G8-AIBT

Data Exploration and Visualization

Isaïa Ebongue: Isaia.ebongue@skema.edu
Arthur Savary: Arthur.savary@skema.edu
Roumouz Mohamed Ahmed Roumouz.mohamedahmed@skema.edu

---

1. Introduction

In our exploration of data science applications in human perception and experience, we undertook the analysis of two distinct yet equally fascinating datasets. The first dataset presented us with an intriguing challenge: understanding the intricate relationship between body metrics, skin color, and clothing choices within the Asian demographic.

Our goal was to develop a systematic approach to classify individuals into seasonal color palettes, potentially revolutionizing how we think about personalized fashion recommendations. The second dataset, MuSe, offered us a window into the emotional landscape of music through 90,001 songs, each characterized by three fundamental dimensions: valence (pleasantness), arousal (intensity), and dominance (control). These sentiment metrics, derived from user-generated tags on Last.fm, provided us with a unique opportunity to quantify something as subjective as musical emotion.

While these datasets might seem disparate at first glance, they share a common thread: both attempts to bring structure and understanding to deeply personal human experiences through data analysis. Through careful application of machine learning techniques and statistical analysis, we aimed to uncover actionable insights that could bridge the gap between data-driven decision making and human intuition in both fashion and music domains.

2. Presentation of the Datasets

Dataset 1:

Profile of Body Metrics and Fashion Colors

The first dataset explores the relationship between body metrics, skin color, and clothing color compatibility, focusing on the Asian demographic. It includes features such as height, weight, skin color, shirt, and trouser colors represented in RGB values. While the RGB values for clothing are randomly generated, skin color data is specifically adapted to reflect common tones in the Asian community.

Number of observations: 25 000

Number of variables: 7

The initial dataset provides information on body metrics and clothing colors for individuals, with the following features:

Variable	Description	Type
Height (in centimeters)	Numeric measurement of an individual's height in centimeters.	Quantitative (Continuous)
Weight (in kilograms)	Numeric measurement of an individual's weight in kilograms	Quantitative (Continuous)
Gender	The gender of the individual, categorized as "Male" or "Female."	Qualitative (Categorical)
BMI	Body Mass Index category, classifying individuals as "Underweight" or "Ideal" based on height and weight.	Qualitative (Categorical)
Skin Color	Skin tone represented by RGB values specific to the Asian demographic	Quantitative (Continuous)
Clothes Color	Shirt color represented by randomly generated RGB values.	Quantitative (Continuous)
Pants Color	Trouser color represented by randomly generated RGB values	Quantitative (Continuous)

Objective:

We aimed to determine an individual's seasonal palette (**Autumn, Winter, Summer, or Spring**) based on their skin tone. Using principles of color theory, we categorized various colors into seasonal groups and classified skin tones into four distinct categories: **fair, light, medium, and tan**. Additionally, we introduced a **harmony score**, derived from color theory, to quantify the compatibility between skin tones and seasonal palettes.

Target Variable:

- **Season Palette**: A categorical variable that assigns skin tones to specific seasonal palettes.
- **Harmony Score**: A numerical measure evaluating the alignment between skin tones and seasonal palettes.

Potential Applications :

- Personalized fashion recommendations.
- Tools for stylists or retail applications.

Source of the dataset:

Kaggle Dataset
Rulanugraha, R. "Profile of Body Metrics and Fashion Colors." Kaggle. Available at: https://www.kaggle.com/datasets/rulanugrh/profile-of-body-metrics-and-fashion-colors/data.

Dataset 2:

The Musical Sentiment Dataset (MuSe)

The second dataset, MuSe, includes sentiment data for 90,001 songs, characterized by three dimensions: valence (pleasantness), arousal (intensity), and dominance (control). These sentiment metrics, derived from user-generated tags on Last.fm, offer a basis for studying the emotional impact of music on listeners.

