**Introduction**

During this course, you added two tools, Python and SQL, to your data science toolkit. In this project, you will apply what you have learned to derive insights and present them to the management.

Scenario

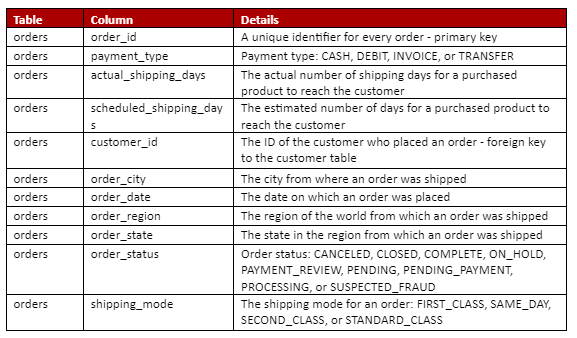
You are working for TerpBuy, a global business-to-consumer and business-to-business platform. It is based in and operates out of College Park, Maryland, USA, and has a distribution facility in Mumbai, India. The company is looking for insights on different aspects of its customers, products, departments, and orders. To help the company, you will have to export data from its data warehouse.

**About the Assignment**

Where do I get the data?

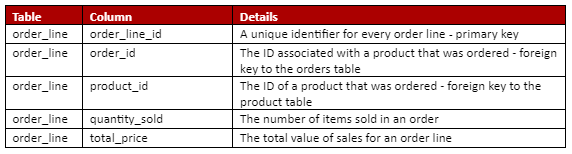
The SQL script with the TerpBuy data set is provided at the bottom of this page. You can download it from there. You should also review this data dictionary, which identifies the attributes for each table. 

**Table 1: orders**



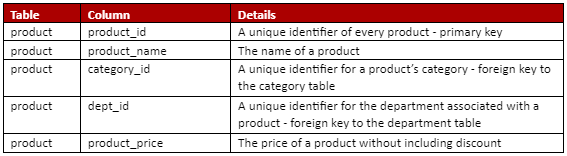
Orders

Table 2: order\_line



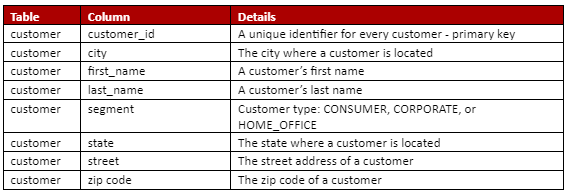
Order\_line

Table 3: Product



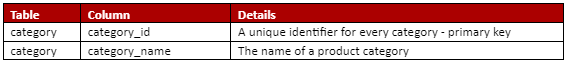
Product

Table 4: customer



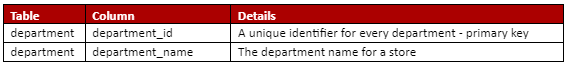
Customer

Table 5: category



Category

Table 6: department



Department

Where do I write code?

Following the directions below, you will first write queries using MySQL Workbench and then take screenshots of the results and add them to a PDF document. Then you will create your own Jupyter Notebook. Finally, you will create a PDF with your executive summary.

Some Tips for the Assignment

Here are some tips for you to complete the assignment:

1. Read the data dictionary thoroughly before starting the assignment. This will give you a good idea of what each database column represents before you begin the analysis.
2. Read all instructions carefully, identify the task(s) to be performed, and proceed to write the required code only then. This assignment is meant to be straightforward. You don’t have to perform additional analyses that are not explicitly requested.
3. There might be some tasks that require you to use functions that you may not have used before. For such tasks, you should rely on the SQL documentation, which you referred to during the modules. Understand that solving this assignment is a part of your learning process, and it includes researching the use of functions and troubleshooting code.
4. For your Jupyter Notebook, always run cells sequentially, or restart the kernel and then run all cells to avoid runtime errors.
5. For many of the questions, there are multiple possible approaches to accomplish what has been asked. In other words, there is more than one correct answer. Additionally, you are expected to document your work appropriately.

How do I submit the assignment?

You need to include three files in a folder:

1. **SQL Query Execution:** A PDF document containing the answers to Part I on the next segment.
2. **Connecting Python to SQL:** A Jupyter Notebook containing the answers to Part II on the next segment.
3. **Executive Summary:** A PDF document containing the answers to Part III on the next segment.

You will have to compress the folder (a zipped folder) containing these files and submit the final zipped folder in the Submission section. Your submission must have three files.

In the next segment, you will find the list of tasks to be completed in the project.

*Disclaimer : This learning module may contain images, both still and moving, including photographs, advertisements, third-party trademarks, video clips etc. that may constitute copyrighted material procured from open-source, public domain, social media, or other print or digital publications, with the sole and bonafide intention of imparting education, disseminating information and illustrating an academic concept or an issue. We believe that this constitutes a 'fair dealing' exception under Section 52 of the Indian Copyright Act, 1957.*

**Project Tasks**

**Part I: SQL Queries**

For each of the queries below, write only one query that answers the question. Your query results must clearly provide the answer. All query results must show text values rather than identifiers. For example, when showing a department in the results, it should show 'Fitness' instead of its corresponding ID value, 2:

1. How many rows of data are stored for each table in the database? List the name of each table followed by the number of rows it has.
2. Which products are considered high-priced products? A high-priced product has a price exceeding $100.00. List the names and prices of the high-priced products.
3. List all orders placed by customers in the state of Florida. Note: The state abbreviation for Florida is 'FL'. Include the customers’ first names, last names, city, and segment, along with the order ID and order date.
4. List all products that fall in one of the following categories: 'Computers', 'Toys', 'Tennis  & Racquet'. Include the products’ names, category, department, and price.
5. TerpBuy is considering reducing its product offerings. Which products have not yet been sold? Include the name, category, and department for each such product.
6. List the names of all cities from where orders are shipped. Also, for such cities, find the number of orders for which shipping was delayed. Sort the list of cities in order from the highest to the least number of shipping orders.
7. How many customers are there in each segment? Show the most popular segment at the top of the result. Incorporate a column alias in the result.
8. How many orders were placed in the first quarter of 2021? Note: A quarter consists of three months. Incorporate a column alias in the result. You can refer to the documentation on date functions provided [here](https://dev.mysql.com/doc/refman/8.0/en/date-and-time-functions.html).
9. List in alphabetical order all states supporting multiple customer segments.
10. To help the commercial sales department with its marketing, find all customers in the corporate segment who have not placed any orders. Include each customers’ first name, last name, street, city, state, and zip code. Sort the results by the last name first and then by the first name.
11. There has been a recall of the product Nike Mens Free 5.0+ Running Shoe. TerpBuy would have to offer a discount coupon to all customers who purchased this product. Find all orders that included this product as a part of the purchase. For all such orders, list the customers’ first names, last names, street, state, zip code, and order date. Each customer can be offered only one discount coupon. Hence, do not list the same customer more than once.
12. Premium customers are those customers who have placed orders with order amounts greater than the average order amount. For each customer, find the first and last names, and the order amount for all orders that exceeded the average order amount.

For each query, include the following items in your document:

1. The corresponding question number
2. The query
3. A legible screenshot of the query text and the executed results showing at least the first 10 rows. The screenshot must show your name and date included as a code comment. If your screenshot is not legible, you will not be awarded any points. You can visit [this](https://www.take-a-screenshot.org/) website for help with taking screenshots.

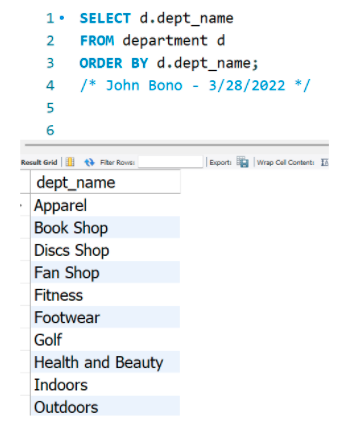
**Failure to include ALL these items will result in a maximum of half-credit for a question.**

As an example, consider that a question has asked you to do this:  
“List all departments in the database in alphabetical order.”

Your solution to this question must appear as shown below in your PDF document.

Query #1

SELECT d.dept\_name  
FROM department d  
ORDER BY d.dept\_name;



Sample

**Part II: Connecting Python to SQL**

All tasks below must be performed in the Jupyter Notebook:

1. Write a query to show the quantity of items sold by each department. Sort the results by department name.
2. Using the query you wrote in Question 1, create a data visualization (e.g., a bar chart) showing all departments and the number of items each of them sold. Using a markdown cell, explain what you observe from the analysis.
3. Write a query to show the number of orders placed in each year in which at least one order was placed. Hint: Search online to learn about MySQL’s YEAR() function to query data.
4. Using the query you wrote in Question 3, create a data visualization (e.g., a line graph) showing all years and the number of orders placed during each year, to see if there is a trend in ordering. Using a markdown cell, explain what you observe from the analysis.

**Part III: Executive Summary**

Based on the data analysis that you performed in the previous parts, write an executive summary highlighting what you believe are the key insights or recommendations that the company should take away from the analysis. The summary must be 200–300 words long.

**Evaluation Rubric**

The weights for the tasks in the assignment are given below.

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Meets Expectations** | **Does Not Meet Expectations** |
| **Part I: Queries (6% each, total 72%)**    If a query meets all expectations but the three items identified in the submission guidelines are not included, then the learner will earn a maximum of 3% for it. | * The query contains syntactically correct SQL code. * The query output correctly answers the question. * The query output does not require an additional manual review of the data. * The query contains all necessary data columns. | * There are minor or major syntax errors in the query. * The query output does not answer the question. * The query output requires an additional manual review of the data to answer the question. * The query does not contain all necessary data columns. |
| **Part II: Query Tasks (4% each, total 8%)** | * The query contains syntactically correct SQL code. * The query provides appropriate data for visualization. * The query contains all necessary data columns. | * There are minor or major syntax errors in the query. * The query output does not provide appropriate data for visualization. * The query does not contain all necessary data columns. |
| **Part II: Python Tasks (5% each, total 10%)** | * The commands are syntactically correct. * The code output correctly answers the question. * The code efficiently utilizes the built-in functions and libraries. * Comments and markdown cells have been used correctly with formatting and without spelling errors. The comments explain why the code is being written. The markdown cells provide a high level of understanding of each analysis that has been performed. | * There are minor or major syntax errors in the code. * The functions used are only partially correct. * The code output either shows more than what the question has asked or less than that. * Unclear results have been provided to address tasks. * Comments and markdown cells have not been used or have been written poorly, thus providing an ambiguous explanation, or they contain spelling errors and/or have been written in a language that is difficult to understand. |
| **Part III: Executive Summary (10%)** | * The executive summary is 200–300 words long. * The summary explains a minimum of three key insights that the company should take away from the data analysis; the insights are clearly supported by the data. * The summary does not contain spelling errors and/or has not been written in a language that is difficult to understand. | * The executive summary either contains fewer than 200 words or more than 300 words. * The summary explains fewer than three key insights that the company should take away from the data analysis, or the insights are not clearly supported by data. * The summary contains multiple spelling errors and/or has been written in a language that is difficult to understand. |