Executive Summary

Data Analyst: Kiet Tuan

Purpose: Real-world datasets are often messy, making data analysis challenging and potentially skewing results. This project aims to demonstrate my ability to clean data and perform data transformations, preparing the dataset for effective visualization about the **customer demographic** of an e-commerce company.

Project goals:

- Check data quality for the table dataset
- Perform data wrangling according to the request
- Making a dashboard for customer general demographic

Scope / Major Project Activities:

Activity	Description
1. Inspect the dataset and load to SQL database	-Assess the dataset -Choose appropriate tables -Mapping tables
2. Data cleaning	-Remove data errors -Filter and sorting -Validate data
3. Data wrangling	-Data structuring, Data integration, Data Enrichment
4. Data visualization	-Create a dashboard about customer demographic

This project does not include: This project is not responsible for collecting primary sources data (surveying, observing, testing, etc.). All the data sources used are secondary datasets.

Data sources used and assessment

In this project, I use the SQL training dataset provided by Paul Scotchford. The dataset is a backup database file type and it is about a hypothesized ecommerce company. The dataset includes various tables from Customers, Geolocation and so on. You can access the dataset here. This database includes multiple tables that are constrained by their primary keys; however, the scope of this project is about customer demographic so I only use two tables which are **Customer and Geolocation**. Upon assessing the quality, I discover that:

-The data columns are in long format

-The data contains many null values

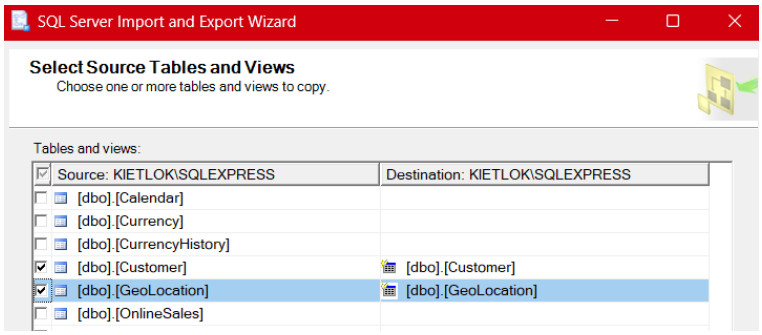
-The rest of the data is in good format

-Since it's a fictional dataset so citation, origin, reliability are not important. This dataset is useful to demonstrate my data wrangling skills and data visualization.

Load the data

After I loaded the ecommerce database to SQL Server studio, I created another database named CustomerDemo and import only Customer and Geolocation tables to the newly created database.

I copy Geolocation and Customer tables to CustomerDemo database by using SQL Import and export wizard.



Data cleaning

```
/* Create new databse named CustomerDemo and import Customer and Geolocation tables
   using SQL Wizard */
CREATE DATABASE CustomerDemo;
USE CustomerDemo;


--Remove unecessary columns
BEGIN transaction
ALTER TABLE [dbo].[Customer]
DROP COLUMN CustomerAlternatekey, Title, NameStyle, Suffix, NumberChildrenAtHome,
                    HouseOwnerFlag, Addressline1, Addressline2, DateFirstPurchase
COMMIT transaction


--Check for nulls
SELECT* FROM Customer
WHERE BirthDate is null or GeographyKey is null or MaritalStatus is null or Gender is null or
         TotalChildren is null or Occupation is null or CommuteDistance is null or
NumberCarsOwned is null or
         CustomerKey is null
         /* I observe that we have 30 null values for GeographyKey. I decided to keep these
records as they only impact the customer location map. Further directions from the
stakeholders/dba are needed */


--Check for duplicates
SELECT CustomerKey, EmailAddress, Phone, COUNT(*) AS counts
FROM [dbo].[Customer]
GROUP BY CustomerKey, EmailAddress, Phone
HAVING COUNT (*) > 1;
                            --I observe no duplicate values


--The stakeholder wants Phone number to be without 1(11), dash '-', and space ' '
BEGIN transaction
UPDATE [dbo].[Customer]
SET Phone = REPLACE(REPLACE(REPLACE (Phone, '1 (11)', ''), '-', ''), ' ', '')
FROM [dbo].[Customer]
COMMIT transaction

--Now check if the phone numbers are 10 digits
SELECT Phone FROM Customer
WHERE LEN(Phone) <> 10;
                                          --All good
```

