A

PROJECT REPORT

ON

“Data science using Python”

SUBMITTED

To

CENTRE FOR ONLINE LEARNING

Dr. D.Y .PATIL VIDYADEETH, PUNE

DPU Dr. D. Y. Patil Vidyapeeth, Pune
(Deemed to be University)
CENTRE FOR ONLINE LEARNING

IN PARTIAL FULFILMENT OF DEGREE OF

MASTERS OF BUSINESS ADMISTRATION

BY

Ruturaj Nitin Bhambure

PRN: 2005020350

BATCH 2020-2022

DPU Dr. D. Y. Patil Vidyapeeth, Pune
(Deemed to be University)
CENTRE FOR ONLINE LEARNING

Dr. D.Y. Patil Vidyapeeth's
CENTRE FOR ONLINE LEARNING,
Sant Tukaram Nagar, Pune.

CERTIFICATE

This is to certify that Mr. Ruturaj Nitin Bhambure PRN - 2005020350 has completed his/her internship at Qollabb Edutech Starting from 14.05.2022 to 10.07.2022. His / Her project work was a part of the MBA (ONLINE LEARNING)

The project is on Data science using python (Data analysis and Data visualization of Spotify Dataset). Which includes research as well as industry practices. He/ She was very sincere and committed in all tasks.

Course Coordinator   Director

____________________   ______________________

Date -

COMPANY LETTER

(TO BE PROVIDED BY THE COMPANY WHERE THE PROJECT WILL BE CARRIED OUT)

To whomsoever it may concern

This is to certify that Mr. Ruturaj Nitin Bhambure PRN - 2005020350 has completed his/her internship at Qollabb Edutech. Starting from 14.05.2022 to 10.07.2022. His / Her project work was a part of the MBA (ONLINE LEARNING)

The project is on Data science using python (Data analysis and Data visualization of Spotify Dataset). Which includes research as well as industry practices. He/ She was very sincere and committed in all tasks.

Signature & Seal of Industry Guide

DECLARATION BY STUDENT

This is to declare that I have carried out this project work myself in part fulfilment of the M.B.A Program of Institute of Distance Learning of Dr. D.Y. Patil Vidyapeeth, Pune – 411018

The work is original, has not been copied from anywhere else, and has not been submitted to any other University / Institute for an award of any degree / diploma.

Date: - Signature:-

Place: Pune Name: Ruturaj Nitin Bhambure

ACKNOWLEDGEMENT

(TO BE GIVEN BY THE SYUDENT)

I would like to express my sincere thanks to Prof. Anand Irabatti, for his/her valuable guidance and support in completing my project. I am also thankful to my mentor Mr. Soham Mohite for guiding me throughout the Online MBA Program.

I would also like to express my gratitude towards our Chancellor Dr. P. D. Patil of Dr. D. Y. Patil Vidyapeeth, Centre for Online Learning for giving me this great opportunity to do a project on Data science using python (Data analysis and Data visualization of Spotify Dataset). Without their support and suggestions, this project would not have been completed.

Place: Pune

Date:

INDEX

EXECUTIVE SUMMARY

Spotify is a proprietary Swedish audio streaming and media services provider founded on 23 April 2006 by Daniel Ek and Martin Lorentzon. It is one of the largest music streaming service providers, with over 422 million monthly active users, including 182 million paying subscribers, as of March 2022. Spotify is listed (through a Luxembourg City-domiciled holding company, Spotify Technology S.A. on the New York Stock Exchange in the form of American depositary receipts.

Spotify offers digital copyright restricted recorded music and podcasts, including more than 82 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. Spotify is currently available in 180+ countries, as of October 2021. Users can search for music based on artist, album, or genre, and can create, edit, and share playlists.

Spotify is available in most of Europe, as well as the Americas and Oceania, with a total availability in 184 markets. The service is available on most devices including Windows, macOS, and Linux computers, iOS and Android smartphones and tablets, smart home devices such as the Amazon Echo and Google Nest lines of products and digital media players like Roku.

Unlike physical or download sales, which pay artists a fixed price per song or album sold, Spotify pays royalties based on the number of artist streams as a proportion of total songs streamed. It distributes approximately 70% of its total revenue to rights holders (often record labels), who then pay artists based on individual agreements. According to Ben Sisario of The New York Times, approximately 13,000 out of seven million artists on Spotify generated $50,000 or more in payments in 2020.

In March 2011, Spotify announced a customer base of 1 million paying subscribers across Europe, and by September 2011, the number of paying subscribers had doubled to two million. In August 2012, Time reported 15 million active users, four million being paying Spotify subscribers. User growth continued, reaching 20 million total active users, including five million paying customers globally and one million paying customers in the United States, in December 2012. By March 2013, the service had 24 million active users, six million being paying subscribers, which grew to 40 million users (including ten million paying) in May 2014, 60 million users (including 15 million paying) in December 2014, 75 million users (20 million paying) in June 2015, 30 million paying subscribers in March 2016, 40 million paying subscribers in September 2016, and 100 million total users in June 2016. In April 2020, Spotify reached 133 million premium users. In countries affected by the COVID-19 pandemic, Spotify registered a fall in users in late February, but it has seen a recovery.

When Spotify hired investor and former Lady Gaga manager Troy Carter as global head of creator services in 2016, what he found upon walking in the door was promising: New data-powered personalized playlists like Discover Weekly, Fresh Finds, and Release Radar were reeling in listeners by the millions. In less than a year, Discover Weekly spun 5 billion songs for over 40 million people—more listeners than Apple Music and Tidal combined. Together with Spotify’s hand-curated playlists, the algorithms are helping elevate the careers of thousands of artists. By mid-2016, more than 8,000 artists saw more than half of their listening come from Discover Weekly alone. Carter’s team is tasked with finding new ways to serve artists, be it with data or new tools for connecting with fans. Despite fresh competition from giants like Amazon and Apple, Spotify keeps growing: It now has over 100 million active listeners and over 40 million paying subscribers. “We aren’t worried about the competition,” Carter told Fast Company in 2016. “It makes us get better and better."

Much of Spotify's success is due to increasingly sophisticated data collection, which allows it to keep releasing new products that captivate its users around a particular mood or moment in time rather than offering the same tired genres.

Gearing up for its highly anticipated IPO, Spotify spent 2017 securing favourable royalty deals with all the major record labels. The company reported more than $3 billion in annual revenue last June, a 52% increase over the previous year. With more than 140 million active users (including more than 70 million premium subscribers), the music-streaming leader has scaled its fan base. Now it’s turning its focus to artists through new initiatives such as Rise, which helps break new acts by integrating them into playlists and marketing them on- and offline; Secret Genius, which shines a spotlight on songwriters through video series, podcasts, and playlists; and RapCaviar Live, a live concert series built around the popular hip-hop playlist. It’s becoming harder for labels alone to market artists, Carter says a year later. “You need other parties equally as invested as the label. We’re stepping up as one of those other parties.”

In this day and age, music streaming apps are preferred amongst most youngsters mainly because it's affordable, it's good quality and it's easy to use. Spotify is amongst the leading music streaming apps, with over 100 million active users and an estimated net value of 8 billion dollars. The membership involves the user paying $9.99 a month, for unlimited music streaming and many other features such as offline playlists. The service is available in 33 countries and is priced at about half of what a regular premium subscription costs. Spotify is available in most of Europe, most of the Americas, Australia, New Zealand, and parts of Asia. It is available for most modern devices, including Windows, macOS, and Linux computers, as well as iOS, Windows Phone and Android smartphones and tablets. Music can be browsed through or searched for by parameters such as artist, album, genre, playlist, or record label. Users can create, edit, and share playlists, share tracks on social media, and make playlists with other users. Spotify provides access to more than 30 million songs.

Many compare the idea of Spotify to iTunes. Spotify completes the job of delivering music in much the same way as iTunes does. Spotify is conveniently located, has a wonderful selection, is compatible with a computer, smartphone, and tablet. Spotify holds a systemic advantage over iTunes in one particular job characteristic of delivering music: relative pricing. While iTunes and Spotify both deliver music over the net, Spotify’s position as a radio service lets it price far below the level of iTunes. For $10 a month, users can gain access to unlimited music as long as they are listening through a Spotify music player.

In addition to its impressive growth and reach, the company’s adoption of new services illustrates the company’s innovative spirit. According to CNBC, the company has announced that its service is now available on Amazon’s Echo, a wireless speaker and voice-command device.

Although Spotify started off as any other mainstream music streaming service, it is definitely a company to watch out for.

CHAPTER 1: INTRODUCTION

INTRODUCTION

Python is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best language used by data scientist for various data science projects/application. Python provide great functionality to deal with mathematics, statistics and scientific function. It provides great libraries to deals with data science application.

One of the main reasons why Python is widely used in the scientific and research communities is because of its ease of use and simple syntax which makes it easy to adapt for people who do not have an engineering background. It is also more suited for quick prototyping.

