**Mata Kuliah Analisis Data Eksploratif**

**Laporan Ujian Akhir Semester (UAS) Take-Home: Melakukan Analisis Data Pada Proses Visualisasi Data Kategorikal dan Numerikal**

**Dosen Pengampu: Wahyu Sri Utami, S.Si., M.Sc**



**Disusun oleh:**

**Lathif Ramadhan (5231811022)**

**PROGRAM STUDI SAINS DATA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS TEKNOLOGI YOGYAKARTA**

**YOGYAKARTA**

**2024**

# Pendahuluan

Data adalah salah satu aset penting dalam era digital yang semakin berkembang pesat saat ini. Dalam konteks analisis data eksploratif, proses pemahaman pola, hubungan, dan tren dari sebuah dataset menjadi langkah awal yang esensial sebelum melangkah ke tahap analisis lanjutan. Mata kuliah Analisis Data Eksploratif memberikan fondasi yang kuat bagi mahasiswa untuk menggali wawasan dari data, memahami struktur data, dan menghasilkan visualisasi yang dapat menyampaikan informasi secara efektif.

Analisis Data Eksploratif (EDA) bertujuan untuk memberikan gambaran menyeluruh dari suatu dataset melalui pengolahan data dan visualisasi yang informatif. Dengan teknik EDA, data dapat dikaji dari berbagai dimensi, baik kategorikal maupun numerikal. Visualisasi seperti bar chart, scatter plot, heatmap, dan box plot memegang peranan penting dalam membantu peneliti memahami distribusi data, mengidentifikasi anomali, serta menjawab pertanyaan yang relevan berdasarkan data yang tersedia.

Dalam tugas Ujian Akhir Semester ini, fokus utama adalah menganalisis dataset Diamonds yang terdiri dari berbagai variabel seperti berat karat, kualitas potongan, warna, kejernihan, harga, dan dimensi berlian. Dataset ini menawarkan peluang untuk mengeksplorasi karakteristik berlian melalui berbagai visualisasi data kategorikal dan numerikal. Dengan ukuran data yang mencakup lebih dari 500 entri, analisis ini memberikan tantangan dan peluang untuk menerapkan metode EDA secara komprehensif.

Persiapan data atau data preparation adalah salah satu langkah krusial dalam analisis ini. Tahapan ini mencakup penanganan missing value, outlier, dan data duplikat, yang bertujuan untuk memastikan keakuratan hasil analisis. Proses ini tidak hanya meningkatkan kualitas data, tetapi juga meminimalkan risiko kesalahan interpretasi yang dapat mempengaruhi hasil akhir visualisasi.

Visualisasi data pada tugas ini tidak hanya berfungsi sebagai alat bantu untuk memahami pola dan distribusi data, tetapi juga menjadi sarana komunikasi yang efektif untuk menyampaikan hasil analisis kepada audiens yang beragam. Kombinasi visualisasi data kategorikal dan numerikal akan memberikan perspektif yang mendalam tentang hubungan antara variabel dalam dataset Diamonds. Proses analisis yang mendalam juga akan dilengkapi dengan interpretasi yang akurat, sehingga menghasilkan laporan yang relevan dan informatif.

Melalui tugas ini, mahasiswa diharapkan mampu mengintegrasikan teori dan praktik dalam analisis data eksploratif. Dengan pendekatan yang sistematis dan metodologis, analisis ini diharapkan tidak hanya memenuhi kebutuhan akademik, tetapi juga memberikan wawasan yang dapat diterapkan dalam konteks dunia nyata, terutama dalam memahami industri berlian yang kompleks.

# Daftar Isi

[Pendahuluan II](#_Toc187019200)

[Daftar Isi III](#_Toc187019201)

[Bab 1: Mengenal Dataset Diamonds 1](#_Toc187019202)

[1. Mengimpor Library 1](#_Toc187019203)

[2. Mengunggah file dari lokal ke Colab dan membaca dataset 2](#_Toc187019204)

[3. Lihat informasi dan ringkasan data (Statistika Deskriptif) 3](#_Toc187019205)

[4. Mengecek Missing Value 5](#_Toc187019206)

[5. Mengecek Outlier 6](#_Toc187019207)

[6. Mengecek Duplicate Value 6](#_Toc187019208)

[Bab 2: Proses dan Teknik pada Data Preparation (Data Cleaning) 7](#_Toc187019209)

[1. Menghapus kolom yang tidak relevan (Feature Selection) 7](#_Toc187019210)

[2. Menangani Missing Values 8](#_Toc187019211)

[a. Mengisi nilai NaN untuk kolom numerik dengan median 8](#_Toc187019212)

[b. Mengisi nilai NaN untuk kolom kategorikal dengan mode 8](#_Toc187019213)

[3. Menangani Outliers 8](#_Toc187019214)

[4. Menangani Duplicates Value 9](#_Toc187019215)

[5. Memperbaiki data yang tidak valid 9](#_Toc187019216)

[Bab 3: Visualisasi Data Numerikal dan Kategorikal 10](#_Toc187019217)

[1. Bar Chart: Harga (Price) Rata-Rata berdasarkan Warna (Color) 10](#_Toc187019218)

[2. Line Graph: Harga (Price) Rata-Rata Berdasarkan Kualitas Potongan (Cut) 12](#_Toc187019219)

[3. Scatter Plot: Hubungan antara Berat Berlian (carat) dan Harga (price) 15](#_Toc187019220)

[4. Heatmap: Korelasi Antara Kolom Carat, Depth, and Table 17](#_Toc187019221)

[5. Boxplot: Kolom Carat, Depth, dan Table 19](#_Toc187019222)

[Kesimpulan 23](#_Toc187019223)

# Bab 1: Mengenal Dataset Diamonds

Dalam analisis data, pemahaman mendalam mengenai dataset yang digunakan merupakan langkah awal yang sangat penting. Dataset *Diamonds*, yang menjadi objek utama dalam laporan ini, diambil dari sumber terpercaya pada platform Kaggle melalui tautan berikut: <https://www.kaggle.com/datasets/shivam2503/diamonds/data>. Dataset ini berisi informasi rinci tentang berlian, mencakup berbagai atribut seperti berat (carat), kualitas potongan (cut), warna (color), kejernihan (clarity), dimensi (x, y, z), dan harga (price). Dengan total 10 kolom data, dataset ini menawarkan peluang analisis yang menarik dan kompleks, baik dari aspek data numerikal maupun kategorikal.

