DAP\_figures

sophie whikehart

2023-12-04

Table of Contents

[Figure 1 3](#_Toc152794026)

[Conduct descriptive analysis for the quantitative variables by creating a summary table for the quantitative variables [age, alcoh, hdl, insulin] 3](#_Toc152794027)

[Figure 2 4](#_Toc152794028)

[Histogram to show distribution of alcohol 4](#_Toc152794029)

[Figure 3 5](#_Toc152794030)

[Histogram to show distribution of hdl 5](#_Toc152794031)

[Figure 4 7](#_Toc152794032)

[Histogram to show distribution of serum insulin levels 7](#_Toc152794033)

[Figure 5 7](#_Toc152794034)

[Frequency distribution of categorical variables 7](#_Toc152794035)

[Figure 6 8](#_Toc152794036)

[Bar chart of Family History of Heart Attack 8](#_Toc152794037)

[Figure 7 9](#_Toc152794038)

[Bar chart on smoking status 9](#_Toc152794039)

[Figure 8 10](#_Toc152794040)

[Bar chart to show alcohol consumption 10](#_Toc152794041)

[Figure 9 11](#_Toc152794042)

[Box plot of CHD and HDL Cholestrol 11](#_Toc152794043)

[Figure 10 12](#_Toc152794044)

[Two sample t-test to determine if HDL cholesterol levels are associated with CHD 12](#_Toc152794045)

[Figure 11 13](#_Toc152794046)

[Final summary table reporting HDl cholesterol, alcohol, serum insulin, smoking status and family history of heart attack by CHD 13](#_Toc152794047)

[Figure 12 14](#_Toc152794048)

[Overall HDL cholesterol by CHD 14](#_Toc152794049)

[Figure 13 15](#_Toc152794050)

[Bar chart on categorized HDL 15](#_Toc152794051)

[Figure 14 16](#_Toc152794052)

[HDL cholesterol by CHD - Family History of Heart Attack = Yes 16](#_Toc152794053)

[Figure 15 17](#_Toc152794054)

[HDL cholesterol by CHD - Family History of Heart Attack = No 17](#_Toc152794055)

knitr::opts\_chunk$set(echo = TRUE,  
 warning = FALSE,  
 message = FALSE,  
 comment = "")  
  
# load libraries   
library(e1071) # Needed for skewness()  
  
library(tidyverse) # Need to summarize data and for %>% function

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2) # Needed to create graphs using the ggplot2 package  
  
library(flextable) # Used to create a nice summary table in Word

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(gtsummary) # Used to create a nice summary table in Word

## #StandWithUkraine  
##   
## Attaching package: 'gtsummary'  
##   
## The following objects are masked from 'package:flextable':  
##   
## as\_flextable, continuous\_summary

#import data   
chdclass <- read.csv(file = "chd3200 (1).csv")  
  
#create factor variables   
# MI family history  
chdclass$fhha.f <- factor(chdclass$fhha, levels = c(0,1),   
 labels = c("No", "Yes"))  
  
# Coronary heart disease (CHD)  
chdclass$incchd.f <- factor(chdclass$incchd, levels = c(0,1),   
 labels = c("No", "Yes"))  
# Smoking status  
chdclass$smoke.f <- factor(chdclass$smoke, levels = c(1,2,3),   
 labels = c("Never", "Former", "Current"))

# Figure 1

## Conduct descriptive analysis for the quantitative variables by creating a summary table for the quantitative variables [age, alcoh, hdl, insulin]

