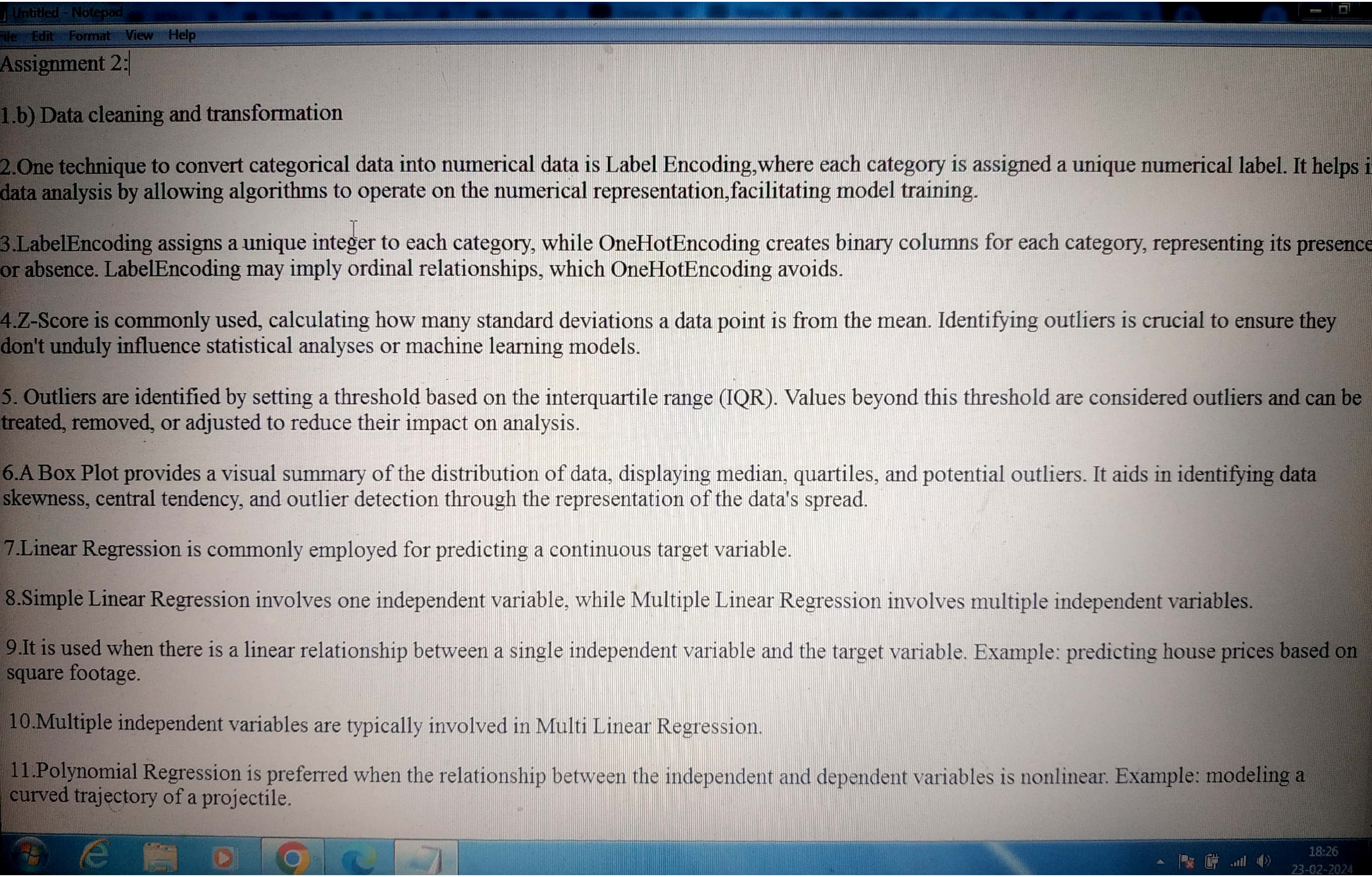


Assignment 2:

1.b) Data cleaning and transformation

2.One technique to convert categorical data into numerical data is Label Encoding,where each category is assigned a unique numerical label. It helps i data analysis by allowing algorithms to operate on the numerical representation,facilitating model training.

3.LabelEncoding assigns a unique integer to each category, while OneHotEncoding creates binary columns for each category, representing its presence or absence. LabelEncoding may imply ordinal relationships, which OneHotEncoding avoids.

4.Z-Score is commonly used, calculating how many standard deviations a data point is from the mean. Identifying outliers is crucial to ensure they don't unduly influence statistical analyses or machine learning models.

5. Outliers are identified by setting a threshold based on the interquartile range (IQR). Values beyond this threshold are considered outliers and can be treated, removed, or adjusted to reduce their impact on analysis.

6.A Box Plot provides a visual summary of the distribution of data, displaying median, quartiles, and potential outliers. It aids in identifying data skewness, central tendency, and outlier detection through the representation of the data's spread.

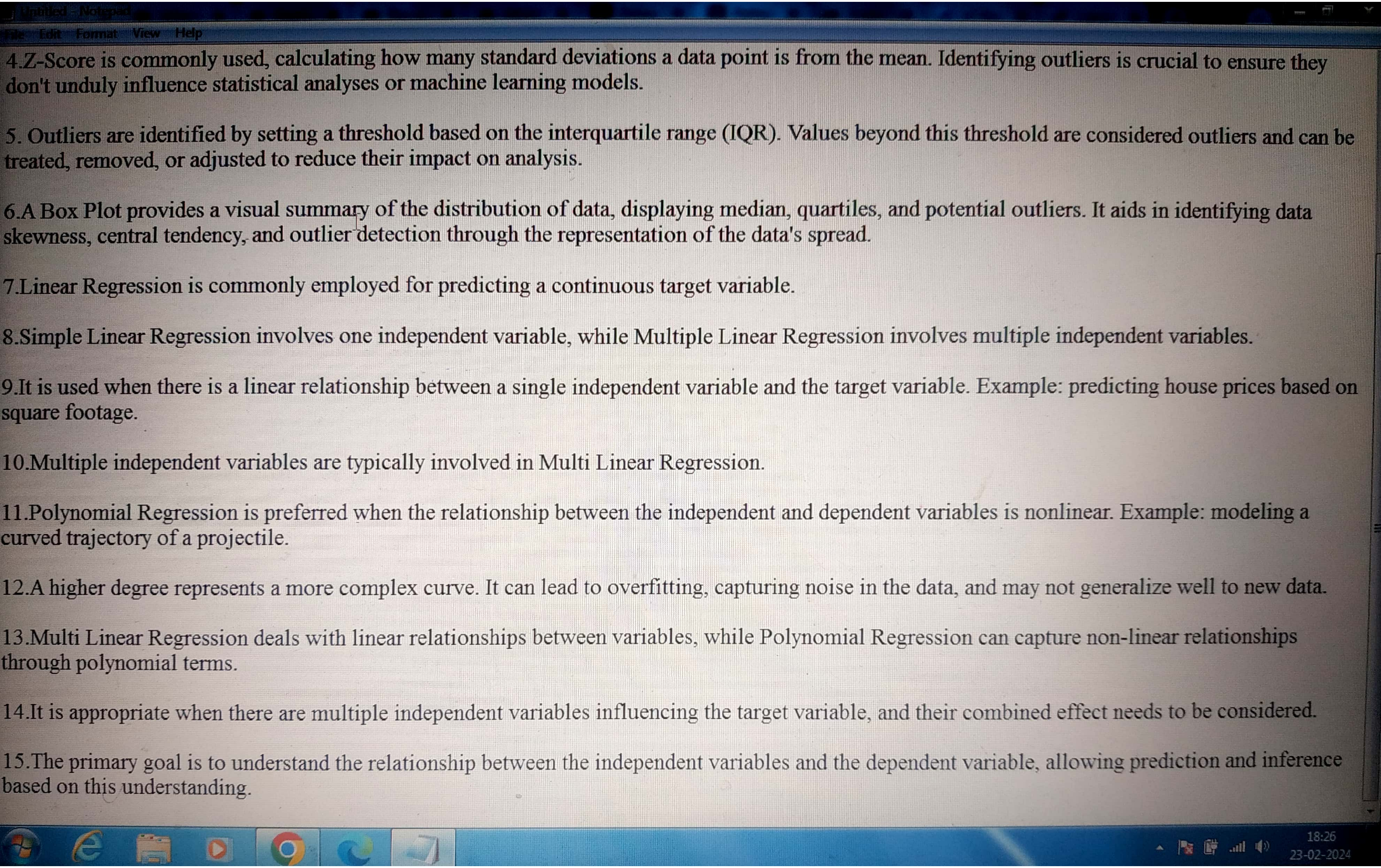
7.Linear Regression is commonly employed for predicting a continuous target variable.

8.Simple Linear Regression involves one independent variable, while Multiple Linear Regression involves multiple independent variables.

9.It is used when there is a linear relationship between a single independent variable and the target variable. Example: predicting house prices based on square footage.

10.Multiple independent variables are typically involved in Multi Linear Regression.

11.Polynomial Regression is preferred when the relationship between the independent and dependent variables is nonlinear. Example: modeling a curved trajectory of a projectile.



4.Z-Score is commonly used, calculating how many standard deviations a data point is from the mean. Identifying outliers is crucial to ensure they don't unduly influence statistical analyses or machine learning models.

5. Outliers are identified by setting a threshold based on the interquartile range (IQR). Values beyond this threshold are considered outliers and can be treated, removed, or adjusted to reduce their impact on analysis.

6.A Box Plot provides a visual summary of the distribution of data, displaying median, quartiles, and potential outliers. It aids in identifying data skewness, central tendency, and outlier detection through the representation of the data's spread.

7.Linear Regression is commonly employed for predicting a continuous target variable.

8.Simple Linear Regression involves one independent variable, while Multiple Linear Regression involves multiple independent variables.

9.It is used when there is a linear relationship between a single independent variable and the target variable. Example: predicting house prices based on square footage.

10.Multiple independent variables are typically involved in Multi Linear Regression.

11.Polynomial Regression is preferred when the relationship between the independent and dependent variables is nonlinear. Example: modeling a curved trajectory of a projectile.

12.A higher degree represents a more complex curve. It can lead to overfitting, capturing noise in the data, and may not generalize well to new data.

13.Multi Linear Regression deals with linear relationships between variables, while Polynomial Regression can capture non-linear relationships through polynomial terms.

14.It is appropriate when there are multiple independent variables influencing the target variable, and their combined effect needs to be considered.

15.The primary goal is to understand the relationship between the independent variables and the dependent variable, allowing prediction and inference based on this understanding.