**Pengenalan Exploratory Data Analysis**

Exploratory Data Analysis (EDA) merupakan tahap eksplorasi data yang telah dibersihkan guna memperoleh*insight*dan menjawab pertanyaan analisis. Pada prosesnya, kita akan sering menggunakan berbagai teknik dan parameter dalam *descriptive statistics* yang bertujuan untuk menemukan pola, hubungan, serta membangun intuisi terkait data yang diolah. Selain itu, tidak jarang kita juga menggunakan visualisasi data untuk menemukan pola dan memvalidasi parameter descriptive statistics yang diperoleh.

Exploratory analysis dan explanatory analysis merupakan dua istilah sering dijumpai dalam proyek analisis data.

Sederhana, exploratory analysis merupakan proses analisis data yang bertujuan untuk mengeksplorasi dan mengenal sebuah data. Pada prosesnya sering kali mulai dengan mendefinisikan berbagai pertanyaan atau hanya menggali data guna menemukan beberapa insight yang menarik dari sebuah data.

Di lain sisi, explanatory analysis merupakan proses analisis data yang bertujuan untuk membagikan beberapa insight yang menarik dari sebuah data. Insight inilah yang akan digunakan untuk menjawab permasalahan bisnis yang dihadapi.

Jika disimpulkan, explanatory analysis merupakan proses penyampaian temuan menarik dari proses exploratory analysis. Proses penyampaian ini tentunya harus diikuti dengan visualisasi data yang baik dan efektif.

**Mendefinisikan Pertanyaan Untuk Explorasi Data**

Sebagai calon praktisi data masa depan, tentunya Anda telah mengetahui bahwa objektif dari sebuah proyek analisis data ialah menjawab tantangan atau permasalahan bisnis. Untuk memastikan pemahaman terhadap sebuah permasalahan bisnis, kita perlu mengurai masalah tersebut ke dalam beberapa pertanyaan analisis.

Problem solving merupakan salah satu keterampilan penting yang harus dimiliki seorang praktisi data. Untuk menyelesaikan sebuah problem, kita harus memahami terlebih dahulu permasalahan yang ingin diselesaikan. Berikut merupakan beberapa kategori problem yang sering dijumpai oleh para praktisi data.

* **Membuat prediksi**  
  Membuat prediksi merupakan salah satu problem yang paling sering dijumpai oleh praktisi data di industri. Jenis problem ini, mengharuskan kita untuk membuat sebuah prediksi berdasarkan hasil analisis dari data sebelumnya.
* **Mengategorikan berbagai hal**  
  Problem lain yang umum dijumpai di industri ialah kebutuhan untuk mengategorikan berbagai hal. Pada problem ini, kita harus menetapkan sebuah informasi ke dalam berbagai grup atau kluster berdasarkan kesamaan yang ada.
* **Menemukan sesuatu yang tidak biasa**  
  Kategori problem selanjutnya ialah mengidentifikasi atau menentukan sesuatu yang tidak biasa. Problem ini, umumnya muncul ketika terdapat sebuah kejadian aneh yang perlu dianalisis lebih lanjut.
* **Mengidentifikasi tema**  
  Permasalahan lain yang mungkin akan Anda temui ialah mengidentifikasi tema. Problem ini sering dijumpai dalam bidang UI/UX. Seorang UI/UX *designer* harus mengidentifikasi tema yang paling tepat dalam merepresentasikan *anonymous survey data*dari para pengguna. Hal ini bertujuan untuk meningkatkan pengalam pengguna ketika menggunakan sebuah produk.
* **Menemukan hubungan**  
  Kategori berikutnya ialah menemukan hubungan atau *connection*. Problem ini mengharuskan kita untuk mengidentifikasi dan menganalisis hubungan dari berbagai data untuk mengatasi suatu permasalahan.
* **Menemukan pola**  
  *Last but not least*, kategori problem yang mungkin Anda temui ialah menemukan pola. Jenis problem ini, membutuhkan analisis mendalam terhadap sebuah data guna memahami pola yang terdapat di dalamnya.

Jika data merupakan building block dalam proyek analisis data, kita dapat menganggap pertanyaan analisis sebagai kompas yang selalu mengarahkan untuk menuju goals dan objektif yang ingin dicapai. Pada proses pembuatan pertanyaan analisis, kita perlu menghindari beberapa poin berikut.

* Hindari membuat pertanyaan yang mengarahkan kita untuk menjawab dengan jawaban tertentu.
* Hindari membuat pertanyaan yang hanya dapat dijawab dengan “ya” atau “tidak.”
* Hindari membuat pertanyaan yang tidak memiliki konteks.

Untuk menghasilkan pertanyaan analisis yang efektif, terdapat beberapa poin yang harus Anda perhatikan seperti berikut.

* **Specific**  
  Pertanyaan yang efektif haruslah bersifat spesifik, artinya sebuah pertanyaan analisis harus fokus pada sebuah topik tertentu. Pertanyaan yang spesifik dapat membantu kita dalam mengumpulkan informasi yang relevan dengan tujuan dari proyek analisis data.
* **Measurable**  
  Selain bersifat spesifik, pertanyaan yang efektif haruslah bisa diukur atau *measurable*. Hal ini tentunya akan mempermudah kita dalam memahami dan menarik insight dari sebuah data.
* **Action-oriented**  
  Pertanyaan yang bersifat *action-oriented* berarti pertanyaan tersebut mampu memberikan informasi yang dapat membantu kita dalam membuat sebuah action-plan.
* **Relevant**  
  Pertanyaan yang relevan berarti pertanyaan tersebut berbobot dan memiliki arti untuk menyelesaikan sebuah permasalahan.
* **Time-bound**  
  Terakhir, pertanyaan yang efektif haruslah bersifat time-bound yang berarti pertanyaan tersebut haruslah memiliki relevansi terhadap waktu saat ini.

**Mengeksplorasi Parameter Statistik dari Data**

Method **describe()** memungkinkan kita untuk melihat rangkuman parameter statistik (mean, median, dll.) dari kolom numerik pada sebuah DataFrame. Jika ingin melihat lebih banyak parameter statistik serta menyertakan kolom non-numerik, Anda bisa menggunakan parameter **include=”all”**.

Selain method **describe()**, pandas juga menyediakan method lain seperti **hist()** untuk membuat plot histogram dari kolom numerik pada sebuah DataFrame. Method ini dapat digunakan untuk memeriksa distribusi data secara cepat.

Jika ingin memeriksa korelasi antar data numerik pada sebuah DataFrame, Anda dapat menggunakan method **corr()** atau **cov()**.

Itulah beberapa method yang sering digunakan untuk mengeksplorasi parameter statistik dari sebuah DataFrame. Sebenarnya masih banyak method lain yang bisa Anda gunakan untuk memeriksa parameter statistik lain. Silahkan kunjungi dokumentasi berikut untuk melihat lebih detail terkait hal tersebut: [dokumentasi pandas.DataFrame](https://pandas.pydata.org/docs/reference/frame.html" \t "_blank).

**Mengelompokkan Data**

Mungkin beberapa dari Anda cukup familier dengan konsep *pivot table* dalam aplikasi pengolahan data seperti Microsoft Excel atau Google Sheets. Ia merupakan tool yang dapat digunakan untuk mengelompokkan data berdasarkan parameter tertentu.

Pada proses analisis data, kita bisa menggunakan konsep pivot table ini untuk menghitung parameter statistik dari kelompok data tertentu. Selain itu, ia juga dapat digunakan untuk kebutuhan visualisasi data dan pembuatan *report*.

Untuk memahami proses pembuatan pivot table, Anda perlu mengenal dua istilah di bawah ini telebih dahulu.

* Key: kolom yang berisi data kategorik dan digunakan untuk mengelompokkan data.
* Value: kolom yang berisi data kuantitatif dan akan dihitung nilai parameter statistiknya

Proses pembuatan pivot table, diawali dengan melakukan split atau proses pengelompokkan pasangan key-values. Tahap berikutnya adalah proses apply atau perhitungan parameter dari setiap kelompok pasangan key-value. Tahap terakhir ialah menggabungkan (*combine*) seluruh key dengan nilai parameternya ke dalam satu tabel. *Tara*…, jadilah sebuah pivot table yang akan membantu kita dalam memahami sebuah data.

Library pandas telah menyediakan sebuah method bernama groupby() yang dapat membantu kita dalam membuat pivot table secara lebih efisien dan scalable. Method ini akan menerima inputan beberapa parameter salah satunya ialah “by”. Parameter tersebut digunakan untuk menentukan key dari pivot table tersebut. Selain itu, untuk menggunakan method ini, kita perlu memadukannya dengan berbagai method untuk menghitung parameter statistik, seperti mean(), mode(), dll.

Untuk membuat pivot table dengan lebih dari satu parameter, kita bisa menggunakan method agg(). Ia memungkinkan kita untuk menentukan beberapa parameter yang ingin digunakan dalam pivot table.

Nah, itulah segenap pembahasan terkait pengelompokkan data atau pembuatan pivot table menggunakan method groupby().  Jika Anda ingin membaca lebih detail terkait method tersebut, silakan kunjungi tautan berikut: [GroupBy documentation](https://pandas.pydata.org/pandas-docs/stable/reference/groupby.html" \t "_blank).