```
--Data Profiling with SQL

--1. Summary Statistics:

--Count of Rows:


SELECT COUNT(*) FROM Customer;  --250

SELECT COUNT(*) FROM Order_;    --250

SELECT COUNT(*) FROM Shipping;  --250


--Count of Distinct Values:


SELECT COUNT(DISTINCT Customer_ID) FROM Customer;  --250

SELECT COUNT(DISTINCT Order_ID) FROM ORDER_;      --250

SELECT COUNT(DISTINCT Shipping_ID) FROM Shipping;   --250


--Basic Descriptive Stats:

--MIN, MAX:


SELECT MIN(Age), MAX(Age) FROM Customer;          --Min=18       Max=80 --No date column like Dataofbirth

SELECT MIN(AMOUNT), MAX(AMOUNT) FROM Order_;       -- No OrderDate  Min= 200 Max=12000
```

```
SELECT MIN(Shipping_ID), MAX(Shipping_ID) FROM Shipping;    -- Min=1  Max=250


--Average Order Amount:

SELECT AVG(Amount) FROM ORDER_;        --Avg=2130


--2. Data Quality Checks:

--Null Values Check:


SELECT COUNT(*) FROM Customer WHERE CUSTOMER_ID IS NULL
                OR FIRST IS NULL OR LAST IS NULL OR AGE IS NULL OR COUNTRY IS NULL;   -- 0 so no nulls
SELECT COUNT(*) FROM ORDER_ WHERE ORDER_ID IS NULL
                OR Customer_ID IS NULL OR ITEM IS NULL OR AMOUNT IS NULL ;          -- 0 so no nulls
SELECT COUNT(*) FROM Shipping WHERE Status IS NULL
                OR Customer_ID IS NULL OR SHIPPING_ID IS NULL;                  -- 0 so no nulls
--Duplicates Detection:


SELECT Customer_ID, COUNT(*) FROM Customer GROUP BY Customer_ID HAVING COUNT(*) > 1;         -- Query produced no
results, meaning no duplicates
```

```
SELECT Order_ID, COUNT(*) FROM ORDER_ GROUP BY Order_ID HAVING COUNT(*) > 1;              -- Query produced no results, meaning no duplicates

SELECT Shipping_ID, COUNT(*) FROM Shipping GROUP BY Shipping_ID HAVING COUNT(*) > 1;        -- Query produced no results, meaning no duplicates


--Range Validation:


--SELECT Date_of_Birth FROM Customer

        --WHERE Date_of_Birth < '1900-01-01' OR Date_of_Birth > CURDATE(); -- this is a general check as we use 1900-01-01 as a default vealue

SELECT Amount FROM ORDER_ WHERE Amount < 0;   -- query produced no results


--3. Data Type and Format Validation:

--Email Format Validation:


--SELECT Email FROM Customer WHERE Email NOT LIKE '%@%.%';


--4. Distribution of Values:

--Frequency Distribution:
```

```
--SELECT Gender, COUNT(*) FROM Customer GROUP BY Gender;

SELECT Status, COUNT(*) FROM Shipping GROUP BY Status ORDER BY COUNT(*) DESC; --Pending=150 Delivered=100


--Percentiles (e.g., Order Amount):

 SELECT PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY Amount) AS median_order_amount FROM ORDER_;   --400



--5. Referential Integrity Checks:

--Foreign Key Consistency (Customer_ID in Orders and Shipping):


SELECT COUNT(*) FROM ORDER_ WHERE Customer_ID NOT IN (SELECT Customer_ID FROM Customer);    --0 so no issues

SELECT COUNT(*) FROM Shipping WHERE Customer_ID NOT IN (SELECT Customer_ID FROM Customer);   --0 so no issues


--6. Data Coverage and Completeness:

--Check for Incomplete Customer Profiles:

SELECT COUNT(*) FROM Customer WHERE CUSTOMER_ID IS NULL

                OR FIRST IS NULL OR LAST IS NULL OR AGE IS NULL OR COUNTRY IS NULL;   -- 0 so no nulls

SELECT * FROM Customer WHERE TRIM(CUSTOMER_ID) = ''

                        OR CUSTOMER_ID LIKE '%[^a-zA-Z0-9 ]%'  -- checks for Special blanks/spaces/character
```