A Project Documentation (Report)

ON

Amazon Product Review Final Project

By

Group – 3

- Rashmi Rajesh Upadhyay
- Preeti Pal
- Adithia V
- Minnu Jacob K.

Towards the Fulfilment of the

Advanced PGP in Data Science and Machine Learning

PROBLEM STATEMENT:

1. Based on the ‘review text’ and ‘summary’ of the data, identify how a review can be classified into a particular category (for example, positive, negative and neutral).
   Analyse the text well to build the categories. In case of many negative sentiments associated with a product, try to find out the reasons.
2. Identify the names of few products by analysing your input text data.
3. Predict future data trends: How the sentiments of reviewers change with time. Suppose input data is provided for the period 1996-2014. Task is to predict the trends after 2014.
4. Can you find any relations/buying trends of customers/any other interesting analysis between given pair of categories (Example, Beauty and Clothing Categories).

Project Objective:

Analysis of the data and make use of best approaches for forecasting.

INTRODUCTION:

Machine Learning:

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

How machine learning works?

- A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabelled, algorithm will produce an estimate about a pattern in the data.

- An Error Function: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

- A Model Optimization Process: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

Machine learning methods:

Supervised machine learning:

Supervised learning, also known as supervised machine learning, is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids over fitting or under fitting. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more.

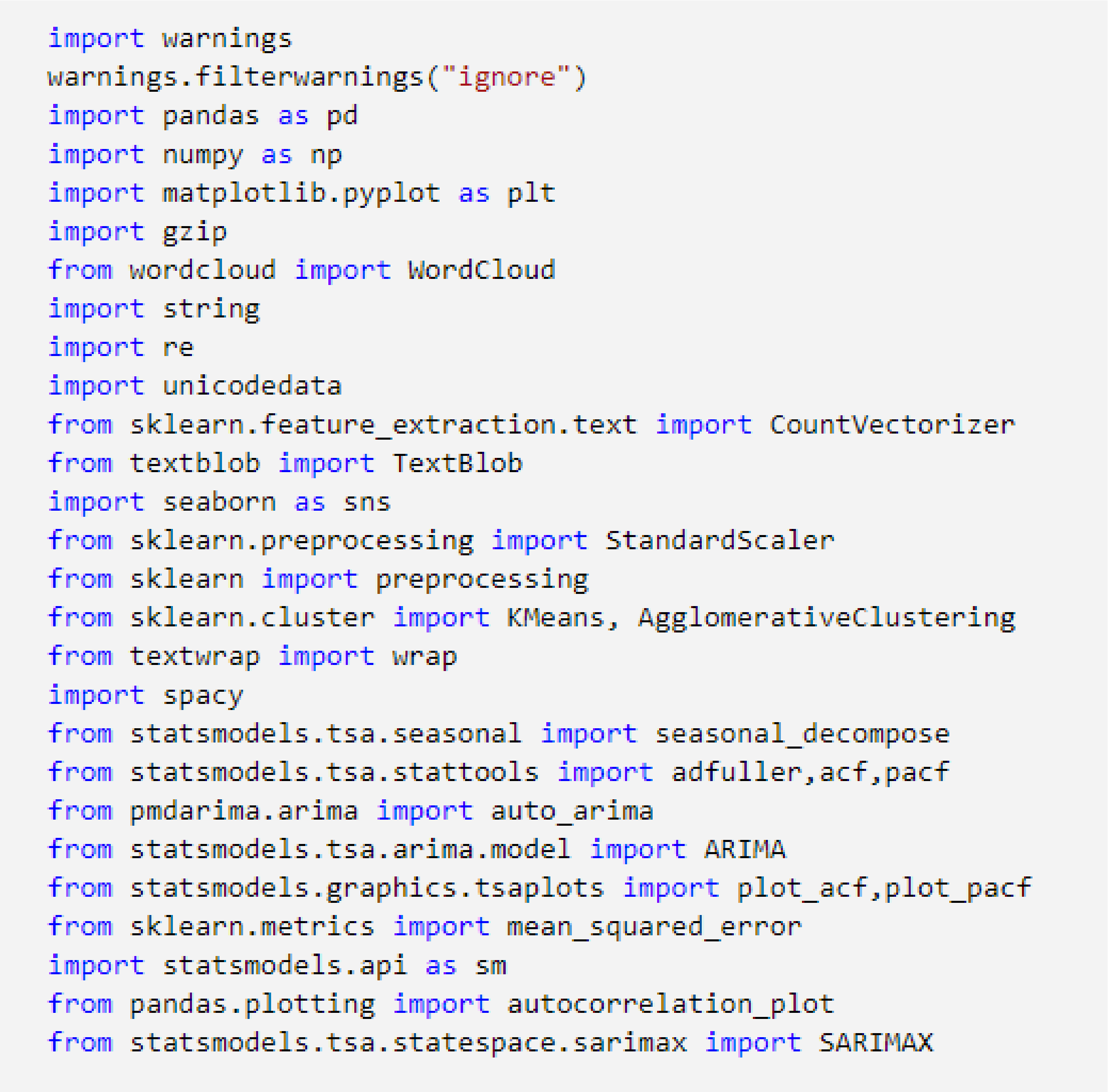
Unsupervised machine learning:

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, image and pattern recognition. It’s also used to reduce the number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more.

Python Libraries:

Python's standard library is very extensive, offering a wide range of facilities. The library contains built-in modules (written in C) that provide access to system functionality such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming. Some of these modules are explicitly designed to encourage and enhance the portability of Python programs by abstracting away platform-specifics into platform-neutral APIs.

Here are some of these that we have used:

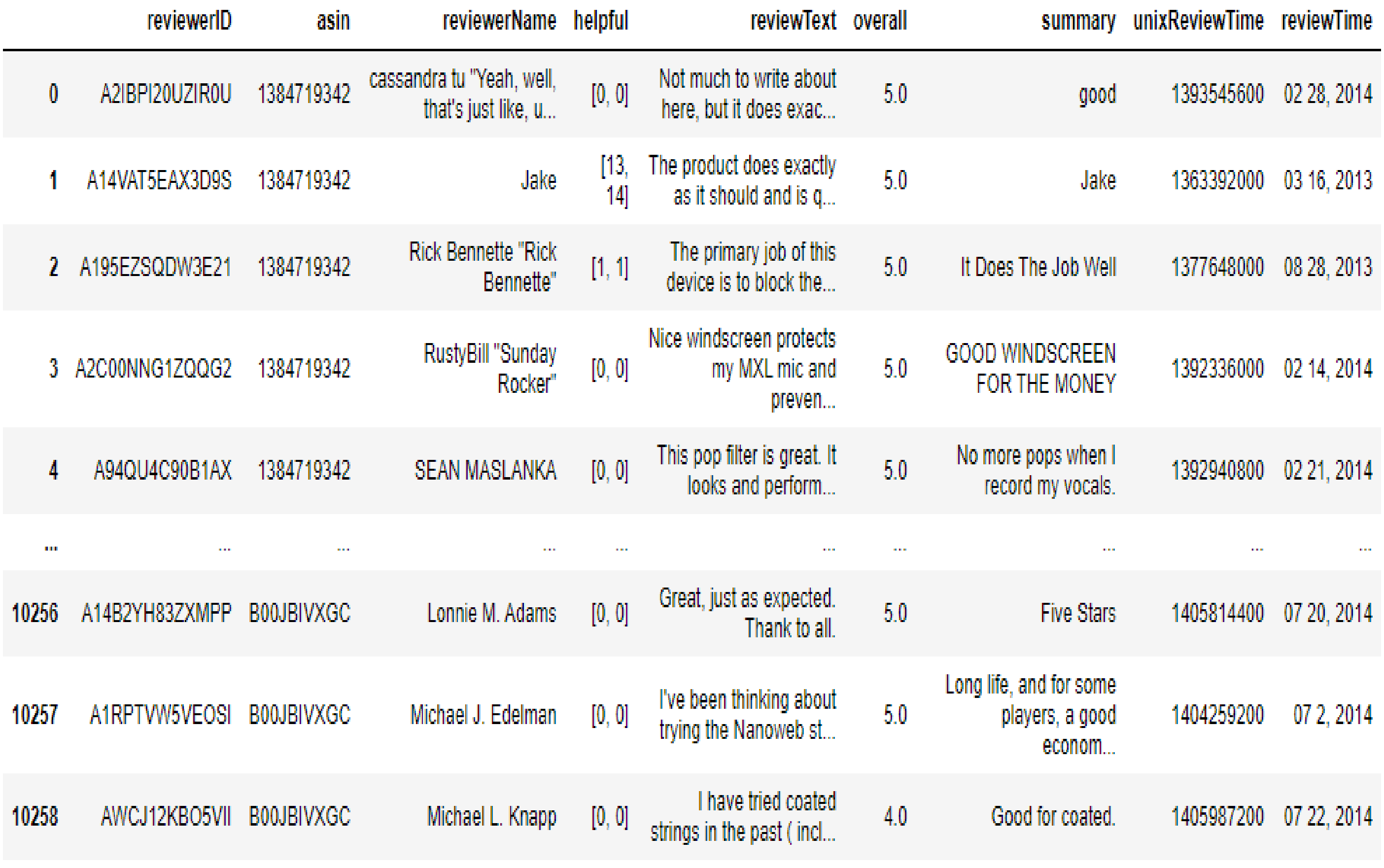


```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import gzip
from wordcloud import WordCloud
import string
import re
import unicodedata
from sklearn.feature_extraction.text import CountVectorizer
from textblob import TextBlob
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
from sklearn.cluster import KMeans, AgglomerativeClustering
from textwrap import wrap
import spacy
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller,acf,pacf
from pmdarima.arima import auto_arima
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm
from pandas.plotting import autocorrelation_plot
from statsmodels.tsa.statespace.sarimax import SARIMAX
```

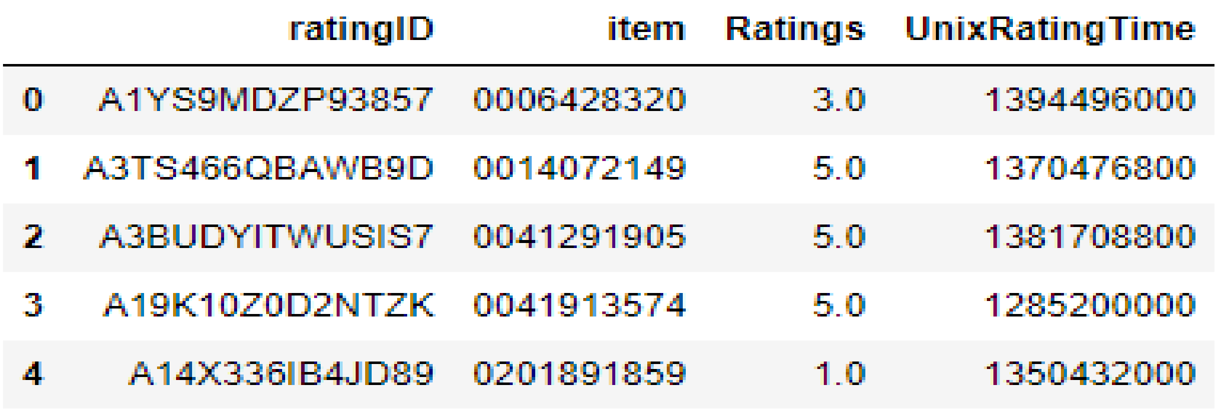
DESCRIPTION OF DATASETS:

We have two amazon data sets: Musical Instruments and Digital Music.

- There are four initial dataset, namely: reviews_Digital_Music, reviews_Musical_Instruments, ratings_Digital_Music, ratings_Musical_Instruments, two product review file, and two product rating file.
- Reviews_Digital_Music and reviews_Musical_Instruments dataset contains reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime.
- Where ratings_Digital_Music and ratings_Musical_Instruments contains ratingID, item, Ratings and UnixRatingTime.
- Here is some glimpse of imported data.



	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A2IBPI20UZIR0U	1384719342	cassandra tu "Yeah, well, that's just like, u...	[0, 0]	Not much to write about here, but it does exac...	5.0	good	1393545600	02 28, 2014
1	A14VAT5EAX3D9S	1384719342	Jake	[13, 14]	The product does exactly as it should and is q...	5.0	Jake	1363392000	03 16, 2013
2	A195EZSQDW3E21	1384719342	Rick Bennette "Rick Bennette"	[1, 1]	The primary job of this device is to block the...	5.0	It Does The Job Well	1377648000	08 28, 2013
3	A2C00NNG1ZQQG2	1384719342	RustyBill "Sunday Rocker"	[0, 0]	Nice windscreen protects my MXL mic and preven...	5.0	GOOD WINDSCREEN FOR THE MONEY	1392336000	02 14, 2014
4	A94QU4C90B1AX	1384719342	SEAN MASLANKA	[0, 0]	This pop filter is great. It looks and perform...	5.0	No more pops when I record my vocals.	1392940800	02 21, 2014
...	...	...	...	...	...	...	...	...	...
10256	A14B2YH83ZXMPP	B00JBIVXGC	Lonnie M. Adams	[0, 0]	Great, just as expected. Thank to all.	5.0	Five Stars	1405814400	07 20, 2014
10257	A1RPTVW5VEOSI	B00JBIVXGC	Michael J. Edelman	[0, 0]	I've been thinking about trying the Nanoweb st...	5.0	Long life, and for some players, a good econom...	1404259200	07 2, 2014
10258	AWCJ12KBO5VII	B00JBIVXGC	Michael L. Knapp	[0, 0]	I have tried coated strings in the past (incl...	4.0	Good for coated.	1405987200	07 22, 2014



	ratingID	item	Ratings	UnixRatingTime
0	A1YS9MDZP93857	0006428320	3.0	1394496000
1	A3TS466QBAWB9D	0014072149	5.0	1370476800
2	A3BUDYITWUSIS7	0041291905	5.0	1381708800
3	A19K10Z0D2NTZK	0041913574	5.0	1285200000
4	A14X336IB4JD89	0201891859	1.0	1350432000

Why product reviews are important in ecommerce?

Review analysis is important because it helps you understand what customers think of your product, app, or service experience. At its heart reviews/ratings is a customer centric activity. It enables you to make decisions on what to improve based on what your customers value.

Data Understanding:

- Collection of relevant data
- Describe the data
- Exploring the data using plots
- Validating the data quality
- After importing all the data, both the review files and ratings file have been concatenated respectively.
- The shape of both concatenated files have been displayed.
- Null value checked and have been treated.
- Data Frames info has been derived.
- Central tendency measures have been calculated.
- Uniqueness of reviewer/rating ID has been checked.
- Distribution of plots on different attributes has been performed to visualize and better understand the data.

Data Analyzing (Methods):

Descriptive Analysis

- Summarizes the data
- Describe the past trends
- Formulating Central Tendency

Exploratory Analysis

- Exploring and figuring patterns
- Discovering correlations

Predictive Analysis

- Predicting future by understanding past patterns

Prescriptive Analysis

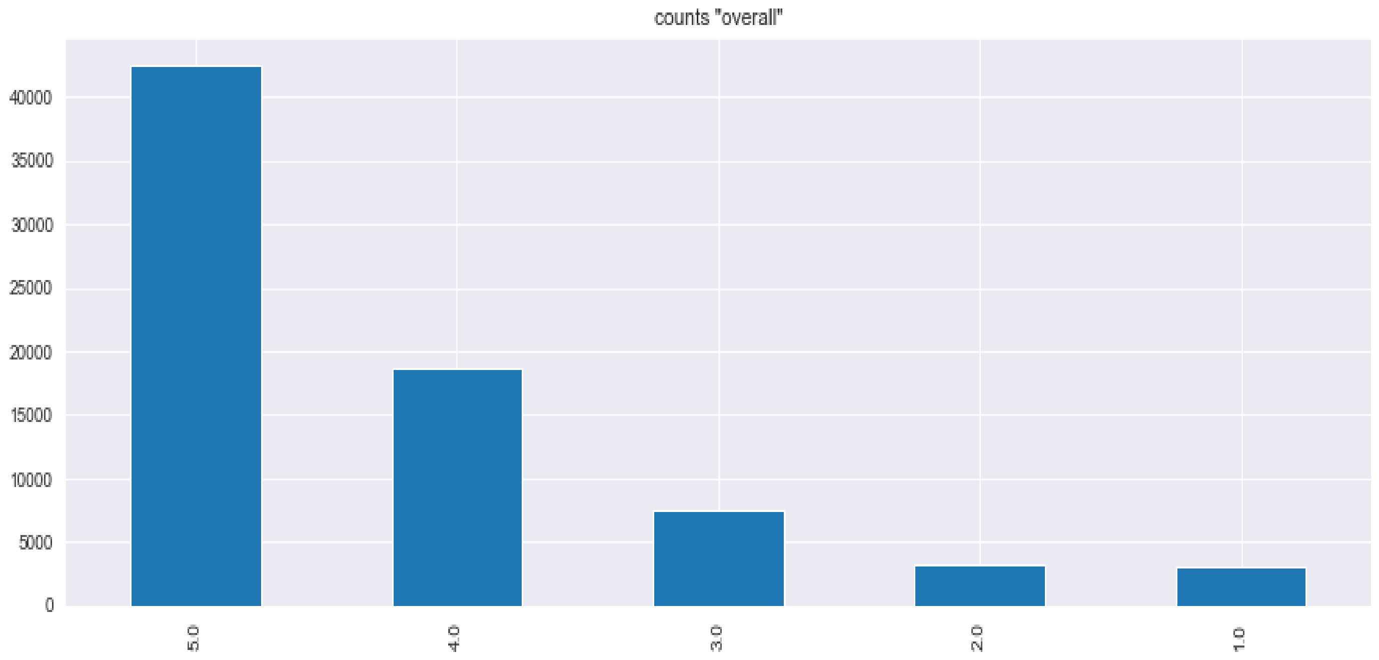
- Decision support
- Recommends action based on forecast

