coursera

Guided Project

# Data Cleaning in Snowflake: Techniques to Clean Messy Data

Estimated Time

120 minutes

**Instructor:**

**Mohamed (Mo) Touiti**

## How Guided Projects work

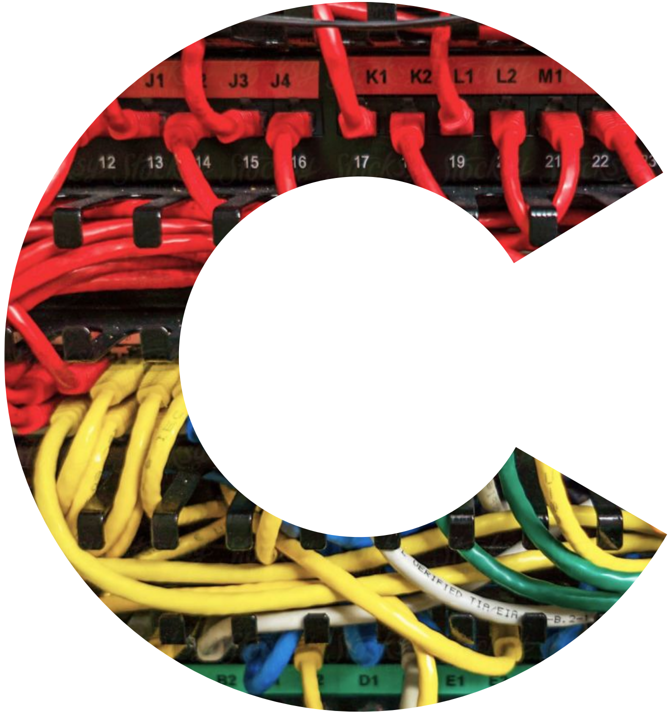
Your workspace is a cloud desktop right in your browser, no download required

In a split-screen video, your instructor guides you step-by-step

# Scenario

For the next marketing campaign, you have been assigned to find the list of inactive customers (didn't make any transactions in the last 90 days). But, like in real life, available customers' data has several challenges like duplicated customers, missing emails, merged columns, non standardized phone numbers and wrong data types .. In addition, additional fields need to be calculated.

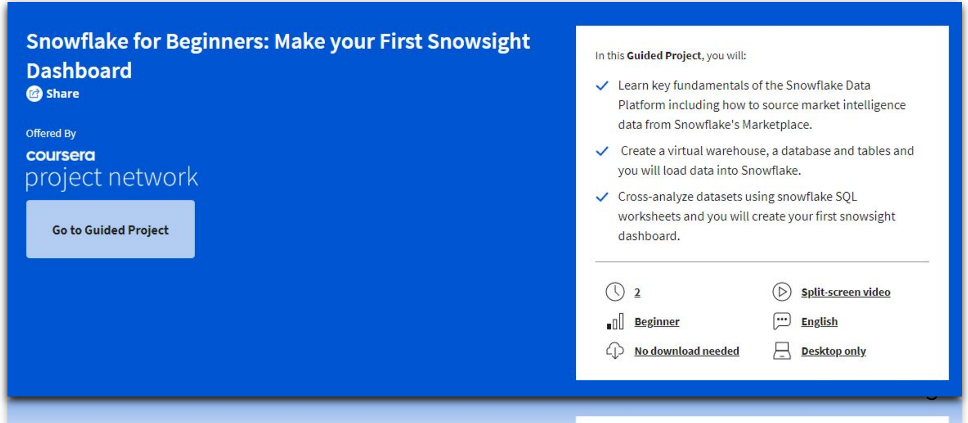
Your task is to reformat and clean data using SQL functions In Snowflake before you can, eventually, find the target list of customers.



# Project Goal

This Project introduces you to one of the most essential skills of any Data Analyst/Data Engineer - Data Preparation and Cleansing.

Throughout a real-life example, you will learn about different forms of messy data and different SQL techniques In Snowflake to solve them.



**Snowflake for Beginners: Make your First Snowsight Dashboard**

Share

Offered By

coursera project network

Go to Guided Project

In this **Guided Project**, you will:

- ✓ Learn key fundamentals of the Snowflake Data Platform including how to source market intelligence data from Snowflake's Marketplace.
- ✓ Create a virtual warehouse, a database and tables and you will load data into Snowflake.
- ✓ Cross-analyze datasets using snowflake SQL worksheets and you will create your first snowsight dashboard.

