Student: Lam Lay
EXST 7005-02

Instructor: Dr. Gentimis
Date: 11/08/2022

Heart Attack Analysis

1. Introduction

Heart disease, also known as cardiovascular disease (CVD), is the leading cause of death globally[1]. About 17.9 million people die from heart disease each year, which is about 32% of all deaths worldwide[1]. In the United State, one person dies every 34 seconds from heart disease[2]. In 2020, about 697,000 Americans died from heart disease, accounting for 1/5 of total number of deaths that year[2]. The total cost of heart disease, including health care services, medicines, and lost productivity due to death, in the United States is about $229 billion each year[2]. There are various factors that can contribute to heart failures. The main purpose of this study is to determine which factors are the most important in detecting heart disease. Determining such factors will help doctors detect heart disease early and thus treat it more effectively.

2. Dataset Description

The dataset is called "Heart Attack Analysis & Prediction Dataset" and obtained from Kaggle[3]. There are 303 observations and 13 variables in the dataset. The response variable/output is binary with 1 representing more chance of heart attack and 0 representing less chance of heart attack. Therefore, this is a classification problem. All variables are described in the Data Description page in Kaggle[4]:

- **Age**: Age of the patient (years)
- **Sex**: Gender of the patient (1 = male, 0 = female)
- **cp**: Chest Pain type chest pain type (0 = typical angina, 1 = atypical angina, 2= non-anginal pain, 3 = asymptomatic)
- **trtbps**: resting blood pressure (in mm Hg)
- **chol**: cholesterol in mg/dl fetched via BMI sensor
- **fbs**: if fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- **restecg**: resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = probable or definite left ventricular hypertrophy by Estes' criteria)
- **thalach**: maximum heart rate achieved
- **exng**: exercise induced angina (1 = yes; 0 = no)
- **old peak**: ST depression induced by exercise relative to rest
- **slp**: the slope of the peak exercise ST segment (0 = unsloping, 1 = flat, 2 = downsloping)
- **caa**: number of major vessels (0-4)
- **thall**: a blood disorder called thalassemia (0 = null, 1 = fixed defect, 2 = normal, 3 = reversable defect)
- **output**: diagnosis of heart disease (angiographic disease status) (0 = < 50% diameter narrowing. less chance of heart disease, 1 = > 50% diameter narrowing. more chance of heart disease)

Table 1. Representation of data set

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

3. Statistical Methods

Hypothesis Testing:

A hypothesis testing is the use of a statistical test to determine whether the sample data sufficiently support a particular hypothesis. The null hypothesis (H_0) represents the status quo, where there is no statistical significance or relationship between two variables. We also have the alternative hypothesis (H_A), which is a statement that contradicts the null hypothesis and is the question that we are testing for. Before the test, we need to set the critical region (alpha level), also known as the rejection region, which is the threshold to determine whether a test statistic is statistically significant. For all of the tests below, I will use alpha level equals 0.05. From the hypothesis testing, we will obtain the probability (p-value) that the null hypothesis is true. If p-value is greater than the alpha level, it means the data do not provide convincing evidence for the alternative hypothesis, so we will stick with the null hypothesis. If p-value is less than the alpha level, we reject the null hypothesis and consider the alternative hypothesis as valid.

Chi-square test of independence:

The chi-square test of independence checks whether there is an association between two categorical/nominal variables. The null hypothesis of this test is that the two variables are independent of each other, and the alternative hypothesis is that they are dependent. A contingency matrix was first created for the *output* variable (risk of heart attack) and other categorical variables (*cp* and *caa*) that we want to test. The matrix contains both the expected and observed values of the numbers of high-risk and low-risk if the variables were independent from each other. From the differences between the observed and expected values, we can calculate the chi-square value, which then gives us the p-value to decide whether the null hypothesis should be rejected or not.

Z-test:

A z-test is a statistical test to determine whether two population means are different or not. To use this test, the variances must be known, the sample size is large enough (greater than or equal to 30), and the population is not extremely skewed and can be approximated by a normal distribution. In this report, a two-proportion z-test, which as its name suggests uses a z-test to compare two proportions, will be used to evaluate whether the women of this dataset are at a higher risk of heart attack than men.

T-test:

When the number of observations is rather low (less than 30) and/or when the population variance is unknown, we should use t-test, rather than z-test. Like a normal distribution, the t-distribution also has a bell shape, but its tails are thicker to resolve the problem with a less reliable estimate of the standard error (since the sample size is small). If the variances of the two

groups are equal to each other, a pooled t-test should be used. If not, an independent t-test will be used. If there are one-to-one correspondence between the observations, we should use a paired t-test.

F-test:
An F-test compares the population variances, while z-test and t-test compare the population means. Defined in the interval [0, +∞), the F distribution is a continuous probability distribution, which is based on the degrees of freedom of the two groups. It is proven mathematically that the F-statistics equals the ratio of two Chi-Squared distributions.

Goodness-of-fit test:
A goodness-of-fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a normal distribution. This test is used to check an assumption of a z-test or t-test, in which the sample distribution should not be extremely skewed. The optimal case is when the distribution is normal. Some of the common goodness-of-fit tests are: Shapiro-Wilk test, Kolmogorov-Smirnov test, Anderson-Darling test, and Cramer-von Mises test. The null hypothesis is that the data comes from a normal distribution. If p-value generated by these tests are greater than the significance alpha level (0.05), we can assume that our data is "normal enough". Otherwise, we reject the null hypothesis.

4. Analysis and Results

4.1. Graphical presentations

First, I created a correlation matrix (Fig. 1) to identify any variable that is highly related to the response (low risk or high risk of heart disease). Then I created stacked histograms for the categorical variables and density plots for continuous variables. From the correlation matrix, it seems like the risk of having heart disease is highly dependent on the types of chest pain. The histogram of this variable (Fig.2) also shows that some types of chest pain, specifically atypical angina or non-anginal pain, contribute to higher risk for heart attack.

Surprisingly, the correlation matrix shows that neither 'cholesterol', 'age', and 'resting blood pressure' variables are significantly related to the risk of heart disease. The density plot of cholesterol variable (Fig.3) also does not demonstrate the common belief that high levels of cholesterol can increase your risk of heart disease[5].

In addition, the histograms of gender (Fig. 4) and number of major blood vessels (Fig. 5) indicates that women of this dataset appear to be at a higher risk of heart attack than men and having more major blood vessels appears to reduce risk for a heart attack, respectively. However, the risk suddenly increases when the number of major blood vessels reaches 4, even though it was decreasing from 0 to 3 (Fig. 5).

4.2.Goals

As a result, we will conduct statistical tests to check if the following hypotheses are true or now:

- the risk for heart attack is dependent on the presence of chest pain
- the risk for heart attack is not dependent on the cholesterol levels
- women of this dataset are at a higher risk of heart attack than men
- having more major blood vessels appears to reduce risk for a heart attack.

