Laporan Praktikum Pertemuan 7

**Data Science Lanjut**

Data Quality With Python For Beginner

Dibuat oleh

Nama : Muhamad faisal halim

Nim : 19.240.0163

Kelas : -

Mata kuliah : Data science lanjut

Mahasiswa pertukaran mahasiswa.

**Universitas muhammadiyah kalimantan timur**

~ stmik widya pratama pekalongan

**Penting**

Mohon maaf pak, pada materi sebelumnya masih terdapat beberapa tugas yang belum saya kumpulkan, yaitu pada module 3,4 dan 6 untuk laporan saya sertakapan pada link berikut [link\_1](https://drive.google.com/drive/folders/1qtaG1pREWQin83ZvYxPP4aF0Cs4iHXZp?usp=sharing) atau <https://bit.ly/laporan-tertinggal-halim0163>.

Semoga dengan saya mencantum laporan sebelumnya ini bisa mengisi kekosongan tugas laporan saya dan menjadi bahan pertimbangan tambahan untuk hal yang diperlukan dalam penilaian akhir matakuliah Data Science lanjut

**Materi dan praktikum**

**Title :** Data quality with python for beginner

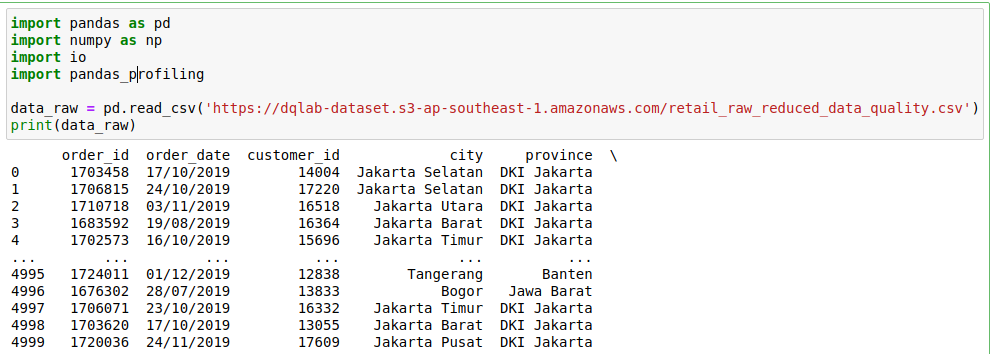
Data profiling

Data profiling adalah kegiatan merangkum dataset menggunakan statistik deskriptif. Yang bertujuan memiliki pemahaman yang kuat tentang data sehingga dapan menyusun framework analist dan memvisualisasikannya.

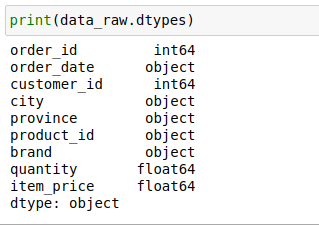
Dalam praktikum kali ini kita masih menggunakan pandas dan numpy, karena memang dari awal sudah dikatakan bahwa kedua package python tersebut sangay membantu dalam data science.

Importing data

Tahap awal pada praktikum ini adalah mengimport dataset kedalam python dan library atau package dala project kitan.



Didalam python khususnya menggunakan pandas kita dapat melakuan pengecekan datatype pada setiap kolom.

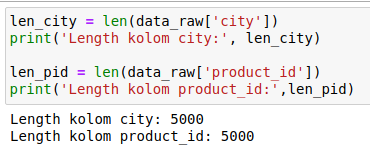


Deskriptif statistik

Pada tahap ini kita akan mempelajari statistika deskriptif yang digunakan untuk dapat memberikan pemahaman lebih mengenai struktur data.

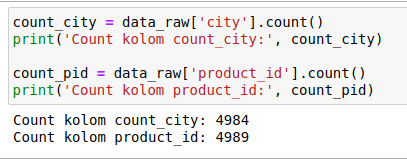
Lenght

Fungsi len digunakan untuk menghitung jumlah pengamatan dalam satu series atau column.



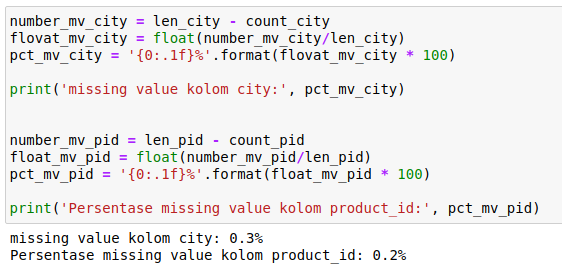
Count

Fungsi ini akan menghitung jumlah pengamatan dalam satu series atau column yang memiliki nilai. Nilai disini bisa diartikan sebagai non null dan bukan missing value.



Missing value

Pada perhitungan ini akan menghasilkan perbedaan antara lenght dan count.

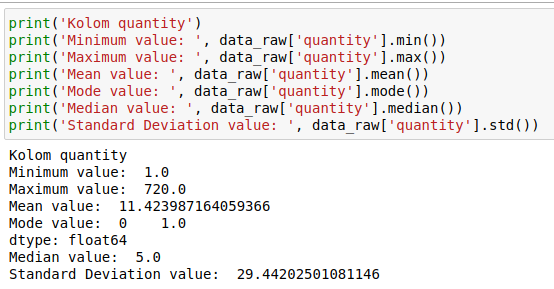


Maximum dan minimum

Fungsi max dan min digunakan untuk mengetahui element terbesar dan terkecil dari suatu kolom di dataframe.

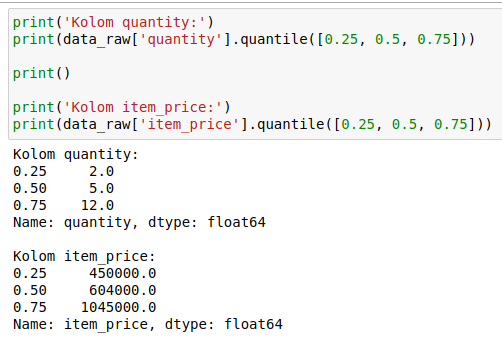
Mean, medium, modus dan standard deviasi

Fungsi mean, medium, modus dan standard deviasi digunakan untuk mengetahui pemusatan data dan persebarannya.



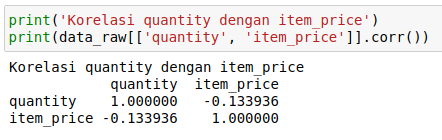
Quantile statistics

Quantiles adalah titik potong yang membagi distribusi dalam ukuran yang sama. Jika akan membagi distribusi menjadi empat grup yang sama, kuantil yang dibuat dinamai quartile. Jika dibagi kedalam 10 sepuluh group yang sama dinamakan percentile. Dalam kasus di bawah ini, ingin membagi distribusi menjadi empat grup atau quartile.



Correlation

Korelasi adalah cara yang tepat untuk menemukan hubungan antara variabel numerik. Koefisien korelasi berkisar antara -1 hingga 1. Korelasi 1 adalah korelasi positif total, korelasi -1 adalah korelasi negatif total dan korelasi 0 adalah korelasi non-linear.



Penggunaan profilling library

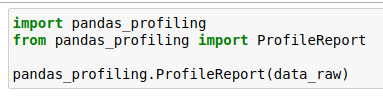
Pada praktik sebelumnya kita tahu bahwa engumpulkan statistik deskriptif dapat menjaro proses yang panjang, pandas profiling library dapat mempersingkat proses tersebut secara otomatis.

Pertama kita install dulu library yang kita butuhskan dengan command berikut.

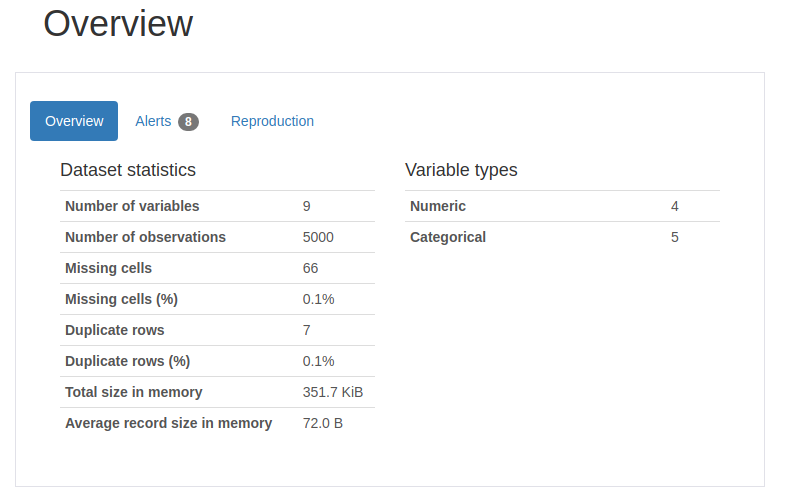
Pip install pandas\_profiling atau

Pip3 install pandas\_profiling

Pastikan bahwa yang kita install adalah versi terbarunya. Untuk penggunaan kodenya bisa kita lihat seperti berikut. Data\_raw diambil dari dataset yang kita import sebelumnya.



Untuk hasil run dari kode diatas terlihat cuplikanya seperti berikut.



Data cleansing

Adalah proses mendeteksi dan memperbaiki catatan yang rusak atau tidak akurat dari kumpulan catatan, tabel, atau basis data dan mengacu pada pengidentifikasian bagian data yang tidak lengkap, tidak benar, tidak akurat, atau tidak relevan dan kemudian mengganti, memodifikasi, atau menghapus datanya.

Missing data

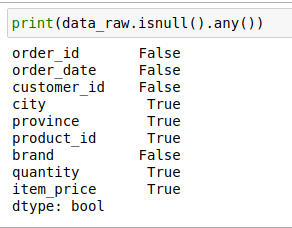
Pada sekarang ini dengan banyaknya data yang ditemukan di dunia pastinya terdapay banyak juga missing value dari data tersebut. Oleh karena itu treatment missing value sangatlah penting, karena missing value dapat mempengaruhi analisis dan mechine learning model.

Ada beberapa cara untuk mengatasi ini,

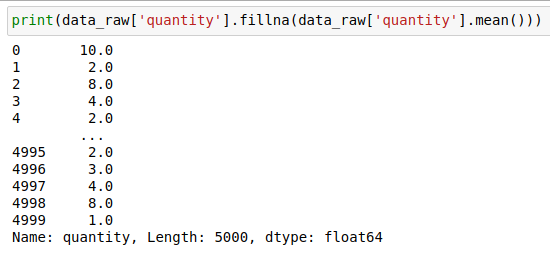
1. Dibiarkan
2. Imputasi
3. Menghapus row yang mengandung missing value

Imputasi adalah cara yang dilakukan untuk mengisi kekosongan data menggunakan teknik tertentu, umumnya menggunakan mean, modus atau mengguankan predictiv modeling.

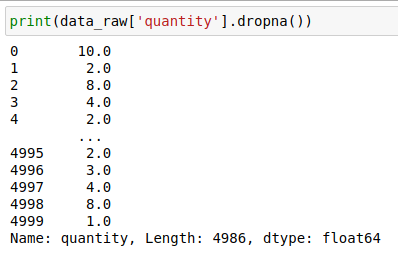
Mengecek kolom yan memiliki missing value



Mengisi missing value pada kolom quantity dengan mean



Menghapus missing value pada kolom quantity



Outlier

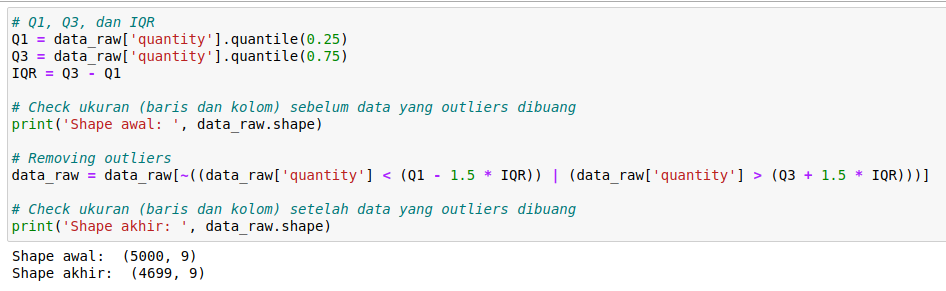
Adalah observasi yang muncul dengan nilai yang sama sekali berbeda dengan sebagian besar nilai lain dalam kelompoknya, biasanya disebut juga dengan nilai ekstrim.

Cara treatment outlier antara lain

1. Dihapus
2. Imputasi
3. Capping
4. Prediction

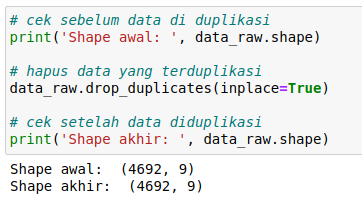
Umumnya outlier dapat ditentukan dengan metric iqr (interquartile range) rumusnya q3 - q1, data suatu observasi dapat dikatakan outlier jika memenuhi syarat-syarat berikut.

|  |
| --- |
| < q1 - 1.5 \* iqr |
| > q3 + 1.5 \* iqr |



Deduplikasi data

Merupakan data dengan kondisi pada row-row tertentu memiliki kesamaan data diseluruh kolomnya.

