Group Members: Sabah Javaid, Anna M. Kobaivanov, Alessandro Cuadros, Kevin Linitz and Han Kim

The Analysis of Cardiovascular Disease:

Analyzing Cholesterol & Glucose Against Other Health Related Factors

Import Data to Clean

Before we could use the analytical abilities that we gained through the course, we first collectively looked for a dataset that captured the interest of all group members. Through Kaggle API, we came across a Cardiovascular Disease dataset. In order to initiate the data cleanup process, we used the pd.read_csv tool to read the online source. We then familiarize ourselves with the data to gain scope of what we intended to adjust and/or remove in the cleanup process.

Data Cleanup

- We checked for null and NaNs values in rows. The dataset did not contain any thus we did not have removed any NaNs or null values.
- We analyzed the dataset with 70,000 rows to find mistakes/inconsistencies in the data, and then found and cleaned the following:
    1) "ap_hi" or systolic blood pressure >= 371. We dropped 39 rows to clean the dataset from this inconsistency. We were left with 69,961 rows.
    2) “ap_lo” or diastolic blood pressure >= 361. We dropped 953 rows to clean the dataset from this inconsistency. We were left with 69,008 rows.
       The highest blood pressure recorded is 370/360. These values were not recorded properly, indeed 180/120 is already a hypertensive crisis and can be conducive to a stroke.
    3) Negative "ap_hi" or systolic blood pressure. We dropped 7 rows and were left with 60,001 rows.
    4) Negative "ap_lo" or diastolic blood pressure. We dropped 16 rows and were left with 68,985.
    5) Found "ap_lo" > "ap_hi": systolic blood pressure cannot be higher than diastolic blood pressure. We dropped 274 rows with this issue and were left with 68,711.
- We dropped the ID column.

Data Preparation

- We created a column named “BMI” using “weight” and “height” columns.
- We converted “gender,” “cholesterol,” and “glucose” to string.
- We created dummy variables for “cholesterol,” and “glucose” and concatenated them in new columns.
- Since the dataset contains “age” in days, we converted this column into float and the values to years.
- We renamed columns “ap_hi” to “systolic_BP,” “ap_lo” to “diastolic_BP,” “alco” to “Alcohol,” “active” to “physical_activity,” and “cardio” to “cardio_disease” for easier reading and analysis.
- We created two additional columns, first creating a function with a condition, and adding it to the dataset as a new column:
    1) “blood_pressure:” which indicates whether the patient has a healthy, elevated, or suffers hypertension based on their systolic and diastolic readings.

2) "BMI_result:" which indicates whether the patient is underweight, normal, or overweight.

Data Analysis, Statistics and Findings

Goal: Understand the sample of patients (Sample and Habits Analysis) that provided information by running statistics of gender, age, and habits such as alcohol intake, whether they are smokers, and whether they exercise and how this is related to diverse health conditions (Health Conditions Analysis) such as glucose, cholesterol, blood pressure, body mass, and finally whether these health conditions are conducive to cardiovascular disease.

- **Sample and Habits Analysis:** We asked the following questions to understand the data sample and found the following results:
  - What is the average age, BMI, height, weight, systolic and diastolic blood pressure by gender?
    The average age of patients is 53 and there is no difference in average by gender. The average BMI for female patients is 1 unit higher than the average for male patients. Since BMI is a function of weight and height, we can conclude that this result is due to male patients being on average 8 cm taller than female patients, even though male patients are 5 kg heavier than female patients. There is no significant difference between the average blood pressure of the patients by gender; however, women are 3 units lower on average for systolic blood pressure than men and 2 units lower on average for diastolic.
  - How many female and male patients are in the clean dataset?
    The information provided is from 44,758 (65%) female and 23,953 (35%) male patients.
  - How many male and female patients have some physical activity, smoke, drink and suffer from a cardiovascular disease (habits df)?
    - ➔ We observe that out of the 44,758 female patients 35,881 have some physical activity, 794 smoke, 1,128 drink alcohol, and 22,025 suffer from cardiovascular disease.
    - ➔ We observe that out of the 23,953 male patients 19,320 have some physical activity, 5,247 smoke, 2,537 drink alcohol, and 11,967 suffer from cardiovascular disease.
    - ➔ We found that there is no difference in terms of physical activity between both genders. Both, most of male and female, have some sort of physical activity (80%). However, bad habits such as smoking (22% males but only 1% females) and drinking (10% males but only 2.5% females) are strongly more predominant in males. Even though males in this sample have strongly more predominant bad habits, there is no significant difference in terms of cardiovascular disease since almost the same percentage of the sample have been affected by cardiovascular disease (males at 50% and females at 49%).
  - What is the average BMI per cholesterol level?
    While the average BMI for men with high cholesterol is 28 and for abnormal cholesterol is 27, for women the average BMI for high cholesterol is 30 and abnormal is 28. It seems that due to its physiological nature females have a higher BMI on average compared to males.

- **Health Conditions Analysis:** We analyzed diverse grouped health conditions and habits, and found the following interesting facts:
    1) Does smoking or drinking affect glucose and cholesterol levels?
       Surprisingly, smoking does not affect glucose or cholesterol levels significantly. Most men and women that smoke have normal cholesterol and glucose. However, it seems smoking is more harmful for females since 85% of men that smoke have normal glucose levels while this percentage is lower at 80% for females. Similarly, only 65% of women that smoke had normal cholesterol levels while this percentage was higher at 74% for smoker males.
       Regarding alcohol consumption, it affects a bit more than smoking and worsens cholesterol more than glucose levels. 84% of male and 79% of female drinkers have normal glucose levels while only 70% of male and 60% of female drinkers have normal cholesterol levels.
    2) Does smoking or drinking lead to cardiovascular disease?
       Somewhat, 47% of people that smoke suffer from cardiovascular disease while 48% of people that drink alcohol suffer from cardiovascular disease.
    3) Does cholesterol make you prone to cardiovascular disease / Do people with cholesterol suffer from cardiovascular disease?
       Somewhat. As we saw, cardiovascular disease affects both genders almost equally (50% males and 49% females). Nevertheless, out of these people suffering from cardiovascular disease only 35% of females and 31% of males have high cholesterol (about ⅓). This proves that the majority, regardless of gender, have normal levels of cholesterol. This raises the question of whether cardiovascular disease might be more due to a congenital issue (hereditary).
    4) Do people with physical activity have lower cholesterol?
       This could be true. Most male (77%) and females (74%) that engage in physical activity have normal cholesterol. This means that exercise could help in normalizing cholesterol levels.
    5) Do people with physical activity tend to be less prone to cardiovascular disease (1) / Do people with cardiovascular disease work out (2)?
        (1) Possibly, yes but the difference is minimal. Out of the 55,201 people that work out 48.5% have a cardiovascular disease while 51.5% do not. Nevertheless, more conclusive analysis is necessary to understand the significance.
        (2) Shockingly, out of all the people who suffer from cardiovascular disease, 79% of them have some physical activity.
    6) Do overweight people suffer from cardiovascular disease? (Look for overweight people by gender)
       Yes, 71% of people who suffer cardiovascular disease are overweight. If we see it by gender, we find that this number is 72% for females and 69% for males. Based on this finding, **being overweight seems to be a causing factor for cardiovascular disease.**
    7) Do people with high blood pressure suffer from cardiovascular disease? (Look for blood pressure people by gender)

Yes, 69% of people with hypertension suffer from cardiovascular disease. If we observe by gender, we find that this number is 63% for females and 66% for males. Based on this finding, **having hypertension can strongly lead to cardiovascular disease.**

The strongest correlations found are:

- Glucose high and cholesterol high = .48
- cardiovascular disease and high cholesterol = .19
- Age and cardiovascular disease = .239
- Blood pressure and cardiovascular disease -> systolic = .43 and diastolic = .34
- People that smoke and drink .34

In addition to prior grouping analysis, we can see that correlations are not strong enough to be as conclusive. However, people that have high glucose are more prone to have high cholesterol. Cardiovascular disease is somewhat related to cholesterol, as in prior grouping analysis we saw that ⅓ of the people with cardiovascular disease have high cholesterol. Age is also slightly related to cardiovascular disease. Blood pressure correlation with cardiovascular disease confirms the findings from grouping analysis that most people with cardiovascular disease have hypertension. Furthermore, having bad habits throws a slight correlation which means that people that smoke tend to drink. However, smoking doesn't affect cholesterol or glucose while drinking might affect cholesterol which could lead to cardiovascular disease. Lastly, physical activity can lower cholesterol but not necessarily avoid cardiovascular disease.

Visualization Analysis

Distribution Plot

In agreement with our statistical analysis, the Distribution Plot shows that most of the population in this dataset fall into the 50-55 age bins.

Hex Plot

The Hex Plot visualization shows us that the majority of the population's weights fall just under 75 kg or 165 pounds.

Gender and BMI Bar Plot

The bar plot visualization shows us BMI by gender. There is no significant difference between the two genders, but as found in the grouping analysis there is a slight increase in BMI in the female population of this dataset.