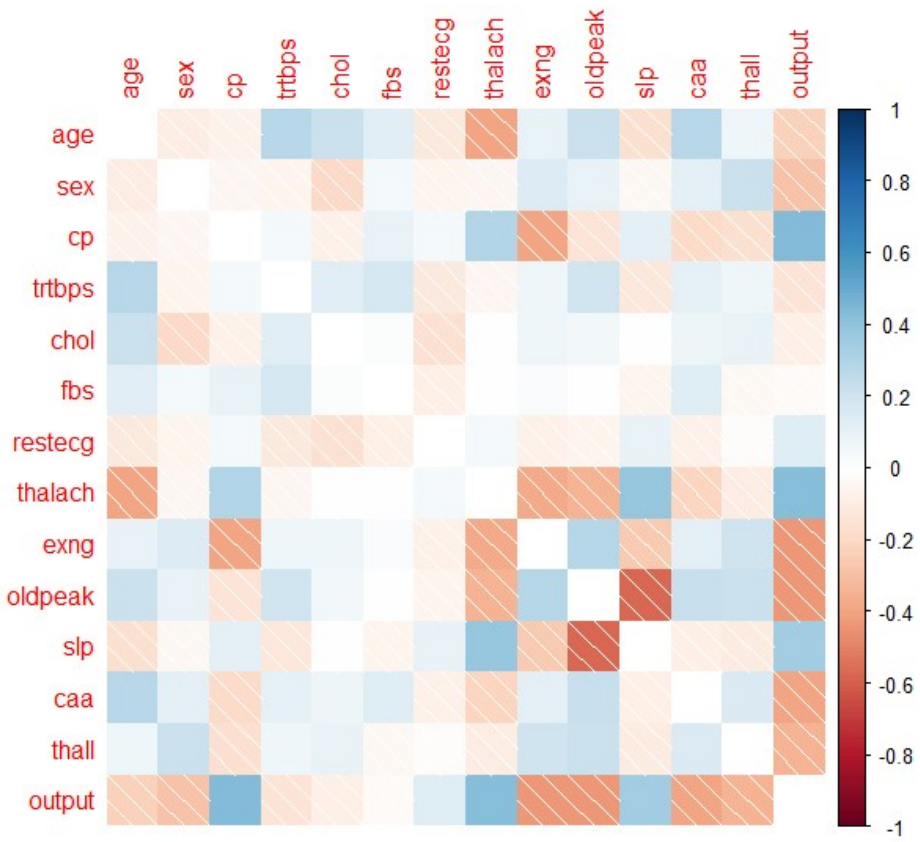

Fig.1. Correlation matrix of all variables in the dataset

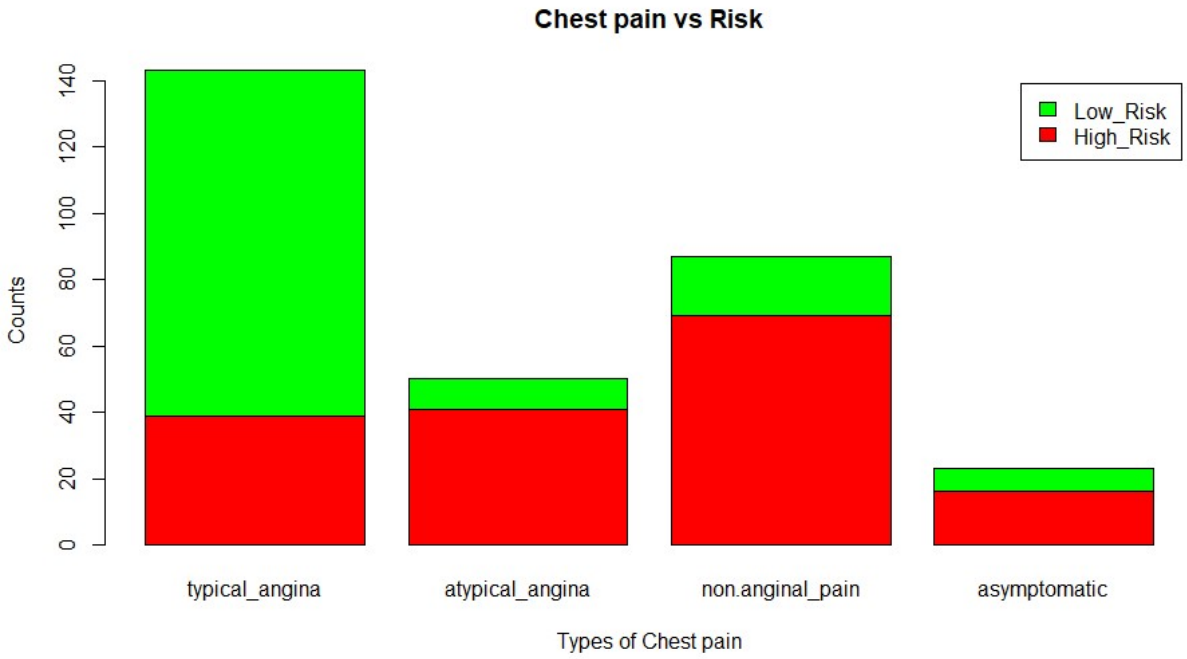

Fig.2. Stacked histogram of different types of chest pain.

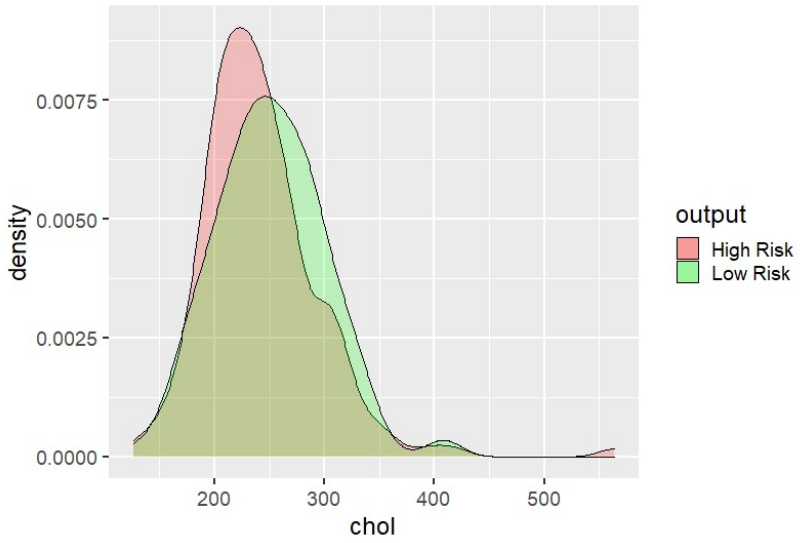

Fig. 3. Density plot of cholesterol levels for high and low risks.

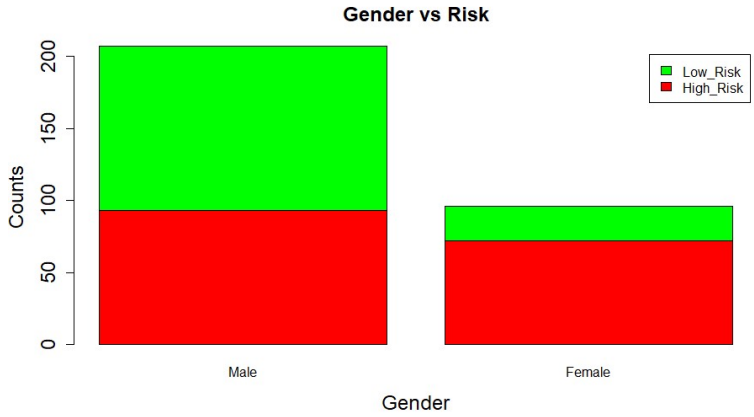

Fig. 4. Stacked histogram of gender.

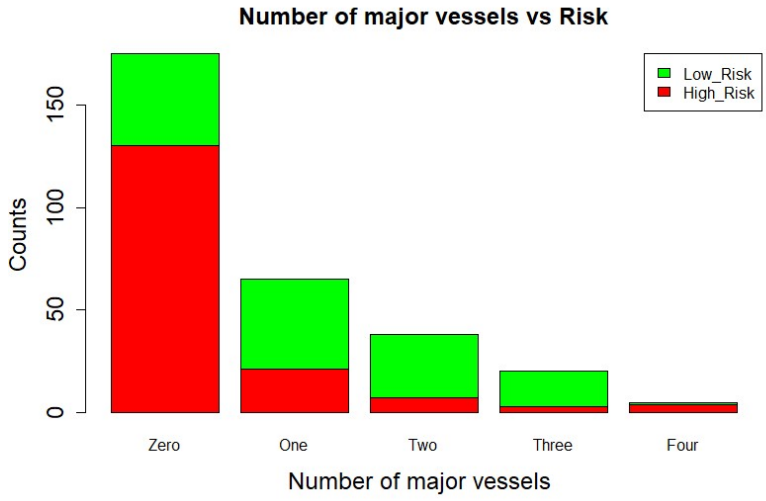

Fig. 5. Stacked histogram of number of major blood vessels.

4.3.Results

4.3.1. Check if the risk for heart attack is dependent on the presence of chest pain.

Because the chest pain variable (*cp*) is a categorical variable, I will use a Chi-square test of independence with the following null and alternative hypotheses.

Chi-square test of independence:

H_0 : The two variables (*output* and *cp*) are independent.
H_A : The two variables (*output* and *cp*) are not independent.

The observed and expected values of the risk of heart attack for different types of chest pain from the chi-square analysis are shown in Table 1. There were more low-risk cases and less high-risk cases associated with typical-angina chest pain than expected. Both atypical-angina and non-anginal pain have more high-risk cases than expected, and their ratio of high-risk over low-risk are 4.5 and 3.8, respectively. The asymptotic people also have more high-risk cases than expected, but not as extreme as atypical-angina and non-anginal pain.

Table 1: Contingency table of observed and expected values for variables chest pain and output from the chi-square test of independence (p-value < 0.001)

		Chest pain				
		Typical angina	**Atypical angina**	**Non-anginal pain**	**Asymptomatic**	**Total**
Output	**Low risk**	104	9	18	7	138
	Expected	65.13	22.77	39.62	10.48	138
	High risk	39	41	69	16	165
	Expected	77.87	27.23	47.38	12.52	165
Total		143	50	87	23	303

$$X = \frac{(104-65.13)^2}{65.13} + \frac{(9-22.77)^2}{22.77} + \frac{(18-39.62)^2}{39.62} + \frac{(7-10.48)^2}{10.48} + \frac{(39-77.87)^2}{77.87} + \frac{(41-27.23)^2}{27.23} + \frac{(69-47.38)^2}{47.38} + \frac{(16-12.52)^2}{12.52} = 81.686$$

Other ways to get Chi-square are: 1) using the *CrossTable()* function in R and then sum all Chi-square contribution values, 2) use *chisq.test()* in R. I tried both and got the same value of chi-square. Because the total chi-square is very large, we expect p-value to be very small:

$$P(X > 81.686\ ) = 0.00001$$

Since our p value is much less than 0.05, we must reject the null hypothesis. This implies that in this experiment, types of chest pain and risk of heart disease are dependent variables. As previously mentioned, the correlation matrix (Fig. 1) and the histogram of chest pain (Fig. 2) show that atypical angina or non-anginal pain contribute to higher risk for heart attack.

4.3.2. Check if the risk for heart attack is not dependent on the cholesterol levels.

The cholesterol variable is continuous, so I will use a t-test to compare the average cholesterol of those with high risk of heart attack to that of those with lower risk. It is believed that high cholesterol levels increase the risk of heart disease, so I will use a one-sided test to see if the people with high risk of heart attack in this dataset has higher average cholesterol.

Assumptions:

1. Independence of observations within the groups
2. Independence of observations between the groups
3. The distributions of cholesterol for high risk and low risk of heart disease should not be extremely skewed. Ideally, they should be normal.

The first two assumptions are based on the design of the experiment and the way the survey was conducted. Since I only have the data, I cannot verify them. Therefore, I will assume that these two assumptions are satisfied. The third assumption can be verified by using a goodness-of-fit test.

Check for normality using a goodness-of-fit test:

H_0: The data for cholesterol comes from a normal distribution
H_A: The data for cholesterol doesn't come from a normal distribution

I used the Kolmogorov-Smirnov test in R to check for the normality of the cholesterol variable and obtained a p-value of 0.3097. The p values are greater than 0.05, so we fail to reject the null hypothesis and can assume that our data is "normal enough".

Check for equal variances using F-test:

H_0 : There is no difference in variance of the cholesterol levels between the high-risk and low-risk populations.
H_A : There is a difference in variance of the cholesterol levels between the high-risk and low-risk populations.

To decide whether I should use a pooled t-test or an independent t-test, I conducted an F-test to check whether the variances are equal or not. If they are equal, a pooled t-test will be use. Otherwise, an independent t-test will be used. The p-value generated by the F-test is 0.3353, which is greater than 0.05. Therefore, we failed to reject the null hypothesis and claim that the variances of the two groups are equal. As a result, a pooled t-test will be used.

Pooled t-test:

H_0: People with higher risk of heart attack has the same average cholesterol as those with lower risk.
H_A: People with higher risk of heart attack has higher average cholesterol than those with lower risk.

$$H_0: \mu_H = \mu_L$$
$$H_A: \mu_H > \mu_L$$

The t-test was done in R, and the calculated p-value is 0.0694. Because our p-value is greater than 0.05, we fail to reject the null hypothesis and claim that there's not enough evidence to show that people with high levels of cholesterol has higher risk of heart disease. This conclusion matches the inference of the density plot (Fig. 3) and correlation matrix (Fig. 1).

