

Intermediate Data Visualization

Advanced Data Visualization

Data Preprocessing for Machine Learning

Learning Progress Review Week 11

By:
MARVEL TEAM

Fikrie | Natalia | Satria

1. *Intermediate Data Visualization*

Kenapa *Data Visualization*?

Data Visualization adalah hal yang sangat penting bagi perkembangan suatu bisnis.

Pengambilan keputusan perusahaan haruslah di dasarkan pada data. Agar lebih mudah di pahami, digunakanlah data visualization tersebut.

Dengan menggunakan elemen visual tersebut, pembaca akan lebih mudah memahami tren, outliers dan pola dalam suatu data.

Apa itu *Data Visualization*?



Visualisasi data atau *data visualization* adalah tampilan berupa grafis atau visual dari informasi dan data. Dengan kata lain, data visualization mengubah kumpulan data menjadi lebih sederhana untuk di tampilkan.

Human Centered Reflection

Purpose

Purpose meliputi 4 hal utama, yaitu:

1. Di mulai dari manakah bila kita ingin memulai visualisasikan dataset yang kita miliki?
2. Masalah apa yang saat ini sedang di analisis untuk di temukan solusinya?
3. Tujuan dari mevisualisasikan dari dataset itu sendiri seperti apa?
4. Apakah secara spesifik bertujuan untuk membuat suatu point tertentu atau ingin membuat cerita dari hasil visualisasi tersebut?

Human Centered Reflection

Audience

- **Siapa target audience kita?** Yang dimana kita akan mePresentasikan hail analysis kita ke client tersebut.
- **Apa yang client butuhkan dari visualisai data?** Di tahap ini kita sebagai Data Scientist akan menganalisa dataset dan akan merubah nya ke visualisasi yang di inginkan client.

Human Centered Reflection

Data

- **Dataset apa saja yang dapat di dimanfaatkan?** Di tahap ini kita sebagai Data Scientist mengolah dan mencari data yang dapat di dimanfaatkan untuk menampilkan visualisasi sesuai keinginan client.
- **Apakah data yang kita gunakan bersifat periodik atau bersifat seperti dashboard atau hanya report?** Kita sebagai Data Scientist harus bisa membedakannya, karena bila tidak hasil visualisasinya tidak akan sesuai tujuan yang di inginkan client.

Human Centered Reflection

Context

- **Dimana kita akan menampilkan hasil visualisasi data?** Di tahap ini kita harus bisa menyesuaikan visualisasinya akan di tampilkan di software atau di website sebagai presentasi ke client atau stakeholder.

Tipe *Data Visualization*

Exploratory Data Analysis

Exploratory Data Analysis mengacu pada proses kritis dalam melakukan investigasi awal pada data untuk menentukan pola, untuk menemukan anomaly, untuk menguji hipotesis dan untuk memeriksa asumsi dengan bantuan statistic ringkasan dan representasi grafis.

Dengan melakukan **Exploratory Data Analysis**, kita dapat lebih memahami kondisi dataset yang kita miliki. Sehingga, kita dapat memulai pembentukan model Machine Learning dengan lebih baik kedepannya.

Tipe *Data Visualization*

Cara melakukan EDA

Dalam melakukan **Exploratory Data Analysis**, kita dapat memulainya dengan menjawab beberapa pertanyaan ini seperti berikut:

- Insight apa yang ingin kita dapatkan dari dataset?
- Bagaimana kondisi dan jenis data yang kita miliki?
- Apakah ada data yang hilang? Bagaimana kita mengatasi data yang hilang?
- Apakah ada data outlier yang di temukan?
- Model Machine Learning apa yang bisa digunakan untuk dataset yang kita miliki?

Type Data Visualization

Tools Exploratory

Libraries : Matplotlib, Seaborn, Ggplot2, Sweetviz, Autoviz.



Software : Tableau, PowerBI



Explanatory Data Analysis

Tujuan **Explanatory Data Analysis** adalah bagaimana kita sebagai Data Scientist mekomunikasikan hasil penemuan atau hasil visualisasi kita secara efisien ke client atau stakeholder untuk di presentasikan.

Cara melakukan Explanatory

Ada beberapa cara yang harus dilakukan seorang Explanatory Data Analysis:

- Membuat narasi atau cerita dari hasil visualisasi data.
- Mengumpulkan insight penting dari proses Explanatory Data Analysis yang relevan dengan narasi atau cerita yang di buat.
- Mengerti siapa audience atau client kita.
- Pemilihan model chart yang mudah di pahami oleh audience.
- Penggunaan warna, ukuran dan penjelasan agar dapat di mengerti sama audience.

Type Data Visualization

Tools Exploratory

Libraries : Dash/Plotly, Bokeh, Streamlit.



Software : Tableau, PowerBI, Google Data Studio, Holistic, Domo



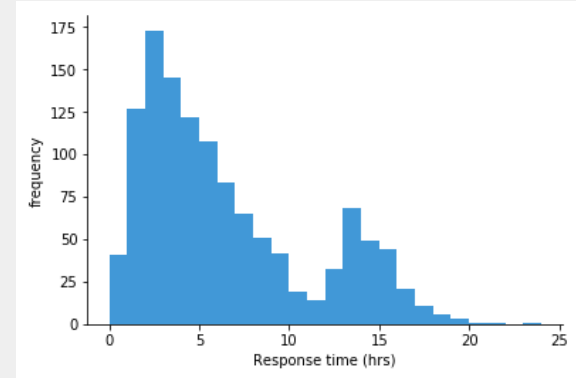
4 Pilar *Data Visualization*

Distribution

Digunakan untuk 1, 2 atau 3 variable yang ingin kita cari dari setiap **distibusinya**.

Distribusi berkaitan dengan kemungkinan terjadinya suatu **outcome** atau suatu nilai dalam suatu distribusi itu sendiri.

Chart yang biasa digunakan pada Distribusi adalah Histogram, density, boxplot dll.



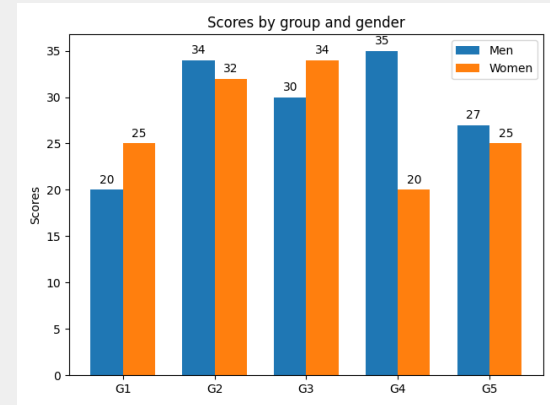
4 Pilar *Data Visualization*

Relationship

Digunakan untuk mencari 2 atau 3 variable yang berbeda.

Relationship biasa digunakan untuk **mencari korelasi** antara 2 variable yang berbeda.

Chart yang dapat digunakan adalah Bar plot, line chart dan bubble plot.



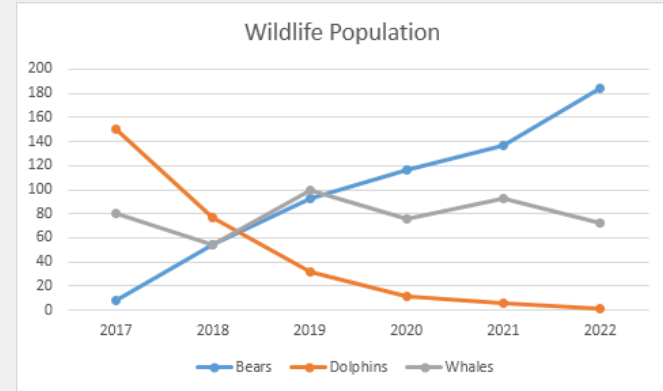
4 Pilar *Data Visualization*

Comparison

Digunakan untuk **membandingkan** antara beberapa kategori atau individual antar kolom.

Teknik **comparison** adalah Teknik yang paling sering digunakan pada visualisasi data yang dimana kita menggunakan 2 atau lebih variable untuk membandingkan hasil.

Chart yang sering digunakan pada comparison adalah Bar plot dan Line chart.



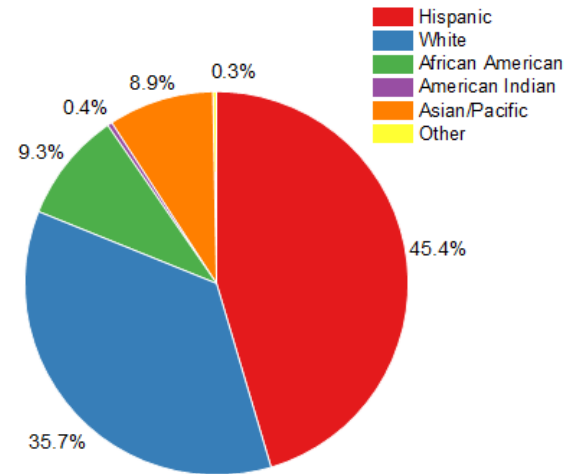
4 Pilar *Data Visualization*

Composition

Digunakan untuk melihat **komposisi** dari beberapa variable yang ada.

Tujuan dari penggunaan Teknik **composition** adalah untuk menampilkan komposisi dari beberapa variable yang telah di pilih.

Pie chart dan stacked plot adalah chart yang paling sering digunakan untuk penggunaan komposisi.



Psychology of Data Visualization

Di dalam data Visualisasi secara **psikologi** elemen-elemen yang ada di dalam satu grup dia akan cenderung bergantung pada karakteristik visualnya.

Psikologi visualisasi data ada 3 diantaranya adalah:

- ❖ *Similarity*
- ❖ *Proximity*
- ❖ *enclosure*

Psychology of Data Visualization

Similarity

Kita harus perhatikan **similarity** di dalam dataset yang kita miliki yang ingin kita visualisasikan.

Contoh :

Jika kita ingin visualisasikan suatu data menggunakan scatter plot, kita perlu menggambarkan suatu area tertentu yang terkumpul dalam beberapa point yang memiliki warna yang sama.

Similarity adalah menunjukkan suatu nilai psikologis dari data visualisasi kita

Psychology of Data Visualization

Proximity

Proximity lebih kuat di bandingkan dengan similarity, yang dimana mata manusia cenderung memperhatikan suatu hal yang berdasarkan kedekatan yang di lihat dengan yang lainnya.

Enclosure

Enclosure adalah bila kita ingin memanfaatkan psikologis similarity dari mata manusia yang ingin menunjukkan bahwa ada suatu grup yang ingin di highlight. Hal ini paling di perlukan untuk menampilkan visualisasi data agar dapat mudah di pahami oleh audience.

2. *Advanced Data Visualization*

Visualisasi Scatter Plot Perbandingan Kuantitatif

Data austin_weather.csv

```
df = pd.read_csv('austin_weather.csv')
df.head()
```

	Date	TempHighF	TempAvgF	TempLowF	DewPointHighF	DewPointAvgF	DewPointLowF	HumidityHighPercent	HumidityAvgPercent	HumidityLowPercent	SeaLevelPressureHighInches	SeaLevelPres:
0	2013-12-21	74	60	45	67	49	43	93	75	57	29.86	
1	2013-12-22	56	48	39	43	36	28	93	68	43	30.41	
2	2013-12-23	58	45	32	31	27	23	76	52	27	30.56	
3	2013-12-24	61	46	31	36	28	21	89	56	22	30.56	
4	2013-12-25	58	50	41	44	40	36	86	71	56	30.41	

Visualisasi Scatter Plot Perbandingan Kuantitatif

Data austin_weather.csv

```
[ ] df2 = df[['DewPointAvgF', 'HumidityAvgPercent', 'WindAvgMPH', 'TempAvgF']].replace(['-'], np.nan)
df2.fillna(method='ffill')
```

replace data bernilai '-' dengan nilai NaN & mengisi nilai nan dengan nilai sebelumnya di row tersebut.

```
[ ] df3 = df2.astype(float)
df3.dtypes
```

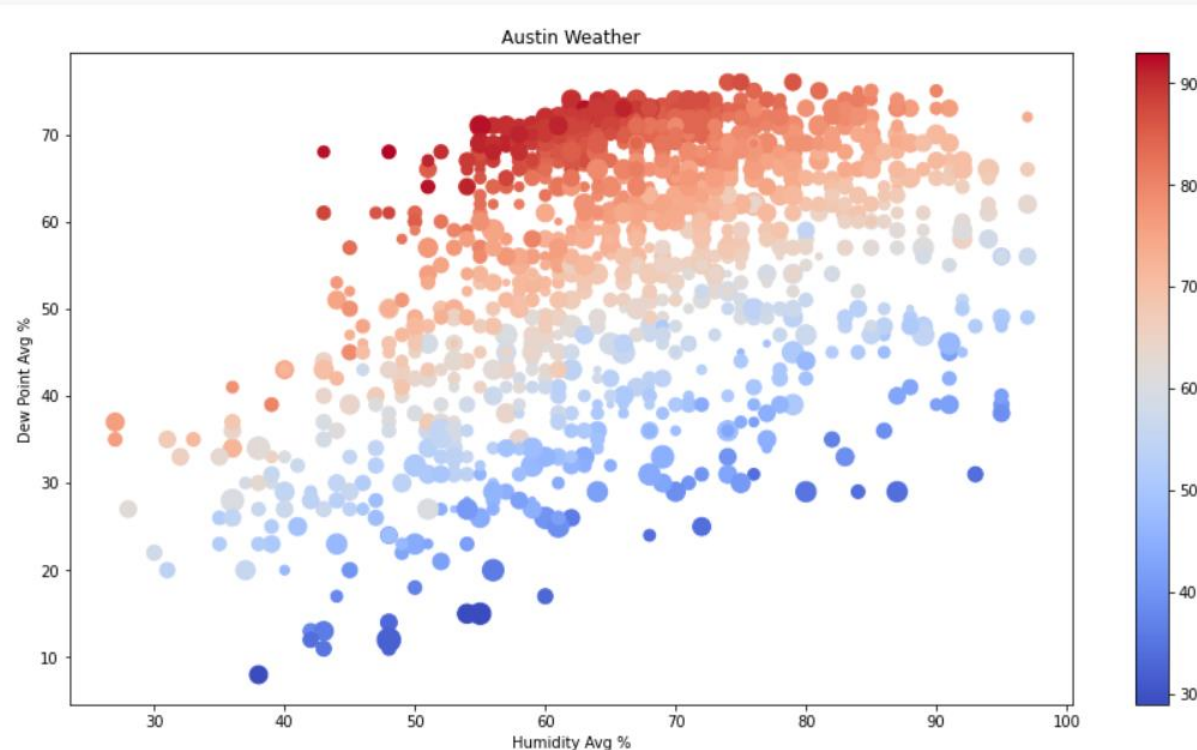
Mengubah tipe data menjadi tipe data float.

```
[ ] fig, ax = plt.subplots(figsize = (15, 8))
a = df3['HumidityAvgPercent']
b = df3['DewPointAvgF']
color = df3['TempAvgF']
size = 20*df3['WindAvgMPH']
d = ax.scatter(a, b, c = color, cmap = 'coolwarm', s = size)

ax.set_xlabel('Humidity Avg %')
ax.set_ylabel('Dew Point Avg %')
ax.set_title('Austin Weather')
fig.colorbar(d)
plt.show()
```

Visualisasi Scatter Plot Perbandingan Kuantitatif

Hasil



Preparing Data

Data vgsales.csv

```
[ ] df = pd.read_csv('vgsales.csv')  
df.head()
```

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37

Preparing Data

Data vgsales.csv

Membuat grup berdasarkan Genre & ambil rata2 penjualan untuk setiap Region

```
[ ] mean_sales = df.groupby('Genre')[['NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales']].mean()  
    display(mean_sales)
```

Hasil →

	NA_Sales	EU_Sales	JP_Sales	Other_Sales
Genre				
Action	0.264726	0.158323	0.048236	0.056508
Adventure	0.082271	0.049868	0.040490	0.013072
Fighting	0.263667	0.119481	0.103007	0.043255
Misc	0.235906	0.124198	0.061967	0.043312
Platform	0.504571	0.227573	0.147596	0.058228
Puzzle	0.212680	0.087251	0.098471	0.021564
Racing	0.287766	0.190865	0.045388	0.061865
Role-Playing	0.219946	0.126384	0.236767	0.040060
Shooter	0.444733	0.239137	0.029221	0.078389
Simulation	0.211430	0.130773	0.073472	0.036355
Sports	0.291283	0.160635	0.057702	0.057532
Strategy	0.100881	0.066579	0.072628	0.016681

Perbandingan Kuantitatif Barplot - Grouping Visualisasi Dengan Barplot

Data vgsales.csv

```
[ ] plt.style.use('ggplot')

x = np.arange(0, 24, 6)

y1 = mean_sales.iloc[0]
y2 = mean_sales.iloc[1]
y3 = mean_sales.iloc[2]
y4 = mean_sales.iloc[3]
y5 = mean_sales.iloc[4]
y6 = mean_sales.iloc[5]
y7 = mean_sales.iloc[6]
y8 = mean_sales.iloc[7]
y9 = mean_sales.iloc[8]
y10 = mean_sales.iloc[9]
y11 = mean_sales.iloc[10]
y12 = mean_sales.iloc[11]

a = mean_sales[['NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales']]
```



```
fig, ax = plt.subplots(figsize = (16,8))

ax.bar(x, y1, width = 0.25, label = 'Action')
ax.bar(x+0.25, y2, width = 0.25, label = 'Adventure')
ax.bar(x+0.5, y3, width = 0.25, label = 'Fighting')
ax.bar(x+0.75, y4, width = 0.25, label = 'Misc')
ax.bar(x+1, y5, width = 0.25, label = 'Platform')
ax.bar(x+1.25, y6, width = 0.25, label = 'Puzzle')
ax.bar(x+1.5, y7, width = 0.25, label = 'Racing')
ax.bar(x+1.75, y8, width = 0.25, label = 'Role-Playing')
ax.bar(x+2, y9, width = 0.25, label = 'Shooter')
ax.bar(x+2.25, y10, width = 0.25, label = 'Simulation')
ax.bar(x+2.5, y11, width = 0.25, label = 'Sport')
ax.bar(x+2.75, y12, width = 0.25, label = 'Strategy')

ax.set_xlabel('Region Sales')
ax.set_ylabel('Mean Sales')
ax.set_title('Mean Sales Video Games by Genre')

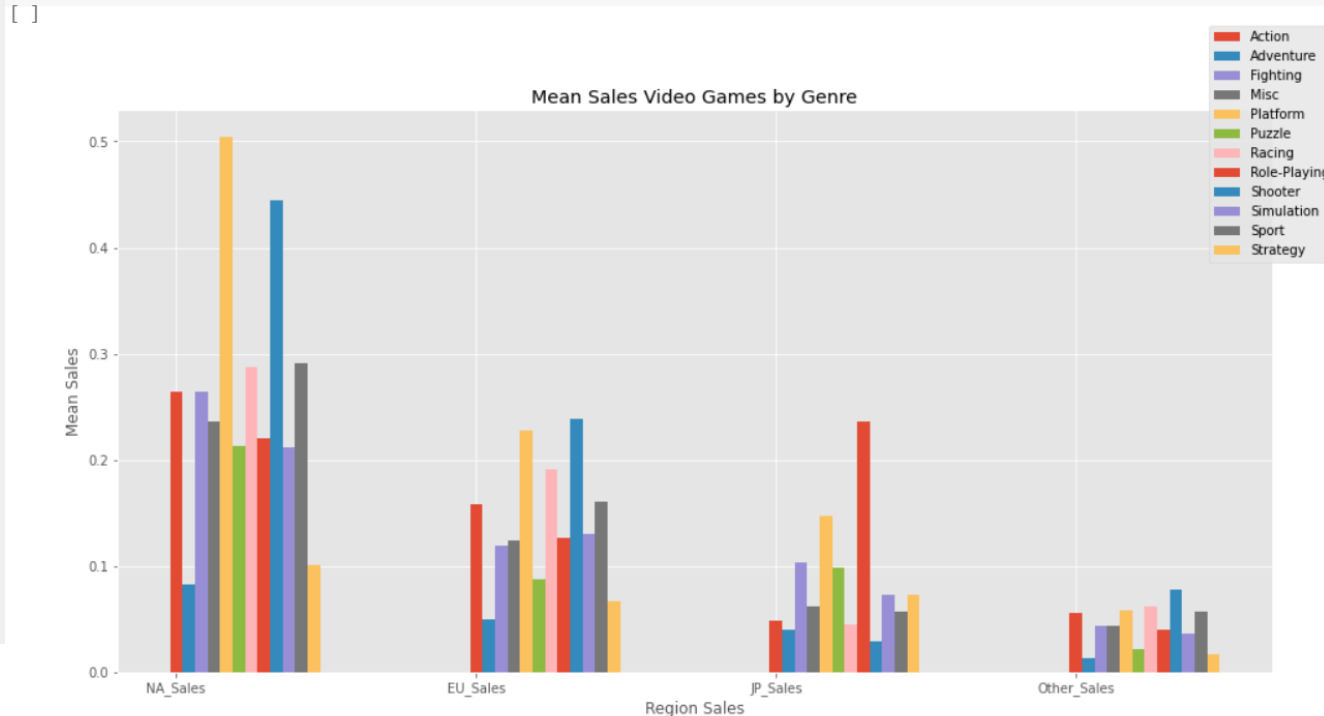
ax.set_xticks(x)
ax.set_xticklabels(a)

ax.legend(loc = 'center', bbox_to_anchor = (1, 0.94))

plt.show()
```

Perbandingan Kuantitatif Barplot - Grouping Visualisasi Dengan Barplot

Hasil



Perbandingan Kuantitatif Barplot - Stack Barplot

Data vgsales.csv

```
[ ] plt.style.use('ggplot')

x = np.arange(0, 24, 6)

fig, ax = plt.subplots(figsize = (16,8))

ax.bar(x, y1, width = 2.5, label = 'Action')
ax.bar(x, y2, width = 2.5, bottom = y1, label = 'Adventure')
ax.bar(x, y3, width = 2.5, bottom = y1+y2, label = 'Fighting')
ax.bar(x, y4, width = 2.5, bottom = y1+y2+y3, label = 'Misc')
ax.bar(x, y5, width = 2.5, bottom = y1+y2+y3+y4, label = 'Platform')
ax.bar(x, y6, width = 2.5, bottom = y1+y2+y3+y4+y5, label = 'Puzzle')
ax.bar(x, y7, width = 2.5, bottom = y1+y2+y3+y4+y5+y6, label = 'Racing')
ax.bar(x, y8, width = 2.5, bottom = y1+y2+y3+y4+y5+y6+y7, label = 'Role-Playing')
ax.bar(x, y9, width = 2.5, bottom = y1+y2+y3+y4+y5+y6+y7+y8, label = 'Shooter')
ax.bar(x, y10, width = 2.5, bottom = y1+y2+y3+y4+y5+y6+y7+y8+y9, label = 'Simulation')
ax.bar(x, y11, width = 2.5, bottom = y1+y2+y3+y4+y5+y6+y7+y8+y9+y10, label = 'Sport')
ax.bar(x, y12, width = 2.5, bottom = y1+y2+y3+y4+y5+y6+y7+y8+y9+y10+y11, label = 'Strategy')

ax.set_xlabel('Region Sales')
ax.set_ylabel('Mean Sales')
ax.set_title('Mean Sales Video Games by Genre')

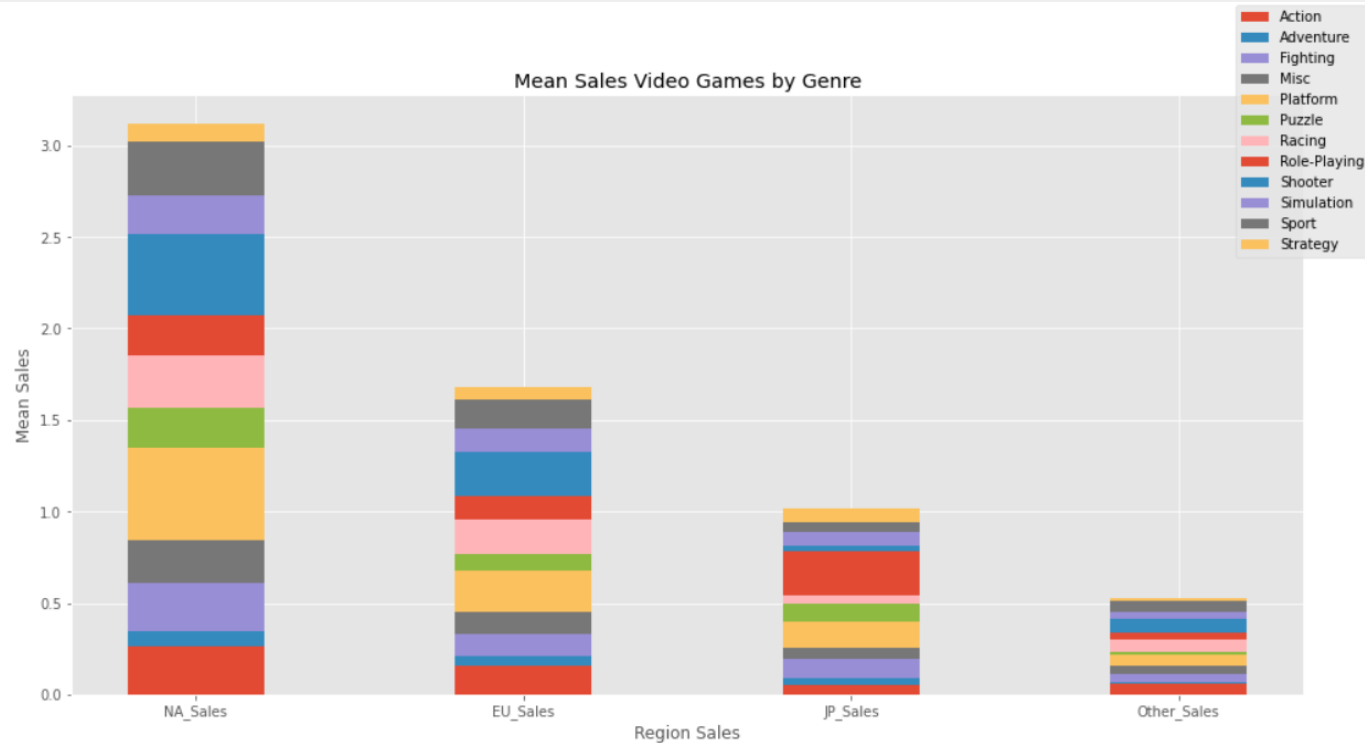
ax.set_xticks(x)
ax.set_xticklabels(a)

ax.legend(loc = 'center', bbox_to_anchor = (1, 0.94))

plt.show()
```

Perbandingan Kuantitatif Barplot - Stack Barplot

Hasil



Scatter Plot

Data <https://raw.githubusercontent.com/vaksakalli/datasets/master/diamonds.csv>

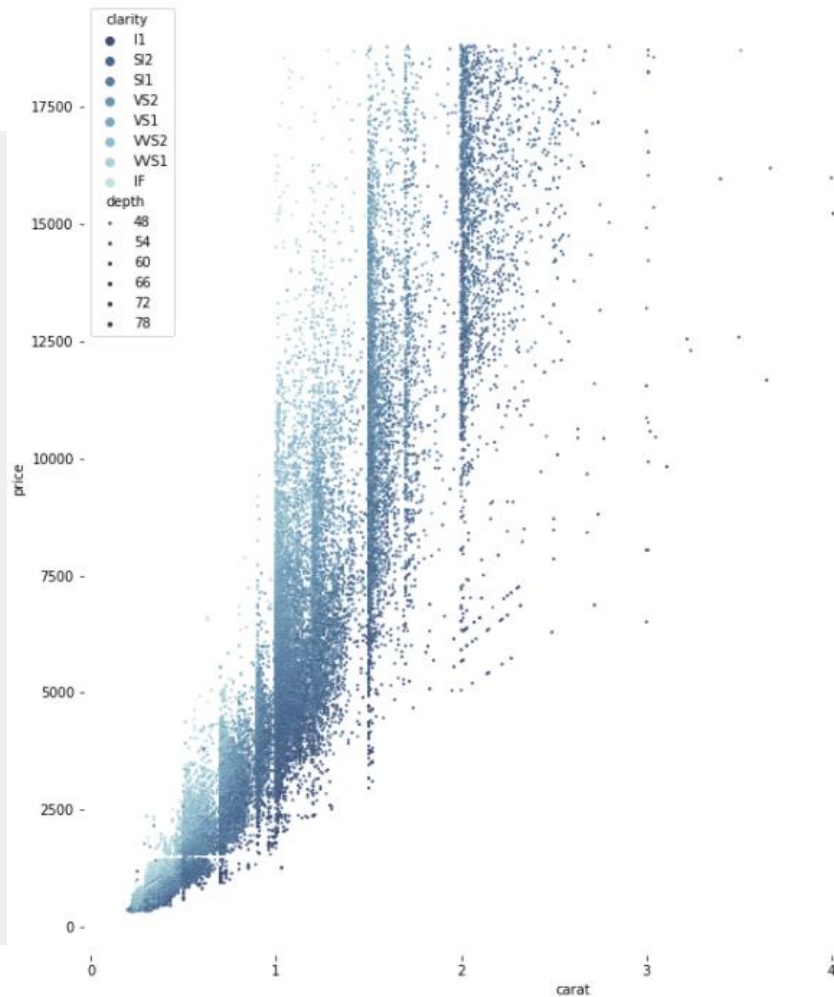
```
[ ] fig, ax = plt.subplots(figsize =(13, 13))
    palette='ch:r=-.2,d=.3_r'

    sns.despine(fig, left=True, bottom=True)
    ranking = ['I1', 'SI2', 'SI1', 'VS2', 'VS1', 'VVS2', 'VVS1', 'IF']
    data = sns.load_dataset('diamonds')
    sns.scatterplot(x = 'carat', y = 'price', size = 'depth',
                    palette = palette, hue = 'clarity',
                    hue_order = ranking, sizes=(1,8),
                    linewidth = 0, ax=ax, data = data)

    plt.show()
```

Scatter Plot

Hasil



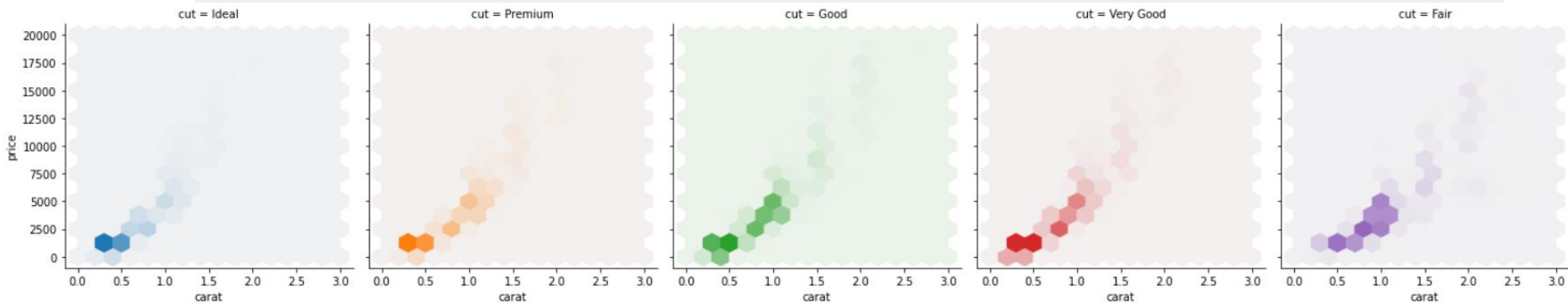
Facet Grid

Data <https://raw.githubusercontent.com/vaksakalli/datasets/master/diamonds.csv>

```
[ ] def hexbin(x, y, color, **kwargs):  
    cmap = sns.light_palette(color, as_cmap=True)  
    plt.hexbin(x, y, gridsize=15, cmap=cmap, **kwargs)  
  
g = sns.FacetGrid(df, col = 'cut', hue = 'cut', height = 4)  
g.map(hexbin, 'carat', 'price', extent = [0, 3, 0, 20000])
```

Facet Grid

Hasil



Pair Plot

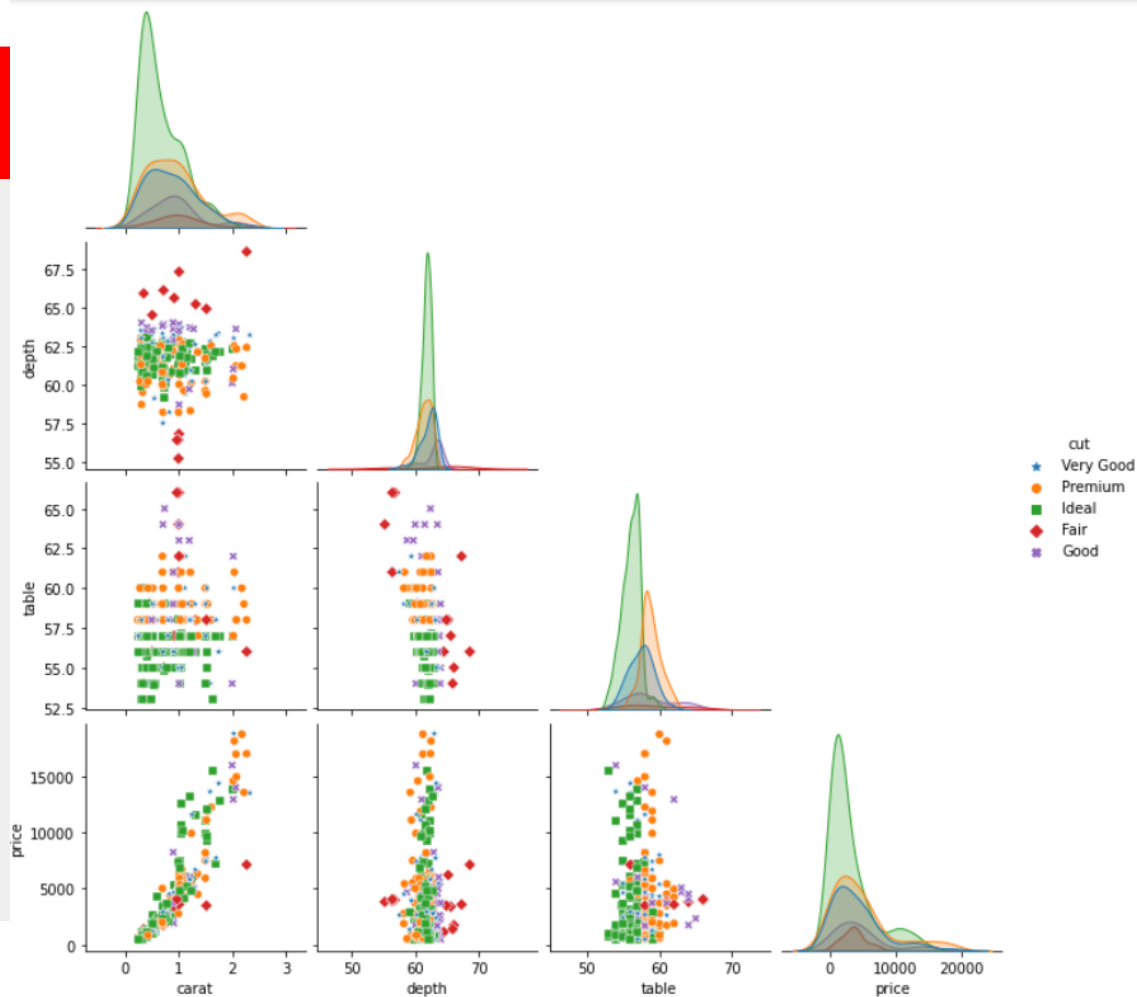
Data <https://raw.githubusercontent.com/vaksakalli/datasets/master/diamonds.csv>

```
[ ] df2 = df.sample(n=300, random_state=123)

data = sns.load_dataset('diamonds')
sns.pairplot(data=df2, vars = ['carat', 'depth', 'table', 'price'], hue = 'cut', markers=['*', 'o', 's', 'D', 'X'], corner = True)
```

Pair Plot

Hasil



Joint Plot

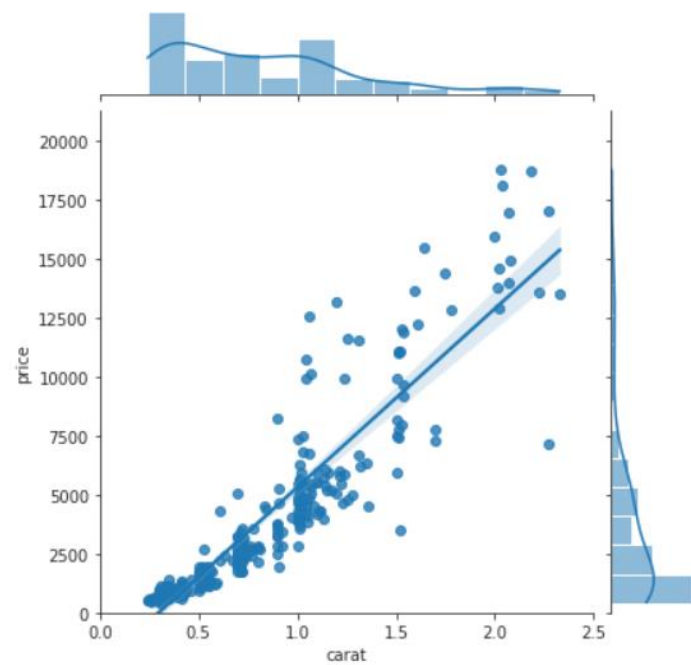
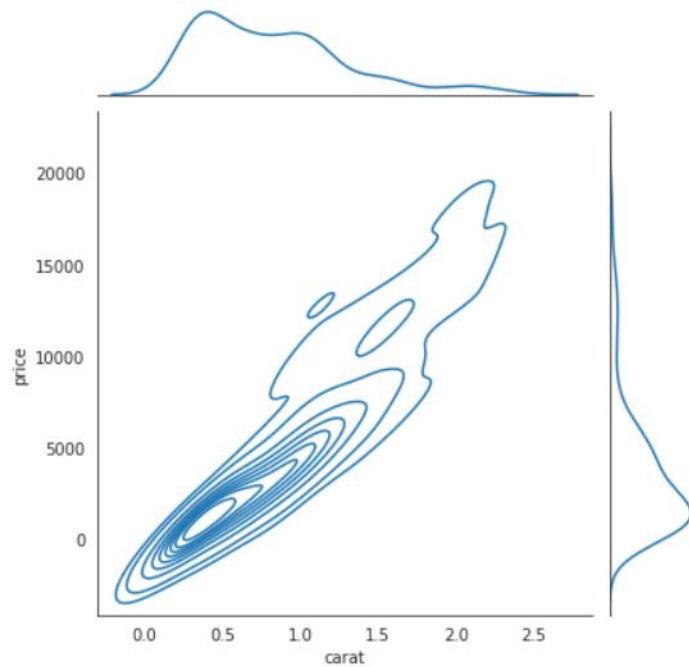
Data <https://raw.githubusercontent.com/vaksakalli/datasets/master/diamonds.csv>

```
[ ] from scipy.stats import pearsonr

with sns.axes_style('white'):
    g = sns.jointplot('carat', 'price', data = df, kind = 'kde')
with sns.axes_style('ticks'):
    g2 = sns.jointplot('carat', 'price', data = df, kind = 'reg')
    g2.ax_marg_x.set_xlim(0.0, 2.5)
    g2.ax_marg_y.set_ylim(0, 21250)
    plt.show()
```

