STAT 3010 PYTHON PROJECT

Rebekah Sander

Abstract

This report provides an exploratory data analysis of The General Social Survey (GSS). The data was collected on a sample of 397 US residents and surveyed different key sociological variables.

11/7/2022

Rebekah Sander **STAT 3010 – Python Project** **11/7/22**

Getting to Know the Dataset:

Table 1: Data Dictionary of Analysis Variables of The GSS

VARIABLE	LABEL	GENERAL TYPE	SPECIFIC TYPE	MEASUREMENT UNITS
sex	What is your biological sex?	Categorical	Nominal	Male or Female
race	What is your biological race?	Categorical	Nominal	AfrAm, Hispanic, Other, or White
degree	What is your highest level of education complete?	Categorical	Nominal	NotHS, HighSchool, JunColl, Bachelor, Graduate
relig	What is your religious affiliation?	Categorical	Nominal	Catholic, Jewish, Other, Protestant
polparty	What is your political party affiliation?	Categorical	Nominal	Democrat, Independent, Republican, Other
cappun	Are you in favor or opposed to capital punishment?	Categorical	Nominal	Favor or Oppose
owngun	Do you own a gun?	Categorical	Nominal	Yes or No
gunlaw	Are you in favor or opposed to gun laws?	Categorical	Nominal	Favor or Oppose
tvhours	How many hours of TV do you typically watch in a day?	Quantitative	Discrete	Hours per day
age	How many full years have you been alive?	Quantitative	Discrete	Years
chldidel	What is your ideal number or children?	Quantitative	Discrete	Integer

Univariate Quantitative Analysis:

1.

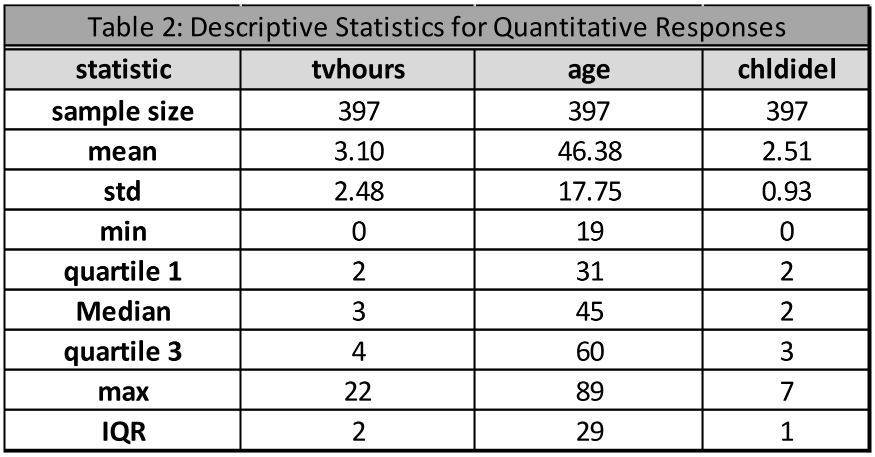
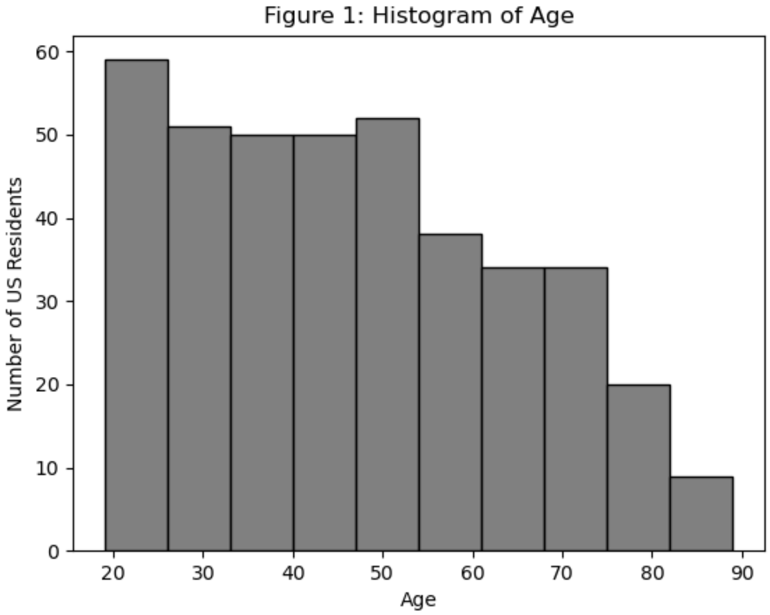


Table 2: Descriptive Statistics for Quantitative Responses

statistic	tvhours	age	chldidel
sample size	397	397	397
mean	3.10	46.38	2.51
std	2.48	17.75	0.93
min	0	19	0
quartile 1	2	31	2
Median	3	45	2
quartile 3	4	60	3
max	22	89	7
IQR	2	29	1

2. **Graphics**

a.



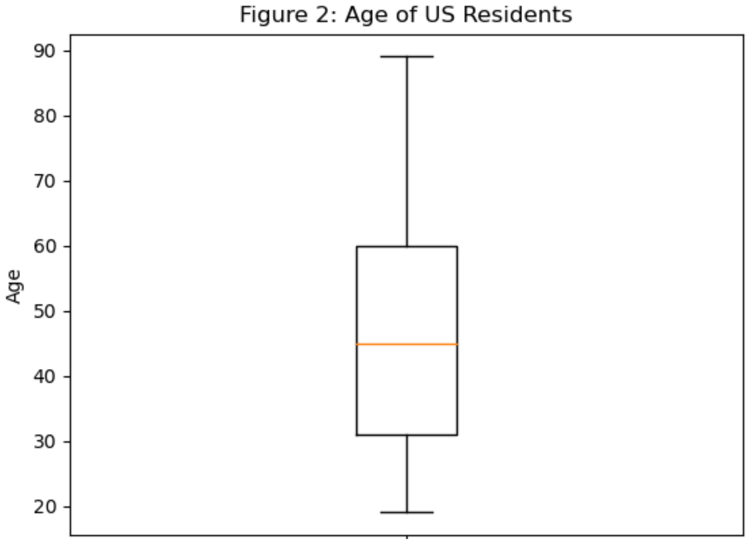


Figure 2: Age of US Residents
Age
90
80
70
60
50
40
30
20

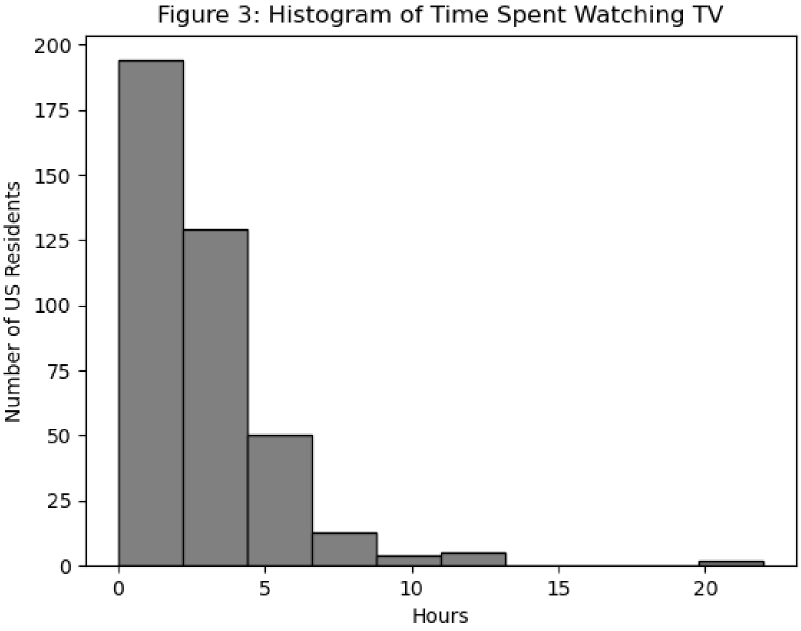


Figure 3: Histogram of Time Spent Watching TV
Number of US Residents
200
175
150
125
100
75
50
25
0
0
5
10
15
20
Hours

b.

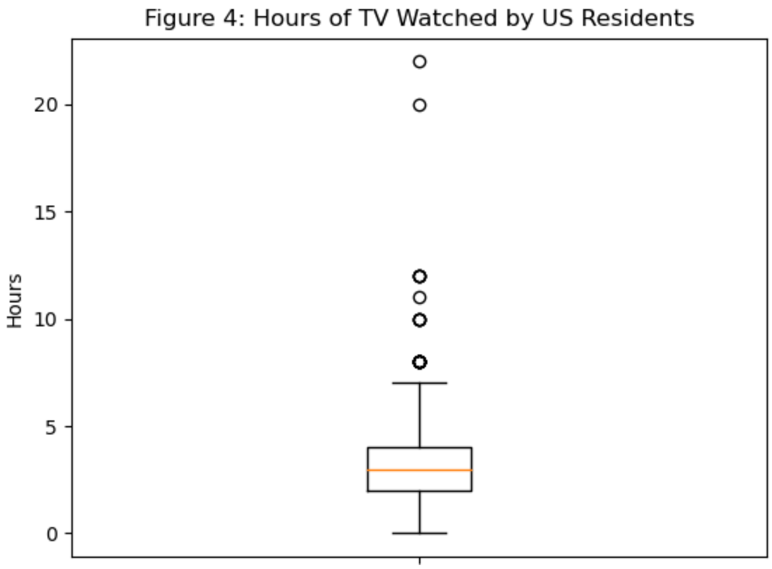


Figure 4: Hours of TV Watched by US Residents
Hours
0
5
10
15
20

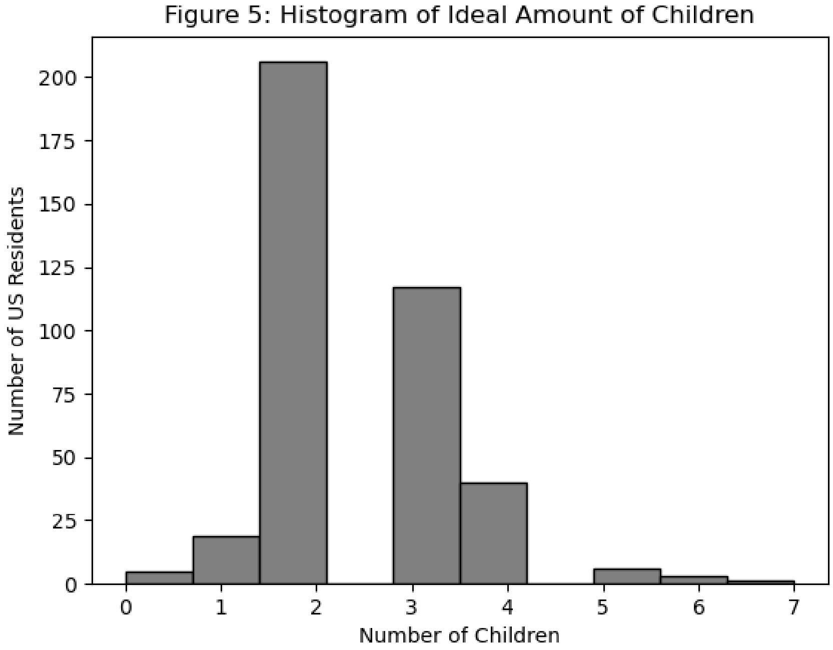
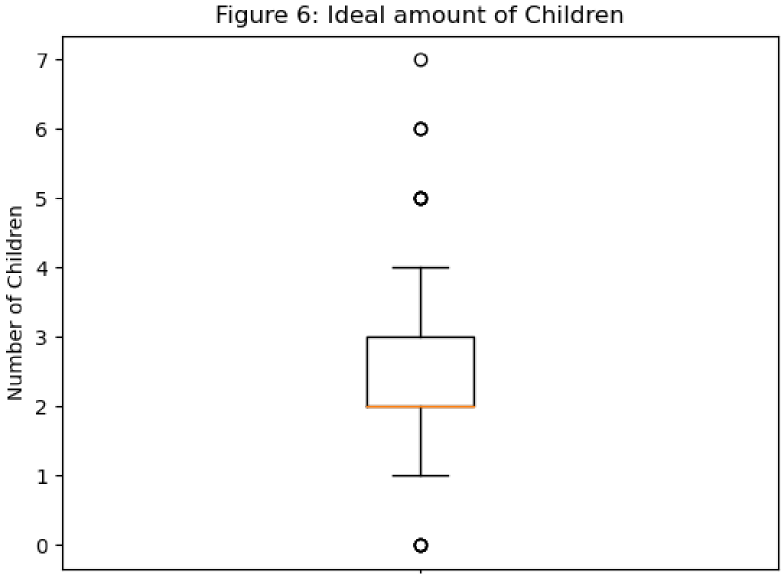


Figure 5: Histogram of Ideal Amount of Children
Number of US Residents
0
25
50
75
100
125
150
175
200
0
1
2
3
4
5
6
7
Number of Children

c.



