1.Data Understanding

This project uses multiple health and demographic datasets from the Demographic and Health Surveys (DHS) for South Africa, covering the years of 1998 and 2016. The datasets include variables such as:

- Maternal and child health indicators; [maternal mortality ratio, pregnancy-related deaths, acute respiratory infection prevalence]
- Fertility measures; [number of children born, general fertility rate]
- Breastfeeding and infant feeding practices;[breastfeeding, complementary feeding, median duration of breastfeeding]
- Literacy levels for women and men aged 15–49
- Socio-demographic characteristics; age groups, survey sample weights

These datasets aim to provide more knowledge into public health outcomes, healthcare access, and socio-demographic trends across different survey years.

2. Data Quality Assessment

Inspection of the following datasets involved checking for completeness, consistency, and correctness:

- **Missing values:** Certain variables, such as confidence intervals for maternal mortality or ARI indicators, contain missing entries. These missing values will be addressed during data preparation.
- **Duplicates:** The datasets contained minimal duplication, as each record corresponds to a unique combination of indicator, characteristic, and survey year.
- **Outliers:** Some numeric variables, such as maternal mortality ratios or extreme fertility counts, may represent true extremes or potential data entry errors. These will require further investigation.

3. Exploratory Data Analysis (EDA)

To better understand the datasets, preliminary analyses were performed using **R**. These include:

3.1 Summary statistics: Calculations of mean, median, minimum, maximum, and standard deviation for numeric variables such as maternal mortality ratios, breastfeeding percentages, and literacy rates. Distribution plots: Histograms and boxplots were produced to examine skewness and identify potential outliers.

3.2 Distribution plots: Histograms and boxplots were produced to examine skewness and identify potential outliers.

3.3 Correlation analysis: Heatmaps were used to explore relationships between variables, for example, maternal literacy and child health indicators.

3.4 Categorical variable visualizations: Bar plots were used to analyse distributions of variables such as age groups, breastfeeding categories, and literacy levels.

These visualizations helped identify key patterns and potential relationships in that; Higher literacy rates among women and men are generally associated with improved child health outcomes and vaccination coverage. Longer durations of breastfeeding correspond with lower rates of bottle feeding among infants. Regions and survey subgroups with higher prevalence of ARI show varied healthcare-seeking behaviour.

4. Hypotheses

4.1 Maternal and child health outcomes are significantly influenced by literacy and education levels.

4.2 Breastfeeding and complementary feeding practices affect child health and nutrition indicators.

4.3 Fertility rates and maternal health indicators vary with socio-demographic factors such as age, region, and access to healthcare.

4.4 Healthcare-seeking behaviour for ARI in children is correlated with caregiver literacy and regional health access.