1. **Problem 1**
   Consider the time-series $(-3,-1,\ 1,\ 3,\ 5,\ 7,\ *)$. Here, a missing entry is denoted by $*$. What would be the estimated value of the missing entry using linear interpolation on a window of the last three values?

$$* = 7 + \frac{(7-5)}{(6-5)} * (7-6) = 9$$

   Therefore, the missing value is 9.

2. **Problem 2**

   (a) The associated task with this dataset is multiclass classification. Change the problem to binary classification and compute the proportion of each class in the binary case? Is this a balanced dataset?
   In order to convert this multiclass problem to a binary classification problem, in Python, I converted the levels of *num* using the following criteria:

$$num_{new} = \begin{cases} 0 & \text{if } num_{old} = 0 \\ 1 & \text{if } 1 \leq num_{old} \leq 4 \end{cases}$$

   From here, because the proportion of patients without the presence of heart disease is not equal to the proportion of patients with the presence of heart disease, this is not a balanced dataset.

   (b) After removing all missing values, we now have 297 patients, down from the earlier 303.

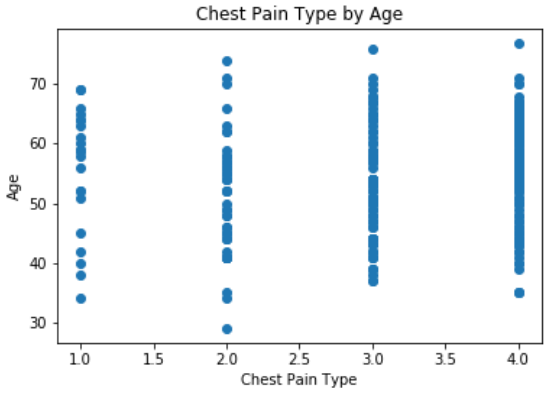
   (c) The means and standard deviations of the two are pretty constant.
   Dropped Data Imputation

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	num
count	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00
mean	54.54	0.68	3.16	131.69	247.35	0.14	1.00	149.60	0.33	1.06	1.60	0.46
std	9.05	0.47	0.96	17.76	52.00	0.35	0.99	22.94	0.47	1.17	0.62	0.50
min	29.00	0.00	1.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	1.00	0.00
25%	48.00	0.00	3.00	120.00	211.00	0.00	0.00	133.00	0.00	0.00	1.00	0.00
50%	56.00	1.00	3.00	130.00	243.00	0.00	1.00	153.00	0.00	0.80	2.00	0.00
75%	61.00	1.00	4.00	140.00	276.00	0.00	2.00	166.00	1.00	1.60	2.00	1.00
max	77.00	1.00	4.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	3.00	1.00

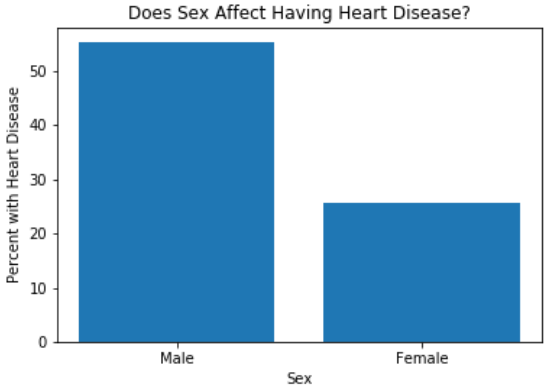
   Mean Data Imputation

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.44	0.68	3.16	131.69	246.69	0.15	0.99	149.61	0.33	1.04	1.60	0.67	4.73	0.46
std	9.04	0.47	0.96	17.60	51.78	0.36	0.99	22.88	0.47	1.16	0.62	0.93	1.93	0.50
min	29.00	0.00	1.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	1.00	0.00	3.00	0.00
25%	48.00	0.00	3.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00	3.00	0.00
50%	56.00	1.00	3.00	130.00	241.00	0.00	1.00	153.00	0.00	0.80	2.00	0.00	3.00	0.00
75%	61.00	1.00	4.00	140.00	275.00	0.00	2.00	166.00	1.00	1.60	2.00	1.00	7.00	1.00
max	77.00	1.00	4.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	3.00	3.00	7.00	1.00

(d) Based on the scatter plot, chest pain is more common among older adults (65+). If also seems like asymptomatic chest pain (value = 4.0) is more frequent than other values. In general, chest pain. Also, typical angina is the least common type of chest pain, but its frequency is more heavily located in the older ages.



