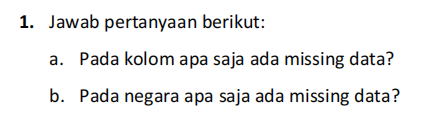
Jawaban Data Mining UTS

Ekky Dimas K (32602200147)



Jawaban :

1. a. Jika dilihat dari data yang belum dibersihkan. Kolom dan negara yang masih missing yang saya temukan seperti :

* Autralia pada kolom unit price
* Singapore pada kolom deskripsi
* EIRE

b. mencari cara dengan jupyter notebook . digunakan untuk memahami seberapa banyak data yang hilang dan memutuskan bagaimana menanganinya.

missing\_data\_per\_column = df.isnull().sum()

print("Missing data per column (original dataset):")

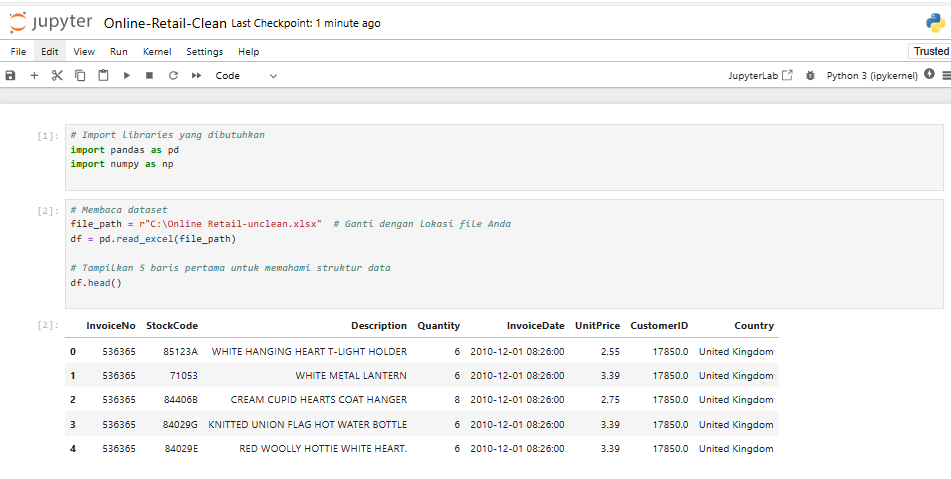
print(missing\_data\_per\_column[missing\_data\_per\_column > 0])

missing\_data\_by\_country = df[df.isnull().any(axis=1)]['Country'].unique()

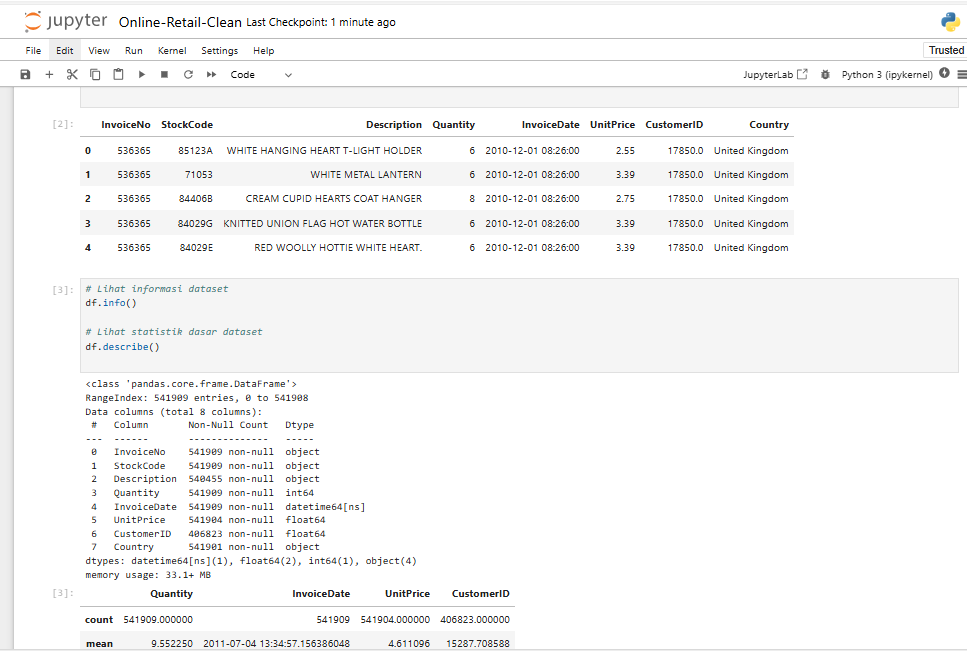
print("Negara dengan missing data (original dataset):")