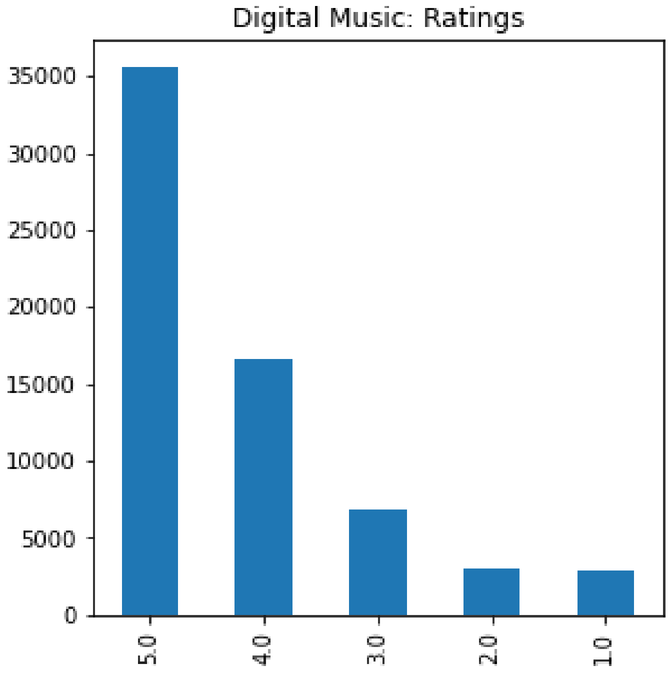
Exploratory Data Analysis (EDA):

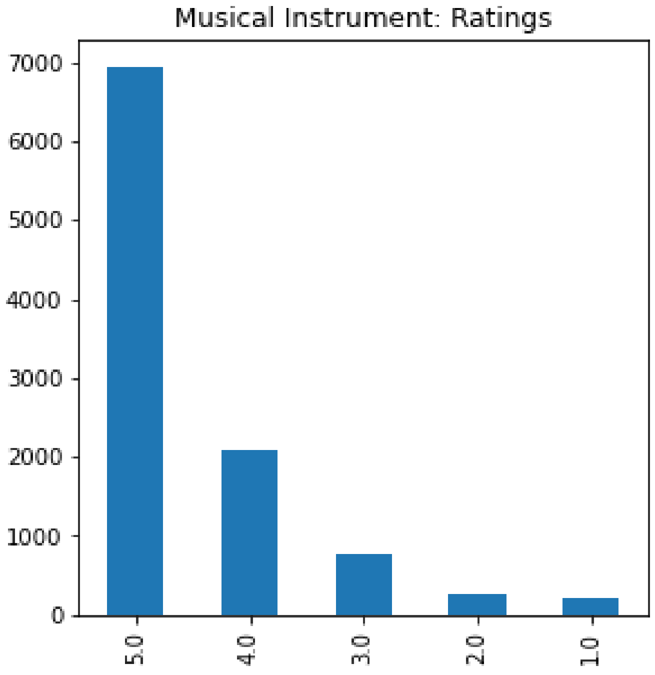
EDA is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

Univariate Analysis:

Univariate analysis is the simplest form of data analysis, where the data being analyzed consists of only one variable. Since it’s a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it. Let us look at a few visualizations used for performing univariate analysis.

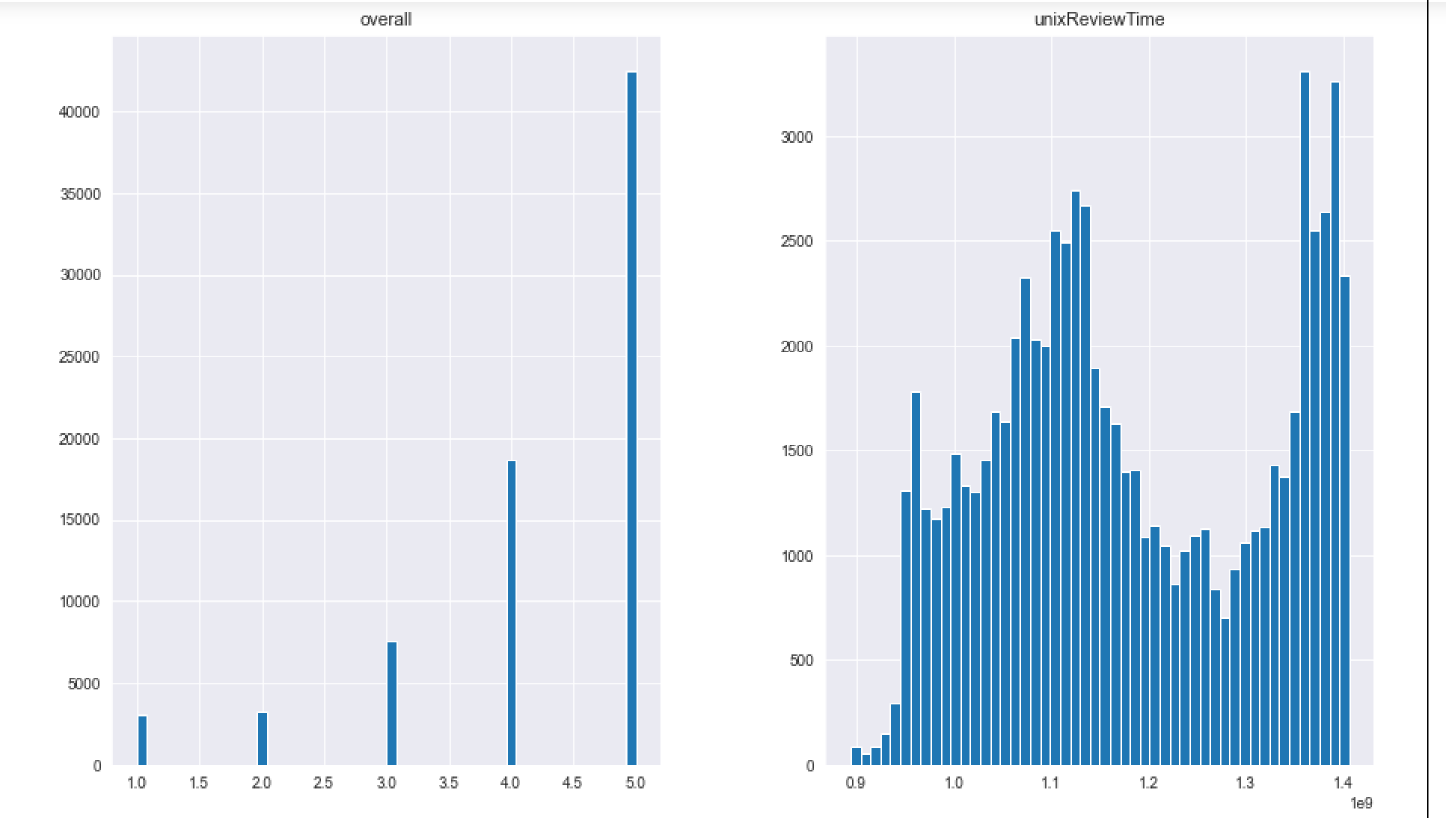






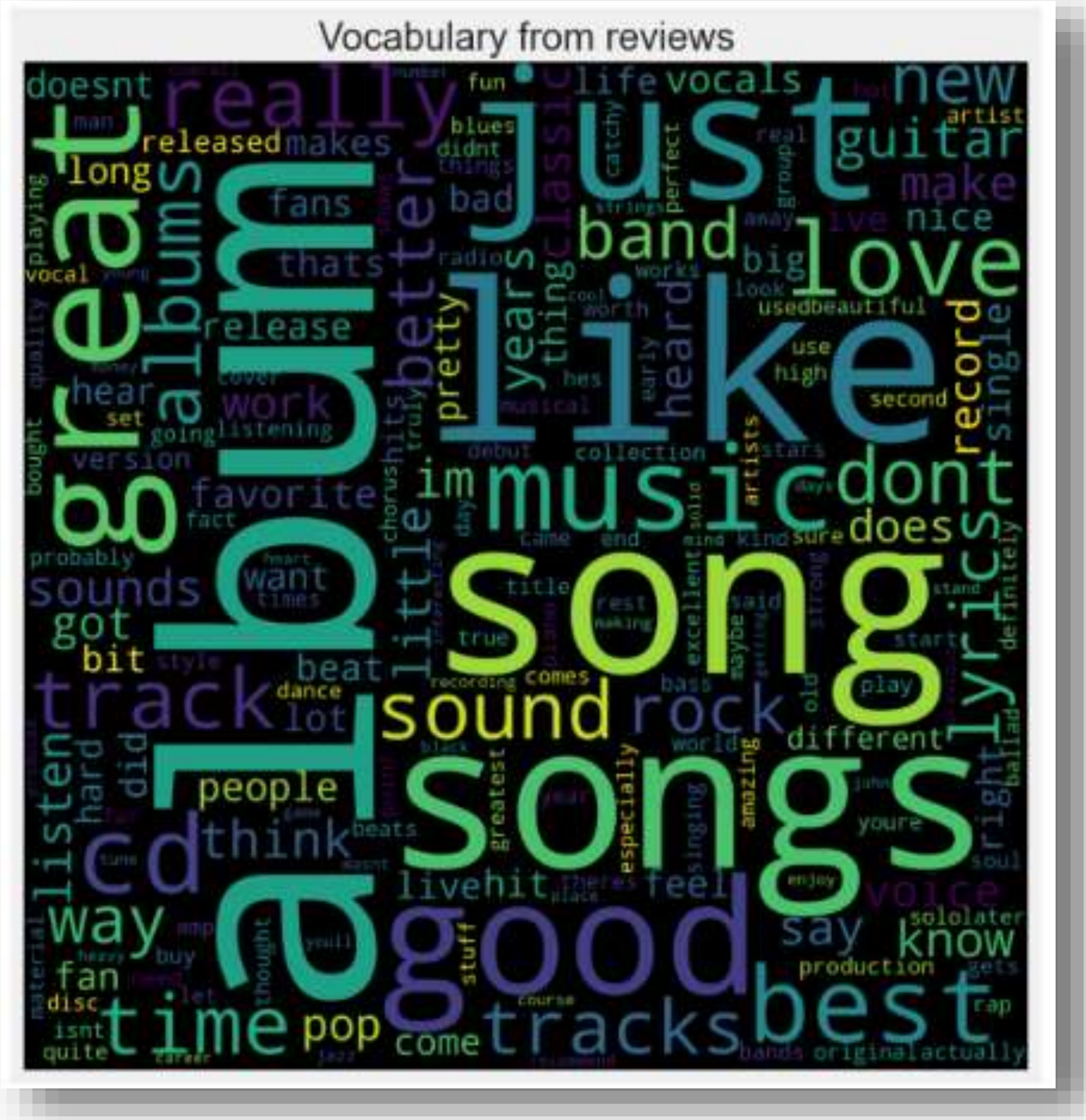
Multivariate Analysis:

Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. This models more realistic applications, where each situation, product, or decision involves more than a single variable. Let us look at a few visualizations used for performing multivariate analysis.



Word Cloud:

Word cloud is a technique for visualising frequent words in a text where the size of the words represents their frequency. Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.


Vocabulary from reviews

Natural Language Processing:

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment.

Sentiment Analysis:

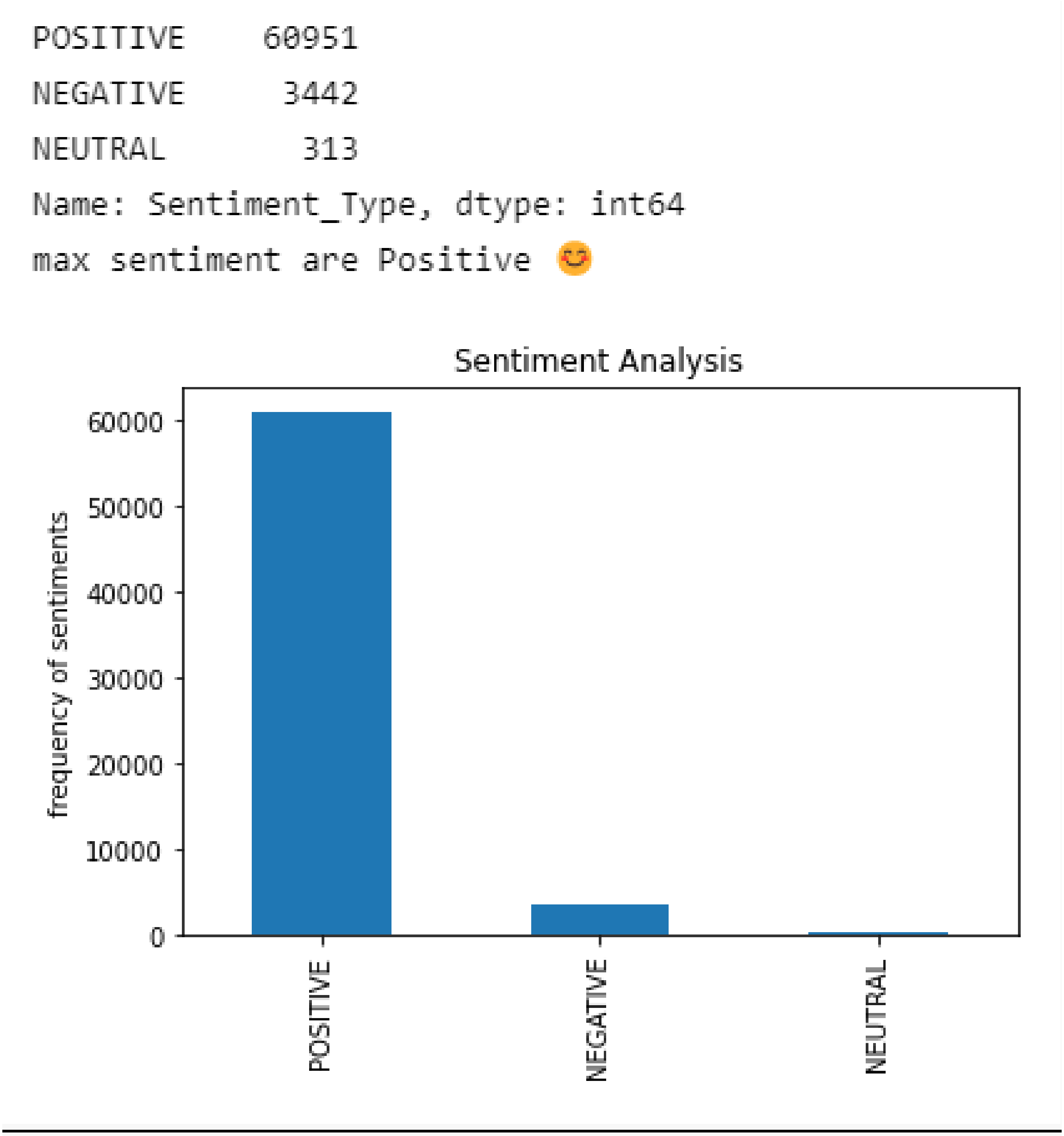
Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

Why Is Sentiment Analysis Important?

Since humans express their thoughts and feelings more openly than ever before, sentiment analysis is fast becoming an essential tool to monitor and understand sentiment in all types of data.

