

Machine learning project

part 1

המחלקה להנדסת תעשייה וניהול בן גוריון



May 10, 2020

חאלד סובח

שירה בן צבי

תוכן עניינים

[הגדרת הבעיה 1](#_Toc40043069)

[תיאור כללי של עולם התוכן הנחקר 1](#_Toc40043070)

[הגדרת שאלת המחקר 1](#_Toc40043071)

[הבנת הנתונים 1](#_Toc40043072)

[תיעוד מקורות הנתונים ומשמעותם 1](#_Toc40043073)

[הסתברויות אפריוריות וקשרים מעניינים בין מאפיינים ועם משתנה המטרה 2](#_Toc40043074)

[קשרים בין מאפיינים 4](#_Toc40043075)

[מאפיינים בעלי השפעה על משתנה המטרה 5](#_Toc40043076)

[איכות הנתונים 6](#_Toc40043077)

[הכנת הנתונים 7](#_Toc40043078)

[בחירת מאפיינים 7](#_Toc40043079)

[טיפול פרטני במאפיינים 8](#_Toc40043080)

# 

# הגדרת הבעיה

## תיאור כללי של עולם התוכן הנחקר

מחלות לב הן סיבת המוות השנייה בישראל לאחר מחלת הסרטן. כמו בארץ כך גם בכל העולם הנושא מהווה דאגה גדולה במערכות הבריאות מכיוון שזהו נטל כבד על החולה ובני משפחתו וכן מערכת הבריאות כולה. לכן נערכים מחקרים רבים על תחלואה ומוות ממחלות לב במטרה לנטר גורמי סיכון באוכלוסייה וקבוצות הנמצאות בסיכון מוגבר לחלות.

## הגדרת שאלת המחקר

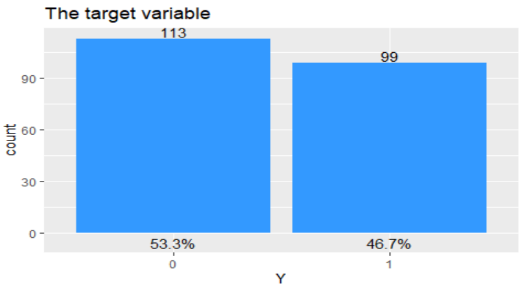
בפרוייקט זה נרצה להיעזר בשיטות של לימוד מכונה על מנת לנבא בצורה המדוייקת ביותר האם למטופל יש סיכון לחלות במחלת לב או לא (התקף לב). זאת נעשה על בסיס מאפיינים (תכונות) שנלקחו ונמדדו עבור כל מטופל ומתועדים בסט הנתונים.

# הבנת הנתונים

## תיעוד מקורות הנתונים ומשמעותם

מקור הנתונים בפרויקט זה הוא סט הנתונים "Heart Disease Data Set" שנלקח ממאגר הנתונים של UCI. סט הנתונים המלא מכיל 76 תכונות (attributes) אך אנו נשתמש בתת-סט ממנו הסוקר 13 תכונות מתוכן. הנתונים נאספו על ידי רופאים על מטופלים במסגרת מחקרים לימודיים. נהיה מעוניינים בעזרת כלים של לימוד מכונה ללמוד את סט הנתונים ועל ידו לנסות לנבא האם המטופל יחווה מחלת לב או לא.

משתנה המטרה Y: בינארי, האם המטופל חווה מחלת לב(1), או לא(0).

**חשוב**: הפכנו את משתנה המטרה: אחרת התכונות סותרות את הידע שכבר יש לנו על מחלות לב– לדוגמא ככל שעולה הגיל כך עולה הסיכוי לחלות במחלת לב [[1]](#endnote-1). על מנת לאשש הסתכלנו על סט הנתונים המקורי וראינו שזה אכן הפוך (נספח 1). כמו כן שינינו את משתנה המטרה ל-Y גדול ע"מ לא להתבלבל עם צירy (נספח 1).

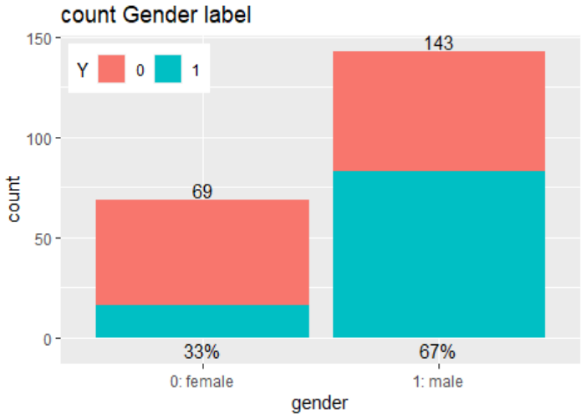
בהסתכלות על התפלגות Y ניתן לראות כמעט שוויון מוחלט בין כמות החולים לכמות הבריאים. ניתן להסיק שאסיפת הנתונים היו ממדגם לא הטרוגני.

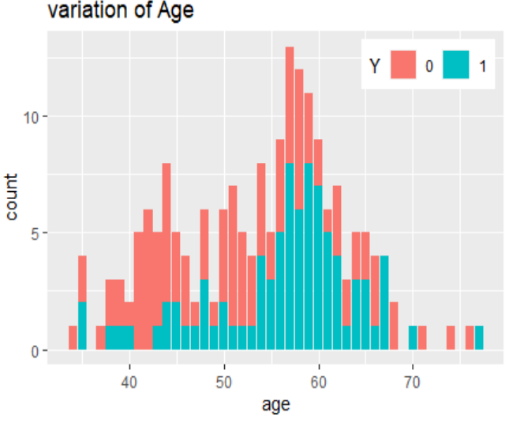
סט הנתונים נכלל בתוכו שני סוגים של משתנים: משתנים רציפים ומשתנים בדידים:

|  |  |  |
| --- | --- | --- |
| **תכונה** | **משמעות** | **סוג** |
| **Id** | מספר זהות. | רציף |
| **Age** | גיל בשנים. | רציף |
| **trestbps** | לחץ דם במנוחה ב- mmHg. | רציף |
| **chol** | הכולסטרול בסרום ב- mg / dl. | רציף |
| **thalach** | פעימות בדקה שהלב שואב תחת לחץ מרבי. | רציף |
| **oldpeak** | דיכוי ST הנגרם על ידי פעילות גופנית יחסית למנוחה. | רציף |
| **Gender** | מין, 1 עבור גבר 0 עבור אישה. | בינארי |
| **FBS** | סוכר בדם בצום. 1 משמעותו גדול מ- 120, 0 קטן או שווה ל-120. | בינארי |
| **CP** | סוג כאב בחזה: 0. אנגינה[[2]](#endnote-2) טיפוסית, 1. אנגינה לא טיפוסית, 2. כאבים שאינם אנגינה 3. אסימפטומטי. | קטגוריאלי |
| **restECG** | פעילות חשמלית של הלב במנוחה: 0.נורמלי, 1.סובל מאי-נורמליות ב-ST-T, 2.מראה היפרטרופיה סבירה. | קטגוריאלי |
| **exang** | תעוקת לב באימון גופני: 0 משמעותו אין, 1 קיימת. | בינארי |
| **Slope** | המדרון בקטע השיא ב-ST. 0.עולה, 1.שטוח, 2.יורד. | קטגוריאלי |
| **CA** | מספר העורקים העיקריים החסומים מתוך 4. 0-3, 4.לא ידוע(נניח ש-4 = נתון חסר). | קטגוריאלי |
| **Thal** | תוצאת בדיקת לחץ Thalium. 0.ערך חסר, 1.פגם קבוע, 2.נורמאלי, 3.פגם הפיך. | קטגוריאלי |

