**Project Based Learning Report**

on

**Spotify Music Analysis Visualization using Python Pandas**

Submitted in the partial fulfillment of the requirements

For the Project based learning in (**Essentials of Data Science**)

in

Electronics & Communication Engineering

By

**2014111113 Simardeep Singh**

**2014111099 Ankur Ranjan**

**2014111089 Manoj Patidar**

Under the guidance of Course In-charge

Prof. Dnyanesh S.Lavhkare

Department of Electronics & Communication Engineering

Bharati Vidyapeeth

(Deemed to be University)

College of Engineering,

Pune – 4110043

**Academic Year: 2021-22**

**Bharati Vidyapeeth**

**(Deemed to be University)**

**College of Engineering,**

**Pune – 411043**

**DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING**

**CERTIFICATE**

Certified that the Project Based Learning report entitled, **“Spotify Music Analysis Visualization using Python Pandas”** is work done by

**2014111113 Simardeep Singh**

**2014111099 Ankur Ranjan**

**2014111089 Manoj Patidar**

in partial fulfillment of the requirements for the award of credits for Project Based Learning (PBL) in **Essentials of Data Science Course** of Bachelor of Technology Semester IV, Electronics & Communication Engineering.

**Date: 21 May 2022**

**Prof. Dnyanesh S.Lavhkare Dr. Tanuja S.Dhope**

**Course In-charge PBL Co-Ordinator**

**Dr. Arundhati A.Shinde**

**Professor & Head**

**ELECTRONICS & COMMUNICATION ENGINEERING**

|  |  |
| --- | --- |
| **Index: -** | |
| Page No. | Contents |
| 1-1 | Problem Statement with Solution |
| 2-5 | Description about project |
| 6-6 | Software Used |
| 7-13 | Results with Analysis |
| 14-14 | Conclusion & Outcome |

**Problem Statement :-**

What is Data Science? Why learn Data Science?

**Solution :-**

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.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply [machine learning](https://www.datarobot.com/wiki/machine-learning/) [algorithms](https://www.datarobot.com/wiki/algorithm/) to numbers, text, images, video, audio, and more to produce [artificial intelligence (AI)](https://www.datarobot.com/wiki/artificial-intelligence/) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate [insights](https://www.datarobot.com/wiki/insights/) which analysts and business users can translate into tangible business value.

Reasons to learn Data Science are: -

1. Learning about data science provides an opportunity for you to recreate yourself.
2. **We live in a digital world, everything is data-driven.** There is data science in **business, accounting, education, science, engineering, healthcare, technology, energy sector, government**, and so on.
3. **Data science is also a very promising field with lots of high paying job opportunities.**
4. **Basic data science skills are important for personal use.**
5. Great potential to branch out with different options.
6. Become a decision-maker, not every job opportunity will give you the power to make informed business decisions. For a data scientist, that is the core responsibility.
7. Less competitive because it is a highly analytical role, competition is less, but demand is not. With a limited talent pool, there is always a challenge for businesses to hire in these roles.

**1**

**Spotify Music Analysis Visualization**

Spotifyis a [proprietary](https://en.wikipedia.org/wiki/Proprietary_software) Swedish [audio streaming](https://en.wikipedia.org/wiki/Streaming_media) and media services provider founded on 23 April 2006 by [Daniel Ek](https://en.wikipedia.org/wiki/Daniel_Ek) and [Martin Lorentzon](https://en.wikipedia.org/wiki/Martin_Lorentzon). It is one of the largest music streaming service providers, with over 422 million monthly [active users](https://en.wikipedia.org/wiki/Active_users), including 182 million paying subscribers, as of March 2022. Spotify is listed (through a [Luxembourg City](https://en.wikipedia.org/wiki/Luxembourg_City)-domiciled [holding company](https://en.wikipedia.org/wiki/Holding_company), Spotify Technology S.A) on the [New York Stock Exchange](https://en.wikipedia.org/wiki/New_York_Stock_Exchange) in the form of [American depositary receipts](https://en.wikipedia.org/wiki/American_depositary_receipt).

Spotify was founded in 2006 in [Stockholm](https://en.wikipedia.org/wiki/Stockholm), Sweden, by [Daniel Ek](https://en.wikipedia.org/wiki/Daniel_Ek), former [CTO](https://en.wikipedia.org/wiki/Chief_technology_officer) of [Stardoll](https://en.wikipedia.org/wiki/Stardoll" \o "Stardoll), and Martin Lorentzon, co-founder of [Tradedoubler](https://en.wikipedia.org/wiki/Tradedoubler" \o "Tradedoubler). According to Ek, the company's title was initially misheard from a name shouted by Lorentzon. Later they thought out a [portmanteau](https://en.wikipedia.org/wiki/Portmanteau) of "spot" and "identify." In February 2009, Spotify opened public registration for the free service tier in the United Kingdom. Registrations surged following the release of the mobile service, leading Spotify to halt registration for the free service in September, returning the UK to an invitation-only policy.

**Datasets: -**

We have downloaded three datasets about the Spotify Music from KAGGLE.com site which are-

artists.csv, tracks.csv and SpotifyFeatures.csv. We have performed analysis visualization on dataset, tracks.csv on Jupyter Notebook.

**Libraries used: -**

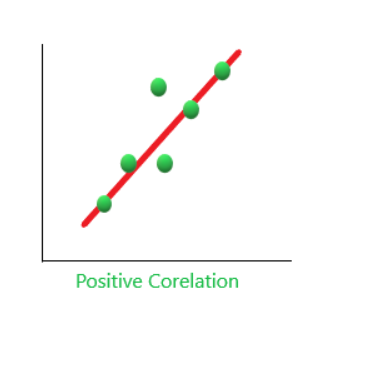
1. Numpy library **-** numpy is used to perform various mathematical operations on arrays.
2. Pandas Library **-** pandas provides various data structures and operations for manipulating numerical data and time series.
3. Matplotlib library from which pyplot module is used for plotting library used for 2D graphics.
4. Seaborn library - seaborn is a library for making statistical graphics in Python.

**2**

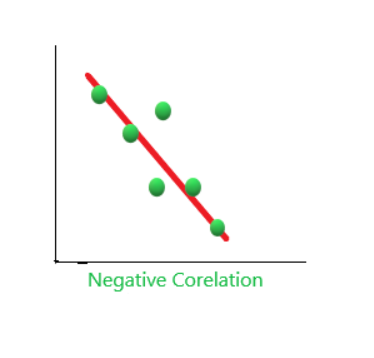
**Correlation: -**

Correlation means an association; it is a measure of the extent to which two variables are related.

1. Positive Correlation: When two variables increase together and decrease together. They are positively correlated. ‘1’ is a perfect positive correlation. For example – demand and profit are positively correlated the more the demand for the product, the more profit hence positive correlation.

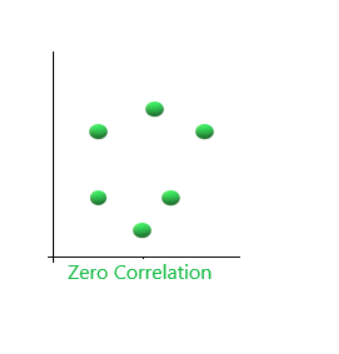


1. Negative Correlation: When one variable increase and the other variable decreases together and vice-versa. They are negatively correlated. For example, If the distance between magnet increases their attraction decreases, and vice-versa. Hence, a negative correlation. ‘-1’ is no correlation.

