UNIT 2 ASSIGNMENT

Managing Your Data in ML

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities, exercises, and coding portion of the assignment. Now, let’s formulate your programming into well-thought responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Week 2 Written Portion

Building a Modeling Dataset

Answer the questions below about building a model dataset and understanding your data through analysis and visualization.

## Questions:

1. What does it mean to have a “modeling dataset”?

|  |
| --- |
| It means having a dataset for machine learning which is clean, consistent, relevant, and large data. |

1. What steps would you take with a raw dataset to end up with a modeling dataset?

|  |
| --- |
| First, I need to understand the data and purpose. By doing so, I could determine label and features. Then remove or replace null values and handle outliers. Find the most correlated features and I might need to split or manipulate the data for better prediction. |

1. What is the difference between nominal data and ordinal data? Explain with an example.

|  |
| --- |
| Both represents categories but ordinal data has orders/rankings. For example, categories of fruits are nominal data because the order does not matter, however, academic degrees are ordinal data because it has an order. |

1. Why is data visualization an important part of the data preparation process?

|  |
| --- |
| Data visualization makes it easier to identify patterns and understand the relationships between features or labels. |

1. What is an outlier?

|  |
| --- |
| Outlier is a data point that differs significantly from other observations so that it can affect the correctness of the prediction. |

1. Name a few libraries used for data analysis and visualization and explain when you would use each library.

|  |
| --- |
| Seaborn: for 2D plotting and finding relationships between columns  Scikit-learn: for finding z-score or Winsorizing outliers |

*To submit this assignment, please refer to the instructions in the course*.