# Assignment-2: Data Wrangling and Regression Analysis

## Section A: Data Wrangling

### 1. What is the primary objective of data wrangling? A) Data visualization B) Data cleaning and transformation C) Statistical analysis D) Machine learning modeling

Answer: Option B is the right answer as data wrangling deals with the data cleaning and in the transformation of raw data into the data that is suitable for analysis.

### 2. Explain the technique used to convert categorical data into numerical data. How does it help in data analysis?

Answer: Label(ordinal) encoding is a technique used to convert the categorical data into numerical data which is the best technique for the data in a certain order so that it could help us in having the similar data to be identified in a easy manner in the analyzing of data.

## 3. How does LabelEncoding differ from OneHotEncoding?

Answer: Label(ordinal) encoding is a technique used to convert the categorical data into numerical data which is best for the data that is in certain order whereas one hot encoding(not ordinal) also converts each categorical data into numerical data but converts into a binary form either 0 or 1 allocated as the 2 categories seperating the single column into certain count of columns depending on the number of rows containing the datasets.

## 4. Describe a commonly used method for detecting outliers in a dataset. Why is it important to identify outliers?

Answer: a.Normal distribution method:This method is used when the data is normal symmetric. b.Inter quartile range :It is the commonly used technique for detecting the outliers in a dataset and it is important to identify the outliers because it might not execute the code as the since data is not coded correctly which is not in the range that is to be present and we use it when our data is skew. c.Boxplot :This method is used to detect the outliers but in a pictorical form.

## 5. Explain how outliers are handled using the Quantile Method.

Answer: We handle with the outliers in the following ways. 1.Trimming: Remove the outliers and we make use of it when the data is sufficient. 2.Capping: Replacing with the extreme values and we make use of it when the data is normal. 3.Treating outliers as missing values and we make use of it when the data is non-normal.

## 6. Discuss the significance of a Box Plot in data analysis. How does it aid in identifying potential outliers?

Answer: Box plots helps us identify outliers where values below Q1-1.5(IQR) and above Q2+1.5(IQR) are considered as the outliers. These values are plotted as data points and fall beyond the boxes above and below showing that they exceed in the range in this way it aid as the identifying potential outliers.

# Section B: Regression Analysis

## 7. What type of regression is employed when predicting a continuous target variable?

Answer: Linear regression is the type of regreesion we consider when predicting a continuous target variable.

## 8. Identify and explain the two main types of regression?

Answer: Linear regression and logistic regression are two types of regression techniques.The linear regression tells about the relationship between feature variable which also be called as an independent varaiable and a dependent variable where both are related to each other linearly. Logistic regression deals with the feature variables(input) and target variables(output) to be related in binary values such as 0 or 1, true or false, etc. This means the target variable can have only two values.

## 9. When would you use Simple Linear Regression? Provide an example scenario. ¶

Answer: Simple linear regression is used to estimate the relationship between two quantitative variables. As an example we can consider how the height of the student and the salary of the parent related to the health of the student where height deals with the health but not the salary.

## 10. In Multi Linear Regression, how many independent variables are typically involved?

Answer: In multi linear regression two or more independent variables are involved.

## 11. When should Polynomial Regression be utilized? Provide a scenario where Polynomial Regression would be preferable over Simple Linear Regression.

Answer: A polynomial regression should be utilized when fitting the non-linear relationship between variables using a non-linear regression line which is not possible with simple linear regression which means that polynomial regression can perform more complex phenomena than linear regression and it is preferable over simple linear regression when the relationship between variables includes curvature which can be called as a scenerio.

## 12. What does a higher degree polynomial represent in Polynomial Regression? How does it affect the model's complexity?

Answer: Higher degree polynomial representation of the polynomial regression basically fits the data better as it increases the curveness of the best fit line where it's complexity gets increased along with the performance.

## 13. Highlight the key difference between Multi Linear Regression and Polynomial Regression.

Answer: Multiple linear regression has multiple independent variables which shows the dependent variable is considered as a linear function of multiple independent variables whereas in polynomial regression the dependent variable is considered as a linear function of the single independent variable where each power of variable demonstrates it as a different variable making each dependent on each other.

## 14. Explain the scenario in which Multi Linear Regression is the most appropriate regression technique.

Answer: We consider a situation in which a multilinear regression is considered as the most appropriate regression technique when the dependent variable is rarely depending on only one variable and in most cases depending on various independent variables.

## 15. What is the primary goal of regression analysis?

Answer: The primary goal of regression analysis is about predicting the value of a dependent variable given the values of one or more independent variables as predictions establishes the relation between the independent and dependent variables.