UJIAN TENGAH SEMESTER (UTS)

Untuk Memenuhi Salah Satu Syarat Penyelesaian

Mata Kuliah Data Warehouse

Program S-1 Teknik Informatika

Oleh

Toni Hari Wibowo

NIM. 20200801314

PROGRAM STUDI S-1 TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS ESA UNGGUL

NOVEMBER 2022

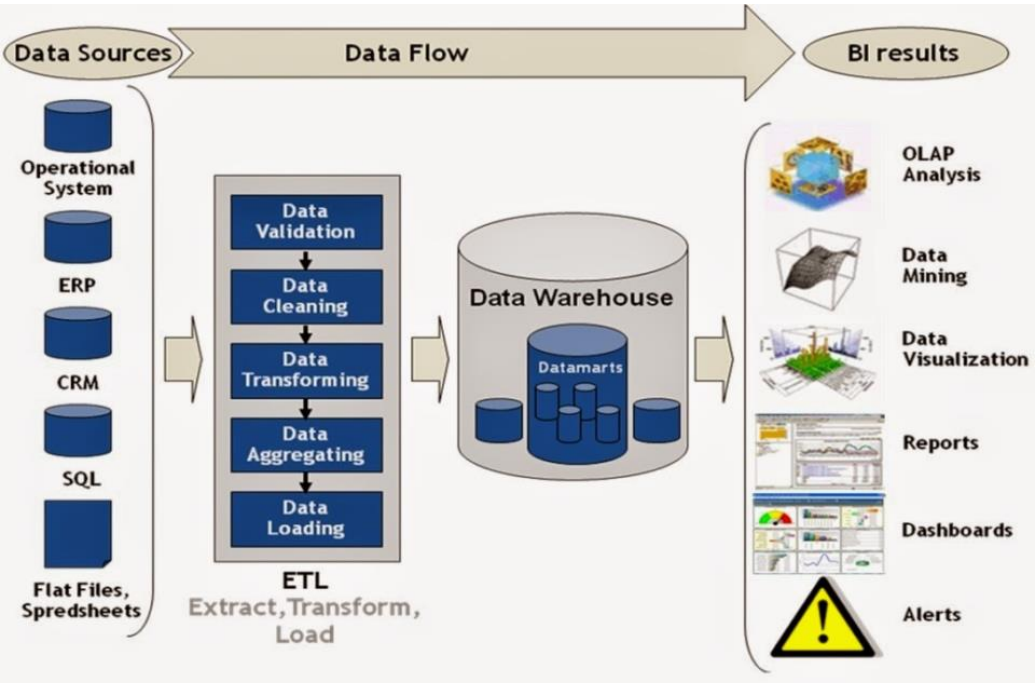
SOAL UTS

Compile lah semua penugasan yang sudah diberikan (dalam group) selama ini sehingga memenuhi ketentuan sebagai berikut:

- Gunakan data riil dari suatu bisnis (industri nya bebas) minimal 5 tahun
- Lakukan analisis kebutuhan dengan minimal mengacu kepada 3 driven untuk mencerminkan kondisi riil dari data yang dimiliki
- Buatlah dokumentasi dari mulai analisis kebutuhan dan disain konseptual dari data yang sudah didapatkan selengkap mungkin

Jawab :

Dari soal diatas maka penulis akan menjawab soal tersebut sebagai berikut, dimana penulis menggunakan dataset yang di ambil dari dataset publik E-commerce Brazilian by Olist. Dataset ini memiliki informasi 100 ribu pesanan dari tahun 2016 hingga 2018 yang dibuat di beberapa pasar di brasil. Fitur-fiturnya memungkinkan melihat pesanan dari berbagai dimensi: dari status pesanan, harga, kinerja pembayaran dan pengiriman hingga lokasi pelanggan, atribut produk, dan akhirnya ulasan yang ditulis oleh pelanggan. Selain itu di dalam dataset ini, olist juga merilis kumpulan data geolokasi yang menghubungkan kode pos brasil dengan koordinat lat/lng. Dalam UTS ini penulis memperoses dataset tersebut melalui alur sebagai berikut :



Gambar 1. Flow

1. **Data Sources**

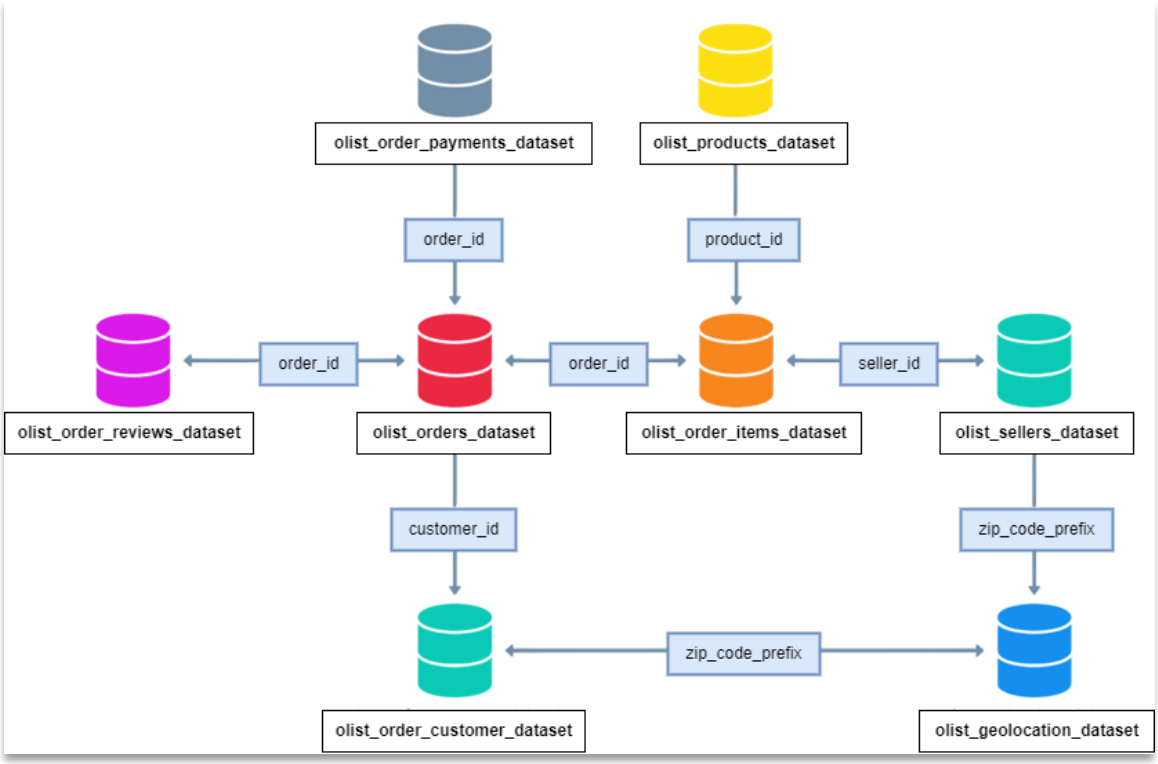
   Data ini di ambil dapat dari berbagai sumber data, dengan contoh lain pada pengejaan UTS ini data di ambil dari situs Kaggle dataset publik dalam bentuk Flat Files yang berformatkan **.csv**

2. **Data Flow**

   Dalam proses ini dataset yang sudah didapatkan akan di proses dengan beberapa tahapan layer sebelum dilakukannya load atau dimasukan kedalam data warehouse, tahap-tahap tersebut diantaranya *Extract, Transform, Load* (ETL) :

   - **Extract**

     Pada layer ini merupakan tahapan *Extracting* data, sebelum data digunakan terlebih dahulu data yang berformatkan file **.csv** tersebut akan dibaca menggunakan program python sehinnga terbentuknya dataframe, berikut adalah gambaran schema dataset yang digunakan:

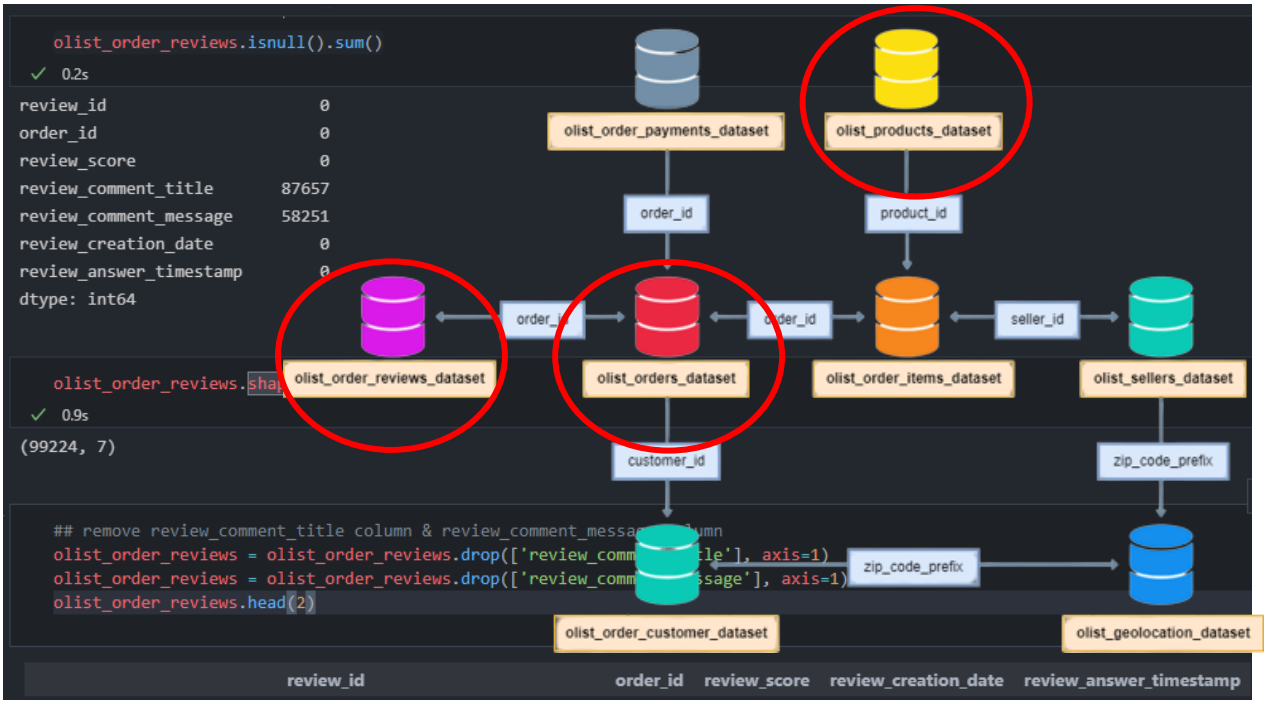


Gambar 2. Schema Dataset

Pada gambar 2 diatas merupakan gambaran dari schema dataset yang berasal dari file e-commerce Brazilian by olist dengan berjumlah 9 file. File-file tersebut nantinya akan diintegrasikan atau digabungkan menjadi satu kesautan yang utuh melalui foreign key yang terdapat pada dataset tersetbut.

- **Transform**

  Sebelum proses penggabungan data, dilakukannya transformasi data, hal tersebut berguna untuk melakukan pembersihan dari baris yang kosong pada data yang akan dipakai untuk divisualisasi dan di analisa.



Gambar 3. Cleaning Data

Pada gambar 3 diatas merupakan gambaran pada saat tahap cleaning dataframe, pada saat dilakukannya pengecekan disetiap dataframe didapatkan 3 dataframe yang memiliki baris yang kosong diantranya yaitu : olist_order_reviews, olist_orders dan olist_products. Dalam program python terdapat beberapa solusi untuk melakukan data cleaning, dan metode tersebut nantinya dapat dipilih dengan menyesuaikan kebutuhan data yang akan dipakai untuk divisualisasi dan di analisa. Berikut adalah metode-metode data cleaning yang terdapat pada program python :

1) Dropna(how=’all)

   jika semua kolom tidak mengandung nilai maka metode ini akan menghapus dari kolom tersebut.

2) Fillna(0)

Lain halnya dengan dropna, jika hanya beberapa nilai yang hilang di kolom, maka metode ini digunakan untuk mengisi nilai NA/NaN menggunakan nilai yang ditentukan.

3) fillna(method='ffill')

   Metode ini merupakan metode duplikasi data dengan mengisikan nilai yang hilang di kolom dengan mengisikan nilai di baris **sebelumnya**.

4) fillna(method='bfill')

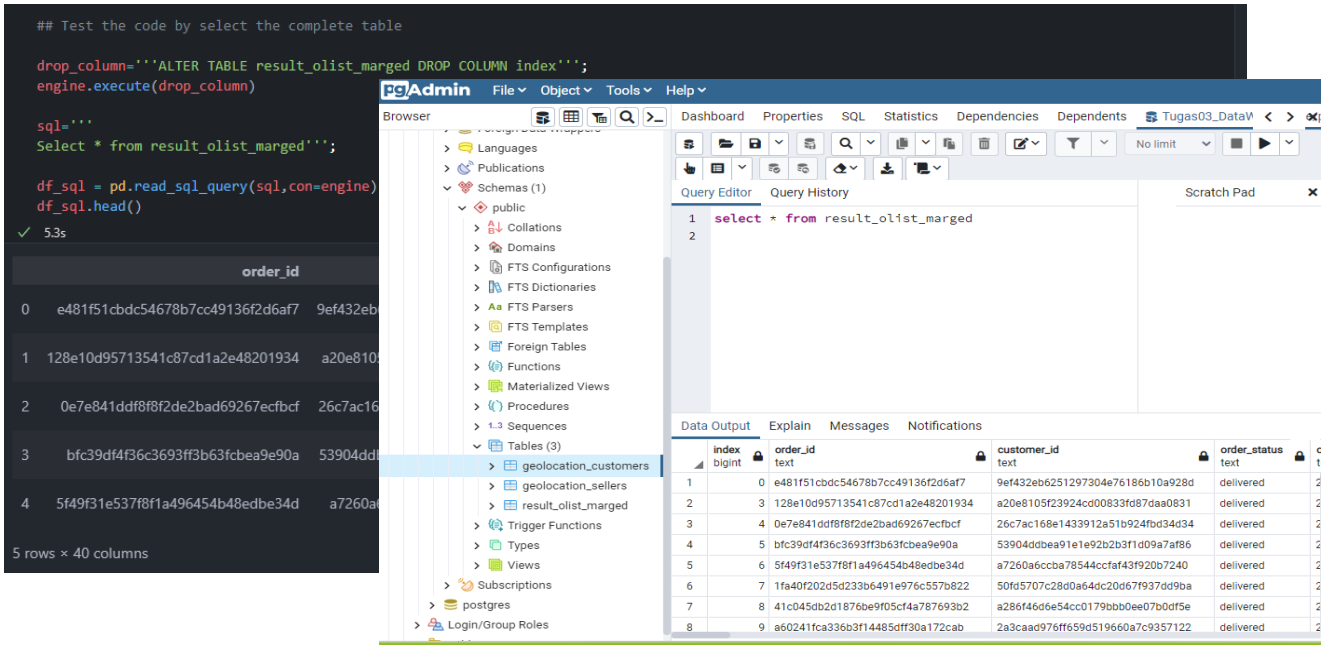
   Metode ini merupakan metode duplikasi data dengan mengisikan nilai yang hilang di kolom dengan mengisikan nilai di baris **setelahnya**.

5) interpolate()

   Metode ini merupakan metode duplikasi data dengan mengisikan nilai yang hilang di kolom dengan mengisikan nilai diantara baris **sebelumnya** dan baris **setelahnya**.

- **Load**

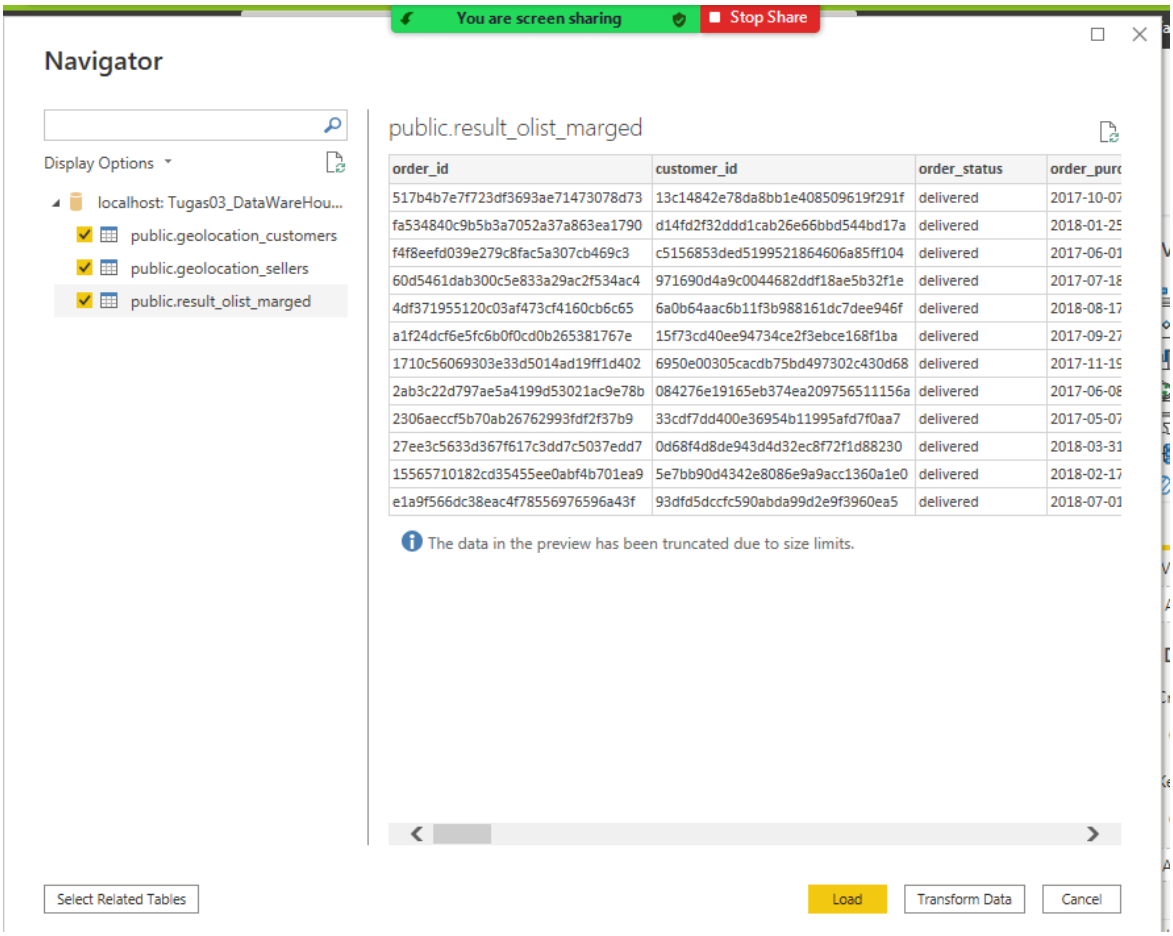
  Pada layer ini merupakan proses melakukan pengiriman data kedalam postgreSQL. Data yang sudah dilakukan pembersihan dan telah di integrasi menjadi satu kesatuan akan di load kedalam database yang ada di postgreSQL. Berikut adalah gambaran pada saat melakukan push kedalam database yang ada di postgreSQL



Gambar 4. Push Data into PostgreSQL

3. Business intelligence (BI) Results

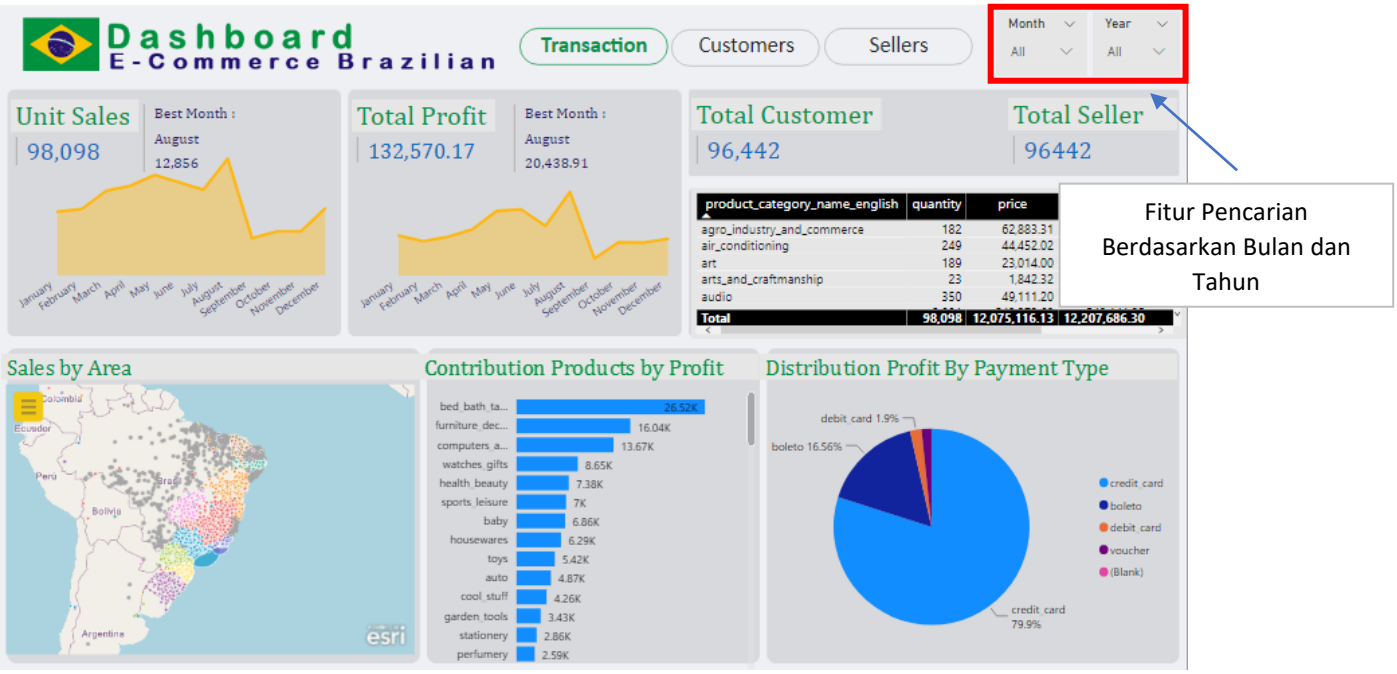
Pada tahap ini merupakan penerapan teknik transformasi data dari data mentah menjadi informasi yang berguna dan bermakna untuk tujuan analisis, salah satu teknik yang digunakan untuk melakukan pengerjaan SOAL UTS ini yaitu membuat dashboard di PowerBI dengan cara menghubungkan database server PostgreSQL dengan PowerBI. Data yang telah disimpan di dalam PostgreSQL nantinya akan di panggil kembali kedalam PowerBI, sehingga perlu dihubungkan antara PostgreSQL dengan PowerBI yang nantinya di dalam PowerBI data tersebut akan di visualisasikan agar dapat dengan mudah di analisis. Berikut adalah gambaran proses hasil load data dari PostgreSQL kedalam PowerBI :



Gambar 5. Load Data Into PowerBI

Hasil

A. Halaman Transaction



Gambar 6. Dashboard PowerBI, Halaman Transaction

Pada gambar 6 diatas merupakan tampilan dari Dashboard yang telah dibuat dan pada gambar tersebut merupakan tampilan infromasi dari Halaman Transaction seperti gambar 6 diatas, penulis membuat beberapa fitur salah satunya untuk pencarian total transaksi berdasarkan bulan dan tahun yang ingin di tentukan, fitur ini penulis buat berdasarkan bulan dan tahun yang di ambil dari kolom **order_delivered_customer** yang ada pada database. Selain itu masih ada beberapa informasi yang penulis tampilkan pada halaman dashboard Transaksi seperti yang ada pada gambar Pada **Gambar 6.** Dashboard PowerBI, Halaman Transaction, yaitu berupa :

- **Unit Sales**

  Informasi Unit Sales yang penulis buat menggunakan visualisasi multi_row card dan juga Area Chart, dari visualisasi multi_row card menggunakan penjumlahan keseluruhan data dari variable order_item_id untuk menghasilkan informasi, selain itu dari visualisasi Area Chart penulis juga menggunakan order_item_id yang di bandingkan dengan variable order_delivered_date dan di filter berdasarkan top 1 untuk pencarian bulan terbaik pada Unit Sales.

- **Total Profit**

  Informasi Total Profit yang penulis buat menggunakan visualisasi multi_row card dan juga Area Chart, dari visualisasi multi_row card menggunakan penjumlahan keseluruhan data dari variable profit, dimana variable profit dibuat setelah pembuatan variable revenue, variable variable tersebut penulis buat terlebih dahulu agar nantinya dapat digunkan untuk memberikan infor masi berupa total profit, berikut adalah rumus yang penulis gunakan :

```
revenue = SUMX('public result_olist_marged', 'public
result_olist_marged'[order_item_id]*'public result_olist_marged'[price])
```

Table 1. Rumus penghitungan revenue

```
profit = SUMX('public result_olist_marged', 'public result_olist_marged'[revenue]-
'public result_olist_marged'[price])
```

Table 2. Rumus penghitungan profit

- **Total Customer**

  Informasi Total Customer yang penulis buat menggunakan visualisasi **multi_row card** yang menggunakan penjumlahan keseluruhan data dari variable **customer_unique_id**, keputusan variable yang penulis gunakan untuk pembuatan informasi tersebut dikarenakan bahwa setiap customer memiliki id yang unique, sehingga tidak ada duplikasi atau redundan pada penghitungan data.

- **Total Seller's**

  Informasi Total Seller's yang penulis buat menggunakan visualisasi **multi_row card** yang menggunakan penjumlahan keseluruhan data dari variable seller_id, dari data yang ada pada kolom seller_id penulis filter terlebih dahulu agar nantinya tidak ada id yang redundan untuk di jumlahkan.

- **Tabel**

  Informasi pada Tabel yang penulis buat terisi dari **product name, quantity, price, revenue, profit hingga gross profit margin(GPM)**. Untuk informasi quantity penulis ambil dari order_item_id dengan alasan penulis berasumsi bahwa dari setiap item yang di

order, maka dari adanya pemesanan barang dan jumlah barang yang harus di sediakan atau dengan kata lain bahwa data dari E-commerce ini melakukan pemesanan seperti **Purches Order**, dimana memesan terlebih dahulu baru adanya penghitungan dari quantity product yang harus di sediakan. penulis mengambil keputusan berikut dikarenakan pada dataset yang penulis gunakan tidak terdapat quantity barang, sehingga penulis mengambil keputsan tersebut. Selain itu Informasi yang penulis tampilkan yaitu quantity, dan revenue penulis gunakan berdasarkan pembuatan variable pada **Table 1.** Rumus penghitungan revenue

Dan **Table 2.** Rumus penghitungan profit, dan ada juga variable baru yang penulis buat yaitu **Gross Profit margin(GPM),** beriku penghitungan rumus dari GPM :

GPM (%) = DIVIDE([profit], [revenue])

Table 3. Rumus penghitungan Gross Profit Margin(GPM)

- **Sales by Area**

  Pada informasi Sales by Area penulis menampilkan persebaran peta dari product yang di order, Informasi tersebut penulis buat menggunakan visualisasi ArcGIS Maps yang berisikan value dari **order_item_id** dan di bandingkan dengan **customer_city** lalu di filter berdasarkan **customer_state** agar peta dari persebaran penjualan product dapat terlihat dengan jelas pada setiap daerah nya.

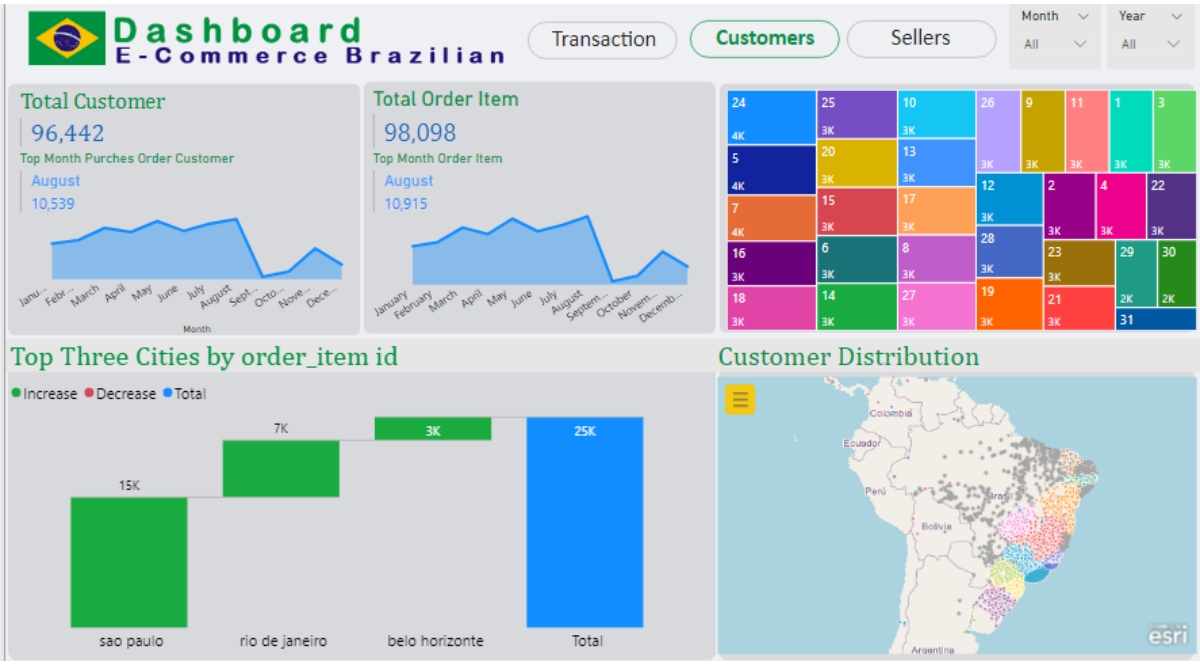
- **Contribution Products by Profit**

  Informasi yang di tampilkan pada Contribution Products by Profit penulis buat menggunakan visualisasi Bar Chart, yang berisikan value berdasarkan perbandingan dari **product_category** dengan **profit**.

- **Distribution Profit by Payment Type**

  Pada Informasi Distributin Profit by Payment Type penulis buat menggunakan visualisasi pie chart yang berisikan value dari **profit** dan di bandingkan dengan **payment_type**

B. Halaman Customer



Gambar 7. Dashboard PowerBI, Halaman Customer

Pada Halaman Customer seperti gambar 7 diatas, juga penulis membuat beberapa fitur salah satunya untuk pencarian total customer berdasarkan bulan dan tahun yang ingin di tentukan, selain itu masih ada beberapa informasi yang penulis tampilkan pada halaman dashboard Transaksi seperti yang ada pada gambar Pada Gambar 7**.** Dashboard PowerBI, Halaman Customer, yaitu berupa :

- **Total Customer**

  Informasi Total Customer yang penulis buat menggunakan visualisasi multi_row card yang menggunakan penjumlahan keseluruhan data dari variable customer_unique_id, keputusan variable yang penulis gunakan untuk pembuatan informasi tersebut dikarenakan bahwa setiap customer memiliki id yang unique, sehingga tidak ada duplikasi atau redundan pada penghitungan data. Selain itu informasi yang penulis berikan pada Total Customer yaitu berupa:

- **Grafik**

  Visualisasi yang penulis gunakan untuk pembuatan grafik yaitu menggunakan Area Chart dimana pada informasi tersebut dapat menampilkan penaikan dan penurunan dari bulan pada saat customer melakukan pemesanan pembeluan.

- **Top Month Purches Order Customer**

  Agar lebih spesifik pada visualisas grafik yang di tampilkan sebelumnya, penulis menampilkan Informasi berupa visualisasi multi-row card di mana nilai dari visualisasi tersebut di isi dari variable top_month_purches_order yang di filter berdasarkan tingkat pertama pada keseluruhan bulan pemesanan pembeluan.

- **Total Order Item**

  Informasi Total_order_item yang penulis buat menggunakan visualisasi multi_row card yang menggunakan penjumlahan keseluruhan data dari variable order_item_id, , selain itu dari visualisasi Area Chart penulis juga menggunakan order_item_id yang di bandingkan dengan variable order_approved_at dan di filter berdasarkan top 1 untuk pencarian bulan terbaik pada Total Order Item.

- **Jumlah pemesanan berdasarkan tanggal**

  Pada informasi jumlah pemesanan berdasarkan tanggal, dibuat dengan menggunakan visualisasi treemap, yang berisi dari nilai penjumlahan total customer dan di bandingkan dengan hari pada kolom order_approved_at.

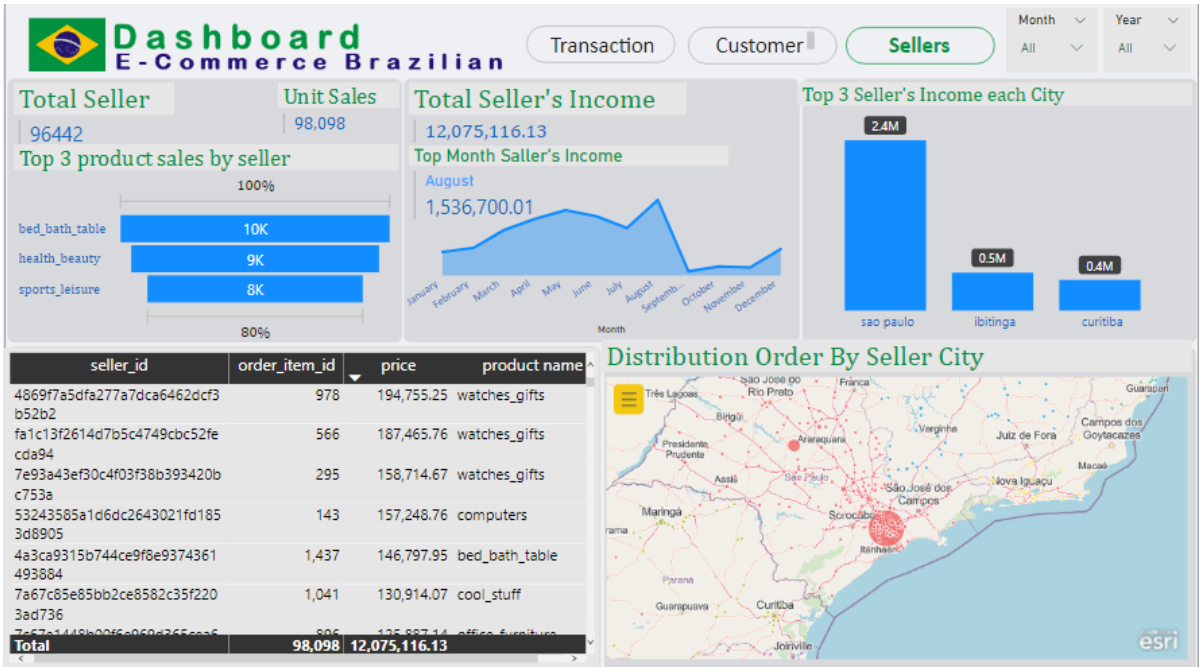
- **Top Three Cities by Order_item_id**

  Dari Informasi Top Three Cities adalah informasi yang menampilkan 3 kota teratas berdasarkan banyaknya jumlah order yang terjadi di dalam database, informasi ini dibuat menggunakan visualisasi waterfall chart yang berisi nilai dari order_item_id yang di bandingkan dengan customer_city lalu di filter berdasarkan 3 kota teratas yang memiliki banyak order.

- **Customer Distribution**

  Pada informasi Customer Distribution penulis menampilkan persebaran peta dari customer, dimana Informasi tersebut penulis buat menggunakan visualisasi ArcGIS Maps yang berisikan value dari jumlah customer lalu di cari berdasarkan customer_city menggunakan latitude dan longitude pada setiap customer.

