Assessment Week 3

Sabrina Ianni 260189219

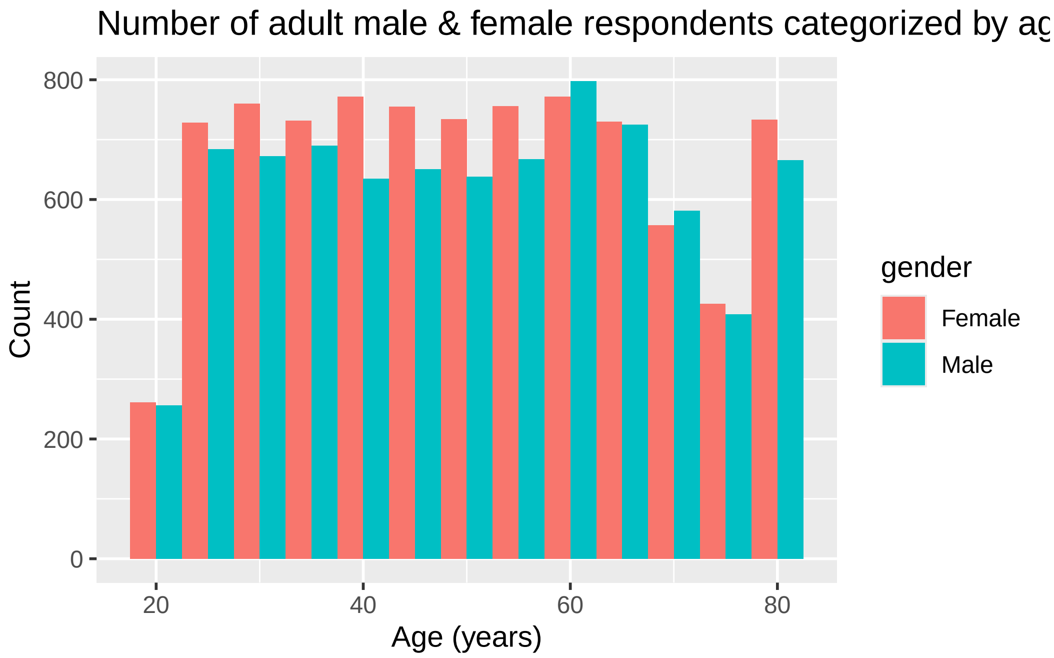
2025-09-21

In this assessment, we will explore different methods of data manipulation through plot reproduction and customization. With `ggplot`, we can gain insights on data using various visualizations.

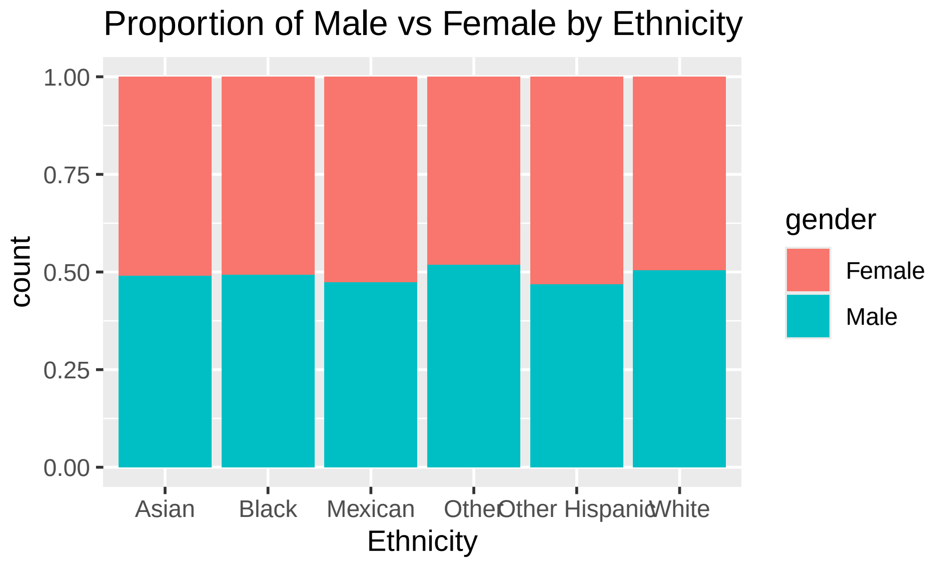
Exercise 1: Reproducing and arranging ggplot2 figures

Reproducing the gender **and the** ***ethnicity*** **distributions using updated dataset `cleaned_NHANES.csv` to include meaningful names and appropriate category labels:**

Gender distribution of adult participants:

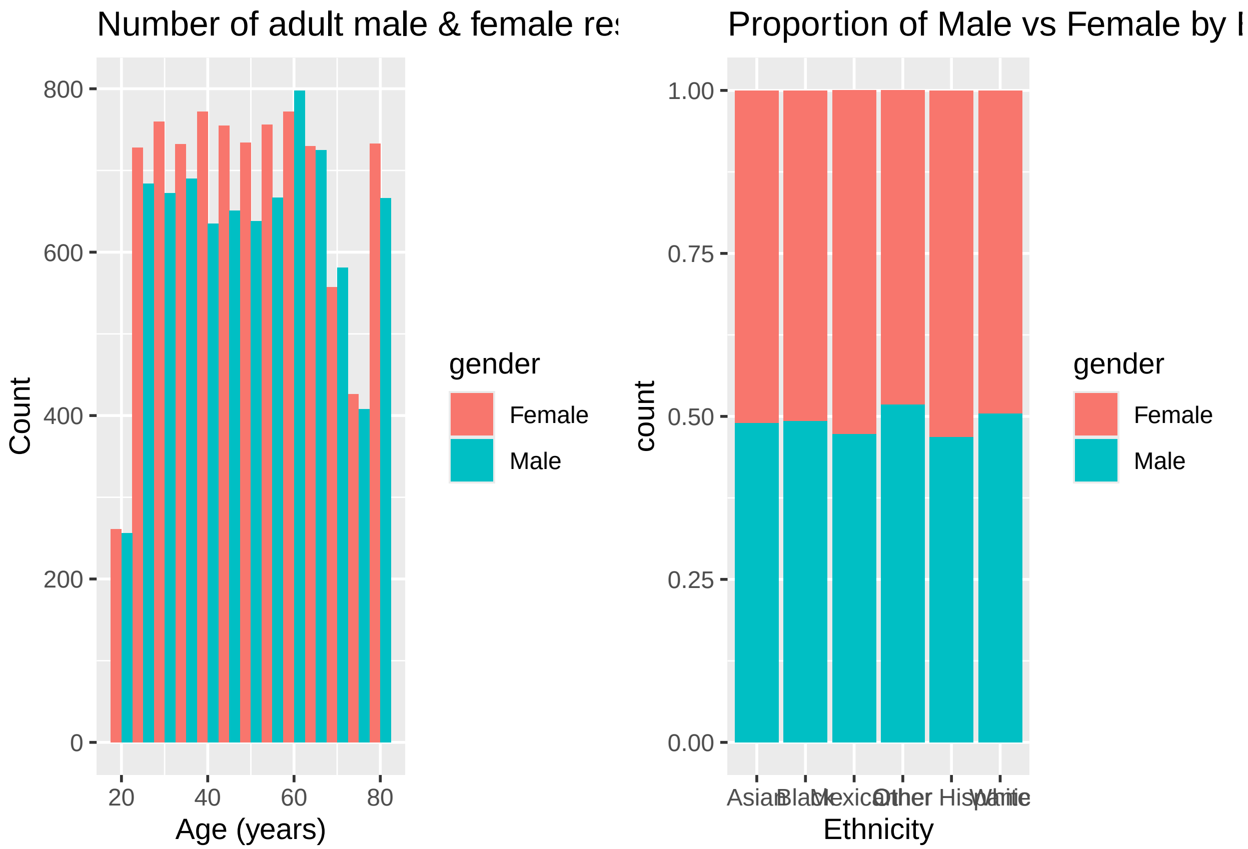


Proportion of male and female respondents categorized by ethnicity:

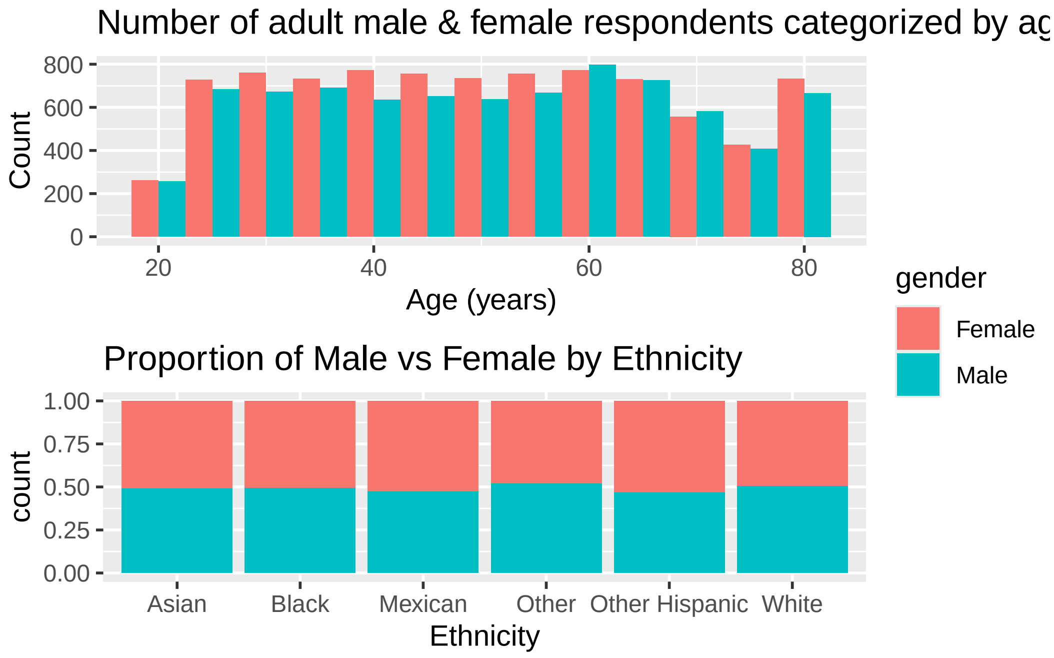


Combining the two plots together using `cowplot` and using `ggarrange()`:

Using `cowplot`:



Now, using `ggarrange()`:

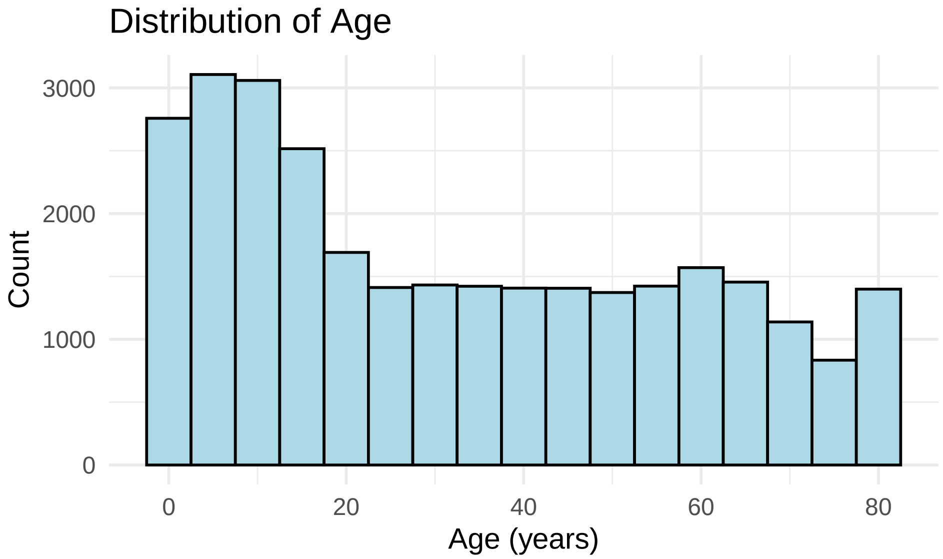


Using cowplot arranges the plots side by side in a sequential row, but will generates two identical legends, taking up space and is redundant. In contrast, using `ggarrange` allows formatting of the layout and legend, creating a visually cleaner, less cluttered appearance.

Exercise 2: Visualizing key characteristics

Exploring the distribution of key variables - specifically: age, gender, and the two ethnicity variables of the `cleaned_NHANES` dataset.

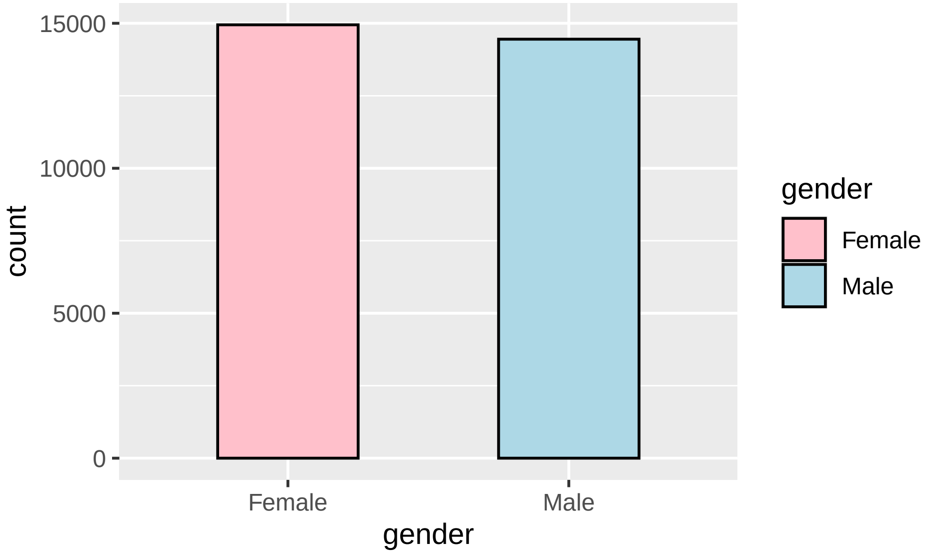
Visualization of the 'age' distribution:



There appears to be more respondents between the ages of 0 to 20 years, in comparison to older age groups.From the ages of 20 to 80, there appears to be similar number of respondents (close to 1500), with slightly less than 1000 respondents for those aged ~75 years of age.

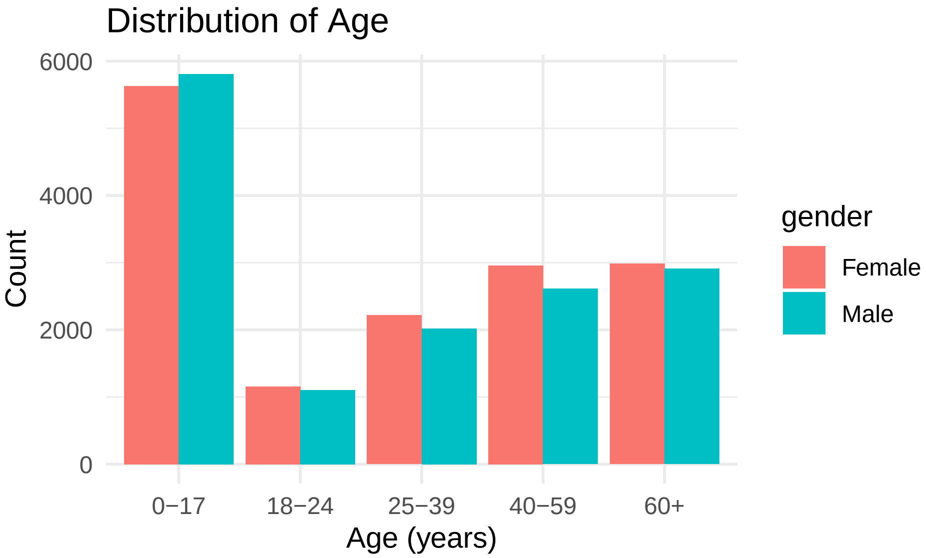
Visualization of the 'gender' distribution:

```
## NULL
```



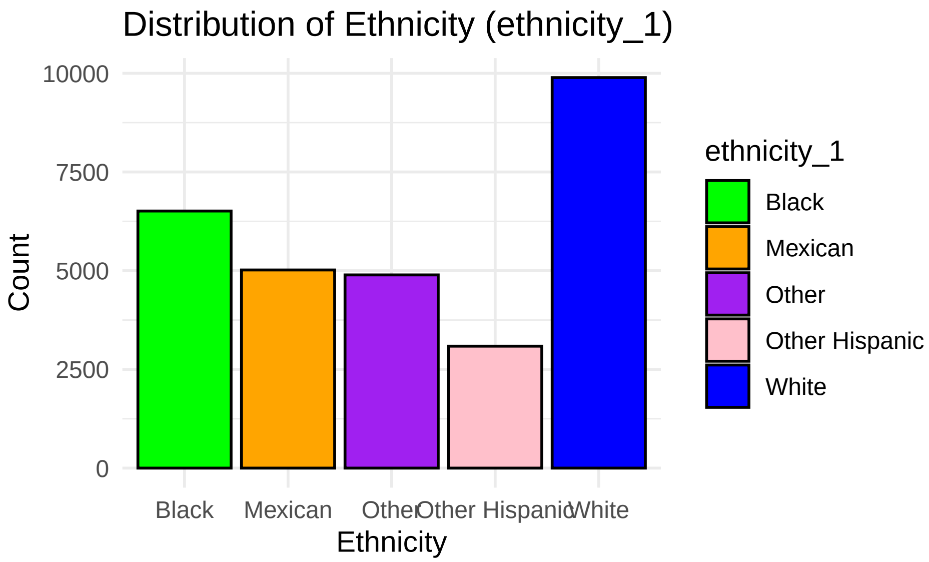
The split between male and female respondents is almost the same, but there are slightly more females in this distribution, with close to 15000 participants in each group.

We can also look at the age vs gender distribution combined:



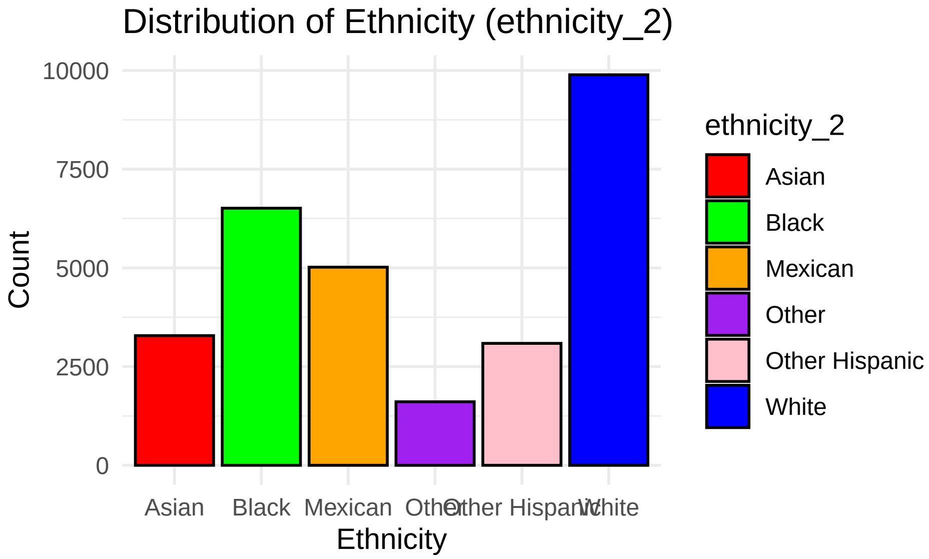
When looking at this distribution, we see an almost even split between male and female. There are slightly more males in the 0-17 age category. This category also has the largest number of participants. The least number of participants is in the 18-24 category, with slightly more females than males in the subsequent age groups from 18 to 60+.

Visualization of the 'ethnicity_1' distribution:



Here, we see close to 10000 respondents in the "white" category, approximately ~ 5800 in the "black" category, close to ~5000 in the "Mexican" and "Other" categories and ~3000 in the "Other Hispanic" category.

Visualization of the 'ethnicity_2' distribution:

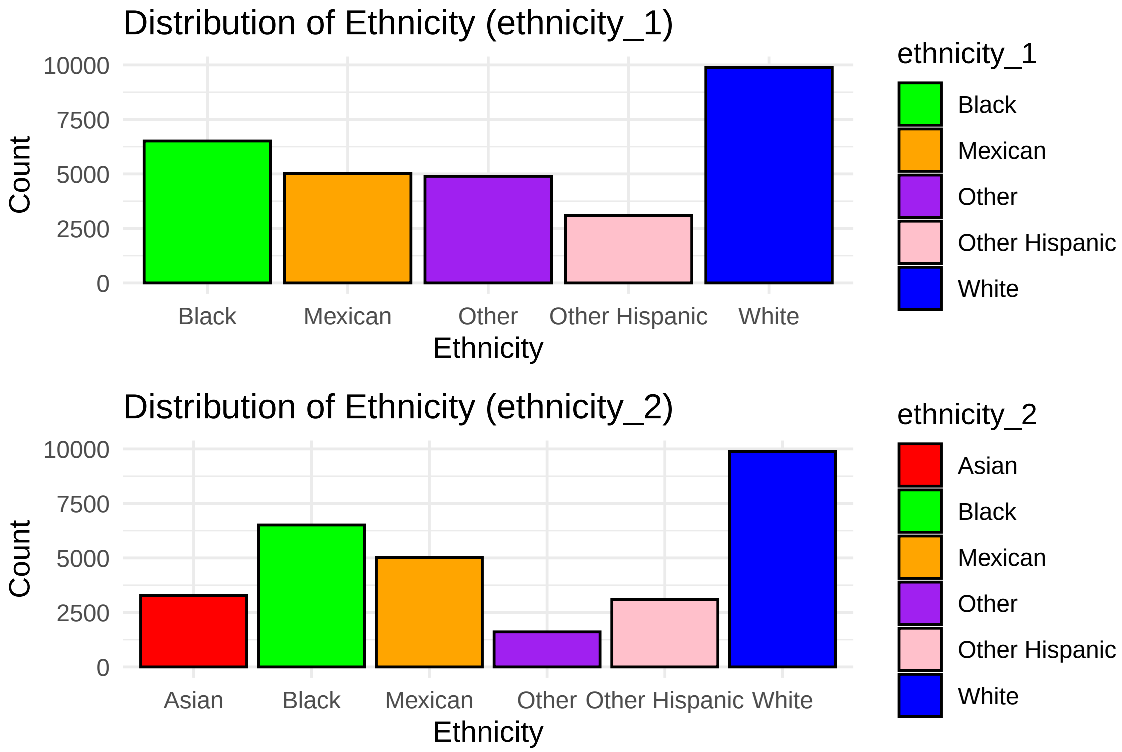


Similarly to `ethnicity_1`, we have the same representation in the "white", "black", and "Mexican" and "Other hispanic" categories. However, we have the addition of the "Asian" category which account for approximately ~3200 participants, and lowering the "Other" category to approximately ~1400.