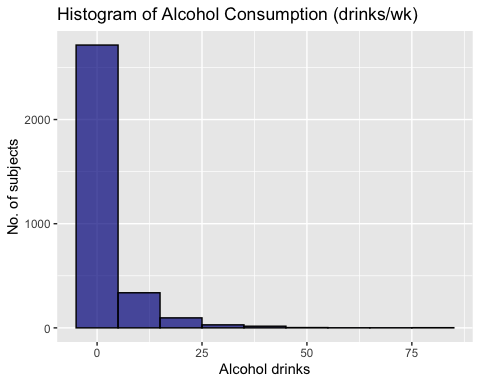
chdclass %>% select(age, alcoh, hdl, insulin) %>%  
 tbl\_summary(missing = "ifany",  
 type = all\_continuous() ~ "continuous2",  
 statistic = all\_continuous() ~ c("{mean} ({sd})",  
 "{median} ({p25}, {p75})",   
 "{min}, {max}",  
 "{skewness}"),  
 digits = list(all\_continuous() ~ 1),   
 label = list(age ~ "Age (y)",  
 alcoh ~ "Alcohol consumption (drinks/wk)",  
 hdl ~ "HDL cholesterol (mg/dL)",  
 insulin ~ "Serum insulin (IU/ml)")  
 ) %>%  
 modify\_header(label = "\*\*Variable\*\*") %>%   
 add\_stat\_label(label = all\_continuous() ~ c("Mean (SD)",   
 "Median (IQR)",  
 "Min to Max", "Skewness"))

| **Variable** | **N = 3,200** |
| --- | --- |
| Age (y) |  |
| Mean (SD) | 71.6 (4.4) |
| Median (IQR) | 71.0 (68.0, 75.0) |
| Min to Max | 65.0, 82.0 |
| Skewness | 0.5 |
| Alcohol consumption (drinks/wk) |  |
| Mean (SD) | 2.6 (6.5) |
| Median (IQR) | 0.0 (0.0, 1.3) |
| Min to Max | 0.0, 77.0 |
| Skewness | 4.4 |
| HDL cholesterol (mg/dL) |  |
| Mean (SD) | 54.9 (15.6) |
| Median (IQR) | 53.0 (44.0, 64.0) |
| Min to Max | 15.0, 149.0 |
| Skewness | 0.9 |
| Serum insulin (IU/ml) |  |
| Mean (SD) | 16.1 (20.0) |
| Median (IQR) | 13.0 (10.0, 17.0) |
| Min to Max | 3.0, 400.0 |
| Skewness | 12.8 |

# Figure 2

# Histogram to show distribution of alcohol

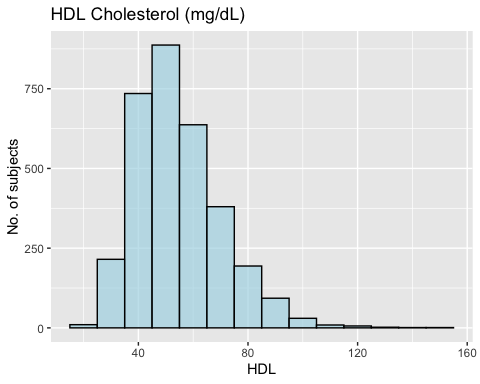
ggplot(data=chdclass, aes(x=alcoh)) +  
geom\_histogram(color="black", fill="blue4", binwidth=10, position="identity", alpha=0.7) +  
labs(x="Alcohol drinks",  
y="No. of subjects",  
title="Histogram of Alcohol Consumption (drinks/wk)")



# Figure 3

# Histogram to show distribution of hdl

ggplot(data=chdclass, aes(x=hdl)) +  
geom\_histogram(color="black", fill="lightblue", binwidth=10, position="identity", alpha=0.7) +  
labs(x="HDL",  
y="No. of subjects",  
title="HDL Cholesterol (mg/dL)")



#create categories  
chdclass$alcoh <- cut(chdclass$alcoh, breaks = c(-Inf,0,7.000,Inf))  
  
#check categories  
table(chdclass$alcoh, exclude=FALSE)

