A DATA-CENTRIC APPROACH TO UNDERSTANDING MEDICATION AND HEALTH INTERACTIONS

INTRODUCTION:

Healthcare is an important field that affects people's well-being and quality of life. As more health records, prescriptions, and patient information are stored digitally, there is a great chance to use this data to make better decisions. Data analytics in healthcare helps doctors, hospitals, and other providers find patterns, make predictions, and improve the way care is given. It can also help lower costs and make treatments more personal and effective. By looking at data related to diseases, medicines, and patient behaviour, we can find useful insights to support better planning and patient care. This project focuses on studying healthcare data to find valuable information that can help improve services and health outcomes.

EXPLORATORY ANALYSIS:

Exploratory Data Analysis (EDA) is the first and most important step in understanding the data. In this healthcare project, EDA helps us explore patient information, drug usage, side effects, and medical conditions to discover hidden patterns and relationships. It involves checking the structure of the data, finding missing values, and using graphs and statistics to better understand the overall trends. Through charts, summaries, and visual tools, we can spot outliers, common health issues, or frequently used medications. EDA helps us prepare the data for further analysis and ensures that we are working with accurate and meaningful information. This step provides a strong foundation for making reliable and data-driven healthcare decisions.

RECORDED INFORMATION:

Recorded information refers to the data that has been collected and stored from various sources over time. In the context of healthcare, this includes details such as drug names, patient conditions, side effects, user ratings, and medical classifications. This information is usually gathered through patient reports, clinical trials, prescriptions, or user reviews. It plays a key role in understanding how drugs perform, what reactions they cause, and how they relate to different medical conditions. By analysing recorded information, we can identify patterns, make informed decisions, and improve the quality of healthcare services. The data source is provided in the form of a CSV (Comma-Separated Values) file, and a brief summary of the column headings is outlined below:

1) **drug_name** – The name of the drug being analysed or reviewed.
2) **medical condition** – The health issue or disease for which the drug is prescribed.
3) **side-effects** – Known or reported side effects associated with the drug.
4) **generic_name** – The non-branded, standard name of the drug’s active ingredient.
5) **drug classes** – The category or classification of the drug based on its function or chemical structure.
6) **brand names** – Commercial brand names under which the drug is sold.
7) **activity** – Could refer to the drug’s pharmacological action or status (e.g., active/inactive).
8) **rx_otc** – Indicates whether the drug is prescription-only (Rx), over-the-counter (OTC), or both.
    a. OTC (Over-the-counter) = Medication that can be purchased without a medical prescription
    b. Rx = Prescription Needed
    c. Rx/OTC = Prescription or Over-the-counter.

9) **pregnancy_category** – The safety classification of the drug for use during pregnancy.

   a. **A**: Studies in pregnant women show the drug is safe during the first trimester, and there's no evidence of risk later in pregnancy either.
   b. **B**: Animal studies show no harm to the baby, but there are no well-controlled studies in pregnant women.
   c. **C**: Animal studies show potential harm to the baby, and there are no good studies in humans—but the drug might still be used if the benefits are greater than the risks.
   d. **D**: There is clear evidence the drug can harm the baby in humans—but it might still be used if the benefits are serious enough to outweigh the risks.
   e. **X**: The drug has been shown to harm the baby in animals or humans, and the risks clearly outweigh any possible benefits—**do not use during pregnancy**.
   f. **N**: The FDA has not yet classified the drug for use in pregnancy.

10) **csa** – Refers to the Controlled Substances Act classification, indicating if the drug is regulated or has abuse potential.

   a. **M**: Drug has multiple schedules based on its form or strength.
   b. **U**: Drug's schedule under the Controlled Substances Act is unknown.
   c. **N**: Drug is **not** controlled under the Controlled Substances Act.
   d. **1**: High abuse potential, **no accepted medical use**, not safe even under medical supervision.
   e. **2**: High abuse potential, **accepted medical use with severe restrictions**, may cause severe dependence.
   f. **3**: Moderate abuse potential, accepted medical use, may cause moderate physical or high psychological dependence.
   g. **4**: Low abuse potential, accepted medical use, may cause limited dependence.
   h. **5**: Lowest abuse potential, accepted medical use, may cause minimal dependence.

11) **alcohol** – Notes any known interactions or warnings related to alcohol use with the drug. (X = Interacts with Alcohol)
12) **related_drugs** – Lists other drugs used to treat the same condition or similar in nature.
13) **medical_condition_description** – A short explanation of the medical condition being treated.
14) **rating** – Average user rating or effectiveness score of the drug (typically from reviews).
15) **no_of_reviews** – The number of user reviews or feedback entries available for the drug.
16) **drug_link** – URL or web link to more information about the drug (possibly from a drug database).
17) **medical_condition_url** – URL or web link to more information about the medical condition.

