1. Data Description:

The NHANES dataset consists of nominal, ordinal, interval, and ratio data. It includes race, health levels, age, and income as respective examples. It consists of survey data gathered by the US National Center for Health Statistics (NCHS). One point of significance for using the NHANES dataset is that nominal data can be compared with ratio data, for example. Income could be viewed in relation to race, and mental health could be viewed through one's level of education. I am curious about how values change over time and therefore am selecting a subset of individuals of Gen Zs, Millennials, and GenXs. Therefore, I can understand how values shift over time by looking at the differences between specific age ranges for generations within the population, and how they differ in terms of exercise, age of sex, smoking, drinking, education, etc. I'm curious if such things change throughout the generations, perhaps indicative of social change and an overall shift from traditional values to what is considered more "progressive." In addition, this sample can be used to help understand people's relationships with technology. For example, how many hours do older generations spend on the computer vs newer generations? This subset can also be used to understand people's attitudes towards physical exercise, for example, earlier generations may have been healthier, in terms of physical and mental health. By using the subset of three consecutive generations, we can begin to uncover some answers. Beyond generational differences in values around smoking and sex, I'm curious about how different factors influence health, for example, whether or not a high income will add or detract from one's overall well-being. There is some bidirectional ambiguity to be noted here, as good health may help generate higher income, or higher income may allow someone resources for better physical health. I will apply this question to mental and physical health. For my three samples, Generation X has 2182 observations, Generation Z has 2193 observations, and millennials have 2164 observations, therefore, there is a relatively equal split in the amount of data coming from these three populations. NHANES has 76 variables, pictured below.

```
 [1] "ID"               "SurveyYr"         "Gender"           "Age"              "AgeDecade"        "AgeMonths"
 [7] "Race1"            "Race3"            "Education"        "MaritalStatus"    "HHIncome"         "HHIncomeMid"
[13] "Poverty"          "HomeRooms"        "HomeOwn"          "Work"             "Weight"           "Length"
[19] "HeadCirc"         "Height"           "BMI"              "BMICatUnder20yrs" "BMI_WHO"          "Pulse"
[25] "BPSysAve"         "BPDiaAve"         "BPSys1"           "BPDia1"           "BPSys2"           "BPDia2"
[31] "BPSys3"           "BPDia3"           "Testosterone"     "DirectChol"       "TotChol"          "UrineVol1"
[37] "UrineFlow1"       "UrineVol2"        "UrineFlow2"       "Diabetes"         "DiabetesAge"      "HealthGen"
[43] "DaysPhysHlthBad"  "DaysMentHlthBad"  "LittleInterest"   "Depressed"        "nPregnancies"     "nBabies"
[49] "Age1stBaby"       "SleepHrsNight"    "SleepTrouble"     "PhysActive"       "PhysActiveDays"   "TVHrsDay"
[55] "CompHrsDay"       "TVHrsDayChild"    "CompHrsDayChild"  "Alcohol12PlusYr"  "AlcoholDay"       "AlcoholYear"
[61] "SmokeNow"         "Smoke100"         "Smoke100n"        "SmokeAge"         "Marijuana"        "AgeFirstMarij"
[67] "RegularMarij"     "AgeRegMarij"      "HardDrugs"        "SexEver"          "SexAge"           "SexNumPartnLife"
[73] "SexNumPartYear"   "SameSex"          "SexOrientation"   "PregnantNow"
```

2. Research Questions:

- How much do different generations sleep on average?
- Is there a relationship between income and health between generations?
- Is there a difference in the relationship between income and health between Generation Z, millennials, and Generation X if a statistical test such as the t-test or ANOVA were to be applied?
- How does the age of first smoking differ between Generation X, millennials, and Generation Z?
- Is there a significant difference in Sex Age between Generation X, Millennials, and Generation Z?
- Does the mean sex age between Generation X, millennials, and Generation Z differ?

3. Analysis:

Is are the mean sex ages between Generation X, millennials, and Generation Z?

1. Objective:

   I will calculate the means of the sex age for 3 generations of Americans (Gen Z, Millennial, GenX).

2. R Code:

```r
# Load in needed libraries
library(NHANES)
library(dplyr)

data(NHANES)

# Create an NHANES dataframe

nhanes_df <- as.data.frame(NHANES)

# Create age bounds so that each generation can be analyzed seperately

age_lower <- 11
age_upper <- 26

# Filter the data by the generation of interest

filtered_data <- nhanes_df %>% filter(SexAge >= age_lower & SexAge <= age_upper)

# Calculate the mean of this generation (Generation Z for the variable of sex age)

genzMean <- mean(filtered_data$SexAge, na.rm = TRUE)

# Print the calculated mean

print(genzMean)

# Repeat this process for the two other age bounds for Generation X and Millenhials who I will compare to each other

age_lower <- 27
age_upper <- 42

filtered_data <- nhanes_df %>% filter(SexAge >= age_lower & SexAge <= age_upper)

millennialMean <- mean(filtered_data$SexAge, na.rm = TRUE)

print(millennialMean)

age_lower <- 43
age_upper <- 58

filtered_data <- nhanes_df %>% filter(SexAge >= 43 & SexAge <= 58)

genxMean <- mean(filtered_data$SexAge, na.rm = TRUE)

print(genxMean)
```

3. Output:

```
[1] 17.17898
>
> age_lower <- 27
> age_upper <- 42
>
> # Filter the data for individuals with age less than 25
> filtered_data <- nhanes_df %>% filter(SexAge >= age_lower & SexAge <= age_upper)
>
> # Calculate the mean of the specified column for the filtered data
> millenialMean <- mean(filtered_data$SexAge, na.rm = TRUE)
>
> # Print the mean value
> print(millenialMean)
[1] 30.53285
>
> age_lower <- 43
> age_upper <- 58
>
> # Filter the data for individuals with age less than 25
> filtered_data <- nhanes_df %>% filter(SexAge >= age_lower & SexAge <= age_upper)
>
> # Calculate the mean of the specified column for the filtered data
> genxMean <- mean(filtered_data$SexAge, na.rm = TRUE)
>
> # Print the mean value
> print(genxMean)
[1] 46
```

4. Interpretation:

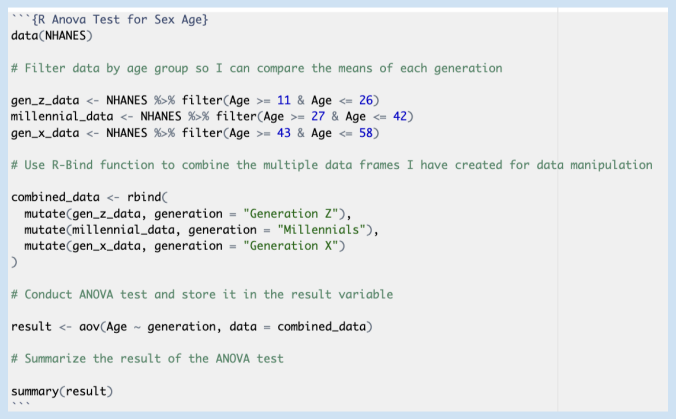
These results reflect drastically different ages of sex between different American generations. It appears that there is a negative correlation between time passing, in other words, the generation that one is born into and the age at which they have sex. This could be reflective of changing values in society towards sex, and the normative age at which one engages in sexual activity. Shown in the calculated means is a negative correlation between age of sex and time (shown by the different individual generations).

Is there a significant difference in Sex Age between Generation X, Millennials, and Generation Z?

1. Objective:

I will run an ANOVA test to compare the sex age between Generations

2. R Code:



```
```{R Anova Test for Sex Age}
data(NHANES)

# Filter data by age group so I can compare the means of each generation

gen_z_data <- NHANES %>% filter(Age >= 11 & Age <= 26)
millennial_data <- NHANES %>% filter(Age >= 27 & Age <= 42)
gen_x_data <- NHANES %>% filter(Age >= 43 & Age <= 58)

# Use R-Bind function to combine the multiple data frames I have created for data manipulation

combined_data <- rbind(
  mutate(gen_z_data, generation = "Generation Z"),
  mutate(millennial_data, generation = "Millennials"),
  mutate(gen_x_data, generation = "Generation X")
)

# Conduct ANOVA test and store it in the result variable

result <- aov(Age ~ generation, data = combined_data)

