NAMA : RISKI ABDILAH PRATAMA

NIM : 20220040091

KELAS : TI22C

TUGAS 5_DATA SCIENCE

TUGAS BAGIAN 1

1. Jelaskan apa yang dimaksud dengan Exploratory Data Analysis (EDA) dan mengapa hal ini penting dalam analisis data?
2. Apa perbedaan antara data kategorikal dan data numerik? Berikan contoh masing-masing.
3. Bagaimana Anda mengidentifikasi dan menangani missing values dalam sebuah dataset?
4. Apa itu outlier dalam konteks analisis data? Mengapa penting untuk mendeteksi dan menangani outlier?

TUGAS BAGIAN 2

Buatlah contoh studi kasus dan selesaikan berdasarkan tahapan proses data understanding.

JAWAB:

Penjelasan Tentang Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) adalah proses awal dalam analisis data yang digunakan untuk memahami dan mengeksplorasi dataset. Tujuan utama EDA adalah untuk memahami struktur data, mendeteksi pola, outlier, dan mengidentifikasi masalah kualitas data seperti missing values. EDA melibatkan penggunaan teknik statistik deskriptif dan visualisasi untuk meringkas karakteristik utama dari dataset tanpa membuat asumsi apa pun.

Pentingnya EDA dalam Analisis Data:

1. **Memahami Distribusi Data**: EDA membantu memahami distribusi dan pola data. Ini memungkinkan kita untuk menentukan apakah data tersebut normal, miring, atau mengandung outlier.
2. **Mengidentifikasi Missing Values**: Menemukan dan menangani missing values sangat penting sebelum melanjutkan analisis yang lebih lanjut.
3. **Mendeteksi Outlier**: EDA membantu mendeteksi outlier, yang dapat mempengaruhi hasil analisis.
4. **Memilih Model yang Tepat**: Dengan memahami hubungan antar variabel, kita bisa memilih teknik modeling yang tepat.
5. **Menghasilkan Hipotesis**: EDA sering membantu menghasilkan hipotesis awal tentang variabel mana yang relevan dan bagaimana hubungan antar variabel bekerja.

Perbedaan Antara Data Kategorikal dan Data Numerik

1. **Data Kategorikal**:
    - **Pengertian**: Data yang berisi kategori atau label. Kategori ini bisa bernilai nominal (tanpa urutan) atau ordinal (dengan urutan).
    - **Contoh**:

        - **Nominal**: Jenis kelamin (Pria, Wanita), Warna mobil (Merah, Biru, Hijau).
        - **Ordinal**: Tingkat pendidikan (SMA, S1, S2, S3).
2. **Data Numerik**:
    - **Pengertian**: Data yang berbentuk angka dan dapat dihitung atau diukur. Data ini dapat berupa data diskrit (nilai yang dapat dihitung) atau data kontinu (nilai yang bisa mengambil setiap angka dalam rentang tertentu).
    - **Contoh**:
        - **Diskrit**: Jumlah anak dalam keluarga, jumlah mobil.
        - **Kontinu**: Tinggi badan, berat badan, suhu.

Cara Mengidentifikasi dan Menangani Missing Values

Identifikasi:

1. **Metode Visualisasi**: Gunakan heatmap atau bar plot untuk melihat pola missing values.
2. **Metode Statistik**: Gunakan fungsi seperti .isnull().sum() dalam Python untuk menghitung jumlah missing values pada setiap kolom.

Penanganan:

1. **Menghapus Data**: Jika jumlah missing values kecil dan tidak signifikan, kita dapat menghapus baris atau kolom yang memiliki missing values.
2. **Mengisi Data (Imputasi)**:
    - **Mean/Median**: Untuk data numerik, missing values dapat diisi dengan mean atau median.
    - **Mode**: Untuk data kategorikal, missing values dapat diisi dengan mode (kategori yang paling sering muncul).
3. **Menggunakan Model**: Jika missing values signifikan, bisa menggunakan model prediktif untuk mengisi data yang hilang berdasarkan nilai variabel lain.
4. **Mark as Missing**: Dalam beberapa kasus, kita dapat menandai nilai yang hilang sebagai kategori tersendiri.

