BIOS 640 Assessment Week 3

William Opoku-Nkoom

2025-09-21

Exercise 1: Reproducing and arranging ggplot2 figures

Fig.1: Distribution of age of respondents

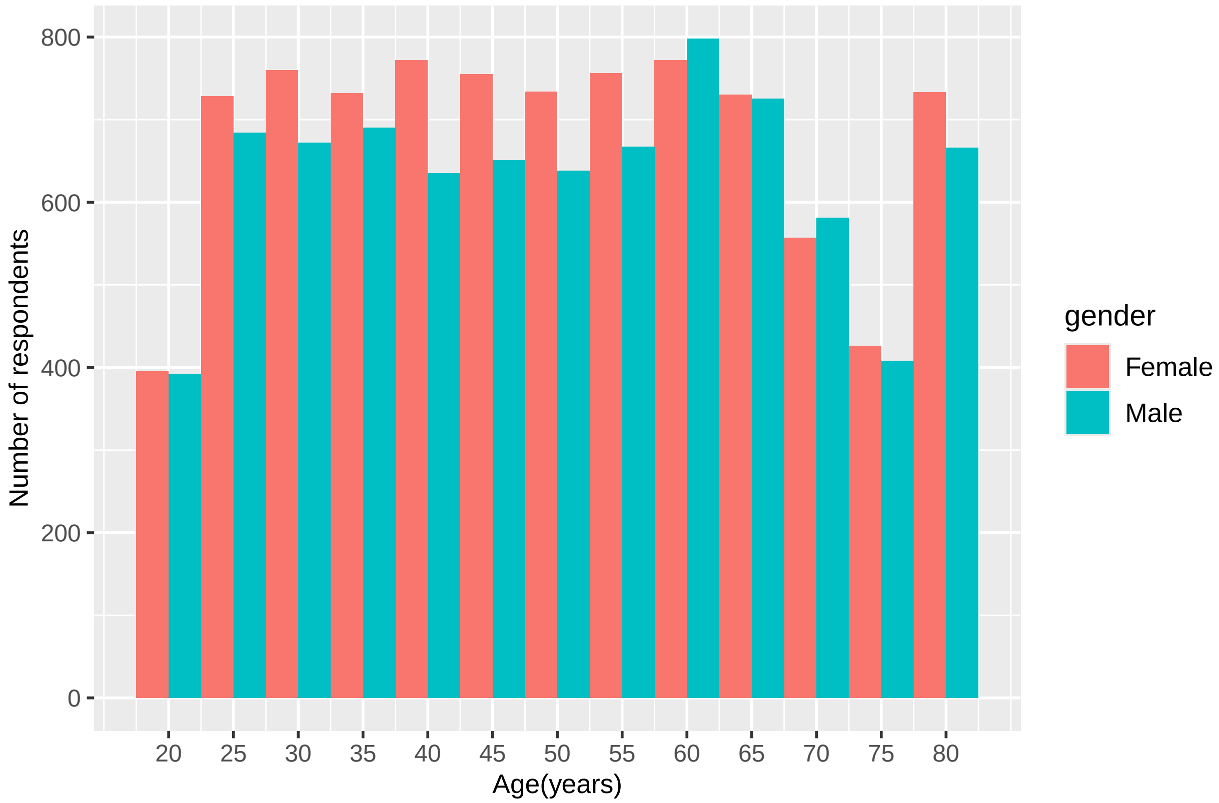


Fig. 2: Repartition of gender by ethnicity

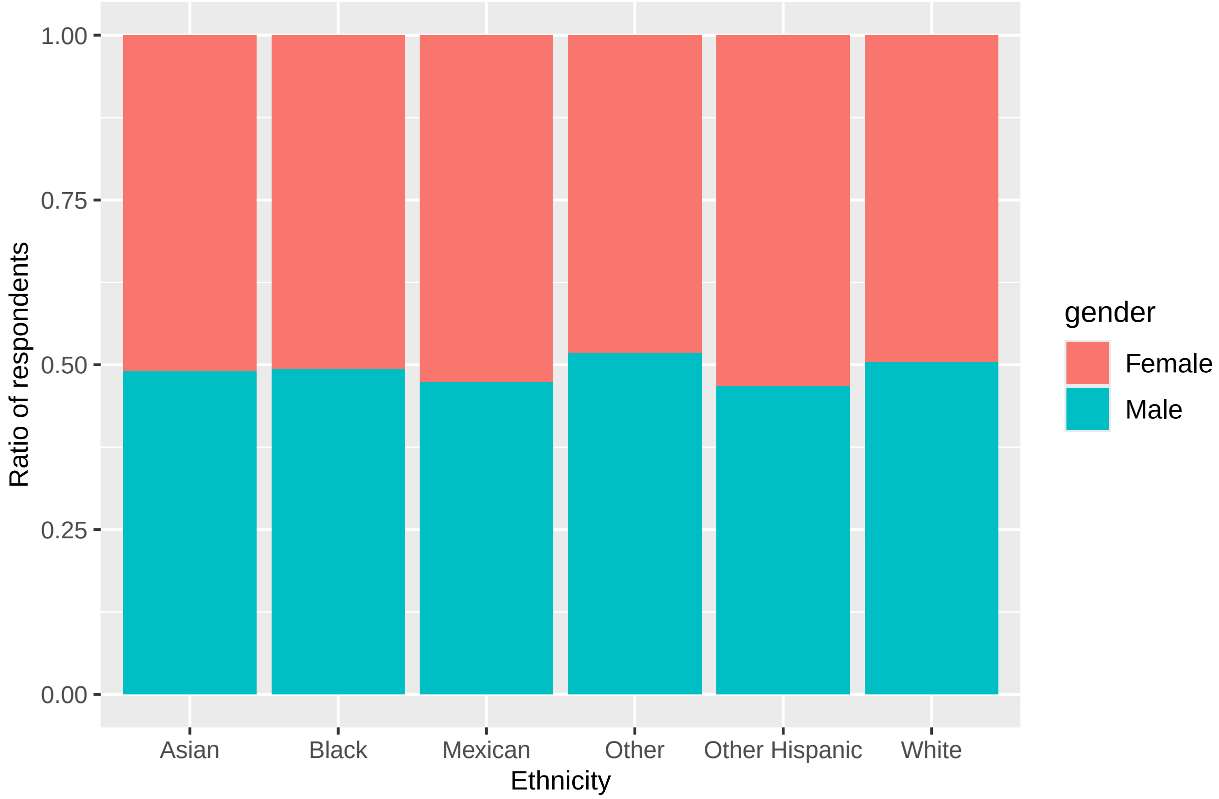


Fig.1: Distribution of age of respondents

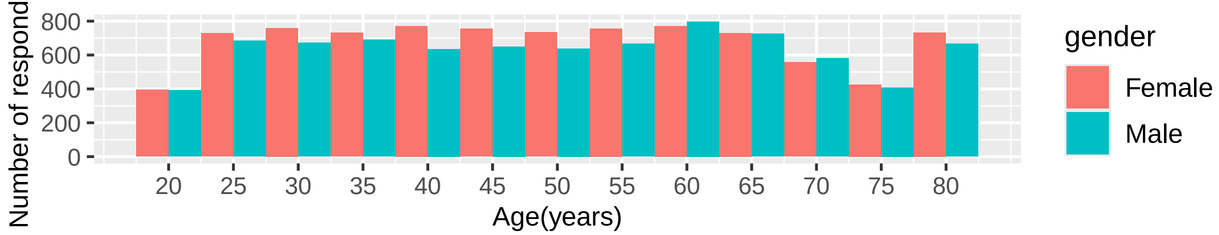
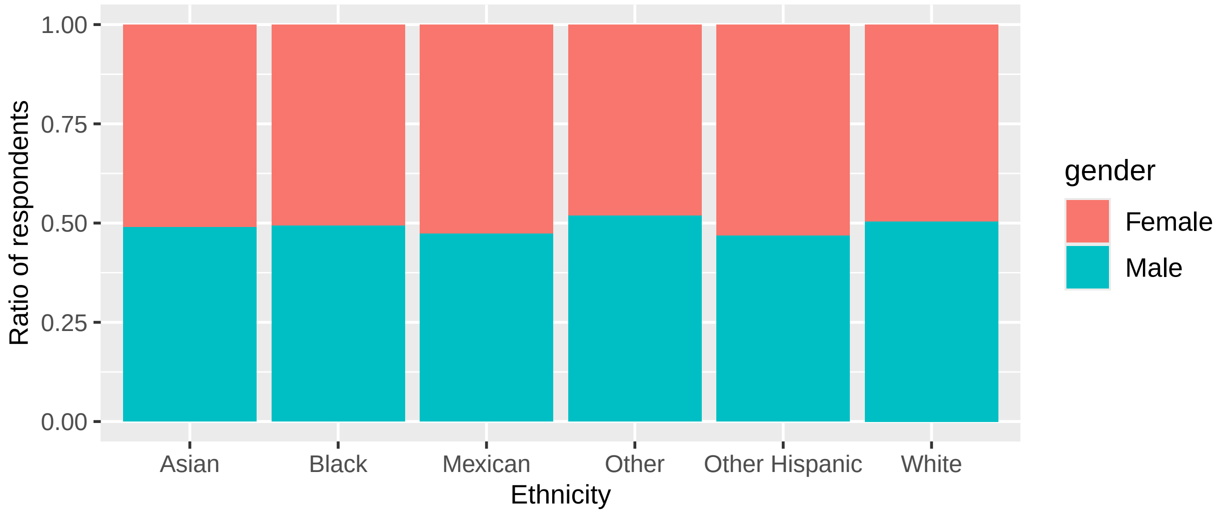


Fig. 2: Repartition of gender by ethnicity



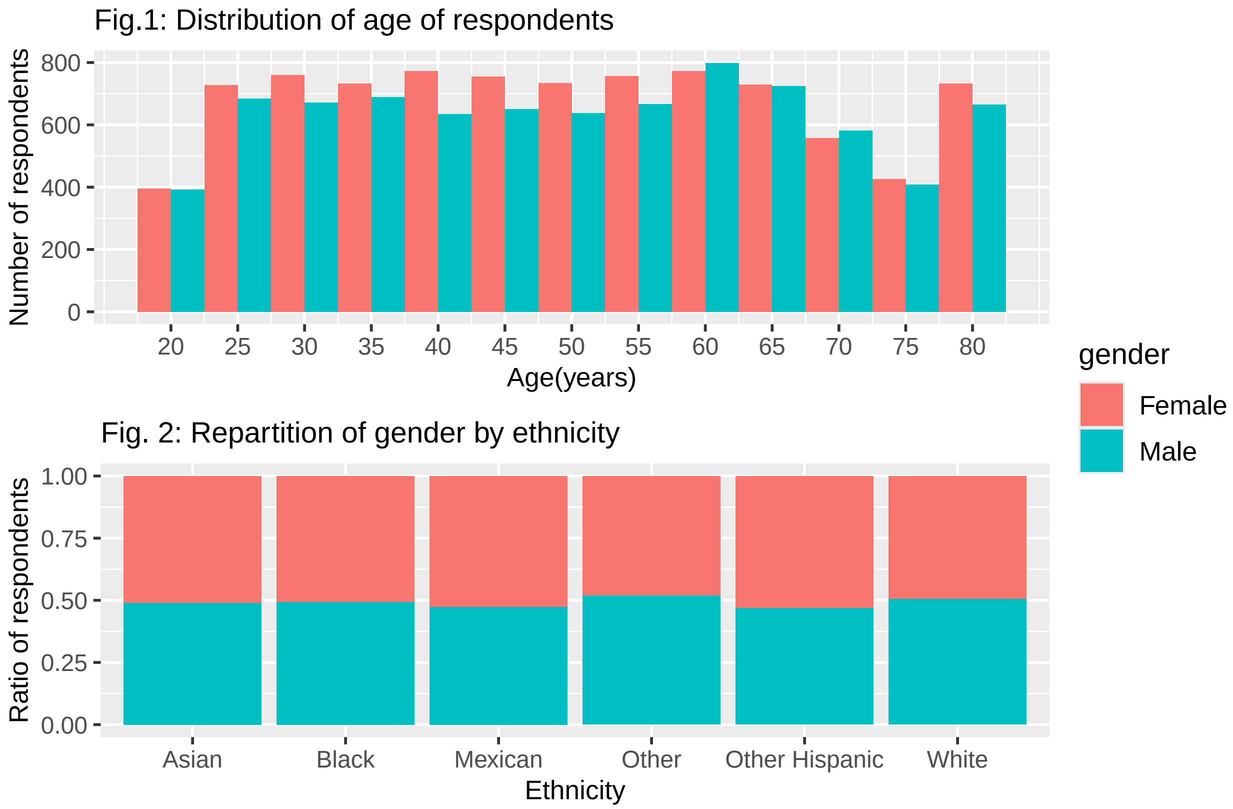


Fig.1: Distribution of age of respondents

Fig. 2: Repartition of gender by ethnicity

Fig. 4: Combined plot with ggarrange

Differences between the outputs from the two packages The layout of the combination produced from the ggpubr package (Fig.4), i.e. with the ggarrange() function is better than that produced from the cowplot package, i.e. with the plot_grid() function. This is because while the former permits you to retain only one legend in the output automatically, this argument is not permitted in the latter function, unless you extract them first and add a legend to the combined plot, which is a manual process and time consuming.

