Medical Data Science

Working with data: homework

Ursula Trinler

deadline: 06.01.2021

Contents

1 Introduction

1.1 Summary

This document includes the cardiovascular disease dataset which is freely available on kaggle (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset). The main research question of the study is whether the variable *cardio* can be explained by variables which have been additionally collected (see Table 2).

1.2 The cardiovascular dataset

The data set consists of 70000 cases, while 13 variables have been included. The first ten rows of the data set can be found in Table (1).

Table 1: First ten rows of the cardiovascular disease data set

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0
8	21914	1	151	67.0	120	80	2	2	0	0	0	0
9	22113	1	157	93.0	130	80	3	1	0	0	1	0
12	22584	2	178	95.0	130	90	3	3	0	0	1	1
13	17668	1	158	71.0	110	70	1	1	0	0	1	0
14	19834	1	164	68.0	110	60	1	1	0	0	0	0

A detailed description of each included variable can be found in in the next table, Table (2).

Table 2: Description of variables

short name	variable	more information
age	Age	in days
gender	Gender	1 = women, 2 = men
height	Body height	in cm
weight	Body weight	in kg
ap_hi	Systolic blood pressure	mmHg
ap_lo	Diastolic blood pressure	mmHg
cholesterol	Cholesterol	1: normal, 2: above normal, 3: well above normal
gluc	Glucose	1: normal, 2: above normal, 3: well above normal
smoke	Smoking	1 = yes, 0 = no
alco	Alcohol consumation	1 = yes, 0 = no
active	Physical activity	1 = yes, 0 = no
cardio	Cardiovascular disease	1 = yes, 0 = no

1.3 Cleaning the data set

Age is displayed in days, however, having age in years is easier to analyze. Therefore, we transform the variable *age* from days into years as follows:

```
cardio_data$age <- round(cardio_data$age/365.25 , 0)
```

For further analysis it is important to also check which data type each variable has been assigned to.

Table 3: Data type of variables

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
integer	numeric	integer	integer	character	integer	integer	integer	integer	integer	integer	integer	integer

Table 3 shows that *weight* is defined as a character, therefore, it needs to be changed to a numeric data type. We also need to be sure, that there are no missing values or big outliers which might induce errors in the data analysis. Therefore, we, firstly, analyze the distribution of continuous variables by visualizing the data using boxplots.

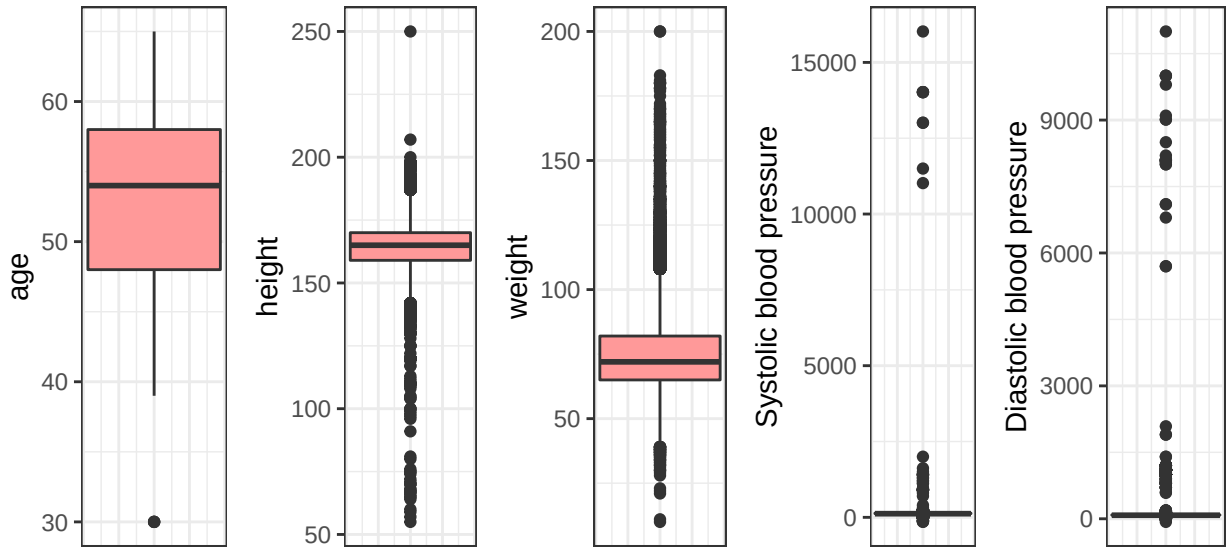

Figure 1: Distribution of continuous variables

Figure 1 shows clearly that there are non-meaningful outliers for the systolic and diastolic blood pressure, as well as for body height and weight as they are outside the healthy and pathologic human range. We define the range of these values as follows:

- Systolic blood pressure: 70 - 190 mmHg
- Diastolic blood pressure: 40 - 100 mmHg
- Weight of an adult person: 40 - 200 kg
- Height of an adult person: 1.40 - 2.20 m

Information about blood pressure range have been taken from *Blood Pressure UK* (http://www.bloodpressureuk.org/your-blood-pressure/understanding-your-blood-pressure/what-do-

the-numbers-mean/), while information of weight and height have been taken from the *NHS*, see https://www.nhs.uk/live-well/healthy-weight/height-weight-chart/.

