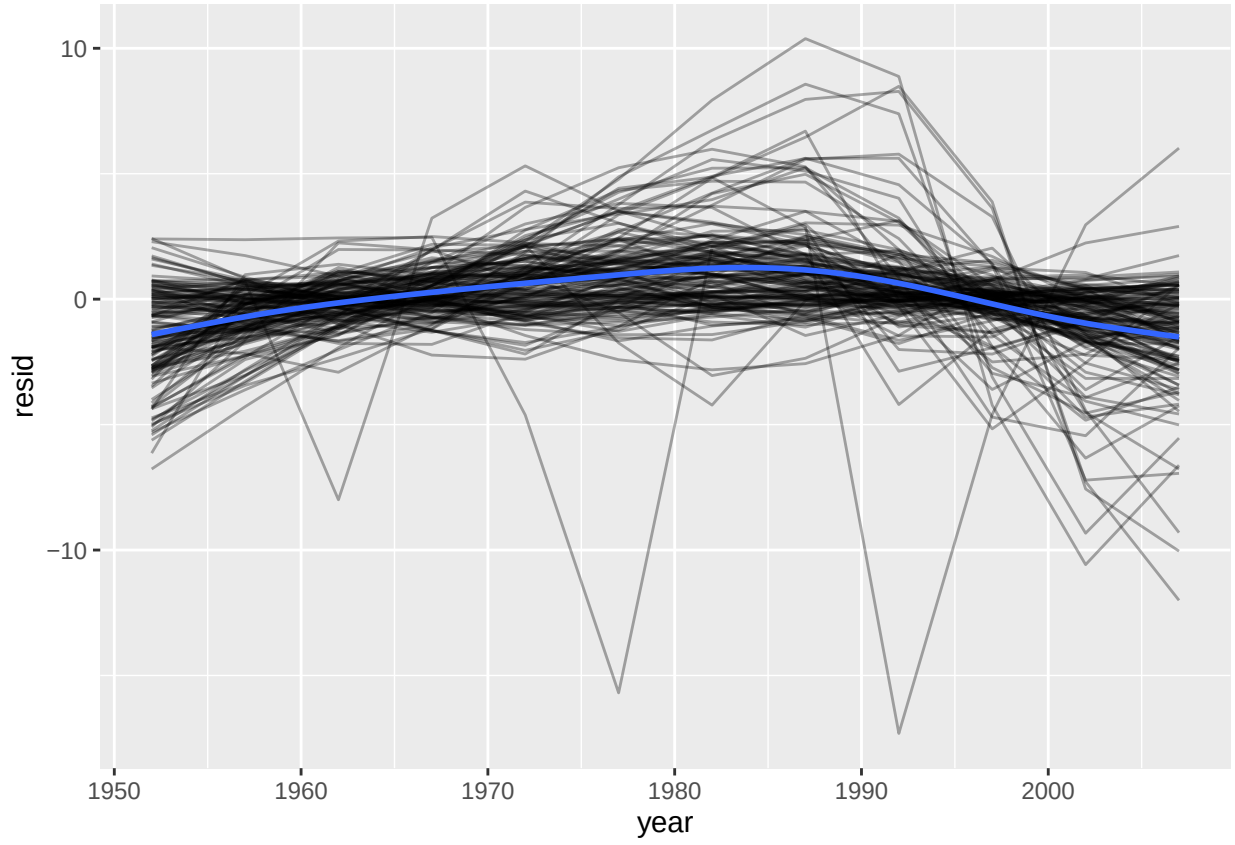
```
## # A tibble: 142 x 4
## # Groups:   country, continent [142]
##    country     continent data              model
##    <fct>       <fct>     <list>            <list>
##  1 Afghanistan Asia      <tibble [12 x 4]> <lm>
##  2 Albania     Europe    <tibble [12 x 4]> <lm>
##  3 Algeria     Africa    <tibble [12 x 4]> <lm>
##  4 Angola      Africa    <tibble [12 x 4]> <lm>
##  5 Argentina   Americas  <tibble [12 x 4]> <lm>
##  6 Australia   Oceania   <tibble [12 x 4]> <lm>
##  7 Austria     Europe    <tibble [12 x 4]> <lm>
##  8 Bahrain     Asia      <tibble [12 x 4]> <lm>
##  9 Bangladesh  Asia      <tibble [12 x 4]> <lm>
## 10 Belgium     Europe    <tibble [12 x 4]> <lm>
## # ... with 132 more rows
```

