Deliverable 3

CSI4142 : Fundamentals of Data Science
winter 2023

School of Electrical Engineering and Computer Science
University of Ottawa

Professeur: Yazan Otoum, Ph.D

Group members :
Souleymane Wilfried Sankara, #300100940
Marie Fabienne Sawadogo; #300101795
Tata Saidatou Berte, #300100935

A) High-level data staging plan

- Data Acquisition : The dataset is about the data scientist jobs salary. So we get the dataset from Kaggle with the following link : https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor?select=glassdoor_jobs.csv

- Data Cleaning : The cleaning of the dataset consisted first of removing the irrelevant column for the analysis. These columns were `"Unnamed: 0","Headquarters","Sector","age",”Company size”,"Salary Estimate","JobDescription","hourly","employer_provided","company_txt","job_state","Competitors","same_state","python_yn","R_yn","spark","aws","excel","seniority","desc_len","num_comp"`.
  After that, we removed the “Founded” and the “Revenue” columns because they had a lot of null values. In this stage we decide to keep the rating column although they have some null values columns because they are relevant for the analysis.
  We removed also the rows with inconsistent values in “Industry”,”Rating” and “Size” columns

- Data transformation : In the original form of the dataset, the “job_simp” column had a lot of null value. We noticed that we can solve this issue by analyzing the first column and using this information to fill the missing rows. We filled the missing rows with the label “Research”.
  We also transform the “location” column data. Indeed, they had the form “City,State”, we then split them into two different columns. The first column was the “City” column and the second one was the “State”. We transformed the “Size” columns data into range values.

- Data storage : We stored the staging data in MySql data database.

DATA QUALITY ISSUE

During the data staging we faced certain issues. The first one was the amount of rows, we only got 742 rows which made our data cleaning more delicate because we didn’t want to remove anymore rows. But the advantage of this dataset was the large number of columns. So we took advantage of that.

In “job_simp” columns we got more than 100 missing values. So to avoid removing them, we used the first column to make some correspondence and complete the missing rows.

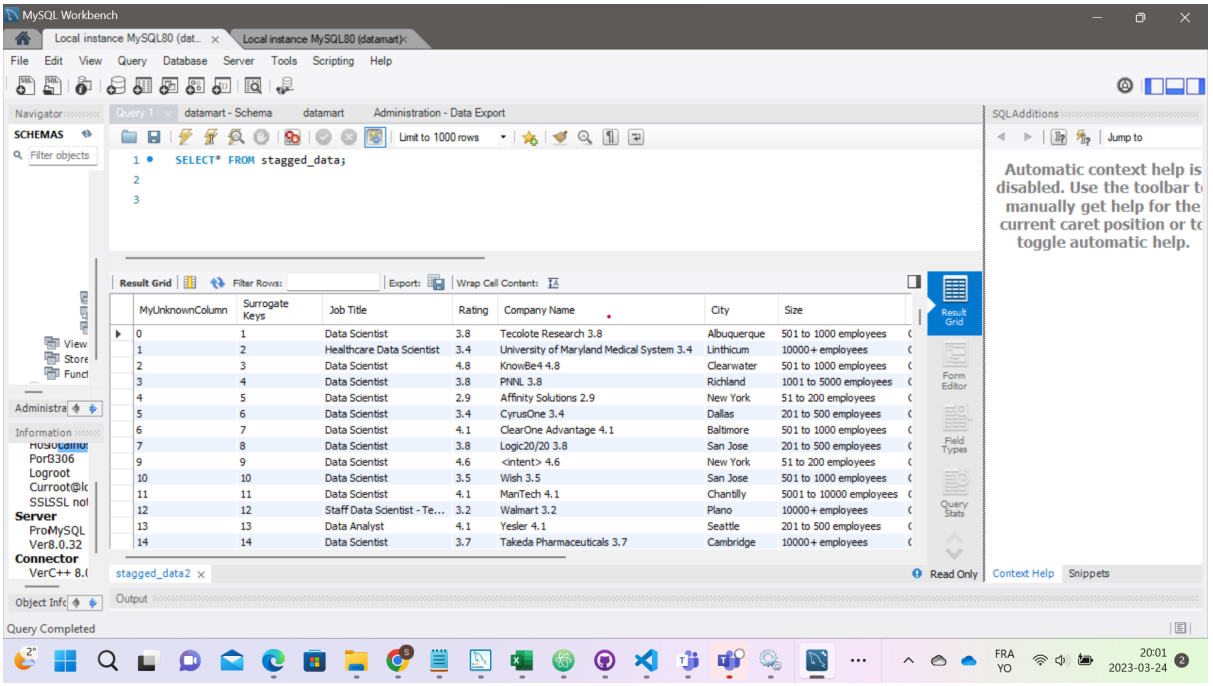
Also in the “Industry”,”Size” and “Rating” columns we gave us some inconsistent values. To solve this problem, we checked the total number of inconsistency rows and the number was very low, less than 10. So we decided to remove them.