(-Inf,0] (0,7] (7, Inf]   
 1537 1240 423

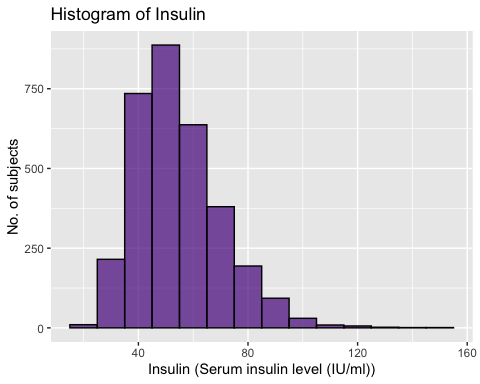
#create a factor variable   
chdclass$alcoh.f <- factor(chdclass$alcoh,  
 levels = c("(-Inf,0]", "(0,7]",  
 "(7, Inf]"),  
 labels = c("0", ">0 to 7", ">7"))  
#check factor variable  
table(chdclass$alcoh.f)

0 >0 to 7 >7   
 1537 1240 423

# Figure 4

## Histogram to show distribution of serum insulin levels

ggplot(data=chdclass, aes(x=hdl)) +  
geom\_histogram(color="black", fill="purple4", binwidth=10, position="identity", alpha=0.7) +  
labs(x="Insulin (Serum insulin level (IU/ml))",  
y="No. of subjects",  
title="Histogram of Insulin")



# Figure 5

## Frequency distribution of categorical variables

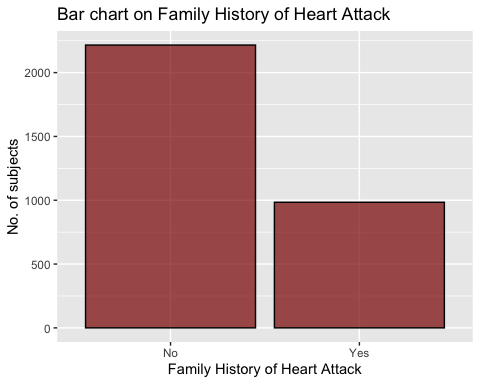
theme\_gtsummary\_journal(journal = "jama") # Use the JAMA theme  
  
chdclass %>% select(fhha.f, incchd.f, smoke.f, alcoh.f) %>%  
 tbl\_summary(missing = "ifany",  
 type = list(fhha.f ~ "categorical",  
 incchd.f ~ "categorical"),  
 digits = list(all\_categorical() ~ c(0,1)),   
 label = list(fhha.f ~ "Family history MI",  
 incchd.f ~ "Coronary heart disease",  
 smoke.f ~ "Smoking status",  
 alcoh.f ~ "Average consumption of alcohol per week")  
 ) %>%  
 modify\_header(label = "\*\*Variable\*\*")

| **Variable** | **N = 3,200** |
| --- | --- |
| Family history MI, n (%) |  |
| No | 2,216 (69.3) |
| Yes | 984 (30.8) |
| Coronary heart disease, n (%) |  |
| No | 2,769 (86.5) |
| Yes | 431 (13.5) |
| Smoking status, n (%) |  |
| Never | 1,558 (48.7) |
| Former | 1,266 (39.6) |
| Current | 376 (11.8) |
| Average consumption of alcohol per week, n (%) |  |
| 0 | 1,537 (48.0) |
| >0 to 7 | 1,240 (38.8) |
| >7 | 423 (13.2) |

# Figure 6

## Bar chart of Family History of Heart Attack

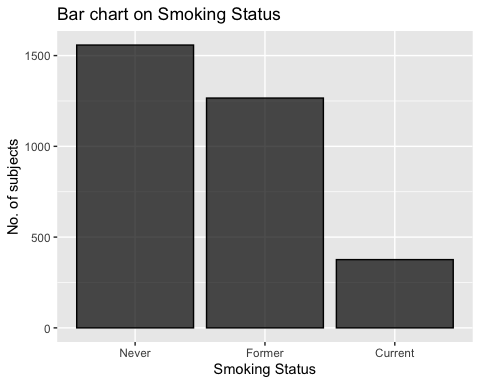
ggplot(data=chdclass, aes(x=fhha.f)) +  
geom\_bar(color="black", fill="red4", binwidth=10, position="identity", alpha=0.7) +  
labs(x="Family History of Heart Attack",  
y="No. of subjects",  
title="Bar chart on Family History of Heart Attack")