In the above distributions for age, gender and ethnicity, we can see there are no missing values in the selected variables.

```
##         age      gender ethnicity_1 ethnicity_2
##           0           0           0           0
```

Now looking at the two ethnicity distributions:



We can see `ethnicity_2` includes "Asian", which is missing in `ethnicity_1`. Having an extra category helps better represent the sample diversity. The `ethnicity_2` distribution separates "Other hispanic" and "Asian", giving a clearer distinction, and more accurate analysis of the data. Removing `ethnicity_1` helps avoid confusion or duplicated information.

```
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
```

Exercise 3: Improving ggplot figure for diet study

*** The orginal figure demonstrates each participants' diet trajectory, however having twenty participants and twenty different colors can clutter the legend and the plot, and can make it more difficult to read the results.**

*** Additionally, with all the lines on one grid, it is difficult to highlight one person's trajectory. Having a plot for each participant can give a better representation of that participant's trajectory.**

*** The color palette used is not visually distinct when plotted on the grid. Some colors are similar and are hard to distinguish when plotted all together.**

*** The axes could be more descriptive:**

- X axis - instead of "Week", can use "Time (weeks)" or "Weeks since start"

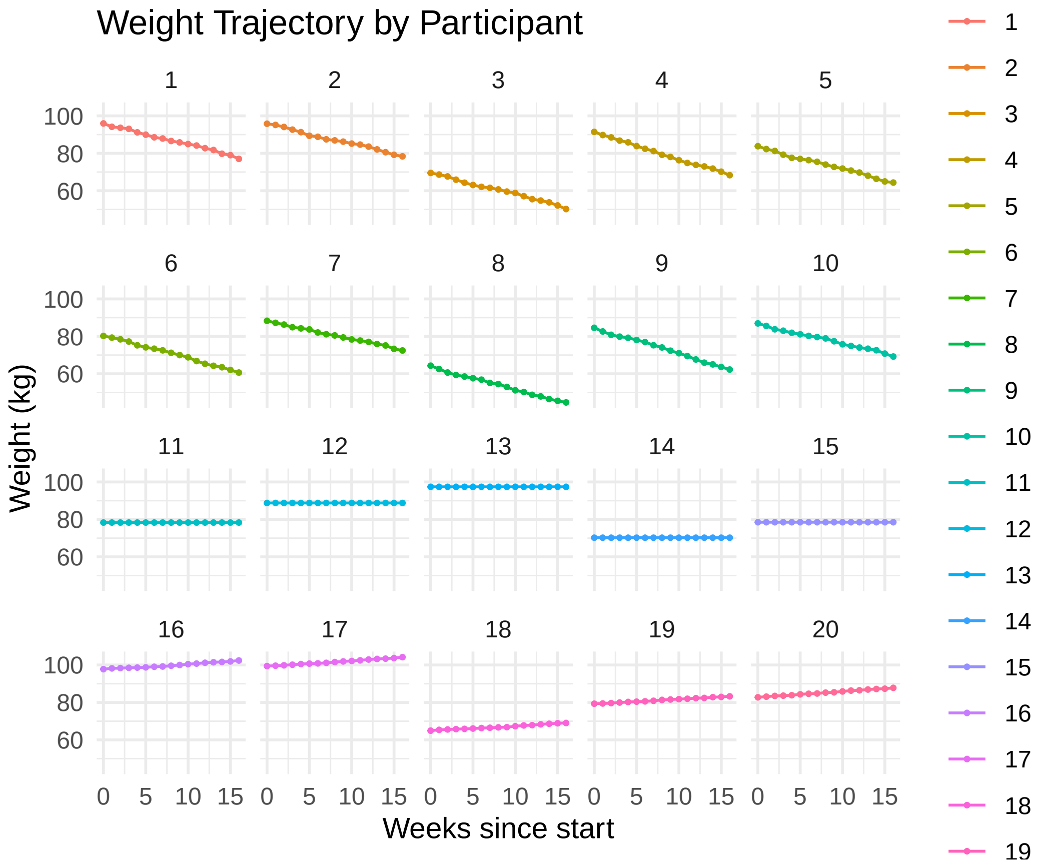
- Y axis - can add units to the "Weight" to make it clearer

*** Can use `theme_minimal` to lighten the background of the grid.**

Here is an example of a clearer represenation of the data, highlighting each participants trajectory. Using the `facet_wrap()` function, and plotting each participant on individual scatter plots to allow for a better visualization of each participant's trajectory.

In this case, the color palette is acceptable, as each line is on a individual grid. However, the legend remains crowded with 20 participants' data being included, but it is easier to visualize.