Box Plot and Violin Plot

The Box Plot shows us that "high" cholesterol is more common for older people, but the "normal" and "abnormal" median fall right under 54 years old.

The Violin Plot is similar to the box plot, but it helps show where the majority lies by analyzing the thickness of the violin. This plot also shows there are more outliers in "normal cholesterol" in younger individuals (35 and below) when compared to "high" and "abnormal" cholesterol levels.

2 Strip Plots

The first strip plot shows the diastolic BP levels of males and females. This helps show that a majority of diastolic BP levels fall within the 75-90 range and that women appear to have slightly lower diastolic BP levels when compared to men.

The second strip plot shows the systolic BP levels of males and females. This helps show that a majority of systolic BP levels fall within the 100-140 range and that women appear to have slightly lower systolic BP levels when compared to men.

Correlation Heat Map

This visualization shows us the correlation of our data in a heat map format. The darker the square and closer to 1, the higher the correlation between the two data points. An example of high correlation is "BMI" and "weight".

Source of Dataset

Since the majority of the group was most interested in analyzing data relevant to the healthcare field, we were able to focus our search for a dataset significantly. Although even through narrowing down our search, we continued our search through the multitudinous datasets on Kaggle and selected the Cardiovascular Disease dataset.

- Why did you pick the dataset(s)?

We collectively chose this dataset due to our common interest in health. When we found this dataset, we all had the same questions and wanted to understand health conditions, habits, and possible drivers for cardiovascular disease better. We decided to analyze the effects of cholesterol, glucose levels, BMI, and blood pressure as well as other health related habits such as smoking, drinking, and physical activity to see if we could find evidence on whether these health conditions and habits triggered cardiovascular disease.

- Information about our **database** and who can use the data:

Once data was cleaned and columns were added and defined in the data frame, an SQLite database was created to export the dataset to a user-created table. The table, *'cardio_data'*, features 20 columns. Each column has its own data type defined and, to protect data integrity, a primary key was set to avoid duplicates. Below is a convenient database dictionary which aims to make the table and its data more readable:

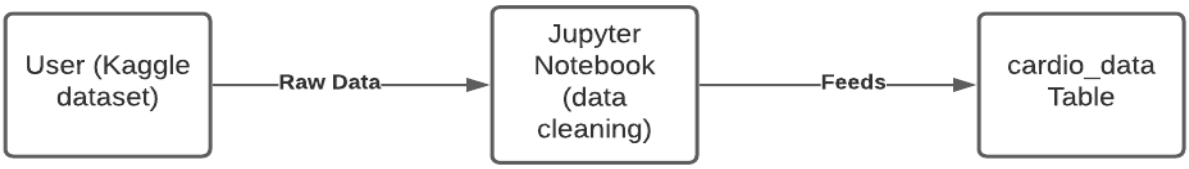
Column	**Data Type**	**Description**
index	integer	Unique identifier for patient record.
age	floating	Patient's age (in years).
gender	character (10)	Patient's gender.

height	integer	Patient's height (in centimeters).
weight	integer	Patient's weight (in kilograms).
systolic_BP	integer	Patient's systolic blood pressure.
diastolic_BP	integer	Patient's diastolic blood pressure.
cholesterol	character (15)	Patient's cholesterol level summary.
gluc	character (15)	Patient's glucose level summary.
smoke	integer	Patient's smoke flag (0 = non-smoker, 1 = smoker).
Alcohol	integer	Patient's alcohol intake flag (0 = no alcohol, 1 = some or abnormal alcohol intake).
physical_activity	integer	Patient's physical activity flag (0 = sedentary, 1 = active).
cardio_disease	integer	Patient's cardiovascular disease flag (0 = no cardiovascular disease, 1 = cardiovascular disease present).
BMI	floating	Patient's BMI (Body Mass Index).
chol_High	integer	Patient's high cholesterol flag (0 = not high, 1 = high).
chol_Normal	integer	Patient's normal cholesterol flag (0 = not normal, 1 = normal).

blood_pressure	character (20)	Patient's blood pressure summary.
BMI_result	character (20)	Patient's BMI level.

SQLite databases are files. When a person that is also skillful with python/database gains access to the machine, where database files reside, they can also further gain access to the database. Since the data does not include any sensitive information (such as names, phone numbers, etc.), we concluded that there was no need to encrypt the database file and risk users running into performance issues. The database and its table serve as an educational source of information that can be used for statistical and reporting purposes.

- Data **flow diagram**



Team Roles

In effort to strive for efficiency, accountability and maximize insight from our analysis, we split the tasks of the project as listed below:

Sabah Javaid	→	Step 1: Import the data to clean
Sabah Javaid and Anna Kobaivanov	→	Step 2: Cleanup & Data Preparation
Anna Kobaivanov	→	Step 3: Statistical analysis (Data analysis + findings)
Kevin Linitz	→	Step 4: Visualization
Alessandro Cuadros	→	Step 5: Export the clean data to a database

Though these were each of our contributions and responsibilities regarding the project, we also worked collectively to overcome any challenges throughout the project in effort to complete our analysis, ensure accuracy in our results and generally fulfill assignment requirements as a team.

- What were the **challenges** you faced?

Some of the few challenges we faced occurred during the cleanup and statistical analysis portions of our analysis. When cleaning our data, our initial cleanup code dropped a significant number of rows from our data however, after running the code line by line and making a few adjustments we found that the issue was in the way in which we set the blood pressure range we wanted to focus on. In our initial code, we intended to drop systolic blood pressure (ap_hi) that exceeded or were equal to 400. We then adjusted our code to drop systolic blood pressure (ap_hi) that exceeded or were equal to 371 and drop diastolic blood pressure (ap_lo) levels that exceeded

or equaled 361. By doing this adjustment, we were able to accurately clean our dataset without losing excessive rows of data in the cleanup process.

We also faced difficulty when proceeding with our statistical analysis. We knew we wanted to incorporate percentages into our findings to discuss and display our results more simply and effectively. Though we weren't well versed in dealing with percentages, we did some external research and attended office hours for 1:1 guidance thereby overcoming this challenge.

Lastly, familiarizing ourselves with the seaborn python library to generate the plots for this project was challenging in and of itself. Several hours of studying the functionalities of the seaborn library were required in the form of taking multiple Datacamp modules as well as utilizing the seaborn cheat sheet on blackboard. While initially challenging, learning the vast array of functionalities within this library became enjoyable after seeing the level of detail and aesthetic appeal of how the plots were being visualized. We specifically enjoyed producing the correlation heat map. With just a few lines of code, we were able to see the correlation of all the data points visualized in one heatmap.

- **Topics** you had to research?

While confirming that the cardiovascular disease dataset was indeed the dataset we would collectively analyze, we researched cardiovascular health and blood pressure reading heavily to guide us in how we would approach our analysis from understanding the impact it has when diagnosed with COVID (Mayo Clinic & Oxford Academic) to understanding the complexity of the disease itself (Mayo Clinic).

Since our data was relatively extensive and required a significant amount of cleanup, analysis, and visualization for our findings, we externally researched how to overcome challenges in analytical steps. An example of this was figuring out how to increase the size of the plot. When producing the plots, it was difficult to view all the data that the plot was attempting to visualize. The solution to this was discovering the fact that running the matplotlib code plt.figure(figsize=(*width*, *height*)) will increase the size of the width and height in inches by inputting the desired plot size by using float values. (8,8) and (10,10) were commonly used in our plots which offered an improved visualization.

- How else can you **expand** your project?

Though we understand that this was a great opportunity to merge cardiovascular disease data with other health related datasets, we decided to analyze the effects of cardiovascular disease alone to acquire a baseline understanding of where cardiovascular health currently lies based on our dataset. By analyzing the Cardiovascular Disease dataset, we immerse ourselves to vast possibilities in which we can expand our analysis through quantifying and contextualizing data into information that is easier to comprehend to the average non-technical person.

However, if we were to expand our project, we could merge our dataset and analysis with many other datasets to gain further insight on the impact of Cardiovascular health. For instance:

1. We could merge our dataset against a COVID-19 dataset to analyze the direct impact of being tested positive for COVID-19 would have against cardiovascular health and its long-term effects.
2. We could also further our analysis of the Cardiovascular Disease dataset by acquiring and incorporating data on family medical history. In this case, we would analyze whether hereditary factors correlate to or impact cardiovascular health.

3. We could also analyze and expand our project based on our findings such as the relationship between being overweight and cardiovascular problems. We could create a binary variable for weight and run a correlation analysis.

References

Data/Python References:(site you researched to learn how to perform a task)

1. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
2. https://docs.python.org/3/library/sqlite3.html
3. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.figure.html

Cardiovascular and Blood Pressure Research Related References:

1. https://academic.oup.com/cardiovascres/advance-article/doi/10.1093/cvr/cvab298/6370961
2. https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-long-term-effects/art-20490351
3. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118
4. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/english_bmi_calculator/bmi_calculator.html
5. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings