Nama : Ghina Khoerunnisa
Program : Data Scientist Intern

Descriptive Statistics

Statistik deskriptif merupakan metode-metode yang berkaitan dengan pengumpulan dan penyajian data sehingga dapat memberikan informasi yang berguna. Metode ini hanya mendeskripsikan kondisi dari data yang sudah dimiliki dan menyajikannya dalam bentuk tabel diagram grafik dalam uraian-uraian singkat dan juga terbatas (Ulya, 2021). Statistik deskriptif dibagi menjadi dua kategori yaitu Measures of central tendency dan Measures of variability (spread). Pada python kita dapat menggunakan package seperti math, statistics, numpy, scipy.stats, dan pandas untuk melakukan statistik deskriptif.

- **Measures of central tendency** menunjukkan nilai tengah atau pusat dari kumpulan data. Berikut merupakan cara mengidentifikasi dan menghitung measures of central tendency:

  Mean/Average: rata-rata aritmatika dari semua elemen kumpulan data. Untuk menghitungnya dapat menggunakan statistics.mean(), sum(x)/len(x), np.mean(x), z.mean()-> z adalah pandas dataframe.

  Weighted Mean: generalisasi dari rata-rata aritmatika yang memungkinkan kita untuk menentukan kontribusi relatif dari setiap titik data ke hasil. Mengalikan setiap titik data dengan bobot yang sesuai, menjumlahkan semua produk, dan membagi jumlah yang diperoleh dengan jumlah bobot. Untuk menghitungnya dapat menggunakan python built-in sum menyesuaikan dengan rumusnya atau menggunakan numpy dengan np.average(y, weights=w).

  Harmonic Mean: rata-rata kebalikan dari rata-rata aritmatika. Harmonic mean dihitung dengan membagi banyaknya nilai di kumpulan data dengan jumlah kebalikan (1 / x_i) dari setiap nilai dalam kumpulan data. Harmonic mean biasa digunakan pada beberapa kasus data misalnya data kecepatan, rasio atau tarif. Ini adalah ukuran yang paling tepat untuk rasio dan tarif karena ini menyamakan bobot setiap titik data. Misalnya, mean aritmatika memberi bobot yang tinggi pada titik data yang besar, sedangkan mean geometrik memberikan bobot yang lebih rendah ke titik data yang lebih kecil (livingeconomyadvisors). Untuk menghitungnya dapat menggunakan statistics.harmonic_mean(x), scipy.stats.hmean(x).

  Geometric Mean: rata-rata yang dihitung dengan mengalikan semua data dalam suatu kelompok sampel, kemudian di akar pangkatkan dengan banyaknya data sampel tersebut. Biasanya digunakan pada perhitungan finansial (saham, aset, dll). Untuk menghitungnya dapat menggunakan scipy.stats.gmean(x).

  Median: elemen tengah dari kumpulan data yang telah disorting. Untuk menghitungnya dapat menggunakan statistics.median(x), np.median(x)

  Mode: Nilai yang paling sering muncul pada kumpulan data. Untuk menghitungnya dapat menggunakan statistics.mode(x) -> tetapi jika menggunakan statistics dengan data yang modusnya ada dua atau lebih maka akan error, scipy.stats.mode(v), atau dengan pandas u.mode().

  Note :

Nilai rata-rata sangat dipengaruhi oleh outlier sehingga tidak cocok untuk menunjukkan central tendency dari data dengan outlier. Pada data dengan outlier sebaiknya menggunakan median karena nilainya tidak dipengaruhi dengan outlier. Sedangkan mode (modus) digunakan untuk data yang bertipe kategorikal.

- **Measures of variability (spread)**, Measures of central tendency tidak cukup untuk menggambarkan data. Diperlukan ukuran variabilitas yang mengukur penyebaran titik data.

  Standard Deviation: positive square root dari sample variance. Standard deviation lebih cocok daripada varians karena memiliki satuan yang sama dengan data points. Untuk menghitungnya dapat menggunakan statistics.stdev(), np.std(y, ddof=1), y.std(ddof=1).

  Variance: Sample variance menunjukkan secara numerik seberapa jauh titik data dari mean. Untuk menghitungnya dapat menggunakan statistics.variance(x), np.var(y, ddof=1), y.var(ddof=1).

  Range: Perbedaan antara nilai terendah dengan nilai tertinggi.

  Skewness: Skewness mengukur asimetri sampel data. Negative skewness menunjukkan bahwa ada ekor dominan di sisi kiri. Positive skewness menunjukkan ekor dominan di sisi kanan. Jika skewness mendekati 0 (misalnya antara −0.5 dan 0.5), maka dataset dianggap cukup simetris. Skewness dapat dilihat dari histogram tetapi dapat juga dengan perhitungan scipy.stats.skew(y, bias=False), z.skew().

  Percentile: Tiap dataset memiliki tiga quartiles, yang merupakan persentil yang membagi dataset menjadi empat bagian. Kuartil pertama membagi sekitar 25% item terkecil dari kumpulan data lainnya. Kuartil kedua/median kira-kira 25% item terletak di antara kuartil pertama dan kedua dan 25% lainnya antara kuartil kedua dan ketiga. Kuartil ketiga membagi sekitar 25% item terbesar dari sisa kumpulan data.

  Covariance dan Correlation Coefficient: Covariance adalah ukuran yang mengukur kekuatan dan arah hubungan antara sepasang variabel dan Correlation Coefficient atau Pearson product-moment correlation coefficient adalah ukuran lain dari korelasi antar data. Jika korelasinya positif, maka kovariansnya juga positif. Jika korelasinya negatif, maka kovariansnya negatif. Jika korelasinya lemah, maka kovariansnya mendekati nol. Correlation Coefficient yang nilainya lebih dari 0 menunjukkan korelasi positif. Jika nilainya kurang dari 0 maka korelasi negatif. Nilai 1 adalah nilai maksimum yang mungkin. Ini menunjukkan hubungan linier positif yang sempurna antara variabel. Nilai -1 adalah nilai minimum yang mungkin. Ini menunjukkan hubungan linier negatif yang sempurna antara variabel. Jika nilainya mendekati 0 maka korelasinya lemah. Untuk menghitung covariance dapat menggunakan np.cov(x_, y_), x__.cov(y__). Dan untuk menghitung correlation coeeficient dapat menggunakan scipy.stats.pearsonr(x_, y_), np.corrcoef(x_, y_).