Assessment 4

Joana Levtcheva, CID 01252821

Q1

Data Summary

```
##   danceability        energy          loudness        speechiness     
##  Min.   :0.0000   Min.   :0.00144   Min.   :-41.808   Min.   :0.00000  
##  1st Qu.:0.4260   1st Qu.:0.32900   1st Qu.:-11.881   1st Qu.:0.03490  
##  Median :0.5430   Median :0.63600   Median : -7.385   Median :0.04490  
##  Mean   :0.5388   Mean   :0.57619   Mean   : -9.485   Mean   :0.07002  
##  3rd Qu.:0.6710   3rd Qu.:0.83075   3rd Qu.: -4.694   3rd Qu.:0.07175  
##  Max.   :0.9710   Max.   :0.99800   Max.   :  0.920   Max.   :0.90500  
##   acousticness          valence           tempo       track_genre       
##  Min.   :0.0000015   Min.   :0.0000   Min.   :  0.0   Length:4526       
##  1st Qu.:0.0228000   1st Qu.:0.1762   1st Qu.: 95.7   Class :character  
##  Median :0.2520000   Median :0.3760   Median :122.0   Mode  :character  
##  Mean   :0.3890598   Mean   :0.4073   Mean   :125.4                     
##  3rd Qu.:0.7900000   3rd Qu.:0.6150   3rd Qu.:158.7                     
##  Max.   :0.9960000   Max.   :0.9830   Max.   :214.0                     
```

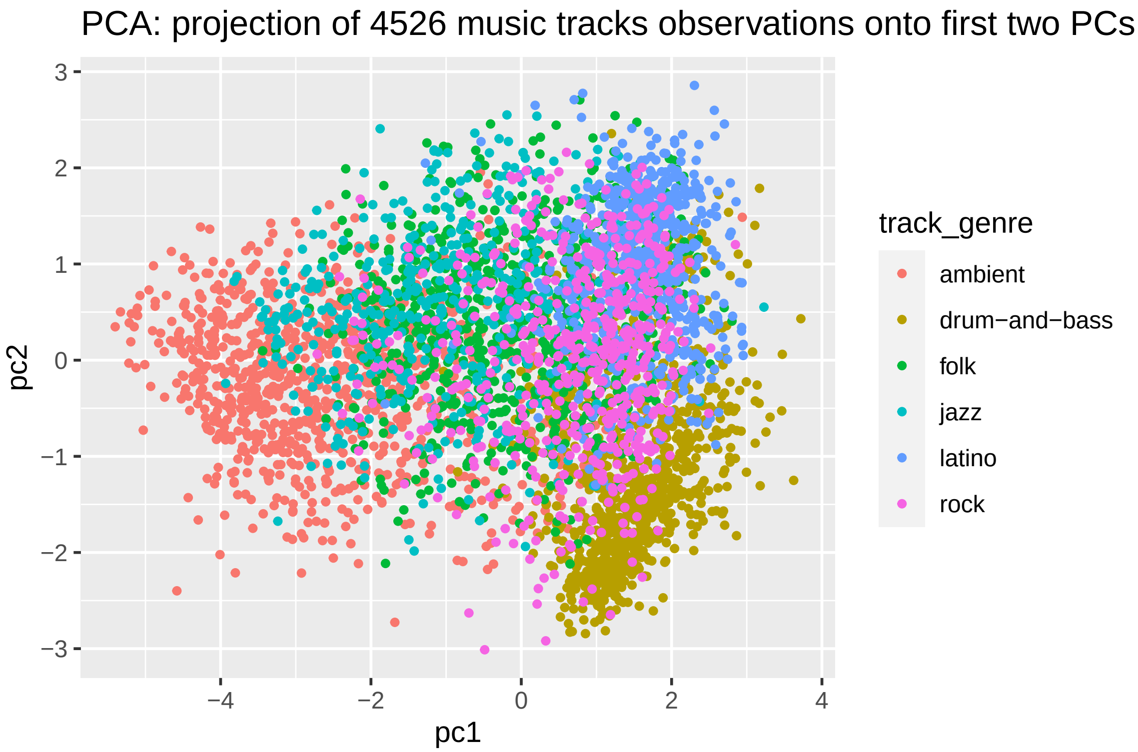
We want to perform principal component analysis on the matrix of feature variables, using the singular value decomposition, so we are going to use the functin `prcomp` which we are going to apply to the scaled matrix of feature variables (leading to having mean 0 for every feature). This is achieved by setting `scale.=TRUE`.

