LAPORAN TUGAS KECIL 2
IF3170 INTELIGENSIA BUATAN

Exploratory Data Analysis

Tahun Akademik 2023/2024

Anggota Kelompok:

Hidayatullah Wildan Ghaly Buchary	13521015
Ahmad Nadil	13521024

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG

2023/2024

DAFTAR ISI

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang

Di era data yang berkembang pesat ini, memiliki kemampuan untuk menggali dan memahami informasi dari dataset menjadi sangat penting. Exploratory Data Analysis (EDA) merupakan langkah awal yang krusial dalam proses analisis data. EDA memberikan pemahaman mendalam tentang aspek-aspek kunci dari dataset, termasuk distribusi, tren, dan pola yang mungkin tidak langsung terlihat. Dalam dunia yang data-driven, teknik-teknik EDA membantu pengambilan keputusan berbasis data dan pengembangan model prediktif yang lebih akurat.

Dataset yang menjadi fokus pada laporan ini berkaitan dengan karakteristik spesifikasi ponsel, yang mencakup berbagai fitur seperti baterai, kapasitas memori, prosesor, dan lainnya. Pemahaman mendalam terhadap data ini sangat bermanfaat, terutama dalam industri telekomunikasi yang kompetitif, dimana penentuan harga yang strategis dan pemahaman konsumen dapat memberikan keunggulan kompetitif yang signifikan.

1.2. Rumusan Masalah

Berdasarkan kebutuhan untuk melakukan analisis eksploratif pada dataset spesifikasi ponsel, terdapat beberapa pertanyaan analisis yang akan dijawab:

a. Bagaimana distribusi statistik dasar dari fitur-fitur numerik dalam dataset?
b. Apakah terdapat nilai duplikat dalam dataset yang dapat mempengaruhi analisis?
c. Berapa banyak nilai yang hilang pada dataset, dan bagaimana sebaiknya kita menanganinya?
d. Apakah terdapat outlier pada fitur-fitur numerik, dan bagaimana pengaruhnya terhadap dataset?
e. Bagaimana distribusi data pada fitur numerik, dan apa yang dapat dikatakan tentang kurtosis dari distribusi tersebut?
f. Bagaimana distribusi fitur kategorikal dan apa insight yang bisa diperoleh dari distribusi tersebut?

g. Apakah terdapat korelasi antara fitur-fitur tersebut dengan kolom target yang bisa membantu dalam prediksi rentang harga?

1.3. Tujuan Penulisan

Tujuan dari laporan ini adalah untuk menerapkan teknik-teknik EDA pada dataset spesifikasi ponsel untuk memperoleh insight yang mendalam. Laporan ini bertujuan untuk:

a. Mengidentifikasi karakteristik utama dataset melalui statistik dasar.
b. Menemukan dan menangani nilai duplikat dan nilai yang hilang untuk memastikan integritas analisis.
c. Mendeteksi outlier dan mengevaluasi pengaruhnya terhadap dataset.
d. Menganalisis distribusi data fitur numerik dan kategorikal untuk menentukan aspek kritis yang mempengaruhi rentang harga.
e. Menjelajahi korelasi antar fitur dengan kolom target untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap penetapan rentang harga ponsel.

BAB II
PEMBAHASAN

2.1. Statistik Dasar

Statistik dasar merupakan alat esensial dalam analisis data, memberikan wawasan penting tentang karakteristik sentral dan variasi dalam suatu dataset. Melalui penggunaan statistik deskriptif, kita dapat merangkum dan menginterpretasikan kumpulan data yang besar dengan cepat, mengidentifikasi tren dan pola yang mungkin tidak segera terlihat. Dalam sub bab ini, kita akan menjelajahi prinsip-prinsip statistik dasar menggunakan data yang terkumpul dalam sebuah file ipynb, memberikan visualisasi yang mudah diinterpretasikan dan membantu kita memahami distribusi data kita dengan lebih baik.

Pertama, ada statistik deskriptif untuk data kuantitatif. Rata-rata, median, dan modus adalah ukuran tendensi sentral yang menggambarkan "titik tengah" dari sebuah dataset. Rata-rata memberikan nilai tengah yang dihitung dengan menjumlahkan semua nilai dan membaginya dengan jumlah total nilai, sedangkan median merupakan nilai tengah dari dataset yang diurutkan. Modus adalah nilai yang paling sering muncul dalam dataset kita. Sebagai contoh, dalam tabel yang disediakan, kita dapat melihat bahwa atribut 'battery_power' memiliki modus 1998, yang menunjukkan bahwa kapasitas baterai sebesar 1998 mAh adalah yang paling umum di antara ponsel-ponsel dalam dataset kita.

