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

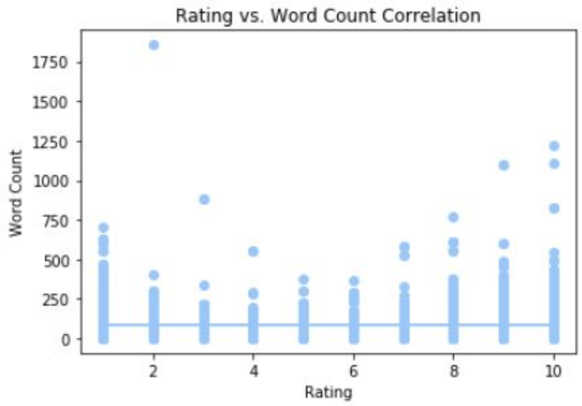
Inferential Statistics

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

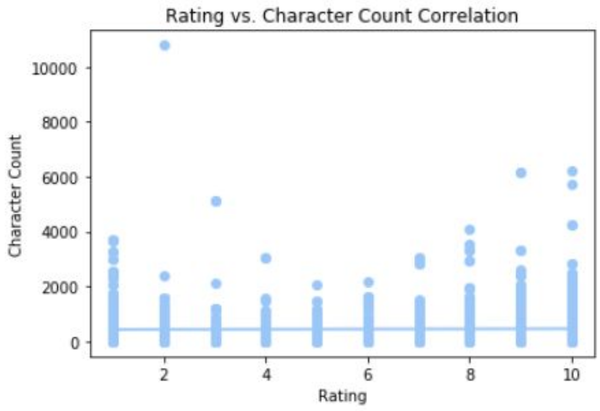
Review Length

The length of reviews were visually explored in the data storytelling step, but I also wanted to inspect statistically. Boxplots were created comparing both the word counts and character counts of reviews with ratings. There was not a clear trend or relationship between review length and rating.

To more clearly inspect any relationship between review length and rating, I calculated regression lines and correlation coefficient (r) for both word count versus rating and character count versus rating. The regression line was almost completely flat and the correlation coefficients were only 0.0267 and 0.0211 respectively. I was surprised there was not a more significant relationship between review length and rating.



```
Correlation coefficient (r): 0.0267
```



```
Correlation coefficient (r): 0.0211
```

Drug Name

A two-sample t-test was performed on each drug name that appeared at least 50 times in the data. The t-test was to discover if specific drugs impacted rating scores and if so, which drugs. Using an alpha of 0.05, 233 drugs had a significant impact on average rating. The 10 drugs with the lowest p_values were *Medroxyprogesterone*, *Depo-Provera*, *Tioconazole*, *Varenicline*, *Alprazolam*, *Chantix*, Nitrofurantoin, *Clonazepam*, *Plan B*, and *Blisovi*.

For each drug with a significant impact on average rating, average difference in rating was calculated (i.e. average rating without drug - average rating with drug). The 10 drugs with the largest increase in average rating were *Privine*, *Zinc Oxide*, *Astelin*, *Acetaminophen / pseudoephedrine*, Chlorpheniramine / *phenylephrine*, *Biafine*, *Niravam*, *Avonex Pen*, *Belladonna / opium*, and *Primatene Mist*. The 10 drugs with the largest decrease in average rating were *Systane*, *Succinylcholine*, *Trimethoprim*, *Chloraseptic Sore Throat Spray*, *Monistat 7*, *Blisovi 24 Fe*, *Rhofade*, *Influenza virus vaccine live trivalent*, *Delsym*, and *Estradiol Patch*. Average difference ranged from an increase of about 3.00 to a decrease of about 4.87.

Medical Condition

Similar to drug names, two-sample t-tests were used on each medical condition that appeared at least 50 times. An alpha of 0.05 was used and 75 conditions had a significant impact on average rating. The 10 conditions with the lowest p-values, starting with the lowest, were *Anxiety*, *Pain*, *Hyperhidrosis*, *HIV Infection*, *Hepatitis* C, *Sinusitis*, *Bacterial Infection*, *High Cholesterol*, *Psoriasis*, and *Overactive Bladder*.

Average difference in rating was calculated for each condition.The 10 conditions with the largest increase in average rating were *Costochondritis*, *Menopausal Disorders*, *B12 Nutritional Deficiency*, *Von Willebrand's Disease*, *Gingivitis*, *NSAID-Induced Gastric Ulcer*, *Dumping Syndrome*, *Biliary Cirrhosis*, *Mucositis*, *Herpes Simplex*, and *Mucocutaneous/Immunocompromised*. The 10 conditions with the largest decrease in average rating were *Bronchospasm Prophylaxis*, *Sore Throat*, *Herpes Zoster / Prophylaxis*, *Macular Edema*, *Pancreatic Exocrine Dysfunction*, *Deep Vein Thrombosis / Recurrent Event*, *Prostatitis*, *Pelvic Inflammatory Disease*, *Body Dysmorphic* Disorder, and *Keratoconjunctivitis Sicca*. Average difference ranged from an increase of about 3.00 to a decrease of about 3.25.

Words

The two-sample t-test was also used to determine significant words used in reviews. [[insert number of words]] had a p-value lower than the chosen alpha value of 0.05. The 10 words with the lowest p-values, starting with the lowest were *moist*, *reported*, *kep*t, *bloat*, *agoraphobia*, *stopped*, *loopy*, *restarted*, *alarming*, and *levoxyl*.

Average difference in rating was calculated for each significant word and the 10 words with the largest increase in average rating were *radio*, *coworker*, *levoxyl*, *aspergers*, *laminectomy*, *loaded*, *managable*, *epileptic*, *simplest*, and *natazia*. The 10 words with the largest decrease in average rating were *moist*, *consists*, *separation*, *tumour*, *wire*, *meanwhile*, *prednisolone*, *manufactured*, *fasciitis*, and *compressed*. Average difference ranged from an increase of about 1.76 to a decrease of about 2.15.