IJC445 Data Visualisation Coursework

Topic: Musical Characteristics and Popularity Dynamics in Billboard Hot-100 Songs (2000–2023)

Student Registration Number: 250124697

Word count:3119

Figure 1: Danceability vs Energy by Song Success

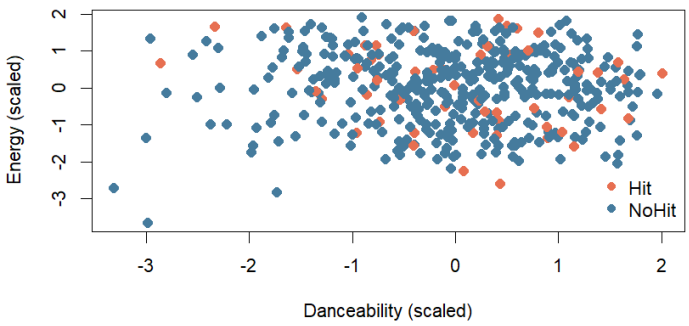


Figure 2: Random Forest Partial Dependence Plots

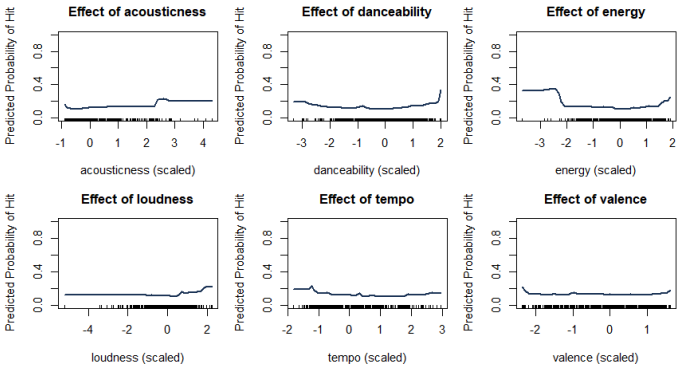


Figure 3: ROC Curve Comparison

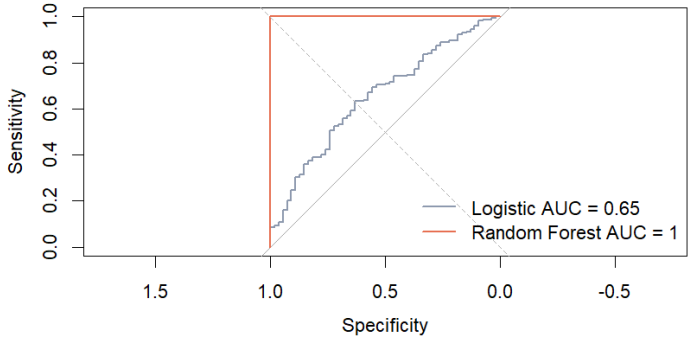
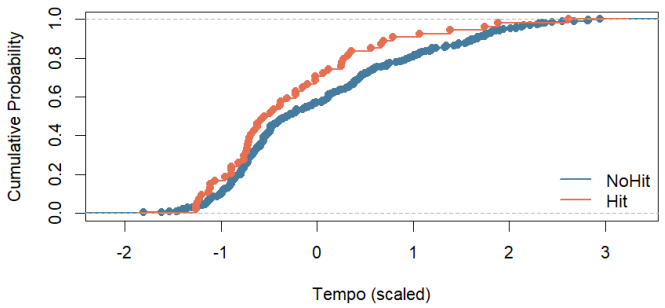


Figure 4: Cumulative Distribution of Tempo by Song Success



Section 1 – Knowledge Building

The growing presence of large music datasets has transformed the analysis of musical success in both academic and music industry contexts. The use of Spotify to measure musical properties via standardized audio features, and the Billboard Hot-100 chart as a widely recognized measure of mainstream success, are two such examples. However, the interaction between measurable audio properties and success on the charts is still a complex and often simplified process. While a large part of the existing literature focuses on accuracy of prediction (Interiano et al., 2018; Pachet & Roy, 2008), defining success as a classification problem, research in visual analytics suggests that prediction alone is insufficient for understanding (Keim et al., 2008; Thomas & Cook, 2006). Visualization is a critical component of sense-making, and particularly so in creative fields where social, cultural, and economic influences also affect outcomes.

The specific question explored by this composite visualization is: "How do the Spotify audio features relate to Billboard Hot-100 Top-10 success, and what can be learned by considering multiple complementary visualizations together?" This question corresponds to the "Ask" level of the ASSERT framework and cannot be answered effectively through a single visualization. Instead, it requires coordinated views that reveal relationships, model behaviour, and distributional patterns simultaneously. Understanding this relationship is significant because simplistic interpretations risk promoting deterministic narratives about musical success.

Figure 1 illustrates an exploratory scatter plot of danceability in opposition to energy. While Top-10 hits are more densely concentrated in areas associated with higher energy and danceability, there is no clear boundary separating them from non-hit songs. This lack of separation is important, as it suggests that these two features alone cannot appropriately explain chart success. The ambiguity in this plot contradicts popular expectations of what makes a piece of music popular and aligns with existing studies that show very limited and context-dependent relationships between audio characteristics and popularity.

Figure 2 further explores by incorporating explanatory insights using Random Forest partial dependence plots. These plots vary from the previous scatter plot in that they show how individual audio features contribute to the predicted probability of success, holding all other features constant. Some features have non-linear effects, with the returns decreasing after a certain level. This suggests that the increase in a particular feature will not have a direct proportionate increase in the chances of success. This also explains the overlap seen in Figure 1.

Figure 3 provides some evaluative insights into the model by comparing the logistic regression and the Random Forest models using the ROC curves. The discriminative power is high for the Random Forest model, suggesting non-linear effects. However, this figure also demonstrates

that improved predictive accuracy does not equate to simple interpretability. Rather, it strengthens the need to combine performance metrics with explanatory visualizations.

Figure 4 provides insight through empirical cumulative distribution functions of tempo. By visualizing entire distributions rather than summary statistics, it reveals subtle but systematic differences between hit and non-hit songs that would otherwise remain hidden.

On a general note, it can be said that the composite visualisation leads to new knowledge where the aspects of ambiguity, non-linearity, predictive evaluation, and distributional nuances. As a result, this leads to an overall understanding of the subject, where it is realized that musical success is linked to complex and interacting combinations of audio features. This highlights the importance of visualisation in the development of new knowledge, rather than merely displaying the findings of the analysis. Perhaps most intriguingly, the development of new knowledge matches the aims of the IJC445 unit, where the role of visualisation is considered to be an analysis method, rather than a communication outcome. Through the gradual development of exploratory, explanatory, evaluative, and distributional perspectives, the analysis moves from surface to deep understanding. This highlights the importance of effective composite visualisations as a means of challenging the status quo, determining the boundaries of analysis, and encouraging reflective thinking, thus illustrating the importance of visualisation as an active agent in the analysis process, rather than a passive display of findings. This is clearly a crucial consideration in the assessment of the narratives developed through data analysis of cultural constructs such as music.

Section 2 – Theoretical Frameworks

In the Ask stage of the ASSERT model, the analysis was informed by a particular and visually centred question, which was: "How do individual Spotify audio features contribute to a song’s chances of appearing in the Billboard Hot-100 Top 10, net of the effects of other features?" This question goes beyond prediction and focuses more on explanation, which in turn necessitates the use of analysis that is visual, rather than merely optional. The focus of analysis in relation to this question also corresponds with the purpose of explanatory visualization, in which the objective is to gain an understanding of the process in which patterns emerge, rather than the potential for prediction that might be achieved. By shifting the focus from classification accuracy to contribution and influence, the analysis also resists the tendency to oversimplify musical success and, instead, situates visualization as a mechanism for structured reasoning and sense-making.

In the Search stage, relevant and reliable data sources had to be selected and identified, which in this analysis were the Spotify audio features and the Billboard Hot-100 rankings, as these are

standard and reliable descriptions of music and are also good indicators of commercial success in mainstream music.

In the Structure stage, data processing had to be carried out in such a way that it allows for easy visualization. In this analysis, this involved combining similar song titles, deleting incomplete data, and making sure that audio features are processed in such a way that it accommodates their varying scales.This step prevents any single feature from dominating the analysis simply because it operates on a larger numerical scale. As a result, the analysis focuses on meaningful relationships between features rather than artefacts of measurement.

In the Explore stage, the Random Forest model was applied since it is appropriate for modeling the non-linear patterns that exist in cultural patterns such as music, where the patterns are not linear or straightforward.The application of the partial dependence plot allows for the exploration of the relationship between the variables and the probability of success, while at the same time considering all the variables. The application of the model ensures that the complexity of the model is taken into account, while at the same time making the model interpretable. A major aspect of this exploration is that the Random Forest model acts as a tool for insight generation and not necessarily a predictive model. Partial dependence plots (Wilkinson, 2005; Wickham, 2010) allow for the exploration of the internal workings of the model, allowing for the determination of thresholds, plateaus, and diminishing returns that would not be observable in the performance of the model.

In the Represent phase, the relationships between the features were visualized using continuous line plots. The plots were chosen because they enable the observation of the trends in the data. Rug plots were added to the horizontal axis to show the distribution of the data points, thus allowing the reader to know the areas of the plot that are based on actual data. From a Grammar of Graphics point of view (Wilkinson, 2005; Wickham, 2010), the visual representation can be characterised as follows: the feature values are represented on the horizontal axis, the predicted probabilities are represented on the vertical axis, and the geometry is a line. The choice of the Cartesian coordinate system and the explicit labels facilitates the choice of visual encoding in a deliberate manner.

Finally, the Tell stage is all about the communication of insight. Looking at the visualisation, it is clear that the influence of audio features on success is non-linear and bounded, with several features having diminishing returns at higher levels. Instead of telling a deterministic story of musical success, the visualisation tells a more complex story in which the audio features are relevant, but only in the context of the whole system.

The use of the ASSERT framework was also iterative, as opposed to linear. This is significant because it reinforces the interdependency of analytical reasoning and visualisation, in that the

results of the analytical process were used to inform the decisions made in the visualisation process. The use of the ASSERT framework as an iterative process also reinforces the role of visualisation in the analytical process.

Section 3 – Accessibility

By the accessibility of the data visualisation, we refer to the ease with which the data visualisation can be perceived, understood, and meaningfully used by various groups of viewers, including the visually impaired, the cognitively impaired, or those with varying levels of data literacy skills (Lundgard & Satyanarayan, 2022; Kim et al., 2021).As has been discussed throughout the module, accessibility is not a choice but a part of effective visualization practices.

This part of the write-up is based on the analysis of Figure 4, which is named "Tempo ECDF." One of the most important benefits of this visualization is the choice of the empirical cumulative distribution function. The function allows the viewer to see the whole distribution and compare it without needing to know complex statistical functions and techniques. This way, the visualisation is made accessible to a wider audience, including those without the technical expertise.

Another benefit of the visualization is the choice of the line geometry, which decreases the level of visual clutter compared to point-dense plots. Lines are very efficient tools for the visualization of the topic, and the cognitive load theory supports the use of lines in the visualization, as has been discussed throughout the lectures.

However, there are accessibility limitations that come with the visualization. The first limitation is that it will not be possible to differentiate between the hit and non-hit songs based on their colors. The visualization will not be accessible to people with color vision deficiency. As a result, it will not be possible to differentiate between the lines. The limitation can be addressed by using redundant coding.

The second limitation is that there is no annotation on the visualization. The visualization does not have text that explains the major differences. As a result, it will only be possible for the viewer to interpret the visualization on their own. Although it is a limitation, it can be addressed by adding annotations.

When analyzing the limitations of the visualization, it is evident that there is an understanding of the literature on accessibility, which is not taken for granted. Accessibility is a continuum, and Figure 4 is a good balance between simplicity and interpretability. The assessment of accessibility not only makes the argument ethical but also methodologically effective.

Furthermore, consideration of the issue of accessibility also raises the issue of the potential for different audiences to interpret a given visualization differently. Even though the visualizations that are created in this type of analysis are intended for an academic audience, they could just as easily be intended for an industry or public presentation situation, where there may be no guarantees of data literacy. In these cases, even a simple visualization could potentially be interpreted incorrectly if the issue of accessibility is not taken into account. The recognition of these potential misinterpretations serves to underscore the ethical imperative of the designer to consider the needs of different audiences. This approach is also in keeping with a more general understanding of the concept of accessibility as a proactive approach rather than a reactive approach.

However, it is worth noting that accessibility goes beyond the visual element of perception and also involves the anticipation of how the visualization would be interpreted by viewers of different data literacy skill levels. Although the visualization in Figure 4 is designed for an academic audience, it would be quite easy to modify it for use in industry or public environments, where it would be unreasonable to assume prior knowledge of ECDFs. In such cases, a lack of explanatory clues or annotations would likely lead to a misunderstanding of the visualization, even if it is technically correct. However, the acknowledgment of this point serves to emphasize the argument that accessibility is not just an issue of accommodating viewers with particular impairments but is also a way of avoiding unnecessary cognitive load on all viewers.

Section 4 – Visualisation Choice

This section will particularly focus on Figure 1: "Danceability vs Energy Scatter Plot" and its justification in terms of the type of visualization used and its relevance to the analysis.

The main aim of using Figure 1 is to explore whether danceability and energy can be used as differentiators for non-hit and hit songs. This kind of analysis can be done using a scatter plot because it can show the relationship between two continuous data sets, and it retains individual data points. This allows for further analysis of correlation, overlap, and clustering in the data.

The retention of individual data points is an important aspect of visualization because ambiguity can sometimes be useful. The overlap of data points for non-hit and hit songs prevents any kind of over-interpretation of results or simplistic thinking. Color coding has been used to highlight differences between songs.

Two kinds of visualisation were considered. One of them, hex bin plot, might help to prevent "overplotting." At the same time, it might hide important individual differences on the level of

songs. Again, since the goal is to visualise ambiguity rather than to avoid it, this option is not so preferable. Another option, density contour plot, might help to visualise distributions on the level of groups. At the same time, it might be less clear to a general audience and hide individual differences, which are important for analytical tasks.

Comparing to the above options, the scatter plot has been identified as the most preferable one. This selection demonstrates an understanding of the need to select a visualisation type depending on the goal of the analysis rather than on convenience. Following classifications of visualisations, scatter plots are recommended for relationship analysis of multivariate data sets.

Yet another important aspect that needs consideration while choosing a visualisation is the nature of reasoning that it will facilitate. The scatter plot in Figure 1 will facilitate exploratory and abductive reasoning, allowing the viewer to formulate their own hypotheses about the data set. This is particularly important in early stages of analysis when ambiguity and uncertainty are beneficial, rather than problematic. More aggregated visualisations, however, may lock in premature and distinct differences and causalities that are unwarranted by the data set. By choosing a visualisation that maintains this ambiguity, honesty in analysis is maintained.

Apart from the importance of clarity and interpretability, the selection of the visualisation will also have an impact on the type of reasoning that is being promoted. The selection of the scatter plot visualisation in Figure 1 supports the use of exploratory and abductive reasoning, as the viewer is able to easily identify the patterns, irregularities, and intersections before drawing an inference. This type of reasoning is critical, especially in the early stages, as ambiguity and unpredictability are valuable and not detrimental. The use of more aggregated data could potentially force the viewer to draw wrong inferences about the clear distinctions and causations that are not being presented. The use of individual data points and intersections maintains the integrity of the analysis and the alignment with the purpose of the analysis.

Section 5 – Implications and Improvements

This section will be dedicated to Figure 3, "ROC Curve Comparison," and will explore some of the ethical issues and how they could be improved.

