Tampereen ammattikorkeakoulu

Ilmanlaadun vaikutus astmaan

55G00FY12-3003 DATA-ANALYYSI JA TEKOÄLYN PERUSTEET

Kata Eho

Heejeong Han

RAPORTTI
Maaliskuu 2025

Tietotekniikan tutkinto-ohjelma

SISÄLLYS

1 JOHDANTO 3
2 Käytetty Data 4
2.1 Käytetyn datan tavoite 4
3 Datan käsittely 5
4 Datan käyttö 6
4.1 Tilastollinen tarkastelu 7
5 Itsearviointi 8
6 Liiitteet 9
6.1 python.py 10
6.2 Air_Quality.csv 15

1 JOHDANTO

Raportointi oppimistehtävä1 – Data-analyysin osiota varten.

2 Käytetty Data

Käytämme työssä dataa: https://catalog.data.gov/dataset/air-quality

Se on datasetti Washingtonin ilmanlaadusta. Käytämme työssä vuosittaista PM2.5 partikkelien määrää ja Asthma lääkärikäyntejä johtuen PM2.5 partikkeleista.

2.1 Käytetyn datan tavoite

Tavoitteena on rakentaa datasta esitys, joka näyttää Astman korrelaation ilmanlaatuun, tarkemmin ilmassa oleviin pienpartikkeleihin (PM2.5)

Tavoite arvosana: 2

3 Datan käsittely

Otimme datan ”pandas” DataFrameen, jonka me jaoimme kahteen dataframeen, ”Particles” ja ”Asthma” jotka pitävät sisällään kaiken datan, joka liittyy PM2.5 partikkeleihin ja kaikki Asthma käyntien datat siihen liittyen.

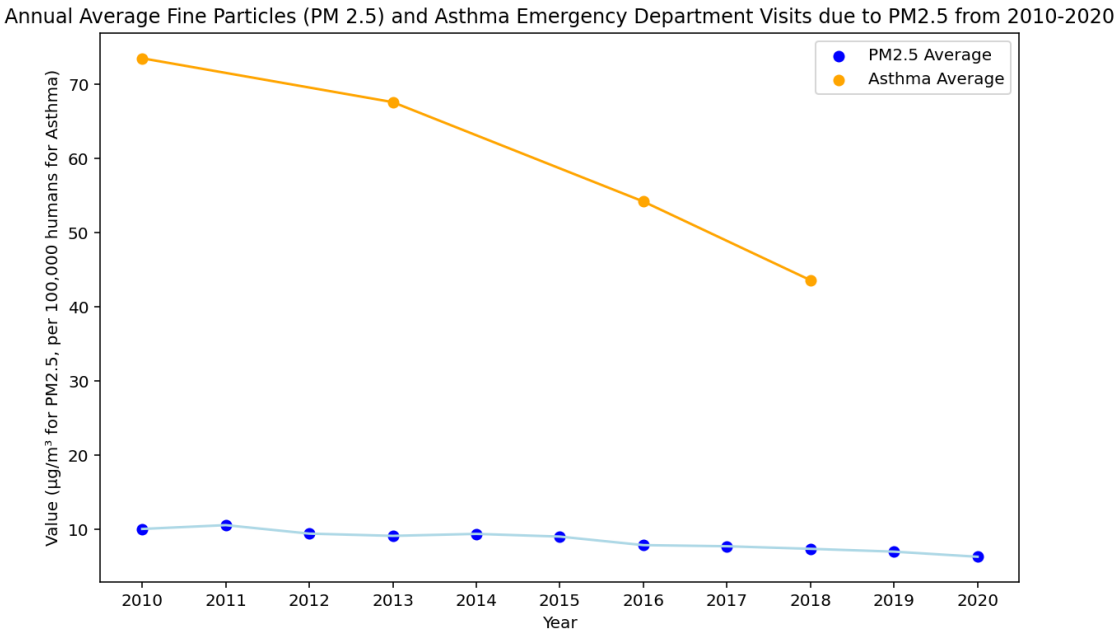
Molemmat dataframet käytiin for loopilla läpi, jossa dataframeista otettiin joka vuodelle tai vuosikategorialle oma dataframe ja dataframet yhdistettiin kahdeksi isommaksi dataframeksi, jossa on vuoden keskiarvo joka vuodelle.

Näistä kahdesta dataframesta vaihdettiin vuositiedot mapilla samaksi, jotta ne saataisiin samaan kuvaajaan. Lopuksi molemmat dataframet aseteltiin kuvaajaksi matplotlib kirjastolla, jossa näkyy vuosittaisten partikkelien määrä ja siihen korreloivat asthmakäyntien määrät.

