STAT 112
Introduction to Data Processing
and Visualization

Final Project
January 15, 2024

Databased Review for Healthy Heart: Lifestyle and Its Effects

By
Cemile BİLGİR
Dilara YILDIRIM
Kemal Can YOLOĞLU
Özgür KILIÇ
Yusuf ÖZCAN

Abstract

This research includes 150 randomly selected people with some lifestyles and their effects on their health. These research's data required precise data tidying and cleaning to be used and being interpreted throughout the research questions examining healthy life with heart disease and cholesterol levels effects additionally some daily life routines such as smoking status, exercise durations and physical activities. As a result of our research questions on the clean data, we interpreted correlations between several numerical and categorical data.

Introduction

In this research we have prepared and analyzed a dataset that contains data from 150 randomly selected people with their individual lifestyles. Main goal of this research is examining the relationship between individual habits and health status with respect to heart health. Original dataset contained 150 observations and 13 variables:

variable name	description	scale
ID	Ordered identifiers for people in this research.	nominal
Age	Ages for people in this research.	ratio, discrete
Gender	Gender of people in this research.	nominal
Blood Pressure	Blood pressure levels of people who attend this research.	ordinal, continuous
Cholesterol Level	Cholesterol levels of people in this dataset.	ordinal, continuous
Family History	Hearth disease situations which have seen from their older family members.	nominal
Smoking Status	Smoking activities of people in this research.	nominal
Physical Activity	Level of daily activities of people in this dataset.	ordinal, continuous
BMI	Body mass index of people in this research.	interval, continuous
Glucose Level	Glucose level in these group members' blood.	interval, continuous
Year	Years of data collection for this research.	ratio, discrete
Hearth Disease	Whether or not have hearth disease.	nominal
Exercise Duration	Exercise duration of people in the dataset.	ratio, continuous

Firstly, these unclean dataset's column names are dirty with some punctuation marks also this dataset has duplicated value. Additionally, some string values such as gender, smoking status, physical activity, and heart disease had dirty values with different written types. Also, some integer values were not suitable for that category name for instance there were some negative values for some category and some values which were not suitable for float values. Moreover, there were some NaN values. We overcame all these problems and mentioned below.

Data Tidying and Cleaning

We have done this step by using two tools which are NumPy and Pandas also parting 3 parts: String formatting, numerical formatting and dealing with missing values included outlier controlling.

String Formatting:

- Firstly, we started by importing NumPy and Pandas libraries to Collab and excel files that we are given.
- Then we wanted to be sure that correct importing, so we looked head and tail of dataset.
- Then we examined the variables and their data types by looking info of dataset.
- After mentioned above steps we wanted to start with column names by renaming them.
- Then we looked at duplicates and dropped them.
- Then we moved on column cells, and we cleaned Gender, Blood Pressure, Cholesterol Level, Family History, Smoking Status, Physical Activity and Heart Disease sections. These are some problems which we cleaned:
- Upper case
- Lower case
- Unfinished data which we can understand which one is.
- White space
- Using _ instead -
- After dealing with all mentioned above problems, we passed through to Numerical formatting by exporting our step 1 clean dataset.

Numerical Formatting:

- After importing the new excel file we looked at the head and tail of the dataset to check that imported correct dataset and we saw that there were some miswritten column names which are Exercise Duration and BMI. We have changed them. Additionally, we have noticed that Collab was adding an unnamed column before the ID section. Thus, we deleted it by using '.drop()' function. Also, there are some changes about spelling of some values in Gender and Smoking Status section. We have also dealt with that issue.
- After that process, we started with Exercise Duration section by looking at the values with using '.values()' function.
- After looking at the values we wanted to replace values with using a dictionary inside the parentheses. Furthermore, we have changed the data type of that section float to integer.
- After the previous step, we have looked at the values and we have seen that there were some negative values. To fix that problem, we replaced them with absolute values.