Pengerjaan analisis ini dilakukan menggunakan Python melalui platform Google Colab. Google Colab dipilih karena kemudahan aksesibilitasnya serta kemampuannya dalam menjalankan kode Python secara efisien, khususnya untuk pengolahan data skala besar. Proses analisis dimulai dari mengimpor library yang relevan untuk manipulasi data, visualisasi, dan analisis statistik. Tahapan awal ini bertujuan untuk mengeksplorasi dataset secara deskriptif guna menggali informasi mendasar tentang struktur dan karakteristik data.

Langkah-langkah utama yang dilakukan meliputi:

1. **Mengimpor library** seperti *pandas*, *numpy*, dan *matplotlib* untuk mendukung pengolahan dan visualisasi data.
2. **Eksplorasi statistik deskriptif**, mencakup mean, median dst.
3. **Penggalian informasi dataset**, termasuk mengevaluasi missing values, outlier, dan duplicate value untuk setiap kolom.

Dengan pendekatan yang sistematis, tahap awal ini diharapkan memberikan wawasan yang jelas mengenai kondisi awal dataset, yang akan menjadi dasar untuk analisis lebih lanjut pada bagian berikutnya

Berikut langkah-langkahnya:

## Mengimpor Library

A screenshot of a computer

Description automatically generated

1. **pandas as pd:**
   * **Fungsi:** Menyediakan struktur data yang efisien dan fleksibel bernama DataFrame untuk memanipulasi dan menganalisis data tabular (data dalam bentuk tabel).
   * **Kegunaan dalam proyek:** Mengimpor, membersihkan, memanipulasi, dan menganalisis data berlian yang disimpan dalam file CSV.
2. **numpy as np:**
   * **Fungsi:** Menyediakan dukungan untuk operasi numerik dan array multidimensi.
   * **Kegunaan dalam proyek:** Melakukan perhitungan matematis, manipulasi array, dan operasi numerik lainnya yang dibutuhkan dalam analisis data, seperti menghitung statistik deskriptif atau menangani outlier.
3. **seaborn as sns:**
   * **Fungsi:** Library visualisasi data berbasis Matplotlib yang menyediakan antarmuka tingkat tinggi untuk membuat grafik statistik yang informatif dan menarik.
   * **Kegunaan dalam proyek:** Membuat berbagai jenis visualisasi data, seperti boxplot, heatmap, scatter plot, dan bar chart, untuk mengeksplorasi dan memahami pola dalam dataset berlian.
4. **matplotlib.pyplot as plt:**
   * **Fungsi:** Library dasar untuk plotting dan visualisasi data di Python.
   * **Kegunaan dalam proyek:** Menyesuaikan plot yang dibuat dengan Seaborn, seperti menambahkan judul, label sumbu, legenda, dan mengubah gaya plot. Juga dapat digunakan untuk membuat plot dasar jika diperlukan.

**Intinya:**

* **Pandas** dan **NumPy** adalah fondasi untuk manipulasi dan analisis data.
* **Seaborn** dan **Matplotlib** digunakan untuk visualisasi data, di mana Seaborn menyediakan antarmuka yang lebih mudah digunakan dan estetis, sementara Matplotlib memungkinkan kustomisasi yang lebih detail.

## Mengunggah file dari lokal ke Colab dan membaca dataset

A screen shot of a computer program

Description automatically generated

## Lihat informasi dan ringkasan data (Statistika Deskriptif)

A screenshot of a computer

Description automatically generated

**Penjelasan Ouput:**

1. **Jumlah Data**: Dataset memiliki **53.940 baris** (entri) dan **11 kolom**.

* Dataset memiliki **53.940 baris**, diberi penomoran dari 0 hingga 53.939.
* Data Columns Dataset terdiri dari **11 kolom**, di mana: **8 kolom bertipe float64 (angka desimal). 3 kolom bertipe object (string/kategorikal)**.

1. **Non-Null Count** : Setiap kolom memiliki jumlah non-null (data tidak kosong) yang berbeda-beda, menunjukkan adanya missing values di beberapa kolom:

* Kolom **Unnamed**: 0 memiliki 48.517 data tidak kosong, artinya 5.426 data kosong (9,98%).
* Kolom **carat** memiliki 48.545 data tidak kosong, artinya 5.398 data kosong (9,94%).
* Kolom **cut** memiliki 48.546 data tidak kosong, artinya 5.397 data kosong (9,94%).
* Kolom **color** memiliki 48.624 data tidak kosong, artinya 5.319 data kosong (9,85%).
* Kolom **clarity** memiliki 48.336 data tidak kosong, artinya 5.607 data kosong (10,39%).
* Kolom **depth** memiliki 48.518 data tidak kosong, artinya 5.425 data kosong (9,98%).
* Kolom **table** memiliki 48.640 data tidak kosong, artinya 5.303 data kosong (9,83%).
* Kolom **price** memiliki 48.564 data tidak kosong, artinya 5.377 data kosong (9,93%).
* Kolom **x** memiliki 48.610 data tidak kosong, artinya 5.333 data kosong (9,88%).
* Kolom **y** memiliki 48.542 data tidak kosong, artinya 5.401 data kosong (9,94%).
* Kolom **z** memiliki 48.558 data tidak kosong, artinya 5.384 data kosong (9,94%).

1. **Dtype** (Tipe Data)

* **float64** (angka desimal/numerik):
  + Kolom: Unnamed: 0, carat, depth, table, price, x, y, z.
  + Biasanya digunakan untuk representasi data numerik seperti ukuran dan harga.
* **object** (kategorikal/string):
  + Kolom: cut, color, clarity(tingkat kejernihan berlian).
  + Berisi data non-numerik seperti kategori atau deskripsi.

A screenshot of a computer

Description automatically generated

**Kolom-kolom dataset**:

* **Unnamed**: 0: Kolom indeks numerik, berasal dari file CSV, ini sebenarnya kolom tanpa nama(tanpa string). Sebagian datanya hilang (missing values).
* **carat**: Berat berlian dalam karat (0.2--5.01). Kolom ini cukup lengkap tetapi mungkin memiliki missing values.
* **cut**: Kualitas potongan berlian (Fair, Good, Very Good, Premium, Ideal), beberapa datanya hilang (missing values).
* **color**: Warna berlian (e.g., E, F, G) from J (worst) to D (best), terdapat missing values.
* **clarity**: Kejernihan berlian (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)), beberapa nilai hilang.
* **depth**: Persentase kedalaman berlianpercentage = z / mean(x, y) = 2 \* z / (x + y) (43--79), menunjukkan banyak nilai kosong.
* **table**: Lebar bagian atas berlian (43--95), beberapa nilai kosong.
* **price**: Harga berlian dalam dolar AS ($326--$18,823), sebagian besar nilainya lengkap.
* **x**, **y**, **z**: Dimensi berlian dalam milimeter (panjang, lebar, tinggi/kedalaman), beberapa memiliki nilai kosong atau tidak valid.
* **x**: length in mm (0--10.74)
* **y**: width in mm (0--58.9)
* **z**: depth in mm (0--31.8)

Tabel Statistika Deskriptif untuk kolom numerik:

A screenshot of a computer

Description automatically generated

## Mengecek Missing Value

A computer screen shot of a program code

Description automatically generated

A black and white table with numbers and letters

Description automatically generated

## Mengecek Outlier

A screen shot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A black rectangle with green text

Description automatically generated

## Mengecek Duplicate Value

A screenshot of a computer

Description automatically generated

# Bab 2: Proses dan Teknik pada Data Preparation (Data Cleaning)

Proses data preparation atau data cleaning merupakan tahap krusial dalam analisis data yang bertujuan untuk memastikan kualitas data sebelum dilakukan analisis lebih lanjut. Data yang bersih dan terstruktur dengan baik menjadi fondasi utama dalam menghasilkan visualisasi yang akurat dan interpretasi yang bermakna. Pada tahap ini, berbagai langkah dilakukan untuk mengidentifikasi dan memperbaiki masalah yang terdapat dalam dataset, seperti nilai yang hilang (*missing values*), data yang tidak relevan, nilai ekstrem (*outliers*), duplikasi data, serta data yang tidak valid.

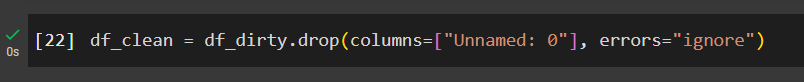
Dalam laporan ini, proses data cleaning meliputi beberapa langkah utama sebagai berikut:

1. **Feature Selection**: Menghapus kolom yang tidak relevan atau tidak memiliki kontribusi signifikan terhadap analisis, seperti kolom *Unnamed: 0* yang hanya berisi indeks numerik tanpa informasi penting.
2. **Penanganan Missing Values**: Mengidentifikasi nilai yang hilang pada setiap kolom dan melakukan penanganan, baik melalui penghapusan data yang tidak lengkap maupun imputasi dengan metode yang sesuai.
3. **Penanganan Outliers**: Mendeteksi dan menangani nilai-nilai ekstrem yang dapat memengaruhi hasil analisis secara signifikan, menggunakan teknik seperti analisis box plot atau perhitungan IQR (*Interquartile Range*).
4. **Duplicate Values**: Menghapus data duplikat yang dapat menyebabkan bias dalam analisis.
5. **Perbaikan Data Tidak Valid**: Mengoreksi nilai yang tidak logis atau berada di luar rentang yang diharapkan, misalnya yang terdapat pada kolom cut, color, dan clarity.

Setiap langkah data cleaning dilakukan secara hati-hati dan sistematis untuk memastikan integritas dataset tetap terjaga. Dengan demikian, hasil analisis pada tahap berikutnya akan lebih representatif dan sesuai dengan tujuan eksplorasi data.

Berikut langkah-langkah pada tahap data preparation atau data cleaning pada dataset Diamonds:

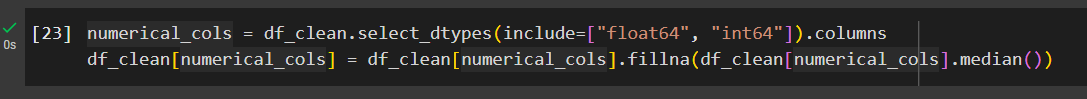
## Menghapus kolom yang tidak relevan (Feature Selection)



Bagian ini menghapus kolom yang tidak relevan dengan analisis, yaitu kolom **Unnamed: 0**. Kolom ini sering kali muncul sebagai index dari dataset yang disertakan saat mengimpor data. **errors="ignore"** memastikan bahwa tidak ada error jika kolom tersebut tidak ada.

## Menangani Missing Values

### a. Mengisi nilai NaN untuk kolom numerik dengan median



Kode ini mencari kolom dengan tipe data numerik (***float64***, ***int64***) dan mengisi nilai yang hilang (***NaN***) dengan median dari masing-masing kolom. Alasan mengapa menggunakan median karena lebih disukai daripada rata-rata untuk menghindari pengaruh nilai ekstrem (***outlier***).

### b. Mengisi nilai NaN untuk kolom kategorikal dengan mode

A screenshot of a computer program

Description automatically generated

Untuk kolom dengan tipe data kategorikal (***object***), kode ini mengisi nilai yang hilang dengan nilai **modus** (nilai yang paling sering muncul) dari masing-masing kolom. Ini membantu untuk mempertahankan distribusi data.

## Menangani Outliers

A computer screen shot of a program code

Description automatically generated

Pada bagian ini menangani *outliers* di seluruh kolom numerik pada dataset menggunakan metode *Interquartile Range (****IQR****)*. Outliers didefinisikan sebagai nilai yang lebih besar dari *Q3 + 3 \* IQR*. Jika nilai price lebih besar dari batas atas ini, maka nilai tersebut akan diganti dengan nilai batas atas.

## Menangani Duplicates Value

A screen shot of a computer

Description automatically generated

Kode ini bertujuan untuk **mengidentifikasi dan menghapus baris-baris duplikat** dalam DataFrame Pandas (*df*) dan menyimpan hasilnya ke DataFrame baru (*df\_no\_duplicates*). Dapat dikatakan bahwa kode ini membersihkan data dengan menghapus baris-baris yang sama persis, sehingga menghasilkan dataset yang lebih ringkas dan akurat.

## Memperbaiki data yang tidak valid

A screen shot of a computer code

Description automatically generated

Kode ini bertujuan untuk **memastikan kualitas dan konsistensi data** pada kolom-kolom kategorikal 'cut', 'color', dan 'clarity' dalam dataset berlian. Proses ini disebut **validasi data**, di mana setiap nilai pada kolom tersebut diperiksa apakah sesuai dengan daftar nilai yang dianggap valid atau sah.

# ****Bab 3: Visualisasi Data Numerikal dan Kategorikal****

## Bar Chart: Harga (Price) Rata-Rata berdasarkan Warna (Color)

A screen shot of a computer code

Description automatically generated

**Breakdown Kode:**

Kode Python di atas menghasilkan grafik batang dengan menggunakan library Seaborn. Berikut penjelasan singkatnya:

* **sns.barplot:** Fungsi ini digunakan untuk membuat grafik batang.
* **x='color', y='price':** Menentukan sumbu x sebagai variabel 'color' (warna) dan sumbu y sebagai variabel 'price' (harga).
* **data=df\_clean:** Menggunakan DataFrame df\_clean sebagai sumber data.
* **order=['D', 'E', 'F', 'G', 'H', 'I', 'J', 'Unknown']:** Mengatur urutan kategori warna pada sumbu x sesuai dengan kualitasnya.
* **plt.title():** Memberikan judul pada grafik.
* **plt.grid(axis='y'):** Menambahkan garis grid pada sumbu y untuk memudahkan pembacaan.
* **plt.text():** Menambahkan teks keterangan di bawah sumbu x untuk menjelaskan skala warna.

**Output Kode:**

A graph of blue bars

Description automatically generated

**Interpretasi Grafik:**

Grafik batang di atas memberikan gambaran visual yang jelas tentang hubungan antara warna berlian dan harga rata-ratanya. Setiap batang pada grafik mewakili satu kategori warna berlian, mulai dari D (warna terburuk) hingga J (warna terbaik), serta kategori "Unknown" untuk berlian dengan warna yang tidak teridentifikasi. Tinggi batang menunjukkan harga rata-rata berlian dengan warna tertentu.

**Temuan Utama:**

* **Warna Mempengaruhi Harga**

Grafik menunjukkan bahwa warna berlian memiliki pengaruh yang sangat signifikan terhadap harga. Berlian dengan warna yang lebih baik (mendekati warna putih murni) cenderung memiliki harga yang lebih tinggi.

* **Tren Penurunan Harga**

Secara umum, terdapat tren peningkatan harga rata-rata dari warna D ke warna J. Hal ini menunjukkan bahwa semakin sedikit warna yang terlihat pada berlian, semakin tinggi nilai jualnya.

* **Variasi Harga**

Terdapat variasi harga yang cukup signifikan antara berbagai kategori warna, terutama antara warna D (terburuk) dan warna J (terbaik).

* **Pengaruh "Unknown":**

Kategori "Unknown" memiliki harga rata-rata yang lebih rendah, yang mungkin disebabkan oleh ketidakpastian mengenai kualitas warna berlian tersebut.

**Implikasi:**

* **Keputusan Pembelian:**

Bagi konsumen, grafik ini menunjukkan bahwa warna adalah salah satu faktor penting yang perlu dipertimbangkan saat membeli berlian. Berlian dengan warna I atau J umumnya dianggap sebagai investasi yang lebih baik karena memiliki potensi nilai jual yang lebih tinggi.

* **Penetapan Harga**

Bagi penjual berlian, grafik ini dapat menjadi referensi dalam menentukan harga jual berlian berdasarkan warna.

* **Analisis Lebih Lanjut**

Grafik ini hanya menunjukkan hubungan antara warna dan harga rata-rata. Untuk analisis yang lebih mendalam, perlu dipertimbangkan faktor-faktor lain seperti kualitas potongan (cut), kejernihan (clarity), dan berat (carat).

Grafik batang ini memberikan visualisasi yang jelas tentang bagaimana warna berlian mempengaruhi harga. Informasi ini sangat berguna bagi konsumen dan pelaku industri perhiasan dalam membuat keputusan yang lebih baik.

## Line Graph: Harga (Price) Rata-Rata Berdasarkan Kualitas Potongan (Cut)

A screen shot of a computer code

Description automatically generated

**Breakdown Kode:**

1. **mean\_price\_by\_cut = df\_clean.groupby('cut')['price'].mean():**
   * **df\_clean.groupby('cut'):** Mengelompokkan data dalam DataFrame df\_clean berdasarkan kolom 'cut'. Ini akan membagi data menjadi beberapa kelompok berdasarkan kualitas potongan.
   * **['price'].mean():** Menghitung rata-rata harga (price) untuk setiap kelompok yang telah dibuat. Hasilnya akan menjadi sebuah Series dengan indeks berupa nilai-nilai unik pada kolom 'cut' dan nilai-nilainya adalah rata-rata harga untuk masing-masing kategori 'cut'.
2. **plt.figure(figsize=(8, 6)):**
   * Membuat sebuah figure (canvas) untuk plot dengan ukuran 8 inci x 6 inci.
3. **plt.plot(mean\_price\_by\_cut.index, mean\_price\_by\_cut.values, marker='o'):**
   * **mean\_price\_by\_cut.index:** Mengambil indeks dari Series mean\_price\_by\_cut, yang merupakan nilai-nilai unik pada kolom 'cut'. Ini akan menjadi nilai pada sumbu x.
   * **mean\_price\_by\_cut.values:** Mengambil nilai-nilai dari Series mean\_price\_by\_cut, yang merupakan rata-rata harga. Ini akan menjadi nilai pada sumbu y.
   * **marker='o':** Menambahkan marker berupa lingkaran pada setiap titik data untuk memperjelas posisi titik data.
4. **plt.title('Line Graph: Harga(Price) rata-rata berdasarkan Cut'):** Memberikan judul pada grafik.
5. **plt.xlabel('Cut'):** Memberikan label pada sumbu x.
6. **plt.ylabel('Average Price'):** Memberikan label pada sumbu y.
7. **plt.grid(True):** Menampilkan grid pada grafik untuk memudahkan pembacaan.
8. **plt.show():** Menampilkan grafik.

**Intinya, kode di atas:**

* Menghitung rata-rata harga berlian untuk setiap kategori kualitas potongan.
* Membuat grafik garis yang menunjukkan hubungan antara kualitas potongan dan harga rata-rata.
* Mengatur tampilan grafik agar lebih mudah dipahami.

**Output Kode:**

A graph with blue lines

Description automatically generated

**Interpretasi Grafik:**

Grafik garis di atas menggambarkan hubungan antara kualitas potongan (cut) dari sebuah berlian dengan harga rata-ratanya. Setiap titik pada grafik mewakili harga rata-rata dari berlian dengan kualitas potongan tertentu, dan garis yang menghubungkan titik-titik tersebut menunjukkan tren umum.

**Temuan Utama:**

* **Kualitas Potongan Mempengaruhi Harga**

Grafik menunjukkan bahwa kualitas potongan memiliki pengaruh yang signifikan terhadap harga berlian. Berlian dengan kualitas potongan yang berbeda memiliki harga rata-rata yang berbeda pula.

* **Premium Memiliki Harga Tertinggi**

Berlian dengan kualitas potongan "Premium" memiliki harga rata-rata yang paling tinggi di antara semua kategori potongan lainnya. Hal ini mengindikasikan bahwa konsumen cenderung membayar harga yang lebih premium untuk berlian dengan kualitas potongan yang sangat baik.

* **Fluktuasi Harga yang Signifikan**

Terdapat fluktuasi harga yang cukup signifikan antara berbagai kategori potongan. Misalnya, terdapat perbedaan yang cukup besar antara harga rata-rata berlian dengan kualitas "Premium" dan "Ideal".

* **Kualitas "Ideal" dan "Good" Mempunyai Harga yang Relatif Rendah**

Berlian dengan kualitas potongan "Ideal" dan "Good" memiliki harga rata-rata yang lebih rendah dibandingkan dengan kategori lainnya. Hal ini menunjukkan bahwa kualitas potongan yang lebih rendah cenderung memiliki nilai jual yang lebih rendah pula.

**Implikasi:**

* **Keputusan Pembelian**

Grafik ini dapat membantu konsumen dalam membuat keputusan pembelian. Konsumen dapat melihat bahwa kualitas potongan memiliki dampak yang signifikan terhadap harga, sehingga mereka dapat mempertimbangkan untuk mengalokasikan anggaran yang lebih besar untuk mendapatkan berlian dengan kualitas potongan yang lebih baik.

* **Strategi Penetapan Harga**

Bagi penjual berlian, grafik ini dapat menjadi referensi dalam menentukan harga jual berlian berdasarkan kualitas potongannya.

* **Analisis Lebih Lanjut**

Grafik ini hanya menunjukkan hubungan antara kualitas potongan dan harga rata-rata. Untuk mendapatkan pemahaman yang lebih komprehensif, perlu dilakukan analisis lebih lanjut dengan mempertimbangkan faktor-faktor lain seperti warna, kejernihan, dan berat.

Grafik line plot ini memberikan gambaran yang jelas tentang pengaruh kualitas potongan terhadap harga berlian. Hasil analisis ini dapat menjadi dasar untuk pengambilan keputusan yang lebih baik, baik bagi konsumen maupun pelaku industri perhiasan.

