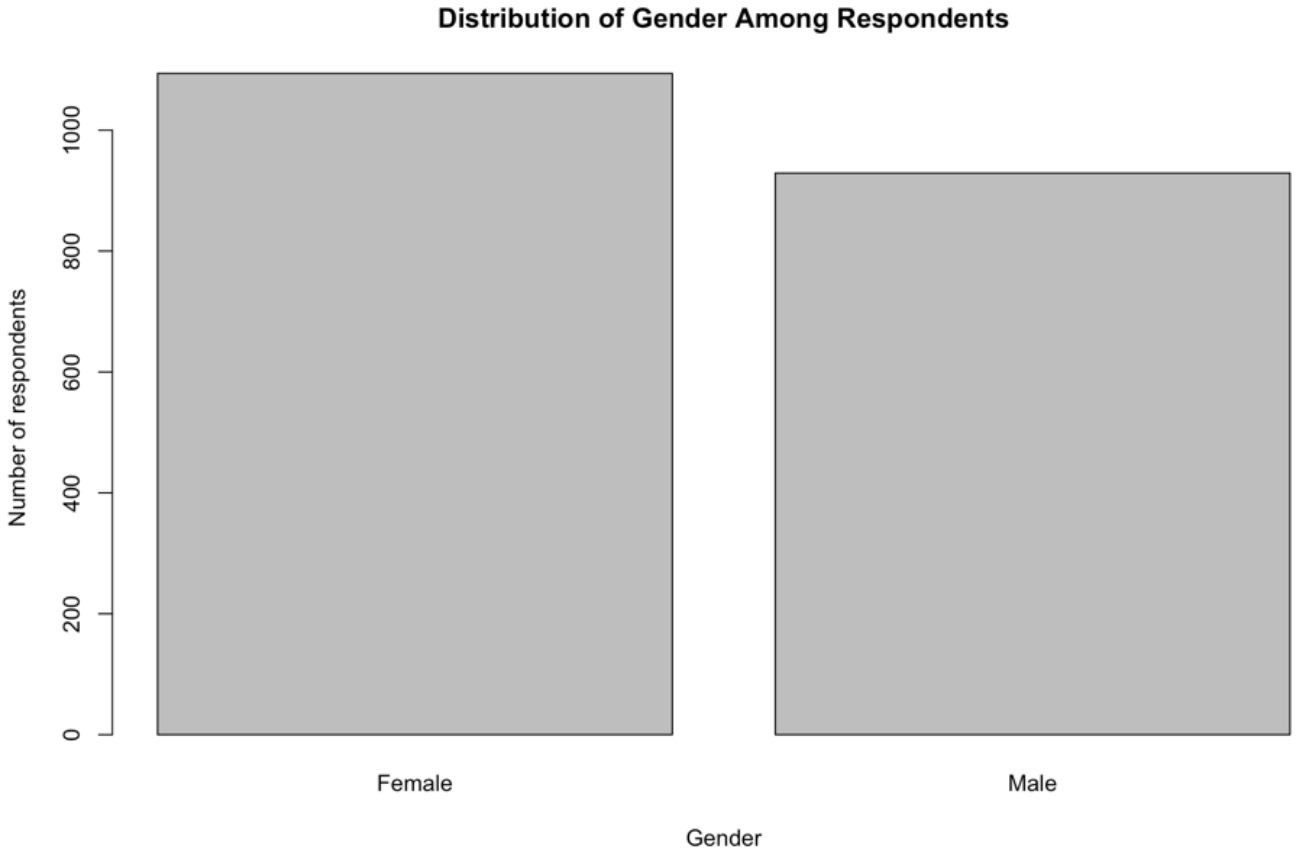
Written answers for Introduction to Statistics - Assignment 1

2.1 Gender (10 points)

1. Describe the variable “newgender" using a bar plot and include the plot in your response. For example, provide a brief description of this variable, including its categories and the number of observations in each category.

Answer:
The variable “newgender” is a **factor created by changing the data type** of “gender” variable, which was originally a character. This is done mainly to make it suitable for running functions in R without taking a lot of compute data and space and it doesn’t have any implications on the data values.
The “newgender” variable has two categories (and no NA’s, just to be clear): Female and Male as seen in the bar plot below. The number of females and males is 1094 and 929, respectively.



