

Imports

```
In [ ]: import pandas as pd
import geopandas as gpd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
from ipabales import init_notebook_mode
import plotly.express as px

init_notebook_mode(all_interactive=True)
```

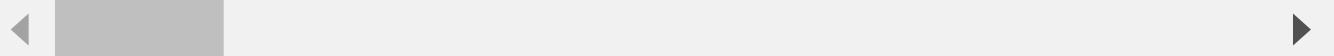
Loading data

```
In [ ]: data = pd.read_excel(
    '../data/who_aap_2021_v9_11august2022.xlsx', sheet_name='AAP_2022_city_v9'
)
data.columns = [
    "WHO Region",
    "ISO3",
    "WHO Country Name",
    "City or Locality",
    "Measurement Year",
    "PM2.5",
    "PM10",
    "NO2",
    "PM2.5 temporal coverage",
    "PM10 temporal coverage",
    "NO2 temporal coverage",
    "Reference",
    "Number and type of monitoring stations",
    "Version of the database",
    "Status",
]
data
```

Out[]: 10 ✓ entries per page

WHO Region	ISO3	WHO Country Name
Eastern Mediterranean Region	AFG	Afghanistan
European Region	ALB	Albania

Showing 1 to 10 of 546 entries (downsampled from 32,191x15 to 546x15 as maxBytes=65536)



Biblioteka `ITables 2.0` ułatwia wyszukiwanie konkretnej wartości w wyświetlanych danych (DataFrame'ach). Dodatkowo pozwala obejrzeć cały output, niezależnie od jego wielkości, co jest niewątpliwym plusem. W kontekście badanych danych używano jej np. do podglądzania, czy dla danej wartości zaszła przewidywana zmiana, co się kryje w innej kolumnie wiersza, gdzie znajduje się szukana wartość oraz do wspomnianego wcześniej przeglądania dłuższych outputów.

Data review

Data types

In []: `data.dtypes`

Out[]: 10 ✓ entries per page

Search:

		0
	WHO Region	object
	ISO3	object
	WHO Country Name	object
	City or Locality	object
	Measurement Year	int64
	PM2.5	float64
	PM10	float64
	NO2	float64
	PM2.5 temporal coverage	float64
	PM10 temporal coverage	float64

Showing 1 to 10 of 15 entries

« < 1 2 > »

Wszystkie typy danych są zgodne z logiką i prezentowaną zawartością.

Basic statistics

In []: `data.describe()`

	Measurement Year	PM2.5	PM10	NO2	PM2.5 temporal coverage
count	32191	15048	21109	22200	7275
mean	2015.579354	22.92032	30.533252	20.619336	90.794096
std	2.752654	17.925906	29.312756	12.133388	14.872681
min	2000	0.01	1.04	0	0
25%	2014	10.35	16.98	12	88.59589
50%	2016	16	22	18.8	97
75%	2018	31	31.3	27.16	99
max	2021	191.9	540	210.68	100



Minimalne i maksymalne wartości w kolumnach "PM2.5 temporal coverage", "PM10 temporal coverage" oraz "NO2 temporal coverage" są zgodne z logiką (wiedząc, że są to procenty rocznego pokrycia pomiarami od 0 do 100), zatem nie ma w tych kolumnach wartości odstających. Kolumna "Measurement Year" i "Version of the database columns",

bazując na przedstawionych podstawowych statystykach, również wydają się nie zawierać wartości odstających. Co do kolumn "PM2.5", "PM10" oraz "NO2" - mogą one zawierać wartości odstające, ale mając na uwadze, że są to dane dotyczące wartości zanieczyszczeń powietrza, nie zaleca się usuwania ich - zdecydowanie zaburzyłoby to dalszą analizę i ograniczyło jej sens.

NAs

In []: `data.isna().sum()`

Out[]: 10 entries per page

Search:

	0
WHO Region	1
ISO3	0
WHO Country Name	0
City or Locality	0
Measurement Year	0
PM2.5	17143
PM10	11082
NO2	9991
PM2.5 temporal coverage	24916
PM10 temporal coverage	26810

Showing 1 to 10 of 15 entries

« < 1 2 > »

Liczba wartości odstających (1) w kolumnie "WHO Region" budzi niepokój i zostanie to później sprawdzone. Wartości odstające w kolumnach "PM2.5", "PM10" i "NO2" wynikają z braku dokonania pomiaru w danym mieście w danym roku i ich typ można określić jako MAR - wartości tych brakuje często dla konkretnych lokalizacji lub lat. Braki w "PM2.5 temporal coverage", "PM10 temporal coverage" i "NO2 temporal coverage" również wynikają z braku wprowadzenia ich do bazy przez dane źródło, choć ich typ przypomina tu bardziej MCAR - nie zależy to od pozostałych danych. 5 NaNs w kolumnie "Reference" również zostanie zbadane w dalszej analizie. Kolumna "Number and type of monitoring stations" zawiera bardzo wiele wartości brakujących, ich typ również najbardziej przypomina MCAR. Kolumna "Status" w 100% składa się z NaNów, więc zostanie ona najprawdopodobniej usunięta na etapie selekcji atrybutów.

Checking the NA in WHO Region column

In []: `data[data["WHO Region"].isna()]`

Out[]:	WHO Region	ISO3	WHO Country Name	City or Locality	Measurement
	24778	NaN	LIE	Liechtenstein	Vaduz

In []:	<code>data["WHO Region"].unique()</code>
Out[]:	<code>array(['Eastern Mediterranean Region', 'European Region', 'Region of the Americas', 'Western Pacific Region', 'South East Asia Region', 'African Region', nan], dtype=object)</code>
In []:	<code>na_slice = data[data["WHO Region"].isna()].copy() na_slice.fillna({"WHO Region": "European Region"}, inplace=True) data.update(na_slice) data[data["ISO3"] == "LIE"]</code>

Out[]:	WHO Region	ISO3	WHO Country Name	City or Locality	Measurement
	24778	European Region	LIE	Liechtenstein	Vaduz

Jedna brakująca wartość w kolumnie "WHO Region" istniała dla Lichtensteina. Wynikało to najprawdopodobniej z błędu przy wprowadzaniu danych do bazy. Jako że Lichtenstein leży na terenie Europy założono, że należy on do tego samego regionu WHO co inne europejskie państwa, więc wypełniono brak odpowiednim regionem.

Checking NAs in Reference column

In []:	<code>data[data["Reference"].isna()]</code>				
Out[]:	WHO Region	ISO3	WHO Country Name	City or Locality	N
	28209	Eastern Mediterranean Region	QAT	Qatar	Doha
	28210	Eastern Mediterranean Region	QAT	Qatar	Doha
	28211	Eastern Mediterranean Region	QAT	Qatar	Doha
	28212	Eastern Mediterranean Region	QAT	Qatar	Doha
	28213	Eastern Mediterranean Region	QAT	Qatar	Doha

Braki w kolumnie "Reference" występują tylko dla Dohy w Katarze. Z powodu niemożności oszacowania możliwego źródła danych z tego obszaru oraz mając na uwadze małe znaczenie tej informacji w kontekście przeprowadzanej analizy, zdecydowano pozostawić ten brak takim, jakim jest.

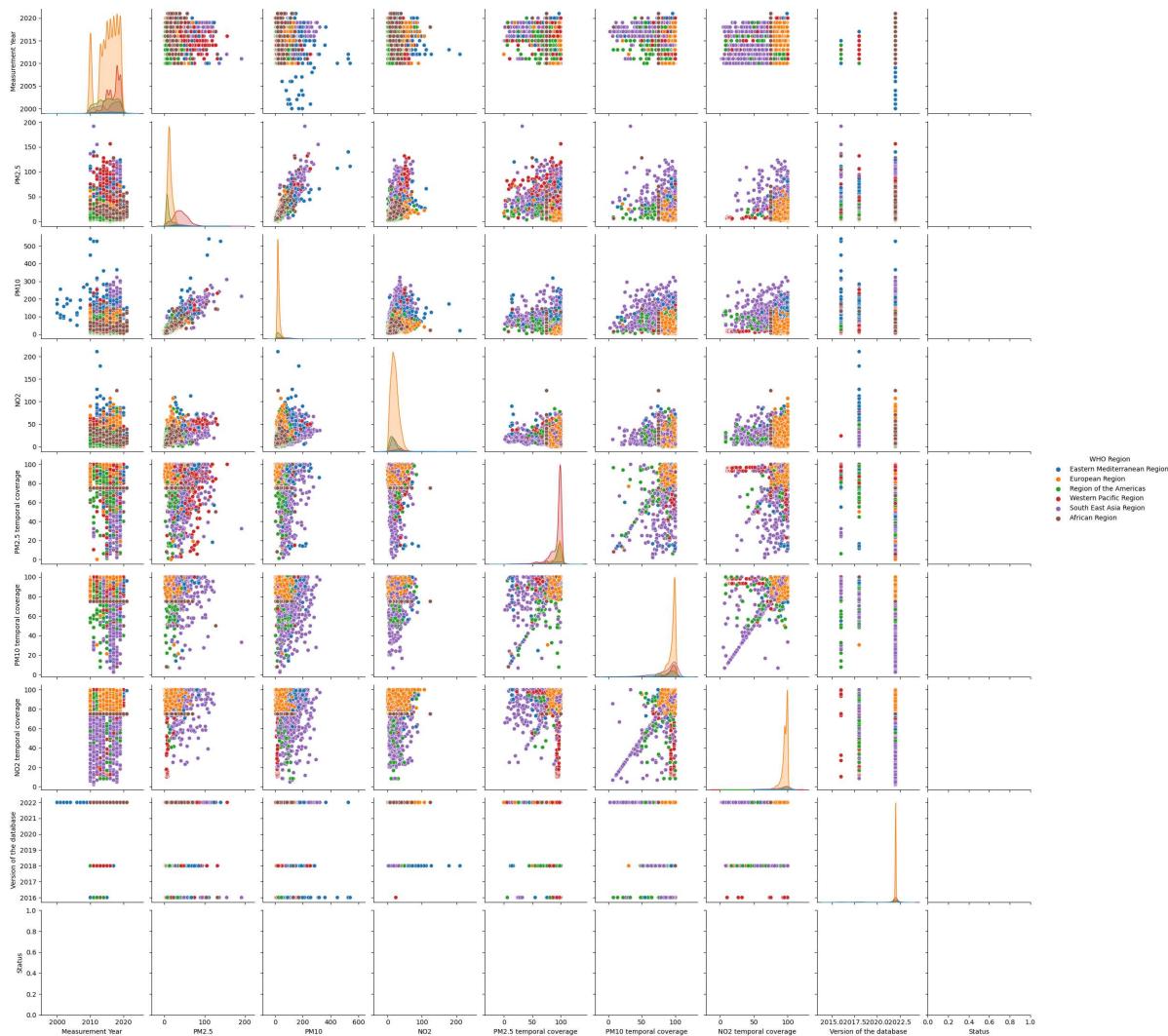
NAs in Number and type of monitoring stations column

Co do braków w kolumnie "Number and types of monitoring stations" liczba stacji może się zmieniać z czasem (dodawane są nowe, bądź aktualne mają awarię), więc uzupełnianie ich tą samą wartością, która znajduje się dla danego roku i miasta może wprowadzać w błąd. Dodatkowo, niektóre źródła nie zapewniają wcale takich informacji, więc biorąc to wszystko pod uwagę nic nie zrobiono w tymi brakami, jednocześnie wiedząc, że usunięcie ich znacznie skróciłoby dostępny zbiór danych. Typ tych braków można zakwalifikować jako MNAR, ponieważ może on zależeć od innych czynników niż dostępne.

Relation between variables

```
In [ ]: sns.pairplot(data, hue="WHO Region").fig.suptitle("Relations and distributions of variables vs WHO region", y=1.05, fontsize=20)
plt.show()
```

Relations and distributions of variables vs WHO region



Z przedstawionej figury można odczytać pewne relacje, np. dodatnia zależność liniowa między wartościami PM2.5 i PM10, co zostanie dogłębniej zbadane w dalszych etapach.

Feature selection and feature engineering

```
In [ ]: data.drop(columns=["Version of the database", "Status"], inplace=True)
```

Zadecydowano o usunięciu kolumny "Status", ponieważ składała się ona w 100% w NaNów, więc nie wnosiła nic do analizy. Dodatkowo usunięto kolumnę "Version of the database", ponieważ nie wpływała ona na wyjaśnienie wartości zanieczyszczeń powietrza.

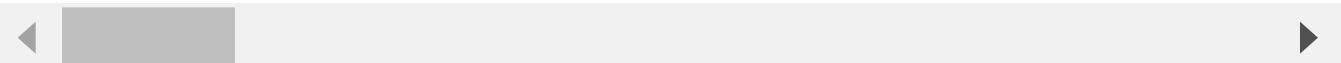
```
In [ ]: def extract_and_sum_numbers(text):
    if pd.notna(text):
        numbers = [
            int(word.split()[0])
            for word in text.split(",")
            if word.strip()[0].isdigit()
        ]
        return sum(numbers)
    else:
        return np.nan

data["Total number of stations"] = data["Number and type of monitoring stations"].apply(extract_and_sum_numbers)
)
data
```

Out[]: 10 entries per page

WHO Region	ISO3	WHO Country Name
Eastern Mediterranean Region	AFG	Afghanistan
European Region	ALB	Albania

Showing 1 to 10 of 585 entries (downsampled from 32,191x14 to 585x14 as maxBytes=65536)



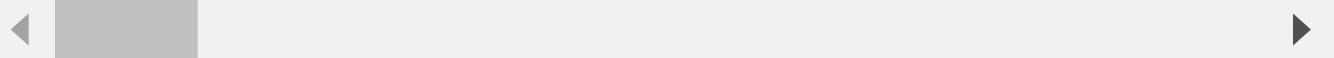
Dodano nową zmienną "Total number of stations", aby zbadać zależność ilości stacji pomiarowych od wartości zanieczyszczeń - czy np. bardziej zanieczyszczone regiony stawiają więcej stacji, aby być świadomym zagrożenia, czy też nie.

```
In [ ]: region_dummies = pd.get_dummies(data["WHO Region"])
data_dummies = pd.concat([data, region_dummies], axis=1)
data_dummies
```

Out[]: 10 ✓ entries per page

WHO Region	ISO3	WHO Country Name
Eastern Mediterranean Region	AFG	Afghanistan
European Region	ALB	Albania

Showing 1 to 10 of 555 entries (downsampled from 32,191x20 to 555x20 as maxBytes=65536)



```
In [ ]: country_dummies = pd.get_dummies(data["WHO Country Name"])
data_dummies = pd.concat([data_dummies, country_dummies], axis=1)
data_dummies
```

Out[]: 10 entries per page

WHO Region	ISO3	WHO Country Name	City or Locality	Measur
Eastern Mediterranean Region	AFG	Afghanistan	Kabul	
European Region	ALB	Albania	Durres	
European Region	ALB	Albania	Durres	
European Region	ALB	Albania	Elbasan	
European Region	ALB	Albania	Elbasan	
European Region	ALB	Albania	Elbasan	
European Region	ALB	Albania	Korce	
European Region	ALB	Albania	Korce	
European Region	ALB	Albania	Vlore	
European Region	ALB	Albania	Vlore	

Showing 1 to 10 of 277 entries (downsampled from 32,191x138 to 277x138 as maxBytes=65536)



Do odpowiednich kopii ogólnego datasetu dodano także dummy values utworzone na podstawie kolumn "WHO Region" oraz "WHO Country Name", aby móc sprawdzić zależność między danym krajem lub regionem (do regresji punktowo-dwuseryjnej).

WHO Region-based analysis

```
In [ ]: xlabel = [x.replace(" ", "\n") for x in data["WHO Region"].unique()]
xlabels

fig, ax = plt.subplots(2, 2, figsize=(18, 15))

sns.histplot(data=data, x="WHO Region", ax=ax[0, 0], shrink=0.9)
ax[0, 0].set_xticks(ticks=range(len(xlabels)), labels=xlabels)
ax[0, 0].set(title="Number of measurements vs WHO Region")

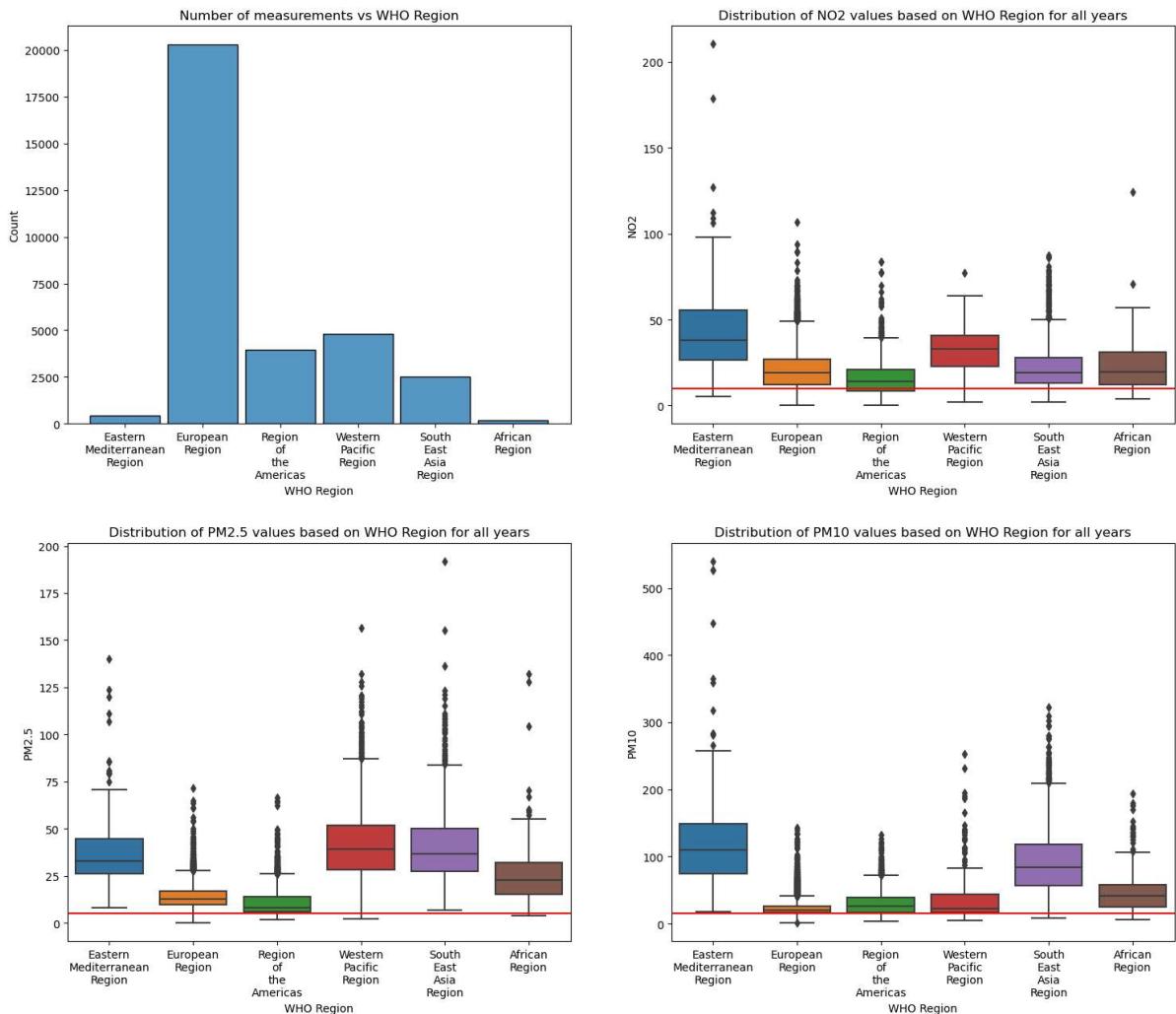
sns.boxplot(data=data, y="NO2", x="WHO Region", ax=ax[0, 1])
ax[0, 1].axhline(y=10, color="r")
ax[0, 1].set_xticks(ticks=range(len(xlabels)), labels=xlabels)
ax[0, 1].set(title="Distribution of NO2 values based on WHO Region for all years")

sns.boxplot(data=data, y="PM2.5", x="WHO Region", ax=ax[1, 0])
ax[1, 0].axhline(y=5, color="r")
ax[1, 0].set_xticks(ticks=range(len(xlabels)), labels=xlabels)
ax[1, 0].set(title="Distribution of PM2.5 values based on WHO Region for all years")

sns.boxplot(data=data, y="PM10", x="WHO Region", ax=ax[1, 1])
ax[1, 1].axhline(y=15, color="r")
ax[1, 1].set_xticks(ticks=range(len(xlabels)), labels=xlabels)
ax[1, 1].set(title="Distribution of PM10 values based on WHO Region for all years")
```

```
plt.suptitle("Data distribution based on WHO Region", fontsize=16)
plt.subplots_adjust(hspace=0.3, wspace=0.2)
plt.show()
```

Data distribution based on WHO Region



Analizę rozpoczęto od generalizacji na regiony WHO. Wykres w lewym górnym rogu przedstawia ilość pomiarów wykonaną dla każdego regionu. Jak widać, zdecydowanie przoduje region europejski, co może świadczyć o największej świadomości ekologicznej w tym regionie. Najgorzej wypadają tu region Wschodniego Morza Śródziemnego oraz Afryka, głównie ze względu na to, że w tych regionach jest wiele krajów biednych, rozwijających się, do których świadomość ekologiczna jeszcze nie dotarła, bądź jest traktowana po macoszemu.\ Pozostałe wykresy przedstawiają rozkład wartości zanieczyszczeń z podziałem na regiony. Czerwona linia oznacza normę uznawaną przez WHO dla każdego z zanieczyszczeń. Pokazuje to jak złym powietrzem oddycha się na całym świecie. Najlepiej pod względem PM_{2.5} wypada region Ameryk oraz Europa, mając jednak na uwadze liczne wartości powyżej 95 percentyla. Najgorsza sytuacja panuje w regionie Zachodniego Pacyfiku (gdzie normy zawyżają Chiny oraz Wietnam) oraz w Azji Południowo-Wschodniej (tutaj głównym winowającą są Indie). W przypadku PM₁₀ sytuacja wygląda podobnie pod względem niskich wartości - tu przodują wciąż Europa i Ameryki, ale tym razem najgorsze powietrze (pod względem PM₁₀) jest w regionie Wschodniego Morza Śródziemnego oraz w Azji Południowo-Wschodniej. Analogicznie badając wartości NO₂.

Correlation

```
In [ ]: data.columns
```

```
Out[ ]: Index(['WHO Region', 'ISO3', 'WHO Country Name', 'City or Locality',
       'Measurement Year', 'PM2.5', 'PM10', 'NO2', 'PM2.5 temporal coverage',
       'PM10 temporal coverage', 'NO2 temporal coverage', 'Reference',
       'Number and type of monitoring stations', 'Total number of stations'],
      dtype='object')
```

```
In [ ]: corr_cols = [
    "Measurement Year",
    "PM2.5",
    "PM10",
    "NO2",
    "PM2.5 temporal coverage",
    "PM10 temporal coverage",
    "NO2 temporal coverage",
    "Total number of stations",
]
```

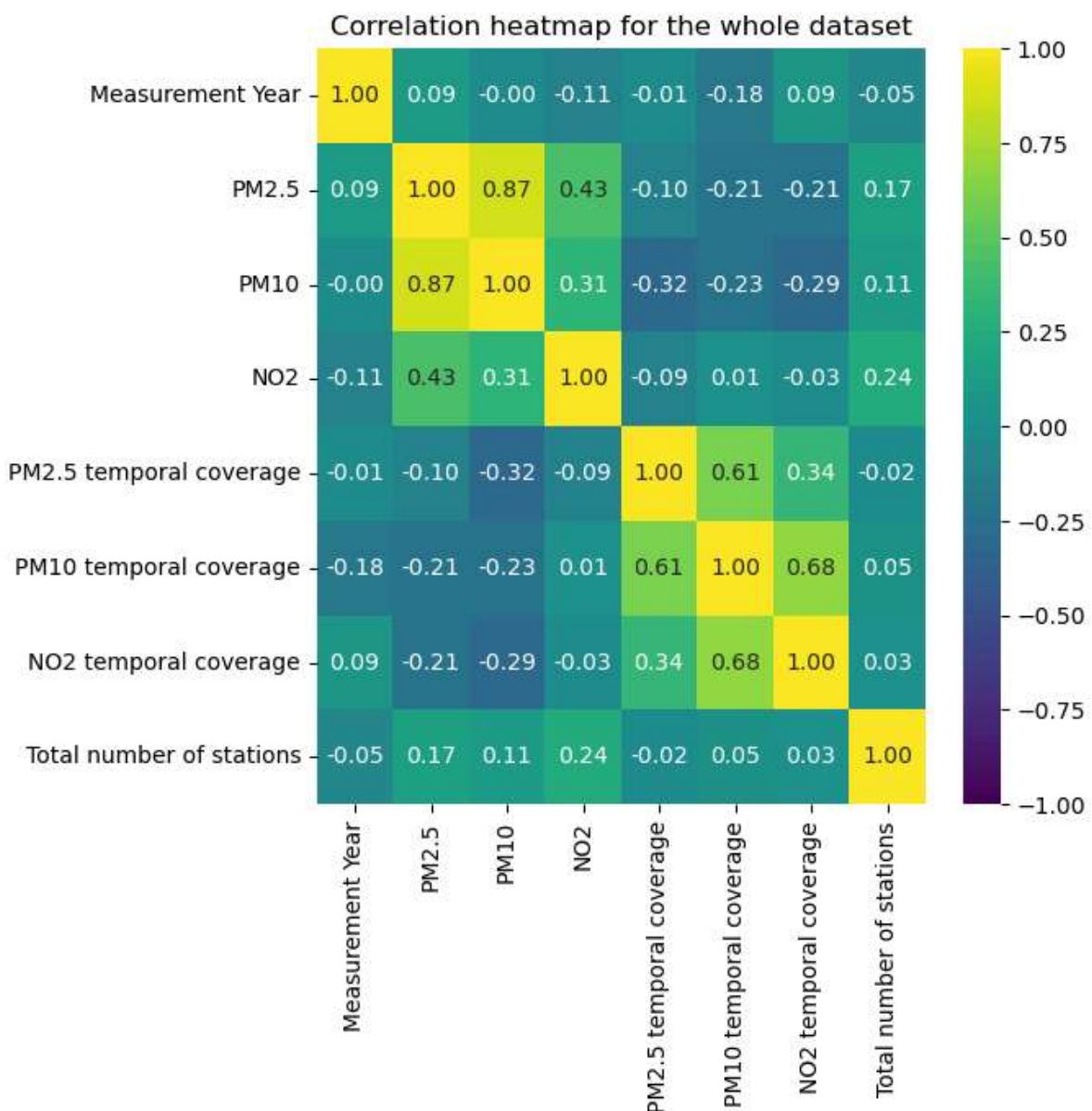
```
corr_df = data[corr_cols].corr()
```

Wybrano jedynie takie kolumny do analizy korelacji liniowej Pearsona, które zawierają wartości ciągłe.

```
In [ ]: fig, ax = plt.subplots(1, 1, figsize=(6, 6))

sns.heatmap(
    corr_df,
    ax=ax,
    vmin=-1,
    vmax=1,
    center=0,
    annot=True,
    fmt=".2f",
    cmap="viridis",
    annot_kws={"fontsize": 10},
)
ax.set(title="Correlation heatmap for the whole dataset")

plt.show()
```



Wartości korelacji liniowej dla całego zbioru danych wyraźnie wskazują istotną zależność między wartościami PM2.5 oraz PM10, a także mniejsze pomiędzy NO2 a PM2.5 oraz PM10. Ponadto można odczytać niewielkie ujemne zależności między wartościami PM2.5, PM10 i NO2 a stopniem rocznego pokrycia każdej z nich - jest to ciekawe, ponieważ pokazuje, że im częściej wykonywano pomiary, tym mniejsze były zanieczyszczenia, co może być pośrednio związane z tym, że większe pokrycie występowało najczęściej w krajach bardziej rozwiniętych. Ponadto występują dodatnie zależności między pokryciami każdego ze współczynników, co jest dość logiczne - jeśli stacja mierzyła jeden, to najczęściej i pozostałe współczynniki, i na odwrót. Występuje też niewielka dodatnia zależność pomiędzy ilością stacji pomiarowych a wartością NO2.

```
In [ ]: i = 0
j = 0

fig, ax = plt.subplots(2, 3, figsize=(23, 15))

for r in data["WHO Region"].unique():
    corr_df = data[data["WHO Region"] == r][corr_cols].corr()
    sns.heatmap(
        corr_df,
        ax=ax[i][j],
        vmin=-1,
```

