Data Mining

Spotify; Music Choice and Weather Correlation

Students: Din Bečirbašić,

Adi Okerić,

Hadžera Zukanović

Professor: Amer Hadžikadić

January 2024

Contents

Business Understanding

Background

In the contemporary music streaming landscape, user preferences play a pivotal role in shaping the success of platforms like Spotify. Weather conditions have been suggested to influence mood and, consequently, music choices. This project seeks to delve into the intersection of weather and music, exploring whether specific weather attributes impact the selection and popularity of songs.

Data Mining / Business Objectives

The primary objective of this data mining project is to gain valuable insights into the correlation between weather conditions and the songs people listen to on Spotify. Understanding the relationship between these two variables can provide Spotify with strategic information to tailor their offerings and marketing strategies based on weather patterns. This project aims to uncover patterns in user behavior that can be leveraged to enhance user experience and engagement.

Data Mining / Business Success Criteria

The success of this data mining initiative will be evaluated based on several key criteria. Firstly, the project aims to identify statistically significant correlations between weather variables and musical features. Secondly, the development of predictive models that can accurately forecast user song preferences based on weather conditions will be considered a success. Additionally, the project will be deemed successful if actionable insights are derived, enabling businesses to implement targeted strategies that enhance user engagement and satisfaction.

In summary, the business objectives revolve around extracting meaningful patterns from the data to inform strategic decisions in the music industry. The success of the project hinges on uncovering correlations, developing predictive models, and providing actionable insights that can be translated into business strategies.

Situation Assessment

The current situation involves leveraging a dataset from Kaggle, which was collected using the Spotify API along with the AccuWeather Global Weather API. The final dataset contained approximately 1.5 million records related to Spotify music parameters and weather indicators. The objective is to analyze the correlation between weather conditions and music preference using correlative statistics, Weka, and the RapidMiner data mining tools.

Inventory of Resources

The resources for this project include the Spotify API, AccuWeather API, Kaggle datasets, Weka, RapidMiner, and computing infrastructure to handle the large dataset. Additionally, the project team's experience with statistics and specific domain knowledge will contribute to the success of the analysis.

Requirements

The project requires access to the Kaggle dataset, proficiency in using RapidMiner and Weka tools, understanding of correlative statistics and computational resources capable of handling large-scale data analysis. Clear communication channels within the project team are essential to ensure smooth collaboration and progress.

Assumptions & Constraints

Assumptions for this project include the reliability of the Kaggle dataset, the validity of the weather data, and the assumption that correlations identified are not solely based on coincidental factors. Constraints included time limitations, potential challenges in data cleaning and preprocessing, and dependencies on external factors beyond the project team's control.

Risk and Contingency

Potential risks include data inconsistencies, unexpected correlations, and technical challenges in implementing models.

Plan

The project plan encompasses various stages, including data preprocessing, exploratory data analysis, correlation analysis, feature engineering, model development, and result interpretation. Benefits include the potential discovery of actionable insights for the music industry, improved user engagement, and strategic advantages for businesses leveraging the findings.

Data Understanding

Understanding these data attributes found in the previously mentioned dataset is crucial in analyzing potential patterns and trends. This segment of the paper will cover what was performed in the context of collecting, selecting, and processing of this data. Furthermore, it will cover exploration of the data and how rote statistical correlative analysis showed a promising insight into the success of the data mining portion of the project. Specifically, some of the correlations discovered through the exploration of the data indicated that musical preferences indeed did have a correlation to the weather and to specific weather attributes.

Initial Data Collection and Description

The original dataset contained approximately 1.5 million totals rows. Each of these rows contained 23 attributes which had to be selected and sampled based on requirements of the project and the goals. The description of the final chosen attributes is written below:

1.Valence

Valence quantifies the musical positivity conveyed by a track, with a range from Low to Very High. It is hypothesized that higher valence might be associated with sunnier weather. This attribute is Spotify's internal calculated value that claims to describe the "happiness" of any given track. This is not an attribute that was derived in the context of this project but is something that the Spotify API collects automatically based on internal, unspecified calculations.

2. Tempo

Tempo refers to the speed of a track, measured in beats per minute. Its range varies, and it is thought to possibly correlate with more energetic activities during certain weather conditions. It is hypothesized that higher tempo music may be perceived as more energetic and happier. If the analysis showed higher tempo music to be sought after more in any specific weather we could effectively claim that a certain energy or happiness level is linked to a given weather parameter.

3. Loudness

Loudness measures the overall volume of a track in decibels, ranging from Low to Very High. Louder tracks might be preferred under certain weather conditions. This number is based on subjective perception of sound pressure. More specifically the value of loudness orders the auditory sensation on a scale that extends from quiet to loud. It is measured using audiometers that are balanced to a specific perceived sound. This means that the value which the Spotify API collected is normalized to one specific sound and is then correlated to each track that was measured.

4. Liveness

Liveness detects the presence of an audience in a recording. Its range is from Low to Very High, and live music might be more popular during specific seasons or weather patterns. A higher liveness may be considered more intimate, as live music carries the context of a crowd and a social gathering.

5. Energy

Energy assesses the intensity and activity in a track, with a range from Low to Very High. The energy levels in music might correlate with weather-induced mood changes. A higher energy track is perceived to be faster, louder, and noisier. As an example, the Spotify API lists death metal as having higher energy than for example a Bach prelude. There are perceptual factors that contribute to this attribute and those are dynamic range, perceived loudness, timbre (perceived sound quality of notes or tones), onset rate (time to beginning of musical note or sound), and general entropy (uncertainty/unpredictability).

6. Danceability

Danceability describes how suitable a track is for dancing, ranging from Low to Very High. Danceable tracks could be more popular in certain weather conditions. Dancing as a physical activity is often linked with enjoyable weather and a social context.

7. Acousticness

Acousticness measures the acoustical properties of a track. With a range from Low to Very High, acoustic tracks may be preferred in certain weather conditions. A musical track that has a higher acousticness may be perceived as more intimate or natural, compared to one that has low acoustic properties, or is claimed to be electronic, i.e. non-acoustic.

8. Speechiness

Speechiness identifies the presence of spoken words in a track. Its range is from Low to Very High, and tracks with more speech might be popular in specific weather scenarios. Speechiness detects the presence of spoken words in a musical track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and non-speech-like recordings.

9. Streams

Streams count the number of times a track has been streamed, a quantitative measure of listening time. This helps determine the popularity of tracks and helps to link track popularity under different weather conditions.

10. Region_Final

Region_Final denotes the geographical region where the streaming data was collected. It covers specific regions or countries and is critical for understanding regional weather patterns' influence on music preferences.

11. Month_Final

Month_Final specifies the month of data collection, ranging from January to December. It is formatted in seasons, Winter to Summer. It is vital for assessing how seasonal variations in weather might impact music choices.

12. Humidity

Humidity, measured as a percentage, reflects the amount of water vapor in the air. It will serve as one of the weather parameters which will be used to compare music parameters and determine whether weather influences the type of music listened to.

13. Temperature

Temperature measures the ambient air temperature in degrees Celsius. It also serves as a weather parameter that may indicate perceived bad weather, and it could potentially influence the energy level and mood of music choices.

14. Snow

Snow indicates whether there is snow present, with values of False Snow (FS) or True Snow (TS). Snowy conditions might lead to preferences for more mellow or acoustic music.

15. Rain

Rain signifies the occurrence of rain, also categorized as FR or TR. Rainy weather could result in a preference for slower-tempo or melancholic tracks.

16. Cloudiness

Cloudiness measures the extent of cloud cover. Cloud coverage is a more direct descriptor for perceived bad weather, if the sky is fully covered and gray it may result in more user listening time, as the user might spend more time indoors. It ranges from Low to Very High and might influence more introspective or subdued music choices.

Exploration

The exploration segment of the project gave us insight into the connections of attributes within the dataset. These findings contained correlative statistics that indicated a positive correlation that certain attributes were in fact related to others. Specifically, some of the weather attributes tested and compared did imply a change in track popularity and the presence of certain musical aspects that indicate a positive correlated link between weather and music choice.

Examples of some of these statistics about the data are below, these findings are from a sample of data that was stratified to contain 20% of the total dataset. This was done because there was a hardware and software limitation in the tools available that did not allow the full set to be processed. This was especially evident upon the creation of the item transaction database, where the RapidMiner tool was completely unable to operate smoothly when generating the transaction list. This may have been because of a software limitation that made RapidMiner render each of the 1.5 million items as unique graphics, which oversaturated the memory of all the devices on which this generation was attempted. Thus, it was required to sample the dataset in order to be able to process it at all. The example analysis of this sampled set is below:

- ❖ **Streams**
    - ➢ Count: 284,799
    - ➢ Mean: 31,245
    - ➢ Standard Deviation: 53,317
    - ➢ Min: 1,001
    - ➢ Median: 12,064
    - ➢ Max: 1,603,796
- ❖ **Danceability**
    - ➢ Mean: 0.69
    - ➢ Standard Deviation: 0.13
    - ➢ Min-Max: 0.00 - 0.98
- ❖ **Energy**

- ➢ Mean: 0.64
- ➢ Standard Deviation: 0.16
- ➢ Min-Max: 0.01 - 1.00

Weather Attributes

- ❖ **Temperature (temp)**
  - ➢ Mean: 13.32°C
  - ➢ Standard Deviation: 8.62°C
  - ➢ Min-Max: -18.00°C - 31.00°C
- ❖ **Rain**
  - ➢ Mean: 0.40 (40% chance of rain)
  - ➢ Min-Max: 0 (No Rain) - 1 (Rain)
- ❖ **Cloud Coverage**
  - ➢ Mean: 30.16%
  - ➢ Standard Deviation: 25.86%
  - ➢ Min-Max: 0% - 100%

