Assignment 2

Feiyu Zheng (fz114)

2/4/2022

1. Download the dataset on restaurant inspection in csv format.

```
# read data from csv file and convert to tibble
dataset <- as_tibble(read_csv("./NYRestaurantInspection2022.csv",
                              na = c("", "N/A"), show_col_types = F))
# show dataset
dataset
```

```
## # A tibble: 373,818 x 26
##       CAMIS DBA             BORO  BUILDING STREET ZIPCODE PHONE `CUISINE DESCR~`
##       <dbl> <chr>           <chr> <chr>    <chr>    <dbl> <chr> <chr>
##  1 50017056 SILVER SPOON D~ Quee~ 5821     JUNCT~   11373 7183~ American
##  2 50081611 HOT GINGER      Quee~ 7332     BELL ~   11364 6467~ Japanese
##  3 50018945 LOI ESTIATORIO  Manh~ 132      WEST ~   10019 2127~ Greek
##  4 50016779 INCREDIBOWL     Quee~ 133-35   ROOSE~   11354 9173~ Chinese
##  5 41432344 LOLLIPOPS ICE ~ Bronx 4120     BAYCH~   10466 7189~ Frozen Desserts
##  6 50078193 99 CENT TASTY ~ Manh~ 383      CANAL~   10013 2129~ Pizza
##  7 50069659 DUNKIN          Quee~ 21102    JAMAI~   11428 7184~ Donuts
##  8 40521003 TOPAZE RESTAUR~ Broo~ 1875     UTICA~   11234 7184~ Caribbean
##  9 41595314 PITA GRILL      Manh~ 1083     2 AVE~   10022 2127~ Jewish/Kosher
## 10 41631475 SUBWAY (+ STAR~ Quee~ 14716    NORTH~   11354 7183~ Sandwiches
## # ... with 373,808 more rows, and 18 more variables: `INSPECTION DATE` <chr>,
## #   ACTION <chr>, `VIOLATION CODE` <chr>, `VIOLATION DESCRIPTION` <chr>,
## #   `CRITICAL FLAG` <chr>, SCORE <dbl>, GRADE <chr>, `GRADE DATE` <chr>,
## #   `RECORD DATE` <chr>, `INSPECTION TYPE` <chr>, Latitude <dbl>,
## #   Longitude <dbl>, `Community Board` <dbl>, `Council District` <chr>,
## #   `Census Tract` <chr>, BIN <dbl>, BBL <dbl>, NTA <chr>
```