Moreover, the output from the former produces equal dimensions (i.e. same width for plot A and B and likewise the length). On the contrary, the dimensions of both plots are not the same from the output of the latter. The latter function shrinks the width of the first plot.

Exercise 2: Visualizing key characteristics

Fig. 5: Distribution of age by gender and ethnicity

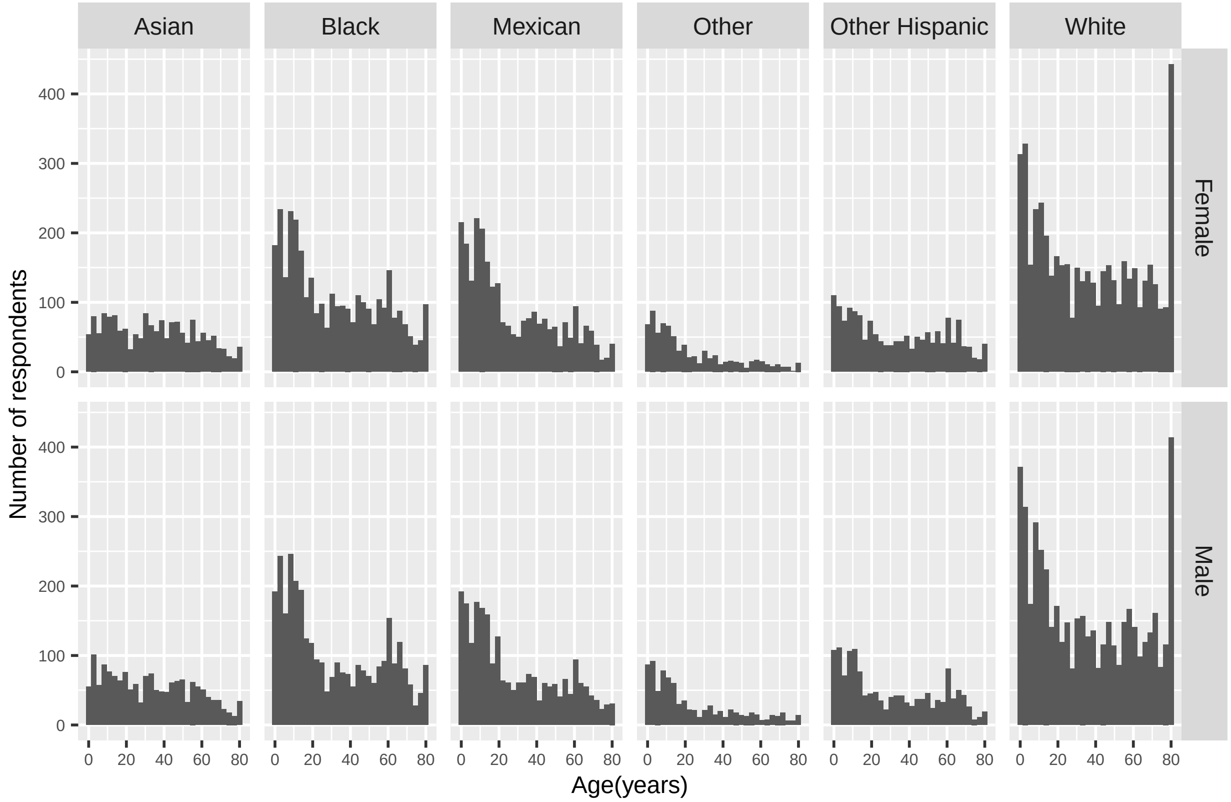


Fig. 6: Repartition of the population by gender

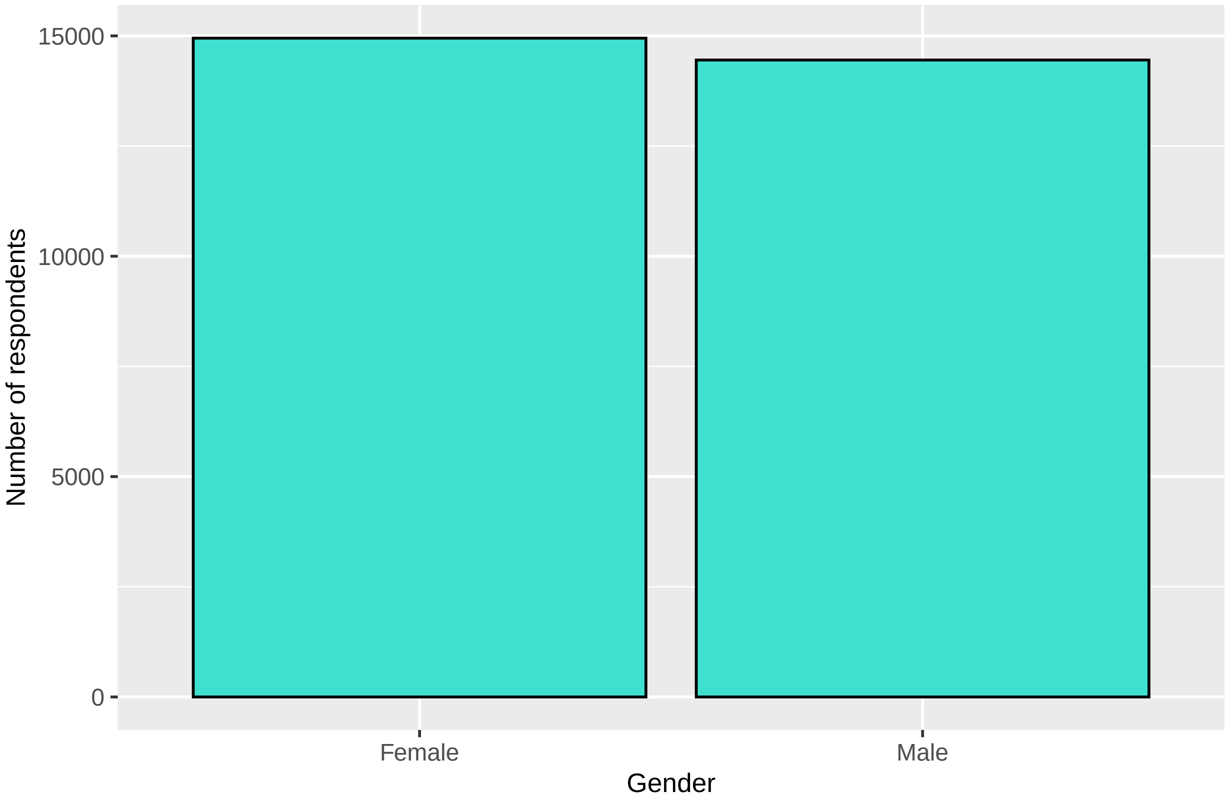


Fig. 7: Repartition of the population by ethnicity_1

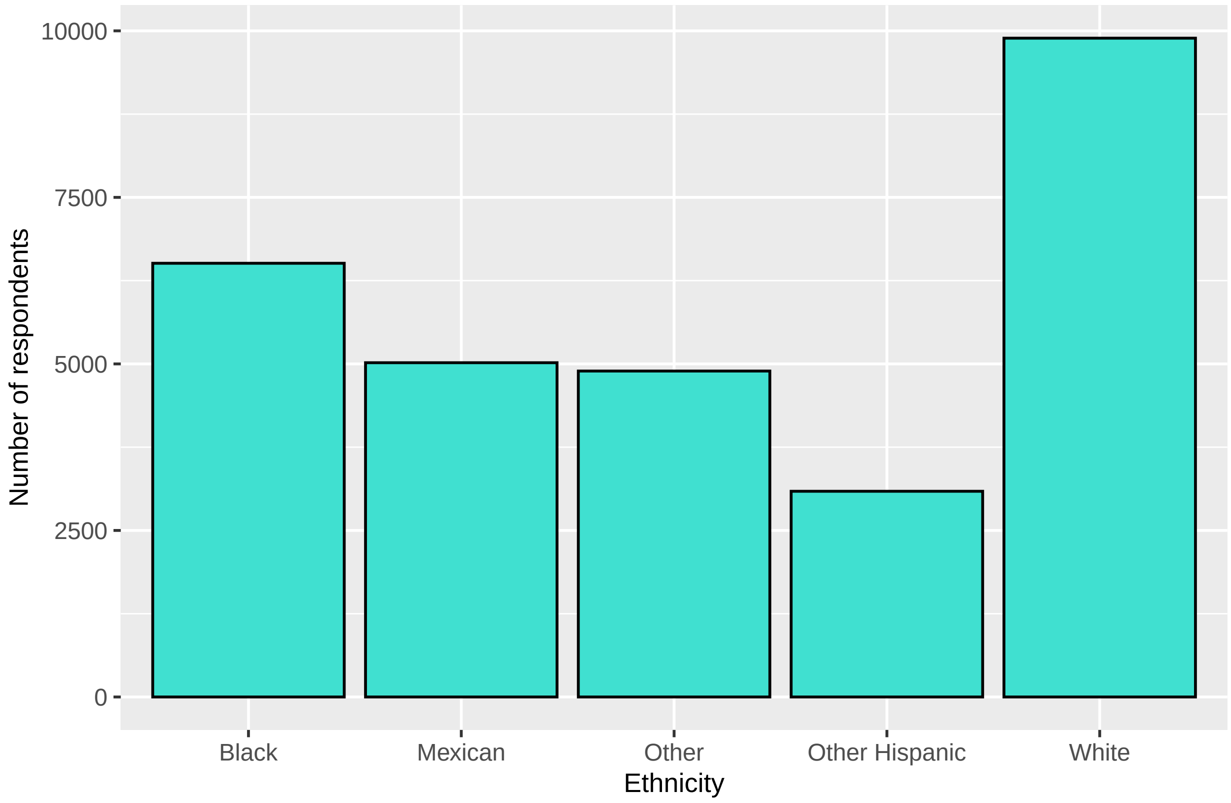
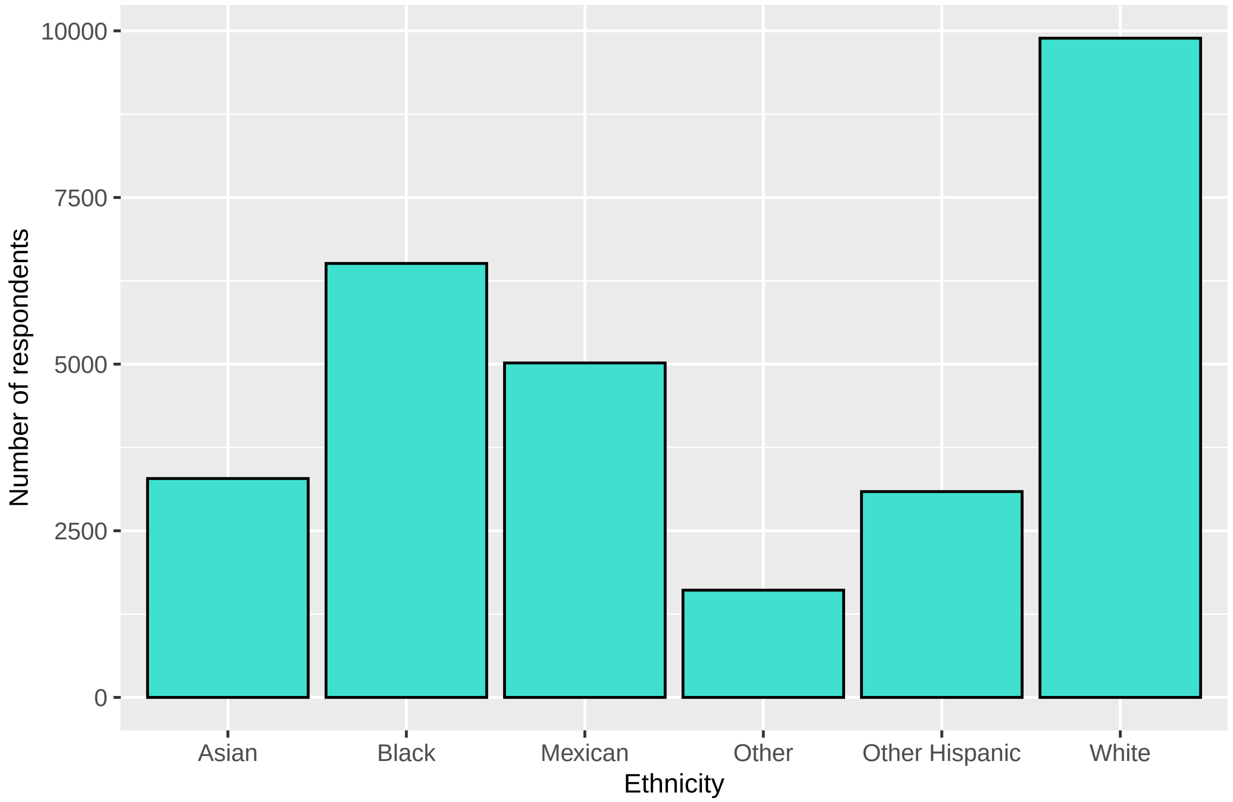