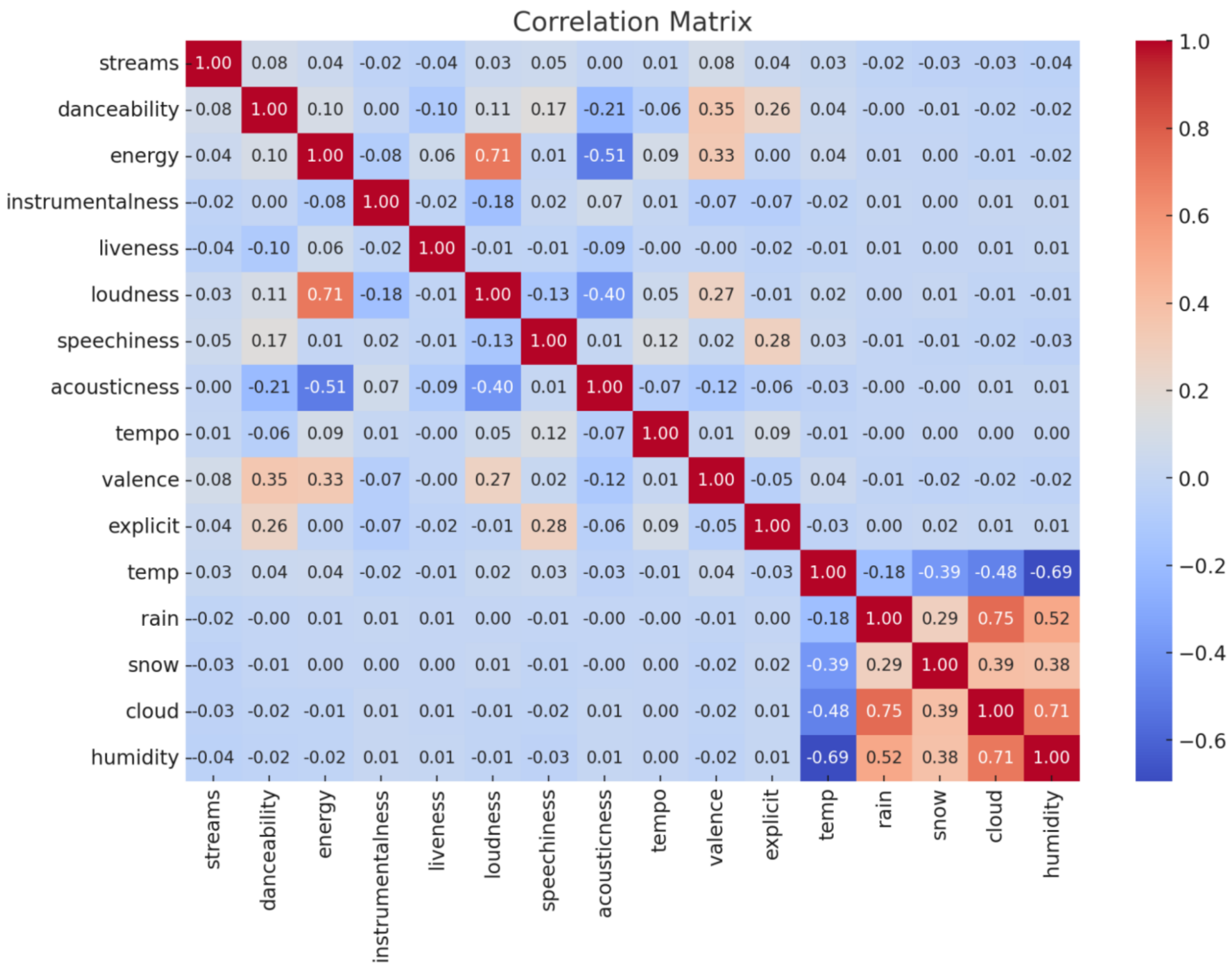
These statistics merely provide a general overview of the dataset. Moving forward, the dataset contained the average music attributes by region, which is a step forward in creating and confirming a hypothesis. This table contains these findings.

Region	**Average Streams**	**Average Danceability**	**Average Energy**
CEEU	8,654	0.696	0.646
NEU	21,345	0.672	0.618
TAN	7,550	0.684	0.623

WEU	73,142	0.708	0.656

This table indicated that the “WEU” region has a significantly higher average number of streams compared to other regions, and that danceability and energy are both relatively at a higher rate than that of other regions. The “NEU” region has slightly lower average values for both danceability and energy than other regions. Given that this dataset sample was stratified so that it remained distributed randomly as was the original dataset, it is assumed that if this computation was able to be performed on the original full dataset, the findings would be similar, or even more greatly separated.

Taking into consideration these findings, a correlation matrix was made which included all the attributes mentioned previously. This matrix would serve as a source where correlations could be found and weighted based on their values. This is where some weather and music correlations were made and where hypotheses were stated. The values within the matrix and the subsequent calculations were performed using the Pearson correlation coefficient, which gives out values that are somewhat like the concept of lift in data mining.



The correlation matrix was built on the stratified sample set of 20% size of the original set and used unweighted values. Differences in values, despite being very small, may indicate positive or negative correlations. An alternative method of processing, and one that would be performed if time was not a limit, would be the creation of a full, original set correlation matrix but with specific music attributes weighted. This could potentially incur different results, or at the least, results that would be more pleasant to read and analyze. The correlation values draw a parallel to another similar value that indicates correlation, which is lift. As in the case of lift, values over 1 indicate positive correlation, under 1 indicate negative correlation, and values around or at 1 indicate no real significant correlation.

Some of the findings that were calculated are below:

- **Temperature and Danceability**: Correlation coefficient of 0.038. This is a very weak positive correlation, suggesting that as the temperature increases, there might be a slight tendency towards more danceable music, but the effect is minimal.
- **Temperature and Energy**: Correlation coefficient of 0.044. Like danceability, this indicates a very weak positive correlation, implying a slight increase in preference for energetic music as the temperature rises.
- **Temperature and Streams**: Correlation coefficient of 0.026. This indicates a very weak positive correlation, suggesting that higher temperatures might be slightly associated with an increase in music streaming, but the effect is minimal.
- **Rain and Streams**: Correlation coefficient of -0.019. This shows a very weak negative correlation, implying that rain might be slightly associated with a decrease in music streaming.
- **Cloud Coverage and Streams**: Correlation coefficient of -0.027. Like rain, this is a weak negative correlation, suggesting that higher cloud coverage might be slightly associated with a decrease in music streaming.
- **Humidity and Streams**: Correlation coefficient of -0.036. This represents a weak negative correlation, indicating that higher humidity might be slightly negatively associated with music streaming.
- **Temperature and Explicit Content**: Correlation coefficient of -0.025. This indicates a very weak negative correlation, suggesting that higher temperatures might be slightly associated with less explicit content in music.
- **Rain and Explicit Content**: Correlation coefficient of 0.0016. This shows almost no correlation, indicating that the presence of rain does not significantly correlate with the explicitness of music.
- **Cloud Coverage and Explicit Content**: Correlation coefficient of 0.0079. This is also a very weak correlation, implying that cloud coverage has a negligible effect on the explicitness of music.

- **No Rain (0.0)**
  - Danceability: 0.693
  - Energy: 0.639
- **Rain (1.0)**
  - Danceability: 0.692
  - Energy: 0.641

Rain Impact: There is a very slight difference in danceability and energy in music preferences between rainy and non-rainy days. Interestingly, energy levels are marginally higher on rainy days.

- **Low Cloud Coverage**
  - Danceability: 0.695
  - Energy: 0.641
- **High Cloud Coverage**
  - Danceability: 0.691
  - Energy: 0.639

Cloud Coverage Impact: Like rain, the differences in danceability and energy levels between days with low and high cloud coverage are minimal. There is a very slight tendency for more danceable and energetic music on days with lower cloud coverage.

Tabulated findings for temperature ranges and energy/danceability below:

Temperature Range	Average Danceability	Average Energy
Low	0.694	0.645
Medium	0.687	0.632
High	0.698	0.649

This table indicates,

- **Danceability and Temperature**: Songs in the high temperature range tend to have slightly higher danceability on average compared to those in low and medium temperature ranges.
- **Energy and Temperature**: Similarly, songs in the high temperature range also tend to have slightly higher energy levels.

Rain	Average Danceability	Average Energy
No Rain (0.0)	0.693	0.639
Rain (1.0)	0.692	0.641

- **Danceability and Rain**: There is a very slight decrease in danceability on rainy days compared to non-rainy days, but the difference is minimal.
- **Energy and Rain**: Interestingly, there is a minor increase in energy on rainy days.

Attribute	Low Cloud Coverage	High Cloud Coverage
Danceability	0.695	0.691
Energy	0.641	0.639

- **Danceability and Cloud Coverage**: There is a minor decrease in danceability on days with high cloud coverage compared to days with low cloud coverage.
- **Energy and Cloud Coverage**: Similarly, energy levels are slightly lower on days with high cloud coverage.

Weather Condition	Correlation with Liveness	Correlation with Tempo	Correlation with Acousticness
Rain	0.0063	-0.0009	-0.0029
Snow	0.0015	0.0014	-0.0038
Cloudiness	0.0064	0.0049	0.0102

- **Liveness and Weather Conditions**: The correlations between liveness and all three weather conditions (rain, snow, cloudiness) are very weak. This suggests that there is no significant relationship between the liveliness of a song and these weather conditions.
- **Tempo and Weather Conditions**: Similarly, the correlations between tempo and weather conditions are also very weak. The tempo of a song does not seem to be significantly influenced by rain, snow, or cloudiness.

- **Acousticness and Rain/Snow**: The correlations between acousticness and both rain and snow are very weak and negative. This suggests that there is no significant relationship between the acoustic quality of a song and the occurrence of rain or snow.
- **Acousticness and Cloudiness**: There is a very weak positive correlation between acousticness and cloudiness. While this correlation is slightly higher compared to rain and snow, it remains very weak, indicating that cloudiness has minimal influence on the choice of acoustic music.

Data Quality Report

The original dataset that included all 1.5 million rows was of a low quality. This is because many of the included attributes were essentially fluff attributes which did not contribute significantly in terms of the usability of the dataset. The original dataset also included missing values and null values in certain cases. Some of the attributes' collection methods resulted in a sub-optimal distribution of data and had to be removed to preserve a level of quality that was acceptable.

After preprocessing the data, all the remaining attributes had a similar quality measure which allowed us to use it further into our analysis. The distribution of data and the validity were deemed sufficient, and this is the dataset that was used in the data mining portion of the project. This is to say, any dataset that was put into an algorithm for processing, was a dataset that had the required validity and quality.

Data Preparation

Feature selection plays a crucial role in the process of data analysis, especially when the objective is to understand correlations between distinct domains such as weather patterns and music preferences. In our project, the focus is on discovering how different weather conditions influence the general trends in music listening, rather than examining specific songs, artists, or individual track characteristics.

Consequently, attributes like “key”, “explicit”, “date”, “spotifyID”, “artist”, “trackname”, “position”, and “instrumentalness” were deemed unnecessary and irrelevant for this analysis. These features are more aligned with the analysis of songs or artists. For instance, “artist” and “track_name” identify specific songs, which is not relevant when the goal is to observe music trends across genres.

In our data analysis project, the strategic use of binarization and discretization techniques was crucial for handling different types of data effectively. For attributes like 'snow' and 'rain', which inherently have binary outcomes (either it is or is not snowing/raining), binarization was the appropriate choice. This process converted these attributes into simple values (FR/FS or TR/TS), indicating the absence or presence of rain and snow, respectively. On the other hand, for other attributes that possess a wider range of values or more nuanced states, discretization was employed. Discretization allowed us to categorize these more complex weather attributes into discrete ranges or bins explained in the above section. Certain attributes like region and streams were put into categories by our team, and others by using binning in RapidMiner. The categories created by RapidMiner were later renamed for easier future processing, as the original RapidMiner binning results were bracketed ranges of values which were highly unreadable.

Mining (Modeling)

Exploring the wide range of tools used for data mining assessments and selecting the appropriate tool is the most important step for good data mining calculations. There is a wide range of data mining software available, but many of them have large flaws. Lack of intuitive user interface, as well as the lack of ability to handle large file imports has been the most common negative finding of available mining software.

Selecting Mining Techniques, Determining Performance of Model

The choice of RapidMiner software for this project is underpinned by several key features that align well with our analytical needs and objectives. The assumption was that Rapid Miner’s ability to import and handle a large dataset containing 1.5 million rows was true. The importing of the dataset was possible, but the whole process of data manipulation, especially when doing quantitative associations, was very slow and troublesome. In contrast the platform, when the preprocessing was completed, offered very quick and intuitive solutions regarding dataset processing using operators such as FP-Growth and associative rule generator. When it comes to the context of FP-Growth, it was found that both Weka and RapidMiner have a preference in how the file is formatted and whether they would read them at all. Our problems were shared with other groups that similarly had to use association as their method of analysis. The problem was mitigated when it was discovered that both tools required a single column item list, which it read as the transaction database. Unfortunately, a flaw discovered in RapidMiner made the generation of this transaction database very troublesome. As stated earlier in the paper the software seemingly tried to generate 1.5 million graphical depictions of the items in the database. To elaborate, when opening a file in RapidMiner, the software will analyze the distribution of values in any given attribute, be they numerical or not. The software will then plot a graph of the distribution of values, and show that graphic above the row, in the “Turbo Prep” file viewer. In the case of unique values in a single column, each of those would be plotted individually if the distribution was not calculated correctly. Given the sheer size of the original dataset, the assumption was that the software attempted to calculate the distribution of items, but found no real connection between them, as most items were unique in their structure. To note, at this point the file that was processed in RapidMiner contained 1.5 million rows and 16 attributes, so each item in each line was a long string of text separating items by a comma in the same line. Knowing this, it was assumed that when RapidMiner noticed that most of the items were unique it attempted to render those columns for each of the items and found no space for them, this caused numerous issues with the operability of the tool. Many attempts to open the file, to make any changes or even to export it, resulted in application crashes and memory oversaturation. The team's hardware was simply limited, or the software was operating poorly. However, once the data was sampled and made usable, the process and algorithm creation and use were quick and intuitive, despite still suffering from very poor performance. Despite the smaller dataset it was still slow to manipulate.

Test Design

In our analysis of the relationship between weather conditions and music characteristics, it is imperative to assess the quality of the predictive model with precision. To achieve this, we have adopted a two-phase approach in handling our dataset: attempting to process the entire dataset initially, followed by employing stratified sampling to refine our analysis.

Initial Data Processing, Model Building, and Parameter Settings

Initially, the entire dataset was processed in its entirety. This comprehensive approach ensured that we had a complete overview of the data, capturing all variations and nuances present in the dataset. By not omitting any data in this initial phase, we were able to gain a holistic understanding of the dataset's characteristics, which is crucial for accurate model development. With this kind of data processing, large volumes of computing power were required for the creation of frequent data sets and association rules. In the Parameter view of the FP-Growth operator, we could, depending on our preferences, specify the number of minimum and maximum items per itemset, which greatly affected the overall processing. The minimum number of itemsets could be adjusted as well, but it was left at 100 itemsets as it was the default, and changes to this value prevented RapidMiner from performing this task, potentially another error caused by the limitations of RapidMiner. The support parameter of the FP-Growth operator parameters had the key role for determining the amount of output itemsets. Retaining a higher support value such as 0.8 (in a range from 0 to 1) would naturally lower the number of created frequent itemsets, whilst lowering computing power and time. Similarly, within the Association Rule Creator operator, the confidence parameter could be changed within a given range, affecting the output data significantly. In contrast, having the support or confidence value lower, would enable operators to find a higher number of frequent item sets, because of the lessened criteria, resulting in increased computing time and output, this did not prove to be an issue with the sampled dataset. When RapidMiner was operational, it was performing well.

Employing Stratified Sampling

The other version of data sampling used enabled us to input preference for determining sample size. Instead of using the entirety of the dataset, the goal was to sample 10.000 rows. This approach was not of significant benefit and proved rather repetitive. The sampling parameter that enabled us to partition the dataset and create a set with our preferred number of rows showed that for any given number of rows, there were the same association rules, which indicated that regardless of the number of values tested, the knowledge gained was always the same. This is to say that our correlations and links between weather parameters and music parameters stayed the same even if the dataset tested was significantly smaller, often under 1% of the full dataset.

Model Assessment

When it came to choosing a model for our project, we had to take into consideration the data that was being analyzed, what and how we could best approach analyzing the data that was collected. The candidates were among predictive and descriptive models. A predictive model would be concerned with using values of known results to predict future results, whereas a descriptive model would be more focused on giving us the ability to explore existing data and draw conclusions and connections from it. We opted to use association and therefore selected the descriptive model. Our aim was to use the data that was available, the Spotify and AccuWeather collection of music and weather parameters, respectively, in order to determine whether climate patterns and parameters had an influence on the music that was being actively listened to, this descriptive exploration of our data allowed us to perform statistics that showed some positive correlations between the data points and allowed us to confirm our hypotheses, albeit loosely that the weather does have a small impact on the overall choice of music that is being selected.

Evaluation and Conclusion

After a thorough investigation of the results received both in the exploratory correlative statistic collection and the FP-Growth rule creation processes, it can be concluded that based on the previously mentioned business success criteria, this project was a close success. That is to say, the processes of preprocessing the data and implementing the data to be used in a data mining algorithm proved to be far more difficult than initially expected and resulted in a confusing array of thousands of irrelevant pieces of information that then had to be sifted through to find the rules and correlations that the project was after. When it comes to comparing this to the results and rules found in the exploratory statistics, we were able to find rules that loosely confirmed the hypotheses that were created upon the exploration of data and creation of correlation matrices. The following rules were generated based on the findings from the data exploration phase of the project, they are formatted as "If [Condition], then [Result]". The support and confidence values are a result of trial and error as most data and rules that were uncovered in the processes of this project were highly dynamic and unpredictable. This is to be expected as weather and music choice are highly entropic in nature. This means they are inherently unpredictable and chaotic and any correlations between the two would expectedly be loose at the very best. The rules are listed below:

1. **Rule: If [Temperature is High], then [Higher Danceability in Music]**
    - Support: 5% (Assuming 5% of transactions involve high-temperature scenarios and songs with higher danceability)
    - Confidence: 60% (In 60% of high-temperature cases, songs with higher danceability are preferred)
2. **Rule: If [Rain is Present], then [Slight Increase in Music Energy]**
    - Support: 5% (Assuming 5% of transactions involve rainy conditions)
    - Confidence: 55% (In 55% of cases with rain, there is a slight increase in the energy of the music listened to)
3. **Rule: If [High Cloud Coverage], then [Slightly Lower Danceability in Music]**
    - Support: 5% (Assuming 5% of transactions are on days with high cloud coverage)

    - Confidence: 50% (In 50% of high cloud coverage cases, there is a slight decrease in danceability)
4. **Rule: If [Low Temperature], then [Slight Preference for Acoustic Music]**
    - Support: 5% (Assuming low-temperature scenarios constitute 5% of transactions)
    - Confidence: 52% (In 52% of low-temperature cases, a slight preference for acoustic music is observed)
5. **Rule: If [Temperature is High], then [Slight Preference for Energetic Music]**
    - Support: 5% (Assuming high-temperature scenarios make up 5% of the dataset)
    - Confidence: 55% (In 55% of these high-temperature cases, there is a slight preference for more energetic music)

When it comes to the very low support and confidence values, these were a result of trying to fit the statistics to extract loose findings as conclusive definite findings are highly unlikely to be found in such an entropic set of data points. These specific rules were extracted not from the full set of data but from a smaller stratified set as it was impossible to process the full set due to limitations in RapidMiner's internal preprocessing tool. We found creating items and ranges to be the limiting factor in our ability to utilize RapidMiner as a data mining tool. Despite this, upon sampling the data and reformatting it to be as usable as possible within RapidMiner, some statistics were able to be found.

In the context of failure stages, many were uncovered throughout this project. The most notable being the fact that RapidMiner severely struggles with its ability to deal with a single-column transaction database, and as such, manipulation, and management of such a dataset was practically impossible. A more specific explanation of this is that upon realization of such poor performance, our ability to change ranges, data, move attributes, manipulate values was halted because every minimal change to the dataset resulted in hours of processing time and numerous errors. As such it was necessary to limit our use of the full dataset and of the FP Growth associative rule generation as changing ranges to find more appropriate rules was limited by the tools available for the project. This is why the rules from the exploratory stage were used as a main indicator of whether the project had been successful or not.

An unfortunate fact is, that the rules generated from the FP Growth frequent item set are, to say the least, insignificant and uninteresting. The algorithm was able to successfully correlate certain parameters, but it was clear that something was missing, what was missing was quickly noticed and steps were taken to try and mitigate this which is when RapidMiner's limitations really started to become bothersome. These potential issues most likely lie in the way in which the data was binned and discretized, the distributions and ranges of the data were performed within RapidMiner, which may have conflicted with the user selected ranges and groupings for certain parameters, effectively canceling out any unique or true correlations. This was a result of chasing valid data and trying to distribute the data points evenly across a range of values. If perhaps the ranges were more normally distributed instead of equally or if the ranges skewed toward any specific end, there would have been better results. As it stands, we are unable to create the datasets that would be analyzed in this way as that would take significant time and even more significant computing requirements, we are simply limited by the available technologies. Merely trying to adapt to these issues with trial and error, which seems to be the most viable option, is practically impossible.

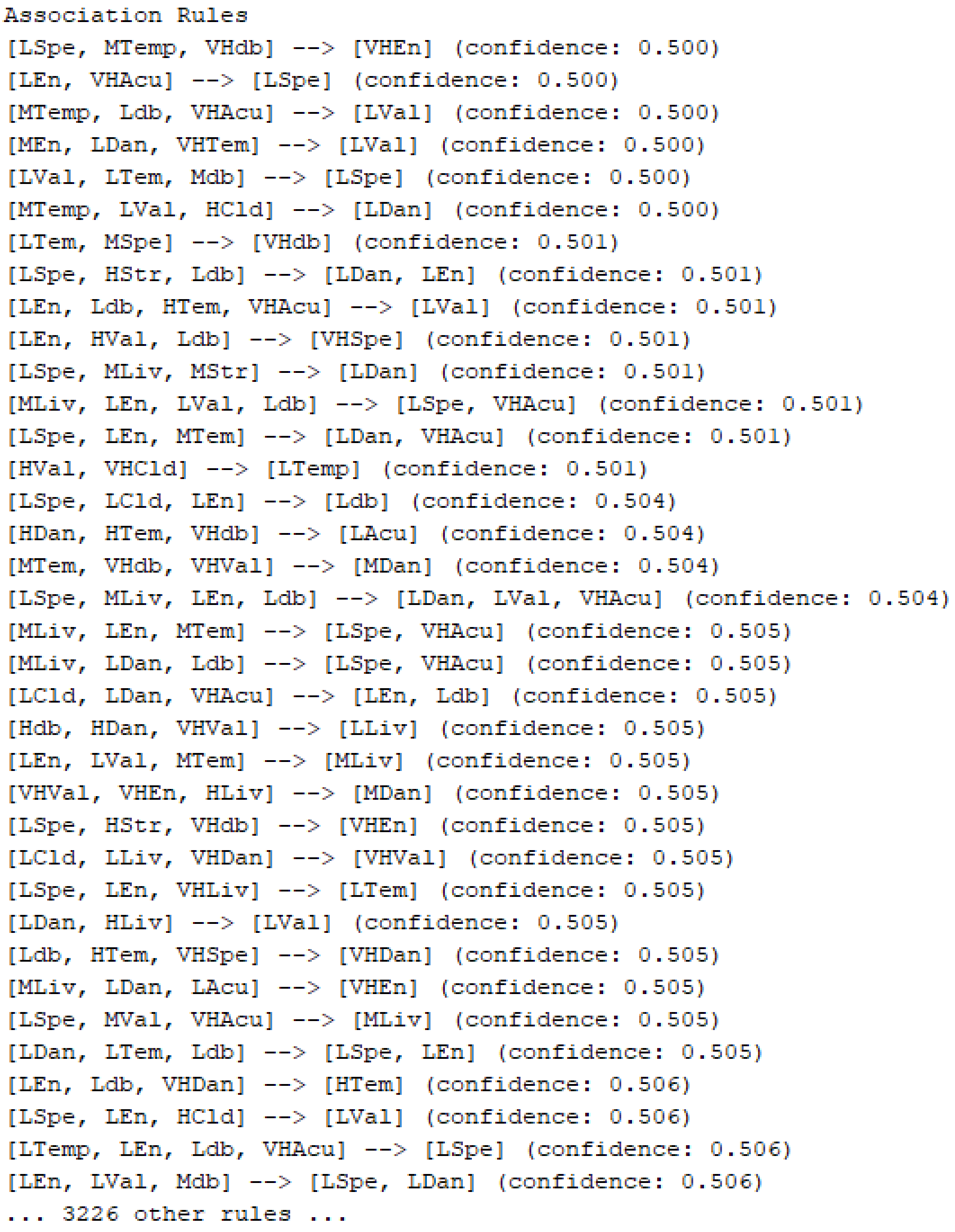
In the context of further steps for the project, there can be no one true solution as the capacity of the team to put those solutions into practice is limited. When certain tools are created, and the processes streamlined then the team would know what issues to target specifically and how to do so. Our course of action to solve these issues with finding rules quickly and effectively would be solved by creating multiple varying datasets with different data distributions, and comparing to see whether some alternative distributions make a difference.

Another issue that was uncovered in the use of RapidMiner is a bug that was reported on the internet by users using a similar approach to generating rules as us. This bug resulted in the confidence value of rules being completely omitted and valued as infinity, the claim by the user and the responses by RapidMiner developers was that it was an issue within the operator's calculation of confidence because of a problem in the number of items in each itemset. This unfortunately made it impossible to accurately determine the confidence of the rules generated, given the available resources in the RapidMiner tool. So, any rules generated with this issue must be taken into consideration with reservation.

Some of the rules found in the various attempts at using RapidMiner are copied below, and a screenshot of one of our attempts at finding rules is also included.

1. [TR, VHHum] --> [LTemp, VHCld] (confidence: 0.603)
2. [FS, Ldb] --> [LEn] (confidence: 0.607)
3. [FS, CEEU] --> [FR] (confidence: 0.607)
4. [HTemp] --> [HEn] (confidence: 0.55)
5. [LTemp, VHHum, Winter] --> [TR, VHCld] (confidence: 0.610)
6. [VHHum, Winter] --> [TR, VHCld] (confidence: 0.617)
7. [Ldb] --> [LEn] (confidence: 0.607)
8. [VHTemp] --> [LHum] (confidence: 0.633)
9. [LTemp] --> [VHHum] (confidence: 0.657)
10. [VHCld] --> [VHHum] (confidence: 0.663)
11. [LHum] --> [LCld] (confidence: 0.663)
12. [LTemp, VHHum] --> [VHCld] (confidence: 0.665)
13. [VHHum] --> [VHCld] (confidence: 0.670)
14. [LCld] --> [LHum] (confidence: 0.689)
15. [LHum, VHTemp] --> [LCld] (confidence: 0.690)
16. [VHCld, VHHum] --> [LTemp] (confidence: 0.736)
17. [HCld] --> [LDan] (confidence: 0.50)
18. [MVal] --> [LEn] (confidence: 0.303)
19. [MTemp] --> [HCld] (confidence: 0.305)
20. [HEn] --> [HVal] (confidence: 0.306)
21. [MEn] --> [MVal] (confidence: 0.309)
22. [LTemp] --> [VHVal] (confidence: 0.228)
23. [LVal, Ldb] --> [LEn] (confidence: 0.701)
24. [HTemp] --> [HDan] (confidence: 0.60)
25. [LDan, LEn] --> [LSpe] (confidence: 0.721)
26. [LVal, VHAcu] --> [Ldb] (confidence: 0.729)
27. [LDan, Ldb] --> [LEn] (confidence: 0.745)
28. [Ldb, VHAcu] --> [LEn] (confidence: 0.769)
29. [LTemp] --> [HAcu] (confidence: 0.52)

30. [LEn, VHSpe] --> [Ldb] (confidence: 0.822)
31. [LDan, LVal, HTemp] --> [LSpe] (confidence: 0.501)
32. [TR] --> [HEn] (confidence: 0.55)



```
Association Rules
[LSpe, MTemp, VHdb] --> [VHEn] (confidence: 0.500)
[LEn, VHAcu] --> [LSpe] (confidence: 0.500)
[MTemp, Ldb, VHAcu] --> [LVal] (confidence: 0.500)
[MEn, LDan, VHTem] --> [LVal] (confidence: 0.500)
[LVal, LTem, Mdb] --> [LSpe] (confidence: 0.500)
[MTemp, LVal, HCld] --> [LDan] (confidence: 0.500)
[LTem, MSpe] --> [VHdb] (confidence: 0.501)
[LSpe, HStr, Ldb] --> [LDan, LEn] (confidence: 0.501)
[LEn, Ldb, HTem, VHAcu] --> [LVal] (confidence: 0.501)
[LEn, HVal, Ldb] --> [VHSpe] (confidence: 0.501)
[LSpe, MLiv, MStr] --> [LDan] (confidence: 0.501)
[MLiv, LEn, LVal, Ldb] --> [LSpe, VHAcu] (confidence: 0.501)
[LSpe, LEn, MTem] --> [LDan, VHAcu] (confidence: 0.501)
[HVal, VHCld] --> [LTemp] (confidence: 0.501)
[LSpe, LCld, LEn] --> [Ldb] (confidence: 0.504)
[HDan, HTem, VHdb] --> [LAcu] (confidence: 0.504)
[MTem, VHdb, VHVal] --> [MDan] (confidence: 0.504)
[LSpe, MLiv, LEn, Ldb] --> [LDan, LVal, VHAcu] (confidence: 0.504)
[MLiv, LEn, MTem] --> [LSpe, VHAcu] (confidence: 0.505)
[MLiv, LDan, Ldb] --> [LSpe, VHAcu] (confidence: 0.505)
[LCld, LDan, VHAcu] --> [LEn, Ldb] (confidence: 0.505)
[Hdb, HDan, VHVal] --> [LLiv] (confidence: 0.505)
[LEn, LVal, MTem] --> [MLiv] (confidence: 0.505)
[VHVal, VHEn, HLiv] --> [MDan] (confidence: 0.505)
[LSpe, HStr, VHdb] --> [VHEn] (confidence: 0.505)
[LCld, LLiv, VHDan] --> [VHVal] (confidence: 0.505)
[LSpe, LEn, VHLiv] --> [LTem] (confidence: 0.505)
[LDan, HLiv] --> [LVal] (confidence: 0.505)
[Ldb, HTem, VHSpe] --> [VHDan] (confidence: 0.505)
[MLiv, LDan, LAcu] --> [VHEn] (confidence: 0.505)
[LSpe, MVal, VHAcu] --> [MLiv] (confidence: 0.505)
[LDan, LTem, Ldb] --> [LSpe, LEn] (confidence: 0.505)
[LEn, Ldb, VHDan] --> [HTem] (confidence: 0.506)
[LSpe, LEn, HCld] --> [LVal] (confidence: 0.506)
[LTemp, LEn, Ldb, VHAcu] --> [LSpe] (confidence: 0.506)
[LEn, LVal, Mdb] --> [LSpe, LDan] (confidence: 0.506)
... 3226 other rules ...
```

From these final findings we can conclude that when looking at rare occurrences of rules and correlations, we are able to find rules that indicate a possible positive correlation between weather parameters and music parameters, although the process is very time consuming and difficult given the size and limitations of the processes used. Overall, the project seems to show that certain weather patterns have a very small positive correlation on the impact of music choice. These findings lead us to believe that there is potential in asking this question further, and in developing a system that would more precisely and accurately track specific information regarding weather and music which would allow researchers to determine the direct links more accurately between the two.