Lab3

Mishu Dhar

Loading Libraries

Part 1:

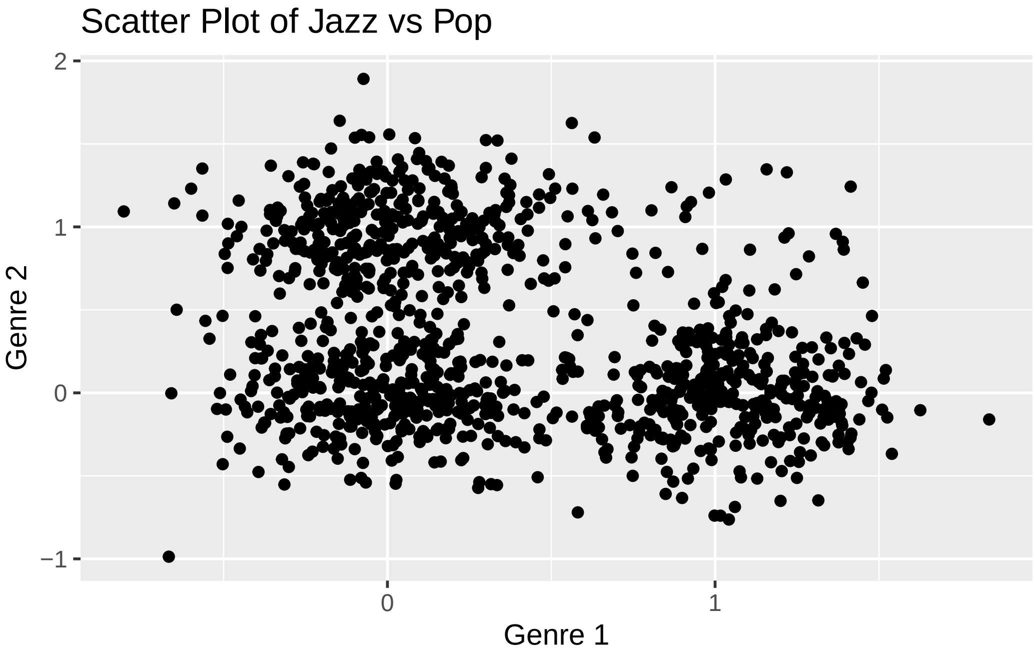
Test Clustering and Influence

Question 1

```
[1] 1075    4
```

There are four (4) columns and 1075 rows in this dataset.

Creating one scatter plot of two genres jazz and pop

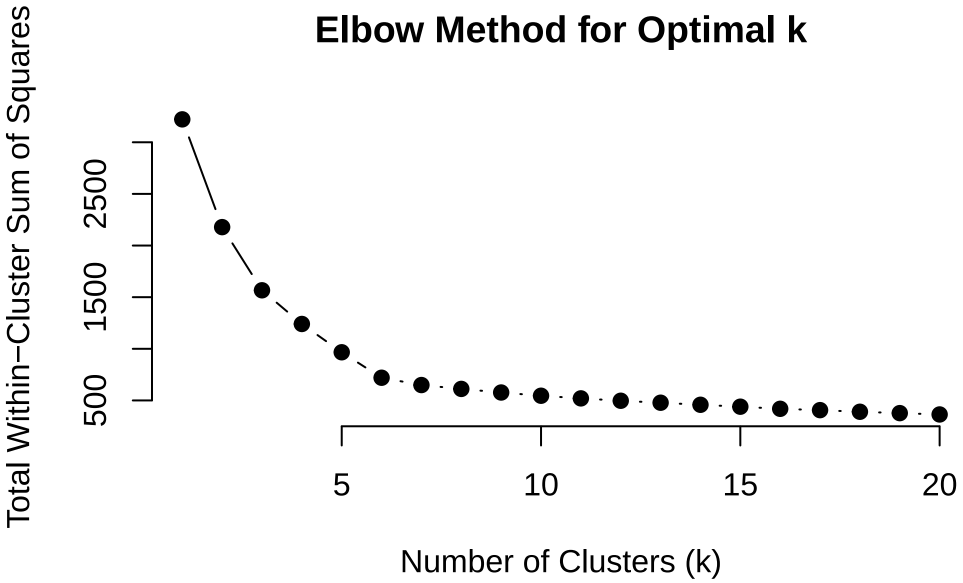


From the scatter plot, it seems that the data points form three distinct groups or clusters. One group in the bottom left, one in the top left, and one in the right side. These clusters suggest that individuals with similar music tastes, such as jazz and pop, tend to group together. The data does not appear randomly scattered but instead forms noticeable patterns, indicating that musical tastes may indeed influence how the data is clustered.

The plot suggests that the data is likely clustered based on musical preferences like jazz and pop.

Question 2:

Question 3: K means clustering



Based on the elbow plot, it looks like 3 is the best number of clusters. This is because the total within-cluster sum of squares (WSS) drops quickly when increasing from 1 to 3 clusters, but after 3, the decrease slows down a lot. This means adding more clusters doesn't really improve the clustering much. So, using 3 clusters is a good choice because it balances keeping the model simple while still capturing most of the variation in the data.

Question 4

```
        jazz        pop       hiphop
```

```
1 -0.6814994 -0.6397617  0.15705668
2  1.2351547 -0.5869824 -0.09087532
3 -0.5621258  1.2334688 -0.06585763
```

The three clusters we found show distinct music preferences. Cluster 1 includes individuals who have a low preference for both jazz and pop, but they seem to like hiphop a little more, though their tastes are generally neutral. Cluster 2 is made up of Jazz lovers, these individuals have a strong preference for jazz, while showing very little interest in pop or hiphop. Cluster 3 is mostly pop fans, as they show a high preference for pop but are not very interested in jazz or hiphop. Overall, these clusters are meaningfully distinct because each group has a clear difference in music tastes, with one group preferring hip-hop slightly, another focused on jazz, and the third leaning towards pop.

Question 5:

```
        jazz        pop      hiphop
1 -0.6634616  0.3705600  0.04963173
2  1.1107205 -0.6203654 -0.08308994
```

When we use k = 2, the clustering shows two main groups. The first group has a low preference for jazz, a moderate preference for pop, and neutral or slightly positive feelings about hiphop. This group seems to have more mixed or neutral music tastes, without strong preferences for any one genre. The second group is made up of people who really like jazz but don't care much for pop or hiphop. Compared to k = 3, where we had three distinct clusters (jazz lovers, pop lovers, and a neutral/hiphop group), using k = 2 combines some of these groups. This changes how we understand the population, as it simplifies the clusters into just two broader groups: jazz lovers and a mixed group with varied or neutral tastes. While k = 2 is simpler, it loses some of the finer details about music preferences that we saw with k = 3.

Question 6

```
Call:
lm(formula = influence ~ factor(cluster_k3) + 0, data = taste_influence)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9226 -0.8025 -0.0355  0.8575  4.0951 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
```

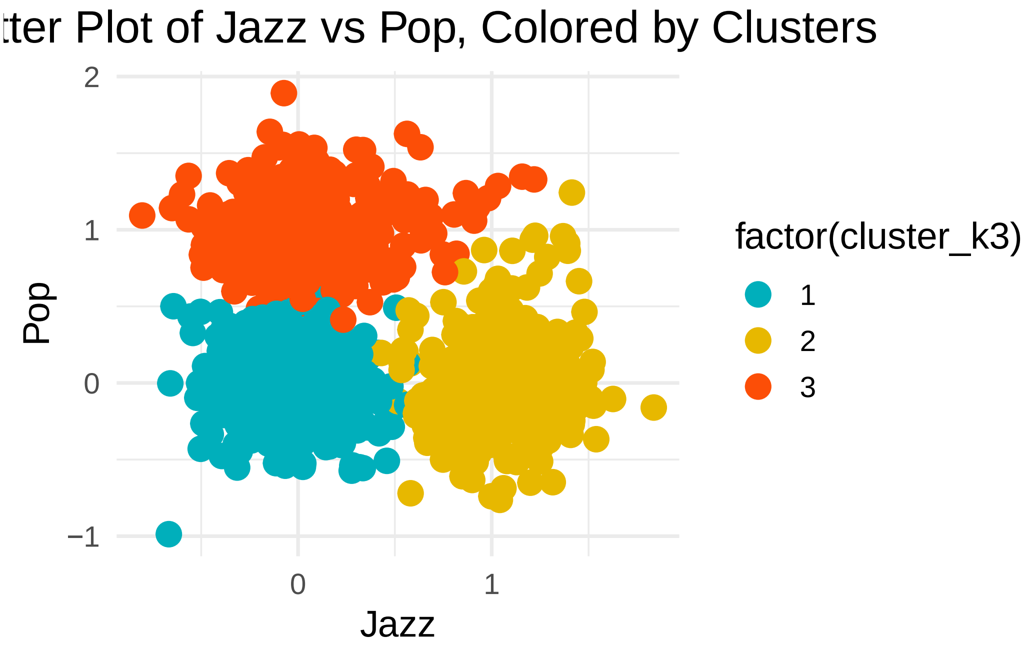
```
factor(cluster_k3)1  1.09530    0.06139   17.84   <2e-16 ***
factor(cluster_k3)2  2.14071    0.06122   34.97   <2e-16 ***
factor(cluster_k3)3  1.30525    0.06147   21.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.162 on 1072 degrees of freedom
Multiple R-squared:  0.6501,    Adjusted R-squared:  0.6491
F-statistic:   664 on 3 and 1072 DF,  p-value: < 2.2e-16
```

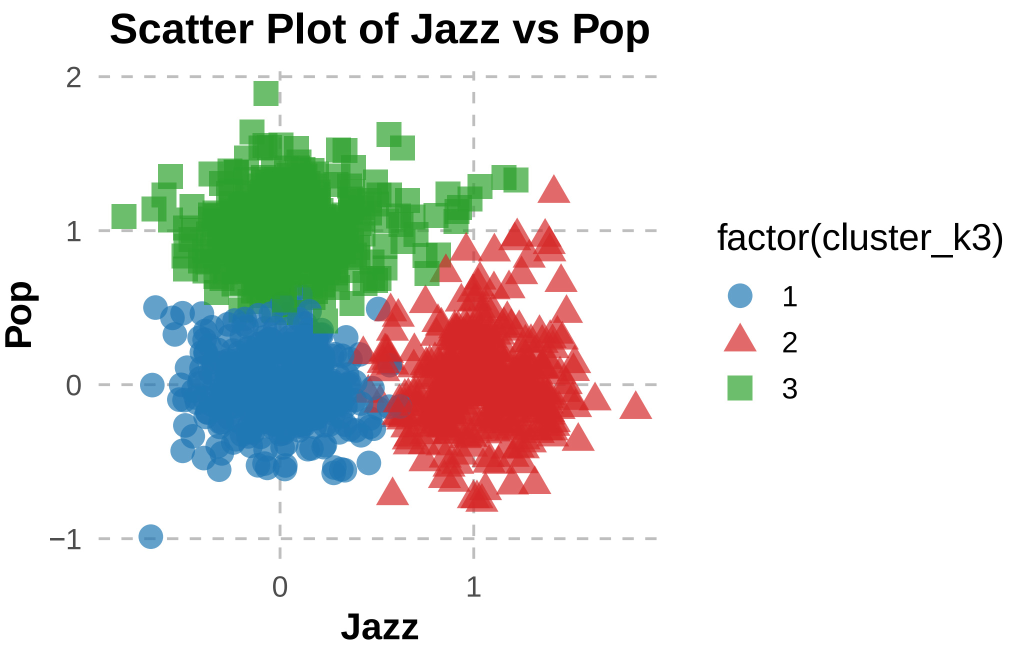
The results show that there are clear differences in influence between the clusters. Cluster has the highest average influence score, meaning that people in this group tend to have more influence on others compared to the other clusters. Cluster 1 has the lowest influence score, while Cluster 3 falls somewhere in the middle. The differences between these groups are statistically significant, meaning that the higher or lower influence scores are unlikely to be due to random chance. Overall, this suggests that people in Cluster 2 are the most influential, while those in Cluster 1 are the least influential.

So yes, there are differences in influence between the clusters.

Question 7



Styling the plot, for better visual.



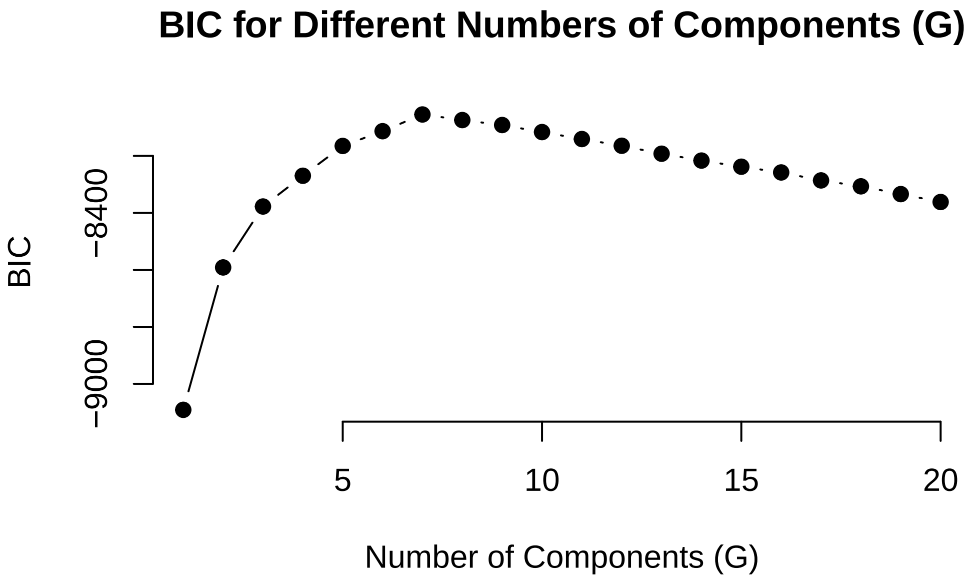
Yes, the k-means algorithm seems to have successfully picked up on the patterns observed earlier. In the initial plot (#1), I noticed multiple clusters based on the relationships between jazz and pop preferences. This plot shows that the clustering algorithm has separated the data into distinct groups based on these preferences, as visualized by the blue, red, and green clusters. The clustering aligns well with the general patterns observed in the earlier scatter plot.

The clusters show clear seperation for the most part. The green cluster (cluster 3) and the blue cluster (cluster 1) have distinct areas with minimal overlap, indicating strong separation. Similarly, the red (cluster 2) is well-defined with clear boundaries on the right side.

There are few overlap between the borders of the clusters, particularly between the blue and red clusters near the middle. This is expected due to plotting the data in 2D. However, overall, the spacing between the clusters is clear enough to distinguish different groups.

In clonclusing, the clusters are generally well separated, though there is some minor overlap near the middle. This suggests that the k-means algorithm effectively identified distinct groups based on musical tastes, with each cluster representing different preferences for jazz, pop, and hiphop.

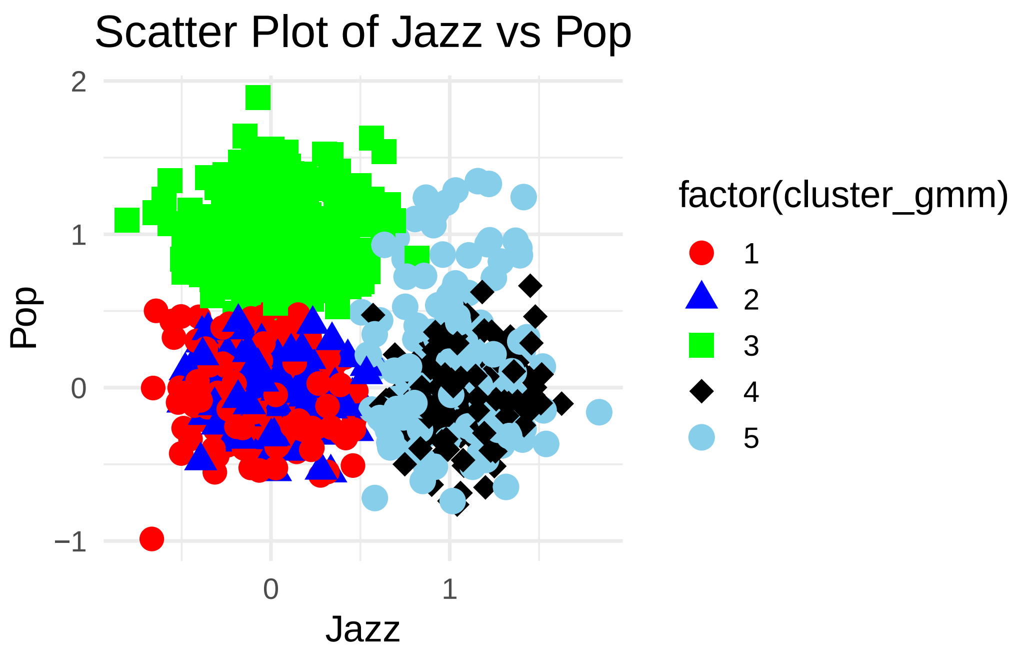
Question 8



This plot shows the BIC values for different numbers of clusters in the Gaussian Mixture Model. The BIC helps us decide the best number of clusters, with lower values being better. From the plot, the BIC is at its lowest point when there are 5 clusters, which means that 5 is likely the best number of clusters to use.

```
             [,1]       [,2]       [,3]       [,4]       [,5]
jazz   -0.6824176 -0.6422952 -0.6283058  1.2729777  1.1705562
pop    -0.6297820 -0.6279485  1.2176767 -0.6460063 -0.3534327
hiphop  0.9115783 -0.8506257 -0.1007189 -0.8537116  0.9320878
```

Ploting:



In the plot, Cluster 3 (green) is the only one that is clearly separated from the others, meaning the individuals in this group have distinct preferences. However, Cluster 1 (red) and Cluster 2 (blue) overlap quite a bit, suggesting that the people in these two groups have similar or mixed tastes, making it hard for the model to differentiate between them. Similarly, Cluster 4 (black) and Cluster 5 (sky blue) also overlap, showing that individuals in these groups have more uncertain or ambiguous preferences. The overlaps between clusters, especially in Clusters 1 and 2, and Clusters 4 and 5, indicate that the boundaries are not as clear, which is expected in a probabilistic model like GMM. This is probably due to the situatin that the model is capturing the uncertainty and the fact that some people's preferences fall between groups.