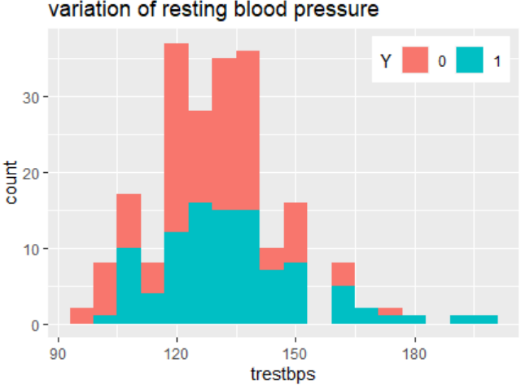
## הסתברויות אפריוריות וקשרים מעניינים בין מאפיינים ועם משתנה המטרה

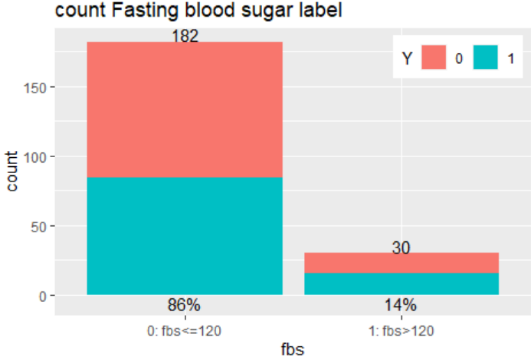
נציג את ההסתברויות האפריוריות שמצאנו לנכון להציג, של המשתנה המוסבר ושל משתנים מסבירים נוספים, וכן האם הן מאוזנות ומייצגות את המציאות:

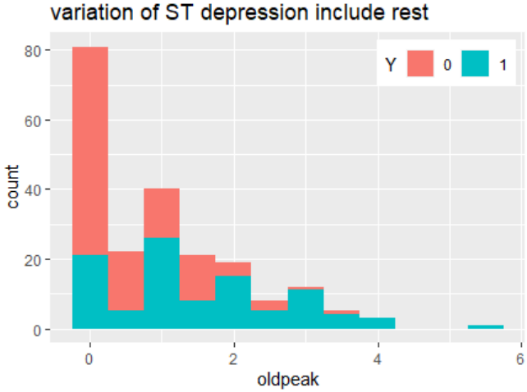
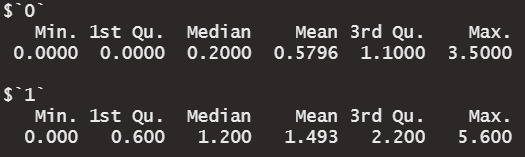
**משתנה מוסבר Y** -ההסתברות לחלות במחלת לב לפי הנתונים הוא 0.467. ניתן לראות שהוא מאוזן, גם ביחס לסט האימון. אבל הוא לא מציאותי. כלומר, ההסתברות האפריורית לחלות במחלת לב קטנה מסיכוי זה, ועומדת על 0.25 עבור בני 80+ ו0.15 עבור בני 60-79 (איגוד הלב האמריקאי, 2015).

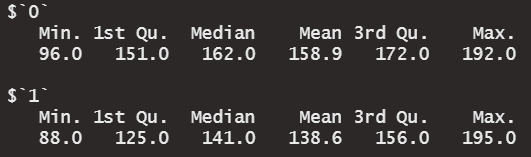
**Gender -** בגרף ניתן לראות שמרבית המדגם הם גברים 67% לעומת 33% נשים. כמובן שזה לא מאוזן, שכן בעולם יש 51% נשים ו-50% גברים. עם זאת, ניתן לראות שישנו חלק גדול יותר מתוך אוכלוסיית הגברים החולים במחלת לב (Y=1). אמנם המשתנה אינו מאוזן אך ההסתברות האפריורית עבור גברים לחלות במחלת לב גבוהה יותר במציאות, כמעט פי 2 מנשים (איגוד הלב האמריקאי, 2015).

**Age** – הגיל המינימלי בסט הנתונים הוא 34. זהו נתון מציאותי שכן ההסתברות האפריורית לחלות במחלת לב מתחת לגיל 39 הוא 0.006[[3]](#endnote-3) . נוסף לכך נראה שמרבית האוכלוסייה החולה נמצאת בין גילאי 45-65, זהו מצב לא מאוזן אך מציאותי – כמו כן נראה שכמות חולי הלב עולה עם הגיל וזהו גם כן נתון מציאותי.

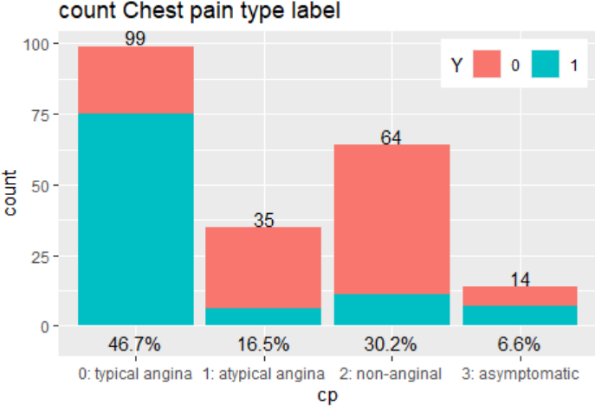
**trestBPS**- לחץ הדם מרוכז בעיקר סביב ערכי 120-140. זה נתון מציאותי שכן לחץ דם תקין הינו עד 130. יש מעט מאוד דגימות מעל לחץ דם 150 ונראה שמרביתם חולים חיוביים. זהו נתון מציאותי גם כן – שכן ידוע שלחץ דם גבוה הוא אחד הסימפטומים למחלת לב (איגוד הלב האמריקאי, 2015).

