Northeastern University

DS5110

Introduction to Data management and Processing

Analysis of NHIS(National Health Interview survey) data and building associated software components

Ankitha Kumari Moodukudru
Deepanshu Parihar
Harish Ramani
Viral Pandey

Contents

1 Summary

The National Health Interview Survey (NHIS) is a yearly cross-sectional household interview survey conducted by the Centers for Disease Control and Prevention's National Center for Health Statistics. Data is collected from a selected sample of households in US by conducting face to face interviews. The survey contains data related to demographics, health status and limitations, injuries, health care access and utilization, health insurance, income and assets. The data set contains responses from seven questionnaires corresponding to household, family, child, injury and disability data. We wanted to explore the following aspects of the data set:

- Compare the mental health of individuals who practice meditation and other relaxation techniques
- Identify factors affecting sleep quality among the participants
- The Affordability of health care and deficiency aiding devices
- Consider the scenario where a business owner wants to explore the data but he/she relies on people with programming skills to help answer important questions. A more intuitive approach would be for the owner to ask what he/she wants as a text query to a bot and get all the visualizations. In this project, we have exposed the NHIS dataset via a conversational bot and suggest frameworks by which this could be generalized for any dataset.

We analyzed the data from adults of the age 18 years and above who participated in the NHIS survey of 2017.

2 Methods

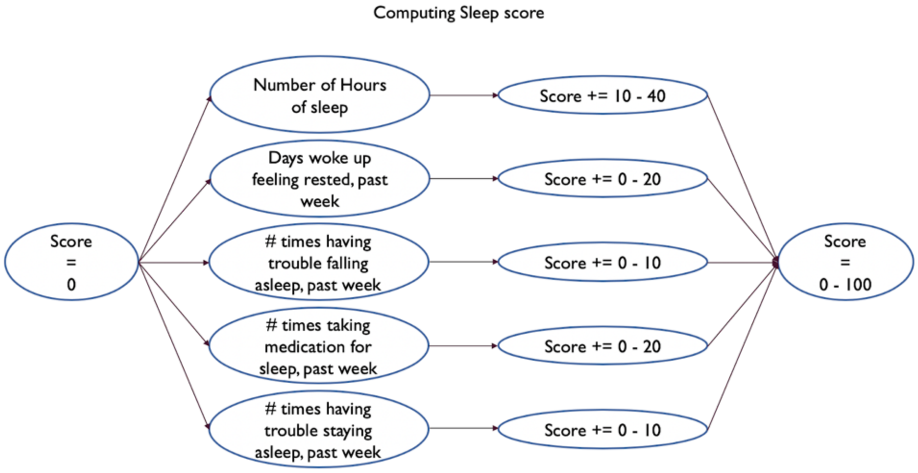
2.1 Analysis of effect of yoga and meditation on emotional well-being:

The survey consisted of user responses that were marked uncertain, not mentioned and refused. These observations were filtered out for the selected columns as they did not favor the analysis. Participants‘ self-reported feelings were captured using a set of 6 questions, each one corresponding to feelings such as sad, nervous, fidgety, worthless, everything required effort and hopeless. Each of these variables contained 5 levels representing the frequency of occurrences of these feelings. Participants' response to practice of yoga or meditation was scattered across multiple columns, each one corresponding to a specific technique of meditation, yoga, tai chi etc.

The variables for yoga and meditation were merged into 2 variables by performing a logical OR on the responses for individual practices. In order to capture the overall psychological distress level of a participant the variables related to feelings were merged into a negative emotion score, that ranged from 0 to 1. This variable was used to visualize the distribution of distress levels in the yoga and meditation practicing groups. We also analyzed the frequency of experiencing these feelings with yoga and meditation by employing the chi-square test of independence with a significance level set at P < 0.05.

2.2 Analysis of quality of sleep and the contributing factors:

The survey consisted of responses on the self reported hours of sleep and other associated attributes such as sleep medication. We collated these columns into one sleep score so that I can compare it with other factors. We weighed all these variables differently and came up sleep score as follows:


Computing Sleep score
Score = 0
Number of Hours of sleep
Days woke up feeling rested, past week
# times having trouble falling asleep, past week
# times taking medication for sleep, past week
# times having trouble staying asleep, past week
Score += 10 - 40
Score += 0 - 20
Score += 0 - 10
Score += 0 - 20
Score += 0 - 10
Score = 0 - 100

Then, based on the initial analysis we chose 60 variables out of the available variables. We plotted them with the sleep score that we devised to check for any linear patterns. We got 23 such variables which showed linear relationship with sleep score. After fitting the linear model with them and plotting the residuals of it with the variables, We chose these variables as predictor variables:

- PHSTAT: How would you describe your health in general?
- ASIMEDC: How worried are you right now about not being able to pay medical costs of a serious illness or accident?
- ASICNHC: How worried are you right now about not being able to pay medical costs for normal health care?

We did cross k fold validation with 10 partitions to check the accuracy of the model.

2.3 Analysis of affordability of health care

Based on the survey responses, we came across certain measures to detect affordability of dental care and eyeglasses. So we defined a new feature called affordability which could represent the affordability of the two responses as one . This feature has values 0,1 or 2 based on the person's ability to afford neither , one or both of eyeglasses and dental care. Then, we narrowed the number of factors affecting affordability by going through the code book and filtering out columns related to vision and dental care for building the model. As the classes we were trying to predict were more than two, we tried a multinomial logistic regression algorithm to fit a model based on Age-Group , medical bills and relative coverage compared to a year ago. The challenge we faced was of data imbalance [Figure 10] as observations for one class was dominating the other two after omitting all rows with NA values in them. This was solved using the upsample function of caret package which replicates previous observations of minority class to build new values to obtain data balance.

