


Rencana Tugas Mandiri ke-2
Mata Kuliah Analisis Data Eksploratif
Sub-CMPK-2 Bentuk-bentuk Transformasi Data dan Analisis Data

	Universitas Pembangunan Nasional Veteran Jawa Timur				
	Fakultas Ilmu Komputer				
	Program Studi Sains Data				
Rencana Tugas Mahasiswa					
Mata Kuliah	Analisis Data Eksploratif				
Kode	SD211123	SKS	3	Semester	5
Dosen Pengampu	Tresna Maulana Fahrudin, S.ST., M.T.				
Bentuk Tugas	Studi Kasus				
Judul Tugas	Tugas 2: Eksplorasi Pertanyaan Dasar pada Data, Penggabungan Data (<i>Merging</i>), Transformasi Data, dan Deteksi Outlier				
Sub CPMK	Sub-CMPK-2 Bentuk-bentuk Transformasi Data dan Analisis Data				
Deskripsi Tugas	Tugas ini bertujuan agar mahasiswa mampu untuk melakukan <ol style="list-style-type: none">Eksplorasi Pertanyaan Dasar pada DataPenggabungan data: Inner Join, Left Join, Right Join, dan Outer JoinTransformasi data: Duplikasi Data, <i>Missing Values</i>, dan Imputasi <i>Missing Values</i>Deteksi outlier: InterQuartile Range dan BoxPlot				
Metode Pengerjaan Tugas	1.	Mencari referensi dan menyelesaikan permasalahan studi kasus yang berkaitan dengan eksplorasi pertanyaan dasar, penggabungan data, transformasi data, dan deteksi outlier			
	2.	Tugas dikerjakan secara individu			
	3.	Tugas diupload di <i>e-learning</i> (masing-masing mahasiswa). Tugas dikumpulkan paling lambat pada H-1 (maksimal 23.59 WIB)			
	4.	Tugas diketik pada kertas A4, <i>font</i> ukuran 12, Times New Roman, dan spasi 1. Tidak ada ketentuan jumlah halaman halaman. Tuliskan nama mahasiswa, NPM, dan kelas paralel			
Bentuk dan Format Luaran	Hasil dokumentasi penyelesaian tugas dengan sistematika dan format yang telah ditentukan				
Indikator, Kriteria dan Bobot Penelitian					
Indikator: Ketepatan dalam mencari solusi/teknik/pendekatan untuk menyelesaikan permasalahan berkaitan dengan eksplorasi pertanyaan dasar pada data, penggabungan data, transformasi data, dan deteksi outlier Kriteria: Non-test Bobot: 4% Rubrik: Holistik					
Lain-Lain					
Daftar Pustaka: S. Kumar Mukhiya., U. Ahmed..2020. <i>Hands-on Exploratory Data Analysis with Python</i> . Birmingham: Packt Publishing					

Implementasikan teknik-teknik eksplorasi pertanyaan dasar pada data, penggabungan data, transformasi data, dan deteksi outlier untuk menyelesaikan berbagai permasalahan studi kasus pada dataset kesehatan seperti pasien kanker payudara, pasien hepatitis, dan juga dataset banking.

1. Eksplorasi Pertanyaan Dasar (Studi Kasus: Breast Cancer Dataset)

age	menopa use	tumor- size	inv- nodes	node- caps	deg- malig	breas t	breast- quad	irra diat	Class
'40-49'	'premeno'	'15-19'	'0-2'	'yes'	'3'	'right'	'left_up'	'no'	'recurrence- events'
'50-59'	'ge40'	'15-19'	'0-2'	'no'	'1'	'right'	'central'	'no'	'no- recurrence- events'
'50-59'	'ge40'	'35-39'	'0-2'	'no'	'2'	'left'	'left_low'	'no'	'recurrence- events'
'40-49'	'premeno'	'35-39'	'0-2'	'yes'	'3'	'right'	'left_low'	'yes'	'no- recurrence- events'
'40-49'	'premeno'	'30-34'	'3-5'	'yes'	'2'	'left'	'right_up'	'no'	'recurrence- events'
...

Catatan: breast-cancer.csv (dataset) – beberapa *missing values* pada dataset ini telah dihapus

- Deskripsikan definisi masing-masing atribut pada dataset kanker payudara berdasarkan pengertian di dalam domain keilmuan medis dan referensi yang dapat dipertanggungjawabkan
- Lakukan eksplorasi pada dataset kanker payudara menggunakan Python dan jawablah pertanyaan sebagai berikut:
 - Berapa banyak pasien yang berumur 50-59 tahun dengan derajat keganasan kanker payudara sebesar 2?
 - Berapa banyak pasien dengan ukuran tumor 0-4 (mm) dengan kondisi kelenjar getah bening belum/tidak menembus kapsul dan sekitarnya (*node-caps*)?
 - Berapa banyak pasien dengan tumor yang melakukan terapi radiasi dengan derajat keganasan kanker sebesar 2 dan 3?
 - Berapa banyak pasien yang memiliki tumor yang terletak di sebelah kanan dan kiri serta tepatnya pada kuadran pusat?
 - Berapa banyak pasien yang sedang premenopause dengan kelenjar getah bening yang mengandung kanker payudara pada *range* 6-8?
 - Berapa banyak pasien yang telah melakukan terapi radiasi, tetapi masih ada kemungkinan terjadi kekambuhan ulang?
 - Berapa banyak pasien yang masih berumur 30-39 tahun dengan kondisi kelenjar getah bening telah menembus kapsul dan sekitarnya (*node-caps*)?

- Berapa banyak pasien yang menopause di atas umur 40 tahun, tetapi tidak melakukan terapi radiasi?
- Berapa banyak pasien dengan ukuran tumor sebesar 50-54 (mm) dengan kelenjar getah bening aksila yang mengandung kanker payudara metastatik sebesar 0-2 (mm)?
- Berapa banyak pasien dengan kelenjar getah bening aksila yang mengandung kanker payudara metastatic sebesar 15-17 (mm) dan dengan kondisi kelenjar getah bening belum/tidak menembus kapsul dan sekitarnya (*node-caps*)?

2. Penggabungan Data Menggunakan Inner Join, Left Join, Right Join, dan Outer Join (Studi Kasus: Breast Cancer Dataset)

PatientID	Tumor Size
1	'15-19'
3	'35-39'
5	'30-34'
7	'40-44'
9	'0-4'
11	'25-29'
15	'25-29'
17	'10-14'
19	'40-44'
21	'40-44'
...	...

PatientID	Tumor Size
4	'35-39'
6	'25-29'
8	'10-14'
12	'15-19'
14	'25-29'
16	'20-24'
18	'15-19'
22	'15-19'
24	'15-19'
26	'10-14'
...	...

PatientID	Breast Quadrant
1	'left_up'
3	'left_low'
7	'left_up'
9	'right_low'
11	'left_low'
13	'central'
17	'left_up'
19	'left_up'
21	'left_up'
23	'left_up'
...	...

PatientID	Breast Quadrant
2	'central'
6	'left_up'
8	'left_up'
10	'left_up'
12	'left_up'
14	'left_up'
16	'central'
18	'left_up'
22	'left_low'
24	'left_low'
...

Catatan: `df1tumor_size_odd.csv`, `df2tumor_size_ev.csv`, `df1breastquad_odd.csv`, dan `df2breastquad_even.csv` (dataset)

- Implementasikan penggabungan data (*merge*) menggunakan Inner Join, Left Join, Right Join, dan Outer Join menggunakan Python untuk menjawab pertanyaan berikut:

- Berapa banyak *record* pasien yang memiliki nilai atribut ukuran tumor dan nilai kuadran kanker payudara secara lengkap keduanya ?
- Berapa banyak *record* pasien yang memiliki nilai atribut ukuran tumor, tetapi tidak memiliki nilai kuadran kanker payudara ? serta berapa banyak *record* nilai atribut kuadran kanker payudara yang NaN tersebut ?
- Berapa banyak *record* pasien yang tidak memiliki nilai atribut ukuran tumor, tetapi memiliki nilai kuadran kanker payudara ? serta berapa banyak *record* nilai atribut ukuran tumor yang NaN tersebut ?
- Jika menggunakan Outer Join, Berapa banyak *record* pasien yang NaN pada masing-masing atribut, baik pada atribut ukuran tumor maupun kuadran kanker payudara?

3. Transformasi Data: Duplikasi Data, *Missing Values*, dan Imputasi *Missing Values*

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Age	Sex	Steroid	Antivirals	Fatigue	Malaise	Anorexia	Liver Big	Liver Firm	Spleen Palpable	Speiders	Ascites	Varices	Bilirubin	Alk Phosphate	SGOT	Albumin	Protim e	Histology	CLASS
1	30	2	1	2	2	2	2	1	2	2	2	2	2	1	85	18	4	?	1	2
2	50	1	1	2	1	2	2	1	2	2	2	2	2	0,9	135	42	3,5	?	1	2
3	78	1	2	2	1	2	2	2	2	2	2	2	2	0,7	96	32	4	?	1	2
4	31	1	?	1	2	2	2	2	2	2	2	2	2	0,7	46	52	4	80	1	2
5	34	1	2	2	2	2	2	2	2	2	2	2	2	1	?	200	4	?	1	2
6	34	1	2	2	2	2	2	2	2	2	2	2	2	0,9	95	28	4	75	1	2
7	51	1	1	2	1	2	1	2	2	1	1	2	2	?	?	?	?	?	1	1
8	23	1	2	2	2	2	2	2	2	2	2	2	2	1	?	?	?	?	1	2
9	39	1	2	2	1	2	2	2	1	2	2	2	2	0,7	?	48	4,4	?	1	2
10	30	1	2	2	2	2	2	2	2	2	2	2	2	1	?	120	3,9	?	1	2
11	39	1	1	1	2	2	2	1	1	2	2	2	2	1,3	78	30	4,4	85	1	2
12	32	1	2	1	1	2	2	2	1	2	1	2	2	1	59	249	3,7	54	1	2
13	41	1	2	1	1	2	2	2	1	2	2	2	2	0,9	81	60	3,9	52	1	2
14	30	1	2	2	1	2	2	2	1	2	2	2	2	2,2	57	144	4,9	78	1	2
15	47	1	1	1	2	2	2	2	2	2	2	2	2	?	?	60	?	?	1	2
16	38	1	1	2	1	1	1	2	2	2	2	1	2	2	72	89	2,9	46	1	2
17	66	1	2	2	1	2	2	2	2	2	2	2	2	1,2	102	53	4,3	?	1	2
18	40	1	1	2	1	2	2	2	1	2	2	2	2	0,6	62	166	4	63	1	2
19	38	1	2	2	2	2	2	2	2	2	2	2	2	0,7	53	42	4,1	85	2	2
20	38	1	1	1	2	2	2	1	1	2	2	2	2	0,7	70	28	4,2	62	1	2
21	22	2	2	1	1	2	2	2	2	2	2	2	2	0,9	48	20	4,2	64	1	2

22	27	1	2	2	1	1	1	1	1	1	1	2	2	1,2	133	98	4,1	39	1	2
23	31	1	2	2	2	2	2	2	2	2	2	2	2	1	85	20	4	100	1	2
24	42	1	2	2	2	2	2	2	2	2	2	2	2	0,9	60	63	4,7	47	1	2
25	25	2	1	1	2	2	2	2	2	2	2	2	2	0,4	45	18	4,3	70	1	2
26	27	1	1	2	1	1	2	2	2	2	2	2	2	0,8	95	46	3,8	100	1	2
27	49	1	1	1	1	1	1	2	1	2	1	2	2	0,6	85	48	3,7	?	1	2
28	58	2	2	2	1	2	2	2	1	2	1	2	2	1,4	175	55	2,7	36	1	2
29	61	1	1	2	1	2	2	1	1	2	2	2	2	1,3	78	25	3,8	100	1	2
30	51	1	1	1	1	1	2	2	2	2	2	2	2	1	78	58	4,6	52	1	2
..
..

