Laporan ini merupakan bagian dari tugas mata kuliah

Data Science sesi 5

Dengan Dosen Pengampu

GINA PURNAMA INSANY, S.ST., M.Kom

Nama : SAILA JULIA

Kelas : TI22A

NIM : 20220040082

PROGRAM STUDI S1 TEKNIK INFORMATIKA
UNIVERSITAS NUSA PUTRA SUKABUMI

DATA SCIENCE SESI 5

Jelaskan apa yang dimaksud dengan Exploratory Data Analysis (EDA) dan mengapa hal ini penting dalam analisis data? Apa perbedaan antara data kategorikal dan data numerik? Berikan contoh masing-masing. Bagaimana Anda mengidentifikasi dan menangani missing values dalam sebuah dataset? 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.

JAWABAN

1. Apa yang dimaksud dengan Exploratory Data Analysis (EDA) dan mengapa hal ini penting dalam analisis data?

Exploratory Data Analysis (EDA) adalah proses analisis data yang digunakan untuk memahami karakteristik data yang dimiliki sebelum menerapkan metode statistik atau model machine learning lebih lanjut. EDA melibatkan penggunaan teknik statistik deskriptif dan visualisasi data (seperti grafik dan diagram) untuk mendapatkan wawasan awal tentang pola, anomali, hubungan antar variabel, serta distribusi data.

Pentingnya EDA:
- Memahami Struktur Data: EDA membantu mengidentifikasi struktur, pola, dan distribusi data.
- Deteksi Outlier dan Missing Values: Mendeteksi nilai yang anomali (outliers) dan nilai yang hilang (missing values) yang dapat memengaruhi hasil analisis.
- Penyederhanaan Data: Mengidentifikasi variabel yang paling penting dan mengurangi kompleksitas dataset.
- Menghasilkan Hipotesis: Mengarahkan pada pertanyaan yang lebih spesifik untuk diuji dalam analisis lanjutan.
- Memvalidasi Asumsi: Memeriksa asumsi yang akan digunakan dalam pemodelan statistik atau machine learning.

2. Apa perbedaan antara data kategorikal dan data numerik? Berikan contoh masing-masing.
- Data Kategorikal : adalah data yang terdiri dari kategori atau label yang tidak memiliki urutan yang inheren atau makna numerik. Data ini digunakan untuk mengelompokkan data ke dalam kategori tertentu.
  - Contoh: Jenis kelamin (Pria, Wanita), Warna (Merah, Biru, Hijau), Status Pernikahan (Lajang, Menikah, Cerai).

- Data Numerik: adalah data yang dapat dihitung atau diukur dan terdiri dari angka. Data ini dapat dibagi lagi menjadi dua jenis: diskrit (hanya mengambil nilai tertentu) dan kontinu(mengambil nilai dalam rentang yang tak terbatas).
  - Contoh: Usia (25, 30, 35), Berat badan (55.5 kg, 70.8 kg), Jumlah penjualan (10, 15, 20).

3. Bagaimana Anda mengidentifikasi dan menangani missing values dalam sebuah dataset?

Identifikasi Missing Values:
- Menggunakan fungsi bawaan dari bahasa pemrograman (seperti `.isnull()` di Python Pandas) untuk memeriksa kolom yang memiliki nilai kosong.
- Menampilkan statistik deskriptif dari dataset yang dapat mengindikasikan adanya data yang hilang.

Cara Menangani Missing Values:
1. Hapus Missing Values:
   - Row Deletion: Menghapus baris yang memiliki nilai kosong jika jumlahnya tidak signifikan.
   - Column Deletion: Menghapus kolom jika sebagian besar datanya hilang.
2. Imputasi:
   - Mean/Median/Mode Imputation: Mengisi nilai yang hilang dengan nilai rata-rata, median, atau modus dari kolom tersebut.

- Predictive Imputation: Menggunakan algoritma regresi atau machine learning untuk memprediksi nilai yang hilang.
- KNN Imputation: Menggunakan teknik K-Nearest Neighbors untuk mengisi nilai yang hilang berdasarkan kedekatan data lain.

4. Apa itu outlier dalam konteks analisis data? Mengapa penting untuk mendeteksi dan menangani outlier?
Outlier adalah nilai atau titik data yang berbeda jauh dari nilai-nilai lain dalam dataset. Outlier bisa muncul karena kesalahan data atau karena data tersebut memang merepresentasikan kejadian yang jarang terjadi.

Pentingnya Deteksi dan Penanganan Outlier:
- Menghindari Distorsi: Outlier dapat mendistorsi hasil analisis statistik, seperti rata-rata dan regresi.
- Identifikasi Kesalahan Data: Deteksi outlier membantu mengidentifikasi kesalahan input data yang perlu diperbaiki.
- Menyoroti Fenomena Menarik: Kadang-kadang outlier menunjukkan fenomena yang jarang atau unik yang perlu dianalisis lebih lanjut.

2. Studi Kasus Berdasarkan Tahapan Proses Data Understanding

Studi Kasus: Analisis Penjualan pada Bisnis Properti

Seorang analis data ditugaskan untuk menganalisis data penjualan dari sebuah perusahaan properti. Tujuan dari analisis ini adalah untuk memahami pola penjualan, mengidentifikasi jenis properti yang paling diminati, serta memberikan rekomendasi strategis untuk meningkatkan penjualan di masa depan.

Tahapan Data Understanding:

1. Data Collection (Pengumpulan Data)
Perusahaan mengumpulkan data dari sistem penjualan mereka selama 2 tahun terakhir. Data tersebut mencakup informasi tentang properti yang terjual, termasuk lokasi, tipe properti, luas bangunan, harga jual, serta tanggal transaksi. Berikut adalah beberapa atribut utama dalam dataset:
- Property_ID: ID unik untuk setiap properti.
- Property_Type: Jenis properti (misalnya, apartemen, rumah tapak, ruko, vila).
- Location: Lokasi properti (nama kota atau area).
- Building_Area: Luas bangunan (dalam meter persegi).
- Land_Area: Luas tanah (dalam meter persegi).
- Price: Harga jual properti.
- Sale_Date: Tanggal terjadinya penjualan.

2. Data Description (Deskripsi Data)
- Melakukan deskripsi statistik untuk memahami distribusi dan karakteristik data. Langkah ini mencakup:
  - Rata-rata, median, minimum, dan maksimum dari harga jual properti.
  - Distribusi luas bangunan dan tanah untuk melihat kisaran dan tren ukuran properti.
  - Jumlah penjualan berdasarkan jenis properti dan lokasi untuk memahami area dan tipe properti yang paling laris.

3. Data Exploration (Eksplorasi Data)
- Menggunakan Exploratory Data Analysis (EDA) untuk mendapatkan wawasan lebih dalam tentang data:
  - Membuat grafik bar untuk melihat perbandingan jumlah penjualan berdasarkan tipe properti (apartemen, rumah, ruko, dll.) dan lokasi.
  - Menggunakan scatter plot untuk memvisualisasikan hubungan antara harga dan luas bangunan. Misalnya, apakah luas bangunan yang lebih besar selalu berhubungan dengan harga yang lebih tinggi?
  - Membuat heatmap untuk memeriksa korelasi antara harga, luas bangunan, dan luas tanah.
  - Menganalisis pola penjualan berdasarkan waktu (misalnya, bulan atau musim) untuk mengetahui kapan

penjualan properti paling tinggi terjadi.

4. Data Quality Verification (Verifikasi Kualitas Data)
- Memeriksa missing values di atribut penting seperti Price, Building_Area, atau Location. Misalnya, jika ada transaksi tanpa harga jual, data tersebut mungkin perlu dihapus atau diperiksa ulang.
- Deteksi dan penanganan outlier pada data harga. Misalnya, jika ada harga jual yang sangat tinggi atau rendah dibandingkan dengan properti serupa, perlu dicari tahu apakah itu memang kesalahan data atau karena adanya fitur khusus pada properti tersebut.
- Memastikan konsistensi format pada kolom Sale_Date agar mudah dianalisis berdasarkan waktu (bulanan atau tahunan).
- Menghilangkan data duplikat, jika ada transaksi yang dicatat lebih dari sekali.

Hasil Analisis dan Wawasan:
Setelah melakukan eksplorasi dan verifikasi data, didapatkan beberapa wawasan penting:
1. Tipe Properti Paling Laris: Rumah tapak (landed house) dan apartemen mendominasi penjualan dalam 2 tahun terakhir, dengan penjualan rumah tapak sedikit lebih tinggi.
2. Lokasi Terpopuler: Penjualan tertinggi terjadi di kota-kota besar seperti Jakarta, Surabaya, dan Bandung. Ada tren peningkatan penjualan di area pinggiran kota karena harga yang lebih terjangkau.
3. Pola Musiman: Penjualan properti meningkat signifikan menjelang akhir tahun, terutama selama kuartal keempat, mungkin karena adanya bonus tahunan dan promo akhir tahun.
4. Faktor Harga: Ada korelasi positif antara luas bangunan dan harga, tetapi beberapa properti dengan luas kecil di area premium memiliki harga yang sangat tinggi.

Rekomendasi Strategis:
1. Fokus pada Tipe Properti Populer: Perusahaan dapat meningkatkan persediaan rumah tapak dan apartemen di area yang paling diminati oleh pembeli.
2. Strategi Promosi di Kuartal Keempat: Mengadakan promosi atau diskon khusus menjelang akhir tahun untuk menarik lebih banyak pembeli.
3. Ekspansi ke Area Pinggiran: Dengan meningkatnya penjualan di area pinggiran, perusahaan dapat mempertimbangkan untuk memperbanyak properti di lokasi tersebut, dengan tetap menawarkan harga yang terjangkau.
4. Penentuan Harga Berdasarkan Lokasi: Menyesuaikan strategi penetapan harga dengan memperhatikan tren harga di berbagai lokasi dan preferensi pasar. Properti di lokasi premium dapat dijual dengan harga lebih tinggi meski luas bangunan tidak terlalu besar.