MEDICINES DETAILS ANALYSIS

FINAL CAPSTONE PROJECT

- By Rina Irene Rafalski
- Intern Code: OGTIPIRDS835
- Mentor: Manisha Anand

Oeson learning

About me

Rina Rafalski
Data Analyst

Certificates

- Microsoft Data Science Professional Certificate
- IBM Certificate Python for Data Science

Experience

- Data Scientist Intern
- 5+ years as Marketing Data Analyst

Areas of Interest

- Data Science
- Machine Learning
- Data Analysis
- Business intelligence

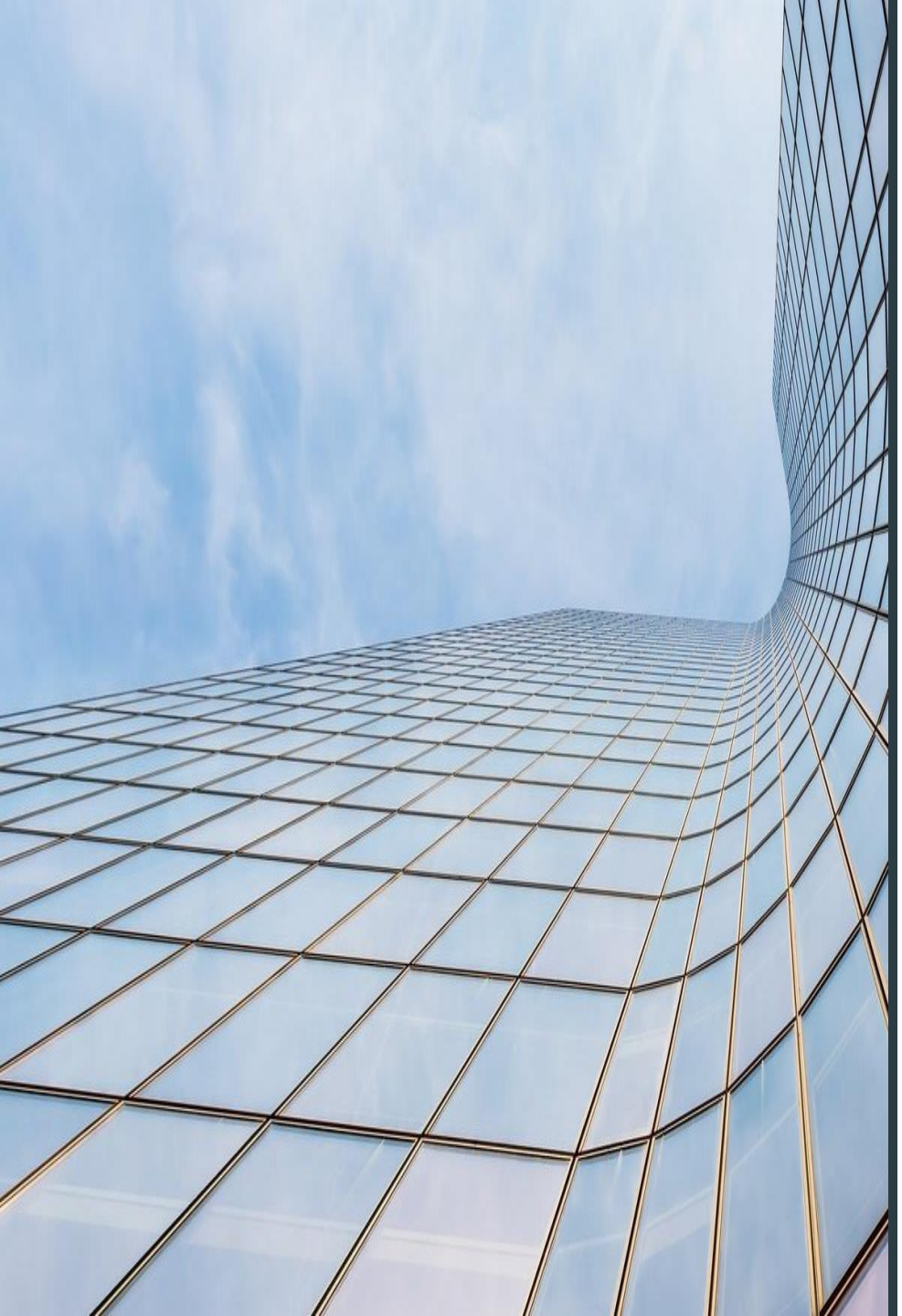
Table of contents

Project Overview

- The pharmaceutical industry faces the challenge of managing vast amounts of data related to the composition, uses, and side effects of medicines. As the number of available medications continues to grow, healthcare providers need reliable insights to prescribe the most effective treatments while minimizing adverse effects. Additionally, understanding the distribution of medicine usage and patient satisfaction can help pharmaceutical companies improve their offerings.
- The objective of the project is to analyze a dataset containing detailed information about over 11,000 medicines, uncover patterns and insights that can help improve decision-making in the healthcare industry and enhance patient outcomes.

Dataset Description

- The dataset Medicine_Details.csv can be downloaded from Kaggle
- The dataset contains a rich repository of information scraped from 1mg, a popular online pharmacy and healthcare platform, covering over 11,000 medicines.
- The dataset includes 9 columns:
  - **Medicine_Name**: Name of the medicine.
  - **Salt_Composition**: The active ingredients in the medicine.
  - **Uses**: Medical conditions or symptoms the medicine is used to treat.
  - **Side_Effects**: Known side effects associated with the medicine.
  - Manufacturer: Company that manufactures the medicine.
  - **Image_URL**: URL to the medicine's image.
  - **Review_Excellent**: Percentage of users who rated the medicine as excellent.
  - **Review_Average**: Percentage of users who rated the medicine as average.
  - **Review_Poor**: Percentage of users who rated the medicine as poor.



1. Data Exploration

Data Exploration

The shape of this dataframe is 93 observations with 14 features.
Data shape: The shape of the dataset is 11825 observations with 9 features.

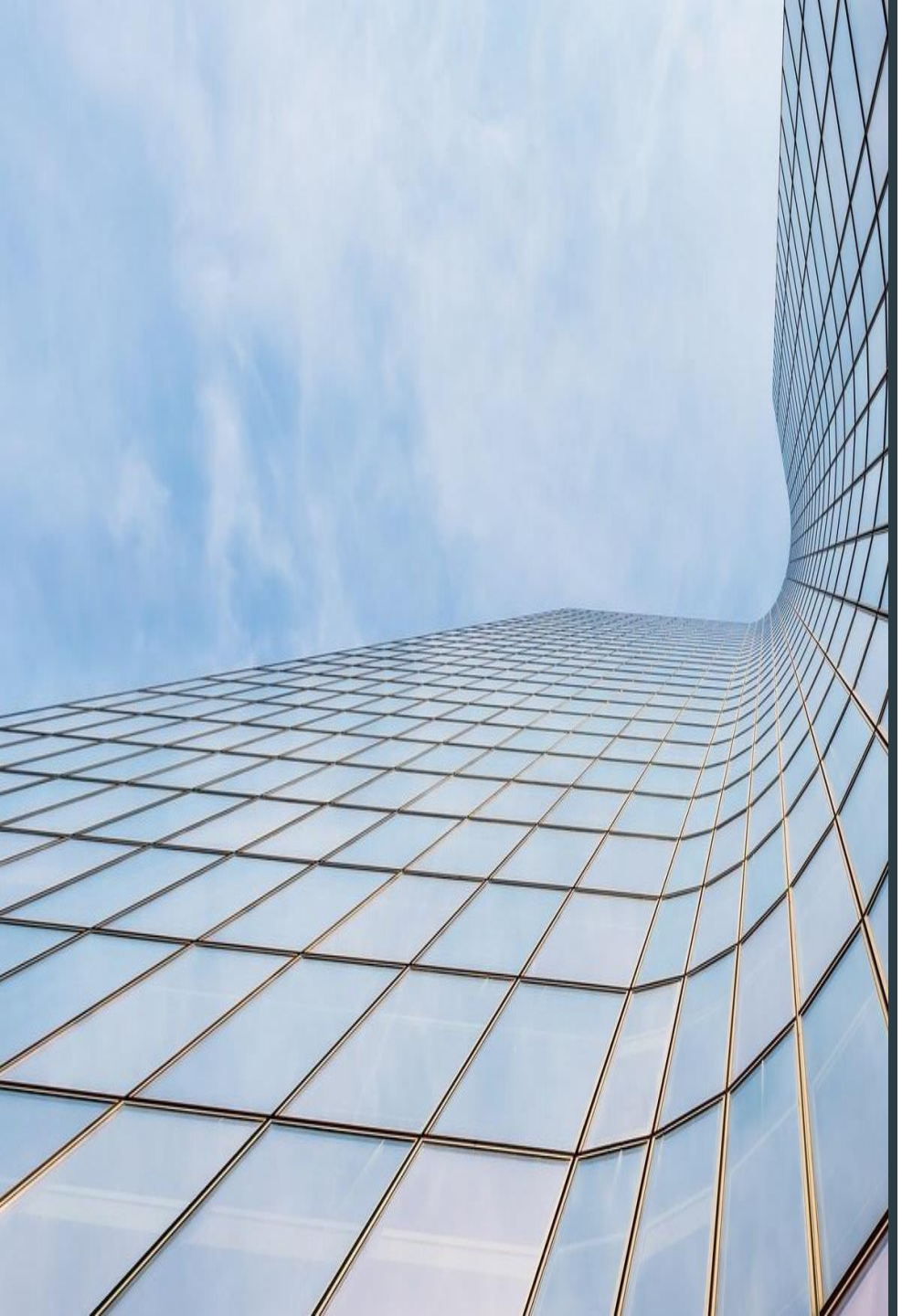
Missing Values: The dataset does not contain any missing values in any of the columns.

Details

- Medicine Name: There are 11,498 unique medicine names out of 11,825 entries. Some medicines have duplicate entries.
- Composition: 3,358 unique compositions are listed. The most common composition is "Luliconazole (1% w/w)," which appears 98 times.
- Uses: There are 712 unique uses, with the most frequent being "Treatment of Type 2 diabetes mellitus" (907 occurrences).
- Side Effects: There are 1,512 unique side effect descriptions, with the most common being "Application site reactions burning irritation..." (390 occurrences).
- Image URL: There are 11,740 unique image URLs. Some images are reused (the most frequent one appears 3 times).

```
Shape of this dataset is (11825, 9).
==================================================
Missing values in any of the columns this dataset are
Medicine Name          0
Composition            0
Uses                   0
Side_effects           0
Image URL              0
Manufacturer           0
Excellent Review %     0
Average Review %       0
Poor Review %          0
dtype: int64
==================================================
```

- Manufacturer: 759 unique manufacturers are listed, with "Sun Pharmaceutical Industries Ltd" being the most frequent (820 occurrences).
- Review Percentages (Excellent, Average, Poor): Reviews range from 0% to 100%. The data seems consistent, but further checks for logical inconsistencies (e.g., sum of review percentages not equal to 100) should be perform



2. Data Preprocessing

Data Preprocessing

Data Cleaning

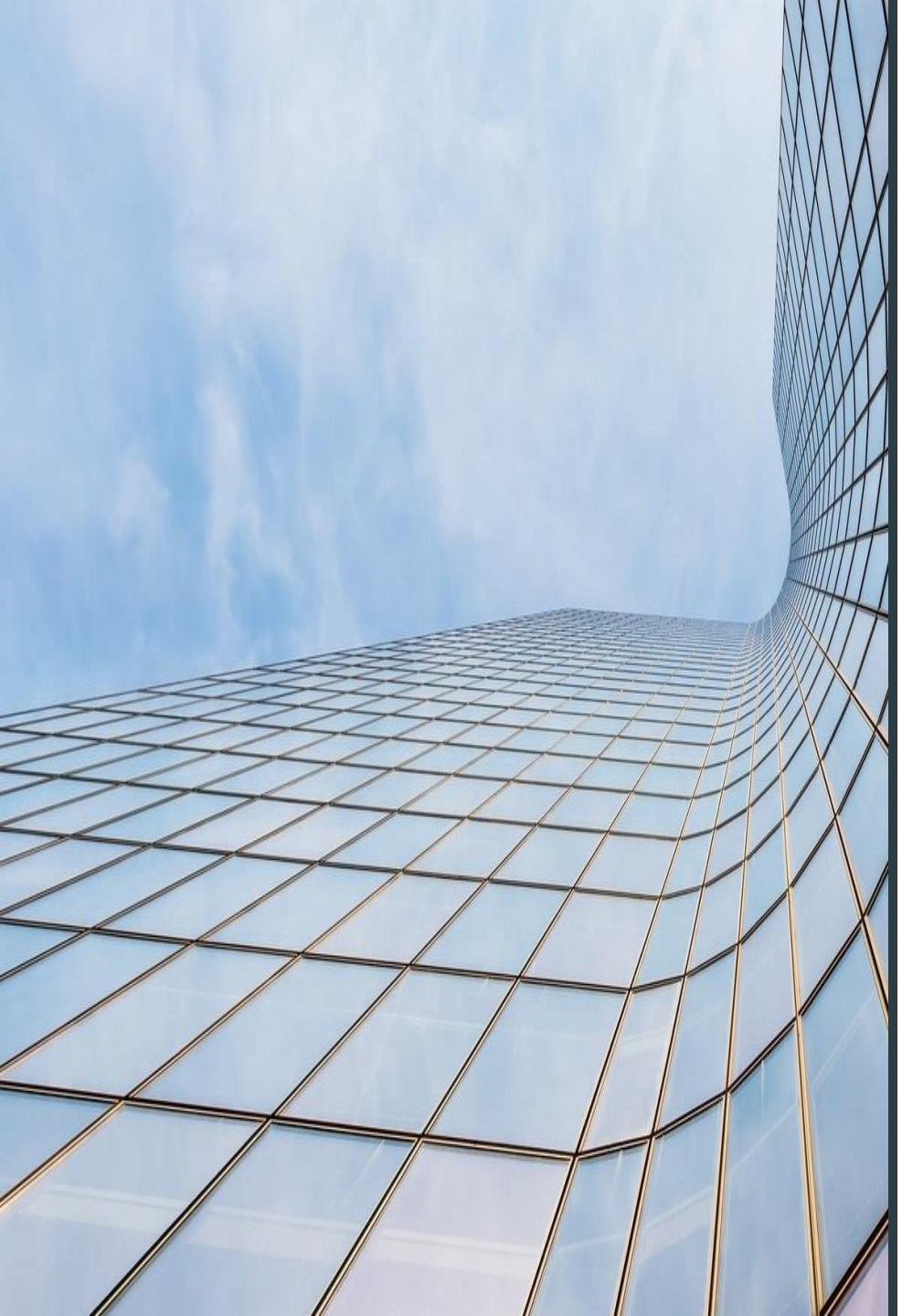
- **Exact Duplicates:** 11825-11741=84 exact duplicates were dropped
- There are no duplicates with different review scores
- **Logical Inconsistencies:** There are no rows where the sum of Excellent/Average/Poor Review % does not equal 100.
  We can trust the aggregate review data for further analysis

Medicine Name	Composition	Uses	Side_effects	Image URL	Manufacturer	Excellent Review %	Average Review %	Poor Review %	Total Review %
Avastin 400mg Injection	Bevacizumab (400mg)	Cancer colon rectum Nonsmall cell lung cancer ...	Rectal bleeding Taste change Headache Noseblee...	https://onemg.gumlet.io/l_watermark_346,w_480,...	Roche Products India Pvt Ltd	22	56	22	100
Augmentin 625 Duo Tablet	Amoxycillin (500mg) + Clavulanic Acid (125mg)	Treatment Bacterial infections	Vomiting Nausea Diarrhea Mucocutaneous candidi...	https://onemg.gumlet.io/l_watermark_346,w_480,...	Glaxo SmithKline Pharmaceuticals Ltd	47	35	18	100
Azithral 500 Tablet	Azithromycin (500mg)	Treatment Bacterial infections	Nausea Abdominal pain Diarrhea	https://onemg.gumlet.io/l_watermark_346,w_480,...	Alembic Pharmaceuticals Ltd	39	40	21	100
Ascoril LS Syrup	Ambroxol (30mg/5ml) + Levosalbutamol (1mg/5ml)...	Treatment Cough mucus	Nausea Vomiting Diarrhea Upset stomach Stomach...	https://onemg.gumlet.io/l_watermark_346,w_480,...	Glenmark Pharmaceuticals Ltd	24	41	35	100
Aciloc 150 Tablet	Ranitidine (150mg)	Treatment Gastroesophageal reflux disease Acid...	Headache Diarrhea Gastrointestinal disturbance	https://onemg.gumlet.io/l_watermark_346,w_480,...	Cadila Pharmaceuticals Ltd	34	37	29	100

Feature Engineering

Creating new features for further analysis

- **User_Satisfaction_Rating** - weighted average score
- **Number_of_Ingredients** - number of active ingredients in Composition
- **Number_of_Side_Effects** - number of reported side effects
- **Number_of_Uses** – number of uses
- **Form_of_Medicine** - Form of medicine from the Medicine Name column
- **Manufacturer_Frequency** - Frequency Encoding of 'Manufacturer'
- **Interaction_Ingredients_SideEffects** - interaction between Number_of_Ingredients and Number_of_Side_Effects
- **Medicine_Form_Popularity** - popularity index of each form of medicine
- **Excellent_Review_Ratio**, **Average_Review_Ratio**, **Poor_Review_Ratio** – ratio of excellent, average, and poor reviews per medicine
- **Medicine_Side_Effect_Counts** - medicine-specific side effect occurrence frequency
- **Interaction_Medicine_Form_Manufacturer** - interaction between Medicine_Form_Popularity and Manufacturer_Frequency



3. Exploratory Data Analysis

Exploratory Data Analysis

Categorical Variables

- **Medicine Name:** There are 11,496 unique medicine names, indicating that the dataset contains a broad range of different medicines. The most frequent medicine name is "Lulifin Cream," appearing 4 times. This suggests that most medicine names are unique or rarely repeated.
- **Form of Medicine:** There are 21 unique forms, such as tablets, injections, syrups, etc., suggesting a diverse range of medicine forms. The most common form is "Tablet," with 5,824 occurrences, indicating that tablets are the predominant form of medicine in the dataset.
- **Manufacturer:** There are 759 unique manufacturers, showing that the dataset represents products from many different companies. "Sun Pharmaceutical Industries Ltd" is the most common manufacturer, appearing 819 times, suggesting that this company has the most products represented in the dataset.

	Medicine Name	Form_of_Medicine	Manufacturer
count	11741	11741	11741
unique	11496	21	759
top	Lulifin Cream	Tablet	Sun Pharmaceutical Industries Ltd
freq	4	5824	819

Numerical Variables

- **User_Satisfaction_Rating:** Mean of 35.47 and a standard deviation of 7.61. The range (16.67 to 50) suggests that user satisfaction varies considerably, indicating a good spread for predictive modeling.
- **Number_of_Ingredients:** Mean of 1.53 and a standard deviation of 0.77, relatively low variance, meaning most medicines have a similar number of ingredients.
- **Number_of_Side_Effects:** Mean of 6.92 and a wider range (1 to 36). Given that side effects likely impact user satisfaction negatively, this feature is probably quite predictive. High variance in this feature suggests it can explain part of the variance in the satisfaction ratings.
- **Number_of_Uses:** Mean of 1.45, standard deviation of 0.99. The low mean indicates that many medicines are used once or infrequently.
- **Manufacturer_Frequency:** Mean of 246.53 and a wide range (1 to 819). A high frequency might suggest popularity or market trust, potentially influencing satisfaction ratings.
- **Interaction_Ingredients_SideEffects:** Mean of 10.52 and a standard deviation of 8.80. A high interaction value suggests that medicines with more ingredients tend to have more side effects, which could negatively impact user satisfaction.

