John Attisha
ECT 584
Final Project

Pew Research Health Tracking Survey

Abstract

Recent advancements in health technology are providing a wealth of benefits for society, whether an individual is in need of medical treatment or simply interested in maintaining a healthy lifestyle. One's demographic profile and general health conditions may play a role in how technology is being used as a health-related resource. In this report, statistical modeling techniques have been used to understand how these variables can be leveraged to predict certain actions and behaviors with regard to using technology to improve one's health. Association rules show that individuals with high blood pressure and who have recently been admitted to an emergency room in the last 12 months are highly likely to use some form of technology to monitor their weight, diet, or exercise regimen. Additionally, individuals who identify as caregivers of other adults with medical conditions or disabilities are highly likely to consult rankings and reviews of doctors online. Clustering techniques identified five main groups of health-technology users. It was found that while 80% of seniors track some type of health metric, only 2% are using a mobile app to do so. Additionally, younger individuals, most of whom live without any serious medical conditions, claim to frequently browse the web and use mobile applications for health-related purposes. As discussed in the full report, insights derived from the application of these statistical algorithms can be highly valuable for a variety of health-related domains, including biotechnology, medical training programs, research, and clinical practice.

Introduction

While the applicability of data science and predictive analytics seems to know no bounds these days, the healthcare industry is particularly well poised to extract and leverage data to improve the well-being of patients in need of treatment and society at large. Health-related websites, wearable technology, and medical imaging are just a few of the many sources of data shaping both behaviors and treatments when it comes to being healthy. As an example, fitness-tracking apps are helping individuals better monitor their exercise routines, check progress, and stay motivated when times get tough. Other advancements in health-related technology, however, do not always lead to an improvement in healthy behaviors. For example, ubiquitous access to the internet leads many individuals who are in need of medical treatment to turn to the web as their first choice for health-related information. In many cases this step is done prior to, or even in place of, seeing a medical professional face to face, and can potentially lead to adverse or harmful outcomes. Thus by being able to understand how certain types of individuals are likely to use technology for health-related purposes, proactive measures can be implemented to ensure better results.

The Pew Internet & American Life Project, a division of the Pew Research Center, recently surveyed over 3,000 individuals to collect demographic information, various health conditions they may have, and how they use technology for health-related purposes. Questions asked from this survey can serve to help understand how different types of demographic profiles and health conditions can be leveraged to predict one's behaviors and actions with regard to using technology and the web as a health-related resource.

A variety of statistical techniques can be applied to the data from this survey. In particular, for purposes of this project, association rules and clustering will be implemented. Association rules will be constructed in order to identify regularities among subgroups of individuals as it pertains to using technology for health reasons. One way this type of information could be helpful is at the clinical level. As an example, doctors and medical professionals may tailor their treatment plans to incorporate the use of technology for patients who are likely to perform certain expected actions online. Clustering techniques will also be performed on this dataset in order to find groups of similar individuals based on the three categories of questions asked in the survey. Analysis of these clusters will allow for the understanding of different general profiles and how they relate to using technology and the web to improve one's health. For example, clustering could be particularly helpful from a public health perspective, by dedicating resources to a grouping of individuals who are in need of tracking health conditions, but are not necessarily accustomed to the advancements in technology to do so.

Data Preparation

The original dataset consisted of 3,014 observations and 82 variables of statistical importance. Each observation represented an individual who completed the survey, and each variable represented a distinct question asked to the interviewee. Questions were asked over the telephone, and covered three broad categories: demographics; general health conditions; and health-related technology use. Out of the 82 variables (i.e. questions), the majority were poised such that the answer was either 1: Yes or 2: No. In order to more properly interpret certain statistical techniques, all “No” values were transformed to be represented by the number 0. By performing this transformation, taking the mean of these variables more intuitively identified the percentage of individuals who responded with “Yes” as an answer. Each question also provided for non-response answers, including "Don't Know" and "Refuse." These answers were typically represented by the integers 8 and 9, respectively. As there were only a small handful of non-response answers for each question, these values were transformed to NA values in the dataset.

Certain observations had to be fully excluded from our dataset. The variable “Intuse” referred to the question *“Do you use the internet, at least occasionally?”* to which respondents were asked to answer either Yes or No. By answering negatively, the majority of subsequent questions (particularly those regarding health-related internet use) were not applicable to the respondent, and thus were not asked of him or her – resulting in a majority of NA values for these observations. Unfortunately, these observations were not of much value to this project and thus were excluded when performing the statistical techniques outlined in this report. 705 individuals responded negatively to this question, resulting in 2,309 remaining eligible observations.

Data value types were treated differently with regard to the two statistical procedures being performed. For association rules, all values were transformed into factors, or categorical data. The variable “Age,” which allowed for a numerical response, was binned in order to properly meet the requirements of association rule data types. For clustering, values remained as integers, and were normalized prior to implementation. Further details are provided in the Clustering portion of the report.

Variables were grouped into three separate matrices: Demographics, General Health, and Health-Related Technology Use. Each row in the three data frames corresponded to the same respondent (i.e. row i in each of the three matrices represented answers from the same individual). A data dictionary identifying the variables associated with each matrix can be found in Appendix 1. A handful of variables from each of the matrices are included below. Each row identifies the variable name, specific question asked, and eligible responses.

Demographic Variables

Number	Variable	Question	Responses
1	Sex	No question asked. Automatically recorded by interviewer.	1: Male 2: Female
5	Mar	"Are you currently married, living with a partner, divorced, separated, widowed, or have you never been married?"	1: Married 2: Living with a partner 3: Divorced 4: Separated 5: Widowed 6: Never been married 7: Single
6	Par	"Are you the parent or guardian of any children under age 18 now living in your household?"	0: No 1: Yes
8	Emplnw	"Are you now employed full-time, part-time, retired, or are you not employed for pay?"	1: Employed full-time 2: Employed part-time 3: Retired 4: Not employed for pay 5: Have own business/ self-employed 6: Disabled 7: Student
9	Ins1a	Do you have "private health insurance offered through an employer or union?"	0: No 1: Yes
16	Inc	"Last year -- that is in 2011 -- what was your total family income from all sources, before taxes? Just stop me when I get to the right category..."	1: <$10,000 2: [$10,000 to $20,000) 3: [$20,000 to $30,000) 4: [$30,000 to $40,000) 5: [$40,000 to $50,000) 6: [$50,000 to $75,000) 7: [$75,000 to $100,000) 8: [$100,000 to $150,000) 9: >=$150,000

General Health Variables

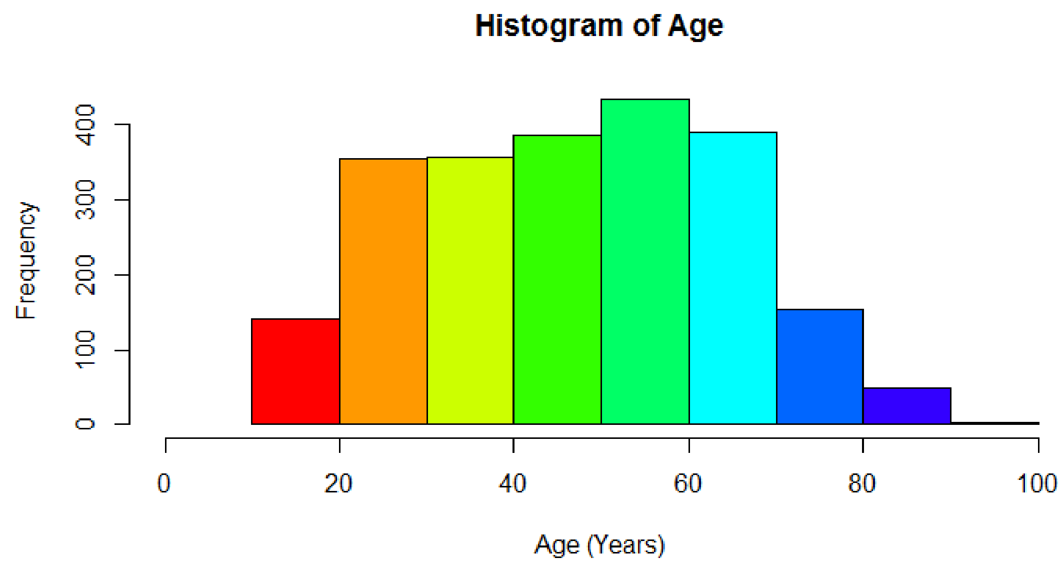
Number	Variable	Question	Responses
1	Q2	"In general, how would you rate your own health — excellent, good, fair, or poor?"	1: Excellent 2: Good 3: Fair 4: Poor
2	Q3a	"Are you now living with diabetes or sugar diabetes?"	0: No 1: Yes
4	Q3c	"Are you now living with asthma, bronchitis, emphysema, or other lung conditions?"	0: No 1: Yes
6	Q3e	"Are you now living with cancer?"	0: No 1: Yes
8	Q4a	"In the last 12 months, have you personally faced a serious medical emergency or crisis?"	0: No 1: Yes
15	Care6	"In the past 12 months, have you provided unpaid care to any child under the age of 18 because of a medical, behavioral, or other condition or disability?"	0: No 1: Yes

Health-Related Technology Variables

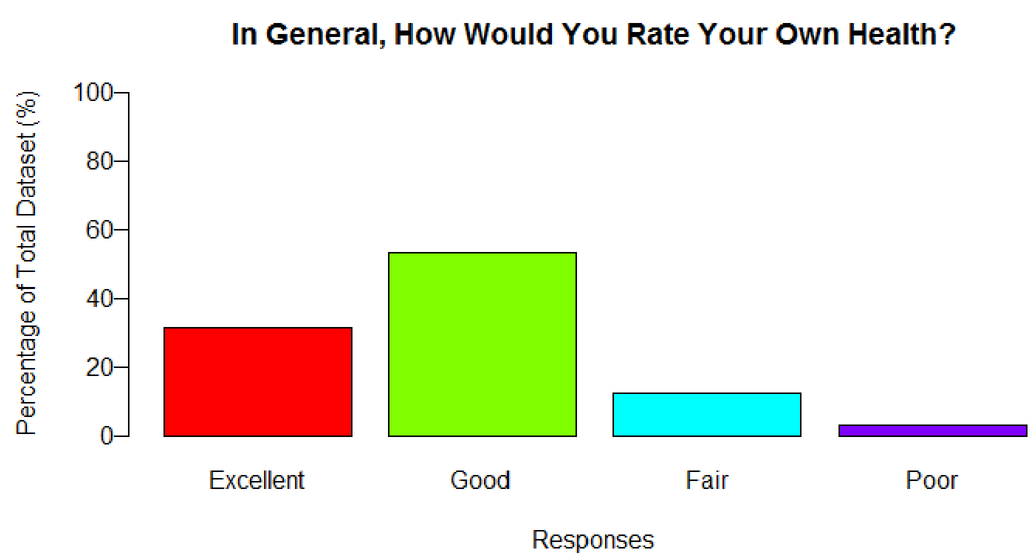
Number	Variable	Question	Responses
1	Q6a	"In the last 12 months, have you looked online for information about a specific disease or medical problem?"	0: No 1: Yes
15	Q10	"When looking for health information online, have you ever been asked to pay for access to something you wanted to see on the internet?"	0: No 1: Yes
25	Q17	"In the last 12 months, have you posted a health-related question online or shared your own personal health experience online in any way?"	0: No 1: Yes
28	Q20a	"Have you consulted online rankings or reviews of doctors or other providers?"	0: No 1: Yes
40	Q24	"Do you currently keep track of your own weight, diet, or exercise routine, or is this not something you currently do?"	0: No 1: Yes
43	Q26m1	"Thinking about the health indicator you pay the most attention to, either for yourself or someone else, how do you keep track of changes? Do you use…"	1: Paper, like a notebook or journal 2: A computer program, like a spreadsheet 3: A Website or other online tool 4: An app or other tool on your phone or mobile device 5: A medical device, like a glucose meter 6: Keep track just in your head 7: Other

Data Analysis

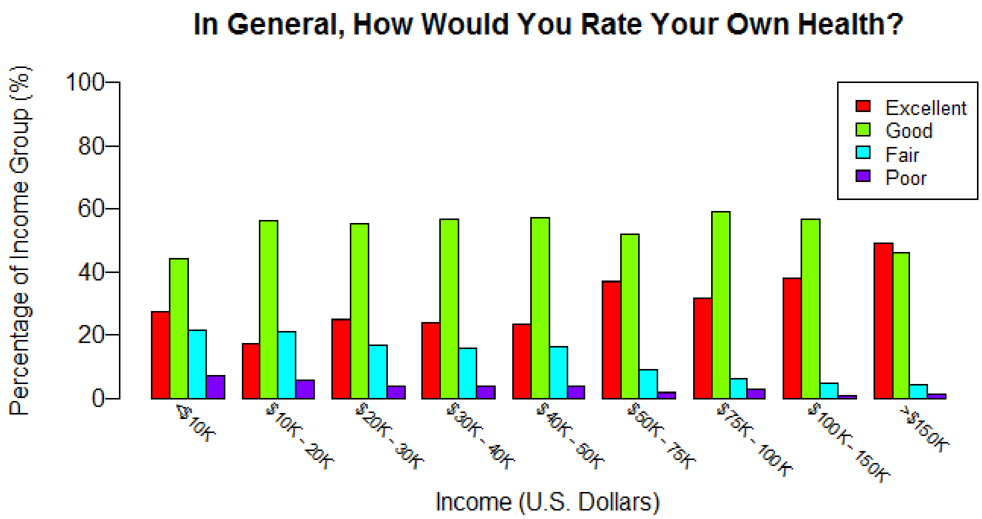
Out of the 2,309 eligible respondents in our analysis, 1,043 are male and 1,266 are female. Ages range from 18 to 92, with a mean of 47, median of 48, and standard deviation of 17.5. As the graph below illustrates, there is a fairly even distribution of ages from 20 to 70 years. We also see a fair number of respondents both below and above this range.



