1. In a matrix like that of transaction data, it is easier to quickly read and analyze the data at hand with just a glance. For example, if we were to look at the transaction data of a store where a value of 1 indicates that it was purchased and a value of 0 that it was not, you can tell that the column populated with majority of 1s indicates that it's a popular item and needs to restocked often at the store.
   However, there can be some disadvantages to a transaction-item matrix. One would be when we have more than 2 dimensions. This makes it difficult to analyze the dataset. For example, with a 2 dimensional dataset we can easily visualize it and understand what's happening. But if we have a 4 dimensional dataset, we can't visualize it in a table like we do with a 2 dimensional matrix. Hence, readability is limited to the size of the matrix.
   Lastly, why a transaction-item matrix is an example of a dataset that has asymmetric discrete features is that most transactions are going to consist only of frequently purchased items. So the matrix will always look skewed with so many items not being purchased. For example, if we look at the transaction-item matrix dataset at a store like Target, we will find that daily essentials like milk, bread, eggs, vegetables, fruits, and toiletries are purchased on a daily basis almost on every bill whereas things like ipad, tv, bedding, and furniture won't be purchased as frequently. When trying to increase sales, this poses a challenge in pre-processing the dataset carefully so that even items that are rarely purchased can be generated as recommendations.

2.

	binary/discrete/cont	qualitative/quantitative	reasoning for ambiguity if any
(a)	discrete/continuous	quantitative	We don't know how the instruments show the readings. So it could show whole numbers or with decimal points also
(b)	discrete	qualitative	N/A
(c)	discrete	quantitative	N/A
(d)	discrete	qualitative	N/A
(e)	continuous	quantitative	N/A

3. Arithmetic mean: To calculate the Arithmetic mean of the speedups we'll first add them up and divide it by the number of items: (2 + 0.5 + 2 + 1 + 10)/5 = 15.5/5 = 3.1.
   Geometric mean: For Geometric mean we find the product of all the numbers and find the nth root of it: $\sqrt[5]{2 * 0.5 * 2 * 1 * 10}$ = $\sqrt[5]{20}$ = 1.82.
   Harmonic mean: This will be the reciprocal of the arithmetic of the reciprocals of all the numbers: 1/(1/2 + 1/0.5 + 1/2 + 1/1 + 1/10) = 1/(0.5 + 2 + 0.5 + 1 + 0.1) = 1/(4.1) = 0.24.

For the given problem, Arithmetic mean doesn't make sense because we have a value of 10 which is away from the rest of the data and Arithmetic mean is affected by outliers. With Harmonic mean, we're mostly dealing with rates. And so even that is suitable for our problem. And finally with Geometric mean, we are dealing with growth, or, in other words, how much better our case in question has become. And we are trying to calculate how much better the new algorithm is when compared to the baseline performance. Therefore, Geometric mean is most suitable for this problem.

4.

Noise	Outliers
Noise are data points created due to things like erroneous data entry, mislabeling, and so on.	Outliers are data points that actually belong to the dataset and although far away from the remaining dataset, they are examples of exceptions in the dataset.
Ideally noise is never desirable in our dataset. In real life applications, when data is passed in real time to our model, the data will be free from error and so our models don't have to be familiar with it. Noisy data can make any existing patterns in the data unclear and hence they should be removed as much as possible in the pre-processing stage itself.	Outliers are examples of exceptions in our use cases. They can be useful for our model to become familiar with as they can be present as input in real life applications.
Noise objects can be like outliers but not become an outlier itself as they are still erroneous entries. They can be far or close to the dataset but still are classified as noise.	Outliers are not always noise objects. Though sometimes it may be difficult for the model that is not complex enough to include them, they are still usually data points that are genuinely different from the rest of the data.
A noise can make a usual value into an unusual one and it can also make an unusual value into a typical one. For example, during the data entry phase instead of entering an outlier value say 120, if it's entered as 12, then that's turning an unusual value into a typical value. The vice versa of the same can also happen turning a typical value into an unusual one.	

5. The following are some of highlights noted about ChatGPT as a search term in Google trends:

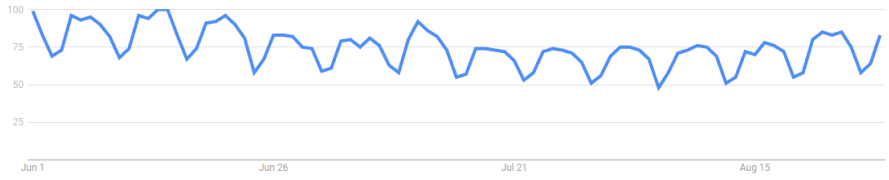
(a) When we compare Google as a search term with ChatGPT, Google is still being used significantly higher than ChatGPT. Although initially Google had issued a red alert after seeing ChatGPT's impact, the search trends show that Google as a search term is dramatically higher than ChatGPT's. ChatGPT might have created a buzz but it's nowhere near overtaking Google. This could be because Google has been the default for so long and gives real time results whereas ChatGPT is trained over data up until 2021. Hence, ChatGPT is for now acting as a very helpful AI assistant but not as a complete replacement for Google.

