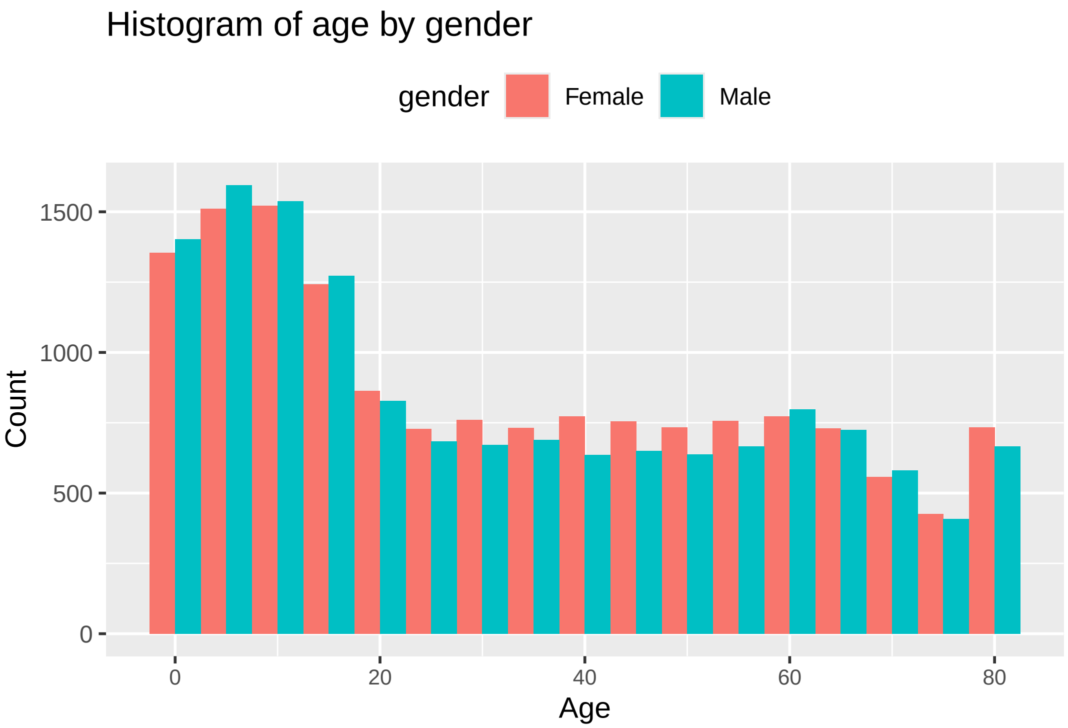
Assessment 3 Alicia Phan

2025-09-19

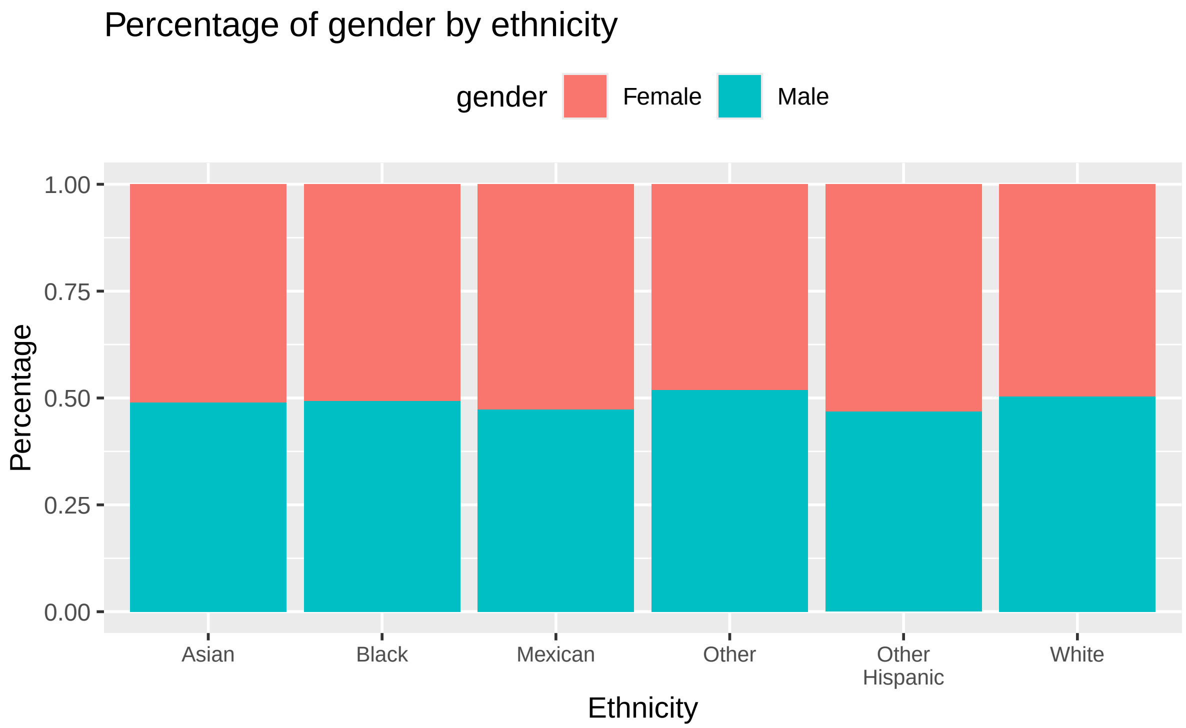
Exercise 1 - Reproducing and arranging ggplot2 figures

Recreating two figures (age and gender, ethnicity and gender)

