Analisis Deskriptif

Pengantar

Data Science

Semester Ganjil 2022 / 2023

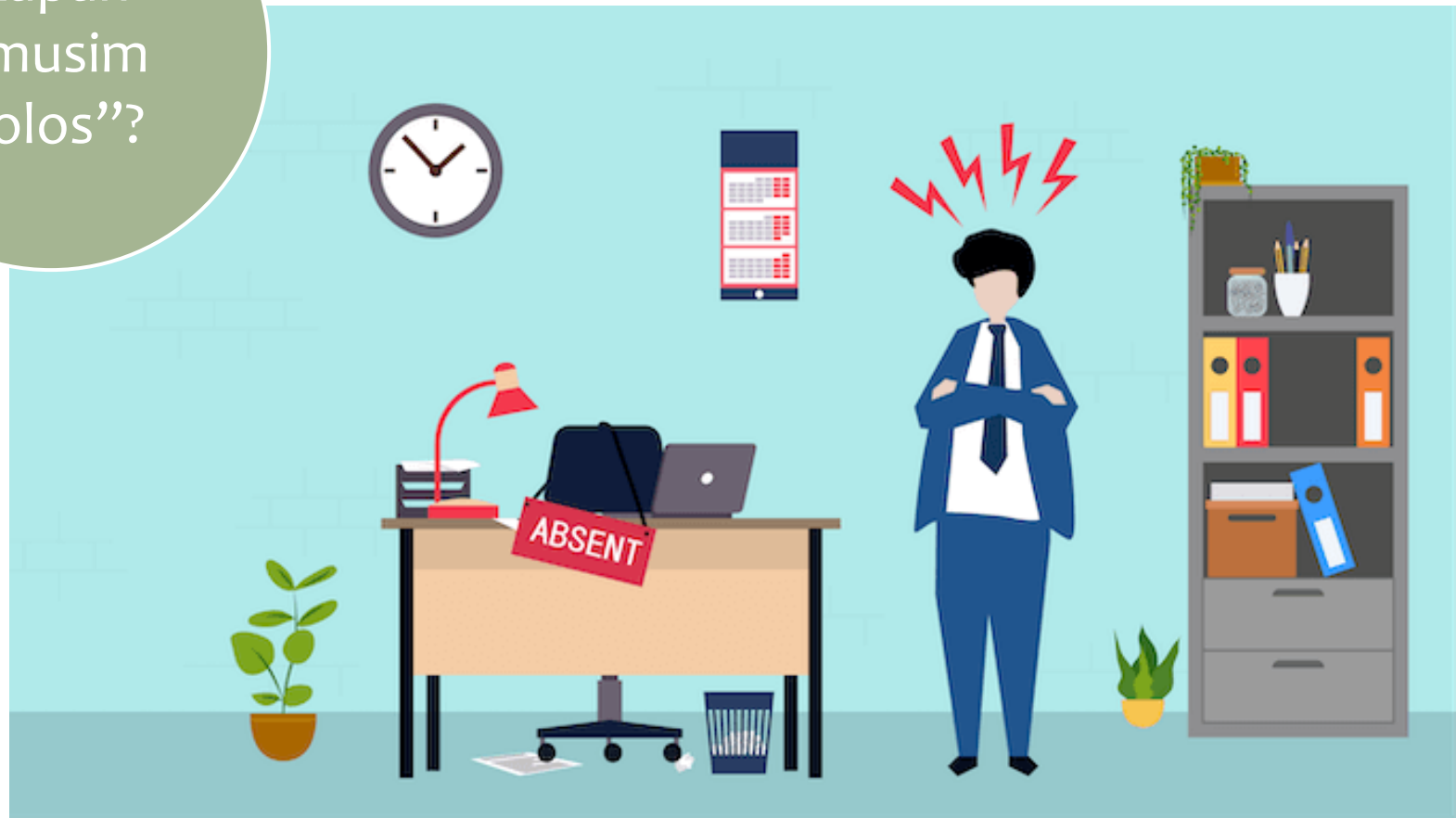
Maria Veronica Claudia M., S.T., M.T.

Absenteeism at Work

Siapa yang sering tidak masuk?

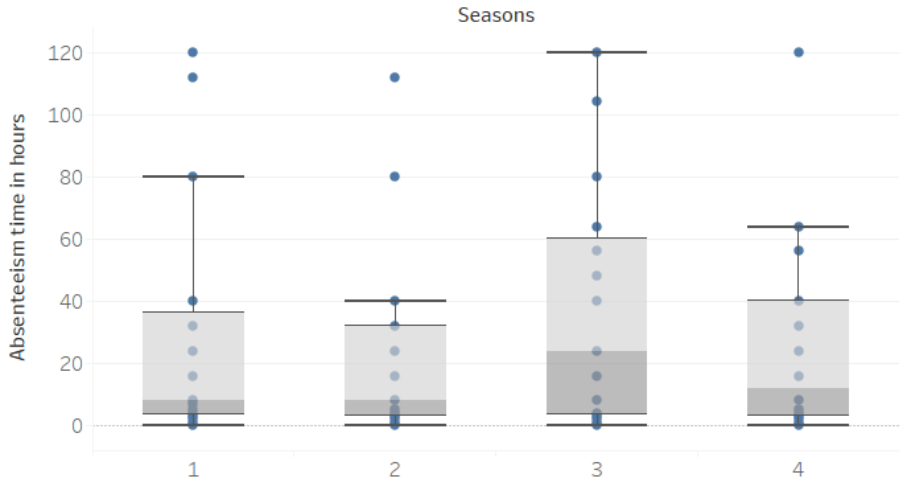
Kapan "musim bolos"?

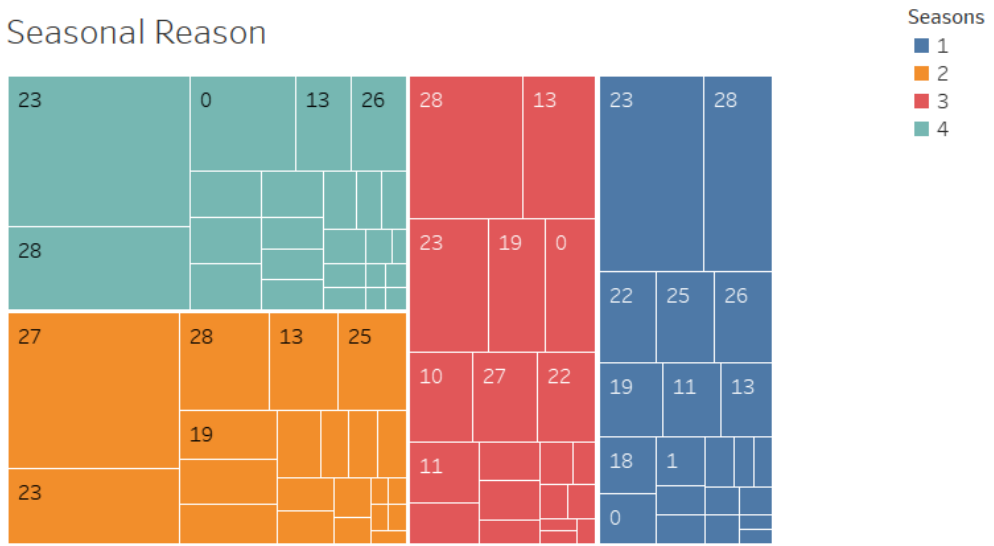
Alasan favorit

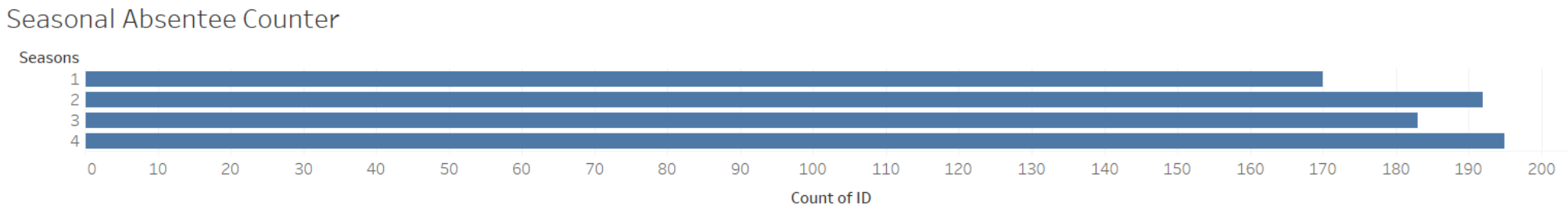


INFORMATIKA UNPAR

Seasonal Reports

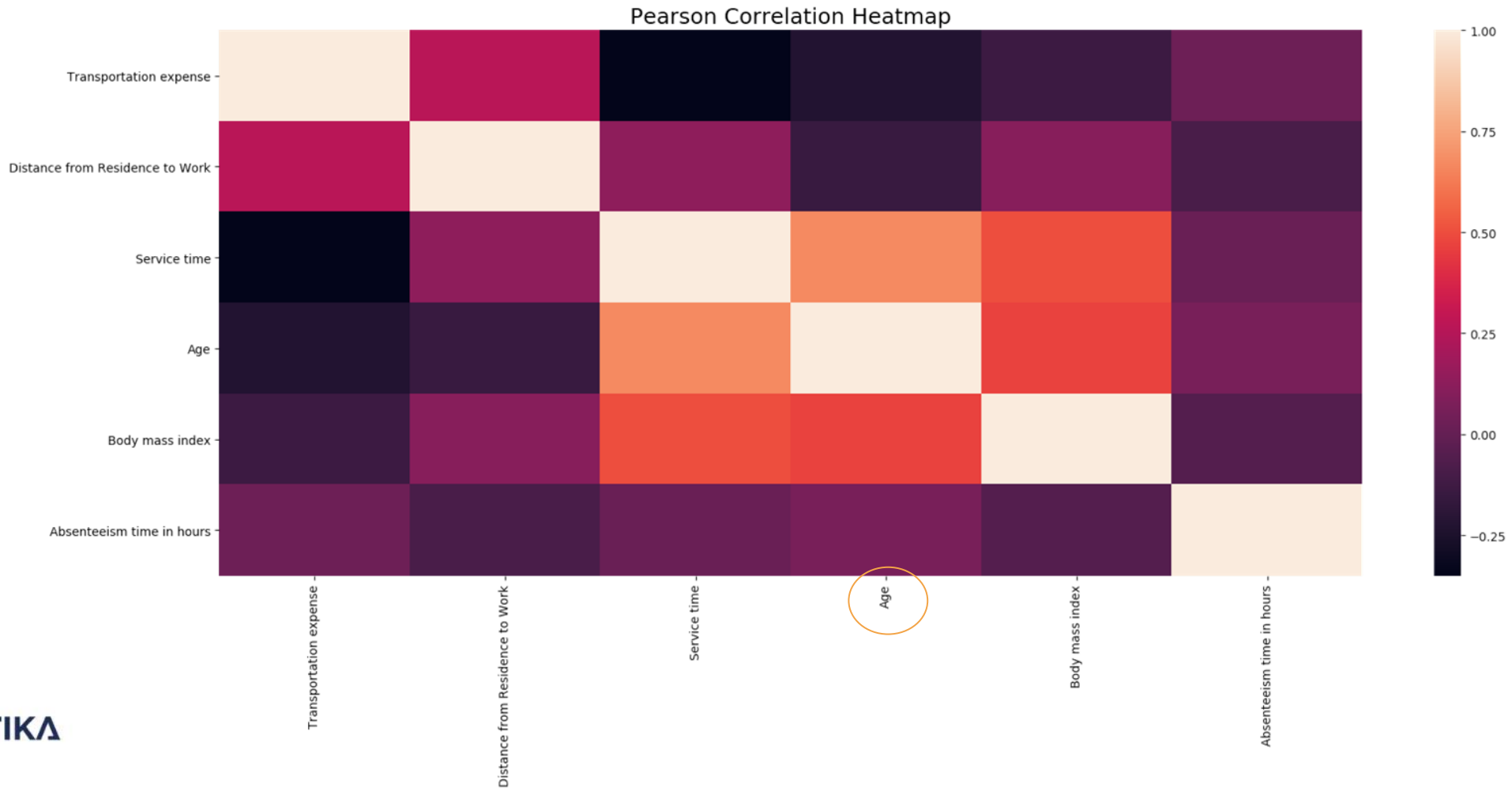






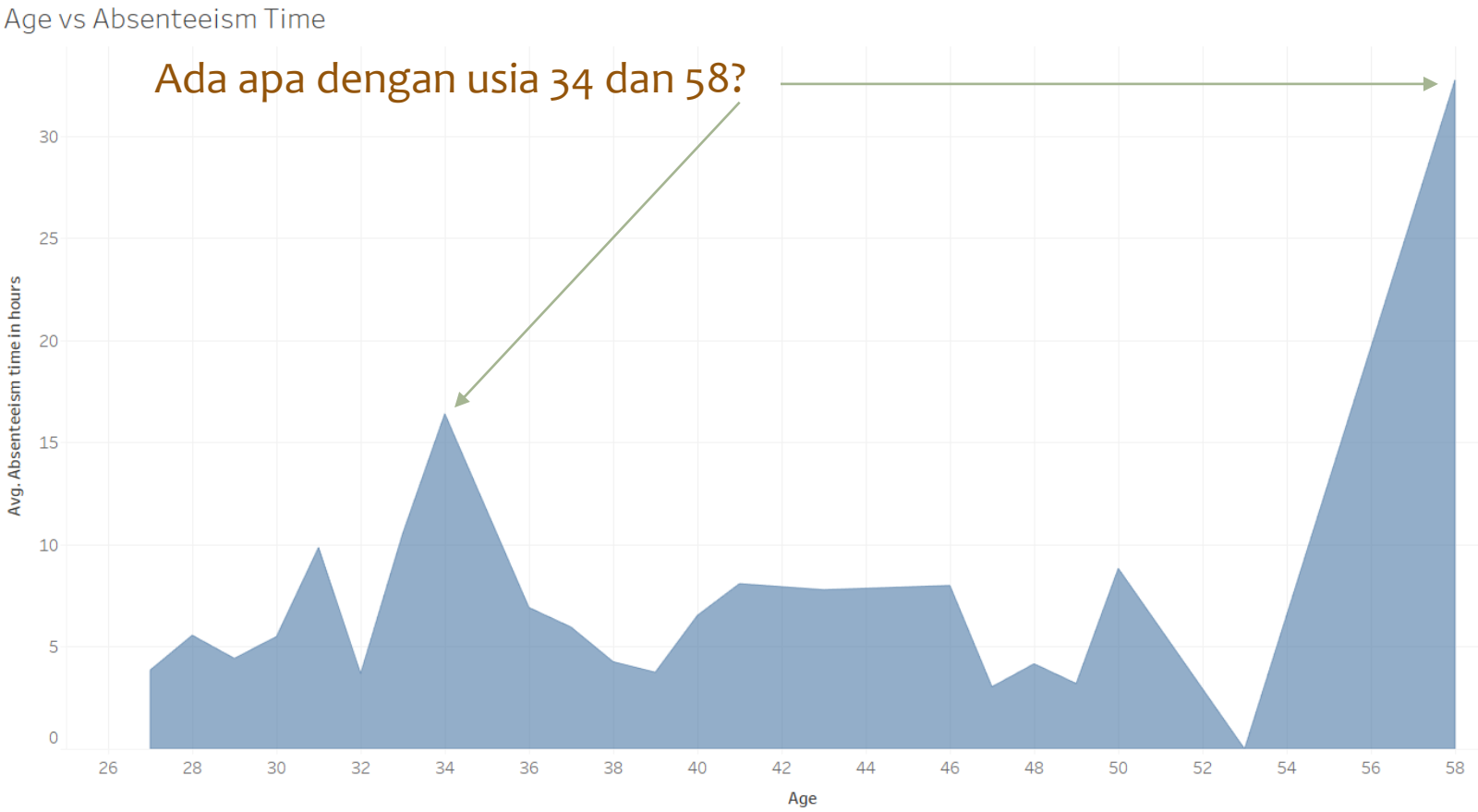
INFORMATIKA UNPAR

Pearson Correlation



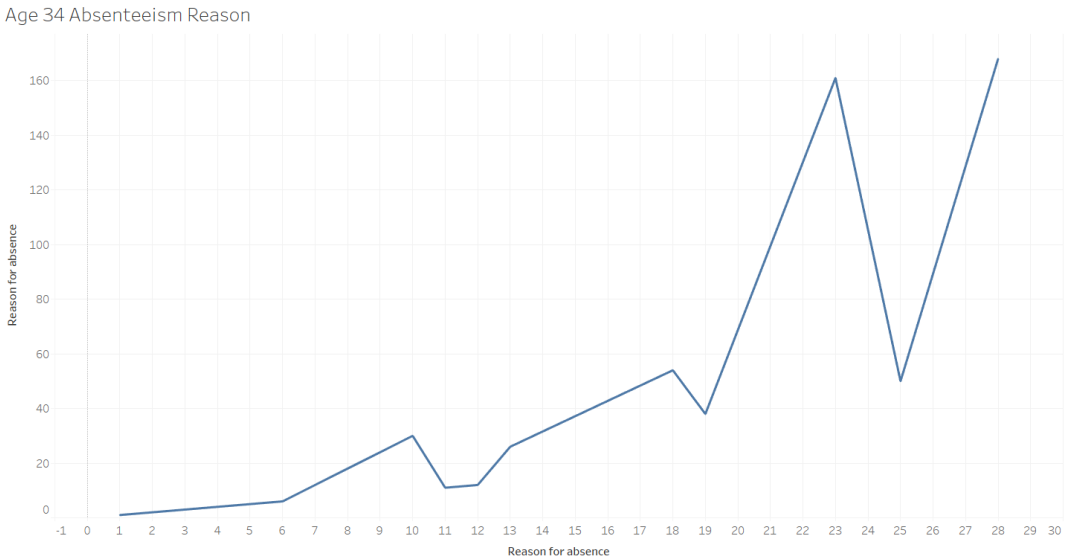
INFORMATIKA UNPAR

Rata-rata per usia

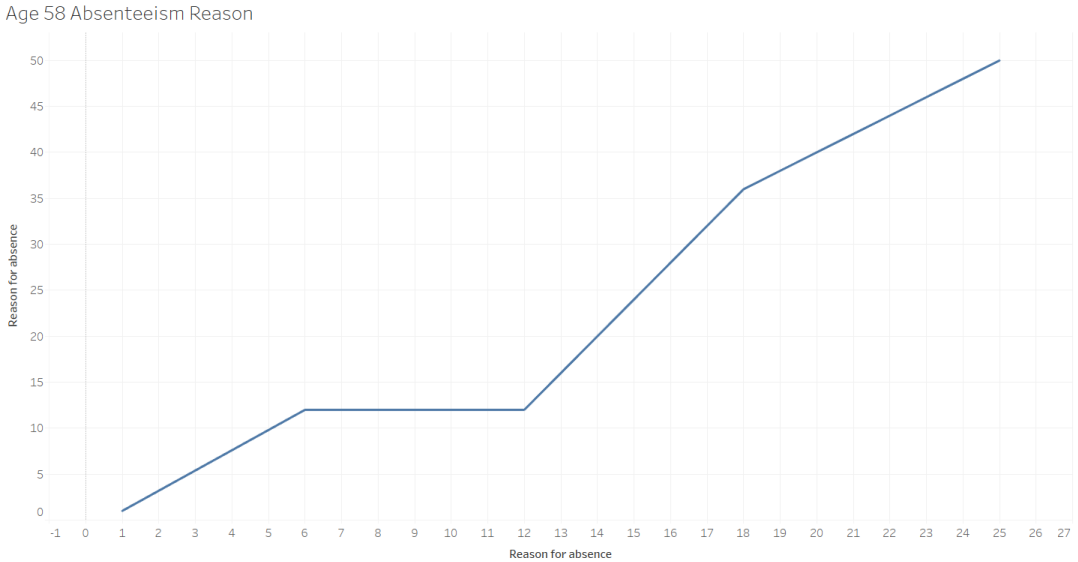


INFORMATIKA UNPAR

Why?



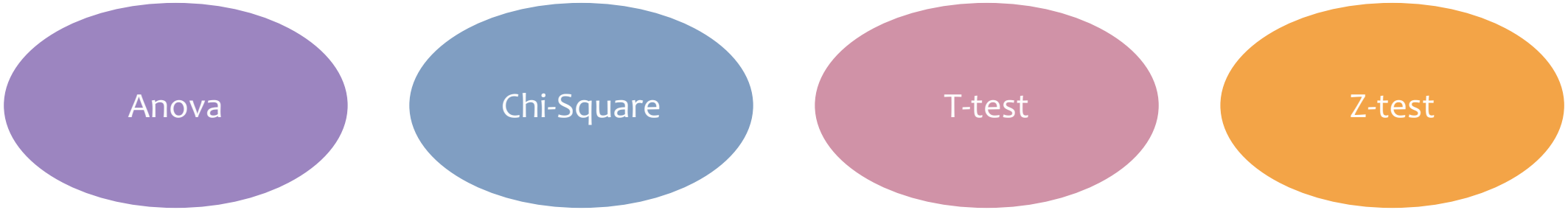
Medical and dental consultation



Laboratory examination

INFORMATIKA UNPAR

Tes-tes Lain



https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740

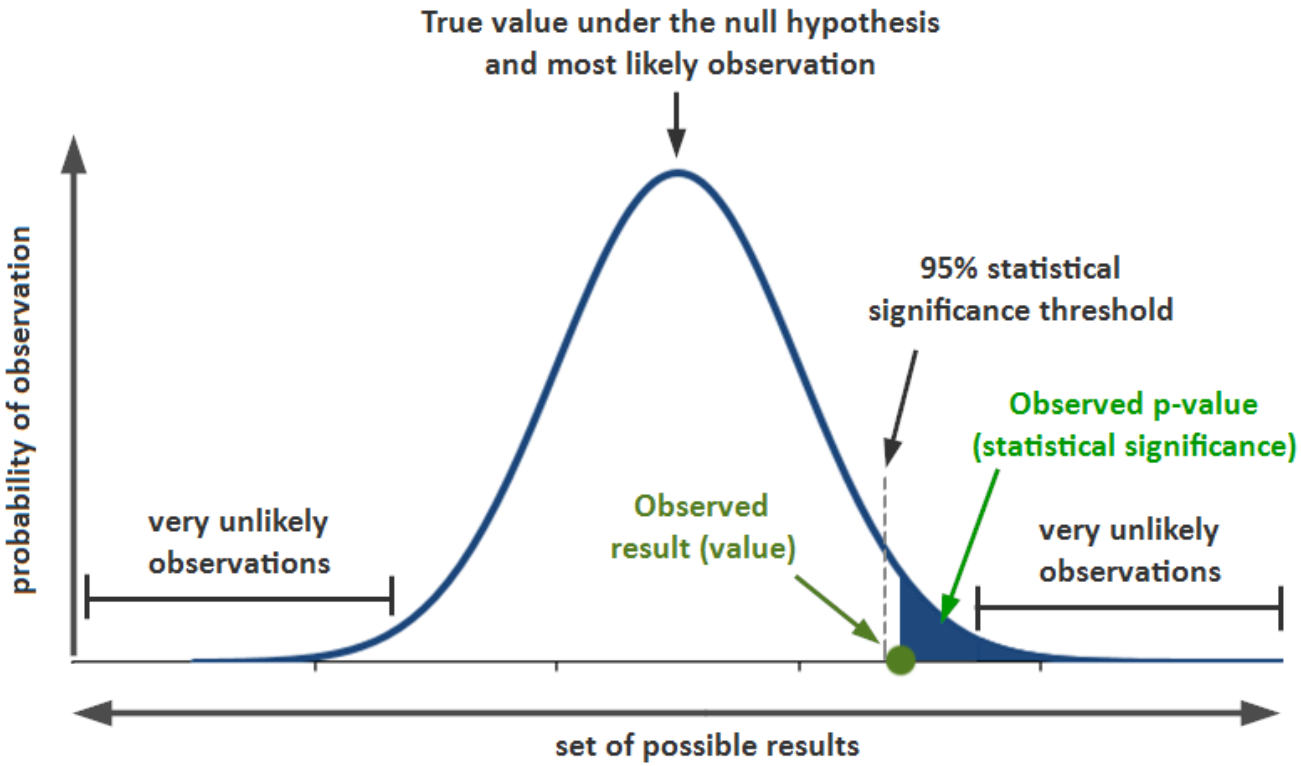
https://math.hws.edu/javamath/ryan/ChiSquare.html

https://medium.com/analytics-vidhya/comprehensive-guide-to-chi-square-tests-for-independence-ff70f5734ad7

Kesimpulan

$$H_0 = independent$$

$$Small\ P - value = !\, H_0$$



Sumber: https://www.simplypsychology.org/p-value.html

Contoh Chi-square

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Question: Are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

Here's the table of expected counts:

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

So, working this out, $\chi^2 = \frac{(60-50.886)^2}{50.886} + \cdots + \frac{(57-48.132)^2}{48.132} = 8.006$

$$\chi^2 = \sum (O-E)^2/E$$

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

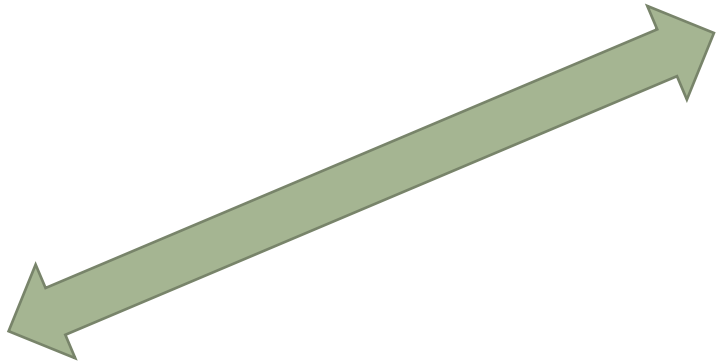
INTRODUCTION TO POPULATION GENETICS, Table D.1

INFORMATIKA UNPAR

Menjelang Ujian:
Ngebut Belajar atau Tidur?

Referensi: Buku DS Bab 2

INFORMATIKA
UNPAR



Apakah ada keterkaitan antara prestasi dengan tidur?

Tahapan Tidur dan Alat Deteksinya

Light Sleep → Deep Sleep → REM

INFORMATIKA UNPAR

Data


Sleep Stages
Nov 14, 2017
Awake
10%
39m
REM
20%
1h 20m
Light
58%
3h 46m
Deep
12%
48m
Awake
REM
Light
Deep
11:24 PM
2 AM
5:57 AM

Eksperimen untuk Pengumpulan Data

88

43

45

Peserta mata kuliah
Pengenalan Kimia Zat Padat

INFORMATIKA
UNPAR

Data Objektif vs Data Subjektif

Pencatatan data dari perangkat Fitbit

Dosen yang sama dan asisten dosen yang berimbang kualitasnya

Soal quiz dan ujian yang sama

CONTOH
Tidur berapa jam?
Tidurnya nyenyak atau tidak?

INFORMATIKA UNPAR

Analisis

Prestasi
VS
Jam Tidur

Prestasi
VS
Durasi

Prestasi
VS
Jam
Bangun

Prestasi
VS
Konsistensi

INFORMATIKA
UNPAR

Prestasi VS Jam Tidur

Bagi jadi 2 kelompok

Lihat perbedaan kedua kelompok

Analisis perbedaan

Apakah orang yang tidur lebih cepat memiliki nilai lebih baik?

INFORMATIKA UNPAR

Bagi Jadi 2 Kelompok

Tidur lebih cepat

Tidur lebih larut

Contoh batasan:

22:00 ; 23:14 ; 00:22 ; 01:47 ; 02:00 ; 02:59 ; 03:12

Median

INFORMATIKA UNPAR

Lihat Perbedaan

Apa betul orang yang tidur lebih cepat memiliki nilai lebih bagus??

Tidur lebih cepat

Rata-rata ($\bar{x}$): 77.25

Tidur lebih larut

Rata-rata ($\bar{x}$): 70.68

INFORMATIKA UNPAR

Faktor 1: Ukuran Sampel

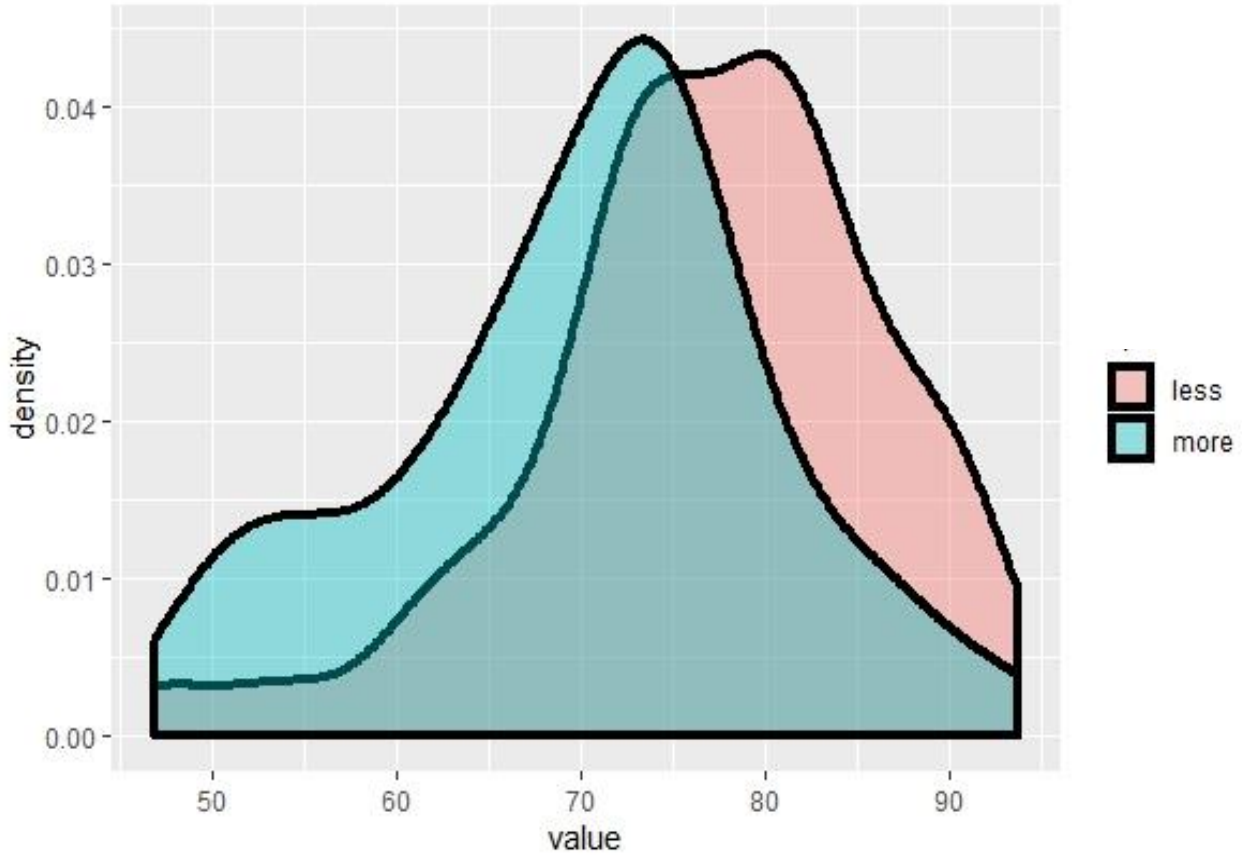
Contoh 1:
Kelompok tidur cepat: 77, 77.5 ($\bar{x}$ = 77.25)
Kelompok tidur larut: 60, 81.36 ($\bar{x}$ = 70.68)

Yang mana yang lebih meyakinkan?

Contoh 2:
Kelompok tidur cepat: 44 siswa dengan $\bar{x} = 77.5$
Kelompok tidur larut: 44 siswa dengan $\bar{x} = 70.68$

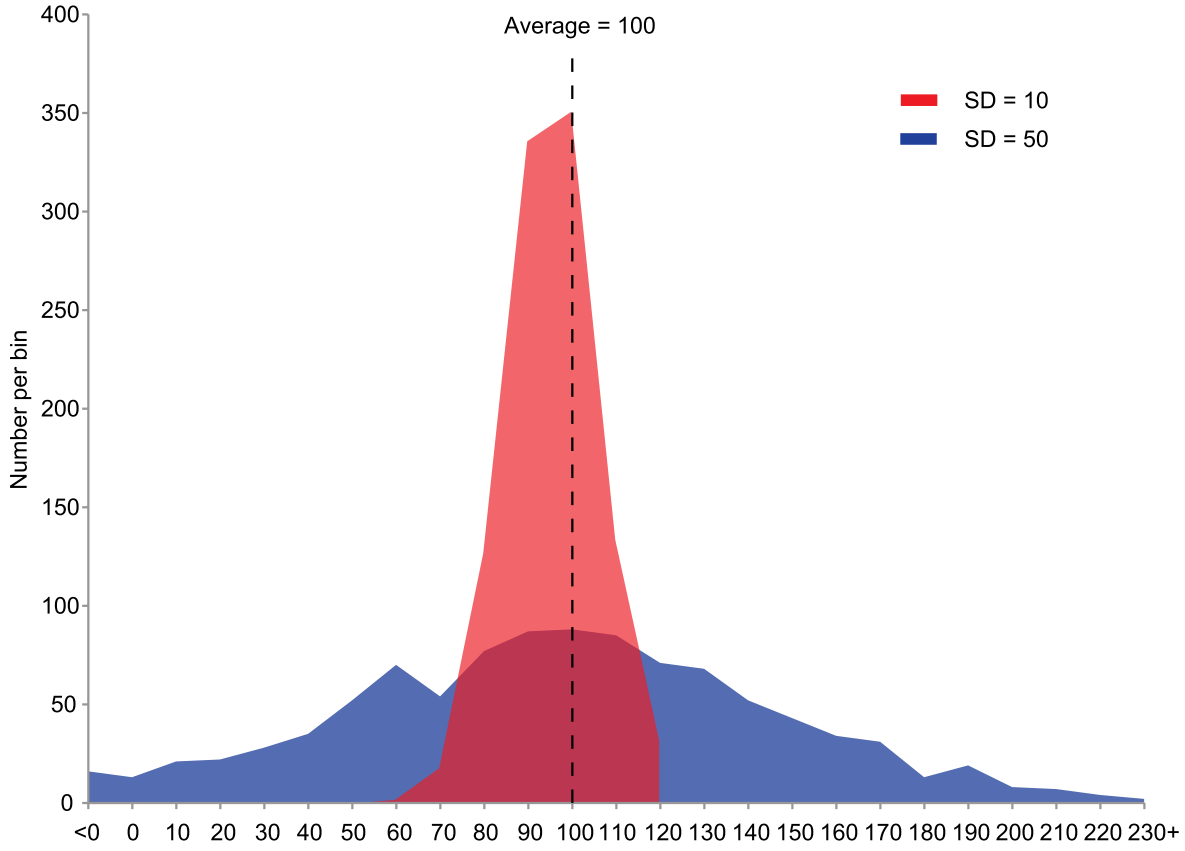
Faktor 2: Distribusi & Simpangan Baku

Jam tidur	Rata-rata	SD
<=1:47 a:m	77.25	13.71
>1:47 a:m	70.68	11.01



INFORMATIKA UNPAR

Ilustrasi Simpangan Baku



INFORMATIKA UNPAR

Analisis Perbedaan: Uji Hipotesa

- Distribusi nilai
- Perbedaan nilai
- Simpangan baku
- Ukuran sampel

$$Small\ P - value = !\ H_0$$

Kelompok yang tidur lebih cepat memiliki nilai yang lebih bagus.

INFORMATIKA UNPAR

Tugas

Lakukan eksplorasi dan *descriptive analysis* terhadap data set *bike sharing* (tersedia di *Google Classroom*). Penjelasan data set sudah disertakan dalam folder.

TOOLS Visualisasi: bebas

Deliverables

1. Laporan dalam bentuk PDF dengan nama file T09_xx.pdf. Sertakan nama anggota kelompok dan NPM dalam laporan.
2. *Workbook* dengan nama file Workbook.xlsx (Ms. Excel), Workbook.twbx (Tableau) dan/atau *workbook* lain. Jika menggunakan Python, kumpulkan *source code*. Beri penomoran pada *sheet* / *dashboard* sesuai penomoran hipotesa atau pertanyaan pada laporan.

Unggah **poin 1** ke *Assignment* **T09** di *Google Classroom*.

Satukan file **poin 2** dalam folder dengan format penamaan T09_xx. Unggah ke *Assignment* **T09-B** di *Google Classroom* dalam bentuk zip.

INFORMATIKA UNPAR

Diskusi, Yuk!

Apakah ada pertanyaan?