(1a) From a new data frame restricted to restaurants in Queens with cuisine equal to "Pizza".

```
# filter restaurants in Queens with "Pizza" as cuisine
queensPizzaRestaurants <- dataset %>%
  filter(BORO == "Queens", `CUISINE DESCRIPTION` == "Pizza") %>%
  as.data.frame()
# show the data
head(queensPizzaRestaurants, 10)
```

```
##       CAMIS                                DBA   BORO BUILDING           STREET
## 1  50010850                  RETRO PIZZA CAFE Queens   41-02A          BROADWAY
## 2  40614489 FRESH MEADOW'S PIZZA & RESTAURANT Queens    19509         69 AVENUE
## 3  50117597                         BARAKAH'S Queens     9002        CORONA AVE
## 4  50117597                         BARAKAH'S Queens     9002        CORONA AVE
## 5  40662141                    J AND D PIZZA Queens    98-53           63 ROAD
```

```
## 6  50044118                   GALLERIA PIZZA Queens     9520         101ST AVE
## 7  50018532         ARTICHOKE BASILLE'S PIZZA Queens     2256           31ST ST
## 8  50076257                99 CENTS HOT PIZZA Queens    16417       JAMAICA AVE
## 9  40366002                  MARGHERITA PIZZA Queens    16304    JAMAICA AVENUE
## 10 50075314                     MARIO'S PIZZA Queens    14929 GUY R BREWER BLVD
##    ZIPCODE      PHONE CUISINE DESCRIPTION INSPECTION DATE
## 1    11103 3476124460               Pizza      11/07/2019
## 2    11365 7182172700               Pizza      04/26/2019
## 3    11373 9293039643               Pizza      01/06/2022
## 4    11373 9293039643               Pizza      01/06/2022
## 5    11374 7182754347               Pizza      03/28/2019
## 6    11416 7188453973               Pizza      12/12/2018
## 7    11105 7182158100               Pizza      02/29/2020
## 8    11432 6314807330               Pizza      08/06/2019
## 9    11432 7186575780               Pizza      01/28/2022
## 10   11434 7186563104               Pizza      11/12/2019
##                                           ACTION VIOLATION CODE
## 1  Violations were cited in the following area(s).            10F
## 2  Violations were cited in the following area(s).            06C
## 3  Violations were cited in the following area(s).            10E
## 4  Violations were cited in the following area(s).            10E
## 5  Violations were cited in the following area(s).            10F
## 6  Violations were cited in the following area(s).            08A
## 7  Violations were cited in the following area(s).            10C
## 8  Violations were cited in the following area(s).            10F
## 9  Violations were cited in the following area(s).            08A
## 10 Violations were cited in the following area(s).            10B
##
## 1  Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surf
## 2
## 3
## 4
## 5  Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surf
## 6
## 7
## 8  Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surf
## 9
## 10                                              Plumbing not properly installed or maintained; anti-s
##    CRITICAL FLAG SCORE GRADE GRADE DATE RECORD DATE
## 1   Not Critical     9     A 11/07/2019  02/04/2022
## 2       Critical    10     A 04/26/2019  02/04/2022
## 3   Not Critical    10     A 01/06/2022  02/04/2022
## 4   Not Critical    10     A 01/06/2022  02/04/2022
## 5   Not Critical     3     A 03/28/2019  02/04/2022
## 6   Not Critical    28  <NA>       <NA>  02/04/2022
## 7   Not Critical    27     B 02/29/2020  02/04/2022
## 8   Not Critical     2     A 08/06/2019  02/04/2022
## 9   Not Critical     9     A 01/28/2022  02/04/2022
## 10  Not Critical    11     A 11/12/2019  02/04/2022
##                             INSPECTION TYPE Latitude Longitude Community Board
## 1          Cycle Inspection / Re-inspection 40.75852 -73.91806              401
## 2          Cycle Inspection / Re-inspection 40.73693 -73.77799              408
## 3  Pre-permit (Operational) / Re-inspection 40.74164 -73.87479              404
## 4  Pre-permit (Operational) / Re-inspection 40.74164 -73.87479              404
```

```
## 5     Cycle Inspection / Initial Inspection 40.73240 -73.85771             406
## 6     Cycle Inspection / Initial Inspection 40.68452 -73.84507             409
## 7          Cycle Inspection / Re-inspection 40.77499 -73.91209             401
## 8          Cycle Inspection / Re-inspection 40.70523 -73.79548             412
## 9          Cycle Inspection / Re-inspection 40.70467 -73.79682             412
## 10    Cycle Inspection / Initial Inspection 40.65758 -73.76753             413
##    Council District Census Tract     BIN        BBL  NTA
## 1                26       015900 4011090 4006770135 QN70
## 2                23       134700 4439390 4071170007 QN41
## 3                25       046100 4045613 4018470049 QN29
## 4                25       046100 4045613 4018470049 QN29
## 5                29       071701 4050419 4020860038 QN18
## 6                32       004001 4189313 4091020010 QN53
## 7                22       011500 4017623 4008440060 QN72
## 8                24       044601 4209587 4097940014 QN61
## 9                27       044601 4216196 4101510001 QN61
## 10               31       032000 4286176 4134100045 QN03
```

