Clustering Analysis of Screen Time and Wellness

Luis Dale Gascon
Computer Science
Towson University
lgascon1@students.towson.edu

***Abstract*—This study examines how device usage affects personal wellness using a dataset that tracks hours spent on devices, technology habits, and related wellness metrics. Drawing inspiration from customer segmentation techniques in marketing, the proposed approach groups users with similar digital behaviors and wellness patterns through probabilistic clustering. Subsequently, a multi-label classifier is trained to recommend personalized strategies aimed at enhancing individual mental well-being.**

***Index Terms*—wellness, clustering, technology usage, mental health, unsupervised learning**

I. Introduction

The rising use of technology has been linked to negative impacts on mental health and overall well-being. This study investigates how individuals utilize technology across various devices, such as phones, laptops, tablets, and televisions, and for different activities including social media, work, entertainment, and gaming. We explore how these usage patterns correlate with wellness metrics such as sleep quality, mood, stress levels, mental health assessments, and other lifestyle factors.

Utilizing this data, the objective is to identify user segments by clustering the dataset via a Gaussian Mixture Model and to train a multi-label classification model. This model aims to recommend personalized strategies that promote healthier technology habits, applying concepts from user segmentation in the field of e-commerce.

II. Related Works

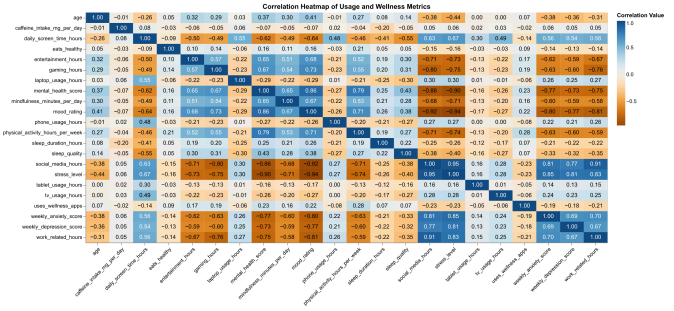
[1] compared several leading clustering algorithms, including K-means, Gaussian Mixture Models (GMM), DBSCAN, agglomerative clustering, and BIRCH, for customer segmentation in the UK retail market. Their findings indicated that GMM, when used alongside PCA for dimensionality reduction, outperformed other methods by achieving a Silhouette Score of 0.80, whereas K-means, BIRCH, and agglomerative algorithms all scored 0.64. A score closer to 1 indicates that the clusters are more defined.

Another study by [2] evaluated clustering performance using both the Silhouette Score and the Davies-Bouldin Index. A higher Davies-Bouldin Index implies that clusters are less compact and not well separated. This research observed that Gaussian Mixture Models encounter difficulties when handling high-dimensional or large-scale datasets, while K-means++ produced more reliable results even in the presence of high dimensionality and noise.

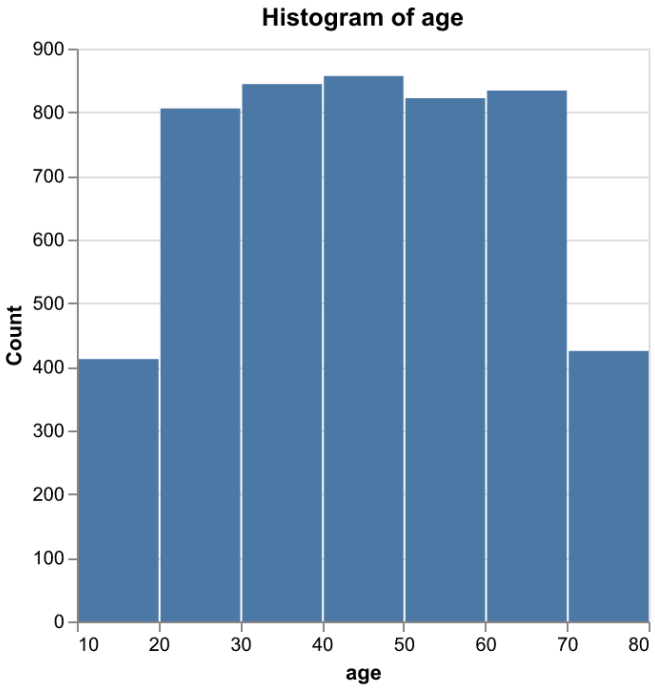
In the context of mental health, research such as this study by [3] explored the relationship between technology use and psychological well-being. The study found that active digital engagement is positively associated with anxiety symptoms, and that access to the internet correlates with higher levels of depression and anxiety, especially among younger individuals. Nevertheless, the paper emphasized that there is no clear causal link established between technology use and mental health outcomes.

