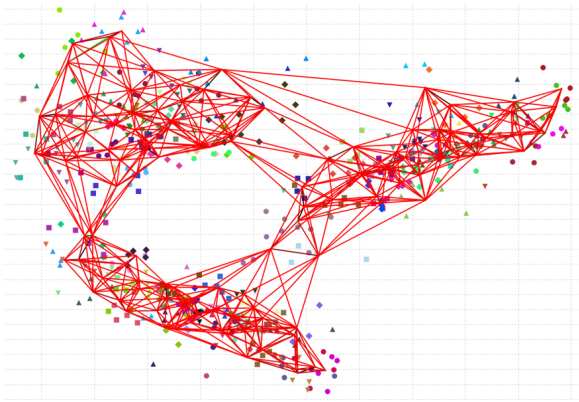
Exploratory Data Analysis on Spotify Metadata

Deepesh.D,Aniket.V,Ayush.T

Abstract:

Data analysis of spotify metadata taken from 1921 to 2020 using feature analysis etc.Investigating which audio features distinguish genres and try to cluster them using Dimension Reduction and Genre Clustering.Descriptive Analysis on Genre,artists and decade attributes.This paper focuses on verifying who the most popular artists are and to analyze the kind of music which goes viral,analyzing the behaviour of various genres over the years and to see how the popularity of various attributes vary with respect to time.

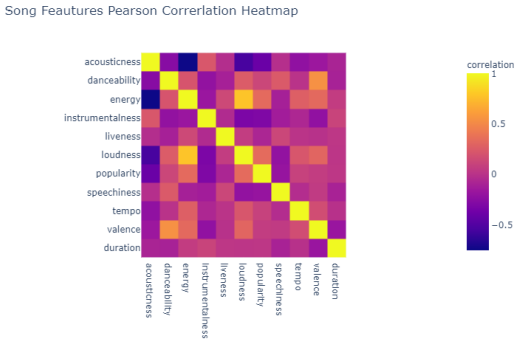


Introduction

Spotify is a well-known music streaming service that requires no introduction. It has expanded to 286 million active users and 130 million premium subscriptions in 14 years. Spotify is used by the average user for 25 hours a month, and 44 percent of users use it on a regular basis .

These figures clearly show that Spotify is a very popular streaming service, which can offer a lot of valuable data which can be studied.

The below image shows Song Features through Pearson Correlation Heatmap



Spotify offers digital copyright restricted recorded music and podcasts, including more than 70 million songs, from record labels and media companies.As a freemium service, basic features are free with advertisements and limited control, while additional features, such as offline listening and commercial-free listening, are offered via paid subscriptions. Users can search for music based on artist, album, or genre, and can create, edit, and share playlists.

Related Work

1. K-means clustering using Spotify song features Aug 27, 2020 - With the help of this paper we figured out how to handle the spotify data and how to perform and inference the Elbow method and genre clustering and Dimensional reduction

2. An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation - The number of songs in the playlist originally, bigger the initial dataset, better the prediction

3. Spotify Sequential Skip Prediction Paper - This paper is mainly useful in teaching us how to make use of the Spotify API to extract data from Spotify in order for processing

4. Music Streaming Sessions Paper - Studying this paper gave us an insight into how users behave, but also taught us that the data regarding this subject is quite scarce, and that only logs are available, when the scope of such data can be much greater. Now we know that the data available for user interaction does not hold much merit, we will not spend a lot of resources trying to visualize it.

5. Visualizing a Million Time Series with the Density Line Chart - This is a paper that deals with visualising extremely Dense Data. As the Spotify Dataset is vast, by studying the techiniques used in this paper to handle the dense data, we can learn a lot. For example, rather than attempting to visualize the entire data and find something, we must look for small patterns within smaller time frames, or lines that behave in particular ways. It also teaches how to use Denselines, should we require that.