After filtering the data, the data set is reduced to 63584, therefore, 6416 cases have been removed. Now, the distribution of these variables looks as follows (Figure 2):

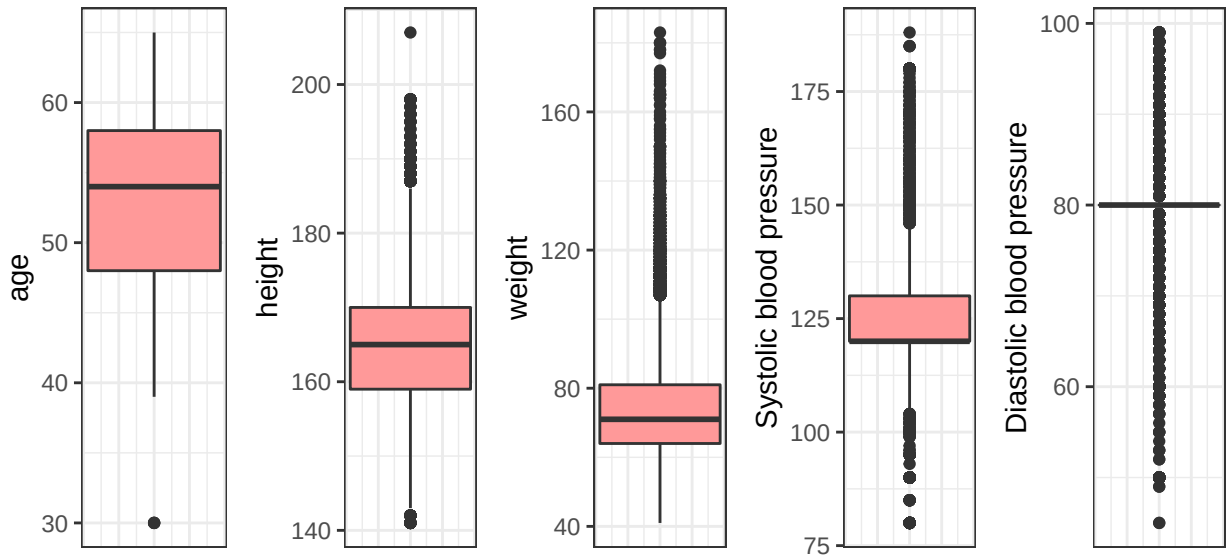

Figure 2: Distribution after removal of outliers

Before we start to analyze the data set in detail we want to visualize the distribution of cases between both cardiovascular groups, to gain an overview if both groups have the same number of participants:

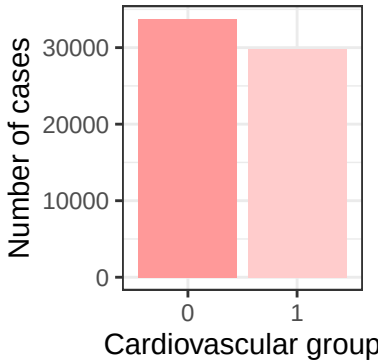

Figure 3: Number of cases per cardiovascular group

Furthermore, we want to see if all data is complete. There are 0 missing cases. Therefore, the data is hereby ready to be used for analysis.

2 Data set maniputation and analysis

2.1 New variable BMI (Task 1)

To understand the influence of the relative weight of a person on cardiovascular diseases, we calculate the BMI and include it in the data set. Table 4 gives an overview of the BMI distribution in both cardio groups (cardio 0 = heathly, cardio 1 = affected).

Table 4: Overview of BMI distribution for each cardio group

Cardio Group	mean	SD	max	min
0	26.4	4.7	68.3	14.6
1	28.2	5.3	64.0	14.5

Cardio group 0 = healthy, 1 = affected
SD = standard deviation

2.2 Correlation between systolic blood pressure and BMI (Task 2)

The BMI might have an influence on the systolic blood pressure. To analyze this we need to correlate both variables (BMI vs. systolic blood pressure) with each other. First we plot the relationship between both variables using a simple scatter plot and include a linear regression line.

