| | | |
|---|---|---|
| | STMIK WIDYA CIPTA DHARMA SAMARINDA | **S1 - TEKNIK INFORMATIKA** |
| PRAKTIKUM PEMBELAJARAN MESIN | **DATA UNDERSTANDING** | LabSheet 02 |
| SEMESTER 5 | | Dosen :<br>**1.** PITRASACHA ADYTIA, MT<br>**2.** WAHYUNI, S.KOM, M.KOM |

# I. Tujuan

1. Mahasiswa mampu melakukan telaah data dengan beberapa metode statistika.

# II. Prosedur Praktikum

## II.1 Telaah Data

### 1. Mengimpor data ke Pandas

Pada tahap ini kita akan coba untuk mengimpor data CSV ke Pandas. Adapun dataset yang akan kita pakai adalah automobileEDA.CSV. Berikut adalah cara untuk mengimpor dataset tersebut.

```
In [2]: import pandas as pd
        path = "D:/dokumen/ngajar/machine_learning/praktikum/automobileEDA.CSV"
        df = pd.read_csv(path)
```

Pada script di atas dilakukan import library pandas. Setelah itu membuat suatu variable untuk menyimpan path dari dataset yang akan digunakan. Sesuaikan path tersebut dengan path yang dimiliki. Setelah itu dilakukan proses impor dataset. Setelah itu kita dapat melihat isi dari dataset dengan dengan cara sebagai berikut.

```
In [3]: df
```

Out[3]:

| | symboling | normalized-losses | make | aspiration | num-of-doors | body-style | drive-wheels | engine-location | wheel-base | length | ... | compression-ratio | horsepower | peak-rpm | city-mpg | highway-mpg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 122 | alfa-romero | std | two | convertible | rwd | front | 88.6 | 0.811148 | ... | 9.0 | 111.0 | 5000.0 | 21 | 27 |
| 1 | 3 | 122 | alfa-romero | std | two | convertible | rwd | front | 88.6 | 0.811148 | ... | 9.0 | 111.0 | 5000.0 | 21 | 27 |
| 2 | 1 | 122 | alfa-romero | std | two | hatchback | rwd | front | 94.5 | 0.822681 | ... | 9.0 | 154.0 | 5000.0 | 19 | 26 |
| 3 | 2 | 164 | audi | std | four | sedan | fwd | front | 99.8 | 0.848630 | ... | 10.0 | 102.0 | 5500.0 | 24 | 30 |
| 4 | 2 | 164 | audi | std | four | sedan | 4wd | front | 99.4 | 0.848630 | ... | 8.0 | 115.0 | 5500.0 | 18 | 22 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 196 | -1 | 95 | volvo | std | four | sedan | rwd | front | 109.1 | 0.907256 | ... | 9.5 | 114.0 | 5400.0 | 23 | 28 |
| 197 | -1 | 95 | volvo | turbo | four | sedan | rwd | front | 109.1 | 0.907256 | ... | 8.7 | 160.0 | 5300.0 | 19 | 25 |
| 198 | -1 | 95 | volvo | std | four | sedan | rwd | front | 109.1 | 0.907256 | ... | 8.8 | 134.0 | 5500.0 | 18 | 23 |
| 199 | -1 | 95 | volvo | turbo | four | sedan | rwd | front | 109.1 | 0.907256 | ... | 23.0 | 106.0 | 4800.0 | 26 | 27 |
| 200 | -1 | 95 | volvo | turbo | four | sedan | rwd | front | 109.1 | 0.907256 | ... | 9.5 | 114.0 | 5400.0 | 19 | 25 |

201 rows × 29 columns

***Silahkan sambal diamati, informasi apa yang di dapat dari gambar di atas?***

## 2. Eksplorasi Data Dasar

Eksplorasi data dasar biasanya dimulai dengan menginspeksi kolom-kolom beserta beberapa baris awal dari data. Untuk melihat beberapa baris awal dari data, fungsi head() dan tail() dari DataFrame. Berikut cara menggunakan fungsi head() dan tail().

```
df.head()
```

```
df.tail()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan.***

Atribut dtypes dari DataFrame menyimpan sebuah Pandas Series yang berisi daftar seluruh kolom di dalam DataFrame yang bersangkutan. Perintah berikut menghasilkan tipe dari df.dtypes, panjang df.dtypes (yakni jumlah kolom dari DataFrame df), serta isi df.dtypes itu sendiri.

```
type(df.dtypes), len(df.dtypes), df.dtypes
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan.***

```
df_noid = df.iloc[:,2:]
df_noid
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan.***

Hal lain yang juga dapat dilakukan adalah menampilkan data dengan mengikuti urutan.

```
In [8]: df_noid = df.iloc[:,2:]
        df1 = df_noid.sort_values(by="horsepower",ascending=True)
        df1.head(10)
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan.***

Selain itu kita juga dapat mengurutkan berdasarkan lebih dari satu kolom, dimana

urutan masing-masing kolom bisa berbeda-beda. Misal kolom satu terurut secara ascending, dan kolom lainnya terurut secara descending.

```
In [9]: df1 = df_noid.sort_values(by=["horsepower", "body-style"], \
         ascending=[False,True])
        df1.head(10)
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan.***

### 3. Deskripsi Data Secara Statistik

Metode eksplorasi data yang lain adalah dengan menerapkan konsep-konsep dari ilmu statistika. Pandas menyediakan cukup banyak fungsi-fungsi statistika yang dapat diterapkan pada suatu DataFrame**.**

**Tabel 1. Fungsi Statistika Pada Pandas**

| Nama fungsi | Fungsi mengembalikan... |
|---|---|
| **count** | banyaknya butir data yang bukan NA (NA = *not available*). |
| **sum** | jumlahan butir-butir data. |
| **mean** | rerata (aritmetik)[8] butir-butir data. |
| **mad** | rerata simpangan absolut (*mean absolute deviation*)[9] dari butir-butir data. |
| **median** | median[10] (aritmetik) dari butir-butir data. |
| **min** | nilai terkecil/minimum dari butir-butir data. |
| **max** | nilai terbesar/maksimum dari butir-butir data. |
| **mode** | nilai modus[11] dari butir-butir data. |
| **abs** | nilai absolut[12] numerik setiap butir data. |
| **prod** | hasil perkalian setiap butir data. |
| **quantile** | nilai pada kuantil[13] tertentu dari butir-butir data; argumen fungsi adalah kuantil yang diinginkan antara 0 hingga 1; nilai kuartil pertama = nilai kuantil pada posisi 0.25. |
| **std** | nilai simpangan baku sampel[14] (bukan populasi) menggunakan koreksi Bessel; kumpulan butir data yang dihitung simpangan bakunya dianggap sebagai kumpulan sampel, bukan seluruh populasi. |

| | |
|---|---|
| **var** | nilai varian sampel;[15] kumpulan butir data yang dihitung varian-nya dianggap sebagia kumpulan sampel. |
| **sem** | galat standar dari rerata.[16] |
| **skew** | nilai ukuran kecondongan[17] (*skewness*) dari distribusi |
| **kurt** | nilai ukuran keruncingan[18] (*kurtosis*) dari distribusi |
| **cumsum** | jumlahan kumulatif data. |
| **cumprod** | perkalian kumulatif data. |
| **cummax** | nilai maksimum kumulatif data |
| **cummin** | nilai minimum kumulatif data. |

Untuk mendapatkan ringkasan statistik secara cepat, DataFrame menyediakan fungsi describe() yang akan menghasilkan sebuah DataFrame baru yang berisi ringkasan statistik dari DataFrame yang padanya describe() diterapkan.

```
In [11]: df_noid.describe()
```

Out[11]:

| | wheel-base | length | width | height | curb-weight | engine-size | bore | stroke | compression-ratio | horsepower | peak-rpm | city-mpg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 201.000000 | 201.000000 | 201.000000 | 201.000000 | 201.000000 | 201.000000 | 201.000000 | 197.000000 | 201.000000 | 201.000000 | 201.000000 | 201.000000 |
| mean | 98.797015 | 0.837102 | 0.915126 | 53.766667 | 2555.666667 | 126.875622 | 3.330692 | 3.256904 | 10.164279 | 103.405534 | 5117.665368 | 25.179104 |
| std | 6.066366 | 0.059213 | 0.029187 | 2.447822 | 517.296727 | 41.546834 | 0.268072 | 0.319256 | 4.004965 | 37.365700 | 478.113805 | 6.423220 |
| min | 86.600000 | 0.678039 | 0.837500 | 47.800000 | 1488.000000 | 61.000000 | 2.540000 | 2.070000 | 7.000000 | 48.000000 | 4150.000000 | 13.000000 |
| 25% | 94.500000 | 0.801538 | 0.890278 | 52.000000 | 2169.000000 | 98.000000 | 3.150000 | 3.110000 | 8.600000 | 70.000000 | 4800.000000 | 19.000000 |
| 50% | 97.000000 | 0.832292 | 0.909722 | 54.100000 | 2414.000000 | 120.000000 | 3.310000 | 3.290000 | 9.000000 | 95.000000 | 5125.369458 | 24.000000 |
| 75% | 102.400000 | 0.881788 | 0.925000 | 55.500000 | 2926.000000 | 141.000000 | 3.580000 | 3.410000 | 9.400000 | 116.000000 | 5500.000000 | 30.000000 |
| max | 120.900000 | 1.000000 | 1.000000 | 59.800000 | 4066.000000 | 326.000000 | 3.940000 | 4.170000 | 23.000000 | 262.000000 | 6600.000000 | 49.000000 |

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan. Informasi apa yang Anda dapatkan?***

```
In [13]: df_noid.describe(include='all')
```

Out[13]:

| | make | aspiration | num-of-doors | body-style | drive-wheels | engine-location | wheel-base | length | width | height | ... | compression-ratio | horsepower | peak-rpm | city-m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 201 | 201 | 201 | 201 | 201 | 201 | 201.000000 | 201.000000 | 201.000000 | 201.000000 | ... | 201.000000 | 201.000000 | 201.000000 | 201.000 |
| unique | 22 | 2 | 2 | 5 | 3 | 2 | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | |
| top | toyota | std | four | sedan | fwd | front | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | |
| freq | 32 | 165 | 115 | 94 | 118 | 198 | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | |
| mean | NaN | NaN | NaN | NaN | NaN | NaN | 98.797015 | 0.837102 | 0.915126 | 53.766667 | ... | 10.164279 | 103.405534 | 5117.665368 | 25.179 |
| std | NaN | NaN | NaN | NaN | NaN | NaN | 6.066366 | 0.059213 | 0.029187 | 2.447822 | ... | 4.004965 | 37.365700 | 478.113805 | 6.423 |
| min | NaN | NaN | NaN | NaN | NaN | NaN | 86.600000 | 0.678039 | 0.837500 | 47.800000 | ... | 7.000000 | 48.000000 | 4150.000000 | 13.000 |
| 25% | NaN | NaN | NaN | NaN | NaN | NaN | 94.500000 | 0.801538 | 0.890278 | 52.000000 | ... | 8.600000 | 70.000000 | 4800.000000 | 19.000 |
| 50% | NaN | NaN | NaN | NaN | NaN | NaN | 97.000000 | 0.832292 | 0.909722 | 54.100000 | ... | 9.000000 | 95.000000 | 5125.369458 | 24.000 |
| 75% | NaN | NaN | NaN | NaN | NaN | NaN | 102.400000 | 0.881788 | 0.925000 | 55.500000 | ... | 9.400000 | 116.000000 | 5500.000000 | 30.000 |
| max | NaN | NaN | NaN | NaN | NaN | NaN | 120.900000 | 1.000000 | 1.000000 | 59.800000 | ... | 23.000000 | 262.000000 | 6600.000000 | 49.000 |

11 rows × 27 columns

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

***Informasi apa yang Anda dapatkan?***

***Apa perbedaan dengan script sebelumnya?***

### 4. Deskripsi Pusat Data: Rerata Aritmetik, Median, dan Modus

Butir-butir data setiap kolom dapat dipandang sebagai sampel dari suatu distribusi statistik tertentu. Deskripsi pusat data pada dasarnya memberikan gambaran mengenai lokasi tempat berkumpulnya kebanyakan butir data pada distribusi tersebut. Terdapat tiga besaran pusat data yang paling banyak dipergunakan, yakni rerata aritmetik (mean), median, dan modus. Pandas mengasumsikan penggunaan konsep modus hanya pada data-data non-numerik. Tidak hanya itu, konsep rerata aritmetik dan median hanya diterapkan pada data-data numerik.

```
In [18]: df_noid.mean()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

```
df_noid[['length', 'width', 'height']].mean()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

```
In [16]: df_noid[['length', 'width', 'height','price']].median()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

```
In [17]: df_noid[['make']].mode()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

### 5. Deskripsi Sebaran Data: Rentang, Kuartil, Simpangan Baku, dan Pencilan

Jika deskripsi pusat data menunjukkan lokasi butir-butir data secara umum berkumpul, maka deskripsi sebaran data menggambarkan seberapa jauh butir-butir data

menyebar dari pusat data. Ada beberapa besaran yang dapat digunakan untuk memberikan gambaran tentang sebaran data. Besaran-besaran tersebut antara lain meliputi rentang, kuantil, simpangan baku, varian, dan pencilan.

- Rentang Rentang (range) atau jangkauan didefinisikan sebagai selisih antara nilai maksimum dan minimum pada kumpulan data. Nilai rentang yang besar dapat menggambarkan bahwa data cenderung tersebar, dan sebaliknya rentang yang kecil dapat menunjukkan bahwa data cenderung mengumpul. Namun demikian, ini tidak sepenuhnya dapat dijadikan pegangan, khususnya jika nilai maksimum atau minimum data ternyata merupakan pencilan. Tidak ada fungsi khusus dalam Pandas untuk menghitung rentang karena ini dengan mudah dapat dihitung memakai fungsi min() dan max(). Di sisi lain, karena nilai rentang hanya bergantung pada dua butir data saja, maka besaran ini biasanya hanya cocok dipakai untuk dataset berukuran kecil.

```
In [21]: df_noid[['price']].max()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

- Sebuah kuantil (quantile) dari suatu kumpulan data didefinisikan sebagai sebuah nilai titik potong yang menentukan berapa banyak butir data yang bernilai lebih kecil darinya dan berapa banyak yang lebih besar darinya. Untuk setiap bilangan bulat $k \geq 2$, $k$-kuantil adalah nilai-nilai yang besarnya membagi himpunan data menjadi $k$ bagian yang berukuran sama. Untuk setiap $k$, terdapat $k - 1$ nilai atau butir data yang berfungsi sebagai $k$-kuantil. Ada beberapa istilah khusus untuk kuantil.
    - 2-kuantil hanya terdiri dari satu butir data yakni median yang membagi kumpulan data menjadi dua bagian yang sama besar: separuh berada di bawahnya dan separuh sisanya berada di atasnya.
    - 4-kuantil terdiri dari tiga titik atau tiga butir data yang disebut kuartil. Ketiga kuartil tersebut secara bersama-sama membagi kumpulan data menjadi empat bagian yang sama besar. Kuartil pertama membagi data sehingga 25% data berada di bawahnya dan 75% data berada di atasnya. Kuartil kedua membagi data sehingga 50% data berada di bawahnya dan 50% berada di atasnya. Lalu, kuartil ketiga membagi data sehingga 75% data berada di bawahnya dan 25% data berada di atasnya.
    - 100-kuantil, disebut juga persentil, meliputi 99 nilai yang secara bersama-sama membagi data menjadi 100 bagian yang sama besar.

```
In [23]: df_noid[['price']].quantile(q=0.75)
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

- Simpangan Baku, Ukuran sebaran data lain yang lazim dipakai adalah simpangan baku dan varian. Varian didefinisikan sebagai rerata dari jumlah kuadrat jarak antara setiap butir data dengan rerata kumpulan data. Pandas menyediakan fungsi var() dan std() untuk menghitung varian dan simpangan baku.

```
In [24]: df_noid[['price']].var(), df_noid[['price']].std()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

### 6. Tabel Frekuensi

Bagian dari analisis data, khususnya untuk kolom-kolom yang bertipe nominal atau adalah menampilkan tabel frekuensinya.

```
In [25]: df_noid['make'].value_counts()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

### 7. Pengelompokan Data Berdasarkan Kolom

Analisis data juga dapat dilakukan dengan mengelompokkan data berdasarkan kolom tertentu.

```
In [26]: df_noid.groupby('make')[['price']]\
          .mean().sort_values(by='price',ascending=False)
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

### 8. Analisis Korelasi

Analisis korelasi dilakukan pada data untuk mengetahui bagaimana hubungan dependensi antar dua buah kolom numerik pada data. Rentang nilai korelasi adalah antara -1 dan 1. Jika nilai korelasi -1 = korelasi negatif, 0 = tidak ada korelasi linear, +1 = korelasi positif.

Korelasi negative berarti hubungan antar kolom tersebut adalah berbanding terbalik atau bertolak belakang. Jika semakin mendekati nilai -1 artinya hubungan antar kolom tersebut bertolak belakang. Semakin tinggi nilai kolom X, maka akan semakin rendah nilai kolom Y.

Korelasi positif artinya hubungan antar kolom tersebut berbanding lurus. Semakin mendekati angka 1, maka hubungan antar kolom tersebut sangat berbanding lurus. Semakin tinggi nilai kolom X, maka akan semakin tinggi nilai kolom Y.

Jika tidak ada korelasi, maka sudah dipastikan bahwa tidak ada hubungan antar kolom

tersebut.

```
In [27]: df_noid.corr()
```

***Silahkan lakukan perintah diatas, dan amati apa yang akan muncul dilembar kerja. Berikan penjelasan!***

***Informasi apa yang Anda dapatkan?***

## III. Laporan Praktikum

Silahkan kumpulkan file berekstensi .ipynb ke asisten lab untuk dinilai

## IV. Referensi


1. **Python Data Science HandBook**
2. **Modul DTS kominfo untuk dosen dan intstruktur**