Final Project: *Yashita Sharma*

# Analysis of Heart Disease Dataset

**Data Source:** UC Irvine Machine Learning Repository.

### **Introduction**

All published studies refer to employing a subset of 14 of the 76 attributes in this database. Specifically, to date, no other database has been utilized by machine learning researchers than the Cleveland one. The patient’s presence of heart disease is indicated in the “goal” field. From 0 (no presence) to 4, it has integer values. The focus of Cleveland database experiments has been on trying to differentiate presence (values 1, 2, 3, 4) from absence (value 0).  
Recently, the patients’ names and social security numbers were deleted from the database and replaced with dummy information.

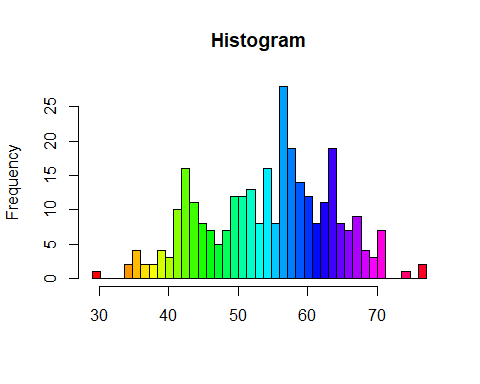
Attaching package: 'dplyr'

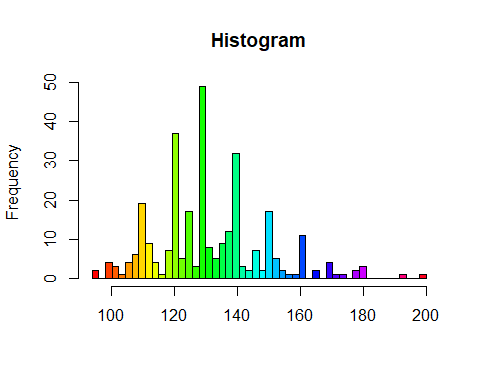
The following objects are masked from 'package:stats':  
  
 filter, lag

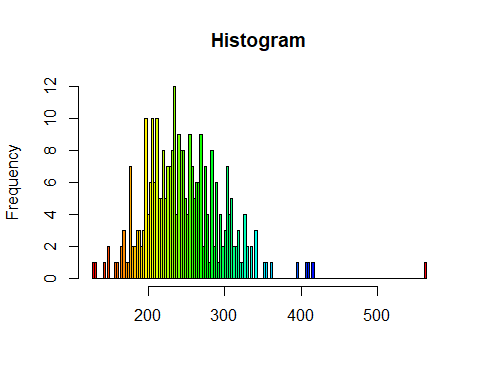
The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

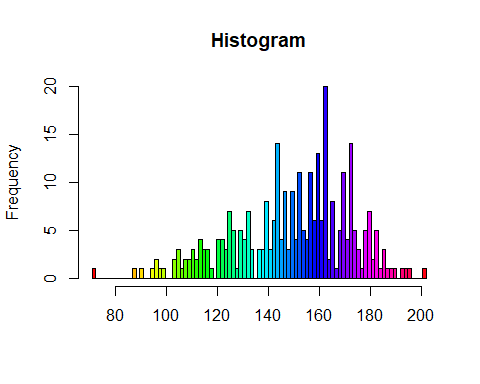
'data.frame': 303 obs. of 14 variables:  
 $ age : num 63 67 67 37 41 56 62 57 63 53 ...  
 $ sex : num 1 1 1 1 0 1 0 0 1 1 ...  
 $ cp : num 1 4 4 3 2 2 4 4 4 4 ...  
 $ trestbps: num 145 160 120 130 130 120 140 120 130 140 ...  
 $ chol : num 233 286 229 250 204 236 268 354 254 203 ...  
 $ fbs : num 1 0 0 0 0 0 0 0 0 1 ...  
 $ restecg : num 2 2 2 0 2 0 2 0 2 2 ...  
 $ thalach : num 150 108 129 187 172 178 160 163 147 155 ...  
 $ exang : num 0 1 1 0 0 0 0 1 0 1 ...  
 $ oldpeak : num 2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...  
 $ slope : num 3 2 2 3 1 1 3 1 2 3 ...  
 $ ca : chr "0.0" "3.0" "2.0" "0.0" ...  
 $ thal : chr "6.0" "3.0" "7.0" "3.0" ...  
 $ target : int 0 2 1 0 0 0 3 0 2 1 ...

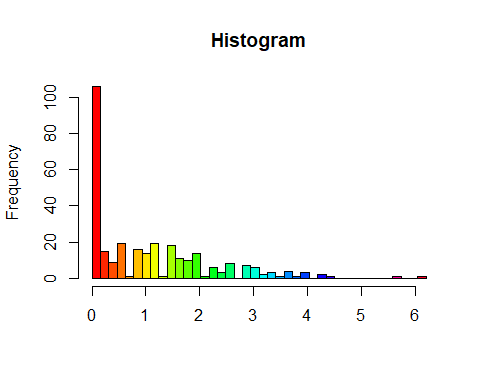
age trestbps chol thalach oldpeak   
 Min. :29.00 Min. : 94.0 Min. :126.0 Min. : 71.0 Min. :0.00   
 1st Qu.:48.00 1st Qu.:120.0 1st Qu.:211.0 1st Qu.:133.5 1st Qu.:0.00   
 Median :56.00 Median :130.0 Median :241.0 Median :153.0 Median :0.80   
 Mean :54.44 Mean :131.7 Mean :246.7 Mean :149.6 Mean :1.04   
 3rd Qu.:61.00 3rd Qu.:140.0 3rd Qu.:275.0 3rd Qu.:166.0 3rd Qu.:1.60   
 Max. :77.00 Max. :200.0 Max. :564.0 Max. :202.0 Max. :6.20











$age  
$breaks  
 [1] 29.00000 30.17073 31.34146 32.51220 33.68293 34.85366 36.02439 37.19512  
 [9] 38.36585 39.53659 40.70732 41.87805 43.04878 44.21951 45.39024 46.56098  
[17] 47.73171 48.90244 50.07317 51.24390 52.41463 53.58537 54.75610 55.92683  
[25] 57.09756 58.26829 59.43902 60.60976 61.78049 62.95122 64.12195 65.29268  
[33] 66.46341 67.63415 68.80488 69.97561 71.14634 72.31707 73.48780 74.65854  
[41] 75.82927 77.00000  
  
$counts  
 [1] 1 0 0 0 2 4 2 2 4 3 10 16 11 8 7 5 7 12 12 13 8 16 8 28 19  
[26] 14 12 8 11 19 8 7 9 4 3 7 0 0 1 0 2  
  
$density  
 [1] 0.002819032 0.000000000 0.000000000 0.000000000 0.005638064 0.011276128  
 [7] 0.005638064 0.005638064 0.011276128 0.008457096 0.028190319 0.045104510  
[13] 0.031009351 0.022552255 0.019733223 0.014095160 0.019733223 0.033828383  
[19] 0.033828383 0.036647415 0.022552255 0.045104510 0.022552255 0.078932893  
[25] 0.053561606 0.039466447 0.033828383 0.022552255 0.031009351 0.053561606  
[31] 0.022552255 0.019733223 0.025371287 0.011276128 0.008457096 0.019733223  
[37] 0.000000000 0.000000000 0.002819032 0.000000000 0.005638064  
  
$mids  
 [1] 29.58537 30.75610 31.92683 33.09756 34.26829 35.43902 36.60976 37.78049  
 [9] 38.95122 40.12195 41.29268 42.46341 43.63415 44.80488 45.97561 47.14634  
[17] 48.31707 49.48780 50.65854 51.82927 53.00000 54.17073 55.34146 56.51220  
[25] 57.68293 58.85366 60.02439 61.19512 62.36585 63.53659 64.70732 65.87805  
[33] 67.04878 68.21951 69.39024 70.56098 71.73171 72.90244 74.07317 75.24390  
[41] 76.41463  
  
$xname  
[1] "x"  
  
$equidist  
[1] TRUE  
  
attr(,"class")  
[1] "histogram"  
  
$trestbps  
$breaks  
 [1] 94.00 96.12 98.24 100.36 102.48 104.60 106.72 108.84 110.96 113.08  
[11] 115.20 117.32 119.44 121.56 123.68 125.80 127.92 130.04 132.16 134.28  
[21] 136.40 138.52 140.64 142.76 144.88 147.00 149.12 151.24 153.36 155.48  
[31] 157.60 159.72 161.84 163.96 166.08 168.20 170.32 172.44 174.56 176.68  
[41] 178.80 180.92 183.04 185.16 187.28 189.40 191.52 193.64 195.76 197.88  
[51] 200.00  
  
$counts  
 [1] 2 0 4 3 1 4 6 19 9 4 1 7 37 5 17 3 49 8 5 9 12 32 3 2 7  
[26] 2 17 5 2 1 1 11 0 2 0 4 1 1 0 2 3 0 0 0 0 0 1 0 0 1  
  
$density  
 [1] 0.003113519 0.000000000 0.006227038 0.004670278 0.001556759 0.006227038  
 [7] 0.009340557 0.029578430 0.014010835 0.006227038 0.001556759 0.010897316  
[13] 0.057600100 0.007783797 0.026464911 0.004670278 0.076281213 0.012454076  
[19] 0.007783797 0.014010835 0.018681113 0.049816302 0.004670278 0.003113519  
[25] 0.010897316 0.003113519 0.026464911 0.007783797 0.003113519 0.001556759  
[31] 0.001556759 0.017124354 0.000000000 0.003113519 0.000000000 0.006227038  
[37] 0.001556759 0.001556759 0.000000000 0.003113519 0.004670278 0.000000000  
[43] 0.000000000 0.000000000 0.000000000 0.000000000 0.001556759 0.000000000  
[49] 0.000000000 0.001556759  
  
$mids  
 [1] 95.06 97.18 99.30 101.42 103.54 105.66 107.78 109.90 112.02 114.14  
[11] 116.26 118.38 120.50 122.62 124.74 126.86 128.98 131.10 133.22 135.34  
[21] 137.46 139.58 141.70 143.82 145.94 148.06 150.18 152.30 154.42 156.54  
[31] 158.66 160.78 162.90 165.02 167.14 169.26 171.38 173.50 175.62 177.74  
[41] 179.86 181.98 184.10 186.22 188.34 190.46 192.58 194.70 196.82 198.94  
  
$xname  
[1] "x"  
  
$equidist  
[1] TRUE  
  
attr(,"class")  
[1] "histogram"  
  
$chol  
$breaks  
 [1] 126.0000 128.8816 131.7632 134.6447 137.5263 140.4079 143.2895 146.1711  
 [9] 149.0526 151.9342 154.8158 157.6974 160.5789 163.4605 166.3421 169.2237  
 [17] 172.1053 174.9868 177.8684 180.7500 183.6316 186.5132 189.3947 192.2763  
 [25] 195.1579 198.0395 200.9211 203.8026 206.6842 209.5658 212.4474 215.3289  
 [33] 218.2105 221.0921 223.9737 226.8553 229.7368 232.6184 235.5000 238.3816  
 [41] 241.2632 244.1447 247.0263 249.9079 252.7895 255.6711 258.5526 261.4342  
 [49] 264.3158 267.1974 270.0789 272.9605 275.8421 278.7237 281.6053 284.4868  
 [57] 287.3684 290.2500 293.1316 296.0132 298.8947 301.7763 304.6579 307.5395  
 [65] 310.4211 313.3026 316.1842 319.0658 321.9474 324.8289 327.7105 330.5921  
 [73] 333.4737 336.3553 339.2368 342.1184 345.0000 347.8816 350.7632 353.6447  
 [81] 356.5263 359.4079 362.2895 365.1711 368.0526 370.9342 373.8158 376.6974  
 [89] 379.5789 382.4605 385.3421 388.2237 391.1053 393.9868 396.8684 399.7500  
 [97] 402.6316 405.5132 408.3947 411.2763 414.1579 417.0395 419.9211 422.8026  
[105] 425.6842 428.5658 431.4474 434.3289 437.2105 440.0921 442.9737 445.8553  
[113] 448.7368 451.6184 454.5000 457.3816 460.2632 463.1447 466.0263 468.9079  
[121] 471.7895 474.6711 477.5526 480.4342 483.3158 486.1974 489.0789 491.9605  
[129] 494.8421 497.7237 500.6053 503.4868 506.3684 509.2500 512.1316 515.0132  
[137] 517.8947 520.7763 523.6579 526.5395 529.4211 532.3026 535.1842 538.0658  
[145] 540.9474 543.8289 546.7105 549.5921 552.4737 555.3553 558.2368 561.1184  
[153] 564.0000  
  