Fig. 8: Repartition of the population by ethnicity_2



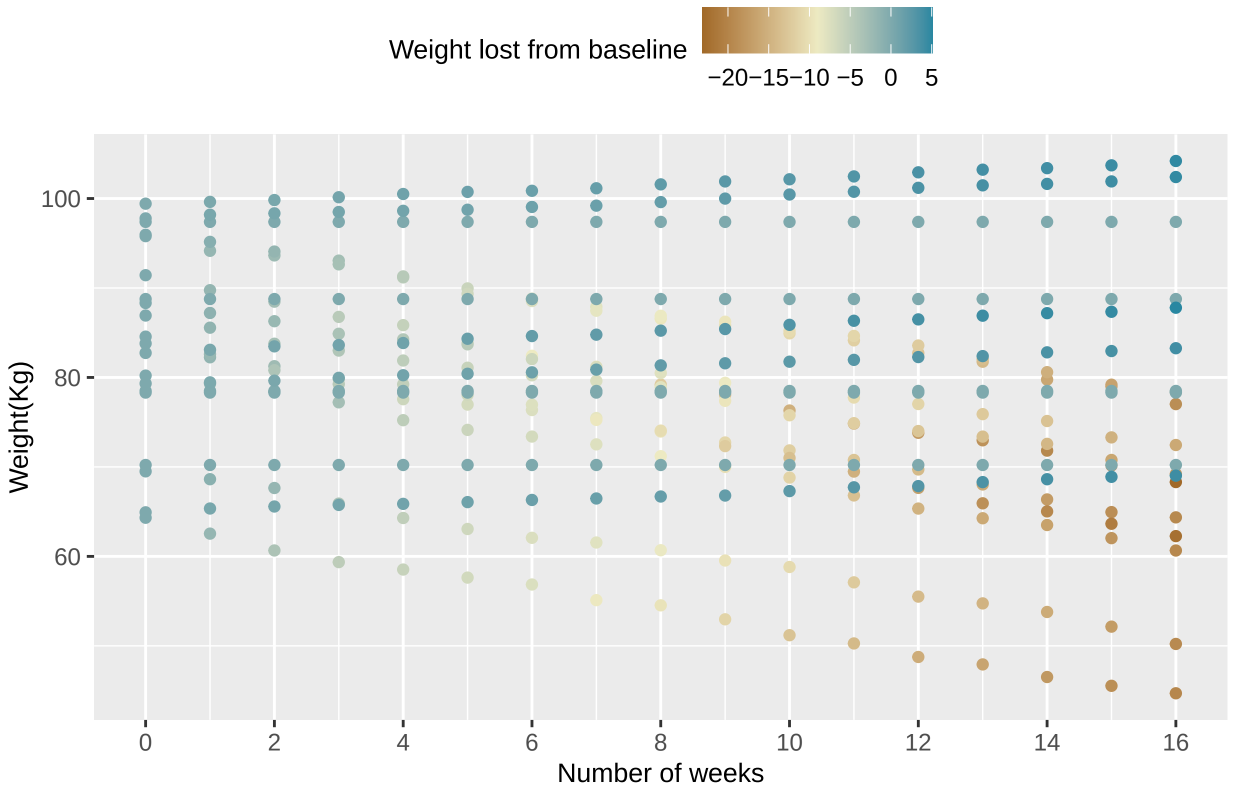
Description of the distributions In general the distribution of age (Fig. 5) is multimodal. This is typically evident among the Whites who dominate in the study population.The multimodal distribution is equally evident among Blacks and Mexicans. In these three population sub-groups, the peaks of the distribution can be observed roughly among respondents aged less or equal 20 years, between 30-40 years, between 50-60 years and 60-70 years. This is very clear among the Whites. The age distributions among Blacks, Mexicans and Other ethnicity, including Other Hispanics could be described as skewed to the right. In other words the number of respondents decreases with age in these study population sub-groups. The distribution of age among the Asians appear to be normal. There is basically no distinction between males and females with respect to their distribution by age.

There are almost equal number of males and females in the study population, even though the females slightly outnumber the males (Fig. 6). As indicated earlier, the Whites dominate, followed by the Blacks, then the Mexicans, then the Others. It is evident that most of the ethnic groups indicated as “Others” are actually Asians (Fig. 8). This distinction is crucial as the Asian population sub-group even outnumber those described as “Other Hispanics”. For this reason, I would say that the second grouping by ethnicity, i.e. ethnicity_2 (Fig. 8) is better than that of the first (Fig. 7).

There are no missing values of these variables, i.e. age, gender, ethnicity_1 and ethnicity_2, in the data set.

Exercise 3: Improving ggplot figure

Fig. 9: Scatter plot with diverging color scales representing weights lost from baseline



The original plot does not demonstrate storytelling visualization characteristics for the following reasons:

1. Creating a legend for individual observations is not appropriate. Legends are usually created for groups for more appealing visual storytelling.
2. Joining the individual points with lines in this context makes it difficult to make any meaning out of it, especially as the lines cross one another
3. The plot does not give any information about the reference point, i.e. the baseline, with respect to other points. This is the purpose of the visualization of this study yet it is missing.
4. No title, no unit for the weight axis, making it difficult for the Figure to stand alone as an independent plot.
5. There is no prudent use of color, as most of the lines even have the same color. As pointed out in the first point the use of color on individual observations is not where the emphasis should be. It is more prudent to use just one color for all observations and distinguish what is critical in this research with a different color, which are the deviated weights (weight loss).

To improve upon the visual storytelling capabilities of this plot, I would go for a plot with diverging color scales which represent the deviation of the weights from the reference point (the baseline, i.e. weight at week 0), throughout the period of the study. The emphasis here is on deviated weights, which is weight loss, colored in the form of sequential color scales to indicate the magnitude of the weights lost by the participants at every measurement.

From this improved plot (Fig. 9), it is evident that the intensity of the brown color which indicates higher magnitude of weight loss, increases as the number of weeks increases for several participants, somewhat depicting the efficacy of the diet in reducing weight. This improved plot has more visual storytelling characteristics compared to the original plot.

Exercise 4: Exploring relationships through visualizations

Fig. 10: Repartition of the population by gender

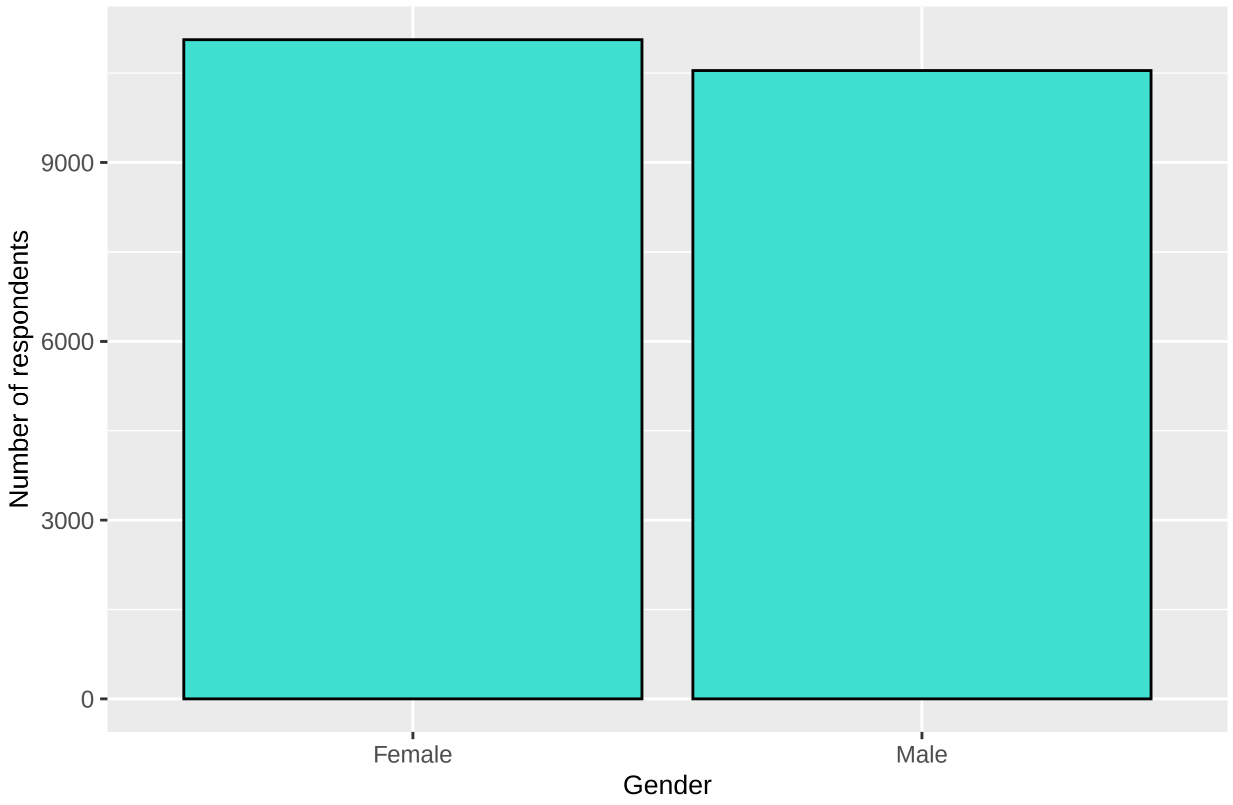


Fig. 11: Repartition of the population by blood presure groups

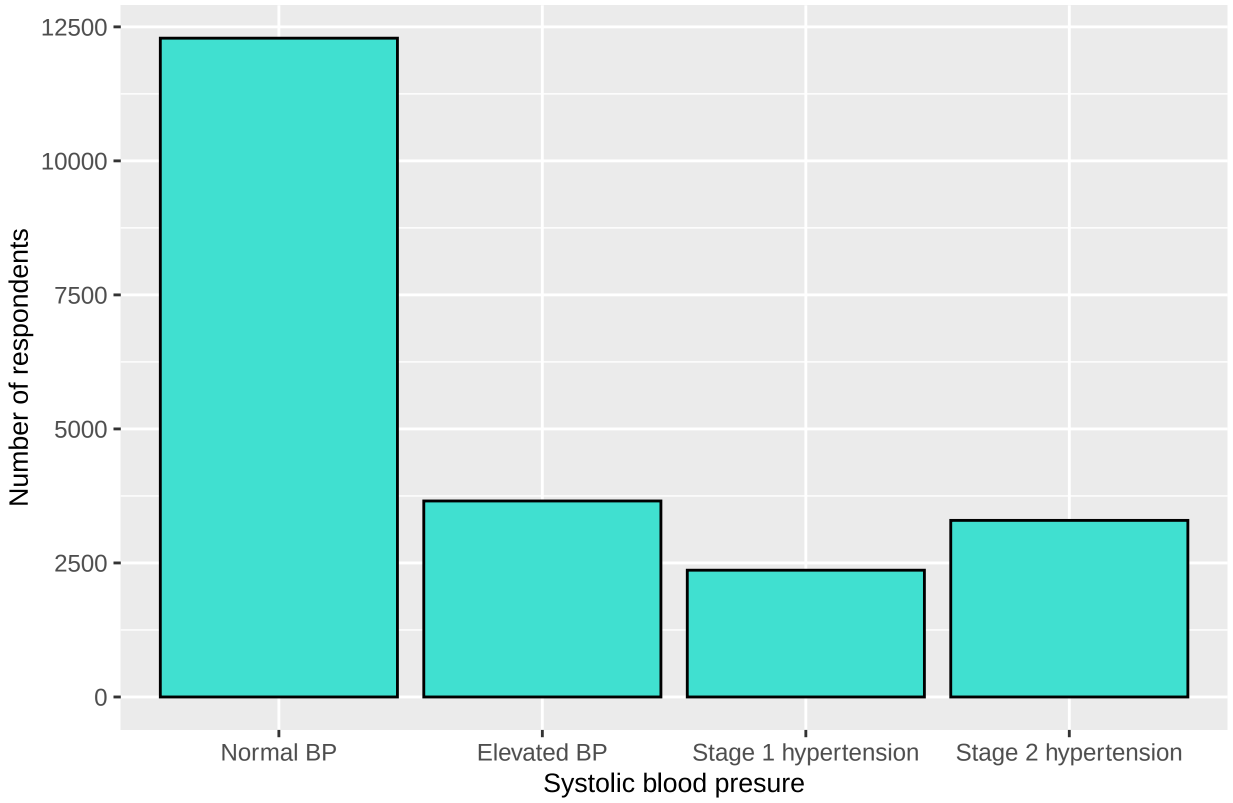


Fig. 12: Distribution of systolic blood pressure of respondents

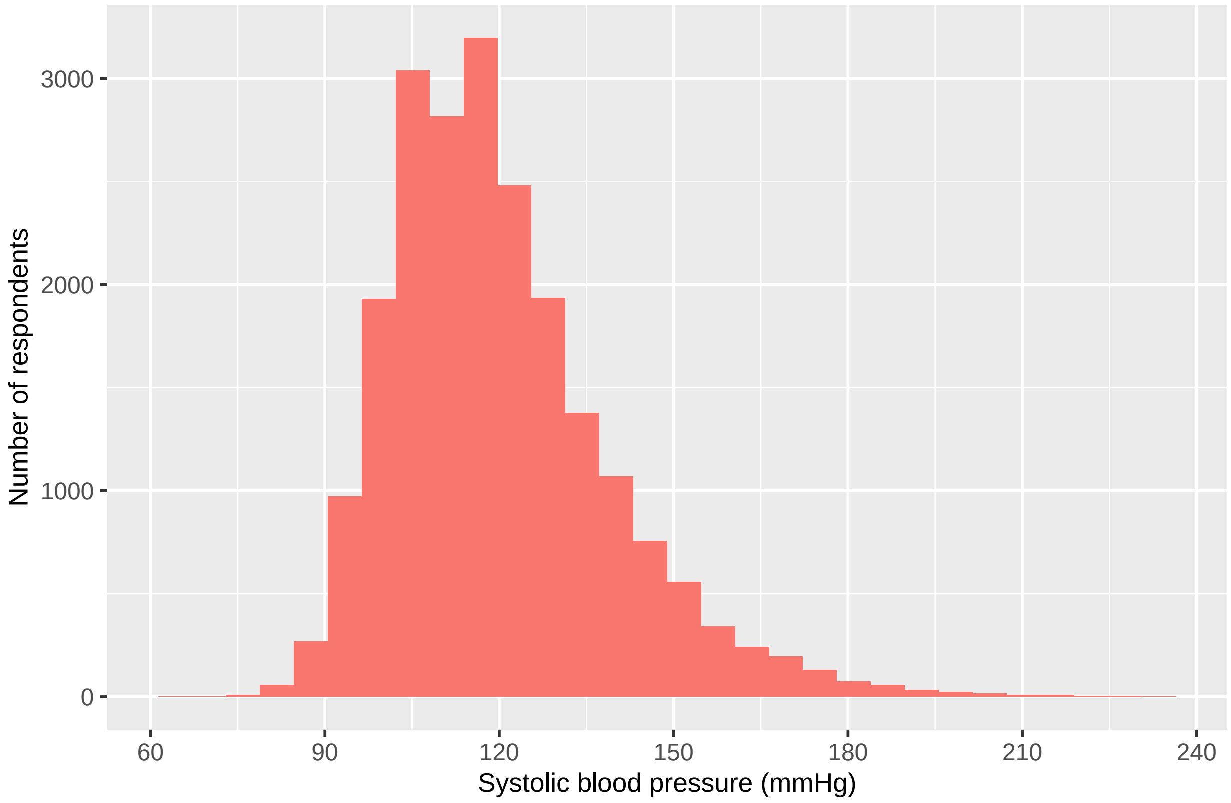


Fig. 13: Boxplot of systolic blood pressure

