Project title: Drugs, side effects and medical condition.

Tools used: R studio.

Objective:

To analyse the relationship between drugs, their side effects, medical conditions treated, and associated user ratings. This EDA project aims to uncover trends in treatment effectiveness and drug classification. Understanding these relationships can inform safer prescriptions and improve drug development.

Dataset overview:

- **Rows:** 2,931
- **Columns:** 17
- **Key Columns:** drug_name, medical_condition, side_effects, drug_classes, rating, no_of_reviews, activity, rx_otc, pregnancy_category

Step by step Guide:

Install and load libraries:

```
install.packages (c("tidyverse", "ggplot2", "dplyr", "readr", "janitor", "corrplot"))
```

Load CSV

```
df=read_csv("C:\\Users\\intel\\OneDrive\\Desktop\\unifiedmentor\\Drugs,Side
Effects\\drugs_side_effects_drugs_com.csv")
```

Clean column names

```
df=clean_names(df)
```

View basic structure

```
Glimpse (df)

Summary (df)
```

Count missing values

```
colSums (is.na(df))
```

Data cleaning and feature engineering

- Convert to numeric

```
df$activity = as.numeric(gsub("%", "", df$activity)) / 100
```

Handling Missing values

- Fill NA in 'rating' and 'no_of_reviews' with 0 or placeholder

```
df$rating[is.na(df$rating)] = 0

df$no_of_reviews[is.na(df$no_of_reviews)] = 0
```

- Fill 'side_effects' and 'related_drugs' with 'Unknown'

```
df$side_effects[is.na(df$side_effects)] = "Unknown"

df$related_drugs[is.na(df$related_drugs)] = "Unknown"
```

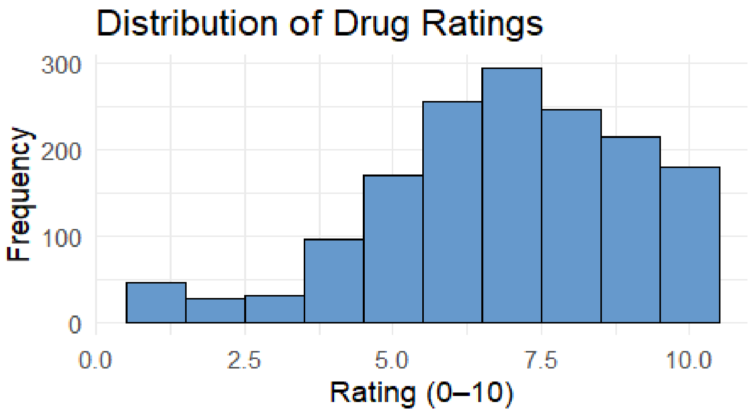
- Fill categorical with 'Unknown'

```
df$generic_name[is.na(df$generic_name)] = "Unknown"
```

```
df$drug_classes[is.na(df$drug_classes)] = "Unknown"
df$rx_otc[is.na(df$rx_otc)] <- "Unknown"
df$pregnancy_category[is.na(df$pregnancy_category)] = "Unknown"
df$alcohol[is.na(df$alcohol)] = "0"
df$alcohol[df$alcohol == "X"] = "1"
df$alcohol = as.numeric(df$alcohol)
summary (df)
```

Distribution of drug rating

```
ggplot(df %>% filter(rating > 0), aes(x = rating)) +
  geom_histogram(binwidth = 1, fill = "#6699CC", color = "black") +
  labs(title = "Distribution of Drug Ratings",
       x = "Rating (0-10)",
       y = "Frequency") +
  theme_minimal()
```



- Most drug ratings fall between 6 and 9, with a strong concentration at 7 and 8. This suggests that users generally rate their medications positively, despite the presence of common side effects like hives or breathing issues.