```
# To compute residuals, we need to unnest the df,
  # so that we do not need to worry about plotting a list of dfs.

by_country <-  by_country %>%
  mutate(resids = map2(data, model, add_residuals)
  )
resids <- unnest(by_country, resids)

# View new df with residuals:

resids
```

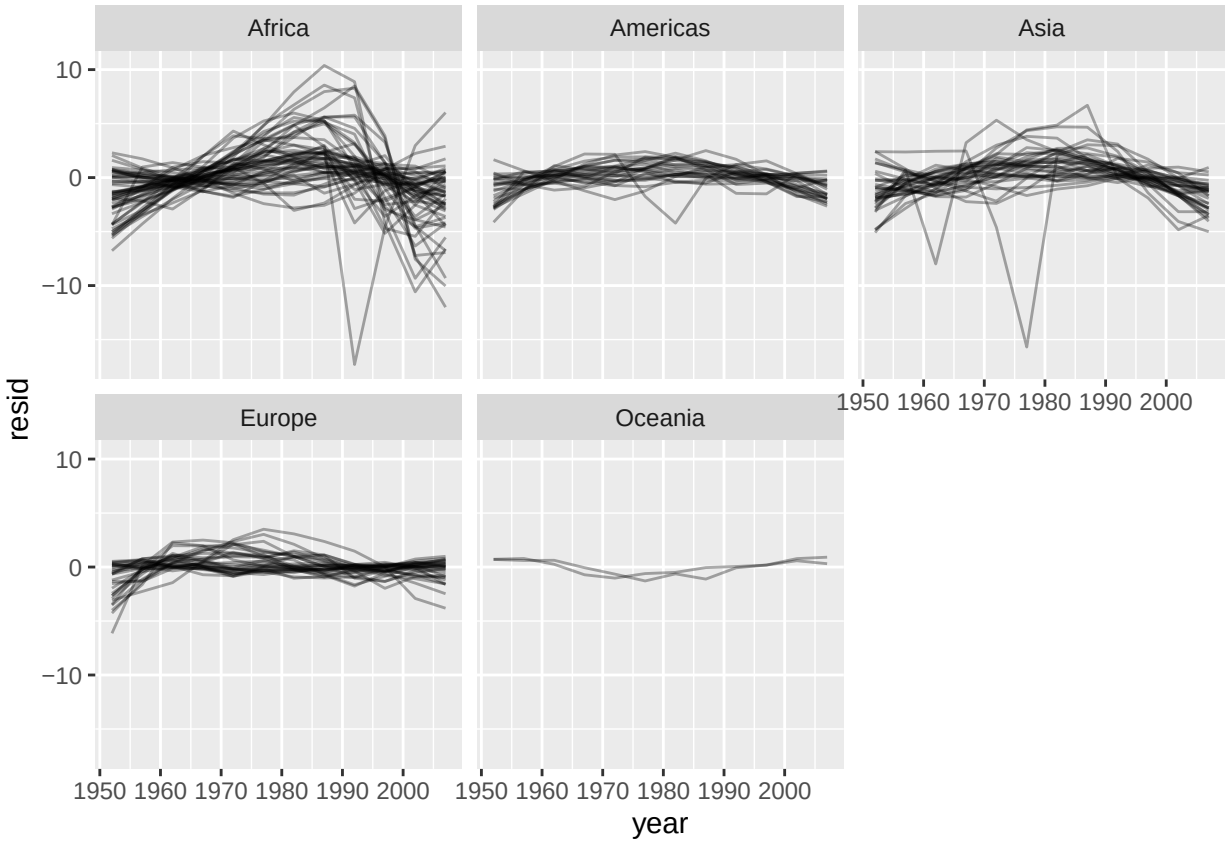
```
## # A tibble: 1,704 x 9
## # Groups:   country, continent [142]
##    country   continent data        model   year lifeExp    pop gdpPercap   resid
##    <fct>     <fct>     <list>      <list> <int>   <dbl>  <int>     <dbl>   <dbl>
##  1 Afghanis~ Asia      <tibble [1~ <lm>    1952    28.8 8.43e6      779. -1.11
##  2 Afghanis~ Asia      <tibble [1~ <lm>    1957    30.3 9.24e6      821. -0.952
##  3 Afghanis~ Asia      <tibble [1~ <lm>    1962    32.0 1.03e7      853. -0.664
```

```
##  4 Afghanis~ Asia      <tibble [1~ <lm>    1967    34.0 1.15e7      836. -0.0172
##  5 Afghanis~ Asia      <tibble [1~ <lm>    1972    36.1 1.31e7      740.  0.674
##  6 Afghanis~ Asia      <tibble [1~ <lm>    1977    38.4 1.49e7      786.  1.65
##  7 Afghanis~ Asia      <tibble [1~ <lm>    1982    39.9 1.29e7      978.  1.69
##  8 Afghanis~ Asia      <tibble [1~ <lm>    1987    40.8 1.39e7      852.  1.28
##  9 Afghanis~ Asia      <tibble [1~ <lm>    1992    41.7 1.63e7      649.  0.754
## 10 Afghanis~ Asia      <tibble [1~ <lm>    1997    41.8 2.22e7      635. -0.534
## # ... with 1,694 more rows
```

```
# Plot all residuals in one plot:

resids %>%
  ggplot(aes(year, resid)) +
  geom_line(aes(group = country), alpha = 1/3) +
  geom_smooth(se = FALSE)
```




```
# Plot residuals based on their country:
resids %>%
  ggplot(aes(year, resid, group = country)) +
  geom_line(alpha = 1/3) +
  facet_wrap(~continent)
```

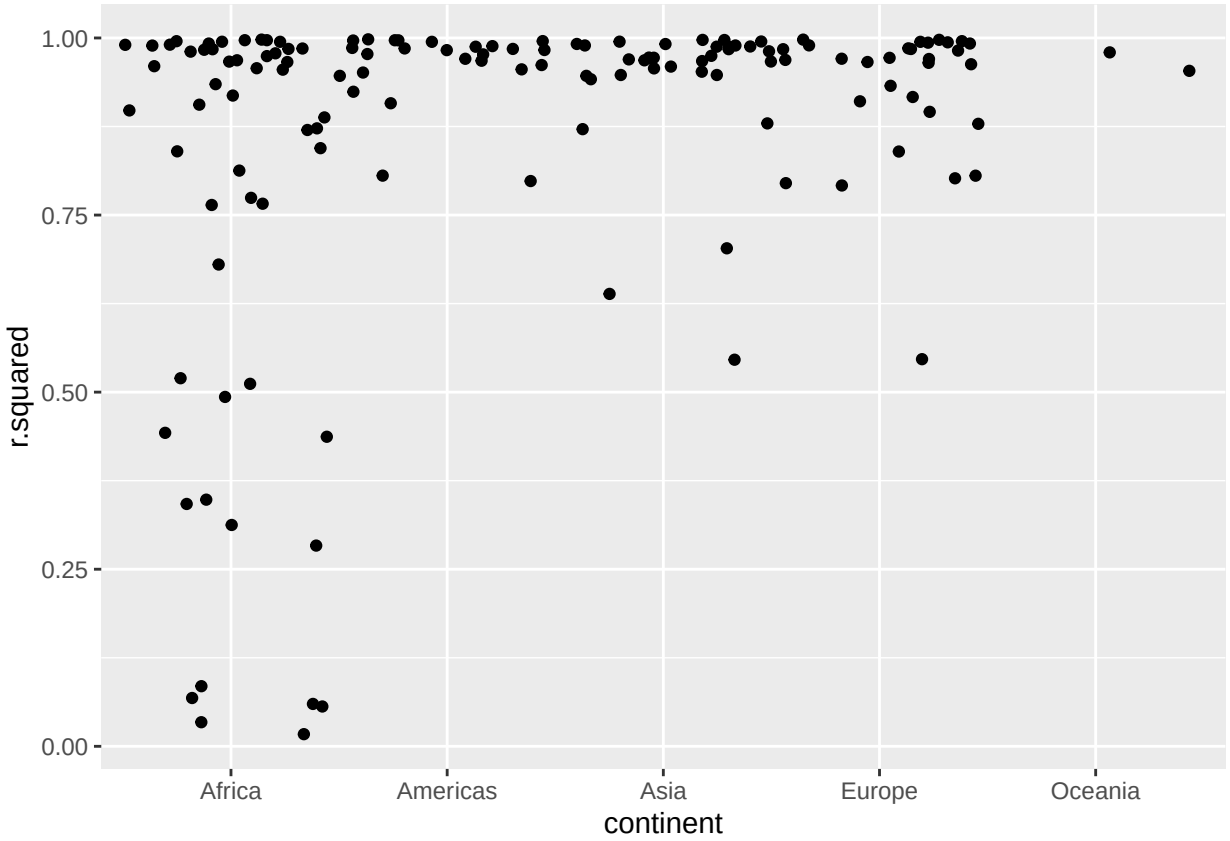

Africa
Americas
Asia
Europe
Oceania
resid
year
10
0
−10
1950 1960 1970 1980 1990 2000

Model Quality

```
# Use mutate() and unnest() to create a data frame with a row for each country:

(by_country_rsq <- by_country %>%
  mutate(glance = map(model, broom::glance)) %>%
  unnest(glance, .drop = TRUE) %>%
  arrange(r.squared))
```

```
