A picture containing shape, arrow

Description automatically generated**Duplication Typecasting**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Gunjan Kumar Gupta**

**Batch Id: DSWDMCSR 300522B**

**Topic: Preliminaries for Data Analysis**

**Problem statement:**

Data collected may have duplicate entries, that might be because the data collected were not at regular intervals or any other reason. To build a proper solution on such data will be a tough ask. The common techniques are either removing duplicates completely or substitute those values with a logical data. There are various techniques to treat these types of problems.

Q1. For the given dataset perform the type casting (convert the datatypes, ex. float to int)

Q2. Check for the duplicate values, and handle the duplicate values (ex. drop)

Q3. Do the data analysis (EDA)?

Such as histogram, boxplot, scatterplot etc

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
| 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 2.55 | 17850 | United Kingdom |
| 536365 | 71053 | WHITE METAL LANTERN | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 12/1/2010 8:26 | 2.75 | 17850 | United Kingdom |
| 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 22752 | SET 7 BABUSHKA NESTING BOXES | 2 | 12/1/2010 8:26 | 7.65 | 17850 | United Kingdom |
| 536365 | 21730 | GLASS STAR FROSTED T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 4.25 | 17850 | United Kingdom |
| 536366 | 22633 | HAND WARMER UNION JACK | 6 | 12/1/2010 8:28 | 1.85 | 17850 | United Kingdom |
| 536366 | 22632 | HAND WARMER RED POLKA DOT | 6 | 12/1/2010 8:28 | 1.85 | 17850 | United Kingdom |

A picture containing shape, arrow

Description automatically generated

**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature of the dataset to create a data dictionary as displayed in the below image:



1. Consider the OnlineRetail.csv dataset
2. Research and perform all possible steps for obtaining solution
3. All the codes (executable programs) should execute without errors
4. Code modularization should be followed
5. Each line of code should have comments explaining the logic and why you are using that function

# -\*- coding: utf-8 -\*-

"""

Created on Wed Jul 6 14:52:25 2022

@author: Gunjan

"""

"""

Q1. For the given dataset perform the type casting (convert the datatypes, ex. float to int)

Q2. Check for the duplicate values, and handle the duplicate values (ex. drop)

Q3. Do the data analysis (EDA)?

Such as histogram, boxplot, scatterplot etc

"""

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Q1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_##

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import math

# load data

df1 = pd.read\_csv(r"D:\DATA Science 360 DigiTMG\Assignment\Data Preprocessing\_Assignments\DataSets-Data Pre Processing\OnlineRetail.csv",encoding='latin1')

df1.dtypes

df1.columns

df2 = df1['UnitPrice']

#df2.dtypes

df2 = df1['c'].astype('int',errors='ignore')

df2.dtype # int32 datype

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Q2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_#

dup = pd.read\_csv(r"D:\DATA Science 360 DigiTMG\Assignment\Data Preprocessing\_Assignments\DataSets-Data Pre Processing\OnlineRetail.csv",encoding='latin1')

dup.columns

dup.drop\_duplicates(subset ="InvoiceNo", keep = False, inplace = True)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Q3\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_#

bx = pd.read\_csv(r"D:\DATA Science 360 DigiTMG\Assignment\Data Preprocessing\_Assignments\DataSets-Data Pre Processing\OnlineRetail.csv",encoding='latin1')

bx.shape

# barplot

plt.bar(height = bx.UnitPrice, x = np.arange(1, 774, 1)) # initializing the parameter

plt.hist(bx.UnitPrice) # histogram

plt.hist(bx.UnitPrice, color='green')

help(plt.hist)

plt.figure()

plt.boxplot(bx.UnitPrice) # boxplot

############### Scatterplot ############3

scp = pd.read\_csv(r"D:\DATA Science 360 DigiTMG\Assignment\Data Preprocessing\_Assignments\DataSets-Data Pre Processing\OnlineRetail.csv",encoding='latin1')

scp.drop\_duplicates(subset ="InvoiceNo", keep = False, inplace = True)

scp.columns

Quantity = scp.Quantity

UnitPrice = scp.UnitPrice

plt.scatter(UnitPrice,Quantity)