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

Inferential Statistics

Overview

My goal in this analysis was to build upon my initial exploratory data analysis (EDA) with a more quantitative evaluation of features. I investigated statistically significant relationships between rating scores and the following:

- Review length (both word and character count)
- Drug name
- Condition
- Review text (i.e. words used)

The correlation between review length and rating was very weak, but I discovered significant average differences in rating for certain drugs, conditions, and words used in reviews

Review Length

I initially explored the relationship between review length and rating during the Data Storytelling step. There wasn't a clear correlation, but I calculated the statistically significant correlation coefficient here to make sure. A p-value of 0.05 was used to determine statistical significance. Word count and character count had statistically significant correlations with rating, but they were very weak, about 0.02 for word count and character count respectively. The length of a review isn't a strong indicator of rating and we likely don't need to include this as a feature in our model.

Drug Name

I then explored whether certain drugs influenced average ratings, and if so how large the effect was. To minimize the effect of outliers and drugs with small sample size I filtered out drugs that appeared less than 50 times. I used a p-value of 0.05 to determine which drugs had a statistically significant average difference in rating. 233 drugs had statistically significant average differences in rating ranging from an average decrease in rating of about 4.87 to an average increase of about 3.00.

Below are the top 5 drugs for largest significant average increase and largest significant average decrease in rating.

Top 5 Drugs with Largest Average Increase in Rating		
	Drug Name	**Avg. Difference (Rating)**
1	Privine	3.00
2	Zinc Oxide	3.00
3	Astelin	3.00
4	Acetaminophen / pseudoephedrine	3.00
5	Chlorpheniramine / phenylephrine	3.00

Top 5 Drugs with Largest Average Decrease in Rating		
	Drug Name	**Avg. Difference (Rating)**
1	Systane	-4.87
2	Succinylcholine	-4.62
3	Trimethoprim	-4.25
4	Chloraseptic Sore Throat Spray	-4.00
5	Monistat 7	-3.97

Medical Condition

Similar to drug names, I explored which medical conditions had a significant average difference in rating. To minimize the effect of outliers and conditions with small sample size, conditions that appeared less than 50 times were not included. Again, a p-value of 0.05 was used. There were 75 conditions with significant average differences ranging from an average decrease in rating of about 3.25 to an average increase in rating of about 3.00.

Below are the top 5 conditions with the largest significant average increase and the top 5 conditions with the largest significant average decrease.

Top 5 Conditions with Largest Average Increase in Rating		
	Medical Condition	**Avg. Difference (Rating)**
1	Costochondritis	3.00
2	Menopausal Disorders	3.00
3	B12 Nutritional Deficiency	3.00
4	Von Willebrand's Disease	3.00
5	Gingivitis	3.00

Top 5 Conditions with Largest Average Decrease in Rating		
	Medical Condition	**Avg. Difference (Rating)**
1	Bronchospasm Prophylaxis	-3.25
2	Sore Throat	-3.11
3	Herpes Zoster, Prophylaxis	-3.00
4	Macular Edema	-2.85
5	Pancreatic Exocrine Dysfunction	-2.50

Words

To determine which words in reviews had the largest impact on rating I again calculated statistically significant average differences. Using a p-value of 0.05, 342 words were found with significant average differences in rating. These significant average differences ranged from an average decrease in rating of about 2.15 to an average increase in rating of about 1.76.

Below are the top 5 words with the largest significant average increase in rating and the top 5 words with the largest significant average decrease in rating.

Top 5 Words with Largest Average Increase in Rating		
	Words	**Avg. Difference (Rating)**
1	Radio	1.76
2	Coworker	1.53
3	Levoxyl	1.41
4	Aspergers	1.40
5	Laminectomy	1.39

Top 5 Words with Largest Average Decrease in Rating		
	Words	**Avg. Difference (Rating)**
1	Moist	-2.15
2	Consists	-1.50
3	Separation	-1.48
4	Tumour	-1.43
5	Wire	-1.43