EDA Project

Joe Stoica, Conor Devins, Geno Kim

29 November, 2018

Statement of goals.

What questions are you trying to address?

Some preliminary questions

1. What is the relationship between perceived side effects and overall satisfaction?

2. What is the relationship between perceived effectiveness and overall satisfaction?

3. What is the relation between perceived side effects and perceived effectiveness?

a. Though intuitively one may assume that more severe side effects would be associated with lower perceived effectiveness, it is also possible that effectiveness-sideeffect trade-offs are nonetheless perceived as a net benefit to the patient (where perceived benefits of the drug for daily life quality outweigh their negative side effects). Are there certain classes of drugs or specific drugs where the relationships are different or does it hold true across all drugs?

4. What is the breakdown of the types of drugs used for treating depression (or other mood disorders).

a. This can be a simple bar chart showing the relative proportions of serotonin-reuptake inhibiting drugs (SRIs), norepinephrine-mechanism drugs (NRIs), Amphetamines (AMP), etc…

5. How do certain classes of drugs compare in terms of their perceived side effects, effectiveness and overall satisfaction?

a. SRIs and NRIs are probably the most popular of options, but are they actually superior along these dimensions of perception?

b. Can probably do Multinomial regression to address this question; if we do this, may be worth excluding some drug classes that have very few cases, and focusing on those classes that represent a sizeable portion of the data. These analyses are covered in Lecture 25, slide 28.

Why do you care?

Why should we care?

Description of your data.

In addition to graphical displays, this should include verbal descriptions of what your variables are, who the individuals in your data from, and how they were selected/sampled. If you have many variables, you don't have to describe all of them, just pick out some key ones.

The data we chose to use is the Drug Review Dataset from the UCI Machine Learning Repository. The data focuses on pharmaceutical drug users ratings and reviews of certain drugs they've taken.

The data was compiled by gathering the reviews from druglib.com, which is "a comprehensive drug database organized by relevance to specific drugs." (TODO make footnote for http://www.druglib.com/). It allows people who have used a specific drug to rate the drug based on their experience.

(TODO explain how we removed people and have that code)

(TODO change observations #) Our data has five columns with 369 different observations:

DrugName: the name of the drug

Satisfaction: Rating (10-point scale, 10 being highest satisfaction)

Effectiveness: 1 - Ineffective 2 - Marginally Effective 3 - Moderately Effective 4 - Considerably Effective 5 - Highly Effective

Side Effects: 1 - Extremely Severe Side Effects 2 - Severe Side Effects 3 - Moderate Side Effects 4 - Mild Side Effects 5 - No Side Effects

Type: Chemical type of the drug

Answering your questions.

This is the most important criterion. It will probably include (but is not limited to) fitting a statistical model or models of some kind, and showing that these models tell you something of interest. You should do the following (not necessarily in this order):

(TODO what's our main question)

1. filter all rows for which the condition is for anything related to "depression" so basically just look for the string "depression" in that column where condition is listed. this includes depression + any other comorbidities
2. delete the comments column where people describe their experience w/ the drug etc.
3. for the drug names (e.g. lexapro etc.) any drug that has a hyphenated add-on such as drugname-xr or whatever, just remove that hyphenated part so you're just left with "drugname", for the simplicity we're gonna ignore different versions of the same drug (paxil-xr vs paxil) just treat paxil-xr as belonging to paxil

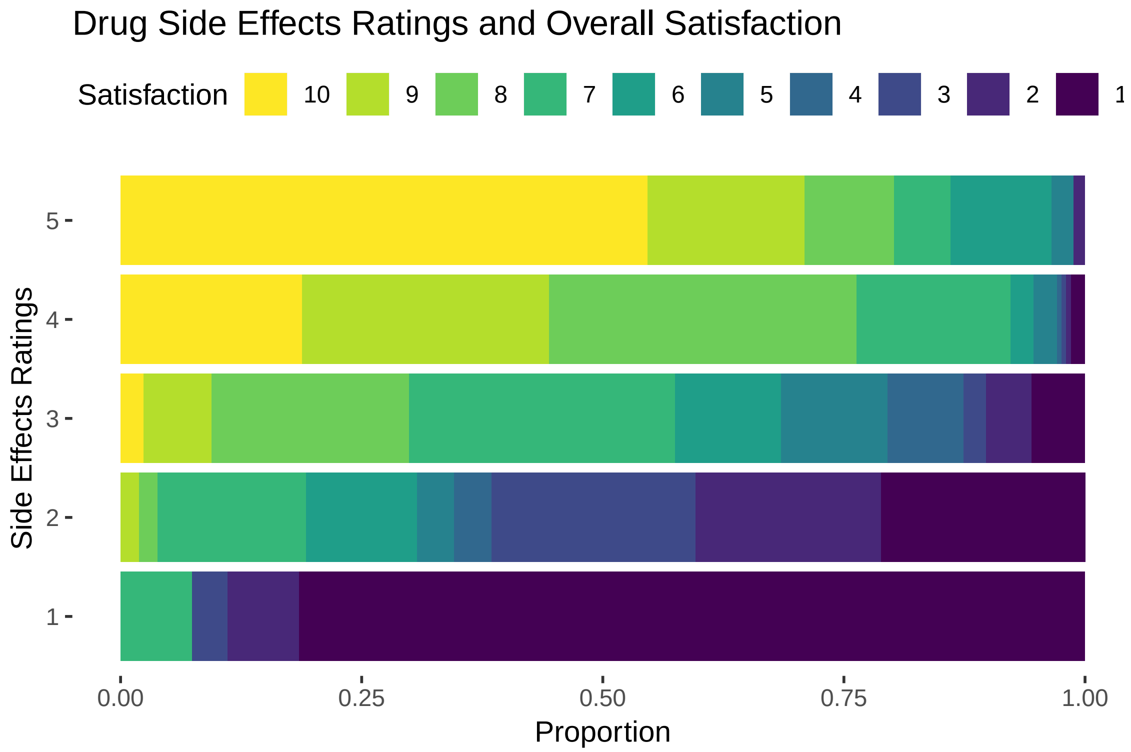
and finally, the column that i added but was not part of the original data set was drug pharmacology (e.g. SRI, NRI, Amphetamines, etc.)

i basically had to manually make a table of contents for which family each drug name belonged to but u won't need to since you can find out that grouping from the csv file i uploaded all those steps summarize what was done to get the data in the form it is now.

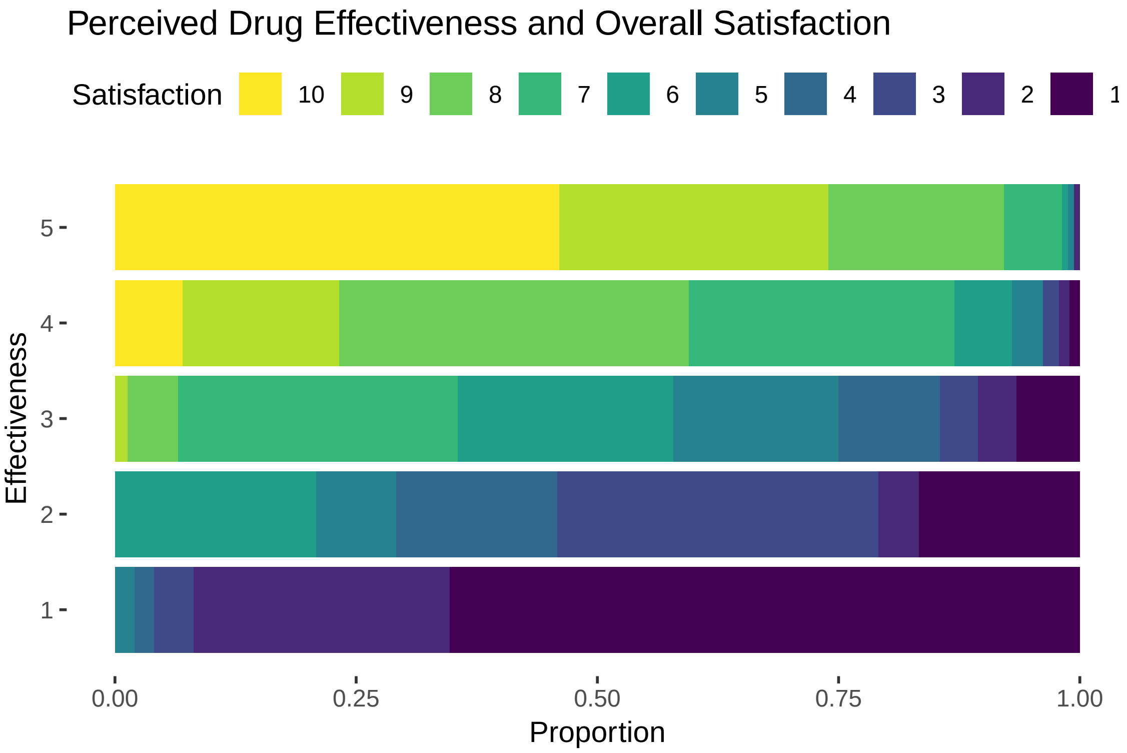
install.packages('SentimentAnalysis') library(SentimentAnalysis)

```
## [1] "X1"                "urlDrugName"       "rating"
## [4] "effectiveness"     "sideEffects"       "condition"
## [7] "benefitsReview"    "sideEffectsReview" "commentsReview"
```

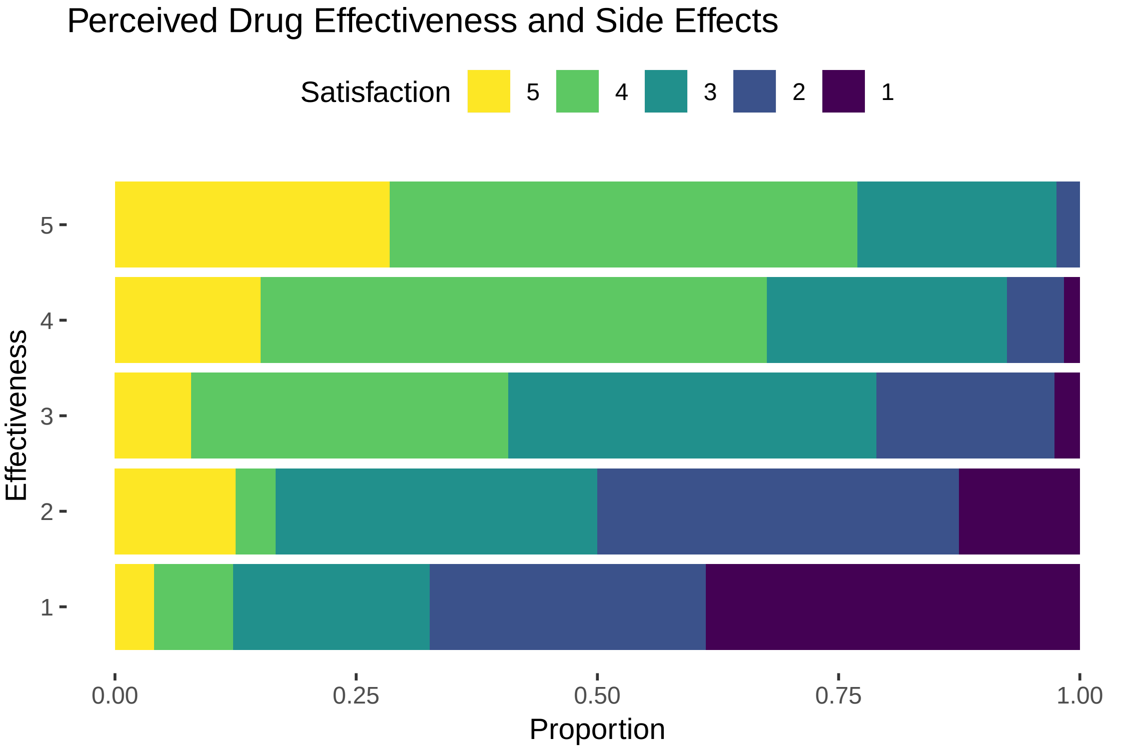
```
## [1] "X1"                "urlDrugName"       "rating"
## [4] "effectiveness"     "sideEffects"       "condition"
## [7] "benefitsReview"    "sideEffectsReview" "commentsReview"
```



