

# **== UJIAN AKHIR SEMESTER ==**

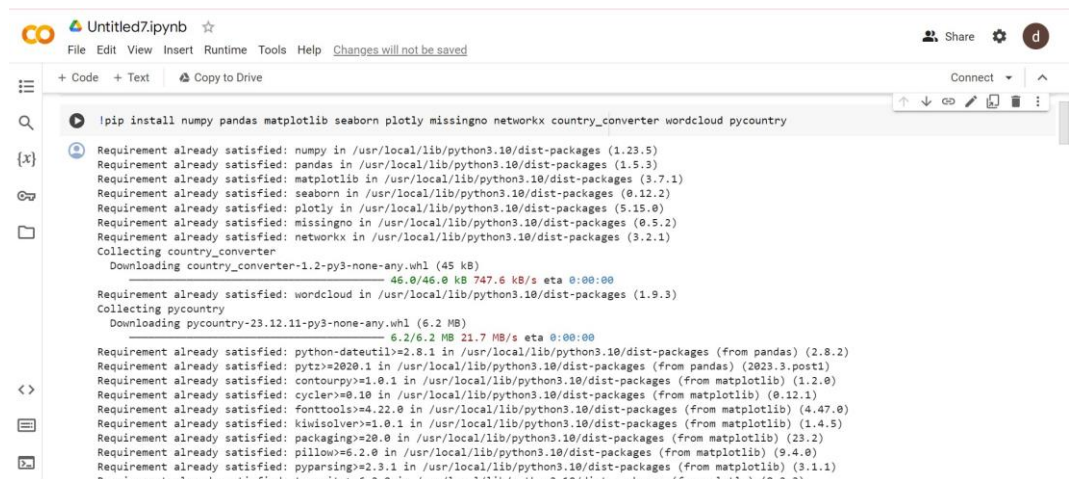
## **= LAPORAN ANALISIS DATA =**

**062140832952/FADHILAH AMALIA PUTRI/5MIM**

**Nama Dataset : EmployeeSampleData**

### ***Informasi proses Processing Data :***

- **Sample Langkah Proseccing Data**



```
Untitled7.ipynb
File Edit View Insert Runtime Tools Help Changes will not be saved
+ Code + Text Copy to Drive
Connect
!pip install numpy pandas matplotlib seaborn plotly missingno networkx country_converter wordcloud pycountry

Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (1.23.5)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (1.5.3)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-packages (0.12.2)
Requirement already satisfied: plotly in /usr/local/lib/python3.10/dist-packages (5.15.0)
Requirement already satisfied: missingno in /usr/local/lib/python3.10/dist-packages (0.5.2)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (3.2.1)
Collecting country_converter
  Downloading country_converter-1.2-py3-none-any.whl (45 kB)
    46.0/46.0 KB 747.6 KB/s eta 0:00:00
Requirement already satisfied: wordcloud in /usr/local/lib/python3.10/dist-packages (1.9.3)
Collecting pycountry
  Downloading pycountry-23.12.11-py3-none-any.whl (6.2 MB)
    6.2/6.2 MB 21.7 MB/s eta 0:00:00
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2023.3.post1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.47.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (23.2)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.1.1)
```

Picture diatas merupakan proses awal dalam pengaktifan pandas supaya dapat connect kedata yang akan dimasukkan, setelah selesai kode itu maka buat lagi kode baru.

```
+ Code + Text Copy to Drive

[ ] # default
import numpy as np
import pandas as pd

# visualization
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import missingno as msno
import networkx as nx
import plotly.graph_objects as go
import country_converter as coco
from wordcloud import WordCloud
import pycountry

[ ] !pip install pycountry

Requirement already satisfied: pycountry in /usr/local/lib/python3.10/dist-packages (23.12.11)

[ ] df = pd.read_csv('data.csv', encoding='windows-1252')
df.shape
```

Kode visualisasi untuk mengimport data yang tersimpan pada pandas, dan setelah itu kita memasukkan data sample yaitu “*EmployeeSampleData*” dengan kode “pd.read\_csv” yang digunakan agar data dapat ditampilkan.

```
+ Code + Text Copy to Drive

[ ] df.columns

Index(['EEID', 'Full Name', 'Job Title', 'Department', 'Business Unit',
      'Gender', 'Ethnicity', 'Age', 'Hire Date', 'Annual Salary', 'Bonus %',
      'Country', 'City', 'Exit Date'],
      dtype='object')
```

Setelah itu buat kode baru lagi yaitu kode untuk menampilkan columns, ketik beberapa columns yang ada pada data yang kita pakai tadi.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   EEID                   1000 non-null  object
1   Full Name              1000 non-null  object
2   Job Title              1000 non-null  object
3   Department             1000 non-null  object
4   Business Unit          1000 non-null  object
5   Gender                 1000 non-null  object
6   Ethnicity              1000 non-null  object
7   Age                   1000 non-null  int64
8   Hire Date              1000 non-null  object
9   Annual Salary          1000 non-null  object
10  Bonus %                1000 non-null  object
11  Country                1000 non-null  object
12  City                   1000 non-null  object
13  Exit Date              85 non-null    object
dtypes: int64(1), object(13)
memory usage: 109.5+ KB
```

Dan tampilannya akan seperti itu.

## Setelah itu kita membuat kode Missing Value

```
[ ] data_na = (df.isnull().sum() / len(df)) * 100
data_na = data_na.drop(data_na[data_na == 0].index).sort_values(ascending=False)[:30]
missing_data = pd.DataFrame({'Missing Ratio' :data_na})
missing_data.head(20)
```

### Missing Ratio

Exit Date	91.5
-----------	------

Missing Value merupakan kode yang digunakan untuk mengidentifikasi untuk data yang hilang.

## Duplicate Value

```
[ ] duplicate_rows_data = df[df.duplicated()]
print("number of duplicate rows: ", duplicate_rows_data.shape)
```

number of duplicate rows: (0, 14)

df.head()

	EEID	Full Name	Job Title	Department	Business Unit	Gender	Ethnicity	Age	Hire Date	Annual Salary	Bonus %	Country	City	Exit Date
0	E02387	Emily Davis	Sr. Manger	IT	Research & Development	Female	Black	55	4/8/2016	\$141,604	15%	United States	Seattle	10/16/2021
1	E04105	Theodore Dinh	Technical Architect	IT	Manufacturing	Male	Asian	59	11/29/1997	\$99,975	0%	China	Chongqing	NaN
2	E02572	Luna Sanders	Director	Finance	Speciality Products	Female	Caucasian	50	10/26/2006	\$163,099	20%	United States	Chicago	NaN
3	E02832	Penelope Jordan	Computer Systems Manager	IT	Manufacturing	Female	Caucasian	26	9/27/2019	\$84,913	7%	United States	Chicago	NaN
4	E01639	Austin Vo	Sr. Analyst	Finance	Manufacturing	Male	Asian	55	11/20/1995	\$95,409	0%	United States	Phoenix	NaN

df.tail()

	EEID	Full Name	Job Title	Department	Business Unit	Gender	Ethnicity	Age	Hire Date	Annual Salary	Bonus %	Country	City	Exit Date
995	E03094	Wesley Young	Sr. Analyst	Marketing	Speciality Products	Male	Caucasian	33	9/18/2016	\$98,427	0%	United States	Columbus	NaN
996	E01909	Lillian Khan	Analyst	Finance	Speciality Products	Female	Asian	44	5/31/2010	\$47,387	0%	China	Chengdu	1/8/2018
997	E04398	Oliver Yang	Director	Marketing	Speciality Products	Male	Asian	31	6/10/2019	\$176,710	15%	United States	Miami	NaN
998	E02521	Lily Nguyen	Sr. Analyst	Finance	Speciality Products	Female	Asian	33	1/28/2012	\$95,960	0%	China	Chengdu	NaN
999	E03545	Sofia Cheng	Vice President	Accounting	Corporate	Female	Asian	63	7/26/2020	\$216,195	31%	United States	Miami	NaN

## Dropna Value Data

```
[ ] df.dropna(inplace=True)

[ ] df.drop_duplicates(inplace=True)

[ ] print(df.info())

<class 'pandas.core.frame.DataFrame'>
Int64Index: 85 entries, 0 to 996
Data columns (total 14 columns):
#   Column              Non-Null Count  Dtype
---  -
0   EEID                 85 non-null    object
1   Full Name            85 non-null    object
2   Job Title            85 non-null    object
3   Department           85 non-null    object
4   Business Unit       85 non-null    object
5   Gender               85 non-null    object
6   Ethnicity            85 non-null    object
7   Age                 85 non-null    int64
8   Hire Date           85 non-null    object
9   Annual Salary       85 non-null    object
10  Bonus %             85 non-null    object
11  Country              85 non-null    object
12  City                 85 non-null    object
13  Exit Date            85 non-null    object
dtypes: int64(1), object(13)
memory usage: 10.0+ KB
None
```

## Gender Value Data

### ✦ Mencari data gender karyawan

```
[ ] gender_distribution = df['Gender'].value_counts()
    print(gender_distribution)
```

```
Male      46
Female    39
Name: Gender, dtype: int64
```

### - Kesimpulan visualisasi

Dapat disimpulkan dari visualisasi data yang dilakukan maka didapatkan deskripsi informasi detail tentang jumlah sample data yang terjadi terkait dengan klasifikasinya terbagi dari bagian, divisi serta scenario yang dipakai.