The Relationship Between Cholesterol Level and the Presence of Heart Disease

Healthcare & Medical Analytics, MSc Business Analytics, Imperial College London, UK

Author: Group 3 | July 2020

Introduction

The presence of a cardiovascular disease in the population varies by living and health conditions. In this study, we examine the relationship between heart disease and cholesterol level, and how this relationship is affected by several factors. The objective is to examine the true effect of cholesterol in the probability of suffering from a heart disease while controlling various socioeconomic variables.

Regarding the reason why cardiovascular disease prevalence was chosen as a topic, approximately 1 in every 4 deaths in the United States occurs due to heart diseases (Centers for Disease Control and Prevention, 2020). As a result, the need to investigate the further reasons why people die from cardiovascular diseases, consist of paramount importance. However, despite the high number of people dying from heart diseases, it is really difficult for patients diagnosed with such a disease to find the appropriate medication, as prescription is often required. Hence, an ease of access to cholesterol medication would be beneficial, as scientifically there are no major adverse side effects to statin (Paul D. Thompson et al., 2016), the drug that lowers cholesterol levels. Therefore, less restrictions should exist for statin resulting in it being easier purchased by cardiovascular disease diagnosed patients.

As said earlier, the level of cholesterol is considered to be one of the most important factors affecting the presence of a heart disease. The relationship between cardiovascular diseases and cholesterol has been examined by several studies and association has been proved to exist (Jaqui Walker, 2013). As seen in previous studies, (Günther Silbernagel et al., 2015), those with high cholesterol absorption experienced more severe cardiovascular diseases, which were more difficult to cure.

Apart from examining how cholesterol level affects the prevalence of heart disease, it is examined how other health or social factors influence that relationship. More specifically, it is assessed how the relationship between cholesterol and heart diseases is affected by respondent's age, height and weight, gender, systolic and diastolic blood pressure, glucose, smoking, alcohol and physical activity.

Variable Selection

In terms of independent variables, a relationship between height and heart diseases seems to exist. Moreover, as Jeonggeun Moon states (2019:p.114), 'if current evidence on the relationship between height and CVD risk is correct, short people could be targeted with more stringent strategies for CVD prevention'.

Regarding further independent variable selection, as Wei Ma states (2017:p.470), 'Routine measurement of interarm blood pressure may provide a simple and effective screening method for the presence of peripheral artery disease', which indicates an association between blood pressure and heart diseases. Similar association between systolic, diastolic blood pressure and cardiovascular events is proven by other studies too (Emmanuelle Vidal-Petiot et al., 2016).

Association between glucose and heart diseases has also been proven (Elisabeth von Gunten et al., 2013) as well as another one between cardiovascular diseases and smoking as David M. Burns states (2003:p.22), 'Cigarette smoking is a major cause of CHD, stroke, aortic aneurysm, and PVD'.

Regarding alcohol consumption, studies have shown a relationship between this and cardiovascular diseases, as Michael V Holmes points out (2014:p.8), 'individuals of European descent with a genetic predisposition to consume less alcohol had a reduced risk of coronary heart disease and ischemic stroke, and lower levels of several established and emerging risk factors for cardiovascular disease'.

In Lori Mosca's research (Lori Mosca's et al., 2011) , it mentions 'more men are living with and dying of coronary heart disease than women and have more hospital discharges for cardiovascular disease and coronary heart disease', which indicates an association between gender and heart disease. Moreover, the Figure 1 (Annual number of adults having diagnosed heart attack or fatal coronary heart disease or CHD by age and sex) in this research suggests age could be another valuable factor.

Association between weight and cardiovascular disease has been affirmed (W B Kannel et al., 1996). Researchers have concluded that the degree of overweight is related to the rate of development of cardiovascular disease.

Lastly, physical activity also seems to be a factor that determines the prevalence of heart diseases as it prevents them (Martha Wrigley et al., 2006).

Descriptive Statistics

Data cleaning is done before any analysis is conducted. For this dataset, we have selected 12 variables selected for the project, which contains the patient's personal information and several health measures. After exploring the dataset, we omitted rows that are unreasonable from common sense. We omitted all heights lower than 140cm and weights lower than 30kg; We omitted all systolic pressure higher than 240mmHg or lower than 80mmHg, and all diastolic pressure higher than 190mmHg or lower than 50mmHg. The cleaned dataset is shown in appendix (Table 1).

Next, we construct the descriptive analysis for the variables shown in Table 2 in appendix. From the table we can see that patients' ages range from 30 to 65, and the majority of the respondents are male. About 70% of respondents have a normal level of cholesterol, this number increased to 84% for glucose level. It is interesting to find that 90% of respondents don't smoke and drink,

and 80% of all respondents go for physical exercise. In this dataset, 50% of respondents have CVD.

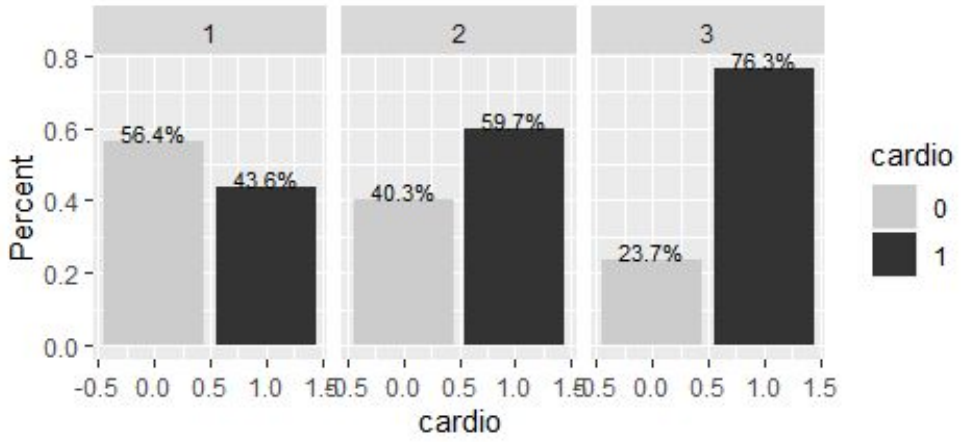

Figure 1: Distribution of target variable

In the three subgroups with different cholesterol levels, the percentages of people having cardiovascular disease increases significantly as the cholesterol increases. 76.3% of people with well above normal cholesterol are diagnosed as cardiovascular. Therefore, it is reasonable to speculate that cholesterol is highly related to cholesterol level.

Regression

Methodology

To estimate the impact of cholesterol levels on whether a person has a cardiovascular disease or not, a regression was conducted. The output of a regression will be able to notify the importance of cholesterol levels on whether a person has a cardiovascular disease or not, and it can tell us the increased likelihood numerically of different cholesterol levels towards getting a cardiovascular disease. There are many regression models, but a logistic regression was decided for the analysis because the target variable - is whether a person has a cardiovascular disease or not - therefore to ensure that the predictions are always between 0 and 1 the logistic regression was used. In regards to data preparation - all missing values were removed, and our regression used 68,589 rows to measure the variable importance. Additionally, before the regression was

run, a sanitation check on the correlation of the independent variables was done to ensure that there was no perfect (or close to perfect) collinearity between the independent variables to ensure that the coefficient estimates are unbiased. After checking the numerical (continuous) variables collinearity between each other (Figure 2), there is no sign of multicollinearity and the chosen independent variables are competent. Additionally, dealing with the categorical data, each categorical column unique value was then converted to an individual column, and one of the columns were dropped to prevent multicollinearity as well.

Output

The output of the logistic regression (Figure 3), has stated that all of the independent variables that were used were all 99% significant, except for the above normal glucose levels. After controlling for other related independent variables, such as the survey participants age, gender, heigh, weight, smoking habits, physical activity habits, etc - the regression results further exemplifies that cholesterol level has a high correlation with the probability in someone getting a cardiovascular disease. Overwhelmingly, both cholesterol levels, cholesterol level above normal and cholesterol well above normal both have the highest coefficients, with 0.3801, and 1.092 respectively. In other words, cholesterol level has the biggest impact on the log odds that a person has a cardiovascular disease or not, for example if someone has a cholesterol level above normal they would have a 0.3801 log odds of getting a cardiovascular disease, and similarity for cholesterol level well above normal - thus, cholesterol levels need to maintained at the normal level to reduce the cardiovascular disease risk.

Limitations

As the majority of the respondents is male, the dataset is imbalanced in regards to gender. The gender skewness would lead to the biased effect of gender on how likely people have cardiovascular disease.

So far, the way we handle missing data was to simply remove all. However, it potentially incurs bias in the estimation of parameters. Without actually investigating the records with missing values, the samples have the chance to lose its capability of representativeness. The technique of data imputation might be considered to handle missing data in the next step and further compare which method will provide better estimation.

From the regression output, the factor of glucose levels turns out to be insignificant in terms of the association to cardiovascular diseases. Therefore, stepwise regression involving multiple variable selections could be another option to build up a more advanced model by filtering unnecessary variables out and highlighting the variables with the optimal estimation.

Additionally, the total/high-density lipoprotein (HDL) cholesterol ratio and the low-density lipoprotein (LDL) cholesterol/HDL cholesterol ratio are two important components to indicate vascular risk (Jesús Millán et al., 2009). By having total cholesterol in the dataset instead of the ratios, the true association between cholesterol and cardiovascular disease would be less significant.

Recommendations

High cholesterol and high blood pressure is closely associated with having cardiovascular disease. Besides, glucose level is a good and sensitive indicator of cardiovascular because as long as glucose deviates from normal level, the risk of cardiovascular increases. In conclusion, a combination of cholesterol, blood pressure and glucose is an accurate health index for defining the high risk group of cardiovascular.

Healthy habits are also very important for reducing the risk of cardiovascular, especially for the elder and the higher group. No smoking, no alcohol and regular physical exercises are proved to be efficient means for preventing cardiovascular.

References

Centers for Disease Control and Prevention. (2020) *Heart Disease in the United States.* Available from: https://www.cdc.gov/heartdisease/facts.htm [Accessed 22th june 2020]

David M. Burns (2003) Epidemiology of Smoking-Induced Cardiovascular Disease. *Progress in Cardiovascular Diseases.* 46 (1), 11-29. Available from: doi:10.1016/S0033-0620(03)00079-3

Elisabeth von Gunten, Julia Braun, Matthias Bopp, Ulrich Keller, David Faeh (2013) J-shaped association between plasma glucose concentration and cardiovascular disease mortality over a follow-up of 32 years. *Preventive Medicine.* 57 (5), 623-628. Available from: doi:10.1016/j.ypmed.2013.08.016

Günther Silbernagel, Günter Fauler, Bernd Genser, Christiane Drechsler, Vera Krane, Hubert Scharnagl, Tanja B. Grammer, Iris Baumgartner, Eberhard Ritz, Christoph Wanner, Winfried März (2015) Intestinal Cholesterol Absorption, Treatment With Atorvastatin, and Cardiovascular Risk in Hemodialysis Patients. *Journal of the American College of Cardiology.* 65 (21), 2291-2298. Available from: doi:10.1016/j.jacc.2015.03.551

Jaqui Walker (2013) Reducing cardiovascular disease risk: cholesterol and diet. *Nursing Standard.* 28 (2), 48-55. Available from: doi:10.7748/ns2013.09.28.2.48.e7747

Jaqui Walker (2013) Cardiovascular event rates and mortality according to achieved systolic and diastolic blood pressure in patients with stable coronary artery disease: an international cohort study. *The Lancet.* 338 (10056), 2142-2152. Available from: doi:10.1016/S0140-6736(16)31326-5

Jeonggeun Moona (2019) The Link between Height and Cardiovascular Disease: To Be Deciphered. *Division of Cardiology, Department of Internal Medicine, Gil Medical Center, Department of Family Medicine, Gil Medical Center, Gachon University College of Medicine, Incheon, South Korea.* 143, 114–115. Available from: doi:10.1159/000502032

Martha Wrigley and Tapesh Pakrashi (2006) Physical activity and cardiovascular disease. *British Journal of Cardiac Nursing.* 1 (8). Available from: doi:10.12968/bjca.2006.1.8.21677

Michael V Holmes et al. (2014) Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ.* 349. Available from: doi:10.1136/bmj.g4164

Millán, J., Pintó, X., Muñoz, A., Zúñiga, M., Rubiés-Prat, J., Pallardo, L. F., Masana, L., Mangas, A., Hernández-Mijares, A., González-Santos, P., Ascaso, J. F., & Pedro-Botet, J. (2009). Lipoprotein ratios: Physiological significance and clinical usefulness in cardiovascular prevention. *Vascular health and risk management*, *5*, 757–765. Available from: doi: 10.2147/VHRM.S6269

Lori Mosca, MD, MPH, PhD, Elizabeth Barrett-Connor, MD, and Nanette Kass Wenger, MD (2011) Sex/Gender Differences in Cardiovascular Disease Prevention. *Circulation AHA Journals*. Available from: doi:10.1161/CIRCULATIONAHA.110.968792

W B Kannel, R B D'Agostino, J L Cobb (1996) Effect of weight on cardiovascular disease. *The American Journal of Clinical Nutrition*, Volume 63, Issue 3,Pages 419S–422S. Available from: doi:10.1093/ajcn/63.3.419

Paul D. Thompson, GregoryPanza, Amanda Zaleski, BethTaylor (2016) Statin-Associated Side Effects. *Journal of the American College of Cardiology.* 67 (20) 2395-2410. Available from: doi:10.1016/j.jacc.2016.02.071

Wei Ma, Baowi Zhang, Ying Yang, Litong Qi, Lei Meng, Yan Zhang, Yong Huo (2017) Correlating the relationship between interarm systolic blood pressure and cardiovascular disease risk factors. *Department of Cardiovascular Disease, Peking University First Hospital, Beijing, China.* 19, 466-471. Available from: doi:10.1111/jch.12987

Appendix

Tables & Figures

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	50	2	168	62	110	80	1	1	0	0	1	0
1	55	1	156	85	140	90	3	1	0	0	1	1
2	52	1	165	64	130	70	3	1	0	0	0	1
3	48	2	169	82	150	100	1	1	0	0	1	1
4	48	1	156	56	100	60	1	1	0	0	0	0
8	60	1	151	67	120	80	2	2	0	0	0	0

Table 1: Cleaned dataset

	Mean	Std.Dev	Min	Median	Max	N.Valid	Pct.Valid
Age	53.3386857	6.7652940	30	54	65	70000	100.00000
Gender	1.3495714	0.4768380	1	1	2	70000	100.00000
Height	164.4656826	7.8204376	140	165	250	69848	99.78286
Weight	74.2110540	14.3863125	30	72	200	69993	99.99000
Ap_hi	127.0174903	17.0724678	80	120	240	69753	99.64714
Ap_lo	81.3967260	9.6681745	50	80	190	68967	98.52429
Chole	1.3668714	0.6802503	1	1	3	70000	100.00000
Gluc	1.2264571	0.5722703	1	1	3	70000	100.00000
Smoke	0.0881286	0.2834838	0	0	1	70000	100.00000
Alco	0.0537714	0.2255677	0	0	1	70000	100.00000
Active	0.8037286	0.3971791	0	1	1	70000	100.00000
Cardio	0.4997000	0.5000035	0	0	1	70000	100.00000

Table 2: Descriptive Analysis

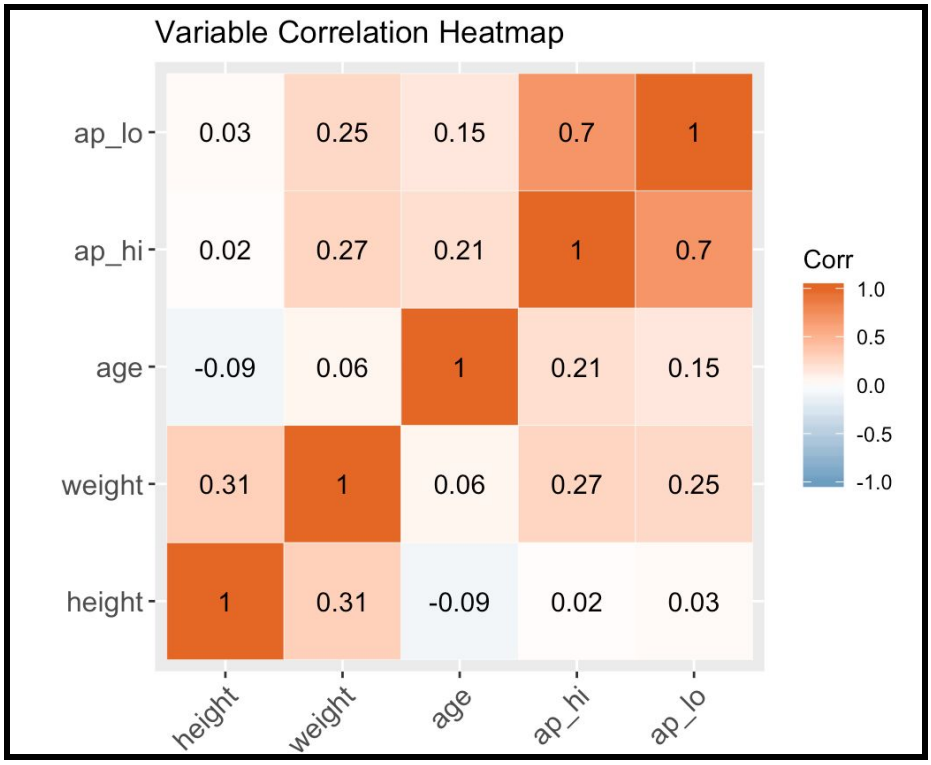

Figure 2: Correlation Heatmap

```
Regression Results
==========================================
                      Dependent variable:
                    ----------------------
                             cardio
------------------------------------------
age                       0.0001***
                          (0.00000)

gender2                     -0.014
                            (0.022)

height                    -0.004***
                            (0.001)

weight                     0.011***
                            (0.001)

ap_hi                      0.053***
                            (0.001)

ap_lo                      0.017***
                            (0.001)

cholesterol2               0.381***
                            (0.027)

cholesterol3               1.092***
                            (0.036)

gluc2                       0.016
                            (0.036)

gluc3                     -0.339***
                            (0.040)

smoke1                    -0.141***
                            (0.035)

alco1                     -0.206***
                            (0.042)

active1                   -0.230***
                            (0.022)

Constant                 -10.839***
                            (0.253)

------------------------------------------
Observations                68,589
Log Likelihood          -38,466.760
Akaike Inf. Crit.       76,961.510
==========================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

Figure 3: Logistic Regression Output

Code

Code for Regression

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(neuralnet)
library(stats)
library(stargazer)
library(readxl)
library(Metrics)
library(forecast)
library(tseries)
library(kableExtra)
library(gridExtra)
library(cowplot)
library(caret)
library(ggcorrplot)
data = read.csv('/Users/nicksidjono/Downloads/cardio_cleaned.csv', sep = ',')
data = data %>% select (-X,-id)
data$gender = as.factor(data$gender)
data$cholesterol = as.factor(data$cholesterol)
data$gluc = as.factor(data$gluc)
data$smoke = as.factor(data$smoke)
data$alco = as.factor(data$alco)
data$active = as.factor(data$active)
data$cardio = as.factor(data$cardio)
data = na.omit(data)
numerical = data %>% select(age, height, weight, ap_hi, ap_lo)
```

```
ggcorrplot(cor(numerical), hc.order = TRUE,
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
    colors = c("#6D9EC1", "white", "#E46726"), lab = TRUE, title = 'Variable
Correlation Heatmap')
glm.fit = glm(cardio ~., data = data, family = binomial)
probit.fit = glm(cardio ~., data = data, family = binomial(link =
"probit"))
stargazer(glm.fit, type = 'text', align = TRUE, title = 'Regression
Results')
```

Code for Descriptive Statistics

```
library(ggplot2)
library(knitr)
library(summarytools)



df_cardio <- read.csv("cardio_train.csv", sep = ";", header = TRUE)



df_cardio["age"] <- round(df_cardio["age"]/365,0)
df_cardio["height"][df_cardio["height"] < 140] <- NA
df_cardio["weight"][df_cardio["weight"] < 30] <- NA
df_cardio["ap_hi"][df_cardio["ap_hi"] > 240 | df_cardio["ap_hi"] < 80] <-
NA
df_cardio["ap_lo"][df_cardio["ap_lo"] > 190 | df_cardio["ap_lo"] < 50] <-
NA

kable(head(df_cardio))

# crate specific variable
age <- df_cardio[,c("age")]
```

```
gender <- df_cardio[,c("gender")]
height <- df_cardio[,c("height")]
weight <- df_cardio[,c("weight")]
ap_hi <- df_cardio[,c("ap_hi")]
ap_lo <- df_cardio[,c("ap_lo")]
chole <- df_cardio[,c("cholesterol")]
gluc <- df_cardio[,c("gluc")]
smoke <- df_cardio[,c("smoke")]
alco <- df_cardio[,c("alco")]
active <- df_cardio[,c("active")]
cardio <- df_cardio[,c("cardio")]

# Providing the labels for different values based on the dictionary
chole_labels <- c("Normal", "Above normal", "Well above normal" )

# Creating frequency table of values
chole_frequency <- table(chole)

# Creating the table of variable and labels
chole_data_frame <- as.data.frame(chole_frequency);
chole_data_frame$Chole_Status <- chole_labels

# Showing the table and information
kable(chole_data_frame, caption = "Cholesterol status")

# Providing the labels for different values based on the dictionary
gluc_labels <- c("Normal", "Above normal", "Well above normal" )

# Creating frequency table of values
gluc_frequency <- table(gluc)

# Creating the table of variable and labels
gluc_data_frame <- as.data.frame(gluc_frequency);
gluc_data_frame$Gluc_Status <- gluc_labels

# Showing the table and information
kable(gluc_data_frame, caption = "Glucose status")

# Providing the labels for different values based on the dictionary
smoke_labels <- c("Non-smoker", "Smoker")

# Creating frequency table of values
```

```r
smoke_frequency <- table(smoke)

# Creating the table of variable and labels
smoke_data_frame <- as.data.frame(smoke_frequency);
smoke_data_frame$Smoke_Status <- smoke_labels

# Showing the table and information
kable(smoke_data_frame, caption = "Smoke status")

# Providing the labels for different values based on the dictionary
alco_labels <- c("Non-drinker", "Drinker")

# Creating frequency table of values
alco_frequency <- table(alco)

# Creating the table of variable and labels
alco_data_frame <- as.data.frame(alco_frequency);
alco_data_frame$Alco_Status <- alco_labels

# Showing the table and information
kable(alco_data_frame, caption = "Alcohol status")

# Providing the labels for different values based on the dictionary
active_labels <- c("No Physical Activity", "With Physical Activity")

# Creating frequency table of values
active_frequency <- table(active)

# Creating the table of variable and labels
active_data_frame <- as.data.frame(active_frequency);
active_data_frame$active_Status <- active_labels

# Showing the table and information
kable(active_data_frame, caption = "Activity status")

age_descr <- descr(age, stats = "common", transpose = TRUE)
gender_descr <- descr(gender, stats = "common", transpose = TRUE)
height_descr <- descr(height, stats = "common", transpose = TRUE)
weight_descr <- descr(weight, stats = "common", transpose = TRUE)
ap_hi_descr <- descr(ap_hi, stats = "common", transpose = TRUE)
ap_lo_descr <- descr(ap_lo, stats = "common", transpose = TRUE)
chole_descr <- descr(chole, stats = "common", transpose = TRUE)
```

```
gluc_descr <- descr(gluc, stats = "common", transpose = TRUE)
smoke_descr <- descr(smoke, stats = "common", transpose = TRUE)
alco_descr <- descr(alco, stats = "common", transpose = TRUE)
active_descr <- descr(active, stats = "common", transpose = TRUE)
cardio_descr <- descr(cardio, stats = "common", transpose = TRUE)

row_names <- c(c("Age", "Gender", "Height", "Weight","Ap_hi", "Ap_lo",
"Chole", "Gluc", "Smoke","Alco","Active","Cardio"))

descr_stats_dataframe <- data.frame(rbind(age_descr, gender_descr,
height_descr, weight_descr, ap_hi_descr, ap_lo_descr, chole_descr,
gluc_descr, smoke_descr, alco_descr, active_descr, cardio_descr), row.names
= row_names)

kable(descr_stats_dataframe, caption = "Descriptive Statistics")
```

```
#Figure 1
library(corrplot)

stats_dataframe <- data.frame(cbind(age, gender, height, weight, ap_hi,
ap_lo, chole, gluc, smoke, alco, active,df_cardio$cardio))
stats_dataframe <- rename(stats_dataframe, cardio = V12)

data1 <- na.omit(stats_dataframe)
ggplot(data1, aes(x=cardio))+
   geom_bar(aes(y = ..prop..,fill = factor(..x..)), stat="count")+
  fill_palette(palette = 'grey') +
   geom_text(aes( label = scales::percent(..prop..),
                  y= ..prop.. ), stat= "count", vjust = 0,size = 3) +
   labs(y = "Percent", fill="cardio") +
   facet_wrap(~chole)
```