Final Project Report
December 08, 2025

David Lewis
lewis3d7@mail.uc.edu

Contents

1. Introduction

According to Drugbank, there are around 20,000 unique drugs used in the healthcare system.[1] To contrast, the average 20 year old American knows around 40,000 words, though not necessarily what each word means.[2] Therefore, if a physician were to learn all of the drug names, they would take up around 1/3 of the physician's entire vocabulary. Having full comprehension of all 20,000 drugs would likely be extremely difficult, if not impossible. Free and open access tools typically do not provide interactive visualizations of drug relationships.[3–5] Instead, they may provide comparison charts or articles.[6]

This project explores the development of an interactive tool meant to aid physicians in medicine selection and comparison. The tool provides interactive visualizations meant to make the comparison of medications easier, showing how different medications relate among their attributes including: user reviews, composition, side effects, and usage.

2. Methods

2.1. Data source

The data were collected from Kaggle using the search term "medicine". The collected dataset is entitled "11000 Medicine details".[7]

After collection of the data, which were contained within a single csv file, the data were processed into an SQLite database using python and the pandas package.[8] The database uses the database architecture defined in Figure 1. The database is then queried by the visualization tool, an R shiny app.[9]

The tool was designed to have two web pages, a "home" page, and a "data viewer" page. The home page would contain simple statistics and visualizations for the data set, including the total number of medicines, while the data viewer page would contain an interactive browser of the medication, with visualizations that change based on which medications are selected from a list. The list would be filterable, and multiple medications could be selected at once. A simple wireframe mockup is provided in Figure 2.

To further expand on the layout plan, the home page would contain 3 histograms that correspond to the available non-numeric variables in the dataset. The summary statistics will include the number of unique medications in the dataset and number of unique values for non-numeric values.

The data viewer page would contain a scatter plot depicting the review percentages given in the dataset. The points in the scatter plot would represent individual medications and would have filters to handle relationships between medications with non-numeric variables.

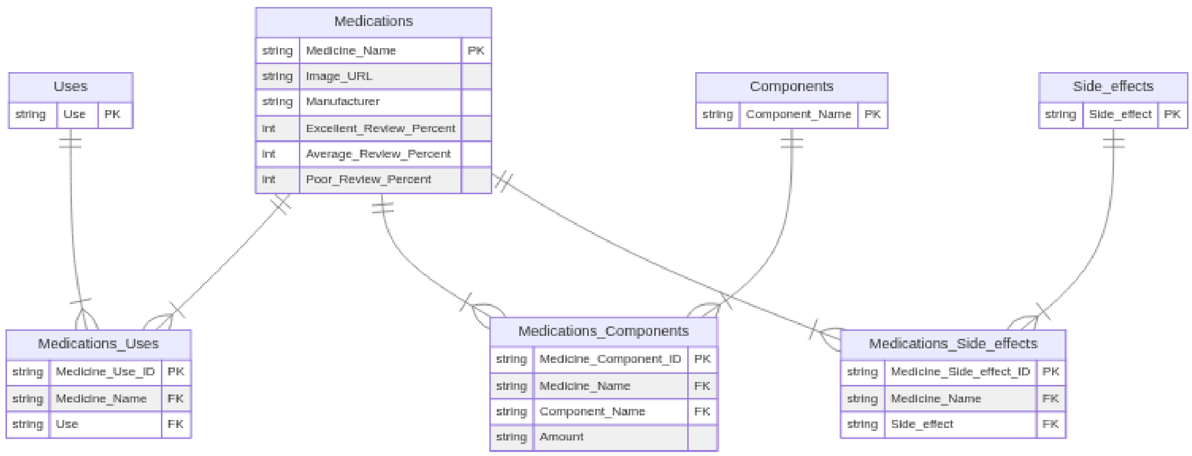

Figure 1: Entity relationship diagram showing architecture of generated SQLite database

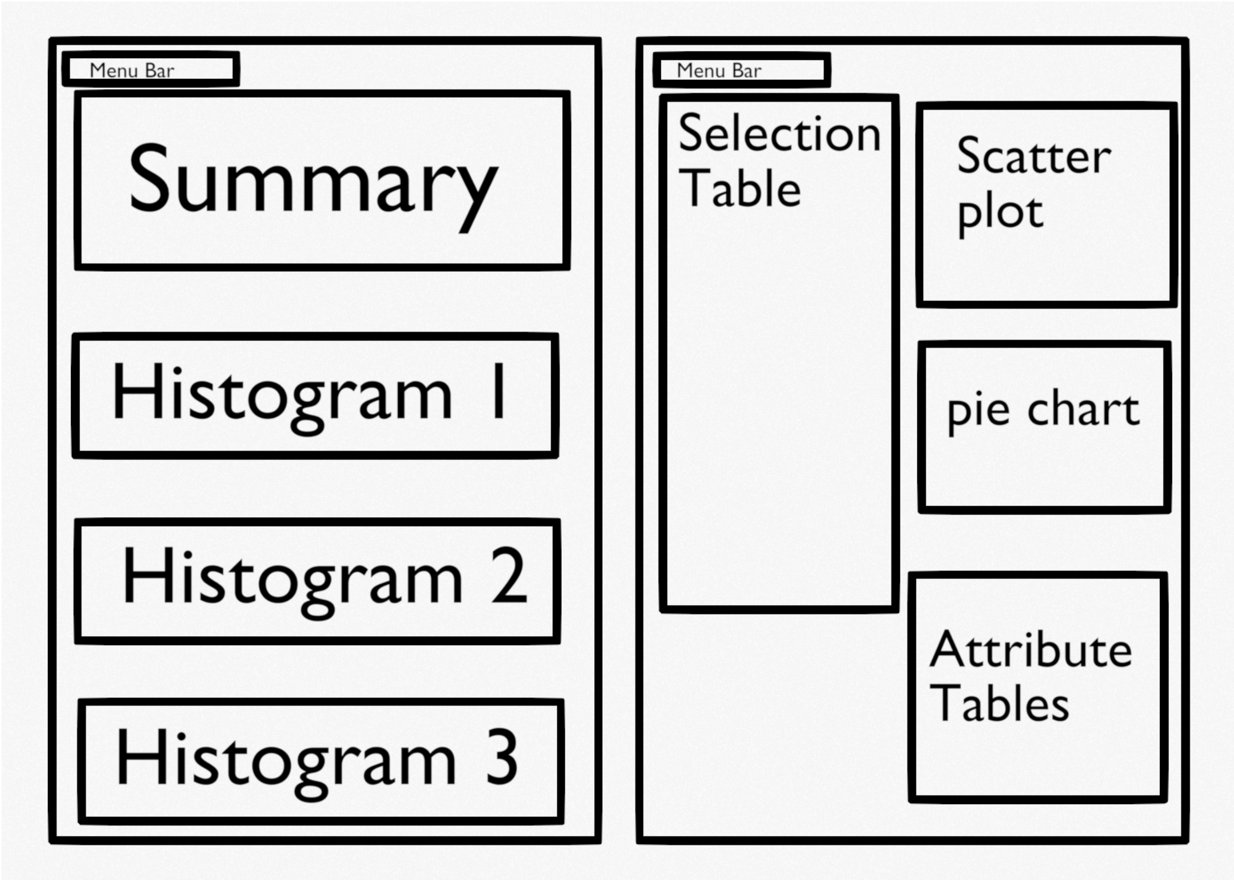

Figure 2: Mockup of shiny app design

3. Results

A live demo of the shiny app can be found on shinyapps.io at the following URL: https://illustratedman-code.shinyapps.io/project/. The source code for the project can be found on Github at the following URL: https://github.com/IllustratedMan-code/ph8093-database-project.

Figure 3 is the first section of the home page, showing the summary statistics for the dataset as well as a histogram of one of the non-numeric variables in the dataset. The histogram is a plotly figure, so it can be interacted with by dragging to select specific regions, zooming in, etc. Additionally, there is a link to the source of the data, which might be helpful for those who want to know more about the composition of these data.

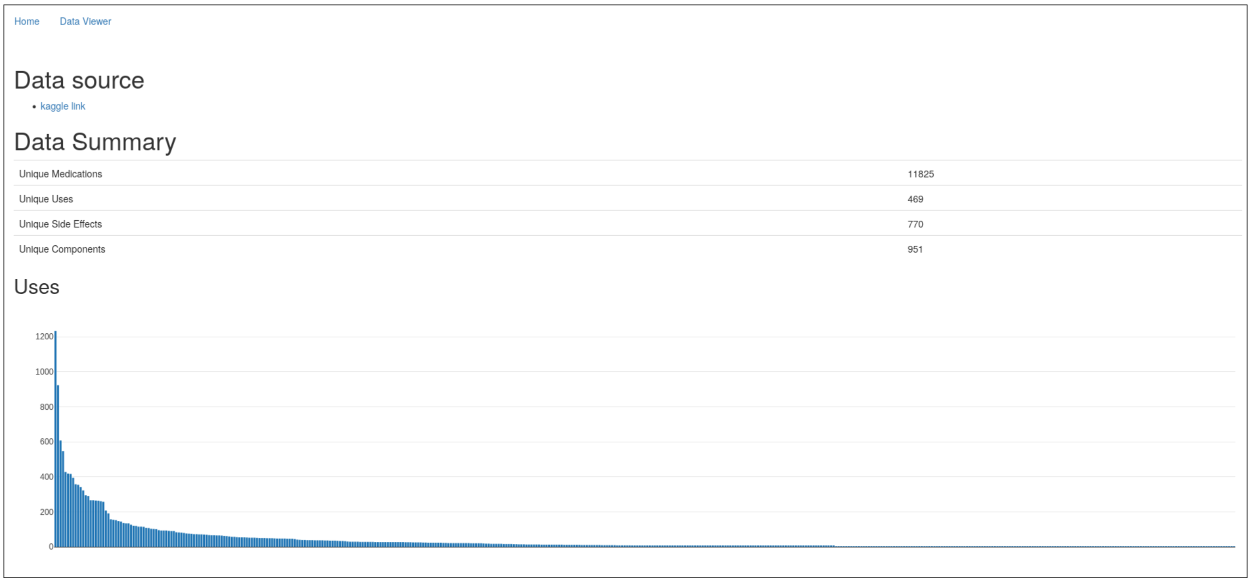

Figure 3: The first section of the home page

Figure 4 shows the second section of the home page, two more histograms showing the remaining non-numeric variables in the dataset. It should be noted that the medications components had the most diversity, with both the largest number of unique values, and lowest maximum value. Side effects, on the other hand, had the least diversity, with over 6000 medications that had nausea as a side effect.

The number of unique values within these non-numeric variables is significantly dwarfed by the number of unique medications. It stands to reason that there are likely medication that share similar non-numeric variables with each other, simply changing the name or composition slightly.

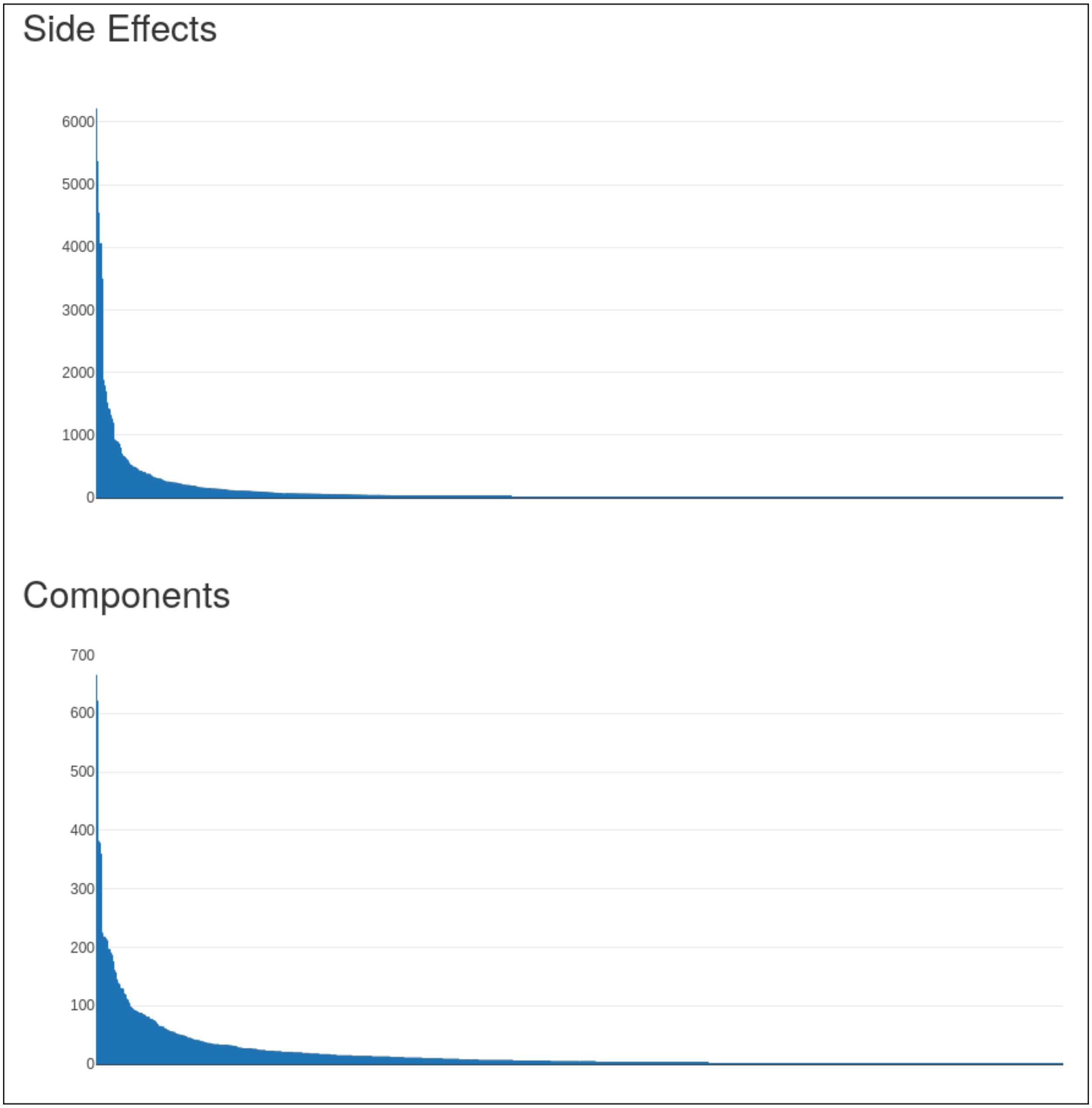

Figure 4: The second section of the home page

Figure 5 shows the first section of the data viewer page. The left side of the page is the table of medications. The table shows the Medicine name, a photo of the medicine, the manufacturer of the medicine, and the only numeric values in the dataset, the review percentages. The review ratings are split into three categories, a poor, average or excellent review.

On the right side of the page, there is a scatter plot that depicts each medication in a semi-3-dimensional way. The poor and average review percentages form the x and y axes respectively. The color depicts the excellent review percentage. Depicting the medications spatially allows one to compare a selected medication with other medication that are related in terms of review percentages. This novel visualization might allow a physician to compare related medications and select an alternative that may accomplish the same result while having better review percentages. To further accomplish this, there are filters that allow one to filter by similar composition, usage, and side effects. For a medication to show up in the filter, it must have at least one similar attribute for each checked box.

Additionally, there is a pie chart reiterating the review data listed in the table. Figure 6 shows a set of tables each showing which components, uses, and side effects each selected medications has.

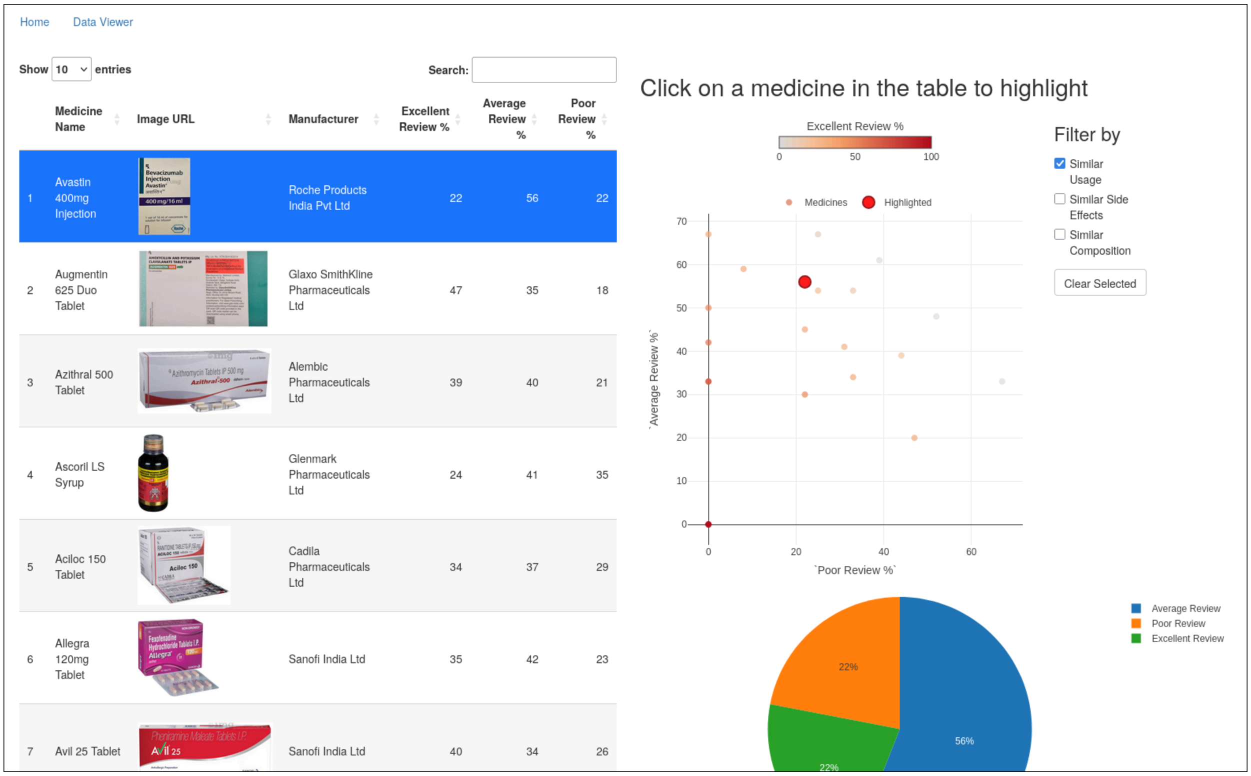

Figure 5: The first section of the data viewer page

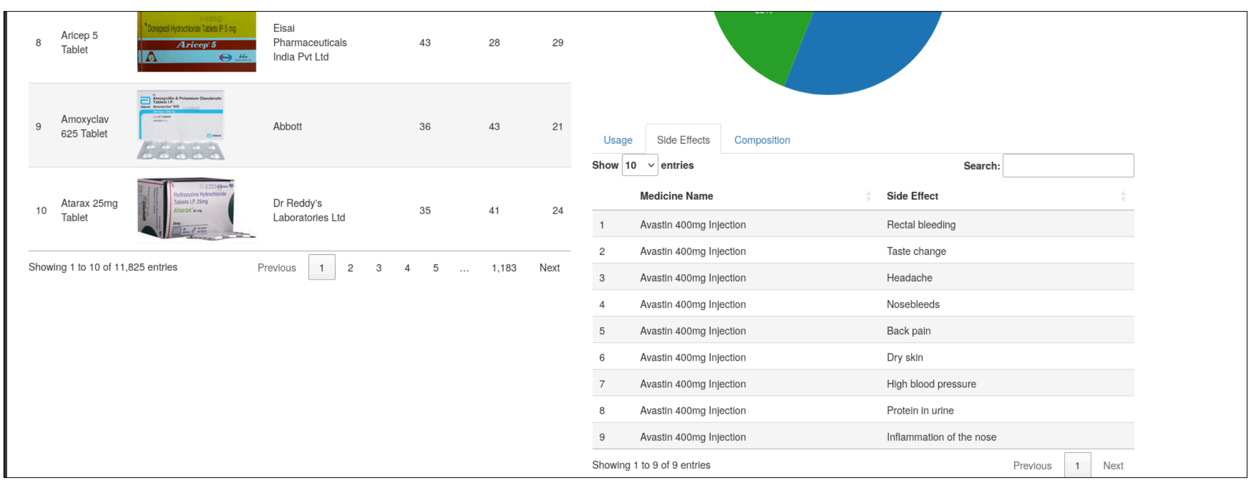

Figure 6: The second section of the data viewer page

4. Discussion and Analysis

This project explores visualizing a large list of medications spatially. Specifically, the scatter plot in Figure 5 provides a way to see the relationship between medications without a comparison chart. However, using only reviews to form this spatial relationship is limiting, it does not represent similar usage for example. A future direction for this type of visualization might be natural language based embeddings that provide a high dimensional spatial relationship between different medications. The dataset was quite limited when it came to different variables. The database that the dataset was scraped from also contains text descriptions of each medication, which could be used to create spatial embeddings.