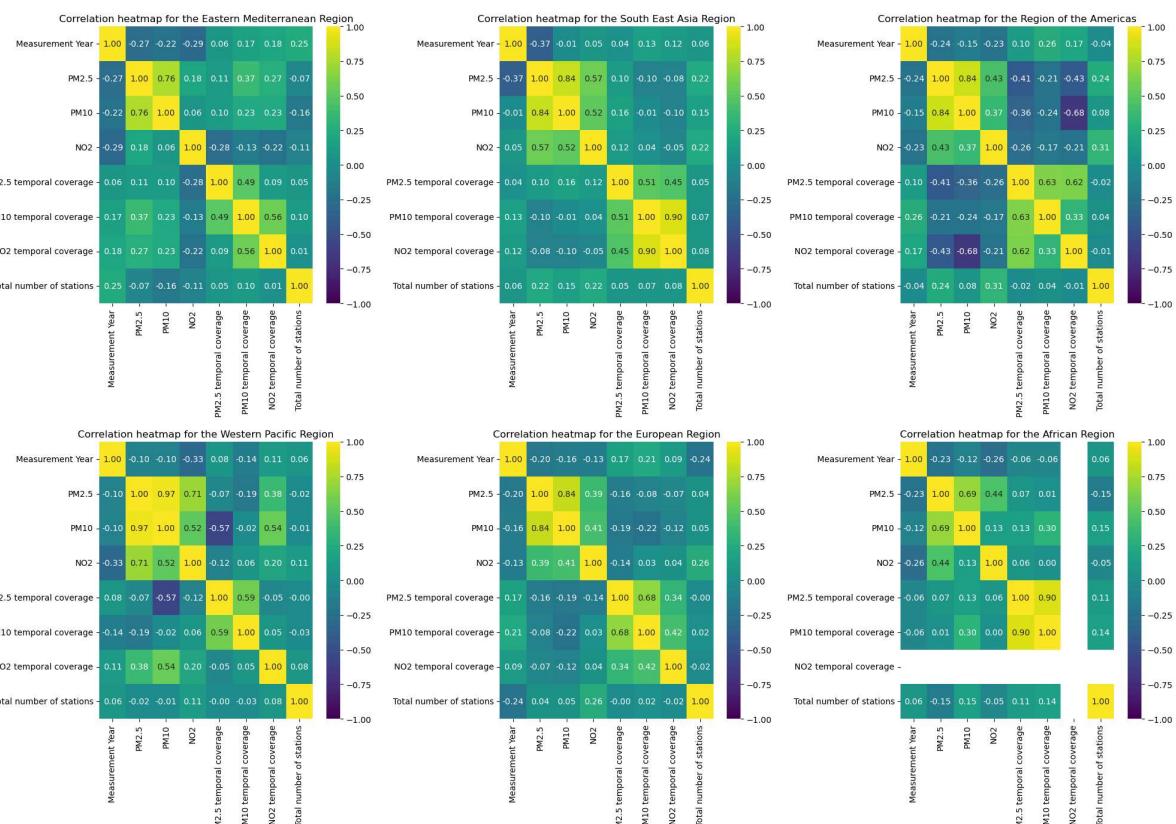
```

    vmax=1,
    center=0,
    annot=True,
    fmt=".2f",
    cmap="viridis",
    annot_kws={"fontsize": 10},
)
ax[i][j].set(title=f"Correlation heatmap for the {r}")
i += 1
i %= 2
j += 1
j %= 3

plt.subplots_adjust(hspace=0.5, wspace=0.5)
plt.suptitle("Correlation heatmaps for each region", fontsize=16)
plt.show()

```

Correlation heatmaps for each region



Na powyższych heatmapach przedstawiono zależności pomiędzy poszczególnymi zmiennymi dla każdego z regionów. Wartości współczynników korelacji nierzaz mocno różnią się od siebie, jednak wszędzie występują duże dodatnie zależności pomiędzy wartościami każdego ze współczynników, szczególnie PM2.5 i PM10. Ponadto w regionie Azji Południowo-Wschodniej występuje ujemna zależność między rokiem pomiaru a wartością PM2.5. Wskazuje to na nieznaczną poprawę jakości powietrza w tym regionie wraz z upływem lat. Dla regionu Ameryk wyróżniają się ujemne zależności między wartościami współczynników a ich stopniem rocznego pokrycia. W regionie europejskim wyróżniają się wysokie wartości współczynników dla stopni pokrycia dla siebie nawzajem, co może wskazywać, że zazwyczaj europejskie czujniki mierzą wszystkie z 3 rodzajów zanieczyszczeń. Dla Afryki widać, że zupełnie nie wypełniano kolumny z rocznym pokryciem pomiarami współczynnika NO2 oraz, że występuje bardzo silna pozytywna zależność między stopniem pokrycia PM2.5 i PM10.

Point-biserial correlation

```
In [ ]: for r in data_dummies["WHO Region"].unique():
    print(f"\n{r}:")
    for c in corr_cols:
        temp_df = data_dummies.dropna(subset=c)
        print(f"{r} vs {c}: {stats.pointbiserialr(temp_df[r], temp_df[c])}")
```

Eastern Mediterranean Region:

Eastern Mediterranean Region vs Measurement Year: SignificanceResult(statistic=-0.020433493981191153, pvalue=0.00024600122593072257)
 Eastern Mediterranean Region vs PM2.5: SignificanceResult(statistic=0.1031366363946883, pvalue=7.199182966700921e-37)
 Eastern Mediterranean Region vs PM10: SignificanceResult(statistic=0.36426022043226786, pvalue=0.0)
 Eastern Mediterranean Region vs NO2: SignificanceResult(statistic=0.17002278268537616, pvalue=1.264128529846678e-143)
 Eastern Mediterranean Region vs PM2.5 temporal coverage: SignificanceResult(statistic=-0.14136809403182024, pvalue=8.62450943600881e-34)
 Eastern Mediterranean Region vs PM10 temporal coverage: SignificanceResult(statistic=-0.07623380154639861, pvalue=2.1558817653567293e-08)
 Eastern Mediterranean Region vs NO2 temporal coverage: SignificanceResult(statistic=-0.007052246771967366, pvalue=0.31995921477155914)
 Eastern Mediterranean Region vs Total number of stations: SignificanceResult(statistic=0.0350995745864689, pvalue=0.0010187075804140191)

European Region:

European Region vs Measurement Year: SignificanceResult(statistic=-0.06181305313485798, pvalue=1.2461816552351145e-28)
 European Region vs PM2.5: SignificanceResult(statistic=-0.5019660662896209, pvalue=0.0)
 European Region vs PM10: SignificanceResult(statistic=-0.4709508188792565, pvalue=0.0)
 European Region vs NO2: SignificanceResult(statistic=-0.034143428574115814, pvalue=3.609955837755273e-07)
 European Region vs PM2.5 temporal coverage: SignificanceResult(statistic=0.08638815294114853, pvalue=1.5723213252180334e-13)
 European Region vs PM10 temporal coverage: SignificanceResult(statistic=0.3171493900311768, pvalue=5.152144759387591e-126)
 European Region vs NO2 temporal coverage: SignificanceResult(statistic=0.36944088974020795, pvalue=0.0)
 European Region vs Total number of stations: SignificanceResult(statistic=-0.02608689386184582, pvalue=0.01463093702593799)

Region of the Americas:

Region of the Americas vs Measurement Year: SignificanceResult(statistic=-0.08773089342540233, pvalue=4.977480469986595e-56)
 Region of the Americas vs PM2.5: SignificanceResult(statistic=-0.26049484419636904, pvalue=6.489615190371197e-232)
 Region of the Americas vs PM10: SignificanceResult(statistic=0.0017056317906980131, pvalue=0.8042925071377917)
 Region of the Americas vs NO2: SignificanceResult(statistic=-0.13093600997685742, pvalue=1.7978730079789976e-85)
 Region of the Americas vs PM2.5 temporal coverage: SignificanceResult(statistic=0.05833064553479305, pvalue=6.404069137069385e-07)
 Region of the Americas vs PM10 temporal coverage: SignificanceResult(statistic=-0.08615753766238893, pvalue=2.4449908290095345e-10)
 Region of the Americas vs NO2 temporal coverage: SignificanceResult(statistic=-0.00721393180807116, pvalue=0.9189685457267649)
 Region of the Americas vs Total number of stations: SignificanceResult(statistic=-0.10006409915353032, pvalue=6.1940915210474116e-21)

Western Pacific Region:

Western Pacific Region vs Measurement Year: SignificanceResult(statistic=0.1907110651035102, pvalue=2.5510330156250427e-261)
 Western Pacific Region vs PM2.5: SignificanceResult(statistic=0.6279332135664273, pvalue=0.0)
 Western Pacific Region vs PM10: SignificanceResult(statistic=0.01417317443144923, pvalue=0.039475752565946924)
 Western Pacific Region vs NO2: SignificanceResult(statistic=0.15804727736142404, pvalue=3.940644188511663e-124)
 Western Pacific Region vs PM2.5 temporal coverage: SignificanceResult(statistic=0.

```
1547084989920406, pvalue=3.3040342092080696e-40)
Western Pacific Region vs PM10 temporal coverage: SignificanceResult(statistic=0.0
58758057555798396, pvalue=1.6109785653262883e-05)
Western Pacific Region vs NO2 temporal coverage: SignificanceResult(statistic=-0.1
327902656663876, pvalue=6.272854169555007e-79)
Western Pacific Region vs Total number of stations: SignificanceResult(statistic=
0.026340006579452183, pvalue=0.013698117063235612)
```

South East Asia Region:

```
South East Asia Region vs Measurement Year: SignificanceResult(statistic=-0.030471
870781641, pvalue=4.5443664453860183e-08)
South East Asia Region vs PM2.5: SignificanceResult(statistic=0.21706787449153575,
pvalue=6.106384672875838e-160)
South East Asia Region vs PM10: SignificanceResult(statistic=0.5869955522796655, p
value=0.0)
South East Asia Region vs NO2: SignificanceResult(statistic=0.040576068267698405,
pvalue=1.467932417415133e-09)
South East Asia Region vs PM2.5 temporal coverage: SignificanceResult(statistic=-
0.34018669806329116, pvalue=1.585732162934143e-196)
South East Asia Region vs PM10 temporal coverage: SignificanceResult(statistic=-0.
25030145383864205, pvalue=1.1634152500678984e-77)
South East Asia Region vs NO2 temporal coverage: SignificanceResult(statistic=-0.3
539678931236533, pvalue=0.0)
South East Asia Region vs Total number of stations: SignificanceResult(statistic=
0.08215839173536814, pvalue=1.3537639103873756e-14)
```

African Region:

```
African Region vs Measurement Year: SignificanceResult(statistic=0.016508611625978
9, pvalue=0.0030561082394488653)
African Region vs PM2.5: SignificanceResult(statistic=0.02214265459012913, pvalue=
0.006600545585984725)
African Region vs PM10: SignificanceResult(statistic=0.0637242950419178, pvalue=1.
908574190231089e-20)
African Region vs NO2: SignificanceResult(statistic=0.015695202706818095, pvalue=
0.019359000014242633)
African Region vs PM2.5 temporal coverage: SignificanceResult(statistic=-0.1655294
2878503737, pvalue=7.465659578923453e-46)
African Region vs PM10 temporal coverage: SignificanceResult(statistic=-0.19710229
395003298, pvalue=2.8680642200596e-48)
African Region vs NO2 temporal coverage: SignificanceResult(statistic=-0.151156300
2336828, pvalue=5.683997130702588e-102)
African Region vs Total number of stations: SignificanceResult(statistic=0.0274676
89198414545, pvalue=0.010150509323086856)
```

Badając wartości współczynników korelacji punktowo-dwuseryjnej zaobserwowano:

- dla regionu Wschodniego Morza Śródziemnego występuje dodatnia zależność wobec wartości PM10 - czyli tutaj są one zazwyczaj większe niż w innych regionach
- w Europie występują wysokie ujemne zależności wobec PM2.5 oraz PM10 - powietrze jest nieco czystsze
- dla regionu Ameryk występuje niewielka ujemna zależność wobec PM2.5
- dla Azji Południowo-Wschodniej zachodzi duża dodatnia zależność wobec wartości PM10 - w tym regionie są one wyższe niż w pozostałych; ponadto występuje tu zazwyczaj wyższe roczne pokrycie pomiarami
- dla Afryki nie zauważono żadnych znaczących zależności

Air pollution across years

```
In [ ]: grouped_df_pm25 = data.groupby([ "WHO Region", "Measurement Year"])[ "PM2.5"].mean(
    numeric_only=True)
```

```

    )
grouped_df_pm10 = data.groupby(["WHO Region", "Measurement Year"])["PM10"].mean(
    numeric_only=True
)
grouped_df_no2 = data.groupby(["WHO Region", "Measurement Year"])["NO2"].mean(
    numeric_only=True
)

```

```

In [ ]: fig, ax = plt.subplots(3, 1, figsize=(15, 12))

sns.lineplot(grouped_df_pm25, x="Measurement Year", y=grouped_df_pm25.values, ax=ax[0])
ax[0].axhline(y=5, color="r")
ax[0].set(title="PM2.5 pullution", ylabel="PM2.5 [( $\mu\text{g}/\text{m}^3$ )"])

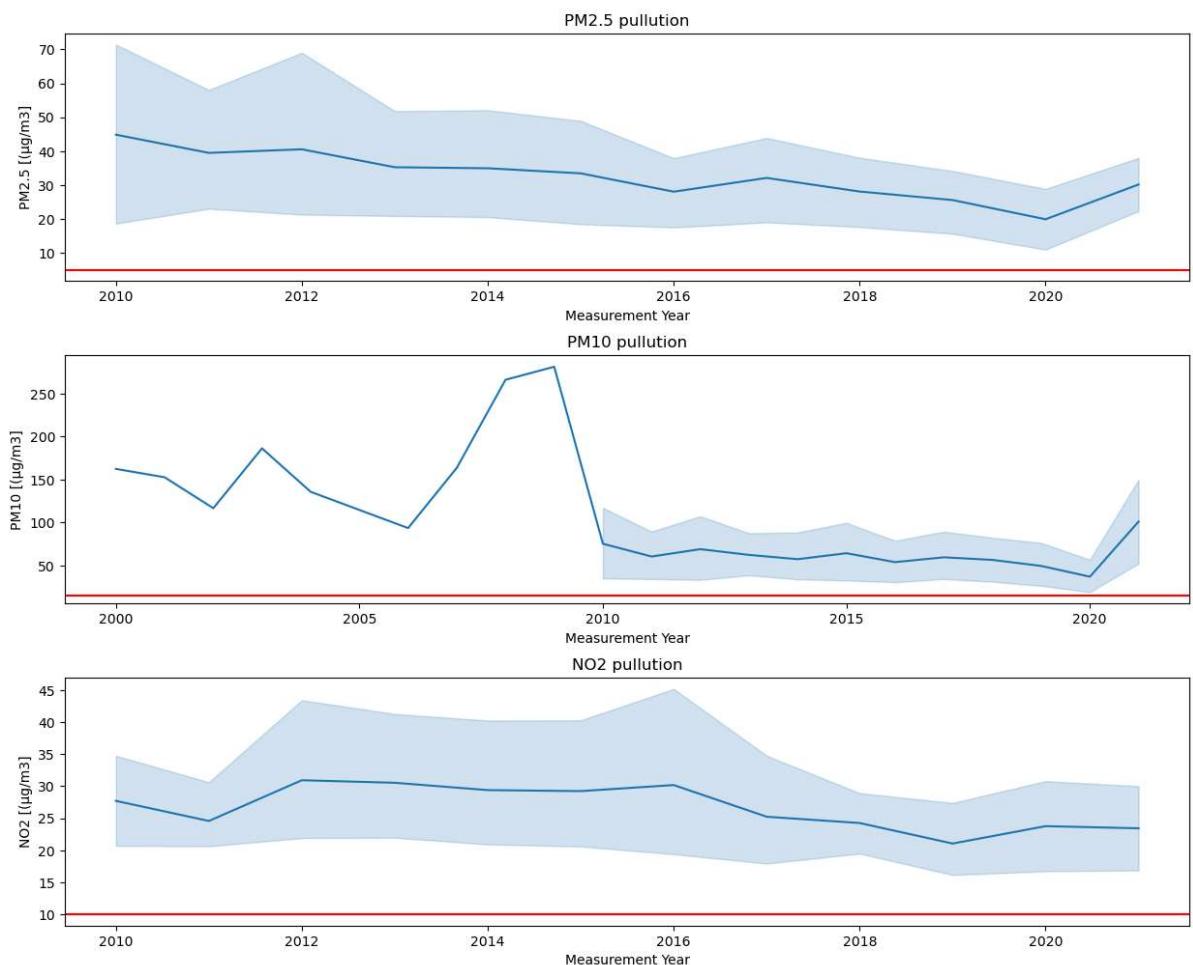
sns.lineplot(grouped_df_pm10, x="Measurement Year", y=grouped_df_pm10.values, ax=ax[1])
ax[1].axhline(y=15, color="r")
ax[1].set(title="PM10 pullution", ylabel="PM10 [ $(\mu\text{g}/\text{m}^3)$ ])

sns.lineplot(grouped_df_no2, x="Measurement Year", y=grouped_df_no2.values, ax=ax[2])
ax[2].axhline(y=10, color="r")
ax[2].set(title="NO2 pullution", ylabel="NO2 [ $(\mu\text{g}/\text{m}^3)$ ])

plt.suptitle("Air pollution across years for all data", fontsize=16)
plt.subplots_adjust(hspace=0.3)
plt.show()