2.4 Conversational Bot

A conversational bot was developed using slack as the interface and dialogflow for processing the texts entered by the user as represented in [Figure 1] The components of the bot are as follows:

- Slack – It is a popular cloud based set of propriety team collaboration tools and services. We use it to create a bot through which we can deliver all the needs of the business owner to visualize and interpret his data by asking questions in natural language.
- Dialogflow – It is a google owned technology that gives a framework for processing natural language. Important terminologies in dialogflow :
  - Entity: Keywords that are specific to a particular application. In our bot, all the names of the columns of the data are specific to our application and they can be treated as entities.
  - Intent: They are basically to match what the user wants. We can create intents by giving some training examples and if the user input matches the common underlying pattern, the intent is triggered. It can be used to parse user texts in a structured manner and get the desired entities.
  - Fulfillment: A separate service or a url could be triggered once an intent is matched. It is called a fulfillment service. In our application, we write our own service and configure the url in dialogflow.
- Application Layer – It is a node server which is the backend for the entire application. It handles requests from slack, and calls dialogflow and parses the response for it to be viewed in slack.
- Data Layer: It is where the R scripts reside that produces the visualization . The visualizations are then put to an S3 bucket and the link is returned back to the user. All

the scripts are exposed as an API (Application Programmable Interface) via Plumber package. All the cleaned data from the analysis are exported in a R package that was created especially for this purpose.

- Workflow: [Figure 2]

3 Results

3.1 Analysis of effect of yoga and meditation on emotional well-being:

- The negative emotion score was plotted against yoga and meditation separately to analyze the distribution of frequencies of emotion scores in each of these categories. It was observed that more participants had higher psychological distress level in the non-yoga and non-meditation participating groups [Figure 3] and [Figure 4]
- 24 % of the US population experience psychological distress of some level. And most of them reported feeling nervous and fidgety more often.
- The chi square test between the frequency of feelings reported vs meditation and yoga (separate), returned a very low p-value, confirming that they are related. [Figure 5] and [Figure 6]

3.2 Analysis of quality of sleep and the contributing factors:

Upon initial exploration of variables present in the data sets, following are some key results:

- We found that people who where financially stable such as having regular income from salary or other sources slept healthy amount others compared to others who were financially not stable [Figure 7].
- Only 62% of the United States population sleeps healthy amount of hours on average. i.e. between 7 and 9 hours.
- Nearly 14% of the population sleeps unhealthy amount of hours i.e. less than 6 hours or more than 10 hours.

The Sleepscore devised using the sleep information was plotted with other variables in the data set giving interesting results about the correlation between them. For example,

- Factors such as marital status was correlated with the sleep score. People who were married had a better sleep score than people who were either divorced, widowed or separated.
- There was no correlation between smoking or alcohol habits and the sleep score.

- We found correlation between the citizenship status of an individual and his/her sleep score.
- Health was strongly correlated with the sleep score. Individuals who felt their health was excellent had way better sleep score than people who thought otherwise [Figure 8].
- Financial conditions like the affordability of medical costs for emergency or normal health care was strongly correlated with sleep score. People who were worried about these costs had lower sleep score than others [Figure 9].

By choosing the appropriate predictor variables based on the analysis and linearity patterns, we fitted a linear model which can predict a sleep score based on the input values for the predictor variables. The model has a test RMSE of 18.83

3.3 Analysis of affordability of health care

- 80% of the survey takers were insured while 20% were uninsured.[Figure 11]
- Private plans were the most used among all health care insurance plans.[Figure 12]. In those plans PPO (employer -provided health insurance) plans dominated all others.[Figure 13]
- The people under PPO plans were not confident in their ability to find affordable health care plans on their own. [Figure 14]
- High costs and loss of job/change of employers were the leading factors for being uninsured in the past year.[Figure 15]
- Among people facing vision issues 90% had their current status of visibility as chronic.[Figure 16]
- Change in health care plans had no effect on affordability of either eyeglasses or dental care.
- There was no linear decrease in either eyeglasses or dental care being not affordable with increasing income levels.[Figure 17]

3.4 Conversational Bot

Sample screen shot of the slack bot as referred in [Figure 18]

4 Discussion

4.1 Analysis of effect of yoga and meditation on emotional well-being:

- The US population is adopting yoga, meditation and associated practices along with being engaged in regular physical activities.
- The data did not contain sufficient information to study the impact of yoga or meditation on the behavioral health.In order to further understand the effect, we need to design an experiment targeted at how often they have been practicing meditation/yoga and also track their feelings over a period of time.

4.2 Analysis of quality of sleep and the contributing factors:

- There were many factors such as alcohol and smoking habits where I thought there might be correlation, but they didn't end up in the model.

4.3 Analysis of affordability of health care

- 80% of the US population is insured under some type of health care which is different from developing countries. Also the transition from one health plan to another is smooth and doesn't cause people much trouble in paying bills.
- The number of people who could afford both eyeglasses and dental care were much more than the other two which shows that they are generally affordable or that the distribution of survey is not even across different sections of the population.
- People with employment need to be taught about finding health insurance so that they feel more confident in finding one as loss of job is one of the major reasons for being uninsured.
- To improve the results, we will try to find more data which could probably give us more observations for cases where affordability has values of 0 and 1. This would help us to build a better predictive model and also solve the issue of data imbalance. While doing this we can also come across more features that can be used as explanatory variables.

4.4 Conversational Bot

- To make it usable for any data we need every data to be structured the same way (Which is nearly impossible) or have an utility that parses the code book info to generate metadata. A sample meta data file can contain these information.

  - Description of the whole data set.
  - A set of tags that describe the data set. This can be used to merge two data sets or columns together in the future.
  - Description for each column.
  - A set of tags from the corpus that describe the column.
  - Data type for each column.
  - Any special values that indicate NA.
  - Levels if its a categorical column.

5 Statement of Contribution

Ankitha Kumari Moodukudru: Performed analysis of yoga, meditation and feelings variables. Wrote functions in the R package corresponding to the analysis performed

Deepanshu Parihar: Responsible for the analysis of health insurance aspect of the data. Built a predicitive model for affordability of denta l care and eyeglasses. Added functions and metadata in the R package for the data related to healthcare.

Harish Ramani: Responsible for creating the conversational bot. Participated in the initial exploratory analysis of the data. Created the initial R package structure and maintained the whole project by reviewing pull requests and adhering to best practices.

Viral Pandey: Responsible for analysis of sleep patterns and associated factors affecting the sleep quality of the participants. Built a linear model that predicts sleep score of an indivual. Added functionality and metadata corresponding to the analysis in the R package.

6 References

- NHIS Raw Data
- Dialogflow documentation
- Plumber API Documentation
- R Package tutorial

7 Appendix

7.1 Links

- R Package
- Chat Bot
- Chat Bot Demo

7.2 Images

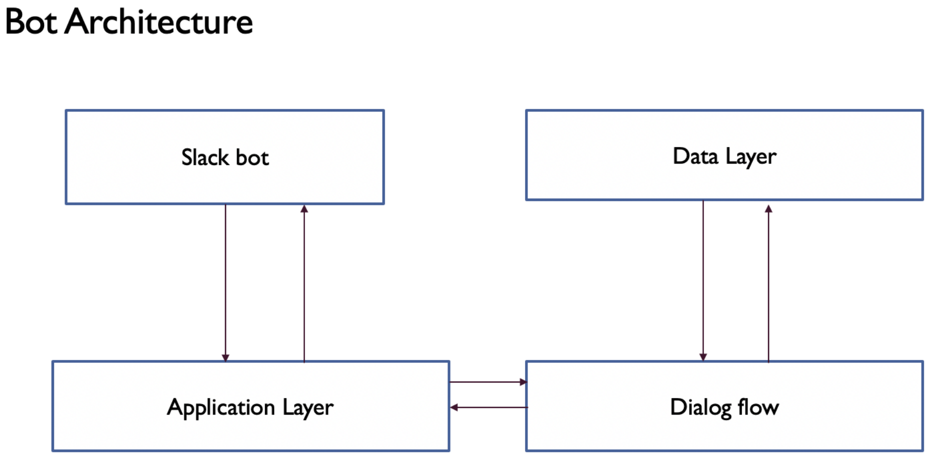


Figure 1: Bot Architecture

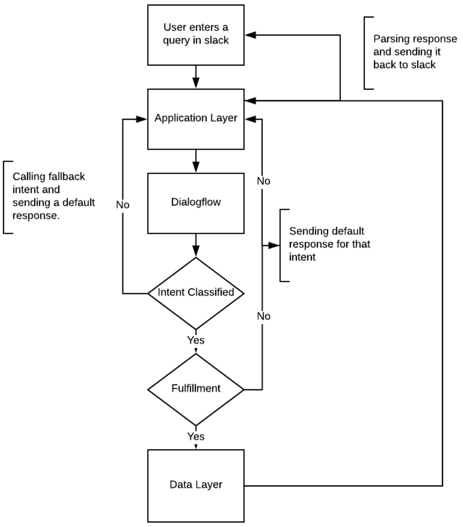

User enters a query in slack
Parsing response and sending it back to slack
Application Layer
Calling fallback intent and sending a default response.
No
Dialogflow
No
Sending default response for that intent
Intent Classified
No
Yes
Fulfillment
Yes
Data Layer

