

HW2 - Nachi Lieder 314399114

Contents

Introduction	1
File reading	2
Logistic Regression Stage	4
Question 2	8
Summary - Q2	13
Future Thoughts	13
Attached code	13

Introduction

In this project we will evaluate a specific dataset and in specific answer the questions to 2 different models that we will assess.

The data we are dealing with questions patients several times over 3 days , and asks them about their thoughts of the future. Among the different sets of features we will be dealing with, we have :

- Day
- Time
- Age
- Gender (male/ female)
- A : agreeableness
- C : conscientiousness
- E : extraversion
- N : neuroticism
- O : openness to experience

Where the last 5 are considered the Big Five traits, and the patients are asked to rate themselves from 1-7.

In the first question , we will be asked to build a model that will answer a binary classification question of whether a patient will think of the future as a funtion of the time of day. We will assess the models results and determine whether there is some sort of statistical significance.

In the second model we will address a question of predicting how many times from the total amount of times that the observants were asked , did they think of the future. We will assess different types of models such as poisson models and quasi-poisson models and perform some sort of comparison.

Among the different analysis we will be performing are over-disperssion tests , tests on means of distributions, anova tests to compare the different models, model selection using Stepwise-AIC . Lets start off with the data exploration prior to the predicitive sections.

File reading

Lets look at the following subset:

Lets look at the time . We are talking about time since midnight , having a maximum at 86400 (seconds in a day). As we see in the plot , there arent any non-rational values in the set. What we can see are three observations that were taken sometime near (0:00 - 0:30). Since the data is cyclic and there is a cycle that needs to be taken into place (value 1 = 86401 = 0:00:01 AM) - I will convert these values to 86400+X to represent midnight since then they will be scaled the same way as the rest and the partition wont be drastic. The next observation occurs around 30000 seconds later => (8:30 AM). There are no missing data points , which is encouraging.

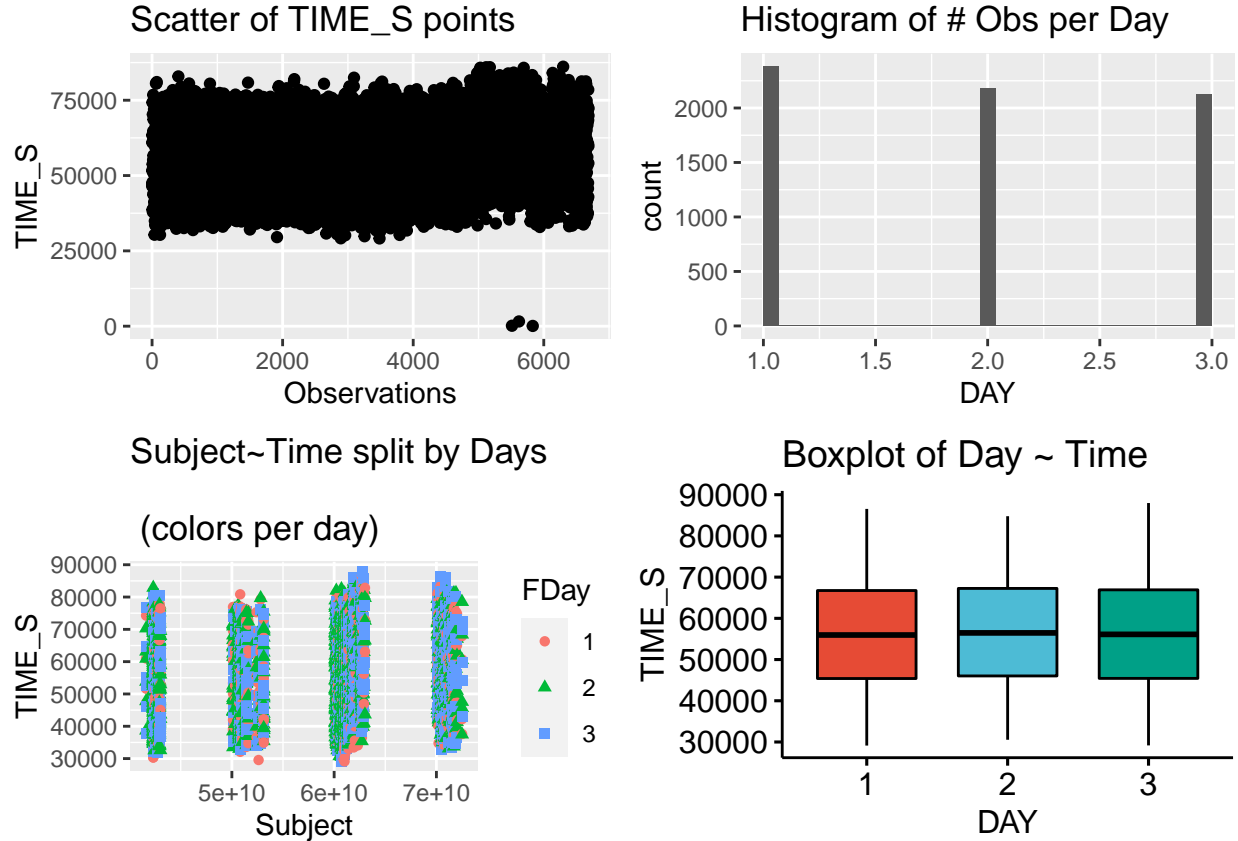
As for the Day variable , there too is no missing data , and the range looks fine [1,3] We are also talking about almost equal sets, having all above 200, though Day1 might have the most.

Regarding the Subject variable , we are looking at 492 subjects, and I wanted to inspect the average time that each subject has its observations measured. Below is the scatter plot with per subject the average time_S. What is interesting here is that there is a large variation of measurement times. We can see averages per Subject_i ranging from 30000 to 80000. We will take this into mind when analyzing the results of our analysis since this can be taken as a bias for the model.

Also , I verified the diversity of the time per day per subject , in the following plot we can see that there is no pattern of time of the day relative to the day of the observation and the subject.

To conclude this part of the analysis I validated that per day , the average time is somewhat similar. In the boxplot below we can see that the medians are very close and we can assume similarity between the distributions of the three classes.

This concludes the exploration analysis per the 4 features



Here we can see the results of the anova test that I performed between Time and the different DAYS , validating the box plot and adding significance , that there is no difference between the different days and their time's mean. We can see that the P val is high and F statistic that is low , and we can assume similarity between the groups.

On the other hand , we can see here the test of the Future ~ Day , where we can see that it is quite significant the difference between the behavior of the subjects in the days with respect to their mean Future. The P val is quite low here which confirms that there is a difference that we should consider in the future and a significantly high F stat which strengthens this claim.

Since we see the significant difference between days in terms of the Future variable means , we will later see the comparison between the upcoming logistic regression of Future ~ Time vs the logistic regression of Future ~ Time + Day.

We will address this enhanced model after we assess the original model.

Table 1: Anova Test TIME_S ~ DAY

Effect	DFn	DFd	F	p	p<.05	ges
DAY	1	6684	0.369	0.543		5.53e-05

Table 2: Anova Test future ~ DAY

Effect	DFn	DFd	F	p	p<.05	ges
DAY	1	6684	26.3	3.07e-07	*	0.004

Logistic Regression Stage

Lets assess a classic Logistic Regression with the Time as the predictor to preedict the probability of a future thought.

I decided to scale the time variable to a range which is closer to the range of the target to optimize computation and convergence of the model.

We can see that standalone , the time has a pretty significant Z val with -6.3 (negative relaionship) , and a very low P val.

We can also see that while splitting the sets to train and test , to evaluate the accuracy , we can see that this specific model is able to acheive an accuracy of 70% (which is subjective to the threhold and benchmarks we are trying to exceed). This could give us an indication of the fitness of the model and another benchmark and pivot point for model selections in the future. Although this model was not fine tuned , it is a benchmark that we can pivot off of while createing more complex models.

The improvement of the dispersion from the null disperssion is quite small , shifting from 8073 to 8033.This shows that the additional input of our single predictor is not quite good enough. our score is equal to 0.004 .

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.893	0.0271	-33	8.08e-239
TIME_S_standardized	-0.171	0.0272	-6.3	2.95e-10

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	8073 on 6685 degrees of freedom
Residual deviance:	8033 on 6684 degrees of freedom

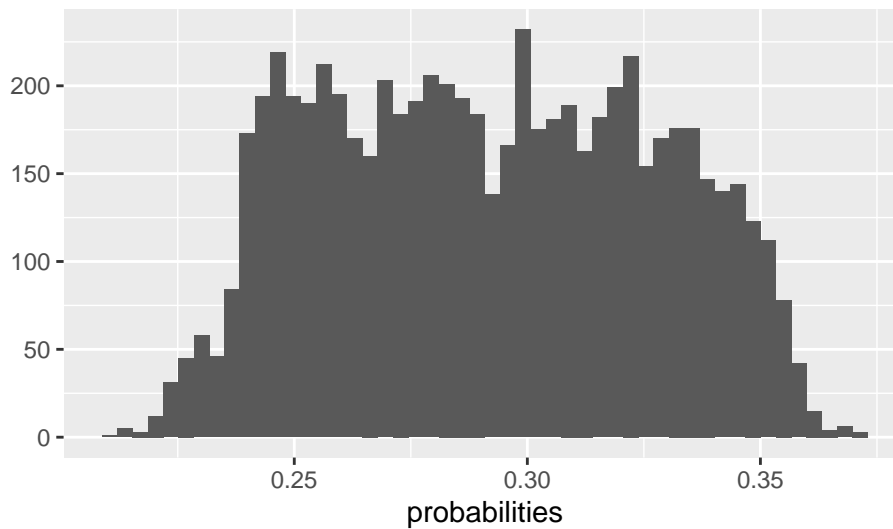
What is interesting here is that while predicting the sample test set , we get a range of probabilities from [0.21,0.37]. A suggestion would be to explore the model with some more predictors. We would expect a broader spread of results rather than the small range. Below we can observe a histogram of the resulted probabilities. We can see that within the range [0.25,0.35] it is prety well spread, and seems gaussian , but not normal.

We can see that generally speaking the range is not well spread , and that probably the model is not complex enough to create a true prediction only using the one predictor and needs more complexity and depth to define a well differentiation. We would assume that given a ratio od 1:3 between positive and negative targeted values would give a wide spread of fitted values , but here we see a dense fit around the distribution of the target (the ratio) so our model emphasizes on the distribution but doest spread the projections wide enough.

Since this is indeed surrounding the 0.3 (which is more or less the mean of the models target) , we can assume that model took into account the distribution, and we would need to create a stronger regularization to the model in order to save this mean of the distributed probabilities , but spread the values over [0,1] with a wider reasonable projection.

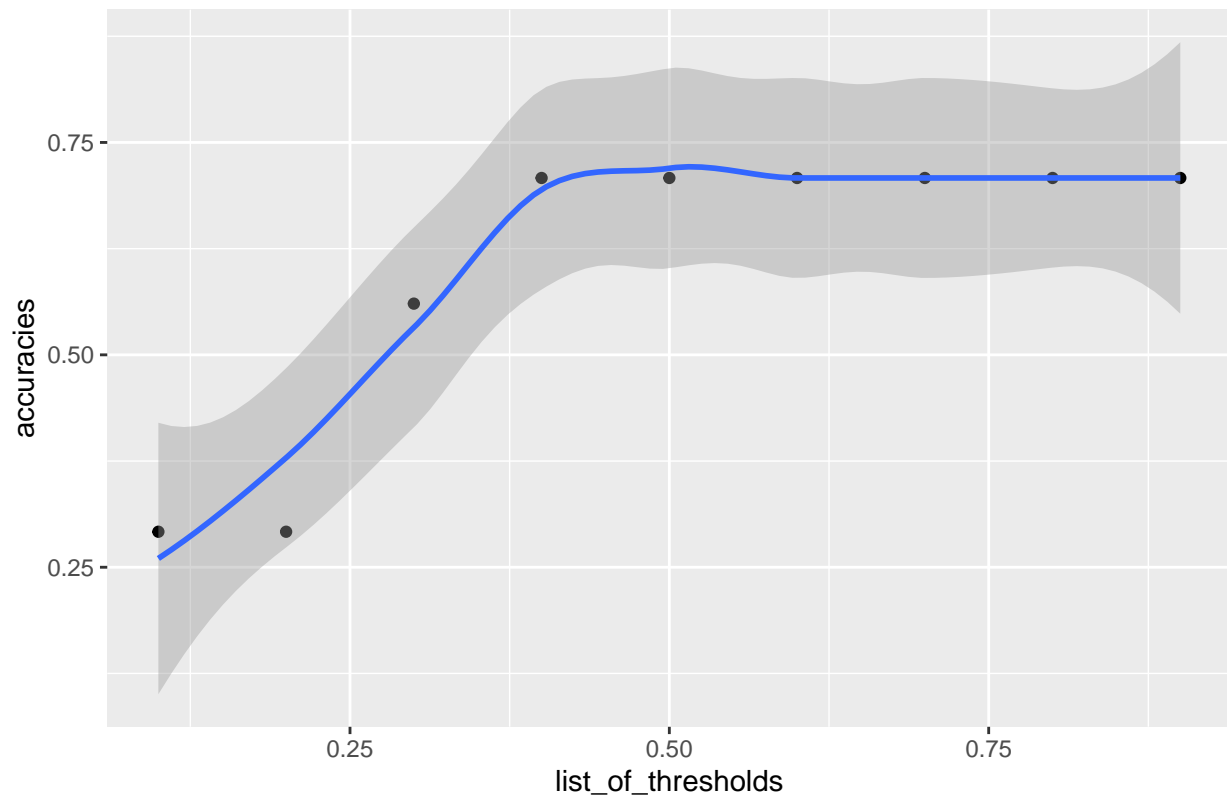
Another option here would be to change the threshold from 0.5 to 0.3 per say and assess accuracy.