6. Analyzing Spotify Data with Pandas - This article shows how to work with data from Spotify and visualize it. It shows how to work with correlations graphs, how to answer questions like “who makes the most Danceable music”, “What are the most Popular genre”, “What is the relationship between Danceability and Energy” using graphing and visualization techniques

7. HITPREDICT: PREDICTING HIT SONGS USING SPOTIFY DATA STANFORD COMPUTER SCIENCE 229: MACHINE LEARNING - With this paper, we understood that should we aim for a prediction using machine learning, LR and NN are the algorithms to go for

8. A survey of Scholarly Data Visualization (2017) - This paper gave us into an insight of how to perform an exploratory survey on

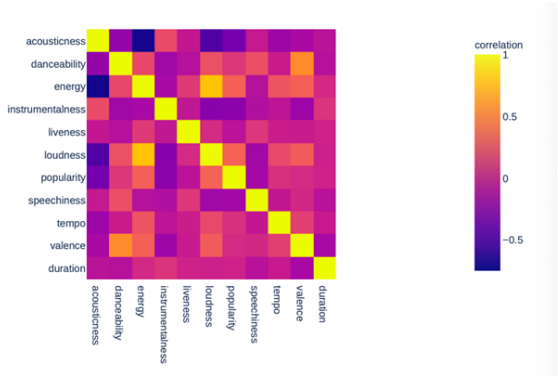
a topic with dense data. We could also learn how to form a Research paper, and how to approach an exploration project in data visualization

9. DataVis Helper: A Tool for Exploring the Design Space of Data Visualization - This paper touches on the most effective Data Visualization methods and tools that one must use while conducting exploratory analysis

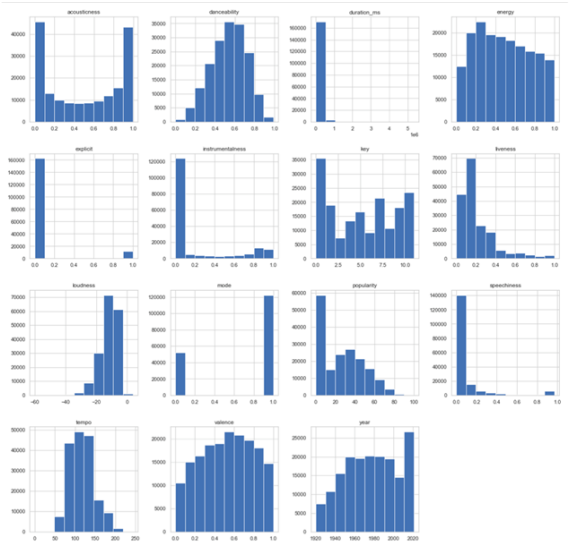
10. College data visualization on the web browser - This paper again gave us an insight on how to deal with dense data and perform exploratory analysis on such dense data, where to look for patterns and how to present them

Proposed Method and Results

1. First we clean the dataset, drop the unwanted columns and null values
2. We print and analyze the Oldest Record and the Most Recent Record
3. We plot a Pearsion Correlation Heatmap between all the attributes and analyze the correlation

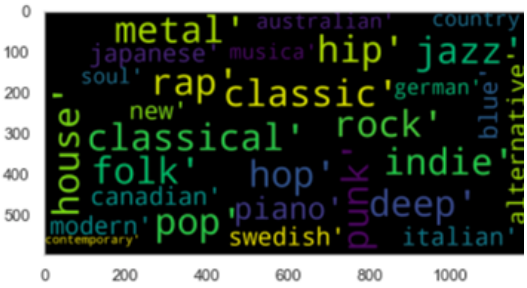


4. Conducted popularity analysis using bar chart to show to artists with popularity by mean, using line chart to compare the loudness with popularity, using a sns.joinplot and bar chart to visualize the popularity of the top 10 artists, and finally Looking at the overall Data Distribution using 15 Histograms

