TUGAS 2:
DATA EXPLORATION AND PREPROCESSING
Due Date: 13 Oktober 2020, pukul 24.00

Kerjakan tugas ini secara kelompok menggunakan Python (Jupyter notebook). Jupyter notebook dikumpulkan lewat ELOK paling lambat hari Selasa, 13 Oktober 2020, pukul 24.00. Jika anda mengerjakan menggunakan Google Colab, pada file yang diunggah tuliskan url dari notebooknya.

Untuk tugas ini diberikan data dari penyewaan video, yang memuat 50 pelanggan reguler. Setiap pelanggan memiliki atribut: *Gender, Income, Age, Rentals* (jumlah video yang disewa pada tahun lalu), *Avg per Visit* (rata-rata video yang disewa per kunjungan selama tahun lalu), *Insidentals* (apakah pelanggan cenderung membeli sesuatu saat menyewa video), dan *Genre* (genre film pilihan pelanggan). File excel Video_Store.xls bisa diunduh dari link di ELOK.

A. Data Exploration

a. Hitung nilai rata-rata, median, dan mode atribut *Income* dan *Age*. Dari ketiga nilai tersebut, tentukan distribusi atribut *Income* dan *Age* miring kekanan (positively skewed) miring kekiri (negatively skewed), atau simetri.
b. Gambarkan histogram dari atribut *Income* dan *Age*. Jelaskan apakah distribusi dari atribut *Income* dan *Age* berdasarkan histogram sama dengan hasil di soal (a).
c. Hitung nilai minimum, maksimum, Q1, Q2, Q3 dari atribut *Income* dan *Age.*
d. Gambarkan boxplot dari *Income* dan *Age.*
e. Hitung rata-rata *Income* pelanggan pria (M) dan pelanggan wanita (F). Rata-rata *Income* mana yang lebih tinggi dari rata-rata keseluruhan pelanggan?
f. Tentukan jenis film (*Genre*) yang paling banyak dipinjam oleh pelanggan pria dan pelanggan wanita.
g. Bandingkan rata-rata *Income* pelanggan pria dan wanita yang menyukai film berjenis komedi (*comedy*).

h. Hitung koefisien korelasi antar atribut numerik dalam data (*Income, Age, Rentals*, *Avg per Visit*). Tentukan atribut mana yang memiliki koefisien korelasi terbesar dan terkecil, dan jelaskan maknanya.
i. Visualisasikan scatter plot antar keempat atribut, yaitu *Income, Age, Rentals*, *Avg per Visit*.
j. Pilih customer yang berharga (valued customer), yaitu customer yang memiliki nilai tinggi pada atribut *Rentals* >= 30. Bandingkan karakteristik dari customer ini dengan karakteristik dari customer lain (yang bukan *valued customer*) dalam hal *Income* dan *Age*.

B. Data Preprocessing

Lakukan data preprocessing pada atribut dari file **Video-Store.xls**. Tambahkan kolom baru dalam file excel sebagai hasil dari preprocessing soal (a), (b), (c), (d), (e).

a. Lakukan smoothing dengan metode *binning* rata-rata pada atribute **Age**. Gunakan bin berukuran 4 (bin depth). Buat grafik atribut **Age** sebelum dan sesudah smoothing.
b. Lakukan *min-max normalization* untuk mentransformasi nilai atribut **Income** ke dalam range [0.0-1.0].
c. Lakukan *z-score normalization* untuk menstandarkan nilai atribut **Rentals**.
d. Ubah atribut **Income** (yang belum dinormalisasi) menjadi diskrit menggunakan kategori berikut:: High = 60K+; Mid = 25K-59K; Low = kurang dari $25K. Hitung jumlah pelanggan untuk masing-masing kategori **Income**.
e. Konversi semua atribut kategorik (*Gender, Incidentals, Genre*) menjadi atribut kontinyu.
f. Buat distance matrik dari pelanggan 1 sampai pelanggan 5 (ukuran matrik adalah 5 x 5) sebelum dan sesudah normalisasi menggunakan min-max (b) dan z-score (c). NOTE: jarak dihitung menggunakan atribut **Income** dan **Rentals**. Apakah distance matrix yang dihasilkan sama?
g. Dalam sheet2 ada data yang nilainya tidak ada (missing value). Isi data yang tidak ada tersebut dengan nilai yang sesuai.