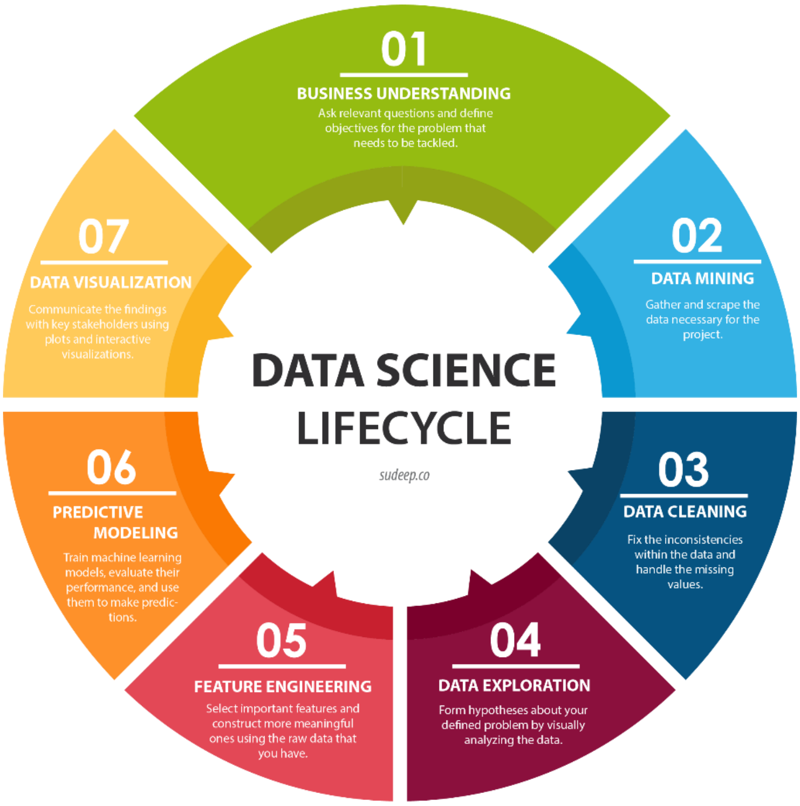
DEFINITION –

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

The data used for analysis can come from many different sources and presented in various formats.

Python for data science

THE DATA SCIENCE LIFECYCLE –


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

1. **Capture**: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.

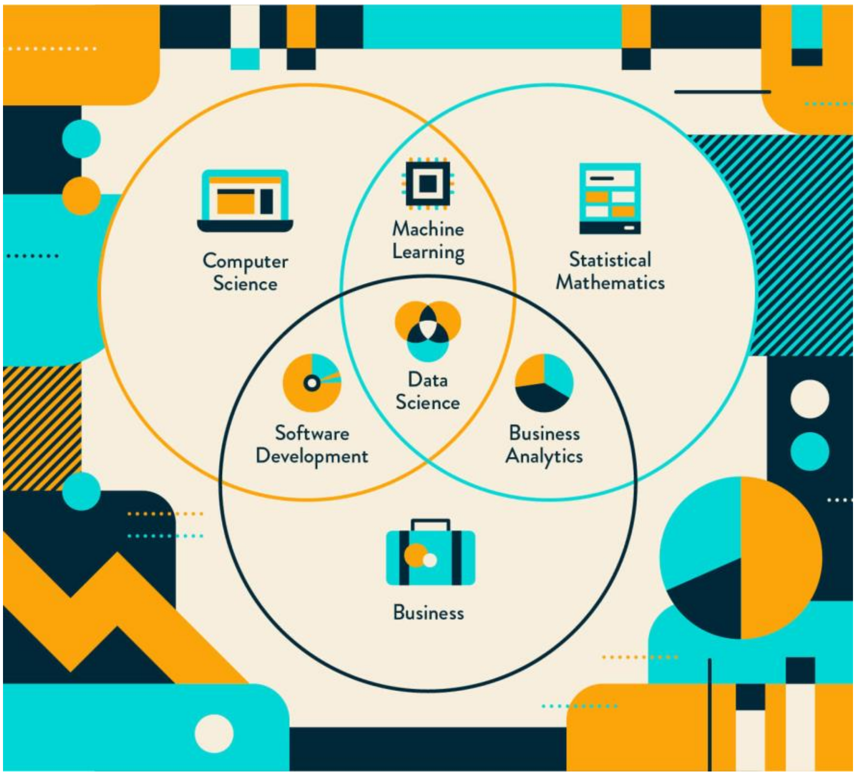
2. **Maintain**: Data Warehousing, Data Cleansing, Data Staging, Data Processing and Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

3. **Process**: Data Mining, Clustering/Classification, Data Modelling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.

4. **Analyse**: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, and Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.

5. **Communicate**: Data Reporting, Data Visualization, Business Intelligence and Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

PREREQUISITES FOR DATA SCIENCE



Here are some of the technical concepts you should know about before starting to learn what is data science.

1. **Machine Learning**

Machine learning is the backbone of data science. Data Scientists need to have a solid grasp of ML in addition to basic knowledge of statistics.

2. **Modelling**

Mathematical models enable you to make quick calculations and predictions based on what you already know about the data. Modelling is also a part of Machine Learning and involves identifying which algorithm is the most suitable to solve a given problem and how to train these models.

3. **Statistics**

Statistics are at the core of data science. A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results.

4. **Programming**

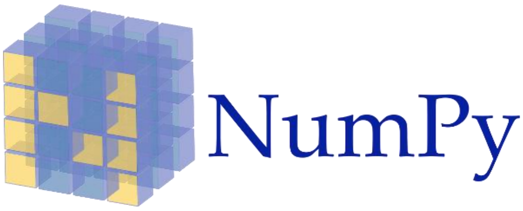
Some level of programming is required to execute a successful data science project. The most common programming languages are Python, and R. Python is especially popular because it’s easy to learn, and it supports multiple libraries for data science and ML.

5. **Databases**

A capable data scientist needs to understand how databases work, how to manage them, and how to extract data from them.

MOST COMMONLY USED LIBRARIES FOR DATA SCIENCE:

1. **Numpy**: Numpy is Python library that provides mathematical function to handle large dimension array. It provides various method/function for Array, Metrics, and linear algebra.



NumPy stands for Numerical Python. It provides lots of useful features for operations on n-arrays and matrices in Python. The library provides vectorization of mathematical operations on the NumPy array type, which enhance performance and speeds up the execution. It’s very easy to work with large multidimensional arrays and matrices using NumPy.

2. **Pandas**: Pandas is one of the most popular Python library for data manipulation and analysis. Pandas provide useful functions to manipulate large amount of structured data. Pandas provide easiest method to perform analysis.

It provide large data structures and manipulating numerical tables and time series data. Pandas is a perfect tool for data wrangling. Pandas is designed for quick and easy data manipulation, aggregation, and visualization. There two data structures in Pandas –

Series – It Handle and store data in one-dimensional data.

Data Frame – It Handle and store Two dimensional data.

3. **Matplotlib**: Matplotlib is another useful Python library for Data Visualization. Descriptive analysis and visualizing data is very important for any organization. Matplotlib provides various method to Visualize data in more effective way. Matplotlib allows to quickly make line graphs, pie charts, histograms, and other professional grade figures. Using Matplotlib, one can customize every aspect of a figure. Matplotlib has interactive features like zooming and planning and saving the Graph in graphics format.

4. **Scipy**: Scipy is another popular Python library for data science and scientific computing. Scipy provides great functionality to scientific mathematics and computing programming. SciPy contains sub-modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, Statmodel and other tasks common in science and engineering.

5. **Scikit** – learn: Sklearn is Python library for machine learning. Sklearn provides various algorithms and functions that are used in machine learning. Sklearn is built on NumPy, SciPy, and matplotlib. Sklearn provides easy and simple tools for data mining and data analysis. It provides a set of common machine learning algorithms to

users through a consistent interface. Scikit-Learn helps to quickly implement popular algorithms on datasets and solve real-world problems.

PURPOSE OF DATA SCIENCE –

The principal purpose of Data Science is to find patterns within data. It uses various statistical techniques to analyse and draw insights from the data. From data extraction, wrangling and pre-processing, a Data Scientist must scrutinize the data thoroughly.

Then, he has the responsibility of making predictions from the data. The goal of a Data Scientist is to derive conclusions from the data. Through these conclusions, he is able to assist companies in making smarter business decisions.

SCOPE OF DATA SCIENCE -

1. **Companies' Inability to handle data -**

Data is being regularly collected by businesses and companies for transactions and through website interactions. Many companies face a common challenge – to analyse and categorize the data that is collected and stored. A data scientist becomes the saviour in a situation of mayhem like this. Companies can progress a lot with proper and efficient handling of data, which results in productivity.

2. **Revised Data Privacy Regulations -**

Countries of the European Union witnessed the passing of the General Data Protection Regulation (GDPR) in May 2018. A similar regulation for data protection will be passed by California in 2020. This will create co-dependency between companies and data scientists for the need of storing data adequately and responsibly. In today's times, people are generally more cautious and alert about sharing data to businesses and giving up a certain amount of control to them, as there is rising awareness about data breaches and their malefic consequences. Companies can no longer afford to be careless and irresponsible about their data. The GDPR will ensure some amount of data privacy in the coming future.

3. **Data Science is constantly evolving -**

Career areas that do not carry any growth potential in them run the risk of stagnating. This indicates that the respective fields need to constantly evolve and undergo a change for opportunities to arise and flourish in the industry. Data science is a broad career path that is undergoing developments and thus promises abundant opportunities in the future. Data science job roles are likely to get more specific, which in turn will lead to specializations in the field. People inclined towards this stream can exploit their opportunities and pursue what suits them best through these specifications and specializations.

4. **An astonishing incline in data growth -**

Data is generated by everyone on a daily basis with and without our notice. The interaction we have with data daily will only keep increasing as time passes. In addition, the amount of data existing in the world will increase at lightning speed. As data production will be on the rise, the demand for data scientists will be crucial to help enterprises use and manage it well.

OBJECTIVES OF THE STUDY –

1. Collecting data

2. Processing data

3. Exploring and visualizing data

4. Analysing (data) and/or applying machine learning (to data)

DATA SCIENCE APPLICATIONS

1. **Fraud and risk detection**: Over the years, financial organizations have learned to analyse the probabilities of risks and defaults through customer profiling, past expenditures, and other variables available through data.

2. **Healthcare**: Data science makes it possible to manage and analyse very large diverse datasets in healthcare systems, drug development, medical image analysis, and more. Recently Data Science approaches were brought in to combat the COVID-19 pandemic. Data Scientists helped in digital contact tracing, diagnosis, risk assessment, resource allocation, estimating epidemiological parameters, drug development, social media analytics, etc.

3. **Internet search**: All search engines, including Google, use data science algorithms to deliver the best result for searched queries within seconds.

4. **Targeted advertising**: Digital ads have a higher call-through rate (CTR) than traditional ads because targeted advertising is based on a user's past behaviour with the help of data science algorithms.