**FBS**- רמת הסוכר בדם בצום מתחלקת לשתי קבוצות – מעל 120 ומתחת ל-120. ניתן לראות שנתון זה ממש לא מאוזן ו86% מהתצפיות בקבוצה של מתחת ל-120. עם זאת, זה נתון מציאותי שכן FBS תקין הוא עד 100, וערכים שבין 100 ל-125 מוגדרים פרה-סוכרת ומעל 126 זוהי כבר סוכרת. על כן הגיוני שמרבית הערכים הם עד 120, אך נראה שבערך מחצית מכל קבוצה הם חולי לב - מה שמעיד שFBS לא מוסיף לנו מידע מיוחד.

**Oldpeak-** נראה שהנתונים לא מאוזנים כלל ומרביתם בין 0 ל-1. לרוב ערך מעל 0.5 הוא סימן לאיסכמיה של שריר הלב, המהווה את הגורם העיקרי לאי ספיקת לב כלילית. לפי התרשים, אפשר להגיד שככל ש- oldpeak גדול יותר, יש סיכוי יותר גדול שהמטופל יהיה אכן חולה לב וזה אכן תואם את המציאות. גם לפי ה-summary, לחולים אכן יש ממוצע גבוה יותר בצורה משמעותית.

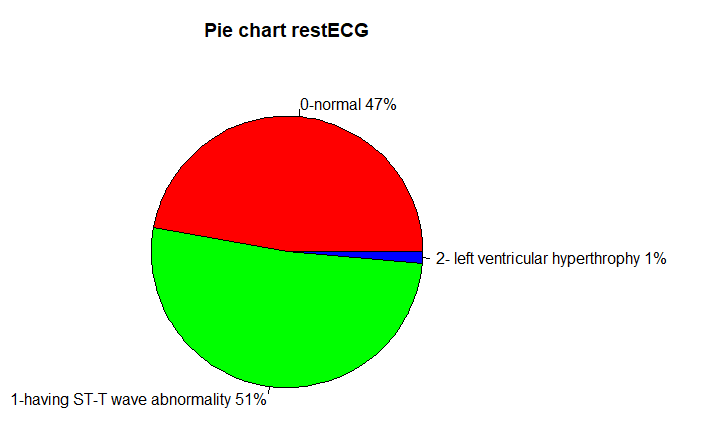
****

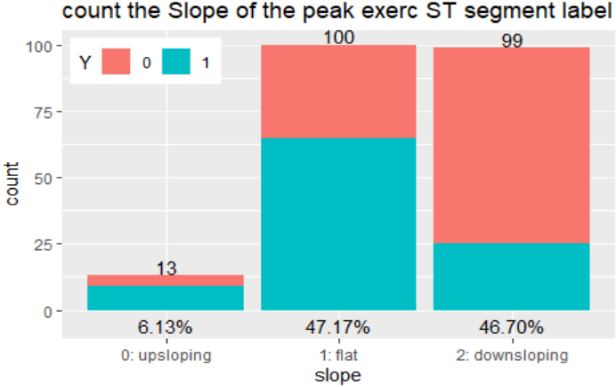
Summary 1 oldpeak

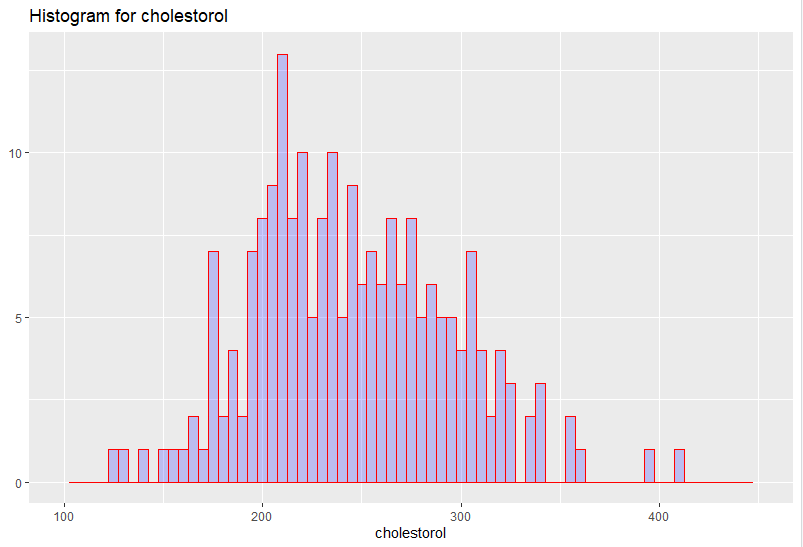
**Thelach** – דופק מקסימלי שהושג. ניתן לראות שהממוצע עבור חולים ובריאים הוא שונה – וכן גם החציון והרבעונים. זהו נתון מציאותי שכן מחקרים הראו שאימון מייעל את פעילות הלב, ומצאו קשר בין חוסר בפעילות גופנית (שריר פחות מאומן, דופק מקסימלי נמוך יותר )לבין מחלות לב שונות.

Summary 2 thelach

**CP-** סוג כאב בחזה. ניתן לראות שהנתונים לא מאוזנים ומרביתם מהקטגוריה של 0 – אנגינה טיפוסית. עם זאת מרבית החולים נופלים תחת קטגוריה זו. ניתן להניח שהמדגם כנראה הגיע מאוכלוסייה של חולים פוטנציאליים ועל כן בגלל זה ריבוי הכאב מסוג אנגינה טיפוסית.

**restECG**- ניתן לראות שהקטגוריות לא מאוזנות, מעל ממחצית המטופלים בעלי אבנורמאליות בבדיקה (קטגוריה 1)., ורק אחוז אחד מקטגוריה 2. זה מחזק את הטענה שהמדגם הגיע כנראה מחולים פוטנציאליים ולא מאוכלוסייה הטרוגנית.

**Slope**- ניתן לראות שהנתונים לא מאוזנים, מרבית המטופלים שייכים לקטגוריות 1 או 2. כמו כן, ניתן לראות בבירור שבקבוצות 0 ו-1 הסיכוי להיות חולה יותר גדול משמעותית מבקבוצה 2.

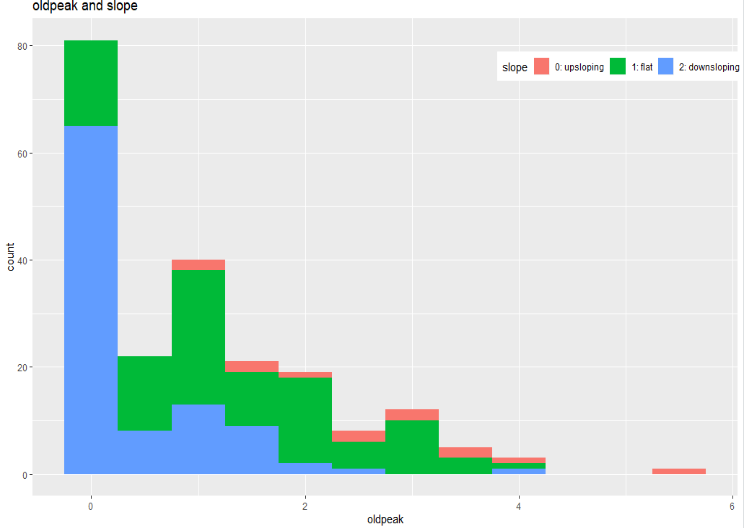
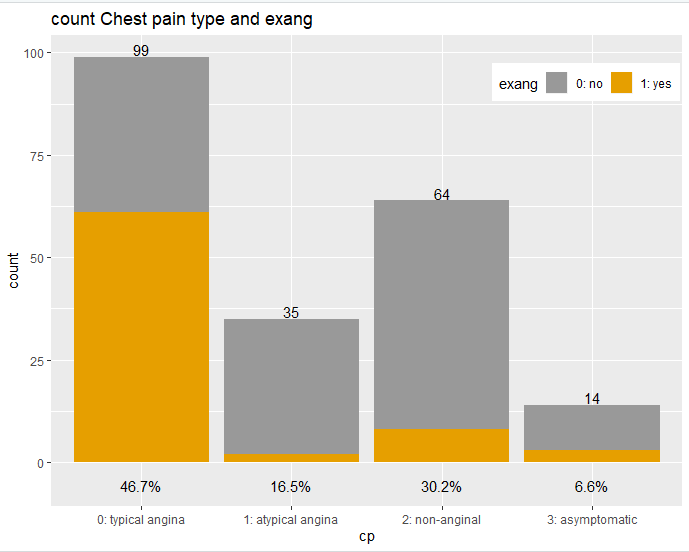
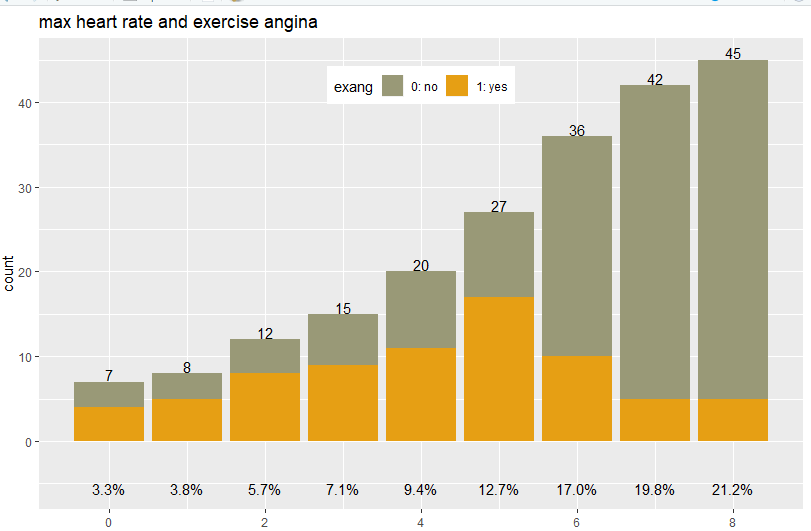
chol- סרום כולסטרול: מציג את הכולסטרול בסרום ב- mg / dl (יחידה).

מדידת ה- LDL (כולסטרול "רע"), HDL (כולסטרול "טוב") וטריגליצרידים נותן מספר שנקרא הכולסטרול הכולל בדם, או כולסטרול בסרום. רמות הכולסטרול בסרום יכולות לעזור להבין את הסיכון להתפתחות מחלות לב בעתיד, כך שרמת כולסטרול גבוהה בסרום מעלה את הסיכון למחלות לב. כולסטרול בריא בסרום מוגדר להיות פחות מ- 200 מ"ג לד"ל.

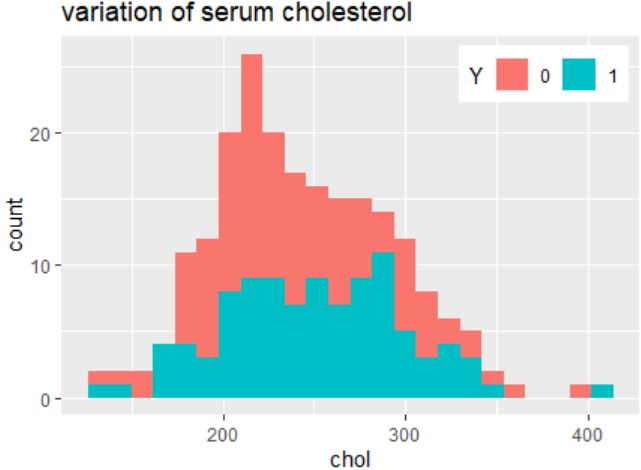
ניתן לראות, שיותר מ-75 אחוז מהמטופלים הם מעל 200, מה שנותן עוד אינדיקציה לכך שסט הנתונים לא הגיע מאוכלוסייה הטרוגנית.

## קשרים בין מאפיינים

על מנת להבין את הקשרים בין המשתנים, ניעזר תחילה בתרשים הנ"ל המראה קורלציה בין 13 המשתנים שלנו. קשרים חזקים יסומנו בעיגולים גדולים בצבעים חזקים ולהפך. מהתרשים ניתן לראות כי אין קורולציה חיובית או שלילית חזקה במיוחד בין אף זוג משתנים (מעל 0.6 בערך מוחלט) אך ניתן לראות כי ישנם מספר זוגות משתנים שכן נראים יותר חזקים בעוצמת הקשרים שלהם: oldpeak, slope, thalach, exang, cp

1. **Oldpeak and slope**: ניתן לראות שאכן מרבית האוכלוסייה עם Oldpeak עד 1 גם נמצאת בקטגוריה 2 של slope – שיפוע יורד. ראינו בסעיף קודם שהאוכלוסייה הזו (בקטגורייה downsloaping) מכילה הכי הרבה "בריאים", ביחס לשתי הקבוצות האחרות וכן שערך מתחת 0.5 ב-oldpeak משמעותו סיכוי קטן יותר לחלות ועל כן הקורלציה הזו הגיונית.
2. **Exang and cp** : ניתן לראות שבקבוצה שחווה אנגינה טיפוסית נמצאים מרבית האוכלוסייה שחווה אנגינה באימון גופני. פרט לכך בשאר קבוצות הCP לא ניתן לראות תאימות לאנגינה באימון גופני.
3. **Thalach and exang**: דופק מקסימלי חילקנו ל-8 קבוצות (על ציר X) לפי קפיצות של 10 פעימות בדקה. ניתן לראות שעבור דופק נמוך נמצאים רוב האוכלוסייה שחווה אנגינה באימון גופני (exang 1), ובקבוצות המגיעות לדופק גבוה (עמודות ימניות) נמצא האחוז הנמוך ביותר של חווי אנגינה באימון גופני. זהו הגיוני שכן אנשים שבכושר ומגיעים לדופק גבוה באימון עם סיכוי קטן יותר לחלות במחלות לב וביניהן אנגינה.

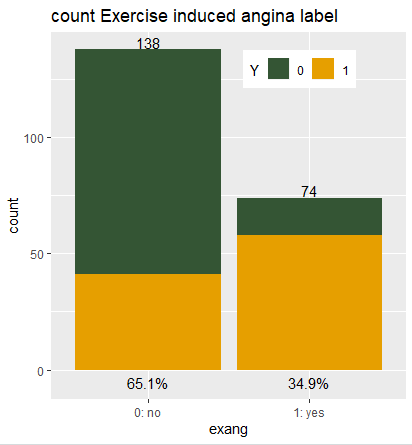
## מאפיינים בעלי השפעה על משתנה המטרה

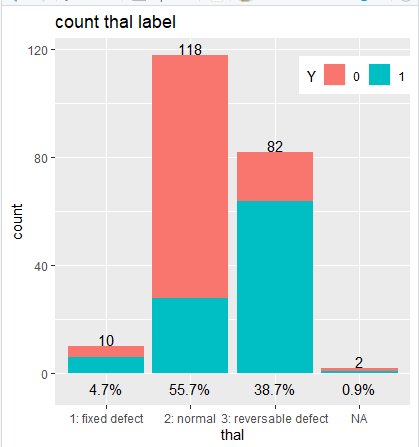
בסעיף הקודם סקרנו את המאפיינים והתפלגותם באוכלוסיית הנתונים (איזון וחוסר איזון), ואת חלקם סקרנו יחד עם ההשפעה על משתנה המטרה. זאת מכיוון שבתכונות רבות שאינן מאוזנות קשה להסביר את חוסר האיזון מבלי להבין את השפעת תכונה זו על משתנה המטרה.

ממחקרים ומאמרים שקראנו, ישנם כמה משתנים מסבירים חשובים בהשפעתם על משתנה המטרה:

**chol** - ככל שסרום הכולסטרול עולה, כך עולה הסיכוי לחלות במחלת לב[[4]](#endnote-4). בגרף התפלגות המאפיין יחד עם סיווג החולים והבריאים ניתן לראות שאין מגמה ברורה מאוד של חולים בעלי כולסטרול גבוה, בניגוד להשערתינו. בנוסף, המתאם בין תכונה זו למשתנה המוסבר היא נמוכה מאוד (0.1-).

**גיל** – ככל שעולה גילינו כך עולה הסיכוי שלנו לחלות במחלת לב. בסעיף קודם ניתן לראות בגרף שנראה שכמות חולי הלב עולה עם הגיל כמו השערתינו, אך לא בצורה מובהקת מאוד כמו שחשבנו. המתאם בין תכונה זו למשתנה המוסבר היא מעט גבוהה יותר מכולסטרול (2.1-).

**Exang**- תעוקת לב (אנגינה) באימון גופני ידועה כסממן למחלת לב. גם בגרף ניתן לראות שהרוב המוחלט של האוכלוסייה בסט הנתונים שחוותה אנגינה באימון גופני גם חלתה. בנוסף לכך, המתאם ביניהם גבוה יחסית – 0.47.

**מין** – בסעיף הקודם בגרף כתבנו שניתן לראות שישנו חלק גדול יותר מתוך אוכלוסיית הגברים החולים במחלת לב (Y=1). ההסתברות עבור גברים לחלות במחלת לב גבוהה יותר במציאות, כמעט פי 2 מנשים (איגוד הלב האמריקאי, 2015). כמו כן, המתאם בין המשתנים האלו הוא גבוה יחסית – 0.33.

**Thal** – בדיקת לחץ תאליום  היא בדיקת הדמיה גרעינית שמראה עד כמה הדם זורם ללב בזמן אימון או במנוחה. מבחינה מחקרית, ככל שזורם פחות דם כך עולה הסיכוי למחלת לב מכיוון שתפקוד הלב אינו תקין. ניתן לראות בגרף שעבור הקבוצה עם פגם הפיך מרביתם גם חולים. כמו כן בקבוצה עם פגם קבוע חלק גדול מהקבוצה גם כן חולה. עם זאת עבור מי שאין לו פגם, הסיכוי נמוך בהרבה לחלות במחלת לב. בנוסף, המתאם בין המשתנים הכי גבוה מכל שאר המאפיינים.

## איכות הנתונים

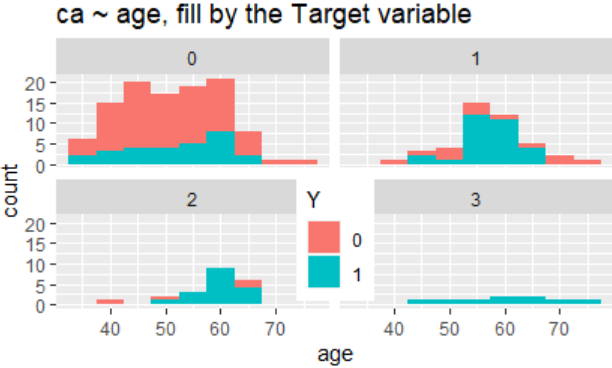
נתונים חסרים:

לפי ההסבר על המשתנים בבסיס הנתונים, ניתן לראות שיש שני מאפיינים שנכללים בתוכם את האפשרות לרשום ערכים חסרים והם:

1. **ca**: יכול לקבל את הערך 4 שמוגדר כלא ידוע, אנחנו בעבודה שלנו נתייחס לערכים האלה כערכים חסרים.
2. **thal**: 0 מוגדר כנתון חסר.

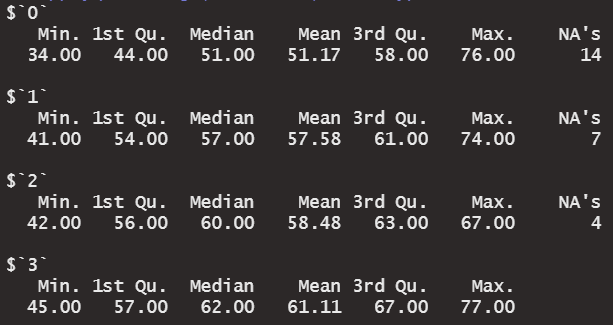
בנוסף לכך, אם נסתכל על סט הנתונים, נשים לב שבמאפיין גיל קיימים גילים לא הגיוניים, לדוגמה-1490(אנחנו נמחק את הגיל אם הוא מעל 100, ונתייחס אליו כנתון חסר). לפי מחקרים והידע האישי שלנו, הגיל הוא מאפיין חשוב, ועוזר מאוד באבחון אם המטופל חולה במחלת לב או לא, ולכן אנחנו ננסה להתמודד עם נתונים אלה.