## Scatter Plot: Hubungan antara Berat Berlian (carat) dan Harga (price)

A screen shot of a computer program

Description automatically generated

**Breakdown Kode:**

* plt.figure(figsize=(8, 6)): Membuat figure dengan ukuran 8 inci x 6 inci.
* plt.scatter(df\_clean['carat'], df\_clean['price'], alpha=0.5): Membuat scatter plot dengan sumbu x adalah kolom 'carat' dan sumbu y adalah kolom 'price' dari dataframe df\_clean. Parameter alpha=0.5 digunakan untuk mengatur transparansi titik-titik data.
* plt.title('Scatter Plot: Carat vs Price'): Memberikan judul pada grafik.
* plt.xlabel('Carat'): Memberikan label pada sumbu x.
* plt.ylabel('Price'): Memberikan label pada sumbu y.
* plt.grid(True): Menampilkan grid pada grafik.
* plt.show(): Menampilkan grafik.

**Output Kode:**

A graph of blue dots

Description automatically generated

**Interpretasi Grafik:**

Grafik scatter plot di atas menggambarkan hubungan antara berat berlian (dalam karat) dengan harga berlian. Setiap titik pada grafik mewakili satu berlian, di mana posisi titik pada sumbu x menunjukkan beratnya (carat) dan posisi titik pada sumbu y menunjukkan harganya.

**Temuan Utama:**

* **Korelasi Positif**

Secara umum, grafik menunjukkan adanya korelasi positif antara berat berlian (carat) dan harga berlian. Artinya, semakin berat sebuah berlian, cenderung semakin tinggi pula harganya. Ini adalah temuan yang intuitif, karena berlian dengan berat yang lebih besar umumnya dianggap lebih langka dan bernilai.

* **Variasi Harga yang Signifikan**

Meskipun terdapat tren peningkatan harga seiring bertambahnya berat, kita juga dapat melihat bahwa ada variasi harga yang sangat signifikan untuk berlian dengan berat yang sama. Hal ini menunjukkan bahwa berat bukanlah satu-satunya faktor yang mempengaruhi harga berlian. Faktor lain seperti kualitas potongan (cut), warna (color), kejernihan (clarity), dan dimensi fisik (x, y, z) juga berperan sangat besar dalam menentukan harga.

* **Konsentrasi Data**

Sebagian besar data terkonsentrasi pada rentang berat 1 hingga 2 karat, dengan harga berkisar antara 2000 hingga 12000. Ini mengindikasikan bahwa sebagian besar berlian yang dianalisis memiliki ukuran menengah keatas dan berada pada kisaran harga yang umum.

**Implikasi:**

* **Pentingnya Faktor Lain**

Grafik ini menegaskan bahwa berat bukanlah satu-satunya faktor penentu harga berlian. Faktor-faktor kualitas lainnya seperti potongan, warna, dan kejernihan memiliki pengaruh yang sangat besar dan dapat menyebabkan variasi harga yang signifikan, bahkan untuk berlian dengan berat yang sama.

* **Segmentasi Pasar**

Data menunjukkan adanya segmentasi pasar yang jelas berdasarkan berat dan harga. Berlian dengan berat di atas 1 karat keatas umumnya memiliki harga yang jauh lebih tinggi dan mungkin ditujukan untuk segmen pasar yang lebih premium.

## Heatmap: Korelasi Antara Kolom Carat, Depth, and Table

**Memahami Heatmap Korelasi**

Heatmap adalah visualisasi yang sangat berguna untuk melihat hubungan antar variabel dalam bentuk matriks. Setiap sel dalam heatmap mewakili korelasi antara dua variabel. Warna pada setiap sel menunjukkan kekuatan dan arah hubungan tersebut. Pada umumnya, warna merah menunjukkan korelasi positif yang kuat, warna biru menunjukkan korelasi negatif yang kuat, dan warna putih mendekati nol menunjukkan tidak ada korelasi.

A screen shot of a computer program

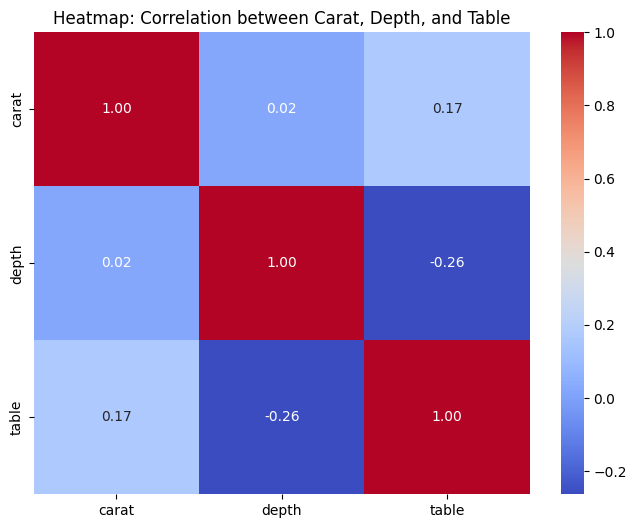
Description automatically generated

**Breakdown Kode:**

Kode Python diatas melakukan beberapa langkah:

1. **Memilih Kolom:** Memilih kolom 'carat', 'depth', dan 'table' dari DataFrame df\_clean untuk dianalisis.
2. **Menghitung Matriks Korelasi:** Menghitung matriks korelasi Pearson antara variabel-variabel yang dipilih.
3. **Membuat Heatmap:** Menggunakan library Seaborn, membuat heatmap untuk visualisasi matriks korelasi.
   * annot=True: Menampilkan nilai korelasi pada setiap sel.
   * cmap='coolwarm': Menggunakan colormap 'coolwarm' untuk menunjukkan korelasi positif (warna merah) dan negatif (warna biru).
   * fmt=".2f": Memformat nilai korelasi hingga 2 angka di belakang koma.