5. **Recommendation systems**: Internet giants as well as other businesses have fervidly made use of recommendation engines to promote their products based on users' previous search results and their interests.

6. **Advanced image, speech, or character recognition**: Facial recognition algorithms on Facebook, speech recognition products, such as Siri, Cortana, Alexa, etc., and Google Lens are all perfect examples of data science applications in image, speech, and character recognition.

7. **Gaming**: Today, games use machine learning algorithms to improve or upgrade themselves as players move up to higher levels. In motion gaming, the opponent (computer) is able to analyse a player's previous moves and accordingly shape up its game. This is all possible because of data science.

8. **Augmented reality (AR)**: Augmented reality promises an exciting future through Data Science. A VR headset, for example, contains algorithms, data, and computing knowledge to offer the best viewing experience.

]

CHAPTER 2: LITERATURE REVIEW

LITERATURE REVIEW

Spotify is a music streaming service which offers alternative and legal ways for users to listen to music. Spotify allows for users to build personalized radio stations and playlists, through both free and paid subscriptions. This music streaming service offers users an opportunity to socialize through music by allowing them to share music on various social media sites such as Facebook and Twitter (Spotify, n.d.).

Spotify's target market mainly consists of traditional early adopters and opinion leaders which includes tech-savvy, music enthusiasts, musicians and record labels. 50 percent of users are ages 18-34, and 24% of Spotify's users are from households making roughly $100,000. Nearly 3 in 5 of Spotify's web visitors are males (Lipsman, 2011). Spotify's main competitors are Pandora and iTunes Radio. Pandora is considered the leader in internet radio with more than 76.2 active monthly users. Unlike Spotify, Pandora users do not have the option to select individual songs to listen to, instead Pandora selects song based on the user's music channel preferences (Couts, 2013). ITunes Radio is significantly less expensive than Spotify at $25 a year. Like Pandora, iTunes radio does not allow users to choose which song they will listen to, instead songs are randomized and based on the users.

Python is one of the most popular and widely used programming languages and has replaced many programming languages in the industry.

There are a lot of reasons why Python is popular among developers and one of them is that it has an amazingly large collection of libraries that users can work with

Here are a few important reasons as to why Python is popular:

1. Python has a huge collection of libraries.

2. Python is a beginner's level programming language because of it simplicity and easiness.

3. From developing to deploying and maintaining Python wants their developers to be more productive.

4. Portability is another reason for huge popularity of Python.

5. Python's programming syntax is simple to learn and is of high level when we compare it to C, Java, and C++.

1. **Numpy:**

Numpy is considered as one of the most popular machine learning library in Python.

TensorFlow and other libraries uses Numpy internally for performing multiple operations on Tensors. Array interface is the best and the most important feature of Numpy.

Features Of Numpy

Interactive: Numpy is very interactive and easy to use.

Mathematics: Makes complex mathematical implementations very simple.

Intuitive: Makes coding real easy and grasping the concepts is easy.

Lot of Interaction: Widely used, hence a lot of open source contribution.

Uses of Numpy?

This interface can be utilized for expressing images, sound waves, and other binary raw streams as an array of real numbers in N-dimensional.

For implementing this library for machine learning having knowledge of Numpy is important for full stack developers.

2. **Pandas:**

Pandas is a machine learning library in Python that provides data structures of high-level and a wide variety of tools for analysis. One of the great feature of this library is the ability to translate complex operations with data using one or two commands. Pandas have so many inbuilt methods for grouping, combining data, and filtering, as well as time-series functionality.

Features of Pandas

Pandas make sure that the entire process of manipulating data will be easier. Support for operations such as Re-indexing, Iteration, Sorting, Aggregations, Concatenations and Visualizations are among the feature highlights of Pandas.

Applications of Pandas?

Currently, there are fewer releases of panda’s library which includes hundreds of new features, bug fixes, enhancements, and changes in API. The improvements in pandas regards its ability to group and sort data, select best suited output for the apply method, and provides support for performing custom types operations.

Data Analysis among everything else takes the highlight when it comes to usage of Pandas. But, Pandas when used with other libraries and tools ensure high functionality and good amount of flexibility.

3. **Seaborn**:

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

4. **Matplotlib:**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

1. Create publication quality plots.

2. Make interactive figures that can zoom, pan, and update.

3. Customize visual style and layout.

4. Export too many file formats.

5. Embed in JupyterLab and Graphical User Interfaces.

6. Use a rich array of third-party packages built on Matplotlib.

5. **Keras:**

Keras is considered as one of the coolest machine learning libraries in Python. It provides an easier mechanism to express neural networks. Keras also provides some of the best utilities for compiling models, processing data-sets, visualization of graphs, and much more.

In the backend, Keras uses either Theano or TensorFlow internally. Some of the most popular neural networks like CNTK can also be used. Keras is comparatively slow when we compare it with other machine learning libraries. Because it creates a computational graph by using back-end infrastructure and then makes use of it to perform operations. All the models in Keras are portable.

Features Of Keras

It runs smoothly on both CPU and GPU.

Keras supports almost all the models of a neural network – fully connected, convolutional, pooling, recurrent, embedding, etc. Furthermore, these models can be combined to build more complex models.

Keras, being modular in nature, is incredibly expressive, flexible, and apt for innovative research.

Keras is a completely Python-based framework, which makes it easy to debug and explore.

6. **SciPy:**

SciPy is a machine learning library for application developers and engineers. However, you still need to know the difference between SciPy library and SciPy stack. SciPy library contains modules for optimization, linear algebra, integration, and statistics.

Features Of SciPy

The main feature of SciPy library is that it is developed using NumPy, and its array makes the most use of NumPy.

In addition, SciPy provides all the efficient numerical routines like optimization, numerical integration, and many others using its specific sub modules.

CHAPTER 3: RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

Research methodology simply refers to the practical “how” of any given piece of research. More specifically, it’s about how a researcher systematically designs a study to ensure valid and reliable results that address the research aims and objectives.

Qualitative, quantitative and mixed-methods are different types of methodologies, distinguished by whether they focus on words, numbers or both. This is a bit of an oversimplification, but it’s a good starting point for understandings. Let’s take a closer look.

Qualitative research refers to research which focuses on collecting and analysing words (written or spoken) and textual data, whereas quantitative research focuses on measurement and testing using numerical data. Qualitative analysis can also focus on other “softer” data points, such as body language or visual elements.

It’s quite common for a qualitative methodology to be used when the research aims and objectives are exploratory in nature. For example, a qualitative methodology might be used to understand peoples’ perceptions about an event that took place, or a candidate running for president.

Contrasted to this, a quantitative methodology is typically used when the research aims and objectives are confirmatory in nature. For example, a quantitative methodology might be used to measure the relationship between two variables (e.g. personality type and likelihood to commit a crime) or to test a set of hypotheses.

Data Science could be a space that incorporates working with colossal sums of information, creating calculations, working with machine learning and more to come up with trade insights. It incorporates working with the gigantic sum of information. Different processes are included to infer the information from the source like extraction of data, information preparation, model planning, model building and many more. The below image depicts the various processes of Data Science.

1. **Discovery -**

To begin with, it is exceptionally imperative to get the different determinations, prerequisites, needs and required budget-related with the venture. You must have the capacity to inquire the correct questions like do you have got the desired assets. These assets can be in terms of individuals, innovation, time and information. In this stage, you too got to outline the trade issue and define starting hypotheses (IH) to test.

2. Information Preparation -

In this stage, you would like to investigate, pre-process and condition data for modelling. You'll be able to perform information cleaning, changing, and visualization. This will assist you to spot the exceptions and build up a relationship between the factors. Once you have got cleaned and arranged the information, it's time to do exploratory analytics on it.

3. Model Planning -

Here, you may decide the strategies and methods to draw the connections between factors. These connections will set the base for the calculations which you may execute within the following stage. You may apply Exploratory Data Analytics (EDA) utilizing different factual equations and visualization apparatuses.

4. Model Building -

In this stage, you'll create datasets for training and testing purposes. You may analyze different learning procedures like classification, association, and clustering and at last, actualize the most excellent fit technique to construct the show.

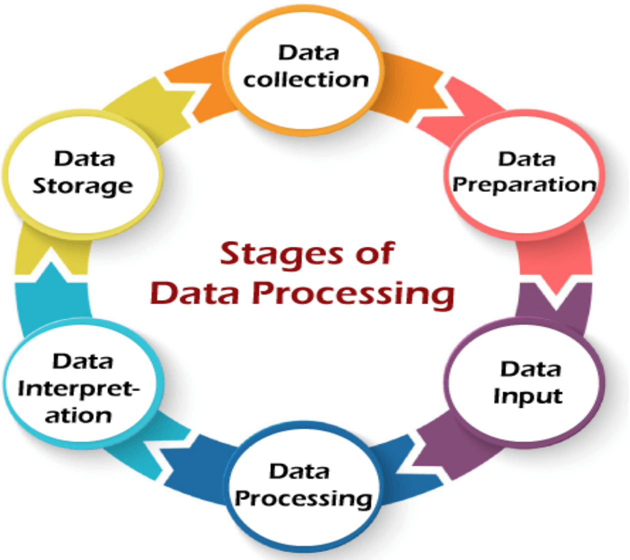
5. Operationalize -

In this stage, you convey the last briefings, code, and specialized reports. In expansion, now a pilot venture is additionally actualized in a real-time generation environment. This will give you a clear picture of the execution and other related limitations.

6. Communicate Results -

Presently, it is critical to assess the outcome of the objective. So, within the final stage, you recognize all the key discoveries, communicate to the partners and decide in the event that the outcomes about the venture are a victory or a disappointment based on the criteria created in Stage 1.

DATA COLLECTION IN DATA SCIENCE –



Data collection is the process of accumulating data that's required to solve a problem statement.

All data science projects (all projects really) start with a problem that needs a solution. There's always something you can solve or improve.

