hubert_mentel_3

November 8, 2024

[1]:
```
#ladowanie biblioteki Pandas
import pandas as pd
```

[3]:
```
#Zadanie 1
df = pd.read_csv('IHME_ORB_C19HSDS_2020_Y2020M12D03.CSV')
df
print(df.head())

# Sprawd podstawowe informacje o danych
print(df.info())

# Wywietl podstawowe statystyki opisowe
print(df.describe())
```

```
      SbjNum NetDuration InterviewTimeVStart InterviewTimeVEnd  \
0  133476254     0:10:14     7/17/2020 13:53   7/17/2020 14:26
1  133281846     0:22:16     7/10/2020 12:53   7/10/2020 14:47
2  133280780     0:19:23     7/10/2020 12:35   7/10/2020 12:54
3  133281834     0:10:11     7/10/2020 10:21   7/10/2020 10:32
4  133491249     0:09:59      7/18/2020 8:27    7/18/2020 8:39

             Date  Srvyr  Country  LANG  R1  R1_5  … G11_Other G11_99  \
0  7/17/2020 8:53   3232        2     1   9  15.0  …       NaN    NaN
1  7/10/2020 7:53   3206        2     4  12  22.0  …       NaN    NaN
2  7/10/2020 7:35   3202        2     3  10  13.0  …       NaN    NaN
3  7/10/2020 5:21   3212        2     1  12   9.0  …       NaN    NaN
4  7/18/2020 3:27   3225        2     3  11  28.0  …       NaN    NaN

   FinalOutcome  NumOfVisits  weight_combined kenya_weight  nigeria_weight  \
0             1            1         0.829860          NaN        0.829860
1             1            1         1.416946          NaN        1.416946
2             1            1         0.883601          NaN        0.883601
3             1            1         1.416946          NaN        1.416946
4             1            1         0.829860          NaN        0.829860

  southafrica_weight  agegroup gk_weight
0                NaN         1  1.555754
1                NaN         2  1.949579
```

```
2                NaN         2  2.151458
3                NaN         2  2.325065
4                NaN         1  1.640484

[5 rows x 247 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3058 entries, 0 to 3057
Columns: 247 entries, SbjNum to gk_weight
dtypes: float64(208), int64(18), object(21)
memory usage: 5.8+ MB
None
             SbjNum        Srvyr      Country         LANG           R1  \
count  3.058000e+03  3058.000000  3058.000000  3058.000000  3058.000000
mean   1.333905e+08  3084.743623     2.012426     2.180510    22.521583
std    1.256690e+05    97.219107     0.817203     2.364438    20.294923
min    1.331715e+08  3001.000000     1.000000     1.000000     1.000000
25%    1.332825e+08  3010.000000     1.000000     1.000000     7.000000
50%    1.333755e+08  3026.000000     2.000000     1.000000    11.000000
75%    1.334985e+08  3212.000000     3.000000     1.000000    49.000000
max    1.336450e+08  3264.000000     3.000000    11.000000    54.000000

              R1_5           R4           R5           R6           R7  …  \
count  1016.000000  3058.000000  3058.000000  3058.000000  3058.000000  …
mean     27.378937     1.503270    34.038914     1.771419    24.448986  …
std      10.088041     0.500071    11.386285     3.130841    26.377909  …
min       9.000000     1.000000    18.000000     1.000000     1.000000  …
25%      18.750000     1.000000    25.000000     1.000000     8.000000  …
50%      28.000000     2.000000    31.500000     2.000000    21.000000  …
75%      35.000000     2.000000    40.000000     2.000000    24.000000  …
max      45.000000     2.000000    99.000000    99.000000    99.000000  …

          G11_96  G11_99  FinalOutcome  NumOfVisits  weight_combined  \
count  32.000000     2.0        3058.0  3058.000000      3058.000000
mean    0.718750     1.0           1.0     1.130150         1.000987
std     0.456803     0.0           0.0     0.449694         0.403105
min     0.000000     1.0           1.0     1.000000         0.799916
25%     0.000000     1.0           1.0     1.000000         0.829860
50%     1.000000     1.0           1.0     1.000000         0.883601
75%     1.000000     1.0           1.0     1.000000         1.000000
max     1.000000     1.0           1.0     5.000000         3.791351

       kenya_weight  nigeria_weight  southafrica_weight     agegroup  \
count   1002.000000     1016.000000              1040.0  3058.000000
mean       1.000000        1.002970                 1.0     1.475147
std        0.568149        0.413589                 0.0     0.653771
min        0.799916        0.829860                 1.0     1.000000
25%        0.799916        0.829860                 1.0     1.000000
50%        0.799916        0.872352                 1.0     1.000000
```

```
75%         1.157689         0.883601                  1.0      2.000000
max         3.791351         2.722043                  1.0      3.000000

          gk_weight
count  3058.000000
mean      2.325065
std       0.858712
min       1.555754
25%       1.753344
50%       2.046237
75%       2.325065
max       7.110619

[8 rows x 226 columns]
```

[5]:
```
# Zadanie 2
mean_weight = df["kenya_weight"].mean()
print(f"Srednia waga kenya : {mean_weight}")

# Oblicz median dla kolumny 'dochd'
median_income = df["nigeria_weight"].median()
print(f"Mediana wagi nigeria: {median_income}")

# Oblicz odchylenie standardowe dla kolumny 'kenya_weight'
std_age = df["kenya_weight"].std()
print(f"Odchylenie standardowe wieku: {std_age}")
```

```
Srednia waga kenya : 0.999999999704591
Mediana wagi nigeria: 0.87235241
Odchylenie standardowe wieku: 0.5681490293617802
```

[7]:
```
# Zadanie 3: Identyfikacja brakujących danych
missing_data_columns = df.isnull().sum()

# Uzupełnienie brakujących danych w kolumnie 'kenya_weight' średnią
df['kenya_weight'].fillna(mean_weight)

# Usunięcie wierszy z brakującymi danymi w kolumnie 'nigeria_weight'
df_cleaned = df.dropna(subset=['nigeria_weight'])

missing_data_columns, df_cleaned.head()
```

[7]:
```
(SbjNum                 0
 NetDuration            0
 InterviewTimeVStart    0
 InterviewTimeVEnd      0
 Date                   0
                       …
```

```
kenya_weight           2056
nigeria_weight         2042
southafrica_weight     2018
agegroup                  0
gk_weight                 0
Length: 247, dtype: int64,
      SbjNum NetDuration InterviewTimeVStart InterviewTimeVEnd  \
0  133476254     0:10:14     7/17/2020 13:53   7/17/2020 14:26
1  133281846     0:22:16     7/10/2020 12:53   7/10/2020 14:47
2  133280780     0:19:23     7/10/2020 12:35   7/10/2020 12:54
3  133281834     0:10:11     7/10/2020 10:21   7/10/2020 10:32
4  133491249     0:09:59      7/18/2020 8:27    7/18/2020 8:39

              Date  Srvyr  Country  LANG  R1  R1_5  … G11_Other G11_99  \
0  7/17/2020 8:53   3232        2     1   9  15.0  …       NaN    NaN
1  7/10/2020 7:53   3206        2     4  12  22.0  …       NaN    NaN
2  7/10/2020 7:35   3202        2     3  10  13.0  …       NaN    NaN
3  7/10/2020 5:21   3212        2     1  12   9.0  …       NaN    NaN
4  7/18/2020 3:27   3225        2     3  11  28.0  …       NaN    NaN

   FinalOutcome  NumOfVisits  weight_combined kenya_weight  nigeria_weight  \
0             1            1         0.829860          NaN        0.829860
1             1            1         1.416946          NaN        1.416946
2             1            1         0.883601          NaN        0.883601
3             1            1         1.416946          NaN        1.416946
4             1            1         0.829860          NaN        0.829860

  southafrica_weight  agegroup gk_weight
0                NaN         1  1.555754
1                NaN         2  1.949579
2                NaN         2  2.151458
3                NaN         2  2.325065
4                NaN         1  1.640484

[5 rows x 247 columns])
```

[9]:
```
# Zadanie 4 wykrywanie wartości odstających
# Oblicz IQR
Q1 = df['nigeria_weight'].quantile(0.25)
Q3 = df['nigeria_weight'].quantile(0.75)
IQR = Q3 - Q1

# Zidentyfikuj wartoci odstajce
outliers = df[(df['nigeria_weight'] < (Q1 - 1.5 * IQR)) | (df['nigeria_weight']␣
 ↪> (Q3 + 1.5 * IQR))]
print("Wartoci odstajce :")
print(outliers)
```

Wartoci odstajce :

	SbjNum	NetDuration	InterviewTimeVStart	InterviewTimeVEnd
1	133281846	0:22:16	7/10/2020 12:53	7/10/2020 14:47
3	133281834	0:10:11	7/10/2020 10:21	7/10/2020 10:32
25	133341539	0:09:55	7/13/2020 9:15	7/13/2020 9:39
43	133530487	0:13:07	7/19/2020 13:37	7/19/2020 13:52
44	133617209	0:14:32	7/23/2020 9:25	7/23/2020 9:43
…	…	…	…	…
983	133536715	0:10:23	7/19/2020 14:41	7/19/2020 15:09
984	133172154	0:10:35	7/6/2020 21:37	7/6/2020 22:11
1001	133292785	0:15:41	7/10/2020 16:37	7/10/2020 17:03
1006	133509771	0:16:05	7/18/2020 20:15	7/18/2020 20:32
1011	133350222	0:18:10	7/13/2020 11:03	7/13/2020 11:24

	Date	Srvyr	Country	LANG	R1	R1_5	…	G11_Other	G11_99
1	7/10/2020 7:53	3206	2	4	12	22.0	…	NaN	NaN
3	7/10/2020 5:21	3212	2	1	12	9.0	…	NaN	NaN
25	7/13/2020 4:15	3232	2	6	14	37.0	…	NaN	NaN
43	7/19/2020 8:37	3230	2	3	10	44.0	…	NaN	NaN
44	7/23/2020 4:25	3204	2	3	11	29.0	…	NaN	NaN
…	…	…	…	…	..	…	…	…	…
983	7/19/2020 9:41	3230	2	3	10	44.0	…	NaN	NaN
984	7/6/2020 16:37	3233	2	1	10	24.0	…	NaN	NaN
1001	7/10/2020 11:37	3225	2	3	11	28.0	…	NaN	NaN
1006	7/18/2020 15:15	3230	2	3	11	27.0	…	NaN	NaN
1011	7/13/2020 6:03	3204	2	3	11	42.0	…	NaN	NaN

	FinalOutcome	NumOfVisits	weight_combined	kenya_weight	nigeria_weight
1	1	1	1.416946	NaN	1.416946
3	1	1	1.416946	NaN	1.416946
25	1	1	1.416946	NaN	1.416946
43	1	1	1.416946	NaN	1.416946
44	1	1	1.416946	NaN	1.416946
…	…	…	…	…	…
983	1	1	1.416946	NaN	1.416946
984	1	1	1.416946	NaN	1.416946
1001	1	1	2.722043	NaN	2.722043
1006	1	1	1.416946	NaN	1.416946
1011	1	1	1.416946	NaN	1.416946

	southafrica_weight	agegroup	gk_weight
1	NaN	2	1.949579
3	NaN	2	2.325065
25	NaN	2	2.538366
43	NaN	2	1.949579
44	NaN	2	1.855697
…	…	…	…
983	NaN	2	1.823179

```
984                NaN         2  1.949579
1001               NaN         3  3.532132
1006               NaN         2  1.855697
1011               NaN         2  7.110619

[157 rows x 247 columns]
```

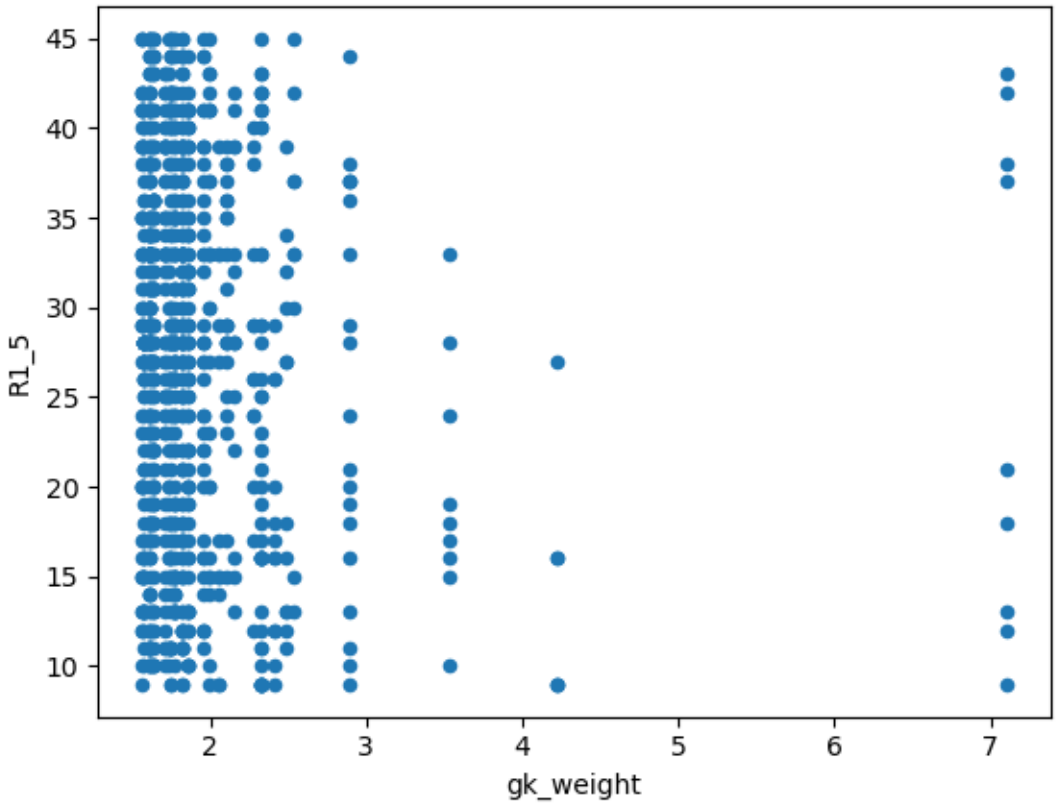
[11]:
```
# Zadanie 5 analiza zależności między kolumnami
# Oblicz macierz korelacji
df_selected = df[['gk_weight', 'R1_5']]

correlation_matrix = df_selected.corr()
print("Macierz korelacji:")
print(correlation_matrix)

# Wykonaj wykres rozrzutu
df.plot.scatter(x='gk_weight', y='R1_5')
```

```
Macierz korelacji:
           gk_weight      R1_5
gk_weight   1.000000 -0.100199
R1_5       -0.100199  1.000000
```

[11]: <Axes: xlabel='gk_weight', ylabel='R1_5'>



[12]:
```
# Zadanie 6 Przekształcanie danych
# Tworzę pomocniczą kolumne duration_minutes która przedstawi dane z␣
 ↪NetDuration w minutach
df['duration_minutes'] = pd.to_timedelta(df['NetDuration']).dt.total_seconds() /
 ↪ 60

# Dodaje nowa kolumne "wielkosc_na_kraj
df['wielkosc_na_kraj'] = df['LANG'] / df['Country']

# Grupuj dane wedug kolumny agegroup i oblicz sredni czas wywiadu
grouped = df.groupby('agegroup')['duration_minutes'].mean()
print("Średni czas wywiadu według grupy wiekowej:")
print(grouped)

# Posortuj dane wedug kolumny agegroup
df_sorted = df.sort_values(by='agegroup', ascending=False)
print("Dane posortowane według grupy wiekowej:")
print(df_sorted.head())
```

```
Średni czas wywiadu według grupy wiekowej:
agegroup
1    11.994770
2    11.925174
3     9.903370
Name: duration_minutes, dtype: float64
Dane posortowane według grupy wiekowej:
         SbjNum NetDuration InterviewTimeVStart InterviewTimeVEnd  \
2506  133196471     0:12:41      7/7/2020 17:22    7/7/2020 17:35
1677  133490122     0:07:41     7/18/2020 14:13   7/18/2020 14:44
1679  133490003     0:11:03     7/18/2020 14:39   7/18/2020 14:50
1680  133489273     0:08:35     7/18/2020 14:09   7/18/2020 14:18
217   133519409     0:13:16     7/19/2020 12:02   7/19/2020 12:17

                Date  Srvyr  Country  LANG  R1  R1_5  … FinalOutcome  \
2506  7/7/2020 10:22   3002        1     1   3   NaN  …            1
1677  7/18/2020 8:13   3030        3     1  50   NaN  …            1
1679  7/18/2020 8:39   3022        3     1  52   NaN  …            1
1680  7/18/2020 8:09   3038        3     1  47   NaN  …            1
217   7/19/2020 7:02   3225        2     3  11  28.0  …            1

     NumOfVisits  weight_combined  kenya_weight  nigeria_weight  \
2506           1         3.791351      3.791351             NaN
1677           1         1.000000           NaN             NaN
1679           1         1.000000           NaN             NaN
1680           1         1.000000           NaN             NaN
```

```
217            1         2.722043           NaN        2.722043

     southafrica_weight  agegroup gk_weight  duration_minutes wielkosc_na_kraj
2506                NaN         3  2.880918         12.683333         1.000000
1677                1.0         3  2.325065          7.683333         0.333333
1679                1.0         3  2.325065         11.050000         0.333333
1680                1.0         3  5.774278          8.583333         0.333333
217                 NaN         3  2.151458         13.266667         1.500000

[5 rows x 249 columns]
```

[ ]: