Palautettava tehtävä 3

June 1, 2025

1 Tehtävä 3: Ristiintaulukoinnit ja dummy-muuttuja

Tässä notebookissa: 1. Luetaan sama Excel-tiedosto (`Opinnäytetyökysely.xlsx`) pandas-DataFrameen, sheet "Kysely". 2. Luodaan ristiintaulukointi **Opiskeluala vs. Sukupuoli**. 3. Lasketaan sarake-normalisoitu prosenttijakauma – miesten ja naisten suhteellinen jakautuminen eri opiskelualoille (sisältää marginaalit). 4. Lasketaan rivinormalisoitu prosenttijakauma – sukupuolijakauma aloittain, ja lisätään kokonaismäärät ("Total count") rivinä. 5. Luodaan dummy-muuttujan **Oliko työ parityö?** taulukko, jossa on sekä määrät että prosenttiosuudet.

1.1 1) Kirjastojen tuonti ja Excel-tiedoston lukeminen

Tuodaan pandas ja pda.ace_tools. Luetaan sheet "Kysely" DataFrameen.
Varmista, että `Opinnäytetyökysely.xlsx` on saman hakemiston juuressa tai päivitä polku.

[16]:
```
import pandas as pd
import ace_tools as tools

# Vaihda tarvittaessa polku, jos Excel ei ole tässä kansiossa
file_path = r"D:\GitHub\PythonDataAnalytics\doc\Opinnäytetyökysely.xlsx"
df = pd.read_excel(file_path, sheet_name="Kysely", engine="openpyxl")

# Tarkistetaan, että DataFrame on ladattu oikein:
df.shape  # (rivit, sarakkeet)
```

[16]:
```
(242, 42)
```

1.2 2) Ristiintaulukointi 1: Opiskeluala vs. Sukupuoli

Lasketaan crosstab, jossa rivinä **Opiskeluala** ja sarakkeina **Sukupuoli**, näyttäen absoluuttiset lukumäärät.

[17]:
```
ct1 = pd.crosstab(df["Opiskeluala"], df["Sukupuoli"])
tools.display_dataframe_to_user("Ristiintaulukointi 1: Opiskeluala vs.␣
↪Sukupuoli", ct1)
```

```
=== Ristiintaulukointi 1: Opiskeluala vs. Sukupuoli ===
Sukupuoli    Mies  Nainen
Opiskeluala
```

```
Kulttuuri      14     110
Liiketalous    20      44
Tekniikka      52       2
```

1.3 3) Ristiintaulukointi 2: Sarake-normalisoitu prosenttijakauma

Lasketaan sama Opiskeluala vs. Sukupuoli, mutta normalisoidaan sarakkeittain (`normalize="columns"`) ja muunnetaan prosenttiluvuiksi.
Sisältää marginaalirivin ("All").

[18]:
```
ct2 = (
    pd.crosstab(
        df["Opiskeluala"],
        df["Sukupuoli"],
        margins=True,
        normalize="columns"
    )
    .mul(100)
    .round(2)
)
tools.display_dataframe_to_user(
    "Ristiintaulukointi 2: Miesten ja naisten suhteellinen jakautuminen (%)",
    ct2
)
```

```
=== Ristiintaulukointi 2: Miesten ja naisten suhteellinen jakautuminen (%) ===
Sukupuoli     Mies  Nainen    All
Opiskeluala
Kulttuuri    16.28   70.51  51.24
Liiketalous  23.26   28.21  26.45
Tekniikka    60.47    1.28  22.31
```

1.4 4) Ristiintaulukointi 3: Rivi-normalisoitu prosenttijakauma ja kokonaismäärät

Lasketaan Opiskeluala vs. Sukupuoli jälleen, mutta normalisoidaan rivittäin (`normalize="index"`) prosentteihin.
Lisätään uusi rivi "Total count", jossa on kunkin sukupuolen absoluuttinen määrä.

[19]:
```
ct3 = pd.crosstab(df["Opiskeluala"], df["Sukupuoli"], normalize="index").
 ↪mul(100).round(2)
totals = df["Sukupuoli"].value_counts()
ct3.loc["Total count"] = totals
tools.display_dataframe_to_user(
    "Ristiintaulukointi 3: Sukupuolijakauma aloittain (%) ja kokonaismäärä",
    ct3
```

```
)
```

```
=== Ristiintaulukointi 3: Sukupuolijakauma aloittain (%) ja kokonaismäärä ===
Sukupuoli     Mies  Nainen
Opiskeluala
Kulttuuri    11.29   88.71
Liiketalous  31.25   68.75
Tekniikka    96.30    3.70
Total count  86.00  156.00
```

1.5 5) Dummy-muuttuja: “Oliko työ parityö?”

Sarakkeen **Oliko työ parityö?** arvot ovat 1 = Kyllä ja 0 = Ei.
Lasketaan ensin lukumäärät, sitten prosenttiosuudet, ja kootaan ne kahden sarakkeen DataFrameen.

[20]:
```
dummy_counts = (
    df["Oliko työ parityö?"]
    .value_counts()
    .rename(index={1: "Kyllä", 0: "Ei"})
)
dummy_prop = (dummy_counts / dummy_counts.sum() * 100).round(2)
dummy_table = pd.DataFrame({"Count": dummy_counts, "Percentage": dummy_prop})

tools.display_dataframe_to_user("Dummy-muuttuja: Oliko työ parityö?",␣
 ↪dummy_table)
```

```
=== Dummy-muuttuja: Oliko työ parityö? ===
                    Count  Percentage
Oliko työ parityö?
Ei                    224       92.56
Kyllä                  18        7.44
```

2 Yhteenveto

Jupyter-notebookissa: - Tuodaan `Opinnäytetyökysely.xlsx` pandas-DataFrameksi (sheet: “Kysely”). - Luodaan ristiintaulukointi **Opiskeluala vs. Sukupuoli** (absoluuttiset lukumäärät). - Lasketaan sarake-normalisoitu prosenttijakauma – miesten ja naisten suhteellinen jakautuminen eri opiskelualoille (sisältää marginaalit). - Lasketaan rivinormalisoitu prosenttijakauma – sukupuolijakauma aloittain prosentteina, ja lisätään “Total count” -rivi kokonaissummien näyttämiseksi. - Luodaan dummy-muuttujan **Oliko työ parityö?** taulukko, jossa esitetään sekä määrät että prosenttiosuudet.