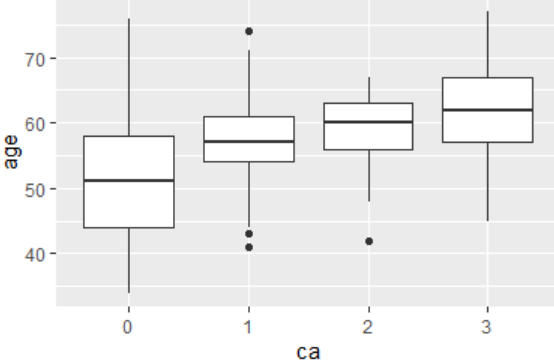
כמו שכתבנו בסעיף קודם, מחקרים הראו שמחלת לב בעלת קשר ישיר לגילו של מטופל, והמאפיין הזה יכול לתרום למערכת להיות מדויקת יותר. ולכן אנו ננסה להשאיר אותם, לטפל בהם ולא למחוק אותם.

**** איך נטפל בהם?

שמנו לב שהגיל של המטופל משתנה בכל קבוצה של המאפיין ca: כלומר, ככל שיש למטופל יותר עורקים חסומים, הוא לרוב מבוגר יותר.

ולכן, אנחנו נתקן את הערכים החריגים בגיל לפי המאפיין ca: נחליף את הנתון החסר בחציון(יותר אמין מהממוצע) ונעגל למטה.

****

****

# הכנת הנתונים

## בחירת מאפיינים

* השמטת המאפיין id - מספר הזיהוי של המטופל: לכל מטופל יש מספר אחר, ולכן נשמיט את המשתנה הזה. אולי אם היה לנו ידע על דרך ייצור המספר, כלומר אם הייתה תלות בזמן, קרובי משפחה וכו', אולי כן היינו משאירים אותו.
* השמטת המאפיין FBS: כאמור, רמת הסוכר בדם בצום מתחלקת רק לשתי קבוצות. כמו כן ראינו שנתון זה ממש לא מאוזן ו86% מהתצפיות בקבוצה של מתחת ל-120. עם זאת, זה נתון מציאותי שכן FBS תקין הוא עד 100, וערכים שבין 100 ל-125 מוגדרים פרה-סוכרת ומעל 126 זוהי כבר סוכרת. הבעיה היא שרק שתי קטגוריות יוצרת מצב בו הגיוני שמרבית הערכים הם עד 120, אך נראה שבערך מחצית מכל קבוצה הם חולי לב - מה שמעיד שFBS לא מוסיף לנו מידע מיוחד. על כן נשמיט אותו.
* נוסף לשני אלו נעדיף לא להשמיט משתנים נוספים. אנחנו לא מומחים בתחום, ולכן נרצה להימנע מהשמטת משתנה שתורם לדיוק בתוצאות הסופיות. כמו כן בסט הנתונים הראשון-האמיתי, היו יותר מ-50 מאפיינים, אך החברה השאירה רק 13 מאפיינים(+id), ולכן ההשערה שלנו כרגע היא שהמאפיינים שיש לנו הם חשובים, וייתכן שעבור מדגם גדול יותר והטרוגני יותר נקבל תוצאות שונות. עבור משתנה הכולסטרול עבורו קיבלנו מתאם נמוך ונדמה לנו שאין הוא תורם רבות למשתנה המוסבר אנו יודעים מידע קודם שזהו מאפיין הכרחי לניבוי מחלת לב ועל כן גם אותו נבחר להשאיר.
* בשני המשתנים הראשוניים שיש בהם ערכים חסרים (thal,ca), נשמיט את התצפיות, אלא אם כן נמצא שיטה טובה למלא אותן. הם בודדים ולכן אפשר למחוק אותם או להשאיר אותם כיוון שיש אלגוריתמים שיודעים עם ערכים חסרים.

## טיפול פרטני במאפיינים

* **trestBPS**

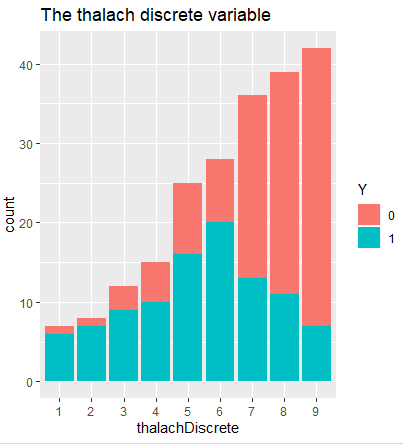
רופאים בדרך כלל מסתכלים באיזה טווח נמצא לחץ הדם[[5]](#endnote-5), כלומר, אם לחץ הדם הינו 120 או 125 מבחינתם זה כמעט זהה. לכן, בעזרת טווחים אלו נהפוך את המאפיין הזה למשתנה קטגוריאלי:

1:(trestbps<120),

2:(120<=trestbps<129),

3:(130<=trestbps<139),

4:(trestbps>=140)

* בשלב זה נעדיף לא להפוך את המאפיין גיל למשתנה קטגוריאלי, כיוון שאנו יודעים שהוא בעל חשיבות רבה במציאות על הסיכוי לחלות במחלת לב. על כן נעדיף כרגע לא לגעת בו. בנוסף לכך, הפיכת משתנה רציף לקטגוריאלי אולי מפשטת יותר, אבל יוצרת איבוד מידע שאולי יגדיל את הטעות מסוג ראשון בהמשך.
* **Thalach** – נבצע דיסקרטיזציה למשתנה הדופק המקסימלי שהושג לפי קפיצות של 10 פעימות לב בדקה, כאשר הקבוצה הראשונה תהיה עד 100, אחריה 110 וכך הלאה.