Histogram of projected values for predictive model

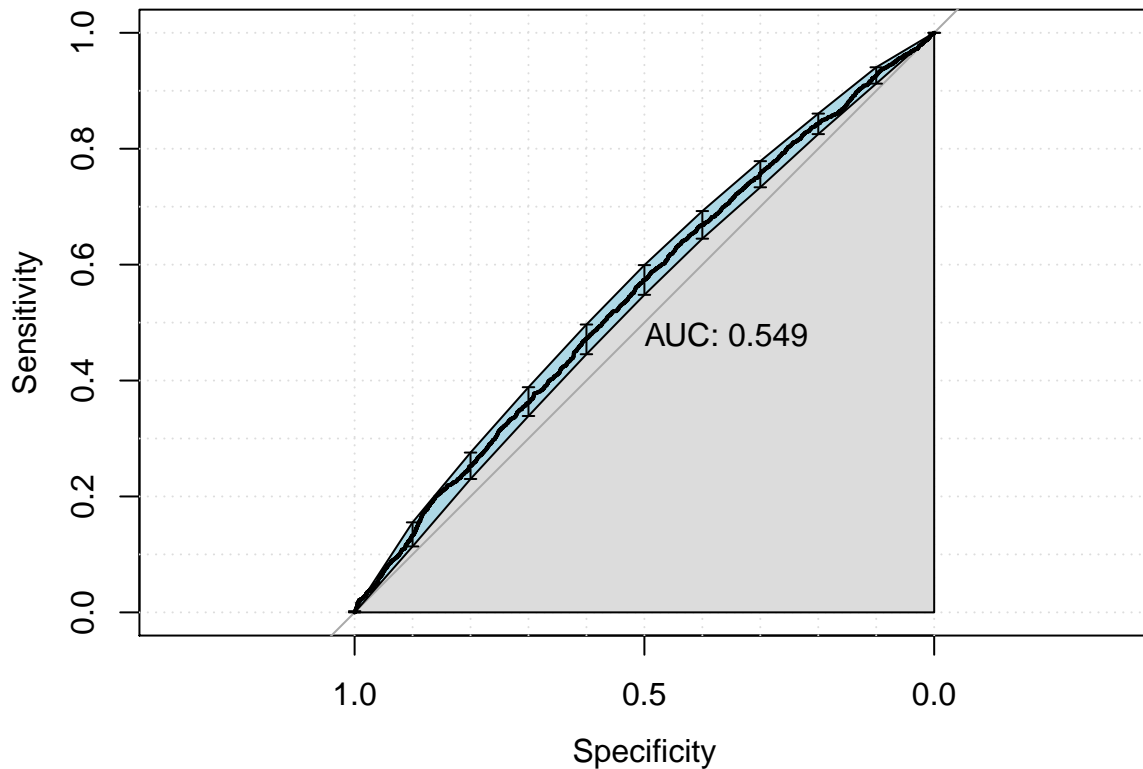


Below we can view the transition and behavior of the model in different thresholds. We can see that once we hit the 50% threshold, we reach a plateau, where this can be due to the distribution of the probabilities which we will address. We may consider a different threshold to uphold different precision/ recall scores, where the balanced accuracy may be around 0.4.

Accuracy Per Threshold



Below we can also assess the AUC of the model. In this assessment we can see a score of ~ 0.549 which is pretty low, and can strengthen our claim that the model is weak and inaccurate only using the given predictor.



```
## [1] 0.004960184
```

Now as stated above , with the findings that we found regarding the potential contribution of the additional predictor “Day” , we will fit the model $\text{Future} \sim \text{Day} + \text{Time}$, and compare the results to assess whether adding this predictor will fit a better model.

First off we can see that the DAY feature has a large absolute Z value , and a very low P val. Also , we can compare the Residual deviance of this model to the previous having an improvment from 8033 to 8008.

We can also compare the AUC and see that below the model provides an AUC of 0.56 vs the 0.54 in the prior model (a slight improvment)

A reccomendation here would be to possibly suggest adding this given Day feature into the model for a better fit of the model.

