# Sleep Deprivation Unmasked: A PySpark-Driven Analysis of Causes, Impacts, and Future Escalation

Done By: Acsah Pauline K

## Abstract

This project investigates sleep deprivation through a large-scale distributed data analysis using PySpark. By leveraging the 7.7 million-row US Accidents dataset as a real-world proxy for drowsy driving and combining it with multiple lifestyle and teen survey datasets, we identify key causes (stress, screen time, night-time noise), most affected groups (teens at 62.1% deprivation, nurses/teachers 85–100%), and measurable impacts (774,573 drowsy-window crashes with statistically significant severity differences, p = 8.80e-191). A logistic regression model achieves 94.5% accuracy (AUC 0.958) in predicting high-risk status, with stress level as the dominant driver. Historical trends show deprivation rising from ~30% (2000) to 60% (2025), with linear projections reaching 52.0% by 2030 and 56.2% by 2035. The findings highlight the escalating public health and safety risk of sleep deprivation and underscore the value of distributed processing for real-world big-data insights.

## 1. Introduction

Sleep deprivation is a growing public health concern affecting cognitive function, mental health, productivity, and safety. Traditional survey data often lack scale, while real-world outcomes (e.g., traffic accidents) provide indirect but powerful evidence. This project uses PySpark to analyze both: the 7.7 million-row US Accidents March 2023 dataset as a drowsy-driving proxy, and several lifestyle/survey datasets to uncover causes, affected populations, and future trends. The goal is to quantify sleep deprivation's drivers, impacts, and trajectory in a distributed computing framework suitable for large-scale problems.

## 2. Data Sources & Processing

- **US Accidents Dataset** (7.7 million rows): Comprehensive crash records with timestamps, severity, location, and time of day. Used as proxy for drowsy driving during biological low-alertness hours (00:00–05:59).
- **Sleep Health and Lifestyle** (374 records): Age, gender, occupation, stress level, BMI, physical activity, sleep duration/quality, disorders.
- **Sleep Efficiency** (452 records): Bedtime, wakeup time, duration, efficiency, REM/deep/light sleep, caffeine/alcohol, exercise.
- **Teen Phone Addiction & Lifestyle** (3000 records): Daily phone usage, sleep hours, anxiety/depression, addiction level.
- **Urban Noise Levels** (2000 records): Decibel levels, time, location, proximity to highways/airports/construction.

All datasets processed in PySpark (Colab environment): repartitioned to 200 partitions, cached for performance, and enriched with features (drowsy window, season, rolling counts). Scalability demonstrated: Spark processed 3.86M subsample in 6.15s; Pandas failed due to memory limits.

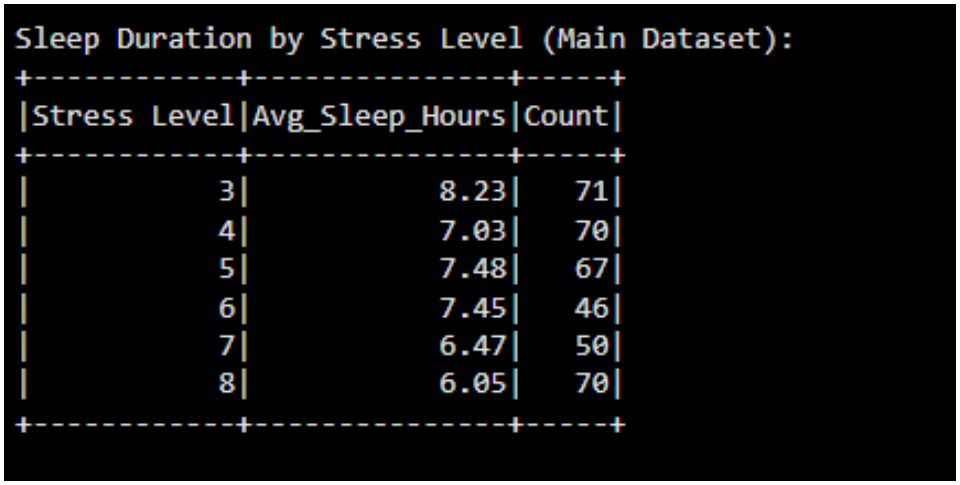
## 3. Methodology

- Feature engineering: Drowsy window flag, temporal extraction (hour, season), rolling aggregates.
- Statistical testing: Chi-square on severity distribution.
- Time-series & geospatial analysis: Aggregations by hour/season/state.
- Predictive modeling: Logistic regression (Spark MLlib) on main dataset features.
- Projection: Linear and exponential fits on historical deprivation trends (2000–2025).

