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: s**

**Batch Id:**

**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)

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

**Answer:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| Invoice no. | Number of the invoice | Quantitative,Discrete, Nominal | Irrelevant |
| Stock Code | Code of stock | Quantitative,Discrete, Nominal | Irrelevant |
| Description | About the data | Qualitative | Irrelevant |
| Quantity | Quantality how much the person bought | Quantitative,Discrete, Nominal | Relevant |
| Invoice date | Data of the invoice | Quantitative,Continuous, Interval | Irrelevant |
| Unit price | Total Cost | Quantitative,Continuous, Ratio | Relevant |
| Customer ID | ID of the customer | Quantitative,Discrete, Nominal | Irrelevant |
| Country | Country of person | Qualitative,Discrete,  Nominal | Relevant |

**BUSINESS PROBLEM**: To build proper solution on the datasets with duplicate values is very tough.

**CONSTRAINT**: Datasets contain duplicate values.

**DATA UNDERSTANDING:**

1. The dataset contains missing values and duplicates.
2. There are columns with float values as well.

**INSIGHTS FROM THE DATA:**

1. Typecasting has been applied to unit price column to change the data type to integer.
2. Typecasting couldn’t be applied to Customer ID column because it has missing values.
3. So mean imputation technique has been applied to fill the missing values.
4. Then typecasting has been done to change the data type to integer.
5. There are 5284 duplicate values which have been dropped.
6. Here I have used Histograms, Scatterplots and Boxplots.
7. Mostly the values lie between 0 to 18000 for quantity column.
8. For unit price the data lies mostly from -500 to 4000.
9. For Customer ID the data are mostly accumulated at the center(15000).

**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