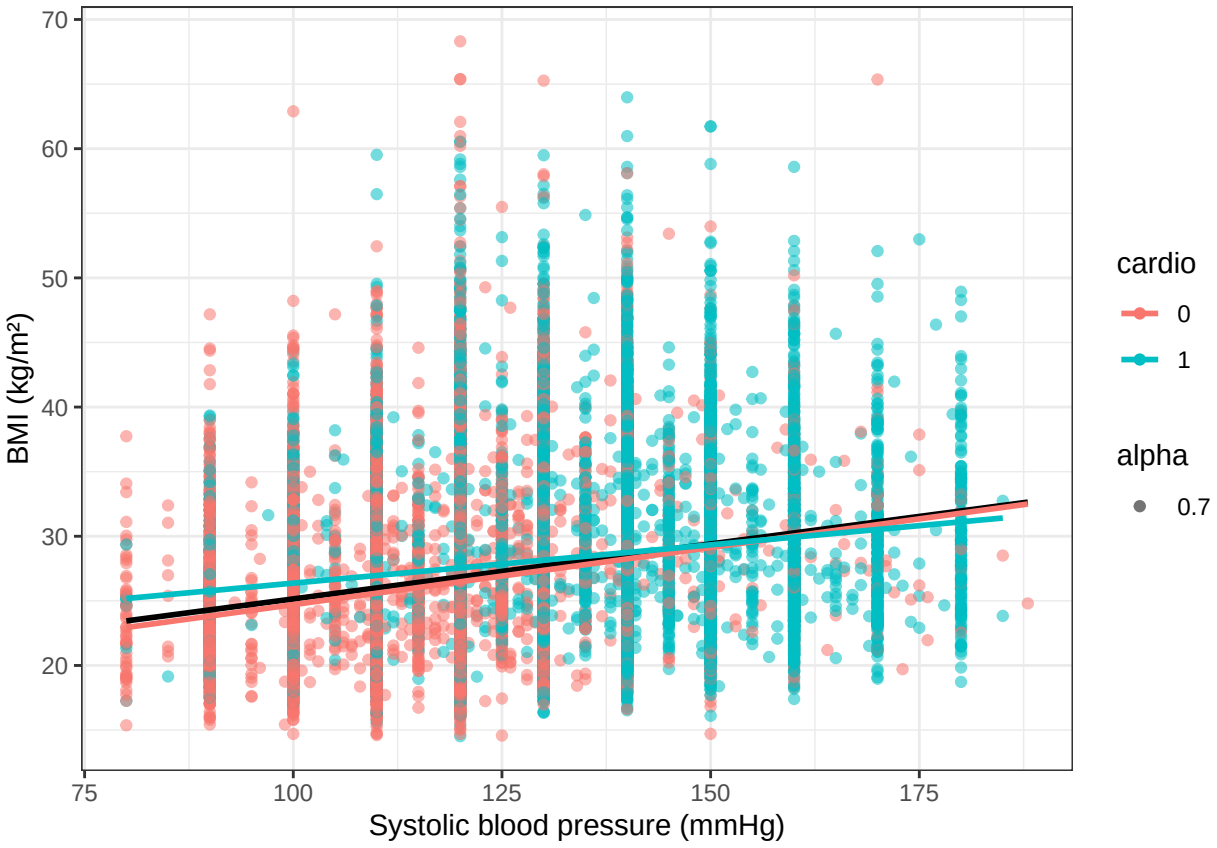

Figure 4: Scatter plot showing the relationship between BMI and systolic blood pressure

The scatter plot (Figure 4) shows a rather weak relationship (black regression line for overall data set), meaning that with increasing systolic blood pressure the BMI is only slightly increasing. People with cardiovascular diseases (cardio = 1 in green), however, tend to have higher systolic blood pressure than people with no cardiovascular disease (cardio = 0 in red), which is indicated by the different distribution over the horizontal axis. Linear regression line of both groups are close to each other, while it is slightly less steep for cardio group 1.

As systolic blood pressure and BMI can be seen as continuous data we can do a Pearson's product-moment correlation to analyze their linear relationship in more detail.

Table 5: BMI vs. systolic blood pressure: Pearson-moment correlation

estimate	0.2405426
statistic	62.4887
p.value	0
parameter	63582
conf.low	0.2332058
conf.high	0.2478519
method	Pearson's product-moment correlation
alternative	two.sided

The results show, that the empiric correlation coefficient is 0.24, which lies within the 95% confidence interval of 0.23 and 0.25, while 0 is not included in the confidence interval. The correlation coefficient of 0.24 shows a positive relationship between BMI and systolic blood pressure, however, is under 0.3, indicating, that the systolic blood pressure and the BMI have only a moderate linear relationship (see also Cohen (1988)). Based on $\alpha = 0.05$ the relationship between both variables is significant as $p <$ 2.22e-16.

The relationship between systolic blood pressure and BMI could be dependent on having or not having cardiovascular diseases. In the above plot (Figure 4) the regression lines for both cardio groups lie quite close together. To be able to analyze, if the relationship between systolic blood pressure and BMI is different between both cardiovascular disease groups, Fisher's z transformation is needed to compare the correlation coefficients of cardio = 0 and cardio = 1 (Diedenhofen 2015). Using the cocor package of (Diedenhofen (2015)), following output is given:

```
##
##   Results of a comparison of two correlations based on independent groups
##
## Comparison between r1.jk = 0.2122 and r2.hm = 0.1655
## Difference: r1.jk - r2.hm = 0.0467
## Group sizes: n1 = 33768, n2 = 29816
## Null hypothesis: r1.jk is equal to r2.hm
## Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)
## Alpha: 0.05
##
## fisher1925: Fisher's z (1925)
##   z = 6.0985, p-value = 0.0000
##   Null hypothesis rejected
##
## zou2007: Zou's (2007) confidence interval
##   95% confidence interval for r1.jk - r2.hm: 0.0317 0.0617
##   Null hypothesis rejected (Interval does not include 0)
```

The results indicate a difference between groups (p-value under $\alpha = 0.05$), i.e. between the correlation between systolic blood pressure and BMI differs between people with (0.17) and without (0.21) cardiovascular diseases, while people with cardiovascular diseases have a smaller correlation coefficient. However, both correlation coefficients stay between 0.1 and 0.3, therefore, they are categorized within the same correlation category (Cohen 1988). This indicates, that the result is not necessarily clinical relevant.

2.3 Correlation between diastolic blood pressure and BMI (Task 3)

Same analysis will be done for the relationship between diastolic blood pressure and BMI.

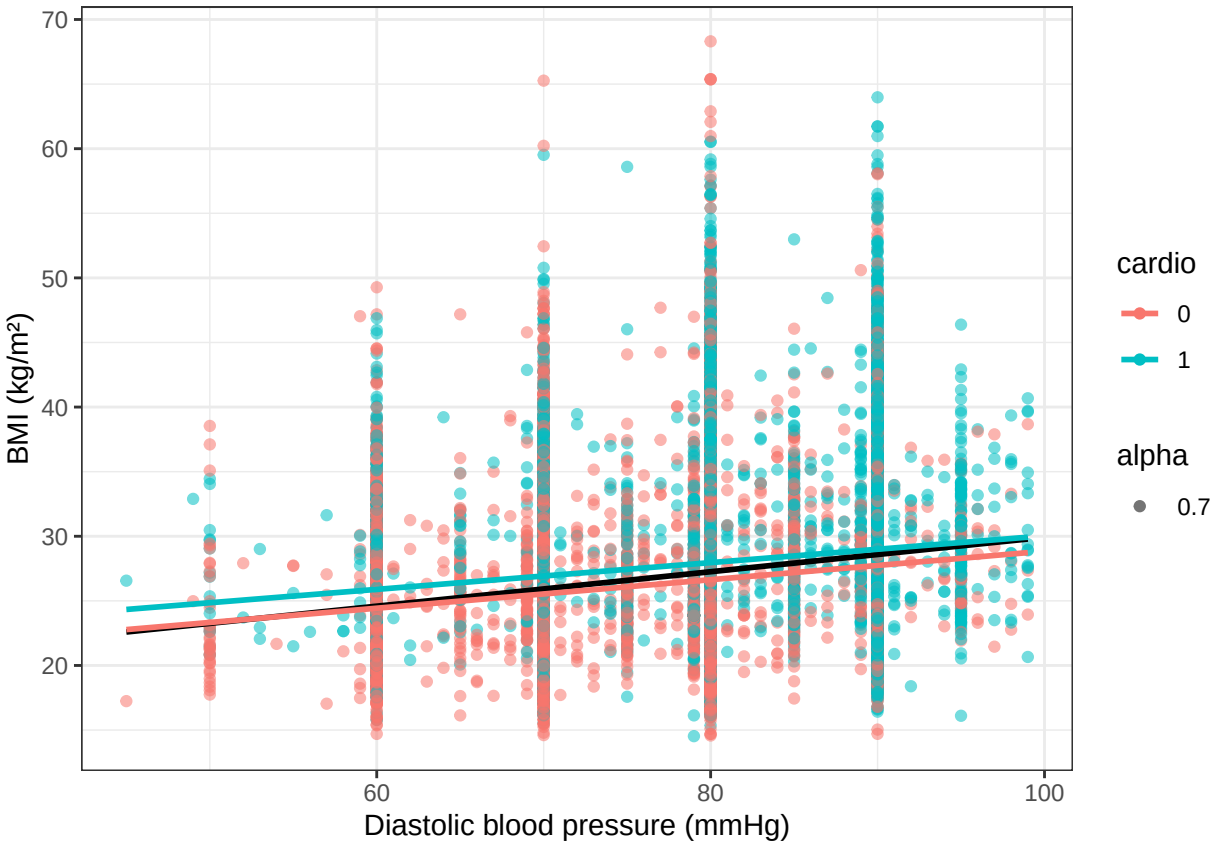

Figure 5: Scatter plot showing the relationship between BMI and diastolic blood pressure

Table 6: BMI vs. diastolic blood pressure: Pearson-moment correlation

estimate	0.2022715
statistic	52.08025
p.value	0
parameter	63582
conf.low	0.194805
conf.high	0.2097146
method	Pearson's product-moment correlation
alternative	two.sided

Similar results are shown for the diastolic blood pressure, resulting in an empiric correlation coefficient of 0.2, which again lies within the 95% confidence interval of 0.19 and 0.21, while 0 is not included in the confidence interval. The confidence interval of 0.2 is again positive and under 0.3, indicating a moderate linear relationship between diastolic blood pressure and BMI (Cohen 1988). The p-value of $p <$ 2.22e-16 shows a significant results based on $\alpha = 0.05$.

