United States Home Health Care Analysis

Yinfeng Zhou

2020/11/26

# Introduction

This project is meant to implement certain data analysis on the United States Home Health Care Dataset. The motivation of this analysis is to help understanding what factors contributes to a better home health care in rating, and thus help getting knowledge on how we should improve home health care and what the future direction will be in this field. The dataset used in this report is downloaded from Data.Medicare.Gov.

The dataset includes a number of variables that provides measurements of the quality of the home care services. Among the variables that are directly related to the home care service, there are 6 binary variables and 17 continuous variables. The binary varibales answer those “Yes or No” questions: Offers Nursing Care Services, Offers Physical Therapy Services, Offers Occupational Therapy Services, Offers Speech Pathology Services, Offers Medical Social Services, Offers Home Health Aide Services. The continuous varibales mainly answer those “How often” questions, using the measure percentage as reported. For example, in the variable How often the home health team began their patients' care in a timely manner, the number indicates the percentage of the home health team having begun their patients’ care in a timely manner.

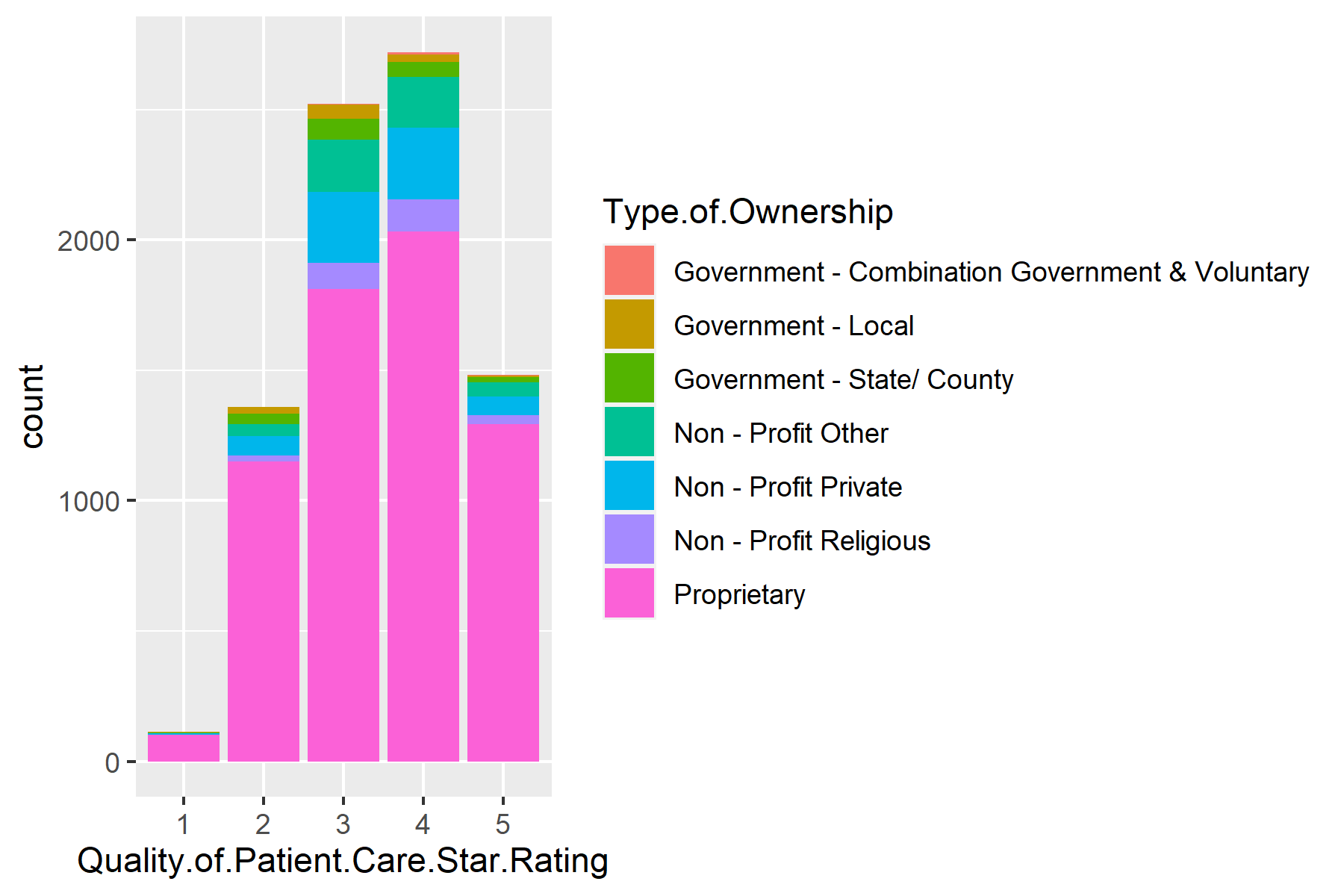
The outcome we are interested in is the Quality of patient care star rating. Although this variable is collected as a numeric rating from 1 through 5, in increments of 0.5, we can treat it as categorical outcomes with 9 possible outcomes.

