DSA 610, Homework #5, Spring 2025

Part I:
Instructions: Answer each discussion question in your own words. You may use posted resources or other online resources to answer the questions (**cite your sources**). Thoroughly explain your responses with a minimum of one paragraph (3-5 sentences) in length. Be thoughtful. We will discuss the answers in class.

1. Many database systems add additional commands to the top of SQL (we can think of them as SQL+) that are particular to those systems. Choose a particular database (such as PostgreSQL, MySQL, etc.) and describe some of these additional commands. What do they do?

2. Describe a potential use case for a cross join.

3. Aggregating data is often used for privacy protection in big data. Describe the advantages and disadvantages of using aggregated data.

4. Joining tables in a database reverses the normalization process we did when we created the database initially. Why would we want to join datasets outside the context of a database setting? Give a use case.

5. As an analyst, most of the time we will be working with a subset of SQL called DQL (Data Query Language). Describe what the other components do (DDL, DML, DCL, TCL) and what kind of data specialist is most likely to use them.

Part II:
Instructions: Use the attached dataset (**health_data_oob.xlsx**) to complete the following tasks in Python. Report your answers to the questions on this homework sheet. Include your Jupyter notebook file along with your homework submission, saved as a PDF. Make sure that any graphs you create are quality graphs with a legend (as needed), axis titles, a descriptive title, appropriate ranges (bar graphs start at 0, etc.). They should be able to stand on their own. You may need to relabel some elements (such as replacing 0s and 1s with string labels).

Use Python to answer the following:

1. Create a pairplot of the data, and summary statistics of each variable. Use this to identify the special codes used in each variable to mark missing data. Replace these codes with NAs. Note: the Blood Pressure variable has been converted to a percentile already, and it is the only useful variable that does not have any missing values. Create a pairplot and summary statistics at the end to confirm all anomalous values have been removed.

2. We modeled imputing the Age variable in lecture. Use that as a model for creating imputations for the Income and Exercise variables.

3. Following the examples in class, assess the impact of the imputation methods used. Which one seems to be the best fit for each variable (Note: I expect that different methods will work better on different variables, so clearly explain your choice for each). Use graphs and various numerical metrics to justify your choice.

4. Think back to the MBA dataset we worked with in project #1 and the previous homework. Which variables could make sense to create dummy variables of and which would not (you can just use the MBA sheet here). You don't need to code anything up, but explain your choices, and how many variables would result from coding up each one.

5. Reload the health dataset from scratch. Use SQL code in Python to illustrate how you can filter or aggregate data. Create a dummy variable for the missing Income data as we did in class for Age. Assess the other variables, including Blood Pressure, based on whether they are missing from Income or not. Are there appreciable differences between those that are missing and those that are not? Provide numerical evidence for you assessment.

Part III:
Work through the tutorials on the site (https://sqlbolt.com/). They are interactive and will give you a chance to practice writing SQL queries on a database about movies. Work through the first 11 lessons and in your homework submission, include screen shots of the completed tasks (the Continue button will be Green when you completed all the tasks on that page). Feel free to do more if you like.