**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 project 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 project 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.

*Complete each project part as you progress through the course. Wait to submit the project until all parts are complete. Begin your course project by completing Part One below. A submit button can be found on the final Course Project assignment page. Information about the grading rubric is available on any of the course project assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the project.*

**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”?

|  |
| --- |
| Data can exist without an intention to perform machine learning tasks. So when we say “modeling dataset”, we are referring to data that has been specifically prepared for use to train or build a machine learning model. |

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

|  |
| --- |
| This data would be designed with specific properties in mind, such as being structured with individual features, a label and an assumption of independence of the rows. Meeting these design specifications can be done with the right sampling methodology and feature engineering. |

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

|  |
| --- |
| These are both categorical data types but have a key distinction. Nominal data represents discrete categories with no sense of ordering, while ordinal data has an ordering. An example of a nominal data point could be “colors of the rainbow,” which 7 discrete values. An example ordinal data point could be the number of stars given to an Amazon review. In the latter, there are 5 categories (1-5) with relative levels of importance. |

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

|  |
| --- |
| Data visualization enables us to identify complex patterns in data that individual statistics do not represent. Data visualization is an effective way to share insights with stakeholders. |

**5.** What is an outlier?

|  |
| --- |
| An outlier is a data point with extreme values that does not sit within the limits of an expected distribution. Outliers can be automatically detected but usually require some subjective judgement about where the outlier boundaries should be. |

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

|  |
| --- |
| Matplotlib is a core Python plotting library that enables a lot of customization and control over the graphics. Seaborn is a wrapper library over Matplotlib that provides an easier interface to work with. Other libraries like Altair and Plotly also exist. Each is designed with a unique focus, such as the simplicity of the APIs or the ability to make interactive plots. |

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