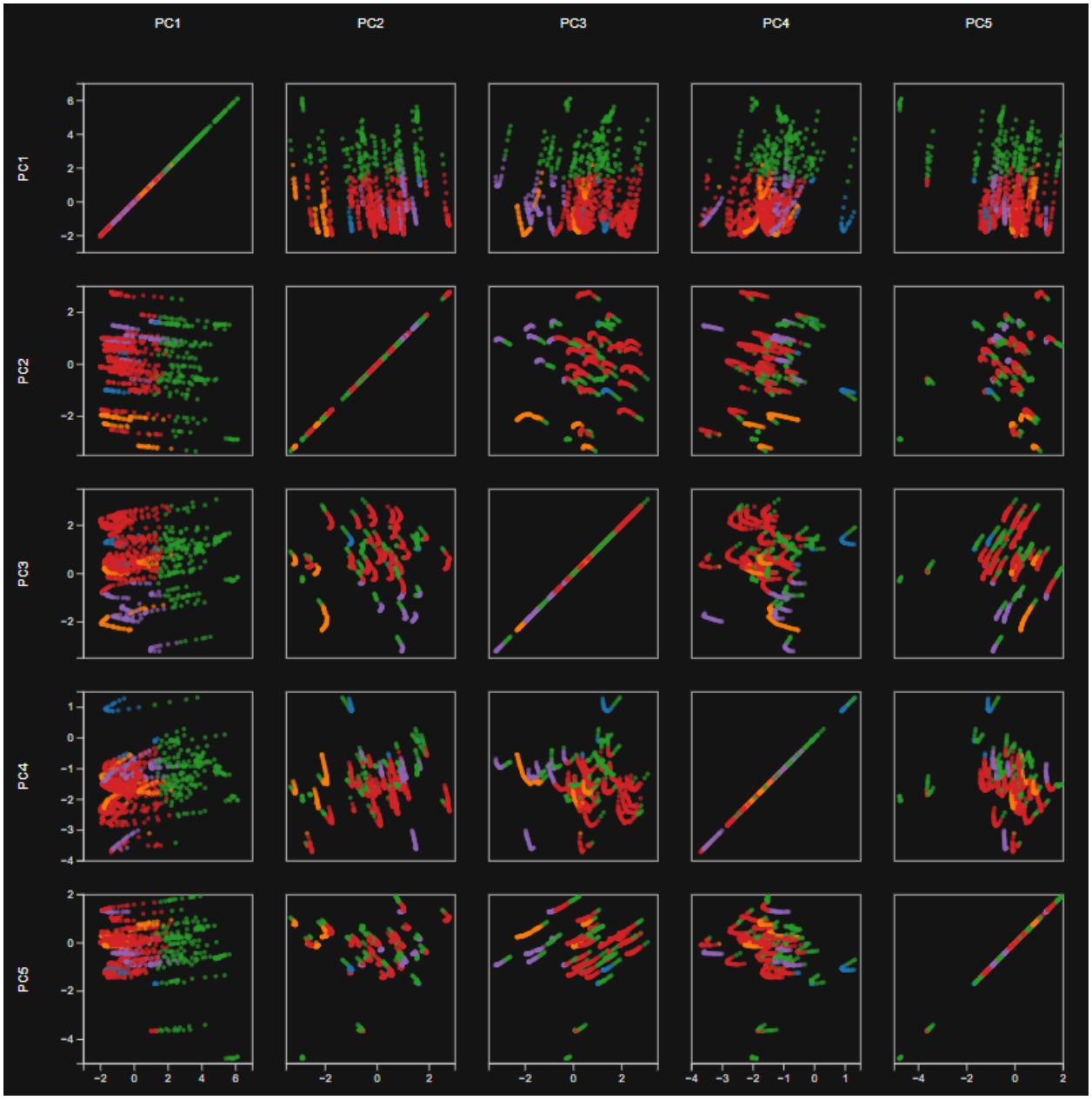
Spotify and Billboard Data Visualization

ScatterPlot Matrix Observations:

- Discrete Variables (e.g., mode, key):
    - The mode variable is binary (0 or 1), with a relatively balanced distribution.
    - The key variable, which represents musical key (0–11), shows a fairly uniform distribution, indicating no particular dominance of any specific key in the dataset.
- Continuous Variables (e.g., speechiness, liveness, valence):
    - Variables like speechiness and liveness are skewed, with most songs having low values for these features. This suggests that highly "spoken" or "live" tracks are less common in the dataset.
    - valence, which measures musical positivity, appears to follow a normal distribution, with most songs having mid-range positivity.
- Valence vs. Danceability:
    - There is a noticeable positive correlation between valence and danceability. Songs that are more "danceable" tend to have higher positivity (happier tone).
- Duration vs. Other Variables:
    - The duration of songs (duration_ms) does not show strong correlations with other variables like valence or danceability.
    - Most songs cluster within a specific range of durations, likely reflecting typical song lengths in popular music.
- Weeks-on-Board vs. Liveness:
    - Songs that remain on the Billboard charts for longer (weeks-on-board) tend to have lower values for liveness. This may suggest that highly "live" tracks are less likely to achieve sustained popularity.
- A clearer focus on fewer variables reveals distinct trends:
    - Songs with higher values for weeks-on-board tend to cluster around specific keys, suggesting certain musical keys may be more commercially successful.
    - The distribution of continuous variables like liveness and speechiness is easier to interpret.
- Trends like the positive correlation between valence and danceability become more evident.
- Distributions of key features like duration_ms and valence remain consistent with observations from other matrices.

PCA Biplot Observations

- For this plot:



- PC1 and PC2 dominate the variance:
    - The spread along PC1 is significant, indicating that this component captures the most variability in the dataset.
- Hypothesis:
    - In this Spotify-Billboard dataset, PC1 **likely** reflects a combination of key musical features (e.g., danceability, valence, speechiness) that broadly differentiate songs based on their overall popularity or "hit" characteristics.
    - Features like danceability, valence, and energy are often correlated in music datasets. These correlations result in a large proportion of shared variance, which is captured by PC169.
    -

    - For example, songs that are highly danceable may also tend to have high valence (positivity), creating a strong shared trend.
- PC2 also contributes substantially but less than PC1, suggesting it captures secondary patterns or relationships.
- Subsequent PCs (PC3, PC4, PC5) show decreasing variance, as expected in PCA. These components capture finer details or noise in the data.
- The separation between clusters is more apparent along higher-order PCs (e.g., PC3 vs. PC4), indicating nuanced groupings.
- Hypothesis:
    - Each subsequent principal component captures progressively less variance. By the time we reach PCs like PC3 or PC4, they represent finer details or noise in the data rather than dominant trends.
    - These PCs may reveal subtle groupings based on niche characteristics (e.g., specific genres or production techniques) rather than broad patterns.