Hands-on training session 5 (Questions)

<u>Please do all of the following tasks by yourself without copying my answer code, unless it is really necessary. As what I said in the class, you cannot learn to ride a bicycle by watching other people riding a bicycle. So please do not use my code, but learn from your own mistakes so that MySQL becomes your own friend and skill.</u> – Your lecturer

- **Task 1:** In the orderdetails table at the car retailer database, please count the number of different prices that each product (productCode) have, and return the productCode of only those who have been sold in more than 5 different prices.

- **Task 2:**
  - Please count how many empty values there are in the column of Sub_product at the cfpb_complaints_2500.
  - Replace these empty values with NULL
  - Drop the records that have NULL values in sub_product from table.

  Note: after doing this task, please drop this incomplete table and import the complete data file from MyCourse again.

- **Task 3:** Which company received the largest amount of complaints from consumers on Wednesday? Assume the Data_received is the date that company received complaints.

- **Task 4:** Which weekday did companies receive the largest amount of complaints?

- **Task 5:** Which month did the companies receive the largest amount of complaints?

- **Task 6:** Considering only the first 7 days of each month, which month did the companies receive the largest amount of complaints?

- **Task 7:** in the table cfpb_complaints_2500, create a new column in the name of ‘datedifference’, this column contains the difference in days between Data_sent_to_company and Data_received.

- **Task 8:** When will be date that you have living in this world for 10,000 days? How about 20,000 days?

- **Task 9**: How many days have you living in this world by today?

- **Task 10:** In the table **products** of the car retailing database, add two new columns of Price_difference, and Value_difference, given that:

  - **Price_difference = (MSRP-buyPrice)**
  - **Value_difference = (MSRP-buyPrice)* quantityInStock**

- **Task 11:** Remove the two new columns of Price_difference, and Value_difference

- **Task 12:** In the Classicmodels database, retrieve the names of the customers who purchased Ferrari in 2004!

Tips:

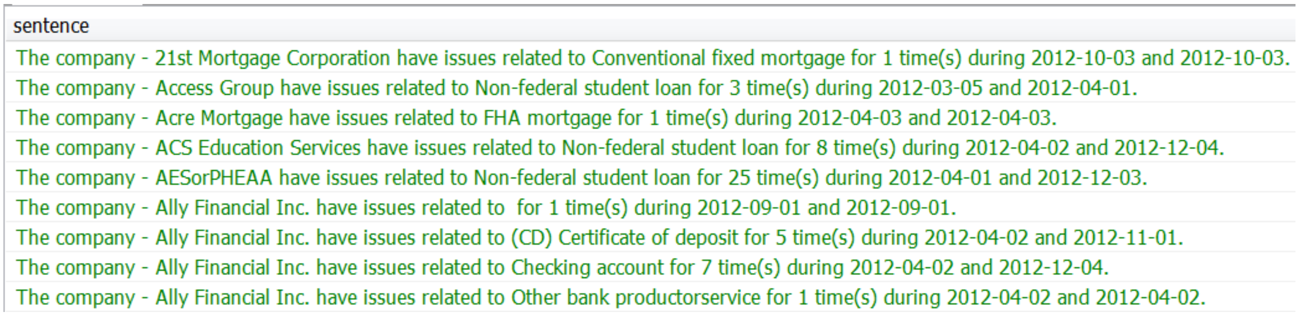
- In the table products, you will find productName including the word “Ferrari”
- In the table orderdetails, you will find the all the orderNumber that have ordered the products including “Ferrari”

- In the table orders, you will find the customerNumber connecting to each orderNumber as well as when the order was made.
- In the table customers, you will find the names of the customers connecting to their customerNumber.

This is a challenging task! Please try to do the task by yourself. However, please also feel free to ask for help!

➢ **Task 13:** In the table [cfpb_complaints_2500], count the frequencies of different **Issues** for different companies and show the starting date and ending date that different issues have been reported, and order the records based on first the company name from A-Z and then on frequencies in a descending manner. Save the results to a new table “Count_result”

➢ **Task 14**: Based on the results in above table “Count_result”, create a report like below.



sentence
The company - 21st Mortgage Corporation have issues related to Conventional fixed mortgage for 1 time(s) during 2012-10-03 and 2012-10-03.
The company - Access Group have issues related to Non-federal student loan for 3 time(s) during 2012-03-05 and 2012-04-01.
The company - Acre Mortgage have issues related to FHA mortgage for 1 time(s) during 2012-04-03 and 2012-04-03.
The company - ACS Education Services have issues related to Non-federal student loan for 8 time(s) during 2012-04-02 and 2012-12-04.
The company - AESorPHEAA have issues related to Non-federal student loan for 25 time(s) during 2012-04-01 and 2012-12-03.
The company - Ally Financial Inc. have issues related to for 1 time(s) during 2012-09-01 and 2012-09-01.
The company - Ally Financial Inc. have issues related to (CD) Certificate of deposit for 5 time(s) during 2012-04-02 and 2012-11-01.
The company - Ally Financial Inc. have issues related to Checking account for 7 time(s) during 2012-04-02 and 2012-12-04.
The company - Ally Financial Inc. have issues related to Other bank productorservice for 1 time(s) during 2012-04-02 and 2012-04-02.