## # A tibble: 142 x 17
## # Groups:   country, continent [142]
##    country continent data  model resids r.squared adj.r.squared sigma statistic
##    <fct>   <fct>     <lis> <lis> <list>     <dbl>         <dbl> <dbl>     <dbl>
##  1 Rwanda  Africa    <tib~ <lm>  <tibb~    0.0172      -0.0811   6.56     0.175
##  2 Botswa~ Africa    <tib~ <lm>  <tibb~    0.0340      -0.0626   6.11     0.352
##  3 Zimbab~ Africa    <tib~ <lm>  <tibb~    0.0562      -0.0381   7.21     0.596
##  4 Zambia  Africa    <tib~ <lm>  <tibb~    0.0598      -0.0342   4.53     0.636
##  5 Swazil~ Africa    <tib~ <lm>  <tibb~    0.0682      -0.0250   6.64     0.732
##  6 Lesotho Africa    <tib~ <lm>  <tibb~    0.0849      -0.00666  5.93     0.927
##  7 Cote d~ Africa    <tib~ <lm>  <tibb~    0.283        0.212    3.93     3.95
##  8 South ~ Africa    <tib~ <lm>  <tibb~    0.312        0.244    4.74     4.54
##  9 Uganda  Africa    <tib~ <lm>  <tibb~    0.342        0.276    3.19     5.20
## 10 Congo,~ Africa    <tib~ <lm>  <tibb~    0.348        0.283    2.43     5.34
## # ... with 132 more rows, and 8 more variables: p.value <dbl>, df <dbl>,
## #   logLik <dbl>, AIC <dbl>, BIC <dbl>, deviance <dbl>, df.residual <int>,
## #   nobs <int>
```

```
# At first glance, the models with the lowest r-squared seem to be all in Africa.
  # I will confirm with a plot as well as find the countries with the lowest 25% of r-squared's

# plot

by_country %>%
  mutate(glance = map(model, broom::glance)) %>%
  unnest(glance, .drop = TRUE) %>%
  arrange(r.squared) %>%
  ggplot(aes(continent, r.squared)) +
    geom_jitter(width = 0.5)
```




```
by_country
```

```
## # A tibble: 142 x 5
## # Groups:   country, continent [142]
##    country     continent data              model  resids
##    <fct>       <fct>     <list>            <list> <list>
##  1 Afghanistan Asia      <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  2 Albania     Europe    <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  3 Algeria     Africa    <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  4 Angola      Africa    <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  5 Argentina   Americas  <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  6 Australia   Oceania   <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  7 Austria     Europe    <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  8 Bahrain     Asia      <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
##  9 Bangladesh  Asia      <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
## 10 Belgium     Europe    <tibble [12 x 4]> <lm>   <tibble [12 x 5]>
## # ... with 132 more rows
```

```
quantile(by_country_rsq $r.squared, probs =  0.25)
```

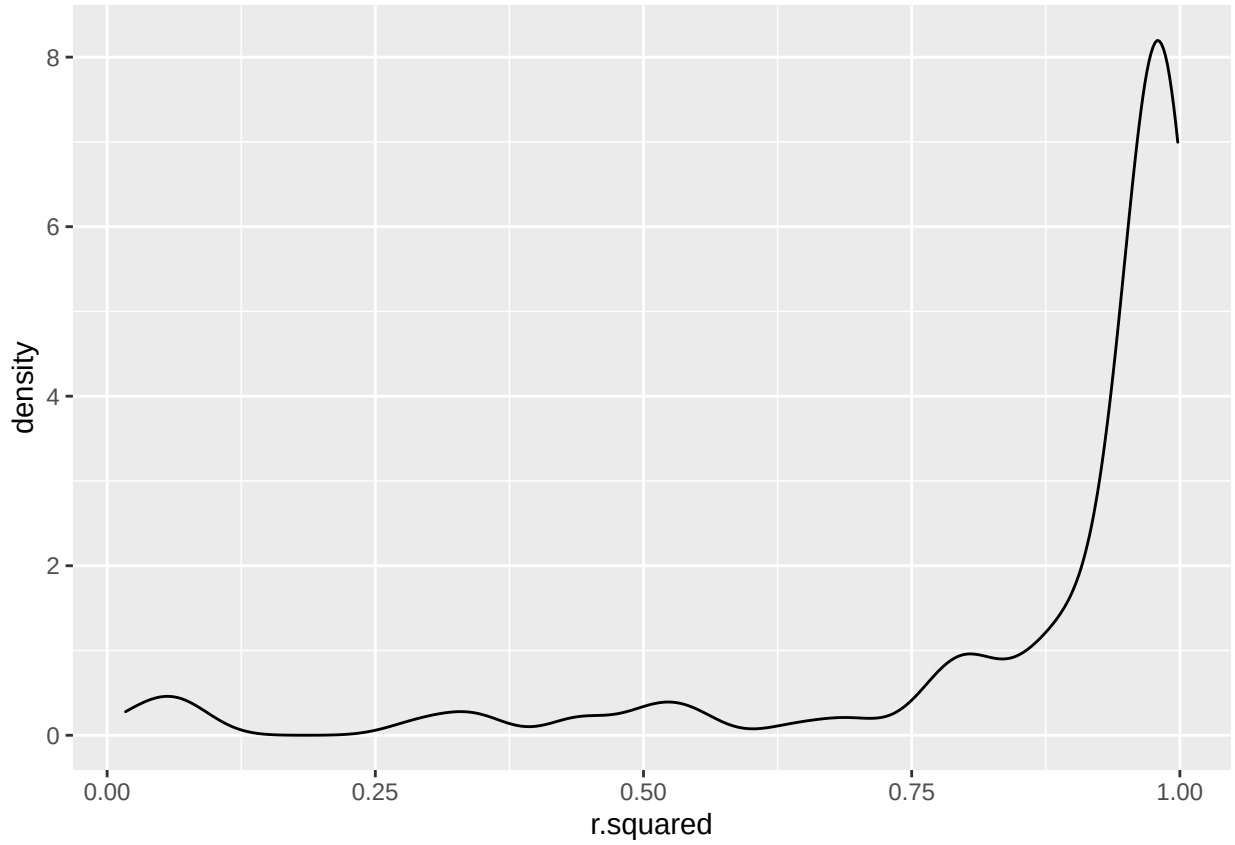
```
##      25%
## 0.873983
```

```
# 0.873983 is the 25% quantile.  This is interesting because this means that 75%
  # of the other r-squared's are larger than 0.873983 when lifeExp is regressed on years
  # which may be considered "good" per se. For fun, lets see the five-number summary
  # for the r-squared variable.  I am suspecting that the distribution of the data is heavily
  # left skewed meaning that the median is greater than the mean.
```

```r
summary(by_country_rsq$r.squared)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01716 0.87398 0.96763 0.86635 0.98754 0.99805
```

```r
