Vrushali Patel

Vallesia Pierre-Louis

Ryan Chen

STA 4164 0001

# Final Report: Predicting Which Factors Contribute Most to User Satisfaction of Pharmaceutical Drugs

## 1. Data Description:

The data is collected and made by the user “the devastator” and it caters to a diverse audience that looks at these certain conditions and medications. The dataset encompasses a comprehensive comparison of the condition, what drug it is, the ease of use, the effectiveness, form, indication, price, reviews, satisfaction, and type that has an intake on public health. Our study's motivation is to understand what factors contribute to patient satisfaction and how the medication is used to use it in the diagnosis, cure, treatment, and prevention of diseases. By analyzing The Drug Performance Evaluation dataset, we seek to identify key determinants to best help us achieve our goal in patient satisfaction. With mass distributions of medication all around the world, it is important to choose the correct form for the patient. With the outcome, we will find out which factors contribute the most to user satisfaction, with the following medication. This also can be a huge thing for pharmaceutical businesses to include when enhancing with their products when selling, they have a higher chance in an elevated level of patient satisfaction. This is a crucial point for anybody in healthcare for professionals that are seeking to improve the overall quality of the delivery and patient outcomes. The dataset is collected through performance evaluation by patients, which are gathered from various sources, such as patient surveys, clinical trials, and healthcare databases. From each of the different variables, it gives insights into the factors that give off patient satisfaction. Patient satisfaction is a vital role in assessing drug performance if the patient is fulfilled with the drug and its outcomes. Additionally, the physical environment is also a significant factor in satisfaction with how everything is present in real life and setting of the overall space around you. In summary, the dataset is a valuable resource for understanding its impact on patient satisfaction and drug performance. The diverse variables are a valuable tool for treatment effectiveness and patient experiences, providing to future researchers and healthcare practices to enhance satisfaction levels and enhancing patient outcomes.

This means it seeks to identify the elements and what influence and effects are experienced by patients. For instance, say a patient used Amoxicillin for Acute Bacterial Sinusitis, patients that have that problem and used that drug compared their effectiveness, form, indication, price, reviews, satisfaction, and type to each other drug and their condition. Ease of use means, rating how easy it was to use the drug. Effectiveness indicates the level of how much the drug helped to alleviate the condition. The form is the physical presentation of the drug, of how the drug came in either cream, tablet, or liquid. The indication, which is if the purpose of the drug is on or off-label. The price of each drug as well, some patients might even look away from getting a drug and take an alternative instead because of the affordability. The reviews indicate the insight of how many people have used the drug and given their experience on it. Satisfaction of how happy and satisfied the patient was when they used the drug and what classification the drug is under. Patients also consider what type of drug it is, whether it is RX, prescription only drug where you need a medical order from a doctor or OTC, over the counter where you can get the drug at any pharmacy. You can obtain certain drugs OTC, but if you want a RX by a doctor and want a higher dosage of a drug, the outcome may lead to a faster recovery.

## 2. Model Diagnostics:

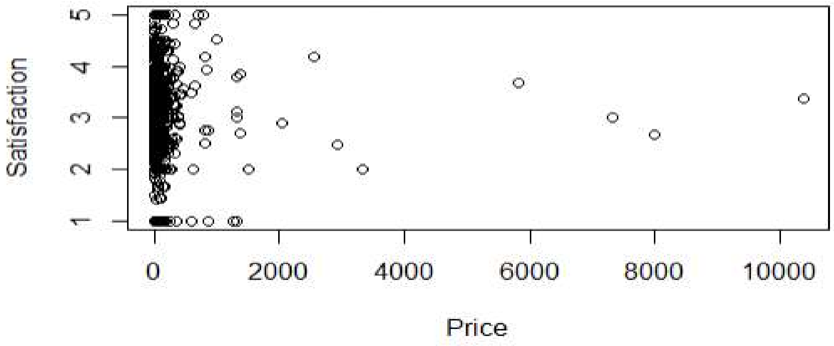
The dataset provides a range of variables. It includes variables such as Condition, Drug, EaseOfUse, Effectiveness, Price, Reviews, Type, and Satisfaction as the response variable. The two variables that will be omitted and removed from the full model will be Condition and Drug name. There are over 300 Drug names and 37 Conditions. Removing them will simplify the dataset and focus on relevant factors that can provide us with a better model that explores the remaining variables. For the indication variable, there are some data that will be omitted since it was left blank to provide a better final model. We will transform the following variables into dummy variables, Type, and Indication.

