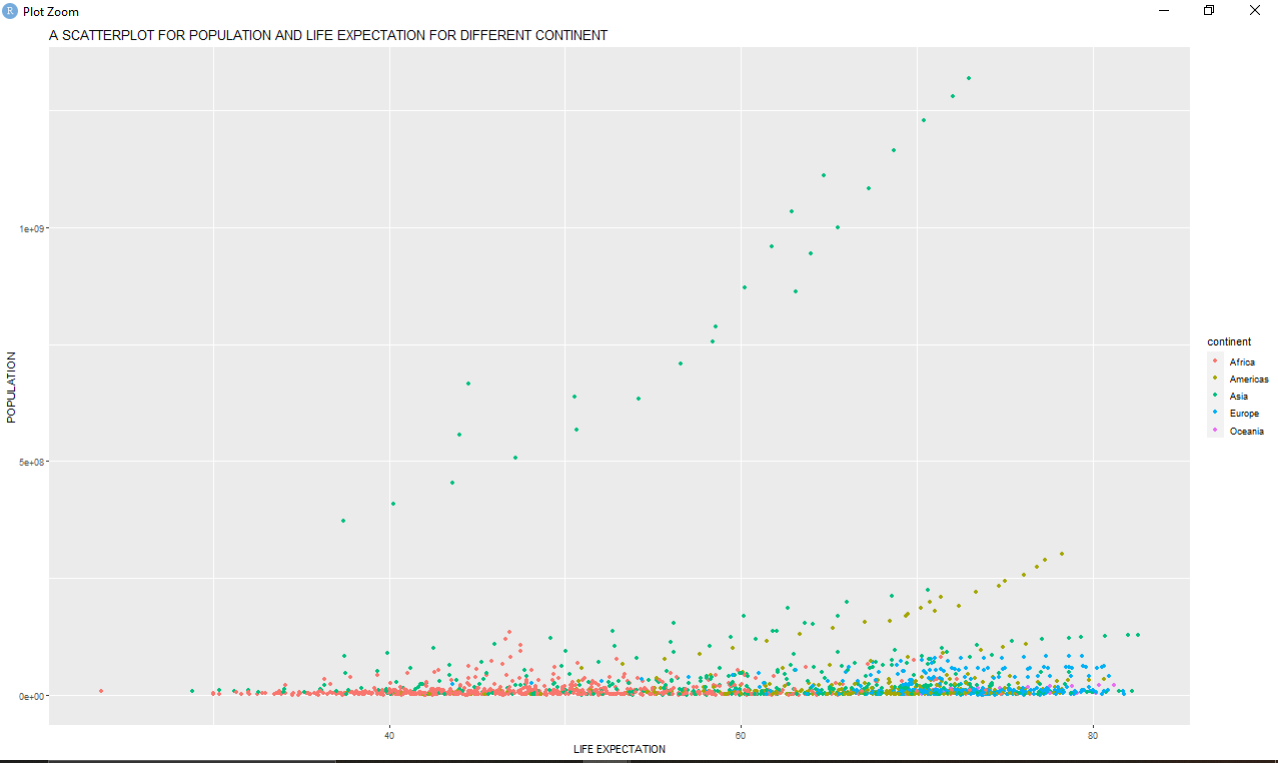
**AN EXTENSIVE EXAMINATION OF EXPANDED DATA SETS FOR FAPMINDER ANALYSIS USING R**

**Introduction**In the field of analysis of data, the importance of strong datasets cannot be emphasised enough. This project expands on the work done in the week prior's assignment, focusing on the detailed examination of the initial evaluation performed on the gapminder data using R. The goal is to enhance the level of our understanding by increasing the amount of information in the data sets by at least two-fold. Our objective is to uncover any subtle details that may have remained hidden in the first study, thereby providing a more thorough comprehension of the deeper trends.  
  
The preliminary analysis utilised scatter plots and box plots to visually represent the correlations within the data. The visualisations were useful tools for extracting insights using the gapminder dataset. Nevertheless, in our quest for a more comprehensive understanding, we acknowledged the necessity to enlarge the dataset and reassess our conclusions. This report provides a detailed comparison of the initial findings with the results obtained from the expanded datasets. We aim to clarify the causes for any differences or similarities discovered between the two investigations by using clear visualisations and detailed explanations.  
  