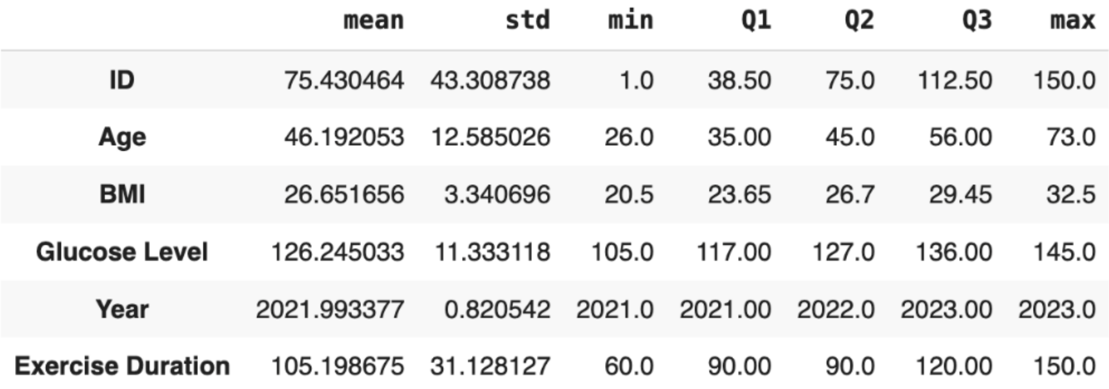
- After the cleaning exercise duration section, we passed other numerical sections but first we looked at the head of the dataset and we saw that age and glucose level sections are float. Thus, we changed them to integers to deal with them easily and in the end, we changed all numeric values to float.
- After dealing with previous sections, we passed to BMI section and first we looked at the values. We have seen that there were some values with using "," to show decimals but it had to be "." For that problem we examined values which were written like this and changed them.
- Before the final step we wanted to change the data type of all numerical variables to float by using '.astype()' function and after that we checked the results by using .info() function.
- After all the mentioned steps, we checked whether any issue exists with any numerical sections and then we exported the current dataset.

NaN Formatting:

- Firstly, we imported the current dataset and we have seen some previous problems which we have already had. Unfortunately, Collab did not save some previous progress, but we have dealt with those problems such as age, exercise duration and glucose level sections datatype and adding unnecessary column by using '.astype()' function and '.drop()' function respectively.
- After doing the mentioned step above and checking head of the dataset, we were not sure that numerical sections' datatype whether integer or float so we used '. astype()' function to change float and then we used .info() function to confirm that issue.
- Before passing thorough to missing value handling we have noticed that we have not controlled any *outlier* on numerical sections so we have controlled step by step all of the numerical values by using '.quantile()' function and other codes which can be seen on GitHub page. Finally, we have seen only one outlier in the exercise duration section, and we replaced an integer which is near the mean value because we wanted to handle with some specific numbers to understand and draw conclusion easily. (Difference is 0.8 and we do not want to round that number up because other values are goes multiple by three.)
- After handling some previous steps' problem, we passed to NaN values by using '.describe().transpose()' and '.isnull().sum()' functions. Then we saw that there were some missing values in one categorical section which is Blood Pressure and 4 numerical sections which are age, BMI, glucose level and exercise duration.
- After examining missing values, first we wanted to start with categorical value by calculating mode value and filling with it.
- After dealing with categorical value, we passed to numerical values by looking at the descriptive statistics table and we wanted to fill NaN values respectively by sections:
- For the age section, mean and mode values were near to each other and we decided to use mode and filled with it.
- For BMI section, again mean and mode values were near, and we have decided to use 26.7 which is coming from rounding mean value which is 26.65.
- For glucose level section, again mean and mode values were near, and we have decided to use mode and filled with it.
- For the exercise duration section, this time we have different mean and mode values and we have decided to use mode value and filled with it.
- After filling NaN values, we wanted to check whether any missing value remained, and we saw none.
- Before the passing final step we wanted to bring out descriptive statistics of numerical variables by using '.transpose()' function and some different codes for the table format we wanted.
- Finally, last time we wanted to look at the first 50 data values and exported the final clean data.

Exploratory Data Analysis

After the cleaning and tidying process, we wanted to look at several research questions by achieving our goal. Questions will be given in every individual step. Before passing through to questions we wanted to explain descriptive statistics of numerical variables that we own from the dataset.