In this report, we are mainly interested in how those continuous variables are correlated with the quality rating, building up a model for prediction and hopefully we can do a causal inference. For convenience, the variables name in this report will be changed as in the Appendix A. The variable name in the plot remains the same as in the dataset.

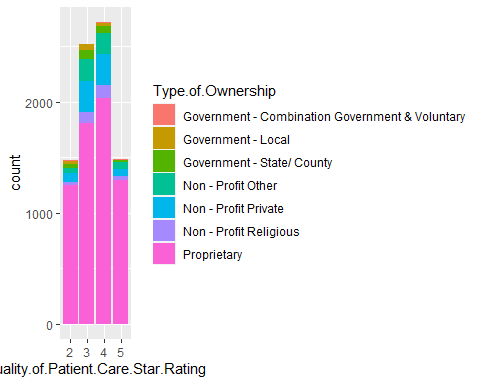
# Method

## Exploratory Data Analysis

In this section, we will apply EDA(Exploratory Data Analysis) on the dataset and get a sense of certain characteristics of the data. To simplify the the analysis, we will take ceiling of the Quality of patient care star rating. For example, 3.5 will be rounded up to 4. Without rounded up, the model fitting will also take too much time (longer than 48 hours using stan\_polr).  
First, we want to get a sense of the distribution of the quality rating by type of ownership.



Since the observations from the first category is much less than the other categories, the rating 1 and 2 are combined as one category, in order for a better model fitting.



From the plot we can tell that counts of rating on both ends are less than the middle ones, which accords with common sense. We then make plots of continuos variables against Rating.

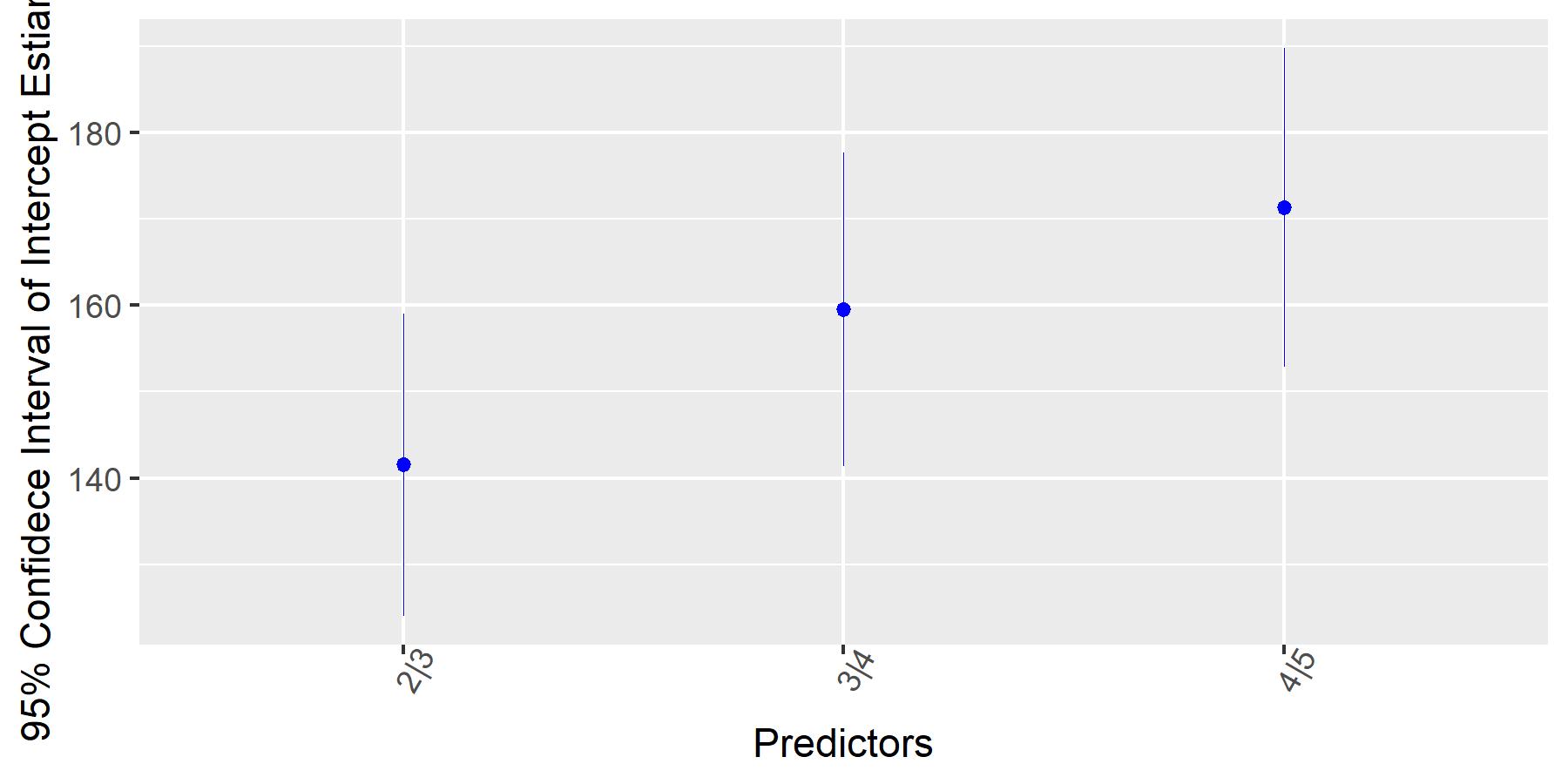
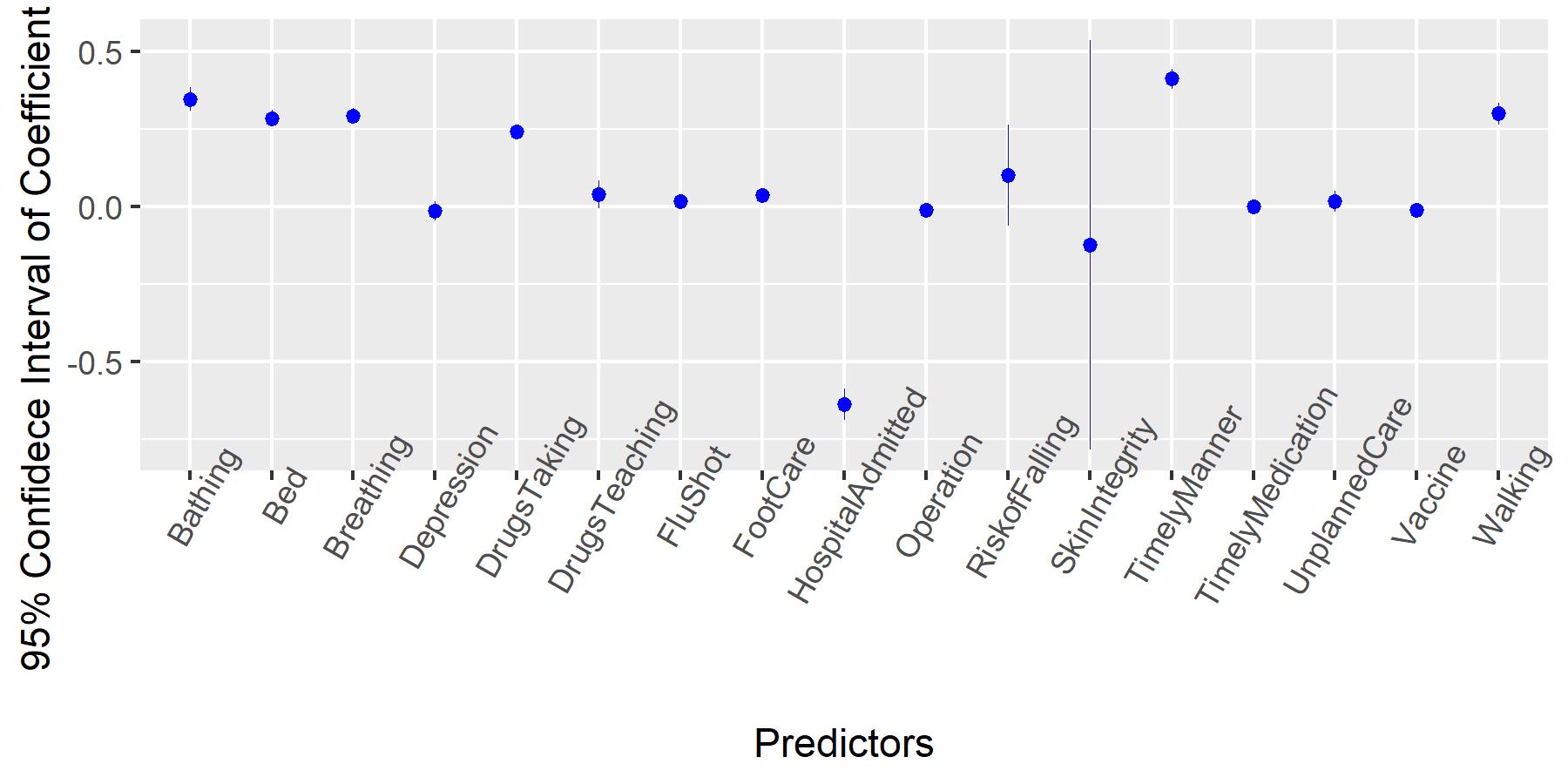
pp1<-ggplot(data=rating,aes(x=TimelyManner,y=Quality.of.Patient.Care.Star.Rating))+geom\_point(size=0.01)+geom\_jitter(size=0.01)+ylab("Rating")

