Drug Performance Evaluation
Exploratory Analysis

Tess Andereson, tanderson4@bellarmine.edu
Matthew Carrico, mcarrico2@bellarmine.edu

I. INTRODUCTION

The data set we chose is on Drug Performance Evaluation. Its main objective is to assess Quality, Cost, and Effectiveness of drugs on the market for 38 common health conditions. The dataset comes from the user, "The Devastator" on Kaggle.com. In this report, we will develop and describe our findings using Exploratory Data Analysis. More specifically, we will look for patterns and relationships between variables to find out what factors are present in most effective drugs, how drug prices relate to other important features of the drugs, how the form of the drug impacts satisfaction and other factors, in addition to other information we can find through EDA.

II. DATA SET DESCRIPTION

This data set contains 685 samples with 10 columns with various data types. It contains no null values or missing data. A complete listing is shown in **Table 1**. The following variables are categorical and are stored as 'objects': Condition, Drug, Form, Indication, and Type. Form, Indication, and Type variables have between 3 and 6 categories that categorize the drugs. The Condition variable has 38 categories, the largest is hypertension with about 15% of the data belonging to that category. The rest of the categories comprise only minor percentages of the data, so there is little to observe between these categories. The drug variable has 470 categories out of 685 entries with the largest category having 8 entries; this category is not useful for our analysis but remains in the data set as an identifier for the drugs.

The other variables are continuous and are stored as 'Float64' numbers. These continuous variables are: Ease of Use, Effective, Price, Reviews, and Satisfaction. Ease of Use, Effectiveness, and Satisfaction are ratings on a scale from 1 to 5 based on customer reviews. The reviews variable contains the number of reviews associated with the drug and the price variable contains the average price of the drug.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Condition	Object	0
Drug	Object	0
Ease of Use	Float64	0
Effective	Float64	0
Form	Object	0
Indication	Object	0
Price	Float64	0
Reviews	Float64	0
Satisfaction	Float64	0
Type	object	0

III. DATA SET SUMMARY STATISTICS

More information on each variable is available in this section. For the continuous variables, **Table 2** shows the summary statistics for each variable. The variables that are ratings on a scale from 1-5 (Ease of Use, Effectiveness, and Satisfaction) have small standard deviations and more balanced distributions with their means being quite similar to their medians. In contrast, the Price and Reviews variables have very large values for standard deviation and significant discrepancies between their means and medians. This suggests a non-Normal distribution and the presence of significant, influential outliers.

For the categorical variables, frequency tables are shown for each relevant variable in **Tables 3a-d**. As described above, the Condition variable has 38 categories, with the largest category comprising only 15% of the data. The rest of the categories make up only minor percentages of the data, so there is little to observe between these categories. Additionally, the drug variable has 470 categories out of 685 entries with the largest category having only 8 entries; this category is not useful for our analysis and does not have a frequency table shown below as each category makes up 1% of the data or less.

Table 2: Summary Statistics for Drug Performance Evaluation

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Ease of Use	685	3.920038	0.894212	1	3.556667	4.05	4.5	5
Effective	685	3.52353	0.954126	1	3	3.6	4.11	5
Price	685	174.21183	667.743466	4	15.49	49.99	145.99	10362.19
Reviews	685	82.64410	374.281398	1	3.0	10.350877	57	4647
Satisfaction	685	3.195699	1.030422	1	2.575	3.2	3.901250	5

Table 3a: Proportions for DRUG PERFORMANCE EVALUATION Condition

Category	*Frequency*	*Proportion (%)*
Hypertension	101	14.74%
Atopic Dermatitis	67	9.78%
Fever	64	9.34%
Gastroesophageal Reflux Disease	54	7.88%
Bacterial Urinary Tract Infection	53	7.74%
Hypercholesterolemia	32	4.67%
Hemorrhoids	31	4.53%
Gout	31	4.53%
Endometriosis	19	2.77%
Pharyngitis due to Streptococcus Pyogenes	19	2.77%
Back Pain	17	2.48%
Diverticulitis of Gastrointestinal Tract	16	2.33%
Bacterial Conjunctivitis	16	2.33%
Flatulence	15	2.19%
Depression	15	2.19%
Edema	15	2.19%
Prevention of Cerebrovascular Accident	14	2.04%
Acute Bacterial Sinusitis	14	2.04%
Vertigo	13	1.90%
Fibromyalgia	12	1.75%
Vulvovaginal Candidiasis	10	1.46%
Adenocarcinoma of Pancreas	8	1.17%
Sore Throat	8	1.17%
Impetigo	5	0.73%
Herpes Zoster	4	0.58%
Scabies	4	0.58%
Genital Herpes Simplex	4	0.58%
Furunculosis	4	0.58%
Chickenpox	3	0.44%
Oral Candidiasis	3	0.44%
Infantile Autism	3	0.44%
Biliary Calculus	2	0.29%
Sleepiness due to Obstructive Sleep Apnea	2	0.29%
Meniere's Disease	2	0.29%
Influenza	2	0.29%
Pyelonephritis	2	0.29%
Colorectal Cancer	1	0.15%

Table 3b: Proportions for DRUG PERFORMANCE EVALUATION Indication

Category	*Frequency*	*Proportion (%)*
On Label	548	80%
Off Label	129	18.83%
\r\r\n	8	1.17%

Table 3c: Proportions for DRUG PERFORMANCE EVALUATION Form

Category	*Frequency*	*Proportion (%)*
Tablet	300	43.80%
Liquid (Drink)	119	17.37%
Cream	90	13.14%
Capsule	73	10.66%
Liquid (Inject)	57	8.32%
Other	46	6.72%

Table 3d: Proportions for DRUG PERFORMANCE EVALUATION Type

Category	*Frequency*	*Proportion (%)*
RX	484	70.66%
OTC	168	24.53%
RX/OTC	28	4.09%
\r\r\n	5	0.73%

Table 4 and the associated Figure 4a describe the relationships that exist between variables in this dataset. Strong relationships are shown on the heatmap in orange and peach, or in the table as values higher than 0.6. Weak relationships are shown in the heatmap in black, or in the table as values close to 0. We conclude that the significant relationships are between Ease of Use and Effectiveness, Effectiveness and Satisfaction, and Ease of Use and Satisfaction. To visualize these relationships, we've included scatterplots in a later section of this report.

Table 4: Correlation Table/Tables

	EaseOfUse	Effective	Price	Reviews	Satisfaction
EaseOfUse	1.000000	0.659237	-0.107480	0.011962	0.650156
Effective	0.659237	1.000000	-0.017532	-0.035802	0.864863
Price	-0.107480	-0.017532	1.000000	-0.024927	-0.024800
Reviews	0.011962	-0.035802	-0.024927	1.000000	-0.084216
Satisfaction	0.650156	0.864863	-0.024800	-0.084216	1.000000

Figure 4a: Heatmap of the Correlation Matrix for Variables in Drug Performance Evaluation

IV. DATA SET GRAPHICAL EXPLORATION

In this section of the report, we will visualize the statistics and relationships we explored earlier in the report.

A. Distributions

These histograms serve as visual representations of the distributions of the continuous variables in the dataset. Much of the information found in these histograms can be understood by reading Table 2 in the previous section which contains minimum and maximum values, as well as the distribution of the data between those values, as well as the variable's measures of center. These histograms, however, provide the same information (less precisely) and it is available at a glance and requires less understanding of the meaning of statistical measures.

Figure 1a: Distribution of Satisfaction Ratings

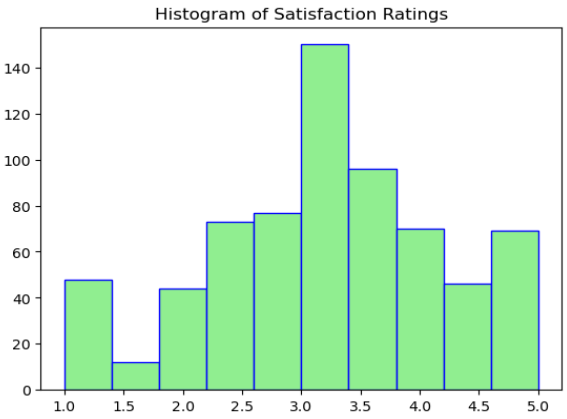
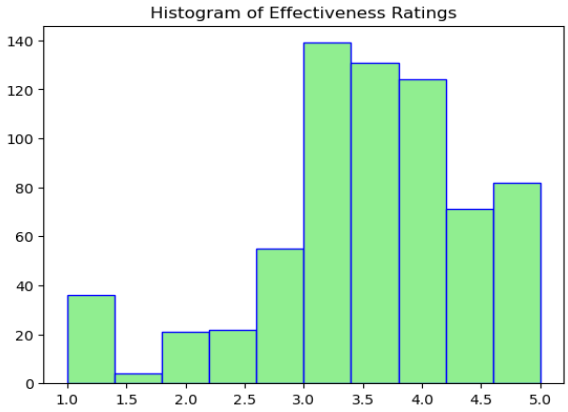


Figure 1b: Distribution of Effectiveness



These distributions of Satisfaction and Effectiveness show a somewhat Normal distribution, though Effectiveness has a bit of a left skew. Still, most of the data are in the middle ratings around 3.0 with fewer observations at the extremes like 1 and 5. This is intuitive – most of the meds will be in the middle with only a few very bad ones and a few really good ones.

Figure 1c: Distribution of Ease of Use Ratings

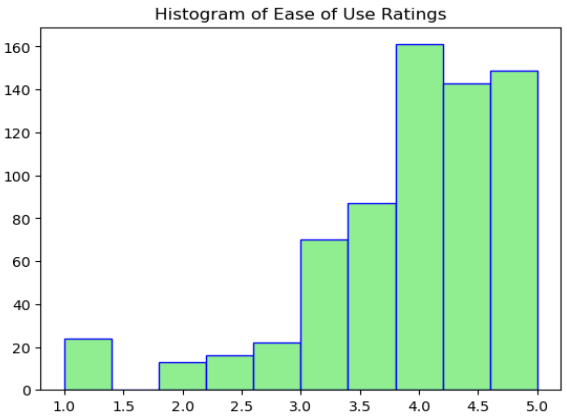
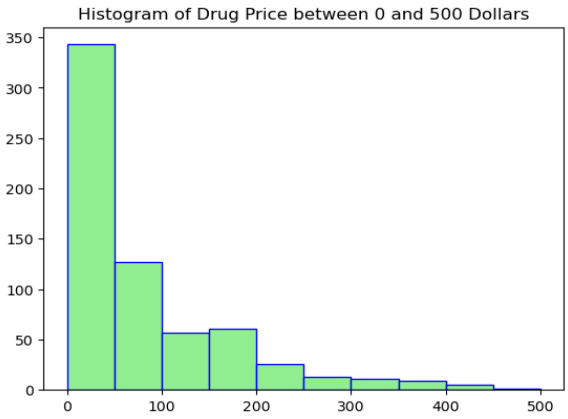
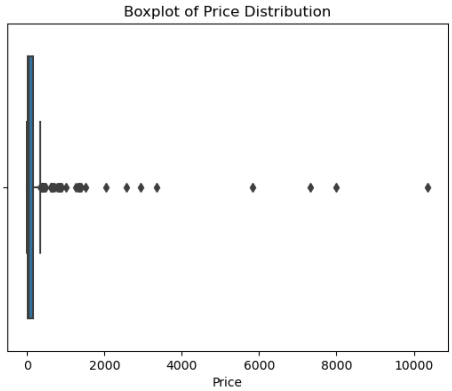


Figure 1d: Distribution of Price Below $500



These distributions of Ease of Use and Drug Price show very skewed distributions. Ease of Use is strongly skewed to the left, meaning there are a lot more observations with high ratings than with low ratings. Price behaves in the opposite way; it is strongly skewed to the right and has more observations with very low prices (around $50) and fewer observations as price increases. The price distribution only contains prices lower than $500 because the range of values is very large as shown in Table 2. We chose to only include the lower three quartiles of this distribution to better represent how the majority of the data behaves, since the presence of large outliers greatly impacts the distribution. This is clear because of the large discrepancy between Mean and Median of price. Below, we include a boxplot of the Price Distribution as it is a better representation of the full distribution. Because of the large frequency of entries around $50 and the infrequent but significant outliers, a histogram of the full range of prices is not very helpful – the outliers aren't clear and much of the data is lost because of the huge range of values.

Figure 1e: Boxplot of Price Distribution



This distribution has much less information about the majority of the data, at a glance because the main box of the boxplot is so condensed on the left. This is because, as mentioned above, the majority of the data occur between $4 and $150. However, this graph is better able to show the presence and influence of the significant outliers that are drugs with prices greater than $300. Each of the diamonds shown outside the box-and-whiskers represents a drug with the corresponding price on the x-axis. Both of these representations of the Price variable are important to understanding how price is distributed in this dataset – the histogram above shows the behavior of most of the observations, while the boxplot to the left shows the behavior or influence of the high-priced drugs that are outliers in this dataset.

B. ScatterPlots / Pairwise Plots (continuous variables)

This figures help visualize the relationships between variables that exist in the data set. As discussed above in Table 4 and Figure 4a, these variables have correlation coefficients with values higher than 6, meaning they have at least a moderately positive relationship. This can be seen in the graphs by the clusters of points mostly forming a diagonal, positively sloping line.

Figure 2: (a) Satisfaction vs Effectivness (b) Effectiveness vs Ease of Use (c) Ease of Use vs Satisfaction

