DATA SCIENCE TASK

IGD Raditya Wibhawa MN

PT Bina Pertiwi

LATIHAN 1

Presentation by IGD Raditya Wibhawa MN

LATIHAN 1

```python
import pandas as pd

# Membaca dataset
df = pd.read_csv('dataset.csv')

def clean(df):
  if(df.isnull().sum()/len(df)) > 0.1:
    # jika null value >10% dari total data maka fill dengan nilai median
    df = df.fillna(df.median())
  elif(df.isnull().sum()/len(df)) <= 0.1:
    # jika null value <10% dari total data maka akan di drop
    df = df.dropna()
  # Menghapus baris yang memiliki nilai yang tidak sesuai
  df = df[df['durasi_film'] > 0]
  df = df[df['kapasitas_auditorium'] > 0]
  df = df[df['tiket_terjual'] > 0]
  df = df[df['harga_tiket'] > 0]

  # Menghapus baris yang duplikat
  df = df.drop_duplicates()
  return df
```

Berikut adalah salah satu metode untuk membersihkan data. Ketika total null value berjumlah 10% dari total data, maka akan diisi dengan nilai median. Ketika total null value dibawah dari 10% dari total data, maka akan di drop dari tabel. Kemudian menghapus baris yang yang durasi_film, kapasitas_auditorium, tiket_terjual, dan harga_tiket yang nilainya 0, dengan pertimbangan semua feature ini tidak mungkin 0. Terakhir menghapus baris yang duplikat

LATIHAN 2

Presentation by IGD Raditya Wibhawa MN

LATIHAN 2

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Membaca dataset
df = pd.read_csv('dataset.csv')

# Membuat histogram untuk melihat distribusi dari jarak, durasi, harga, driver_rating, dan customer_rating
plt.figure(figsize=(10, 6))
plt.subplot(2, 3, 1)
sns.histplot(data=df, x='jarak')
plt.title('Distribusi Jarak')

plt.subplot(2, 3, 2)
sns.histplot(data=df, x='durasi')
plt.title('Distribusi Durasi')

plt.subplot(2, 3, 3)
sns.histplot(data=df, x='harga')
plt.title('Distribusi Harga')

plt.subplot(2, 3, 4)
sns.histplot(data=df, x='driver_rating')
plt.title('Distribusi Driver Rating')

plt.subplot(2, 3, 5)
sns.histplot(data=df, x='customer_rating')
plt.title('Distribusi Customer Rating')

plt.tight_layout()
plt.show()
```

Code disebelah untuk membuat visualisasi histogram dari 5 variable. Histogram ini berfungsi untuk melihat persebaran distribusi dari masing masing variable

LATIHAN 2

```
# Membuat scatter plot untuk melihat hubungan antara jarak dan harga
sns.scatterplot(data=df, x='jarak', y='harga')
plt.title('Hubungan antara Jarak dan Harga')
plt.show()

# Membuat scatter plot untuk melihat hubungan antara jarak dan durasi
sns.scatterplot(data=df, x='jarak', y='durasi')
plt.title('Hubungan antara Jarak dan Durasi')
plt.show()

# Membuat scatter plot untuk melihat hubungan antara durasi dan harga
sns.scatterplot(data=df, x='durasi', y='harga')
plt.title('Hubungan antara Durasi dan Harga')
plt.show()
```

Code disebelah adalah untuk membuat Scatter Plot dari 2 variable numerik. Scatter Plot sendiri berfungsi untuk menunjukkan apakah ada korelasi antara 2 variable numerik

LATIHAN 2

```
# Membuat heatmap untuk melihat korelasi antara fitur-fitur dalam dataset
corr = df.corr()
sns.heatmap(corr, annot=True)
plt.title('Korelasi antara Fitur-fitur dalam Dataset')
plt.show()
```

Code diatas untuk membuat visualisasi heatmap yang berfungsi untuk melihat korelasi antar variable dengan seluruh variable lainnya. Visualisasi Heatmap berguna untuk mengidentifikasi pola tertentu berdasarkan warna yang muncul pada visualisasinya, biasanya heatmap menggunakan warna merah ketika ada hubungan yang kuat, sedangkan biru untuk hubungan yang lemah

LATIHAN 2

```
# membuat barplot antara driver_rating dan harga
plt.bar(df['driver_rating'], df['harga'])
plt.show()

# membuat barplot antara driver_rating dan jarak
plt.bar(df['driver_rating'], df['jarak'])
plt.show()

# membuat barplot antara driver_rating dan durasi
plt.bar(df['driver_rating'], df['durasi'])
plt.show()
```

Code disebelah untuk membuat visualisasi barplot yang bertujuan untuk membandingkan nilai numerik antar kategori, disini saya membuat 3 visualisasi barplot, yaitu antara driver_rating dan harga, driver_rating dan jarak, driver_rating dan durasi.

LATIHAN 2

```
# membuat boxplot antara customer_rating dan harga
sns.boxplot(x = 'customer_rating', y = 'harga', data = df)
plt.show()

# membuat boxplot antara customer_rating dan durasi
sns.boxplot(x = 'customer_rating', y = 'durasi', data = df)
plt.show()

# membuat boxplot antara customer_rating dan jarak
sns.boxplot(x = 'customer_rating', y = 'jarak', data = df)
plt.show()
```

Code disebelah untuk membuat 3 visualisasi boxplot. Variable yang dibandingkan adalah customer_rating dan harga, customer_rating dan durasi, customer_rating dan jarak. Boxplot sendiri berguna untuk mengecek adanya outlier, serta menunjukkan distribusi dari variable numerik masing masing kategori customer_rating/

LATIHAN 3

Presentation by IGD Raditya Wibhawa MN

LATIHAN 3

```python
import pandas as pd
from scipy.stats import skew, kurtosis

# Membaca dataset
df = pd.read_csv('dataset.csv')

# Menghitung statistik deskriptif untuk fitur numerik
desc = df[['umur', 'lama_bekerja', 'gaji']].describe()
print(desc)
```

Descriptive Statistics

Code disebelah menggunakan df.describe() untuk menunjukkan Statistik Deskriptif yang terdiri dari :

1. Count : Jumlah baris data yang tidak mengandung NaN
2. Mean : Nilai rata-rata
3. Standard Deviation : Simpangan baku dari mean
4. Min : Nilai terkecil tiap kolom
5. 25% : Q1(Kuartil pertama) dalam suatu kolom
6. 50% : Q2(Kuartil kedua) dalam suatu kolom
7. 75% : Q3(Kuartil ketiga) dalam suatu kolom
8. Max : Nilai terbesar tiap kolom

LATIHAN 3

```
# Hitung skewness
skew = df[['umur', 'lama_bekerja', 'gaji']].skew()
print(skew)
```

Descriptive Statistics

Code disebelah untuk menunjukkan Statistik Deskriptif berupa :

Skewness : Skewness menunjukkan apakah asimetri distribusi dari nilai sebuah variable tailnya ke arah kanan(Skewness Positif), tailnya ke arah kiri(Skewness Negatif), atau skewness nol yang artinya data terdistribusi dengan baik.

LATIHAN 3

```
# Calculate interquartile range (IQR)
q1_umur = df['umur'].quantile(0.25)
q3_umur = df['umur'].quantile(0.75)
iqr_umur = q3_umur - q1_umur

q1_lama_bekerja = df['lama_bekerja'].quantile(0.25)
q3_lama_bekerja = df['lama_bekerja'].quantile(0.75)
iqr_lama_bekerja = q3_lama_bekerja - q1_lama_bekerja

q1_gaji = df['gaji'].quantile(0.25)
q3_gaji = df['gaji'].quantile(0.75)
iqr_gaji = q3_gaji - q1_gaji

print(iqr_umur)
print(iqr_lama_bekerja)
print(iqr_gaji)
```

Descriptive Statistics

Code disebelah untuk menunjukkan Statistik Deskriptif berupa :

Interquartile Range(IQR) : IQR digunakan untuk menemukan rentang dimana sebagian besar data berada. Untuk menemukan IQR cukup dengan mengurangkan hasil dari Q3 dengan Q1.

LATIHAN 3

Inference Statistics

```
# Chi-square test
import pandas as pd
from scipy.stats import chi2_contingency
df = pd.read_csv("dataset.csv")
table = pd.crosstab(df.jenis_kelamin, df.pendidikan)
chi2, pval, dof, expected = chi2_contingency(table)
print(pval)
```

Code disebelah untuk menunjukkan Statistik Inferensi berupa :

Chi-Square test : Chi-square test adalah Hypothesis Testing yang digunakan untuk menentukan apakah ada hubungan asosiasi yang signifikan antara 2 kategori yang berbeda

LATIHAN 3

```python
# two sample t-test

import pandas as pd
from scipy.stats import ttest_ind

df = pd.read_csv("dataset.csv")

laki_laki = df.gaji[df['jenis_kelamin'] == 'L']
perempuan = df.gaji[df['jenis_kelamin'] == 'P']

tstat, pval = ttest_ind(laki_laki, perempuan)
print(pval)
```

Inference Statistics

Code disebelah untuk menunjukkan Statistik Inferensi berupa :

Two sample t-test : Two sample t-test digunakan untuk menentukan apakah ada perbedaan yang signifikan pada gaji individu dengan jenis kelamin laki laki dan perempuan. Two sample t-test digunakan apabila kategori didalam suatu variable itu adalah binary

LATIHAN 3

```
# ANOVA dan Tukey's Range Test

from scipy.stats import f_oneway

df = pd.read_csv('dataset.csv')

s1 = df.gaji[df['pendidikan'] == 'S1']
s2 = df.gaji[df['pendidikan'] == 'S2']
s3 = df.gaji[df['pendidikan'] == 'S3']

f_stat, pval = f_oneway(s1,s2,s3)
```

```
# Jika pval <= 0.05, berarti ada perbedaan yang signifikan
# antara S1, S2, S3 terhadap gaji yang diperoleh

from statsmodels.stats.multicomp import pairwise_tukeyhsd
tukey_results = pairwise_tukeyhsd(df.gaji, df.pendidikan, 0.05)
print(tukey_results)
```

Inference Statistics

Code disebelah untuk menunjukkan Statistik Inferensi berupa :

ANOVA dan Tukey's Range test : ANOVA dan tukey's range test digunakan untuk menunjukkan apakah ada perbedaan yang signifikan antara kategori yang berjumlah 3 atau lebih. Code disebelah bertujuan untuk menunjukkan apakah ada perbedaan yang signifikan antara gaji terhadap 3 kategori pendidikan. Jika pval ANOVA menunjukkan nilai <=0.05, dilanjutkan dengan menggunakan Tukey's Range Test untuk menunjukkan gaji dari tingkat pendidikan mana yang memiliki perbedaan secara signifikan

LATIHAN 3

```python
from scipy.stats import binomtest

sample_data = df.sample(n=100)

# Perform a two-sided binomial test
p_value = binomtest(len(sample_data['jenis_kelamin'] == 'P'), n=100, p=0.25, alternative='two-sided')

print(p_value)
```

Inference Statistics

Code diatas untuk menunjukkan Statistik Inferensi berupa :

Binomial Test : Binomial test adalah test yang digunakan untuk menunjukkan apakah sebuah sample dapat merepresentasikan hypothesis dari keseluruhan data. Sample yang diambil adalah 100. Asumsikan :

1. Null Hypothesis : 25% dari total pekerja adalah perempuan
2. Alternative Hypothesis : Tidak benar 25% dari total pekerja adalah perempuan

Dengan menggunakan code diatas, apabila p_value <= 0.05, maka dengan sample yang digunakan, Null Hypothesis dapat dibantah. Begitu pula sebaliknya.

LATIHAN 4

Presentation by IGD Raditya Wibhawa MN

LATIHAN 4

Penjelasan model yang digunakan

Model Machine Learning yang akan saya gunakan adalah KMeans Clustering, task yang diberikan untuk menentukan individu mana yang tertarik pada penawaran kartu kredit terbaru, namun karena tidak ada feature/kolom yang bisa dijadikan label, maka saya memutuskan untuk menggunakan Clustering untuk mengelompokkan mana yang tertarik dan tidak tertarik

Presentation by IGD Raditya Wibhawa MN

LATIHAN 4

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

```
df = pd.read_csv('dataset.csv')
```

Disini saya akan menggunakan Clustering untuk memprediksi individu mana saja yang tertarik pada penawaran kartu kredit baru karena tidak ada kolom yang bisa dijadikan label apakah tertarik atau tidaknya. Code disamping berikut untuk mengimport library yang diperlukan dan membaca dataset yang digunakan. Untuk code selanjutnya, disini saya berasumsi kalau data telah bersih dan tinggal di transformasi.

LATIHAN 4

```
# Convert gender values to binary (0 for Pria, 1 for Wanita)
gender_mapping = {'Pria': 0, 'Wanita': 1}
df['jenis_kelamin'] = df['jenis_kelamin'].map(gender_mapping)

# Standardize the 'pendapatan' and 'pengeluaran_bulanan' columns using StandardScaler
scaler = StandardScaler()
df[['usia','pendapatan', 'pengeluaran_bulanan']] = scaler.fit_transform(df[['usia','pendapatan', 'pengeluaran_bulanan']])
```

Pada bagian ini saya mengubah nilai categorical di kolom jenis kelamin menjadi numeric, yang tadinya Pria dan Wanita, jadi 0 dan 1 agar mudah dibaca ketika penerapan clusteringnya. Setelah itu saya menerapkan standarisasi untuk kolom usia, pendapatan, pengeluaran bulanan.

LATIHAN 4

```
features = ['usia', 'jenis_kelamin', 'pendapatan', 'pengeluaran_bulanan', 'jml_kartu_kredit']
X = df[features]
```

Pada bagian ini saya memilih usia, jenis kelamin, pendapatan, pengeluaran bulanan dan jumlah kartu kredit sebagai fitur

LATIHAN 4

```python
# Determine the optimal number of clusters using the Elbow Method
inertia_values = []
possible_clusters = range(1, 11)  # Try different numbers of clusters

for k in possible_clusters:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    inertia_values.append(kmeans.inertia_)

