2025

Oppimistehtävä 1

JAAKKO OJA

Tämä datasetti valikoitui Kagglesta ja on nimeltään Sleep Health and Digital Screen Exposure Dataset!

Se sisältää tietoja, jotka auttavat ymmärtämään unen laadun, stressitasojen ja digitaalisten näyttöjen käytön välistä suhdetta. Datasetti sisältää seuraavat sarakkeet:

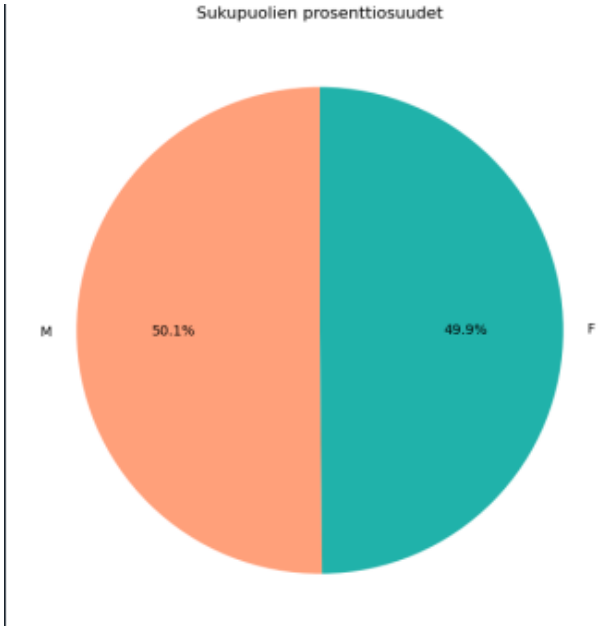
- Gender: Sukupuoli (M/F)
- Age: Ikä (vuosina)
- Sleep Duration: Keskimääräinen unen kesto (tunneissa)
- Sleep Quality: Itsearvioitu unen laatu (asteikolla 1-5)
- Stress Level: Stressitaso (asteikolla 1-5)
- Blood Pressure: Verenpaine (systolinen/diastolinen mmHg)
- Heart Rate: Leposyke (bpm)
- Daily Steps: Keskimääräinen päivittäinen askelmäärä
- Physical Activity: Fyysisen aktiivisuuden taso (kvantifioitu mitta)
- Height: Pituus (cm)
- Smoking: Tupakoiko henkilö (Y/N)
- Medical Issue: Onko henkilöllä olemassa olevia sairauksia (Y/N)
- Ongoing Medication: Käyttääkö henkilö lääkitystä (Y/N)
- Smart Device Before Bed: Käyttääkö henkilö digitaalisia laitteita ennen nukkumaanmenoa (Y/N)
- Average Screen Time: Keskimääräinen päivittäinen näyttöaika (tunneissa)
- Blue-Light Filter: Käyttääkö henkilö sinivalosuodatinta (Y/N)
- Discomfort/Eye-Strain: Kokee silmien epämukavuutta tai rasitusta (Y/N)
- Redness in Eye: Kokee silmien punoitusta (Y/N)
- Itchiness/Irritation in Eye: Kokee silmien kutinaa tai ärsytystä (Y/N)
- Dry Eye Disease: Kuivasilmäisyyden diagnoosi tai oireet (Y/N)

Datasetti sisältää useita rivejä, jotka edustavat yksilöiden terveystietoja, mutta tarkkaa rivimäärää ei ole mainittu, eikä tätä löytynyt etsinnästä huolimatta. Excel-tiedosto on kumminkin yli 1300 kilotavun kokoinen, eli kyllä useita rivejä dataa on!