```

Air pollution across years for all data



Dla całości danych przedstawiono wartości współczynników zanieczyszczenia powietrza dla kolejnych lat. Jasnoniebieska obwoluta pokazuje przedziały ufności, a raczej w tym przypadku - maksymalne i minimalne wartości średnich dla każdego z regionów. Czerwoną

linią oznaczona normę przyjętą przez WHO dla każdego ze współczynników. W przypadku każdego z nich widać generalną (delikatną) tendencję spadkową w ostatnich latach, jednak z przyrostem w ostatnim badanym roku dla PM2.5 oraz PM10.

```
In [ ]: i = 0
j = 0

fig, ax = plt.subplots(3, 2, figsize=(15, 12))

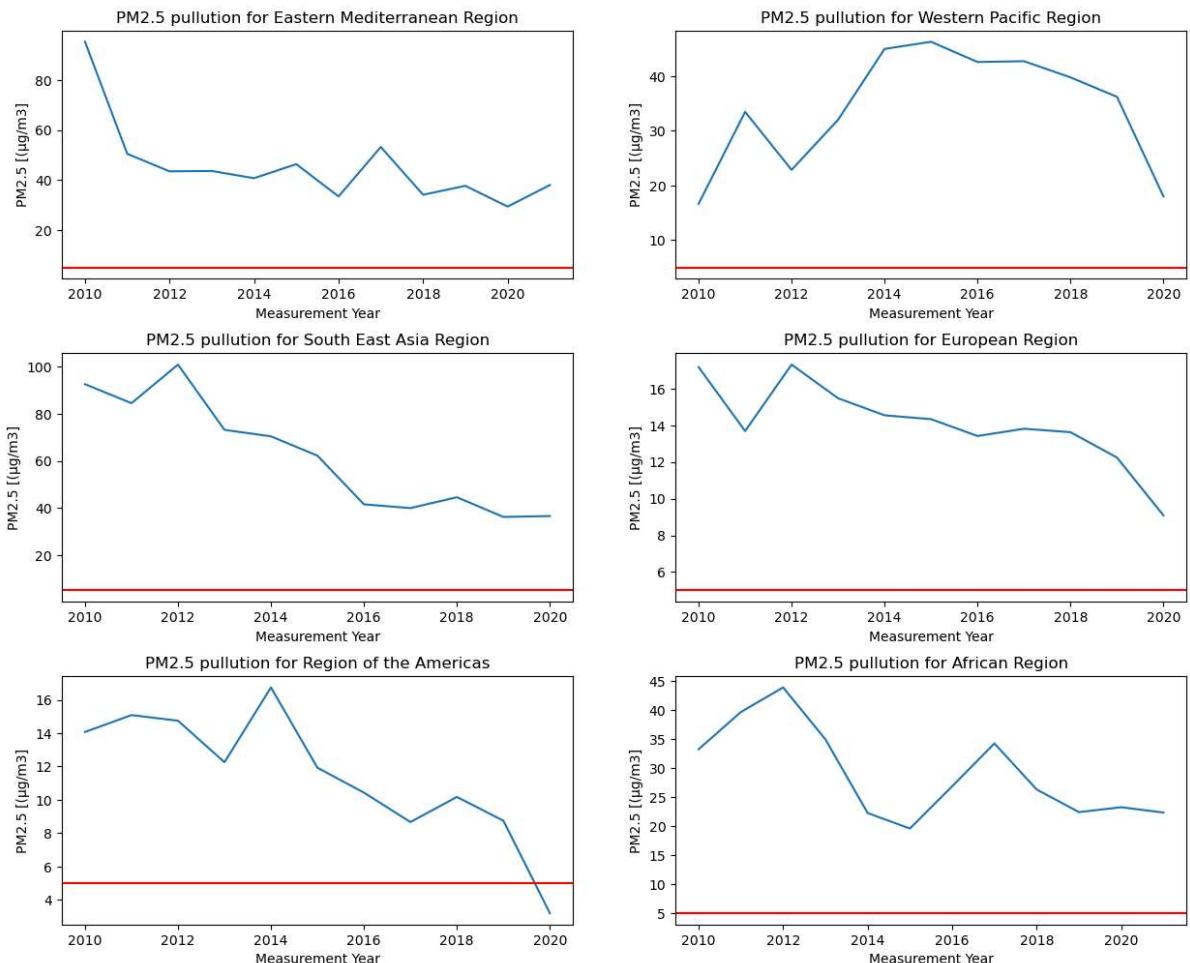
for r in data["WHO Region"].unique():

    sub_grouped_df = grouped_df_pm25[
        grouped_df_pm25.index.get_level_values("WHO Region").isin([r])
    ]

    sns.lineplot(
        sub_grouped_df,
        x="Measurement Year",
        y=sub_grouped_df.values,
        ax=ax[i, j],
    )
    ax[i, j].axhline(y=5, color="r")
    ax[i, j].set(title=f"PM2.5 pollution for {r}", ylabel="PM2.5 [(\mu g/m³")]
    i += 1
    i %= 3
    j += 1
    j %= 2

plt.suptitle("PM2.5 pollution across years for all regions", fontsize=16)
plt.subplots_adjust(hspace=0.3)
plt.show()
```

PM2.5 pollution across years for all regions



Badając samo PM2.5 dla każdego z regionów można wykryć wyraźny ujemny trend w regionach Azji Południowo-Wschodniej, Europie, Amerykach, a w ostatnich latach także w regionie Zachodniego Pacyfiku. W regionie afrykańskim wartości zmieniają się w sposób dość losowy. Niestety, w każdym z regionów norma jest znaczająco przekroczona (oprócz ostatniego roku w Amerykach). Zazwyczaj, średnie wartości zanieczyszczeń PM2.5 znacznie różniły się między regionami na korzyść Europy i Ameryk.

```
In [ ]: i = 0
j = 0

fig, ax = plt.subplots(3, 2, figsize=(15, 12))

for r in data["WHO Region"].unique():

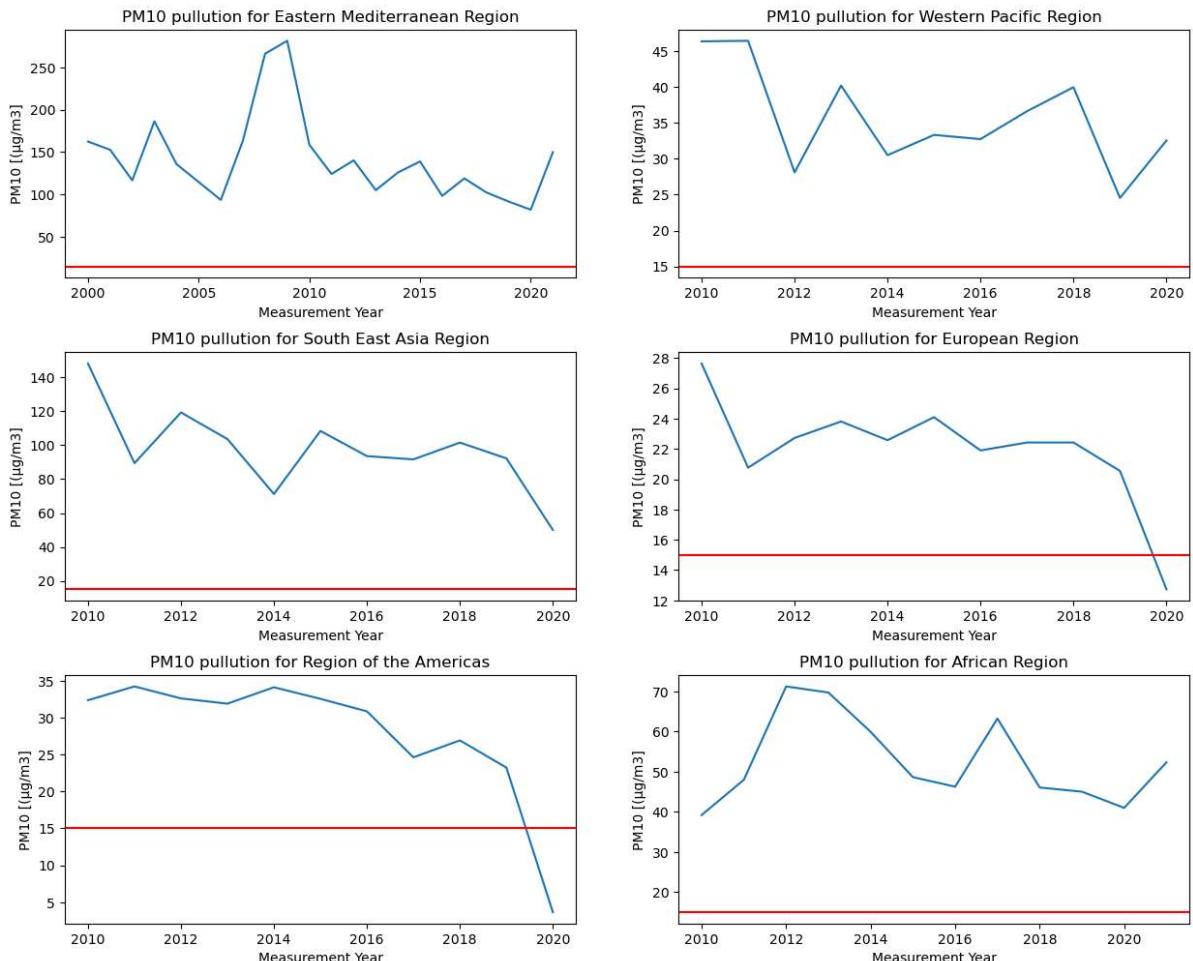
    sub_grouped_df = grouped_df_pm10[
        grouped_df_pm10.index.get_level_values("WHO Region").isin([r])
    ]

    sns.lineplot(
        sub_grouped_df,
        x="Measurement Year",
        y=sub_grouped_df.values,
        ax=ax[i, j],
    )
    ax[i, j].axhline(y=15, color="r")
    ax[i, j].set(title=f"PM10 pollution for {r}", ylabel="PM10 [(µg/m³)]")
    i += 1
    i %= 3
```

```
j += 1
j %= 2

plt.suptitle("PM10 pollution across years for all regions", fontsize=16)
plt.subplots_adjust(hspace=0.3)
plt.show()
```

PM10 pollution across years for all regions



W przypadku PM10 ujemny trend występuje jedynie w Europie oraz Amerykach. Poza tymi regionami utrzymuje się on na mniej więcej stałym poziomie, ze znaczącym skokiem w okolicach roku 2009 dla regionu Wchodniego Morza Śródziemnego. Norma nie została średnio przekroczona tylko w ostatnich latach dla Europy i Ameryk. Warto zwrócić uwagę na podziałkę osi Y. W rejonach Wschodniego Morza Śródziemnego oraz Azji Południowo-Wschodniej zanieczyszczenie przyjmowało ogromne wartości w porównaniu z pozostałymi.

```
In [ ]: i = 0
j = 0

fig, ax = plt.subplots(3, 2, figsize=(15, 12))

for r in data["WHO Region"].unique():

    sub_grouped_df = grouped_df_no2[
        grouped_df_no2.index.get_level_values("WHO Region").isin([r])
    ]

    sns.lineplot(
        sub_grouped_df,
        x="Measurement Year",
        y="PM10 [µg/m³]",
        color="blue"
    )

    if j == 0:
        plt.title(f"PM10 pollution for {r} region")
    else:
        plt.title("")

    j += 1
```

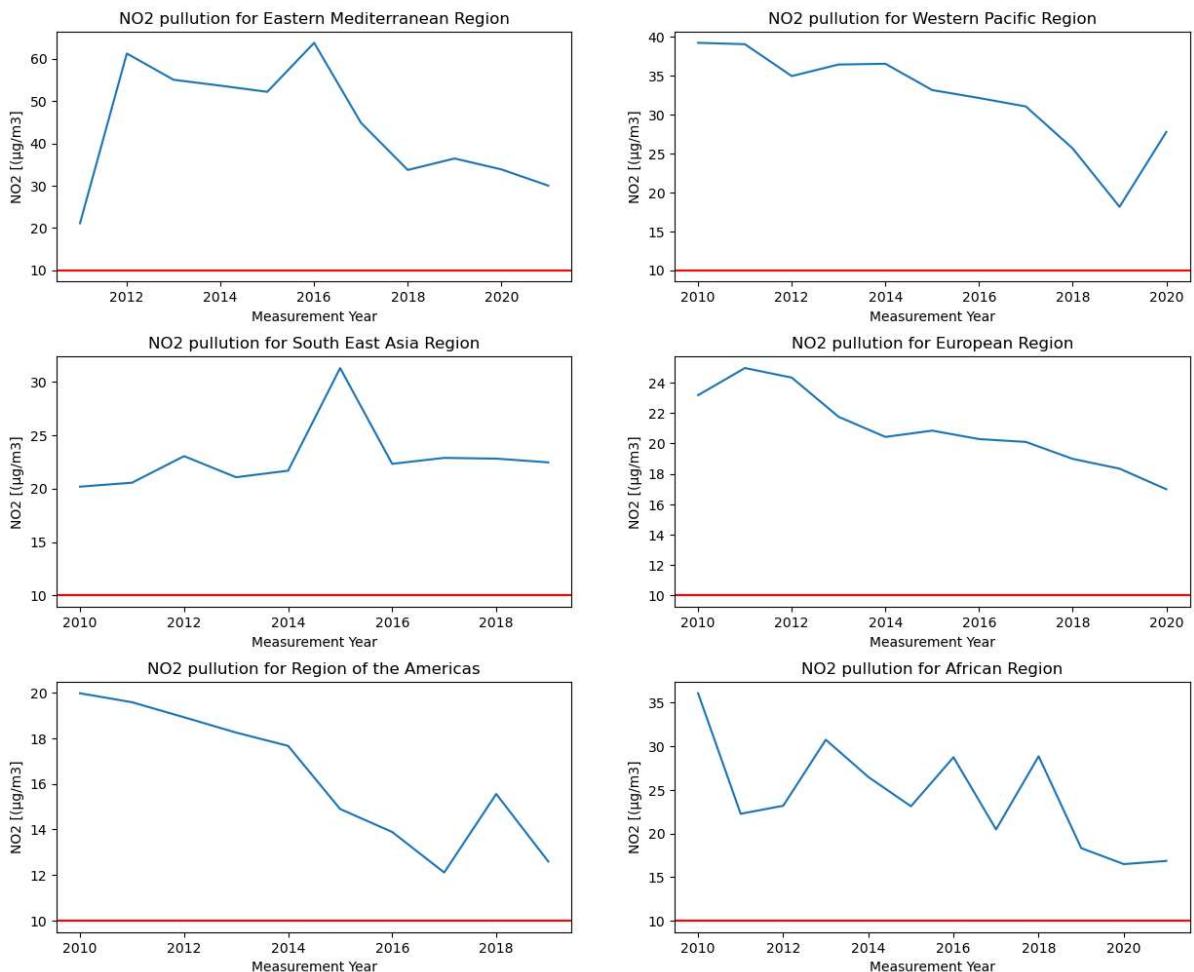
```

        y=sub_grouped_df.values,
        ax=ax[i, j],
    )
    ax[i, j].axhline(y=10, color="r")
    ax[i, j].set(title=f"NO2 pollution for {r}", ylabel="NO2 [(µg/m³)]")
    i += 1
    i %= 3
    j += 1
    j %= 2

plt.suptitle("NO2 pollution across years for all regions", fontsize=16)
plt.subplots_adjust(hspace=0.3)
plt.show()

```

NO2 pollution across years for all regions



Wartości NO₂ charakteryzuje niewielki ujemny trend dla każdego z regionów oprócz Azji Południowo-Wschodniej, gdzie istnieje niewielki trend wzrostowy. Najwyższe średnie wartości były rejestrowane w rejonie Wschodniego Morza Śródziemnego.

