Data Analysis and Visualization

Occupancy detection using PIR sensors

Semester Project

Laraib Sultana	**NUM-BSCS-2022-06**
Samreen Kazmi	**NUM-BSCS-2022-52**
Mariyam Noor ul ain	**NUM-BSCS-2022-25**

Project Objective

The objective of this project is to implement and evaluate a lifelong learning-based occupancy detection system on edge devices, as proposed in existing research, and to further explore its performance in real-world scenarios. The project aims to train and compare multiple machine learning models—including the lifelong learning model—with a focus on analyzing detection accuracy, adaptability over time, and computational efficiency. Additionally, the project will investigate the temporal effects on occupancy patterns (e.g., time of day or day of the week) to assess how time influences model predictions. This comparative and time-based analysis will help determine the most effective approach for reliable and context-aware occupancy detection in smart environments.

Project Phases

Phase 2: Exploratory Data Analysis (EDA)

- **Data Description:**

  The dataset consists of 7,651 time-indexed observations spanning from August to October 2024, with 61 features including timestamp, temperature, 55 PIR motion sensor readings, and derived time-based fields like day of the week and month. The Label column serves as the target variable with three classes (0, 1, 2), though it shows class imbalance. All data types are appropriately assigned, with numerical features normalized between 0 and 1, and no missing or duplicate records present. Summary statistics indicate a consistent structure across PIR features, and the dataset appears well-prepared for modeling or further analysis.

  Type: pandas.core.frame.DataFrame

  Entries (rows): 7,651

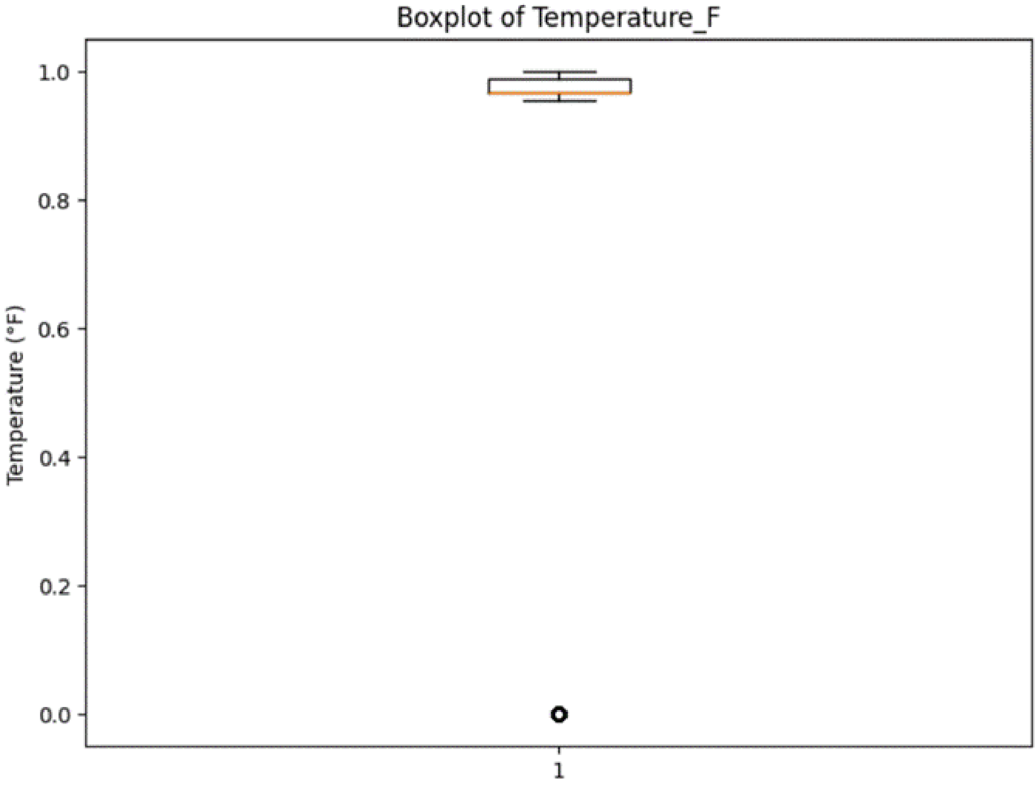
Datetime index range: 2024-08-08 19:19:56 to 2024-10-08 04:35:21

Columns (variables): 61

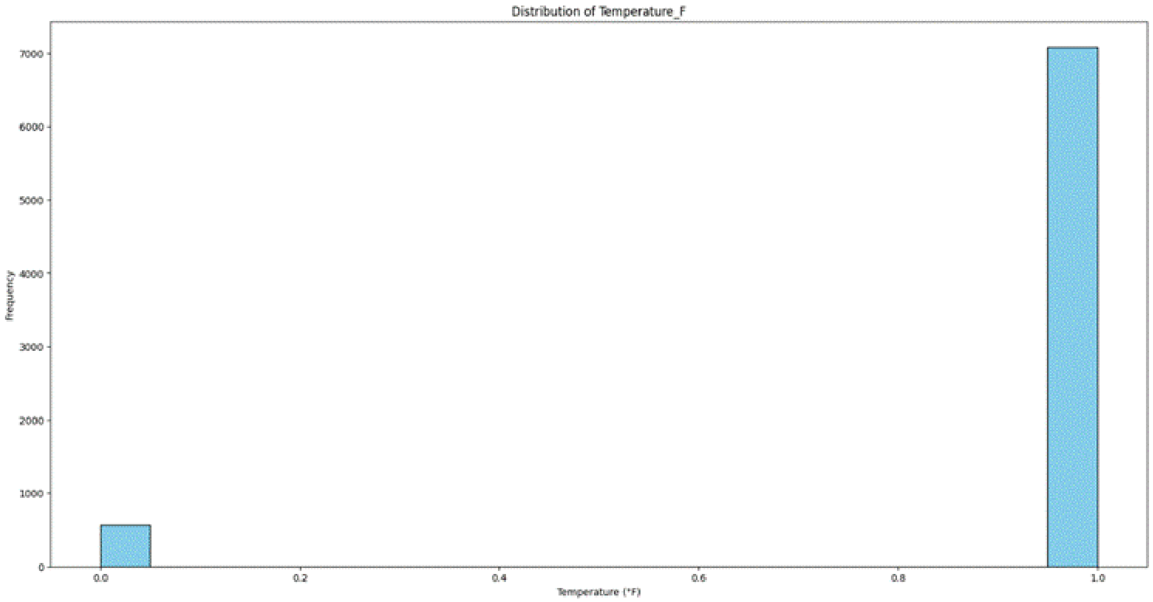
Duplicate rows: 0

- **Visualizations:**

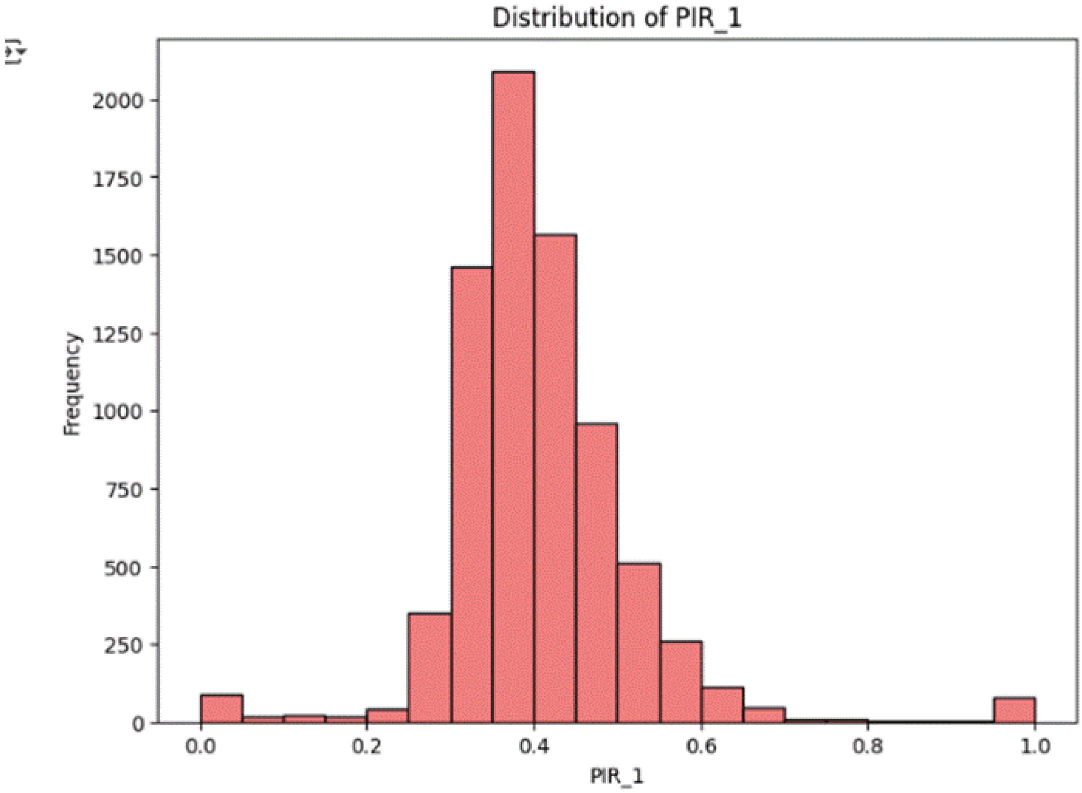
Box plot

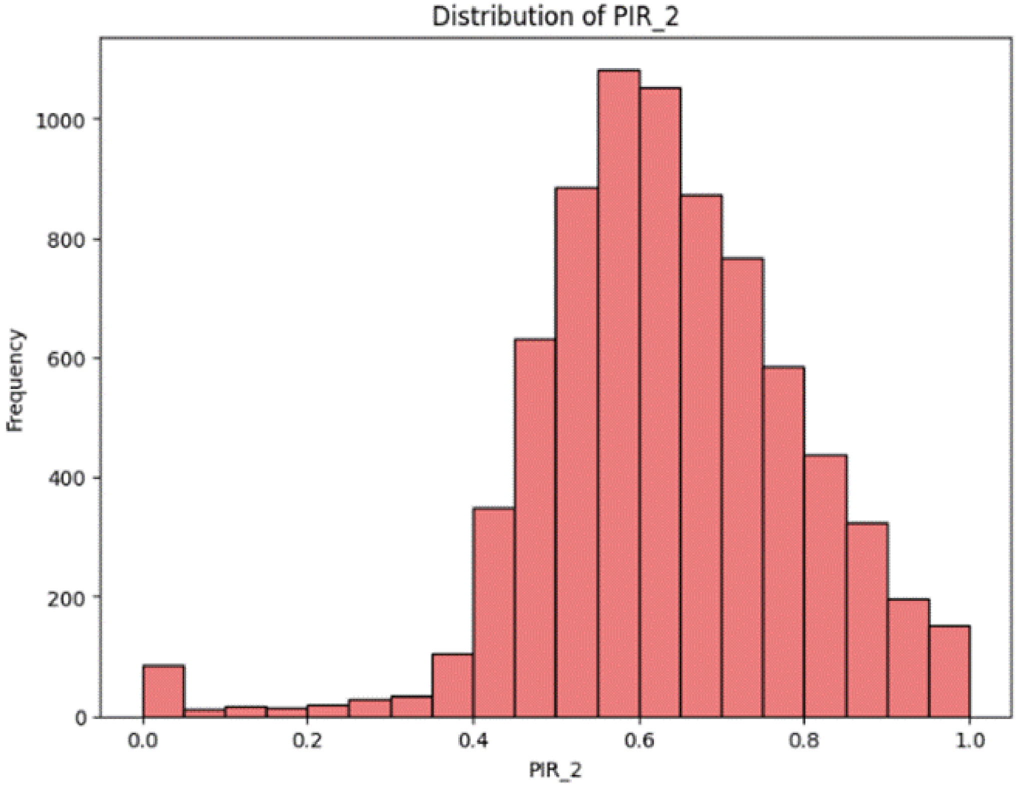


Histogram:

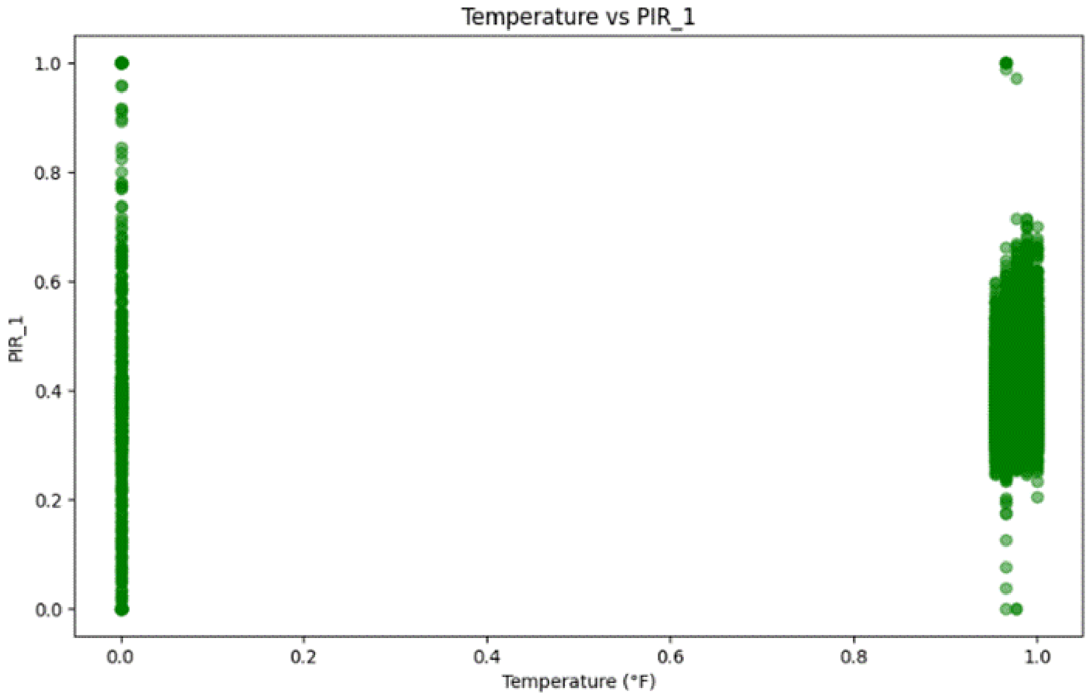


```
for col in pir_columns
```





Scatter Plot:



- **Initial Findings:**

General Overview

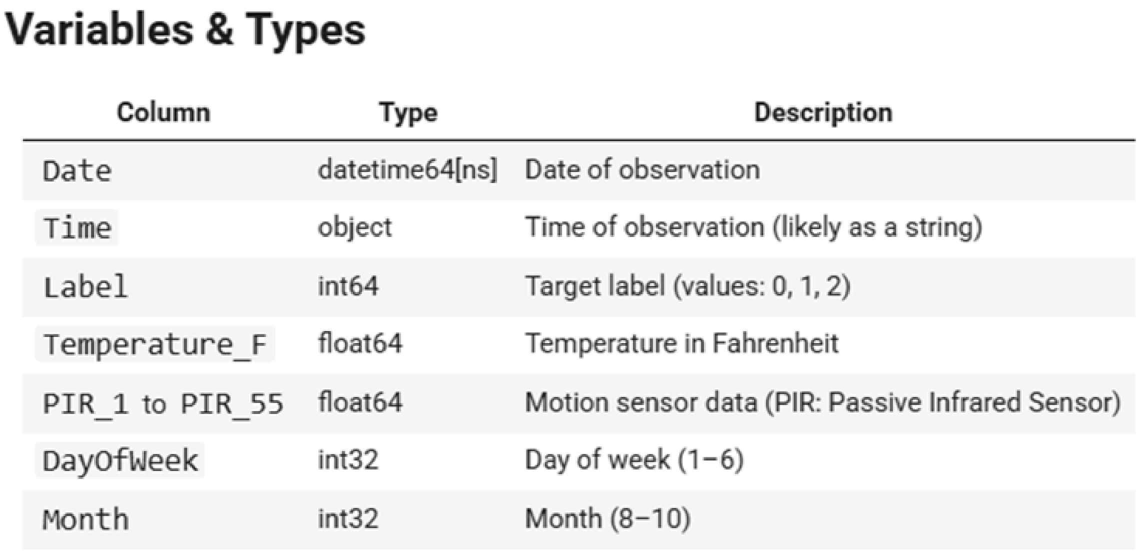
Type: pandas.core.frame.DataFrame

Entries (rows): 7,651

Datetime index range: 2024-08-08 19:19:56 to 2024-10-08 04:35:21

Columns (variables): 61

Duplicate rows: 0



Variables & Types

Column	Type	Description
`Date`	datetime64[ns]	Date of observation
`Time`	object	Time of observation (likely as a string)
`Label`	int64	Target label (values: 0, 1, 2)
`Temperature_F`	float64	Temperature in Fahrenheit
`PIR_1` to `PIR_55`	float64	Motion sensor data (PIR: Passive Infrared Sensor)
`DayOfWeek`	int32	Day of week (1–6)
`Month`	int32	Month (8–10)

Summary Statistics

General stats for key variables:

- **Label** (target):
  - Mean: 0.258
  - Std: 0.584
  - Min: 0
  - Max: 2
  - Median (50%): 0
- **Temperature_F**:
  - Mean: 0.903
  - Std: 0.257
  - Range: 0.0 to 1.0 *(likely normalized)*
- **PIR_1**:
  - Mean: 0.407
  - Std: 0.112
  - Range: 0.0 to 1.0
- **PIR_2 to PIR_55**:
  - Similar structure to PIR_1

- Generally normalized in [0,1]
- Means range: ~0.4 to ~0.7
- Std dev ranges: ~0.1 to ~0.18

- **DayOfWeek**:
  - Integer range: 1 (Monday?) to 6
  - Mean: 4.9 (suggesting skew toward later weekdays)
- **Month**:
  - Range: 8 to 10
  - Mean: ~9.0
  - Reflects the date range of August to October 2024

Observations

- PIR sensor data (55 features) likely represent presence or motion in different zones or time slices.
- No missing values in any column — complete dataset.
- Temperature values seem normalized (0 to 1), which is common in preprocessing.
- The target variable Label has class imbalance — most values are 0.

Phase 3: Data Preprocessing

1. Handling Missing Values

- **Observation**: No missing values were found in any of the 61 columns.
- **Action**: No imputation or deletion was necessary.

2. Duplicate Removal

- **Observation**: No duplicate rows were detected in the dataset.
- **Action**: No rows were removed.
- **Resulting Shape**: (7,651 rows, 61 columns)

3. Outlier Treatment

- **Observation**: Extreme values were present in Temperature_F and the PIR sensor readings.
- **Action**: Winsorization was applied at the 1st and 99th percentiles to cap outliers.

- **Validation**: Boxplots were used for visual inspection before and after winsorization.

4. Feature Scaling

- **Columns Scaled**: Temperature_F and all PIR sensor columns (PIR_1 to PIR_55)
- **Scaling Method**: MinMaxScaler (scales data to a 0–1 range)
- **Rationale**: Standardizing the feature range improves model performance and convergence.

5. Categorical Encoding

- **Encoded Column**: Label
- **Encoding Method**: Label Encoding (retained as 0, 1, 2)
- **Rationale**: Converts categorical values to numeric for compatibility with machine learning models.

6. Time Feature Engineering

- **New Features Created**:
  - DayOfWeek: Extracted from Date
  - Month: Extracted from Date
  - DateTime: Combined Date and Time, set as index
- **Rationale**: Time-based features provide insight into temporal patterns and trends.

7. Data Subsetting

- **Action**: No subsetting applied; the full dataset was retained.
- **Rationale**: Preserves the complete data for training and analysis.

Summary Table

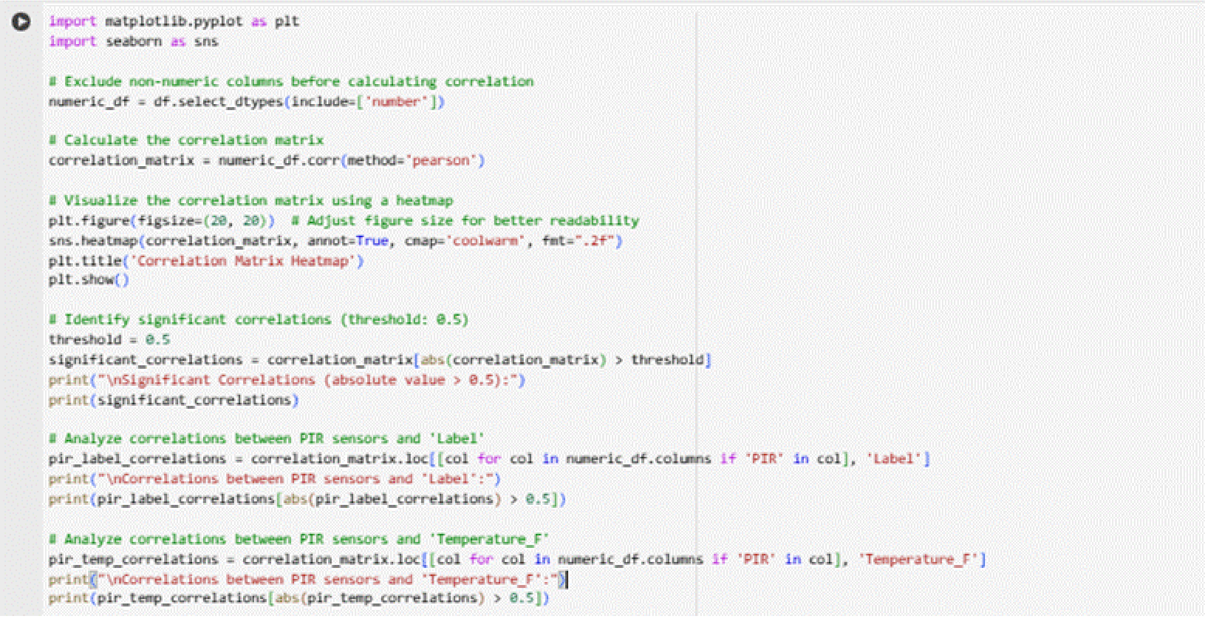
Step	**Action Taken**
Missing values	**None Found**
Duplicate rows	**None Removed**
outliers	**Minorized at 1st and 99th percentile**

Feature scaling	**Minmax Scaler**
Encoding	**Label Encoding on Label**
Time Feature	**Added day of week,Month,Date,time**
Data Subsetting	**none**

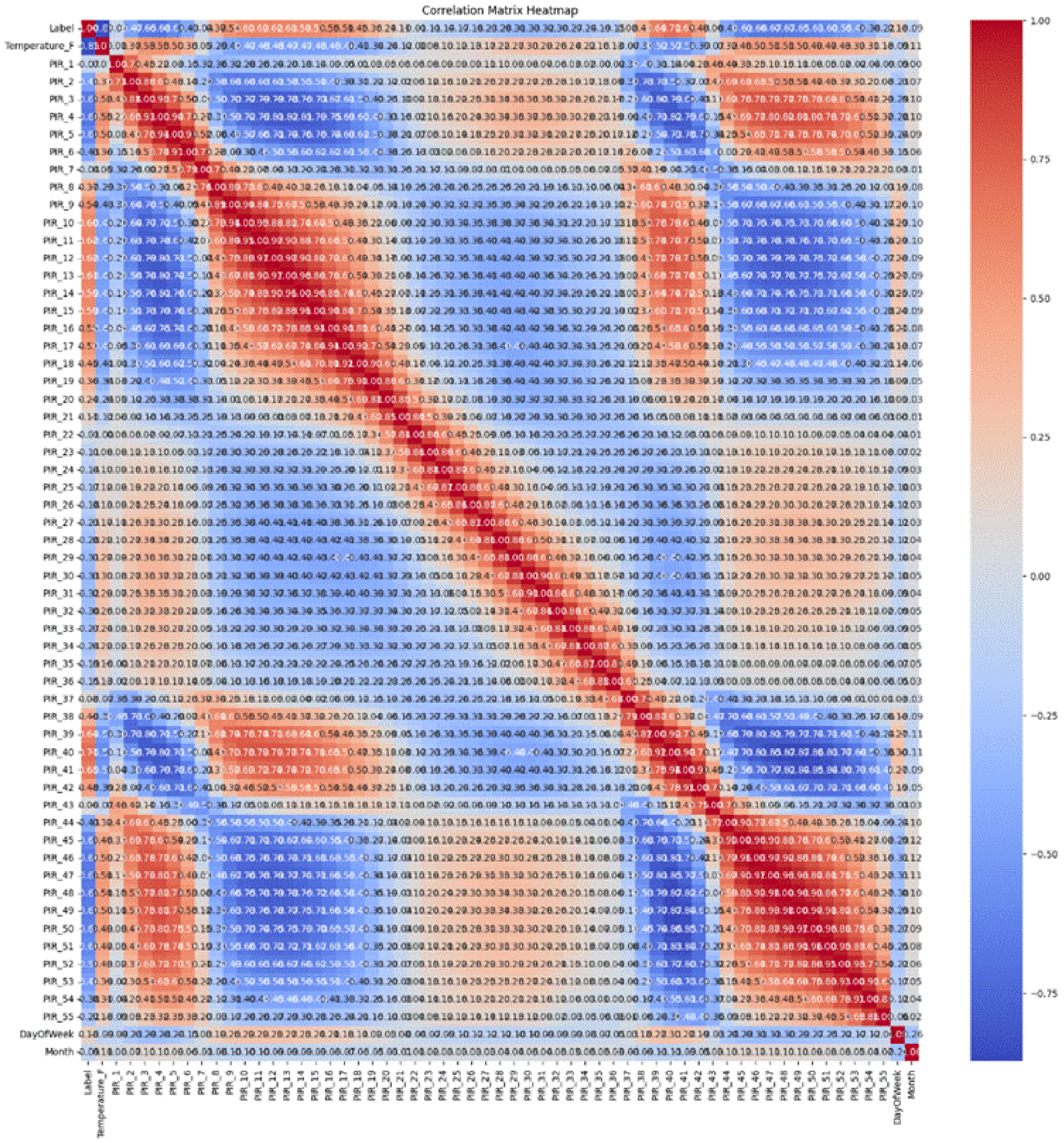
Phase 4: Correlation Analysis

Correlation Matrix:

Code:



```
import matplotlib.pyplot as plt
import seaborn as sns

# Exclude non-numeric columns before calculating correlation
numeric_df = df.select_dtypes(include=['number'])

# Calculate the correlation matrix
correlation_matrix = numeric_df.corr(method='pearson')

# Visualize the correlation matrix using a heatmap
plt.figure(figsize=(20, 20))  # Adjust figure size for better readability
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix Heatmap')
plt.show()

# Identify significant correlations (threshold: 0.5)
threshold = 0.5
significant_correlations = correlation_matrix[abs(correlation_matrix) > threshold]
print("\nSignificant Correlations (absolute value > 0.5):")
print(significant_correlations)

# Analyze correlations between PIR sensors and 'Label'
pir_label_correlations = correlation_matrix.loc[[col for col in numeric_df.columns if 'PIR' in col], 'Label']
print("\nCorrelations between PIR sensors and 'Label':")
print(pir_label_correlations[abs(pir_label_correlations) > 0.5])

# Analyze correlations between PIR sensors and 'Temperature_F'
pir_temp_correlations = correlation_matrix.loc[[col for col in numeric_df.columns if 'PIR' in col], 'Temperature_F']
print("\nCorrelations between PIR sensors and 'Temperature_F':")
print(pir_temp_correlations[abs(pir_temp_correlations) > 0.5])
```



Explanation:

The correlation matrix heatmap reveals several important relationships among the variables. The target variable, **Label**, shows a strong **negative correlation** with **Temperature_F** (around -0.85), as well as with PIR sensors such as **PIR_3 to PIR_6** and **PIR_45 to PIR_53**, suggesting that higher temperatures and certain sensor zones are associated with lower activity. In contrast, **Label** is **positively correlated** with sensors **PIR_9 to PIR_17** and **PIR_39 to PIR_41**, indicating areas where increased sensor activity corresponds to the target behavior. The temperature variable itself is **positively correlated** with **PIR_3 to PIR_6** and **negatively correlated** with **PIR_39 to PIR_41**, further highlighting environmental influence on movement patterns. Distinct **sensor clusters** emerge in the matrix, including **PIR_1 to PIR_6**, **PIR_9 to PIR_17**, and **PIR_46 to PIR_53**, each showing strong internal correlations, which may reflect shared physical zones or similar functional roles. **DayOfWeek** and **Month** have generally weak correlations with other variables, as expected given their temporal nature. These insights help identify potentially **redundant features**, suggest which sensors are most relevant for **predictive modeling**, and provide context for understanding how **temperature may suppress activity**, all of which are crucial for effective data-driven decision making.

- Significant Correlations:

  Some notable patterns:

  Label vs Other Variables

  - Label has strong **negative correlations** with:
    - Temperature_F: **-0.85**
    - Several PIR sensors: PIR_3 (-0.65), PIR_4 (-0.68), PIR_5 (-0.61), PIR_45 (-0.61), PIR_46 to PIR_53 (ranging from -0.55 to -0.67)
  - Label has strong **positive correlations** with:
    - PIR_9 to PIR_17: values around **0.51 to 0.62**
    - PIR_39, PIR_40, PIR_41: values up to **0.71**

  PIR Sensors vs Temperature

  - Some PIR sensors (e.g., PIR_3, PIR_4, PIR_40, etc.) are **positively correlated** with Temperature_F (e.g., PIR_3: 0.53, PIR_4: 0.55).
  - A few PIR sensors show **negative correlation** with Temperature_F (e.g., PIR_39: -0.52, PIR_40: -0.57).

  PIR Sensor Clustering

  - Strong **inter-PIR correlations**, especially among:
    - PIR_1, PIR_2, PIR_3, PIR_4, PIR_5, PIR_6: indicating these may be spatially close or measuring similar activity.
    - PIR_10–PIR_17 and PIR_46–PIR_53: high mutual correlations suggest functional groupings.

- **Key Insights**
  - There is strong evidence that activity (as measured by PIR sensors) and temperature are associated with the Label.
  - Negative correlation between Label and Temperature_F suggests that higher temperatures may be associated with one class (e.g., inactivity).

Clusters of PIR sensors are highly correlated with each other and with the Label, indicating their potential as predictors in classification or pattern recognition tasks.