Question 9

```

Call:
lm(formula = influence ~ jazz + pop + hiphop + uncertainty, data = taste_influence)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4181 -0.6166  0.0223  0.6152  2.6519 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.76499    0.05220  14.656   <2e-16 ***
jazz         0.93311    0.05652  16.510   <2e-16 ***
pop          0.12238    0.05759   2.125   0.0338 *  
hiphop       1.07338    0.05205  20.622   <2e-16 ***
uncertainty -3.26656    0.29321 -11.141   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9608 on 1070 degrees of freedom
Multiple R-squared:  0.4071,	Adjusted R-squared:  0.4049 
F-statistic: 183.7 on 4 and 1070 DF,  p-value: < 2.2e-16
```

The results show that people who prefer hiphop have the greatest influence, with their influence score increasing by 1.07 for each unit of hip-hop preference. Jazz also has a strong effect, with each unit increase in jazz preference raising the influence score by 0.93. Pop has a smaller positive effect, increasing influence by 0.12 for each unit of pop preference. However, uncertainity plays a negative role. Individuals with higher uncertainty in their cluster assignment see a decrease of 3.37 in their influence score for each unit increase in uncertainty. This means that people whose musical tastes are less clear or more mixed are generally less influential compared to those with more defined preferences for a particular genre.

Part 2

Question 1

```
Number of rows: 200

Number of columns: 25

'data.frame':	200 obs. of  25 variables:
 $ taste_jazz             : num  0.514 -0.939 -0.9 2.165 -1.004 ...
 $ taste_classical        : num  1.903 -0.205 -0.715 1.775 -1.029 ...
 $ taste_blues            : num  0.782 -1.606 -0.408 2.231 -0.906 ...
 $ taste_pop              : num  -0.994 1.438 0.301 -2.323 -0.202 ...
 $ taste_country          : num  -0.645 1.414 0.143 -2.461 0.405 ...
 $ taste_raegge           : num  -1.1967 1.3209 -0.6888 -3.0995 -0.0218 ...
 $ income                 : num  0.5622 0.4726 -0.7719 0.0369 -1.2684 ...
 $ nbhood_avg_income      : num  0.982 0.509 -1.413 -0.103 -0.458 ...
 $ education               : num  0.0869 0.1443 -1.0826 0.2845 -0.964 ...
 $ nbhood_avg_education    : num  0.2151 -0.0758 -1.2908 0.4931 -0.9584 ...
 $ nhood_crime             : num  -0.519 -0.8322 0.2236 -0.3893 -0.0971 ...
 $ nbhood_unemployment     : num  -0.0422 0.0237 0.18 -0.4525 0.105 ...
 $ nhbood_avg_temp         : num  -5.981 -0.869 11.162 8.901 20.295 ...
 $ nhbood_pop              : num  9.974 9.515 9.904 -0.294 10.083 ...
 $ nhbood_nr_lights        : num  10.779 7.081 -0.973 14.623 10.184 ...
 $ nhbood_nr_pizzerias     : num  16.84 2.65 7.95 15.46 12.23 ...
 $ city_avg_taste_jazz     : num  8.83 11.76 10.3 10.26 9.49 ...
 $ city_avg_taste_classical: num  9.26 11.62 9.28 10.21 9.69 ...
 $ city_avg_taste_blues    : num  8.23 9.45 9.04 9.57 8.7 ...
 $ city_avg_taste_pop      : num  11.49 9.46 10.35 9.61 10.78 ...
 $ city_avg_taste_country  : num  10.56 8.48 11.05 10.51 9.69 ...
 $ city_avg_taste_raegge   : num  11.09 9.15 10.81 9.27 10.24 ...
 $ taste_film_action       : num  0.0239 1.266 0.8066 -0.3661 -0.533 ...
 $ taste_film_romcom       : num  -0.1989 0.0613 0.0154 -0.9403 0.9833 ...
 $ taste_film_documentary  : num  0.481 -1.161 -1.246 1.015 -0.829 ...
```

The dataset has 200 rows and 25 columns. It includes information about people's music and film preferences, like how much they like jazz, pop, or action movies. It also contains details about their personal situation, like their income and education level, as well as information about their neighborhood, such as average income, education, crime rates, and the number of pizzerias. Additionally, the dataset has information about city-wide averages for music tastes and other neighborhood features, like temperature and population.

Question 2

```
                                PC1           PC2          PC3          PC4
taste_jazz              -1.130734e-02 -0.0023440600 -0.014436501 -0.008909244
taste_classical         -7.785547e-03  0.0075890578 -0.011243550 -0.010266701
taste_blues             -1.094852e-02  0.0036306694 -0.017388315 -0.001064776
taste_pop               -7.826432e-05 -0.0116550537  0.028459962 -0.008757703
taste_country            1.995948e-03 -0.0017553217  0.024188748  0.002956621
taste_raegge             4.944613e-03 -0.0147713147  0.025235482 -0.001484971
income                   5.135923e-03  0.0159201632 -0.023204570  0.004816634
nbhood_avg_income        6.926408e-03  0.0164944063 -0.018198678  0.002946876
education               -2.588437e-03  0.0090498047 -0.013038645  0.004062818
nbhood_avg_education    -5.864163e-03  0.0133028161 -0.010250448  0.001596353
nhood_crime              4.772406e-03 -0.0105344496  0.015323486 -0.012135919
nbhood_unemployment      4.138436e-03 -0.0096372982  0.013718324 -0.018383328
nhbood_avg_temp          4.802633e-02  0.5671282890  0.807538197  0.148611391
nhbood_pop              -2.325264e-01  0.4612561449 -0.155298040 -0.840024982
```

```
nhbood_nr_lights         -9.704813e-01 -0.1031315060  0.094911157  0.193648996
nhbood_nr_pizzerias      -2.950091e-02  0.6727586139 -0.555567040  0.479546204
city_avg_taste_jazz      -5.433612e-03 -0.0054625216  0.019601534 -0.008823974
city_avg_taste_classical  7.555348e-03 -0.0079776471  0.008232628 -0.015469707
city_avg_taste_blues      3.312960e-03 -0.0006816765  0.026113769 -0.027082346
city_avg_taste_pop       -6.397509e-03  0.0016809645 -0.017835786  0.031700528
city_avg_taste_country    3.592366e-03  0.0052711518 -0.005572193  0.031518718
city_avg_taste_raegge     1.429978e-03  0.0080628332 -0.013574051  0.015636057
taste_film_action         1.278440e-02 -0.0064266489  0.006172812 -0.009046225
taste_film_romcom        -8.944392e-03 -0.0154741603  0.003254971  0.021808982
taste_film_documentary   -5.620100e-03  0.0218192137 -0.012201239 -0.012146298
                                  PC5           PC6          PC7          PC8
taste_jazz               -0.391356896  0.1007131225 -0.138824400 -0.012237116
taste_classical          -0.353542324  0.0642878780 -0.176239846  0.012043016
taste_blues              -0.400089780  0.0717880600 -0.112471653 -0.045529973
taste_pop                 0.360573852 -0.0381043486  0.137084433 -0.023295118
taste_country             0.351138565 -0.0481445179  0.154380297  0.056135797
taste_raegge              0.366432559 -0.0587096832  0.159771305  0.032995999
income                   -0.098207691 -0.4030137669  0.132322032  0.026352692
nbhood_avg_income        -0.118070537 -0.3827974154  0.109550937  0.054578050
education                -0.119228791 -0.3414803329  0.124205027  0.009023843
nbhood_avg_education     -0.119504991 -0.3449084112  0.093427566  0.002757513
nhood_crime               0.105666397  0.3784314147 -0.149358221  0.021643243
nbhood_unemployment       0.109582504  0.3793168798 -0.113090635  0.001716456
nhbood_avg_temp          -0.027878766 -0.0113752045 -0.022070363 -0.001929894
nhbood_pop                0.035182999 -0.0115116986 -0.021443441  0.003988627
nhbood_nr_lights          0.001928062  0.0008131054  0.009878499  0.016213339
nhbood_nr_pizzerias       0.031100684  0.0626442486  0.045833218  0.014885787
city_avg_taste_jazz      -0.104188562  0.1469711754  0.381870614 -0.003249313
city_avg_taste_classical -0.133858683  0.1351414138  0.328425364  0.030538239
city_avg_taste_blues     -0.126604602  0.1602678761  0.321220376  0.001964281
city_avg_taste_pop        0.127842901 -0.1384285186 -0.390955012  0.060101036
city_avg_taste_country    0.092178022 -0.1351107272 -0.337439533 -0.020260831
city_avg_taste_raegge     0.100903423 -0.1773130795 -0.391350702  0.067751894
taste_film_action        -0.046440164  0.0479066415  0.002293344  0.810254873
taste_film_romcom         0.102078272 -0.0289704767 -0.006994258 -0.328035158
taste_film_documentary   -0.059566515  0.0030829813 -0.012834657 -0.463067926
                                  PC9         PC10          PC11         PC12
taste_jazz                0.012189257 -0.425216599  1.044962e-01 -0.045424729
taste_classical          -0.079291288 -0.367242396  2.167876e-02  0.080408701
taste_blues              -0.038703128 -0.302309194  1.180720e-01 -0.135252242
taste_pop                 0.076428274 -0.415785136  1.610499e-01 -0.046969496
taste_country             0.014479256 -0.380940697 -4.589530e-06  0.135381137
```

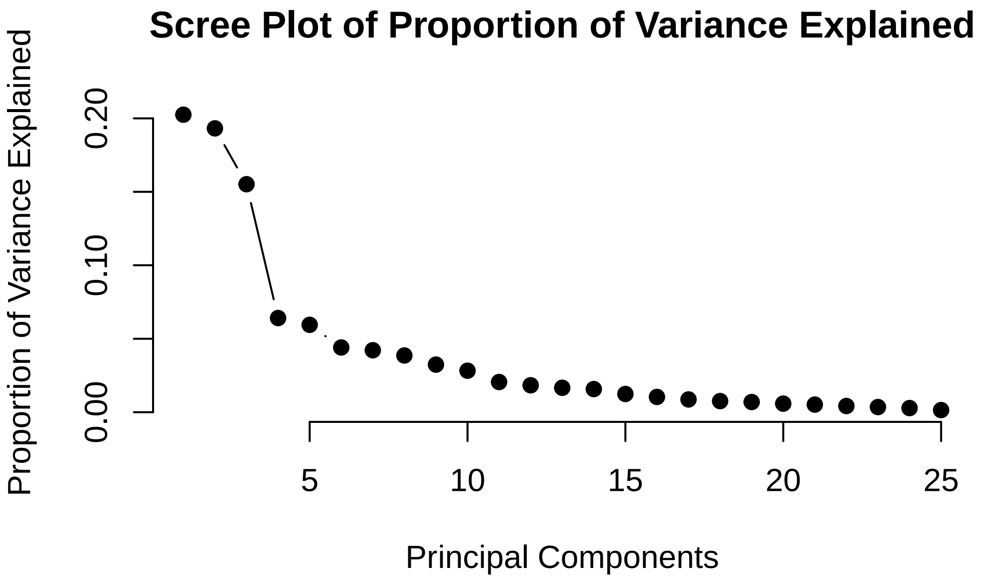
```
taste_raegge              0.085488497 -0.340749399  1.196841e-01 -0.088687815
income                   -0.034162964 -0.120763527 -5.040623e-01  0.022145255
nbhood_avg_income        -0.040982051 -0.061196459 -5.286246e-01  0.055115907
education                 0.001224664 -0.096356808  2.033939e-01  0.257679475
nbhood_avg_education      0.003561027 -0.109214186  1.779074e-01  0.204464711
nhood_crime              -0.007826796 -0.141948124 -3.851475e-01  0.307973946
nbhood_unemployment      -0.053422440 -0.165763993 -3.313715e-01  0.243385152
nhbood_avg_temp          -0.005267221  0.003053709 -4.798490e-03 -0.007793975
nhbood_pop               -0.027456210  0.005927121  1.119172e-02 -0.006761677
nhbood_nr_lights          0.010908013  0.007399605 -1.409646e-02  0.010823942
nhbood_nr_pizzerias      -0.007511522 -0.015304380  9.379607e-03  0.003765053
city_avg_taste_jazz      -0.022553920  0.038647180  2.014991e-02  0.032775852
city_avg_taste_classical -0.033068691  0.100046747  1.210566e-01  0.391433751
city_avg_taste_blues     -0.058196923  0.123697289  4.530256e-02  0.080104810
city_avg_taste_pop        0.002951694  0.035599262 -5.213114e-02 -0.382976129
city_avg_taste_country    0.034199035 -0.062989775  1.484362e-01  0.369177765
city_avg_taste_raegge    -0.030824772  0.195130442  1.407785e-01  0.477096880
taste_film_action         0.098790836  0.001693627  3.098243e-02 -0.060801204
taste_film_romcom        -0.713151169 -0.026616109  3.477667e-02 -0.056778901
taste_film_documentary    0.668215811  0.022405077 -7.370666e-02  0.022515418
                                  PC13         PC14         PC15         PC16
taste_jazz                0.0945098623 -0.013086909 -0.149658514  0.163440235
taste_classical           0.1757688069  0.062877277  0.087451202 -0.178489760
taste_blues               0.0648342457 -0.033402225 -0.037738194  0.115410162
taste_pop                 0.0503016269  0.014645813 -0.005225807  0.269256653
taste_country             0.1282251198 -0.116293737 -0.079255088  0.144287450
taste_raegge              0.1615205066  0.032623450 -0.062357571 -0.275376910
income                    0.1249857525  0.049210913 -0.120567373  0.042685501
nbhood_avg_income         0.1368478333  0.063594821 -0.044101620  0.008648714
education                -0.4164517948 -0.165652022  0.114162916  0.127947806
nbhood_avg_education     -0.4478274134 -0.205163957  0.115303886 -0.078875979
nhood_crime              -0.2870477449 -0.162320988  0.038370726  0.055300659
nbhood_unemployment      -0.2100359202 -0.030564154  0.050839789  0.031975768
nhbood_avg_temp           0.0015206997 -0.001074768  0.013404815  0.003886063
nhbood_pop               -0.0041584700  0.015168026 -0.003530627  0.002782233
nhbood_nr_lights          0.0060759029 -0.005191697  0.001127424 -0.007376924
nhbood_nr_pizzerias       0.0026664089 -0.006733946 -0.003943703 -0.003399659
city_avg_taste_jazz      -0.1956050709  0.684453556 -0.180183374  0.379652794
city_avg_taste_classical  0.2432368415 -0.085518058  0.240917507  0.309722787
city_avg_taste_blues      0.0242657841 -0.434685429 -0.768834203 -0.076375060
city_avg_taste_pop       -0.2366634911 -0.137290620 -0.218677433  0.548310947
city_avg_taste_country   -0.1758410054  0.435696253 -0.415443357 -0.289133078
city_avg_taste_raegge     0.4351319970 -0.037833942 -0.082473039  0.304521155
```

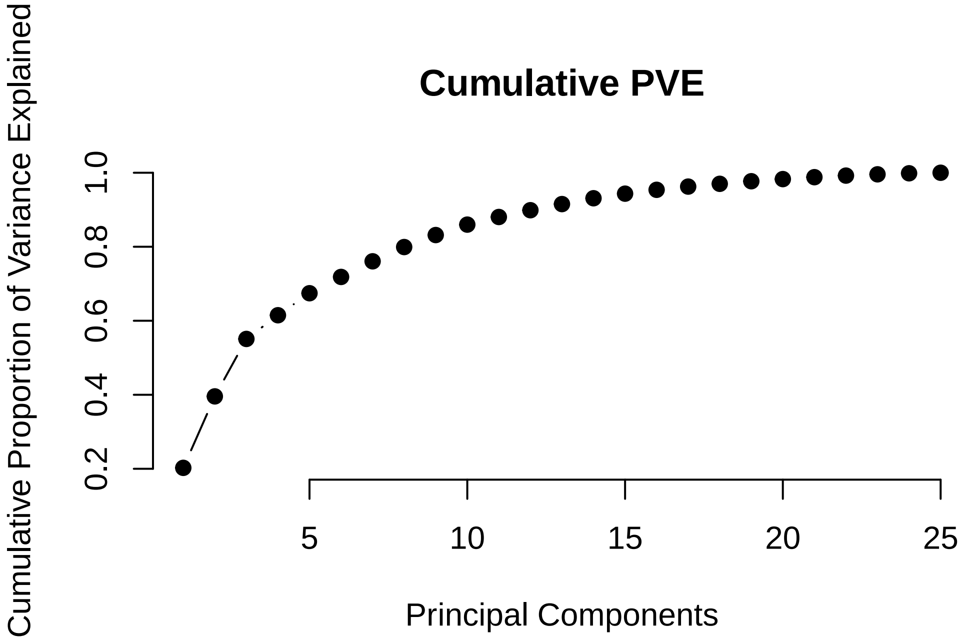
```
taste_film_action        -0.0244521968  0.024476726  0.031105685 -0.077017010
taste_film_romcom         0.0008967707  0.010294159  0.037293311 -0.037497639
taste_film_documentary    0.0334808895 -0.028529647  0.008440275  0.009441012
                                 PC17          PC18          PC19          PC20
taste_jazz                0.213777447 -0.1209619800  2.791794e-02 -0.4576144545
taste_classical          -0.009685631 -0.3448142951 -2.452772e-01 -0.0316089090
taste_blues              -0.342617715  0.4937705961  2.021310e-01  0.4953394878
taste_pop                 0.319446212 -0.3949907321  5.554183e-02  0.4854592235
taste_country             0.184202857  0.4901687602  2.485906e-01 -0.3731365656
taste_raegge             -0.626580360 -0.0347523483 -3.369478e-01 -0.0947829654
income                    0.023986209 -0.1109964809  1.099165e-01  0.1886574195
nbhood_avg_income        -0.003608526  0.1023788763 -1.926573e-01 -0.0497640341
education                -0.104603809 -0.0502094142  5.692946e-02  0.0895553662
nbhood_avg_education     -0.110145636 -0.0782875935 -4.668685e-05 -0.1829050409
nhood_crime              -0.113663068 -0.0225789575  6.515083e-02  0.0791070856
nbhood_unemployment      -0.117785459 -0.0617976735 -6.230234e-02 -0.0002724561
nhbood_avg_temp          -0.001121923  0.0007497393 -2.859762e-03  0.0005528799
nhbood_pop                0.001378507  0.0130684023 -8.090538e-03  0.0003101993
nhbood_nr_lights          0.002131427 -0.0006493079 -3.072023e-03  0.0042503686
nhbood_nr_pizzerias      -0.002330654 -0.0121944810 -6.170627e-04  0.0022646803
city_avg_taste_jazz      -0.250270932 -0.1099311765  1.112666e-01 -0.1432638027
city_avg_taste_classical  0.132876129  0.2177888542 -5.793821e-01  0.1063795430
city_avg_taste_blues     -0.027704353 -0.1310122870 -2.777185e-02  0.0673441099
city_avg_taste_pop       -0.076863125  0.0470334410 -4.332274e-01 -0.0639658790
city_avg_taste_country    0.258551494  0.2415163667 -2.340900e-01  0.1681133020
city_avg_taste_raegge    -0.319957961 -0.2069852649  2.536519e-01 -0.0515740692
taste_film_action         0.027288797 -0.0556511164  7.608949e-03  0.0440615589
taste_film_romcom         0.001221406 -0.0559338433 -3.987620e-02 -0.0211362237
taste_film_documentary   -0.028905591 -0.0527100849 -4.279142e-02 -0.0272231472
                                 PC21         PC22          PC23          PC24
taste_jazz                0.529869503  0.023860092 -0.0317398791  0.1030058925
taste_classical          -0.636872298  0.087166784  0.1332744126 -0.0182236461
taste_blues               0.003801776 -0.130211887 -0.0706654722 -0.0617922287
taste_pop                 0.061572910 -0.249160234 -0.0032105409 -0.0237483997
taste_country            -0.365001465  0.079116543  0.0449854712 -0.0719341365
taste_raegge              0.212023507  0.088643065 -0.0063042539  0.1145423554
income                    0.023194314  0.500602255 -0.4115094326  0.1042934347
nbhood_avg_income         0.103113409 -0.537817248  0.3768551941 -0.1032672878
education                 0.112759079  0.387285138  0.5574521657 -0.0850158989
nbhood_avg_education     -0.116706041 -0.392860759 -0.5386739293 -0.0009299925
nhood_crime              -0.015818802 -0.098709860  0.0969599828  0.6228775300
nbhood_unemployment       0.143993930  0.092418294 -0.1391675628 -0.7061048377
nhbood_avg_temp           0.009594324  0.005535916 -0.0018003374  0.0005786653
```

```
nhbood_pop                0.004524952  0.003608993  0.0004922846  0.0075129681
nhbood_nr_lights          0.002908450  0.001090385  0.0007537353 -0.0014948624
nhbood_nr_pizzerias       0.004883096 -0.005053748  0.0017211244  0.0006245696
city_avg_taste_jazz      -0.148925303 -0.055607034  0.0013841940  0.0290460743
city_avg_taste_classical  0.067758029  0.068523537 -0.1572014224  0.0733397706
city_avg_taste_blues     -0.091418193 -0.044841797  0.0468240022 -0.0650147275
city_avg_taste_pop       -0.163933467  0.065975613 -0.0369206924  0.0035008496
city_avg_taste_country    0.067007460  0.025395660 -0.0360502851  0.0278359419
city_avg_taste_raegge     0.015244217 -0.091013034 -0.0221836757 -0.0413325084
taste_film_action         0.006870541 -0.066765813  0.0177215640 -0.1189213884
taste_film_romcom         0.028963484 -0.072655201  0.0351401853 -0.0754237337
taste_film_documentary   -0.073497322 -0.057699032  0.0386584142 -0.1349646020
                                 PC25
taste_jazz                0.035036513
taste_classical          -0.073660871
taste_blues               0.035121668
taste_pop                -0.052714374
taste_country             0.064581269
taste_raegge             -0.001086311
income                    0.110642049
nbhood_avg_income        -0.092391788
education                 0.025548622
nbhood_avg_education     -0.017335666
nhood_crime               0.119820061
nbhood_unemployment      -0.117683087
nhbood_avg_temp           0.001602152
nhbood_pop                0.003570177
nhbood_nr_lights         -0.001701038
nhbood_nr_pizzerias      -0.002769651
city_avg_taste_jazz       0.024950609
city_avg_taste_classical  0.060862184
city_avg_taste_blues      0.003178758
city_avg_taste_pop        0.013471479
city_avg_taste_country    0.045313826
city_avg_taste_raegge     0.020817680
taste_film_action         0.539803387
taste_film_romcom         0.587742288
taste_film_documentary    0.539782669
```

Why it is problematic?

Performing PCA without standardizing the data is problematic because the variables in the dataset are likely on different scales. For example, inclome might be measured in thousands, while music tastes (like jazz or pop) are measured on a smaller scale. Neighborhood features, like the number of pizzerias or the temperature, are on entirely different scales as well. Without standardization, PCA may overemphasize variables with larger scales, leading to misleading results where those variables dominate the principal components. It's important to standardize the data (scale all variables to have mean = 0 and standard deviation = 1) before applying PCA, so each variable contributes equally regardless of its original scale.





Based on the scree plot and the cumulative PVE plot, it is clear that keeping around 5 principal components is a good choice. The scree plot shows that the amount of variance explained by each component drops quickly after the first 4 or 5, and then flattens out. The cumulative PVE plot also shows that the first 5 components explain about 80% of the total variance. After that, adding more components doesn't make much difference. So, keeping 5 components will capture most of the important information in the data without making it too complex.

Question 4

```
                             PC1          PC2          PC3          PC4
taste_jazz          -0.165598284 -0.315796907  0.188815999  0.008645344
taste_classical     -0.186439330 -0.282337949  0.220597843  0.034722749
taste_blues         -0.197560893 -0.315298024  0.166961965 -0.021606793
taste_pop            0.218843741  0.270293966 -0.188021321 -0.046135465
taste_country        0.205061458  0.277200903 -0.206469793  0.021552088
taste_raegge         0.205283637  0.279349334 -0.207769006  0.016002989
income              -0.334078396  0.119288960 -0.200122814  0.036660873
nbhood_avg_income   -0.340921831  0.100533268 -0.182486414  0.070252855
education           -0.341538161  0.082650349 -0.207921612  0.023059624
nbhood_avg_education -0.349351488  0.090315358 -0.180684902  0.005989955
nhood_crime          0.329488500 -0.103288082  0.217348454  0.014278135
nbhood_unemployment  0.346934465 -0.113247576  0.191109908 -0.006426588
nhbood_avg_temp      0.033748193  0.010148644 -0.045986155 -0.050748419
```

```
nhbood_pop               -0.036709545 -0.046398844 -0.014724121 -0.211302634
nhbood_nr_lights         -0.006280977 -0.004203786  0.018981095 -0.189007367
nhbood_nr_pizzerias      -0.110745654  0.026056839  0.061859046 -0.186292300
city_avg_taste_jazz       0.095354373 -0.243906113 -0.308085582 -0.023559477
city_avg_taste_classical  0.067175435 -0.280305639 -0.289200006  0.047647626
city_avg_taste_blues      0.097818272 -0.278059681 -0.271576796  0.001225083
city_avg_taste_pop       -0.076748148  0.263499641  0.319545930  0.060475896
city_avg_taste_country   -0.091673374  0.247513332  0.295280303  0.009527403
city_avg_taste_raegge    -0.112170433  0.248201800  0.290919907  0.084001215
taste_film_action         0.020026975 -0.063442131  0.002728314  0.711698717
taste_film_romcom         0.049294821  0.113532618 -0.006786904 -0.227452260
taste_film_documentary   -0.059787432 -0.053353936  0.029209916 -0.546605735
                                   PC5          PC6          PC7          PC8
taste_jazz               -4.807130e-02  0.025543417 -0.140837350 -0.142106313
taste_classical          -6.322271e-02  0.095356195 -0.038907687 -0.153191757
taste_blues              -8.290845e-02  0.032036861 -0.071417700 -0.105090297
taste_pop                 4.656127e-02  0.034126013 -0.167741357 -0.116564042
taste_country             5.467145e-02  0.093755637 -0.057777495 -0.064802474
taste_raegge              5.140658e-02 -0.007247030 -0.142725658 -0.073238896
income                   -3.582694e-05 -0.006680535  0.047836117  0.052807736
nbhood_avg_income         6.787990e-03 -0.003253640  0.076182387  0.034012305
education                -2.630559e-02  0.012127159 -0.065499618 -0.072905931
nbhood_avg_education     -1.014768e-02  0.053499861 -0.068511997 -0.104959353
nhood_crime               4.273240e-02  0.020770471  0.037623211  0.001020844
nbhood_unemployment       1.377604e-02  0.044884436  0.048100974  0.027252395
nhbood_avg_temp           2.151341e-01  0.347025929  0.434541120 -0.746805691
nhbood_pop                3.700403e-01  0.537389689 -0.011283582  0.250357144
nhbood_nr_lights         -1.519667e-01  0.553902590 -0.599386642 -0.010956630
nhbood_nr_pizzerias       2.468936e-01  0.222752692  0.458837297  0.468049875
city_avg_taste_jazz      -4.694237e-02  0.027932233 -0.001003856 -0.005052657
city_avg_taste_classical -3.406103e-02 -0.062782227  0.067283691  0.067222239
city_avg_taste_blues     -1.690488e-02  0.044324249  0.091318079 -0.062684045
city_avg_taste_pop       -5.199699e-03  0.034654530 -0.049194413  0.026651062
city_avg_taste_country    5.539677e-03 -0.048194573  0.027168304 -0.180133673
city_avg_taste_raegge     2.086289e-02  0.024152551  0.056432431 -0.019419872
taste_film_action         2.548018e-01  0.162714507 -0.094841874  0.070810008
taste_film_romcom        -6.802694e-01  0.188473815  0.274049680  0.051454075
taste_film_documentary    4.207922e-01 -0.367800546 -0.197083056 -0.108920697
                                  PC9         PC10         PC11        PC12
taste_jazz               -0.173808413  0.310590176  0.053029964  0.12129885
taste_classical          -0.238881520  0.266189767  0.040736108  0.01708830
taste_blues              -0.092859254  0.256145577  0.031460000  0.18805728
taste_pop                -0.190560982  0.287651054  0.115514176  0.14657546
```

```
taste_country            -0.147476133  0.337185817  0.058667369 -0.05611331
taste_raegge             -0.126183740  0.251011863  0.045731768  0.16638854
income                   -0.146823188  0.101254293 -0.425554337 -0.22189388
nbhood_avg_income        -0.136131787  0.045255727 -0.453994676 -0.27787291
education                -0.055794699  0.062884104  0.316896354 -0.16853823
nbhood_avg_education     -0.078560078  0.051974069  0.265269540 -0.13172324
nhood_crime              -0.158214235  0.121448596 -0.133673040 -0.44750833
nbhood_unemployment      -0.196466305  0.140368216 -0.116551898 -0.38550231
nhbood_avg_temp           0.171100639 -0.053907642 -0.122207250  0.04493805
nhbood_pop               -0.531540679 -0.318380826  0.104054153  0.07167598
nhbood_nr_lights          0.437911294 -0.002823064 -0.111668071 -0.19241727
nhbood_nr_pizzerias       0.363739791  0.479933849  0.119969828  0.03035797
city_avg_taste_jazz       0.185856110  0.085830548  0.005005366 -0.17355705
city_avg_taste_classical  0.055208911 -0.052959268  0.321024507 -0.24834936
city_avg_taste_blues      0.006038633 -0.178141258  0.056420322  0.03011061
city_avg_taste_pop        0.090402904 -0.023969727 -0.189166531  0.21878969
city_avg_taste_country    0.046084617  0.079917626  0.331670086 -0.36791221
city_avg_taste_raegge    -0.003415377 -0.263000335  0.278308842 -0.21457552
taste_film_action         0.091982439  0.034888888 -0.002527119  0.05235425
taste_film_romcom        -0.145136771 -0.017150243  0.015276714  0.06630590
taste_film_documentary    0.061396622 -0.006925673 -0.061936508 -0.04484737
                                 PC13        PC14         PC15         PC16
taste_jazz                0.084782734 -0.02949072  0.109790707  0.178418491
taste_classical           0.204885836  0.01547718 -0.026644776 -0.154869725
taste_blues               0.039884406 -0.03348638  0.023658567  0.135071748
taste_pop                 0.062355312  0.01922825 -0.023507452  0.246264399
taste_country             0.147562290 -0.18620345  0.076047326  0.122555660
taste_raegge              0.122135263  0.02237394  0.100109161 -0.195040863
income                    0.141953055 -0.01269950  0.120184666  0.062041692
nbhood_avg_income         0.174872908 -0.01101349  0.053444154  0.026096484
education                -0.370856363 -0.10975775 -0.148536658  0.110308314
nbhood_avg_education     -0.428981122 -0.12326185 -0.116520624 -0.112783149
nhood_crime              -0.262382763 -0.18704712 -0.064886872  0.034406584
nbhood_unemployment      -0.185873430 -0.05807765 -0.072751139  0.027030057
nhbood_avg_temp           0.014221153 -0.02351276 -0.150852742  0.007764799
nhbood_pop               -0.006309821  0.19969601  0.016130433  0.049587748
nhbood_nr_lights          0.085689266 -0.12226176  0.013691823 -0.084087410
nhbood_nr_pizzerias       0.004096595 -0.10018175  0.056633591 -0.031140398
city_avg_taste_jazz      -0.044185783  0.65922327  0.043749445  0.428703834
city_avg_taste_classical  0.344049886 -0.21969729 -0.282174333  0.241801714
city_avg_taste_blues     -0.124376126 -0.37331608  0.770138815  0.087532876
city_avg_taste_pop       -0.286450449 -0.09285865  0.047684285  0.637081900
city_avg_taste_country    0.026141928  0.39474037  0.450765885 -0.142944090
```

```
city_avg_taste_raegge     0.452896654 -0.19556200  0.018912912  0.291559645
taste_film_action        -0.029456699  0.04234429 -0.008233693 -0.088904378
taste_film_romcom        -0.007063703  0.03115654 -0.028625290 -0.032961208
taste_film_documentary    0.028550583 -0.05719423 -0.010223560  0.010850422
                                 PC17        PC18        PC19         PC20
taste_jazz                0.141258893 -0.14184850  0.09640072 -0.216704523
taste_classical          -0.090775277 -0.29161749 -0.28936466 -0.303651393
taste_blues              -0.184801798  0.50322867  0.19787587  0.503427743
taste_pop                 0.223198873 -0.50307975 -0.06348292  0.420678479
taste_country             0.214744591  0.33480573  0.44702904 -0.416862492
taste_raegge             -0.569047967  0.23936316 -0.42198158 -0.012176435
income                   -0.002300923 -0.12787656  0.03014562  0.191900461
nbhood_avg_income         0.042099291  0.11885457 -0.13942573 -0.026112309
education                -0.098578755 -0.04170411  0.05630122  0.218947051
nbhood_avg_education     -0.113063015 -0.02509947 -0.02636801 -0.281341409
nhood_crime              -0.103658525 -0.01528071  0.04207905  0.083109550
nbhood_unemployment      -0.114414105 -0.01263256 -0.08659129  0.044339101
nhbood_avg_temp           0.006760906  0.01723026 -0.02483635  0.035775023
nhbood_pop                0.056112717  0.10980293 -0.03643447  0.009139158
nhbood_nr_lights          0.016962320 -0.01278544 -0.03363350  0.040653957
nhbood_nr_pizzerias      -0.053488509 -0.09353098 -0.03032707  0.026113646
city_avg_taste_jazz      -0.289563938 -0.06432175  0.10564009 -0.170915387
city_avg_taste_classical  0.265076709  0.22702066 -0.44060320  0.029066321
city_avg_taste_blues     -0.070071893 -0.10550152 -0.06353635  0.009694349
city_avg_taste_pop        0.059325957  0.15075485 -0.38719938 -0.173309975
city_avg_taste_country    0.304301938  0.18091434 -0.17698949  0.116018107
city_avg_taste_raegge    -0.454590007 -0.18888229  0.22724983 -0.011515355
taste_film_action         0.013026390 -0.06404510 -0.02491066  0.034923153
taste_film_romcom        -0.005143275 -0.03440860 -0.05409334 -0.019068788
taste_film_documentary   -0.038178774 -0.03726415 -0.05141659 -0.063004524
                                 PC21         PC22          PC23         PC24
taste_jazz                0.695599574  0.001545591  0.0368753328 -0.122432027
taste_classical          -0.527950339 -0.202649911 -0.0745390465  0.022630387
taste_blues              -0.215656868  0.208907317  0.0040639328  0.073986524
taste_pop                -0.099625702  0.258046170 -0.1121526814  0.034906117
taste_country            -0.203080759 -0.130698947 -0.0038298274  0.071855156
taste_raegge              0.230993554 -0.074636996  0.0410290525 -0.125864851
income                   -0.047790064 -0.159647814  0.6422831774 -0.130901850
nbhood_avg_income         0.113982728  0.222658507 -0.6048932718  0.124903940
education                 0.077127160 -0.611258201 -0.2357092646  0.066437968
nbhood_avg_education     -0.053422614  0.594139615  0.2166077865  0.009240723
nhood_crime              -0.069142958  0.056743715 -0.1538235385 -0.629354561
nbhood_unemployment       0.146440188 -0.014597864  0.1865553484  0.683979724
```

```
nhbood_avg_temp          0.064082797 -0.023257621  0.0324273884 -0.005855453
nhbood_pop               0.022991938 -0.020295988  0.0059053985 -0.049911735
nhbood_nr_lights         0.010061410 -0.009559630  0.0001918179  0.012325878
nhbood_nr_pizzerias      0.030914874  0.027745967 -0.0287695874 -0.003374108
city_avg_taste_jazz     -0.082155445  0.036569405 -0.0286610492 -0.027795411
city_avg_taste_classical 0.023188733  0.038044187  0.1259045026 -0.063590304
city_avg_taste_blues    -0.081535141  0.005453017 -0.0523354110  0.061453980
city_avg_taste_pop      -0.108380415 -0.051073192  0.0508533095 -0.004468997
city_avg_taste_country   0.002428511  0.007918055  0.0321587197 -0.024181700
city_avg_taste_raegge    0.031474403  0.101502582 -0.0199249305  0.043276561
taste_film_action       -0.005271302  0.045919873 -0.0603741813  0.131937843
taste_film_romcom        0.036069702  0.032503297 -0.0740540134  0.075126342
taste_film_documentary  -0.048719778  0.012019021 -0.0703748280  0.137254347
                                PC25
taste_jazz               0.038011426
taste_classical         -0.075606014
taste_blues              0.037925833
taste_pop               -0.054340071
taste_country            0.063982482
taste_raegge            -0.001230300
income                   0.123984442
nbhood_avg_income       -0.100905274
education                0.021956319
nbhood_avg_education    -0.014464037
nhood_crime              0.122905019
nbhood_unemployment     -0.116358452
nhbood_avg_temp          0.011553028
nhbood_pop               0.022880574
nhbood_nr_lights        -0.013249304
nhbood_nr_pizzerias     -0.019606293
city_avg_taste_jazz      0.024753321
city_avg_taste_classical 0.053724010
city_avg_taste_blues     0.002160835
city_avg_taste_pop       0.012704576
city_avg_taste_country   0.041331122
city_avg_taste_raegge    0.021721983
taste_film_action        0.575572697
taste_film_romcom        0.556740091
taste_film_documentary   0.530241983
```