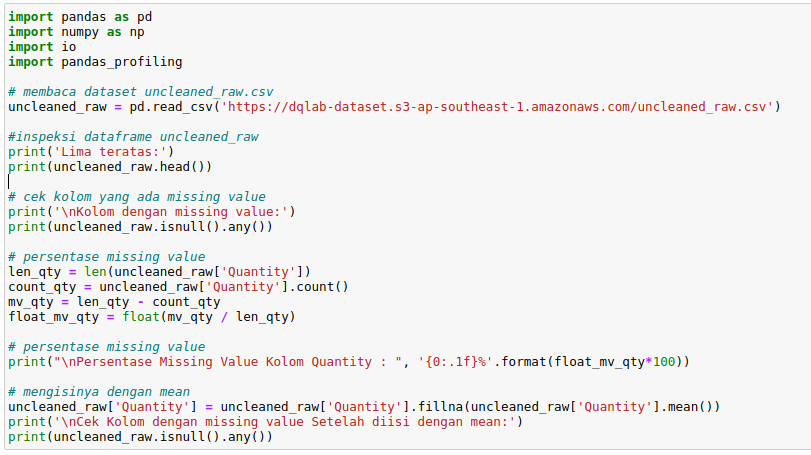


Case studi

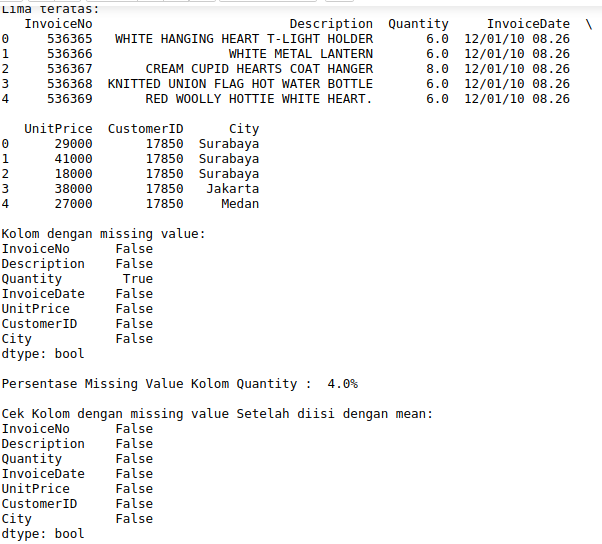
Profiling data

1. Import data kedalam variable uncleaned\_csv
2. Inspeksi data
3. Cek kolom yang mengandung missing value, jika ada berapa persen missing value tersebut.
4. Mengisi missing value dengan mean

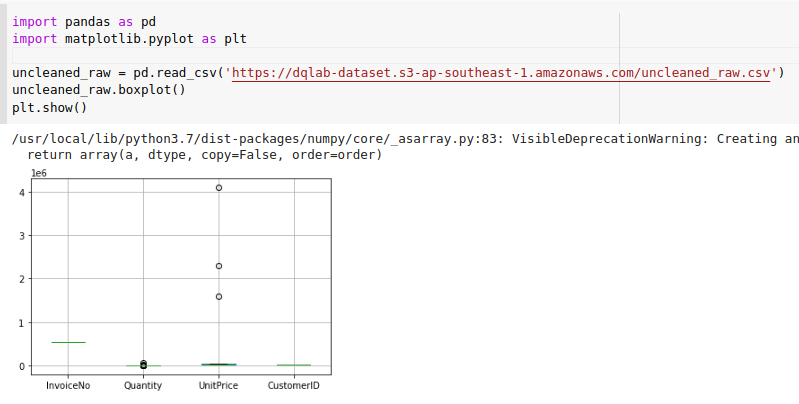
Hasil pengerjaan



Hasil output



1. Mengetahui kolom yang memiliki outlier, gunakan visualisasi dengan boxplot pada dataframe



1. Melakukan removing outliers pada kolom unitprice
2. Ceking duplikasi dan lakukan deduplikasi dataset tersebut