Histogram of age by gender

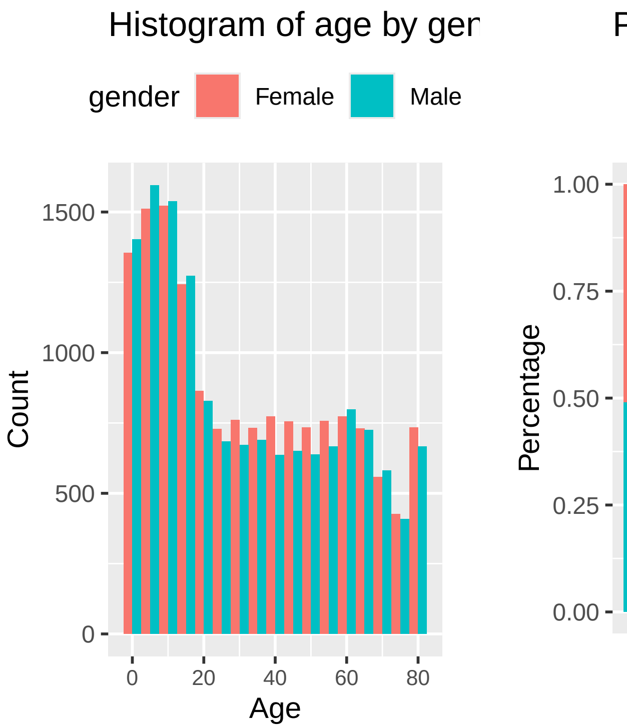


Percentage of gender by ethnicity

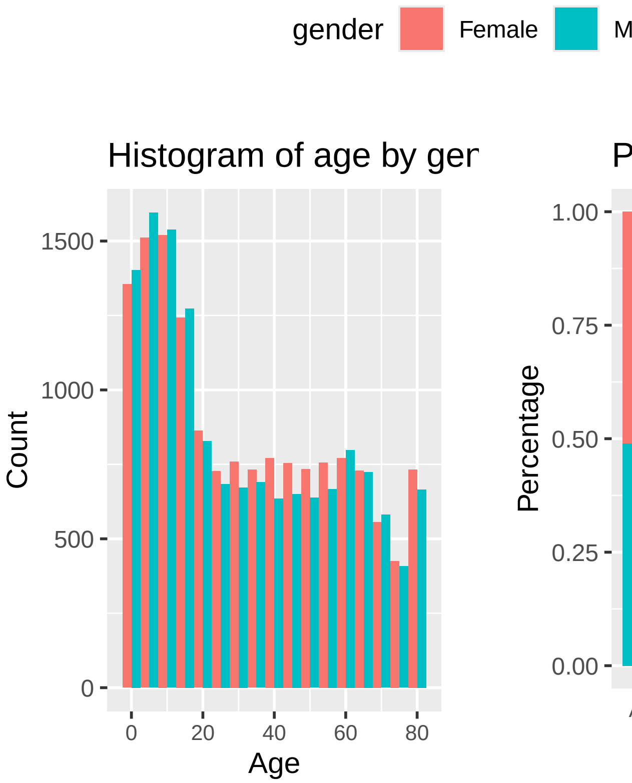


Combining the two plots:

Please refer to the combined plot below showing the two methods used. The two methods are easy to use to combine multiple plots to one plot. However, there were several additions added to the original plot to ensure readability of the plots. For both methods, the x axis for the second plot showing ethnicity showed some overlapping text due to the plot size when combined. From this, label_wrap function was used to wrap the x labels and ensure readability. The ggarrange method shows easier way to combine the two plots with shared legend, the two plots here shared the legend of gender column, which is faciliated through the field "common.legend = TRUE" field. In comparison, the plot_grid function from cowplot package does not have this field; however, there is a solution to add common legend by extracting the legend.