(1b) What are the 5 most frequently inspected restaurants (use the variable "DBA" in the data frame)?

```
frequentlyInspected <- queensPizzaRestaurants %>%
  group_by(DBA) %>% # group by the name of restaurants
  count(DBA, sort = TRUE) %>% # compute the frequency of inspection and sort in descending order
  head(5) # choose the 5 most frequently inspected restaurants

# show the data
frequentlyInspected
```

```
## # A tibble: 5 x 2
## # Groups:   DBA [5]
##   DBA                   n
##   <chr>             <int>
## 1 DOMINO'S            110
## 2 PAPA JOHN'S         100
## 3 PAPA JOHN'S PIZZA    70
## 4 LA BELLA PIZZA       56
## 5 ROSA'S PIZZA         52
```

As the above result shows, the 5 most frequently inspected restaurants are DOMINO'S, PAPA JOHN'S, PAPA JOHN'S PIZZA, LA BELLA PIZZA, and ROSA'S PIZZA.

(1c) On what dates has pizza parlor "SUSANO'S PIZZERIA & RESTAURANT" been inspected?

```
susanosPizza <- queensPizzaRestaurants %>%
  filter(DBA == "SUSANO'S PIZZERIA & RESTAURANT") %>% # filter by name
  select(DBA, `INSPECTION DATE`) %>% # choose column DBA and INSPECTION DATE
  distinct() %>% # only show distinct date
  arrange(desc(mdy(`INSPECTION DATE`))) # sort the data by INSPECTION DATE in
                                        # descending order
# show the result
susanosPizza
```

```
##                              DBA INSPECTION DATE
```

```
## 1  SUSANO'S PIZZERIA & RESTAURANT      01/08/2020
## 2  SUSANO'S PIZZERIA & RESTAURANT      12/09/2019
## 3  SUSANO'S PIZZERIA & RESTAURANT      08/14/2019
## 4  SUSANO'S PIZZERIA & RESTAURANT      07/31/2019
## 5  SUSANO'S PIZZERIA & RESTAURANT      03/25/2019
## 6  SUSANO'S PIZZERIA & RESTAURANT      03/14/2019
## 7  SUSANO'S PIZZERIA & RESTAURANT      09/25/2018
## 8  SUSANO'S PIZZERIA & RESTAURANT      09/11/2018
## 9  SUSANO'S PIZZERIA & RESTAURANT      04/13/2018
## 10 SUSANO'S PIZZERIA & RESTAURANT      03/15/2018
## 11 SUSANO'S PIZZERIA & RESTAURANT      03/01/2017
```

The above result lists the dates when the pizza parlor "SUSANO'S PIZZERIA & RESTAURANT" was inspected.