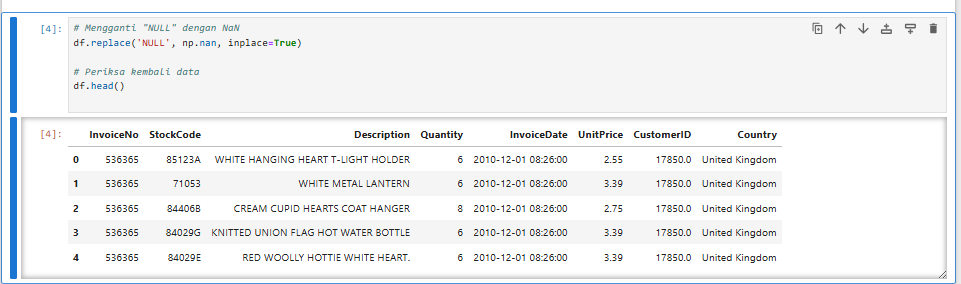
print(missing\_data\_by\_country)

1. Cara saya melakukan data cleansing dengan menggunakan jupyter notebook dengan run per cell nya. Dengan penjelasan sebagai berikut. Penjelasan sudah sudah terintegrasi dalam komentar jupyter notebook



Gambar 1. 1 Import library

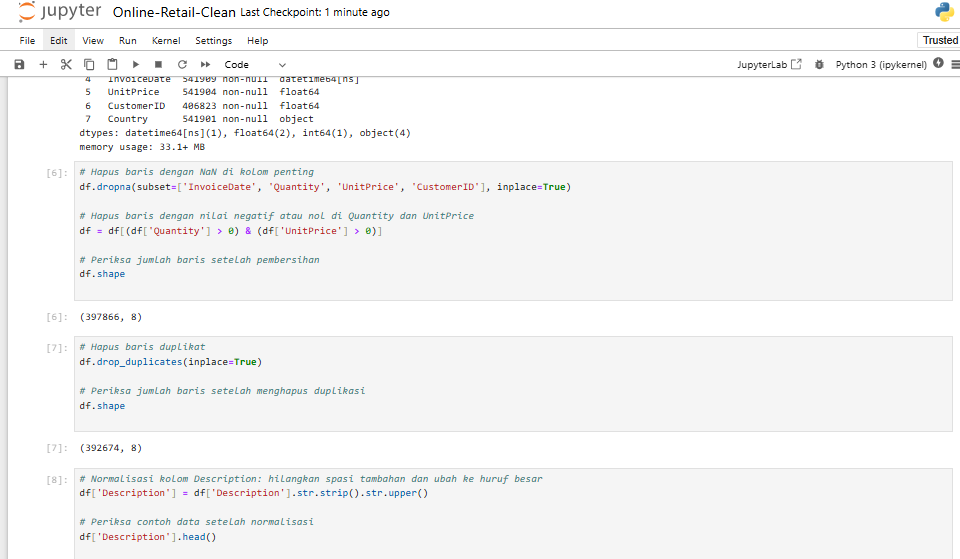




Gambar 1. 2 untuk memeriksa file Kembali



Gambar 1. 3 mengintegrasikan date time



Gambar 1. 4 Menghapus duplikasi



Gambar 1. 5 Menormalisasi huruf file, menyimpan hasil dan mendownload file excel

1. Jika Anda mengisi missing data dengan data baru, jelaskan bagaimana anda mendapatkan

setiap value untuk data baru tersebut masing-masing!

1. Kolom deskripsi . Mengisi nilai kosong berdasarkan data lain dalam baris yang sama.

df['Description'] = df.groupby('StockCode')['Description'].transform(lambda x: x.fillna(x.mode()[0] if not x.mode().empty else "UNKNOWN"))

1. Kolom Unitprice . Mengisi nilai kosong dengan rata-rata atau median berdasarkan produk (StockCode) yang sama.

df['UnitPrice'] = df.groupby('StockCode')['UnitPrice'].transform(lambda x: x.fillna(x.median()))

df['UnitPrice'] = df['UnitPrice'].fillna(df['UnitPrice'].median()) # Backup dengan median global

1. Kolom CustumerID . dengan cara Menandai transaksi tanpa CustomerID sebagai transaksi "anonim" atau mengganti dengan nilai paling umum.

df['CustomerID'] = df.groupby('Country')['CustomerID'].transform(lambda x: x.fillna(x.mode()[0] if not x.mode().empty else 99999))

df['CustomerID'] = df['CustomerID'].fillna(99999) # Backup dengan nilai default