2

Split-screen video

Beginner

English

No download needed

Desktop only

# Tasks Summary

S STRUCTURE

SQL String functions to remove unwanted characters and split rows to multiple columns.

C CONFORMITY

Extract dates from Text fields then use SQL date functions for comparisons and calculations.

D DUPLICATIONS

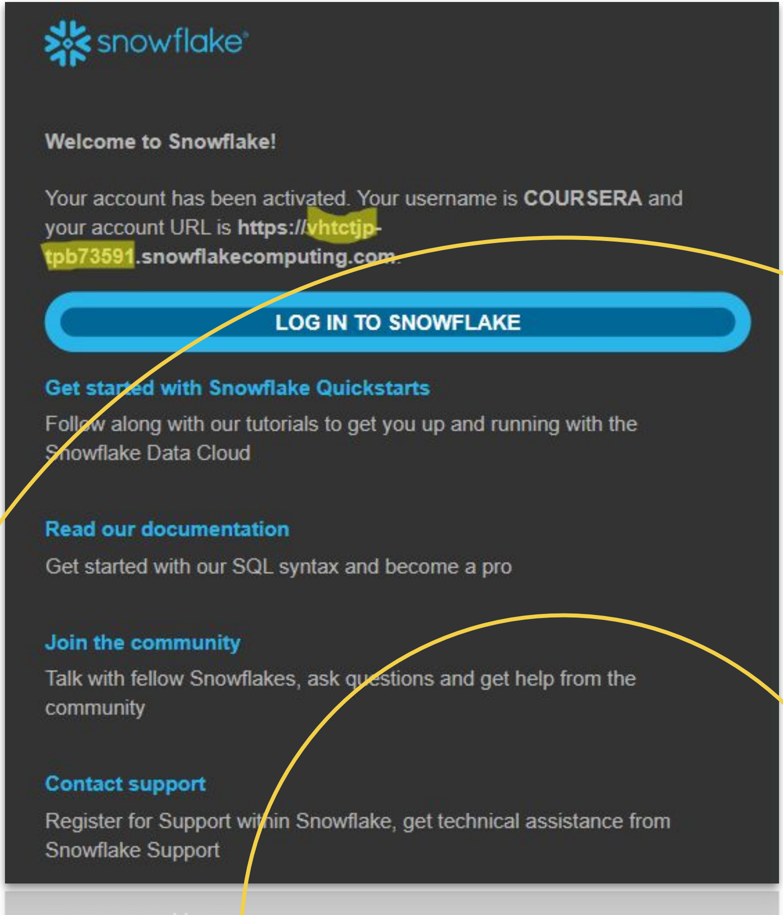
Identify and correct missing and duplicated data.

S SCALE

Build a View to scale the work, then query data to find list of inactive customers (didn't transact in the last 90 days).

# Task 1

## Load Project Data



**Welcome to Snowflake!**

Your account has been activated. Your username is **COURSERA** and your account URL is **https://vhtctjp-tpb73591.snowflakecomputing.com**.

**Get started with Snowflake Quickstarts**

Follow along with our tutorials to get you up and running with the Snowflake Data Cloud

**Read our documentation**

Get started with our SQL syntax and become a pro

**Join the community**

Talk with fellow Snowflakes, ask questions and get help from the community

**Contact support**

Register for Support within Snowflake, get technical assistance from Snowflake Support

# Task Summary

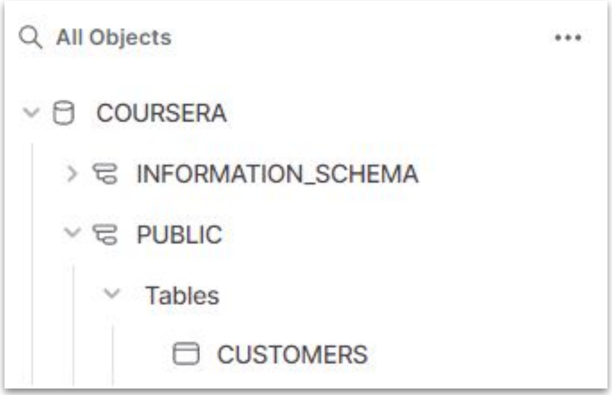
## Task #1

### Task Goal

Load Project Data

### Key Takeaways

- There are several ways to load data into a snowflake table, we used INSERT statement in this project.
- Always make sure to set up the context of your worksheet (Role, Virtual Warehouse, Database, Schema)



# Task 2

## Investigate data quality issues

- □ SELECT … RANDOM() .. LIMIT
- □ Automatic Contextual Statistics in Snowflake UI

# Task Summary

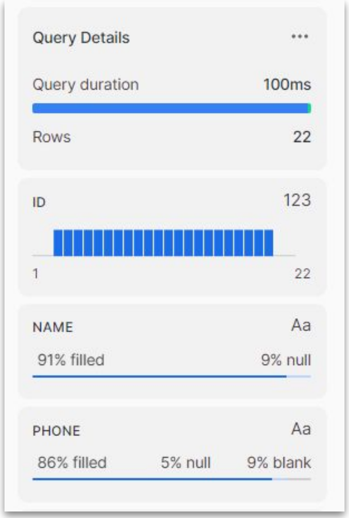
## Task #2

### Task Goal

Investigate Data Quality Issues

### Key Takeaways

- Tip : Always think first to to analyse visually small samples of the dataset to detect quality issues.
- The function RANDOM() generates a random value for each row in the table. The ORDER BY clause sorts all rows in the table by the random number generated by the RANDOM() function.
- The LIMIT ‘n’ clause picks the first ‘n’ row in the result set sorted randomly.
- Use the Contextual Statistics in Snowflake UI to investigate your dataset.