# Summarize the result of the ANOVA test

summary(result)
```
```

3. Output:

```
               Df  Sum Sq Mean Sq F value Pr(>F)
generation      2 1111103  555552   26021 <2e-16 ***
Residuals    6536  139546      21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Interpretation of findings

This ANOVA test produced an incredibly small p-value of <2e^-16 which converts to 0.0000000000000002. This means that there is statistical significance between the three samples from Generation X, Millennials, and Generation Z.

How does the age of first smoking differ between Generation X, millennials, and Generation Z?

1. Objective:

I will calculate the means of the smoking age for 3 generations of Americans (Gen Z, Millennial, GenX).

I will also conduct an ANOVA test to compare the smoking age between the three generations that make up my subsets

2. R Code:

```r
```{r Smoking and Generation}

# Create a data frame containing NHANES data to be later manipulated

nhanes_df <- as.data.frame(NHANES)
print(NHANES$SmokeAge)

# Store upper and lower bounds for each generation in variables to be later used to sort by generation

xLower <- 11
xUpper <- 26

# Filter the NHANES data using a pipe operator to only look at people who's ages fall in a particular range

filtered_data <- nhanes_df %>% filter(SmokeAge >= xLower & SmokeAge <= xUpper)

# Calculate the mean of smoking age for the particular generation that I'm observing using the mean() operator

genzMean <- mean(filtered_data$SmokeAge, na.rm = TRUE)

# This process is then repeated for the two other generations that are under observation.

zLower <- 27
zUpper <- 42

filtered_data <- nhanes_df %>% filter(SmokeAge >= zLower & SmokeAge <= zUpper)

millennialMean <- mean(filtered_data$SmokeAge, na.rm = TRUE)

mLower <- 43
mUpper <- 58

filtered_data <- nhanes_df %>% filter(SmokeAge >= mLower & SmokeAge <= mUpper)

genxMean <- mean(filtered_data$SexAge, na.rm = TRUE)

# Load in required packages for statistical tests

library(tibble)
library(dplyr)

print(genzMean)
print(millennialMean)
print(genxMean)


# Use pipe operators to establish the age ranges for each generation which will be used in the ANOVA test

gen_z_data <- NHANES %>% filter(Age >= 11 & Age <= 26)
millennial_data <- NHANES %>% filter(Age >= 27 & Age <= 42)
gen_x_data <- NHANES %>% filter(Age >= 43 & Age <= 58)

# Combine data using the rbind function, which will group the information from each generation into one variable to be observed.

combined_data <- rbind(
  mutate(gen_z_data, generation = "Generation Z"),
  mutate(millennial_data, generation = "Millennials"),
  mutate(gen_x_data, generation = "Generation X")
)


# Conduct an ANOVA test using the mean_data tibble which holds data on the means of the three generations I'm observing. This value is then stored in result variable

result <- aov(SmokeAge ~ generation, data = combined_data)

# Summarize the results of the ANOVA test

summary(result)
```

3. Output: ANOVA test and calculated mean values for smoking age between Generation X, Millennials, and Generation Z.

```
> print(genzMean)
[1] 17.1335
> print(millennialMean)
[1] 31.54015
> print(genxMean)
[1] 16.15385
>
> gen_z_data <- NHANES %>% filter(Age >= 11 & Age <= 26)
> millennial_data <- NHANES %>% filter(Age >= 27 & Age <= 42)
> gen_x_data <- NHANES %>% filter(Age >= 43 & Age <= 58)
>
>
> combined_data <- rbind(
+   mutate(gen_z_data, generation = "Generation Z"),
+   mutate(millennial_data, generation = "Millennials"),
+   mutate(gen_x_data, generation = "Generation X")
+ )
>
> result <- aov(SmokeAge ~ generation, data = combined_data)
>
> summary(result)
             Df Sum Sq Mean Sq F value   Pr(>F)
generation    2    934   466.8   19.69 3.38e-09 ***
Residuals  2130  50508    23.7
```