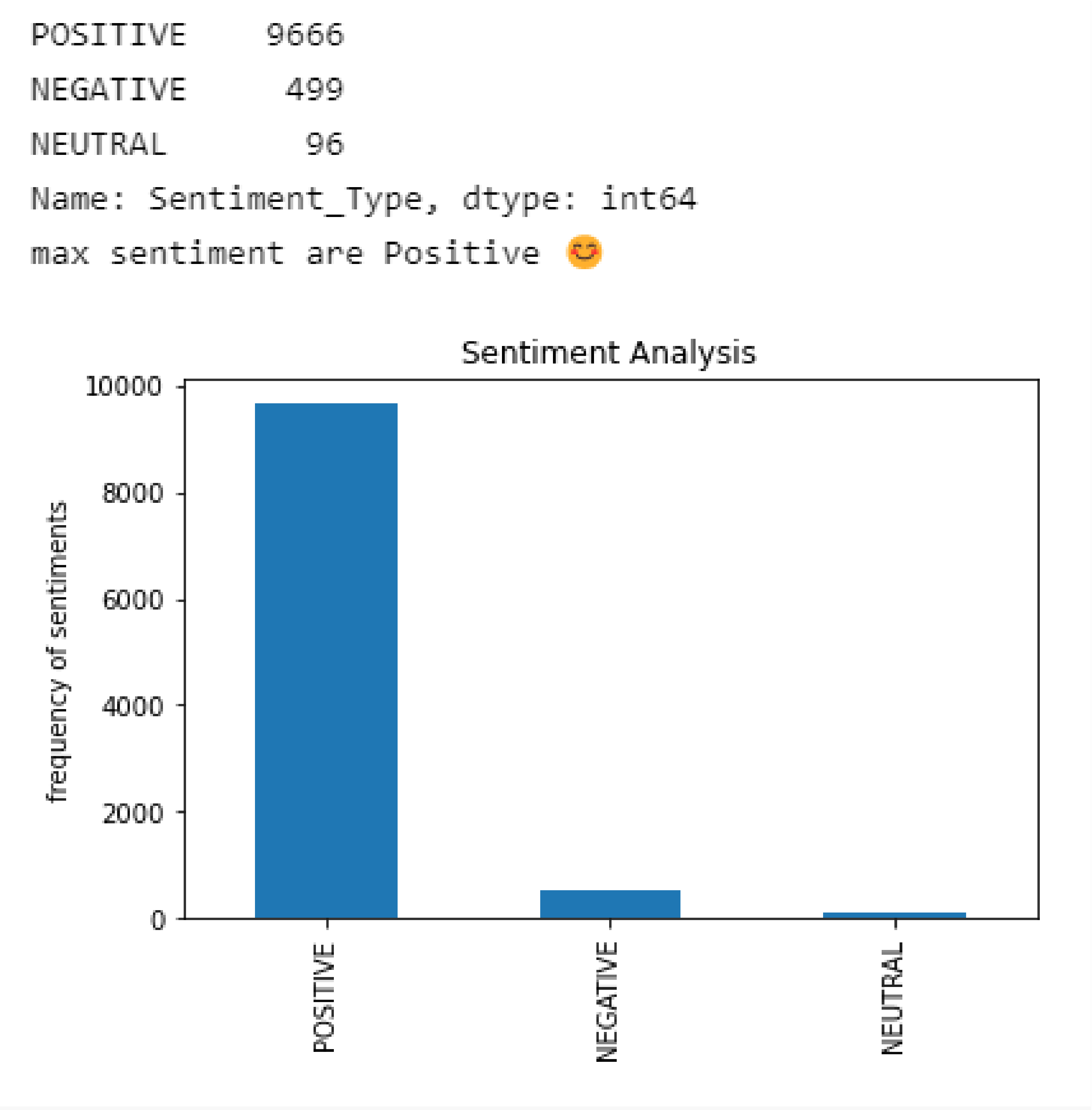
Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers' needs.

Sentiment Analysis for Digital Music Data:



```
POSITIVE    60951
NEGATIVE     3442
NEUTRAL       313
Name: Sentiment_Type, dtype: int64
max sentiment are Positive 😊
```

Sentiment Analysis for Musical Instruments Data:



```
POSITIVE    9666
NEGATIVE     499
NEUTRAL       96
Name: Sentiment_Type, dtype: int64
max sentiment are Positive 😊
```

CLASSIFICATION:

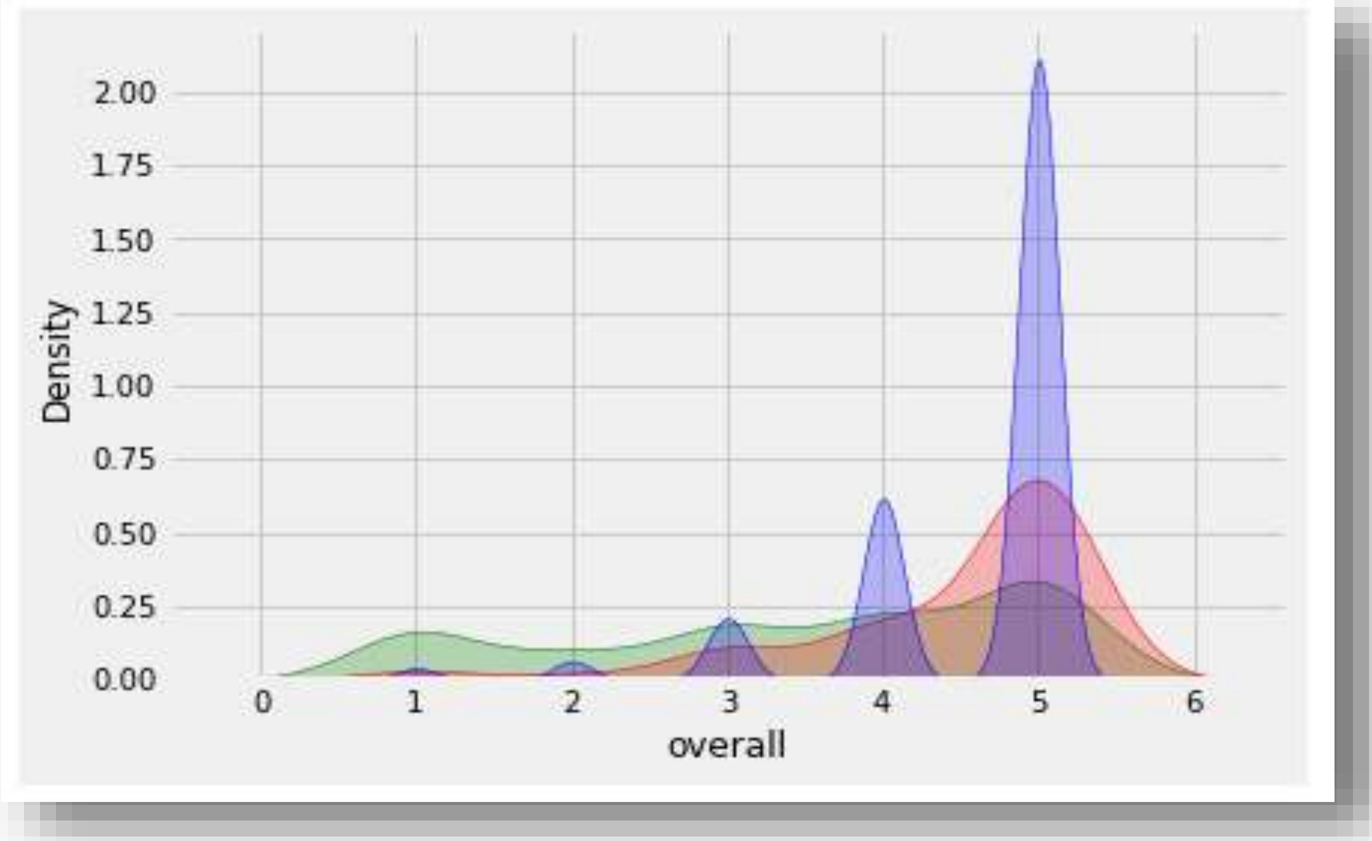
Classification is a task of Machine Learning which assigns a label value to a specific class and then can identify a particular type to be of one kind or another. The most basic example can be of the mail spam filtration system where one can classify a mail as either "spam" or "not spam.

Classification usually refers to any kind of problem where a specific type of class label is the result to be predicted from the given input field of data.

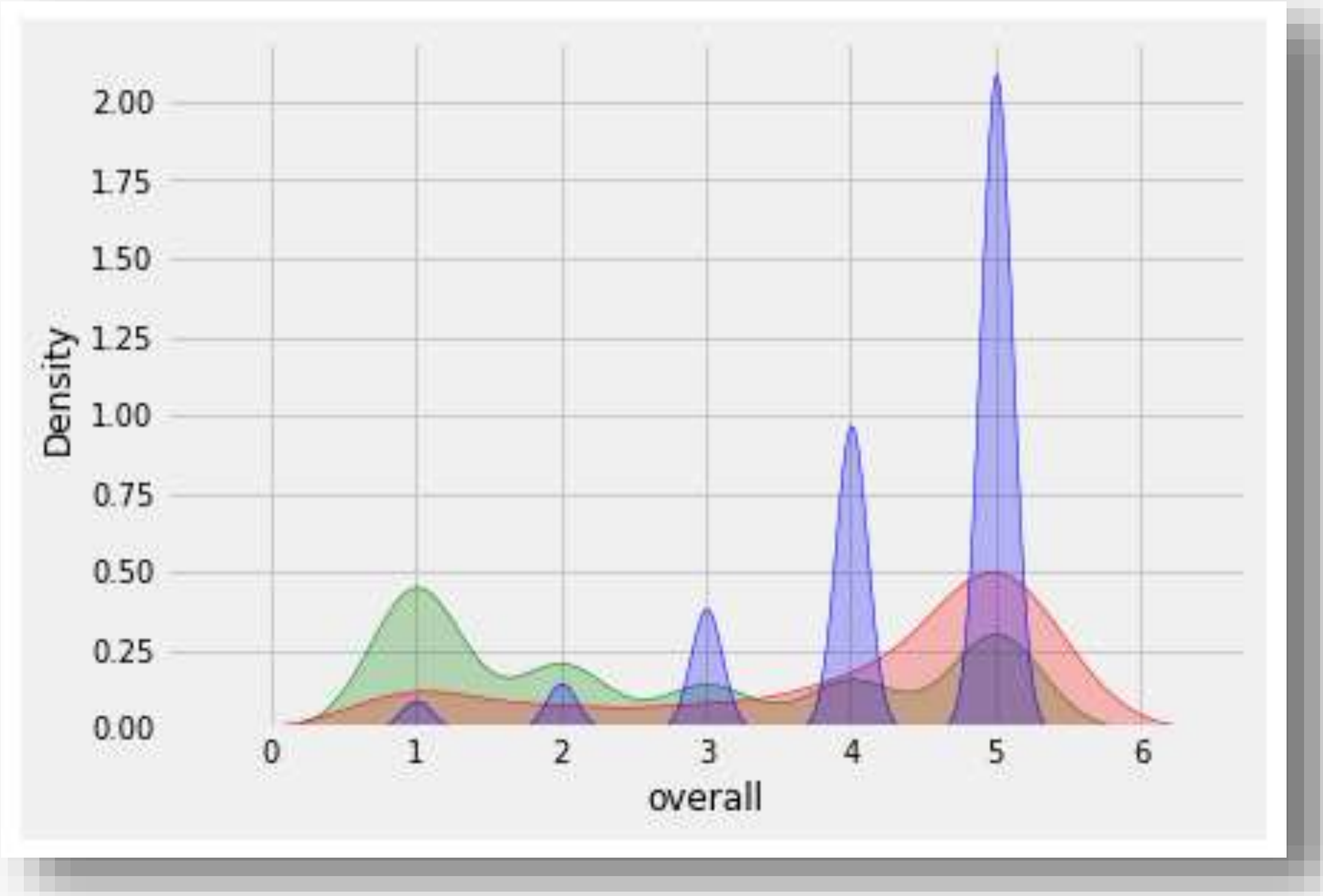
Some types of Classification challenges are:

- Classifying emails as spam or not.
- Classify a given handwritten character to be either a known character or not.
- Classify recent user behaviour as churn or not.

Classification for Digital Music Data:



Classification for Musical Instruments Data:

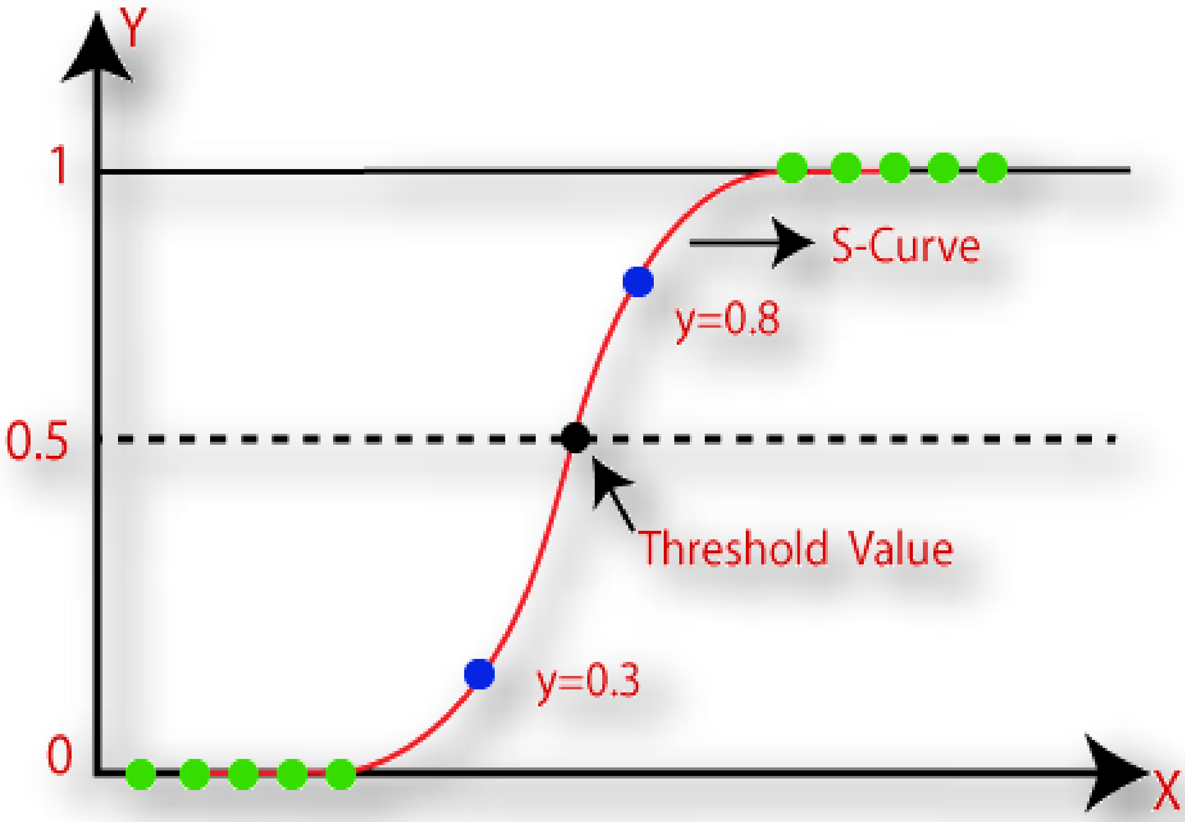


Logistic Regression:

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving

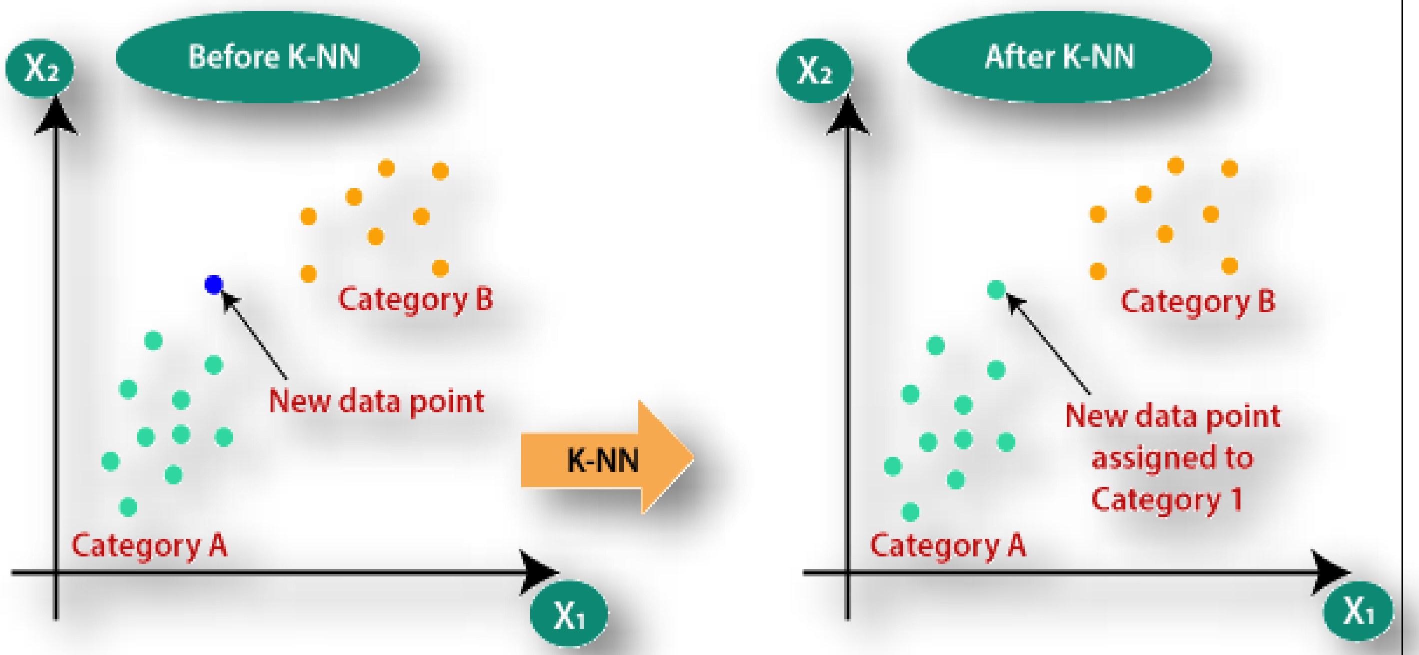
Regression problems, whereas **Logistic regression is used for solving the classification problems**.

- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



K-Nearest Neighbor (KNN):

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.


Before K-NN
X₂
Category B
New data point
Category A
X₁
K-NN
After K-NN
X₂
Category B
New data point assigned to Category 1
Category A
X₁

KNN Classifier
Input value
Predicted Output

Cross-Validation:

Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. We can also say that it is a technique to check how a statistical model generalizes to an independent dataset.

K-Fold Cross-Validation

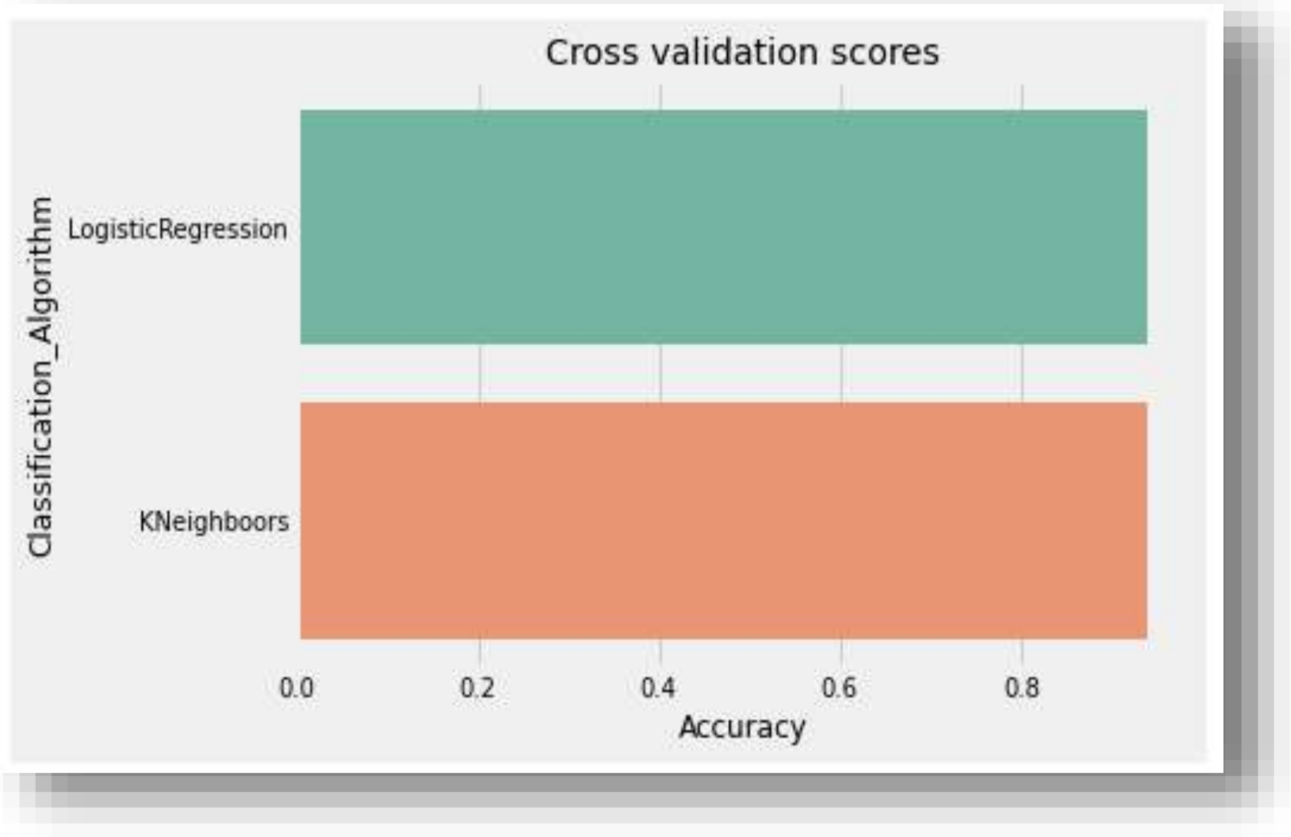
K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called folds. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set. This approach is a very popular CV approach because it is easy to understand, and the output is less biased than other methods.

Stratified k-fold cross-validation

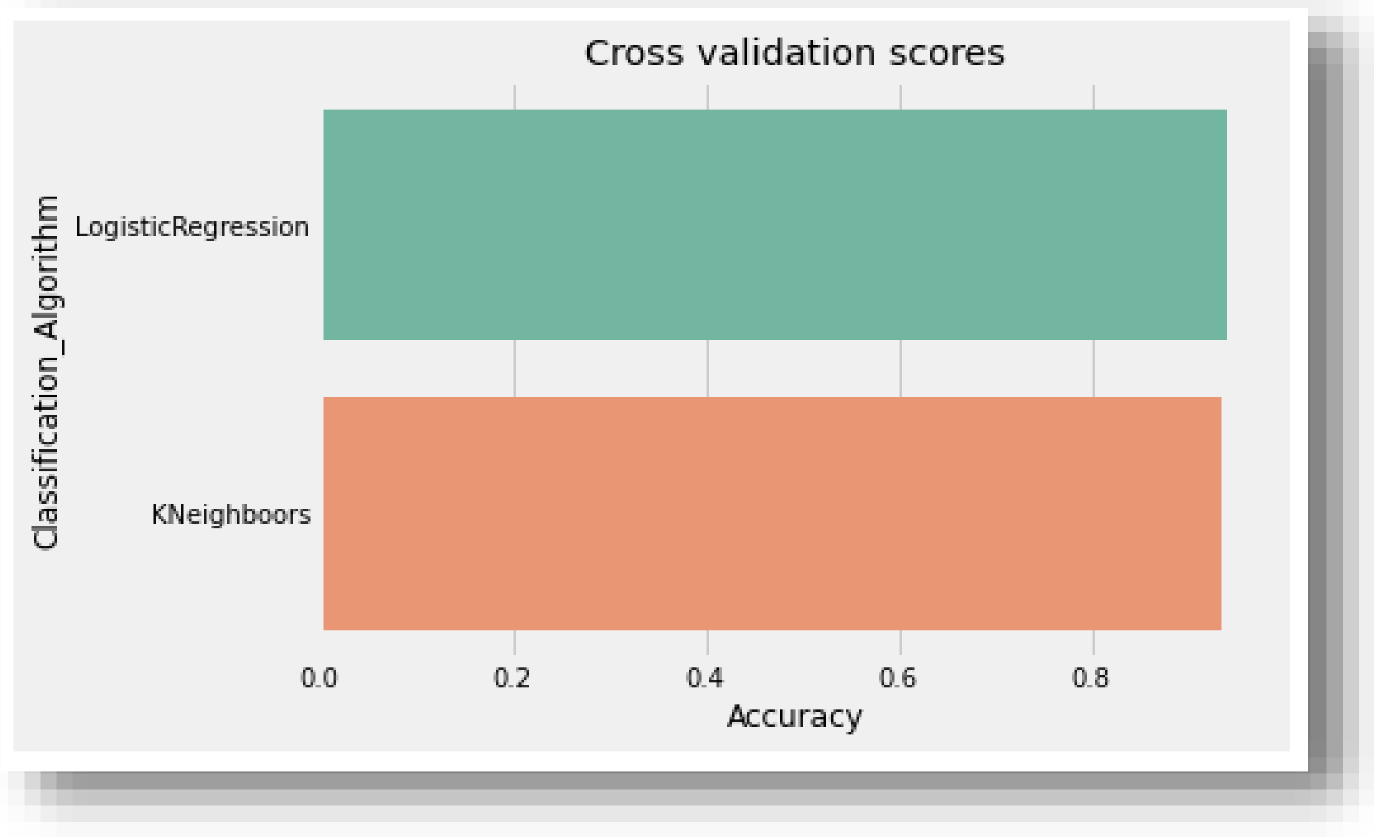
This technique is similar to k-fold cross-validation with some little changes. This approach works on stratification concept, it is a process of rearranging the data to ensure that each fold or group is a good representative of the complete dataset. To deal with the bias and variance, it is one of the best approaches.

It can be understood with an example of housing prices, such that the price of some houses can be much high than other houses. To tackle such situations, a stratified k-fold cross-validation technique is useful.

Cross- Validation for Digital Music Data:



Cross- Validation for Musical Instruments Data:



Applications of Cross-Validation:

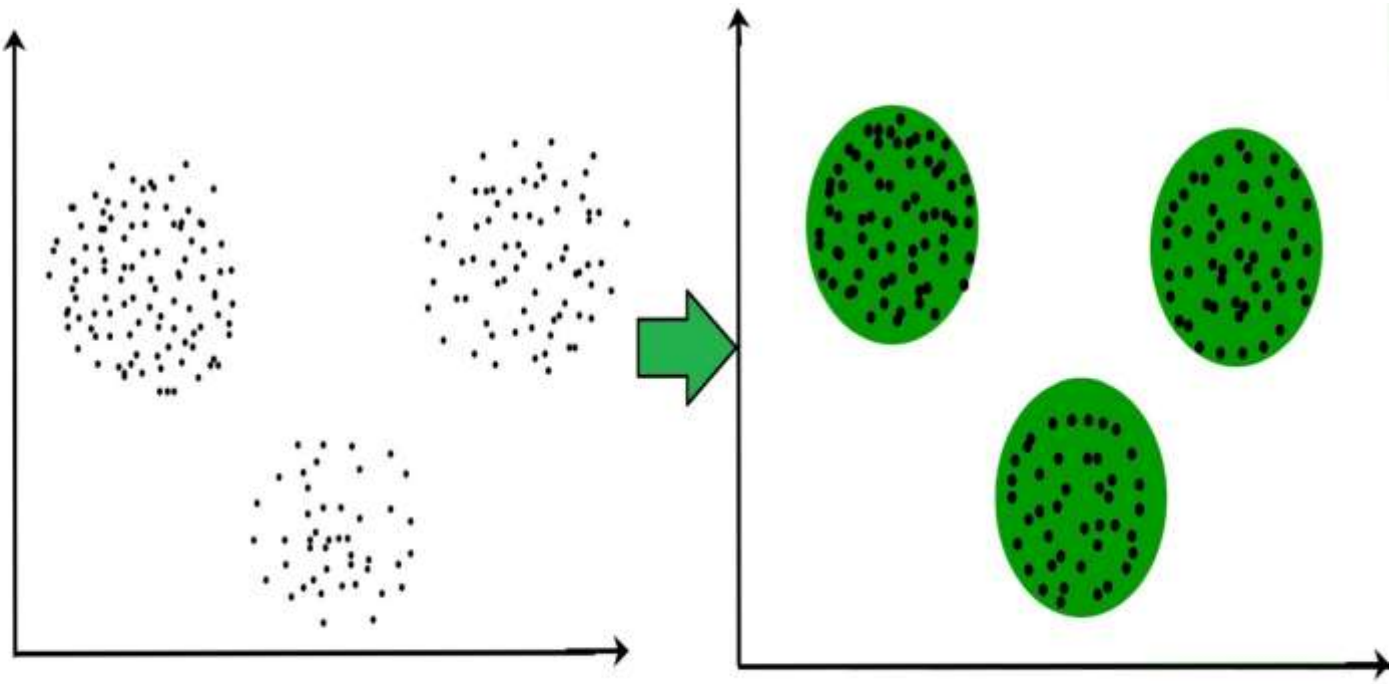
- This technique can be used to compare the performance of different predictive modelling methods.
- It has great scope in the medical research field.
- It can also be used for the meta-analysis, as it is already being used by the data scientists in the field of medical statistics.

CLUSTERING:

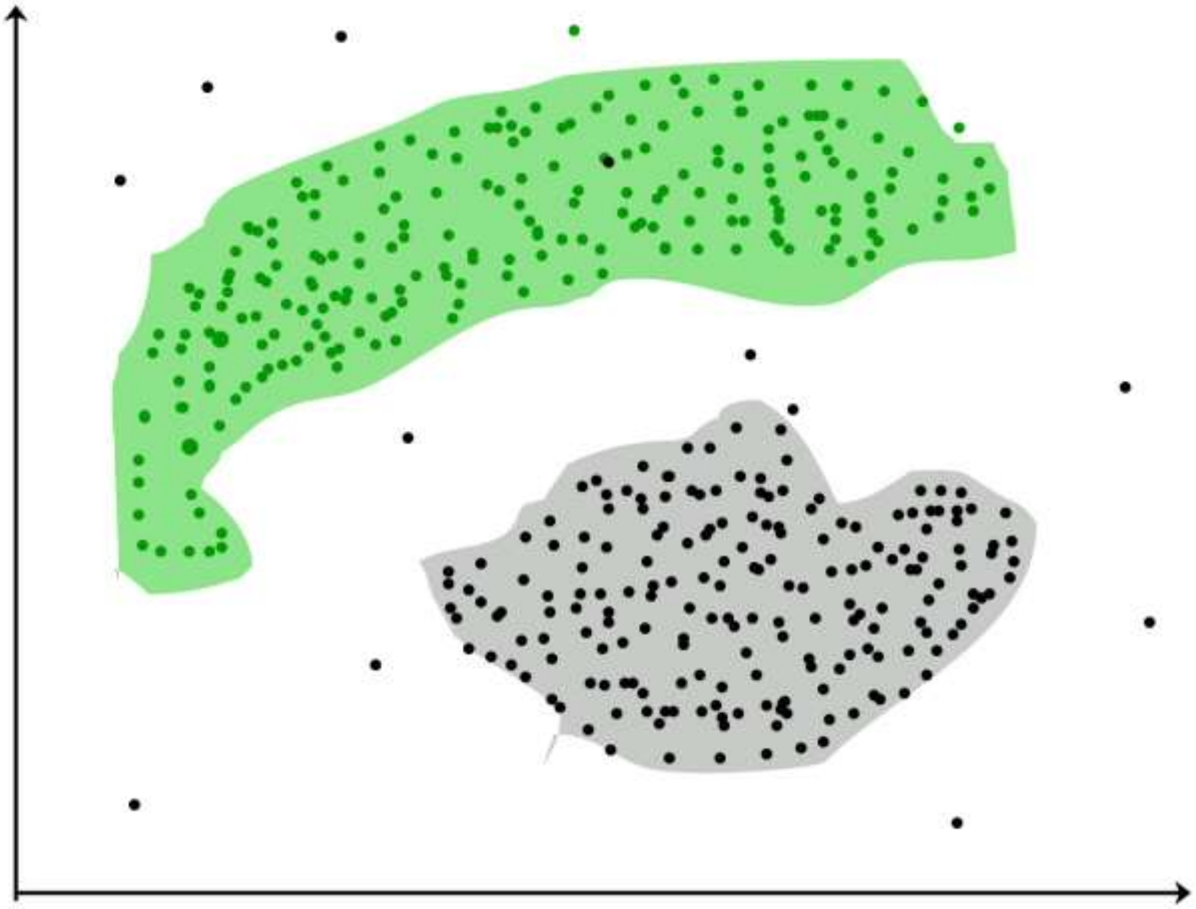
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



It is not necessary for clusters to be spherical. Such as:



Why clustering is important?

Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Clustering Algorithms:

K-means clustering algorithm:

It is the simplest unsupervised learning algorithm that solves clustering problem. K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

Applications of Clustering in different fields:

Marketing: It can be used to characterize & discover customer segments for marketing purposes.

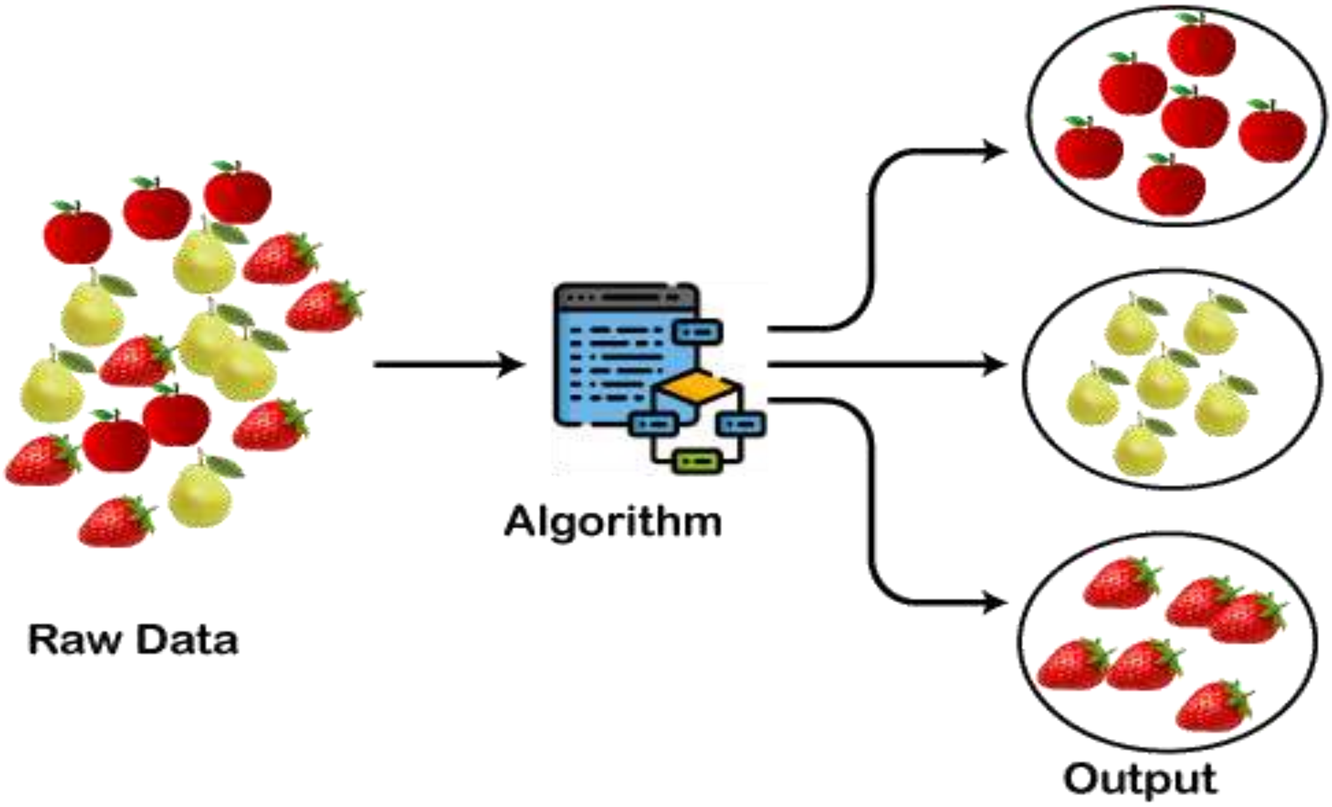
Biology: It can be used for classification among different species of plants and animals.

Libraries: It is used in clustering different books on the basis of topics and information.

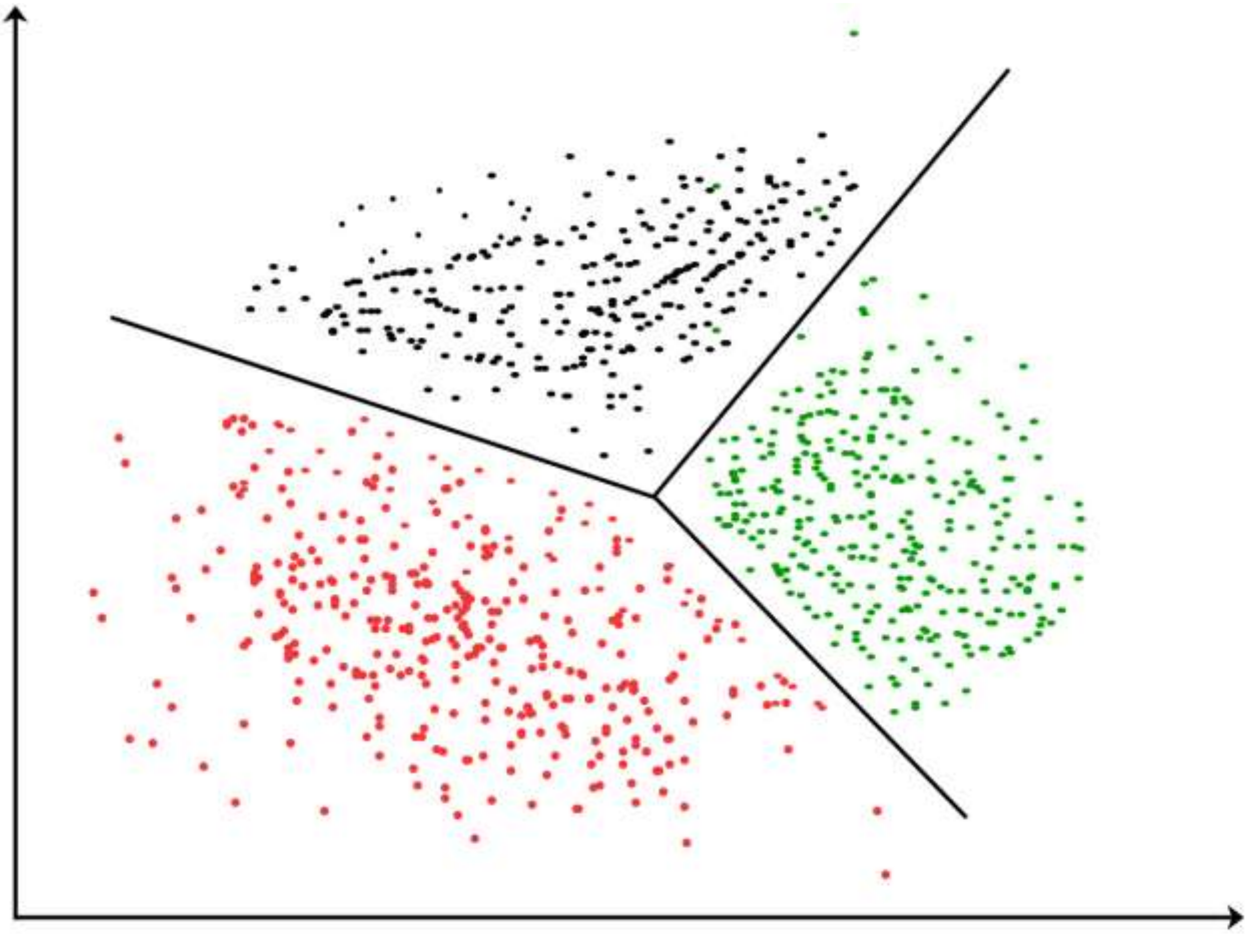
Insurance: It is used to acknowledge the customers, their policies and identifying the frauds.

City Planning: It is used to make groups of houses and to study their values based on their geographical locations and other factors present.

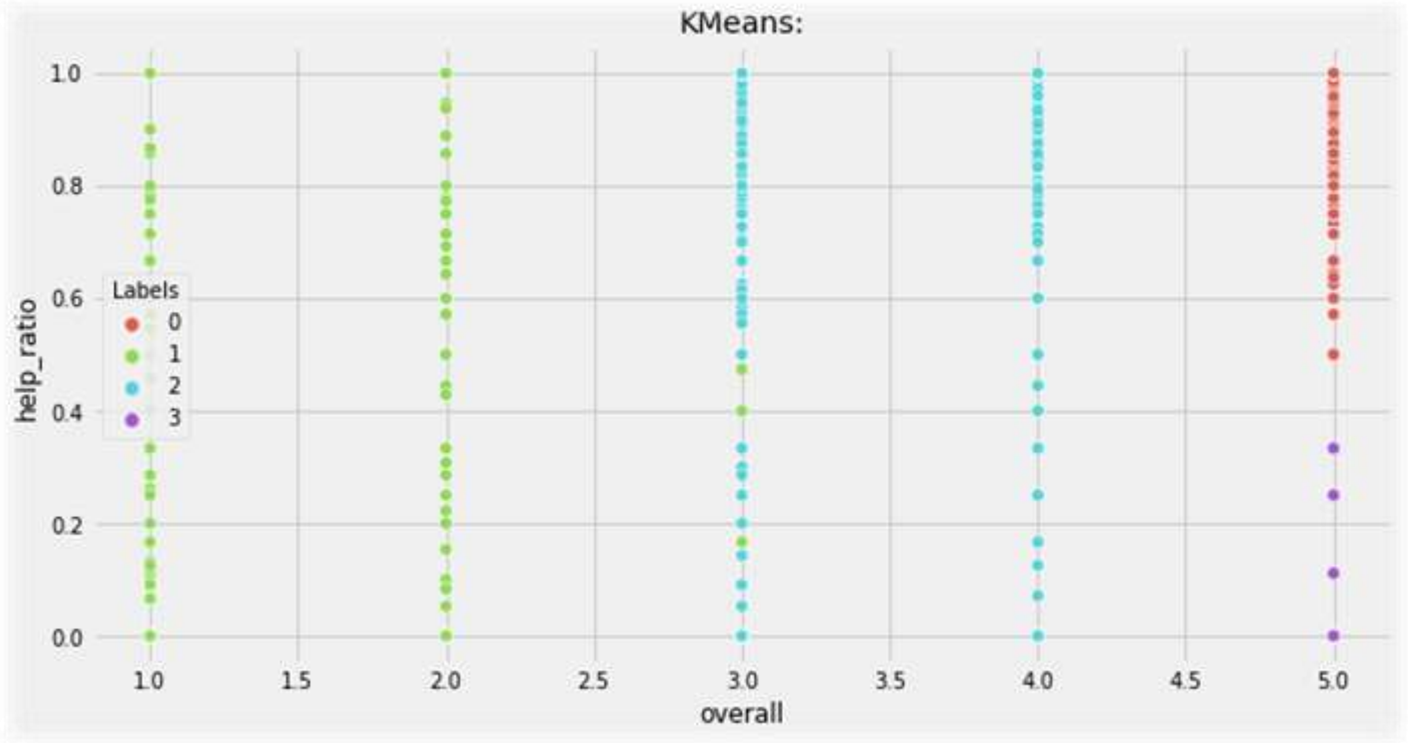
Earthquake studies: By learning the earthquake-affected areas we can determine the dangerous zones.



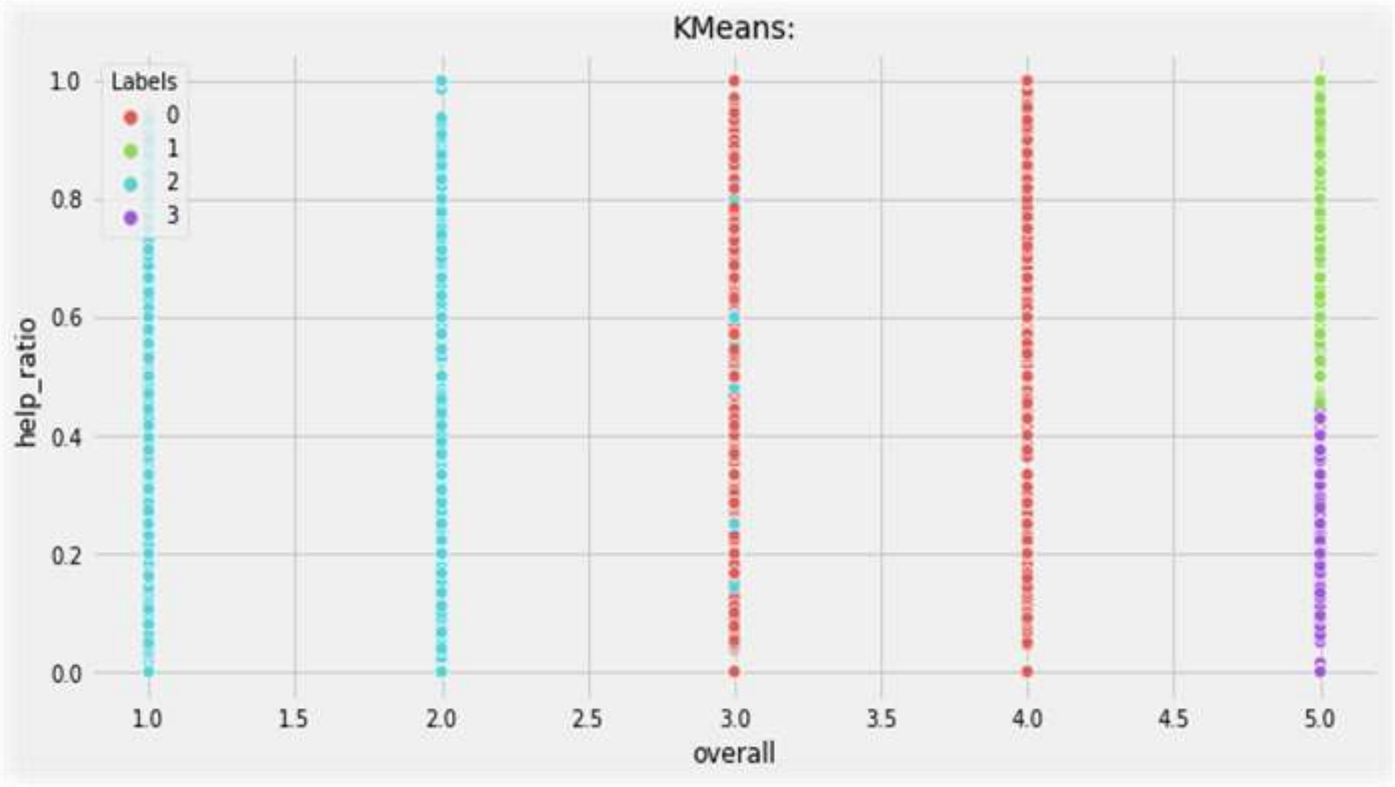
Before K-Means K-Means After K-Means Algorithm



K-means clustering on musical instruments data:



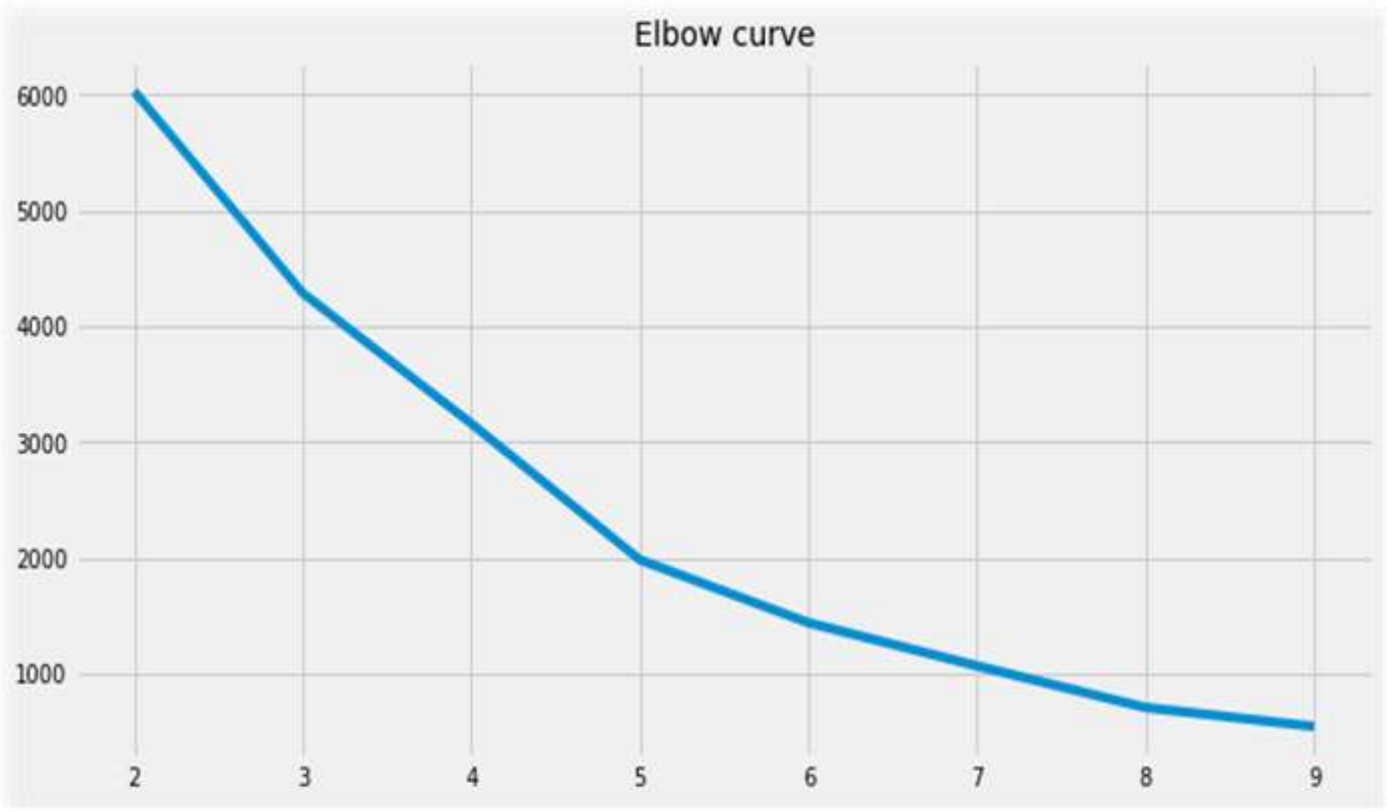
K-means clustering on digital music data:



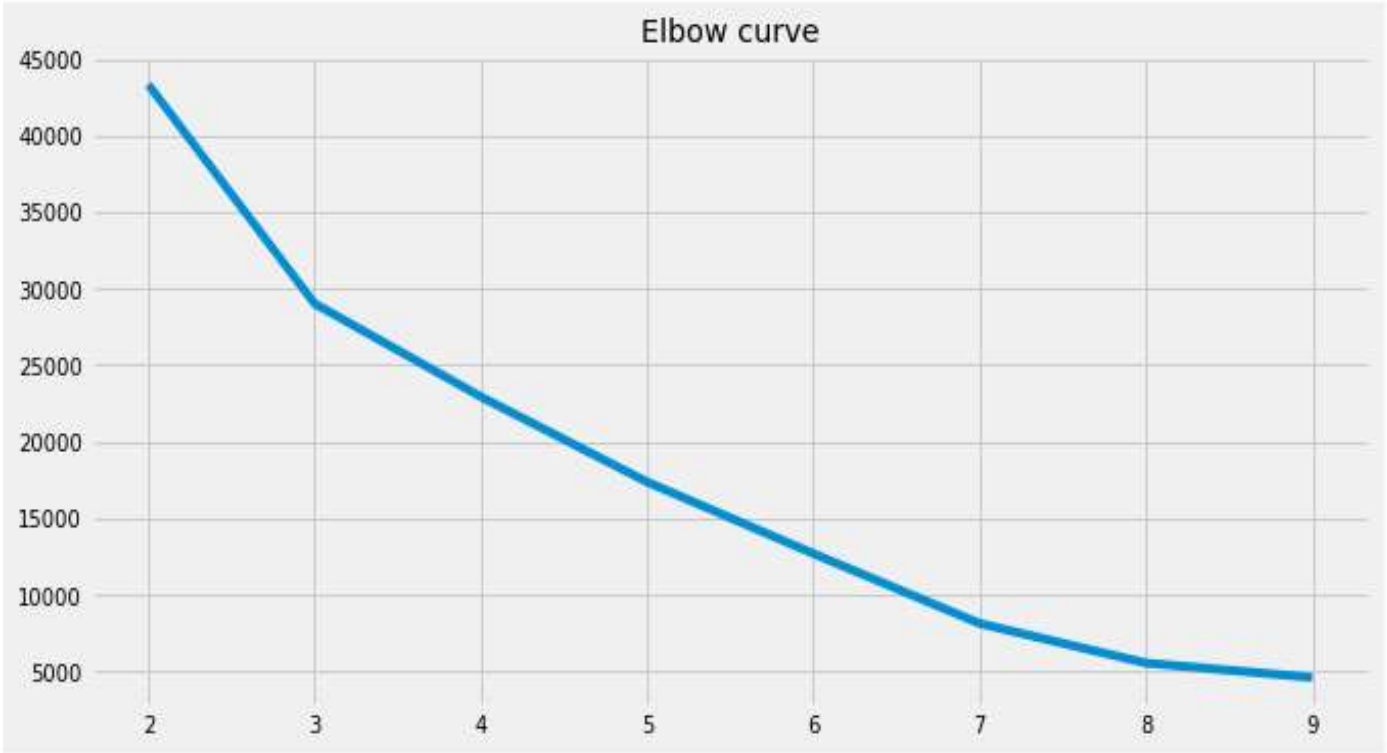
ELBOW METHOD OR ELBOW CURVE:

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of *k*. As you know, if *k* increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as *k* increases. The value of *k* at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

Elbow Curve for Musical Instrument Data:



Elbow Curve for Digital Music Data:



TIME SERIES ANALYSIS:

- A time series is nothing but a sequence of various data points that occurred in a successive order for a given period of time.
- A Time-Series represents a series of time-based orders. It could be Years, Months, Weeks, Days, Hours, Minutes, and Seconds.
- A time series is an observation from the sequence of discrete-time of successive intervals.
- A time series is a running chart.
- The time variable/feature is the independent variable and supports the target variable to predict the results.
- Time Series Analysis (TSA) is used in different fields for time-based predictions – like Weather Forecasting, Financial, Signal processing, Engineering domain – Control Systems, Communications Systems etc.
- Using ARIMA, SARIMA models, we could predict the future (forecasting).