Query Details

Query duration 100ms

Rows 22

ID 123

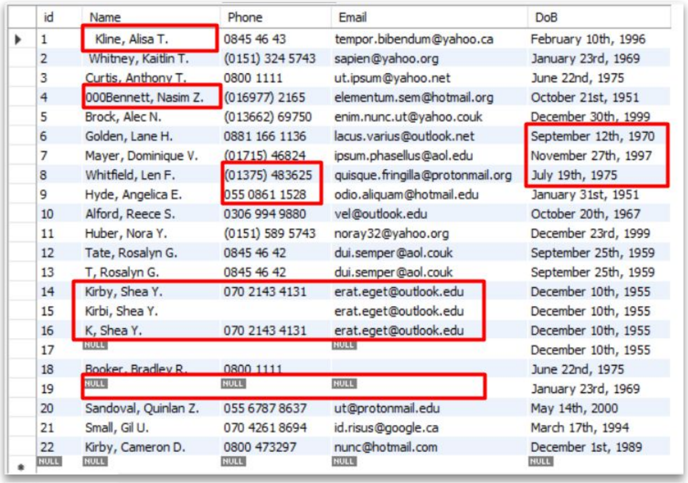
1 22

NAME Aa

91% filled 9% null

PHONE Aa

86% filled 5% null 9% blank



| id | Name | Phone | Email | DoB |
|---|---|---|---|---|
| 1 | Kline, Alisa T. | 0845 46 43 | tempor.bibendum@yahoo.ca | February 10th, 1996 |
| 2 | Whitney, Kaitlin T. | (0151) 324 5743 | sapien@yahoo.org | January 23rd, 1969 |
| 3 | Curtis, Anthony T. | 0800 1111 | ut.ipsum@yahoo.net | June 22nd, 1975 |
| 4 | 000Bennett, Nasim Z. | (016977) 2165 | elementum.sem@hotmail.org | October 21st, 1951 |
| 5 | Brock, Alec N. | (013662) 69750 | enim.nunc.ut@yahoo.couk | December 30th, 1999 |
| 6 | Golden, Lane H. | 0881 166 1136 | lacus.varius@outlook.net | September 12th, 1970 |
| 7 | Mayer, Dominique V. | (01715) 46824 | ipsum.phasellus@aol.edu | November 27th, 1997 |
| 8 | Whitfield, Len F. | (01375) 483625 | quisque.fringilla@protonmail.org | July 19th, 1975 |
| 9 | Hyde, Angelica E. | 055 0861 1528 | odio.aliquam@hotmail.edu | January 31st, 1951 |
| 10 | Alford, Reece S. | 0306 994 9880 | vel@outlook.edu | October 20th, 1967 |
| 11 | Huber, Nora Y. | (0151) 589 5743 | noray32@yahoo.org | December 23rd, 1999 |
| 12 | Tate, Rosalyn G. | 0845 46 42 | dui.semper@aol.couk | September 25th, 1959 |
| 13 | T, Rosalyn G. | 0845 46 42 | dui.semper@aol.couk | September 25th, 1959 |
| 14 | Kirby, Shea Y. | 070 2143 4131 | erat.eget@outlook.edu | December 10th, 1955 |
| 15 | Kirbi, Shea Y. | | erat.eget@outlook.edu | December 10th, 1955 |
| 16 | K, Shea Y. | 070 2143 4131 | erat.eget@outlook.edu | December 10th, 1955 |
| 17 | NULL | | NULL | December 10th, 1955 |
| 18 | Booker, Bradley R. | 0800 1111 | | June 22nd, 1975 |
| 19 | NULL | NULL | NULL | January 23rd, 1969 |
| 20 | Sandoval, Quinlan Z. | 055 6787 8637 | ut@protonmail.edu | May 14th, 2000 |
| 21 | Small, Gil U. | 070 4261 8694 | id.risus@google.ca | March 17th, 1994 |
| 22 | Kirby, Cameron D. | 0800 473297 | nunc@hotmail.com | December 1st, 1989 |
| NULL | NULL | NULL | NULL | NULL |

# Task 3

## Remove unwanted characters

- □ TRIM ()
- □ CONCAT ()

# Task Summary

## Task #3

### Task Goal

Remove unwanted characters

### Key Takeaways

- Start with CONCAT() since spaces are hard to visually spot.
- In Arguments you can define  one or more characters to remove from the left and right side of expression
- The default value is SPACE, i.e. if no characters are specified, all leading and trailing blank spaces are removed.
- TRIM will remove Leading and Trailing characters, if you want to remove only leading characters use LTRIM and Trailing characters use RTRIM.

# Task 4

## Extract First and Last Names

❑ SPLIT_PART()

# Task Summary

## Task #4

### Task Goal

Extract First and Last Names

### Key Takeaways

- When using the function SPLIT_PART(), if the count parameter is positive, everything to the left of the final delimiter (counting from the left) is returned.
- If count is negative, everything to the right of the final delimiter (counting from the right) is returned.
- SPLIT_PART() is 1-based → 0 is treated as 1.

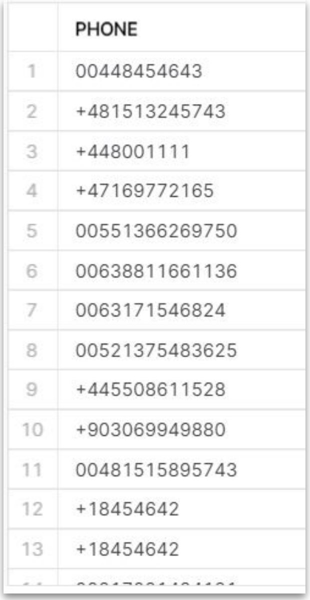
GUIDED PROJECT

# Practice Activity

This task is optional and ungraded. The goal is to check your understanding.

# Practice Task

## Standardize Phone column



|  | PHONE |
|---|---|
| 1 | 00448454643 |
| 2 | +481513245743 |
| 3 | +448001111 |
| 4 | +47169772165 |
| 5 | 00551366269750 |
| 6 | 00638811661136 |
| 7 | 0063171546824 |
| 8 | 00521375483625 |
| 9 | +445508611528 |
| 10 | +903069949880 |
| 11 | 00481515895743 |
| 12 | +18454642 |
| 13 | +18454642 |

### Things to Note

Import SQL Worksheet Practice Task 1 from Project files.

In this practice use LTRIM function to remove Zeros as well as Plus Sign from the left side.