The new and improved figure:



Exercise 4: Exploring relationships through visualizations

Now we will be looking at the relationships between age, gender, and systolic blood pressure from our `cleaned_NHANES` dataset. Because our original dataset has four different systolic blood pressure (SBP) readings, we will use the average value of systolic blood pressure for each participant. Here is an example of our new categorical variable `sbp_avg` (average systolic blood pressure):

```
##    sbp_avg
## 1 112.6667
## 2 157.3333
## 3 142.0000
```

Now we want to create a new dataset that includes individuals who have non-missing values for age, gender, and average systolic blood pressure.

```
# Use filter() function to create a new dataset
cleaned_NHANES_filtered <- cleaned_NHANES_3 %>%
  filter(!is.na(age), !is.na(gender), !is.na(sbp_avg))
```

To have a better understanding what the average SBP scores mean for the participants, we will categorize the average SBP measures them into the following categories:

Normal : SBP $<$ 120mmHg

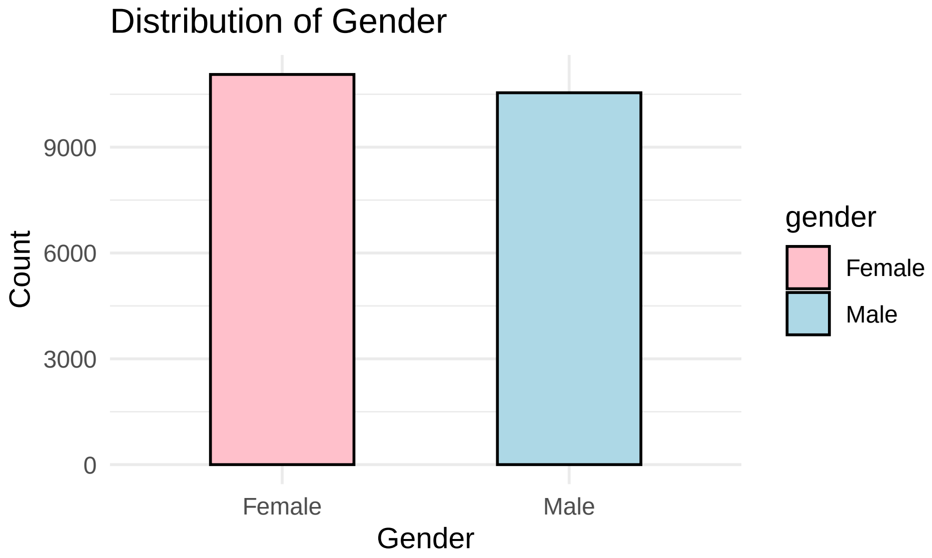
Elevated: SBP = 120 - 129 mmHg

Stage 1 hypertension: SBP = 130 - 139 mmHg

Stage 2 hypertension: SBP $>$ or = 140 mmHg

Looking at our new data we can see how some of the distribution changed:

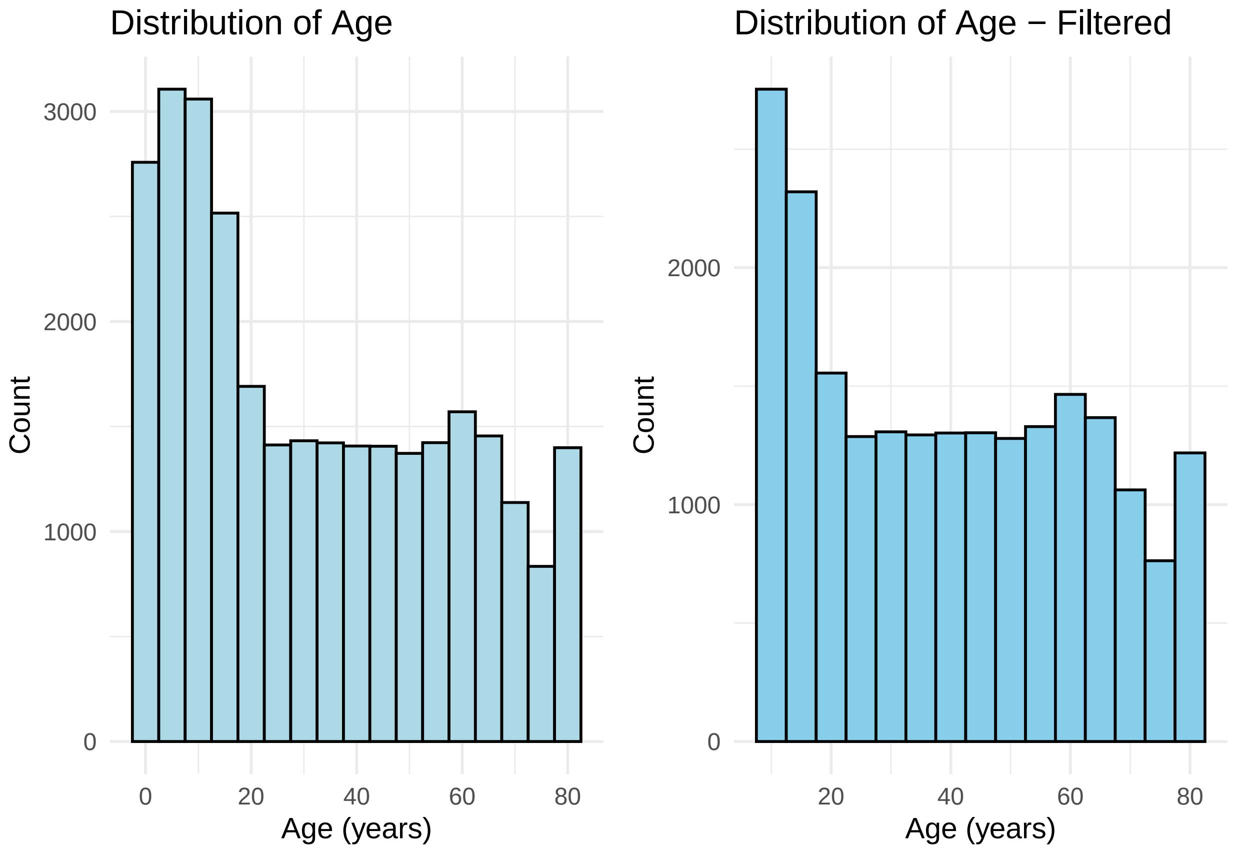
Here is the gender distribution for our new sample:



We can see we still have similar proportion of male to female, but the total number of participants is less due to only including those participants with non missing data (about 11000 vs 15000 from the original data).

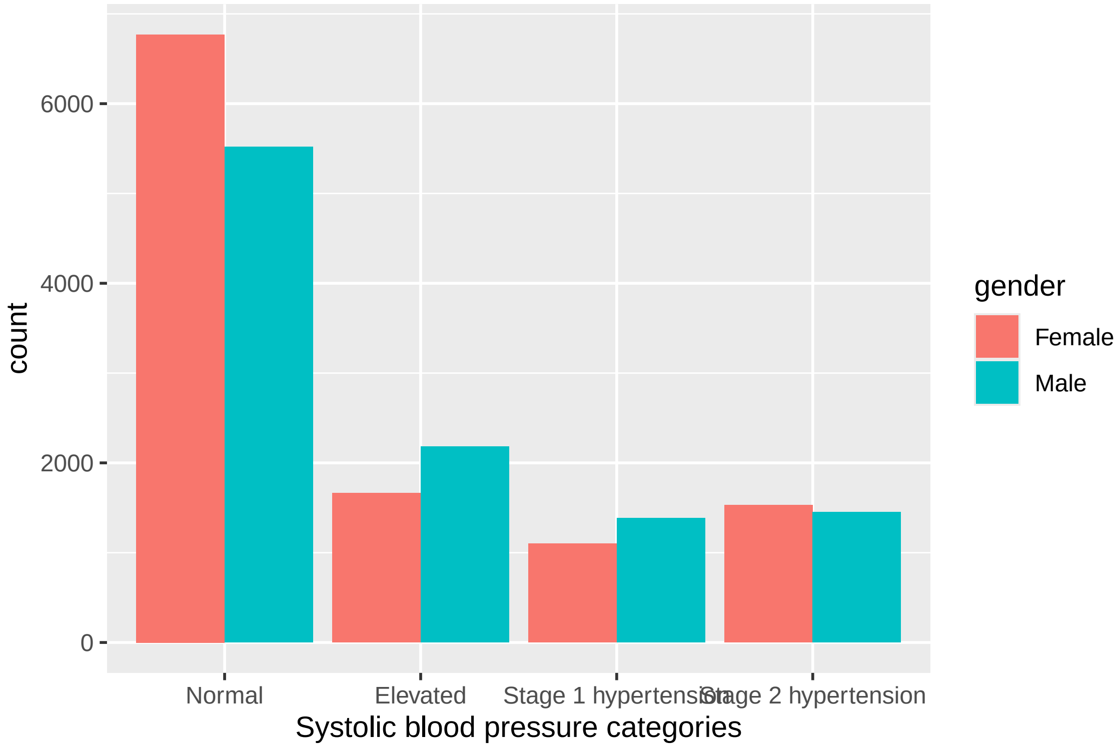
Now we'll look at the age distribution of our data. We see how the distribution remains nearly the same as our original data, with most participants aged from 0-20. However, there are much less participants in the younger age groups from our filtered data.

When comparing the age groups between the two datasets, we can see the differences here:


Distribution of Age

Distribution of Age − Filtered

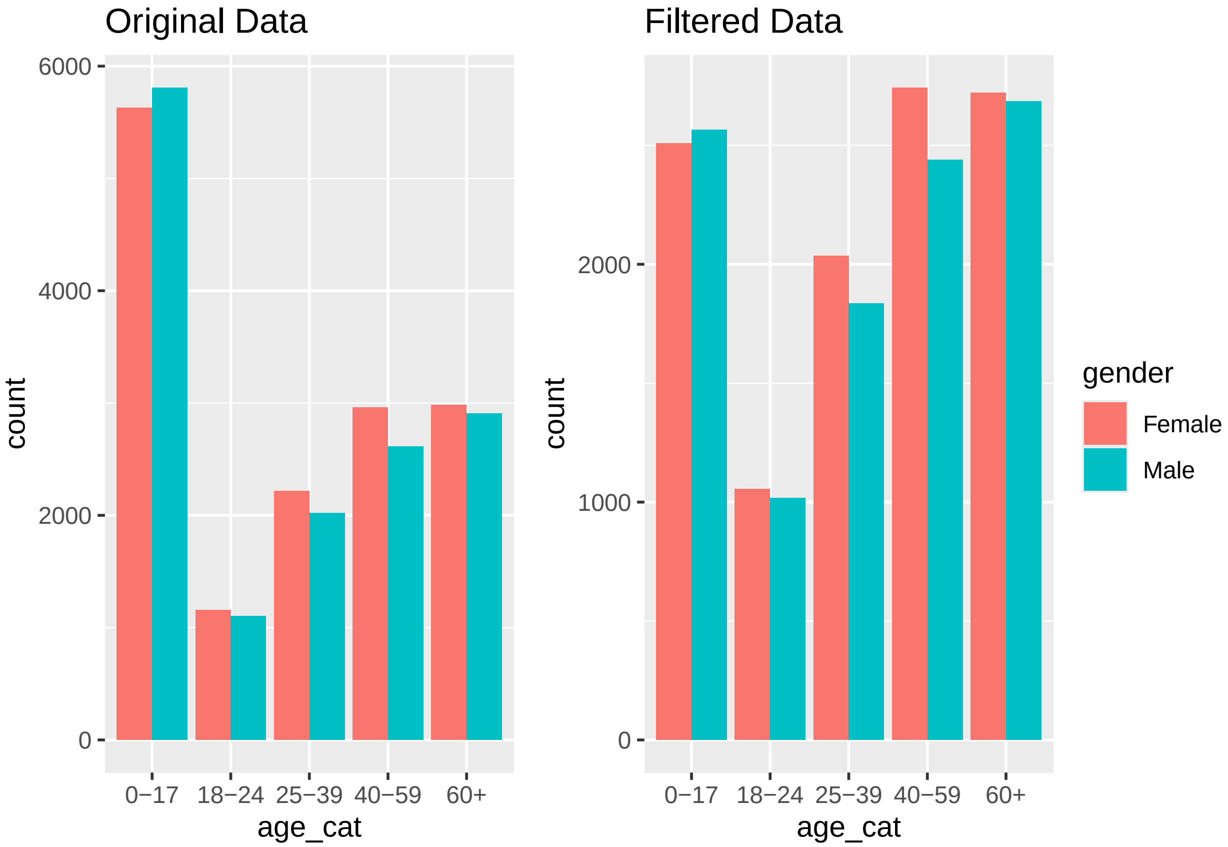
Next, we will look at the differences in blood pressure between males and females:



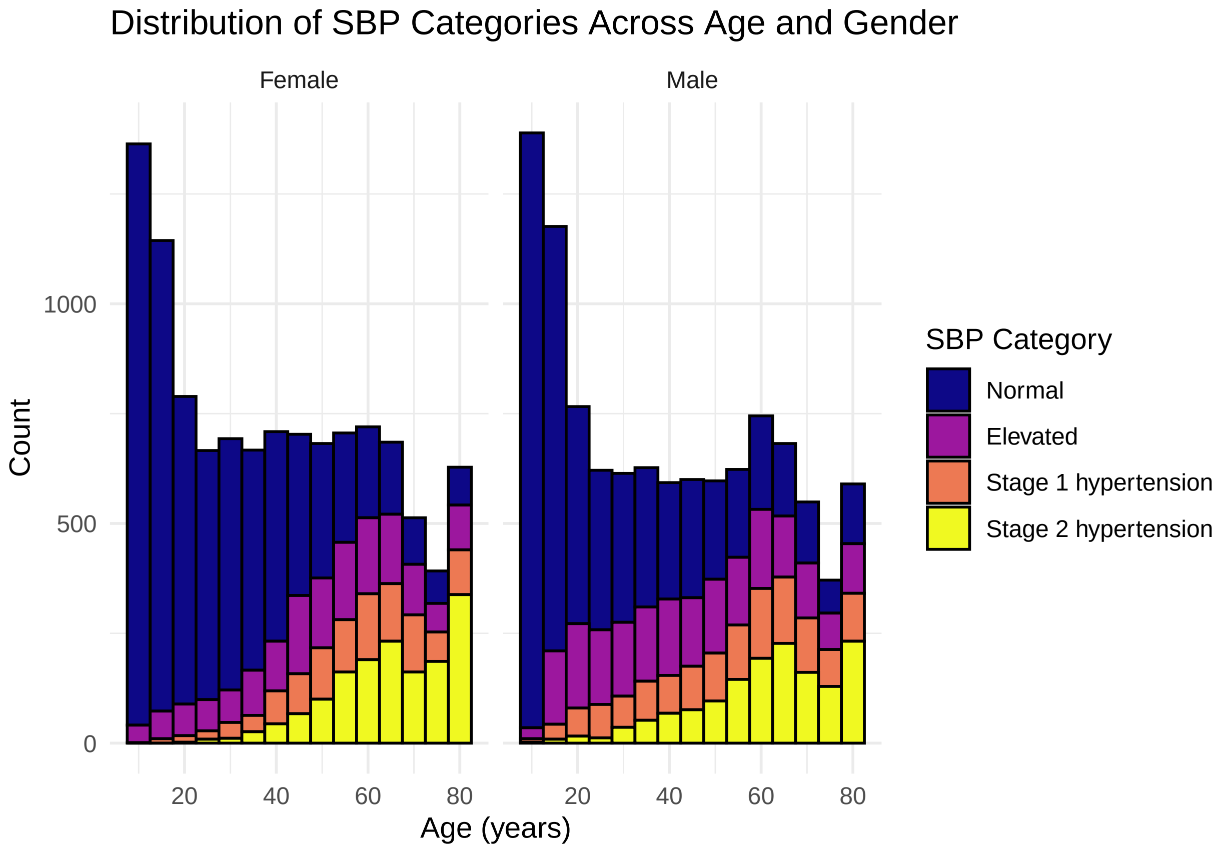
We can see from this bar plot above most people categorized in the "normal" group, with more females than male. This trend is likely explained by the younger age distribution of participants, as younger individuals are generally healthier.

There are more men than women in the "Elevated" and "Stage 1 hypertension" groups, with slightly more women than men in the "Stage 2 hypertension" category.

Similarly, we can analyze mutiple variables from our original and filtered data. Looking at the age and gender distribution below, we see similar counts in males and females across all age categories, and again, we now see a larger proportion of the participants aged 25 to 60+ in the filtered data (those with non missing values):



Lastly, we will look at how hypertension changes across age in men and women:



When looking at the data above, there are similar counts in both males and females. Secondly, we see the younger participants have "Normal" blood pressure in both men and women, as expected in younger, generally healthier individuals. However, we see more men than women developping elevated SBP earlier on in life. Across both genders, we see higher number of individuals with "Stage 1" and "Stage 2" hypertension as age increases.

End of analysis