Top side effects

- Split side effects text and count frequency

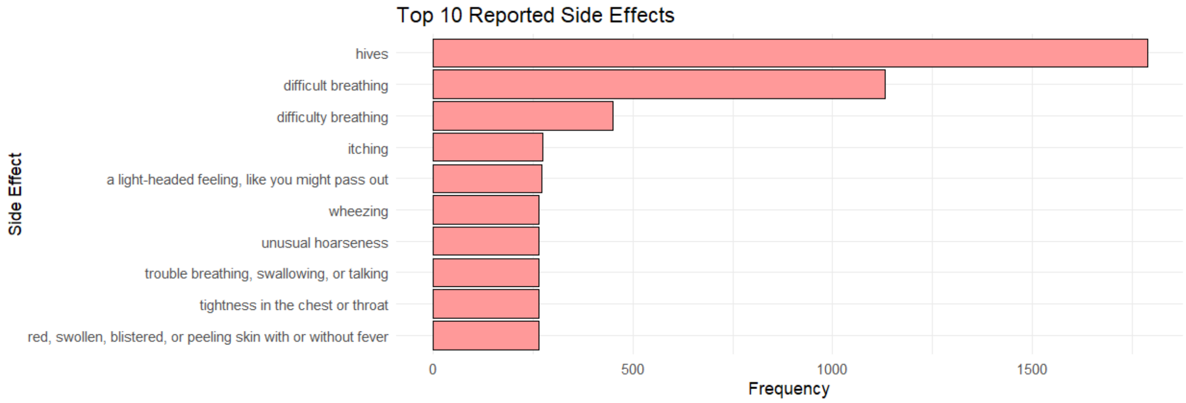
```
Library (tidyverse)
top_side_effects =df %>%
  filter(!is.na(side_effects)) %>%
  mutate(side_effects = strsplit(side_effects, ";")) %>%
  unnest(side_effects) %>%
  mutate(side_effects = str_trim(tolower(side_effects))) %>%
  count(side_effects, sort = TRUE)
```

- View top 10

```
head(top_side_effects, 10)
```

- Plot: Top 10 Side Effects

```r
library(ggplot2)

top_side_effects %>%

  top_n(10, n) %>%

  ggplot(aes(x = reorder(side_effects, n), y = n)) +

  geom_bar(stat = "identity", fill = "#FF9999", color = "black") +

  coord_flip() +

  labs(title = "Top 10 Reported Side Effects",

       x = "Side Effect",

       y = "Frequency") +

       theme_minimal()
```



- This chart highlights the most frequently reported side effects. Hives is the most common, followed by breathing difficulties and itching — these flags were used in further feature engineering.

Top drug classes

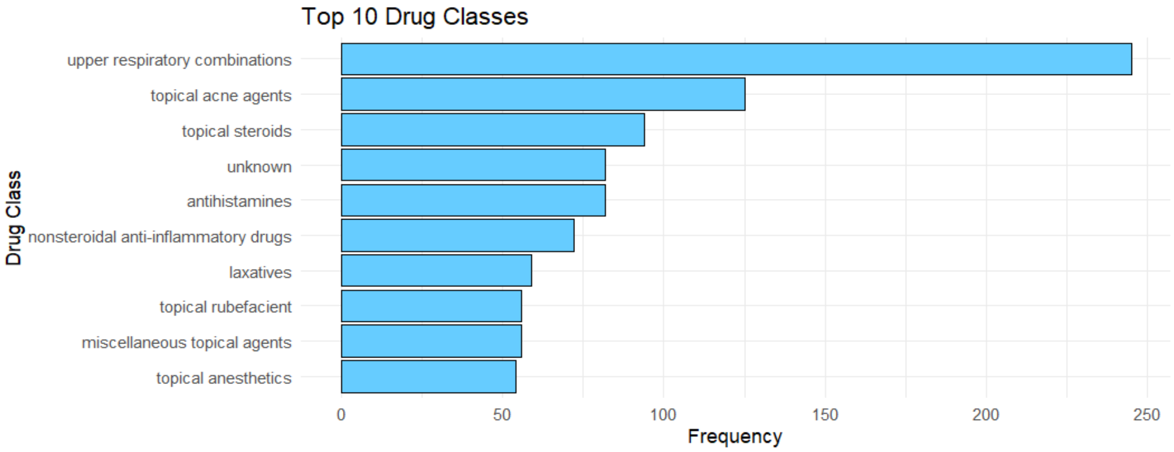
```r
top_drug_classes = df %>%

  filter(!is.na(drug_classes)) %>%

  mutate(drug_classes = strsplit(drug_classes, ",")) %>%

  unnest(drug_classes) %>%

  mutate(drug_classes = str_trim(tolower(drug_classes))) %>%

  count(drug_classes, sort = TRUE)
```