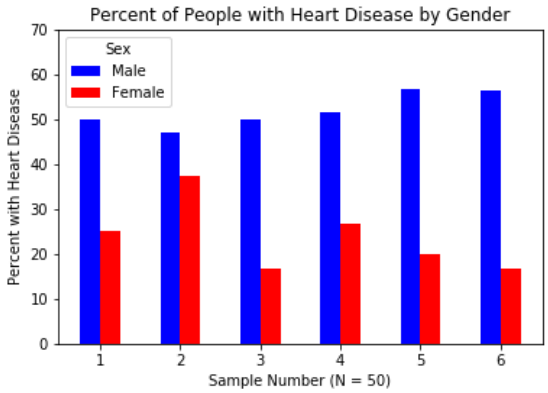
(e) Based on the bar plot, heart disease is much more common in men than women. That being said, it may be possible that *reported* heart disease is more common in men than women.



(f) i. Based on the table below, because each sample has an approximately 60-40 split, it is safe to say that the samples are balanced.

	Heart Disease Present	Heart Disease Not Present
0	0.44	0.56
1	0.44	0.56
2	0.42	0.58
3	0.44	0.56
4	0.42	0.58
5	0.42	0.58

ii.

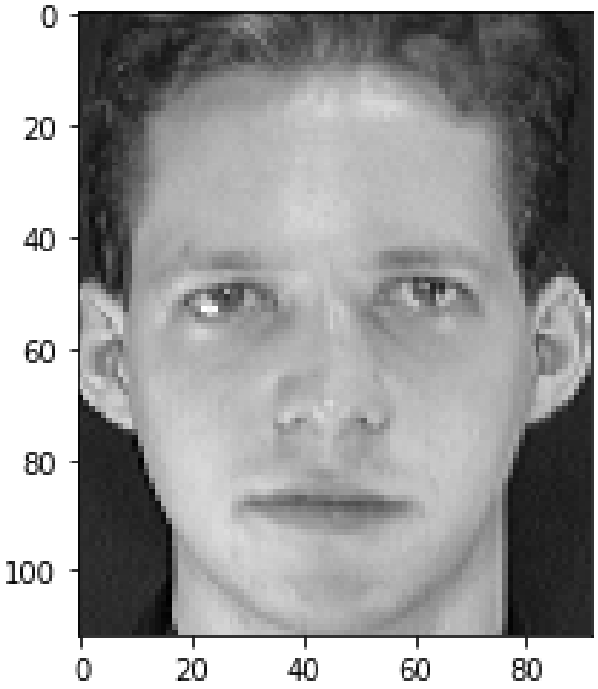


(g) Based on the bar plot, heart disease is still more common in men than women. The disparity is much greater in samples 3, 5, and 6 than in sample 2. Samples 1 and 4 most closely resemble the box plot of the full sample.

