Project Report 1

Zesen Zhuang

February 1, 2023

Task 1

Metric	**Value**
Min	0
Average	234,408.55
Max	10,435,467

Table 1: Statistics for original data

Table 1 shows the statistics for the original data.

Metric	**Value**
Min	107,582
Average	226,899.35
Max	349,583

Table 2: Statistics for processed data

Table 2 shows the statistics for the data after removing the outliers.

Outliers detection

q1	**median**	**q3**	**iqr**
198333.0	224840.0	258834.0	60501.0

Table 3: Quantiles for the original data

Table 3 shows the quantiles for the original data. The interquartile range (IQR) is calculated as $q3 - q1 = 258834.0 - 198333.0 = 60501.0$. The lower and upper bounds are calculated as $q1 - 1.5 \times iqr = 198333.0 - 1.5 \times 60501.0 = 107582.5$ and $q3 + 1.5 \times iqr = 258834.0 + 1.5 \times 60501.0 = 349583.5$. The outliers are defined as the data points that are less than the lower bound or greater than the upper bound. Finally, there are **559989** songs removed.

Task 2

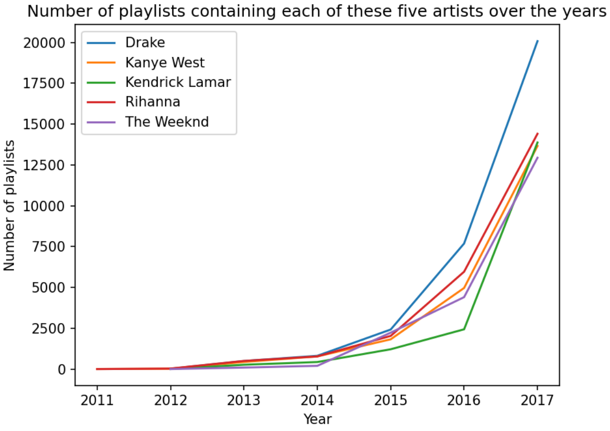

Figure 1: Top 5 artists

Figure 1 shows the number of playlists containing each of the top 5 artists over the years. The top 5 artists are **Drake**, **Kanye West**, **Kendrick Lamar**, **Rihanna**, and **The Weeknd**. **Drake** is the most popular artist, and all artist have a significant increase in the number of playlists containing them after 2016.

Task 3

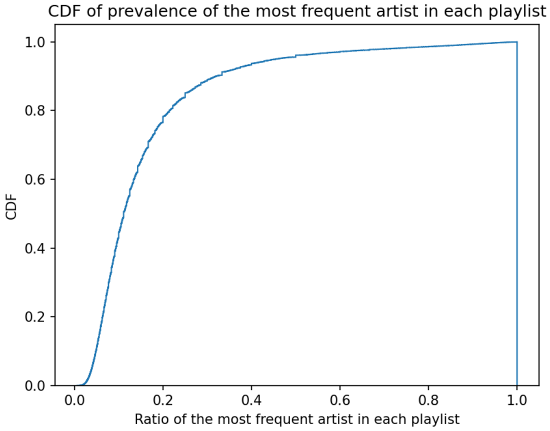

Figure 2: Artist prevalence CDF

Figure 2 shows the CDF of the artist prevalence. The artist prevalence is defined as the fraction of songs by the most frequent artist. The result shows that most playlists have a low artist prevalence (less than 0.5), which means that the playlists are more diverse.