3. Respectively,

a. For the age of US residents who participated in this survey, the mean is the best representation of central tendency since the data is unimodal and only slightly skewed right. The average age of participants in this survey is 46.38. Because there are no outliers in the age data, the standard deviation of 17.75 is the best representation of dispersion. There are no missing values for this variable.

b. For the daily average time spent watching tv, the median is the best representation of central tendency since the data is skewed right. The median amount of hours participants spent watching tv is 3 hours a day. Because there are multiple outliers in the data, the interquartile range of 2 is the best representation of dispersion. There are outliers of US residents watching on average 10-20 hours of TV per day. There are no values missing for this variable.

c. For the ideal number of children wanted by the GSS participants, the median is the best representation of central tendency since the data is skewed right. The median number of the ideal amount of children was observed to be 2. Because there are outliers in the data, the interquartile range of 1 is the best representation of dispersion. There are outliers occurring at 0 and 5-7. There are no missing values for this variable.

Univariate Categorical Analysis:

4.

a.

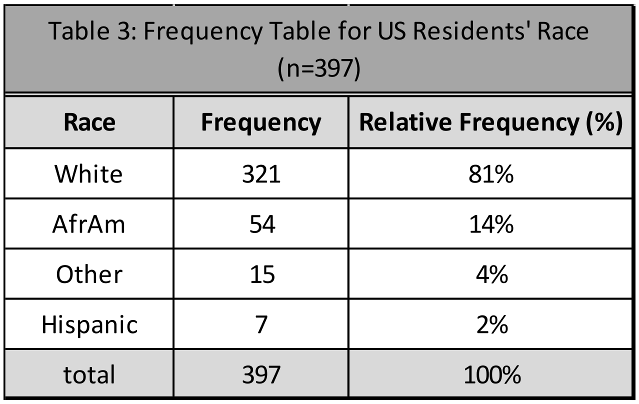


Table 3: Frequency Table for US Residents' Race (n=397)

Race	Frequency	Relative Frequency (%)
White	321	81%
AfrAm	54	14%
Other	15	4%
Hispanic	7	2%
total	397	100%

b.

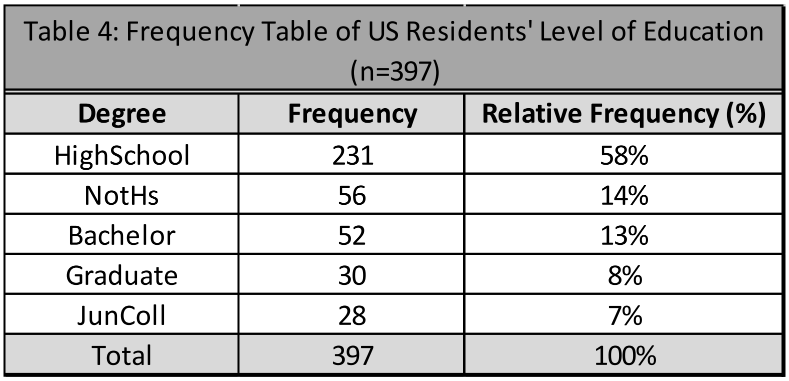


Table 4: Frequency Table of US Residents' Level of Education (n=397)

Degree	Frequency	Relative Frequency (%)
HighSchool	231	58%
NotHs	56	14%
Bachelor	52	13%
Graduate	30	8%
JunColl	28	7%
Total	397	100%

c.

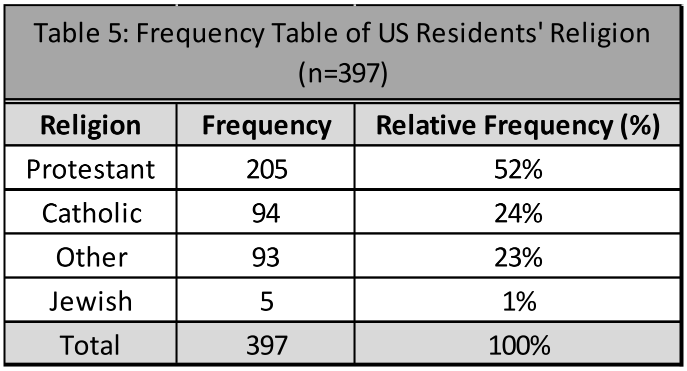
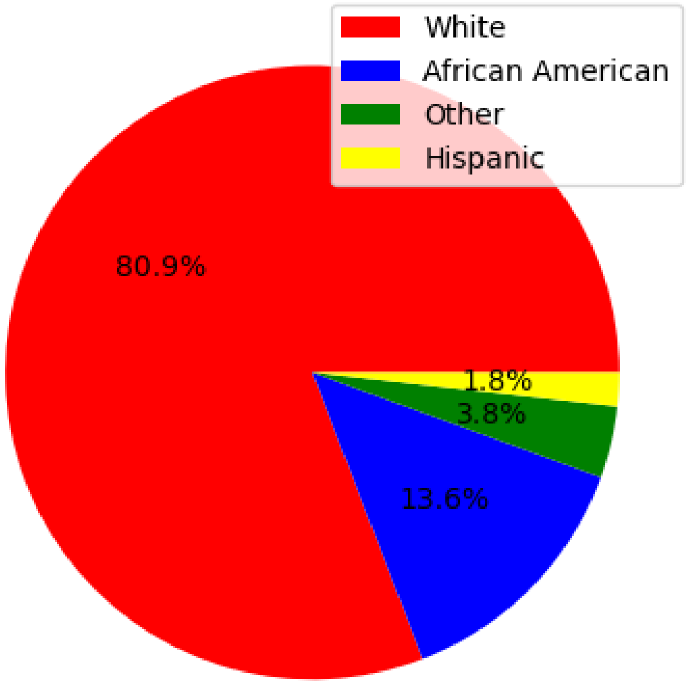


Table 5: Frequency Table of US Residents' Religion (n=397)

Religion	Frequency	Relative Frequency (%)
Protestant	205	52%
Catholic	94	24%
Other	93	23%
Jewish	5	1%
Total	397	100%

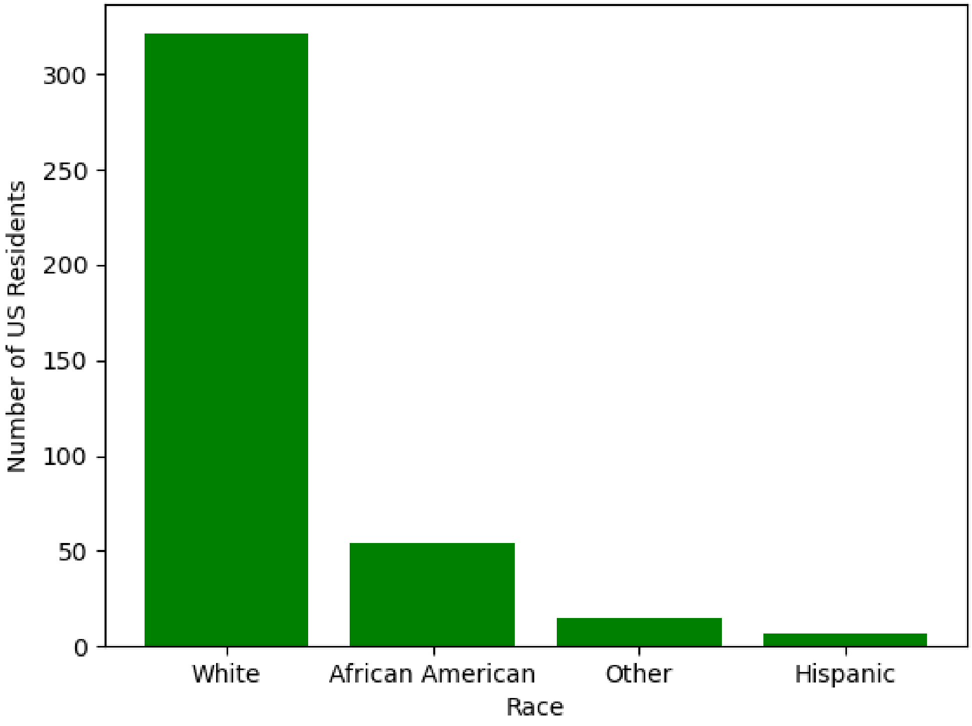
5.

Figure 7: Pie Chart of US Resident Race (n=397)



a.

Figure 8: Bar Chart of US Resident Race (n=397)



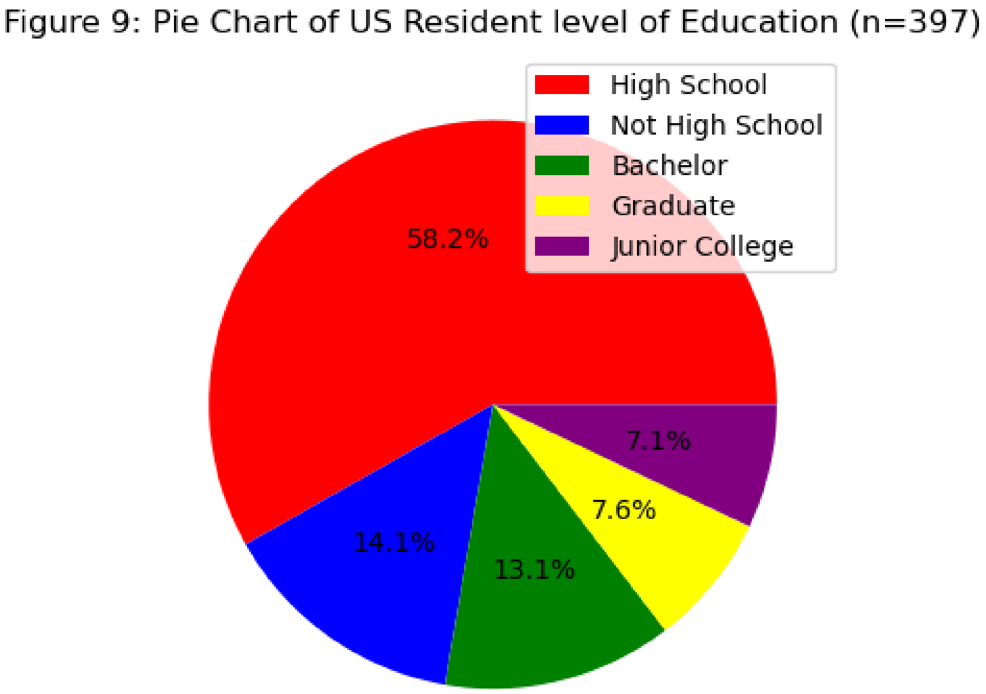


Figure 9: Pie Chart of US Resident level of Education (n=397)

b.

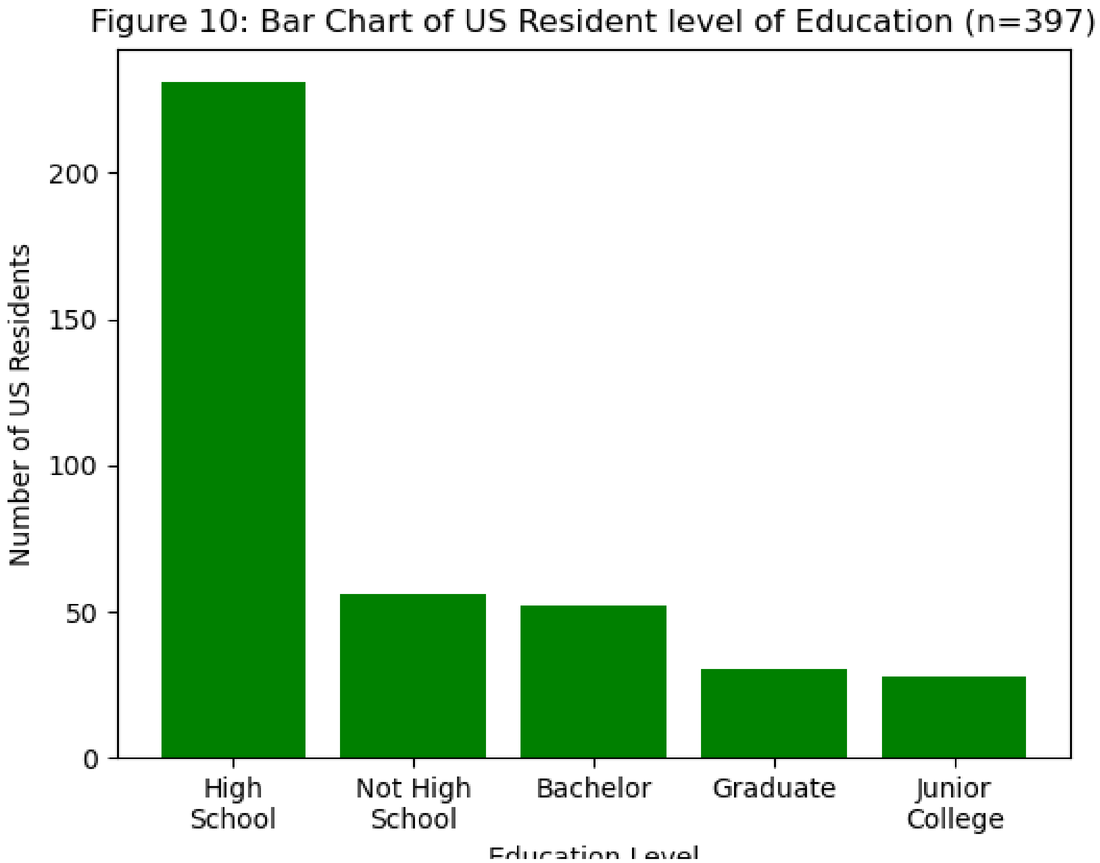
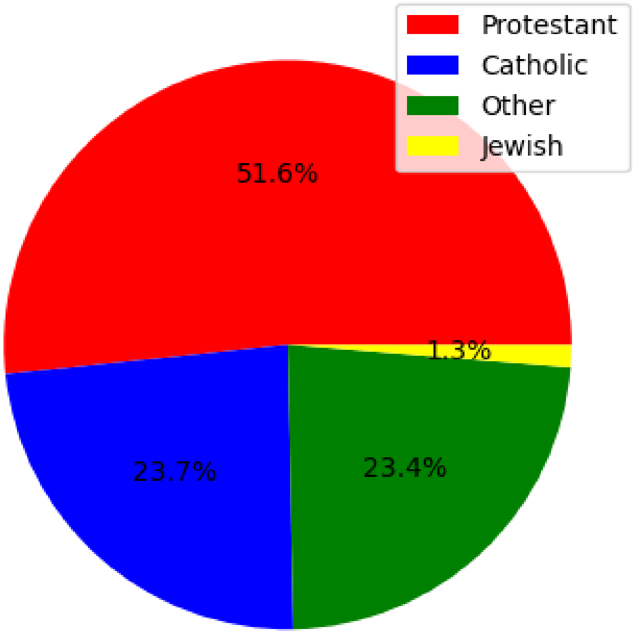


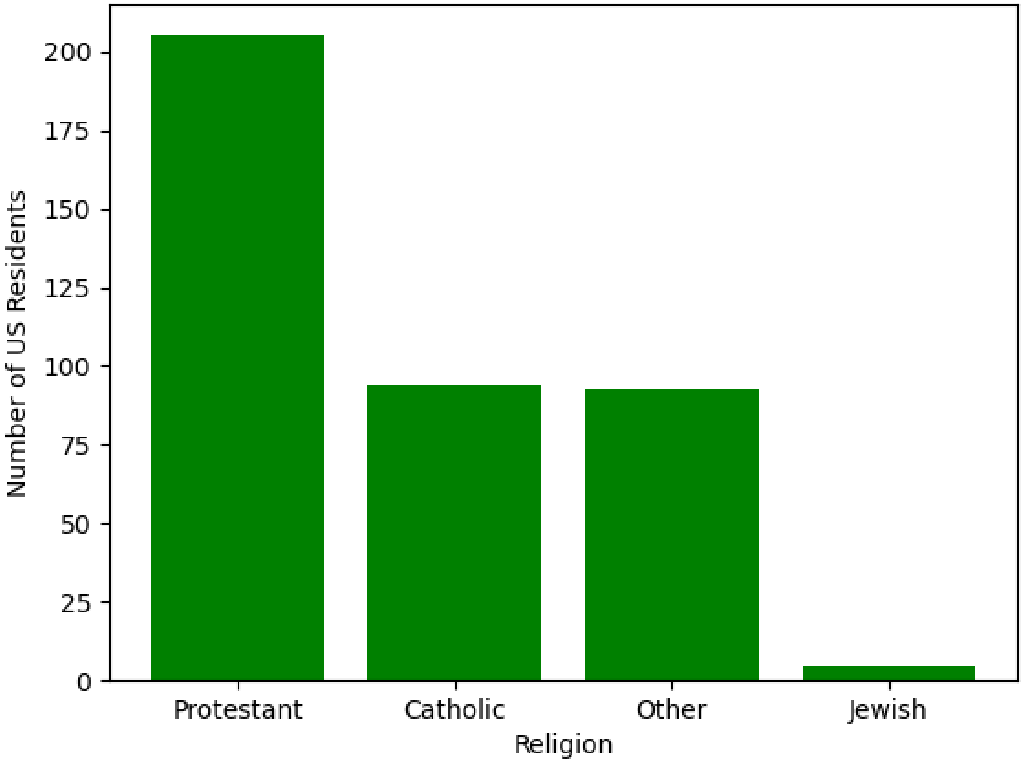
Figure 10: Bar Chart of US Resident level of Education (n=397)

Figure 11: Pie Chart of US Resident Religion (n=397)



c.

Figure 12: Bar Chart of US Resident Religion (n=397)



6. For the US resident participants in the GSS,

a. The distribution of race is unequally distributed. We observe the very significant mode to be 80.9% white.

b. The distribution of the level of education has a mode of 58.2% of residents whose highest degree is high school. Disregarding the mode, the other levels of education were approximately equally distributed.

c. The distribution of religion is unequally distributed and has a mode of 51.6% whom are protestant. Catholic and “other” approximately make up the other half of the data with Jewish being the least represented.

Bivariate Analysis 2 Categorical:

7.

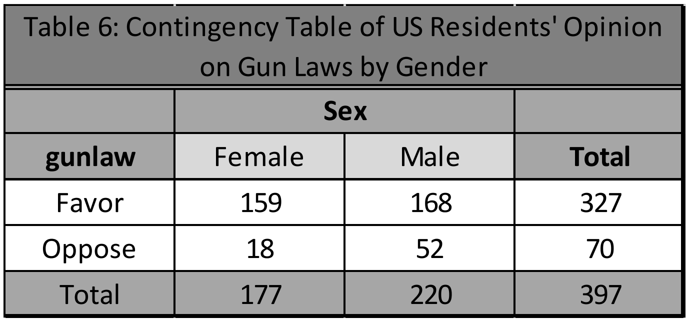


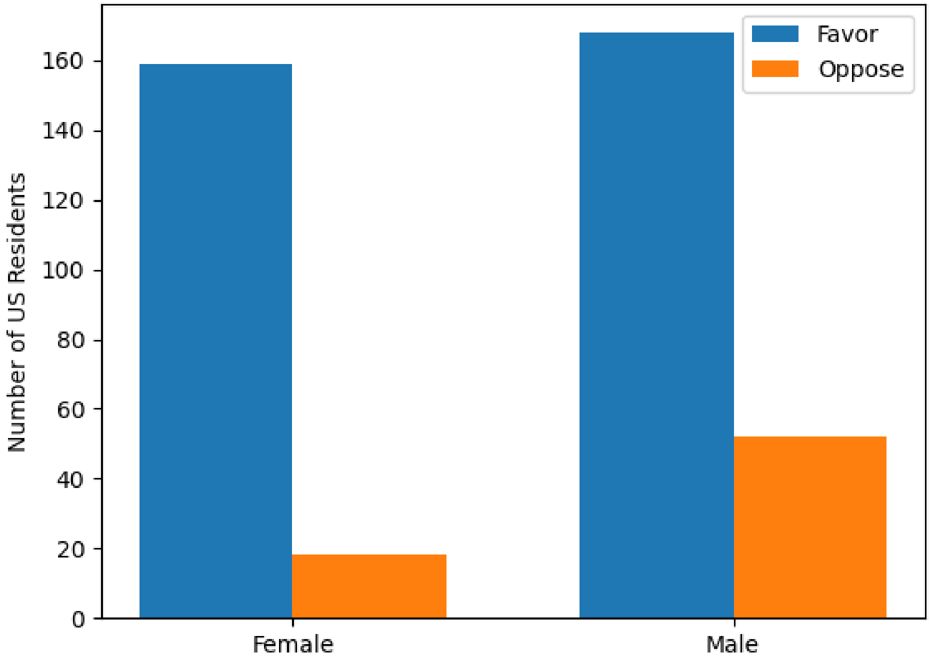
Table 6: Contingency Table of US Residents' Opinion on Gun Laws by Gender

	Sex		
gunlaw	Female	Male	**Total**
Favor	159	168	327
Oppose	18	52	70
Total	177	220	397

8. The survey was made up of 55% male and 45% female US residents. Of those residents, 82% are in favor of gun laws and 18% are opposed. Of those who are female, 90% are in favor and 10% are opposed. Of those who are male, 76% are in favor while 24 are opposed. We can observe an interesting relationship between gender and the opinion on gun laws as it is clear more females are in favor than the males.

9.

Figure 13: Side by Side Bar Chart of US Resident by Gun Law Opinion and Gender(n=397)



a.

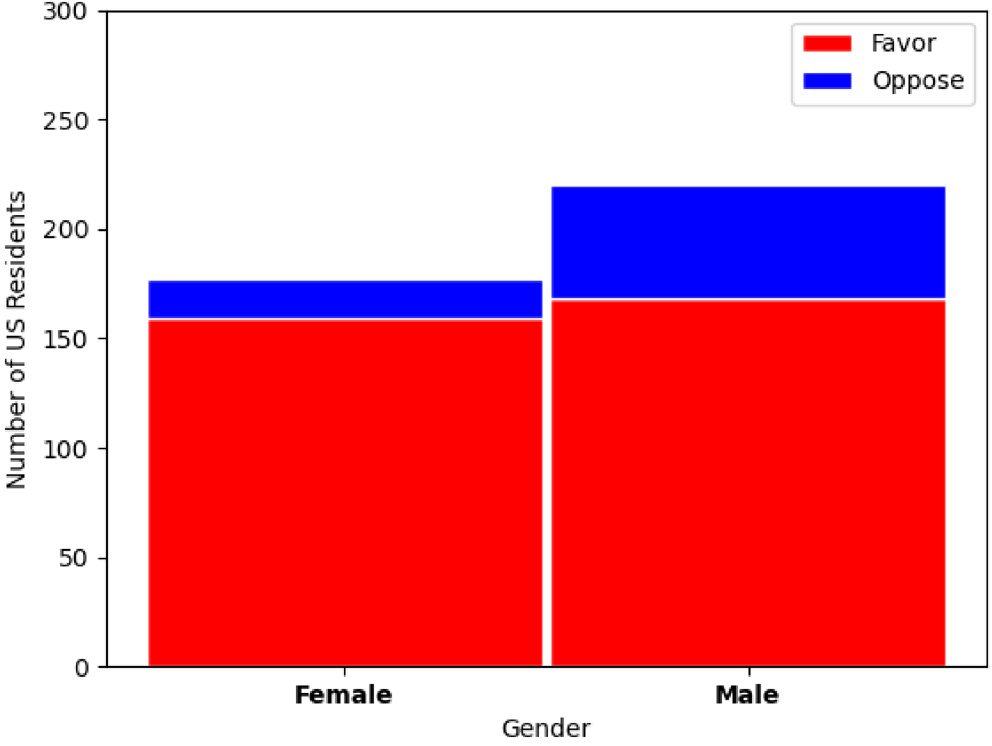
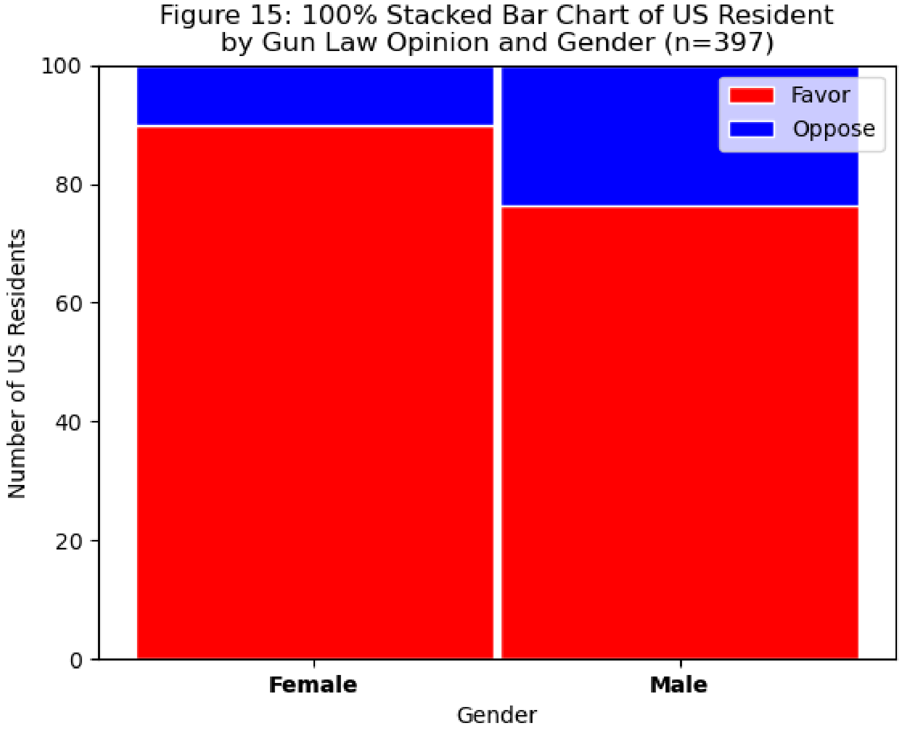


Figure 14: Stacked Bar Chart of US Resident by Gun Law Opinion and Gender (n=397)

b.



c.

10. The grouped side by side bar chart does a good job showing frequencies within a certain category. For example, it is easier to see just how many more females are opposed to gun laws compared to those females in favor of gun laws. This representation though does not do the best at comparing males and females as there are an unequal amount of each. The stacked bar chart does a better job at comparing the amount of females vs males while still easily comparing views on gun laws. Though the stacked bar chart still does not give the full picture. To get the full picture, the 100% stacked bar chart is best as it takes the relative frequencies to make the females and males comparable. It was not until observing the 100% stacked bar chart that we could really see that females are more in favor than males when it comes to gun laws.

Bivariate Analysis 1 Cat 1 Quant:

11.

a.

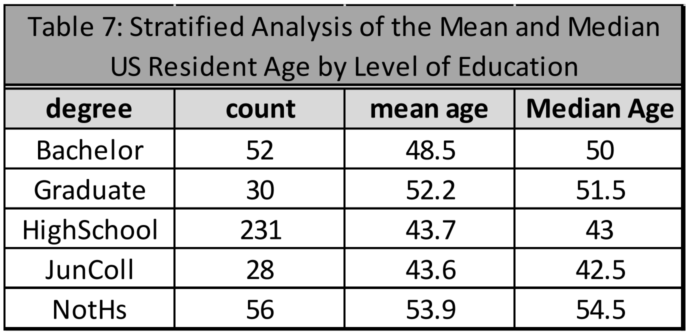


Table 7: Stratified Analysis of the Mean and Median US Resident Age by Level of Education

degree	count	mean age	Median Age
Bachelor	52	48.5	50
Graduate	30	52.2	51.5
HighSchool	231	43.7	43
JunColl	28	43.6	42.5
NotHs	56	53.9	54.5

b.

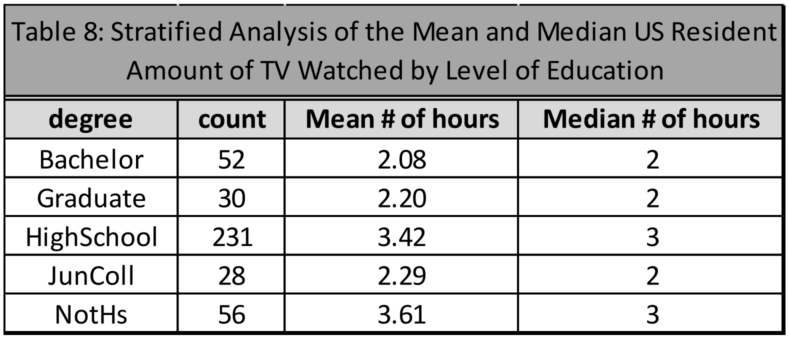


Table 8: Stratified Analysis of the Mean and Median US Resident Amount of TV Watched by Level of Education

degree	count	Mean # of hours	Median # of hours
Bachelor	52	2.08	2
Graduate	30	2.20	2
HighSchool	231	3.42	3
JunColl	28	2.29	2
NotHs	56	3.61	3

c.

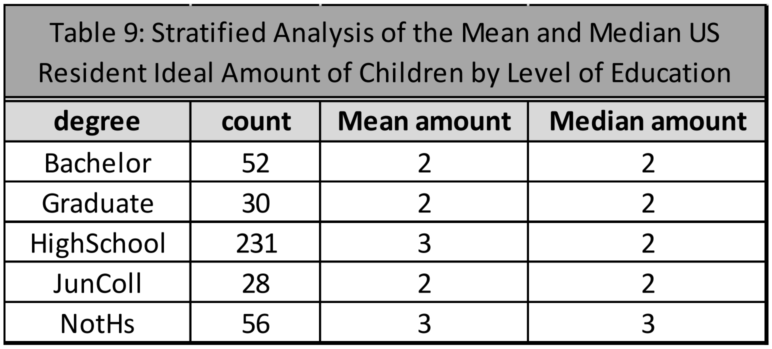
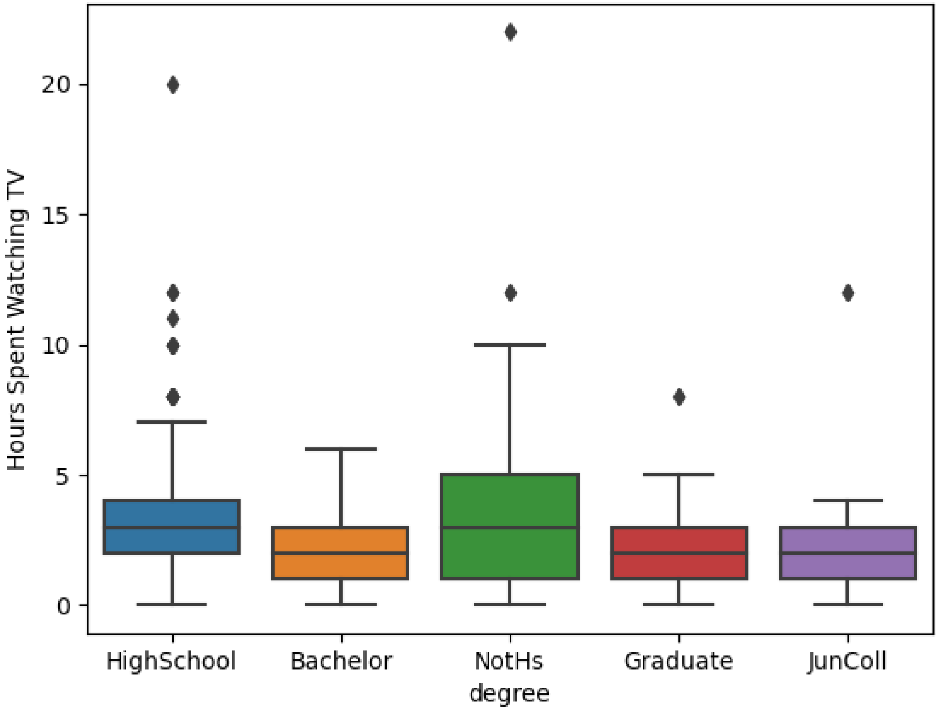


Table 9: Stratified Analysis of the Mean and Median US Resident Ideal Amount of Children by Level of Education

degree	count	Mean amount	Median amount
Bachelor	52	2	2
Graduate	30	2	2
HighSchool	231	3	2
JunColl	28	2	2
NotHs	56	3	3

12.

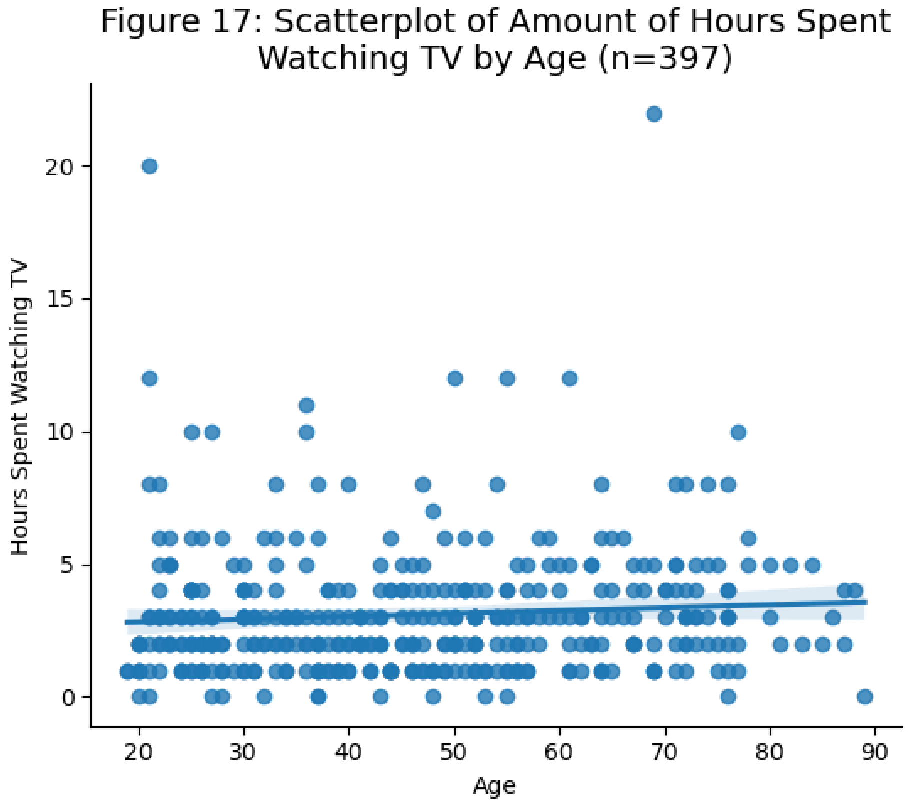
Figure 16: Side by Side Box Plot of the Amount of Hours Spent Watching TV by Level of Education



13. In regards to the daily average time spent watching tv, he participants whose highest level of education is high school is extremely skewed right with outliers at 10-20 hours. We see a similar distribution for the highest level of education being less than a high school degree. Though, those who did not reach a high school degree differ from the other distributions such that the data is more spread out. The participants whose highest level of education were graduate and junior college, have a similar distribution. Both are skewed right due to slight outliers nearing 10 hours. What is interesting to observe, is the distribution of the participants whose highest level of education is a bachelor's degree. Those with a bachelor have no outliers in how many hours of tv they watch, in addition, the distribution is slightly skewed right with the median being less than 5 hours and the max amount being a little more than 5 hours. From the data, we can observe more of the relationship between hours of tv watched and the level of education of a participant. It seems that the higher the degree, the less tv is being watched. Of course, there are some exceptions.

Bivariate Analysis 2 Quant:

14.



15. According to figure 17, there is a strong, slightly positive correlation between the average amount of hours spent watching tv and age of GSS participants. It is clear to see from the figure that how much tv is being watched varies about the same no matter the age. We see a majority of the data watching under 5 hours of tv per day.

Confidence Intervals:

16. Confidence intervals for US Residents' Ideal amount of Children

- **a.** 90% CL [0.99 , 4.04]
- **b.** 95% CL [0.69 , 4.33]
- **c.** 99% CL [0.12 , 4.91]

17. As the confidence level increases, we can observe that the confidence interval becomes bigger. That is, we are more confident about a value being included in a broader interval. The narrower the interval, the less confident we become. For a 95% confidence level, we are 95% confident that the population mean for the ideal number of children is between 0 and 4 children. We must round down since children must be whole numbers.

Variable Creation:

18. Using the age variable, I created a categorical variable named age_cat.

19. Age can be nice as a categorical variable as we can create meaningful intervals. Since the data does not include anyone under the age of 18, I decided to cut the age variable into:

- **a.** 18-24 as "young adults"
- **b.** 25-34 as "middle adults"
- **c.** 35-54 as "older adults"
- **d.** 55+ as "seniors"

20.

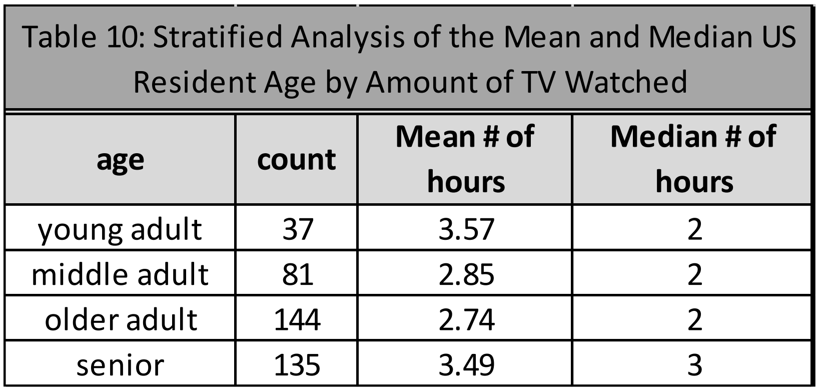


Table 10: Stratified Analysis of the Mean and Median US Resident Age by Amount of TV Watched

age	count	Mean # of hours	Median # of hours
young adult	37	3.57	2
middle adult	81	2.85	2
older adult	144	2.74	2
senior	135	3.49	3

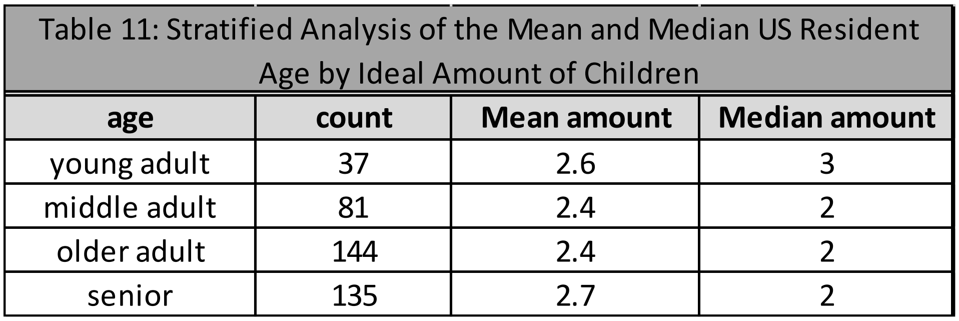


Table 11: Stratified Analysis of the Mean and Median US Resident Age by Ideal Amount of Children

age	count	Mean amount	Median amount
young adult	37	2.6	3
middle adult	81	2.4	2
older adult	144	2.4	2
senior	135	2.7	2