OCTOBER 2025

GB CORP DATA ASSESSMENT

MARINA SAFWAT

github repo: https://github.com/WA-on-test2/Sales_Full_Data_Project.git

//

PART A
DATA QUALITY ISSUES

1. Inconsistency
    - Mixed Upper and Lower cases
    - Mixed Languages e.g. some names are written in Arabic and others in English
    - Emails had some letters in Arabic and had random spaces
    - Formats were so various e.g. Phone no. & Dates even some dates were written in English Except for the month is in Arabic
    - The currency for example some entries had just value and others had the currency stated(Mixed Ar&En)
    - Value Representation was various e.g. Gender&Payment method & Status
2. DUPLICATE ISSUES
    - Columns that must be unique had some duplicates e.g. Customer ID
    - The governorate was repeated in Arabic, in different styles, and we had lots of redundancy that needs to be addressed, because if ignored no meaningful analysis can be done.
    - Duplicate categories with spelling variations & some were in Arabic

PART A

DATA QUALITY ISSUES

3. MISSING DATA ISSUES (NULLs)
   - Some columns were severely missing and cannot be imputed, we need to get the value from the source again for example if it is a system e.g. names, contact info like email, phone no.

4.CALCULATION/FORMULA ERRORS
   - Some column were derived from other through formula and was computed wrong

5. DATA TYPE ISSUES
- Numbers were stored as text, and that is a major problem in case we needed to perform calculations on them.

6. Lack of Standardization
- even the product SKU and the customer ID which suppose to be generated by a system non-human were also various and lacked having a standard style.

7.COMPLETENESS ISSUES
- Missing crucial fields.

WHY this happened?

I THINK THE ERRORS ARE BECAUSE OF THE LACK OF CONTROL ON THE ENTRIES AND THE HUMAN MISTAKES, SYSTEM INTEGRATION PROBLEMS, LACK OF DATA GOVERNANCE

PART A

HOW TO GET EXTRA DATA?

1.Surveys

- (About their experience, what they want to change, what qualifications does the company offer that keeps the customers loyal)

2.Frequency of their purchases

- (Asses their loyalty)

3. Extract Implicit feedback

- from the customers by collecting data for their behavior(Tracking clicks, engagement level, time spent on a page....) and build a feedback matrix that can be used for building a recommendation system.

4.Seasonal Pattern of purchases

5.Collect data about their profession

6. Wishlist or Saved Items to capture their purchase intent and may be put some discounts on those items

7.Birthday Personal celebrations or for example send 20% off on the customer birthday, keep him loyal

PART A

CONTROLS TO ADD ON THE SCREEN

1. Mandatory input (if the user doesn’t enter a field error message arises)
2. for the phone se regex for example its mandatory to write country’s code then 10 numbers only
3. for the email put regex to control the format,only English letters , the position of the @, the domain all of these are controls
4. If someone enter in the gender “انثى” arise error
5. may be even put as much as we can of menus and choices between F, M (Drop down Lists)
6. Reduce the calculations done by human ,Let the system calculates
7. Ask the user for confirmation before submitting

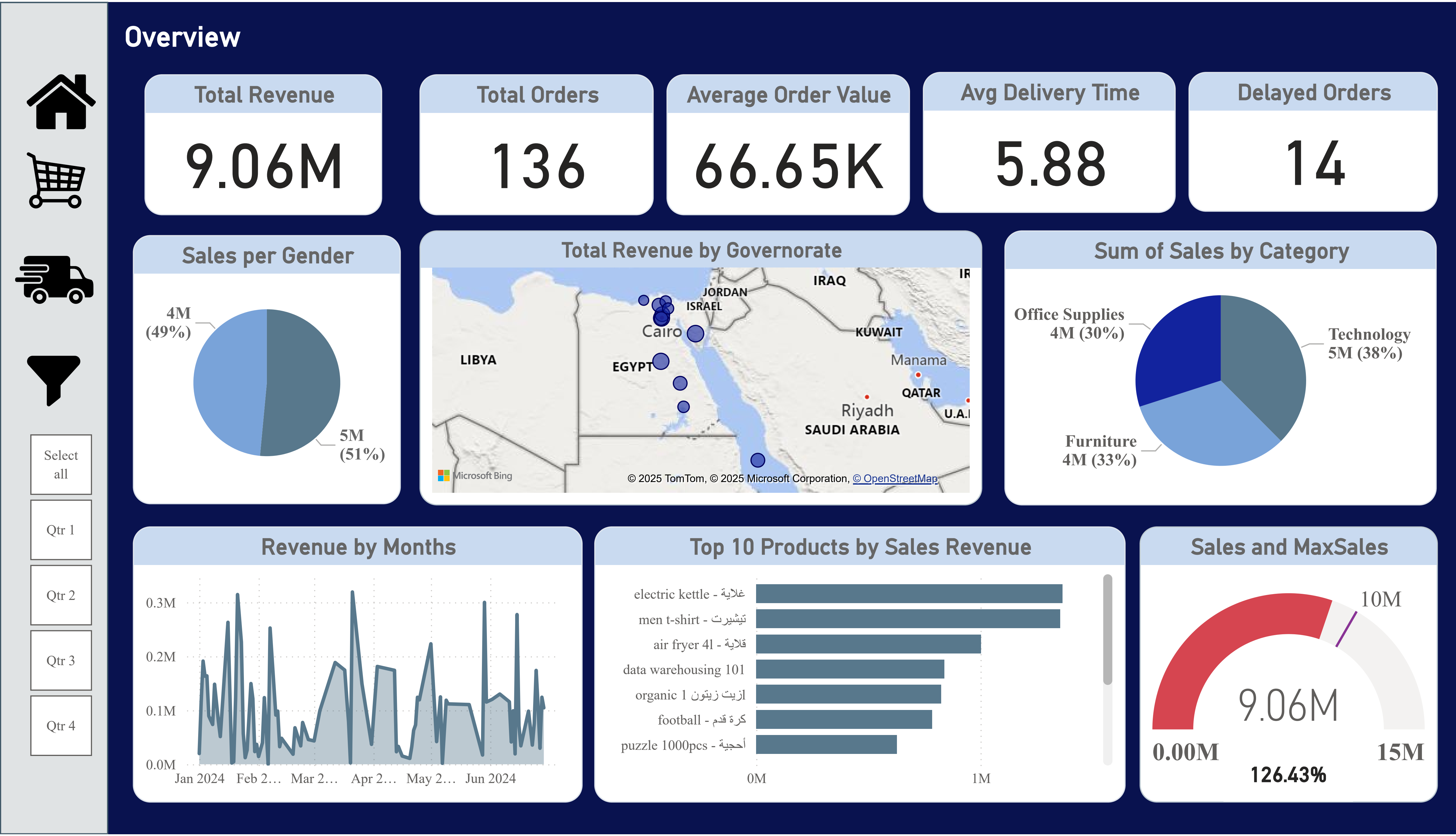
PART B

ALL THE INSIGHTS REQUIRED IN THIS PART ARE IN THE DASHBOARD

SALES
DASHBOARD

OVERVIEW

SHIPPING



Overview

Total Revenue 9.06M

Total Orders 136

Average Order Value 66.65K

Avg Delivery Time 5.88

Delayed Orders 14

Shipping

Select all

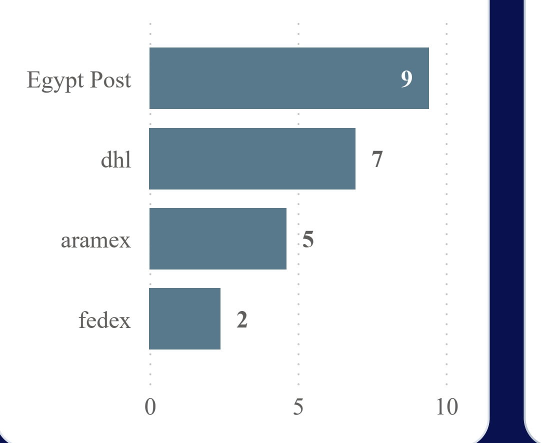
Qtr 1

Qtr 2

Qtr 3

Qtr 4

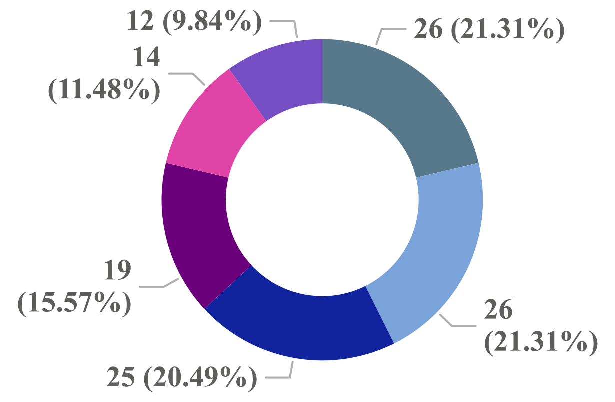
Avg Shipping duration



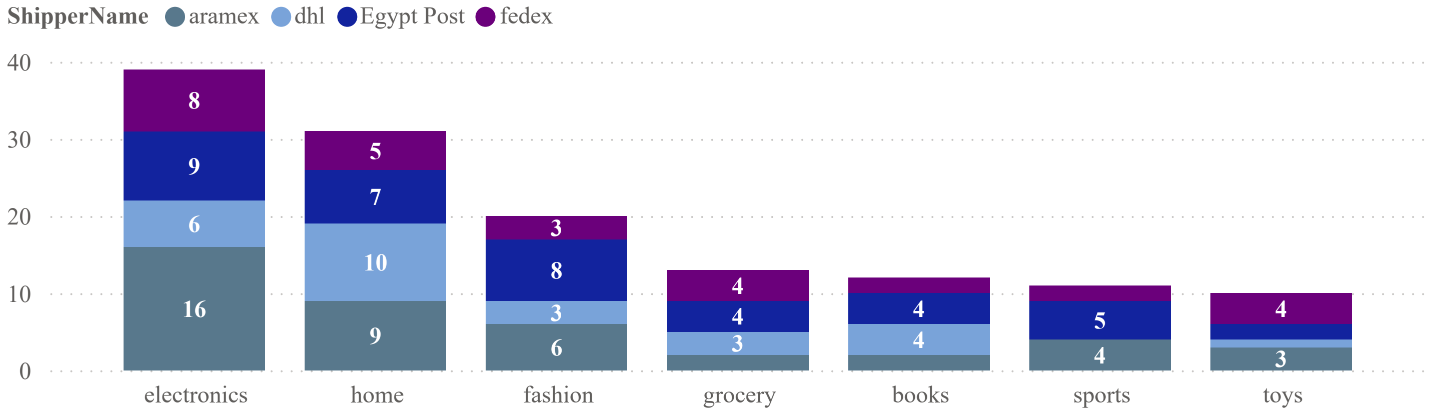
Delivery Performance by Location

IRAQ
JORDAN
ISRAEL
Cairo
KUWAIT
LIBYA
EGYPT
Manama
QATAR
Riyadh
SAUDI ARABIA
U.A.
Microsoft Bing
© 2025 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap

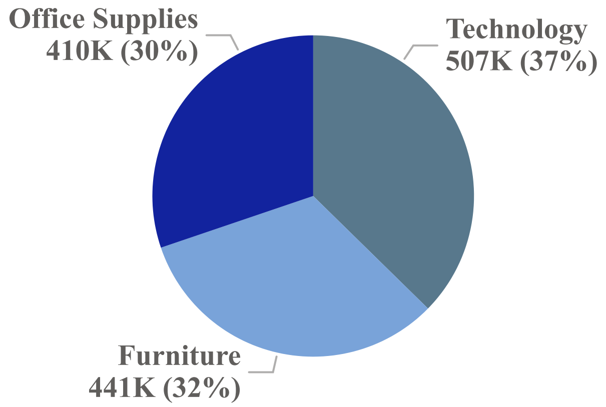
Delivery Performance



Orders Distribution Across Categories and Shipping Companies



Shipping by Category



Select all

Qtr 1

Qtr 2

Qtr 3

Qtr 4

Payment & Risk Analysis

Unpaid Orders - Revenue at Risk

CustomerName	Month	Sum of TotalAmount	PaymentStatus	Pr
ahmed ali	April	33,111.54	unpaid	air
ahmed ali	May	75.00	unpaid	org
ahmed fathi		31,041.74	unpaid	pu
ahmed ibrahim	March	163,504.61	unpaid	foo
ahmed mahmoud	March	62,163.35	unpaid	org
ahmed nasr	February	32,584.31	unpaid	foo
Total		**3,903,271.18**		

Underperforming Products

ProductName	Category	Total Orders	TotalRevenue	Average Order Va
labtop i7 16gb - لابتوب	electronics	3	208,691.61	69,563
bluetooth سماعة	electronics	9	480,671.31	53,407
laptop i7 16gb - لابتوب	electronics	9	486,297.22	54,033
smartphone a55 - هاتف	electronics	11	530,021.51	48,183
smartphone a55 - تليفون	electronics	7	561,258.58	80,179
puzzle 1000pcs - أحجية	toys	10	627,715.66	62,771
Total		**136**	**9,064,869.06**	**66,653**

