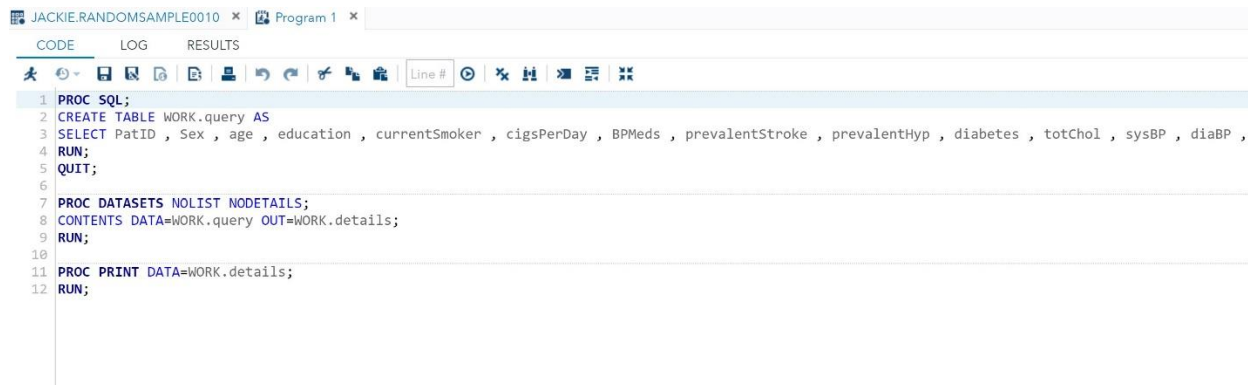


## 1) SAS code for SRS of n = 500



```

1 PROC SQL;
2 CREATE TABLE WORK.query AS
3 SELECT PatID , Sex , age , education , currentSmoker , cigsPerDay , BPMeds , prevalentStroke , prevalentHyp , diabetes , totChol , sysBP , diaBP ,
4 RUN;
5 QUIT;
6
7 PROC DATASETS NOLIST NODETAILS;
8 CONTENTS DATA=WORK.query OUT=WORK.details;
9 RUN;
10
11 PROC PRINT DATA=WORK.details;
12 RUN;

```

## 2) Table containing demographic variables

**a) Create an appropriate graphical display for the demographic categorical variable(s). Interpret your graph(s).**

The demographic variables in this data set are sex, age, and education. Sex is a qualitative, categorical variable (further classified as dichotomous or binary, since only two categories within the variable). In this study, age is a quantitative, discrete variable, since they only counted whole numbers for age. Education is a qualitative, ordinal variable, as it is a character variable, but has a natural ordering in terms of level of education received.

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	263	52.60	263	52.60
1	237	47.40	500	100.00

Interpretation: For sex, 0 = female, and 1 = male, so 52.60% of the participants are female, while the remaining 47.40% are male.

<b>age</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
33	1	0.20	1	0.20
34	3	0.60	4	0.80
35	3	0.60	7	1.40
36	8	1.60	15	3.00
37	13	2.60	28	5.60
38	21	4.20	49	9.80
39	21	4.20	70	14.00
40	25	5.00	95	19.00
41	17	3.40	112	22.40
42	20	4.00	132	26.40
43	15	3.00	147	29.40
44	20	4.00	167	33.40
45	23	4.60	190	38.00
46	26	5.20	216	43.20
47	16	3.20	232	46.40
48	28	5.60	260	52.00
49	11	2.20	271	54.20
50	12	2.40	283	56.60
51	17	3.40	300	60.00
52	18	3.60	318	63.60
53	13	2.60	331	66.20
54	10	2.00	341	68.20
55	23	4.60	364	72.80
56	18	3.60	382	76.40
57	18	3.60	400	80.00
58	9	1.80	409	81.80
59	9	1.80	418	83.60
60	16	3.20	434	86.80
61	12	2.40	446	89.20
62	13	2.60	459	91.80
63	14	2.80	473	94.60
64	7	1.40	480	96.00
65	4	0.80	484	96.80
66	6	1.20	490	98.00
67	6	1.20	496	99.20
68	3	0.60	499	99.80
69	1	0.20	500	100.00

Interpretation: The age is at exam time. The age 48 appeared the most (frequency = 28). Both the ages 33 and 69 appeared the least frequently, with frequency = 1 for both.

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	214	42.80	214	42.80
2	152	30.40	366	73.20
3	74	14.80	440	88.00
4	60	12.00	500	100.00

Interpretation: Education used the following scale: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College. Participants identifying as only having received some high school education (1) was the highest percent of participants at 42.80%, and the lowest percentage (12.00%) belonged to those identifying as having college education.

**b) Create a summary table for the demographic continuous variable(s): Your table should include three (3) measures of central tendency and two (2) measures of dispersion of your choice. Interpret these summary measures.**

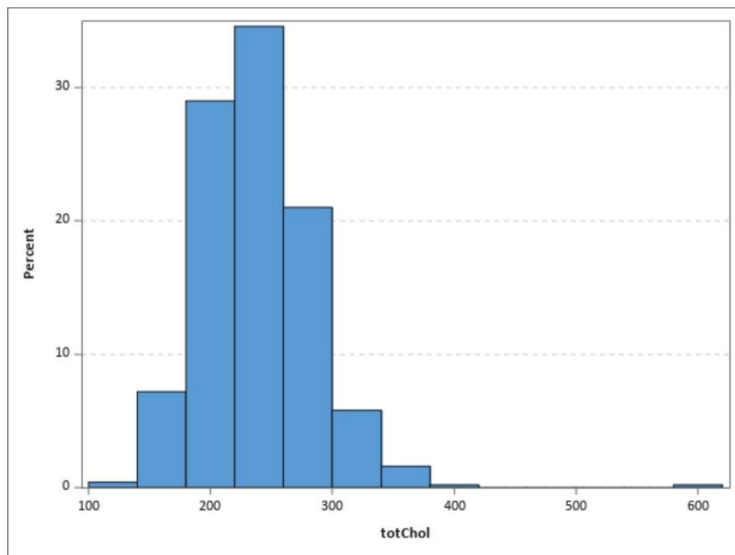
Summary Table for Demographic continuous variables

Variable	Mean	Minimum	Maximum	Median	Mode	Range
Sex	0.4740000	0	1.0000000	0	0	1.0000000
age	49.3940000	33.0000000	69.0000000	48.0000000	48.0000000	36.0000000
education	1.9600000	1.0000000	4.0000000	2.0000000	1.0000000	3.0000000

Interpretation: For the variable sex, calculating the mean and median does not make sense, as it is a qualitative, categorical variable, so computing measures of central tendency will not accurately describe the data set. However, the mode is useful, as it shows which value was present the most in the data set. Since 0 represents female, females were represented more in the group. For the variable age, it is appropriate to calculate the mean, since of the wide range of data points (minimum = 33; maximum = 69). It is also necessary to point at that the mean and median are similar values, possibly suggesting some form of normalcy. When looking at education, the mode is an important statistic, as it represents which education level appeared the most within the group. Similar as with age, the mean and median for education are close to each other, but with the mode being different, there is not as strong of an argument for normalcy.

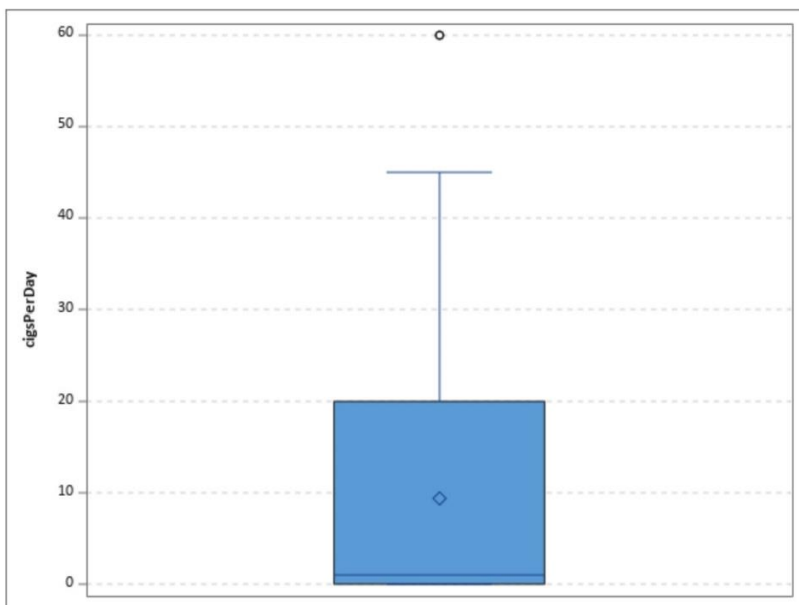
**3) Using the sample data, you've created in question 1, do the following:**

