Laporan Praktikum Pertemuan 5

Data Science Lanjut

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

Note

Data pada praktikum ini disamakan dengan data yang ada pada contoh yang diberikan di Openlearning UMKT.

Materi dan Praktikum

Title : Statistic Using Python For Data Science

Pengenalan Pandas

Walaupun sudah pernah dibahas sebelumnya namun tidak ada salahnya saya tuliskan lagi disini. Pandas adalah sebuah librari berlisensi BSD dan open source yang menyediakan struktur data dan analisis data yang mudah digunakan dan berkinerja tinggi untuk bahasa pemrograman Python. Dengan kata lain, Pandas adalah librari analisis data yang memiliki struktur data yang diperlukan untuk membersihkan data mentah ke dalam sebuah bentuk yang cocok untuk analisis (yaitu tabel).

Struktur data dasar pandas dinamakan DataFrame, yaitu sebuah koleksi kolom berurutan dengan nama dan jenis, dengan demikian merupakan sebuah tabel yang tampak seperti database dimana sebuah baris tunggal mewakili sebuah contoh tunggal dan kolom mewakili atribut tertentu

Pengenalan Statistika

Statistika adalah sebuah ilmu yang mempelajari bagaimana cara merencanakan, mengumpulkan, menganalisis, lalu menginterpretasikan, dan akhirnya mempresentasikan data. Singkatnya, statistika adalah ilmu yang bersangkutan dengan suatu data. Istilah "Statistika" berbeda dengan "Statistik". Statistika pada umumnya bekerja dengan memakai data numerik yang di mana adalah hasil cacahan maupun hasil pengkuran yang dilakukan dengan menggunakan data kategorik yang diklasifikasikan menurut sebuah kriteria tertentu.

Konsep dasar Statistika

- Observasi
  adalah suatu unit yang diukur dengan data, misal : warga, siswa, hewan.
- Populasi
  adalah koleksi dari keseluruhan observasi, misal : semua siswa disekolah, semua spesies macan, semua kendaraan di pekalongan.
- Sampel
  adalah sub kolehsi dari populasi. misal : beberapa siswa disekolah, 3 spesies macan, 100 kendaraan di pekalongan

Ada dua kategori data pada statistika khususnya pada sampel atau populasi. yaitu data kuantitatif dan data kualitatif.
Data Kualitatif

Data kualitatid atau data kategorik adalah daya yang didapat dari hasil mengkategorikan atau menjelaskan suatu attribut dari populasi atau sampel. misal : nama jala, golongan darah.

Data Kuantitatif

Data yang dipertolah hasil perhitungan disuatu populasi, data ini selalu berbentuk angka. misal : Data Gaji Karyawan, Data Populasi Suatu kota atau negara. Data ini bisa dibagi lagi menjadi dua, yaitu data diskrit dan data kontinu.

Kedua data tersebut sama-sama dari hasil perhitungan namun jika hasil perhitungan yang ada dapat memuat rasio, desimal atau bilangan irasional adalah data kontinu, dan begitujuga sebaliknya. contoh :

- Data Diskrit : Jumlah telepon yang diterima CS perhari.
- Data Kontinu : berat bada, gaji, tinggu bada, waktu.

Statistika Deskriptif

Statistika deskriptif adalah metode yang berkaitan dengan pengumpulan / penyajian data hingga memberi informasi yang berguna. Contoh statistika ini : Diagram, Tabel, Grafik.

Dengan Statistika deskriptif, kumpulan data bisa tersaji dengan ringkas dan rapi serta mampu memberikan informasi inti dari kumpulan data yang ada. Informasi yang diperoleh dari statistika deskriptif ini antara lain ukuran pemusatan data, ukuran penyebaran data, serta kecenderungan suatu gugus data.

Statistika Inferensial

Statistik inferensial yaitu sebuah sebuah metode yang mampu dipakai untuk menganalisis kelompok kecil dari data induknya atau sample yang diambil dari populasi sampai pada peramalan dan penarikan kesimpulan pada kelompok data induknya atau populasi.

Skala Pengukuran dalam Statistika

Dalam hasil penelitian Stanley Smith Stevens (1946), dia membuat klasifikasi skala pengukuran penelitian sosial menjadi empat jenis skala pengukuran yaitu skala nominal, skala ordinal, skala interval dan skala rasio, ke empat jenis skala tersebut yang menjadi acuan sampai saat ini. sumber statmat.net

1. Skala Nominal

   Skala nominal merupakan jenis skala pengukuran yang termasuk kedalam kategori atau kelompok dari suatu subyek. Misalnya, dapat anda lihat pada variabel jenis kelamin, dimana pengelompokan umumnya hanya menjadi dua, yaitu laki-laki (L) dan perempuan(P) yang masing-masing diberi misal 1 dan 2.

2. Skala Ordinal

   Skala ordinal merupakan salah satu jenis skala pengukuran dimana lambang-lambang bilangan hasil pengukurannya berupa urutan atau tingkatan. Uji statistik yang sesuai adalah modus, median, distribusi frekuensi.

3. Skala Interval

   Merupakan jenis skala pengukuran yang mempunyai karakteristik mirip dengan skala ordinal yaitu memiliki urutan tertentu. Sifat lain yang melekat pada skala interval adalah adanya satuan skala (scale unit). Uji statistik yang sesuai adalah semua uji statistik kecuali uji yang mendasarkan pada rasio seperti koefisien variasi.