Joint Plot

Hasil



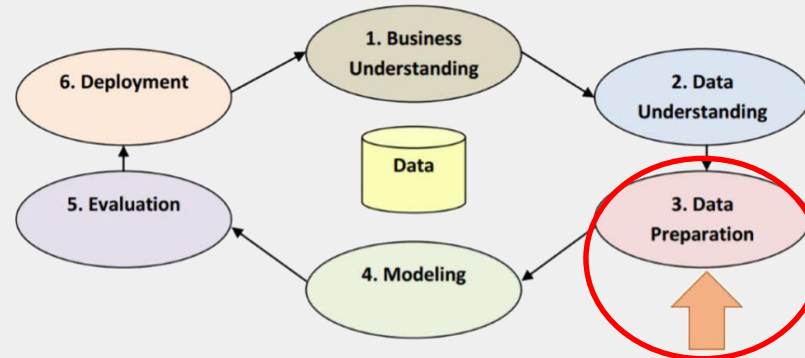
3. *Data Preprocessing for Machine Learning (with Python)*

Apa itu *Data Preprocessing*?



- **Data preprocessing** adalah proses manipulasi dataset sebelum diinput kedalam model.
- Proses ini bisa juga disebut dengan langkah awal untuk mengambil semua informasi yang tersedia dengan cara membersihkan, memfilter, dan menggabungkan data-data tersebut.

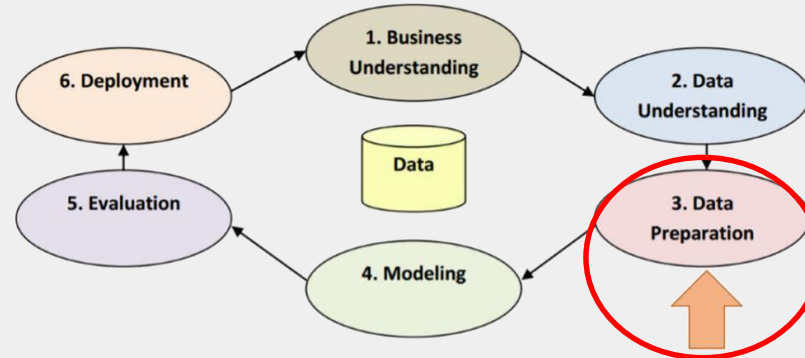
Tujuan *Data Preprocessing*



Tujuan *Data preprocessing* adalah :

- Data bisa digunakan untuk melatih model
- Meningkatkan efisiensi model
- Meningkatkan performa model

Tujuan *Data Preprocessing*

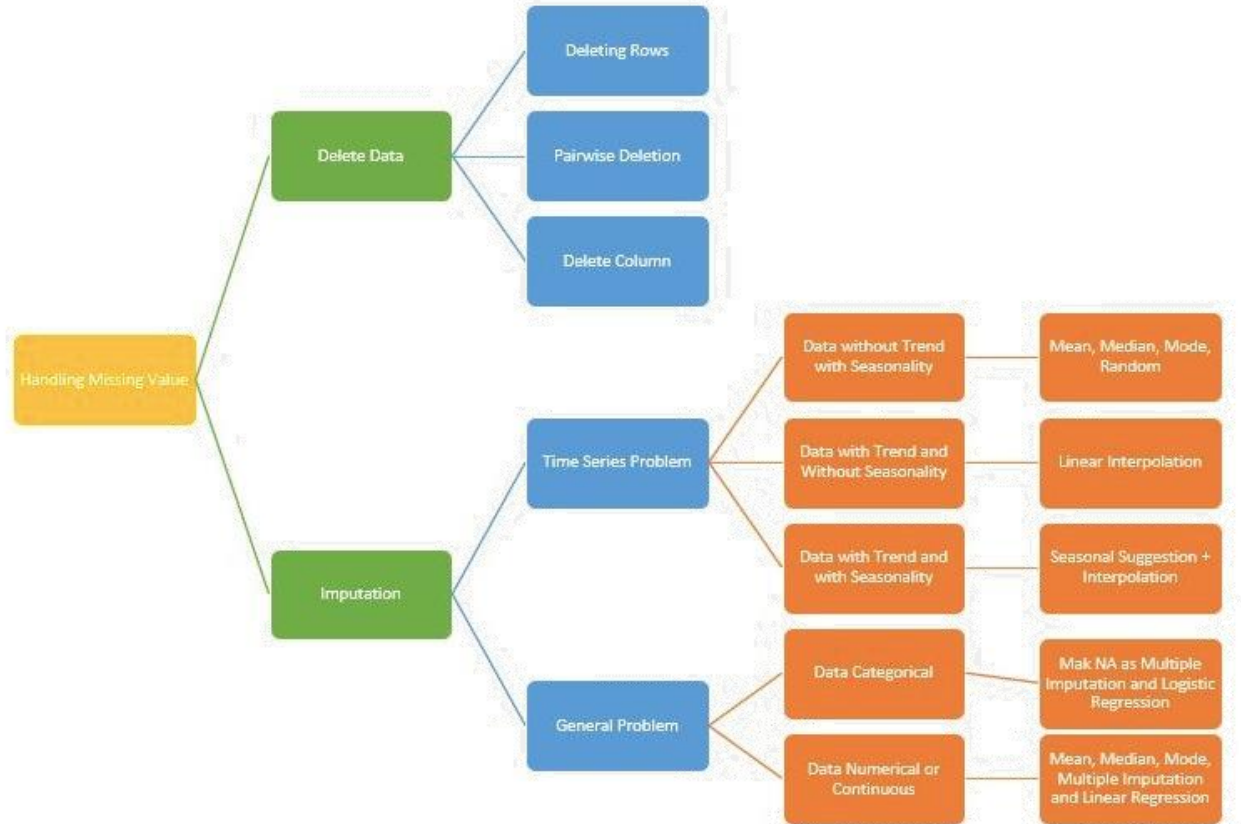


Tahapan *Data preprocessing* adalah :

1. *Data Cleaning* (Menangani *missing values*)
2. Menangani fitur kategorikal
3. Transformasi data (*Scaling* / Normalisasi)
4. Bagi menjadi *train* dan *test data*.

