Drug Reviews from WebMD

NYCDSA Web Scraping Project

Michael Chin

2/14/2018

Why scrape for WebMD reviews?

- Insights:
  - Public perception
  - Factors affecting drug satisfaction
- Useful for:
  - Drug company marketing
  - Medical care professionals
- Not useful for:
  - People in need of medical attention

Drug reviews Scraped

Cymbalta™ duloxetine

4589 reviews

Lilly

Pristiq

1169 reviews

Pfizer

Pfizer

3047 reviews

Depression Drug Market expected to be valued at 16.8 billion USD by 2020

WebMD review format

From each review we can extract:

- Condition
- Age (category)
- Gender
- Review date
- Treatment time
- Ratings (focus on satisfaction)
- Comment

Total observations scraped: 8303 over 3 medications

Observations & Questions

General description of dataset:

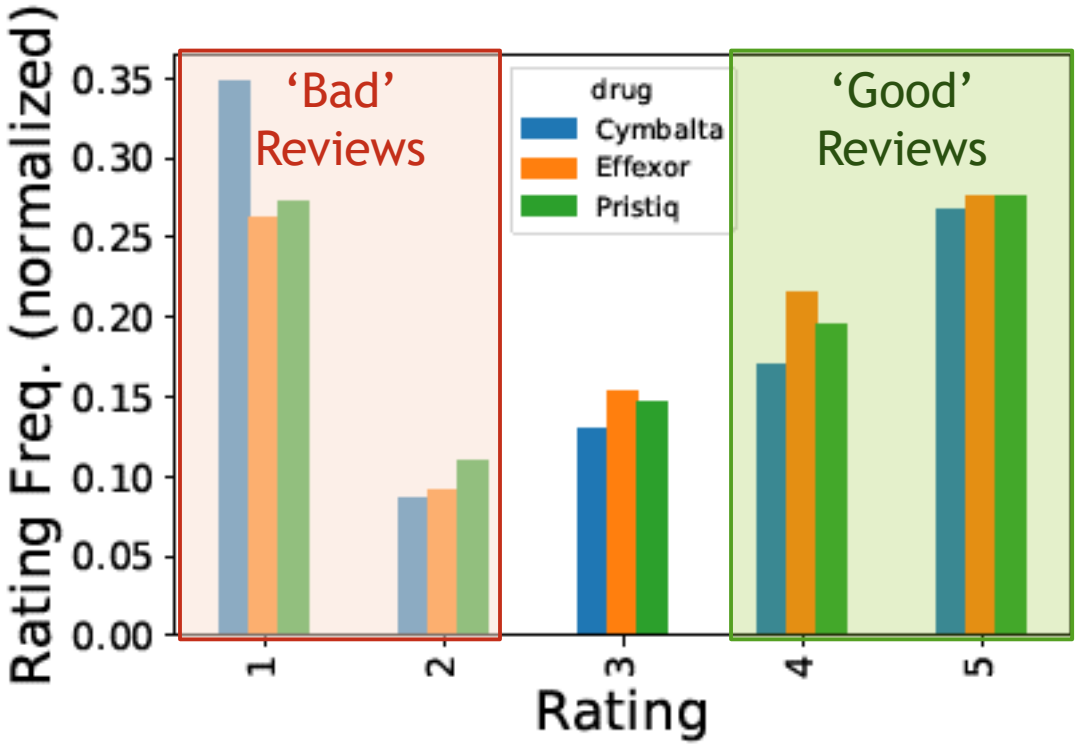
Drug	Most frequent			Avgerage
	Condition	Age range	Gender	Satisfaction
Cymbalta	Depression	45 – 54	Female	2.9 / 5
Effexor	Depression	45 – 54	Female	3.1 / 5
Pristiq	Depression	45 – 54	Female	3.1 /5

Question: Is there really a difference in satisfaction between these three drugs?