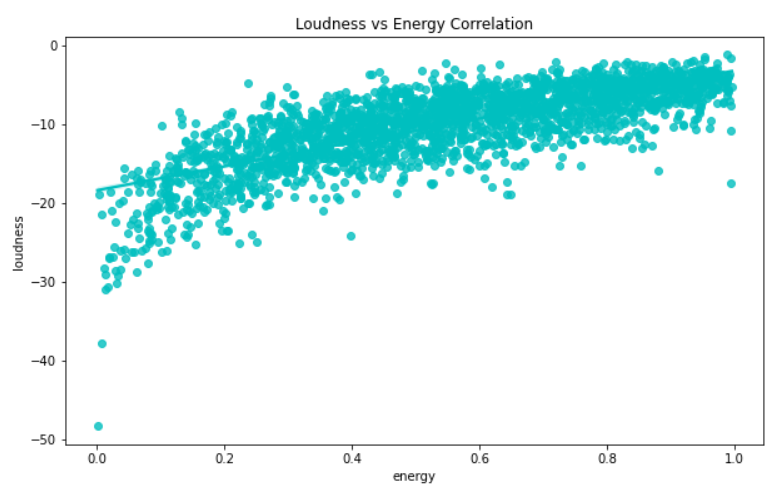


**3**

1. Zero Correlation (No Correlation): When two variables don’t seem to be linked at all. ‘0’ is a perfect negative correlation. For Example, the amount of tea you take and level of intelligence.



**Regression: -**

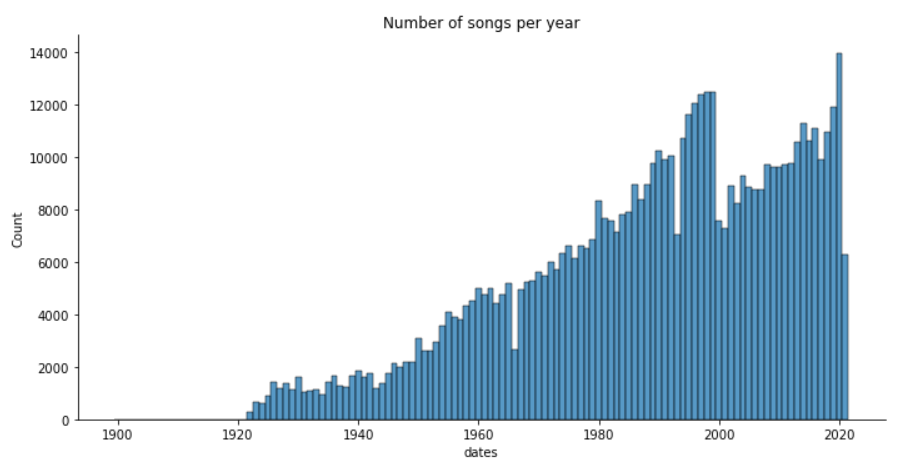
Regression, one of the most common types of [machine learning](https://www.datarobot.com/wiki/machine-learning/) [models,](https://www.datarobot.com/wiki/model/) estimates the relationships between variables. Whereas [classification](https://www.datarobot.com/wiki/classification/) models identify which category an observation belongs to, regression models estimate a numeric value.

The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses. Regression plots as the name suggests creates a regression line between 2 parameters and helps to visualize their linear relationships.

**4**

**Distribution/Density Plot: -**

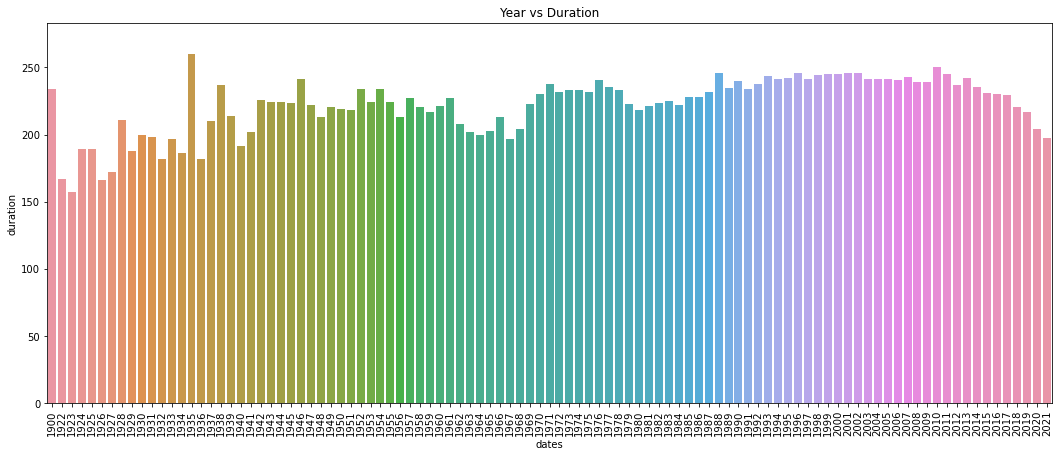
A distribution or density plot depicts the distribution of data over a continuous interval. A density plot is like a smoothed histogram and visualizes the distribution of data over a continuous interval. So, a density plot also gives into what might be the distribution of the population.



**Bar plot: -**

The bar chart is a frequency chart for a qualitative variable. A bar chart can be used to access the most-occurring and least-occurring categories within a dataset. A bar graph is a data visualization technique that can be used to represent numerical values in a dataset to show how different datapoints vary from each other. It utilizes a bar as a measure of magnitudes. The bigger the bar, the higher the number.

To draw a bar chart, call ‘barplot()’ of the seaborn library.

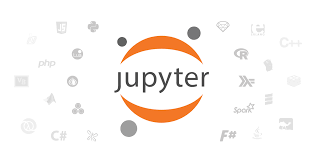
****

**5**

**Software Used: -**

**Jupyter Notebook: -**

Jupyter Notebook (formerly IPython Notebooks) is a [web-based interactive](https://en.wikipedia.org/wiki/Web_application) computational environment for creating [notebook](https://en.wikipedia.org/wiki/Notebook_interface) documents. A Jupyter Notebook document is a browser-based [REPL](https://en.wikipedia.org/wiki/Read%E2%80%93eval%E2%80%93print_loop) containing an ordered list of input/output cells which can contain code, text (using [Markdown](https://en.wikipedia.org/wiki/Markdown)), mathematics, [plots](https://en.wikipedia.org/wiki/Plot_(graphics)) and [rich media](https://en.wikipedia.org/wiki/Interactive_media). Underneath the interface, a notebook is a [JSON](https://en.wikipedia.org/wiki/JSON) document, following a versioned schema, usually ending with the “. ipynb" extension. Jupyter notebooks are built upon a number of popular [open-source](https://en.wikipedia.org/wiki/Open-source_software) libraries.

Jupyter Notebook can connect to many *kernels* to allow programming in different languages. A Jupyter kernel is a program responsible for handling various types of requests ([code execution](https://en.wikipedia.org/wiki/Execution_(computing)), [code completions](https://en.wikipedia.org/wiki/Autocomplete), inspection), and providing a reply. Kernels talk to the other components of Jupyter using [ZeroMQ](https://en.wikipedia.org/wiki/ZeroMQ" \o "ZeroMQ), and thus can be on the same or [remote machines](https://en.wikipedia.org/wiki/Remote_computer). Unlike many other Notebook-like interfaces, in Jupyter, kernels are not aware that they are attached to a specific document, and can be connected to many clients at once. Usually kernels allow execution of only a single language, but there are a couple of exceptions. By default Jupyter Notebook ships with the IPython kernel. As of the 2.3 release (October 2014), there are 49 Jupyter-compatible kernels for many programming languages, including Python, R, Julia and Haskell.