III. Dataset

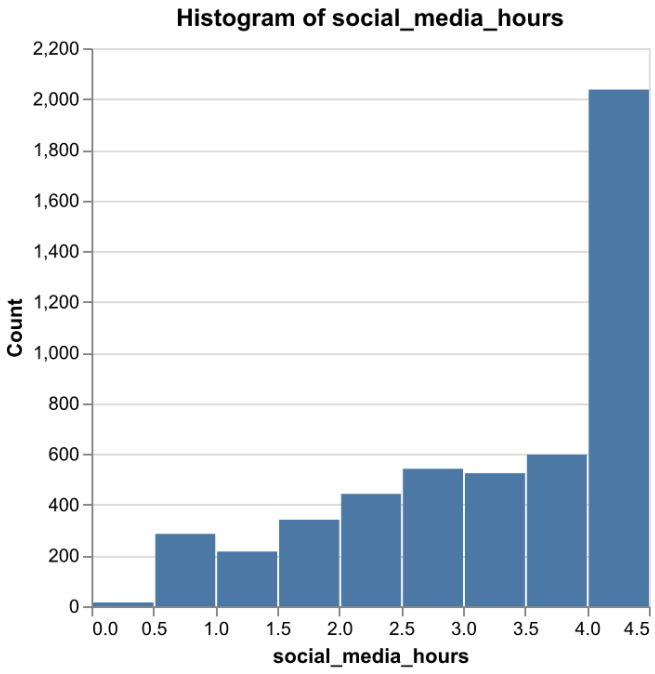
The dataset utilized in this study comprises 5,000 observations and is publicly available from Kaggle. It details device usage hours, usage types, and a range of wellness metrics, including sleep quality, mood, stress levels, mental health scores, healthy eating habits, caffeine intake, weekly anxiety, weekly depression, and mindfulness.

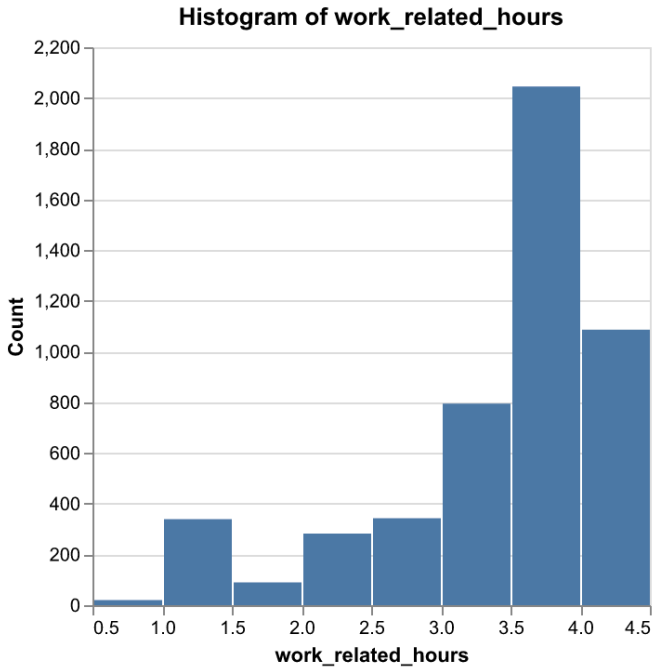


The dataset covers a good sample of age ranges.



An interesting pattern that we noticed is that the hours spent on social media and work related tasks are both negatively skewed.





IV. Approach

The methods for this experiment consists of four sequential stages designed to transform raw usage data into user segments. First, data preprocessing is performed to standardize numerical features and encode categorical variables. Second, dimensionality reduction is applied using Principal Component Analysis (PCA) to address the high-dimensional nature of the dataset. Third, Gaussian Mixture Models (GMM) are utilized to cluster users into distinct behavioral segments based on probabilistic assignments. Finally, a multi-label classification model is trained from the labels assigned from the cluster generated from the previous step.

