Final Project: Preliminary Analysis Report on Drugs, Side Effects and Medical Condition dataset

GROUP 11: Yash S, Trusha S, Neer B

2024 – 2 – 16

Northeastern University: College of Professional Studies

Introduction

Introduction

The objective of this report is to explore and analyze the "Drugs, Side Effects, and Medical Condition" dataset to understand the relationship between drugs classes and their associated side effects, the medical condition that they are prescribed for, their ratings and number of reviews that they received, this is essential for evaluating their safety and efficiency. Initially we have performed ***visualization*** of the numerical variables and categorical variables to uncovering patterns between them and then trying to understand whether they have any significant association between them or not. For this we use statistical such as ***Correlation Analysis, Chi-squared tests and Analysis of Variance (ANOVA).*** Finally, our goal is to predict the drug efficiency(rating) of a particular drug using ***GLM and Regularization techniques***. By doing so we aim to derive valuable insights that can help pharmaceutical companies take informed decisions for the future of healthcare.

Dataset Overview

The dataset comprises of information on various drug types, their side effects, the medical condition that they are prescribed for, various safety classifications, their rating and the number of reviews.

Key Variables:

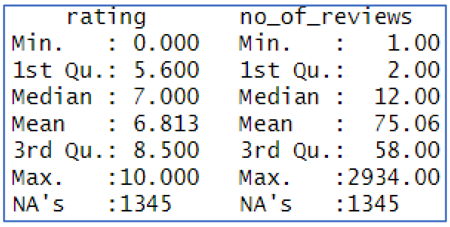
1. Nominal (Categorical variable):-
   a) **Drug Name (*drug_name*):** This variable contains the name of the drug.
   b) **Medical Condition(*medical_condition*):** This variable contains the medical condition that the drug is used to treat.
   c) **Side Effects (*side_effects*):** This variable contains the most common side effects associated with these drugs.
   d) **Drug Classes (*drug_classes*):** This variable classifies the drugs based on the medical condition they treat.
   e) **Prescription or Over-the-counter (*rx_OTC*):** This variable indicates whether the drug is prescription based or over the counter type.
   f) **Alcohol (*alcohol*):** This variable contains information that indicates whether the drug interacts with alcohol.
2. Ordinal (Categorical variable):-
   a) **Pregnancy Category (pregnancy_category):** This variable contains information about the safety classification of the drug's use during pregnancy,
   b) **Controlled Substances Act Schedule (*csa*):** This variable contains information about the scheduled classification of drugs under Controlled Substance Act (Schedule I-IV).
3. Continuous Variable:-
   a) **Rating (rating):** This variable contains the user ratings of the drug's effectiveness on a scale of 1 to 10.
4. Discrete Variable:-
   a) **Number of Reviews (*no_of_reviews*):** This variable contains the information about the number of user ratings of that particular drug.

Dataset Cleaning Steps

The dataset initially showed 1,345 values missing from ***ratings*** and ***number of reviews***, these rows with NA's were dropped, this was done to focus on the patterns within the available data. Many primary fields such as ***side effects, drug classes, prescription or over-the-counter, pregnancy categories*** and ***alcohol interactions*** had quite a few missing values as well that were not recognised by the interpreter but manually identified and imputed with "Unknown" to maintain data consistency and integrity.

Original Dataset: 17 fields, 2,931 rows.

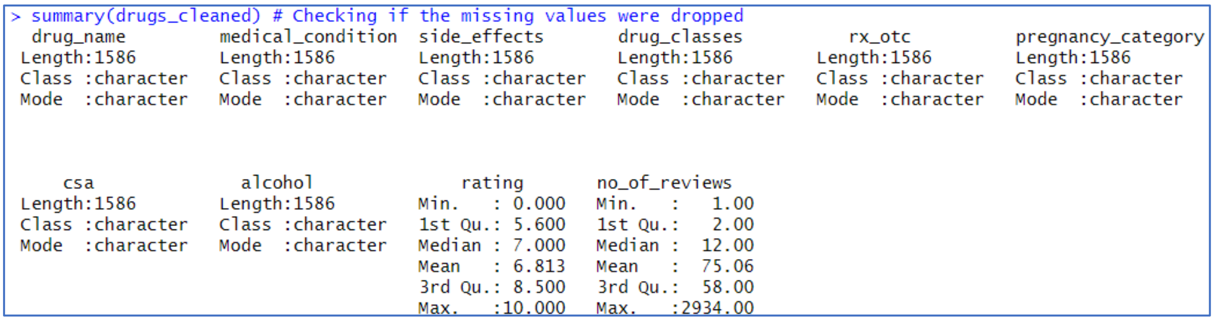
Cleaned Dataset: 10 fields, 1,420 rows.



```
     rating         no_of_reviews    
 Min.   : 0.000   Min.   :   1.00  
 1st Qu.: 5.600   1st Qu.:   2.00  
 Median : 7.000   Median :  12.00  
 Mean   : 6.813   Mean   :  75.06  
 3rd Qu.: 8.500   3rd Qu.:  58.00  
 Max.   :10.000   Max.   :2934.00  
 NA's   :1345     NA's   :1345     
```

Explanatory Data Analysis

1.Descriptive Statistics of Key Variables:



```
> summary(drugs_cleaned) # Checking if the missing values were dropped
  drug_name         medical_condition  side_effects       drug_classes          rx_otc          pregnancy_category
 Length:1586        Length:1586        Length:1586        Length:1586        Length:1586        Length:1586       
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  



     csa              alcohol              rating       no_of_reviews    
 Length:1586        Length:1586        Min.   : 0.000   Min.   :   1.00  
 Class :character   Class :character   1st Qu.: 5.600   1st Qu.:   2.00  
 Mode  :character   Mode  :character   Median : 7.000   Median :  12.00  
                                       Mean   : 6.813   Mean   :  75.06  
                                       3rd Qu.: 8.500   3rd Qu.:  58.00  
                                       Max.   :10.000   Max.   :2934.00  
```

As we can see that each following categorical variables have 1,586 entries:

a. Drug Name [Nominal]
b. Medical Condition [Nominal]
c. Side Effects [Nominal]
d. Drug Classes [Nominal]
e. Prescription [Nominal]
f. Pregnancy Category [Ordinal]
g. Controlled Substance Act (CSA) Schedule [Ordinal]
h. Alcohol [Nominal]

And we can interpret the numerical variables as follows:

a. Rating [Continuous]

- The ratings have a range of **1 to 10** and the average rating is **6.8** with a median of **7**.

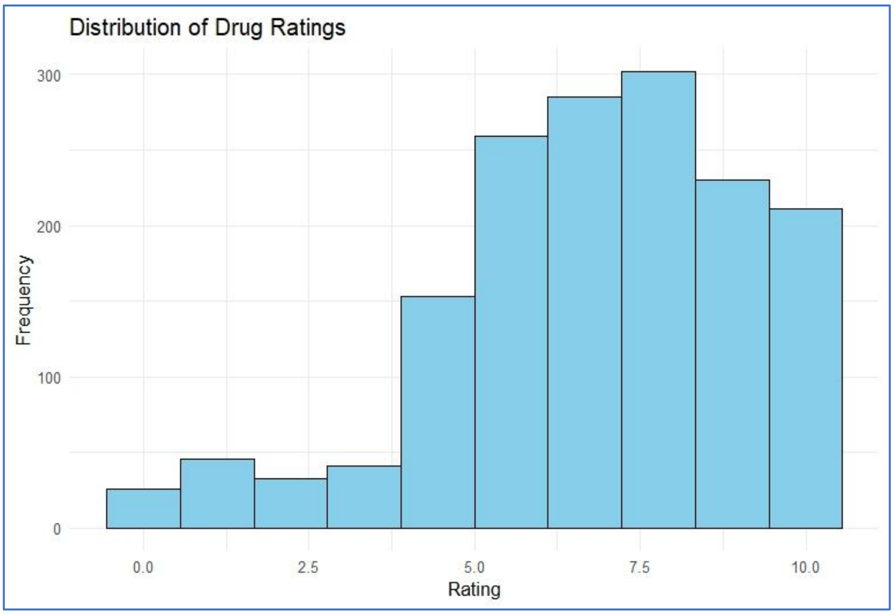
b. Number of Reviews [Discrete]

- The average number of reviews are 75 with a median of 12 and minimum & maximum number of reviews are 1 and 2,934 respectively.

2. Visualizations:

a. Drug Rating Distribution:

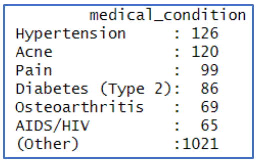
This graph shows the distribution of the drug ratings ranging from 1 to 10. As we can see the graph is skewed to the left with majority of the rating clustered near 7.5 (higher end of the scale). This graph suggests people only tend to share their experiences when they feel extremely positive or negative about the particular drug.



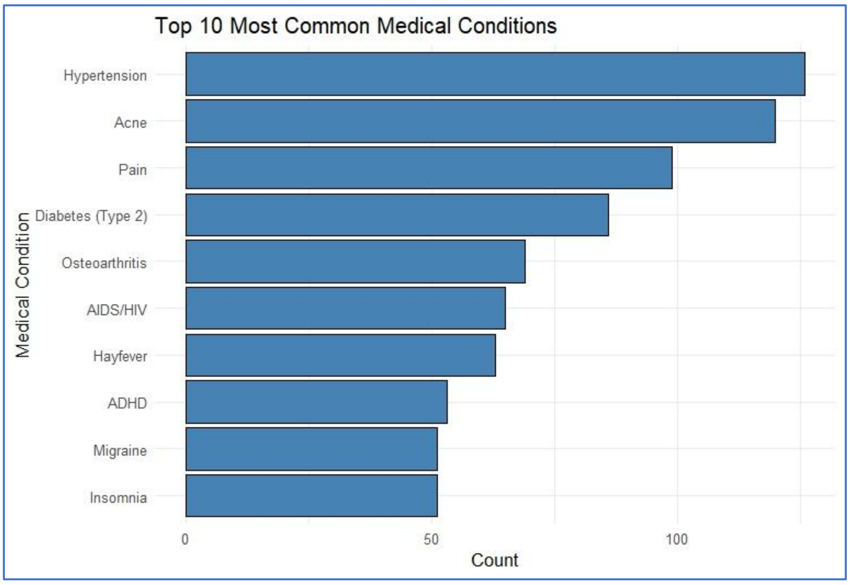
The chart shows distibution of drug ratings

b. Most Common Medical Conditions

This graph shows top ten identified most common medical conditions. These are Hypertension (126), Acne (120), Pain(99), Diabetes [Type 2] (86), Osteoarthritis (69), AIDS/HIV (65), Hayfever (63), ADHD (52), Migraine (50) and Insomnia (50). We can see that there is a significant drop from Hypertension (126) to Insomnia (50)



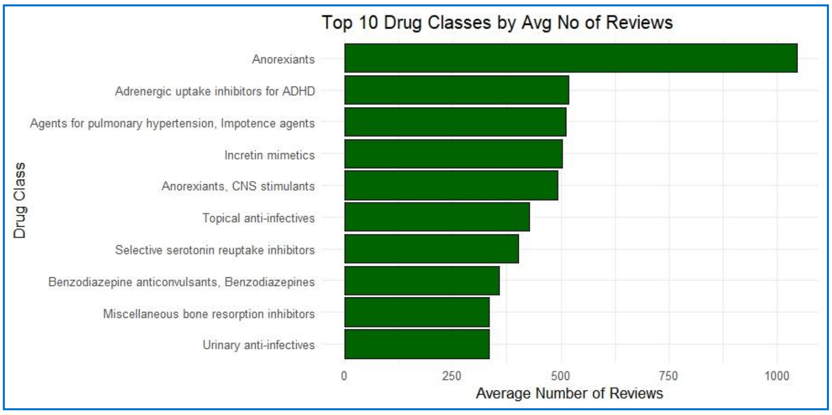
```
           medical_condition
Hypertension      : 126
Acne              : 120
Pain              :  99
Diabetes (Type 2):  86
Osteoarthritis    :  69
AIDS/HIV          :  65
(Other)           :1021
```



The chart shows top 10 most common medical conditions

c. **Distribution of drug classes by the average number of reviews:**

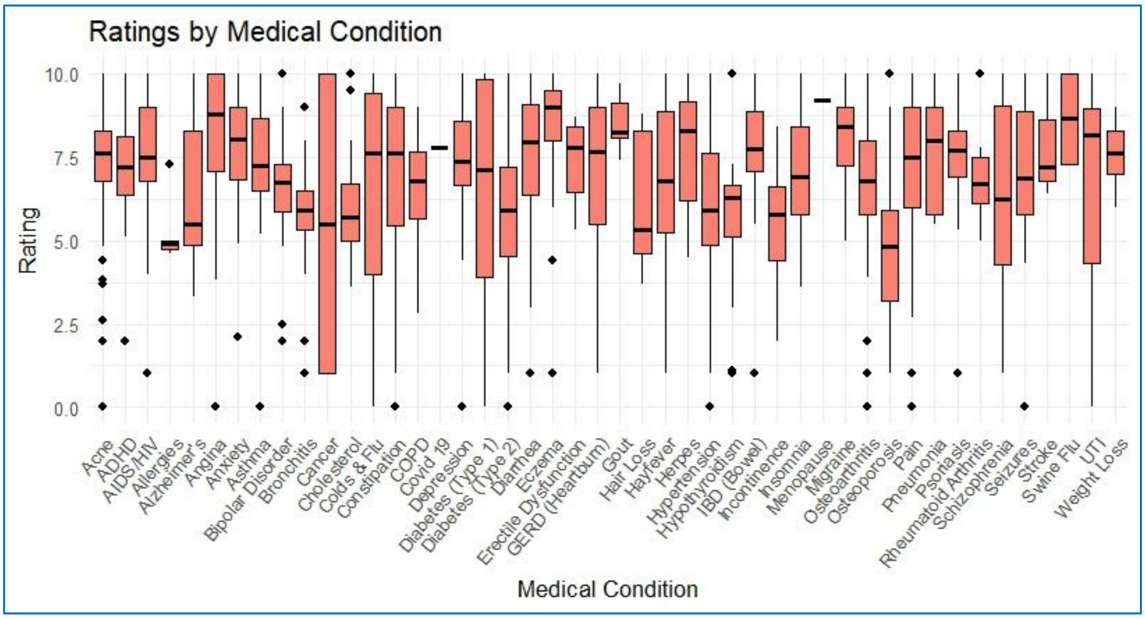
This graph indicates the distribution of the different drug classes by the average number of reviews by the users. We can see Anorexiants, adrenergic uptake inhibitors for ADHD and agents for pulmonary hypertension are top three drug classes that have been used by most of the users.



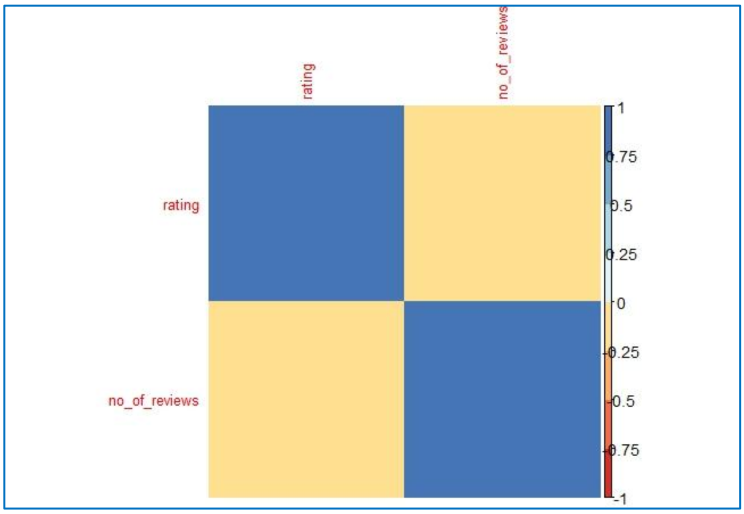
Distribution of drug classes by the average number of reviews

d. **Boxplot of Ratings by Medical Conditions:**

In the below chart we can see the rating for each medical conditions with their corresponding boxplots. We can see the some of the medical conditions such as Anxiety, Asthma, GERD (Heartburn), etc. have less variability in rating whereas medical conditions such as Cholesterol, Diabetes (Type 2), Hypertension, etc. have higher variability. We can also see that many of the medical conditions have lower outliers meaning most of them must have had some side effects and only a few of these medical conditions such as Bipolar Disorder, Hyperthyroidism, Osteoporosis, etc. have upper outliers indicating a large number of happy reviews from the customers.



Correlation Matrix:



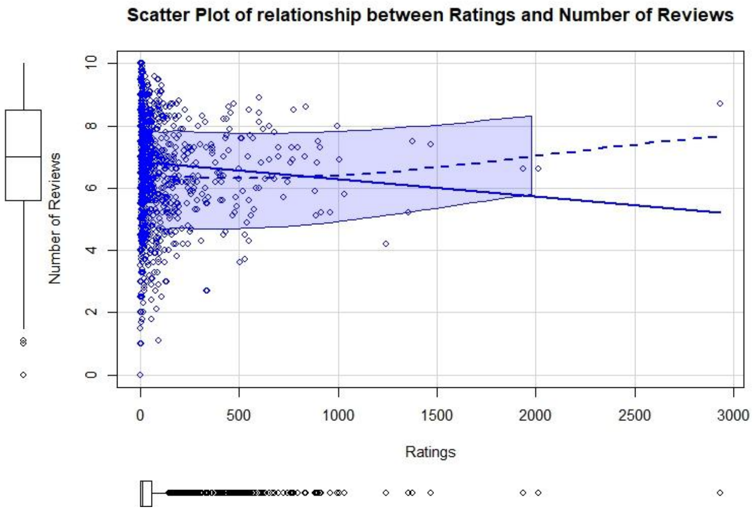
Correlation Matrix Plot

In this correlation matrix, we can see that two numerical variables are paired up with each other in every different possible way and that every combination is explored, each box is filled with only two colours. As we can see, the yellow blocks have opposite relationship, meaning when one goes up the other goes down, usually the darker the shade is the weaker the bond, but here we notice only one shade because we have only two numerical variables. The perfect diagonal blocks always show perfect a positive correlation since it is compared with itself.

From the above correlation matrix following were the key inferences made:

i. **Applications received (Apps) and Accepted Applications (Accept):**

- Low Correlation (-0.0421029)
- Scatterplot showing a positive correlation between Applications received and Accepted applications, this simply means that acceptance rate improves when more applications are received.



The graph indicates a negative correlation between Ratings and Number of Reviews.

Hypothesis Testing:

Chi-Square test for independence –

While performing these tests we found their approximations to be unreliable so we performed Cramer’s V test to compliment our test and raised our significance level to 90%.

Drug Classes and Medical Conditions

a) *Null Hypothesis(H_0):* There is **no** significant association between drug classes and medical condition.

Alternative Hypothesis(H_1): There is **a** significant association between drug classes and medical condition.

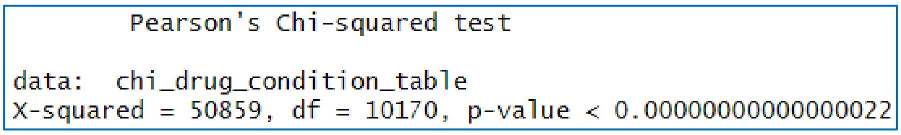
b) *Significance Level* = 0.10

Degree of Freedom = 10,170

Critical Value = *10,353.2*

c) *Chi-Square Value* = 50,859

p-value $< 2.2 * 10^{-16}$



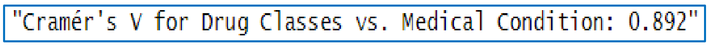
```
        Pearson's Chi-squared test

data:  chi_drug_condition_table
X-squared = 50859, df = 10170, p-value < 0.00000000000000022
```

d) The chi-squared test statistic value is ***greater than*** the critical value, and the p-value is ***less than*** the level of significance.

e) *Conclusion:*

Since the p-value is **less than** the level of significance, we **reject** the Null Hypothesis. There is enough evidence to support that there is **a** significant association between drug classes and medical conditions.

f) We further support our decision with Cramer’s V test:



```
"Cramér's V for Drug Classes vs. Medical Condition: 0.892"
```

The value is **0.892** which is very close to 1, this signifies a **strong correlation** between the drug classes and medical conditions.

Drug Classes and Medical Conditions

a) *Null Hypothesis(H_0):* There is **no** significant association between drug classes and side effects.

Alternative Hypothesis(H_1): There is **a** significant association between drug classes and side effects.

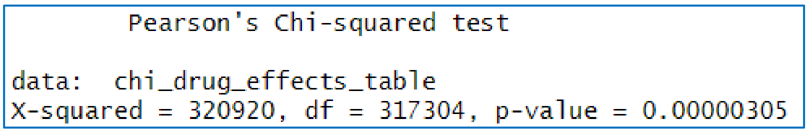
b) *Significance Level* = 0.10

Degree of Freedom = 317,304

Critical Value = *318,325.3*

c) *Chi-Square Value* = 320,920

p-value = 0.00000305

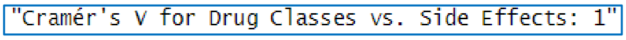


```
        Pearson's Chi-squared test

data:  chi_drug_effects_table
X-squared = 320920, df = 317304, p-value = 0.00000305
```

d) The chi-squared test statistic value is ***greater than*** the critical value, and the p-value is ***less than*** the level of significance.

e) *Conclusion:*
Since the p-value is **less than** the level of significance, we **reject** the Null Hypothesis. There is enough evidence to support that there is **a** significant association between drug classes and medical conditions.

f) We further support our decision with Cramer’s V test:



```
"Cramér's V for Drug Classes vs. Side Effects: 1"
```

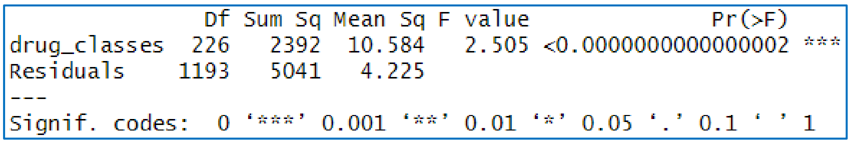
The value is 1, this signifies a **very strong correlation** between the drug classes and side effects.

One-way ANOVA – Drug Ratings across different Drug Classes

a) *Null Hypothesis(H_0):* There is no difference in mean ratings among different drug classes.
Alternative Hypothesis(H_1): There is a significant difference in mean ratings among different drug classes.

b) *Significance Level* = 0.10
Degree of Freedom Numerator = 226
Degree of Freedom Denominator = 1,193
Critical Value = 1.136

c) *F Value* = 2.505
p-value < $2*10^{-16}$



```
               Df Sum Sq Mean Sq F value               Pr(>F)    
drug_classes  226   2392  10.584   2.505 <0.0000000000000002 ***
Residuals    1193   5041   4.225                                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

d) The F test statistic value is ***greater than*** the critical value, and the p-value is ***less than*** the level of significance.

e) *Conclusion:*
Since the p-value is **less than** the level of significance, we **reject** the Null Hypothesis. There is enough evidence to support that there is **a** significant difference in mean ratings among different drug classes.

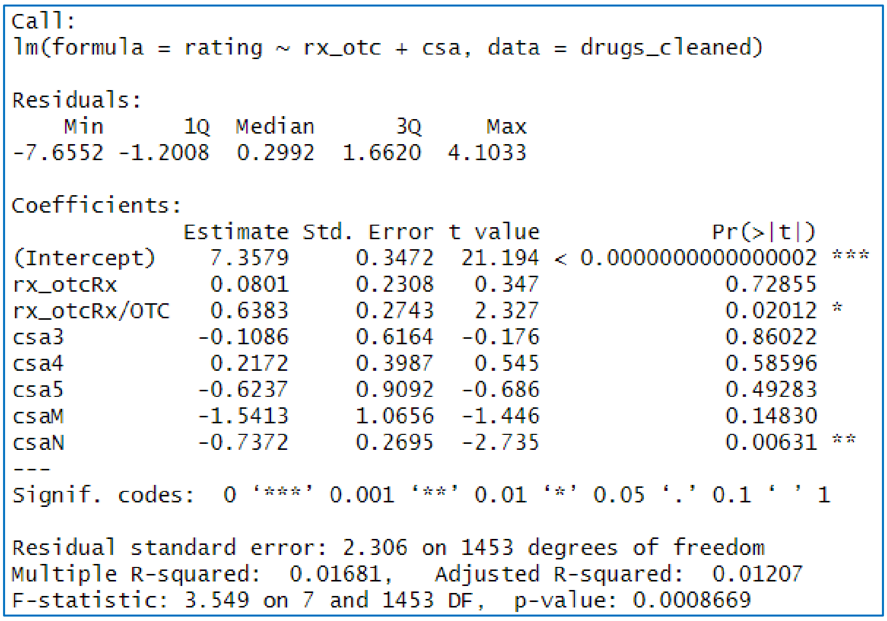
Regression Analysis:

a) Data Partitioning

The dataset was split into 65% of training data and 35% of testing data, this was done using the “caret” package ensuring a balanced distribution of public and private colleges (Target Variable).

b) Step-wise Feature Selection

The method for feature selection was step-wise selection which uses AIC-based forward and backward selection to select the most influential predictors.



```
Call:
lm(formula = rating ~ rx_otc + csa, data = drugs_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6552 -1.2008  0.2992  1.6620  4.1033 

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)    
(Intercept)    7.3579     0.3472  21.194 < 0.0000000000000002 ***
rx_otcRx       0.0801     0.2308   0.347              0.72855    
rx_otcRx/OTC   0.6383     0.2743   2.327              0.02012 *  
csa3          -0.1086     0.6164  -0.176              0.86022    
csa4           0.2172     0.3987   0.545              0.58596    
csa5          -0.6237     0.9092  -0.686              0.49283    
csaM          -1.5413     1.0656  -1.446              0.14830    
csaN          -0.7372     0.2695  -2.735              0.00631 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.306 on 1453 degrees of freedom
Multiple R-squared:  0.01681,	Adjusted R-squared:  0.01207 
F-statistic: 3.549 on 7 and 1453 DF,  p-value: 0.0008669
```

The regression equation for the above model is as follows:

$$\text{rating}=7.3579+0.0801\cdot \text{rx\_otc}+0.6383\cdot \text{rx\_otcRx/OTC}-0.1086\cdot \text{csa3}+0.2172\cdot \text{csa4}-0.6237\cdot \text{csa5}-1.5413\cdot \text{csaM}-0.7372\cdot \text{csaN}$$

We can interpret the logistic regression coefficients as follows:

i. **Intercept (7.3579):**

This value represents the expected value of ratings when all the other predictors are zero. Although this is not meaningful for any interpretation, it is part of the regression equation. NOTE: This was clearly seen in the first visualization.

ii. **Prescription or Over the counter(0.0801):**

For every one unit increase in rx_otc the ratings are also expected to increase.

iii. **Prescription or Over the counter(RxOTC) (0.6383):**

For every one unit increase in rx_otc(Rx_OTC) the ratings are also expected to increase.

iv. **CSA Schedule 3 (0.1086)**

For every one unit increase in CSA Schedule 3 the ratings are also expected to increase.

v. **CSA Schedule 4 (0.2172):**

For every one unit increase in CSA Schedule 4 the ratings are also expected to increase.

vi. **CSA Schedule 5 (0.6237):**

For every one unit increase in CSA Schedule 5 the ratings are also expected to increase.

vii. **CSA Schedule M (1.5413):**

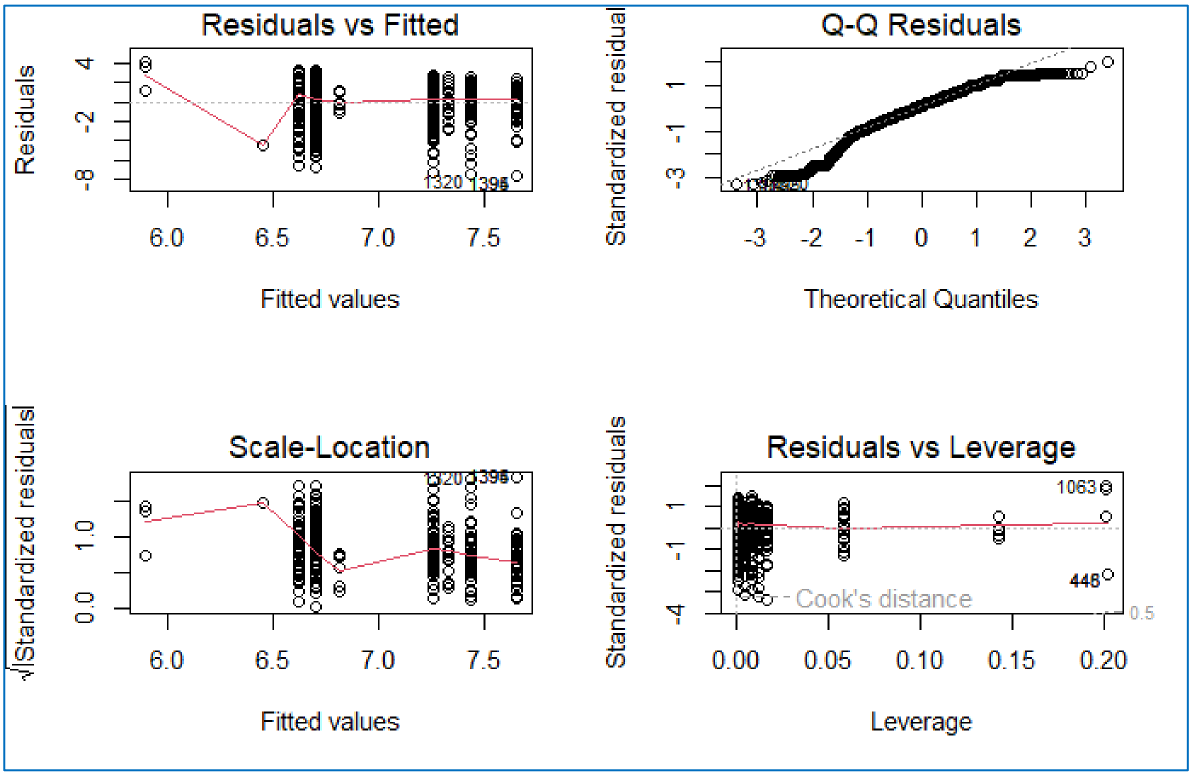
For every one unit increase in CSA Schedule M the ratings are also expected to increase.

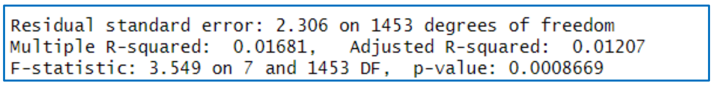
viii. **CSA Schedule N (0.7372):**

For every one unit increase in CSA Schedule N the ratings are also expected to increase.

c) Diagnostic Plots and Refinement

i. Residual vs Fitted: There is a visible trend of randomly scattered vertical points and they are not distributed alone the horizontal line of zero suggesting potential issues with linearity and homoscedasticity.
ii. Q-Q Residuals: The Q-Q plot shows deviation from the line, indicating the residuals are not perfectly normally distributed.
iii. Scale Location: We can see the points are not equally spread and vertically shaped, and the pattern suggests outliers.
iv. Residuals vs Leverage: There are very few points with high leverage that are potential outliers as shown by Cook's Distance, but these do not impact the regression model significantly.



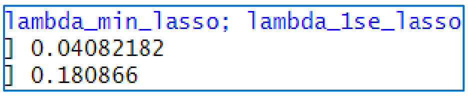


```
Residual standard error: 2.306 on 1453 degrees of freedom
Multiple R-squared:  0.01681,    Adjusted R-squared:  0.01207
F-statistic: 3.549 on 7 and 1453 DF,  p-value: 0.0008669
```

This graph confirms what we infer from the Multiplied R^2 value **(0.01681)** this model only predicts 1.68% variability which does not capture the data well, and the Adjusted R^2 value **(0.01207)** suggests that the model reflects lower explanatory powers. This warrants approach of the best regularization technique LASSO regression

d) LASSO Regression

The LASSO (Least Absolute Shrinkage and Selection Operator) Regression applies the L1 regularization, this shrinks the coefficients to zero, this effectively selects only the most important features and drops the rest.

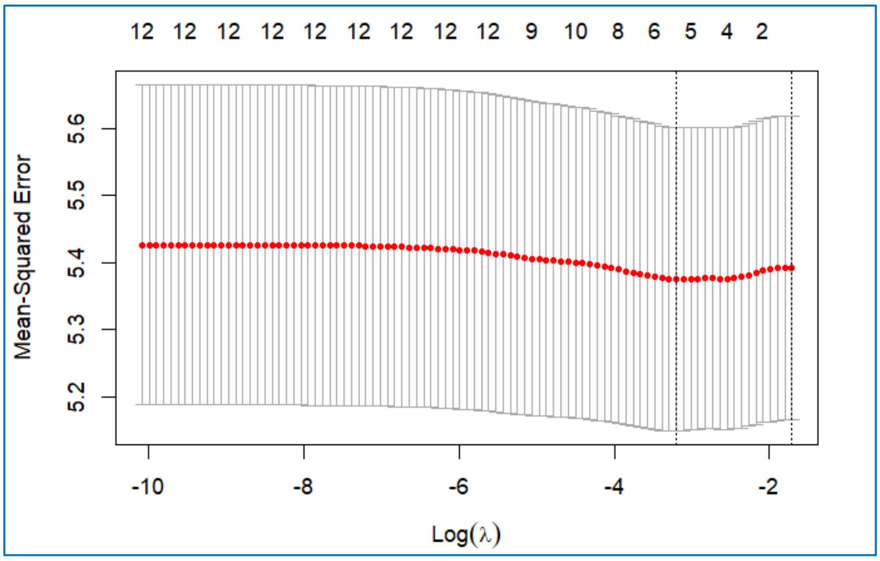


```
lambda_min_lasso; lambda_1se_lasso
] 0.04082182
] 0.180866
```

Lambda.min (0.0408): This value of lambda minimizes the MSE on the cross-validated data. In this case at lambda.min is equal to **0.0408**. This value results in the most complex model with best performance on the training data
Lambda.1se (0.1808): This value of lambda is within one standard error that minimizes the average of the cross-validation error. In this case lambda.1se is equal to **0.1808**. This value results in a simpler model that balances the bias and the variance, reducing overfitting.

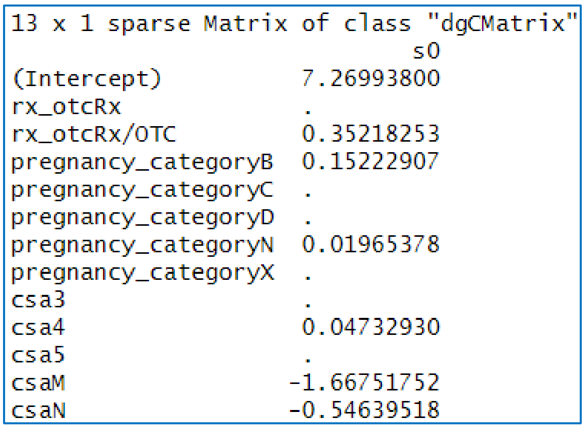
While lambda.min has lower regularization strength as compared to lambda.1se, the models trained with lambda.min will have the lowest possible error on the training data but the models trained with lambda.1se might generalise new data better as it employs a higher degree of regularization, which will help prevent overfitting.

Visualization:
When we plot this function and interpret the graph, we can see a trade-off between the Mean-Square Error and the logarithm of the regularization parameter [Log(λ)], the two vertical dashes line are the lambda.min [approx. log(0.0408) ≈-1.77] and lambda.1se [approx. log(0.1808) ≈ 0.56], as the log(λ) increases the mean-squared error stays relatively stable but begins to dip then rise sharply after a point indicating that increasing regularization strength can cause the model to underfit.



Graph showing relationship between MSE and Log(λ)

Model Evaluation:
Since we aim for the lowest possible error on the training data, we select the lambda.min as the optimal value of lambda to fit a Least Absolute Shrinkage and Selection Operator (LASSO) regression model, following are the interpretation of the variables selected.



```
13 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)          7.26993800
rx_otcRx             .
rx_otcRx/OTC         0.35218253
pregnancy_categoryB  0.15222907
pregnancy_categoryC  .
pregnancy_categoryD  .
pregnancy_categoryN  0.01965378
pregnancy_categoryX  .
csa3                 .
csa4                 0.04732930
csa5                 .
csaM                -1.66751752
csaN                -0.54639518
```

The regression equation for the above is:
y=7.27+0.35·rx_otcRx/OTC+0.15·pregnancy_categoryB+0.02·pregnancy_categoryN+0.05·csa4−1.67·csaM−0.55·csaN

We can interpret this equation as the follows:

i. **Intercept (7.27):** The starting value of the outcome when all predictors are zero.

ii. **Prescription or Over-the-Counter(Rx/OTC) (0.35):** A positive impact on the outcome by prescription and over-the-counter medication.

iii. **Pregnancy Category B (0.15):** Slight increase in the outcome due to drugs classified in pregnancy category B.

iv. **Pregnancy Category N (0.02):** A small positive effect of medications without specific pregnancy categories.

v. **CSA 4 (0.05):** Marginally raises the outcome due to Schedule IV controlled substances.

vi. **CSA M (-1.67):** Significant reduction in the outcome when a controlled substance in the "M" category is present.

vii. **CSA N (-0.55):** A notable negative influence on the outcome by drugs in the "N" category.

The following intercepts are set to zero as they do not significantly contribute in the model under the chosen regularization method.

i. *rx_otcRx*

ii. *pregnancy_categoryC*

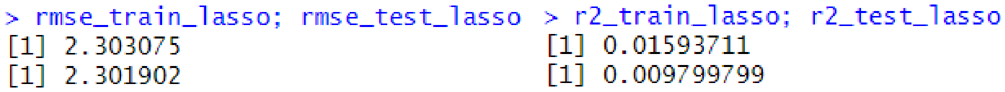
iii. *pregnancy_categoryB*

iv. *pregnancy_categoryX*

v. *csa3*

vi. *csa5*

Performance Evaluation:



```
> rmse_train_lasso; rmse_test_lasso  > r2_train_lasso; r2_test_lasso
[1] 2.303075                         [1] 0.01593711
[1] 2.301902                         [1] 0.009799799
```

i) Root Mean Squared Error (RMSE):

- Training: 2.303075
- Testing: 2.301902

ii) Coefficient of Determination (R^2):

- Training: 0.01593711
- Testing: 0.009799799

These values obtained are low, indicating that our LASSO model also has a low explanatory power for the drug ratings.

e) Model Evaluation

This concludes that both the multiple linear regression and LASSO models have low accuracy and explanatory power for predicting the drug rating (target variable), which leads to the suggestion, that the predictors used in the models may not be sufficient or appropriate for capturing the variability in the rating. Advanced modelling techniques such as Decision trees and Random forests can be used to boost the accuracy for our target variable.

Conclusion

Conclusion:

In this analysis, we successfully identified significant association between drug classes, medical conditions and side effects, highlighted the variations in the drug ratings, we developed a linear regression model and performed LASSO regression as well, and compared its performance metrics to figure out both models have low accuracy and explanatory powers for predicting the target. Decision trees and Random forests can help improve the accuracy and would be an asset.

Key Findings:

- We found weak correlation between ratings and number of reviews indicating that more reviews do not mean that the drug is rated higher.
- We noticed a strong correlation between the drug classes, medical conditions and the side effects.
- We determined using ANOVA that not all drug classes have the same rating; some tends to be lower and vice-versa.

Works Cited

- Bluman, A. (2018). *Elementary statistics: A step-by-step approach* (10th ed.). McGraw Hill.
- Kabacoff, R. I. (2022). *R in action: Data analysis and graphics with R and tidyverse* (3rd ed.). Manning Publications.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R. Springer.*
- R-bloggers. (2021, October 6). Lambda.min, lambda.1se, and cross-validation in LASSO (binomial response). R-bloggers. https://www.r-bloggers.com/2021/10/lambda-minlambda-1se-and-cross-validation-in-lasso-binomial-response/
- Kaggle. (n.d.). Drugs, Side Effects, and Medical Conditions Dataset. Retrieved from https://www.kaggle.com/datasets/jithinanievarghese/drugs-side-effectsand-medical-condition
- Eric, U. (2023). Drug Prescription Model [Kaggle notebook]. Retrieved February 15, 2025, from https://www.kaggle.com/code/ugonnaeric/drug-prescription-model
- Seo, H. (2023). Medicine Recommendation [Kaggle notebook]. Retrieved February 15, 2025, from https://www.kaggle.com/code/hyeonseo6101/medicine-recommendation

Appendix

R Code:

```
#Authors: Yash S, Trusha S, Neer B

#Created: 2025-01-30

#Edited: 2025-02-15

#Course: ALY6015

#Final Project


cat("\014") # clears console

rm(list = ls()) # clears global environment

try(dev.off(dev.list()["RStudioGD"]), silent = TRUE) # clears plots

try(p_unload(p_loaded(), character.only = TRUE), silent = TRUE) # clears packages

options(scipen = 100) # disables scientific notion for entire R session


library(pacman)

p_load(tidyverse, corrplot, RColorBrewer, caret, car, rcompanion, vcd, leaps, glmnet)


drugs <- read_csv("drugs_side_effects_drugs_com.csv")


# EDA

summary(drugs) # 1345 NA values found, others exist but is not shown


# Key columns identified:

# Categorical Variables: drug_name, medical_condition, side_effects, drug_classes, rx_otc,
pregnancy_category, csa, alcohol

# Numerical Variables: rating, no_of_reviews

# Missing values manually identified:

# side_effects, drug_classes, brand_names, rx_otc, pregnancy_category, alcohol, rating, and
no_of_reviews.
```

```
# Imputing NA values
drugs_cleaned <- drugs %>%
  mutate(
    side_effects = ifelse(is.na(side_effects), "Unknown", side_effects),
    drug_classes = ifelse(is.na(drug_classes), "Unknown", drug_classes),
    rx_otc = ifelse(is.na(rx_otc), "Unknown", rx_otc),
    pregnancy_category = ifelse(is.na(pregnancy_category), "Unknown", pregnancy_category),
    alcohol = ifelse(is.na(alcohol), "Unkonwn", alcohol),
    )

# Selecting the key columns for analysis
drugs_cleaned <- drugs %>%
  select(
    drug_name, medical_condition, side_effects, drug_classes, rx_otc, pregnancy_category, csa, rating, no_of_reviews, activity
  ) %>% drop_na
summary(drugs_cleaned) # Checking if the missing values were dropped

# Visualizations
# Distribution of Ratings
ggplot(drugs_cleaned, aes(x = rating)) +
  geom_histogram(bins = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Drug Ratings", x = "Rating", y = "Frequency") +
  theme_minimal()

# Top 10 most common medical conditions
drugs_cleaned %>%
  count(medical_condition, sort = TRUE) %>%
  top_n(10) %>%
  ggplot(aes(x = reorder(medical_condition, n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black") +
  coord_flip() +
  labs(title = "Top 10 Most Common Medical Conditions", x = "Medical Condition", y = "Count") +
  theme_minimal()


# Top 10 Drug Classes by average number of reviews
drugs_cleaned %>%
  group_by(drug_classes) %>%
  summarise(avg_reviews = mean(rating, na.rm = TRUE)) %>%
  top_n(10, avg_reviews) %>%
  ggplot(aes(x = reorder(drug_classes, avg_reviews), y = avg_reviews)) +
  geom_bar(stat = "identity", fill = "darkgreen", color = "black") +
  coord_flip() +
  labs(title = "Top 10 Drug Classes by Avg No of Reviews",
       x = "Drug Class", y = "Average Number of Reviews") +
  theme_minimal()


# Boxplot of ratings by medical condition
ggplot(drugs_cleaned, aes(x = medical_condition, y = rating)) +
  geom_boxplot(fill = "salmon", color = "black") +
  labs(title = "Ratings by Medical Condition", x = "Medical Condition", y = "Rating") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 50, hjust = 1))


# # A# Count the occurrences of each drug class
# top_10_counts <- head(sort(table(drugs_cleaned$drug_classes), decreasing = TRUE), 10)
#
# # Create the bar plot
# ggplot(data.frame(Drug_Class = names(top_10_counts), Count =
as.numeric(top_10_counts)),
#       aes(x = reorder(Drug_Class, Count), y = Count)) +
#   geom_bar(stat = "identity", fill = "skyblue") +
#   labs(title = "Top 10 Most Frequent Drug Classes",
#       x = "Drug Class",
#       y = "Count")+
#   theme_minimal() +
#   coord_flip()
################################################################################
#####
# Questions to Explore:
# 1. Are there significant correlations between variables (Numerical variables)?
# Selecting the numeric columns for correlation analysis
numeric_data <- drugs_cleaned %>% select(rating, no_of_reviews)
# Calculating the correlation matrix
correlation_matrix <- cor(numeric_data, use = "complete.obs")
print(correlation_matrix)
# Plot of correlation matrix
corrplot(correlation_matrix, method = "color",
        tl.cex = 0.7,
        number.cex = 0.7,
        col = brewer.pal(n = 8, name = "RdYlBu"))


# Scatter plot showing correlation between the Drug ratings and Number of Reviews
scatterplot(rating ~ no_of_reviews, data = drugs_cleaned, xlab = "Ratings (1-10)", ylab =
"Number of Reviews (Count)",
          main = "Scatter Plot of relationship between Ratings and Number of Reviews")


################################################################################
#####
```

```
# 2. Is there a significant association between specific drug classes and medical conditions?
# Is there a significant association between specific drug classes and certain side effects?
# Correlation Analysis for categorical (nominal) variables using Cramér's V
# Converting the variables as factors
drugs_cleaned <- drugs_cleaned %>%
  mutate(
    medical_condition = as.factor(medical_condition),
    side_effects = as.factor(side_effects),
    drug_classes = as.factor(drug_classes)
  )

# Drug Classes vs. Medical Condition
# Chi-square Test
# a)
# Null Hypothesis (H0): There is no significant association between drug classes and medical condition.
# Alternative Hypothesis (H1): There is a significant association between drug classes and medical condition
# b)
# Degrees of Freedom = 10170
# Critical Value = 10353.2
# c)
# Creating a contingency table

# Chi-Square Value (X-squared) = 55067
# p-value < 0.00000000000000022
# d)
# Chi-Square Value (55067) > Critical Value (10769.83)
# p-value (< 0.00000000000000022) < α (0.10)
# e)
# Conclusion:

# Since p-value (< 0.00000000000000022) is less than α = 0.10.

# We reject the Null Hypothesis.

# There is evidence to claim that there is a significant association

# between drug classes and medical conditions.


# Since Chi-squared approximation may be incorrect, we can further confirm this by

# calculating Cramér's V for drug classes vs. medical condition

v_test_drug_condition <- cramerV(chi_drug_condition_table)

print(paste("Cramér's V for Drug Classes vs. Medical Condition:",
round(v_test_drug_condition, 3)))


# Drug Classes vs. Side Effect

# Chi-square Test

# a)

# Null Hypothesis (H0): There is no significant association between drug classes and side
effects.

# Alternative Hypothesis (H1): There is a significant association between drug classes and side
effects.

# b)

# Degrees of Freedom = 361998

# Critical Value = 363398.7

# c)

# Creating a contingency table

chi_drug_effects_table <- table(drugs_cleaned$drug_classes, drugs_cleaned$side_effects)

# Chi-Square Independence test

chi_test_drug_effect <- chisq.test(chi_drug_effects_table)

chi_test_drug_effect

# Chi-Square Value (X-squared) = 368108

# p-value = 0.0000000000004621
# d)
# Chi-Square Value (368108) > Critical Value (363398.7 )
# p-value (0.0000000000004621) < α (0.05)
# e)
# Conclusion:
# Since p-value (0.0000000000004621) is less than α = 0.05.
# We reject the Null Hypothesis due to lack of evidence.
# There is a significant association between drug classes and side effects.

# Since Chi-squared approximation may be incorrect, we can further confirm this by
# calculating Cramér's V for drug classes vs. side effects
v_test_drug_effects <- cramerV(chi_drug_effects_table)
print(paste("Cramér's V for Drug Classes vs. Side Effects:", round(v_test_drug_effects, 3)))

#############################################################################
#####
# 3. Compare mean drug efficacy ratings across different drug categories to identify
# any significant variations.
# ANOVA
# a)
# H0: There is no difference in mean ratings among different drug classes.
# H1: There is a significant difference in mean ratings among different drug classes.
# b)
# Degree of Freedom Numerator = 226
# Degree of Freedom Denominator = 1193
# Critical Value = 1.13577
# c)
# One-way Anova
anova_drug_class <- aov(rating ~ drug_classes, data = drugs_cleaned)
summary(anova_drug_class)
```

```
# F-value = 2.505

# p-value <0.0000000000000002

# d)

# F-value(2.565) > Critical value(1.255)

# p-value (0.0000000000000002) < Significance level(0.05)

# e)

# Conclusion:

# Since F-value(2.565) is more than Critical value(1.255).

# We reject the Null Hypothesis


####################################################################################
#####

# End of Preliminary Analysis

####################################################################################
#####

# Future Exploration:

# Predicting Drug Effectiveness (Rating)

# Model: Linear Regression

# Regularization: LASSO (Least Absolute Shrinkage and Selection Operator)

# Response Variable: rating (continuous)

# Predictors: pregnancy_category, rx_otc

# Use Case: Determine how different drug features impact their rating.

####################################################################################
#####

# Splitting the data into train and test set

# Maintaining a % of event rate 70/30 split

set.seed(123)

trainIndex <- createDataPartition(drugs_cleaned$rating, p = 0.65, list = FALSE, times = 1)

train_set <- drugs_cleaned[trainIndex, ]

test_set <- drugs_cleaned[-trainIndex, ]
```

```
# Selecting the key columns for analysis (these columns ensure no noise is introduced to the model)

drugs_cleaned <- drugs %>%

  select(

    rx_otc, pregnancy_category, csa, rating,

  )


# Step-wise Feature Selection

model_step <- step(lm(rating ~ ., data = drugs_cleaned), direction = "both")

step_summary <- summary(model_step)

step_summary

step_summary$r.squared


# # All subset regression

# best_subset <- regsubsets(rating ~., data = drugs_cleaned, nvmax = 3)

# reg_summary <- summary(best_subset)

# reg_summary

# # Best model by Mallow's Cp and BIC

# which.min(reg_summary$cp)

# which.max(reg_summary$adjr2)


# Plot diagnostic graphs for the regression model

par(mfrow = c(2, 2))

plot(model_step)

dev.off()


# # Check for multicollinearity using VIF

# # Now, running VIF function

# vif(model_step)
```

```
############################################################################
#####
# LASSO Regression
x_train <- model.matrix(rating ~ ., data = train_set)[,-1]
y_train <- train_set$rating
x_test <- model.matrix(rating ~ ., data = test_set)[,-1]
y_test <- test_set$rating

# Finding best values for lambda
set.seed(123)
lasso_model <- cv.glmnet(x_train, y_train, alpha = 1)  # Alpha = 1 for LASSO
# Comparing Lambda values
lambda_min_lasso <- lasso_model$lambda.min
lambda_1se_lasso <- lasso_model$lambda.1se
lambda_min_lasso; lambda_1se_lasso

# Plotting the LASSO regression model
plot(lasso_model)

# Fitting a LASSO regression model with minimum lambda value
lasso_fit <- glmnet(x_train, y_train, alpha = 1, lambda = lambda_min_lasso)
# Checking Coefficients
coef(lasso_fit)

# Determining RMSE for training set
pred_train_lasso <- predict(lasso_fit, s = lambda_min_lasso, newx = x_train)
rmse_train_lasso <- sqrt(mean((y_train- pred_train_lasso)^2))

# Determining RMSE for testing set
```

```
pred_test_lasso <- predict(lasso_fit, s = lambda_min_lasso, newx = x_test)
rmse_test_lasso <- sqrt(mean((y_test- pred_test_lasso)^2))

rmse_train_lasso; rmse_test_lasso

# Determining R-squared for training data
# Computing R-squared for LASSO Regression
ss_res_lasso_train <- sum((y_train- pred_train_lasso)^2)
ss_tot_lasso_train <- sum((y_train- mean(y_train))^2)
r2_train_lasso <- 1- (ss_res_lasso_train / ss_tot_lasso_train)

# Determining R-squared for testing data
ss_res_lasso_test <- sum((y_test- pred_test_lasso)^2)
ss_tot_lasso_test <- sum((y_test- mean(y_test))^2)
r2_test_lasso <- 1- (ss_res_lasso_test / ss_tot_lasso_test)

r2_train_lasso; r2_test_lasso

###########################################################################
#####
###########################################################################
#####
# End of Final Project
```