Tilastollista tarkastelua varten teimme seuraavat toimenpiteet:

- Teimme listan datojen yhteisistä vuosista
- Suodatimme datan käyttämään yhteisiä vuosia
- Järjestimme datan vuosijärjestykseen
- Teimme numPy arrayt arvojen käyttämiseen testissä
- Ajoimme datat pearsonin korrelaatiotestin läpi
- Tarkastelimme ja vertasimme P-arvoa haluttuun tarkastelutarkkuuteen

4 Datan käyttö



Kuva1. PM2.5 partikkelien ja korreloivien astma käyntien määrät kuvaajana

Y-akselilla on määrät, niin partikkelien määrä mikrogrammoina ja astmakäyntien määrä per 100000 ihmistä. X-Akselilta voimme lukea mittausvuoden. Kaikki arvot ovat keskimääriä siltä vuodelta kaikilta alueilta.

Kuten kuvasta näkee, partikkelimäärän pienentymisellä on vaikutus astmakäyntien määrään, molemmat ovat laskeneet joka vuosi. Kuvaajasta toki hieman vaikea huomata partikkelien laskua, sillä partikkelien tarkasteltavat erot ovat suhteessa pieniä astmakäynteihin

Käsiteltyä dataa voisi esimerkiksi käyttää tutkimaan partikkelien vaikutusta astmaan. Alkuperäisestä datasetistä löytyy myös muita ilmanlaatuun liittyviä arvoja, esimerkiksi ajetut kilometrit autoilla, joita voisi käyttää lisäämään tutkielman efektiä.

4.1 Tilastollinen tarkastelu

Kuvaajaa silmillä tarkasteltaessa voimme huomata jo korrelaation Astma käyntien ja ilmassa olevien PM2.5 hiukkasten välillä. Teimme tätä väitettä tukemaan tilastollisen tarkastelun Pearsonin korrelaatiotestillä.

```
# Pearson's correlation coefficient
corr_coeff, p_value = pearsonr(pm25_values, asthma_values)

print(f"Pearson correlation coefficient: {corr_coeff:.3f}")
print(f"P-value: {p_value:.3f}")

# Tulosten tulkinta
if p_value < 0.05:
    print("Correlation has statistical significance.")
else:
    print("Correlation doesn't have statistical significance.")
```

Kuva2. Koodia tilastollisesta tarkastelusta

```
Pearson correlation coefficient: 0.979
P-value: 0.021
Correlation has statistical significance.
```

Kuva3. tilastollisen tarkastelun koodin ulostulo.

Tuloksista voimme todeta, että näiden kahden asian välillä on tilastollisesti merkitsevä korrelaatio, koska testin P-arvo on pienempi kuin haluttu tarkastelun tarkkuus (5%). Hiukkasmäärän pienentyessä myös astmakäyntien tarve pienenee.

5 Itsearviointi

Onnistuimme mielestämme hyvin, saimme tehtyä sen mitä lähdimme tekemään, eli esittelemään ilmanlaadun ja astman korrelaatiota. Mielestämme työ on 2. pisteen arvoinen. Siinä on datan käsittely ja tilastollinen tarkastelu.

Kehityskohteita ovat selkeästi tuotettu Python koodi, sillä aika paljon joutui netistä selaamaan tietoa, että sai asioita aikaiseksi. Huomasimme, että dataframejen pyörittely on kuitenkin jäänyt päähän melko hyvin, eikä niiden kanssa tullut ongelmaa. Myös ongelmanratkaisu tässä tehtävässä oli hyvällä tasolla. Alkuun raavimme takaraivoja hieman, kun mietimme mitenkä dataa olisi järkevä lähteä kasaamaan, mutta päädyimme hyvään vaihtoehtoon hetken miettimisen jälkeen.

Tilastolliseen tarkasteluun jouduimme käyttämään apuna ChatGPT generatiivista tekoälytyökalua.

6 Liiitteet

Liitteinä käytetty alkuperäinen data ja käsittelyyn liittyvä python koodi

6.1 python.py

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import pearsonr



df = pd.read_csv('Air_Quality.csv')
desc = df.describe()

print(df.count())

average_rows = {}

particles = df.loc[df['Name'] == 'Fine particles (PM 2.5)']



for year in range(2010, 2021):
    year_str = f'Annual Average {year}'
    annual_data = particles[particles['Time Period'] == year_str]

    if not annual_data.empty:
        # Calculate the mean of 'Data Value'
        average_value = annual_data['Data Value'].mean()

        # Get the entire row with the mean value
        annual_average_row = annual_data.iloc[[0]]  # Take the first row and modify
it
        annual_average_row['Data Value'] = average_value

        average_rows[year] = annual_average_row

average_df = pd.concat(average_rows.values(), ignore_index=True)
```

```
particle_map = {
    'Annual Average 2010': '2010',
    'Annual Average 2011': '2011',
    'Annual Average 2012': '2012',
    'Annual Average 2013': '2013',
    'Annual Average 2014': '2014',
    'Annual Average 2015': '2015',
    'Annual Average 2016': '2016',
    'Annual Average 2017': '2017',
    'Annual Average 2018': '2018',
    'Annual Average 2019': '2019',
    'Annual Average 2020': '2020'
    }

average_df['Time Period'] = average_df['Time Period'].map(particle_map)



## Asthma

asthma = df.loc[df['Name'] == 'Asthma emergency department visits due to PM2.5']

asthma_average_rows = {}

for year in range(2009, 2019):
    year_str = f'{year}-{year+2}'
    asthma_annual_data = asthma[asthma['Time Period'] == year_str]

    if not asthma_annual_data.empty:
        # Calculate the mean of 'Data Value'
        asthma_average_value = asthma_annual_data['Data Value'].mean()

        # Get the entire row with the mean value
```

```
        asthma_annual_average_row = asthma_annual_data.iloc[[0]]  # Take the first row and modify it
        asthma_annual_average_row['Data Value'] = asthma_average_value

        asthma_average_rows[year] = asthma_annual_average_row


asthma_average_df = pd.concat(asthma_average_rows.values(), ignore_index=True)

asthma_map = {
    '2009-2011': '2010',
    '2012-2014': '2013',
    '2015-2017': '2016',
    '2017-2019': '2018'
    }

asthma_average_df['Time Period'] = asthma_average_df['Time Period'].map(asthma_map)



# Create a DataFrame for plotting
plot_df = pd.DataFrame({
    'Year': average_df['Time Period'],
    'Average PM2.5': average_df['Data Value']
})

asthma_plot_df = pd.DataFrame({
    'Year': asthma_average_df['Time Period'],
    'Average Asthma Visits': asthma_average_df['Data Value']
})

# Plot the data with lines connecting points
plt.figure(figsize=(10, 6))
```

```
plt.scatter(plot_df['Year'], plot_df['Average PM2.5'], color='blue', label='PM2.5 Average')
plt.plot(plot_df['Year'], plot_df['Average PM2.5'], linestyle='-', marker='', color='lightblue')

## Asthma
plt.scatter(asthma_plot_df['Year'], asthma_plot_df['Average Asthma Visits'], color='orange', label='Asthma Average')
plt.plot(asthma_plot_df['Year'], asthma_plot_df['Average Asthma Visits'], linestyle='-', marker='', color='orange')

# Add labels and title
plt.xlabel('Year')
plt.ylabel('Value (µg/m³ for PM2.5, per 100,000 humans for Asthma)')
plt.title('Annual Average Fine Particles (PM 2.5) and Asthma Emergency Department Visits due to PM2.5 from 2010-2020')
plt.legend()

plt.show()



## Statistics
common_years = ['2010', '2013', '2016', '2018']

# Filter Data
pm25_common = average_df[average_df['Time Period'].isin(common_years)]
asthma_common = asthma_average_df[asthma_average_df['Time Period'].isin(common_years)]

# Arrange data
pm25_common = pm25_common.sort_values(by='Time Period')
asthma_common = asthma_common.sort_values(by='Time Period')

# Numpy arrays for testing
pm25_values = pm25_common['Data Value'].to_numpy()
```

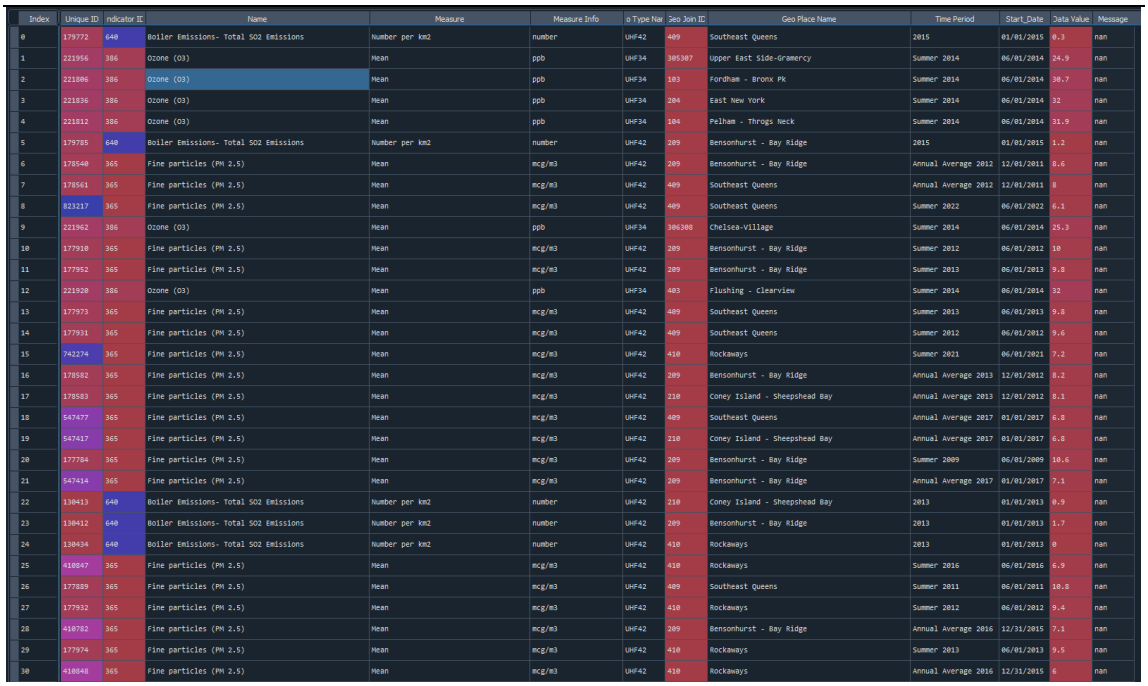
```
asthma_values = asthma_common['Data Value'].to_numpy()