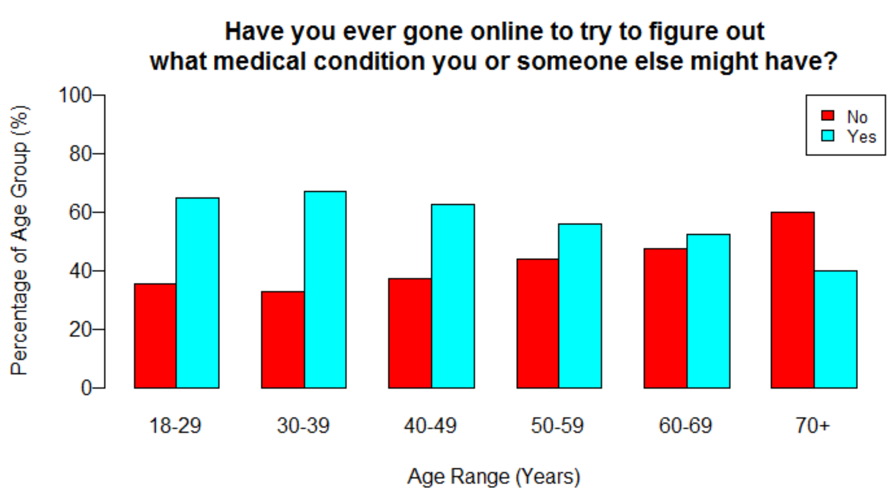
The chart below shows user responses to the question, *"In general, how would you rate your own health?"* Over half of the respondents answered with "Good", while 30% answered with "Excellent" and the remaining 20% answered with either "Fair" or "Poor."



The chart below depicts responses to the same question, broken down by income. “Good” seems to hover between 40% and 60% regardless of income level. As income increases, however, we see a noticeable increase in the percentage of “Excellent” responses. “Fair” and “Poor” both appear to decrease with an increase in income, accounting for only 5% of the 174 individuals who claim to make over $150,000.



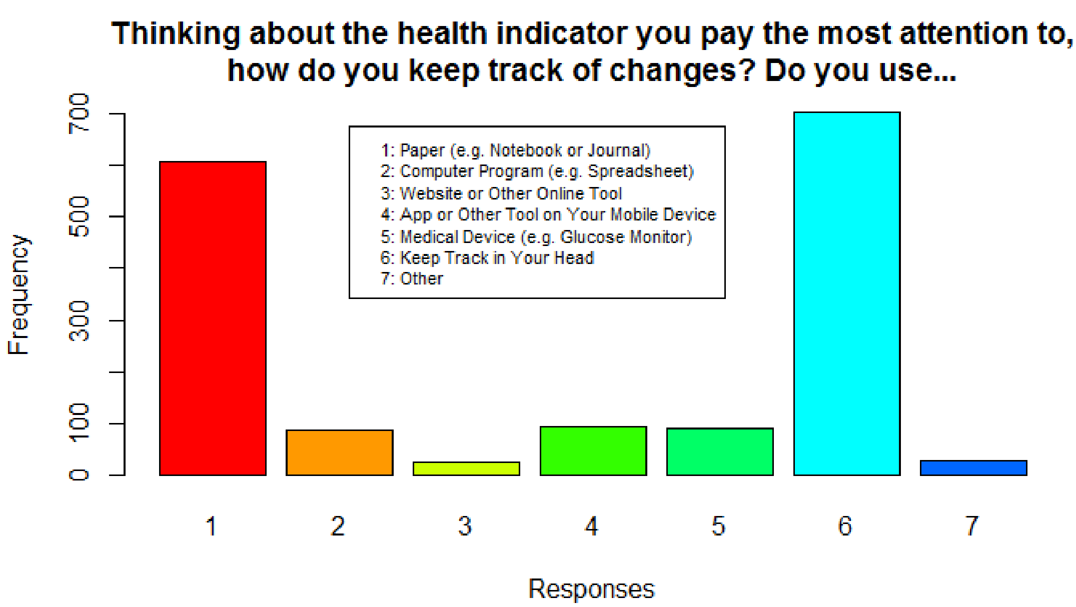
In response to the question, *“Have you ever gone online specifically to try to figure out what medical condition you or someone else might have?”* 990 individuals responded with “Yes” and 710 with “No.” Broken down by age, as seen below, we find that younger individuals are more likely to answer positively to this question, while older people are more likely to claim to have never gone online to research a medical condition.



With regard to tracking one's own health information, 65% of respondents say that they keep track of some type of personal health metric. Specifically, in response to the question *"Do you currently keep track of your own weight, diet, or exercise routine, or is this not something you currently do?"* we see the following responses:

"Do you currently keep track of your own weight, diet, or exercise routine?"	
Response	*Frequency*
Yes	1,496
No	811
NA	2

Out of the 1,496 positive responses, the majority of individuals who track health information do so either using paper (e.g. with a notebook or journal), or simply by using their head. The full range of responses can be seen in the chart below.



