Task 1

• Data Exploration

Table_1 Content: Customers data **(Esraa)**

1-table name: Customers

2-columns name & description & datatypes:

A- Customer_id → an identifier for each customer (nvarchar).
B- customer_unique_id → a unique identifier for each customer (nvarchar).
C- customer_zip_code_prefix → The zip code prefix of the customer (float).
D- customer_city → The city where the customer is located (nvarchar).
E- customer_state → The state where the customer is located (nvarchar).

3-Column's statistics (Max, Min, mean):

→ customer_zip_code_prefix :

	max_code	min_code	avg_code
1	99990	1003	35137.4745829185

4-Tables Shape (A*B):

→ (99441*5).
→ Rows: 99441 Columns: 5.

5-# Of NULL VALUES:

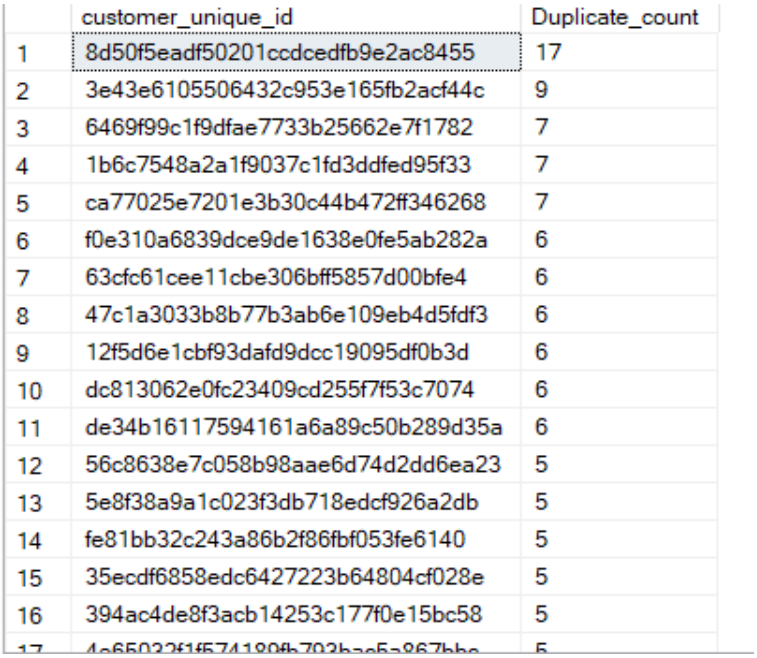
→ there are no missing values.

	missing_customer_id	missing_customer_unique_id	missing_zip_code_prefix	missing_customer_city
1	0	0	0	0

6-# Of Duplicates:

→ There are duplicates in the (customer_unique_id).

Sample:



	customer_unique_id	Duplicate_count
1	8d50f5eadf50201ccdcedfb9e2ac8455	17
2	3e43e6105506432c953e165fb2acf44c	9
3	6469f99c1f9dfae7733b25662e7f1782	7
4	1b6c7548a2a1f9037c1fd3ddfed95f33	7
5	ca77025e7201e3b30c44b472ff346268	7
6	f0e310a6839dce9de1638e0fe5ab282a	6
7	63cfc61cee11cbe306bff5857d00bfe4	6
8	47c1a3033b8b77b3ab6e109eb4d5fdf3	6
9	12f5d6e1cbf93dafd9dcc19095df0b3d	6
10	dc813062e0fc23409cd255f7f53c7074	6
11	de34b16117594161a6a89c50b289d35a	6
12	56c8638e7c058b98aae6d74d2dd6ea23	5
13	5e8f38a9a1c023f3db718edcf926a2db	5
14	fe81bb32c243a86b2f86fbf053fe6140	5
15	35ecdf6858edc6427223b64804cf028e	5
16	394ac4de8f3acb14253c177f0e15bc58	5
17	4e65022f1f574189fb702bae5a867bbe	5

→ There are no fully duplicated rows in the Table.

Table_2 Content: Geolocation data **(Ganna)**

1-table name: Geolocation

2-columns name & description & datatypes:

A- Geolocation_zip_code_prefix → represent the zip code for a specific geographic location (float).
B- Geolocation_lat → represent the latitude coordinate of the geographic location (float).
C- Geolocation_Ing → represent the longitude coordinate of geographic location (float).
D- Geolocation_city → represent the city name associated with the geographic location (nvarchar).
E- Geolocation_state → represent the state region associated with geographic location (nvarchar).

3- Column's statistics (Max, Min, mean):

- For two columns (string or nvarchar):

(We will calculate only Max and Min)

→ Geolocation_city:

	max_geo	min_geo
1	zortea	* cidade

→ Geolocation_state:

	max_geo_state	min_geo_state
1	TO	AC

- For three columns (float):

(We will calculate Max, Min & Average)

→ Geolocation_zip_code_prefix:

	max_code	min_code
1	99990	1001

	avg_code
1	36574.1664658661

→Geolocation_lat:

	max_lat	min_lat
1	45.0659331826969	-36.6053744107061

	avg_lat
1	-21.1761529103838

→Geolocation_Ing:

	max_lng	min_lng
1	121.105393810577	-101.466766449314

	avg_lng
1	-46.3905413209361

4- Tables Shape (A*B):

→(1000163*5).
→Rows: 1000163 Columns:5

5- # Of NULL VALUES:

→There are no missing values.

	column_name	missing_count
1	geolocation_zip_code_prefix	0
2	geolocation_lat	0
3	geolocation_lng	0
4	geolocation_city	0
5	geolocation_state	0

6- # Of Duplicates:

→ At this table there’re many duplicates in all columns.

Sample:

	geolocation_zip_code_prefix	geolocation_lng	geolocation_city	geolocation_state	duplicate_count
1	1001	-46.634338178054	-23.5504977069075	SP	8
2	1001	-46.6340269977783	-23.5513366552888	SP	3
3	1001	-46.6339695601014	-23.5498252739339	SP	2
4	1001	-46.6335594782337	-23.5492919999999	SP	6
5	1002	-46.6354211019966	-23.5483179780714	SP	2
6	1002	-46.6350723907899	-23.5485510657856	SP	6
7	1002	-46.6340036282243	-23.548878451007	SP	2
8	1003	-46.6371570123492	-23.5489006527786	SP	3
9	1003	-46.6361721618096	-23.5489604779272	SP	2
10	1003	-46.6351826122684	-23.5490444526804	SP	4
11	1004	-46.6353705534363	-23.5506985670263	SP	2
12	1004	-46.6353234066016	-23.5507654329736	SP	2
13	1004	-46.6348218584454	-23.5491813805181	SP	3
14	1004	-46.6343066683305	-23.5495349480928	SP	5

Table_3 Content: Order items data. **(Ganna & Esraa)**

1- table name: Order items

2- columns name & description & datatypes:

A- Order_id → represent the unique identifier for an order (nvarchar).
B- Order_item_id → represent the unique identifier for an item within an order (float).
C- Product_id → represent the unique identifier for a product (nvarchar).
D- Seller_id → represent the unique identifier for the seller of the product (nvarchar).
E- Shipping_limit_date → represent the deadline by which the item should be shipped (date).
F- Price → represent the price of the product in the order (float).
G- Freight_value → represent the shipping cost for the item (float).

3- Column's statistics (Max, Min, mean):

→order_item_id

	max_order_item_id	min_order_item_id	avg_order_item_id
1	21	1	1.19783399911229

→Price:

	max_price	min_price	avg_price
1	6735	0.85	120.653739014773

→freight_value:

	max_value	min_value	avg_value
1	409.68	0	19.9903199289859

4- Tables Shape (A*B):

→(112650*7).
→Rows: 112650 Columns:7.

5- #Of NULL VALUES:

→ There’re no missing values.

	column_name	missing_count
1	order_id	0
2	order_item_id	0
3	product_id	0
4	seller_id	0
5	shipping_limit_date	0
6	price	0
7	freight_value	0

6-# Of Duplicates:

→ at the all columns there’re no duplicates but at column (order_id) , it contains duplicates.

order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value	duplicate_count

Sample:

	order_id	duplicate_count
1	0008288aa423d2a3f00fcb17cd7d8719	2
2	00143d0f86d6fbd9f9b38ab440ac16f5	3
3	001ab0a7578dd66cd4b0a71f5b6e1e41	3
4	001d8f0e34a38c37f7dba2a37d4eba8b	2
5	002c9def9c9b951b1bec6d50753c9891	2
6	002f98c0f7efd42638ed6100ca699b42	2
7	003324c70b19a16798817b2b3640e721	2
8	00337fe25a3780b3424d9ad7c5a4b35e	2
9	003822434f91204da0a51fe4cf2aba18	2
10	003f201cdd39cdd59b6447cff2195456	2
11	005059edee63c8c708ba61910793b31b	2
12	00526a9d4ebde463baee25f386963ddc	4
13	00571ded73b3c061925584feab0db425	2
14	005d9a5423d47281ac463a968b3936fb	3