Figure 2: Flow chart for Bot

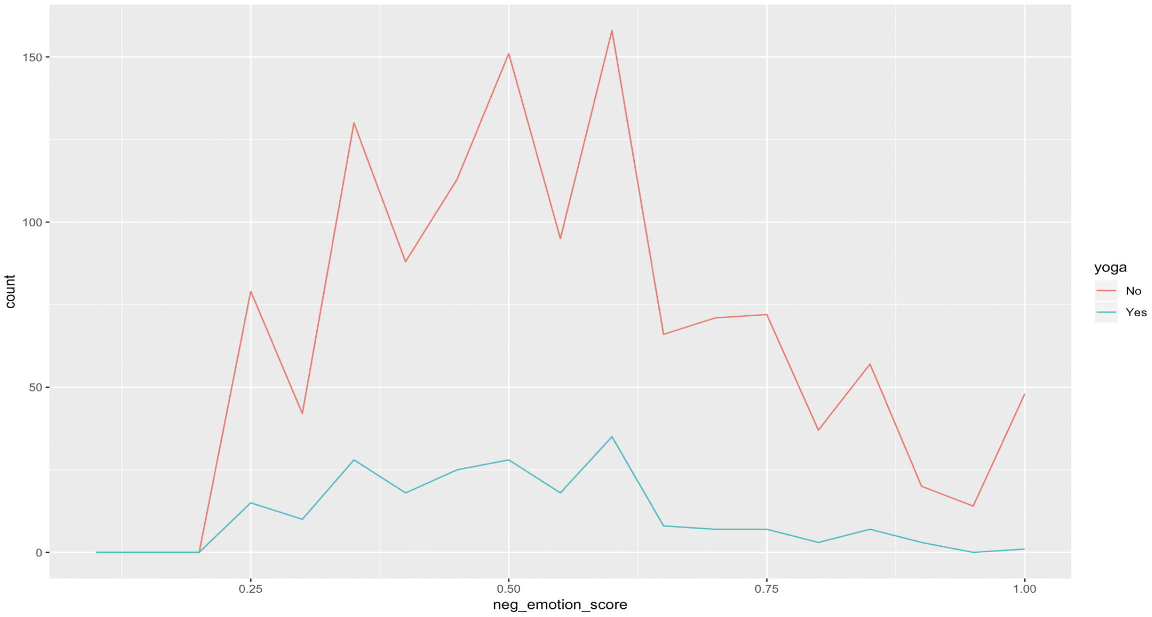


Figure 3: Yoga Vs Negative emotion score

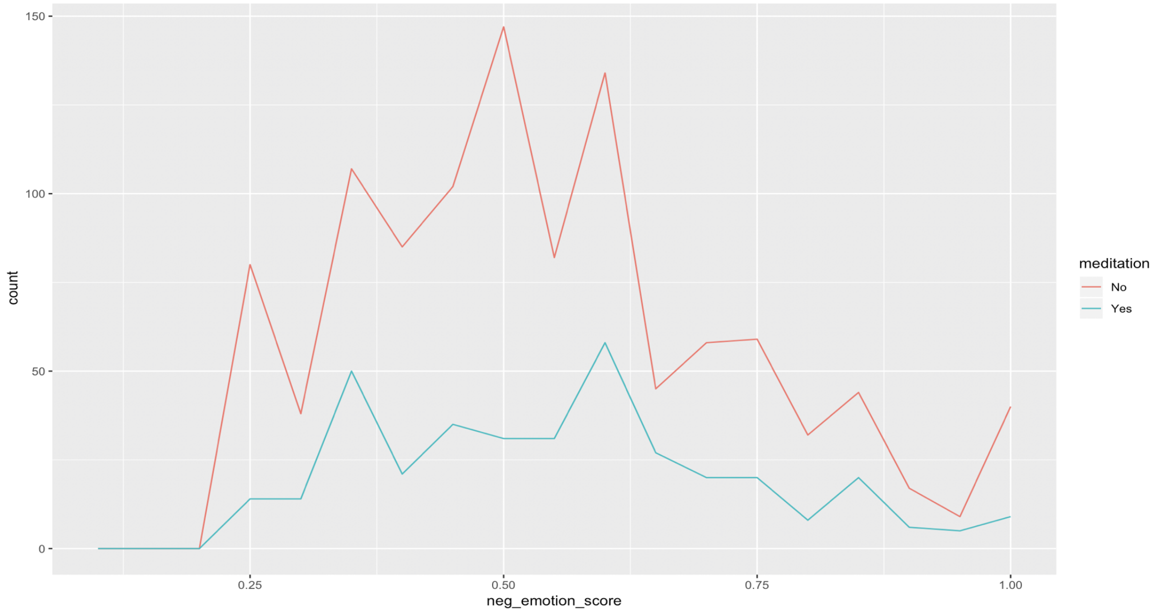
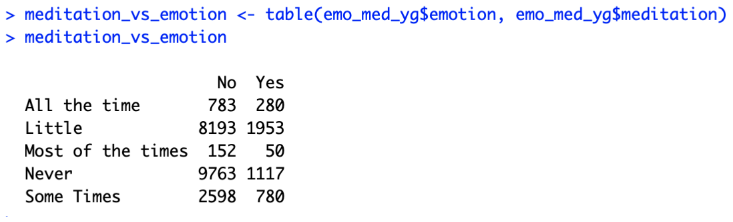


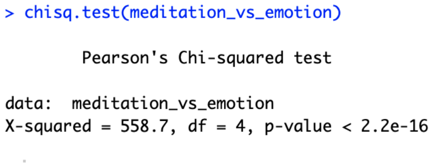
Figure 4: Meditation Vs Negative emotion score



```
> meditation_vs_emotion <- table(emo_med_yg$emotion, emo_med_yg$meditation)
> meditation_vs_emotion

                     No  Yes
  All the time      783  280
  Little           8193 1953
  Most of the times  152   50
  Never            9763 1117
  Some Times       2598  780
```

Figure 5: Frequency table



```
> chisq.test(meditation_vs_emotion)

        Pearson's Chi-squared test

data:  meditation_vs_emotion
X-squared = 558.7, df = 4, p-value < 2.2e-16
```

Figure 6: Chi Square test results

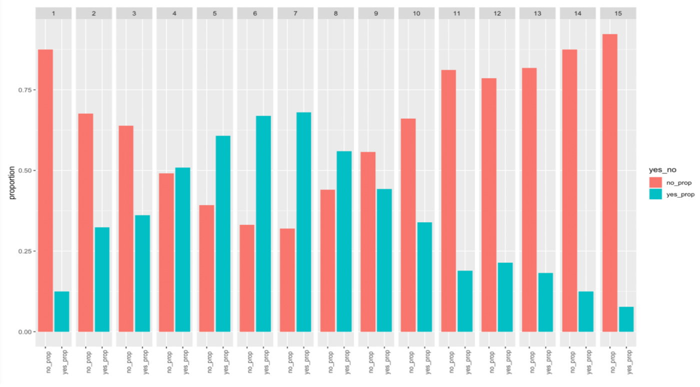


Figure 7: Proportion of people based on whether they have income from regular salary faceted by number of hours of sleep

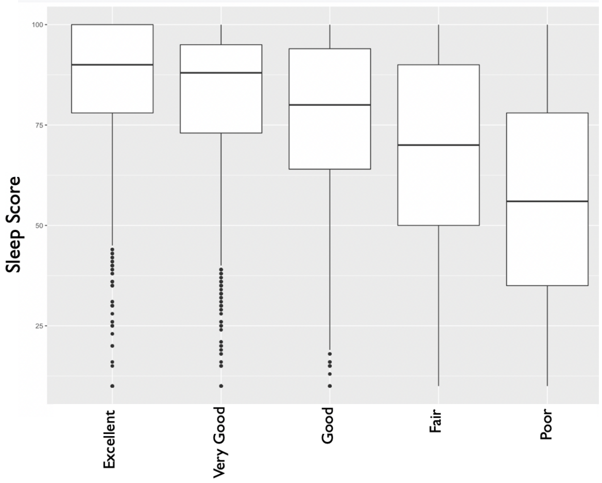


Figure 8: Sleep score Vs Health status

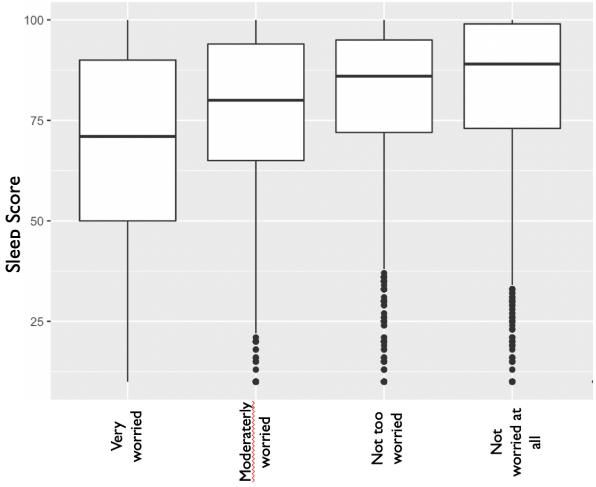


Figure 9: Sleep score Vs Affordability for normal health care

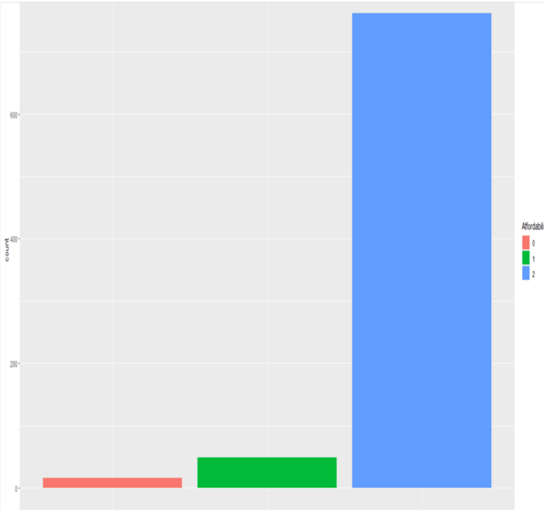


Figure 10: Data Imbalance

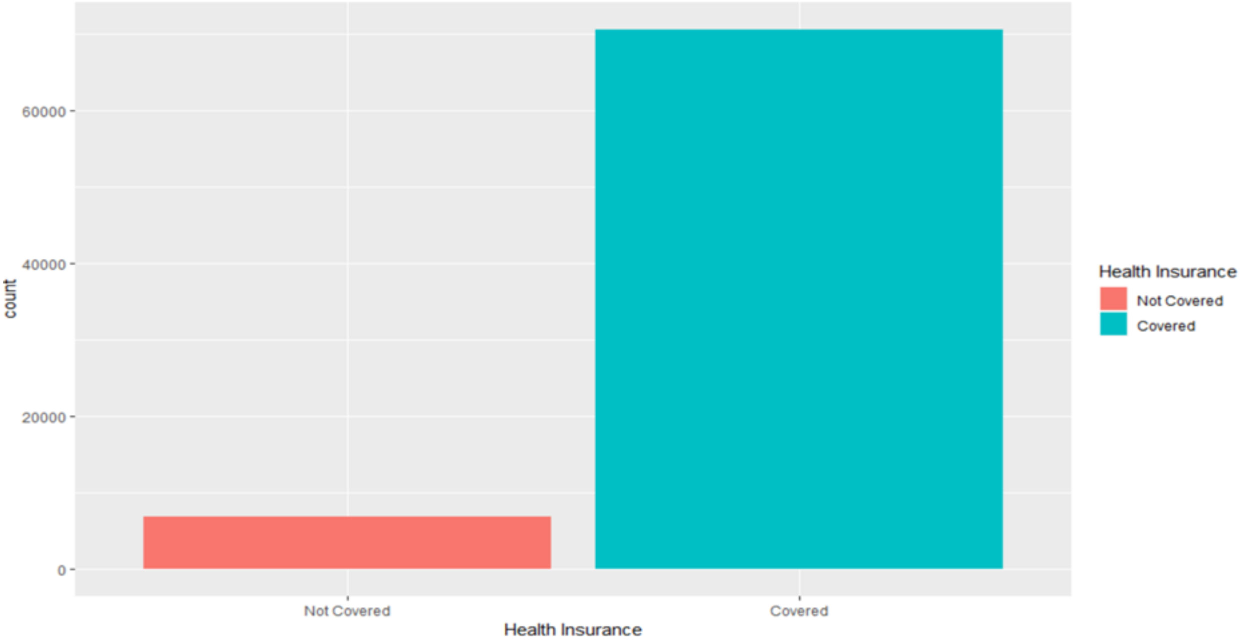


Figure 11: Insured Vs Not Insured