Identify a problem statement

The most vital step is to identify and pinpoint the exact question that needs to be answered.

For example, let's say your online cat food business is not producing enough sales. Your problem statement would be: find ways to attract more customers and improve your sales.

You can work backward once you briefly identify your problem and solution. In this case, you can start off by taking a look at the audience you are targeting.

Maybe you need to target a wider age group, or you may want to learn more about what type of cat owners shop online, such as their geographic location, gender, ethnicity, and so on.

Collecting more data is often about collecting the right type of data. Thus, the first step is to understand what problem needs solving and how you can go about solving it.

Determine what type of data is needed

The next step is to consider what type of data you must collect.

Is it quantitative or qualitative?

Accessing and processing quantitative data is easier because it involves raw numbers and digits. On the other hand, processing qualitative data, such as customer reviews or feedback, is more complex.

Segregating the different types of data from the moment of data collection can be useful while performing data processing down the line.

Decide on your data sources

Once you have an idea about what data you need, start looking into whether the data is within your organization or if you'll require third-party or external data.

In most cases, the smart thing to do is to acquire external data. This acquisition will keep you on par with your competitors, who will probably also invest in third-party data. You must be willing to buy data and keep your legal team close.

At this point, it's important to draw your attention to the ethical issues relevant to data collection and data privacy.

Make sure your audience is fully aware of the data you're collecting about them. You don't want to fall into a data scandal, such as the one in which Facebook and Cambridge Analytical were involved. If your organization is buying data from another corporation, your legal team must be careful to consider all data privacy clauses.

In addition to that, collecting data from government organizations is also common. Some data scientists also use surveys to collect data.

Another practice is to build a user persona based on existing data. For instance, your organization has insights into the type of people who buy sports gear. Such information can get used to create a user persona for people with varied interests. This process is common when there is not enough data available.

Create a timeline

Now it's time to identify the time frame within which the data is most useful.

For example, do you need end-to-end data about how a customer lands on an e-commerce website? Or do you need relevant parts about the user's search history, geography, and background?

Identifying the timeline is key to getting the exact type of data you need to solve your problem statement.

A potential lead may generate data at different stages, and it's your job to effectively evaluate which data is most relevant.

Collect your data

To effectively collect data, devise a plan that addresses all the questions relevant to securely collecting data.

If you're collecting data from a third party or a stakeholder, make sure all requirements and privacy issues get considered.

Additionally, create a plan for how you will store the data. Make sure your organization has the right tools and infrastructure to manage and process the data.

You also need to establish a systematic approach for storing all the different types of data so that you can later combine and further process them.

For example, storing transactional data can be relatively easier since there are tons of tools that arrange such data in a tabular format. On the other hand, unstructured data can be relatively difficult to manage and store due to its loose format.

Therefore, you must devise a plan to collect your data and make the processing simpler.

Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Below are the steps involved in data cleaning –

1. Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyse. For example, if you want to analyse data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performing dataset.

2. **Fix structural errors**

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabelled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analysed as the same category.

3. **Filter unwanted outliers**

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analysing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn't mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

4. **Handle missing data**

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.

As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.

As a third option, you might alter the way the data is used to effectively navigate null values.

5. **Validate and QA**

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

CHAPTER 4: DATA ANALYSIS

DATA ANALYSIS

Visualization
Statistical Modeling
Statistical Computing
Data Technology
Data Research
Data Consulting
Real-world Applications
Scientific Methods
Data Science

Data analysis is the practice of working with data to glean useful information, which can then be used to make informed decisions.

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts," Sherlock Holme's proclaims in Sir Arthur Conan Doyle's A Scandal in Bohemia.

This idea lies at the root of data analysis. When we can extract meaning from data, it empowers us to make better decisions. And we're living in a time when we have more data than ever at our fingertips.

Companies are widening up to the benefits of leveraging data. Data analysis can help a bank to personalize customer interactions, a health care system to predict future health needs, or an entertainment company to create the next big streaming hit.

The World Economic Forum Future of Jobs Report 2020 listed data analysts and scientists as the top emerging job, followed immediately by AI and machine learning specialists, and big data specialists.

Data analysis process

As the data available to companies continues to grow both in amount and complexity, so too does the need for an effective and efficient process by which to harness the value of that data. The analysis method typically moves through several iterative phases. Let's take a closer look at each.

1. Identify the business question you'd like to answer. What problem is the company trying to solve? What do you need to measure, and how will you measure it?

2. Collect the raw data sets you'll need to help you answer the identified question. Data collection might come from internal sources, like a company's client relationship management (CRM) software, or from secondary sources, like government records or social media application programming interfaces (APIs).

3. Clean the data to prepare it for analysis. This often involves purging duplicate and anomalous data, reconciling inconsistencies, standardizing data structure and format, and dealing with white spaces and other syntax errors.

4. Analyse the data. By manipulating the data using various data analysis tools and techniques, you can begin to find trends, correlations, outliers, and variations that begin to tell a story. During this stage, you might use data mining to discover patterns within databases or data visualization software to help transform data into an easy-to-understand graphical format.

5. Interpret the results of your analysis to see how well the data answered your original question. What recommendations can you make based on the data? What are the limitations to your conclusions?

Data Visualization -

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modelled, it must be visualized for conclusions to be made. Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data visualization is important for almost every career. It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders. It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.

Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning (ML) algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

Data analytics

Companies around the globe generate vast volumes of data daily, in the form of log files, web servers, transactional data, and various customer-related data. In addition to this, social media websites also generate enormous amounts of data.

Companies ideally need to use all of their generated data to derive value out of it and make impactful business decisions. Data analytics is used to drive this purpose.

Data analytics is the process of exploring and analysing large datasets to find hidden patterns, unseen trends, discover correlations, and derive valuable insights to make business predictions. It improves the speed and efficiency of your business.

Businesses use many modern tools and technologies to perform data analytics. This is data analytics for beginners, in a nutshell.

Python for data science

System configuration –

Laptop brand - Lenovo

Windows edition – Windows 8

Processor – Intel® Pentium® CPU 2020M @2.40GHz

Hardware requirements -

Installed RAM – 6GB

Storage – 450 GB

System type – 64-bit Operating system, x64 based processor

Software requirements –

Python 3.10 and Jupyter notebook application which will be used for data processing, data operations, data analysis and data visualization.

Python for data science

Spotify dataset –

The dataset contains 20 columns:

1. id
2. name
3. popularity
4. duration_ms
5. explicit
6. artists
7. id_artists
8. release_date
9. danceability
10. energy
11. key
12. loudness
13. mode
14. speechiness
15. acousticness
16. instrumentalness
17. liveness
18. valence
19. tempo
20. time_signature

Data analysis and Data visualization

The Spotify dataset has been collected from keggle.com and is then used for analysis.

Now the required important libraries are imported.

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Now the dataset will be read through the jupyter notebook and will use the code to display the first five rows of the dataset. The dataset is in the CSV format.

In [2]:

```
df_tracks = pd.read_csv("C:\\Users\\Mahesh\\Downloads\\tracks.csv")
df_tracks.head()
```

Out[2]:

	id	name	popularity	duration_ms	explicit	artists	id_artists	release_date
0	35iwgR4jXetI318WEWsa1Q	Carve	6	126903	0	['Uli']	['45tIt06XoI0Iio4LBEVpls']	1922-02-22
1	021ht4sdgPcrDgSk7JTbKY	Capítulo 2.16 - Banquero Anarquista	0	98200	0	['Fernando Pessoa']	['14jtPCOoNZwquk5wd9DxrY']	1922-06-01
2	07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	181640	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	1922-03-21
3	08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	176907	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	1922-03-21

4	08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	163080	0	['Dick Haymes']	['3BiJGZsyX9sJchTqcSA7Su']	1922

To be continued…

danceability	**energy**	**key**	**loudness**	**mode**	**speechiness**	**acousticness**	**instrumentalness**	**liveness**	**valence**	**tempo**	**time_signature**
0.645	0.4450	0	-13.338	1	0.4510	0.674	0.7440	0.151	0.127	104.851	3
0.695	0.2630	0	-22.136	1	0.9570	0.797	0.0000	0.148	0.655	102.009	1
0.434	0.1770	1	-21.180	1	0.0512	0.994	0.0218	0.212	0.457	130.418	5
0.321	0.0946	7	-27.961	1	0.0504	0.995	0.9180	0.104	0.397	169.980	3
0.402	0.1580	3	-16.900	0	0.0390	0.989	0.1300	0.311	0.196	103.220	4

Now we will check the info about the number of rows and columns and null values in the dataset.

In [3]:

```
df_tracks.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 586672 entries, 0 to 586671
Data columns (total 20 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   id                586672 non-null  object
 1   name              586601 non-null  object
 2   popularity        586672 non-null  int64
 3   duration_ms       586672 non-null  int64
 4   explicit          586672 non-null  int64
 5   artists           586672 non-null  object
 6   id_artists        586672 non-null  object
 7   release_date      586672 non-null  object
 8   danceability      586672 non-null  float64
 9   energy            586672 non-null  float64
 10 key                586672 non-null  int64
 11 loudness           586672 non-null  float64
 12 mode               586672 non-null  int64
 13  speechiness       586672 non-null  float64
 14  acousticness      586672 non-null  float64
 15  instrumentalness  586672 non-null  float64
 16  liveness          586672 non-null  float64
 17 valence            586672 non-null  float64
 18 tempo              586672 non-null  float64
 19  time signature    586672 non-null  int64
dtypes: float64(9), int64(6), object(5)
memory usage: 89.5+ MB
```

In [4]:

```
sorted_df = df_tracks.sort_values("popularity", ascending = True).head(10)
sorted_df
```

Out[4]:

	id	name	popularity	duration_ms	explicit	artists	id_artists	release_date
546130	181rTRhCcggZPwP2TUcVqm	Newspaper Reports On Abner, 20 February 1935	0	896575	0	['Norris Goff', 'Chester Lauck', 'Carlton Bric...	['3WCwCPDMpGzrt0Qz6quumy', '7vk8UqABg0Sga78GI3...	1935-02-20
546222	0yOCz3V5KMm8l1T8EFc60i	恋は水の上で	0	188440	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949
546221	0y48Hhwe52099UqYjegRCO	私の誕生日	0	173467	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949
546220	0xCmgtf9ka07hkZg3D6PaV	エル・チョクロ (EL CHOCLO)	0	205280	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949
546219	0tBXS3VuCPX7KWUFH2nros	恋は不思議なもの	0	185733	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949

546218	0qrKnQtYDVJhKFAXTHYVS9	ゆうべはどうしたの(WHATS A MALLAU)	0	183427	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949
546217	0nqsDxOeKSwEzp3AUQAAqS	Screen Director's Playhouse, Music For Million...	0	1767071	0	['Wilms Herbert', 'June Allyson', 'Joseph Kear...	['2rbm8QWvmnVwxFo84EVM1h', '4yW5adMgyIfHFzaL9i...	1949-04-10
546216	0kGEdsxVLYjCdfxM9tbezd	ブルーマンボ	0	162147	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949
546215	0bc3PUZurUUXrY7yqoOxjq	Screen Director's Playhouse, Trade Winds direc...	0	1776652	0	['Wally Maher', 'Tay Garnett', 'Lurene Tuttle'...	['7hkhJTTI3VnUGVWUt8SJXT', '3kYeeIpRCgJz4fQYDv...	1949-05-19
546214	0Wwm0ruSjYMIiWG0nyAI1F	Screen Director's	0	1767576	0	['Joseph Gra	['6GK59BC4LJzqR0OpHAX2S3',	1949-05-08

		Playhouse, It's A Wonderful ...				nby', 'Jimmy Stewart', 'Irene Tedr...	'58BzBaExrnrx898sby...	

To be continued…

danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
0.595	0.262	8	-17.746	1	0.9320	0.993	0.007510	0.0991	0.320	79.849	4
0.418	0.388	0	-8.580	1	0.0358	0.925	0.000014	0.1050	0.439	94.549	4
0.642	0.178	5	-11.700	1	0.0501	0.993	0.000943	0.0928	0.715	119.013	4
0.695	0.467	0	-12.236	0	0.0422	0.827	0.000000	0.0861	0.756	125.941	4
0.389	0.388	2	-8.221	1	0.0351	0.869	0.000000	0.0924	0.372	72.800	4
0.631	0.249	5	-11.883	1	0.0355	0.951	0.000000	0.0814	0.517	131.097	4
0.533	0.317	7	-13.047	1	0.9180	0.682	0.000000	0.3330	0.336	76.836	4
0.529	0.546	0	-6.462	0	0.0418	0.784	0.000000	0.3750	0.903	128.604	4
0.599	0.321	0	-15.428	0	0.9330	0.808	0.000000	0.5570	0.379	93.025	4
0.645	0.341	8	-12.177	1	0.8670	0.690	0.000000	0.1530	0.431	117.591	4

In [5]:

df_tracks.describe().transpose()

Out[5]:

	count	**mean**	**std**	**min**
popularity	586672.0	27.570053	18.370642	0.0
duration_ms	586672.0	230051.167286	126526.087418	3344.0
explicit	586672.0	0.044086	0.205286	0.0
danceability	586672.0	0.563594	0.166103	0.0
energy	586672.0	0.542036	0.251923	0.0
key	586672.0	5.221603	3.519423	0.0
loudness	586672.0	-10.206067	5.089328	-60.0
mode	586672.0	0.658797	0.474114	0.0
speechiness	586672.0	0.104864	0.179893	0.0
acousticness	586672.0	0.449863	0.348837	0.0
instrumentalness	586672.0	0.113451	0.266868	0.0
liveness	586672.0	0.213935	0.184326	0.0
valence	586672.0	0.552292	0.257671	0.0
tempo	586672.0	118.464857	29.764108	0.0
time_signature	586672.0	3.873382	0.473162	0.0

To be continued…

25%	**50%**	**75%**	**max**
13.0000	27.000000	41.00000	100.000
175093.0000	214893.000000	263867.00000	5621218.000
0.0000	0.000000	0.00000	1.000
0.4530	0.577000	0.68600	0.991
0.3430	0.549000	0.74800	1.000
2.0000	5.000000	8.00000	11.000
-12.8910	-9.243000	-6.48200	5.376
0.0000	1.000000	1.00000	1.000
0.0340	0.044300	0.07630	0.971

0.0969	0.422000	0.78500	0.996
0.0000	0.000024	0.00955	1.000
0.0983	0.139000	0.27800	1.000
0.3460	0.564000	0.76900	1.000
95.6000	117.384000	136.32100	246.381
4.0000	4.000000	4.00000	5.000

In [6]:

```
most_popular = df_tracks.query("popularity>90", inplace = False).sort_values("popularity", ascending = False)
most_popular[:10]
```

Out[6]:

	id	**name**	**popularity**	**duration_ms**	**explicit**	**artists**	**id_artists**	**release_date**
93802	4iJyoBOLtHqaGxP12qzhQI	Peaches (feat. Daniel Caesar & Giveon)	100	198082	1	['Justin Bieber', 'Daniel Caesar', 'Giveon']	['1uNFoZAHBGtllmzznpCI3s', '20wkVLutqVOYrc0kxF...	2021-03-19
93803	7lPN2DXiMsVn7XUKtOW1CS	drivers license	99	242014	1	['Olivia Rodrigo']	['1McMsnEElThX1knmY4oliG']	2021-01-08
93804	3Ofmpyhv5UAQ70mENzB277	Astronaut In The Ocean	98	132780	0	['Masked Wolf']	['1uU7g3DNSbsu0QjSEqZtEd']	2021-01-06

92810	5QO79kh1waicV47BqGRL3g	Save Your Tears	97	215627	1	['The Weeknd']	['1Xyo4u8uXC1ZmMpatF05PJ']	2020-03-20
92811	6tDDoYIxWvMLTdKpjFkc1B	telepatía	97	160191	0	['Kali Uchis']	['1U1el3k54VvEUzo3ybLPlM']	2020-12-04
92813	0VjIjW4GlUZAMYd2vXMi3b	Blinding Lights	96	200040	0	['The Weeknd']	['1Xyo4u8uXC1ZmMpatF05PJ']	2020-03-20
93805	7MAibcTli4IisCtbHKrGMh	Leave The Door Open	96	242096	0	['Bruno Mars', 'Anderson .Paak', 'Silk Sonic']	['0du5cEVh5yTK9QJze8zA0C', '3jK9MiCrA42lLAdMGU...	2021-03-05
92814	6f3Slt0GbA2bPZlz0aIFXN	The Business	95	164000	0	['Tiësto']	['2o5jDhtHVPhrJdv3cEQ99Z']	2020-09-16
91866	60ynsPSSKe6O3sfwRnIBRf	Streets	94	226987	1	['Doja Cat']	['5cj0lLjcoR7YOSnhnX0Po5']	2019-11-07
92816	3FAJ6O0NOHQV8Mc5Ri6ENp	Heartbreak Anniversary	94	198371	0	['Giveon']	['4fxd5Ee7UefO4CUXgwJ7IP']	2020-03-27

To be continued…

danceability	**energy**	**key**	**loudness**	**mode**	**speechiness**	**acousticness**	**instrumentalness**	**liveness**	**valence**	**tempo**	**time_signature**
0.677	0.696	0	-6.181	1	0.1190	0.32100	0.000000	0.4200	0.464	90.030	4

0.585	0.436	10	-8.761	1	0.0601	0.72100	0.000013	0.1050	0.132	143.874	4
0.778	0.695	4	-6.865	0	0.0913	0.17500	0.000000	0.1500	0.472	149.996	4
0.680	0.826	0	-5.487	1	0.0309	0.02120	0.000012	0.5430	0.644	118.051	4
0.653	0.524	11	-9.016	0	0.0502	0.11200	0.000000	0.2030	0.553	83.970	4
0.514	0.730	1	-5.934	1	0.0598	0.00146	0.000095	0.0897	0.334	171.005	4
0.586	0.616	5	-7.964	1	0.0324	0.18200	0.000000	0.0927	0.719	148.088	4
0.798	0.620	8	-7.079	0	0.2320	0.41400	0.019200	0.1120	0.235	120.031	4
0.749	0.463	11	-8.433	1	0.0828	0.20800	0.037100	0.3370	0.190	90.028	4
0.449	0.465	0	-8.964	1	0.0791	0.52400	0.000001	0.3030	0.543	89.087	3

In [7]:

```
df_tracks.set_index("release_date", inplace=True)
df_tracks.index=pd.to_datetime(df_tracks.index)
df_tracks.head()
```

Out[7]:

	id	**name**	**popularity**	**duration_ms**	**explicit**	**artists**	**id_artists**	**danceability**
release_date								
1922-02-22	35iwgR4jXetI318WEWsa1Q	Carve	6	126903	0	['Uli']	['45tIt06XoI0Iio4LBEVpls']	0.645

1922-06-01	021ht4sdgPcrDgSk7JTbKY	Capítulo 2.16 - Banquero Anarquista	0	98200	0	['Fernando Pessoa']	['14jtPCOoNZwquk5wd9DxrY']	0.695
1922-03-21	07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	181640	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.434
1922-03-21	08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	176907	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.321
1922-01-01	08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	163080	0	['Dick Haymes']	['3BiJGZsyX9sJchTqcSA7Su']	0.402

To be continued…

energy	**key**	**loudness**	**mode**	**speechiness**	**acousticness**	**instrumentalness**	**liveness**	**valence**	**tempo**	**time_signature**
0.4450	0	-13.338	1	0.4510	0.674	0.7440	0.151	0.127	104.851	3
0.2630	0	-22.136	1	0.9570	0.797	0.0000	0.148	0.655	102.009	1
0.1770	1	-21.180	1	0.0512	0.994	0.0218	0.212	0.457	130.418	5
0.0946	7	-27.961	1	0.0504	0.995	0.9180	0.104	0.397	169.980	3
0.1580	3	-16.900	0	0.0390	0.989	0.1300	0.311	0.196	103.220	4