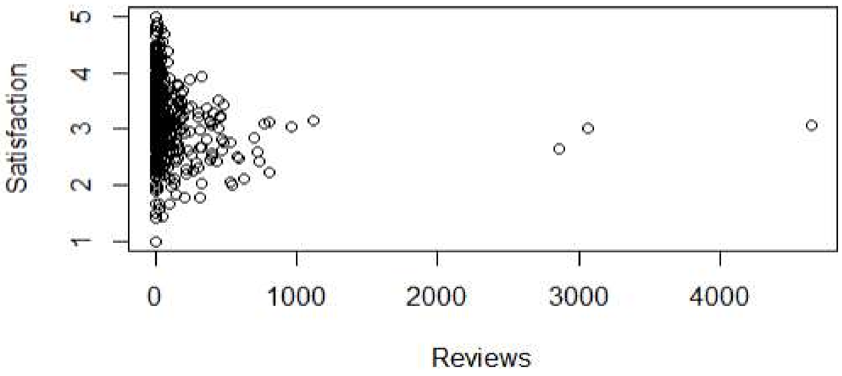
Vrushali Patel

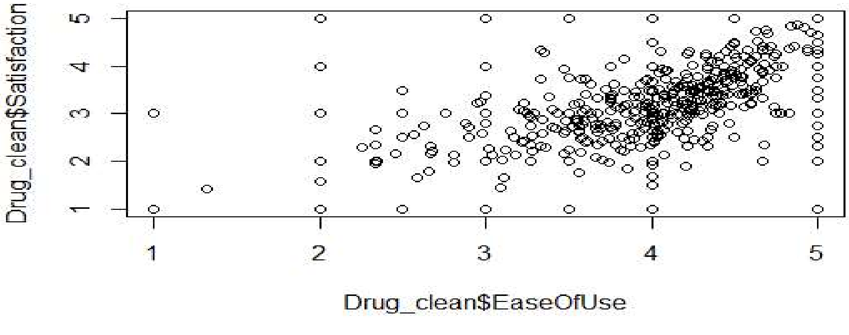
Vallesia Pierre-Louis

Ryan Chen

STA 4164 0001





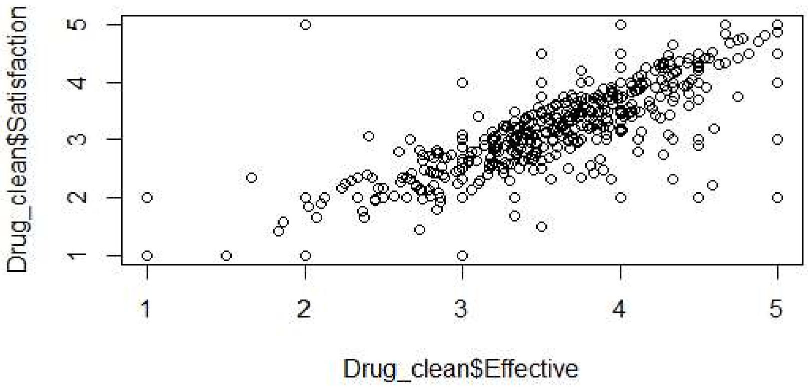


Vrushali Patel

Vallesia Pierre-Louis

Ryan Chen

STA 4164 0001

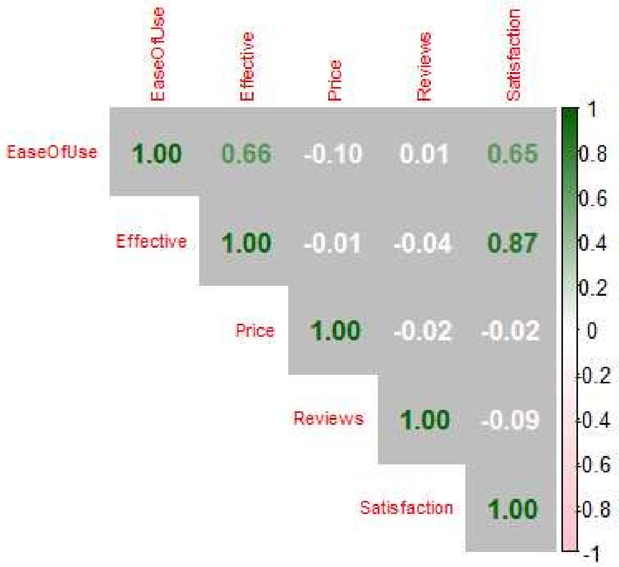


An obvious outlier that is seen is for the drugs with prices over $5000 and drugs with over 3000 reviews. Initially, we wanted to remove these high prices, but there are many patients whose medication costs are at those prices. The price of Paclitaxel-Protein Bound, a drug that is used for breast and pancreatic cancer patients, is $10362.19 and their satisfaction is also important, therefore, we decided not to remove the outlier prices.

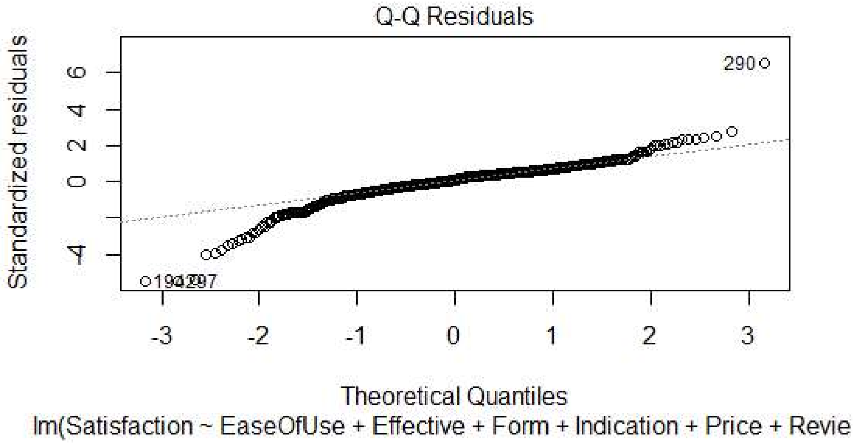
To determine the remaining outliers, we performed a logistic regression of the full model and determined the Cook's distance. The cook's distance showed no outliers, as the values were greater than one. We followed with the leverage. The leverage's threshold was 0.02790698. There were 74 observations that surpassed that threshold, thus being a potential outlier. The Jackknife's residual was 1.963707. There were 34 observations that exceeded that value. There were 8 observations that fell into the Leverage and Jackknife. We decided not to remove the outliers as we looked over them, and they seemed plausible, and the Cook's Distance was sufficient.

A correlation plot model was built to observe and estimate the linearity of each variable and satisfaction. In the correlation plot, 1 represents perfect positive linearity, -1 represents perfect negative linearity and 0 indicates no linear correlation.  The variables that have a near perfect positive relationship with Satisfaction are EaseOfUSe and Effective. The linearity assumption is being violated from the variables with correlations that are close to zero, such as

Price and Review, which predicts that we might have to remove those variables. The correlation plot alone is not enough evidence to remove the variables, but it hints on what direction to go.



A VIF test was performed and there were no values above 10, indicating that collinearity was not violated. To tackle our normality, we performed a Shapiro test and plotted the full model. Upon viewing the plots, we can observe that the normality is slightly violated on the QQ plot.

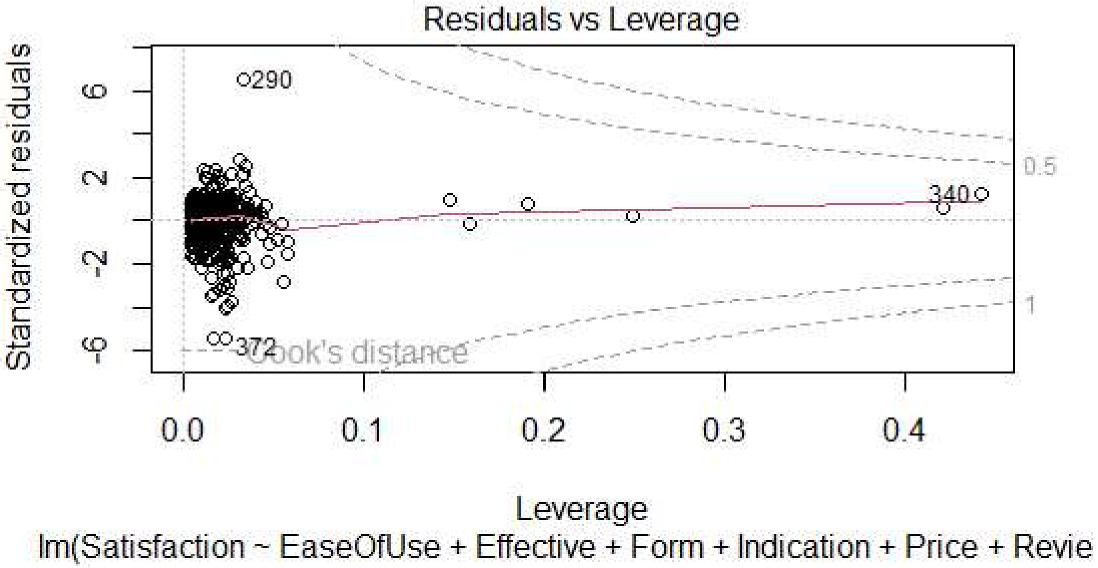


Vrushali Patel

Vallesia Pierre-Louis

Ryan Chen

STA 4164 0001



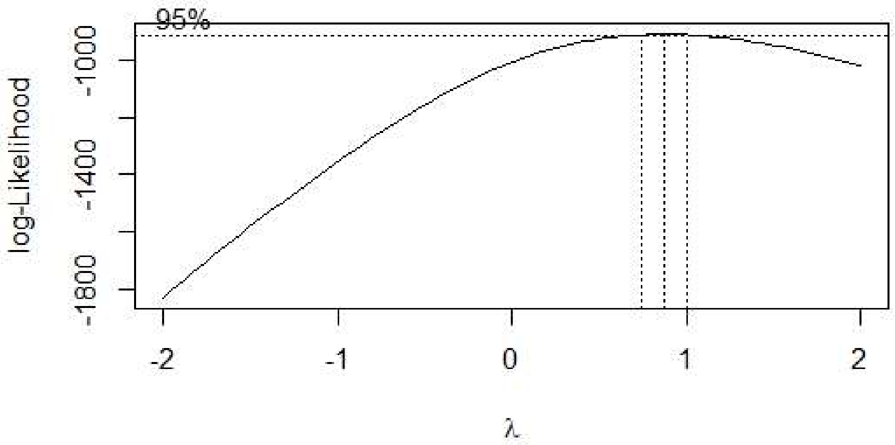
Heteroscedasticity is present in the Residual vs Leverage plot which implies a non-constant variance. In the Shapiro-Wilk test, the W is reported as 0.8827. This value is close to 1. The closer a data is to 1, the more the data resembles a normal distribution. The p-value that is associated is extremely small. This suggests strong evidence against the null hypothesis, indicating the residuals are not normally distributed. The solution to both is to use transformation.

Vrushali Patel

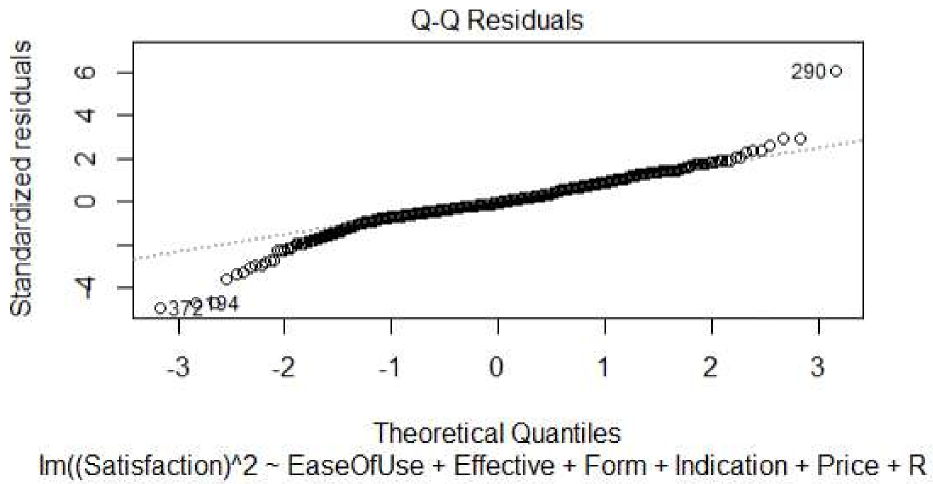
Vallesia Pierre-Louis

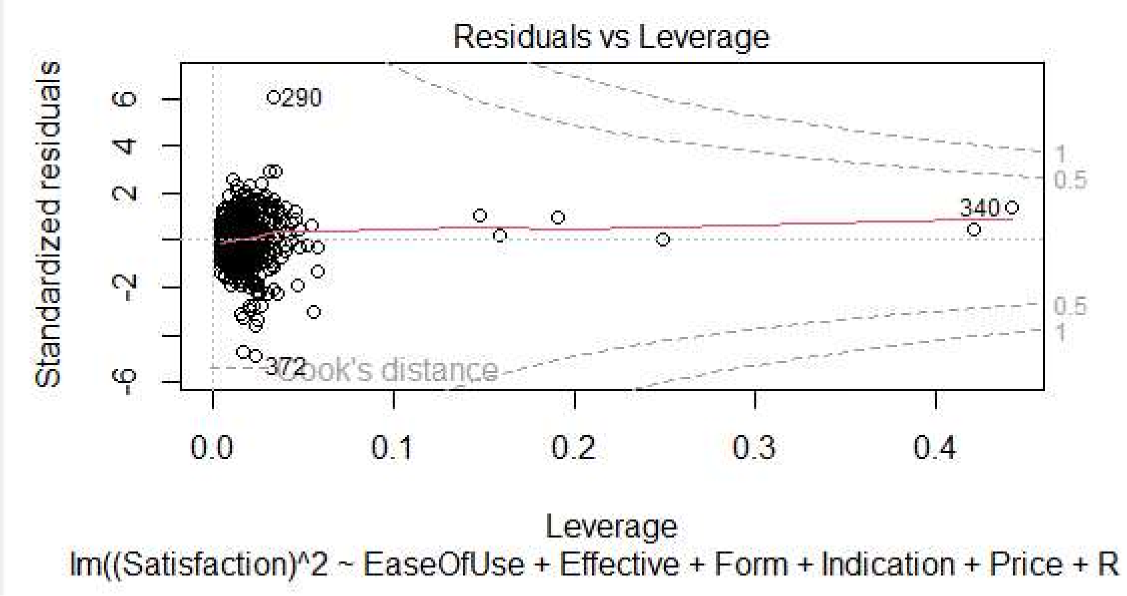
Ryan Chen

STA 4164 0001



A Box Cox test was performed on the model to determine the transformation type. Upon transforming the Satisfaction variable in the model, it was revealed that there was not much of an improvement to the plot.



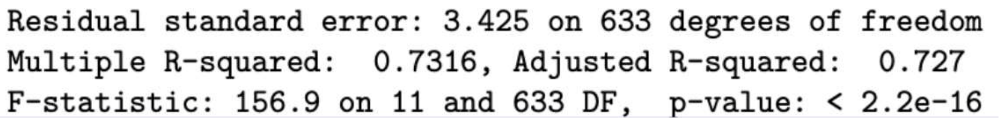


Since the plots did not have a significant change with the transformations, we decided to continue without the transformations.

### 3. Model Selection:

After performing all our model diagnostics and seeing the values given to us, we looked over all the possible models to see the criteria for each and judge which one would be best for us. We decided to use backwards elimination because the criteria this model uses are the F-statistic and the p-value, which according to our model diagnostics were good to use as the F-statistics was significant due to the p-value. This model would be best for us as there were some variables that were quite insignificant relative to others so removing these one at a time and reevaluating our model create the best model for our dataset. Our full model to start will consist of a response variable satisfaction with the following predictor variables, EaseOfUse, Effective, FormCream, FormLiquid(Drink), FormLiquid(Inject), FormOther, FormTablet, IndicationOnLabel, Price, Reviews, and TypeRx. With this full model we would move forward and in creating our final model with backwards elimination. Starting the backwards elimination process we looked at the p-values of the variables and removed them if they were deemed not significant. Through this process we removed the following variables, FormLiquid(Inject), FormOther, FormTablet,Price, and Reviews. The resultant model after would include the predictor variables of EaseOfUse, Effective, FormCream, FormLiquid(Drink), IndicationOnLabel, and TypeRx, with the response variable of Satisfaction as stated before.

Statistics from full model use for selection



```
Residual standard error: 3.425 on 633 degrees of freedom
Multiple R-squared:  0.7316, Adjusted R-squared:  0.727
F-statistic: 156.9 on 11 and 633 DF,  p-value: < 2.2e-16
```

## 4. Results

### Summary and Interpretations:

Final Model:

Satisfaction = -049258 + 0.11392 ( EaseOfUse) + 0.90651 (Effective) + 0.20885 (FormCream) + 0.16081 (FormLiquid (Drink)) + 0.17707 (IndicationOn Label) - 0.21956 (TypeRx)

From our model we can say that the following factors EaseOfUse, Effectiveness, Form (cream and liquid (Drink)), Indication (on Label), Type (Rx) contribute the most to user satisfaction of pharmaceutical drugs. This model shows us that a higher EaseOfUse rating and a higher Effectiveness rating both contribute to a higher overall Satisfaction rating. If the form is either cream such as an ointment or a drinkable liquid, then these will also contribute to a higher overall satisfaction rating. If the drug is a prescription, then it will contribute to a lower overall satisfaction rating. From the coefficients we can also see that effectiveness has the higher overall contribution to user satisfaction as it has the higher absolute value of a coefficient. Our AIC and BIC values decreased drastically from 916.8616 and 974.881 respectively before the model to 324.3032 and 352.6646 respectively after the model. This is good as the lower values show that the final model if better fitting and will be better for new data. Our R-squared value went from 0.727 in the full model to 0.8424 in the final model. This shows improvement in the model as a higher R-squared value shows strong relation between the response and predictor variables, a well fit model, and shows a model is better overall for predicting. We also ended up with a higher F-Statistic which shows increased levels of significance within the model.

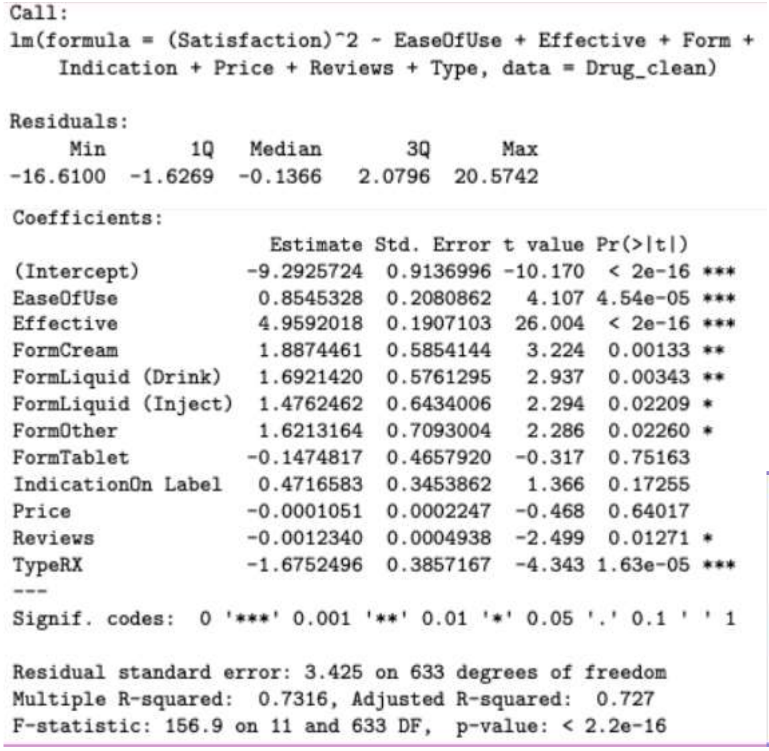
Vrushali Patel

Vallesia Pierre-Louis

Ryan Chen

STA 4164 0001

Full Model Statistics before

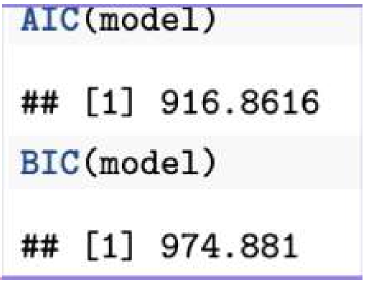


```
Call:
lm(formula = (Satisfaction)^2 ~ EaseOfUse + Effective + Form +
    Indication + Price + Reviews + Type, data = Drug_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-16.6100  -1.6269  -0.1366   2.0796  20.5742

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -9.2925724  0.9136996 -10.170  < 2e-16 ***
EaseOfUse            0.8545328  0.2080862   4.107 4.54e-05 ***
Effective            4.9592018  0.1907103  26.004  < 2e-16 ***
FormCream            1.8874461  0.5854144   3.224  0.00133 **
FormLiquid (Drink)   1.6921420  0.5761295   2.937  0.00343 **
FormLiquid (Inject)  1.4762462  0.6434006   2.294  0.02209 *
FormOther            1.6213164  0.7093004   2.286  0.02260 *
FormTablet          -0.1474817  0.4657920  -0.317  0.75163
IndicationOn Label   0.4716583  0.3453862   1.366  0.17255
Price               -0.0001051  0.0002247  -0.468  0.64017
Reviews             -0.0012340  0.0004938  -2.499  0.01271 *
TypeRX              -1.6752496  0.3857167  -4.343 1.63e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.425 on 633 degrees of freedom
Multiple R-squared:  0.7316, Adjusted R-squared:  0.727
F-statistic: 156.9 on 11 and 633 DF,  p-value: < 2.2e-16
```



```
AIC(model)

## [1] 916.8616
BIC(model)

## [1] 974.881
```

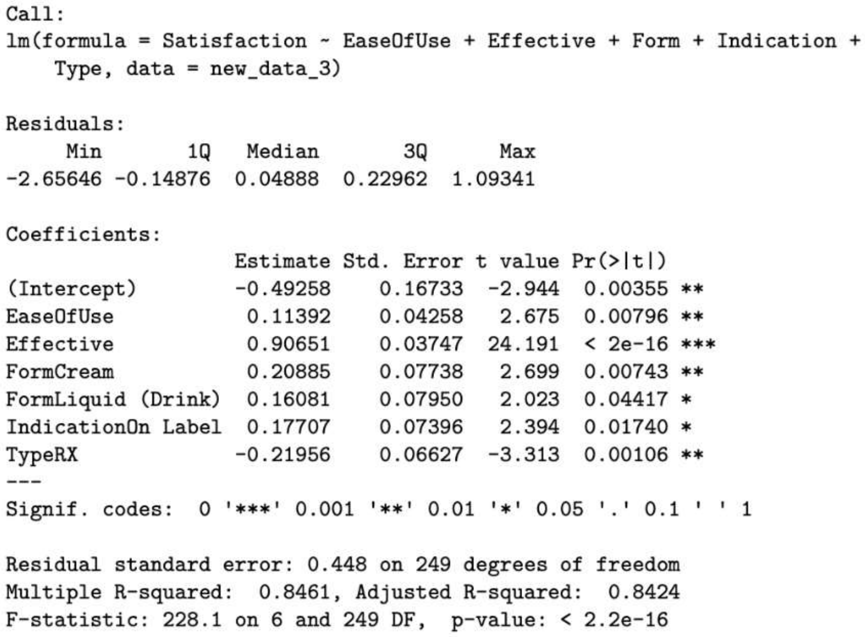
Final Model Statistics

Vrushali Patel

Vallesia Pierre-Louis

Ryan Chen

STA 4164 0001

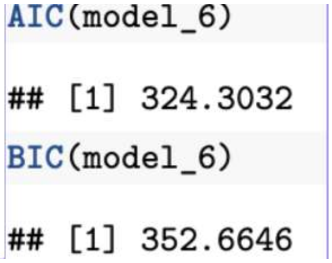


```
Call:
lm(formula = Satisfaction ~ EaseOfUse + Effective + Form + Indication +
    Type, data = new_data_3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.65646 -0.14876  0.04888  0.22962  1.09341

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.49258    0.16733  -2.944  0.00355 **
EaseOfUse           0.11392    0.04258   2.675  0.00796 **
Effective           0.90651    0.03747  24.191  < 2e-16 ***
FormCream           0.20885    0.07738   2.699  0.00743 **
FormLiquid (Drink)  0.16081    0.07950   2.023  0.04417 *
IndicationOn Label  0.17707    0.07396   2.394  0.01740 *
TypeRX             -0.21956    0.06627  -3.313  0.00106 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.448 on 249 degrees of freedom
Multiple R-squared:  0.8461, Adjusted R-squared:  0.8424
F-statistic: 228.1 on 6 and 249 DF,  p-value: < 2.2e-16
```



```
AIC(model_6)

## [1] 324.3032
BIC(model_6)

## [1] 352.6646
```

## 5. Conclusion and Limitations:

In conclusion, our results do satisfy the motivation of our study of patient satisfaction. Several factors contribute to user satisfaction, specifically, EaseOfUse, Effectiveness, Form (cream and liquid(Drink)), Indication(on Label), Type(Rx). Higher ratings in specifics like EaseOfUse and Effectiveness correspond to greater satisfaction, while the form of the drug being cream, or liquid tends to yield rates of higher satisfaction. These findings helped answer our initial question, along with our motivation to find out which factors contribute most to patient satisfaction. The factors that came out make sense as people would want a drug that is easy to use and effective. People prefer using easy to apply creams or easy to inject liquids that we can drink compared to other forms such as tablets to swallow or liquids to inject. Indication on the label is important so you can read what you are taking. Surprisingly for us a prescription led to lower user satisfaction, which we found confusing as it would be doctor approved, but this could be a factor in other things such as limiting accessibility to people. Because of our reduced AIC and BIC values, we have made many improvements, like for instance, we suggest better fitting and predictive capabilities. As our R-squared value increased in the final model, it indicates a stronger relationship between the response variables and predictors, but it can be improved upon. Some limitations we faced were some issues with data and that the solutions might be out of the realm of knowledge of this class. Our data failing normality was one struggle we were never able to figure out. In this project we learned more proficiency programing with R, how to properly handle datasets, more in-depth knowledge of statistical values and how they can be used in a project and how differing schedules can results in difficulty for working on group projects.

**References:**

Kaggle Dataset: Drug Performance Evaluation Dataset

**We, the project team members, certify that below is an accurate account of the percentage of effort contributed by each team member in the project and report.**

| Project Team Member | Percentage of Total Effort |
|---|---|
| Vallesia Pierre-Louis | 33.33 |
| Vrushali Patel | 33.33 |
| Ryan Chen | 33.33 |