Jordan Black

CGS 2020 202501

Dr. Lakshmi Prayaga

2/15/2025

**Module 3 Assignment**

1. **Give three examples of data that can be viewed as Big data:**

**Big Data refers to enormous data sets that are complex and challenging to process using traditional data-processing tools.**

**Retail and E-commerce Data:**

 Corporations like Amazon gather extensive data about customer purchase patterns, website interactions, and product evaluations. The data enables companies to refine supply chain operations and deliver personalized recommendations.

**Healthcare Data**: Medical records, genomic sequences, and real-time health monitoring produce enormous quantities of data. Analyzing this data enables the creation of personalized treatments and elevates the standard of patient care.

**Financial Services Data**: Financial institutions and banks run extensive analyses on transaction records, market feeds, and customer profiles. Financial institutions rely on transactional and market data to protect against fraud and manage risks while delivering personalized banking services.

1. **What is the difference between supervised and unsupervised learning?**

**Supervised Learning:** During training on labeled data, the model associates each input with its corresponding correct output. Classification algorithms recognize spam emails, while regression algorithms estimate property values.

**Unsupervised Learning:** The model receives unlabeled data to discover patterns or structures autonomously without specific instructions. Clustering applied to customer segmentation alongside dimensionality reduction through PCA for feature extraction provides examples of unsupervised learning techniques.

**For the questions below use the PPT for data scrubbing**

1. **What should the missing values be filled with?**

Appropriate values based on the data type and context must fill in missing values.

Mean or Median: The mean or median should be applied to numerical data when the distribution is symmetrical. Use the median for skewed data.  
  
Mode: When working with categorical data, we replace missing values with the most common value, which is known as the mode.  
  
The provided R script requires you to substitute the missing **Solar. R** data points with the column's average value.

1. **Would normalization be sufficient for scaling the temperature values? Why or why not?**

**Normalization** transforms data into a specific range, usually between 0 and 1. It benefits scenarios with consistent value ranges but becomes problematic when outliers exist. Normalization fails to produce accurate results when temperature data includes outliers because these extreme values will warp the range**.**

**Standardization** through Z-score normalization becomes the better choice in these circumstances because it adjusts the data to achieve a mean of 0 and a standard deviation of 1, decreasing the impact of outliers.

1. **For interpretation by most algorithms, how should you modify the weather column?**

For most machine learning algorithms, the "weather" column needs to be transformed from categorical data such as "sunny," "rainy," and "cloudy" into numerical values.

**One-Hot Encoding:** Transform each weather category into binary columns that represent the presence or absence of that category (e.g., "sunny" = 1 while "not sunny" = 0).

**Label Encoding**: Each category should be given its distinct numerical identifier (e.g., "sunny" = 1, "rainy" = 2, "cloudy" = 3).  
  
Algorithms requiring non-ordinal input data benefit from One-Hot Encoding because it transforms categorical variables into binary columns.

**6. Use the r script I provided and add the following**

[R CODE ASSIGNMENTs\Updated\_CGS2020\_Module3 (1).R](R%20CODE%20ASSIGNMENTs/Updated_CGS2020_Module3%20(1).R)

1. **Remove the NA values from the Solar. R column and replace them with the mean values of the column:**

**R: CODE**

**#replace NA values in Solar.R with the mean**

**Air1$Solar.R [is.na(air1$Solar.R)] <- mean(air1$Solar.R, na.rm = TRUE)**

**b. Replace numbers 6,7,8 and 9 from the month column to July, August, September and November as we did to replace 5 with May**

**R CODE**

**b. #replace numbers in the Month column, with actual month names**

**air$Month <- gsub(6, “June”, air1$Month)**

**air$Month <- gsub(7, “July”, air1$Month)**

**air$Month <-gsub(8, “August”, air1$Month)**

**air$Month <- gsub(9, “September”, air1$Month)**

A computer code with black text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Print a summary of the air1 file, and submit  the screenshot of the summary:**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Screenshot of the Code:A screenshot of a computer

AI-generated content may be incorrect.**