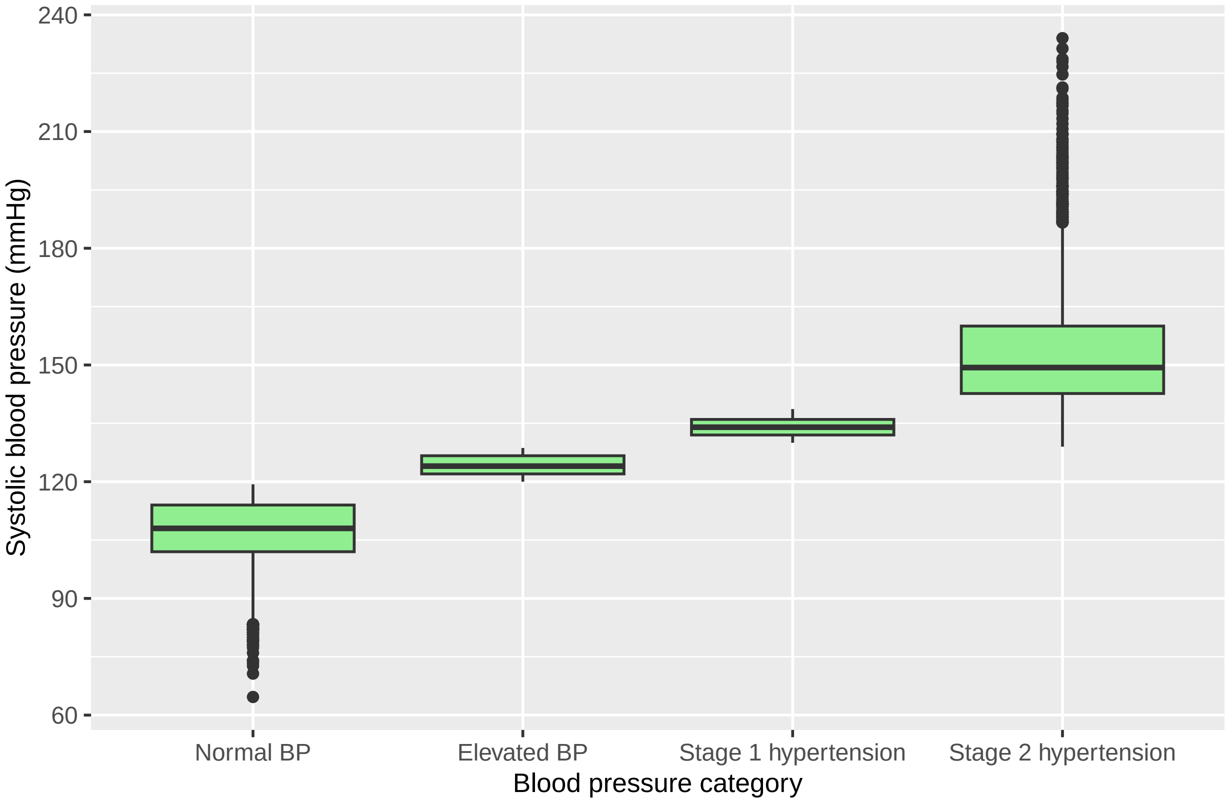
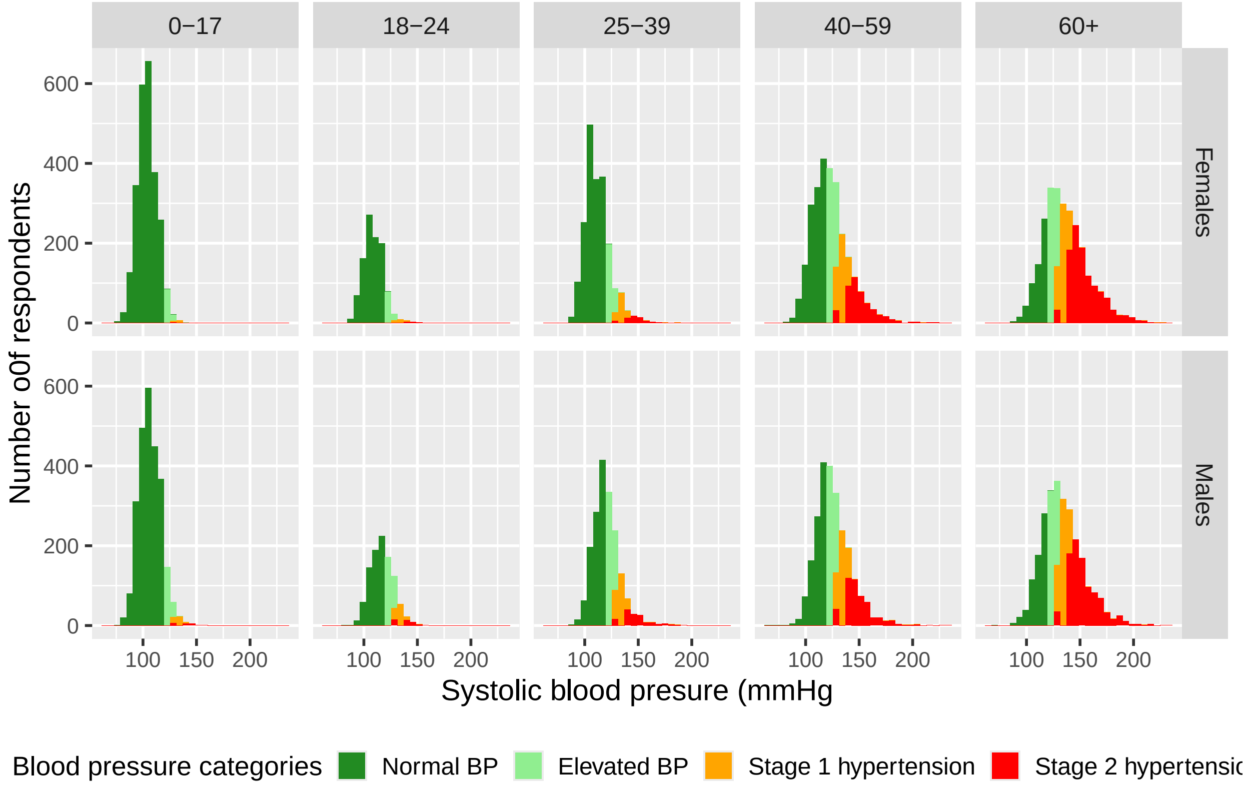


Fig. 14: Changes in hypertension across age and gender groups



Report The sample size reduced from 29,400 to 21,604 after filtering out missing values in the age, gender and average systolic blood pressure variables. A common distribution in both data sets (old and new), which is the distribution of gender indicates no marked difference (Fig. 6 and Fig. 10). The distribution of gender in both data sets is similar, meaning the data did not really change after filtering and that, comparable inferences could be made from both data sets.This commonality or the persistence of the distribution in gender could be attributed to the large sample size, as the sample size still remained significantly high after applying the filtering arguments.

The distribution of blood pressure in the study population is right skewed from the histogram plot (Fig. 12). This is more evident when stratified by blood pressure categories in the natural order as shown in the bar plot (Fig.11).From the plots there are more individuals with normal blood pressure but relatively fewer number of individuals described as hypertensives, that is with high blood pressure readings.

To analyse these categories further, a boxplot (Fig. 13) was used to view the distribution of the central tendencies or some summary statistics. The plot shows the 25th percentle, the median and the 75th percentile of each blood pressure group. It further gives information about the nature of the distribution of each category and the existence of extreme values or outliers for each. From this plot it is evident that the individuals with normal blood pressure show a left skewed distribution while the stage 2 hypertensives show right skewed distribution and more outliers. The elevated and and the stage 1 blood pressure groups both appear to show normal distribution (Fig. 13).

The plot of the hypertension categories by gender and sex (Fig. 14) clearly shows the vulnerability of hypertension in older people. The risk of hypertension increases with age, with significantly very high proportions in people aged 60 years and above and quiet high in ages 40 to 59 years. The risk appears to remain the same by gender across age groups, but it is slightly higher in males compared to females, especially from 18 year and above. At the latter ages, however, females appear to be at higher risk compared to males.