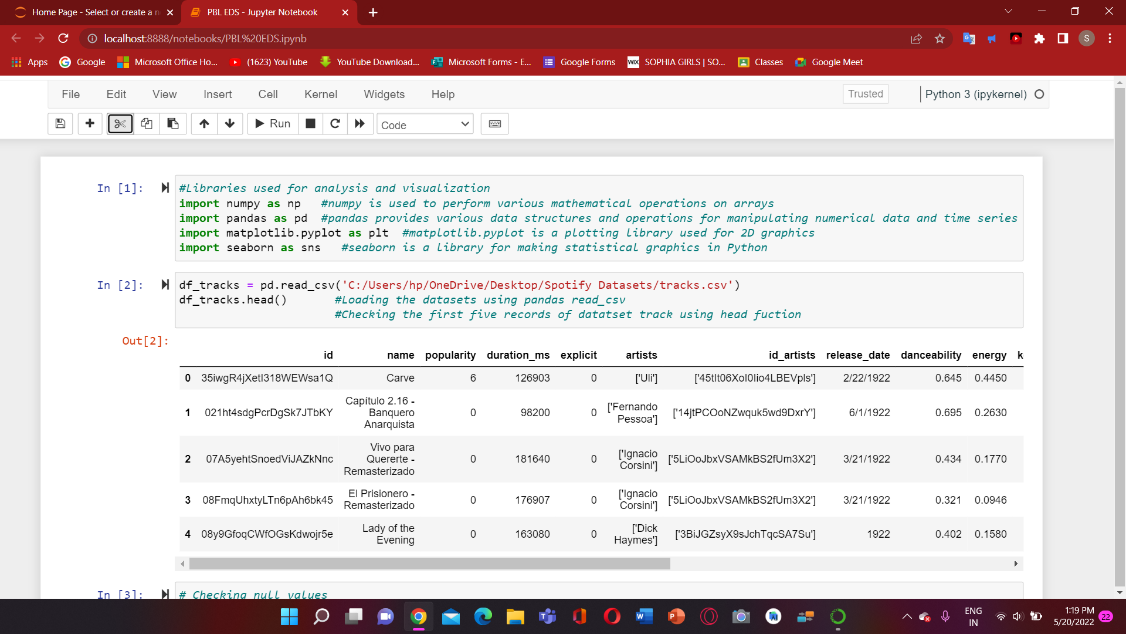
**6**

**Result with Analysis**

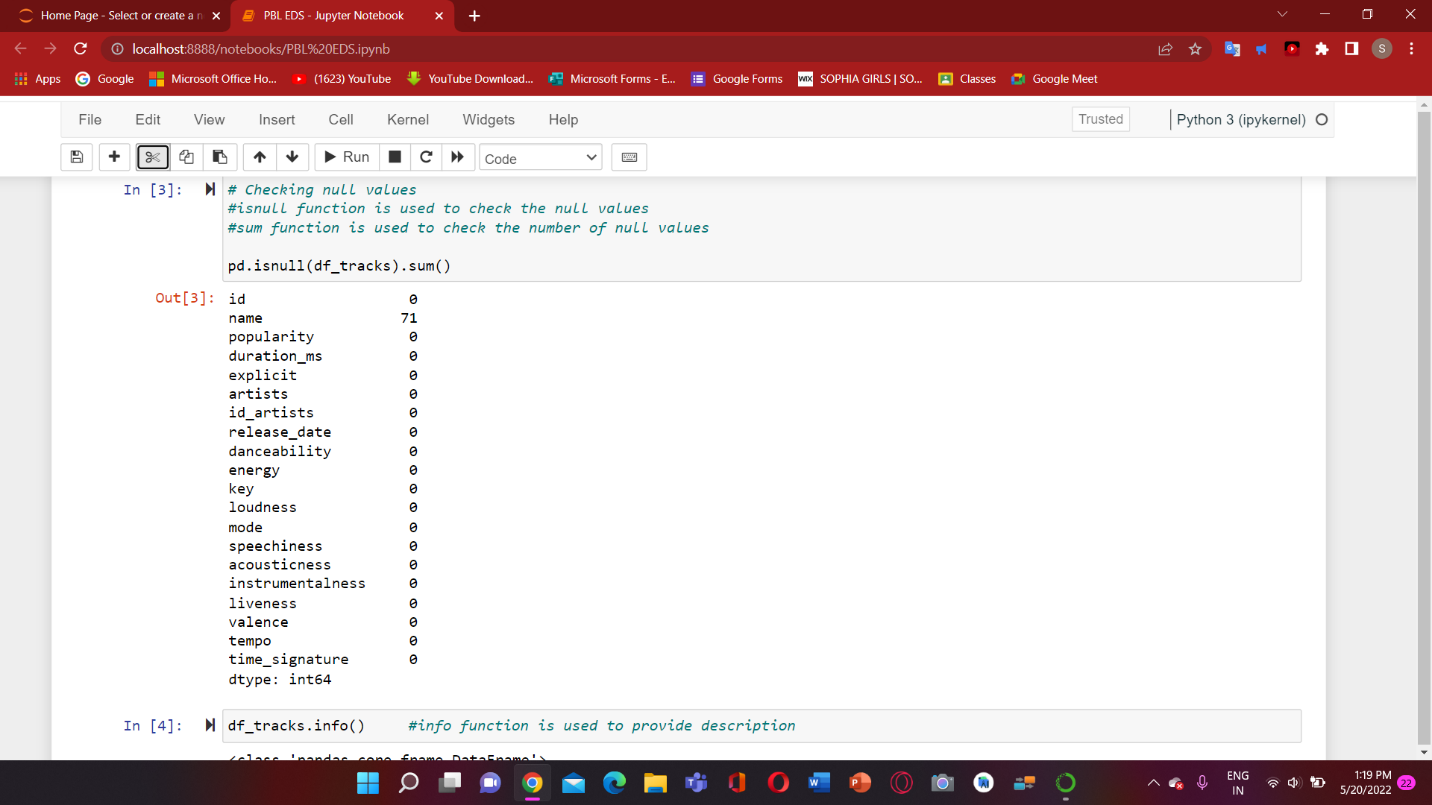
**Analysis of the code: -**

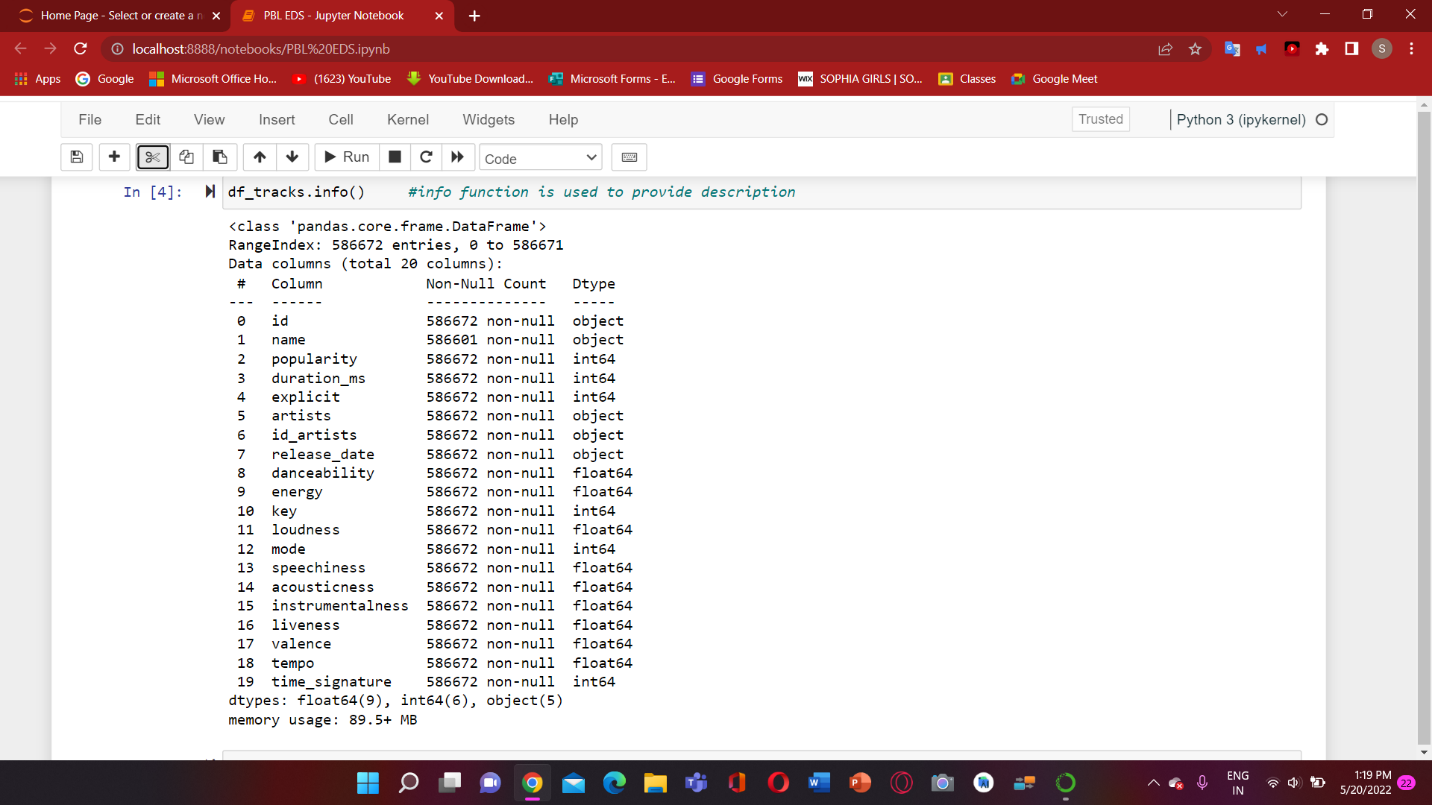
* First, we have imported four libraries – numpy as np, pandas as pd, matplotlib.pyplot as plt and seaborn library as sns.
* Secondly, we have loaded our dataset – tracks.csv using read\_csv() function of pandas library and used head() function for displaying first five rows of the dataset.
* Then, for checking null values in the dataset, we have used isnull() function of pandas library.
* For having the top 10 least popular songs, we have used sorted function to sort the information from the dataset. Then wehave set a condition to display only top 10 songs which have the popularity greater than 90.Then, we have converted the duration column given in milliseconds into seconds by using the lambda function.
* After this, we have created a Correlation Map using heatmap() function of Seaborn library and then , we have created a Regression Plot using regplot() function of Seaborn library.
* Then, we used displot() function of Seaborn library, to create a Distribution plot for the ‘number of songs per year’. At last, we used barplot() function of Seaborn library to create a barplot.