4. Skala Rasio

   Skala rasio adalah jenis skala pengukuran yang menghasilkan data dengan mutu yang paling tinggi. Perbedaan skala rasio dengan skala interval terletak pada keberadaan nilai nol (based value). Pada skala rasio, nilai nol bersifat mutlak, tidak seperti pada skala interval. Data yang dihasilkan oleh skala rasio adalah data rasio. Tidak ada pembatasan terhadap alat uji statistik yag sesuai.

Python Packaage Untuk Statistika.

Didalam bahasa pemrogaman python terdapat beberapa package terkenal bantuan untuk kita melakukan perhitungan statistika. berikut adalah beberapa daftar package terkenal dalam python untuk melakukan perhitungan statistika.

1. **Numpy**, digunakan untuk analisis data numerik dan perhitungan berbasis vektor dan matrix.
2. **Pandas**, untuk pengolahan data dalam tabel. (tabular data)
3. **matplotlib**, untuk melakukan penggambaran grafik. pembantu pembuatan grafik informatika.
4. **statsmodel**, digunakan untuk melakukan pengujian hipotesis, eksplorasi data ataupun pemodelan statistika.
5. **scipy**, digunakan untuk melakukan uji statistika, juga dapat digunakan untk pemodelan statistika.

Ukuran Pusat (Meansures Of Central Tendency)

Adalah statistika deskriptif yang dapat memambantu kita mengindentifikasi kasus-kasus tipikan dalam sebuah sample atau populasi, terdapat beberapa ukuran pusat yang dapat digunakan untuk manganalisa data. mean, median, dan modus.

Mean (Rata-rata)

Adalah salah satu ukuran pusat yang nilainya diperoleh dengan menjumlahkan semua titik nilai dan membanginya dengan jumlah datanya.

Untuk perhitungan mean dalam python kita dapat menggunakan fungsi `.mean()` pada numpy.

```
import numpy as np
import pandas as pd

data = pd.read_csv("./Dataset/dataset_statistic.csv", sep=';')
produk_a = data[data['Produk'] == 'A']

print (produk_a['Pendapatan'].mean())
print (np.mean(produk_a['Pendapatan']))
```

```
550000.0
550000.0
```

Median

Adalah suatu ukuran pusat yanng nilainya terletak di tengah data, misal data 1,2,3,4,5 maka nilai medianya adalah 3, namun jika kita memiliki data 1,2,3,4 maka nilai tengahnya adalah “(2 + 3) / 2” = 2,5.

Dalam python khusunya pada package numpy kita dapat menggunakan fungsi `.median()`, contohnnya seperti berikut.

```
import numpy as np
import pandas as pd

data = pd.read_csv("./Dataset/dataset_statistic.csv", sep=';')
produk_a = data[data['Produk'] == 'A']

print (produk_a['Pendapatan'].median())
print (np.median(produk_a['Pendapatan']))
```

```
600000.0
600000.0
```

Modus

Modus biasa dikatakan sebagai data yang memiliki frekuensi kemunculan terbanyak. misal kita memliki data 2,2,2,2,2,4,4,3,3,3,5,6,6 maka modus dari data tersebut adalah 2, karena kemunculan angka 2 ada sebanyak 4 kali.

Numpy python juga menyediakan fungsi untuk dapat kita mendapatkan modus. untuk kodenya bisa dilihat pada gambar dibawah.

```
import numpy as np
import pandas as pd

data = pd.read_csv("./Dataset/dataset_statistic.csv", sep=';')
print (data['Produk'].value_counts())
```

```
D    5
A    4
B    4
C    4
E    3
Name: Produk, dtype: int64
```

Kuantil

Adalah nilai-nilai data yang membagi data yang telah diurutkan sebelumnya menjadi beberapa bagian yang sama besar ukuranya. ada beberapa ukuran fraktil ini.

Kuartil - Adalah kuantil yang membagi data menjadi empat bagian sama bersar, kuartil ke-2 adalah median dari data tersebut.
Besil - Adalah kuantil yang membagi data menjadi 10 bagian sama besar.
Persentil - Adalah kuantil yang membagi data menjadi 100 bagian sama besar.

berikut adalah contoh kuantil dalam python.

```
import numpy as np
import pandas as pd

raw_data = pd.read_csv("./Dataset/dataset_statistic.csv", sep=';')
print (raw_data['Pendapatan'].quantile(q = 0.5))
print (np.quantile(raw_data['Pendapatan'], q=0.5))
```

```
875000.0
875000.0
```

Ada kalanya kita ingin menghitung sekaligus beberapa ukuran, misalnya menghitung nilai mean sekaligus menghitung nilai median. Kita dapat melakukan kedua hal tersebut dengan menggunakan method `.agg()` pada objek pandas DataFrame sebagaimana contoh berikut:

```python
import numpy as np
import pandas as pd

data = pd.read_csv("./Dataset/dataset_statistic.csv", sep=';')
print (data[['Pendapatan', 'Harga']].agg([np.mean, np.median])); print()
print (data[['Pendapatan', 'Harga', 'Produk']].groupby('Produk').agg([np.mean, np.median]))
```

```
        Pendapatan     Harga
mean     1160000.0  197500.0
median    875000.0  200000.0

       Pendapatan                Harga
             mean     median      mean    median
Produk
A         550000.0   600000.0  100000.0  100000.0
B         875000.0   875000.0  150000.0  150000.0
C         850000.0   900000.0  200000.0  200000.0
D        2100000.0  1200000.0  250000.0  250000.0
E        1200000.0  1100000.0  300000.0  300000.0
```