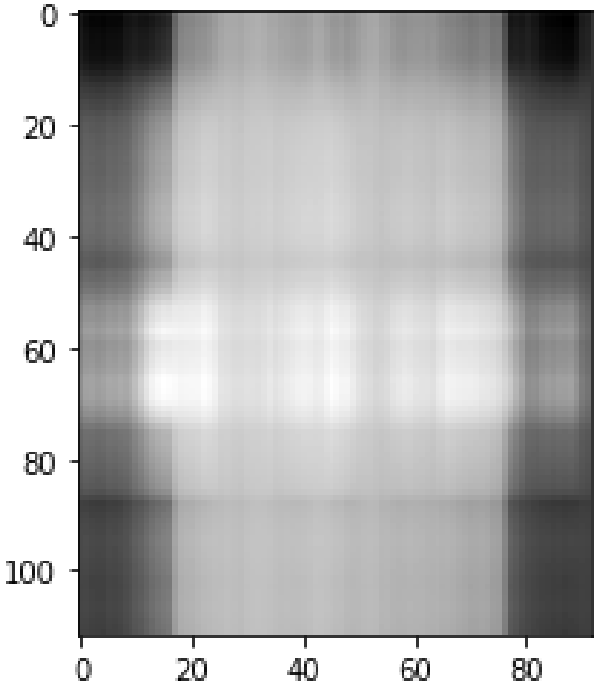
3. **Problem 3**
You are given a set of m objects that is divided into K groups, where the i-th group is of size m_i . If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)
When we randomly select $n\frac{m_i}{m}$ elements from m_i, we guarantee that we have at least one observation from each of our K groups, but if we randomly select n observations, the only way to *guarantee* that we have an observation from each group is if we have m elements. However, that is not allowed because we want $n < m$ elements.

4. **Problem 4**
Download the image hw2_2020_problem4_Face.pgm from the class homework data folder. Find a PCA package and use it to compute eigenvectors and eigenvalues for this image.

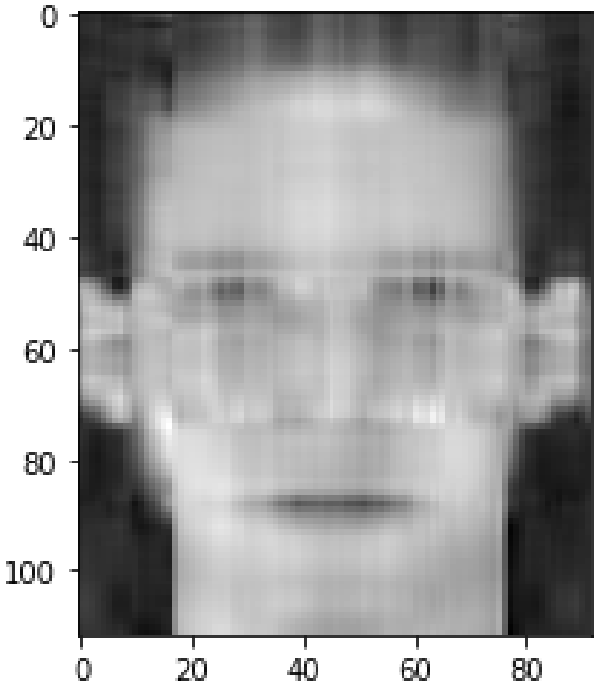
(a) i. Original Image:



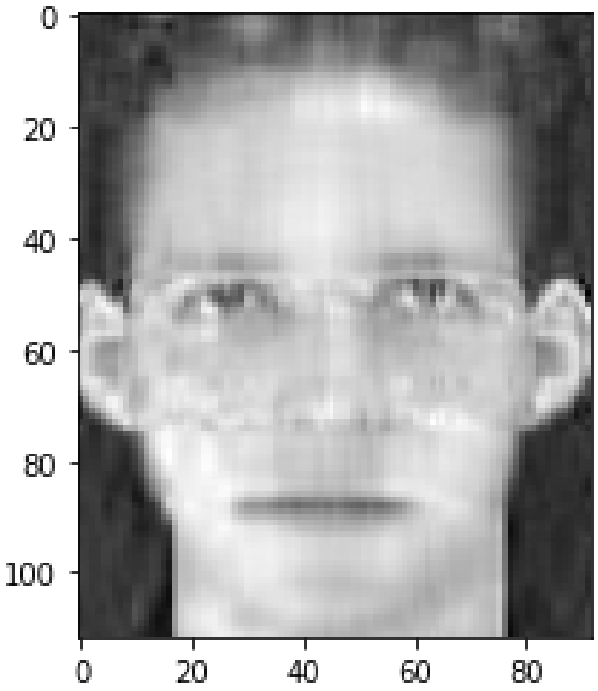
ii. 1 PC Image:



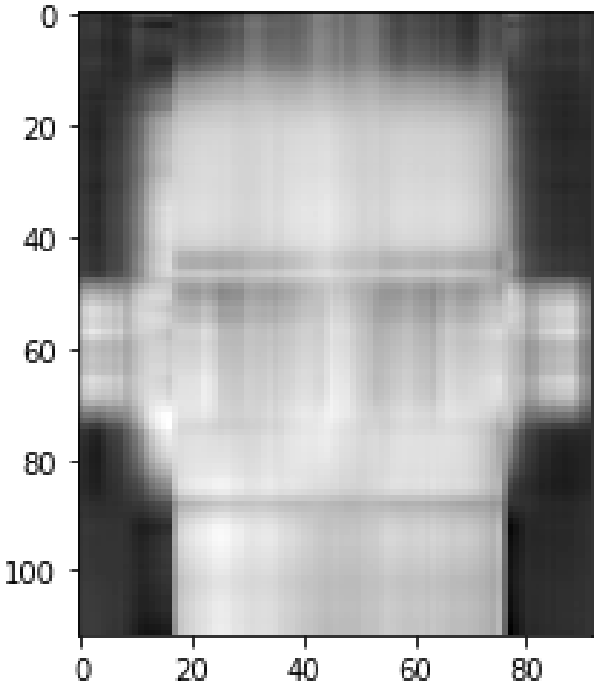
iii. 5 PC Image:



iv. 10 PC Image:



(b) To keep 80% of the variance, we need 3 principal components:



5. **Problem 5**

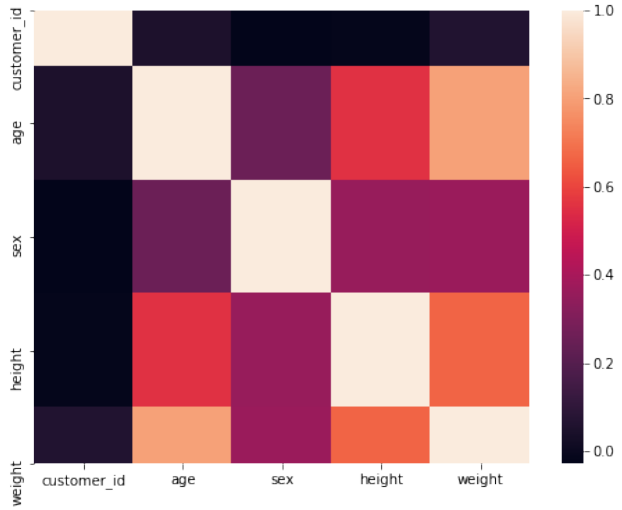
(a)

Column Name	Number of Missing Values
customer_id	0
customer_name	0
age	0
sex	0
height	2976
weight	2817
membership_type	1455

(b) A naive method of handling NaN's in our data is to just remove any columns with NaN values. The benefits are that the program could run more quickly with less observations, and as long as we have enough observations to not significantly reduce the power of the program (which it doesn't in this case since we still have about 2000 observations), our results *should be* consistent to having a larger dataset of full values. That being said, losing observations hurts because variation is information, and these observations with missing data might be valuable information.

(c) I would argue that a better approach would be to impute these values. First, we find the values for *membership_type* because that is "easiest." Which type of member a person is seems to be related to his or her age, which makes sense considering the labels are "kids," "youths," and "adults." After discovering the minimum ages that show up in our initial data set of youth and adults, I was able to relabel those values in our dataset. Now, I imputed the missing values

for the height and weight categories. When considering imputation techniques, I first thought that it would be best to replace the missing values with the averages and continue on with analysis, until I recognized that age, sex, height, and weight are all interrelated according to health professionals. After creating a heatmap of correlations, age, sex, height, and weight were all non un-related, so I thought it would be best to impute these values based on a regression framework.

