CS51: Correlation and Regression Report

A Critical Analysis of the Heart Attack Analysis and Prediction Dataset

Minerva University

CS51: Formal Analyses

Professor E. Volkan.

Rachael Chew

January 30, 2022

Part 1: Introduction

Every 45 seconds, one person in the United States has a heart attack, whereby 659,000 people die from heart diseases each year (Centers for Disease Control and Prevention, 2021). This report explores the Heart Attack Analysis & Prediction Dataset based on sample data obtained from a Cleveland database.

In particular, this report examines the relationship between age and maximum heart rate (HR_{max}), whereby a study has shown that a higher HR_{max} has a higher risk of mortality following a heart attack (Citroner, 2019). This analysis aims to answer the following question: *Is there a negative linear association between age and maximum heart rate in the population of men in Cleveland in 1988?*

By constructing a simple linear regression model, the dataset can be analysed to see if there is evidence to suggest that maximum heart rate decreases as age increases, potentially because older people lead more sedentary lifestyles whereby research has shown that a lack of exercise could lower HR_{max} (McNamara, 2014).

Part 2: Dataset

This dataset was obtained from Dr Robert Detrano, from the Cleveland Clinic Foundation in 1988 (Detrano, 1988), who collected 76 variables from 303 patients, such as their age, sex, cholesterol levels and HR_{max}. The data can be found here and the first few rows of the dataset are displayed in Appendix A.

Of the 76 variables, only 3 variables: sex, age and HR_{max}, are key variables of interest. Sex is a dummy variable, also known as a binary variable, represented as a Boolean value (1 = male; 0 = female). Since it represents 2 different categories of sex, it is treated as a qualitative categorical variable. Preprocessing of data was conducted for 2 purposes:

1. To filter out only data for males, which accounted for 207 instances.
2. To check for missing values, where it was concluded that there were none (Appendix A).

Age is the predictor variable (independent) while HR_{max} is the response variable (dependent) because HR_{max} is the hypothesised effect that varies according to age. HR_{max} is measured in beats per minute (bpm). HR_{max} and age are both collected as discrete quantitative variables because the integer values are assumed to a specified degree of precision, dictated by a strict finite set of values the variable can assume. Each count has a strict preceding and succeeding integer number, where no other values lie between them (Appendix A). This set of quantitative variables is bounded by a lower bound of 0, since 0 bpm and 0 years old are impossible.[1]

[1] **#variables**: I clearly identify and classify the variables relevant to my regression analysis - sex as a dummy variable (to filter relevant data for the analysis), and heart rate and age as quantitative discrete variables. I explain the identification of age and HR_{max} as whole numbers, because they can only assume specific values with a specific degree of precision with strict succeeding and preceding values, and further describe their lower bounds. I describe and explain the identified response (dependent) and predictor (independent) variables that will inform the analysis.

Part 3: Methods and Analysis

3.1 Summary Statistics

The dataset was read into Python using the pandas package for analysis. To begin the logistic regression, descriptive statistics for the data sample were examined. Appendix A illustrates general summary statistics for the HR_{max} of the sample of 207 men. The relevant summary statistics for age and HR_{max} are displayed in Table 1 and their sample distributions are displayed in Figures 1 and 2.

Table 1: Summary statistics for age and HR_{max} (beats per minute (bpm))

	Age	**HR_{max} (bpm)**
Count	$n_1 = 207$	$n_2 = 207$
Mean	$\bar{x} = 53.76$	$\bar{y} = 148.96$
Standard Deviation	$s_x = 8.88$	$s_y = 24.13$

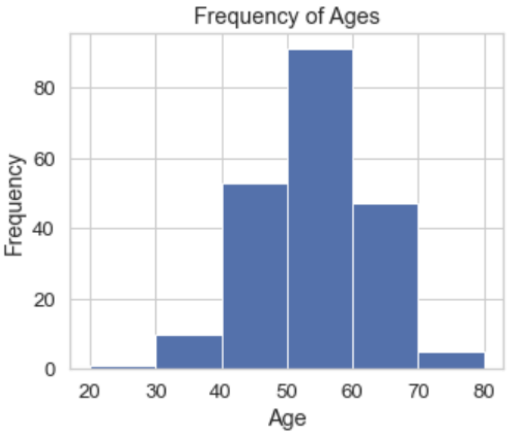


Figure 1: Histogram for Age

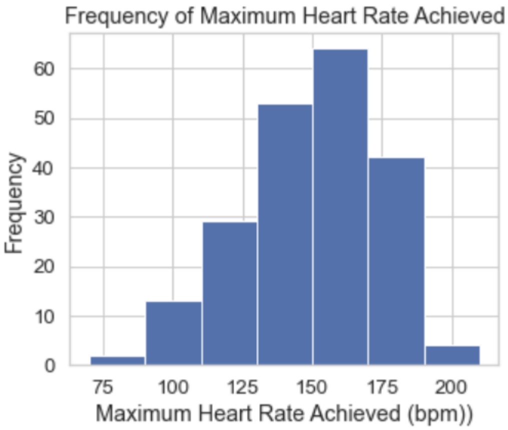


Figure 2: Histogram for HR_{max}

3.2 5 Conditions for Simple Linear Regression Model

To proceed with constructing the simple linear regression model, the following 5 conditions need to be satisfied to confirm if running a regression model is a suitable choice of association between age and HR_{max}.

Condition 1: Linearity

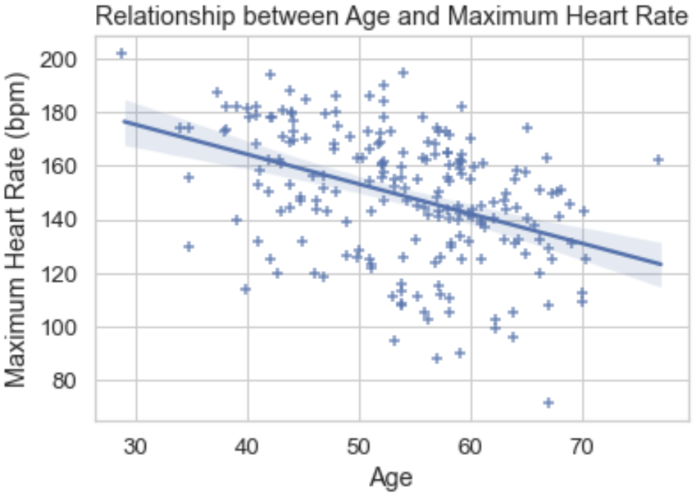


Figure 3: Correlation plot between age and HR_{max}

This linearity condition is satisfied upon inspection of Figure 3. The trend of data reveals a negative linear association between age and HR_{max}, whereby as age increases, HR_{max} decreases.