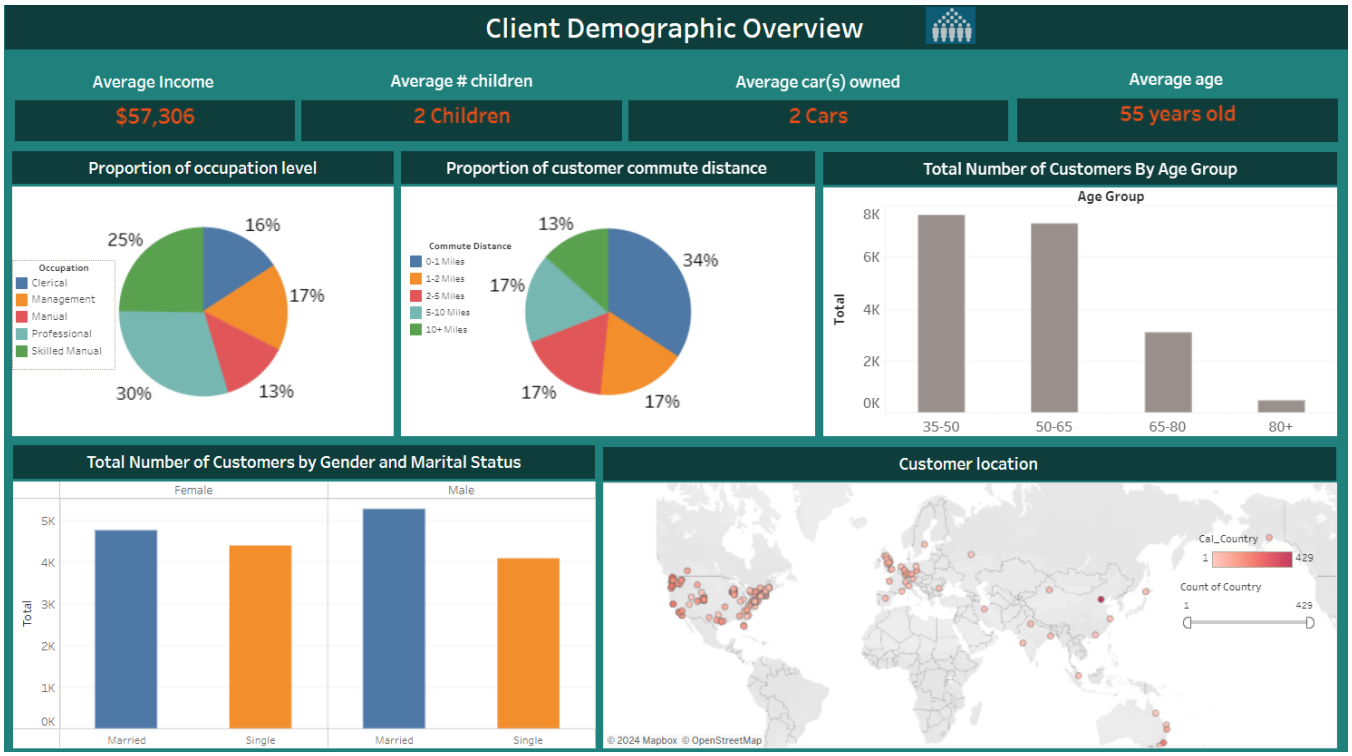
Data wrangling

```
--Add a customer Age column
BEGIN transaction
ALTER TABLE [dbo].[Customer]
ADD Age smallint
COMMIT transaction;

BEGIN transaction
UPDATE [dbo].[Customer]
SET Age = 2024 - Year(BirthDate);
COMMIT transaction;

/* Add an age classification:
                        35-50
                        50-65
                        65-80
                        80+
*/

SELECT MIN(Age) AS Min_Age, MAX(Age) AS Max_Age from customer
 --Min age is 38 and max age is 108

 BEGIN transaction
 ALTER TABLE [dbo].[Customer]
 ADD Age_Group VARCHAR(10);
 COMMIT transaction;

 BEGIN transaction
 UPDATE [dbo].[Customer]
 SET Age_Group = CASE
       WHEN Age BETWEEN 35 AND 50 THEN '35-50'
       WHEN Age BETWEEN 50 AND 65 THEN '50-65'
       WHEN Age BETWEEN 65 AND 80 THEN '65-80'
       WHEN Age > 80 THEN '80+'
END;
COMMIT transaction;

--The stakeholder wants a full name column
SELECT
       Customerkey,
       FirstName,
       MiddleName,
       LastName,
       CONCAT (Firstname, ' ', MiddleName, ' ', LastName) AS FullName

FROM[dbo].[Customer]
ORDER BY CustomerKey
```

```sql
--The stakeholder also wants to extract the name from customer email domain.
SELECT
       Customerkey,
       FirstName,
       MiddleName,
       LastName,
       EmailAddress,
       REPLACE (EmailAddress, '@awari.com.au' , ' ' ) AS Email_Name

FROM [dbo].[Customer]
ORDER BY CustomerKey

--Reference the location for each customer using Geolocation table
BEGIN transaction
ALTER TABLE [dbo].[Customer]
ADD City VARCHAR(50),
    StateProvince VARCHAR(50),
       Country VARCHAR(50)
COMMIT transaction;

BEGIN transaction
UPDATE [dbo].[Customer]
SET City = Geo.City,
       StateProvince = Geo.StateProvinceName,
       Country = Geo.CountryRegionName

FROM Customer cs INNER JOIN GeoLocation Geo ON
       cs.GeographyKey = Geo.GeographyKey;

COMMIT transaction;

--Copy the data to Excel for data visualization
SELECT
       MaritalStatus, Gender, Age_Group, Age, YearlyIncome, TotalChildren,
       NumberCarsOwned, Occupation, CommuteDistance, City, StateProvince, Country

FROM [dbo].[Customer]
ORDER BY CustomerKey
```

Data visualization

This is the dashboard I made using Tableau. You can access the dashboard here.



Observations

- Proportion of Occupation Level: A pie chart showing the distribution of clients across different occupation levels, with the highest proportion in Professional (30%) and the lowest in Manual (13%).
- Proportion of Customer Commute Distance: Another pie chart illustrating the distances clients commute, with the largest group commuting 0-1 miles (34%).
- Total Number of Customers by Age Group: A bar chart indicating that most clients fall within the 35-50 (7625) and 50-65 (7280) age groups.
- Total Number of Customers by Gender and Marital Status: A grouped bar chart revealing that married males constitute the largest group (over 5,000), followed by married females, single females, and single males.
- Customer Location: A map showcasing the geographical distribution of clients, predominantly concentrated in the United States and Europe, with a significant cluster in the US. Additionally, a noticeable number of customers are located in the East coast of Australia.