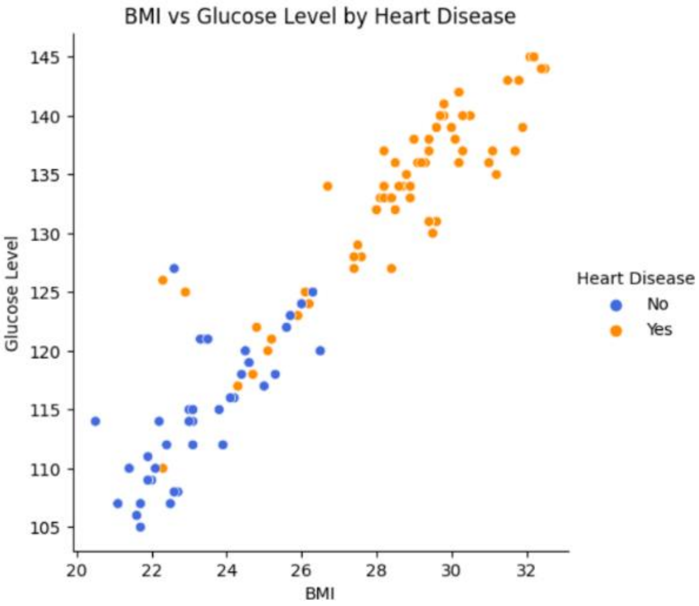


	mean	std	min	Q1	Q2	Q3	max
ID	75.430464	43.308738	1.0	38.50	75.0	112.50	150.0
Age	46.192053	12.585026	26.0	35.00	45.0	56.00	73.0
BMI	26.651656	3.340696	20.5	23.65	26.7	29.45	32.5
Glucose Level	126.245033	11.333118	105.0	117.00	127.0	136.00	145.0
Year	2021.993377	0.820542	2021.0	2021.00	2022.0	2023.00	2023.0
Exercise Duration	105.198675	31.128127	60.0	90.00	90.0	120.00	150.0

By looking at the ID row it can be understood that this research includes data for 150 individuals. The youngest person of the sample is 26 years old while the oldest is 73 years old and the mean age of the sample is 46.2. Standard deviation of body mass indexes is 3.34 kg/m^2 which means nearly 68% of the individuals have body mass indexes within the range of 23.3-30.04 kg/ m^2 while the minimum body mass index is 20.5 and the maximum is 32.5. The standard deviation for glucose level which is 11.3 mg/dL represents a reasonable variability around the mean, but some individuals may have significantly higher or lower glucose level considering the minimum and maximum values for glucose level which are 105 and 145 mg/dL. When the year row is taken into consideration, it is possible to state that this research includes three years these years are 2021, 2022 and 2023. Lastly, the mean of exercise duration is 105.2 minutes while the standard deviation is 31.1 the shortest exercise duration is 60 minutes and the longest is 150 minutes. These descriptive statistics will be interpreted more in depth in the following questions by using other variables such as heart disease status, smoking status, gender, and family history.

1. Is there a relationship between BMI and glucose levels and how do these two variables affect the risk of having heart disease?

In this scatter plot, the graph uses BMI values on the x-axis, glucose level values on the y-axis, and the dots are colored according to the heart disease status. The orange color represents "Yes", and the blue color represents "No".



So, if we analyze the graph, we see a positive association between BMI and glucose level, and generally, as these two variables increase, the rate of having heart disease has also increased. Especially, all individuals with a BMI above 28 and a glucose level above 125 have a heart disease. On the other hand, individuals with low BMI and low glucose levels do not usually have heart disease, although there are a few exceptions (these exceptions may have different genetic causes from BMI and glucose levels). Moreover, in the range where BMI levels are approximately between 24 to 26, and glucose levels are around 115 to 125, individuals with and without heart disease have shown a mixed distribution. Therefore, in this interval there is also a positive relationship between BMI and glucose levels, but we cannot make a comment regarding heart disease.

In conclusion, according to this plot there is a positive association between BMI and glucose level, and individuals with high levels of these two variables tend to have a heart disease.

