

A PYTHON PROJECT



Cardiovascular diseases and
other types of illnesses

TEAM: TOBACCO KILLS
MEMBER: PHAM THU TRANG



Table of Contents

01

INTRODUCTION

02

DATASET

03

**IN-DEPTH
DATA ANALYSIS**

04

VISUALIZATION

05

**LOGISTIC
REGRESSION**

06

CONCLUSION



01

INTRODUCTION

- This project aims to **study the connection between different types of illnesses (including cardiovascular disease) biological information and more important, living habits.**
- Health-related problems are very relevant to our lives and there are tons of bad living habits that would lead to different types of illnesses in the long run.
- **Main issues:**
 - Sex & Age to Heart Disease
 - Smoking & Alcohol Drinking to Heart Disease
 - BMI to Heart Disease, Stroke, Diabetes, Kidney

02

DATASET

Step 1:

- No NaN values
- 319 795 participants
- 18 variables

Step 2: Remove 5 variables

- Physical Health
- Mental Health
- Race
- Walking Difficulty
- Skin Cancer

Independent variables		Dependent variables
Information	Living habits	Illnesses
Age	Smoking	Heart Disease
Sex	Alcohol Consumption	Stroke
BMI	Physical Activity	Kidney Disease
	Sleep Time	Asthma
		General Health
		Diabetes

02

DATASET

Step 3: Create new variables

Life Style:

- 4. Healthy + Physical Activity = Yes Sleep Time = 7-9 hrs
- 3. Healthy - Physical Activity = No Sleep Time = 7-9 hrs
- 2. Unhealthy - Physical Activity = Yes Sleep Time != 7-9 hrs
- 1. Unhealthy + Physical Activity = No Sleep Time != 7-9 hrs

Age (groups):

- 1. 25 - 34
- 2. 25 - 44
- 3. 45 - 54
- 4. 55 - 64
- 5. 65 - 79
- 6. 80+

BMI groups:

- 1. Underweight BMI < 18.5
- 2. Healthy weight BMI in [18.5, 25)
- 3. Overweight BMI in [25, 30)
- 4. Obesity BMI > 30

Independent variables		Dependent variables
Information	Living habits	Illnesses
Age	Smoking	Heart Disease
Sex	Alcohol Consumption	Stroke
BMI	Physical Activity	Kidney Disease
BMI groups	Sleep Time	Asthma
	Life Style	General Health
		Diabetes

03

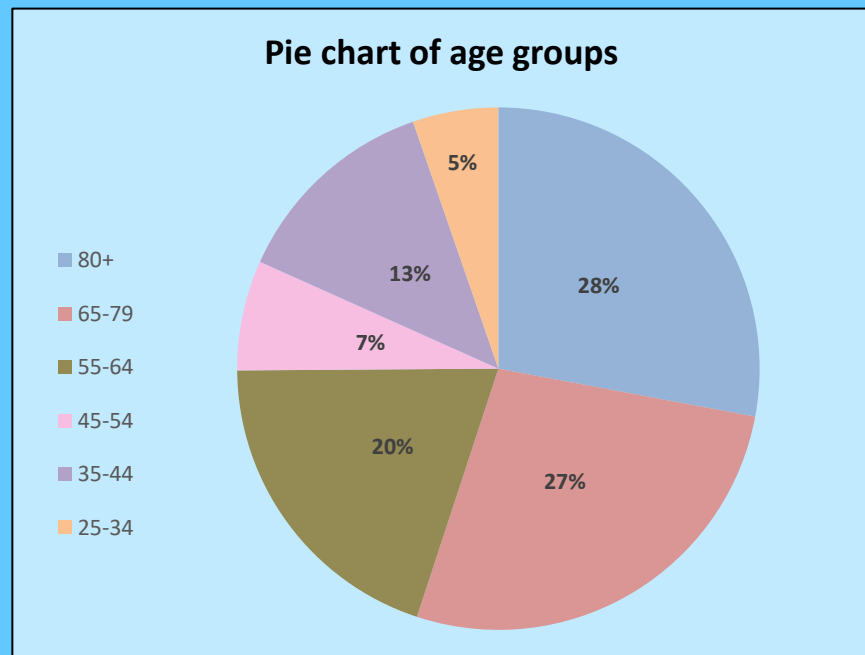
IN-DEPTH DATA ANALYSIS



Sex	Number of people	%
Female	167805	52.472678
Male	151990	47.527322

BMI group	Number of people	%
Overweight	114512	35.807939
Obesity	102842	32.158727
Healthy weight	97331	30.435435
Underweight	5110	1.597899

About 75% of participants are above 54 years old.
More than half of participants are above 65 years old.



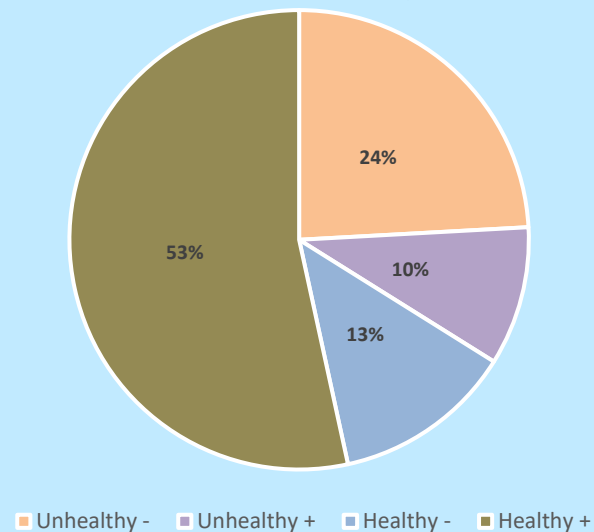
03

IN-DEPTH DATA ANALYSIS

Smoking	Number of people	%
No	187887	58.752326
Yes	131908	41.247674

Alcohol	Number of people	%
No	298018	93.190325
Yes	21777	6.809675

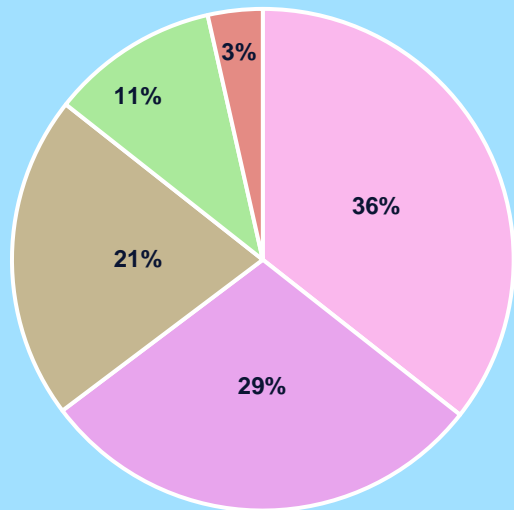
Pie chart of life style



03

IN-DEPTH DATA ANALYSIS

Pie chart of general health



Very good Good Excellent Fair Poor

Heart Disease	Number of people	%
No	292422	91.440454
Yes	27373	8.559546

Stroke	Number of people	%
No	307726	96.22602
Yes	12069	3.77398

Asthma	Number of people	%
No	276923	86.593912
Yes	42872	13.406088

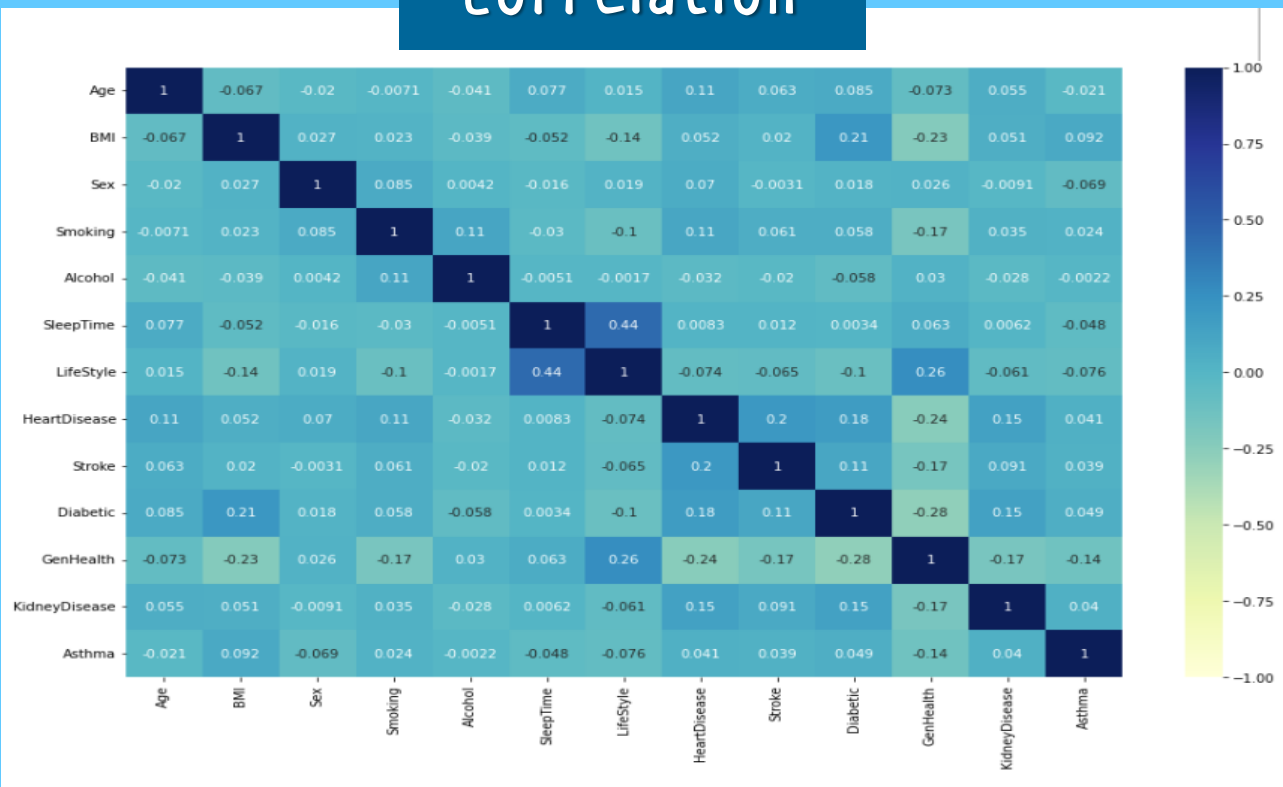
Diabetic	Number of people	%
No	272212	85.120781
Yes	40802	12.758799
Prediabetes	6781	2.120421

Kidney Disease	Number of people	%
No	308016	96.316703
Yes	11779	3.683297

03

IN-DEPTH DATA ANALYSIS

correlation



03

IN-DEPTH DATA ANALYSIS

correlation

There is **no significant correlations** between the variables being studied

- **Life Style and General Health:** weak positive correlation coefficient (+0.26)
- **BMI and Diabetic:** weak positive correlation coefficient (+0.21)
- **BMI and General Health:** weak negative correlation coefficient (-0.23)
- **Heart Disease and Stroke:** weak positive correlation coefficient (+0.2)
- **Heart Disease, Kidney Disease and Diabetic** all share weak positive correlations with each other.

Since most the dependent variables are not continuous, the Pearson correlation can be used to propose hypotheses, but not to draw concrete conclusions.