2. Report the type of this variable and provide the appropriate measure(s) of central tendency and dispersion.

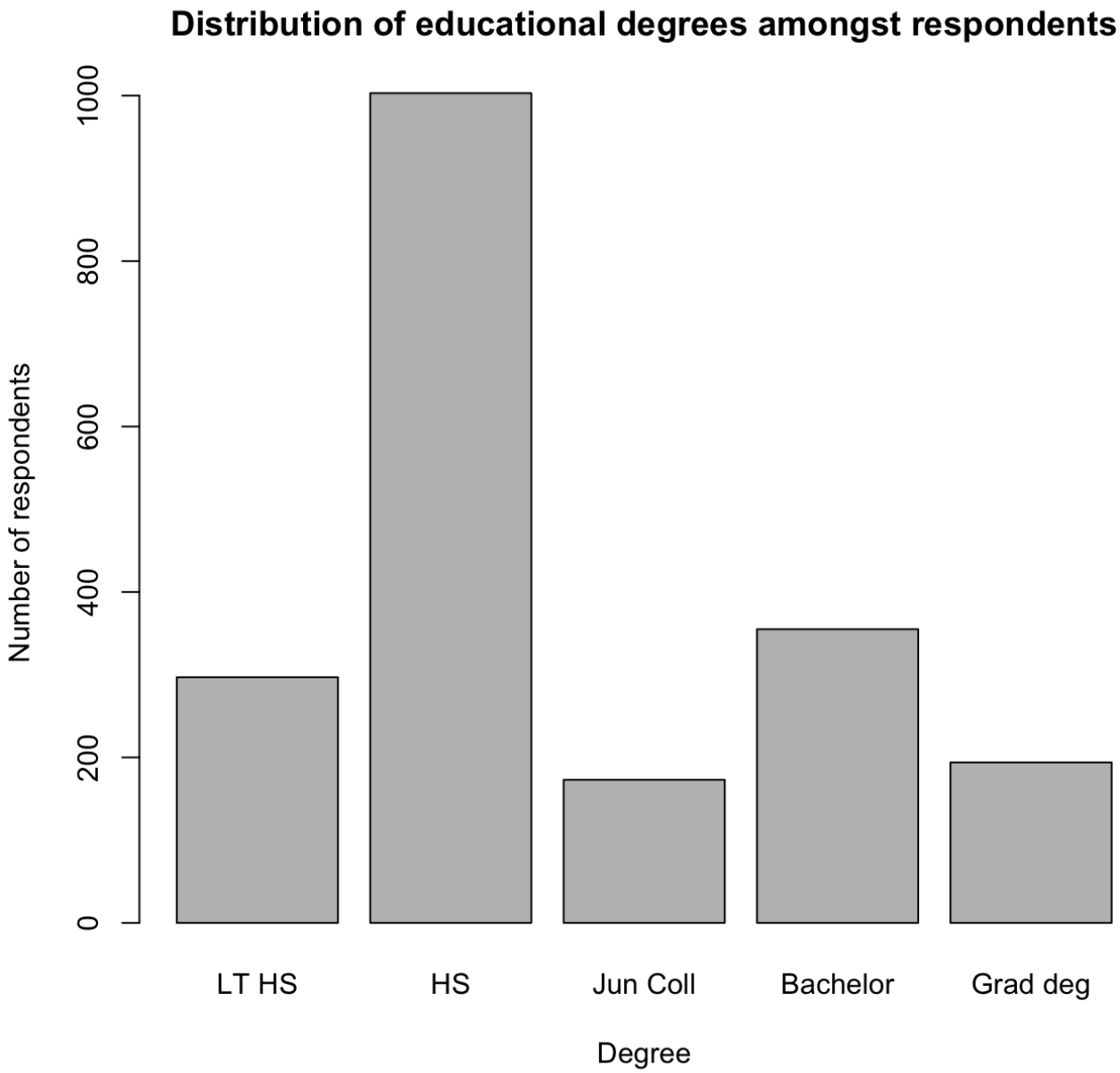
Answer:
The “newgender” variable is nominal/categorical, meaning it has **limited categories with no inherent rank or order.** As a result, we cannot determine its central tendency or dispersion. We can only identify the mode, which is the category with the highest frequency. In this case, the category is Female, as illustrated in the plot.

2.2 Respondents' Educational Attainment (20 points)

1. Describe the variable “newdegree" using a bar plot and include the plot in your response.

Answer:

Similar to Q1, the variable "newdegree" is **a factor** created by converting the data type of the "degree" variable, which was originally a character type. However, this variable contained NA values, which were removed prior to conducting any analysis in R.
The "newdegree" variable is **an ordinal variable** which has five categories: Left High School (LT HS), High School (HS), Junior College (Jun Coll), Bachelor's (Bachelor), and Graduate degree (Grad deg), as illustrated in the bar graph below. The counts for each category are as follows: LT HS - 297, HS - 1003, Jun Coll - 173, Bachelor - 355, and Grad deg - 194, respectively.



2. Report the type of this variable and provide the appropriate measure(s) of central tendency and dispersion.

Answer:
The “newdegree” is a ordinal variable with 5 categories. Ordinal variables have a **limited number of categories with an inherent order** among them**.** However, the **difference** between two consecutive categories is not necessarily equal.
Hence, it’s not possible to find its mean, variance or standard deviation for these variables but we can find the mode, range and the median. The mode for “newdegree” variable is HS at 1003, meaning most of the respondents are from High School. The range of this variable is 1 to 5 and the median is the category, High School.

2.3 Hours per day watching TV (30 points)

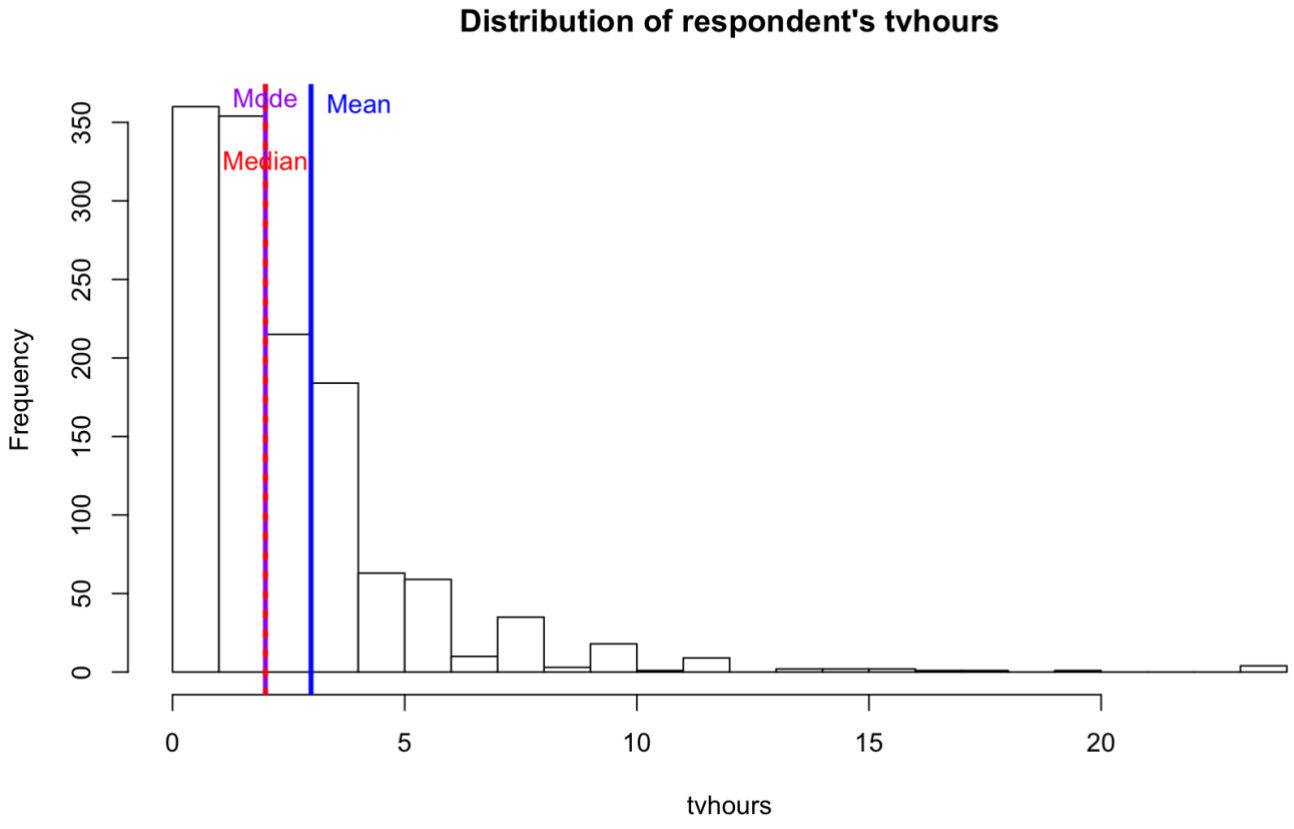
1. Report all relevant measures of central tendency and dispersion for the variable “tvhours."

Answer:
Below are relevant measures of central tendency and dispersion of the variable “tvhours”:

a) Mean: 2.982
b) Mode: 2
c) Median: 2.000
d) Range: 0.000 to 24.000
e) Variance: 7.07
f) Standard Deviation: 2.66

2. Determine whether the distribution of the variable “tvhours" is normal, positively skewed, or negatively skewed. Provide an explanation with a histogram included in your written paper file.

Answer:
The mode and median of the distribution of television viewing hours are both 2, while the mean is 2.982. The distribution is positively skewed, as the mean exceeds both the median and the mode. This indicates that, although there is a higher frequency of lower values, the **presence of outliers (check for higher values of tvhours beyond 20)** is causing the mean to shift towards the right tail of the histogram. Since the mean is calculated by summing all values and dividing by the total number of values, it is particularly sensitive to outliers, giving us a positively skewed distribution on the histogram as shown below.



2.4 Education and Health Care System (40 points)

Imagine you are investigating the connection between individual's educational levels ("newdegree") and their views on the health care system ("newhealth") using *ourdata.RData*. The independent variable (cause) is the educational level, while the dependent variable (outcome) is the opinion on the health care system.

1. Formulate a null hypothesis and an alternative hypothesis.

Answer:
Null hypothesis (H_0): There is no relationship between educational level and opinion on the health care system
Alternative hypothesis (H_A): There is a relationship between educational level and opinion on the health care system

2. Report the p-value. At a significance level of 5%, discuss whether you can reject the null hypothesis and provide reasons.

Answer:
After performing the Chi-squared test on the variables "newdegree" and "newhealth," we obtained a p-value $< 2.2 \times 10^{-16}$, with the test statistic (χ^2) equal to 128.01 and degrees of freedom (df) equal to 12. At the significance level of 5%, the p-value is 0.05. The p-value represents the probability of observing the results assuming that the null hypothesis is true. A lower p-value (<0.05) indicates that we can state with 95% confidence that there is a likelihood of a relationship, rather than this result occurring by random chance.

In this test, the p-value <<< 0.05 indicates that there is a strong evidence against the null leading us to reject the null in favour of the alternative hypothesis.

3. Provide conclusions about the relationship between individuals' education levels and their opinions on the healthcare system based on the chi-square test results.

Answer:
I have taken the liberty to use functions such as chi.residuals, chi.observed, and chi.expected in my code to better understand the reasons behind the statistical significance of the alternative hypothesis and to clearly identify the differences.
The chi-squared test indicates that individuals with lower educational qualifications, such as those who left high school, completed high school, or attended junior college, are more likely to rate the healthcare system as Fair, Good, or Poor, and less likely to rate it as Excellent. In contrast, individuals with higher educational degrees, such as a Bachelor's or Graduate Degree, are more inclined to rate the healthcare system as Excellent or Good, and less likely to rate it as Poor or Fair. This suggests that access to the healthcare system may be limited for those with lower educational qualifications, contributing to their dissatisfaction; conversely, those with higher education may have better access and be more satisfied. Additionally, the highest positive residual value (5.1532211) for the Fair rating among individuals who left high school may indicate that they perceive themselves as "undeserving" of healthcare benefits.
Please find the residuals table on which I am basing my above suggestions:

	ourdata$newdegree (Independent variable)					
ourdata$ newhealth (Dependent variable)		LT HS	HS	Jun Coll	Bachelor	Grad deg
	EXCELLENT	-4.0017456	-0.4016063	-0.4973194	2.9661610	2.2883628
	FAIR	5.1532211	-0.4946629	-0.3869414	-1.8245086	-2.3701793
	GOOD	-2.6762672	1.0511394	0.8643716	-0.2672823	0.4605566
	POOR	6.5647350	-1.3095801	-0.7520817	-2.0424604	-1.6753318