Week 3 Assessment

Gopika Balasubramanian

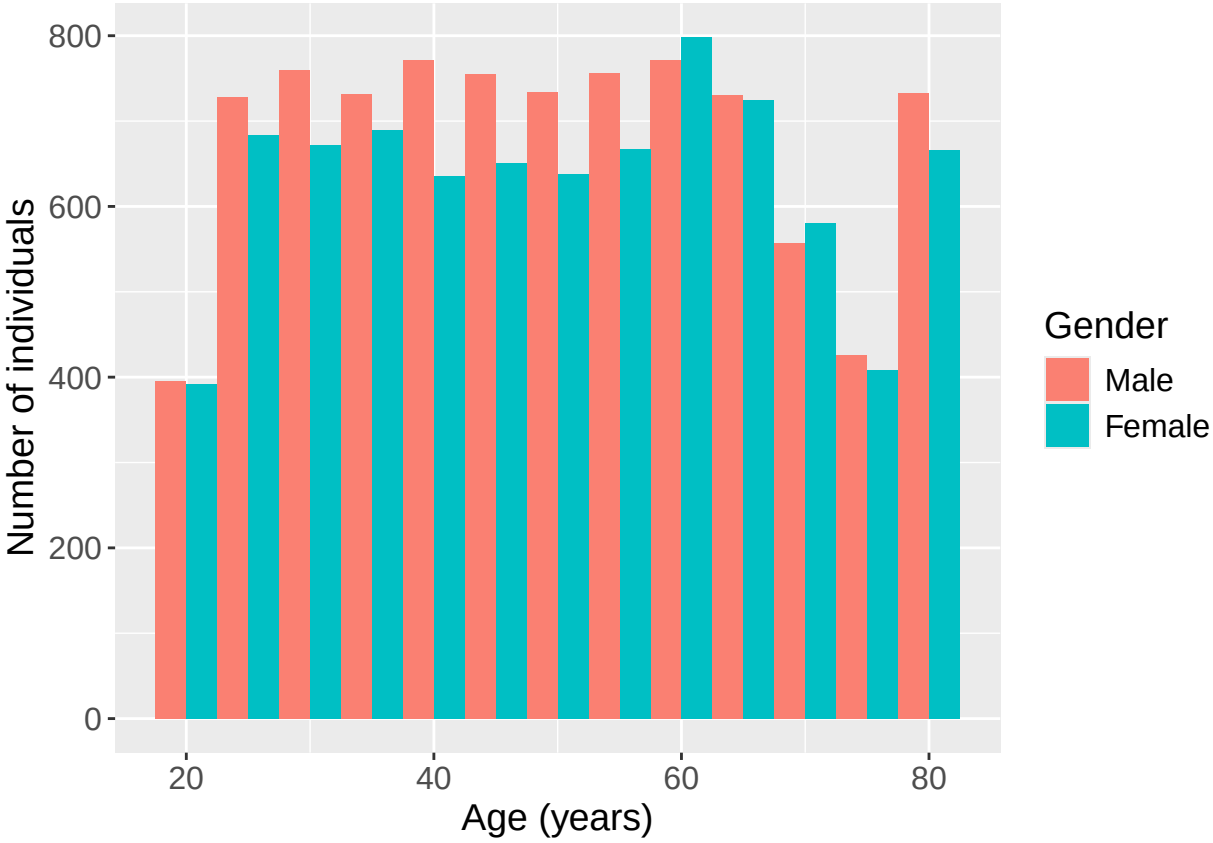
2025-09-21

Exercise 1

Reproducing and arranging ggplot2 figures

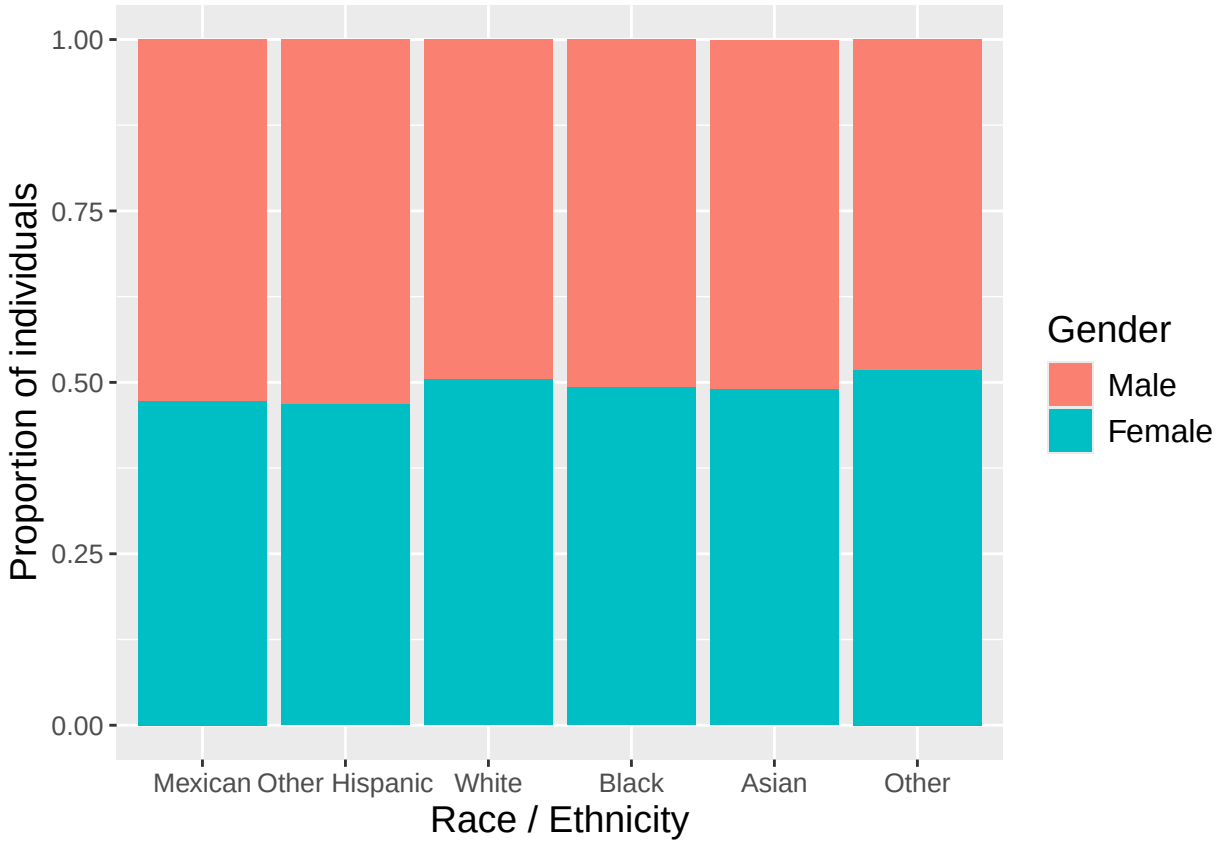
Plot 1

As a first step, the required dataset 'cleaned_NHANES.csv' was imported. The first plot provided shows the age distribution stratified by gender (RIAGENDR). In the NHANES raw data, as per the data dictionary for RIAGENDR variable, Male is coded as 1 and Female is coded as 2. To retain the same colour coding in the reproduced plot, the fill colours were assigned manually. Further, as the provided plot only included age distribution from 20 years onwards, age was filtered to 20 or above in the dataset before plotting. Binwidth was set to 5 to ensure each bin covered 5 years, axes and category labels were appropriately added.



