Scenario

You work at a radio station and your manager gave you the task of doing a deep dive into the chart-topping hits of a chosen artist! You are to create a summary of their most successful tracks played on the station.

Research Goals

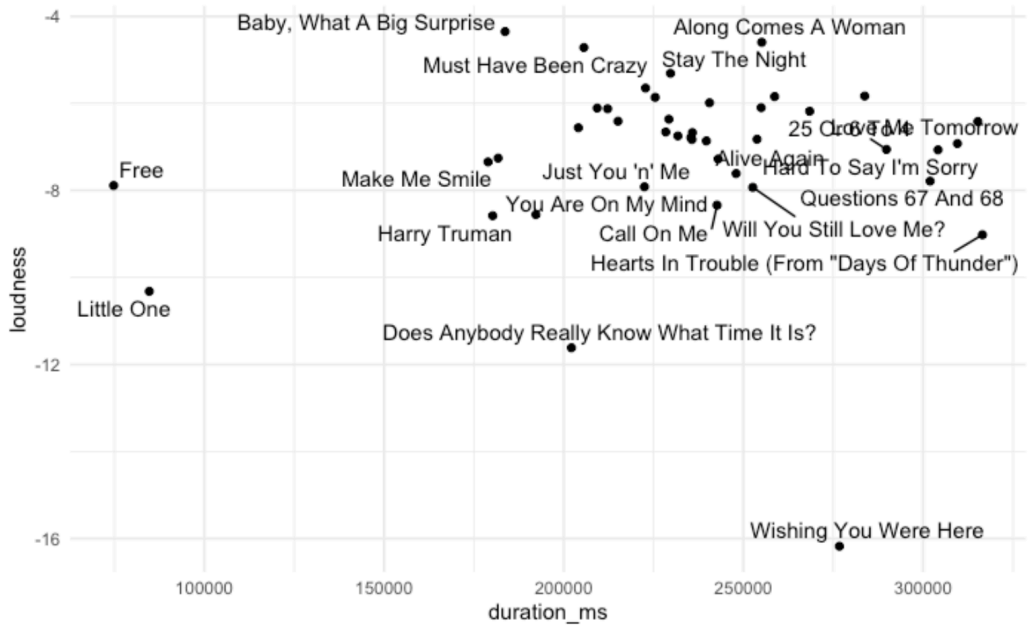
Using Billboard Top 100 data, we are looking to use clustering and principal component analysis to find similarities among Chicago's most popular songs. We are going to study the components of Chicago's most popular songs in order to find what makes their songs so successful and create a summary for our radio station.

Data

The original data we are accessing is a collection of 24,313 songs that have been featured on the Billboard Top 100 Charts. The focus of our data is the 42 Chicago songs that have been featured on the Billboard Top 100 list. In our initial data exploration, we narrowed down some of our variables and decided to deselect the `spotify_popularity` and `billboard_weeks` variables. These two variables are not necessarily "features" of the song that Chicago created, but instead are signifiers of their popularity among radio listeners.

Before running our algorithms, there are a few groups in which the feature variables may be grouped. Potential groups for these predictors include:

1. Musical analytics include duration of the song, key, mode, tempo, time signature, and loudness.
2. Musical features, including danceability, energy, speechiness, acousticness, instrumentalness, liveness, and valence.



According to the scatter plot of duration in milliseconds and loudness, the songs *Free, Little One, Does Anybody Really Know What Time It Is?,* and *Wishing You Were Here* are further apart from other points. This implies that very short and quiet songs are distinguishing features that separate the outliers.

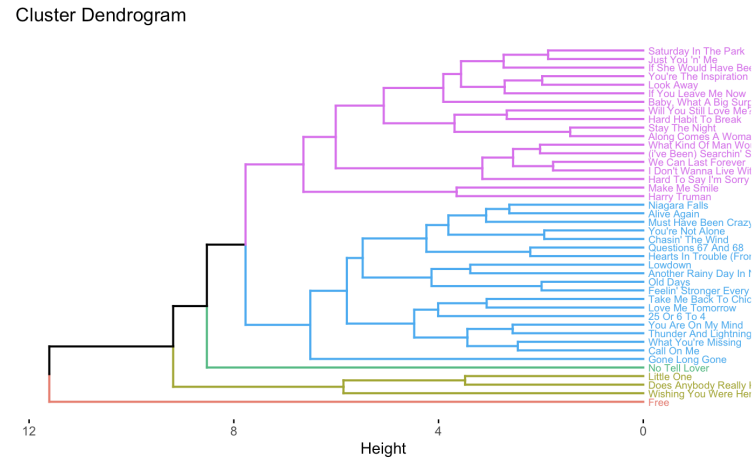
```
my_artist %>%
  ggplot(aes(x = duration_ms, y = loudness)) +
  geom_point() +
  ggrepel::geom_text_repel(aes(label = song)) +
  theme_minimal()
```

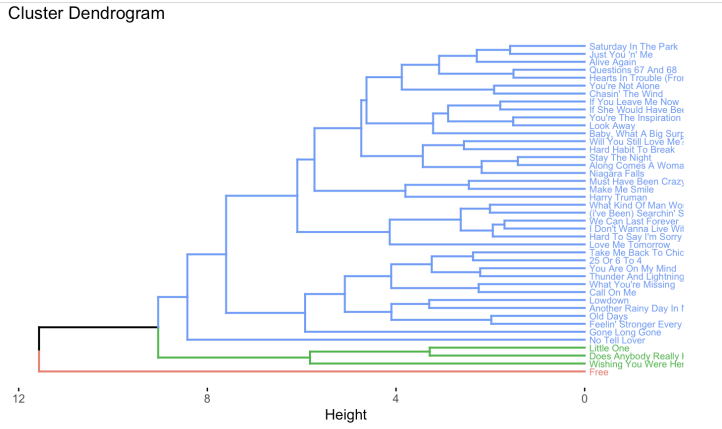
Clustering Analysis

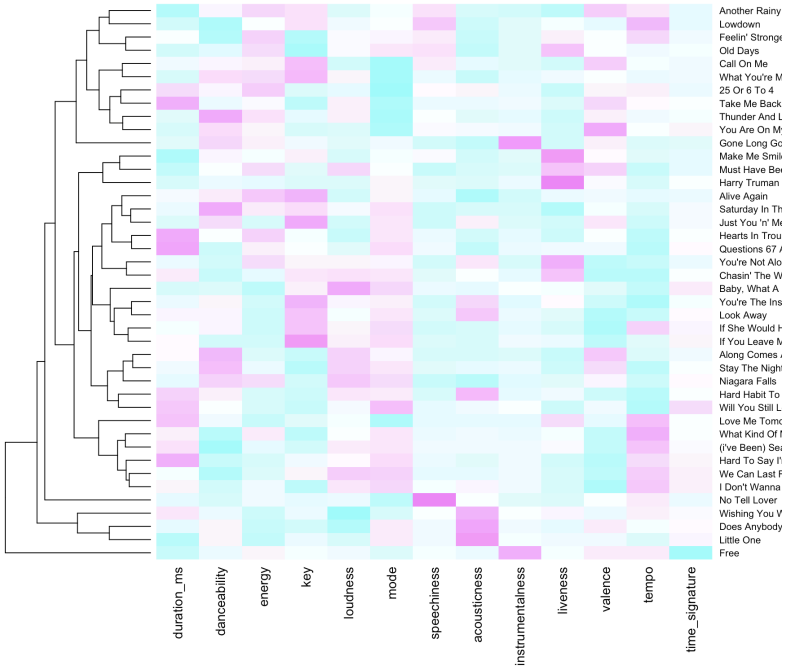
We created a hierarchical clustering model of the band Chicago's most popular songs using complete linkage, which defines the distance between two clusters as the maximum distance between their individual elements. This approach allowed us to examine the relationships between songs based on their musical attributes in a structured and interpretable way. Initially, we identified three clusters, but one was significantly larger than the others. However, one of

these clusters was significantly larger than the other two, suggesting that it might be capturing more variability than is ideal for interpretation.

We found that if we increased the number of clusters to five it would better represent the variability between clusters. When we made this adjustment, the previously large cluster was split into three more specific subclusters. Interestingly, one of these new subclusters contained only a single song, indicating a potential outlier. The other two subclusters were almost evenly split. We used the mean of the variables within each cluster to characterize them. Notably, one of the two larger subclusters has songs with high mode values, while the other contains songs with lower mode values, highlighting a key difference in musical structure.





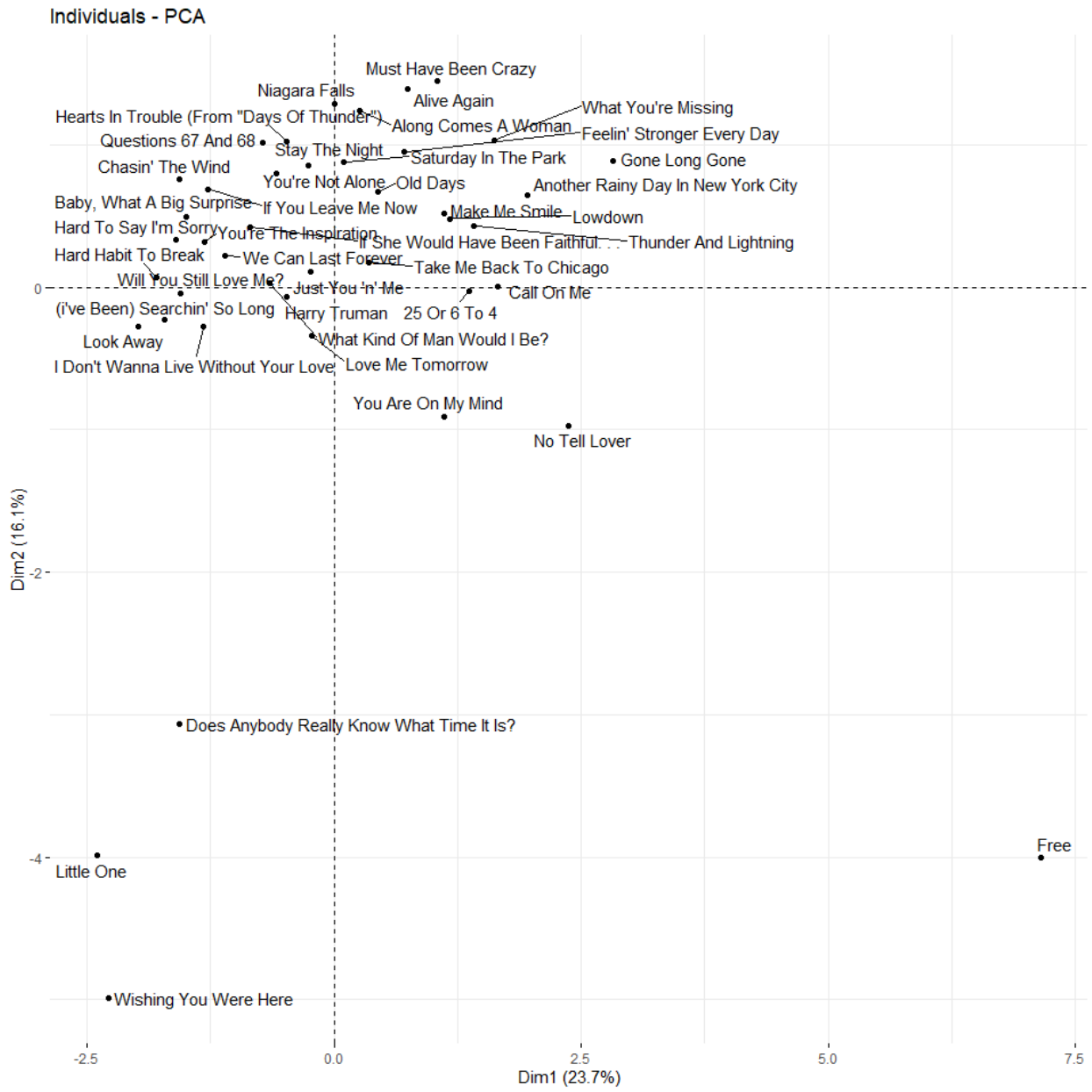


Using the dendrogram and the heatmap, we identified five clusters of Chicago songs based on unique features. Our first cluster contains songs including “If You Leave Me Now”, “You’re The Inspiration”, and “Hard To Say I’m Sorry”. On average, the songs in the first cluster tend to be longer with higher modality, meaning that generally, these songs are in a major key and are “happier” songs. Our second cluster contains songs including “25 Or 6 To 4”, “Feelin’ Stronger Everyday”, and “Old Times”. Songs in the second cluster are higher energy, low modality songs, meaning that these songs tend to be in a minor key and are “sadder” songs. Our third cluster contains three songs: “Does Anybody Really Know What Time It Is?”, “Little One”, and “Wishing You Were Here”. Compared to other Chicago songs, these songs are shorter, with low energy and high acousticness. Our fourth cluster is the song “Free”, a song with high instrumentalness, valence, and energy. Finally, our fifth cluster is the song “No Tell Lover”, which is a song with low instrumentalness but high tempo and speechiness.

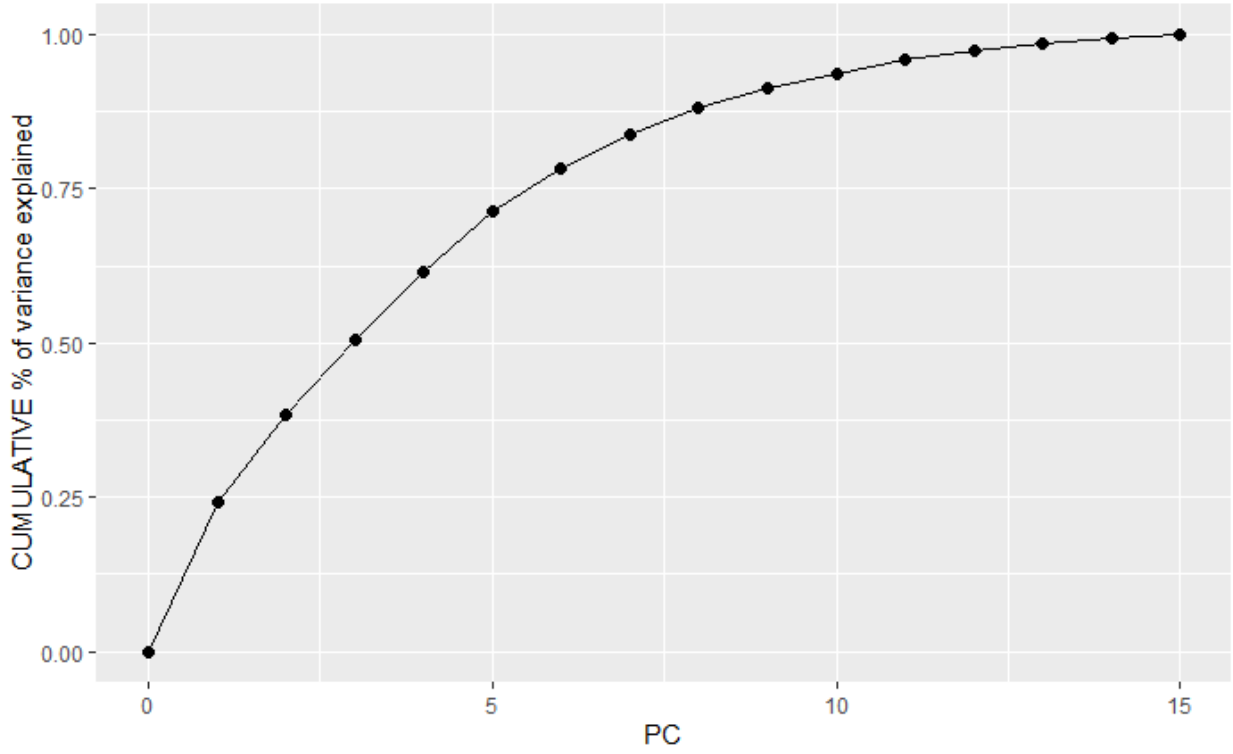
Dimension Reduction

We conducted a principal component analysis (PCA) of our data, which is a machine learning method for reducing dimensions of our data into features that are not correlated with one another, which allows us to identify patterns in the data and lower computational intensity of later models. PCA works by making uncorrelated linear combinations, called principal components (PCs), of the variables in our dataset, where each PC retains unique information from the data. This causes the first few PCs to retain more information than the last few PCs.

If we plot our first two PCs, we get the following score plot.



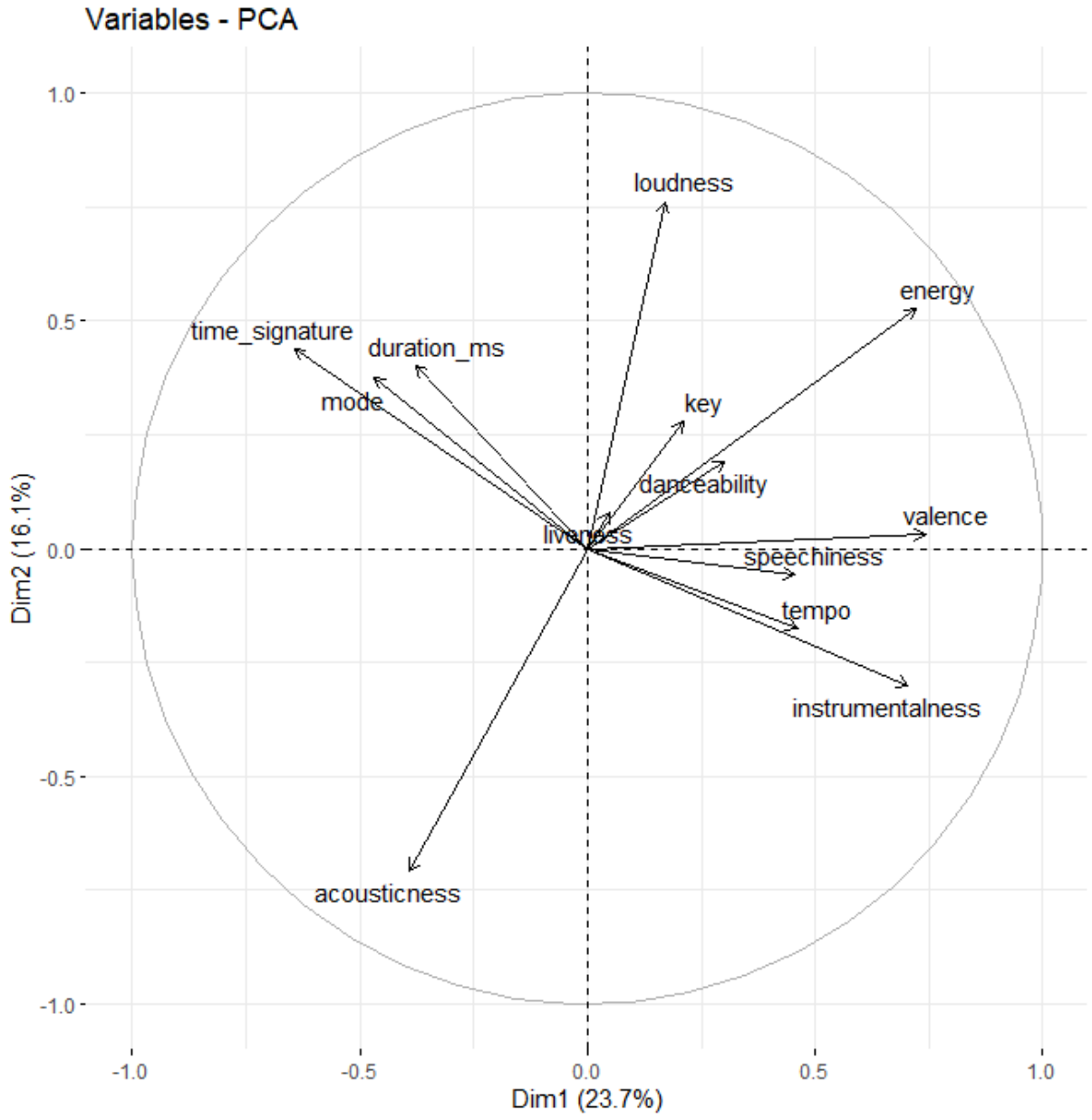
From this score plot, we can see that most of Chicago's songs are relatively similar, at least relative to each other, but there are a few outliers, such as *Free*, *Wishing You Were Here*, *Little One*, and *Does Anybody Really Know What Time It Is?*. It's important to note that this structure is only present using the first two PCs, which cumulatively contain less than 40% of the original variation in the dataset. As shown in our scree plot showing cumulative % of variance explained as a function of PCs, we don't explain more than 75% of the variance until we include 6 PCs.



In general, the clusters we see in our PCA score plot agree with our hierarchical clustering results.

Looking at the loadings plot for the first two PCs, multiple features appear to be highly correlated with one another, indicated by the similar direction for multiple

vectors. An example of this is the strong positive correlation amongst time signature, duration, and mode which are all negatively correlated with tempo and instrumentalness. These strong correlations indicate that we could reduce the number of features, but this claim is based on less than 40% of the original information/variance of the data.



Implementation Subsections

Hierarchical Code:

```
hier_model <- hclust(dist(scale(my_artist)), method = "complete")

cluster_data <- my_artist %>%

  mutate(hier_cluster_k = as.factor(cutree(hier_model, k = 5)))


fviz_dend(hier_model, k = 5, horiz = TRUE, cex = .5)

library(tidyverse)

cluster_data_hier <- my_artist %>%

  mutate(cluster = as.factor(cutree(hier_model, k = 5))) %>%

  group_by(cluster) %>%

  summarize_all(mean)


print(cluster_data_hier)


HEAT MAP


heatmap(scale(data.matrix(my_artist)), Colv = NA, col = cm.colors(256))
```

PCA Code:

```
my_artist <- my_artist %>%

  column_to_rownames("song") %>%

  select(-spotify_popularity, - billboard_weeks)  # remove popularity because
these are already Chicago's most popular songs
```

```
artist_pca <- prcomp(my_artist, scale = TRUE, center = TRUE)  # perform PCA

fviz_pca_ind(artist_pca, repel = TRUE)  # plot PC2 vs PC1

# plot cumulative % explained vs PC

artist_pca %>%

  tidy(matrix = "eigenvalues") %>%

  rbind(0) %>%

  ggplot(aes(y = cumulative, x = PC)) +

    geom_point(size = 2) +

    geom_line() +

    labs(y = "CUMULATIVE % of variance explained")

fviz_pca_var(artist_pca, repel = TRUE)  # loadings plot
```

Conclusions

Write 1–2 paragraphs about your overall takeaways about your data that address your research goals and any limitations of your analyses.

Interpret all evidence in the context of the data, remaining mindful of the target audience of your report.

The research goal is to analyze 42 of the most successful tracks by Chicago and analyze their similarities and differences. Using clustering techniques such as hierarchical and k-means, as well as principal component analysis, we found that 39 of the 42 songs (92.86%) fall into a large cluster, indicating that these songs share similar features. The clustering approach with five clusters painted a better picture of similarities and differences within the dominant cluster. For instance, longer, high-modality songs versus high-energy, low-modality songs.

The dimension reduction approach with PCA yielded an identical conclusion to the three-cluster analysis. It revealed that most of the songs share similar features, as shown in the score plot, with the exception of *Free* and three other

songs. The significant discovery was the weighting of variables in forming two PCs, notably, energy contributes the most while liveness contributes the least in the variance.

Contributions

Describe each student's concrete contribution to this assignment. Please be specific and honest.

At minimum, your summary should include answers to the following questions:

- *Who took the lead on each visualization, method, or section?*
- *Who was responsible for reviewing/revising each visualization, method, or section?*
- *What other roles/responsibilities did you take on with respect to your collaboration?*

Nick – Developed the clustering algorithms and associated visualizations, including dendrograms and heat maps. Wrote the section detailing the implementation of hierarchical clustering methods. Revised research goals and data section.

Holden – Composed the research goals and data summary sections. Contributed to the interpretation of the clustering results. Revised PCA implementation and interpretation.

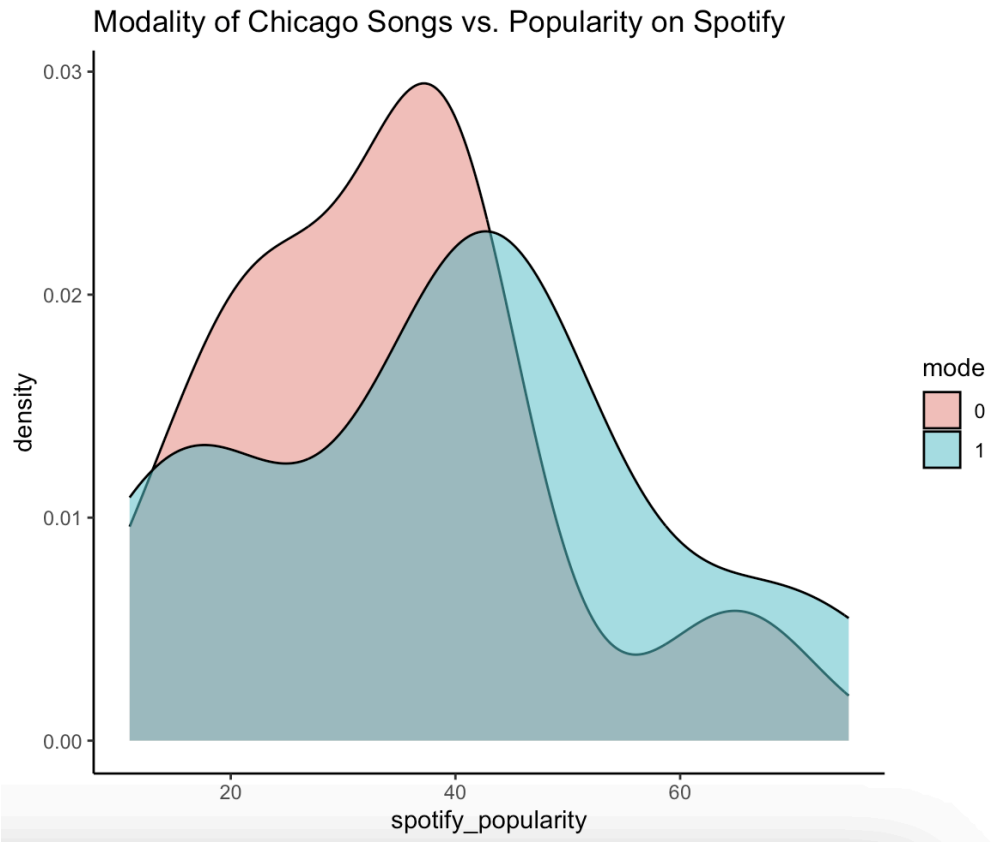
Will – Implemented the Principal Component Analysis (PCA) algorithm and created relevant visualizations. Wrote the sections on PCA implementation and interpretation. Final revision of the report and compiling for submission.

David – Wrote the conclusion and contribution sections. Produced visualizations for the data context section and provided accompanying explanations. Revised clustering implementation and interpretation.

Appendix

Use this section to provide any code, visualizations, etc. that you would like to share as a supplementary resource but were not part of your main narrative above.

I recommend using the `eval: false` *code chunk option in this section so that the code appears but doesn't actually run. This will be particularly helpful in the case of any computationally intensive approaches.*



Although our further exploration of the data will not contain the `spotify_popularity` variable, below is a plot exploring the relationship between the mode and the song's popularity on Spotify. Generally, the mode determines whether a song is in the major key, and tends to be a happier song, or the minor key, and is therefore a sadder song.

Code:

```
my_artist <- my_artist %>%

  mutate(mode = as.factor(mode))
```

```
ggplot(my_artist, aes(x = spotify_popularity, fill = mode))+
  geom_density(alpha = 0.5) +
  theme_classic() +
  labs(title = "Modality of Chicago Songs vs. Popularity on Spotify")
```

When creating the clustering, we also attempted a k-means model. The K-means model provided good insight into the relationship between the clusters and the variables but we chose to stick with hierarchical clustering as it provided a dendrogram that helped us better visualize and interpret the clusters. We found that we were able to get comparable results when taking the cluster means to those in the k-means approach.