PRIMED

A medicine Search System

Milestone #1: Data Preparation

Pedro Simões, up202403063@up.pt
Miguel Garrido, up202108889@up.pt
Emanuel Maia, up202107486@up.pt
Guilherme Martins, up202403106@up.pt

Table of contents

01
Introduction

02
Theme Selection

03
Data Collection

04
Pipeline Description

05
Data Model

06
Dataset Characterization

07
Conclusion

01

Introduction

02

Theme Selection

03

Data Collection

Data Selection

Medicine

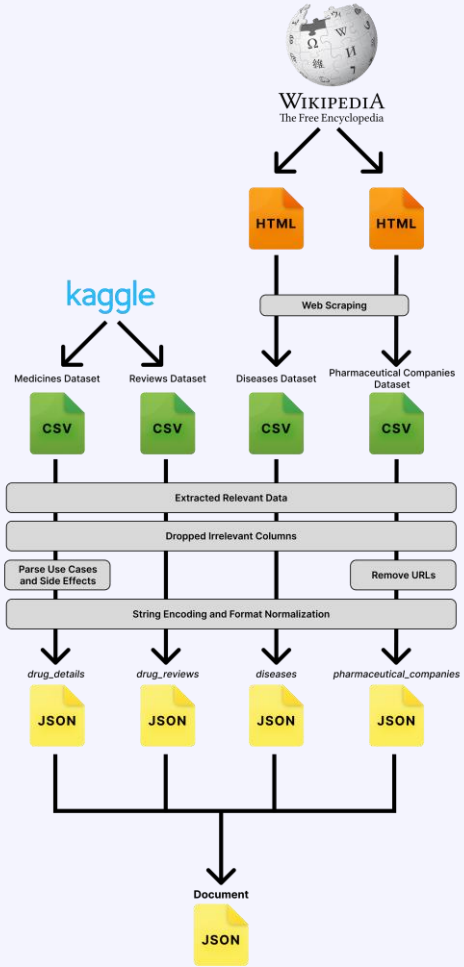
Disease

Pharmaceutical Companies

Reviews

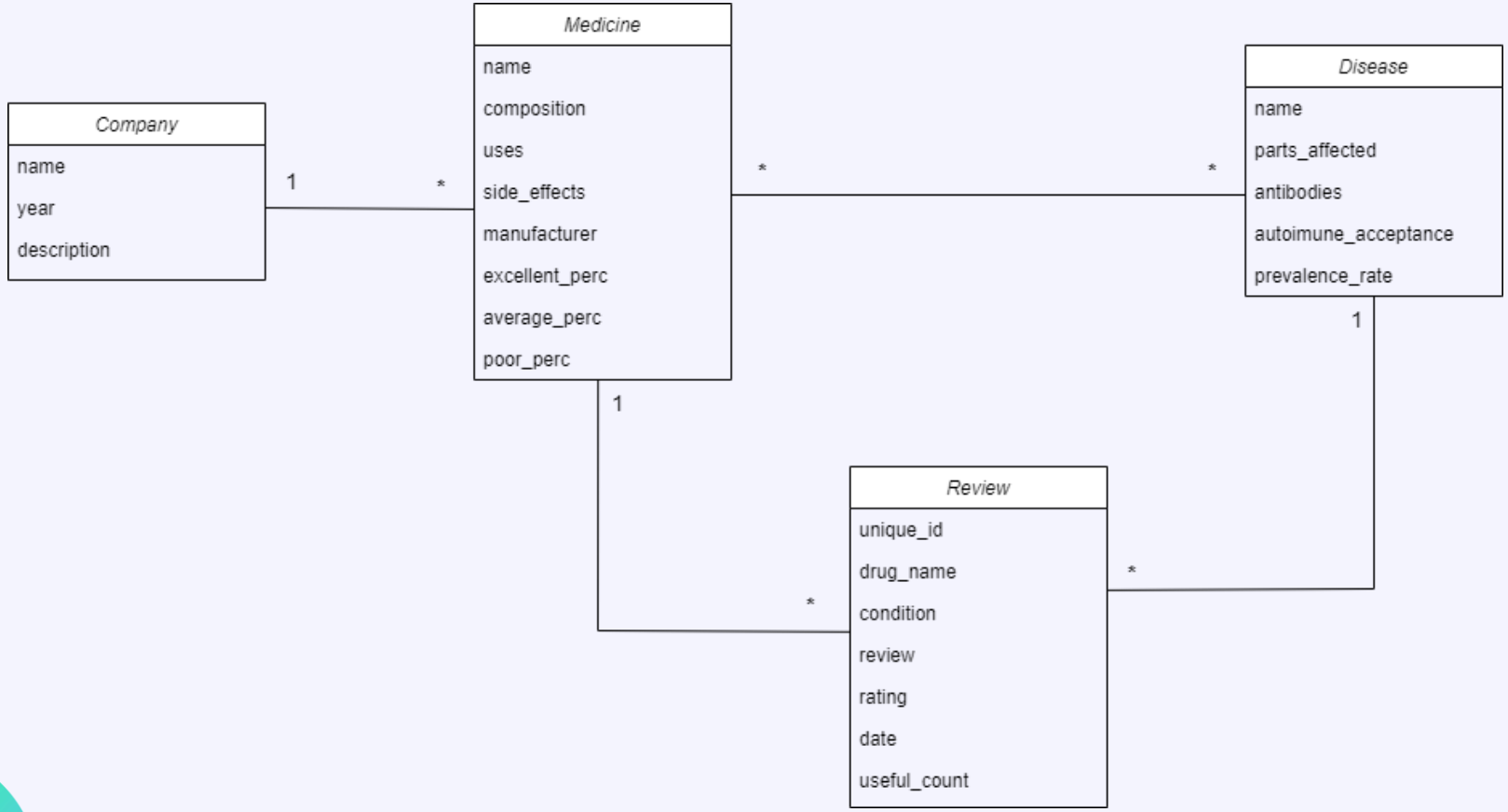
04

Pipeline Description


WIKIPEDIA
The Free Encyclopedia
HTML
HTML
Web Scraping
kaggle
Medicines Dataset
Reviews Dataset
Diseases Dataset
Pharmaceutical Companies Dataset
CSV
CSV
CSV
CSV
Extracted Relevant Data
Dropped Irrelevant Columns
Parse Use Cases and Side Effects
Remove URLs
String Encoding and Format Normalization
drug_details
drug_reviews
diseases
pharmaceutical_companies
JSON
JSON
JSON
JSON
Document
JSON

