Drug Performance Evaluation
Exploratory Analysis

Bobby Gabriel, bgabriel@bellarmine.edu
Ashley Ridley, aridley@bellarmine.edu

I. INTRODUCTION

This dataset is called the “Drug Performance Evaluation” and it assesses quality, cost and effectiveness of a certain drug. It contains drug performance metrics for 37 common conditions and can be found at https://www.kaggle.com/datasets/thedevastator/drug-performance-evaluation?select=Drug_clean.csv. We chose this dataset because it looked like a good dataset to compare variables and it contained a mixture of object and int data types.

II. DATA SET DESCRIPTION

Narrative summary of the data set: e.g. this data set contains 685 samples with 10 columns with either data type object or data type float64. Additionally, each column is completely filled and contains no missing data. A complete listing is shown in **Table 1**.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
V1 Condition	Nominal/object	0%
V2 Drug	Nominal/object	0%
V3 EaseOfUse	Interval/float64	0%
V4 Effective	Interval/float64	0%
V5 Form	Nominal/object	0%
V6 Indication	Ordinal/object	0%
V7 Price	Ratio/float64	0%
V8 Reviews	Ratio/float64	0%
V9 Satisfaction	Interval/float64	0%
V10 Type	Nominal/object	0%

III. Data Set Summary Statistics

The summary of statistics on each float64 variable will be calculated with pandas and include the count, mean, standard deviation, minimum, maximum, and the percentiles for 25, 50, and 75. These values can be found in table 2. For the categorical variables, we will be recording the frequency they appear and the proportion. This data can be found on Table 3.

Table 2: Summary Statistics for XXX (name of dataset)

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
V3 EaseOfUse	*685*	*3.92*	*.89*	*1*	*3.56*	*4.05*	*4.50*	*5*
V4 Effective	*685*	*3.52*	*.95*	*1*	*3.00*	*3.60*	*4.11*	*5*
V7 Price	*685*	*174.21*	*667.74*	*4*	*15.49*	*49.99*	*145.99*	*10362.19*
V8 Reviews	*685*	*82.64*	*273.28*	*1*	*3.00*	*10.35*	*57.00*	*4647.00*
V9 Satisfaction	*685*	*3.19*	*1.03*	*1*	*2.58*	*3.20*	*3.90*	*5*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for Condition (There are far too many conditions so I will list the top 10)

Category	*Frequency*	*Proportion (%)*
Hypertension	*101*	*14.7*
Atopic Dermatitis	*67*	*9.8*
Fever	*64*	*9.3*
Gastroesophageal reflux disease	*54*	*7.8*
Bacterial Urinary Tract Infection	*53*	*7.7*

Hypercholesterolemia	*32*	*4.6*
Hemorrhoids	*31*	*4.5*
Gout	*31*	*4.5*
Endometriosis	*19*	*2.7*
Steptococcus Pyogenes	*19*	*2.7*

Proportions for Drug (There are 470 different drugs, I will list the top 5)

Category	*Frequency*	*Proportion (%)*
Niacin	*8*	*1.2*
Naproxen Sodium	*7*	*1.0*
Hydrocortisone	*7*	*1.0*
Ibuprofen	*6*	*0.8*
Amoxicillin-Pot Clavulanate	*5*	*0.7*

Proportions for 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*

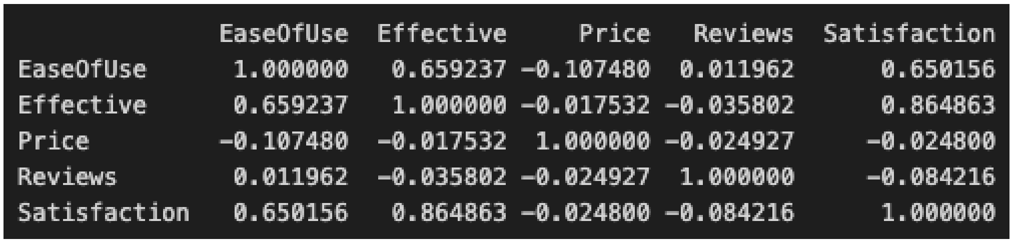
Proportions for Indication

Category	*Frequency*	*Proportion (%)*
On Label	*548*	*80*
Off Label	*137*	*20*

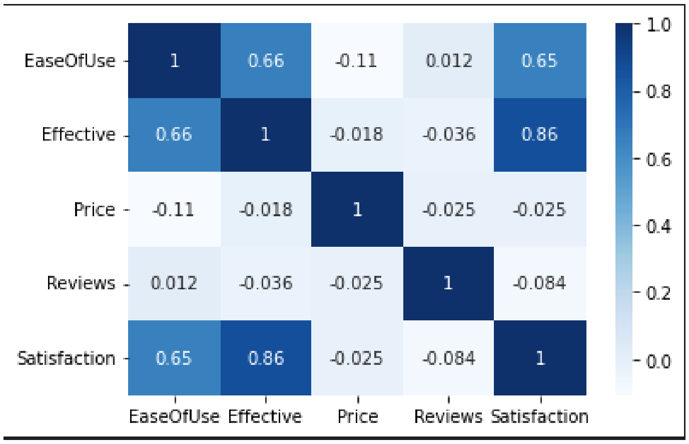
Proportions for Type

Category	*Frequency*	*Proportion (%)*
RX	*484*	*70.66*
OTC	*168*	*24.52*
RX/OTC	*32*	*4.78*

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
```



IV. DATA SET GRAPHICAL EXPLORATION

For the sections below, we will look at both the continuous and categorical variables and compare how they relate to one another. We will try to find notable values, variables that have a strong/weak correlation, and overall draw conclusions for the dataset.

A. Distributions

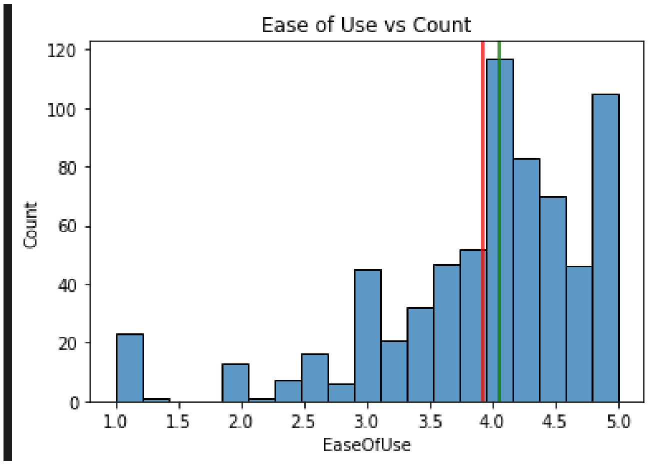


Figure 1: Histogram Comparison of EaseOfUse vs Count from dataset. The red line specifies the mean. The green line specifies the median.

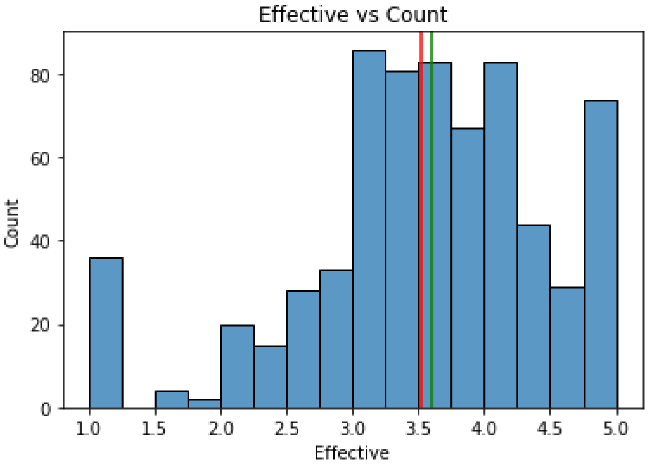


Figure 2: Histogram comparison of Effectiveness vs Count. The red line specifies the mean. The green line specifies the median.

B) Scatter Plots and PairPlots

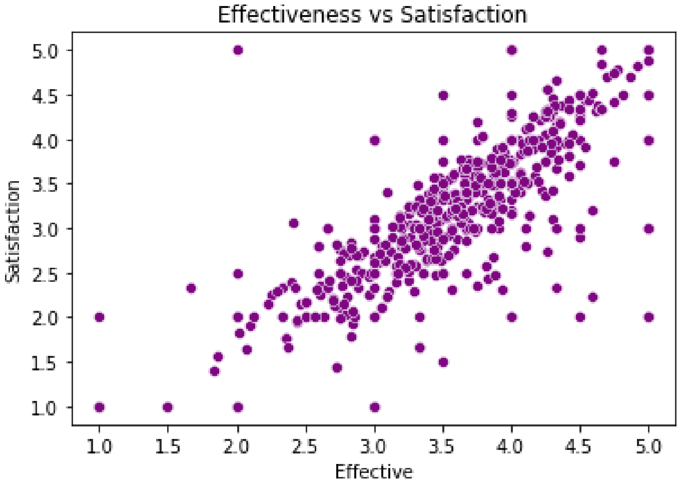


Figure 3: Scatter plot of Effectiveness vs Satisfaction

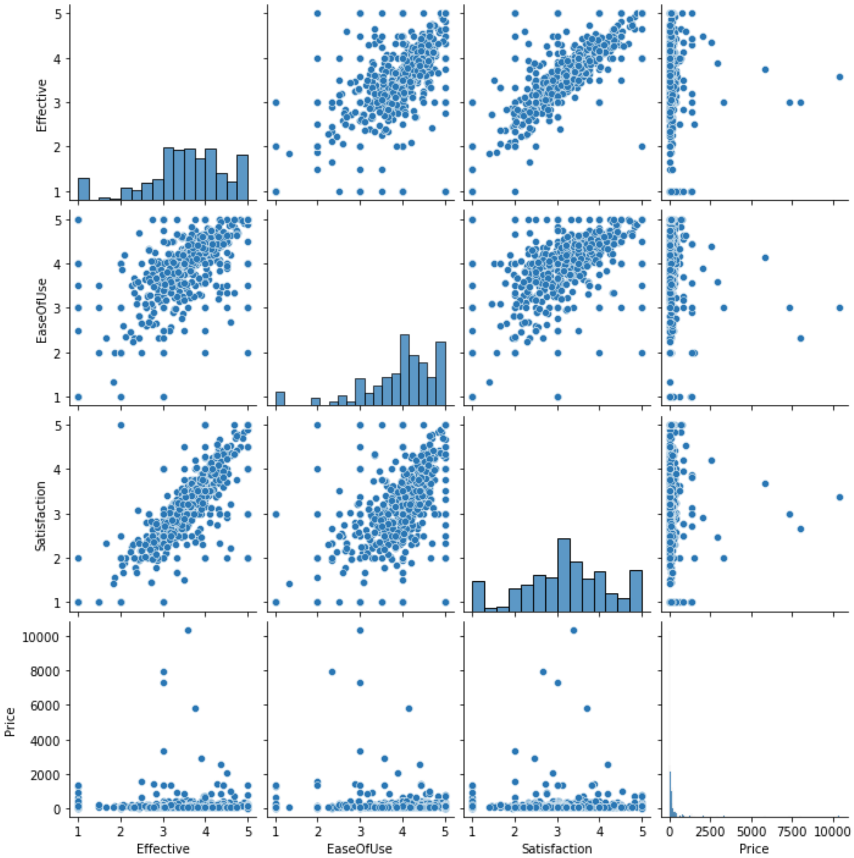


Figure 4: Pair plot of Effectiveness, Satisfaction, EaseOfUse, Price compared against one another.

Barcharts (categorical variables)

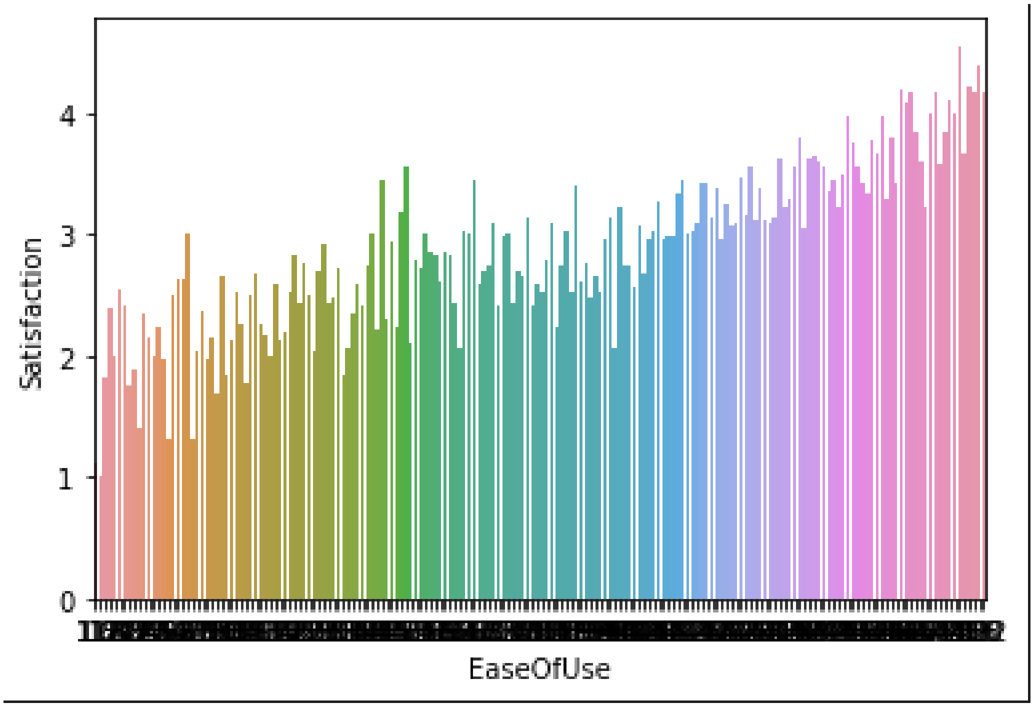


Figure 5: Bar plot of EaseOfUse vs Satisfaction

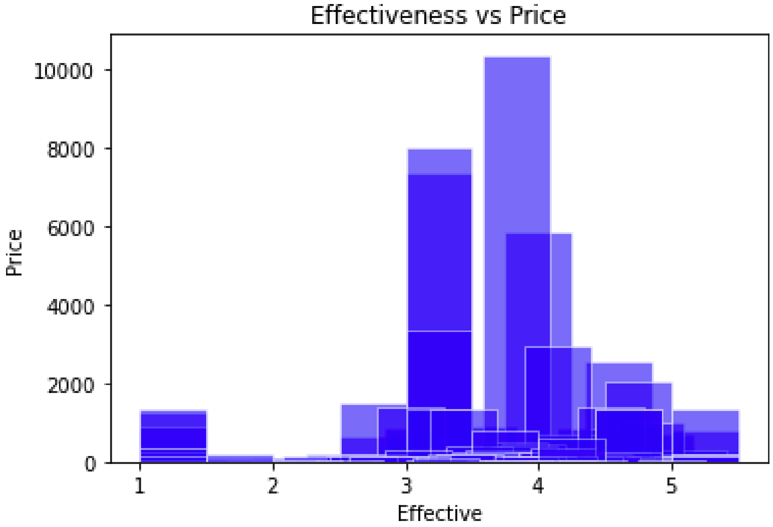


Figure 6: Bar Plot of Effectiveness vs Price

V. SUMMARY OF FINDINGS

Overall, this dataset was very nice to work with as it had enough values and a mixture of continuous and categorical variables to effectively compare columns. It was also nice because it started out with no null values which made it easy to start our data analysis and begin to check our findings. Based on our graphs from the figures above, there are many conclusions that can be made. To start with our scatter plot of Effectiveness vs Satisfaction, there seems to be a linear correlation in how satisfied the customer is compared to the effectiveness of the drug. The next result I discovered was that it is not necessarily the case that if a drug is more effective that the price will be higher. If we refer to Figure 6, it seems that the most expensive drugs are only in the effectiveness level of around 3 to 4 out of 5.

The final result I would like to note is from figure 2. I thought it was very reassuring seeing such a high count of effective drugs in the world today. The graph is left skewed, and the mean is around 3.5/5 which is more than the halfway point. These are good characteristics when dealing with drugs that could save people’s lives.