

Data Analysis Final Assignment Report

Team: Fantastic Three

Md Ehtashamul Huque & *MdShamsurRahmanShishir* & *SheikhMd.Nayeem*

1 Contributions

Clearly state each team member's specific contributions. Be concrete.

- Md Ehtashamul Huque:
 - Dataset selection and acquisition
 - Data quality analysis and preprocessing pipeline
 - Missing-value handling and outlier analysis
- Md Shamsur Rahman Shishir:
 - Visualizations and exploratory data analysis (EDA)
 - Probability analysis tasks
 - Law of Large Numbers (LLN) and Central Limit Theorem (CLT) demonstrations
- Sheikh Md. Nayeem:
 - Regression modeling and interpretation
 - Model evaluation and comparison
 - Report writing and figure polishing

2 Dataset Description

- Dataset name and source (Kaggle): IoT Telemetry Sensor Dataset (public IoT sensor dataset)
- The dataset consists of timestamped sensor readings collected continuously from multiple IoT devices, making it suitable for temporal pattern analysis and forecasting.
- Time period covered and sampling frequency: From 2020-07-12 00:01 to 2020-07-20 00:03 (approximately 8 days) with high-frequency asynchronous sampling.
- Key variables analyzed: Temperature, Humidity, CO, Smoke, LPG, Motion (binary), Light (binary), Device ID
- Size and structure:
 - Number of observations (rows): 405,184
 - Number of features (columns): 9
 - Target variable(s): Temperature
- Missing data summary: 0 missing data
- Any known limitations or caveats: Short observation window (8 days) and potential sensor noise due to real-world deployment.

3 Task 1. Data Preprocessing and Basic Analysis

3.1 Basic statistical analysis using pandas

- Descriptive stats (mean, std, min, max, quantiles) were computed for all numeric sensor variables.
- Grouped summaries by device and by day revealed differences in temperature and gas sensor distributions.

3.2 Original data quality analysis including visualization

- Outliers and suspicious values were identified using boxplots and percentile-based thresholds.
- Consistency checks confirmed correct timestamp ordering and absence of duplicates.

3.3 Data preprocessing

- Cleaning steps performed: Timestamp conversion and sorting.
- Outlier handling: Extreme values beyond reasonable physical limits were removed using percentile thresholds.
- Feature engineering: Standardization applied for PCA and regression models.
- Final dataset shape after preprocessing: 405,184 rows and 9 columns with no missing values.

3.4 Preprocessed vs original data visual analysis

- Before vs after comparison plots showed reduced skewness and improved distribution stability.
- Minor smoothing occurred, but overall signal structure was preserved.

4 Task 2. Visualization and Exploratory Analysis

4.1 Time series visualizations

- Plot of main variable(s) over time: Temperature and humidity plots revealed clear diurnal cycles.
- Daily cycles were consistent across days.

4.2 Distribution analysis with histograms

- Histograms for key numeric variables showed mild skewness in temperature and heavy tails in gas sensors.
- Gas sensors exhibited occasional extreme values.

4.3 Correlation analysis and heatmaps

- Pearson correlation was used due to approximately linear relationships.
- Strong correlations were observed among CO, Smoke, and LPG sensors.

4.4 Daily pattern analysis

- Aggregation method: Hourly mean aggregation.
- Stable daily temperature cycles were observed.
- Temperature patterns were stable; gas sensors were more variable.

4.5 Summary of observed patterns

- Statement 1 (True): **Temperature follows a daily cycle.** Evidence: Hourly mean plots.
- Statement 2 (True): **Gas sensors are strongly correlated.** Evidence: Pearson correlation heatmap.
- Statement 3 (False): **Motion events occur uniformly over time.** Evidence: Event density varies by hour.

5 Task 3. Probability Analysis

5.1 Threshold-based probability estimation

- Threshold(s) choice: Upper quantiles of temperature distribution.
- Estimate probabilities of exceeding thresholds: Empirical probabilities computed from data.
- Visual support: Empirical CDF plots.

5.2 Cross tabulation analysis

- variables: Motion and Light events.
- Co-occurrence was higher than expected under independence.

5.3 Conditional probability analysis

- Define events A and B : A = Motion detected, B = Light on.
- Compute and interpret $P(A)$, $P(B)$, $P(A | B)$, $P(B | A)$.
- comparison and conclusion: $P(A | B) > P(A)$, indicating dependence; Bayes' rule verified empirically.

5.4 Summary of observations from each probability task

- Threshold probability: High-temperature events occur infrequently.
- Crosstab: Motion and light frequently co-occur.
- conditional probability: Event variables are dependent.

6 Task 4. Statistical Theory Applications

6.1 Law of Large Numbers (LLN) demonstration

- Variable: Temperature is stable and continuously measured.
- Experiment: Sample mean plotted as n increases.
- Plot and short interpretation: Sample mean converges to a stable value.

6.2 Central Limit Theorem (CLT) application

- Sampling procedure: Repeated random sampling with increasing sample sizes.
- Show distribution of sample means for increasing n .
- Plot(s): Histograms of sample means approaching normality.

6.3 Result interpretation

- LLN : Large samples yield reliable averages.
- CLT : Sample means approximate a normal distribution with minor deviations due to noise.

7 Task 5. Regression Analysis

7.1 Linear or Polynomial model selection

- Define target y and predictors X : y = Temperature; X = sensor readings, events, device ID.
- Motivation for linear vs polynomial: Linear baseline compared with non-linear models.
- Time-aware split used to avoid data leakage.

7.2 Model fitting and validation

- Fit procedure and preprocessing: Scaling and feature preparation.
- Validation method: Chronological holdout split.
- Metrics report: Standard regression performance metrics.
- Residual analysis: Residual plots examined for systematic error.

7.3 Result interpretation and analysis

- Main effects and practical meaning: Gas sensors and humidity were important predictors.
- Failure cases: Linear models struggled with non-linear relationships.

8 Bonus Tasks

- Q-Q plot with explanation: Regression residuals showed mild deviations from normality.

9 Key Findings and Conclusions

- Main findings from preprocessing and EDA: Clear temporal patterns and sensor correlations.
- Main findings from probability tasks: Motion and light events are dependent.
- Main findings from LLN and CLT: Empirical confirmation of statistical theory.
- Main findings from regression: Random Forest outperformed linear models.
- Limitations: Short time span and sensor noise.
- Further work depends on time: Longer datasets and sequence models.

10 Reproducibility Notes

- Exact dataset source link and version or download date: Public IoT telemetry dataset (downloaded 2020).
- Key libraries used and versions: pandas, NumPy, matplotlib, scikit-learn, SciPy.