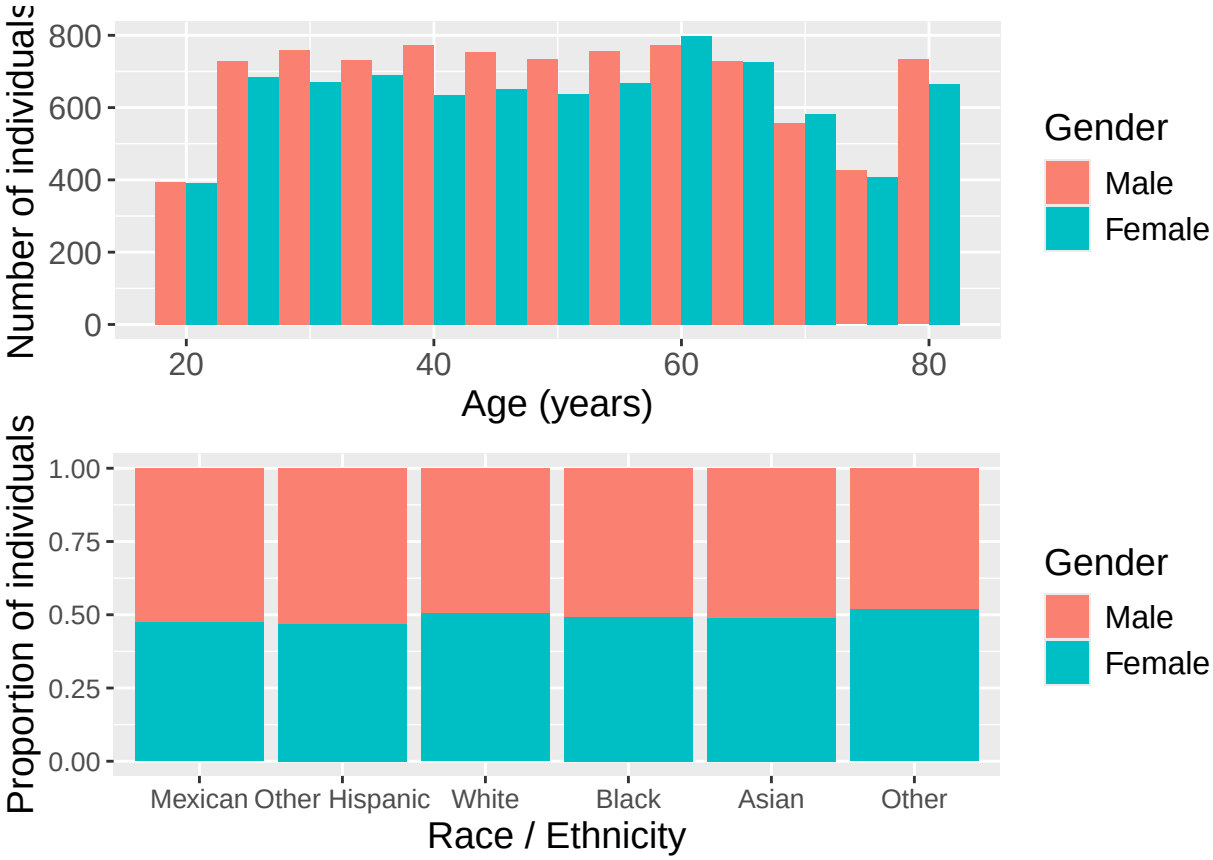
Plot 2

Before the modification of the second plot, the unique values in ethnicity_1 and ethnicity_2 variables were compared to identify which corresponded to RIDRETH3 in the raw data. To match the order of categories of the given plot, the categories were re-levelled. Position = 'fill' was provided within the geom_bar() input to ensure proportion was plotted.



Using cowplot to combine both plots

The plot_grid() function from cowplot package was used to combine the graphs.

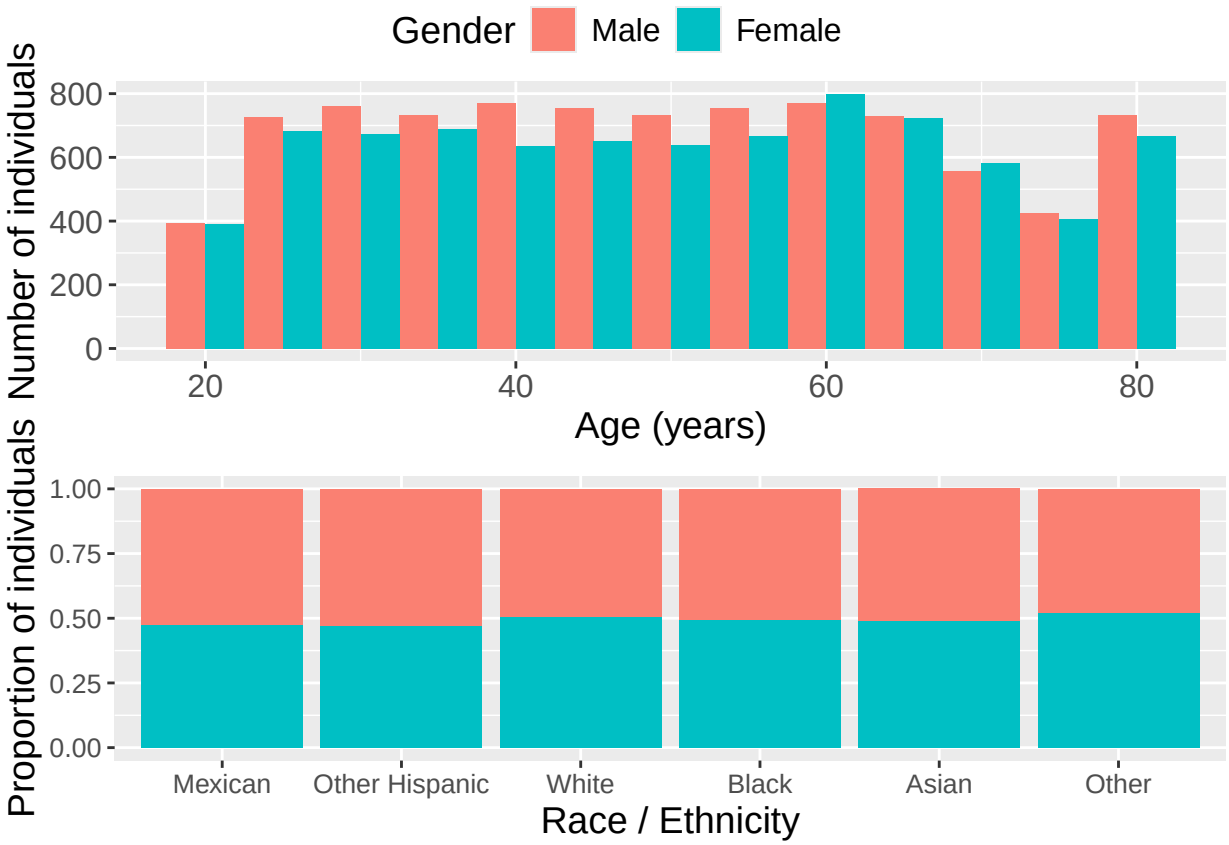


Using ggarrange() to combine both plots

Next, the same two plots were combined using ggarrange().