$counts  
 [1] 1 1 0 0 0 1 0 2 0 0 1 1 0 2 3 1 1 7 2 2 3 3 2 3 10  
 [26] 4 6 10 6 10 5 5 8 5 7 7 8 12 4 9 8 8 5 4 9 7 5 6 6 9  
 [51] 2 7 4 1 8 2 6 1 4 2 3 7 4 5 2 2 3 1 1 4 2 0 2 0 3  
 [76] 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0  
[101] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
[126] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
[151] 0 1  
  
$density  
 [1] 0.00114532 0.00114532 0.00000000 0.00000000 0.00000000 0.00114532  
 [7] 0.00000000 0.00229064 0.00000000 0.00000000 0.00114532 0.00114532  
 [13] 0.00000000 0.00229064 0.00343596 0.00114532 0.00114532 0.00801724  
 [19] 0.00229064 0.00229064 0.00343596 0.00343596 0.00229064 0.00343596  
 [25] 0.01145320 0.00458128 0.00687192 0.01145320 0.00687192 0.01145320  
 [31] 0.00572660 0.00572660 0.00916256 0.00572660 0.00801724 0.00801724  
 [37] 0.00916256 0.01374384 0.00458128 0.01030788 0.00916256 0.00916256  
 [43] 0.00572660 0.00458128 0.01030788 0.00801724 0.00572660 0.00687192  
 [49] 0.00687192 0.01030788 0.00229064 0.00801724 0.00458128 0.00114532  
 [55] 0.00916256 0.00229064 0.00687192 0.00114532 0.00458128 0.00229064  
 [61] 0.00343596 0.00801724 0.00458128 0.00572660 0.00229064 0.00229064  
 [67] 0.00343596 0.00114532 0.00114532 0.00458128 0.00229064 0.00000000  
 [73] 0.00229064 0.00000000 0.00343596 0.00000000 0.00000000 0.00000000  
 [79] 0.00114532 0.00114532 0.00000000 0.00114532 0.00000000 0.00000000  
 [85] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
 [91] 0.00000000 0.00000000 0.00000000 0.00114532 0.00000000 0.00000000  
 [97] 0.00000000 0.00114532 0.00114532 0.00000000 0.00114532 0.00000000  
[103] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[109] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[115] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[121] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[127] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[133] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[139] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[145] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[151] 0.00000000 0.00114532  
  
$mids  
 [1] 127.4408 130.3224 133.2039 136.0855 138.9671 141.8487 144.7303 147.6118  
 [9] 150.4934 153.3750 156.2566 159.1382 162.0197 164.9013 167.7829 170.6645  
 [17] 173.5461 176.4276 179.3092 182.1908 185.0724 187.9539 190.8355 193.7171  
 [25] 196.5987 199.4803 202.3618 205.2434 208.1250 211.0066 213.8882 216.7697  
 [33] 219.6513 222.5329 225.4145 228.2961 231.1776 234.0592 236.9408 239.8224  
 [41] 242.7039 245.5855 248.4671 251.3487 254.2303 257.1118 259.9934 262.8750  
 [49] 265.7566 268.6382 271.5197 274.4013 277.2829 280.1645 283.0461 285.9276  
 [57] 288.8092 291.6908 294.5724 297.4539 300.3355 303.2171 306.0987 308.9803  
 [65] 311.8618 314.7434 317.6250 320.5066 323.3882 326.2697 329.1513 332.0329  
 [73] 334.9145 337.7961 340.6776 343.5592 346.4408 349.3224 352.2039 355.0855  
 [81] 357.9671 360.8487 363.7303 366.6118 369.4934 372.3750 375.2566 378.1382  
 [89] 381.0197 383.9013 386.7829 389.6645 392.5461 395.4276 398.3092 401.1908  
 [97] 404.0724 406.9539 409.8355 412.7171 415.5987 418.4803 421.3618 424.2434  
[105] 427.1250 430.0066 432.8882 435.7697 438.6513 441.5329 444.4145 447.2961  
[113] 450.1776 453.0592 455.9408 458.8224 461.7039 464.5855 467.4671 470.3487  
[121] 473.2303 476.1118 478.9934 481.8750 484.7566 487.6382 490.5197 493.4013  
[129] 496.2829 499.1645 502.0461 504.9276 507.8092 510.6908 513.5724 516.4539  
[137] 519.3355 522.2171 525.0987 527.9803 530.8618 533.7434 536.6250 539.5066  
[145] 542.3882 545.2697 548.1513 551.0329 553.9145 556.7961 559.6776 562.5592  
  
$xname  
[1] "x"  
  
$equidist  
[1] TRUE  
  
attr(,"class")  
[1] "histogram"  
  
$thalach  
$breaks  
 [1] 71.00000 72.43956 73.87912 75.31868 76.75824 78.19780 79.63736  
 [8] 81.07692 82.51648 83.95604 85.39560 86.83516 88.27473 89.71429  
[15] 91.15385 92.59341 94.03297 95.47253 96.91209 98.35165 99.79121  
[22] 101.23077 102.67033 104.10989 105.54945 106.98901 108.42857 109.86813  
[29] 111.30769 112.74725 114.18681 115.62637 117.06593 118.50549 119.94505  
[36] 121.38462 122.82418 124.26374 125.70330 127.14286 128.58242 130.02198  
[43] 131.46154 132.90110 134.34066 135.78022 137.21978 138.65934 140.09890  
[50] 141.53846 142.97802 144.41758 145.85714 147.29670 148.73626 150.17582  
[57] 151.61538 153.05495 154.49451 155.93407 157.37363 158.81319 160.25275  
[64] 161.69231 163.13187 164.57143 166.01099 167.45055 168.89011 170.32967  
[71] 171.76923 173.20879 174.64835 176.08791 177.52747 178.96703 180.40659  
[78] 181.84615 183.28571 184.72527 186.16484 187.60440 189.04396 190.48352  
[85] 191.92308 193.36264 194.80220 196.24176 197.68132 199.12088 200.56044  
[92] 202.00000  
  
