**Exercise- Data Cleaning Exercise**

Please use the screenshots ONLY as a reference. The written instructions have to be followed AS written.

PLEASE MAKE SURE YOU SUBMIT ALL THE SCREENSHOTS WITH TIMESTAMPS AT THE BOTTOM RIGHT (WINDOWS USERS) OR THE TOP RIGHT (MAC OSX USERS) OR ELSE YOU WILL AUTOMATICALLY QUALIFY FOR A DISCOUNT.

### Objective:

The objective of this exercise is to develop skills on how to cleanse data set in Python.

**Prerequisite:** Install Anaconda 2019.03 Installer, Python 3.7 version

In order to install, please click on the below link.

<https://www.anaconda.com/distribution/>

Then, launch **Jupyter Notebook** in the anaconda Navigator.

**A screenshot of a cell phone

Description automatically generated**

* In the Jupyter notebook ‘Files’ tab, Navigate to data folder and create new Python 3 notebook.

**A screenshot of a cell phone

Description automatically generated**

* Rename the file in given format by clicking on “Untitled” (top left corner) in the newly created notebook.

A screenshot of a cell phone

Description automatically generated

**Instructions:**

**Step 1: Import Required libraries**

**Create a Heading -** To Create a heading, write your headings in Markdown cells using # characters:

For example,we write the following code in the cell, then click on run to see the output.

* Remember to check the output of the written code block we need click on RUN.

For example, we write the following code in the cell, then click on run to see the output

A screenshot of a cell phone

Description automatically generated

After clicking on Run, the code will look like this:-

A screenshot of a cell phone

Description automatically generated

This is the heading of the section we will work on now. Throughout this assignment, you will have to create heading in the similar manner for every section of code you will write. This helps to keep the code organized.

Write the following code in the next cell, this will increase the width of the working area to the max. 

After you run this, write the following to import all the required Libraries.

A screenshot of a cell phone

Description automatically generated

**Question:**

1. **Take a screen shot of your page, showing page Title (your name) and page background.**

**Step 2: Setting Directory and Reading File**

Add the following code.

**A screenshot of a cell phone

Description automatically generated**

**Code Explanation**:

* + - Path- It is the address to the folder where the dataset is present.
    - Data\_ Dir - The exact path of the Dataset.
    - Data\_ Output - Path and name of the output file ( you will use this in the later part of the assignment ).

Using this code we set the directory as the path, and read the file from there.

**# Exploring the Data Set.**

Add the following code to see the different sheets present in the data.

**A screenshot of a cell phone

Description automatically generated**

**Question:**

**b. How many Sheets are present, also name them ?**

It’s time to explore the data elements present in the sheets.

**A screenshot of a social media post

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**Head** is function used in python to see the top rows of the dataset.You can see what are all the columns present.

**Step 3: Data Manipulation**

Removing Blank Rows - To Remove blank rows first we need to check weather blank rows are present or not.

A close up of a logo

Description automatically generated

This code will tell the number of columns we have , what are they and the number of non-null values.

**Question:**

1. **How many non-null rows are present in the data set?**

Use the below code to see the number of null Values present in each column.

**A screenshot of a cell phone

Description automatically generated**

**As you can see in the output that we have null values in every column. But there are some entries where all the values of a row are NULL. Now you will delete such rows. To do so, add the subsequent code.**

**A screenshot of a cell phone

Description automatically generated**

**Now, let’s check the null values present.**

**A screenshot of a cell phone

Description automatically generated**

**Question:**

1. **How many empty rows were present in the data set before we removed them?**

Let’s check the same for the products’ sheet.

**A screenshot of a cell phone

Description automatically generated**

**Question:**

1. **Using the above code tell how many empty rows are present in the Products sheet? Paste the Screenshot of the output showing page title (your name).**

**Step 4: Imputing the blank cells in the price column**

As you can see,Quite a few rows have missing values in the price column. Since the total number of missing values exceed 2% of the entire dataset, we cannot delete the rows. To clean the data, we impute values into the blank cells. The method to be used for imputation can be decided by looking at the histogram.

A skewed histogram suggests adopting the median as the method for imputation, where as a normal distribution leans towards adopting the mean.

To plot a histogram create bins that range from 1 to 27. Then compute Mean and Median. To achieve the results add this to the code.

A screenshot of a cell phone

Description automatically generated

PS - Ignore the warning messages.

**Question:**

1. **Paste a screenshot of the histogram and explain the methodology chosen for imputation.**

As you have chosen the imputing method. Let’s replace the null values in the price column with imputed values.

Use the first line if you want to use median , use second if Mean is your chosen imputing method.

A screenshot of a cell phone

Description automatically generated

Now, check the number of Null values present . Use -



**Question:**

1. **Paste a screenshot of the screen ( display name ) with the imputed method code showing the result of the ‘ .isna().sum ’ function. How many null values are there?**

**Step 5: Removal of Duplicate Values**

**Hint:** Check for countries with spelling errors and correct the data. Also check for duplicate entries for countries if any. Make sure only one record exists for each country (delete any one duplicate entry)

To list the unique values in the country column use-

**A close up of a logo

Description automatically generated**

**Question:**

1. **List the country with duplicate entry/ spell check error.**

**Remove the Spelling error by executing the following code-**

**A close up of a logo

Description automatically generated**

**Question :**

1. **Again, run the unique value command and paste the Screenshot of the whole screen. Explain what happened.**

Since, we have updated the sheet for the spelling errors, let’s check for the Duplicate entries.

A screenshot of a cell phone

Description automatically generated

This code will show the number of duplicate entries present. A duplicate entry here means when 2 rows have exact same entries for all columns.

**Question:**

1. **How many duplicate entries are present?**

In order to see the duplicate rows use the following code.

A screenshot of a cell phone

Description automatically generated

**Question:**

1. **Explain the outcome of the executed query, also paste the Screenshot of the code and the output using whole screen.**

**T**o remove the duplicates from the data frame use the code -

**A screenshot of a cell phone

Description automatically generated**

**Question:**

1. **Use the duplicated(). Sum() function, to check the duplicate entries present. Paste the Screenshot of the code and the output using whole screen.**

**Step 6: Merging both the sheets present in the Excel file.**

Create the heading by using “# Merging Both data Sets” query and run it as a. heading.

Use the following code to join both

**A screenshot of a cell phone

Description automatically generated**

Finally, we will sort the whole file by country column.

A screenshot of a cell phone

Description automatically generated

**Question:**

1. **Paste the screenshot of the head of df\_full .**

As, you can see that our file is now sorted, we have removed all the duplicates. It’s time for us to export the output in a file.

Execute the following code,

A picture containing screenshot

Description automatically generated

Make sure the data\_output is set correctly.

A screenshot of a cell phone

Description automatically generated

**Instructions:**

1. Submit the assignment document in Microsoft word.
2. Submit excel file
3. Submit .ipynb file created

Make sure you submit all the files.