# Figure 7

## Bar chart on smoking status

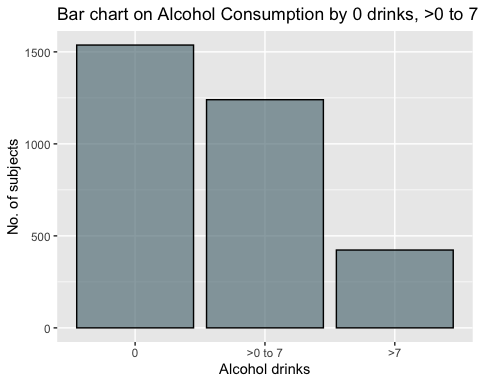
ggplot(data=chdclass, aes(x=smoke.f)) +  
geom\_bar(color="black", fill="black", binwidth=10, position="identity", alpha=0.7) +  
labs(x="Smoking Status",  
y="No. of subjects",  
title="Bar chart on Smoking Status")



# Figure 8

## Bar chart to show alcohol consumption

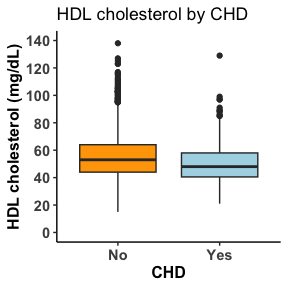
ggplot(data=chdclass, aes(x=alcoh.f)) +  
geom\_bar(color="black", fill="lightblue4", binwidth=10, position="identity", alpha=0.7) +  
labs(x="Alcohol drinks",  
y="No. of subjects",  
title="Bar chart on Alcohol Consumption by 0 drinks, >0 to 7 and >7 drinks per week")



# Figure 9

## Box plot of CHD and HDL Cholestrol

ggplot(data=chdclass, aes(x=incchd.f, y=hdl, fill=incchd.f)) +  
 geom\_boxplot() +  
 labs(x="CHD",  
 y="HDL cholesterol (mg/dL)",  
 title="HDL cholesterol by CHD") +   
 scale\_y\_continuous(limits=c(0, 140), breaks=seq(0,150,20)) +   
 theme\_classic() +  
 theme(legend.position="none") +  
 theme(  
 legend.position = "none",  
 legend.title = element\_text(colour="black", size=10,   
 face="bold"),  
 legend.text = element\_text(colour="black", size=10,   
 face="bold"),  
 axis.title.x = element\_text(face="bold", size=12),  
 axis.text.x = element\_text(face="bold", size=11),  
 axis.title.y = element\_text(face="bold", size=12),  
 axis.text.y = element\_text(face="bold", size=11)) +  
 scale\_fill\_manual(values=c("orange","lightblue"))



# Figure 10

## Two sample t-test to determine if HDL cholesterol levels are associated with CHD

# t.test   
t.test(hdl ~ incchd.f, conf.level=0.95, var.equal = FALSE, data=chdclass, alternative= "two.sided")

Welch Two Sample t-test  
  
data: hdl by incchd.f  
t = 6.1893, df = 600.74, p-value = 1.119e-09  
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0  
95 percent confidence interval:  
 3.190391 6.156090  
sample estimates:  
 mean in group No mean in group Yes   
 55.54099 50.86775

# Figure 11

## Final summary table reporting HDl cholesterol, alcohol, serum insulin, smoking status and family history of heart attack by CHD

theme\_gtsummary\_journal(journal = "jama") # Use the JAMA theme  
  
chdclass %>% select(incchd.f, age, alcoh.f, hdl, insulin, smoke.f, fhha.f) %>%  
 tbl\_summary(by = incchd.f,  
 missing = "ifany",  
 statistic = list(age ~ c("{median} ({p25}, {p75})"),  
 hdl ~ c("{mean} ({sd})"),  
 insulin ~ c("{median} ({p25}, {p75})")  
 ),  
 percent = "row",  
 type = list(fhha.f ~ "categorical",  
 alcoh.f ~ "categorical"),  
 digits = list(all\_categorical() ~ c(0,1)),   
 label = list(hdl ~ "HDL cholesterol (mg/dL)",  
 insulin ~ "Serum insulin (IU/ml)",  
 smoke.f ~ "Smoking status",  
 age ~ "Age (years)",  
 alcoh.f ~ "Average number of alcoholic drinks per week",  
 fhha.f ~ "Family history of heart attack")  
 ) %>%  
 add\_overall() %>%  
 modify\_header(label = "\*\*Variable\*\*") %>%  
 modify\_spanning\_header(stat\_1 ~ "Coronary Heart Disease",  
 stat\_2 ~ "Coronary Heart Disease") %>%   
 add\_p()