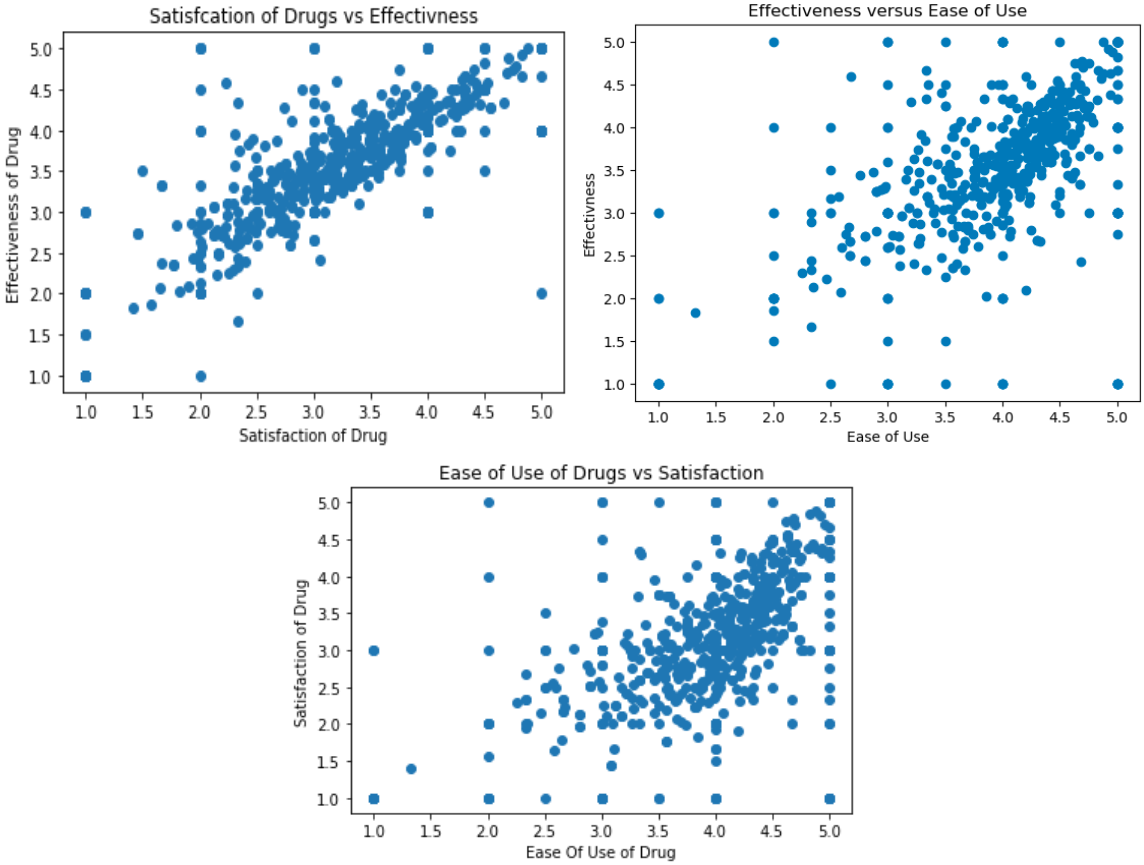
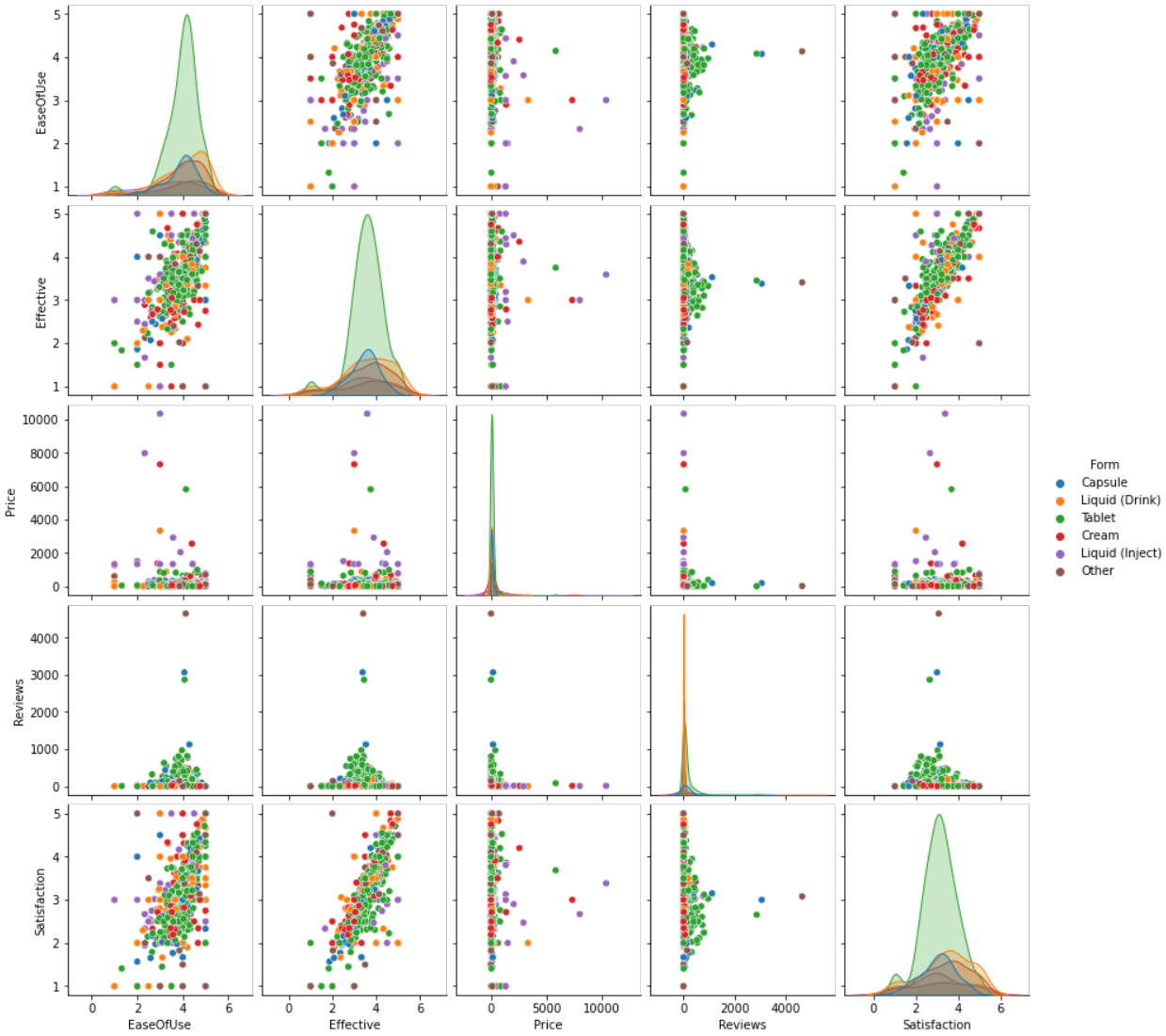


Figure 3: Pairwise Plot of the Data Set



C. *Barcharts (categorical variables)*

Figure 4: Ease of Use by Form of Drug

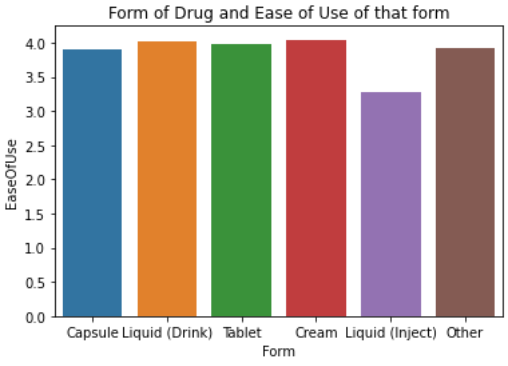


Figure 5: Effectiveness by Form of Drug

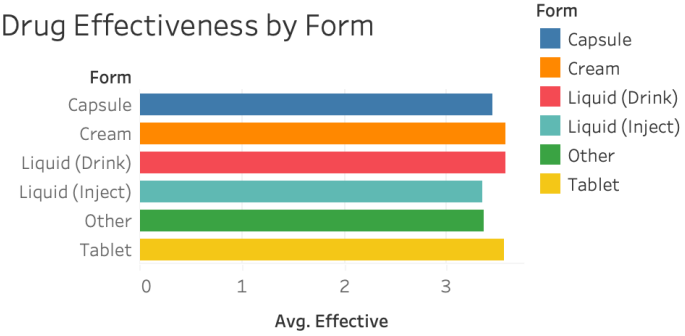


Figure 6: Average Price by Form

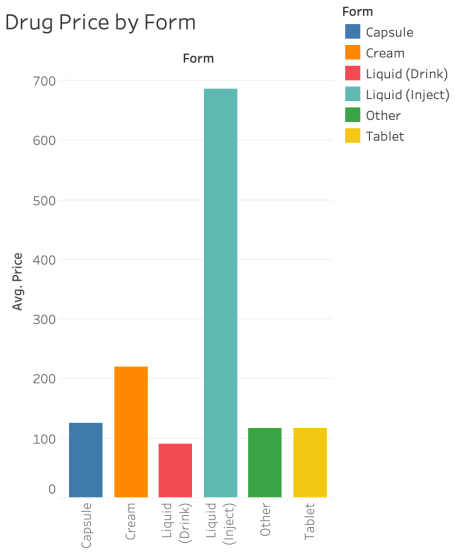


Figure 7: Average Price by Type

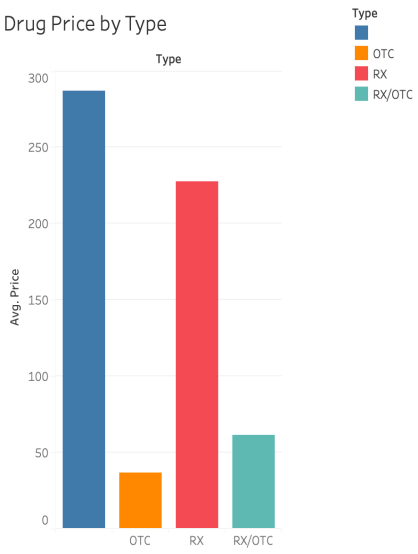


Figure 8: Ease of Use by Form

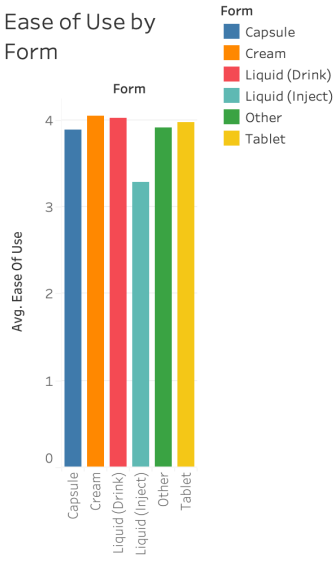
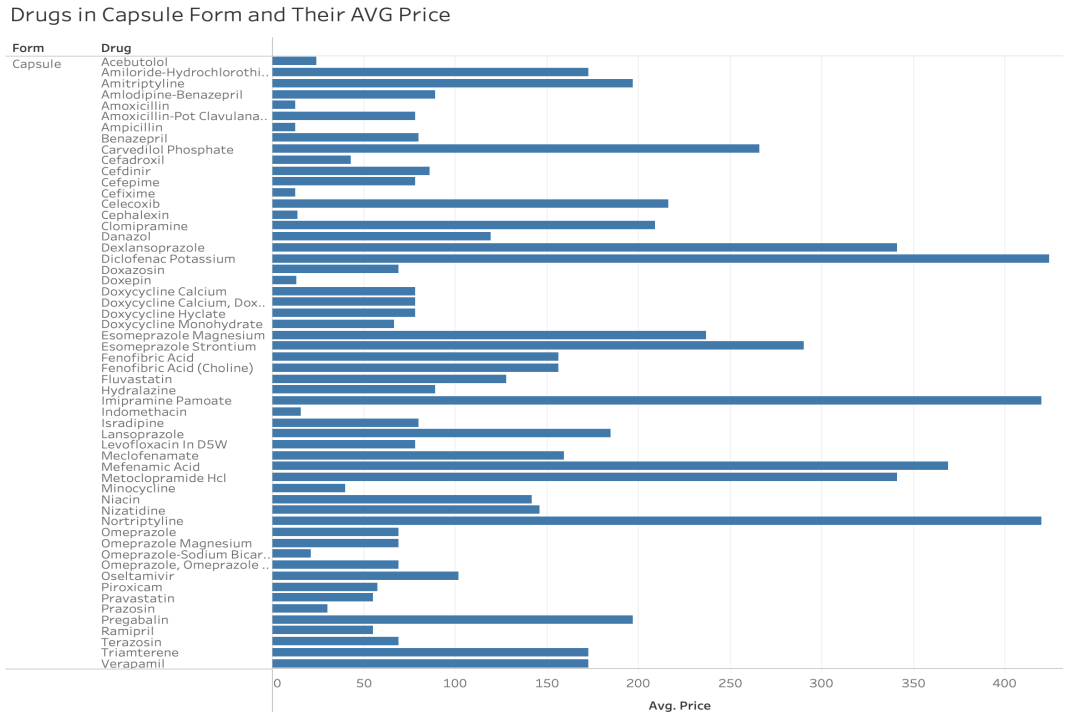