1. Goldberg, R. J., Gore, J. M., Gurwitz, J. H., Alpert, J. S., Brady, P., Strohsnitter, W., ... & Dalen, J. E. (1989). The impact of age on the incidence and prognosis of initial acute myocardial infarction: the Worcester Heart Attack Study. *American heart journal*, *117*(3), 543-549.‏ [↑](#endnote-ref-1)
2. אנגינה היא תעוקת לב [↑](#endnote-ref-2)
3. <https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_449846.pdf> [↑](#endnote-ref-3)
4. <https://www.tasmc.org.il/Laboratory-tests/Pages/Serum-Cholesterol.aspx> [↑](#endnote-ref-4)
5. <https://www.tasmc.org.il/Be-Well/InterestAreas/pressure/Pages/HBP.aspx>

   # תיעוד קוד

   ################################

   ##### ML Project - Part A #####

   ################################

   library(dplyr)

   ## Please put all the datasets in the same folder(Set As Working Directory)

   #Uploading the dataset:

   Heart <- read.csv('Xy\_train.csv')

   ##################################

   ##### Working with NA Values #####

   ##################################

   #make missing values as NA:

   Heart$age <- ifelse(Heart$age > 100,NA,Heart$age) #age column:

   Heart$ca <- ifelse(Heart$ca == 4,NA,Heart$ca) #ca column:

   Heart$thal <- ifelse(Heart$thal == 0,NA,Heart$thal) #thal column:

   #Check NA values:

   library(Amelia)

   missmap(Heart, col = c('yellow','black'),y.at = c(1),y.labels = c(''))

   #Duplicated rows

   which(duplicated(Heart) | duplicated(Heart[nrow(Heart):1, ])[nrow(Heart):1])

   #Missing data in specific row:

   row.missing <- function(Column){

   MissingValues <- c()

   for(i in 1:length(Column)){

   if(is.na(Column[i])){

   MissingValues <- c(MissingValues,i)

   }

   }

   print(MissingValues)

   }

   row.missing(Heart$thal)

   row.missing(Heart$age)

   row.missing(Heart$ca)

   ################################################

   ##### Preparing the Catergorical variables #####

   ################################################

   #Change the categorical variables from int to factor (look at: str(Heart))

   #Option 1: We won't use it !

   #cols <- c("gender","cp","fbs","restecg","exang","slope","ca","thal")

   #Heart[cols] <- lapply(Heart[cols], factor)

   #Option 2: Use it :)

   ### Change factors names in columns:

   Heart <- Heart %>%

   mutate(

   gender = case\_when(

   gender == 0 ~ "0: female",

   gender == 1 ~ "1: male"

   ),

   fbs = case\_when(

   fbs == 0 ~ "0: fbs<=120",

   fbs == 1 ~ "1: fbs>120"

   ),

   exang = case\_when(

   exang == 0 ~ "0: no",

   exang == 1 ~ "1: yes"

   ),

   cp = case\_when(

   cp == 0 ~ "0: typical angina",

   cp == 1 ~ "1: atypical angina",

   cp == 2 ~ "2: non-anginal",

   cp == 3 ~ "3: asymptomatic"

   ),

   restecg = case\_when(

   restecg == 0 ~ "0: normal",

   restecg == 1 ~ "1: ST-T wave abnormality",

   restecg == 2 ~ "2: hypertrophy"

   ),

   slope = case\_when(

   slope == 0 ~ "0: upsloping",

   slope == 1 ~ "1: flat",

   slope == 2 ~ "2: downsloping"

   ),

   thal = case\_when(

   thal == 1 ~ "1: fixed defect",

   thal == 2 ~ "2: normal",

   thal == 3 ~ "3: reversable defect"

   ),

   gender = as.factor(gender),

   fbs = as.factor(fbs),

   exang = as.factor(exang),

   cp = as.factor(cp),

   slope = as.factor(slope),

   ca = as.factor(ca),

   thal = as.factor(thal),

   restecg=as.factor(restecg)

   )

   # The Target Variable:

   Heart <- Heart %>% rename(Y = y) #change target name(To make it easy:

   #y is y axis,Y is the Target variable)

   Heart$Y <- ifelse(Heart$Y == 1,0,1) #CHANGE Y

   Heart$Y <- as.factor(Heart$Y) #change Y to factor

   Yprior<-length(which(Heart$Y==1))/length(Heart$Y) #Prior Probability

   min(Heart$age)

   max(Heart$age)

   ########################################

   ########## Data Visualization ##########

   ########################################

   library(ggplot2)

   ###############################

   ##### the Target variable #####

   ###############################

   # Y label

   ggplot(Heart,aes(Y)) + geom\_bar(fill = '#3399FF') + #3399FF #9933FF

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('The target variable') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5))

   #################################

   ##### Categorical variables #####

   #################################

   #Gender

   ggplot(Heart,aes(gender)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count Gender label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.15, 0.9),legend.direction = "horizontal")

   # cp

   ggplot(Heart,aes(cp)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count Chest pain type label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.847, 0.9),legend.direction = "horizontal")

   # fbs

   ggplot(Heart,aes(fbs)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.05)) + #add counting

   ggtitle('count Fasting blood sugar label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.847, 0.9),legend.direction = "horizontal")

   # restecg

   slices<-c(sum(Heart$restecg==0),sum(Heart$restecg==1),sum(Heart$restecg==2))

   lbls=c("0-normal","1-having ST-T wave abnormality","2- left ventricular hyperthrophy")

   pct <- round(slices/sum(slices)\*100)

   lbls <- paste(lbls, pct) # add percents to labels

   lbls <- paste(lbls,"%",sep="") # ad % to labels

   pie(slices,labels = lbls, col=rainbow(length(lbls)),

   main="Pie chart restECG")

   ggplot(Heart,aes(restecg)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count Resting electrocardiographic label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.847, 0.9),legend.direction = "horizontal")

   # exang

   ggplot(Heart,aes(exang)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count Exercise induced angina label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.7, 0.9),legend.direction = "horizontal")+

   scale\_fill\_manual(values=c("#345534", "#E69F00"))

   # slope

   ggplot(Heart,aes(slope)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count the Slope of the peak exerc ST segment label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.15, 0.9),legend.direction = "horizontal")

   # ca

   ggplot(Heart,aes(ca)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count number of major Vessels label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.847, 0.9),legend.direction = "horizontal")

   # thal

   ggplot(Heart,aes(thal)) + geom\_bar(aes(fill = Y)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count thal label') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.847, 0.9),legend.direction = "horizontal")

   ################################

   ##### Contiunous variables #####

   ################################

   # age

   ggplot(Heart,aes(x=age,fill = Y)) + geom\_bar() + #Used

   ggtitle('variation of Age') +

   theme(legend.position = c(0.835, 0.9),legend.direction = "horizontal")

   tapply(Heart$age, Heart$Y, summary) #Used.

