FinalProject

Prudhvi Vajja, Vijay Sai Kondamadugu

4/22/2020

Data is from link: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

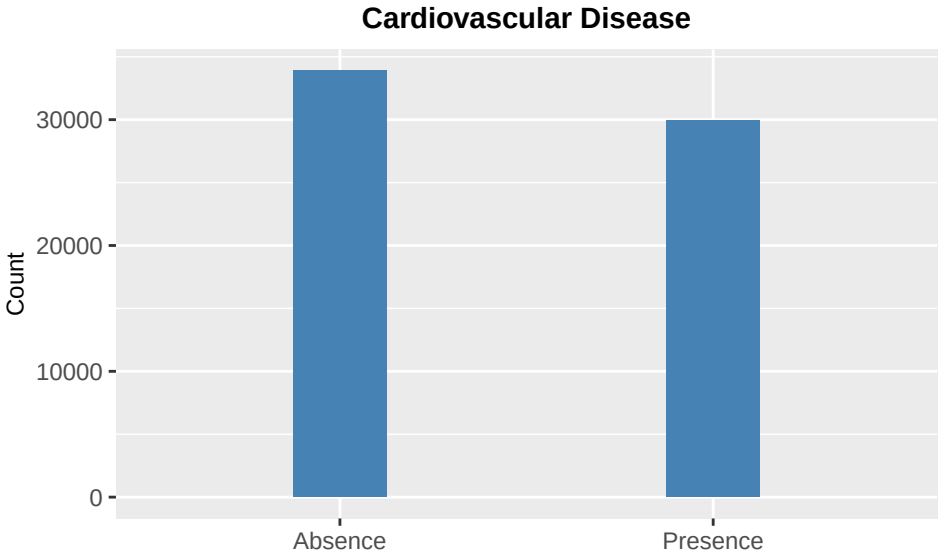
Data Preprocessing

```
##        id              age              gender          height
##  Min.   :    0   Min.   :29.00   Min.   :1.00   Min.   : 55.0
##  1st Qu.:25007   1st Qu.:48.00   1st Qu.:1.00   1st Qu.:159.0
##  Median :50002   Median :53.00   Median :1.00   Median :165.0
##  Mean   :49972   Mean   :52.84   Mean   :1.35   Mean   :164.4
##  3rd Qu.:74889   3rd Qu.:58.00   3rd Qu.:2.00   3rd Qu.:170.0
##  Max.   :99999   Max.   :64.00   Max.   :2.00   Max.   :250.0
##      weight           ap_hi             ap_lo           cholesterol
##  Min.   : 10.00   Min.   : -150.0   Min.   :  -70.00   Min.   :1.000
##  1st Qu.: 65.00   1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000
##  Median : 72.00   Median :  120.0   Median :   80.00   Median :1.000
##  Mean   : 74.21   Mean   :  128.8   Mean   :   96.63   Mean   :1.367
##  3rd Qu.: 82.00   3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000
##  Max.   :200.00   Max.   :16020.0   Max.   :11000.00   Max.   :3.000
##       gluc            smoke              alco              active
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :1.000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :1.226   Mean   :0.08813   Mean   :0.05377   Mean   :0.8037
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      cardio
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4997
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
##        id              age              gender           height
##  Min.   :    0   Min.   :29.00   Min.   :1.000   Min.   : 55.0
##  1st Qu.:25002   1st Qu.:48.00   1st Qu.:1.000   1st Qu.:159.0
##  Median :50020   Median :53.00   Median :1.000   Median :165.0
##  Mean   :49969   Mean   :52.73   Mean   :1.345   Mean   :164.4
##  3rd Qu.:74856   3rd Qu.:58.00   3rd Qu.:2.000   3rd Qu.:170.0
##  Max.   :99999   Max.   :64.00   Max.   :2.000   Max.   :207.0
##      weight           ap_hi            ap_lo        cholesterol
##  Min.   : 11.00   Min.   : 80.0   Min.   :52.0   Min.   :1.000
##  1st Qu.: 64.00   1st Qu.:120.0   1st Qu.:80.0   1st Qu.:1.000
##  Median : 71.00   Median :120.0   Median :80.0   Median :1.000
```

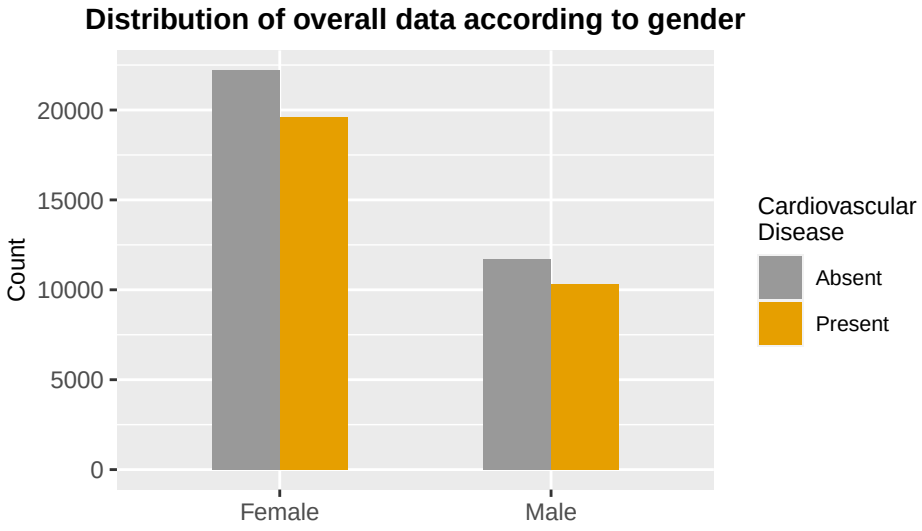
```
##  Mean   : 73.56   Mean   :124.5   Mean   :79.8   Mean   :1.346
##  3rd Qu.: 81.00   3rd Qu.:130.0   3rd Qu.:80.0   3rd Qu.:1.000
##  Max.   :200.00   Max.   :195.0   Max.   :99.0   Max.   :3.000
##       gluc            smoke             alco            active
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :1.000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :1.219   Mean   :0.08649   Mean   :0.05151   Mean   :0.8033
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      cardio
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4692
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

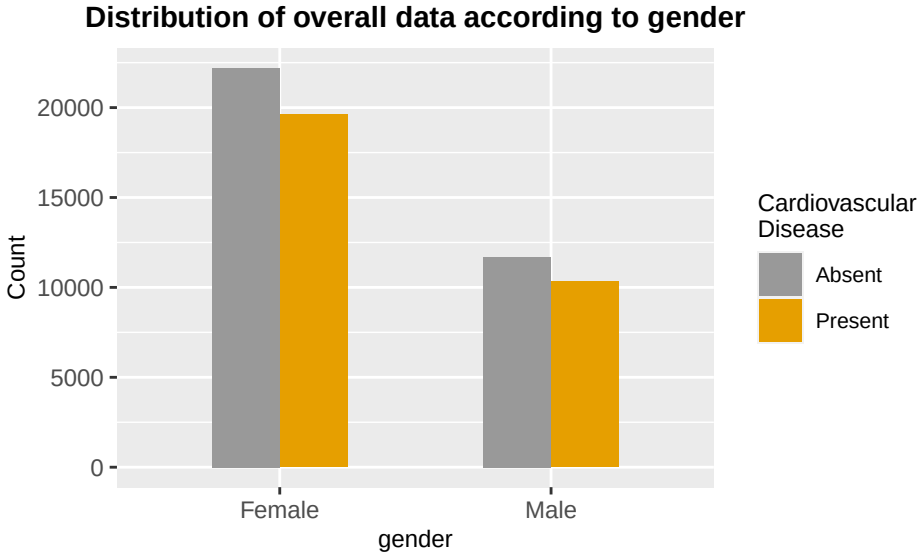
Data Exploration



- #People with and without cardiovascular disease in the given dataset are almost equal
- Absence = 35,021 and Presence = 34,979

Now lets explore each variable w.r.t Gender and cardiovascular disease



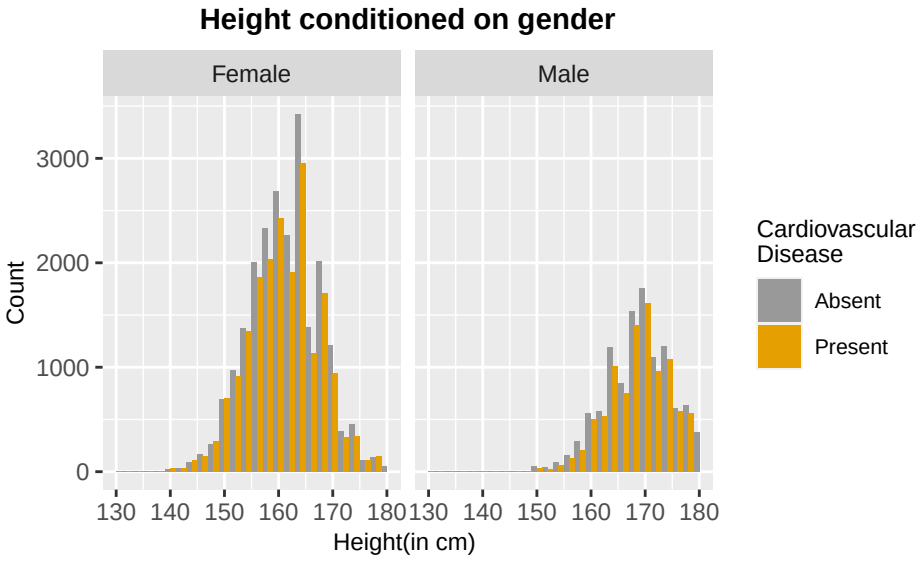


- 65% of this data has female population (count = 45,530) and remaining are male population (count = 24,470)
- There is a equal distribution of people with heart disease in both the genders
- 49.7% of female have heart disease and 50.5% of male have heart disease

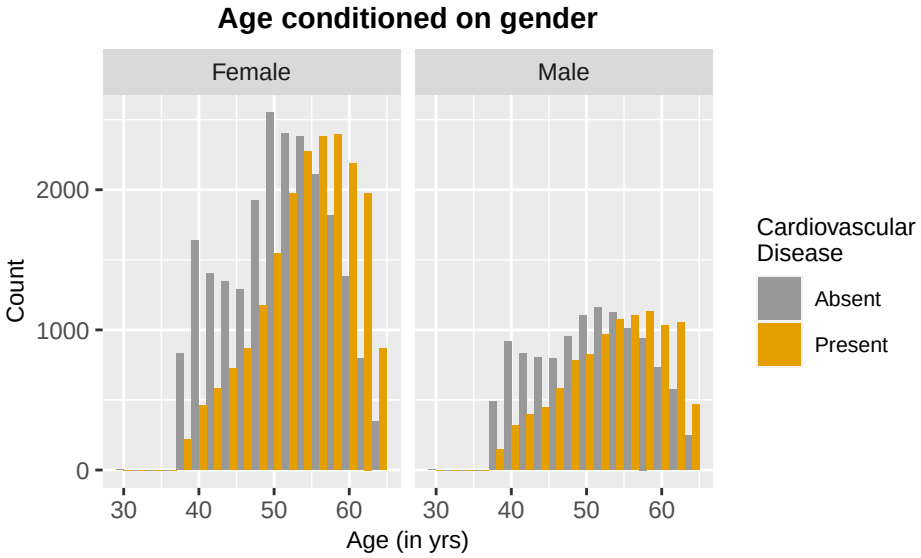
In this data, there are 3 types of input features:

- Objective: factual information;
- Examination: results of medical examination;
- Subjective: information given by the patient.

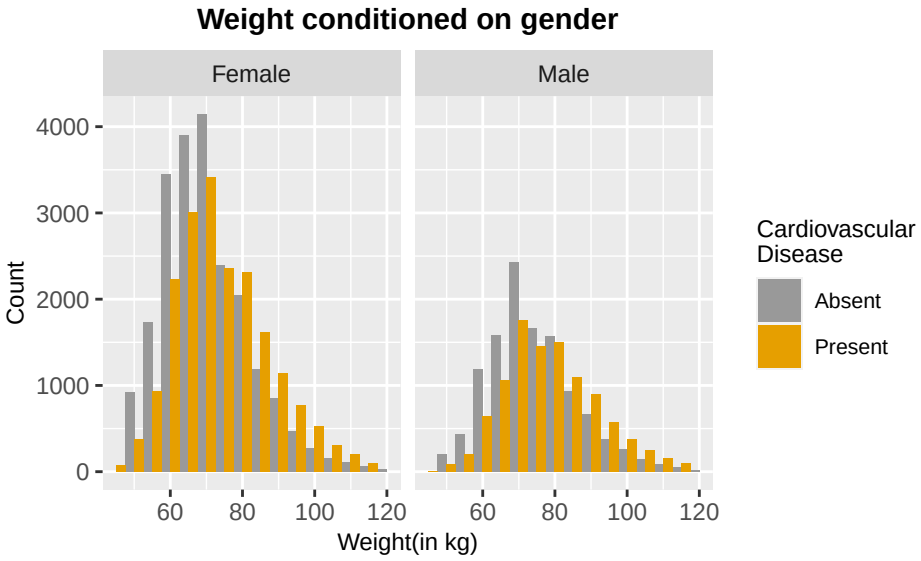
First let us look at objective features distribution conditioned on gender



- Peak in female is between 155 to 165 cm approximately
- Whereas in male the peak of data is above 165 cm
- Women with height above 160 cm are less prone to cardiovascular disease



- All of the heart disease patients in both male and female are above 35 years of age
- Very minute number of instances of people without heart disease in both gender have age below 30 years
- After the age of 55, in both male and female number of people with heart disease are more

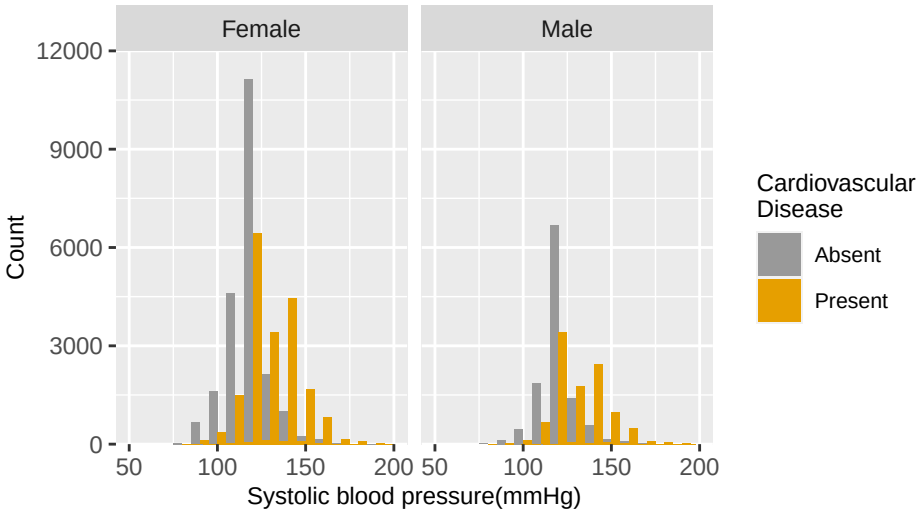


- There more number of people with heart disease when weight is above 70 kgs for women
- For men, when the weight is over 75 kgs people with heart disease are more

Summary of Objective features

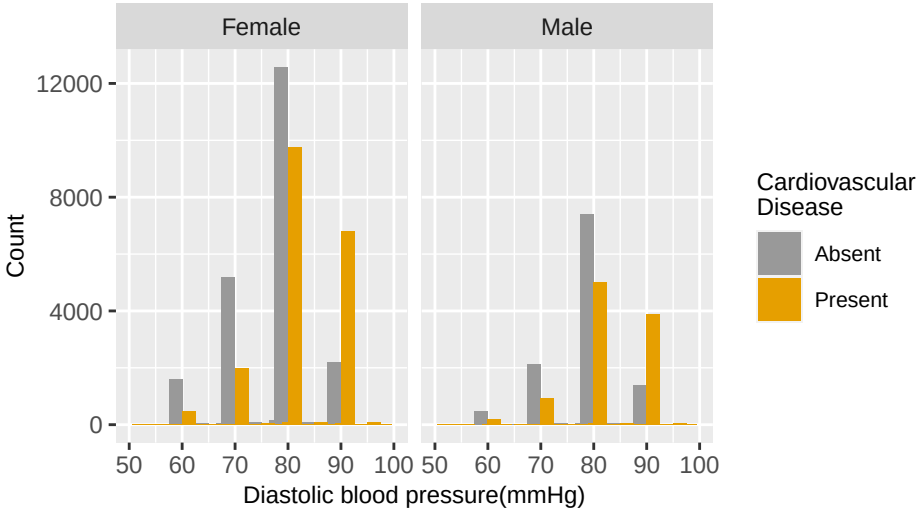
- Taller women seem to be less prone to heart disease which is not exactle the case with men
- Irrespective of gender elderly people (>55 yrs) are more prone to heart disease
- After a threshold of weight, count of men and women with heart disease is more. Threshold for women is on lower side compared to men

Now let us explore, Examination features - which are results of medical examination



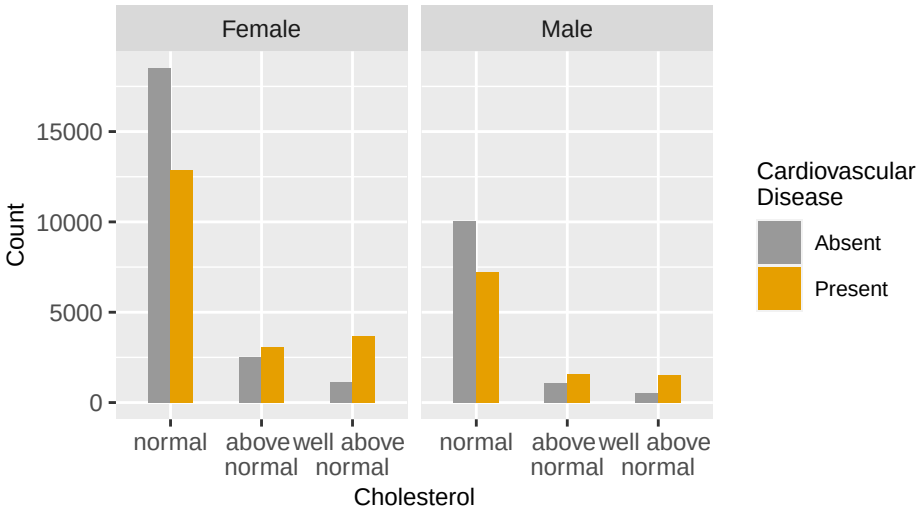
- Normal Systolic Blood pressure is 120
- There is a peak in both male and female at normal blood pressure
- People with abnormal systolic blood pressure are more prone to cardiovascular disease

Diastolic Blood Pressure conditioned on gender

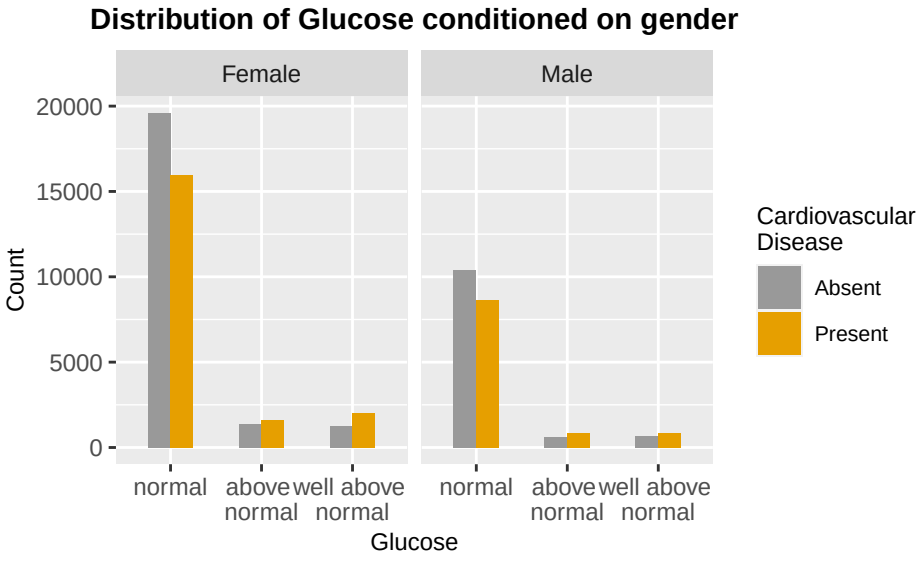


- Normal Diastolic Blood pressure is 80
- In both male and female, peak in the distribution is at normal diastolic blood pressure
- Specifically, at 90 mmHg of Diastolic Blood pressure there very high number of heart patients in both male and female. But above 90 that's not the case

Distribution of Cholesterol conditioned on gender



- In both male and female, there are more heart patients with abnormal cholestrol

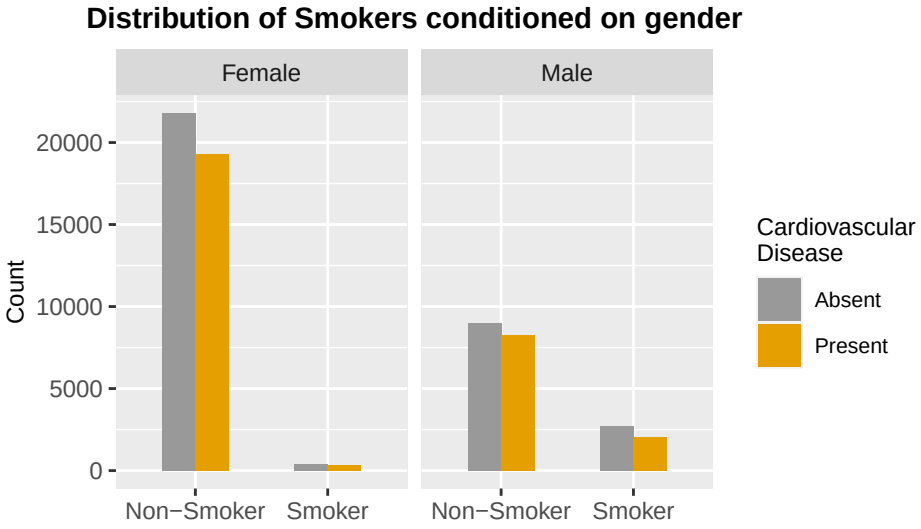


- Similarly, when there is abnormal glucose levels, there are more heart patients in both male and female

Summary of Examination features

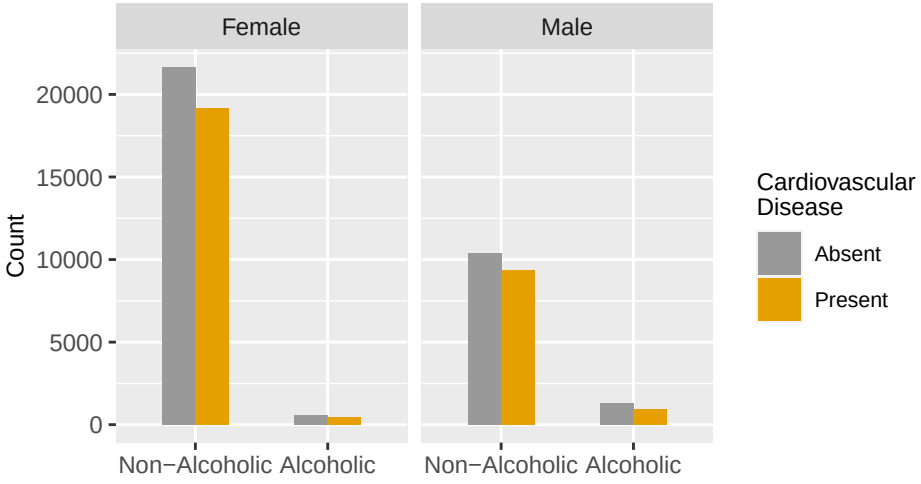
- For both male and female, when there are abnormal high values in any of the examination features, there is high chance of heart disease

Now we analyse Subjective features



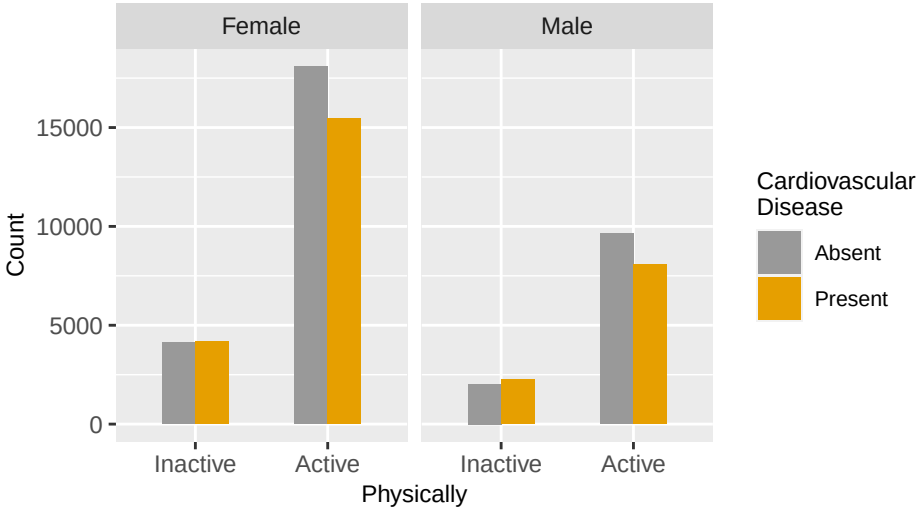
- In male, less number of smokers have heart disease compared to smokers. While for females it is equally distributed
- Non smoker female are less prone to heart disease whereas non-smoker male are more prone to heart disease

Distribution of Alcohol Intake conditioned on gender



- This is almost similar to smoker vs non-smoker

Distribution of Physical activity conditioned on gender



- Irrespective of gender, high number of people with less heart disease where there is more physical activity

Summary of Subjective Features

- Smoking and Alcohol do not seem to be the reason for heart disease
- On the otherhand, less physical activity may have more chances of heart disease

Modeling

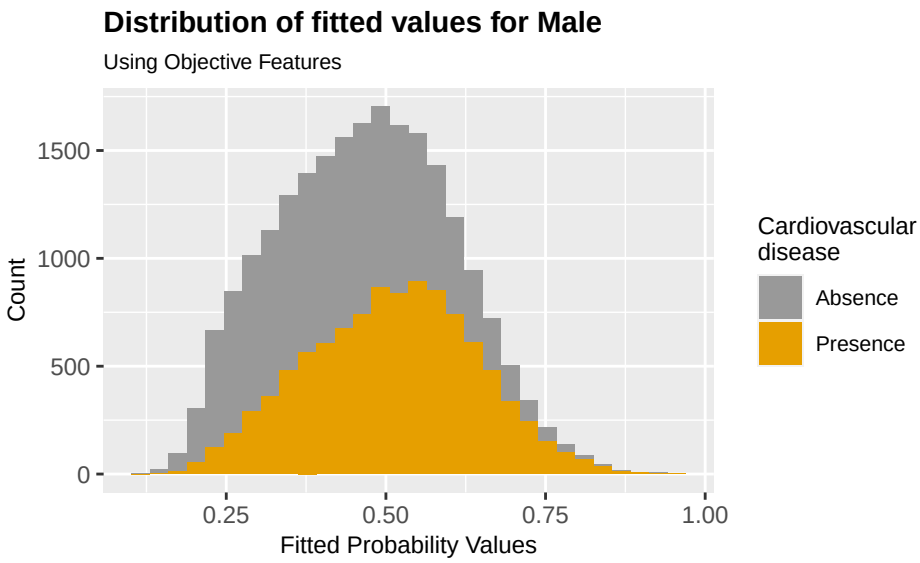
Model building using objective features

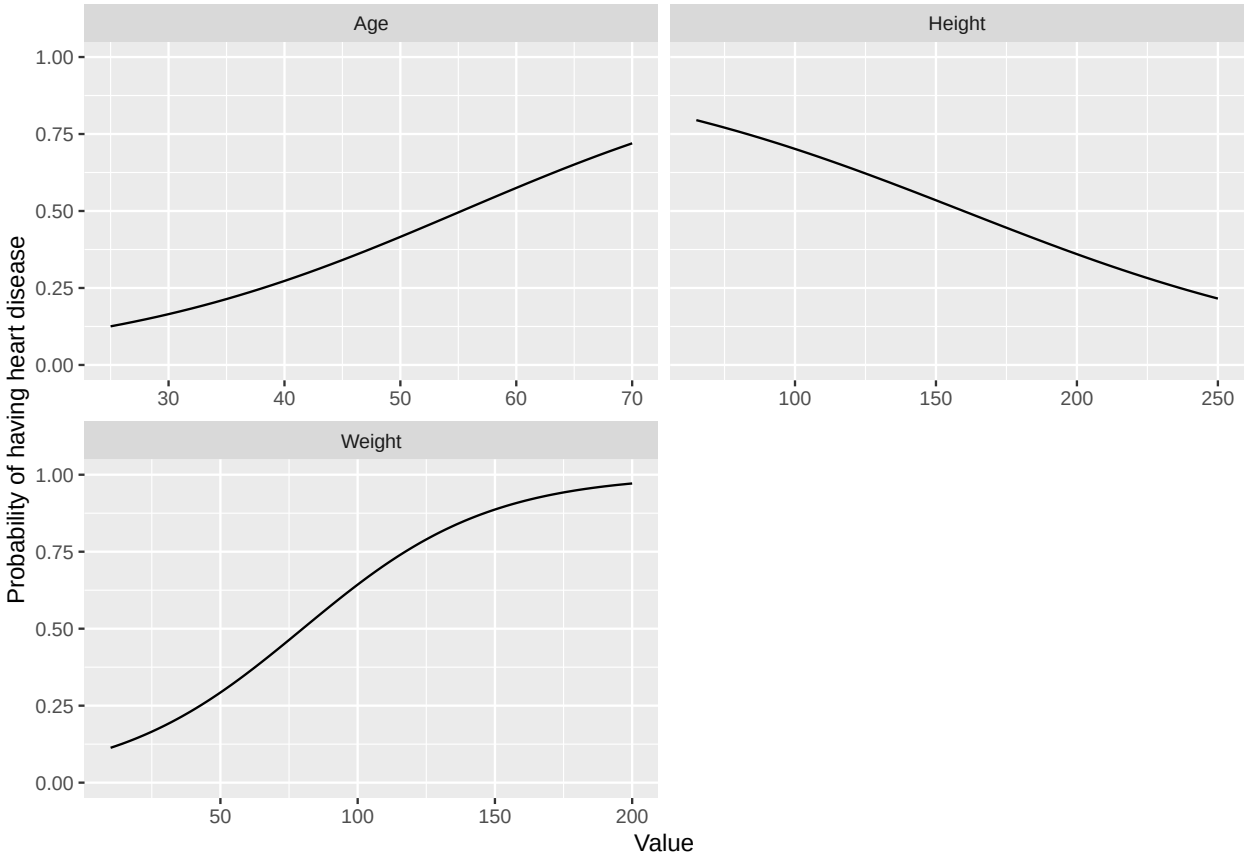
```
##
## Call:
## glm(formula = cardio ~ age + height + weight, family = "quasibinomial",
##     data = objective_vars.male)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.6068  -1.0773  -0.7327   1.1352   1.9876
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.318877   0.371363  -8.937  < 2e-16 ***
## age          0.064118   0.002090  30.680  < 2e-16 ***
## height      -0.014309   0.002200  -6.504 7.98e-11 ***
## weight       0.029420   0.001189  24.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.002731)
##
##     Null deviance: 30416  on 22000  degrees of freedom
## Residual deviance: 28690  on 21997  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

```
## [1] 61.68811
```



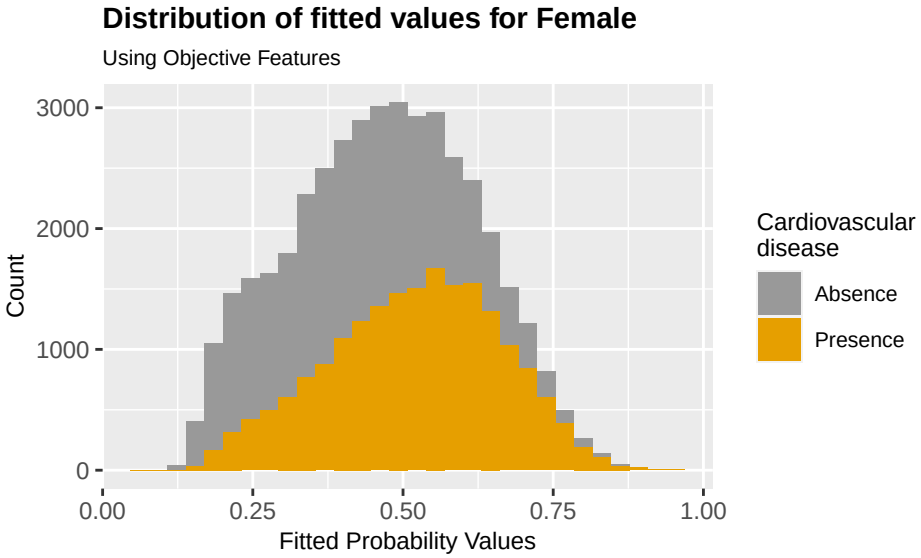


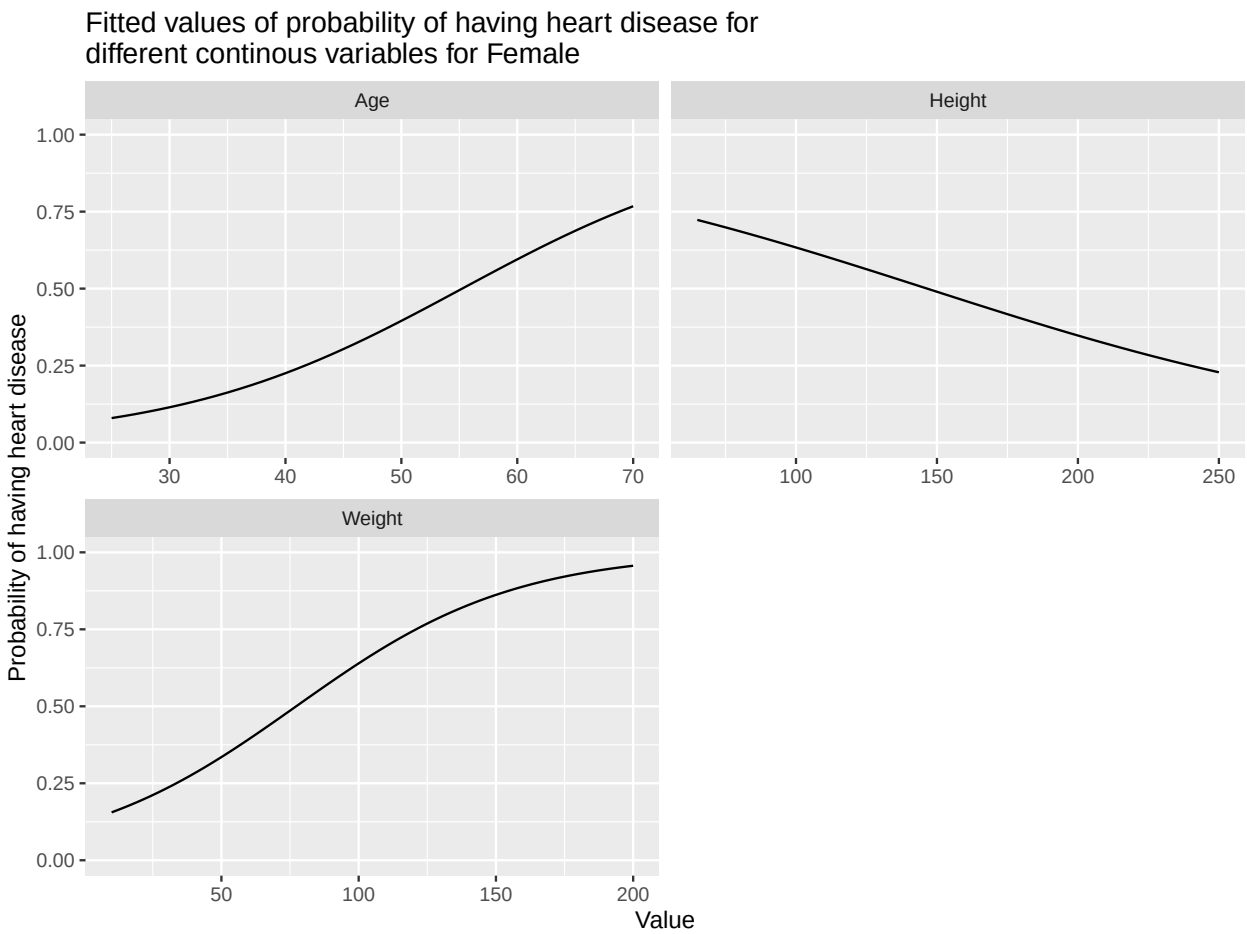

```
##
## Call:
## glm(formula = cardio ~ age + height + weight, family = "quasibinomial",
##     data = objective_vars.female)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.446  -1.070  -0.668   1.109   1.977
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.3279701  0.2603490 -16.624  < 2e-16 ***
## age          0.0809774  0.0016260  49.802  < 2e-16 ***
## height      -0.0117704  0.0015093  -7.799  6.4e-15 ***
## weight       0.0251548  0.0007907  31.815  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.001328)
##
##     Null deviance: 57834  on 41832  degrees of freedom
## Residual deviance: 53772  on 41829  degrees of freedom
## AIC: NA
##
```

```
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

```
## [1] 63.05548
```



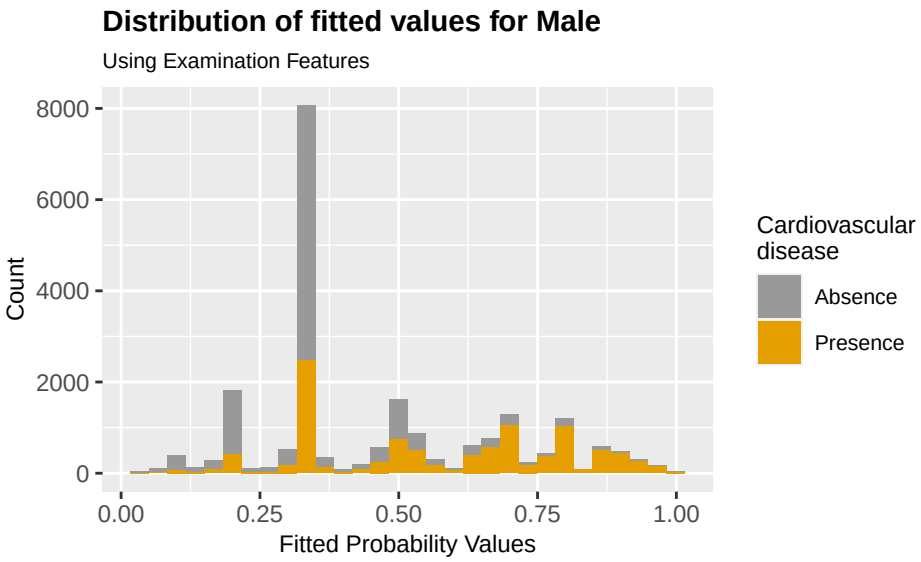


Model building using examination features

```
##
## Call:
## glm(formula = cardio ~ ap_hi + ap_lo + cholesterol + gluc, family = "quasibinomial",
##     data = examination_vars.male)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1479  -0.9229  -0.6240   0.9616   2.5003
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.027118   0.208265 -48.146  < 2e-16 ***
## ap_hi         0.066353   0.001716  38.676  < 2e-16 ***
## ap_lo         0.012046   0.002863   4.207  2.6e-05 ***
## cholesterol   0.576042   0.028524  20.195  < 2e-16 ***
## gluc         -0.108113   0.032213  -3.356 0.000792 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.063727)
##
##     Null deviance: 30416  on 22000  degrees of freedom
## Residual deviance: 25890  on 21996  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

```
## [1] 70.71497
```

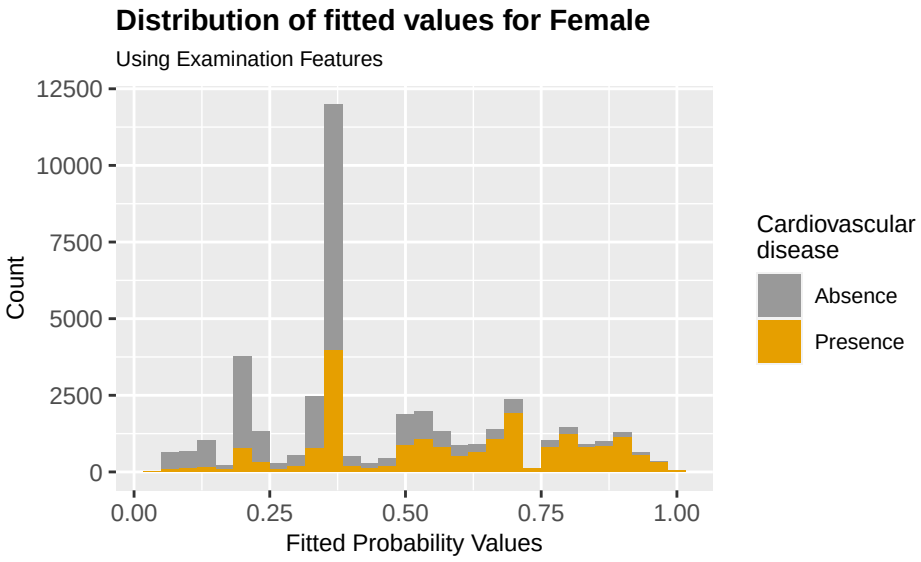



```
##
## Call:
## glm(formula = cardio ~ ap_hi + ap_lo + cholesterol + gluc, family = "quasibinomial",
##     data = examination_vars.female)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.1355  -0.9532  -0.5038   0.9302   2.6375
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.023969   0.142657 -70.266  < 2e-16 ***
## ap_hi         0.065343   0.001255  52.046  < 2e-16 ***
## ap_lo         0.013990   0.002068   6.765 1.35e-11 ***
## cholesterol   0.583734   0.020140  28.985  < 2e-16 ***
## gluc         -0.073360   0.022922  -3.200  0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.062683)
##
##     Null deviance: 57834  on 41832  degrees of freedom
## Residual deviance: 48251  on 41828  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

```
## [1] 71.46033
```



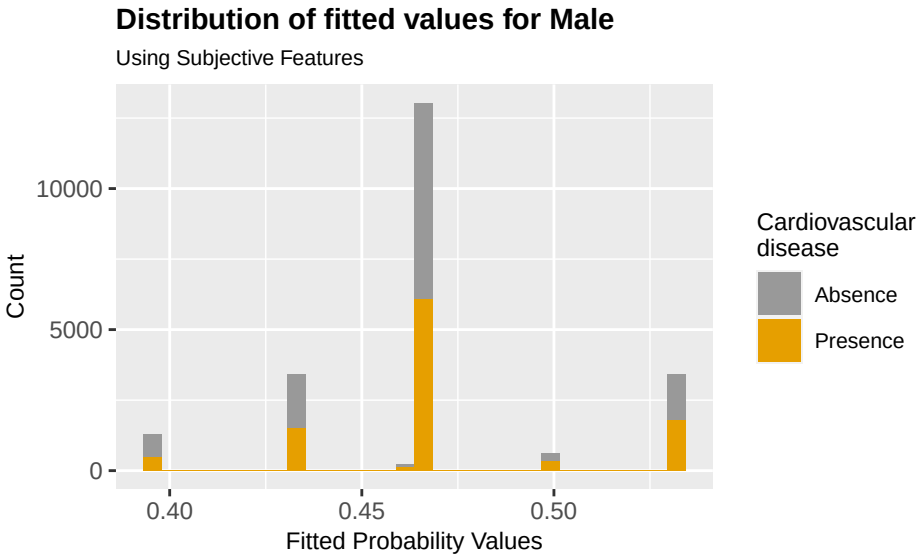
Model building using subjective features

```
##
## Call:
## glm(formula = cardio ~ alco + smoke + active, family = "binomial",
##     data = subjective_vars.male)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.236  -1.122  -1.007   1.234   1.358
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  0.13740    0.03125   4.397 1.10e-05 ***
## alco        -0.14312    0.04831  -2.963  0.00305 **
## smoke       -0.14027    0.03539  -3.964 7.37e-05 ***
## active      -0.26873    0.03423  -7.851 4.14e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30416  on 22000  degrees of freedom
## Residual deviance: 30310  on 21997  degrees of freedom
## AIC: 30318
##
## Number of Fisher Scoring iterations: 3
```

```
## [1] "Accuracy"
```

```
## [1] 53.95664
```

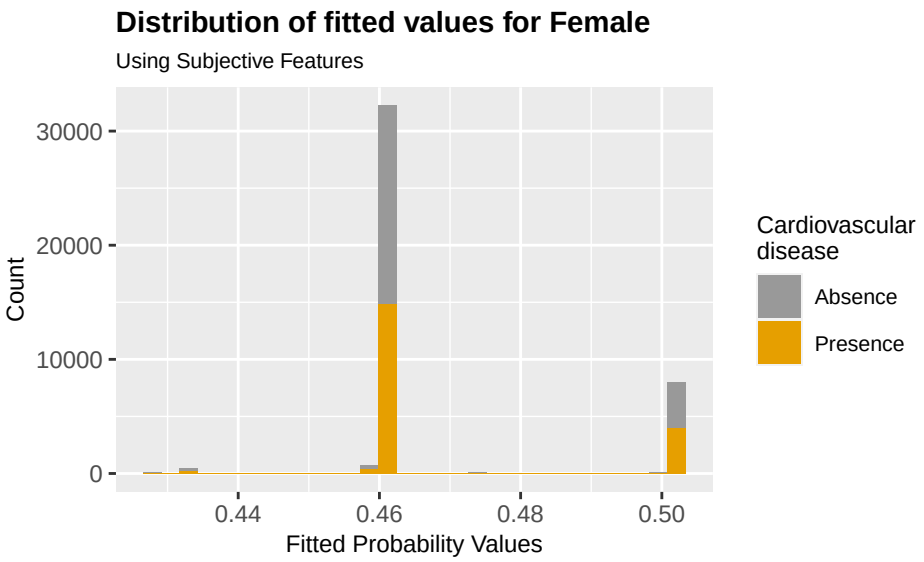



```
##
## Call:
## glm(formula = cardio ~ alco + smoke + active, family = "binomial",
##     data = subjective_vars.female)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.183  -1.113  -1.113   1.244   1.301
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.01283    0.02204   0.582    0.560
## alco        -0.01306    0.06402  -0.204    0.838
## smoke       -0.11908    0.07587  -1.570    0.117
## active      -0.16697    0.02455  -6.800 1.04e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 57834  on 41832  degrees of freedom
## Residual deviance: 57785  on 41829  degrees of freedom
## AIC: 57793
##
## Number of Fisher Scoring iterations: 3
```

```
## [1] "Accuracy"
```

```
## [1] 53.21397
```



Testing Models

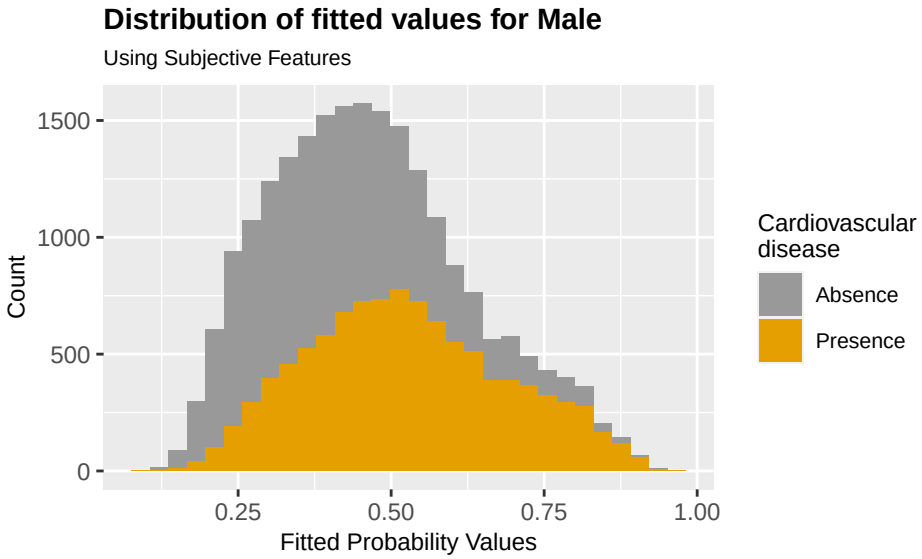
Testing with all the variables:

```
##
## Call:
## glm(formula = cardio ~ alco + smoke + active, family = "binomial",
##     data = subjective_vars.male)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.236  -1.122  -1.007   1.234   1.358
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.13740    0.03125   4.397 1.10e-05 ***
## alco        -0.14312    0.04831  -2.963  0.00305 **
## smoke       -0.14027    0.03539  -3.964 7.37e-05 ***
## active      -0.26873    0.03423  -7.851 4.14e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30416  on 22000  degrees of freedom
## Residual deviance: 30310  on 21997  degrees of freedom
```

```
## AIC: 30318
##
## Number of Fisher Scoring iterations: 3
```

```
## [1] "Accuracy"
```

```
## [1] 64.05618
```

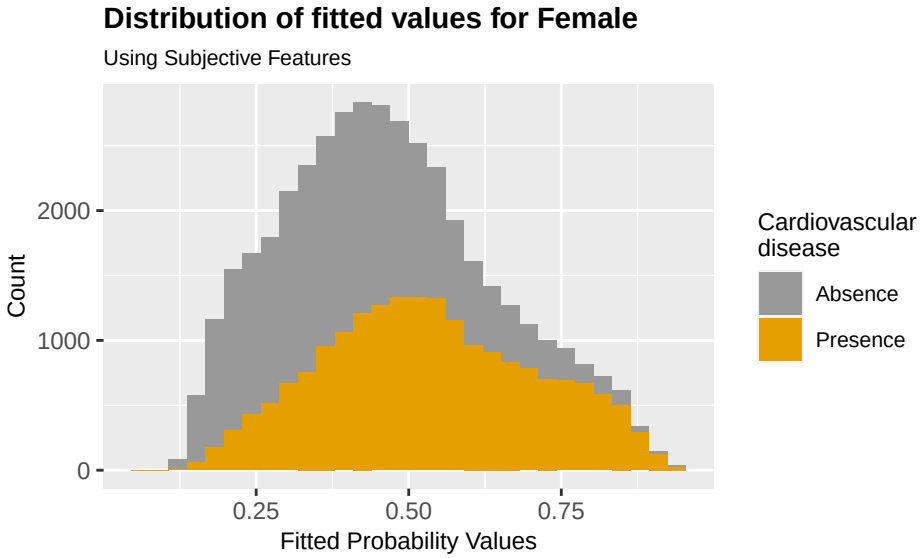



```
##
## Call:
## glm(formula = cardio ~ alco + smoke + active + age + weight +
##     height + cholesterol + gluc, family = "binomial", data = all_vars.female)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4651  -1.0260  -0.6468   1.1115   2.0931
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.0656464  0.2660565 -19.040  < 2e-16 ***
## alco        -0.1583253  0.0691854  -2.288   0.0221 *
## smoke       -0.1405657  0.0819779  -1.715   0.0864 .
## active      -0.1697179  0.0262028  -6.477 9.35e-11 ***
## age          0.0749223  0.0016540  45.296  < 2e-16 ***
## weight       0.0220994  0.0008064  27.404  < 2e-16 ***
## height      -0.0074592  0.0015268  -4.886 1.03e-06 ***
## cholesterol  0.6255468  0.0185805  33.667  < 2e-16 ***
## gluc        -0.0982202  0.0211372  -4.647 3.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 57834  on 41832  degrees of freedom
## Residual deviance: 52358  on 41824  degrees of freedom
## AIC: 52376
##
```

```
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

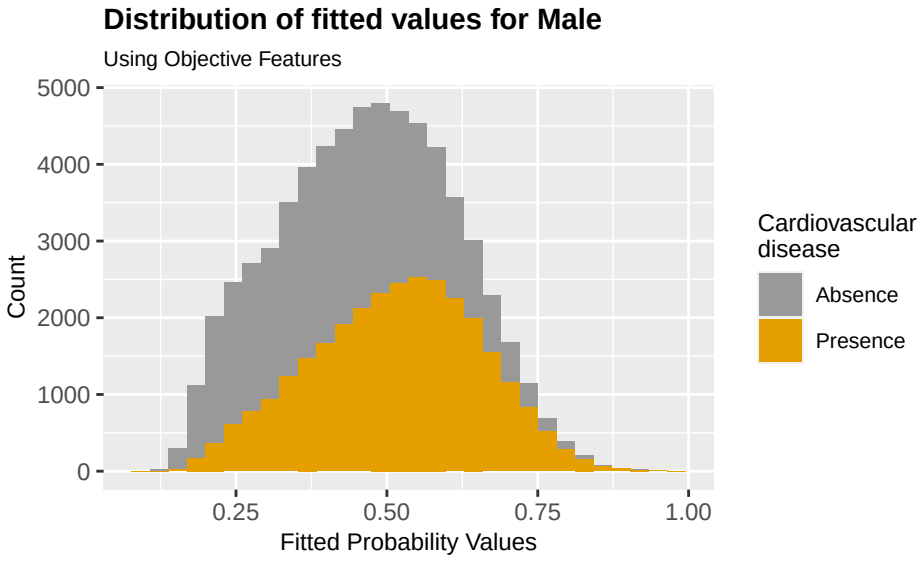
```
## [1] 64.89853
```




```
##
## Call:
## glm(formula = cardio ~ weight + gender + height + age, family = "binomial",
##     data = objective_vars.male)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5476  -1.0726  -0.6931   1.1192   1.9893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.039220   0.208506 -19.372   <2e-16 ***
## weight       0.026567   0.000653  40.683   <2e-16 ***
## genderMale   0.007746   0.020076   0.386      0.7
## height      -0.012104   0.001232  -9.824   <2e-16 ***
## age          0.074653   0.001281  58.285   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88250  on 63833  degrees of freedom
## Residual deviance: 82512  on 63829  degrees of freedom
## AIC: 82522
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

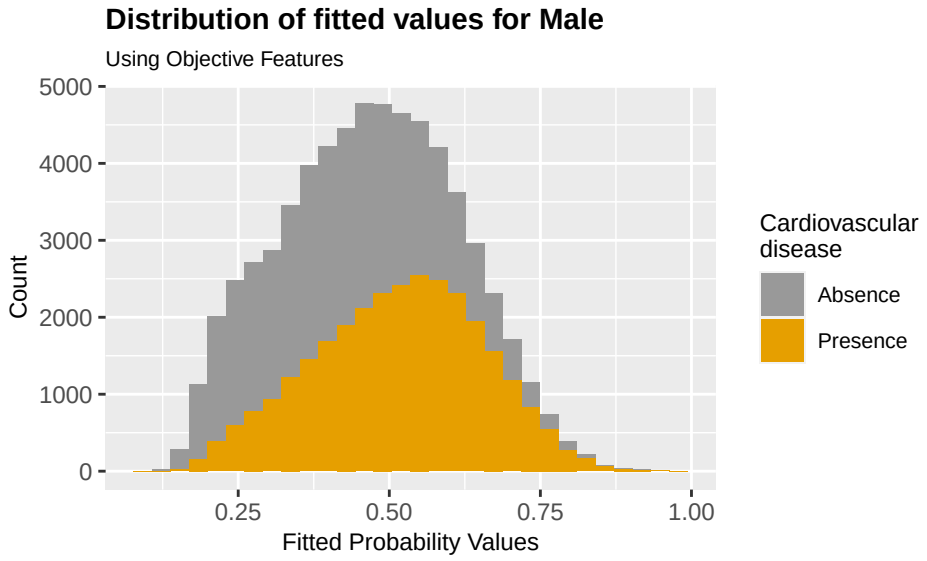
```
## [1] 62.55287
```




```
##
## Call:
## glm(formula = cardio ~ weight * gender + height + age, family = "binomial",
##     data = objective_vars.male)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6402  -1.0737  -0.6926   1.1181   2.0223
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.8725803  0.2155656 -17.965  < 2e-16 ***
## weight             0.0252357  0.0007826  32.245  < 2e-16 ***
## genderMale        -0.2939802  0.1009338  -2.913  0.00358 **
## height            -0.0125679  0.0012422 -10.118  < 2e-16 ***
## age                0.0747297  0.0012813  58.322  < 2e-16 ***
## weight:genderMale  0.0040747  0.0013357   3.051  0.00228 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88250  on 63833  degrees of freedom
## Residual deviance: 82502  on 63828  degrees of freedom
## AIC: 82514
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

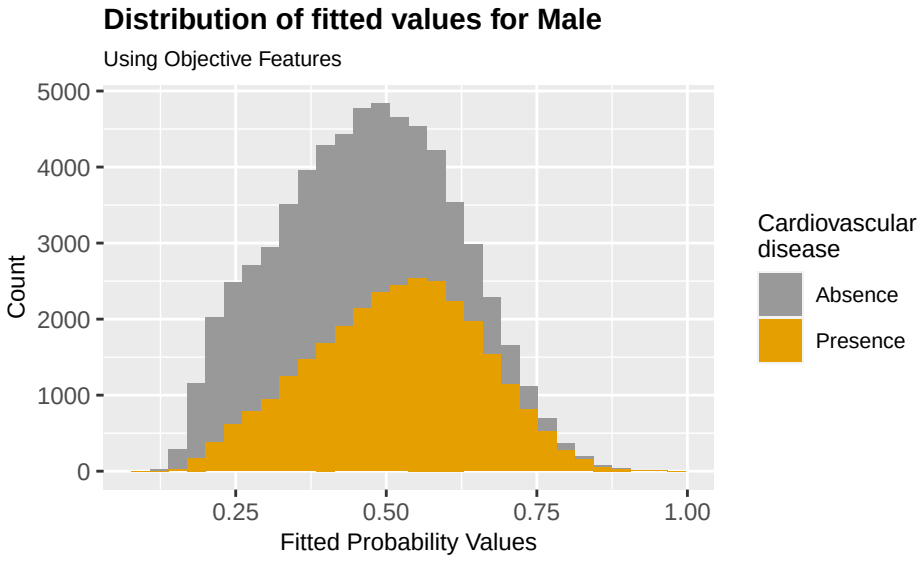
```
## [1] 62.60613
```




```
##
## Call:
## glm(formula = cardio ~ height * gender + weight + age, family = "binomial",
##     data = objective_vars.male)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.542  -1.073  -0.693   1.119   1.991
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.9460237  0.2518301 -15.669   <2e-16 ***
## height            -0.0126628  0.0014954  -8.468   <2e-16 ***
## genderMale        -0.2654889  0.4141730  -0.641    0.522
## weight             0.0265172  0.0006573  40.341   <2e-16 ***
## age                0.0746653  0.0012810  58.287   <2e-16 ***
## height:genderMale  0.0016373  0.0024790   0.660    0.509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88250  on 63833  degrees of freedom
## Residual deviance: 82511  on 63828  degrees of freedom
## AIC: 82523
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

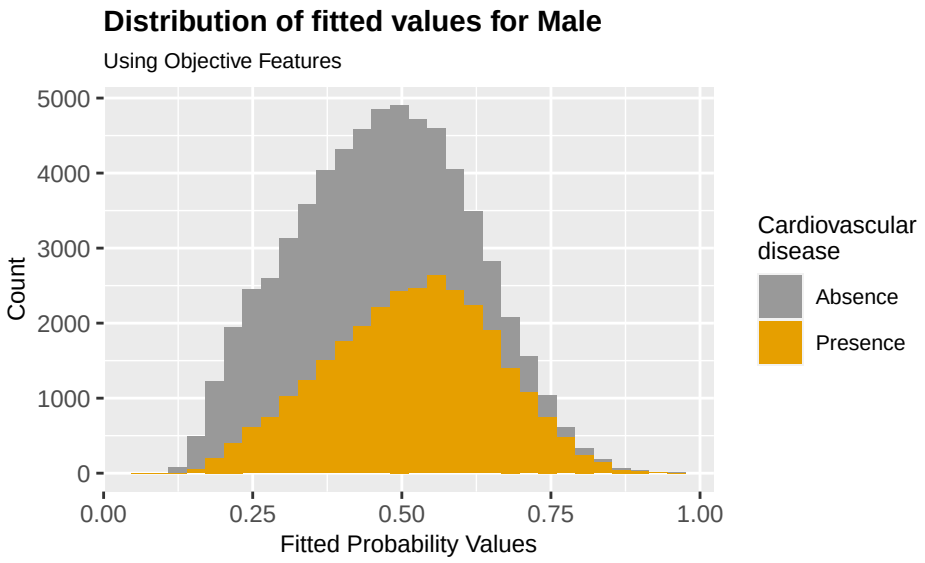
```
## [1] 62.57324
```




```
##
## Call:
## glm(formula = cardio ~ age * gender + height + weight, family = "binomial",
##     data = objective_vars.male)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5001  -1.0731  -0.6894   1.1189   1.9891
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.3613917  0.2147229 -20.312  < 2e-16 ***
## age             0.0809609  0.0016257  49.800  < 2e-16 ***
## genderMale      0.9029735  0.1409001   6.409 1.47e-10 ***
## height         -0.0121520  0.0012327  -9.858  < 2e-16 ***
## weight          0.0264881  0.0006535  40.531  < 2e-16 ***
## age:genderMale -0.0168934  0.0026323  -6.418 1.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88250  on 63833  degrees of freedom
## Residual deviance: 82471  on 63828  degrees of freedom
## AIC: 82483
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Accuracy"
```

```
## [1] 62.57324
```


Distribution of fitted values for Male
Using Objective Features
5000
4000
3000
2000
1000
0
Count
0.00
0.25
0.50
0.75
1.00
Fitted Probability Values
Cardiovascular disease
Absence
Presence