|  |  |  |
| --- | --- | --- |
| **Nama:**  Adriansyah Maulana Putra  **NIM:** 064002200046 | C:\Users\RPL-SI 02\Pictures\288px-Trisakti_Logo.svg.png | **MODUL 8**  **Nama Dosen:**  **Dedy Sugiarto** |
| **Hari/Tanggal:**  **Hari,** Senin 7 Agustus **2023** | **Praktikum Statistika** | **Nama Asisten Labratorium**  **1. Elen Fadilla Estri**  **064002000008**  **2. Rukhy Zaifa Aduhalim**  **064002000041** |

**Eksplorasi Data Menggunakan Python**

1. **Teori Singkat**

histogram berguna untuk memberikan gambaran ukuran tendensi sentral dan kesimetrisan data pengamatan. Penyajian grafis lainnya yang bisa merangkum informasi lebih detail mengenai distribusi nilai-nilai data pengamatan adalah Box and Whisker Plots atau lebih sering disebut dengan BoxPlot atau Box-Plot (kotak-plot) saja. Seperti namanya, Box and Whisker, bentuknya terdiri dari Box (kotak) dan whisker.

Box-plot atau boxplot (juga dikenal sebagai diagram box-and-whisker) merupakan suatu box (kotak berbentuk bujur sangkar). Boxplot adalah salah satu cara dalam statistik deskriptif untuk menggambarkan secara grafik dari data numeris melalui lima ukuran sebagai berikut:

* Nilai observasi terkecil,
* Kuartil terendah atau kuartil pertama (Q1), yang memotong 25% dari data terendah
* Median (Q2) atau nilai pertengahan,
* Kuartil tertinggi atau kuartil ketiga (Q3), yang memotong 25% dari data terbesar
* Nilai observasi terbesar.

Dalam boxplot juga ditunjukkan, jika ada, nilai outlier dari observasi. Boxplot dapat digunakan untuk menunjukkan perbedaan antara populasi tanpa menggunakan asumsi distribusi statistik yang mendasarinya. Karenanya, boxplot tergolong dalam statistik non-parametrik. Jarak antara bagian-bagian dari box menunjukkan derajat dispersi (penyebaran) dan skewness (kecondongan) dalam data. Dalam penggambarannya, boxplot dapat digambarkan secara horizontal maupun vertikal.

1. **Alat dan Bahan**

Hardware : Laptop/PC

Software : R Studio

1. **Elemen Kompetensi**
   1. Latihan pertama – Praktikum

1. Buka Jupyter Notebook atau Google Colab di Browser

|  |
| --- |
|  |

2. Lalu jalankan script berikut dan berikan output (gunakan nama variable data dengan nama masing-masing)

|  |
| --- |
| import pandas as pd  from pandas.tools import plotting  import matplotlib.pyplot as plt  import numpy as np  from sklearn.model\_selection import train\_test\_split, cross\_val\_score, KFold, GridSearchCV  from sklearn.linear\_model import LogisticRegression  from sklearn.tree import DecisionTreeClassifier,export\_graphviz  from sklearn.metrics import confusion\_matrix, accuracy\_score  from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifier |

Output:

|  |
| --- |
|  |

3. Script

|  |
| --- |
| data\_nama = pd.read\_csv('C:/prakstatik/houseprice.csv')  data\_nama.head(5) |

Output:

|  |
| --- |
|  |

4. Boxplot dari harga rumah/Price

|  |
| --- |
| col\_list=['Price']  numhouse = data\_nama[data\_nama.columns[data\_nama.columns.isin(col\_list)]]  plt.figure(figsize=(10,5))  numhouse.boxplot(sym='r\*', grid=False)  plt.show() |

Output:

|  |
| --- |
|  |

5. Histogram dari Price

|  |
| --- |
| plt.figure(figsize=(15,5))  plt.subplot(121)  data\_nama['Price'].plot.hist(bins=10, title='Price')  plt.show() |

Output:

|  |
| --- |
|  |

6. Scatter dari Price

|  |
| --- |
| col\_list=['Price', 'SqFt']  numhouse = data\_nama[data\_nama.columns[data\_nama.columns.isin(col\_list)]]  numhouse.plot.scatter(x='SqFt', y='Price') |

Output:

|  |
| --- |
|  |

7. Group Boxplot Berdasarkan Bedrooms

|  |
| --- |
| col\_list=['Price', 'Bedrooms']  numhouse = data\_nama[data\_nama.columns[data\_nama.columns.isin(col\_list)]]  plt.figure(figsize=(15,5))  numhouse.boxplot(by='Bedrooms')  plt.show() |

Output:

|  |
| --- |
|  |

* 1. Latihan Kedua – Tugas

1. Buatlah Scatter Plot Harga Rumah Berdasarkan Bedrooms!

Scritpt:

|  |
| --- |
| col\_list=['Price', 'Bedrooms']  numhouse = data\_adrian[data\_adrian.columns[data\_adrian.columns.isin(col\_list)]]  numhouse.plot.scatter(x='Bedrooms',y='Price') |

Output:

|  |
| --- |
|  |

Penjelasan: : Memanggil kolom ‘Price’ dan ‘Bedrooms’. Mengambil kolom dari dataframe data\_adrian lalu memeriksa dengan metode isin(col\_list) apakah data yang ingin dipanggil (kolom Price dan Bedrooms di col\_list) ada dalam dataframe data\_adrian. Memanggil metode plot.scatter() dari dataframe numhouse serta menempatkan kolom Bedrooms ke sumbu x dan kolom Price ke sumbu y.

2. Buatlah Scatter Plot Harga Rumah berdasarkan Bathrooms!

Script:

|  |
| --- |
| col\_list=['Bathrooms', 'Bathrooms']  numhouse = data\_adrian[data\_adrian.columns[data\_adrian.columns.isin(col\_list)]]  numhouse.plot.scatter(x='Bathrooms',y='Bathrooms') |

Output:

|  |
| --- |
|  |

Penjelasan: Memanggil kolom ‘Price’ dan ‘Bathrooms’. Mengambil kolom dari dataframe data\_fahmi lalu memeriksa dengan metode isin(col\_list) apakah data yang ingin dipanggil (kolom Price dan Bathrooms di col\_list) ada dalam dataframe data\_adrian. Memanggil metode plot.scatter() dari dataframe numhouse serta menempatkan kolom Bathrooms ke sumbu x dan kolom Price ke sumbu y

3. Buatlah Grouped Boxplot berdasarkan Bathrooms dengan Price sebagai perbandingannya!

Script:

|  |
| --- |
| col\_list=['Price', 'Bathrooms']  numhouse = data\_adrian[data\_adrian.columns[data\_adrian.columns.isin(col\_list)]]  plt.figure(figsize=(15,5))  numhouse.boxplot(by='Bathrooms')  plt.show() |

Output:

|  |
| --- |
|  |

Penjelasan: Memanggil kolom ‘Price’ dan ‘Bathrooms’. Mengambil kolom dari dataframe data\_adrian lalu memeriksa dengan metode isin(col\_list) apakah data yang ingin dipanggil (kolom Price dan Bathrooms di col\_list) ada dalam dataframe data\_adrian. Membuat gambar plot dengan ukuran lebar 15 dan tinggi 5. Menggunakan dataframe numhouse yang berisi kolom Price dan Bathrooms ditampilkan dengan metode boxplot(). Menampilkan gambar plot box yang sudah dibuat.

1. **File Praktikum**

Github Repository:

|  |
| --- |
|  |

1. **Soal Latihan**

Soal:

1. Apa yang dimaksud Exploratory Data Analysis?
2. Mengapa EDA diperlukan melakukan dalam melakukan analisis data?

Jawaban:  
1. Exploratory Data Analysis (EDA) adalah suatu pendekatan dalam analisis data yang bertujuan untuk memahami karakteristik, pola, dan hubungan dalam data. EDA melibatkan penerapan berbagai teknik statistik dan visualisasi untuk mengidentifikasi informasi penting, anomali, pola tersembunyi, serta sifat-sifat menarik lainnya dari data. Tujuan utama dari EDA adalah untuk mendapatkan wawasan awal tentang data sebelum melakukan analisis lebih lanjut atau membangun model statistik.

2. EDA diperlukan dalam melakukan analisis data karena memiliki beberapa manfaat penting:

* Pemahaman Data Awal: EDA membantu Anda untuk memahami struktur dan karakteristik dasar dari data Anda. Ini membantu dalam mengidentifikasi jenis data, nilai yang hilang, outliers, dan pola-pola awal yang mungkin tidak terlihat pada pandangan pertama.
* Pendekatan Visual: Teknik visualisasi dalam EDA, seperti plot grafik, histogram, scatter plot, dan lainnya, membantu menggambarkan data dengan cara yang mudah dimengerti. Hal ini memudahkan identifikasi pola, hubungan, dan distribusi data secara visual.
* Pengidentifikasian Anomali: EDA memungkinkan Anda untuk mengidentifikasi anomali atau nilai yang tidak wajar dalam data, seperti outliers. Anomali ini bisa berpotensi mempengaruhi hasil analisis atau model yang akan dibangun.
* Pemilihan Fitur (Feature Selection): Melalui EDA, Anda dapat menganalisis korelasi antara fitur-fitur (kolom-kolom) dalam data dan memutuskan fitur-fitur mana yang paling relevan untuk analisis atau model selanjutnya.
* Persiapan Data yang Baik: EDA membantu dalam persiapan data yang lebih baik sebelum melakukan analisis lebih lanjut. Ini melibatkan pembersihan data, mengisi nilai yang hilang, atau mengubah format data agar lebih sesuai dengan kebutuhan analisis.
* Hipotesis Awal: EDA dapat membantu dalam membentuk hipotesis awal tentang hubungan antara variabel atau asumsi tentang data. Hipotesis-hipotesis ini kemudian dapat diuji melalui analisis lebih lanjut.
* Penghematan Waktu dan Sumber Daya: Dengan memahami data dengan baik melalui EDA, Anda dapat menghindari kesalahan atau kesalahan yang mungkin terjadi dalam analisis lebih lanjut. Ini dapat menghemat waktu dan sumber daya yang mungkin diperlukan untuk perbaikan.

1. **Kesimpulan**
   1. Dalam pengerjaan praktikum Statistika, kita melakukan analisis karakteristik, hubungan, dan pola yang ada dalam data.
   2. Kita juga dapat mengetahui cara untuk menampilkan hubungan antar data dengan diagram scatter serta boxplot
2. **Cek List (✔)**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Penyelesaian** | |
| **Selesai** | **Tidak Selesai** |
| **1.** | Latihan Pertama |  |  |
| **2.** | Latihan Kedua |  |  |

1. **Formulir Umpan Balik**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Waktu Pengerjaan** | **Kriteria** |
| **1.** | Latihan Pertama | 20 Menit | Menarik |
| **2.** | Latihan Kedua | 15 Menit | Menarik |

Keterangan:

1. Menarik
2. Baik
3. Cukup
4. Kurang