As health tracking software and mobile applications continue to develop and gain traction among individuals who use a smart phone, one could imagine that the responses to this question will evolve. Nevertheless, there could be a number of reasons why individuals are shying away from using the internet and mobile apps to track their health. Software may be cumbersome to use compared to other means, and for some individuals it may require a steep learning curve at the onset. Elder individuals who are required to track certain health metrics might not be familiar with recent advances in technology. Additionally, individuals may be concerned about the privacy of their data, especially due to the fact that health information can be very personal. Anxiety over potentially lost or stolen data may lead certain individuals to rely on more secure methods of tracking their health.

Association Rules

Development of association rules can allow us to understand what types of demographic profiles and health conditions may be useful in predicting specific types of health-related technology use. Therefore, in constructing these rules, the antecedent will consist of variables from two matrices: Demographics and General Health. The consequent of our rules will be limited to variables in our Health-Related Technology Use matrix. With almost 50 variables in our Health-Tech matrix, the tail of our association rules can vary widely. Therefore, we will focus on the following variables for the consequent of our rules:

- Q20a: *"Have you consulted online rankings or reviews of doctors or other providers?"*
- Q24: *"Do you currently keep track of your own weight, diet, or exercise routine, or is this not something you currently do?"*
- Q26m1: *"Thinking about the health indicator you pay the most attention to, either for yourself or someone else, how do you keep track of changes?"*

With regard to each of the questions above, rules will first be developed by focusing only on the Demographic variables for the antecedent. The same procedure will be performed using only the General Health variables. Finally, both matrices will be grouped together, in an attempt to uncover further notable rules not found by using each set of variables alone. Rules found to be interesting from each of the three procedures will be extracted and analyzed in the report below. The "arules" package in

R, specifically the "apriori" command, will be used to construct the association rules [1]. The full list of rules can be reviewed in Appendix 2.

After constructing rules with Question Q20a as the consequent (*"Have you consulted online rankings or reviews of doctors or other providers?")*, a number of noteworthy rules were identified. The table below includes three of these rules, each taken from a different set of starting antecedent variables.

Variables	Antecedent	Consequent	Support	Confidence	Lift
Demographic Only	Rule 2: Q1=3, Emplnw=1, Race=1, Educ2=8	Healthtech$q20a=1	0.010	0.511	2.84
General Health Only	Rule 37: Q3a=0, Q3c=0, Q4a=0, Q5c=3, Care2=1, Care9a=1, Care9b=1	Healthtech$q20a=1	0.010	0.581	3.23
Both Matrices Combined	Rule 53: Smph=1, Mar=1, Ins1a=1, Q3a=0, Care9a=1, Care9b=1	Healthtech$q20a=1	0.023	0.509	2.83

Using only demographic variables for our antecedent, we find the following rule: White individuals who are highly educated (i.e. have a Master's degree or higher), have full-time employment, and rate their quality of life as "Good," are likely to access online reviews of doctors. Interestingly, these types of individuals had two better options with which to rate the quality of their lives ("Excellent" and "Very Good"). This may indicate that these individuals are dealing with an issue in their lives, potentially medically-related. They likely have the means to decide which doctors they would like to see, and are using the internet as a resource to identify those physicians who have relatively high rankings. Out of 47 individuals who met the antecedent, 24 went to on meet the consequent of the rule. Note that the "apriori" command within the "arules" package in R calculates Lift as the confidence of the consequent with regard to the antecedent (as noted in the table above) divided by the confidence of the consequent with regard to the entire dataset [2].

Using only health variables to construct our association rules, we find that individuals who provide care to an adult relative to help them take care of themselves, and who also state that the internet has been helpful in their ability to cope with the stress of being a caregiver, are likely to access online reviews of doctors. These types of individuals are caring for adults with medical conditions or disabilities, a task which can certainly be stressful at times. Stating that they find the internet to be helpful in managing this stress indicates that they may often turn to the web to find helpful health-related resources, such as searching online reviews to ensure that their loved ones receive the best care possible.

Finally, using both sets of variables we find that individuals who are married, have private health insurance from an employer, and who feel the internet has been helpful to care for someone in need of medical attention as well as to cope with the stress of being a caregiver, are likely to use the internet to search for reviews and rankings of doctors. This association rule is somewhat related to the previous rule, however it also includes attributes describing the individual's relationship status and health insurance policies. By combining both sets of variables, we were able to identify a more descriptive group of individuals which are likely to search online reviews of doctors, at a higher support than the previous two rules.

With regard to Question Q24 as the consequent of our association rules (*"Do you currently keep track of your own weight, diet, or exercise routine, or is this not something you currently do?")* a number of notable rules were developed, three of which can be seen in the table below. The full list of rules can be found in Appendix 3.