For better looking

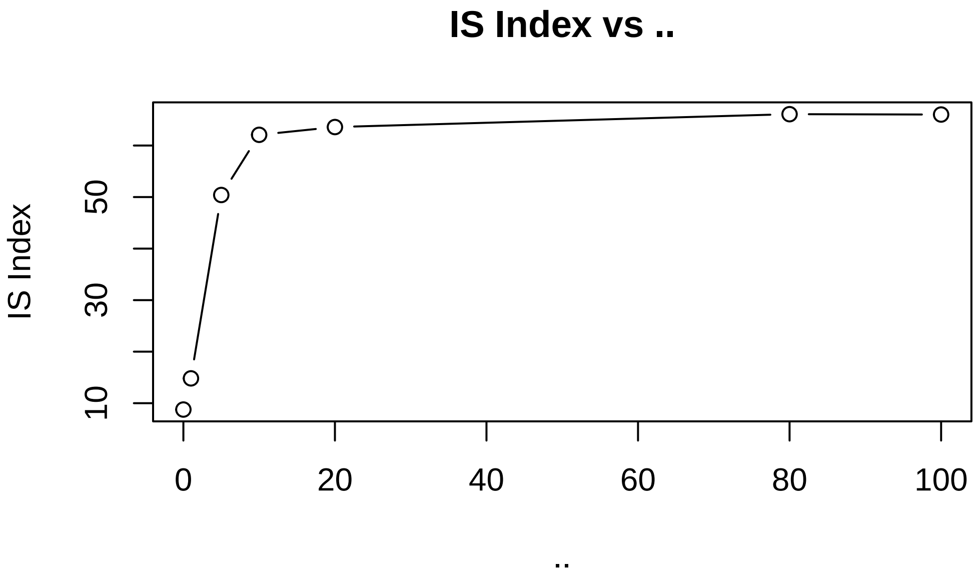
```
                                  PC1          PC2          PC3          PC4
```

```
taste_jazz               -0.165598284 -0.315796907  0.188815999  0.008645344
taste_classical          -0.186439330 -0.282337949  0.220597843  0.034722749
taste_blues              -0.197560893 -0.315298024  0.166961965 -0.021606793
taste_pop                 0.218843741  0.270293966 -0.188021321 -0.046135465
taste_country             0.205061458  0.277200903 -0.206469793  0.021552088
taste_raegge              0.205283637  0.279349334 -0.207769006  0.016002989
income                   -0.334078396  0.119288960 -0.200122814  0.036660873
nbhood_avg_income        -0.340921831  0.100533268 -0.182486414  0.070252855
education                -0.341538161  0.082650349 -0.207921612  0.023059624
nbhood_avg_education     -0.349351488  0.090315358 -0.180684902  0.005989955
nhood_crime               0.329488500 -0.103288082  0.217348454  0.014278135
nbhood_unemployment       0.346934465 -0.113247576  0.191109908 -0.006426588
nhbood_avg_temp           0.033748193  0.010148644 -0.045986155 -0.050748419
nhbood_pop               -0.036709545 -0.046398844 -0.014724121 -0.211302634
nhbood_nr_lights         -0.006280977 -0.004203786  0.018981095 -0.189007367
nhbood_nr_pizzerias      -0.110745654  0.026056839  0.061859046 -0.186292300
city_avg_taste_jazz       0.095354373 -0.243906113 -0.308085582 -0.023559477
city_avg_taste_classical  0.067175435 -0.280305639 -0.289200006  0.047647626
city_avg_taste_blues      0.097818272 -0.278059681 -0.271576796  0.001225083
city_avg_taste_pop       -0.076748148  0.263499641  0.319545930  0.060475896
city_avg_taste_country   -0.091673374  0.247513332  0.295280303  0.009527403
city_avg_taste_raegge    -0.112170433  0.248201800  0.290919907  0.084001215
taste_film_action         0.020026975 -0.063442131  0.002728314  0.711698717
taste_film_romcom         0.049294821  0.113532618 -0.006786904 -0.227452260
taste_film_documentary   -0.059787432 -0.053353936  0.029209916 -0.546605735
                                   PC5
taste_jazz               -4.807130e-02
taste_classical          -6.322271e-02
taste_blues              -8.290845e-02
taste_pop                 4.656127e-02
taste_country             5.467145e-02
taste_raegge              5.140658e-02
income                   -3.582694e-05
nbhood_avg_income         6.787990e-03
education                -2.630559e-02
nbhood_avg_education     -1.014768e-02
nhood_crime               4.273240e-02
nbhood_unemployment       1.377604e-02
nhbood_avg_temp           2.151341e-01
nhbood_pop                3.700403e-01
nhbood_nr_lights         -1.519667e-01
nhbood_nr_pizzerias       2.468936e-01
city_avg_taste_jazz      -4.694237e-02
```

```
city_avg_taste_classical -3.406103e-02
city_avg_taste_blues     -1.690488e-02
city_avg_taste_pop       -5.199699e-03
city_avg_taste_country    5.539677e-03
city_avg_taste_raegge     2.086289e-02
taste_film_action         2.548018e-01
taste_film_romcom        -6.802694e-01
taste_film_documentary    4.207922e-01
```

The first principal component (PC1) is mainly influenced by socio-economic factors and neighborhood characteristics. Higher income (-0.33), education (-0.34), and average neighborhood income (-0.34) contribute negatively, while higher crime (0.33) and unemployment rates (0.35) contribute positively. The second principal component (PC2) is driven by music preferences. People who prefer jazz (-0.32), classical (-0.28), and blues (-0.32) are on the negative side, while those who prefer pop (0.27), country (0.28), and reggae (0.28) are on the positive side. PC2 separates different music tastes, with minor contributions from socio-economic factors like income (0.12). The third principal component (PC3) is also a mix of music preferences and socio-economic factors. Those who prefer jazz (0.19), classical (0.22), and blues (0.17) are on the positive side, while pop (-0.19), country (-0.21), and reggae (-0.21) fans, along with higher income (-0.20) and education (-0.21), tend to be on the negative side. PC3 also shows that higher crime (0.22) and unemployment rates (0.19) are linked to the positive side. The fourth principal component (PC4) is dominated by film preferences, with action film fans (0.71) on the positive side and those who prefer documentaries (-0.55) and romantic comedies (-0.23) on the negative side. Music and socio-economic factors play a minimal role in this component. For PC5, the values are very low.

Question 5



Based on the Is Index vs plot, the best value for appears to be around 10. The IS index increases sharply from = 0 to = 10, indicating that this value provides a good balance between maintaining explained variance and enforcing sparsity in the model. After = 10, the IS index flattens out, meaning that increasing further does not lead to any meaningful improvement. This suggests that choosing a larger value, such as 20 or higher, would not provide additional benefits. Therefore, = 10 is the most appropriate choice to maximize the model's performance while simplifying the interpretation.

```
                            PC1         PC2       PC3         PC4
taste_jazz           0.00000000 -0.40944779 0.0000000 0.000000000
taste_classical      0.00000000 -0.40055033 0.0000000 0.000000000
taste_blues          0.00000000 -0.41758646 0.0000000 0.000000000
taste_pop            0.00000000  0.40676154 0.0000000 0.000000000
taste_country        0.00000000  0.39713415 0.0000000 0.000000000
taste_raegge         0.00000000  0.41580207 0.0000000 0.000000000
income              -0.40514515  0.00000000 0.0000000 0.000000000
nbhood_avg_income   -0.41113209  0.00000000 0.0000000 0.000000000
education           -0.41401382  0.00000000 0.0000000 0.000000000
nbhood_avg_education -0.39028352  0.00000000 0.0000000 0.000000000
nhood_crime          0.40884101  0.00000000 0.0000000 0.000000000
nbhood_unemployment  0.41838243  0.00000000 0.0000000 0.000000000
```

```
nhbood_avg_temp           0.00000000  0.03836199  0.0000000  0.004499147
nhbood_pop                0.00000000  0.00000000  0.0000000  0.000000000
nhbood_nr_lights          0.00000000  0.00000000  0.0000000 -0.189249995
nhbood_nr_pizzerias      -0.03007174  0.00000000  0.0567533  0.000000000
city_avg_taste_jazz       0.00000000  0.00000000 -0.4048178  0.000000000
city_avg_taste_classical  0.00000000  0.00000000 -0.3973007  0.000000000
city_avg_taste_blues      0.00000000  0.00000000 -0.4083402  0.000000000
city_avg_taste_pop        0.00000000  0.00000000  0.4350293  0.000000000
city_avg_taste_country    0.00000000  0.00000000  0.3926632  0.000000000
city_avg_taste_raegge     0.00000000  0.00000000  0.4060506  0.000000000
taste_film_action         0.00000000  0.00000000  0.0000000  0.883012110
taste_film_romcom         0.00000000  0.00000000  0.0000000 -0.428740047
taste_film_documentary    0.00000000  0.00000000  0.0000000 -0.025214723
                                PC5
taste_jazz                0.0000000
taste_classical           0.0000000
taste_blues               0.0000000
taste_pop                 0.0000000
taste_country             0.0000000
taste_raegge              0.0000000
income                    0.0000000
nbhood_avg_income         0.0000000
education                 0.0000000
nbhood_avg_education      0.0000000
nhood_crime               0.0000000
nbhood_unemployment       0.0000000
nhbood_avg_temp           0.1306847
nhbood_pop                0.3998735
nhbood_nr_lights          0.0000000
nhbood_nr_pizzerias       0.2693230
city_avg_taste_jazz       0.0000000
city_avg_taste_classical  0.0000000
city_avg_taste_blues      0.0000000
city_avg_taste_pop        0.0000000
city_avg_taste_country    0.0000000
city_avg_taste_raegge     0.0000000
taste_film_action         0.0000000
taste_film_romcom        -0.2518780
taste_film_documentary    0.8288820
```

