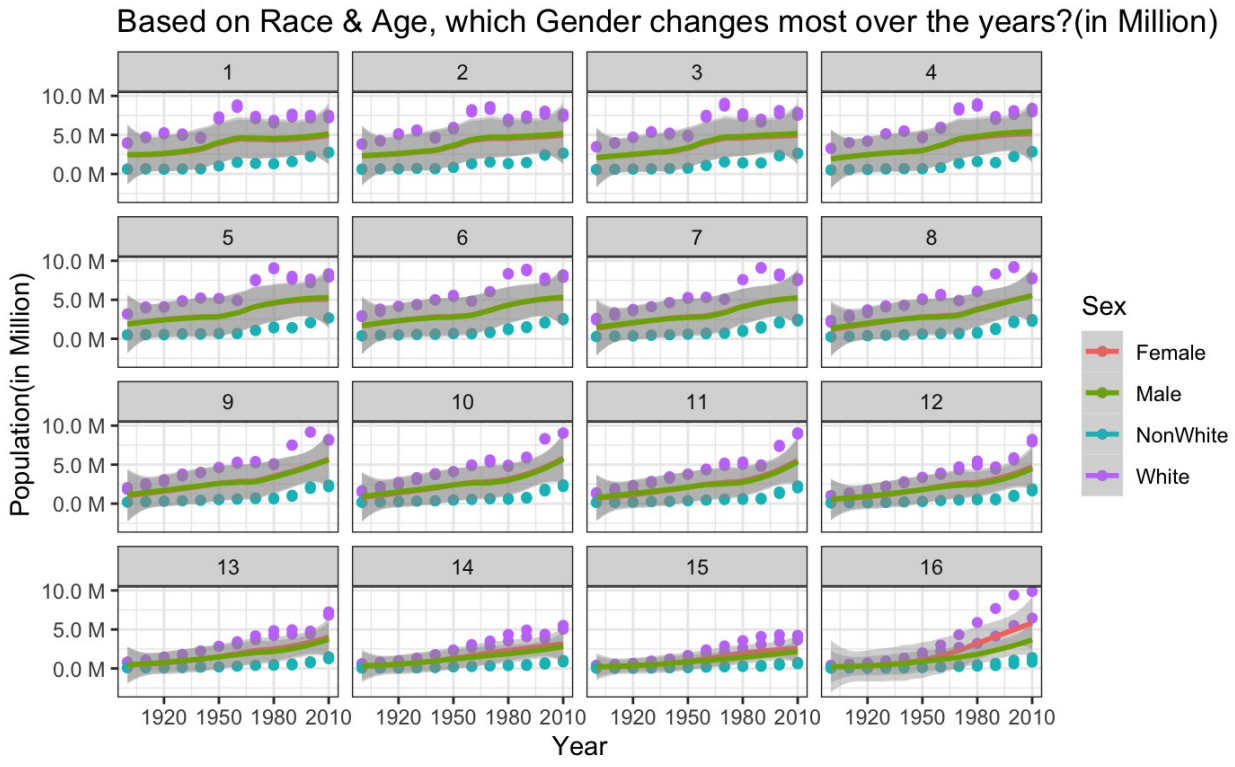
My part2 question: Based on Race and AgeGroup, which gender has a greater change in population over the year?

1.Re-state your question:

My revised question is 'Based on Race and Age Group, which gender has a greater change in population over the year?'. I choose this topic because this question can cover all the variables provided by dataset, including 'Age', 'Year', 'Gender', 'Sex', and 'Population', which can give a clear overview of the dataset. This question can be answered visually because 'ggplot2' and 'plotly' in R and Python can both visualize it with colorful labels and sufficient details. Instead of just looking the original dataset, we can use programming software to show the trend of gender population changing from 1900 to 2010. Also, this question can be answered visually in showing the patterns of population in different races with two colors. Last, but not least, it can be visualized in multiples plots based on Age Groups (from 1 to 16) to show a combination of both Race and Gender. Use visualizations, the information can be broken down into manageable categories and easier for readers to understand the relationship between variables. Besides, by visualizing, this question can be more attractive to readers.

2. Design a static visualization:



3.Provide a short narrative write-up:

For the question 'Based on Race and Age Group, which gender has a greater change in population over the year', I mainly use R and ggplot2 to do the visualization. Before I create visualization, I took a look for my part1's sketches to find a better inspiration. For part1, I discovered the changes of gender population over the years, and I sketched a line graph, a bar plot, and a box plot. For all these three plots, the most appropriate one is line graph which has

both sufficient details and shows a clear trend pattern. However, they all have advantage and disadvantages. To dig deep based on the 'part2-analytical-dataset', I then revised the question to get more relationship between variables and 'gender trend' in populations. Also, I start to think about to combine the advantages of line, bar, and box plots to create a better visualization which not only shows a clear pattern but also easier for readers to compare.

The first step I choose to visualize is the relationship between Year and population. This is because using these two quantitative variables can provides us with an overview of the trend from the year 1900 to 2010. For sketches in part 1, I found that although using 'Year' instead of 'Age' as x axis can be more helpful to compare the development of 'population' in a broader view of fluctuation. Instead of just getting rid of variable 'Age Group', I also start think about to generate multiple plots by different Age Groups. After find the feasibility, I use 'facet_wrap' function in ggplot2 to build, and found that it really helps in giving a stronger out frame of the visualization.

After deciding the out frame, I then start to thinks about the ways to visualize the relationship between the chosen variables. In part1, the sketches of line plot split variable 'sex' as male and female and using two colors to compare. This method works very well on showing different trends in one plot, I then adopt the greatest advantage of this, and use green line stands for 'Male' and red line stands for 'Female'. Moreover, adopting the advantage of box plot that can show the maximum and minimum value, I use scatter plot to visualize more information about the dataset. Also, have the experience of using too large scales of labels in part1, I set axis in 'million' for convenience.

Last but not least, I want to discover the relationships between multiple variables such as 'Sex' and 'Year', 'Race' and 'Year', 'Sex' and 'Age', as well as 'Race' and 'Age'. The most efficient way I found is that I can utilize the function of ggplot2 to set the color of scatter point by 'Race' (popular for 'white' and blue for 'nonwhite'). Based on these we can see the population change of gander based on different races. Also, it becomes convenient to compare the total change difference by each age group (from 1 to 16). The sketches created in part1 did help a lot on inspiration on this revised question and served as a very good base for digital visualization.