A. Preprocesing

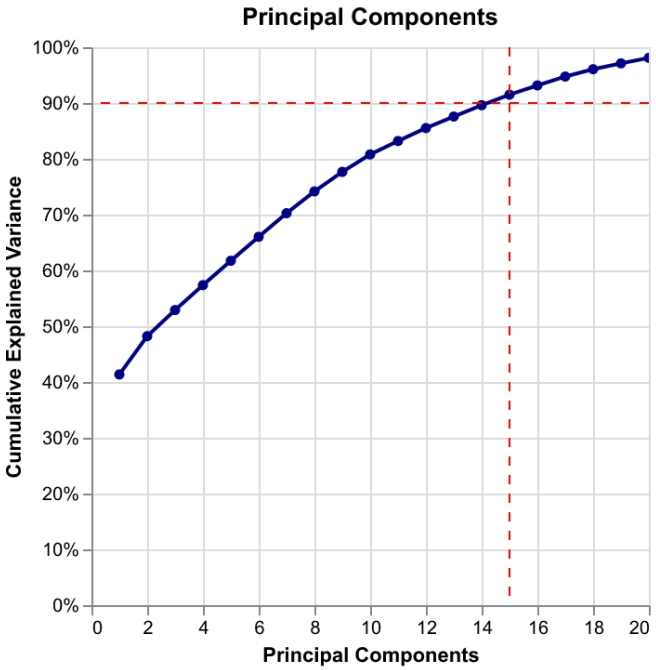
Dimensionality reduction techniques are sensitive to the magnitude of values, so numerical features are standardized to ensure each feature contributes equally to the analysis. This is done using sklearn's StandardScaler().

Categorical features, such as location type and gender, are encoded via one-hot encoding. This is done using sklearn's OneHotEncoder(). Location type and gender have very low cardinality.

a) *Dimensionality Reduction:*

Based on the methodology from [1] and the insights from [2] regarding GMMs and high dimensionality, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset prior to clustering.

To determine the optimal number of components to retain, the cumulative explained variance was analyzed for components ranging from 1 to 20, with the objective of retaining more than 90% of the variance.



According to our findings, 15 is the optimal number of principal components to maintain a variance of 90% for this dataset.

b) *Clustering:*

We propose an approach that combines clustering and classification for user segmentation. To capture overlapping user groups, we will be utilizing Gaussian Mixture Models (GMMs), which offer soft clustering as opposed to K-Means, which provides hard clusters. This is important because individuals often display multiple, overlapping behaviors in their technology use.

Gaussian Mixture Models assumes that data is a mix of several different normal distributions, also known as a Guassian. This clustering algorithm uses the Expectation-Maximization algorithm to assign data points to clusters probabilistically.

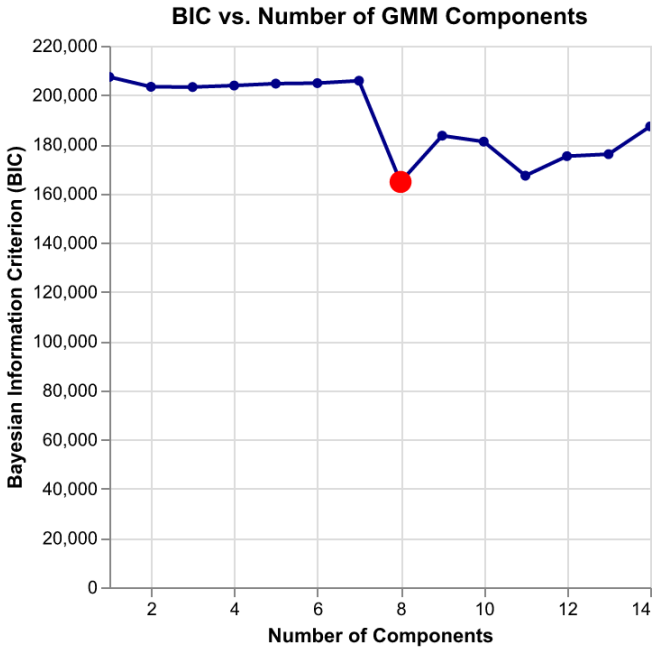
The expectation section takes a guess and decides the probability of an observed data point based on the following parameters:

- Mean (μ_k): Center of a cluster
- Covariance of components (Σ_k): Shape of the cluster
- Mixing weights (π_k): Prior probability that a random data point belongs to a cluster

The maximization step updates the above parameters to optimize the log-likelihood, which fits the cluster to the data as much as possible.

This process continues until the modifications of the maximization step no longer meets the convergence threshold.

Before running the clustering algorithm, the optimal number of Gaussian components was determined by evaluating a range of components and examining changes in the Bayesian Information Criterion (BIC), as demonstrated by [4]. The component count yielding the lowest BIC score was selected, as it represents the best fit to the data while minimizing the number of clusters to prevent overfitting.



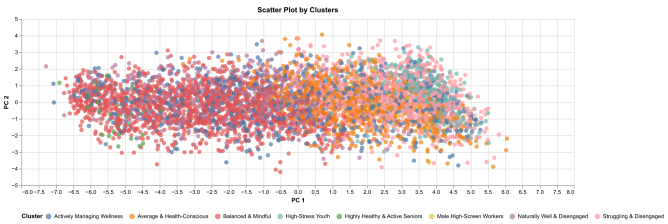
To ensure reproducibility, a fixed random_state was applied. Without this parameter, the probabilistic nature of GMM resulted in varying BIC scores across runs. Initial evaluations suggested 11 clusters; however, after setting random_state to 42, the model deterministically identified 8 clusters as having the lowest BIC score.

Once the clusters were formed based on their probabilities, labels were assigned to each group based on the observed characteristics shared by the data points.

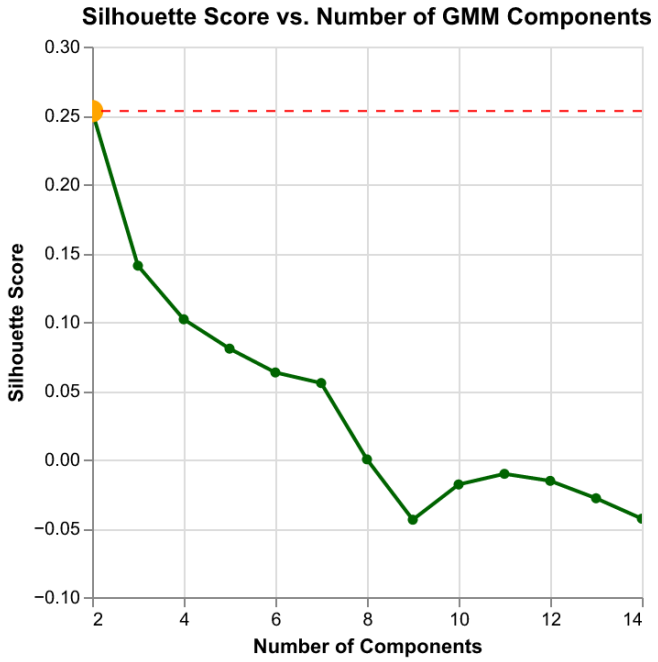
To expedite the data analysis process and the generation of cluster labels, a Large Language Model (LLM), specifically Gemini-3, was utilized. As [5] highlights, LLMs excel at providing meaningful summaries of complex datasets due to their ability to identify intricate relationships across multiple variables.

The following are the cluster labels that Gemini-3 produced:

- Struggling & Disengaged
- High-Stress Youth
- Male High-Screen Workers
- Actively Managing Wellness
- Highly Healthy & Active Seniors
- Average & Health-Conscious
- Balanced & Mindful
- Naturally Well & Disengaged



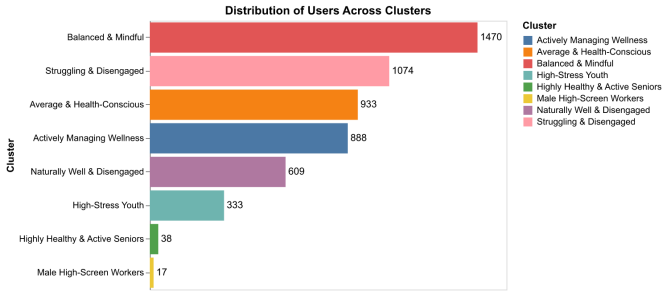
The Silhouette Score of the GMM cluster is 0.0000660399, which is negligible. This indicates that the clusters are significantly overlapping and not well-separated in the feature space.



Despite the low clustering metric score, we will continue with our proposed solution. Once probabilities are generated for each data point, one or more cluster labels are assigned based on a probability threshold.

Initially, the probability threshold was set to 40%, but this resulted in only 2% of the entire dataset having more than one label assigned. The threshold was iteratively decreased in increments of 10%, yielding the following distribution:

Label count	40%	30%	20%	10%
1 label	4,912	4,798	4,639	4,418
2 labels	88	202	360	580
3 labels	0	0	1	2



B. *Multi-label Classification*

Before we can proceed to train a model for multi-label classification, we first have to encode our labels via MultiLabelBinarizer which maps a list of labels into a binary matrix that will then be passed for training our multi-label logistic regression model.

sklearn's MultiOutputClassifier allows us to extend logistic regression to support multi-label classification. The above strategy fits a logistic regression classifier per target label.

C. *Results*

To evaluate the performance of the multi-label classifier, the Hamming loss between the test values and predictions was calculated. Hamming loss represents the fraction of labels that are incorrectly predicted.

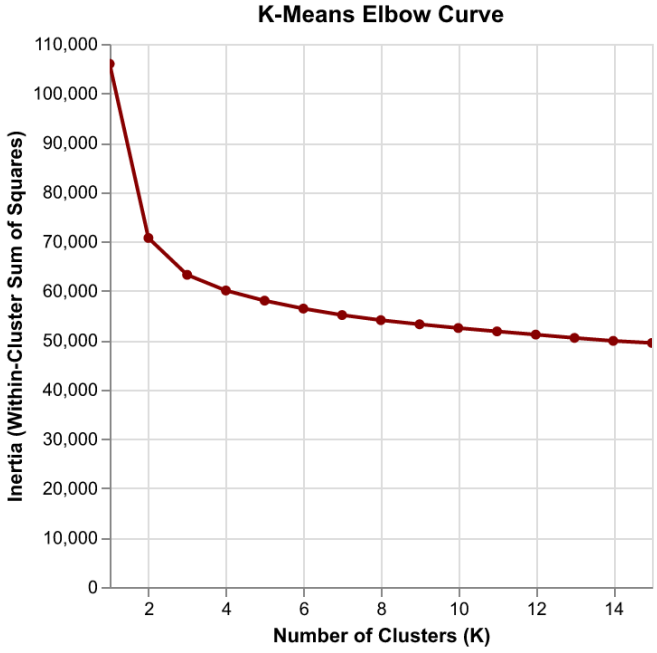
The model achieved an average Hamming loss of 0.013, indicating the rate of incorrect prediction of only at 1%.

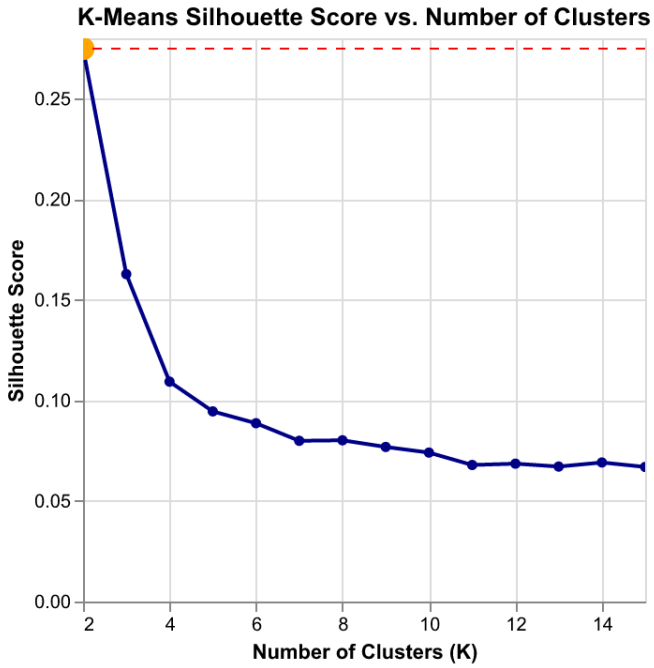
Even when evaluated against a strict metric for multi-label classification, such as subset accuracy, the model achieved a score of 90%.

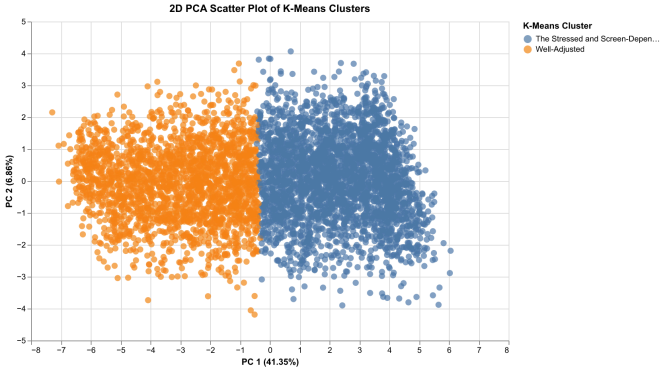
While these metrics suggest high performance, an analysis of the label distribution, where the majority of users belong to a single label, suggests the task effectively resembles multi-class classification rather than true multi-label classification.

D. *K-Means*

Since the majority of the datasets are only mapped with a single label, we were curious to see how a hard clustering algorithm such as K-Means would segment our data. Plotting the elbow curve shows 2 as a good number for clusters. The same applies for the silhoutte score. The silhoutte score is highest at 2 clusters.







Setting the number of cluster to 2 for sake of achieving a higher silhoutte score would be detrimental to our analysis and model. The only insight that this tells us is that higher screen time is unhealthy and reduces our classifier to a binary classifier with uninteresting predictions.

TABLE I
AVERAGE VALUES OF INTEREST BETWEEN THE TWO CLUSTERS

Feature	Cluster 0 (Balanced)	Cluster 1 (Stressed)
Cluster Size	2,143	2,857
Age	51.5	39.6
Mental Health Score (Higher is Better)	76.75	55.78
Mood Rating (Higher is Better)	7.08	2.47
Stress Level (Higher is Worse)	2.88	7.85
Weekly Anxiety Score (Higher is Worse)	4.36	11.84
Daily Screen Time (hours)	3.75	6.00
Social Media Hours	2.10	4.16
Physical Activity (hours/week)	4.53	1.26
Sleep Duration (hours)	7.53	7.25
Mindfulness (mins/ day)	24.33	14.22

V. Conclusion

This study demonstrates how unsupervised learning can uncover patterns between technology use and personal wellness. Using PCA and Gaussian Mixture Models, we identified eight overlapping user segments, each representing distinct digital and wellness profiles.

Although the multi-label classification model achieved strong performance (90% accuracy, 1% Hamming loss), it is not an indication that it's a strong multi-label classifier since the training data mainly consists of single labels. The majority of single-label assignments indicate that user behaviors are easily separated, and that its Silhouette Score may not be ideal for evaluating GMM clusters with the data that we have.

Overall, this framework offers a starting point for personalized wellness recommendations. Future work will focus on building interactive tools and refining clustering methods by experimenting with its hyper parameters to further fit the clusters with the data by comparing BIC scores or trying out a different soft clustering algorithm to better support healthier technology habits.

VI. Member responsibilities

As this was done by a single author, all of the work presented in the project was done by the sole author alone.

References

[1] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.

[2] L. S. Ling and C. T. Weiling, "Enhancing Segmentation: A Comparative Study of Clustering Methods," *IEEE Access*, vol. 13, no. , pp. 47418–47439, 2025, doi: 10.1109/ACCESS.2025.3550339.

[3] J. Lee and Ž. Žarnic, "The impact of digital technologies on well-being: Main insights from the literature," *OECD Papers on Well-being and Inequalities*, no. 29, 2024, doi: 10.1787/cb173652-en.
[4] V. Lavorini, "Gaussian mixture model clusterization: How to select the number of components (clusters)." [Online]. Available: https://medium.com/data-science/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4
[5] P. MJ Lindeman, "Using an LLM for Data Analysis: Your AI path to faster insights." [Online]. Available: https://www.quadratichq.com/blog/using-an-llm-for-data-analysis-your-ai-path-to-faster-insights