Final Project Proposal
October 30, 2025

David Lewis
lewis3d7@mail.uc.edu

Contents

1. Introduction

1.1. Background

There are lots of medicines that do different things. It is hard for physicians to pick between each of the different medications. Simplifying the list of medications is therefore important to patient care.

1.2. Research questions

- Which medications with the same compositions are the highest rated?

- How does the usage of the medication correlate with user reviews?

- How do side effects correlate?

- How can a physician easily browse for medications?

2. Methods

2.1. Data collection

The data were collected from Kaggle using the search term "medicine". The dataset is located at the following URL: https://www.kaggle.com/datasets/singhnavjot2062001/11000-medicine-details

2.2. Data Manipulation

The data was originally a single CSV. There were many "many-to-many" relationships, which I've extracted (reverse group-by?) into separate tables.

The Medications table is shown below:

	Medicine Name	Manufacturer	Excellent Review %	Average Review %	Poor Review %
0	Avastin 400mg Injection	Roche Products India Pvt Ltd	22	56	22
1	Augmentin 625 Duo Tablet	Glaxo SmithKline Pharmaceuticals Ltd	47	35	18
2	Azithral 500 Tablet	Alembic Pharmaceuticals Ltd	39	40	21
3	Ascoril LS Syrup	Glenmark Pharmaceuticals Ltd	24	41	35
4	Aciloc 150 Tablet	Cadila Pharmaceuticals Ltd	34	37	29
5	Allegra 120mg Tablet	Sanofi India Ltd	35	42	23
6	Avil 25 Tablet	Sanofi India Ltd	40	34	26
7	Aricep 5 Tablet	Eisai Pharmaceuticals India Pvt Ltd	43	28	29
8	Amoxyclav 625 Tablet	Abbott	36	43	21
9	Atarax 25mg Tablet	Dr Reddy's Laboratories Ltd	35	41	24
10	Azee 500 Tablet	Cipla Ltd	37	38	25

> "I'm hiding the `Image URL` column here""
>
> — David Lewis

Some stats from the tables:

2.2.1. Medications table

	Unique Counts
Medicine Name	11498
Image URL	11740
Manufacturer	759
Excellent Review %	85
Average Review %	79
Poor Review %	81

2.2.2. Medications Uses Junction table

	Unique Counts
Medicine Uses ID	17811
Medicine Name	11498
Use	469

2.3. ER diagram
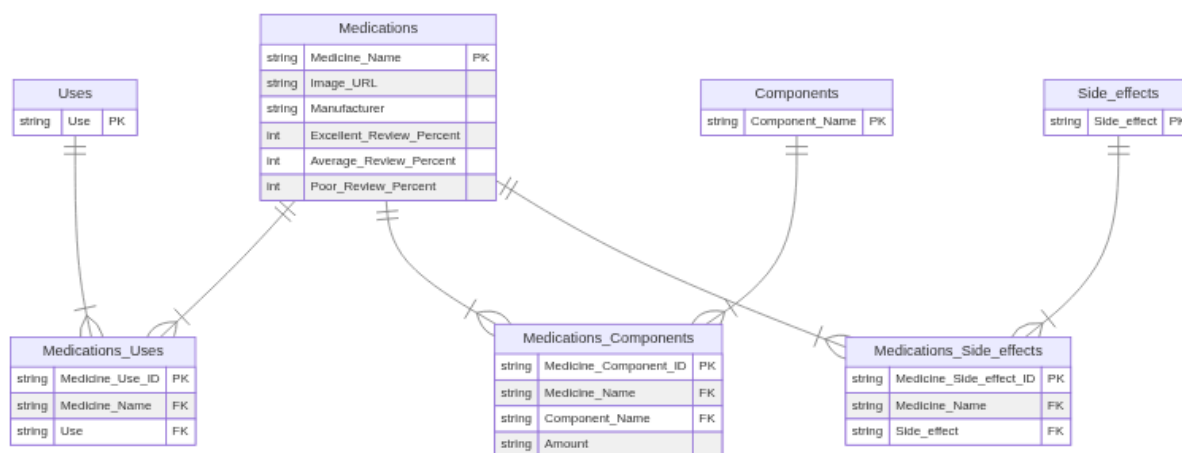
This ER diagram makes sense because the original dataset in csv had columns that were lists with non-unique elements, i.e. a many-to many relationship. This database takes those columns and extracts the columns as their own tables, connecting them via a junction table.

2.4. Data analysis

- Statistical analysis to determine effect of Use/Composition/Side effects on Review

- Compare reviews of medicines with similar compositions (determine what "similar" means in this context)

2.5. R Packages

2.5.1. Statistical packages

- Tidyverse

- Plotly

Pretty sure I can do all the statistical analysis I need to without any extra packages. Tidyverse is a family of data manipulation/visualization tools and plotly is for interactive visualization.

2.5.2. Shiny stuff

- Shiny

- Shiny Router

- Box

Shiny should be self-explanatory, but shiny router and box allow for actual modularization within the R app (i.e. I don't have to store everything in the same file).