### Pro Tip

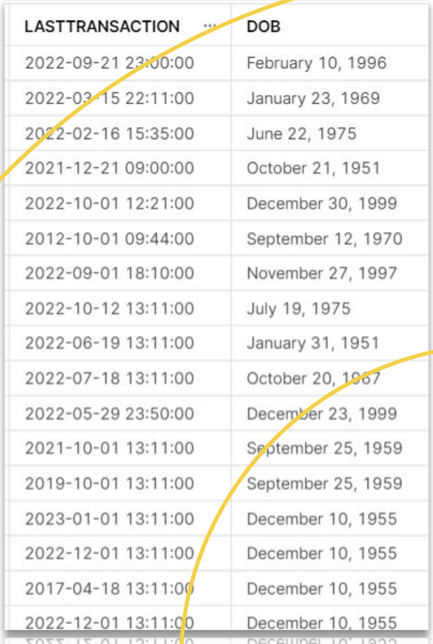
Always use documentation to read examples

*(Pause the video to complete the task and unpause to see the solution once the task is complete)*

# Task 5

## Extract date from text

- TO_DATE()



| LASTTRANSACTION | DOB |
|---|---|
| 2022-09-21 23:00:00 | February 10, 1996 |
| 2022-03-15 22:11:00 | January 23, 1969 |
| 2022-02-16 15:35:00 | June 22, 1975 |
| 2021-12-21 09:00:00 | October 21, 1951 |
| 2022-10-01 12:21:00 | December 30, 1999 |
| 2012-10-01 09:44:00 | September 12, 1970 |
| 2022-09-01 18:10:00 | November 27, 1997 |
| 2022-10-12 13:11:00 | July 19, 1975 |
| 2022-06-19 13:11:00 | January 31, 1951 |
| 2022-07-18 13:11:00 | October 20, 1967 |
| 2022-05-29 23:50:00 | December 23, 1999 |
| 2021-10-01 13:11:00 | September 25, 1959 |
| 2019-10-01 13:11:00 | September 25, 1959 |
| 2023-01-01 13:11:00 | December 10, 1955 |
| 2022-12-01 13:11:00 | December 10, 1955 |
| 2017-04-18 13:11:00 | December 10, 1955 |
| 2022-12-01 13:11:00 | December 10, 1955 |

# Task Summary

## Task #5

### Task Goal

Extract date from text

### Key Takeaways

- `YYYY-MM-DD` is the ISO Date Format .
- Always check documentation to find format Elements.

# Task 6

## Add new computed column

## “Days Since Last Transaction”

❑ DATEDIFF()

❑ CURRENT_DATE()

# Task Summary

## Task #6

### Task Goal

Add new calculated Column “Days Since Last Transaction”

### Key Takeaways

- Using the Datediff() function calculate number of days between LastTransaction and Current Date.
- <date_or_time_part> can be in days, months years …

GUIDED PROJECT

# Practice Activity

This task is optional and ungraded. The goal is to check your understanding.

# Practice Task

## Calculate Customers Age

### Things to Note

Use DATEDIFF Function
Combine all what you have learned to write one
Select Statement which shows Customer details.

### Pro Tip

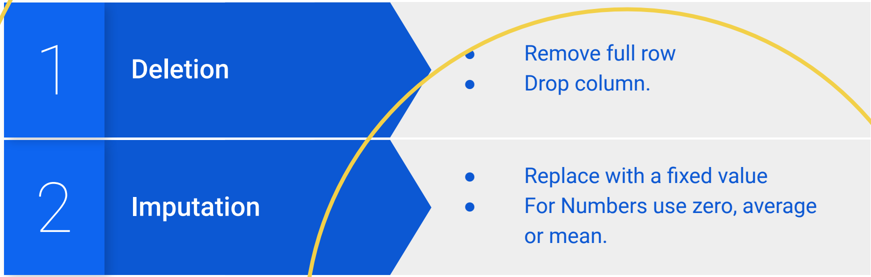
Instead of days use Years.

*(Pause the video to complete the task and unpause to see the solution once the task is complete)*

# Task 7

## Deal with missing values

- □ IS NULL
- □ IFF()



1 Deletion

- Remove full row
- Drop column.

2 Imputation

- Replace with a fixed value
- For Numbers use zero, average or mean.

