Institute of Health Informatics

Introduction to Data Pipeline – Analysis of Sleep Variations (KS4)

This toolkit is intended to support teachers delivering enrichment programmes.

Learning objectives:

1. Students will have an introduction to the data collection and processing pipeline used in data science experiments.
2. Students will be able to collect and organise quantitative data related to sleep patterns and present it in a clear and organised manner.
3. Students will be able to identify patterns and trends within a dataset, drawing conclusions about sleep variations within their class and discussing potential implications.

Related subjects:

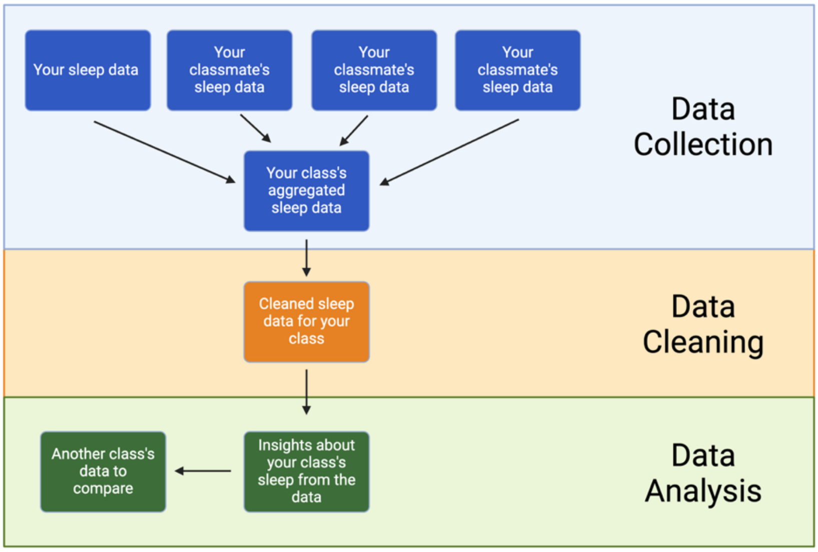
- Mathematics
- Computer Science
- Science/ Biology/ Physics
- Statistics
- Personal & Health Education

Materials:

- Clock
- Pen and paper
- Spreadsheet (e.g. MS Excel, Google Sheets)
- Computer programme for plotting data (e.g. MS Excel, Google Sheets, Python, R)

In this series of activities students will encounter three stages of the data pipeline:

1. Data collection
2. Information assimilation (data cleaning and processing)
3. Knowledge creation (data analysis)


Your sleep data
Your classmate's sleep data
Your classmate's sleep data
Your classmate's sleep data
Your class's aggregated sleep data
Data Collection
Cleaned sleep data for your class
Data Cleaning
Another class's data to compare
Insights about your class's sleep from the data
Data Analysis

Activity 1: Data collection at home

Relevant subjects/topics

- Mathematics
- Science
    - Data collection

Introduction

In the data collection section of this exercise, each student will be asked to collect longitudinal sleep data at home over seven nights. For this, they should each record the time at which they go to sleep and the time at which they wake up.

Data collection – instructions for students

1. Complete the following table (either on paper or using a spreadsheet)

Night number	Date	Bedtime/sleep time (HH:MM)	Wake-up time (HH:MM)
1			
2			
3			
4			
5			
6			
7			

2. After completing the table, submit your data to your teacher, who will aggregate and pull together data from the whole class

Discussion points

- How do we measure what time we go to sleep? Do we note down the time we go to bed, a short time after that, or do we ask someone else to check what time we fall asleep?
- What was your experience of making daily recordings of data (collecting longitudinal data)? Did you manage to collect complete data over the week?
- Why is it important that everyone measures in the same way?

Activity 2: Data cleaning and plot creation

Relevant subjects/topics:

- Mathematics
- Science
  - Data visualisation
  - Data analysis and evaluation
- Computer Science
  - Programming

Introduction

In this section, the collated student data will be cleaned to ensure that it is consistent and properly formatted. Plots will then be created using spreadsheet software to visualise the data. This can either be completed by each of the students themselves, or by the teacher in front of the class as a group exercise.

Data processing – preparation

1. The data from all students should be aggregated into a single spreadsheet, and anonymised (remove student names, initials, etc). You may use the spreadsheet provided here https://github.com/ucl-ihi/NSL-2024-teacherspack/raw/main/data_activity_spreadsheet.xlsx

Data processing – instructions for students (or for teacher in front of class)

1. Clean and process your data. e.g. calculating hours of sleep from sleep and wake-up times, average hours of sleep per student, check that all data is in the correct format etc.
2. Visualise your data as various graph types using the spreadsheet software plotting functionality. There are step-by-step instructions for plotting two graphs below: histogram of hours asleep, histogram of bedtime.

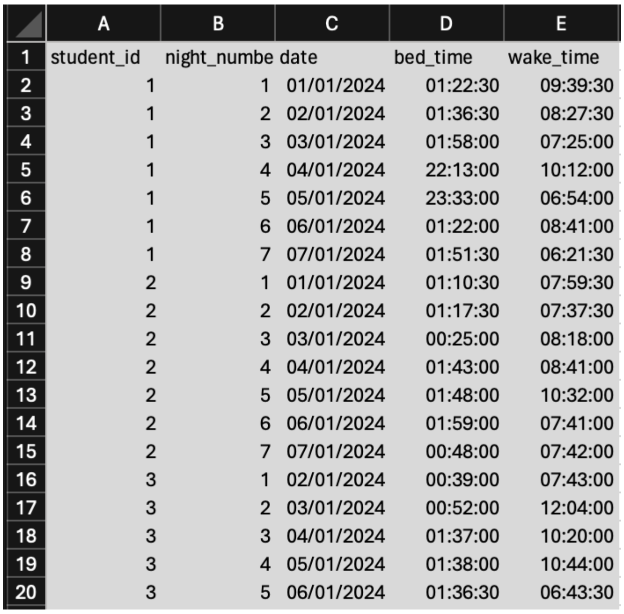
Discussion points

- How do you feel about sharing your deidentified data? How would you feel if the data you shared was identifiable to you? What are the advantages and disadvantages of identifiable and deidentified data?
- Why is it important to clean the data? What would happen if there was some incorrectly formatted data? This could be illustrated by introducing some data artifacts, for example:
  - Entries missing the sleep time and only including the wake-up time
  - Entries with the time recorded in the incorrect format, e.g. “eight” instead of 8.
- Do we want to stratify (split up) the data before we plot it? Can we plot weekend data separately to the weekday data and what would this show us? Should we average hours of sleep by individual students, or by day of the week, or not average it at all?
- What kinds of plots would work well for the data that we are displaying? What would we be able to see from the plot using a histogram vs a pie chart vs a scatter plot?

Plotting worked example (spreadsheet) for Activity 2

The final Excel spreadsheet is found here: https://github.com/ucl-ihi/NSL-2024-teacherspack/raw/main/data_activity_excel_plot.xlsx

The spreadsheet begins with just the data, which is in five columns: student_id, night_number, date, bed_time, wake_time

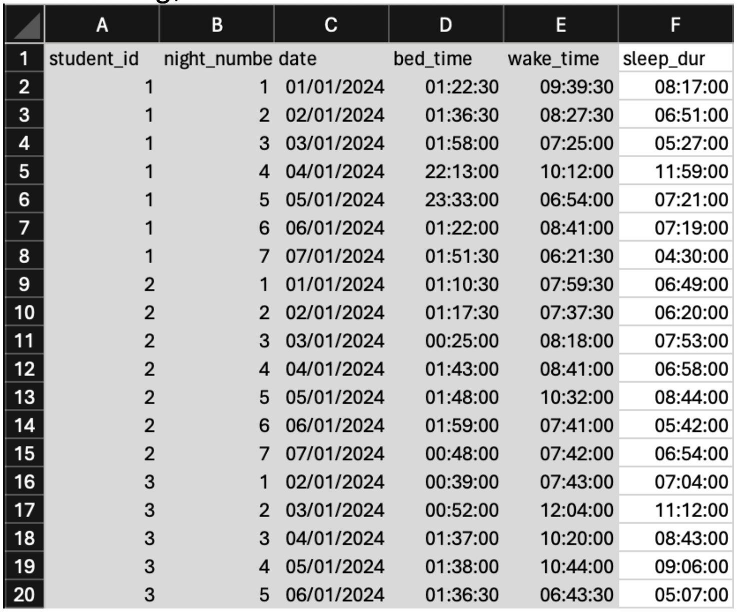


	A	B	C	D	E
1	student_id	night_numbe	date	bed_time	wake_time
2	1	1	01/01/2024	01:22:30	09:39:30
3	1	2	02/01/2024	01:36:30	08:27:30
4	1	3	03/01/2024	01:58:00	07:25:00
5	1	4	04/01/2024	22:13:00	10:12:00
6	1	5	05/01/2024	23:33:00	06:54:00
7	1	6	06/01/2024	01:22:00	08:41:00
8	1	7	07/01/2024	01:51:30	06:21:30
9	2	1	01/01/2024	01:10:30	07:59:30
10	2	2	02/01/2024	01:17:30	07:37:30
11	2	3	03/01/2024	00:25:00	08:18:00
12	2	4	04/01/2024	01:43:00	08:41:00
13	2	5	05/01/2024	01:48:00	10:32:00
14	2	6	06/01/2024	01:59:00	07:41:00
15	2	7	07/01/2024	00:48:00	07:42:00
16	3	1	02/01/2024	00:39:00	07:43:00
17	3	2	03/01/2024	00:52:00	12:04:00
18	3	3	04/01/2024	01:37:00	10:20:00
19	3	4	05/01/2024	01:38:00	10:44:00
20	3	5	06/01/2024	01:36:30	06:43:30

Starting a cell with = in Excel or Google Sheets will calculate the formula entered. For example, typing 1+1 in a cell and pressing enter will show 1+1, whereas typing =1+1 in a cell and pressing enter will show 2.

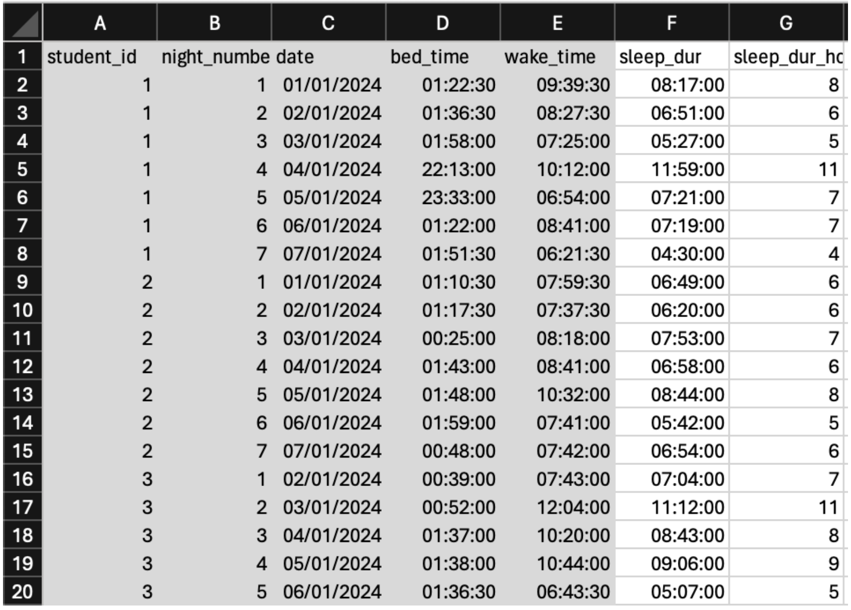
Plot 1: histogram of hours asleep

1. Create a new column called sleep_dur which will contain the sleep duration calculated from the wake_time and bed_time. Use the equation (24 + wake_time) - bed_time, for example, in cell F2, enter `=(24+E2)-D2` .The 24 enables the subtraction to give a positive number even when the person slept in the previous day (e.g. 6 – 23), while maintaining the same meaning, because the hours of the clock is of modulo 24.



	A	B	C	D	E	F
1	student_id	night_numbe	date	bed_time	wake_time	sleep_dur
2	1	1	01/01/2024	01:22:30	09:39:30	08:17:00
3	1	2	02/01/2024	01:36:30	08:27:30	06:51:00
4	1	3	03/01/2024	01:58:00	07:25:00	05:27:00
5	1	4	04/01/2024	22:13:00	10:12:00	11:59:00
6	1	5	05/01/2024	23:33:00	06:54:00	07:21:00
7	1	6	06/01/2024	01:22:00	08:41:00	07:19:00
8	1	7	07/01/2024	01:51:30	06:21:30	04:30:00
9	2	1	01/01/2024	01:10:30	07:59:30	06:49:00
10	2	2	02/01/2024	01:17:30	07:37:30	06:20:00
11	2	3	03/01/2024	00:25:00	08:18:00	07:53:00
12	2	4	04/01/2024	01:43:00	08:41:00	06:58:00
13	2	5	05/01/2024	01:48:00	10:32:00	08:44:00
14	2	6	06/01/2024	01:59:00	07:41:00	05:42:00
15	2	7	07/01/2024	00:48:00	07:42:00	06:54:00
16	3	1	02/01/2024	00:39:00	07:43:00	07:04:00
17	3	2	03/01/2024	00:52:00	12:04:00	11:12:00
18	3	3	04/01/2024	01:37:00	10:20:00	08:43:00
19	3	4	05/01/2024	01:38:00	10:44:00	09:06:00
20	3	5	06/01/2024	01:36:30	06:43:30	05:07:00

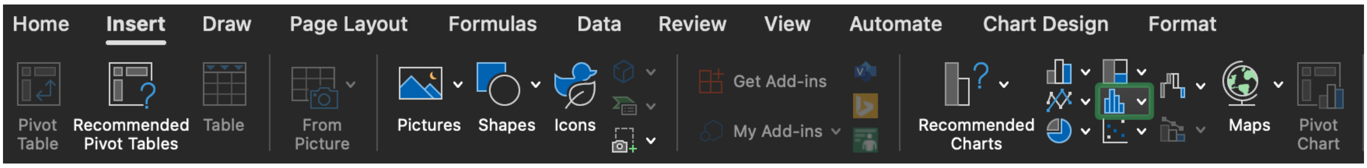
2. Create a new column called sleep_dur_hour which will contain the sleep duration rounded down to the hour. Use the HOUR function on the sleep_dur column, which returns the hour of a time, for example, in cell G2, enter `=HOUR(F2)` This is the column we will plot, you may now use plot method 1 (step 3 and 4) or plot method 2 (steps 5-7).

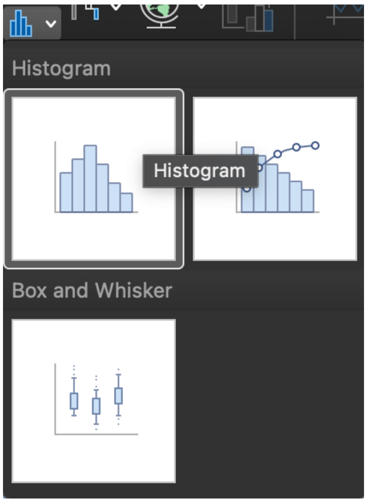


	A	B	C	D	E	F	G
1	student_id	night_numbe	date	bed_time	wake_time	sleep_dur	sleep_dur_hc
2	1	1	01/01/2024	01:22:30	09:39:30	08:17:00	8
3	1	2	02/01/2024	01:36:30	08:27:30	06:51:00	6
4	1	3	03/01/2024	01:58:00	07:25:00	05:27:00	5
5	1	4	04/01/2024	22:13:00	10:12:00	11:59:00	11
6	1	5	05/01/2024	23:33:00	06:54:00	07:21:00	7
7	1	6	06/01/2024	01:22:00	08:41:00	07:19:00	7
8	1	7	07/01/2024	01:51:30	06:21:30	04:30:00	4
9	2	1	01/01/2024	01:10:30	07:59:30	06:49:00	6
10	2	2	02/01/2024	01:17:30	07:37:30	06:20:00	6
11	2	3	03/01/2024	00:25:00	08:18:00	07:53:00	7
12	2	4	04/01/2024	01:43:00	08:41:00	06:58:00	6
13	2	5	05/01/2024	01:48:00	10:32:00	08:44:00	8
14	2	6	06/01/2024	01:59:00	07:41:00	05:42:00	5
15	2	7	07/01/2024	00:48:00	07:42:00	06:54:00	6
16	3	1	02/01/2024	00:39:00	07:43:00	07:04:00	7
17	3	2	03/01/2024	00:52:00	12:04:00	11:12:00	11
18	3	3	04/01/2024	01:37:00	10:20:00	08:43:00	8
19	3	4	05/01/2024	01:38:00	10:44:00	09:06:00	9
20	3	5	06/01/2024	01:36:30	06:43:30	05:07:00	5

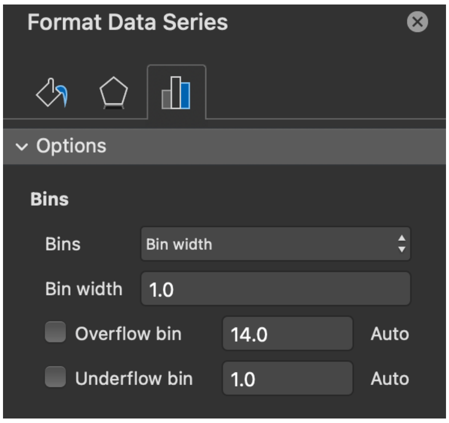
<u>Plot method 1 (step 3 and 4):</u> use the Histogram chart type

3. Select all of column G, navigate to the Insert menu at the top, select histogram.





4. Double click on the bars of the chart to open the Format Data Series side menu. Customise as you see fit (for example, bin width 1 would give one bar per hour).



Format Data Series

Options

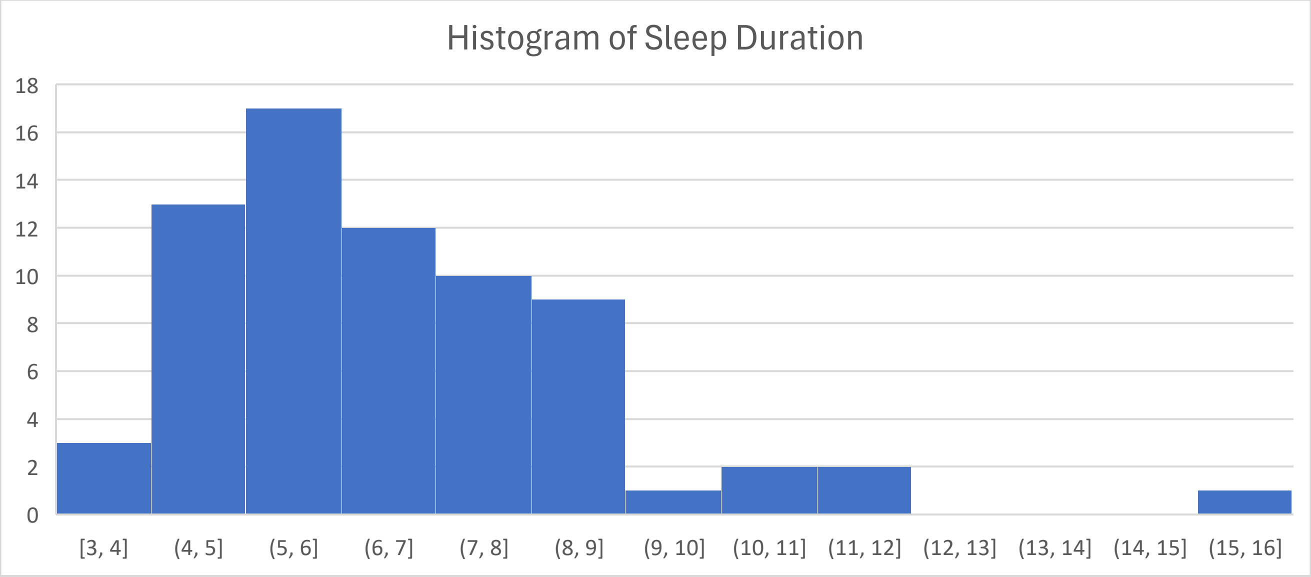
Bins

Bins: Bin width

Bin width: 1.0

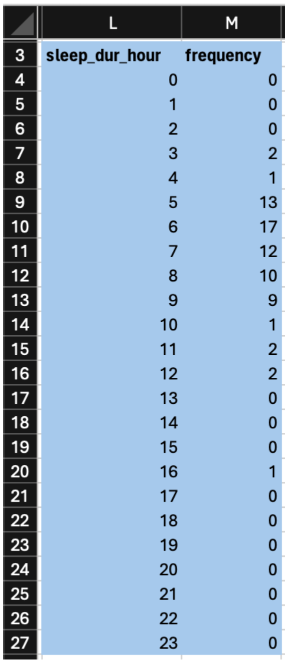
□ Overflow bin 14.0 Auto

□ Underflow bin 1.0 Auto



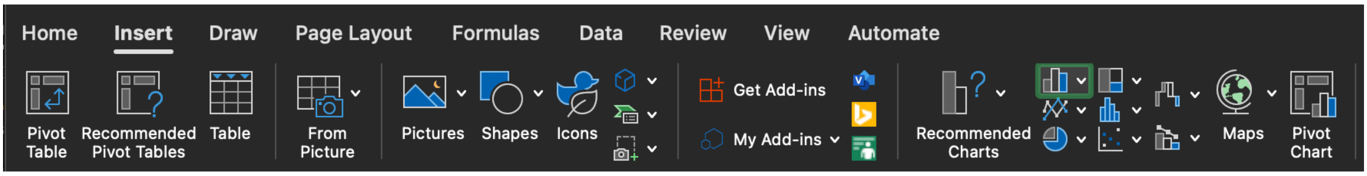
<u>Plot method 2 (steps 5-7):</u> create a new table to count the frequency of each hour manually then plot.

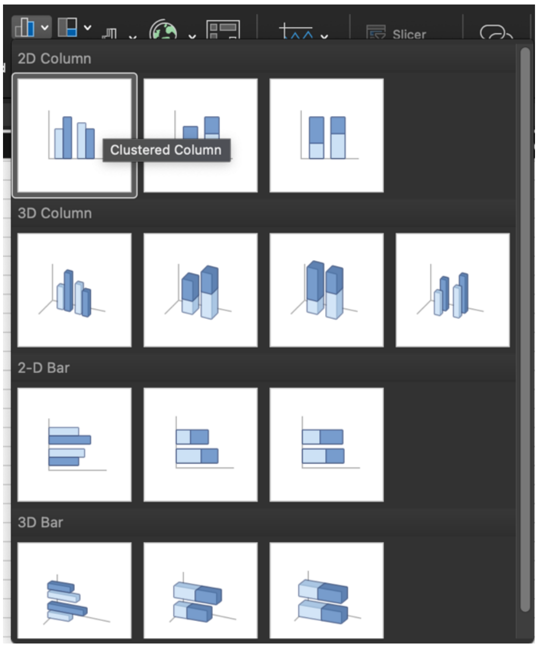
5. In a new column, list the number 0 to 23 to represent all hours.
6. Create a column called frequency which will count the number of occurrences of each hour in the sleep_dur_hour column. Use the COUNTIF function to count cells that match a condition. If we list all hours in column L and have the sleep_dur_hours in column G, in cell M4, use `=COUNTIF($G$2:$G$71,L4)`. Note: the $ will give an absolute cell reference, thus references will not move around when cells are duplicated.

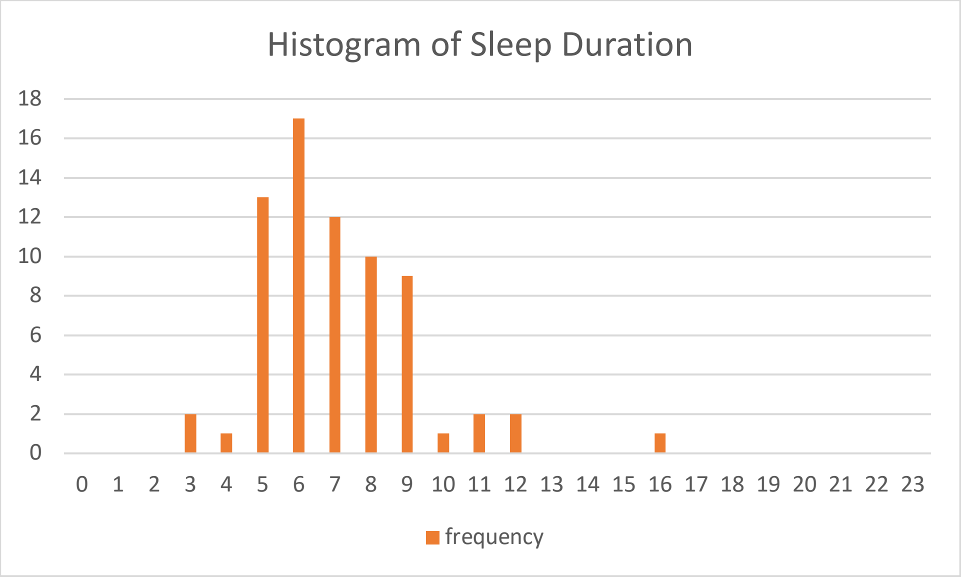


	L	M
3	sleep_dur_hour	frequency
4	0	0
5	1	0
6	2	0
7	3	2
8	4	1
9	5	13
10	6	17
11	7	12
12	8	10
13	9	9
14	10	1
15	11	2
16	12	2
17	13	0
18	14	0
19	15	0
20	16	1
21	17	0
22	18	0
23	19	0
24	20	0
25	21	0
26	22	0
27	23	0

7. Select the cells of frequency, navigate to the Insert menu at the top, select 2D column.






Histogram of Sleep Duration
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
frequency

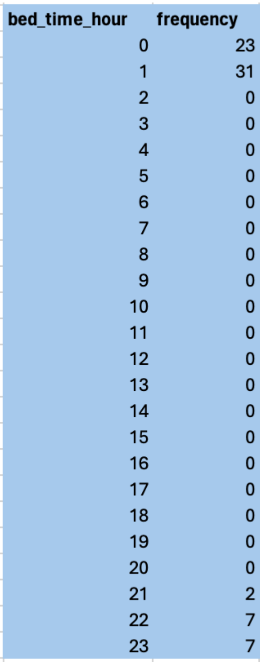
Plot 2: histogram of bedtime (steps 8-16)

To create plot 2, first follow the instructions to create plot 1 (steps 1-7).

8. Create a new column called bed_time_hour which will contain the bed_time rounded down to the hour. Use the HOUR function on the bed_time column, which returns the hour of a time, for example, in cell H2, enter `=HOUR(D2)` This is the column we will plot.
9. You may use create a histogram using the same methods as with plot 1 (**Plot method 1**: steps 3-4 or **Plot method 2**: steps 5-7) or plot method 3 below (steps 10-16).

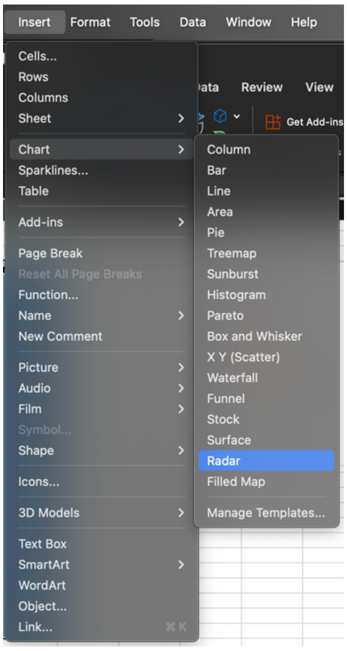
Plot method 3 (steps 10-16): to create a more accurate representation of the cyclic nature of the clock hours, we can use a polar plot.

10. Follow steps 5 and 6 above (Plot 1 method 2) to obtain the following table with two columns bed_time_hour and frequency using COUNTIF

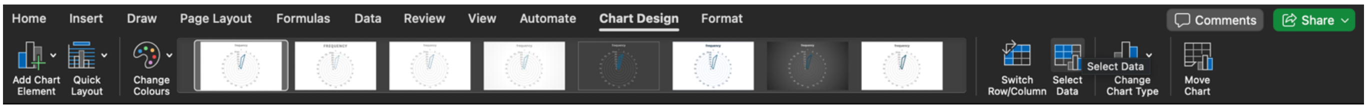


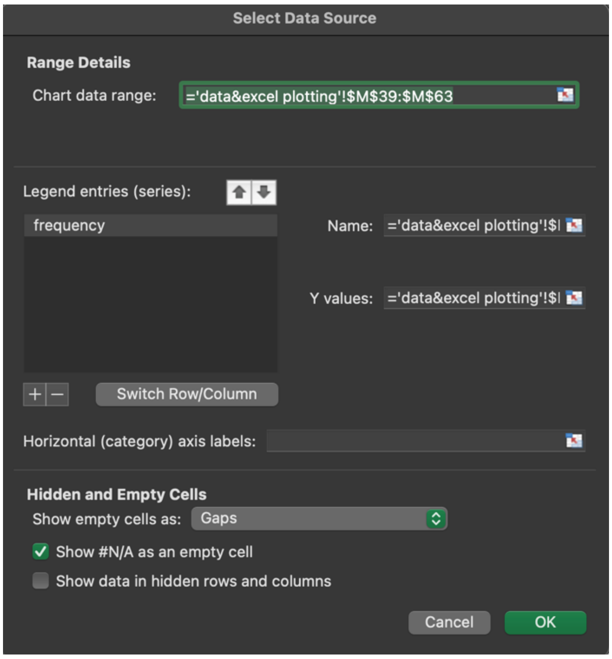
bed_time_hour	frequency
0	23
1	31
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	2
22	7
23	7

11. Select the cells of frequency, navigate to the Insert menu at the top bar, then Chart > Radar



12. The labels around the circle may be incorrect (starting at 1), to use the correct labels (starting at 0), select Chart Design menu and Select Data. Then update the Horizontal (category) axis labels to the cells with bed_time_hour listing 0 to 23.




Select Data Source

Range Details

Chart data range: ='data&excel plotting'!$M$39:$M$63

Legend entries (series): frequency

Name: ='data&excel plotting'!$I

Y values: ='data&excel plotting'!$I

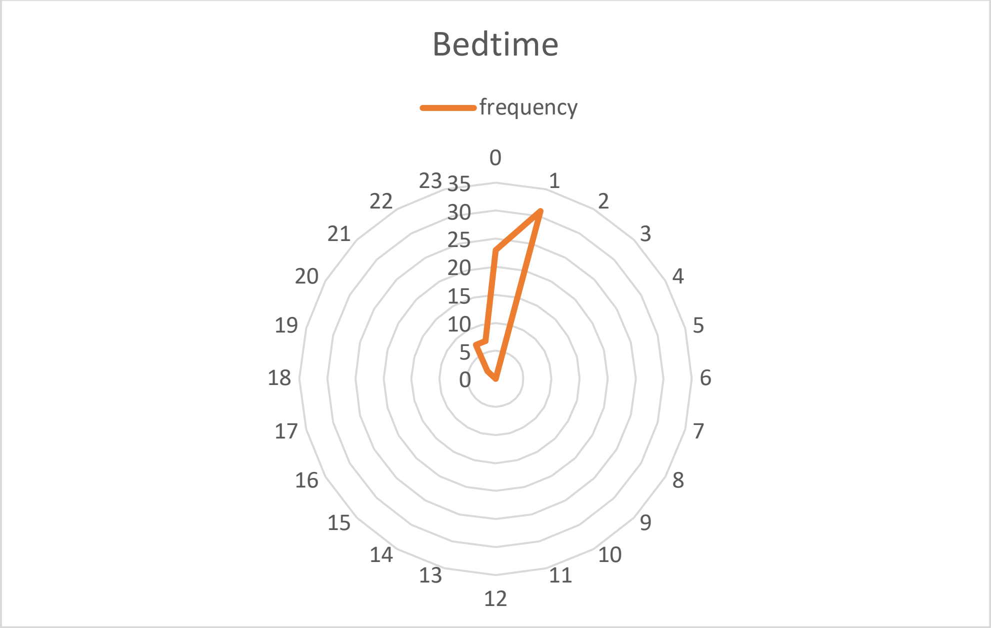
Horizontal (category) axis labels:

Hidden and Empty Cells

Show empty cells as: Gaps

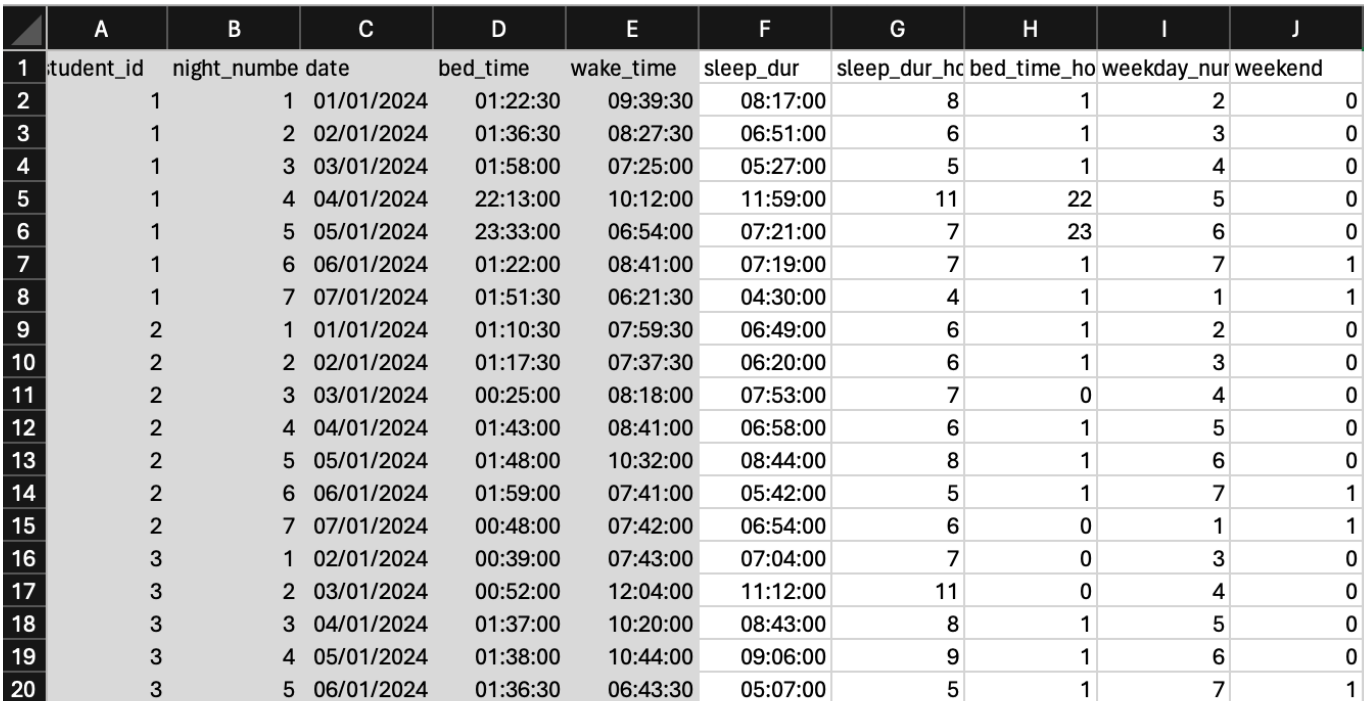
☑ Show #N/A as an empty cell

□ Show data in hidden rows and columns



13. If you would like to take the plot further and make a two in-depth plot to explore the bed times on weekdays and weekends, create a column called weekday_num which will contain the day of the week, numbers 1 (Sunday) through 7 (Saturday). Use the WEEKDAY function on the date column. For example, in cell I2, enter `=WEEKDAY(C2)`

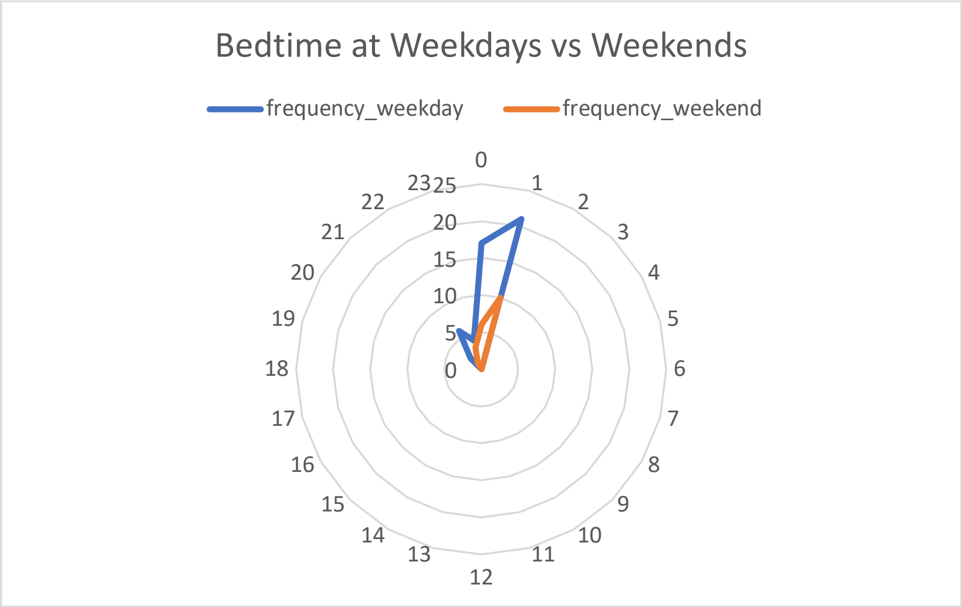
14. Create a column called weekend which will have “0” to represent a weekday and “1” to represent a weekend, this is called a binary column. We will use the OR logic, meaning as long as one of the conditions are met, we will give “1”. Add two COUNTIF together, one to count when the weekday_num is “1” (Sunday) and the other when weekday_num is “7” (Saturday). For example, in cell J2, enter `=COUNTIF(I2,1) + COUNTIF(I2,7)`



	A	B	C	D	E	F	G	H	I	J
1	tudent_id	night_numbe	date	bed_time	wake_time	sleep_dur	sleep_dur_hc	bed_time_ho	weekday_nur	weekend
2	1	1	01/01/2024	01:22:30	09:39:30	08:17:00	8	1	2	0
3	1	2	02/01/2024	01:36:30	08:27:30	06:51:00	6	1	3	0
4	1	3	03/01/2024	01:58:00	07:25:00	05:27:00	5	1	4	0
5	1	4	04/01/2024	22:13:00	10:12:00	11:59:00	11	22	5	0
6	1	5	05/01/2024	23:33:00	06:54:00	07:21:00	7	23	6	0
7	1	6	06/01/2024	01:22:00	08:41:00	07:19:00	7	1	7	1
8	1	7	07/01/2024	01:51:30	06:21:30	04:30:00	4	1	1	1
9	2	1	01/01/2024	01:10:30	07:59:30	06:49:00	6	1	2	0
10	2	2	02/01/2024	01:17:30	07:37:30	06:20:00	6	1	3	0
11	2	3	03/01/2024	00:25:00	08:18:00	07:53:00	7	0	4	0
12	2	4	04/01/2024	01:43:00	08:41:00	06:58:00	6	1	5	0
13	2	5	05/01/2024	01:48:00	10:32:00	08:44:00	8	1	6	0
14	2	6	06/01/2024	01:59:00	07:41:00	05:42:00	5	1	7	1
15	2	7	07/01/2024	00:48:00	07:42:00	06:54:00	6	0	1	1
16	3	1	02/01/2024	00:39:00	07:43:00	07:04:00	7	0	3	0
17	3	2	03/01/2024	00:52:00	12:04:00	11:12:00	11	0	4	0
18	3	3	04/01/2024	01:37:00	10:20:00	08:43:00	8	1	5	0
19	3	4	05/01/2024	01:38:00	10:44:00	09:06:00	9	1	6	0
20	3	5	06/01/2024	01:36:30	06:43:30	05:07:00	5	1	7	1

15. Create two frequency columns, one for weekdays and one for weekends. Here we use COUNTIFS which can take multiple conditions. The first condition is the same as before, the second condition is to filter for whether the weekend column says 0 or 1. If Column H contains the data to summarise (bed_time_hour), column J contains the data to filter by (weekend), and cell U40 is an hour of the day (e.g. 0), in cell V2 we have the total nights slept beginning at midnight on weekdays (0), enter `=COUNTIFS($H$2:$H$71,U40,$J$2:$J$71,"=0")` . In cell W2 we have the total nights slept beginning at midnight on weekends (1), enter `=COUNTIFS($H$2:$H$71,AD40,$J$2:$J$71,"=1")`

16. Repeat steps 11 and 12 to make the polar histogram plot



Activity 3: Data analysis, knowledge and insights extracted from your own data and sample data

Relevant subjects/topics:

- Mathematics/Statistics
  - Graphs
  - Samples
- Science
  - Data visualisation
  - Data analysis and evaluation
- Computer Science
  - Programming

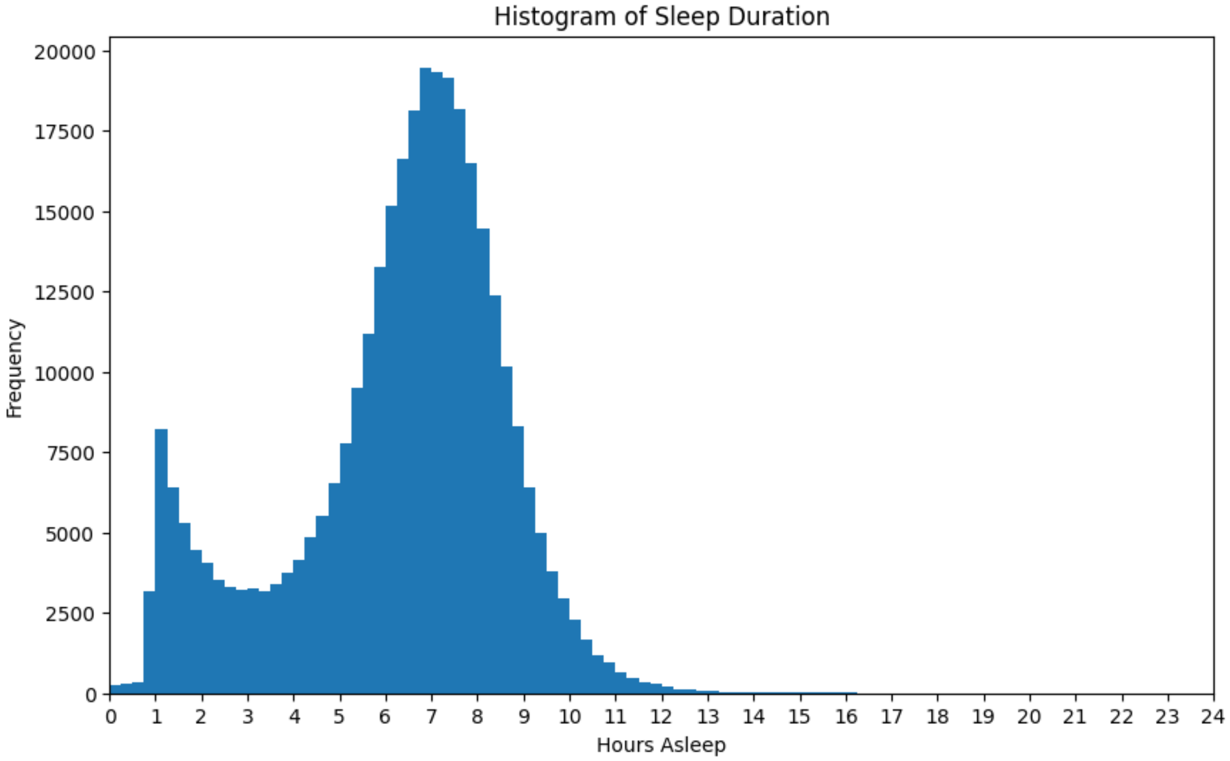
Introduction

To provide a comparison of your activity and sleep data with more people, we will use Fitbit data from the NetHealth Project (https://sites.nd.edu/nethealth/). The project contains data from around 700 undergraduate students at the University of Notre Dame, USA, who enrolled in the Autumn of 2015 and graduated in the Spring of 2019. Using the Fitbit smartwatch, their physical activity and sleep were tracked and summarised to daily measurements.

This activity will compare your own data with your class's data then with the NetHealth data.

Data analysis of NetHealth students – presented by teachers

This histogram plot generally shows a Normal distribution (bell curve) centred around 7 hours of sleep. More specifically, the distribution is called a Bimodal distribution, with two peaks at 1 hour and 7 hours of sleep. 1 hour of sleep is likely naps or returned to bed after waking up.

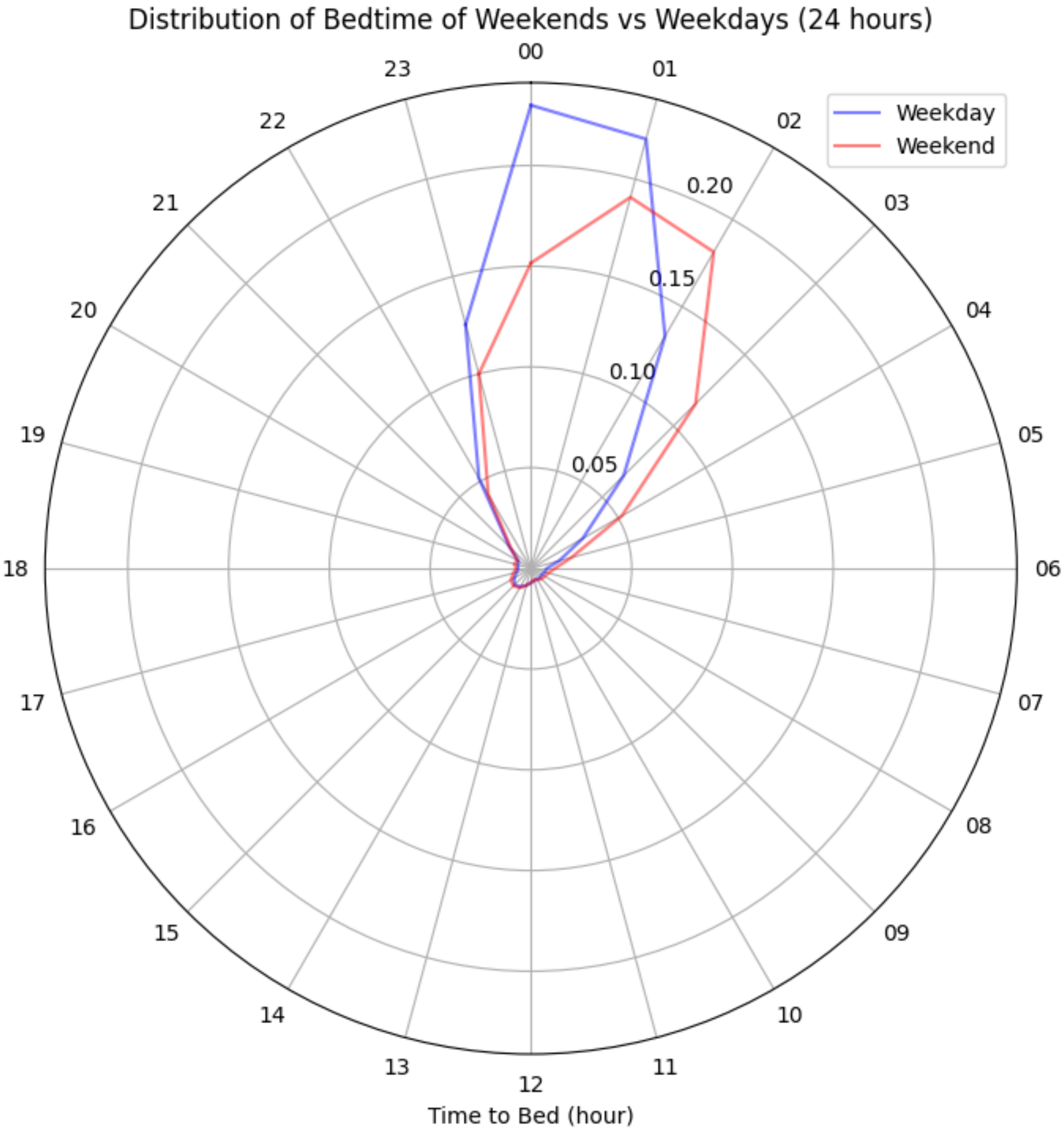


This polar plot (using the circular polar axis as to the cartesian XY axis) shows the density histogram of how many people slept at which hour of the day. The further away from the centre, the more people slept at that hour. The motivation for using a polar axis to represent the hours of the day instead of the regular XY axis, is that the hours are cyclic; using a polar axis can ensure that 11pm, midnight, 1am are neighbours in the visualisation.

Instead of frequency in the above histogram, this plot uses density, which enables more direct comparison of the overall pattern and shape. Since there are more weekdays than weekends, there is also more data from weekdays than weekends. Plotting the frequency will result in a large loop for weekdays and a small loop for weekends.

There is a clear difference in the distribution of bedtime between weekdays and weekends. On weekends, people tend to go to bed later than on weekdays, around midnight on weekdays and around 1am to 2am on weekends. There is also greater variability in the bedtime on weekends compared to weekdays, shown by the rounder loop.

Idea for activity: students can have a go at understanding the plot before it is explained.



Data analysis – instructions for students

1. Draw a line with your average (mean) hours of sleep across the week. How does this compare with the data that you see in the histogram?
2. Plot your weekday and weekend sleep times separately. Do you see that you slept later, or for less time on weekends than on weekdays?
3. Compare your class's data to the data from another class or the NetHealth class, or your individual data to one of your friend's – does it look similar or different? Why do you think this might be?

Data analysis (coding) – for teachers and advanced students

All the Python code to process the NetHealth data and recreate the plots are found here: https://github.com/ucl-ihi/NSL-2024-teacherspack/blob/main/data_activity_nethealth_analysis.ipynb

Resources created by:

Dr Kevin Tsang (IHI, UCL)

Nikita Hukerikar (IHI, UCL)

Reviewed by:

Dr Kate Honeyford (GBSH, UCL)

Dr Matthew Darlison (IHI, UCL)