Exercise- Data Cleansing Exercise

Student Name: **Student Id:**

Date:

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

Objective:

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

Instructions:

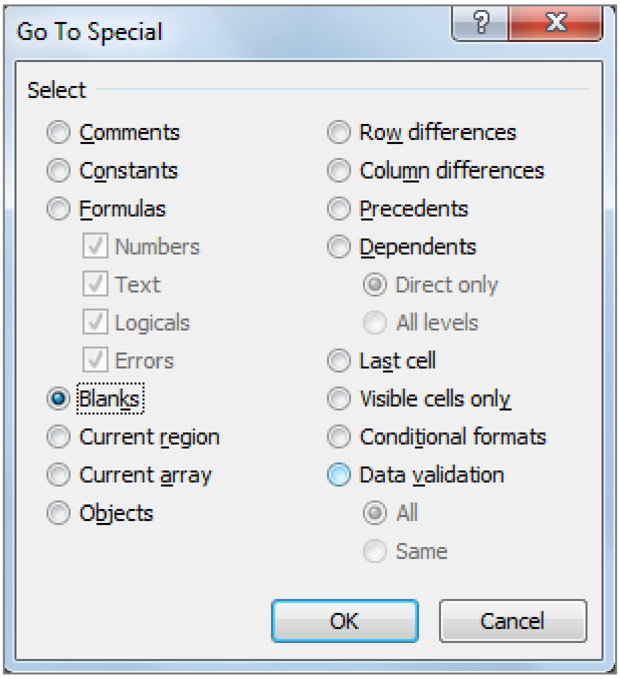
Copy the random number generated in cell J6 and paste the value using the option "Value" under Paste Options in the cell J6 itself.

LEAVE THE NUMBER ON THE CELL IN J6 WITHOUT FORMATTING.

Step 1: Removal of Blank Rows or Columns

Check for blank rows if present and delete them. Only delete rows which are entirely blank.

(Use Find & Select → Go to Special → Select Blanks)



Select

○ Comments
○ Constants
○ Formulas
☑ Numbers
☑ Text
☑ Logicals
☑ Errors
● Blanks
○ Current region
○ Current array
○ Objects

○ Row differences
○ Column differences
○ Precedents
○ Dependents
● Direct only
○ All levels
○ Last cell
○ Visible cells only
○ Conditional formats
○ Data validation
● All
○ Same

Question:

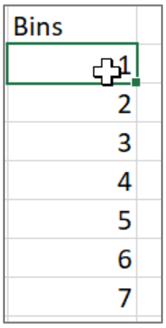
a) **What is the value in cell D439899? Paste a screenshot of the same.**

Note: Don’t sort the data by any column

Step 2: Imputing the blank cells in the price column

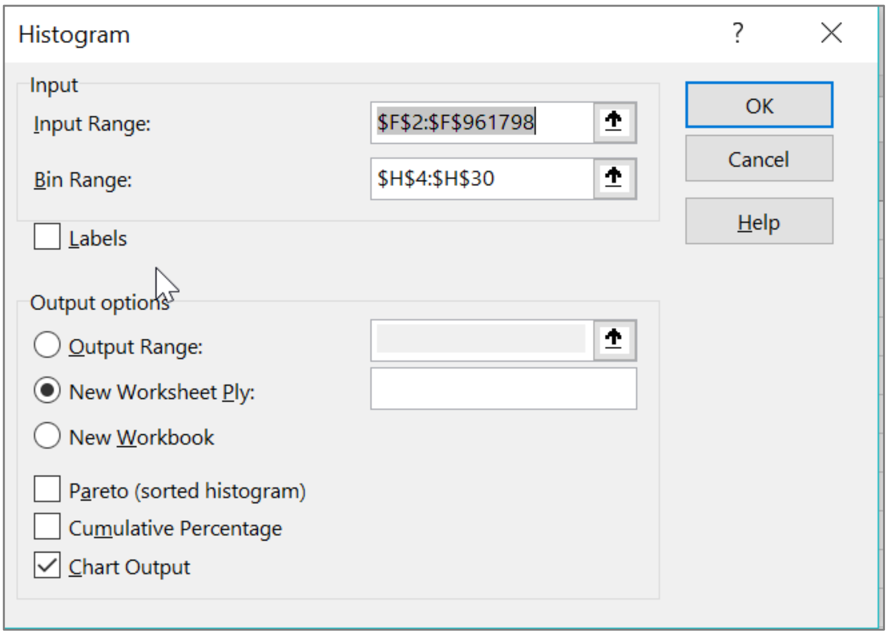
Hint 1: Quite a few rows have missing values in the price column. Since the total number of missing values exceed 2% of the entire dataset, we cannot delete the rows. To clean the data, we impute values into the blank cells. The method to be used for imputation can be decided by looking at the histogram.
A skewed histogram suggests adopting the median as the method for imputation, where as a normal distribution leans towards adopting the mean.

Hint 2: To plot a histogram, include the Analysis ToolPak in the Add-Ins. Create bins that range from 1 to 27 (Highest price in our dataset) in a new column. It should have an interval of 1.



Bins
1
2
3
4
5
6
7

Once the bins have been allotted, go to the Data Tab in the Menu Bar. Select Data Analysis and select the Histogram Option. Fill in the Input Range and Bin Range according to your worksheet and select the New Worksheet Ply option. Select the Chart Output option and Click on Ok. This should create a histogram along with the frequency distribution table in a new worksheet. Make your decision accordingly.



Histogram

Input

Input Range: $F$2:$F$961798

Bin Range: $H$4:$H$30

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

Question:

b) **Paste a screenshot of the histogram and explain the methodology chosen for imputation. (The title of the chart should be your name and give appropriate axis titles)**

Question:

c) **Paste a screenshot of the table with the imputed values**

Step 3: Bring all the cells in a single format

Hint 1: At the first glance of the entire data set, all the cells appear in a different color (You will find blocks of colored cells that highlight formatting problems) and random conditional formatting in different sets of cells. Fix that as an initial step.

There are several approaches to remove conditional formatting. In this exercise, you need to select **Conditional Formatting option -- > Clear Rules > Clear Rules from Entire Sheet** and **Clear > 'Clear Formats'** option available in the Home tab.

Hint 2: Convert Cells into correct Numerical Format. The format to be used for Price is Number with 2 decimal places, for Product Code and Quantity is General and for Obs Date is Custom.

Also check for alignment and bring all the data in one common numerical format after cleansing.

Question:

d) **Paste a screenshot of the table after initial cleaning.**

Note: Initial part of the table along with the random number should be visible.

Question:

e) **What is the total number of rows (including header) present in the data sheet?**

Step 3: 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)

Question:

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

Step 4: Change all text to one common case (Arial and Font size - 12)

Step 5: Use VLOOKUP to include the Product Name table on the Products sheet into a new column beside Product Code in the Observations sheet.

Question:

g) **Paste a screenshot showing the changes made.**

Note: Initial part of the table along with the random number should be visible.

Save the entire data set.

Sort the data by country column

Question:

h) **What is the Product Name and the Quantity of Product Code 44 in Islamabad? Paste a screenshot of the same.**

Attach only your assignment document (only the answers) on eLearning.