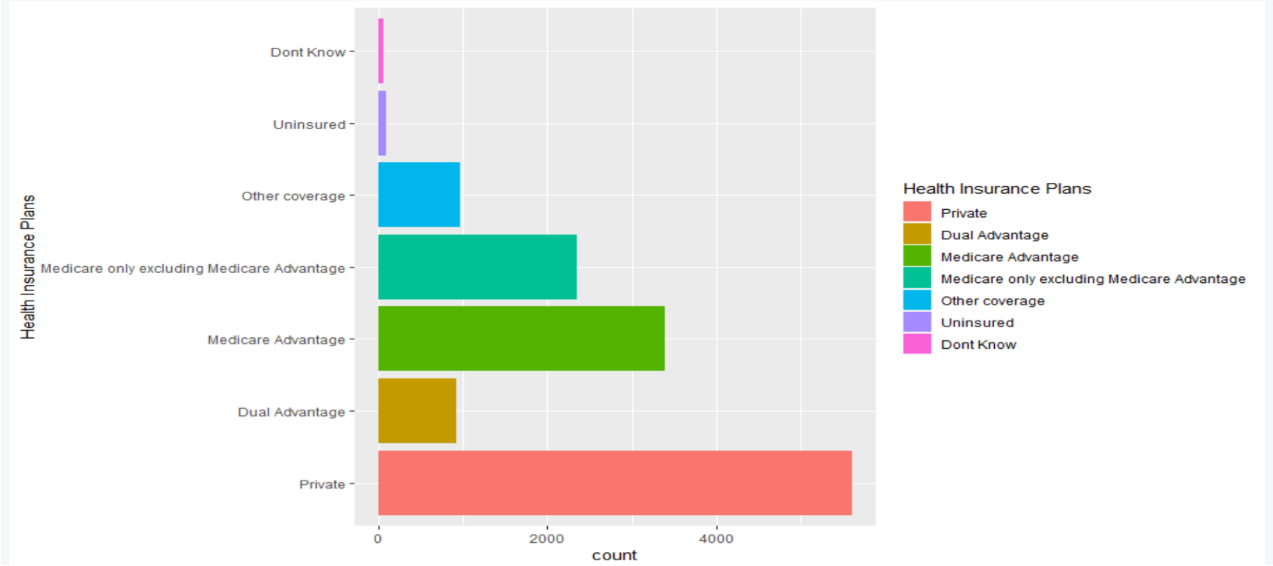


Figure 12: Health care insurance plans

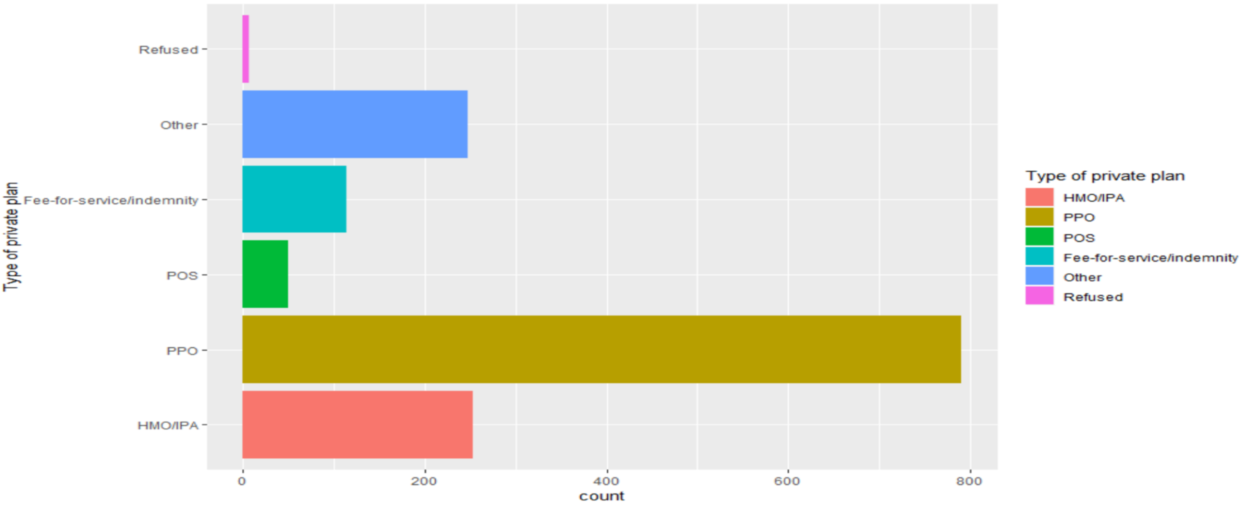


Figure 13: Private insurance plans

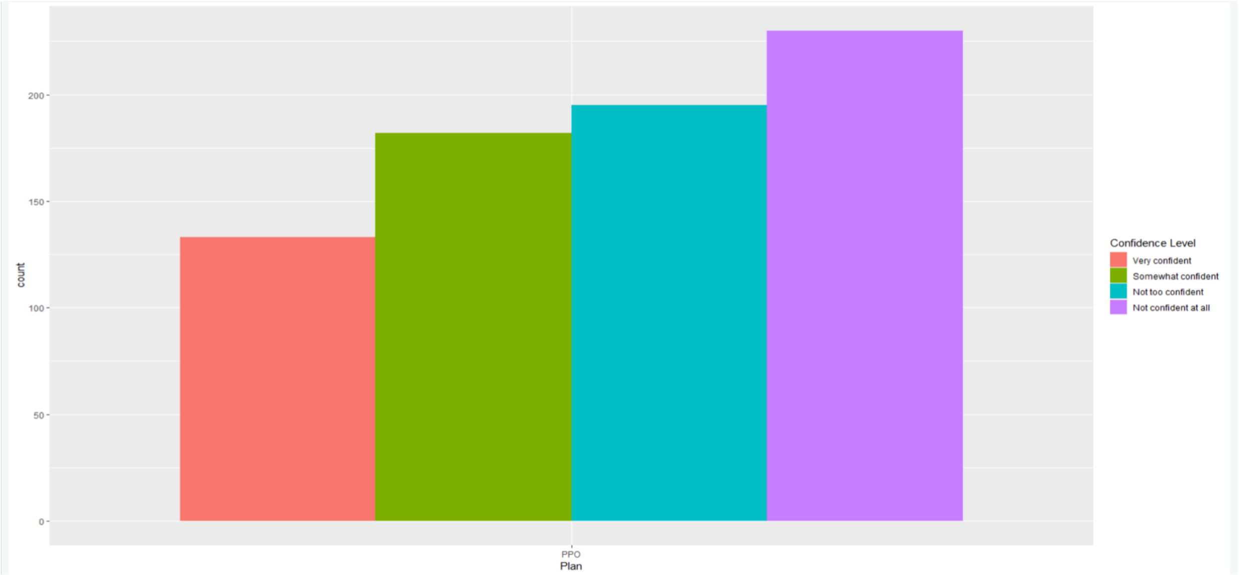


Figure 14: Ability to afford other health care plans

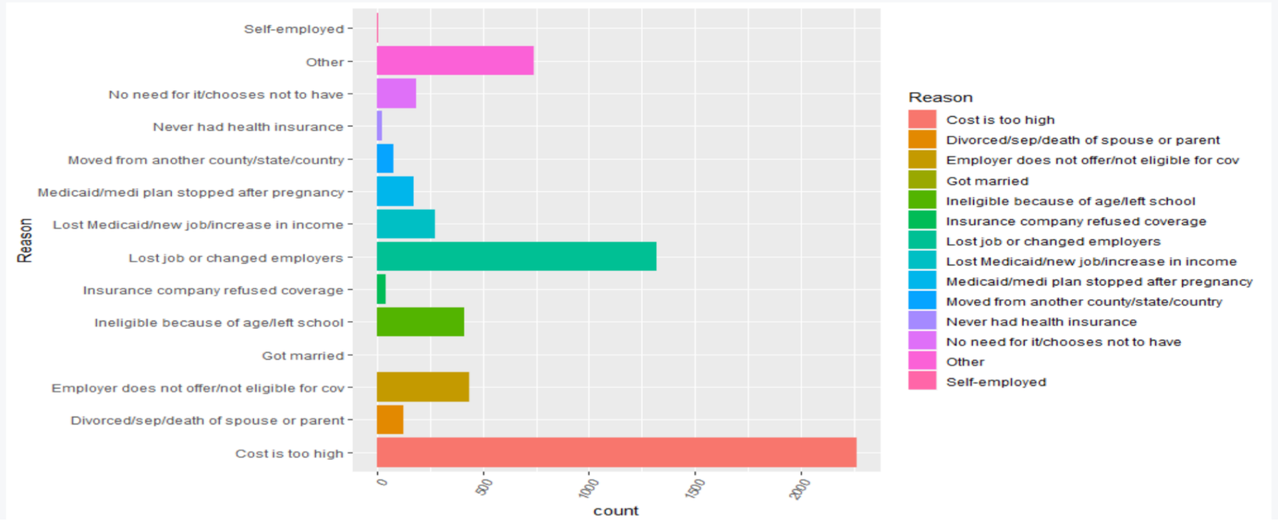


Figure 15: Reasons for not being insured

