Ein Bild, das Text, Schrift, Logo, Screenshot enthält.

Automatisch generierte Beschreibung

**Big Data – Final Report**

Computer Science

Cooperative State University Heidenheim

submitted by

Michael Behringer

Submission date: 11.09.2024

Matriculation Number, Course 1788307, TINF2021AI

Reviewer: Amin Mobasheri (Dr. rer. nat.)

# Introduction

The Report is about a dataset collected in our Big Data lecture. This dataset was collected using an Arduino system equipped with two sensors: the ADT7410 temperature sensor and the VEML7700 light sensor. The project’s objective is to monitor environmental changes in temperature and light around a personal computer (PC) setup.

## Sensors and Data Collection

Temperature Sensor (ADT7410): The ADT7410 temperature sensor captures temperature readings in degrees Celsius. It was positioned near the PC's exhaust to measure how hot the area becomes when the PC is in use.

Light Sensor (VEML7700): The VEML7700 light sensor records light intensity, providing values between 0 and 1023, where 0 represents the brightest condition and 1023 represents the darkest. The sensor was placed on the desk to track changes in ambient light levels within the room.

## Data Transmission

Data from both sensors is collected every minute and transmitted via MQTT to an external MQTT broker. This allows for real-time monitoring and analysis of environmental conditions surrounding the PC setup.

## Project Goals

The main goals of this project are:

* To explore how light levels in the room relate to the temperature produced by the PC’s exhaust.
* To identify patterns in working hours by analyzing data from various times of the day and different days of the week.
* To investigate how sunlight impacts the room's environment based on the window’s orientation.
* Using outside temperature data, potentially explore whether we can detect when the window is open or closed. Changes in the room's temperature compared to the external environment may offer insights into whether the window is allowing airflow and sunlight to enter, affecting both the temperature and light levels indoors.

By analysing this data, we aim to gain insights into how light and temperature fluctuate during the day, how these changes are affected by PC usage, and how sunlight plays a role in altering room conditions. This could further help in understanding work habits and the relationship between artificial light, natural light, and heat within a workspace environment.

# Data Cleaning & Preparation

## Corrupted Data

To ensure data quality, we identified and removed erroneous data points:

Initial Setup Data: The first 5,000 data entries from both the temperature and light sensors were discarded. These entries contained incorrect readings due to calibration issues during the initial setup phase.

Light Sensor Spikes: The light sensor data was closely analyzed for sudden and unrealistic changes. A threshold was set to flag data points where the light level changed by more than 400 units within a 5-minute interval, which is uncommon even with artificial lighting. These flagged entries were removed as they likely represented sensor anomalies.

## Anomalies/ Sudden Spikes or Drops

Anomalies were found primarily in the light sensor data, with a few issues in the temperature data:

Light Sensor Anomalies: Significant anomalies were detected between May 23, 2024, and May 27, 2024. During this period, the light sensor recorded extreme fluctuations, with changes exceeding 400 units in just 5 minutes. This behavior was deemed unrealistic and this entire time window was excluded from the dataset.

Temperature Sensor Anomalies: Although the temperature sensor performed consistently, some negative values were recorded, which is outside the expected range and likely a result of sensor malfunctions. These erroneous readings were removed from the dataset. All valid temperature readings, including values up to 50°C, were retained, as they reflect the typical range for the PC’s exhaust temperature and workload.

Figur 1 visualizes these unrealistic light changes, in the end of May:

Ein Bild, das Text, Screenshot, Diagramm, Electric Blue (Farbe) enthält.

Automatisch generierte Beschreibung

Figure 1: Unrealistic light changes

## Periods of inactivity

By checking both the light and temperature sensor values for long periods of inactivity, we can observe that both sensors have about 50 instances of missing values (no new value in the past 5 minutes) over a period of 4 months.

These periods of inactivity can be used to remove spikes and drops in the values.

* Total number of inactivity periods (5 Minutes): 43
* Total number of inactivity periods (15 Minutes): 18
* Total number of inactivity periods (60 Minutes): 13

Inactivity periods of less than 5 minutes were considered acceptable, as temperature and light conditions are unlikely to change significantly within such a short timeframe. These short gaps were retained, ensuring that any legitimate changes, such as artificial light turning on or off, would still be captured by the dataset. Even with longer gaps, the overall trends in temperature and light levels would remain largely unaffected, ensuring that the integrity of the analysis and the evaluation of the data remains accurate and meaningful.

# Exploring Data Analysis

In this section, we analyze the temperature and light intensity data collected over the measurement period. Statistical methods were applied to both datasets to extract meaningful insights, including measures of central tendency (mean, median), variability (standard deviation, range, interquartile range), and data distribution patterns.

## Temperature Data Analysis

The temperature data includes 166,627 data points recorded between April 2, 2024, and July 31, 2024. To gain a deeper understanding of how temperature fluctuates in the monitored environment, we calculated the following statistical metrics:

* Average Temperature: 25.21°C, indicating a typical working temperature near the PC's exhaust.
* Median Temperature: 23.90°C, showing that half of the recorded temperatures were below this value.
* Standard Deviation: 5.03°C, suggesting a moderate spread around the average, likely reflecting varying PC usage levels.
* Minimum Temperature: 17.04°C, representing periods of low PC activity or cooler room temperatures.
* Maximum Temperature: 54.61°C, likely during periods of heavy PC usage or poor ventilation.
* Range: 37.57°C, highlighting the total variability in the dataset.
* Interquartile Range (IQR): 3.77°C, showing that the middle 50% of temperature readings were tightly clustered between 22.25°C and 26.02°C.

These figures provide a clear overview of the temperature trends, with most values clustering around the mid-20s, punctuated by occasional spikes likely related to PC workloads.

## Light Intensity Data Analysis

The light intensity data comprises 166,478 data points from the same period. The data shows a broad range of values, from complete darkness to the sensor’s maximum detectable light levels. Key metrics calculated include:

* Average Light Intensity: 839,74, which indicates a generally low-light environment, though this is skewed by the high number of extreme darkness levels.
* Median Light Intensity: 1023, reflecting that the room was often dark or the sensor detected minimal light during periods of inactivity.
* Standard Deviation: 237.75, suggesting wide variability in the light levels, likely corresponding to alternating periods of natural sunlight, artificial light, and darkness.
* Minimum Light Intensity: 1023, representing darkness or very low ambient light conditions.
* Maximum Light Intensity: 0, the sensor's maximum value, showing the brightest conditions.
* Range: 1023.00, covering the entire spectrum of possible light levels.
* Interquartile Range (IQR): 334.00, showing significant variation between the lower 25th percentile (0.00) and the upper 75th percentile (334.00), likely capturing the effect of both natural and artificial lighting.

The light intensity data demonstrates considerable fluctuations, potentially reflecting changes in day/night cycles, artificial light usage, and sunlight entering through windows.

## Box Plot

Box plots summarize the distribution of temperature and light intensity data, displaying the median, interquartile range (IQR), and potential outliers. Figur 2 is ideal for quickly understanding the spread and central tendency of my Data.

Ein Bild, das Text, Diagramm, parallel, Reihe enthält.

Automatisch generierte Beschreibung

Figure 2: Box Plot

Components of a Box Plot:

* Box
  + The box represents the interquartile range (IQR), which contains the middle 50% of the data. It spans from the first quartile (Q1) (25th percentile) to the third quartile (Q3) (75th percentile).
  + The line inside the box marks the median (50th percentile), showing where the center of the data lies.
* Whiskers
  + The whiskers extend from the box to the smallest and largest values that are within 1.5 times the IQR from Q1 and Q3. These are called the "inner fences."
  + Any data points within this range are considered part of the typical data variation.
* Outliers
  + Data points that fall outside the whiskers (more than 1.5 times the IQR from Q1 or Q3) are plotted as individual points. These are considered outliers—unusually high or low values that might indicate anomalies or extreme conditions in the data (such as a sensor malfunction or unusual environmental changes).

Although the Temperature Box Plot showing Outliers at about 32 degrees Celsius, these values are no outliers in my Dataset, because that are the Temperature Values, when the PC is running.

## Histogram

A histogram provides a graphical representation of the distribution of numerical data by grouping data into bins (or intervals). It shows how often data points fall within a certain range and helps in understanding the frequency distribution of the data.

Ein Bild, das Text, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Figure 3: Histogram Temperature

When examining the distribution of temperatures, it becomes evident that the data predominantly feature room temperatures ranging from 21 to 27 degrees Celsius. Another notable peak is observed at around 35 degrees Celsius, with a variation of plus or minus two degrees, which corresponds to the exhaust heat from the PC.

Ein Bild, das Text, Diagramm, Reihe, Screenshot enthält.

Automatisch generierte Beschreibung

Figure 4: Histogram Light

When examining the distribution of light values, it is noticeable that the maximum value of 1023 frequently occurs. This is because 1023 represents absolute darkness, and the sensor often fails to detect anything in low-light conditions.

# Time series analysis

In this chapter, we delve into the time series analysis of the temperature and light intensity data collected over the study period. Time series analysis is essential for understanding how these environmental variables evolve over time and identifying any trends, seasonal patterns, or periodic fluctuations.

### Trend

Ein Bild, das Reihe, Diagramm, Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Reihe, Diagramm, Text enthält.

Automatisch generierte Beschreibung

## Time Series Analysis

Ein Bild, das Text, Screenshot, Rechteck, Design enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Schrift, Design enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Rechteck, Schrift enthält.

Automatisch generierte Beschreibung

Ein Bild, das Screenshot, Text, Schrift, Reihe enthält.

Automatisch generierte Beschreibung