Comparing the plots combined via two methods

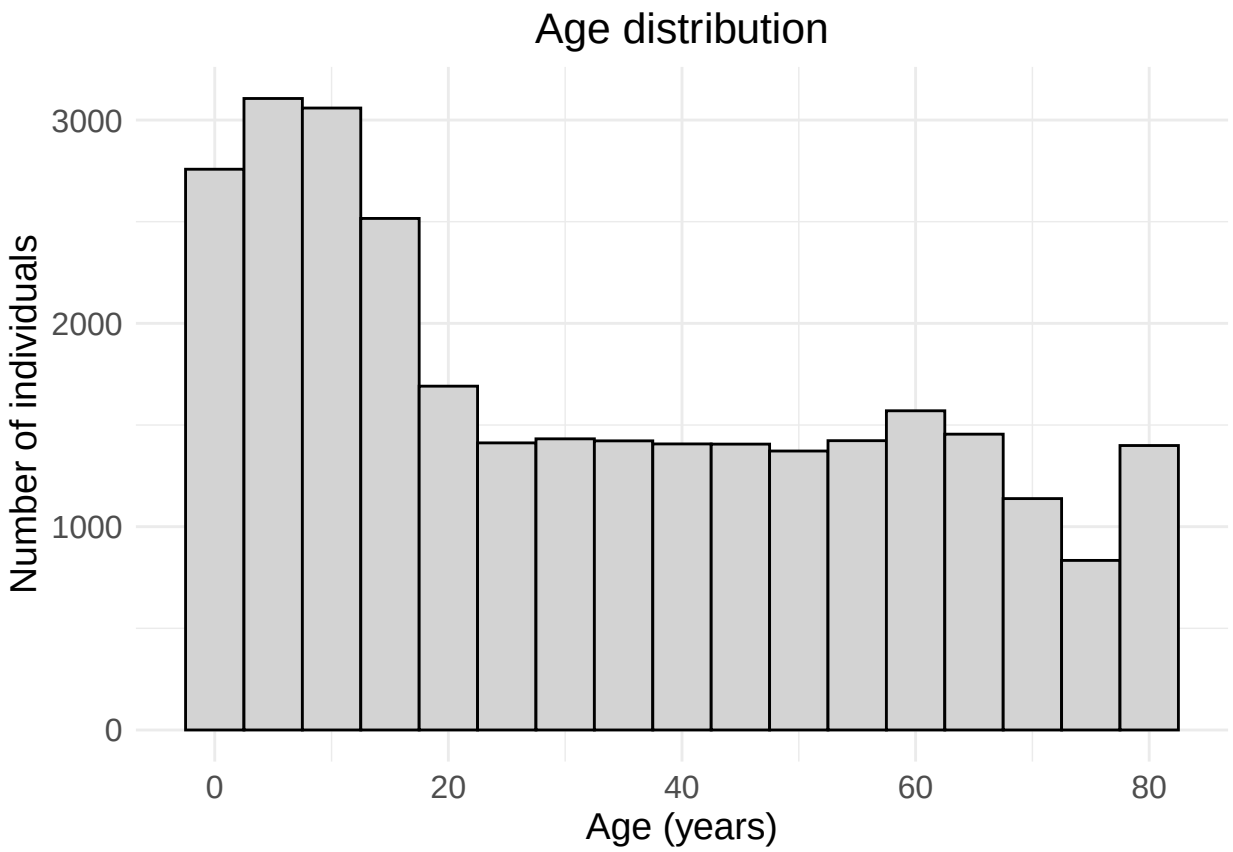
On comparing the two methods for combining plots, the layout provided by ggarrange looks better since it saves space and avoids redundancy by retaining the legend only once. Thus, for cases like this where the legends across the plots to be combined is the same, the ggrrange() function offers an efficient way to avoid legend repition by setting common.legend = TRUE.



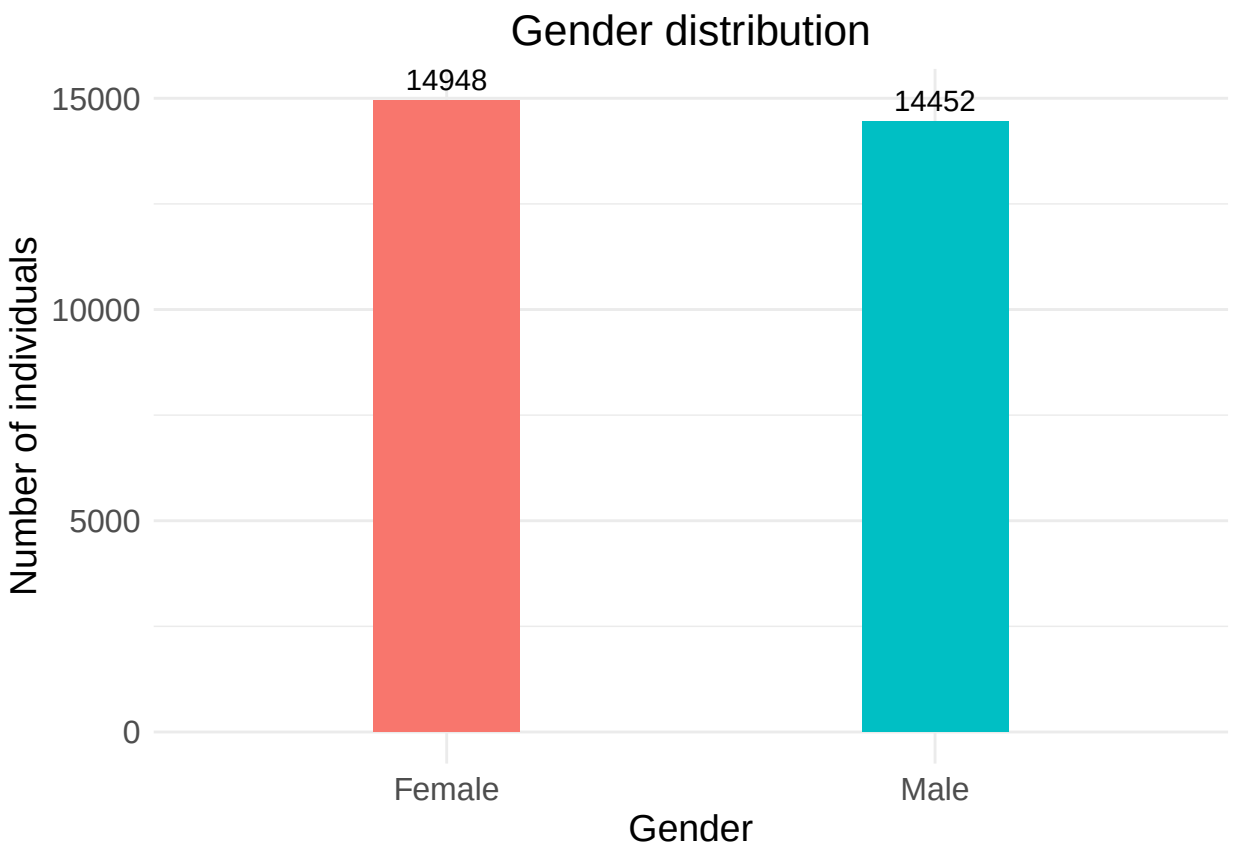
Exercise 2

Visualizing key characteristics

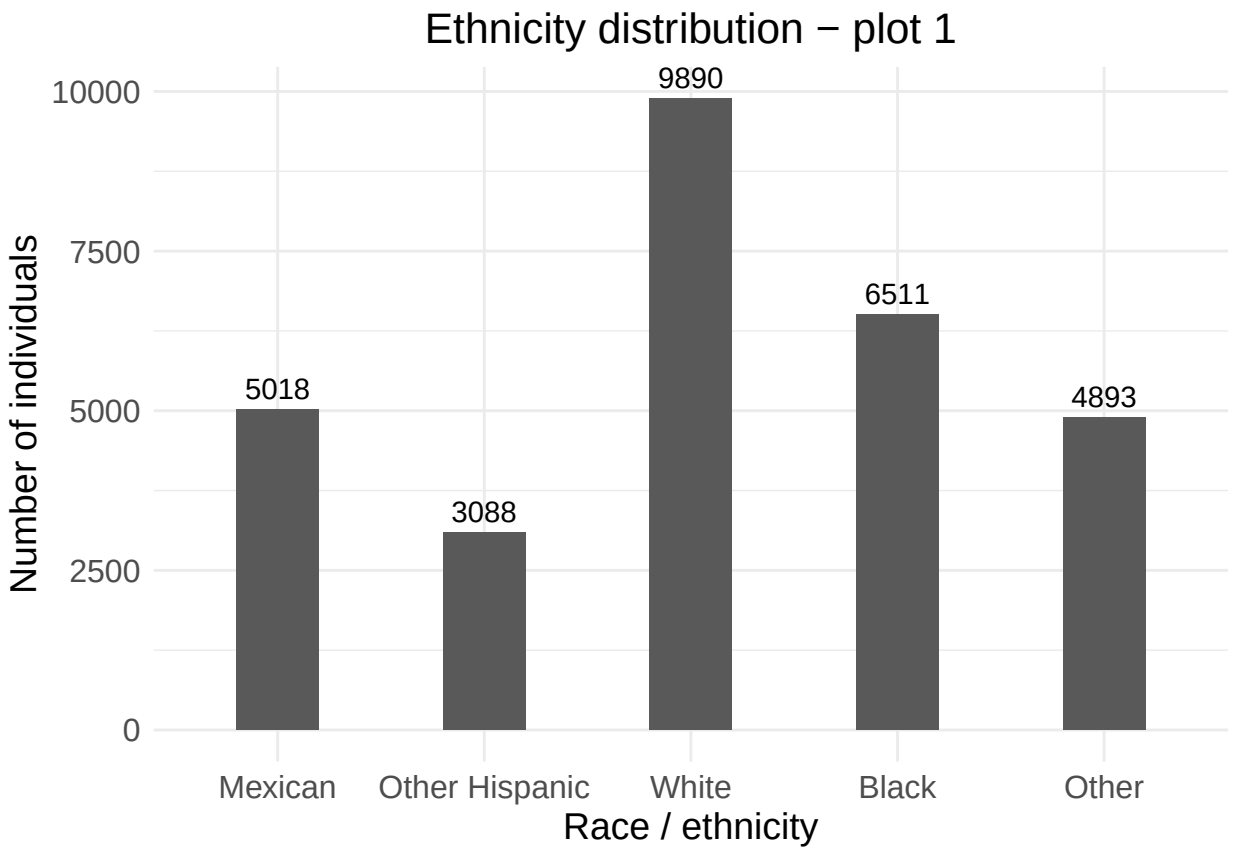
Age: Peak of the distribution occurs within the younger age groups, particularly around 10-15 years, indicating more individuals belong to these age groups. There is a sharp decline by 25 years old, after which the distribution is more or less even or consistent across age groups until 60 years old. After 60 years, there seems to be a declining trend, with fewer individuals comprising older age groups. Interestingly, the counts near 80 years seems to be against this trend with an increase - however, it must be noted that they way the data is coded - all above 80 years were also bunched up in this category. Therefore, the increase at 80 years might just be an artefact. There were no missing values in this column as checked using sum(is.na()).



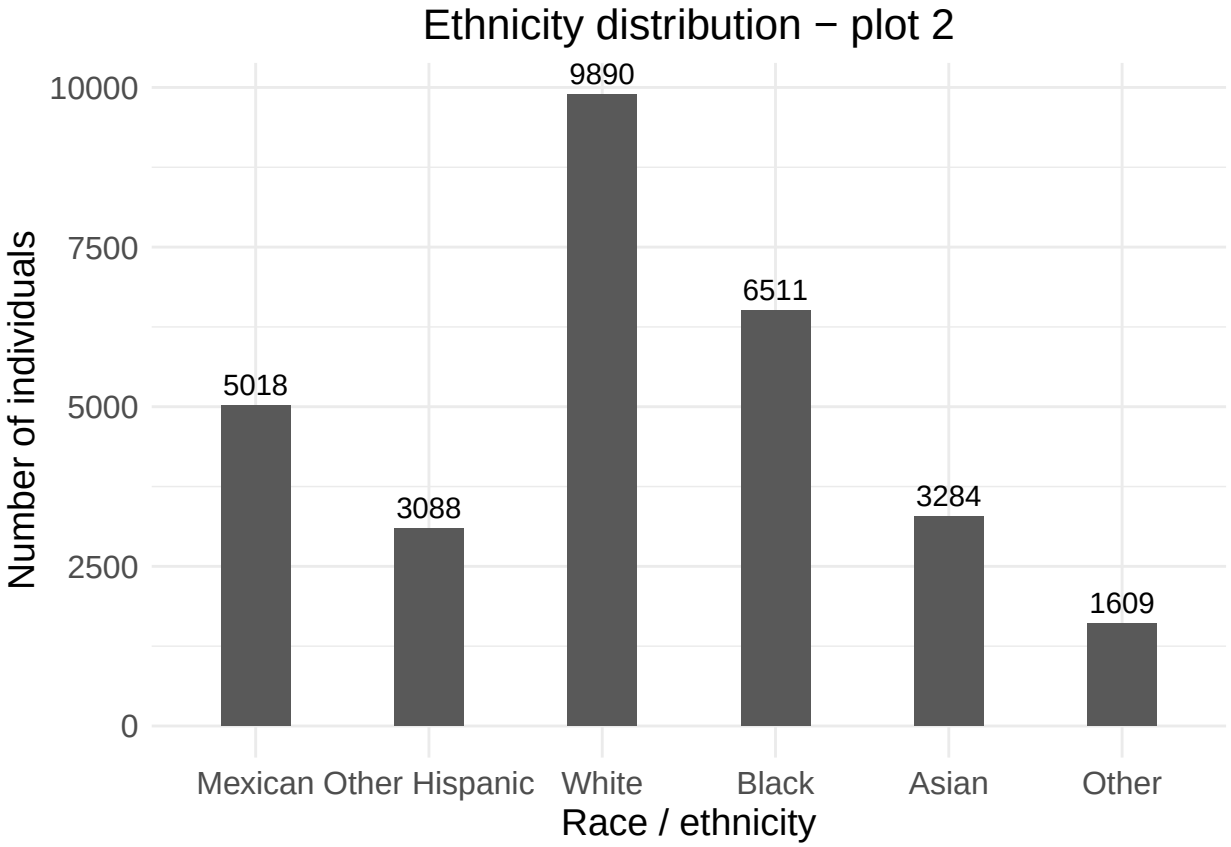
Gender: There are 14948 males and 14452 females in the dataset, as displayed in the bar plot. There were no missing values in gender variable.