C. Halaman Sellers



Gambar 8. Dashboard PowerBI, Halaman Sellers

Pada Halaman Seller's seperti gambar 8 diatas, juga penulis membuat beberapa fitur dihalaman ini yaitu berupa :

- **Total Seller's**

  Informasi Total Seller's yang penulis buat menggunakan visualisasi multi_row card yang menggunakan penjumlahan keseluruhan data dari variable seller_id, dari data yang ada pada kolom seller_id penulis filter terlebih dahulu agar nantinya tidak ada id yang redundan untuk di jumlahkan.

- **Unit Sales**

  Informasi Unit Sales yang penulis buat menggunakan visualisasi multi_row card dan juga Area Chart, dari visualisasi multi_row card menggunakan penjumlahan keseluruhan data dari variable order_item_id untuk menghasilkan informasi total dari item yang di order.

- **Top 3 Product sales by sellers**

  Informasi Tope 3 Product sales by sellers yang penulis buat menggunakan visualisasi funnel, dimana visualisasi tersebut menggambarkan tingkatkan dari product yang banyak

di jual oleh seller. Dari visualisasi tersebut penulis menggunakan nilai order_item_id yang di bandingkan berdasarkan product_category_name

- **Total Seller's Income**

  Informasi Total Seller's Income yang penulis buat menggunakan visualisasi multi_row card yang menggunakan penjumlahan keseluruhan data dari variable price, dimana keputusan dari penggunaan variable tersebut, penulis berasumsi bahwa price adalah harga awal yang di jual dari seller's maka dari itu E-Commerce ini menjual dengan harga lebih dari harga awal yang dijual dari sellers.

  - **Grafik**

    Pada Informasi Grafik ini penulis menampilkan Informasi tersebut berupa grafik penaikan dan penurunan dari setiap bulan yang di hasilkan oleh seller's, untuk membuat visualisasi ini menggunakan Area Chart yang berisi nilai dari price yang di bandingkan dengan bulan pada setaip order_delivered.

  - **Top Month Saller's Income**

    Untuk menghasilkan informasi yang lebih spesifik pada grafik tersebut maka dari itu penulis membuat visualisasi Top Month Saller's Income, dimana informasi ini memberikan berupa total angka Income dari bulan yang tertinggi.

- **Top 3 Seller's Income each City**

  Dari Informasi Top 3 Seller's Income each City adalah informasi yang menampilkan 3 kota teratas berdasarkan banyaknya jumlah Income yang terjadi di dalam database, informasi ini dibuat menggunakan visualisasi Line and Cluster Column atau sama dengan diagram batang, dari visualisasi ini berisi nilai dari total keseluruhan price yang di bandingkan dengan seller's_city lalu di filter berdasarkan 3 kota teratas yang memiliki banyak Income.

- **Table**

  Informasi pada Tabel yang penulis buat pada halaman Seller's terisi dari Seller_id, Order_item_id, Price, dan Product_name, penulis memilih menampilkan seller_id dikarenakan di dataset yang penulis gunakan tidak terdapat dari nama seller, maka dari itu penulis membuat keputusan untuk menampilkan kolom tersebut, dari informasi tersebut nantinya dapat di cari lebih lanjut untuk penelusuran nama dari seller berdasarkan pencarian seller_id, dengan catatan jika dataset yang digunakan sudah benar-benar lengkap atau sesuai dengan database pada E-Commerce Brazilian.

- **Distribution Order by Seller City**

  Pada informasi Distribution Order by Sellery City penulis menampilkan persebaran peta, dimana Informasi tersebut penulis buat menggunakan visualisasi ArcGIS Maps yang berisikan value dari order_item_id lalu di cari berdasarkan seller_city menggunakan latitude dan longitude pada setiap customer dan di filter berdasarkan seller_state agar informasi dapat terpapar dengan jelas dari setiap perbedaan daerah yang ada pada di peta tersebut.

Kesimpulan

Dari penjelasan pada saat proses pengambilan data hingga pembuatan informasi data dalam bentuk Visual di dalam powerBI penulis menyimpulkan, Dashboard Visual dari dataset E-Commerce Brazilian By Olist telah dibuat, dashboard tersebut dibuat guna untuk memenuhi syarat dalam pengerjaan SOAL UTS ini dan juga berguna untuk menampilkan transformasi data dari data mentah menjadi informasi yang berguna dan bermakna untuk tujuan analisis.

Dari hasil analisis yang penulis lakukan, dimana pada dataset E-Commerce Brazilian By Olist, jumlah setiap tahun seperti customer, product, seller dan order barang terus meningkat, dari hasil peningkatannya tersebut banyak terjadi penjualan di bulan **Agustus** dari tahun **2016 s/d 2018**, dan untuk tanggal yang paling unggul pada saat pemesanan yaitu pada tanggal **24 di bulan Agustus**. Dari situ penulis berasumsi bahwa ada **hari penting** di dalam bulan Agustus sehingga dapat menghasilkan jumlah penjualan yang sangat tinggi, selain itu category product yang paling laris di jual yaitu category product **bed bath table** yang berada di tingkatan pertama dengan jumlah **9,593** dari total kesuluran category_product **98,098**.