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: ANANDAKRISHNAN K V**

**Batch Id: 19042021**

**Topic: Preliminaries for Data Analysis**

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.

**Problem statement:**

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

**Ans:-**

**Python code:-**

################## Type casting###############################################

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

from sklearn.impute import SimpleImputer as si

### import data set OnlineRetail ###

df = pd.DataFrame(OnlineRetailcsv)

d = df.copy(deep=True)

df1 = df.copy(deep=True)

df1.dtypes

### UNitPrice & CustomerID columns datas are float64 type. so will convert those to int64

### remove the perticular rows having NA or missins values. Because type casting is not possible on NA datas

df1.isna() ### missing values treat as True

df1.isna().sum() ### total number of missing values ## df1.notna()=>Treat missing values as False

df1.shape ### shape of the dataframe

df2 = pd.DataFrame(df1.dropna(how='any')) ### drop the row datas contain missing values ; how = any => if any of the row data is NA then complete row will get eleminate ; shape => showing shape

### for subset or column name mentioned operation => df5 = pd.DataFrame(df1.dropna(subset=["Description","CustomerID"],how='any'))

df2.shape

#type casting

# Now we will convert 'float64' into 'int64' type.

df2.UnitPrice = df2.UnitPrice.astype('int64')

df2.CustomerID = df2.CustomerID.astype('int64')

df2.dtypes

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

**Ans:-**

**Python code:-**

################## Type casting & duplicate Treatment###########################

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

from sklearn.impute import SimpleImputer as si

### import data set OnlineRetail ###

df = pd.DataFrame(OnlineRetailcsv)

d = df.copy(deep=True)

df1 = df.copy(deep=True)

df1.dtypes

### UNitPrice & CustomerID columns datas are float64 type. so will convert those to int64

### remove the perticular rows having NA or missins values. Because type casting is not possible on NA datas

df1.isna() ### missing values treat as True

df1.isna().sum() ### total number of missing values ## df1.notna()=>Treat missing values as False

df1.shape ### shape of the dataframe

df2 = pd.DataFrame(df1.dropna(how='any')) ### drop the row datas contain missing values ; how = any => if any of the row data is NA then complete row will get eleminate ; shape => showing shape

### for subset or column name mentioned operation => df5 = pd.DataFrame(df1.dropna(subset=["Description","CustomerID"],how='any'))

df2.shape

#type casting

# Now we will convert 'float64' into 'int64' type.

df2.UnitPrice = df2.UnitPrice.astype('int64')

df2.CustomerID = df2.CustomerID.astype('int64')

df2.dtypes

#Identify duplicates records in the data

duplicate = df2.duplicated()

duplicate

sum(duplicate)

#Removing Duplicates

df2\_dup = df2.drop\_duplicates()

df2\_dup

## conclusion

print("the data set without any duplicates:", df2\_dup)

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