PROJECT REPORT

Course name: TKO_7093 Statistical Data Analysis

Submission đate: 22.10.2025

Group project 016

Team members:

Thu Vu - 2507135

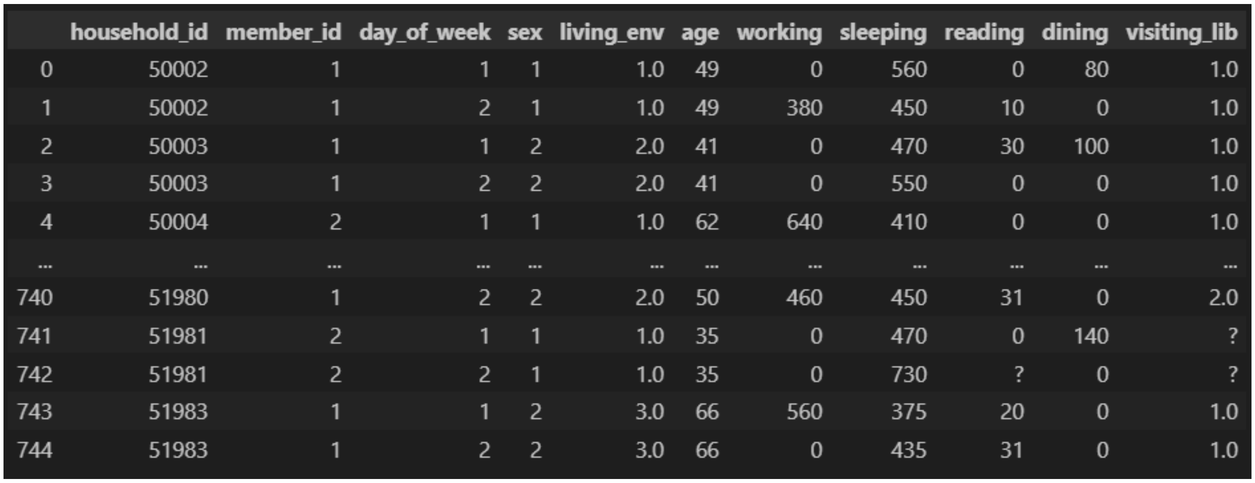
Ha Nguyen - 2514146

Table of Contents

1. Data preparation

1.1 Loading Data

The first step is loading and processing the data. The column names and their descriptions are provided in the **“**habit.txt**”** file. After defining the column names, the data is read into a DataFrame, as shown below:



	household_id	member_id	day_of_week	sex	living_env	age	working	sleeping	reading	dining	visiting_lib
0	50002	1	1	1	1.0	49	0	560	0	80	1.0
1	50002	1	2	1	1.0	49	380	450	10	0	1.0
2	50003	1	1	2	2.0	41	0	470	30	100	1.0
3	50003	1	2	2	2.0	41	0	550	0	0	1.0
4	50004	2	1	1	1.0	62	640	410	0	0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
740	51980	1	2	2	2.0	50	460	450	31	0	2.0
741	51981	2	1	1	1.0	35	0	470	0	140	?
742	51981	2	2	1	1.0	35	0	730	?	0	?
743	51983	1	1	2	3.0	66	560	375	20	0	1.0
744	51983	1	2	2	3.0	66	0	435	31	0	1.0

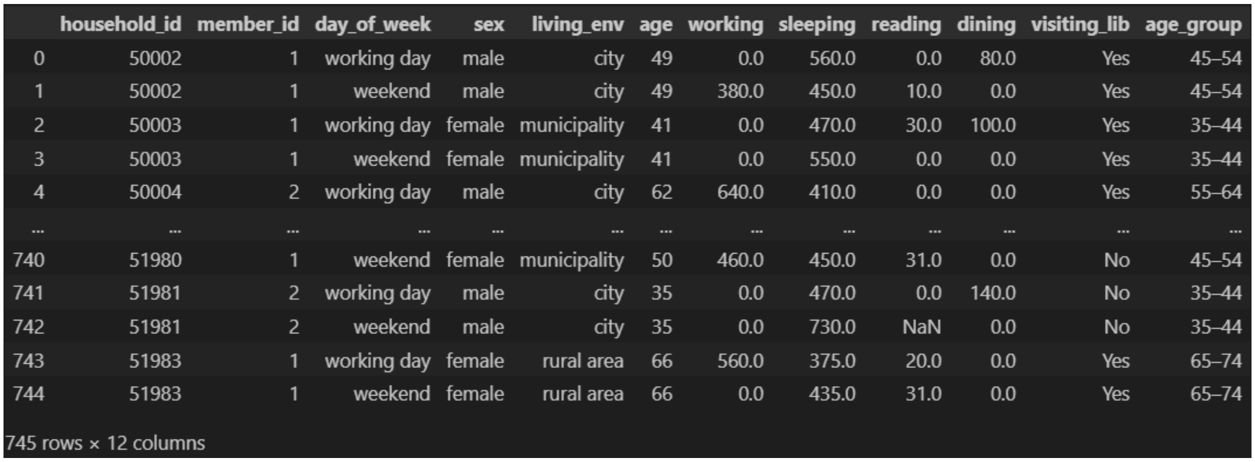
Picture 1. Raw Data Frame

1.2 Cleaning Data

After the initial processing, the data still requires cleaning before analysis. All activity-related columns (except *visiting_lib*) are converted to numeric values, with missing or invalid entries replaced by 0.

For the *visiting_lib* column, the data is converted to binary values: 1 (yes) and 2 (no). This adjustment was made because the processed data showed several cases where library visits lasted only 1 or 2 minutes, which seemed impossible.

Next, the categorical variables — *day_of_week*, *sex*, and *living_env* — are handled and updated in the DataFrame. Finally, an additional column for *age groups* is created based on defined bins and labels.



	household_id	member_id	day_of_week	sex	living_env	age	working	sleeping	reading	dining	visiting_lib	age_group
0	50002	1	working day	male	city	49	0.0	560.0	0.0	80.0	Yes	45–54
1	50002	1	weekend	male	city	49	380.0	450.0	10.0	0.0	Yes	45–54
2	50003	1	working day	female	municipality	41	0.0	470.0	30.0	100.0	Yes	35–44
3	50003	1	weekend	female	municipality	41	0.0	550.0	0.0	0.0	Yes	35–44
4	50004	2	working day	male	city	62	640.0	410.0	0.0	0.0	Yes	55–64
...	...	...	...	...	...	...	...	...	...	...	...	...
740	51980	1	weekend	female	municipality	50	460.0	450.0	31.0	0.0	No	45–54
741	51981	2	working day	male	city	35	0.0	470.0	0.0	140.0	No	35–44
742	51981	2	weekend	male	city	35	0.0	730.0	NaN	0.0	No	35–44
743	51983	1	working day	female	rural area	66	560.0	375.0	20.0	0.0	Yes	65–74
744	51983	1	weekend	female	rural area	66	0.0	435.0	31.0	0.0	Yes	65–74

745 rows × 12 columns

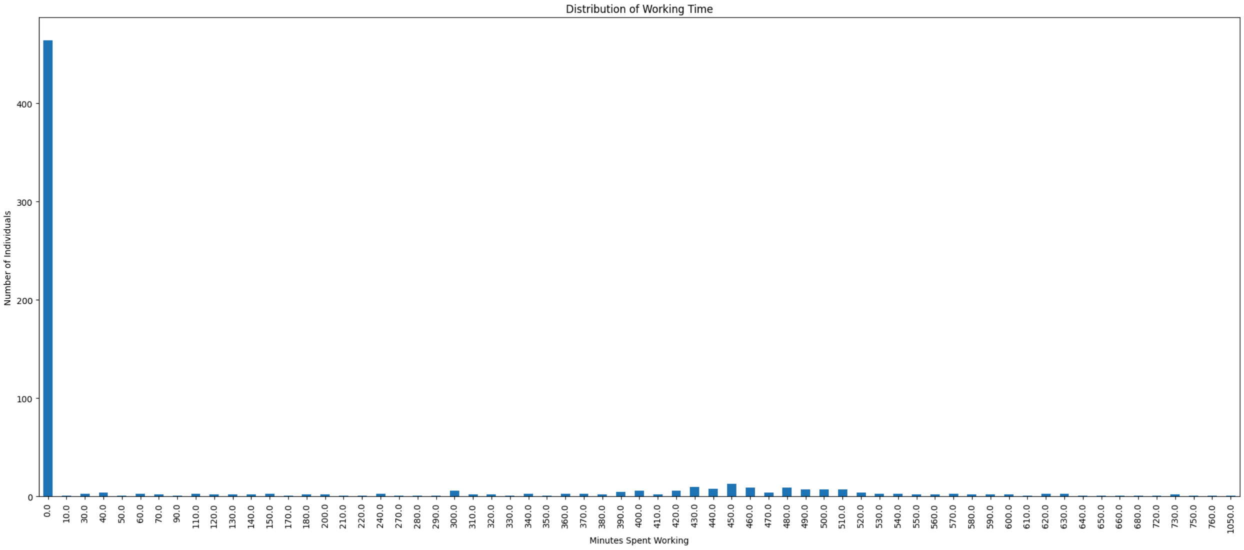
Picture 2. Cleaned Data Frame

2. Characteristics of Individuals in the Dataset

2.1 Characterize the individuals who spend time reading, sleeping, working, dining, and visiting the library using Descriptive Statistics

To provide an overview of how Finnish individuals allocate their time, descriptive statistics were calculated for five key daily activities: working, sleeping, reading, dining, and visiting libraries. These summaries describe the central tendency, variability, and distributional patterns of each activity, offering a general picture of typical and exceptional behavior in the population.

- Working:

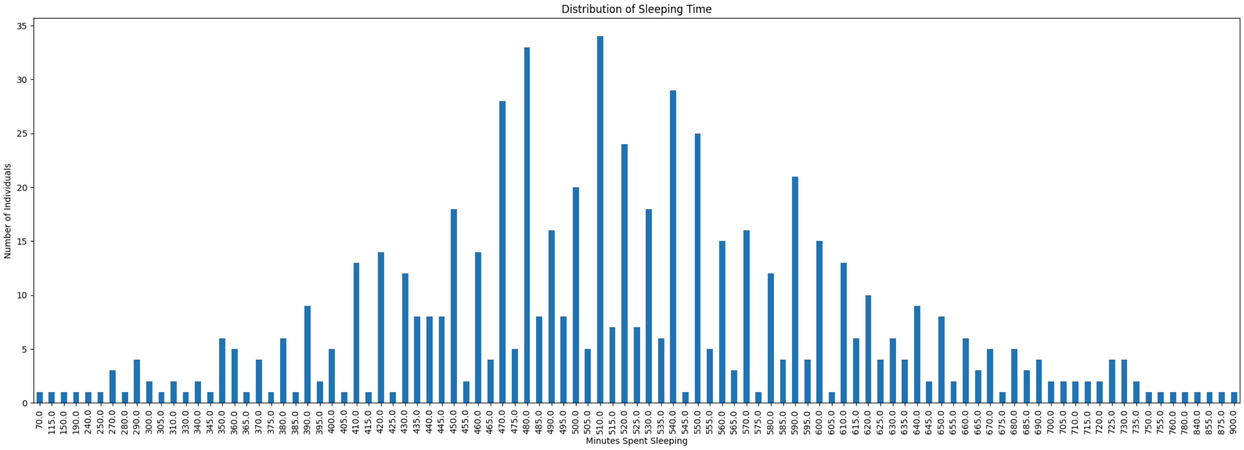


Picture 3. Distribution of Working Time

Most individuals spend very little time working on a typical day (median = 0), but the mean is 120 minutes because a smaller number of people work much longer, pulling the average up. The high standard deviation (208) and large range (1050) indicate high variability. The distribution is right-skewed since mean > median.

Working: mean=120.6 min, median=0.0, SD=208.1, range=1050.0, IQR=200.0

- Sleeping:

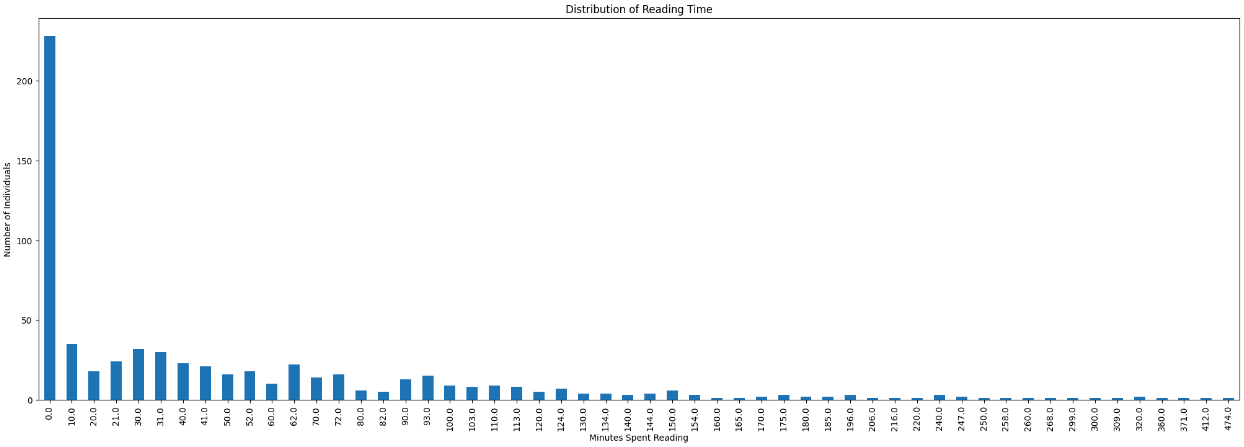


Picture 4. Distribution of Sleeping Time

Finnish individuals sleep on average 520 minutes per day (≈8.7 hours). The median is close to the mean, suggesting the distribution is roughly symmetric. The IQR of 115 minutes shows moderate variability in sleeping habits, while the large range (830 minutes) indicates some extreme cases (very short or long sleepers).

Sleeping: mean=520.0 min, median=515.0, SD=100.7, range=830.0, IQR=115.0

- Reading:

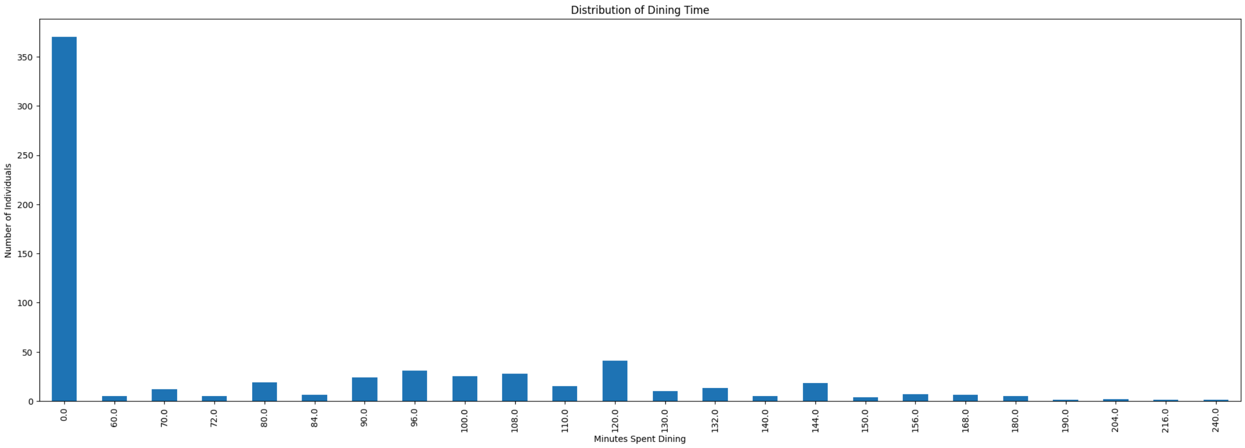


Picture 5. Distribution of Reading Time

On average, individuals spend 48 minutes reading per day. The median is lower than the mean, showing a right-skewed distribution, where some people read a lot more than most. The IQR of 70 minutes indicates moderate variability in typical reading time, but the very large range (404 minutes) suggests a few extreme cases.

Reading: mean=48.0 min, median=30.0, SD=64.0, range=404.0, IQR=70.0

- Dining:

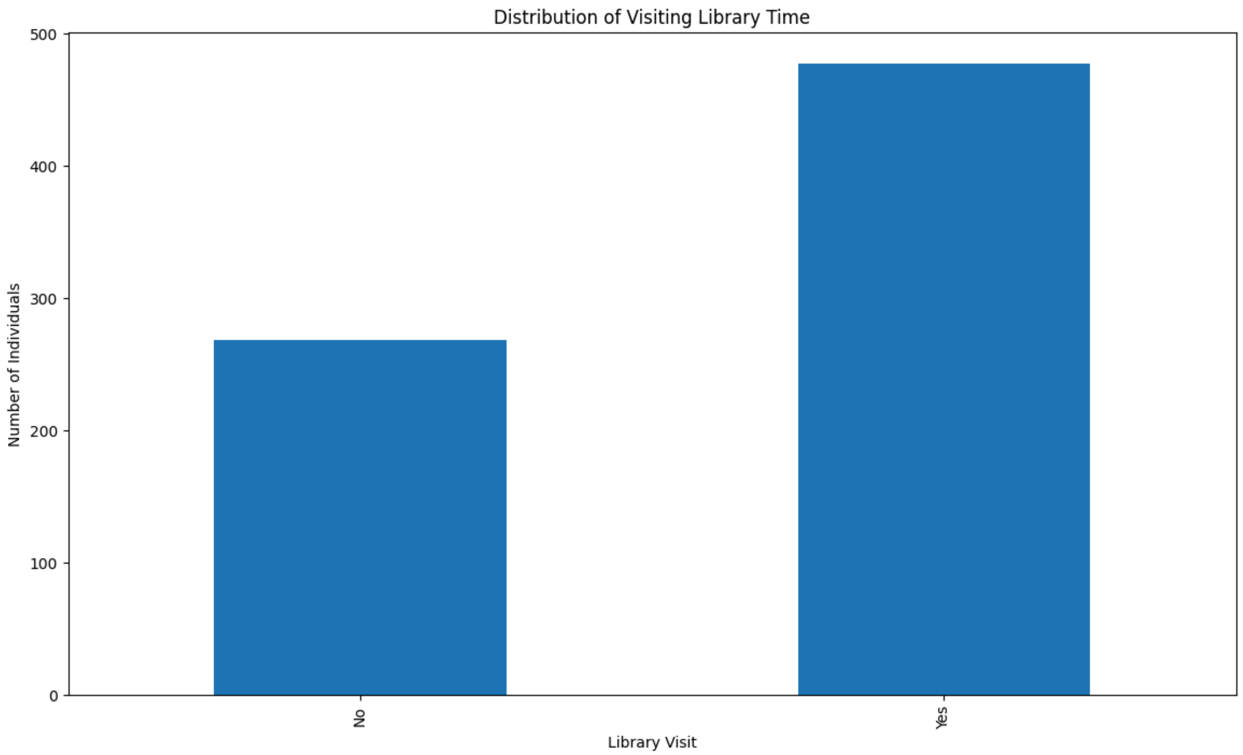


Picture 6. Distribution of Dining Time

Most individuals do not spend much time dining on a typical day (median = 0), but the mean is higher because some spend substantial time dining. The distribution is heavily right-skewed, with high variability (SD = 58.6) and a wide range of 240 minutes.

Dining: mean=48.5 min, median=0.0, SD=58.6, range=240.0, IQR=100.0

- Visiting Library:



Picture 7. Distribution of Visiting Library Time

About two-thirds of individuals report visiting the library. This categorical variable shows that library visits are fairly common, but a minority (36%) do not visit at all.

Visiting Library:

- Yes: 477 (64.0%)
- No: 268 (36.0%)

2.2 Characterize the individuals who spend time reading, sleeping, working and dining at the restaurant by different groups using Principal Component Analysis (PCA)

Explained variance ratio by each principal component:

- PC1: 0.3579 - explains 35.79% of the total variance

- PC2: 0.2690 - explains 26.90% of the total variance

Together, these two components explain approximately 62.7% of all variation in people's time-use behavior.

Loadings:

	PC1	PC2
Sleeping	-0.64	-0.3
Reading	-0.27	0.72
Dining	-0.08	-0.61
Working	0.71	-0.07

PC1:

- High loadings: working = +0.71
- Strong negative loadings: sleeping = - 0.64

Interpretation:

- PC1 reflects a “work–rest trade-off”.
- High PC1 scores: More time working, less time sleeping → Work-oriented pattern.
- Low PC1 scores: More time sleeping, less time working → Rest-oriented or low-workload pattern.

PC2:

- High loadings: reading = +0.72
- Strong negative: dining = −0.61, sleeping = −0.30

Interpretation:

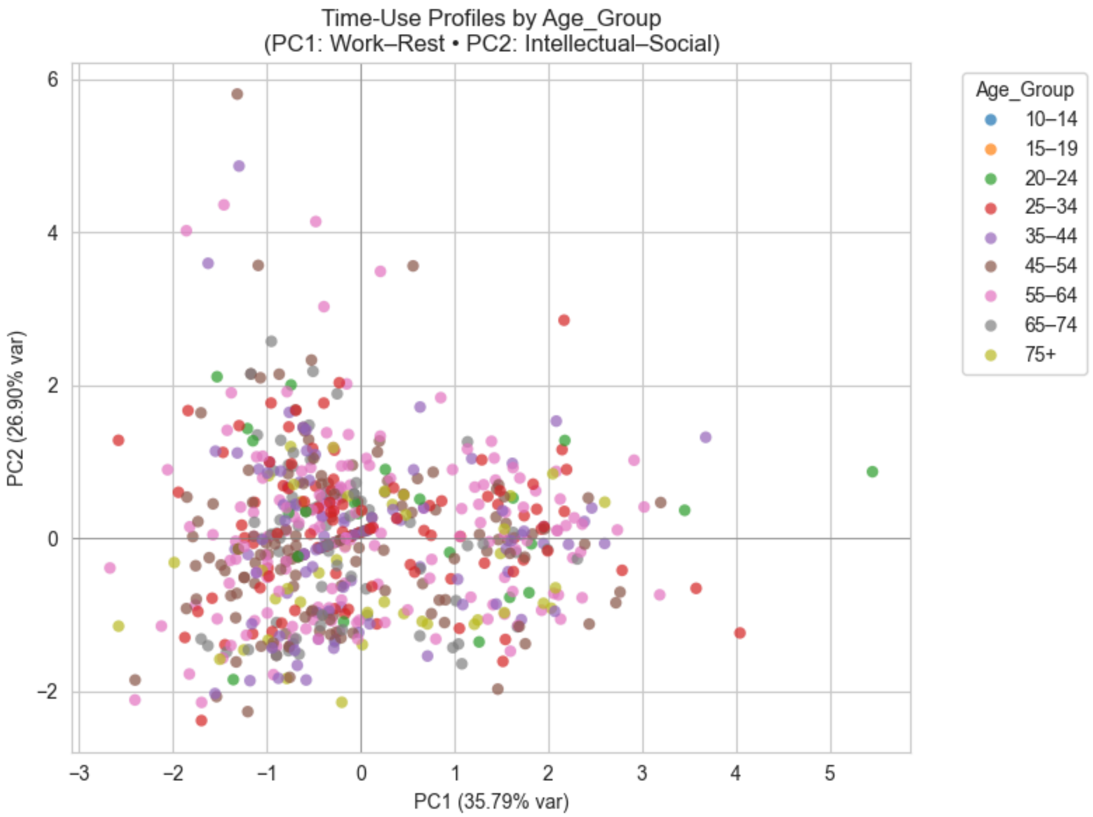
- PC2 captures an “intellectual vs. social/leisure” axis.
- High PC2 scores: More time reading, less dining at the restaurant or sleeping → Intellectual leisure
- Low PC2 scores: More time dining at the restaurant and sleeping → Social/relaxation pattern.

Combine both components:

Quadrant in (PC1, PC2) plane	Lifestyle interpretation
High PC1, High PC2	Work-focused and intellectually engaged (working + reading, little sleep).
High PC1, Low PC2	Work-focused but socially active (working + dining out, less reading).
Low PC1, High PC2	Rest-oriented but intellectually engaged (more sleep + reading, little work).
Low PC1, Low PC2	Rest-oriented and socially relaxed (more sleep + dining out).

Scatter plots by group variables: age group, living environment, sex, and day of week

a. Age group



Picture 8. PCA results based on Age Group

a1. Young Adults (20–24 years old):

- PC1: Clustered near center or slightly negative → Less work, more sleep.
- PC2: Clustered near center → Balanced between reading and dining at the restaurant.

This group tends to spend more time resting and reading, as many are not yet working full-time.

a2. Working-Age Adults (25 – 54 years old):

- PC1: Shift rightward → More work, less sleep.
- PC2: Centered between -1 and +2 → Mixed leisure preference.

This age range is engaged in active work life, with no dominant leisure pattern (balanced between reading and dining out).

a3. Older adults (55-74 years old):

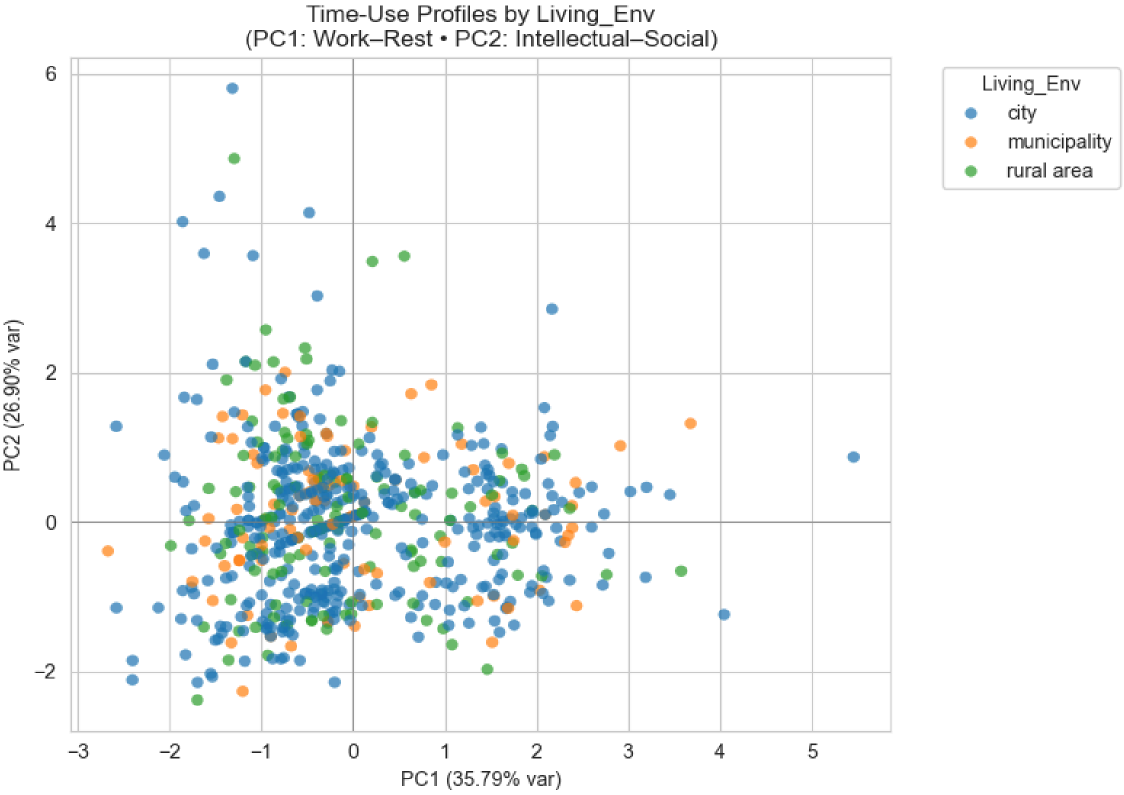
- Dots are concentrated closer to the center and slightly upper-left (high PC2, low PC1).

This age group focuses more on rest and intellectual engagement.

a4. Elderly (75+ years old):

This group falls within the lower-left quadrant (low PC1, low PC2), indicating a rest-oriented and social lifestyle.

b. Living environment



Picture 9. PCA results based on Living Environment

b1. City

- PC1: Slight right spread → More work.
- PC2: Several dots higher → More reading.

City residents are more work-focused and intellectually engaged.

b2. Municipality

- Orange dots are concentrated closer to the center

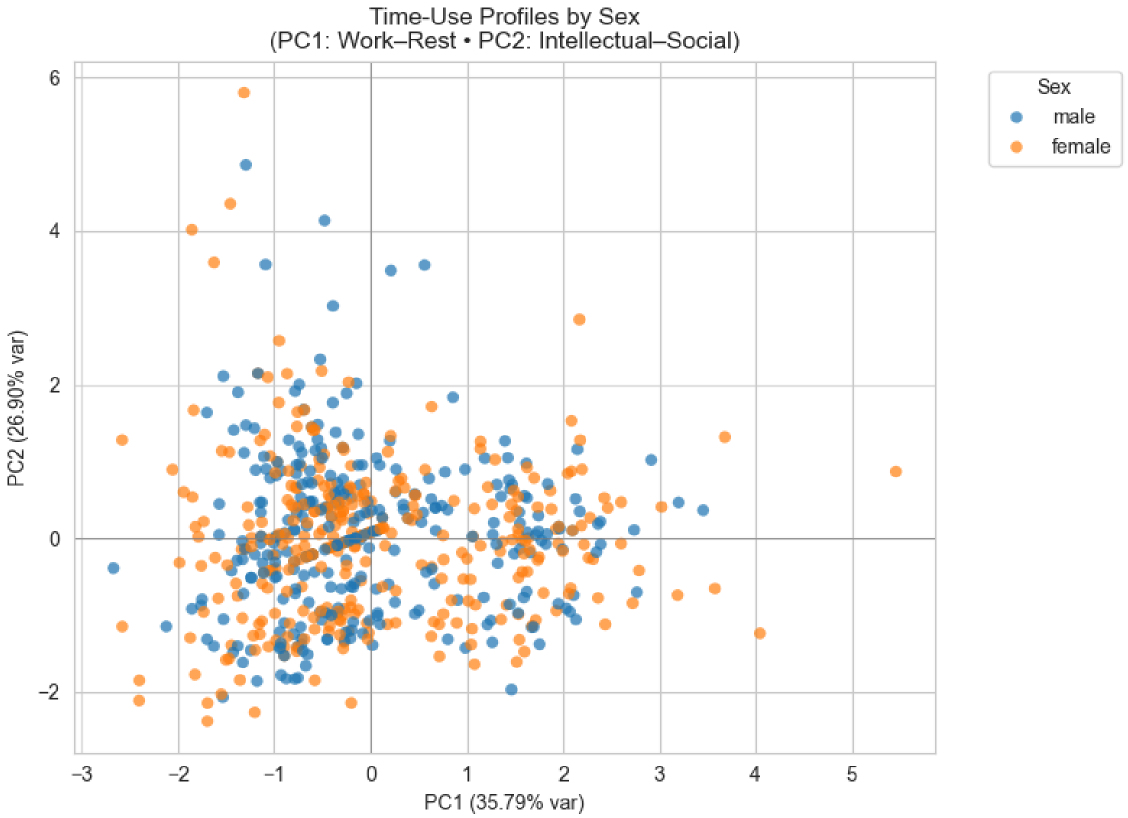
Municipality residents show a mixed lifestyle with no strong skew toward any activity.

b3. Rural area

Slight tendency toward lower left (low PC1, low PC2) → Less work, more rest and socializing.

Rural residents have a social and rest-oriented lifestyle, with a greater focus on dining at the restaurant.

c. Sex



Picture 10. PCA results based on Sex

c1. Male

- Slight spread to the right (higher PC1) → More work-oriented.
- Some points higher on PC2 → More reading.

Men tend to spend more time working and engaging in intellectual activities.

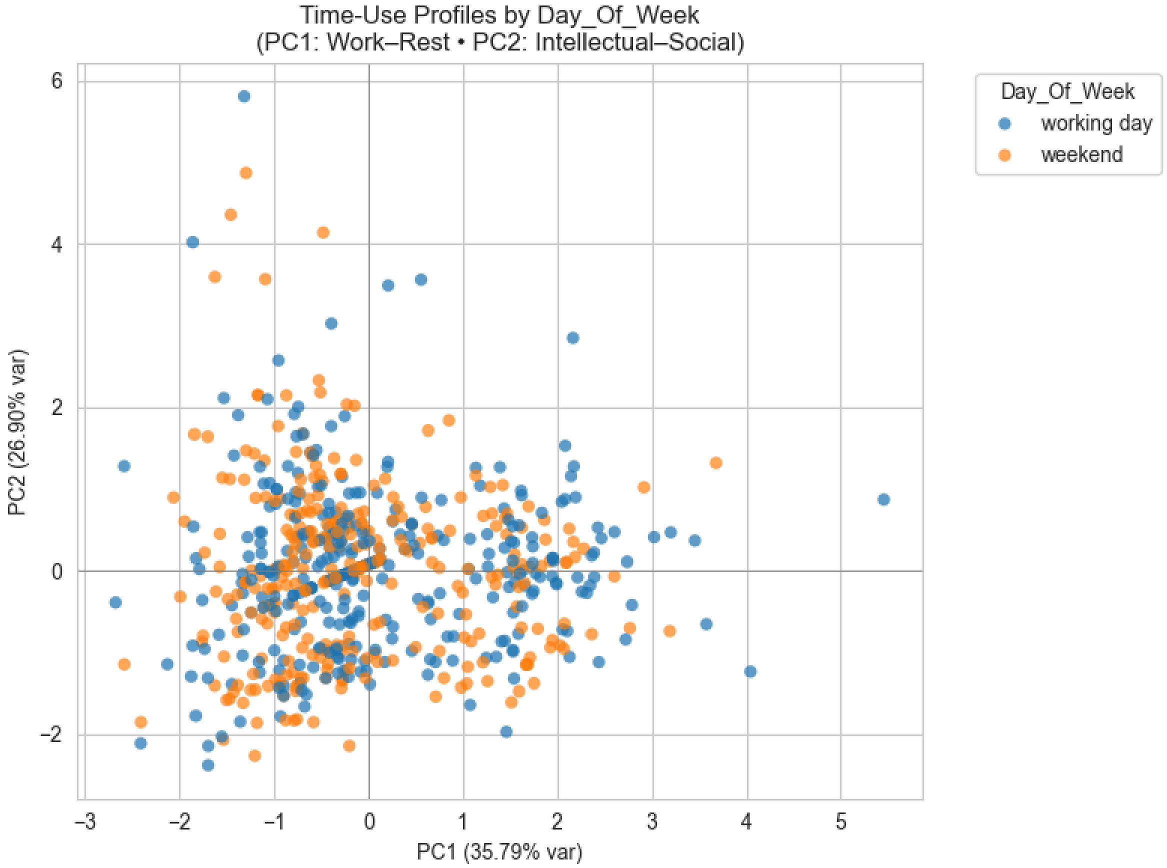
c2. Female

- Clustered more toward the center or left → Balanced work/rest.

- Some dots lower on PC2 → More dining out.

Women appear to spend less time at work, relatively more time in social activity, with a balanced or rest-oriented profile overall.

d. Day of week



Picture 11. PCA results based on Day of Week

d1. Working day

- Greater spread along PC1, with points extending further right and into extreme positions → More working, less sleeping.

- The spread across PC2 suggests a mixture of intellectual and social activities during after-work hours.

On working days, individuals’ routines are work-dominated.

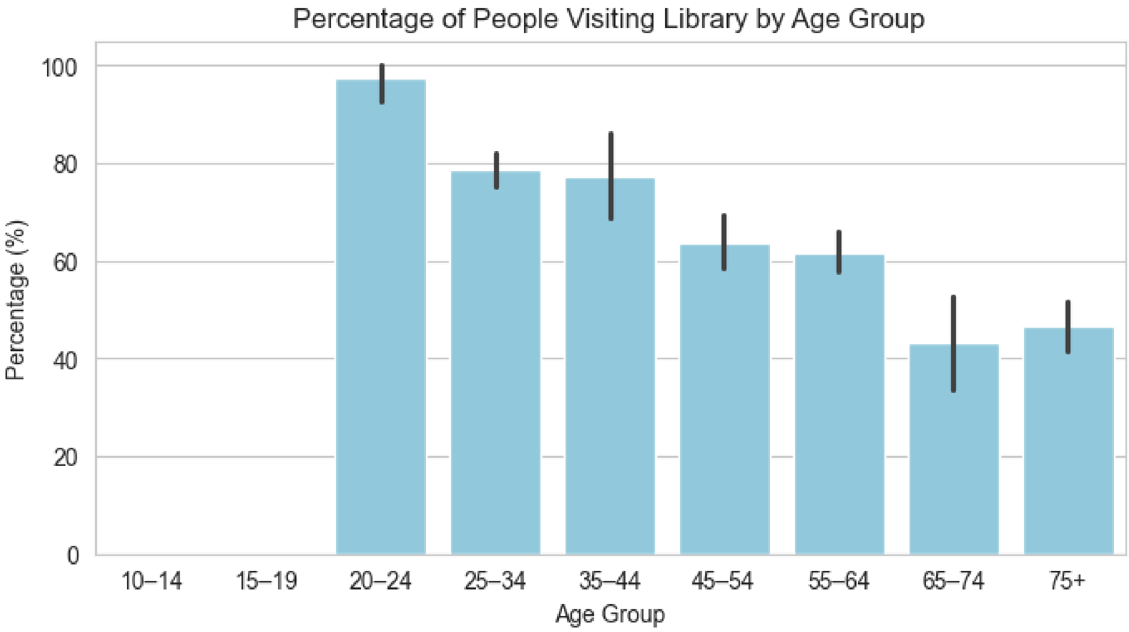
d2. Weekend

- Clustered left on PC1 → More resting, less working.
- Many appear higher on PC2 → More reading.
- Some extend downward (low PC2) → More dining out.

People rest more and engage in intellectual or social activities during weekends.

2.3 Characterize the individuals who spend time visiting the library by different groups

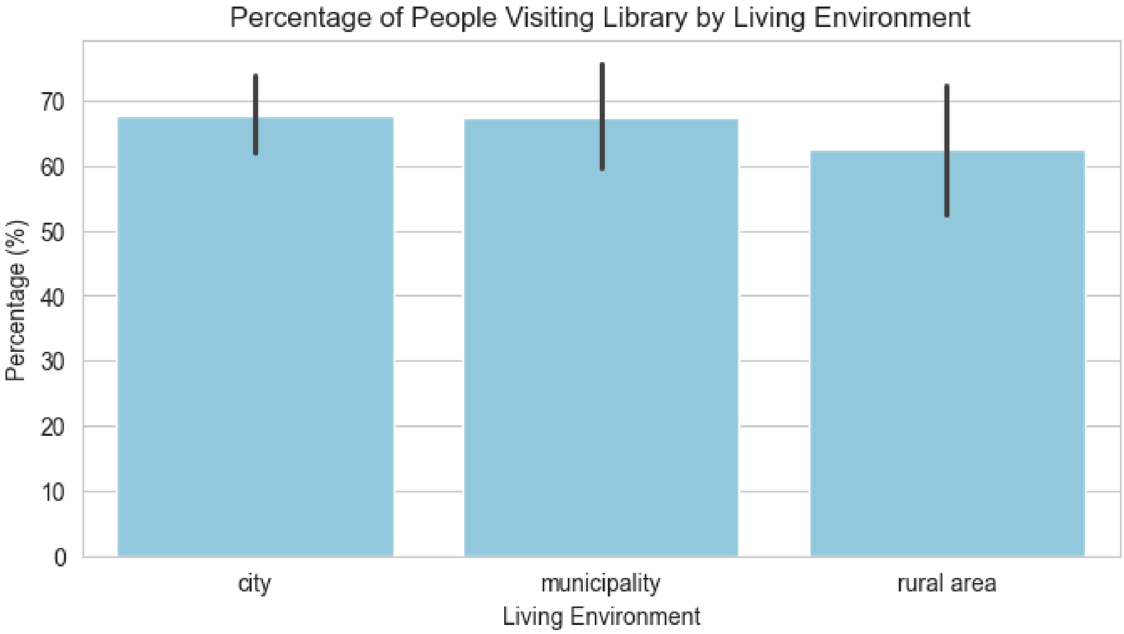
a. Age group



Picture 12. Percentage of People Visiting the Library by Age Group

The highest proportion of library visitors is in the 20–24 age group and the lowest rate of visiting libraries is in the 65-74 age group. There is a downward trend as age increases after 20-24 years old.

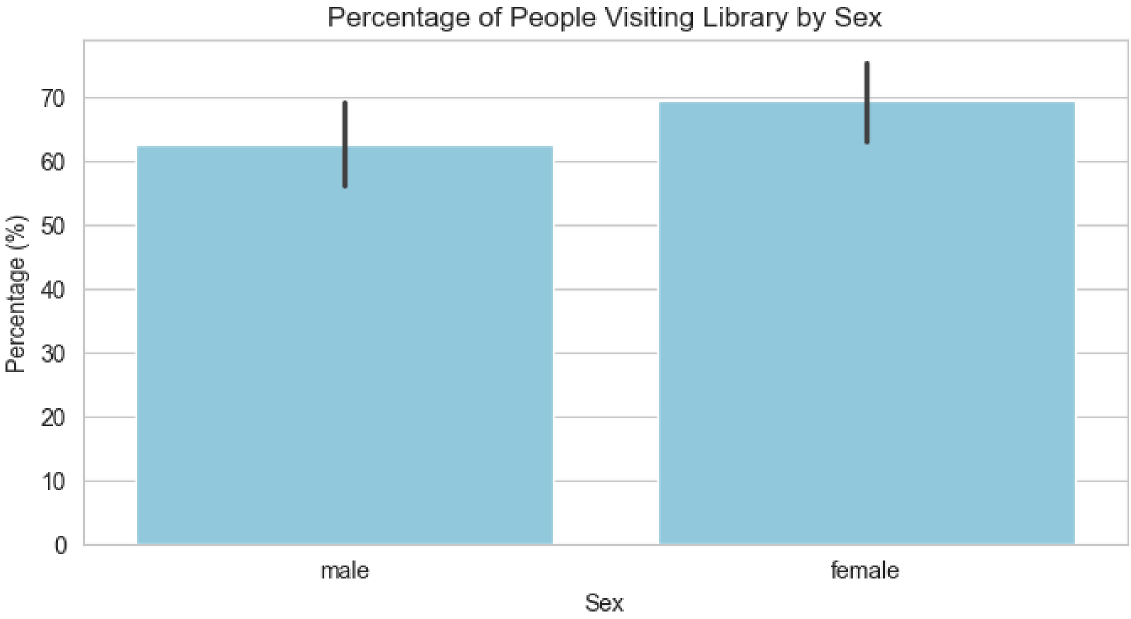
b. Living environment



Picture 13. Percentage of People Visiting the Library by Living Environment

City residents and municipality residents have the same proportion of people visiting the library, while rural areas show lower rates.

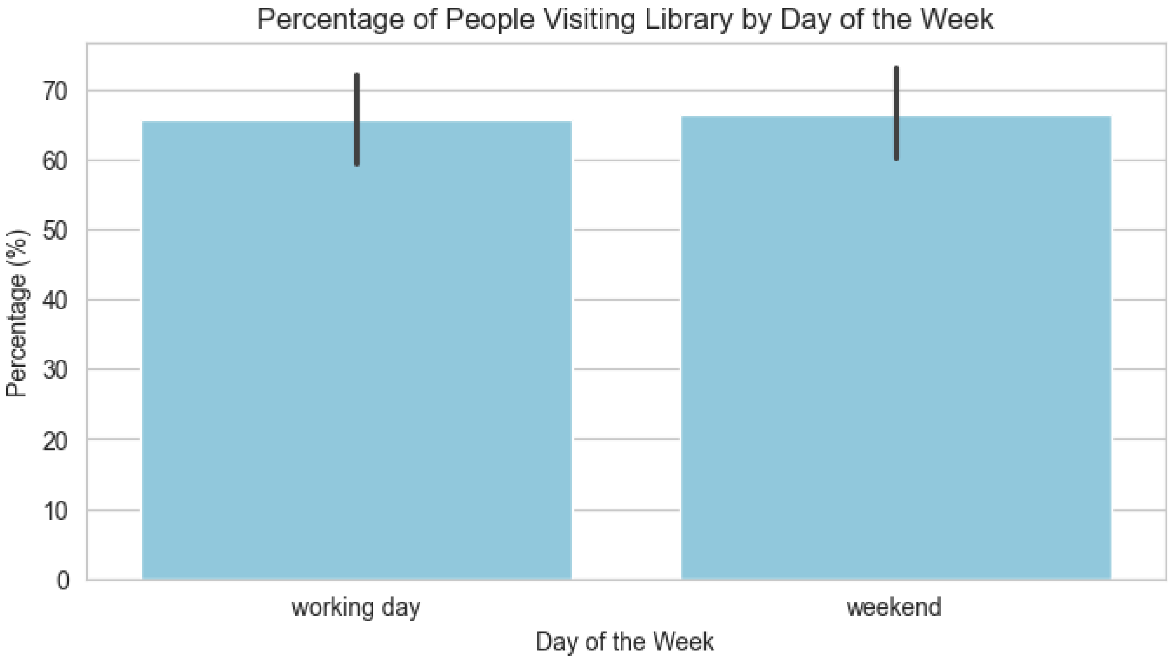
c. Sex



Picture 14. Percentage of People Visiting the Library by Sex

Females (≈70%) show a slightly higher rate of visiting libraries than males (≈63%).

d. Day of week



Picture 15. Percentage of People Visiting the Library by Day of Week

The rate of visiting libraries is very similar between working days and weekends (around 65–67%).

3. Average Daily Time Spent by Finnish Households on Each Activity

Time-use data were collected at the household level and expressed as daily mean minutes spent on each activity. For continuous variables (time spent), 95% t-based confidence intervals were calculated. For the binary outcome of library visitation, Wilson confidence intervals were computed to provide more accurate coverage.

The mean daily time and its 95% confidence interval that Finnish households allocate to activities, including sleeping, reading, dining at the restaurant and working are summarized below:

Activities	n_households	mean	lower	upper
sleeping	337	520.24	512.17	528.30
reading	337	48.60	42.15	55.04
dining	336	49.13	44.45	53.81
working	337	121.65	105.69	137.60

Interpretation:

- Sleeping: On average, households spent 520.24 minutes per day (approximately 8.7 hours) sleeping (95% CI: 512.17 – 528.30 minutes; *n* = 337). The narrow confidence interval suggests that sleep duration is relatively consistent across households.
- Reading: Households spend an average of 48.6 minutes per day reading (95% CI: 42.15-55.04 minutes; n=337). The moderate width of the CI suggests some variation in reading habits among households.
- Dining at restaurant: Households spend an average of 49.13 minutes per day on dining at restaurant activity (95% CI: 44.45-53.81 minutes; n=336). The narrow CI indicates relatively consistent eating habits across households, with little variation.

- Working: The mean time spent working was 121.65 minutes per day, or approximately 2.0 hours (95% CI: 105.69-137.60 minutes; n=337). The wide CI reflects large differences among households.

The library visiting proportion among Finnish households:

Activities	n_households	proportion	lower	upper
Visiting library	378	0.6402	0.5906	0.687

Interpretation:

- Visiting library: The library visiting proportion revealed that 64.02% of households visited a library (95% CI: 59.06% - 68.70%; n=378). This represents approximately two out of every three households, indicating that library use is fairly common among households. The confidence interval of approximately 8% demonstrates a reasonable precision estimate.