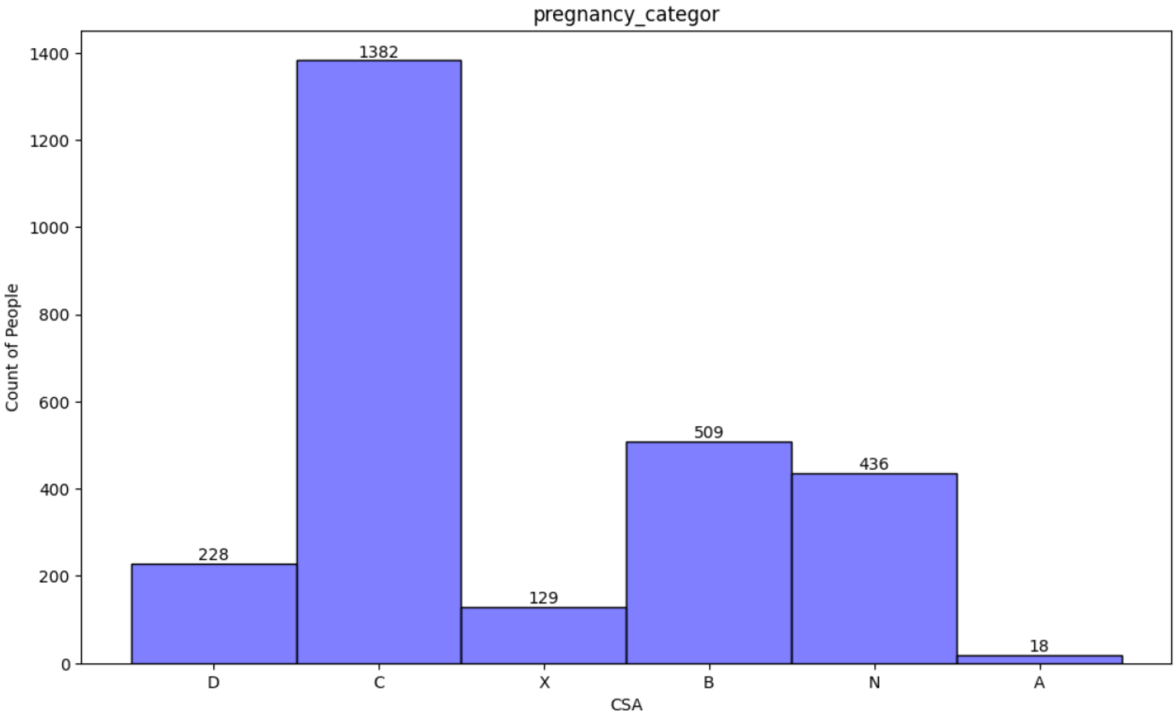
Diagnostic Conclusion:

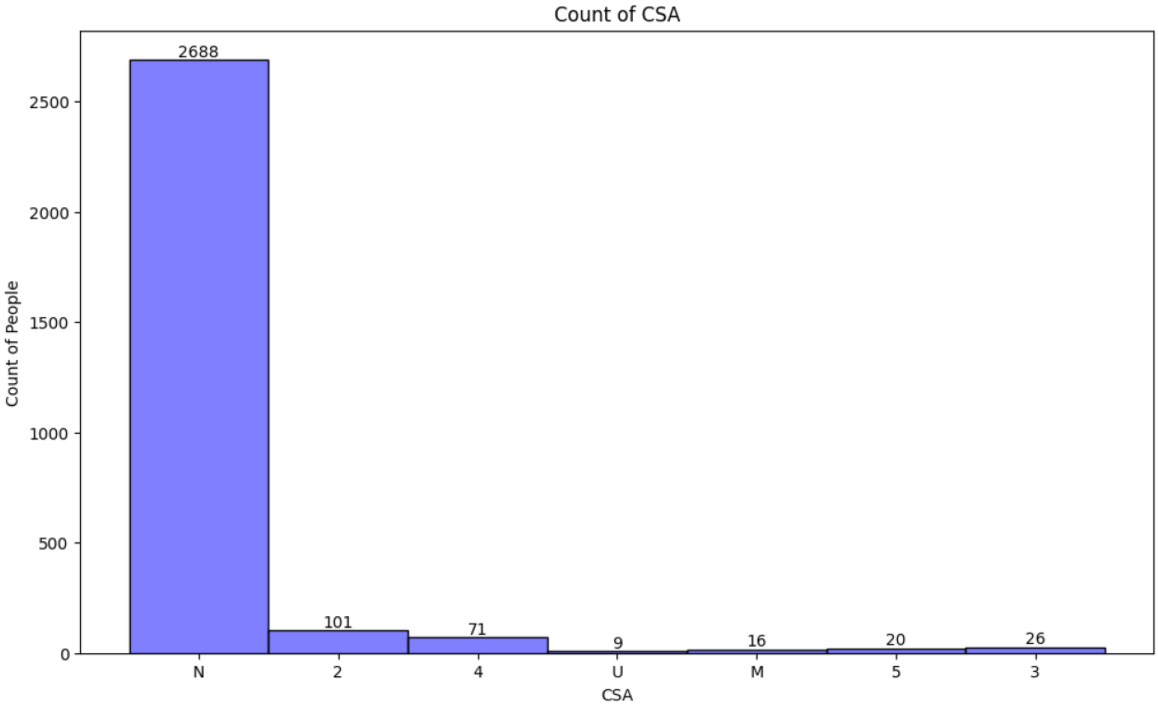
Data Cleaning:

Data cleaning is the process of minimizing noise or inconsistencies in a dataset. Noise includes unwanted formats, incorrect data types, or missing values that complicate analysis. Missing values are either removed if irrelevant or replaced (e.g., marked as "unknown") if they contribute to the analysis. For example, apart from unique IDs, most duplicate or null values may still hold analytical value. Data cleaning may also involve correcting values—such as adjusting an "activity" value from 0.87 to 87%. Additionally, unique values in certain columns, like **CSA schedule**, **pregnancy category**, and **alcohol interaction**, are identified and reviewed.

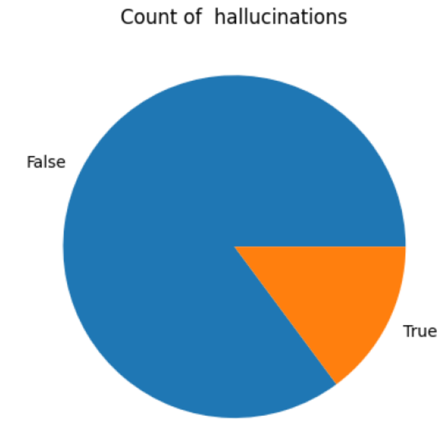
Data Analysis and Visualization:

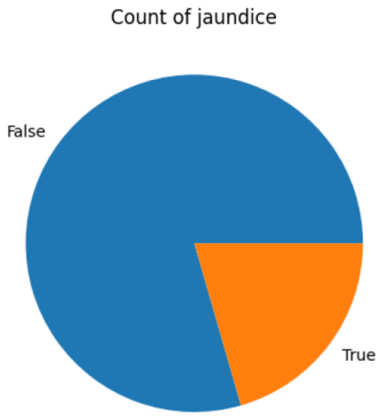
- The dataset contains **2931 rows** and **17 columns**.
- Categorical columns such as **pregnancy category**, **CSA (Controlled Substances Act) schedule**, and **alcohol interaction** were analysed.

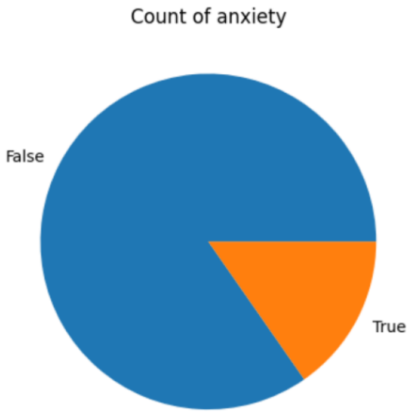


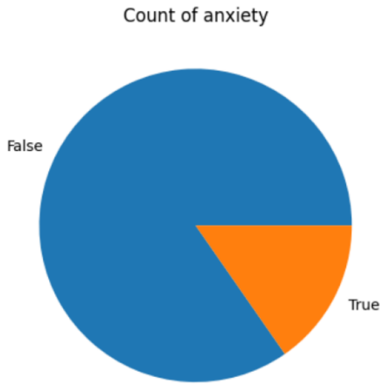


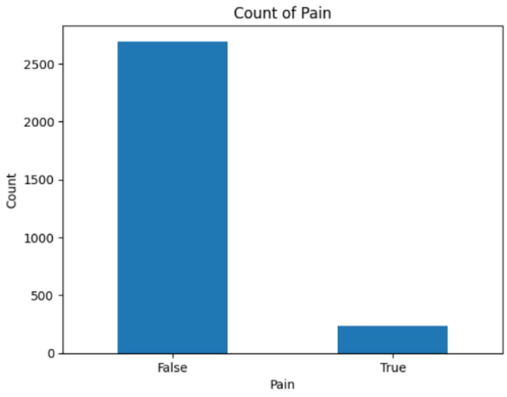
- The number of patients affected by various medical conditions was counted based on these categories, and the data was visualized using **Matplotlib** in **Python**.