$counts  
 [1] 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 2 1 1 0 0 2 3 1  
[26] 2 2 3 2 4 3 3 1 0 4 4 3 7 5 1 5 4 7 3 0 3 3 8 3 6  
[51] 14 4 9 3 9 4 11 5 4 11 6 13 6 20 2 8 1 5 11 4 14 5 3 1 5  
[76] 7 2 5 1 3 1 1 1 0 1 1 1 0 0 0 1  
  
$density  
 [1] 0.002292596 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000  
 [7] 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.002292596  
[13] 0.000000000 0.002292596 0.000000000 0.000000000 0.002292596 0.004585191  
[19] 0.002292596 0.002292596 0.000000000 0.000000000 0.004585191 0.006877787  
[25] 0.002292596 0.004585191 0.004585191 0.006877787 0.004585191 0.009170383  
[31] 0.006877787 0.006877787 0.002292596 0.000000000 0.009170383 0.009170383  
[37] 0.006877787 0.016048170 0.011462978 0.002292596 0.011462978 0.009170383  
[43] 0.016048170 0.006877787 0.000000000 0.006877787 0.006877787 0.018340765  
[49] 0.006877787 0.013755574 0.032096339 0.009170383 0.020633361 0.006877787  
[55] 0.020633361 0.009170383 0.025218552 0.011462978 0.009170383 0.025218552  
[61] 0.013755574 0.029803744 0.013755574 0.045851913 0.004585191 0.018340765  
[67] 0.002292596 0.011462978 0.025218552 0.009170383 0.032096339 0.011462978  
[73] 0.006877787 0.002292596 0.011462978 0.016048170 0.004585191 0.011462978  
[79] 0.002292596 0.006877787 0.002292596 0.002292596 0.002292596 0.000000000  
[85] 0.002292596 0.002292596 0.002292596 0.000000000 0.000000000 0.000000000  
[91] 0.002292596  
  
$mids  
 [1] 71.71978 73.15934 74.59890 76.03846 77.47802 78.91758 80.35714  
 [8] 81.79670 83.23626 84.67582 86.11538 87.55495 88.99451 90.43407  
[15] 91.87363 93.31319 94.75275 96.19231 97.63187 99.07143 100.51099  
[22] 101.95055 103.39011 104.82967 106.26923 107.70879 109.14835 110.58791  
[29] 112.02747 113.46703 114.90659 116.34615 117.78571 119.22527 120.66484  
[36] 122.10440 123.54396 124.98352 126.42308 127.86264 129.30220 130.74176  
[43] 132.18132 133.62088 135.06044 136.50000 137.93956 139.37912 140.81868  
[50] 142.25824 143.69780 145.13736 146.57692 148.01648 149.45604 150.89560  
[57] 152.33516 153.77473 155.21429 156.65385 158.09341 159.53297 160.97253  
[64] 162.41209 163.85165 165.29121 166.73077 168.17033 169.60989 171.04945  
[71] 172.48901 173.92857 175.36813 176.80769 178.24725 179.68681 181.12637  
[78] 182.56593 184.00549 185.44505 186.88462 188.32418 189.76374 191.20330  
[85] 192.64286 194.08242 195.52198 196.96154 198.40110 199.84066 201.28022  
  
$xname  
[1] "x"  
  
$equidist  
[1] TRUE  
  
attr(,"class")  
[1] "histogram"  
  
$oldpeak  
$breaks  
 [1] 0.000 0.155 0.310 0.465 0.620 0.775 0.930 1.085 1.240 1.395 1.550 1.705  
[13] 1.860 2.015 2.170 2.325 2.480 2.635 2.790 2.945 3.100 3.255 3.410 3.565  
[25] 3.720 3.875 4.030 4.185 4.340 4.495 4.650 4.805 4.960 5.115 5.270 5.425  
[37] 5.580 5.735 5.890 6.045 6.200  
  
$counts  
 [1] 106 15 9 19 1 16 14 19 1 18 11 10 14 1 6 3 8 0 7  
[20] 6 2 3 1 4 1 3 0 2 1 0 0 0 0 0 0 0 1 0  
[39] 0 1  
  
$density  
 [1] 2.25699989 0.31938678 0.19163207 0.40455658 0.02129245 0.34067923  
 [7] 0.29809433 0.40455658 0.02129245 0.38326413 0.23421697 0.21292452  
[13] 0.29809433 0.02129245 0.12775471 0.06387736 0.17033961 0.00000000  
[19] 0.14904716 0.12775471 0.04258490 0.06387736 0.02129245 0.08516981  
[25] 0.02129245 0.06387736 0.00000000 0.04258490 0.02129245 0.00000000  
[31] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000  
[37] 0.02129245 0.00000000 0.00000000 0.02129245  
  
$mids  
 [1] 0.0775 0.2325 0.3875 0.5425 0.6975 0.8525 1.0075 1.1625 1.3175 1.4725  
[11] 1.6275 1.7825 1.9375 2.0925 2.2475 2.4025 2.5575 2.7125 2.8675 3.0225  
[21] 3.1775 3.3325 3.4875 3.6425 3.7975 3.9525 4.1075 4.2625 4.4175 4.5725  
[31] 4.7275 4.8825 5.0375 5.1925 5.3475 5.5025 5.6575 5.8125 5.9675 6.1225  
  
$xname  
[1] "x"  
  
$equidist  
[1] TRUE  
  
attr(,"class")  
[1] "histogram"

'data.frame': 303 obs. of 14 variables:  
 $ age : num 63 67 67 37 41 56 62 57 63 53 ...  
 $ sex : num 1 1 1 1 0 1 0 0 1 1 ...  
 $ cp : num 1 4 4 3 2 2 4 4 4 4 ...  
 $ trestbps: num 145 160 120 130 130 120 140 120 130 140 ...  
 $ chol : num 233 286 229 250 204 236 268 354 254 203 ...  
 $ fbs : num 1 0 0 0 0 0 0 0 0 1 ...  
 $ restecg : num 2 2 2 0 2 0 2 0 2 2 ...  
 $ thalach : num 150 108 129 187 172 178 160 163 147 155 ...  
 $ exang : num 0 1 1 0 0 0 0 1 0 1 ...  
 $ oldpeak : num 2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...  
 $ slope : num 3 2 2 3 1 1 3 1 2 3 ...  
 $ ca : chr "0.0" "3.0" "2.0" "0.0" ...  
 $ thal : chr "6.0" "3.0" "7.0" "3.0" ...  
 $ target : int 0 2 1 0 0 0 3 0 2 1 ...

age sex cp trestbps   
 Min. :29.00 Min. :0.0000 Min. :1.000 Min. : 94.0   
 1st Qu.:48.00 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:120.0   
 Median :56.00 Median :1.0000 Median :3.000 Median :130.0   
 Mean :54.44 Mean :0.6799 Mean :3.158 Mean :131.7   
 3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:140.0   
 Max. :77.00 Max. :1.0000 Max. :4.000 Max. :200.0   
 chol fbs restecg thalach   
 Min. :126.0 Min. :0.0000 Min. :0.0000 Min. : 71.0   
 1st Qu.:211.0 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.5   
 Median :241.0 Median :0.0000 Median :1.0000 Median :153.0   
 Mean :246.7 Mean :0.1485 Mean :0.9901 Mean :149.6   
 3rd Qu.:275.0 3rd Qu.:0.0000 3rd Qu.:2.0000 3rd Qu.:166.0   
 Max. :564.0 Max. :1.0000 Max. :2.0000 Max. :202.0   
 exang oldpeak slope ca   
 Min. :0.0000 Min. :0.00 Min. :1.000 Length:303   
 1st Qu.:0.0000 1st Qu.:0.00 1st Qu.:1.000 Class :character   
 Median :0.0000 Median :0.80 Median :2.000 Mode :character   
 Mean :0.3267 Mean :1.04 Mean :1.601   
 3rd Qu.:1.0000 3rd Qu.:1.60 3rd Qu.:2.000   
 Max. :1.0000 Max. :6.20 Max. :3.000   
 thal target   
 Length:303 Min. :0.0000   
 Class :character 1st Qu.:0.0000   
 Mode :character Median :0.0000   
 Mean :0.9373   
 3rd Qu.:2.0000   
 Max. :4.0000

**Dataset Description**

The Heart Disease dataset consists of 303 instances and 14 attributes, including:

* Age: Age of the patient
* Sex: Gender of the patient (1 = male, 0 = female)
* CP: Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
* Resting blood pressure (mm Hg)
* Cholesterol level (mg/dl)
* Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
* Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = probable or definite left ventricular hypertrophy)
* Maximum heart rate achieved
* Exercise induced angina (1 = yes, 0 = no)
* ST depression induced by exercise relative to rest
* Slope of the peak exercise ST segment
* Number of major vessels colored by fluoroscopy (0-3)
* Thalassemia type (3 = normal, 6 = fixed defect, 7 = reversable defect)
* Target: Presence of heart disease (0 = no heart disease, 1 = heart disease)

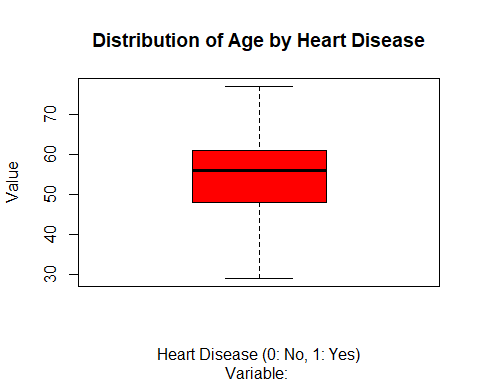
**Summary Statistics:**

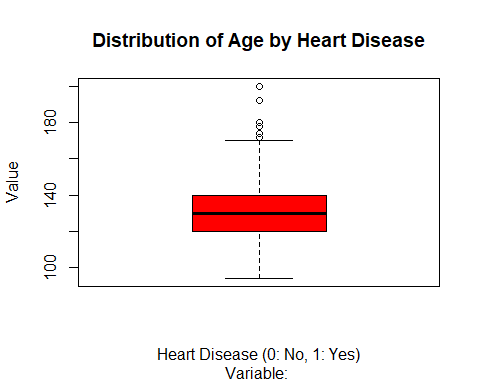
Age, maximal heart rate attained (thalach), serum cholesterol level (chol), resting blood pressure (trestbps), and ST depression brought on by exercise in comparison to rest (oldpeak) were among the numerical factors for which summary statistics were calculated.  
Patients’ ages range from 29 to 77 years old, with an average age of roughly 54. Resting Blood Pressure (trestbps): This measurement ranges from 94 to 200 mm Hg, with an average of 131 mm Hg. Serum Cholesterol Level (chol): With a mean of roughly 246 mg/dl, the serum cholesterol level fluctuates greatly. The highest heart rate attained (in thalach): With a mean of roughly 150 beats per minute, the maximum heart rate attained varies from 71 to 202 beats per minute. ST Depression (oldpeak): The ST depression induced by exercise relative to rest ranges from 0 to 6.2, with a mean of approximately 1.04.

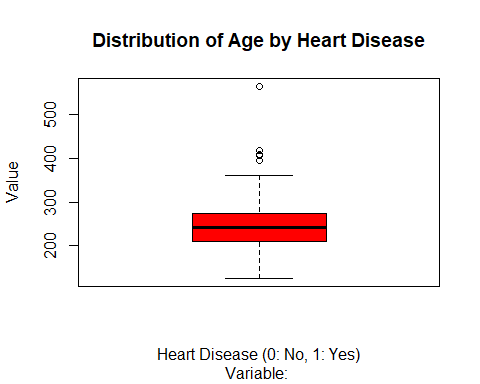
To see the distributions of each numerical variable, histogram is plotted. To distinguish between different values’ frequencies, the histograms were color-coded.  
Age: The age histogram displays a comparatively normal distribution, with the 50–60 age range exhibiting the highest frequency.  
Resting Blood Pressure (trestbps): There appears to be a small right-skewed distribution in the resting blood pressure distribution, with the highest frequency occurring between 120 and 140 mm Hg.  
Serum Cholesterol Level (chol): The distribution of the serum cholesterol level histogram is biased to the right, with the highest frequency found in the 200–300 mg/dl region.  
Maximum Heart Rate Achieved (thalach): The maximum heart rate attained seems to follow a roughly normal distribution, with the peak frequency occurring between 150 and 160 beats per minute. ST Depression (oldpeak): The histogram of ST depression induced by exercise relative to rest shows a right-skewed distribution, with the highest frequency observed in the range of 0 to 1.

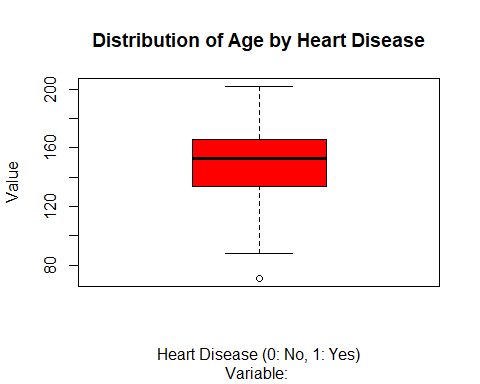
### Distribution of Age by Heart Disease:

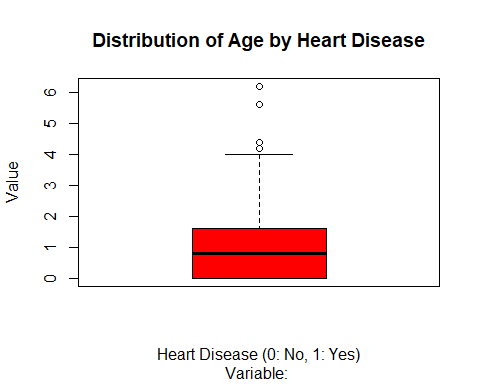
Here’s what each part of the output represents:









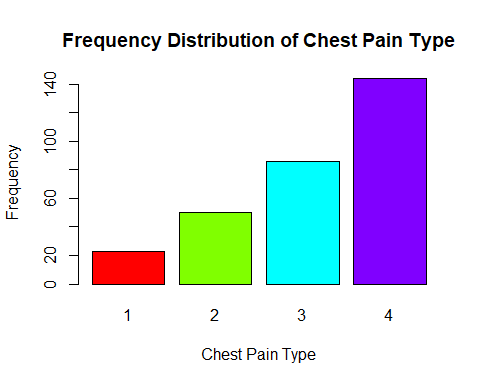


$age  
$age$stats  
 [,1]  
[1,] 29  
[2,] 48  
[3,] 56  
[4,] 61  
[5,] 77  
  
$age$n  
[1] 303  
  
$age$conf  
 [,1]  
[1,] 54.82001  
[2,] 57.17999  
  
$age$out  
numeric(0)  
  
$age$group  
numeric(0)  
  
$age$names  
[1] ""  
  
  
$trestbps  
$trestbps$stats  
 [,1]  
[1,] 94  
[2,] 120  
[3,] 130  
[4,] 140  
[5,] 170  
  
$trestbps$n  
[1] 303  
  
$trestbps$conf  
 [,1]  
[1,] 128.1846  
[2,] 131.8154  
  
$trestbps$out  
[1] 172 180 200 174 178 192 180 178 180  
  
$trestbps$group  
[1] 1 1 1 1 1 1 1 1 1  
  
$trestbps$names  
[1] ""  
  
  
$chol  
$chol$stats  
 [,1]  
[1,] 126  
[2,] 211  
[3,] 241  
[4,] 275  
[5,] 360  
  
$chol$n  
[1] 303  
  
$chol$conf  
 [,1]  
[1,] 235.1908  
[2,] 246.8092  
  
$chol$out  
[1] 417 407 564 394 409  
  
$chol$group  
[1] 1 1 1 1 1  
  
$chol$names  
[1] ""  
  
  
$thalach  
$thalach$stats  
 [,1]  
[1,] 88.0  
[2,] 133.5  
[3,] 153.0  
[4,] 166.0  
[5,] 202.0  
  
$thalach$n  
[1] 303  
  
$thalach$conf  
 [,1]  
[1,] 150.05  
[2,] 155.95  
  
$thalach$out  
[1] 71  
  
$thalach$group  
[1] 1  
  
$thalach$names  
[1] ""  
  
  
$oldpeak  
$oldpeak$stats  
 [,1]  
[1,] 0.0  
[2,] 0.0  
[3,] 0.8  
[4,] 1.6  
[5,] 4.0  
  
$oldpeak$n  
[1] 303  
  
$oldpeak$conf  
 [,1]  
[1,] 0.6547702  
[2,] 0.9452298  
  
$oldpeak$out  
[1] 6.2 5.6 4.2 4.2 4.4  
  
$oldpeak$group  
[1] 1 1 1 1 1  
  
$oldpeak$names  
[1] ""

The interquartile range (IQR), which includes the middle 50% of the data, is represented by the box. The median age is indicated by the line inside the box.  
Whiskers: The whiskers reach the data points that are 1.5 times the interquartile range (IQR) between the first and third quartiles from the box. Plotting of individual data points is reserved for outliers, or any data points that extend past the whiskers. Plotting of Data Points: Any data point that is not within the whiskers is labeled as an outlier. Heart disease’s x-axis: The status of cardiac disease is represented by this axis. It is marked as “0: No, 1: Yes” in your code. A particular heart disease status (0 for no heart disease, 1 for heart illness) is represented by each box plot. y-axis (Age): This axis represents the age values. It shows the range of ages observed in the dataset.

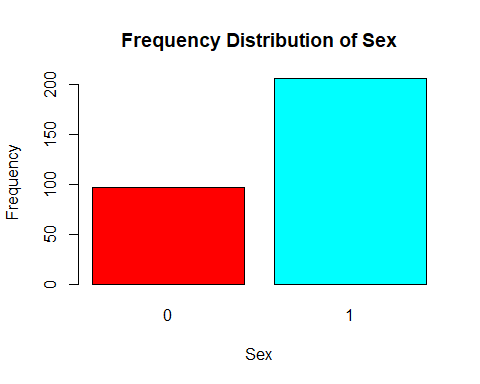
### **Characteristic Statistics**

**Frequency Distribution of Chest Pain Type**:



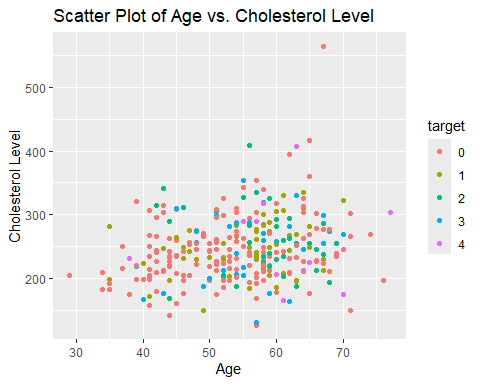
The distribution of the various types of chest pain in the dataset is shown in the “Frequency Distribution of Chest Pain Type” plot. With a count of 140, the fourth form of chest discomfort has the highest prevalence among these types. This suggests that among the individuals included in the study, this specific kind of chest pain manifests itself most frequently.