Comparing the correlation coefficient between healthy and affected participants, following output is shown according to Diedenhofen (2015):

```
##
##   Results of a comparison of two correlations based on independent groups
##
## Comparison between r1.jk = 0.1734 and r2.hm = 0.1424
## Difference: r1.jk - r2.hm = 0.031
## Group sizes: n1 = 33768, n2 = 29816
## Null hypothesis: r1.jk is equal to r2.hm
## Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)
## Alpha: 0.05
##
## fisher1925: Fisher's z (1925)
##   z = 3.9989, p-value = 0.0001
##   Null hypothesis rejected
##
## zou2007: Zou's (2007) confidence interval
##   95% confidence interval for r1.jk - r2.hm: 0.0158 0.0462
##   Null hypothesis rejected (Interval does not include 0)
```

The correlation coefficient between diastolic blood pressure and BMI differs between people with (0.14) and without (0.17) cardiovascular diseases seems to differ again (p-value under $\alpha = 0.05$), while people with cardiovascular diseases have a again a smaller correlation coefficient. Though, both correlation coefficients are again categorized between 0.1 and 0.3 (Cohen 1988).

2.4 Restrict data to 95% quantile threshold of blood pressure and BMI and repeat 2.2 and 2.3 (Task 4)

The 95% quantile of BMI, systolic and diastolic blood pressure are 36.72, 150 and 90, respectively. The data set is filtered to exclude cases above these thresholds, while systolic and diastolic thresholds are used separately as a filter, which results in two different data sets (cardio95 systolic and diastolic data set).