Figure 3 compares the prediction capabilities of the Logistic Regression and Random Forest models. However, it could be argued that notwithstanding the higher figures of the AUC for the Random Forest model, there are some ethical issues with regard to the prediction capabilities. As was discussed during the lectures, it must be noted that there are no objective data and

visualisations. Although it is true that they could be right from an objective point of view, it is also true that they could be wrong, as it was discussed earlier, particularly with regard to creative industry sectors such as music, which is based on factors such as culture, society, and economics.

Possible Enhancements:

If factors such as genre, marketing strategies, and social media are taken into consideration, it could be more beneficial for the user, as they will be able to get a better understanding of the success.

Interactive visualisations could be used to avoid cognitive overload. Third, accessibility could be improved through redundant encodings and annotation of all figures. Finally, longitudinal analysis could be used to understand how musical trends and feature relevance change over time.

The process of considering implications and improvements serves to reinforce that visualisations are rarely an isolated entity, divorced from their social and organisational context. Once a visualisation has been constructed, it will be subject to a number of possible uses, such as simplification.

For instance, ROC curves and other plots that are performance-oriented might have been selectively used to make claims of predictive certainty when, in fact, the analysis is more complex than what is being presented. The recognition of this point only serves to emphasize the importance of careful design and framing. Recommendations for improvement through the use of additional contextual information, explanations, and other forms of presentation, therefore, not only improve the quality of analysis but also help to ensure that it is done in a more responsible and ethical manner.

By considering the implications and suggestions for improvement, there is an implicit recognition that visualisations can be understood and used in ways other than their original context of analysis. Performance-oriented visualisations such as ROC curves, though valuable in assessing models, can be easily abused in presenting false senses of prediction certainty. (Hand, 2009). Selbst et al. (2019) argue that abstraction in sociotechnical systems, like the reduction of music success to audio features, neglects important social and cultural aspects.This is especially true in areas like music, where there are always cultural, social, and economic aspects at play that cannot be captured by audio features. The awareness of the potential for abuse also relates to the role of the visualisation designer in terms of proper framing and explanation. Furthermore, the designer must also acknowledge that there is the potential for models such as the highly successful Random Forests to obscure the ‘human’ aspect of music. In this manner, by incorporating qualitative metadata such as ‘cultural trend reports’ or ‘marketing spend’ into

future visualisations, it is possible to transcend algorithmic forecasts and towards a more holistic perspective of the socio-technical process of musical popularity. Recommendations for improvement are not only aimed at increasing accuracy in analysis but are also aimed at preventing misuse.

References:

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM, 53*(6), 59–67. https://doi.org/10.1145/1743546.1743567

Knaflic, S. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley.

Munzner, T. (2014). *Visualization analysis and design*. CRC Press.

Buja, A., Cook, D., & Swayne, D. F. (1996). Interactive high-dimensional data visualization. Journal of Computational and Graphical Statistics, 5(1), 78-99.

Cairo, A. (2016). The truthful art: Data, charts, and maps for communication. New Riders.

Correll, M. (2019). Ethical dimensions of visualization research. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13.

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). Spotify teardown: Inside the black box of streaming music. MIT Press.

Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten (2nd ed.). Analytics Press.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis.

Communications of the ACM, 55(4), 45-54.

Huang, W., Eades, P., & Hong, S. H. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. Information Visualization, 8(3), 139-152.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. Royal Society Open Science, 5(5), 171274.

Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In Information visualization (pp. 154-175). Springer.

Kennedy, H., Hill, R. L., Aiello, G., & Allen, W. (2016). The work that visualisation conventions do. Information, Communication & Society, 19(6), 715-735.

Kim, N. W., Joyner, S. C., Riegelhuth, A., & Kim, Y. (2021). Accessible visualization: Design space, opportunities, and challenges. Computer Graphics Forum, 40(3), 173-188.

Lundgard, A., & Satyanarayan, A. (2022). Accessible visualization via natural language descriptions: A four-level model of semantic content. IEEE Transactions on Visualization and Computer Graphics, 28(1), 1073-1083.

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Okabe, M., & Ito, K. (2008). Color universal design (CUD): How to make figures and presentations that are friendly to colorblind people. J Fly Data Info, 1-20.

Pachet, F., & Roy, P. (2008). Hit song science is not yet a science. Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), 355-360.

Prey, R. (2018). Nothing personal: Algorithmic individuation on music streaming platforms. Media, Culture & Society, 40(7), 1086-1100.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68.

Sharpe, L. T., Stockman, A., Jägle, H., & Nathans, J. (1999). Opsin genes, cone photopigments, color vision, and color blindness. In Color vision: From genes to perception (pp. 3-51). Cambridge University Press.

Sweller, J., van Merriënboer, J. J., & Paas, F. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251-296.

Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. IEEE Computer Graphics and Applications, 26(1), 10-13.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

Ware, C. (2020). Information visualization: Perception for design (4th ed.). Morgan Kaufmann.

Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3-28.

Wilkinson, L. (2005). The grammar of graphics (2nd ed.). Springer.

Appendix

```
# IJC445 – Music Data Analysis & Visualisation
# Dataset: Billboard Hot-100 with Spotify Audio Features
#installations
install.packages(c(
  "data.table",
```

```
  "tidyverse",
  "caret",
  "pROC",
  "randomForest"
))

# 0. Clear workspace to avoid conflicts

rm(list = ls())
# 1. Load required libraries

library(data.table)   # Fast data loading
library(tidyverse)    # Data manipulation & plotting
library(caret)        # Machine learning framework
library(pROC)         # ROC & AUC analysis
library(randomForest) # Random Forest model
# 2. Load dataset

bb_raw <- fread("BillboardDataset.csv", encoding = "UTF-8")

# 3. Select relevant columns for analysis

bb <- bb_raw[, c(
  "song",
  "band_singer",
  "ranking",
```

```
  "year",
  "lyrics",
  "danceability",
  "energy",
  "loudness",
  "speechiness",
  "acousticness",
  "instrumentalness",
  "liveness",
  "valence",
  "tempo",
  "duration_ms"
), with = FALSE]

# 4. Rename columns for clarity

bb <- bb %>%
  rename(
    artist = band_singer,
    rank   = ranking
  )

# 5. Define Spotify audio feature variables

audio_features <- c(
  "danceability","energy","loudness","speechiness",
```

```
  "acousticness","instrumentalness",
  "liveness","valence","tempo"
)

# 6. Remove rows with missing audio feature values

bb <- bb %>%
  drop_na(all_of(audio_features))

# 7. Aggregate duplicate song entries
#    - Best (minimum) chart rank
#    - Earliest year
#    - Mean audio features

bb_clean <- bb %>%
  group_by(song, artist) %>%
  summarise(
    rank   = min(rank, na.rm = TRUE),
    year   = min(year, na.rm = TRUE),
    lyrics = first(lyrics),
    across(all_of(audio_features), mean),
    .groups = "drop"
  )

# 8. Create binary success variable
#    Hit = Top 10
```

```
#    NoHit = Outside Top 10

bb_clean <- bb_clean %>%
  mutate(hit = if_else(rank <= 10, "Hit", "NoHit"))

# 9. Standardise audio features

bb_clean <- bb_clean %>%
  mutate(across(all_of(audio_features), scale))

# 10. Create modelling dataset

model_data <- bb_clean[, c("hit", audio_features)]
# 11. Train predictive models

set.seed(123)

# 12. Cross-validation setup
ctrl <- trainControl(