While the dataset does contain a considerable amount of medications, more do exist[1], and more will be created every year. A more sophisticated visualization tool would actively query web databases to remain up to date.

Shiny is quite limited in flexibility and certain optimizations were difficult if not impossible to complete. For example, certain expensive operations will freeze the user interface in between inputs. While this can be somewhat solved by the futures package, fine tuned control of the execution stack can't be controlled from within R. Multiple invocations of a function in the current implementation will all execute even if the results are not required. So the user interface can still freeze under certain circumstances.

In future iterations of this project, shiny and R would be dropped in favor of a more powerful and flexible framework and language. For example, rust, which has advanced multithreading and asynchronous support, could be used by compiling to webassembly.

That being said, the spatial visualization provides an easy-to-use way to compare different medications without prior knowledge of said medications. Further optimizations would serve to make this visualizations more intuitive and useful.

5. Response to feedback

> “
>
> I recommend removing the medication images, … you may also add Therapeutic and Pharmaceutical classes as filters …
>
> ”
>
> — Jason

I'm not going to remove the images because one of the class comments actually preferred it with the images, so seems the feedback is split on this point. Additionally, I do not have any data for additional classifications of the medications. All variables are represented in the visualizations/tables.

> “… By adding proper sections and heading to each table, visualizations can be made more intuitive. …”
>
> — unknown classmate

I wasn't exactly sure what to write for each section, though I did add directions to get a user started in the data viewer. The home page already had section labels.

> "… could you filter the top X so that it's easier to have and see what's in each bar."
>
> — unknown classmate

This is already supported by plotly, by dragging over the desired region.

> "You have a very interesting and interactive graph that is very small. I would highlight that on the page (make it bigger?) and minimize the pie chart's placement on the page."
>
> — Scott

The plot is now larger and more prominent on the page.

> "This application is very useful. By adding proper sections and headings to each table, visualizations can be made more intuitive. Does your data have generic medication names as a column? It might be nice to include that as part of your table! As well, consider combining all three of your tables together in your visualizations."
>
> — unknown classmate

There are no generic medication names. All columns are represented in some way in the app. I did make the table into a tabbed interface, so there is less scrolling, but the tables are incompatible with each other, so I can't strictly combine them.

> "Cool app! I liked that you got images for these meds embedded into your table. The similarity plot/map is really cool. Maybe the tables that you have down in the bottom could be pie chart/bar charts too?"
>
> — unknown classmate

I'm not really sure how the tables could be represented as bar charts or pie charts as most of the time, the usage, composition, etc. are unique text names, not numeric values.

6. References

1. Statistics | DrugBank Online. Accessed December 2, 2025. https://go.drugbank.com/stats

2. Brysbaert M, Stevens M, Mandera P, Keuleers E. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Front Psychol.* 2016;7:1116. doi:10.3389/fpsyg.2016.01116[1]

3. Knox C, Wilson M, Klinger C, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research.* 2024;52(D1):D1265-D1275. doi:10.1093/nar/gkad976[2]

4. WebMD - Better information. Better health. WebMD. Accessed December 2, 2025. https://www.webmd.com/

5. RxList - The Internet Drug Index for prescription drug information, interactions, and side effects. RxList. Accessed December 2, 2025. https://www.rxlist.com/

6. Drugs Comparison Index: Find Drug vs. Drug, Side Effects, Uses, Interactions, More. RxList. Accessed December 2, 2025. https://www.rxlist.com/drugs_comparison/article.htm

7. 11000 Medicine details. Accessed December 2, 2025. https://www.kaggle.com/datasets/singhnavjot2062001/11000-medicine-details

8. team T pandas development. pandas-dev/pandas: Pandas. Published online September 30, 2025. doi:10.5281/zenodo.17229934[3]

9. Shiny. Shiny. Accessed December 2, 2025. https://shiny.posit.co/

[1] https://doi.org/10.3389/fpsyg.2016.01116
[2] https://doi.org/10.1093/nar/gkad976
[3] https://doi.org/10.5281/zenodo.17229934