E.g. The company - **21st Mortgage Corporation** have issues related to **Conventional fixed mortgage** for **1** time(s) during **2012-10-03** and **2012-10-03**.

➢ **Task 15**. In the [tripadvisor_data_for_handson_assignment_ONLY] dataset, please evaluate whether and how the use of mobile device to write a review affects how people gave a rating to a hotel.

- The code to do the task is easy, but I hope you could think about why and how this happens. When review platform allows people to submit review via mobile phone, how will it affect the ratings on a hotel and review length? This actually becomes a real business intelligence question that you need to get an answer by yourself.

- **Task 16.** Research shows that the review given at the same month of lodging vs. different month of lodging would be significantly different. Please find evidence from our data. Please also check how the review generation method (via mobile phone) affect the reported difference.

- the column ‘review_date’ provides the date of when the review is provided.

- the columns ‘year_stayed’ and ‘review_date’ provides the information of which month the accommodation in a hotel took place.

- **Task 17.** Please import ‘tripadvisor_hotel_sample.sql’ to your database. This data table provides the information of hotels, while the [tripadvisor_data_for_handson_assignment_ONLY] table provides the information of reviews on hotels. Please get the list of hotels that are in [tripadivsor_hotel_sample] table, but not in [tripadvisor_data_ for_handson_assignment_ONLY] table.

- **Task 18.** In the tripadvisor_hotel_sample table, please find out whether different hotels are actually using the same hotel name? Please sort the result by having the most frequently used name appear first.

- **Task 19.** Through the above task, we find different hotels using the same name. However, whether could different hotels with the same name appear in the same city?

- **Task 20.** Please write ONE query to count the number of hotels with different hotel stars, calculate the average length of hotel name, average length of hotel address, and think about **why** these differences exist among different stars of hotels.

- **Task 21**. In the ‘orders’ table of ‘classicmodels’ database, we can find time (orderDate) of product orders (orderNumber) that were made by each customer (customerNumber). Could you please identify the **first** order (in term of orderDate) made by each customer in the database?

- **Task 22.** In the answer of task 21, what would be the results, if we replace “IN” with “NOT IN”.

- **Advance tasks**

- Take the advance tasks only if you have interest to do so. It is not within the requirement of course, but only for those who have interests to explore the potential of MySQL.

- The questions below are above the requirement for MySQL beginners. The purpose to offer these extra tasks is simply to show what you can achieve by using MySQL, and how MySQL would be used for developing business intelligence.

- You do not need to code by yourself. It would be sufficient if you can understand the code.

- **Task 1.** If you check Mycourse page, you will find a paper [How do tourists evaluate Chinese hotels at different cities? Mining online tourist reviewers for new insights]. By quickly browsing the paper, you will find that the most popular words describe the conditions of tourism in a city. **In this vein, how can we obtain the most popular words and their frequency across a large amount of review via the use of MySQL?** There are some analysis and data manipulation done in the paper, and now I would like you to duplicate the analysis via the use of our course dataset. Specifically, based on the review titles with maximal 10 words:

- Please ignore those titles like '<U+7ACB>', which is probably caused by non-English words.

1) Please provide me the most popular 30 words used in review titles. Remove the punctuation like [, ! “”].

2) Please compare the ten most popular obtained from review title that are written by the authors whose 'author_location' is empty VS. those whose 'author_location' is not empty.

3) What are the ten most popular words co-occurring with the word “great”? What are the ten most popular words co-occurring with the word “terrible”?

4) What are the ten most popular words for 5-star overall rating review titles? What are the ten most popular words for 1-star overall rating review titles?

- **Task 2.** There is a encrypt method, which is substitution cipher [or letter replacement, see. https://en.wikipedia.org/wiki/Substitution_cipher]. There is a method to decrypt the text that were encrypted by this method, which is frequency analysis [see. https://en.wikipedia.org/wiki/Letter_frequency]. In other words, given a sufficient long text that is encrypted, I can easily decrypt it by comparing the frequency of letters.

Here, as a way of training, my question is that can you calculate the frequency of the letter used in the TripAdvisor review titles, e.g. based on the left ten letters of each review title to reduce the difficulty. Can you do this? Code to answer this question is not provided.

- **Task 3.** What are the most popular words that have been used to name hotels? For instance, what are the most popular words in the name of 5-star hotels? How about 1-star hotels?

Code to answer this question is not provided.