Latihan Data Cleaning

Link untuk Data: https://github.com/drninhayati/PPK-AI-DAN-DATA-ANALITIK

Ringkasan Hasil Cleaning

- **Tarikh** → standard YYYY-MM-DD.
- **Harga** → nombor RM tanpa simbol.
- **Produk** → hanya “Laptop Acer / Printer HP / Mouse Logi”.
- **Quantity** → semua nombor, tiada teks.
- **Cawangan** → konsisten (Kuala Lumpur, Penang, Johor Bahru).
- **Tiada duplikasi, tiada data tidak relevan, tiada ejaan salah.**

1. Missing Values (Nilai Kosong)

Isu: Ada kolum kosong pada Tarikh, Produk, Quantity, Harga_Unit, Cawangan.
Contoh:

- OrderID 1014 (Tarikh kosong)
- OrderID 1005 (Produk kosong)
- OrderID 1019 (Cawangan kosong)

Langkah Cleaning:

1. **Kenal pasti nilai kosong**
   - Gunakan filter Blanks pada Excel atau formula:
     - **=COUNTBLANK(A2:G2)**
   - untuk kira berapa sel kosong pada setiap baris.
2. **Pilih kaedah pembetulan:**
   - **Isi dengan 0** (jika numeric seperti Quantity/Jumlah).
   - **Isi dengan “Unknown”** untuk kategori teks (contoh Produk).
   - **Isi dengan N/A** jika maklumat tidak wujud.
   - **Isi dengan nilai imputasi** (Mean, Median, atau Mode) → guna fungsi:
     - =IF(ISBLANK(D2),AVERAGE(D$2:D$100),D2)
   - untuk isi Quantity kosong dengan purata.
3. **Hasil Akhir:** Semua sel kosong sudah diganti dengan nilai yang sesuai.

2. Data Duplikasi

Isu: Rekod yang sama berulang (cth: OrderID 1008 & 1003, OrderID 1018 & 1006).
Langkah Cleaning:

1. Pilih semua data → pergi ke menu **Data → Remove Duplicates**.

2. Tandakan semua kolum kecuali OrderID (sebab ID unik boleh lain tapi data sama).
3. Klik OK. Excel akan paparkan berapa rekod duplikasi dibuang.
4. Jika perlu **semak dahulu** sebelum buang → guna formula:
    - **=COUNTIFS(B:B,B2,C:C,C2,D:D,D2)**
5. untuk kira rekod berapa kali muncul.
6. **Hasil Akhir:** Dataset tinggal rekod unik sahaja.

3. Format Data

Isu:

- Tarikh bercampur format → 01/01/2025, 20250120, 12.01.2025.
- Harga bercampur → RM2,500, 2500, 1,20.

Langkah Cleaning:

1. **Tarikh**
    - Pilih kolum tarikh → Data → Text to Columns → Date → YMD.
    - Tukar semua tarikh ke format ISO:
    - =TEXT(B2,"yyyy-mm-dd"
2. **Harga**
    - Buang simbol RM & , dengan **Find & Replace**.
    - Tukar semua jadi **Number** (Right-click → Format Cells → Number).
    - Jika ada format pelik seperti 1,20 → tukar ke 120. (Cara Lain Klik Sini)
    - Jika ada USD → buat kolum baru Harga_Unit_RM:
        - **=IF(RIGHT(E2,3)="USD",VALUE(LEFT(E2,LEN(E2)-3))*4.67,E2)**
3. **Hasil Akhir:** Semua tarikh dalam YYYY-MM-DD & semua harga dalam nombor RM.

4. Outlier / Nilai Pelik

Isu:

- OrderID 1006: Quantity = 200 (sangat besar).
- OrderID 1016: Harga = RM999,999.

Langkah Cleaning:

1. Gunakan **Conditional Formatting → Highlight Cell Rules → Greater Than** untuk highlight nilai pelik (contoh >100 untuk Quantity).
2. Semak data asal → sahkan dengan pemilik data (adakah betul atau typo).
3. Jika **typo** → betulkan kepada nilai normal (cth: 200 → 2).
4. Jika **benar** (contoh jualan besar) → kekalkan tapi tandakan sebagai special case.
5. **Hasil Akhir:** Dataset bersih daripada nilai ekstrem yang tidak logik.

5. Unit & Skala Tidak Seragam

Isu: OrderID 1007 harga dalam USD.

Langkah Cleaning:

1. Tambah kolum baru Harga_Unit_RM.
2. Formula:
    - `=IF(RIGHT(E2,3)="USD",VALUE(LEFT(E2,LEN(E2)-3))*4.67,E2)`
3. *(andaikan tukaran 1 USD = RM4.67)*.
4. Buang kolum lama Harga_Unit bila semua data sudah konsisten.
5. **Hasil Akhir:** Semua harga dalam satu unit sahaja iaitu RM.

6. Data Tidak Relevan

Isu:

- OrderID 1009: Produk = Nasi Lemak.
- OrderID 1017: Produk = Durian.

Langkah Cleaning:

1. Gunakan **Filter** → tapis produk selain “Laptop Acer / Printer HP / Mouse Logi”.
2. Tandakan & buang rekod tidak relevan.
3. Jika mahu simpan, pindahkan ke sheet “Data Lain”.
4. **Hasil Akhir:** Dataset hanya mengandungi produk relevan.

7. Kesalahan Ejaan / Label

Isu:

- “Mouse Logii” → “Mouse Logi”.
- “Johorr” → “Johor Bahru”.

1. **Buat senarai rujukan (dictionary)**
    - Sheet baru: Dictionary.
    - Kolum A = Ejaan Salah
    - Kolum B = Ejaan Betul
2. Mouse Logii | Mouse Logi
3. Johorr | Johor Bahru
4. **Tambah kolum baru di dataset utama** (contoh kolum C = Nama Produk Asal).
5. Guna formula **VLOOKUP** untuk mapping semula:
    - =IFERROR(VLOOKUP(C2, Dictionary!A:B, 2, FALSE), C2)
        - C2 = nilai asal yang nak dibetulkan.
        - Dictionary!A:B = rujukan jadual ejaan.
        - 2 = ambil kolum kedua (Ejaan Betul).
        - IFERROR(..., C2) = jika tiada padanan, kekalkan ejaan asal.
6. Copy hasil formula → **Paste Values** untuk kekalkan data bersih.

✅ Hasil Akhir

- Semua nama produk & cawangan sudah standard:

- Mouse Logii → Mouse Logi
- Johorr → Johor Bahru

8. Formula UPPER / LOWER

- =UPPER(A2) → Semua huruf jadi BESAR (**LAPTOP ACER**)
- =LOWER(A2) → Semua huruf jadi kecil (**laptop acer**)
- =PROPER(A2) → Huruf pertama setiap perkataan jadi besar (**Laptop Acer**)

8. Konsistensi Kategori

Isu: Johor Bahru ditulis berbeza → “Johor Bahru”, “Johor BHRU”, “J.Bahru”.

Langkah Cleaning:

1. Gunakan **Pivot Table** → dapatkan senarai unik cawangan.
2. Standardkan semua ke bentuk konsisten (“Johor Bahru”).
3. Untuk elak kesilapan masa depan → guna **Data Validation (Drop-down List)** dengan kategori rasmi:
   - Kuala Lumpur
   - Penang
   - Johor Bahru
4. **Hasil Akhir:** Semua kategori konsisten dan sah.

Jenis Chart dalam Excel & Kegunaan

1. Column Chart (Bar Menegak)

- **Kegunaan:** Bandingkan nilai antara kategori.
- **Contoh:** Jumlah jualan mengikut produk (Laptop Acer, Printer HP, Mouse Logi).

2. Bar Chart (Bar Mendatar)

- **Kegunaan:** Sama seperti Column Chart tetapi mendatar. Sesuai jika nama kategori panjang.
- **Contoh:** Jumlah jualan mengikut cawangan (Kuala Lumpur, Penang, Johor Bahru).

3. Line Chart

- **Kegunaan:** Tunjuk trend perubahan dari masa ke masa.
- **Contoh:** Trend jualan harian sepanjang Januari 2025.

4. Pie Chart

- **Kegunaan:** Menunjukkan peratus sumbangan setiap kategori.
- **Contoh:** Peratus sumbangan produk terhadap jumlah keseluruhan jualan.

5. Doughnut Chart

- **Kegunaan:** Sama seperti Pie Chart tetapi dalam bentuk cincin. Lebih sesuai untuk banding beberapa siri data.
- **Contoh:** Sumbangan jualan produk mengikut cawangan.

6. Area Chart

- **Kegunaan:** Menunjukkan magnitud trend dalam tempoh masa dengan kawasan berwarna.
- **Contoh:** Perubahan kumulatif jualan sepanjang minggu.

7. Scatter (X-Y) Chart

- **Kegunaan:** Untuk analisis hubungan (correlation) antara dua pembolehubah.
- **Contoh:** Hubungan Quantity dengan Jumlah Jualan.

8. Combo Chart

- **Kegunaan:** Gabungkan dua carta berbeza dalam satu (contoh Column + Line).
- **Contoh:** Column untuk Jumlah Jualan + Line untuk Quantity terjual.

9. Funnel Chart (Excel versi terbaru / Office 365)

- **Kegunaan:** Menunjukkan proses step-by-step yang semakin mengecil.
- **Contoh:** Bilangan jualan dari prospek → tempahan → bayaran.

10. Treemap & Sunburst

- **Treemap:** Tunjukkan data hierarki dengan blok berwarna.
- **Sunburst:** Tunjukkan data hierarki dalam bentuk cincin berlapis.
- **Contoh:** Produk → Sub-produk → Jualan.

📊Pivot Table

Berdasarkan data hasilkan graf di bawah:

1. Jumlah Jualan Mengikut Produk
2. Jumlah Jualan Mengikut Cawangan
3. Trend Jualan Harian (Mengikut Tarikh)
4. Perbandingan Quantity vs Jumlah Jualan
5. Penjualan Produk Mengikut Cawangan (Matrix)

6. Top 3 Order Tertinggi (OrderID)
7. Purata Quantity & Purata Nilai Jualan per Produk
8. Peratus Sumbangan Produk kepada Jumlah Keseluruhan

1. Jualan Mengikut Produk

- Fields:
  - Rows: Produk
  - Values: Jumlah (Sum)
- Objektif: Lihat produk mana paling tinggi jumlah jualan (Laptop Acer / Printer HP / Mouse Logi).
- Aktiviti: Pelajar bandingkan top-selling product.

2. Jualan Mengikut Cawangan

- Fields:
  - Rows: Cawangan
  - Values: Jumlah (Sum)
- Objektif: Kenal pasti cawangan paling aktif menjual.
- Aktiviti: Minta pelajar buat chart (bar/column) untuk visualkan perbandingan.

3. Trend Jualan Mengikut Tarikh

- Fields:
  - Rows: Tarikh
  - Values: Jumlah (Sum)
- Objektif: Analisis trend jualan harian sepanjang Januari 2025.
- Aktiviti: Tukar ke Pivot Chart (Line) untuk lihat pola kenaikan/penurunan jualan.

4. Analisis Quantity vs Jumlah Jualan

- Fields:
  - Rows: Produk
  - Values: Quantity (Sum), Jumlah (Sum)
- Objektif: Bandingkan jumlah unit terjual dengan hasil jualan.
- Aktiviti: Diskusikan produk margin tinggi (Laptop Acer → nilai tinggi walau unit sedikit).

5. Penjualan Produk Mengikut Cawangan

- Fields:
  - Rows: Cawangan
  - Columns: Produk
  - Values: Quantity (Sum)
- Objektif: Lihat setiap cawangan jual produk apa paling banyak.
- Aktiviti: Tanya pelajar – adakah semua cawangan fokus pada produk sama?

6. Top 3 Order Tertinggi

- Fields:
  - Rows: OrderID
  - Values: Jumlah (Sum)
- Objektif: Kenal pasti order mana yang paling besar nilainya.
- Aktiviti: Susun descending → highlight top 3 sales.

7. Perbandingan Unit vs Nilai Jualan

- Fields:
  - Rows: Produk
  - Values: Quantity (Average), Jumlah (Average)
- Objektif: Faham beza purata unit terjual dengan purata nilai RM.
- Aktiviti: Bincang mengapa produk murah (Mouse Logi) tinggi quantity tapi nilai kecil.

8. Contribution % (Peratus Sumbangan)

- Fields:
  - Rows: Produk
  - Values: Jumlah (Show Values As → % of Grand Total)
- Objektif: Lihat peratus sumbangan setiap produk pada total jualan.
- Aktiviti: Pelajar buat pie chart → perbandingan visual lebih jelas.

Aktiviti Pivot Table	Fields (Row/Column/Values)	Chart Sesuai	Kegunaan
1. Jumlah Jualan Mengikut Produk	Rows = Produk, Values = Jumlah (Sum)	Column Chart	Bandingkan produk mana paling tinggi nilainya.
2. Jumlah Jualan Mengikut Cawangan	Rows = Cawangan, Values = Jumlah (Sum)	Bar Chart	Bandingkan prestasi setiap cawangan (sesuai untuk nama kategori panjang).
3. Trend Jualan Harian (Mengikut Tarikh)	Rows = Tarikh, Values = Jumlah (Sum)	Line Chart	Analisis trend kenaikan/penurunan jualan mengikut masa.
4. Perbandingan Quantity vs Jumlah Jualan	Rows = Produk, Values = Quantity (Sum), Jumlah (Sum)	Combo Chart (Column + Line)	Bandingkan bilangan unit terjual vs nilai jualan (margin analysis).
5. Penjualan Produk Mengikut Cawangan (Matrix)	Rows = Cawangan, Columns = Produk, Values = Quantity (Sum)	Clustered Column Chart	Nampak produk mana paling laris dalam setiap cawangan.
6. Top 3 Order Tertinggi (OrderID)	Rows = OrderID, Values = Jumlah (Sum), Sort = Descending	Bar Chart (Top 3)	Kenal pasti order paling besar → highlight 3 teratas.
7. Purata Quantity & Purata Nilai Jualan per Produk	Rows = Produk, Values = Quantity (Average), Jumlah (Average)	Clustered Column Chart	Bandingkan purata unit terjual dengan purata nilai RM.
8. Peratus	Rows = Produk,	Pie Chart /	Tunjuk peratus

Sumbangan Produk kepada Jumlah Keseluruhan	Values = Jumlah (% of Grand Total)	Doughnut Chart	sumbangan setiap produk terhadap total jualan keseluruhan

Contoh: AI Prompt untuk menulis result and discussion

*Berdasarkan carta/graf yang diberikan, tuliskan bahagian **Results and Discussion** secara akademik dan berimpak tinggi. Pastikan:*

1. **Results** – Huraikan dapatan kuantitatif dan kualitatif daripada carta, sertakan trend utama, perbandingan, dan nilai penting.
2. **Discussion** – Kaitkan dapatan dengan teori atau kajian lepas, jelaskan sebab-sebab kemungkinan, serta implikasi praktikal dan akademik.
3. Gunakan **academic conjunctions** seperti *furthermore, in contrast, significantly, consequently, therefore, notably*.
4. Pastikan gaya penulisan sesuai untuk **high-impact journal** (formal, kritikal, berstruktur).
5. Tutup dengan pernyataan tentang **impak keseluruhan dapatan** terhadap bidang kajian