Project Title	Drugs, Side Effects and Medical Condition
Tools	Google Colab
Technologies	Machine Learning, Python, SQL, Excel
Domain	Data Science
Project Difficulties level	Intermediate

Name: Sachin Rathod

UM ID: UMID07112569651

Ph. No: +91 8310619278

Email: sachinmrathod067@gmail.com

Internship Role: Data Science

Contents

Introduction:

The pharmaceutical and healthcare industry generates vast amounts of data related to drugs, their medical applications, side effects, safety classifications, and patient feedback. Analyzing this data effectively is critical for improving drug safety, enhancing treatment outcomes, and supporting informed decision-making for healthcare professionals and patients alike.

This project, **"Drug Recommendation and Risk Analysis Using Side Effects and Medical Conditions"**, aims to analyze and extract meaningful insights from a real-world dataset containing information about various drugs, the medical conditions they treat, associated side effects, regulatory classifications, and user ratings. The dataset has been sourced from publicly available drug information platforms and includes both clinical and user-generated attributes, making it suitable for exploratory data analysis and predictive modeling.

The primary focus of this project is to understand the relationship between **drug effectiveness, safety risks, and patient perception**. By combining exploratory data analysis (EDA), feature engineering, and machine learning techniques, the project evaluates how factors such as pregnancy category, alcohol interaction, controlled substance classification (CSA), and number of reviews influence drug ratings. A risk score is also engineered to quantify potential safety concerns associated with drug usage.

Additionally, this project incorporates professional data visualizations to present trends and patterns in a clear and interpretable manner. A machine learning model is developed to predict drug ratings based on selected features, demonstrating how data-driven approaches can support drug evaluation and recommendation systems.

Overall, this project showcases the application of **data analytics and machine learning in healthcare**, emphasizing data cleaning, visualization, feature engineering, and predictive modeling. The insights derived from this analysis can assist in identifying high-risk drugs, understanding patient feedback trends, and supporting evidence-based decision-making in pharmaceutical analytics.

Objectives:

The objective of this project is to perform an in-depth analysis of drugs, their side effects, and the medical conditions they are prescribed for, using data analytics and machine learning techniques. The project aims to derive actionable insights that help evaluate drug effectiveness, safety risks, and patient perception.

The specific objectives of this project are:

- To explore and understand the distribution of drugs across various medical conditions and drug classes.
- To analyze the relationship between drug ratings and key safety factors such as pregnancy category, alcohol interaction, and Controlled Substances Act (CSA) classification.
- To engineer a risk score that quantifies potential safety concerns associated with drug usage.
- To identify patterns and trends in user ratings and review counts across different medical conditions.
- To develop a machine learning model to predict drug ratings based on selected features such as medical condition, number of reviews, and risk score.
- To visualize insights using professional graphs and charts for effective communication of findings.
- To demonstrate the application of data-driven techniques in healthcare and pharmaceutical analytics.

Data Science Application:

- **Exploratory Data Analysis (EDA)**
  To understand the distribution of drugs across medical conditions, drug classes, and safety categories.
- **Risk Analysis**
  To evaluate drug safety using pregnancy category, alcohol interaction, and Controlled Substances Act (CSA) classifications.
- **Trend Analysis**
  To identify patterns in drug usage, ratings, and review counts across different medical conditions.
- **Predictive Modeling**
  To build a machine learning model for predicting drug ratings based on medical condition, risk score, and user engagement.

- **Feature Engineering**
  To create a composite risk score that quantifies potential safety concerns associated with drugs.
- **Data Visualization & Insight Generation**
  To present analytical findings using professional plots for effective interpretation and decision-making.

Column Descriptors:

The dataset used in this project contains detailed information about pharmaceutical drugs, their medical applications, safety attributes, and user feedback. Each column provides specific insights that contribute to the analysis and predictive modeling tasks.

Column Details

- **Drug_name**: Name of the drug as commonly prescribed or marketed.
- **Generic_name**: The chemical or generic name of the drug, independent of brand names.
- **Medical_condition**: The medical condition or disease for which the drug is prescribed.
- **Medical_condition_description**: A brief textual description of the medical condition, including symptoms and related information.
- **Side_effects**: Reported side effects associated with the drug, ranging from mild to severe.
- **Drug_classes**: Classification of the drug based on its pharmacological category or mechanism of action.
- **Brand_names**: Commercial brand names under which the drug is available in the market.
- **Activity**: Represents recent user activity or popularity of the drug, expressed as a percentage.
- **RX_OTC**: Indicates whether the drug requires a prescription (**Rx**), is available over-the-counter (**OTC**), or both.
- **Pregnancy_Category**: FDA pregnancy risk classification of the drugs
  - **A, B** – Low risk
  - **C** – Moderate risk
  - **D, X** – High risk
  - **N** – Not classified
- **CSA**: Controlled Substances Act (CSA) schedule indicating the potential for drug abuse and regulatory control.

- **Alcohol**: Indicates whether the drug interacts with alcohol (**1 = Interaction**, **0 = No Interaction**).
- **Related_drugs**: List of drugs related by chemical structure, therapeutic use, or treatment similarity.
- **Rating**: Average user rating of the drug on a scale of 1 to 10, reflecting effectiveness and tolerability.
- **No_of_reviews**: Total number of user reviews submitted for the drug.
- **Drug_link**: URL linking to detailed drug information.
- **Medical_condition_url:** URL linking to detailed information about the medical condition.

Ethical Considerations and Data Privacy:

This project follows ethical data science practices in the analysis of healthcare-related data. The dataset used consists of publicly available and anonymized drug information and aggregated user feedback, with no inclusion of personally identifiable information (PII). As a result, the analysis complies with data privacy standards and does not violate regulations such as GDPR or HIPAA. The insights generated are intended solely for educational and analytical purposes and should not be interpreted as medical advice. Potential biases arising from subjective user ratings and uneven data distribution are acknowledged and addressed through careful data preprocessing and responsible interpretation. All data handling was performed in a secure environment, ensuring transparency, fairness, and responsible communication of findings.

Drugs, Side Effects and Medical Condition

Project Overview:

The **Drug Recommendation and Risk Analysis Using Side Effects and Medical Conditions** project focuses on analyzing pharmaceutical drug data to understand drug effectiveness, safety risks, and patient perception. The project utilizes a real-world dataset containing information about drugs, their medical applications, side effects, regulatory classifications, and user ratings.

The analysis begins with data cleaning and preprocessing to handle missing values, standardize attributes, and prepare the dataset for analytical and machine learning tasks. Exploratory Data Analysis (EDA) is performed to identify trends in drug usage across medical conditions, examine rating distributions, and evaluate the impact of safety factors such as pregnancy category, alcohol interaction, and Controlled Substances Act (CSA) classification.

Feature engineering techniques are applied to create a composite risk score that quantifies potential drug safety concerns. Professional data visualizations are used to effectively communicate patterns and relationships within the dataset. Additionally, a machine learning model is developed to predict drug ratings based on selected features, demonstrating the application of predictive analytics in healthcare data science.

The project is implemented using Python in a Google Colab environment, ensuring reproducibility and scalability. Overall, this project showcases how data science techniques can be applied to healthcare and pharmaceutical datasets to support informed decision-making, risk assessment, and analytical insight generation.

Project Steps:

1)Understanding the Problem:

The first step of the project is to clearly understand the problem of analyzing pharmaceutical drug data. The focus is on evaluating drug effectiveness, safety risks, and user ratings. Key factors such as side effects, medical conditions, and regulatory classifications are identified. This step helps in defining the scope and objectives of the project.

2) Dataset Collection and Preparation

The dataset is collected from publicly available sources related to drugs and medical conditions. Initial preparation involves loading the dataset and understanding its structure, columns, and data types. Any inconsistencies in the data format are identified at this stage. This step ensures the dataset is ready for further analysis

3) Data Cleaning and Preprocessing

Data cleaning is performed to handle missing values, remove inconsistencies, and correct data types. Categorical variables are encoded, and numerical features are standardized where required. This step improves data quality and ensures accurate analysis. Clean data is essential for reliable model performance.

4) Exploratory Data Analysis and Visualization

Exploratory Data Analysis (EDA) is carried out to identify trends and patterns in the dataset. Visualizations such as bar charts, box plots, and heatmaps are used to understand relationships between variables. This step helps in uncovering insights related to drug ratings, medical conditions, and safety factors. EDA also supports better feature selection.

5) Feature Engineering and Model Evaluation

New features such as a risk score are created using safety-related attributes like pregnancy category and alcohol interaction. A machine learning model is then trained to predict drug ratings. The model's performance is evaluated using appropriate metrics to assess accuracy and reliability. Final results are interpreted and documented for reporting.

Source code and Output

In [1]: # Import Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, r2_score
```

In [2]: # Load the dataset

```
data = pd.read_csv('/content/drugs_side_effects_drugs_com.csv')
data.head()
```

	drug_name	medical_condition	side_effects	generic_name	drug_classes	brand_names	activity	rx_otc	pregnancy_category	csa	alcohol	related_drugs
0	doxycycline	Acne	(hives, difficult breathing, swelling in your ...	doxycycline	Miscellaneous antimalarials, Tetracyclines	Acticlate, Adoxa CK, Adoxa Pak, Adoxa TT, Alod...	87%	Rx	D	N	X	amoxicillin: https://www.drugs.com/amoxicillin...
1	spironolactone	Acne	hives ; difficulty breathing; swelling of your...	spironolactone	Aldosterone receptor antagonists, Potassium-sp...	Aldactone, CaroSpir	82%	Rx	C	N	X	amlodipine: https://www.drugs.com/amlodipine.h...
2	minocycline	Acne	skin rash, fever, swollen glands, flu-like sym...	minocycline	Tetracyclines	Dynacin, Minocin, Minolira, Solodyn, Ximino, V...	48%	Rx	D	N	NaN	amoxicillin: https://www.drugs.com/amoxicillin...
3	Accutane	Acne	problems with your vision or hearing; muscle o...	isotretinoin (oral)	Miscellaneous antineoplastics, Miscellaneous u...	NaN	41%	Rx	X	N	X	doxycycline: https://www.drugs.com/doxycycline...
4	clindamycin	Acne	hives ; difficult breathing; swelling of your ...	clindamycin topical	Topical acne agents, Vaginal anti-infectives	Cleocin T, Clindacin ETZ, Clindacin P, Clindag...	39%	Rx	B	N	NaN	doxycycline: https://www.drugs.com/doxycycline...

In [3]:

```
data.info()

data.shape
```

(2931, 17)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2931 entries, 0 to 2930
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   drug_name                       2931 non-null   object
 1   medical_condition               2931 non-null   object
 2   side_effects                    2807 non-null   object
 3   generic_name                    2888 non-null   object
 4   drug_classes                    2849 non-null   object
 5   brand_names                     1718 non-null   object
 6   activity                        2931 non-null   object
 7   rx_otc                          2930 non-null   object
 8   pregnancy_category              2702 non-null   object
 9   csa                             2931 non-null   object
 10  alcohol                         1377 non-null   object
 11  related_drugs                   1462 non-null   object
 12  medical_condition_description   2931 non-null   object
 13  rating                          1586 non-null   float64
 14  no_of_reviews                   1586 non-null   float64
 15  drug_link                       2931 non-null   object
 16  medical_condition_url           2931 non-null   object
dtypes: float64(2), object(15)
memory usage: 389.4+ KB
```

In [4]:

data.describe(include='all')

```
data.describe(include='all')
```

	drug_name	medical_condition	side_effects	generic_name	drug_classes	brand_names	activity	rx_otc	pregnancy_category	csa	alcohol	related_
count	2931	2931	2807	2888	2849	1718	2931	2930	2702	2931	1377	
unique	2912	47	2759	1392	274	1552	93	3	6	7	1	
top	triamcinolone	Pain	hives ; difficult breathing; swelling of your ...	diphenhydramine	Upper respiratory combinations	Acne-Clear, Benzac AC, BenzePrO, Benziq, Brevo...	0%	Rx	C	N	X	doxyc https://www.drugs.com/doxycy
freq	3	264	10	17	245	10	895	1998	1382	2688	1377	
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

In [5]: # Data Cleaning

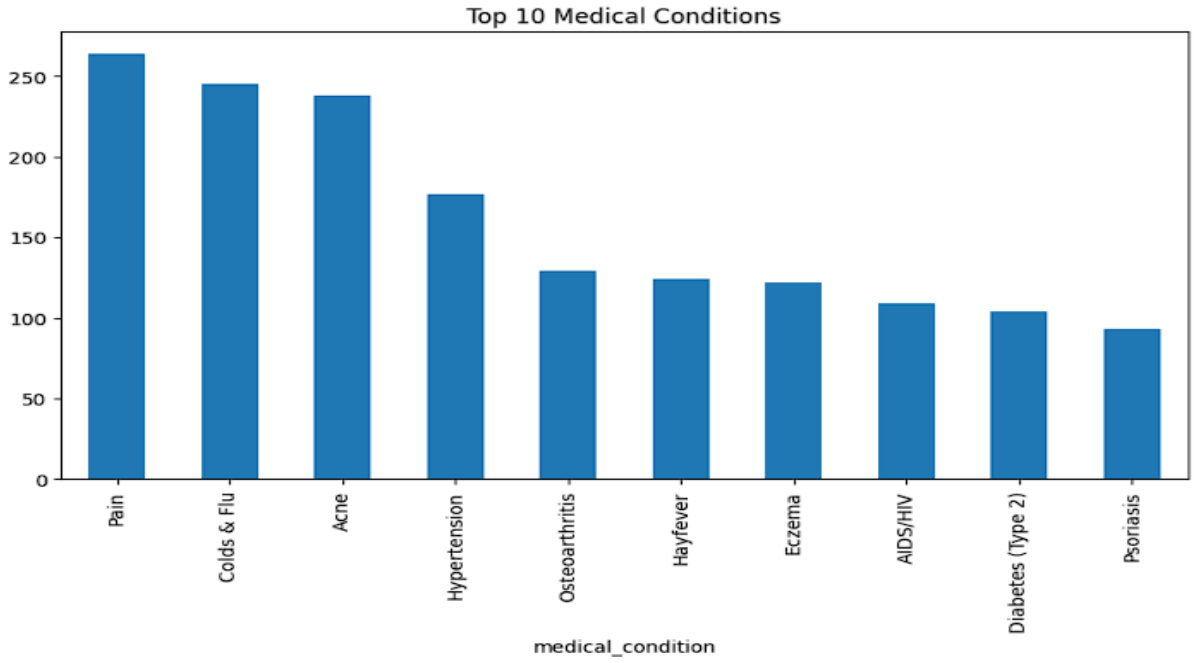
```
data['side_effects'] = data['side_effects'].fillna('Unknown')
data['related_drugs'] = data['related_drugs'].fillna('Unknown')
data['rating'] = data['rating'].fillna(0)
data['no_of_reviews'] = data['no_of_reviews'].fillna(0)

data['alcohol'] = data['alcohol'].fillna(0)
data['alcohol'] = data['alcohol'].replace({'X':1})

data['rx_otc'] = data['rx_otc'].fillna('Unknown')
data['pregnancy_category'] = data['pregnancy_category'].fillna('Unknown')
data['drug_classes'] = data['drug_classes'].fillna('Unknown')
```

In [6]: # Feature Engineering

```
# Risk Score

data['risk_score'] = (
    data['alcohol'] +
    data['pregnancy_category'].isin(['D','X']).astype(int) +
    data['csa'].isin(['1','2']).astype(int)
)
```

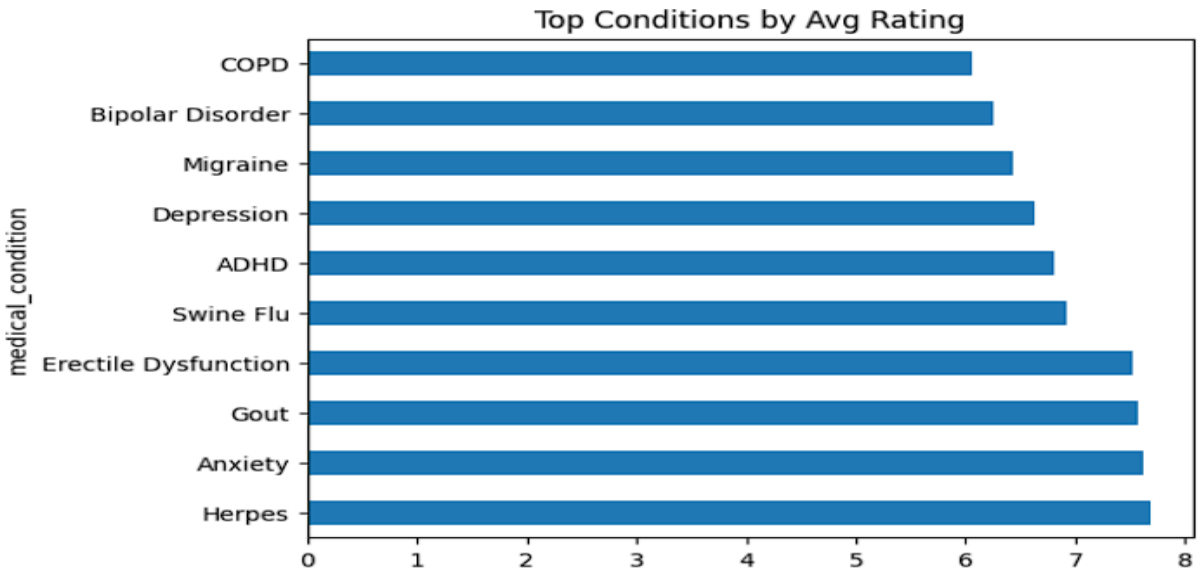
In [7]: # Exploratory Data Analysis

```
# Top Medical Conditions
plt.figure(figsize=(10,5))
data['medical_condition'].value_counts().head(10).plot(kind='bar')
plt.title("Top 10 Medical Conditions")
plt.show()
```



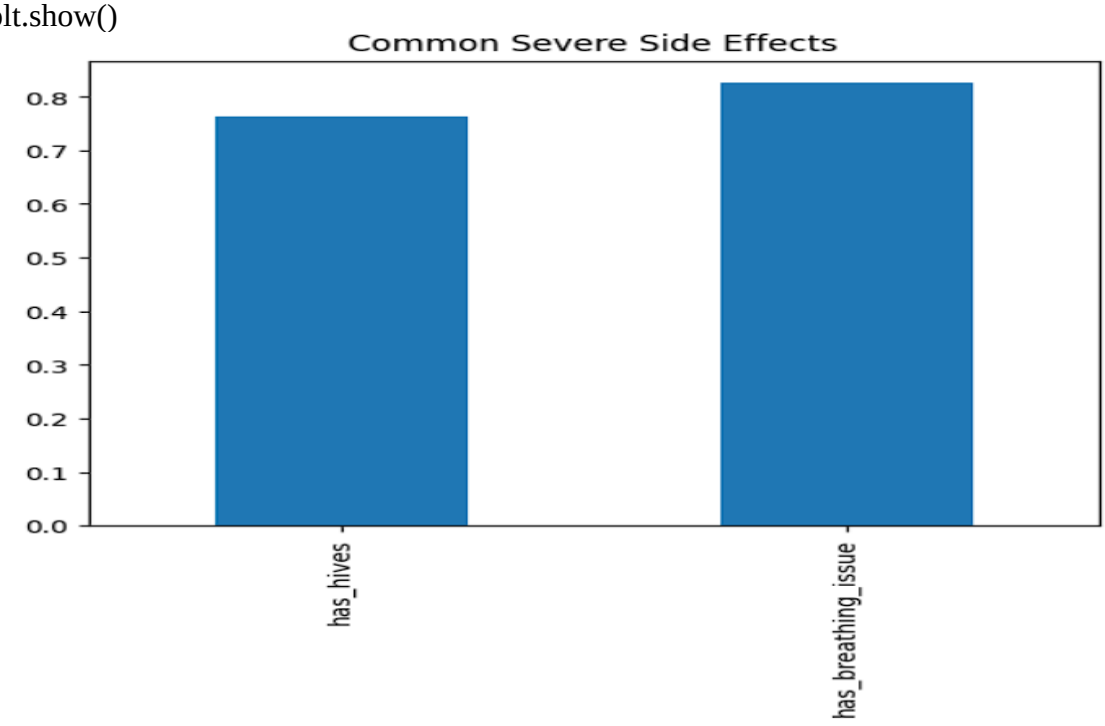
In [8]: # Average Rating by Condition

```
avg_rating =
data.groupby('medical_condition')['rating'].mean().sort_values(ascending=False).head(10)
avg_rating.plot(kind='barh', title="Top Conditions by Avg Rating")
plt.show()
```



In [9]: # Side Effect Severity Analysis

```
data['has_hives'] = data['side_effects'].str.contains('hives', case=False)
data['has_breathing_issue'] = data['side_effects'].str.contains('breathing', case=False)

data[['has_hives','has_breathing_issue']].mean().plot(kind='bar')
plt.title("Common Severe Side Effects")
plt.show()
```



In [10]: # Encoding for ML

```
le = LabelEncoder()

cols = ['generic_name','medical_condition','drug_classes','rx_otc','pregnancy_category','csa']
for col in cols:
    data[col] = le.fit_transform(data[col])
```

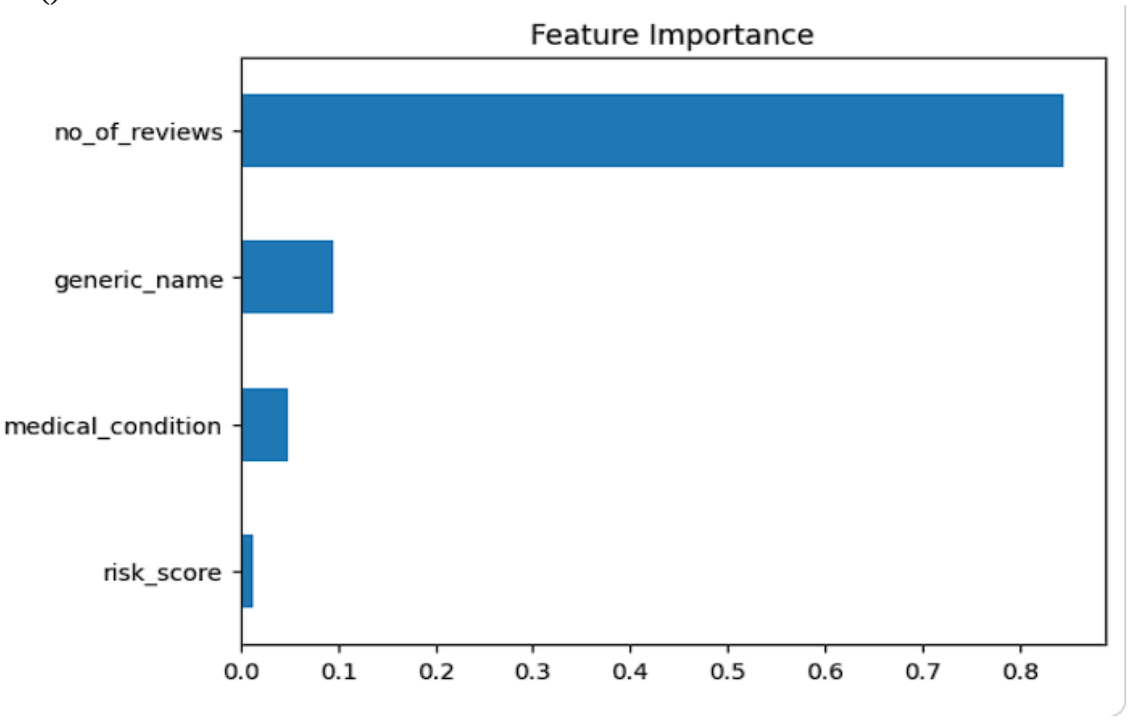
In [11]: # ML Model – Drug Rating Prediction

```
X = data[['generic_name','medical_condition','no_of_reviews','risk_score']]
y = data['rating']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

pred = model.predict(X_test)

print("MAE:", mean_absolute_error(y_test, pred))
print("R2 Score:", r2_score(y_test, pred))
```

OUTPUT: **MAE** : 0.7657304710527028 **R2 Score** : 0.8435452004004039

In [12]: # Feature Importance

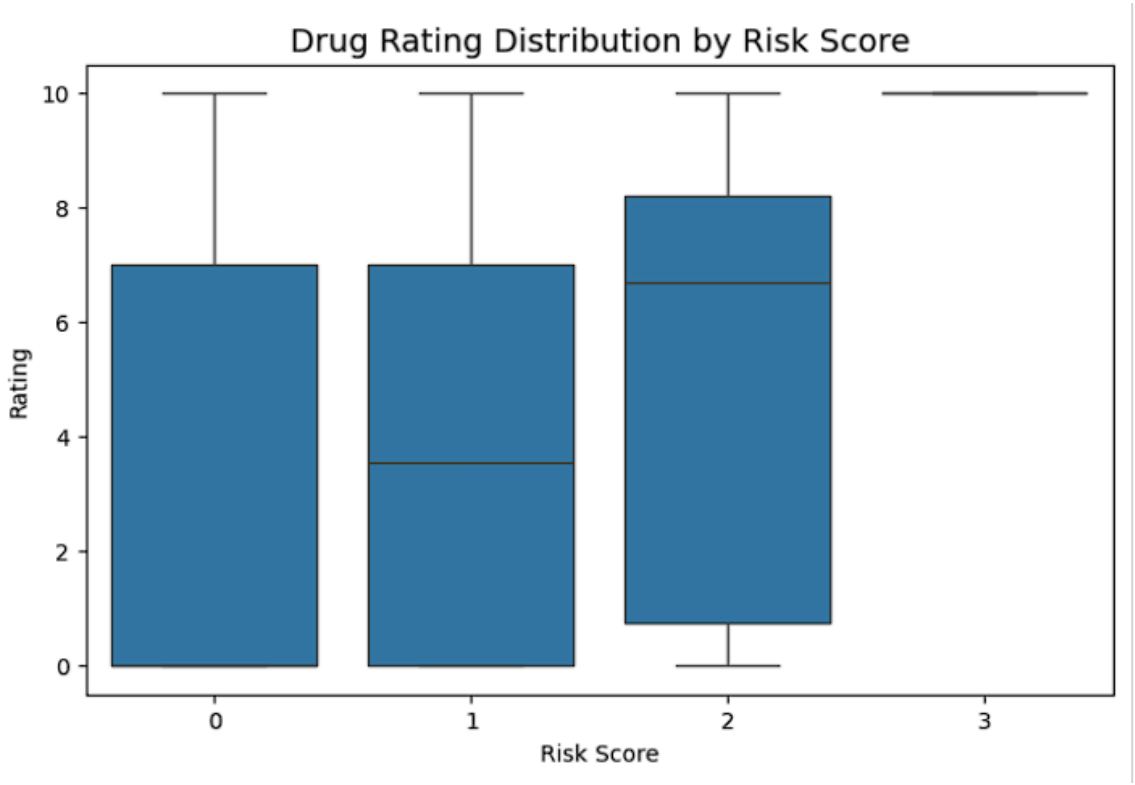
```
importance = pd.Series(model.feature_importances_, index=X.columns)
importance.sort_values().plot(kind='barh', title="Feature Importance")
plt.show()
```



In [13]: # Rating vs Risk Score

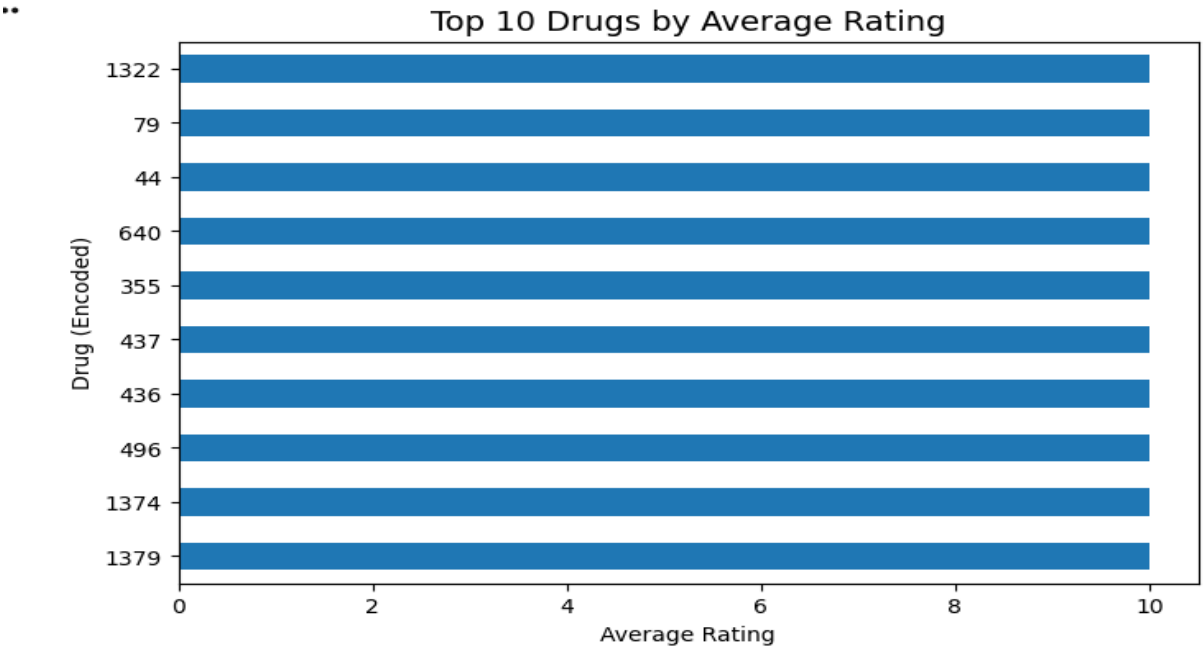
```
# Shows how drug risk impacts user ratings
plt.figure(figsize=(8,5))
sns.boxplot(x='risk_score', y='rating', data=data)
plt.title("Drug Rating Distribution by Risk Score", fontsize=14)
plt.xlabel("Risk Score")
plt.ylabel("Rating")
plt.show()
```

..



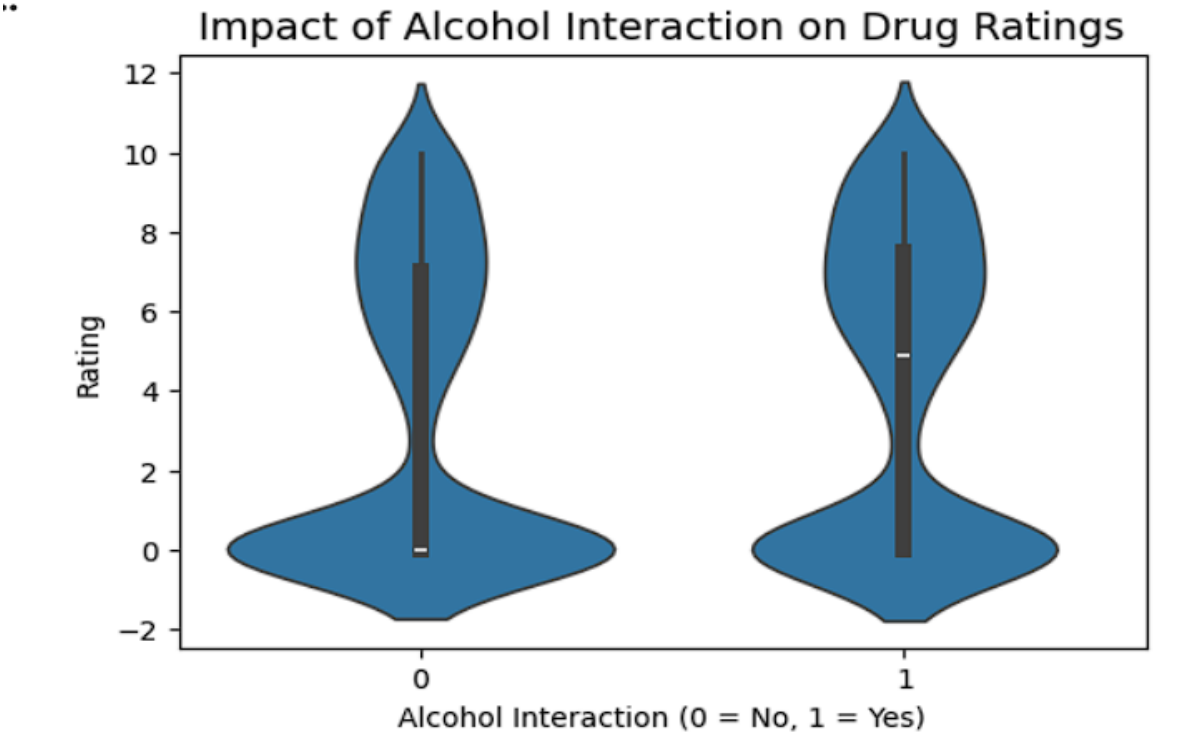
In [14]: # Top 10 Drugs by Average Rating

```
# Identifies most trusted drugs
top_drugs =
data.groupby('generic_name')['rating'].mean().sort_values(ascending=False).head(10)

plt.figure(figsize=(8,5))
top_drugs.plot(kind='barh')
plt.title("Top 10 Drugs by Average Rating", fontsize=14)
plt.xlabel("Average Rating")
plt.ylabel("Drug (Encoded)")
plt.show()
```