Apa Itu Outlier dan Pentingnya dalam Analisis Data

Outlier adalah titik data yang berada jauh dari sebagian besar data lainnya dalam suatu dataset. Outlier bisa disebabkan oleh kesalahan pengukuran, entri data yang salah, atau fenomena yang sebenarnya ada.

Pentingnya Mendeteksi dan Menangani Outlier:

1. **Pengaruh pada Statistik**: Outlier dapat mempengaruhi nilai mean, variansi, dan korelasi, yang menyebabkan kesimpulan yang salah.
2. **Mengganggu Model Prediktif**: Outlier bisa membuat model prediktif tidak akurat, terutama dalam regresi dan machine learning.
3. **Kualitas Data**: Menangani outlier penting untuk meningkatkan kualitas dan akurasi analisis data.

Cara Menangani Outlier:

1. **Menghapus Outlier**: Jika outlier disebabkan oleh kesalahan atau entri data yang tidak valid.
2. **Transformasi Data**: Gunakan transformasi seperti logaritma atau square root untuk mengurangi efek outlier.
3. **Menggunakan Metode Robust**: Algoritma seperti median atau regresi robust lebih tahan terhadap outlier.

Tugas Bagian 2: Studi Kasus dengan Tahapan Proses Data Understanding

Studi Kasus: Analisis Penjualan Produk di Toko Elektronik

Latar Belakang:
Sebuah toko elektronik ingin meningkatkan penjualan produk mereka. Toko ini memiliki dataset yang mencakup informasi seperti:

- Jenis produk
- Harga produk
- Jumlah produk yang terjual
- Tanggal penjualan
- Lokasi toko

Tujuan:
Memahami pola penjualan produk dan mengidentifikasi faktor-faktor yang mempengaruhi penjualan untuk membuat keputusan strategis.

Tahapan Proses Data Understanding

1. **Pengumpulan Data**: Dataset yang tersedia mencakup informasi dari berbagai cabang toko selama 2 tahun terakhir.
2. **Eksplorasi Data (EDA)**:
    - **Statistik Deskriptif**: Hitung mean, median, dan variansi untuk jumlah penjualan dan harga produk.
    - **Visualisasi Data**:
        - Buat histogram untuk melihat distribusi harga dan jumlah penjualan.
        - Buat bar chart untuk melihat kategori produk yang paling banyak terjual.
    - **Deteksi Missing Values**:
        - Cek apakah ada missing values pada harga atau jumlah produk yang terjual. Jika ada, imputasi dengan nilai mean (untuk harga) atau median (untuk jumlah penjualan).
    - **Deteksi Outlier**: Gunakan boxplot untuk mendeteksi outlier pada harga produk atau jumlah penjualan.
3. **Pemodelan Awal**:
    - Buat scatter plot untuk melihat hubungan antara harga produk dan jumlah penjualan.
    - Gunakan regresi linier sederhana untuk memodelkan hubungan antara harga produk dan jumlah penjualan.
4. **Menghasilkan Hipotesis**:
    - Apakah harga yang lebih tinggi mengurangi jumlah penjualan?
    - Apakah produk dengan kategori tertentu lebih laku dibandingkan yang lain?
5. **Evaluasi Awal**:
    - Lakukan evaluasi berdasarkan visualisasi dan hasil pemodelan.

- Jika ditemukan masalah seperti outlier atau distribusi yang tidak normal, lakukan penyesuaian sebelum melakukan analisis lebih lanjut.

Kesimpulan Sementara: Berdasarkan analisis awal, produk dengan harga yang lebih tinggi tampaknya memiliki penjualan yang lebih rendah, tetapi ada beberapa kategori produk yang tetap laku meskipun harganya tinggi. Selanjutnya, dapat dilakukan analisis lanjutan untuk memahami faktor-faktor seperti lokasi toko dan tren musiman yang mempengaruhi penjualan.