Initial Hypothesis

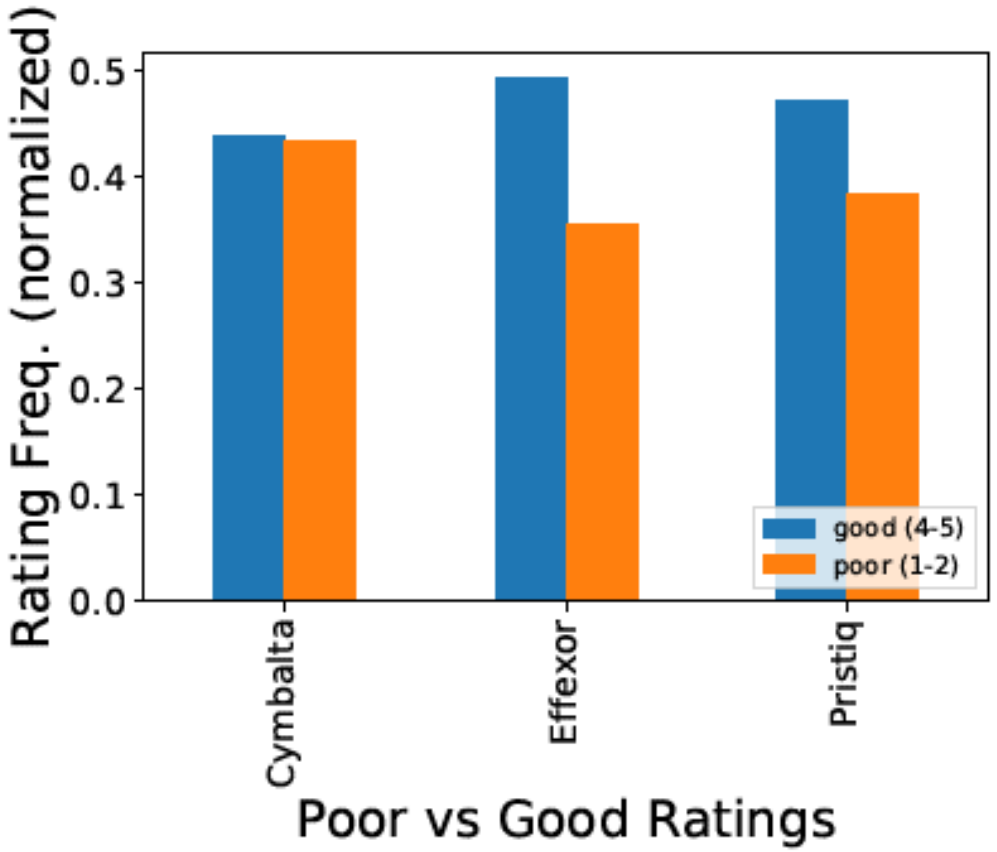
Hypothesis: ‘Satisfaction’ is significantly different between each drug

Problem: ‘Satisfaction’ distribution not normal



Workaround: Bin population into ‘Poor’ (1 – 2) and ‘Good’ (4 – 5) and test frequency of categories

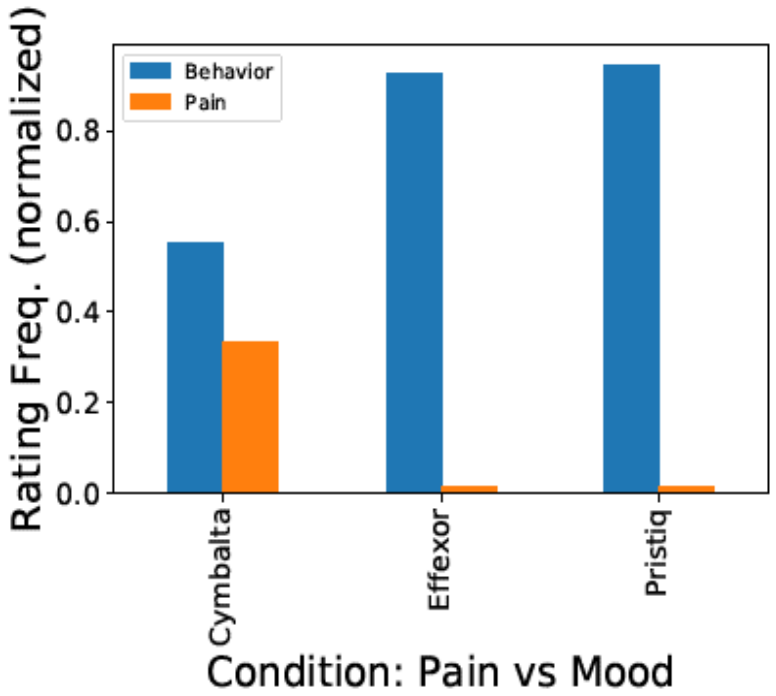
Comparing ratio of ‘Good’ to ‘Bad’ reviews across drugs

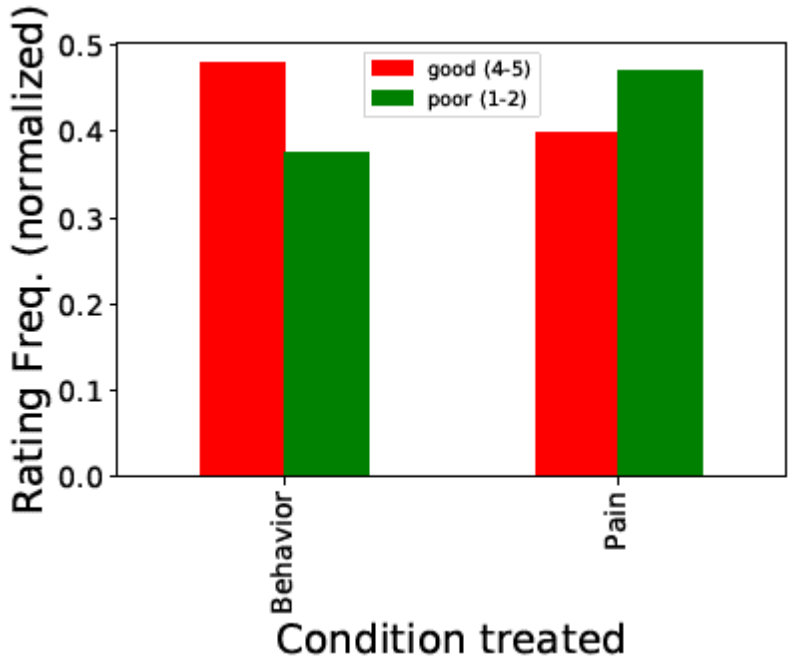


Chi Sq. Stat.	**38.6**
P - value	4.2 e-9

Cymbalta receives more ‘Poor’ reviews than competing products. Why?

A look at conditions being treated

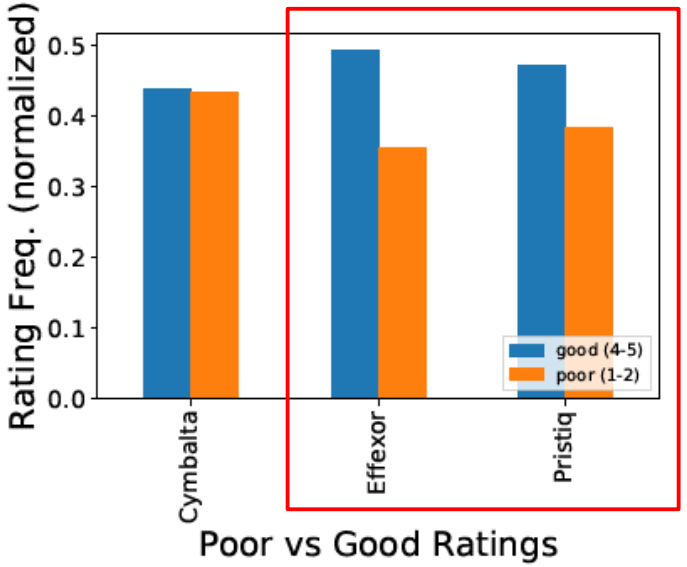


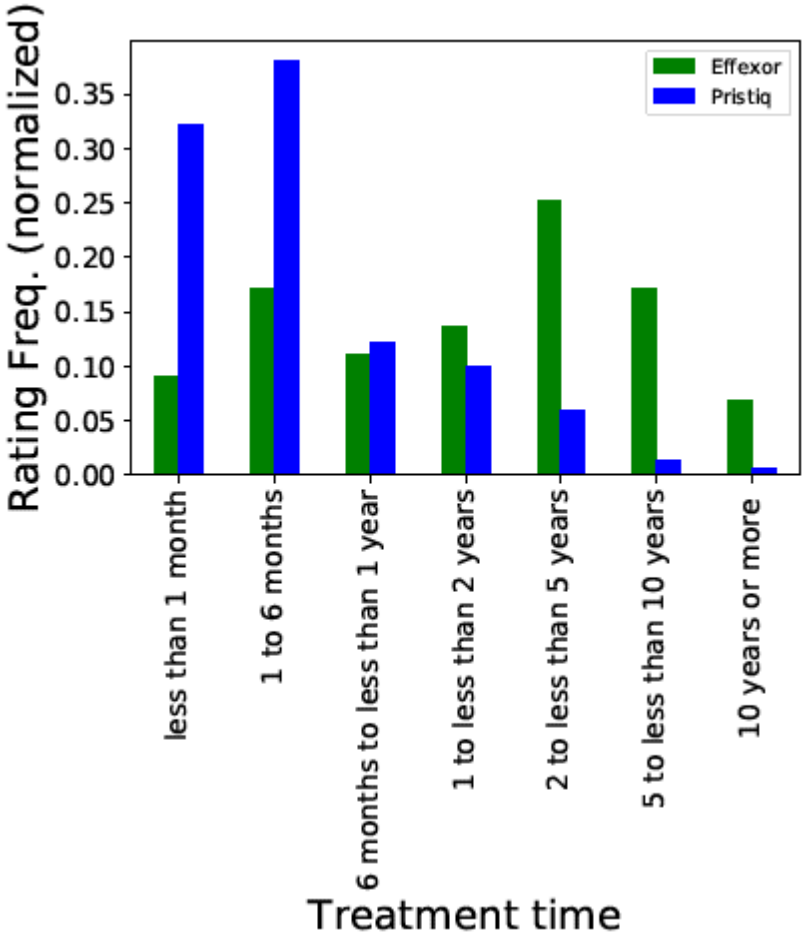


Test condition	Drug vs Condition	Condition vs Rating
Chi Sq. Stat.	38.6	41.1
P - value	4.2 e-9	1.4 e-10

Cymbalta's low ratings stem from customer dissatisfaction with it as a pain medication. Effexor and Pristiq are rarely prescribed for that purpose

Identifying factors differentiating Effexor and Pristiq





Treatment time may be a factor in reviewer response to Effexor or Requires further investigation

Conclusions & Next Steps

- Despite similar demographics, not all SNRI's are the same in reviewers eyes
- Patient satisfaction of Cymbalta driven lower from poor reviews associated with pain treatment

- What can be done next:
  - Logistic Regression Model

Tools

Scrapy

Python Library

NumPy &

Pandas

Data clean-up

matplotlib

SciPy