In [8]:

```
df_tracks["duration"]= df_tracks["duration_ms"].apply(lambda x: round(x/1000))
df_tracks.drop("duration_ms", inplace=True, axis=1)
df_tracks.duration.head()
```

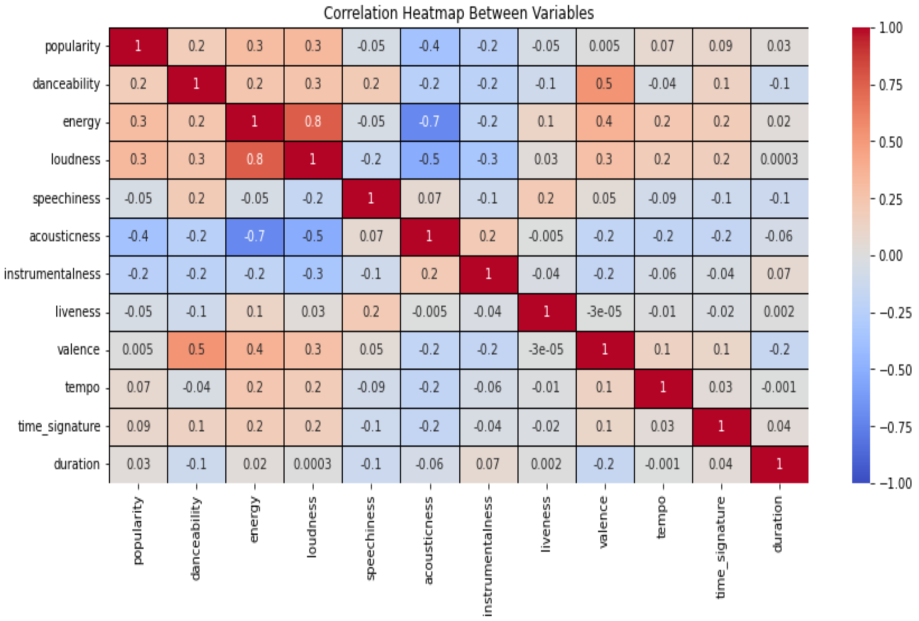
Out[8]:

```
release_date
1922-02-22    127
1922-06-01     98
1922-03-21    182
1922-03-21    177
1922-01-01    163
Name: duration, dtype: int64
```

In [9]:

```
corr_df=df_tracks.drop(["key", "mode", "explicit"], axis=1).corr(method="pearson")
plt.figure(figsize=(14,6))
heatmap=sns.heatmap(corr_df,annot=True, fmt=".1g", vmin=-1, vmax=1, center=0,
cmap="coolwarm", linewidths=1, linecolor="Black")
heatmap.set_title("Correlation Heatmap Between Variables")
heatmap.set_xticklabels(heatmap.get_xticklabels(), rotation=90)
```

```
[Text(0.5, 0, 'popularity'),
 Text(1.5, 0, 'danceability'),
 Text(2.5, 0, 'energy'),
 Text(3.5, 0, 'loudness'),
 Text(4.5, 0, 'speechiness'),
 Text(5.5, 0, 'acousticness'),
 Text(6.5, 0, 'instrumentalness'),
 Text(7.5, 0, 'liveness'),
 Text(8.5, 0, 'valence'),
 Text(9.5, 0, 'tempo'),
 Text(10.5, 0, 'time_signature'),
 Text(11.5, 0, 'duration')]
```



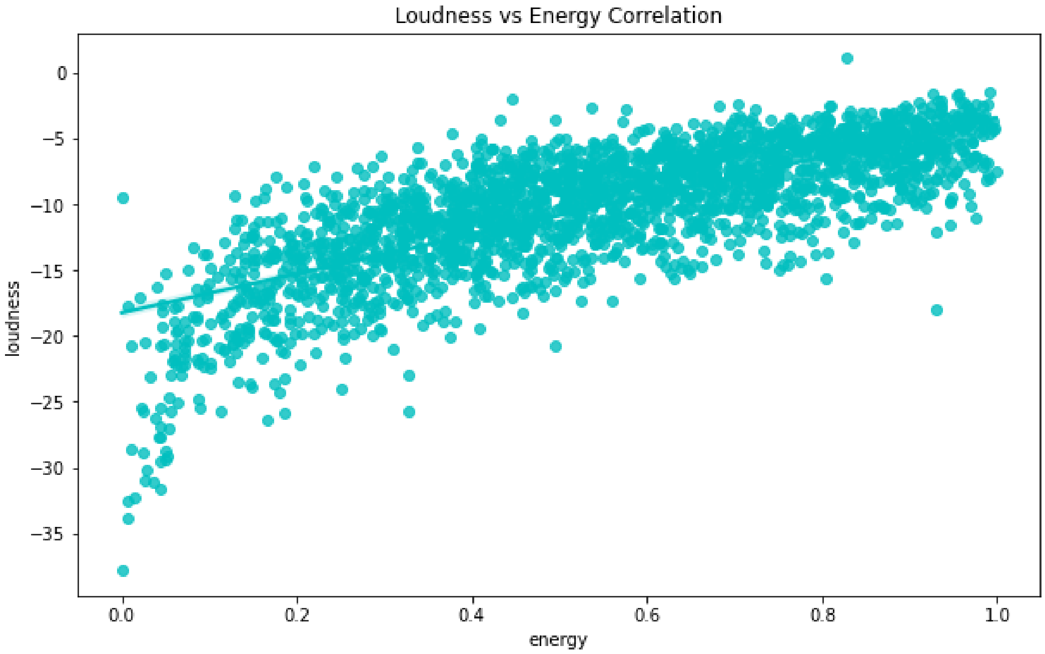
	popularity	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	duration
popularity	1	0.2	0.3	0.3	-0.05	-0.4	-0.2	-0.05	0.005	0.07	0.09	0.03
danceability	0.2	1	0.2	0.3	0.2	-0.2	-0.2	-0.1	0.5	-0.04	0.1	-0.1
energy	0.3	0.2	1	0.8	-0.05	-0.7	-0.2	0.1	0.4	0.2	0.2	0.02
loudness	0.3	0.3	0.8	1	-0.2	-0.5	-0.3	0.03	0.3	0.2	0.2	0.0003
speechiness	-0.05	0.2	-0.05	-0.2	1	0.07	-0.1	0.2	0.05	-0.09	-0.1	-0.1
acousticness	-0.4	-0.2	-0.7	-0.5	0.07	1	0.2	-0.005	-0.2	-0.2	-0.2	-0.06
instrumentalness	-0.2	-0.2	-0.2	-0.3	-0.1	0.2	1	-0.04	-0.2	-0.06	-0.04	0.07
liveness	-0.05	-0.1	0.1	0.03	0.2	-0.005	-0.04	1	-3e-05	-0.01	-0.02	0.002
valence	0.005	0.5	0.4	0.3	0.05	-0.2	-0.2	-3e-05	1	0.1	0.1	-0.2
tempo	0.07	-0.04	0.2	0.2	-0.09	-0.2	-0.06	-0.01	0.1	1	0.03	-0.001
time_signature	0.09	0.1	0.2	0.2	-0.1	-0.2	-0.04	-0.02	0.1	0.03	1	0.04
duration	0.03	-0.1	0.02	0.0003	-0.1	-0.06	0.07	0.002	-0.2	-0.001	0.04	1

In [10]:

```
sample_df = df_tracks.sample(int(0.004*len(df_tracks)))
print(len(sample_df))
2346
```

In [11]:

```
plt.figure(figsize=(10,6))
sns.regplot(data = sample_df, y = "loudness", x = "energy", color = "c").set(title = "Loudness vs Energy Correlation")
[Text(0.5, 1.0, 'Loudness vs Energy Correlation')]
```

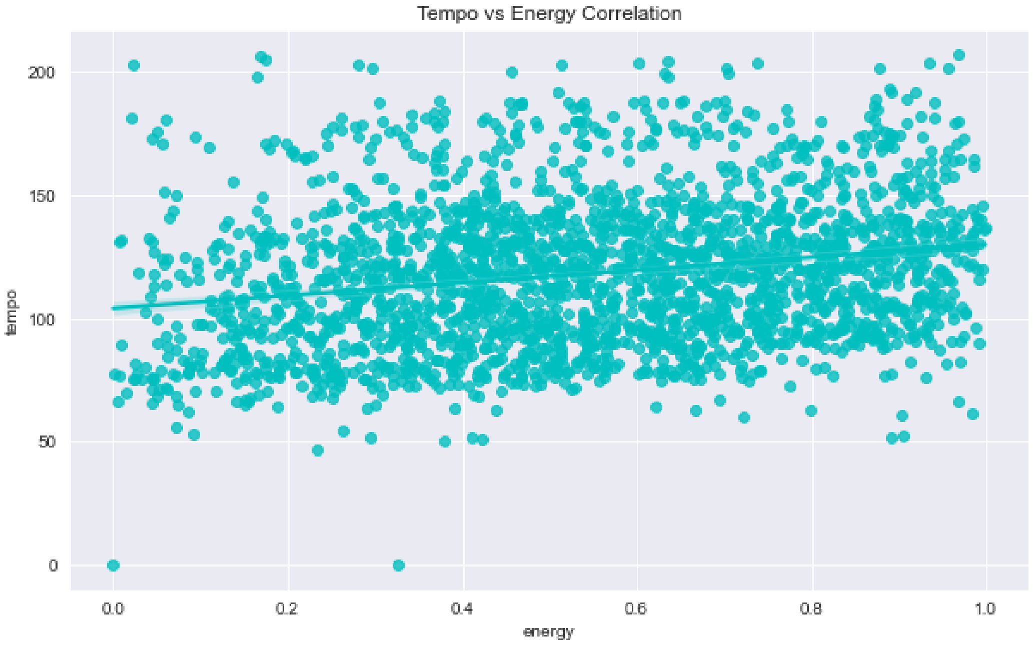


In [19]:

```
plt.figure(figsize=(10,6))
sns.regplot(data = sample_df, y = "tempo", x = "energy", color = "c").set(title =
"Tempo vs Energy Correlation")
```

