

CMSC 210: Assignment #1

Data Preprocessing & Unsupervised Learning

Data Preprocessing & Preliminary Data Mining

Data ingested from any source is inherently “dirty” for many reasons. According to Kim et al. (2003)¹, this is a major problem that is only starting to be recognized. The primary reason, however, why most (if not all) of the data are dirty is either due to human error, lack of standardization, or even technical issues.

In this assignment, our data was acquired by surveying Digital Health Enthusiasts on social media using Google Forms and has gathered a total of 57 various responses. But, while the survey was developed with a standard for data encoding in mind, encoding errors were still observed. Moreover, technical issues during dissemination and additional export columns when using Google Forms were inevitable. Hence, the survey data is currently dirty and needs pre-processing. Table 1 below shows a matrix of the data we’ve gathered that needs pre-processing, reasons for pre-processing, and actions on handling this before using it for actual analyses.

Table 1. Data pre-processing tasks.

No.	Data	Reasons for Pre-processing	Action on how to clean the data
1	Column names	Long column names	Rename column names for easier dataset management during analyses. The questions column can be codified to q1-q17.
2	Q5 - column L	Mislabeled entries	Replace values like “Not Applicable. (If you answered “No” in Q4)” with “Not Applicable”
3	Q12 - column S Q15 - column V	Misspelled values due to wrong set-up in the form	Replace values like “Note concerned” with “Not concerned” and “Consolation Convenience” with “Consultation Convenience”
4	Q7-10 columns N-Q	Anomalies in quantitative data	Identify and filter out data outliers (if necessary)
5	Q15-17 columns V-X	Multi-valued attributes in a column.	Transform comma-separated values into an array split by “,” remove unnecessary white spaces and NA values, and group similar/related values. This is done during Association Rule Mining Analysis.
6	Timestamp, Email Address, and Name columns, etc.	Unnecessary columns for analysis	Remove and retain columns that are only necessary for analysis. Technically, all columns will be considered but for general housekeeping columns A-C are primarily irrelevant.
7	All	Unsuitable data format upon reading the flat file as a data frame.	Identify and transform data types into a suitable file format

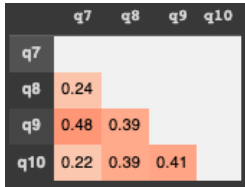
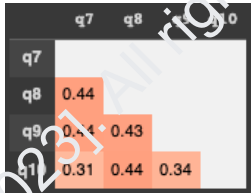

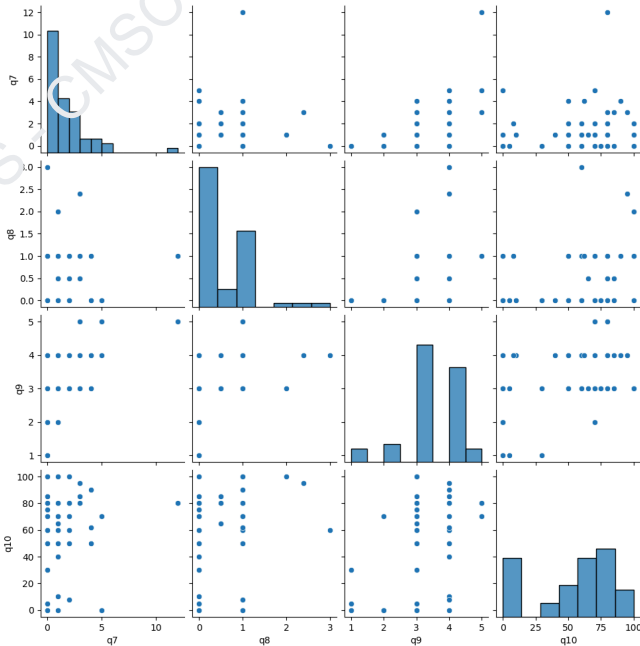
¹ Kim, W., Choi, B.J., Hong, E.K. *et al.* A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery* 7, 81–99 (2003). <https://doi.org/10.1023/A:1021564703268>

Please note however that duplicates and missing data for this dataset are presumed to be unlikely because the Google form was set up not to allow users to take the survey more than once, and all questions in the survey are required (except for the name, which is optional).

