

EDA Project

Joe Stoica, Conor Devins, Geno Kim

26 November, 2018

Statement of goals.

What questions are you trying to address?

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

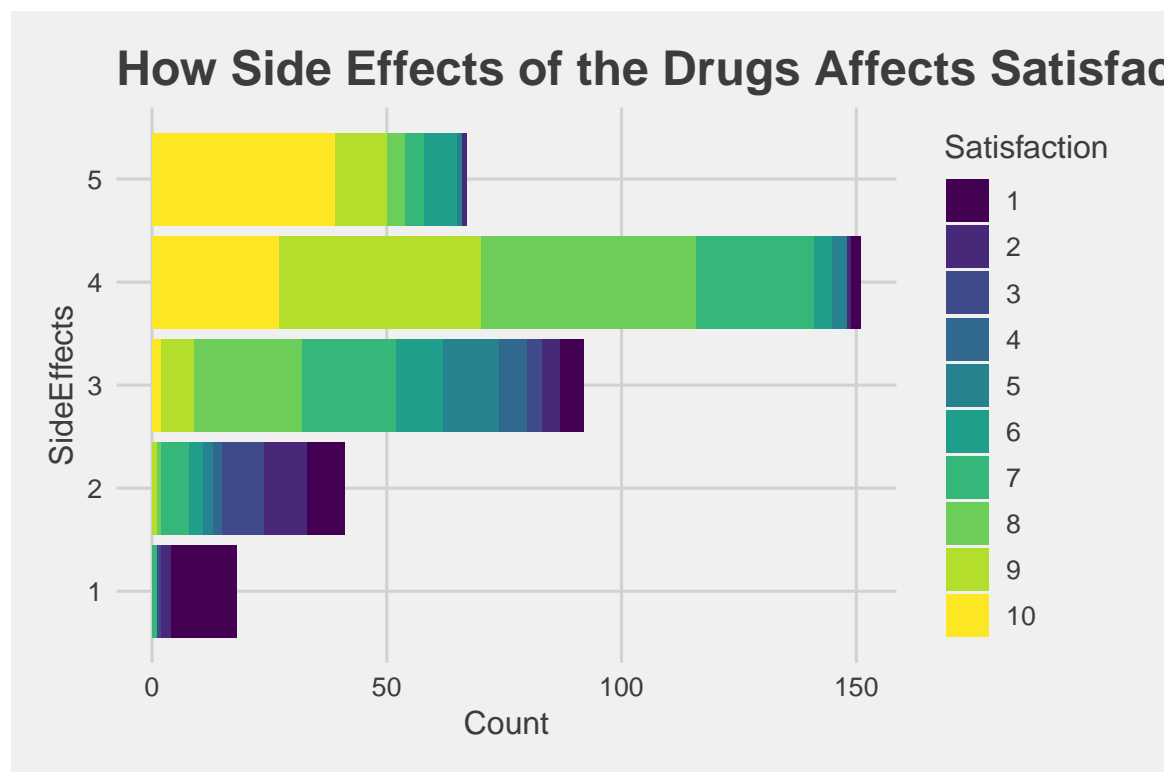
Side Effects: 1 - Extremely Severe Side Effects 2 - Severe Side Effects 3 - Moderate Side Effects 4 - Mild Side Effects 5 - No Side Effects

Type: Chemical type of the drug

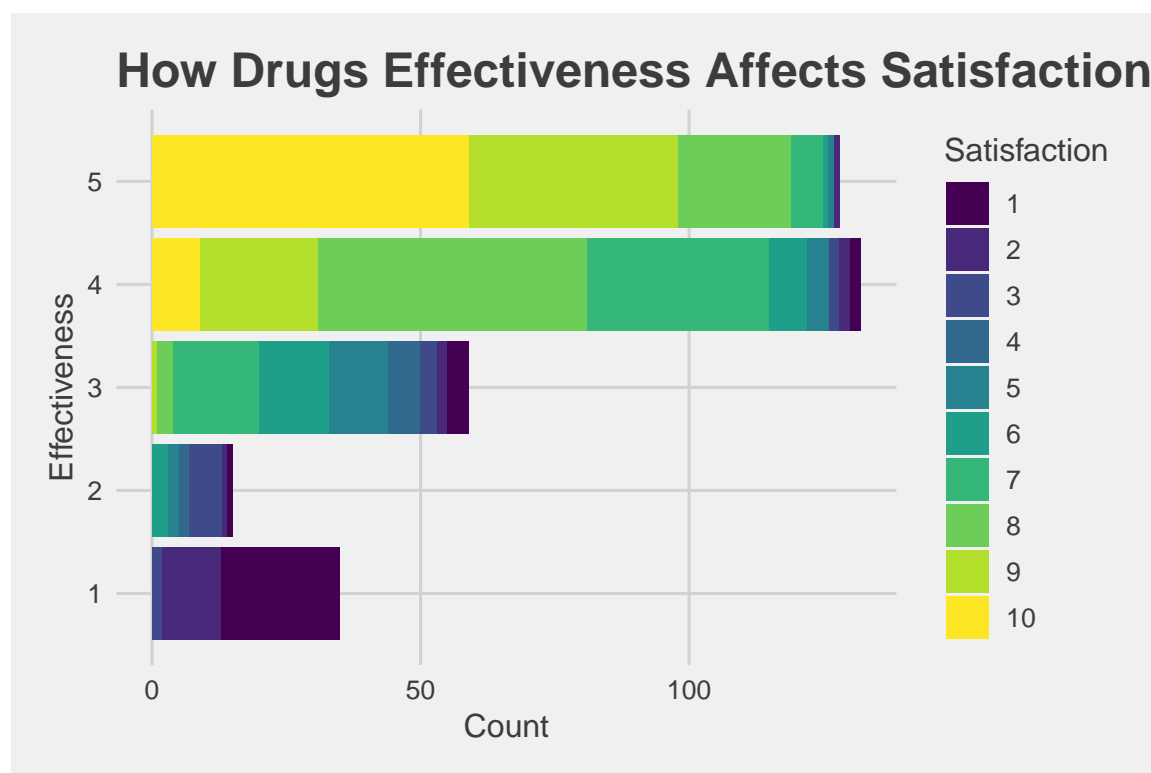
Answering your questions.

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)



The above plot shows that the worse the side effects are, the least satisfied the subjects were.



The above plot shows that the more effective , the least satisfied the subjects were.

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.