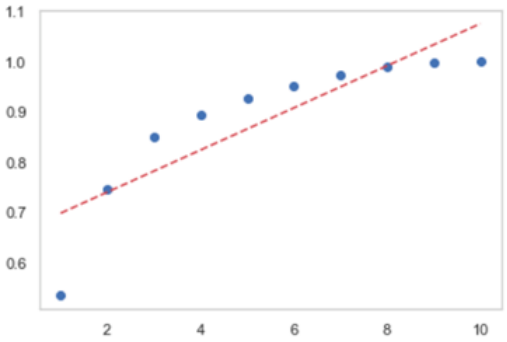


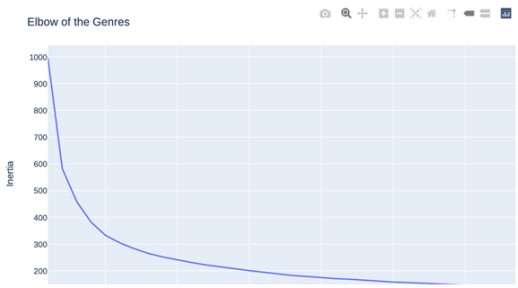
6. Compare and Check the correlation of specific features, like Danceability vs Speechiness
7. Using a Pie chart we measured the Popularity of the Keys and Modes

8. We the made a Wordcloud to see the most frequently used words



9. Finally we used Principal Component Analysis to cluster the various attributes, and we picked the first four components which covered 90% of the space. We analyzed the correlation. We performed Elbow analysis, and Figured that 20 clusters will be idea. But after performing Clustering, interesting results were not found.





Conclusion

EDA on Spotify dataset to analyze the data and plot graphs for showing correlation between song features,top artists with popularity by mean,loudness vs popularity,name of top 10 songs and artists by popularity,overall data distribution,danceability vs speechiness,Popularity vs

Acousticness,genre popularity,keys and modes,wordcloud for most common genres,analyzation by year and finally pca implementation for reducing the components.

We found out that Loudness correlates with energy, Iann Dior is the most popular Artist by mean, Danceability vs Speechiness have a perfect positive correlation, wherease popularity and acousticness have a negative correlation. G# is the most used key and Minor is the more popular mode. Classic, Hip, Classical are the three most frequently occuring words.

This analysis has shown how the enormous data collected by spotify holds so many interesting and valuable patterns and trends which one could examine and utilize to further optimize the app for a smoother user experience. All these trends are invisible to the human eye, and can only be understood by computers if the data remains numerical, but through data visualization, these patterns reveal themselves, so that the human eye can understand and work with such data, because while machine learning is a very helpful tool, human surveliance and intervention is always necessary, and this intervention can only be possible with the help of Data Visualization.

References

[1] G. Bonnin and D. Jannach. Automated generation of music playlists: Survey and experiments. ACM Computing Surveys, 47(2):26, 2015. [2] C. J. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical report, June 2010. [3] R. Burke. Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 12(4):331–370, Nov 2002. [4] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani. Recsys challenge 2018: Automatic music playlist continuation. In Proceedings of the 12th ACM Conference on Recommender Systems, RecSys ’18, Vancouver, BC, Canada, 2018. [5] R.-C. Chen, L. Gallagher, R. Blanco, and J. S. Culpepper. Efficient cost-aware cascade ranking in multi-stage retrieval. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’17, pages 445–454, New

York, NY, USA, 2017. ACM. [6] S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. Playlist prediction via metric embedding. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’12

Sentence

Exploratory Data Analysis on Spify MetadataDeepesh.D,Aniket.V,Ayush.T Abstract:Data analysis of spotify metadata taken from1921 to 2020 using feature analysisetc.Investigating which audio featuresdistinguish genres and try to cluster themusing Dimension Reduction and GenreClustering.Descriptive Analysis onGenre,artists and decade attributes.Thispaper focuses on verifying who the mostpopular artists are and to analyze the kind ofmusic which goes viral,analyzing thebehaviour of various genres over the yearsand to see how the popularity of variousattributes vary with respect to time.IntroductionSpotify is a well-known music streamingservice that requires no introduction. It hasexpanded to 286 million active users and130 million premium subscriptions in 14years. Spotify is used by the average user for25 hours a month, and 44 percent of usersuse it on a regular basis .These figures clearly show that Spotify is avery popular streaming service, which canoffer a lot of valuable data which can bestudied.The below image showsSong Featuresthrough Pearson Correlation Heatmap Spotify offers digital copyright restrictedrecorded music and podcasts, includingmore than 70 million songs, from recordlabels and media companies.As a freemiumservice, basic features are free withadvertisements and limited control, whileadditional features, such as offline listeningand commercial-free listening, are offeredvia paid subscriptions. Users can search formusic based on artist, album, or genre, andcan create, edit, and share playlists.Related Work1. K-means clustering using Spotify songfeatures Aug 27, 2020 - With the help of thispaper we figured out how to handle thespotify data and how to perform andinference the Elbow method and genreclustering and Dimensional reduction2. An Analysis of Approaches Taken in theACM RecSys Challenge 2018 for AutomaticMusic Playlist Continuation - The number ofsongs in the playlist originally, bigger theinitial dataset, better the prediction3. Spotify Sequential Skip Prediction Paper- This paper is mainly useful in teaching ushow to make use of the Spotify API toextract data from Spotify in order forprocessing4. Music Streaming Sessions Paper -Studying this paper gave us an insight intohow users behave, but also taught us that thedata regarding this subject is quite scarce,and that only logs are available, when thescope of such data can be much greater.Now we know that the data available foruser interaction does not hold much merit,we will not spend a lot of resources trying tovisualize it.5. Visualizing a Million Time Series withthe Density Line Chart - This is a paper thatdeals with visualising extremely Dense Data.As the Spotify Dataset is vast, by studyingthe techiniques used in this paper to handlethe dense data, we can learn a lot. Forexample, rather than attempting to visualizethe entire data and find something, we mustlook for small patterns within smaller timeframes, or lines that behave in particularways. It also teaches how to use Denselines,should we require that.6. Analyzing Spotify Data with Pandas -This article shows how to work with datafrom Spotify and visualize it. It shows howto work with correlations graphs, how toanswer questions like “who makes the mostDanceable music”, “What are the mostPopular genre”, “What is the relationshipbetween Danceability and Energy” usinggraphing and visualization techniques7. HITPREDICT: PREDICTING HITSONGS USING SPOTIFY DATASTANFORD COMPUTER SCIENCE 229:MACHINE LEARNING - With this paper,we understood that should we aim for aprediction using machine learning, LR andNN are the algorithms to go for8. A survey of Scholarly Data Visualization(2017) - This paper gave us into an insightof how to perform an exploratory survey on2 a topic with dense data. We could also learnhow to form a Research paper, and how toapproach an exploration project in datavisualization9. DataVis Helper: A Tool for Exploring theDesign Space of Data Visualization - Thispaper touches on the most effective DataVisualization methods and tools that onemust use while conducting exploratoryanalysis10. College data visualization on the webbrowser - This paper again gave us aninsight on how to deal with dense data andperform exploratory analysis on such densedata, where to look for patterns and how topresent themProposed Method and Results1.Firstwecleanthedataset,droptheunwantedcolumnsandnullvalues2.WeprintandanalyzetheOldestRecordandtheMostRecentRecord3.WeplotaPearsionCorrelationHeatmapbetweenalltheattributesandanalyzethecorrelation4.Conductedpopularityanalysisusingbarcharttoshowtoartistswithpopularitybymean,usinglinecharttocomparetheloudnesswithpopularity,usingasns.joinplotandbarcharttovisualizethepopularityofthetop10artists,andfinallyLookingattheoverallDataDistributionusing 15Histograms6.CompareandCheckthecorrelationofspecificfeatures,like Danceability vs Speechiness7.UsingaPiechartwemeasuredthePopularityoftheKeysandModes3 8.WethemadeaWordcloudtoseethe most frequently used words9.FinallyweusedPrincipalComponentAnalysistoclusterthevariousattributes,andwepickedthefirstfourcomponentswhichcovered90%ofthespace.Weanalyzedthecorrelation.WeperformedElbowanalysis,andFiguredthat20clusterswillbeidea.But afterperformingClustering,interestingresultswere not found.ConclusionEDAonSpotifydatasettoanalyzethedataandplotgraphsforshowingcorrelationbetweensongfeatures,topartistswithpopularitybymean,loudnessvspopularity,nameoftop10songsandartistsbypopularity,overalldatadistribution,danceabilityvsspeechiness,Popularityvs4

Acousticness,genrepopularity,keysandmodes,wordcloudformostcommongenres,analyzationbyyearandfinallypcaimplementationforreducingthecomponents.WefoundoutthatLoudnesscorrelateswithenergy,IannDioristhemostpopularArtistbymean,DanceabilityvsSpeechinesshaveaperfectpositivecorrelation,whereasepopularityandacousticnesshaveanegativecorrelation.G#isthemostusedkeyandMinoristhemorepopularmode.Classic,Hip,Classicalarethethreemostfrequentlyoccuring words.Thisanalysishasshownhowtheenormousdatacollectedbyspotifyholdssomanyinterestingandvaluablepatternsandtrendswhichonecouldexamineandutilizetofurtheroptimizetheappforasmootheruserexperience.Allthesetrendsareinvisibletothehumaneye,andcanonlybeunderstoodbycomputersifthedataremainsnumerical,butthroughdatavisualization,thesepatternsrevealthemselves,sothatthehumaneyecanunderstandandworkwithsuchdata,becausewhilemachinelearningisaveryhelpfultool,humansurvelianceandinterventionisalwaysnecessary,andthisinterventioncanonlybepossiblewiththehelp of Data Visualization.References[1]G.BonninandD.Jannach.Automatedgenerationofmusicplaylists:Surveyandexperiments.ACMComputingSurveys,47(2):26,2015.[2]C.J.Burges.Fromranknettolambdaranktolambdamart:Anoverview.Technicalreport,June2010.[3]R.Burke.Hybridrecommendersystems:Surveyandexperiments.UserModelingandUser-AdaptedInteraction,12(4):331–370,Nov2002.[4]C.-W.Chen,P.Lamere,M.Schedl,andH.Zamani.Recsyschallenge2018:Automaticmusicplaylistcontinuation.InProceedingsofthe12thACMConferenceonRecommenderSystems,RecSys'18,Vancouver,BC,Canada,2018.[5]R.-C.Chen,L.Gallagher,R.Blanco,andJ.S.Culpepper.Efficientcost-awarecascaderankinginmulti-stageretrieval.InProceedingsofthe40thInternationalACMSIGIRConferenceonResearchandDevelopmentinInformationRetrieval,SIGIR'17,pages445–454,New5 York,NY,USA,2017.ACM.[6]S.Chen,J.L.Moore,D.Turnbull,andT.Joachims.Playlistpredictionviametricembedding.InProceedingsofthe18thACMSIGKDDInternationalConferenceonKnowledgeDiscovery and Data Mining, KDD '126

Report Title:	Report
Report Link: (Use this link to send report to anyone)	https://www.check-plagiarism.com/plag-report/28346dd2b19ecf7bc537e6fb5c65e9089a2661622744382
Report Generated Date:	03 June, 2021
Total Words:	640
Total Characters:	7498
Keywords/Total Words Ratio:	0%
Excluded URL:	No
Unique:	**91%**
Matched:	**9%**

Sentence wise detail:

Exploratory Data Analysis on Spify MetadataDeepesh. D,Aniket. V,Ayush.

T Abstract:Data analysis of spotify metadata taken from1921 to 2020 using feature analysisetc.

Investigating which audio featuresdistinguish genres and try to cluster themusing Dimension Reduction and GenreClustering.

Descriptive Analysis onGenre,artists and decade attributes.

Thispaper focuses on verifying who the mostpopular artists are and to analyze the kind ofmusic which goes viral,analyzing

thebehaviour of various genres over the yearsand to see how the popularity of variousattributes vary with respect to time.

IntroductionSpotify is a well-known music streamingservice that requires no introduction.

It hasexpanded to 286 million active users and130 million premium subscriptions in 14years.

Spotify is used by the average user for25 hours a month, and 44 percent of usersuse it on a regular basis .

These figures clearly show that Spotify is avery popular streaming service, which canoffer a lot of valuable data which can bestudied.

The below image showsSong Featuresthrough Pearson Correlation Heatmap Spotify offers digital copyright restrictedrecorded music and podcasts, includingmore than 70 million songs, from recordlabels and

media companies.

As a freemiumservice, basic features are free withadvertisements and limited control, whileadditional features, such as offline listeningand commercial-free listening, are offeredvia paid subscriptions. (0)

Users can search formusic based on artist, album, or genre, andcan create, edit, and share playlists. Related Work1.

K-means clustering using Spotify songfeatures Aug 27, 2020 - With the help of thispaper we figured

out how to handle thespotify data and how to perform andinference the Elbow method and genreclustering and Dimensional reduction2. (1)

An Analysis of Approaches Taken in theACM RecSys Challenge 2018 for AutomaticMusic Playlist

Continuation - The number ofsongs in the playlist originally, bigger theinitial dataset, better the

prediction3.

Spotify Sequential Skip Prediction Paper- This paper is mainly useful in teaching ushow

to make use of the Spotify API toextract data from Spotify in order forprocessing4.

Music Streaming Sessions Paper -Studying this paper gave us an insight intohow users behave, but also taught us that thedata

regarding this subject is quite scarce,and that only logs are available, when thescope of such data can be much greater.

Now we know that the data available foruser interaction does not hold much merit,we will not spend a lot of resources trying tovisualize it.5.

Visualizing a Million Time Series withthe Density Line Chart - This is a paper thatdeals with visualising extremely Dense Data.

As the Spotify Dataset is vast, by studyingthe techiniques used in this paper to handlethe dense data, we can learn a lot.

Forexample, rather than attempting to visualizethe entire data and find something, we mustlook for small patterns within smaller timeframes, or lines that behave in

particularways.

It also teaches how to use Denselines,should we require that.6.

Analyzing Spotify Data with Pandas -This article shows how to work with datafrom Spotify and visualize it.

It shows howto work with correlations graphs, how toanswer questions like “who makes the mostDanceable

music”, “What are the mostPopular genre”, “What is the relationshipbetween Danceability and Energy” usinggraphing and visualization

techniques7.

HITPREDICT: PREDICTING HITSONGS USING SPOTIFY DATASTANFORD COMPUTER SCIENCE 229:MACHINE LEARNING - With this paper,we understood that

should we aim for aprediction using machine learning, LR andNN are the algorithms to go for8.

A survey of Scholarly Data Visualization(2017) - This paper gave us into an insightof how to perform an exploratory survey on2 a topic with dense data.

We could also learnhow to form a Research paper, and how toapproach an exploration project in datavisualization9.

DataVis Helper: A Tool for Exploring theDesign Space of Data Visualization - Thispaper

touches on the most effective DataVisualization methods and tools that onemust use while conducting exploratoryanalysis10. (2)

College data visualization on the webbrowser - This paper again gave us aninsight on how to deal with

dense data andperform exploratory analysis on such densedata, where to look for patterns and how topresent themProposed Method and

Results1.

Firstwecleanthedataset,droptheunwantedcolumnsandnullvalues2.

WeprintandanalyzetheOldestRecordandtheMostRecentRecord3.

WeplotaPearsionCorrelationHeatmapbetweenalltheattributesandanalyzethecorrelation4.

Conductedpopularityanalysisusingbarcharttoshowtoartistswithpopularitybymean,usinglinecharttocomparetheloudnesswithpopularity,usingasns.

joinplotandbarcharttovisualizethepopularityofthetop10artists,andfinallyLookingattheoverallDataDistributionusing15Histograms6.

CompareandCheckthecorrelationofspecificfeatures,like Danceability vs Speechiness7.

UsingaPiechartwemeasuredthePopularityoftheKeysandModes3 8.

WethemadeaWordcloudtoseethe most frequently used words9.

FinallyweusedPrincipalComponentAnalysistoclusterthevariousattributes,andwepickedthefirstfourcomponentswhichcovered90%ofthespace. Weanalyzedthecorrelation. (3)

WeperformedElbowanalysis,andFiguredthat20clusterswillbeidea.

