# Viikko 38 -tehtävät

## Tehtävä 1

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sb

import openpyxl as xl

from scipy.stats import chi2\_contingency, pearsonr, spearmanr

data\_df = pd.read\_excel("./work/viikko4/datasets/tt.xlsx")

original\_data\_df = data\_df.copy()

data\_df.info()

data\_desc = data\_df.describe()

data\_df.hist()

koulutus\_map = {1: 'Peruskoulu', 2: '2. Aste', 3: "Korkeakoulu", 4: "Ylempi korkeakoulu"}

sukupuoli\_map = {1: 'Mies', 2: 'Nainen'}

data\_df['koulutus'] = data\_df['koulutus'].map(koulutus\_map)

data\_df['sukup'] = data\_df['sukup'].map(sukupuoli\_map)

frekvenssi\_df = pd.crosstab(index=data\_df['koulutus'], columns='Lukumäärä')

frekvenssi\_df['%'] = (frekvenssi\_df['Lukumäärä'] / frekvenssi\_df['Lukumäärä'].sum()) \* 100

Kuva, joka sisältää kohteen teksti, kuvakaappaus, Fontti, numero

Kuvaus luotu automaattisesti

sb.barplot(data=frekvenssi\_df, x='Lukumäärä', y=frekvenssi\_df.index)

Kuva, joka sisältää kohteen teksti, kuvakaappaus, Suorakaide, viiva

Kuvaus luotu automaattisesti

## Tehtävä 2

frekvenssi\_df = pd.crosstab(index=data\_df['koulutus'], columns=data\_df['sukup'])

Kuva, joka sisältää kohteen teksti, kuvakaappaus, Fontti, numero

Kuvaus luotu automaattisesti

## Tehtävä 3

res = chi2\_contingency(frekvenssi\_df)

res.pvalue



Riippuvuus ei merkitsevä, koska p-arvo noin suuri

frekvenssi\_df.plot.barh(ylabel="Koulutus", xlabel="Lukumäärä")

Kuva, joka sisältää kohteen teksti, kuvakaappaus, viiva, numero

Kuvaus luotu automaattisesti

## Tehtävä 4

df\_corr = original\_data\_df[['sukup', 'ikä', 'perhe', 'koulutus', 'palkka']]

corr = df\_corr.corr()

sb.heatmap(corr, annot=True)

Kuva, joka sisältää kohteen teksti, kuvakaappaus, neliö, Suorakaide

Kuvaus luotu automaattisesti

pearsonr(df\_corr['ikä'], df\_corr['palkka'])



spearmanr(df\_corr['ikä'], df\_corr['palkka'])



sb.regplot(x=df\_corr['ikä'], y=df\_corr['palkka'])

Kuva, joka sisältää kohteen kuvakaappaus, viiva, diagrammi, Tontti

Kuvaus luotu automaattisesti

Tulosten perusteella iän ja palkan välillä on tilastollisesti merkittävä, mutta heikko positiivinen korrelaatio.