**Assignment 1 (To be completed individually) /17**

**CSCI 528**

**Name: Sahil Sajan**

**Due:** Sept 26, 2024 (11:59PM) on Moodle  
**Instructions**: Answer each of the following questions and submit to the course’s Moodle page. You can answer the questions on MS Word or print the assignment, complete it with a pencil, and then upload a scanned copy of the assignment.

1. **(2 points)** For each given task below, label them as an example of classification, clustering, association rule mining, or an anomaly detection task. (hint: there is one example of each task type).
   1. Predict which car will sell to a given customer  
      Classification\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. Identify sharks that change their behaviour when near other sharks  
      Association rule mining\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   3. Detect sharks that have an abnormal speed  
      Anomaly detection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   4. Group together all similar vehicles

Clustering\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **(4 points)** Classify the following examples as nominal, ordinal, interval, or ratio. Some cases may have more than one interpretation, so briefly indicate your reasoning. Your answer should include three different values as seen in the example. You **must** describe why you classified each attribute with what you selected. Classify each one based on the type that they can be with the most properties, for example if a data type could be ordinal, then label it as ordinal and not nominal.

Example:

Number of students in a course

(**ratio**)

My reasoning would then go here for why I chose ratio.

* 1. Number of courses registered by a student in a given semester.  
     (ratio)  
     You can count the number of courses that a student has enrolled in, additionally zero courses here means the student enrolled in no courses so there is a meaningful zero point. We can also compare 2 or more students based on the number of courses they took (a student who enrolled in 2 courses took twice the number of courses that a student who only enrolled in 1 course took).
  2. Temperature in Kelvin.  
     (ratio)

Temperature in Kelvin falls under ratio attribute as Kelvin has an absolute zero point (0 kelvin means absence of heat altogether). Also, we can compare 2 different temperatures which are measured in Kelvin (200 kelvin is twice as hot as 100 kelvin).

* 1. Temperature in Celsius  
     (interval)

Celsius, unlike in Kelvin, does not denote absolute absence of heat when we state 0 degrees Celsius. This is the reason that although we can compare differences between 2 temperature points in Celsius, we cannot state that 200 degrees Celsius is twice as hot as 100 degrees Celsius as it does not scale off absolute zero. This is why it falls under interval attributes and not ratio.

* 1. Rank in the military.  
     (ordinal)

A rank in military is an ordinal attribute as there is a clear order in the ranking. We know which rank is greater or lesser than the other, however, we cannot say how much greater one rank is than the other. That’s what makes the attribute ordinal.

* 1. Social security number.  
     (nominal)

Social security number is just a number where the value of the number does not matter. It is simply used for identification. That makes this attribute nominal.

* 1. Movie ratings provided by users (bad, ok, good, great).   
     (ordinal)

These are not numeric ratings so we rule out interval and ratio. We have a ranking of the ratings where great is better than good which is better than ok and bad ranks the lowest. However, we cannot quantify how much better one rank is than the other, so it is ordinal.

* 1. The items on someone’s Wishlist.  
     (nominal)

The items on a wishlist are only names of products which have no order or ranking between them. This is why they are nominal attributes.

* 1. Speed of a car  
     (ratio)

Speed of a car has an absolute 0 (0 kmph, for example, means no movement at all). Additionally, we can compare 2 or more speed values and say that 40kmph is twice as fast as 20kmph. This means it is a ratio attribute.

1. **(1 point)** What is an example, where using identification numbers of something can assist with predictions? In other words, can we get any additional information from only looking at the identification number of a person or an item. (hint: Look at your own student ID). Be descriptive.  
     
   Identification numbers can help us in various areas with predicting useful data. The StFX student ID, for example, has our student number which begins with our year of enrollment. For example, my ID number, 202405476 indicates that I have enrolled in the year 2024. Using this information from my ID card, we can already get additional information such as my year of enrollment. We can use this information to get a predicted graduation year if we know the ID card belongs to a student in the master’s program so he is expected to graduate 2 years from his date of enrollment which will be 2026 in this example.
2. (**2 points**) On the course’s Moodle page, there is a file called day.csv, which contains information regarding bike sharing data. The **csv**, refers to comma-separated values and is a format to store information. You can open this file in MS Excel or any text editor. This dataset was originally from https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset. Answer the following questions regarding this dataset (day.csv).  
   1. How many attributes are in represented in this dataset, including the ID?  
      There are 16 attributes in the dataset (represented by the number of columns).
   2. How many objects/records/samples/instances are in this dataset? (objects, records, samples, and instances all refer to the same thing here).  
      There are 731 objects in the dataset (represented by the number of rows).
3. **(8 points**) Complete and submit the python script **analyze\_bike\_sharing.py** that reads in the file **day.csv** and includes the following functions:  
   1. **count\_samples(my\_data)**  
      This function counts the number of instances (samples) in the dataset and returns the number of instances. Assume each row in the dataset is a single instance. The my\_data parameter in this and all functions will be the first item of the tuple returned by **read\_data(file\_path).**
   2. **get\_min\_max(my\_data, column\_idx)**   
      This function gets the min value and max value of a column that is referenced by the index column\_idx**.** This value should return a tuple (min\_value, max\_value)
   3. **get\_mean\_median\_standard\_deviation(my\_datam, column\_idx)**  
      This function returns the a tuple that contains the (mean, median, and sample standard deviation) of the column referenced by column\_idx.
   4. **get\_most\_similar\_sample(my\_data, sample)**  
      This function finds the sample the is the most similar to the given sample (ignoring the data attribute). Use cosine similarity to measure the similarity.
   5. **get\_pearson\_correlation(my\_data, attr\_a\_idx, attr\_b\_idx)**  
      This function calculates the Pearson correlation between two attributes, given by their index numbers (attr\_a\_idx, attr\_b\_idx).
   6. **find\_attribute\_with\_largest\_correlation(my\_data, my\_data\_labels, candidate\_attrs, target\_attr)**This function finds the attribute from a list of candidate attributes (candidate\_attrs) that has the largest absolute Pearson correlation with the target attribute (target\_attr). My\_data\_labels is returned from **read\_data(file\_path)** as the second item of the tuple and is a dictionary that maps the name of an attribute to it’s attribute index.

The function **read\_data(file\_path)** is given to you to read in the dataset and store it in a dictionary. The keys in the dictionary are the instance ids and each value is a list of the attributes’ values for the instance. The index for the attributes starts at 0 and refers to the date attribute. All attributes are in order as seen in the original dataset. You may **not** modify the **read\_data(file\_path)** function.

There are checks in the test portion of the code. You will receive 1 point for each check that you get correct.

# IMPORTANT for your code!

* To run the code, you will need to run from command line:
  + python .\analyze\_bike\_sharing.py .\day.csv
* I will run the code in the same manner and therefore, it must work this way.
* The code submission must be a .py file and you don’t need to submit the data.csv file.
* You may **NOT** import any additional libraries or packages that are not already imported with the starter code.
* If you import additional libraries, your code will automatically be marked out of 50% and all code that uses the library will be marked as incorrect.
* Any attempt to modify the declaration of functions, such as the parameters that it uses, will be marked as incorrect.
* Any modification to the test code (all code below the line ###TESTS) will result in a **0** on the entire coding assignment.
* Code that can’t be run due to a syntax error will be marked out of 50%. If you are unable to get a function working, have it return a 0 for all expected int values and “none” for all expected string values. This should make your code not give a syntax error and not be marked out of 50% because of it.
* Your submitted code must **NOT** contain any additional print statements than what was given in the starter code.
* Have fun with the assignment and the data!