Condition 2: Independent Observations

To satisfy independence, the sample data has to consist of less than 10% of the male population in Cleveland. This condition is checked by noting that the Cleveland population size was 1,690,000 in 1988 (Macrotrends, 2021), whereby approximately 50% is assumed to be male. Since the sample data of 207 males is approximately 0.02% of the male population, the condition of sample size <10% population is satisfied.

Condition 3: Normality of residuals

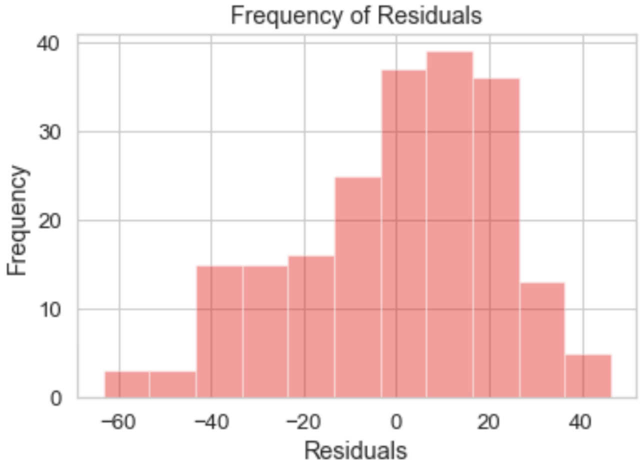


Figure 4: Frequency Distribution of Residuals

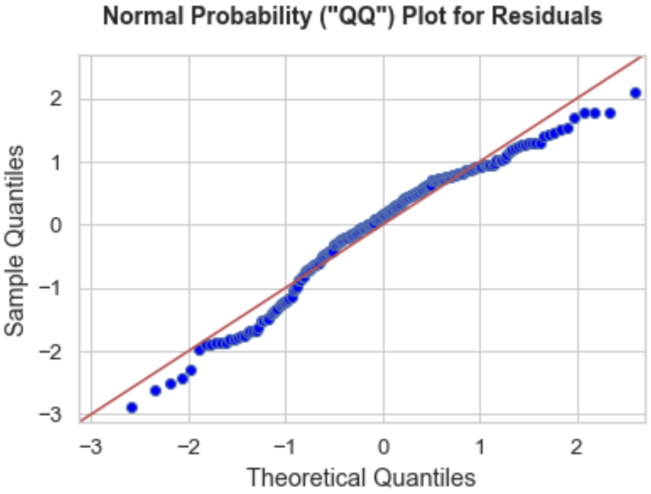


Figure 5: Normal Probability Plot for Residuals

As observed from the bell-shaped unimodal histogram (Figure 4), the residual distribution has a slight left-skew (non-normality). This aligns with the Q-Q plot (Figure 5), where the blue data points slightly deviate from the red line representing a standard normal distribution. Non-normality of residuals reflects that the error in the following model is inconsistent across the range of observed sample data, hence the predictive ability of the following regression model is inconsistent across the full range of the outcome variable (HR_{max}). Since the normality of residuals is unsatisfied, the predictive power of the following model decreases.

Condition 4: Equal Variance

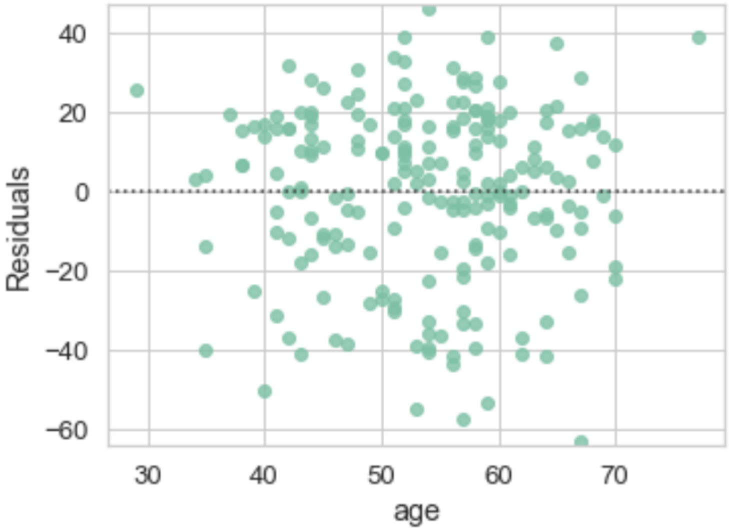


Figure 6: Residual Plot

From Figure 6, the approximately consistent variability of residuals has a rectangular-shaped distribution about the least-squares line (at y=0). This homoscedastic distribution of residuals means that the variance of the dependent variable (HR_{max}) is roughly equal for all the data, hence this condition is satisfied.

Condition 5: Random Observations

This condition cannot be checked, since the data collection methods for this dataset was unknown. Even if the distribution of HR_{max} is slightly normal (Figure 2), it is extremely plausible that the dataset was collected with a natural underlying structure between observations. For instance, observations that occurred consecutively during mandatory checkups for men with poorer health or a history of heart conditions could result in an inaccurate assumption of randomness. This leads to a sample dataset revealing a linear association between age and HR_{max}, when there could be none in the population because heart condition could be an extraneous variable affecting the trend of decreasing HR_{max} with age. The data could be sourced from a single clinic, where inferences from one clinic sample could poorly represent the Cleveland population because this clinic could be skewed towards a demographic of men. Acknowledging the plausibility of a lack of randomness would lower the predictive power of the model[2].

3.3 Pearson's Correlation Coefficient, r-value

To assess the strength of the linear association between age and HR_{max}, Pearson's correlation coefficient (r-value) is computed. Computing Pearson's r uses the formula $R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$, whereby x represents age and y represents HR_{max}. x_i and y_i represent each data point, $\bar{x}$ and $\bar{y}$ represent the sample means, while s_x and s_y are the sample deviations. The correlation coefficient is computed in Python (see Appendix B), producing an r-value of -0.409, whereby the absolute value of r = | -0.409 | = 0.409. Since $0.3 < 0.409 < 0.5$, it can be interpreted that the r-value is low, and hence the strength of the negative linear association between age and HR_{max} is weak (Moore, 2013). Since only 3 of 5 conditions for plotting a least-squares line are fulfilled, the accuracy of the r-value is questionable, since it may lack representation of the entire population of men in Cleveland, and skewed towards this particular sample of men.

While this analysis reveals a weak linear association, causation cannot be inferred between age and HR_{max}, because correlation does not imply causation. Underlying extraneous variables, for example, fitness levels or cholesterol levels could be the true cause of HR_{max} to decrease while age increases. For instance, because older people lead more sedentary lifestyles, their heart capacity decreases as a direct result of lack of physical activity (Citroner, 2019), hence the lowered HR_{max}. A controlled

[2] **#dataviz**: I clearly illustrate the data in the form of a table, histograms, scatter plots and a 'Q-Q' plot to easily identify the features and relationships between each variable. Histograms (Figs. 1, 2 and 4) were selected to visualise frequency distributions of dependent and independent variables, and residuals - which was effective in identifying the left skew in its distribution to evaluate the condition of normality. This observation aligned with that in the Q-Q plot (Fig. 5) where data points deviated from the straight line, reflecting deviation from normality. I intentionally utilised scatter plots in Figs. 3 and 6 to observe and evaluate the relationships between two numeric variables (independent and dependent variables) in Fig 3, and residual and independent variable in Fig 6 to understand their variance in relation to the least-squares line.

experiment is needed to establish causal relationships. In the context of this data collection, it is possible that this sample of men underwent a medical examination because they were specially selected for follow-up from a history of a type of heart condition, or unhealthy track record. In reality, age and HR_{max} could potentially lack a linear association, as levels of fitness could be the extraneous variable causing this association in this particular sample of men[3].

3.4 Regression Equation and Coefficient of Determination, R^2

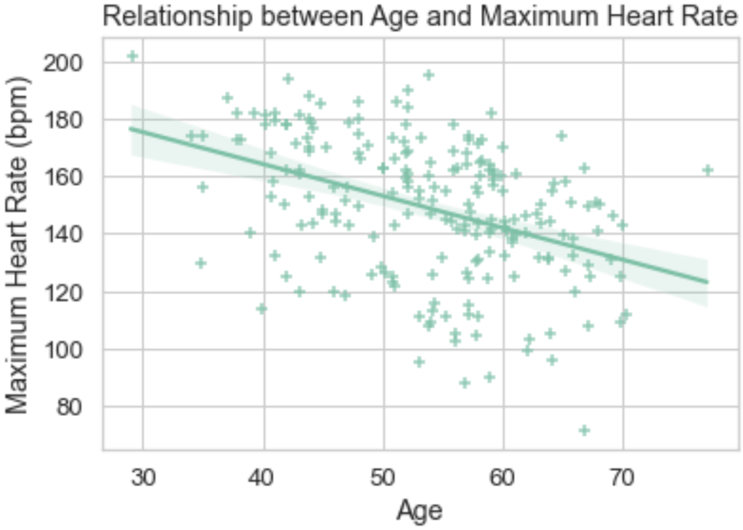


Figure 7: Regression plot of Age and HR_{max}

Since the r-value only shows the degree of association, a regression model was constructed to predict the variation in HR_{max} (dependent variable) based on age (independent variable). An equation for the regression line was constructed: HR_{max} = -1.11 * age + 208.625. This means that for every 1-year increase in age, there is a 1.11 bpm decrease in HR_{max}. The slope parameter is -1.11 bpm.

[3] **#correlation:** The correlation coefficient was computed to produce -0.409, and explained to suggest a weak linear association between age and HR_{max}, in the context of the data collected. The distinction between correlation and causation is clearly explained, that causation cannot be drawn from this analysis because of the need to conduct a controlled experiment to draw causal links. Upon looking at the data visualisations in Section 3.2, the reliability of the r-value is questioned because only 3 of 5 conditions for the least-squares line are fulfilled. I clearly identified extraneous variables such as cholesterol and fitness levels that could be the underlying cause of the association, because physical activity generally decreases with age, and a lack thereof decreases the capacity of the heart, leading to lowered HR_{max}.

The coefficient of determination, R-squared, uses the formula $R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$ to evaluate how much of the variability in HR_{max} (dependent variable) can be explained by the regression model. Upon computation in python, $r^2 = 0.167$, means that 16.7% of the variance in HR_{max} can be explained by the change in age. The rest of the variance remains unexplained and is attributed to other factors that are not included in the regression equation. Since $r^2 = 0.167$ is low, the regression model has low explanatory power. This is also observed from the data points that do not cluster tightly around the regression line in Figure 7[4]. The equation also allows one to predict the value of HR_{max} for an age that falls within the dataset range of ages.

3.5 Hypothesis Testing

To assess the statistical significance of the slope coefficient (-1.11) of age (independent variable) in the regression model, the following hypotheses are set up:

H0: there is no linear association between age and HR_{max} in the population of men in Cleveland in 1988; β1 = 0.

HA: there is a negative association between age and HR_{max} in the population of men in Cleveland in 1988; β1 < 0.

The T-score is first computed using the formula $t = \frac{b_1 - 0}{\mathrm{SE}(b_1)}$, where SE follows the formula used in the confidence interval computation. See Appendix C for the calculation in Python. The T-score of -6.41 represents the number of standard deviations from the mean in the t-distribution, and it results in a one-tailed p-value of $9.44e^{-10} < 0.05$. This p-value of $9.44e^{-10}$ represents the probability of obtaining

[4] **#regression:** The regression model was constructed to show the relationship between age (independent variable) and HR_{max} (dependent variable). The regression equation: HR_{max} = *-1.11 * age + 208.625* was explained to reveal that for every 1-year increase in age, there is a 1.11 bpm decrease in HR_{max}. The explanatory power of the model is evaluated based on the low R^2 value (0.167), whereby only 16.7% of the variation HR_{max} can be explained by the model, hence the regression model is evaluated to possess weak explanatory power.

an effect at least as extreme as the one in the sample data, assuming the truth of the null hypothesis. Since the p-value of $9.44e^{-10}$ < 0.05, it is statistically significant at α= 0.05, hence it is concluded that there is sufficient evidence to reject the null hypothesis that the slope of the regression line is 0.

In this context, a failure to identify an association between age and HR_{max}, (Type II error, failing to reject a false null hypothesis) has more dangerous implications than a Type I error (rejecting a true null hypothesis) because of potential higher risk of heart attacks, which could have been preventable by mandatory checks directed at men in the age range of extreme low and high HR_{max} that could increase their risk of heart attacks.[5]

[5] **#significance**: A measure of statistical significance was obtained from the t-score and p-value. The p-value of 9.44e-10 < 0.05 at alpha = 0.05 was interpreted to explain the presence of statistical significance, and hence there is sufficient evidence to reject the null hypothesis that the slope of the regression line is 0, hence a linear association exists between age and $HR_{max.}$ This increased the strength of the regression model since the slope of the independent variable in the regression model is significant. Lastly, implications of Type II error is analysed to be more severe because of the potential lack of health checks, which could be mitigated if HR_{max} is predicted to be associated with age.