Correlation Analysis among Numeric Variables

Telemed Frequency (q7) x Avg. App Usage (q8) x Trust Rating (q9) x Digital Health Literacy (q10)

Table 2. Correlation analysis summary.

Method	Pearson (Parametric)	Spearman (Non-Parametric)
Correlation Coefficient Only		
P-value (Top) Correlation Coefficient (Bottom)		
Paired Comparison		

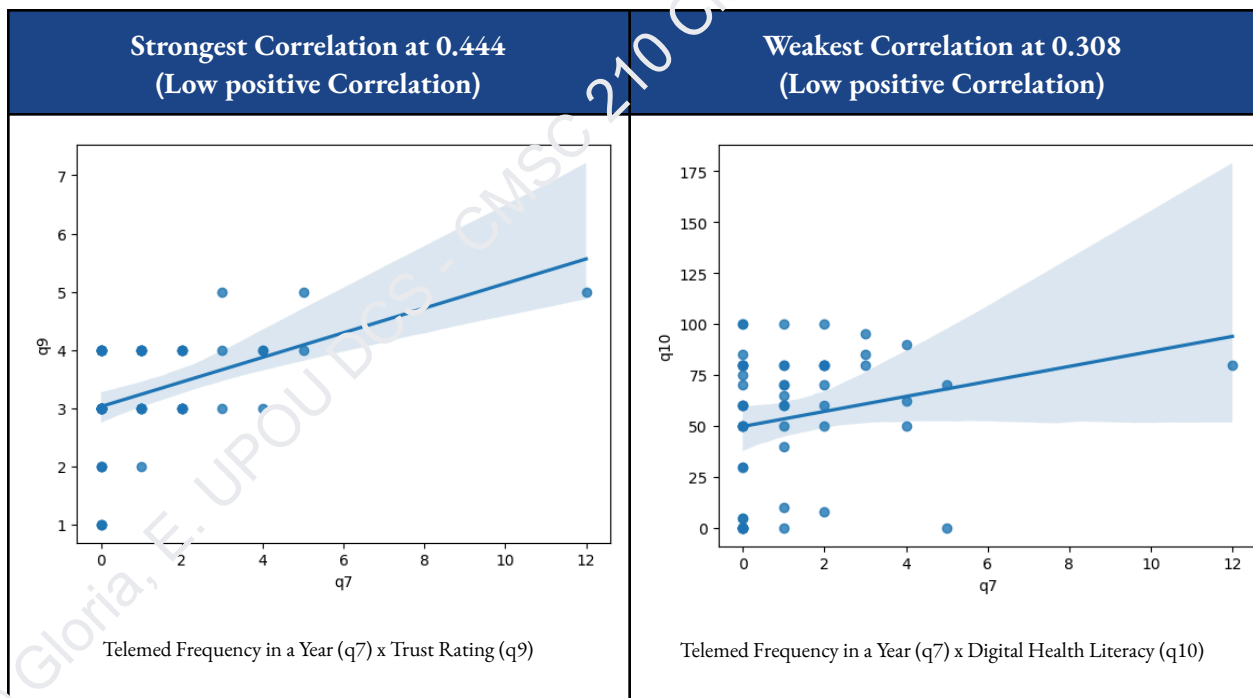
In this section, two Correlation methods, one parametric (Pearson) and a non-parametric (Spearman), were used to show the positive or negative relationship among the numerical variables (q7-q10 or Questions 7 - 10,

respectively) in the dataset. The results have shown that for the Pearson test, there are four significant paired comparisons: q9 VS q7, q9 VS q8, q10 VS q8, and q10 VS q9, whereas, for the non-parametric, all paired comparisons have shown statistical significance. Between these two methods, however, the results for the Pearson correlation coefficient are less likely to be accurate due to its inherent assumptions that the data should be 1) normally distributed, 2) linear, and 3) homoscedastic. The paired comparison chart shows that at least one of these assumptions has been violated. Thus, the Spearman results will be used in this analysis.