(b) Although China has the highest traffic for ChatGPT as a search term, when we observe the trend in China alone, we see that it's actually declining faster than any other region. Other regions, the trend is either almost the same or increasing compared to the past few months. Especially since February 2023, there was a sudden steep decline.

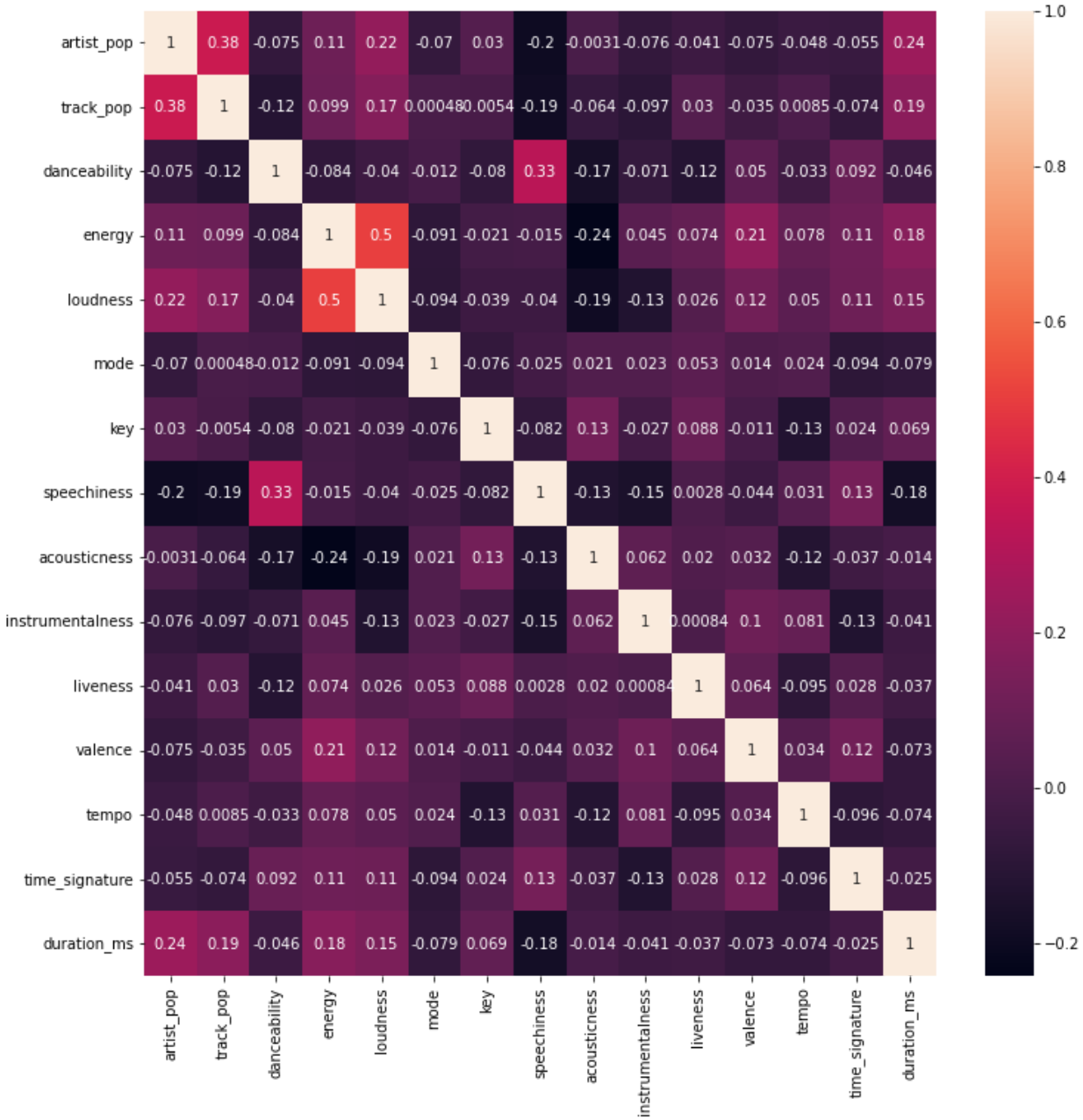
(c) A lot of African countries show no sign of using ChatGPT. This might mean that there is a major problem of lack of infrastructure and accessibility to computers and the internet in Africa. ChatGPT has the potential of dramatically accelerating an individual's productivity and efficiency. This means a lot of our time and energy can be channelized on things that actually matter. But a whole country is unable to leverage this, a major problem that needs to be addressed.