**a. Create a histogram of the variable 'serum total cholesterol' (TotChol).**



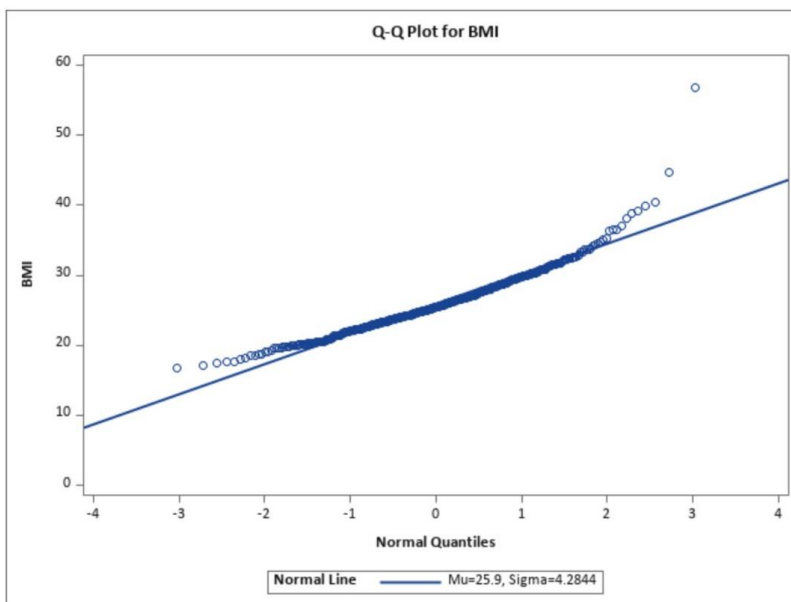
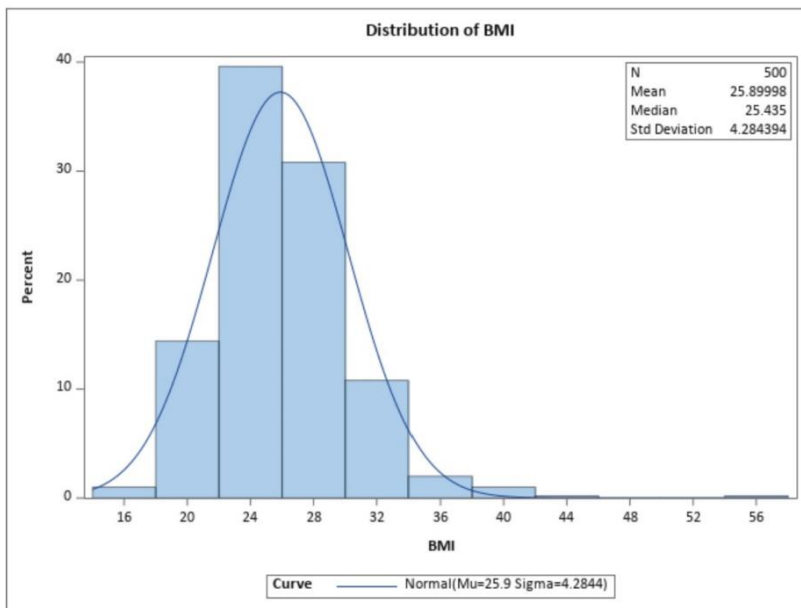
Interpretation: The histogram appears to be fairly even distributed, but there is one far right outlier that makes it seem somewhat skewed right. The mean, median, and mode are somewhat similar, but the mean will be pulled to the right due to the outlier, while the outlier will not have an affect on the median and mode.

