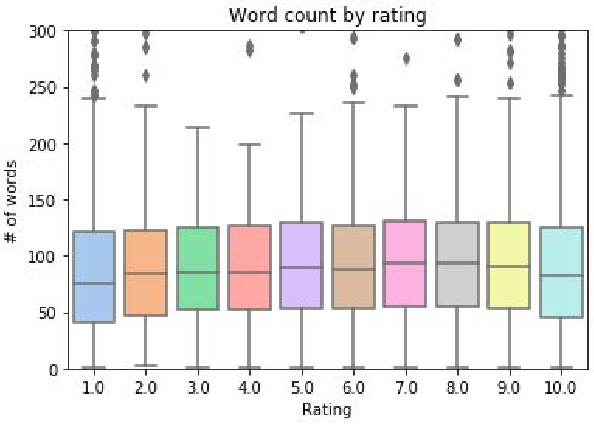
Lucas Reynolds
Springboard Data Science Career Track
January 2020 Cohort
Capstone Project 1

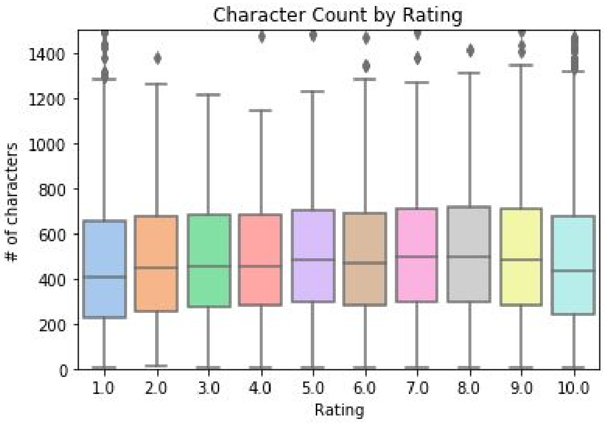
Data Storytelling

Problem Statement: Can text provided in reviews from medicine users be used to predict the rating they will provide?

Review Length

The length of reviews were first compared to see if there were any significant differences between ratings. A chart with boxplots comparing character counts for each rating and a sorted list of median character counts for each rating were created. The same was done for word count. There was not a large enough difference between review lengths of different ratings to help us predict ratings based on review length.





Drug Name

The median rating was calculated for different drugs and two lists were created, one for drugs with median ratings of 8 and above, and one for drugs with median ratings that were less than 8. Only drugs that appeared more than 50 times were included in these lists to avoid spelling errors or drug names with small sample sizes.

The 10 drugs with the highest median ratings were Relpax, Cialis, Diazepam, Polyethylene glycol 3350, Propofol, Propranolol, Cyproheptadine, Klonopin, Coblicistat/elvitegravir/emtricitabine/tenofovir, Clonazepam, and Ricatriptan all with median ratings of 10.

The 10 drugs with the lowest median ratings were Xarelto, Levora, and Intuniv with median ratings of 7.5, Abreva, Keppra, Microgestin Fe, Metroprolol, Methocarbamol, Lupron Depot, Loestrin 24 Fe, and Lo Loestrin with median ratings of 7.0.

Medical Condition

Similar to drug names, the median rating was calculated for different medical conditions. A list for conditions with median ratings of 8 and above and a list for conditions with median ratings that were less than 8 were both created. Only conditions that appeared more than 50 times were included to avoid spelling errors or conditions with small sample sizes.

The 10 conditions with the highest median ratings were Head Lice, Cluster Headaches, Headache, Rhinitis, HIV Infection, COPD, Gout, Prevention of Bladder Infection, Chronic Idiopathic Constipation, Chronic Myelogenous Leukemia, and Cold Sores all with a median rating of 10.

The 10 conditions with the lowest median ratings were Deep Vein Thrombosis, Menstrual Disorders, Endometriosis, Glaucoma Open Angle, Birth control, Benign Prostatic Hyperplasia, High Blood Pressure, Bacterial Infection, Autism, Neuropathic Pain, and Ovarian Cysts all with a median rating of 7.

Words

Bar graphs and lists with the top 10 most frequent words were created for each rating. Initial comparisons showed a lot of overlap in most frequent words between ratings. Words with the most overlap were added as stop words to the text cleaning function. The top 10 words for each rating are listed below.

1.0	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**	**10.0**
like	months	months	months	months	first	first	effects	years	years
months	like	like	first	first	months	effects	first	effects	effects
took	first	first	like	effects	effects	months	years	first	first
never	pill	effects	effects	like	like	like	months	months	life
would	effects	pill	pill	pill	years	years	like	like	like
first	would	would	period	feel	weight	feel	feel	feel	feel
pill	back	period	month	years	feel	also	also	would	months

effects	got	also	feel	period	period	weight	weeks	back	would
doctor	period	feel	also	also	also	pill	would	one	one
one	weeks	month	weight	weight	week	back	weight	weeks	back

Word clouds were also created to visually display the most common words for each rating.

As I learn additional techniques for data storytelling relevant to NLP I’ll come back and utilize them here.