TKO7093 – STATISTICAL DATA ANALYSIS

Project Report

Group 160

Ayana Kotuwegoda Guruge – 2406865

Sheheryar Wahidi – 2413773

Yagya Yadav – 2409273

Abstract

The purpose of this study is to have a closer understanding and analysis of daily time usage of individuals in Finland with certain activities collected such as working, sleeping, reading, dining out and visiting the library. The goals are to study the characterization of the population, making calculations of average times spent on the above-mentioned activities, comparisons with the gender, living environments as well as the day of the week such as the weekdays and the weekends. It is also instructed to find the relationships between these activities. The techniques employed include descriptive analysis, estimation with confidence intervals, hypothesis testing, as well as multivariate analysis. The results of this study have shown that sleeping is the most important factor in time usage, followed by other activities. The difference between weekdays and weekends is very low. The effect of the living environment on reading is also significant. The results of this study provide a systematic view of the time usage of the people of Finland. Statistical theories have been applied to real-life scenarios.

Contents

1. Introduction

This report was prepared for the Head of Research to analyze survey data about the daily habits of individuals in Finland. Our team functioned as a Data Scientist group to investigate how much time people spend on activities like work, sleep, and leisure, and whether these patterns change depending on the person's background or living environment.

The main aims of this analysis are as follows:

- Describe the demographic characteristics in the dataset.
- Estimate how much time individuals in households spend on average times on daily activities.
- Examine time use is different from living environments (city, municipality and rural) and by days of the week (weekend and weekdays).
- Examine and explore associations between the activities using correlation and principal component analysis.

1.1. Tools used

To manage our workflow and shared coding, we used the following tools:

- **GitHub:** For shared coding and version control.
- **Visual Studio:** For sandbox coding and local testing.
- **OneDrive and Word:** For collaborative report writing.
- **Zoom and WhatsApp:** For daily updates and group discussions.

2. Methodologies used and justification

For our analysis, we designed a workflow that moves from essential data cleaning to advanced multivariate patterns, ensuring every step aligns with the statistical principles covered in the course.

To start, we handled the Data Preparation through Household-Level Aggregation. Because the raw data contained multiple entries per household, we had to group them to avoid pseudo-replication. This ensures that each observation is independent. A fundamental assumption for the tests we ran later.

We then moved into Descriptive Statistics to get a baseline of the data. Since time-use variables are often plagued by skewness and zero-inflation (lots of people spending zero minutes on certain tasks), we prioritized medians and boxplots over simple means. To quantify our uncertainty, we calculated Confidence Intervals for the averages, sticking to the parametric methods emphasized in our lectures rather than using Bayesian or bootstrap alternatives.

For our comparative analysis, we chose tests based on the specific "behavior" of the data:

- **Temporal Comparisons:** When checking Weekday vs. Weekend differences, we used **Welch's t-test**. It's more reliable than a standard t-test because it's robust against **heteroscedasticity** (unequal variances) and different sample sizes.
- **Environmental Comparisons:** To see if living location (City vs. Rural) mattered, we used the Kruskal–Wallis test. Since our data wasn't normally distributed, this non-parametric approach was the safest bet. When we found significant results, we ran Dunn's test for post-hoc analysis to pinpoint which groups differed while keeping our Type I error rate under control.
- **Library Usage:** Because visiting the library is an occasional event, we treated it as a binary outcome and used a Chi-square (x^2) test to check for independence between groups.

Finally, we used Multivariate Exploration to see how activities interact. We started with Correlation Analysis to find overlapping behaviors and finished with Principal Component Analysis (PCA). PCA was our go-to for dimensionality reduction, helping us visualize broad "time-use profiles" without needing to force the data into rigid clusters.

By choosing these specific methods, our group ensured that the results are not just descriptive, but statistically defensible.

3. Data Preparation and Cleaning

3.1. Followed steps.

- Data loaded from habits.data into a DataFrame matching the variables.
- Missing values are checked and not removed.
- Codes are converted into readable labels.
- Checking all demographic values for sensible values.
- Checking the validity of the ranges.
- Replacing the invalid values of the variables accordingly.

Following were the initial findings.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 745 entries, 0 to 744
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   kohde   745 non-null    int64
 1   jasen   745 non-null    int64
 2   pvknro  745 non-null    int64
 3   sp      745 non-null    int64
 4   ASALUE  745 non-null    float64
 5   IKAL1   745 non-null    int64
 6   A1      741 non-null    object
 7   A2      737 non-null    object
 8   A3      733 non-null    object
 9   A4      736 non-null    object
 10  A5      703 non-null    float64
dtypes: float64(2), int64(5), object(4)
memory usage: 64.2+ KB
```

	kohde	jasen	pvknro	sp	ASALUE	IKAL1	A1	A2	A3	A4	A5
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

3.2. Addressing missing values

We checked for missing values in each column before performing any transformations. The results were as follows:

- kohde: 0
- jasen: 0
- sp: 0
- ASALUE: 0
- IKAL1: 0
- A1 (Sleep): 4
- A2 (Work): 8
- A3 (Reading): 12
- A4 (Restaurant): 9
- A5 (Library): 42

We decided that missing values are not always 0 minutes spent; they can be recording errors. Therefore, we kept them as NaN values. Using errors="coerce" during our conversion ensured that calculations remain correct because incorrect or missing values are now treated uniformly as NaN without creating new, artificial data.

3.3. Data type conversions and cleaning

We checked all demographic values for sensible values and verified the validity of ranges. Based on this audit, we made the following corrections:

1. Age Group (IKAL1): We found values over 20, even though the habits.txt file says the range should be 1-9. We assumed these were actual ages and sorted them back into the 1-9 groups.
2. Library (A5): We assumed the float values in this column were hours spent in the library. We converted these to 1 (Yes) and created an A5_binary variable to depict whether a person visited the library or not.
3. Mapping: We converted sp, pvp, ASALUE, and IKAL1 into categorical data types so they are treated correctly in descriptive statistics and statistical tests.

Below is the summary of converted data types with the new columns.

Note: The age_group and the A5_binary did not replace the IKAL1 and A5 respectively to keep the original integrity of the data file.

[7]:

	kohde	jasen	pvknro	sp	ASALUE	IKAL1	A1	A2	A3	A4	A5	\
0	50002	1	1	1	1.0	49	0.0	560.0	0.0	80.0	1.0	
1	50002	1	2	1	1.0	49	380.0	450.0	10.0	0.0	1.0	
2	50003	1	1	2	2.0	41	0.0	470.0	30.0	100.0	1.0	
3	50003	1	2	2	2.0	41	0.0	550.0	0.0	0.0	1.0	
4	50004	2	1	1	1.0	62	640.0	410.0	0.0	0.0	1.0	

	age_group	A5_binary
0	45-54	1.0
1	45-54	1.0
2	35-44	1.0
3	35-44	1.0
4	55-64	1.0

4. Task 1: Characterizing Individuals in the data

We analyzed the individual-level data to understand the demographic makeup of the survey participants.

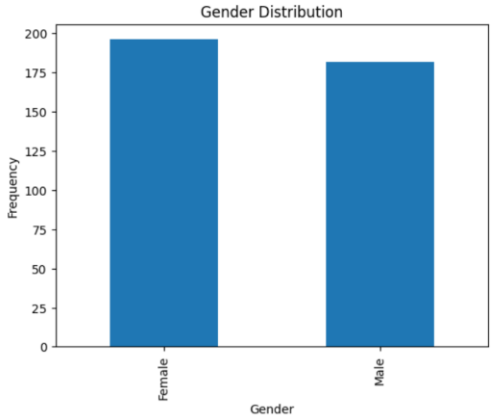
3.1 Unique Individuals

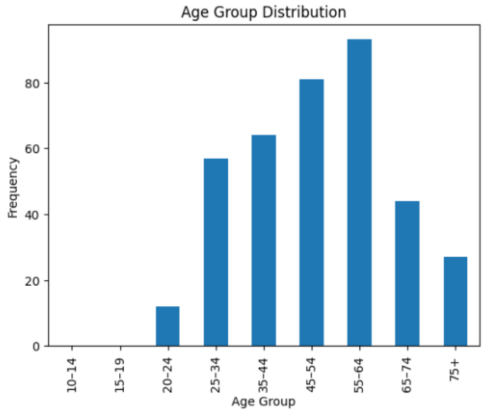
Although the data has 745 observations, these are person-day entries. By identifying individuals through kohde and jasen IDs, we found there are 378 unique individuals in the dataset. The higher count of 745 is due to the same individuals providing both weekend and weekday entries.

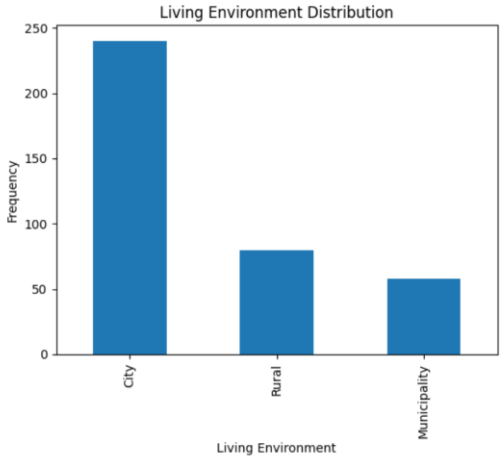
3.2 Demographic Distribution

The analysis shows a balanced and diverse group:

- Gender: The distribution is balanced, with females making up about 52% and males 48%.
- Age: Most individuals belong to middle-aged groups, specifically 45–64 years. Younger and oldest age groups are less represented.
- Environment: Most individuals live in cities, with smaller proportions in rural areas or municipalities.







3.3 Household Size

We examined household size by counting individuals per kohde. The results show that every household in the dataset consists of a single individual (Mean = 1.0). This means that household-level averages will correspond directly to individual observations in our next tasks.

4. Task 2: Estimation of average daily time spent on activities

The goal of this task was to figure out how much time Finnish households actually spend on different activities in a typical day. Since our data prep showed that each household in this study is just one person, these averages represent individual habits across the country.

4.1 Methods and Calculation

We calculated the daily averages in two ways. First, we looked at a simple unweighted mean. Then, we calculated a weighted average to account for the fact that a week has five weekdays and two weekend days. This weighting is important because it prevents the weekend data from having too much influence on the overall daily average.

To make our estimates more reliable, we used bootstrap resampling to generate 95% confidence intervals. This method helps us understand the range where the true population average is likely to fall, rather than just giving up a single number.

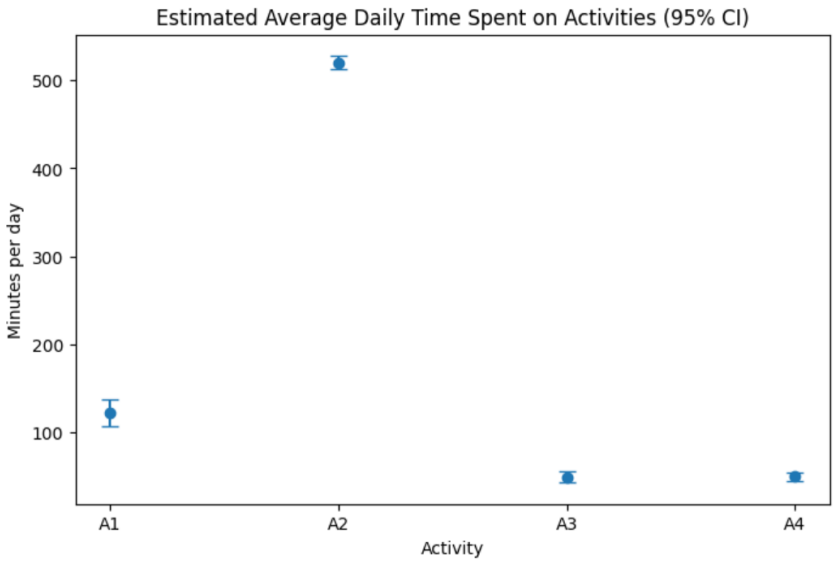
4.2 Statistical Estimates of Time Use

Our analysis showed that sleeping and working are by far the biggest parts of the day for people in the survey.

A 95% confidence interval is used. This is assuming the following:

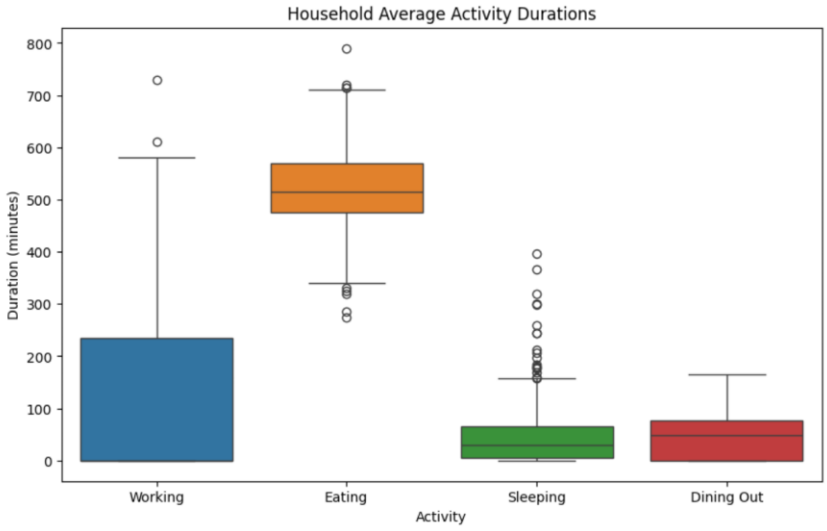
- Central limit theorem where large enough sample size is considered.
- And taking household independently.

Activity	**Weighted Mean (min/day)**	**95% Bootstrap CI (min)**
A1 (Sleeping)	517.2 minutes	509.0 – 526.0
A2 (Working)	120.6 minutes	108.3 – 138.8
A3 (Reading)	45.9 minutes	42.6 – 56.4
A4 (Restaurant)	44.9 minutes	40.6 – 49.2



4.3 Visualizing the Distributions

Looking at the boxplot for household activities, we can see quite different patterns for each habit.



- Sleeping: This is the most consistent activity. The box is tight and the mean is high, showing that almost everyone in the sample gets a similar amount of sleep, around 8.6 hours.

- Working: This shows the most variation. There are many zeros, meaning many people didn't work on their survey day, but there are also high outliers for people who worked long hours.
- Reading and Restaurant visits: Both are zero-inflated, which just means a lot of people didn't do these activities at all during the survey period.
- Library visits: This activity was so rare that we only had 10 households with valid duration data. This resulted in an extremely wide and unreliable confidence interval of 0 to 81 minutes, so we must interpret the library results with caution.

4.4 Conclusion on Weighting

When comparing our weighted and unweighted results, we found that for things like sleep and work, the difference was exceedingly small. This suggests these habits are stable throughout the week. However, the differences were more noticeable for leisure activities like reading, which seem to fluctuate more depending on whether it is a weekday or a weekend.

5. Task 3 – Differences Between Groups

This is to examine the time spent on activities between groups in Finland.

The differences that are analyzed will be as follows:

- Between weekdays and weekends.
- Between living environments.
- The analysis is done at the household level to get independent observations.

Parametric statistical methods are applied to compare group means. Although individual activity durations are skewed and include many zero values, inference is based on household-level means. Given the moderate sample size, the Central Limit Theorem supports the use of mean-based inference.

5.1. Methods used

- Independent samples t-tests to compare weekends and weekdays.
- One-way ANOWA to compare living environments.
- 95% confidence intervals to quantify the estimation uncertainty.

5.2. Weekday vs. Weekend

```
Weekday Data Sample:
   kohde   pvknro     A1     A2    A3     A4
0  50002  Weekday    0.0  560.0   0.0   80.0
2  50003  Weekday    0.0  470.0  30.0  100.0
4  50004  Weekday  640.0  410.0   0.0    0.0
6  50005  Weekday    0.0  540.0  40.0    NaN
8  50006  Weekday    0.0  540.0   0.0   90.0

Weekend Data Sample:
   kohde   pvknro     A1     A2    A3     A4
1  50002  Weekend  380.0  450.0  10.0    0.0
3  50003  Weekend    0.0  550.0   0.0    0.0
5  50004  Weekend    0.0  550.0  72.0  108.0
7  50005  Weekend    0.0  550.0  52.0  108.0
9  50006  Weekend    0.0  530.0  62.0    0.0

T-test Results between Weekday and Weekend Activities:
  Activity  t-statistic   p-value
0       A1     0.324070  0.745988
1       A2    -0.834543  0.404282
2       A3    -1.657364  0.097940
3       A4    -2.775214  0.005682
```

- Differences between weekdays and weekends were assessed using Welch's independent samples t-tests, which are robust to unequal variances.
- No statistically significant differences were found for working (A1), sleeping (A2), or reading (A3) ($p > 0.05$).
- Dining out (A4) shows a statistically significant difference ($t = -2.78$, p is around 0.006), indicating higher average dining time on weekends.
- Overall, routine daily activities remain stable across the week, while leisure-related activities vary by day type.

5.3. Differences in Activity Time by Living Environment

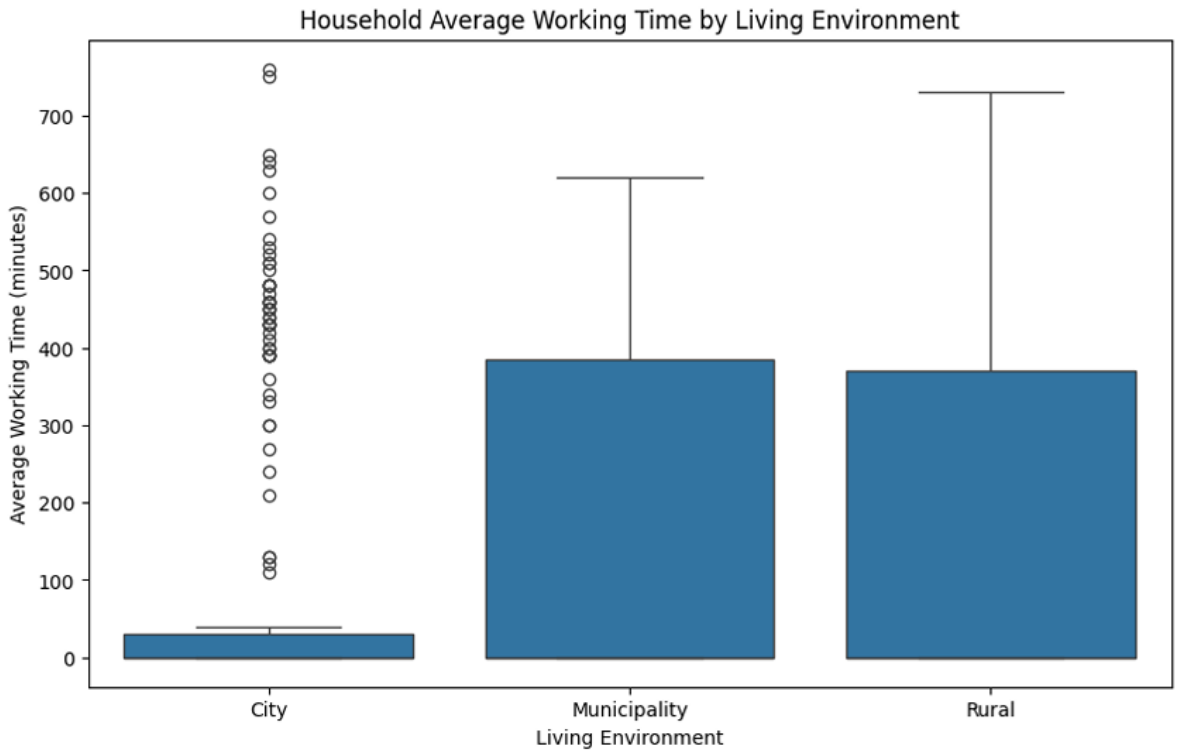
Because activity durations are skewed and contain many zeros, non-parametric methods were used. Kruskal–Wallis tests were applied to compare activities across living environments.

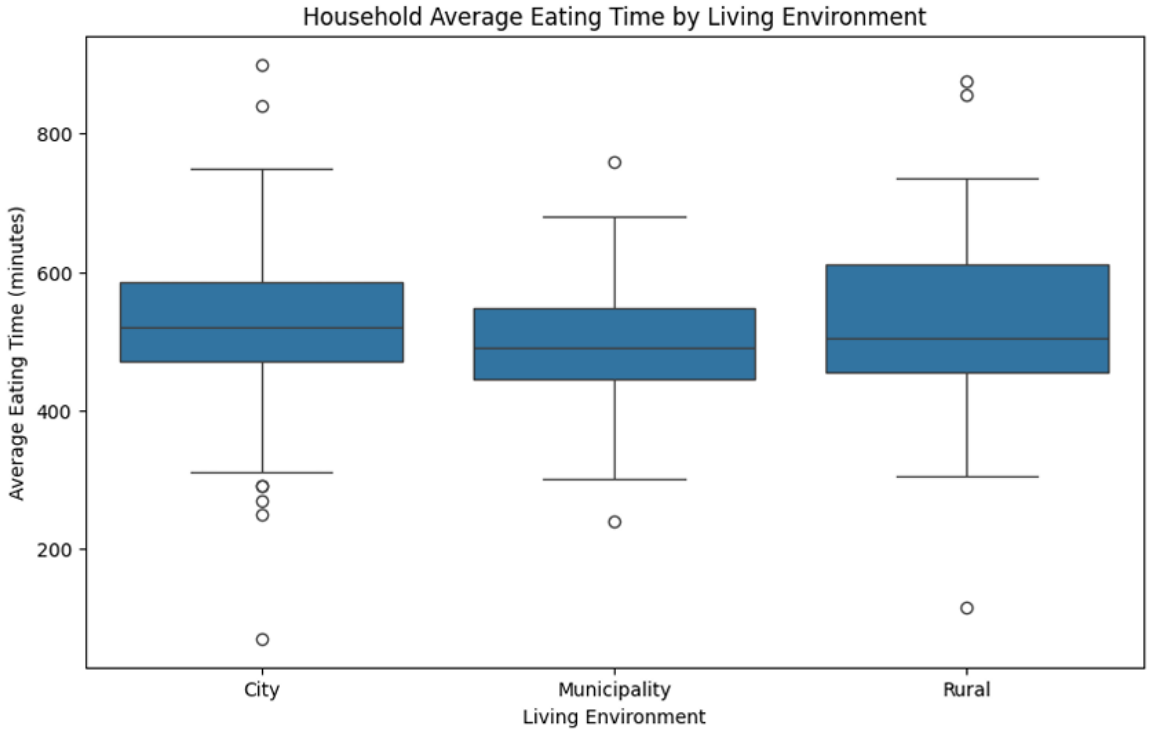
- A statistically significant difference was found for reading time (A3).
- No significant differences were observed for working, sleeping, or dining out.

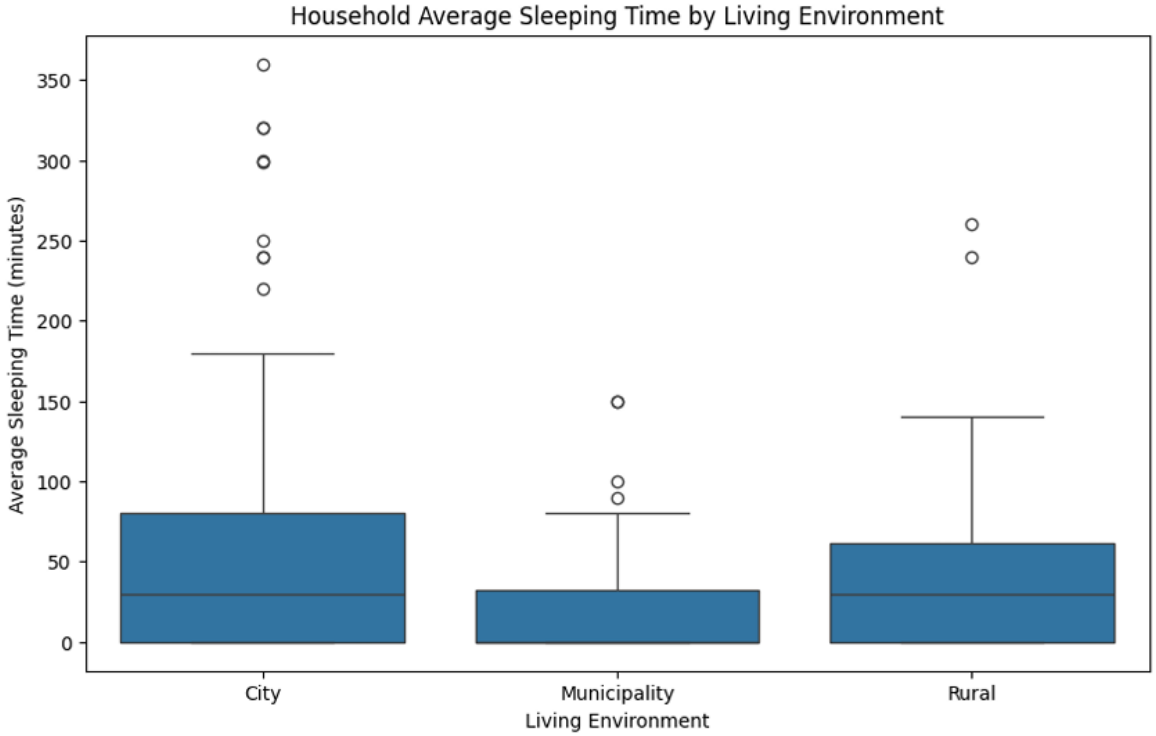
A post-hoc Dunn's test revealed that the difference in reading time is driven by city versus municipality households.

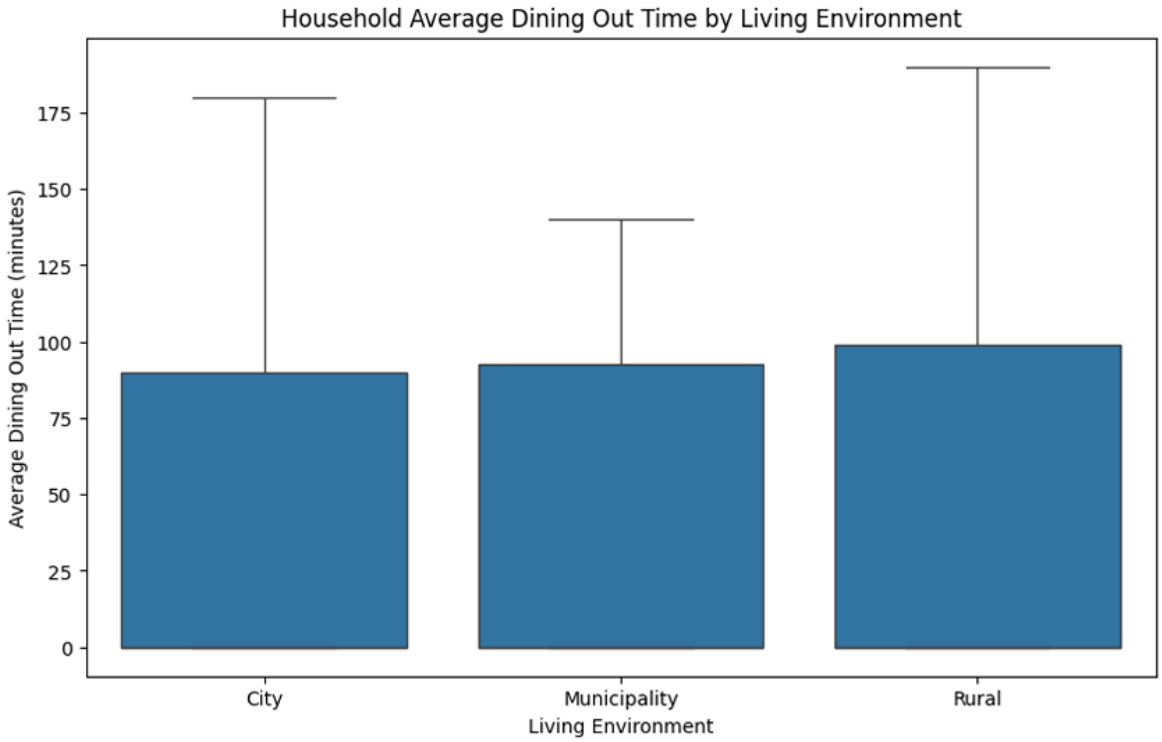
Kruskal-Wallis ANOVA Results across Living Environments:

	Activity	H-statistic	p-value
0	A1	5.931665	0.051518
1	A2	2.686724	0.260967
2	A3	8.899015	0.011684
3	A4	0.611458	0.736586




Household Average Eating Time by Living Environment
Average Eating Time (minutes)
800
600
400
200
City
Municipality
Rural
Living Environment


Household Average Sleeping Time by Living Environment
Average Sleeping Time (minutes)
350
300
250
200
150
100
50
0
City
Municipality
Rural
Living Environment




```
Dunn's Post-hoc Test Results for A3 (Sleeping Time):
                    City  Municipality    Rural
City            1.000000      0.026893  1.00000
```


[redacted]

```
Municipality  0.026893      1.000000  0.12023
Rural         1.000000      0.120230  1.00000
```

5.4. Visiting the library (A5)

Since A5 was converted to a binary data type in the data cleaning and preparation stage due to the inconsistencies, it cannot be compared with means or ANOVA-type methods.

Therefore, for further examination of how it has been differed across groups, categorical methods like cross-tabulations and Pearson's chi-square test are used for comparison rates.

- Between living environments.
- Between weekdays and weekends.

```
Crosstabulation of Visiting the library by living environment:
A5_binary     1.0  2.0
ASALUE
City          311  128
Municipality   79   35
Rural          87   63

Chi-square Test Results by living environments:
Chi-square statistic:  8.584
Degrees of freedom: 2
p-value:  0.0137

Crosstabulation of Visiting the library by Weekday/Weekend:
A5_binary  1.0  2.0
pvknro
Weekday    239  116
Weekend    238  110

Chi-square Test Results by Weekday/Weekend:
Chi-square statistic:  0.049
Degrees of freedom: 2
p-value:  0.8243
```

6. Task 4: Associations between activities in the Finnish population

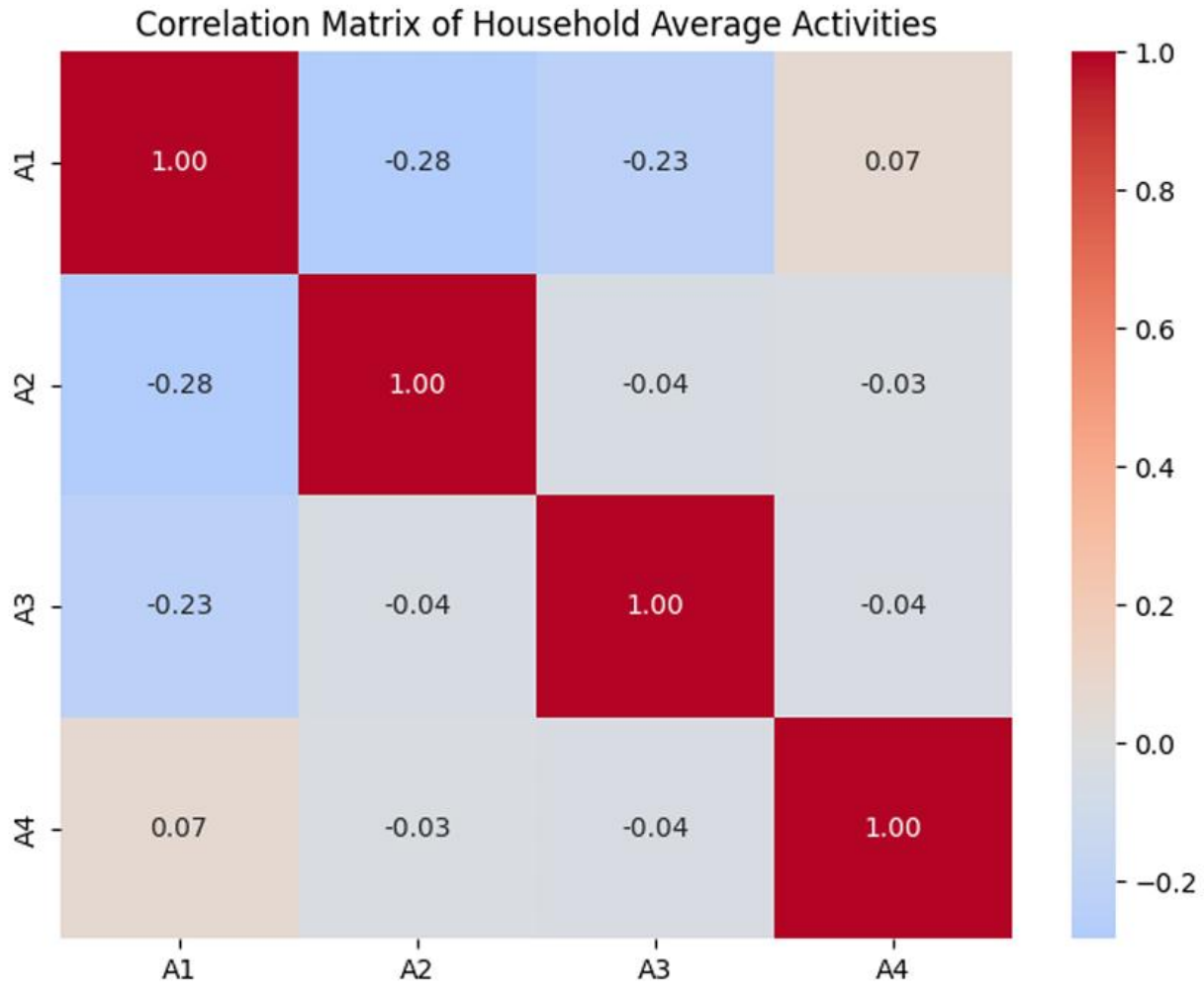
Examining which daily activities are associated with others. The goal is to identify patterns such as whether working time is related to dining out or sleeping etc.

6.1. Data used

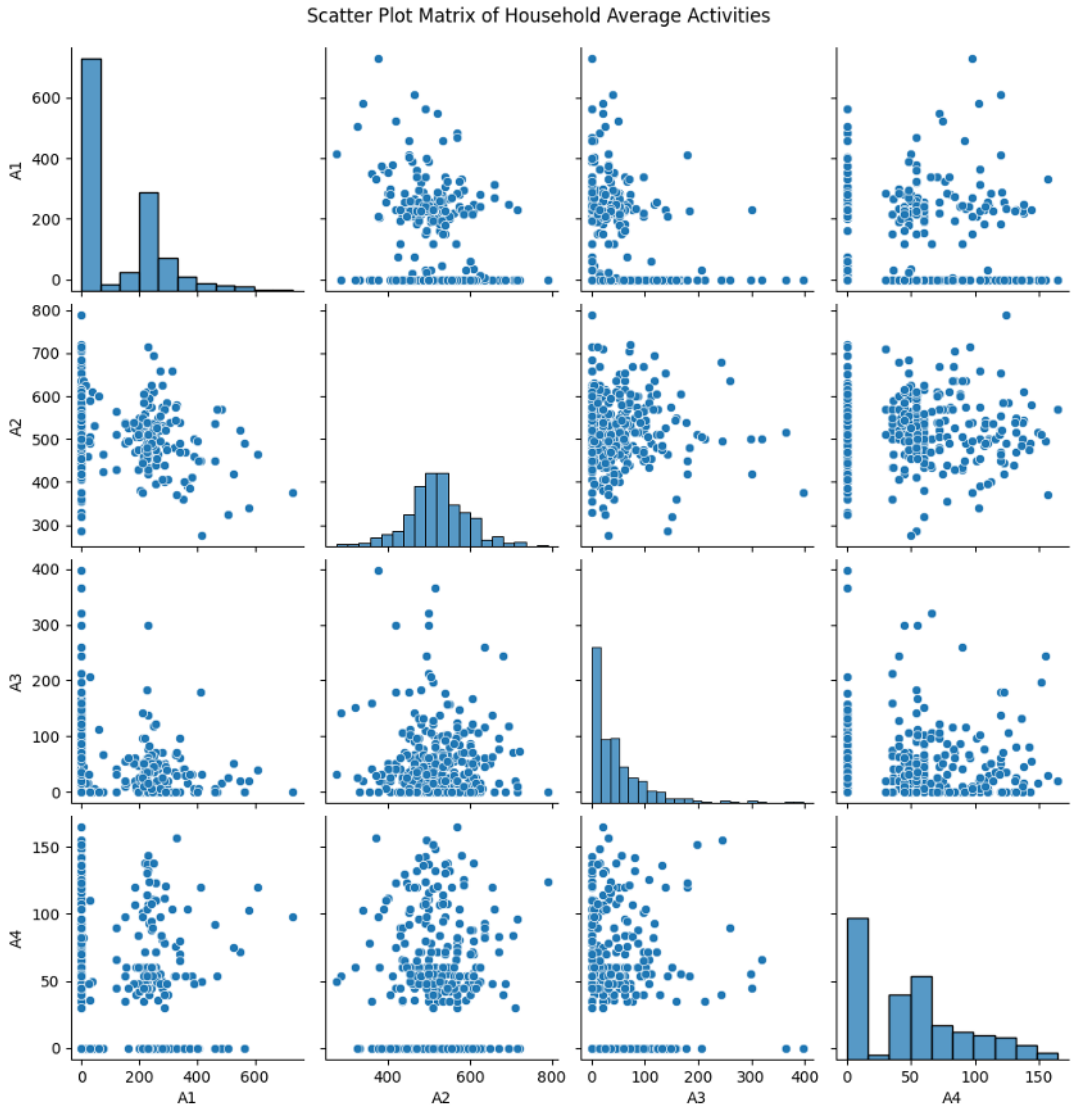
- Durations of household-level activities.
- Activities

Correlation analysis shows weak associations between most activities, indicating that time spent on one activity does not strongly predict time spent on others.

- Working time is moderately negatively correlated with sleeping and reading.
- Dining out shows near-zero correlation with all other activities.



A scatter plot matrix further illustrates these weak and mostly linear relationships.



Above correlation can be observed as follows:

- The scatter plot matrix visualizes pairwise relationships between household-level average activity durations.
- Most activity pairs show diffuse, cloud-like patterns, confirming the weak correlations observed numerically.
- Strong zero-inflation is visible for working (A1), reading (A3), and dining out (A4), indicating that many households do not engage in these activities daily.
- Sleeping time (A2) shows a more concentrated distribution, reflecting its regular and essential nature.

- No clear linear or non-linear structures dominate the plots, suggesting that activity behaviors vary independently across households.

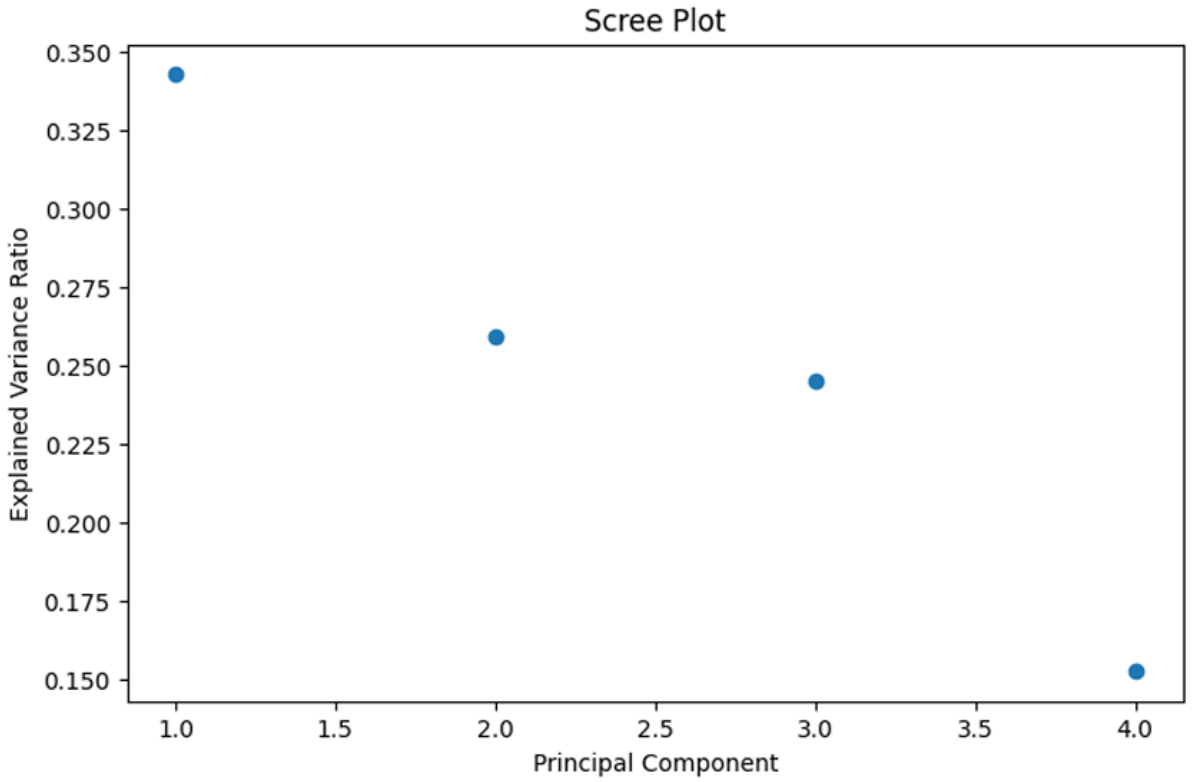
Overall, the low correlations suggest that each activity captures a different aspect of daily time use, and hence the use of multivariate techniques such as PCA to study underlying activity patterns further.

7. Principal Component Analysis (PCA)

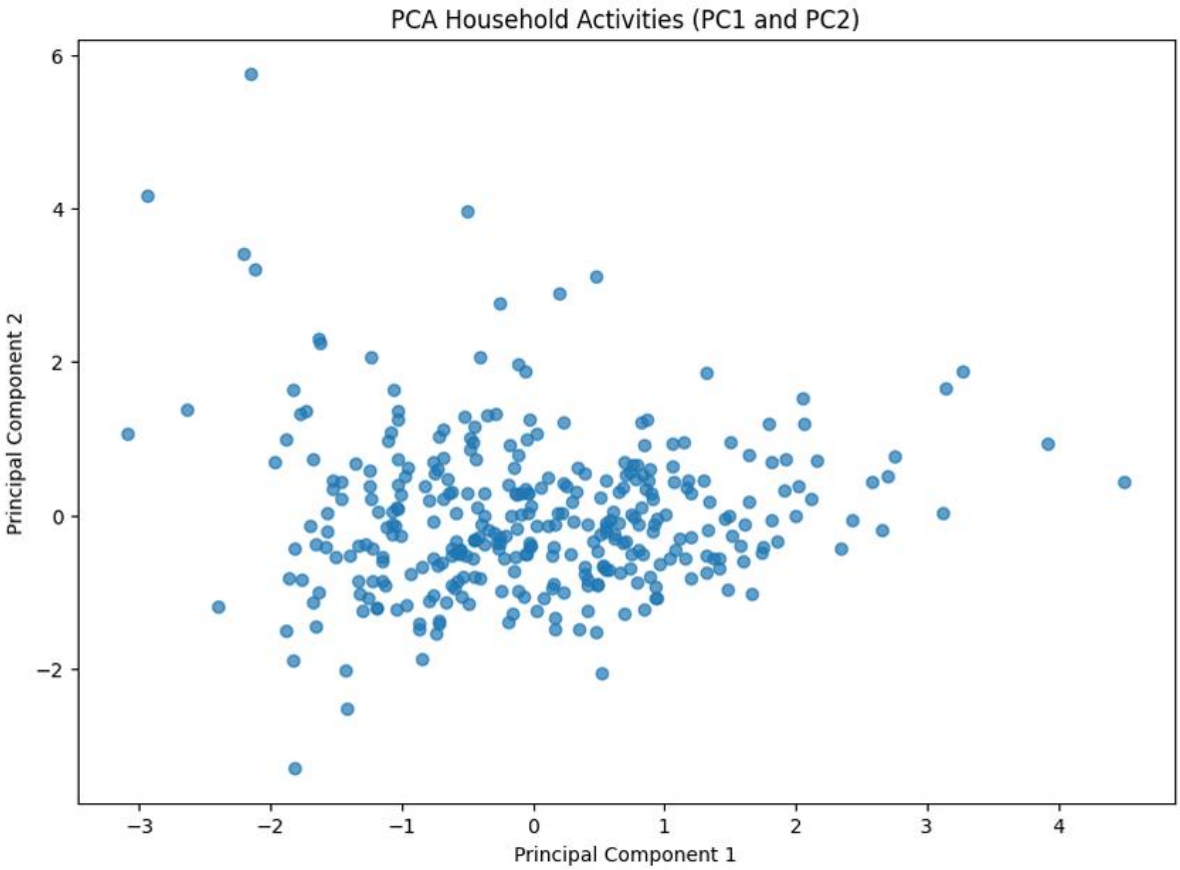
PCA was applied to standardized household-level activity averages to explore underlying patterns.

- The first two components explain about 60% of the total variance.
- Loadings indicate that PC1 contrasts working with sleeping and reading.
- The PCA scatter plot shows a continuous distribution, with no clear clusters.

```
PC1: Explained variance = 0.343, Cumulative = 0.343
PC2: Explained variance = 0.259, Cumulative = 0.602
PC3: Explained variance = 0.245, Cumulative = 0.848
PC4: Explained variance = 0.152, Cumulative = 1.000
```




```
PCA Loadings:
         PC1       PC2       PC3       PC4
A1  0.710639  0.012470 -0.122302  0.692733
A2 -0.533103 -0.637129 -0.010121  0.556565
A3 -0.409898  0.737619  0.281346  0.456888
A4  0.206824 -0.223225  0.951727 -0.040125
```



The PCA plot shows a continuous spread of households with no clear group separation, indicating gradual variation in activity patterns. Thus, PCA mainly serves as a dimensionality-reduction and visualization tool rather than a clustering method on its own.

PCA is therefore used as a dimensionality-reduction and visualization tool rather than a clustering method.

8. Conclusion

This analysis examined daily time-use patterns of individuals and households in Finland using survey data and a combination of descriptive, inferential, and multivariate methods.

Because no clear separation is visible in the PCA space and no clustering objective is defined in the research questions, no clustering algorithm was applied.

Sleeping accounts for the largest share of daily time, followed by working, while reading and dining out occur less frequently and are often zero. Average time estimates are stable and precisely estimated for core activities.

Differences between groups are limited. Dining out is the only activity that differs significantly between weekdays and weekends, with higher values on weekends. By living environment, only sleeping time shows a significant difference, mainly between city and municipality households. Other activities remain largely consistent across groups.

Library visits differ by living environment but not by day of the week, indicating location-related effects rather than weekly routines.

Correlations between activities are generally weak, suggesting that activities represent distinct aspects of daily time use. PCA supports this by showing gradual variation without clear clustering. Because no clear separation is visible in the PCA space and no clustering objective is defined in the research questions, no clustering algorithm was applied.

Overall, Finnish daily routines are highly stable, with only modest variation in discretionary activities depending on context.

9. Use of AI

AI was used for following purposes and following only.

- To fact check theories in the lecture materials.
- To find solutions for syntax and logical errors (Not as whole code!)
- To properly structure the approach of the whole notebook and the report.
- Used Gamma to assist in making the presentation slides.