UDACITY
Part of Accenture

Sleep Health and Lifestyle Analysis

Descriptive Statistics

Data Scientist: Edward Amankwah

Agenda

Results of sleep health and lifestyle analysis

- Data Description
- Typical Amount of Physical Activity
- Number of Daily Steps
- Distribution of Heart Rates

Data Description

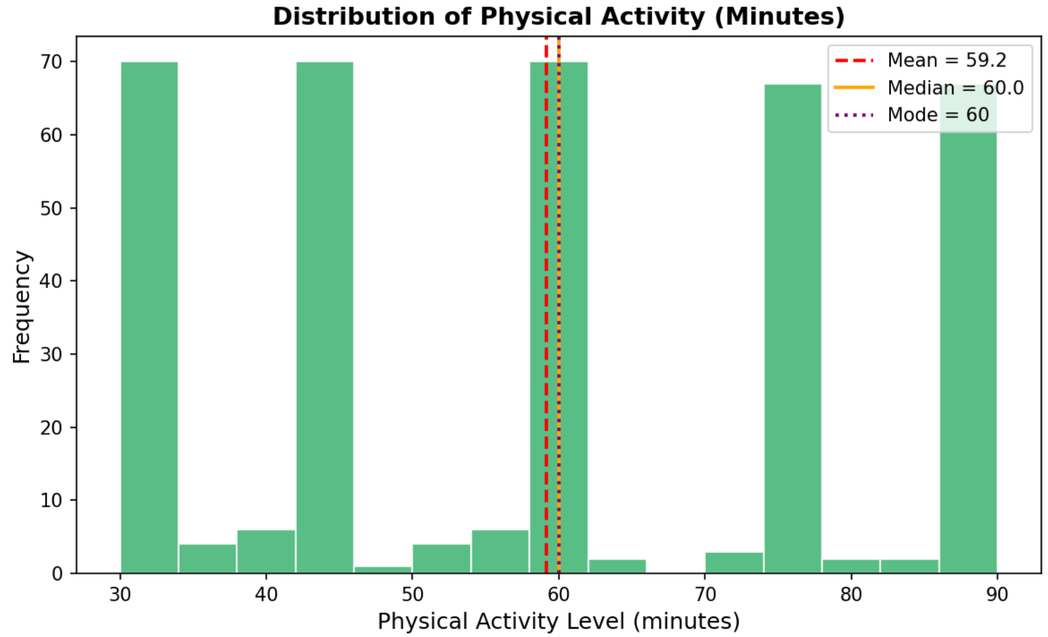
Type	Examples from the dataset
Continuous	Sleep Duration (e.g., 6.1, 7.8 hours — measured on a real-number scale) and Age (recorded as a decimal-capable value)
Integer	Physical Activity Level (whole minutes, e.g., 30, 42, 60) and Daily Steps (whole-number counts, e.g., 3000, 10000)
Ordinal categorical	Quality of Sleep (rated 1–10, ordered but intervals not guaranteed equal) and Stress Level (1–10 ordered scale)
Nominal categorical	Gender (Male/Female — no natural order) and Occupation / BMI Category / Sleep Disorder (unordered labels)

Note:

- BMI Category (Normal, Overweight, Obese) could be argued as ordinal rather than nominal, since there is an implied health progression across those categories.
- In the context of this project, it was classified as nominal, which is a defensible and common choice,

Typical Amount (Minutes) of Physical Activity

Measure	Value
Mean	59.17 minutes
Median	60.00 minutes
Mode	60 minutes

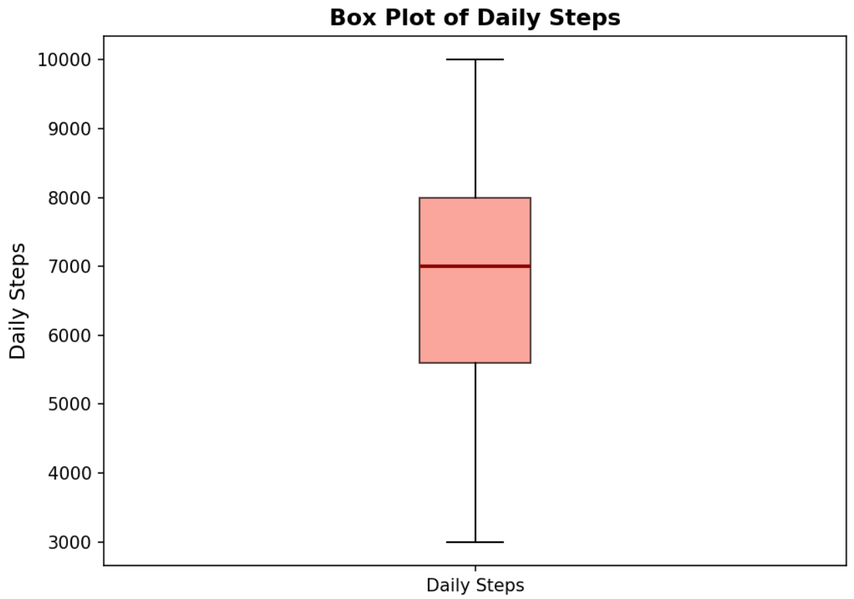


What this tells us about skewness:

- The mean (59.17), median (60), and mode (60) are nearly identical, which is the hallmark of a **symmetric, roughly normal distribution**.
- When mean < median < mode, a distribution is left-skewed; when mean > median > mode, it is right-skewed.
- Here the three measures of center converge, and the computed skewness is only **−0.07**, confirming the data is very nearly symmetric with essentially no skew.

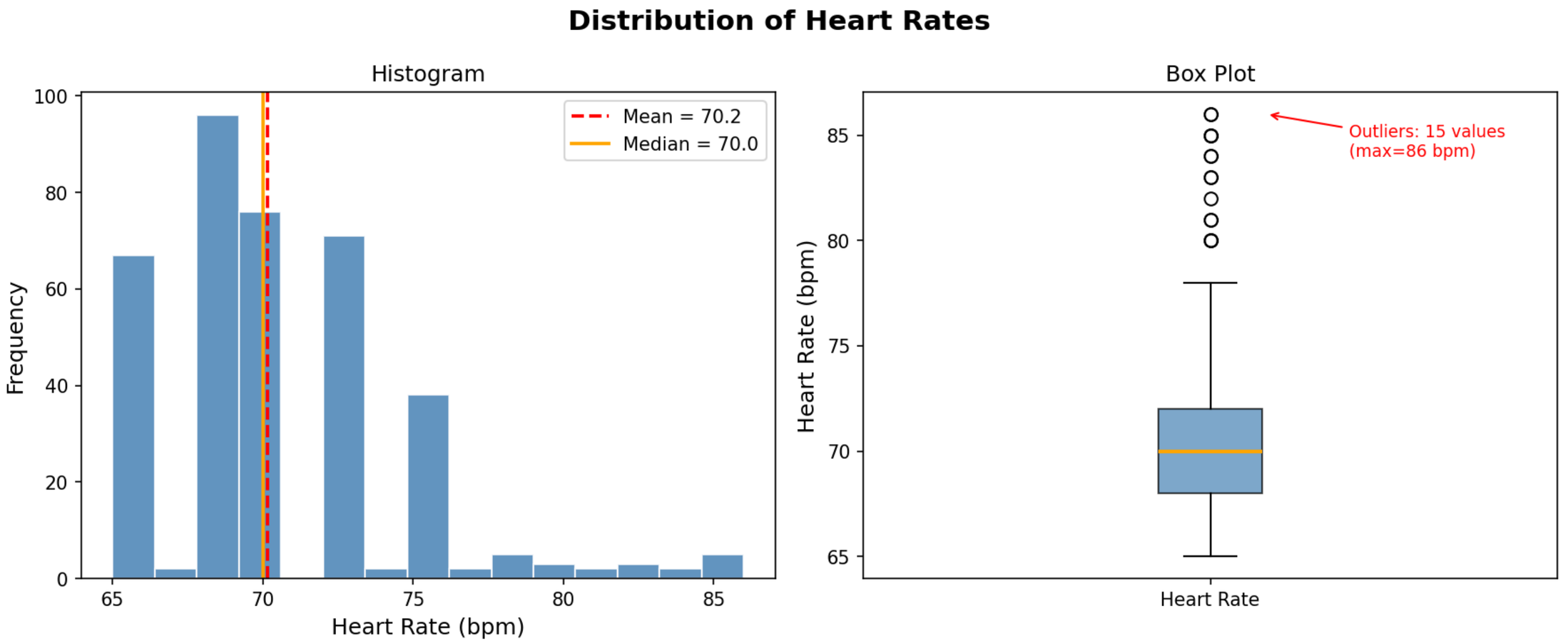
Analysis of Daily Steps Taken

Measure of Spread	Value
Standard Deviation	1,617.92 steps
Minimum	3,000 steps
Maximum	10,000 steps
Range	7,000 steps
Variance	2,617,651
Q1 (25th percentile)	5,600 steps
Q3 (75th percentile)	8,000 steps
IQR (Q3 − Q1)	2,400 steps
Skewness	+0.18 (very slight right skew)



- IQR of 2,400 is a more robust measure of spread than the range because it is resistant to extreme values
  - middle 50% of participants walk between 5,600 and 8,000 steps per day.
- Standard deviation of ~1,618 steps confirms moderate variability.
- Mean (6,817) is only slightly above the median (7,000), consistent with the near-zero positive skew.

Distribution of Heart Rates



See the description of next slide

Distribution of Heart Rates

Shape:

- Distribution is right-skewed (positively skewed), with a skewness of +1.22.
- Bulk of heart rates cluster between 65–75 bpm, but a tail extends to the right toward higher values.
- Notably, the mean (70.17 bpm) exceeds the median (70.0 bpm), which is the classic signature of right skew
    - mean is pulled toward the tail.

Outliers:

- Using the IQR fence method (Q1 = 68, Q3 = 72, IQR = 4, upper fence = 78 bpm), there are 15 outliers
    - all on the high side, with values ranging from 80 to 86 bpm.
- These high-end heart rate values are visible as individual dots above the whisker in the boxplot.

Project Stand Out: Mean vs. Median → Distribution Shape

The relationship between the mean and median is one of the most reliable indicators of a distribution's shape. Here's the rule:

- Mean > Median → right skew (positive) — the mean is pulled toward a high-value tail
- Mean ≈ Median → symmetric (approximately normal)
- Mean < Median → left skew (negative) — the mean is pulled toward a low-value tail

The table below is how that plays out across the key variables in this dataset:

Project Stand Out: Mean vs. Median → Distribution Shape

Variable	Mean	Median	Mode	Mean vs. Median	Shape
Physical Activity (min)	59.17	60.00	60	Mean < Median	Slight left skew (skewness = −0.07, nearly symmetric)
Daily Steps	6,816.84	7,000.00	—	Mean < Median	Slight left skew (skewness = +0.18, nearly symmetric)
Heart Rate (bpm)	70.17	70.00	—	Mean > Median	Right skew (skewness = +1.22, notable tail toward high bpm)
Age	42.18	43.00	—	Mean < Median	Slight left skew
Quality of Sleep	~7.3	7.0	—	Mean > Median	Slight right skew

The key takeaway:

- For physical activity, the mean, median, and mode are nearly identical (59.17 / 60 / 60), so the distribution is essentially symmetric.
- For heart rate, the mean sits above the median and 15 high-end outliers pull the mean upward
  - a typical example of how outliers distort the mean more than the median, revealing right skew.
  - precisely why the median is the preferred measure of center for skewed data
    - resistant to the pull of extreme values.

Multiple Measures of Center

Rather than reporting just one, here are all three measures of center for each key variable:

Variable	Mean	Median	Mode
Physical Activity	59.17 min	60.0 min	60 min
Daily Steps	6,816.84	7,000	8,000
Heart Rate	70.17 bpm	70.0 bpm	70 bpm
Age	42.18 yrs	43.0 yrs	—

- Reporting all three matters because they each tell a different story:
  - mode reveals the most common value, the median is robust to outliers, and the mean incorporates every data point but is sensitive to extreme values.
- When all three agree (as with physical activity), you can be confident the distribution is symmetric.
- When they diverge (as with heart rate), you know a skew or outliers are present.

Multiple Measures of Spread

Measure	Physical Activity	Daily Steps	Heart Rate
Range	75 − 30 = 45 min	10,000 − 3,000 = 7,000	86 − 65 = 21 bpm
Standard Deviation	20.78 min	1,617.92 steps	4.14 bpm
Variance	431.7	2,617,651	17.1
Q1	45 min	5,600	68 bpm
Q3	75 min	8,000	72 bpm
IQR	30 min	2,400 steps	4 bpm

Why report multiple measures?

- Range is simple but is entirely determined by just two extreme values, so one outlier can make it misleading.
- Standard deviation gives an average distance from the mean but is also sensitive to outliers.
- IQR is the most robust : describes the spread of the middle 50% of the data and is unaffected by outliers.
- For heart rate particularly, the IQR (4 bpm) paints a very different picture than the range (21 bpm), because those 15 outliers inflate the range dramatically while the IQR remains stable.
- Together, these measures give a complete, picture of variability

References

1. **Centers for Disease Control and Prevention. (2022).** Sleep and sleep disorders: Data and statistics. U.S. Department of Health and Human Services. https://www.cdc.gov/sleep/data-statistics.html
2. **Field, A. (2018).** Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications. [Foundational reference for measures of centre, spread, and distributional shape interpretation including skewness diagnostics.]
3. **Hunter, J. D. (2007).** Matplotlib: A 2D graphics environment. Computing in Science and Engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55 [Reference for histogram and boxplot visualisation methodology used in this analysis.]
4. **Kaggle. (2023).** Sleep health and lifestyle dataset. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset [Primary dataset used in all statistical analyses presented in this report.]
5. **McKinney, W. (2022).** Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media. [Reference for pandas-based statistical computation and data manipulation methods.]
6. **National Sleep Foundation. (2023).** Sleep health index. Sleep Foundation. https://www.sleepfoundation.org [Contextual norms for sleep duration and quality benchmarks applied in interpretation of findings.]
7. **NumPy Development Team. (2023).** NumPy documentation: Statistical functions. https://numpy.org/doc/stable/reference/routines.statistics.html [Reference for variance, standard deviation, percentile, and skewness computation.]
8. **pandas Development Team. (2023).** pandas documentation: DataFrame.describe and related methods. https://pandas.pydata.org/docs/ [Reference for summary statistics generation and data type handling.]
9. **Triola, M. F. (2021).** Elementary statistics (14th ed.). Pearson Education. [Core reference for variable type classification, the IQR fence outlier detection method, skewness interpretation, and the mean-median-mode relationship as a distributional shape indicator.]
10. **World Health Organisation. (2019).** Sleep and health: A public health perspective. WHO Press. [Epidemiological context for sleep disorder prevalence and health impact across global adult populations.]