In order to evaluate the proportion of variance in these features that is retained in the first three principal components let's take a look at the summary of the applied PCA:

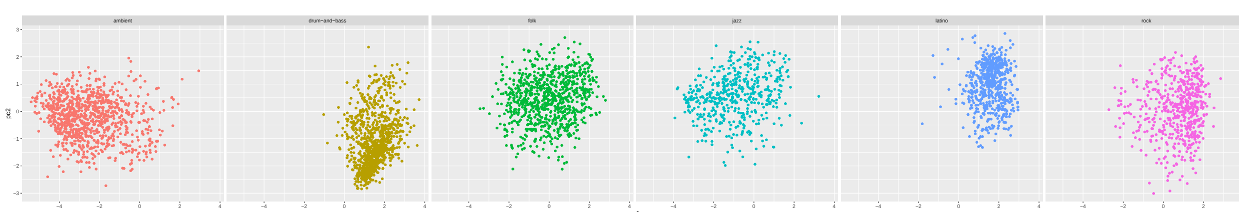
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8589 1.1147 0.9487 0.80804 0.63932 0.49532 0.30809
## Proportion of Variance 0.4936 0.1775 0.1286 0.09328 0.05839 0.03505 0.01356
## Cumulative Proportion  0.4936 0.6712 0.7997 0.89300 0.95139 0.98644 1.00000
```

From the `Cumulative Proportion` we can conclude that the first three components together explain 79.97% of the variability.

Now, let's visualise the projection of the feature matrix onto its first two principal components. First, we are going to visualise the categories in one plot, choosing to represent each data point as a point object, and each point from a category will have its unique colour. As `track_genre` is a discrete variable, by default `ggplot2` will choose the colours for encoding to be equidistant points on an HSL colour circle. Because of the higher density of the data I have chosen to set a smaller size for the points so that there is more visibility if some categories are overlapping.



In the previous plot, given the higher density of the data and the fact that the categories are not very distinct, it is hard to see the exact categories outlines. Also, as we increase the number of colours required, then, we reduce the distinction between the chosen colours. Looking at the plot and legend we can claim that `folk`, `jazz`, and `latino` have colours which can't be easily distinguished, especially when these categories are somewhat overlapping (folk and jazz for example). So, we might want to have a look at every track genre individually. This can be done with a facet plot. It preserves the dimensions of the original plot, and by also plotting them on one row we can still observe the whole picture while allowing for better individual category visibility, and eventually uncovering hidden parts of one category under another one. Moreover, the colours for every category are set to be the same as in the common plot which allows to make a fast connection between the plots.



Q2

Part a

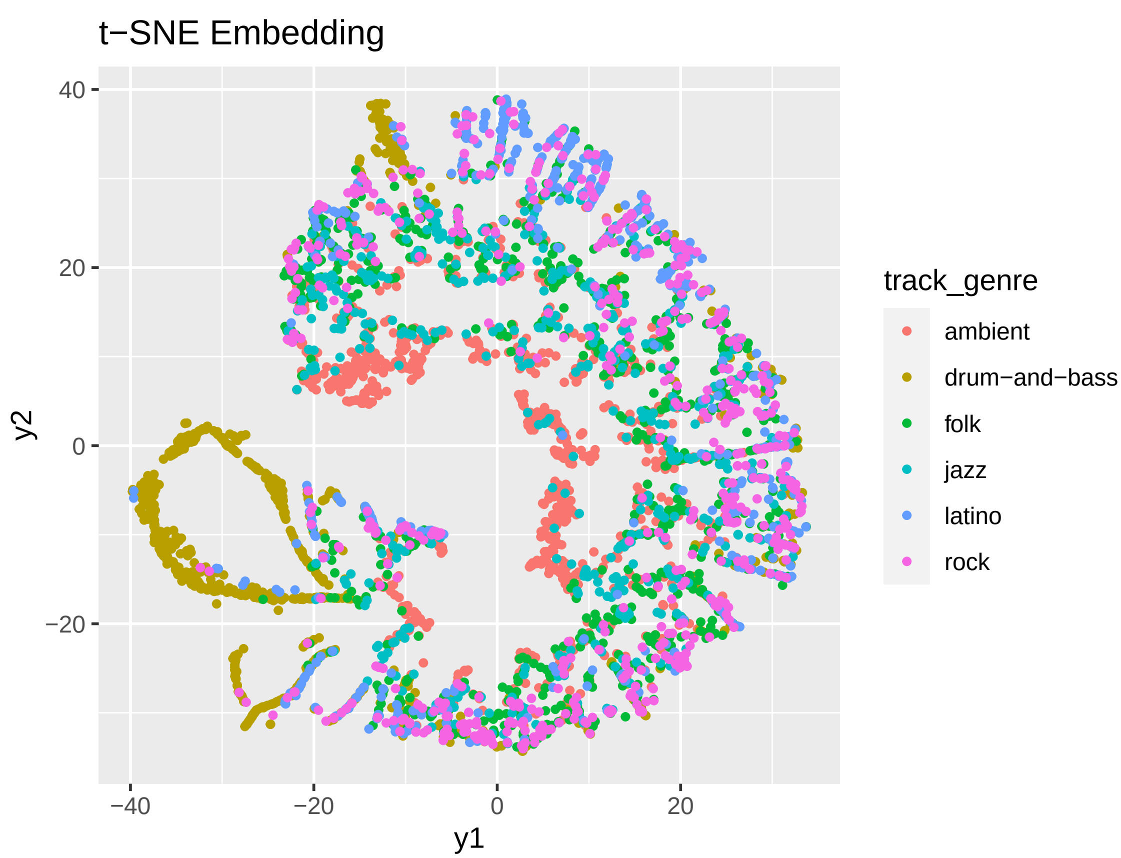
t-distributed stochastic neighbour embedding (t-SNE) is a dimensionality reduction technique that aims to preserve dissimilarity between observations in a high-dimensional dataset.

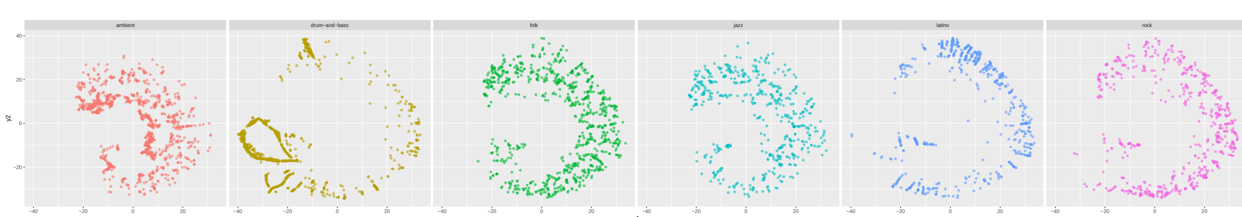
Suppose the dataset has n observations in m-dimensional space. The method works by first constructing the joint discrete probability distribution $P = \{p_{ij}, i, j = 1, .., n\}$, where $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$, with $p_{j|i}$ being the conditional probabilities of point i picking j as its neighbour over the observations in the m-dimensional space, using the Euclidean distances, under a Gaussian centred at point i and a given variance σ_i^2; For a given embedding in $d << m$, we then similarly construct the discrete joint probability distribution $Q := \{q_{ij}, i, j = 1, ..., n, i \neq j\}$, defining the similarity between points in the d-dimensional embedded space using a t-distribution with 1 degree-of-freedom. An optimal embedding is then found by minimising the Kullback-Leibler divergence of the joint probability distribution P from the joint probability distribution Q.

The perplexity parameter is a tunable hyperparameter associated with the construction of the similarity probabilities $p_{ij}, i, j = 1, ..., n$, that balances attention between small-scale and large-scale structure in the original dataset: low values of the perplexity will result in embeddings that preserve the small-scale structure of the data; high values of the perplexity will result in embeddings that focus on dissimilarities in the data over large scales.

Part b

I experimented with perplexity values mainly in the range [5, 50], and I also explored values outside of this range. In my opinion values around 50 produce a better result rather than values signaficantly smaller or larger. Small values of perplexity emphasise more local disimilarities which typically leads to not so strongly defined clusters, and in oure case almost completely fails to divide the genre classes. For larger values than 50 we do not get any further improvements, only increasing the processing time.





The PC projection produces overlapping “round” groups next to each other. The track genres `ambient` and `drum-and-bass` seem to be more distinct than the others. Although they are also overlapping with other genres, these parts are with lower density, and the most distinct parts are with higher density. Latino and rock are overlapping for the bigger part of their areas, as well as there is significant overlapping between jazz and folk.

The t-SNE embedding produces groups in a lunar form which are also overlapping from the outside to the inside. Again, the two more distinct groups are `ambient` and `drum-and-bass`. As in the PC, they are overlapping with other genres bet the higher density of the groups is in the distinct regions. Folk and jazz, as well as again latino and rock are respectively overlapping for most of their areas.

The groups in the PC projection seem to be more well defined and distinguishable even when there is overlapping, compared to the produced groups from the t-SNE embedding which are smaller, thinner, in lunar form, and almost all of them are overlapping.