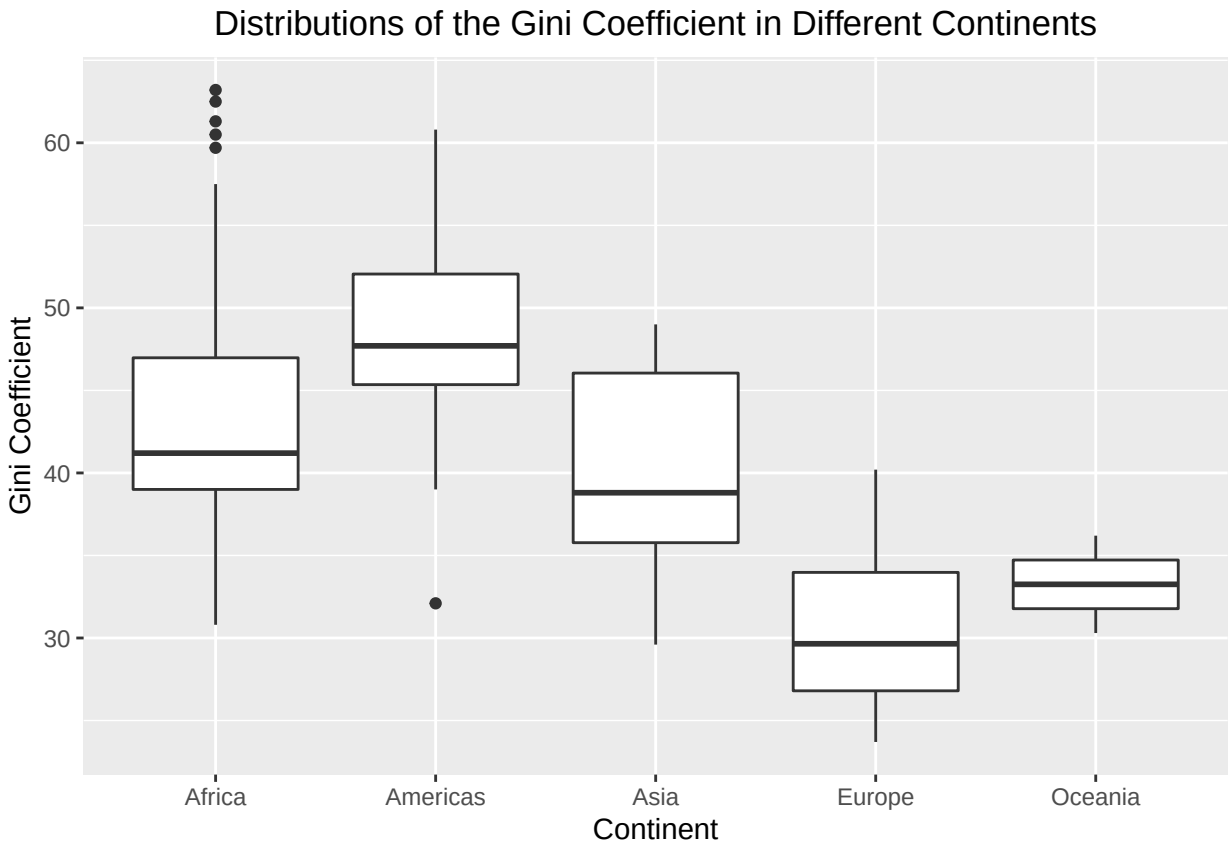
2. The file "gapminder_2007_gini.tsv" is in the Files > Lecture materials > Lecture2_Jan31 folder. It is a subset of the 2007 Gapminder data merged with recent Gini coefficient data.

```
# read the data from the tsv file
gapminder2007 <- read_tsv("./gapminder_2007_gini.tsv", show_col_types = F)
# show the data
gapminder2007
```

```
## # A tibble: 108 x 7
##    country                 continent  year lifeExp       pop gdpPercap  gini
##    <chr>                   <chr>     <dbl>   <dbl>     <dbl>     <dbl> <dbl>
##  1 Albania                 Europe     2007    76.4   3600523     5937.  29
##  2 Algeria                 Africa     2007    72.3  33333216     6223.  35.3
##  3 Argentina               Americas   2007    75.3  40301927    12779.  45.8
##  4 Australia               Oceania    2007    81.2  20434176    34435.  30.3
##  5 Austria                 Europe     2007    79.8   8199783    36126.  29.2
##  6 Bangladesh              Asia       2007    64.1 150448339     1391.  32.1
##  7 Belgium                 Europe     2007    79.4  10392226    33693.  25.9
##  8 Benin                   Africa     2007    56.7   8078314     1441.  36.5
##  9 Bolivia                 Americas   2007    65.6   9119152     3822.  46.6
## 10 Bosnia and Herzegovina Europe     2007    74.9   4552198     7446.  36.2
## # ... with 98 more rows
```

(2a) Create a plot to compare the distributions (e.g., central tendency, dispersion) of the Gini coefficient in different continents. (Hint: Use a boxplot)