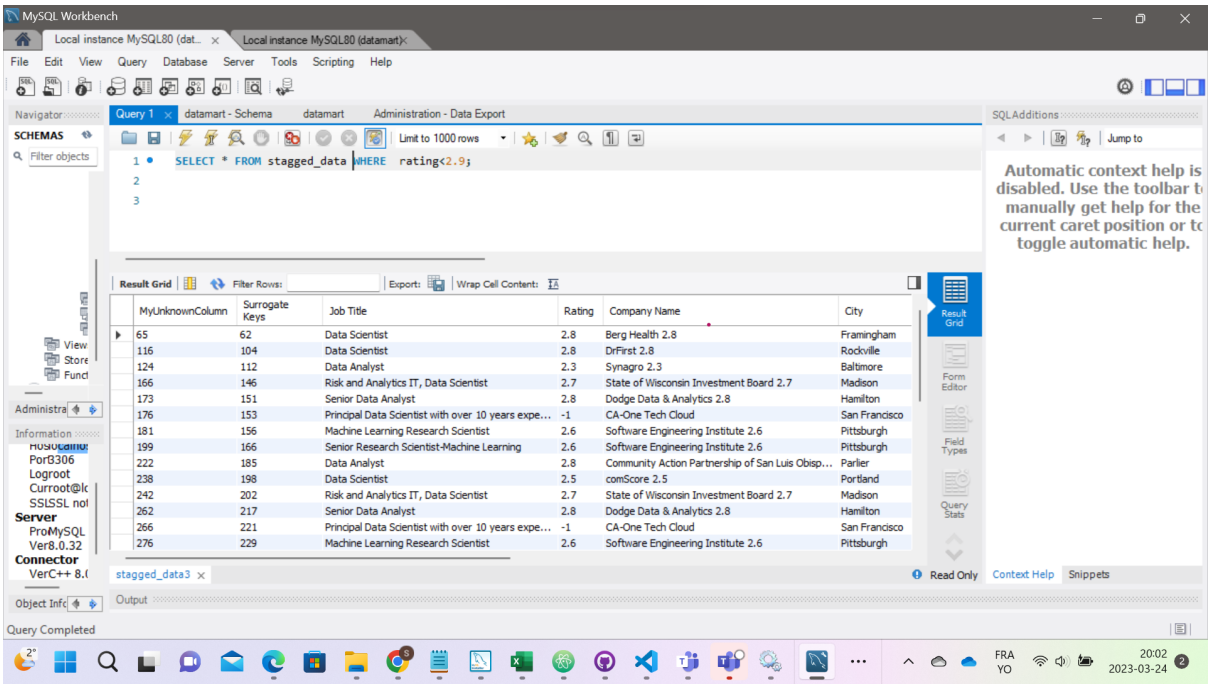
MySQL database

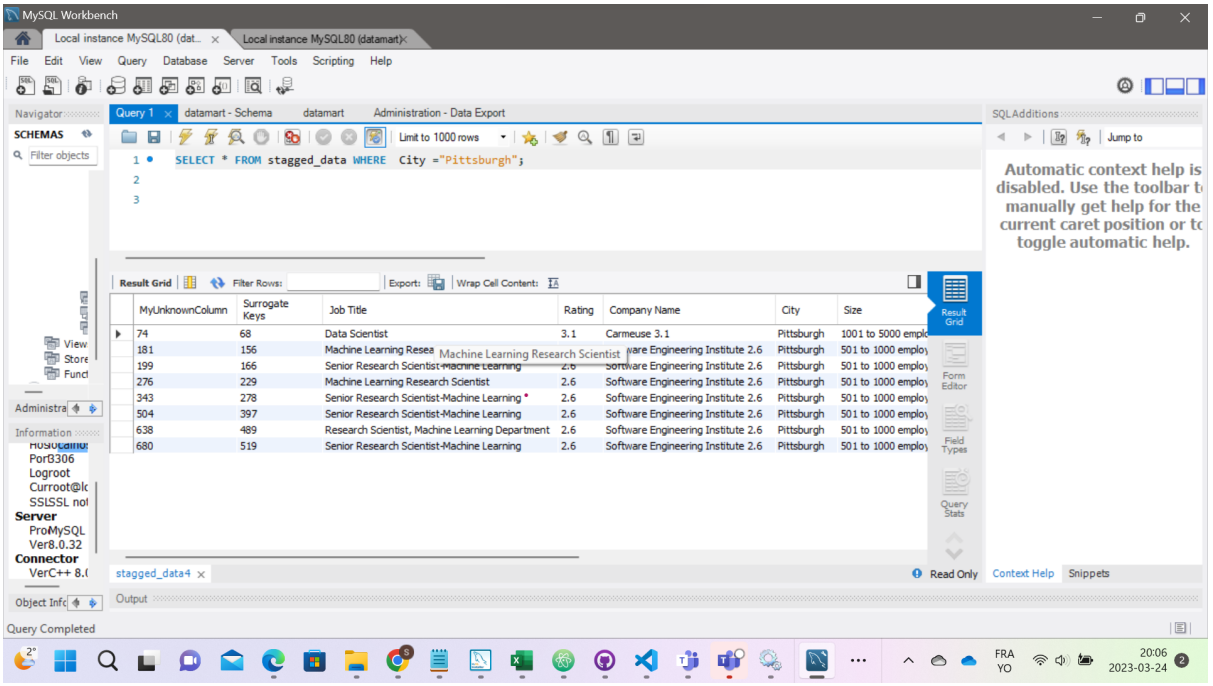
We created the database using MySQL.

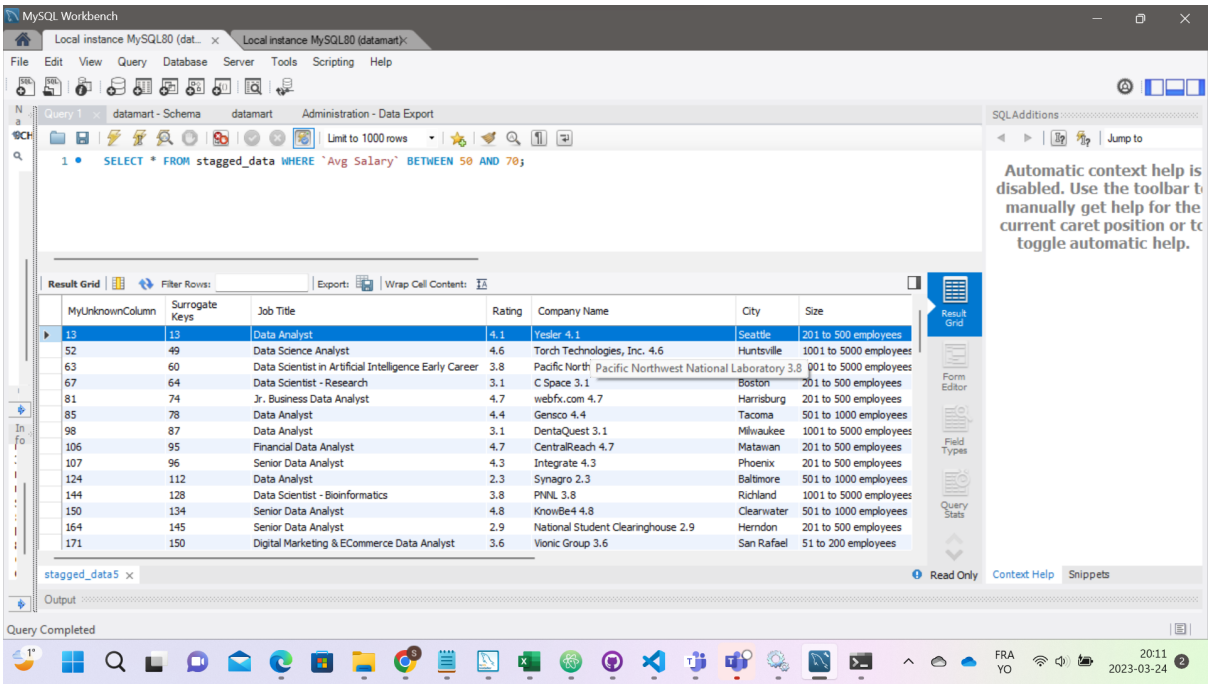
````CREATE TABLE stagged_data (
    SurrogateKey INT PRIMARY KEY,
    JobTitle VARCHAR(255),
    Rating FLOAT,
    CompanyName VARCHAR(255),
    City VARCHAR(255),
    Size VARCHAR(255),
    TypeOfOwnership VARCHAR(255),
    Industry VARCHAR(255),
    MinSalary INT,
    MaxSalary INT,
    AvgSalary FLOAT,
    JobClassification VARCHAR(255),
    State CHAR(2)
);

