# 📊 CORD-19 Dataset Analysis – Final Report

## 1. Data Overview

* Dataset size: 1056660 rows, 19 columns
* Key fields explored: title, abstract, publish\_time, journal, source\_x

## 2. Cleaning & Preparation

* Extracted year from publish\_time
* Added abstract\_word\_count
* Saved cleaned dataset as cleaned\_cord19.csv

## 3. Descriptive Statistics

* **Papers per Year:** Strong increase after 2020 due to COVID-19.
* **Top Journals:** bioRxiv, PLoS One, Sci Rep among most frequent.
* **Top Sources:** Medline, PMC, WHO collections.

## 4. Text Mining on Abstracts

* Preprocessed abstracts (lowercased, stopwords removed).
* Word frequency analysis highlighted key terms (e.g., “covid”, “patients”, “virus”).
* WordCloud showed major research themes visually.

## 5. Topic Modeling (LDA)

* Extracted thematic clusters from abstracts.
* Topics included:
  + Clinical research and patients
  + Public health and epidemiology
  + Virology and molecular studies
  + Vaccination and immunity
  + Global impact and policy

## 6. Key Insights

* **COVID-19 drove a massive publication spike starting 2020.**
* **Preprints (bioRxiv, medRxiv) were crucial early sources.**
* **Research covered clinical, public health, and biological aspects.**
* **WordCloud + LDA confirm diverse but COVID-focused themes.**

## ✅ Conclusion

The CORD-19 dataset provides an invaluable resource for tracking the scientific response to COVID-19. Our analysis combined data cleaning, exploratory analysis, and text mining to uncover trends and themes in global research output.