Part 4: Results and Conclusions

The association between age and HR_{max} have been examined based on this dataset, by calculating Pearson's r, and constructing a regression model. Based on the results, the regression model has limited utility, because of the plausibility of extraneous variables (cholesterol and fitness levels) and biased selection of sample (men with a history of heart diseases) causing the apparent negative linear association between age and HR_{max}. However, in the case that this dataset is indeed random and could be representative of the population of men in Cleveland, coupled with the regression equation evaluated to be statistically significant, the model can be utilised to predict HR_{max} based on their age. For example, it can be used to filter out specific medical tests that are of urgent need for men with higher or lower HR_{max} respectively, because high and low maximums pose different types of risks to being heart attack prone (McNamara, 2014). This model can be utilised alongside other regression models, to draw out combinations of variables that would place men in greater risk of cardiovascular diseases, and hence need more frequent health checkups. Since the dataset only accounted for men between the ages of 29 to 77, the model cannot be extrapolated and generalised to predict the HR_{max} outside of this age range.

This conclusion is an inductive statistical generalisation because it is drawn from observations of a sample of instances (207 men), which is a small subset of the population, used to draw a conclusion about the larger population of men in Cleveland in 1988 (approximately 845,000 men). The generalisation is strong because it is determined by conclusions drawn from the significance level of 5%. Since one can be 95% confident that the model produces an outcome HR_{max} that lies within the intervals, one can hence make a strong inference of the population. However, this strength is weakened because the sample is relatively small (0.02% of the population), hence the inductive conclusion about the population is unlikely to follow from the unrepresentative samples in the premises, which are the statistical test and regression model. Secondly, it is unreliable because the unknown data collection method could result in the biased collection of data (only unhealthy men) as mentioned previously, and hence unable to accurately reflect the HR_{max} of the population. Since only 3 of 5 conditions were satisfied to utilise the least-squares regression, the ultimate limitations of the dataset make the inference about the population unreliable. Drawing too strong a conclusion from this inadequate sample of 207 about the population of 845,000 men would lead to the "hasty generalisation" fallacy. More evidence will be needed to enhance the reliability of the data, but the truth in the population cannot be 100% guaranteed.[6]

Word Count: 1398 words

[6] **#induction**: I clearly identify the type of inductive reasoning used (statistical generalisation) because conclusions drawn about the larger population are extrapolated from a subset of the population, which means that there is an inductive leap in reasoning. I clearly explain why the inductive reasoning is strong because of the methods of analysis (95% confident about the conclusions drawn) but lacks in reliability because of the lack of information about the data collection methods. I explain that more evidence will be needed to enhance the induction, but the truth in the claims cannot be guaranteed.

References

Center for Disease Control and Prevention. (2021, September 27). Heart disease facts. Retrieved December 12, 2021, from https://www.cdc.gov/heartdisease/facts.htm.

Citroner, G. (2019, November 13). Yes, you still should get to the gym after you turn 60. Healthline. Retrieved December 12, 2021, from https://www.healthline.com/health-news/people-over-age-60-should-work-out-more-not-less.

Detrano, R. (1988). UCI Machine Learning Repository: Heart disease data set. Retrieved December 12, 2021, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Kim, J. (2015). How to choose the level of significance: A pedagogical note. Retrieved December 12, 2021, from https://mpra.ub.uni-muenchen.de/66373/1/MPRA_paper_66373.pdf.

MacroTrends. (2021). Cleveland metro area population 1950-2021. Retrieved December 12, 2021, from https://www.macrotrends.net/cities/22959/cleveland/population.

McNamara, D. J. (2014). Dietary cholesterol, heart disease risk and cognitive dissonance. Proceedings of the Nutrition Society, 73(2), 161–166. http://doi.org/10.1017/S0029665113003844

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). The basic practice of statistics (6th ed.). New York, NY: W. H. Freeman and Company.

Rahman, R. (2021, March 22). Heart attack analysis &amp; prediction dataset. Kaggle. Retrieved December 12, 2021, from https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

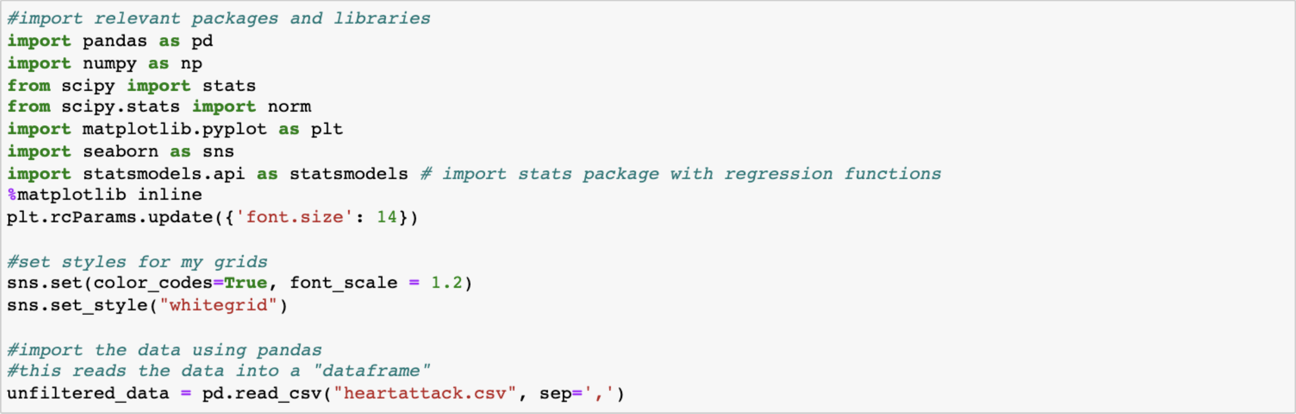
Reflection

I utilised Kosslyn's principle of building on prior associations (Kosslyn, 2017), by drawing on fall semester concepts of statistical significance by using p-values, to assess how significant the evidence is to reject or fail to reject my null hypothesis regarding the slope of my regression equation. I combined a correlation analysis with an evaluation of statistical significance, to see how significant the computed linear association between two variables are.

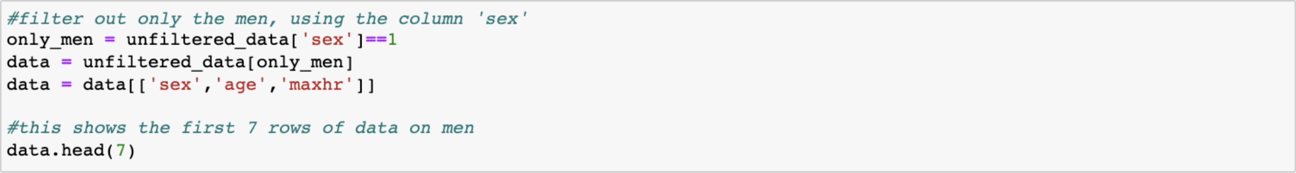
(74 words)

Appendix

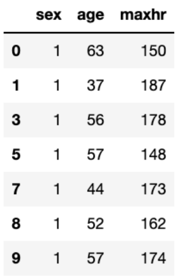
Appendix A: Import, Analyze, and Visualize Data