Most and least polluted regions

PM2.5

```
In [ ]: grouped_region_pm25 = data.groupby("WHO Region")["PM2.5"].mean()
pd.DataFrame(grouped_region_pm25).sort_values(by="PM2.5")
```

Out[]:

WHO Region	PM2.5
Region of the Americas	11.38941
European Region	14.002193
African Region	27.25712
Eastern Mediterranean Region	38.063258
Western Pacific Region	40.324123
South East Asia Region	42.841212

Zgodnie z poprzednimi spostrzeżeniami, najniższe średnie zanieczyszczenie PM2.5 występuje w regionie Ameryk oraz Europy, a najwyższe w rejonie Zachodniego Pacyfiku oraz Azji Południowo-Wschodniej.

```
In [ ]: grouped_region_pm25 = data.groupby("WHO Region")["PM2.5"].median()
pd.DataFrame(grouped_region_pm25).sort_values(by="PM2.5")
```

Out[]:

WHO Region	PM2.5
Region of the Americas	8
European Region	12.81
African Region	22.86
Eastern Mediterranean Region	33
South East Asia Region	36.5
Western Pacific Region	39.11

Mediana zanieczyszczeń PM2.5 nie odbiega znacząco od średniej, co oznacza, że całe regiony są równomierne zanieczyszczone PM2.5.

PM10

```
In [ ]: grouped_region_pm10 = data.groupby("WHO Region")["PM10"].mean()
pd.DataFrame(grouped_region_pm10).sort_values(by="PM10")
```

Out[]:

WHO Region	PM10
European Region	23.013879
Region of the Americas	30.677423
Western Pacific Region	32.781478
African Region	51.906625
South East Asia Region	94.626582
Eastern Mediterranean Region	121.317951

Rejony najmniej zanieczyszczone cząsteczkami PM10 są analogiczne do tych wspomianych w przypadku PM2.5. Jednakże tutaj Europa ma pod tym względem lepsze powietrze niż Ameryki, które średnio tylko odrobinę przegrywają z Zachodnim Pacyfikiem, który miał najgorszy wynik PM2.5. Zdecydowanie najbardziej zanieczyszczonym regionem PM10 jest Wschodnie Morze Śródziemne.

```
In [ ]: grouped_region_pm10 = data.groupby("WHO Region")["PM10"].median()
pd.DataFrame(grouped_region_pm10).sort_values(by="PM10")
```

Out[]:

WHO Region	PM10
European Region	20.74
Western Pacific Region	22.4
Region of the Americas	26.085
African Region	41.39
South East Asia Region	84
Eastern Mediterranean Region	109.8

Wyniki bazujące na medianie nieco odbiegają od tych opartych o średnią, co wskazują na istotne różnice w zanieczyszczeniach wewnętrz regionów z pojedynczymi krajami/miastami znacząco zanieczyszczonymi.

NO2

```
In [ ]: grouped_region_no2 = data.groupby("WHO Region")["NO2"].mean()
pd.DataFrame(grouped_region_no2).sort_values(by="NO2")
```

Out[]:

WHO Region	NO2
Region of the Americas	15.75789
European Region	20.38904
South East Asia Region	22.099986
African Region	23.160242
Western Pacific Region	31.970357
Eastern Mediterranean Region	45.715101

Najczystsze regiony pod względem NO2 to również Europa i Ameryki, a najbardziej zanieczyszczona - Zachodni Pacyfik oraz Wschodnie Morze Śródziemne.

```
In [ ]: grouped_region_no2 = data.groupby("WHO Region")["NO2"].median()
pd.DataFrame(grouped_region_no2).sort_values(by="NO2")
```

Out[]:

WHO Region	NO2
Region of the Americas	14
European Region	18.98
South East Asia Region	19.33
African Region	19.51
Western Pacific Region	33
Eastern Mediterranean Region	38.05

Mediana prezentuje się tutaj bardzo podobnie do średniej, tak jak w przypadku PM2.5.

Stations analysis

```
In [ ]: grouped_region_no_stations = data.groupby("WHO Region")[
    "Total number of stations"
].mean()
pd.DataFrame(grouped_region_no_stations).sort_values(by="Total number of stations")
```

Out[]:

WHO Region	Total number of stations
Region of the Americas	1.598307
European Region	2.093603
Western Pacific Region	2.484848
South East Asia Region	2.516116
African Region	2.670455
Eastern Mediterranean Region	2.801105

Co ciekawe najmniejszą średnią liczbę stacji posiada region Ameryk, a największą - region Wschodniego Morza Śródziemnego. Należy mieć tu jednak na uwadze liczne braki w tej kolumnie.

```
In [ ]: grouped_region_no_stations = data.groupby("WHO Region")[
    "Total number of stations"
].median()
pd.DataFrame(grouped_region_no_stations).sort_values(by="Total number of stations")
```

Out[]:

WHO Region	Total number of stations
Eastern Mediterranean Region	1
European Region	1
Region of the Americas	1
Western Pacific Region	1
African Region	2
South East Asia Region	2

Medianą wyników prezentuje się już bardziej jednostajnie, wciąż mniej więcej utrzymując kolejność regionów.

Country-based analysis

```
In [ ]: fig, ax = plt.subplots(2, 2, figsize=(40, 32))

sns.histplot(data=data, x="WHO Country Name", ax=ax[0, 0], shrink=0.7)
ax[0, 0].set_xticklabels(ax[0, 0].get_xticklabels(), rotation=90)
ax[0, 0].set(title="Number of measurements vs Country")

sns.boxplot(data=data, y="NO2", x="WHO Country Name", ax=ax[0, 1], width=0.7)
ax[0, 1].axhline(y=10, color="r")
ax[0, 1].grid(axis="x")
ax[0, 1].set_xticklabels(ax[0, 1].get_xticklabels(), rotation=90)
```

```

ax[0, 1].set(title="Distribution of NO2 values based on Country for all years")

sns.boxplot(data=data, y="PM2.5", x="WHO Country Name", ax=ax[1, 0], width=0.7)
ax[1, 0].axhline(y=5, color="r")
ax[1, 0].grid(axis="x")
ax[1, 0].set_xticklabels(ax[1, 0].get_xticklabels(), rotation=90)
ax[1, 0].set(title="Distribution of PM2.5 values based on Country for all years")

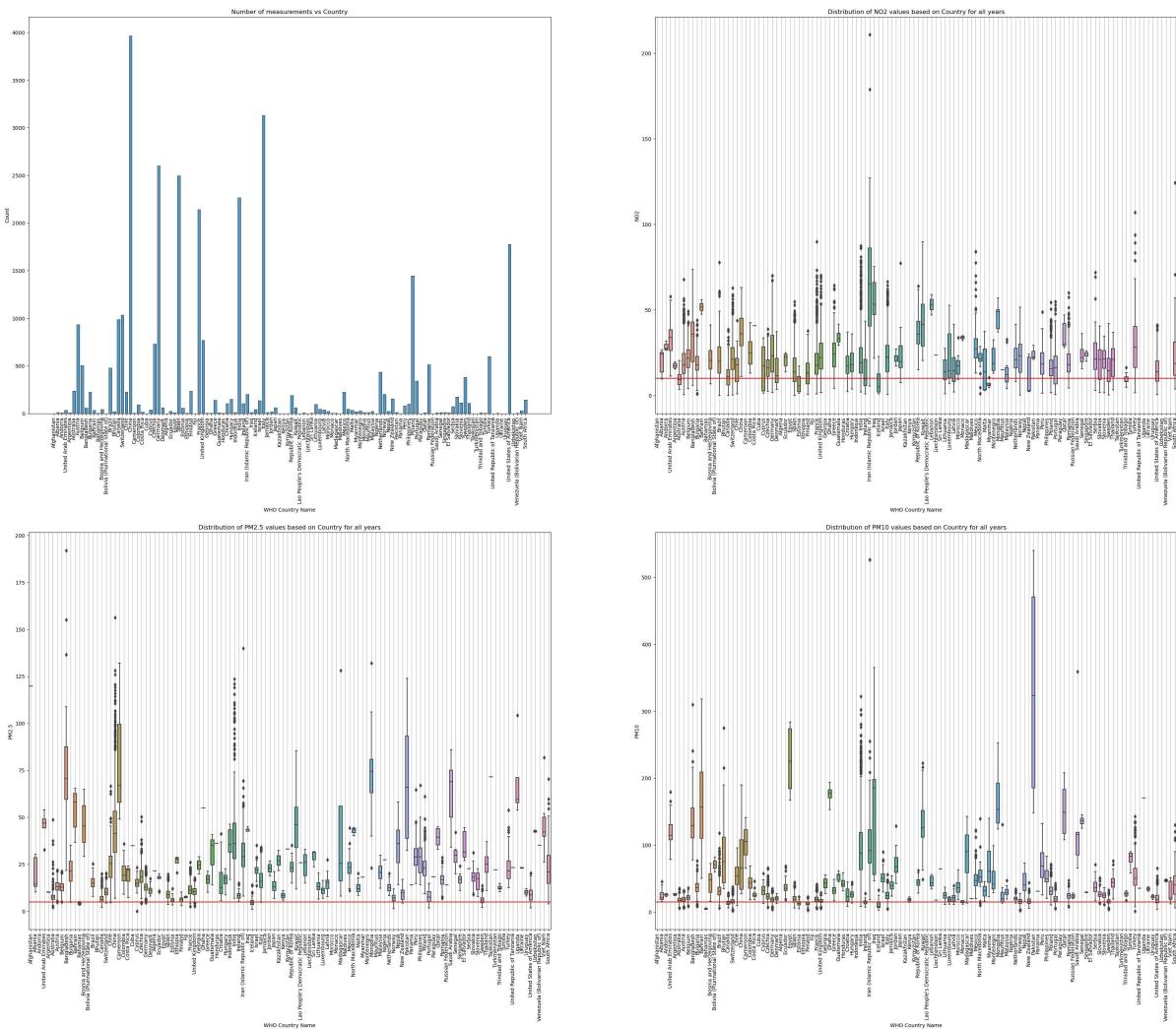
sns.boxplot(data=data, y="PM10", x="WHO Country Name", ax=ax[1, 1], width=0.7)
ax[1, 1].axhline(y=15, color="r")
ax[1, 1].grid(axis="x")
ax[1, 1].set_xticklabels(ax[1, 1].get_xticklabels(), rotation=90)
ax[1, 1].set(title="Distribution of PM10 values based on Country for all years")

plt.suptitle("Data distribution based on Country", fontsize=20)
plt.subplots_adjust(hspace=0.3)
plt.show()

```

C:\Users\adamp\AppData\Local\Temp\ipykernel_40448\2476146839.py:4: UserWarning: FixedFormatter should only be used together with FixedLocator
 ax[0, 0].set_xticklabels(ax[0, 0].get_xticklabels(), rotation=90)

Data distribution based on Country



Przechodząc do analizy opartej o kraje, ponownie rozpoczęto od zbadania rozkładu danych. Z histogramu w lewym górnym rogu widać, że w liczbie pomiarów zdecydowanie przodują Chiny znane z ogromnego zanieczyszczenia powietrza. Zaraz za nimi: (co ciekawe) Włochy, Niemcy, Hiszpania i Indie.\ Pod względem rozkładu wartości zanieczyszczeń PM2.5 można wyróżnić kraje z niewielkimi IQR jak np. Grecja, Malta czy USA, a także z bardzo dużymi,

twarzyszącej im przy tym wysokiej medianie, tutaj: Bangladesz, Chiny, Kamerun, Mongolia, Pakista, Senegal. Chiny oraz Indie mają przy tym bardzo wiele wartości odstających powyżej 95 percentyla, co sugeruje, że istnieje wiele miast z bardzo wysokim zanieczyszczeniem PM2.5.\ Badając rozkład PM10 uwagę zwraca na siebie szczególnie Pakistan z ogromnym IQR oraz bardzo wysoką medianą. Ponadto duże zanieczyszczenia mają także: Bahrajn, Egipt, Ghana, Irak, Mongolia i Katar.\ W przypadku NO₂ wartości są już mniej zróżnicowane, jednak przoduje tutaj Iran, Irak oraz Bahrajn - wszystkie położone geograficznie bardzo blisko siebie.

Most and least polluted countries (across all years)

Pokuszono się o wyspecyfikowanie najmniej oraz najbardziej zanieczyszczonych państw pod względem każdego ze współczynników, zarówno pod względem średniej oraz mediany.