Catatan: Hepatitis.xls (dataset)

- Deskripsikan definisi masing-masing atribut pada dataset hepatitis berdasarkan pengertian di dalam domain keilmuan medis dan referensi yang dapat dipertanggungjawabkan
- Lakukan transformasi data menggunakan Python dengan langkah-langkah sebagai berikut:
 - Lakukan pemeriksaan apakah terdapat duplikasi data (*row*) pada dataset hepatitis
 - Lakukan analisis deskriptif berapa banyak atribut yang memiliki *missing values*
 - Lakukan imputasi *missing values* dengan berbagai teknik yang efektif dan baik untuk mengisi atribut-atribut yang NaN, misalnya menggunakan *mean*, *median*, *modus*, *clustering*, *regression*, maupun metode taksiran dan prediksi lainnya

4. Deteksi Outlier: InterQuartile Range dan BoxPlot

ATM Name	Transaction Date	No Of Withdrawals	Total amount Withdrawn (in Rupee)	Total amount Withdrawn (in Rupiah)	Weekday
Big Street ATM	01/01/11	50	123800	23136982	Saturday
Mount Road ATM	01/01/11	253	767900	143512831	Saturday
Airport ATM	01/01/11	98	503400	94080426	Saturday
KK Nagar ATM	01/01/11	265	945300	176667117	Saturday
Christ College ATM	01/01/11	74	287700	53768253	Saturday
Big Street ATM	02/01/11	17	52800	9867792	Sunday
Mount Road ATM	02/01/11	194	529300	98920877	Sunday
Airport ATM	02/01/11	67	268600	50198654	Sunday
KK Nagar ATM	02/01/11	260	809400	151268766	Sunday
Christ College ATM	02/01/11	80	300000	56067000	Sunday
Big Street ATM	03/01/11	24	88100	16465009	Monday
Mount Road ATM	03/01/11	246	897100	167659019	Monday
Airport ATM	03/01/11	109	603900	112862871	Monday

KK Nagar ATM	03/01/11	309	1333100	249143059	Monday
Christ College ATM	03/01/11	108	522800	97706092	Monday
Big Street ATM	04/01/11	34	101600	18988024	Tuesday
Mount Road ATM	04/01/11	246	826000	154371140	Tuesday
Airport ATM	04/01/11	113	541600	101219624	Tuesday
KK Nagar ATM	04/01/11	225	999400	186777866	Tuesday
Christ College ATM	04/01/11	143	737400	137812686	Tuesday
Big Street ATM	05/01/11	30	98000	18315220	Wednesday
Mount Road ATM	05/01/11	211	754400	140989816	Wednesday
...

Catatan: ATM-Transaction.csv (dataset)

- Deskripsikan definisi masing-masing atribut pada dataset ATM Transaction berdasarkan pengertian di dalam domain keilmuan *banking* dan referensi yang dapat dipertanggungjawabkan
- Lakukan analisis deskriptif dan deteksi outlier menggunakan Python dan menjawab pertanyaan berikut:
 - Bank mana yang teramai dan tersepi bagi nasabah untuk melakukan penarikan uang?
 - Pada hari apa ATM paling sering dan pada hari apa yang paling jarang dikunjungi oleh nasabah untuk melakukan penarikan uang?
 - Lakukan deteksi *outlier* untuk melihat potensi adanya *fraud* berdasarkan atribut **Total amount Withdrawn (in Rupiah)** pada dataset ATM Transaction menggunakan metode InterQuartileRange dan visualisasikan menggunakan BoxPlot serta sebutkan nama **No Of Withdrawals** dan **nama ATM-nya** yang terdeteksi *fraud*.