# Plot the Elbow Method graph
plt.plot(possible_clusters, inertia_values, marker='o')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia (Within-Cluster Sum of Squares)')
plt.title('Elbow Method')
plt.show()
```

Pada bagian ini saya menggunakan Elbow Method untuk menentukan berapa jumlah cluster yang paling optimal

LATIHAN 4

Training

```
# Train KMeans clustering model
k = 2  # Number of clusters
kmeans = KMeans(n_clusters=k, random_state=42)
df['cluster'] = kmeans.fit_predict(X)
```

Pada bagian ini saya menggunakan KMeans untuk menentukan cluster dari feature yang telah ditentukan sebelumnya, untuk jumlah clusternya sendiri saya memilih 2 karena yang dibutuhkan adalah kelompok yang tertarik dan tidak tertarik

LATIHAN 4

Evaluation

```
# Calculate silhouette score to evaluate the quality of clustering
silhouette_avg = silhouette_score(X, df['cluster'])
print("Silhouette Score:", silhouette_avg)
```

Untuk evaluasi KMeans sendiri saya menggunakan Silhoutte Score. Silhoutte Score ini menentukan seberapa akurat hasil clustering yang telah ditentukan tadi

LATIHAN 4

Validation

```
from sklearn.model_selection import cross_val_score

# Assuming X contains your feature data
k = 2  # Number of clusters

kmeans = KMeans(n_clusters=k, random_state=42)
silhouette_scores = cross_val_score(kmeans, X, cv=10, scoring='neg_mean_silhouette_score')

# Calculate the mean silhouette score
mean_silhouette_score = silhouette_scores.mean()

print("Mean Silhouette Score : ", mean_silhouette_score)
```

Untuk memvalidasi hasil Silhoutte Score, saya menggunakan 10 Fold Cross Validation agar memastikan hasilnya sesuai dengan yang diharapkan

LATIHAN 5

Presentation by IGD Raditya Wibhawa MN

LATIHAN 5

Beberapa insight yang telah ditemukan dari Visualisasi data adalah :

1. Brand yang paling sering di beri rating adalah SEPHORA COLLECTION
2. Brand yang paling sering di review adalah SEPHORA COLLECTION
3. Brand yang memiliki keuntungan paling besar adalah TOM FORD
4. Kategori produk yang paling laku secara berurutan berasal dari segmen (Parfum, serum wajah, moisturizer, value and gift sets, cologne)
5. Produk yang paling sering di review berasal dari segmen PARFUM
6. Brand parfum yang paling laku adalah TOM FORD
7. Untuk brand SEPHORA COLLECTION sendiri, yang paling laku berasal dari segmen Face Brushes

Presentation by IGD Raditya Wibhawa MN

LATIHAN 5

Beberapa insight yang telah ditemukan dari Visualisasi data adalah :

8. Beberapa segmen dari brand TOM FORD yang paling laku adalah secara berturut turut adalah (Perfume, Cologne, Body Mist & Hair Mist, Rollerballs & Travel Size, Lotions & Oil)
9. Variable numerik seperti price, number_of_reviews, dan rating memiliki jumlah yang tidak terdistribusi dengan baik

LATIHAN 5

Beberapa insight yang ditemukan dari Machine Learning Model:

1. Feature yang akan digunakan antara lain adalah : brand, category, size, rating, number_of_reviews, love, price, marketingFlags, dan how_to_use
2. Untuk labelsnya adalah exclusive
3. Pertimbangan tidak menggunakan feature lain karena ketidakseimbangan antar data didalam kolom tersebut
4. Tujuan dari penggunaan Machine Learning pada case ini adalah untuk mengklasifikasi product mana yang bersifat eksklusif

Presentation by IGD Raditya Wibhawa MN

LATIHAN 5

Beberapa insight yang ditemukan dari Machine Learning Model:

Berikut 6 Machine Learning Model yang digunakan beserta Accuracynya :

1. Logistic Regression, Accuracy : 75.3%
2. Decision Tree, Accuracy : 74.3%
3. Random Forest Classifier, Accuracy : 73.9%
4. SVM, Accuracy : 75.6%
5. K-Nearest Neighbor, Accuracy : 74.3%
6. Gradient Boosting, Accuracy : 74.9%

LATIHAN 5

Hasil : Paling optimal dengan menggunakan SVM Classifier dengan accuracy 75.6%

Hambatan : Data yang digunakan sangat tidak seimbang, bahkan ketika telah melakukan grouping untuk mengatasi Categorical Feature yang memiliki perbedaan hingga ribuan jenis

Improvisasi : Mengambil lebih banyak data lagi/Menggunakan metode sampling untuk memperoleh data secara seimbang

Presentation by IGD Raditya Wibhawa MN

Link GITHUB code, pdf, dan data yang digunakan untuk task ini :
https://github.com/radityawibhawa/Sephora_Collection_Analysis