PM2.5

```
In [ ]: grouped_country_pm25 = data.groupby("WHO Country Name")["PM2.5"].mean()
grouped_country_pm25 = (
    pd.DataFrame(grouped_country_pm25).sort_values(by="PM2.5").dropna()
)
grouped_country_pm25.index[:3].to_list()

Out[ ]: ['Bahamas', 'Iceland', 'Estonia']
```

```
In [ ]: grouped_country_pm25.index[-3:].to_list()

Out[ ]: ['Bangladesh', 'Cameroon', 'Afghanistan']
```

```
In [ ]: grouped_country_pm25 = data.groupby("WHO Country Name")["PM2.5"].median()
grouped_country_pm25 = (
    pd.DataFrame(grouped_country_pm25).sort_values(by="PM2.5").dropna()
)
grouped_country_pm25.index[:3].to_list()

Out[ ]: ['Bahamas', 'Iceland', 'Estonia']
```

```
In [ ]: grouped_country_pm25.index[-3:].to_list()

Out[ ]: ['Tajikistan', 'Mongolia', 'Afghanistan']
```

Mediana nie wprowadziła tu znacznych zmian w stosunku do średniej, przez co można uznać Bahamy, Islandię i Estonię za najmniej zanieczyszczone kraje cząsteczkami PM2.5, a Afganistan wraz z towarzyszącymi mu na listach państwami - za najbardziej.

PM10

```
In [ ]: grouped_country_pm10 = data.groupby("WHO Country Name")["PM10"].mean()
grouped_country_pm10 = (
    pd.DataFrame(grouped_country_pm10).sort_values(by="PM10").dropna()
)
grouped_country_pm10.index[:3].to_list()

Out[ ]: ['Bahamas', 'Iceland', 'Estonia']
```

```
In [ ]: grouped_country_pm10.index[-3:].to_list()
```

```
Out[ ]: ['Ghana', 'Egypt', 'Pakistan']
```

```
In [ ]: grouped_country_pm10 = data.groupby("WHO Country Name")["PM10"].median()
grouped_country_pm10 = (
    pd.DataFrame(grouped_country_pm10).sort_values(by="PM10").dropna()
)
grouped_country_pm10.index[:3].to_list()
```

```
Out[ ]: ['Bahamas', 'Iceland', 'Finland']
```

```
In [ ]: grouped_country_pm25.index[-3:].to_list()
```

```
Out[ ]: ['Tajikistan', 'Mongolia', 'Afghanistan']
```

Pod względem PM10 również najlepiej wypadają Bahamy, Islandia ora Estonia, a biorąc pod uwagę medianę, także Finlandia. Najbardziej zanieczyszczone kraje to ponownie: Afganistan, Mongolia, Tadżykistan oraz Ghana i Egipt.

NO2

```
In [ ]: grouped_country_no2 = data.groupby("WHO Country Name")["NO2"].mean()
grouped_country_no2 = pd.DataFrame(grouped_country_no2).sort_values(by="NO2").dropna()
grouped_country_no2.index[:3].to_list()
```

```
Out[ ]: ['Estonia', 'Myanmar', 'Iceland']
```

```
In [ ]: grouped_country_no2.index[-3:].to_list()
```

```
Out[ ]: ['Lebanon', 'Iraq', 'Iran (Islamic Republic of)']
```

```
In [ ]: grouped_country_no2 = data.groupby("WHO Country Name")["NO2"].median()
grouped_country_no2 = pd.DataFrame(grouped_country_no2).sort_values(by="NO2").dropna()
grouped_country_no2.index[:3].to_list()
```

```
Out[ ]: ['New Zealand', 'Iceland', 'Estonia']
```

```
In [ ]: grouped_country_no2.index[-3:].to_list()
```

```
Out[ ]: ['Lebanon', 'Iraq', 'Iran (Islamic Republic of)']
```

Badając wartości NO2, ponownie najczystsza jest Estonia i Islandia, a dodatkowo Birma i Nowa Zelandia, a najbardziej zanieczyszczone: Iran, Irak i Liban.

Biggest progress and regress in air pollution

Zbadano, czy na przestrzeni lat jakieś państwa odnotowały postęp w zakresie czystego powietrza, czy też wręcz przeciwnie. Dodatnia wartość różnicy oznacza progress, a ujemna - regress.

PM2.5

```
In [ ]: grouped = data.groupby(["WHO Country Name", "Measurement Year"]).mean(numeric_only=True)
grouped_pm25 = grouped.dropna(subset="PM2.5")
```

```

earliest = grouped_pm25.groupby(level=0)["PM2.5"].first()
latest = grouped_pm25.groupby(level=0)["PM2.5"].last()

difference = earliest - latest # + --> progress; - --> regress

difference

```

Out[]: 10 entries per page

Search:

PM2.5	
WHO Country Name	
Afghanistan	0
Albania	11.47
Algeria	0
Argentina	-0.15
Australia	7.731667
Austria	7.270416
Bahamas	2.03
Bahrain	14.5075
Bangladesh	6.17
Belgium	5.796286

Showing 1 to 10 of 106 entries

« < 1 2 3 4 5 ... 11 > »

In []: difference.sort_values()[:3]

Out[]:

PM2.5	
WHO Country Name	
Indonesia	-29.35
Madagascar	-24
Mongolia	-13.966667

In []: difference.sort_values()[-3:]

Out[]:

WHO Country Name	PM2.5
Viet Nam	42.405
Uganda	44.15
Pakistan	63.043333

Jak widać największy regres pod względem PM2.5 odnotowały: Indonezja, Madagaskar oraz Mongolia, a progres: Wietnam, Uganda i Pakistan.

PM10

```
In [ ]: grouped_pm10 = grouped.dropna(subset="PM10")

earliest = grouped_pm10.groupby(level=0)[ "PM10" ].first()
latest = grouped_pm10.groupby(level=0)[ "PM10" ].last()

difference = earliest - latest # + --> progress; - --> regres

difference
```

Out[]: 10 entries per page

Search:

WHO Country Name	PM10
Albania	-13.91
Andorra	2.92
Argentina	2.37
Australia	0.902971
Austria	7.610359
Bahamas	2.16
Bahrain	46.543333
Bangladesh	26.70625
Belgium	7.428993
Bhutan	-111.5

Showing 1 to 10 of 102
entries

« < 1 2 3 4 5 ... 11 > »

In []: difference.sort_values()[:3]

Out[]:

WHO Country Name	PM10
Bhutan	-111.5
Egypt	-93.5
Madagascar	-93

In []: difference.sort_values()[-3:]

Out[]:

WHO Country Name	PM10
China	117.516667
Iraq	153.63
Pakistan	247.163333

Pod względem cząsteczek PM10 ogromny regres zarejestrowano w: Bhutanie, Egipcie i na Madagaskarze, a równie duży progres w: Chinach, Iraku i Pakistanie.

NO2

```
In [ ]: grouped_no2 = grouped.dropna(subset="NO2")
earliest = grouped_no2.groupby(level=0)[ "NO2" ].first()
latest = grouped_no2.groupby(level=0)[ "NO2" ].last()
difference = earliest - latest # + --> progress; - --> regres
difference
```

Out[]: 10 ✓ entries per page

Search:

NO2	
WHO Country Name	
Albania	-6.115
Andorra	0.63
Argentina	-2.9
Australia	-1.861667
Austria	5.462345
Bahrain	0
Bangladesh	-3.527143
Belgium	6.281629
Bosnia and Herzegovina	4.726667
Brazil	7.590804

Showing 1 to 10 of 77
entries

« < 1 2 3 4 5 ... 8 >

»

In []: difference.sort_values()[:3]

Out[]:

NO2	
WHO Country Name	
Mexico	-21.44
Iran (Islamic Republic of)	-18.68125
United Arab Emirates	-12.878333

In []: difference.sort_values()[-3:]

Out[]:

NO2	
WHO Country Name	
Latvia	21.58
South Africa	22.011538
Kuwait	29.182364

W przypadku wartości NO2 regres odnotowano w: Meksyku, Iranie i Zjednoczonych Emiratach Arabskich, a postęp w: Litwie, Południowej Afryce oraz Kuwejcie.

Stations analysis

```
In [ ]: summed_stations_df = data.groupby("WHO Country Name")["Total number of stations"].sum()
mean_stations_df = data.groupby("WHO Country Name")["Total number of stations"].mean()

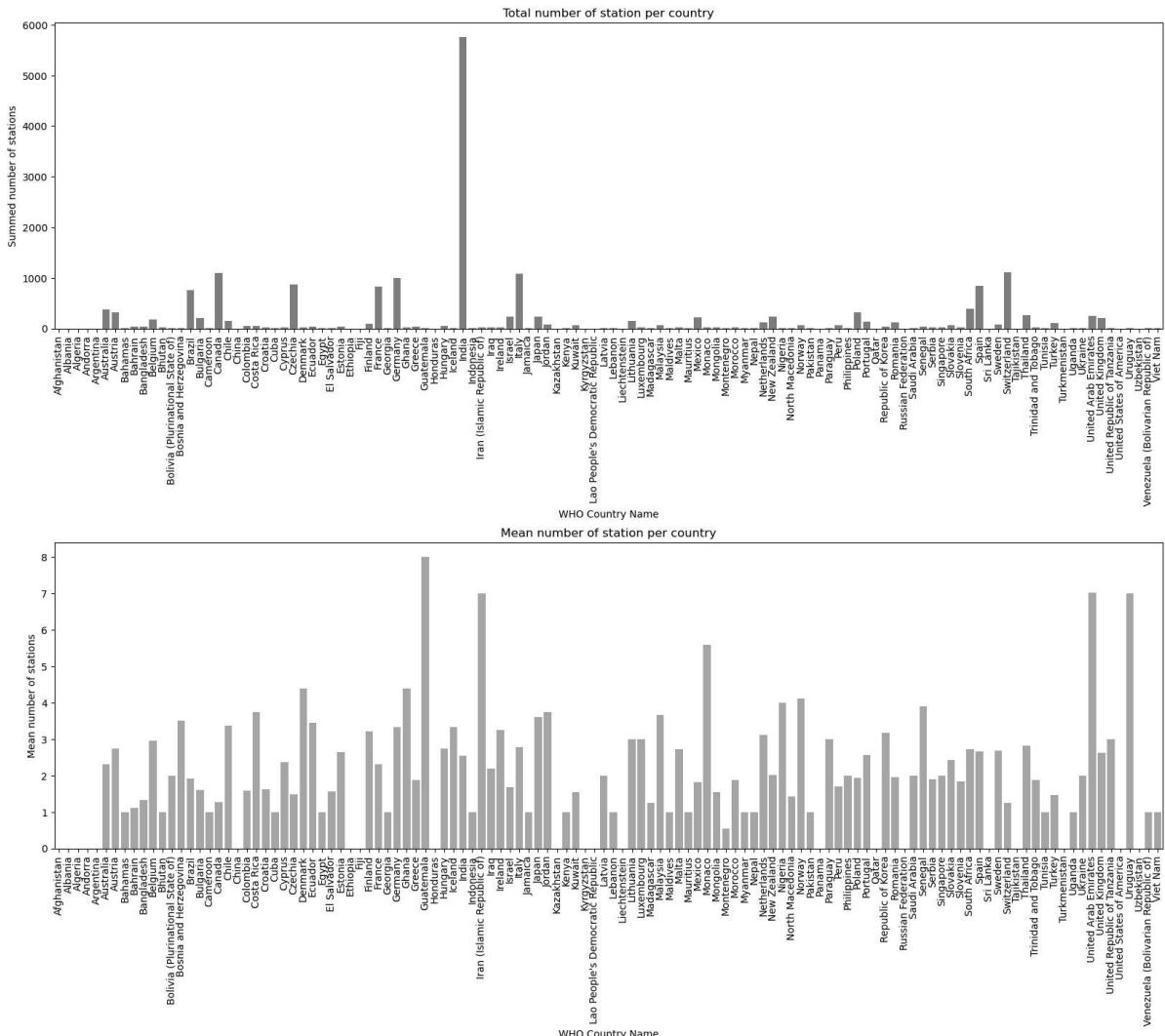
fig, ax = plt.subplots(2, 1, figsize=(20, 15))

sns.barplot(
    x=summed_stations_df.index, y=summed_stations_df.values, ax=ax[0], color="gray"
)
ax[0].set_xticklabels(ax[0].get_xticklabels(), rotation=90)
ax[0].set(
    title="Total number of station per country", ylabel="Summed number of stations"
)

sns.barplot(
    x=mean_stations_df.index, y=mean_stations_df.values, ax=ax[1], color="darkgrey"
)
ax[1].set_xticklabels(ax[1].get_xticklabels(), rotation=90)
ax[1].set(title="Mean number of station per country", ylabel="Mean number of statics")

plt.suptitle("Stations distribution", fontsize=16)
plt.subplots_adjust(hspace=0.7)
plt.show()
```

Stations distribution



Podjęto się również analizy ilości stacji pomiarowych w każdym z krajów, zarówno pod względem ich całkowitej ilości, jak i średniej. W sumarycznej ilości wyróżniają się Indie, a poza nimi, choć znacznie mniej: Szwajcaria, Włochy, Niemcy, Czechy, Kanada, Francja i Hiszpania. W średniej ilości stacji prowadzi za to Gwatemala, a zaraz za nią są: Iran, Zjednoczone Emiraty Arabskie oraz Urugwaj. Należy pamiętać, że dane dotyczące ilości stacji mogą mieć wysoki bias, ze względu na dużą ilość wartości brakujących typu MCAR.

Air pollution map

PM2.5

```
In [ ]: grouped_country_pm25 = data.groupby(["WHO Country Name", "ISO3"])["PM2.5"].mean()
df_grouped_country_pm25 = grouped_country_pm25.reset_index()

fig = px.choropleth(
    df_grouped_country_pm25,
    locations="ISO3",
    locationmode="ISO-3",
    color="PM2.5",
    hover_name="WHO Country Name",
    color_continuous_scale=px.colors.sequential.Sunsetdark,
)

fig.update_layout(
    title_text="Average concentration of PM2.5 across all years",
)
fig.show()
```

Mapa świata z naniesionymi wartościami zanieczyszczeń PM2.5 wyraźnie przedstawia całą Azję Południową jako najbardziej zanieczyszczoną. Najlepiej prezentuje się Ameryka Północna oraz Europa Północna i Zachodnia.

PM10

```
In [ ]: grouped_country_pm10 = data.groupby(["WHO Country Name", "ISO3"])["PM10"].mean()
df_grouped_country_pm10 = grouped_country_pm10.reset_index()

fig = px.choropleth(
    df_grouped_country_pm10,
    locations="ISO3",
    locationmode="ISO-3",
    color="PM10",
    hover_name="WHO Country Name",
    color_continuous_scale=px.colors.sequential.Sunsetdark,
)

fig.update_layout(
    title_text="Average concentration of PM10 across all years",
)
fig.show()
```

Podobnie jak mapa dla PM2.5, mapa dla PM10 pokazuje te same siedliska zanieczyszczeń, dodając przy tym niektóre państwa afrykańskie. Należy zauważyć, jak przerażająco

prezentuje się ona dla Pakistanu.

NO2

```
In [ ]: grouped_country_no2 = data.groupby(["WHO Country Name", "ISO3"])["NO2"].mean()
df_grouped_country_no2 = grouped_country_no2.reset_index()

fig = px.choropleth(
    df_grouped_country_no2,
    locations="ISO3",
    locationmode="ISO-3",
    color="NO2",
    hover_name="WHO Country Name",
    color_continuous_scale=px.colors.sequential.Sunsetdark,
)

fig.update_layout(
    title_text="Average concentration of NO2 across all years",
)

fig.show()
```

Dla NO₂, rozkład pomiędzy krajami jest już nieco bardziej równomiarny, ze zdecydowanie wyróżniającym się Iranem, a także Mongolią i Chinami.

City-based analysis

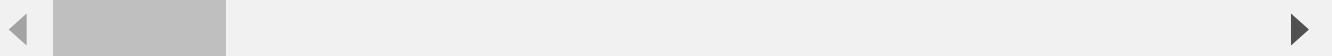
Correlation (point-biserial)

```
In [ ]: cities_corr_df = data.copy()
cities_corr_df["is_India"] = cities_corr_df["WHO Country Name"].apply(
    lambda x: 1 if x == "India" else 0
)
cities_corr_df["is_China"] = cities_corr_df["WHO Country Name"].apply(
    lambda x: 1 if x == "China" else 0
)
cities_corr_df
```

Out[]: 10 ✓ entries per page

WHO Region	ISO3	WHO Country Name
Eastern Mediterranean Region	AFG	Afghanistan
European Region	ALB	Albania

Showing 1 to 10 of 512 entries (downsampled from 32,191x16 to 512x16 as maxBytes=65536)



Z racji tego, że spodziewano się, iż miasta chińskie oraz indyjskie będą wyróżniały się większym zanieczyszczeniem, utworzono zmienne kategoryczne mówiące o przynależności do nich.

```
In [ ]: for i in ["is_India", "is_China"]:
    print(f"\n{i}:")
    for c in corr_cols:
        temp_df = cities_corr_df.dropna(subset=c)
        print(f"{i} vs {c}: {stats.pointbiserialr(temp_df[i], temp_df[c])}")
```

```

is_India:
is_India vs Measurement Year: SignificanceResult(statistic=-0.033722921784423314,
pvalue=1.4305626692267871e-09)
is_India vs PM2.5: SignificanceResult(statistic=0.16586006739676543, pvalue=2.8436
607862728626e-93)
is_India vs PM10: SignificanceResult(statistic=0.5670443433607244, pvalue=0.0)
is_India vs NO2: SignificanceResult(statistic=0.03616984654129489, pvalue=7.020482
2265107e-08)
is_India vs PM2.5 temporal coverage: SignificanceResult(statistic=-0.3074306764651
945, pvalue=4.829027810058524e-159)
is_India vs PM10 temporal coverage: SignificanceResult(statistic=-0.15350792280808
412, pvalue=9.846775449244917e-30)
is_India vs NO2 temporal coverage: SignificanceResult(statistic=-0.336410605955611
56, pvalue=0.0)
is_India vs Total number of stations: SignificanceResult(statistic=0.0846226099494
6087, pvalue=2.148838202164032e-15)

```

```

is_China:
is_China vs Measurement Year: SignificanceResult(statistic=0.22373939837460843, p
value=0.0)
is_China vs PM2.5: SignificanceResult(statistic=0.6725632858954628, pvalue=0.0)
is_China vs PM10: SignificanceResult(statistic=0.06486150067896079, pvalue=3.98289
492759937e-21)
is_China vs NO2: SignificanceResult(statistic=0.14946638223898057, pvalue=4.439268
1451651326e-111)
is_China vs PM2.5 temporal coverage: SignificanceResult(statistic=0.14599731680496
67, pvalue=5.979949758409239e-36)
is_China vs PM10 temporal coverage: SignificanceResult(statistic=nan, pvalue=nan)
is_China vs NO2 temporal coverage: SignificanceResult(statistic=nan, pvalue=nan)
is_China vs Total number of stations: SignificanceResult(statistic=nan, pvalue=na
n)

```

```
c:\Users\adamp\anaconda3\envs\ml_moje\lib\site-packages\scipy\stats\_stats_py.py:4
424: ConstantInputWarning:
```

An input array is constant; the correlation coefficient is not defined.

```
c:\Users\adamp\anaconda3\envs\ml_moje\lib\site-packages\scipy\stats\_stats_py.py:4
424: ConstantInputWarning:
```

An input array is constant; the correlation coefficient is not defined.

```
c:\Users\adamp\anaconda3\envs\ml_moje\lib\site-packages\scipy\stats\_stats_py.py:4
424: ConstantInputWarning:
```

An input array is constant; the correlation coefficient is not defined.

Wartości współczynników korelacji punktowo-dwuseryjnej pokazały, że gdy dane miasto należy do Indii, istnieje duża szansa, że jest bardzo zanieczyszczone częsteczkami PM10. W przypadku, gdy miasto należy do Chin, będzie ono najprawdopodobniej wysokie stężenie PM2.5.

Most and least polluted cities (across all years)

PM2.5

```
In [ ]: grouped_city_pm25 = data.groupby(["WHO Country Name", "City or Locality"])[
    "PM2.5"
].mean()
grouped_city_pm25 = pd.DataFrame(grouped_city_pm25).sort_values(by="PM2.5").dropna()
grouped_city_pm25.index[:10].to_list()
```

```
Out[ ]: [('Portugal', 'Porto Santo'),
          ('Norway', 'Todalen'),
          ('Sweden', 'Bredkalen'),
          ('Australia', 'Emu River'),
          ('Australia', 'St Helens'),
          ('Iceland', 'Suðurnesjabær'),
          ('Iceland', 'Husavik'),
          ('Australia', 'Fingal'),
          ('Norway', 'Birkeland'),
          ('Canada', 'Steeper')]
```

```
In [ ]: grouped_city_pm25.index[-10:].to_list()
```

```
Out[ ]: [('China', 'Hetian Shi'),
          ('Pakistan', 'Rawalpindi'),
          ('China', 'Hotan'),
          ('China', 'Kashi Shi'),
          ('India', 'Delhi'),
          ('India', 'Hapur'),
          ('India', 'Noida'),
          ('India', 'Agra'),
          ('Afghanistan', 'Kabul'),
          ('Cameroon', 'Bamenda')]
```

```
In [ ]: grouped_city_pm25 = data.groupby(["WHO Country Name", "City or Locality"])[
          "PM2.5"]
          ].median()
grouped_city_pm25 = pd.DataFrame(grouped_city_pm25).sort_values(by="PM2.5").dropna()
grouped_city_pm25.index[:10].to_list()
```

```
Out[ ]: [('Portugal', 'Porto Santo'),
          ('Norway', 'Todalen'),
          ('Sweden', 'Bredkalen'),
          ('Australia', 'Emu River'),
          ('Australia', 'St Helens'),
          ('Iceland', 'Suðurnesjabær'),
          ('Australia', 'Fingal'),
          ('Canada', 'Steeper'),
          ('Canada', 'Labrador City'),
          ('Iceland', 'Husavik')]
```

```
In [ ]: grouped_city_pm25.index[-10:].to_list()
```

```
Out[ ]: [('China', 'Hetian'),
          ('China', 'Hetian Shi'),
          ('Pakistan', 'Rawalpindi'),
          ('China', 'Hotan'),
          ('India', 'Agra'),
          ('India', 'Delhi'),
          ('India', 'Hapur'),
          ('India', 'Noida'),
          ('Afghanistan', 'Kabul'),
          ('Cameroon', 'Bamenda')]
```

Podobnie jak w przypadku analizy na podstawie krajów, tak i tutaj przedstawiono najbardziej i najmniej zanieczyszczone miasta badając i średnią i medianę zanieczyszczeń. Widać, że najczystsze pod względem PM2.5 są miasta na północy Europy oraz wybrane lokalizacje w Australii i Kanadzie. Najbardziej zanieczyszczone są miasta chińskie, indyjskie oraz Bamenda w Kamerunie, stolica Afganistanu - Kabul oraz pakistańskie Rawalpindi.

PM10

```
In [ ]: grouped_city_pm10 = data.groupby(["WHO Country Name", "City or Locality"])[
    "PM10"
].mean()
grouped_city_pm10 = pd.DataFrame(grouped_city_pm10).sort_values(by="PM10").dropna()
grouped_city_pm10.index[:10].to_list()
```

```
Out[ ]: [('Switzerland', 'Jungfraujoch'),
('Iceland', 'Suðurnesjabær'),
('Norway', 'Todalen'),
('Finland', 'Muonio'),
('Sweden', 'Bredkalen'),
('Germany', 'Garmisch-Partenkirchen'),
('Norway', 'Hurdal Municipality'),
('Bahamas', 'Nassau'),
('Norway', 'Birkeland'),
('Sweden', 'Malmberget')]
```

```
In [ ]: grouped_city_pm10.index[-10:].to_list()
```

```
Out[ ]: [('Egypt', 'Greater Cairo'),
('Bahrain', 'Ras Hayan'),
('Iran (Islamic Republic of)', 'Boshehr'),
('India', 'Jharia'),
('Bahrain', 'Hamad Town'),
('Saudi Arabia', 'Al Jubail'),
('Iraq', 'Basra'),
('Pakistan', 'Rawalpindi'),
('Iran (Islamic Republic of)', 'Zabol'),
('Pakistan', 'Peshawar')]
```

```
In [ ]: grouped_city_pm10 = data.groupby(["WHO Country Name", "City or Locality"])[
    "PM10"
].median()
grouped_city_pm10 = pd.DataFrame(grouped_city_pm10).sort_values(by="PM10").dropna()
grouped_city_pm10.index[:10].to_list()
```

```
Out[ ]: [('Switzerland', 'Jungfraujoch'),
('Iceland', 'Suðurnesjabær'),
('Norway', 'Todalen'),
('Finland', 'Muonio'),
('Sweden', 'Bredkalen'),
('Germany', 'Garmisch-Partenkirchen'),
('Norway', 'Hurdal Municipality'),
('Bahamas', 'Nassau'),
('Sweden', 'Malmberget'),
('Italy', 'Alpe Devero')]
```

```
In [ ]: grouped_city_pm10.index[-10:].to_list()
```

```
Out[ ]: [('Bahrain', 'Ras Hayan'),
('Iran (Islamic Republic of)', 'Boshehr'),
('Egypt', 'Greater Cairo'),
('India', 'Jharia'),
('Bahrain', 'Hamad Town'),
('Saudi Arabia', 'Al Jubail'),
('Iraq', 'Basra'),
('Pakistan', 'Rawalpindi'),
('Iran (Islamic Republic of)', 'Zabol'),
('Pakistan', 'Peshawar')]
```

Dla PM10 wyniki prezentują się podobnie, dodając do "czystych miast" Jungfraujoch w Szwajcarii, Garmisch-Partenkirchen w Niemczech oraz Nassau na Bahamach. Najbardziej

zanieczyszczone są miasta leżące całkiem blisko siebie w Azji Południowo-Wschodniej oraz Egipcie.

NO2

```
In [ ]: grouped_city_no2 = data.groupby(["WHO Country Name", "City or Locality"])["NO2"].mean()
grouped_city_no2 = pd.DataFrame(grouped_city_no2).sort_values(by="NO2").dropna()
grouped_city_no2.index[:10].to_list()
```

```
Out[ ]: [('Canada', 'Aylesford'),
          ('Canada', 'Southampton'),
          ('Switzerland', 'Jungfraujoch'),
          ('Norway', 'Hemnes'),
          ('Sweden', 'Bredkalen'),
          ('Norway', 'Todalen'),
          ('Austria', 'Rauris'),
          ('Italy', "Sant'eufemia A Maiella"),
          ('Italy', 'Seulo'),
          ('Finland', 'Muonio')]
```

```
In [ ]: grouped_city_no2.index[-10:].to_list()
```

```
Out[ ]: [('Mexico', 'Zona Metropolitana Del Valle De Toluca'),
          ('Brazil', 'Belo Horizonte'),
          ('Kuwait', 'Ali Subah Al-Salem'),
          ('Iran (Islamic Republic of)', 'Tehran'),
          ('Iran (Islamic Republic of)', 'Varamin'),
          ('Iran (Islamic Republic of)', 'Orumye'),
          ('Iran (Islamic Republic of)', 'Pakdasht'),
          ('Iran (Islamic Republic of)', 'Hamedan'),
          ('Iran (Islamic Republic of)', 'Yasoj'),
          ('Iran (Islamic Republic of)', 'Arak')]
```

```
In [ ]: grouped_city_no2 = data.groupby(["WHO Country Name", "City or Locality"])[
          "NO2"]
          ].median()
grouped_city_no2 = pd.DataFrame(grouped_city_no2).sort_values(by="NO2").dropna()
grouped_city_no2.index[:10].to_list()
```

```
Out[ ]: [('Canada', 'Aylesford'),
          ('Canada', 'Southampton'),
          ('Switzerland', 'Jungfraujoch'),
          ('Norway', 'Hemnes'),
          ('Sweden', 'Bredkalen'),
          ('Norway', 'Todalen'),
          ('Austria', 'Rauris'),
          ('Italy', "Sant'eufemia A Maiella"),
          ('Italy', 'Seulo'),
          ('Finland', 'Muonio')]
```

```
In [ ]: grouped_city_no2.index[-10:].to_list()
```

```
Out[ ]: [('Mexico', 'Zona Metropolitana Del Valle De Toluca'),
          ('Brazil', 'Belo Horizonte'),
          ('Kuwait', 'Ali Subah Al-Salem'),
          ('Iran (Islamic Republic of)', 'Varamin'),
          ('Iran (Islamic Republic of)', 'Tehran'),
          ('Iran (Islamic Republic of)', 'Orumye'),
          ('Iran (Islamic Republic of)', 'Pakdasht'),
          ('Iran (Islamic Republic of)', 'Hamedan'),
          ('Iran (Islamic Republic of)', 'Yasoj'),
          ('Iran (Islamic Republic of)', 'Arak')]
```

W przypadku NO₂ także przodują miasta północy, dodając do nich austriackie Rauris, a także włoskie Sant'eufemia A Maniella i Seulo. Najbardziej zanieczyszczone są praktycznie wszystkie miasta Iranu oraz Ali Subah Al-Salem w Kuwejcie, Belo Horizonte w Brazylii oraz Zona Metropolitana Del Valle De Toluca w Meksyku.

Biggest progress and regress in air pollution

PM2.5

```
In [ ]: grouped = data.groupby([
    "WHO Country Name", "City or Locality", "Measurement Year"
]).mean(numeric_only=True)
grouped_pm25 = grouped.dropna(subset="PM2.5")

earliest = grouped_pm25.groupby(level=(0, 1))["PM2.5"].first()
latest = grouped_pm25.groupby(level=(0, 1))["PM2.5"].last()

difference = earliest - latest # + --> progress; - --> regres
difference
```

Out[]: 10 entries per page

Search:

WHO Country Name	City or Locality	PM2.5
Afghanistan	Kabul	0
Albania	Durres	0
Albania	Korce	1.7
Albania	Vlore	0
Albania	Vrith	0
Algeria	Algiers	0
Argentina	Buenos Aires	-0.15
Australia	Adelaide	1.16
Australia	Albury	0
Australia	Armidale	0

Showing 1 to 10 of 4,038 entries

« < 1 2 3 4 5 ... 404 > »

```
In [ ]: difference.sort_values()[:10]
```

Out[]:

WHO Country Name	City or Locality	PM2.5
Pakistan	Lahore	-55.88
Philippines	Nia Road	-31
China	Xinfu District, Xinzhou	-26.45
Saudi Arabia	Yanbu	-25
China	Weifang	-24.4
India	Singrauli	-24.34
Madagascar	Antananarivo	-24
China	Hetian	-24
China	Hotan	-23.92
China	Zhanhe Qu	-23.16

In []: difference.sort_values()[-10:]

Out[]:

WHO Country Name	City or Locality	PM2.5
China	Weibin Qu	44.93
China	Shijiazhuang	47.33
China	Hengshui	47.34
China	Beijing	48.5
South Africa	Waterberg	51.56
Bangladesh	Barisal	58.23
China	Xingtai	59
China	Baoding	59.67
Bangladesh	Rajshahi	64.05
Iran (Islamic Republic of)	Zabol	70.61

Przechodząc do analizy progresu i regresu, widać, że w przypadku PM2.5 po obu stronach barykady znajdują się miasta chińskie, w tym sama stolica. Największy postęp dokonał się w irańskim Zabol, a zwiększenie zanieczyszczenia w pakistańskim Lahore.

PM10

In []: grouped_pm10 = grouped.dropna(subset="PM10")

```

earliest = grouped_pm10.groupby(level=(0, 1))["PM10"].first()
latest = grouped_pm10.groupby(level=(0, 1))["PM10"].last()

difference = earliest - latest # + --> progress; - --> regres

difference

```

Out[]:

10 entries per page

Search:

WHO Country Name	City or Locality	PM10
Albania	Durres	-6.91
Albania	Korce	5.1
Albania	Vlore	-7.46
Albania	Vrith	0
Andorra	Escaldes-Engordany	2.92
Argentina	Buenos Aires	2.37
Australia	Adelaide	-4.11
Australia	Albany	-1.6
Australia	Albury	-7.62
Australia	Armidale	0

Showing 1 to 10 of 4,426
entries

« < 1 2 3 4 5 ... 443 >
»

In []: difference.sort_values()[:10]

Out[]:

WHO Country Name	City or Locality	PM10
Bhutan	Pasakha	-170
Bahrain	Ma'ameer	-147.55
India	Gorakpur	-140.67
India	Singrauli	-109
Egypt	Greater Cairo	-105
Madagascar	Antananarivo	-93
Egypt	Delta Region	-82
India	Imphal	-80
India	Bhopal	-72.04
India	Jodhpur	-71.61

In []: difference.sort_values()[-10:]

Out[]:

WHO Country Name	City or Locality	PM10
North Macedonia	Tetove	89.54
Bangladesh	Khulna	92.49
Kuwait	Al Ahmadi	93.01
Kuwait	Al-Ahmadi	93.01
Mauritius	Beau Bassin/Rose Hill	93.98
Iran (Islamic Republic of)	Gachsaran	97.58
Turkey	Batman	97.65
India	Nanded	118.67
Iran (Islamic Republic of)	Qazvin	125.8
Iraq	Baghdad	151.13

W przypadku PM10 ogromny regres zaszedł w Pasakha w Bhutanie, w Ma'ameer w Bahrajnie i Gorakpurze w Indiach, a progres: w stolicu Iraku, w Qazvin w Iranie oraz w Nanded w Indiach.

NO2

In []: grouped_no2 = grouped.dropna(subset="NO2")

```

earliest = grouped_no2.groupby(level=(0, 1))["NO2"].first()
latest = grouped_no2.groupby(level=(0, 1))["NO2"].last()

difference = earliest - latest # + --> progress; - --> regres

difference

```

Out[]:

10 entries per page

Search:

WHO Country Name	City or Locality	NO2
Albania	Durres	1.85
Albania	Elbasan	-0.74
Albania	Korce	0
Albania	Vlore	0
Albania	Vrith	0
Andorra	Escaldes-Engordany	0.63
Argentina	Buenos Aires	-2.9
Australia	Adelaide	-2.64
Australia	Brisbane	-1.21
Australia	Canberra	0

Showing 1 to 10 of 4,403
entries

« < 1 2 3 4 5 ... 441 >
»

In []: difference.sort_values()[:10]

Out[]:

WHO Country Name	City or Locality	NO2
India	Pune	-48.67
Mexico	Mexico City	-47.39
India	Karimnagar	-43
India	Ramagundam	-43
India	Warangal	-41.5
Iran (Islamic Republic of)	Pakdasht	-36.36
India	Khammam	-35.5
Bangladesh	Barisal	-35.06
India	Delhi	-32.56
Iran (Islamic Republic of)	Ahvaz	-30.26

In []: difference.sort_values()[-10:]

Out[]:

WHO Country Name	City or Locality	NO2
Germany	Leonberg	34.1
Italy	Ferentino	34.13
China	Guangzhou	36
France	La Mulatiere	37.15
Germany	Freiburg Im Breisgau	43.79
Serbia	Novi Sad	47.55
Brazil	Betim	47.98
Iran (Islamic Republic of)	Shiraz	51.73
Japan	Kagoshima City	62.16
Iran (Islamic Republic of)	Varamin	79.12

Dla wartości NO2 zaszły mniejsze zmiany niż dla PM10 oraz PM2.5, ale największy regres zanotowano głównie w miastach indyjskich, a progres w Varamin w Iranie, Kagoszimie w Japonii oraz stolicy Iranu.

Stations analysis

```
In [ ]: summed_by_cities_stations_df = (
    data.groupby(["WHO Country Name", "City or Locality"])["Total number of stations"]
    .sum()
    .reset_index()
)

summed_by_cities_stations_df.sort_values("Total number of stations", ascending=False
                                         :10
)
```

Out[]:

	WHO Country Name	City or Locality	Total number of stations
3609	India	Kolkata	278
6467	United Arab Emirates	Abu Dhabi	113
2308	Czechia	Prague	105
423	Brazil	Rio De Janeiro	86
434	Brazil	Sao Paulo	86
3561	India	Hyderabad	86
3485	India	Chennai	81
3510	India	Delhi	77
5615	South Africa	Gert Sibande	77
3590	India	Kanpur	76

Tutaj zbadano, w których miastach na świecie znajduje się najwięcej stacji pomiarowych wedle dostępnych danych. Jak widać najwięcej jest ich w Kalkucie, Abu Dhabi oraz Pradze.

Air pollution map

```
In [ ]: # from geopy.geocoders import Nominatim
# from geopy.exc import GeocoderTimedOut

# geolocator = Nominatim(user_agent="EDA")

# data_Lon_Lat = data.copy()

# cities_list = data_Lon_Lat["City or Locality"].unique()
# cities_df = pd.DataFrame({"City": cities_list, "Longitude": None, "Latitude": None}

# def get_coordinates(city):
#     try:
#         location = geolocator.geocode(city)
#         if location:
#             return location.Longitude, location.Latitude
#         else:
#             return None, None
#     except GeocoderTimedOut as e:
#         print("Error: geocode failed on input %s with message %s" % (city, e.message))

# cities_df[["Longitude", "Latitude"]] = cities_df["City"].apply(get_coordinates)
```

```
#     Lambda x: pd.Series(get_coordinates(x))
# )
# cities_df
```

W celu wykonania mapy miast i zanieczyszczeń próbowało zmapować miasta na współrzędne geograficzne, co ostatecznie się nie udało ze względu na zbyt dużą liczbę zapytań (zbyt dużą liczbę miast w analizowanym zbiorze danych).

W związku z tym wspomóżono się plikiem dostępnym pod tym adresem:

<https://simplemaps.com/data/world-cities>

Get cities coords

```
In [ ]: cities_gdf = gpd.read_file("../data/worldcities.csv")

data_lat_lng = data.merge(cities_gdf, left_on="City or Locality", right_on="city")
print(len(data_lat_lng["City or Locality"].unique()))
print(len(data["City or Locality"].unique()))
data_lat_lng
```

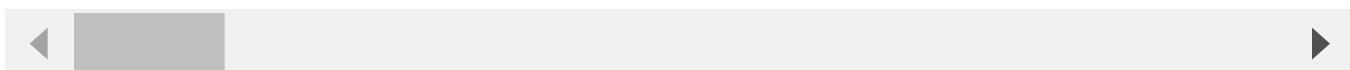
2535

6874

Out[]: 10 entries per page

	WHO Region	ISO3	WHO Country Name	City or Locality
0	Eastern Mediterranean Region	AFG	Afghanistan	Kabul
1	European Region	ALB	Albania	Elbasan
2	European Region	ALB	Albania	Elbasan
3	European Region	ALB	Albania	Elbasan
4	European Region	AND	Andorra	Escaldes-Engordany
5	European Region	AND	Andorra	Escaldes-Engordany
6	European Region	AND	Andorra	Escaldes-Engordany
7	European Region	AND	Andorra	Escaldes-Engordany
8	European Region	AND	Andorra	Escaldes-Engordany
9	European Region	AND	Andorra	Escaldes-Engordany

Showing 1 to 10 of 315 entries (downsampled) from 17,602x26 to 315x26 as maxBytes=65536)



W pliku tym znajdują się wszystkie stolice państw oraz stanów/województw (prawie połowa z analizowanych), zatem uznano ten zbiór za reprezentatywny.

PM2.5

```
In [ ]: grouped_city_pm25 = data_lat_lng.groupby(
    ["WHO Country Name", "City or Locality", "lat", "lng"]
```

```

)[ "PM2.5"].mean()
df_grouped_city_pm25 = grouped_city_pm25.reset_index().dropna()

fig = px.scatter_geo(
    df_grouped_city_pm25,
    lat=df_grouped_city_pm25["lat"],
    lon=df_grouped_city_pm25["lng"],
    color="PM2.5",
    color_continuous_scale=px.colors.sequential.Sunsetdark,
    hover_name="City or Locality",
)

fig.update_layout(
    title_text="Average concentration of PM2.5 across all years",
)
fig.show()

```

Utworzona mapa zanieczyszczeń poszczególnych miast pokazuje niektóre miasta Afryki oraz całą Azję Południową jako najbardziej zanieczyszczone, a Amerykę Północną, Australię oraz Europę Północną jako te z najlepszym powietrzem pod względem PM2.5.

PM10

```

In [ ]: grouped_city_pm10 = data_lat_lng.groupby(
    ["WHO Country Name", "City or Locality", "lat", "lng"]
)[ "PM10"].mean()
df_grouped_city_pm10 = grouped_city_pm10.reset_index().dropna()

fig = px.scatter_geo(
    df_grouped_city_pm10,
    lat=df_grouped_city_pm10["lat"],
    lon=df_grouped_city_pm10["lng"],
    color="PM10",
    color_continuous_scale=px.colors.sequential.Sunsetdark,
    hover_name="City or Locality",
)

fig.update_layout(
    title_text="Average concentration of PM10 across all years",
)
fig.show()

```

Mapa PM10 prezentuje się analogicznie do mapy z PM2.5. Tutaj jedynie ze względu na dostępność danych o PM10 oraz o miastach, to głównie Indie i Pakistan są wskazane jako najbardziej zanieczyszczone.

NO2

```

In [ ]: grouped_city_no2 = data_lat_lng.groupby(
    ["WHO Country Name", "City or Locality", "lat", "lng"]
)[ "NO2"].mean()
df_grouped_city_no2 = grouped_city_no2.reset_index().dropna()

fig = px.scatter_geo(
    df_grouped_city_no2,
    lat=df_grouped_city_no2["lat"],

```

```
    lon=df_grouped_city_no2["lng"],  
    color="NO2",  
    color_continuous_scale=px.colors.sequential.Sunsetdark,  
    hover_name="City or Locality",  
)  
  
fig.update_layout(  
    title_text="Average concentration of NO2 across all years",  
)  
  
fig.show()
```

W przypadku NO2 nie można określić konkretnego regionu jako najbardziej zanieczyszczonego, może oprócz Teheranu w Iranie. Tutaj niestety wartości NO2 są wszędzie podobnie wysokie.

Na pytania 1-4, a także częściowo kolejne, wypowiedziano się w trakcie analizy. Poniżej odpowiedzi na pytania 5-7.

1. Na podstawie przeprowadzonej analizy widać wyraźnie dodatnią zależność pomiędzy wartościami PM2.5 oraz PM10, co jest zgodne z codzienymi obserwacjami - zanieczyszczenia ta powstają w wyniku emisji spalin samochodowych oraz paleniu w piecach, zatem oba są produkowane w mniej więcej podobny sposób. Ponadto można wskazać rejony Bliskiego Wschodu oraz Azji Południowo-Wschodniej jako te najbardziej zanieczyszczone. Co ważne, często (choć nie zawsze) są to regiony biedne, kraje rozwijające się, co można dość łatwo połączyć. W takich miejscowościach panuje zdecydowanie mniejsza świadomość ekologiczna, jest mniejszy współczynnik ludzi wykształconych, a żyjący tam ludzie mają często po prostu inne priorytety. W związku z tym regiony te od lat są zanieczyszczone, co koreluje się z niskim poziomem życia.
2. Ten rodzaj danych może mieć jedynie bardzo ogólne znaczenie jako dane wejściowe do modelu ML. Dane te są zagregowane do roku, co jest dość dużym stopniem generalizacji, ponadto występuje tu wiele braków i można by przewidzieć jedynie ogólne trendy, tzn. które miejsca ogólnie staną się jeszcze bardziej, lub też nieco mniej zanieczyszczone. Należałyby dodać choćby dane miesięczne, aby np. móc predykować w jakich okresach roku należy nosić maseczki antysmogowe w danych miastach, czy też wprowadzać inne obostrzenia. Niewątpliwie jednak dane te mogą pomóc wytworzyć pewien ogólny model zanieczyszczeń.
3. Przedstawiona eksploracyjna analiza danych pokazuje jak wiele można uzyskać z praktycznie surowych danych. Jesteśmy w stanie zobaczyć trendy zanieczyszczeń, ocenić najbardziej oraz najmniej zanieczyszczone miejsca, a także określić relacje między zmiennymi w oparciu o różne poziomy generalizacji. Dowiedzono, że Europa Północna słynie z czystego powietrza, podobnie jak Kanada oraz Australia i Nowa Zelandia. Ponadto przedstawiono rejon Iranu oraz krajów sąsiadujących (także "przez morze") jako te najbardziej skażone dwutlenkiem azotu. Również ukazano, że Indie oraz Chiny są świadome stopnia swojego zanieczyszczenia cząsteczkami PM2.5 oraz PM10 - to tam znajduje się najwięcej czujników jednocześnie wskazujących najwyższe zanieczyszczenie. Widać także, że oba te kraje starają się polepszyć sytuację w niektórych miastach, co przedstawia różnica w zanieczyszczeniach na korzyść ostatnich lat dla wybranych chińskich i indyjskich miast. Tak przeprowadzone EDA, pomaga ocenić, które zmienne

uwzględnić w modelu, oraz który stopień generalizacji przyjąć. Pomaga także w ocenie danych wyjściowych z modelu - po porównaniu ich z tego typu analizą danych historycznych możemy ogólnie ocenić, czy model działa dobrze, czy też nie.