Selanjutnya, ada ukuran variabilitas seperti rentang (range), kuartil, interquartile range (IQR), variansi, dan standar deviasi. Rentang memberikan gambaran kasar tentang sebaran data dengan mengurangkan nilai terkecil dari nilai terbesar. Kuartil membagi dataset menjadi empat bagian yang sama, dan IQR (diferensial antara kuartil ketiga dan pertama) menunjukkan seberapa tersebar data di antara nilai tengah 50% dari data kita. Variansi dan standar deviasi memberikan pengukuran lebih lanjut tentang seberapa jauh nilai-nilai dalam set data menyimpang dari rata-rata. Dalam konteks dataset kita, jika kita melihat pada 'mobile_wt', standar deviasi yang tinggi akan mengindikasikan berat ponsel yang bervariasi secara signifikan dari rata-rata.

Lebih lanjut, ada ukuran bentuk distribusi yaitu skewness dan kurtosis. Skewness adalah ukuran asimetri distribusi sekitar mean-nya. Sebuah distribusi dengan ekor panjang di sisi kiri akan memiliki skewness negatif, sementara ekor panjang di sisi kanan menunjukkan skewness positif. Kurtosis menggambarkan tingkat puncak atau datarnya distribusi relatif terhadap distribusi normal. Distribusi dengan kurtosis tinggi memiliki puncak yang lebih tajam dan ekor yang lebih berat, sedangkan distribusi dengan kurtosis rendah akan lebih datar dengan ekor yang lebih ringan. Misalnya, pada 'ram' pada dataset kita menunjukkan kurtosis yang rendah (-1.186141e+00), hal ini menandakan adanya kecenderungan pada nilai RAM tidak terlalu jauh dari nilai rata-ratanya.

Terkait dengan data kategorikal atau biner, kita fokus pada frekuensi dan modus. Frekuensi adalah jumlah kemunculan setiap kategori, dan modus adalah kategori yang paling sering muncul. Dalam file ipynb yang disebutkan, kita dapat melihat, misalnya, bahwa mayoritas ponsel memiliki dukungan 'four_g', dengan modus menunjukkan nilai 1 (yaitu, keberadaan fitur 4G).

Setiap elemen statistik ini menyumbang ke pemahaman yang lebih komprehensif tentang dataset dan membantu dalam pengambilan keputusan berbasis data. Pada dasarnya, dengan memadukan semua ukuran ini, kita dapat membentuk gambaran yang jelas tentang data yang kita analisis, yang terdokumentasi secara rinci dalam tabel pada file ipynb yang disertakan. Ini bukan hanya memudahkan kita untuk memahami tren dan pola dalam data, tetapi juga untuk menyajikan temuan ini kepada pemangku kepentingan dengan cara yang dapat diakses dan mudah dipahami.

2.2. Duplikasi Nilai

Dalam langkah penting dari proses pembersihan data, kami melakukan pemeriksaan terhadap nilai-nilai duplikat dalam dataset. Duplikasi data bisa terjadi karena berbagai alasan, seperti kesalahan saat entri data, penggabungan dataset dari berbagai sumber, atau kesalahan dalam proses ekstraksi data. Duplikat yang tidak terdeteksi dapat menyebabkan bias dalam analisis statistik dan mengganggu validitas model prediksi yang kita kembangkan, karena dapat memberikan bobot tambahan pada data tertentu yang tidak seharusnya terjadi.

Setelah menjalankan kode yang dirancang khusus untuk mengidentifikasi dan menghitung baris duplikat, kami menemukan bahwa tidak ada baris dalam dataset kami yang

sepenuhnya identik, yang menegaskan bahwa setiap entri dalam dataset merupakan unik. Temuan ini menunjukkan bahwa dataset telah terlindungi dari salah satu masalah umum yang sering terjadi dalam tahapan pengumpulan dan pengolahan data, yaitu redundansi data. Dengan demikian, keaslian dari setiap rekaman data terjamin, yang meningkatkan integritas dataset dan membuatnya lebih handal untuk dijadikan bahan analisis mendalam. Untuk pemahaman yang lebih mendalam mengenai mekanisme dan logika di balik kode yang telah digunakan untuk mencapai kesimpulan ini, rincian yang lebih ekstensif dan penjelasan langkah demi langkah tersedia dalam file ipynb yang menyertai.

2.3. Missing Value

Analisis terhadap missing values merupakan langkah penting untuk memastikan kualitas data sebelum melangkah ke proses analisis lebih lanjut. Nilai NA pada atribut data dapat diinterpretasikan sebagai missing value atau nilai yang sah tergantung pada konteks atribut tersebut. Walaupun begitu, perlu dibedakan missing value dan wrong value. Berikut adalah peninjauan masing masing atribut numerik:

a. **battery_power**: Jika nilai 0 muncul, ini tidak masuk akal dalam konteks praktis karena baterai tidak mungkin memiliki kapasitas 0 mAh. Ini bisa jadi wrong value.
b. **clock_speed**: Nilai 0 bisa saja merupakan wrong value karena mikroprosesor tidak mungkin memiliki kecepatan jam 0 GHz.
c. **fc (Front Camera)**: Nilai 0 bisa menunjukkan bahwa ponsel tersebut tidak memiliki kamera depan, sehingga bisa dianggap sebagai nilai yang sah.
d. **int_memory**: Nilai 0 bisa saja wrong value, karena tidak mungkin ada ponsel tanpa memori internal sama sekali.
e. **m_dep (Mobile Depth)**: Jika nilai 0 muncul, ini mungkin wrong value karena semua ponsel akan memiliki ketebalan yang lebih besar dari 0 cm.
f. **mobile_wt**: Nilai 0 tidak masuk akal karena semua ponsel memiliki berat, jadi ini bisa jadi wrong value.
g. **n_cores**: Nilai 0 di sini tidak masuk akal karena prosesor harus memiliki setidaknya satu core. Ini bisa dianggap sebagai wrong value.

h. **pc (Primary Camera)**: Nilai 0 bisa menunjukkan bahwa ponsel tersebut tidak memiliki kamera utama, yang mungkin jarang tapi mungkin ada pada model-model tertentu, atau bisa juga dianggap sebagai wrong value.
i. **px_height**: Tinggi resolusi piksel tidak mungkin 0, karena ini berarti layar tidak akan menampilkan apapun. Nilai ini bisa dianggap wrong value.
j. **px_width**: Sama seperti px_height, nilai 0 tidak mungkin dan mungkin merupakan wrong value.
k. **ram**: Nilai 0 untuk RAM tidak masuk akal karena ponsel tidak akan bisa beroperasi tanpa RAM. Ini kemungkinan merupakan wrong value.
l. **sc_h (Screen Height)**: Nilai 0 tidak masuk akal karena layar tidak bisa tidak memiliki tinggi, sehingga ini kemungkinan merupakan wrong value.
m. **sc_w (Screen Width)**: Sama seperti sc_h, nilai 0 di sini juga tidak masuk akal dan bisa jadi merupakan wrong value.
n. **talk_time**: Waktu bicara 0 jam tidak masuk akal karena ini berarti ponsel tidak dapat digunakan untuk berbicara setelah diisi penuh. Ini kemungkinan merupakan wrong value.

Secara keseluruhan, dalam set data yang berhubungan dengan spesifikasi teknis perangkat seperti ponsel, nilai 0 bisa saja merupakan wrong value, kecuali dalam kondisi tertentu di mana 0 dapat menunjukkan ketiadaan fitur tertentu (misalnya, fc dan pc jika ponsel benar-benar tidak memiliki kamera).

Berbeda dengan wrong value, missing value hanya mencari data yang benar-benar hilang atau NA. Kami telah melakukan analisis missing value pada file ipynb yang dilampirkan. Berdasarkan analisis atribut numerik yang telah dilakukan pada file ipynb, tidak ditemukan adanya missing value.

Analisis missing value penting karena nilai yang hilang dalam dataset bisa menyebabkan interpretasi yang salah saat melakukan analisis data dan bisa mengganggu keakuratan model prediktif yang mungkin kami kembangkan menggunakan dataset ini. Mengetahui lokasi tepat dari nilai yang hilang memungkinkan kami untuk mengambil keputusan informasi tentang bagaimana cara terbaik untuk mengatasi masalah ini - apakah itu melalui penghapusan baris data, pengisian nilai yang hilang dengan teknik imputasi, atau mungkin menggunakan metode analisis data yang bisa menangani missing values secara inheren.

Selain melakukan analisis untuk missing values pada atribut numerik, kami juga memeriksa atribut non-numerik dengan mencari nilai 'NA'. Berdasarkan hasil yang didapatkan dari menjalankan kode pencarian *missing value*, kami tidak menemukan indikasi adanya data yang hilang pada atribut non-numerik tersebut. Ini mengindikasikan bahwa, setidaknya untuk variabel non-numerik, dataset telah tampak lengkap tanpa adanya kekosongan data yang harus ditangani lebih lanjut.

2.4. Pendeteksian Outlier

Outlier merupakan pengamatan yang signifikan berbeda dari data lainnya dan dapat menunjukkan variabilitas ekstrem atau kesalahan pengukuran. Untuk mengidentifikasi outlier, kami menerapkan metode IQR yang mengukur penyebaran statistik dan menentukan batas bawah dan atas. Nilai di luar rentang ini dianggap sebagai outlier.

Dari analisis yang dilakukan, outlier terdeteksi pada kolom fc (Resolusi kamera depan). Beberapa nilai yang jauh melampaui rentang umum ini termasuk resolusi kamera depan dengan megapiksel yang sangat tinggi, seperti yang ditunjukkan pada baris ke-53, 98, 413, dan seterusnya. Nilai-nilai ini jauh melebihi kuartil atas dikalikan 1.5, yang mungkin menandakan adanya spesifikasi khusus atau kesalahan penginputan data.

Hasil deteksi outlier ini membawa implikasi penting dalam pengolahan data selanjutnya. Khususnya, analisis yang lebih mendalam diperlukan untuk memutuskan apakah outlier-outlier ini harus dipertahankan, yang mungkin merepresentasikan produk yang unik di pasar, atau dihapus, untuk meminimalkan distorsi pada analisis data selanjutnya.

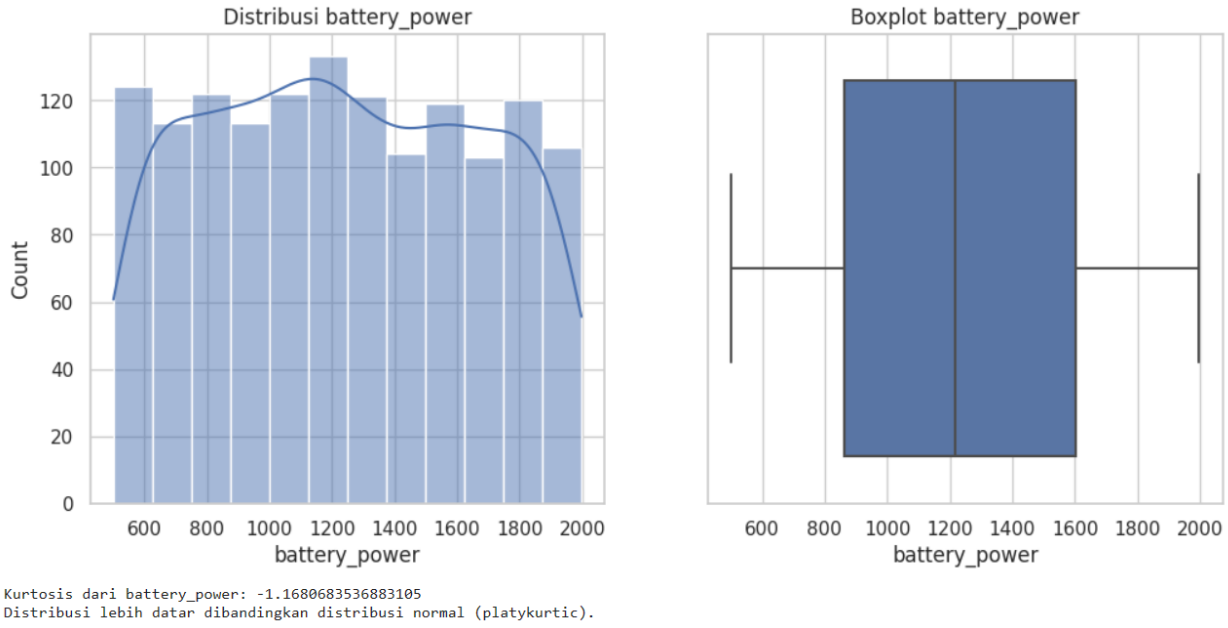
Rincian tabel outlier untuk kolom fc dan lainnya dapat ditemukan pada output fungsi find_outliers dalam file ipynb yang menyertai laporan ini. Tabel tersebut memberikan informasi lengkap mengenai baris data yang terindikasi sebagai outlier.

2.5. Analisis Data pada Kolom Numerik dan Non-Numerik

Dalam analisis data, visualisasi merupakan komponen penting yang memungkinkan kita untuk mengamati distribusi, tren, dan outlier dalam data. Kode yang disajikan di atas merupakan serangkaian instruksi yang dirancang untuk memberikan representasi grafis dari distribusi kolom numerik dan non-numerik dalam sebuah DataFrame. Menggunakan pustaka matplotlib dan seaborn, kode ini secara sistematis menghasilkan histogram dan boxplot untuk kolom-kolom

numerik. Histogram memperlihatkan frekuensi data yang dibagi dalam beberapa bin dan menambahkan estimasi densitas kernel untuk mengilustrasikan distribusi data. Boxplot memberikan ringkasan lima angka dari distribusi data, menampilkan median, kuartil, dan outlier. Analisis kurtosis juga dilakukan untuk setiap kolom numerik, memberikan insight tentang kecerdasan puncak distribusi dibandingkan dengan distribusi normal.