4.3.3 Check if women of this dataset are at a higher risk of heart attack than men.

A contingency table between the two variables (*sex* and *output*) was created as shown in Table 2. Upon initial observation, more than half of the female group are at high risk of heart disease, while more than half of the male group have low risk.

Table 2: Contingency table between gender and risk

		Gender		
		Female	**Male**	**Total**
Output	**Low Risk**	24	114	138
	High Risk	72	93	165
Total		96	207	303

From Table 2, we can find the percentages for higher risk among the two genders as follows:

$$p_F = \frac{72}{96} = 0.75$$

$$p_M = \frac{93}{207} = 0.449$$

Hypothesis:

H_0: the proportion of people with higher risk of heart disease among female students is the same as that of male student.

H_A: the proportion of people with higher risk of heart disease among female students is higher than that of male student.

$$H_0: p_F = p_M$$
$$H_A: p_F > p_M$$

Assumptions:

As mentioned above, we can assume that there is independence within and between because they are based on the design of the experiment and the way the survey was conducted. What we can verify is whether there are enough responses in all possible categories. Since both samples for males and females have less than 1000 observations, each group needs to have at least 10 observed successes and 10 observed failures. According to Table 3, this condition is satisfied as follows:

- Male and higher risk 93 > 10
- Male and lower risk 114 > 10
- Female and higher risk 72 > 10
- Female and lower risk 24 > 10

Two-proportion z-test:
Using the following equations, we calculated that the z-score equals to 5.359476.

$$SE = \sqrt{\frac{p_m(1-p_m)}{n_m} + \frac{p_f(1-p_f)}{n_f}} \quad \text{(eqn. 1)}$$

$$Z = \frac{p_f - p_m}{SE} \quad \text{(eqn. 2)}$$

Because this is a one-sided test, p-value = P(Z > 5.359476) = (1 – P(Z < 5.359476)) = $4.17 * 10^{-8}$. Because p-value is much less than 0.05, we must reject the null hypothesis. Therefore, we conclude that in this sample the proportion of having higher risk of heart attack is statistically significantly higher for female patients than for male patients.

4.3.3. Check if having more major blood vessels appears to reduce risk for a heart attack.

Similar to chest pain (*cp*), the number of major blood vessels (*caa)* is a categorical variable. Therefore, I will use a Chi-square test to check for the independence between this variable (*caa*) and the response (*output*).

Chi-square test of independence:
H_0 : The two variables (*output* and *caa*) are independent.
H_A : The two variables (*output* and *caa*) are not independent.

The observed and expected values of the risk of heart attack for different number of major blood vessels from the chi-square analysis are shown in Table 3. There were more high-risk cases and less low-risk cases associated with having no major blood vessels than expected. The number of high-risk cases steadily decreases from having 0 to 3 major blood vessels, but it suddenly increases more than expected when the number of major blood vessels equals to 4.

Table 3: Contingency table of observed and expected values for variables *caa* and *output* from the chi-square test of independence (p-value < 0.001)

		Number of major blood vessels					
		0	**1**	**2**	**3**	**4**	**Total**
Output	**Low risk**	45	44	31	17	1	138
	Expected	79.70	29.60	17.31	9.11	2.28	138
	High risk	130	21	7	3	4	165
	Expected	95.30	35.40	20.69	10.89	2.72	165
Total		175	65	38	20	5	303

I used the same methods in section 4.3.1 to calculate Chi-square, which equals to 74.367, another large number. Hence, we expect p-value to be very small:

$$P(X > 74.367) = 2.712 * 10^{-15} < 0.00001$$

Since our p value is much less than 0.05, we must reject the null hypothesis. This implies that in this experiment, the risk of heart attack is dependent on the number of major blood vessels a person has. As previously mentioned, the histogram of *caa* (Fig. 5) shows that having more major blood vessels seems to reduce risk for a heart attack.

Discussion and Conclusions:

Based on the p-values generated in the abovementioned four tests, I conclude that the risk of having heart disease is dependent on the presence of chest pain and the number of major blood vessels a person has. Specifically, if a person has atypical angina or non-anginal pain, he/she is more likely to have high risk for heart attack as shown in figure 2. On the other hand, having more major blood vessels appears to reduce risk for a heart attack. This makes sense because if the blood vessels function well and are not blocked, nothing will trigger an abnormal heart rhythm, which later results in a cardiac arrest[6]. I am not sure why this dataset shows that the risk increases when the number of major blood vessels equals to 4. I assume it is related to the researcher's definition of major blood vessels, but I do not have this information. However, my chi-square test has showed that the variable *caa* and the risk of heart attack are dependent of each other.

The p-value also shows that the risk for heart attack and cholesterol levels are independent of each other. This is surprising because it is well-known that high levels of cholesterol can increase your risk of heart disease[5]. However, this finding is only limited to the dataset used in this report. We are not given any information about how the experiment/survey was conducted and how they measure the cholesterol levels of the participants. As a result, we should not generalize the findings of this dataset to the real world.

Finally, it seems that women in this dataset are at a higher risk of heart attack than men. This is actually proven in several studies: because women have smaller arteries and are also more likely to suffer from sleep issues than men, they are at higher risk for heart disease[7,8].

Reference

1. World Health Organization. (n.d.). *Cardiovascular diseases*. World Health Organization. Retrieved September 26, 2022, from https://www.who.int/health-topics/cardiovascular-diseases/
2. Centers for Disease Control and Prevention. (2022, July 15). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved September 26, 2022, from https://www.cdc.gov/heartdisease/facts.htm
3. Rahman, R. (2021, March 22). *Heart attack analysis & prediction dataset*. Kaggle. Retrieved September 26, 2022, from https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download
4. Rahman, R. (2021, March 22). *Heart attack analysis & prediction dataset*. Kaggle. Retrieved September 26, 2022, from https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion/329925?resource=download&search=old
5. Mayo Foundation for Medical Education and Research. (2021, July 20). *High cholesterol*. Mayo Clinic. Retrieved November 15, 2022, from https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800#:~:text=Your%20body%20needs%20cholesterol%20to,to%20flow%20through%20your%20arteries.
6. Joy, K. (2018, February 9). *A doctor explains what makes a heart attack a 'widowmaker'*. Widowmaker Heart Attack: Symptoms, Risk Factors & Treatments Explained | Michigan Medicine. Retrieved November 15, 2022, from https://healthblog.uofmhealth.org/widowmaker-heart-attack-explained#:~:text=%E2%80%9CYou%20could%20go%20from%2020,that%20is%20fatal%20within%20minutes.
7. *Why women are having more heart attacks than men*. Why Women are Having More Heart Attacks Than Men: North Valley Women's Care: OBGYNs. (n.d.). Retrieved November 15, 2022, from https://www.northvalleywomenscare.com/blog/why-women-are-having-more-heart-attacks-than-men#:~:text=Women%20are%20officially%20having%20more,of%20death%20in%20women%20worldwide.
8. Chinnaiyan, K. (2018, February 12). *Why are women at higher risk than men for heart disease?* Beaumont Health. Retrieved November 15, 2022, from https://www.beaumont.org/health-wellness/blogs/why-are-women-at-higher-risk-than-men-for-heart-disease#:~:text=Women%20have%20smaller%20arteries%20than,arteries%20that%20feed%20the%20heart.