Continued to the next page >>

Exploratory Data Analysis

>> Continued from the previous page >>

- **Medicine_Form_Popularity**: Mean is 0.32, with a maximum of 0.50. This suggests that some forms of medicine are more popular than others, potentially influencing satisfaction.
- **Excellent_Review_Ratio**: On average, about 38.5% of reviews for each medicine are categorized as "Excellent". A minimum value of 0 indicates that some medicines received no excellent reviews, while the maximum value of 1 shows that some medicines received only excellent reviews.
- **Average_Review_Ratio**: On average, around 35.8% of the reviews are categorized as "Average". The maximum value of 0.88 suggests that for some medicines, the majority of reviews are "Average", while others might have no average reviews.
- **Poor_Review_Ratio**: On average, around 25.7% of the reviews are categorized as "Poor". A maximum of 1 indicates that some medicines only received poor reviews, while others received none.
- **Medicine_Side_Effect_Counts**: On average, each medicine is associated with about 7 side effects. A minimum of 1 indicates that some medicines have very few side effects, while others have up to 36 side effects. The high standard deviation of 4.49 suggests that some medicines have significantly more side effects than others.
- **Interaction_Medicine_Form_Manufacturer**: Mean: 83.99, Std (Standard Deviation): 108.14. The high mean and large standard deviation indicate that this interaction is highly variable across medicines, with some combinations of medicine forms and manufacturers being far more common than others.

	User_Satisfaction_Rating	Number_of_Ingredients	Number_of_Side_Effects	Number_of_Uses
count	11741.000000	11741.000000	11741.000000	11741.000000
mean	35.465605	1.529427	6.918576	1.452943
std	7.606346	0.769249	4.307200	0.998860
min	16.666667	1.000000	1.000000	0.000000
25%	31.333333	1.000000	4.000000	1.000000
50%	35.500000	1.000000	6.000000	1.000000
75%	40.333333	2.000000	9.000000	2.000000
max	50.000000	9.000000	36.000000	8.000000

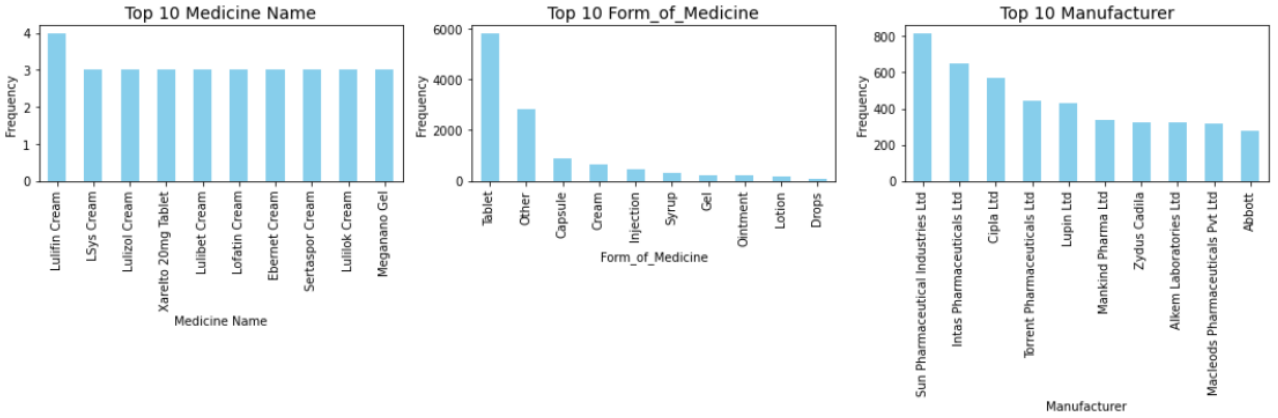
Manufacturer_Frequency	Interaction_Ingredients_SideEffects	Medicine_Form_Popularity	Excellent_Review_Ratio
11741.000000	11741.000000	11741.000000	11741.000000
246.534963	10.516055	0.315832	0.385051
247.717093	8.804890	0.192172	0.251922
1.000000	1.000000	0.000085	0.000000
31.000000	5.000000	0.076143	0.220000
162.000000	8.000000	0.240610	0.340000
336.000000	14.000000	0.496040	0.510000
819.000000	90.000000	0.496040	1.000000

Average_Review_Ratio	Poor_Review_Ratio	Medicine_Side_Effect_Counts	Interaction_Medicine_Form_Manufacturer
11741.000000	11741.000000	11741.000000	11741.000000
0.357835	0.257114	7.124691	83.991382
0.182640	0.239491	4.491668	108.142250
0.000000	0.000000	1.000000	0.001533
0.270000	0.000000	4.000000	4.571587
0.350000	0.220000	6.000000	32.242569
0.470000	0.350000	10.000000	136.906993
0.880000	1.000000	36.000000	406.256367

Exploratory Data Analysis

Categorical Variables

- **Medicine Name**: The most common medicine in the dataset is "Lulifin Cream", appearing 4 times, the low frequency indicates that most medicine names are unique or nearly unique.
- **Form of Medicine**: The most common form of medicine is "Tablet," with a significantly higher frequency than other forms. This suggests that tablets are the most common form in the dataset.
- **Manufacturer**: The most common manufacturer is "Sun Pharmaceutical Industries Ltd" with the highest number of medicines (800 products), followed by "Cipla Ltd" and "Cadila Ltd."with around 600-700 products each. This suggests that a few manufacturers have a more extensive product range in the dataset.



Numerical Variables

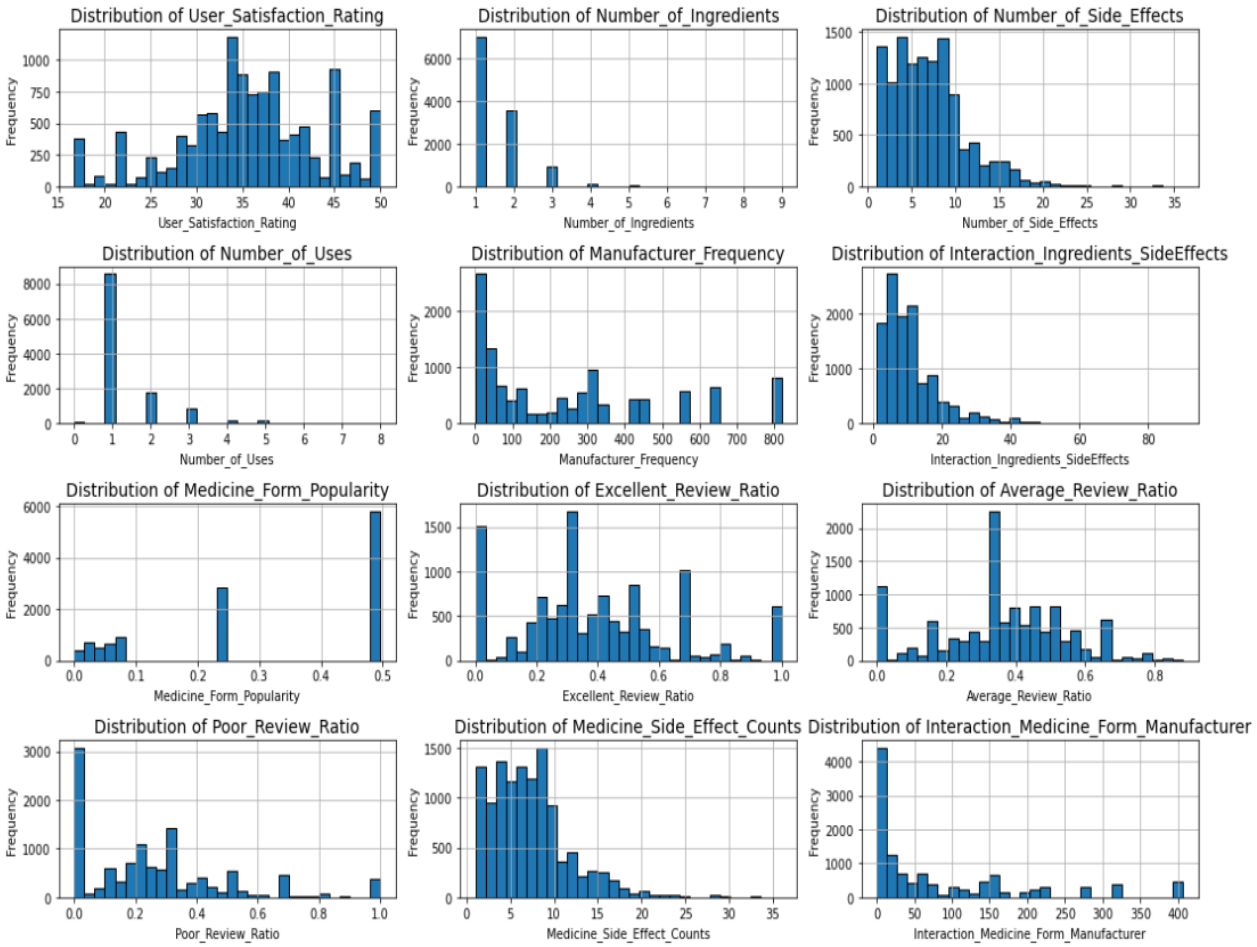
- User_Satisfaction_Rating: normally distributed with a slight skew to the left (lower ratings are less common than higher ratings). The distribution has several peaks, suggesting multiple modes, which could indicate different groups or clusters of user satisfaction levels
- Number_of_Ingredients: The distribution is highly right-skewed, with the majority of medicines containing only 1 ingredient. It could suggest that simpler medicines are more common or preferred.
- Number_of_Side_Effects: This distribution is also right-skewed, with a peak at around 5-10 side effects. This feature is likely a strong predictor, as the presence of many side effects could negatively affect user satisfaction.
- Number_of_Uses: The distribution is also right-skewed, with most medicines having 1-2 uses. The steep drop-off suggests that the majority are specialized for 1 or 2 purposes.
- Manufacturer_Frequency (Top 10 Manufacturer): The most common manufacturer is "Sun Pharmaceutical Industries Ltd" with the highest number of medicines (800 products), followed by "Cipla Ltd" and "Cadila Ltd."with around 600-700 products each. This suggests that a few manufacturers have a more extensive product range in the dataset.

Continued to the next page >>

Exploratory Data Analysis

>> Continued from the previous page >>

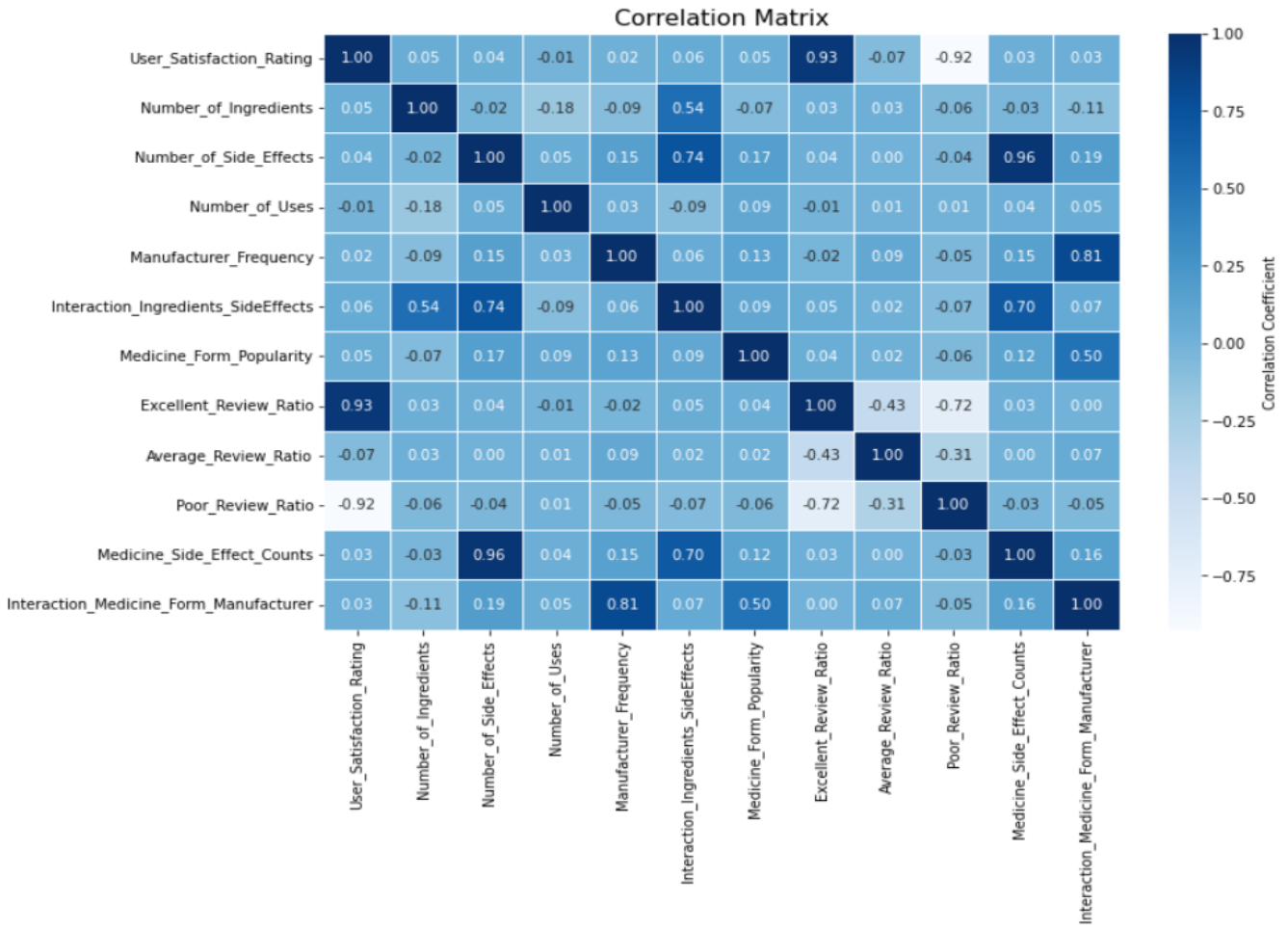
- **Interaction_Ingredients_SideEffects:** This distribution is also right-skewed, with most values being relatively low, indicating that medicines with a higher number of ingredients and side effects are less common.
- **Medicine_Form_Popularity:** The most common form of medicine is "Tablet. This suggests that tablets are the most common form in the dataset.
- **Excellent_Review_Ratio**: Slightly left-skewed, with many medicines having an excellent review ratio between 0.2 and 0.6, it suggests that majority of medicines have a balanced distribution of reviews.
- **Distribution of Average_Review_Ratio**: The distribution is roughly normal, centered around 0.4, meaning that many medicines have an average reviews.
- **Distribution of Poor_Review_Ratio**: This distribution is right-skewed, with a large number of medicines having a low poor review ratio (close to 0), it suggests that many medicines are generally well-received.
- **Distribution of Medicine_Side_Effect_Counts**: The distribution is right-skewed, with the majority of medicines having between 4 and 10 side effects.
- **Distribution of Interaction_Medicine_Form_Manufacturer**: The distribution is heavily right-skewed, with the majority of values concentrated between 0 and 100. This interaction term shows that certain medicine forms and manufacturers dominate the dataset.



Exploratory Data Analysis

Correlation Analysis

- **Review-Based Features** (e.g., Excellent_Review_Ratio and Poor_Review_Ratio) are by far the most important predictors of user satisfaction. This suggests that user feedback (in the form of reviews) captures a significant amount of the variation in satisfaction.
- **Side Effects and Ingredients:** The weak correlations for the number of ingredients and side effects suggest that these features alone may not be sufficient to predict user satisfaction. You may want to explore more detailed features, such as side effect severity or specific ingredient interactions.
- **Weakly Correlated Features:** Features like Number_of_Uses, Manufacturer_Frequency, and Medicine_Form_Popularity exhibit little to no correlation with satisfaction. These features may not contribute much individually to the prediction of satisfaction but could become more important when interacting with other features.### Observations:



Exploratory Data Analysis

Side Effects Analysis

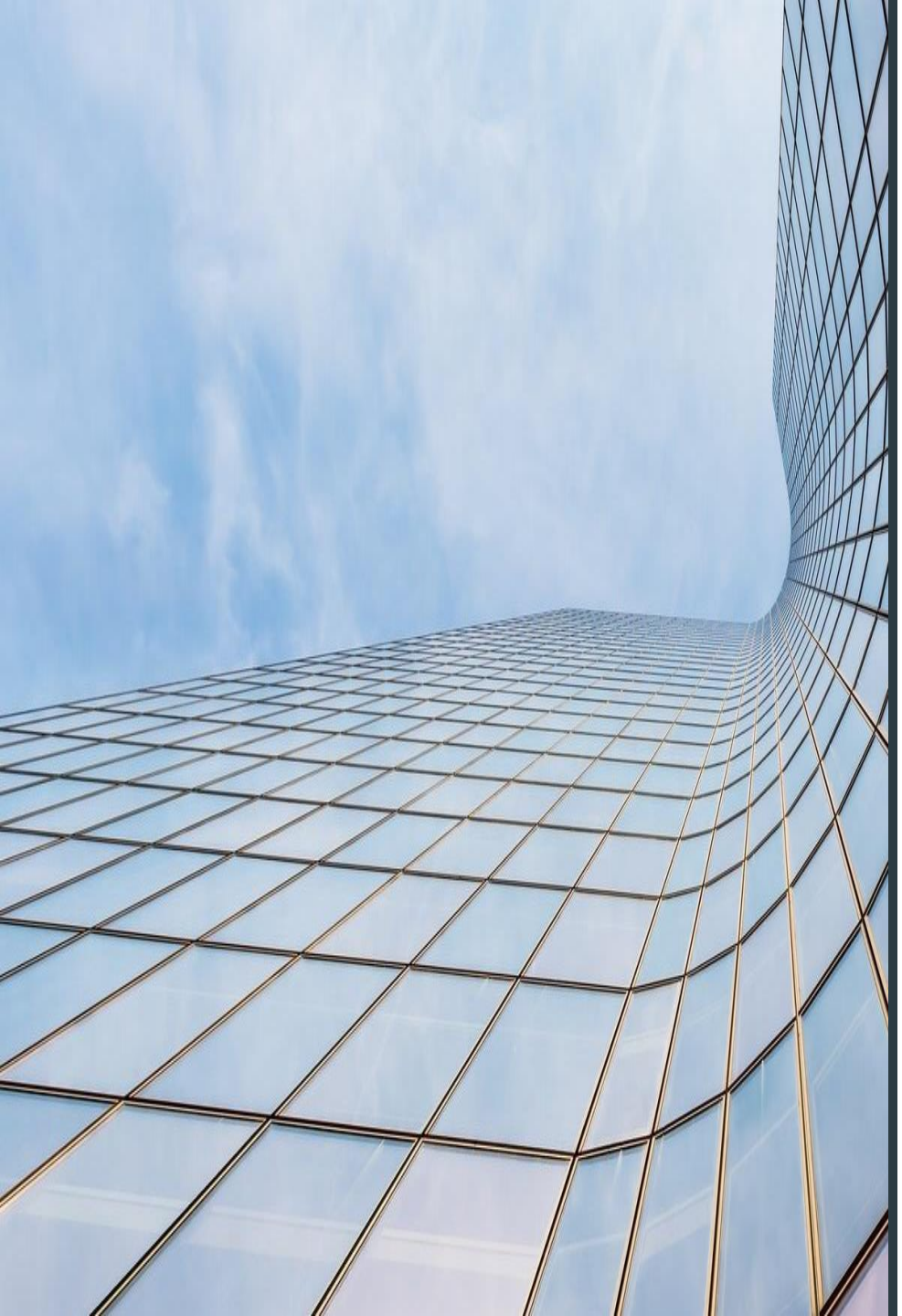
The top 5 most common side effects and their associated statistics:

1. **Nausea**: Appears 6,170 times across 6,110 different medicines, including "Augmentin 625 Duo Tablet," "Azithral 500 Tablet," etc.
2. **Headache**: Appears 5,336 times in 5,286 medicines, such as "Avastin 400mg Injection," "Ascoril LS Syrup," etc.
3. **Diarrhea**: Appears 4,520 times across 4,470 medicines, including "Augmentin 625 Duo Tablet," "Azithral 500 Tablet," etc.
4. **Dizziness**: Appears 4,035 times across 4,003 medicines, including "Ascoril LS Syrup", "Allegra 120mg Tablet" etc.
5. **Vomiting:** Appears 3,473 times across 3,429 medicines, including "Augmentin 625 Duo Tablet", "Ascoril LS Syrup" etc.