|  | | Coronary Heart Disease | |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Overall**, N = 3,200 | **No**, N = 2,769 | **Yes**, N = 431 | **p-value**1 |
| Age (years), Median (IQR) | 71 (68, 75) | 71 (68, 75) | 72 (69, 77) | <0.001 |
| Average number of alcoholic drinks per week, n (%) |  |  |  | 0.046 |
| 0 | 1,537 (100.0) | 1,311 (85.3) | 226 (14.7) |  |
| >0 to 7 | 1,240 (100.0) | 1,078 (86.9) | 162 (13.1) |  |
| >7 | 423 (100.0) | 380 (89.8) | 43 (10.2) |  |
| HDL cholesterol (mg/dL), Mean (SD) | 55 (16) | 56 (16) | 51 (14) | <0.001 |
| Serum insulin (IU/ml), Median (IQR) | 13 (10, 17) | 12 (9, 17) | 13 (10, 19) | 0.002 |
| Smoking status, n (%) |  |  |  | 0.005 |
| Never | 1,558 (100.0) | 1,376 (88.3) | 182 (11.7) |  |
| Former | 1,266 (100.0) | 1,065 (84.1) | 201 (15.9) |  |
| Current | 376 (100.0) | 328 (87.2) | 48 (12.8) |  |
| Family history of heart attack, n (%) |  |  |  | <0.001 |
| No | 2,216 (100.0) | 1,947 (87.9) | 269 (12.1) |  |
| Yes | 984 (100.0) | 822 (83.5) | 162 (16.5) |  |
| 1Wilcoxon rank sum test; Pearson's Chi-squared test | | | | |

# Figure 12

## Overall HDL cholesterol by CHD

#Categorizing HDL cholesterol  
chdclass$hdlcat <- cut(chdclass$hdl, breaks = c(-Inf,40,Inf))  
  
table(chdclass$hdlcat, exclude = FALSE)

(-Inf,40] (40, Inf]   
 550 2650

# Create a factor variable  
chdclass$hdlcat.f <- factor(chdclass$hdlcat,   
 levels = c("(-Inf,40]", "(40, Inf]"),   
 labels = c("<40", ">=40"))  
  
# Check factor variable  
table(chdclass$hdlcat.f)

<40 >=40   
 550 2650

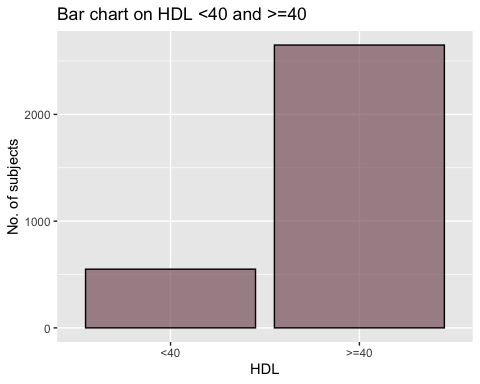
#Create cross-tab for overall HDL cholesterol and CHD incidence  
chdclass |>   
 select(incchd.f, hdlcat.f) |>  
 tbl\_summary(by = incchd.f,  
 missing = "no",  
 percent = "row",  
 statistic = all\_categorical() ~ "{n} ({p}%)",  
 label = list(incchd.f ~ "Incidence of CHD",  
 hdlcat.f ~ "HDL cholesterol")) %>%  
 modify\_spanning\_header(stat\_1 ~ "Overall HDL and CHD")

|  | Overall HDL and CHD |  |
| --- | --- | --- |
| **Characteristic** | **No**, N = 2,769 | **Yes**, N = 431 |
| HDL cholesterol, n (%) |  |  |
| <40 | 442 (80%) | 108 (20%) |
| >=40 | 2,327 (88%) | 323 (12%) |