95% quantile of overall data of systolic blood pressure and BMI

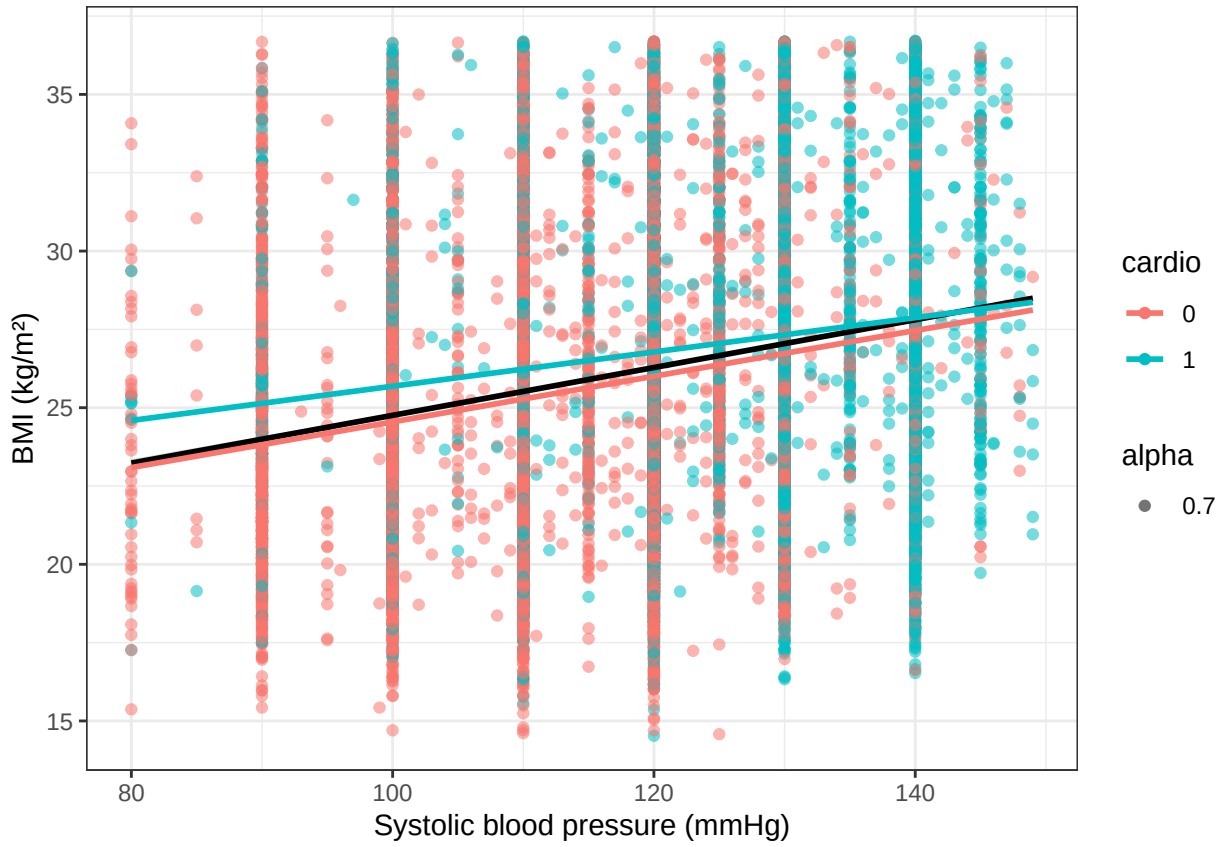

Figure 6: Scatter plot showing the relationship between BMI and systolic blood pressure (below 95% quantile threshold)

Table 7: 95% BMI vs. systolic blood pressure: Pearson-moment correlation

estimate	0.2141019
statistic	51.73774
p.value	0
parameter	55718
conf.low	0.2061652
conf.high	0.2220104
method	Pearson's product-moment correlation
alternative	two.sided

```
##
##   Results of a comparison of two correlations based on independent groups
##
## Comparison between r1.jk = 0.1873 and r2.hm = 0.1494
```

```
## Difference: r1.jk - r2.hm = 0.0379
## Group sizes: n1 = 32027, n2 = 23693
## Null hypothesis: r1.jk is equal to r2.hm
## Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)
## Alpha: 0.05
##
## fisher1925: Fisher's z (1925)
##   z = 4.5489, p-value = 0.0000
##   Null hypothesis rejected
##
## zou2007: Zou's (2007) confidence interval
##   95% confidence interval for r1.jk - r2.hm: 0.0215 0.0542
##   Null hypothesis rejected (Interval does not include 0)
```

95 quantile of overall data of diastolic blood pressure and BMI

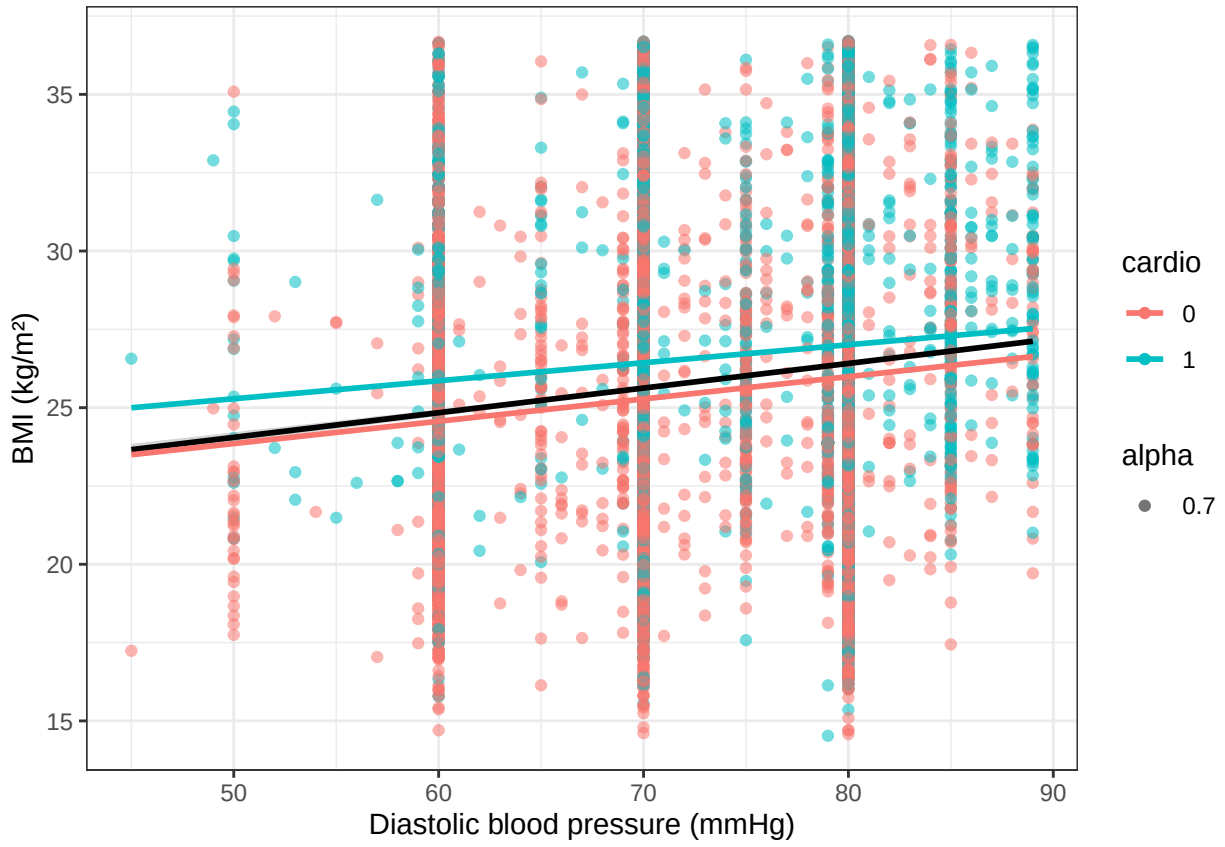

Figure 7: Scatter plot showing the relationship between BMI and diastolic blood pressure (below 95% quantile threshold)