Components of Time Series Analysis:

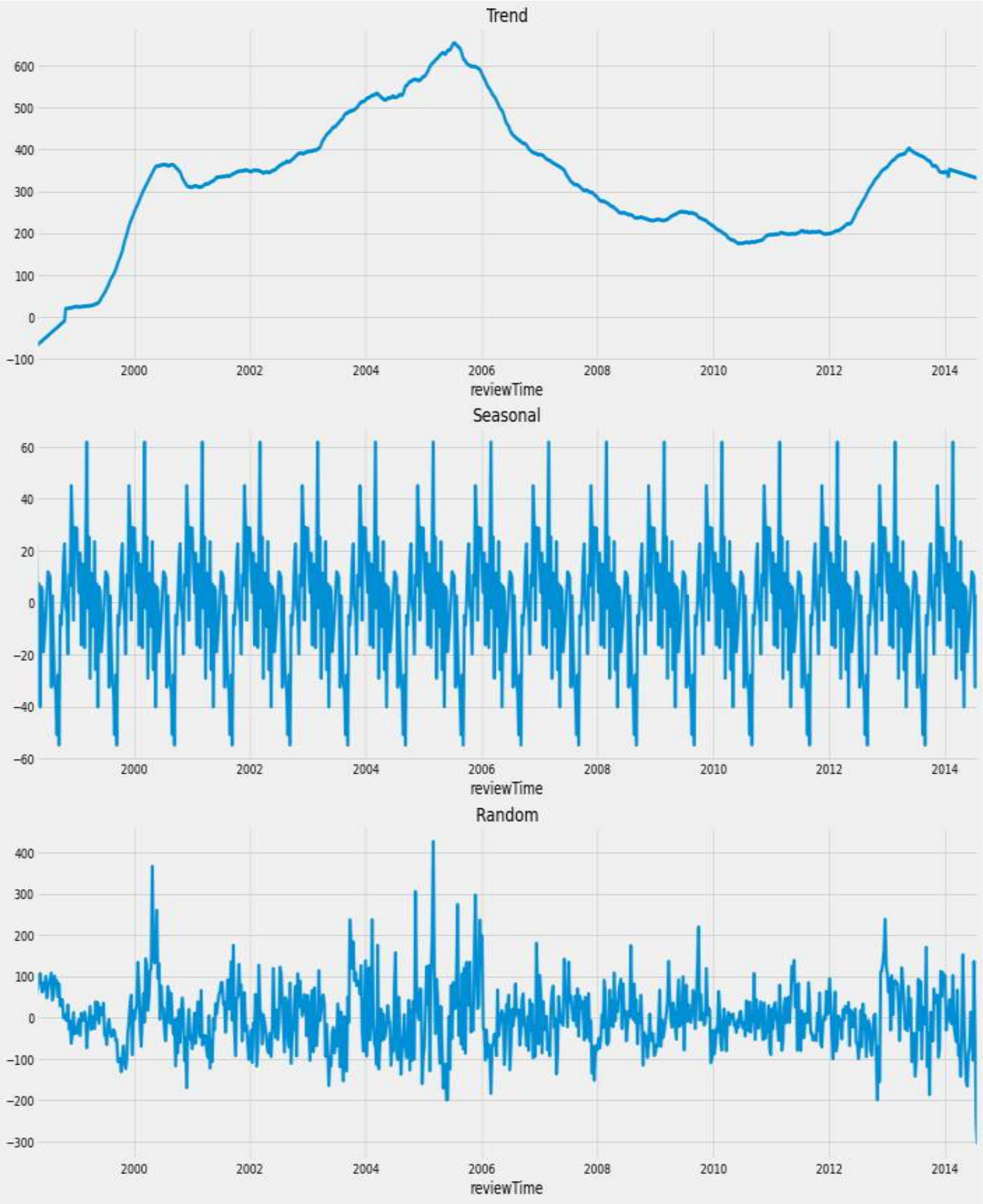
Trend: In which there is no fixed interval and any divergence within the given dataset is a continuous timeline. The trend would be negative or positive or null trend.

Seasonality: In which regular or fixed interval shifts within the dataset in a continuous timeline.

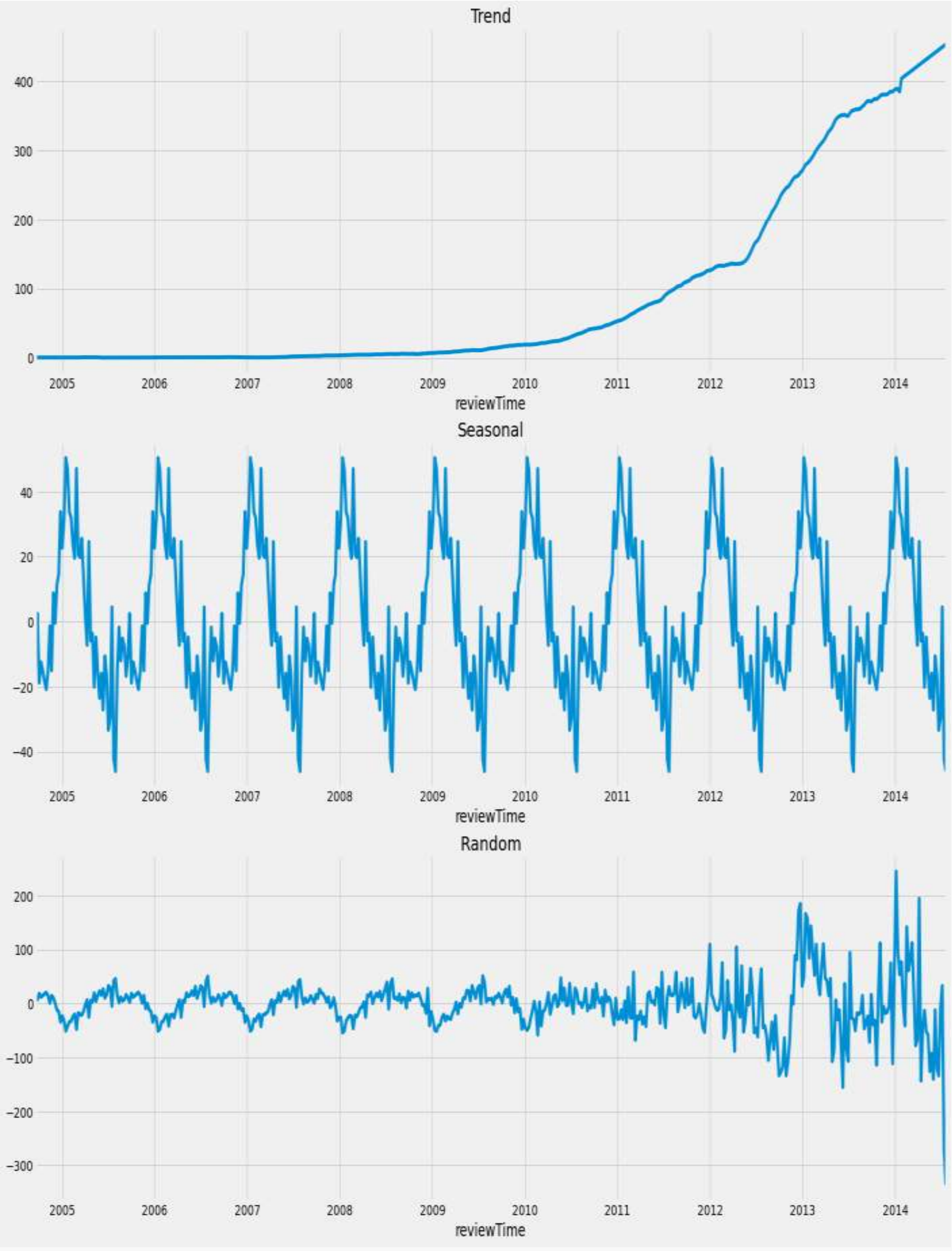
Cyclical: In which there is no fixed interval, uncertainty in movement and its pattern.

Irregularity: Unexpected situations/events/scenarios and spikes in a short time span.

For Digital Music Data:



For Musical instruments Data:



Data Types of Time Series:

Stationary: A dataset that should follow the below thumb rules, without having Trend, Seasonality, Cyclical, and Irregularity component of time series.

- The Mean value of them should be completely constant in the data during the analysis.
- The Variance should be constant with respect to the time-frame.
- The Covariance measures the relationship between two variables.

Non- Stationary: This is just the opposite of Stationary.

So how to make data stationary?

The most common approach is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.

ARIMA (Auto Regressive Integrated Moving Average) MODEL:

The ARIMA methodology is a statistical method for analyzing and building a forecasting model which best represents a time series by modeling the correlations in the data. Owing to purely statistical approaches, ARIMA models only need the historical data of a time series to generalize the forecast and manage to increase prediction accuracy while keeping the model parsimonious.

ARIMA is actually a class of models that 'explains' a given time series based on its own past values, that is its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

We can model any Time Series that are non-seasons exhibiting patterns and not a random white noise with ARIMA models.

Autoregression (AR):

Refers to a model that shows a changing variable that regresses on its own lagged, or prior, values. A statistical model is autoregressive if it predicts future values based on past values. For example, an autoregressive model might seek to predict a stock's future prices based on its past performance.

Integrated (I):

Represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).

Moving average (MA):

Moving Average:

Incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

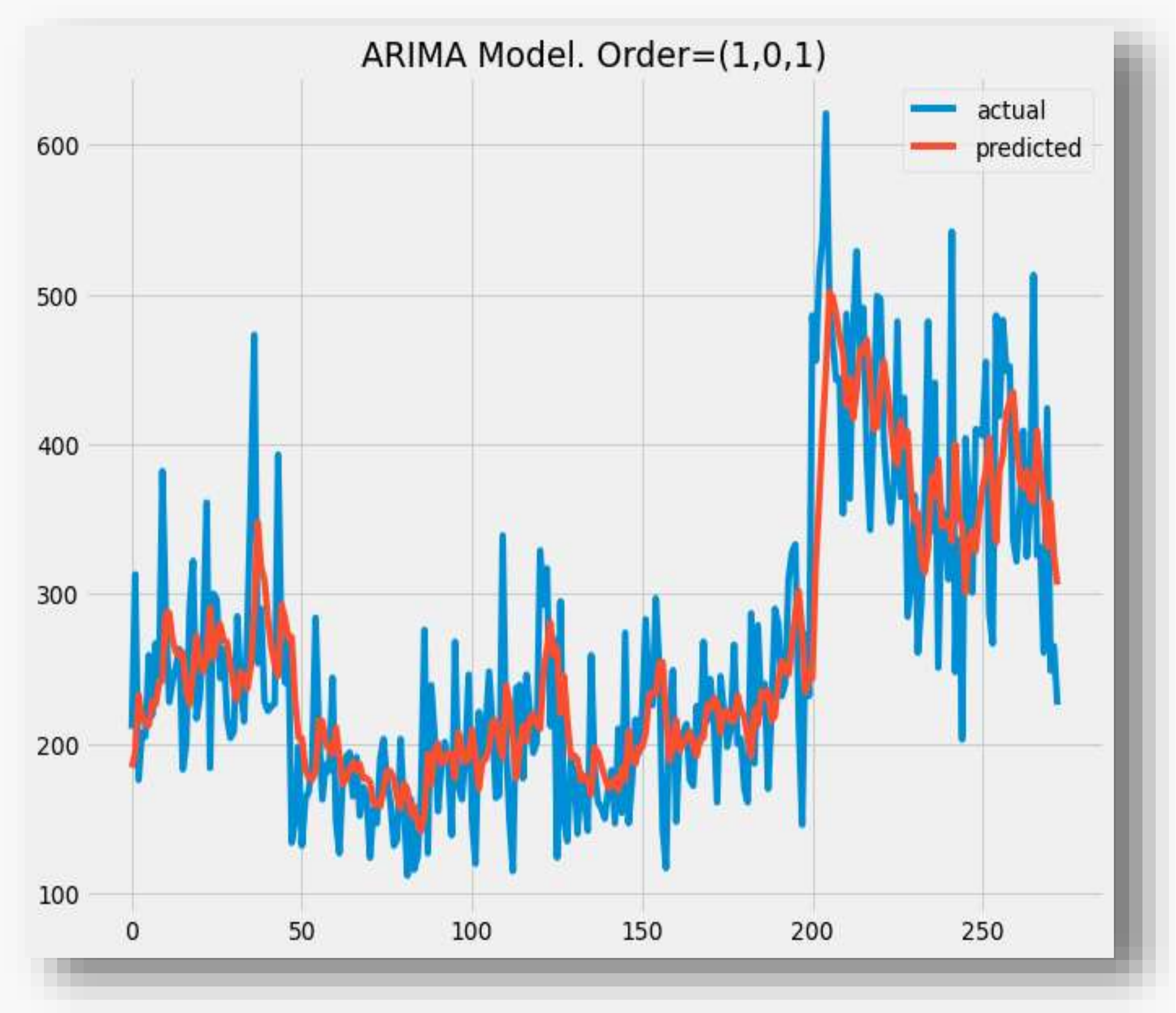
The commonly used time series method is Moving Average. This method is slick with random short-term variations. Relatively associated with the components of time series.

The Moving Average (MA) (Or) Rolling Mean: In which MA has calculated by taking averaging data of the time-series, within k periods.