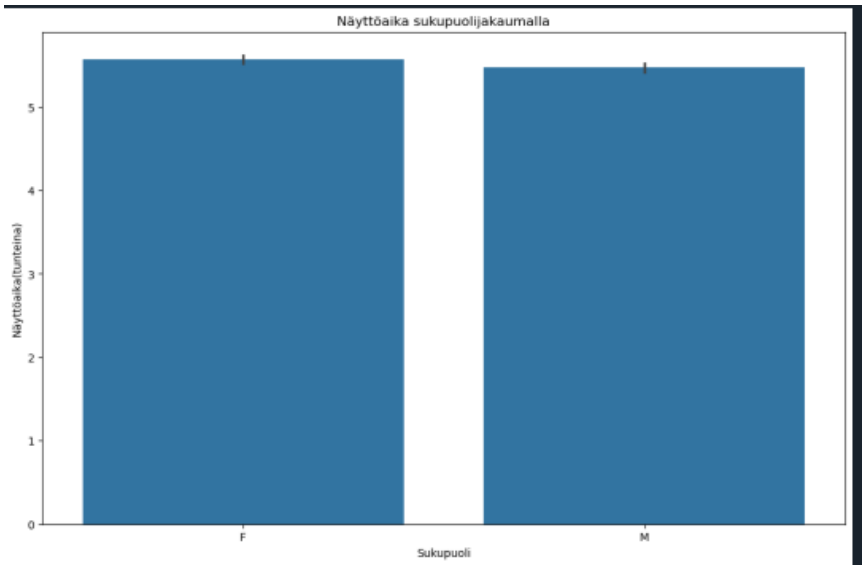
Työn tavoite: Omalla kohdalla tämän työn tavoite on oppia analysoimaan dataa paremmin ja oppia tuntemaan python-ohjelmointikieltä.

Oman työn arviointi: Arvioin omaa työtäni kriittisesti mutta selkeästi. Minusta työni täyttää 2-pisteen vaatimuksen sillä vaaditut asiat löytyvät. Kehityskohteiksi voisin luetella korrelaatio-heatmapin visualisoinnin, tätä en saanut toimimaan datasetin Y/N-arvojen takia yrityksestä huolimatta.

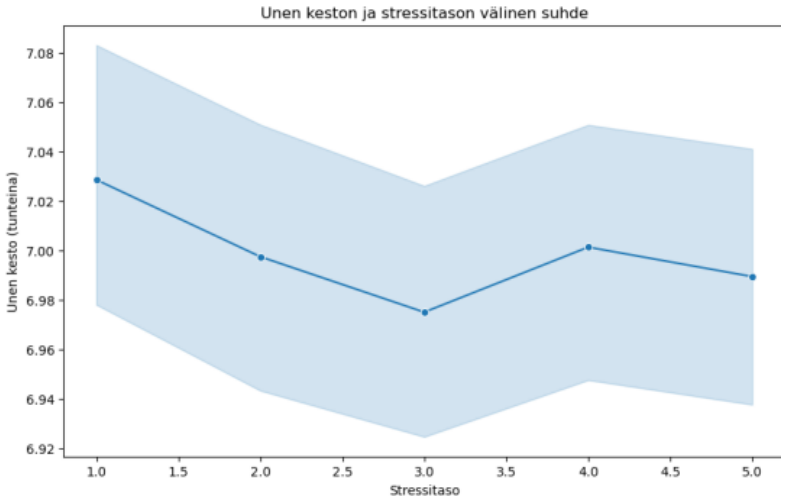
Datan visualisoinnit



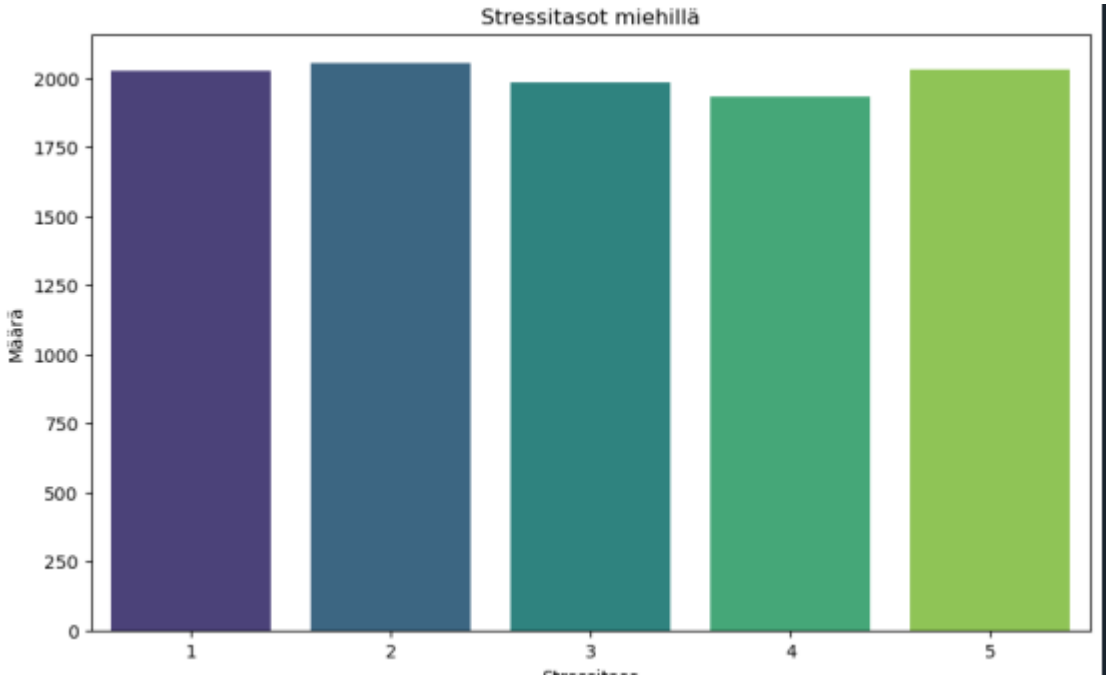
Datasetissä on tämän mukaan miesten ja naisten osuus aika tarkasti 50/50.



Näyttöaika sukupuolijakaumalla. Tässäkin hyvin tasaista, mutta huomaamme, että naiset käyttävät päivässä enemmän puhelinta kuin miehet.



Tässä on kuvattuna unen keston ja stressitason välinen suhde. Huomattakoon, että stressitaso on tässä välillä 1-5, numeron 1 ollessa LOW ja numeron 5 ollessa HIGH. Tässä datasetissä on havaittu, että henkilön mennessä nukkumaan hänellä on jo valmiiksi jonkinlainen stressitaso.



Tässä on visualisoituna miesten stressitaso. Kuvaajan alaosassa lukee ’Stressitaso’. Tästä voidaan hyvin päätellä, että stressitasoja 3-4 on hienoisesti vähemmän kuin tasoja 1-2. Huomataan myös että kovinta stressitasoa 5 on toiseksi eniten tässä datasetissä.

Datan tilastollista analyysiä:

```
##Tilastollinen analyysi
##Esimerkkinä tässä stressitason ja unen määrän yhteys Pearsonin korrelaatiokertoimella
corr, p_val = pearsonr(df["Stress level"], df["Sleep duration"])
print(f"Korrelaatiokerroin: {corr}, p-arvo: {p_val}")
```

```
Korrelaatiokerroin: -0.006088251100650671, p-arvo:
0.38925848131273727
```

Tässä datasetissä tulokset ovat vähintään erikoisia. Stressin ja unen määrän välinen korrelaatiokerroin on negatiivinen (-0.006) eli näiden kahden muuttujan välillä ei ole lineaarista yhteyttä. P-arvo on myös huomattavasti suurempi kuin normi 5% arvo, mistä voidaan todeta että tulos ei ole tilastollisesti merkitsevä. Loppukaneettina todettakoon, että stressillä ei tässä datasetissä ei näytä olevan merkittävää vaikutusta unen määrään. Vähän kyllä saa epäillä tätä datasettiä.

Tein jonkun ihmeellisen muutoksen tässä kohtaa, enkä saanut tätä korrelaatiokerrointa enään toimimaan. Otin siittä kumminkin näyttökuvan ennen muutoksia ja on tässä työssä nyt sitten esillä tämän vuoksi. Tämän vuoksi tein tästä t-testin!

T-testi unen laadun ja unen määrän välillä:

```
T-testi: -253.21212658830166, p-arvo: 0.0
Voimme hylätä nollahypoteesin: Unen laadun ja unen määrän välillä on
tilastollisesti merkitsevä ero.
```

Pientä pohdintaa tästä. Näiden kahden muuttujan välillä on merkittävä yhteys. Käytin tässä merkitsevyystasona 5% lukuarvoa. P-arvoksi tälle datasetille saatiin tasan 0, jonka perusteella voidaan hylätä nollahypoteesi.