Unpaid Orders %

64

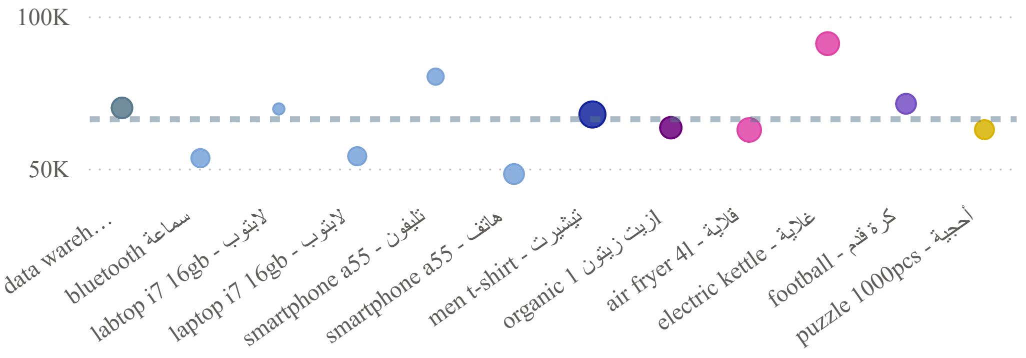
Revenue at Risk

5.97M

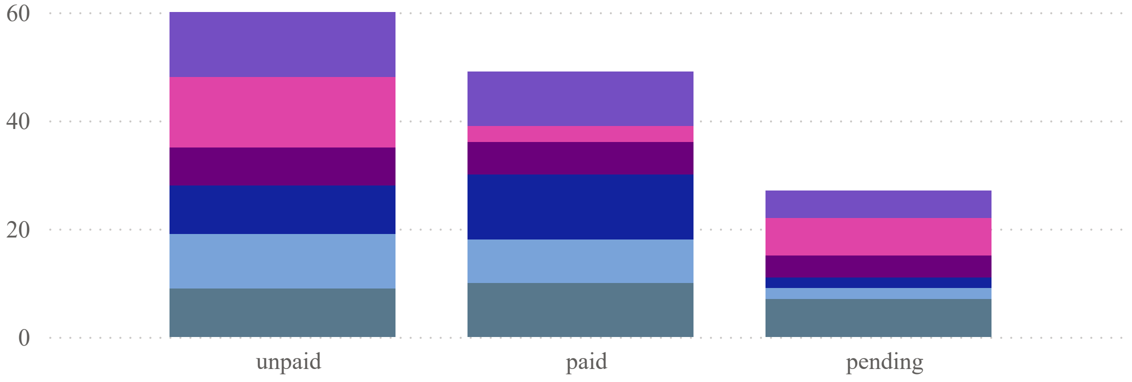
Avg Order Value

66.65K

Products Above/Below Average AOV



Payment Status Analysis



PART C

DATA QUALITY

Check and fix data quality issues based on the DAMA dimensions (accuracy, completeness, etc.). Produce three key items: a DQ report, a cleansing guide (step-by-step specs), and a working script or SQL.

- The report is in the next page
- both of the cleaning and script are in the repo

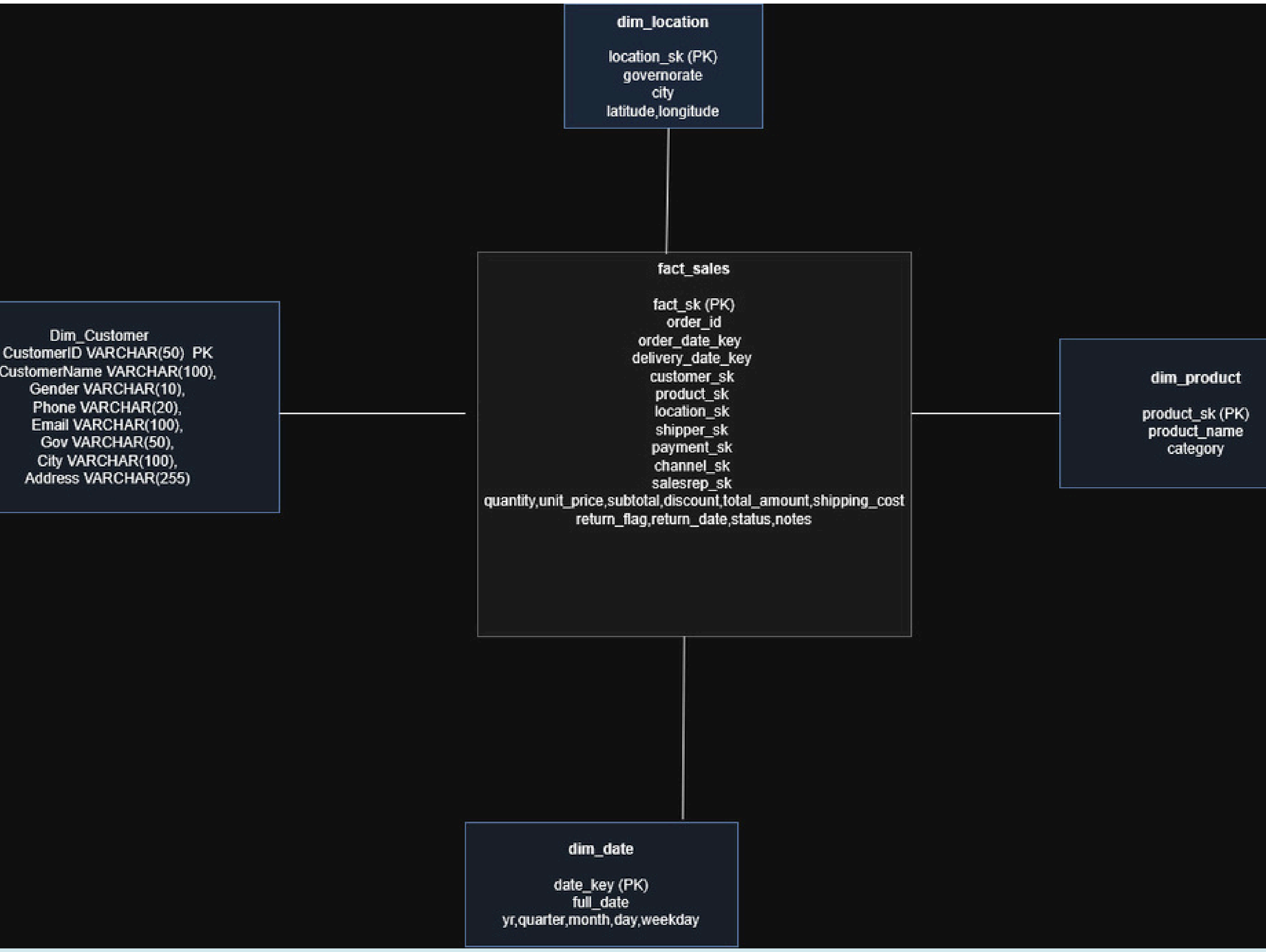
1) DATA QUALITY REPORT

Data Quality Violation	How did it show in the dataset
1.Accuracy	TotalAmount(Wrong calculations) UnitPrice (wrong prices) Quantity Data entries are in mixed languages - Decrease the accuracy of the entries as it losses its value and become not readable
2.Completeness	This dataset had lots of missing values that were crucial and cannot be imputed, how would you impute customerID unless you go back to the system (tried using the customer sheet but it didn’t fill all the Nulls in the Sales_Order table regarding the customer ID ProductSKU , Email (Lots of Nulls), Phone DeliveryDate (Nulls), Category (but could be imputed), Address ProductDescription , ShippingCost
3.Consistency	data was very far away from being consistent in the same column entries vary a lot their is an obvious example in the category column ,("الكترونيات" variations: "Electronics" vs "Electronic" vs 50%) Phone (+20, 012, 010"), CustomerName (capitalization:"), PaymentStatus Nasr City" vs "Nasr", Email , Channel Status
4.Validity	The data isn’t even valid Email(most of the email either wrong fromate or filler test.test) Phone , CustomerID , ProductSKU, DeliveryDate some of them are before OrderDate , Quantity (negative values), UnitPrice (zero/negative), Discount (exceeds subtotal)
5.Uniqueness	There are many duplicates even in the Customer ID,some personal info is duplicated, CustomerName (duplicate variations: "Ahmed Ali" vs "Ahmad Aly"), ProductName (English/Arabic duplicates), ProductSKU (duplicate formats), Category, OrderID
6.Timeliness	Data is not even valid for the moment to be timeliness but it doesn’t satisfy this constraint and some delivery date is earlier than order date

PART D

1) THE DATA MODEL AND DESIGN

I UPLOADED IT CLEARER ON THE GITHUB REPO



dim_location

location_sk (PK)
governorate
city
latitude,longitude

fact_sales

fact_sk (PK)
order_id
order_date_key
delivery_date_key
customer_sk
product_sk
location_sk
shipper_sk
payment_sk
channel_sk
salesrep_sk
quantity,unit_price,subtotal,discount,total_amount,shipping_cost
return_flag,return_date,status,notes

Dim_Customer
CustomerID VARCHAR(50) PK
CustomerName VARCHAR(100),
Gender VARCHAR(10),
Phone VARCHAR(20),
Email VARCHAR(100),
Gov VARCHAR(50),
City VARCHAR(100),
Address VARCHAR(255)

dim_product

product_sk (PK)
product_name
category

dim_date

date_key (PK)
full_date
yr,quarter,month,day,weekday

//

PART D

THE CODES ETL& THE MODELLING IN THE WAREHOUSE ARE ON GITHUB

Thank you