An example of continuous variables versus Quality of Patient Care Star Rating are plotted above. The rest will be put in the Appendix B. In general, for the variables RiskofFalling, DrugsTeaching, RiskofFalling, Depression, FootCare, the points concentrate on the right side, indicating these services are usually provided with a large percentage. Variables FluShot and Vaccine, while also show a tendency on a large percentage of service, they also have a larger dispersion towards left. In addition, the dots also have a tendency of concentrating on the upper-right, indicating that there should be some positive correlation between the variables and the ratings.  
Moreover, Walking, Bed, Bathing, Breathing, Operation, DrugsTaking are showing a more obvious positive correlation. The HospitalAdmitted and SkinIntegrity show negative correlations, which is consistent with what the variables represent.

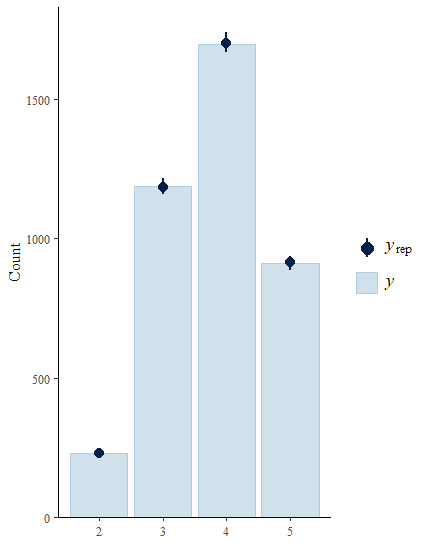
## Modelling

### Ordinal Categorical Logistic Regression Model

In this section, we will fit an ordinal categorical logistic model on the dataset. The stan\_plor() function will be used to build the model. To avoid wasting too much time on model fitting, we will randomly draw 4000 observations for fitting.

 From the coefficent plot, we can extract some useful information. First, the SkinIntegrity have a large standard error, indicating that there is a large uncertainty in the estimate. Several coefficients, such as Depression, Flushot,Vaccine,Operation,UnplannedCare,SkinIntegrity,TimelyMedication, RiskofFalling have 95% confidence interval crossing 0, indicating that we can not safely reject that these coefficients should be 0. The result is also consistent with what we found in the EDA, where under these predictors, the Rating doesn’t show a noticable positive or negative correlation.

## Validation



In the ppc bar plot, the darker blue dots indicate the medians of yrep, and the intervals indicates the uncertainty intervals. From the posterior predictive check, the plot clearly tells us that the model fits quite well.