- ➢ View top 10

```r
head(top_drug_classes, 10)
```

- ➢ Plot: Top 10 Drug Classes

```r
top_drug_classes %>%

  top_n(10, n) %>%

  ggplot(aes(x = reorder(drug_classes, n), y = n)) +

  geom_bar(stat = "identity", fill = "#66CCFF", color = "black") +

  coord_flip() +

  labs(title = "Top 10 Drug Classes",

       x = "Drug Class",
       y = "Frequency") +
  theme_minimal()
```



- Upper respiratory combinations, topical acne agents, and topical steroids were the most common classes, indicating a prevalence of conditions like cold/flu and acne.

Top medical condition

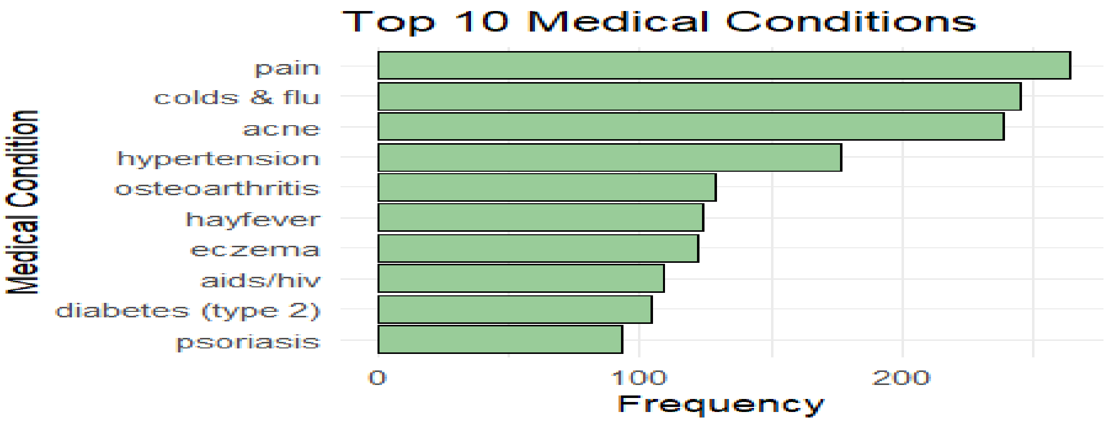
```
top_medical_conditions <- df %>%
  count(medical_condition = tolower(medical_condition), sort = TRUE)
```

➢ View top 10

```
head(top_medical_conditions, 10)
```

➢ Plot: Top 10 Medical Conditions

```
top_medical_conditions %>%
  top_n(10, n) %>%
  ggplot(aes(x = reorder(medical_condition, n), y = n)) +
  geom_bar(stat = "identity", fill = "#99CC99", color = "black") +
  coord_flip() +
  labs(title = "Top 10 Medical Conditions",
       x = "Medical Condition",
       y = "Frequency") +
  theme_minimal()
```



- Pain, colds & flu, and acne were the most commonly treated conditions, which aligns with the most frequent drug classes and side effects.

Feature engineering

- Lowercase for consistent pattern matching

```
df$side_effects = tolower(df$side_effects)

df$drug_classes = tolower(df$drug_classes)

df$medical_condition = tolower(df$medical_condition)
```

- Top 3 Side Effects Flags

```
df$has_hives = grepl("hives", df$side_effects)

df$has_difficult_breathing = grepl("difficult breathing|difficulty breathing", df$side_effects)

df$has_itching = grepl("itching", df$side_effects)
```

- Top 3 Drug Class Flags

```
df$is_upper_respiratory = grepl("upper respiratory combinations", df$drug_classes)

df$is_topical_acne = grepl("topical acne agents", df$drug_classes)

df$is_topical_steroid = grepl("topical steroids", df$drug_classes)
```

- Top 3 Medical Condition Flags

```
df$has_pain = grepl("pain", df$medical_condition)

df$has_colds_flu = grepl("colds & flu", df$medical_condition)

df$has_acne = grepl("acne", df$medical_condition)
```

- Create a numeric-only dataframe for correlation

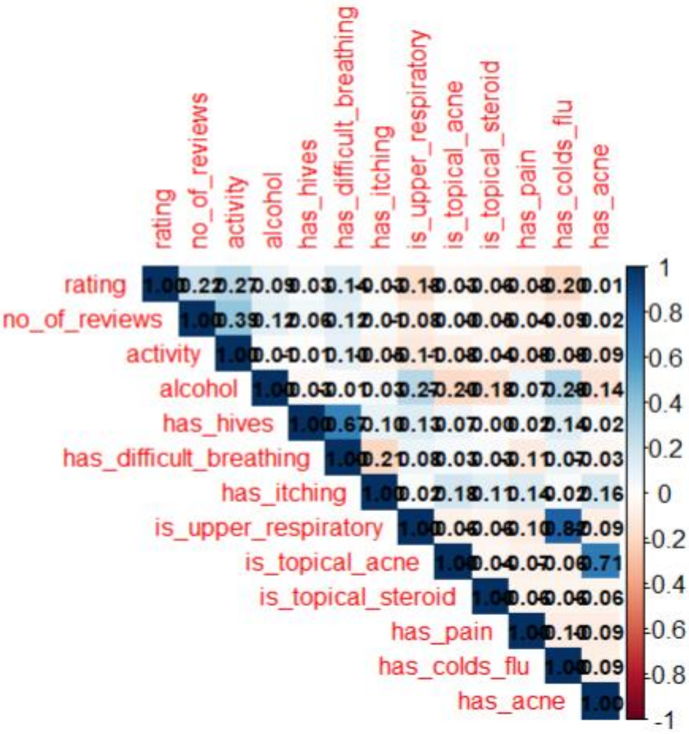
```
df_numeric = df %>%

  select(rating, no_of_reviews, activity, alcohol,

         has_hives, has_difficult_breathing, has_itching,

         is_upper_respiratory, is_topical_acne, is_topical_steroid,

         has_pain, has_colds_flu, has_acne) %>%

  mutate(across(everything(), as.numeric))  # Convert to numeric

install.packages("corrplot")

library(corrplot)
```

Compute correlation matrix

```
cor_matrix = cor(df_numeric, use = "complete.obs")
```

- Plot heatmap

```
corrplot(cor_matrix, method = "color", type = "upper",

         tl.cex = 0.8, number.cex = 0.7, addCoef.col = "black")
```



- Rating has moderate positive correlation with activity (0.27) and review count (0.22), while common side effects show weak or no correlation with ratings.

Conclusion:

This analysis uncovered meaningful insights about drugs, side effects, and the medical conditions they treat. Despite frequent reporting of side effects like hives and breathing difficulty, user ratings remain generally positive, indicating tolerance or perceived effectiveness. Certain drug classes, like topical acne agents, were strongly aligned with their corresponding conditions, validating the dataset's consistency. Moderate correlations between rating and factors like review count and activity suggest they may influence perceived drug effectiveness.

Overall, the analysis demonstrates how structured EDA, feature engineering, and correlation mapping can help identify patterns in health-related datasets and guide further modeling or medical research.