Out[19]:

```
[Text(0.5, 1.0, 'Tempo vs Energy Correlation')]
```

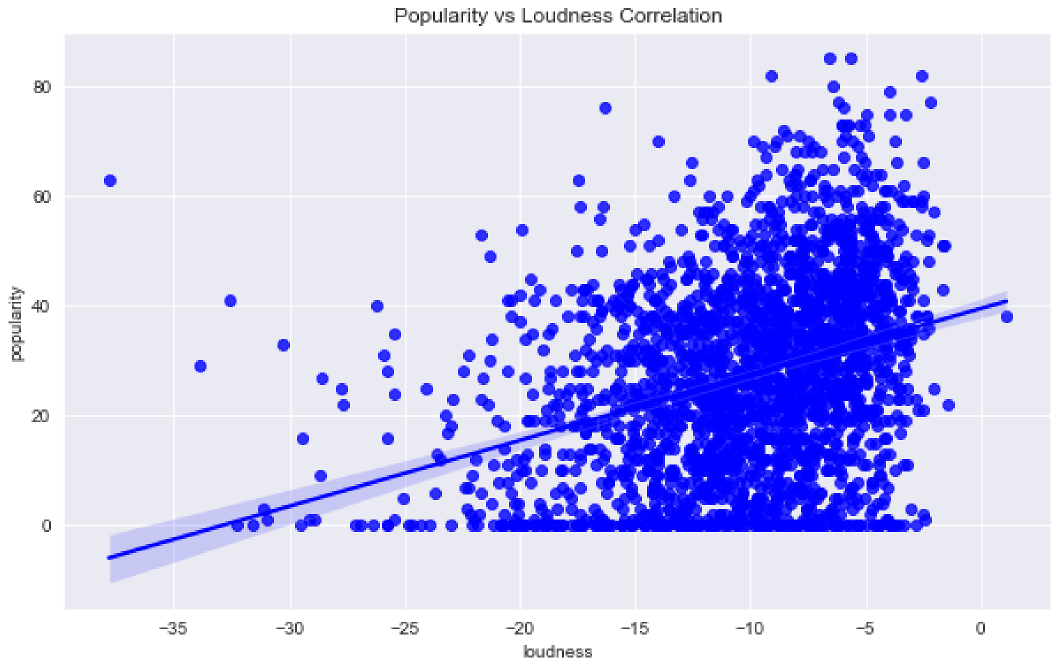


In [20]:

```
plt.figure(figsize=(10,6))
sns.regplot(data = sample_df, y = "popularity", x = "loudness", color = "b").set(title =
"Popularity vs Loudness Correlation")
```

Out[20]:

```
[Text(0.5, 1.0, 'Popularity vs Loudness Correlation')]
```

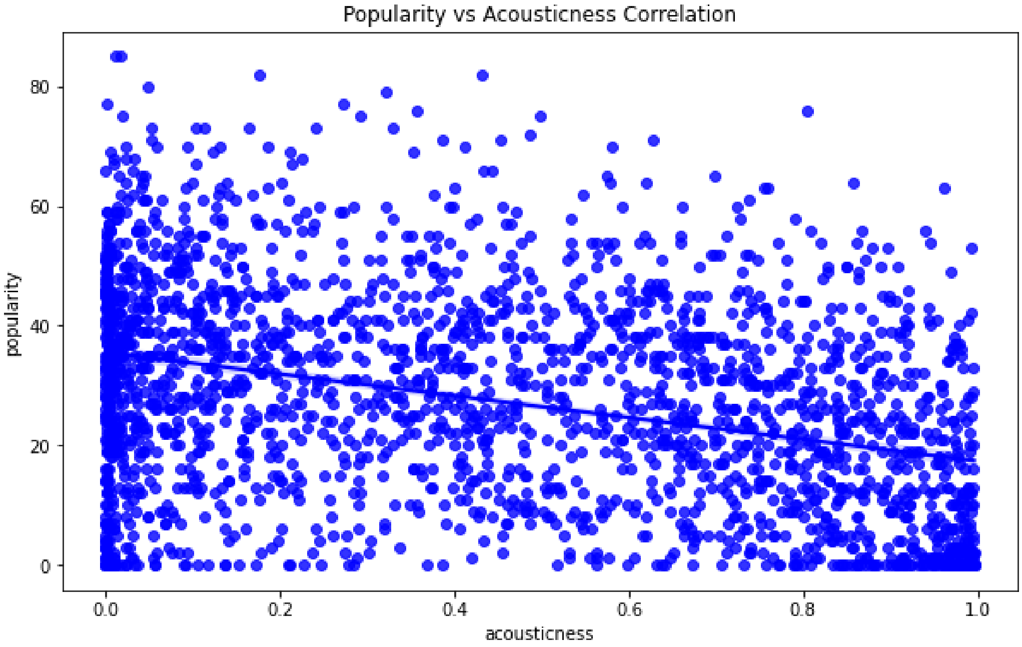


In [12]:

```
plt.figure(figsize=(10,6))
sns.regplot(data = sample_df, y = "popularity", x = "acousticness", color =
"b").set(title = "Popularity vs Acousticness Correlation")
```

Out[12]:

```
[Text(0.5, 1.0, 'Popularity vs Acousticness Correlation')]
```

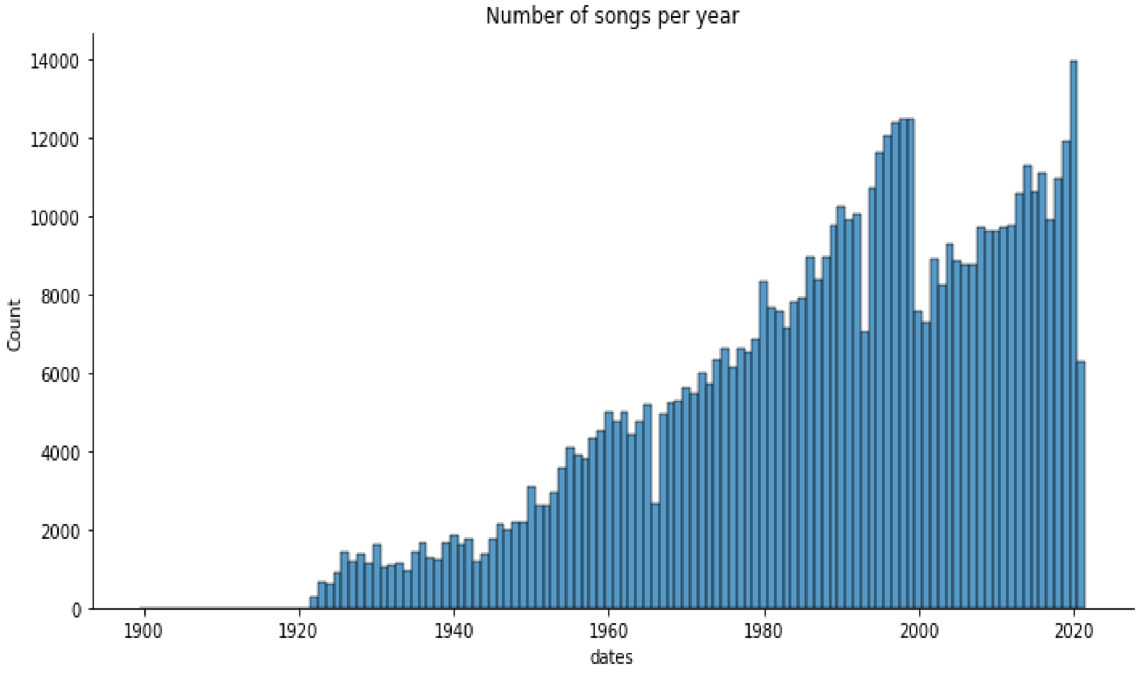


In [13]:

```
df_tracks['dates']=df_tracks.index.get_level_values('release_date')
df_tracks.dates=pd.to_datetime(df_tracks.dates)
years=df_tracks.dates.dt.year
sns.displot(years,discrete=True,aspect=2,height=5,kind="hist").set(title="Number of
songs per year")
```

Out[13]:

```
<seaborn.axisgrid.FacetGrid at 0xe7ebebee80>
```

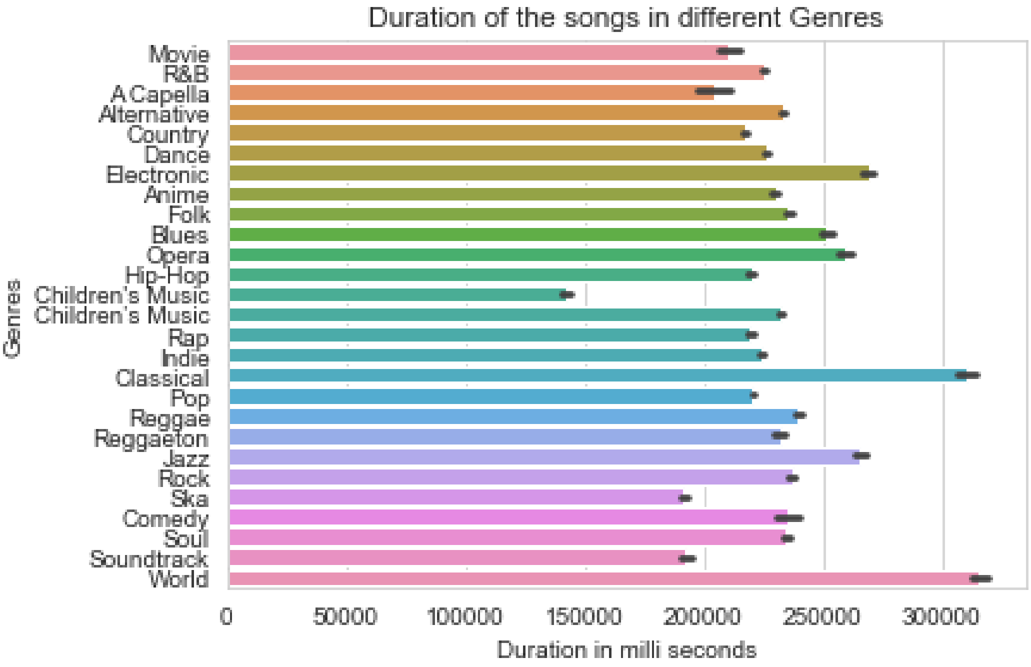


In [17]:

```
plt.title("Duration of the songs in different Genres")
sns.color_palette("rocket", as_cmap= True)
sns.barplot(y='genre', x='duration_ms', data=df_genre)
plt.xlabel("Duration in milli seconds")
plt.ylabel("Genres")
```