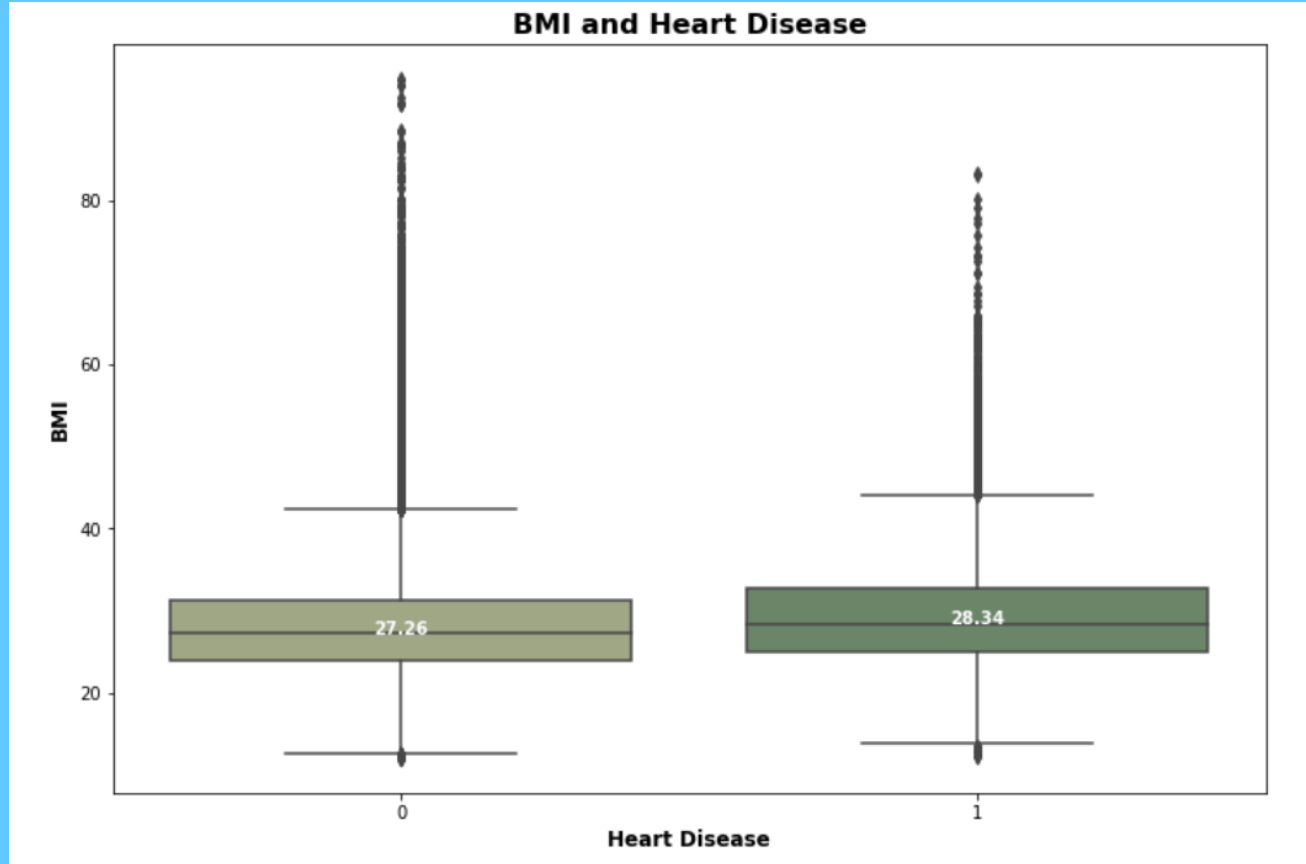
04

VISUALIZATION

1



BMI of people who 'ARE DIAGNOSED' with heart disease are **slightly higher** than those of people who 'ARE NOT DIAGNOSED' with heart disease



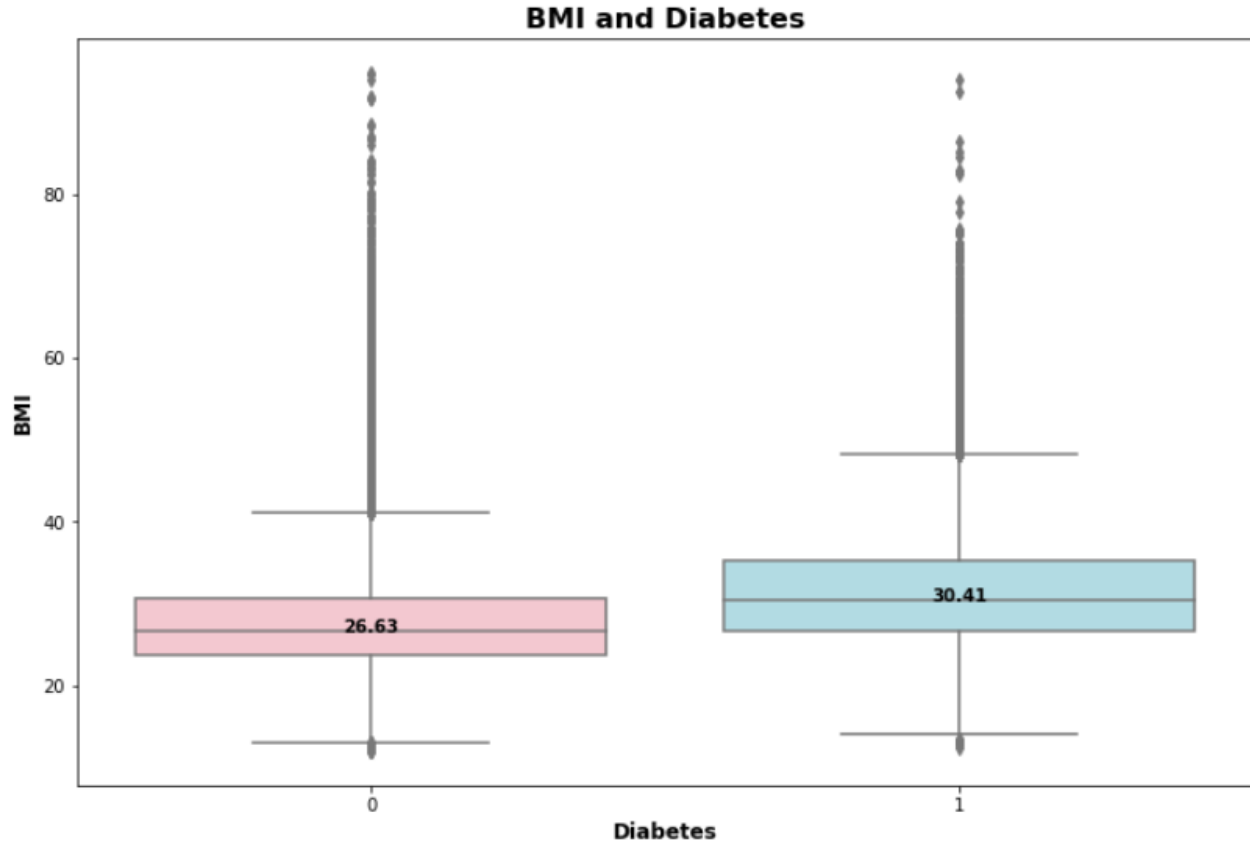
04

VISUALIZATION

2



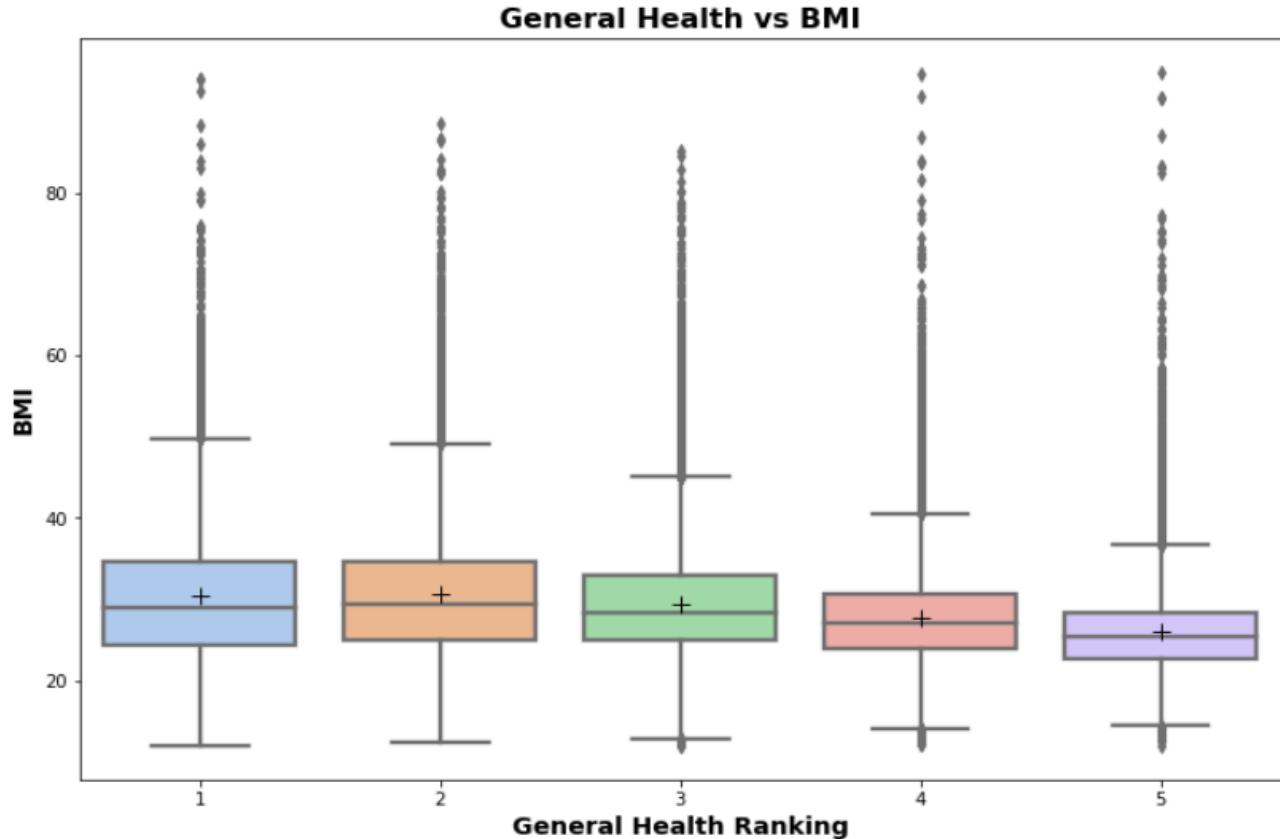
BMI of people who 'ARE DIAGNOSED' with diabetes are **significantly higher** than those of people who 'ARE NOT DIAGNOSED' with diabetes



04

VISUALIZATION

3

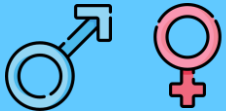


BMI of participants experience **downward trend** as the level of General Health go up.

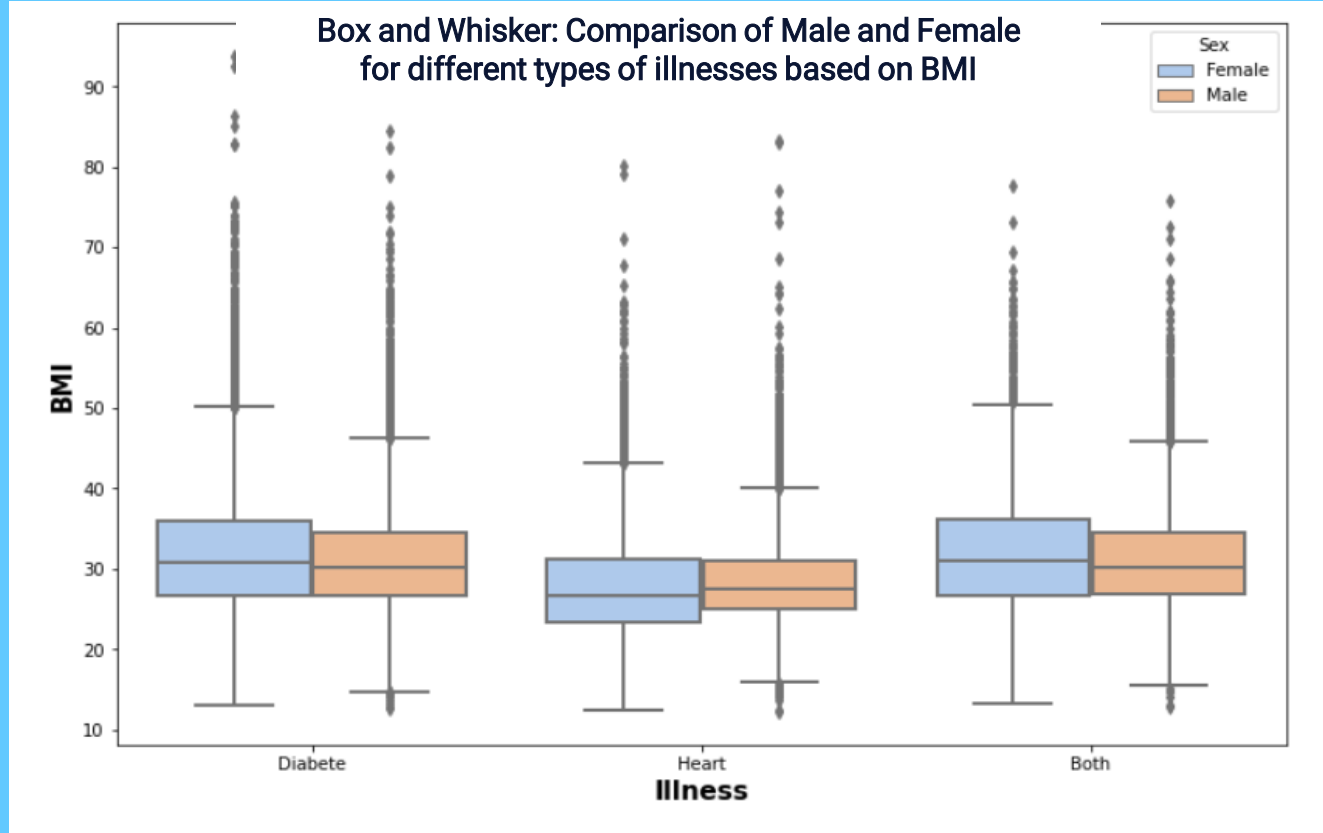
04

VISUALIZATION

4



On average, there is **no significant difference** between BMI index of men and women for diabetes, heart disease or both.

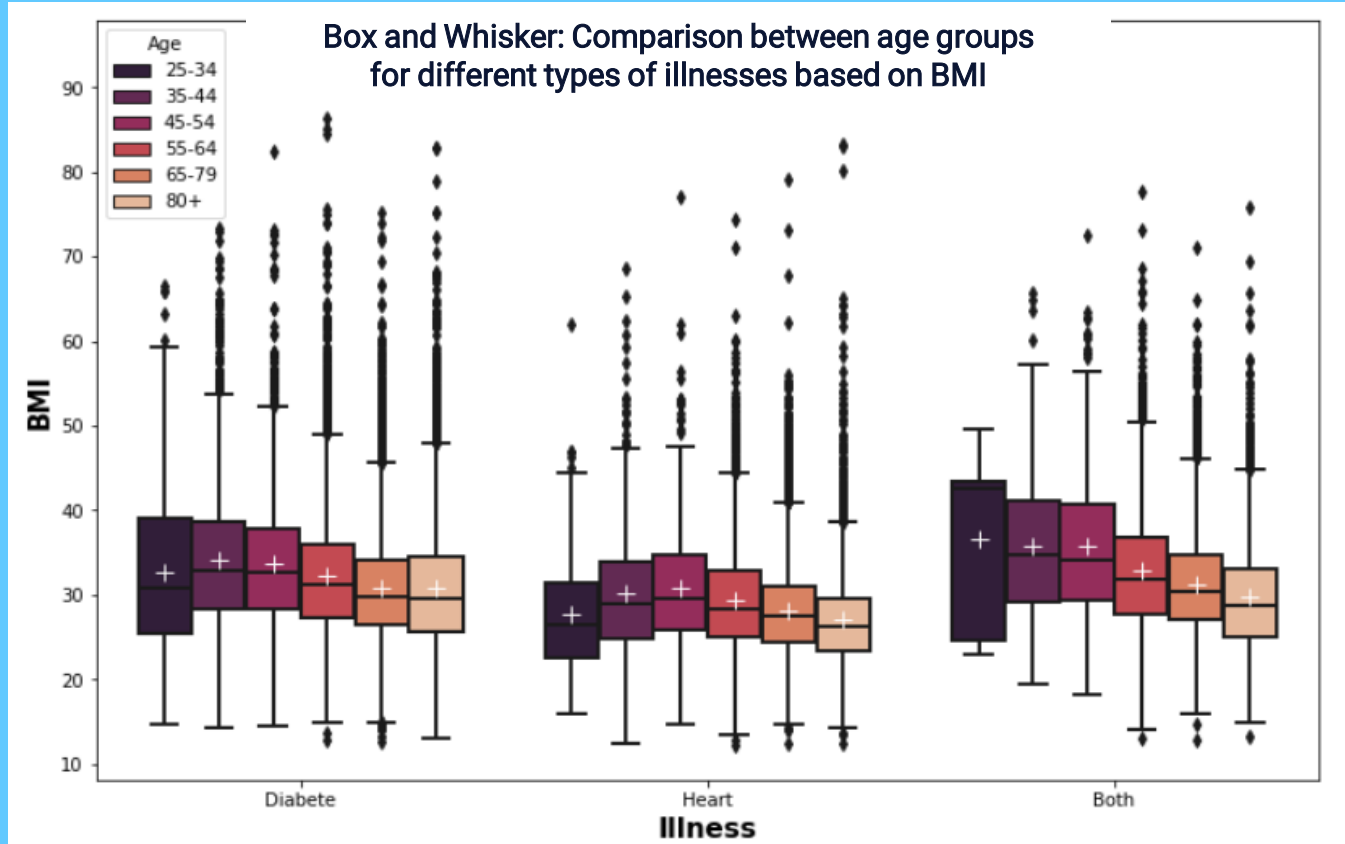


04

VISUALIZATION

5

On average, all participants who have diabetes and/or heart disease have their BMI > 25, which means most of them are overweight and/or obese. The younger half of participants (25-54 years old) have more BMI (body fat) than the older half.

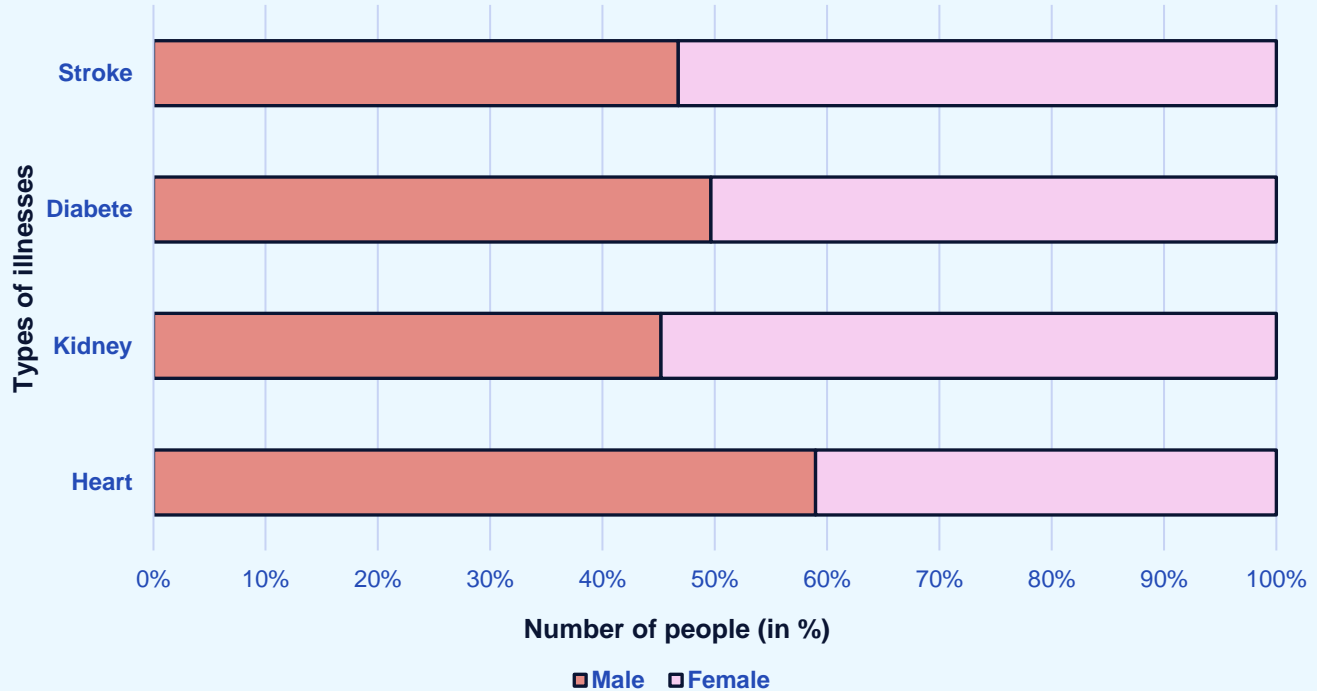


04

VISUALIZATION

6

Comparison between male and female for different types of illnesses

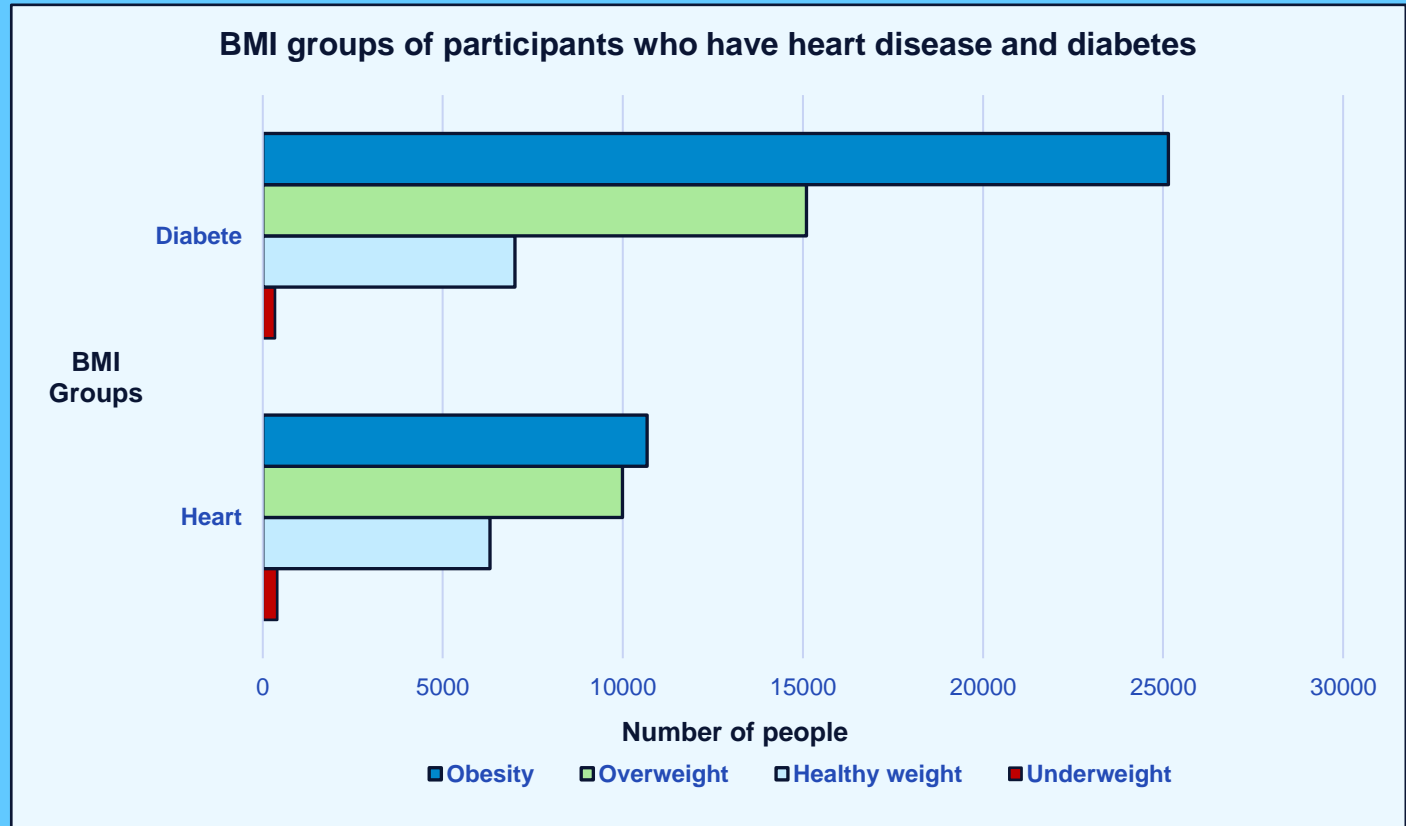


- Stroke, Diabetes, Kidney: The difference in the number of women and men that are diagnosed with these diseases is not significant.
- Heart: More men have heart disease than women.

04

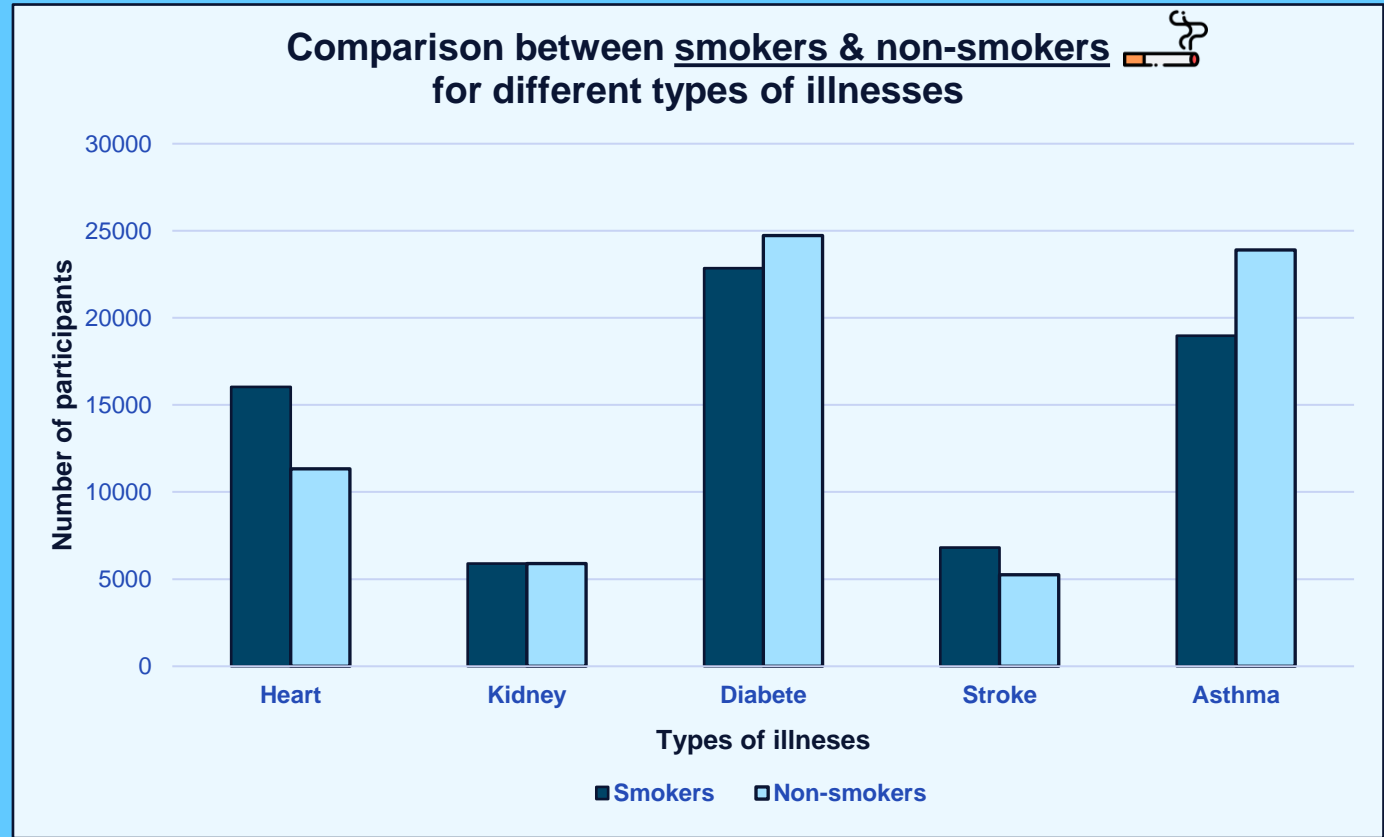
VISUALIZATION

7



- **Diabetes:** The number of **obese** people surge up, significantly higher than any other BMI group. The share of **obese and overweight** people are much more.
- **Heart:** The **obese and overweight** groups are still larger in number compared to the healthy group.

- **Heart:** The number of smokers are moderately higher than non-smokers.
- **Asthma:** People who have asthma are more non-smokers than smokers. However, non-smokers can also be passive smokers (who smell the smokes), and **it is more dangerous to be passive smokers than active smokers.**
- No significant differences in the number of smokers and non-smokers for ppl who have **kidney disease, stroke or diabetes**

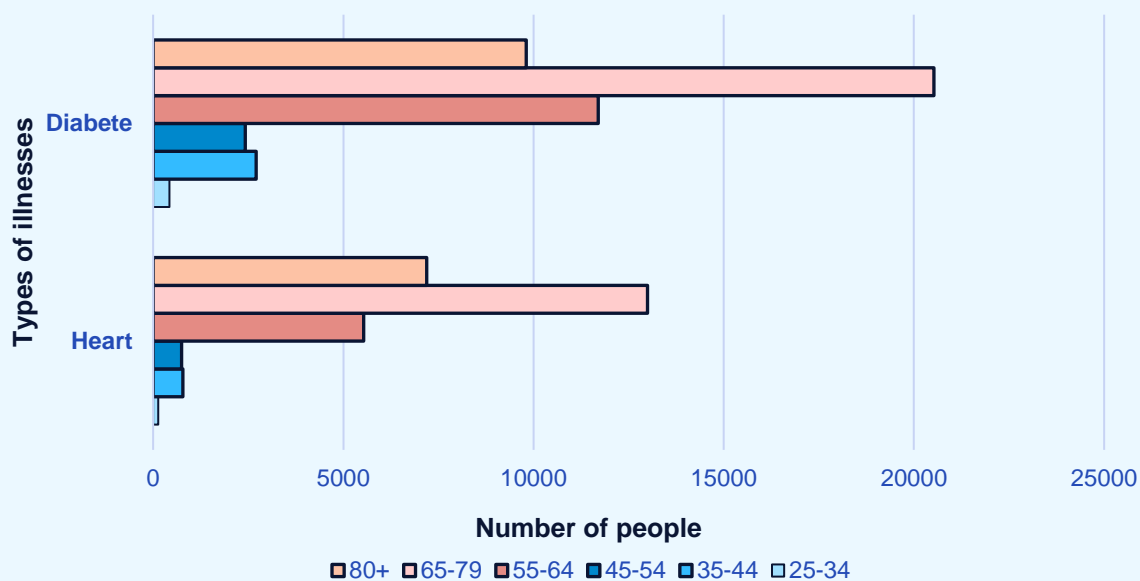


04

VISUALIZATION

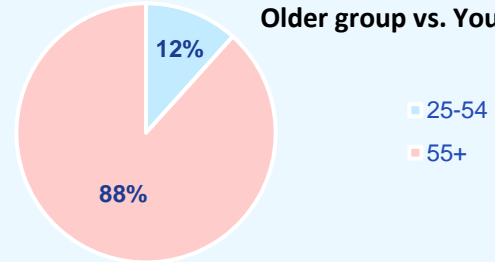
9

Comparison between age groups
for different types of illnesses



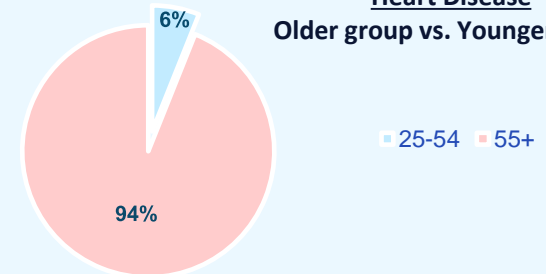
Diabetes

Older group vs. Younger group



Heart Disease

Older group vs. Younger group



05

LOGISTIC REGRESSION

Coefficients of independent (X -) variables with Heart Disease

- A higher index of **age** is associated with a 42.39% probability of heart disease
- For people with **diabetes**, there is a 41.85% risk of having heart disease compared to those who are not diabetic.

X - variables	Coefficients (β)	Risk ($e^{\beta} - 1$)*100%
Age	0.35342649	42.39
Diabetes	0.34956856	41.85
Stroke	0.30909753	36.22
Sex	0.25749593	29.37
Kidney Disease	0.21239515	23.66
Asthma	0.09839852	10.34
BMI	0.06462406	6.68



05

LOGISTIC REGRESSION

Accuracy score: 0.9131114562378178
Precision score: 0.5306122448979592
Recall score: 0.05885857262004051
f1 score: 0.10596310596310597

True Negative: 87109
True Positive: 494
False Positive: 437
False Negative: 7899

- The logistic regression model predicted **accurately 91.31% of the times** whether a heart disease based on their age, BMI, sex, whether they smoke, drink alcohol, is diabetic, had a stroke, has a kidney disease or asthma.
- 56.06% of people predicted to **have** heart disease actually **do have heart disease**
- $(100 - 5.8)\% = 94.2\%$ of people predicted to **not have** heart disease actually **do have heart disease**
- F1 score = 10.6% which is very low


- 87109 people were predicted to not have heart disease, and it's true that they don't.
- 494 people were predicted to have heart disease, and it's true that they do.
- 7899 people were predicted to **not have** heart disease, but in fact they do.
- 437 people were predicted to **have** heart disease, but in fact they **don't**.



06

CONCLUSION

main

1. The correlation results aren't significant, but they correspond the common knowledge about the effects of bad habits to the risk of having illnesses.
 2. Most people who have diabetes and/or heart disease are **either overweight or obese**.
 3. Stroke, diabetes, kidney disease, heart disease **can occur to anyone regardless of age and sex**. However, people more than 55 years old are more likely to have diabetes and/or heart disease.
 4. For **heart disease**, the number of smokers are moderately higher than non-smokers.
 5. The logistic regression model is better at predicting if a person does NOT have a heart disease compared to if a person actually has a heart disease. This is true because of the dataset's imbalance.
- 



06

CONCLUSION

additional

1. In general, people who have heart disease and/or diabetes may have **a higher BMI** than those who do not have heart disease.
2. In general, people who have a high level of general health have a lower BMI than the others.
3. Surprising, people who have asthma are more non-smokers than smokers. However, it is important to know that non-smokers can also be passive smokers (who smell the smokes), and **it is more dangerous to be passive smokers than active smokers**.
4. Most of the time the model predicts accurately whether a person has a heart disease, but many people who are positive with heart disease were predicted to not have it, which is dangerous.



06


CONCLUSION

dataset

Two reasons why the results of the research should be taken as a guidance, not as specific conclusions:

- The dataset's **imbalance**
- Too many binary and categorical variables

Three reasons why the dataset is realistic:

- There cannot be too much people who have cardiovascular diseases in real life.
 - The number of participants is big: 319795.
 - The dataset is extracted from a real survey in 2020.
- 

REAL INFORMATION

“Tobacco doubles the risk of stroke and significantly increases the risk of having cardiovascular and kidney diseases, asthma and diabetes”



THANK YOU FOR YOUR ATTENTION

and stop smoking. tobacco kills.