Table 8: 95% BMI vs. systolic blood pressure: Pearson-moment correlation

estimate	0.1201098
statistic	26.27285
p.value	4.816847e-151
parameter	47157
conf.low	0.111205
conf.high	0.1289954
method	Pearson's product-moment correlation
alternative	two.sided

```
##
##   Results of a comparison of two correlations based on independent groups
##
## Comparison between r1.jk = 0.1183 and r2.hm = 0.0754
## Difference: r1.jk - r2.hm = 0.0429
## Group sizes: n1 = 29297, n2 = 17862
## Null hypothesis: r1.jk is equal to r2.hm
## Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)
## Alpha: 0.05
##
## fisher1925: Fisher's z (1925)
##   z = 4.5630, p-value = 0.0000
##   Null hypothesis rejected
##
## zou2007: Zou's (2007) confidence interval
##   95% confidence interval for r1.jk - r2.hm: 0.0245 0.0614
##   Null hypothesis rejected (Interval does not include 0)
```

2.5 Summary relationship bwetween blood pressure and BMI

There seems to be only a moderate relationship between blood pressure and BMI. Although correlation coefficients do differ between cardio groups (without disease > with disease) according to the Pearson-moment correlation, all correlation coefficients stay within the range of 0.1-0.3, therefore, in the same category according to Cohen (1988). It has to be noted, that diastolic blood pressure versus BMI of the affected group while applying a 95% threshold, resulted even in a correlation coefficient lower than 0.1 (0.08).

The following table (Table **??**) summarizes the findings of this relationship. Also, including a threshold of 95% quantile does not change anything regarding the relation of blood pressure and BMI. In general, however, the correlation coefficients decrease slightly for every comparison. This can be explained by

	Overall	With disease	Without disease	p-Value
Systolic vs. BMI	0.24	0.17	0.21	< 2.22e-16
Diastolic vs. BMI	0.20	0.14	0.17	< 2.22e-16
Systolic95 vs. BMI95	0.21	0.15	0.19	< 2.22e-16
Diastolic95 vs. BMI95	0.12	0.08	0.12	< 2.22e-16

2.6 Distribution of age within both cardio groups (Task 5)

Another theory is that age might be differently distributed in both cardiovascular groups (with vs. without disease). The age distribution within both groups can be extracted from Table 9.

Table 9: Overview of age distribution for each cardio group

Cardio Group	mean	SD	max	min
0	51.6	6.8	65	30
1	55.0	6.3	65	39

Cardio group 0 = healthy, 1 = affected
SD = standard deviation

Furthermore, the average age difference between both groups is 3.35, the population with cardiovascular diseases being older than the healthy population. The boxplot and density curve of Figure 8 strengthen this hypothesis.

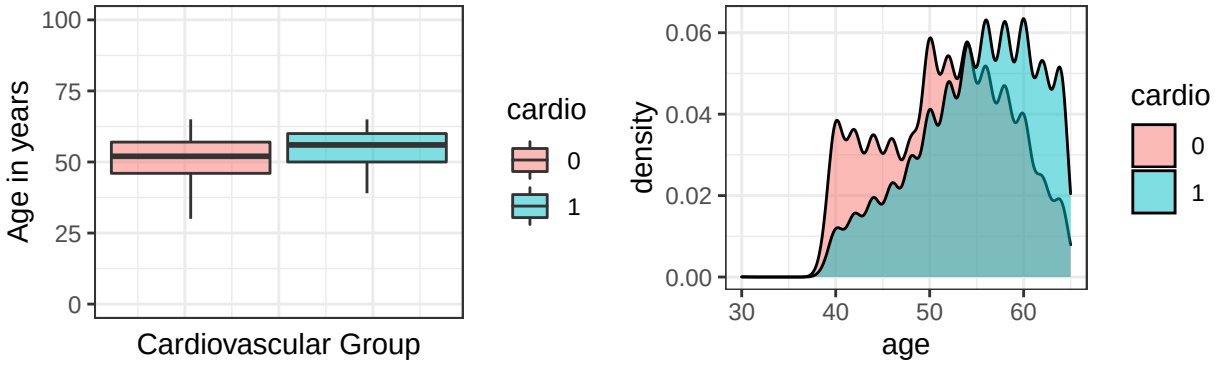

Figure 8: Distribution of age

2.7 Visualize the distribution of age for gender and cardio (Task 6)

While age seems to be slightly different distributed between cardio groups, gender looks fairly equally distributed between healthy and affected persons (Figure 9)