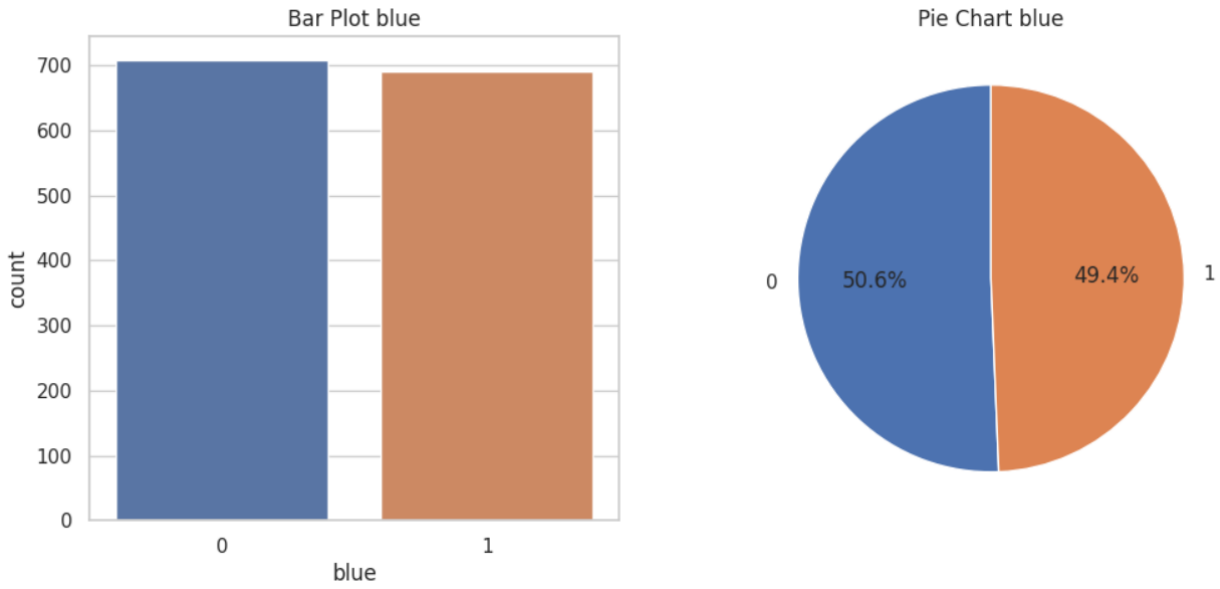
Untuk kolom non-numerik, yang seringkali berisi data kategorikal atau biner, barplot digunakan untuk menampilkan jumlah observasi per kategori, dan piechart menunjukkan proporsi masing-masing kategori sebagai persentase dari total. Ini memberikan pemahaman yang cepat tentang dominasi kategori dalam dataset. Setiap visualisasi dilengkapi dengan judul yang jelas untuk meningkatkan keterbacaan dan interpretasi. Kode ini dirancang untuk dieksekusi dalam lingkungan Jupyter Notebook, di mana penggunaan %matplotlib inline menjamin bahwa plot terintegrasi langsung ke dalam tampilan notebook, membuat proses analisis lebih interaktif dan mudah diakses.



Gambar 2.5.1. Gambar distribusi numerik battery_power

Gambar 2.5.1 merupakan salah satu contoh nilai numerik yang telah dianalisis di file ipynb. Dari data tersebut kita bisa melihat distribusi battery_power dalam bentuk diagram dan boxplot. Selain itu kita juga bisa melihat nilai kurtosis dari battery_power adalah 1.17 yang berarti

Distribusi memiliki ekor yang lebih berat dan puncak yang lebih tajam dibandingkan distribusi normal. Lebih lengkapnya dapat dilihat di file ipynb yang dilampirkan.



Gambar 2.5.2. Gambar distribusi non numerik blue

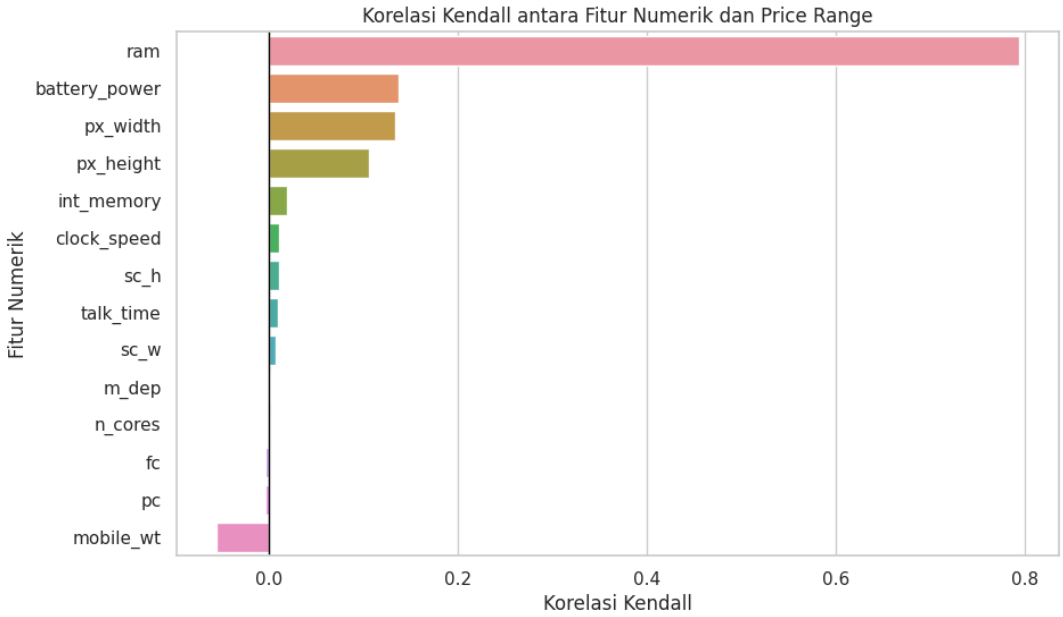
Gambar 2.5.2 merupakan salah satu contoh nilai non numerik yang telah dianalisis di file ipynb. Terdapat dua diagram yang dapat dilihat yaitu bar plot dan pie chart untuk blue. Blue disini merupakan ada tidaknya bluetooth pada ponsel, 0 berarti tidak ada dan 1 berarti ada. Pada pie chart dapat terlihat bahwa ponsel yang tidak memiliki bluetooth ada 50.6% dan yang memiliki fitur bluetooth adalah 49.4%. Lebih lengkapnya bisa dilihat di file ipynb yang dilampirkan.

2.6. Analisis Korelasi

Pemahaman tentang hubungan antar fitur sangat penting dalam analisis data eksploratif, khususnya untuk menilai bagaimana fitur berkorelasi dengan variabel target. Dalam konteks ini, variabel target adalah price_range. Kami menggunakan korelasi Spearman karena tidak mengasumsikan distribusi normal dari data dan dapat menangkap hubungan monoton, baik linear maupun non-linear.

Fungsi correlation menghitung dan memvisualisasikan matriks korelasi dengan metode Spearman antara price_range dan fitur lainnya dalam dataset. Kami menggunakan heatmap dari seaborn untuk mempermudah interpretasi visual dari korelasi tersebut. Nilai korelasi yang dekat

dengan 1 atau -1 menunjukkan hubungan yang kuat, positif atau negatif, sedangkan nilai yang dekat dengan 0 menunjukkan tidak adanya hubungan yang kuat.



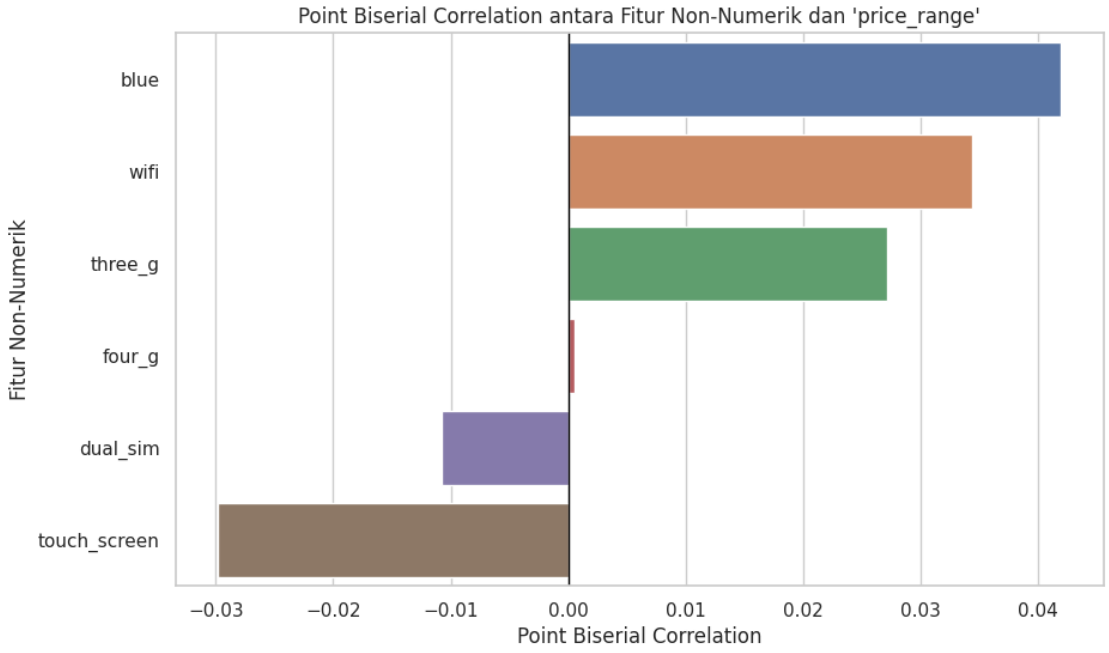
Gambar 2.6.1. Gambar korelasi kolom numerik dengan price range