The above plot shows that the worse the side effects are, the least satisfied the subjects were. (TODO more explanation maybe)



The above plot shows that the more effective, the more satisfied the subjects were. (TODO more explanation maybe)



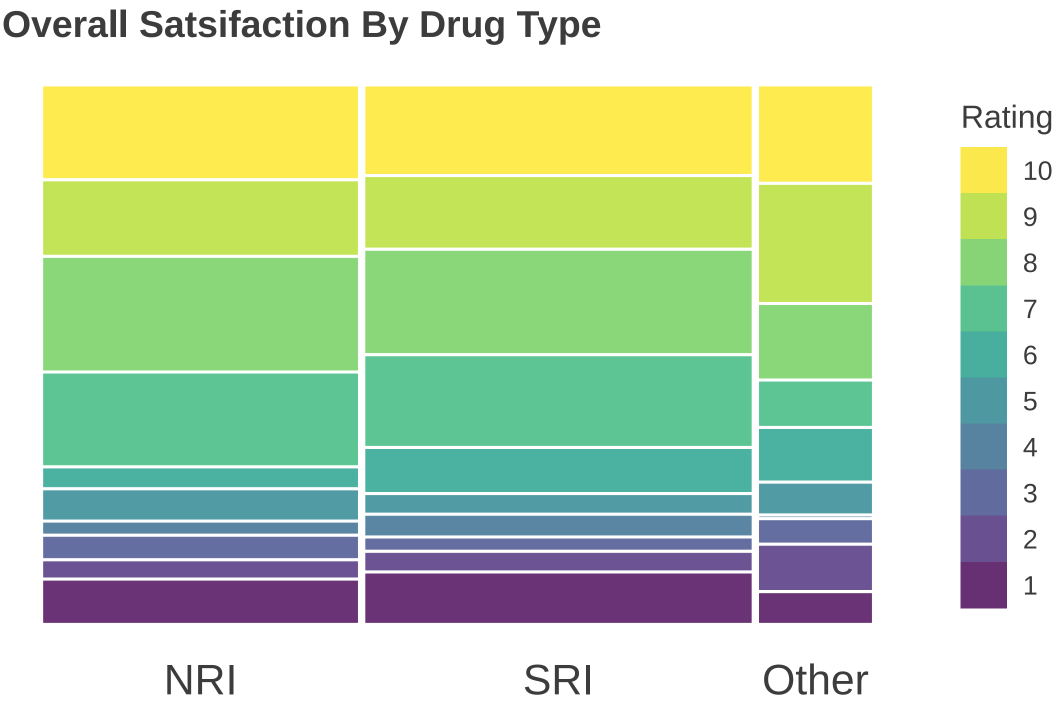
State answers to your questions;

Describe how you came to these answers;

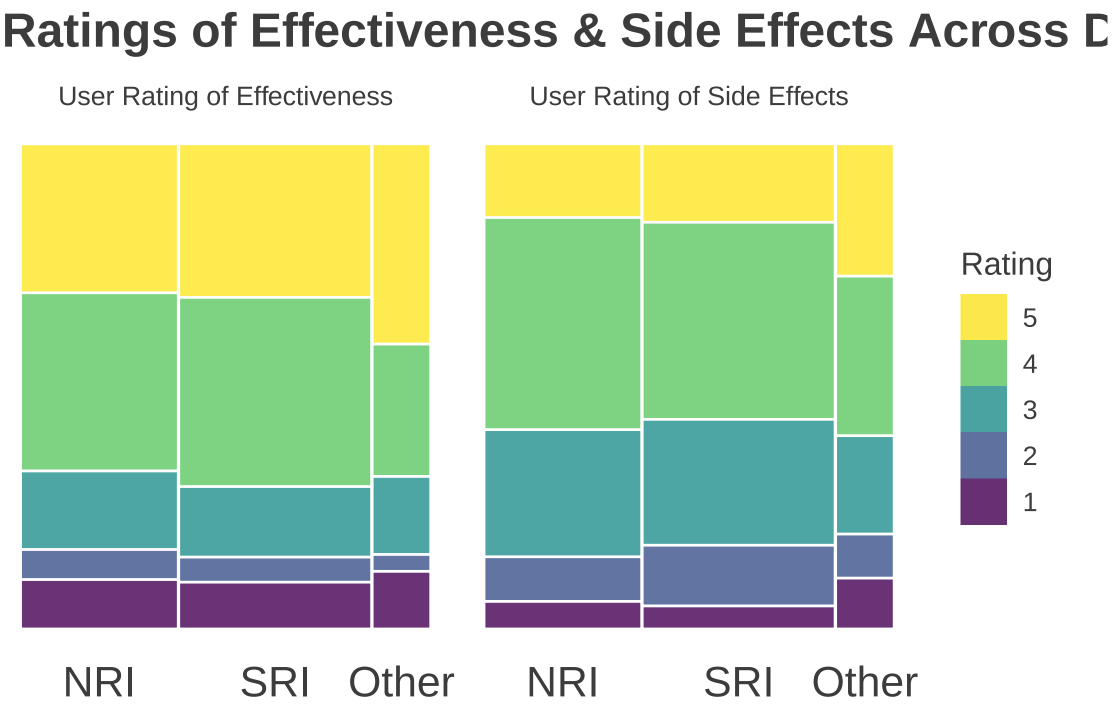
Explore the implications to your answers. For example, if your answer is a non-trivial model, plot the fit and describe what's going on in words.