**Output Kode:**



**Interpretasi Grafik:**

Berdasarkan heatmap yang dihasilkan, kita dapat menarik beberapa kesimpulan:

* **Korelasi Positif Antara Carat dan Table**

Terdapat korelasi positif yang lemah antara variabel carat dan table. Ini berarti bahwa secara umum, berlian dengan berat (carat) yang lebih besar cenderung memiliki ukuran meja (table) yang lebih besar pula. Namun, korelasi ini tergolong lemah, sehingga tidak semua berlian dengan carat besar memiliki table yang besar.

* **Korelasi Negatif Lemah Antara Depth dan Table**

Terdapat korelasi negatif yang lemah antara variabel depth dan table. Artinya, berlian dengan depth (kedalaman) yang lebih besar cenderung memiliki ukuran meja yang lebih kecil. Namun, seperti halnya korelasi antara carat dan table, korelasi ini juga tergolong lemah.

* **Tidak Ada Korelasi yang Signifikan Antara Carat dan Depth**

Nilai korelasi antara carat dan depth sangat mendekati nol, yang menunjukkan bahwa tidak ada hubungan linear yang signifikan antara kedua variabel ini. Artinya, berat berlian tidak secara langsung mempengaruhi kedalamannya.

**Implikasi**

* **Hubungan yang Kompleks**

Hubungan antara carat, depth, dan table pada berlian adalah kompleks dan tidak selalu linear. Faktor-faktor lain seperti kualitas potongan (cut) dan kejernihan (clarity) juga dapat mempengaruhi hubungan antara ketiga variabel ini.

* **Pemodelan**

Informasi dari heatmap ini dapat digunakan untuk membangun model prediksi harga berlian. Variabel yang memiliki korelasi yang kuat dengan harga dapat dijadikan sebagai fitur dalam model.

* **Analisis Lebih Lanjut**

Untuk memahami hubungan antara variabel-variabel ini secara lebih mendalam, perlu dilakukan analisis lebih lanjut, seperti analisis regresi atau analisis komponen utama.

Heatmap korelasi adalah alat yang sangat berguna untuk mengeksplorasi hubungan antara variabel-variabel numerik. Dalam konteks data berlian, heatmap ini membantu kita memahami bagaimana carat, depth, dan table saling terkait. Informasi ini dapat digunakan untuk berbagai tujuan, seperti membangun model prediksi harga, melakukan analisis lebih lanjut, atau membuat keputusan bisnis yang lebih baik.

## Boxplot: Kolom Carat, Depth, dan Table

**Pendahuluan**

Grafik boxplot di atas memberikan gambaran visual yang sangat baik mengenai distribusi dari tiga variabel numerik penting pada data berlian, yaitu:

* **Karat:** Berat berlian
* **Depth:** Proporsi tinggi berlian terhadap diameternya
* **Table:** Lebar bagian atas berlian

Setiap boxplot mewakili distribusi dari satu variabel, memberikan informasi tentang pusat data, sebaran data, dan adanya outlier.

A screen shot of a computer program

Description automatically generated

**Breakdown Kode:**

 **Membuat Figure dan Axes:**

* fig, axes = plt.subplots(1, 3, figsize=(15, 6)):
  + plt.subplots(1, 3): Membuat sebuah figure dengan 1 baris dan 3 kolom subplot. Artinya, kita akan membuat 3 boxplot dalam satu gambar.
  + figsize=(15, 6): Mengatur ukuran figure menjadi 15 inci lebar dan 6 inci tinggi.
  + fig: Merupakan objek figure yang menampung seluruh subplot.
  + axes: Merupakan array yang berisi objek axes untuk setiap subplot.

 **Membuat Boxplot:**

* sns.boxplot(x=df\_clean['carat'], ax=axes[0]):
  + sns.boxplot: Fungsi dari library Seaborn untuk membuat boxplot.
  + x=df\_clean['carat']: Menentukan data yang akan diplot adalah kolom 'carat' dari DataFrame df\_clean.
  + ax=axes[0]: Menentukan bahwa boxplot ini akan ditempatkan pada subplot pertama (indeks 0).
* Begitu pula untuk sns.boxplot(x=df\_clean['depth'], ax=axes[1]) dan sns.boxplot(x=df\_benar['table'], ax=axes[2]), hanya saja kolom dan subplot yang berbeda.

 **Memberikan Judul:**

* axes[0].set\_title('Boxplot: Kolom Carat'): Memberikan judul "Boxplot: Kolom Carat" pada subplot pertama. Begitu pula untuk subplot lainnya.

 **Menampilkan Plot:**

* plt.tight\_layout(): Menyesuaikan layout plot agar tidak ada bagian yang tumpang tindih.
* plt.show(): Menampilkan plot yang telah dibuat.

**Output Kode:**

A blue rectangular object with black lines

Description automatically generated

**Interpretasi Grafik:**

Dari ketiga boxplot yang disajikan, kita dapat menarik beberapa kesimpulan:

1. **Tidak Ada Outlier yang Signifikan:**
   * **Carat:** Boxplot untuk kolom carat menunjukkan bahwa sebagian besar data terkonsentrasi di sekitar median, dengan sedikit variasi. Tidak ada outlier yang terlihat secara signifikan, yang mengindikasikan bahwa berat berlian dalam dataset cukup seragam.
   * **Depth:** Boxplot untuk kolom depth juga menunjukkan distribusi data yang cukup rapat di sekitar median, tanpa adanya outlier yang mencolok. Ini mengindikasikan bahwa kedalaman berlian dalam dataset relatif konsisten.
   * **Table:** Boxplot untuk kolom table juga menunjukkan distribusi data yang cukup baik, tanpa adanya outlier yang signifikan. Ini mengindikasikan bahwa ukuran meja (table) berlian dalam dataset relatif stabil.
2. **Distribusi Data yang Cukup Simetris:**
   * Bentuk kotak pada ketiga boxplot menunjukkan bahwa distribusi data cenderung simetris. Artinya, jumlah data di bawah median kira-kira sama dengan jumlah data di atas median.

