**Assignment for the course MySQL for Data Analytics 2023**

Student ID: \_\_\_\_\_\_\_\_\_\_\_ Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* **In business, the result is often more important than the process.** If the final answer is wrong, you will get only a few points even if most of your MySQL codes are correct.
* **Please do not copy codes from other students.** Cheating (such as copying) on assignments will lead to course failure.
* **You can add new columns to a table or create additional tables to support your analysis.**
* **The assignment has a total of 88 points, which will be standardized to 100 points for grading.**

1. In the table orders at the car retailer (**classicmodels**) database, what is the customerNumber of the customer who has the highest frequency of placing orders to the company in 2004 [orderDate in 2004]? You don’t need to consider whether the products have been finally shipped or not.

The customerNumber of the customer: \_\_\_\_\_\_\_(6 points)

The highest frequency is: \_\_\_\_\_\_ (6 points)

*The MySQL code that generates the result:*

1. In the **classicmodels** database, customer names and customer numbers can be found [see table “customers”]. Customers make different payments on different dates [See table “payments”]. Please specify the names of two customers who are most **often** (count frequency) to make payments during the weekend [12 points].

Name of Customer 1: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Customer 2: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Note: i) Please provide “customerName”, not the contact names from table “customers”

*The MySQL code that generates the result:*

1. In the classicmodels database, one sales representative is responsible for one or several customers [see “salesRepEmployeeNumber” in table “customers”]. Customers make different payments on different dates [See table “payments”]. In other words, we can say sales representatives help the company get customers to make payments. Now the question is, who is the sales representative that brings the most **revenue** [or **the total amount of the payments from customers in the table payments**] to the company?

Note: i) Ignoring those customers who are not assigned to any sales representative – they have not made any purchases yet.

The salesRepEmployeeNumber of the sales representative who brings the most **revenue** is:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ [12 points].

*The MySQL code that generates the result:*

1. In the **cfpb\_complaints\_2500** database, you can find “closed with relief” in **Company\_response** column. Please find the **name of the company** that has the highest **ratio** of cases that is ‘**closed with relief’**? Only those companies with more than 30 cases [all different kinds of cases, no matter whether they are featured with ‘closed with relief’ or not] in the database are considered (10 points).

- **ratio** [or percentage]for an individual company is calculated as:

The amount of its cases featured with ‘closed with relief’ / the amount of its total cases

The name of the company: \_\_\_\_\_\_\_\_\_ (5 points)

The ratio of the cases for the company: \_\_\_\_\_\_\_\_\_ (5 points)

*The MySQL code that generates the result:*

1. In the **cfpb\_complaints\_2500** database, many complaints are related to ‘loan’ (those cases where the word ‘loan’ is included in the column ‘Issue’). Please specify **the name of the company** that has the most issues related to ‘**loan’** on **Wednesday (DATA\_received)**. Only complaints with the column ‘**State’** starting with character “**A**” are considered (8 points).

The name of the company: \_\_\_\_\_\_\_\_\_ (8 points)

*The MySQL code that generates the result:*

1. In the Chile database, let’s assume that an income less than 10,000 is a low income; an income between 10,000 and 100,000 is a middle income; an income higher than 100,000 is a high income. We would like to know whether the income level and the statusquo have a certain relationship for females who voted yes to Pinochet. To answer this question, you need to provide the average statusquo value for the females who voted yes to Pinochet in correspondence to their different income levels. (9 points)

[Please carefully read the question so that you will not miss any important condition when answering the question; Please provide **three** digits after the decimal point in the results]

Suggestion: you may need to update the table by adding a new column of income\_level to answer the question.

|  |  |
| --- | --- |
| Income level | Mean statusquo |
| High\_income |  |
| Middle\_income |  |
| Low\_income |  |

*The MySQL code that generates the result:*

1. Based on the use of cfpb consumer complaint database (2500 rows), please count the frequency of the complaints that satisfy the following three conditions at the same time: i) consumer finally disputed with the company (*Consumer\_disputed*); and ii) were received on Friday (*Data\_received*) and iii) the difference between *Data\_received* and *Data\_sent\_to\_company* is more than 5 days. (10 points)

Please specify the name of the company that has the biggest amount of the above-mentioned complaints.