Appendix:
R code

```
# ====================================================
# Data Cleaning
rm(list = ls())  # remove all variables stored previously

library(Hmisc)

heart = read.csv('heart.csv')  # import data
nrow(heart)                    # number of rows
summary(heart)

str(heart)          # type/class of each variable

colSums(is.na(heart))  # No missing values

heart2 = unique(heart) # Removing any duplicate rows => it seems like there's no duplicate rows in the original dataset

colSums(heart==0)   # trtbps and chol do not have any zero values => good

# ========================================================
# Graphical Presentation
attach(heart)

# Fig.1: Correlation plots
library(corrplot)
library(RColorBrewer)
M = cor(heart)
corrplot(M, method = 'shade', diag = FALSE) # colorful number

# Fig.2: Histogram of chest pain
table(heart$output, heart$cp)        # Contingency table for risk vs chest pain
dat_cp = read.table(text ="typical_angina atypical_angina non-anginal_pain asymptomatic
High_Risk 39 41 69 16
Low_Risk 104 9 18 7 ", header= TRUE)
barplot(as.matrix(dat_cp),col=c("red","green"),legend = TRUE,
        main = "Chest pain vs Risk", xlab = "Types of Chest pain", ylab = "Counts",)

# Fig.3: Density plot of cholesterol
HighRisk_dat = subset(heart, output == 1)
LowRisk_dat = subset(heart, output == 0)

library(ggplot2)
ggplot() +
```

```
  geom_density(aes(chol, fill = 'High Risk'), alpha = .2, data = HighRisk_dat) +
  geom_density(aes(chol, fill = 'Low Risk'), alpha = .2, data = LowRisk_dat) +
  scale_fill_manual(name = "output", values = c("red", "green"))

# Fig.4: Histogram of gender
table(heart$output, heart$sex)       # Contingency table for risk vs sex
dat_sex = read.table(text ="Male Female
High_Risk 93 72
Low_Risk 114 24 ", header= TRUE)
barplot(as.matrix(dat_sex),col=c("red","green"),legend = TRUE,
        main = "Gender vs Risk", xlab = "Gender", ylab = "Counts",
        cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)

# Fig.5:  Histogram of caa
table(heart$output, heart$caa)       # Contingency table for risk vs caa
dat_caa= read.table(text ="Zero One Two Three Four
High_Risk 130 21 7 3 4
Low_Risk 45 44 31 17 1 ", header= TRUE)
barplot(as.matrix(dat_caa),col=c("red","green"),legend.text = TRUE, args.legend = list(x =
"topright"),
        main = "Number of major vessels vs Risk", xlab = "Number of major vessels ", ylab =
"Counts",
        cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)

# ============================================================
# presence of chest pain vs risks
table(heart$output, heart$cp)       # Contingency table for risk vs chest pain

library(gmodels)                          # calculate chi-square
CrossTable(heart$output, heart$cp)
total_chi_squared = 23.200+8.329+11.801+1.153+19.404+6.966+9.870+0.964
total_chi_squared
1 - pchisq(total_chi_squared, df = 1, lower.tail=TRUE)

chisq.test(heart$output, heart$cp)

# ================================================================
# Cholesterol vs output
boxplot(heart$chol, ylab = "chol")
out <- boxplot.stats(heart$chol)$out
out_ind <- which(heart$chol %in% c(out))
out_ind
heart_noOutliers = heart[-c(out_ind),]

# check normality (Goodness of fit test)
ks.test(heart$chol, 'pnorm', mean(heart$chol), sd(heart$chol))     # Kolmogorov–Smirnov test
```

```
# Check the variances
var.test(chol ~ output, heart, alternative = "two.sided")

# Because the variances are equal, we use pooled t-test.
t.test(heart$chol ~ heart$output, var.equal=TRUE, alternative="greater")

# ============================================================
# Females present in this data set appear to be at a higher risk of heart attack than males.
table(heart$output, heart$sex)       # Contingency table for risk vs gender

n_m = 114 + 93
n_f = 24 + 72
p_m = 93/n_m
p_f = 72/n_f

SE=sqrt(p_m*(1-p_m)/n_m+p_f*(1-p_f)/n_f)
Z=(p_f-p_m)/SE

p = 1-pnorm(Z)           # one-sided test

# ================================================================
# caa vs output
table(heart$output, heart$caa)       # Contingency table for risk vs caa

chisq.test(heart$output, heart$caa)

library(gmodels)
CrossTable(heart$output, heart$caa)
total_chi_squared2 = 15.110+7.001+10.834+6.836+0.716+ 12.637+5.855+9.061+5.717+0.599
total_chi_squared2
1 - pchisq(total_chi_squared2, df = 1, lower.tail=TRUE)
```