4. Interpretation:

These results are interesting as they reflect the “pendulum” nature of trends. For example, in fashion, what is in style, might be in fashion at one point, then go out of fashion, then come back into style in future years, as is true with the Y2K style. This goes for music, aesthetics, clothing, etc. I think these results reflect that while smoking was popular with Generation X, it went out of fashion for millennials, perhaps due to anti-smoking campaigns or perhaps because of people deeming it less socially acceptable. Subsequently, it appears that Generation Z has a much lower smoking age, indicative of a similar popularity of smoking early to that of Generation X. After running an ANOVA test, it was also evident that the data can be deemed significant as the calculated p-value was 0.00000000338, meaning that there is an incredibly small chance of a type 1 error, that is, incorrectly rejecting the null hypothesis that states no significant difference between generation and smoking.
```

Is there a relationship between income and health between generations? and

Is there a difference in the relationship between income and health between Generation Z, millennials, and Generation X?

1. Objective:

I will calculate the correlation between income and sleep between Generation X, millennials, and Generation Z.

2. R Code: 3. Output: Display the results from your R code (e.g., R output, or a graph).

```r
```{r Income vs Sleep}

# Load in necessary packages to conduct my analysis

library(ggplot2)
library(NHANES)

data(NHANES)

# Use a pipe operator to search the NHANES data set for people who's ages fall in my desired range for analysis

filtered_data <- NHANES %>% filter(Age <= 26 & Age >= 11)

# Select my variables of interest (in this case I'm only going to look at Income and Sleep)

selected_data <- filtered_data[, c("HHIncomeMid", "BPSysAve", "BPDiaAve", "SleepHrsNight")]

# The na.omit function will ignore data who's values are null, therefore, I'll only being looking at existing data

selected_data <- na.omit(selected_data)

# Calculate a simple correlation between income and sleep for my specific generation

correlation <- cor(selected_data$HHIncomeMid, selected_data$SleepHrsNight)

# Print this correlation

print(correlation)

# Use a pipe operator to search the NHANES data set for people who's ages fall in my desired range for analysis

# The same steps will be repeated for Millennials and Generation X

filtered_data <- NHANES %>% filter(Age >= 27 & Age <= 42)

selected_data <- filtered_data[, c("HHIncomeMid", "BPSysAve", "BPDiaAve", "SleepHrsNight")]

selected_data <- na.omit(selected_data)

correlation <- cor(selected_data$HHIncomeMid, selected_data$SleepHrsNight)

print(correlation)

# Use a pipe operator to search the NHANES data set for people who's ages fall in my desired range for analysis

filtered_data <- NHANES %>% filter(Age <= 58 & Age >= 43)

selected_data <- filtered_data[, c("HHIncomeMid", "BPSysAve", "BPDiaAve", "SleepHrsNight")]

selected_data <- na.omit(selected_data)

correlation <- cor(selected_data$HHIncomeMid, selected_data$SleepHrsNight)

print(correlation)
```
```

```
[1] 0.07666751
[1] 0.0558749
[1] 0.1759847
```

(The generations move chronologically forward from bottom to top so the bottom statistic is for Generation X and the top statistic is for Generation Z)

3. Interpretation:

Through the data looking at income and health between 3 generations, it appears that while earlier generations seemed to prioritize work over other departments in their life, newer generations may have access to information concerning health that influences their behavior. This can be seen in the high positive correlation between income and physical health seen through the amount of sleep one gets. It appears that as modern science develops, and newer generations continue to place value on physical wellbeing, the more money one makes, the more they prioritize their health, vs earlier generations, where it can be seen that there isn't as strong of a correlation between income and wellbeing.

How much do different generations sleep on average?

1. Objective: Clearly state your goal (e.g., “I will calculate the mean age for physically active adults”)

   I will calculate the mean hours of sleep for Generation X, millennials, and Generation Z.

2. R Code: (following the same logic as calculating the correlations in the previous section)

```r
```{r Generation vs. Sleep }
library(ggplot2)
library(NHANES)

data(NHANES)

# Filter the NHANES data to only observe Generation Z using a pipe operator and filter function

filtered_data <- NHANES %>% filter(Age <= 26 & Age >= 11)

# Use filtered_data variable to gather data from the SleepHrsNight column to be analyzed

selected_data <- filtered_data[, c("SleepHrsNight")]

# Omit rows with missing values

selected_data <- na.omit(selected_data)

# Calculate the mean of selected data and store it into the mean variable.

mean <- mean(selected_data$SleepHrsNight)

# Print the calculated mean

print(mean)

# Filter the NHANES data to only observe the Millennial generation using a pipe operator and filter function

filtered_data <- NHANES %>% filter(Age >= 27 & Age <= 42)

# Use filtered_data variable to gather data from the SleepHrsNight column to be analyzed

selected_data <- filtered_data[, c("SleepHrsNight")]

# Omit rows with missing values

selected_data <- na.omit(selected_data)

# Calculate the mean of selected data and store it into the mean variable.

mean <- mean(selected_data$SleepHrsNight)

# Print the calculated mean

print(mean)

# Filter the NHANES data to only observe the Generation X using a pipe operator and filter function

filtered_data <- NHANES %>% filter(Age <= 58 & Age >= 43)

# Use filtered_data variable to gather data from the SleepHrsNight column to be analyzed

selected_data <- filtered_data[, c("SleepHrsNight")]

# Omit rows with missing values

selected_data <- na.omit(selected_data)

# Calculate the mean of selected data and store it into the mean variable.

mean <- mean(selected_data$SleepHrsNight)

# Print the calculated mean

print(mean)
```
```

3. Output: Means of Each generation (Top value is generation Z, middle value is the Millenial generation, bottom value is Generation X)

```
[1] 7.134655
[1] 6.882407
[1] 6.715204
```

4. Interpretation:

As seen here, the average or mean sleep has increased from Generation X to Generation Z. This goes along with the hypothesized value of health that has become more prevalent in current generations versus earlier ones. This could be attributed to the advent of the internet, as well as modern science where we have sound reasoning behind why sleep is important. Because we have access to such information, more and more people are likely prioritizing and emphasizing the amount of sleep that they get, which provides reasoning for the steady increase in the amount of sleep between Generation X, millennials, and Generation Z.

1. Objective: Clearly state your goal (e.g., “I will calculate the mean age for physically active adults”)

I will run an ANOVA test to compare the relationship between hours of sleep, one’s income, and which generation one belongs to. I’m doing this to assess whether values and culture around work ethic changes between consecutive generations, that is if people tend to sleep more and make more money in more recent generations, as opposed to older generations who have a stronger philosophy around working hard and sleeping less.

2. R Code:

```r
```{r t-test and anova}

# Load in required packages for t-test and ANOVA test

library(ggplot2)
library(NHANES)
library(dplyr)

data(NHANES)

# use a pipe operator and mutate function to filter the data by the three generations I'm interested in analyzing

NHANES <- NHANES %>%
  mutate(generation = case_when(
    Age >= 11 & Age <= 26 ~ "Generation X",
    Age >= 27 & Age <= 42 ~ "Millennials",
    Age >= 43 & Age <= 58 ~ "Generation Z",
    TRUE ~ "Other"
  ))

# na.omit function is used to handle missing values so that my statistical test does not produce any errors due to null data values / missing data values

gen_x_data <- NHANES %>% filter(generation == "Generation X")
gen_x_sleep <- na.omit(gen_x_data$SleepHrsNight)

millennials_data <- NHANES %>% filter(generation == "Millennials")
millennials_sleep <- na.omit(millennials_data$SleepHrsNight)

gen_z_data <- NHANES %>% filter(generation == "Generation Z")
gen_z_sleep <- na.omit(gen_z_data$SleepHrsNight)

# Mean function is applied to calculate the means for each generation's sleep

mean_gen_x <- mean(gen_x_sleep)
mean_millennials <- mean(millennials_sleep)
mean_gen_z <- mean(gen_z_sleep)

# Cat function is used to concatenate some text with the actual means calculate to make the presentation of my results more accessible

cat("Mean Sleep Hours (Generation X):", mean_gen_x, "\n")
cat("Mean Sleep Hours (Millennials):", mean_millennials, "\n")
cat("Mean Sleep Hours (Generation Z):", mean_gen_z, "\n")

# Conducting t-tests to compare the means of each generation's sleep. Since t-tests compare two means, I've conducted this multiple times as there are three subsets of interest.

t_test_x_m <- t.test(gen_x_sleep, millennials_sleep)
t_test_m_z <- t.test(millennials_sleep, gen_z_sleep)
t_test_x_z <- t.test(gen_x_sleep, gen_z_sleep)

print("Results from t-test:")
print(t_test_x_m)
print(t_test_m_z)
print(t_test_x_z)

# Here, I conducted an anova test to get a broader picture of statistical significance of the relationship bewteen sleep and which generation one comes from

anova_result <- aov(SleepHrsNight ~ generation, data = NHANES)

# Summary of the ANOVA results

summary(anova_result)
```
```

3. Output:

```
Mean Sleep Hours (Generation X): 7.134655
Mean Sleep Hours (Millennials): 6.882407
Mean Sleep Hours (Generation Z): 6.715204
[1] "T-Test Results:"

        Welch Two Sample t-test

data:  gen_x_sleep and millennials_sleep
t = 5.5214, df = 2862, p-value = 3.665e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1626674 0.3418275
sample estimates:
mean of x mean of y
 7.134655  6.882407


        Welch Two Sample t-test

data:  millennials_sleep and gen_z_sleep
t = 4.3024, df = 4325.7, p-value = 1.727e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.09101298 0.24339302
sample estimates:
mean of x mean of y
 6.882407  6.715204


        Welch Two Sample t-test

data:  gen_x_sleep and gen_z_sleep
t = 9.0183, df = 2974.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3282536 0.5106472
sample estimates:
mean of x mean of y
 7.134655  6.715204

              Df Sum Sq Mean Sq F value Pr(>F)
generation     3    199   66.34   37.09 <2e-16 ***
Residuals   7751  13864    1.79
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2245 observations deleted due to missingness
```

4. Interpretation:

As seen here, the p-values produced by each of the three statistical tests are less than my chosen alpha value of 0.05. Therefore, we can conclude that there is a significant difference between the mean amount of sleep for Generation X, Millennials, and Generation Z. In addition, the ANOVA test also produced a p-value less than 0.05, therefore, there is an incredibly low chance that observed differences are due to chance, but rather, they are due to some form of correlation between the generation one belongs to, and how much sleep they get.

4. Conclusion

The findings gathered in this statistical analysis show the multifaceted ways through which the values of subsequent generations diverge in terms of behaviors, morals, and work ethic. In the first statistical analysis, it was discovered that as generations progressed forward, the age at which people first had sex decreased. This could be due to cultural shifts, and straying away from more conservative and traditional values which emphasized chastity and waiting for such a connection until marriage. In modern culture, it is no longer looked down upon to do this, whereas in the past, it might have been condemned and therefore the behavior was less common, as seen in the results where generation X had a much higher mean sex age. This statistic may be somewhat inaccurate of the exact number, as the age 47 is incongruent with the average age of marriage around the time of Generation X.

Another interesting finding is the pendulum nature of trends. This can be seen in the smoking habits of Generation X, millennials, and Generation Z. While smoking was very popular with Generation X, having an average first smoking age of 16.15, it appeared to taper out for Millennials, who had a first smoking age of 31.54. Then, in Generation Z, the age of first smoking decreased back to 16.15, even lower than for Generation X. This elucidates the pendulum nature of trends, in this case, smoking, and that at one point, specifically in the millennial generation, people smoked for the first time at a much older age, perhaps suggesting that it went out of fashion.

Whereas in one department i.e. smoking, newer generations will be more prone to risk-taking, in other realms like looking at how income influences one's health, it appears that newer generations take the two hand in hand. Through the data looking at income and health between 3 generations, it appears that while earlier generations seemed to prioritize work over other departments in their life, newer generations may have access to information concerning health that influences their behavior. This can be seen in the high positive correlation between income and physical health seen through the amount of sleep one gets. It appears that as modern science develops, and newer generations continue to place value on physical wellbeing, the more money one makes, the more they prioritize their health, vs earlier generations, where it can be seen that there isn't as strong of a correlation between income and wellbeing, meaning that in some cases people may have placed more priority on producing a high income vs. taking care of themself. In conclusion, values and norms shift as time passes, as seen in looking at habits concerning smoking, sex, and health in regard to the three most recent generations.