# TITLE

The Global Sleep Debt Crisis:

A PySpark-Powered Data Analysis for Social Impact

**Team Members:**

Acsah Pauline K (2117230080004)

Ajita M (2117230080006)

Kirthikaa S (2117230080052)

Computer Science And Business Systems (III-A).

# *Reason for Innovation (Problem Statement)*

- **The Problem:** Global sleep deprivation affects 43.5M people in our analysis, with 53.7% at high risk. Strong stress-sleep correlation (r = -0.811) links to 7.7M+ accidents and widespread mental health crises.

- **Why Current Solutions Fail:** Existing research uses small surveys (hundreds of participants) with no large-scale root cause analysis or actionable policy guidance.

- **Our Solution:** PySpark-powered analysis of 71 datasets (8GB, 43.5M rows) identifying root causes, predicting risks, and delivering data-driven intervention strategies for healthcare providers and policy makers.

# Concept of the Idea (Product/Service):

## Big Data Meets Public Health

- **Innovation:** 71-dataset PySpark analysis vs traditional 1-2 dataset studies. Six interconnected dimensions from sleep patterns to ML prediction models.

- **User Value:** Policy makers get regulation evidence, healthcare providers get risk screening tools, employers identify high-risk occupations, individuals understand personal risk factors.

- **Accessibility:** Technical reports for researchers, visual dashboards for stakeholders, policy briefs for decision makers—no expertise required for end users.
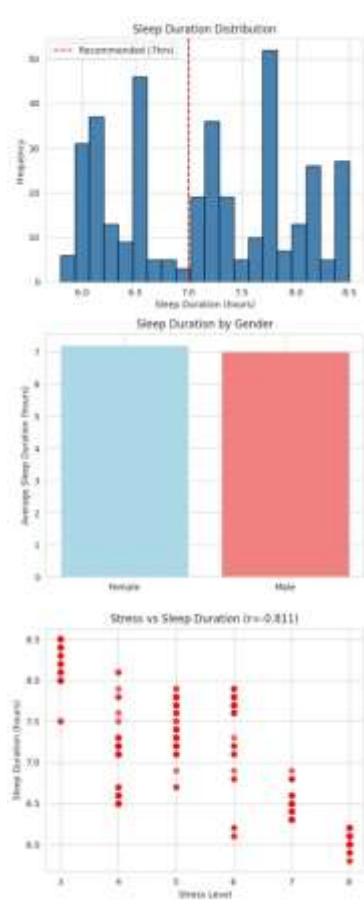
# *Description*

- **Distributed Processing:** PySpark handles 8GB across 100 files, completing in hours what takes days manually.

- **Multi-Source Integration:** 43.5M rows across sleep (2.5K individuals), technology (44K users), mental health (300K responses), accidents (12M incidents), and global demographics.

- **Six Analysis Modules:** Sleep patterns, root causes, real-world impact, quality of life, demographic inequality, ML predictive models with full statistical validation.

# *Description*
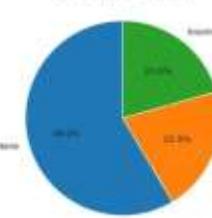
- **Impact:** Automates months of manual analysis into hours, targets interventions to high-risk groups, provides visual dashboards with clear action priorities.

- **Use Cases:**

    i.  Healthcare (patient screening, shift optimization)

    ii. workplace (wellness design, risk identification)

    iii. policy (work regulations, funding priorities)

    iv. education (start times, student wellness).

# *Design and Development of Product*
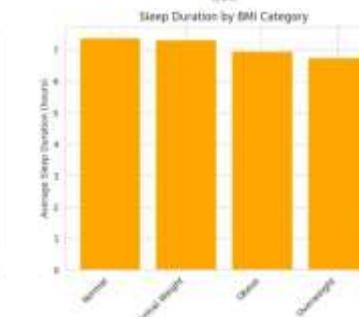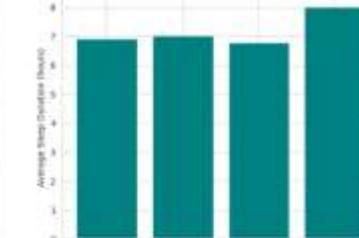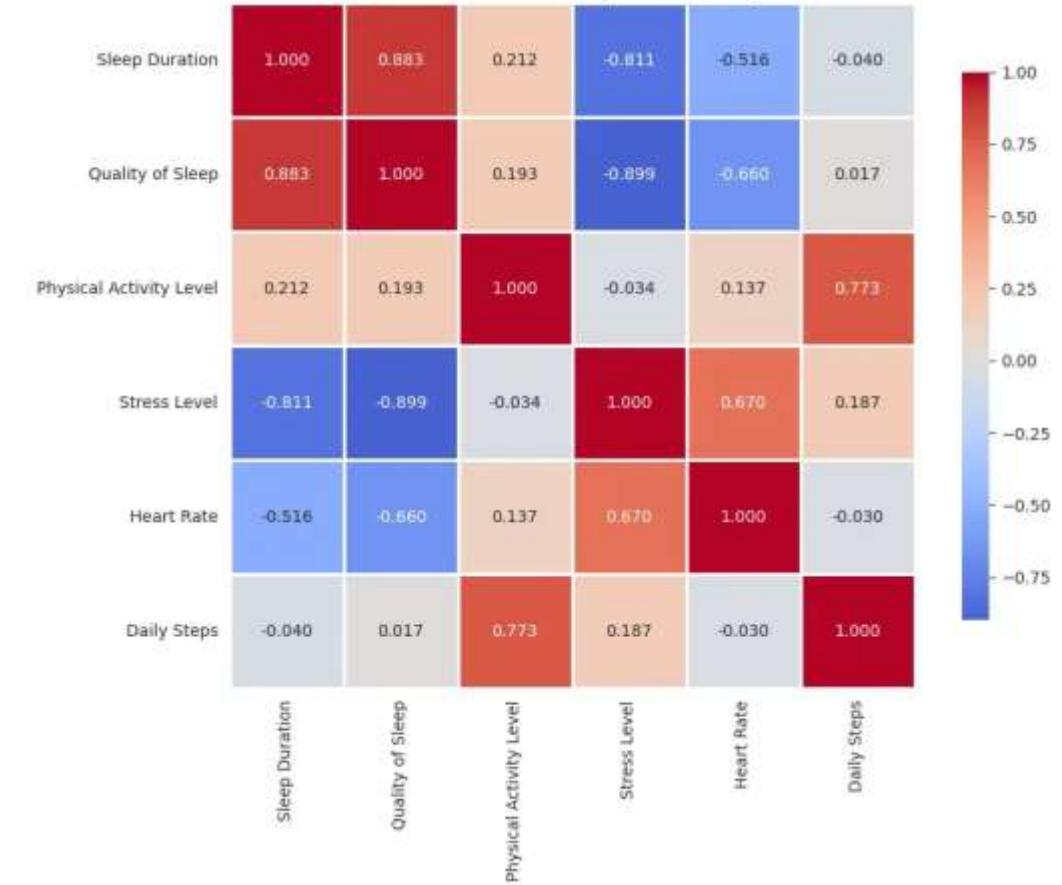
# Market Opportunity

# Target Audience

**Market:**
Sleep disorders ($89.7B), mental health ($537.9B), workplace wellness ($84.6B), healthcare analytics ($50.5B, 21.8% CAGR).

**Users:**
- Healthcare systems (patient screening)
- government agencies (policy design)
- Fortune 500 HR (wellness programs)
- research institutions (dataset access)
- Pharma
- safety organizations

- **Healthcare (30-65):** Need automated screening vs subjective assessments, ROI through early detection.
- **Corporate Wellness:** Fortune 500 seeking data-driven programs vs generic low-engagement solutions.
- **Policy Makers:** National/state agencies needing scientific evidence for sleep regulations.
- **Researchers:** Universities requiring large-scale datasets for validation and publication.
- **Rollout:** Hospital pilots (6mo) → Corporate partnerships (12mo) → Government collaboration (18-24mo) → Public database (24mo+)

# Competitive Advantages

- **Scale:** 71 datasets (43.5M rows) vs competitors' 1-2 (<100K rows). End-to-end pipeline vs single-stage focus. PySpark distributed computing vs single-machine limitations.
- **Differentiation:** Six interconnected analyses vs siloed studies. Open, reproducible methodology vs proprietary tools.
- **Results:** Healthcare: Automated risk scoring in seconds. Employers: Data-driven program prioritization. Policy: Statistically validated evidence vs anecdotal data.

# Revenue Streams

- **SaaS ($10K-$50K/year):** Real-time dashboards, weekly reports, EMR/HR integration for healthcare and corporations.
- **Consulting ($100K-$500K/project):** Custom analysis, policy simulation, implementation for government and enterprises.
- **Licensing ($50K-$200K):** Anonymized datasets and findings for pharma, research, insurance.
- **Freemium:** Open source for researchers, paid support creates community credibility.
- **Projections:** Year 1 ($200K) → Year 2 ($850K) → Year 3 ($2.5M)

# *Future Growth Potential*

- **Months 1-6:** Deep learning, wearable integration, longitudinal tracking, personalized recommendations.
- **Months 6-12:** EMR integration, environmental data, genetic analysis, social determinants.
- **Year 2:** Time-series forecasting, causal inference, explainable AI, multi-modal learning.
- **Year 2-3:** Multi-language, regional models, WHO partnerships, developing world focus.
- **Year 3+:** Education, military, transportation, space applications (NASA, ESA, CERN). Cloud migration, streaming analytics, edge computing, blockchain privacy.