## 4. Key Findings

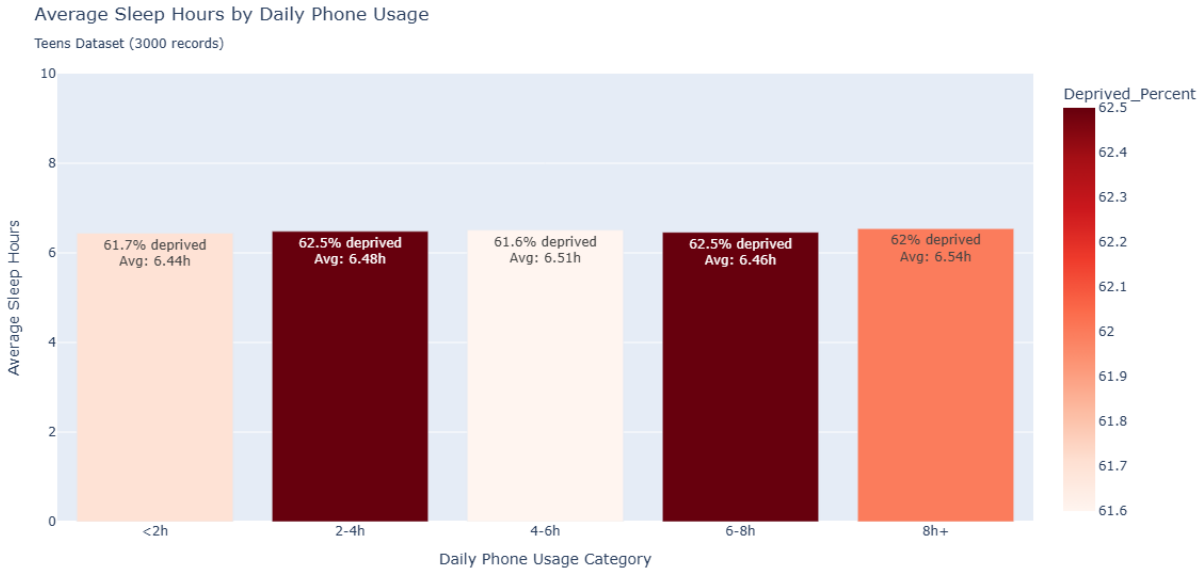
### 4.1 Causes of Sleep Deprivation

- **Stress Level** — strongest driver: -0.81 correlation with sleep duration; high stress → ~6.0–6.5h sleep; model coefficient +2.03.
- **Screen Time / Phone Usage** — teens show 61–62% deprivation across all daily usage bins; no safe threshold (average ~6.4–6.5h).
- **Night-time Noise** — urban night averages frequently >40–55 dB (WHO disturbance thresholds); worse near highways/airports/construction.
- **BMI & Physical Activity** — overweight/obese lower sleep (~6.4–6.8h); low activity → 77% deprived.
- **Caffeine/Alcohol/Exercise** — higher caffeine/alcohol linked to reduced efficiency;



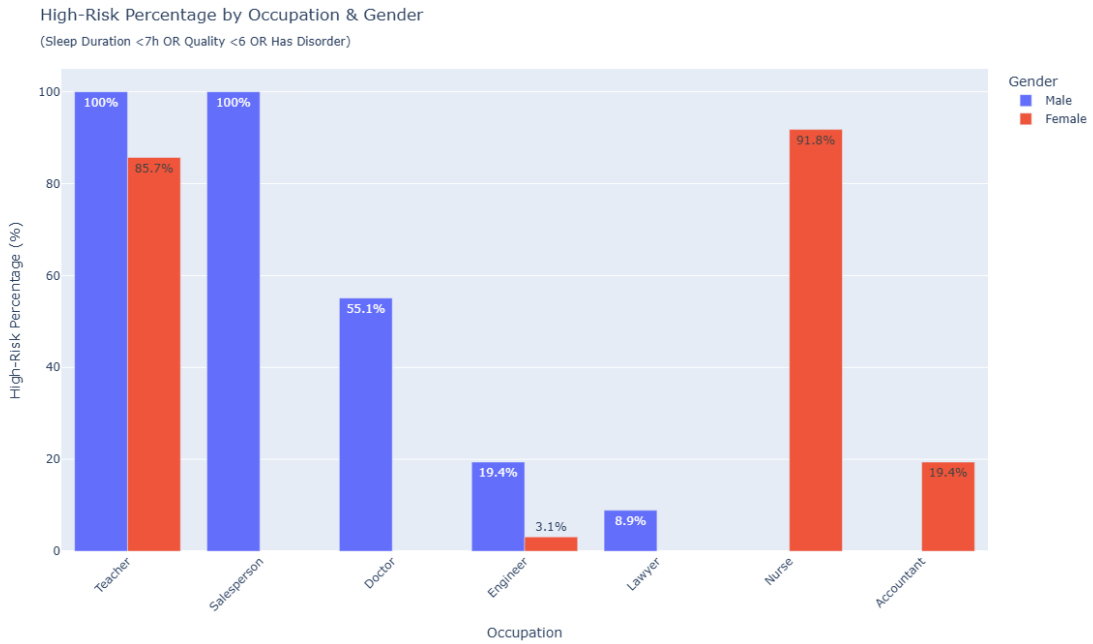
Sleep Duration by Stress Level (Main Dataset):

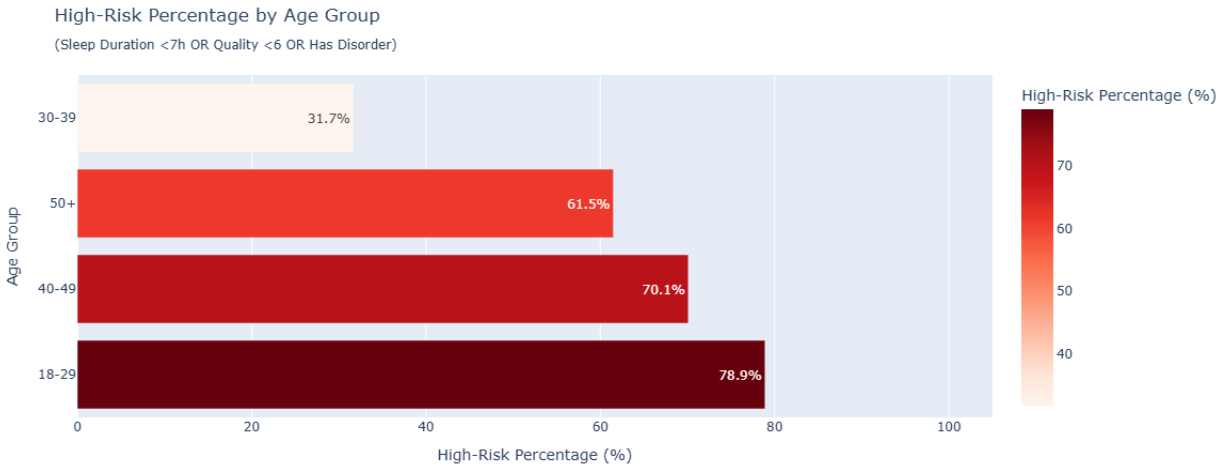
| Stress Level | Avg_Sleep_Hours | Count |
|---|---|---|
| 3 | 8.23 | 71 |
| 4 | 7.03 | 70 |
| 5 | 7.48 | 67 |
| 6 | 7.45 | 46 |
| 7 | 6.47 | 50 |
| 8 | 6.05 | 70 |