Ethnicity_1: According to the distribution of ethnicities (by ethnicity_1), persons identifying as White comprised the majority, followed by Black and Mexican. Other races also comprised a substantial number of individuals, greater than those in Other Hispanic category alone. There was no missing data in this column.



Ethnicity_2: According to the distribution of ethnicities (by ethnicity_2 which included Asian as a category), persons identifying as White still comprised the majority, followed by Black, Mexican, Asian, Other Hispanic and Other races. There was no missing data in this column. It is observed that 'Asian' was a subset of the 'Other' category as coded in ethnicity_1. There were no missing values in this column.



All plots were saved using ggsave() function. It was decided to keep the ethnicity_2 variable over ethnicity_1 as it provided more granular data. If needed 'Asian' could later be categorized as 'Other' in future, but getting information on 'Asian' from 'Other' would be less feasible.

Exercise 3

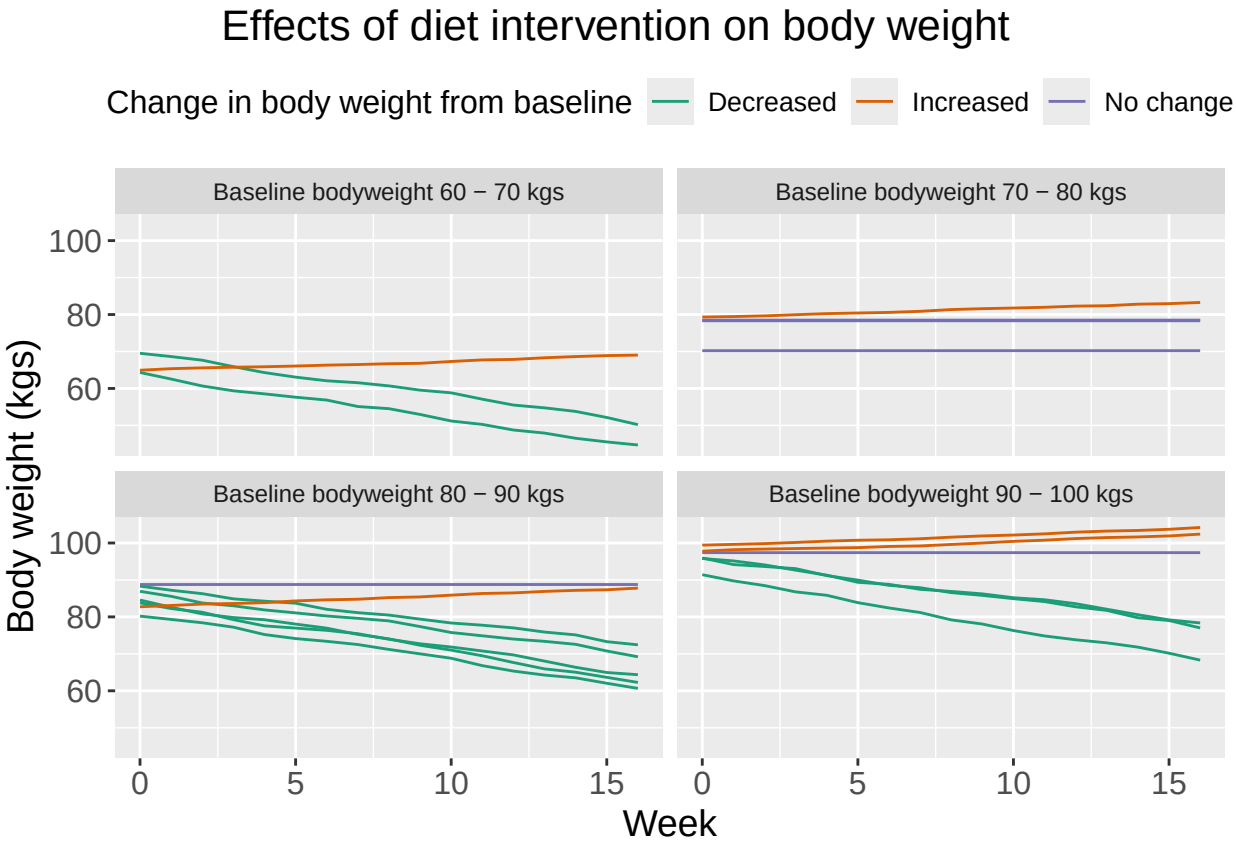
Improving ggplot figure

The graph produced by the colleague seems a bit cluttered with many criss-crossing lines and is hard to follow in terms of whether the participants' body weights are increasing or decreasing or unchanged. While an attempt has been made to show the change in weight data for every participant, it is hard to take away the key message from the current plot. While x and y axis seem mostly okay, and need minor adjustments (the units of body weight could be added, however this is not provided in the dataset either, but is likely in kilograms).

To improve the plot, the idea was to facet the current plot (weight over weeks) by participant baseline weight categories and further to highlight the data by whether the body weight for the participant increased, decreased, or remained unchanged from Week 0 to Week 16. This was done as the goal of the plotting project was to see how the diet intervention impacted the change in body weight over period of time compared to baseline. Further, faceting would help make the data less crowded overall and while also providing an opportunity to check if the patterns of weight change have any relationship with the baseline weight categories.