The Spearman correlation results showed the strongest correlation at 0.444 or a Low Positive Correlation between the respondents' frequency of telemedicine consults in a year versus the trust rating of Telemedicine. The weakest correlation, on the other hand, is at 0.308, or a Low Positive Correlation between the respondents' frequency of telemedicine consults in a year versus their digital health literacy.

Table 3. Paired comparison with the strongest and weakest correlation coefficient.

The results also showed a noticeable outlier on the x-axis. However, it was not omitted in the analysis as the



proponents believe that this is a natural part of the population being studied. This is a fraction of the Filipino people who are still considered as Digital Health Enthusiasts on social media but are more frequent in seeking healthcare services probably because of their current health condition (i.e., Chronic Disease Management,

Regular Health Screenings, Prescription Renewal, Family Planning, Rehabilitation, etc.), work affiliations, or even wellness and lifestyle guidance (i.e., Online workout, training, fitness & nutrition, etc.).

Discussion

Strongest Correlation (q7 VS q9): A correlation coefficient of 0.444 suggests a positive correlation between the two variables. This means that as the frequency of telemedicine usage in a year increases, the trust rating of telemedicine also tends to increase. This could suggest that increased use leads to trust. Similar to other services in the market, this makes sense because increased exposure to a service often leads to a better understanding of benefits and limitations, which leads to trust-building. Similarly, consumer experience matters as Telemedicine is a relatively new approach to seeking healthcare and as people become more accustomed to it through frequent usage. The positive correlation could also suggest a growing acceptance of Telemedicine as this becomes more integrated in our healthcare delivery system. This is also supported by the fact that more healthcare providers are offering such services with the support of various programs by the Philippine Department of Health (i.e., Policies on recognizing Telemedicine Providers since the COVID-19 pandemic, UHC Law) and Philippine Health Insurance Corporation (i.e., Policies on eHealth, FMP, integration) among others.

In contrast, however, it is essential to note that correlation does not imply causation. While there is a positive correlation, it does not mean that increased usage of telemedicine directly causes an increase in trust. Other factors, such as quality of care received, privacy, and prior experience with telemedicine, which were not captured in this survey, could also be considered. More importantly, the correlation suggests that there is some level of trust-building with increased usage. But, trust levels can still vary among individuals. For example, some people may still trust telemedicine significantly even with no or infrequent use, while others might need more exposure to develop.

Weakest Correlation (q7 VS q10): A correlation coefficient of 0.308 suggests a positive but relatively weak relationship between telemedicine frequency and digital health literacy. This means there is some degree of association between these two variables, but it is not robust or highly influential. We could say that, to some extent, an increase in the frequency of telemedicine usage in a year is associated with a slight improvement in digital health literacy. This may be because individuals who use telemedicine more often have more exposure to digital health tools and may gradually become more comfortable and skilled in using them. People who use

telemedicine frequently may have more opportunities for education and training, which can improve their digital health literacy.

Given that the correlation coefficient shows a relatively weak association between telemedicine frequency and digital health literacy, this could suggest that other factors influence digital health literacy beyond just usage. It should be noted that not all individuals who use telemedicine may experience the same improvement in digital health literacy. There are factors like age, education, and prior technology experience that could have more impact on this relationship. While digital health literacy is a multifaceted concept, it encompasses various skills and competencies related to, but not limited to, using digital tools and understanding of health information. More importantly, while telemedicine can improve literacy, it may only cover some aspects of digital health literacy, including privacy, evaluating the credibility of online health information, and managing personal health records.