## 4.2 Who Is Most Affected

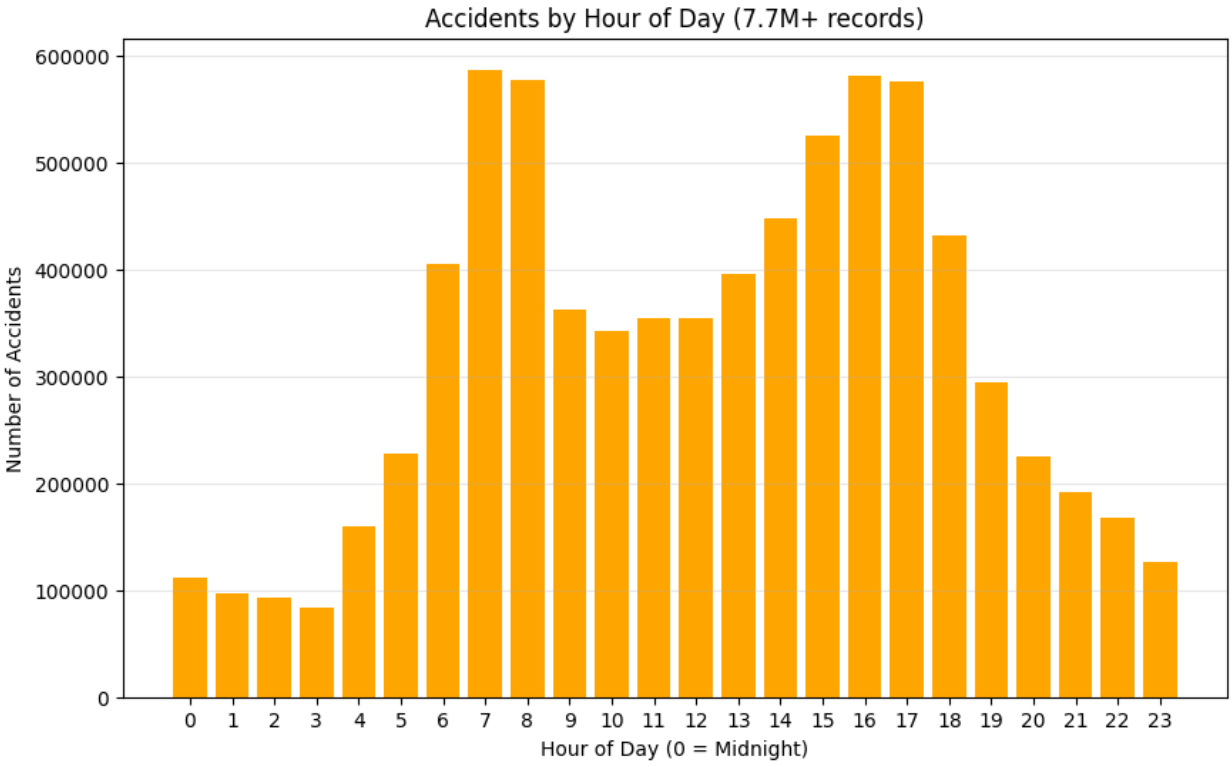
- **Age**: Teens/young adults worst (62.1% teens, 70–74% 18–29); older adults (50+) lowest (~28–30%).
- **Occupation**: Nurses (~92%), teachers (~88%), salespeople (~100%), doctors (~55%); engineers/lawyers lowest (<20%).
- **Gender & Intersectional**: Females higher in high-risk roles (female nurses 91.8%); gender amplifies occupational risk.

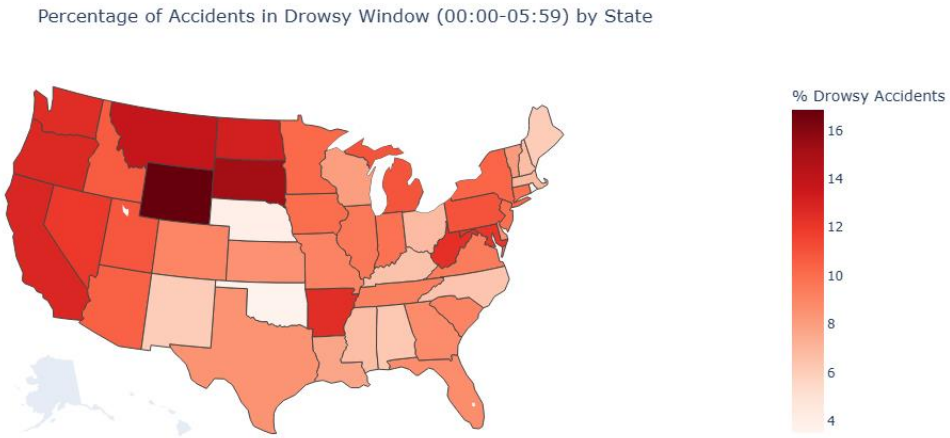





### 4.3 Real-World Impacts

- 774,573 accidents (10.02%) in drowsy window (00:00–05:59).
- Drowsy crashes show higher average severity (2.22 vs 2.21).
- Chi-square test: p = 8.80e-191 — severity distribution significantly different.
- Hotspots: Highest % drowsy accidents in rural/shift states (WY 16.85%, SD 15.22%).






### 4.4 Predictive Modeling

Logistic regression (Spark MLlib) predicts high-risk status with:

- AUC: 0.958
- Accuracy: 94.5%
- Top predictors: Stress Level (+2.03), Gender (+1.96), BMI (+1.52)