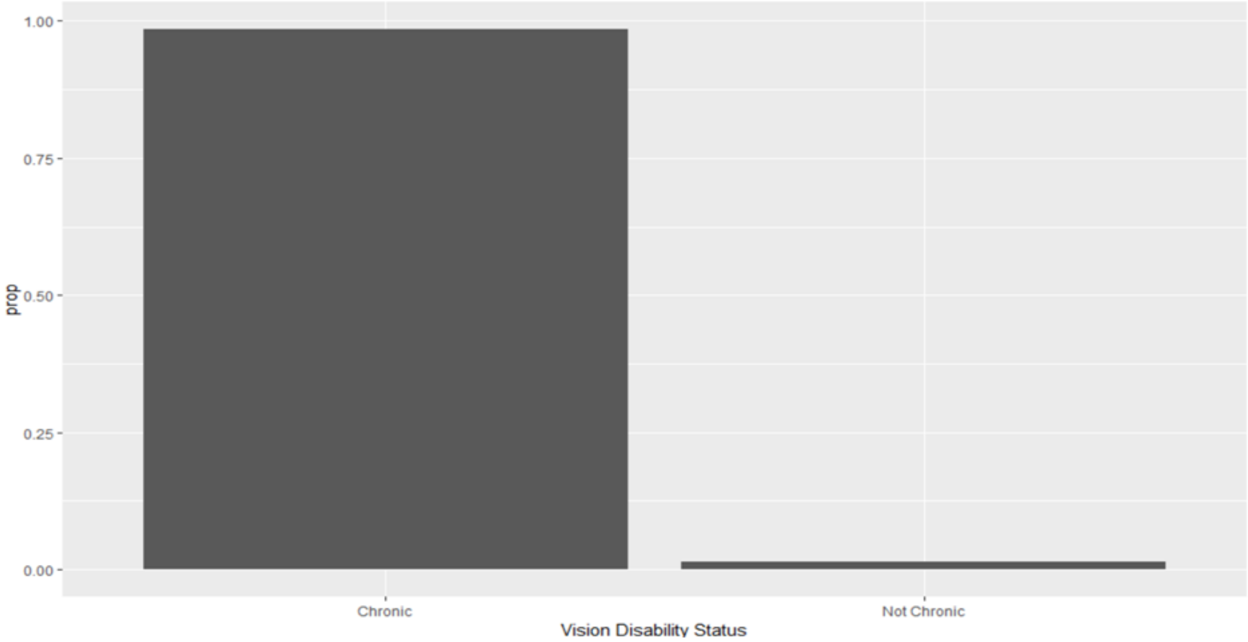


Figure 16: Current status of Visibility

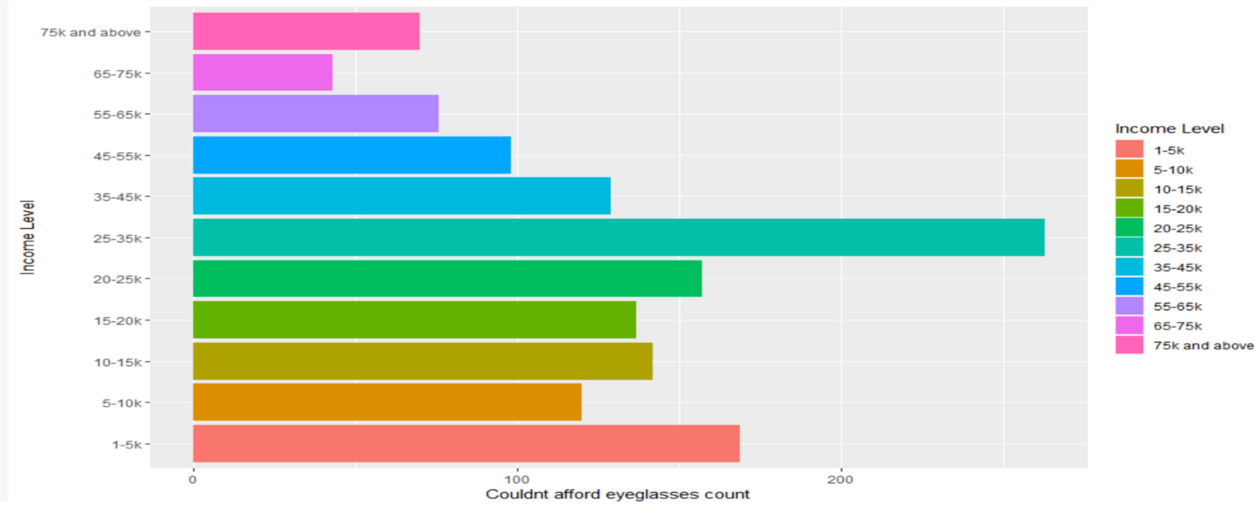
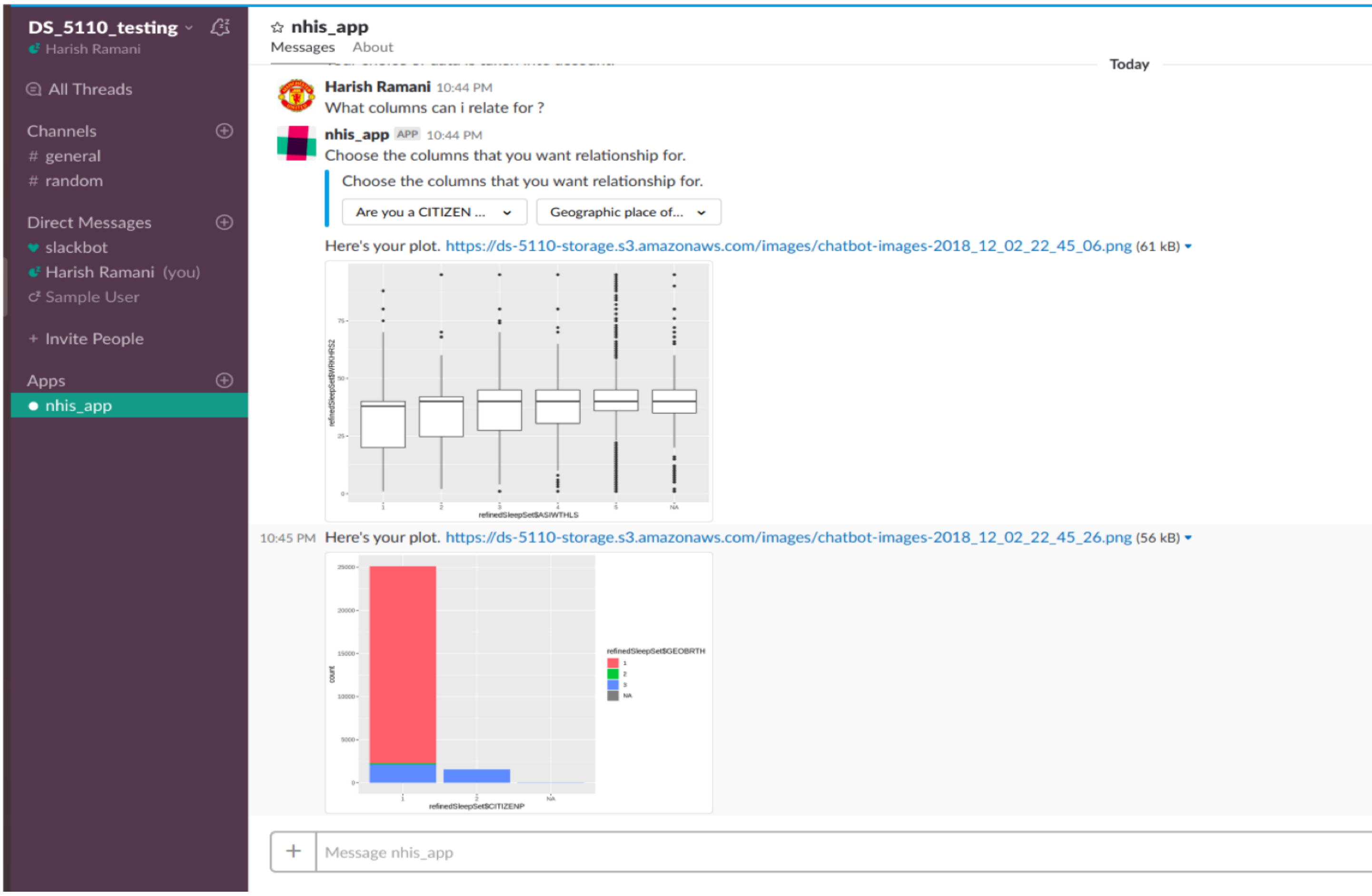


Figure 17: Income vs Ability to afford Eyeglasses



Harish Ramani 10:44 PM

What columns can i relate for ?

nhis_app APP 10:44 PM

Choose the columns that you want relationship for.

Choose the columns that you want relationship for.

Are you a CITIZEN ... ⌄ | Geographic place of... ⌄

Here's your plot. https://ds-5110-storage.s3.amazonaws.com/images/chatbot-images-2018_12_02_22_45_06.png (61 kB) ▾

10:45 PM Here's your plot. https://ds-5110-storage.s3.amazonaws.com/images/chatbot-images-2018_12_02_22_45_26.png (56 kB) ▾

Figure 18: Slack Bot