The name of the company: \_\_\_\_\_\_\_\_\_\_\_\_(5 points)

Frequency of the mentioned complaints of the company: \_\_\_\_\_\_\_\_\_\_\_(5 points)

*The MySQL code that generates the result:*

1. In the data “tripadvisor\_review\_sample\_without\_reviewtext”, based on the review titles that have **at least two words** [a total of 5 points]:
2. Regarding the **first** word used in the review title, what is the most popular word [case-insensitive]?

The word is: \_\_\_\_\_\_\_\_\_\_\_[1 points] The frequency of the word is: \_\_\_\_\_\_\_\_\_\_[1 points]

1. Regarding the **second** word used in the review title, what is the most popular word [case-insensitive]?

The word is: \_\_\_\_\_\_\_\_\_\_\_[1 points] The frequency of the word is: \_\_\_\_\_\_\_\_\_[1 points]

1. For those titles that **starts with** the words “bad” OR “terrible”, what are the most popular **second** word [case-insensitive]:  
   The word is: \_\_\_\_\_\_\_\_\_\_\_[1 points]  
     
   Note: Before answering question 6, please clean the title based on the following two requirements.

i) Please remember to first **remove** six types of punctuation marks from the title, including:

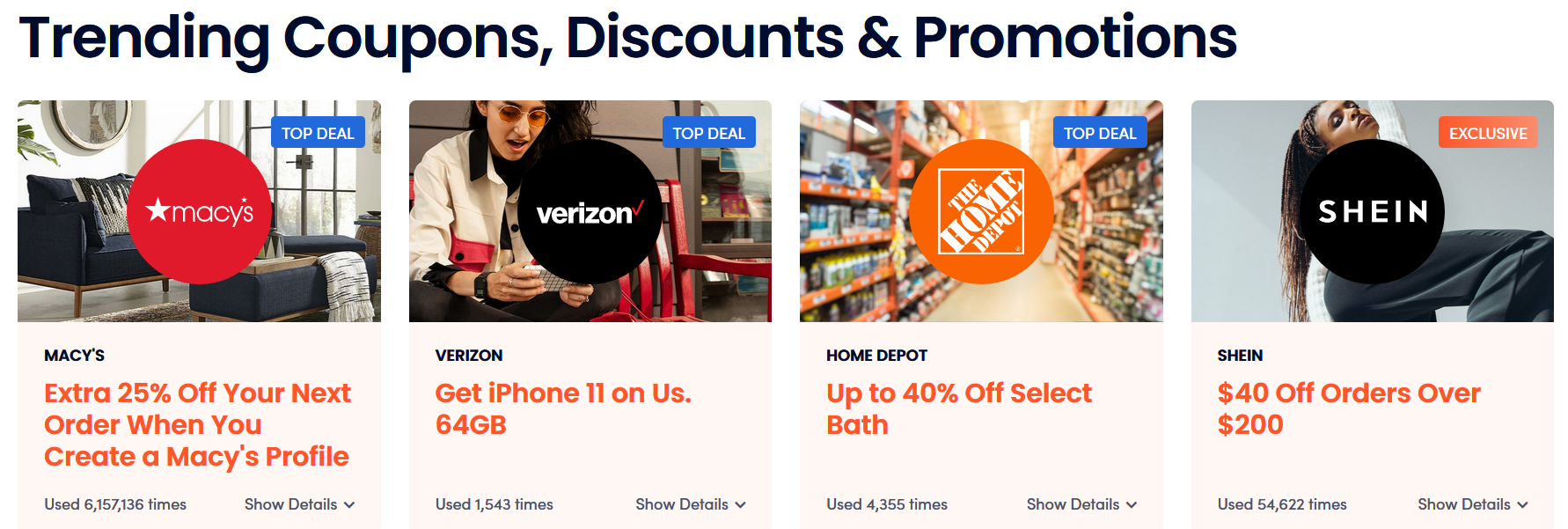
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| “ | ” | " | - | , | ! |

*In order to obtain consistent results*, please do not remove more punctuation marks than we specified above and also do not replace the punctuation with an empty space.  
 ii) Please remove empty spaces from both sides of the title.

*The MySQL code that generates the result:*

1. The following screenshot shows the sale volume of a coupon site (<https://www.savings.com/>). Assuming that you have collected the sale volume information of each coupon on this website **on a daily basis**. For instance, the **total sale volume** for the Coupon “Macy’s” is 6,157,136 at the data collection date.

You want to calculate the daily sale volume for each coupon of this website (maybe because this website is the competitor of your company) for further analysis. How can you do that task using MySQL? **(5 points)**



Based on the data of the **total sale volume** for each coupon (like **Table A)**. Please calculate **the daily sale volume**, as shown in **Table B**.

**Explanation**: For instance, the **daily** sale volume for CouponA on ‘2021-09-05’ is 25, because the **total** sale volume for CouponA is 37 on ‘2021-09-**05’** and 12 on ‘2021-09-**04’.** The difference in total sale volumes between the two dates are the daily sale volume, which is   
37 – 12 = 25.

|  |  |
| --- | --- |
|  |  |
| **Row Data: Table A** | **Result: Table B** |

**Requirement**:

1. Please write code to generate the results as shown in Table B – the code will be the answer.
2. You can download Table A from Mycourse (coupon\_sale\_volume.sql)

*The MySQL code that generates the result* **(5 points)***:*

1. Assuming you are now a business analyst offering consultant service to the tourism minister of Finland **(a total of 5 points)**. The minister wants to know how tourists travel within Finland between different cities.

* Specifically, **for those tourists whose** **first visit** to Finland is Helsinki city [i.e., the first review in the database (in terms of review\_date) is about a hotel in **Helsinki\_Uusimaa**], which city would most likely be visited by those tourists in the future?

1. **For those tourists whose** **first visit** to Finland is Helsinki city [i.e., the first review in the database (in terms of review\_date) is about a hotel in **Helsinki\_Uusimaa]**, they also visited “**Rovaniemi\_Lapland”** for \_\_\_\_\_\_\_\_ **times in the future** (2 points).
2. **For those tourists whose** **first visit** to Finland is to Helsinki city [i.e. first review in the database (in terms of review\_date) is about a hotel in **Helsinki\_Uusimaa]**, \_\_\_\_\_\_\_\_\_\_\_ of **them** also visited **both** “**Saariselka\_Lapland”** and “**Rovaniemi\_Lapland**” cities **in the future** (3 points).

Note:

* Please read the assignment questions carefully!
* Please download the “assignment\_tourist\_Finland.sql” dataset from MyCourse.
  + It would be good to remove previous review-related tables before you import the database file so that you won’t mix the current assignment tables with previous review-related tables.
  + The sql file contains two tables of “hotel” [431 records] and “review2” [56,709 records]. Column ‘Id’ of the table ‘hotel’ is connected to the column ‘hotel\_id’ of the table ‘review2’
* It may happen that a traveler wrote multiple reviews about hotels in different cities as his/her **first** reviews (on the same but earliest review date). If one of these ‘first’ reviews includes Helsinki city, the traveler should be counted as a tourist whose first visit to Finland is Helsinki city. If one of these ‘first’ reviews includes ‘Rovaniemi\_Lapland’ city, that trip is **NOT** considered visiting ‘Rovaniemi\_Lapland’ city. “**In the future**” means future trips after these first reviews that were written on the same but earliest review day - *I know this does not sound so logical, but it is good to increase the difficulty of the assignment question for training your mind).*
* Assume that different values in the column “city” of the “hotel” table represent different cities. For instance, “**Saariselka**” and “**Rovaniemi**” are two different cities.