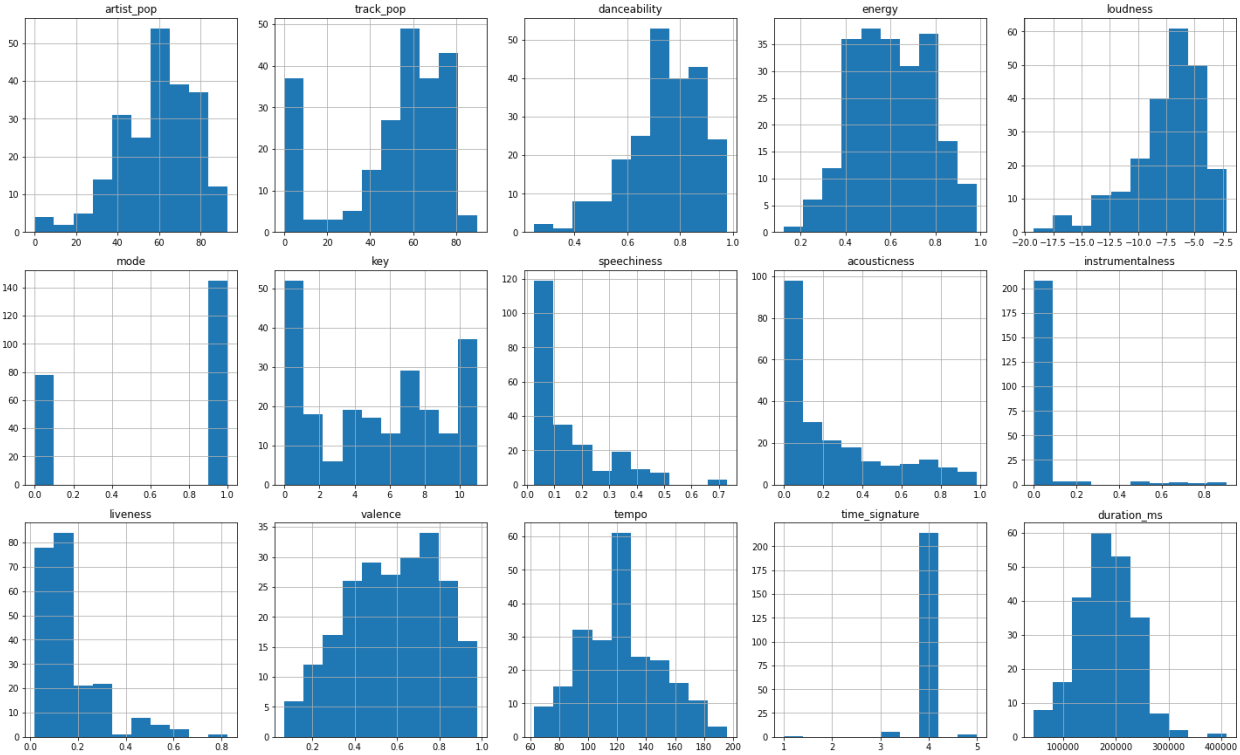
(d) For some reason, there is a huge difference in the number of people searching ChatGPT in Top 2 countries, but only at first glance. If we change the time period from the default 12 months to past 90 days, we observe the difference to be less significant. This could mean that although China has been ahead for a while, other countries for some reason only more recently started using it. At this rate, there is a possibility that over the course of next few months, another country will be above China in search traffic for ChatGPT as a term.

(e) Also a fun additional observation that can be made is how the graph looks for the past 90 days like a sine wave.



6.


artist_pop 1 0.38 -0.075 0.11 0.22 -0.07 0.03 -0.2 -0.0031 -0.076 -0.041 -0.075 -0.048 -0.055 0.24
track_pop 0.38 1 -0.12 0.099 0.17 0.00048 0.0054 -0.19 -0.064 -0.097 0.03 -0.035 0.0085 -0.074 0.19
danceability -0.075 -0.12 1 -0.084 -0.04 -0.012 -0.08 0.33 -0.17 -0.071 -0.12 0.05 -0.033 0.092 -0.046
energy 0.11 0.099 -0.084 1 0.5 -0.091 -0.021 -0.015 -0.24 0.045 0.074 0.21 0.078 0.11 0.18
loudness 0.22 0.17 -0.04 0.5 1 -0.094 -0.039 -0.04 -0.19 -0.13 0.026 0.12 0.05 0.11 0.15
mode -0.07 0.00048 -0.012 -0.091 -0.094 1 -0.076 -0.025 0.021 0.023 0.053 0.014 0.024 -0.094 -0.079
key 0.03 -0.0054 -0.08 -0.021 -0.039 -0.076 1 -0.082 0.13 -0.027 0.088 -0.011 -0.13 0.024 0.069
speechiness -0.2 -0.19 0.33 -0.015 -0.04 -0.025 -0.082 1 -0.13 -0.15 0.0028 -0.044 0.031 0.13 -0.18
acousticness -0.0031 -0.064 -0.17 -0.24 -0.19 0.021 0.13 -0.13 1 0.062 0.02 0.032 -0.12 -0.037 -0.014
instrumentalness -0.076 -0.097 -0.071 0.045 -0.13 0.023 -0.027 -0.15 0.062 1 0.00084 0.1 0.081 -0.13 -0.041
liveness -0.041 0.03 -0.12 0.074 0.026 0.053 0.088 0.0028 0.02 0.00084 1 0.064 -0.095 0.028 -0.037
valence -0.075 -0.035 0.05 0.21 0.12 0.014 -0.011 -0.044 0.032 0.1 0.064 1 0.034 0.12 -0.073
tempo -0.048 0.0085 -0.033 0.078 0.05 0.024 -0.13 0.031 -0.12 0.081 -0.095 0.034 1 -0.096 -0.074
time_signature -0.055 -0.074 0.092 0.11 0.11 -0.094 0.024 0.13 -0.037 -0.13 0.028 0.12 -0.096 1 -0.025
duration_ms 0.24 0.19 -0.046 0.18 0.15 -0.079 0.069 -0.18 -0.014 -0.041 -0.037 -0.073 -0.074 -0.025 1
artist_pop
track_pop
danceability
energy
loudness
mode
key
speechiness
acousticness
instrumentalness
liveness
valence
tempo
time_signature
duration_ms
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


artist_pop
track_pop
danceability
energy
loudness
mode
key
speechiness
acousticness
instrumentalness
liveness
valence
tempo
time_signature
duration_ms

7. The problem that this research paper is addressing is to reduce exposure to harmful content on social media while balancing the interests of creators, users, and platforms. To achieve this they use an algorithm called Gamine, a fast greedy algorithm that reduces the exposure to harmful content via edge rewiring, i.e., replacing certain recommendations by others. And it works both with and without quality constraints on recommendations. They confirm that it is effective, robust, and efficient in practice. The dataset they use to achieve this is a combination of synthetic and real-world data. All of the programming was done in Python and on a high end computer. It's also worth noting that they mention all of the code, datasets and results are publicly available. After the implementation the authors had evaluated the variation of performance based on modeling choices and input parameters on Gamine. Then they demonstrated the effectiveness of Gamine in reducing exposure to harm compared to existing methods and baselines. Next, they checked to make sure that Gamine is scalable in theory and practice. They also noted the features that made reducing exposure to harm easier and harder on different datasets. Finally, they had mentioned guidelines for reducing exposure to harm in recommendation graphs under budget constraints. Overall, they observed that Gamine robustly reduces the expected total exposure to harm, and that it changes its behavior predictably under parameter variations. Gamine also gives recommendations that are only 5% less relevant than the original recommendations. So it can be relied upon as a replacement to the original recommender except it reduces the exposure to harmful content significantly. When compared to its contenders, Gamine offers more reliable performance and achieves stronger harm reduction than its contenders.

   Personally, I like this paper. Initially the paper even mentions the author of another paper who's research inspired this paper. However, the author attacks the limitations of the original paper saying it fails in some specific ways. And the author then proceeds to say that they will address all of these limitations. This I found very peculiar. But I think that this kind of research can be useful as companies are always focused on how to gain more customers and increase their profits and the consequences of these efforts can be quite concerning. Humanity has made dramatic improvements and has come very far from where it started, however, the side effects of it are not being addressed. I think such kind of research should be encouraged more. Additionally, I also think the paper was very technical and challenging for a layman person to read. If the paper had some simpler explanations to help understand what they're doing, then I believe it can reach more people and inspire them to work on such things as well. I also find it very interesting how they provide a lead at the end to carry on this research further. It shows that there is more potential to this work and is also a calling that this needs more people working on it. Overall, the impact of this paper, in my opinion, can be huge to better steer social media platforms in being responsible with the kind of influence they have on the world.