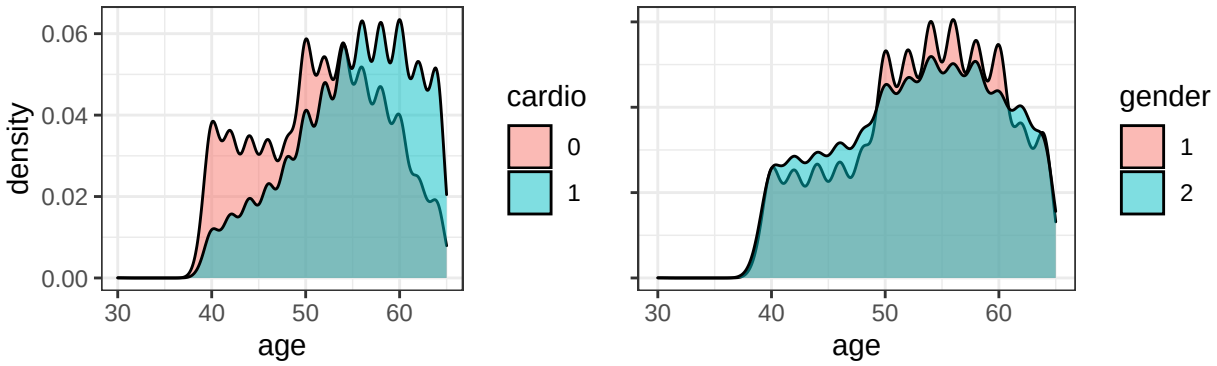

Figure 9: Age distribution for gender and cardio

2.8 Additionally include variable glucose into plot of 2.6 (Task 7)

Taking, furthermore, the different types of glucose into account a slight different distribution between group 1-2 and group 3 can be visually detected (Figure 10), while especially older people seem to have a glucose level well above the normal level.

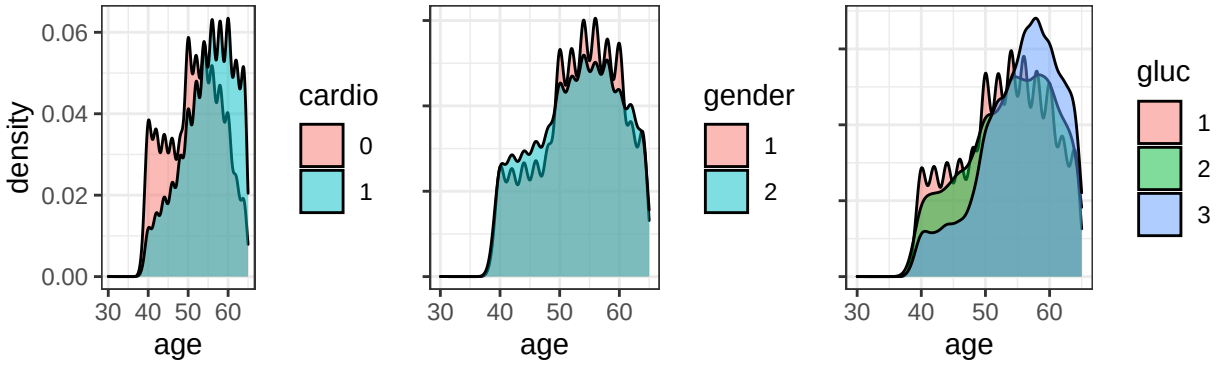

Figure 10: Age distribution for cardio, gender and glucose

2.9 Risk factors smoking, alcohol and physical activity (Task 8)

Further risk factors, which might trigger cardiovascular diseases may be smoking, alcohol abuse or insufficient physical activity. Table 10 summarizes the distribution of these variables between the two types of *cardio*. To quantify the differences between both groups of cardio, a Chi-Square test will be undertaken for each risk factor.

It can be concluded, that, interestingly, the cardio group 0, which are people without cardiovascular diseases do smoke more and more frequently abuse alcohol as well as are less physically active than participants with cardiovascular diseases.

Table 10: Relative frequency of risk factors in percent

Cardio group	Smoking	Alcohol	Physical activity
Cardio 0	9.26	5.51	18.13
Cardio 1	8	4.75	21.38
Chi-Square	1.6856e-08	1.3309e-05	< 2.22e-16

References

Cohen, Jacob. 1988. *Statistical Power Analysisfor the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum Associates.

Diedenhofen, Jochen, Birk AND Musch. 2015. "Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations." *PLOS ONE* 10 (4): 1–12. https://doi.org/10.1371/journal.pone.0121945.