رواد مصر الرقمية

وزارة الاتصـــــــالات
وتكنولوجيا المعلومات
Ministry of Communications
and Information Technology

Phase 1

Data Preparation & Exploratory Data Analysis Report

DS healthcare Cairo
CAI3_AIS4_S2

Under the supervision of Dr. Mahmoud Talaat

Dataset

Final_Augmented_dataset_Diseases_and_Symptoms

Objective:

To explore, clean, and prepare a large-scale medical symptom dataset for multi-class classification by disease category.

Executive Summary

We successfully loaded and prepared a dataset of 246,945 patient records, each associated with a disease and 377 binary symptom indicators. Key steps included:

- **Data Cleaning**

  Removal of 57,298 duplicate entries.

- **Data Filtering**

  Retained only diseases with 800 or more occurrences, reducing the number of unique diseases from 773 to 105.

- **Feature Engineering**

  Categorized all 105 diseases into 10 medical specialties (e.g., Respiratory, Neurological, etc.).

- **EDA & Insights**

  Generated visualizations to understand the distribution of diseases, categories, and most common symptoms.

- **Data Splitting**

  Split the master dataset into 10 category-specific subsets for potential focused modeling.

Initial Dataset Overview

246,945 entries (rows), 378 columns

Columns

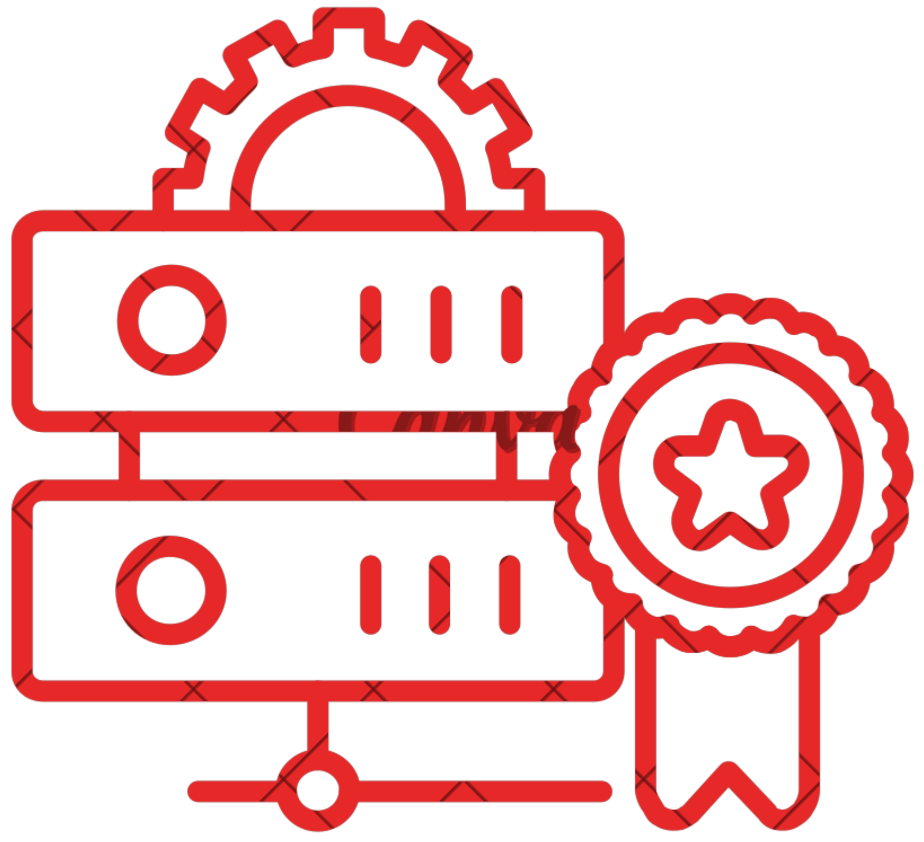
- 1 column for the disease name (diseases) and 377 binary columns (1/0) indicating the presence or absence of a specific symptom.

Data Types

- 377 int64 (symptoms), 1 object (disease name).

Data Quality Check

1. **Missing Values:** Zero missing values found across all 378 columns. This is a significant strength of the dataset.
2. **Duplicates**: 57,298 duplicate rows were identified and removed, resulting in a final dataset of 90,789 unique records.



Data Filtering - Disease Frequency

Initial Diversity

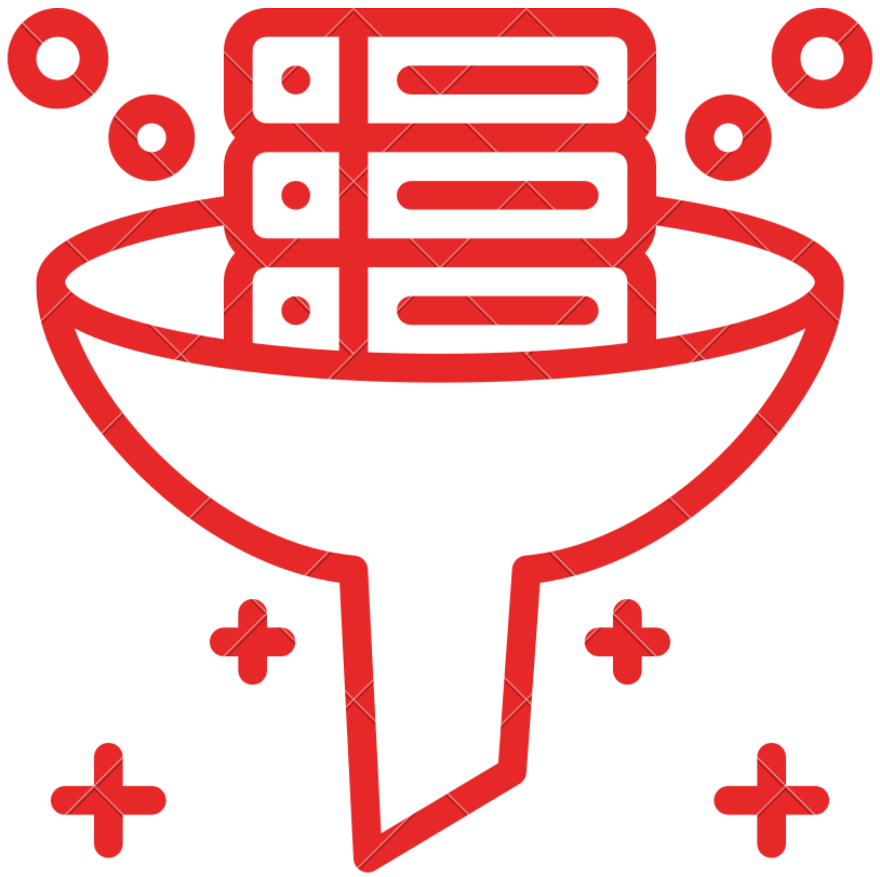
- The dataset contained 773 unique diseases.

Issue

- Many diseases had very low representation, which could lead to poor model performance for those classes.

Solution

A frequency filter was applied. Only diseases with 800 or more recorded cases were kept.



Result

- The number of unique diseases was reduced from 773 to 105. This ensures a more robust and balanced dataset for training classification models. Examples of diseases near the cutoff are shown (e.g., Seborrheic Dermatitis: 800, Acute Stress Reaction: 799).

Feature Engineering

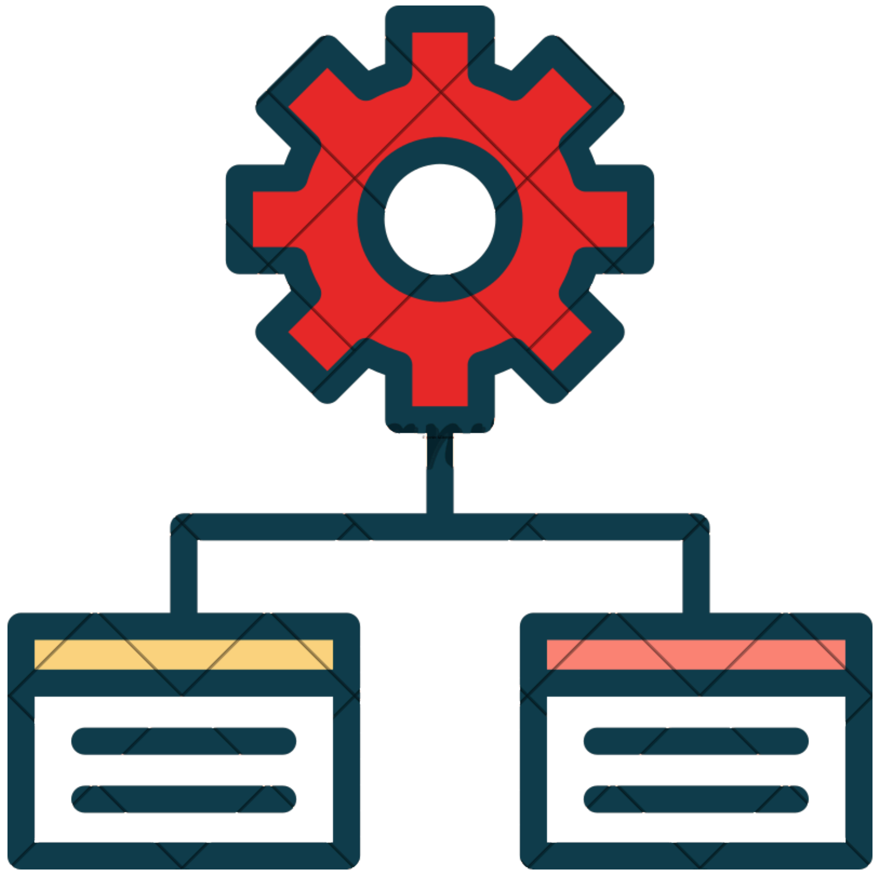
Objective

- To add a higher-level, clinically meaningful label for more interpretable analysis and modeling.

Method

- Each of the 105 diseases was manually mapped to one of **10 medical categories**:

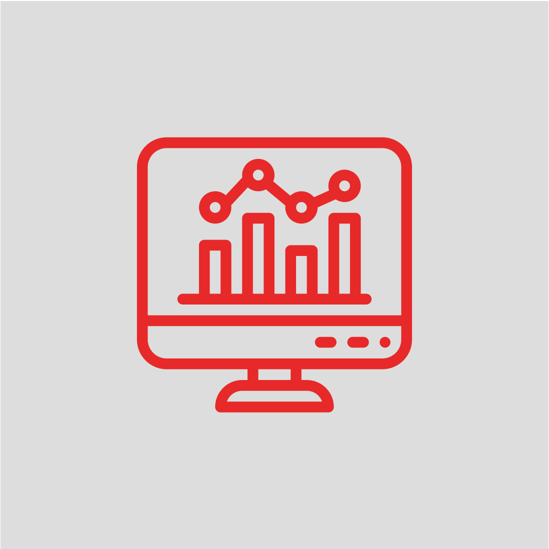
1. Respiratory
2. Gastrointestinal and Liver
3. Urinary and Reproductive
4. Cardiovascular
5. Neurological and Psychiatric
6. Skin
7. Musculoskeletal
8. Eye, Ear and Nose
9. Infectious Diseases
10. Others (for injuries, allergies, etc.)



Data Preparation for Modeling

Label Encoding

The disease names were encoded into numerical values using a LabelEncoder from scikit-learn. The encoder was saved (label_encoder.pkl) and a mapping dictionary was exported to JSON (disease_mapping.json) for future use.



Data Splitting

The master DataFrame was split into 10 separate DataFrames based on the disease category (e.g., Neuro, Gast, Skin).

Final Cleaning

For each category-specific DataFrame, any symptom column that had only a single unique value (i.e., was constant and provided no information) was dropped. This optimizes the feature set for each category's model.

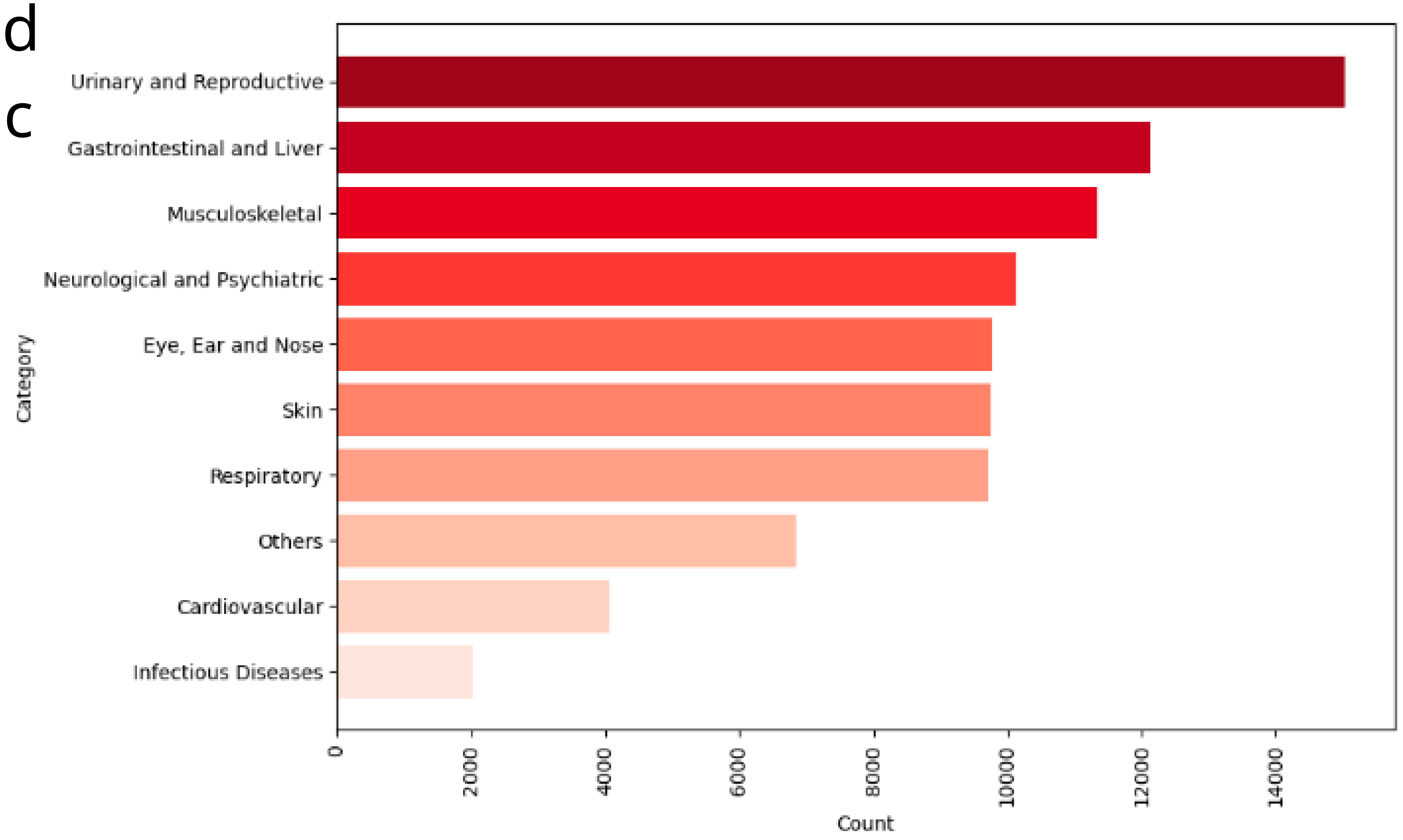
EDA

(Exploratory Data Analysis)

Category Distribution

The dataset is not perfectly balanced across categories, which is a realistic .representation of medical data

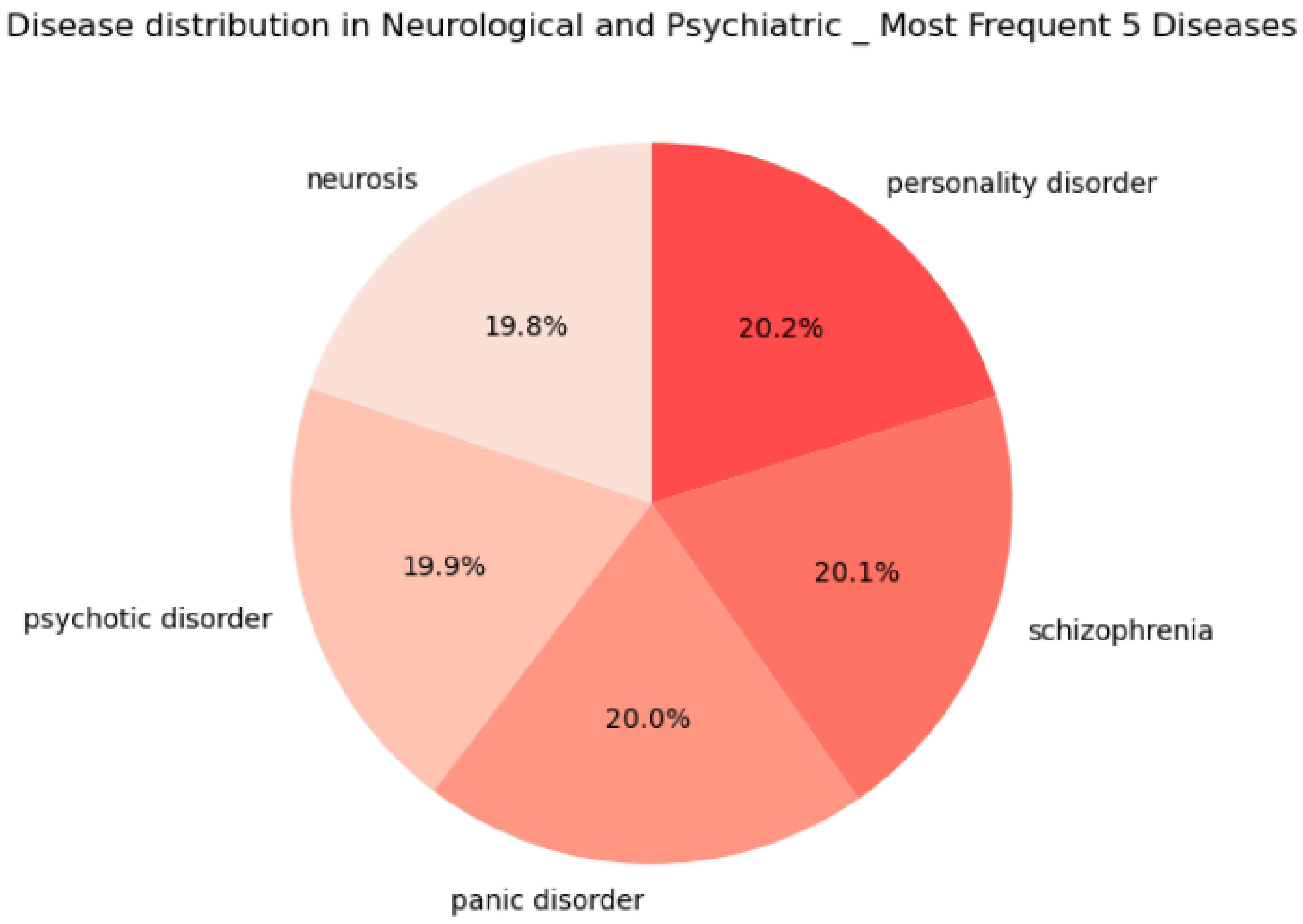
Finding: The Urinary and Reproductive category is the most prevalent (15,021 cases), while Infectious Diseases is the least prevalent (2,034 cases). This distribution must be considered during model training to avoid bias.



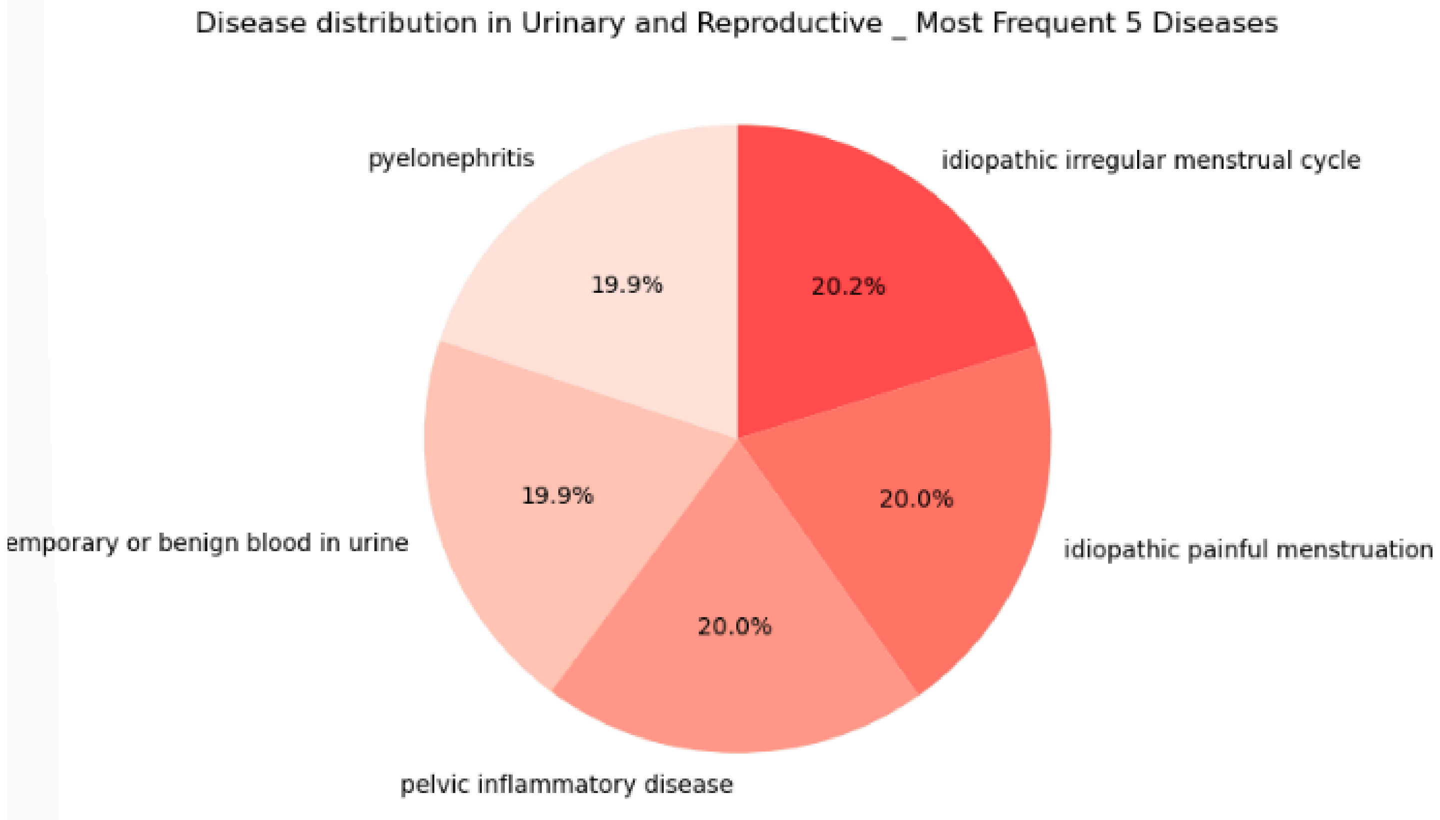
Top Diseases by Category

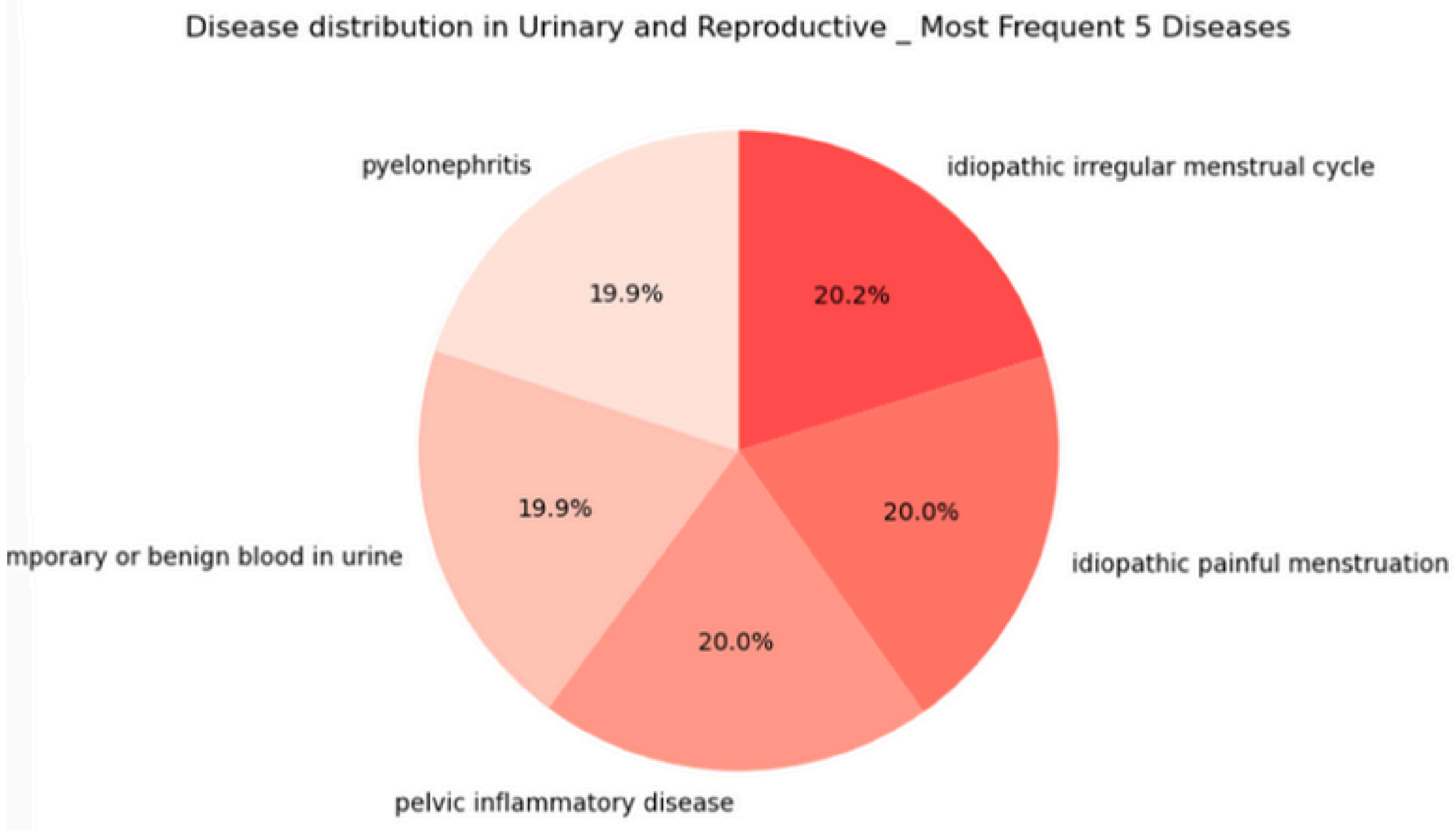
Most categories are dominated by a few :common diseases. For example

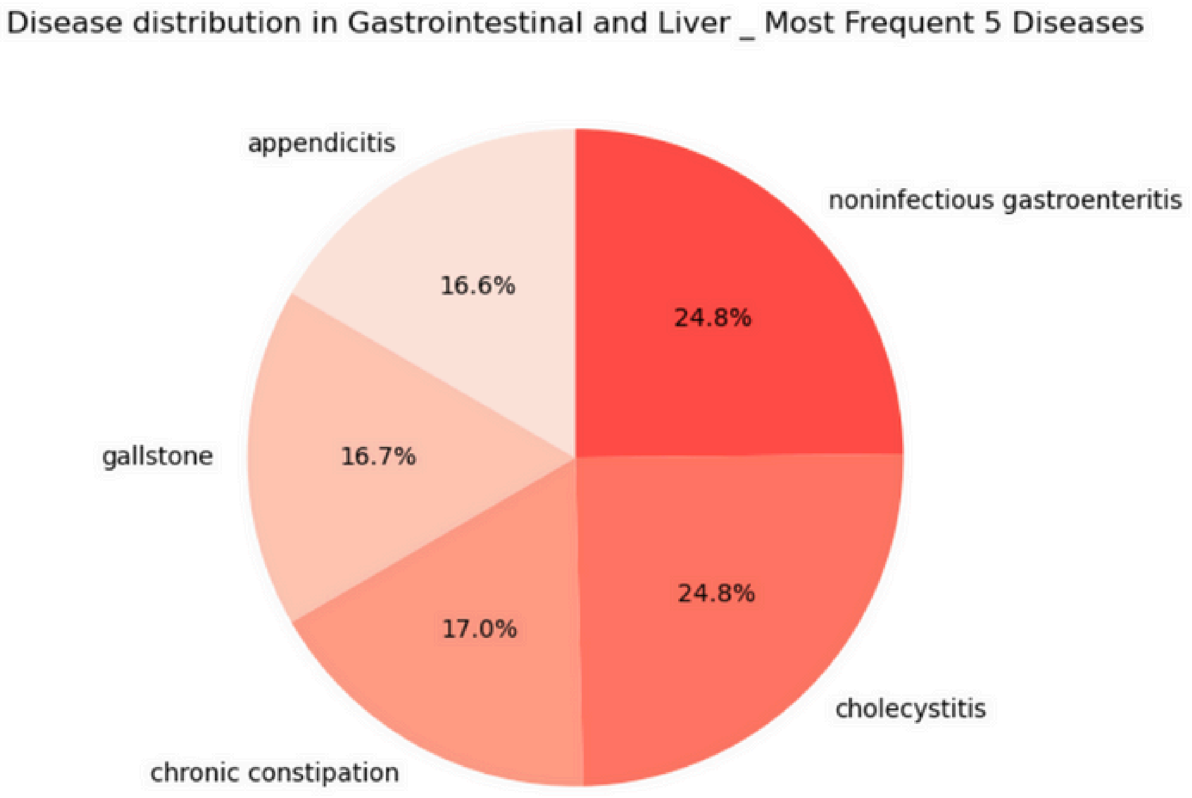
a. **Infectious Diseases**: is almost split between Sepsis (~60%) and Strep Throat (~40%).
b. **Neurological/Psychiatric**: is led by Neurosis, Psychotic Disorder, and Personality Disorder.
c. **Respiratory**: is led by Asthma and the Common Cold.

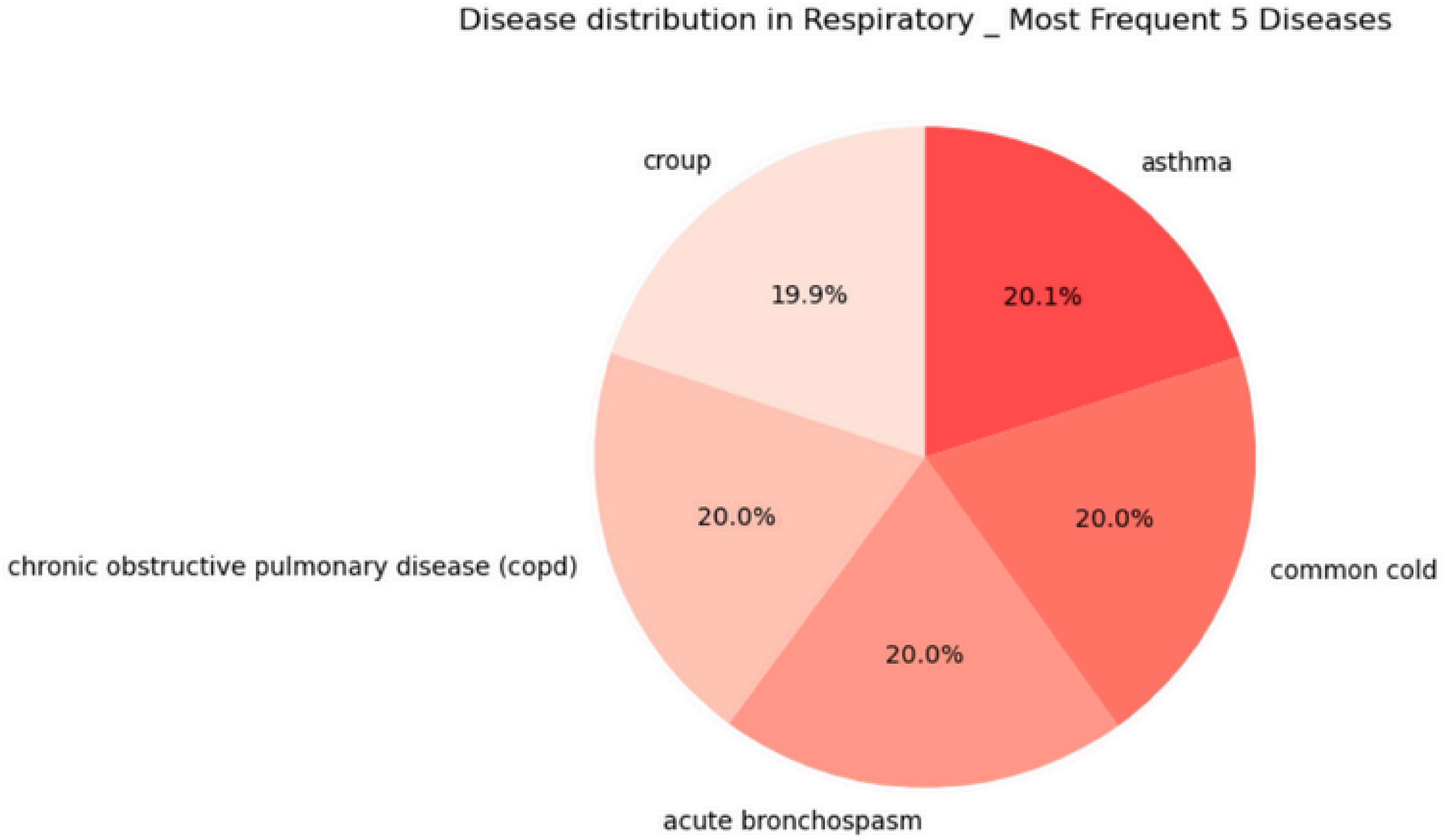


Top Diseases by Category

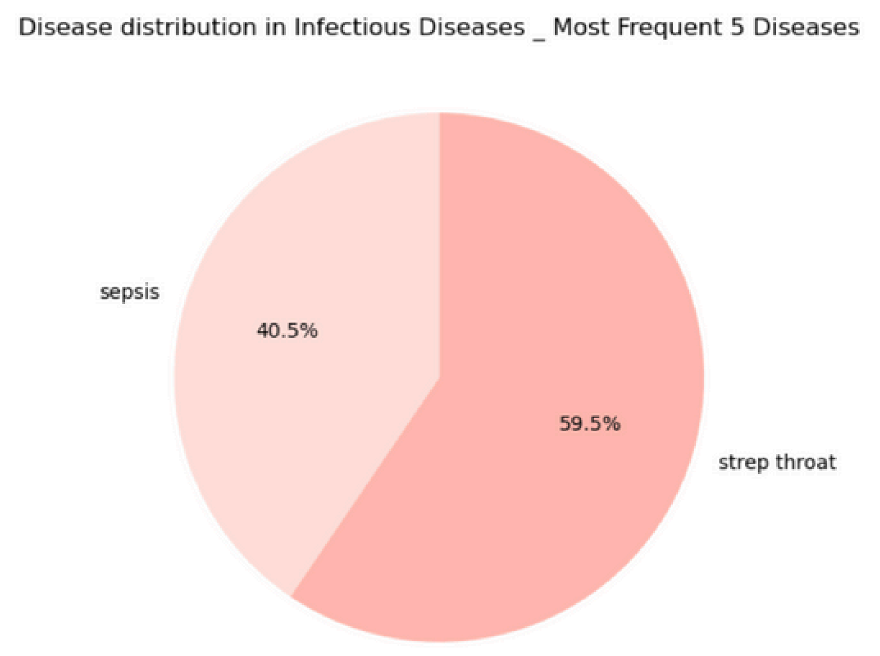


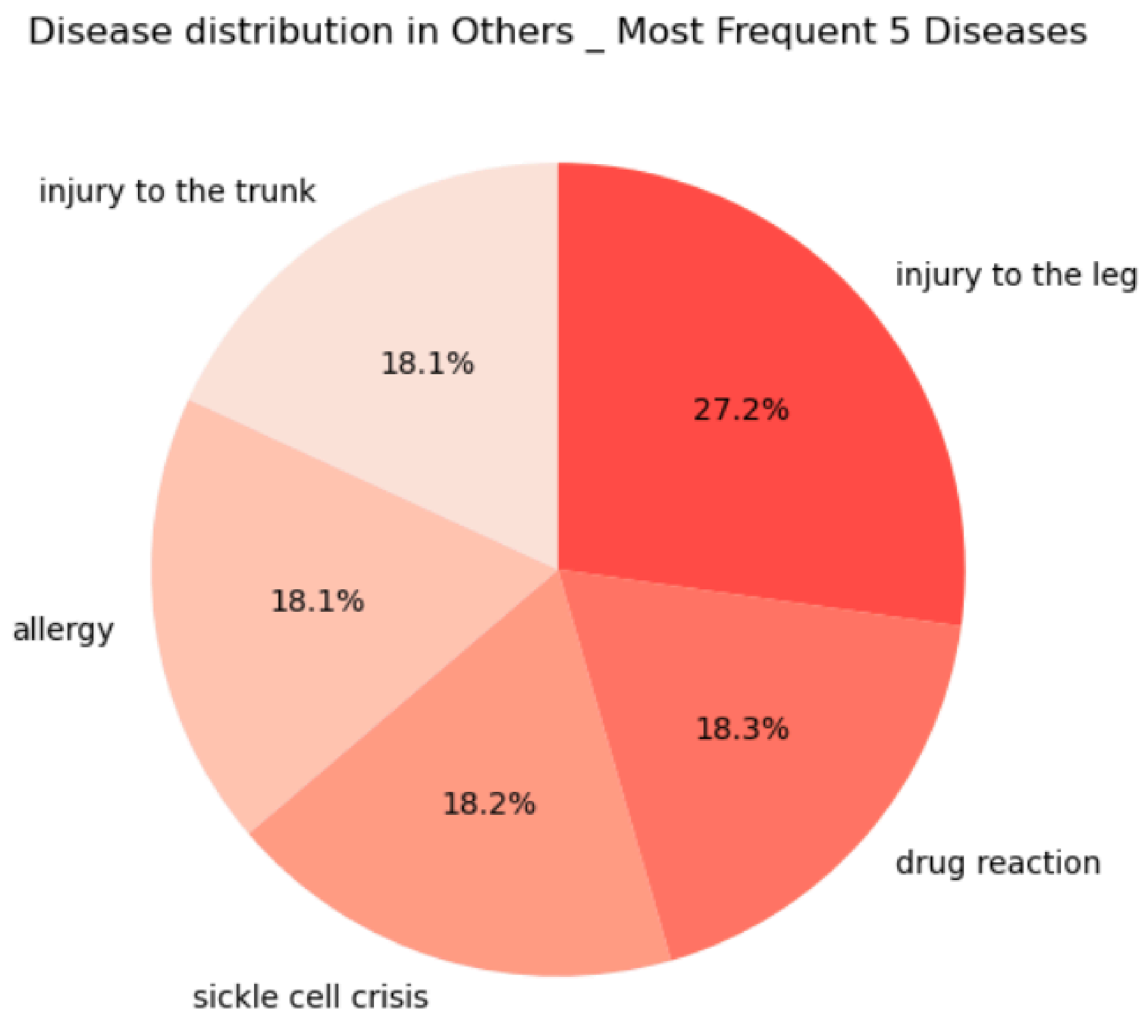


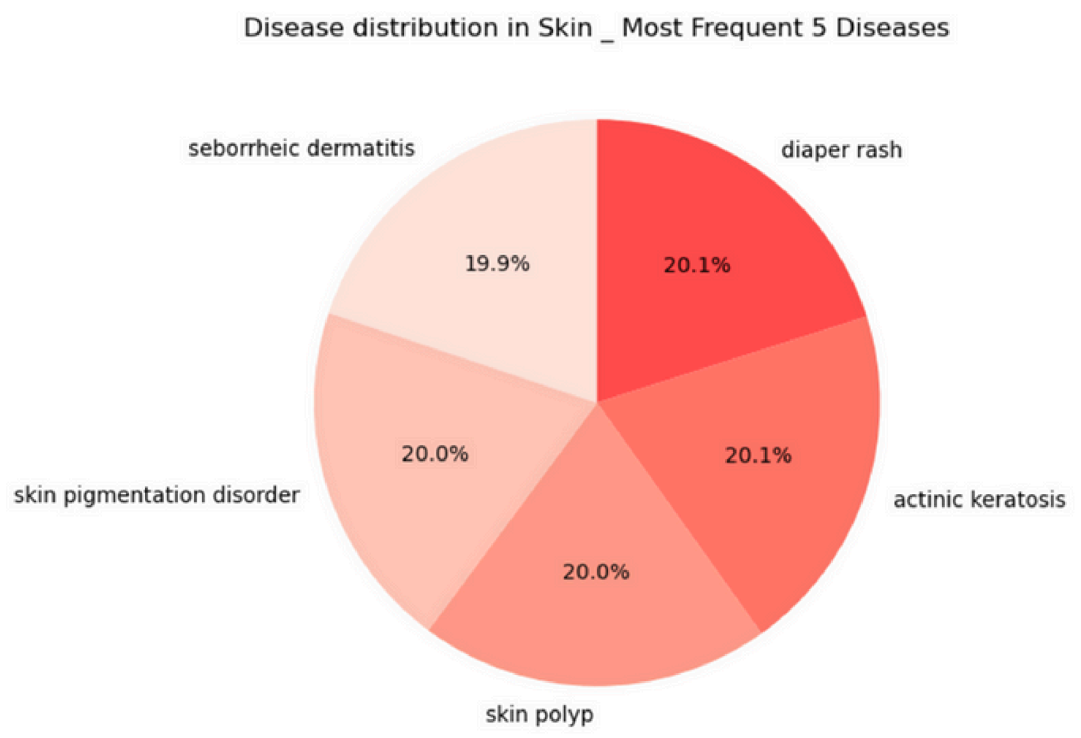


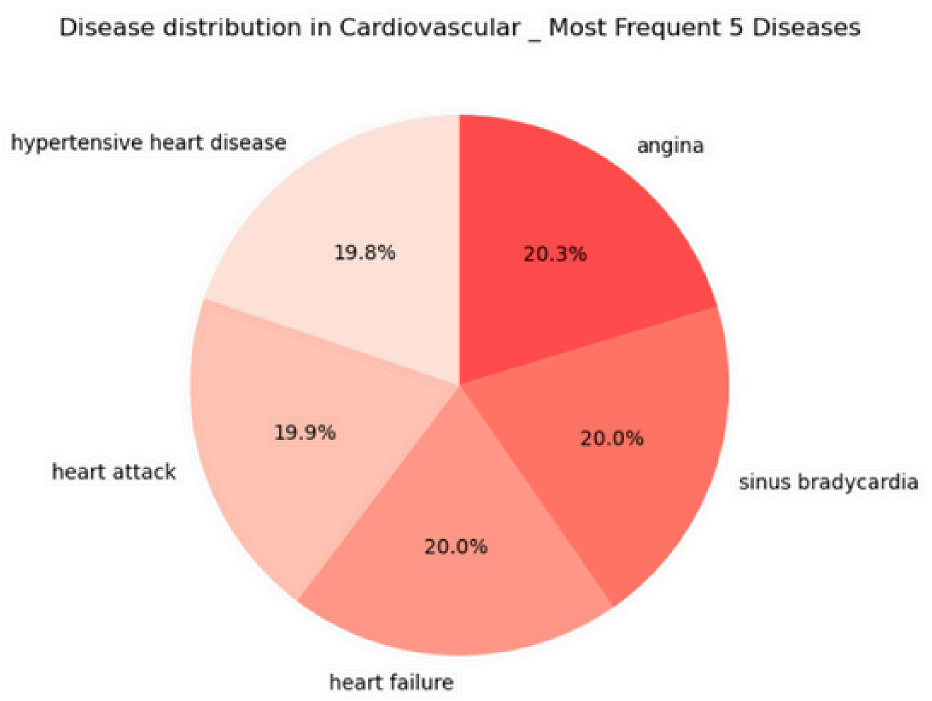


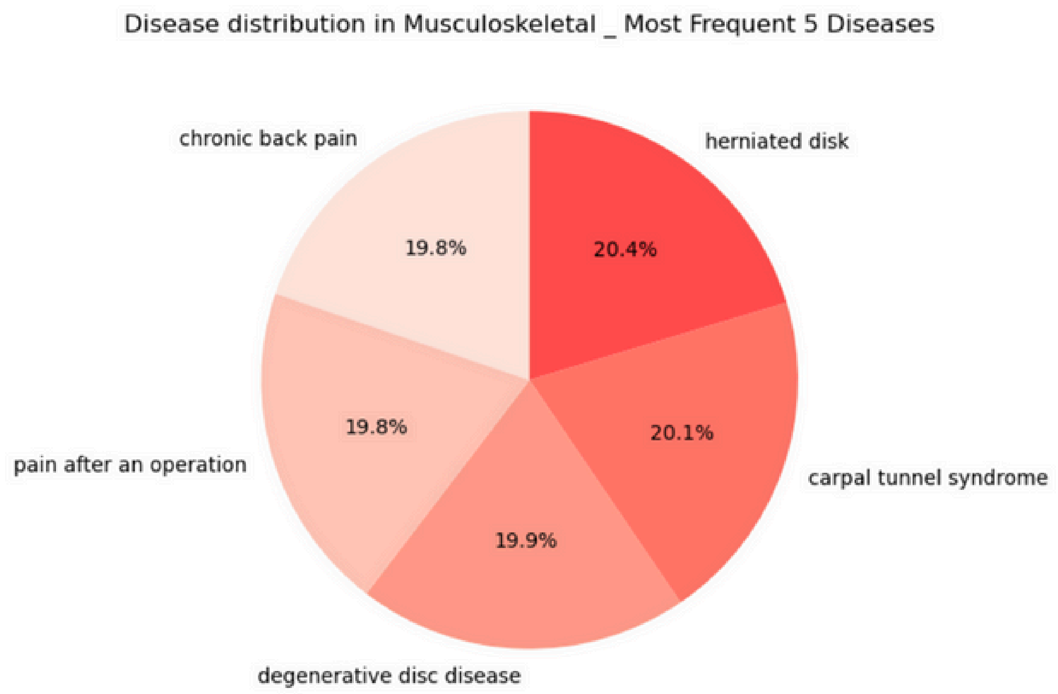
Top Diseases by Category



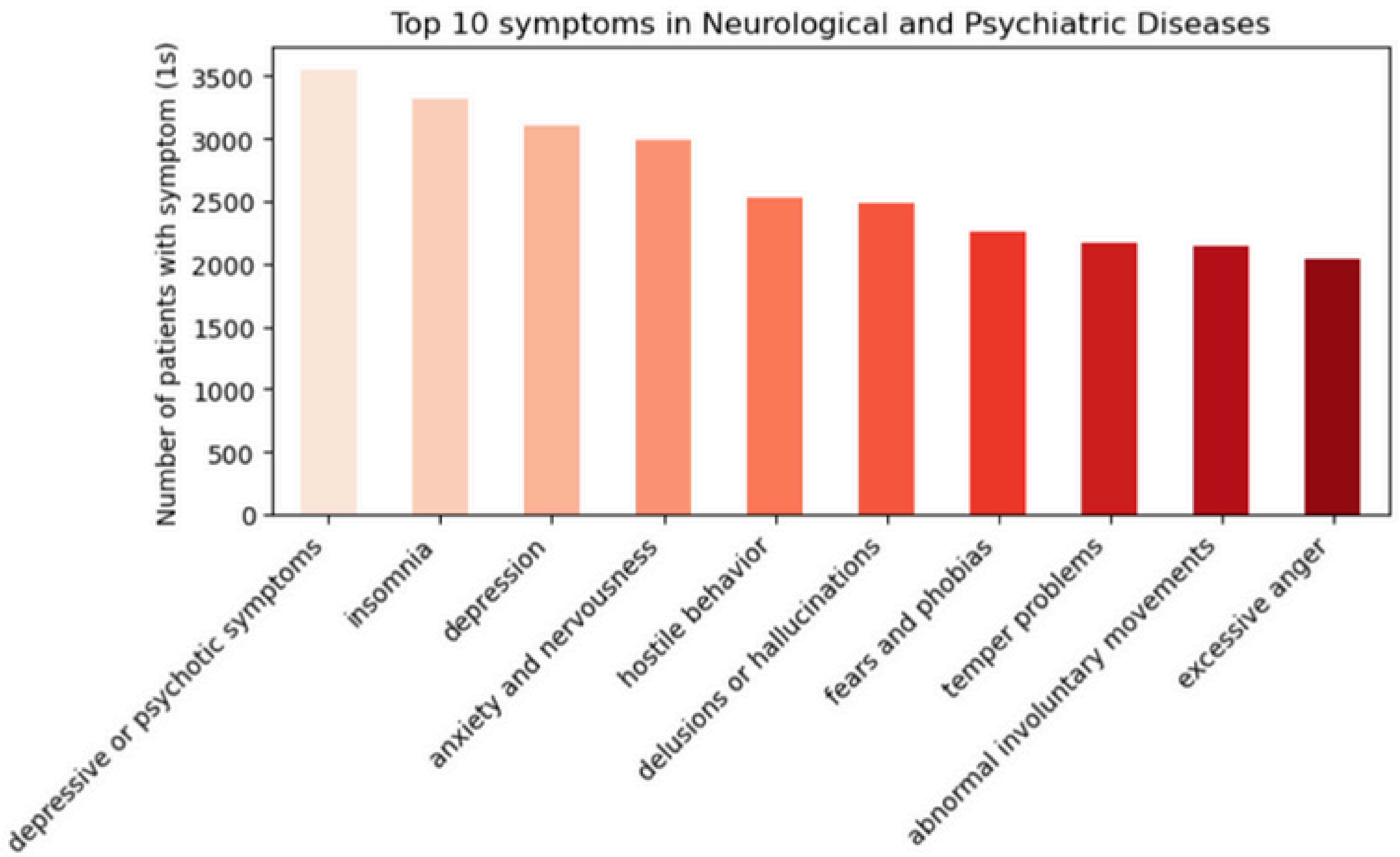






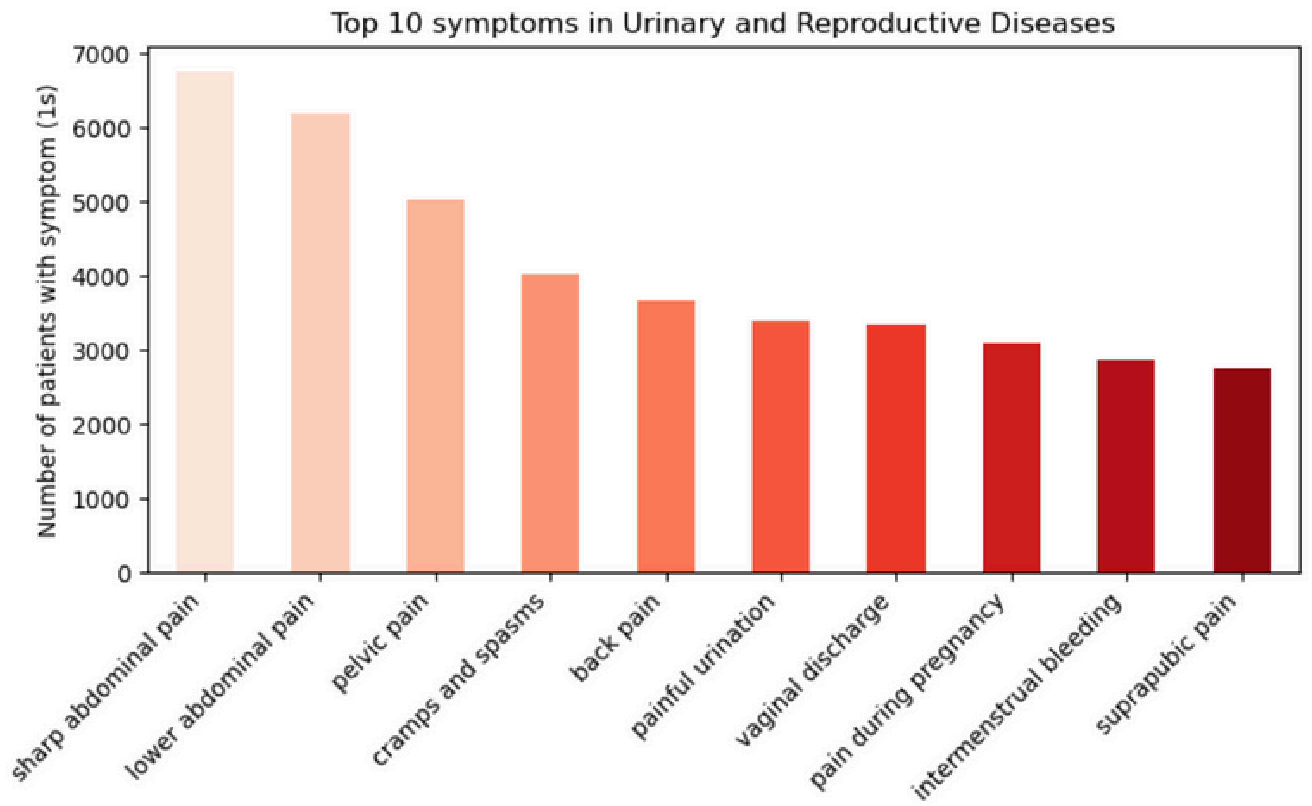


Symptom Analysis



Symptoms are highly specific to their category, validating the clinical logic of .the categorization

