Analysis Albert Oleksy 12/2/18 5th period Stats
R sq = 12.9 % (or .122) Bivariate Data Project
Regression equation
10.75 + .423 x
R = 34.9% (or .349)

In this study, the population of citizens in urban cities is considered, with a sample of the pedestrians in 36 urban cities who walk through main streets during the day. This study seeks to find if there is a correlation between walking pace in American cities and their respective heart healths. These 36 cities range across America, with 12 from the Midwest, 6 from the Northeast, 12 from the west, and 6 from the South. The 2 states that are not contiguous, Alaska and Hawaii, are not represented. Not every contiguous state is represented either. The study includes 7 cities from California, which may hurt the representation of the whole country. A more representative state may provide more urban cities, with representation from each state in proportion to their state population, as well as by including Alaska and Hawaii.

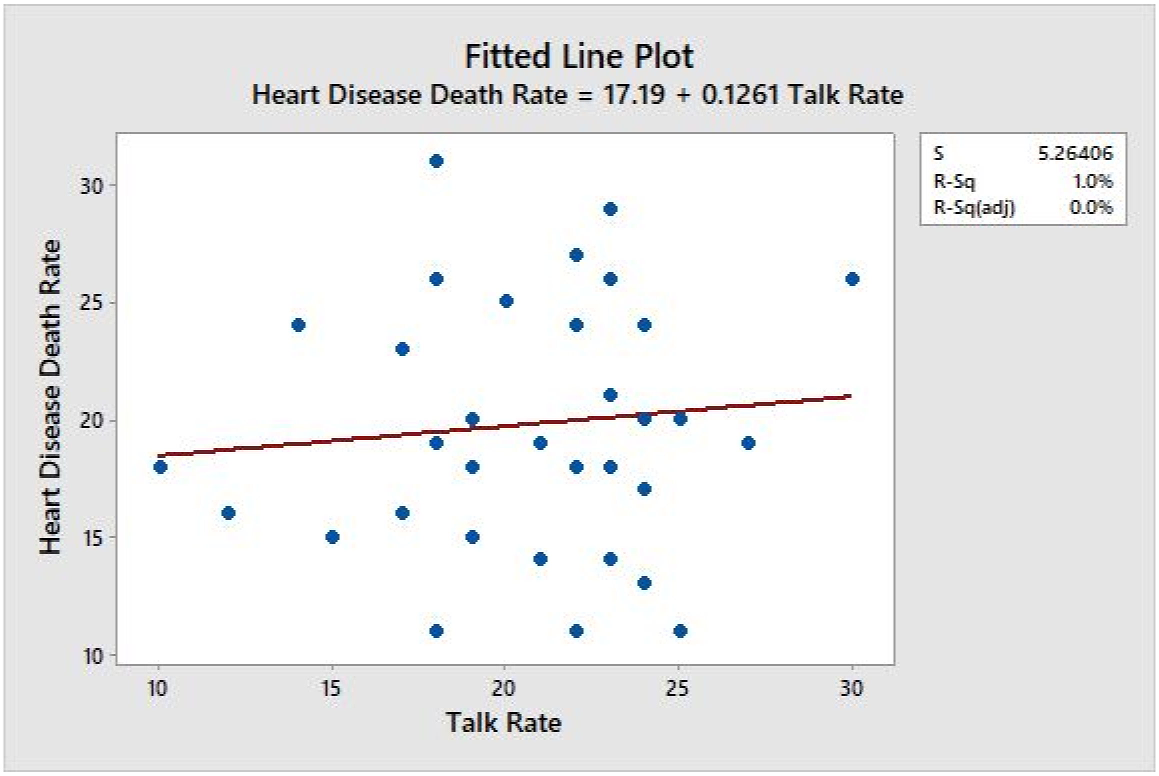
In the regression analysis, especially significant were the correlation coefficient (r) and the coefficient of determination (r^2). Since r was .349, the data weakly correlates in a linear fashion, but in the positive direction. So, as walking rate increases, so does the death rate of heart disease, which supports the idea that a faster pace of life causes more heart disease. R^2 was .122, which shows that the proportion of y (heart disease death rate) that could be attributed to the linear relationship of the two variables of data (walking rate and heart disease death rate) is also quite low (as it's on a scale of 0 to 1), showing that perhaps a least squares line doesn't represent the data very well, and there isn't a strong linear correlation. There doesn't seem to be a relationship between walking pace and heart disease death rate.