Number of observations: 90 001

Number of variables: 10

Description of Variables :

Variable	Description	Type
Lastfm_url	The Last.fm webpage link for the song, providing metadata or user-generated information about the track.	Qualitative (Categorical)
Track	The title of the song.	Qualitative (Categorical)
Artist	The name of the artist or musical group that performed the song	Qualitative (Categorical)
Seeds	The initial emotion-related keyword(s) used to gather data about the song, indicating the song's emotional tag(s).	Qualitative (Categorical)
Number_of_emotion_tags	The count of distinct words or tags that contributed to the emotion score of the song.	Quantitative (Discrete)
Valence_tags	A numeric score representing the pleasantness of the song, where higher values indicate a more positive emotional tone.	Quantitative (Continuous)
Arousal_tags	A numeric score representing the intensity or energy of the song, where higher values indicate greater emotional excitement.	Quantitative (Continuous)

Dominance tags	A numeric score representing the level of control or influence associated with the song, where higher values indicate greater emotional dominance or power.	Quantitative (Continuous)
Mbid	The unique identifier for the song in the MusicBrainz database.	Qualitative (Categorical)
Spotify_id	The unique identifier for the song in the Spotify catalog.	Qualitative (Categorical)

Objective:

- The aim is to quantify music's emotional influence by analyzing the interplay between valence, arousal, and dominance. Regression analysis will be employed to predict an overall emotion score for each song based on these dimensions.

Potentiel Applications :

- Enhancing music recommendation systems.
- Insights into how music affects emotions during different times of the year.
- Quantitative measures for music therapy or emotional studies.

Source: Cakiki, C. "Muse: The Musical Sentiment Dataset." Kaggle. Available at: https://www.kaggle.com/datasets/cakiki/muse-the-musical-sentiment-dataset

3. Objective and Chosen Approach

Dataset 1 :

The objective of this study is to analyze the relationship between skin tone, clothing colors, and seasonal palettes to explore patterns of color harmony and fashion preferences based on color theory. The data contains variables such as skin tone (derived from RGB values), clothing and pants colors, and categories based on seasonal palettes (Autumn, Winter, Summer, Spring). Additionally, a harmony score was calculated to assess compatibility between skin tones and seasonal palettes.

The chosen approach involves several steps:

1. **Data Processing**: RGB values for skin, clothes, and pants were normalized and converted into HLS values to extract lightness and hue.

2. **Grouping:** Skin tones were categorized as Fair, Light, Medium, or Tan, while colors were grouped into seasonal palettes based on lightness and hue characteristics.

3. **Harmony Scoring**: A matrix based on color theory was used to assign harmony scores for combinations of skin tones and seasonal palettes.

4. **Multivariate Analysis**: Multiple Correspondence Analysis (MCA) was conducted to examine the relationships among categorical variables, including BMI, gender, skin tone, clothing and harmony levels.

Dataset 2:

The aim of this study is to examine the relationship between musical genres, artists, and the emotions evoked by songs, as expressed through the dimensions of arousal, dominance, and valence. The data contains the variables valence (“the pleasantness of a stimulus”), arousal (“the intensity of emotion provoked by a stimulus”) and dominance ("the degree of control exerted by a stimulus"), as well as detailed information about the songs such as, Spotify id, artist, genre the number of seeds (number of emotions generated by the song).

The chosen approach involves several steps:

1. **Data processing:** we only work on a sample of 5000 songs in this case because working on the entire dataset, required too much processing power, and we also modified the dataset by creating id for each artist and each genre to have numerical data that are better suited for our case, and create a new variable “mean emotions”, that is a mean of the value of the arousal, valence and dominance, to make our analysis easier.
2. **Univariate regressions:** we first tried to perform 2 univariate linear models to determine which variable influence the most the emotions. We tried to predict our new variable mean emotion with the genre, after that with the emotions.
3. **Multivariate regressions:** after that we performs 4 multivariate linear models to have the best possible predictions. We first tried to determine the mean emotions variable by using the genre and the artist and after that we tried to predict separately the arousal, dominance and the valence.
4. **Data cleansing:** we filtered the dataset to only keep the songs that have one seed and selected the top group.
5. **Group analysis**: we grouped the songs by their emotion tags, with each group composed of the number of songs by tags and the mean values of valence, arousal and dominance.

4. Data Visualization

Dataset 1:

The analysis revealed distinct patterns in the relationships between skin tones, seasonal palettes, and harmony levels. Key findings include the alignment of **Fair skin tones** with **High harmony levels** and **summer palettes**, while **Tan skin tones** showed stronger compatibility with **Autumn palettes**. Imbalances in the dataset, such as the dominance of Autumn in both clothes and pants palettes, were observed, raising questions about their potential impact on the results and the MCA analysis. The MCA further highlighted clustering patterns, such as the association of "Underweight" BMI with **Low harmony** and **Winter palettes**, and demonstrated the significant variance explained by the first few dimensions, underscoring the need for focused dimensional analysis.