Out[17]:

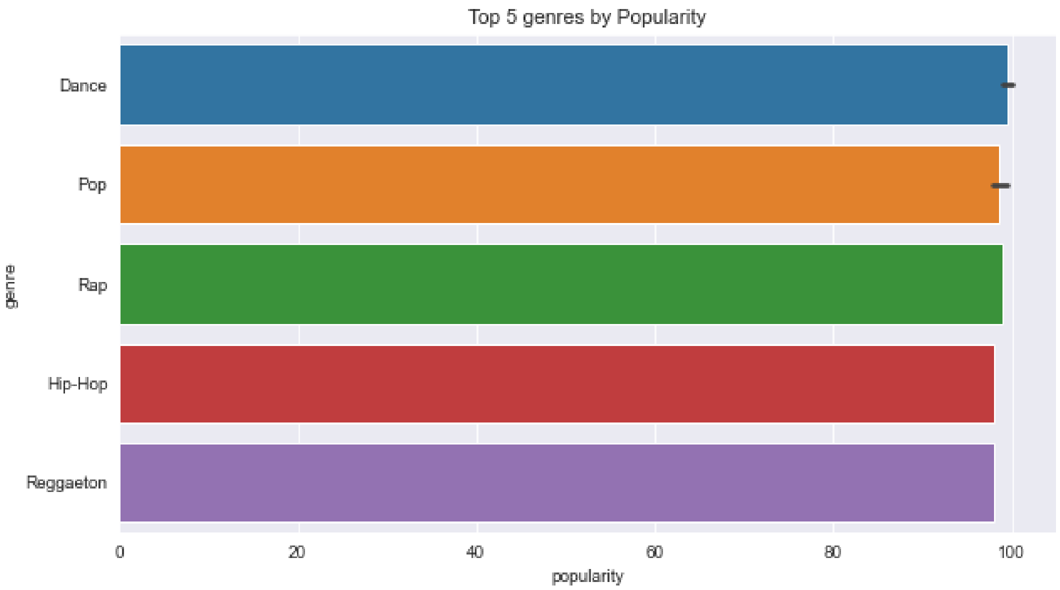
```
Text(0, 0.5, 'Genres')
```



In [18]:

```
sns.set_style(style = "darkgrid")
plt.figure(figsize=(10,5))
famous = df_genre.sort_values("popularity", ascending = False).head(10)
sns.barplot(y='genre', x='popularity', data = famous).set(title="Top 5 genres by Popularity")
Out[18]:

[Text(0.5, 1.0, 'Top 5 genres by Popularity')]
```



CHAPTER 5: FINDINGS, SUGGESTIONS, RECOMMENDATION

FINDINGS, SUGGESTIONS, RECOMMENDATION

1. The dataset has total 20 columns and 586673 rows.
2. Then we sorted the column name popularity in the ascending order and found out that the artist ['Hibari Misora'] and ['Wilms Herbert' and ['June Allyson', 'Joseph Kear...] are the artist who are having the least popular songs.
3. In the step 6 we sorted the column name popularity in which we put an argument and displayed the rows with greater than 90% popularity, and sorted the values in descending order. The artist 'Justin Bieber', 'Daniel Caesar', 'Giveon's song Peaches has popularity as 100 and following to that the next song with popularity 99 is 'Olivia Rodrigo’s Drivers licence and the third in the list is Masked Wolf’s Astronaut In The Ocean.
4. For more clear understanding the duration in milli seconds is converted into seconds.
5. In the step 9 we have used to code to find out correlation between the different fields of the table and some attributes have positive correlation with each other while some have negative.
6. In step 11 we have created the data visualization in which the regression plot has been plotted between the loudness and energy with loudness on y axis and energy on x axis. The energy is independent variable and loudness is dependent variable and found out that the correlation is positive i.e. as the energy increases the loudness also increases.
7. In step 12 we have created the data visualization in which the regression plot has been plotted between the tempo and energy with tempo on y axis and energy on x axis. The energy is independent variable and tempo is dependent variable and found out that the correlation is positive i.e. as the energy increases the tempo also increases.
8. In step 13 we have created the data visualization in which the regression plot has been plotted between the popularity and loudness with popularity on y axis and loudness on x axis. The loudness is independent variable and popularity is dependent variable and found out that the correlation is positive i.e. as the loudness increases the popularity also increases.
9. Contrary to the above steps now we have created the data visualization in which the regression plot has been plotted between the popularity and acousticness with popularity on y axis and acousticness on x axis. The acousticness is independent variable and popularity is dependent variable and

found out that the correlation is Negative i.e. as the acousticness increases the popularity decreases.

10. In the step 13 we have plotted a histogram to check the number of songs that has been released over the time period. The histogram shows that in the early years 1920’s very less songs were released and as the time passed the number of songs released per year has increased a lot showing the highest number of songs were released in the year 2019.
11. In the further step we have imported another dataset which include the column genre so we can gather some information based on the genre of the song.
12. In the step 17 we have used seaborn library for displaying the bar plot for getting information about duration of the songs in different genre and so have found that the genre ‘world’ has the songs which are having the longest duration and following to that is the genre ‘indie’. And the songs which have the shortest duration are having the genre ‘children’s music’
13. In the last step it is found out the top 5 genres by popularity and the top most genre is Dance and following to that the next most popular genre is pop.

The top 5 genres according to their rank are:

1. Dance
2. Pop
3. Rap
4. Hip hop
5. Reggaeton

CHAPTER 6: CONCLUSION

CONCLUSION

A. Spotify has almost most of the songs that are released since 1920. And so it's vital for the company to manage its database, maintain the quality of its app and the features.

B. The popularity of the song depends upon its other attributes such acousticness, energy, loudness, tempo etc.

C. There are more and more songs releasing in the recent years.

D. The duration of the songs is found less in the genre of children's music so the songs that are made for kids are shorter while the songs of the genre "world" were longer in duration.

E. The trends that were calculated in the project could change in the upcoming years as the people and their choices changes through the generations.

BIBLIOGRAPHY

- https://www.google.co.in/ - Used for searching information.
- https://www.geeksforgeeks.org/ - used for learning about data science
- https://www.kaggle.com/datasets - Used for to select dataset
- https://pandas.pydata.org/docs/ - Pandas Documentation
- https://numpy.org/doc/ - Numpy Documentation
- https://matplotlib.org/stable/index.html - Matplotlib Documentation
- https://www.geeksforgeeks.org/python-visualize-missing-values-nan-values-using-missingno-library/ - Missingno Information
- https://ipython.readthedocs.io/en/stable/interactive/plotting.html - matplotlib inline Documentation
- https://plotly.com/python/plotly-express/ - Plotly express in Python

REFERENCES

- https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science
- https://en.wikipedia.org/wiki/Spotify
- https://www.geeksforgeeks.org/python-for-data-science/
- https://www.geeksforgeeks.org/data-cleansing-introduction/
- https://www.interviewbit.com/blog/python-libraries/
- C:\Users\Mahesh\Downloads\tracks.csv
- C:\Users\Mahesh\Downloads\SpotifyFeatures.csv
- https://www.dataquest.io/blog/what-is-data-science/
- https://www.heavy.ai/learn/data-science
- https://towardsdatascience.com/why-data-science-succeeds-or-fails-c24edd2d2f9
- https://www.simplilearn.com/
- https://www.youtube.com/

ANNEXURE

A – QUESTIONNAIRE

1. Which songs are more popular on spotify and how spotify can make use of this information?

2. Which artists are delivering the most popular songs?

3. How spotify can use the information to be unique from the other music streaming apps?

4. Which genre has more popularity?

5. How spotify can increase it subscribers and put on advertisements on the specific songs or genre which are more popular?

6. Which attribute has a correlation with another attribute and how to use this information for making the application more delightful?

7. Many songs were released in the recent years and many will in the upcoming years so spotify needs to manage its huge data in the clouds?

8. How it could make use of top 5 genres for advertising its products and offering subscriptions?

9. Spotify could add more feature to its app so the listeners can make adjustments in the stereo settings?

10. How the popularity of the songs depend upon the artist?

11. How the loudness of the song depends upon the energy?

SCOPE FOR FUTURE STUDY

1. There is a lot of scope for future study and much analysis could be done on the data.
2. As the years passes the more and more songs will be released and more people would overcome as an artist and thus the information conclusions found now will be changed in the future.
3. In the upcoming generations of people the music taste of people could change and so as the popularity of specific song or genre will change. So a lot new findings could be found in the future.
4. Spotify would need to use the updated data so need to analyse their data and find result and make use of it for product marketing and targeting the customers and offer subscriptions.

Photographs, Drawings

https://www.google.co.in/ - All the photographs that are used in the project are used from Google Images.