The principal loadings from the sparse PCA show that each component focuses on fewer key variables, making it easier to interpret compared to standard PCA. For example, PC1 is mainly influenced by socio-economic factors such as income (-0.41), neighborhood average income (-0.41), education (-0.41), and neighborhood crime (0.41), highlighting a balance between socio-economic status and crime/unemployment. PC2 focuses on music preferences, with negative loadings for jazz (-0.41), classical (-0.40), and blues (-0.42), and positive loadings for pop (0.41), country (0.40), and reggae (0.42). PC3 reflects on music preferences, contrasting pop (0.44) and country (0.39) with jazz (-0.40), and classical (-0.40). PC4 is driven by film preferences, with a strong positive loading for action films (0.88) and negative loadings for romantic comedies (-0.43). Lastly, PC5 highlights documentary preferences (0.83) and neighborhood population (0.40). Sparse PCA is easier to interpret because it zeros out irrelevant variables, but this simplification might cause it to miss subtle patterns, making standard PCA more comprehensive in capturing all relationships.

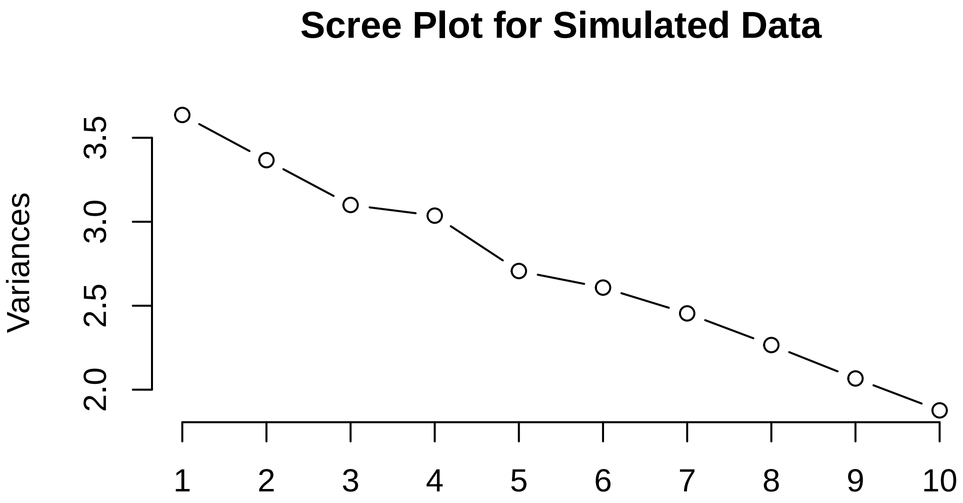
Question 6

```
Importance of components:
                           PC1     PC2   PC3     PC4     PC5     PC6     PC7
Standard deviation     1.90678 1.83487 1.761 1.74253 1.64522 1.61491 1.56665
Proportion of Variance 0.07272 0.06734 0.062 0.06073 0.05413 0.05216 0.04909
Cumulative Proportion  0.07272 0.14005 0.202 0.26278 0.31691 0.36907 0.41816
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     1.50530 1.43766 1.36997 1.34751 1.28997 1.25454 1.21005
Proportion of Variance 0.04532 0.04134 0.03754 0.03632 0.03328 0.03148 0.02928
Cumulative Proportion  0.46348 0.50482 0.54235 0.57867 0.61195 0.64343 0.67271
                          PC15   PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     1.18449 1.1489 1.12536 1.04347 1.01842 0.98788 0.96464
Proportion of Variance 0.02806 0.0264 0.02533 0.02178 0.02074 0.01952 0.01861
Cumulative Proportion  0.70077 0.7272 0.75250 0.77427 0.79502 0.81454 0.83315
                          PC22    PC23   PC24    PC25   PC26    PC27    PC28
Standard deviation     0.94120 0.90424 0.8426 0.83099 0.7936 0.74099 0.71343
Proportion of Variance 0.01772 0.01635 0.0142 0.01381 0.0126 0.01098 0.01018
Cumulative Proportion  0.85086 0.86722 0.8814 0.89523 0.9078 0.91880 0.92898
                          PC29    PC30    PC31    PC32    PC33    PC34    PC35
Standard deviation     0.68671 0.66873 0.62157 0.59379 0.54226 0.51834 0.50577
Proportion of Variance 0.00943 0.00894 0.00773 0.00705 0.00588 0.00537 0.00512
Cumulative Proportion  0.93841 0.94736 0.95508 0.96214 0.96802 0.97339 0.97851
                          PC36    PC37    PC38    PC39    PC40    PC41    PC42
Standard deviation     0.43831 0.42681 0.40214 0.35759 0.32958 0.29891 0.26237
Proportion of Variance 0.00384 0.00364 0.00323 0.00256 0.00217 0.00179 0.00138
Cumulative Proportion  0.98235 0.98599 0.98923 0.99178 0.99396 0.99574 0.99712
                          PC43    PC44    PC45    PC46    PC47    PC48     PC49
Standard deviation     0.24441 0.21431 0.14431 0.09695 0.07673 0.04672 0.005841
Proportion of Variance 0.00119 0.00092 0.00042 0.00019 0.00012 0.00004 0.000000
```

```
Cumulative Proportion  0.99831 0.99923 0.99965 0.99984 0.99996 1.00000 1.000000
                            PC50
Standard deviation     7.327e-16
Proportion of Variance 0.000e+00
Cumulative Proportion  1.000e+00
```

After estimating the standard PCA on the simulated dataset, we find that PCA is not effective in significantly reducing the dimensionality of the data. The reason is that the variance is spread out across many components. For example, the first principal component (PC1) only explains around 7.27% of the total variance, and even with the first 9 components, we only capture about 50% of the variance. To explain almost all the variance (close to 100%), we need to use all 50 components.

This is probably because the dataset was generated with random, independent variables that do not have any inherent correlations or patterns. In such cases, no few components can capture most of the data's variance, which makes PCA less useful for dimensionality reduction in this scenario.



Quiz

Question 1

The correct answers are a and c

a. In supervised learning, observations are assigned to predefined categories based on labeled data. Unsupervised learning does not have predefined labels; instead, it finds patterns or clusters without prior knowledge of categories.

c. Supervised learning uses labeled data with known outcomes (ground truth), while unsupervised learning works with unlabeled data, where there is no ground truth to guide the learning process.

Question 2

Correct answers are a and d.

a. PCA helps reveal underlying structures or patterns in the data by identifying principal components, which are combinations of the original variables that capture the most variance.

d. A primary purpose of PCA is to reduce the dimensionality of a dataset by identifying the most important components that explain the majority of the variance.

Question 3.

When selecting the number of clusters (or dimensions) in unsupervised learning, we usually seek to balance two competing forces:

Model complexity (or number of clusters/dimensions): Increasing the number of clusters or dimensions typically improves how well the model fits the data, capturing more details and nuances. However, this can lead to overfitting, where the model starts capturing noise and spurious patterns rather than the true underlying structure.

Simplicity and Interpretatbility: Fewer clusters or dimensions lead to a simpler, more interpretable model that generalizes better to new data. However, this may result in underfitting, where important patterns or structures in the data are missed.

In conlusion, the goal is to find a balance between capturing sufficient structure to accurately represent the data (model complexity) and keeping the model simple enough to avoid overfitting and ensure interpretability (simplicity). This is often done using techniques like the elbow method or analyzing the proportion of variance explained in PCA.

Question 4

The correct answers are b, c, and d.

a. When we lack domain knowledge, following the elbow criterion is generally a good practice, as it helps us rely on quantitative methods. So, this option is not correct. (False)

b. If we have substantial domain knowledge and a clear hypothesis, we can choose a different number of clusters or dimensions than what the elbow criterion suggests. Our prior understanding of the problem might guide us toward a specific number of clusters or components that better fit our hypothesis. (True)

c. If interpretability is not a priority, and our goal is to maximize predictive performance, we may look for more clusters or dimensions than the elbow criterion suggests, focusing instead on the model's predictive accuracy. (True)

d. If our goal is visualization then we may prioritize reducing the number of dimensions to 2 or 3 for easy plotting, even if the elbow criterion suggests more dimensions. This choice is driven by the need for clear, simple visualizations rather than strict adherence to the elbow criterion.