

Project name: Submission Technical Exercise Result

Project Summary

- Pada tahap awal Gathering Data, sangat penting untuk memastikan bahwa format nama file dataset yang sesuai dengan yang diperlukan agar data dapat terbaca dengan baik. Dalam masalah ini, dataset yang benar adalah 'sleep_health.csv'. Namun terdapat kesalahan penulisan pada format file tersebut yang menyebabkan dataset tidak terbaca saat menjalankan code. Untuk mengantisipasi masalah ini,, pastikan untuk memeriksa dan memperbaiki penulisan nama file agar sesuai dengan format yang ditentukan, agar proses pembacaan data dapat berjalan tanpa kendala.
- kesalahan karena jumlah subplot yang ingin dibuat melebihi kapasitas grid yang telah ditentukan. Kode menggunakan `plt.subplot(2, 4, i + 1)`, yang membuat grid dengan 2 baris dan 4 kolom, sehingga hanya dapat menampung maksimal 8 subplot. Saat iterasi mencapai kolom ke-9, program mencoba membuat subplot ke-9, yang menghasilkan `ValueError`. Untuk mengatasi masalah ini, perlu dilakukan perubahan pada argumen di fungsi `plt.subplot()`. Mengubah argumen dari `plt.subplot(2, 4, i + 1)` menjadi `plt.subplot(3, 3, i + 1)` mengubah grid menjadi 3 baris dan 3 kolom, yang dapat menampung hingga 9 subplot.
- Menggunakan `df_clean.drop(columns=['person_id'], inplace=True)` hasilkan kesalahan `KeyError` karena kolom `person_id` tidak ada dalam Data Frame `df_clean`. Hapus atau komentari kode tersebut, karena kolom `person_id` tidak perlu dihapus.
- Kode `df_clean = pd.concat([df_clean, df_clean['blood_pressure'].str.split('/', expand=True)], axis=1).drop('blood_pressure', axis=1)` menghasilkan `KeyError` karena kolom `blood pressure` tidak ada. Periksa keberadaan kolom dengan `if 'blood_pressure' in df_clean.columns:`. Jika ada, pisahkan kolom; jika tidak, tampilkan pesan bahwa kolom telah diproses.
- Kode `plt.subplot(1, 3, i + 1)` menghasilkan kesalahan karena indeks subplot (`i + 1`) melebihi jarak yang valid untuk grid 1x3. Kesalahan ini terjadi saat `i + 1` menjadi 4, yang tidak valid untuk grid yang hanya memiliki 3 kolom. Ubah kode dalam fungsi `plt.subplot()` dari `plt.subplot(1, 3, i + 1)` menjadi `plt.subplot(1, len(filtered_categorical_columns), i + 1)`. Perubahan ini menyesuaikan jumlah kolom grid secara dinamis berdasarkan jumlah kolom dalam `filtered_categorical_columns`.

Error Notes

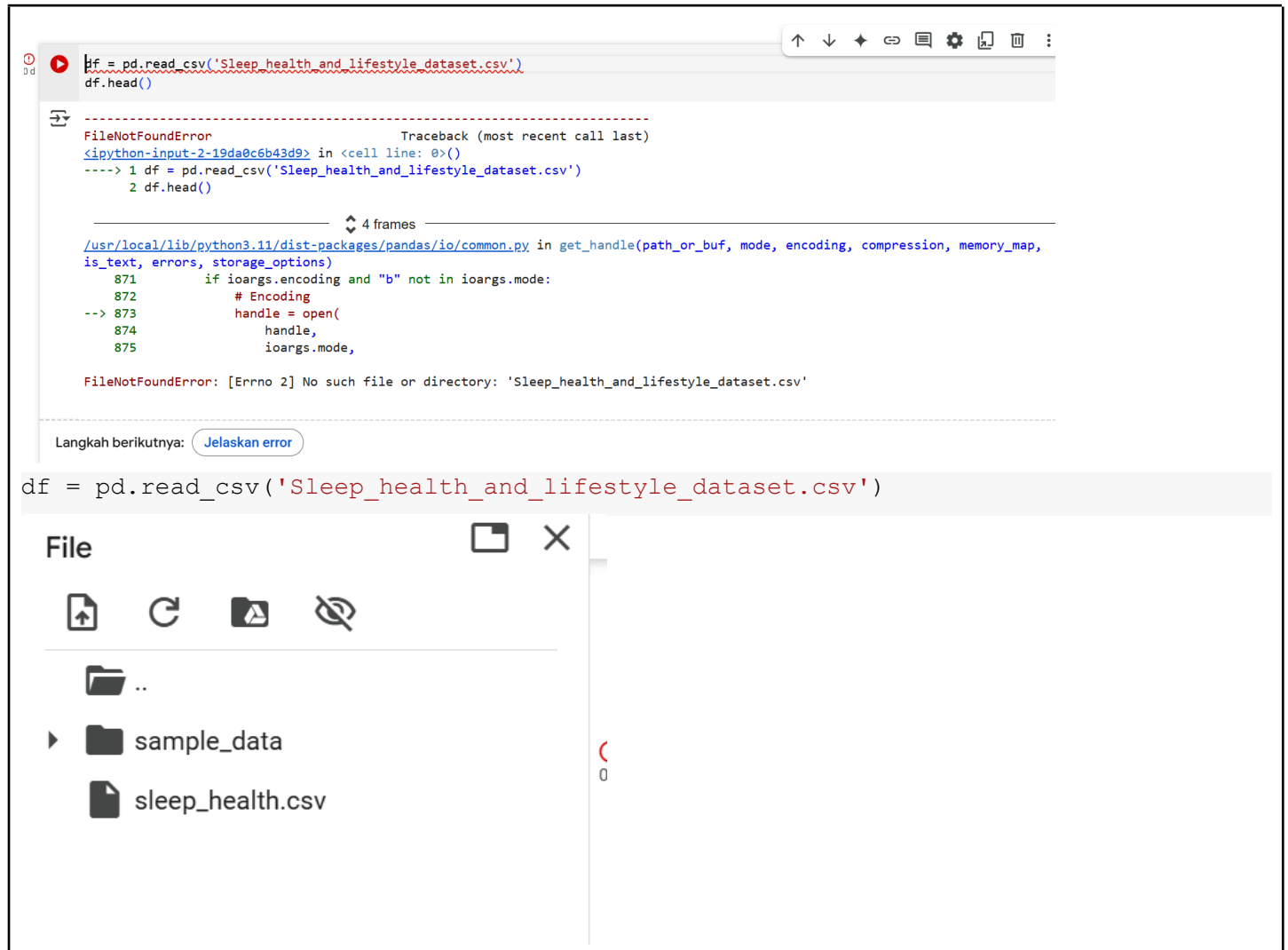
Pada project yang diperiksa terjadi sebuah error saat

- Gathering Data : Kesalahan saat membaca dataset dikarenakan salah penulisan yang tidak sesuai
- Assessing Data : kesalahan karena jumlah subplot yang ingin dibuat melebihi kapasitas grid yang telah ditentukan.
- Cleaning Data : Kesalahan pada saat penghapusan kolom yang seharusnya tidak perlu dihapus karena memang sudah tidak ada kolom tersebut di `df_clean`
- Feature Engineering : Kesalahan kolom `blood pressure` tidak ada. Periksa keberadaan kolom dengan `if 'blood_pressure' in df_clean.columns:`
- Exploratory Data Analysis di bagian Univariate Analysis : Kode `plt.subplot(1, 3, i + 1)` menghasilkan kesalahan karena indeks subplot (`i + 1`) melebihi jarak yang valid untuk grid 1x3. Kesalahan ini terjadi saat `i + 1` menjadi 4, yang tidak valid untuk grid yang hanya memiliki 3 kolom.

saya mengatasinya dengan cara

- Pemeriksaan format penulisan dataset
- Validasi Kolom sebelum melakukan operasi pada kolom, selalu memeriksa keberadaannya untuk menghindari KeyError.
- Mengubah parameter pada fungsi subplot agar sesuai dengan jumlah kolom yang ada, mencegah ValueError.
- Menambahkan log atau pesan kesalahan untuk memudahkan pelacakan masalah di setiap tahap

Code Review



The screenshot shows a Jupyter Notebook interface. The top part displays a code cell with the following code:

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

Below the code cell, a traceback for a `FileNotFoundError` is shown. The error message is: `FileNotFoundError: [Errno 2] No such file or directory: 'Sleep_health_and_lifestyle_dataset.csv'`. The traceback indicates that the error occurred in the `read_csv` function of the `pandas` library.

Below the error message, there is a button labeled "Jelaskan error".

Below the error message, the code cell is shown again with the following code:

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

Below the code cell, a file explorer window is open, showing the contents of the `sample_data` directory. The files listed are `sleep_health.csv` and `sleep_health_and_lifestyle_dataset.csv`.

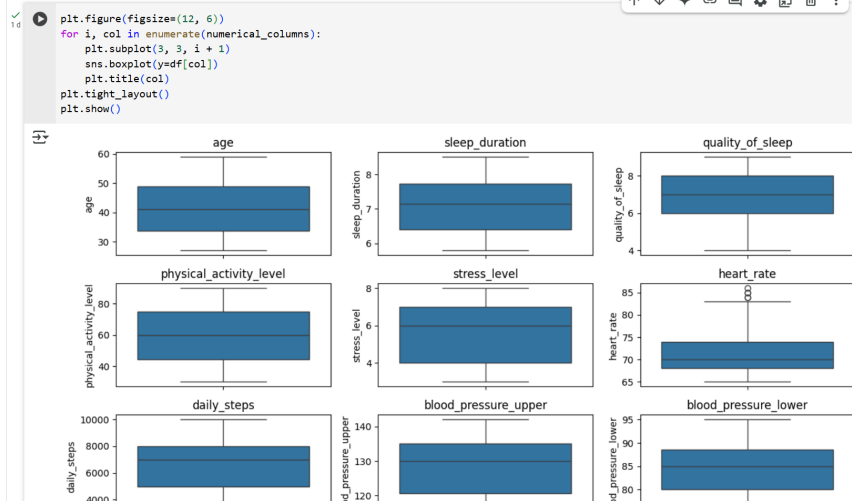
Pembacaan data pada saat awal bagian Gathering Data harus sesuai dengan format nama pada dataset yang disediakan, pada kasus ini format nama pada dataset adalah 'sleep_health.csv' namun pada terdapat kesalahan penulisan format nama pada dataset sehingga dataset tidak terbaca/mengalami error, maka dari itu harus diubah menjadi sesuai format seperti dibawah ini :



```
code plt.figure(figsize=(12, 6))
for i, col in enumerate(numerical_columns):
    plt.subplot(2, 4, i + 1)
    sns.boxplot(y=df[col])
    plt.title(col)
    plt.tight_layout()
    plt.show()
```

Kesalahan yang muncul karena code yang dibuat lebih banyak subplot daripada tata letak grid yang telah ditentukan dengan code 'plt.subplot(2, 4, i+1)'. Baris yang dibuat membuat grid 2x4, yang dapat menampung maksimal 8 subplot. Namun code yang dibuat melakukan iterasi melalui numerical_columns, dan ketika i menjadi 8 dan code mencoba membuat subplot ke-9 yang menghasilkan ValueError.

feedback solusi yang sesuai letak grid di plt.subplot() untuk mengakomodasi semua kolom numerik yang di buat plot seperti di bawah ini :



Perubahan `plt.subplot()` dengan mengubah argumen yang sebelumnya '`plt.subplot(2, 4, i+1)`' menjadi `plt.subplot(3, 3, i+1)`. Perubahan ini akan membuat grid 3x3 (9 subplot), yang akan cukup untuk mengakomodasi semua kolom numerik

```
df_clean.drop(columns=['person_id'], inplace=True)
df_clean.info()
```

```
-----
KeyError                                Traceback (most recent call last)
<ipython-input-24-1c366183cd24> in <cell line: 0>()
----> 1 df_clean.drop(columns=['person_id'], inplace=True)
      2 df_clean.info()
```

```
-----
3 frames -----
/usr/local/lib/python3.11/dist-packages/pandas/core/indexes/base.py in drop(self, labels, errors)
    7068         if mask.any():
    7069             if errors != "ignore":
-> 7070                 raise KeyError(f"{labels[mask].tolist()} not found in axis")
    7071             indexer = indexer[~mask]
    7072             return self.delete(indexer)

KeyError: "[ 'person_id' ] not found in axis"
```

Langkah berikutnya: [Jelaskan error](#)

code `df_clean.drop(columns=['person_id'], inplace=True)`
Kesalahan yang terjadi pada kasus ini "KeyError: "['person_id'] not found in axis" yang mengidentifikasi bahwa kolom `person_id` tidak ada di data frame `df_clean`.

feedback di kasus ini kolom '`person_id`' memang tidak ada di dalam daftar kolom di data frame `df_clean`, solusi nya adalah tidak perlu untuk melakukan penghapusan kolom tersebut seperti di bawah ini :

```
print(df_clean.columns)
df_clean.info()
```

```
Index(['gender', 'age', 'occupation', 'sleep_duration', 'quality_of_sleep',
       'physical_activity_level', 'stress_level', 'bmi_category', 'heart_rate',
       'daily_steps', 'sleep_disorder', 'blood_pressure_upper',
       'blood_pressure_lower', 'blood_pressure_category', 'age_group'],
      dtype='object')
<class 'pandas.core.DataFrame'>
RangeIndex: 132 entries, 0 to 131
Data columns (total 15 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   gender                                132 non-null    object
1   age                                  132 non-null    int64
2   occupation                            132 non-null    object
3   sleep_duration                        132 non-null    float64
4   quality_of_sleep                      132 non-null    int64
5   physical_activity_level               132 non-null    int64
6   stress_level                         132 non-null    int64
7   bmi_category                         132 non-null    object
8   heart_rate                           132 non-null    int64
9   daily_steps                          132 non-null    int64
10  sleep_disorder                       132 non-null    object
11  blood_pressure_upper                 132 non-null    int64
12  blood_pressure_lower                 132 non-null    int64
13  blood_pressure_category              132 non-null    object
14  age_group                           132 non-null    object
dtypes: float64(1), int64(8), object(6)
memory usage: 15.6+ KB
```

```
df_clean = pd.concat([df_clean, df_clean['blood_pressure'].str.split('/', expand=True)], axis=1).drop('blood_pressure', axis=1)
df_clean = df_clean.rename(columns={0: 'blood_pressure_upper', 1: 'blood_pressure_lower'})
df_clean['blood_pressure_upper'] = df_clean['blood_pressure_upper'].astype(int)
df_clean['blood_pressure_lower'] = df_clean['blood_pressure_lower'].astype(int)

df_clean.head()
```

```
Traceback (most recent call last)
/usr/local/lib/python3.11/dist-packages/pandas/core/indexes/base.py in get_loc(self, key)
    3804     try:
-> 3805         return self._engine.get_loc(casted_key)
    3806     except KeyError as err:
        except KeyError as err:

index.pyx in pandas._libs.index.IndexEngine.get_loc()

index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'blood_pressure'

The above exception was the direct cause of the following exception:

KeyError                                Traceback (most recent call last)
      3810     ):
      3811         raise InvalidIndexError(key)
-> 3812     raise KeyError(key) from err
      3813 except TypeError:
      3814     # If we have a listlike key, _check_indexing_error will raise

KeyError: 'blood_pressure'
```

code `df_clean = pd.concat([df_clean, df_clean['blood_pressure'].str.split('/', expand=True)], axis=1).drop('blood_pressure', axis=1)`
Kesalahan `KeyError: 'blood_pressure'` karena tidak ditemukan di `df_clean` dataframe ketika mencoba mengakses nya dengan menggunakan code `df_clean['blood_pressure']`. Hal ini mungkin terjadi karena kolom tersebut sudah tidak ada/dihapus saat Featuring Engineering sebelumnya dan memiliki kolom blood pressure kategori yang lain.

feedback : solusi yang diberikan membuat code pemeriksaan terlebih dulu apakah kolom tersebut masih ada di data frame `df_clean` atau tidak. Code pemeriksaan ini dilakukan dengan kondisi `if 'blood_pressure' in df_clean.columns:`. Jika kolom tersebut ada maka akan menjalankan proses Feature Engineering (Memisahkan nilai tekanan darah 'blood_pressure_upper' dan 'blod_pressure_lower' lalu mengubah tipe data kolom menjadi integer)

