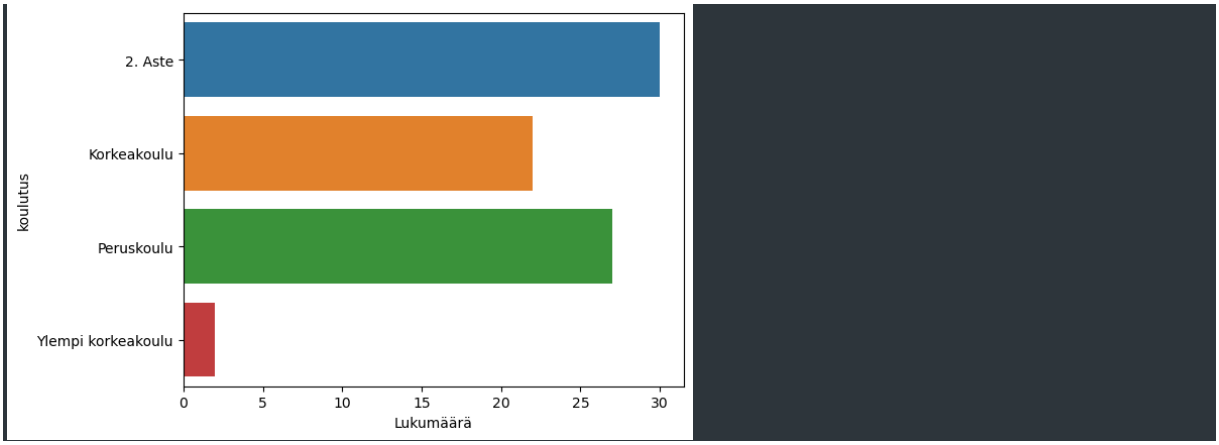
Viikko 38 -tehtävät

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
```

col_0	Lukumäärä	%
koulutus		
2. Aste	30	37.037037
Korkeakoulu	22	27.160494
Peruskoulu	27	33.333333
Ylempi korkeakoulu	2	2.469136

```
sb.barplot(data=frekvenssi_df, x='Lukumäärä', y=frekvenssi_df.index)
```



Tehtävä 2

```
frekvenssi_df = pd.crosstab(index=data_df['koulutus'], columns=data_df['sukup'])
```

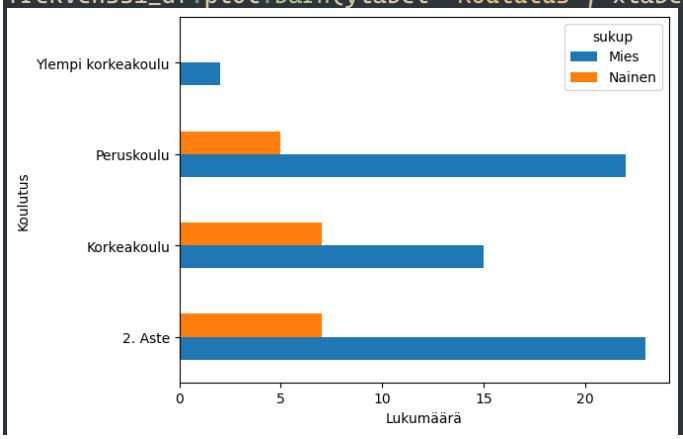
sukup	Mies	Nainen
koulutus		
2. Aste	23	7
Korkeakoulu	15	7
Peruskoulu	22	5
Ylempi korkeakoulu	2	0

Tehtävä 3

```
res = chi2_contingency(frekvenssi_df)

res.pvalue
```

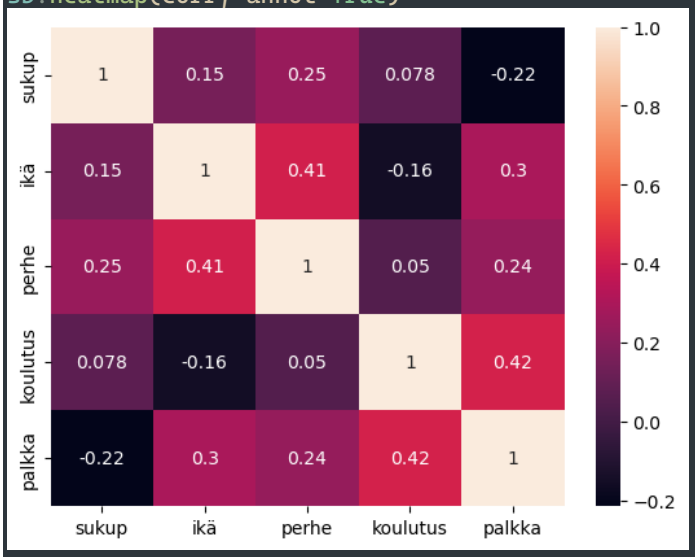
```
0.6070173075042058
```

```
Riippuvuus ei merkitsevä, koska p-arvo noin suuri


frekvenssi_df.plot.barh(ylabel="Koulutus", xlabel="Lukumäärä")
```



Tehtävä 4

```
df_corr = original_data_df[['sukup', 'ikä', 'perhe', 'koulutus', 'palkka']]

corr = df_corr.corr()

sb.heatmap(corr, annot=True)
```



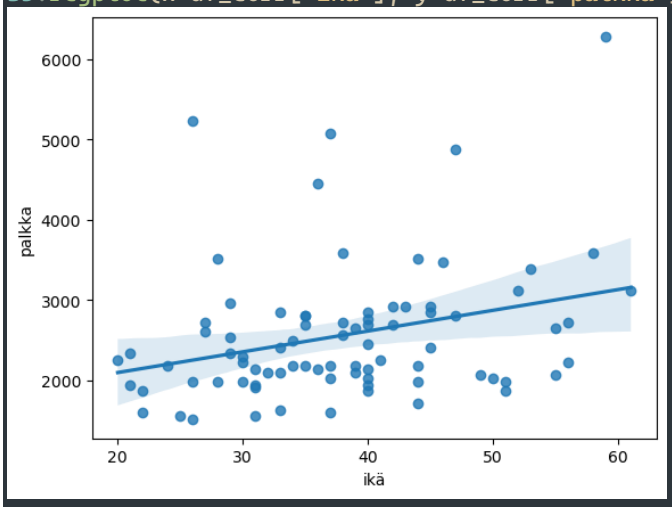

```
pearsonr(df_corr['ikä'], df_corr['palkka'])
```

```
PearsonRResult(statistic=0.2968719048961778, pvalue=0.0067615145801082485)
```

```
spearmanr(df_corr['ikä'], df_corr['palkka'])
```

```
SignificanceResult(statistic=0.3095587335779997, pvalue=0.004654685549528727)
```

```
sb.regplot(x=df_corr['ikä'], y=df_corr['palkka'])
```



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