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: Vishvash C**

**Batch Id: 23012024**

**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 for any other reason. Building a proper solution on such data will be a tough ask. The common techniques are either removing duplicates completely or substituting those values with 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 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 the 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.

import pandas as pd

data = pd.read\_excel(r"C:/Users/Lenovo/Downloads/Study material/EDA/InClass\_DataPreprocessing\_datasets/Online Retail1.xlsx")

data.dtypes

'''

CustomerID is float- Python automatically identify the data types by interpreting the values.

As the data for CustomerID is numeric Python detects the values as float64.

From measurement levels prespective the CustomerID is a Nominal data as it is an identity for each employee.

If we have to alter the data type which is defined by Python then we can use astype() function

'''

# Convert 'int64' to 'str' (string) type.

data.CustomerID = data.CustomerID.astype('str')

data['CustomerID']

data.dtypes

### Identify duplicate records in the data ###

# Duplicates in rows

duplicate = data.duplicated() # Returns Boolean Series denoting duplicate rows.

duplicate

sum(duplicate)

pd.set\_option('display.max\_columns', None)

# Removing Duplicates

data = data.drop\_duplicates() # Returns DataFrame with duplicate rows removed.

# Parameters

duplicate = data.duplicated()

sum(duplicate)

# Duplicates in Columns

# We can use correlation coefficient values to identify columns which have duplicate information

# Correlation coefficient

'''

Ranges from -1 to +1.

Rule of thumb says |r| > 0.85 is a strong relation

'''

data[['UnitPrice','Quantity']].corr()

import seaborn as sns

import matplotlib.pyplot as plt

sns.boxplot(data)

sns.boxplot(data=data, x='Country', y='UnitPrice')

plt.title('Box Plot of UnitPrice by Country')

plt.xticks(rotation=45)

plt.show()

sns.histplot(data)

sns.histplot(data=data, x='Quantity', bins=10, kde=True)

plt.title('Histogram of Quantity')

plt.show()

sns.countplot(data=data, x='Country')

plt.title('Count of Transactions by Country')

plt.xticks(rotation=45)

plt.show()

sns.scatterplot(data=data, x='Quantity', y='UnitPrice')

plt.title('Scatter Plot of Quantity vs. UnitPrice')

plt.show()

Output:

data.dtypes

Out[193]:

InvoiceNo object

StockCode object

Description object

Quantity int64

InvoiceDate datetime64[ns]

UnitPrice float64

CustomerID object

Country object

dtype: object

data = pd.read\_excel(r"C:/Users/Lenovo/Downloads/Study material/EDA/InClass\_DataPreprocessing\_datasets/Online Retail1.xlsx")

data.dtypes

Out[195]:

InvoiceNo object

StockCode object

Description object

Quantity int64

InvoiceDate datetime64[ns]

UnitPrice float64

CustomerID float64

Country object

dtype: object

data.CustomerID = data.CustomerID.astype('str')

data['CustomerID']

data.dtypes

Out[196]:

InvoiceNo object

StockCode object

Description object

Quantity int64

InvoiceDate datetime64[ns]

UnitPrice float64

CustomerID object

Country object

dtype: object

duplicate = data.duplicated() # Returns Boolean Series denoting duplicate rows.

duplicate

sum(duplicate)

Out[197]: 36

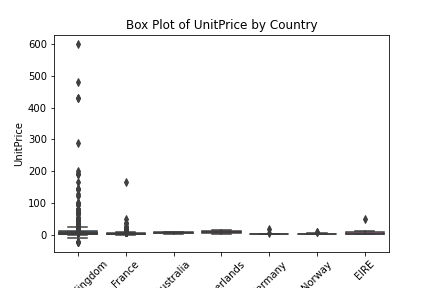
data = data.drop\_duplicates() # Returns DataFrame with duplicate rows removed.

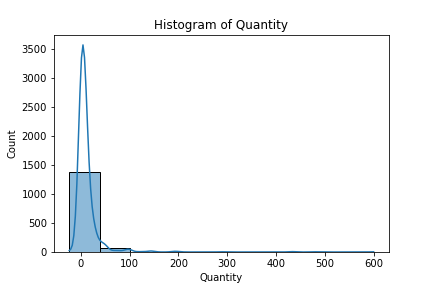
# Parameters

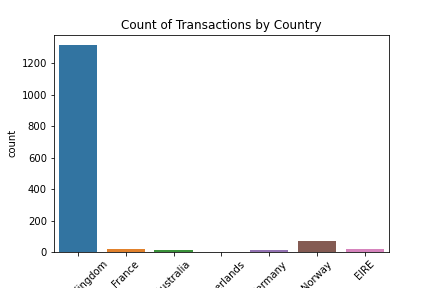
duplicate = data.duplicated()

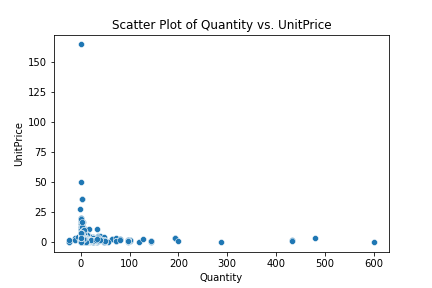
sum(duplicate)

Out[198]: 0









|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance­­­­** |
| **ID** |  | **Quantitative/ Nominal** | **Irrelevant (ID does not provide useful information)** |
| InvoiceNo | Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. | Nominal | Irrelevant |
| StockCode | Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product. | Nominal | Relevant |
| Description | Product (item) name. Nominal. | Nominal | Relevant |
| Quantity | The quantities of each product (item) per transaction. Numeric. | Nominal | Relevant |
| InvoiceDate | Invoice Date and time. Numeric, the day and time when each transaction was generated. | Quantitative | Relevant |
| UnitPrice | Unit price. Numeric, Product price per unit in sterling. | Quantitative | Relevant |
| CustomerID | Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer. | Nominal | Relevant |
| Country | Country name. Nominal, the name of the country where each customer resides. | Nominal | Relevant |