**b. Create a boxplot of the variable ‘number of cigarettes smoked per day’ (CigsPerDay).**



Interpretation: There appears to be an outlier, the value at the very top.

**c. Investigate the normality assumption of the variable ‘body mass index’ (BMI).**



**d. Compute frequencies and percentages for the variables CurrentSmoker, PrevalentHyp, and CHD.**

currentSmoker	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	249	49.80	249	49.80
1	251	50.20	500	100.00

Interpretation: For this variable, 0 = nonsmoker, 1 = smoker. The data is almost split evenly, with 49.80% being nonsmokers and 50.20% being smokers.

<b>prevalentHyp</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>0</b>	334	66.80	334	66.80
<b>1</b>	166	33.20	500	100.00

Interpretation: prevalentHyp stands for prevalence of hypertension, so whether a person has hypertension or not. For this data set, 0 = no, and 1 = yes. 66.80% of people do not have hypertension, while 33.20% do.

<b>CHD</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>0</b>	412	82.40	412	82.40
<b>1</b>	88	17.60	500	100.00

Interpretation: CHD is an outcome variable, whether the patient developed coronary heart disease or not (0 = no; 1 = yes). This data was collected after a 10-year period. Of the patients, 82.40% did not develop coronary heart disease, while 17.60% did.

---

Jacqueline DiFulvio  
MPH 612  
Data Analysis Project Part 2

## **Introduction:**

The Framingham Heart Study (FHS) was conducted in Framingham, Massachusetts with data collection beginning in 1948. The study's goal was to identify factors that contribute to cardiovascular disease. Our own study had to questions of focus. The first was is there evidence that most participants in the Framingham Heart Study were current smokers at the baseline screening exam? The second question was does being overweight increase the risk for heart disease? Using the data collected in the FHS, we will compute statistical tests to answer our questions by randomly selected a sample pool of 500 from the data of FHS.

## **Methods:**

For the first research question focusing on smoking status, a two-tailed t-test was conducted. We ran tests for normality in the SAS software to verify that all assumptions were met. Our hypotheses for the test were  $H_0: \mu = 1$  ;  $H_a: \mu \neq 1$ . For the second research question on weight increasing risk for heart disease, an upper one-tail t-test was conducted. Again, we ran tests for normality in the SAS software to verify that all assumptions were met. The hypotheses for this test were  $H_0: \mu_1 - \mu_2 = 0$  ;  $H_a: \mu_1 - \mu_2 > 0$ .

## Results:

Table 1 shows the results from the description of the sample characteristics,  $n = 500$ . For sex, participants were roughly equal, but more participants were female (56%) than male (44%). Similar to sex, smoking status was also roughly equal at the status of exam. 52.80% were non-smokers, while 47.20% were smokers. The category with the most drastic difference was if the patient developed CHD, with 16.40% developed CHD and 83.60% not developing CHD.

**Table 1. Description of the Sample Characteristics,  $n = 500$**

Characteristics	Frequency (n)	Percent (%)
<b>Participant Sex</b>		
Male	220	44.00
Female	280	56.00
<b>Participant smoking status at exam</b>		
Nonsmoker	264	52.80
Smoker	236	47.20
<b>Patient developed Coronary Heart Disease</b>		
Yes	82	16.40
No	418	83.60
<b>Body Mass Index (kg/m<sup>2</sup>), mean (std)</b>	25.6771000 (4.3326932)	

\*std = Standard deviation

Table 2 shows the key statistics outputs from the t-tests ran for both study questions. For current smoking, the p-value is reported as  $Pr > |t|$  and the value is  $<0.001$ . Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. We conclude that we have enough evidence that the status of smoking is different than 1 (smoker), therefore majority of participants were not smokers at the baseline exam. For BMI affects on CHD, the p-value was 0.0324, which again is less than the significance level of 0.05, so we reject the null hypothesis. We conclude that we have enough evidence that being overweight does affect risk of developing CHD.

**Table 2. Summarization of statistical results**

	Mean (Std Dev)	Lower 95%, Upper 95%	t Value	Pr >  t
<b>Current Smoking</b>	0.4720000 (0.4997153)	0.4280923, 0.5159077	-23.63	<0.001
<b>BMI affects CHD</b>	-1.2434 (3.6376)	-2.3815, -0.1053	-2.15	0.0324

## Discussion:

Based on the statistical results, the majority of participants were not smokers at the time of the baseline screening exam. Additionally, it was found that being overweight, which we classified as having a BMI greater than 25, does increase the risk for heart disease. This may suggest that even if a participant was not a smoker at the time of baseline screening, if they were overweight, they would have an increased risk of heart disease.

## Appendix:

- a) SAS code for one-way frequencies for Table 1

```
proc freq data=JACKIE.FRAMINGHAMM_500;  
    tables Sex currentSmoker CHD / nocum plots=none;  
run;
```

- b) SAS code for summary statistics for BMI for Table 1

```
proc means data=JACKIE.FRAMINGHAMM_500 chartype mean std n vardef=df;  
    var BMI;  
run;
```

- c) SAS code and outputs for 95% confidence interval for the proportion of participants who were current smokers at exam, and hypothesis test.

```
proc means data=JACKIE.FRAMINGHAMM_500 chartype mean std stderr vardef=df c1m  
    alpha=0.05;  
    var currentSmoker;  
run;
```

```
/* Test for normality */
```

```
proc univariate data=JACKIE.FRAMINGHAMM_500 normal mu0=1;  
    ods select TestsForNormality;  
    var currentSmoker;  
run;
```

```
/* t test */
```

```
proc ttest data=JACKIE.FRAMINGHAMM_500 sides=2 h0=1 plots(showh0);  
    var currentSmoker;  
run;
```

Analysis Variable : currentSmoker				
Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
0.4720000	0.4997153	0.0223479	0.4280923	0.5159077

N	Mean	Std Dev	Std Err	Minimum	Maximum
500	0.4720	0.4997	0.0223	0	1.0000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
0.4720	0.4281 0.5159	0.4997	0.4705 0.5328

DF	t Value	Pr >  t
499	-23.63	<.0001

Tests for Normality			
Test	Statistic	p Value	
Shapiro-Wilk	W	0.635401	Pr < W <.0001
Kolmogorov-Smirnov	D	0.355553	Pr > D <.0100
Cramer-von Mises	W-Sq	14.66693	Pr > W-Sq <.0050
Anderson-Darling	A-Sq	90.03206	Pr > A-Sq <.0050



- d) SAS code and outputs for hypothesis to verify whether being overweight increases the risk for coronary heart diseases with the new data pool for BMI > 25.

```
/* Test for normality */
proc univariate data=JACKIE.OVERWEIGHT normal mu0=0;
  ods select TestsForNormality;
  class CHD;
  var BMI;
run;

/* t test */
proc ttest data=JACKIE.OVERWEIGHT sides=U h0=0 plots=none;
  class CHD;
  var BMI;
run;
```

Variable: BMI  
CHD = 0

Variable: BMI  
CHD = 1

Tests for Normality					Tests for Normality				
Test	Statistic		p Value		Test	Statistic		p Value	
Shapiro-Wilk	W	0.853316	Pr < W	<0.0001	Shapiro-Wilk	W	0.668295	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.137613	Pr > D	<0.0100	Kolmogorov-Smirnov	D	0.209061	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.282657	Pr > W-Sq	<0.0050	Cramer-von Mises	W-Sq	0.529718	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7.988977	Pr > A-Sq	<0.0050	Anderson-Darling	A-Sq	3.403024	Pr > A-Sq	<0.0050

Variable: BMI

CHD	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		207	28.5492	3.2254	0.2242	25.0300	43.4800
1		49	29.7927	5.0373	0.7196	25.1000	56.8000
Diff (1-2)	Pooled		-1.2434	3.6376	0.5779		
Diff (1-2)	Satterthwaite		-1.2434		0.7537		

CHD	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
0		28.5492	28.1072 28.9912	3.2254	2.9417 3.5700
1		29.7927	28.3458 31.2395	5.0373	4.2007 6.2931
Diff (1-2)	Pooled	-1.2434	-2.1975 Infity	3.6376	3.3469 3.9840
Diff (1-2)	Satterthwaite	-1.2434	-2.5034 Infity		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	254	-2.15	0.9838
Satterthwaite	Unequal	57.642	-1.65	0.9478

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	48	206	2.44	<.0001

# Study for correlation between age and serum total cholesterol (mg/dL)

## Introduction

Having a higher serum total cholesterol level increases the risk for heart disease (Schober et al., 2007, p. 1). While serum total cholesterol levels have been studied and shown to increase mortality among young and middle-aged people, the association of serum total cholesterol and mortality as age increases may be downplayed (Liang et al., 2017, p. 2). Here I investigate if there is a relationship between increasing age and serum total cholesterol levels.

## Methods

I chose a sample of 500 participants from the Framingham Heart Study, centered in Framingham, Massachusetts, focusing on the variable age to try to answer the question if there is a linear relationship between age and serum total cholesterol (mg/dL). A scatter plot of age versus total serum cholesterol was plotted. I hypothesized that there will be a linear relationship between age and serum total cholesterol. The null hypothesis will show that there is no linear relationship between the two variables. Considering the hypothesis, a linear regression of scatter plot was performed ( $\hat{Y} = 185.53762 + 1.01701 \cdot X$ ) and the results are shown in Table 1.

## Graphs/Tables

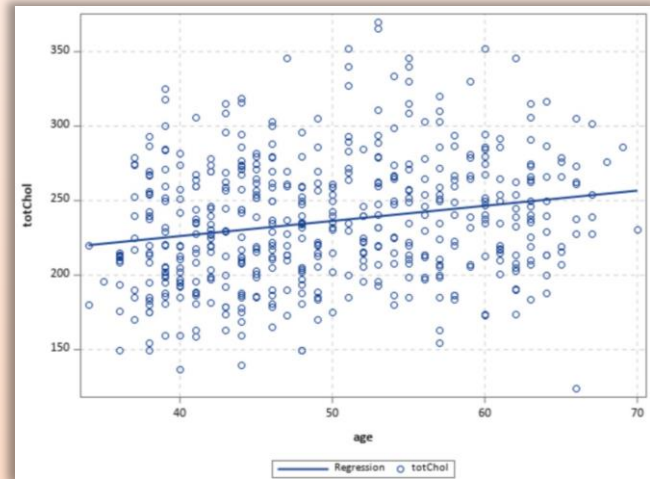


Figure 1. Scatter plot of age and serum total cholesterol and associated linear regression

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38551	38551	22.84	<.0001
Error	498	840554	1687.85948		
Corrected Total	499	879105			
Root MSE		41.08357	R-Square	0.0439	
Dependent Mean		236.00600	Adj R-Sq	0.0419	
Coeff Var		17.40785			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	185.52762	10.72085	17.31	<.0001
age	1	1.01701	0.21280	4.78	<.0001

Table 1. Linear regression parameters of age and serum total cholesterol.

## Key Findings/Conclusions/Concepts from Course

A linear regression gives a correlation factor of  $r=0.2095$  ( $r^2=0.0439$ ). The intercept is 185.5376, which means that on average, there is 185.5376 mg/dL of serum total cholesterol at the starting age. The 1.0170 slope of this linear regression shows the rate of increase with age. Coefficient of determination  $r^2=0.0439 < 1$  shows that the data is very scattered around the regression line and only 4.39% of variation along the regression line can be explained by variable age. The correlation factor of 0.2095 shows that there is no obvious linear relationship between these two variables, and only about 20.95% of the variation in serum total cholesterol is explained by age. P-value =  $<0.0001$  shows that there is strong evidence to reject the null hypothesis. I conclude that age may be a good predictor of serum total cholesterol level (mg/dL).

Data Analysis performed by: Jacqueline DiFulvio

### Data Sources:

Liang, Y., Vetrano, D. L., & Qiu, C. (2017). Serum total cholesterol and risk of cardiovascular and non-cardiovascular mortality in old age: a population-based study. *BMC geriatrics*, 17(1), 294.

Schober, S. E., Carroll, M. D., Lacher, D. A., Hirsch, R., & Division of Health and Nutrition Examination Surveys (2007). High serum total cholesterol – an indicator for monitoring cholesterol lowering efforts: U.S. adults, 2005-2006. *NCHS data brief*, (2), 1-8.