The revised plot makes it clearer to understand at the outset that not all patients experienced a decrease in body weight. There were at least few patients for whom the body weight increased or remained unchanged over the course of time at week 16 compared to week 0. Further, more patients in this experiment started

with a baseline weight of 80 to 90 kgs and of them most experienced a decrease in body weight by week 16 (5 of 7 participants). In 60 - 70 kgs baseline weight category, 2 of 3 participants experienced a decline, while none of the 4 participants with 70 - 80 kgs baseline weight experienced decline. More sample sizes in these categories are needed to investigate whether diet intervention would produce a considerable effect in these patients. Interestingly, for those participants with higher end baseline weight category (90 - 100 kg), there was a 50-50 pattern of diet decreasing vs. remaining unchanged / increasing. Therefore, while overall around 50% of the participants have experienced a decline in body weight by week 16, further study could be done to evaluate if these patterns are affected by body weight at baseline.



Exercise 4

Exploring relationships through visualizations

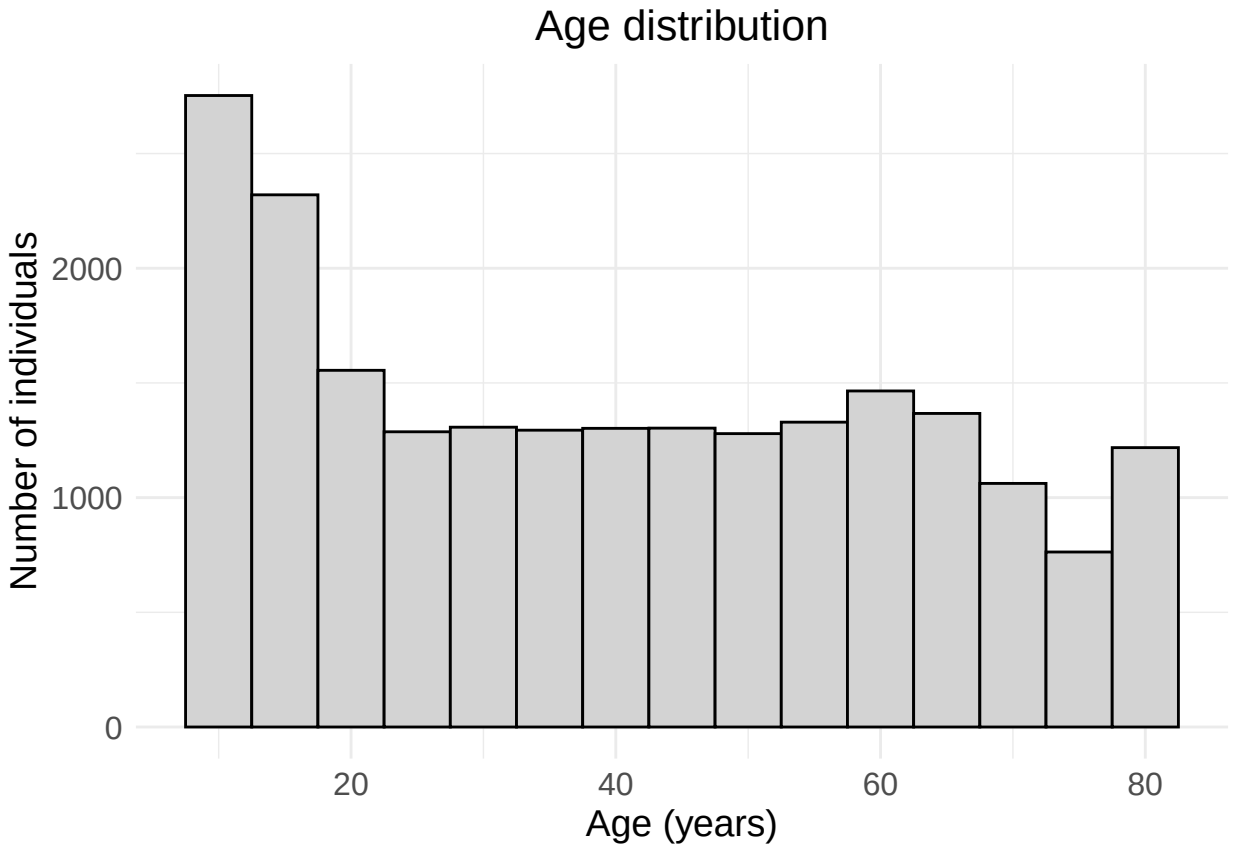
The objective of the present report was to evaluate the distribution of individuals across hypertension categories (as indicated and characterised by systolic blood pressure measurements) within the NHANES data set. Further, the prevalence of hypertension across demographic variables like age and gender were also investigated.

Only those participants with at least one systolic blood pressure measurement data (out of four possible) were considered as part of the analysis. Further, to ensure full correspondence of blood pressure data and demographic data, individuals with missing age or gender data were not considered.

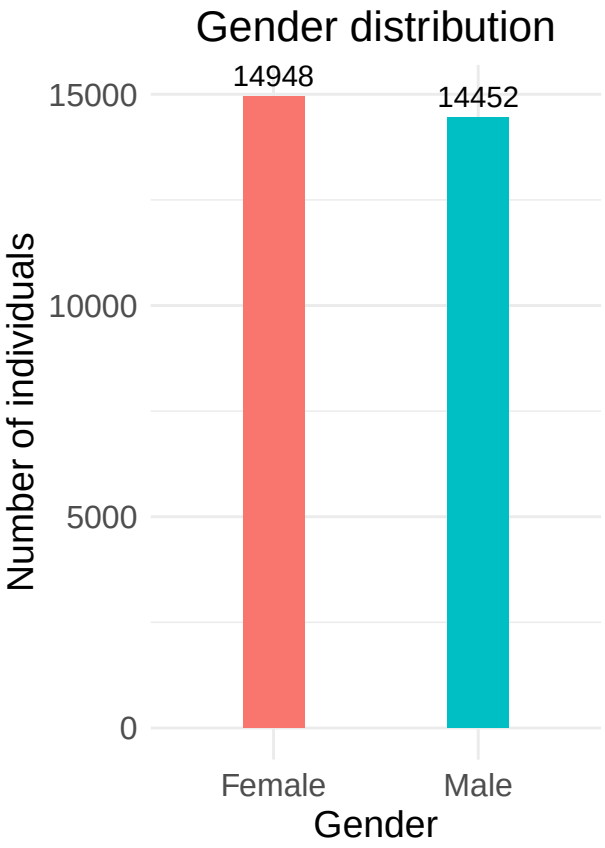
An average of all available systolic blood pressure measurements was computed per individual, which was used as the main blood pressure measure to categorize patients across different hypertension categories as per indicated clinical cut-offs. The categories were: Normal (SBP $<$ 120 mm Hg), Elevated (120 - 129 mm Hg), Stage 1 hypertension (130 - 139 mm Hg), and Stage 2 hypertension ($>=$ 140 mmHg).