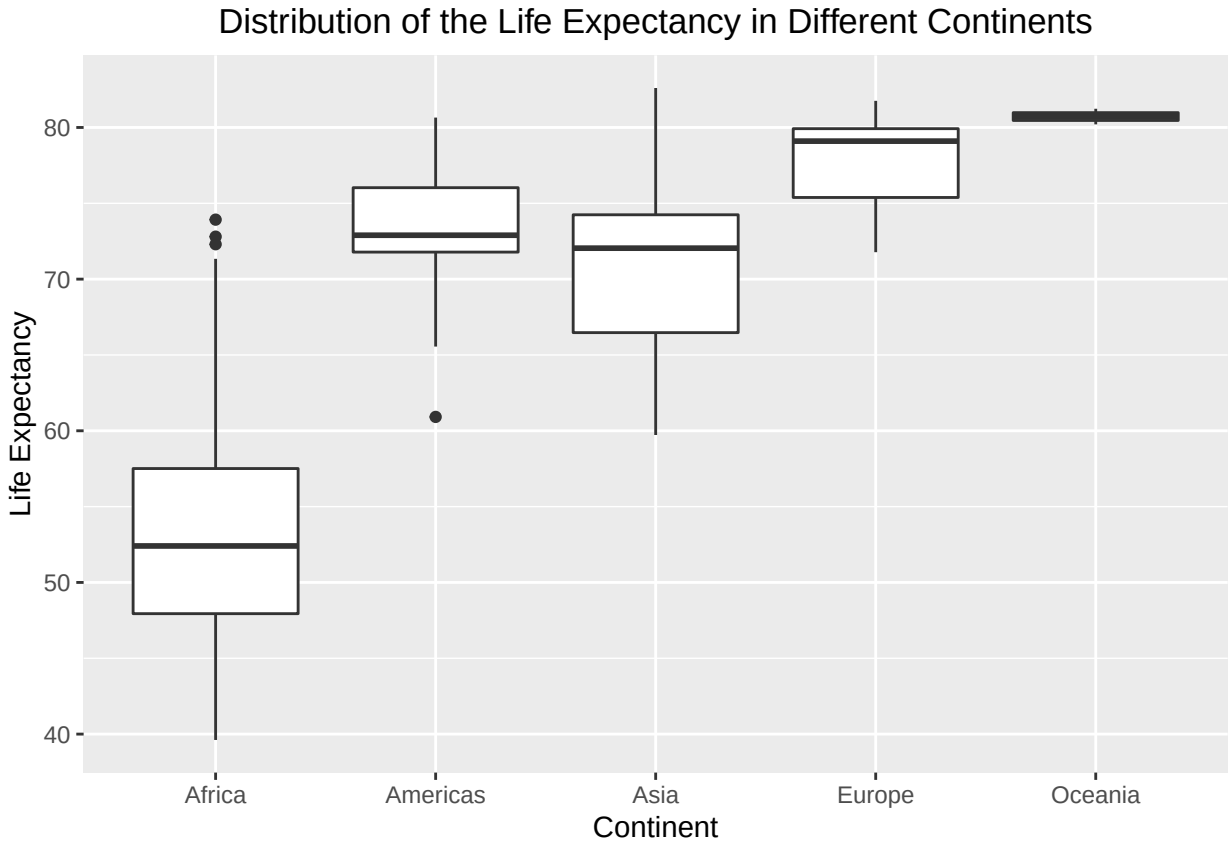
```
gapminder2007 %>%
  ggplot(aes(x = continent, y = gini)) +
  geom_boxplot() + # box plot
  labs(
    x = "Continent",
    y = "Gini Coefficient",
    title = "Distributions of the Gini Coefficient in Different Continents"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```



(2b) Does the Gini coefficient appear to have any impact on the life expectancy in 2007? Explain your answer using a plot, classified by continents.

We first draw the box plot of the life expectancy and then compare it with the above gini coefficient graph.

```
gapminder2007 %>%
  ggplot(aes(x = continent, y = lifeExp)) +
  geom_boxplot() + # box plot
  labs(
    x = "Continent",
    y = "Life Expectancy",
    title = "Distribution of the Life Expectancy in Different Continents"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```



Based the above box plot about life expectancy and the box plot in 2a about the Gini coefficient, we can see that the Gini coefficient can reflect the life expectancy. A low Gini coefficient may indicate a relative high life expectancy and when people's Gini coefficient tend to be equal, their life expectancy tend to be the same age as well.