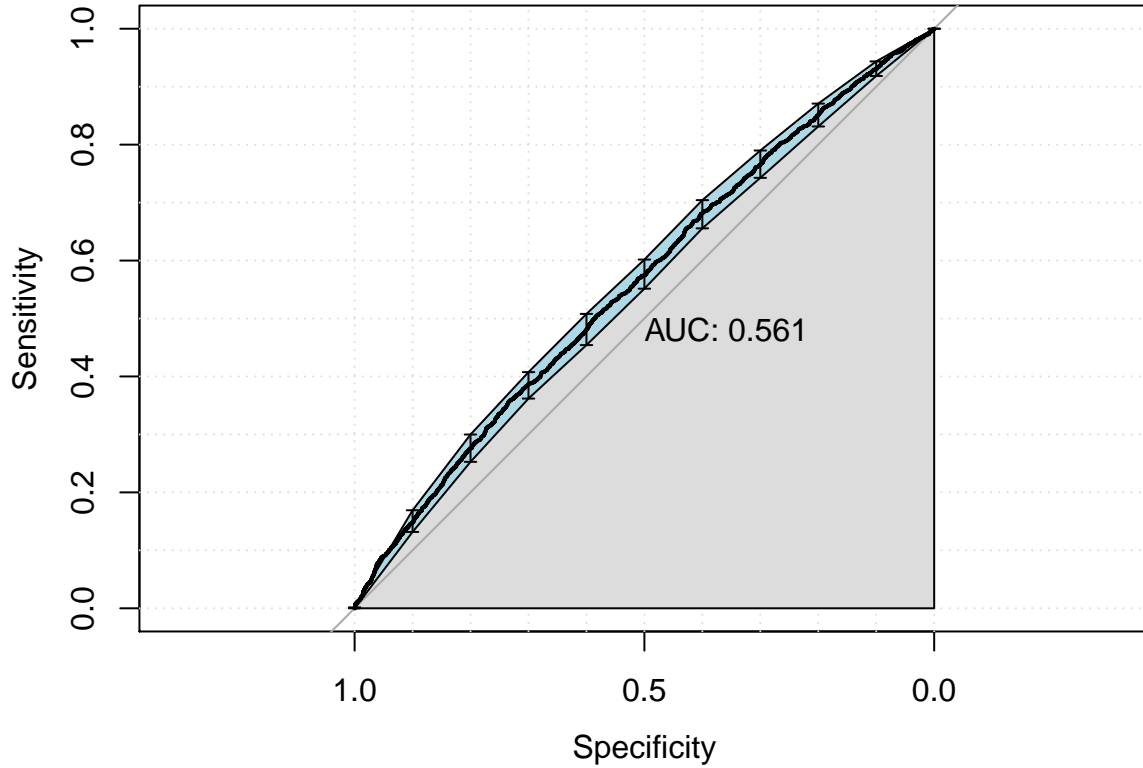
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.567	0.0689	-8.22	1.97e-16
TIME_S_standardized	-0.171	0.0272	-6.28	3.48e-10
DAY	-0.168	0.0331	-5.08	3.86e-07

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	8073 on 6685 degrees of freedom
Residual deviance:	8008 on 6683 degrees of freedom

Table 7: Anova Test - Enhanced model vs LR

Resid. Df	Resid. Dev	Df	Deviance
6683	8008	NA	NA
6684	8033	-1	-25.9



The next step is to try to evaluate the targeted variable using a mixed model with time of day as a fixed effect

When observing the standard error we can see that given the day within the subject ,there is a much lower variance , than just the subject as it is. There is a difference of 0.08 to 0.53. This means that the model and its variance can be explained by fitting an intercept specifically per subject , and that the main noise indeed comes from the Day.

In addition , while looking at the fixed effects , we can see that both the regular intercept and also the Time (standardized) have significant estimates. So per subject we receive its own intercept based on the usual intercept + the time standardized which is tailor made per subject. The coefficients are obviously equal to all in both models.

We can see that the AIC in the mixed model evaluation is slightly lower than the linear model above , shifting from 8037 to 7803 , indicating that this model is slightly more predictive and well fit.

The addition of adding the level of the DAY into the mixed model gives only a slight if not any improvement. What is interesting is that adding another model with only the Day as the mixed random variable , we get a higher AIC which indicates that the major factor here is indeed the subject (which makes sense)

Generalized linear mixed model fit by maximum likelihood (Laplace

```
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: future ~ TIME_S_standardized + (1 | Subject/DAY)
## Data: newdata
##      AIC      BIC    logLik deviance df.resid
## 7803.723 7830.954 -3897.861  7795.723     6682
## Random effects:
## Groups      Name      Std.Dev.
## DAY:Subject (Intercept) 0.2885
## Subject      (Intercept) 0.7291
## Number of obs: 6686, groups: DAY:Subject, 1379; Subject, 492
## Fixed Effects:
##      (Intercept)  TIME_S_standardized
##      -0.9946      -0.1903
```

To compare in depth these two models (linear model vs the mixed affect model) I performed an anova test which summarizes the comparison. Here we can see the difference between the AICs , as well as the LogLikelihood . Also notice the Chi Squared high level , and very low P val which conclude the rejection of the null hypothesis of equal models in terms of the deviance and addition of new features. This concludes that the addition of the predictor Subject and Day are significantly considered contributors.

In terms of agreement , it seems like the two models somewhat behave the same more or less , with one having a slight improvement. The difference here would be that for the advanced mixed model , we have a more tailor made model , fitting the intercept to have a more individual set , where each subject and observation would have its own additional contribution to the intercept. By lowering the generalization we are able to receive slightly better results.

Table 8: Anova Test - Mixed model vs LR

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
fit.lr	2	8037	8051	-4017	8033	NA	NA	NA
mixed_model	4	7804	7831	-3898	7796	238	2	2.43e-52

Question 2

I will retrieve the 7 given features (age , sex , and the big five features) and create an aggregated dataframe representing the aggregation grouped by subjects , and their descriptive analysis.

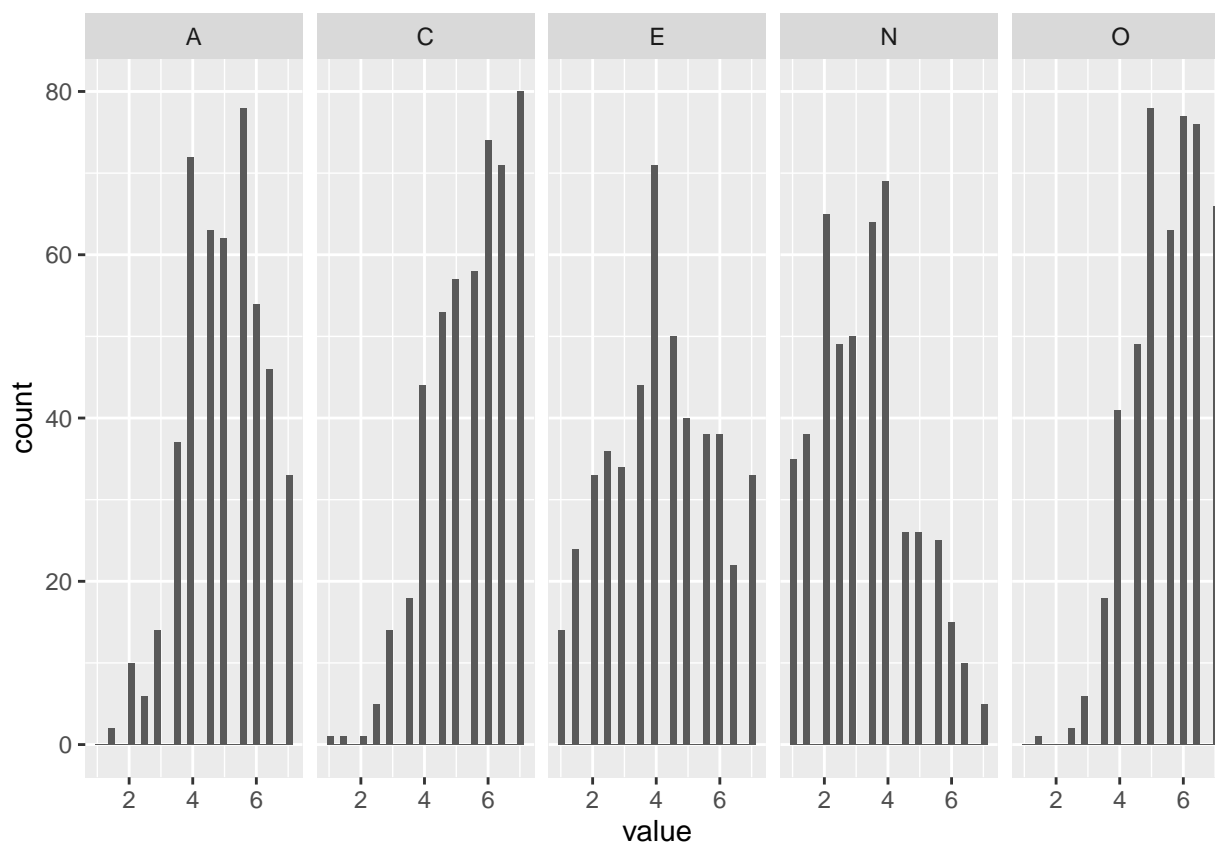
while viewing out the aggregation count , we can see that different subjects have different amounts of observations. The smallest set was 1 observation per subject , and largest 18.

Lets try to observe each subjects big five by representing its categories in the form of means per subject. The following set of histograms per each category represents the means distribution . As you can see there are very different distributions per category.

For an instance , categories A, E are much more gaussian ,while the density of C and O are emphasized on the right, and category N on the left in terms of skewness.

My personal opinion in the matter: Since the scale of the big 5 is from 1-7 , we can say that traits O and C (openness to experiences and conscientiousness) were found to have much higher ranks. These are self imaging traits that are expected to have higher values. Values of agreeableness and Extraversion are with slightly lower means in overall and representing a slightly more unbiased set of questions. Neuroticism is considered a negative trait in some way, and like the first two , is expected to have a more biased look at themselves in terms of self image.

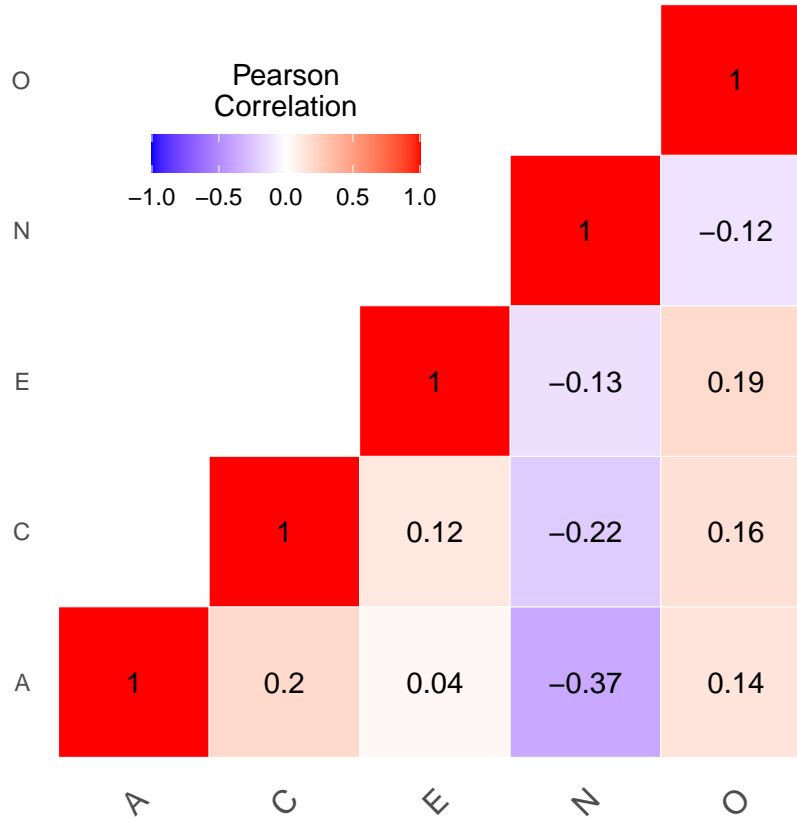
Lets take this behavior under consideration while building our model.



Lets build a model in the form that will answer the following question: Predict the number of Future thoughts per subject using the descriptive information regarding the big 5 traits + age and gender.

We will start off using the means of the Big 5 per subject, and add a numeric value of the age and a categorical predictor of the gender. I will attempt to fit Poisson model and evaluate the following. Next I will add and define the offset of the number of observations per Subject into the model. Last, I will fit a Stepwise AIC model to perform a proper model selection and filter the remaining contributing columns.

First off we will observe the histogram of the target variable, and a heatmap of the pearson correlations between the averages of the Big 5. Bellow are the plots. We can see from the heatmap that there is a slightly strong negative correlation between N and the other 4 traits , in particular with A. We would expect the model selection to remove one of these two due this finding and inner correlation. This is interesting that there is such a large absolute correlation between N and A (neuroticism and agreeableness). From the definition of the two traits , we can see that these traits sort of contradict each other - one being friendly/compassionate , and the other sensitive/nervous) which makes sense that these would be negatively correlated - meaning , the more neurotic the less friendly and sensitive.



To perform the model selection I decided to run a Step AIC (both directions) on a Poisson model , starting off with the former 7 predictors, with the intention to reduce the dimension by selecting the optimal fit in terms of predictors using the factor of AIC.

Below are the steps of the model selection, following with the final selected models representation.

From the results Below we can see that the model selection has dropped the Gender and N predictors , and has resulted in a pretty significant prediction model. We can see that with removing the following 2 predictors , we can see a slight decrease in the AIC.

Its interesting to see that the Z value for the age is negative , and significant reflecting on a negative relation to the predictor (lower the age , the more people think of the future).

Another interesting finding is the relation between “O” and the target , where O represents Openness to experiences. We find that the test’s results indicates that the higher the grade the observant gives itself to this predictor (O) , the more they think of the future. One of the more interesting facts that this trait includes is active imagination , which we can derive a strong connection to future thinking.

On the other hand , a negative relationship we see is with predictor “C” , being Conscientiousness. Under this definition , we see this as a trait of carefulness and deliberation. This is surprising, since we would expect that people that are careful and think carefully before they act would have a high relationship with the number of times they thought of the future.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-16.2	0.171	-95	0
A	0.0512	0.0208	2.46	0.014
C	-0.0669	0.0207	-3.23	0.00123
E	0.0454	0.0152	2.99	0.00283

	Estimate	Std. Error	z value	Pr(> z)
O	0.186	0.0228	8.16	3.37e-16
age	-0.00709	0.00235	-3.01	0.00261

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	7104 on 476 degrees of freedom
Residual deviance:	6991 on 471 degrees of freedom

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	469	6990	8384
- N	1	0.121	470	6990	8382
- sex	1	1.24	471	6991	8381

I decided to look at the selected model and compare to the default one, using an Anova to test the deviance of the two models. We can see a result that we will not reject the null hypothesis , and most likely treat these two models with the same deviance.

Table 12: Anova offset W/O Step AIC

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
469	6990	NA	NA	NA
471	6991	-2	-1.36	0.507

Next we will check out the model of the quasi poisson. As we can see , there can be an assumption that there is overdispersion in the data. I also performed an overdispersion test to verify this. We can see that the dispersion ratio is not close to 1 (actual result is 53696), and since its significantly larger than one , it indicates an over dispersion. We also can identify this by the very low P value (~ 0).

Also ,below are 3 histograms. these plots represent the distribution of the values of the following:

- Number of observations per class of value of the future thinking target.
- Histogram of the number of observations.
- Histogram of the ratios between # of future thoughts and # of total times asked.

By looking at these three plots we can get a full picture of the targeted variable and its relation to the offsetting of the number of times questioned.

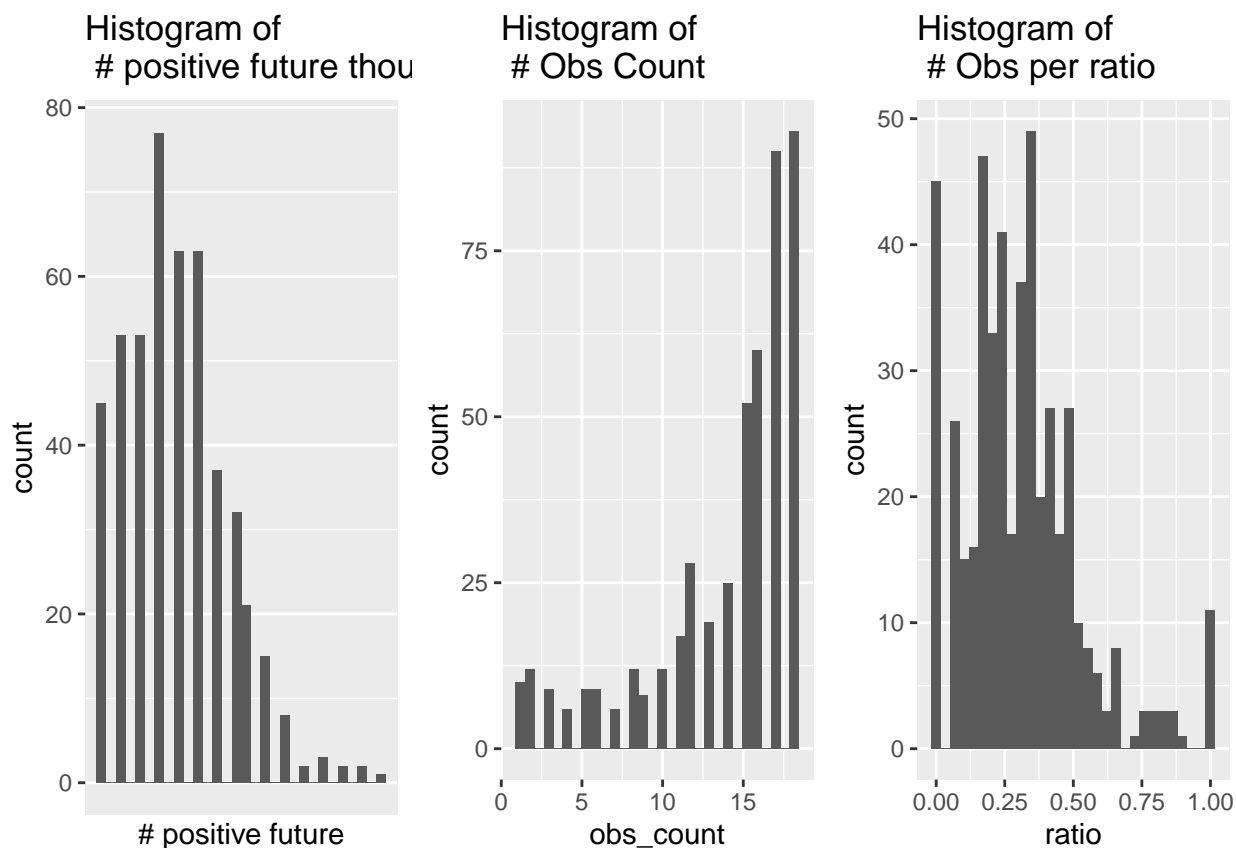
From the left histogram we can see that there is a skewed plot with a dense left side. Also we can see that there is a slight inflation of zeros. This represents our target variable and we can see here a clear overdispersion of the data.

The middle plot represents the amount of times the patients were asked in total. This is interesting since we would expect this plot to behave the same way as the former (left) plot. Instead we see an emphasis to the right with a small left tail, showing that most patients were asked more than 10 times.

The right plot represents the ratio between the two . This is more gaussian , though we can see that the majority have a lower ratio than 0.5. Another point here is that there are about 10 bservations that have a 100% ratio , meaninig that they thought of the future everytime sampled. This represents $\sim 2\%$ of the sampled patients.

Following these three plots we will attempt a quasipoisson model and compare the results. First off ,we can see that the estimates stay the same (by definition) and due to the factorization of the quasi-poisson model , the std errors start to grow , and T values become insignificant. This is as expected due to the overdispersion of the data. We can also see that the dispersion parameter is 53756.73 which is significantly high.

We are assuming that the fitted values will stay the same and that the only effect will take place in the factorization which we can see from the change in the std error.



```
## # Overdispersion test
##
##      dispersion ratio =      53695.739
##      Pearson's Chi-Squared = 25183301.423
##      p-value =          < 0.001
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-16.2	39.7	-0.41	0.682
A	0.0512	4.83	0.0106	0.992
C	-0.0669	4.8	-0.0139	0.989
E	0.0454	3.53	0.0129	0.99
O	0.186	5.3	0.0352	0.972
age	-0.00709	0.546	-0.013	0.99

(Dispersion parameter for quasipoisson family taken to be 53756.73)

Null deviance:	7104 on 476 degrees of freedom
Residual deviance:	6991 on 471 degrees of freedom

Summary - Q2

To sum up this research , we attempted to predict the number of times a patient has future thoughts, using the Big 5 traits plus some demographic information. I tried to use the poisson and Quasipoisson , and the results indicate that the models have room for improvement with additional touches. We can assume that there might be a different fit that may help.

Future Thoughts

We saw here an attempt to fit a poisson model on the given predictors. A couple thoughts for the future would be to possibly engineer the predictors and refine them. The current predictors were the averages based on a small set per subject, where the variance might be high per subject and this may have affected the predictivnes and fitness. With more data per subject this might have been improved.

Another thought would be to add the another dimension of featurization to the model per predictor . By doing this we would create a slightly more complex and lower the generaliztion of the model. We would need to do this without harming the dimensionality of the model.

We can attempt to fit a different type of model that may give us better results (such as a negative binomeal , exponential etc...) and evaluate the different results.

Attached code

Below is the attached code from which I performed the analysis.

```
# import libraries
library(rmarkdown)
library(dplyr)
library(ggplot2)
library(tidyr)
library(pivottabler)
library(gtsummary)
library(ggpubr)
library(ggfortify)
library(cluster)
library(MASS)
library(lmtest)
library(fBasics)
library(rcompanion)
library(gridExtra)
library(cowplot)
library(kableExtra)
library(haven)
library(tidyverse)
library(rstatix)
library(ggpubr)
```

```

library(lme4)
library(reshape2)
library(kableExtra)
library(pander)
library(performance)
library(pROC)

# reading the data and storing in df
#setwd("School/courses/applied_stats/p2_2")

path = file.path( "ETT_ESM_Study1.sav")
df = read_sav(path)

# convert time to numeric
df$TIME_S <- as.numeric(df$TIME_S)

myvars <- c("TIME_S", "DAY", "Subject","future")
newdata <- df[myvars]

par(mfrow=c(1,2))

newdata$idu <- row.names(newdata)
# if row numbers are integers (most likely!)
newdata$Observations <- as.numeric(row.names(newdata))

# plot scatter of time points
p1 <- ggplot(newdata, aes(x=Observations, y=TIME_S )) + geom_point() + ggtitle('Scatter of TIME_S points')

number_of_missing <- sum(is.na(newdata$TIME_S))

# find all observations that are below 3500 - outliers (3 of them)
index <- newdata$TIME_S < 3500
# round them to the next day to get better spread
newdata[index,]$TIME_S <- newdata[index,]$TIME_S + 86400

#plot number of observations pe day
p2 <- ggplot(newdata, aes(x=DAY)) + geom_histogram() + ggtitle('Histogram of # Obs per Day')

number_of_unique_subjects <- length(unique(newdata$Subject))

grouped_by_subject <- newdata %>%
  group_by(Subject) %>%
  summarise_at(vars(TIME_S), funs(mean(., na.rm=TRUE)))

newdata$FDay <- factor(newdata$DAY)
# Plot Subject ~ Time split by Days (colors)
p3 <- ggplot(newdata, aes(x=Subject, y=TIME_S, color=FDay, shape=FDay)) +
  geom_point() + ggtitle('Subject~Time split by Days
  \n (colors per day) ')

```

```

# boxplot of 3 days
p4<- ggboxplot(
  data =newdata, x = "DAY", y = "TIME_S",
  fill = "DAY", palette = "npg", legend = "none",
  ggtheme = theme_pubr()
)+ggtitle('Boxplot of Day ~ Time')

grid.arrange(p1, p2,p3,p4,ncol = 2, nrow = 2)

# run anova tests on time~day and future~day
res.aov_time <- newdata %>% anova_test(TIME_S ~ DAY)
res.aov_future <- newdata %>% anova_test(future ~ DAY)

# plot out the anova tests
panderOptions("digits", 3)
pander(res.aov_time,caption = "Anova Test TIME_S ~ DAY")
pander(res.aov_future,caption = "Anova Test future ~ DAY")

# standardize the time data
newdata$TIME_S_standardized <- scale(newdata$TIME_S)
# factorize the subject data
newdata$Subject = factor(newdata$Subject)

# fit a Logistic regression on dataset
fit.lr <- glm(future ~TIME_S_standardized, data = newdata, family = binomial)
pander(summary(fit.lr))

# evaluate simple prediction on same set
probabilities <- fit.lr %>% predict(newdata, type = "response")
par( mfrow = c(1,1) )
#hist(probabilities)
qplot(probabilities, bins = 50, main ='Histogram of projected values for predictive model') # 'qplot' a

# calculate the accuracy of the model
list_of_thresholds = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)
accuracies = c()
for(i in list_of_thresholds){

newdata$model_prob <- predict(fit.lr, newdata, type = "response")
newdata <- newdata %>% mutate(model_pred = 1*(model_prob > i) + 0,
                             visit_binary = 1*(future == 1) + 0)
newdata <- newdata %>% mutate(accurate = 1*(model_pred == future))

accuracies <- c(accuracies, sum(newdata$accurate)/nrow(newdata))
}

df_thresholds <- data.frame(thresholds = c(list_of_thresholds),accuracy = c(accuracies))

```

```

ggplot(df_thresholds, aes(x=list_of_thresholds, y=accuracies)) +
  geom_point()+
  geom_smooth()+
  ggtitle('Accuracy Per Threshold') # for the main title

ROC_obj <- roc(newdata$future,newdata$model_prob,
  smoothed = TRUE,
  ci=FALSE,
  ci.alpha=0.9,
  stratified=FALSE,
  plot=TRUE,
  auc.polygon=TRUE,
  max.auc.polygon=FALSE,
  grid=TRUE,
  print.auc=TRUE,
  show.thres=FALSE)
sens.ci <- ci.se(ROC_obj)
#par( mfrow = c(1,1) )

plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")

#grid.arrange(p1,ncol = 1, nrow = 1)

#mean(probabilities)

# run non standardized glm (for internal comparisons)
mod.0 <- glm(newdata$future ~newdata$TIME_S, family = binomial)
with(summary(mod.0), 1 -deviance/ null.deviance)

# plot results for glm
#par(mfrow=c(2,2))
#plot(fit.lr)

# split to train and test
train_index <- sample(1:nrow(newdata), 0.8 * nrow(newdata))
test_index <- setdiff(1:nrow(newdata), train_index)

# Build X_train, y_train, X_test, y_test
X_train <- newdata[train_index,'TIME_S']
y_train <- newdata[train_index, "future"]

X_test <- newdata[test_index, 'TIME_S']
y_test <- newdata[test_index, "future"]

# fit on train set
fit_train <- glm(y_train$future ~X_train$TIME_S, family = binomial)

```



```

# project on test set
probabilities <- fit_train %>% predict(X_test, type = "response")
pred_binary <- probabilities>.5
# mean(pred_binary == y_test)

# fit a Logistic regression on dataset
fit.enhanced_lr <- glm(future ~TIME_S_standardized + DAY, data = newdata, family = binomial)
pander(summary(fit.enhanced_lr))

pander(anova(fit.enhanced_lr, fit.lr),caption='Anova Test - Enhanced model vs LR')

probabilities <- fit.enhanced_lr %>% predict(newdata, type = "response")
ROC_obj <- roc(newdata$future,probabilities,
  smoothed = TRUE,
  ci=FALSE,
  ci.alpha=0.9,
  stratified=FALSE,
  plot=TRUE,
  auc.polygon=TRUE,
  max.auc.polygon=FALSE,
  grid=TRUE,
  print.auc=TRUE,
  show.thres=FALSE)
sens.ci <- ci.se(ROC_obj)
#par( mfrow = c(1,1) )

plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")

# perform mixed model
mixed_model <- glmer(
  future ~ TIME_S_standardized + (1|Subject/DAY),
  data = newdata,
  family = binomial(link = "logit")
)
mixed_model

# plot anova of mixed vs GLM
pander(anova(mixed_model, fit.lr),caption='Anova Test - Mixed model vs LR')

cols_big_five <- c('Subject','age','sex','A','C','E','N','O','future')
big_five <- df[cols_big_five]

big_five$sex <- factor(big_five$sex)

# group by subject and view mean
a <- big_five %>% group_by(Subject)%>% summarise(mean=(mean('E')))

# group by subject and view mean

```

```

means <- aggregate(big_five, list(Subject = big_five$Subject), mean)
# select the relevant columns
means_sample_Set <- means[c('A','C','E','N','O')]
# histograms of means
means_sample_Set %>% select_if(is.numeric) %>% gather(cols, value) %>% ggplot(aes(x = value)) + geom_l

# build the testing set
# y variable (counts of futures)
sums<- aggregate(big_five[c('Subject','future')], list(Subject = big_five$Subject), FUN=sum)

# number of observations per subject
lens<- aggregate(big_five[c('Subject','future')], list(Subject = big_five$Subject), FUN=length)

#create the X var
x <- means[c('Subject','A','C','E','N','O')]
y <- sums[c('Subject','future')]
# merge the two together for assignment and alignment
merged <- merge(x,y,by="Subject")

# create mini df with age and gender to append
age_and_sex <- big_five[c('Subject','age','sex')]
age_and_sex <- age_and_sex %>% dplyr::distinct(age_and_sex$Subject , .keep_all = TRUE)

# append to the df the sex and age
merged <- inner_join(merged,age_and_sex,by="Subject")

# append the num of observations per subject for ffset
lens <- lens[c('Subject','future')]
merged <- merge(merged,lens,by="Subject")

# rename columns
merged <- merged %>%
  rename(
    future_sum = future.x ,
    obs_count = future.y
  )

# drop observations with null values
merged <- drop_na(merged)
#split again to predictors and y
x <- merged[c('A','C','E','N','O','age','sex')]
y <- merged$future

# create dataset with all columns
dd <- merged[c('A','C','E','N','O','age','sex','future_sum','obs_count')]

# plot histogram of counts of futures
p1<-ggplot(merged, aes(x=future)) + geom_histogram() +ggtitle('Histogram of # of True future thoughts ')

# funtion to get upper triangle of corr matrix

```

```

get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}

# fun to reorder the matrix
reorder_cormat <- function(cormat){
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}

# correlation matrix
cors <- cor(merged[c('A','C','E','N','O')], method = c( "pearson"))
upper_tri <- get_upper_tri(cors)
upper_tri <- round(upper_tri, digits = 2)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

# plot heat map of corr matrix
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
    midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.5, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
    title.position = "top", title.hjust = 0.5))

# print out the summary of the poisson regression with a log offset of the counts per observation
m0 <- glm(future_sum ~ A+C+E+N + O+ age+sex + offset(obs_count),poisson, dd)
# Stepwise Poisson model - both ways

step.model <- stepAIC(m0, direction = "both",
  trace = FALSE)

```

```

pander(summary(step.model))
pander(step.model$anova)

# graphic representation of predictions vs actual
predicted_vals <- predict(step.model,dd,type = "response")
plot(dd$future_sum,predicted_vals,xlab="Actual Number of Future Obs", ylab="Predicted Number of Future Obs", col="red")
abline(a=dd$future_sum, b=predicted_vals, col="red")

# anova stepwise model vs classic poisson
pander(anova(m0, step.model, test="Chisq"), caption="Anova offset W/O Step AIC")

# histogram of num of obs per day
dd$ratio <- dd$future_sum / dd$obs_count
p1 <-ggplot(dd, aes(x=future_sum)) + geom_histogram() +ggtitle('Histogram of \n # positive future thoughts')

p2 <-ggplot(dd, aes(x=obs_count)) + geom_histogram() +ggtitle('Histogram of \n # Obs Count')

p3<- ggplot(dd, aes(x=ratio)) + geom_histogram() +ggtitle('Histogram of \n # Obs per ratio')
grid.arrange(p1, p2,p3, ncol = 3, nrow = 1)

# run overdispersion test
#pander(check_overdispersion(m0), caption = 'Over Dispersion Test')

l<- check_overdispersion(m0)
print(l)
#c(l$dispersion_ratio,l$chisq_statistic,l$p_value)

# run quasipoisson
m2 <- glm(future_sum ~ A+C+E + 0+ age + offset(obs_count),quasipoisson, dd)
pander(summary(m2), caption ="Quasi-Poisson Model")

# use anova test to compare stepwise model to quasi
#pander(anova(m2,step.model), caption ="Anova test - Quasi vs Stepwise")

```