Assignment Chapter 4 - Data Wrangling with SQL

## Instructions

1. This assignment is split into 2 parts. For Part 1, no dataset is required. For part 2 you will need to use the boston\_crime.csv dataset that was used during the SQL demonstration lessons.
2. Please answer the questions in the boxes provided.
3. Please submit the assignment through the TalentLabs Learning System.

## Part 1: SQL Queries

**Question 1.1:**

Complete the query below to load data without duplicates.

|  |
| --- |
| SELECT  DISTINCT \*  FROM  dataset.tableName |

**Question 1.2:**

Write a query to select all columns from “cars.database”, and all rows which have missing values in the “mileage” column.

|  |
| --- |
| SELECT \*  FROM  cars.database  WHERE  mileage is NULL; |

**Question 1.3:**

Following on from question 1.2, write a query to replace the missing values in the mileage column with 0 for rows where the column “condition” has values equal to “new”.

|  |
| --- |
| UPDATE cars.database  SET mileage = 0  WHERE  mileage is NULL  AND  condition = “new”; |

**Question 1.4:**

Write a query to select 3 columns (“Date”, “Purchase\_Price”, “Purchase\_Desc”) from the following table: shop.history. Filter the query to only include data for dates (in “Date” column) between Jan 1st 2019 and April 1st 2022. Finally, order the resulting table by the “Purchase\_Price” column with the highest value first.

|  |
| --- |
| # Assuming data type of Date column is Timestamp,  # if it’s not timestamp but date, just replace “CAST(Date AS date)” to “Date” instead  SELECT  Date, Purchase\_Price, Purchase\_Desc  FROM shop.history  WHERE CAST(Date AS date) BETWEEN  DATE(“2019-01-01”)  AND  DATE(“2019-04-01”)  ORDER BY Purchase\_Price DESC |

## Part 2 – Data Wrangling with SQL

For part 2 of this assignment you will need to use the boston\_crime.csv dataset. Make sure your data set id is boston, and the table name is crime (FROM boston.crime).

**Question 2.1:**

How many entries (rows) does this dataset contain?

|  |
| --- |
| 319073 |

**Question 2.2:**

How many unique offense codes are present within the data? Use the Group By command to find your answer. In the box below, please provide your answer to the question and the query used.

|  |
| --- |
| Unique offense codes: 425  Query:  SELECT  CODE, COUNT(CODE)  FROM boston.offense\_codes  GROUP BY CODE |

**Question 2.3:**

Find out how many OFFENSE\_DESCRIPTION entries contain the word “ASSAULT” as the first word?

|  |  |
| --- | --- |
| – e.g. ASSAULT - AGGRAVATED - BATTERY |  |

In the box below, please provide your answer to the question and the query used.

|  |
| --- |
| OFFENSE\_DESCRIPTION entries containing “Assault” as the first word:  Query:  # method 1: via Regex  SELECT \* FROM boston.crime  WHERE REGEXP\_CONTAINS(OFFENSE\_DESCRIPTION, "^ASSAULT");  # method 2: via LIKE operator  SELECT \* FROM boston.crime  WHERE OFFENSE\_DESCRIPTION LIKE "ASSAULT%"; |

**Question 2.4:**

Make a new column called TIME which contains the time of the offense from the OCCURRED\_ON\_DATE column. (Hint: you will need to use the CAST and SUBSTR functions together)

In the box below, please provide the query used as well as a screenshot of the query results containing the new TIME column. The column should look like the one in the Sample Screenshot below.

|  |
| --- |
| Query:  SELECT  \*, SUBSTR(CAST(OCCURRED\_ON\_DATE AS STRING),12,8) AS TIME  FROM boston.crime;  Screenshot: |

## 