3. Using the original gapminder data frame, please generate a data frame with a new variable called gdp by multiplying the population size by the gdp per capita. To make those large numbers more understandable, please form an additional new variable called gdp_ratio equal to the gdp divided by the gdp of the United States in 2007. Find the median gdp_ratio by continent and year, and then plot the median gdp_ratio over time, distinguishing the continents. Please use both points and lines for the plot.

Install the original gapminder package and show the data.

```
gapminder # the original gapminder
```

```
## # A tibble: 1,704 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.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
```

```
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

Add two columns: gdp and gdp_ratio.

```
# convert tibble to data frame
gapminderWithGDP <- as.data.frame(gapminder)
# generate a new column called gdp
gapminderWithGDP$gdp <- gapminderWithGDP$pop * gapminderWithGDP$gdpPercap
# compute and get the gdp of the United States in 2007
us2007GDP <- gapminderWithGDP %>%
  filter(year == 2007, country == "United States") %>%
  pull(gdp)

# generate a new column called gdp_ratio
gapminderWithGDP$gdp_ratio <- round(gapminderWithGDP$gdp / us2007GDP, 5)

# show the sorted data by continent, year and gdp_ratio
gapminderWithGDP %>%
  arrange(continent, year, gdp_ratio) %>%
  head(10)
```

```
##                  country continent year lifeExp     pop gdpPercap       gdp
## 1  Sao Tome and Principe    Africa 1952  46.471   60011  879.5836  52784691
## 2                Comoros    Africa 1952  40.715  153936 1102.9909 169790013
## 3               Djibouti    Africa 1952  34.812   63149 2669.5295 168578117
## 4      Equatorial Guinea    Africa 1952  34.482  216964  375.6431  81501035
## 5                 Gambia    Africa 1952  30.000  284320  485.2307 137960781
## 6          Guinea-Bissau    Africa 1952  32.500  580653  299.8503 174108987
## 7                Lesotho    Africa 1952  42.138  748747  298.8462 223760205
## 8               Botswana    Africa 1952  47.622  442308  851.2411 376510766
## 9              Swaziland    Africa 1952  41.407  290243 1148.3766 333308277
## 10               Eritrea    Africa 1952  35.928 1438760  328.9406 473266516
##    gdp_ratio
## 1      0e+00
## 2      1e-05
## 3      1e-05
## 4      1e-05
## 5      1e-05
## 6      1e-05
## 7      2e-05
## 8      3e-05
## 9      3e-05
## 10     4e-05
```

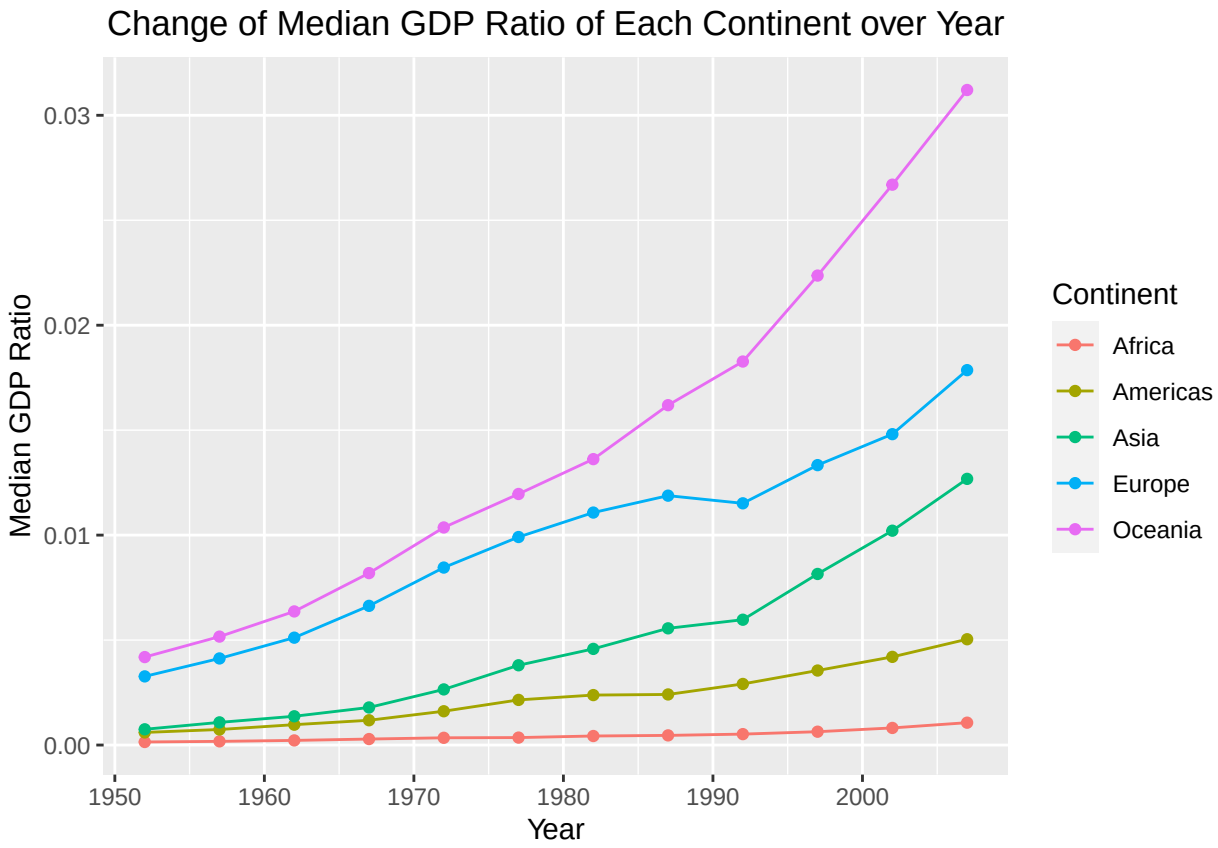
Compute the median gdp_ratio by continent and year and plot the data via points and lines.