Namun jika tidak ada maka akan menampilkan pesan “The ‘bood_pressure’ column has already been processed” dan tidak akan menjaankan proses feature engineering lagi seperti yang terlihat di bawah ini :

```
if 'blood_pressure' in df_clean.columns:
    df_clean = pd.concat([df_clean, df_clean['blood_pressure'].str.split('/', expand=True)], axis=1).drop('blood_pressure', axis=1)
    df_clean = df_clean.rename(columns={0: 'blood_pressure_upper', 1: 'blood_pressure_lower'})
    df_clean['blood_pressure_upper'] = df_clean['blood_pressure_upper'].astype(int)
    df_clean['blood_pressure_lower'] = df_clean['blood_pressure_lower'].astype(int)
else:
    print("The 'blood_pressure' column has already been processed.")

df_clean.head()
```

The 'blood_pressure' column has already been processed.

	gender	age	occupation	sleep_duration	quality_of_sleep	physical_activity_level	stress_level	bmi_category	heart_rate	dail
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	77	
1	Male	28	Doctor	6.2	6	60	8	Normal	75	
2	Male	28	Sales Representative	5.9	4	30	8	Obese	85	
3	Male	28	Software Engineer	5.9	4	30	8	Obese	85	
4	Male	29	Teacher	6.3	6	40	7	Obese	82	

Langkah berikutnya: [Buat kode dengan df_clean](#) [Lihat plot yang direkomendasikan](#) [New interactive sheet](#)

```

categorical_columns = df_eda.select_dtypes(include=['object', 'category']).columns
filtered_categorical_columns = categorical_columns.drop(['occupation', 'blood_pressure_category'])

plt.figure(figsize=(12, 4))
for i, col in enumerate(filtered_categorical_columns):
    plt.subplot(1, 3, i + 1)
    sns.countplot(x=df_eda[col])
    plt.title(col)
    plt.xlabel('')
    plt.ylim(0, df_eda[col].value_counts().max()*1.08)
    for p in plt.gca().patches:
        plt.gca().annotate(int(p.get_height()), (p.get_x() + p.get_width() / 2., p.get_height()),
                           ha='center', va='center', xytext=(0, 5), textcoords='offset points')

plt.tight_layout()
plt.show()

```

```

-----
ValueError                                Traceback (most recent call last)
<ipython-input-43-8c1ef9eb97d5> in <cell line: 0>()
      4 plt.figure(figsize=(12, 4))
----> 5 for i, col in enumerate(filtered_categorical_columns):
      6     plt.subplot(1, 3, i + 1)
      7     sns.countplot(x=df_eda[col])
      8     plt.title(col)

1 frames
/usr/local/lib/python3.11/dist-packages/matplotlib/gridspec.py in _from_subplot_args(figure, args)
    587     else:
    588         if not isinstance(num, Integral) or num < 1 or num > rows*cols:
--> 589             raise ValueError(
    590                 f"num must be an integer with 1 <= num <= {rows*cols}, "
    591                 f"not {num!r}"
)
ValueError: num must be an integer with 1 <= num <= 3, not 4

```

code : `plt.subplot(1, 3, i + 1)`

kesalahan yang terjadi harus berupa bilangan bulat dengan $1 \leq \text{num} \leq 3$, not 4 bahwa kode in mencoba membuat subplot dengan indeks (num) yang berada di luar rentang yang valid untuk kisi-kisi subplot ditentukan. `plt.subplot(1, 3, i+1)` mendefinisikan grid dengan 1 baris dan 3 kolom $i+1$ mewakili indeks dari subplot yang dibuat. Kesalahan terjadi ketika $i + 1$ menjadi 4, yang berada di luar rentang yang valid untuk grid 1x3 (rentang yang valid adalah 1 hingga 3). Hal ini terjadi karena `filtered_categorical_columns` berisi 4 item, dan perulangan mengulanginya, sehingga menghasilkan indeks 4 pada perulangan terakhir.

feedback : Perubahan yang paling penting adalah di dalam fungsi `plt.subplot()` yang sebelumnya `plt.subplot(1, 3, i + 1)` menjadi `plt.subplot(1, len(filtered_categorical_columns), i + 1)`. Secara dinamis menyesuaikan grid subplot berdasarkan jumlah kolom di `filtered_categorical_columns`. Sekarang, jika ada 3 kolom yang difilter, akan memiliki grid 1x3; jika ada 4 kolom yang difilter, dan akan memiliki grid 1x4, dan seterusnya. Hal ini memastikan bahwa indeks ($i + 1$) selalu berada dalam rentang yang valid, sehingga mencegah terjadinya `ValueError`.

code

feedback