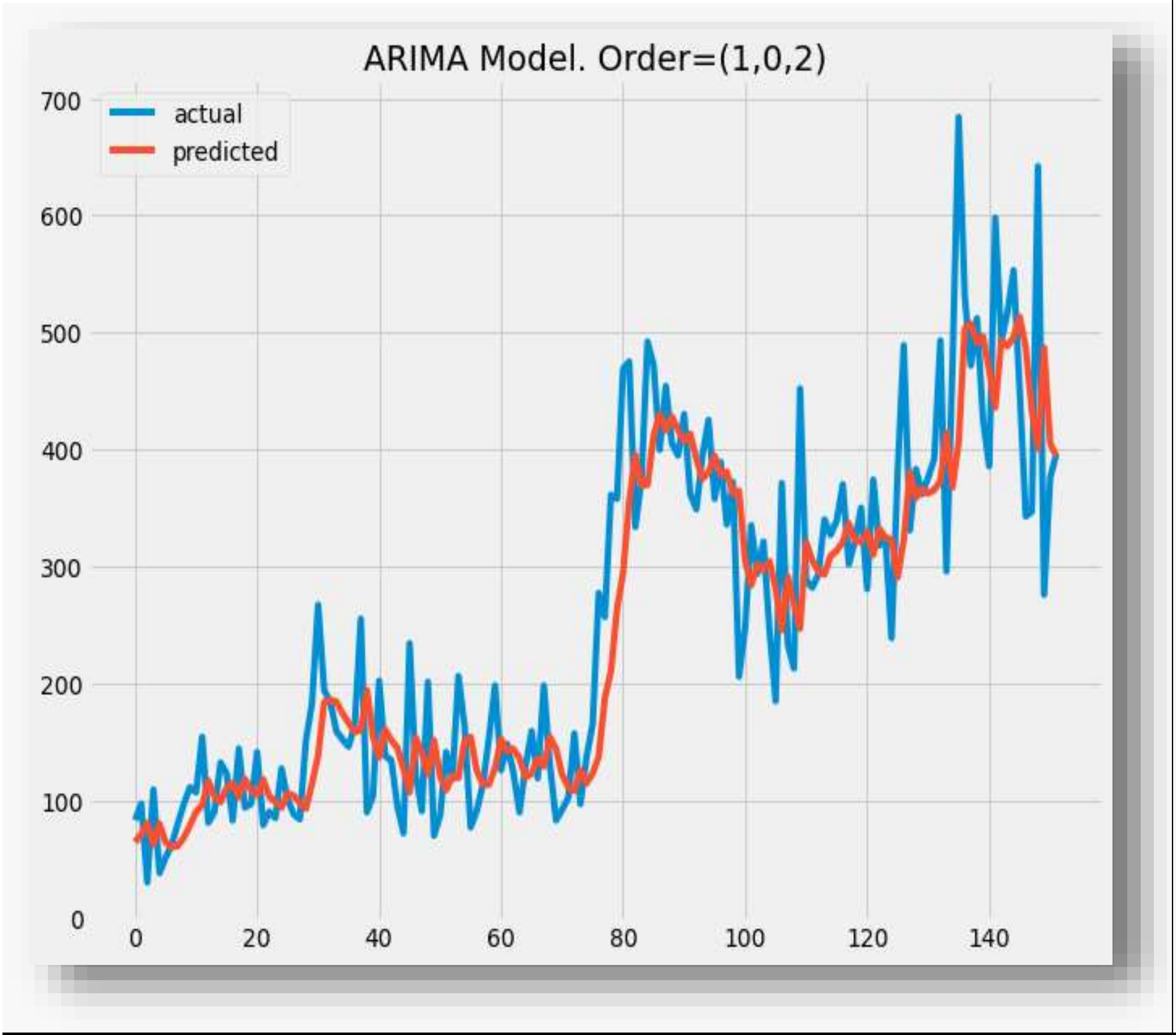
What does the p, d and q in ARIMA model mean?

- The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.
- ‘p’ is the order of the ‘Auto Regressive’ (AR) term. It refers to the number of lags of Y to be used as predictors.
- ‘q’ is the order of the ‘Moving Average’ (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

ARIMA Model for Digital Music Data:



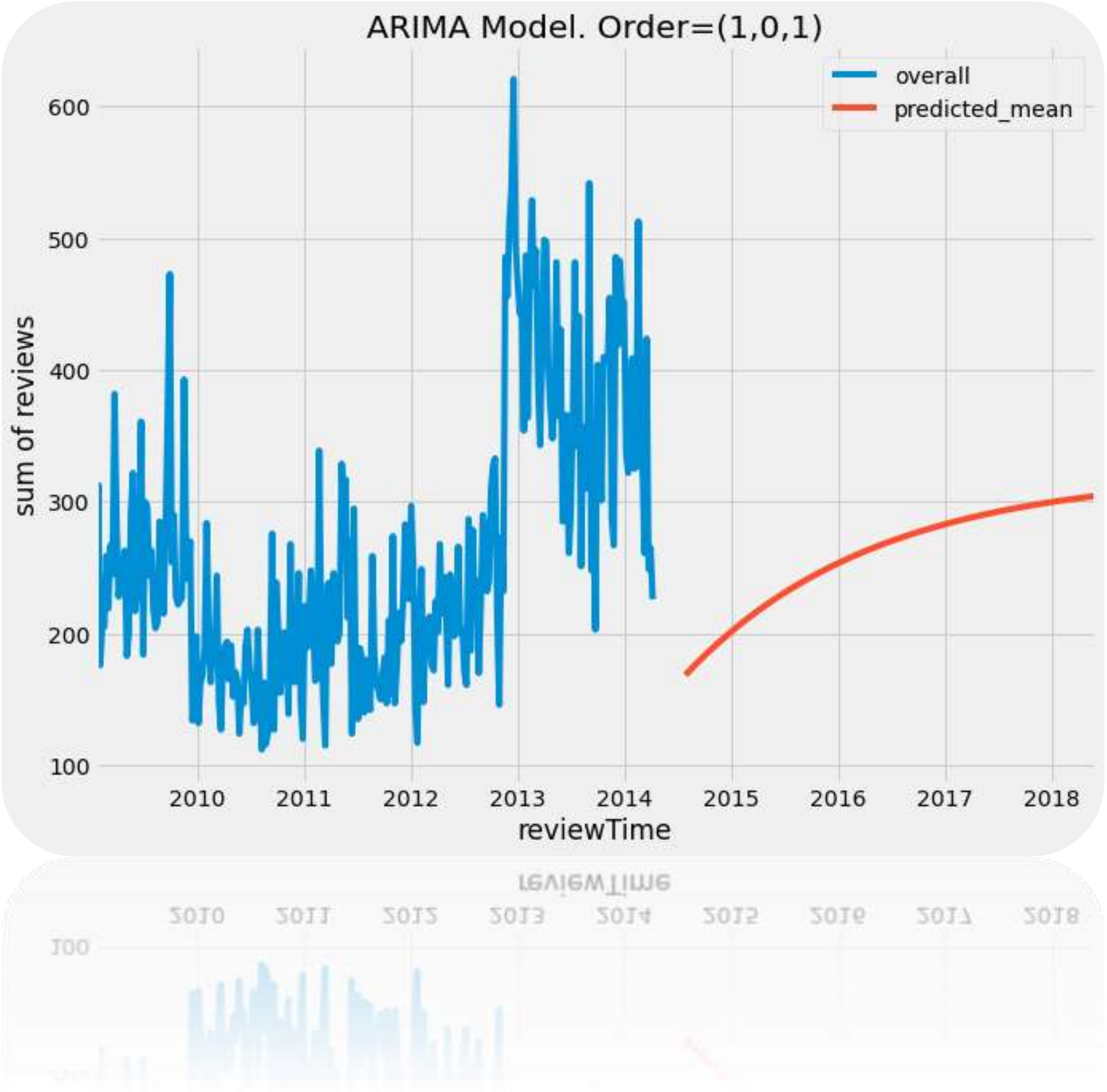
ARIMA Model for Musical Data:



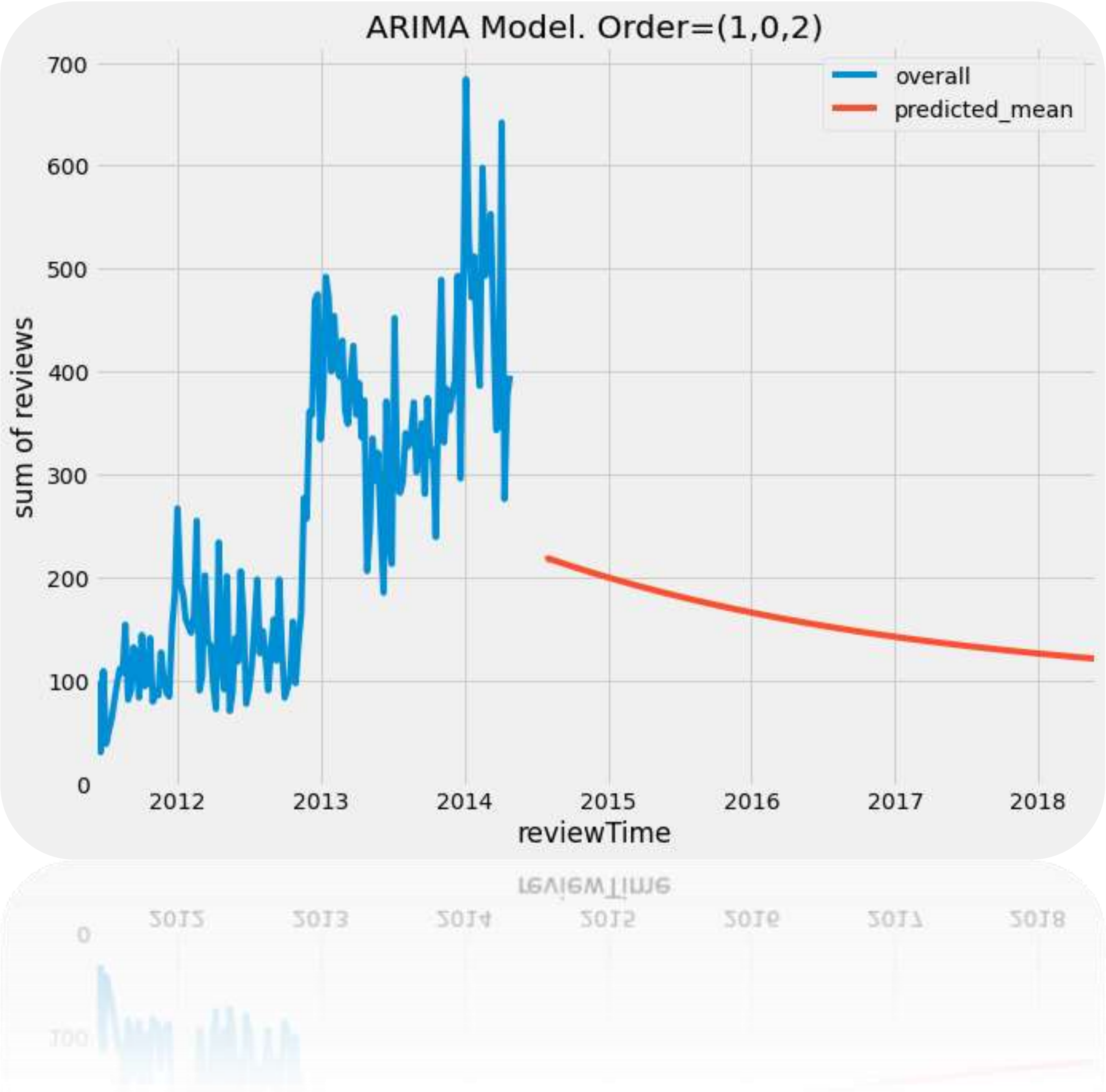
What is the need for ?

Since forecasting a Time Series, such as Sales and Demand, is often of incredible commercial value, which increases the need for forecasting.

Forecasting using ARIMA Model for Digital Music Data:



Forecasting using ARIMA Model for Musical Instruments Data:



SARIMA (Seasonal Auto Regressive Integrated Moving Average) MODEL:

If a Time Series has seasonal patterns, we have to insert seasonal periods, and it becomes **SARIMA**, short for **'Seasonal ARIMA'**.

There are three terms characterizing model: p, q, and d

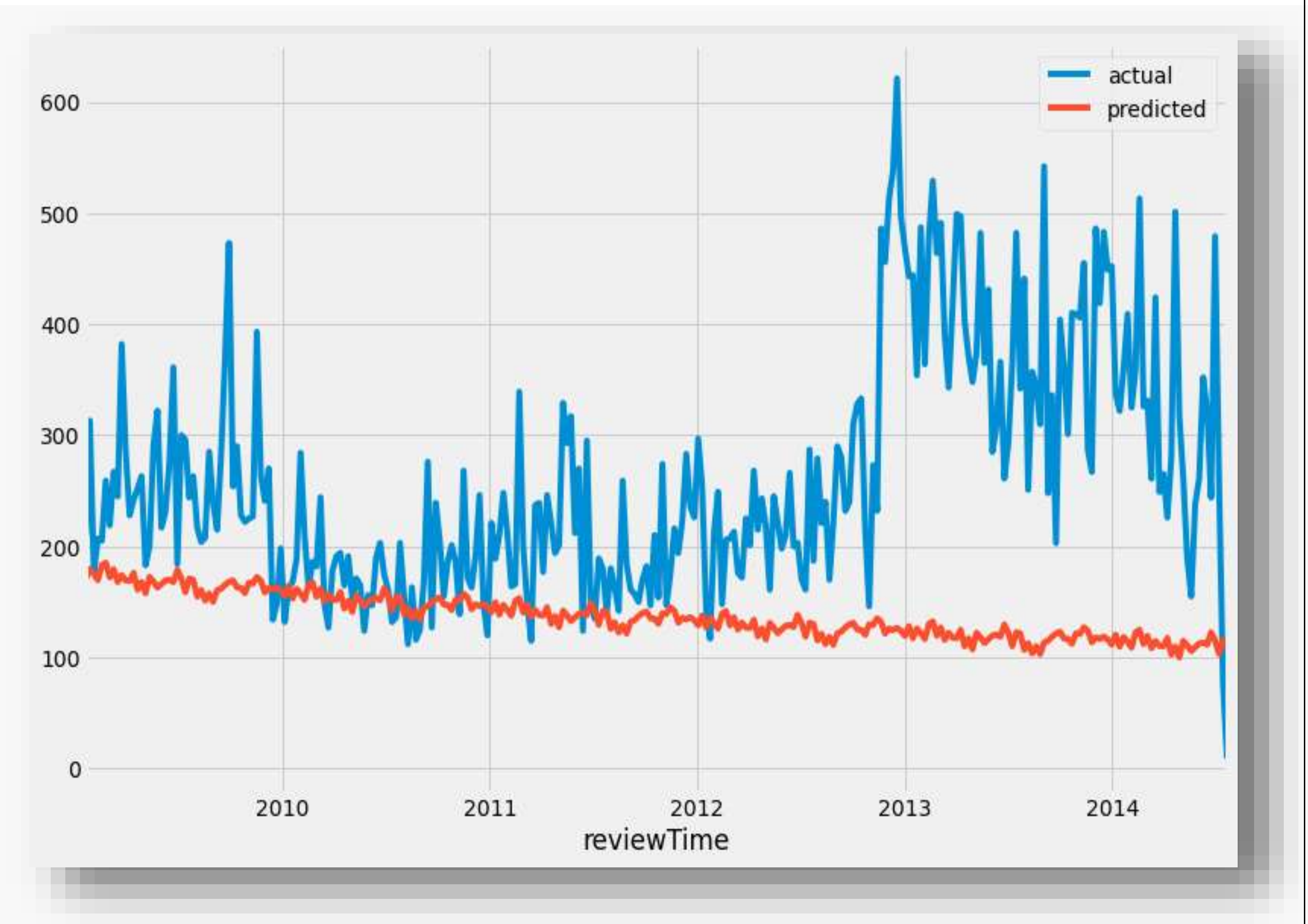
Where,

p = the order of the AR (auto regressive) term

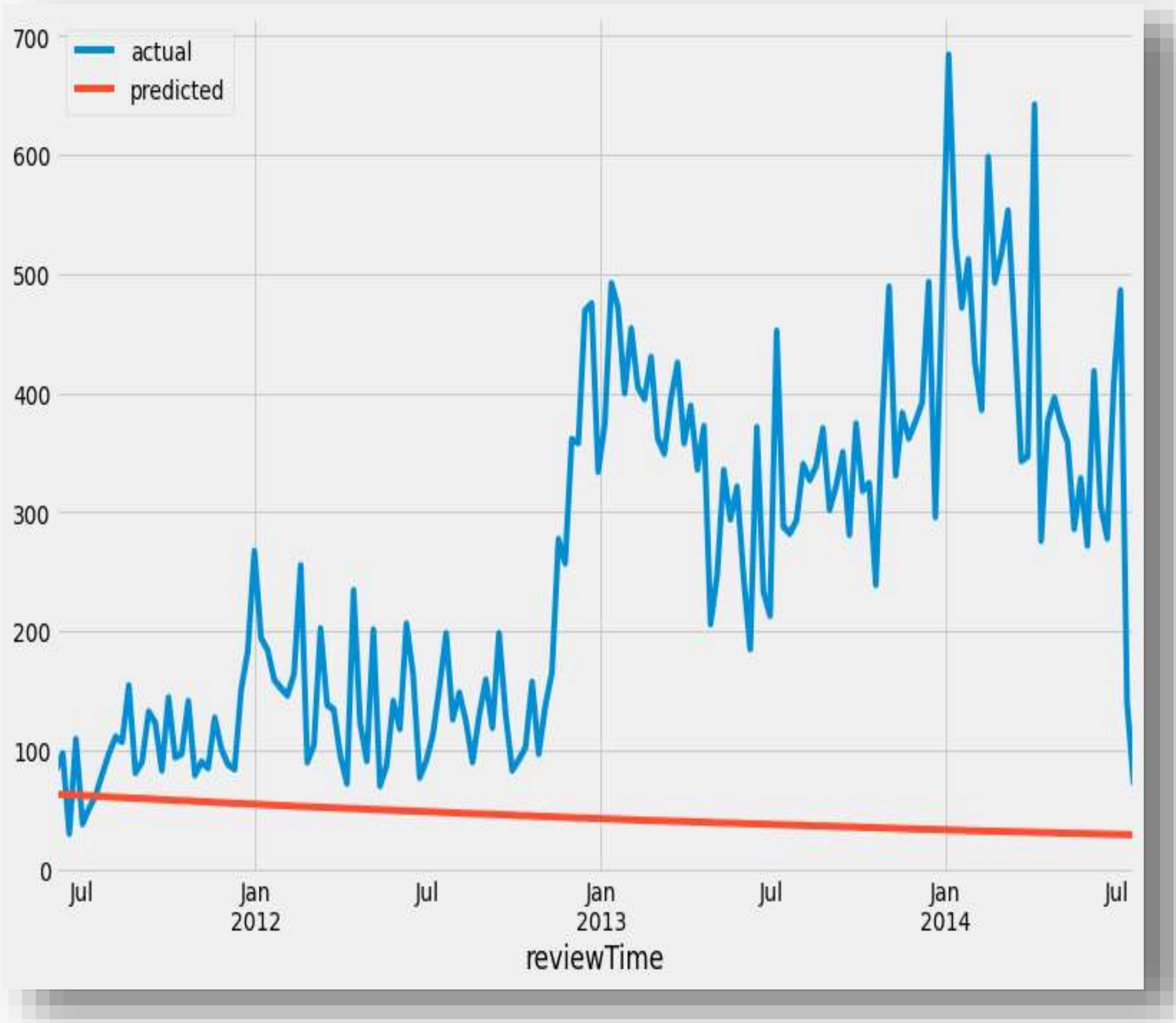
q = the order of the MA (moving average) term

