**UTS DATA MINING DAN BUSINESS INTELLIGENCE**

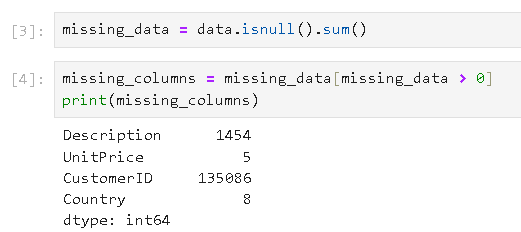
**NAMA : ASSYFA FEBRIWANTI**

**NIM : 32602200046**

File Online Retail-unclean.xlxs berisi tentang transaksi barang retail oleh customer pada berbagai negara dan waktu. Di dalam file tersebut ada beberapa missing data, lakukan data cleaning pada missing data tersebut, dan jawab pertanyaan-pertanyaan dibawah :

1. Jawab pertanyaan berikut :
2. Pada kolom apa saja ada missing data?

Jawab :



Berdasarkan hasil diatas, kolom yang terdapat missing data, antara lain :

* Kolom Description
* Kolom UnitPrice
* Kolom CustomerID
* Kolom Country

1. Pada negara apa saja ada missing data?

Jawab :



Berdasarkan hasil diatas, negara yang terdapat missing data, antara lain :

* Australia
* Bahrain
* Canada
* EIRE
* France
* Hongkong
* Israel
* Portugal
* Singapore
* Switzerland
* United Kingdom
* Unspecified

1. Lakukan dan jawab :
2. Lakukan data cleaning pada missing data tersebut.
3. Jelaskan bagaimana anda melakukan data cleaning?

Jawab :



Berdasarkan hasil cleaning data diatas, saya melakukan cleaning data pada :

* Kolom Description dengan menghapus baris data yang kolom Description nya kosong.
* Kolom UnitPrice dengan mengisi mean atau rata-rata data UnitPrice di kolom UnitPrice yang kosong.
* Kolom CustomerID dengan mengisi angka 0 di kolom CustomerID yang kosong.
* Kolom Country dengan mengisi modus dari data Country di kolom Country yang kosong.

Dengan cleaning data diatas, maka sudah tidak ada missing data pada file.

1. Jika anda mengisi missing data dengan data baru, jelaskan bagaimana anda mendapatkan setiap value untuk data baru tersebut masing-masing!

Jawab :

1. Kolom UnitPrice : Saya mengisi kolom UnitPrice dengan mean dari data UnitPrice karena nilai yang ada memiliki kemungkinan besar berada di sekitar rata-rata dataset yang ada, dan juga kolom UnitPrice penting untuk analisis penjualan. Kemudian data yang ada pada UnitPrice tidak terlalu terpengaruh oleh outlier, dan saya juga memilih solusi yang sederhana tanpa mempengaruhi hasil analisis.
2. Kolom CustomerID : Saya mengisi kolom CustomerID dengan angka 0, karena jumlah missing data nya sekitar 25% dari keseluruhan data, jadi jika dihapus maka akan berpengaruh karena menjadi kehilangan sebagian besar data, karena CustomerID bersifat unik yang dimana tidak bisa menggunakan mean, modus atau median untuk mengisi missing data nya karena 1 CustomerID hanya bisa digunakan oleh satu baris data.
3. Kolom Country : Saya mengisi kolom Country dengan modus dari data Country karena transaksi yang data negara nya kosong memiliki kemungkinan besar berasal dari negara yang paling sering muncul (negara dengan transaksi terbanyak). Seperti pada kolom Country modus nya adalah “United Kingdom”, karena mayoritas transaksi berasal dari negara tersebut, itu juga termasuk solusi yang sederhana dan efektif karena nilai kosong nya hanya sedikit, yaitu hanya 8 data.