Tampereen ammattikorkeakoulu

OPPIMISTEHTÄVÄ

Data-analyysi

Hanna Rantanen
Jenna Räty

DATA-ANALYYSI JA TEKOÄLYN PERUSTEET
Lokakuu 2022

Tietotekniikka
Tietoliikennetekniikka ja tietoverkot

1 DATAN KÄYTTÖTARKOITUS JA TYÖN TAVOITE

1.1 Kuvaus datasta

Oppimistehtävän datasetti löydettiin Kaggle-sivustolta. Kyseinen datasetti on nimeltään '*Suicidal Behaviours Among Adolescents*', ja se tarkastelee tekijöitä, jotka ennustavat nuorten itsetuhoista käyttäytymistä. Dataa on kerätty 27 maasta käyttäen Global School-based Student Health Survey-kyselylomaketta (GSHS).

Datasetin sarakkeet ovat seuraavat:

- Country
- Year
- Age Group
- Sex
- Currently_Drink_Alcohol
- Really_Get_Drunk
- Overwieght
- Use_Marijuana
- Have_Understanding_Parents
- Missed_classes_without_permssion
- Had_sexual_relation
- Smoke_cig_currently
- Had_fights
- Bullied
- Got_Seriously_injured
- No_close_friends
- Attempted_suicide

Datasetti on csv-tiedosto, jota voidaan tarkastella Excelissä. Rivejä tiedostossa on 107kpl ja sarakkeita 17kpl. Csv-tiedostoissa data on tallennettu siten, että jokaisella rivillä on peräkkäin datan arvot pilkulla erotettuna, vastaten sarakkeiden järjestystä. Country-sarakkeeseen on tallennettu maa (esim. 'Argentina'), Year-sarakkeessa on vuosiluku väliltä 2010–2018, Age Group-sarakkeessa on ikäryhmä väliltä 13–15 tai 16–17, sekä Sex-sarakkeessa sukupuoli ('Male' tai 'Female'). Lopuissa sarakkeista, jotka kuvaavat itsetuhoisiin yrityksiin vaikuttavia tekijöitä sekä itsetuhoisten yritysten esiintymistä, arvot on esitetty prosenttilukuina kerätystä aineistosta. Numeroarvot eivät siis esitä ilmiöiden esiintymisen absoluuttista määrää, vaan prosentuaalista osuutta kyseisessä maassa kyseisenä vuonna tietyssä ikäryhmässä ja sukupuolessa esiintyneistä arvoista.

1.2 Kuvaus datan esikäsittelystä

Ensiksi haluttiin tarkistaa, onko datasetissä NaN-arvoja. Tämä katsottiin isnull()-funktiolla. Tästä saatiin tieto, että aineistossa on 6 NaN-arvoa. Nämä arvot muutettiin nollaksi fillna()-funktiolla. Tämän jälkeen tarkistettiin jälleen isnull():n avulla, että NaN-arvoja ei enää ole. Myös info()-funktion avulla voitiin todeta, että NaN-arvoja ei aineistossa enää ole.

Describe()-funktiolla aineistoa tarkasteltaessa havaittiin, että Currently_Drink_Alcohol-sarakkeessa on virheellinen prosenttiluku 548 %. Tässä on varmaankin tarkoitus olla arvo 54.8 %. Mikäli virheellinen arvo olisi jätetty aineistoon, se olisi vääristänyt aineistosta tehtäviä päätelmiä ja tuloksia. Arvo korjattiin loc-toiminnon avulla.

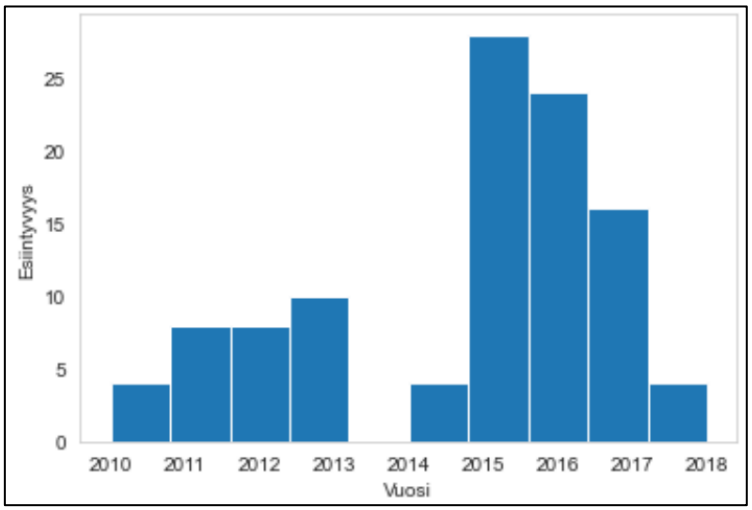
Datasta tarkasteltiin itsetuhoisten yritysten määrää käyttäen nlargest()- ja nsmallest()-funktioita. Huomattiin, että prosentuaalisesti eniten itsetuhoisia yrityksiä oli Samoalla vuonna 2011, 13–15-vuotiaiden poikien keskuudessa (67,2 %). Prosentuaalisesti vähiten itsetuhoisia yrityksiä oli Indonesiassa vuonna 2015, 16–17-vuotiaiden tyttöjen keskuudessa (2.7 %). Mean()-funktion avulla laskettiin keskiarvo itsetuhoisten yritysten sarakkeesta, tuloksena oli 14.5 %.

Unique()-funktion avulla voitiin tarkastella, mistä eri maista aineisto on kerätty. Funktio tulostaa useasti esiintyvien arvojen joukosta niiden uniikit arvot. Tässä tarkastelussa aineiston maiden todettiin olevan 'Argentina', 'Barabados', 'Benin', 'Bhutan', 'Brunei Darussalam', 'Dominican Republic', 'Fiji Islands', 'Indonesia', 'Jamaica', 'Kiribati', 'Laos', 'Malaysia', 'Mauritus', 'Mongolia', 'Namibia', 'Nepal', 'Peru', 'Samoa', 'Seychelles', 'Suriname', 'Thailand', 'Timor-Leste', 'Trinidad and Tobago', 'Tuvalu', 'Uruguay', 'Vanuatu' ja 'Wallis and Futuna'. Huomataan, että aineiston maat sijaitsevat maapallon eteläisellä puoliskolla sekä päiväntasaajan tuntumassa.

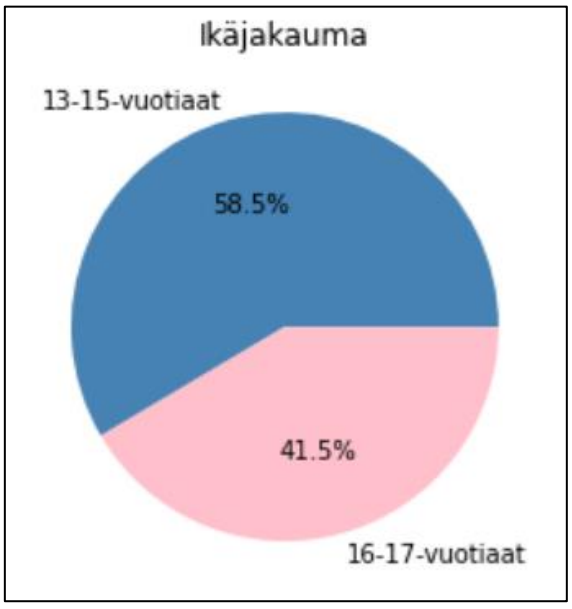
2 DATAN VISUALISOINTI JA ANALYSOINTI

2.1 Datan visualisointi

Havainnollistettiin aineiston vuosilukuja histogrammin avulla, joka esitelty alla. Kuvasta voidaan nähdä, että aineistoa on kerätty eniten vuosina 2015 ja 2016, sekä vähiten vuosina 2010, 2014 ja 2018.

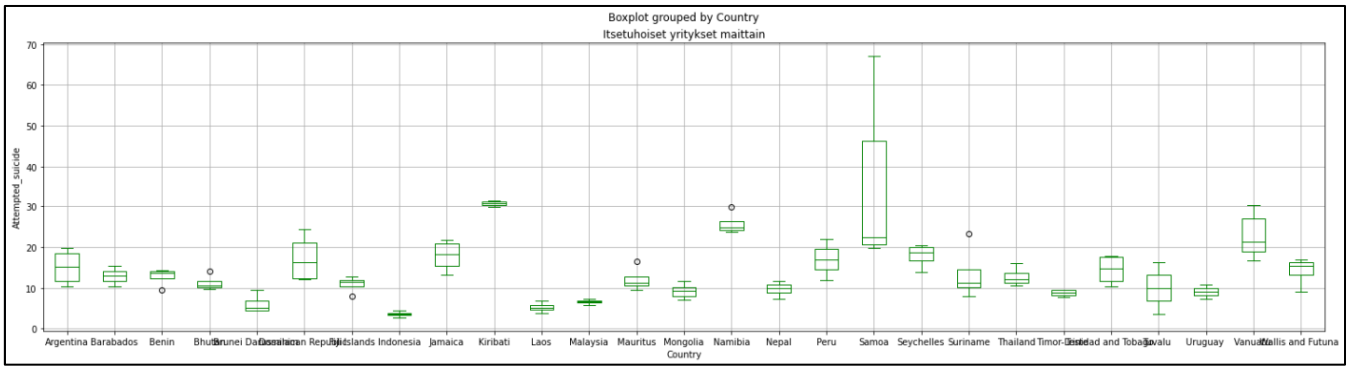


Alla olevassa kuvassa on visualisoitu aineiston ikäjakauma: 13–15-vuotiaita on hieman yli puolet aineistosta. Tämä on ymmärrettävää, sillä kyseiseen ikähaarukkaan kuuluu 13-, 14- ja 15-vuotiaita, kun vanhempien nuorten ikähaarukkaan kuuluu vain 16- ja 17-vuotiaita.



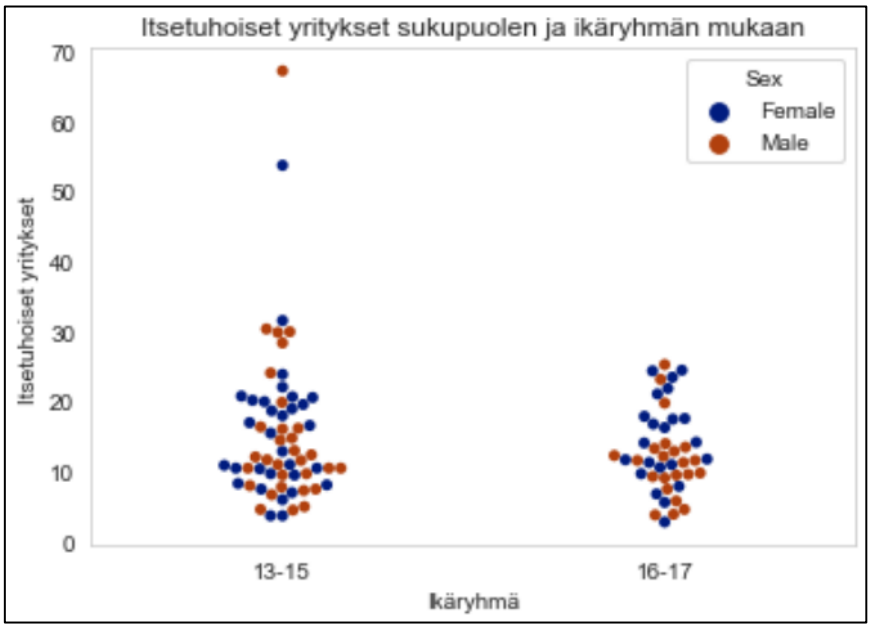
Alla olevassa kuvassa on visualisoitu itsetuhoisten yritysten prosentuaalista määrää vuosittain. Kuvasta voidaan nähdä, että esimerkiksi vuonna 2011 oli pari poikkeavan suurta prosenttilukua itsetuhoisten yritysten määrässä. Taas vuonna 2014 yritysten prosentuaalinen määrä oli muita vuosia pienempi.