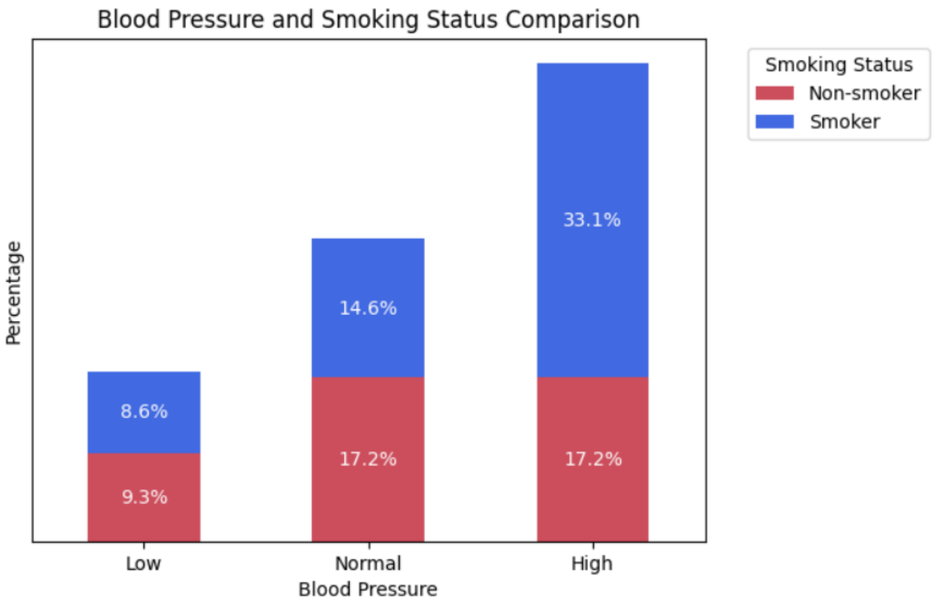
Here is part from academical research about BMI and heart diseases:

Eleven studies with 18,984 subjects were included in this study. The G-2548A (rs12112075), rs7799039, and A19G (rs2167270) polymorphisms of the leptin gene (but not the Lys656Asn (rs1805094) polymorphism) are associated with an increased risk of cardiovascular disease. Our pooled analysis revealed an association between the G-2548A (rs12112075) polymorphism and heart disease, high BMI, and obesity. This indicates that individuals carrying the AA allele are at an increased risk for heart disease, high BMI, and obesity. People with heart failure and coronary artery disease did not have the rs7799039 polymorphism or its alleles linked to them.

[Gene polymorphism of leptin and risk for heart disease, obesity, and high BMI: a systematic review and pooled analysis in adult obese subjects - PubMed (nih.gov))]

This research shows that there is a direct association between BMI and heart disease which supports our findings about relationship between BMI value and heart disease.

2. What is the relationship between blood pressure and smoking status?

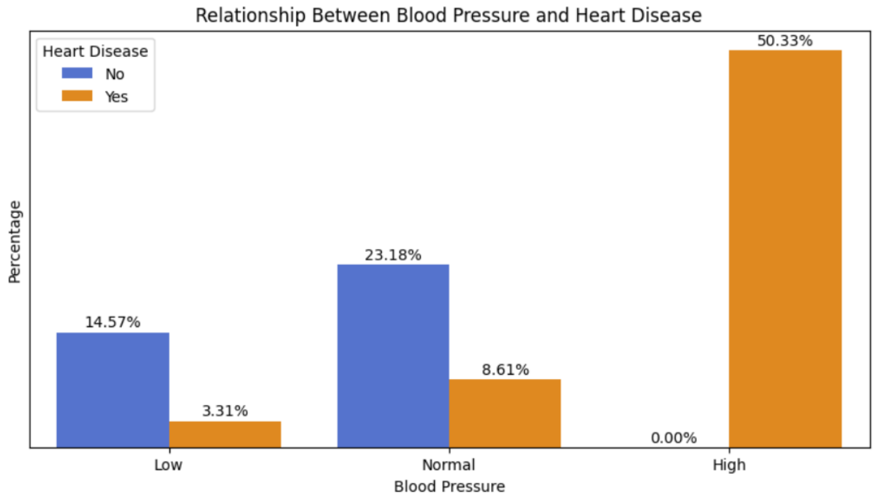


We utilized a stacked bar chart to compare two categorical variables in our dataset, smoking status, and blood pressure.

In this chart, we can observe the distribution of smoker and non-smoker groups for each blood pressure level relative to the total number of individuals. When comparing individuals with low blood pressure, we observe a slight difference (%0.7), with non-smokers being more prevalent. From this, we can interpret that there is no significant impact of smoking status on individuals with low blood pressure. When comparing individuals with normal blood pressure levels, we observe that the percentage of non-smokers is 2.6% higher than smokers.

In this case, we can interpret that the probability of having normal blood pressure increases in the absence of smoking. When examining individuals with high blood pressure, we observe a significant difference, with smokers constituting a markedly higher percentage (15.9%) than non-smokers. It can be clearly stated that the probability of having high blood pressure increases when smoking is present.

We have seen the adverse effects of smoking on blood pressure.
Now, another question we are curious about in our research is whether there is a relationship between blood pressure and heart disease.