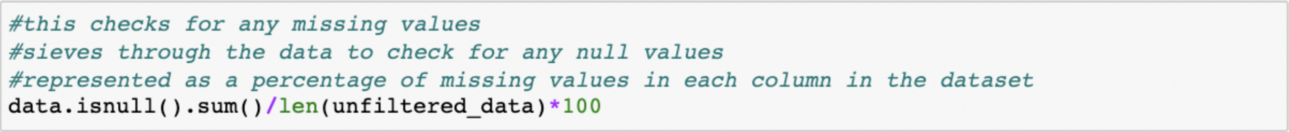
```
#import relevant packages and libraries
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as statsmodels # import stats package with regression functions
%matplotlib inline
plt.rcParams.update({'font.size': 14})

#set styles for my grids
sns.set(color_codes=True, font_scale = 1.2)
sns.set_style("whitegrid")

#import the data using pandas
#this reads the data into a "dataframe"
unfiltered_data = pd.read_csv("heartattack.csv", sep=',')
```



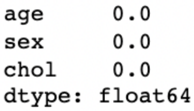
```
#filter out only the men, using the column 'sex'
only_men = unfiltered_data['sex']==1
data = unfiltered_data[only_men]
data = data[['sex','age','maxhr']]

#this shows the first 7 rows of data on men
data.head(7)
```



	sex	age	maxhr
0	1	63	150
1	1	37	187
3	1	56	178
5	1	57	148
7	1	44	173
8	1	52	162
9	1	57	174



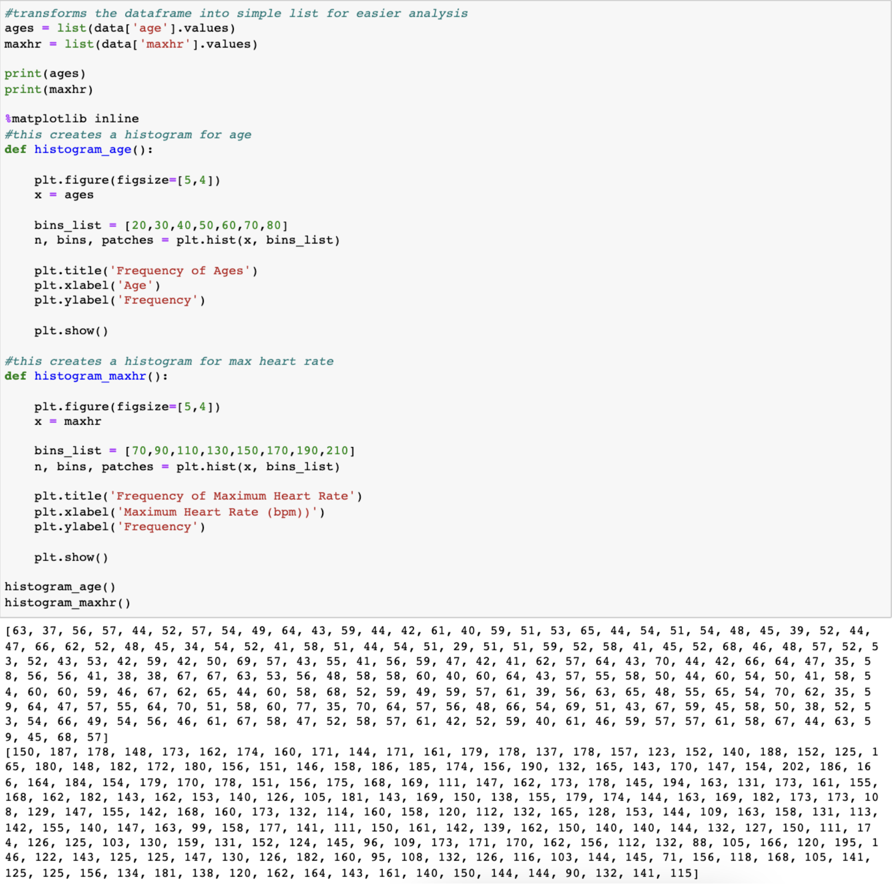
```
#this checks for any missing values
#sieves through the data to check for any null values
#represented as a percentage of missing values in each column in the dataset
data.isnull().sum()/len(unfiltered_data)*100
```



```
age      0.0
sex      0.0
chol     0.0
dtype: float64
```

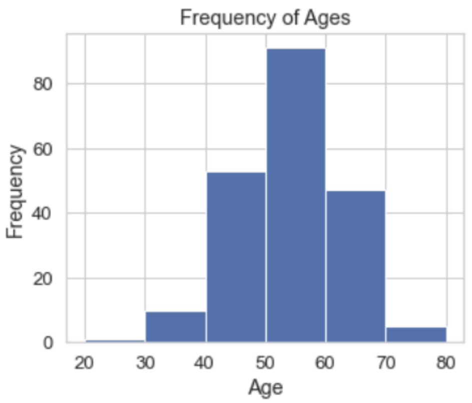
```
#print the summary statistics (count, mean, SD, etc.)

#as explained in the documentation, Bessel's correction is used by default.
#this means that the denominator in the calculation of standard deviation is n-1
#(https://pandas/pydata.org/panadas-docs/stable/generated/pandas.DataFrame.std.html)

data.describe()
```

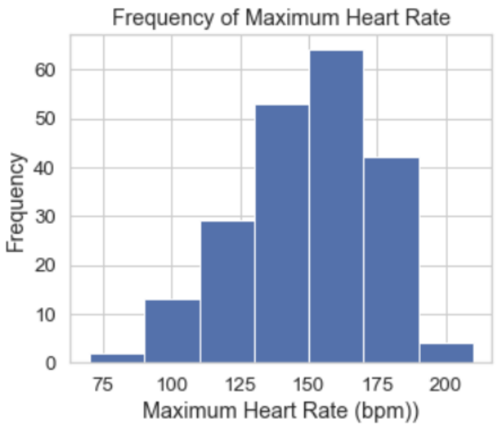
	sex	age	maxhr
count	207.0	207.000000	207.000000
mean	1.0	53.758454	148.961353
std	0.0	8.883803	24.130882
min	1.0	29.000000	71.000000
25%	1.0	47.000000	132.000000
50%	1.0	54.000000	151.000000
75%	1.0	59.500000	168.000000
max	1.0	77.000000	202.000000



```
#transforms the dataframe into simple list for easier analysis
ages = list(data['age'].values)
maxhr = list(data['maxhr'].values)

print(ages)
print(maxhr)

%matplotlib inline
#this creates a histogram for age
def histogram_age():

    plt.figure(figsize=[5,4])
    x = ages

    bins_list = [20,30,40,50,60,70,80]
    n, bins, patches = plt.hist(x, bins_list)

    plt.title('Frequency of Ages')
    plt.xlabel('Age')
    plt.ylabel('Frequency')

    plt.show()

#this creates a histogram for max heart rate
def histogram_maxhr():

    plt.figure(figsize=[5,4])
    x = maxhr

    bins_list = [70,90,110,130,150,170,190,210]
    n, bins, patches = plt.hist(x, bins_list)

    plt.title('Frequency of Maximum Heart Rate')
    plt.xlabel('Maximum Heart Rate (bpm))')
    plt.ylabel('Frequency')

    plt.show()

histogram_age()
histogram_maxhr()
```

```
[63, 37, 56, 57, 44, 52, 57, 54, 49, 64, 43, 59, 44, 42, 61, 40, 59, 51, 53, 65, 44, 54, 51, 54, 48, 45, 39, 52, 44,
47, 66, 62, 52, 48, 45, 34, 54, 52, 41, 58, 51, 44, 54, 51, 29, 51, 51, 59, 52, 58, 41, 45, 52, 68, 46, 48, 57, 52, 5
3, 52, 43, 53, 42, 59, 42, 50, 69, 57, 43, 55, 41, 56, 59, 47, 42, 41, 62, 57, 64, 43, 70, 44, 42, 66, 64, 47, 35, 5
8, 56, 56, 41, 38, 38, 67, 67, 63, 53, 56, 48, 58, 58, 60, 40, 60, 64, 43, 57, 55, 58, 50, 44, 60, 54, 50, 41, 58, 5
4, 60, 60, 59, 46, 67, 62, 65, 44, 60, 58, 68, 52, 59, 49, 59, 57, 61, 39, 56, 63, 65, 48, 55, 65, 54, 70, 62, 35, 5
9, 64, 47, 57, 55, 64, 70, 51, 58, 60, 77, 35, 70, 64, 57, 56, 48, 66, 54, 69, 51, 43, 67, 59, 45, 58, 50, 38, 52, 5
3, 54, 66, 49, 54, 56, 46, 61, 67, 58, 47, 52, 58, 57, 61, 42, 52, 59, 40, 61, 46, 59, 57, 57, 61, 58, 67, 44, 63, 5
9, 45, 68, 57]
[150, 187, 178, 148, 173, 162, 174, 160, 171, 144, 171, 161, 179, 178, 137, 178, 157, 123, 152, 140, 188, 152, 125, 1
65, 180, 148, 182, 172, 180, 156, 151, 146, 158, 186, 185, 174, 156, 190, 132, 165, 143, 170, 147, 154, 202, 186, 16
6, 164, 184, 154, 179, 170, 178, 151, 156, 175, 168, 169, 111, 147, 162, 173, 178, 145, 194, 163, 131, 173, 161, 155,
168, 162, 182, 143, 162, 153, 140, 126, 105, 181, 143, 169, 150, 138, 155, 179, 174, 144, 163, 169, 182, 173, 173, 10
8, 129, 147, 155, 142, 168, 160, 173, 132, 114, 160, 158, 120, 112, 132, 165, 128, 153, 144, 109, 163, 158, 131, 113,
142, 155, 140, 147, 163, 99, 158, 177, 141, 111, 150, 161, 142, 139, 162, 150, 140, 140, 144, 132, 127, 150, 111, 17
4, 126, 125, 103, 130, 159, 131, 152, 124, 145, 96, 109, 173, 171, 170, 162, 156, 112, 132, 88, 105, 166, 120, 195, 1
46, 122, 143, 125, 125, 147, 130, 126, 182, 160, 95, 108, 132, 126, 116, 103, 144, 145, 71, 156, 118, 168, 105, 141,
125, 125, 156, 134, 181, 138, 120, 162, 164, 143, 161, 140, 150, 144, 144, 90, 132, 141, 115]
```





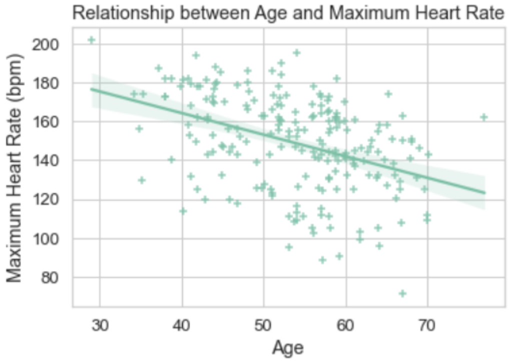
Appendix B: Pearson’s Correlation Coefficient

```
#create the functions
#this function prints the pearson's correlation value for two columns
#define the function, it takes two arguments (two columns from the dataset printed above: age and maxhr)
def pcorr(column_a, column_b):
    print("The pearson's r value comparing", column_a , "to", column_b , "is:")
    print(round(data[column_a].corr(data[column_b]),3)) #computes r of both variables, rounds to three decimal places
    print("")

# this function prints pearson's r for the two selected columns
# it also prints a scatterplot of the data with a trendline
# the plot introduces "jitter" to offset the values, so that they are not printed on top of one another

#function to plot the two variables form above in a scatterplot
def plotpcorr(column_a, column_b):
    pcorr(column_a, column_b)
    sns.set_palette("Set2")
    sns.regplot(x= column_a, y= column_b, data=data, marker="+", x_jitter=.25, y_jitter=.25)

#ensures that the above functions run
print("Functions loaded.")
```

```
Functions loaded.
```

```
plt.title('Relationship between Age and Maximum Heart Rate')
plotpcorr('age', 'maxhr')
plt.ylabel('Maximum Heart Rate (bpm)')
plt.xlabel('Age')
```

```
The pearson's r value comparing age to maxhr is:
-0.409


Text(0.5, 0, 'Age')
```