As we delve into the complexities of this extensive research, we rely on the gapminder dataset as the foundation of our exploration. The next sections will describe the progression from the first scatter plots and box plots to the enhanced visualisations obtained from the extended datasets. We want to go beyond data accumulation and instead focus on extracting significant insights, identifying previously unnoticed trends, and making valuable contributions to the continuously developing field of data analysis.

We start the assignment by providing the initial plot created in the previous assignment. These plots created were a scatterplot and a boxplot.

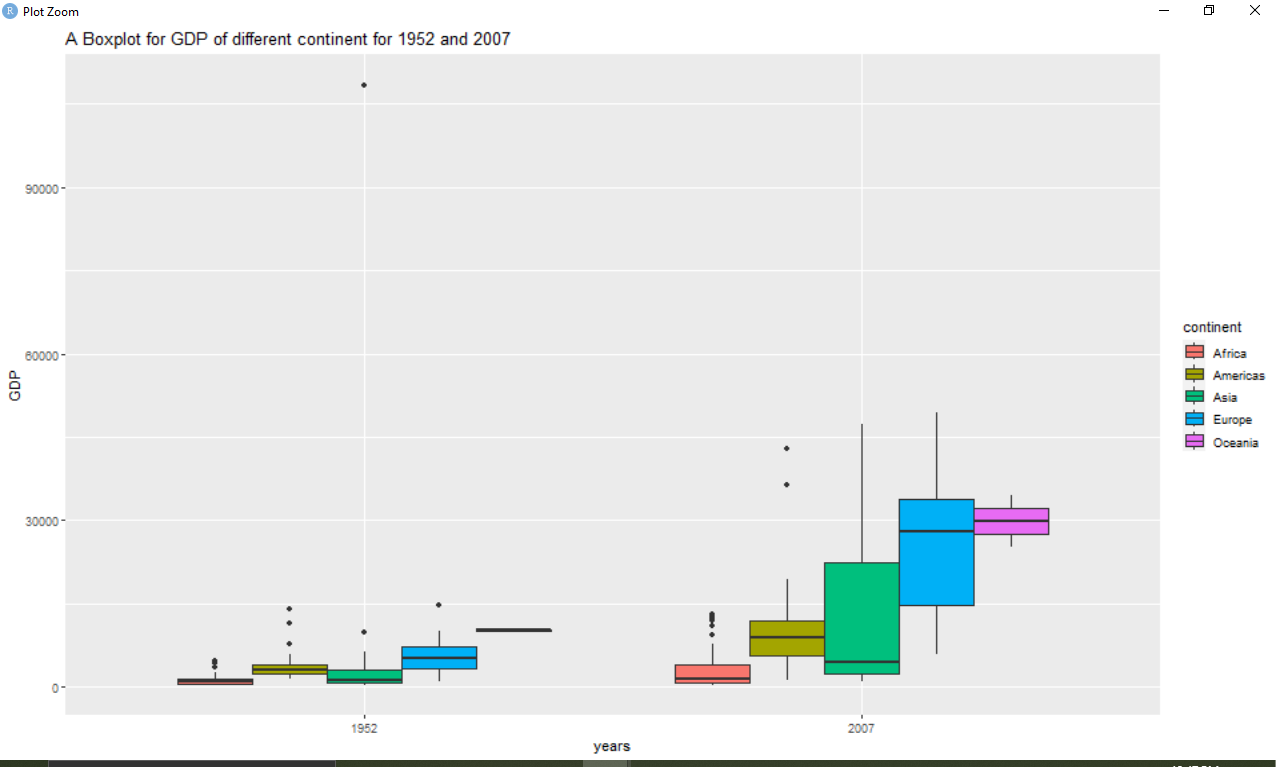
**Initial plots from the previous assignment**



The scatterplot examination of population, lifespan, and continent variables offers valuable insights into the interrelationships among these elements. Below are the main discoveries as per the previous assignment:

1. Correlation between Population and Life Expectancy: - A conspicuous trend is noticed wherein lower life expectancy (under 40) are linked to shorter life expectancies.   
Specifically, if the population is below 40, the combined population of continents within this range typically falls below 500 million.   
2. Africa Continent: - Africa is distinguished as the continent with the most minimal population and life expectancy. In the scatterplot, Africa has a minimum population of 7,290,203 and a minimum average life expectancy of 23.6.   
3. Asia Continent: - Asia is the region with the largest documented population and longest life expectancy. In the scatterplot, the maximum population recorded for Asia is 127,467,972, whereas the greatest lifespan is 82.6.   
4. Europe Continent: - Europe is renowned for its comparatively low population size in relation to other continents, while boasting a much greater life expectancy. This implies that despite Europe having a relatively lower population in comparison to other continents, it has a considerably greater life expectancy.   
  
These observations emphasize the intricate relationship between population density and lifespans across various continents. The discrepancies between continents highlight the contrasting demographic and health attributes of different locations, emphasizing the significance of taking into account many factors for a thorough comprehension of global demographic patterns.

Next plot produced in the previous assignment is a boxplot which is shown below.



The analysis of the plot, with GDP on the y-axis and the year variable on the x-axis (considering years 1952 and 2007), reveals insightful observations about the economic landscape of different continents.

In the year 1952, the disparities in GDP among continents are evident. Notably, Africa emerges as the continent with the lowest GDP during this period. In contrast, Oceania stands out with the highest GDP count, closely followed by the Europe continent. This snapshot provides a historical perspective on the economic conditions across continents, showcasing the varying levels of development.

Shifting the focus to the year 2007, a notable transformation in the GDP landscape is observed. Oceania takes the lead as the continent with the highest GDP, followed closely by Europe and Asia. In contrast, Africa finds itself at the lower end in terms of GDP among the continents. This comparison between 1952 and 2007 highlights the dynamic nature of economic growth, with certain continents experiencing substantial advancements while others undergo more moderate changes.

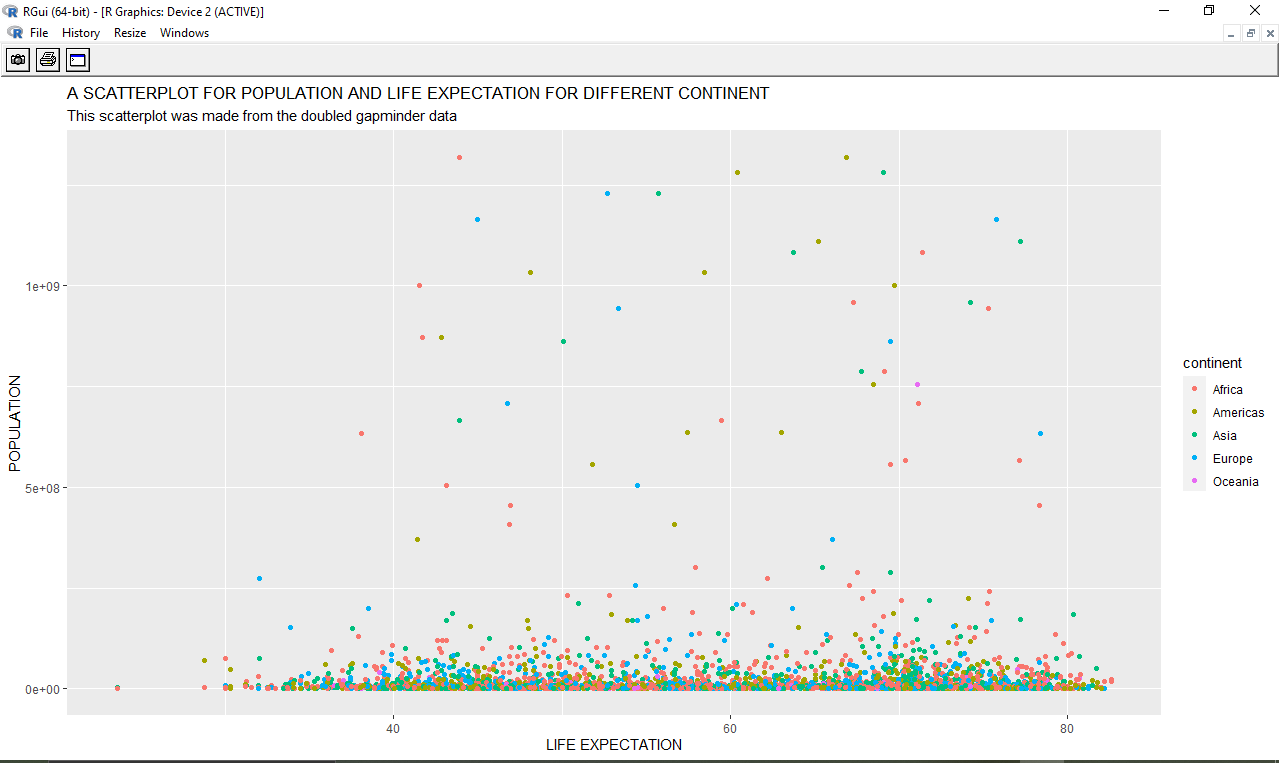
A key observation is the drastic increase in GDP for all continents over the considered period. Asia stands out with the highest GDP increase, indicating rapid economic growth. Europe follows closely, showcasing significant advancements as well. On the other hand, the remaining continents exhibit a comparatively slower rate of GDP increase. This pattern underscores the diverse economic trajectories of continents, with some experiencing remarkable progress while others evolve at a more gradual pace. Overall, the analysis provides valuable insights into the changing economic landscape over the selected years, offering a nuanced understanding of global economic trends.

Next we are going to increase our data twice as much and then recreate the above plots and try to spot the differences.

**Doubling the data**

We create a doubling data function named `doubling\_data\_func` which accepts an initial data frame (`original\_data`) as an argument. This function expands the size of the data frame by adding identical data to itself, then thereafter rearranges the values in each variable separately using a for loop. The function creates a new data frame called `doubled\_data` that has the same dimensions as the original data and gives column names to it. The for loop sequentially traverses each variable in the original data, merging the original variable values with their randomly rearranged equivalents using the `sample` function, and subsequently assigning the rearranged values to the corresponding variables in the new data frame. The resultant data frame, named `doubled\_data`, contains twice as many rows as the original data frame. Additionally, the values in each variable have been rearranged. Ultimately, the function is executed on a data frame named `gapminder\_data`, and the outcome is saved in a fresh data frame named `doubled\_gapminder\_data`.

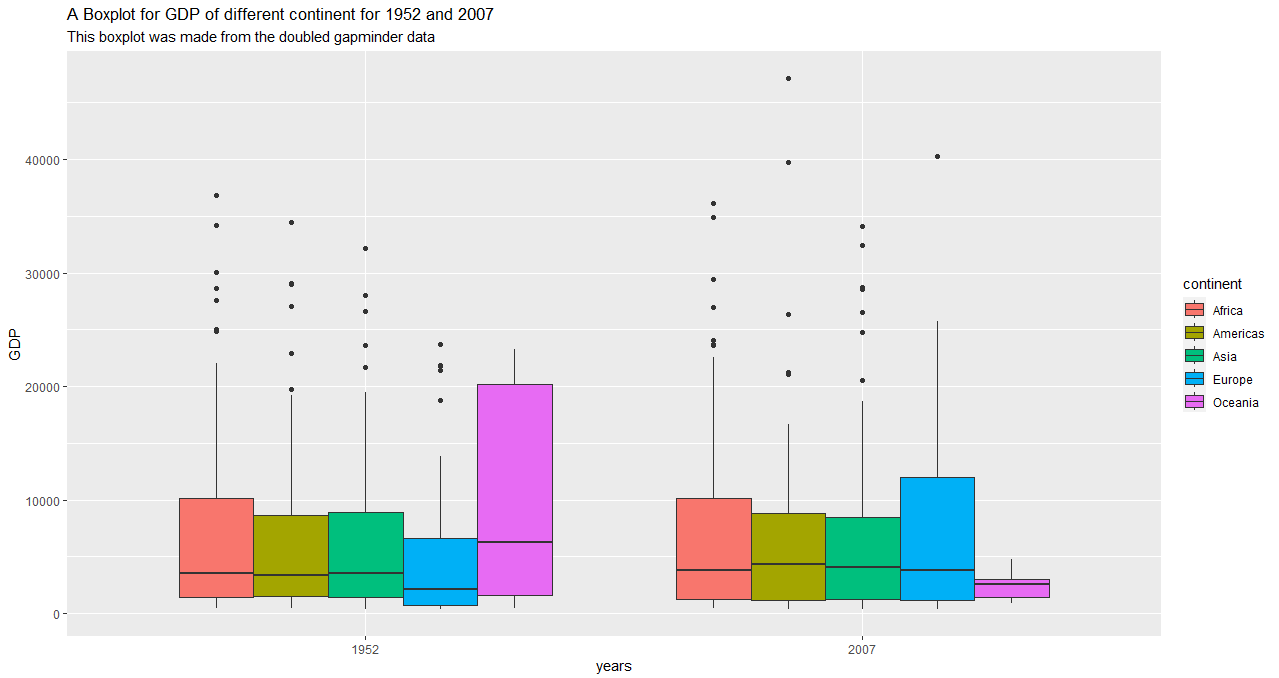
Using the new data we create new plots with similar codes displayed above. New scatterplot is shown below.