d = the number of differences required to make the time series stationary

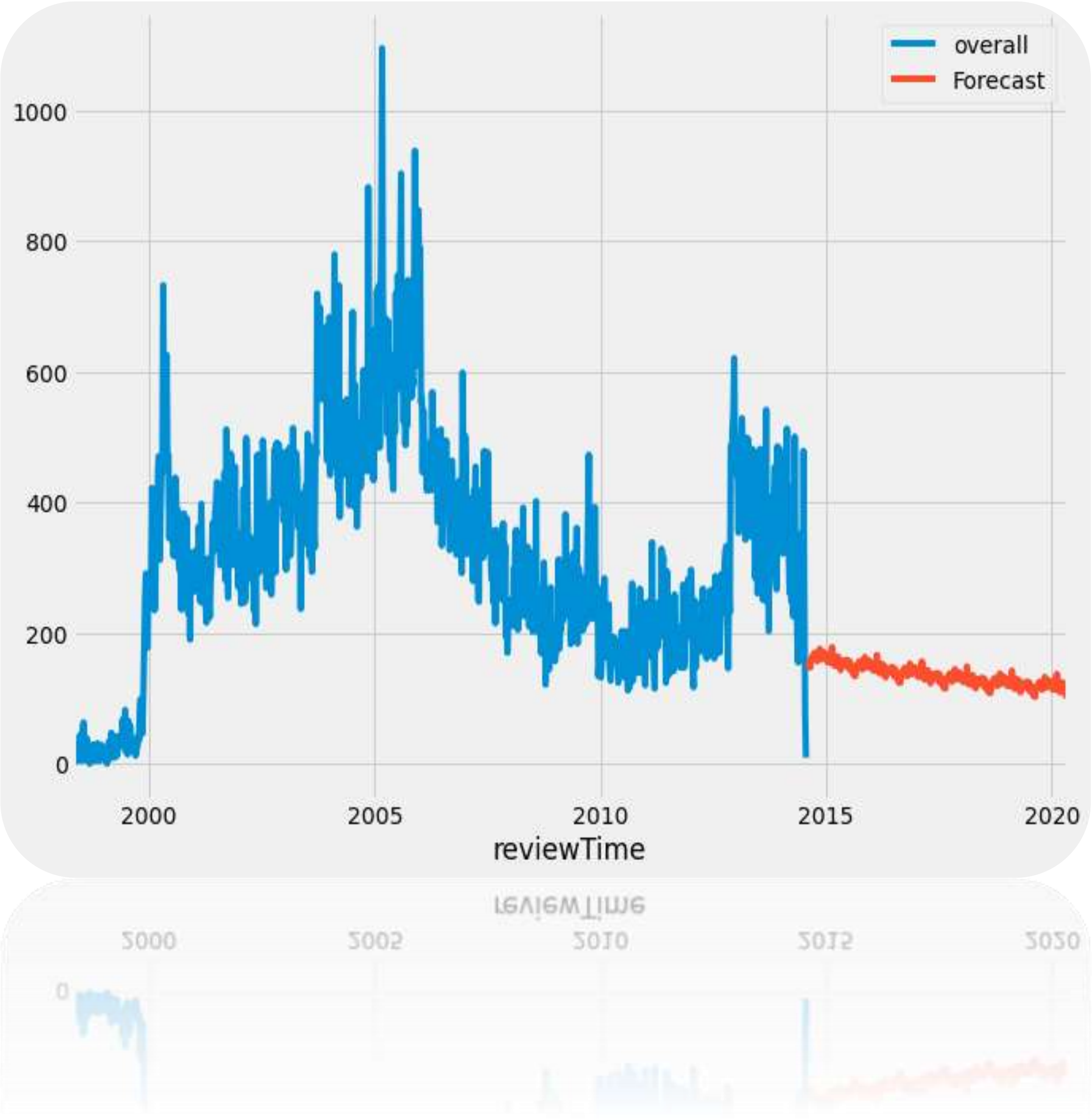
SARIMA Model for Digital Music Data:



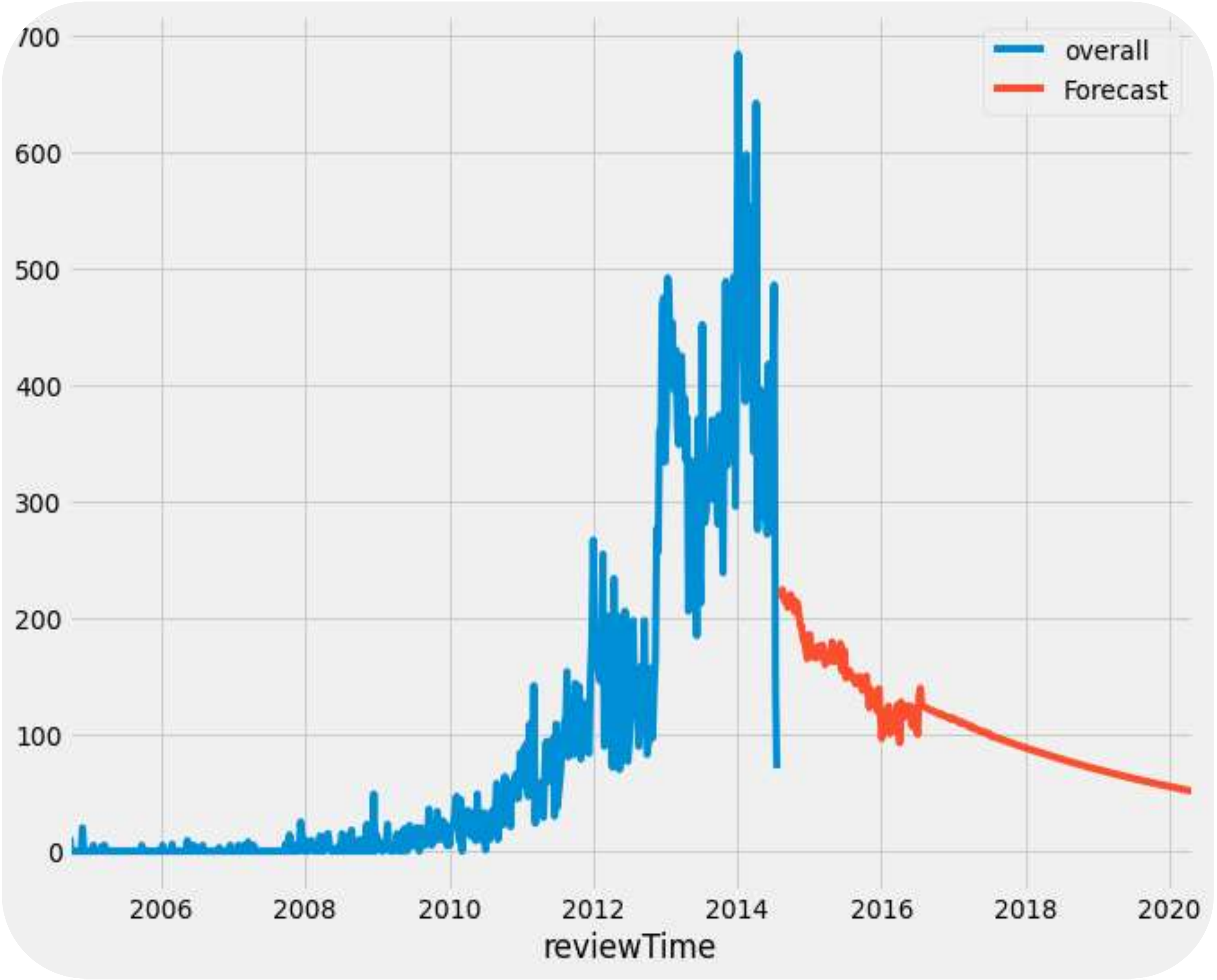
SARIMA Model for Musical Instruments Data:



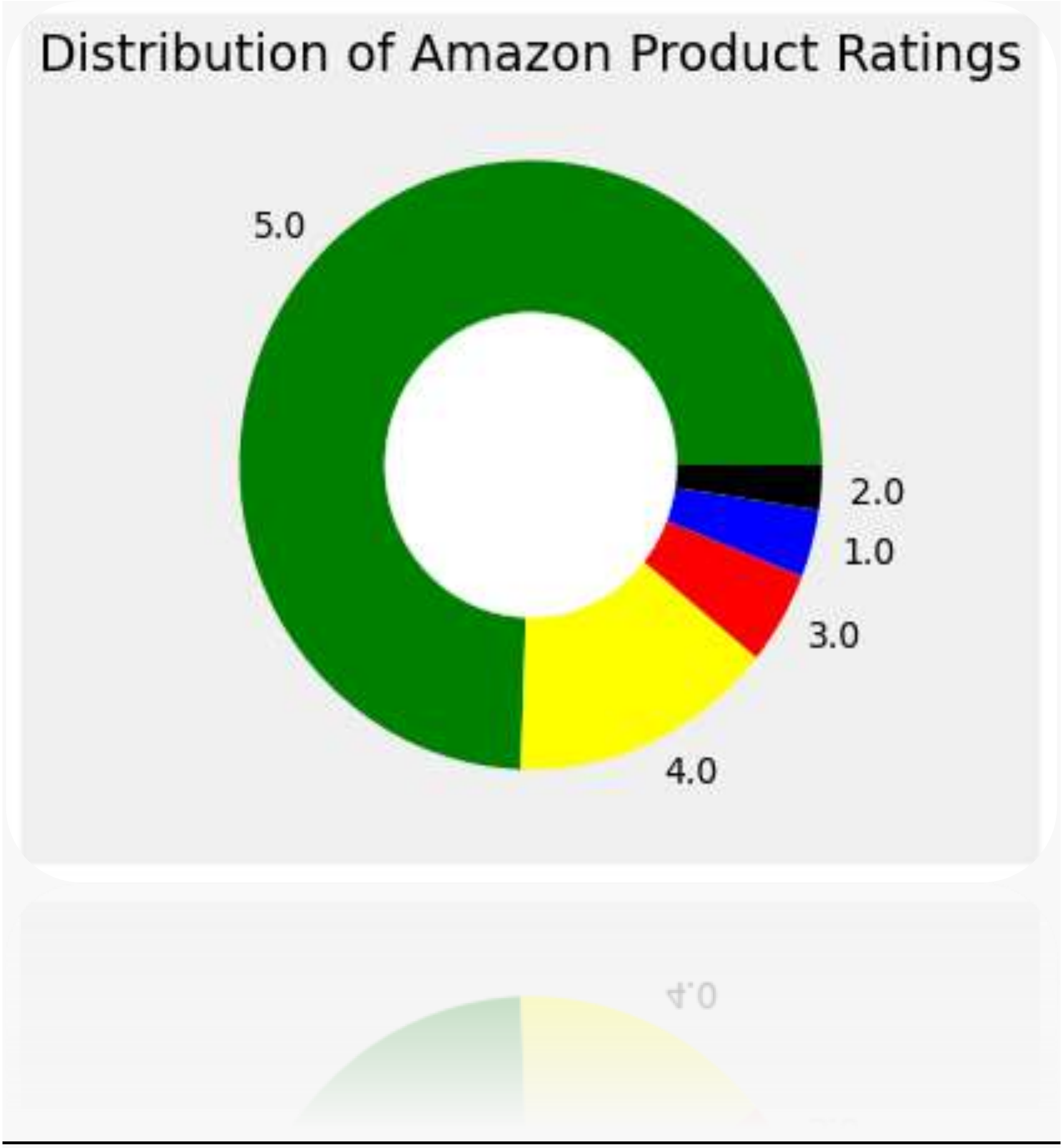
Forecasting using SARIMA Model for Digital Music Data:



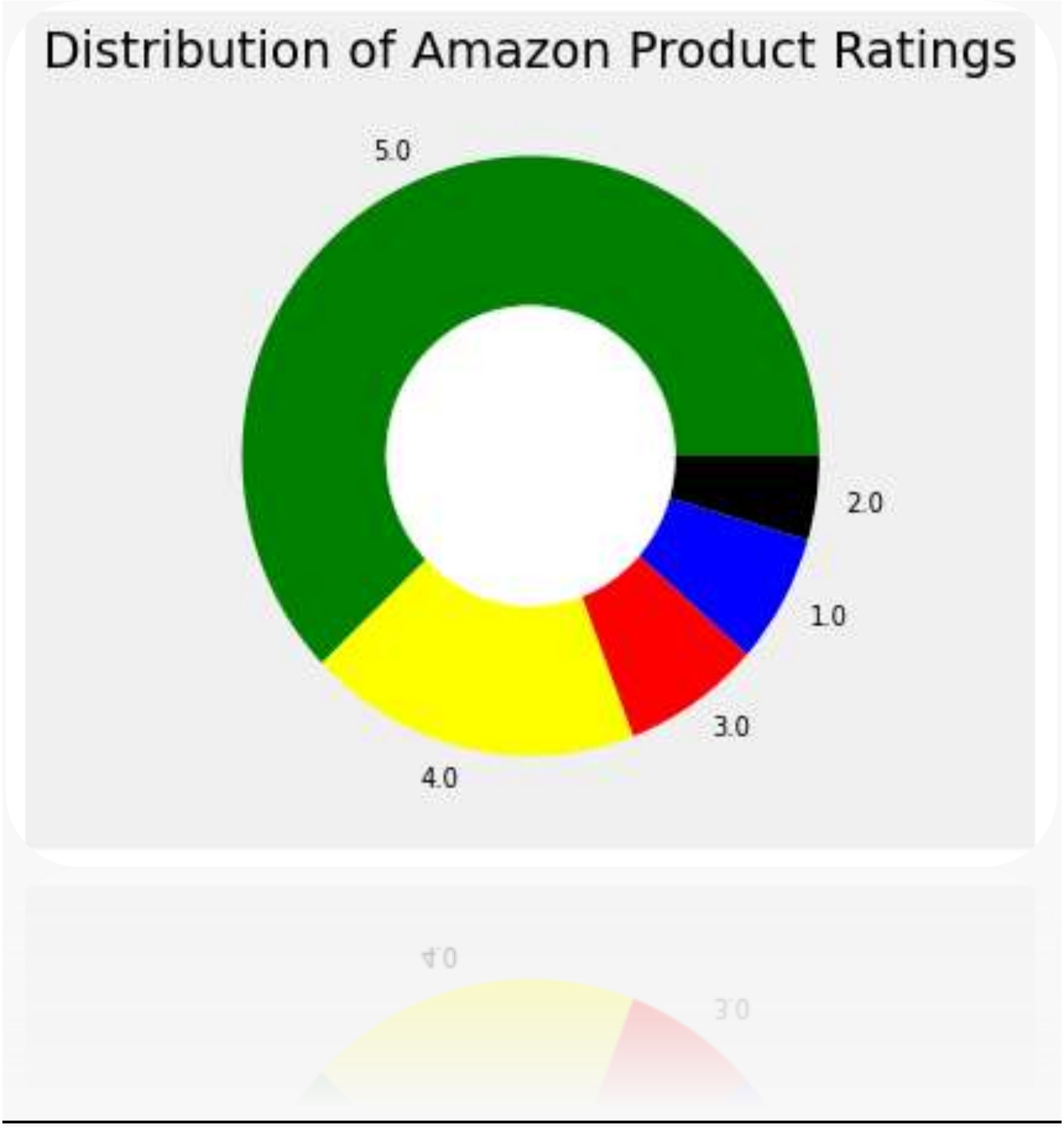
Forecasting using SARIMA Model for Musical Instruments Data:



Distribution of Ratings for Digital Music:



Distribution of Ratings for Musical Instruments:



CONCLUSION:

- By sentiment analysis we found that most of the reviews are positive for both Digital Music and Musical Instrument products.
- After 2005, from our graphs, we can see that there is dip in trend till 2010 in digital music. Which may be result of recession. Although we see both digital music and musical instruments is growing up after 2010.
- We can see in trend that In 2007 the market for digital music was declining the reason is that Amazon introduced music in mp3 format which enabled or allowed the consumers to easily transfer music among a variety of devices – from iPods to personal computers and compact discs.
- With the help of time series we can see there is chance of downward trend among buyers after 2015-16. To overcome this, introduction to new products or ideas may help.
- Our graphs also show the jumping rate of interest in buyers in musical instruments year on year. The idea of improving existing products may help to capture more market share by Amazon. For e.g. bringing new companies that are into same music business, this may provide more options for buyers.

REFERENCES:

www.geeksforgeeks.org

www.javatpoint.com

www.analyticsvidhya.com

www.capitalone.com