**Screenshots of code: -**

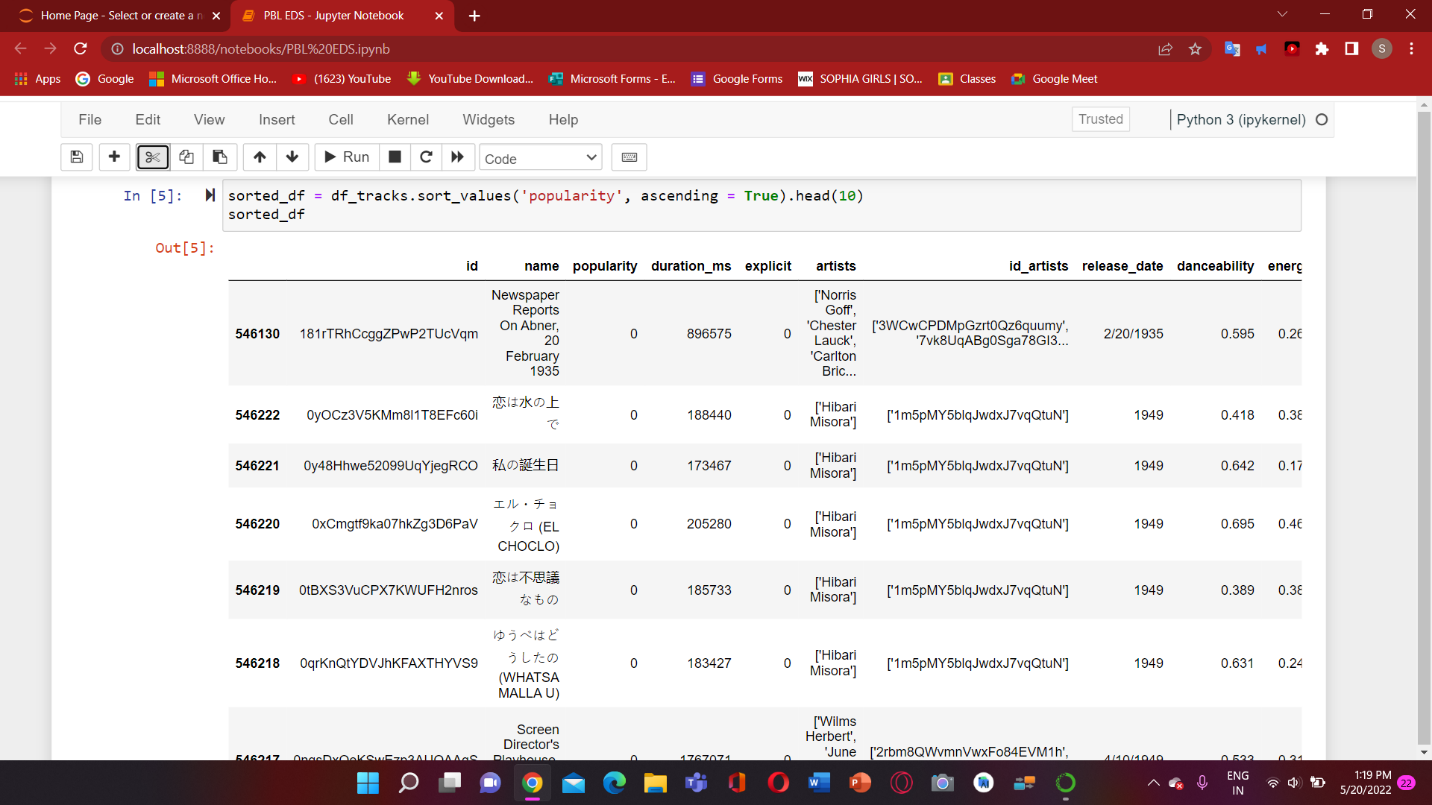


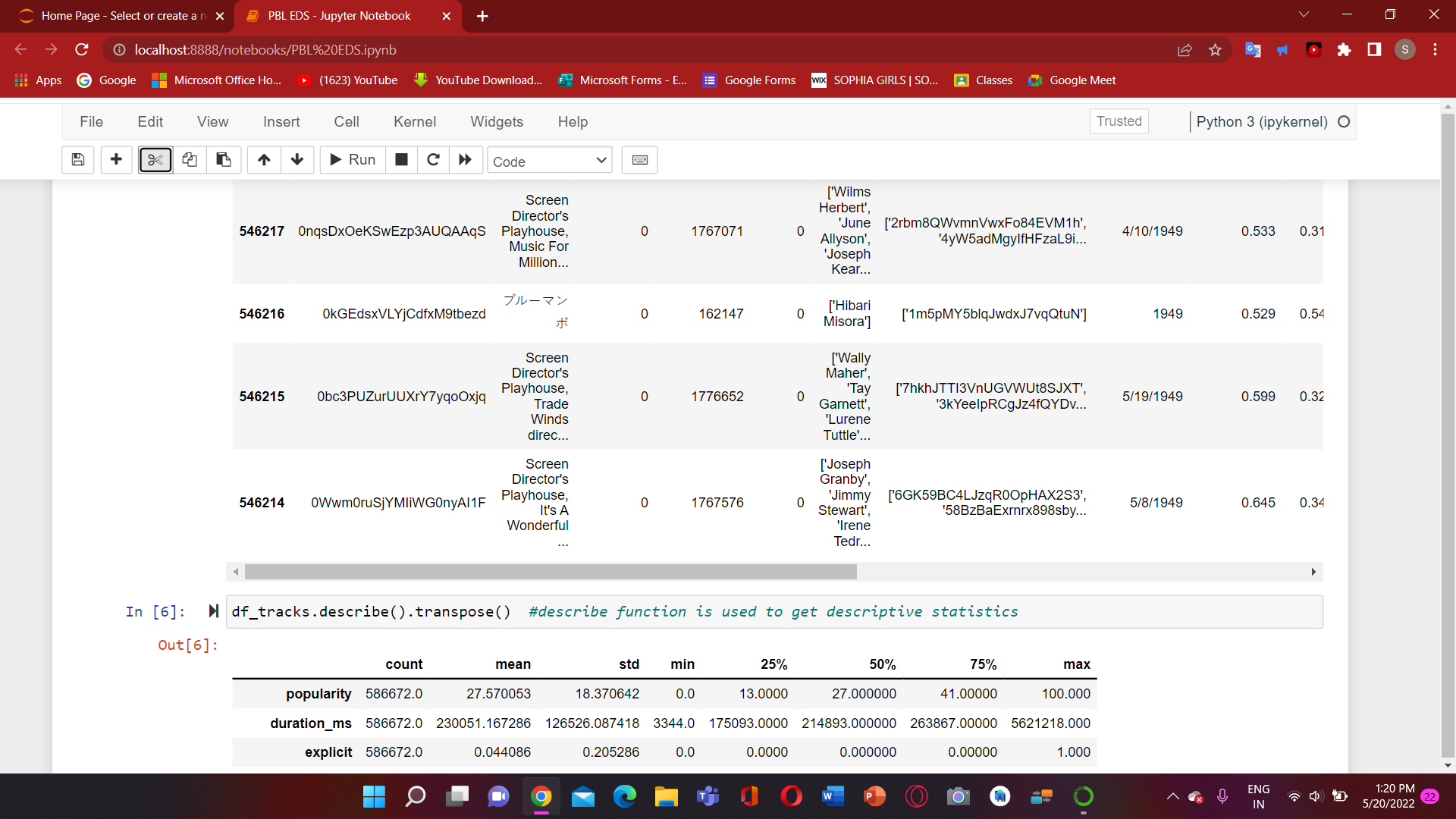
**7**



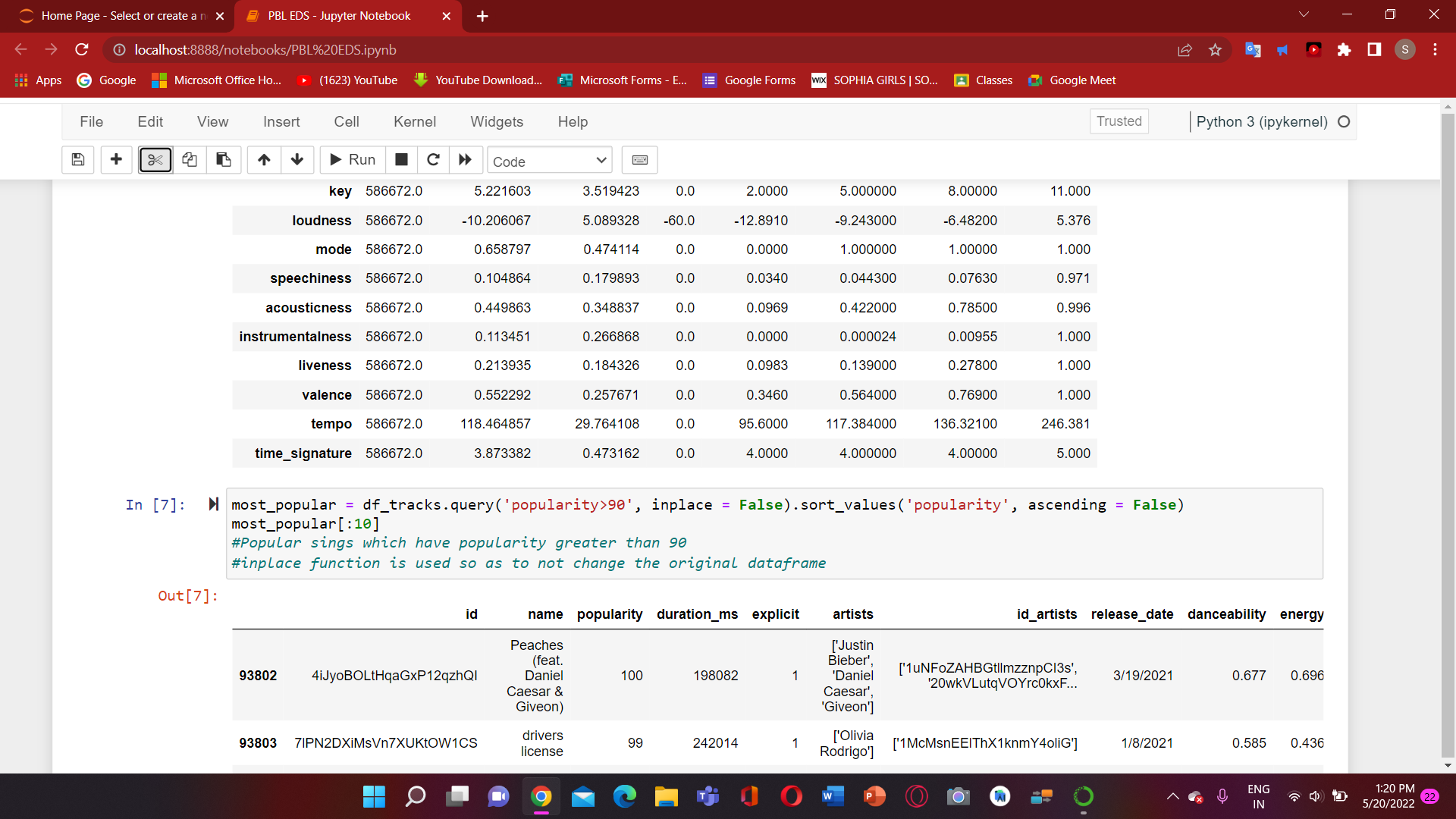


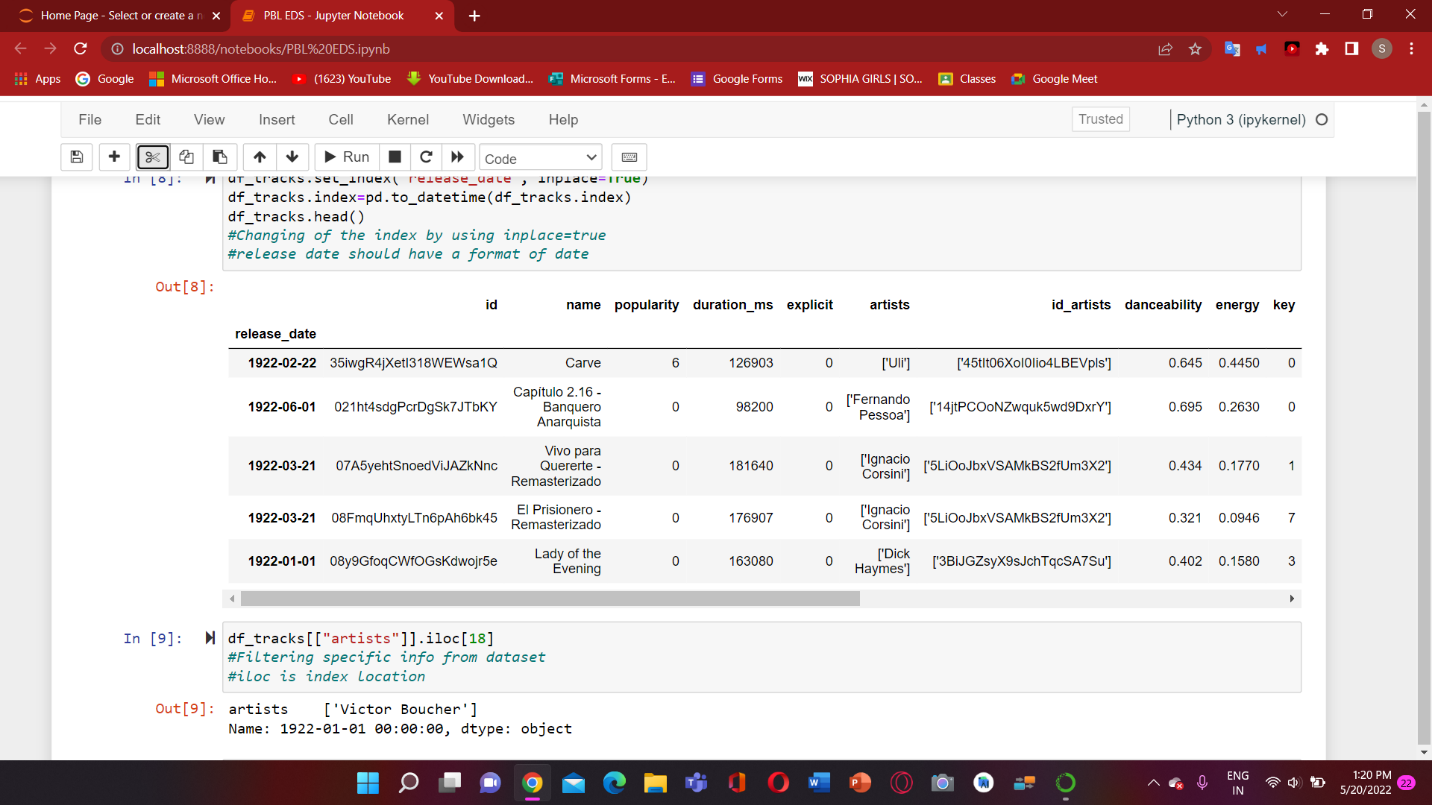
**8**



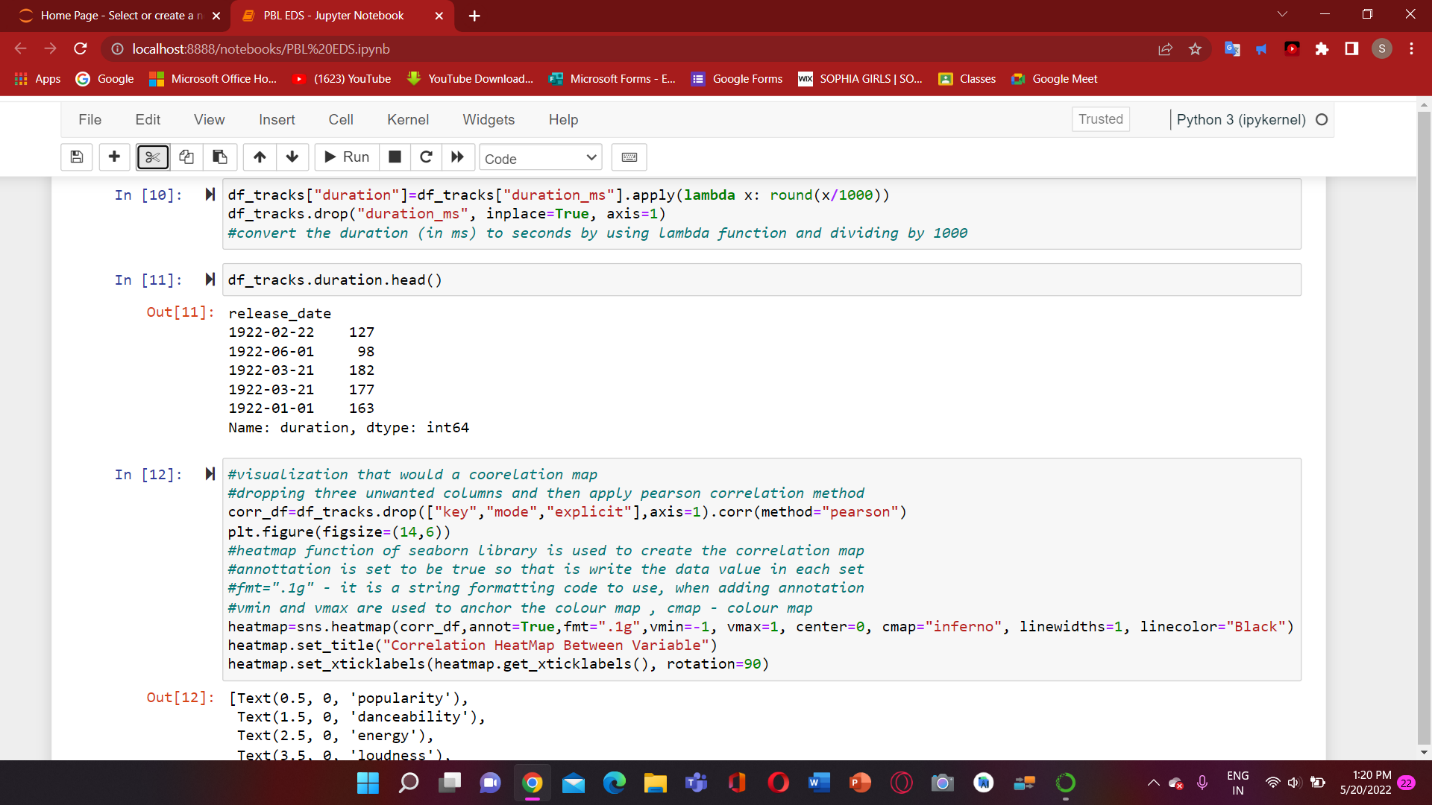


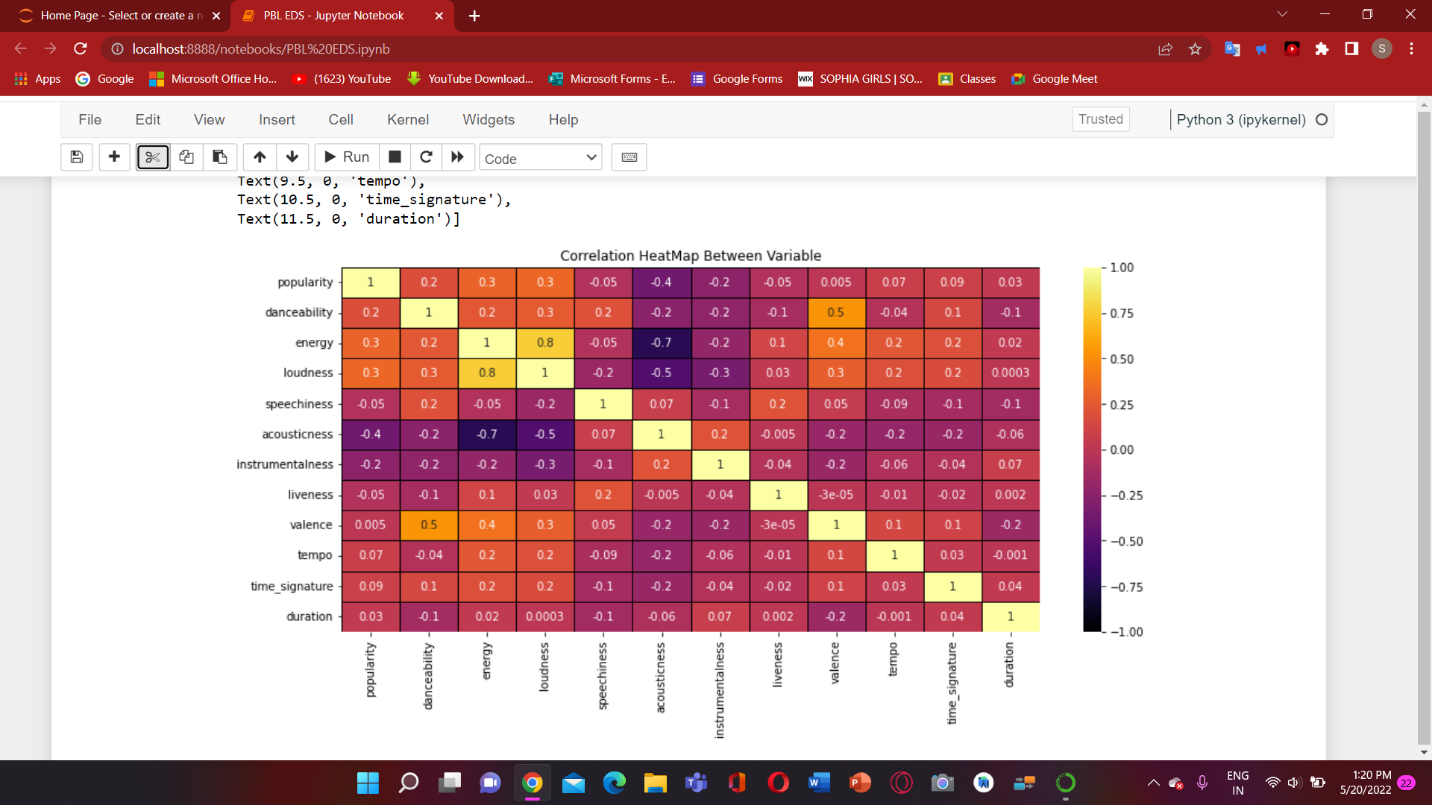
**9**



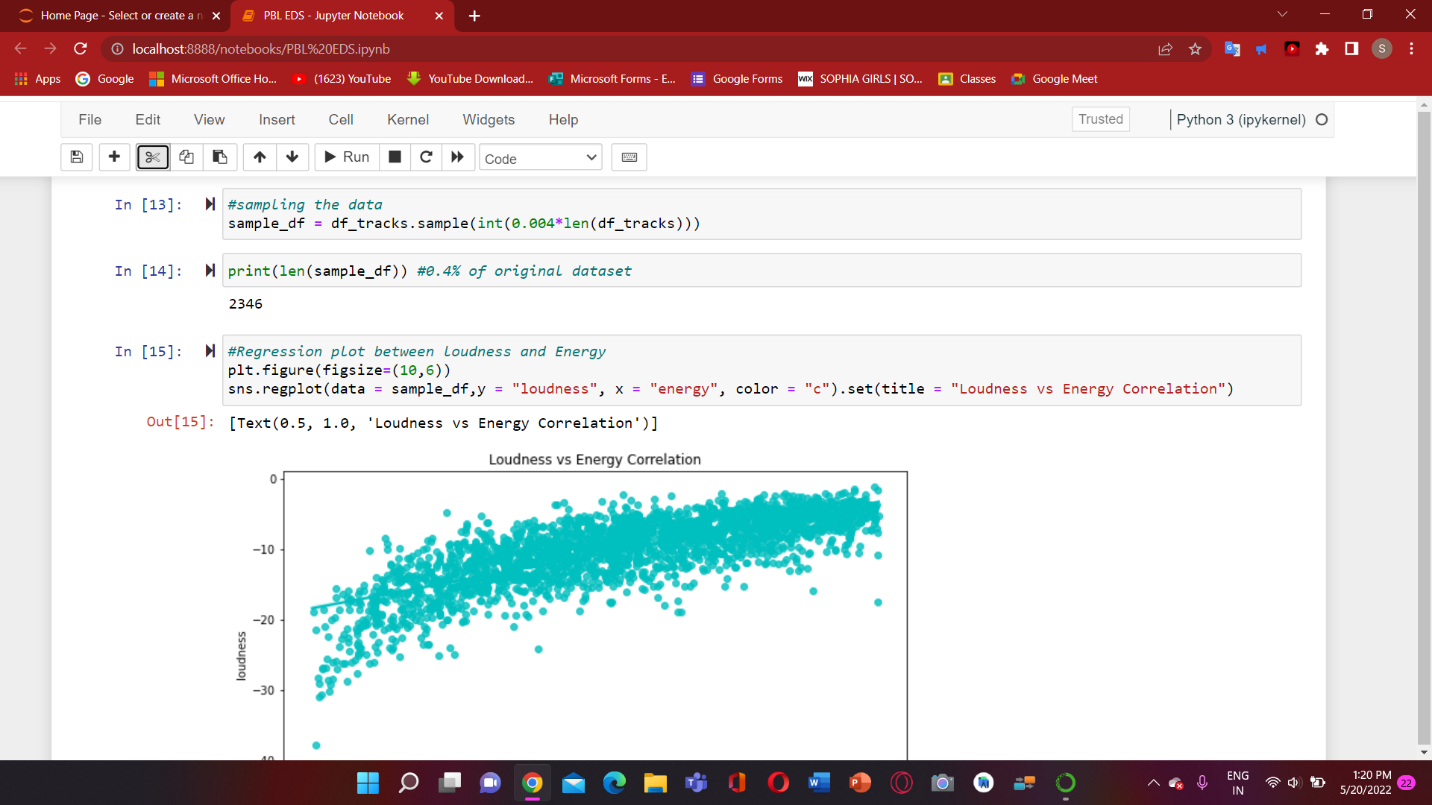


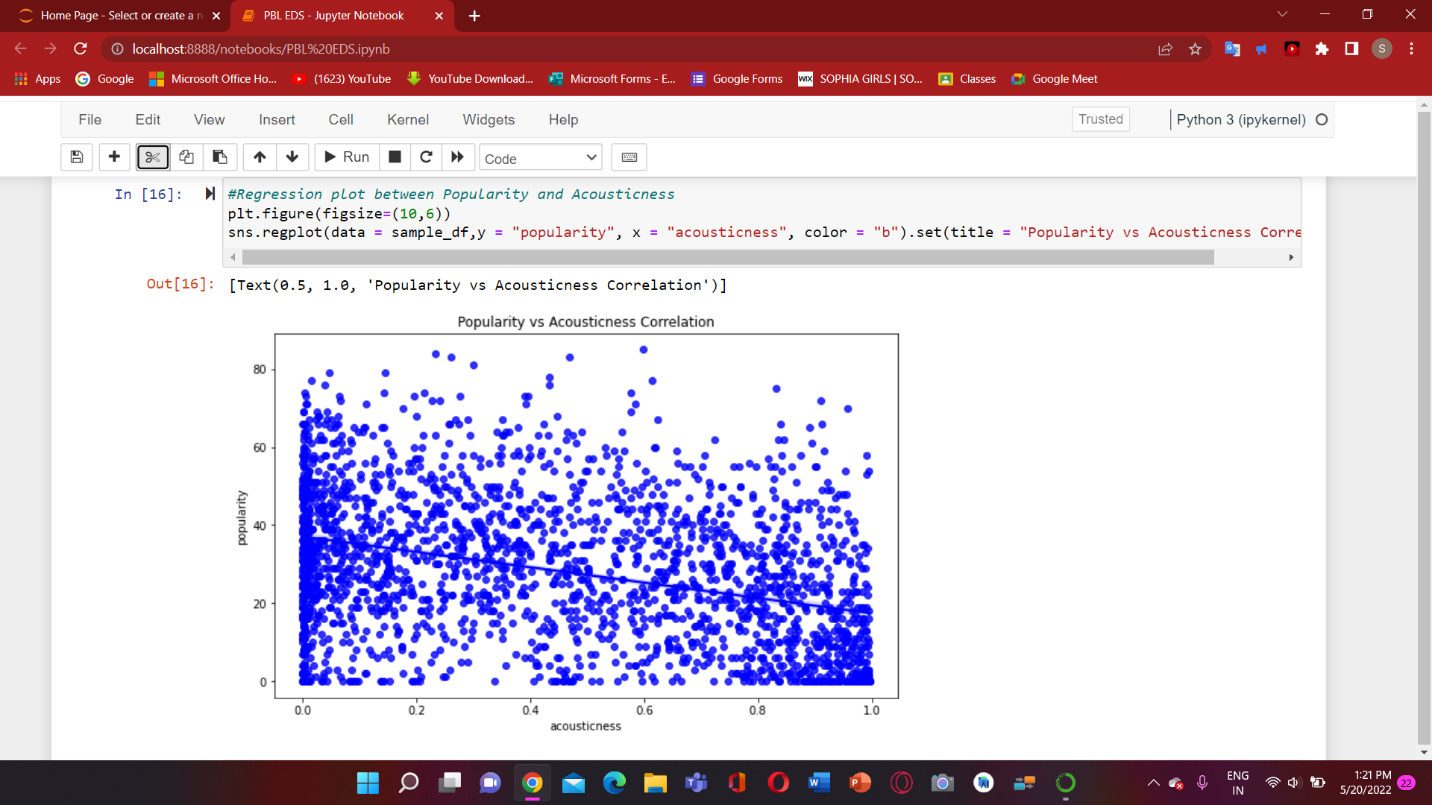
**10**



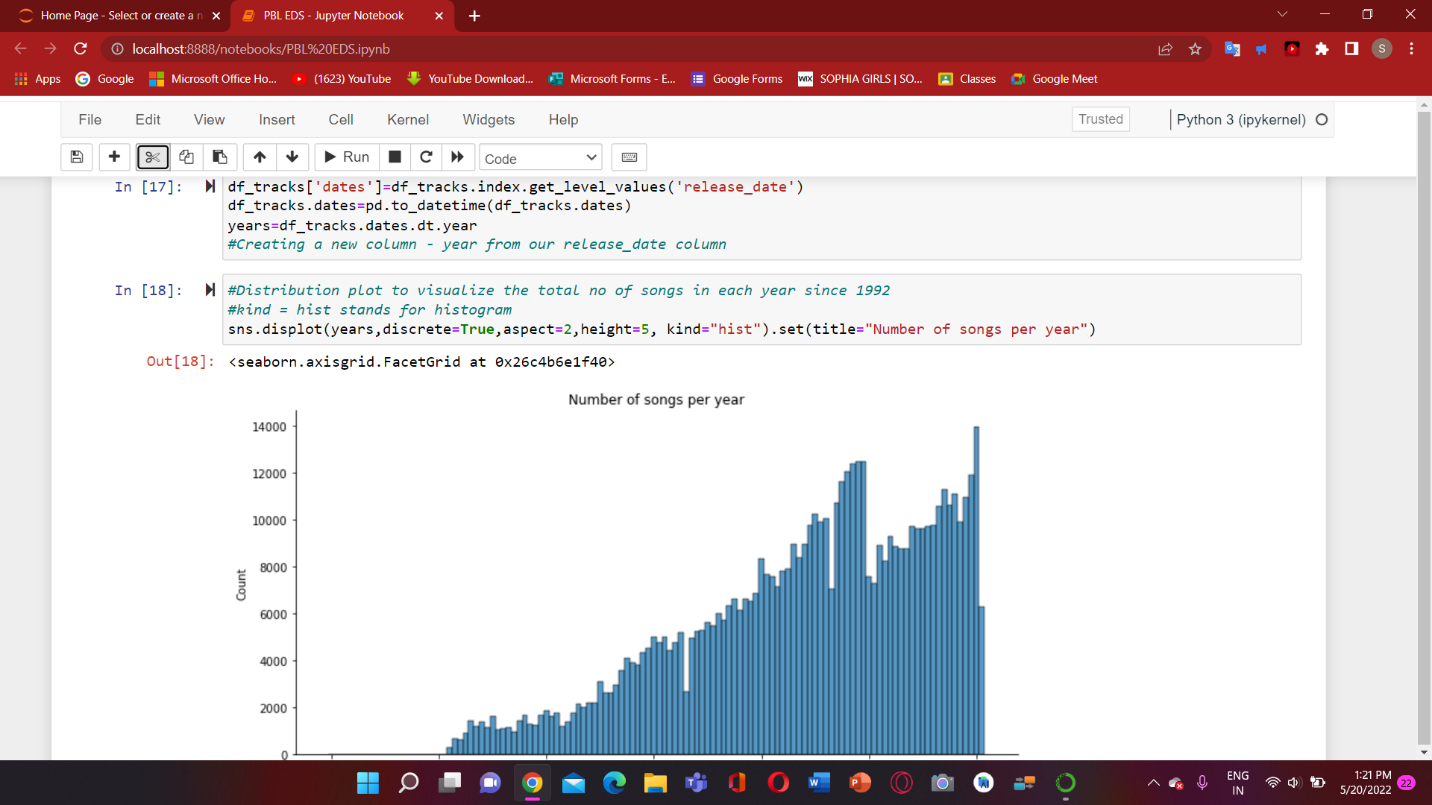