Gambar 2.6.1 memperlihatkan sebuah diagram korelasi yang mengilustrasikan hubungan antara atribut-atribut numerik dan rentang harga (price range) suatu produk. Dalam diagram tersebut, korelasi antara kapasitas RAM dan rentang harga menonjol secara signifikan, dengan nilai mendekati 0.8. Hal ini mengindikasikan bahwa ada kecenderungan kuat dimana semakin besar kapasitas RAM, maka rentang harga yang dikaitkan juga semakin tinggi.

Di sisi lain, atribut seperti daya baterai (battery_power), lebar piksel (px_width), dan tinggi piksel (px_height) juga menunjukkan hubungan positif terhadap rentang harga, namun dengan intensitas yang lebih rendah dibandingkan dengan RAM. Indikasi ini menunjukkan bahwa walaupun terdapat kenaikan nilai untuk daya baterai, lebar, dan tinggi piksel yang terkait dengan peningkatan rentang harga, pengaruhnya tidak sebesar yang diberikan oleh kapasitas RAM. Terlihat juga adanya korelasi negatif antara berat ponsel (mobile_wt) dan price range, yang menyiratkan bahwa ponsel dengan berat yang lebih tinggi cenderung memiliki rentang harga yang lebih rendah.

Kolom lain seperti internal memory (int_memory), kecepatan prosesor (clock_speed), tinggi layar (sc_h), dan beberapa lainnya, memiliki korelasi yang sangat rendah atau mendekati nol dengan rentang harga. Hal ini menyiratkan bahwa besar atau kecilnya nilai atribut-atribut

tersebut mungkin tidak memiliki pengaruh yang signifikan terhadap rentang harga suatu ponsel. Jika ada pengaruh, kemungkinan itu sangat minimal.



Gambar 2.6.2. Gambar korelasi kolom non numerik dengan price range

Pada Gambar 2.6.2, diagram korelasi disajikan untuk menunjukkan hubungan antara berbagai atribut numerik dan rentang harga suatu produk. Meskipun nilai-nilai korelasi yang diwakili dalam diagram tersebut tidak menunjukkan signifikansi yang kuat, beberapa atribut tetap menunjukkan hubungan positif dengan rentang harga. Secara khusus, fitur seperti Bluetooth (blue), Wi-Fi (wifi), dan dukungan jaringan 3G (three_g) tampak memiliki korelasi positif dengan rentang harga, meskipun tidak begitu kuat. Di sisi lain, korelasi antara dukungan jaringan 4G (four_g) dengan rentang harga hampir tidak ada, dengan nilai korelasi mendekati nol. Sementara itu, fitur dual SIM (dual_sim) dan layar sentuh (touch_screen) terlihat memiliki korelasi negatif, menunjukkan bahwa semakin tinggi kehadiran fitur-fitur ini, mungkin terdapat kecenderungan ringan terhadap penurunan dalam rentang harga.

BAB III
KESIMPULAN

3.1. Kesimpulan

Analisis data eksploratif yang telah kami lakukan mengungkapkan wawasan mendalam tentang karakteristik data ponsel yang terkandung dalam dataset. Statistik dasar menonjolkan variasi yang signifikan pada spesifikasi seperti RAM dan daya baterai, yang berkontribusi besar terhadap penetapan rentang harga. Integritas data diperkuat oleh konsistensi yang terlihat dalam bentuk ketiadaan nilai duplikat dan missing values, yang memberikan dasar yang solid untuk analisis lanjutan.

Dalam proses deteksi outlier, kami mengidentifikasi beberapa nilai ekstrem, khususnya pada kolom resolusi kamera depan. Outlier-outlier ini menuntut analisis lebih lanjut untuk memastikan apakah mereka merupakan indikasi dari spesifikasi yang sangat tinggi atau sekadar kesalahan data, yang penting untuk diatasi guna memastikan keakuratan analisis kita.

Pemeriksaan distribusi data pada kolom numerik memberikan wawasan tentang bentuk distribusi dan kecenderungan yang ada, yang krusial untuk pemilihan teknik statistik atau algoritma machine learning yang tepat. Adapun analisis pada data kategorikal membantu kita mengerti sebaran dan potensi pengaruh fitur-fitur seperti Bluetooth, Wi-Fi, dan dukungan jaringan 3G atau 4G terhadap rentang harga.

Gambar 2.6.1 dan Gambar 2.6.2 dari analisis korelasi menambahkan lapisan pemahaman tambahan, dengan menunjukkan bahwa sementara atribut seperti RAM memiliki korelasi yang sangat tinggi dengan rentang harga, atribut lain seperti kapasitas baterai, dimensi piksel, dan fitur konektivitas seperti Bluetooth, Wi-Fi, dan dukungan jaringan 3G menunjukkan hubungan yang lebih moderat. Fitur seperti dukungan 4G tidak menunjukkan korelasi yang signifikan, sedangkan dual SIM dan layar sentuh tampaknya memiliki korelasi negatif dengan rentang harga.

Kesimpulan yang telah kami tarik dari analisis ini memandu kami untuk lebih memahami faktor-faktor yang mempengaruhi penetapan harga ponsel, dan hal ini akan sangat berharga dalam pengembangan model prediktif harga yang akurat. Langkah berikutnya adalah

memanfaatkan wawasan ini untuk merancang dan menerapkan model pembelajaran mesin yang dapat mengintegrasikan nuansa kompleks dari data yang telah kami analisis.

3.2. Saran

Seiring dengan kemajuan analisis kami, kami telah sampai pada beberapa temuan kunci yang mungkin memerlukan pertimbangan lebih detail. Untuk memaksimalkan efektivitas dan ketepatan dari kesimpulan kami, kami mengajukan sejumlah rekomendasi yang diharapkan dapat mengembangkan pemahaman kita lebih dalam. Berikut adalah daftar saran yang telah kami susun, yang kami percaya akan membawa analisis ini ke level yang lebih tinggi dalam hal keakuratan dan relevansi.

- Mengingat pentingnya data yang bersih untuk analisis, disarankan agar data masa depan menjalani proses pembersihan dan normalisasi yang ketat untuk meminimalisir dampak outlier. Proses ini bisa melibatkan teknik pencarian dan imputasi nilai yang lebih canggih untuk menangani nilai yang ekstrem atau tidak konsisten.
- Disarankan untuk melakukan penelitian lebih lanjut dengan dataset yang lebih besar dan beragam untuk mengkonfirmasi temuan ini. Penelitian tersebut dapat meliputi data dari berbagai sumber dan segmen pasar untuk meningkatkan keumuman model yang dikembangkan.
- Pengaruh fitur kategorikal seperti Wi-Fi, Bluetooth, dan dukungan jaringan 3G/4G terhadap harga perlu dianalisis lebih mendalam. Hal ini dapat melibatkan penggunaan teknik pengkodean kategorikal yang lebih maju seperti 'one-hot encoding' atau 'mean encoding' dalam proses pembelajaran mesin.
- Untuk menambah kedalaman analisis, saran lain adalah mengintegrasikan dataset ini dengan data pasar eksternal, seperti tren konsumen, data penjualan, dan siklus rilis produk, yang dapat memberikan konteks tambahan pada faktor-faktor yang mempengaruhi harga.
- Seiring perkembangan teknologi, fitur baru terus diintegrasikan ke dalam ponsel. Penting untuk terus mengevaluasi dampak dari fitur-fitur ini terhadap rentang harga dan memperbarui model prediktif sesuai dengan tren terkini.

REFERENSI

[1] Bhandari, P. (2023, June 21). How to find outliers: 4 ways with examples & explanation. Scribbr. https://www.scribbr.com/statistics/outliers/

[2] Radhi, M., Amalia, A., Sitompul, D. R. H., Sinurat, S. H., & Indra, E. . (2022). ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS (EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK. Jurnal Sistem Informasi Dan Ilmu Komputer Prima(JUSIKOM PRIMA), 4(2), 23 -27. https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475