# Result

From the EDA part, we could tell that the From the modelling part, the ordinal categorical logistic regression model fits well on the dataset. Some variables have a siginificant coefficient estimate on 5% siginificance level, while variables such as Depression, Flushot,Vaccine,Operation,UnplannedCare,SkinIntegrity,TimelyMedication, have confidence interval crossing 0. On the other hand, Bathing, Bed, Breathing, DrugsTaking, TimelyManner, Walking have relatively large coefficient estimates. From the validation part, we use ppc bar plot to show how well the model fits. From the plot, the medians of predictive values yrep are basically the same as the counting of observative values y, indicating the model fitting good on the datset

# Discussion

The results drawn above shows that, for some variables, they might not have siginificant impacts on the Rating, such as Depression and all those insignificant variables in 5% level shown in the results. For those significant ones, there are two groups of them. First, for those having low estimates, they are showing a positive correlation with Rating, but they contribute small to the increase on the Rating. Meanwhile, for those having high estimates, the model predicts a larger increase on Rating when they are increasing. Therefore, by improving home health care serivce such as helping patients at getting in and out of bed (Bed), beginning patient’s care in a timely manner (TimelyManner), the team should be expected to have a higher Rating from the patient.

### Appendix A

Rating: Quality of patient care star rating, a numeric rating from 1 through 5, in increments of 0.5. Factored in this report.

TimelyManner : How often the home health team began their patients’ care in a timely manner.

DrugsTeaching: How often the home health team taught patients (or their family caregivers) about their drugs

RiskofFalling: How often the home health team checked patients’ risk of falling

Depression: How often the home health team checked patients for depression

FluShot: How often the home health team determined whether patients received a flu shot for the current flu season

Vaccine: How often the home health team made sure that their patients have received a pneumococcal vaccine (pneumonia shot)

FootCare: With diabetes, how often the home health team got doctor’s orders, gave foot care, and taught patients about foot care

Walking: How often patients got better at walking or moving around

Bed: How often patients got better at getting in and out of bed

Bathing: How often patients got better at bathing

Breathing: How often patients’ breathing improved

Operation: How often patients’ wounds improved or healed after an operation

DrugsTaking: How often patients got better at taking their drugs correctly by mouth

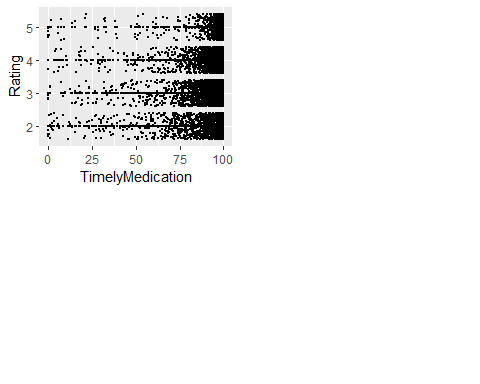
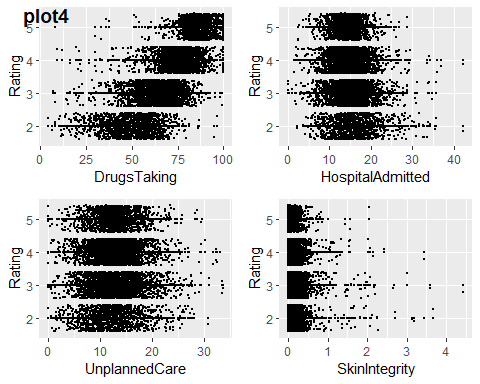
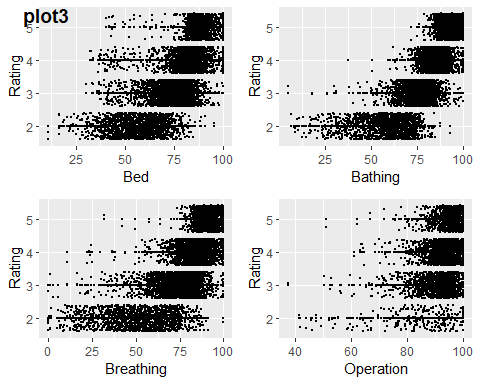
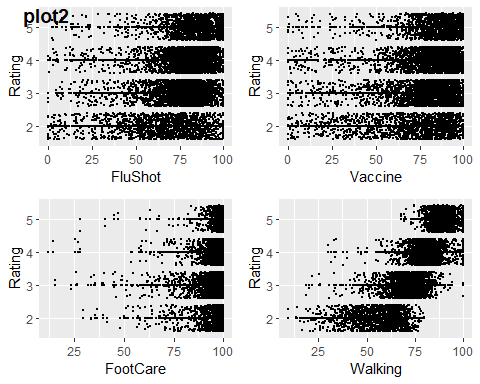
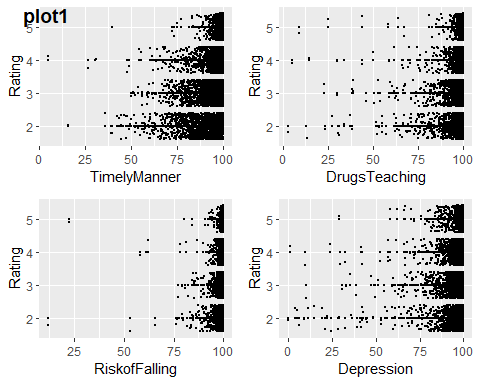
HospitalAdmitted: How often home health patients had to be admitted to the hospital