Identification of work left to do/limitations.

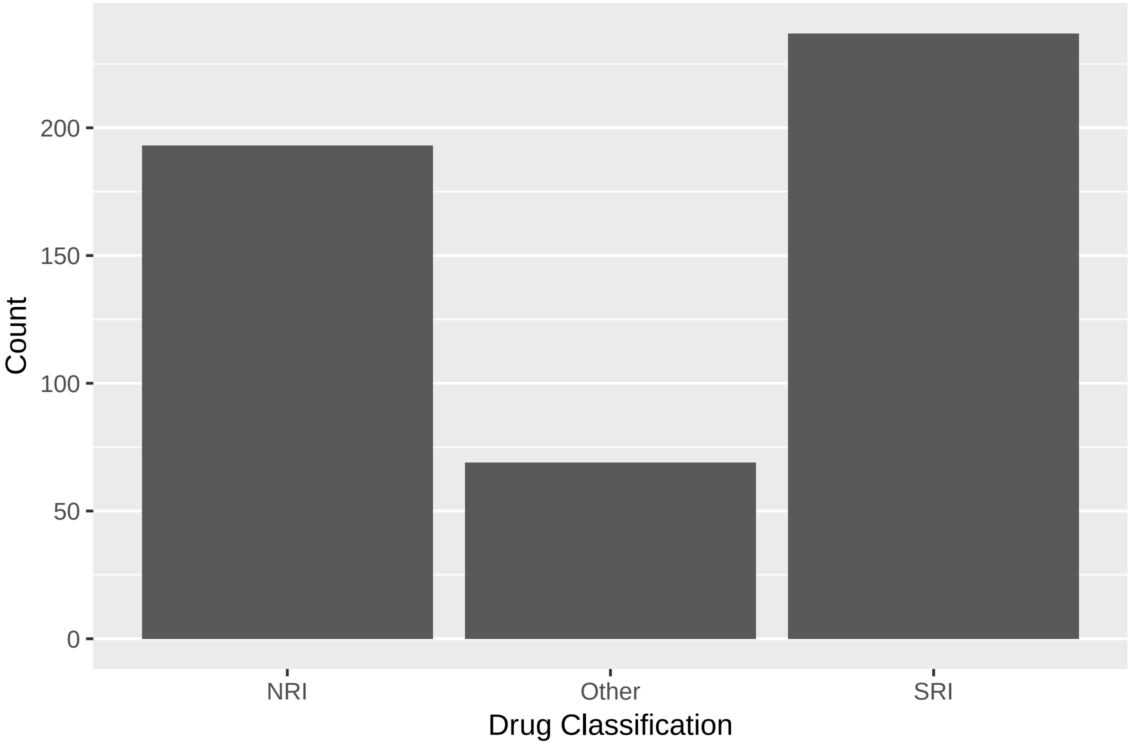
It's EDA, so we don't require perfection. However, you should have a clear idea of what the imperfections in your work are (what doesn't fit well? what other variables would you really want to know?), and how they could potentially be addressed.

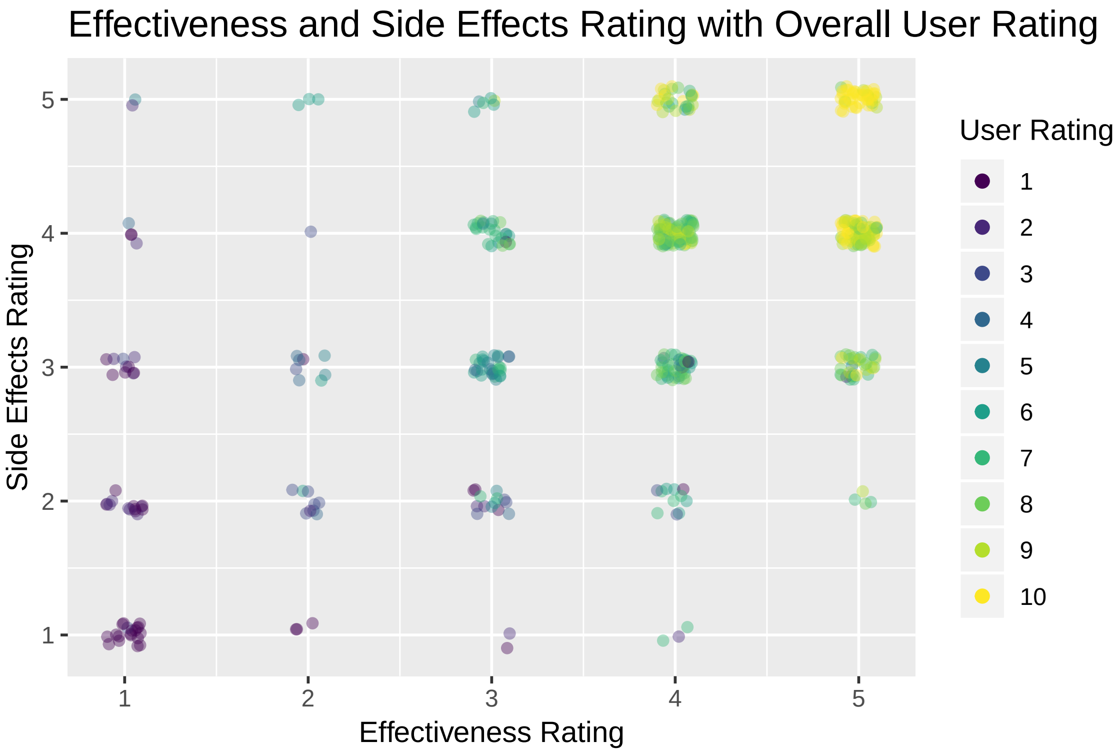


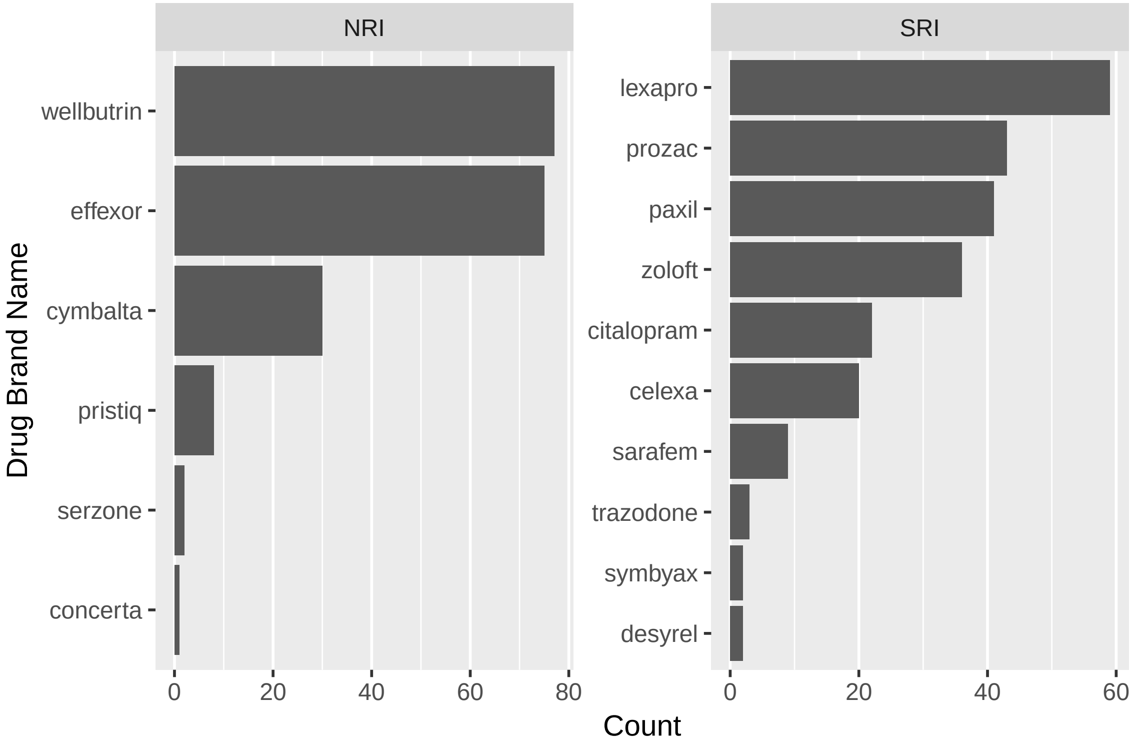
Mosaic plot of user ratings of drug effectiveness and side effects by drug type (SRI vs NRIs)



Ratings of Effectiveness & Side Effects Across D

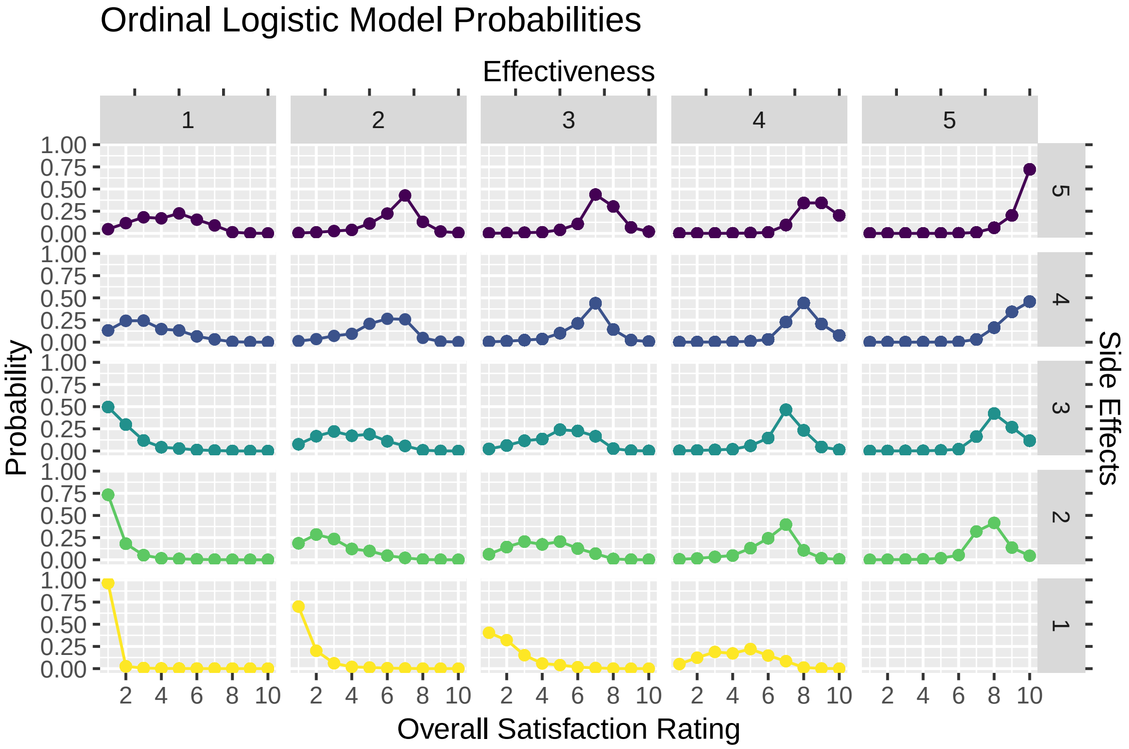




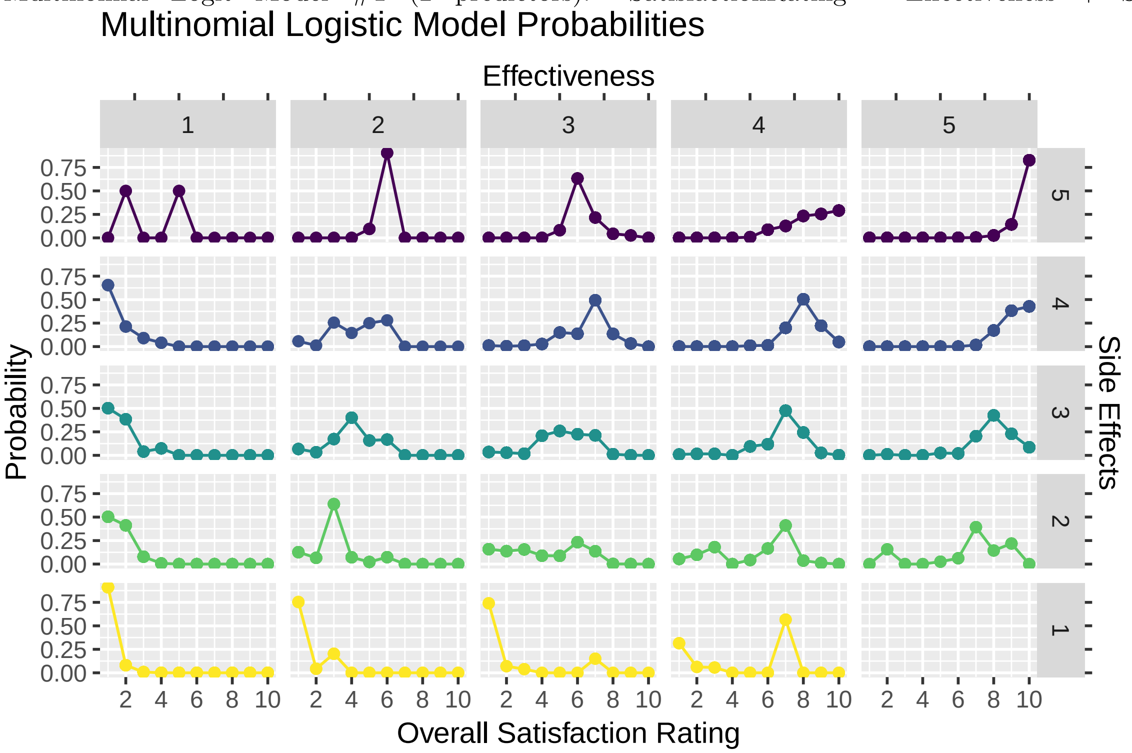


I created three models all trying to predict rating. They all have side effects and effectiveness. The first model only uses those two, the second model takes into consideration drug type, and the third model takes into consideration the drug itself. The first model has the lowest Akaike information criterion (AIC). "[It] is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.(stolen from wikipedia (https://en.wikipedia.org/wiki/Akaike_information_criterion)".

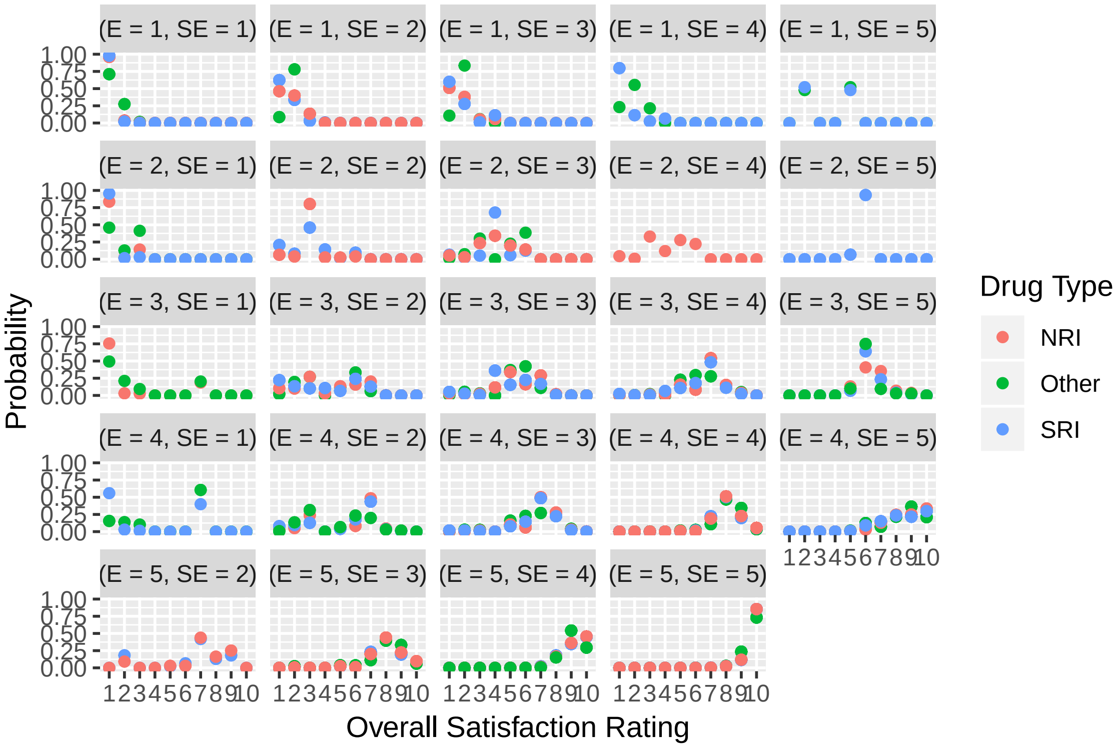
So even though it doesnt have the lowest residual deviance, it is still the best model to use. Adding the sentiment analysis might improve it, however.



Multinomial Logit Model #1 (2 predictors): SatisfactionRating ~ Effectiveness + SideEffects



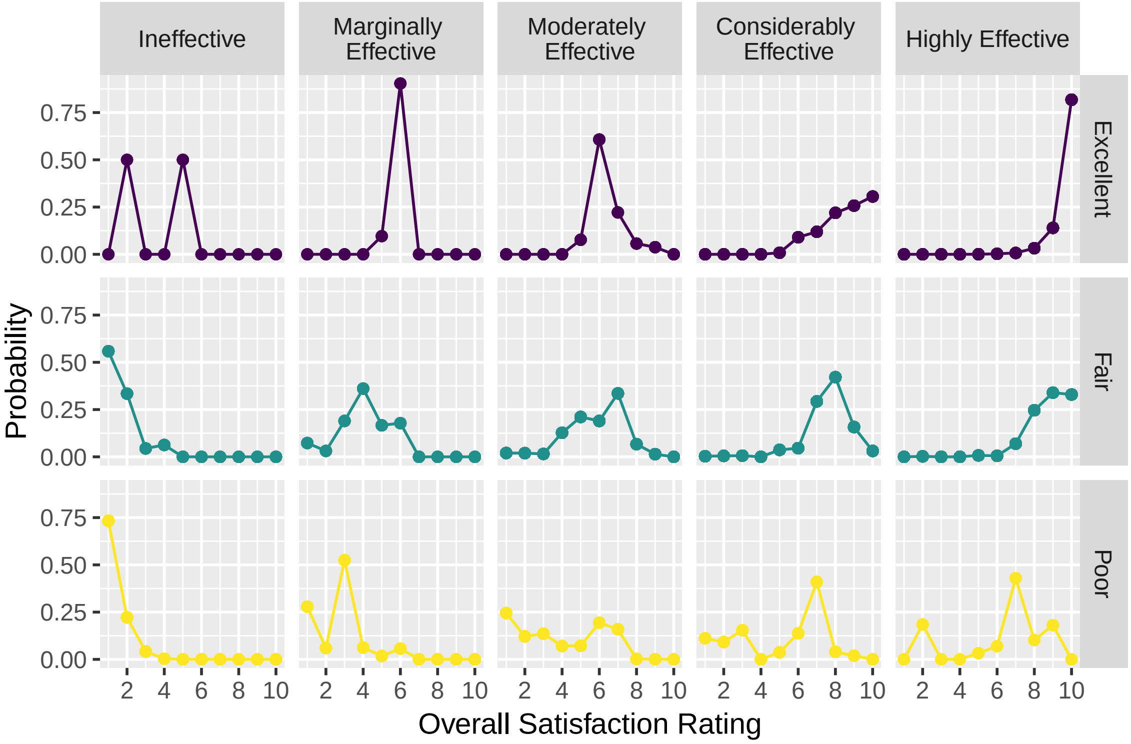
Multinomial Logit Model #2 (3 Predictors): SatisfactionRating ~ Effectiveness + SideEffects + DrugType



Multinomial Logit Model #3 (3 Predictors): SatisfactionRating ~ Effectiveness + SideEffects + DrugBrand

Plot multinomial model fits with fewer category levels

```
##
## Call:
## VGAM::vglm(formula = factor(rating) ~ as.factor(effectiveness) +
##     as.factor(sideEffects), family = "multinomial", data = df2)
##
##
## Pearson residuals:
##                         Min       1Q     Median         3Q    Max
## log(mu[,1]/mu[,10]) -2.1300 -0.06521 -1.304e-02  8.719e-06 17.698
## log(mu[,2]/mu[,10]) -1.5206 -0.12957 -2.535e-02 -5.657e-03 20.816
## log(mu[,3]/mu[,10]) -1.2363 -0.09222 -2.223e-02  3.938e-08 12.689
## log(mu[,4]/mu[,10]) -0.9373 -0.01436  3.939e-08  2.129e-05  3.474
## log(mu[,5]/mu[,10]) -1.8415 -0.21729 -9.015e-02 -1.984e-02 11.494
## log(mu[,6]/mu[,10]) -1.9186 -0.18692 -1.056e-01 -1.215e-02 13.720
## log(mu[,7]/mu[,10]) -2.6233 -0.53348 -1.263e-01  7.401e-09  3.585
## log(mu[,8]/mu[,10]) -2.7857 -0.53309 -2.035e-01  8.049e-09  5.469
## log(mu[,9]/mu[,10]) -2.3709 -0.39107 -2.474e-01  7.612e-09  4.483
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                   -2.1424  3291.9604  -0.001  0.99948
## (Intercept):2                   14.9002  1244.8543      NA       NA
## (Intercept):3                   -1.9972  3095.1555  -0.001  0.99949
## (Intercept):4                   -1.9740  3065.7896  -0.001  0.99949
## (Intercept):5                   14.9002  1244.8541      NA       NA
## (Intercept):6                   -4.7054 11015.3153   0.000  0.99966
## (Intercept):7                   -4.6142 10533.1207   0.000  0.99965
## (Intercept):8                   -3.8325  7191.2084  -0.001  0.99957
## (Intercept):9                   -4.3465  9237.1557   0.000  0.99962
## as.factor(effectiveness)2:1    -17.7369  1998.0387  -0.009  0.99292
## as.factor(effectiveness)2:2    -18.0768  1998.0389  -0.009  0.99278
## as.factor(effectiveness)2:3    -14.2425  1998.0388  -0.007  0.99431
## as.factor(effectiveness)2:4    -13.9489  1998.0389  -0.007  0.99443
## as.factor(effectiveness)2:5     -0.2140  1966.8914   0.000  0.99991
## as.factor(effectiveness)2:6     21.6322 11119.9906   0.002  0.99845
## as.factor(effectiveness)2:7      0.8559 11946.3131   0.000  0.99994
## as.factor(effectiveness)2:8      0.8766  8481.5301   0.000  0.99992
## as.factor(effectiveness)2:9      1.7249 10579.7589   0.000  0.99987
## as.factor(effectiveness)3:1    -19.6701  1519.1622  -0.013  0.98967
## as.factor(effectiveness)3:2    -19.1775  1519.1622  -0.013  0.98993
## as.factor(effectiveness)3:3    -17.3983  1519.1624  -0.011  0.99086
## as.factor(effectiveness)3:4    -15.6322  1519.1625  -0.010  0.99179
## as.factor(effectiveness)3:5     -0.6183  1470.5014   0.000  0.99966
## as.factor(effectiveness)3:6     21.0543 11043.0493   0.002  0.99848
## as.factor(effectiveness)3:7     19.9551 10562.1248   0.002  0.99849
## as.factor(effectiveness)3:8     17.8113  7233.6571   0.002  0.99804
## as.factor(effectiveness)3:9     17.9140  9270.2263   0.002  0.99846
## as.factor(effectiveness)4:1    -37.6879  1309.6778  -0.029  0.97704
## as.factor(effectiveness)4:2    -36.6881  1309.6778  -0.028  0.97765
## as.factor(effectiveness)4:3    -34.5015  1309.6779  -0.026  0.97898
## as.factor(effectiveness)4:4    -51.2314  2568.8583      NA       NA
## as.factor(effectiveness)4:5    -18.5145  1244.8543  -0.015  0.98813
## as.factor(effectiveness)4:6      3.4834 11015.3153   0.000  0.99975
## as.factor(effectiveness)4:7      3.6715 10533.1207   0.000  0.99972
## as.factor(effectiveness)4:8      3.5004  7191.2084   0.000  0.99961
## as.factor(effectiveness)4:9      4.1742  9237.1557   0.000  0.99964
## as.factor(effectiveness)5:1    -57.4085  2684.9841      NA       NA
## as.factor(effectiveness)5:2    -39.8334  1309.6781  -0.030  0.97574
## as.factor(effectiveness)5:3    -54.6724  2757.6572      NA       NA
## as.factor(effectiveness)5:4    -53.8252  2977.9903      NA       NA
## as.factor(effectiveness)5:5    -22.4732  1244.8546  -0.018  0.98560
## as.factor(effectiveness)5:6     -1.0445 11015.3154   0.000  0.99992
## as.factor(effectiveness)5:7     -0.1276 10533.1207   0.000  0.99999
## as.factor(effectiveness)5:8      0.6001  7191.2084   0.000  0.99993
## as.factor(effectiveness)5:9      2.5797  9237.1557   0.000  0.99978
## as.factor(sideEffects)Fair:1    37.5221  3192.0689   0.012  0.99062
## as.factor(sideEffects)Fair:2    19.9663   950.5085      NA       NA
## as.factor(sideEffects)Fair:3    34.8395  2988.7275   0.012  0.99070
## as.factor(sideEffects)Fair:4    35.1659  2958.3104   0.012  0.99052
## as.factor(sideEffects)Fair:5     3.7834     1.1910   3.177  0.00149 **
## as.factor(sideEffects)Fair:6     1.6106     0.6853   2.350  0.01877 *
## as.factor(sideEffects)Fair:7     3.1886     0.6102   5.226 1.73e-07 ***
## as.factor(sideEffects)Fair:8     2.9413     0.4970   5.918 3.26e-09 ***
## as.factor(sideEffects)Fair:9     1.7969     0.3925   4.579 4.68e-06 ***
## as.factor(sideEffects)Poor:1    57.3819  3285.2999   0.017  0.98606
## as.factor(sideEffects)Poor:2    39.1441  1227.6710   0.032  0.97456
## as.factor(sideEffects)Poor:3    54.3744  3088.0996   0.018  0.98595
## as.factor(sideEffects)Poor:4    51.9300  3058.6703   0.017  0.98645
## as.factor(sideEffects)Poor:5    20.0558   871.9965   0.023  0.98165
## as.factor(sideEffects)Poor:6    18.9879   871.9959   0.022  0.98263
## as.factor(sideEffects)Poor:7    19.7973   871.9958      NA       NA
## as.factor(sideEffects)Poor:8    16.8534   871.9963   0.019  0.98458
## as.factor(sideEffects)Poor:9    15.9582   871.9962   0.018  0.98540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  9
##
## Residual deviance: 1327.147 on 4428 degrees of freedom
##
## Log-likelihood: -663.5733 on 4428 degrees of freedom
##
## Number of iterations: 19
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', '(Intercept):5', 'as.factor(effectiveness)4:4', 'as.factor(effectiveness)5:1', 'as.
##
## Reference group is level  10  of the response
```



Multinomial logit model 1 and 2 not very different based on AIC (1380.239 vs 1383.26 respectively), whereas model 3 has lowest deviance but considerably higher AIC (1879.28). Could go with m1, since it has lowest AIC.