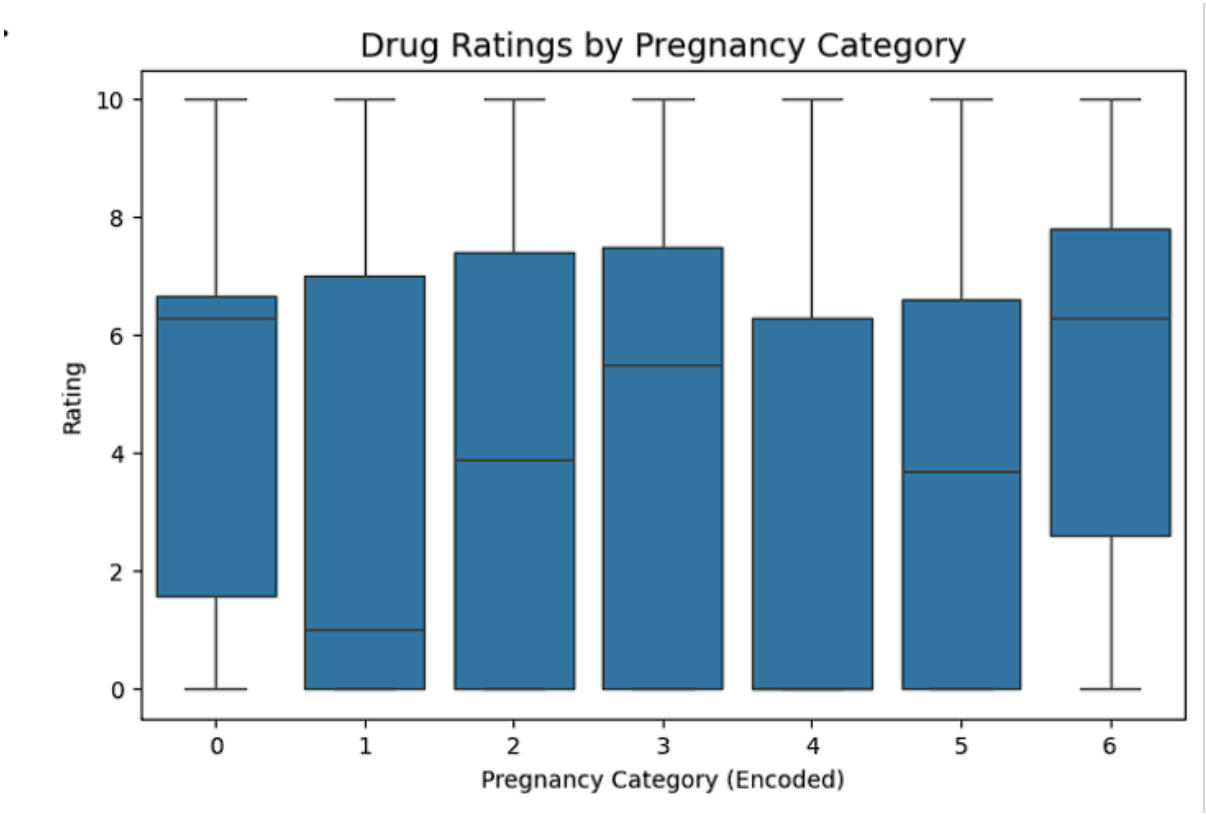
In [15]: # Alcohol Interaction Impact on Ratings

```
# Healthcare safety visualization
plt.figure(figsize=(6,4))
sns.violinplot(x='alcohol', y='rating', data=data)
plt.title("Impact of Alcohol Interaction on Drug Ratings", fontsize=14)
plt.xlabel("Alcohol Interaction (0 = No, 1 = Yes)")
plt.ylabel("Rating")
plt.show()
```



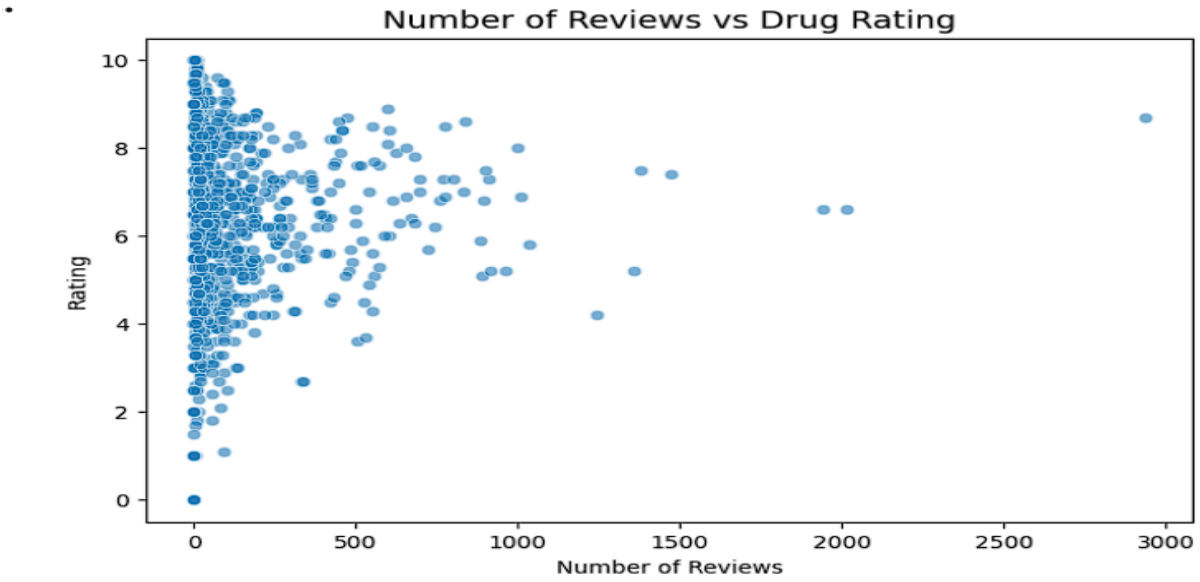
In [16]: # Pregnancy Category vs Rating

```
# Regulatory + healthcare analysis
plt.figure(figsize=(8,5))
sns.boxplot(x='pregnancy_category', y='rating', data=data)
plt.title("Drug Ratings by Pregnancy Category", fontsize=14)
plt.xlabel("Pregnancy Category (Encoded)")
plt.ylabel("Rating")
plt.show()
```