````


MySQL Workbench
Local instance MySQL80 (dat...
Local instance MySQL80 (datamart)
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
Query 1
datamart - Schema
datamart
Administration - Data Export
Limit to 1000 rows
1 • SELECT* FROM stagged_data;
2
3
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
MyUnknownColumn | Surrogate Keys | Job Title | Rating | Company Name | City | Size
0 | 1 | Data Scientist | 3.8 | Tecolote Research 3.8 | Albuquerque | 501 to 1000 employees
1 | 2 | Healthcare Data Scientist | 3.4 | University of Maryland Medical System 3.4 | Linthicum | 10000+ employees
2 | 3 | Data Scientist | 4.8 | KnowBe4 4.8 | Clearwater | 501 to 1000 employees
3 | 4 | Data Scientist | 3.8 | PNNL 3.8 | Richland | 1001 to 5000 employees
4 | 5 | Data Scientist | 2.9 | Affinity Solutions 2.9 | New York | 51 to 200 employees
5 | 6 | Data Scientist | 3.4 | CyrusOne 3.4 | Dallas | 201 to 500 employees
6 | 7 | Data Scientist | 4.1 | ClearOne Advantage 4.1 | Baltimore | 501 to 1000 employees
7 | 8 | Data Scientist | 3.8 | Logic20/20 3.8 | San Jose | 201 to 500 employees
9 | 9 | Data Scientist | 4.6 | <intent> 4.6 | New York | 51 to 200 employees
10 | 10 | Data Scientist | 3.5 | Wish 3.5 | San Jose | 501 to 1000 employees
11 | 11 | Data Scientist | 4.1 | ManTech 4.1 | Chantilly | 5001 to 10000 employees
12 | 12 | Staff Data Scientist - Te... | 3.2 | Walmart 3.2 | Plano | 10000+ employees
13 | 13 | Data Analyst | 4.1 | Yesler 4.1 | Seattle | 201 to 500 employees
14 | 14 | Data Scientist | 3.7 | Takeda Pharmaceuticals 3.7 | Cambridge | 10000+ employees
stagged_data2
Read Only
Context Help
Snippets
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or tc toggle automatic help.
Output
Query Completed
20:01
2023-03-24


MySQL Workbench
Local instance MySQL80 (dat...
Local instance MySQL80 (datamart)
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
Query 1
datamart - Schema
datamart
Administration - Data Export
Limit to 1000 rows
1 • SELECT * FROM stagged_data WHERE rating<2.9;
2
3
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
MyUnknownColumn | Surrogate Keys | Job Title | Rating | Company Name | City
65 | 62 | Data Scientist | 2.8 | Berg Health 2.8 | Framingham
116 | 104 | Data Scientist | 2.8 | DrFirst 2.8 | Rockville
124 | 112 | Data Analyst | 2.3 | Synagro 2.3 | Baltimore
166 | 146 | Risk and Analytics IT, Data Scientist | 2.7 | State of Wisconsin Investment Board 2.7 | Madison
173 | 151 | Senior Data Analyst | 2.8 | Dodge Data & Analytics 2.8 | Hamilton
176 | 153 | Principal Data Scientist with over 10 years expe... | -1 | CA-One Tech Cloud | San Francisco
181 | 156 | Machine Learning Research Scientist | 2.6 | Software Engineering Institute 2.6 | Pittsburgh
199 | 166 | Senior Research Scientist-Machine Learning | 2.6 | Software Engineering Institute 2.6 | Pittsburgh
222 | 185 | Data Analyst | 2.8 | Community Action Partnership of San Luis Obisp... | Parlier
238 | 198 | Data Scientist | 2.5 | comScore 2.5 | Portland
242 | 202 | Risk and Analytics IT, Data Scientist | 2.7 | State of Wisconsin Investment Board 2.7 | Madison
262 | 217 | Senior Data Analyst | 2.8 | Dodge Data & Analytics 2.8 | Hamilton
266 | 221 | Principal Data Scientist with over 10 years expe... | -1 | CA-One Tech Cloud | San Francisco
276 | 229 | Machine Learning Research Scientist | 2.6 | Software Engineering Institute 2.6 | Pittsburgh
stagged_data3
Read Only
Context Help
Snippets
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or tc toggle automatic help.
Output
Query Completed
20:02
2023-03-24


MySQL Workbench
Local instance MySQL80 (dat...
Local instance MySQL80 (datamart)
File Edit View Query Database Server Tools Scripting Help
Query 1
datamart - Schema
datamart
Administration - Data Export
SCHEMAS
Limit to 1000 rows
1 • SELECT * FROM stagged_data WHERE City ="Pittsburgh";
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
MyUnknownColumn | Surrogate Keys | Job Title | Rating | Company Name | City | Size
74 | 68 | Data Scientist | 3.1 | Carmeuse 3.1 | Pittsburgh | 1001 to 5000 empl
181 | 156 | Machine Learning Resea | | ware Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
199 | 166 | Senior Research Scientist-Machine Learning | | Software Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
276 | 229 | Machine Learning Research Scientist | 2.6 | Software Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
343 | 278 | Senior Research Scientist-Machine Learning | 2.6 | Software Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
504 | 397 | Senior Research Scientist-Machine Learning | 2.6 | Software Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
638 | 489 | Research Scientist, Machine Learning Department | 2.6 | Software Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
680 | 519 | Senior Research Scientist-Machine Learning | 2.6 | Software Engineering Institute 2.6 | Pittsburgh | 501 to 1000 employ
Machine Learning Research Scientist
stagged_data4
Read Only
Context Help
Snippets
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or t toggle automatic help.
Output
Query Completed
FRA YO
20:06
2023-03-24


MySQL Workbench
Local instance MySQL80 (dat...
Local instance MySQL80 (datamart)
File Edit View Query Database Server Tools Scripting Help
Query 1
datamart - Schema
datamart
Administration - Data Export
Limit to 1000 rows
1 • SELECT * FROM stagged_data WHERE `Avg Salary` BETWEEN 50 AND 70;
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
MyUnknownColumn | Surrogate Keys | Job Title | Rating | Company Name | City | Size
13 | 13 | Data Analyst | 4.1 | Yesler 4.1 | Seattle | 201 to 500 employees
52 | 49 | Data Science Analyst | 4.6 | Torch Technologies, Inc. 4.6 | Huntsville | 1001 to 5000 employees
63 | 60 | Data Scientist in Artificial Intelligence Early Career | 3.8 | Pacific North | | 001 to 5000 employees
67 | 64 | Data Scientist - Research | 3.1 | C Space 3.1 | Boston | 201 to 500 employees
81 | 74 | Jr. Business Data Analyst | 4.7 | webfx.com 4.7 | Harrisburg | 201 to 500 employees
85 | 78 | Data Analyst | 4.4 | Gensco 4.4 | Tacoma | 501 to 1000 employees
98 | 87 | Data Analyst | 3.1 | DentaQuest 3.1 | Milwaukee | 1001 to 5000 employees
106 | 95 | Financial Data Analyst | 4.7 | CentralReach 4.7 | Matawan | 201 to 500 employees
107 | 96 | Senior Data Analyst | 4.3 | Integrate 4.3 | Phoenix | 201 to 500 employees
124 | 112 | Data Analyst | 2.3 | Synagro 2.3 | Baltimore | 501 to 1000 employees
144 | 128 | Data Scientist - Bioinformatics | 3.8 | PNNL 3.8 | Richland | 1001 to 5000 employees
150 | 134 | Senior Data Analyst | 4.8 | KnowBe4 4.8 | Clearwater | 501 to 1000 employees
164 | 145 | Senior Data Analyst | 2.9 | National Student Clearinghouse 2.9 | Herndon | 201 to 500 employees
171 | 150 | Digital Marketing &ECommerce Data Analyst | 3.6 | Vionic Group 3.6 | San Rafael | 51 to 200 employees
Pacific Northwest National Laboratory 3.8
stagged_data5
Read Only
Context Help
Snippets
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or t toggle automatic help.
Output
Query Completed
FRA YO
20:11
2023-03-24