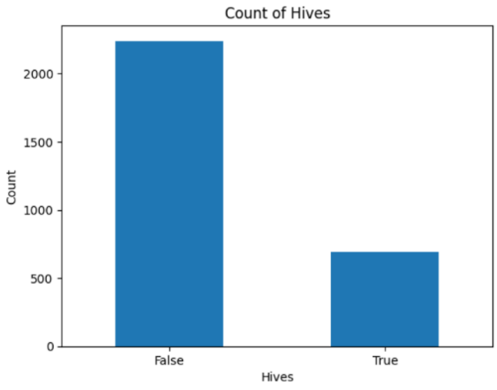




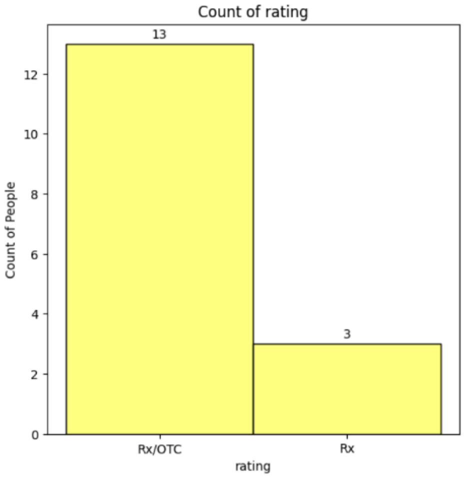


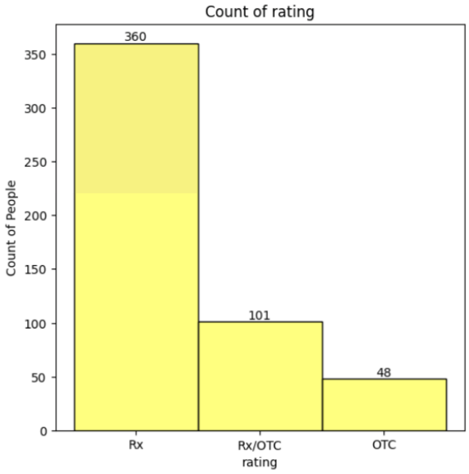




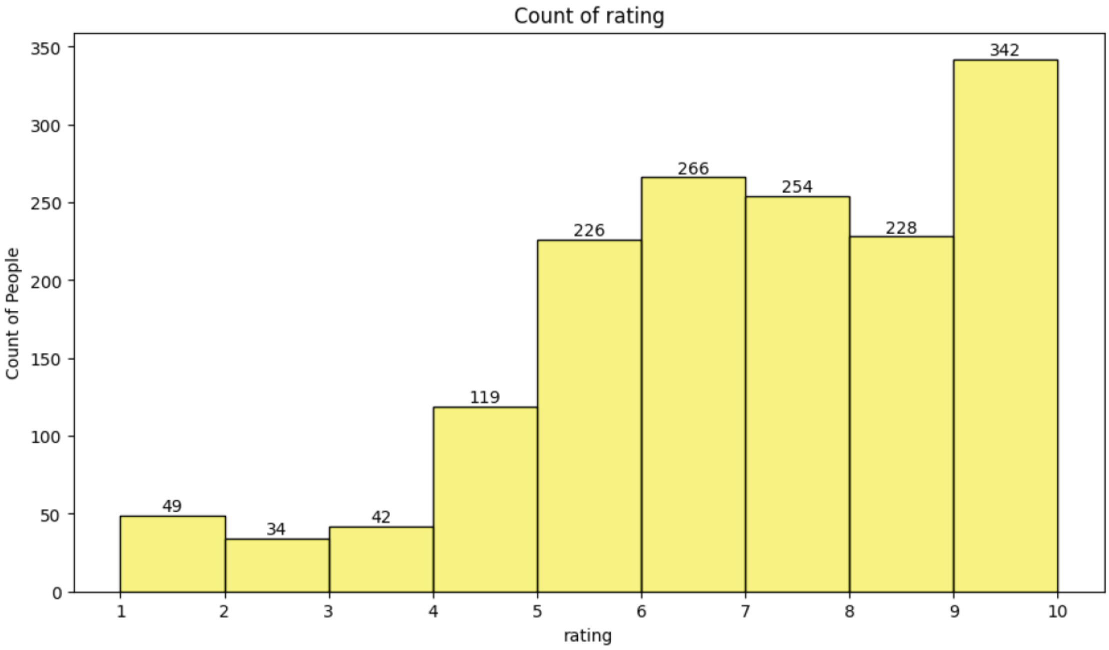


- The availability of medicines was examined under different classifications such as **RX (prescription-only)**, **OTC (over-the-counter)**, and **RX/OTC (both)**.





- Ratings of different range are analysed and their counts are validated.



End Statement:

This Python-based visualization provided a basic understanding of **side effects**. Using functions like **def (define function)** and **input ()**, data filtering was made easier, allowing for more effective visualization.