The data and the graph used to represent the data show that while there is a positive linear correlation between walk rate and heart disease death rate, it is quite weak. A large regression for many of the points causes a low determination as well. This is likely due to other factors that vary from city to city (confounding variables) such as health and diet of the subjects, weather in the city, etc. These variables may affect the heart health of a city's citizens, and are not related to the pace of life idea that the original study focuses on. The original study also considers other variables that affect the pace of life of a city, including talking and banking speed. Since these differ greatly from city to city as well, and the pace of walking (which can't represent pace of life in its entirety) doesn't show a strong correlation, no strong connection between the more broad subject of pace of life and heart health can be found from the walking rate dependant variable either.

The original study looked to find a correlation between pace of life and heart health by taking data of big cities. The walking pace was measured by taking the time, with a stopwatch, that pedestrians took to walk 60 feet in a city. This could cause confounding variables, as people may walk at different paces in different parts of the city (for example, near area of work, people may walk faster to be on time to their job). To remedy this, the study records walking pace in multiple different areas of the city, instead of just a main downtown street, as the original study did. Furthermore, the original study only took measurements on clear, sunny days in attempt to keep confounding variables constant (this strategy is called blocking.) However, some cities have different climates, and a clear sunny day in one city may be more unusual than in another. For example, Los Angeles walking pace may not be affected by a clear sunny day because those days are common, but in Boston, where the climate is colder, people may be more affected by the unusually hot temperature. For this reason, the study should have measured walking pace during an average weather day in each respective city.

I originally used the variable of talking pace, which was found by asking clerks about the differences in mail services provided, and recording the syllables divided by the time of their response. However, this yielded no interesting data, as the the coefficient of determination was 1.0% (R-sq). The graph did not show any patterns that could be nonlinear, either, so there seemed to be no correlation between talking pace and heart health. (That graph is provided below.)

The only pattern found in the study was the normal distribution of walking pace. As expected, more extreme walking rates (whether they are more or less than the average) are less frequent, and the middle values (that are near both the mean and the median) and more frequent, with frequency decreasing at a fairly constant rate the further from the middle they extend.



Below is the data set (the values have been standardized)
36 urban cities represent the data points

Talk Rate	Heart Disease Death Rate	Walk
24	24	28
23	29	23
18	31	24
23	26	28
30	26	22

24	20	25
24	17	26
21	19	30
18	26	22
22	24	22
23	26	23
20	25	25
23	14	23
25	11	18
27	19	27
14	24	22
24	20	23
24	13	22
25	20	23
19	18	12
17	16	23
18	19	20
17	23	20
18	11	22
22	27	14
23	18	20
19	15	17
19	20	26
22	18	19
23	21	23

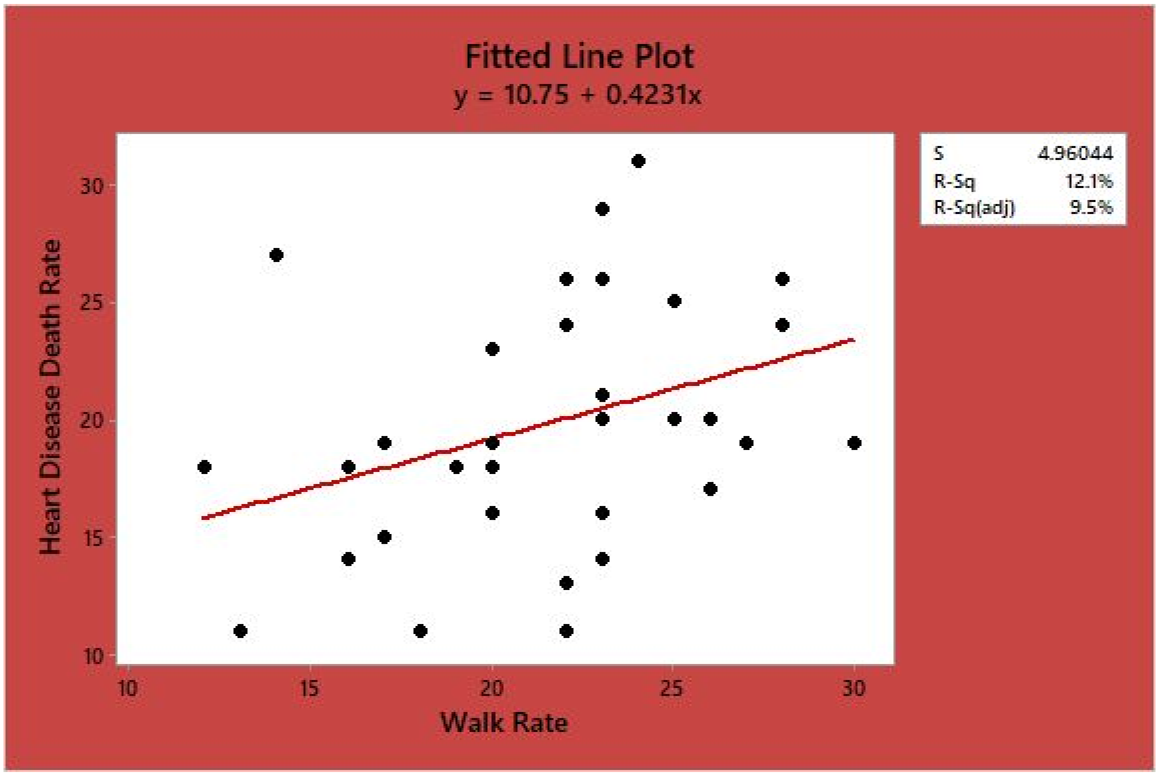
22	11	13
21	14	16
18	19	17
15	15	17
10	18	16
12	16	20

(Original Study and Data Citation: http://stat552.cwick.co.nz/homeworks/pace-of-life.pdf)

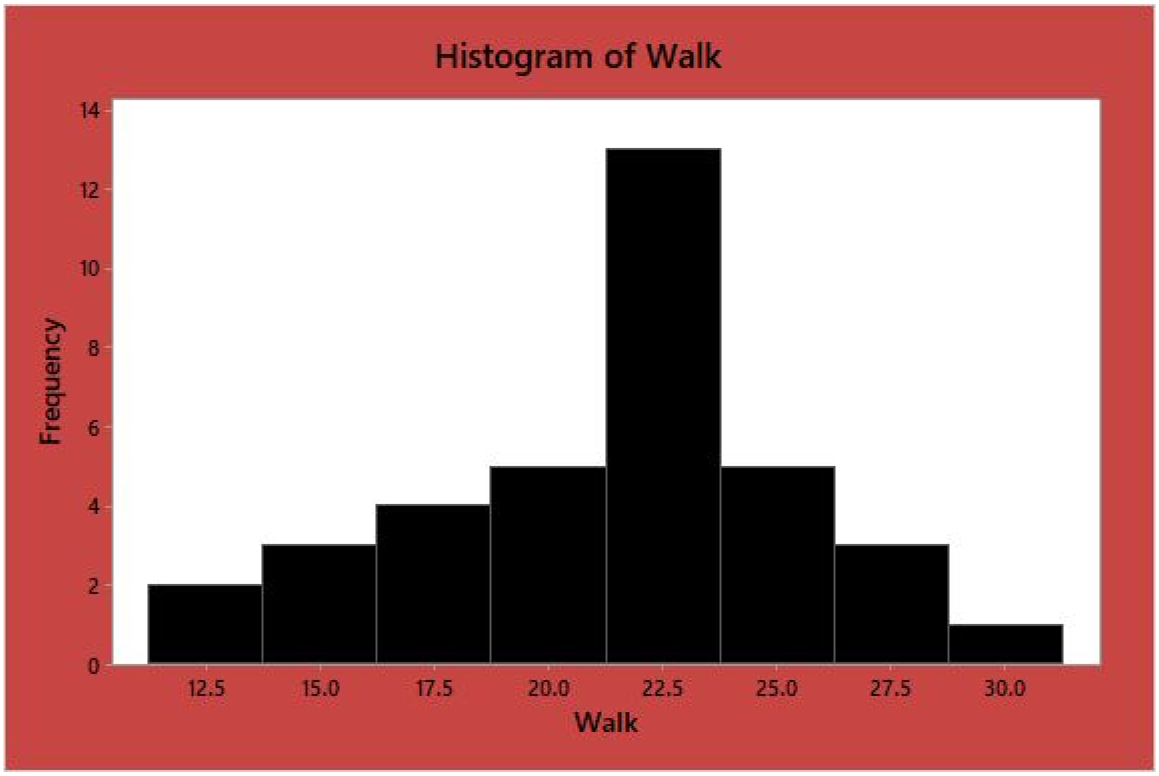
Below lists the descriptive statistics. Most of the data of the walking variable lie around the mean and median, which is represented by nearly a bell curve on the corresponding histogram. The corresponding histogram is provided in the infographic. (All graphs in the infographic are also provided below in this analysis)

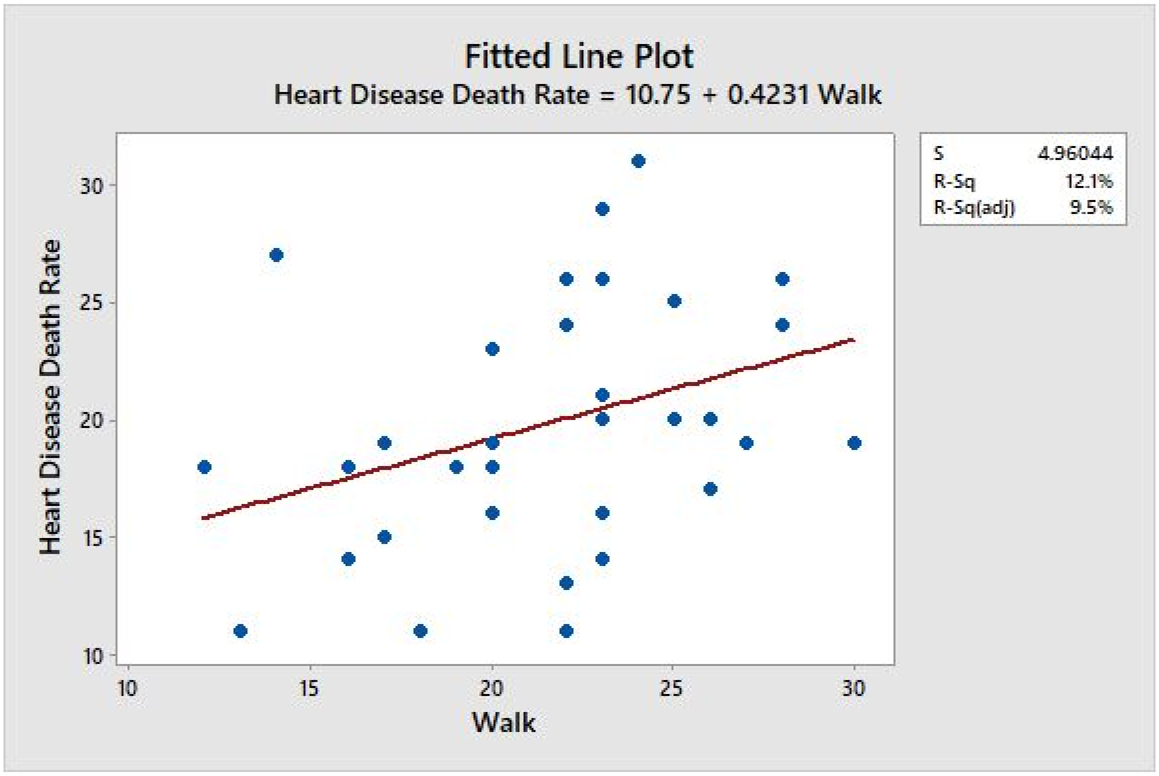
Walk:
Mean: 21.417
St. Dev.: 4.285
Min: 12.000
Q1: 18.250
Median: 22.000
Q3: 23.750
Max.:30.000

Heart Disease Death Rate:
Mean: 19.806
St. Dev.: 5.214
Min: 19.806
Q1: 16.000
Median: 19.000
Q3: 24.000
Max: 31.000



Need 2 graphs





Link to Infographic:
https://create.piktochart.com/output/34353035-biv-data-proj-walking-rate-vs-heart-disease

Regression Analysis: Heart Disease Death Rate versus Talk Rate

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	9.489	9.489	0.34	0.562
Talk Rate	1	9.489	9.489	0.34	0.562
Error	34	942.150	27.710		
Lack-of-Fit	13	241.050	18.542	0.56	0.862
Pure Error	21	701.100	33.386		
Total	35	951.639			

Model Summary

S	R-sq	R-sq (adj)	R-sq(pred)
5.26406	1.00%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	17.19	4.56	3.77	0.001	
Talk Rate	0.126	0.215	0.59	0.562	1.00

Regression Equation

Heart Disease Death Rate	17.19 + 0.126 Talk Rate

Fits and Diagnostics for Unusual Observations

Obs	Heart Disease Death Rate	Fit	Resid	Std Resid
3	31.00	19.46	11.54	2.24
5	26.00	20.	5.	1.05

```
                 9      0
                 7      3

 3      18.      1      -        -0.1
 5      00       8      0         0
                 .      .
                 4      4
                 5      5

                                          R  Large residual

                                           X  Unusual X
```