Variables	Antecedent	Consequent	Support	Confidence	Lift
Demographic Only	Rule 38: Mar=1, Par=0, Hh1=2, Race=1, AgeBin=60-69	Healthtech$q24=1	0.052	0.753	1.16
General Health Only	Rule 181: Q3b=1, Q3e=0, Q4b=1, Q5a=2, Care6=0	Healthtech$q24=1	0.029	0.791	1.22
Both Matrices Combined	Rule 16: Hh1=2, Ins1b=1, Hisp=0, Race=1, Q3d=0, Q4c=0	Healthtech$q24=1	0.060	0.822	1.27

Using only demographic variables, we see that individuals who are white, married, have a total of two adults in their household and no children under 18 at home, and are between 60 and 69 years of age are likely to keep track of their weight, diet, or exercise routine. This could be due to the fact that these individuals are reaching an age where it is imperative for them to track certain health metrics due to medical conditions or simply for their own well-being. Additionally, these individuals may be retired and have a certain amount of free time on their hands, allowing them to track information which could simply be beneficial for them to know. Out of the 158 individuals who met the antecedent of this rule, 119 also met the consequent, resulting in a confidence of 0.753.

Using only health variables, we see that individuals who are living with high blood pressure, and have either gone to the ER or been hospitalized unexpectedly, are highly likely to keep track of their weight, diet, or exercise regimen. Keeping track of these health metrics are certainly beneficial for individuals with this type of health condition, particularly if they are prone to being hospitalized. By ensuring these individuals utilize a modern technology health-tracking device, medical professionals would also benefit by having objective, accurate logs of their patients' health.

Finally, using both demographic and health variables, we see that individuals who are white and have purchased private health insurance themselves, yet do not have any heart conditions nor have had any significant change in their health in the last 12 months, are highly likely to keep track of exercise, weight, or diet routines. Out of the 169 individuals meeting the antecedent of this rule, 139 also met the consequent, resulting in a confidence of 0.822. Perhaps this type of individual is simply choosing to be proactive about his or her health. Alternatively, tracking these metrics may have motivated certain individuals to improve their lifestyle, resulting in better health measurements. One additional explanation may be that this type of person could have an incentive from his or her health insurance provider to track certain health metrics in exchange for a subsidy or decrease in insurance premiums.

Regarding Question Q26m1 as the consequent of our association rules (*“Thinking about the health indicator you pay the most attention to, either for yourself or someone else, how do you keep track of changes?”*), we also see a number of noteworthy associations. Recall from the Data Analysis section of this report that the majority of answers were either “1: Paper (e.g. a notebook or journal)” or “6: Keep track just in your head.” In terms of constructing association rules, all other possible answers would require a very low minimum support threshold. In the rules below, the consequent will be restricted to answer “1: Paper.” As before, the table below highlights three rules, each of which are taken from a different set of starting antecedent variables. The full list of rules can be found in Appendix 4.

Variables	Antecedent	Consequent	Support	Confidence	Lift
Demographic Only	Rule 18: Sex=2, Ins1c=0, Ins1d=1, Race=1, Educ2=3	Healthtech$q26m1=1	0.011	0.634	2.41
General Health Only	Rule 2: Q3a=1, Q3b=1, Q5a=2, Care9b=2	Healthtech$q26m1=1	0.011	0.758	2.88
Both Matrices Combined	Rule 55: Sex=2, Par=0, Q5a=2, Care2=1, Care7=1	Healthtech$q26m1=1	0.022	0.607	2.31

Focusing on demographic variables, we see that individuals who are women, white, have Medicare, and whose education level is classified as “some college” tend to keep track of health indicators on paper, e.g. by using a notepad or journal. As these individuals have Medicare as their form of health insurance, the fact that they are relatively elderly compared to other age groups in the dataset may indicate why paper is the preferred tracking device for health metrics. As discussed before, these individuals may not be familiar with technological advances in health tracking applications, let alone adept at using them. Similarly, simply tracking health metrics in their head (the other most popular answer to this question), may not suffice as an accurate long term option.

Using only health variables, we see that individuals who have both diabetes and high blood pressure, and who tend to get health care information from their doctor in person, are likely to use something like a journal or notebook to keep track of health metrics. It may be the case that software and applications for tracking health conditions like diabetes and high blood pressure are simply not well-developed or easily understandable to be widely accepted by the majority of individuals with these medical ailments.

Finally, combining both the demographic and health variables, we see that women who are taking care of an adult at home, as well as managing their medications, and do not have children in their household, are likely to use some type of journal or notepad in order to track various health metrics. This may fit the profile of an older female individual taking care of her spouse, whose children have left the household by this point in their lives. We see that this person is also managing her spouse's medication, indicating the need to track this information to ensure proper adherence to a medication schedule. This individual may feel apprehensive or uneasy about using a technological device to track medication schedules, for fear of lost or stolen health data. Additionally, tracking this information in one's head may not be a sufficiently accurate option.

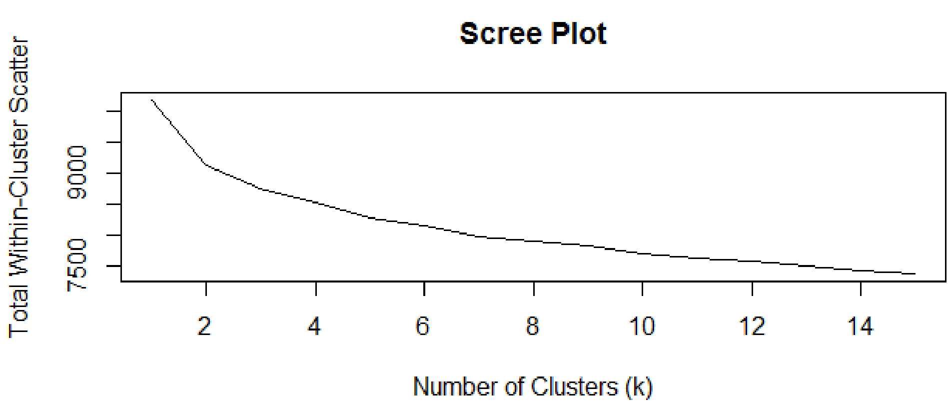
Clustering

Recall that the goal of clustering is to develop groups of objects with high intra-class similarity and low inter-class similarity. Clustering techniques can be performed on our dataset to discover collections of individuals who are similar based on Demographic, General Health, and Health Technology variables. Using domain expertise, these clusters can be inspected to gain a broader understanding of the general distinctions among different groups of health-tech users. Additionally, clusters can serve as input for various supervised prediction or classification models.

A variety of partitioning (k-means, k-medoids) and hierarchical (agglomerative) algorithms were explored to determine the best groupings of clusters for the dataset. While each approach has strengths and weaknesses, k-means typically resulted in a more comprehensible group of clusters. K-means clustering was performed on three differently combined sets of data: Demographic and Health Technology variables only; General Health and Health Technology variables only; and all three sets of variables combined together. Each of the three implementations provided distinctive insightful findings. For purposes of this project, results from performing k-means clustering on the entire dataset (i.e. all three matrices combined) will be analyzed and interpreted below.

As mentioned in the Data Preparation section of this report, the dataset was pre-processed such that with regard to the Yes/No questions, No = 0 and Yes = 1. Therefore, after running the k-means algorithm, the cluster means for these questions represented the percentage of individuals in a particular cluster who responded with "Yes" as an answer. Other variables not in the [0,1] range were normalized using decimal scaling. This form of normalization allowed for a more intuitive understanding of the cluster means for these variables, since the original categorical levels (i.e. answers) were most often represented by integers. Observations containing NA values were eliminated from the dataset prior to running the k-means algorithm. Variables with a high percentage of NA values were eliminated altogether.

By examining the total within-class scatter for various levels of k, one can better determine the appropriate number of clusters to use for a dataset. The scree plot below shows a relative leveling off of the within-class scatter at around five clusters [3]. After exploring the resulting clusters for various levels of k, and running each algorithm multiple times to obtain to best possible groupings, it was determined that the most appropriate number of clusters for the dataset was indeed five.



Observations were fairly evenly distributed among the five clusters. Three clusters described below identify notable profiles of individuals as portrayed by their demographics, health conditions, and health-related technology use. Appendix 5 contains full technical details, including cluster means, for each grouping of individuals modeled over variables in the dataset.

Cluster 1 appears to identify the oldest grouping of individuals in our data, with an average age of 65. They are mostly retired, and 75% have Medicare as a form of health insurance. 63% of this cluster lives with high blood pressure, and 25% live with diabetes. Two out of three individuals in this cluster have looked online about a specific disease, and almost half have looked online for a specific medical treatment. While 80% keep track of their weight, calories, or exercise regimen, only 2% of individuals in this cluster claim to use mobile apps to track these health metrics! Only 28% claim to use the internet at all on mobile devices. It could be the case that these individuals are not familiar with the variety of health tracking apps available to them. Alternatively, this cluster may be familiar with health tracking apps, but are not adequately knowledgeable of how to use them, as most of these people likely learned how to use internet-connected devices at a late age. Doctors and medical professionals might consider spending time recommending that these individuals learn some of the more advantageous health tracking applications. After learning how the apps are properly used, individuals may find these options to be relatively easier than the more traditional options of tracking one's health. This would also benefit medical professionals by having a more accurate log of their patients' health metrics.

Cluster 2 appears to identify the youngest group of individuals out of our five clusters, with an average age of 33. These individuals are typically working either full or part-time, and have health insurance either through an employer or by means of purchasing it themselves. They typically rate their health as being either "Excellent" or "Good." Only 15% claim to be living with high blood pressure, a stark contrast to the 63% seen in Cluster 1. Nevertheless, when asked if they have gone online to research a specific medical problem, 78% responded in the affirmative. 25% have also looked online for information about pregnancy and childbirth, the highest percentage of all five clusters. 60% claim to track their weight, diet, or exercise routine, and 22% have mobile apps that help manage their health. These individuals appear to be the healthiest of our five clusters, but still seem to be open to the idea of using the internet and mobile applications for health-related purposes. People in this cluster are likely highly adept at navigating the web, and keep track of their health metrics not because they have to, but because it's a relatively trivial task.

In Cluster 5 we see individuals who have an average age of 43, are typically employed, highly educated, make a relatively high income, and have health insurance through their employer. 11% claim to have diabetes and 21% have high blood pressure. As expected, both of these values fall somewhere between the values seen in our previous two clusters. 80% track either their weight, diet, or exercise, and 28% have health-related apps on their phone. Over 70% claim to have consulted online reviews of drugs or medical treatments, noticeably higher than both Cluster 1 (15%) and Cluster 2 (11%). Here we see individuals who are not only adept at using the internet, but are also reaching a point where certain medical conditions are potentially becoming more prevalent in their lives. If there was ever a target market to whom a tech-company should advertise its health-related technology, this is it. Out of all five clusters, these individuals are some of the most active users of technology with regard to health-related purposes.

Executive Summary and Conclusion

Implementation of statistical algorithms, including k-means clustering and apriori association rules, have allowed for the identification of distinct subgroups of individuals and how they behave with regard to using technology for health purposes. Association rules were developed to identify who would be likely to (1) consult rankings and reviews of doctors; (2) keep track of weight, diet, or exercise routines; and (3) keep tracks of these health metrics via a notepad or journal as opposed to other means.

While not exclusive to this subgroup, individuals who identify as caregivers of adults with medical conditions are fairly likely to consult reviews of doctors and medical professionals online. Regarding the tracking of health metrics, we found that individuals with certain medical conditions including diabetes and high blood pressure are likely to keep track of their weight, diet, or exercise regimen. We also found a contrasting subgroup of individuals with little to no medical conditions who are also likely to track these health metrics. Further study would be needed to identify if these individuals are perhaps simply being proactive about their health, or if the actual act of tracking one's health metrics leads to a healthier lifestyle. Finally, we found that elder individuals are likely to keep track of these health metrics in a notepad or journal, as opposed to other means. This could be due to a lack of knowledge about health tracking software, or out of concern for keeping this information private.

By performing k-means clustering on our dataset, we were able to identify a number of distinguishable groups of individuals based on their demographic, general health, and health- technology attributes. Cluster 1 identified a collection of older individuals, many of which live with one or more medical conditions. While 80% of this cluster keeps track of certain health metrics, only 2% choose to use a mobile app to do so. Cluster 2 identified the youngest of the five groups, most of whom live without any major medical conditions. Nevertheless, these individuals frequently browse the web and use mobile applications for health-related purposes. Perhaps they are comfortable enough with using technology that performing these tasks requires little effort compared to other clusters of individuals. Finally, Cluster 5 identified a group of middle-aged individuals. A fair portion of this cluster (though not has high as Cluster 1) appears to live with one or more medical conditions. These individuals are most likely to use the internet and mobile applications for health related purposes. This could be due to the fact that they are not only adept at using most forms of technology, but are also approaching a stage in life where caring for one's health becomes more of a necessity.

This information could be highly valuable for a number of health-related sectors. The health-tech industry could benefit by developing and marketing medical tracking devices for individuals who are in need of monitoring certain health metrics, but are currently choosing to do so via more traditional means. Medical researchers could mine the massive amount of data that is a byproduct of the widespread acceptance and use of health tracking software and applications. Doctors and medical staff could alter treatment plans by leveraging the increased knowledge extracted from their patients' health tracking technology. Most importantly, the public can benefit by understanding their own health conditions more accurately, working with their medical providers to formulate precise and individualized lifestyle plans so as to maximize one's health for the long-term future.

References:

[1] Zhao, Yanchang. Association Rules. Rdatamining.com: R and Data Mining. 2015.
http://www.rdatamining.com/examples/association-rules

[2] Borgelt, Christian. Lift Value. Apriori – Find Frequent Item Sets and Association Rules with the Apriori Algorithm. 2010.
http://www.borgelt.net/doc/apriori/apriori.html#lift

[3] Kabacoff, Robert I. Cluster Analysis. Quick-R – Accessing the Power of R. 2014.
http://www.statmethods.net/advstats/cluster.html