All supporting graphs for this section can be found in the appendix attached to this report, providing detailed visualizations of these findings.

Graph Number	**Description**	**Purpose**	**Interpretation**
1	Bar plot of skin tone distribution.	Visualizes the distribution of skin tones (Fair, Light, Medium, Tan) in the dataset.	Identifies the most prevalent skin tones and highlights the balance or imbalance in skin tone representation.
2	Bar plot of clothes season distribution.	Displays the distribution of clothes seasons (Spring, Summer, Autumn, Winter).	Shows which seasonal palette dominates the dataset, e.g., Autumn appears to be the most common.
3	Bar plot of pants season distribution.	Illustrates the distribution of pants colors classified into seasonal palettes.	Highlights the prevalence of pants seasonal classifications, showing any over- or under-representation of specific seasons.
4	Color harmony matrix (heatmap).	Represents harmony scores between skin tones (Fair, Light, Medium, Tan) and clothes seasons (Spring, Summer, Autumn, Winter).	Highlights which skin tones harmonize best with specific seasonal palettes, based on color theory.

Graph Number	**Description**	**Purpose**	**Interpretation**
5	MCA variable plot for "Clothes Season, Harmony Level, and Skin Tone."	Displays the relationships between categorical variables in the MCA space.	Visualizes the associations and clustering of categories like "Summer" and "High Harmony" or "Autumn" with "Tan."
6	MCA variable plot for "Clothes Season, Pants Season, Harmony Level, and Skin Tone."	Shows the relationships between these variables in the MCA space.	Adds the interaction between pants season and the other variables, revealing more nuanced associations.
7	MCA variable plot for "BMI, Gender, Clothes Season, Pants Season, Harmony Level, and Skin Tone."	Explores all variables in the MCA space to identify the most important relationships and clusters.	Incorporates BMI and gender to examine whether these features add to the clustering of categories in the MCA.
8	Scree plot for "Clothes Season, Harmony Level, and Skin Tone."	Shows the proportion of variance explained by each MCA dimension for this subset of variables.	Highlights the dimensions that explain the majority of variance, e.g., the first few dimensions are most influential.
9	Scree plot for "Clothes Season, Pants Season, Harmony Level, and Skin Tone."	Visualizes the explained variance distribution for these variables.	Demonstrates how adding pants season affects the distribution of variance explained by the MCA dimensions.
10	Scree plot for "BMI, Gender, Clothes Season, Pants Season, Harmony Level, and Skin Tone."	Shows the explained variance for all variables in the dataset included in the MCA.	Indicates the variance distribution when all variables are analyzed, revealing the importance of each dimension in capturing the dataset's complexity.

Dataset 2:

The second dataset focuses on the emotional dimensions of music tracks based on attributes like valence, arousal, dominance, and tags associated with genres, artists, and emotional seeds. It provides a detailed analysis of how these attributes influence emotions and offers visual representations through various graphs. Below is a summary table for each graph, ordered logically based on insights.

Graph number	Title	Description and Interpretation
1	QQ Plots for Residuals	Displays QQ plots for the residuals of the multivariate models (emotions mean, arousal, dominance, and valence). The plots assess the normality of residuals for model quality.
2	Actual vs Predicted Values	Shows scatter plots comparing actual vs predicted values for emotions mean, arousal, dominance, and valence. A red line indicates perfect predictions.
3	Correlation Heatmap	Illustrates the correlations between emotional dimensions (valence, arousal, dominance, and mean emotions). Strong correlations are highlighted with red shades.
4	Valence vs Arousal in Emotion Tags	Highlights the relationship between valence and arousal for different emotion groups. Bubble sizes represent the number of songs in each group.
5	Arousal vs Dominance in Emotion Tags	Explores the relationship between arousal and dominance for different emotion groups. Labels provide clarity for dominant tags.
6	Valence vs Dominance in Emotion Tags	Examines the relationship between valence and dominance for the same emotion groups.
7	Artists with Both Sleazy and Fierce Songs	Highlights artists with both "sleazy" and "fierce" songs, showcasing their average emotional values (arousal and valence). Bubble size represents the number of songs.

