# UJIAN TENGAH SEMESTER (UTS)

# DATA MINING

# DATA CLEANING



Oleh:

Gabriel Batubara 14116039

**PROGRAM STUDI TEKNIK INFORMATIKA INSTITUT TEKNOLOGI SUMATERA**

**2020**

# Kementrian Pendidikan dan Kebudayaan INSTITUT TEKNOLOGI SUMATERA

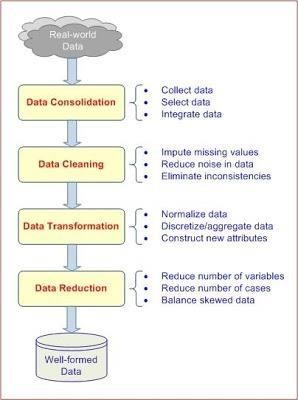
Jalan Terusan Ryacudu, Desa Way Hui, Kecamatan Jati Agung, Lampung Selatan 35365 Telepon : (0721) 8030188, Email : [pusat@itera.ac.id,](mailto:pusat@itera.ac.id) Website : http//[www.itera.ac.id](http://www.itera.ac.id/)

## UJIAN TENGAH SEMESTER

|  |  |  |
| --- | --- | --- |
| Mata Kuliah Mining  Kode Mata Kuliah  Hari/Tanggal | : Data Warehouse / Data  : IF3017  : Rabu, 21 Oktober 2020 | Dosen : Meida Cahyo Untoro, S.Kom.,M.Kom.  Sifat Ujian : Individu dan Buka Catatan  Waktu : 150 Menit |

**SOAL DISKRIPSI**

Maksud dari persiapan data (atau yang lebih dikenal dengan pra-pemrosesan data) adalah mengambil data yang diidentifikasi pada tahap sebelumnya dan menyiapkan nya untuk analisa dengan menggunakan metode-metode DM. Dibandingkan dengan tahapan-tahapan lainnya dalam CRISP-DM, pra-pemrosesan data menyita waktu dan usaha paling banyak; banyak orang percaya bahwa tahap ini bertanggungjawab atas sekitar 80 persen dari total waktu yang diluangkan untuk project DM. Penyebab dari usaha yang sedemikian besar itu yang dihabiskan untuk tahap ini adalah karena data riil (di ‘real-world’) yang ada pada umumnya tidak lengkap (tidakadanya nilai pada atribut-atributnya, tidakadanya atribut tertentu yang menjadi perhatian, atau hanya berisi data yang sudah ringkas dan digabungkan), ‘noisy’ (berisi data yang error atau data yang tidak diinginkan), dan data yang tidak konsisten (berisi data yang berbeda antara kode-kode dan nama- nama). **Gambar 1 menunjukkan empat langkah utama yang dibutuhkan untuk mengonversi data mentah riil menjadi dataset yang bisa digali, jelaskan 4 proses tersebut dan berikan contoh data yang digunakan?**



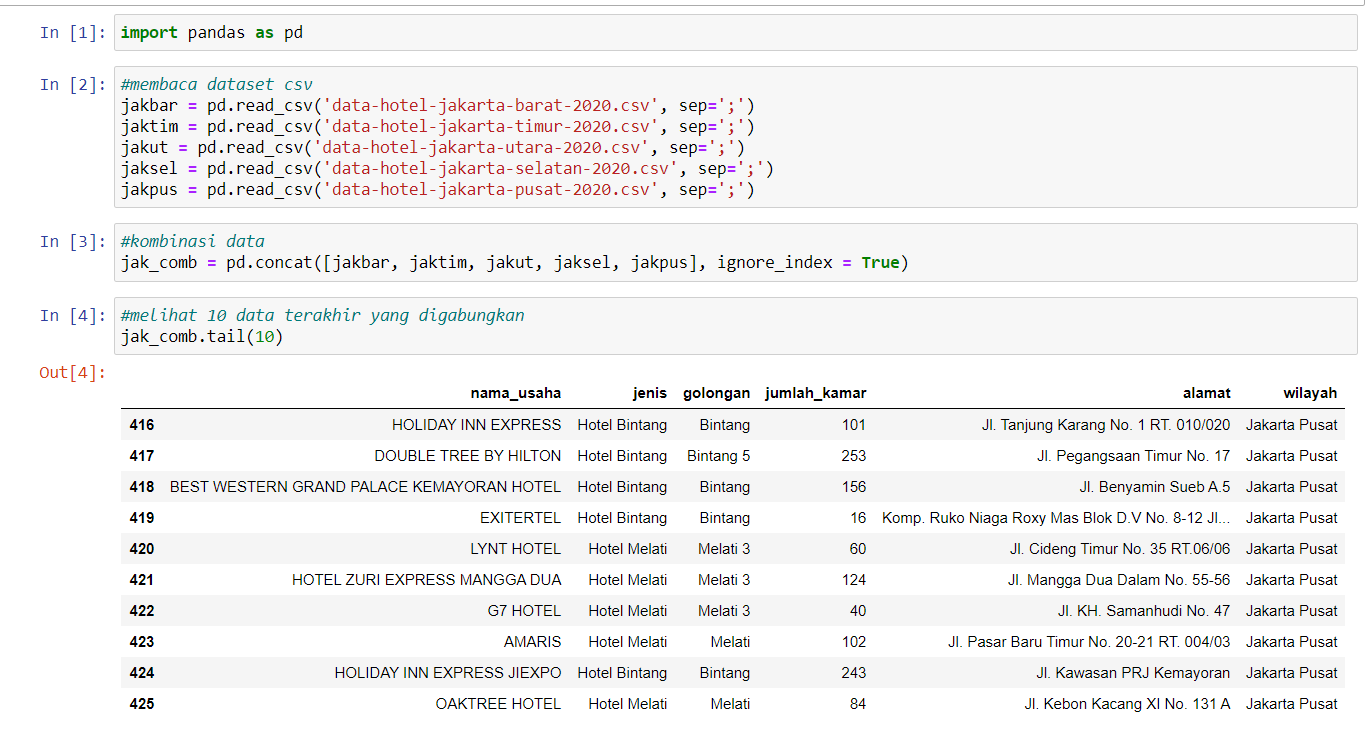
Gambar 1. Tahap-tahap pra-pemrosesan data

|  |  |  |
| --- | --- | --- |
| **Tugas utama** | **Contoh Sub-tugas** | **Contoh Metode yang popular** |
| **Pembersihan data (Data cleaning)** | Mengurusi nilai-nilai yang hilang/kosong dalam data  Mengidentifikasi dan mengurangi noise dalam data  Menemukan dan menghilangkan data yang salah | Mengisi nilai yang hilang (imputations) dengan nilai yang tepat (mean, median, min/max, mode, etc); menulis ulang nilai yang hilang dengan konstanta seperti  “ML”; menghapus baris pada nilai yang hilang; dibiarkan saja.  Mengidentifikasi nilai-nilai yang seharusnya tidak seharusnya di dalam data dengan teknik-teknik statistic sederhana (seperti rerata dan standard deviasi) atau dengan analisa kluster; setelah diidentifikasi bisa menghilangkan nilai-nilai tersebut atau memodifikasi/menghaluskan nya dengan ‘binning’, ‘regression’, atau ‘simple averages’.  Mengidentifikasi nilai-nilai yang salah dalam data (selain data yang tidak seharusnya), seperti nilai-nilai ganjil, labellabel kelas yang tidak konsisten, distribusi  ganjil; |

|  |  |  |
| --- | --- | --- |
| **Tugas utama** | **Contoh Sub-tugas** | **Contoh Metode yang popular** |
|  |  | setelah teridentifikasi, gunakan domain expertise 9si ahlinya dari domain tersebut) untuk membetulkan nilai-nilai tersebut atau menghapus baris-baris yang memiliki nilai yang salah tersebut. |

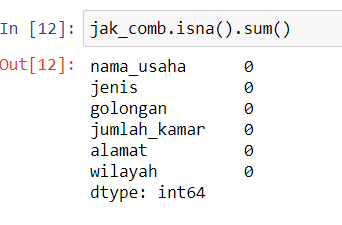
## Ketentuan dalam menggerjakan Diskripsi Kasus Data mining pada periode Ujian Tengah Semester 2020/2021

1. Dataset yang digunakan merupakan hasil dari pencarian data secara real baik langsung ataupun dari internet.
2. Tugas utama dilakukan dengan menggunakan jupyter notebook atau Google Colab
3. Dapat menggunakan SQL atau yang lain dalam preprosesing sesuai dengan kebutuhan dan mekanisme metode yang dipilih
4. Setiap kelompok tidak diperkenankan dalam menangani atau menyelesaikan UTS menggunakan Tugas Utama yang sama (Misalnya, 2 Orang menggunakan Data Cleaning)
5. Penilaian akan dilakukan berdasarkan tool yang dipakai dan explorasi detail serta penjelesan yang tepat.



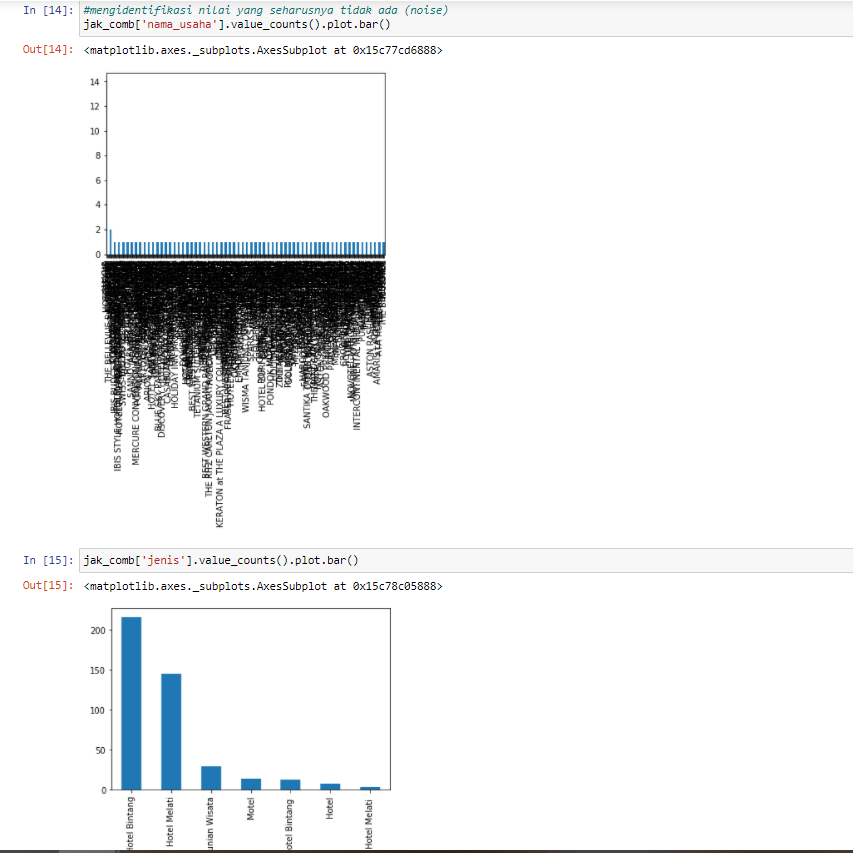
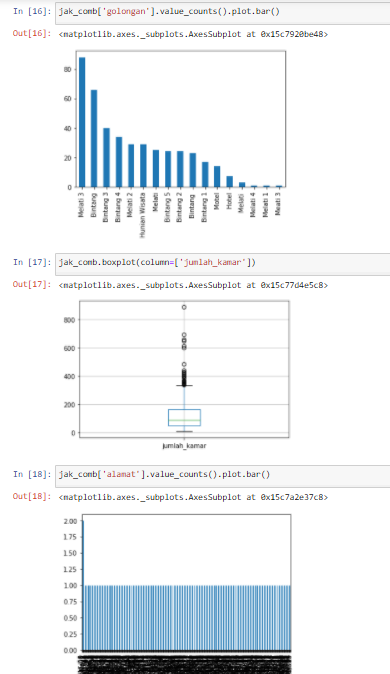
Gambar 1

Pada Gambar 1, kita memasukkan library yang digunakan dan membaca dataset yang digunakan. Pada kasus ini data yang say gunakan dipecah kemudian disatukan kembali ke dalam variable ‘jak\_comb’ dan kemudian menampilkan datanya.

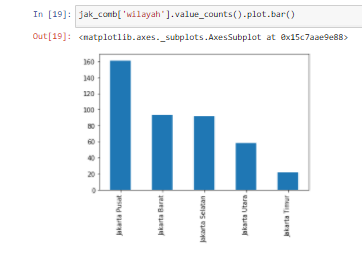


Gambar 2

Selanjutnya pada gambar 2 kita mencari missing value pada data, pada kasus ini data yang digunakan tidak terdapat ‘missing value’.

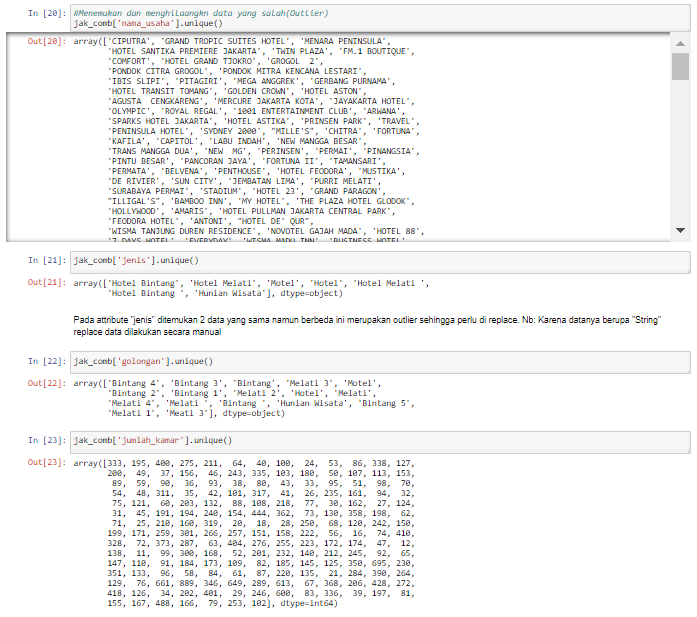
 

Gambar 3 (a) dan (b)



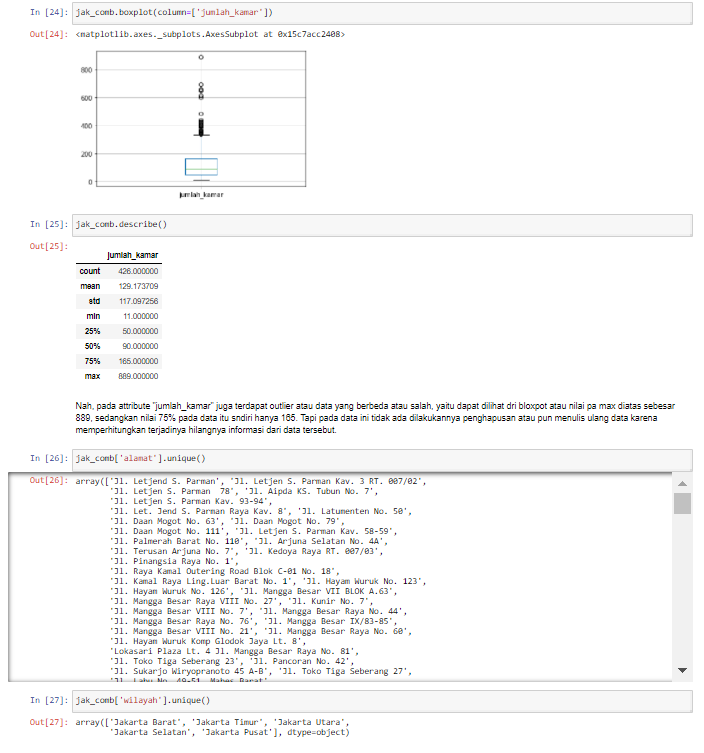
Gambar 3(c)

Selanjutnya mencari nilai yang seharusnya tidak ada (noise), cara mencarinya yaitu dengan memviualisasikan datanya dari setiap attribut yang ada, nah dari ketiga gambar tersebut yaitu gambar a, b, dan c dapat diidentifikasikan bahwa pada setiap atribut pada data yang digunakan tidak ada ditemukannya data noise.



Gambar 4

Selanjutnya adalah mencari data yang salah (outlier) pada data, pada gambar 4 dapat ditemukan data yang salah pada atribut ‘jenis’ yaitu value ‘Hotel Bintang’ dengan ‘Hotel Bintang ’. Pada kasus saya data harus di-*relpace,* namun karena pada atribut ini datanya berupa ‘String’ maka cara me-*replace* nya saya lakukan dengan cara manual tidak dengan coding karena keterbatasan saya dalam menggunakan coding python.



Gambar 5

Untun attribut selanjutnya dapat dilihat pada gambar 5. Pada gambar diatas dapat ditemukan lagi *outlier* pada atribut ‘jumlah\_kamar’. Nah, pada attribute ‘jumlah\_kamar’ terdapat outlier atau data yang salah, yaitu dapat dilihat dari bloxpot atau nilai max diatas sebesar 889, sedangkan nilai 75% pada data itu sendiri hanya 165. Tapi pada data ini tidak ada dilakukannya penghapusan atau pun menulis ulang data karena memperhitungkan terjadinya hilangnya bebrapa informasi dari data tersebut.