****

**Difference between the scatterplot from original and the scatterplot from the doubled data**

The changes between the original data scatter plot and the data scatterplot that was duplicated are significant and indicate the effects of doubled data set on visualization. Population and life expectancy are clearly correlated in the first figure, especially when life duration is below 40, which causes populations to drop below 500 million. On the other hand, the following graphic shows data points that are randomly distributed and do not show any pattern in the relationship between life expectancy and population size. The life expectancy deviates from the original patterns and differs greatly between continents. In terms of how the African continent is shown, the original image links Africa to the regions with the lowest life expectancy and demographics, but the replicated data plot displays data points randomly, making it difficult to discern any distinct traits of the continent's people. Asia is shown as having the largest population in the first picture, but the second plot shows tightly packed data values for Asia with a population under 500 million, reflecting similar trends in other continents. Duplicating data skews the information, producing a disjointed depiction that hides important patterns and characteristics of the continents. For effective analysis, it is crucial to maintain the initial data's integrity because any changes, such duplication, might compromise the dataset's correctness and result in incorrect conclusions. The scatterplots' deceptive portrayal highlights the possibility of misinterpreting separate things, such continents, and distorting relationships between variables. In order to preserve important patterns and insights and guarantee accurate results in visualization of data and analysis, it is imperative to maintain the original dataset.

Next we create new boxplot from the doubled data as shown below.

****

**Differences between the initial gapminder data’s boxplot and the doubled gapminder data boxplot**

The discrepancies between the boxplots produced with the original Gapminder dataset and the modified dataset, in which the data was doubled, highlight serious distortions in GDP figures. 1952 was shown in the original boxplot as having the lowest GDP—less than 10,000—in Africa and the greatest GDP in Oceania. But the doubled data boxplot showed Africa to have an atypically high GDP of about 10,000, suggesting that the data change caused a distortion. The data boxplot was modified to show that Europe had a higher GDP than other continents and that Oceania had the lowest GDP, which contradicted the previous plot's observation of a changing GDP landscape. The contrast of GDP changes between 1952 and 2007 was also fraudulently depicted; the second boxplot, which showed a decline for several continents, failed to indicate the anticipated increase. These differences highlight the negative effects of double the data, which result in connections changing, significant information being lost, and inaccurate comparisons of continents and their gross domestic product (GDP) patterns over time. Accurate understanding of the dynamics of the world economy depends on maintaining the authenticity of the original dataset.

**Conclusion**

To sum up, the examination of the variations between the original as well as doubled data scatter plots and boxplots reveals the significant influence of changing datasets on data processing and visualization. The effects of increasing the data are obvious whether looking at GDP differences over time or identifying trends in the connections between demographic and life expectancy between different continents. The significance of maintaining the accuracy of the underlying datasets is shown by the skewed demographic features, distorted relationships, and misrepresented GDP numbers. Such modifications run the danger of producing inaccurate interpretations and conclusions in addition to obscuring important information. Accurate evaluations of financial and demographic tendencies are hampered by the unpredictability and obscurity of the visuals created from the doubled data. In order to guarantee the accuracy of insights derived from visual representations and the dependability of studies, it is imperative to preserve data authenticity. In data science, following the original dataset exactly is still crucial to making wise choices and arriving at reliable results.