**George Brown College**

**Centre for Arts, Design and Information Technology**

**Machine Learning Process – Working with pandas (Assignment 03 – Lab Exercise)**

Your Name:

Your Student ID:

The goal of this lab-exercise is to practice with pandas functions.

Please follow the following steps using the Jupyter Notebook or any other environment to implement steps. At the end you need to present the codes to the instructor and upload your ipynb file to designated dropbox.

Please note that this is an individual lab-exercise, and everyone needs to finish the lab-exercise and submit their tasks individually.

The individual submission is mandatory, while everyone will open and run their assignments in the class to demonstrate what they have learned from it.

**Let’s start the lab-exercise:**

1. Download the csv file from the following URL and store it in a DataFrame named, myDF.

<http://tiny.cc/lab-exercise-01>

1. Show the first five rows of the DF.
2. Does the first row contain the column names? If not, please assign the following column names to them:

[ID, Range, Value, Machine Type, Zip code]

1. Add your name and student number on top of your dataset in your local file and show your DF with your name and student number as part of the dataset.
2. Now, use a method to skip the first row and show the DF as it was before adding your information.
3. Show the type of your DF.
4. Select the ‘Value’ column using two different commands. Which one is more object-oriented style?
5. What is the type of this selected column?
6. Can we apply both methods (mentioned in question 7) for fetching the ‘Zip code’ column? Which other column has the same status? Why we cannot use both method for fetching its content?
7. Create a new Series called ‘Brand’ by combining 'Machine Type' and 'Zip code' Series.
8. Create a new Series called ‘Total’ by multiplying ‘Range’ and ‘Value’ Series.
9. Use the following commands:
   1. myDF.head()
   2. myDF.describe()
   3. myDF.describe
   4. myDF.shape
   5. myDF.dtypes

Why some have parentheses and some don’t? What is the difference between b. and c.?

1. How can we use the describe() method to show the statistics for the object (not numerical) columns?
2. Rename the ‘ID’ to ‘My ID’ in the DF without replacing the ID Series name in the original DF. Show the unchanged DF by displaying the top 5 rows.
3. Rename the ‘ID’ to ‘My ID’ **inplace** and show the changed column name using the head() method.
4. Show the DF columns.
5. Change all the column names to ['id', 'range', 'value', 'machine type', 'zip code', 'brand', 'total'] and display the new column names.
6. Replace all the spaces in the column names with ‘\_’ and show the changed column again.
7. Delete the 'brand', 'total' columns permanently.
8. Delete row number 3 and row number 5 from the DF permanently and show the top 5 rows.
9. Sort the value Series in ascending order, showing only the value column.
10. Sort the value Series in ascending order, showing the whole DataFrame.
11. Sort the value Series in ascending order, showing the whole DataFrame, **inplace**.
12. Sort myDF DataFrame based on the range and value in ascending and descending order, respectively. (range in ascending and value in descending order).
13. Display those rows of myDF that their range is between 250 and 350.
14. Bonus Question: Display those rows of myDF that their range is between 250 and 350 in an ascending sorted format.
15. Display only the id and value columns of those rows of myDF that their range is between 250 and 350.
16. Display only the id, value and machine\_type columns of the of those rows of myDF that their range is between 250 and 350 and their ‘machine\_type’ is equal to ‘H’.
17. Display those rows of myDF that their ‘machine\_type’ is equal to ‘H’ or ‘R’ or ‘X’. (entire DataFrame)
18. Show the myDF columns’ datatype. Display (select) only the numerical columns.

Thank you and good luck,

Reza

Out beyond ideas of wrong and right, there is a field. I'll meet you there.

~Rumi