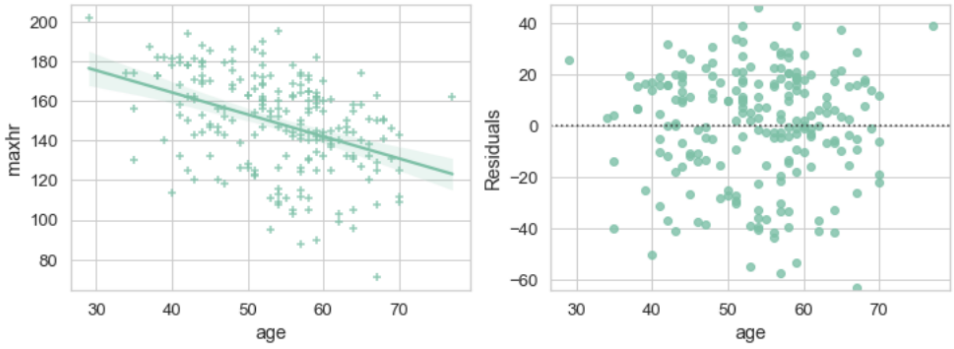


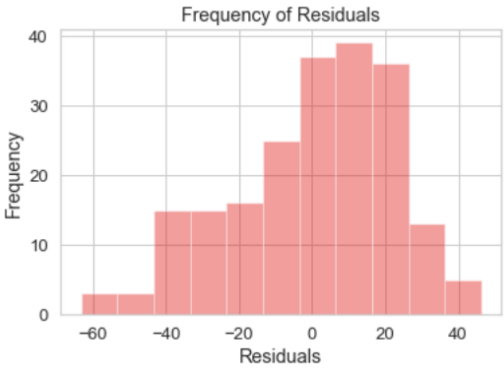
Appendix C: Coefficient of Determination

```
#code adapted from CS51 Session 4 - (2.2) Synthesis: Correlation, Regression, and Statistics
def regression_model(column_x, column_y):
    # this function uses built in library functions to create a scatter plot,
    # plots of the residuals, compute R-squared, and display the regression eqn

    # fit the regression line using "statsmodels" library:
    X = statsmodels.add_constant(data[column_x])
    Y = data[column_y]
    regressionmodel = statsmodels.OLS(Y,X).fit() #OLS stands for "ordinary least squares"

    # extract regression parameters from model, rounded to 3 decimal places:
    Rsquared = round(regressionmodel.rsquared,3)
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    # make plots:
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))

    sns.regplot(x=column_x, y=column_y, data=data, marker="+", ax=ax1) # scatter plot
    sns.residplot(x=column_x, y=column_y, data=data, ax=ax2) # residual plot, to evaluate normality condition
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)

    plt.figure() # histogram, for frequency of residuals
    plt.ylabel('Frequency')
    plt.title('Frequency of Residuals')
    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red') # histogram

    # print the results:
    print("R-squared = ",Rsquared)
    print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)
```

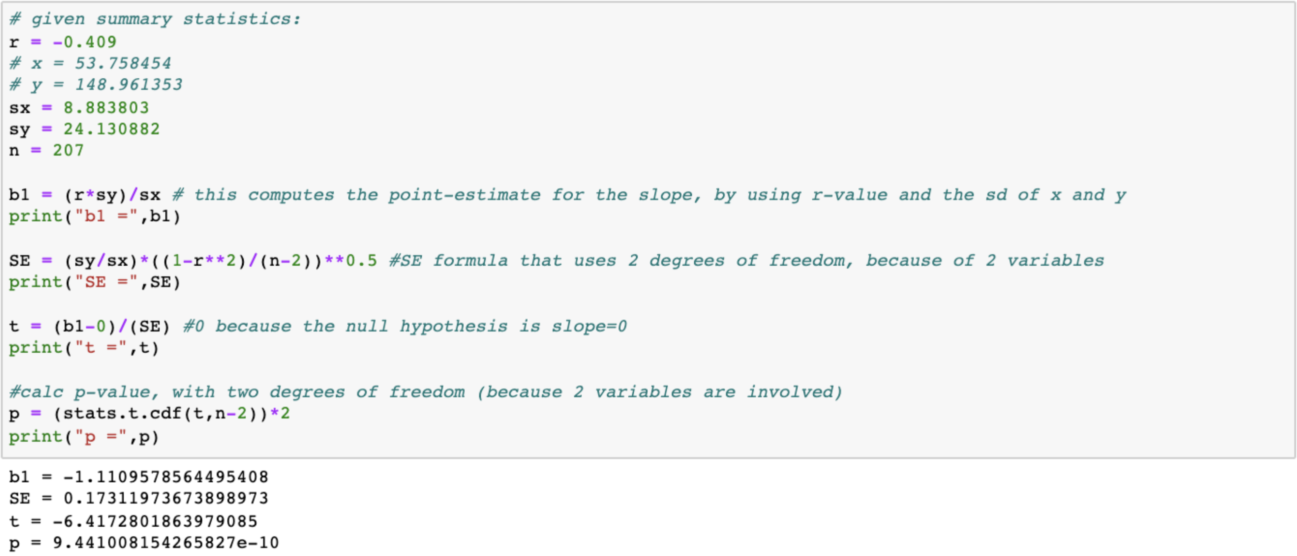
```
regression_model('age', 'maxhr')
```

```
R-squared =  0.167
Regression equation: maxhr =  -1.11 * age +  208.625
```