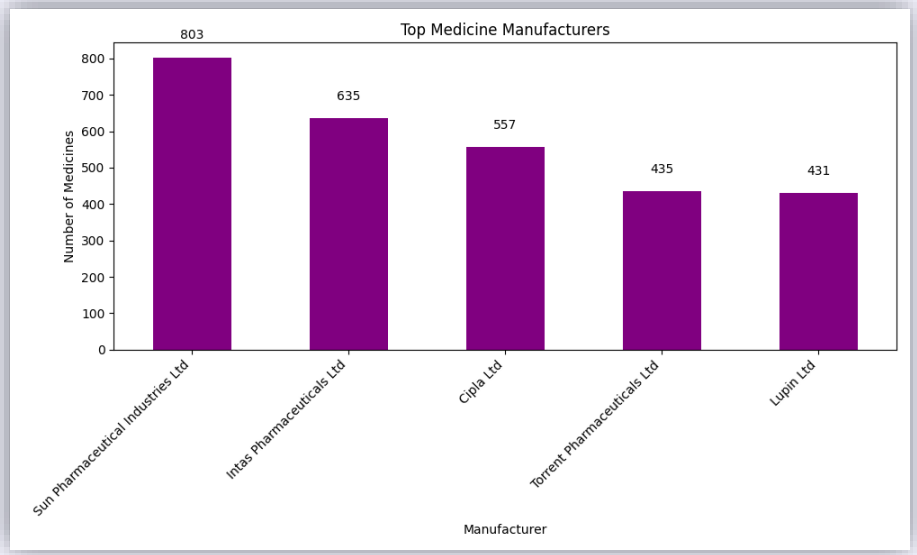
05

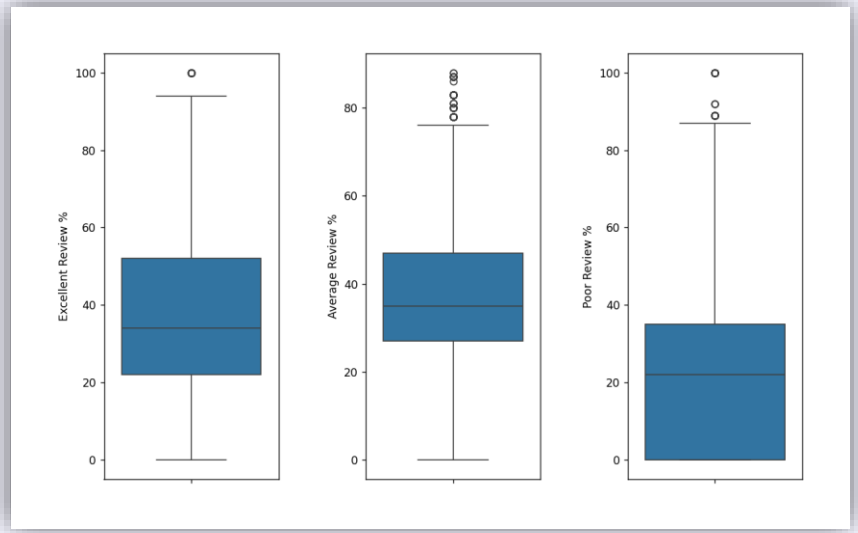
Conceptual Data Model

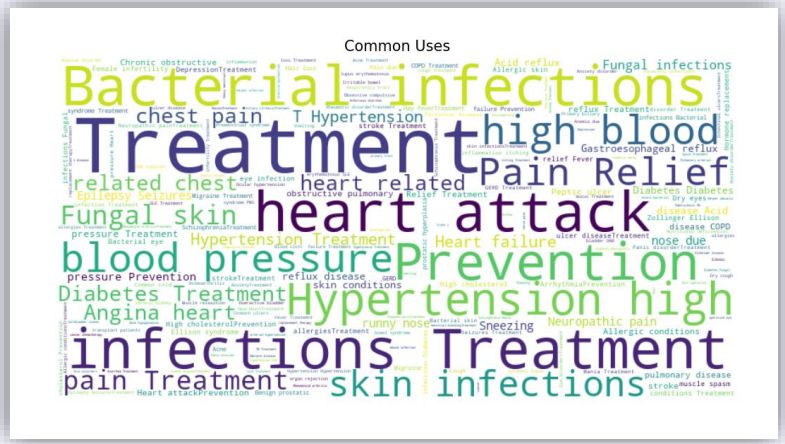

Company
name
year
description
1
*
Medicine
name
composition
uses
side_effects
manufacturer
excellent_perc
average_perc
poor_perc
*
*
Disease
name
parts_affected
antibodies
autoimune_acceptance
prevalence_rate
1
1
*
Review
unique_id
drug_name
condition
review
rating
date
useful_count
*

06

Dataset Characterization


Top Medicine Manufacturers
Number of Medicines
Manufacturer
803
635
557
435
431
Sun Pharmaceutical Industries Ltd
Intas Pharmaceuticals Ltd
Cipla Ltd
Torrent Pharmaceuticals Ltd
Lupin Ltd


Excellent Review %
Average Review %
Poor Review %


Common Uses
Bacterial infections
Treatment
high blood
Pain Relief
heart attack
Prevention
Hypertension high
infections Treatment
skin infections
chest pain
Hypertension
related chest
heart related
Fungal skin
blood pressure
Diabetes Treatment
Angina heart
pain Treatment
Fungal infections

07

Conclusion

References

1. **Kaggle**. https://www.kaggle.com/, 2024.
2. **Wikipedia**. https://www.wikipedia.org/, 2024.
3. **Kaggle**. 11000 medicine details. https://www.kaggle.com/datasets/singhnavjot2062001/11000-medicine-details, 2024.
4. **Creative Commons**. Cc0 1.0 universal. https://creativecommons.org/publicdomain/zero/1.0/.
5. **Wikipedia**. https://en.wikipedia.org/wiki/List_of_autoimmune_diseases, 2024.
6. **Creative Commons**. Creative commons attribution-sharealike 4.0 interna- tional license. https://en.wikipedia.org/wiki/Wikipedia:Text_of_the_Creative_Commons_Attribution-ShareAlike_4.0_International_License.
7. **Wikipedia**. https://en.wikipedia.org/wiki/List_of_pharmaceutical_companies, 2024.
8. **Surya Kallumadi and Felix Grer**. Drug reviews. https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com, 2018. UC Irvine Machine Learning Repository.
9. **UC Irvine**. https://archive.ics.uci.edu/.
10. **Creative Commons**. Attribution 4.0 international. https://creativecommons.org/licenses/by/4.0/legalcode.
11. **Python Software Foundation**. Python Programming Language. Python Software Foundation, 2023. Version 3.x.
12. **The pandas development team**. pandas: Python Data Analysis Library, 2023. Version 2.x.
13. **Takahiko Wada**. unidecode: Text transliteration for Python, 2023. Version 1.x.
14. **Python Software Foundation**. html: HTML and XHTML processing for Python, Python Standard Library.
15. **Python Software Foundation**. json: JSON encoder and decoder for Python, 2023. Python Standard Library.
16. **Python Software Foundation**. csv: CSV file reading and writing for Python, 2023. Python Standard Library.
17. **Pritha Bhandari**. How to find interquartile range (iqr) | calculator & examples. https://www.scribbr.com/statistics/interquartile-range/, 2023.