DSA 610, Homework #4, 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 prefer to use a database system to manage your data versus just a spreadsheet? Or vice versa? What are the strengths and weaknesses of each?

2. Select a specific type of non-relational database and describe how it is different from a relational database. What are the advantages and disadvantages of each? Give a scenario where the non-relational type would be preferred.

3. What are some methods we can use for dealing with missing values? Describe at least three different strategies. When is each method appropriate?

4. What are some strategies for identifying and dealing with outliers? Why might we not want to remove them? Why might we want to remove them?

5. What are some of the challenges data analysts face when dealing with datetimes? Give some specific examples.

Part II:
Instructions: Use the attached dataset (**BW_MBAdata.xlsx** that we also used for Project #1) 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. The final python example from lecture provided you code for downloading and importing into Python a dataset on air quality that is a time series. The code also creates the time-based data that you need. Choose one of the other variables in the dataset and make a line graph of the data.

2. Continuing with the same variable from the air quality data, perform the following tasks:
   a) Decompose the time series into trend, seasonal and noise.
   b) Calculate the first difference. Is the original data stationary? Is the differenced data stationary? If the answer to both is no, calculate one more difference and check that also.
   c) Create an ACF and PACF graph.

3. Create a SARIMA model of the time series. You may need to experiment with parameters to determine the best parameters for your model. Plot the results.

4. Using the MBA data (the same that we used in Project #1), which variables are missing values? How many values are missing? Why might we not want to drop the data from the dataset entirely?

5. Choose one of the variables with missing values. Impute the missing values in the dataset using mean, median, regression and one other method (explain what it is and why you chose it). What are the results of the imputation and the impact on the variable? Compare summary statistics before and after. What do you notice?

6. Consider the average starting base salary (of the MBA Data sheet). Identify any outliers. How would you handle them in this data set? Rescale or transform the variables? Remove them? Some other method? Explain your reasoning. Compare the results before and after your treatment. Did it improve the situation or no change?

Part 3:
Use the attached file (MedicalSummaryForm.docx). Create a sample conceptual, and then a logical model for a database needed to store the data collected by this form. Think carefully about the entities and relationships. Identify any potential normalization issues. Identify any primary and foreign keys. You can use an online tool for your diagram. For each Field you identify, what kind of data type is it?