```
%pip install ucimlrepo
%pip install -U ydata-profiling
from ucimlrepo import fetch_ucirepo
import pandas as pd
from ydata_profiling import ProfileReport
from scipy import stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import wilcoxon, shapiro
```

```
Requirement already satisfied: ucimlrepo in /usr/local/lib/python3.11/dist-packages (0.0.7)
Requirement already satisfied: pandas>=1.0.0 in /usr/local/lib/python3.11/dist-packages (from ucimlrepo) (2.2.2)
Requirement already satisfied: certifi>=2020.12.5 in /usr/local/lib/python3.11/dist-packages (from ucimlrepo) (2025.6.15)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.0->ucimlrepo) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.0->ucimlrepo) (2.9.0.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.0->ucimlrepo) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.0->ucimlrepo) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas>=1.0.0->ucimlrep
Requirement already satisfied: ydata-profiling in /usr/local/lib/python3.11/dist-packages (4.16.1)
Requirement already satisfied: scipy<1.16,>=1.4.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.15.3)
Requirement already satisfied: pandas!=1.4.0,<3.0,>1.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.2.2)
Requirement already satisfied: matplotlib<=3.10,>=3.5 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (3.10.0)
Requirement already satisfied: pydantic>=2 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.11.7)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (6.0.2)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (3.1.6)
Requirement already satisfied: visions<0.8.2,>=0.7.5 in /usr/local/lib/python3.11/dist-packages (from visions[type_image_path]<0.8.2,>=0
Requirement already satisfied: numpy<2.2,>=1.16.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.0.2)
Requirement already satisfied: htmlmin==0.1.12 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.1.12)
Requirement already satisfied: phik<0.13,>=0.11.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.12.4)
Requirement already satisfied: requests<3,>=2.24.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.32.3)
Requirement already satisfied: tqdm<5,>=4.48.2 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (4.67.1)
Requirement already satisfied: seaborn<0.14,>=0.10.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.13.2)
Requirement already satisfied: multimethod<2,>=1.4 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.12)
Requirement already satisfied: statsmodels<1,>=0.13.2 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.14.4)
Requirement already satisfied: typeguard<5,>=3 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (4.4.4)
Requirement already satisfied: imagehash==4.3.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (4.3.1)
Requirement already satisfied: wordcloud>=1.9.3 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.9.4)
Requirement already satisfied: dacite>=1.8 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.9.2)
Requirement already satisfied: numba<=0.61,>=0.56.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.60.0)
Requirement already satisfied: PyWavelets in /usr/local/lib/python3.11/dist-packages (from imagehash==4.3.1->ydata-profiling) (1.8.0)
Requirement already satisfied: pillow in /usr/local/lib/python3.11/dist-packages (from imagehash==4.3.1->ydata-profiling) (11.2.1)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2<3.2,>=2.11.1->ydata-profiling) (3
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profiling
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profiling) (0
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profilin
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profilin
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profiling)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profiling
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profi
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba<=0.61,>=0.56.0->ydata-p
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas!=1.4.0,<3.0,>1.1->ydata-profiling) (
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas!=1.4.0,<3.0,>1.1->ydata-profiling)
Requirement already satisfied: joblib>=0.14.1 in /usr/local/lib/python3.11/dist-packages (from phik<0.13,>=0.11.1->ydata-profiling) (1.5
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profiling) (0.
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profiling) (2.3
Requirement already satisfied: typing-extensions>=4.12.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profiling)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profiling) (
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-prof
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-profiling) (3.10
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-profiling)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-profiling)
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.11/dist-packages (from statsmodels<1,>=0.13.2->ydata-profiling) (1
Requirement already satisfied: attrs>=19.3.0 in /usr/local/lib/python3.11/dist-packages (from visions<0.8.2,>=0.7.5->visions[type_image_
Requirement already satisfied: networkx>=2.4 in /usr/local/lib/python3.11/dist-packages (from visions<0.8.2,>=0.7.5->visions[type_image_
Requirement already satisfied: puremagic in /usr/local/lib/python3.11/dist-packages (from visions<0.8.2,>=0.7.5->visions[type_image_path
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib<=3.10,>=3.5->y
```

```
obesity = fetch_ucirepo(id=544)  # Load the Obesity dataset by ID
df = pd.concat([obesity.data.features, obesity.data.targets], axis=1)

#transforming target categorical variable to numeric
column_rename_map = {
    'FAVC': 'FAVC_FrequentHighCaloricFood',
    'FCVC': 'FCVC_VegetableConsumptionFreq',
    'NCP': 'NCP_NumberOfMainMeals',
    'CAEC': 'CAEC_BetweenMealSnacking',
    'CH2O': 'CH2O_DailyWaterIntake',
    'SCC': 'SCC_CalorieMonitoring',
    'FAF': 'FAF_PhysicalActivityFreq',
    'TUE': 'TUE_ScreenTimeHours',
    'CALC': 'CALC_AlcoholConsumption',
    'MTRANS': 'MTRANS_TransportationMode',
    'NObeyesdad': 'NObeyesdad_ObesityLevel'
}

#df = df.rename(columns=column_rename_map)

#print(df.columns.tolist())
df=df.rename(columns=column_rename_map)
profile = ProfileReport(df, title="YData Profiling Report")
profile.to_notebook_iframe()
```

```
Summarize dataset: 100%                                          90/90 [00:13<00:00,  3.72it/s, Completed]

  0%|          | 0/17 [00:00<?, ?it/s]
 24%|██▎       | 4/17 [00:00<00:00, 34.12it/s]
 47%|████▋     | 8/17 [00:00<00:00, 32.03it/s]
 71%|███████   | 12/17 [00:00<00:00, 29.15it/s]
100%|██████████| 17/17 [00:00<00:00, 32.17it/s]
Generate report structure: 100%                                       1/1 [00:05<00:00,  5.42s/it]

Render HTML: 100%                                             1/1 [00:03<00:00,  3.86s/it]
```

YData Profiling Report   Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Overview

Brought to you by YData

Overview   Alerts 12   Reproduction

Dataset statistics

Number of variables	17
Number of observations	2111
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	9
Duplicate rows (%)	0.4%
Total size in memory	280.5 KiB
Average record size in memory	136.1 B

Variable types

Categorical	5
Numeric	8
Boolean	4

Variables

Select Columns

```
    # to find the duplicate
    duplicates = df[df.duplicated(keep=False)]

    # Display the duplicates
    #print(f" Total duplicate rows: {len(duplicates)}")
```

```
print(duplicates)
#print(duplicates.count())
#duplicates.to_csv('duplicates_output.csv', index=False) # output as csv
cleaned_df=df.drop_duplicates(keep='first')
df=cleaned_df
print(df.info())
```

```
     Gender   Age  Height  Weight family_history_with_overweight  \
97   Female  21.0    1.52    42.0                             no
98   Female  21.0    1.52    42.0                             no
105  Female  25.0    1.57    55.0                             no
106  Female  25.0    1.57    55.0                             no
145    Male  21.0    1.62    70.0                             no
174    Male  21.0    1.62    70.0                             no
179    Male  21.0    1.62    70.0                             no
184    Male  21.0    1.62    70.0                             no
208  Female  22.0    1.69    65.0                            yes
209  Female  22.0    1.69    65.0                            yes
282  Female  18.0    1.62    55.0                            yes
295  Female  16.0    1.66    58.0                             no
309  Female  16.0    1.66    58.0                             no
443    Male  18.0    1.72    53.0                            yes
460  Female  18.0    1.62    55.0                            yes
466    Male  22.0    1.74    75.0                            yes
467    Male  22.0    1.74    75.0                            yes
496    Male  18.0    1.72    53.0                            yes
523  Female  21.0    1.52    42.0                             no
527  Female  21.0    1.52    42.0                             no
659  Female  21.0    1.52    42.0                             no
663  Female  21.0    1.52    42.0                             no
763    Male  21.0    1.62    70.0                             no
764    Male  21.0    1.62    70.0                             no
824    Male  21.0    1.62    70.0                             no
830    Male  21.0    1.62    70.0                             no
831    Male  21.0    1.62    70.0                             no
832    Male  21.0    1.62    70.0                             no
833    Male  21.0    1.62    70.0                             no
834    Male  21.0    1.62    70.0                             no
921    Male  21.0    1.62    70.0                             no
922    Male  21.0    1.62    70.0                             no
923    Male  21.0    1.62    70.0                             no

    FAVC_FrequentHighCaloricFood  FCVC_VegetableConsumptionFreq  \
97                            no                            3.0
98                            no                            3.0
105                          yes                            2.0
106                          yes                            2.0
145                          yes                            2.0
174                          yes                            2.0
179                          yes                            2.0
184                          yes                            2.0
208                          yes                            2.0
209                          yes                            2.0
282                          yes                            2.0
295                           no                            2.0
309                           no                            2.0
443                          yes                            2.0
460                          yes                            2.0
466                          yes                            3.0
467                          yes                            3.0
496                          yes                            2.0
523                          yes                            3.0
527                          yes                            3.0
659                          yes                            3.0
663                          yes                            3.0
```

```
#transforming target categorical variable to numeric

df['NObeyesdad_ObesityLevel']=df['NObeyesdad_ObesityLevel'].str.strip() # to remove whitespace after/before NObeyesdad
# use manual mapping
obesity_mapping= {
    'Insufficient_Weight': 0,
    'Normal_Weight':1,
    'Overweight_Level_I':2,
    'Overweight_Level_II':3,
    'Obesity_Type_I':4,
    'Obesity_Type_II':5,
    'Obesity_Type_III':6
    }

df['NObesyesdad_encoded'] = df['NObeyesdad_ObesityLevel'].map(obesity_mapping)
df['NObeyesdad_ObesityLevel']=df['NObesyesdad_encoded']
```

```
print(df['NObeyesdad_ObesityLevel'])

df=df.drop('NObesyesdad_encoded', axis=1)
print(df)
#print(df.columns.tolist())
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
/tmp/ipython-input-96-279510089.py in <cell line: 0>()
      1 #transforming target categorical variable to numeric
      2 
----> 3 df['NObeyesdad_ObesityLevel']=df['NObeyesdad_ObesityLevel'].str.strip() # to remove whitespace after/before NObeyesdad
      4 # use manual mapping
      5 obesity_mapping= {

                                   3 frames
/usr/local/lib/python3.11/dist-packages/pandas/core/strings/accessor.py in _validate(data)
    243 
    244         if inferred_dtype not in allowed_types:
--> 245             raise AttributeError("Can only use .str accessor with string values!")
    246         return inferred_dtype
    247 

AttributeError: Can only use .str accessor with string values!
```

```
# normality test of continuous variable FCVC_VegetableConsumptionFreq
mean_FCVC=df['FCVC_VegetableConsumptionFreq'].mean()
print(f" Mean FCVC_VegetableConsumptionFreq : {mean_FCVC :.2f}")

# Shapiro-Wilk test on the full column
stat, p_value= shapiro(df['FCVC_VegetableConsumptionFreq'])
print(f"Shapiro-Wilk Test= p- value = {p_value}")
if p_value>0.05:
  print("Data is normallly distributed")
else:
  print("Data is not normally distributed")

# normality test of NCP_NumberOfMainMeals
mean_NCP=df['NCP_NumberOfMainMeals'].mean()
print(f"Mean NCP_NumberOfMainMeals: {mean_NCP:.2f}")
stat, p_value= shapiro(df['NCP_NumberOfMainMeals'])
print(f"Shapiro-Wilk test:p-value ={p_value}")
if p_value>0.05:
  print("Data is normally distributed")
else:
  print("Data is not normally distributed")

# normality test of FAF_PhysicalActivityFreq
mean_FAF=df['FAF_PhysicalActivityFreq'].mean()
print(f"Mean FAF_PhysicalActivityFreq: {mean_FAF:.2f}")
stat, p_value= shapiro(df['FAF_PhysicalActivityFreq'])
print(f"Shapiro-Wilk test:p-value ={p_value}")
if p_value>0.05:
  print("Data is normally distributed")
else:
  print("Data is not normally distributed")


# normality test of CH2O_DailyWaterIntake
mean_CH2O=df['CH2O_DailyWaterIntake'].mean()
print(f"Mean CH2O_DailyWaterIntake: {mean_CH2O:.2f}")
stat, p_value= shapiro(df['CH2O_DailyWaterIntake'])
print(f"Shapiro-Wilk test:p-value ={p_value}")
if p_value>0.05:
  print("Data is normally distributed")
else:
  print("Data is not normally distributed")
```

```
 Mean FCVC_VegetableConsumptionFreq : 2.42
Shapiro-Wilk Test= p- value = 4.380314063568239e-41
Data is not normally distributed
Mean NCP_NumberOfMainMeals: 2.70
Shapiro-Wilk test:p-value =2.821421039483737e-49
Data is not normally distributed
Mean FAF_PhysicalActivityFreq: 1.01
Shapiro-Wilk test:p-value =1.128105084773664e-32
```

```
Data is not normally distributed
Mean CH2O_DailyWaterIntake: 2.00
Shapiro-Wilk test:p-value =2.9302074063783406e-29
Data is not normally distributed
```

```
# applying one hot encoding to caregorical variable CAEC_BetweenMealSnacking
df['CAEC_BetweenMealSnacking_Sometimes'] = (df['CAEC_BetweenMealSnacking'] == 'Sometimes').astype(int)
df['CAEC_BetweenMealSnacking_Frequently'] = (df['CAEC_BetweenMealSnacking'] == 'Frequently').astype(int)
df['CAEC_BetweenMealSnacking_Always'] = (df['CAEC_BetweenMealSnacking'] == 'Always').astype(int)
df['CAEC_BetweenMealSnacking'] = (df['CAEC_BetweenMealSnacking'] == 'No').astype(int)
print(df)
```

```
      Gender        Age    Height      Weight family_history_with_overweight  \
0     Female  21.000000  1.620000   64.000000                            yes
1     Female  21.000000  1.520000   56.000000                            yes
2       Male  23.000000  1.800000   77.000000                            yes
3       Male  27.000000  1.800000   87.000000                             no
4       Male  22.000000  1.780000   89.800000                             no
...      ...        ...       ...         ...                            ...
2106  Female  20.976842  1.710730  131.408528                            yes
2107  Female  21.982942  1.748584  133.742943                            yes
2108  Female  22.524036  1.752206  133.689352                            yes
2109  Female  24.361936  1.739450  133.346641                            yes
2110  Female  23.664709  1.738836  133.472641                            yes

     FAVC_FrequentHighCaloricFood  FCVC_VegetableConsumptionFreq  \
0                              no                            2.0
1                              no                            3.0
2                              no                            2.0
3                              no                            3.0
4                              no                            2.0
...                           ...                            ...
2106                          yes                            3.0
2107                          yes                            3.0
2108                          yes                            3.0
2109                          yes                            3.0
2110                          yes                            3.0

      NCP_NumberOfMainMeals  CAEC_BetweenMealSnacking SMOKE  ...  \
0                       3.0                         0    no  ...
1                       3.0                         0   yes  ...
2                       3.0                         0    no  ...
3                       3.0                         0    no  ...
4                       1.0                         0    no  ...
...                     ...                       ...   ...  ...
2106                    3.0                         0    no  ...
2107                    3.0                         0    no  ...
2108                    3.0                         0    no  ...
2109                    3.0                         0    no  ...
2110                    3.0                         0    no  ...

      SCC_CalorieMonitoring FAF_PhysicalActivityFreq  TUE_ScreenTimeHours  \
0                        no                 0.000000             1.000000
1                       yes                 3.000000             0.000000
2                        no                 2.000000             1.000000
3                        no                 2.000000             0.000000
4                        no                 0.000000             0.000000
...                     ...                      ...                  ...
2106                     no                 1.676269             0.906247
2107                     no                 1.341390             0.599270
2108                     no                 1.414209             0.646288
2109                     no                 1.139107             0.586035
2110                     no                 1.026452             0.714137

      CALC_AlcoholConsumption MTRANS_TransportationMode  \
0                          no     Public_Transportation
1                   Sometimes     Public_Transportation
2                  Frequently     Public_Transportation
3                  Frequently                   Walking
4                   Sometimes     Public_Transportation
```