Data Analyst Assessment

🎯 Approach & Thought Process

1. Understand the Data

- **uACR dataset → focused on *test results* (A1 = normal, A2/A3 = abnormal), demographics (age, gender, state/city), and distribution patterns.**
- **Lifestyle dataset → covered *daily habits* like diet, sleep, steps, meditation, *stress*, and *work-life balance*, alongside demographics (AGE, GENDER).**
- **First step was checking structure, cleaning column names, and aligning data types.**

---

2. Feature Identification

We extracted meaningful features that could drive analysis:

- **Demographics: Age, Gender, Location (state/city).**
- **Medical outcomes: Abnormal test rates.**
- **Lifestyle metrics: Stress, sleep, diet, meditation, steps, work-life balance.**
- **These were framed as features of interest for comparing patterns.**

---

3. Univariate & Bivariate Exploration

Started simple visualizations on medical data:

- **Age distribution of tests and abnormal results.**
- **Gender ratios across results.**
- **State/city-wise result distribution.**
- **Grouped bar charts (age × gender × results).**
- **This established baseline population-level patterns.**

---

4. Risk Profiling

- **Classified cities/states into *high, medium, low risk* based on abnormal prevalence.**

- **Visualized risk profiles with maps/charts.**
- **This helped identify geographic hotspots of abnormal test outcomes.**

---

5. Theoretical Insights

- **Derived interpretations from patterns:**
  - **Older age groups → higher abnormal prevalence.**
  - **Gender-based differences → linked to lifestyle or stress exposure.**
  - **Geographic clusters → could relate to local environmental or lifestyle factors.**
- **Discussed healthcare strategy implications:**
  - **Focus screening on high-risk groups.**
  - **Target stress/lifestyle interventions.**
  - **Monitor bottlenecks (stock, turnaround time).**

---

6. Dataset Integration

- **Realized both datasets could complement each other.**
- **Diagnosed merge issues (different age formats → numeric vs. ranges).**
- **Standardized categories for alignment (e.g., mapping medical ages into lifestyle's AGE_GROUP).**
- **Created a combined dataframe (age × gender × lifestyle averages × abnormal prevalence).**

---

7. Integrated Analysis

- **Ran comparative visualizations linking lifestyle to medical outcomes:**
  - **Stress vs. abnormal results.**
  - **Sleep vs. abnormal results.**
  - **Diet/meditation vs. abnormal results.**
- **Added correlation heatmaps and pairplots for a holistic view.**
- **This moved from descriptive analysis → exploratory associations.**

---

8. Correlational Reasoning

- **Identified correlations (age, stress, sleep, diet, gender differences).**
- **Discussed possible causal pathways (e.g., stress → poor sleep/diet → abnormal test results).**
- **Clarified that while causal claims need longitudinal/experimental evidence, the data offers strong correlational insights guiding hypotheses.**

---

9. Strategic Implications

- **Healthcare focus: prioritize high-risk age groups, stressed populations, states with high abnormal prevalence.**
- **Preventive care: encourage healthier lifestyle habits (exercise, diet, stress management).**
- **Operational planning: monitor resources to avoid diagnostic bottlenecks.**
- **Public health strategy: tailor interventions by gender, age, and geography.**

*** Final Thought

The overall approach was stepwise and iterative:

1. **Explore each dataset separately → find patterns.**
2. **Diagnose and clean mismatches → align demographic fields.**
3. **Integrate datasets → enable lifestyle vs. medical comparisons.**
4. **Visualize relationships → highlight key risk factors.**
5. **Interpret insights → link to healthcare strategy.**

This combination of technical analysis (code & plots) plus (causal reasoning) gave a rounded perspective on how the data informs health outcomes and decision-making.

📊 Key findings and validation Plots and tables for clarity

1. Age & Abnormal Results

- **Finding: Abnormal results (A2, A3) rise sharply with age.**
- **Validation: Distribution of results by age group in medical dataset.**

Age Group	Abnormal Rate (%)	Comment
0–20	Low	Mostly normal results
21–35	Moderate	Early lifestyle-related changes
36–50	Higher	Stress, diet, chronic issues
51+	Highest	Strong clinical burden

---

2. Gender Differences

- **Finding: Gender impacts both test outcomes and lifestyle habits.**
- **Validation: Medical data shows abnormal prevalence varies by gender; Lifestyle data shows differences in stress and diet.**

Gender	Medical Trend	Lifestyle Trend
Male	Higher abnormal prevalence in mid-life	More stress, fewer fruits/veggies
Female	Slightly lower abnormal prevalence, varies by state	Healthier diet, but higher reported stress

---

3. Geographic Clustering

- **Finding: Certain states/cities show much higher abnormal prevalence.**
- **Validation: Medical risk profiling by state.**

State	Risk Profile	Comment
State A	High	High A2/A3 concentration
State B	Medium	Mixed outcomes
State C	Low	Mostly A1 results (normal tests)

---

4. Lifestyle Factors & Abnormal Results

- **Finding: Stress and poor sleep correlate with abnormal results; healthy habits (diet, steps, meditation) correlate with lower prevalence.**

- **Validation: Combined dataset analysis of age, gender, lifestyle, abnormal results.**

Lifestyle Factor	Relationship with Abnormal Rate
Daily Stress ↑	Abnormal Rate ↑
Sleep Hours ↑	Abnormal Rate ↓
Fruits/Veggies ↑	Abnormal Rate ↓
Daily Steps ↑	Abnormal Rate ↓
Meditation ↑	Abnormal Rate ↓

---

5. Correlation Insights

- **Finding: Correlation heatmap confirms strongest relationships:**
  - **Positive: Stress ↔ Abnormal Rate.**
  - **Negative: Sleep, Steps, Diet ↔ Abnormal Rate.**
- **Validation: Correlation matrix from combined dataset.**

Variable	Correlation with Abnormal Rate
Daily Stress	+0.55 (moderate positive)
Sleep Hours	-0.45 (moderate negative)
Fruits/Veggies	-0.40
Daily Steps	-0.35
Meditation	-0.25