UnplannedCare: How often patients receiving home health care needed urgent, unplanned care in the ER without being admitted

SkinIntegrity: Changes in skin integrity post-acute care: pressure ulcer/injury

TimelyMedication: How often physician-recommended actions to address medication issues were completely timely

### Appendix B: EDA Plots



### Appendix B: Code

care<-read.csv2("Home\_Health\_Care\_Agencies.csv",header=TRUE,sep=",")  
care%<>%filter(Quality.of.Patient.Care.Star.Rating!="")  
sum(care$Footnote.for.quality.of.patient.care.star.rating=="")==nrow(care)  
care%<>%select(!starts\_with("Footnote"))  
  
for(i in 17:34){  
care[,i]%<>%as.numeric()  
}  
for(i in 8:14){  
 care[,i]%<>%as.factor()  
}  
care$Date.Certified%<>%mdy()  
care$Quality.of.Patient.Care.Star.Rating%<>%as.factor()  
#extract subset about quality rating  
rating<-care[,c(2,9:34)]  
colnames(rating)<-c("State","Type.of.Ownership","Offers.Nursing.Care.Services","Offers.Physical.Therapy.Services", "Offers.Occupational.Therapy.Services" , "Offers.Speech.Pathology.Services" , "Offers.Medical.Social.Services" , "Offers.Home.Health.Aide.Services", "Date.Certified" , "Quality.of.Patient.Care.Star.Rating", "TimelyManner","DrugsTeaching","RiskofFalling","Depression","FluShot","Vaccine","FootCare","Walking","Bed","Bathing","Breathing","Operation","DrugsTaking","HospitalAdmitted","UnplannedCare","SkinIntegrity","TimelyMedication")  
rating$Type.of.Ownership%<>%factor()  
rating$Quality.of.Patient.Care.Star.Rating%<>%ceiling()%<>%factor()  
#Bar Plot1  
ggplot(data=rating,aes(x=Quality.of.Patient.Care.Star.Rating))+geom\_bar(stat="count",aes(fill=Type.of.Ownership))+geom\_vline(aes(xintercept=mean(Quality.of.Patient.Care.Star.Rating)),color="red")  
#Bar Plot2  
rating$Quality.of.Patient.Care.Star.Rating[which(rating$Quality.of.Patient.Care.Star.Rating==1)]<-2  
rating$Quality.of.Patient.Care.Star.Rating%<>%factor()  
ggplot(data=rating,aes(x=Quality.of.Patient.Care.Star.Rating))+geom\_bar(stat="count",aes(fill=Type.of.Ownership))+geom\_vline(aes(xintercept=mean(Quality.of.Patient.Care.Star.Rating)),color="red")  
  
##Modelling  
rown<-sample(rownames(rating),4000)  
fitset<-rating[rown,10:27]  
colnames(fitset)[1]<-"Rating"  
m2<-stan\_polr(Rating~TimelyManner+DrugsTeaching+RiskofFalling+Depression+FluShot+Vaccine+FootCare+Walking+Bed+Bathing+Breathing+Operation+DrugsTaking+HospitalAdmitted+UnplannedCare+SkinIntegrity+TimelyMedication,data=fitset,prior=NULL,refresh=0)  
summary(m2,digits=2)  
  
#Condidence Interval  
interval<-posterior\_interval(m2)  
interval<-data.frame(name=rownames(interval),interval)  
  
#Validation  
predy<-posterior\_predict(m2)  
n\_sims<-nrow(predy)  
subset<-sample(n\_sims,100)  
yrep<-as.data.frame(lapply(data.frame(predy[subset,]),as.numeric))  
ppcplot<-ppc\_bars(as.numeric(temp$Rating)+1,as.matrix(yrep))