<u>Otos datasta(kuvana 20 ensimmäistä riviä) ja python-koodi:</u>

	Gender	Age	...	Itchiness/Irritation in eye	Dry Eye Disease
0	F	24	...	N	Y
1	M	39	...	Y	Y
2	F	45	...	N	N
3	F	45	...	Y	N
4	F	42	...	N	Y
5	F	42	...	Y	Y
6	M	26	...	Y	Y
7	M	33	...	N	Y
8	M	36	...	Y	N
9	M	33	...	Y	Y
10	F	21	...	Y	Y
11	M	26	...	N	Y
12	M	42	...	Y	Y
13	F	28	...	N	Y
14	M	41	...	N	Y
15	F	37	...	N	Y
16	M	32	...	N	N
17	F	25	...	Y	Y
18	F	31	...	N	Y
19	F	23	...	N	Y
20	F	28	...	Y	Y

```
## Oppimistehtava 1
## Dataset: Sleep Health and Digital Screen Exposure Dataset(Kaggle)
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import ttest_ind

df = pd.read_csv("C:/Users/Jaakko/Desktop/Data-analyysi ja tekoälyn
perusteet/Oppimistehtava 1/Dry_Eye_Dataset.csv")
print(df.info())
##Yleisinfo datasta
print(df.head())
##Perustilastointia
print(df.describe())
print(df.columns)

##Datan esikäsittely
print(df.isnull().sum())
print(df['Stress level'].value_counts())
print(df['Sleep duration'].describe())
df = df.dropna()
# Otos datasta (50 ensimmäistä riviä)
df_sample = df.head(50)
print(df_sample)
```

```
## Halutaan muuttaa Y/N-arvot numeerisiksi arvoiksi
yn_columns = ['Smoking', 'Medical issue', 'Ongoing medication', 'Smart device before bed',
        'Blue-light filter', 'Discomfort Eye-strain', 'Redness in eye',
        'Itchiness/Irritation in eye', 'Dry Eye Disease']

for col in yn_columns:
   df[col] = df[col].map({'Y': 1, 'N': 0})

## Varmistetaan, että kaikki Y/N-arvot on muutettu oikein, sillä tässä datasetissä näitä Y/N-arvoja oli huiman paljon.
print(df[yn_columns].head())

## Muutetaan sukupuoli numeeriseksi arvoksi
df['Gender'] = df['Gender'].map({'M': 1, 'F': 0})
## Stressitasojen normalisointi, lisätään hieman satunnaisuutta
df["Stress level"] = df["Stress level"] + np.random.randint(-1, 2, df.shape[0])
df["Stress level"] = df["Stress level"].clip(1, 5)
## Unen keston korjaus, jos datasetissä on outoja arvoja
df.loc[df["Sleep duration"] < 3, "Sleep duration"] = df["Sleep duration"].median()
df.loc[df["Sleep duration"] > 14, "Sleep duration"] = df["Sleep duration"].median()

## Sukupuolten prosenttiosuudet
gender_counts = df['Gender'].value_counts()
plt.figure(figsize=(10, 8))
gender_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90, colors=['#FFA07A', '#20B2AA'])
plt.title('Sukupuolien prosenttiosuudet')
plt.ylabel('')
plt.show()
```

```
## Aloitetaan datan visualisointi perinteisellä stressin vaikutuksella unen määrään tunteina.
plt.figure(figsize=(10, 6))
sns.lineplot(x='Stress level', y='Sleep duration', data=df, marker='o')
plt.title('Unen keston ja stressitason välinen suhde')
plt.xlabel('Stressitaso')
plt.ylabel('Unen kesto (tunteina)')
plt.show()

## Päivittäinen näyttöaika sukupuolijakaumalla. Huomataan, ettei isoa eroa ole.
plt.figure(figsize=(13, 8))
sns.barplot(x='Gender', y='Average screen time', data=df)
plt.title('Näyttöaika sukupuolijakaumalla')
plt.xlabel('Sukupuoli')
plt.ylabel('Näyttöaika (tunteina)')
plt.show()

## Miesten stressitasojen analysointia
df_male = df[df['Gender'] == 1]
# Tehdään visualisointi stressitasoista miehillä
plt.figure(figsize=(10, 6))
sns.countplot(x='Stress level', data=df_male, palette='viridis')
plt.title('Stressitasot miehillä')
plt.xlabel('Stressitaso')
plt.ylabel('Määrä')
plt.show()

## Tilastollinen analyysi: T-testin käyttö unen laadun ja unen määrän välillä
```

```
# Tarkistetaan, että sarakkeet ovat numeerisia ja niissä ei ole puuttuvia arvoja
print(df["Sleep quality"].head())
print(df["Sleep duration"].head())

# Suoritetaan t-testi unen laadun ja unen määrän välillä käyttäen 5% merkitsevyystasoa
t_stat, p_val = ttest_ind(df["Sleep quality"], df["Sleep duration"])
alpha = 0.05

print(f"T-testi: {t_stat}, p-arvo: {p_val}")
if p_val < alpha:
    print("Voimme hylätä nollahypoteesin: Unen laadun ja unen määrän välillä on tilastollisesti merkitsevä ero.")
else:
    print("Emme voi hylätä nollahypoteesia: Unen laadun ja unen määrän välillä ei ole tilastollisesti merkitsevää eroa.")
```