Chi-Square Test between Telemedicine Use and Online Health Communications Membership

Figure 1. Chi-square test summary of results and decision rule.

```
Chi-square statistic: 0.020026350461132995
Critical value: 3.841458820694124
p-value: 0.8874635150406981
Significance level (alpha): 0.05
Degree of Freedom: 1

Decision with Chi Square Statistic >= Critical value:
Fail to reject H0. No sufficient evidence to say that there is a relationship between the 2 categorical variables.

Decision with P-value <= alpha:
Fail to reject H0. No sufficient evidence to say that there is a relationship between the 2 categorical variables.
```

Table 4. Chi-square contingency table between observed (vs expected).

Q4: Tried Telemedicine	Q6: Participate in Online Health Communities		Sum (Row)
	Yes	No	
Yes	22 (21.79)	5 (5.21)	27
No	24 (24.21)	6 (5.79)	30
Sum (Column)	46	11	57

Discussion

The results of the Chi-square test suggest that there is no sufficient evidence to say that there is a relationship between participating in online health communities or forums and having tried telemedicine for medical or

wellness consultations. This is because the chi-square statistic is close to zero, and the p-value is much higher than the significant level (α). This indicates that the two variables are likely independent of each other. Simply put, participation in online health communities does not appear to be associated with trying/using telemedicine. While these findings are based on the specific data and analysis, looking at other factors that are likely more influential in an individual's decision to adopt telemedicine for healthcare consultations is also recommended.

On a practical note, these results suggest that the respondents' decision to use telemedicine is not significantly influenced by whether or not they participate in online health communities or forums. Other factors, such as access to technology, personal preferences, and healthcare needs, likely play a more significant role in telemedicine adoption. It is also important to note that while this statistical analysis found no meaningful relationship, the real-world relationship between online health community participation and telemedicine use can be more complex. Individuals may use various sources of information and support, including online communities when making healthcare decisions.

Association Rule Mining for Reasons for Telehealth Use and Barriers to Digital Health Adoption

Reasons for Telehealth Use

Minimum Support = 0.2 | Minimum Confidence = 0.6

Table 5. Association rule mining summary result table for *Reasons for Telehealth use*.

Rule	Support Count	Support	Confidence	Rank
Avoiding In-Person Visit -> Time Saving	23	0.41	0.70	1
Consultation Convenience -> Time Saving	19	0.34	0.70	2
Avoiding In-Person Visit AND Consultation Convenience -> Time Saving	12	0.21	0.86	3

Discussion - Reasons for Telehealth Use

The results of association rule mining provided valuable insights into the relationship for telehealth use and the perceived benefits, particularly related to timesaving. These results highlight the significance of convenience and the perceived time-saving benefits of telehealth usage.

Avoiding In-Person Visit -> Time Saving: This rule indicates that a significant portion (41%) of respondents mentioned “Avoiding In-Person Visit” as a reason for using telehealth. Furthermore, there is a high confidence of 70% that when people use telehealth to avoid in-person visits, they believe it saves them time. In other words, there is a strong association between avoiding in-person visits and the perception of time savings. This suggests that many users use telehealth to save time they would otherwise spend traveling to a physical healthcare facility.

Consultation Convenience -> Time Saving: Similar to the first rule, this highlights the importance of convenience in telehealth usage. It indicates that 34% of telehealth users find “Consultation Convenience” to be a reason for using telehealth. With a confidence of 70%, it suggests that those who consider telehealth consultations convenient also believe it saves them time. The association between consultation convenience and time savings reinforces the idea that telehealth is valued for its ability to provide efficient and time-saving healthcare access.

Avoiding In-Person Visit AND Consultation Convenience -> Time Saving: This rule combines both “Avoiding In-Person Visit” and “Consultation Convenience” as reasons for using telehealth. It indicates that when users mention both these reasons, there is a strong association (confidence of 86%) with the belief that it saves them time. Simply put, avoiding in-person visits and finding convenient consultations strongly leads to the perception of time savings. This suggests that the perceived time savings are exceptionally high when both reasons align.

