Capstone Project II - Data Analysis

Hi nama saya Dwi Pamuji, Berikut ini adalah Capstone Project II – Data Analysis. Pembahasan akan saya mulai dari :

* + General Data Information
  + SQL
  + Data Manipulation
  + Data Visualization & Statistics

Untuk detail setiap bagiannya dapat melihat informasi dibawah ini.

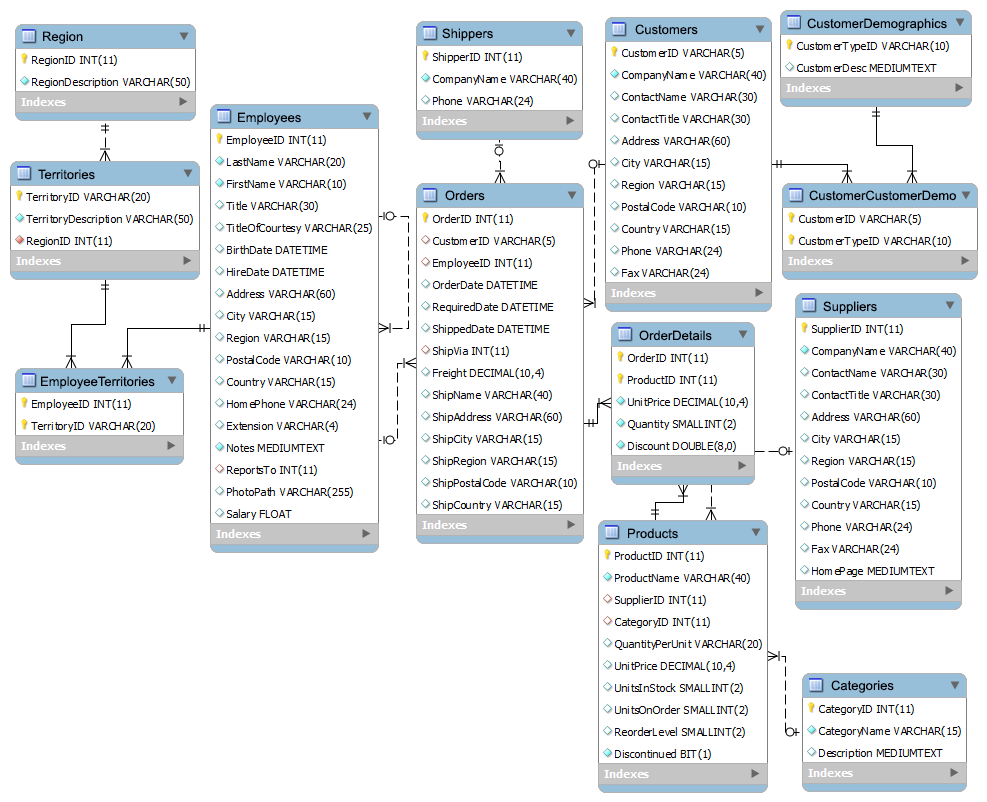
1. GENERAL DATA INFORMATION
   1. Context

Northwind merupakan perusahaan yang bergerak pada bidang impor-ekspor makanan dan minuman berlokasi di Amerika dan tidak memproduksi barangnya sendiri, Perusahaan Northwind menjual Kembali produk dari supplier ke konsumen. Produk yang dijual perushaan Northwind berupa makanan dan minuman dengan 77 jenis produk yang berbeda, yang di kelompokan atau dikategorikan menjadi 8 kategori. Perushaan ini mengimpor dari 29 perusahaan dari berbagai negara. Dan memiliki total pelanggan sebanyak 91 perusahan di berbagai negara. Dari database tersebut saya akan berfokus kepada supplier dan mengambil insight-insight yang berhubungan dengan perusahan-perusahaan supplier.

* 1. Database Information

Penjelasan mengenai data data setiap table pada database northwind

* 1. Hubungan Antar Tabel



berikut ini adalah Entity Relationship Diagram (ERD) untuk database "Northwind"

1. SQL
   1. Menghubungkan Ke Database

Membuat koneksi dari SQL ke Python agar dapat melakukan query di jupyterlab

* 1. Membuat Function Query
     1. Data Order, Detail Supplier dan Total Penjualan

\*\*Pada Analasis ini saya akan melakukan analisa yang berfokus kepada ```Supplier```\*\*

Data pertama ini merupakan data utama yang nantinya akan dianalisa lebih lanjut. Data ini merupakan gabungan dari 4 tabel, yaitu tabel ```Orders```, ```OrderDetails```, ```Product```, dan ```Suppliers```. Masing-masing dari setiap tabel tersebut diambil beberapa kolomnya dan tidak diambil secara keseluruhan. Informasi-informasi yang dianggap penting saja lah yang diambil. Informasi yang diambil antara lain adalah :

- OrderID dari tabel Orders

- OrderDate dari tabel Orders

- RequiredDate dari tabel Orders

- ShippedDate dari tabel Orders

- SupplierID dari tabel Suppliers

- Supplier\_CompanyName adalah CompanyName dari tabel Suppliers

- ContactName dari tabel Suppliers

- ContactTitle dari tabel Suppliers

- Address dari tabel Supplier

- City dari tabel Supplier

- Region dari tabel Supplier

- Country dari tabel Supplier

- ProductID dari tabel Products

- ProductName dari tabel Products

- Harga\_Jual adalah UnitPrice dari Tabel Products yang di \*\*definisikan sebagai harga jual perusahaan kepada customer\*\*

- Harga\_Beli adalah UnitPrice dari Tabel OrderDetails yang di \*\*definisikan sebagai harga beli perusahaan dari supplier\*\*

- Quantity dari tabel OrderDetails

Selain dari tabel, terdapat sebuah kolom juga yang dinamakan Total\_Harga yang merupakan hasil perkalian dari Harga\_Jual dengan Quantity.

Semua informasi tersebut kemudian dijadikan dalam sebuah DataFrame yang nantinya akan diolah informasinya.

* + 1. Total Keuntungan Setiap Produk

\*\*Pada Analasis ini saya akan melakukan analisa yang berfokus kepada ```Supplier```\*\*

Data pertama ini merupakan data utama yang nantinya akan dianalisa lebih lanjut. Data ini merupakan gabungan dari 4 tabel, yaitu tabel ```Orders```, ```OrderDetails```, ```Product```, dan ```Suppliers```. Masing-masing dari setiap tabel tersebut diambil beberapa kolomnya dan tidak diambil secara keseluruhan. Informasi-informasi yang dianggap penting saja lah yang diambil. Informasi yang diambil antara lain adalah :

- OrderID dari tabel Orders

- OrderDate dari tabel Orders

- RequiredDate dari tabel Orders

- ShippedDate dari tabel Orders

- SupplierID dari tabel Suppliers

- Supplier\_CompanyName adalah CompanyName dari tabel Suppliers

- ContactName dari tabel Suppliers

- ContactTitle dari tabel Suppliers

- Address dari tabel Supplier

- City dari tabel Supplier

- Region dari tabel Supplier

- Country dari tabel Supplier

- ProductID dari tabel Products

- ProductName dari tabel Products

- Harga\_Jual adalah UnitPrice dari Tabel Products yang di \*\*definisikan sebagai harga jual perusahaan kepada customer\*\*

- Harga\_Beli adalah UnitPrice dari Tabel OrderDetails yang di \*\*definisikan sebagai harga beli perusahaan dari supplier\*\*

- Quantity dari tabel OrderDetails

Selain dari tabel, terdapat sebuah kolom juga yang dinamakan Total\_Harga yang merupakan hasil perkalian dari Harga\_Jual dengan Quantity.

Semua informasi tersebut kemudian dijadikan dalam sebuah DataFrame yang nantinya akan diolah informasinya.

* + 1. Transaksi yang lokasi Negara Customer dan Negara Supplier Sama

Data ketiga ini menginformasikan bahwa terdapat Order yang memiliki lokasi negara customer dan negara supplier sama

1. DATA MANIPULATION
2. Apakah ada anomalies ? (Missing Value, Data Formating, Data Double, Outlier)
3. Bagaimana menangani anomalies tersebut
4. Buktikan jika sudah tidak ada anomalis
   1. Melihat Data Sekilas Dari General Info

Terlihat sekilas bahwa secara keseliruhan terdapat 2155 baris data dengan total 18 kolom. Setiap kolomnya memiliki tipe data yang berbeda-beda. Ada Object, integer, dan datetime.

Pertama-tama mari berfokus pada non-null values atau data yang tersedia pada setiap kolomnya. Jika melihat informasi tersebut, terdapat 1 kolom atau feature yang memiliki data tidak lengkap yaitu pada kolom Region. Feature Region kehilangan lebih dari 65% data.

***Kesimpulan pertama adalah bahwa terdapat missing value yang harus ditanggulangi.***

Kedua kita akan Berfokus pada features berikut ini:

1. Harga\_Beli
2. Harga\_Jual
3. Total\_Harga

Pada tipe data Harga\_Jual, Harga\_Beli, dan juga Total\_Harga. Ketiga feature ini merupakan feature yang seharusnya memiliki tipe data numerik (dibuktikan pada preview data di bagian sebelumnya), sedangkan yang terbaca tipe data dari ketiga feature ini adalah object. Artinya, ketiga feature ini tidak dianggap memiliki komponen data yang numerik. Tentu saja hal tersebut harus ditanggulangi, mengingat ke depannya data yang bersifat numerik ini akan digunakan.

***Kesimpulan keduanya adalah terdapat features yang memiliki tipe data yang salah dan harus diubah sesuai dengan tipe data seharusnya.***

* 1. ANOMALIES

Pada bagian ini saya akan mencari anomali anomali pada data dan melakukan penanganannya

* + 1. Missing Value
    2. Handling Missing Value
    3. Recheck Missing Value
    4. Mengubah Tipe Data yang Salah
    5. Data Duplicate
  1. Membuat Feature

Pada Bagian ini saya akan menambahkan beberapa feature agar Analisa menjadi lebih mudah dan efisien

* + 1. Feature Profit\_Each, Total\_Harga, dan Profit
* Profit Each = Menginformasikan keuntungan per produk tiap transaksi (Harga Jual - Harga Beli)
* Total Harga = Menginformasikan total penjualan tiap transaksi (Harga Jual x Quantity)
* Profit = Menginformasikan keuntungan tiap transaksi (Profit Each x Quantity)
  + 1. Feature ‘ProsessingDate’

ProsessingDate = pengurangan antara requiredDate dan juga shippedDate

* 1. Check Cleaned Data

Mengecek Kembali data yang sudak ‘Bersih’

* + 1. Preview Cleaned Data
    2. General Info Cleaned Data
  1. Outlier

Mengecek apakah ada data outlier dan berada dimana? dan bagaimana menanganinya

1. DATA VISUALIZATION DAN STATISTIK

Data Visualization :

Data Visulization menggunakan Python

Insight, dan sehala informasi di jelaskan melalui grafik untuk memecahkan masalah

Statistic :

Membandingkan Value/Proportion dengan methode yang sesuai

Korelasi antara focus analisis dengan variable lain

* 1. Data Visualization

melakukan analisis data dengan menggunakan visual sebagai medianya. Di sini, kita akan melakukan visualisasi data yang berfokus pada **Supplier**

* + 1. Total Penjualan Pertahun
    2. Top 5 Supplier Berdasarkan Penjualan
    3. Worst 5 Supplier Berdasarkan Penjualan
    4. 10 Product Terlaris
    5. Top 3 Supplier berdasarkan product terlaris
    6. Top Supplier berdasarkan profit yang dihasilkan dari penjualan produknya
  1. STATISTIC
     1. Perbedaan Penjualan Produk (Total\_Harga) Tiap Supplier
     2. Distribusi Feature
     3. Korelasi Antar Feature