We compared the variables of blood pressure and heart disease using a clustered bar chart. Clearly, as the level of blood pressure increases, the probability of having heart disease also increases. Moreover, we observe that all individuals in our study who have high blood pressure are also suffering from heart disease.

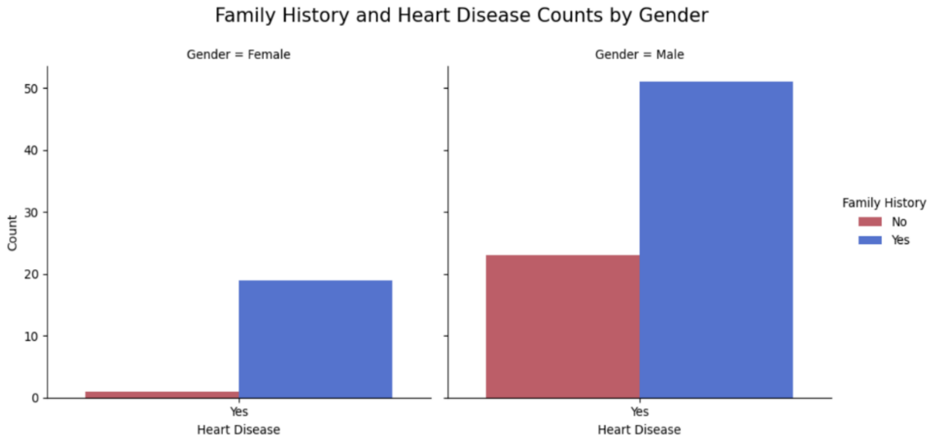
After interpreting the relationship between smoking and blood pressure, we then examined the relationship between blood pressure and heart disease.

Clearly, we observed a distinct connection between them. Therefore, this constitutes significant developments in our research.

3. How does family history affect heart disease in different genders?

In the bar chart, blue bars show people with a family history of heart disease, while pink bars show those without such a history. The first two bars are classified as females, and the last two bars are classified as males.

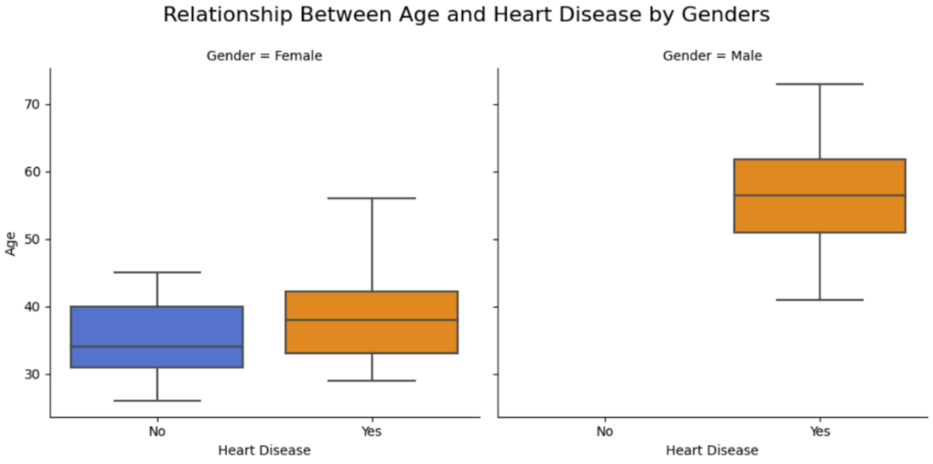
By looking at the count values of the bars in this bar chart, it becomes evident that, in general, males are more predisposed to have heart disease than females, because in both cases (with or without a family history of heart disease), the values in the bars of males are higher than those for females. On the other hand, when we correlate the number of heart diseases with family history, the risk of developing this disease for females increases by approximately 20 times if there is a family history, whereas for males, the increase is only 2 times.



Considering this evaluation, we can say that although heart diseases are more common in males than females, a family history of heart disease affects females more than males. In other words, when a male and a female with a family history of heart disease are compared (all other factors are assumed to be equal), the female is more likely to have heart disease.