```
# suppress the warning message of dplyr
options(dplyr.summarise.inform = F)

# compute the median gdp_ratio by continent and year
gapminderWithMedian <- gapminderWithGDP %>%
```

```
  group_by(year, continent) %>%
  summarize(median_gdp_ratio = median(gdp_ratio))

# show the sorted data by continent and year
gapminderWithMedian %>% arrange(continent, year)
```

```
## # A tibble: 60 x 3
## # Groups:   year [12]
##     year continent median_gdp_ratio
##    <int> <fct>                <dbl>
##  1  1952 Africa            0.000145
##  2  1957 Africa            0.000175
##  3  1962 Africa            0.00022
##  4  1967 Africa            0.000285
##  5  1972 Africa            0.000345
##  6  1977 Africa            0.000355
##  7  1982 Africa            0.00043
##  8  1987 Africa            0.00046
##  9  1992 Africa            0.00052
## 10  1997 Africa            0.000635
## # ... with 50 more rows
```

```
# plot the median gdp_ration over time in different continents
gapminderWithMedian %>%
  ggplot(aes(x = year, y = median_gdp_ratio, color = continent)) +
  geom_line() + # line plot
  geom_point() + # point plot
  labs(
    x = "Year",
    y = "Median GDP Ratio",
    color = "Continent",
    title = "Change of Median GDP Ratio of Each Continent over Year"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```


Change of Median GDP Ratio of Each Continent over Year
Median GDP Ratio
0.03
0.02
0.01
0.00
1950
1960
1970
1980
1990
2000
Year
Continent
Africa
Americas
Asia
Europe
Oceania