CORD-19 Data Analysis Report

Introduction

This report presents a beginner-friendly analysis of the CORD-19 research dataset focusing on the metadata file. The goal was to load, clean, analyze, and visualize the data to extract meaningful insights about COVID-19 research publications.

Key Findings

- Most publications are concentrated between 2020 and 2021, reflecting the surge in research during the pandemic.
- Top journals publishing COVID-19 research include major medical and scientific publishers such as BMJ, The Lancet, and PLOS ONE.
- Word frequency analysis of paper titles highlights common terms like 'COVID-19', 'coronavirus', 'pandemic', and 'SARS-CoV-2'.
- The dataset includes a wide range of sources, showcasing the collaborative global response to the pandemic.

Reflection

Working on this assignment provided hands-on experience with the data science workflow: loading real-world datasets, handling missing data, and performing exploratory analysis. One of the main challenges was managing the large dataset size, which required working with samples to save space and speed up analysis. Another challenge was cleaning inconsistent date formats. Through this project, I gained practical skills in Pandas, Matplotlib, and Streamlit, and learned the importance of incremental debugging and focusing on core objectives when working with large and complex datasets.