**Instructions:**

**You can use Word, Excel, Power Point, R and/or Python to answer the questions in this exam. There are a total of five (5) multi-part questions, with point values noted for each question. You must use Excel if it is specified in the problem statement.**

**Please show your calculations, or the details of your program(s) for each problem. You must supply the R/Python programs, and the programs should be commented so that each step is clearly explained.**

**Combine all of your answers/files into a single zipped file and post the zipped file to CANVAS.**

**#1** (10 Points)

**Is the following function a proper distance function? Why? Explain your answer. Measure the distance between (0, 0, 0) and (0, 1, 0)**

**#2** (15 Points)

**An employee of a company is traveling to either England, Italy, or Spain. The employee can travel to only one country. There is a 50% chance the employee will go to England and a 20% chance to Italy. Assume the chances of contracting COVID to be proportional to the prevalence of the disease in each country, given in the table below. For example, the chances of contracting COVID in England is 1200/1,000,000.**

|  |  |
| --- | --- |
|  | **Prevalence** |
|  | Cases |
|  | **Per Million** |
| **England** | 1200 |
| **Italy** | 1400 |
| **Spain** | 1800 |

**What are the chances that the employee will contract COVID while travelling?**

**Assume that the employee has traveled to Europe and contracted COVID, what is the probability that he/she traveled to England?**

**#3** (20 Points)

**Load the “IBM\_attrition\_v3B” dataset, from the raw\_data module in CANVAS, into R/Python. This is a fictional, IBM supplied, employee data set. Perform the EDA analysis by:**

1. **Summarizing each column (e.g., min, max, mean)**
2. **Identifying missing values**
3. **Replacing the numerical missing values with the “mode” of the corresponding columns**
4. **Displaying the scatter plot of “Age”, “MonthlyIncome” and “YearsAtCompany”, one pair at a time**
5. **Showing box plots for columns: “Age”, “MonthlyIncome”, and “YearsAtCompany”**

**#4** (25 Points)

**Use Excel and the “****IBM\_Attrition\_missingB.csv” file (Excel file containing another variation of the fictional employee dataset) to replace the missing “MonthlyIncome” using knn (k=3).**

**#5 (**30 Points)

**Load the CANVAS “IBM\_attrition\_v2B.CSV” dataset into R/Python. Remove the missing values. Construct a knn (k=3) model to classify attrition (attrition=’yes’) based on the other variables. Predict attrition for a random sample (30%) of the data (test dataset). Measure the accuracy, precision, recall and F1 of the model.**

**Hint (see ‘ifelse’ function in R)**

**Dataset dependency: IBM\_attrition\_V2B, IBM\_attrition\_v3B, IBM\_Attrition\_missing.csvB**