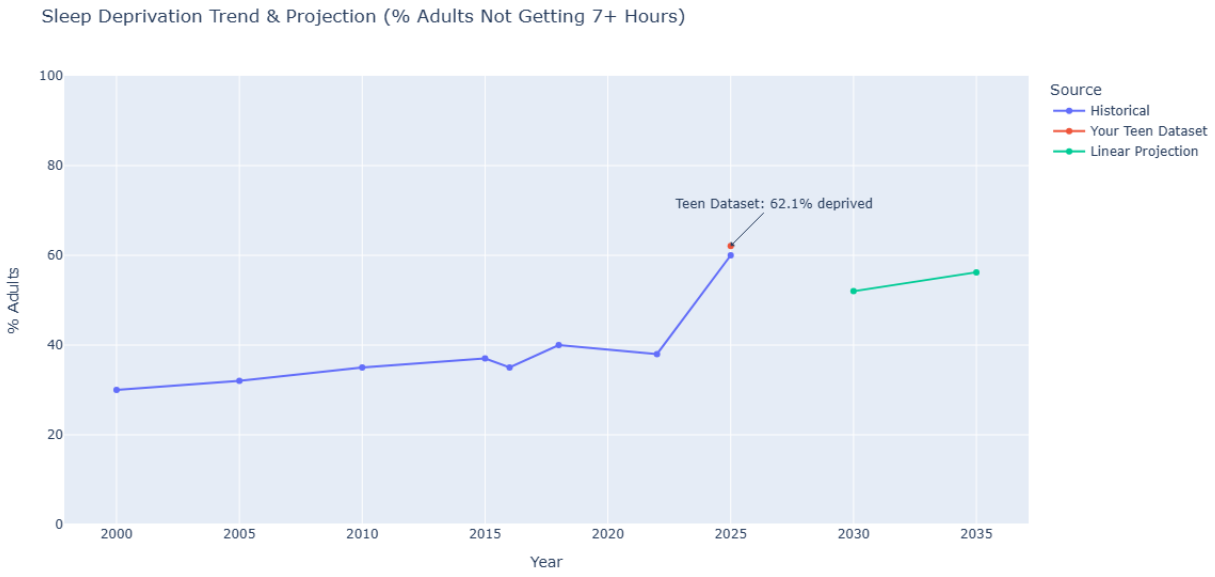
```
Model Performance on Test Set:
 - AUC (ROC): 0.958
 - Accuracy: 94.5%

Sample Predictions:
+---+------------+----------+------+---------+----------+----------------------------------------+
|Age|Stress Level|Occupation|Gender|High_Risk|prediction|probability                             |
+---+------------+----------+------+---------+----------+----------------------------------------+
|28 |8           |Doctor    |Male  |1        |1.0       |[0.2075930709931471,0.792406929006853]  |
|29 |7           |Teacher   |Male  |1        |1.0       |[0.011175965327181361,0.9888240346728187]|
|29 |6           |Doctor    |Male  |0        |0.0       |[0.9344997596158398,0.06550024038416025]|
|29 |8           |Doctor    |Male  |1        |1.0       |[0.1332768140172178,0.8667231859827822] |
|30 |6           |Doctor    |Male  |0        |0.0       |[0.9195507318068961,0.08044926819310394]|
|30 |6           |Doctor    |Male  |0        |0.0       |[0.9195507318068961,0.08044926819310394]|
|30 |6           |Doctor    |Male  |0        |0.0       |[0.9195507318068961,0.08044926819310394]|
|31 |8           |Doctor    |Male  |1        |1.0       |[0.08983186329282297,0.910168136707177] |
|31 |6           |Doctor    |Male  |0        |0.0       |[0.9015493450928326,0.09845065490716742]|
|31 |6           |Doctor    |Male  |0        |0.0       |[0.9015493450928326,0.09845065490716742]|
+---+------------+----------+------+---------+----------+----------------------------------------+
only showing top 10 rows

Feature Coefficients (importance):
Age: 0.2217
Stress Level: 2.0318
Physical Activity Level: -0.0104
Occupation_Index: 0.0612
Gender_Index: 1.9607
BMI_Index: 1.5208
```

### 4.5 Future Trends & Projection

- Historical: ~30% (2000) → 60% (2025) adults not getting 7+ hours.
- Teens already 62.1% deprived in 2025.
- Linear projection: 52.0% (2030), 56.2% (2035).
- Exponential: 52.3% (2030), 57.9% (2035) — suggests possible acceleration.




## 5. Conclusion & Implications

Sleep deprivation is driven primarily by stress, exacerbated by modern screen time, environmental noise, and lifestyle factors. Teens and shift workers (especially female nurses/teachers) face the highest risk, with real-world safety consequences (774K drowsy crashes) and a projected rise to 52–58% of adults by 2035. The 94.5% accurate predictive model highlights stress as the most actionable target. Distributed processing enabled analysis at scale — a critical capability for large societal datasets.

Distributed frameworks like PySpark are essential for uncovering patterns in massive real-world data. Future work could include integrating wearables or real-time noise monitoring for personalized risk alerts.

## 6. References & Tools

- Datasets: Kaggle US Accidents, Sleep Health, Sleep Efficiency, Teen Phone Addiction, Urban Noise Levels
- Tools: PySpark, Plotly, Pandas, SciPy, HTML/CSS (Dashboard)
- GitHub: https://github.com/acsahpauline/sleepspark