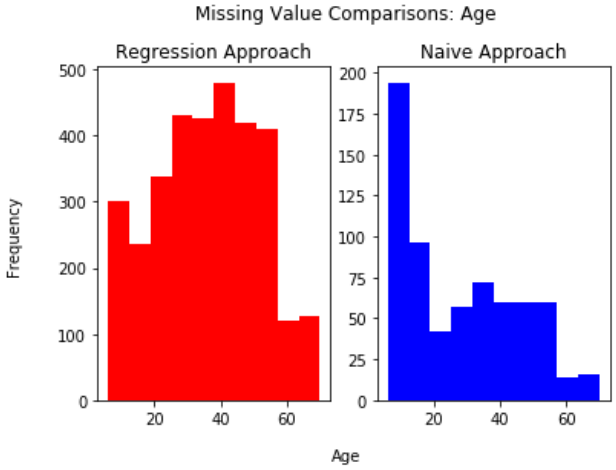


Using the naive method dataset, I discovered parameters for predicting height and weight as follows:

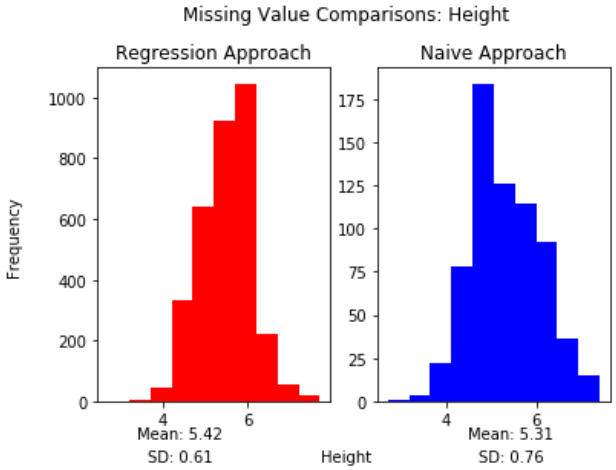
$$\hat{height} = 0.00332973age + 0.24414104sex + 0.00791759weight$$
$$\hat{weight} = 1.9005861age + 13.94502402sex + 19.80550666height$$

These coefficients make sense if you think about it: height increases with age, men are generally heavier than women, and taller people tend to be slightly heavier. Older people are definitely taller than younger, men are on average taller than women, and people who are taller also tend to weigh more. This method comes with one caveat though: if both height and weight are missing, neither are imputable via this regression method and therefore observations with both data missing were removed from out dataset. Using these equations, I was able to impute missing values and retain more data. We now have no missing values.

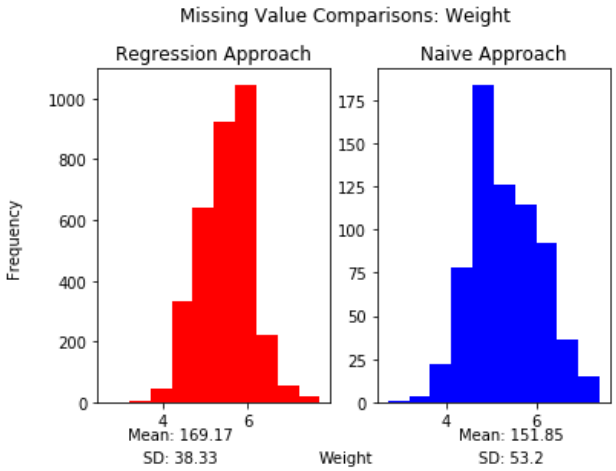
(d) i. Based on the bar plot, the naive imputation is much more right skewed than the regression-based imputation approach.



ii. Based on the bar plot, heights seem to be relatively consistent across imputation methods. Even the means and standard deviations are similar.



iii. The regression-based imputation method yields a tighter fit for weights than dropping values.



iv. Adult males are the most common type of gym customer in the naive approach, while the regression approach has a relatively equal number of adult

men and women. Additionally, both follow a similar trend: the majority of customers are adults - kids and youths are not as common. This makes sense considering the general demographics of a gym.