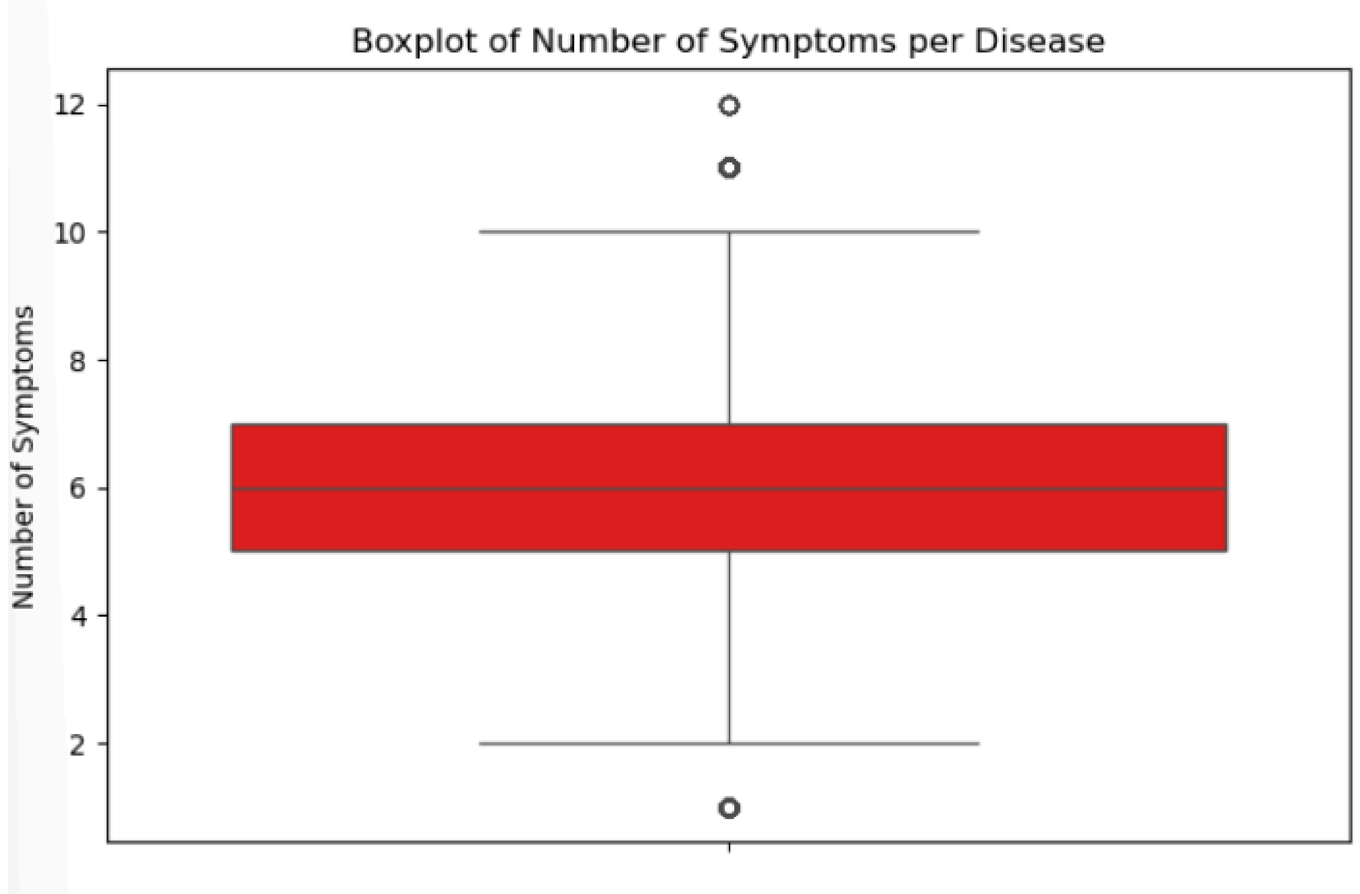
a. **Neurological/Psychiatric**: Top symptoms are psychological (e.g., depressive symptoms, anxiety, delusions).
b. **Gastrointestinal**: Top symptoms are physical and localized (e.g., abdominal pain, vomiting, diarrhea).
c. **Cardiovascular**: Top symptoms are cardio-respiratory (e.g., chest pain, shortness of breath, palpitations).
d. **Skin**: Top symptoms are dermatological (e.g., skin lesion, swelling, rash).

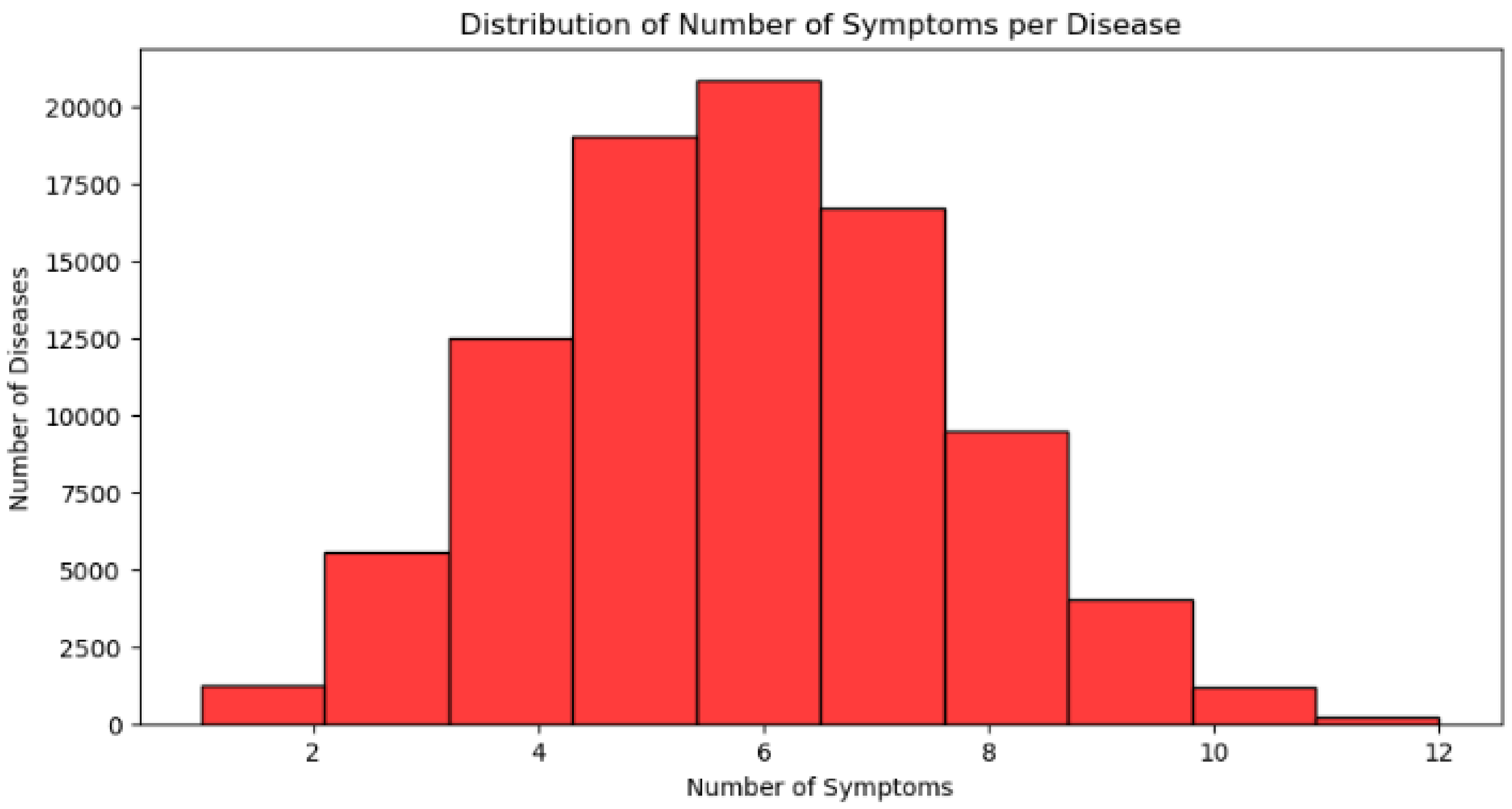


Symptom Count Analysis

To understand the complexity of cases in the dataset by analyzing the number of .symptoms per record

Finding: The num_symptoms feature, engineered by summing all symptom flags, could be a useful feature for .model training





Dataset

Doctors Clinics

Objective:

To clean, enrich, and analyze a dataset of Egyptian doctors, including their specialization, ratings, and clinic locations, with a focus on geocoding for spatial analysis.