  method = "cv",
  number = 10,
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)

# 13.Logistic Regression model
```

```
logit_model <- train(
  hit ~ .,
  data = model_data,
  method = "glm",
  family = binomial,
  trControl = ctrl,
  metric = "ROC"
)

# 14.Random Forest model
rf_model <- train(
  hit ~ .,
  data = model_data,
  method = "rf",
  trControl = ctrl,
  metric = "ROC",
  tuneLength = 5
)

# 15. Model evaluation using ROC & AUC

roc_logit <- roc(
  model_data$hit,
  predict(logit_model, model_data, type = "prob")[, "Hit"]
)
```

```
roc_rf <- roc(
  model_data$hit,
  predict(rf_model, model_data, type = "prob")[, "Hit"]
)

# 16. Define colour palette

hit_col   <- "#E76F51"
nohit_col <- "#457B9D"

# FIGURE 1: Danceability vs Energy Scatter Plot

cols <- ifelse(model_data$hit == "Hit", hit_col, nohit_col)

plot(
  model_data$danceability,
  model_data$energy,
  col = cols,
  pch = 16,
  cex = 1.2,
  xlab = "Danceability (scaled)",
  ylab = "Energy (scaled)",
  main = "Figure 1: Danceability vs Energy by Song Success"
)

legend(
```

```
  "bottomright",
  legend = c("Hit", "NoHit"),
  col = c(hit_col, nohit_col),
  pch = 16,
  bty = "n"
)

# FIGURE 2: Random Forest Partial Dependence Plots

rf_vis_data <- bb_clean %>%
  select(hit, all_of(audio_features)) %>%
  mutate(across(all_of(audio_features), as.numeric)) %>%
  as.data.frame()

rf_vis_data$hit <- factor(rf_vis_data$hit, levels = c("NoHit", "Hit"))

rf_vis <- randomForest(
  hit ~ .,
  data = rf_vis_data,
  ntree = 500
)

par(
  mfrow = c(2, 3),
  mar = c(4, 4, 3, 1),
  oma = c(0, 0, 2, 0)
```

```
)

for (feat in c("acousticness", "danceability", "energy",
               "loudness", "tempo", "valence")) {

  grid_vals <- seq(
    min(rf_vis_data[[feat]]),
    max(rf_vis_data[[feat]]),
    length.out = 50
  )

  pd_vals <- sapply(grid_vals, function(v) {
    temp <- rf_vis_data
    temp[[feat]] <- v
    mean(predict(rf_vis, temp, type = "prob")[, "Hit"])
  })

  plot(
    grid_vals,
    pd_vals,
    type = "l",
    lwd = 2,
    col = "#1D3557",
    xlab = paste(feat, "(scaled)"),
    ylab = "Predicted Probability of Hit",
    main = paste("Effect of", feat),
```

```
    ylim = c(0, 1)
  )

  rug(rf_vis_data[[feat]])
}

mtext(
  "Figure 2: Random Forest Partial Dependence Plots",
  outer = TRUE,
  cex = 1.2,
  font = 2
)

par(mfrow = c(1, 1))

# FIGURE 3: ROC Curve Comparison

plot(
  roc_logit,
  col = "#8D99AE",
  lwd = 2,
  main = "Figure 3: ROC Curve Comparison"
)

plot(
  roc_rf,
```

```
  col = hit_col,
  lwd = 2,
  add = TRUE
)

abline(a = 0, b = 1, lty = 2, col = "grey70")

legend(
  "bottomright",
  legend = c(
    paste("Logistic AUC =", round(auc(roc_logit), 2)),
    paste("Random Forest AUC =", round(auc(roc_rf), 2))
  ),
  col = c("#8D99AE", hit_col),
  lwd = 2,
  bty = "n"
)

# FIGURE 4: Tempo Cumulative Distribution (ECDF)

plot(
  ecdf(model_data$tempo[model_data$hit == "NoHit"]),
  col = nohit_col,
  lwd = 2,
  main = "Figure 4: Cumulative Distribution of Tempo by Song Success",
  xlab = "Tempo (scaled)",
```

```
  ylab = "Cumulative Probability"
)

lines(
  ecdf(model_data$tempo[model_data$hit == "Hit"]),
  col = hit_col,
  lwd = 2
)

legend(
  "bottomright",
  legend = c("NoHit", "Hit"),
  col = c(nohit_col, hit_col),
  lwd = 2,
  bty = "n"
)
```

GITHUB LINK:-

https://github.com/honeymotwani/IJC-445-Data-Visualisation.git