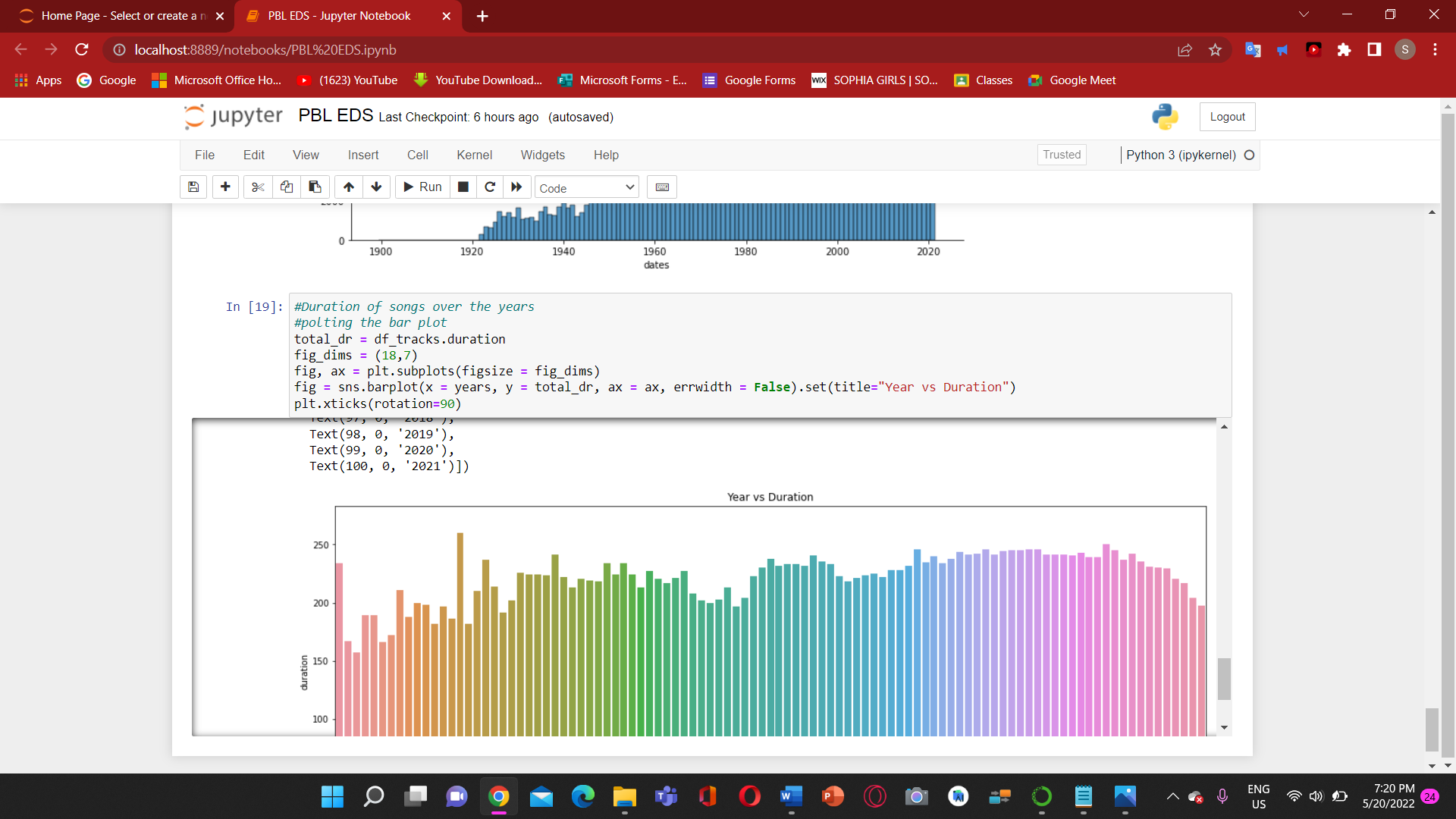
**11**





**12**





**13**

**Project Outcome: -**

From this project, we learnt to describe a flow process for data science problems and classified data science problems into standard typology. We also learnt about correlating results to the solution approach followed and assessing the solution approach.

**Project Conclusion: -**

From this project, we gained the knowledge of software – Jupyter Notebook. We learnt to analyse the datasets and afterwards, visualizing them. We learnt about various plots which are – Correlation plots, Regression plot, Distribution plot and Bar chart.

**14**