# Figure 13

## Bar chart on categorized HDL

ggplot(data=chdclass, aes(x=hdlcat.f)) +  
geom\_bar(color="black", fill="pink4", binwidth=10, position="identity", alpha=0.7) +  
labs(x="HDL",  
y="No. of subjects",  
title="Bar chart on HDL <40 and >=40")



# Figure 14

## HDL cholesterol by CHD - Family History of Heart Attack = Yes

#Categorizing HDL cholesterol  
chdclass$hdlcat <- cut(chdclass$hdl, breaks = c(-Inf,40,Inf))  
  
table(chdclass$hdlcat, exclude = FALSE)

(-Inf,40] (40, Inf]   
 550 2650

# Create a factor variable  
chdclass$hdlcat.f <- factor(chdclass$hdlcat,   
 levels = c("(-Inf,40]", "(40, Inf]"),   
 labels = c("<40", ">=40"))  
  
# Check factor variable  
table(chdclass$hdlcat.f)

<40 >=40   
 550 2650

#Create cross-tab for FFHA = YES by HDL cholesterol and CHD incidence  
chdclass %>%   
 filter(fhha.f == "Yes") %>%  
 select(incchd.f, hdlcat.f) %>%  
 tbl\_summary(by = incchd.f,  
 missing = "no",  
 percent = "row",  
 statistic = all\_categorical() ~ "{n} ({p}%)",  
 label = list(incchd.f ~ "Incidence of CHD",  
 hdlcat.f ~ "HDL cholesterol (mg/dL)")) %>%  
modify\_header(label = "\*\*Variable\*\*") %>%  
 modify\_spanning\_header(stat\_1 ~ "Family History of Heart Attack = Yes to CHD")

|  | Family History of Heart Attack = Yes to CHD |  |
| --- | --- | --- |
| **Variable** | **No**, N = 822 | **Yes**, N = 162 |
| HDL cholesterol (mg/dL), n (%) |  |  |
| <40 | 135 (75%) | 44 (25%) |
| >=40 | 687 (85%) | 118 (15%) |

# Figure 15

## HDL cholesterol by CHD - Family History of Heart Attack = No

#Categorizing HDL cholesterol  
chdclass$hdlcat <- cut(chdclass$hdl, breaks = c(-Inf,40,Inf))  
  
table(chdclass$hdlcat, exclude = FALSE)

(-Inf,40] (40, Inf]   
 550 2650

# Create a factor variable  
chdclass$hdlcat.f <- factor(chdclass$hdlcat,   
 levels = c("(-Inf,40]", "(40, Inf]"),   
 labels = c("<40", ">=40"))  
  
# Check factor variable  
table(chdclass$hdlcat.f)

<40 >=40   
 550 2650

#Create cross-tab for FFHA = NO by HDL cholesterol and CHD incidence  
chdclass %>%   
 filter(fhha.f == "No") %>%  
 select(incchd.f, hdlcat.f) %>%  
 tbl\_summary(by = incchd.f,  
 missing = "no",  
 percent = "row",  
 statistic = all\_categorical() ~ "{n} ({p}%)",  
 label = list(incchd.f ~ "Incidence of CHD",  
 hdlcat.f ~ "HDL cholesterol (mg/dL)")) %>%  
modify\_header(label = "\*\*Variable\*\*") %>%  
 modify\_spanning\_header(stat\_1 ~ "Family History of Heart Attack = No to CHD")

|  | Family History of Heart Attack = No to CHD |  |
| --- | --- | --- |
| **Variable** | **No**, N = 1,947 | **Yes**, N = 269 |
| HDL cholesterol (mg/dL), n (%) |  |  |
| <40 | 307 (83%) | 64 (17%) |
| >=40 | 1,640 (89%) | 205 (11%) |