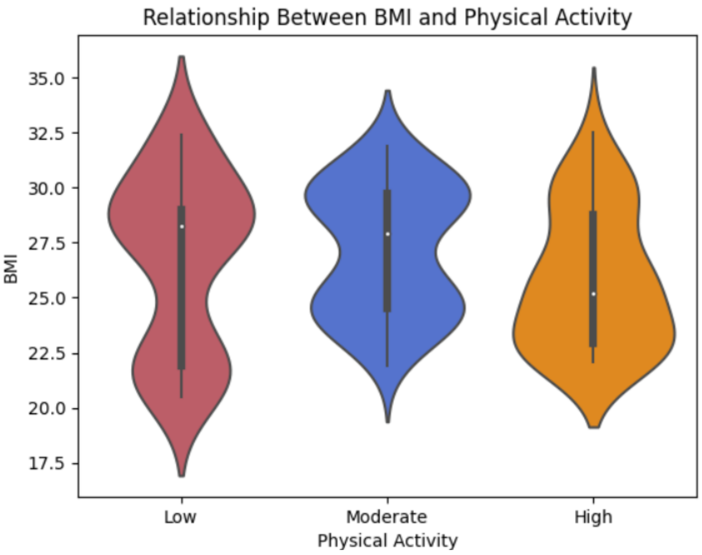
4. How is the distribution of ages by gender and heart disease status?

When the box plot for males is examined, it is possible to state that among males that the research is conducted there is not a single male who does not suffer from heart disease, which seems uncommon. One of the reasons for this could be about the target audience of this research for males, for instance, the research may be conducted on only who suffers from heart disease for males. On the other hand, unlike males, in the research there are females who suffer from heart disease and who does not.



There is a lot of variability in age for females and males because boxes are quite wide, and whiskers cover a long length. If we look at the plot for females median age for females with heart disease is higher than median age for females without heart disease so it can be said that females with heart disease tend to be older than females without heart disease so age could be one of the factors for heart disease. If we compare males and females who suffer from heart disease, we can see that males who suffer from heart disease tend to be older than females who suffer from heart disease.

5. What is the relationship between BMI and physical activity?



This violin plot compares the body mass index distributions of individuals with different levels of physical activity which are low, moderate, high. It is known that the width of the violins indicates the spread of BMI distribution among individuals at that level of physical activity.

When analyzing the low physical activity group, it can be observed that BMI values show a broader range compared to the other groups. Also, the upper section of the violin plot is wider than the lower section, it can be interpreted that individuals with low physical activity generally tend to have higher BMI values. The median value of BMI for a low physical activity group is between 27.5 and 30. When individuals with moderate and high levels of physical activity are examined, it is seen that the range of BMI values is approximately the same. When examining the violin plot of the moderate physical activity group, it is observed that the width of the lower and upper sections is nearly equal. Therefore, that indicates a more balanced distribution of BMI levels. The median value of BMI for this group is very close to that of the low physical activity group and approximately 27.5.

Finally, when we examine the group with high physical activity, the lower section of the violin plot is wider than the upper section. Therefore, we can infer that individuals with high physical activity generally tend to have lower BMI values. Also, the median value of BMI for this final group is almost equal to 25.

As a result, we can conclude that individuals with high levels of physical activity tend to have lower BMI values, indicating that they have a more ideal weight. This contributes to a more favorable health condition.

Conclusion

To sum up, in this research 5 questions were asked to examine the relationship between individual habits and health status with respect to heart health. It is understood that individuals who has BMI value above 28 kg/ m^2 and glucose level above 125 mg/dL tend to have heart disease so, high BMI and glucose levels increases the risk of heart disease. We examined the relationship between blood pressure and smoking status followed by relationship between blood pressure and heart disease, we saw that smoking increases the probability of having a high blood pressure while high blood pressure increases the probability of having heart disease. So, we understood that smoking can increase blood pressure and affect heart health. Then, affect of the family history on heart disease is examined and it concluded that family history is a critical factor for heart disease and nearly all of the females who suffer from heart disease has a family history. Distribution of ages by gender and heart disease is examined and it is understood that variability in age is wide for both genders, older females tend to have a heart disease compared to older ones while examining the distribution we have noticed that among males that the research is conducted there is not a single male who does not suffer from heart disease which is interesting. Finally, we have examined the relationship between BMI level and physical activity to conclude our research and we understood that individuals with high levels of physical activity tend to have lower BMI values which contributes to heart health status.

References

Khaki-Khatibi, F., Shademan, B., Gholikhani-Darbroud, R., Nourazarian, A., Radagdam, S., & Porzour, M. (2022). Gene polymorphism of leptin and risk for heart disease, obesity, and high BMI: a systematic review and pooled analysis in adult obese subjects. *Hormone molecular biology and clinical investigation*, 44(1), 11–20. https://doi.org/10.1515/hmbci-2022-0020