Using plot_grid() function from cowplot package

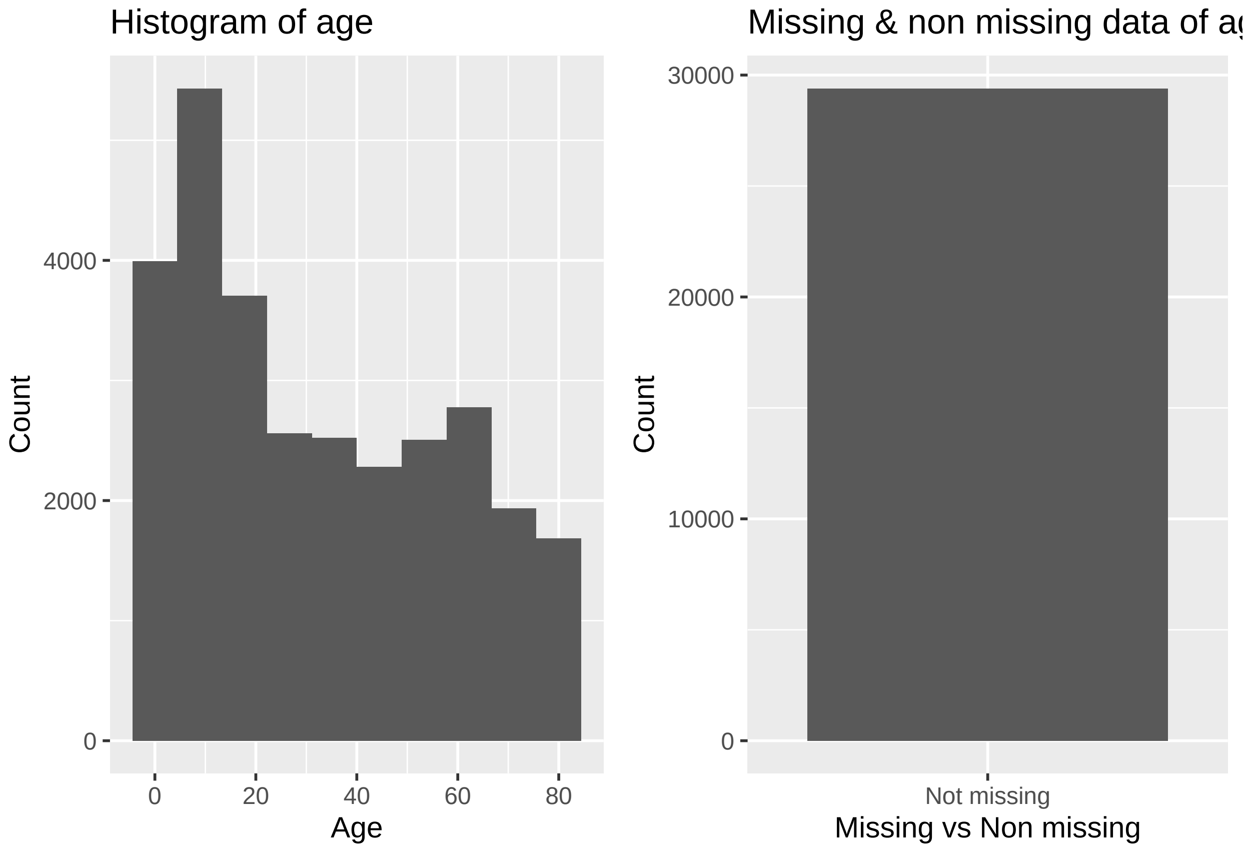


Using ggarrange() function from ggpubr package

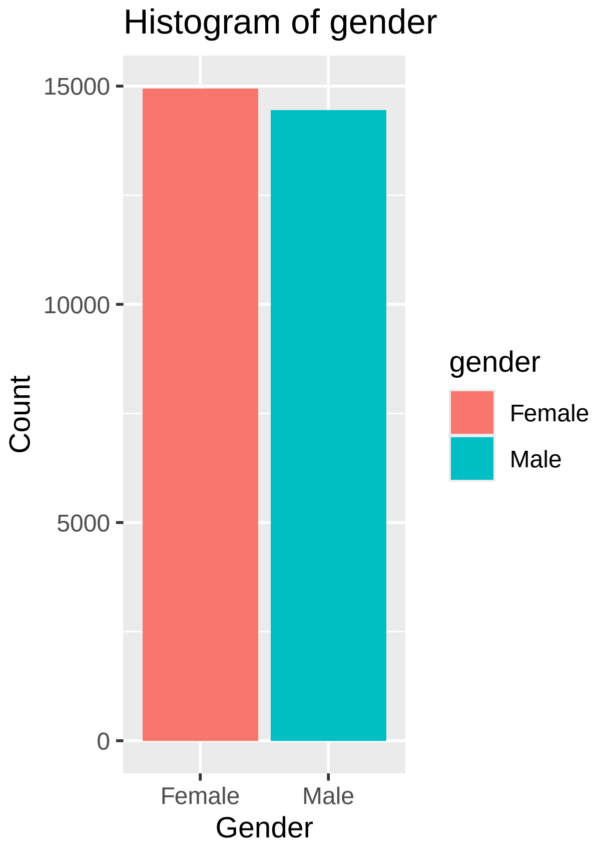
Exercise 2 - Visualizing key characteristics

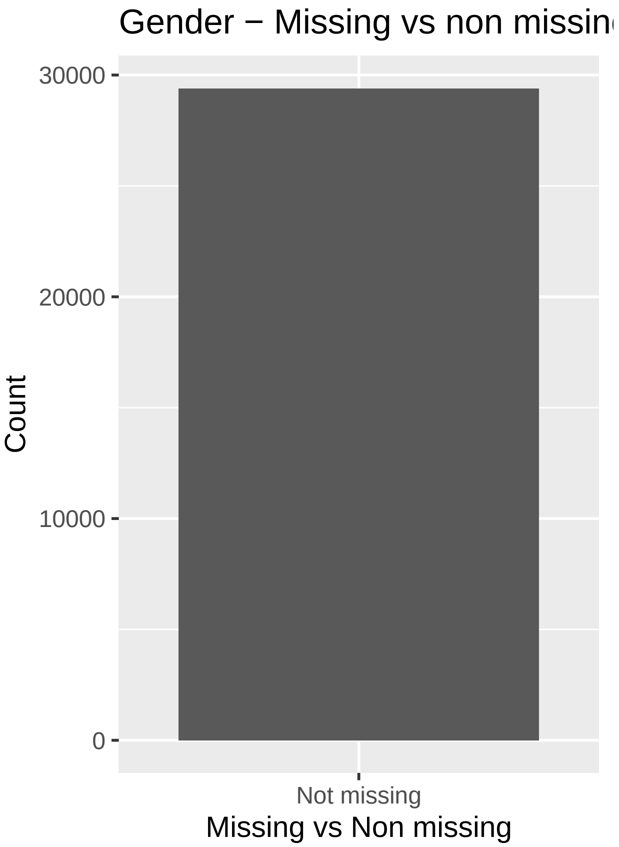
Visualisations of the following variables: Age, gender, ethnicity variables

Age


Histogram of age
Count
4000
2000
0
0
20
40
60
80
Age
Missing & non missing data of ag
Count
30000
20000
10000
0
Not missing
Missing vs Non missing

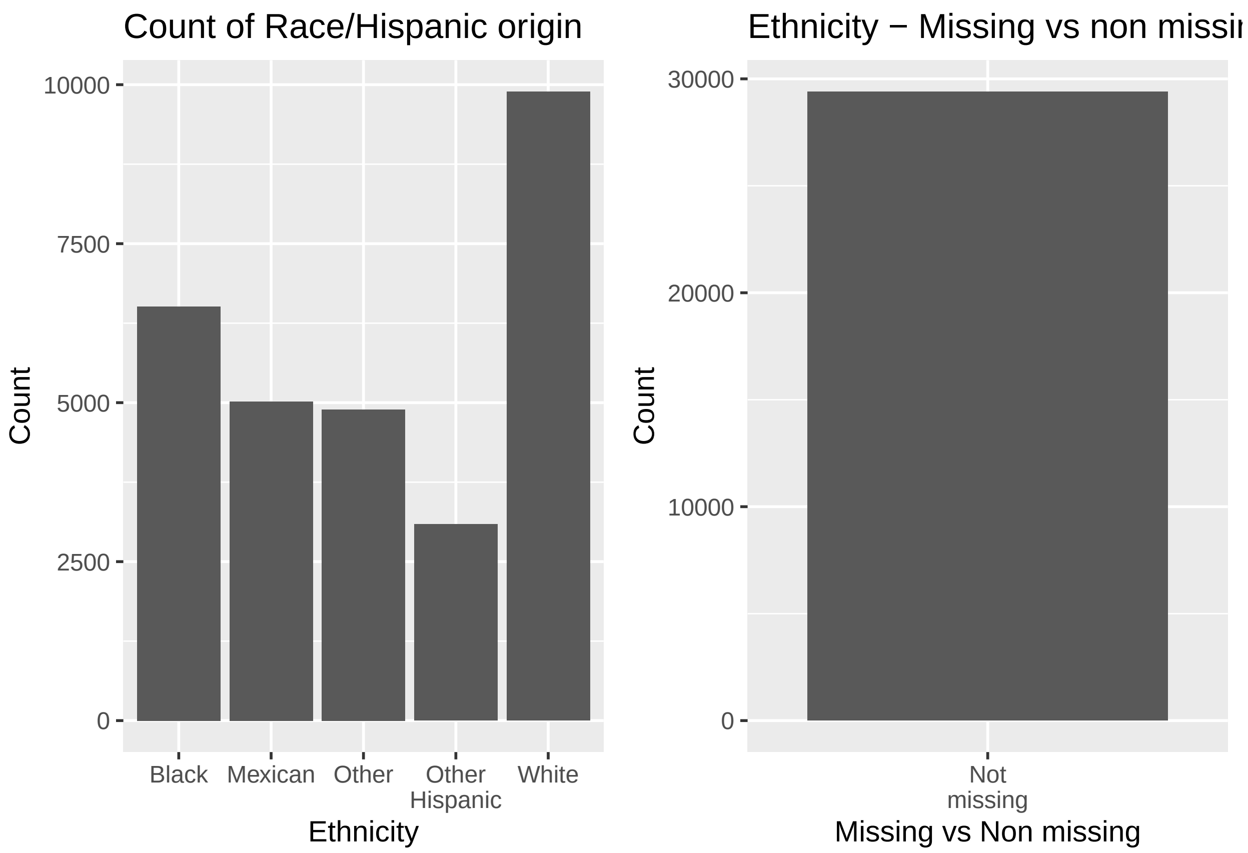
Gender



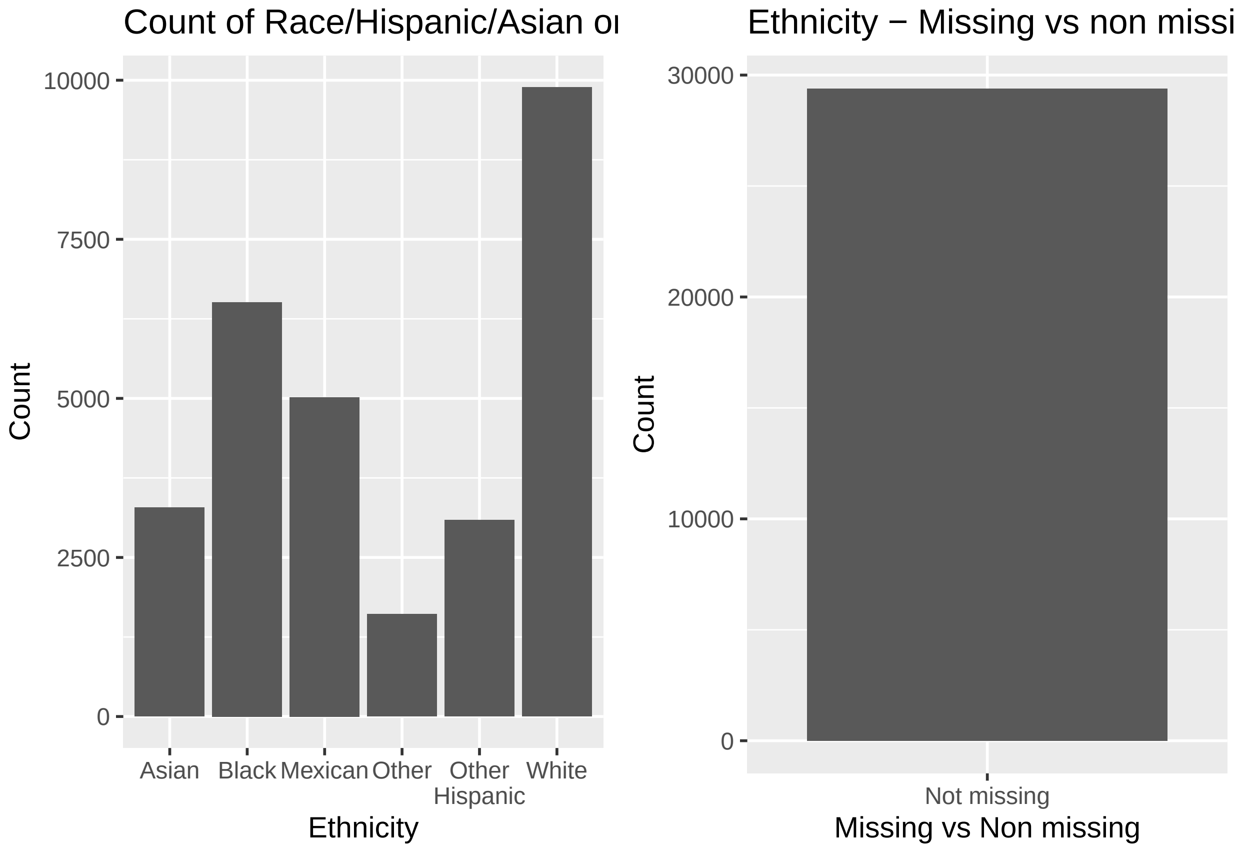


Ethnicity variables

a. Ethnicity 1



b. Ethnicity 2



```r
ggsave("age_combined.png", age_combined)
ggsave("gender_combined.png", gender_combined)
ggsave("ethnicity1.png", ethnicity1_combined)
ggsave("ethnicity2.png", ethnicity2_combined)
```

Based on the results of both variable, it is shown that there were 5000 observation of “other” category in the first variable, whereas only around 1000 observations of “other” category in the second one. From this, it is more preferable to select the variable Ethnicity 2 to include more details counts of the ethnicity variable that includes Asian category. The Ethnicity 2 variable is able to explain more of the data. Therefore, the ethnicity 1 variable will be removed.

```r
nhanes_data <- nhanes_data %>%
  select(-ethnicity_1)
```

Exercise 3 - Improving ggplot figure

For interpretation of the gragh, the graph is missing descriptive elements like title of gragh, legend of the graph, weight measurement value. In addition, there were too many colors used for the graph for each participant; the colors used were rainbow-color, which might not be friendly for color deficiency readers.

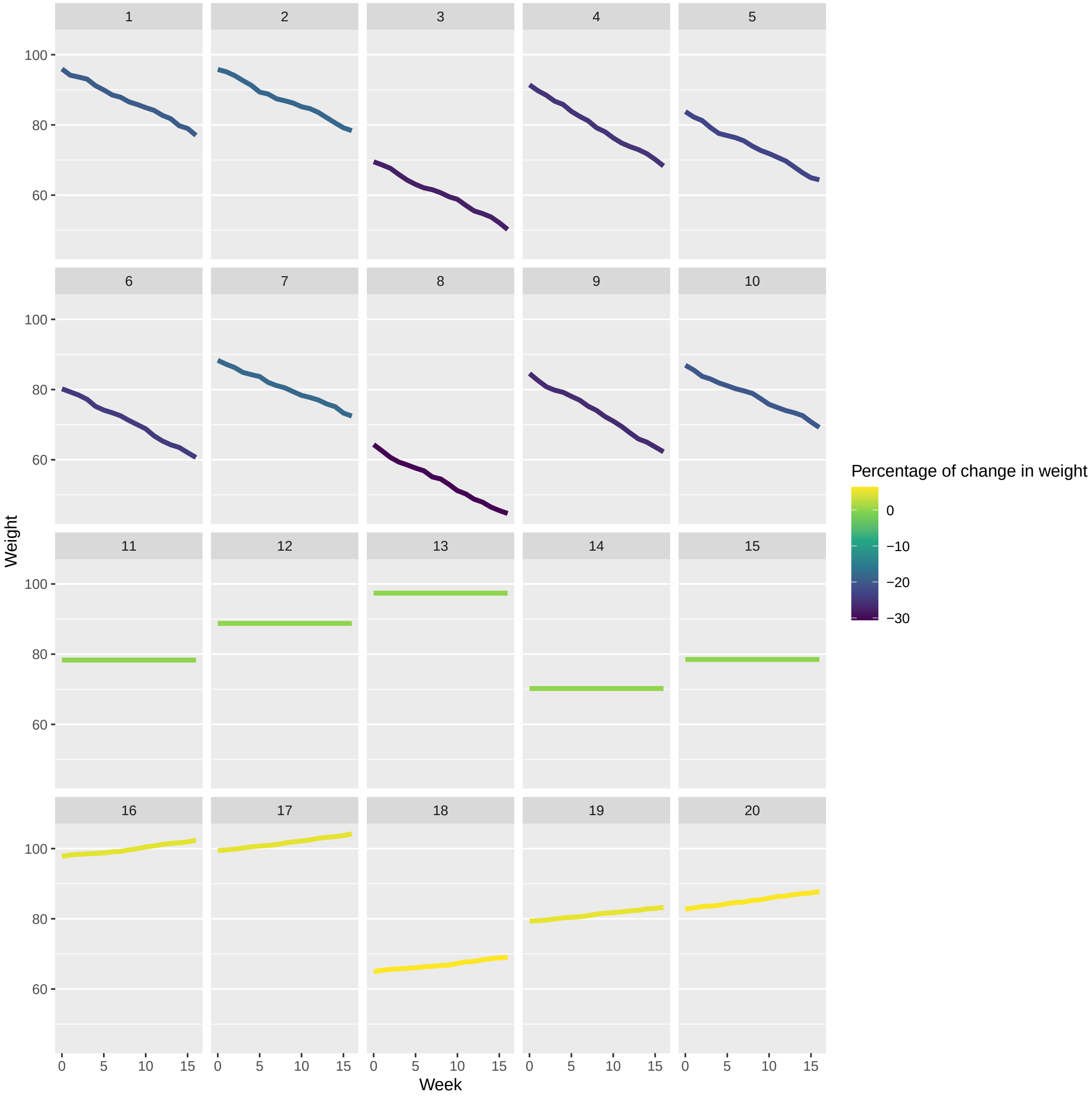
For visual representation, the area for the color is very minimal for each participant, and there is overlapping between lines. In addition, the background grid has too many lines. In addition, the lines are too thin as well, make it hard for line visibility.

Regarding data completeness, it seems that the legend for participant 1 is missing. This is due to the length of the legend that does not fit the plot. This also causes the confusion and the completeness of the data when presented in the plot.

The following suggestions are proposed:

- Title, Legend name, and weight measurement value should be added.
- Background grid should be simplified.
- If possible, the information on how much the participant increase/decrease can be shown, and that information on each participant should be available. Proposition on creating categories of increase/decrease from beginning to end for each participants.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Exercise 4 - Exploring relationships through visualizations

The report aims to understand the relationships between age, gender, and systolic blood pressure (SBP). In order to explore the relationships between variables, it is important to verify the data completeness and the data itself. /

Below is a summary of the NHANES data that will be used for analysis showing the missing data and completion rate of each variable. Based on the data, we noted that there are four SBP measurements representing different time of measurement. In addition, the SBP measurements include missing data where the completion rates are 70%, 72.3%, 72.1%, and 4.7% for measurements from wave 1 to 4 respectively. Furthermore, both gender and age variables have completion rate of 100% meaning no missing data.

Table 1: Data summary

Name	nhanes_data_analysis_raw
Number of rows	29400
Number of columns	7
Column type frequency:	
character	1
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gender	0	1	4	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
patient_id	0	1.00	88256.50	8487.19	73557	80906.75	88256.5	95606.25	102956	
systolic_bp_1	8784	0.70	119.94	18.91	66	106.00	116.0	130.00	236	
systolic_bp_2	8120	0.72	119.99	19.11	66	106.00	116.0	130.00	238	
systolic_bp_3	8182	0.72	119.66	18.87	62	106.00	116.0	130.00	232	
systolic_bp_4	28007	0.05	128.16	23.81	76	110.00	126.0	144.00	234	
age	0	1.00	32.52	24.91	0	10.00	28.0	54.00	80	

Based on the preliminary data completeness analysis above, the data transformation was conducted to take average of all four SBP measurements to obtain a new SBP variable as average SBP. Regarding the average SBP variable, the variable is missing only if four of the measurements are missing. Subsequently, in order to ensure the quality of analysis, we filtered out the observations where average SBP variable, age or gender variable is missing. Noted that the age and gender variables are completed per above observation, therefore, we only filtered out observations where average SBP is missing.

```r
nhanes_complete <- nhanes_data_analysis_raw %>%
  mutate(avg_sbp = rowMeans(select(nhanes_data_analysis_raw, systolic_bp_1,
                                   systolic_bp_2, systolic_bp_3, systolic_bp_4)
                            , na.rm = TRUE)) %>%
  filter(!is.na(avg_sbp)) %>%
  select(patient_id, age, gender, avg_sbp)

skim(nhanes_complete)
```

Table 4: Data summary

Name	nhanes_complete
Number of rows	21604
Number of columns	4

Column type frequency:	
character	1
numeric	3
Group variables	None

Variable type: character

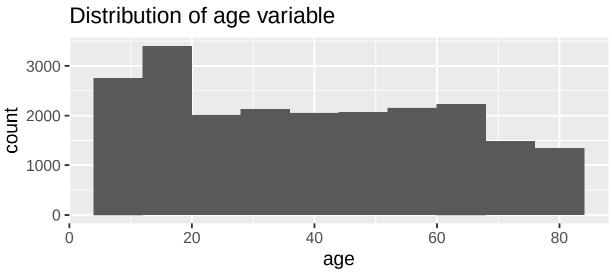
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gender	0	1	4	6	0	2	0

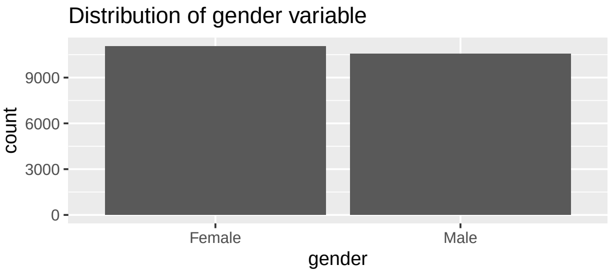
Variable type: numeric

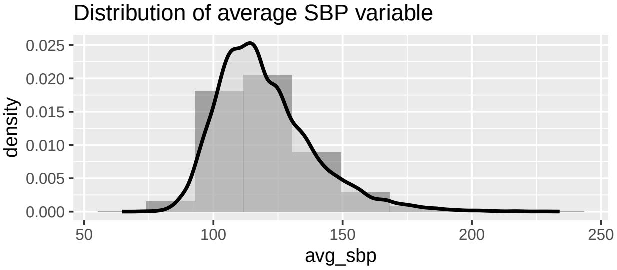
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
patient_id	0	1	88206.39	8461.65	73557.00	80869.75	88179.50	95496.75	102956	
age	0	1	39.92	22.34	8.00	18.00	39.00	60.00	80	
avg_sbp	0	1	120.07	18.94	64.67	106.67	116.67	130.00	234	

After the data cleaning, 7796 observations were removed from the original dataset to remove the missing data. Skim() function was used to ensure the data completeness of the three variables: Age, gender, avg_sbp.

In order to further understand the dataset, we need to explore the distribution of each of the variable.

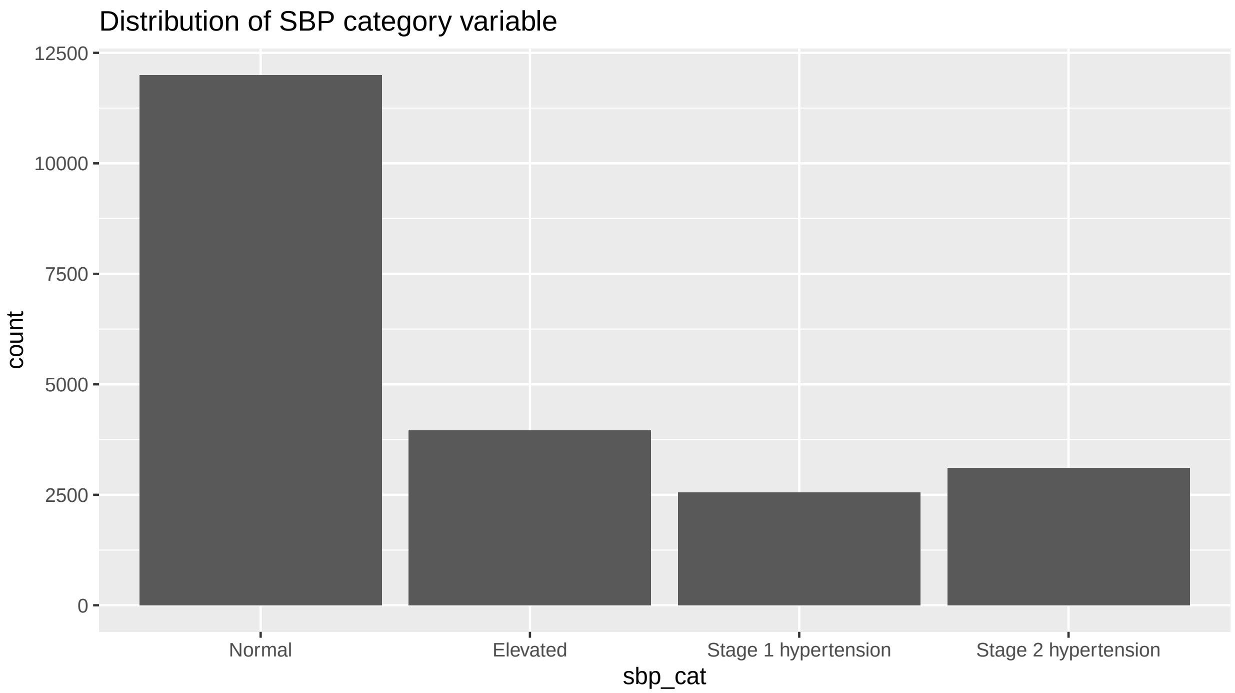




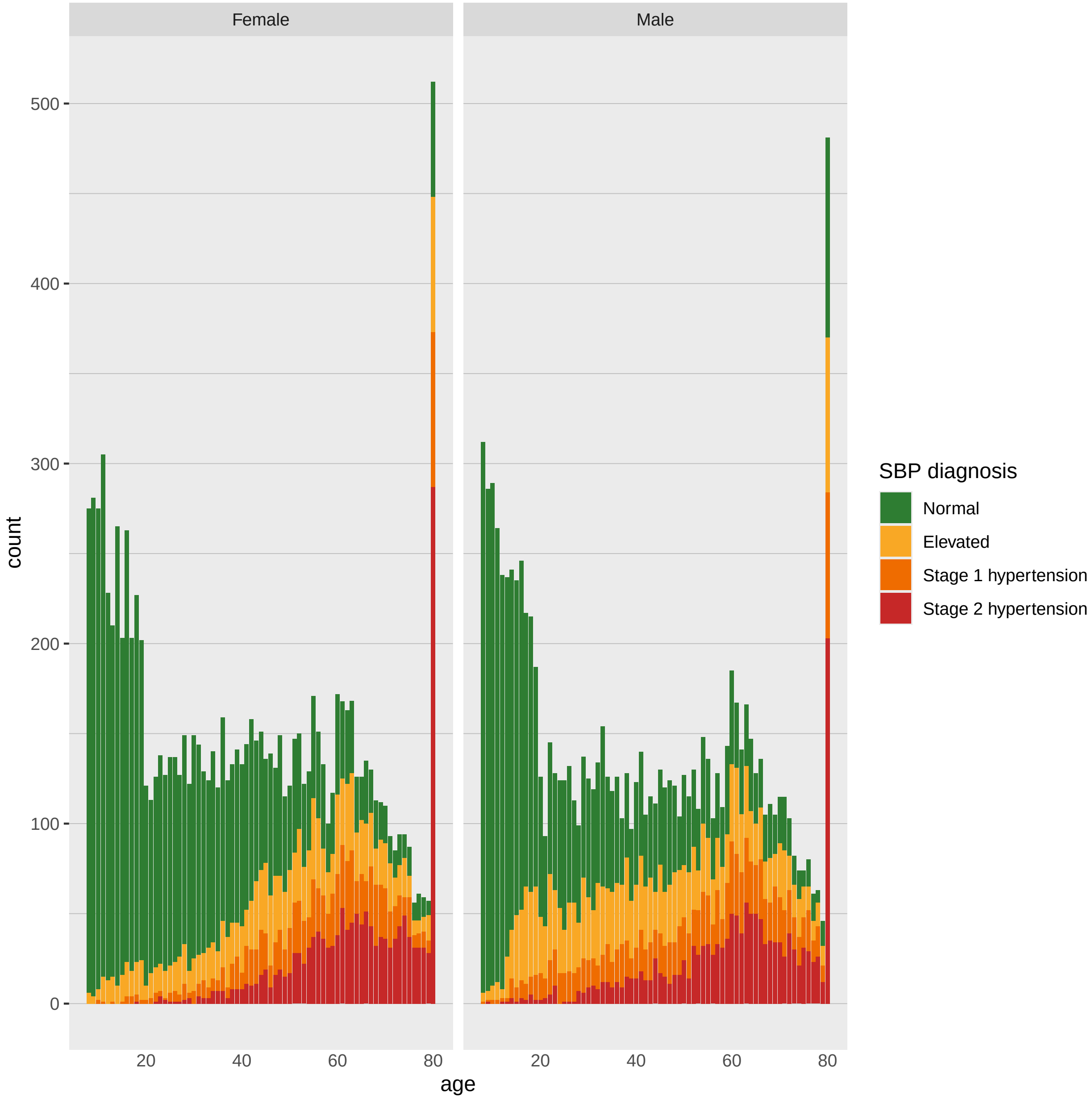


We noted that average SBP is a continuous numeric variable with right skewed distribution. However, looking at only the distribution, there is limited information on where the value of SBP is considered normal. From this, a new categorical variable is created to classify the average SBP observations to the level severity as follows:

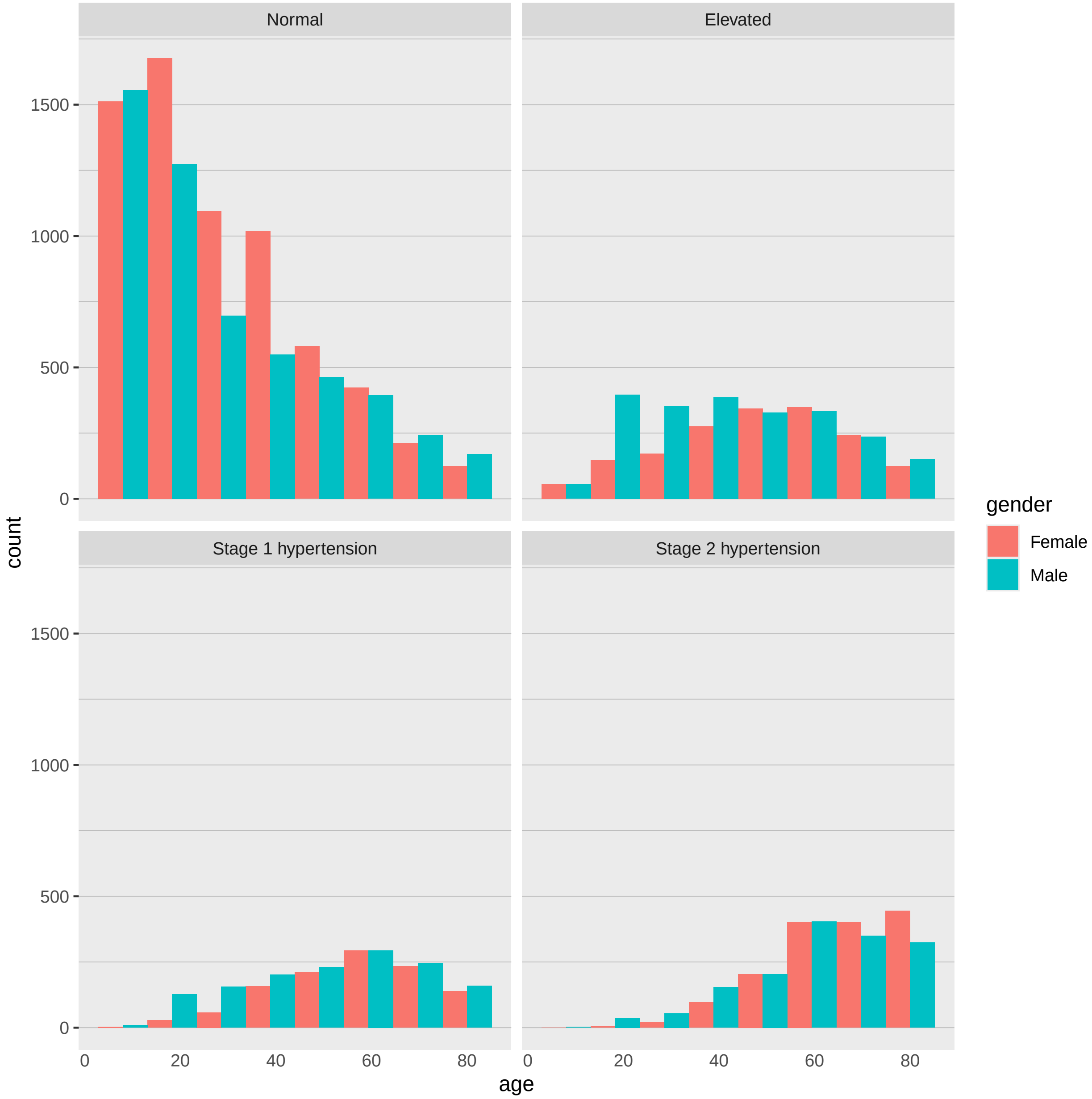
- Normal SBP < 120 mm Hg
- Elevated: 120   SBP   129 mm Hg
- Stage 1 hypertension: 130   SBP   139 mm Hg
- Stage 2 hypertension: SBP   140 mm Hg



Finally, we will explore the relationships between age, gender, and blood pressure diagnosis. We first compared the SBP diagnosis with gender facet. There is an apparent trend that there are more observations for Elevated, Stage 1 and Stage 2 hypertension diagnosis for older age groups than younger ones starting at 50 and 60 years old. We also observed that the distribution of Elevated diagnosis shows less left skew for Male category than Female category. This means that there are more observations that male patients have higher blood pressure at the age of 20-30 years old. Similarly, Stage 1 hypertension for male is also less left skewed comparing to women, with more observations of stage 1 hypertension diagnosis for men than women at around 20-30 years old.



Secondly, we compared again age and gender but with SBP diagnosis as facet. This further confirmed our observations above where there are more male observations with Elevated diagnosis and Stage 1 hypertension at around 20-30 years old, and more female observations with normal diagnosis at the same age group. In contrast, for the age group from 50 years old and older, there is no notable pattern observed between the genders.



In conclusion, we can see that there are more Elevated, Stage 1 and Stage 2 diagnosis for older age groups. In addition, there is a notable observation where male patients have higher blood pressure for younger age groups around 20 - 30 years old. (**caliendo2008?**)

References