Common Side Effects and Associated Medicines

	Side_effects_splitted	Count
0	Nausea	6170
1	Headache	5336
2	Diarrhea	4520
3	Dizziness	4035
4	Vomiting	3473
5	Abdominal pain	1868
6	Sleepiness	1775
7	Constipation	1683
8	Fatigue	1510
9	Stomach pain	1419

	Medicines	Number of Medicines
0	[Augmentin 625 Duo Tablet, Azithral 500 Tablet...	6110
1	[Avastin 400mg Injection, Ascoril LS Syrup, Ac...	5286
2	[Augmentin 625 Duo Tablet, Azithral 500 Tablet...	4470
3	[Ascoril LS Syrup, Allegra 120mg Tablet, Alleg...	4003
4	[Augmentin 625 Duo Tablet, Ascoril LS Syrup, A...	3429
5	[Azithral 500 Tablet, Azee 500 Tablet, Augment...	1832
6	[Arkamin Tablet, Alex Junior Syrup, Aptimust S...	1761
7	[Atarax 25mg Tablet, Altraday Capsule SR, Atar...	1665
8	[Arkamin Tablet, Amitone 10mg Tablet, Ascoril ...	1500
9	[Ascoril LS Syrup, AldigesicSP Tablet, AF Kit ...	1411



4. Model Building and Evaluation

Model Building and Evaluation

Predicting user satisfaction ratings using the following features:

Target variable: User_Satisfaction_Rating

Features:

- Number_of_Ingredients
- Number_of_Side_Effects
- Number_of_Uses
- Manufacturer_Frequency
- Interaction_Ingredients_SideEffects
- Medicine_Form_Popularity
- Excellent_Review_Ratio
- Medicine_Side_Effect_Counts
- Interaction_Medicine_Form_Manufacturer

Model Selection

The dataset contains a mix of numerical features, interaction terms, and non-linear relationships and contains some skewed features. To align with the characteristics of the dataset were chosen 3 following models:

1. **Random Forest**
   - Captures complex, non-linear relationships between features without needing feature transformations
   - Handles of skewed features
   - Provides feature importance insights

2. **Gradient Boosting Regressor**
   - Boosted performance over Random Forests
   - Handles weak correlations
   - Focuses on high accuracy
   - Has flexibility in model tuning

Neural Networks (Keras with TensorFlow)
- Has capability to model complex non-linear relationships
- Suitable for large feature set
- Flexible - useful when the relationships in the dataset are not straightforward or linear.

Model Building and Evaluation

Random Forest Model

Data Preparation for Modeling

```python
# Import Necessary Libraries
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Select features and target
X = df_cleaned[['Number_of_Ingredients', 'Number_of_Side_Effects', 'Number_of_Uses',
                'Manufacturer_Frequency', 'Interaction_Ingredients_SideEffects',
                'Medicine_Form_Popularity', 'Excellent_Review_Ratio',
                'Interaction_Medicine_Form_Manufacturer']]

y = df_cleaned['User_Satisfaction_Rating']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Train the Random Forest Regressor

```python
# Initialize RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)

# Train the model
rf_model.fit(X_train, y_train)
```

```
RandomForestRegressor(random_state=42)
```

Predict and Evaluate Model Performance

```python
# Predict on test data
y_pred = rf_model.predict(X_test)

# Evaluate model performance
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Random Forest Regressor Mean Squared Error (MSE): {mse:.2f}')
print(f'Random Forest Regressor R-squared (R²): {r2:.2f}')
```

```
Random Forest Regressor Mean Squared Error (MSE): 6.97
Random Forest Regressor R-squared (R²): 0.88
```

Observations

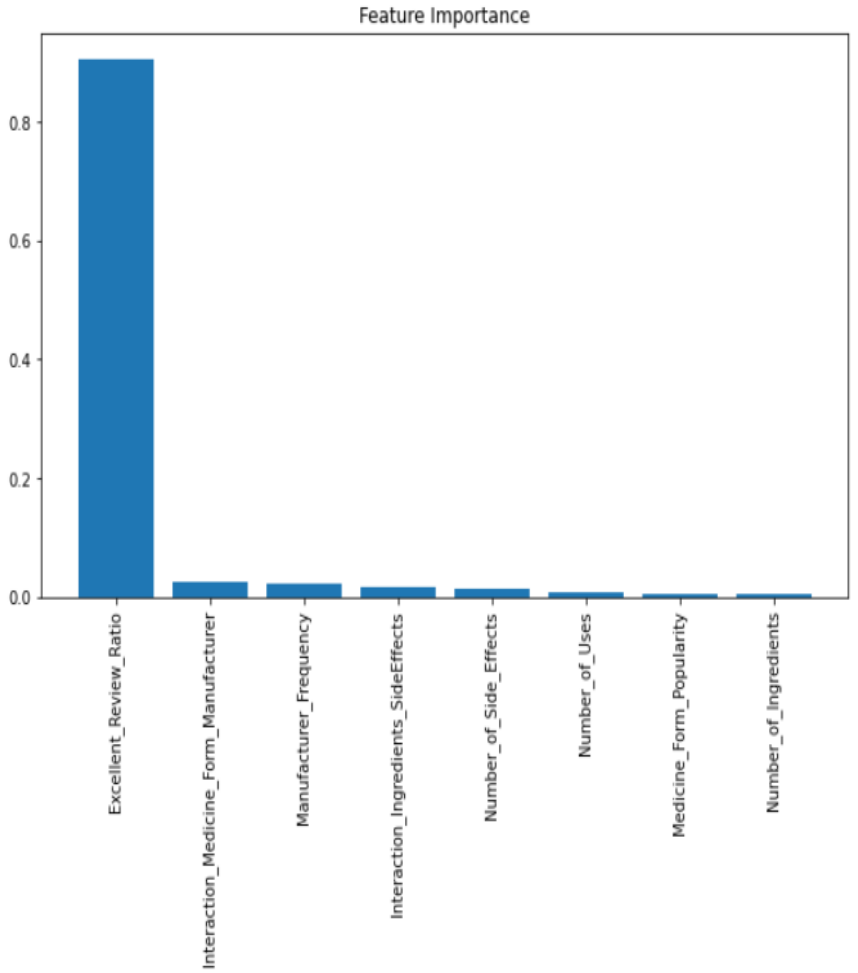
- **Mean Squared Error (MSE): 6.97:** It indicates that, on average, the squared difference between the predicted and actual values of User_Satisfaction_Rating is relatively small. This suggests that the model is making fairly accurate predictions.
- **R-squared (R^2): 0.88:** The model explains 88% of the variance in the target variable (User_Satisfaction_Rating). This is a strong result, indicating that the Random Forest model is doing a good job of capturing the underlying patterns in the data.

Model Building and Evaluation

Feature Importance

- Excellent_Review_Ratio (0.9051): the most important predictor of User_Satisfaction_Rating.
- Interaction_Medicine_Form_Manufacturer (0.0262): comes in second, but with a much smaller importance score, indicates that certain combinations may affect user satisfaction.
- Manufacturer_Frequency (0.0221): has a modest influence on the prediction, suggestes that brand reputation or recognition may have a small impact on satisfaction.
- Interaction_Ingredients_SideEffects (0.0153): has some influence, indicates that how ingredients and side effects interact may affect user satisfaction.
- Number_of_Side_Effects (0.0131): has a small but noticeable impact on satisfaction, which aligns with the intuition that more side effects might reduce satisfaction.
- Other Features: Number_of_Uses (0.0082), Medicine_Form_Popularity (0.0059), and Number_of_Ingredients (0.0039) contribute very little to the model, indicating that these features don't have much predictive power for user satisfaction.

```
Feature ranking:
1. Feature Excellent_Review_Ratio (0.9051)
2. Feature Interaction_Medicine_Form_Manufacturer (0.0262)
3. Feature Manufacturer_Frequency (0.0221)
4. Feature Interaction_Ingredients_SideEffects (0.0153)
5. Feature Number_of_Side_Effects (0.0131)
6. Feature Number_of_Uses (0.0082)
7. Feature Medicine_Form_Popularity (0.0059)
8. Feature Number_of_Ingredients (0.0039)
```



Model Building and Evaluation

Gradient Boosting Regressor

```python
from sklearn.ensemble import GradientBoostingRegressor

# Initialize the Gradient Boosting Regressor
gbr_model = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)

# Train the model
gbr_model.fit(X_train, y_train)

# Predict on the test set
y_pred_gbr = gbr_model.predict(X_test)

# Evaluate model performance
mse_gbr = mean_squared_error(y_test, y_pred_gbr)
r2_gbr = r2_score(y_test, y_pred_gbr)

print(f'Gradient Boosting Regressor Mean Squared Error (MSE): {mse_gbr:.2f}')
print(f'Gradient Boosting Regressor R-squared (R²): {r2_gbr:.2f}')
```

```
Gradient Boosting Regressor Mean Squared Error (MSE): 5.88
Gradient Boosting Regressor R-squared (R²): 0.90
```

Neural Network using Keras with TensorFlow

```python
# Build the Neural Network model
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(100, input_shape=(X_train_scaled.shape[1],), activation='relu')) # Input layer with 100 neur

model.add(tf.keras.layers.Dense(50, activation='relu'))  # Hidden layer with 50 neurons
model.add(tf.keras.layers.Dense(1))  # Output layer

# Compile the model
model.compile(optimizer='adam', loss='mean_squared_error')

# Train the model
history = model.fit(X_train_scaled, y_train, validation_split=0.2, epochs=100, batch_size=32, verbose=0)

# Predict on the test set
y_pred_keras = model.predict(X_test_scaled)

# Evaluate model performance
mse_keras = mean_squared_error(y_test, y_pred_keras)
r2_keras = r2_score(y_test, y_pred_keras)

print(f'Neural Network (Keras) Mean Squared Error (MSE): {mse_keras:.2f}')
print(f'Neural Network (Keras) R-squared (R²): {r2_keras:.2f}')
```

```
74/74 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
Neural Network (Keras) Mean Squared Error (MSE): 6.51
Neural Network (Keras) R-squared (R²): 0.89
```

Model Building and Evaluation

Model Evaluation

Random Forest Regressor:

- Mean Squared Error (MSE): 6.97
- R-squared (R^2): 0.88
- Random Forest performs well with an R^2 of 0.88, indicating that it explains 88% of the variance in user satisfaction.

Gradient Boosting Regressor:

- Mean Squared Error (MSE): 5.88
- R-squared (R^2): 0.90
- Gradient Boosting performs slightly better than Random Forest, with a lower MSE and a higher R^2, explaining 90% of the variance in user satisfaction. This shows that Gradient Boosting can capture slightly more complex relationships in the data.

Neural Network (Keras):

- Mean Squared Error (MSE): 6.44
- R-squared (R^2): 0.89
- The Neural Network performs quite well, with an R^2 of 0.89 and an MSE of 6.51. It's comparable to Random Forest and Gradient Boosting, but does not outperform Gradient Boosting.

Model	Mean Squared Error (MSE)	R-squared (R^2)
Random Forest Regressor	6.97	0.88
Gradient Boosting Regressor	5.88	0.90
Neural Network (Keras)	6.51	0.89

Conclusion

- **Exploratory Data Analysis**
  - Tablets are the most common form of medicine
  - Sun Pharmaceutical Industries Ltd is the top manufacturer
  - User satisfaction ratings are normally distributed with a slight left skew
  - Most medicines have 1-2 ingredients and 1-2 uses
- **Side Effects Analysis**
  - Top 5 most common side effects: Nausea, Headache, Diarrhea, Dizziness, Vomiting
- **Modeling Results**
  - Three models tested: Random Forest, Gradient Boosting, and Neural Network
  - Gradient Boosting performed best with R^2 of 0.90
  - All models explained at least 88% of variance in user satisfaction
- **Feature Importance**
  - Excellent_Review_Ratio was the most important predictor
  - Interaction between medicine form and manufacturer was second most important
  - Number of side effects had a small but noticeable impact on satisfaction
- **Practical Implications:**
  - User reviews are crucial in predicting satisfaction
  - Combination of medicine form and manufacturer influences user satisfaction
  - Number of side effects impacts user satisfaction, but less than expected
- **Future Directions:**
  - Exploring more detailed features like side effect severity or specific ingredient interactions
  - Investigating why some features (e.g., Number_of_Uses) showed weak correlations with satisfaction

THANK YOU

CONTACT

Rina Irene Rafalski
@ rinaraf@gmail.com