1 Introduction

This project considers 3 different techniques for time series data analysis. They are clustering, anomaly detection and forecasting. The main motivation of choosing time series is my personal experience in a real time data analytics and real time anomaly detection as a software engineer who delivers distributed systems in a big data domain. For me, it has been always interesting what do data scientist do to develop models which can work with different metrics or time series.
As a dataset was chosen spotify dataset which is available in kaggle [1]. This dataset has the following columns: track id, name, country, date, positions, streams, artists, artists genres, duration, explicit.
In general this result can be used for music streaming platforms to see how amount of listeners are increasing or decreasing. It can help to analyze and predict revenue or detect competitors.

2 Data preparation

Initially plan was to analyze what music genres get streamed over time period. After several attempts faced with some problems, is that was quite complicated to normalize genres for each data record. Example of a genre feature: ['taiwan pop', 'zhongguo feng', 'mandopop', 'c-pop']. It was not that complicated to retrieve a main genre or give some genre for very specific genres to avoid problem with shor time series. Cleaning code was excepting too much time, and for this reason country feature was selected. As a result, for time series was selected sum of streams for specific countries. A country with date and streams gave good enough data which is possible to use for analysis. Since spotify had different number if countries over different periods, it was decided to fill such countries with 0 streams in periods when countries were not used.

name	country	date	position	streams	artists
Crackküche	de	2021-04-15	82	625718	['Haftbefehl']
WATER	jp	2019-01-31	171	50896	['Suchmos']
Gözleri Aşka Gülen	tr	2018-11-15	59	185439	['Nilipek.']
Gözleri Aşka Gülen	tr	2018-11-22	133	111159	['Nilipek.']
Gözleri Aşka Gülen	tr	2018-11-29	166	96204	['Nilipek.']
Gözleri Aşka Gülen	tr	2018-12-06	184	90088	['Nilipek.']

Figure 1: Example of dataset.

3 Time Series Clustering

In this section two different clustering techniques are provided. On the Figure 2 it can be clearly seen how well clusters united is subgroups. Couple outliers like ad (Andorrda), ru kr, ua are in their own branches, since these countries have joined Spotify later comparing to others. The following cluster is implemented with DTW distance.

However, Hierarchical clustering which is on the Figure 3 shows stream groups for different countries from different angle. On a top left cluster, it can be seen how countries joined Spotify with a high increase in streams. The following cluster is implemented with Shape-base distance. This clustering groups similar times series into subgroups.

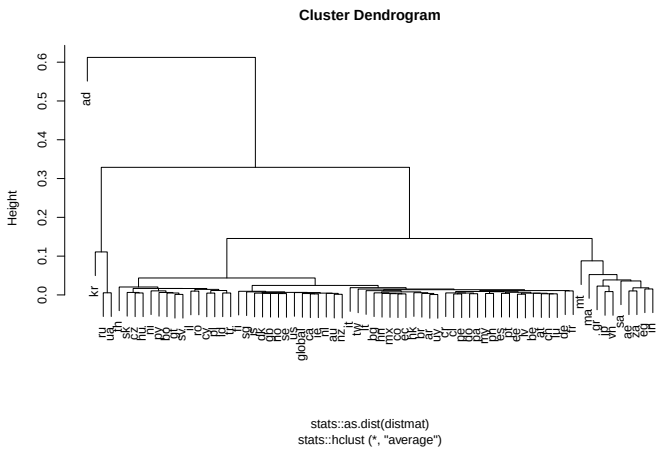

Figure 2: Partitional clustering for overal streams of grouped countures.

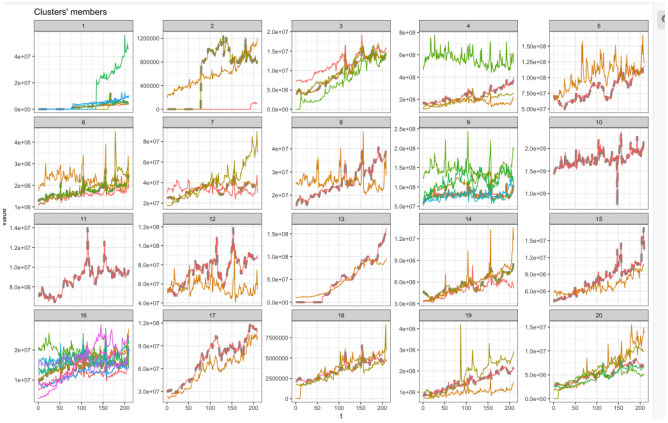

Figure 3: Hierarchical clustering for overal streams of grouped countures.

4 Time Series Outliers

Specifically for time series, outliers usually represent spikes and cavity of a times series figure. For example here was considered streams from Japan Spotify listeners. It can be clearly seen that they have started becoming active listeners after 2020. There are also spikes detected by the algorithm. Figure 4. Also, for detection outliers in times series, spikes and cavities are not always the best metrics. It's important to be able to detect a certain patterns with using other techniques like correlation. This specific example with listeners is focused on a growth or a fall of listeners.

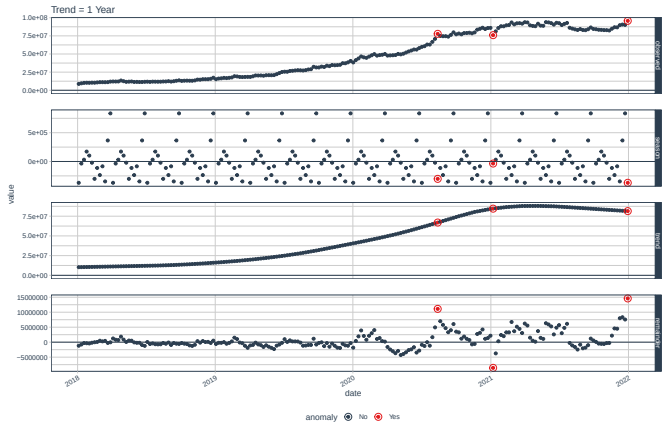

Figure 4: Anomaly detection for Japan Spotify listeners from 2018 to 2022.

5 Time Series Forecasting

The following forecast is done with Arima model. Figure 5. On the Figure 4 we can see how Arima predicted active Japan Spotify listeners for the period between 2020–2021. To predict a growth of listeners a history of two years between 2018 and 2020 was used.

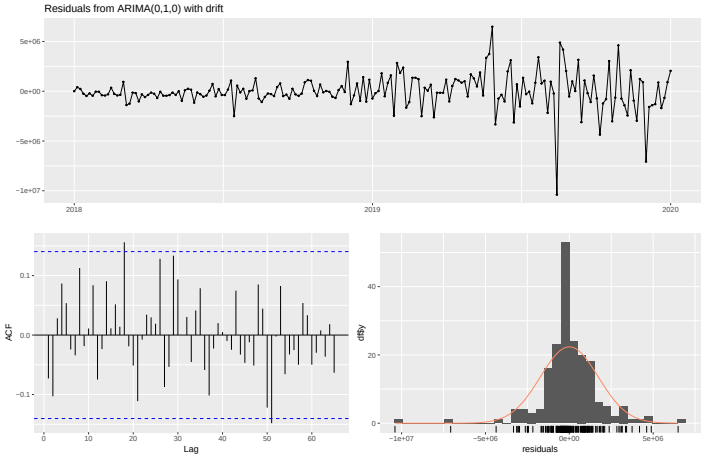

Figure 5: Forecast based on Arima model for 2022 for Japan Spotify listeners.

6 Summary

This project examines three techniques for time series data analysis: clustering, anomaly detection, and forecasting. The Spotify dataset is used to analyze how the amount of listeners change over time. The dataset has columns such as track id, name, country, date, positions, streams, artists, artist genres, duration, and explicit. The data preparation includes summing streams for specific countries and filling in missing data for countries that did not exist in certain periods. The time series clustering section uses two different techniques, DTW distance and Shape-base distance, to group similar time series into subgroups. The time series outliers section focuses on detecting spikes and cavities in the data, as well as using correlation to detect patterns. Finally, the forecast is done using the Arima model to predict the growth of Japan Spotify listeners for the period between 2020-2021.[?]

References

[1] *Spotify track chart dataset 2014-2022.* https://www.kaggle.com/datasets/jfreyberg/spotify-chart-data.