Executive Summary

We processed a dataset of 1,210 initial doctor listings. Key steps included:

- **Data Cleaning**

  Handled missing values in critical columns (specialization, avg_rate, clinic_location).

- **Data Filtering**

  Retained only diseases with 800 or more occurrences, reducing the number of unique diseases from 773 to 105.

- **Feature Engineering**

  Mapped each doctor's specialization to one of 9 broader medical categories (e.g., Cardiovascular, Skin) for high-level analysis

- **Geocoding Challenge**

  Successfully retrieved latitude and longitude coordinates for clinic locations, a complex process involving multiple steps to handle incomplete addresses.

Initial Dataset Overview

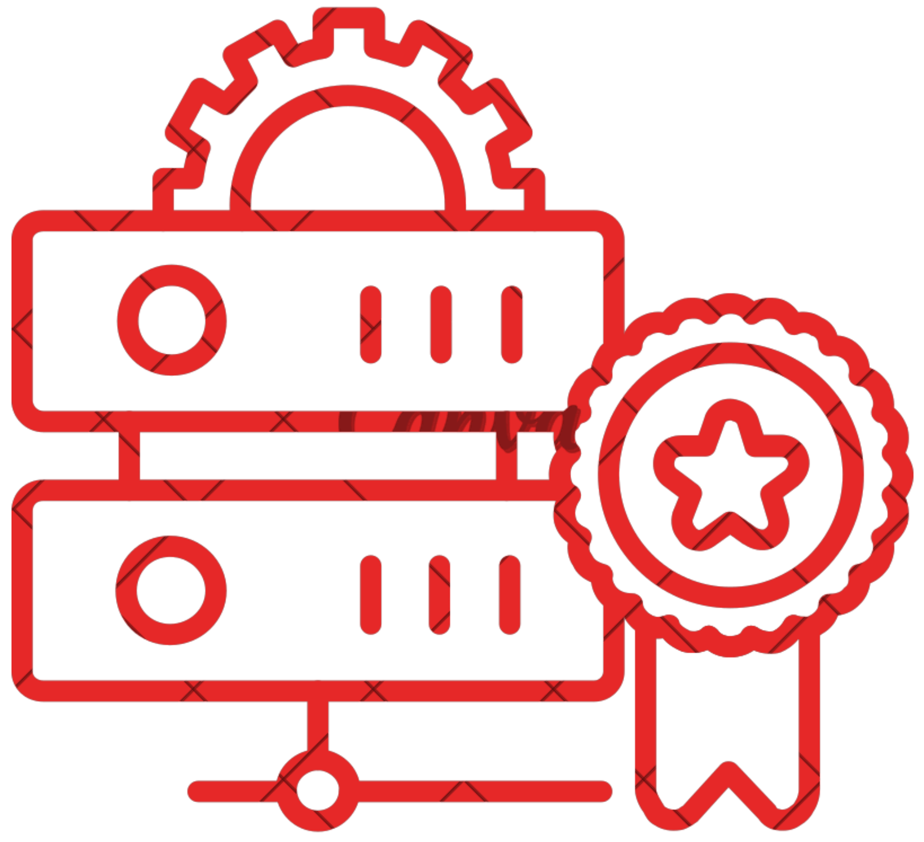
1,210 entries (rows), 10 original columns.

Columns

- specialization: Doctor's field (e.g., Dermatologist, Cardiologist).
- avg_rate: Average patient rating (Float64).
- clinic_location: Address or area of the clinic (String).
- Other columns: fees, waiting_time, rate_count, doctor_views, pages.

Data Quality Check

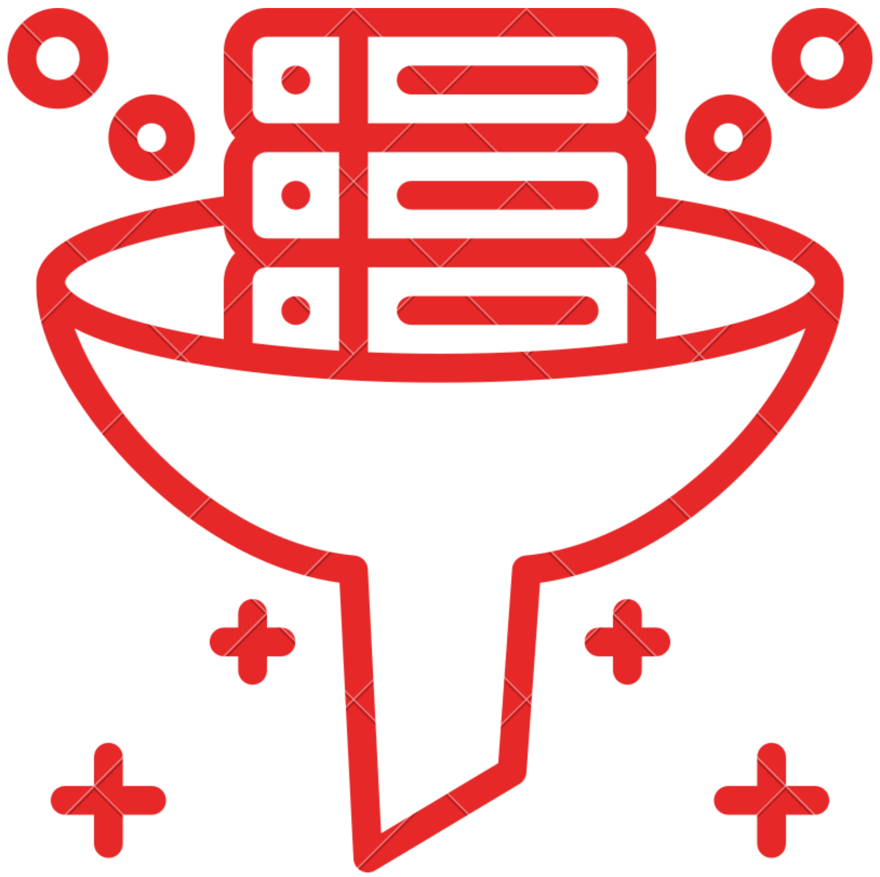
1. **Missing Values**: Present in several columns, most critically in clinic_location (283 missing) and avg_rate (59 missing).
2. **Summary Stats**: The avg_rate is very high, with a mean of 4.76 and 75% of doctors having a perfect 5.0 rating, indicating a potential rating bias.



Data Cleaning & Imputation

Handling Missing Values:

- **avg_rate**: 59 missing values filled with the column mean (~4.76).
- **clinic_location**: 283 rows with missing location were dropped entirely, as they are crucial for geocoding. This left 927 records to work with.



Feature Engineering - Categorization

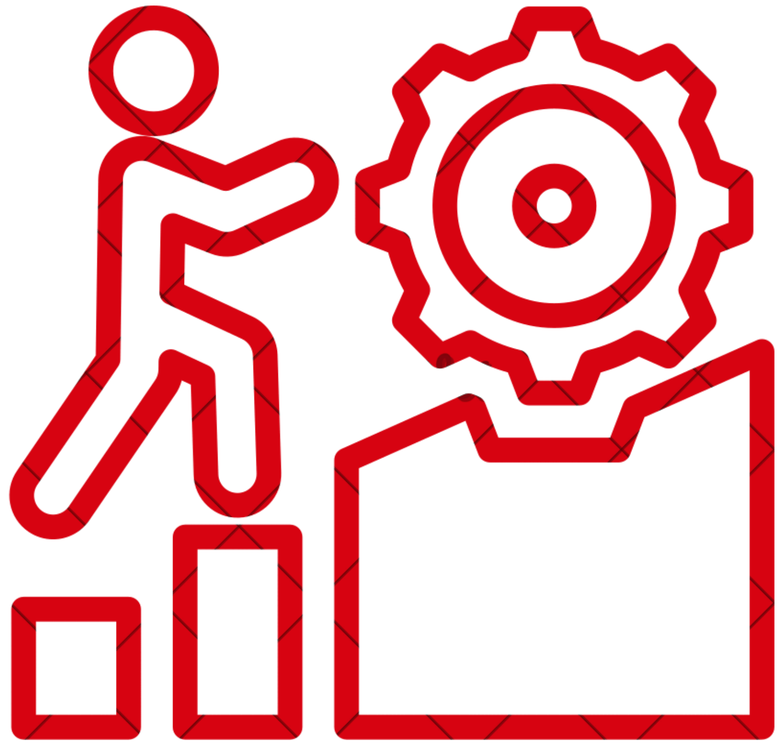
- A manual mapping dictionary was created to assign each specialization to a broader Category (e.g., 'Dermatologist' -> 'Skin', 'Cardiologist' -> 'Cardiovascular').
- This created a new, highly useful column for aggregated analysis.

The Geocoding Process

Challenge

- Converting text-based clinic_location entries (e.g., "Nasr City : Zaker Hussien Street") into numerical latitude and longitude.

Process



- **Attempt 1 (Exact Match):** Tried to geocode the full clinic_location string. Result: 396 out of 927 addresses failed.
- **Attempt 2 (General Area)**: For failed addresses, extracted the general area (e.g., "Nasr City" from "Nasr City : Zaker Hussien...") and geocoded that.
- **Result**: Successfully obtained coordinates for the majority of records. The final latitude and longitude columns were created by combining results from both attempts.

Final Cleaning

- Invalid coordinates (outside Egypt's approximate boundaries) were set to NaN. Remaining missing coordinates were set to 0 as a placeholder for EDA.

Final Dataset Structure

Final Columns

2750 rows × 7 columns

after handling duplicates and expansions during merge operations

	Doctor_Name	specialization	avg_rate	clinic_location	Category	latitude	longitude
0	Salim El-Shazly	Physiotherapist	5.0	El-Mansoura	Musculoskeletal	36.786091	9.900016
1	Salim El-Shazly	Physiotherapist	5.0	El-Mansoura	Musculoskeletal	36.786091	9.900016
2	Salim El-Shazly	Physiotherapist	5.0	El-Mansoura	Musculoskeletal	36.786091	9.900016
3	Salim El-Shazly	Physiotherapist	5.0	El-Mansoura	Musculoskeletal	36.786091	9.900016
4	Salim El-Shazly	Physiotherapist	5.0	El-Mansoura	Musculoskeletal	36.786091	9.900016
...	...	...	...	...	...	...	...

Ready for Analysis

- The dataset is now clean, categorized, and contains spatial data for mapping.

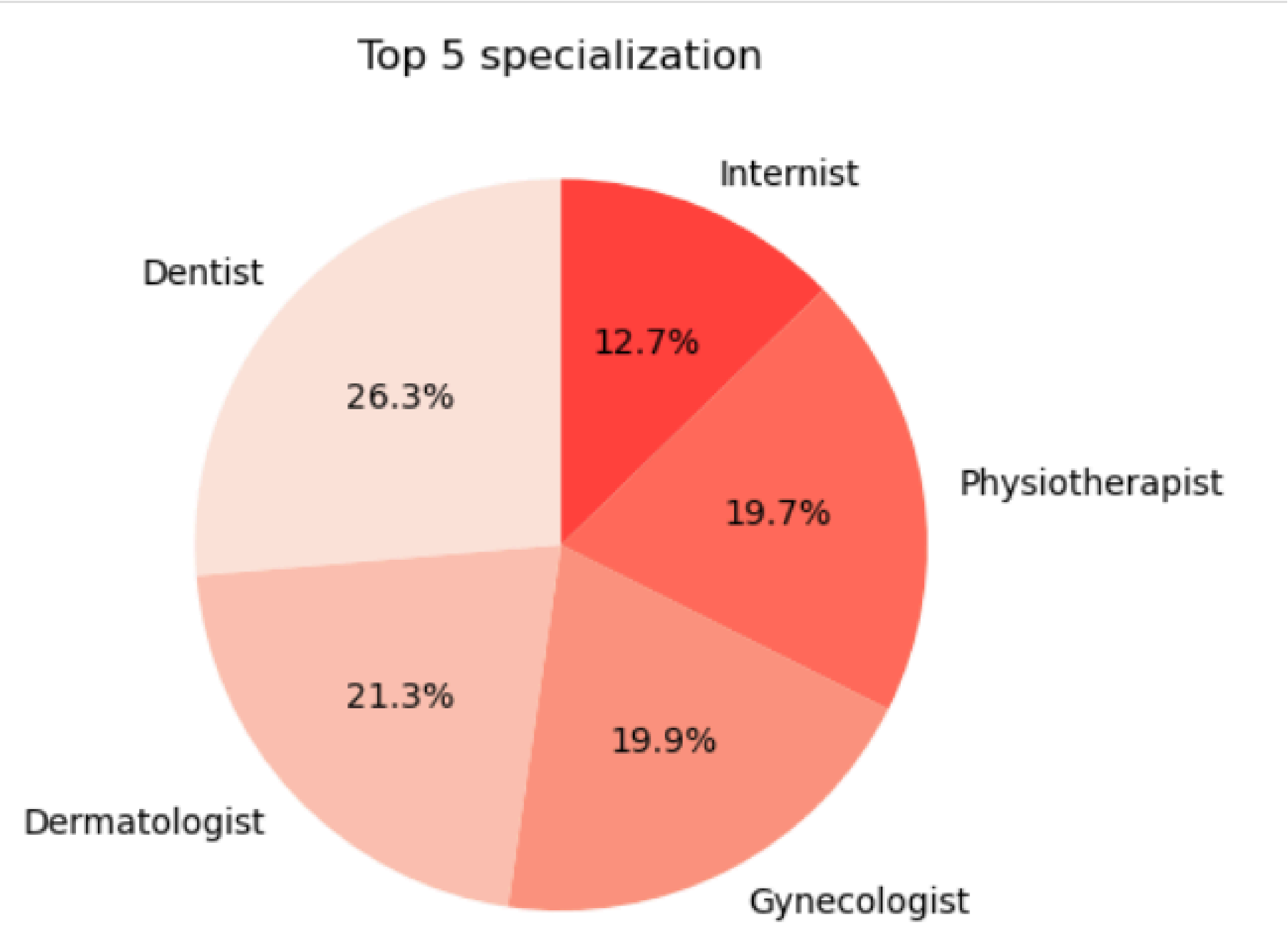
EDA

(Exploratory Data Analysis)

Specialization Distribution

The dataset is dominated by a few specializations. Dentists and Internists together make up nearly 40% of all .listings

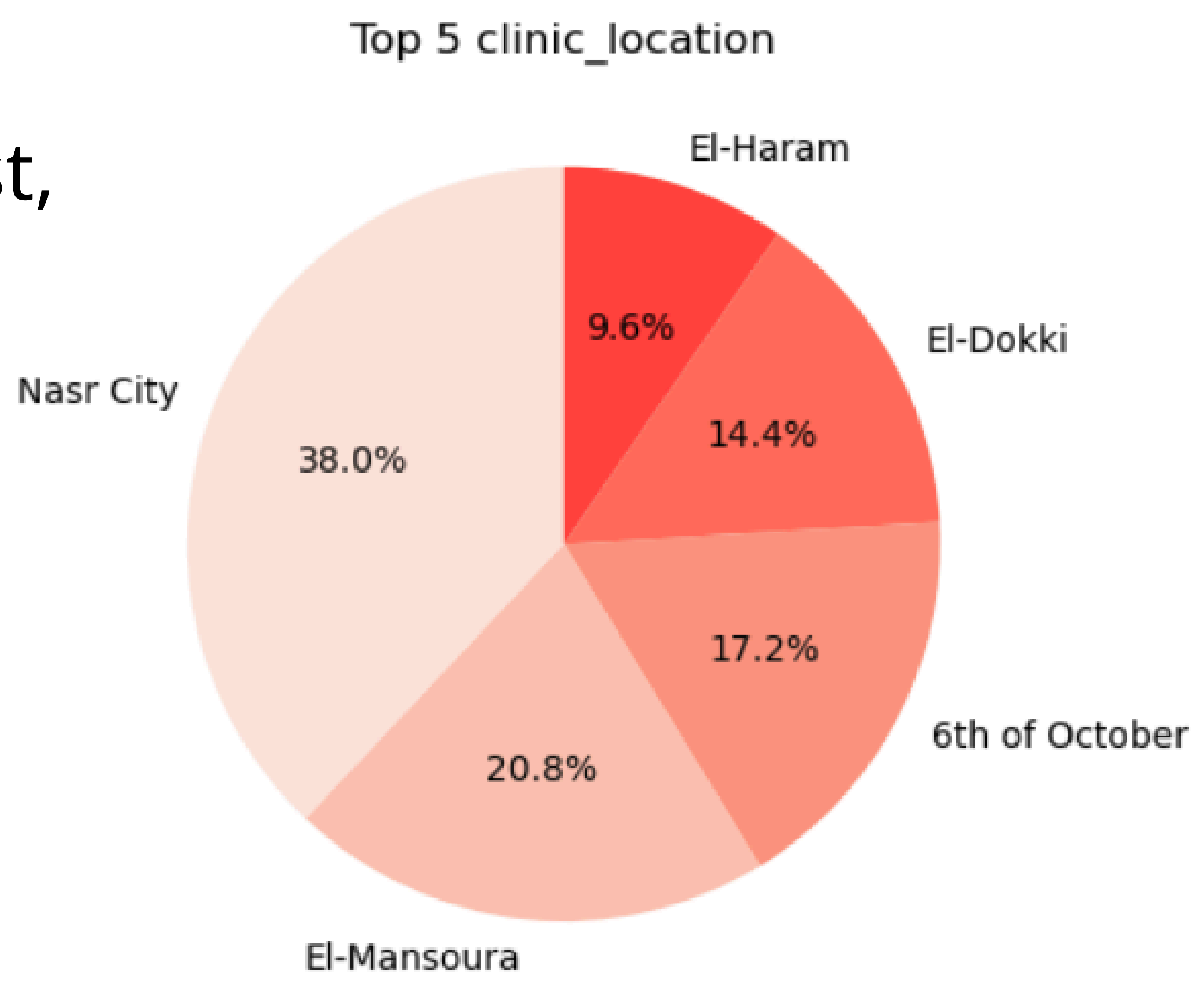
This indicates the dataset may not be perfectly representative of all medical fields in Egypt and is skewed towards more common general and dental practices.



Category Distribution

The "Others" category is the largest, which includes specializations like Dentist, Internist, and Nutritionist. This is consistent with the finding that Dentists and Internists are the .most common specializations

Following "Others", Musculoskeletal (e.g., Orthopedists, Physiotherapists) and Urinary and Reproductive (e.g., Gynecologists, Urologists) are the next largest categories.



Spatial Distribution

Doctors are heavily concentrated in Cairo and the Greater Cairo area (including Giza, 6th of October City). Significant clusters are also seen in .Alexandria and the Nile Delta region (e.g., Mansoura, Tanta)

There is a clear urban-rural divide in healthcare provider availability, which is a common challenge in many countries.

Next Step?

1. Proceed with building and training machine learning models for multi-class disease prediction.
2. Address class imbalance during model training.
3. Use Doctors data as a foundation for a doctor search and recommendation tool based on specialization, location, and patient ratings.

Thank You