

```

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as stats

df = pd.read_excel('tt.xlsx')

print(df.describe())
print(df.info())
df.hist()
plt.show()

print(df['palkka'].nlargest(5))
print(df['ikä'].nsmallest(5))

#t1
koulutus = ['Peruskoulu', '2. aste', 'Korkeakoulu', 'Ylempi korkeakoulu']
sukup = ['Mies', 'Nainen']

edu_df = pd.crosstab(index=df['koulutus'], columns='Lukumäärä')
edu_df.index = koulutus
edu_df.columns.name = ''
print(edu_df)

n = edu_df['Lukumäärä'].sum()
edu_df['%'] = (edu_df['Lukumäärä'] / n) * 100
print(edu_df.round(1))

sns.barplot(x='Lukumäärä', y=edu_df.index, data=edu_df)
plt.show()

#t2
gedu_df = pd.crosstab(index=df['koulutus'], columns=df['sukup'])
gedu_df.index = koulutus
gedu_df.columns = sukup

#t3
p = stats.chi2_contingency(gedu_df)[1]
if p>0.05:
    print(f'Riippuvuus ei ole tilastollisesti merkitsevä, p={p}')
else:
    print(f'Riippuvuus on tilastollisesti merkitsevä, p={p}')

# matplotlib

gedu_df.plot(kind='barh')
plt.show()

# seaborn

gedu_df.reset_index(inplace=True)
gedu_df.rename(columns={'index': 'Koulutus'}, inplace=True)

gedu_df = pd.melt(gedu_df, id_vars='Koulutus', var_name='Sukupuoli', value_name='Lukumäärä')

sns.barplot(x='Lukumäärä', y='Koulutus', data=gedu_df, hue='Sukupuoli')
plt.show()

#t4

df_corr = df[['sukup', 'ikä', 'perhe', 'koulutus', 'palkka']]
print(df_corr.corr())
sns.heatmap(df_corr.corr(), annot=True)
plt.show()

pearsonr = stats.pearsonr(df['ikä'], df['palkka'])
print(f'\npearson: {pearsonr}')

spearmanr = stats.spearmanr(df['ikä'], df['palkka'])
print(f'\nspearman: {spearmanr}')

```