# Task Summary

## Task #7

### Task Goal

Deal with Missing Values

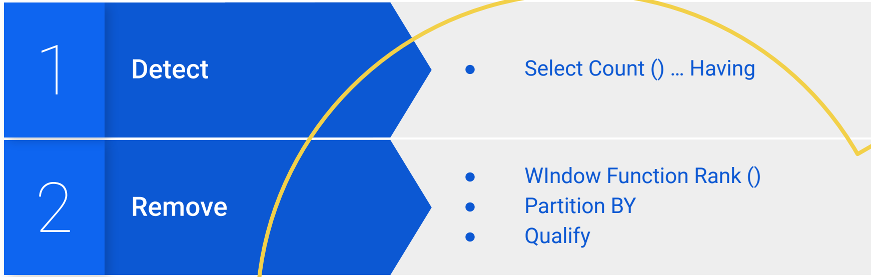
### Key Takeaways

- Missing values can be either Null values or empty fields (Blank).
- There are 2 main strategies when tackling missing values :
  - Imputation (replace it with other values)
  - Deletion (remove the entire column or row).
- **Warning**: When deleting an entire row or column be careful … you might lose some useful data from the dataset.

# Task 8

## Eliminate duplications

- ❑ SELECT COUNT() .. HAVING
- ❑ RANK()
- ❑ Window Function with PARTITION BY Clause
- ❑ QUALIFY



1 Detect
- Select Count () … Having

2 Remove
- WIndow Function Rank ()
- Partition BY
- Qualify

# Task Summary

## Task #8

### Task Goal

Eliminate duplications

### Key Takeaways

- When dealing with duplicate rows, think first of which columns to use as identifier.
- Take a decision on which rows to keep.

# Task 9

## Export list of inactive customers

- □ CREATE VIEW
- □ SELECT

# Task 9

## Export list of inactive customers

- ❑ CREATE VIEW
- ❑ SELECT

# Task Summary

## Task #9

### Task Goal

Export list of inactive customers

### Key Takeaways

- Build View to scale your work and reuse the same data for future analysis.
- Querying views is more simple and easy.

# Congratulations !!

GUIDED PROJECT

# Cumulative Challenge

This challenge is optional and ungraded.
The goal is to build your confidence.

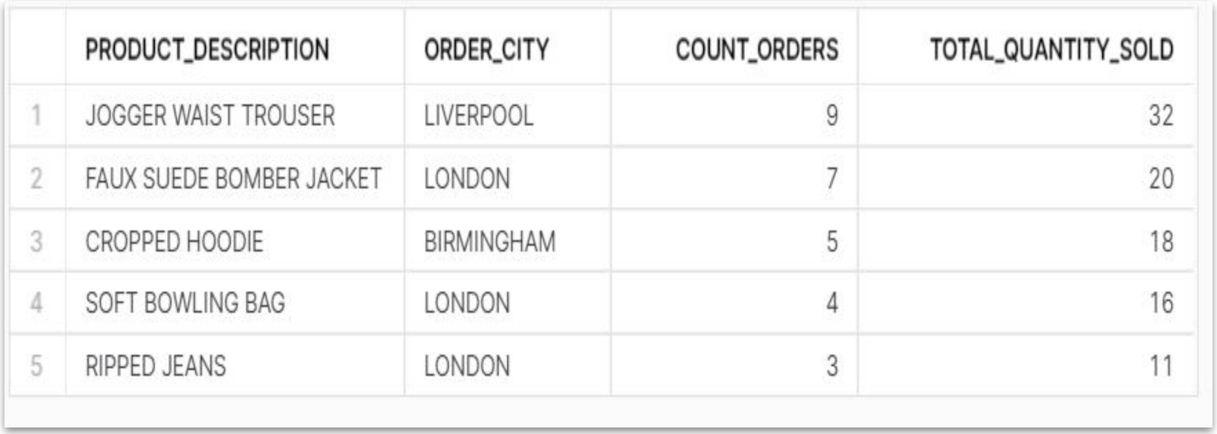
# Scenario/ Challenge

**Find the list of Top 5 Sold Products per City during Jan 2023**

## Your Task

After the great work you did finding the list of Inactive Customers, your manager has asked you to use the Orders dataset to find Top 5 Sold Products per City during the month of January 2023.