Alla olevassa kuvassa on visualisoitu itsetuhoisten yritysten prosentuaalista määrää maittain. Kuvasta voidaan nähdä, että Indonesiassa itsetuhoisten yritysten prosentuaalinen määrä on muita matalampi, kun taas Samoassa määrä on suurempi. Aiemmin pienintä ja suurinta prosenttilukua itsetuhoisten yritysten määrässä tarkasteltaessa todettiin, että pienin prosenttilukema esiintyi Indonesiassa ja suurin Samoassa: tästä voidaan siis päätellä, että kyseisissä maissa on todellisuudessakin pienimmät ja suurimmat arvot.

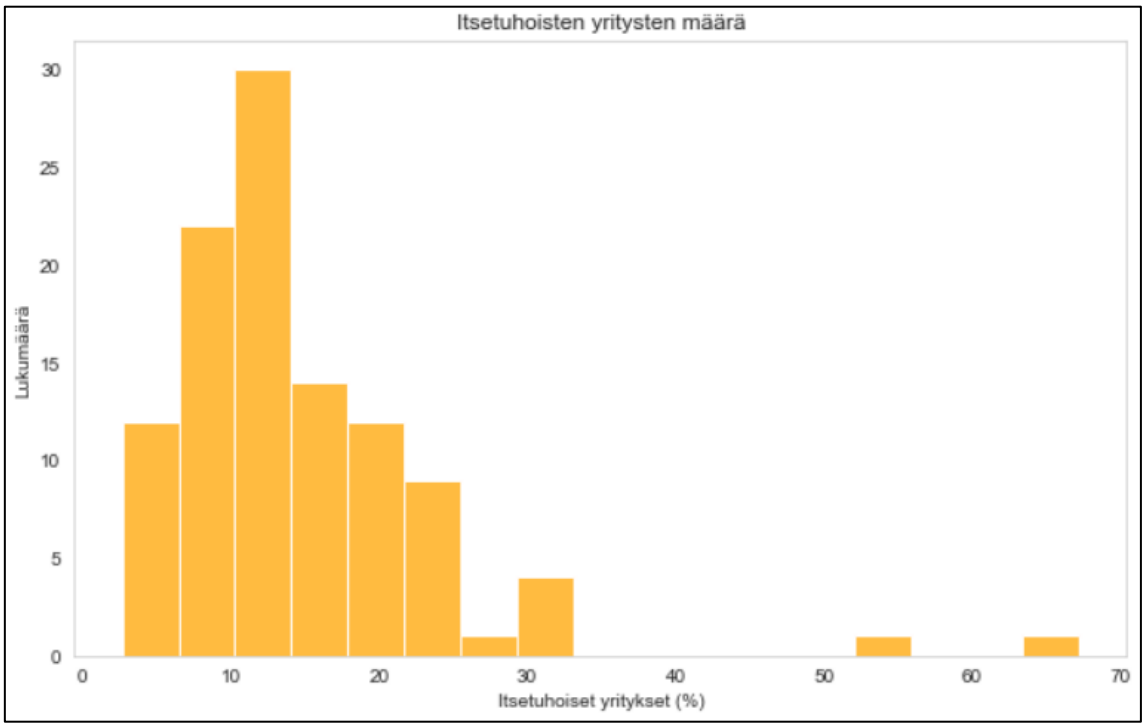


Alla olevassa kuvassa on otettu seabornin swarmplot aineistosta. Tässä tarkastellaan itsetuhoisten yritysten prosentuaalista määrää suhteessa ikäryhmiin sekä sukupuoliin. Kuvaa tarkasteltaessa huomataan, että suurimmat prosenttilukemat itsetuhoisissa yrityksissä esiintyvät 13–15-vuotiaiden ikäryhmässä. Nämä poikkeavan suuret prosenttiluvut kuuluvat tytölle ja pojalle. Muuten sukupuolten välisistä eroista kuvassa on nähtävissä se, että alhaisempien prosenttilukujen kohdalla vaikuttaisi olevan enemmän poikia kuin tyttöjä, ja itsetuhoisten yritysten prosentuaalisen määrän

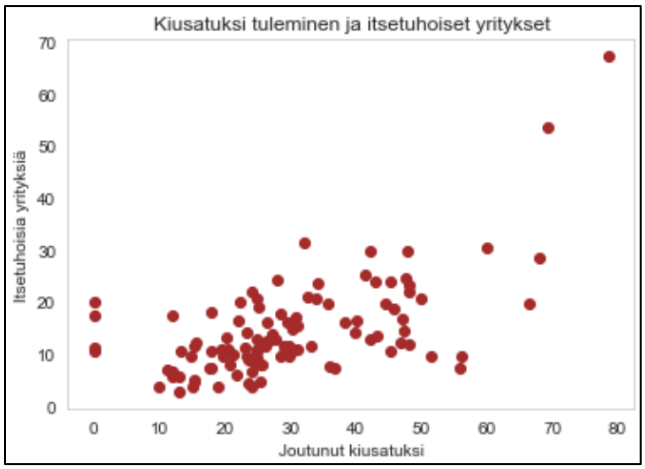
lisääntyessä myös tyttöjen osuus datapisteistä lisääntyy. Selkeästi kuvasta näkyy myös se, että itsetuhoisten yritysten datapisteitä on ylipäätään enemmän 13–15-vuotiaiden keskuudessa kuin 16–17-vuotiaiden ikäryhmässä.



Alla olevassa kuvassa on visualisoitu itsetuhoisten yritysten prosentuaalista määrää. Kuvasta nähdään, että itsetuhoisten yritysten määrästä suurin osa sijoittuu välille 5 % - 15 % aineiston kohteista. Dataa aiemmin tarkasteltaessa laskettiin keskiarvoksi tälle sarakkeelle 14.5 %, mikä täsmää tässä visualisoituun tulokseen.

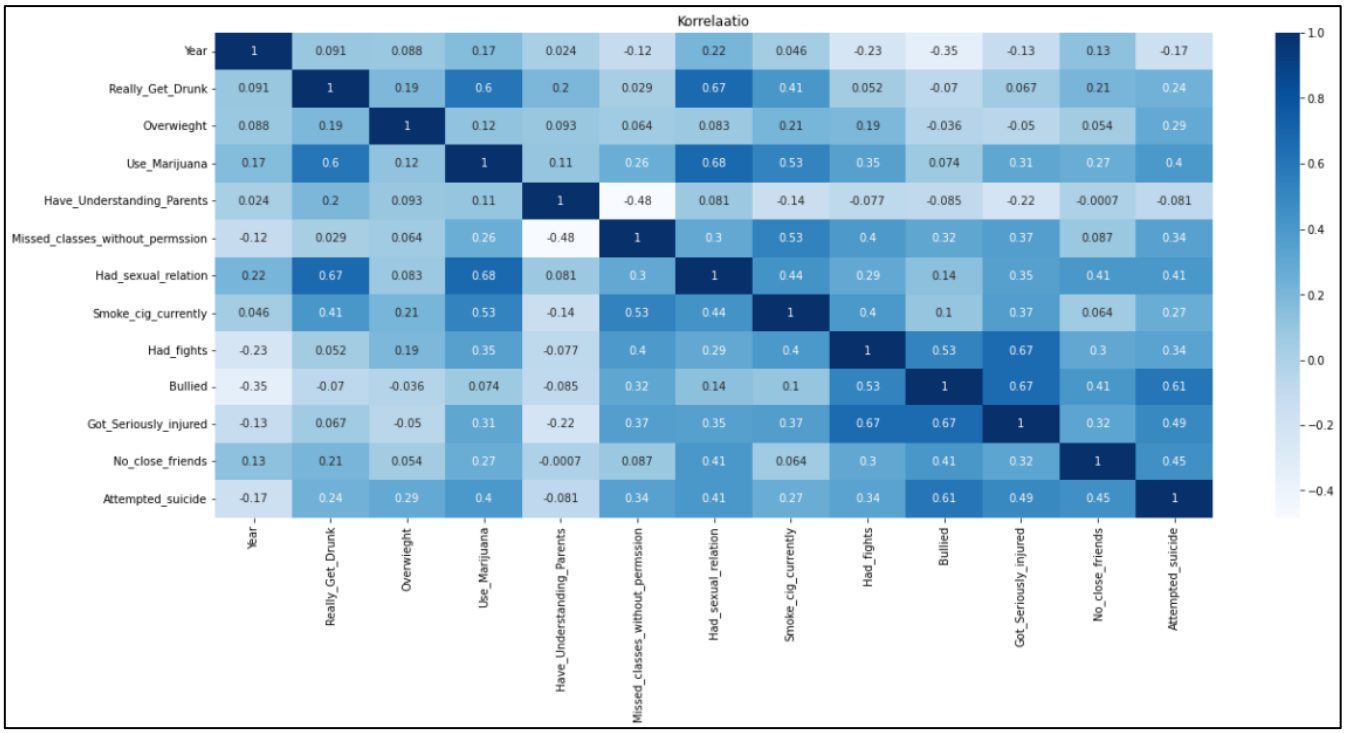


Alla olevissa kuvissa on tarkasteltu työn tekijöiden hypoteesia siitä, mitkä tekijät vaikuttavat eniten nuorten itsetuhoisten yritysten esiintymiseen. Näiden tekijöiden epäiltiin olevan ilman läheisiä ystäviä oleminen sekä kiusatuksi tuleminen. Lineaarisuutta on nähtävissä näistä kahdesta hieman selkeämmin kiusatuksi tulemisen ja itsetuhoisten yritysten kuvaajan välillä.

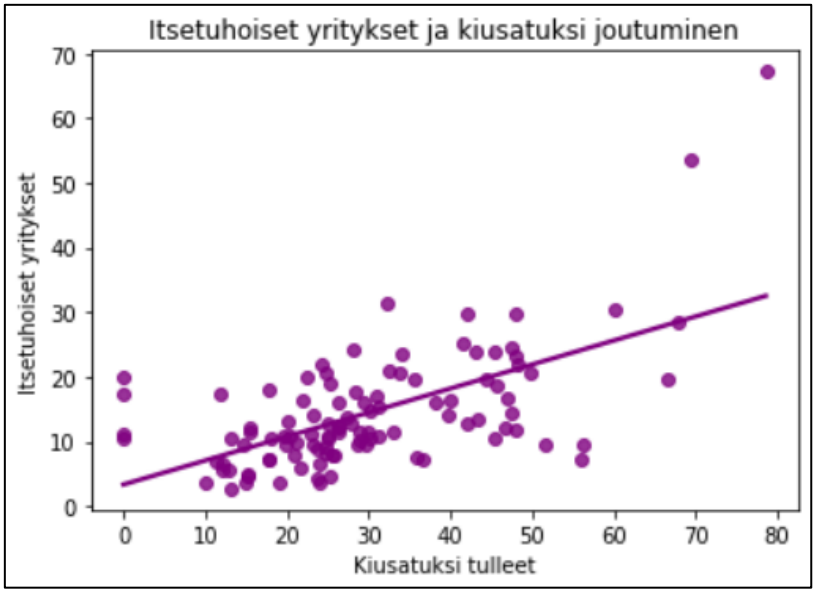


2.2 Datan analysointi

Alla olevasta korrelaatiokuvasta voidaan nähdä, että suurimmat korrelaatioluvut ovat luokkaa 0.67–0.68. Nämä korrelaatiot ovat seksisuhteiden ja marihuanan käytön välillä sekä esimerkiksi tappeluissa olleiden ja vakavasti loukkaantuneiden välillä. Itsetuhoisten yritysten sarakkeen kanssa korreloi vahvimmin kiusatuksi tuleminen (korrelaatio 0.61). Heatmapista nähdään myös, että ymmärtäväiset vanhemmat sekä itsetuhoiset yritykset korreloivat negatiivisesti.