Graph number	Title	Description and Interpretation
8	Emotional Dimensions of Music Groups	Combines multiple graphs (valence vs arousal, arousal vs dominance, valence vs dominance) to provide a holistic view of emotional dimensions.

6. Presentation and Interpretation of Results

Dataset 1:

In our analysis of the Profile of Body Metrics and Fashion Colors dataset, we encountered several fascinating challenges that led us to develop innovative approaches for understanding the relationships between skin tones and clothing colors. The initial dataset presented us with an interesting puzzle: raw RGB values without any inherent correlation between clothing and skin colors, as these combinations were randomly generated. We knew we needed to establish a meaningful framework to analyze these relationships effectively.

Our first step involved transforming the RGB values into a more intuitive color space. We opted for HLS (Hue, Lightness, Saturation), which allowed us to classify both skin tones and clothing colors in a way that aligned with human perception. Through this transformation, we categorized skin tones into four groups - Fair, Light, Medium, and Tan - primarily based on their lightness values. Similarly, we organized clothing colors into seasonal palettes - Spring, Summer, Autumn, and Winter - using a combination of hue and lightness characteristics.

Skin tone	*Best Seasonal Palettes*	*Harmony level (in average)*	*Recommendation*
Fair	Spring, Summer	High	Light and bright colors work best.
Light	Spring, Summer	High	Soft and warm tones are most suitable.
Medium	Autumn, Winter	Medium	Earthy and deep shades are ideal.
Tan	Autumn, Winter	Low	Rich, dark colors enhance the tone.

A crucial aspect of our dataset that we carefully considered was its inherent imbalance. The distribution showed a significant overrepresentation of Fair skin tones (12,272 samples) compared to Light (4,655) and Tan (4,743). Similarly, in clothing seasons, Autumn dominated with 8,008 samples, while other seasons had around 4,500 samples each. The pants seasons showed an interesting bi-modal distribution, with Autumn and Winter (6,838 and 6,880 respectively) having significantly more samples than Spring and Summer (3,941 and 4,011). As a group, we made a conscious decision to maintain this imbalance rather than resampling the data. We believed that preserving these distributions, despite their bias, would allow us to analyze the dataset more truthfully and potentially reveal important insights about representation in fashion data.

The Multiple Correspondence Analysis (MCA) results proved particularly illuminating. We conducted three different analyses, progressively increasing the complexity of our variable sets. In the simplest analysis, focusing solely on basic seasonal relationships, we observed a clear separation between warm and cool seasons along the first dimension, explaining 24% of the variance. When we added clothing types, the explained variance decreased to 16.8%, but we gained richer insights into how different clothing categories clustered with specific seasons. In our most complex analysis, including all variables, the first dimension's explained variance dropped to 12.9%, revealing more nuanced relationships between skin tones, clothing types, and seasons.

The correlation heatmap we generated provided particularly actionable insights. For Fair skin tones, we discovered strong harmony scores (0.8-0.9) with Spring and Summer palettes, while Tan skin tones showed higher compatibility (0.8-0.9) with Autumn and Winter collections. Light skin tones demonstrated remarkable versatility, maintaining good harmony scores across multiple seasons, though they peaked with Spring and Summer palettes.

One of the most interesting patterns emerged in the MCA plots: winter and autumn categories consistently clustered together on the left side of the plots, while spring and summer grouped on the right, regardless of how many variables we included. This stability suggests a fundamental distinction between warm and cool seasonal palettes that persists even as we add more complexity to our analysis. The declining variance explained in our MCA dimensions as we added variables (from 24% to 12.9%) reflects the increasing complexity and nuance of our analysis, rather than a limitation.

What makes these findings particularly valuable is their practical applicability. While our initial dataset required significant transformation to yield meaningful results, the final analysis provides a robust framework for understanding color harmony in fashion. These insights can be particularly valuable for fashion retailers, personal stylists, and even in developing AI-driven

fashion recommendation systems. Our analysis provides a data-driven approach to what has traditionally been a largely intuitive process of color matching in fashion.

Perhaps most importantly, our decision to maintain the data imbalance allows us to acknowledge and understand potential biases in fashion data collection and representation. This awareness can inform future research and practical applications, ensuring that fashion recommendations and analyses better serve diverse populations while remaining grounded in the realities of current fashion data landscapes.

Dataset 2:

In our analysis of the MuSe dataset, we embarked on an interesting journey to understand the emotional dimensions of music through data analysis. Working with a substantial dataset of 90,001 songs, we faced our first challenge in managing computational constraints, which led us to focus on a carefully selected sample of 5,000 songs. This decision, while practical, provided us with a manageable dataset that still maintained statistical significance for our analysis.

We began by restructuring the data to better suit our analytical needs. This involved creating numerical identifiers for artists and genres and introducing a new variable - the mean emotions score - which is the three primary emotional dimensions: valence (pleasantness), arousal (intensity), and dominance (control). This pre-processing step was crucial for our subsequent analysis.

Our initial approach involved two univariate regression analyses to understand which factors most strongly influence emotional responses to music. The results were quite revealing when we attempted to predict mean emotions using genre alone, we achieved an R-squared value of 0.20, indicating that genre explains about 20% of the emotional variance. However, when we used artist as the predictor, the R-squared value jumped to 0.75, suggesting that the artist has a much stronger influence on a song's emotional impact than its genre.

Building on these findings, we developed four multivariate regression models that combined both genre and artist information. These models showed impressive improvements in predictive power. The model predicting mean emotions achieved an R-squared of 0.77, while the individual models for valence, arousal, and dominance showed even stronger results, with R-squared values of 0.80, 0.76, and 0.80 respectively. The RMSE values, ranging from 0.35 to 0.55, indicated reasonably good prediction accuracy across all models.

Perhaps the most intriguing part of our analysis emerged when we went deeper into the emotional dimensions of specific songs. By examining cases where songs were tagged with single

emotional descriptors (seeds), we uncovered fascinating patterns. The most frequent emotional tags in our dataset included 'sleazy' (812 songs), 'martial' (583 songs), and 'technical' (530 songs), each associated with distinct emotional profiles. For instance, 'sacred' songs showed high valence (6.95) with moderate arousal (5.0), while 'aggressive' songs displayed low valence (3.08) but high arousal (5.87).

A particularly interesting finding emerged when we noticed an overlap between 'sleazy' and 'fierce' tags in the emotional space. This led us to investigate artists who had songs tagged with both descriptors. We found five such artists, including Kelis, PENTAGONIC, and Ke$ha, each showing unique patterns in how they navigate these emotional territories. Kelis, for example, maintained relatively high valence (5.69) while delivering high arousal (5.70), suggesting a more positive interpretation of these emotional states.

While our regression models didn't achieve perfect prediction accuracy (possibly due to our sample size limitations), they revealed important insights about how artists and genres influence musical emotion. The scatter plots we created for the most frequent emotional keywords provided a visual understanding of how different emotional descriptors cluster in the three-dimensional space of valence, arousal, and dominance.

This analysis opens up exciting possibilities for future research.

6. Conclusion

In examining both the Fashion Season Palette and MuSe datasets, we've uncovered fascinating parallels in how data science can illuminate subjective human experiences - whether in fashion or music. Both analyses demonstrated the power of converting qualitative human perceptions into quantifiable metrics, though each presented unique challenges. In fashion, we successfully mapped the relationship between skin tones and seasonal colour palettes, creating a framework for personalized style recommendations. Similarly, in music, we quantified emotional responses through valence, arousal, and dominance, revealing how artists and genres shape our emotional experiences. While the fashion analysis showed strong correlations between specific skin tones and seasonal palettes, the music analysis demonstrated that artists have a stronger influence on emotional impact than genres alone. Both studies highlight an important truth about human perception: while deeply personal and subjective, these experiences can be meaningfully analysed through data, creating frameworks that bridge the gap between individual preference and systematic understanding. This suggests promising directions for future research in both fields, particularly in developing more sophisticated recommendation systems that account for both objective measurements and subjective human experiences.

7. Data Sources and Consulted References

Rulanugraha, R. "Profile of Body Metrics and Fashion Colors." *Kaggle*. Available at: https://www.kaggle.com/datasets/rulanugrh/profile-of-body-metrics-and-fashion-colors/data.

Cakiki, C. "Muse: The Musical Sentiment Dataset." *Kaggle*. Available at: https://www.kaggle.com/datasets/cakiki/muse-the-musical-sentiment-dataset.

The Concept Wardrobe. "Seasonal Color Analysis: Which Color Season Are You?" Available at: https://theconceptwardrobe.com/colour-analysis-comprehensive-guides/seasonal-color-analysis-which-color-season-are-you.