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”?

|  |
| --- |
| A modeling dataset is one that is prepared properly for training a model. It comes from raw data that is then processed appropriately. It usually consists of a dependent variable that the model will try to predict or classify based on the independent data given to it. |

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

|  |
| --- |
| In order to arrive at a modeling dataset from a raw dataset you must first gather the data and handle any missing data. You need to clean and process the data by handing outliers, transforming it when necessary, and removing any points that do not meet the characteristics of the data set. Next, you need to split the data so that a portion of it can be used for training, another for validation, and lastly one for testing. Once the data is properly organized, transformed, and split you have a modeling dataset that you can work with to train your model. |

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

|  |
| --- |
| Nominal data has categories with no order or rank to them. The categories are instead qualitative and distinct from one another. One example of this is eye color, with the categories being ‘brown, blue, hazel’ and so on. Ordinal data, on the other hand, represents categories with a distinct ranking. An example of this could be a job position, starting from ‘entry level’ and ranging to ‘senior’. Therefore, the data has an inherent order, unlike nominal data. |

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

|  |
| --- |
| Data visualization through graphs and plots is important to the data preparation process because it makes it easier to identify patterns, trends, and outliers in large data sets. It is not possible to fully understand large data sets by purely looking at a table of data, and trends are more visible when you are able to look at the big picture instead of pieces of information at a time. An example of this would be a large set of data showing the population of people in every 1000 sq ft of the United States, A visual representation of the population density through the intensity of a certain color would make it easily visible what areas of the country have the largest and smallest populations. |

1. What is an outlier?

|  |
| --- |
| An outlier is a statistical anomaly that does not follow the general trend of a set of data. For example, if data points are generally following a linear regression, the outlier would be a point far from the line of points. Because it is not statistically significant, it can often be ignored. |

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

|  |
| --- |
| A few libraries used for data analysis are NumPy, Pandas, and Matplotlib. NumPy allows for heavy computation. With it, you can work with large arrays and matrices and solve complex systems and equations. Pandas allow you to create data structures such as DataFrames and Series so that you can filter, transform, and further manipulate data. Matplotlib is a plotting and graphing tool and is often used to create visual representations of equations and data so you can further analyze it. |

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