# Pearson's correlation coefficient
corr_coeff, p_value = pearsonr(pm25_values, asthma_values)

print(f"Pearson correlation coefficient: {corr_coeff:.3f}")
print(f"P-value: {p_value:.3f}")

# Tulosten tulkinta
if p_value < 0.05:
    print("Correlation has statistical significance.")
else:
    print("Correlation doesn't have statistical significance.")
```

6.2 Air_Quality.csv



Index	Unique ID	ndicator IE	Name	Measure	Measure Info	o Type Nar	Seo Join ID	Geo Place Name	Time Period	Start_Date	Data Value	Message
0	179772	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	409	Southeast Queens	2015	01/01/2015	0.3	nan
1	221956	386	Ozone (O3)	Mean	ppb	UHF34	305307	Upper East Side-Gramercy	Summer 2014	06/01/2014	24.9	nan
2	221806	386	Ozone (O3)	Mean	ppb	UHF34	103	Fordham - Bronx Pk	Summer 2014	06/01/2014	30.7	nan
3	221836	386	Ozone (O3)	Mean	ppb	UHF34	204	East New York	Summer 2014	06/01/2014	32	nan
4	221812	386	Ozone (O3)	Mean	ppb	UHF34	104	Pelham - Throgs Neck	Summer 2014	06/01/2014	31.9	nan
5	179785	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	209	Bensonhurst - Bay Ridge	2015	01/01/2015	1.2	nan
6	178540	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Annual Average 2012	12/01/2011	8.6	nan
7	178561	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409	Southeast Queens	Annual Average 2012	12/01/2011	8	nan
8	823217	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409	Southeast Queens	Summer 2022	06/01/2022	6.1	nan
9	221962	386	Ozone (O3)	Mean	ppb	UHF34	306308	Chelsea-Village	Summer 2014	06/01/2014	25.3	nan
10	177910	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Summer 2012	06/01/2012	10	nan
11	177952	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Summer 2013	06/01/2013	9.8	nan
12	221920	386	Ozone (O3)	Mean	ppb	UHF34	403	Flushing - Clearview	Summer 2014	06/01/2014	32	nan
13	177973	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409	Southeast Queens	Summer 2013	06/01/2013	9.8	nan
14	177931	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409	Southeast Queens	Summer 2012	06/01/2012	9.6	nan
15	742274	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	410	Rockaways	Summer 2021	06/01/2021	7.2	nan
16	178582	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Annual Average 2013	12/01/2012	8.2	nan
17	178583	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	210	Coney Island - Sheepshead Bay	Annual Average 2013	12/01/2012	8.1	nan
18	547477	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409	Southeast Queens	Annual Average 2017	01/01/2017	6.8	nan
19	547417	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	210	Coney Island - Sheepshead Bay	Annual Average 2017	01/01/2017	6.8	nan
20	177724	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Summer 2009	06/01/2009	10.6	nan
21	547414	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Annual Average 2017	01/01/2017	7.1	nan
22	130413	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	210	Coney Island - Sheepshead Bay	2013	01/01/2013	0.9	nan
23	130412	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	209	Bensonhurst - Bay Ridge	2013	01/01/2013	1.7	nan
24	130424	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	410	Rockaways	2013	01/01/2013	0	nan
25	410847	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	410	Rockaways	Summer 2016	06/01/2016	6.9	nan
26	177889	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409	Southeast Queens	Summer 2011	06/01/2011	10.8	nan
27	177932	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	410	Rockaways	Summer 2012	06/01/2012	9.4	nan
28	410782	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209	Bensonhurst - Bay Ridge	Annual Average 2016	12/31/2015	7.1	nan
29	177974	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	410	Rockaways	Summer 2013	06/01/2013	9.5	nan
30	410848	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	410	Rockaways	Annual Average 2016	12/31/2015	6	nan