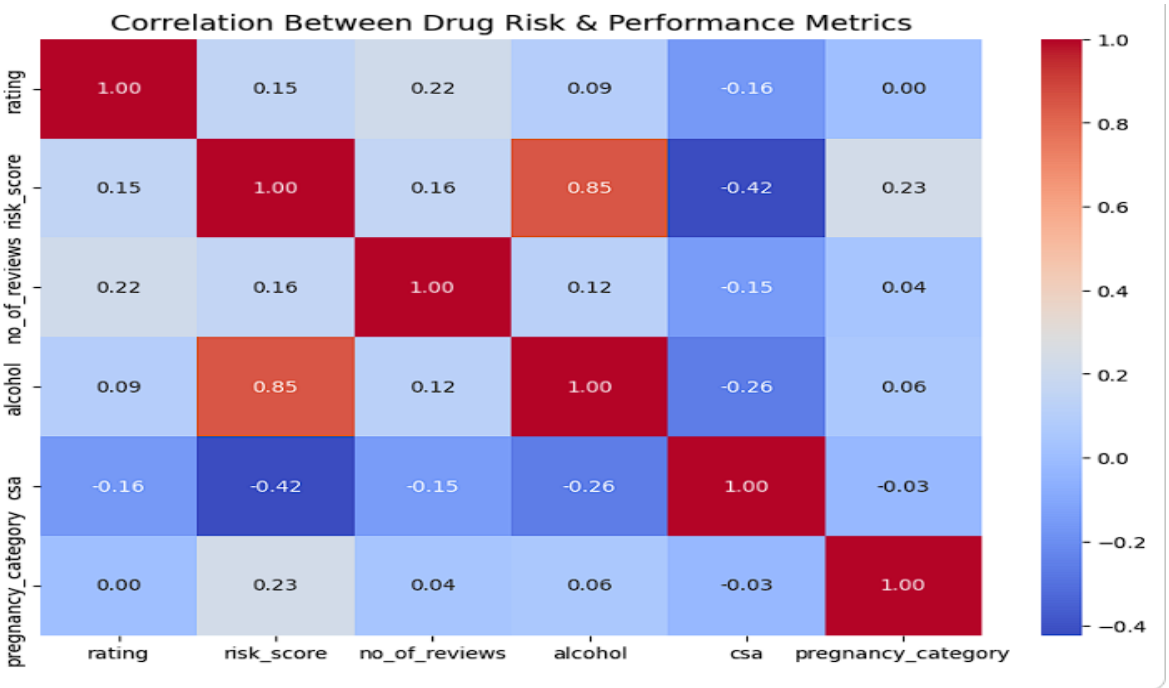
In [17]: # Review Count vs Rating (Popularity vs Quality)

```
# Very professional scatter plot
plt.figure(figsize=(8,5))
sns.scatterplot(x='no_of_reviews', y='rating', data=data, alpha=0.6)
plt.title("Number of Reviews vs Drug Rating", fontsize=14)
plt.xlabel("Number of Reviews")
plt.ylabel("Rating")
plt.show()
```



In [18]: # Correlation Heatmap (Clean Version)

```
# Final summary visualization
plt.figure(figsize=(10,7))
sns.heatmap(
    data[['rating','risk_score','no_of_reviews','alcohol','csa','pregnancy_category']].corr(),
    annot=True,
    cmap='coolwarm',
    fmt=".2f"
)
plt.title("Correlation Between Drug Risk & Performance Metrics", fontsize=14)
plt.show()
```



Conclusion:

This project successfully demonstrates the application of data science and machine learning techniques in analyzing pharmaceutical drug data. By exploring the relationships between drugs, medical conditions, side effects, and safety classifications, meaningful insights were derived regarding drug effectiveness and potential risks. The use of exploratory data analysis and professional visualizations helped in identifying important trends and patterns within the dataset.

Feature engineering played a key role in enhancing the analysis, particularly through the creation of a risk score that combined multiple safety-related factors. The machine learning model developed for predicting drug ratings showed that attributes such as medical condition, number of reviews, and safety risk significantly influence user perception of drugs. Although the model is not intended for clinical decision-making, it effectively illustrates the potential of predictive analytics in healthcare data analysis.

Overall, this project highlights the importance of clean data, thoughtful feature design, and responsible interpretation of results in healthcare analytics. The findings from this study can support further research in drug safety analysis and recommendation systems. The project also provides a strong foundation for future enhancements using advanced models, real-time data, and deeper clinical insights.

LINK:

Drugs, Side Effects and Medical Condition arrow_drop_up