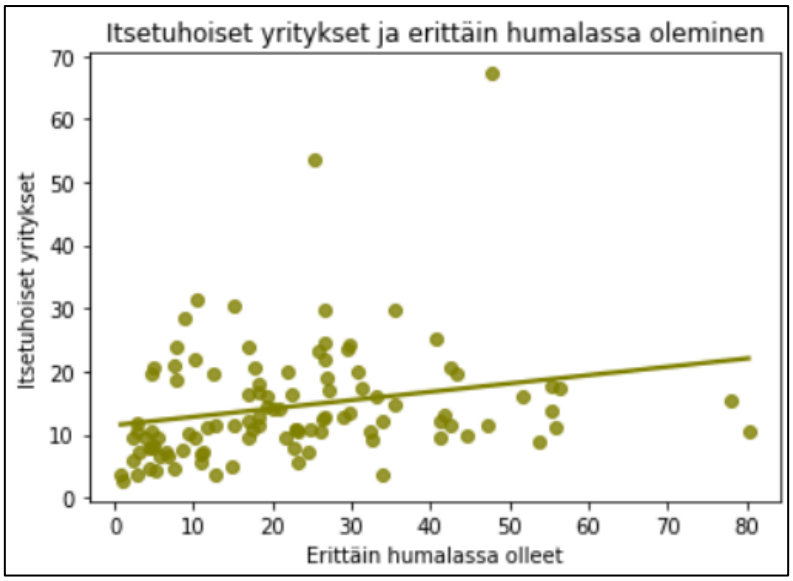
Tarkasteltiin siis kiusatuksi tulemisen ja itsetuhoisten yritysten välistä suhdetta, sillä tässä todettiin edellisessä vaiheessa olleen merkitsevä korrelaatio (0.61). Sovitettiin regressiosuora pistepilveen seabornin regplot-rutiinilla, mikä näkyy alla olevassa kuvassa.



Korrelaatiota tarkasteltiin khiin neliö-testin, pearsonr-rutiinin ja spearmanin avulla. Tuloksista nähdään, että p-arvo on hyvin pieni, mikä kertoo korrelaation olevan tilastollisesti merkitsevä (alle 0,05).

```
Riippuvuus on tilastollisesti merkitsevä, p=2.030917266150951e-21
Korrelaatio (pearsonr): (0.6054239272785555, 6.2082104896602445e-12)
Korrelaatio (spearman): SpearmanrResult(correlation=0.537943303060889, pvalue=2.7362027705158876e-09)
```

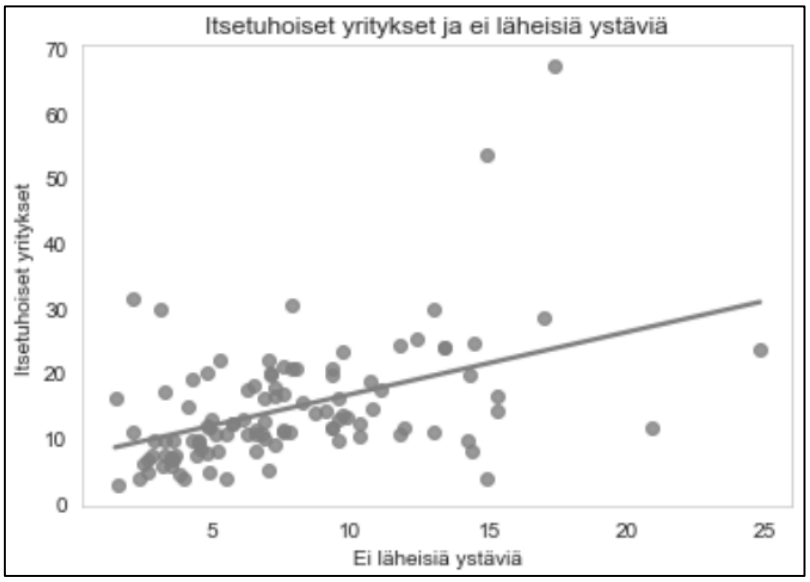
Tarkasteltiin myös vähemmän korreloivien tekijöiden regressiosuoraa, erittäin humalassa oleminen sekä itsetuhoiset yritykset. Alla olevasta kuvasta nähdään, että suora ei nouse yhtä jyrkästi, mikä kertoo pienemmästä korrelaatiosta (0.24).



Kun tarkasteltiin korrelaatiota, voitiin todeta, että myös tässä phiin neliö-testin mukaan riippuvuus on tilastollisesti merkitsevä. Huomataan kuitenkin, että sekä pearsonr-rutiinin että spearmanin mukaan korrelaatio on merkittävästi heikompi, ja myös p-arvo on suurempi vaikkakin vielä merkitsevä.

```
Riippuvuus on tilastollisesti merkitsevä, p=1.8748339119460856e-55
Korrelaatio (pearsonr): (0.23564554514645403, 0.015028094479063784)
Korrelaatio (spearman): SpearmanrResult(correlation=0.3803703708532896, pvalue=5.77103925520266e-05)
```

Seuraavaksi tarkasteltiin vielä toisena hypoteesina olleen, ilman läheisiä ystäviä olemisen ja itsetuhoisten yritysten, yhteyttä. Pistepilvisovitus on esitetty alla olevassa kuvassa.



Korrelaatiota tarkasteltaessa todettiin riippuvuuden olevan tilastollisesti merkitsevä. Sekä pearsonr-rutiinin että spearmanin mukaan korrelaatio oli suurempi kuin erittäin humalassa olemisen ja itsetuhoisten yritysten välillä, samoin p-arvot olivat pienemmät ja siten merkitsevämmät.

```
Riippuvuus on tilastollisesti merkitsevä, p=0.00048022545087301306
Korrelaatio (pearsonr): (0.4525151520708665, 1.1160512647704182e-06)
Korrelaatio (spearman): SpearmanrResult(correlation=0.49767046957587185, pvalue=5.712459846353397e-08)
```

Lopputuloksena voidaan todeta, että korrelaatio kiusatuksi tulemisen ja itsetuhoisten yritysten välillä oli aineiston tekijöiden välillä suurin. Tämä voidaan havaita myös tarkastelemalla yllä esitettyjen regressiosuorien jyrkkyyttä: mitä jyrkempi suora, sitä suurempi korrelaatio. Lisäksi tarkastellut numeroarvot p-arvojen ja korrelaatiolukujen osalta tukevat tätä päätelmää.

3 TYÖN ARVIOINTI

Tekijöiden mielestä tässä oppimistehtävässä on käytetty laajasti kurssin data-analyysiosuudessa opittuja tekniikoita ja menetelmiä. Kurssin harjoitustehtävien lisäksi työssä käytettiin lähteenä internetiä. Dataa on pyritty visualisoimaan sekä analysoimaan erilaisin keinoin, jotta sen ominaisuudet sekä siitä tulkittavat asiat on saatu monipuolisesti esitettyä. Dataa on tarkasteltu kuvien lisäksi myös erilaisten numeeristen metriikoiden avulla, esimerkiksi korrelaatiolukujen ja p-arvojen kautta.