# as suspected the median: 0.96763 is greater than the mean: 0.86635 of the data.  To show visually:

ggplot(by_country_rsq, aes(x=r.squared)) + geom_density()
```




```r
# Instead we'll look at the bottom 15%

quantile(by_country_rsq $r.squared, probs =  0.15)
```

```
##       15%
## 0.7672385
```

```r
rsq_bottom_15<- filter(by_country_rsq, r.squared < 0.7672385 )

view(rsq_bottom_15 %>%
  arrange(desc(r.squared)))

# Now let's look at the bottom 5%

quantile(by_country_rsq $r.squared, probs =  0.05)
```

```
##        5%
## 0.3139529
```

```r
rsq_bottom_05<- filter(by_country_rsq, r.squared < 0.3139529 )

view(rsq_bottom_05 %>%
  arrange(desc(r.squared)))
```

```r
table(rsq_bottom_15$continent == "Africa") # 18 TRUE
```

```
##
## FALSE  TRUE
##     4    18
```

```r
table(rsq_bottom_15$continent == "Asia") # 3 TRUE
```

```
##
## FALSE  TRUE
##    19     3
```

```r
table(rsq_bottom_15$continent == "Europe") # 1 TRUE
```

```
##
## FALSE  TRUE
##    21     1
```

```r
table(rsq_bottom_15$continent == "South America")  # 0 TRUE
```

```
##
## FALSE
##    22
```

```r
table(rsq_bottom_15$continent == "North America") # 0 TRUE
```

```
##
## FALSE
##    22
```

```r
table(rsq_bottom_15$continent == "Oceania") # 0 TRUE
```

```
##
## FALSE
##    22
```

```r
# Higher outbreaks of disease such as the HIV/AIDS epidemic and events such as the Rwandan genocide
  # - events that decreased life expectancy in some African countries
    #(more variation other than what we have controlled for)
  # - contribute to the low R-squared of the models
    # (especially for a continent like Africa which overall has low levels of development
    # when compared to the rest of the world )
```

```
# Though the graphics are not great for analysis, we can see the how spread out the
  # outliers are for North America. This isn't necessary bad, because it means that
  # potentially a country's GDP is rapidly increasing from one year to the next.
  # In this particular visual, we  group the box plots by continent.
  # However, because of this visual, we can begin to investigate why
  # some country's GDP are rapidly increasing across continents.
ggplot(gapminder3, aes(x = continent, y = gdp, colour = continent)) + geom_boxplot()
```