ButafterperformingClustering,interestingresultswere not found.

ConclusionEDAonSpotifydatasettoanalyzethedataandplotgraphsforshowingcorrelationbetweensongfeatures,topartistswithpopularitybymean,loudnessvspopularity,nameoftop10songsandartistsbypopularity,overalldatadistribution,danceabilityvsspeechiness,Popularityvs4
Acousticness,genrepopularity,keysandmodes,wordcloudformostcommongenres,analyzationbyyearandfinallypcaimplementationforreducingthecomponents.

WefoundoutthatLoudnesscorrelateswithenergy,IannDioristhemostpopularArtistbymean,DanceabilityvsSpeechinesshaveaperfectpositivecorrelation,whereasepopularityandacousticnesshaveanegativecorrelation.

G#isthemostusedkeyandMinoristhemorepopularmode.

Classic,Hip,Classicalarethethreemostfrequentlyoccuring words.

Thisanalysishasshownhowtheenormousdatacollectedbyspotifyholdssomanyinterestingandvaluablepatternsandtrendswhichonecouldexamineandutilizetofurtheroptimizetheappforasmootheruserexperience.

Allthesetrendsareinvisibletothehumaneye,andcanonlybeunderstoodbycomputersifthedataremainsnumerical,butthroughdatavisualization,thesepatternsrevealthemselves,sothatthehumaneyecanunderstandandworkwithsuchdata,becausewhilemachinelearningisaveryhelpfultool,humansurvelianceandinterventionisalwaysnecessary,andthisinterventioncanonlybepossiblewiththehelp of Data Visualization. References[1]G. BonninandD. Jannach.

Automatedgenerationofmusicplaylists:Surveyandexperiments.

ACMComputingSurveys,47(2):26,2015.[2]C. J. Burges.

Fromranknettolambdaranktolambdamart:Anoverview.

Technicalreport,June2010.[3]R. Burke. Hybridrecommendersystems:Surveyandexperiments. (4)

UserModelingandUser-AdaptedInteraction,12(4):331–370,Nov2002.[4]C.-W. Chen,P. Lamere,M. Schedl,andH. Zamani.

Recsyschallenge2018:Automaticmusicplaylistcontinuation.

InProceedingsofthe12thACMConferenceonRecommenderSystems,RecSys'18,Vancouver,BC,Canada,2018.[5]R.-C. Chen,L. Gallagher,R. Blanco,andJ. S. Culpepper.

Efficientcost-awarecascaderankinginmulti-stageretrieval.

InProceedingsofthe40thInternationalACMSIGIRConferenceonResearchandDevelopmentinInformationRetrieval,SIGIR'17,pages445–454,New5 York,NY,USA,2017. ACM.[6]S. Chen,J. L. Moore,D. Turnbull,andT. Joachims.
Playlistpredictionviametricembedding. (5)

InProceedingsofthe18thACMSIGKDDInternationalConferenceonKnowledgeDiscovery and Data Mining, KDD '126

Match Urls:

0: https://account.microsoft.com/account/ManageMyAccount?destrt=services-landing

1: https://medium.com/@chitu_rk/dimensionality-reduction-what-problem-does-it-solve-862364980a2f

2: https://pubmed.ncbi.nlm.nih.gov/23243055/

3: https://www.jaad.org/article/S0190-9622(10)01704-4/pdf

4: http://www.microlinkcolleges.net/elib/files/undergraduate/Photography/504705.pdf

5: https://www.bibsonomy.org/bibtex/48e43cae971c85967d4ef52e6cfd1836

Keywords Density

One Word	2 Words	3 Words
data 7.57%	spotify data 0.8%	exploratory data analysis 0.2%
popular 3.39%	data visualization 0.8%	datavisualization9 datavis helper 0.2%
spotify 3.19%	dense data 0.8%	scholarly data visualization 0.2%
paper 2.59%	paper gave 0.6%	2017 paper gave 0.2%
popularity 2.39%	data learn 0.4%	insightof perform exploratory 0.2%

Plagiarism Report

By check-plagiarism.com