DSA 610, Homework #3, 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. When would you choose a relational database over a non-relational database, and vice-versa? Consider factors like data structure, scalability needs, query complexity, and the type of analysis you intend to perform. Give specific examples of use cases for each.

2. How can relational and non-relational databases be used together in a modern data architecture? Discuss the concept of polyglot persistence and its implications.

3. Discuss the key considerations when choosing a graph type for data visualization. How do the type of data (e.g., categorical, numerical, time series), the message you want to convey, and the audience influence your choice?

4. What are the key properties of a well-normalized database? How does normalization improve data integrity, reduce redundancy, and simplify data maintenance? Discuss the trade-offs involved in database normalization. What are the potential performance implications of highly normalized databases, and how can these be mitigated?

5. What are the ethical considerations involved in data cleaning and transformation? How can we ensure that these processes do not introduce bias or distort the data in ways that could lead to misleading conclusions?

Part II:
Instructions: Use the attached dataset (**beer_preference_data_hw1.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 scatterplot for salary vs. age with (at least two) different plotting packages. Create variations with color-coding by gender, or marital status. Apply a regression line to your graph. Then make a jointplot of the same data.

2. Make a histogram of salary. What is the shape of the distribution?

3. Create comparative boxplot of age by marital status. Does marital status appear to affect the age?

4. Using the gapminder dataset or the mpg dataset we imported for examples in the lecture. Use the data from one of these datasets to experiment with graphs we created in class, such as

contour plots, 3D graphs, pairplots, or maps. You can use the examples we did with a different dataset than what we used in the example in class modify the graph we made with different variables and coloring schemes.

5. Create a word cloud of your choice of text data. It can be a website you scrape from, or a text file you upload from your computer. You can use AI to help you modify our code examples as needed. What does the word cloud tell you about the semantic content of the text?

6. Using the karate club data we plotted in class, starting from one of the examples, modify the graph (change the labels, color, layout, etc.). AI or the package documentation can assist you in making changes. Did any of your changes make the graph more or less readable? Why or why not?