Kehityskohteena tämän työn myötä voisi olla visualisointitekniikoiden monipuolisempi käyttö. Vaikka työssä käytettiin useita visualisointitapoja, tekijät ovat tietoisia, että niitä olisi olemassa vielä lukuisia muitakin.

Onnistumisia tämän työn osalta oli datan tarkoituksenmukainen visualisointi sekä oletettujen lopputulosten saavuttaminen datan analysoinnin sekä numeeristen metriikoiden tarkastelun myötä. Myös datan kuvaus ja esikäsittely oli kattava ja perusteellinen. Oppimistehtävän raportti on huolellisesti ja johdonmukaisesti laadittu, ja se etenee selkeästi havainnollistavien kuvien ja sanallisen kuvauksen myötä. Työn tekijöiden mielestä valittu datasetti oli synkästä aiheestaan huolimatta mielenkiintoinen ja sopi oppimistehtävän aineistoksi hyvin. Oppimistehtävä koettiin tekijöiden toimesta onnistuneeksi ja se sisältää 2 pisteen työlle vaadittavat ominaisuudet.

4 LIITTEET

Liite 1. Otos datasta

Index	Country	Year	Age Group	Sex	Currently_Drink_Alcohol	Really_Get_Drunk	Overwieght	Use_Marijuana	Have_Understanding_Parents	Missed_classes_without_permssion	Had_sexual_relation	Smoke_cig_curren	Had_fights	Bullied	Got_Seriously_injured	No_close_friends	Attempted_suicide
0	Argentina	2018	13-15	Female	50.3	30.7	27.8	7.9	41.5	24.7	25.7	16.8	17.2	0	27.5	4.8	19.9
1	Argentina	2018	13-15	Male	44.9	26.1	39.1	8.4	44.5	27.9	38.4	12.1	33.2	0	37.4	5.5	10.4
2	Argentina	2018	16-17	Female	67.2	56.3	22.5	21.9	37.1	34	59.1	28.5	15.1	0	30.1	6.3	17.4
3	Argentina	2018	16-17	Male	68.1	55.8	27.9	27	39.8	39.4	68.6	28	33.6	0	40.3	6.6	11.2
4	Argentina	2012	13-15	Male	49.3	28.9	35.9	10.6	46.2	32	43.5	17	44.2	42.1	24.8	6.1	12.9
5	Argentina	2012	13-15	Female	50.7	26.8	21.8	6.5	49.9	29.4	30.7	20.5	24.7	25.2	24.2	4.3	18.9
6	Barabados	2011	13-15	Male	29.4	80.2	32.3	15.6	48.1	14.6	32.9	0	42.7	30.3	49.1	6.8	10.4
7	Barabados	2011	13-15	Female	26.6	78	39.1	7.4	44.8	11.7	13.5	0	29.7	31.1	41.1	8.3	15.4
8	Benin	2016	13-15	Male	38	19.3	12.7	0.8	35.7	18.8	31.8	5.1	32.1	47.4	48.4	10.8	14.4
9	Benin	2016	13-15	Female	42.4	16.8	18.3	0.2	39.7	6.5	14.2	1.3	27.3	51.5	40.4	14.2	9.6
10	Benin	2016	16-17	Male	46.5	29.6	3.3	2.1	33.3	17.3	35.7	6.9	28.8	43.3	51	9.7	13.4
11	Benin	2016	16-17	Female	36.1	20	14.7	0.6	34.1	20.6	40.1	2	17.8	39.8	39.5	15.3	14.1
12	Bhutan	2016	13-15	Male	26	22.8	9.7	21.7	38.7	28.4	18.4	32.5	51.7	31.2	50.8	7.8	10.9
13	Bhutan	2016	13-15	Female	11.7	9.3	15.7	3.6	45.9	19.3	7.8	9.6	34.3	28.9	38.9	10.3	10.3
14	Bhutan	2016	16-17	Male	41.2	44.6	5.9	35.5	41	31.6	27	43.8	46.6	21.2	46.7	6.9	9.7
15	Bhutan	2016	16-17	Female	21.4	20.9	13.2	7.4	48.1	20.6	10	17.1	29.6	23.2	39.1	9.1	14
16	Brunei Darussalam	2014	13-15	Male	4.4	4.5	37.4	0.6	31.4	38.5	11.5	13.9	31.9	25.3	37.1	2.7	4.5
17	Brunei Darussalam	2014	13-15	Female	3.2	2.3	34.8	0	25.7	35.5	8.5	4.3	17.1	21.7	22.8	2.5	5.9
18	Brunei Darussalam	2014	16-17	Male	9.4	7.5	32.5	0.5	30.8	39.6	16.7	26.5	27.2	15.3	38.9	4.9	4.5
19	Brunei Darussalam	2014	16-17	Female	2.1	2.4	33.6	1	25.4	36.6	11.6	6.3	13.1	14.8	21.1	2.9	9.6
20	Dominican Republic	2016	13-15	Male	35.8	26.3	31.7	4.5	48.8	25.6	47	8.1	32.4	26.3	38.5	6.9	12.3
21	Dominican Republic	2016	13-15	Female	40.4	21.9	33	4.9	43.8	23	16.9	6.7	19.2	22.3	30.3	7.1	20.1
22	Dominican Republic	2016	16-17	Male	49.3	34	17.1	4	43.6	28.3	59.6	9.7	28.7	26.4	35.3	5.7	12.1
23	Dominican Republic	2016	16-17	Female	50.2	29.8	27.8	3.1	42.8	24.6	26.6	9	14.6	28.1	24.6	11.8	24.3
24	Fiji Islands	2016	13-15	Male	15.5	12.7	27.6	7.2	43.7	25.1	22	11.9	43.1	33.1	55.7	9.3	11.6
25	Fiji Islands	2016	13-15	Female	9.3	4.5	27.4	2.5	53.7	20.6	9.9	4.3	24.2	25.7	38.7	6.6	8
26	Fiji Islands	2016	16-17	Male	27.5	26.5	22.1	13.7	40.4	27.4	32.4	22.3	39.3	25	54.2	9.6	12.8
27	Fiji Islands	2016	16-17	Female	14.8	11.8	36.9	4.2	50.6	20.4	12.4	10	24.7	23	37.2	7.6	11.2
28	Indonesia	2015	13-15	Male	6.1	5.1	15.7	2.1	31.8	23.5	6.9	21.5	35.9	23.7	39.6	3.8	4.4
29	Indonesia	2015	13-15	Female	1.4	0.8	14.9	0.7	36.2	16.9	4	1.5	13.1	19	21.2	2.4	3.6
30	Indonesia	2015	16-17	Male	12.1	12.7	12.6	2.9	30.2	26.1	7.4	33	23.5	24	34.8	4	3.7
31	Indonesia	2015	16-17	Female	2.2	1.1	13.1	0.2	38.8	15.8	2.5	1.5	7	13	18.4	1.6	2.7
32	Jamaica	2017	13-15	Male	54.8	33.2	22.8	20.9	34.1	30.2	60.6	17.9	44.3	26.3	45.4	9.6	16.1
33	Jamaica	2017	13-15	Female	36.1	17.8	28.2	13.3	31.6	19.7	15.5	8.7	25.4	24.8	34	8	20.6
34	Jamaica	2017	16-17	Male	63.3	41.7	16.6	37.8	32.4	32.8	73.9	20.7	34.6	20.1	45	9.9	13.2
35	Jamaica	2017	16-17	Female	44.9	26.5	24.1	19.4	26.2	32.1	45	14	19.7	24.2	32.5	7	21.8
36	Kiribati	2011	13-15	Male	43.7	35.4	31.9	6.8	14.8	38.1	37.2	34.3	43.3	42.1	64.2	3.1	29.8
37	Kiribati	2011	13-15	Female	19.3	10.4	46.4	1.6	15.4	30.1	9.8	19.5	28.5	32.2	53.2	2.1	31.5
38	Laos	2015	13-15	Male	18.9	14.8	11.3	0.6	18.5	39.9	12.5	5.7	12	15.2	20	7	4.9
39	Laos	2015	13-15	Female	20.8	10.9	11.8	0.1	16.5	36	6.9	1.6	8.7	11.3	16.9	3.5	6.9
40	Laos	2015	16-17	Male	36.1	34	10.3	1.2	15.5	45.4	24.7	12.1	6.4	9.9	16.1	5.5	3.8
41	Laos	2015	16-17	Female	35.7	23.2	11	0.3	16.7	39.1	10.7	1.2	3.5	12	15.2	3.5	5.5
42	Malaysia	2012	13-15	Male	9.3	6.8	25.3	1.4	32.6	30	9.5	17.1	38.5	24	43.7	3.6	6.6
43	Malaysia	2012	13-15	Female	5.7	3	22.2	0.4	32.3	25.4	6.9	1.9	21.9	17.8	28.4	2.8	7.4
44	Malaysia	2012	16-17	Male	13.6	11	11.9	1.5	31.4	36.9	9.7	27	29.1	12.9	39.7	3.2	5.7
45	Malaysia	2012	16-17	Female	7.4	5.8	11.2	0.3	29.6	34.1	7.5	2.5	16.4	12	26.4	2.7	6.7
46	Mauritus	2017	13-15	Male	20.2	18.2	28.2	8.5	39.7	25.8	22.5	19.6	42.7	29	46.9	11.9	11.5
47	Mauritus	2017	13-15	Female	21.2	17	21.8	2.1	39.3	17.2	10.2	11.5	22.6	22	31.6	7.3	16.5

Liite 2. Python-koodi

```
"""
Data-analyysi ja tekoälyn perusteet
Oppimistehtävä 1
Hanna Rantanen & Jenna Räty
201227
"""

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as stats
from scipy.stats.stats import pearsonr
from scipy.stats import chi2_contingency

#Luetaan aineisto dataframeen
df = pd.read_csv('GHSH_Pooled_Data1.csv')

###### DATAN ESIKÄSITTELY ##############

#Tarkistetaan onko NaN-arvoja
print (df.isnull().values.any())

#Katsotaan NaN-arvojen lukumäärä
print (df.isnull().sum().sum())

#Korvataan NaN-arvot nollalla
df = df.fillna(0)

#Tarkistetaan onko NaN-arvoja
print (df.isnull().values.any())

#Katsotaan NaN-arvojen lukumäärä
print (df.isnull().sum().sum())

#info()-funktiolla voidaan vielä tarkastella datan rivimäärää, onko null-arvoja sekä datatyyppiä
print (df.info())

#Tarkastellaan Currently_Drink_Alcohol-sarakkeen arvoja
print (df['Currently_Drink_Alcohol'].describe())

# --> huomataan, että sarakkeessa virheellinen prosenttilukema 548% --> oikea lukema varmaankin 54.8%

#Korvataan virheellinen prosenttilukema
df.loc[df['Currently_Drink_Alcohol'] == 548, 'Currently_Drink_Alcohol'] = '54.8'
```

```
#Tarkastellaan itsetuhoisten yritysten määrää
print (df['Attempted_suicide'].nlargest(5))
print (df['Attempted_suicide'].nsmallest(5))

#Tulostetaan keskiarvo itsetuhoisten yritysten sarakkeelle
print (df['Attempted_suicide'].mean())

#Tarkastellaan mitä eri maita aineistossa on
print (pd.unique(df.Country))

###### DATAN VISUALISOINTI ##################

#Tarkastellaan vuosilukuväliä jolla data on kerätty
print (df['Year'].nlargest(1))
print (df['Year'].nsmallest(1))

#Havainnollistetaan dataa histogrammin avulla
plt.hist(df['Year'])
plt.ylabel('Esiintyvyys')
plt.xlabel('Vuosi')
plt.show()

#Visualisoidaan aineiston ikäjakauma
labels = ["13-15-vuotiaat", "16-17-vuotiaat"]
agegroup = df.groupby('Age Group').size()
agegroup.plot(kind='pie', autopct='%.1f%%', labels=labels, ylabel='', colors=['steelblue', 'pink'])
plt.title('Ikäjakauma')
plt.show()

#Visualisoidaan vuosittaiset itsetuhoiset yritykset
plt.scatter(df['Year'], df['Attempted_suicide'], color='red')
plt.title('Itsetuhoiset yritykset vuosittain')
plt.xlabel('Year')
plt.ylabel('Attempted_suicide')
plt.show()

#Visualisoidaan itsetuhoiset yritykset maittain
df_cat = df.select_dtypes(include=object)
df.boxplot(column = 'Attempted_suicide', by = 'Country', figsize=(25,6), color='green')
plt.title('Itsetuhoiset yritykset maittain')
plt.ylabel('Attempted_suicide')
plt.show()

#Swarmplot itsetuhoisista yrityksistä suhteessa sukupuoleen ja ikäryhmään
sns.swarmplot(data=df, x='Age Group', y='Attempted_suicide', hue='Sex', orient='v', palette='dark')
plt.title('Itsetuhoiset yritykset sukupuolen ja ikäryhmän mukaan')
plt.xlabel('Ikäryhmä')
plt.ylabel('Itsetuhoiset yritykset')
```

```
plt.show()

#Visualisoidaan pylväsdiagrammilla itsetuhoisten yritysten määrää
plt.figure(figsize=(10,6))
sns.set_style('whitegrid', {'axes.grid' : False})
sns.histplot(data=df, x='Attempted_suicide', color='orange')
plt.title('Itsetuhoisten yritysten määrä')
plt.xlabel('Itsetuhoiset yritykset (%)')
plt.ylabel('Lukumäärä')
plt.show()

#Visualisoidaan ei läheisiä ystäviä omaavien ja itsetuhoisen yrityksen tehneiden suhdetta
x = df.No_close_friends
y = df.Attempted_suicide
plt.scatter(x,y, color='brown')
plt.title('Ei läheisiä ystäviä ja itsetuhoiset yritykset')
plt.ylabel('Itsetuhoisia yrityksiä')
plt.xlabel('Ei läheisiä ystäviä')
plt.show()

#Visualisoidaan ei kiusatuksi tulleiden ja itsetuhoisen yrityksen tehneiden suhdetta
x = df.Bullied
y = df.Attempted_suicide
plt.scatter(x,y, color='brown')
plt.title('Kiusatuksi tuleminen ja itsetuhoiset yritykset')
plt.ylabel('Itsetuhoisia yrityksiä')
plt.xlabel('Joutunut kiusatuksi')
plt.show()

###### DATAN ANALYSOINTI ####################

#Tarkastellaan aineiston korrelaatioita
df_corr = df.corr()
plt.figure(figsize=(20,8))
sns.heatmap(df_corr, annot=True, cmap='Blues')
plt.title('Korrelaatio')
plt.show()
#Sovitetaan regressiosuora pistepilveen (kiusatuksi tuleminen & itsetuhoiset yritykset)
sns.regplot(x='Bullied', y='Attempted_suicide', data=df, ci=None, color='purple')
plt.title('Itsetuhoiset yritykset ja kiusatuksi joutuminen')
plt.xlabel('Kiusatuksi tulleet')
plt.ylabel('Itsetuhoiset yritykset')
plt.show()

#Katsotaan korrelaatio itsetuhoisten yritysten ja kiusatuksi tulemisen välillä
p = stats.chi2_contingency(df[['Attempted_suicide', 'Bullied']])[1]
if p > 0.05:
    print (f"Riippuvuus ei ole tilastollisesti merkitsevä, p={p}")
```

```
else:
    print (f'Riippuvuus on tilastollisesti merkitsevä, p={p}')
#Korrelaatio käyttäen pearsonr-rutiinia
p2 = stats.pearsonr(df['Attempted_suicide'], df['Bullied'])
print (f'Korrelaatio (pearsonr): {p2}')
#Korrelaatio käyttäen spearman-rutiinia
p3 = stats.spearmanr(df['Attempted_suicide'], df['Bullied'])
print (f'Korrelaatio (spearman): {p3}')

#Sovitetaan regressiosuora pistepilveen (erittäin humalassa oleminen & itsetuhoiset yritykset)
sns.regplot(x='Really_Get_Drunk', y='Attempted_suicide', data=df, ci=None, color='olive')
plt.title('Itsetuhoiset yritykset ja erittäin humalassa oleminen')
plt.xlabel('Erittäin humalassa olleet')
plt.ylabel('Itsetuhoiset yritykset')
plt.show()

#Katsotaan korrelaatio itsetuhoisten yritysten ja erittäin humalassa olemisen välillä
pp = stats.chi2_contingency(df[['Attempted_suicide', 'Really_Get_Drunk']])[1]
if pp > 0.05:
    print (f'Riippuvuus ei ole tilastollisesti merkitsevä, p={pp}')
else:
    print (f'Riippuvuus on tilastollisesti merkitsevä, p={pp}')
#Korrelaatio käyttäen pearsonr-rutiinia
pp2 = stats.pearsonr(df['Attempted_suicide'], df['Really_Get_Drunk'])
print (f'Korrelaatio (pearsonr): {pp2}')
#Korrelaatio käyttäen spearman-rutiinia
pp3 = stats.spearmanr(df['Attempted_suicide'], df['Really_Get_Drunk'])
print (f'Korrelaatio (spearman): {pp3}')

#Sovitetaan regressiosuora pistepilveen (ei läheisiä ystäviä & itsetuhoiset yritykset)
sns.regplot(x='No_close_friends', y='Attempted_suicide', data=df, ci=None, color='gray')
plt.title('Itsetuhoiset yritykset ja ei läheisiä ystäviä')
plt.xlabel('Ei läheisiä ystäviä')
plt.ylabel('Itsetuhoiset yritykset')
plt.show()

#Katsotaan korrelaatio itsetuhoisten yritysten ja ilman läheisiä ystäviä olemisen välillä
ppp = stats.chi2_contingency(df[['Attempted_suicide', 'No_close_friends']])[1]
if ppp > 0.05:
    print (f'Riippuvuus ei ole tilastollisesti merkitsevä, p={ppp}')
else:
    print (f'Riippuvuus on tilastollisesti merkitsevä, p={ppp}')
#Korrelaatio käyttäen pearsonr-rutiinia
ppp2 = stats.pearsonr(df['Attempted_suicide'], df['No_close_friends'])
print (f'Korrelaatio (pearsonr): {ppp2}')
#Korrelaatio käyttäen spearman-rutiinia
ppp3 = stats.spearmanr(df['Attempted_suicide'], df['No_close_friends'])
print (f'Korrelaatio (spearman): {ppp3}')
```