   # trestbps

   ggplot(Heart,aes(x=trestbps,fill=Y)) + geom\_histogram(binwidth = 6) +

   ggtitle('variation of resting blood pressure') +

   theme(legend.position = c(0.835, 0.9),legend.direction = "horizontal")

   tapply(Heart$trestbps, Heart$Y, summary)

   # chol

   qplot(Heart$chol,

   geom="histogram",

   binwidth = 5,

   main = "Histogram for cholestorol",

   xlab = "cholestorol",

   fill=I("blue"),

   col=I("red"),

   alpha=I(.2),

   xlim=c(100,450))

   ggplot(Heart,aes(x=chol,fill=Y)) + geom\_histogram(binwidth = 5)+ #Used.

   ggtitle('variation of serum cholesterol') +

   theme(legend.position = c(0.835, 0.9),legend.direction = "horizontal")

   tapply(Heart$chol, Heart$Y, summary)

   # thalach

   ggplot(Heart,aes(x=thalach,fill=Y)) + geom\_histogram(binwidth = 3) + #Used.

   ggtitle('variation of Maximum heart rate achieved') +

   theme(legend.position = c(0.165, 0.9),legend.direction = "horizontal")

   tapply(Heart$thalach, Heart$Y, summary)

   # oldpeak

   ggplot(Heart,aes(x=oldpeak,fill=Y)) + geom\_histogram(binwidth = 0.5) + #Used.

   ggtitle('variation of ST depression include rest') +

   theme(legend.position = c(0.835, 0.9),legend.direction = "horizontal")

   tapply(Heart$oldpeak, Heart$Y, summary)

   ####################################################

   #####couples of attributes with high correlation###

   #oldpeak and slope

   ggplot(Heart,aes(x=oldpeak,fill=slope)) + geom\_histogram(binwidth = 0.5) + #Used.

   ggtitle('oldpeak and slope') +

   theme(legend.position = c(0.835, 0.9),legend.direction = "horizontal")

   tapply(Heart$oldpeak, Heart$slope, summary)

   # cp and exang

   ggplot(Heart,aes(x=cp)) + geom\_bar(aes(fill = exang)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('count Chest pain type and exang') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.847, 0.9),legend.direction = "horizontal")+

   scale\_fill\_manual(values=c("#999999", "#E69F00"))

   #thalach groups and exang

   ggplot(Heart,aes(x=findInterval(Heart$thalach, c(100, 110, 120,130,140,150,160,170)))) + geom\_bar(aes(fill = exang)) + #Used.

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) + #add counting

   ggtitle('max heart rate and exercise angina') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))), #add percentage

   stat='count',position=position\_fill(vjust=-5.5)) +

   theme(legend.position = c(0.5, 0.9),legend.direction = "horizontal")+

   scale\_fill\_manual(values=c("#999977", "#E69F14"))

   ####################################################

   ##### Comparing the Target variable in training and test set #####

   Ytest <- read.csv('preds\_sample.csv')

   Ytest$y <- as.factor(Ytest$y)

   Y\_Train <- ggplot(Heart,aes(Y)) + geom\_bar(fill = '#3399FF') +

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) +

   ggtitle('Target variable from training set') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))),

   stat='count',position=position\_fill(vjust=-5.5))

   y\_Test <- ggplot(Ytest,aes(y)) + geom\_bar(fill = '#9933FF') +

   geom\_text(stat = 'count',aes(label =..count.., vjust = -0.1)) +

   ggtitle('Target variable from test set') +

   geom\_text(aes(label=scales::percent(..count../sum(..count..))),

   stat='count',position=position\_fill(vjust=-3))

   library("gridExtra")

   grid.arrange(Y\_Train,y\_Test,nrow = 1)

   ####################################################

   #######################

   ##### Correlation #####

   #######################

   library(corrplot)

   library(polycor)

   library(GGally)

   library(corrr)

   Heart2<-read.csv('Xy\_train.csv')

   M <- cor(na.omit(select(Heart2,c(age,trestbps,chol,thalach,oldpeak,cp,restecg,gender,fbs,exang,slope,ca,thal))))

   corrplot.mixed(M, lower.col = "black",

   number.cex = 0,order="hclust",tl.cex = 1)

   #or:

   ggpairs(select(Heart,c(age,trestbps,chol,thalach,oldpeak,cp,restecg,gender,fbs,exang,slope,ca,thal)))

   ## Big Plots ##

   hetcor(Heart)

   ggpairs(Heart)

   ####correlations with Y

   MM<-cor(Heart2,Heart2$Y,method="kendall")

   corela<-MM[,15]

   #####################################

   ##### Working with missing data #####

   #####################################

   #Dropping id column:

   library(dplyr)

   Heart <- select(Heart,-id)

   #Dropping rows with NA columns:

   Heart <- Heart %>% filter(!is.na(ca))

   Heart <- Heart %>% filter(!is.na(thal))

   # Fill NA values in age column:

   ggplot(Heart,aes(ca,age)) + geom\_boxplot()

   ggplot(Heart,aes(x=age,fill=Y)) + geom\_histogram(binwidth = 5) +

   facet\_wrap(~ ca) + ggtitle('ca ~ age, fill by the Target variable') +

   theme(legend.position = c(0.5, 0.3))

   tapply(Heart$age, Heart$ca, summary)

   ageMed <- as.data.frame(Heart %>% group\_by(ca) %>% summarise(MED = median(age,na.rm = T)))

   for(i in 1:length(Heart$age)){

   if(is.na(Heart[i,1])){#just for missing values

   Heart[i,1] <- ageMed[as.integer(Heart[i,12])+1,2] #change missing value by age-ca median.

   }

   }

   ##########################

   ##### Discretization #####

   ##########################

   # BloodPressureRange

   BloodPressure <- function(tres){

   if(tres<120){

   tres <- 1

   }

   else if(tres<129){

   tres <- 2

   }

   else if(tres<139){

   tres <- 3

   }

   else{

   tres <- 4

   }

   }

   Heart$BloodPressureRange <- as.factor(sapply(Heart$trestbps,BloodPressure))

   ggplot(Heart,aes(BloodPressureRange,fill = Y)) + geom\_bar() +

   ggtitle('The BloodPressureType variable')+

   scale\_fill\_manual(values=c("#999900", "#E99954"))

   # Thalach discrete

   maxh <- function(tres){

   if(tres<=100){

   tres <- 1

   }

   else if(tres<=110){

   tres <- 2

   }

   else if(tres<=120){

   tres <- 3

   }

   else if(tres<130){

   tres <- 4

   }

   else if(tres<=140){

   tres <- 5

   }

   else if(tres<=150){

   tres <- 6

   }

   else if(tres<=160){

   tres <- 7

   }

   else if(tres<=170){

   tres <- 8

   }

   else{

   tres <- 9

   }

   }

   Heart$thalachDiscrete <- as.factor(sapply(Heart$thalach,maxh))

   ggplot(Heart,aes(thalachDiscrete,fill = Y)) + geom\_bar() +

   ggtitle('The thalach discrete variable') [↑](#endnote-ref-5)