Expected Outcome →



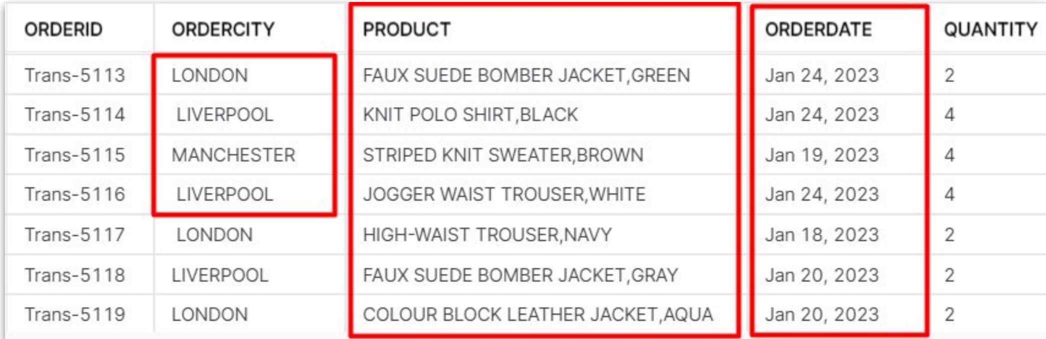
| | PRODUCT_DESCRIPTION | ORDER_CITY | COUNT_ORDERS | TOTAL_QUANTITY_SOLD |
|---|---|---|---|---|
| 1 | JOGGER WAIST TROUSER | LIVERPOOL | 9 | 32 |
| 2 | FAUX SUEDE BOMBER JACKET | LONDON | 7 | 20 |
| 3 | CROPPED HOODIE | BIRMINGHAM | 5 | 18 |
| 4 | SOFT BOWLING BAG | LONDON | 4 | 16 |
| 5 | RIPPED JEANS | LONDON | 3 | 11 |

# Scenario/ Challenge

## Find the list of Top 5 Sold Products per City during Jan 2023

## Steps

1. Import worksheet “Cumulative Challenge.sql” into your Snowflake Workspace.
2. Run the script to create new table ORDERS then Load the ORDERS dataset.
3. Write a Select * statement to print the dataset, then Investigate data quality issues.
4. Write a new Select statement to extract the following columns :
   a. PRODUCT_DESCRIPTION: Extract product description without color from PRODUCT using SPLIT_PART
   b. ORDER_CITY: use TRIM to remove Leading Spaces from ORDERCITY
   c. ORDER_DATE: convert ORDERDATE into a Date Column using TO_DATE
   d. ORDERID and QUANTITY : keep these columns as is.
5. Modify previous query to add :
   a. Filter dates Between '2023-01-01' and '2023-01-31'
   b. GROUP BY PRODUCT_DESCRIPTION, ORDER_CITY
   c. COUNT_ORDERS = COUNT(ORDERID)
   d. TOTAL_QUANTITY_SOLD= SUM(QUANTITY)
   e. ORDER BY TOTAL_QUANTITY_SOLD DESC
   f. LIMIT result to 5
6. Extract the result as CSV.



| ORDERID | ORDERCITY | PRODUCT | ORDERDATE | QUANTITY |
|---|---|---|---|---|
| Trans-5113 | LONDON | FAUX SUEDE BOMBER JACKET,GREEN | Jan 24, 2023 | 2 |
| Trans-5114 | LIVERPOOL | KNIT POLO SHIRT,BLACK | Jan 24, 2023 | 4 |
| Trans-5115 | MANCHESTER | STRIPED KNIT SWEATER,BROWN | Jan 19, 2023 | 4 |
| Trans-5116 | LIVERPOOL | JOGGER WAIST TROUSER,WHITE | Jan 24, 2023 | 4 |
| Trans-5117 | LONDON | HIGH-WAIST TROUSER,NAVY | Jan 18, 2023 | 2 |
| Trans-5118 | LIVERPOOL | FAUX SUEDE BOMBER JACKET,GRAY | Jan 20, 2023 | 2 |
| Trans-5119 | LONDON | COLOUR BLOCK LEATHER JACKET,AQUA | Jan 20, 2023 | 2 |

# Congratulations !!

Instructor:

Mohamed (Mo) Touiti