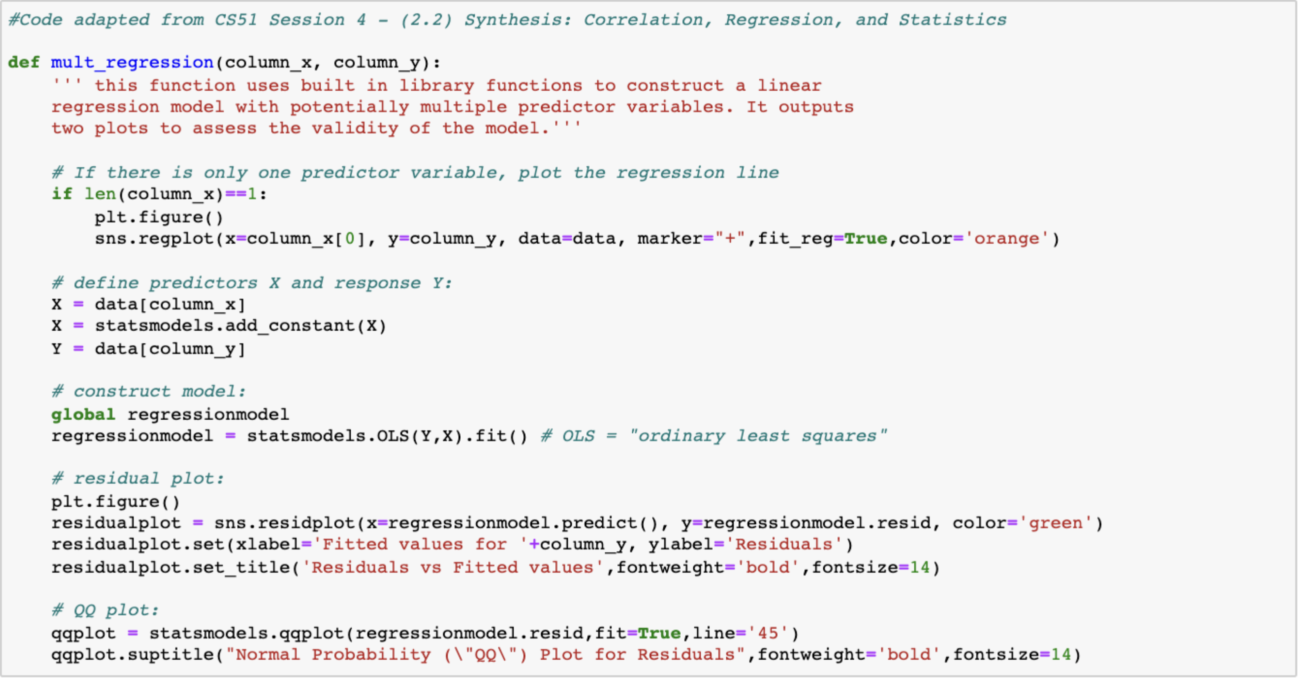
Appendix D: Hypothesis Testing



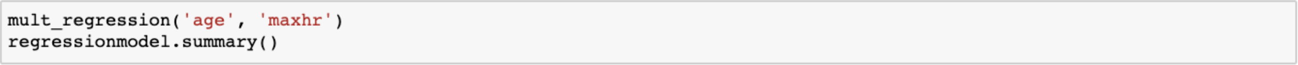
```
# given summary statistics:
r = -0.409
# x = 53.758454
# y = 148.961353
sx = 8.883803
sy = 24.130882
n = 207

b1 = (r*sy)/sx # this computes the point-estimate for the slope, by using r-value and the sd of x and y
print("b1 =",b1)

SE = (sy/sx)*((1-r**2)/(n-2))**0.5 #SE formula that uses 2 degrees of freedom, because of 2 variables
print("SE =",SE)

t = (b1-0)/(SE) #0 because the null hypothesis is slope=0
print("t =",t)

#calc p-value, with two degrees of freedom (because 2 variables are involved)
p = (stats.t.cdf(t,n-2))*2
print("p =",p)
```

```
b1 = -1.1109578564495408
SE = 0.17311973673898973
t = -6.4172801863979085
p = 9.441008154265827e-10
```

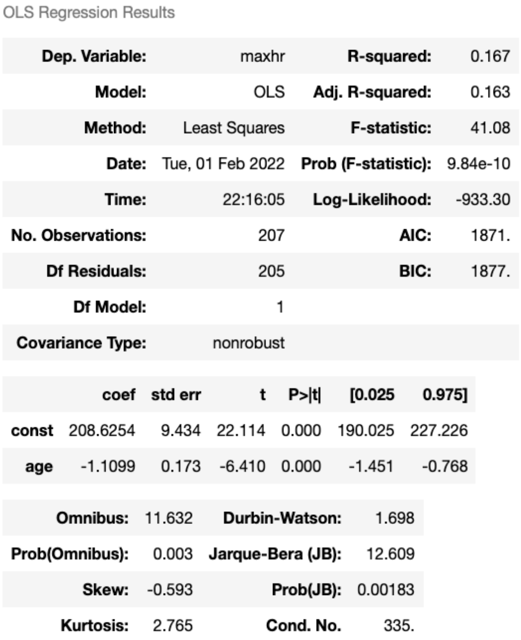
Appendix E: Summary Table



```
#Code adapted from CS51 Session 4 - (2.2) Synthesis: Correlation, Regression, and Statistics

def mult_regression(column_x, column_y):
    ''' this function uses built in library functions to construct a linear
    regression model with potentially multiple predictor variables. It outputs
    two plots to assess the validity of the model.'''

    # If there is only one predictor variable, plot the regression line
    if len(column_x)==1:
        plt.figure()
        sns.regplot(x=column_x[0], y=column_y, data=data, marker="+",fit_reg=True,color='orange')

    # define predictors X and response Y:
    X = data[column_x]
    X = statsmodels.add_constant(X)
    Y = data[column_y]

    # construct model:
    global regressionmodel
    regressionmodel = statsmodels.OLS(Y,X).fit() # OLS = "ordinary least squares"

    # residual plot:
    plt.figure()
    residualplot = sns.residplot(x=regressionmodel.predict(), y=regressionmodel.resid, color='green')
    residualplot.set(xlabel='Fitted values for '+column_y, ylabel='Residuals')
    residualplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)

    # QQ plot:
    qqplot = statsmodels.qqplot(regressionmodel.resid,fit=True,line='45')
    qqplot.suptitle("Normal Probability (\"QQ\") Plot for Residuals",fontweight='bold',fontsize=14)
```



```
mult_regression('age', 'maxhr')
regressionmodel.summary()
```

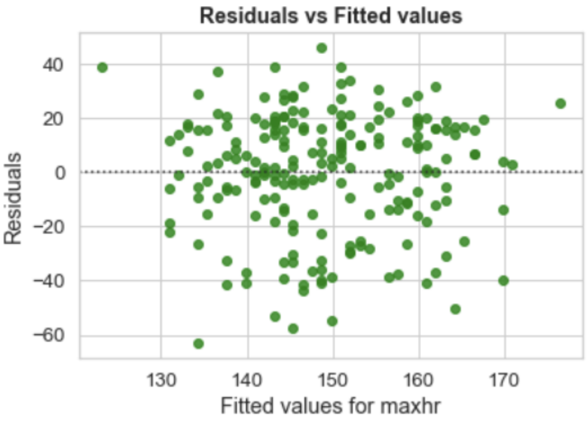


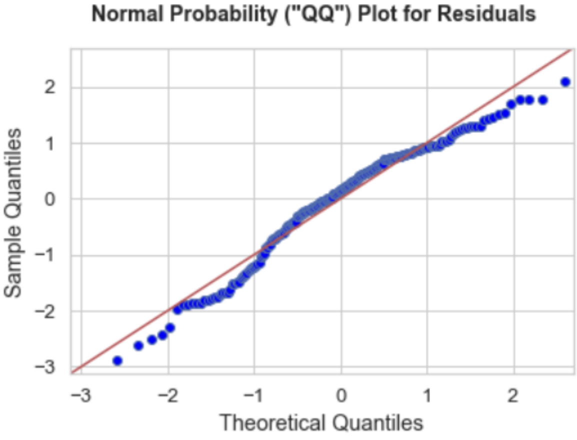
OLS Regression Results

Dep. Variable:	maxhr	**R-squared:**	0.167
Model:	OLS	**Adj. R-squared:**	0.163
Method:	Least Squares	**F-statistic:**	41.08
Date:	Tue, 01 Feb 2022	**Prob (F-statistic):**	9.84e-10
Time:	22:16:05	**Log-Likelihood:**	-933.30
No. Observations:	207	**AIC:**	1871.
Df Residuals:	205	**BIC:**	1877.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	208.6254	9.434	22.114	0.000	190.025	227.226
age	-1.1099	0.173	-6.410	0.000	-1.451	-0.768

Omnibus:	11.632	**Durbin-Watson:**	1.698
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	12.609
Skew:	-0.593	**Prob(JB):**	0.00183
Kurtosis:	2.765	**Cond. No.**	335.


Residuals vs Fitted values
Residuals
Fitted values for maxhr


Normal Probability ("QQ") Plot for Residuals
Sample Quantiles
Theoretical Quantiles