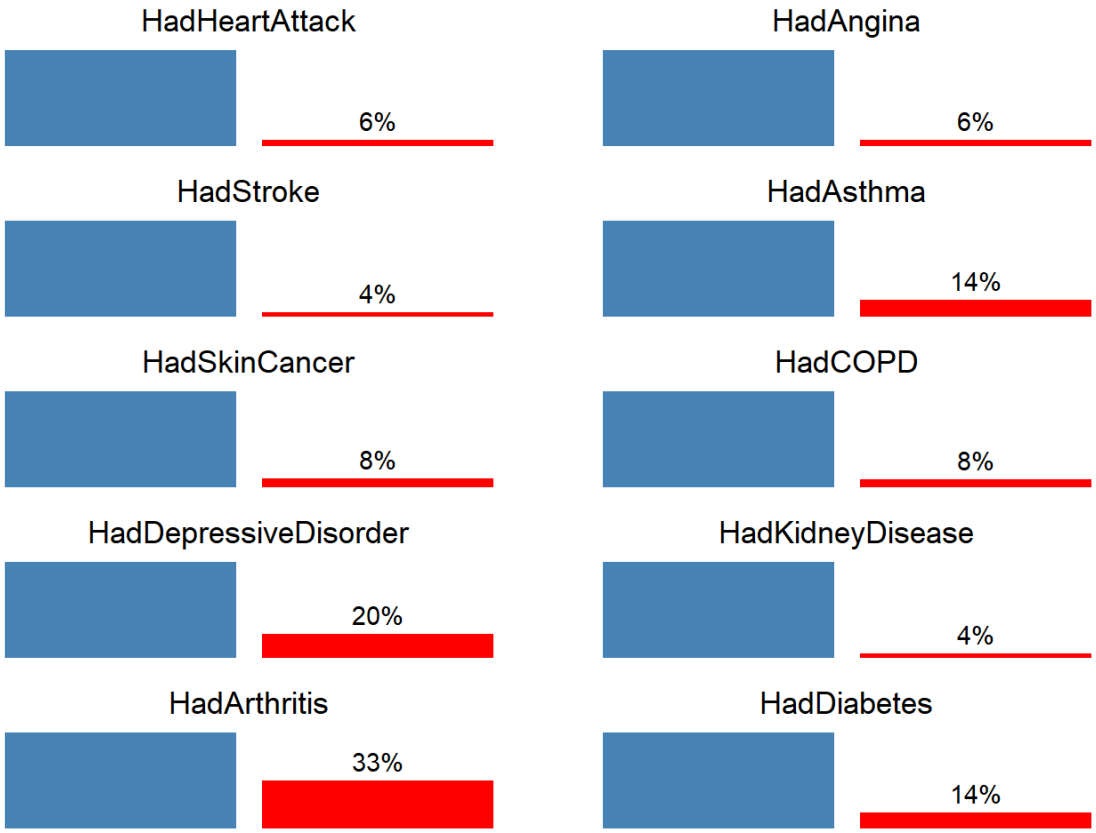
We want to see if there is any relationship between different ethnicities and the probability of suffering from various diseases collected in the dataset. To do this, we plan to use some clustering models ignoring the race label and then compare the clustering results with it.

As clustering algorithms are relatively expensive, and to avoid going completely blind, we decided to make a first pass using **linear regression**. We left that variable in categorical mode so that the algorithm automatically creates dummy variables for each category.

Here came our first problem: to **avoid multicollinearity**, R decides to ignore the first level, as it can be explained linearly by the absence of the rest. We chose to keep the NA values in that column but label them to be their first level, thus being ignored by the “glm” algorithm.

We reviewed the **balance** of all diseases and the probability of suffering from them. In general, it is **around 4-8%**, and in the most extreme case (**HadArthritis**), it is **33%**. Since balancing will get us relatively small samples, we will continue as is, although we will create a balanced model for that disease and compare both results.



Some diseases may be directly related. For example, heart attack, angina, and stroke are circulatory problems, while asthma and COPD are respiratory/lung related. Therefore, we will include only one disease in the dependent variable, and the rest of the columns in the independent variable.

Once we have the results, focusing on the **p-value** and the **estimate**, we can draw several **conclusions**. We summarize them below, you can find more details in the corresponding .Rmd and .html files.

```
[1] "GLM for HadArthritis unbalanced ( 12 )"

Call:
glm(formula = formula, family = binomial, data = df)

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
GeneralHealth                                    -0.1903273  0.0046738 -40.722  < 2e-16 ***
HadDepressiveDisorder                             0.4516615  0.0107875  41.869  < 2e-16 ***
DifficultyWalking                                 0.7380827  0.0124843  59.121  < 2e-16 ***
AgeCategory                                       0.2453016  0.0016982 144.447  < 2e-16 ***
Gender                                           -0.3979480  0.0108109 -36.810  < 2e-16 ***
PhysicalHealthDays                                0.0178631  0.0005239  34.094  < 2e-16 ***
HadAsthma                                         0.3456244  0.0110825  31.187  < 2e-16 ***
LastCheckupTime                                   0.1418907  0.0047679  29.760  < 2e-16 ***
(Intercept)                                      -3.8579970  0.1337599 -28.843  < 2e-16 ***
SleepHours                                       -0.0726057  0.0026032 -27.891  < 2e-16 ***
PneumoVaxEver                                     0.2370474  0.0089789  26.400  < 2e-16 ***
RemovedTeeth                                      0.1945269  0.0082492  23.581  < 2e-16 ***
ChestScan                                         0.1917944  0.0083786  22.891  < 2e-16 ***
SmokerStatus                                      0.1484133  0.0083060  17.868  < 2e-16 ***
HIVTesting                                        0.1584012  0.0092735  17.081  < 2e-16 ***
HadSkinCancer                                     0.2028097  0.0130166  15.581  < 2e-16 ***
DifficultyConcentrating                           0.1884455  0.0139657  13.493  < 2e-16 ***
DeafOrHardOfHearing                               0.1615307  0.0130313  12.396  < 2e-16 ***
PhysicalActivities                                0.1161866  0.0095789  12.129  < 2e-16 ***
HadKidneyDisease                                  0.2113426  0.0178806  11.820  < 2e-16 ***
BMI                                               0.0229594  0.0019932  11.519  < 2e-16 ***
CovidPos                                          0.1013863  0.0089643  11.310  < 2e-16 ***
AlcoholDrinkers                                   0.0917674  0.0082728  11.093  < 2e-16 ***
WeightInKilograms                                 0.0069588  0.0006459  10.773  < 2e-16 ***
HadCOPD                                           0.1400427  0.0145011   9.657  < 2e-16 ***
FluVaxLast12                                      0.0819770  0.0086118   9.519  < 2e-16 ***
RaceEthnicityCategoryHispanic                    -0.2462517  0.0261286  -9.425  < 2e-16 ***
RaceEthnicityCategoryWhite only, Non-Hispanic     0.1953790  0.0218216   8.953  < 2e-16 ***
RaceEthnicityCategoryMultiracial, Non-Hispanic    0.2675031  0.0346993   7.709 1.27e-14 ***
DifficultyErrands                                -0.1075132  0.0175578  -6.123 9.16e-10 ***
HadAngina                                         0.0990223  0.0168418   5.880 4.11e-09 ***
MentalHealthDays                                  0.0030059  0.0005602   5.366 8.04e-08 ***
State                                             0.0011400  0.0002452   4.650 3.32e-06 ***
HeightInMeters                                   -0.3236017  0.0775169  -4.175 2.99e-05 ***
DifficultyDressingBathing                         0.0833145  0.0233335   3.571 0.000356 ***
TetanusLast10Tdap                                -0.0404826  0.0113689  -3.561 0.000370 ***
HadStroke                                        -0.0617222  0.0179256  -3.443 0.000575 ***
HadHeartAttack                                   -0.0483571  0.0171411  -2.821 0.004786 **
RaceEthnicityCategoryBlack only, Non-Hispanic     0.0676925  0.0255319   2.651 0.008019 **
HadDiabetes                                      -0.0278666  0.0109342  -2.549 0.010817 *
HighRiskLastYear                                 -0.0599146  0.0238690  -2.510 0.012068 *
RaceEthnicityCategoryOther race only, Non-Hispanic -0.0701777  0.0289765  -2.422 0.015440 *
BlindOrVisionDifficulty                           0.0332361  0.0173576   1.915 0.055519 .
ECigaretteUsage                                  -0.0107229  0.0098422  -1.089 0.275942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 542171  on 426042  degrees of freedom
Residual deviance: 418308  on 425999  degrees of freedom
AIC: 418396

Number of Fisher Scoring iterations: 5
```

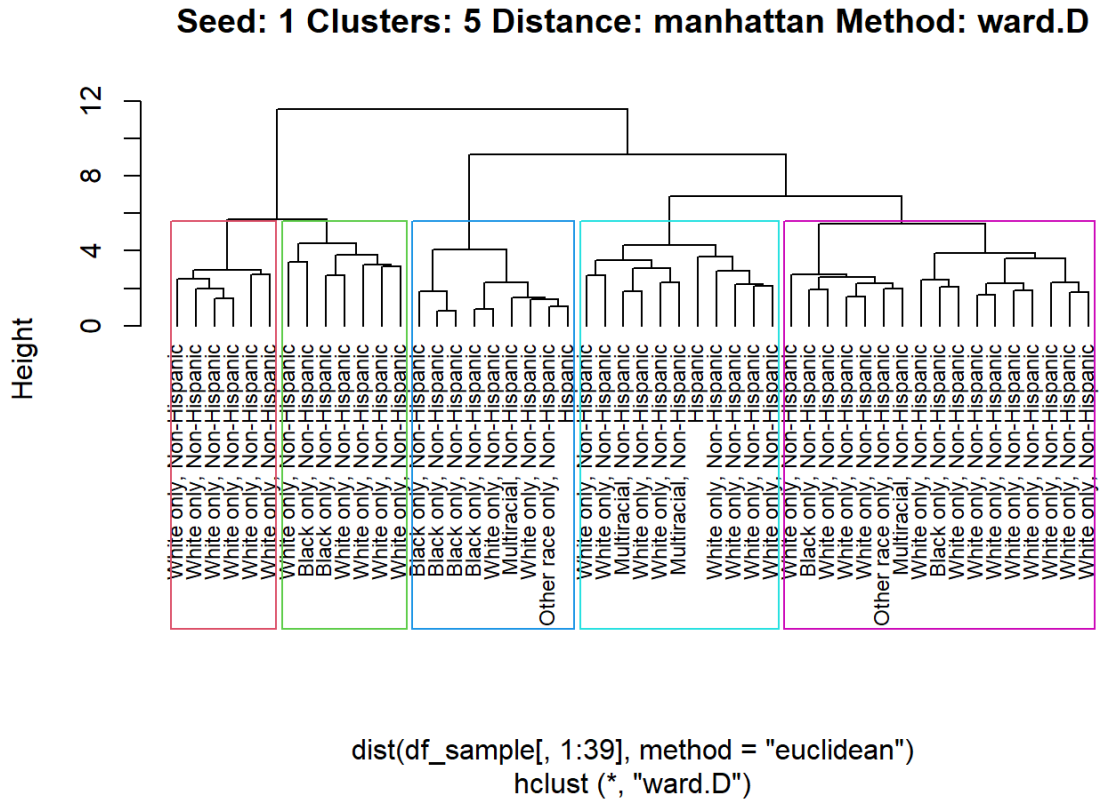
```
[1] "GLM for HadArthritis balanced ( 11 )"

Call:
glm(formula = HadArthritis ~ ., family = binomial, data = df_balanced)

Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
DifficultyWalking                                0.7336674  0.0155397  47.213  < 2e-16 ***
AgeCategory                                      0.2543665  0.0019608 129.725  < 2e-16 ***
GeneralHealth                                   -0.1988636  0.0055041 -36.130  < 2e-16 ***
HadDepressiveDisorder                            0.4438486  0.0128289  34.598  < 2e-16 ***
Gender                                          -0.3809506  0.0126232 -30.179  < 2e-16 ***
PhysicalHealthDays                               0.0191187  0.0006450  29.640  < 2e-16 ***
HadAsthma                                        0.3458622  0.0132257  26.151  < 2e-16 ***
LastCheckupTime                                  0.1400611  0.0053964  25.955  < 2e-16 ***
SleepHours                                      -0.0714054  0.0030996 -23.037  < 2e-16 ***
PneumoVaxEver                                    0.2259960  0.0106361  21.248  < 2e-16 ***
(Intercept)                                     -3.2365891  0.1569929 -20.616  < 2e-16 ***
RemovedTeeth                                     0.1923713  0.0096497  19.935  < 2e-16 ***
ChestScan                                        0.1952010  0.0098684  19.780  < 2e-16 ***
SmokerStatus                                     0.1694579  0.0097951  17.300  < 2e-16 ***
HIVTesting                                       0.1657986  0.0108261  15.315  < 2e-16 ***
HadSkinCancer                                    0.2114828  0.0159254  13.280  < 2e-16 ***
DifficultyConcentrating                          0.1991530  0.0168471  11.821  < 2e-16 ***
DeafOrHardOfHearing                              0.1824431  0.0160532  11.365  < 2e-16 ***
PhysicalActivities                               0.1089865  0.0114002   9.560  < 2e-16 ***
AlcoholDrinkers                                  0.0922440  0.0097259   9.484  < 2e-16 ***
CovidPos                                         0.0984305  0.0104868   9.386  < 2e-16 ***
HadCOPD                                          0.1649975  0.0180135   9.160  < 2e-16 ***
BMI                                              0.0217044  0.0023731   9.146  < 2e-16 ***
WeightInKilograms                                0.0069667  0.0007661   9.094  < 2e-16 ***
HadKidneyDisease                                 0.1985421  0.0221779   8.952  < 2e-16 ***
RaceEthnicityCategoryHispanic                   -0.2558798  0.0302612  -8.456  < 2e-16 ***
RaceEthnicityCategoryWhite only, Non-Hispanic    0.1986944  0.0255158   7.787 6.86e-15 ***
FluVaxLast12                                     0.0782034  0.0100719   7.764 8.20e-15 ***
RaceEthnicityCategoryMultiracial, Non-Hispanic   0.2728735  0.0405743   6.725 1.75e-11 ***
MentalHealthDays                                 0.0037373  0.0006679   5.595 2.20e-08 ***
HadAngina                                        0.1120495  0.0209820   5.340 9.28e-08 ***
DifficultyErrands                               -0.0989162  0.0216410  -4.571 4.86e-06 ***
State                                            0.0010355  0.0002877   3.599 0.000320 ***
HeightInMeters                                  -0.3016369  0.0909817  -3.315 0.000915 ***
TetanusLast10Tdap                               -0.0404774  0.0133594  -3.030 0.002446 **
HadStroke                                       -0.0607631  0.0221552  -2.743 0.006095 **
RaceEthnicityCategoryOther race only, Non-Hispanic -0.0881426  0.0334689  -2.634 0.008449 **
DifficultyDressingBathing                        0.0740045  0.0294897   2.510 0.012090 *
HadHeartAttack                                  -0.0498154  0.0211353  -2.357 0.018425 *
BlindOrVisionDifficulty                          0.0395983  0.0212538   1.863 0.062446 .
HighRiskLastYear                                -0.0405738  0.0270361  -1.501 0.133427
RaceEthnicityCategoryBlack only, Non-Hispanic    0.0399967  0.0298394   1.340 0.180114
HadDiabetes                                     -0.0104482  0.0132150  -0.791 0.429160
ECigaretteUsage                                 -0.0075014  0.0115366  -0.650 0.515544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 393361  on 283749  degrees of freedom
Residual deviance: 299512  on 283706  degrees of freedom
AIC: 299600

Number of Fisher Scoring iterations: 4
```

Different ethnicities are not in the top half of the list of variables that predict different diseases. General health, age, and gender are usually more important (among the top 5 or 10 in most cases). But in almost all cases, we see some kind of genetic influence. This is especially relevant for skin cancer, where 4 of the 5 ethnic groups are among the top 15 relevant variables out of 44 total. In the rest of the models, the first one rarely appears before that same number (15).

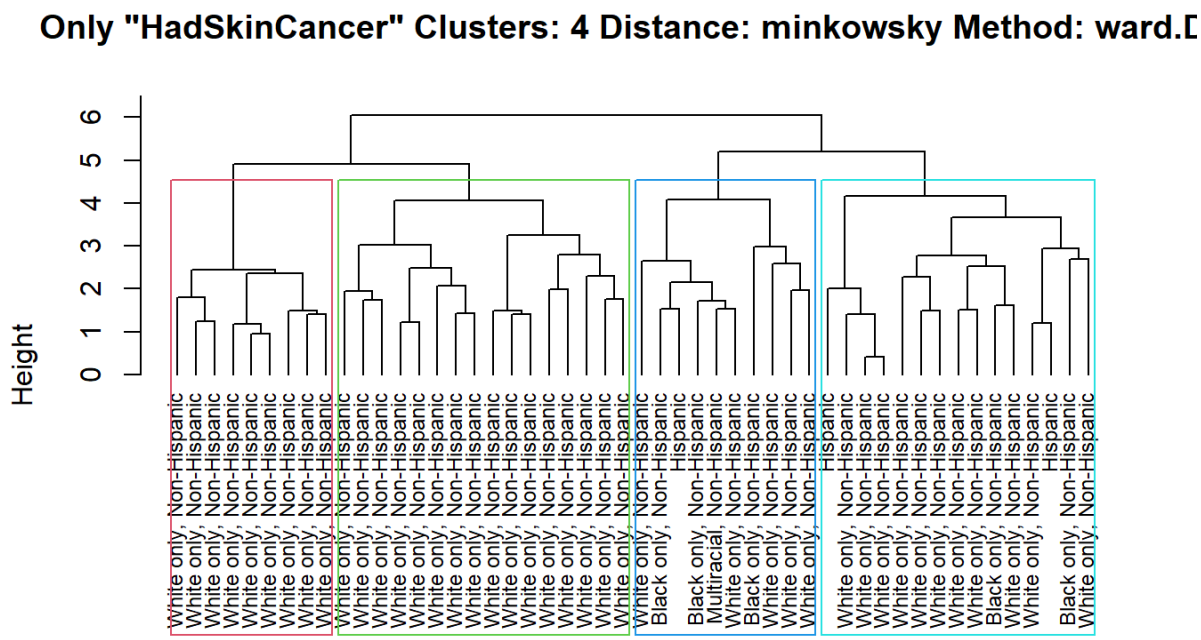
```
[1] "GLM for HadSkinCancer ( 5 )"

Call:
glm(formula = formula, family = binomial, data = df)

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
AgeCategory                                      0.2900259  0.0031455  92.204  < 2e-16 ***
(Intercept)                                     -8.4424325  0.2279626 -37.034  < 2e-16 ***
RaceEthnicityCategoryBlack only, Non-Hispanic   -2.1555945  0.0738637 -29.183  < 2e-16 ***
HadArthritis                                     0.2454783  0.0129034  19.024  < 2e-16 ***
PneumoVaxEver                                    0.2627184  0.0141923  18.511  < 2e-16 ***
RaceEthnicityCategoryHispanic                   -0.9782193  0.0556312 -17.584  < 2e-16 ***
AlcoholDrinkers                                  0.2232837  0.0128865  17.327  < 2e-16 ***
RaceEthnicityCategoryOther race only, Non-Hispanic -1.1170087  0.0669225 -16.691  < 2e-16 ***
PhysicalActivities                               0.2437231  0.0154365  15.789  < 2e-16 ***
FluVaxLast12                                     0.1866596  0.0141900  13.154  < 2e-16 ***
HadKidneyDisease                                 0.2914119  0.0226742  12.852  < 2e-16 ***
HeightInMeters                                   1.5155526  0.1302486  11.636  < 2e-16 ***
LastCheckupTime                                  0.0968944  0.0085571  11.323  < 2e-16 ***
ChestScan                                        0.1389014  0.0130143  10.673  < 2e-16 ***
RaceEthnicityCategoryWhite only, Non-Hispanic    0.3673576  0.0355130  10.344  < 2e-16 ***
State                                           -0.0036461  0.0003795  -9.607  < 2e-16 ***
RemovedTeeth                                    -0.1200012  0.0129128  -9.293  < 2e-16 ***
CovidPos                                         0.1242243  0.0141489   8.780  < 2e-16 ***
PhysicalHealthDays                               0.0056692  0.0007851   7.221 5.15e-13 ***
HadAngina                                        0.1516344  0.0213099   7.116 1.11e-12 ***
HadAsthma                                        0.1084288  0.0175752   6.169 6.85e-10 ***
DeafOrHardOfHearing                              0.1023988  0.0172330   5.942 2.82e-09 ***
HIVTesting                                       0.0912601  0.0154930   5.890 3.85e-09 ***
HadDepressiveDisorder                            0.1001597  0.0172949   5.791 6.99e-09 ***
DifficultyWalking                               -0.1021287  0.0184294  -5.542 3.00e-08 ***
ECigaretteUsage                                  0.0820903  0.0164332   4.995 5.87e-07 ***
RaceEthnicityCategoryMultiracial, Non-Hispanic  -0.2993421  0.0676965  -4.422 9.79e-06 ***
HadDiabetes                                     -0.0730963  0.0168033  -4.350 1.36e-05 ***
DifficultyErrands                               -0.1132567  0.0268281  -4.222 2.43e-05 ***
Gender                                          -0.0716392  0.0173943  -4.119 3.81e-05 ***
WeightInKilograms                               -0.0046743  0.0011444  -4.085 4.41e-05 ***
SmokerStatus                                    -0.0455828  0.0128156  -3.557 0.000375 ***
HadStroke                                        0.0797038  0.0243147   3.278 0.001045 **
GeneralHealth                                   -0.0233192  0.0074226  -3.142 0.001680 **
HadHeartAttack                                  -0.0618357  0.0231409  -2.672 0.007537 **
MentalHealthDays                                 0.0022439  0.0009255   2.425 0.015329 *
DifficultyDressingBathing                        0.0768195  0.0337948   2.273 0.023019 *
DifficultyConcentrating                         -0.0435165  0.0227850  -1.910 0.056149 .
BMI                                             -0.0053036  0.0035504  -1.494 0.135223
TetanusLast10Tdap                                0.0218751  0.0179329   1.220 0.222529
HighRiskLastYear                                -0.0568987  0.0502703  -1.132 0.257696
HadCOPD                                          0.0199068  0.0203619   0.978 0.328248
SleepHours                                       0.0022989  0.0041367   0.556 0.578394
BlindOrVisionDifficulty                          0.0026884  0.0263754   0.102 0.918815
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 239910  on 426042  degrees of freedom
Residual deviance: 200330  on 425999  degrees of freedom
AIC: 200418

Number of Fisher Scoring iterations: 7
```

Knowing that there is a certain relationship between both data, we will start the clustering process using a hierarchical model (**hclust**). The first steps are to eliminate the “NA’s” values that the race variable still contains. Since they are only 3%, they are directly removed from the dataset. Next, we normalize the data, thus avoiding variables with high values (height, weight, etc.) producing biases during the analysis.

The dataset includes more than 400,000 rows, making it unfeasible for direct study and processing. Therefore, we will rely on the dendrogram comparing the separation into clusters with the ethnicity label. We will make several for loops trying different random seeds with small samples with **all distance and agglomeration methods**.

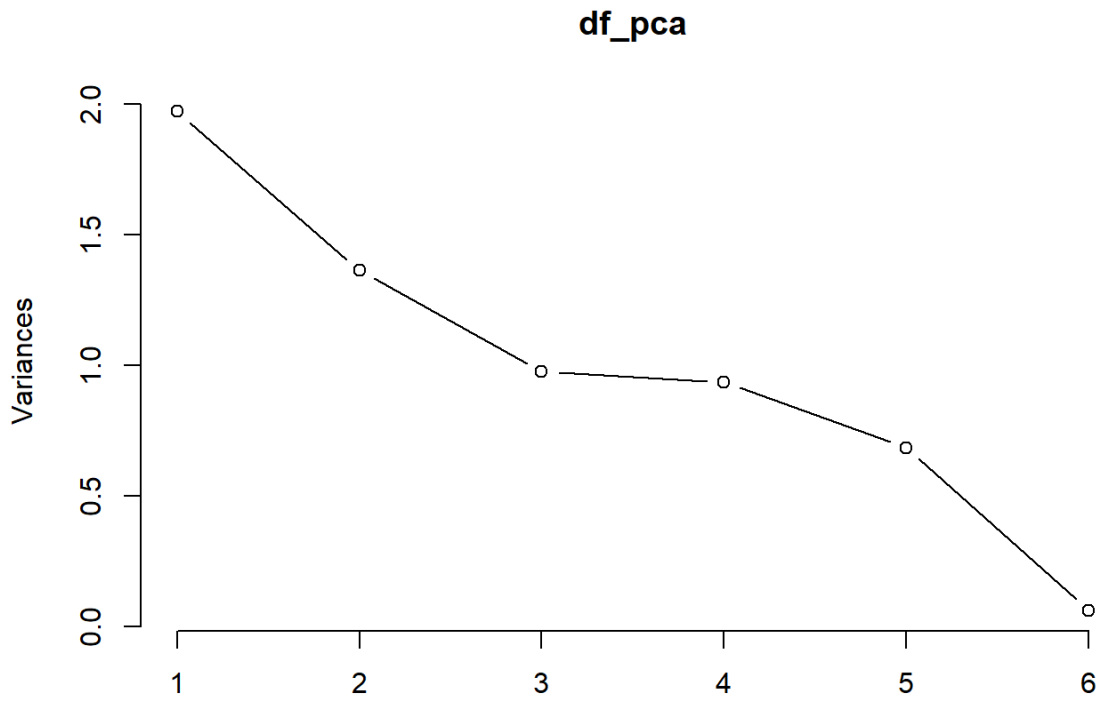


For each sample, we will keep the number of **different ethnicities** in it and use it as the **k value** when cutting the dendrogram. We have also tried to reduce the number of columns to consider, using only skin cancer as the disease, as we already saw that it is the one in which races have the greatest influence.



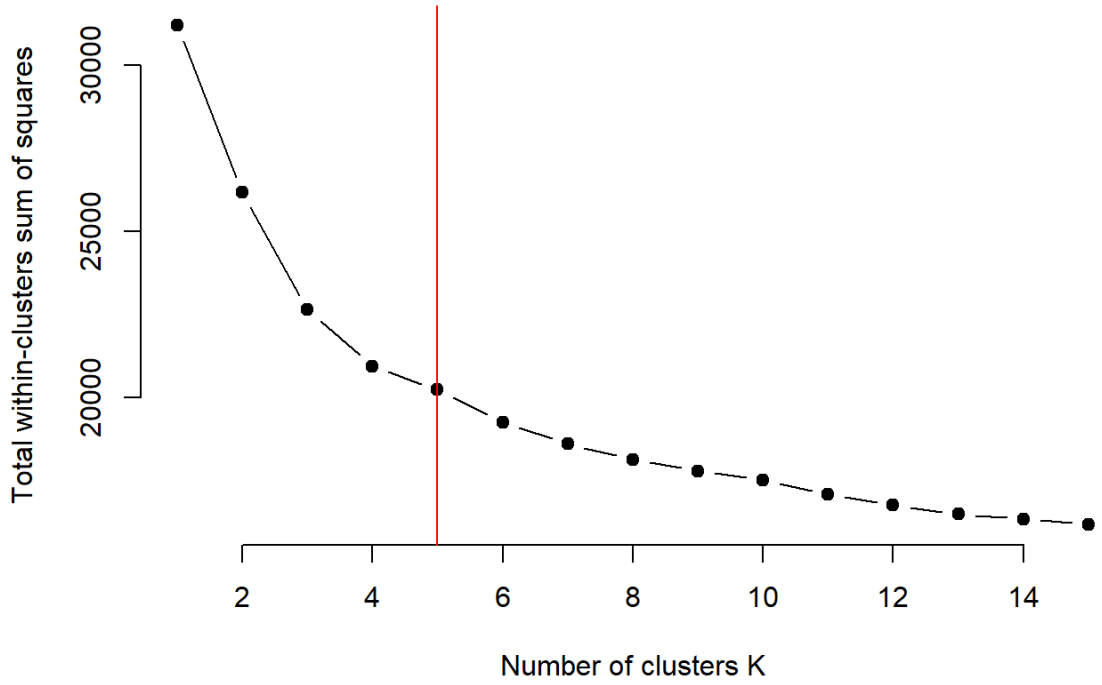
In no case were acceptable, much less satisfactory, groupings obtained. It is true that when working with sub-samples, we may have simply had bad luck with the pseudo-random number generator.

Finally, we also tried **K-means**. We started with a **PCA** analysis of the “real” numerical variables to try to reduce them, which are: “PhysicalHealthDays”, “MentalHealthDays”, “SleepHours”, “HeightInMeters”, “WeightInKilograms”, and “BMI”. Thanks to the **elbow plot**, we saw that we could reduce them to only 3 components. Another possible approach would have been to directly eliminate height and weight, as BMI is obtained using a formula using both.



Having already gathered enough evidence that there was no strong relationship between races and diseases, we decided to drastically reduce the dataset to only one disease, skin cancer, and reduce the number of rows to 4,000 records to avoid the relatively expensive K-means algorithm. Even so, our computer took almost an hour to calculate the Hopkins statistic.

In this reduced sample, we obtained an **H value of approximately 0.7**, indicating that the dataset was clusterable. We counted the number of unique races in the sample and used another elbow graph to determine the optimal number of clusters, obtaining a promising result: 5 is the number of ethnicities and one of the points where the curve began to “flatten.”



Finally, we labeled each row according to the cluster in which it was grouped, and using a matrix, we checked if they corresponded to the race labels, obtaining a result consistent with what we saw with the previous methods: despite being able to be grouped into 5 clusters, they do not match the race values.

```
##
##                                    1   2   3   4   5
##   Black only, Non-Hispanic        84  61  80  51  58
##   Hispanic                        78  88 115  49  56
##   Multiracial, Non-Hispanic       25  14  22  12  13
##   Other race only, Non-Hispanic   67  68  45  18  39
##   White only, Non-Hispanic       845 678 617 448 369
```

Since we have obtained compatible results using these approaches, we can conclude that although ethnicity can influence certain diseases, it is not a determining factor. Otherwise, we could have continued mining the data with other algorithms (PAM or DBSCAN) and even more methods (classification and association rules).