The strong association rule reveals that users value telehealth for its convenience and time-saving potential. These insights can guide service improvements (i.e., Efficiency and User Experience on telehealth platforms) and communication strategies (i.e., emphasizing comfort and time-saving benefits as a persuasive strategy) to better meet user expectations and needs in telehealth. Moreover, recognizing the importance of these factors, telehealth services can be customized to provide flexible scheduling and streamlined processes that align with users' expectations.

Barriers to Digital Health Adoption

Minimum Support = 0.05 | Minimum Confidence = 0.6

Table 6. Association rule mining summary result table for *Barriers to Digital Health Adoption*.

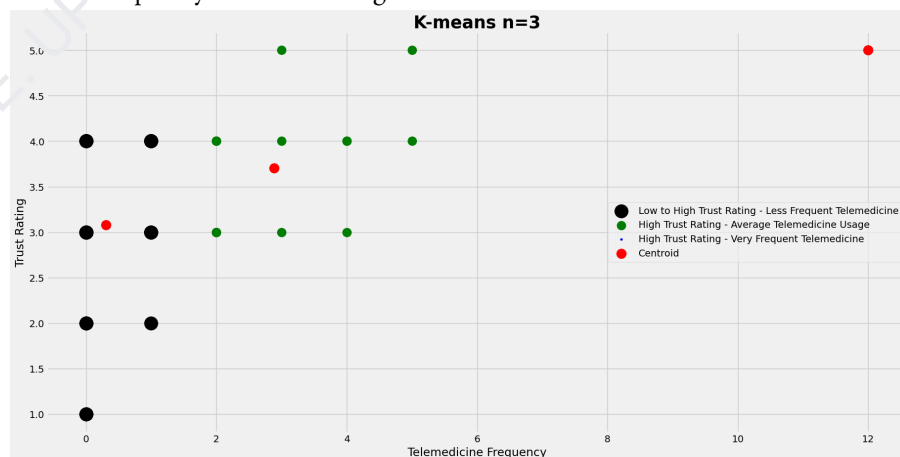
Rule	Support Count	Support	Confidence	Rank
Limited Access AND Privacy Concerns -> Limited Awareness	3	0.056	0.6	1
Limited Access AND Limited Awareness -> Privacy Concerns	3	0.056	0.6	2
Privacy Concerns AND Limited Awareness -> Limited Access	3	0.056	0.6	3

Barriers to Digital Health Adoption Discussion

The results of the association rule mining analysis provided insights into the barriers to digital health adoption and their interrelationships. The analysis identified the top three most vital association rules, each involving different permutations of barriers. However, the results show that all three rules have the same support and confidence value. Firstly, the support value is only at ~6% of the total transactions (54), which is pretty low. This value is also very close to the minimum support set. Second, while frequency is low, reliability for these three rules is high, with a confidence level of 60%. Lastly, we could only conclude that the results were highly influenced for the following reasons: limited variation or overlap in the data, limited data due to a small sample, and complex combinations are rare in the dataset.

K-Means Clustering Analysis of Telemed Frequency and Trust Rating

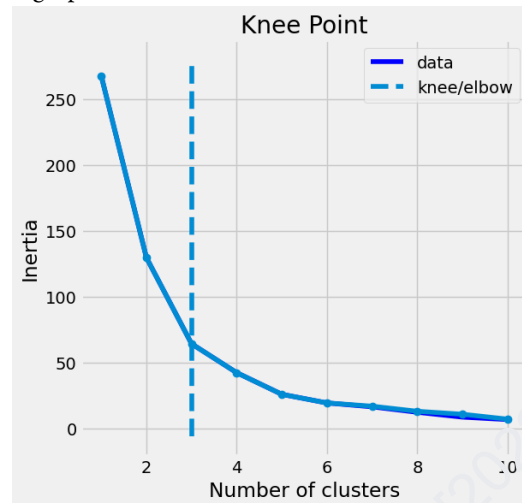
Figure 2. Telemedicine frequency and trust rating clusters at $k = 3$.