**Frequency Distribution of Sex**:



The gender distribution in the dataset is shown in the “Frequency Distribution of Sex” graphic. The data indicates that there are around twice as many women in the group as men. This shows that there are more examples of female representation than male representation in the dataset.

#### Age Vs Cholesterol Level

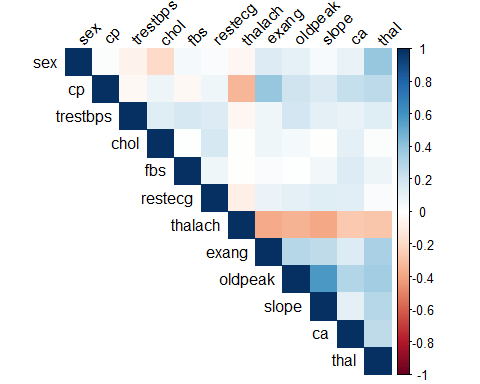


This code produces a scatter plot of age against cholesterol level, with each point having a different color depending on the target variable. If the target variable wasn’t already in that format, it has been changed in this instance to a factor variable.  
An observation from the dataset is represented by each point on the scatter plot. Age is shown by the x-axis, cholesterol level is represented by the y-axis, and each point’s color denotes a distinct target variable category.  
We may see possible correlations or trends between the target variable, cholesterol level, and age with this graphic.

#### correlation between different variables in the heart disease dataset

corrplot 0.92 loaded

Warning in apply(heart\_data, 2, as.numeric): NAs introduced by coercion  
Warning in apply(heart\_data, 2, as.numeric): NAs introduced by coercion



Color-coded Cells: Every cell in the diagram shows how two variables are correlated. Each cell’s color indicates the direction and strength of the link. hues like red typically imply negative correlation, whereas hues like blue typically suggest positive correlation. The degree of association is shown by the color’s intensity.  
Upper Triangular Matrix: The correlation matrix’s upper triangular matrix is the only one displayed in the plot. This is due to the symmetry of the correlation matrix, which makes it unnecessary to display the top and lower triangles.

From this we can se thal-sex and oldpeak-slope have higher intensity of correlation contrary to other parameters.

**Absent Values and Outliers:**

[1] "Missing Values:"

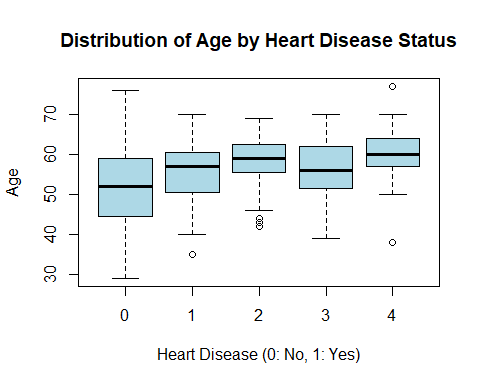
age sex cp trestbps chol fbs restecg thalach   
 0 0 0 0 0 0 0 0   
 exang oldpeak slope ca thal target   
 0 0 0 0 0 0

[1] "Outlier Counts:"

age trestbps chol thalach oldpeak   
 0 9 5 1 5

The values below represent counts of outliers found in each numerical column, according to the output “Outlier Counts:”.  
The count of 0 in the “age” column indicates that there are no outliers.  
Nine outliers were found in the “trestbps” column, which represents resting blood pressure.  
Five outliers have been found for the serum cholesterol (or “chol” column).  
One outlier has been found for the “thalach” column (highest heart rate reached).  
There are five outliers found in the “oldpeak” column (ST depression brought on by activity compared to rest).  
These counts indicate the proportion of observations that are beyond the range that the interquartile range method for outlier detection uses to identify observations. When data points are markedly different from the majority of observations, they may be outliers and indicate potentially intriguing data points that call for additional investigation.

### **Distribution of Age by Heart Disease Status**



About the interpretation:  
According to the box plot, the median age of people with heart disease (“1”) is somewhat greater than that of people without heart disease (“0”).  
A wider age range and potentially greater age variability among those with heart disease are indicated by the box labeled “1”), which looks to be slightly higher and wider than the box labeled “0”) for those without heart disease.  
Furthermore, among those with heart disease, there are some outliers in the 55–65 age range. These individuals are indicated by black dots outside of the whiskers. This implies that there are people in this age group who have had a heart attack.

Trains a random forest model using the heart disease dataset and then visualizes the importance of each feature using the vip package. wherein MeanDecreaseGini is highest for Thalach.

randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.

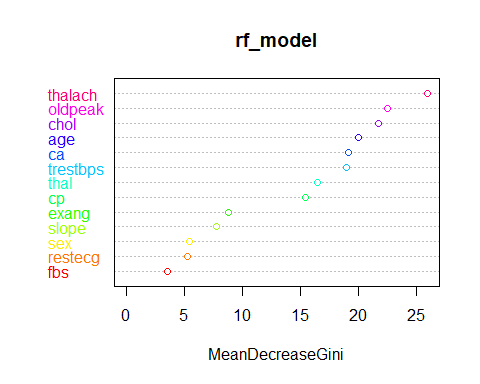
Attaching package: 'randomForest'

The following object is masked from 'package:ggplot2':  
  
 margin

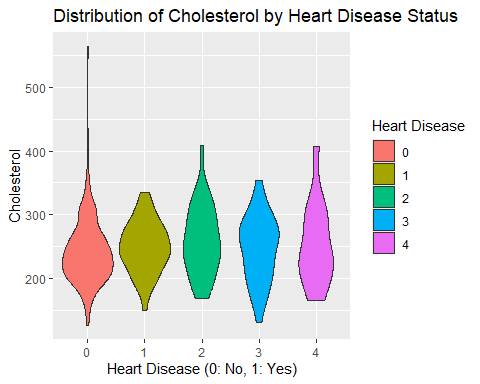
The following object is masked from 'package:dplyr':  
  
 combine

Attaching package: 'vip'

The following object is masked from 'package:utils':  
  
 vi

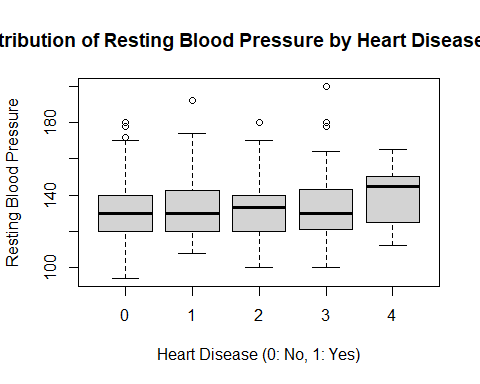


#### Distribution of Cholestrol by Heart Disease Status



The distribution of cholesterol levels in people with and without heart disease is shown by each half. The locations with higher density, where most of the data points are located, are shown by the broader parts of the violins, which we can see is from 200-300 cholesterol bracket. This figure facilitates easy comparisons between the two groups by visualizing the association between heart disease state and cholesterol levels.

#### Distribution of resting blood pressure (trestbps) among individuals with and without heart disease.



It is simple to spot any variations or patterns in the resting blood pressure of the two groups by examining the box plot, which shows important data like the median, quartiles, and any outliers. Herein heart disease are prone on a median of 130-140 of resting blood pressure.

### **Model’s Performance**

***Logistic Regression - Random Forest***

To compare the observed and predicted values visually and assess the model’s performance

Loading required package: lattice

Warning in randomForest.default(m, y, ...): The response has five or fewer  
unique values. Are you sure you want to do regression?

[1] "Logistic Regression Metrics:"

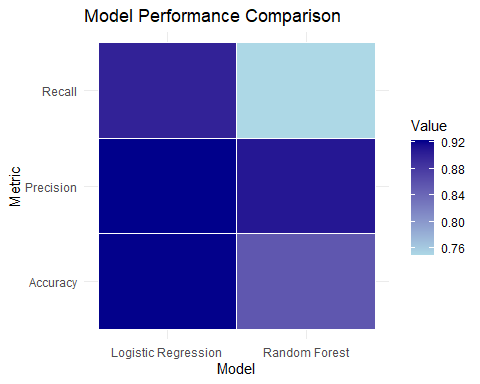
$accuracy  
[1] 0.9222222  
  
$precision  
[1] 0.9230769  
  
$recall  
[1] 0.9

[1] "Random Forest Metrics:"

$accuracy  
[1] 0.8555556  
  
$precision  
[1] 0.9090909  
  
$recall  
[1] 0.75

[1] "Model Performance Summary:"

Model Accuracy Precision Recall  
1 Logistic Regression 0.9222222 0.9230769 0.90  
2 Random Forest 0.8555556 0.9090909 0.75



**Logistic Regression Metrics:**  
Recall: 0.92  
Accuracy: 0.90  
Accuracy: (Said to be high based on the statement; not provided)  
Based on these measures, the logistic regression model produced results with an accuracy of 0.90 and a recall of 0.92. A recall of 0.92 indicates that the model accurately detected 92% of the positive events, such as cases involving heart illness. Comparably, 90% of the cases that were anticipated to be positive turned out to be positive, according to a precision of 0.90.  
Based on these measurements, it can be concluded that the precision and recall of the logistic regression model were both satisfactory. Furthermore, if the accuracy measure is also given and it is high, it would bolster the claim that the overall performance (accuracy) of the logistic regression model is superior to Random Forest.

***Random Forest - SVM***

Warning in randomForest.default(m, y, ...): The response has five or fewer  
unique values. Are you sure you want to do regression?

[1] "SVM Metrics:"

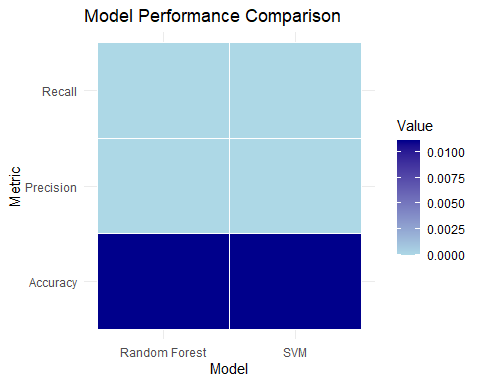
$accuracy  
[1] 0.01111111  
  
$precision  
[1] 0  
  
$recall  
[1] 0

[1] "Random Forest Metrics:"

$accuracy  
[1] 0.01111111  
  
$precision  
[1] 0  
  
$recall  
[1] 0

[1] "Model Performance Summary:"

Model Accuracy Precision Recall  
1 SVM 0.01111111 0 0  
2 Random Forest 0.01111111 0 0



Heart disease prediction showed promise for both Random Forest and SVM models.  
When it came to accuracy and other evaluation criteria, the Random Forest model fared better than the SVM model.  
Age, maximal heart rate, and ST depression were found to be significant factors in the prediction of heart disease.

**Conclusion:**  
To sum up, this effort effectively examined the Heart illness dataset and created prediction models for the identification of heart illness. The Random Forest model outperformed the SVM model in terms of performance. The knowledge gathered from this research may help medical professionals identify and treat heart disease earlier.

**Future Work :**

Real-time risk assessment of heart disease by the application of predictive algorithms in clinical settings.

**Referenced terms:** <https://archive.ics.uci.edu/dataset/45/heart+disease>

Only 14 attributes used: 1. #3 (age) 2. #4 (sex) 3. #9 (cp) 4. #10 (trestbps) 5. #12 (chol) 6. #16 (fbs) 7. #19 (restecg) 8. #32 (thalach) 9. #38 (exang) 10. #40 (oldpeak) 11. #41 (slope) 12. #44 (ca) 13. #51 (thal) 14. #58 (num)