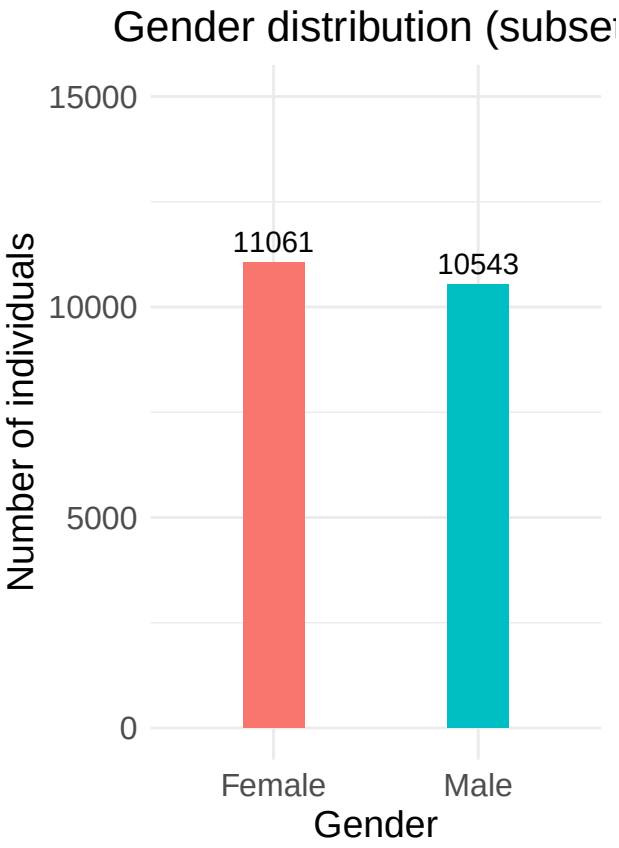
The filtered and modified datasets contains 21,604 unique observations corresponding to individuals with non-missing average systolic blood pressure measurement and age, gender data. The reduction from 29,400 observations in the original data set indicates that 7,796 rows were discarded due to missing age, gender, or average systolic blood pressure measurements. Few characteristics of the sample data set were investigated via visualizations.

Age: age distribution of individuals in the modified data set was similar to the full data set, however general decrease in counts across all ages and particularly among younger (< 15 year olds) was noted.

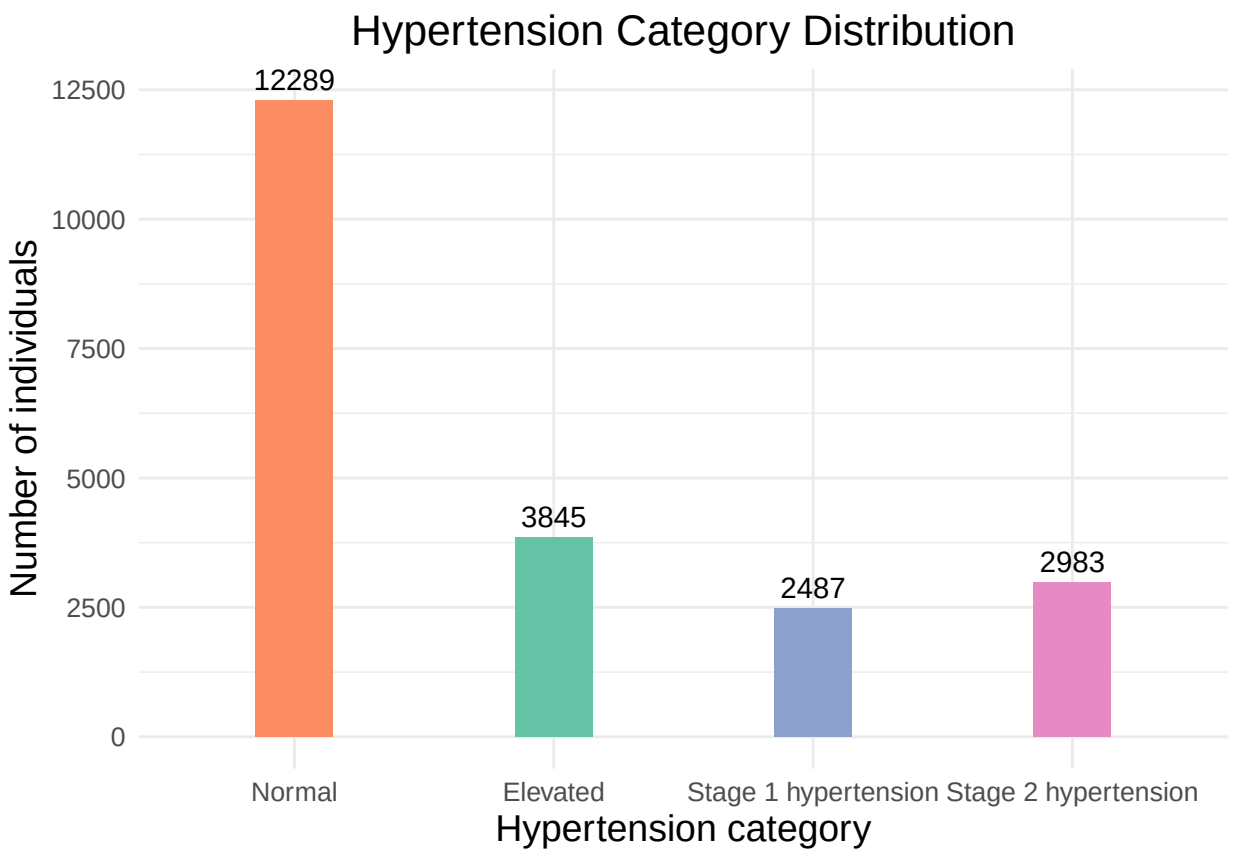


Gender: The distribution of males and females were similar in full dataset and the present subset data, with slightly higher number of females than males in both. The removal of rows impacted both genders similarly.

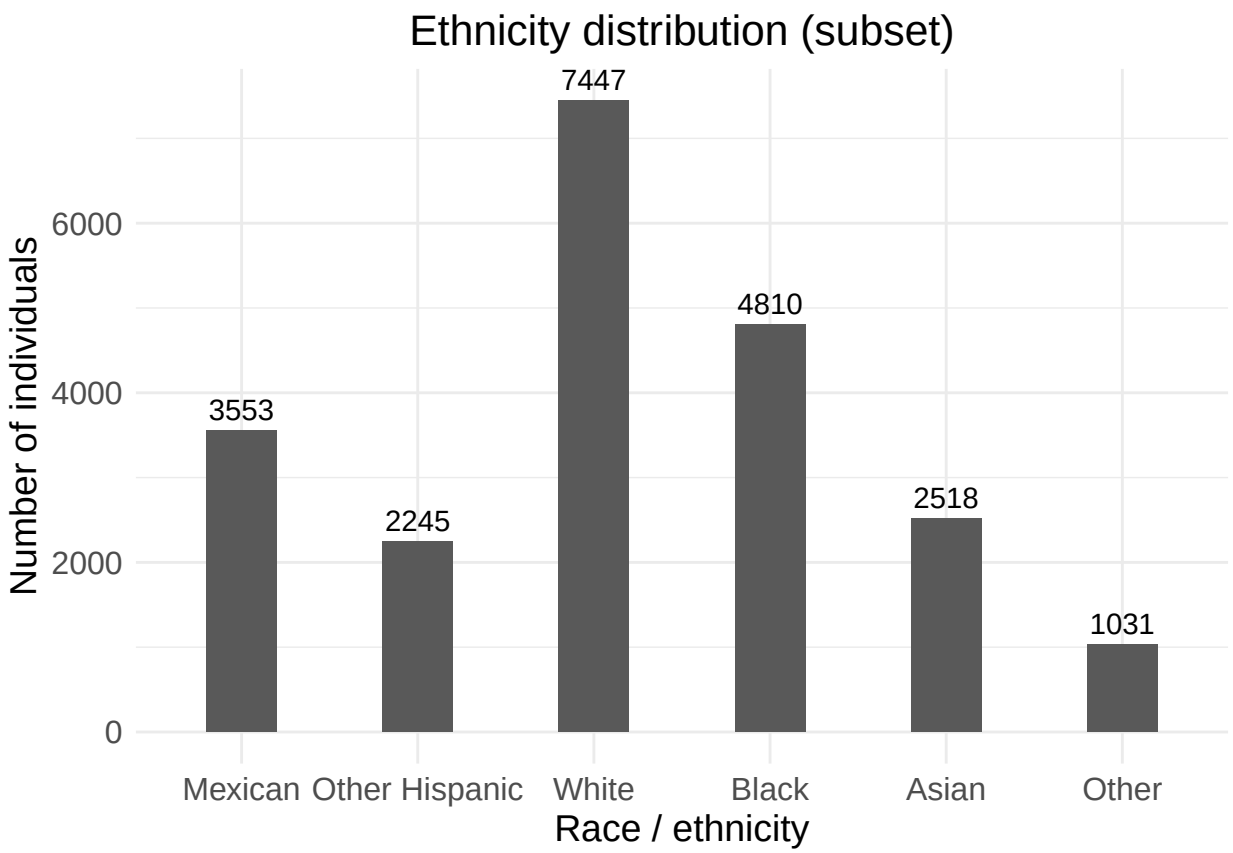




Hypertension categories: Most individuals had normal average SBP, followed elevated SBP. Slightly more individuals had Stage 2 hypertension compared to Stage 1 hypertension.

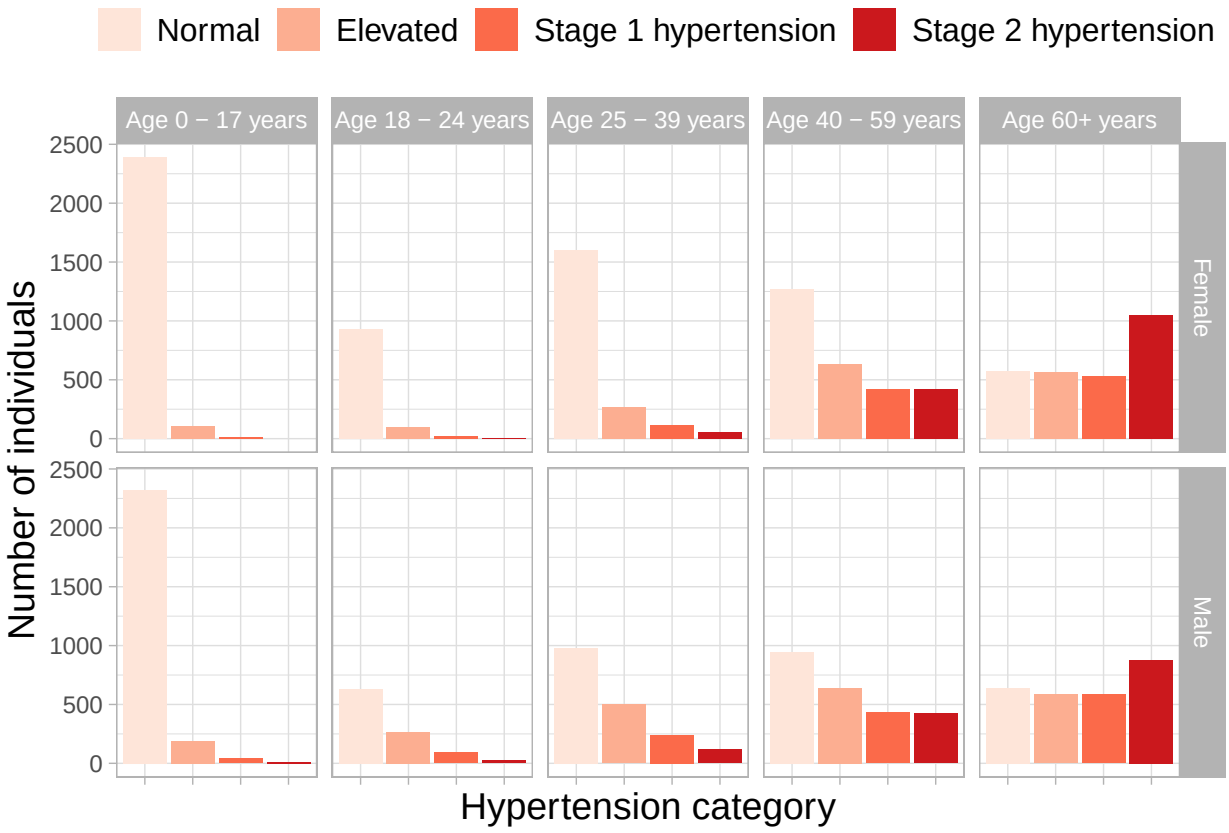


Ethnicity: ethnicity distribution was similar to the full dataset. White comprised the majority, followed by Black, Mexican, Asian, Other Hispanic, and Other categories.



The pattern of prevalence of hypertension stages across age and gender categories were investigated via visualization. Most individuals in young age groups (< 24 years old), irrespective of gender, typically had normal average SBP. In these age groups (0 - 17 and 18 - 24 years), elevated SBP or stage 1 or stage 2 hypertension were less prevalent. The shift in pattern was noticeable for the 25 - 40 years age group with more individuals having elevated SBP and stage 1 or 2 hypertension (elevated SBP was higher than stage 1 or 2). The progressive shift in pattern continued into the next age group where prevalence of non-normal SBP categories became more pronounced. However, for both these age categories - participants with normal SBP were still the majority. In the oldest age group, of 60 years or older individuals, the prevalence of stage 2 hypertension was very pronounced and comprised the majority while those with normal SBP, elevated SBP or stage 1 hypertension were similarly distributed.

The age related differences noticed were consistent for both males and females. In terms of differences in patters by gender, one noticeable finding was that difference between number of individuals with normal SBP and non-normal SBP tended to be sharper for females compared to males for those in 18 - 24, 25 - 39, and 40 - 59 year groups (difference was less for males). This was not noticeably different for those in the youngest age group. In the oldest age group, the difference between those in stage 2 hypertension vs. other categories was more stark for females compared to males.


Normal
Elevated
Stage 1 hypertension
Stage 2 hypertension
Age 0 − 17 years
Age 18 − 24 years
Age 25 − 39 years
Age 40 − 59 years
Age 60+ years
Female
Male
Number of individuals
2500
2000
1500
1000
500
0
Hypertension category