**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: Shabezade Srikanth**

**Batch Id: DS 07102022**

**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)

Answer: import pandas as pd

import numpy as np

data1=pd.read\_csv('F:/Assignments/Assignment 4 data pre processing/DataSets/OnlineRetail.csv',encoding='unicode\_\_escape')

data1.head()

data1.head(10)

data1.dtypes

data1['UnitPrice']=data1['UnitPrice'].astype('int64') #here float gets converted to int64

data1['UnitPrice'].dtypes

from sklearn.impute import SimpleImputer

mean\_imputer=SimpleImputer(missing\_values=np.nan, strategy='mean')

data1['CustomerID']=pd.DataFrame(mean\_imputer.fit\_transform(data1[['CustomerID']]))

data1['CustomerID'].isnull().sum

data1['CustomerID']=data1['CustomerID'].astype('int64')

data1['CustomerID'].dtypes

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

Answer: duplicate=data1.duplicated()

sum(duplicate)

data2=data1.drop\_duplicates()

Q3. Do the data analysis (EDA)?

Answer: import matplotlib.pyplot as plt

import numpy as np

plt.hist(data1.Quantity);plt.show()

plt.boxplot(data1.Quantity);plt.show()

plt.hist(data1.UnitPrice);plt.show()

plt.boxplot(data1.UnitPrice);plt.show()

plt.hist(data1.CustomerID);plt.show()

plt.boxplot(data1.CustomerID);plt.show()

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

**A picture containing shape, arrow

Description automatically generatedGrading Guidelines:**

**Note: 1. An Assignment submission is considered complete only when successful executable code(s), and documentation explaining the applied solution and results are provided. Failing to submit either of them will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline date will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline date
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of questions asked in assignments are completed. (or)
  + All assignments were submitted, however, after the given deadline
* **A picture containing shape, arrow

  Description automatically generatedGrade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of questions asked in assignments are completed. (or)
  + Less than 80% of questions asked in assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):** Assignments submitted after the Deadline and with 50% or less of questions
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of questions asked in the assignments are submitted after the deadline (OR)
  + Less than 30% of questions asked in the assignments are submitted before deadline

**Grade F: (< 50):** Copied submission or No submission