**Implikasi**

* **Kualitas Data:** Ketiadaan outlier yang signifikan pada ketiga variabel ini menunjukkan bahwa data yang digunakan untuk analisis memiliki kualitas yang baik. Tidak adanya outlier yang ekstrem dapat mengurangi bias dalam analisis.
* **Konsistensi Data:** Distribusi data yang relatif seragam pada ketiga variabel menunjukkan bahwa data yang dikumpulkan memiliki tingkat konsistensi yang baik. Hal ini memperkuat reliabilitas analisis yang dilakukan.
* **Asumsi Analisis:** Ketiadaan outlier dan distribusi data yang cukup simetris mendukung asumsi-asumsi yang sering digunakan dalam analisis statistik, seperti asumsi normalitas.

Berdasarkan analisis boxplot, dapat disimpulkan bahwa data yang digunakan untuk analisis memiliki kualitas yang baik dan distribusi yang relatif normal. Ketiadaan outlier yang signifikan pada variabel carat, depth, dan table mengindikasikan bahwa data tersebut dapat diandalkan untuk analisis lebih lanjut.

# Kesimpulan

Berdasarkan analisis visualisasi yang dilakukan pada data berlian, kita dapat menarik beberapa kesimpulan yang relevan mengenai hubungan antar variabel dan kualitas data. Berikut adalah temuan utama dari analisis tersebut:

1. **Scatter Plot: Carat vs Price**
   * **Korelasi Positif**: Ada hubungan positif antara berat berlian (carat) dan harga berlian, yang berarti semakin berat berlian, semakin tinggi harganya. Ini merupakan temuan yang logis, karena berlian dengan ukuran yang lebih besar cenderung lebih bernilai.
   * **Variasi Harga**: Walaupun ada tren harga yang meningkat seiring dengan bertambahnya berat, ada variasi harga yang signifikan, yang menunjukkan bahwa faktor lain seperti kualitas potongan (cut), warna (color), dan kejernihan (clarity) turut mempengaruhi harga berlian.
   * **Segmentasi Pasar**: Data menunjukkan adanya konsentrasi pada berlian dengan berat antara 1 hingga 2 karat yang cenderung memiliki harga pada rentang menengah. Ini mungkin mencerminkan pasar yang lebih luas dan lebih umum.
2. **Heatmap: Korelasi Antara Carat, Depth, dan Table**
   * **Korelasi Positif Lemah Antara Carat dan Table**: Secara umum, berlian dengan berat yang lebih besar cenderung memiliki ukuran meja yang lebih besar. Namun, korelasi ini cukup lemah.
   * **Korelasi Negatif Lemah Antara Depth dan Table**: Terdapat hubungan negatif yang lemah antara kedalaman berlian dan ukuran meja, yang berarti semakin dalam berlian, ukuran meja cenderung lebih kecil.
   * **Tidak Ada Korelasi yang Signifikan Antara Carat dan Depth**: Tidak ditemukan hubungan linear yang kuat antara berat berlian dan kedalamannya.
3. **Boxplot: Distribusi Variabel Carat, Depth, dan Table**
   * **Ketiadaan Outlier Signifikan**: Tidak ada outlier yang mencolok pada ketiga variabel (carat, depth, dan table). Hal ini menunjukkan bahwa data dalam dataset ini cukup konsisten dan tidak ada nilai ekstrem yang dapat mempengaruhi analisis.
   * **Distribusi Simetris**: Ketiga variabel menunjukkan distribusi yang cukup simetris, yang mendukung asumsi normalitas dalam analisis statistik.

**Implikasi**

* **Faktor Lain dalam Penentuan Harga**: Dari hasil scatter plot, terlihat bahwa harga berlian dipengaruhi oleh lebih banyak faktor selain berat, seperti kualitas potongan, warna, dan kejernihan. Oleh karena itu, untuk prediksi harga yang lebih akurat, model prediktif sebaiknya mempertimbangkan variabel-variabel tersebut.
* **Kualitas Data**: Ketiadaan outlier yang signifikan dan distribusi data yang cukup simetris menunjukkan bahwa data yang digunakan dalam analisis ini memiliki kualitas yang baik, yang dapat memberikan dasar yang kuat untuk analisis lebih lanjut.
* **Pengembangan Model**: Hasil dari heatmap dapat digunakan untuk membangun model prediksi harga berlian. Variabel dengan korelasi yang kuat dengan harga dapat digunakan sebagai fitur dalam model tersebut.

**Rekomendasi**

* **Segmentasi Pasar**: Berdasarkan temuan dari scatter plot, perusahaan berlian dapat mempertimbangkan untuk mengidentifikasi segmen pasar berdasarkan berat berlian dan harga, dengan fokus pada berlian dengan berat antara 1 hingga 2 karat sebagai segmen yang lebih luas.
* **Analisis Lanjutan**: Untuk lebih memahami hubungan antara faktor-faktor yang mempengaruhi harga berlian, disarankan untuk melakukan analisis regresi atau menggunakan teknik analisis multivariat seperti analisis komponen utama (PCA).

Secara keseluruhan, analisis ini memberikan wawasan yang berguna untuk memahami faktor-faktor yang mempengaruhi harga berlian dan dapat digunakan untuk membantu keputusan bisnis dan pengembangan model prediksi harga.