2.6. Interactive Visualizations

- Searchable histogram (i.e. histogram changes) for each component's dosages
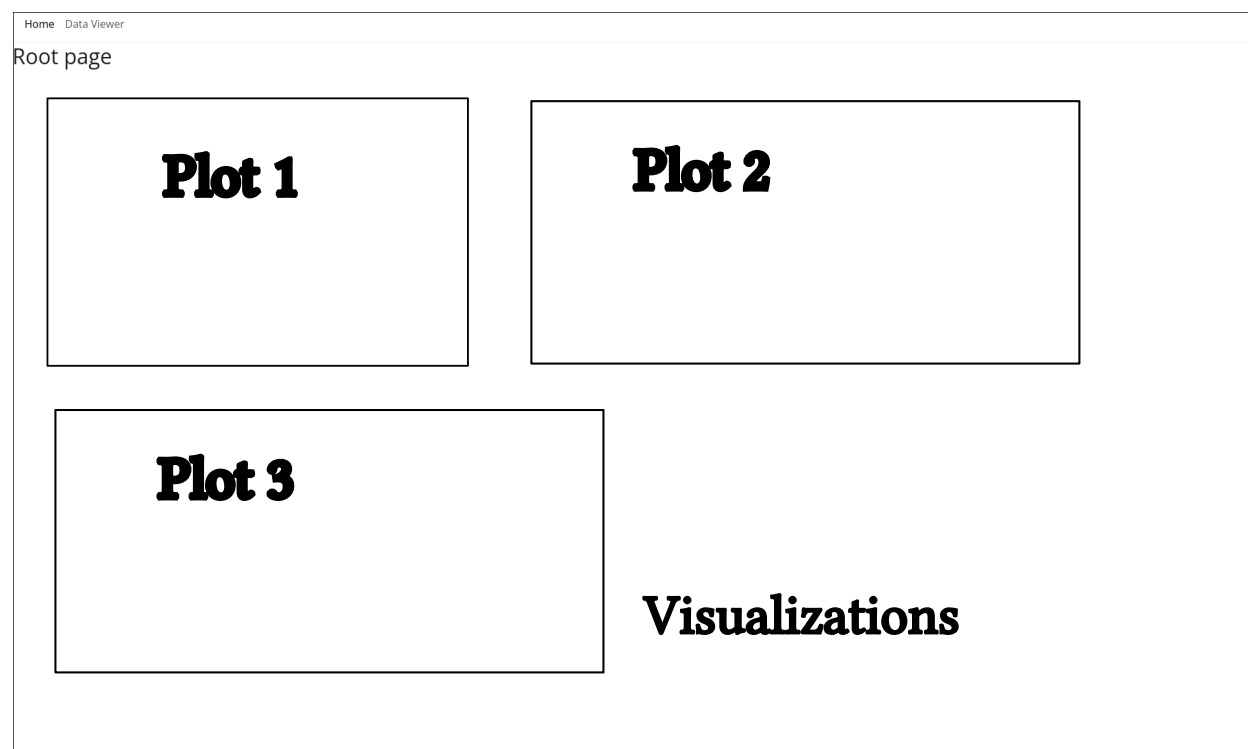
- Searchable component to review histogram

2.7. Layout

1 tab for non-interactive/interactive data vizualization, another for browsing the data.

> "I'm not really sure *which* plots I'm using yet, but there will be plots on the first page, data browser on the second"
>
> — David Lewis

2.7.1. Page 1

2.7.2. Page 2

Home   Data Viewer

Show 10 ˅ entries   Search: _____

	Medicine.Name	Composition	Uses	Side_effects	Image.URL	Manufacturer	Excellent.Review..	Average.Review..	Poor.Review..
1	Avastin 400mg Injection	Bevacizumab (400mg)	Cancer of colon and rectum Non-small cell lung cancer Kidney cancer Brain tumor Ovarian cancer Cervical cancer	Rectal bleeding Taste change Headache Nosebleeds Back pain Dry skin High blood pressure Protein in urine Inflammation of the nose		Roche Products India Pvt Ltd	22	56	22
2	Augmentin 625 Duo Tablet	Amoxycillin (500mg) + Clavulanic Acid (125mg)	Treatment of Bacterial infections	Vomiting Nausea Diarrhea Mucocutaneous candidiasis		Glaxo SmithKline Pharmaceuticals Ltd	47	35	18
3	Azithral 500 Tablet	Azithromycin (500mg)	Treatment of Bacterial infections	Nausea Abdominal pain Diarrhea		Alembic Pharmaceuticals Ltd	39	40	21
4	Ascoril LS Syrup	Ambroxol (30mg/5ml) + Levosalbutamol (1mg/5ml) + Guaifenesin (50mg/5ml)	Treatment of Cough with mucus	Nausea Vomiting Diarrhea Upset stomach Stomach pain Allergic reaction Dizziness Headache Rash Hives Tremors Palpitations Muscle cramp Increased heart rate		Glenmark Pharmaceuticals Ltd	24	41	35

2.8. non-interactive visualizations (ggplot/plotly)

In general, these will be visualizations for the entirety of the data

- Some sort of scatter plot showing relationships between medications (UMAP/PCA) based on various metrics (need to define similarity metric)

- Will try to use plotly for everything, since it allows for interactive inspection of points

> "I have some other ideas for this, but nothing super detailed yet."
>
> — David Lewis

2.9. Filters

- I'm not going to filter the data at all, since there is no statistical score to do so with, i.e. p value.

2.10. How proposed visualizations help answer the research questions?

- The interactive data viewer will help physicians connect medications to symptoms/reviews/etc.

- Interactive/non interactive visualizations may show how some medications may be really well received by some, but terrible for others

- UMAP/PCA plot may show distinct groups of medicines in terms of side effect, uses, etc. May be able to select alternative medications based on groups.