Discussion

Rationale for Selecting K

Figure 3. Knee and elbow method graph.



Selecting the appropriate value of k (the number of clusters) in K-means clustering is a critical step in the clustering process. In this case, we have used the knee and elbow methods, which yielded a k -value of 3 and an inertia (a.k.a. a within-cluster sum of squares or WCSS) of roughly 64.37, which provides reasonably compact clusters. The elbow point on the graph is the point at which the rate of decrease in inertia starts to slow down, indicating diminishing returns from increasing the number of clusters. Our results suggested that three clusters might be an appropriate choice. This is further complemented by using the knee method via the 'kneelocator()' function built-in in Python, which offers a data-driven approach to confirm the elbow method's findings. This helped in precisely analyzing the elbow point by looking at the curvature of the inertia curve. Therefore, the choice of $k=3$ is supported by both knee and elbow methods and the inertia value. Selecting this value is also reasonable and data-driven, balancing simplicity and meaningful clustering for further analysis or decision-making.

K-Means Cluster (K=3)

The results of the K-means clustering with $K=3$ and the associated descriptions of the clusters are as follows:

Low to High Trust Rating - Less Frequent Telemedicine Cluster in black: This is the largest cluster with 37 respondents and a centroid of (0.31, 3.08). It comprises respondents with a range of trust ratings described as "Low to High Trust Rating". However, these respondents are less frequent users of telemedicine.

This cluster suggests that individuals within it have varying degrees of trust in telemedicine but tend to use it less frequently. It may include those who are cautious or skeptical about telemedicine but are open to it. Others might also have adequate information and capacity to pay (both via out-of-pocket or financing - HMO/Insurance/Philhealth) for telemedicine consultations but due to other factors they might still prefer face-to-face consultations or other means of delivery. It is also good to note the significance of Filipino health seeking behavior in this cluster. Since it has been known that a lot of Filipinos has the tendency to delay preventive healthcare measure until the illness becomes evident² despite digital technologies becoming a staple in our daily living and access to various healthcare services over the internet.

High Trust Rating - Average Telemedicine Usage Cluster in green: This cluster has a total of 17 respondents and a centroid of (2.88, 3.71). Respondents in this cluster have high trust ratings and are described as “high trust ratings but average telemedicine users”. This suggests that they trust telemedicine and are more inclined to use it, but their usage frequency is moderate to typical. They are comfortable with telemedicine and may use it as a regular part of their healthcare, but not to an extreme degree. Such users would most likely be those who are comfortable with the convenience of seeking prescription, second opinion, medical certificates, or for common illnesses that are not life threatening.

High Trust Rating - Very Frequent Telemedicine Usage Cluster in blue: This cluster has only one respondent, which makes it an outlier. This fact has already been established on page 3 of this report. It has a centroid of (12, 5), indicating a high trust rating and widespread use of telemedicine. This individual is a high-trust user who is highly engaged with telemedicine, which makes them stand out from the rest.

These clustering descriptions provide valuable insights into the behavior and attitudes of the respondents toward telemedicine. For example, Low to High Trust Rating - Less Frequent Telemedicine Cluster in black may benefit from educational initiatives to increase telemedicine adoption while High Trust Rating - Average Telemedicine Usage Cluster in green may require further strategies to enhance their telemedicine experience. High Trust Rating - Very Frequent Telemedicine Usage Cluster in blue is an outlier and could be examined to understand the reasons behind their unique behavior.

² Allan B. De Guzman, Neil Angelo S. Ho & Mariz Dyan M. Indunan (2021) A choice experiment of the health-seeking behavior of a select group of Filipino nursing students, International Journal of Health Promotion and Education, 59:4, 198-211, DOI: [10.1080/14635240.2020.1730704](https://doi.org/10.1080/14635240.2020.1730704)