4. Differences in Activities Across Living Environments and Days of the Week in Finland

First, we tested the normality of each continuous activity variable (sleeping, reading, dining at the restaurant, and working) across living environments and days of the week using the Shapiro–Wilk test. The results indicated that none of the activity distributions were normally distributed ($p < 0.05$).

For differences across living environments (city, municipality, rural area), we applied the Kruskal–Wallis test.

For differences between days of the week (working day vs weekend), we used the Wilcoxon signed-rank test since data from the same participants were available for both conditions, making it a paired comparison between two related samples.

For the categorical activity of visiting the library, we used Pearson's Chi-Squared test to examine.

The differences between activities by living environment are summarized below:

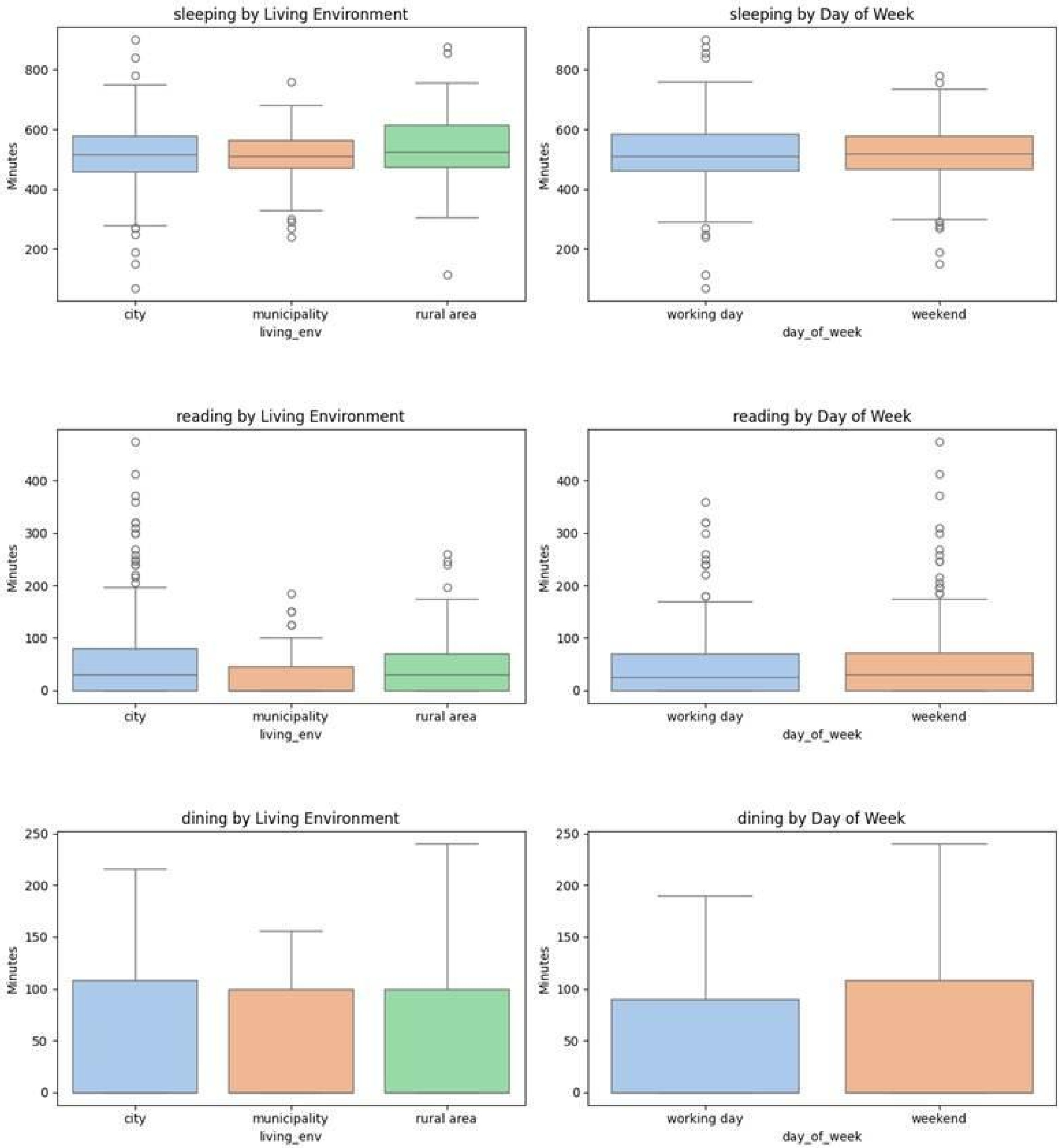
Statistical test	Activities	P value	Interpretation
Kruskal–Wallis test	Sleeping	0.0688	Sleep time seems relatively consistent across groups.
Kruskal–Wallis test	Reading	0.0004	Reading time differs significantly by living environment.
Kruskal–Wallis test	Dining at the restaurant	0.6624	Dining out habit is similar regardless of where people live.

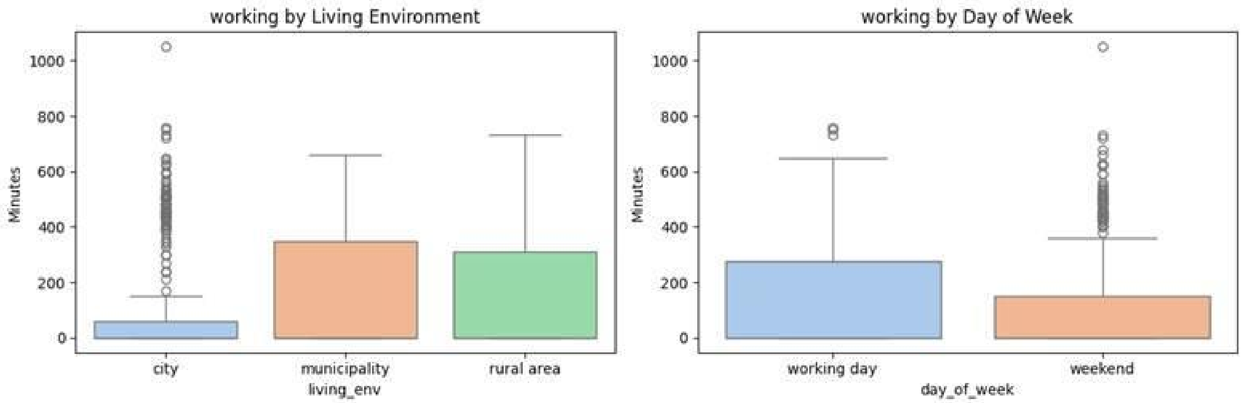
Kruskal–Wallis test	Working	0.1401	Working time shows no significant difference between environments.
Pearson’s Chi-Squared test	Visiting library	0.0258	The proportion of people visiting libraries varies by living environment.

The difference between activities by day of week is summarized below:

Statistical test	**Activities**	**P value**	**Interpretation**
Wilcoxon signed-rank test	Sleeping	0.6116	Sleeping time does not differ significantly between working days and weekends.
Wilcoxon signed-rank test	Reading	0.004	Reading time differs significantly by day of the week.
Wilcoxon signed-rank test	Dining at the restaurant	0.0019	Time spent dining at restaurants differs between working days and weekends.
Wilcoxon signed-rank test	Working	0.9204	Working time shows no significant difference between day types.
Pearson’s Chi-Squared test	Visiting library	0.7739	The proportion of people visiting libraries does not differ between working days and weekends.

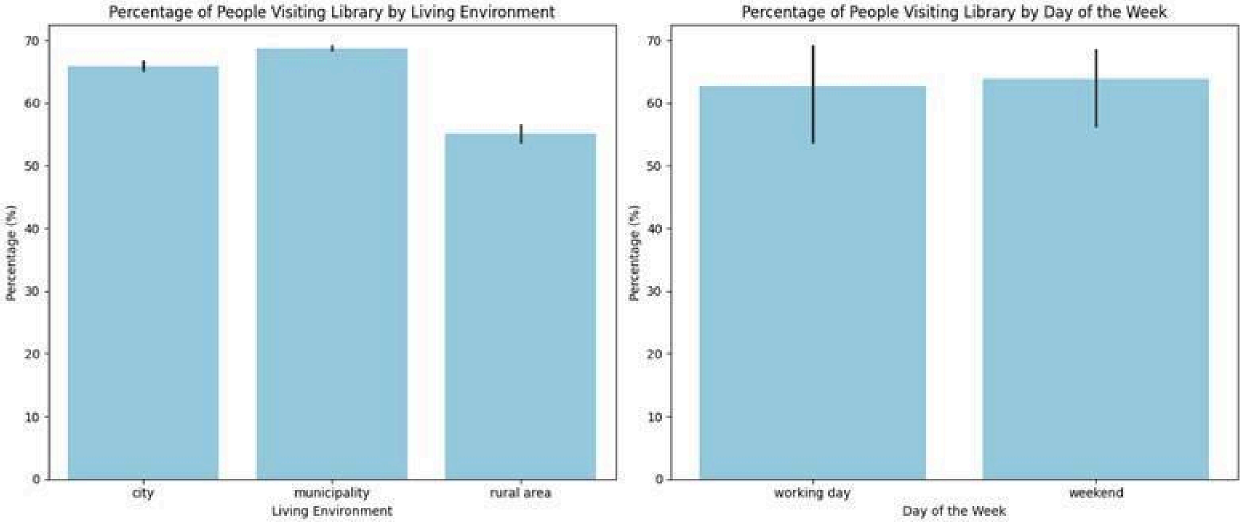
Visualization for activities by living environment and day of week:


sleeping by Living Environment
Minutes
800
600
400
200
city
municipality
rural area
living_env
sleeping by Day of Week
Minutes
800
600
400
200
working day
weekend
day_of_week
reading by Living Environment
Minutes
400
300
200
100
0
city
municipality
rural area
living_env
reading by Day of Week
Minutes
400
300
200
100
0
working day
weekend
day_of_week
dining by Living Environment
Minutes
250
200
150
100
50
0
city
municipality
rural area
living_env
dining by Day of Week
Minutes
250
200
150
100
50
0
working day
weekend
day_of_week



Picture 16. Boxplot for activities by Living Environment and Day of Week

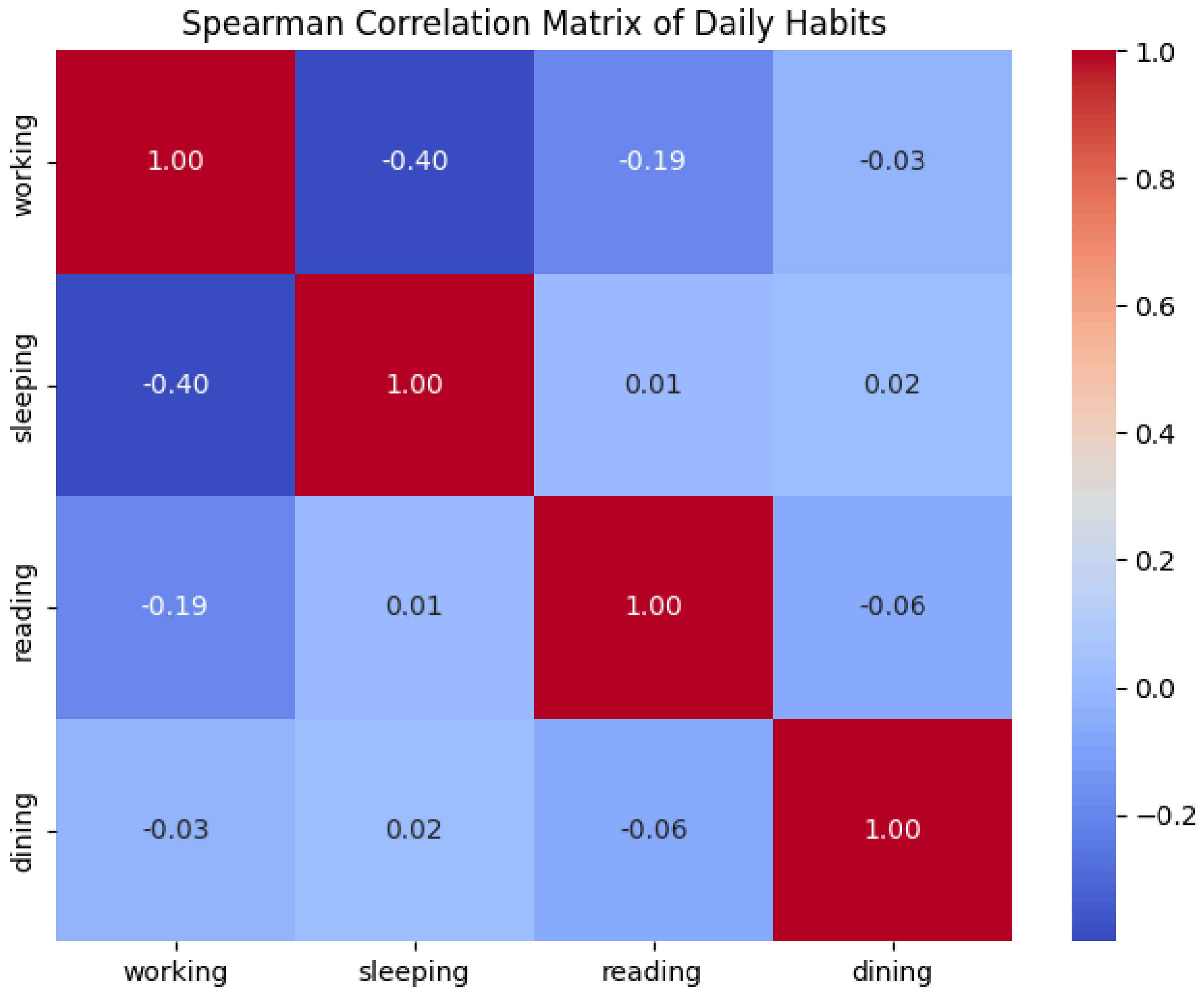
Visualization for the proportion of visiting the library by living environment and day of week:



Picture 17. Bar plot of the proportion of Visiting Library by Living Environment and Day of Week

5. Associations Between Activities in the Finnish Population

- Correlation between working, sleeping, reading, and dining:



Picture 18. Heatmap of Spearman Correlation Matrix of Daily Habits

Working and sleeping have a moderate negative correlation (r = −0.40) → Individuals who spend more time working tend to sleep less.

Working and reading are weakly negatively correlated (r = −0.19), suggesting that more working time is slightly associated with less reading.

Dining is almost uncorrelated with other activities (|r| < 0.06), indicating that meal time is relatively independent of work, rest, and leisure.

All other correlations are very small, implying weak associations between most non-work activities.

- Correlation between working, sleeping, reading, and dining grouped by visiting the library:

Visiting library		**Sleeping**	**Reading**	**Dining**	**Working**
YES	**Sleeping**	1.000000	0.031662	0.048499	-0.373168
	Reading	0.031662	1.000000	-0.076862	-0.223734
	Dining	0.048499	-0.076862	1.000000	0.011526
	Working	-0.373168	-0.223734	0.011526	1.000000
NO	**Sleeping**	1.000000	0.005825	0.012146	-0.410273
	Reading	0.005825	1.000000	-0.060686	-0.172934
	Dining	0.012146	-0.060686	1.000000	-0.049272
	Working	-0.410273	-0.172934	-0.049272	1.000000

Spearman correlation analysis revealed a moderate negative relationship between working and sleeping time (ρ ≈ −0.4) in both library visitors and non-visitors. Other activity pairs showed only weak or negligible associations, suggesting that time spent on working and sleeping are the most interrelated daily habits, regardless of library visitation.

Principal Component Analysis (PCA) supported these findings, identifying two main behavioral dimensions that explained 62.7 % of the variance. The first component (PC1) represented a Work–Rest balance, contrasting working (+0.71) with sleeping (−0.64), and the second component (PC2) captured a Leisure–Routine dimension, driven mainly by reading (+0.72) and dining (−0.61). Together, these results suggest that Finnish individuals’ daily habits are structured primarily around balancing work and rest, and to a lesser extent, around the contrast between leisure and routine activities.