Assignment: Exploratory Data Analysis with SQL

Estimated time needed: 60 minutes

Introduction

Using this R notebook you will perform exploratory data analysis using SQL queries with the RSQLite R package. You will be graded on the accuracy of your results as well as the content of your SQL queries.

Establish your SQL Lite connection

Install RSQLite package

The RSQlite package needs to be installed in your notebook. Let's load the RSQLite package by clicking on the following cell and executing it (Shift+Enter):

In [ ]:
```
install.packages("https://cran.r-project.org/src/contrib/Archive/RSQLite/RSQLite
```

Restart Kernel

After installing the RSQLite package, it is necessary to restart R Kernel. Click **Kernel** > **Restart Kernel** from the main menu. This will register the newly installed packages. Now proceed to load the RSQLite package.

Load RSQLite

Load the 'RSQLite' library, and use the 'dbConnect( )' function as you did in the previous labs to establish the connection to your SQLite database.
You are now ready to start running SQL queries using the RSQLite library as you did in Course 3.

In [ ]:
```
library("RSQLite")
```

In [ ]:
```
# provide your solution here to connect db
```

Download the following csv files:

- WORLD_CITIES
- BIKE_SHARING_SYSTEMS
- CITIES_WEATHER_FORECAST
- SEOUL_BIKE_SHARING

and load the csv's into 4 tables as mentioned below

- SEOUL_BIKE_SHARING
- CITIES_WEATHER_FORECAST
- BIKE_SHARING_SYSTEMS
- WORLD_CITIES

> Hint : Use the read_csv() function and dbWriteTable() functions

Task 1 - Record Count

Determine how many records are in the seoul_bike_sharing dataset.

Solution 1

In [ ]:
```
# provide your solution here
```

Task 2 - Operational Hours

Determine how many hours had non-zero rented bike count.

Solution 2

In [ ]:
```
# provide your solution here
```

Task 3 - Weather Outlook

Query the the weather forecast for Seoul over the next 3 hours.

Recall that the records in the CITIES_WEATHER_FORECAST dataset are 3 hours apart, so we just need the first record from the query.

Solution 3

In [ ]:
```
# provide your solution here
```

Task 4 - Seasons

Find which seasons are included in the seoul bike sharing dataset.

Solution 4

In [ ]:
```
# provide your solution here
```

Task 5 - Date Range

Find the first and last dates in the Seoul Bike Sharing dataset.

Solution 5

In [ ]:
```
# provide your solution here
```

Task 6 - Subquery - 'all-time high'

determine which date and hour had the most bike rentals.

Solution 6

In [ ]:
```
# provide your solution here
```

Task 7 - Hourly popularity and temperature by season

Determine the average hourly temperature and the average number of bike rentals per hour over each season. List the top ten results by average bike count.

Solution 7

In [ ]:
```
# provide your solution here
```

Task 8 - Rental Seasonality

Find the average hourly bike count during each season.

Also include the minimum, maximum, and standard deviation of the hourly bike count for each season.

> Hint : Use the SQRT(AVG(col*col) - AVG(col)*AVG(col)) function where col refers to your column name for finding the standard deviation

Solution 8

In [ ]:
```
# provide your solution here
```

Let's explore a bit and see what might be the most significant contributing factors in terms of the provided data.

Task 9 - Weather Seasonality

Consider the weather over each season. On average, what were the TEMPERATURE, HUMIDITY, WIND_SPEED, VISIBILITY, DEW_POINT_TEMPERATURE, SOLAR_RADIATION, RAINFALL, and SNOWFALL per season?

Include the average bike count as well , and rank the results by average bike count so you can see if it is correlated with the weather at all.

Solution 9

In [ ]:
```
# provide your solution here
```

Task 10 - Total Bike Count and City Info for Seoul

Use an implicit join across the WORLD_CITIES and the BIKE_SHARING_SYSTEMS tables to determine the total number of bikes avaialble in Seoul, plus the following city information about Seoul: CITY, COUNTRY, LAT, LON, POPULATION, in a single view.

Notice that in this case, the CITY column will work for the WORLD_CITIES table, but in general you would have to use the CITY_ASCII column.

Solution 10

In [ ]:
```
# provide your solution here
```

Task 11 - Find all city names and coordinates with comparable bike scale to Seoul's bike sharing system

Find all cities with total bike counts between 15000 and 20000. Return the city and country names, plus the coordinates (LAT, LNG), population, and number of bicycles for each city.

Later we will ask you to visualize these similar cities on leaflet, with some weather data.

Solution 11

In [ ]:
```
# provide your solution here
```

In [ ]:
```
close(conn)
```

Author(s)

Jeff Grossman

Lakshmi Holla

Other Contributor(s)

Malika Singla

In [ ]: