Report on Health Survey Analysis

🧾 **Comprehensive Report on "Health Survey Analysis"**

---

1. Dataset Overview

The analysis is based on the dataset titled *district_health_survey_updated_2024.csv*.
This dataset compiles vital information on various health and social indicators across multiple Indian States and Union Territories. It serves as an essential source to study the interrelationship between education, sanitation, nutrition, early marriage, and women's welfare indicators.

The key columns used in the analysis are as follows:

- **State/UT** represents the name of the Indian State or Union Territory.
- **Female population age 6 years and above who ever attended school (%)**, renamed as *School_Attendance*, measures the percentage of females aged six and above who have attended school at least once.
- **Population living in households that use an improved sanitation facility (%)**, renamed as *Improved_Sanitation*, reflects the percentage of the population with access to improved sanitation systems.
- **Households using clean fuel for cooking (%)**, renamed as *Clean_Fuel*, shows the proportion of households that rely on clean and sustainable cooking fuels such as LPG or electricity.
- **Women age 20–24 years married before age 18 years (%)**, renamed as *Early_Marriage*, indicates the percentage of women who were married before reaching the legal age of 18.
- **Current use of family planning methods – any modern method (%)**, renamed as *Modern_FP*, measures the adoption of modern contraceptive or family planning methods among currently married women aged 15–49.
- **Children under 5 years who are stunted (height-for-age) (%)**, renamed as *Stunting*, highlights the percentage of children under five whose growth is stunted due to malnutrition or poor health conditions.

These columns collectively capture multiple dimensions of health, education, sanitation, and social development, offering a holistic view of state-level performance.

---

2. Data Loading and Cleaning Operations

The project starts by initializing a SparkSession to perform large-scale data processing using the PySpark framework.
The dataset is imported using the spark.read.csv() function, which automatically detects the column types and reads data efficiently with the header option enabled.

Next, only the essential columns are selected from the dataset to maintain focus and reduce redundancy. The column names are simplified for better readability and consistency across subsequent analyses.

A crucial step involves cleaning and transforming the data using a custom function named clean_and_cast_column().
This function performs several actions:

1. **Removes unwanted characters** like parentheses or special symbols.
2. **Replaces invalid entries** (for example, "*") with None or null values.
3. **Casts text-based numerical data** into float values to ensure compatibility with mathematical operations.

Missing or invalid data are handled gracefully to avoid skewing analytical results.
This ensures that the final dataset is clean, standardized, and suitable for computation and visualization.

---

3. Data Analysis Operations

After the data cleaning stage, the next phase involves performing statistical analysis to uncover trends and relationships.
The cleaned dataset is grouped by *State/UT*, and averages are calculated for all key indicators.
These aggregated statistics are computed using Spark's groupBy() and agg() functions, which efficiently handle large datasets.

Once aggregation is completed, the Spark DataFrame is converted into a Pandas DataFrame for easier manipulation and visualization.
Descriptive statistics, such as means and correlations, are computed to identify patterns and dependencies among variables like school attendance, sanitation, and child nutrition.

The correlation matrix highlights how improvements in education and sanitation are associated with reductions in early marriage and child stunting rates.

---

4. Data Visualization

Visualization plays a key role in interpreting the results.
The project employs **Matplotlib** and **Seaborn** libraries to present data insights visually.
Several types of plots are generated:

- **Bar charts** are used to compare indicators across states.
- **Heatmaps** visually represent the correlation strengths between multiple indicators.
- **Pair plots** allow for multi-dimensional observation of relationships among health and social indicators.

For example, the Seaborn pair plot in the notebook demonstrates how *School_Attendance*, *Clean_Fuel*, and *Stunting* interact with one another, revealing that higher education and better sanitation typically correlate with lower stunting rates.

These plots make complex relationships easily understandable and highlight outlier states that perform significantly better or worse in certain aspects.

---

5. Insights Derived

The analysis produces several meaningful insights:

1. **Education is a strong positive influencer.**
   States with higher female school attendance rates generally demonstrate better sanitation access and greater use of clean fuels.
2. **Early marriage correlates with poor social outcomes.**
   States with a higher percentage of early marriages tend to show lower access to sanitation and clean fuel facilities, reflecting social inequality and gender imbalance.
3. **Nutrition outcomes are tied to education and awareness.**
   The stunting rate among children under five decreases significantly in regions where women's education levels and family planning awareness are higher.
4. **Sanitation and clean fuel access are interlinked.**
   Improved sanitation and clean cooking fuel usage often coexist, indicating better overall living standards in those states.

Overall, these findings emphasize the importance of education and empowerment in improving health and social conditions.

---

6. Tools and Libraries Used

The analysis is powered by multiple tools and libraries.
PySpark is used for data extraction, cleaning, and distributed computation.
Pandas handles data manipulation once the dataset is smaller and easier to process locally.
Matplotlib and **Seaborn** are responsible for generating static and aesthetic visualizations.
NumPy supports numerical operations and ensures smooth handling of arrays and mathematical transformations.

The combination of these tools enables efficient, scalable, and visually clear analysis.

---

7. Final Output

The final output of the notebook includes a clean, standardized dataset with renamed columns, computed state-wise averages for all indicators, and multiple visual representations.
These visualizations highlight relationships, patterns, and dependencies among key variables, helping policymakers, researchers, or analysts identify areas needing attention.

The notebook concludes with comprehensive graphical outputs, statistical results, and summarized insights that together form a clear narrative of India's district-level health and development patterns.

---

8. Limitations of the Analysis

While the analysis provides valuable insights, several limitations must be acknowledged:

1. **Data Availability and Completeness:**
   The dataset may have missing or incomplete entries for some states or indicators, which can affect the reliability of averages and correlations.

2. **Temporal Limitation:**
   The dataset appears to represent a single year or a limited time frame (2024), and hence does not capture long-term trends or improvements over time.

3. **Data Source Quality:**
   The accuracy of findings depends on the quality and consistency of the data collected from official health surveys. Errors or biases in data collection could propagate into the analysis.

4. **Causation vs Correlation:**
   The analysis identifies statistical relationships but does not prove causation. For example, higher school attendance correlates with better sanitation but may not directly cause it.

5. **Geographic and Socioeconomic Diversity:**
   India's states differ widely in culture, economy, and geography. Aggregated statistics might mask local variations and disparities within individual states or districts.

6. **Static Visualization:**
   The analysis uses static plots, which, while effective for explanation, may not offer interactive exploration that dynamic dashboards could provide (e.g., Power BI or Tableau).

7. **Computational Constraints:**
   Though PySpark handles large data efficiently, processing time and memory constraints may limit the scale or complexity of analysis, especially on local systems.

---

9. Conclusion

In summary, this health survey analysis successfully integrates PySpark's processing power with Python's visualization capabilities to derive meaningful insights from a large dataset. Despite some limitations related to data completeness and interpretation, the project demonstrates a strong analytical framework for understanding key social and health indicators in India.
It offers a foundation for policymakers, data analysts, and researchers to build upon with more granular data and time-based trends in future studies.