Missing Values Handling

Bagan cara penanganan missing values:



Missing Values Handling

Cara umum penanganan missing values	Kelebihan	Kekurangan
Hapus baris	<ul style="list-style-type: none"> Tanpa manipulasi data asli 	<ul style="list-style-type: none"> Kehilangan informasi Jumlah data yang dilatih untuk model akan turun sehingga mengurangi kinerja model
Mengisi menggunakan mean atau median untuk fitur numerikal	<ul style="list-style-type: none"> Mencegah kehilangan informasi 	<ul style="list-style-type: none"> Dapat menimbulkan kebocoran data Bias terhadap pencilan (gunakan median jika ada pencilan)
Mengisi menggunakan mode/kategori baru untuk fitur kategorikal	<ul style="list-style-type: none"> Mencegah kehilangan informasi 	<ul style="list-style-type: none"> Menambah perhitungan saat encoding

Label Encoding

- **Label encoding** mengubah setiap nilai dalam kolom menjadi angka yang berurutan.
- Baik digunakan pada fitur kategorikal nominal atau biner.
- Dalam Python, dapat menggunakan library Sci-kit learn dengan metode LabelEncoder.

State (Nominal Scale)		State (Label Encoding)
Maharashtra		3
Tamil Nadu		4
Delhi		0
Karnataka		2
Gujarat		1
Uttar Pradesh		5

One Hot Encoding

- **One hot encoding** adalah teknik yang merubah setiap nilai di dalam kolom menjadi kolom baru dan mengisinya dengan nilai biner yaitu 0 atau 1.
- Dalam Python, terdapat 2 cara yaitu :
 1. Menggunakan Scikit-learn dengan metode OneHotEncoder.
 2. Pandas metode get_dummies.

One Hot Encoding

Contoh :

Dataset

	fruit	size
0	apple	large
1	apple	medium
2	banana	small
3	orange	large
4	banana	medium
5	apple	small

Menggunakan OneHotEncoder dari Scikit-learn

```
encoder = OneHotEncoder()
df_fruit_encoded = pd.DataFrame(encoder.fit_transform(df[['fruit']]).todense(),
                                columns=encoder.get_feature_names())
df_fruit_encoded
```

	x0_apple	x0_banana	x0_orange
0	1.0	0.0	0.0
1	1.0	0.0	0.0
2	0.0	1.0	0.0
3	0.0	0.0	1.0
4	0.0	1.0	0.0
5	1.0	0.0	0.0

One Hot Encoding

Contoh :

Dataset

	fruit	size
0	apple	large
1	apple	medium
2	banana	small
3	orange	large
4	banana	medium
5	apple	small

Menggunakan get_dummies dari Pandas

```
pd.get_dummies(df['size'])
```

	large	medium	small
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	1	0
5	0	0	1

Feature Scaling (Standardization, Normalization)

- **Feature Scaling** adalah suatu cara untuk membuat data numerikal pada dataset memiliki rentang nilai (skala) yang sama sehingga tidak ada satu variabel data yang mendominasi variabel data lainnya.
- Dapat dilakukan 2 cara :
 - a. Normalization
 - b. Standardization

Feature Scaling (Standardization, Normalization)

Normalisasi (*normalization*) adalah teknik penskalaan dimana nilai – nilai digeser dan diubah skalanya menjadi kisaran antara 0 dan 1. Ini juga dikenal sebagai penskalaan Min-Max.

Normalisasi dalam Python dan library Scikit-learn :

```
from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
train_X = scaler.fit_transform(train_X)  
test_X = scaler.transform(test_X)
```

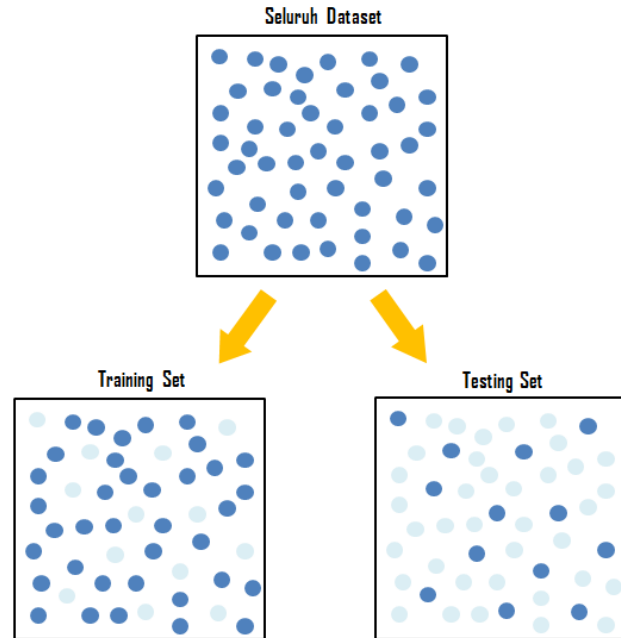
Feature Scaling (Standardization, Normalization)

Standardisasi (*standardization*) adalah teknik penskalaan fitur sehingga memiliki mean (rata-rata) = 0 dan standar deviasi = 1. Secara teknis, standardisasi memusatkan dan menormalkan data dengan mengurangi mean dan membaginya dengan standar deviasi.

Standardisasi dalam Python dan library Scikit-learn :

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
train_X = scaler.fit_transform(train_X)
test_X = scaler.transform(test_X)
```

Train-Test Split



- **Train-Test Split** adalah salah satu metode yang dapat digunakan untuk mengevaluasi performa model *machine learning*.
- Metode ini membagi dataset menjadi dua bagian, yaitu :
 - a. Train data**, digunakan untuk fit model *machine learning*
 - b. Test data**, digunakan untuk mengevaluasi hasil *fit* model tersebut.

Special Thanks to :



Slide template by SlideCarnival