Figure 9: Satisfaction Ratings by Form

Figure 10: Drugs in Capsule Form and their AVG Price



D. *Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.*

Figure 11: Pie Chart for Prescription versus Over the Counter

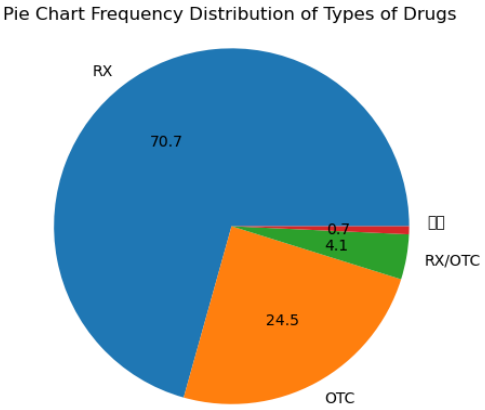


Figure 12: Effectiveness of Drugs that fight a Fever

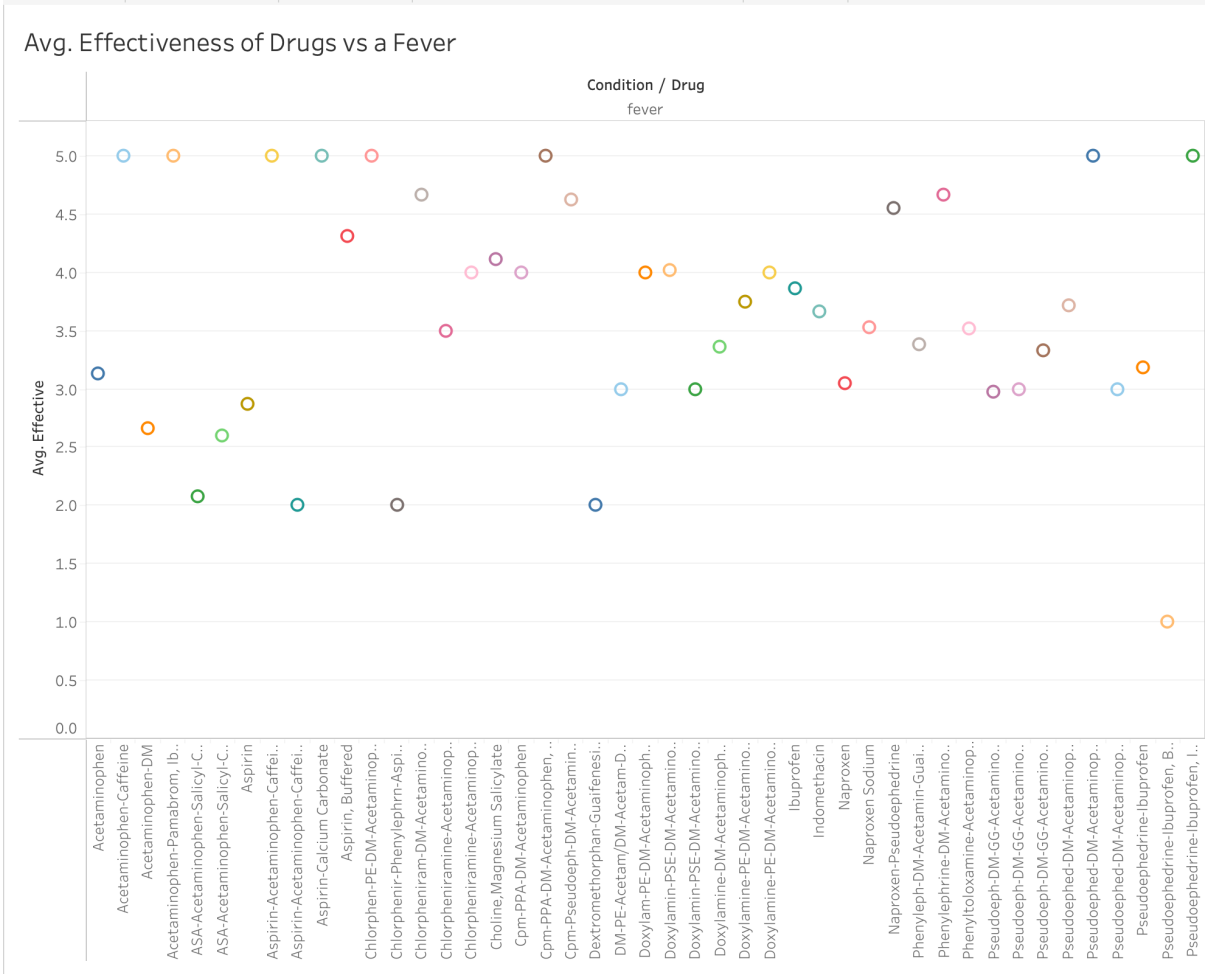


Figure 13: Drug Price by Condition

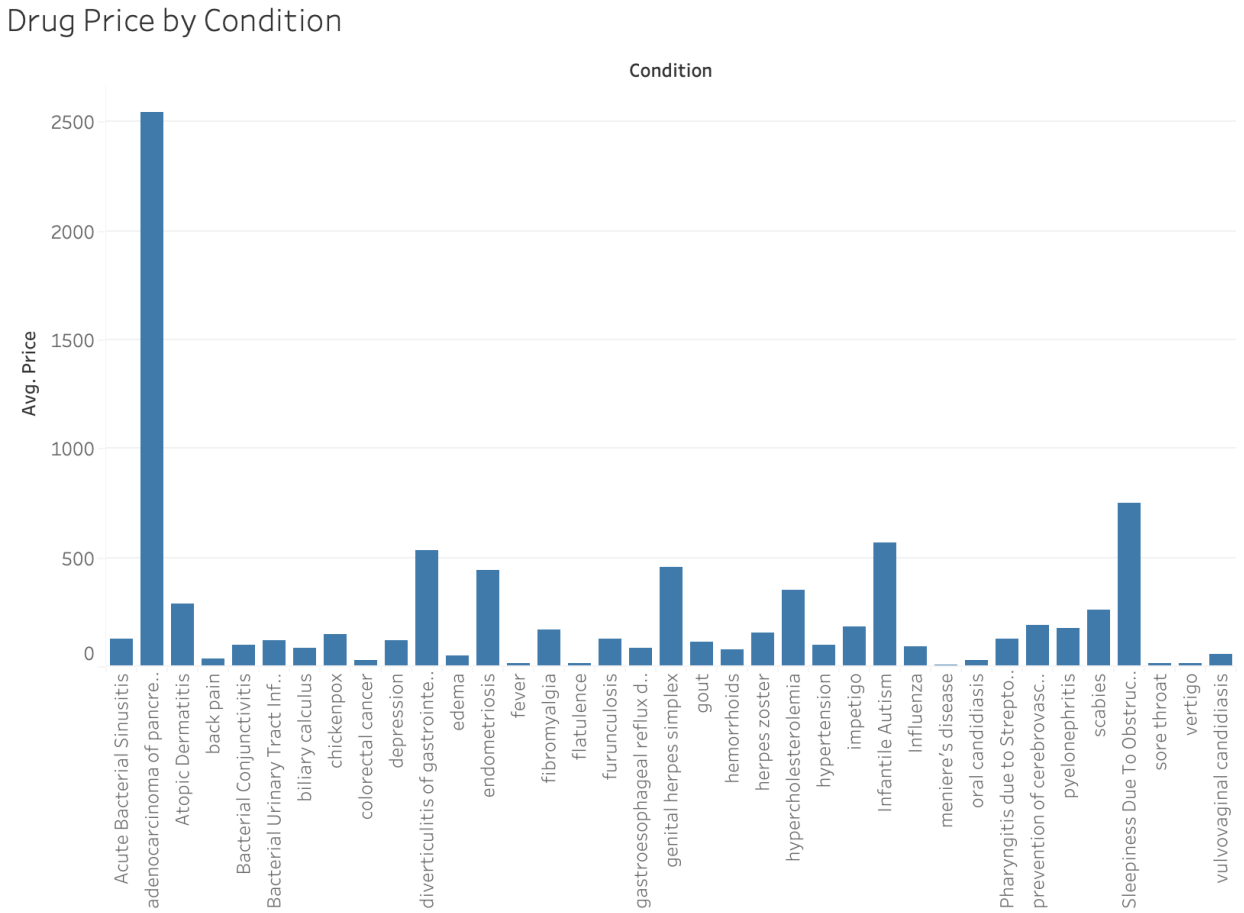
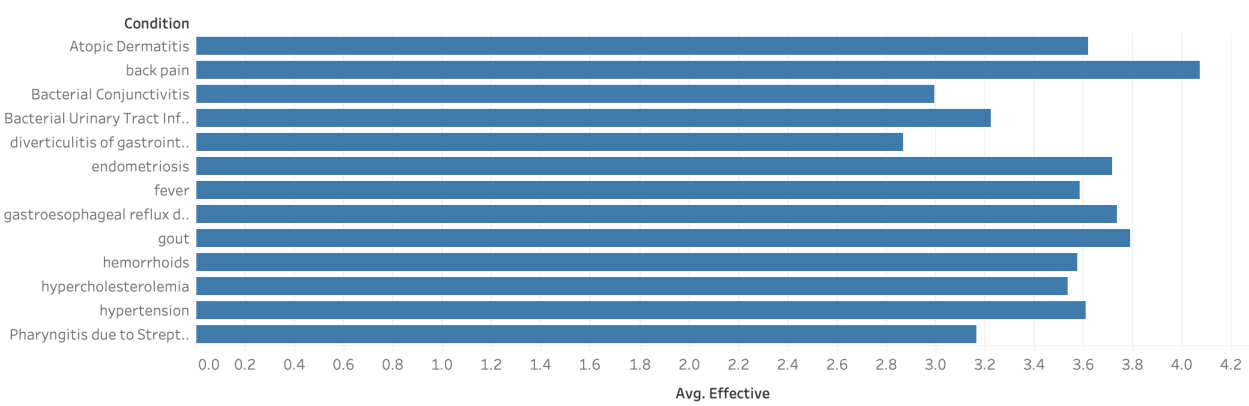


Figure 14: Drug Effectiveness by Condition (filtered for Conditions with Count>15)



V. SUMMARY OF FINDINGS

After analyzing this dataset using Exploratory Data Analysis, we find that most of the medication in this dataset are moderately effective, easy to use, and inexpensive. We also see that there is a strong linear relationship between Satisfaction ratings and Effectiveness ratings which is intuitive – people are more satisfied when they find the drug to be effective. We also learned that most of the drugs from this dataset are distributed by a pharmacy and not available over the counter.